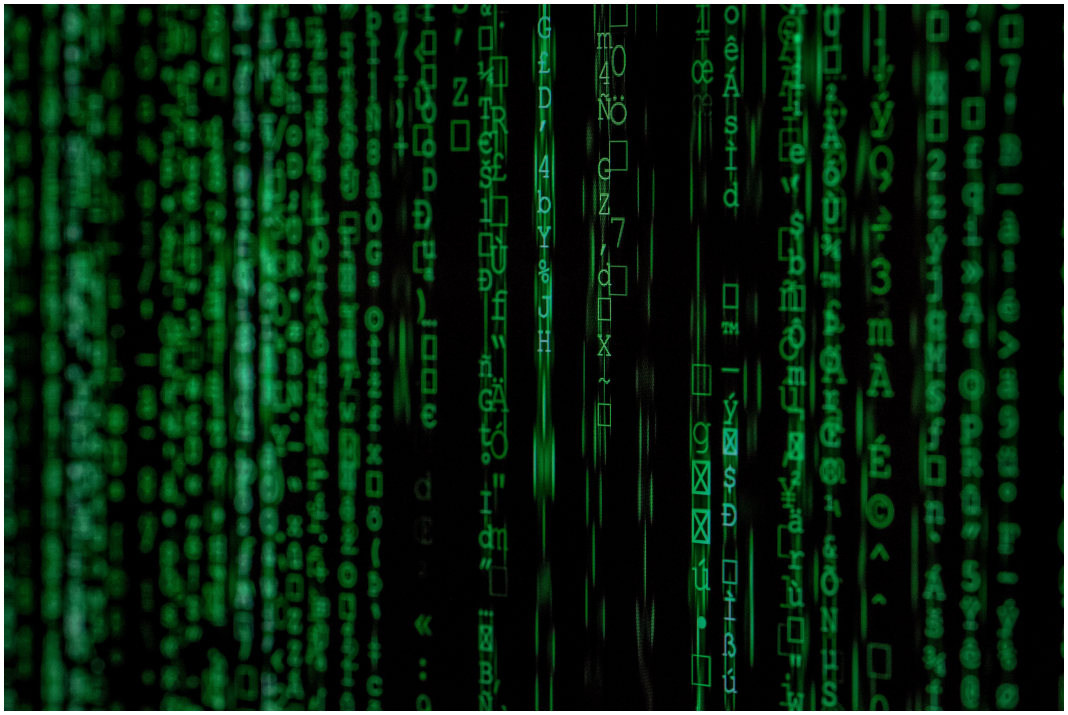


# Quality Attributes of Data in Distributed Deep Learning Architectures

Master's thesis in Computer science and engineering

SHAMEER KUMAR PRADHAN

SAGAR TUNGAL

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2021

Master's thesis 2021

# Quality Attributes of Data in Distributed Deep Learning Architectures

SHAMEER KUMAR PRADHAN
SAGAR TUNGAL

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2021

Quality Attributes of Data in Distributed Deep Learning Architectures
SHAMEER KUMAR PRADHAN
SAGAR TUNGAL



Supervisor: Eric Knauss, Ph.D. and Hans-Martin Heyn, Ph.D. Department of Computer Science and Engineering
Advisor: Olof Eriksson, Stefan Andersson, and Oliver Brunnegard, Veoneer Sweden AB
Examiner: Christian Berger, Ph.D. Department of Computer Science and Engineering

Master's Thesis 2021
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: *Matrix Background* by Markus Spiske (`https://www.pexels.com/@markusspiske`)

Typeset in LaTeX
Gothenburg, Sweden 2021

Quality Attributes of Data in Distributed Deep Learning Architectures
SHAMEER KUMAR PRADHAN, SAGAR TUNGAL
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

# Abstract

Large volume of data is generated by different systems. Intelligent systems such as autonomous driving uses such large volume of data to train their artificial intelligence models. However, good quality data is one of the foremost needs of any system to function in an effective and safe manner. Especially in critical systems such as those related with autonomous driving, quality data becomes sacrosanct as fault in such systems could result in fatal accidents. In this thesis, a Design Science Research is conducted to identify challenges related with data quality of a distributed deep learning system. The challenges are identified by conducing interviews with five experts from autonomous driving domain as well as through literature review. The challenges and their severity are validated using a survey. After identification of the challenges, five artifact components are developed that relate with assessing and improving data quality. The artifact components include *Data Quality Workflow*, *List of Challenges*, *List of Data Quality Attributes*, *List of Data Quality Attribute Metrics*, and *Potential Solutions*. The abstract artifact components and concrete implementation of those components are devised and validated using second round of interviews. In the third iteration of this study, the final artifact components are validated through a focus group session with experts and survey. Furthermore, the artifact also presents the information regarding which challenges affect which data quality attributes. This association between challenges and attributes are also validated in the focus group session. The results depict that most of the challenge - attribute association presumed by the researchers of this thesis are valid. Similarly, the templates developed for the artifact components are regarded as appropriate as well. A contribution of this thesis study towards the body of software engineering and requirements engineering research is the comprehensive and unified "Data Quality Assessment and Maintenance Framework" developed as a series of artifact components in this thesis. This framework can be used by researchers and practitioners to improve processes related with data quality as well as enhance data quality of the systems they develop.

Keywords: Data quality, Data, Data quality attributes, Data quality challenges, Data quality workflow, Data quality assessment, Data quality maintenance, Design science research, Artifacts, Template, Deep learning, Distributed architecture, Distributed deep learning architecture, Advanced driver assistance systems

# Acknowledgements

I would like to thank many people who have supported me in my endeavor to complete this thesis. First, I would like to provide my heartfelt gratitude to my supervisors, Eric and Hans-Martin, in the academic side and Olof, Stefan, and Oliver, in the industrial side at Veoneer. This thesis would not have been successful without their constant guidance and feedback. Interview, survey, and focus group participants also deserve my gratification. I'd also like to express my gratitude to my family, including my parents, guardians, my sister, and P, for the love, dedication, and support shown towards me. They have been by my side night and day through thick and thin. Similarly, I'd like to thank my thesis partner and pal Sagar for collaborating with me in the thesis. Finally, I'd like to thank Chalmers for proving me the opportunity to pursue a master's degree in the department. Ad infinitum. Godspeed.

Shameer Kumar Pradhan, Gothenburg, September 15, 2021

I would like to extend my sincere gratitude to my academic supervisors Eric Knauss and Hans-Martin Heyn for their constant support and guidance through out this thesis project. I would like to thank my industry supervisors Olof Eriksson, Stefan Andersson, and Oliver Brunnegard for giving the opportunity to carry out thesis work at Veoneer AB and providing timely feedback. I would like to thank all the interview and survey participants for sharing their insights and expertise for conducting the study. I would also like to thank my friend and thesis partner Shameer Kumar Pradhan for great collaboration.
Finally, I would like to thank my parents for all the love, support, and encouragement through out this journey.

Sagar Tungal, Gothenburg, September 2021

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| $ACC$ | Adaptive Cruise Control |
| $AD$ | Autonomous Driving |
| $ADAS$ | Advanced Driver-Assistance System |
| $AEB$ | Automatic Emergency Braking |
| $AI$ | Artificial Intelligence |
| $ANN$ | Artificial Neural Network |
| $CNN$ | Convolutional Neural Network |
| $DQAMF$ | Data Quality Assessment and Maintenance Framework |
| $DSR$ | Design Science Research |
| $ISA$ | Intelligent Speed Adaptation |
| $LDWS$ | Lane Departure Warning System |
| $ML$ | Machine Learning |
| $OEM$ | Original Equipment Manufacturer |

# 1

# Introduction

Almost all software systems are designed to take in input data, process it, and produce output. According to International Data Corporation (2020), the global installed base of storage capacity is 6.8 zettabytes in 2020. The size is expected to grow by 17.8% annually over the next five years. Memon et al. (2017) identifies a number of applications of big data such as in agriculture, banking, education, chemistry, etc. Systems based on machine learning and deep learning require a large amount of data for training the algorithms.

Critical distributed deep learning applications such as ADAS rely on a large amount of data generated by a number of sensors. ADAS, which is a part of AD is designed to make driving comfortable and safe by assisting the driver make the right decisions (Ziębiński et al. 2017). It helps during situations such as overtaking other vehicles, parking, and detecting obstacles and slippery roads that might go unnoticed to the naked eye. In addition, ADAS can also, independently from the driver, mitigate potentially dangerous situation. Examples include automatic emergency braking systems or lane change support with intervention. To enable all these functions, ADAS receives a significant amount of data from several sources for analysis. Since the driver's decision during certain situations would be based on analysis performed by ADAS, it is very important that the data captured is trustworthy, timely, and of the required quality to make the right decision. A lack of quality data might compromise the decision-making capabilities of the driver in AD, which can result in a fatal accident.

In this masters thesis, the researchers try to establish a framework that serves as a benchmark and helps the stakeholders to have a single-point of reference for the right data quality attributes and challenges associated with them in a distributed deep learning system. The framework helps the stakeholders to identify challenges pertaining to data quality, data quality attributes affected by those challenges, and data quality metrics associated with the attributes. The framework is devised with ADAS as the reference application.

## 1.1 Statement of the Problem

Distributed deep learning systems such as ADAS would need a large amount of data for analysis. The data is gathered via different devices and sensors such as lidar, satellite crash sensors, night vision sensor, radar, vision systems, etc. in the context of AD and other sensors in a different context. The data is fed into an electronic control unit for analysis. The distributed deep learning system associated with the collaborating case company uses four categories of data namely driver data, vehicle data, surround data, and global data. There are numerous sub-types of those data as well. For instance, surround data can include the distance between the vehicle and a nearby object. Furthermore, some data could be collected and stored for future use while others might be required urgently for the development of autonomous driving functionalities.

As mentioned in Section 2.1, there are are a number of data management procedures that can be followed. The procedures can be implemented in different applications such as those mentioned in Section 2.2. There are few frameworks and procedures developed for data quality as mentioned in Section 2.3. Although data is important for effective analysis, there is no proper procedure to determine and manage the quality of the data. There is a need of a framework for defining relevant data quality attributes for the kinds of data that various hardware and sensors collect in a deep learning system. There should be a workflow for data quality assessment. Currently, most of the information regarding data quality assessment for the distributed deep learning system at hand is based on the expertise of the people employed by the case company. There is no central repository of such information. This results in a number of challenges, which in turn affects data quality in a negative fashion.

Problem statement: No general data quality assessment model and an apt approach for systematic management of the quality of input data for distributed deep learning system for autonomous driving exists currently.

## 1.2 Statement of Purpose

The aim of this study is to comprehend the data quality requirements in a distributed deep learning system such as ADAS and develop an artifact that assists in data quality assessment. Deep learning systems can be implemented in highly-critical applications like AD. A slight divergence from the standard can mean the difference between a safe journey and a fatal accident. Primacy of quality of data fed into the central controlling unit cannot be overstated. Therefore, by using the artifact proposed in the thesis, relevant parties shall be able to understand the attributes that are important for maintaining quality of data and the challenges affecting data quality. Not only should they be able to understand the quality of the data independently, it should also be possible for them to understand the co-dependencies between data quality attributes and data quality challenges.

## 1.3 Case Company

This thesis is produced in association with Veoneer Sweden AB (NYSE: VNE, Nasdaq Stockholm: VNE SDB), an automotive technology company spun off from Autoliv in 2018. It is headquartered in Stockholm, Sweden. As of the writing of this thesis, Veoneer has 7,500 employees spread across 11 countries with 6 manufacturing sites and 22 technical centers. Veoneer is focused on advanced driver-assistance with the purpose to create trust in mobility. The company designs, manufactures, and sells software and hardware systems for occupant protection, ADAS, and collaborative and automated driving to OEMs. They work with cutting edge technologies like vision systems, radar, lidar, thermal sensing, electronic controls, and human-machine interface. The company has clocked more than a billion revenue. This thesis aims to augment the data quality assessment procedure in organizations such as Veoneer in development of distributed deep learning systems such as ADAS.

## 1.4 Research Questions

The principal goals of this study are to understand the needs of data quality for distributed deep learning systems, recognize the challenges related to data quality, and devise an artifact that assist in data quality assessment and solution. These goals can be fulfilled by answering the following research questions of the thesis.

**Research Question 1 (RQ1)**: What are relevant challenges of managing data quality requirements when developing large systems based on distributed deep learning?

Answering RQ1 would help identify the challenges pertaining to data quality. Identification of such challenges can, in turn, help devise solutions for those challenges.

**Research Question 2 (RQ2)**: What constitutes a data quality framework for developing large systems based on distributed deep learning?

RQ2 would help devise a series of components for a data quality framework. The goal of this framework is to help researchers and practitioners determine data quality requirements including data quality challenges, data quality attributes, metrics, and solutions to those challenges.

**Research Question 3 (RQ3)**: To what extent can relevant challenges of managing data quality requirements be mitigated by a data quality framework for developing systems based on deep learning?

Answering RQ3 would make sure that the developed framework is effective in managing the data quality requirements and mitigate the challenges associated with them.

## 1.5 Scope and Limitations

The study was mostly carried out carried out mostly based on the expert interviews and literature review. However, the researchers have interviewed experts from AD domain only. Additionally, most of the interviewees are from the same company — Veoneer — and only two are from different companies. Even then, one of those companies is highly related with the case company. The researchers only had access to a limited number of interviewees. Even though a broad section of experts were interviewed from various teams (fields), there still is a chance that experts from other fields like data management, data collection, data labelling, as well as vehicle owners, etc. are missed.

The data collected is mostly from past experiences of the experts and does not necessarily take into consideration future data quality challenges and attributes which can evolve. A generic framework for data quality will be a useful tool in system design process. However, due to the limitations mentioned above, the outcome of the study might be inclined towards automotive systems and hence, may need further development to cover more applications.

The scope of the study is limited to establishing a data quality framework; and hence, does not relate to individual data types produced by individual sensors. The researchers look into data quality requirements by first looking into data quality attributes and data quality challenges. The study tries to identify how data quality challenges affect the data quality attributes and try to fix the challenges by devising solutions. When the challenges are reduced or mitigated, it improves the quality of data attributes which in turn improves overall data quality of a system.

# 2

# Background and Related Work

In this chapter, the previous research and concepts related to the study are discussed. The various literature in connection to the study topic are reviewed. The important information related to the study is highlighted along with the contribution of this thesis to the study topic. In Section 2.1 the various techniques and processes of data management are discussed. In Section 2.2, the concept of deep learning is discussed and how data is helpful in the success of deep learning systems are presented. Finally, in Section 2.3, the need for good data quality and various data quality metrics are presented.

For background study, online research paper search was performed with terms such as *procedures of data management*, *data quality attributes*, *data quality metrics*, *data quality framework*, *applications of deep learning systems*, and *data in deep learning systems*. The criteria employed to source the research papers were finding recent publications (although there are few exceptions) and referring those papers that have been cited by many other papers.

## 2.1 Procedures of Data Management

Michener (2015) devises set of ten "simple rules" that a data management plan should follow. They include rules relating to *determination of requirements*, *identifying the required data*, *data organization*, *data documentation*, *data quality assurance*, *data storage and preservation*, *project data policies*, *data dissemination*, *roles and responsibilities*, and *budgeting*. According to Michener (2015), the plan provides "logical and comprehensive" method of managing data in a system. He goes on to say that a plan should be a "living document" and hence, should be updated periodically.

The processes associated with data and the tasks performed by development team can be scattered among the three stages of the "step-by-step checklist" of data management proposed by Tavakoli et al. (2006). The three stages are, namely, *preparatory* stage, *data organization* stage, and *analysis and dissemination* stage. During the *preparatory* stage, the project is initiated, requirements are defined, data collection procedure is set, and personnel are trained. Similarly, during *data organization* stage, data collection, data entry, data manipulation, and backup are performed.

Similarly, baseline data analysis and archiving are done during the final stage of *analysis and dissemination*.

Hu et al. (2014) devise a big data lifecycle "value chain" that includes *data generation*, *data acquisition*, *data storage*, and *data analytics*. A number of sources are used to generate the data in a big data application. For instance, in the context of ADAS, different sensors such as vision sensors, lidar, radar, and ultrasonic sensors are used to generate data (Kukkala et al. 2018). As per Hu et al. (2014), data acquisition involves collection, transmission, and preprocessing of data. The data, once collected, have to be stored in some long-term storage systems for future referencing. Finally, analytical models can be developed and applied on the data to get insight regarding certain topics.

Another study suggests a "logical stream" of data quality management (Sun & Wang n.d.). The stream consists of five parts that flow one after the other. They include *data production*, *data process*, *data storage*, *data sharing*, and *data use*. According to the authors of the paper, all of the "living periods" of data should include the five parts of the stream.

According to Laudon & Laudon (2009), any information system needs to reflect over three dimensions, namely, the *people*, the *technology*, and the *organizational processes*. By focusing on these dimensions, the effectiveness of an information system can be improved. *People* dimension relates to the humans that devise, develop, test, and use the system; and their training, ability, and cognition to perform those activities. *Organization processes* dimension involves the activities done by organizations and people to achieve a certain goal. In the context of data management, it could include tasks such as data collection, requirements identification, etc. Lastly, tools and technologies used to collect, store, transfer, process, and analyze the data can be considered as part of the *Technology* dimension. Furthermore, machine learning algorithms and neural networks for deep learning can be considered as part of the *Technology* dimension as well (Rogério & Hirama 2015).

Prasad et al. (2011) propose a data cleansing workflow that traverses from *investigation*, *standardization*, *matching*, and *survivorship*. Data quality requirements are identified during the *investigate* stage of the workflow. Rule sets are customized, and data is transformed in a uniform format during *standardization*. Similar data are identified in the *matching* stage. During *survivorship*, the customer "decides which data to be retained after deduplication" (Juddoo 2015).

*Balanced Scorecard* is a tool that allows to reflect upon from four perspectives – *financial*, *customer*, *internal business*, and *learning and growth* (Zizlavsky 2013). The *Balanced Scorecard* is used to measure the performance of any business or project in terms of a set of performance metrics developed from an organization's vision and strategy (Gawankar et al. 2015).

The Open Measured Data Management Working Group (2021) has developed a vendor-neutral platform called OpenMDM to manage measured data. This platform

is primarily used by automotive companies to build in-house applications. It can 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 the descriptions.

## 2.2 Data in Deep Learning Systems

In recent years, data is playing a major role in decision making in almost every walk of life. Chen et al. (2009), in their article, define data as "computerized representations of models and attributes of real or simulated entities." Data can be of various forms, structures, numbers, pictures, which can be collected or recorded from the real or simulated environment. Data are "recorded (captured and stored) symbols and signal readings", where symbols comprise of words, numbers, diagrams, and images whereas signals include sensor readings (Liew 2007).

Shrestha & Mahmood (2019) describe deep learning as a subset of machine learning, that has significant impact in areas like healthcare, autonomous vehicles, and sentiment prediction. Deep learning utilizes nodes and networks that resemble a human brain. It also enables unsupervised learning from large unstructured or unlabeled data sets. Feature extraction in deep learning algorithms is automated as they have a layered architecture of data representation that is inspired by the working of human brain. The extraction of the high- and low-level features is done at the last layers and lower layers respectively (Pouyanfar et al. 2018).

Data plays an important role in the success of deep learning learning algorithms (Roh et al. 2019). Huge amount of data is collected from various sources and fed to deep learning systems. The systems analyze and make proper decision based on the information that is carried by data. As data grows, management of data and its quality becomes a challenge. Issues pertaining to collecting, processing, analyzing, sharing, and deploying data sets have become prevalent. There has been limited research and study towards data quality, though deep learning models are widely used in various applications (Raj et al. 2019). To harvest information from large non-traditional data is not easy. Therefore, it needs advanced technologies and interdisciplinary teams working in close collaboration (Chen & Lin 2014).

ADAS is one of the applications in which distributed deep learning architecture is used to make proper decisions and assist the driver in various situations. Deep learning models like ANN and CNN have proven to be highly beneficial solutions for the complex ADAS tasks like ACC, LDWS, ISA, etc (Borrego-Carazo et al. 2020). ADAS helps in safety and security of the occupants of the vehicle as well as pedestrians and other vehicles. Guda et al. (2018) propose ADAS application using deep learning that include features like drowsiness detection, traffic sign detection, etc. For deep learning systems to make appropriate decisions, the data fed to them must be trustworthy, timely, and of the required quality. Any deviation from the data quality could lead to bad training of the deep learning systems which could, in turn, lead to accidents and loss of life.

Esteva et al. (2019) in their article discuss the benefits of deep learning in the health care domain. They further state that, Computer Vision in health care is helpful in identifying malignant tumors in the patient's radiograph. Tasks like object classification, detection, and segmentation are handled by Computer Vision. Deep learning helps in predictive analysis of the disease and enables physicians to make better decisions on the treatment (Muniasamy et al. 2019).

## 2.3 Data Quality Attributes and Metrics

The literature reviewed for the following section pertain to data quality attributes and metrics associated with them. First, data quality is defined. Then, need of good data quality is presented for systems such as ADAS and for machine learning algorithms in general. After that, data quality from a perspective of software engineering is provided. Review of literature that study the requirements for data quality and challenges pertaining to data quality is provided next. This is followed by review of literature associated with data quality frameworks. Finally, reporting mechanism of data quality is presented.

Earley & Henderson (2017) define data quality as "the planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meet the needs of data consumers." Deep learning systems such as ADAS require a large volume of different types of data gathered via various sensors. ADAS needs to process the gathered data and make decisions. In functions such as AEB, the decision has to be an appropriate one. Such decisions can only be made if the input data is of good quality.

Sessions & Valtorta (2006), in their research paper, study the effects of data quality on machine learning algorithms. They start with an assumption that quality of data is necessary to produce a more accurate result and then, set out to verify it. They also develop procedures of creating "more robust and useful" algorithms by using data quality assessments. They regard accuracy as the primary measure of data quality in their study. They develop three Bayesian networks to determine the importance of accurate data in Prototypical Constraint-based (PC) algorithm. By analyzing the results of the tests in the Bayesian networks, they determine that data quality has "an enormous effect on the results and efficiency of Bayesian network learning algorithms" Sessions & Valtorta (2006).

Bobrowski et al. (1970) study data quality from a perspective of software engineering. They stress the significance of data quality in design, validation, and implementation of software. They propose three key areas in which data quality activities should focus – *data quality metrics*, *data quality and testing*, and *data quality in software development process*. They conclude the paper by stating that data quality is crucial to the work of software engineers.

Heinrich et al. (2018) provide five requirements for data quality metrics in their research paper. They use a decision-oriented framework, which includes a decision

matrix, to devise the five requirements. They state that data quality metric requirements should make sure that "the metric values can support decision-making under uncertainty." Furthermore, they emphasize that inefficient and impractical metrics should be extracted by managing data quality in "economically oriented" manner. They also test the requirements on metrics such as timeliness, completeness, reliability, correctness, and consistency.

Cai & Zhu (2015) study characteristics of big data environments, present the challenges, and develop a data quality framework in their proceedings paper. They, then, develop a dynamic assessment process for data quality using the framework. They demarcate the quality criteria for big data through the framework using five dimensions, which are subdivided into nine "elements." The elements are further divided into 25 indicators. Following the division, they propose a data quality assessment process, which can be visually presented using a flowchart.

A data quality assessment and monitoring framework is devised in a research by Batini et al. (2007). In the study, they improve upon Basel II operational risk evaluation methods and develop an assessment methodology called ORME-DQ. The assessment methodology is divided into four phases for data quality *risk prioritization*, *risk identification*, *risk measurement*, and *risk monitoring*. Then, they develop a framework consisting of five modules to support the phases of the methodology.

Fletcher (1998) develops a data quality framework for distributed computing environment. He devises two different models, one for data quality and the other for distributed environment. Then, he combines the two models to propose a measure he terms as the "Data Quality Risk Exposure Level" (DQREL). He also discusses the appropriate applications of the DQREL framework.

A paper by Kahn et al. (2015) studies reporting mechanism of data quality in distributed networks. They present the need to have a data quality reporting guideline so that data source can be determined as acceptable. They develop a model that "captures the flow of data from data originator across successive data stewards and finally to the data consumer." They refer clinical practice as an example of the area in which poor data quality posses risk. Alongside the framework, they propose 20 data quality reporting recommendations.

# 3

# Method

In the following chapter, research method employed in this study is discussed in detail. Section 3.1 discusses the concept of DSR method and its usefulness in this study topic. In Sections 3.1.1, 3.1.2, and 3.1.3, various phases of DSR are discussed. The study carried out during each phase of DSR is presented in those sections.

## 3.1 Design Science Research

Vaishnavi & Kuechler (2004), in their research study, state that DSR is focused on creating and evaluating novel and creative artifacts. DSR is applicable in wide range of fields in addition to software, human-computer interaction, and system design methodologies. DSR helps in creating knowledge and devising solutions for existing problems.

In their article, Gacenga et al. (2012) establish a "performance measurement framework" for Information Technology Service Management investments in organizations using DSR. For their research, they collect data via survey questionnaires, literature review, and case studies.

Knauss (2020) argues that DSR is appropriate for those master's thesis that intend to devise solutions for problems with practical relevance. He denotes such thesis as "Constructive Masters Thesis". Industry proposal expects a solution towards data quality in deep learning systems. There has been a limited research and work in this area. The result from the study in this thesis contributes value to business as well as to research in data quality in deep learning systems. Since the goal of this thesis is to devise a solution for understanding data quality and challenges associated with them in a distributed deep learning system, DSR is chosen as the primary methodology for this study.

The process of devising solutions and artifact development in DSR is usually achieved in three iterative cycles as shown in Figure 3.1. The iterative process helps in improving and evolving the artifact in each cycle based on the feedback from experts.

According to Guideline 5 mentioned by Knauss (2020), all research questions were worked upon in each iteration. During the first iteration the focus was on RQ1 for problem identification with the help of interviews and literature review. The sec-

ond iteration addressed RQ2 about developing a framework through brainstorming, literature review, and exchange of ideas within the team. Finally, in the third iteration, the focus was on RQ3 for evaluation of the artifact that was created during the second iteration.

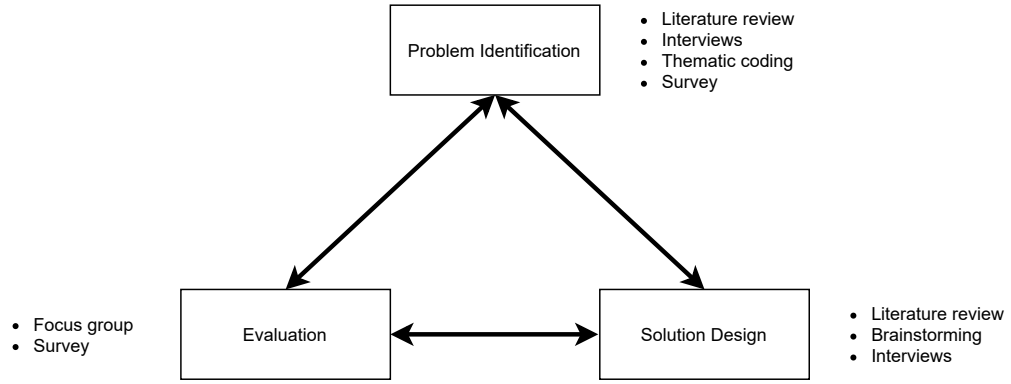



**Figure 3.1:** Stages of Design Science Research

### 3.1.1 Problem Identification

As discussed in Section 3.1, the first iteration begins with problem identification. The focus is on RQ1 to identify relevant challenges of managing data quality requirements when developing large systems based on distributed deep learning. Knauss (2020) argues that during the first cycle of the iteration, it is important to understand the exact problem so that the study does not deviate from the actual problem. It is also important to evaluate the solutions that already exist. Further research is completely dependent on the identified problems; hence, it is extremely important to identify the right problems (Rai 2017). The problem identification process included formulation of interview questions to mine required information, conduction of interviews, literature review, thematic analysis, and survey.

The first iteration involved interviews and literature review as the primary source for identifying data quality challenges. A total of 27 data quality challenges were identified and prioritized. The interviews were conducted via Microsoft Teams[1], an online communication tool. The interviews were recorded and transcribed. The data quality challenges were segregated using data-driven thematic analysis. To identify more pressing data quality challenges, a survey was sent out to deep learning and AD experts to identify those challenges.

### 3.1.2 Solution Design

In the second iteration, RQ2 was primarily addressed. In this iteration, the goal is to design and develop solutions to the identified problems from the first iteration. The solution can be in the form of a construct, a model, a method, an instantiation,

[1] `https://www.microsoft.com/en-ww/microsoft-teams/group-chatTeams-software`

or a design theory (vom Brocke & Maedche 2019). The developed novel artifact is designed to meet the stakeholder requirements and resolve the identified challenges.

In this phase, a *Data Quality Workflow* was developed which is discussed in detail in Section 4.2.1. The data quality challenges that were identified during the first iteration were segregated into different challenge sets using mind maps. An initial template for the *List of Challenges* artifact component was developed to document all the challenges identified through interviews and literature review. A template for *List of Data Quality Attributes* was created to list all the data quality attributes along with their definition and the challenges that affect the data quality attributes. An Ishikawa diagram, also known as a fishbone diagram, was also developed to divide the challenges into various challenge sets. Finally, a template for potential solutions and the potential solutions themselves to handle the data quality challenges were developed. The potential solutions were identified through literature review and group discussion. The steps regarding the implementation of the potential solutions were also listed.

During the second iteration, *List of Challenges* template, *List of Data Quality Attributes* template, *List of Data Quality Attributes Metrics* template, and *Potential Solutions* template, were refined and concrete versions were developed. The templates are presented in Tables 4.1, 4.3, 4.5, and 4.7. The challenge sets were identified and defined. The challenges were grouped and listed under each of the challenge set. Also the data quality attributes without metrics were identified and listed.

### 3.1.3 Evaluation

Evaluation is a process of observing and measuring the artifact's efficacy to support a solution to the problem (Peffers et al. 2006). Peffers et al. (2006) further discuss that evaluation is performed by the comparison of actual results derived by using the artifact to objectives of the solution. The constructed artifact should be evaluated based on the suggested criteria of solution (Vaishnavi & Kuechler 2004). During this stage the developed artifact related to a study topic is evaluated based on the application of the artifact to the challenges identified. It is then observed if the artifact is able to achieve the intended results.

During the second iteration interviews, the *List of Challenges* artifact component and the template were presented to interviewees to elicit their opinion and expertise. It was attempted to validate individual challenges in the list and solutions developed so far. Using their feedback the *List of Challenges*, *Potential Solutions* and other components of the artifacts were further refined.

The association between data quality challenges and data quality attributes that was established through brainstorming by listing the challenge set under particular data quality attribute were validated by participants using Mentimeter [2] in a focus group session. The association of challenges with the data quality attributes was evaluated with the experts using Boolean questions (yes or no). If a participant answers with a

[2] `https://www.mentimeter.com`

"yes", then it denotes an agreement that a particular challenge affects the attribute and if the participant chooses to answer a "no", then it denotes a denial that a challenge affects the attribute.

A survey was also sent out to the participants for similar data collection as that of the focus group session. In addition to the questions asked in the focus group session, in the survey, the fields of the templates of *List of Challenge*, *List of Data Quality Attribute*, *List of Data Quality Attribute Metrics*, and *Potential Solutions* were validated in order to check if they are relevant to the study and are helpful in deriving intended results. The opinion from the experts was also gathered to understand, if any other fields are required in the artifact template or that are relevant to the study in this thesis that help to solve challenges.

## 3.2 Interviews

Qualitative interview is considered as an approach to collect data that addresses a number of research questions (McGrath et al. 2018). In this thesis, interviews were conducted with the case company experts to better understand and investigate various data quality challenges. Majid et al. (2017) state that "qualitative interviews offer rich and detailed information in understanding people's experiences." In this thesis study, interviews helped to mine data quality challenges and data quality attributes in the field of deep learning based on the experience of the experts in automotive domain.

The selected interviewees were experts who had years of experience in deep learning and who were willing to take part in the interviews. Interviews were carried out remotely via Microsoft Teams due to COVID-19 restrictions and mostly lasted for an hour. As remote online interviews reduce time and cost (Farooq & de Villiers 2017), the interviews conducted also had these benefits. Before the start of the interviews, a consent form was filled out by the interviewee regarding recording the interview and use of data collected during interview for further research. Consent form provides the participants an assurance that the information would be used in ethical manner (Illing 2013). The consent form is presented in Appendix A.1.1. Interviewees were assured of data confidentiality. It was agreed with the interviewees to share the results of research with them after the study is complete.

Farooq & de Villiers (2017) state that a well developed interview guide helps in building rapport with the interviewees. They also argue that feedback received from interviews can be helpful in further refining and rephrasing of the interview questions. Based on the outcome of previous interviews, questions were tuned accordingly to fill the knowledge gap for further interviews. The questions were designed based on the goal to find answers to the research questions. The interviews were conducted in two cycles. The interview questions are presented in Appendix A.1.2.

During the first iteration interviews, data quality challenges were identified which addressed RQ1. The questions were formulated to mine data quality challenges. In

the second iteration interviews, the questions were designed to brainstorm about potential components of the artifact. The most pressing challenges were identified and the potential solutions to the challenges were validated during the second cycle interview. This helped to address the RQ2.

Table 3.1 shows the list of first iteration interviewees in the order of interviews conducted.

**Table 3.1:** List of Interviewees in First Iteration Interviews

| Name | Role | Team |
|---|---|---|
| Interviewee A | Research Specialist | Research |
| Interviewee B | Functional Safety Engineer | Driver Assistance Systems |
| Interviewee C | Feature Tech Lead | Vision Pre-Development |
| Interviewee D | Group Manager | ADAS Platform Development |
| Interviewee E | Technical lead | AI and ML |

Table 3.2 shows the list of second iteration interviewees in the order of interviews conducted.

**Table 3.2:** List of Interviewees in Second Iteration Interviews

| Name | Role | Team |
|---|---|---|
| Interviewee F | Development Manager for Road Traffic Management | Traffic Management |
| Interviewee G | Product Owner | Ground Truth |
| Interviewee H | Engineering Technical Fellow | Research and Innovation |

## 3.3 Thematic Analysis

Thematic analysis is a qualitative analysis tool to classify qualitative data into themes, which enables "to associate an analysis of the frequency of a theme with one of the whole content" (Alhojailan 2012). Thematic analysis is also an important tool for analysis of qualitative data since such data needs rigorous, systematic analytical procedures to be regarded as trustworthy (Nowell et al. 2017). Labra et al. (2019) describes six phases of thematic analysis, namely in order, *familiarization with collected data*, *generating initial codes*, *searching for themes*, *reviewing themes*, *defining and naming themes*, and *presenting and discussing results*. Thematic analysis has been applied in fields such as social work (Labra et al. 2019), nursing (Vaismoradi et al. 2016), tourism research (Walters 2016), and education (Xu & Zammit 2020).

As this research is iterative in nature, thematic analysis was conducted two times. The researchers, more or less, followed the phases described by Labra et al. (2019). After the conduction of the interviews, they were transcribed immediately in both of the iterations. The researchers divided the content of each interview transcription independently. This familiarized the researchers with the collected data. Data-driven

coding (Gibbs 2007) was used in the thematic analysis of first-iteration interviews. Furthermore, descriptive coding (Linneberg & Korsgaard 2019) and analytic coding (Gibbs 2007) were applied as the methods of coding. In descriptive coding, codes are assigned based on a summary of the content (Linneberg & Korsgaard 2019). If a collection of statement includes phrases or keywords that can be extracted as codes, descriptive coding method was used. Similarly, codes are assigned as analytical extractions during analytical coding (Gibbs 2007). So, in the same collection of statement, if that collection can be analyzed and coded as an abstract concept, analytic coding was also used.

The transcriptions and codes were recorded in a spreadsheet document in a chronological order. Along with the transcriptions, a standard form containing the name, position of the interviewees in their company, their experience, etc. were also recorded. Both researchers coded the data in separate spreadsheets independently. Same collection of statements could be coded with one or more codes. Then the researchers conducted "code combination" meetings in which they discussed their respective codes and formulated final "combined" codes. After this step, themes were searched for, reviewed, and defined using the final codes. In this iteration, thirteen themes were identified and defined. These themes were subdivided into a number of codes in each theme. For example, a theme called *Applications* can be divided into codes such as *AEB*, *ACC*, etc. The themes are further explained in Section 4.1.1 of this thesis. The themes and their associated codes are presented in Appendix A.5.

The second iteration of the research pertained to figuring out potential solutions to the challenges identified in the first iteration. Thematic analysis performed in this iteration used two coding techniques. First, analytic coding (Gibbs 2007) was used to assigned analytical codes to the data. Secondly, deductive coding (Seale 2017), also known as concept-driven coding (Gibbs 2007), was used. In deductive coding, a set of themes and codes are predetermined and data are coded based on those. The researchers, in this thesis, used four deductive codes. They related to reconfirmation of an already-identified challenge by the interviewee, reconfirmation of a proposed solution, problem with the already-identified challenge, and problem with the proposed solution.

## 3.4 Survey

According to Isaac & Michael (1997), a survey research is useful “to answer questions that have been raised, to solve problems that have been posed or observed, to assess needs and set goals, to determine whether or not specific objectives have been met, to establish baselines against which future comparisons can be made, to analyze trends across time, and generally, to describe what exists, in what amount, and in what context.” Surveys give a "snapshot" of the situation at a certain point in time (Denscombe 1998). An advantage of a survey research is that such study generates data that resembles the real world (Kelley et al. 2003). As any researcher can collect data from a large number of people by using survey as a research tool, it is applicable in this thesis as well.

During the first iteration of this study, five people were interviewed. While eliciting the challenges of assessing data quality, most of the challenges were mentioned by only one interviewee (i.e. only one interviewee mentioned a particular challenge; others did not mention it as a challenge). This resulted in there being a large number of challenges. However, appropriate level of severity was difficult to deduce from the interview responses. In order to mitigate this, a survey was conducted with the same interview participants and the members of the VEDLIoT Requirements Engineering Workgroup, an European Union-funded research and innovation program (`https://vedliot.eu`). As questionnaire development is a multistage process (Pew Research Center n.d.), the final question set for the survey was developed after a few design iterations of the questionnaire. The first version of the question set asked the participants to rank each identified challenge using a Likert scale of range 1 – 10. Cox III (1980) states that "beyond a certain limit an increase in the number of response alternatives becomes meaningless..." He also provides recommendations regarding the number of alternatives. The suggested range is between five to nine alternatives. However, he cautions against overuse of neutral category (e.g., a value of 3 in a scale of 1 – 5) (Cox III 1980).

The first version of the survey questionnaire was modified. In the modified version, 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. Similarly, they were asked to rate the categories, themselves, as well. The ranking was based on a Likert scale of range 1 through 6 – 1 being the least severe challenge and 6 being the most severe challenge. A scale with even number of alternatives was deliberately selected so as to induce the participants to "pick a side." The modified version of the survey was sent to the participants. The survey was created, sent, and collated using an online tool, Microsoft Forms. An online survey has a number of benefits including faster delivery and timeliness, convenience, easy data entry and analysis, and provisions of non-skippable mandatory questions (Evans & Mathur 2005).

A survey was also conducted during the third iteration of this study. A comprehensive survey questionnaire was sent to the members of the VEDLIoT Requirements Engineering Workgroup. This survey attempted to validate the components of the artifact. First, it asked the participants to provide Boolean response to appropriateness of individual fields for the templates of the artifact components. For e.g., it asked whether a *Description* field is appropriate in *List of Challenges* artifact component. In the survey, templates for *List of Challenges*, *List of Data Quality Attributes*, *List of Data Quality Attribute Metrics*, and *Potential Solutions* artifact component templates are validated for appropriateness.

Similarly, the participants were asked to rank the challenges in the same manner as in the survey in the first iteration (i.e., ranking of individual challenges in each challenge set and Likert scale ranking for challenge sets). This was done to gather more data regarding severity of the challenges.

Likewise, the direct relationship between the data quality challenges and AI models

were also validated during the final survey. The relationship denotes that the AI models are directly affected by certain data quality challenges.

Finally, the participants were asked to provide Boolean value for association between challenges and data quality attributes. 20 data quality challenges were presented to the participants for validating that the challenges affect the attributes or not. The rest of the challenges, except those that were validated in the focus group, were not validated for association with data quality attributes because, in the opinion of the researchers in this study, there were no clear connection on how those challenges could affect any data quality attribute. The challenges that were presented in the survey included *Data Delay*, *Data Drop*, *Incomplete Data*, *Low Labeled Data Volume*, *Data Acquisition*, *Data Ownership*, *Imbalanced Dataset*, *Redundant Data*, *Improper Data Transfer*, *Manual Data Collection*, *Manual Data Labeling*, *Regulatory Compliance*, *New Data Types*, *Data Dependent on External Conditions*, *Incompatible Data Formats*, *Outlier Data*, *Unstructured Data*, *Lack of Good Data from Simulations*, *Incorrect Labeling*, and *Noise*.

### 3.4.1 Challenge Score

During the first iteration, 27 data quality challenges were identified through interviews and literature review. A way to rank the challenges was necessary for effective analysis. *Challenge Score* ranks the identified challenges in terms of its 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. As stated in Section 3.4, the survey contained two types of questions. One type of question asked the participants to provide a significance value based on a Likert scale to five sets of challenges. Another type of question asked to rank individual challenges inside the five sets of challenges.

As there are two types of response from two types of question, they needed to be combined in some manner for both of them to be useful. *Challenge Score* combines both types of responses in one final value. For each respondent, the value they provide for the overall sets of challenges are recorded. The highest ranked challenge in a challenge set is given the highest numerical value corresponding to the number of challenges in that challenge set. Decreasing numerical values are assigned to remaining challenges in the particular challenge set. For e.g., if there are 4 challenges in a challenge set, the highest ranked challenge is given a value of 4, second highest ranked is given a value of 3, and so on.

For each individual challenge, the assigned numerical value is multiplied with the value given by that particular participant for the challenge set of that particular challenge. This is done 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 with the total number of challenges in the particular challenge set and further by dividing the result by the total number of participants. This is done to normalize the final value.

Below is an example of the calculation of *Challenge Score*. Here, *A*-*F* are survey participants.

**Table 3.3:** Ranking for an Individual Challenge

| Challenge Set | Challenge | A | B | C | D | E | F | Challenge Score |
|---|---|---|---|---|---|---|---|---|
| Data Management | Manual Data Labeling | 10 | 3 | 8 | 10 | 9 | 9 | 3.217 |

**Table 3.4:** Ranking for a Challenge Set

| Challenge Set | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Data Management | 4 | 3 | 4 | 4 | 6 | 2 |

In Table 3.3, survey participants A and D ranked *Manual Data Labeling* as the most pressing challenge among *Data Management* challenges; hence it is given a value of 10 as there are 10 challenges in that challenge set. Survey participants E and F ranked this challenge second, so it is given a value of 9. For survey participant C, it was the third most pressing; and hence a value of 8 is given. For survey participant B, this challenge was the third from the last. Hence, it was given a value of 3.

Similarly, as shown in Table 3.4, survey participants A, C, and D gave *Data Management* challenge set a value of 4 in the Likert scale. Survey participant B gave a value of 3, survey participant E gave a value of 6, survey participant F gave a value of 2.

Finally, sum of the product is calculated. The formula for this particular example is (Table 3.3 Column *A* * Table 3.4 Column *A*) + (Table 3.3 Column *B* * Table 3.4 Column *B*) + (Table 3.3 Column *C* * Table 3.4 Column *C*) + (Table 3.3 Column *D* * Table 3.4 Column *D*) + (Table 3.3 Column *E* * Table 3.4 Column *E*) + (Table 3.3 Column *F* * Table 3.4 Column *F*).

Numerically, (10*4) + (3*3) + (8*4) + (10*4) + (9*6) + (9*2) = 193

Then, to normalize the result, the value is divided by the total number of challenges in the particular challenge set. Here, as there are 10 *Data Management* challenges, the sum is divided by 10. Furthermore, the value is also divided by the total number of participants in the survey. So, the result is divided by 6.

Numerically, (193/10)/6 = 3.217. This is the final *Challenge Score*.

The same technique is used to calculate the *Challenge Score* for other challenges as well. This technique is also utilized in the calculation of *Challenge Score* from responses of the survey in the third iteration.

## 3.5 Focus Group

Freitas et al. (2021) in their article state that "focus group is a type of in-depth interview accomplished in a group defined with respect to proposal, size, composition, and interview procedures." Focus group is a qualitative research method which is performed in combination with other research techniques. Focus group can be performed by gathering people of similar research interests or expertise in a group (Freitas et al. 2021). In their article they further discuss that the participant group is provided an environment to register their ideas and opinions spontaneously and facilitate interaction focused on a topic. Focus group is helpful in investigating new fields through idea generation.

A focus group session was conducted to validate the artifact components during the third iteration of this thesis. A group of people from academia and industry who shared an interest in the research study of this thesis took part in the session. The session was conducted for two hours with 5 participants. The participants were presented with a set of questions to brainstorm regarding the association between the challenges and the data quality attributes. They also shared their ideas and thoughts through discussion. One of the participants was not available for the entirety of the session due to scheduling conflicts. Mentimeter was used to present the questions to the participants as well as collect the response. The response data was later analyzed to derive conclusions from the focus group session.

# 4

# Results

In this chapter, the results of each iteration of the study is presented. The results are presented for each of the research questions relating to the problem, solution, and evaluation stages of a typical DSR. In Section 4.1, results of the first iteration are presented. As such, the themes identified through thematic coding are presented in that section. Section 4.2 presents the results of the second iteration i.e., design of the artifact as a solution. Lastly, in Section 4.3, the results from the evaluation of the artifact are provided. It includes analysis of the focus group response and the survey response.

## 4.1 Iteration 1 - Problem (RQ1)

### 4.1.1 Identified Themes

#### 4.1.1.1 Applications

This theme refers to a set of decision making applications in the context of ADAS, that assist and help driver navigate during uncertain and complex situations. This helps in the safety of driver, passengers and pedestrians. Performance of such applications is extremely important and is dependent on the quality of data they are fed. Examples can be object detection, braking request management, vehicle detection, lane departure assist. Subpar quality performance of the above mentioned applications may lead to mishaps and loss of lives.

#### 4.1.1.2 Challenges

This theme includes the statements in which interviewee discuss about the challenges associated with data quality and assessment of quality of data. One of the primary themes this study focuses on is *Challenges*. Some of the codes in this theme include data delay, expensive procedure, unclear traffic signs, and so on. As part of the analysis of theme, the codes are subdivided into challenges that the artifact can try to solve and those that it cannot solve. Moreover, those challenges that can be solved are partitioned into *system challenges* and *data challenges*. Data challenges directly affect the behavior of the deep learning algorithms. System challenges are those challenges that are pertinent to data quality but do not directly affect the deep learning algorithms. Furthermore, the challenges are also categorized together into

sub-themes. They are – *data availability* challenges, *data management* challenges, *data source* challenges, *data structure* challenges, and *data trust* challenges.

#### 4.1.1.3 Current Procedures

This theme helps in understanding the procedures and methods followed, and the tools used currently in various teams and companies to gather requirements, collect data, set data requirements, ensure data quality, and test. It helps in identifying advantages and challenges in the current procedures followed for ensuring data quality and other operations that directly or indirectly affects such data quality. For e.g., current procedure in one the company to ensure data quality in ADAS is to drive a vehicle for 10,000 Kms in different geographical locations, which is time consuming and expensive. The codes in this theme are further subdivided into three broad topics – the processes associated with data (e.g., data collection, data storage, data review, data security, etc.), the tasks the development team performs (e.g., project approval, function development, peer review, etc.), and the tools used (e.g., data blacklist, written documentation, etc.).

#### 4.1.1.4 Data Assumptions

This theme relates to the assumptions made regarding data by function developers and function development companies. This theme has three codes in it: labeled data, trust in bounding boxes, and trust in sensor calibration. Function developers depend on the data collected from a number of sensors. They regard these data as correct and develop the functions based on the data. In other words, they assume that these data are correct.

#### 4.1.1.5 Data Types

Data types is an important theme in this thesis study. Several important data types that are encountered and collected for the better performance of the safety functions are identified and listed through interviews and literature reviews. It is important to train the systems with various data types depending on the contexts of the use of safety systems. Few examples for data types in context of ADAS are position, velocity, weight of vehicle, size of object, etc.

#### 4.1.1.6 Extra Info

All the codes that does not fall in any of the identified themes are grouped under *Extra Info* theme. They are of less relevance and do not provide valuable information for study in this thesis. Examples for *Extra Info* are team structure, customers, predictable event, vehicle dynamics model, etc.

#### 4.1.1.7 Goals

Another important theme is the *goals* theme. In this theme, those interview statements that relate to the goals of the system they are working on, the goals of their team, and the goals of their particular job are categorized together. Example codes

include crash avoidance, improve data quality, proper function switch off, correct system behavior, etc. These goals facilitate the development of the artifact since it is trying to help the stakeholders achieve their goals.

#### 4.1.1.8 Hardware Components

Hardware components is a theme that comprises of various hardware used for data gathering. The quality of data is dependent on the quality of hardware. Subpar quality hardware may lead to noise and other quality issues in data which in turn effect deep learning algorithm performance. For e.g., Low quality night vision camera may not capture the right images. Low resolution data might lead to bad training of algorithms and effect the algorithm performance.

#### 4.1.1.9 Impact of Low Data Quality

This theme collates the statements that relate to the consequences of low data quality. A code in this theme, "improper algorithm training" causes underfitting, overfitting, false positives, and false negatives. These are also codes in the theme. Similarly, an example of another code is aggressive system behavior. This theme is useful to understand the problems causes by training deep learning models using data with subpar quality.

#### 4.1.1.10 Metrics

This is also an important theme as it organizes those statements that mention metrics that can be used to measure data quality. Some of the codes in this theme include availability, signal-to-noise ratio, accuracy, standard deviation, etc. This theme was referenced alongside *challenges* theme during analysis of the challenges to correlate them with appropriate metrics.

#### 4.1.1.11 Nature of Data

This theme relates to those statements that discuss about various nature of data. The data types, themselves, are collected in a separate theme called *data types*. Examples of nature of data codes include rare data, outlier data, third-party data, computed data, difficult-to-label data. The aim of this theme is categories the types of data into broad sub-themes. For e.g., velocity is coded in theme called *data types*, but certain values of velocity can be an outlier or can be computed or so on.

#### 4.1.1.12 Solutions

This is also one of the important themes since it groups the statements that discuss about tools and techniques currently used to assess data quality in the interviewees' organizations and teams. Additionally, this theme also collects those statements in which potential solutions that have not yet been applied to evaluate data quality are discussed. Example codes in this theme are heuristics, automated data analysis, data review, data contract, simplified reporting, quick feedback loop, and so on.

This theme is referenced along with challenges and metrics themes to develop the artifact of this study.

#### 4.1.1.13 Team Structure

Team structure is a theme in which it is determined regarding the team that is responsible for gathering data requirements, team responsible to set data quality requirements, collecting data, and developing functions. It also helps in identifying whose say matters more in setting data quality requirements. It further helps in understanding how the data is handled and transferred between various teams and its implications. For e.g., one of the interviewee mentioned that, it is the product owner who is responsible for setting data quality requirements and he/she also gets help from the developers.

## 4.2 Iteration 2 - Solution (RQ2)

In this section, a series of artifact components are devised and presented. The artifact components can be grouped into a unified and comprehensive framework called ***"Data Quality Assessment and Maintenance Framework***." The components include *Data Quality Workflow*, *List of Challenges*, *List of Data Quality Attributes*, *List of Data Quality Attribute Metrics*, and *Potential Solutions*.

### 4.2.1 Data Quality Workflow

The steps taken in this thesis to assess data quality, data quality attributes, and potential solutions can be presented in the form of a workflow. The workflow is aptly titled as *Data Quality Workflow*. It can, in itself, be one of the components of the artifact developed in this thesis as it provides a proper method of management of data quality in any deep learning systems such as ADAS. Following are the steps in the workflow.

1. **Identify data quality challenges**

   Challenges concerning to data quality can be identified from a number of sources. Primary sources of data collection such as interview, field study, survey, etc. can be utilized to identify the challenges. Research papers, books, etc. can be used as second-hand sources as well. As stated in the method chapter of this thesis, a number of interviews were conducted to understand data quality challenges. Similarly, seven research papers were referred to collect potential data quality challenges as well. Furthermore, the collected challenges can be divided into different categories. In this thesis, they were categorized in five groups relating with *data availability*, *data management*, *data source*, *data structure*, and *data trust*.

2. **Collect and organize data quality attributes**

   Data quality attributes can be collected from various sources such as research

papers, proceedings papers, books, standards, technical reports, Internet articles, interviews, etc. For the purpose of this thesis, four research papers and one technical report were referred to collect data quality attributes. Data quality attributes were also elicited from interviews. Identical data quality attributes can be represented by single data quality attribute. Differently phrased data quality attributes can also be represented by a single attribute; however, this is not done in this thesis. For e.g., *understandability* and *ease of understanding attributes* can be represented by the same attribute.

3. **Associate data quality challenges and data quality attributes**

   Once data quality challenges and data quality attributes have been identified, they can be associated with each other. The association, here, 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). However, there can be data quality attributes that are not affected by any identified challenge and data quality challenges that do not affect any attribute.

4. **Define data quality attribute metrics**

   In this step, metrics to measure data quality attributes are formulated. The metrics help to put a quantitative value to the attributes. For 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. For 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 application.

5. **Identify solutions for data quality challenges**

   A way of improving data quality attribute metrics, and thus, improving data quality attributes, is to determine 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 result in lesser data to be dropped, thus improving *completeness* (attribute). There are a number of sources to identify solutions such as research papers, technical reports, books, etc. Teams can also brainstorm and devise potential new solutions for the challenges as well. An effective way to validate potential solutions is to implement them as tests in part of a system. For the purpose of this thesis, they were validated through expert interviews and focus group session.

6. **Present to stakeholders**

As the final step, identified data quality challenges, identified data quality attributes, and potential solutions should be presented to appropriate stakeholders. They could be higher management, other colleagues, or customers. Suitable form of presentation should also be decided.

The steps mentioned above do not necessarily have to be linear. They can be executed in parallel. Especially, steps 1 and 2 can be done in parallel. Similarly, steps 3, 4, 5 can be done in any order and can be performed in linear or parallel fashion. The steps can also be iterative, i.e., one can cycle through a single step more than once.

The workflow is shown in Figure 4.1. Here, solid arrow depicts steps that need to follow another step. Dashed arrow represents steps that can be done in parallel. Curved arrow going towards the same box means the step is iterative.


**Figure 4.1:** Data Quality Workflow Solution

### 4.2.2 List of Challenges

After analyzing the survey response from the first iteration and the interview responses from the second iteration of this study, a number of challenges were identified. However, in this artifact component, not all of them are listed and discussed. Those challenges that the researchers deemed as *System Challenges* are not mentioned in this component as they pertain to the system and not data. These challenges can be seen in the figure presented in Appendix A.2 - *System Challenges* box.

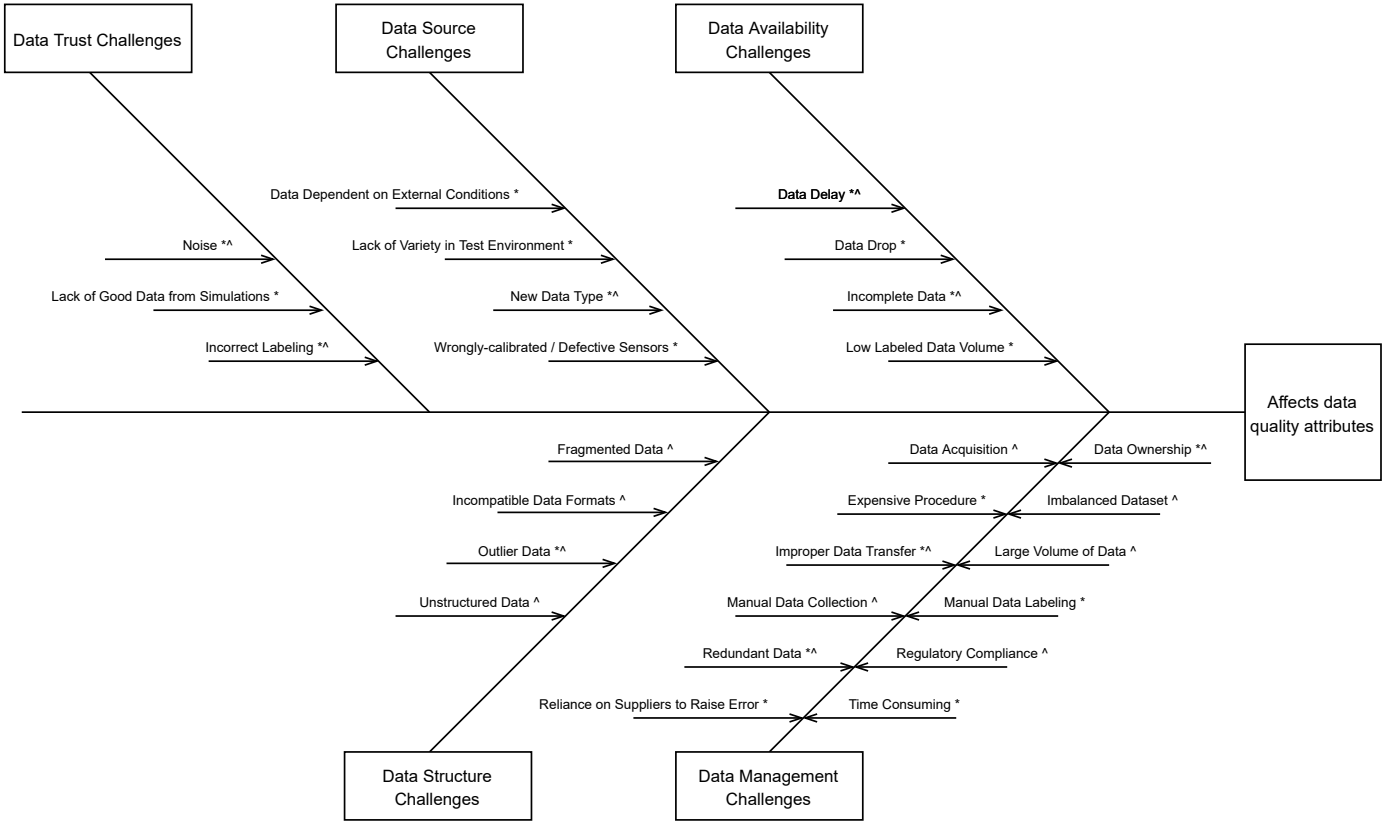
Six challenges that were initially identified as part of *Data Challenges* were also removed from the list of challenges. Reasons for the removal include high vagueness of the challenge, realization that the challenge is not a challenge itself, obscurity, etc. The removed challenges are *Uncertain Data Quality Identification*, *Fast Increasing Data*, *Wrong Metadata*, *Reliance on a Single Data Source*, *Data Mix-up*, and *Fake Data*. These are still presented in the *Challenge Score* ranking in Appendix A.3. Two new challenges - *Expensive Procedure* and *Time Consuming* were added after the second iteration because almost all of the experts mentioned them as challenge. All of the challenges are retrieved from the list of initial challenges as shown in the *Data Challenges* box in A.2.

**Table 4.1:** Template for *List of Challenges* Artifact Component

| Field | Description |
|---|---|
| Name | Name of the data quality challenge |
| Reference | Reference that denotes the identification of the challenge |
| Description | Description of the data quality challenge |
| Directly affects AI Functions | Boolean value to denote whether the data quality challenge directly affects AI functions or not |
| Challenge Score | A calculated value that denotes the ranking of the data quality challenge in terms of severity |
| Responsible Stakeholder | People and/or department in an organization responsible to handle the challenge.<br><br>**Note:** This field is not implemented in the following list of challenges as the values for these fields should be provided by implementer and not the researchers. |
| Impact Level | Degree to which the challenge affects the AI models. Could have values such as HIGH, MEDIUM, LOW.<br><br>**Note:** This field is not implemented in the following list of challenges as the values for these fields should be provided by implementer and not the researchers. |

**Table 4.2:** Definition of Challenge Sets and List of Challenges

| Challenge Set | Definition | List of Challenges |
|---|---|---|
| Data Availability Challenges | These challenges affect the availability of data during processing by AI models. | Data Delay, Data Drop, Incomplete Data, Low Labeled Data Volume |
| Data Management Challenges | These challenges relate to the management of data and management operations performed on data. | Data Acquisition, Data Ownership, Expensive Procedure, Imbalanced Dataset, Improper Data Transfer, Large Volume of Data, Manual Data Collection, Manual Data Labeling, Redundant Data, Regulatory Compliance, Reliance on Suppliers to Raise Error, Time Consuming |
| Data Source Challenges | These are the data quality challenges caused due to the source of the data. | Data Dependent on External Conditions, Lack of Variety in Test Environment, New Data Type, Wrongly-calibrated / Defective Sensors |
| Data Structure Challenges | These challenges are related to the format and structure of the data. | Fragmented Data, Incompatible Data Format, Outlier Data, Unstructured Data |
| Data Trust Challenges | These challenges are caused due to lack of transparency in the data and lack of its quality to extract meaningful information. | Noise, Lack of Good Data from Simulations, Incorrect Labeling |




* denotes challenges identified through interviews with experts

^ denotes challenges identified through literature review

**Figure 4.2:** Identified Challenges Divided in Challenge Sets

**Note:**

- In the following list, an asterisk (*) denotes that the challenge shares its rank with other challenge(s).
- In "Directly affects AI Functions" field, the response from the survey in the third iteration is provided. The response is in the format of the number of "yes" and "no" answered by the survey participants.

#### 4.2.2.1 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. For e.g., a delay can occur in data transmission from sensor to long-term storage, sensor to deep learning functions, long-term storage to deep learning functions, etc. Similarly, there can also be a delay in the reception of 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 every now and then. This causes tracking of data to be difficult and disrupts management and processing of data. This will hinder the training of deep learning models. For e.g., dropping three frames in a 30 second clip would mean losing 0.7 seconds. This is a problem for algorithmic correctness.
**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), Azeroual & Abuosba (2017)
**Description:** This challenge is similar to data drop in the sense that both are caused by missing data. 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, data drop occurs when there is 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:** In the training dataset, the volume of the data that is labeled is significantly lesser than the volume of the data that is unlabeled. Since a large volume

of data is unlabeled, the unlabeled data is useless and the deep learning models cannot be properly trained. For 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*)

#### 4.2.2.2 Data Management Challenges

**Name:** Data Acquisition
**Reference:** Kruse et al. (2016)
**Description:** There are a number of ways data can be acquired, such as, real-world collection, simulations, and third-party purchase. Identifying the method to follow in any given context is a challenge. Similarly, the required procedures for each method need to be determined prior to the acquisition. Furthermore, methods for collection, transmission, and storage of data should also be defined.
**Directly affects AI functions:** 2 "Yes", 2 "No"
**Challenge score:** Survey 1 - 0.850 (Rank 31/31), Survey 2 - 1.100 (Rank 23/25)

**Name:** Data Ownership
**Reference:** Interviewee E, Kruse et al. (2016)
**Description:** This challenge pertains to the determination of the owner of the data and necessary consent to use the data owned by a different party. For e.g., data generated by a vehicle in operation is owned by the owner of that vehicle. If the vehicle manufacturer wishes to use data for improving the deep learning models, it has to identify and execute the appropriate steps it needs to take to be able to legally use that data.
**Directly affects AI functions:** 1 "Yes", 3 "No"
**Challenge score:** Survey 1 - 1.817 (Rank 19/31), Survey 2 - 2.400 (Rank 12/25)

**Name:** Expensive Procedure
**Reference:** Interviewee A, Interviewee B, Interviewee C, Interviewee D
**Description:** Procedures such as data collection, data labeling, simulations, and data management are expensive. The cost increases, especially when tasks have to be done manually or be done by a specialized system.
**Directly affects AI functions:** 0 "Yes", 4 "No"
**Challenge score:** Survey 1 - Not Available (challenge added in second iteration only), Survey 2 - 1.925 (Rank 19/25)

**Name:** Imbalanced Dataset
**Reference:** Corrales et al. (2016), Azeroual & Abuosba (2017)
**Description:** Imbalanced dataset or unevenly represented data leads to bias in the neural networks. If a dataset contains records that skew towards certain label(s),

the training will be skewed towards those labels. This could lead to biased decisions to be made in the future.
**Directly affects AI functions:** 4 "Yes", 0 "No"
**Challenge score:** Survey 1 - 2.967 (Rank 6/31), Survey 2 - 3.025 (Rank 6/25)

**Name:** Improper Data Transfer
**Reference:** Interviewee E, Kruse et al. (2016)
**Description:** If proper standards and methods are not followed when transmitting data between the sensors, from sensors to data storage, from sensors to deep learning functions, and to-and-fro data storage and deep learning functions, it could lead to situations such as data corruption.
**Directly affects AI functions:** 3 "Yes", 1 "No"
**Challenge score:** Survey 1 - 1.667 (Rank 21/31), Survey 2 - 1.950 (Rank 17/25)

**Name:** Large Volume of Data
**Reference:** Cai & Zhu (2015), Kruse et al. (2016)
**Description:** A large data volume makes it difficult to assess the quality of the data within an acceptable timeframe and cost. As the amount of data increases, it becomes challenging to "collect, clean, integrate, and obtain the necessary high-quality data"
**Directly affects AI functions:** 2 "Yes", 2 "No"
**Challenge score:** Survey 1 - 2.250 (Rank 15/31), Survey 2 - 2.650 (Rank 8/25)

**Name:** Manual Data Collection
**Reference:** Gao et al. (2016)
**Description:** In many applications, data has to be collected manually. For e.g., in the context of ADAS, vehicles have to be driven in different situations and environments in order to gather data. This is a time consuming and expensive method of data collection.
**Directly affects AI functions:** 2 "Yes", 2 "No"
**Challenge score:** Survey 1 - 2.500 (Rank 12/31), Survey 2 - 2.425 (Rank 11/25) (Jointly as *Manual Data Collection and Labeling*)

**Name:** Manual Data Labeling
**Reference:** Interviewee C, Interviewee D
**Description:** Most of the time, the collected data has to be manually labeled. Apart from being time consuming and expensive, manual data labeling is also prone to human errors. Adequately trained and enough manpower is required for manual labeling of data. For e.g., manually labeling a car as a car and a truck as a truck in every frame is difficult.
**Directly affects AI functions:** 2 "Yes", 2 "No"
**Challenge score:** Survey 1 - 3.217 (Rank 5/31), Survey 2 - 2.425 (Rank 11/25) (Jointly as *Manual Data Collection and Labeling*)

**Name:** Redundant Data
**Reference:** Interviewee E, Corrales et al. (2016), Azeroual & Abuosba (2017)
**Description:** This challenge causes similar problem as that caused by an imbalanced dataset. If there is duplicate data, neural networks will be biased towards the labels that are more prevalent than those that are not. This could lead to biased decisions to be made in the future by the system.
**Directly affects AI functions:** 2 "Yes", 2 "No"
**Challenge score:** Survey 1 - 1.367 (Rank 27/31), Survey 2 - 0.575 (Rank 25/25)

**Name:** Regulatory Compliance
**Reference:** Kruse et al. (2016)
**Description:** In many instances where one is working with data, there are rules and regulations one has to be in compliance. Identifying the regulations, required processes, and executing the compliance tasks are challenging.
**Directly affects AI functions:** Not applicable (excluded from the survey due to technical error)
**Challenge score:** Survey 1 - 1.867 (Rank 17/31), Survey 2 - Not Available (due to technical error, see Appendix A.6 for information)

**Name:** Reliance on Suppliers to Raise Error
**Reference:** Interviewee B
**Description:** Function developers are reliant on the suppliers of the sensors to detect erroneous data and raise such errors. There is no appropriate processes within the function development organization to identify the errors and thus, are dependent on the sensor suppliers.
**Directly affects AI functions:** 1 "Yes", 3 "No"
**Challenge score:** Survey 1 - 2.583 (Rank 11/31), Survey 2 - 2.225 (Rank 14/25)

**Name:** Time Consuming
**Reference:** Interviewee A, Interviewee B, Interviewee C, Interviewee D
**Description:** Procedures such as data collection and data labeling consume a long time for completion. Futhermore, if they are to be done manually, it takes even more time. Fixing errors with the procedures take time as well.
**Directly affects AI functions:** 1 "Yes", 3 "No"
**Challenge score:** Survey 1 - Not Available (challenge added in second iteration only), Survey 2 - 2.350 (Rank 13/25)

#### 4.2.2.3 Data Source Challenges

**Name:** Data Dependent on External Conditions
**Reference:** Interviewee B

**Description:** Sometimes data could be affected by external conditions. For example, during bad weather like heavy snow, heavy rain, sandstorms and when it is extremely dark, the camera or the sensor would not be much of a use and the data (pictures and video streams) captured by the camera would be of low quality. This would affect the performance of deep learning algorithms if trained with the low quality data gathered during such weather condition.
**Directly affects AI functions:** 4 "Yes", 0 "No"
**Challenge score:** Survey 1 - 3.238 (Rank 4/31), Survey 2 - 1.937 (Rank 18/25)

**Name:** Lack of Variety in Test Environment
**Reference:** Interviewee B
**Description:** There is lack of variety data from different environment sources across globe in which the vehicle has to be tested. An example can be, if the car is only tested in Sweden then there might be no test data of the vehicle behavior in desert conditions.
**Directly affects AI functions:** 4 "Yes", 0 "No"
**Challenge score:** Survey 1 - 3.786 (Rank 2/31), Survey 2 - 3.625 (Rank 3/25*)

**Name:** New Data Type
**Reference:** Interviewee D, Cai & Zhu (2015), Juddoo (2015), Kruse et al. (2016)
**Description:** Data is collected from various data source which can be of different types that increases the difficulty in integration and processing of the data in a short timeframe. An example source of new data types can be a camera that collects large amount of different types of data every single moment. It can collect new data such as color, different vehicle size, traffic signals, etc.
**Directly affects AI functions:** 2 "Yes", 2 "No"
**Challenge score:** Survey 1 - 2.405 (Rank 14/31), Survey 2 - 1.438 (Rank 22/25)

**Name:** Wrongly-calibrated / Defective Sensors
**Reference:** Interviewee C, Eur (2020)
**Description:** Wrongly-calibrated or defective sensors cause the value of the data collected differ from the real value. If that data is used to train AI models, it will not be effective since the training is done with wrong data. When real data is encountered by the AI models in the future, it will not perform as it should since it was trained with wrong data.
**Directly affects AI functions:** 4 "Yes", 0 "No"
**Challenge score:** Survey 1 - 2.643 (Rank 9/31), Survey 2 - 3.625 (Rank 3/25*)

#### 4.2.2.4 Data Structure Challenges

**Name** Fragmented Data (this challenged was initially called *Data Fragmentation* and was written in that form in Survey 1 and the initial collection of challenges. It was renamed to *Fragmented Data* later as it was realized that the earlier term was

actually a solution to a different problem unrelated with this thesis.)
**Reference:** ISO (2019)
**Description:** After data collection, it might be stored in different locations. While retrieving the data, data might need to be accessed from different locations and sources. This can be a challenge in training AI systems.
**Directly affects AI functions:** 1 "Yes", 3 "No"
**Challenge score:** Survey 1 - 1.333 (Rank 28/31*), Survey 2 - 2.000 (Rank 15/25*)

**Name:** Incompatible Data Formats
**Reference:** Kruse et al. (2016), Gao et al. (2016)
**Description:** There are predefined data formats that any data should be in so that the AI models can be trained using those data. However, if certain data is not in the predefined formats, the AI models would not be properly trained as they would be expecting data in certain format. There might be a need of manual intervention when AI models receive data in incompatible formats.
**Directly affects AI functions:** 1 "Yes", 3 "No"
**Challenge score:** Survey 1 - 1.333 (Rank 28/31*), Survey 2 - 2.438 (Rank 10/25)

**Name:** Outlier Data
**Reference:** Interviewee E, Corrales et al. (2016)
**Description:** The data that significantly differs from rest of the data is considered to be an outlier. Such data may significantly affect the analysis capability of the deep learning applications. However, not all outliers are bad. An example can be, when someone is trying to hack a car. This data can be an outlier as it does not necessarily happen frequently. However, if it is to happen, it should be detected as a real intrusion and not as an anomaly in data.
**Directly affects AI functions:** 2 "Yes", 2 "No"
**Challenge score:** Survey 1 - 1.433 (Rank 25/31), Survey 2 - 2.500 (Rank 9/25)

**Name:** Unstructured Data
**Reference:** Kruse et al. (2016)
**Description:** Deep learning algorithms mostly work well with structured data. Unstructured data is completely useless and training the algorithms with such data would lead to poor performance of the deep learning applications. For e.g., the pixels in the pictures are structured and deep learning algorithms perform well with structured data.
**Directly affects AI functions:** 1 "Yes", 3 "No"
**Challenge score:** Survey 1 - 1.833 (Rank 18/31), Survey 2 - 1.813 (Rank 21/25)

#### 4.2.2.5 Data Trust Challenges

**Name:** Incorrect Labeling
**Reference:** Interviewee C, Interviewee D, Interviewee E, Corrales et al. (2016),

Azeroual & Abuosba (2017)
**Description:** If dataset contains incorrect labels, it will affect the training. Neural networks do not have a way to segregate data in terms of correctness of labels. Incorrectly labeled data is useless as using such data could make the neural networks function in unintended manners. For e.g., as an incorrect label, a traffic light can be labeled as normal electric pole instead.
Directly affects AI functions: 4 "Yes", 0 "No"
**Challenge score:** Survey 1 - 2.667 (Rank 8/31), Survey 2 - 3.750 (Rank 1/25*)

**Name:** Lack of Good Data from Simulations
**Reference:** Interviewee D
**Description:** Simulations can be used as a technique for data collection, however, the industry strongly believes that data cannot be simulated in a good way. Simulated data might non cover all the scenarios and hence industry prefers collecting lot of random, real world data.
**Directly affects AI functions:** 4 "Yes", 0 "No"
**Challenge score:** Survey 1 - 1.800 (Rank 20/31), Survey 2 - 1.833 (Rank 20/25)

**Name:** Noise
**Reference:** Interviewee C, Corrales et al. (2016), Gupta & Gupta (2019)
**Description:** It refers to the data that is irrelevant and meaningless. There are various kind of noise that effects the performance of the deep learning algorithms which reduces the application's accuracy and predictive capability. Therefore it is extremely important to identify and handle the data noise. Data with low signal to noise ration is an example for noisy data
**Directly affects AI functions:** 2 "Yes", 2 "No"
**Challenge score:** Survey 1 - 2.452 (Rank 13/31), Survey 2 - 2.917 (Rank 7/25)

### 4.2.3 List of Data Quality Attributes

In this section, 82 data quality attributes are presented. Along with them, definitions of the attributes and challenges affecting those attributes are also provided.

**Table 4.3:** Template for *List of Data Quality Attributes* Artifact Component

| Field | Description |
|---|---|
| Name | Name of the data quality attribute |
| Reference | Reference that denotes the identification of the attribute |
| Definition | Description of the data quality attribute |
| Challenges affecting the DQ Attribute | List of challenges that affect the data quality attribute |

**Note:**

- NA: Not Applicable
- The numbers in the brackets are the weighted average values for the challenge-attribute association calculated from the results of the focus group session and survey 2.
- The first number inside the brackets denotes the weighted average from the results of the focus group and the second number denotes the weighted average from the results of survey 2.
- If there is no weighted average from either focus group or survey, the space is left blank. For example, (, 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 Section 4.3.1.

**Table 4.4:** 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 & 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 & Zhu (2015), Sidi et al. (2012), Wang & Strong (1996), ISO (2019), Eur (2020), CDDQ (2017) | *The conditions and modalities by which users can access, use and interpret data.* (Eur 2020),<br><br>*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) |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Accuracy** | Interviewees, Cai & Zhu (2015), Bobrowski et al. (1970), Sidi et al. (2012), Wang & Strong (1996), ISO (2019) | *The degree to which data values correctly represents real-world entities.* (Earley & Henderson 2017),<br><br>*The extent to which data are correct, reliable, and certified free of error.* (Wang & Strong 1996),<br><br>*Accuracy of data is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure.* (Eur 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. (1970), Wang & Strong (1996) | *The number of facts stored.* (Bobrowski et al. 1970),<br><br>*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 & Strong (1996) | *The extent to which the quantity or volume of available data is appropriate.* (Wang & Strong 1996) | NA |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Auditability** | Cai & 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) |
| **Authorization** | Cai & Zhu (2015) | *It refers to whether an individual or organization has the right to use the data.* Cai & Zhu (2015) | NA |
| **Availability** | Interviewees, Cai & Zhu (2015), Sidi et al. (2012), ISO (2019) | *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 2019) | 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 & Strong (1996), Cai & Zhu (2015), ISO (2019) | *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 2019),<br><br>*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) |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Clarity / Interpretability / Unambiguous** | Bobrowski et al. (1970), Sidi et al. (2012), Wang & Strong (1996), Eur (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** | Eur (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.* Eur (2020) | NA |
| **Comment** | Eur (2020) | *Supplementary descriptive text which can be attached to data or metadata.* (Eur 2020) | NA |
| **Completeness** | Interviewees, Cai & Zhu (2015), Bobrowski et al. (1970), Sidi et al. (2012), Wang & Strong (1996) | *Refers to whether all required data is present.* (Earley & Henderson 2017),<br><br>*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 (2019) | *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 (2019) | Data Ownership (, 1), Regulatory Compliance (, 1) |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Conciseness / Concise Representation** | Bobrowski et al. (1970), Sidi et al. (2012), Wang & 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 |
| **Confidentiality** | Eur (2020), ISO (2019) | *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.* (Eur 2020)<br><br>*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 2019) | Data Ownership (, 0.66), Regulatory Compliance (, 0.66) |
| **Consistency / Uniformity** | Cai & Zhu (2015), Bobrowski et al. (1970), Sidi et al. (2012), ISO (2019), Earley & Henderson (2017), CDDQ (2017) | *Can refer to ensuring that data values are consistently represented within a dataset and between datasets, and consistently associated across datasets.* (Earley & Henderson 2017),<br><br>*Measures whether or not data is equivalent across systems or location of storage.* (CDDQ 2017) | Data Drop (0.8, 1), Improper Data Transfer (0.6, 0.66), Incompatible Data Formats (, 0.66), Incomplete Data (0.6, 1) |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Consistency and Synchronization** | Earley & Henderson (2017), CDDQ (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 & 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** | Eur (2020) | *Individual or organisational contact points for the data or metadata, including information on how to reach the contact points.* (Eur 2020) | Regulatory Compliance (, 0.66) |
| **Correctness** | Interviewees, Bobrowski et al. (1970) | *Every set of data stored represents a real world situation.* (Bobrowski et al. 1970) | 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** | Eur (2020) | *Cost associated with the collection and production of a statistical product and burden on respondents.* (Eur 2020) | see *Cost Effectiveness* for the challenges affecting this attribute |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Cost Effectiveness** | Wang & 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, ) |
| **Currency / Currentness** | Sidi et al. (2012), Earley & Henderson (2017), CDDQ (2017), ISO (2019) | *The measure of whether data values are the most up-to-date version of the information.* (Earley & Henderson 2017),<br><br>*The degree to which data has attributes that are of the right age in a specific context of use.* (ISO 2019) | 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** | Eur (2020) | *Any change in a value of a statistic released to the public.* (Eur 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 |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Definition / Documentation** | Cai & 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 |
| **Ease of Manipulation** | Pipino et al. (2003), Sidi et al. (2012) | *The extent to which data is easy to manipulate and apply to different same format.* (Pipino et al. 2003) | NA |
| **Ease of Operation** | Wang & 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 |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **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 (2019) | *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 2019) | 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 & 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) |
| **Flexibility** | Wang & 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. 2003), (Sidi et al. 2012) | NA |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Frequency of Dissemination** | Eur (2020) | *The time interval at which the statistics are disseminated over a given time period.* (Eur 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** | Eur (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.* (Eur 2020) | Regulatory Compliance (, 1) |
| **Integrity** | Cai & Zhu (2015), Sidi et al. (2012), CDDQ (2017) | *Measures the structural or relational quality of datasets.* (CDDQ 2017) | NA |
| **Integrity or Coherence** | See *Integrity* and *Coherence* | | |
| **Latency** | Earley & Henderson (2017) | *The time between when the data was created and when it was made available for use.* (Earley & Henderson 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. 2008), (Sidi et al. 2012) | NA |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Lineage** | CDDQ (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.* (CDDQ 2017) | Data Acquisition (1, 1), Data Ownership (, 0.66), Regulatory Compliance (, 0.66) |
| **Metadata** | Cai & 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** | Eur (2020) | *The date on which the metadata element was inserted or modified in the database.* (Eur 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 |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Objectivity** | Bobrowski et al. (1970), Sidi et al. (2012), Wang & 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) |
| **Portability** | ISO (2019) | *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 2019) | 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. (1970), ISO (2019) | *The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.* (ISO 2019) | 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** | Eur (2020) | *Time lag between the actual delivery of the data and the target date when it should have been delivered.* (Eur 2020) | NA |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Quality Management** | Eur (2020) | *Systems and frameworks in place within an organisation to manage the quality of statistical products and processes.* (Eur 2020) | NA |
| **Readability** | Cai & 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 |
| **Reasonability** | Earley & Henderson (2017) | *Asks whether a data pattern meets expectations.* (Earley & Henderson 2017) | Data Drop (0.4, 0.5), Incomplete Data (0.8, 0.5) |
| **Recoverability** | ISO (2019) | *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 2019) | NA |
| **Reference Period** | Eur (2020) | *The period of time or point in time to which the measured observation is intended to refer.* (Eur 2020) | NA |
| **Release Policy** | Eur (2020) | *Rules for disseminating statistical data to all interested parties.* (Eur 2020) | Regulatory Compliance (, 0) |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Relevance** | Cai & Zhu (2015), Bobrowski et al. (1970), Sidi et al. (2012), Wang & Strong (1996), Eur (2020) | *The extent to which data are applicable and helpful for the task at hand.* (Wang & Strong 1996),<br><br>*The degree to which statistical information meet current and potential needs of the users.* (Eur 2020) | New Data Type (, 0.33) |
| **Reliability** | Cai & Zhu (2015), Bobrowski et al. (1970), Sidi et al. (2012) | *Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value.* (Eur 2020) | Data Drop (0.8, 1), Improper Data Transfer (0.8, 0.66), Incomplete Data (0.8, 1), Incorrect Labeling (1, 1) |
| **Representation** | CDDQ (2017) | *Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).* (CDDQ 2017) | NA |
| **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. 2008), (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 |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Statistical Presentation** | Eur (2020) | *Description of the disseminated data which can be displayed to users as tables, graphs or maps.* (Eur 2020) | NA |
| **Statistical Processing** | Eur (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 (Eur 2020). | NA |
| **Structure** | Cai & Zhu (2015) | *It refers to the level of difficulty in transforming semi-structured or unstructured data to structured data through technology.* (Cai & Zhu 2015) | Unstructured Data (, 0.66) |
| **Timeliness** | Cai & Zhu (2015), Bobrowski et al. (1970), Sidi et al. (2012), Wang & Strong (1996), Earley & Henderson (2017), CDDQ (2017) | *Length of time between data availability and the event or phenomenon the data describe.* (Eur 2020),<br><br>*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 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| 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 & Strong (1996), ISO (2019) | *The extent to which data are well documented, verifiable, and easily attributed to a source.* (Wang & Strong 1996),<br><br>*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 2019) | Data Acquisition (0.8, 1), Data Ownership (, 0.66), Regulatory Compliance (, 0.66) |
| **Transactability** | (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 |
| **Unambiguous** | Bobrowski et al. (1970) | *Each piece of data has a unique meaning.* (Bobrowski et al. 1970) | NA |
| **Understandability / Ease of Understanding** | Sidi et al. (2012), Wang & Strong (1996), ISO (2019) | *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 2019) | Incomplete Data (0.8, 0.75) |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Uniqueness** | Earley & Henderson (2017) | *No entity exists more than once within the dataset.* (Earley & Henderson 2017) | Redundant Data (, 1) |
| **Unit of Measure** | Eur (2020) | *The unit in which the data values are measured.* (Eur 2020) | NA |
| **Usability** | Interviewees, Cai & Zhu (2015), Bobrowski et al. (1970), 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 helpf.ul 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** | Earley & Henderson (2017), CDDQ (2017) | *Refers to whether data values are consistent with a defined domain of values.* (Earley & Henderson 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 & Strong (1996) | *The extent to which data are beneficial and provide advantages from their use.* (Wang & Strong 1996) | NA |

**Table 4.4:** List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

| DQ Attribute | Source | Definition | Challenge affecting the DQ attribute |
|---|---|---|---|
| **Variety of Data Sources** | Wang & 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** | Earley & Henderson (2017) | *Remain current for a short period.* (Earley & Henderson 2017) | NA |

### 4.2.4 List of Data Quality Attribute Metrics

Following section presents metrics associated with data quality attributes. It also provides formula for computation of those metrics. The data quality attributes without an applicable metric are listed as well.

**Table 4.5:** Template for *List of Data Quality Attribute Metrics* Artifact Component

| Field | Description |
|---|---|
| DQ Attribute | Name of the data quality attribute |
| Metric | Name of the data quality attribute metric |
| Formula | Formula used to calculate a given metric |

**Table 4.6:** 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<br><br>Degree of mandatory data completeness | number of available data records / total number of data records<br><br>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 |

**Table 4.6:** List of Data Quality Attribute Metrics

| DQ Attribute | Metric | Formula |
|---|---|---|
| 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 |
| Integrity or Coherence | (see *Integrity*) | |
| 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 |
| Timeliness and Availability | (see *Timeliness* and *Availability* | |
| Transactability | (see *Effectiveness*) | |
| Uniqueness | (see *Duplication*) | |
| Usefulness | (see *Effectiveness*) | |
| Variety of Data Sources | Number of data sources | number of data sources |

**Table 4.6:** List of Data Quality Attribute Metrics

| DQ Attribute | Metric | Formula |
|---|---|---|
| Volatility | Degree of volatility | number of data records that change within a given time period / total number of data records |

A number of data quality attributes do not have an applicable metric. A reason behind the lack of metric is that these attributes do not have a discernible numeric value associated with them. For e.g., *Comment* do 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

### 4.2.5 Potential Solutions

**Table 4.7:** Template for *Potential Solutions* Artifact Component

| Field | Description |
|---|---|
| Name | Name of the potential solution |
| Challenge it Tries to Solve | Denotes the challenge(s) a particular solution tries to solve |
| Requirement Specifications | These are the requirements that should be specified before implementing the solution |
| Implementation Details | This presents the stepwise implementation of the solution |

#### 4.2.5.1 Auto Increasing Sequential Number

**Challenge it Tries to Solve:** Data Drop
**Requirement Specifications:**

1. Add new field for auto-increasing sequential number in database.

**Implementation Details:**

- To handle data drop, firstly, the above mentioned requirement specification should be satisfied.
- Check if there is a field for auto-increasing sequential number in the dataset.
- If there exists a field for auto-increasing sequential number, check if each record in the dataset has a value in auto-increasing sequential number field.
- If the above condition is true, then, check if the auto-increasing sequential number field is in sequence and no values are missing.

- If each of the above conditions are not satisfied, then notify the appropriate stakeholders.
- Finally, if auto-increasing sequential number field is in sequence and no values are missing, then start processing the data.

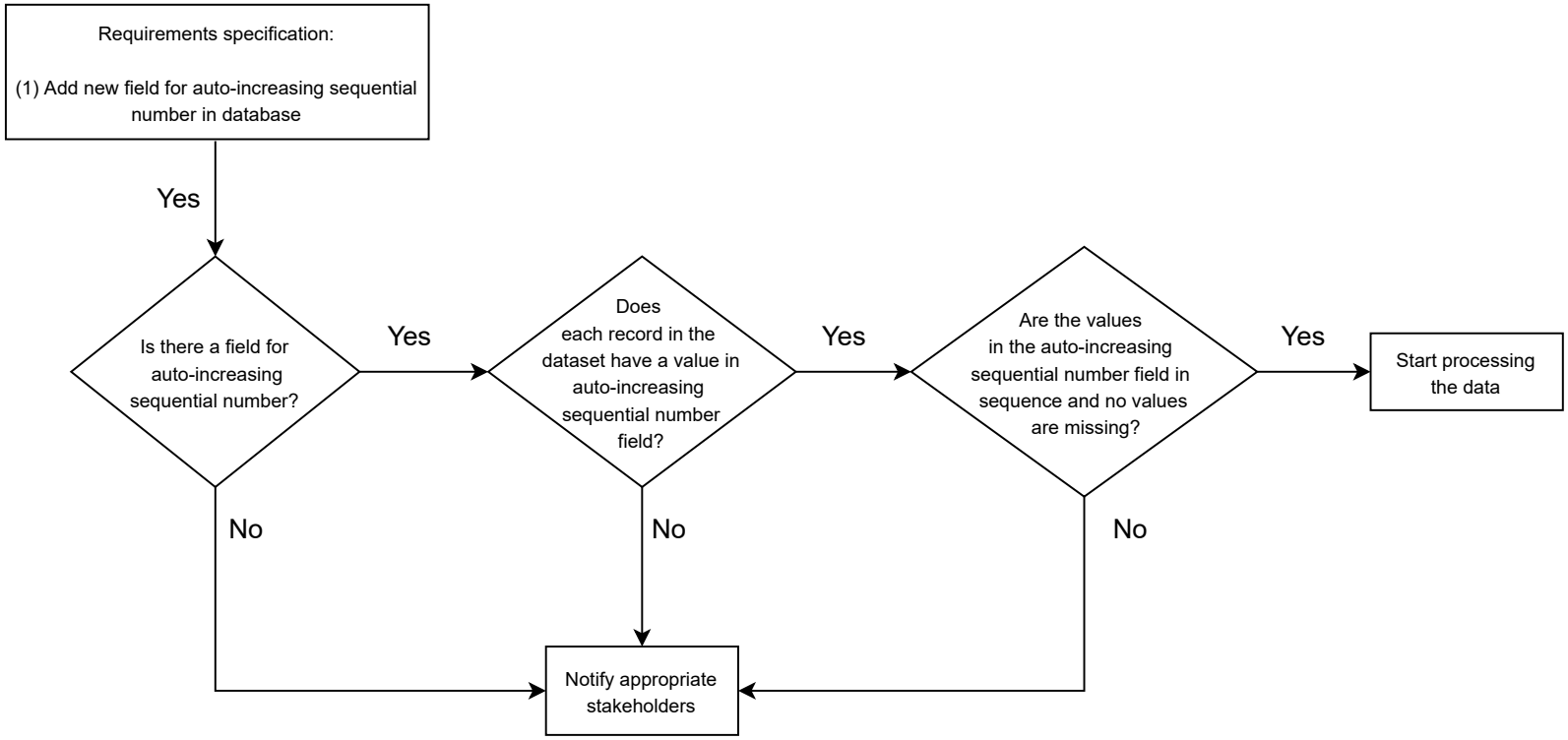



**Figure 4.3:** Flowchart for *Auto Increasing Sequential Number* Solution

#### 4.2.5.2 Automated Labeling

**Challenge it Tries to Solve:** Low Labeled Data Volume, Manual Data Labeling
**Requirement Specifications:**

1. Decide if the labeling is to be done image series or video

**Implementation Details:**

- First, above mentioned requirement specifications should be completed.
- While labeling each frame in an image series or a video, the labeler has to draw bounding box around each object manually. In this solution, instead of labeling each object in each frame, the labeler needs to label only few frames. Then, the pixels related with a particular object is tracked in succeeding frames until that object is no longer visible. For example, a vehicle's pixels can be labeled as belonging to that vehicle and those pixels can be tracked until that vehicle is visible in the frames. If new objects are introduced in the frame, they need to be manually labeled. Again, keep on tracking their pixel until they are visible in the succeeding frames.
- Cruz-Sandoval et al. (2019) describes two approaches of semi-automated data labeling in healthcare sector.
  - First, they propose using gesture recognition with smartwatches to label data. Gesture recognition software uses data from sensors such as accelerometer and gyroscope in a smartwatch. This approach works by recognizing a limited set of pre-determined "discrete" gestures. They use Dynamic Time Warping algorithm to determine if a signal contains a gesture. Next, they use Support Vector Machines and Sequential Minimal Optimization algorithms to recognize the gestures. Two experiments, in which 15 participants were asked to perform six gestures using different smartwatches, were conducted to assess this approach. With this experiment, they determined that by using automatic gesture recognition, "the system can reduce the burden of online, self-labeling".
  - Secondly, they study smart microphones to label "audible home activities". They use an Intelligent System for Sound Annotation (ISSA) to label audio. ISSA includes a sound detector, an audio classifier, and voice assistant functions. It uses already-trained models for audio classification. In an experiment to test this approach, they collected 10 samples of 2.5 to 5.2 seconds in range, from eight homes activities. With 3D direction, ISSA's accuracy was 87.27
- Namatevs et al. (2019) propose a set of neural networks to count moving objects in video. In their article, they cite the example of vehicle detection and enumeration. They use ImageNet dataset for object detection and MS COCO dataset for image captioning. They state that by dividing an image into

multiple grids, a “faster and more accurate convolution” can be formulated. Objects are detected from image pixels themselves in this technique. After object detection, they use a recurrent neural network with long short-term memory (RNN-LSTM) to count the objects.

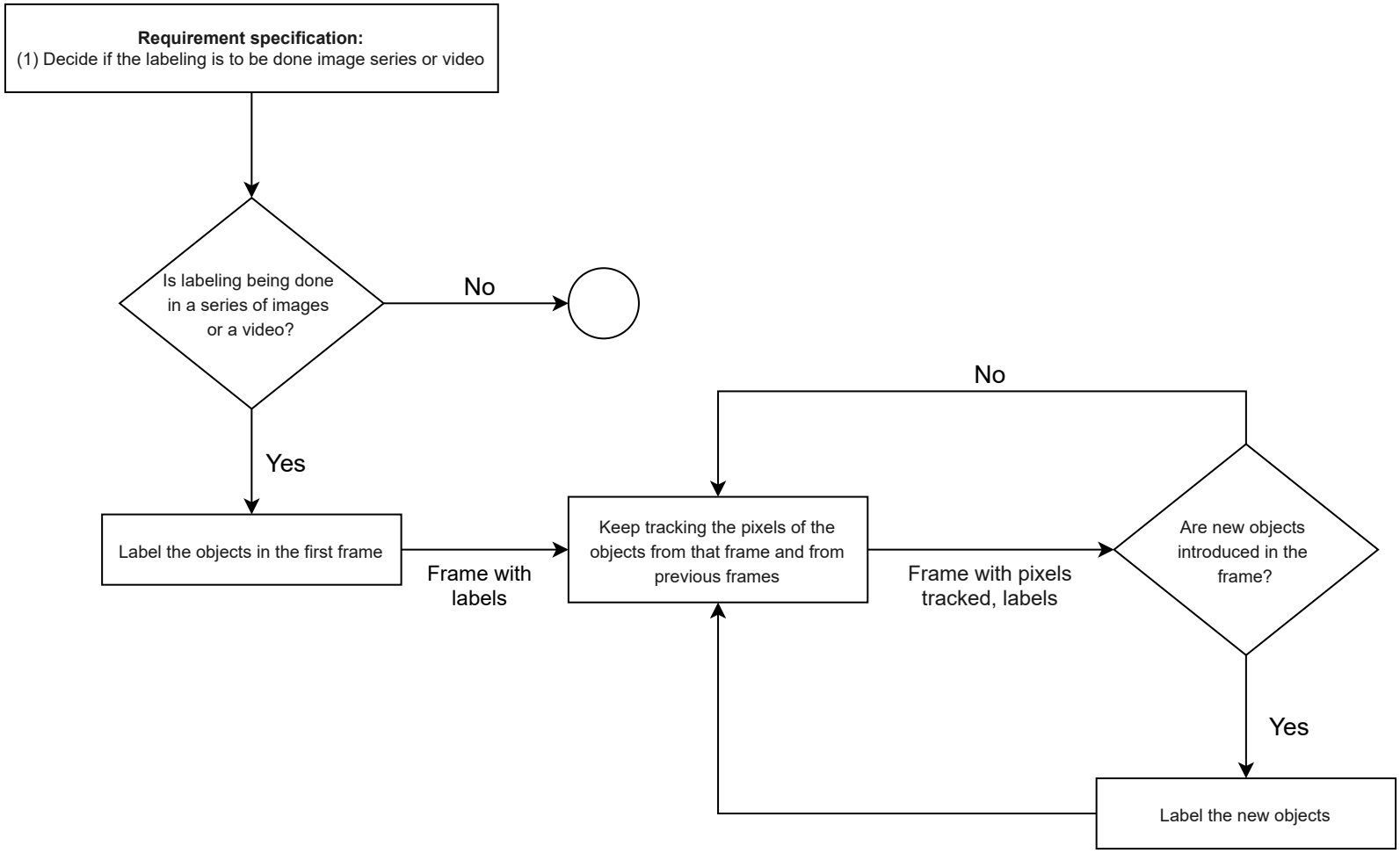



**Figure 4.4:** Flowchart for *Automated Labeling* Solution

#### 4.2.5.3 Continuous Data Processing

**Challenge it Tries to Solve:** Data Delay
**Requirement Specifications:**

1. Add new fields for departure timestamp and arrival timestamp in 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 the initial stage of processing.
- 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 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 data departure timestamp is there or not.
      - If data departure timestamp exists, compute the total time taken by finding the difference between arrival time and departure time.
      - Check if the time taken is within the acceptable range.
      - If it is within the acceptable range, stop.
      - 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.

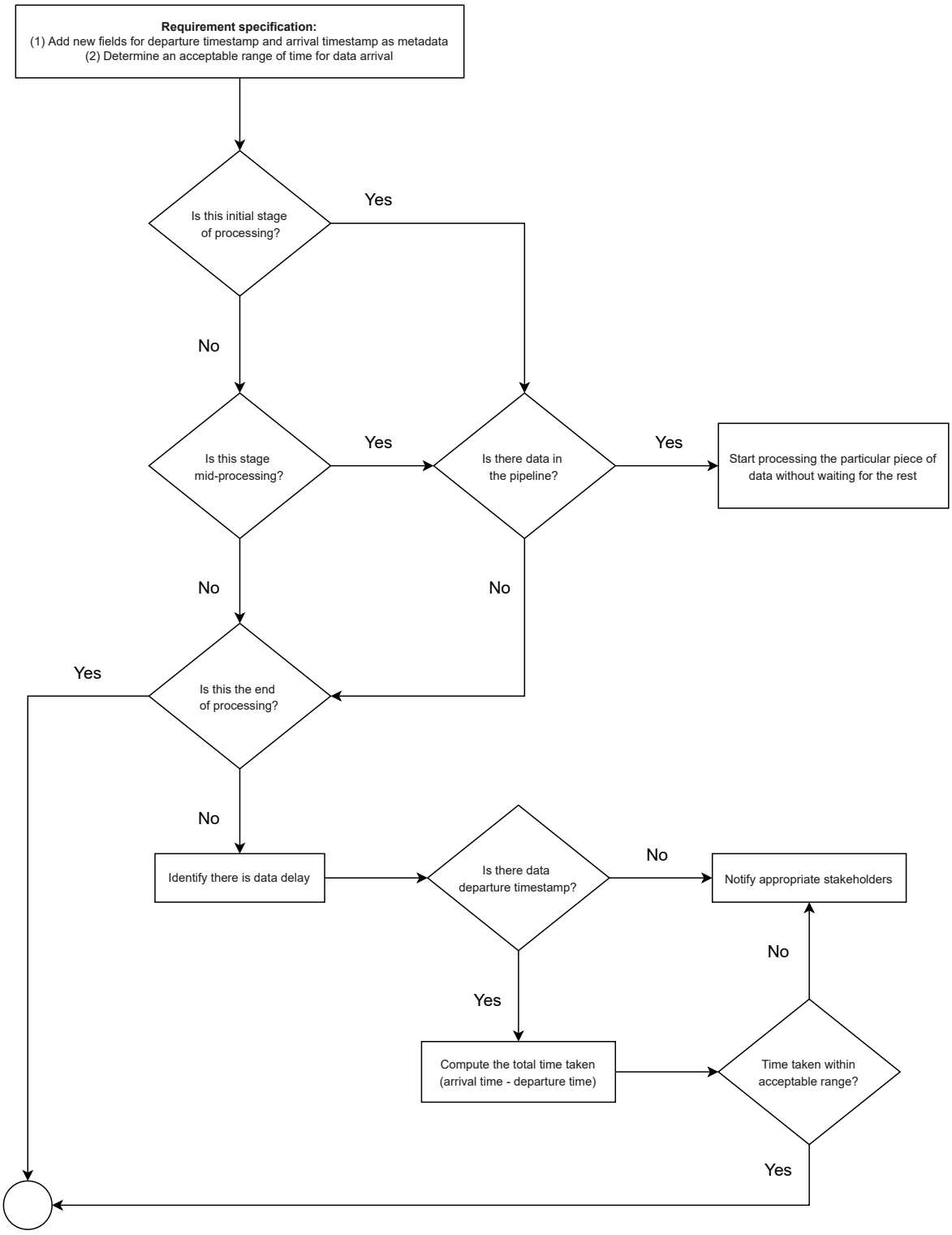



**Figure 4.5:** Flowchart for *Continuous Data Processing* Solution

#### 4.2.5.4 Corroboration of Data with Central Data Repository

**Challenge it Tries to Solve:** Data Dependent on External Conditions
**Requirement Specifications:**

1. Define the central data repository, its structure, and address,
2. Define the procedure if central data repository cannot be contacted,
3. Define the way AI disengagement notification is sent to the user

**Implementation Details:**

- First, above mentioned requirement specifications should be completed.
- Check if a central repository of data exists.
- If a central repository of data exists, contact that repository. Check if the required data is available in the repository.
  - If the required data is available in the repository, fetch the data, process it, and take appropriate steps.
  - If the required data is not available in the repository, disengage AI.
- If a central repository of data does not exist, check if the system is behaving properly.
  - If the system is behaving properly, process the existing data and take appropriate steps.
  - If the system is not behaving properly, disengage AI.
- When AI is disengaged, send a notification of the disengagement to the user.

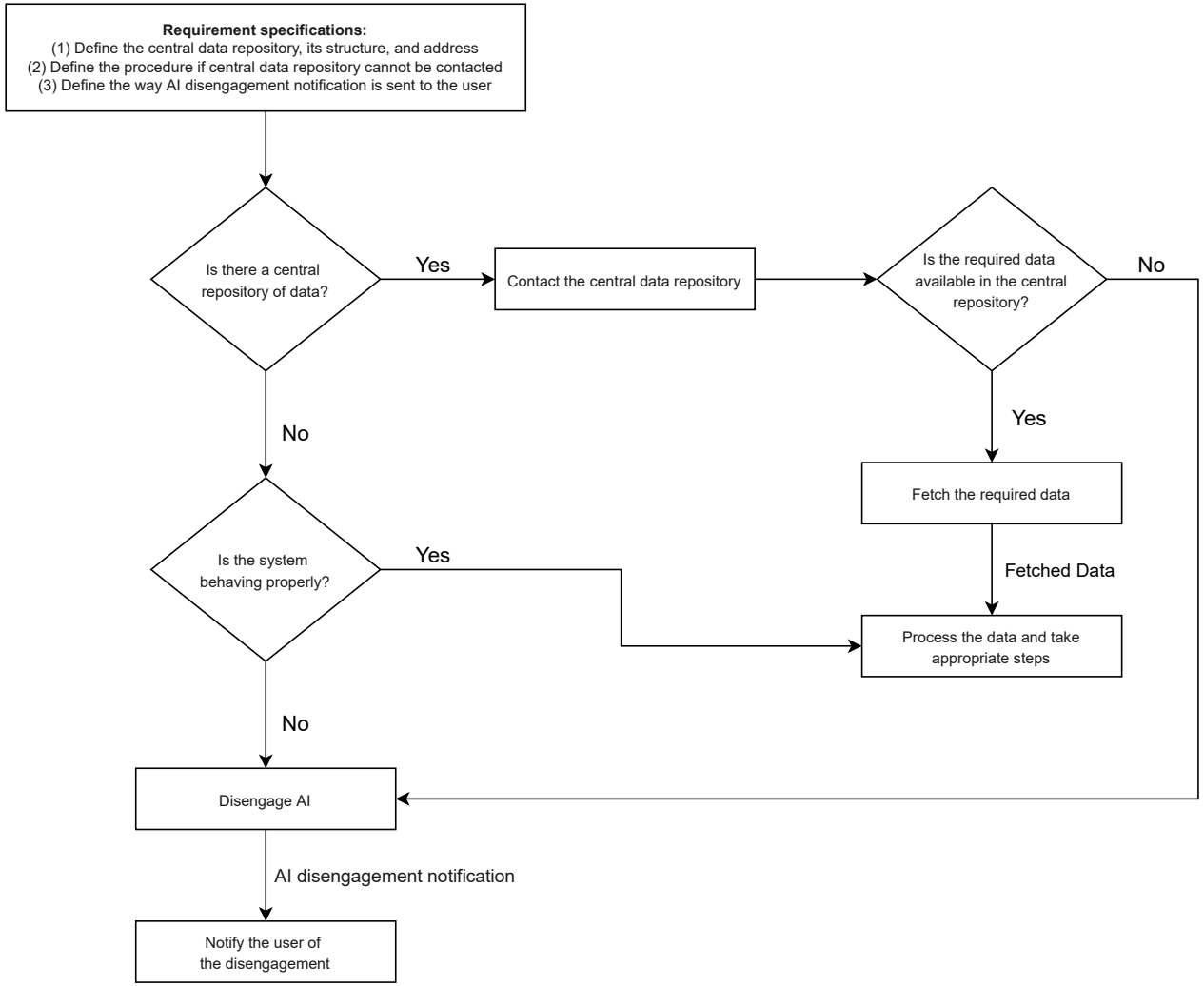



**Figure 4.6:** Flowchart for *Corroboration of Data with Central Data Repository* Solution

#### 4.2.5.5 Data Acquisition Solution Task

**Challenge it Tries to Solve:** Data Acquisition
**Requirement Specifications:**

1. Decide if data is collected from the sensor or purchased from third party sellers.
2. Specify data and data quality requirements.
3. Define methods for collection, transmission, and storage of data.

**Implementation Details:** NA

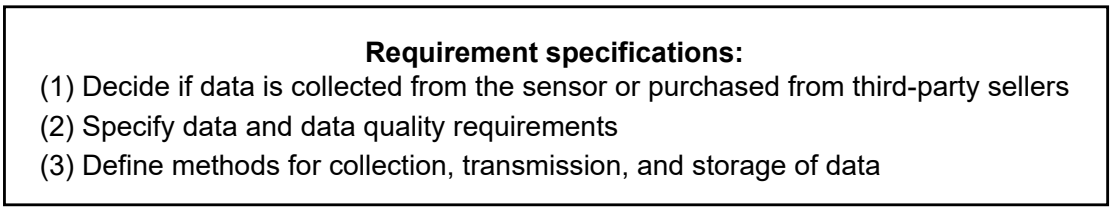



**Figure 4.7:** Requirement Specifications for *Data Acquisition* Solution Task

#### 4.2.5.6 Data Filter

**Challenge it Tries to Solve:** Reliance on Suppliers to Raise Error
**Requirement Specifications:**

1. Decide if data is collected from the sensor or purchased from third-party sellers,
2. Define the data requirements that should be checked by the filters,
3. Define the thresholds the data needs to comply with

**Implementation Details:**

- First, above mentioned requirement specifications should be completed.
- Check if the data is collected through sensors or comes from third-party sellers.
- If the data is collected through sensors, send it through the filters assigned for sensor-collected data.
- If the data is coming from third-party sensors, send it through the filters assigned for purchased data.
- After the data goes through either of the filters, check if there is any error in the filtration process.
- If there is any error in the filtration process, notify appropriate stakeholders. Else, start processing the remaining data.

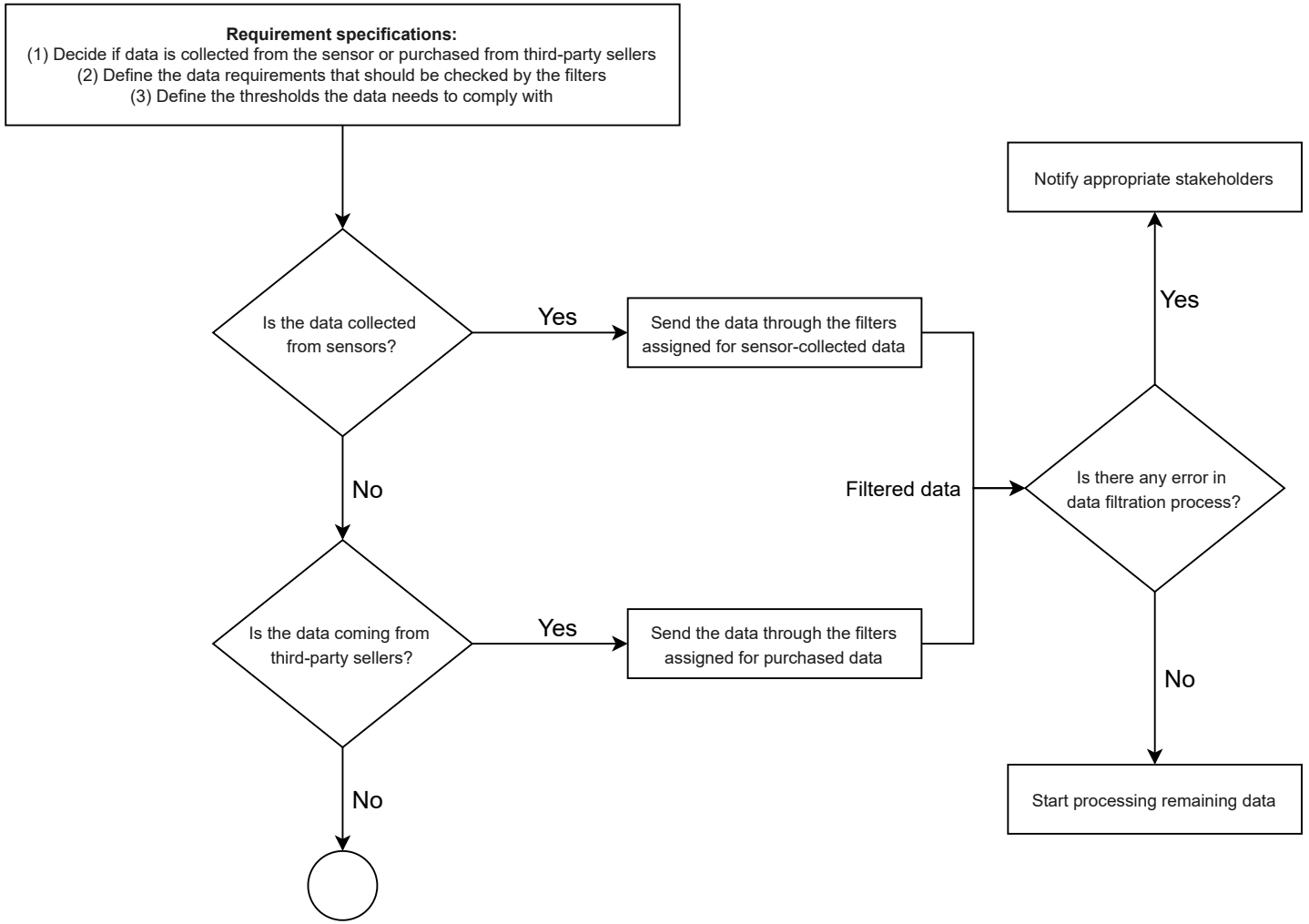



**Figure 4.8:** Flowchart for *Data Filter* Solution

#### 4.2.5.7 Data Level Methods and Algorithm Level Methods

**Challenge it Tries to Solve:** Imbalanced Dataset
**Requirement Specifications:** Not Applicable
**Implementation Details:**

- Kotsiantis et al. (2005), in their article, discuss *Data level* methods and *Algorithm level* methods to handle imbalanced dataset.

  Data level methods have various sampling techniques to handle imbalance dataset. These techniques are

  1. *Undersampling* randomly eliminates majority of class examples to achieve balance class distribution. The idiosyncrasies of the machine learning algorithm is overcome by balancing out the dataset.

  2. *Oversampling* does random replication of minority class examples to balance class distribution.

  3. *Feature Selection* framework separates the features for positive and negative classes and then combines them explicitly.

  Algorithm level methods comprises of three techniques to handle imbalance datasets. They are

  1. In *Threshold method*, a score is produced by neural networks that determines the degree to which an example belongs to the class.

  2. In *One-class learning*, the problem involves separating the single target class of samples from the novel samples that does not belong to the same class of the training set (Noumir et al. 2012).

  3. *Cost-sensitive learning* is an approach to incorporate cost in decision making to improve classifier performance. This is done to define fixed and unequal misclassification costs between classes.

#### 4.2.5.8 Identify Mandatory and Optional Fields

**Challenge it Tries to Solve:** Incomplete Data
**Requirement Specifications:**

1. Decide which fields should be mandatory and optional

**Implementation Details:**

- First, above mentioned requirement specifications should be completed. This is done to identify fields that are mandatory for effective functioning of deep learning algorithms.

- When data is collected and arrives at the function, check if all mandatory fields have data or not. If all mandatory fields have data in each record, start processing the data.
- If not, decide whether to remove the record or replace the value.
- Use listwise deletion or pairwise deletion to delete a record (Kang 2013) (Houari et al. 2016).
  - In listwise deletion, the record with missing data is delete in its entirety. Analysis is done with remaining data. According to Kang (2013), this method is suitable only if there is a large sample of data.
  - In pairwise deletion, a record is deleted only if "the particular data-point needed to test a particular assumption is missing" (Kang 2013).
  - If any other data is missing, the record is still used for analysis as those data do not affect the particular test assumption.
  - This method "preserves more information" than listwise deletion (Kang 2013).
- For replacing a value use method such as mean substitution, regression imputation, last observation carried forward, and multiple regression (Kang 2013).
  - In mean substitution, the missing value is replaced by the mean of all values in a particular field. Huisman (2000) improved the mean substitution by devising a formula for a Corrected Item Mean Substitution (CIM). According to him, the CIM "replaces missing values by the item mean which is corrected for the 'ability' of the respondent, i.e., the score on the observed items of the respondent compared with the mean score on these items."

$$CIM = \left( \frac{\#obs\,(j) \times PM}{\sum_{j \in obs(i)} IM} \right) \times IM$$

**Figure 4.9:** Corrected Item Mean Substitution (CIM)

    Here, PM is participant's mean and IM is item mean. Item mean is the mean of the "observed cases". Participant's mean is the mean of respondents who "answered a specific question" (Béland et al. 2018).
  - In regression substitution, a generated regression equation is used for missing value prediction. The missing value is replaced by the imputed

value (Chhabra et al. 2019). An example of the regression equation is linear regression ($y = ax + b$, where $y$ is the imputed value, $x$ is an auxiliary variable, $a$ is a coefficient, and $b$ is a constant)

  - In last observed carried forward method, a missing value is replaced by the last observed value in the particular field from the same object (Kang 2013). For example, if a value of 50.0 is measured in the previous observation, that is recorded as the value of current observation if the value in current observation is missing.

  - In multiple imputation, the missing values are replaced with "plausible" values calculated from an imputation model. This is done multiple times. Then, statistical analysis of interest is performed in each dataset independently. Then, a "single MI estimate" is derived by combining independent estimates (Rezvan et al. 2015).

- Cox et al. (2014) uses a number of methods to handle missing data in their paper. The methods include listwise deletion, pairwise deletion, mean imputation, regression imputation, hot-deck imputation, dummy-variable adjustment, maximum likelihood, and multiple imputation. They use survey data of over 5,000 students from 33 institutions to illustrate that "different missing-data approaches can lead to substantively meaningful differences when interpreting results." They also provide guidelines on appropriate methods of handling missing data. According to Cox et al. (2014), there is no "best" method of handling missing data, but a suitable method is "context specific".

- Jakobsen et al. (2017) analyzed 166 studies with keywords *"missing data"*, *"randomi*"*, and *"statistical analysis"* to study missing data handling techniques for randomized clinical trials. They take into consideration "strengths and limitations" of best-worst and worst-best sensitivity analyses, multiple imputation, and full information maximum likelihood methods. They also depict a flowchart of the steps of handling missing data.

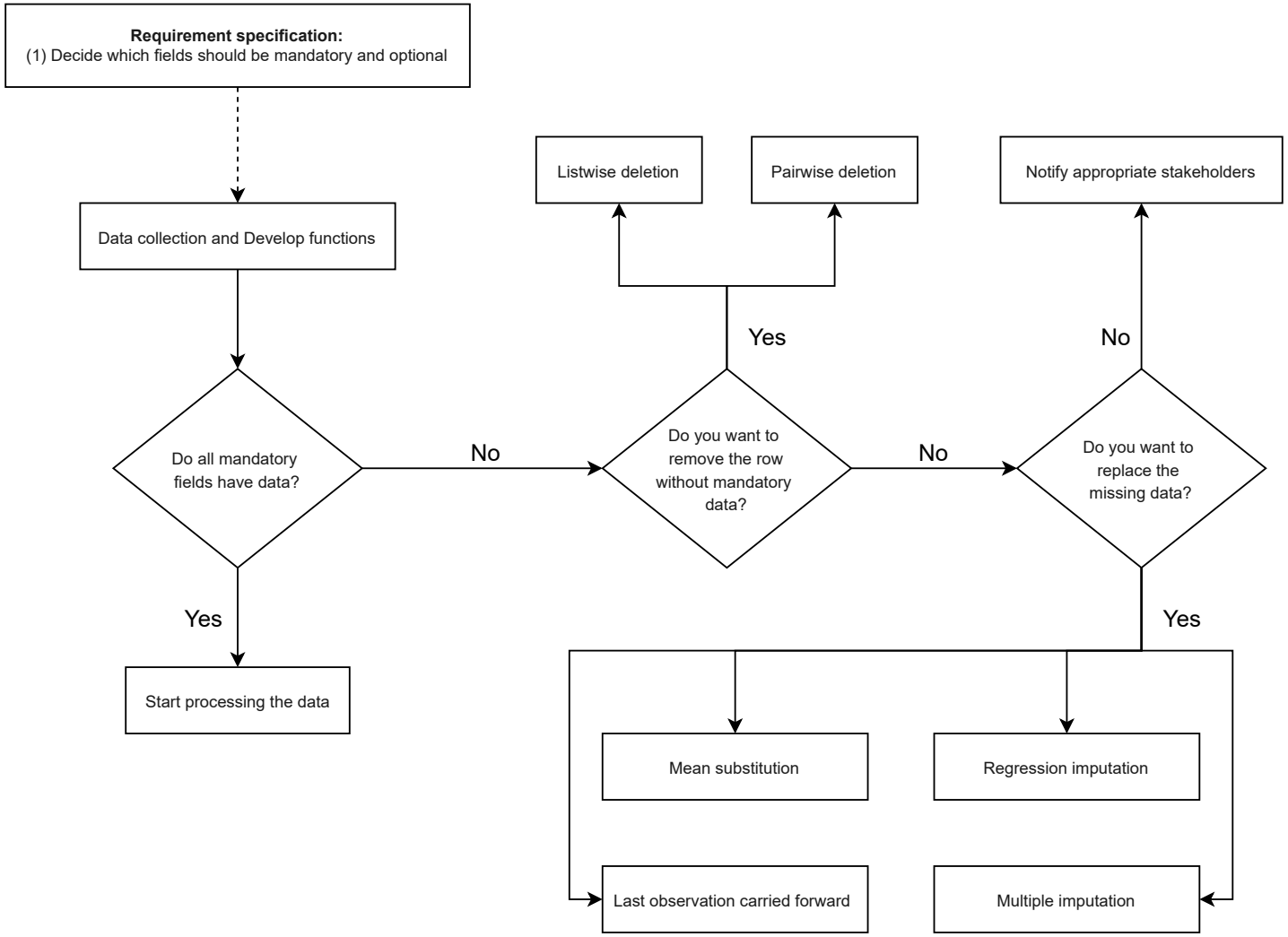



**Figure 4.10:** Flowchart for *Identify Mandatory and Optional Fields* Solution

#### 4.2.5.9 Improper Data Transfer Solution Task

**Challenge it Tries to Solve:** Improper Data Transfer
**Requirement Specifications:**

1. Specify the data transmission standards to follow.
2. Specify the modes of data transmission.
3. Specify the process of data transmission.

**Implementation Details:** NA

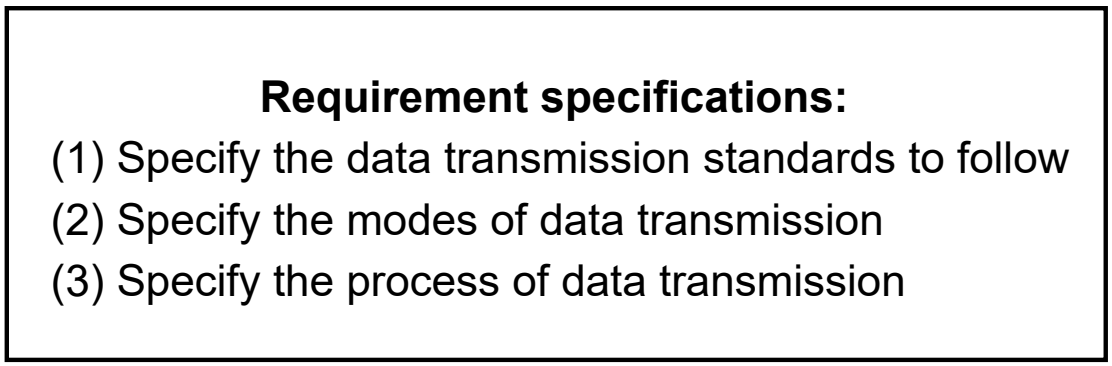



**Figure 4.11:** Requirement Specifications for *Improper Data Transfer* Solution Task

#### 4.2.5.10 Outlier Techniques

**Challenge it Tries to Solve:** Outlier Data
**Requirement Specifications:** Not Applicable
**Implementation Details:**

- According to Anscombe (1960), outlier data can be divided into two main categories - those arising from errors in data and those arising from inherent variability of the data.
- To handle outlier data, first, decide whether the step is to identify outlier data or treat outlier data.
- Two methods to identify outlier data include determining data outside interquartile range (Kwak & Kim 2017) and regression analysis (Gentleman & Wilk 1975).
  - To use the first method, "measure the distance between a data point and the center of all data points to determine an outlier. Based on this method, the data points that do not fall within three SD of the mean are identified as outliers" Kwak & Kim (2017). They also propose to use median and quartile range as alternatives as these statistics "are less senstive to outliers" (Kwak & Kim 2017).

  - Regression analysis, on the other hand, utilizes simple residuals that are "adjusted by the predicted values, and standardized residuals against the observed values to detect outliers" (Gentleman & Wilk 1975).

- There are also a number of ways to treat outlier data. Some of them are mentioned in this thesis. They are *Least trimmed squares* (Rousseeuw & Leroy 1987), *Windsorization* (Kwak & Kim 2017) (Osborne & Overbay 2004), *Least median of squares* (Rousseeuw & Leroy 1987), *Robust estimation method* (Osborne & Overbay 2004), *Trimming* (Kwak & Kim 2017), *Transformation* (Osborne & Overbay 2004), and *Truncation* (Osborne & Overbay 2004).

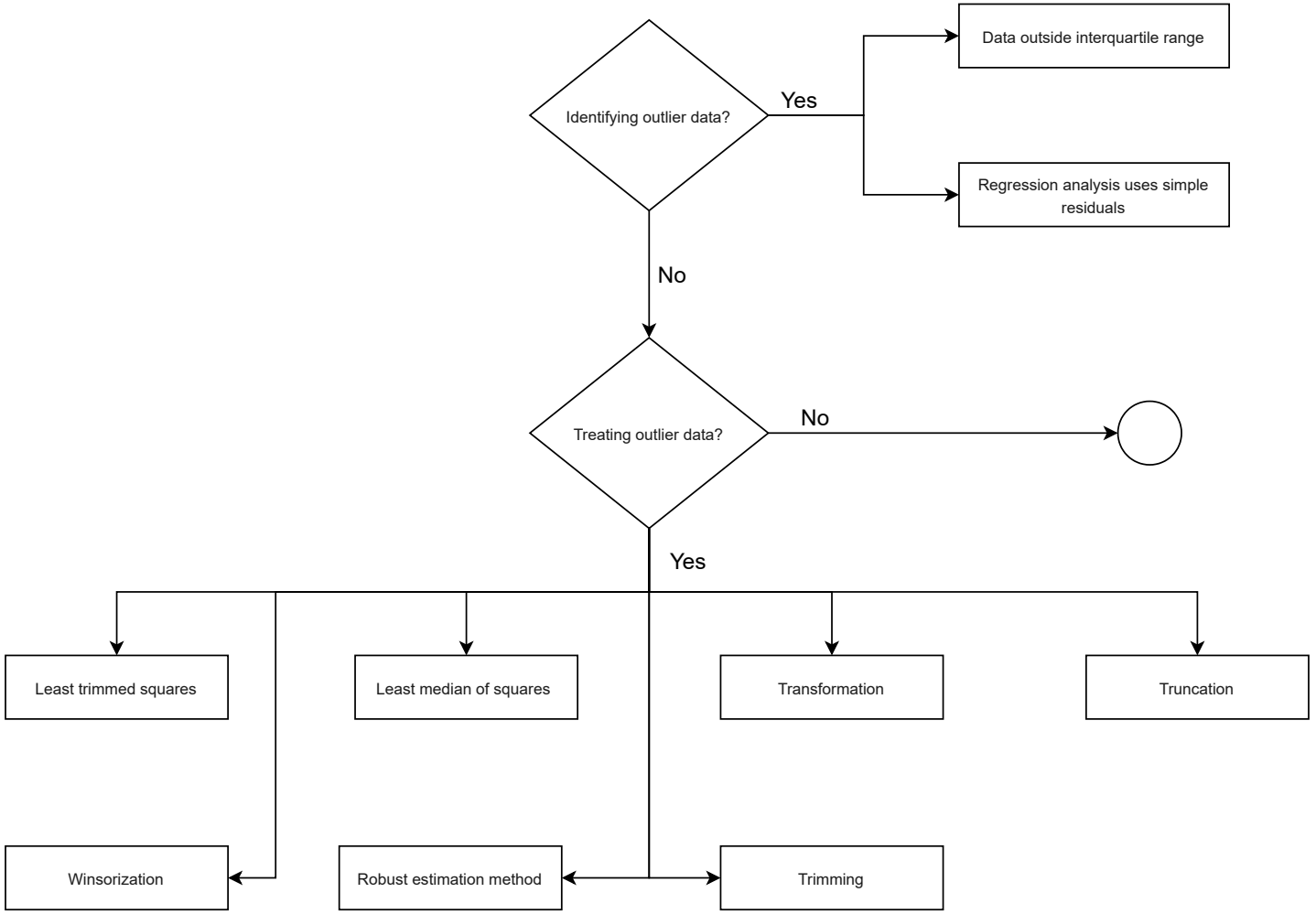



**Figure 4.12:** Flowchart for *Outlier Techniques* Solution

#### 4.2.5.11 Pair-wise Attribute Algorithm

**Challenge it Tries to Solve:** Noisy Data
**Requirement Specifications:** Not Applicable

**Implementation Details:**

- To handle noisy data, firstly, noise instance is identified using Pairwise Attribute Algorithm (PANDA) and checked if the noise exists or not (Gupta & Gupta 2019).
- If noise exists then, (Gupta & Gupta 2019) in thier article discuss that either ignore the noisy data, filter the noisy data, or alter the noisy data. The process of removing the attribute values that are erroneous from data set is termed as filtering. Alteration of noisy data is also called as polishing or data scrubbing or relabeling.

#### 4.2.5.12 RIASC Tool for Removing Redundancies (RTRR)

**Challenge it Tries to Solve:** Redundant Data
**Requirement Specifications:** Not Applicable
**Implementation Details:**

- DeCastro-García et al. (2018) developed a tool called *RIASC tool for removing redundancies* to "remove any unnecessary data, to compute the level of the redundancy, and to recover the original and filtered database" in a vector database.
- *RIASC tool* has three key features to remove redundancy, firstly, by removing the redundant variables to clean the input source. Second is to erase the redundant variables. Finally, the graphs are generated by variable recognition and eliminating the redundancies in the later stage.

#### 4.2.5.13 Test Environments

**Challenge it Tries to Solve:** Lack of Variety in Test Environment, Manual Data Collection
**Requirement Specifications:**

1. Define the types of environment in which in which data should be collected in, depending on the context.
2. Determine if real-world data collection and/or simulated environment data collection is suitable for the context.

**Implementation Details:**

- Check if the data should be collected from real-world environment or simulated environment.
- If the data should be collected from either of the environments, then collect data from different contexts.
- In terms of geographic location, data can collected from desert, highway, mountain, rural, snowy, and urban locations.
- In terms of temporal context, data can be collected during daytime, dawn, dusk, and nighttime.
- Vehicular data using different types of vehiles can be collected.
- In terms of the context of traffic density, data can be collected when traffic is congested, heavy, low, moderate, and when there are pedestrians.
- In terms of the context of weather, data can be collected in weather situations like cloudy, drizzle, fog, hurricane, overcast, partly cloudy, rainy, sandstorm, snowy, stormy, and sunny.

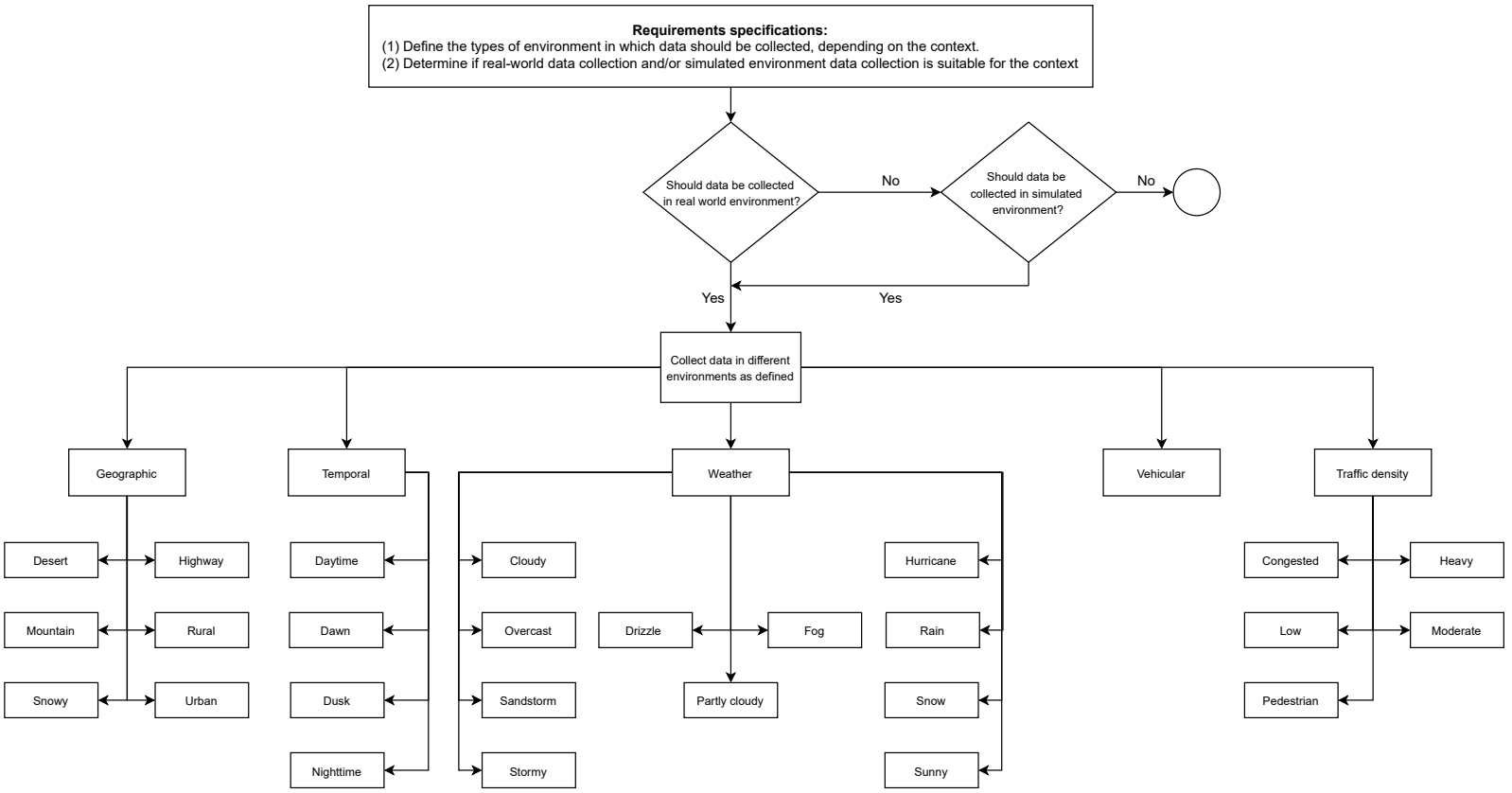



**Figure 4.13:** Flowchart for *Test Environments* Solution

## 4.3 Iteration 3 - Evaluation (RQ3)

### 4.3.1 Focus Group Results

A focus group session was conducted in the third iteration of this study. Five experts participated in the session. 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 than from the ones in the first iteration of this study. The second type of questions relates to the validation of the association between data quality challenges and data quality attributes.

The way the challenges were ranked in the focus group was different from the way they were ranked during the surveys. Unlike in the surveys, the ranking from the focus group session portrays the overall ranking of the challenges 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 focus group. Ranking of the challenge sets using a Likert scale is presented in Table 4.8. Ranking for challenges in each challenge set are presented in Tables 4.9, 4.10, 4.11, 4.12, and 4.13 . These rankings are without *Challenge Score* because of aforementioned reason. Detailed data from the focus group session is presented in Appendix A.6.

**Table 4.8:** 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 |

**Table 4.9:** Ranking of *Data Availability* Challenges

| Rank | Challenge |
|---|---|
| 1 | Incomplete Data |
| 2 | Low Labeled Data Volume |
| 3 | Data Drop |
| 4 | Data Delay |

**Table 4.10:** Ranking of *Data Management* Challenges

| Rank | Challenge |
|---|---|
| 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 |

**Table 4.11:** Ranking of *Data Source* Challenges

| Rank | Challenge |
|---|---|
| 1 | Lack of Variety in Test Environment |
| 2 | Data Dependent on External Conditions |
| 3 | New Data Types |
| 4 | Wrongly-calibrated / Defective Sensor |

**Table 4.12:** Ranking of *Data Structure* Challenges

| Rank | Challenge |
|---|---|
| 1 | Unstructured Data |
| 2 | Outlier Data |
| 3 | Incompatible Data Formats |
| 4 | Data Fragmentation |

**Table 4.13:** Ranking of *Data Trust* Challenges

| Rank | Challenge |
|---|---|
| 1 | Incorrect Labeling |
| 2 | Lack of Good Data from Simulations |
| 3 | Noise |

Fifteen data quality challenges were presented for validation regarding their association with data quality attributes. The number of participants ranged between 4-5 as some participants had to leave mid-session. In the opinion of the researchers of this thesis, the remaining challenges, except for the ones that are presented in

Survey 2, but not in focus group, do not affect any of the data quality attributes presented in *List of Data Quality Attributes* artifact component. So, they are not presented for validation.

107 data quality challenge-attribute associations were presented for validation during the focus group. The number of associations for each challenge is shown in Table 4.14. From the table, it can be seen that the experts regarded only 4 challenge-attribute associations as not valid (i.e., the initial supposition that the challenges affect the attributes for 4 of the attributes are not valid in expert opinion). Similarly, for 30 challenge-attribute associations, there was unanimity (i.e., all of the experts present in the focus group session regarded a particular challenge affects a particular attribute).

For 45 challenge-attribute associations, there were more than half, but not all, of the experts in the focus group regarding a particular challenge does affect a particular attribute (weighted average $> 0.5$). Similarly, for 26 challenge-attribute associations, there were more than half, but not all, of the experts in the focus group regarding a particular challenge does not affect a particular attribute (weighted average $< 0.5$). Only for *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*.

**Table 4.14:** Number of Associations of Data Quality Challenge and Data Quality Attributes and Weighted Average of Whether The Challenges Affect The Attributes (Yes-No)

| Data Quality Challenge | Number of Data Quality Attributes | 0 | 0.2 | 0.25 | 0.4 | 0.5 | 0.6 | 0.75 | 0.8 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Data Delay | 9 | 2 | 0 | 1 | 0 | 2 | 0 | 1 | 0 | 3 |
| Data Drop | 15 | 0 | 3 | 0 | 4 | 0 | 3 | 0 | 4 | 1 |
| Incomplete Data | 16 | 1 | 2 | 0 | 0 | 0 | 4 | 0 | 5 | 4 |
| Low Labeled Data Volume | 8 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 2 | 3 |
| | | | | | | | | | | |
| Data Acquisition | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 3 |
| Imbalanced Dataset | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| Improper Data Transfer | 7 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 3 | 0 |
| Manual Data Collection | 4 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Manual Data Labeling | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| Redundant Data | 4 | 0 | 1 | 0 | 1 | 0 | 2 | 0 | 0 | 0 |
| | | | | | | | | | | |
| Data Dependent on External Conditions | 3 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 |
| | | | | | | | | | | |
| Outlier Data | 6 | 1 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | | | |
| Incorrect Labeling | 10 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 8 |
| Lack of Good Data from Simulations | 6 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 4 | 0 |
| Noise | 5 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 1 |

### 4.3.2 Survey Results

In this thesis, the study was carried to understand the data quality challenges and how they affect the data quality attributes and devise appropriate solutions for the challenges. After the challenges were identified, five artifacts were developed. Out of them four are based on templates. The templates have various fields depending on the context of use. A survey was sent to industry partners to validate if the fields are appropriate for the templates. The survey results are presented in Tables A.29, A.30, A.31, and A.32.

The fields presented for validation of the template for *List of Challenge* artifact component were *Name*, *Sources*, *Description*, *Whether the challenge directly affects AI function*, and *Challenge Score*. There was unanimity among the participants for all of the fields except for the *Sources* field to be appropriate in this template. Only 75% of the participants agreed that the *Sources* field is appropriate for this template.

The fields presented for validation of the template for *List of Data Quality Attributes* artifact component were *Name, Sources, Definition,* and *Which challenges affect the data quality attribute*. There was unanimity among the participants for all the fields except for *Sources* field to be appropriate in this template. Only 50% of the participants agreed that the *Sources* field is appropriate for this template.

The fields presented for validation of the template for *List of Data Quality Attribute Metrics* artifact component were *Data quality attribute, Metric*, and *Formula*. There was unanimity among the participants for all the fields to be appropriate in this template.

The fields presented for validation of the template for *Potential Solutions* artifact component were *Name, Requirements specification*, and *Implementation details*. There was unanimity among the participants for all the fields to be appropriate in this template.

The participants in this survey were also asked for any fields that they think would be relevant for the fields. Out of the four participants, two provided response for the questions that tried to elicit novel fields. The new suggestions are shown in Table 4.15. These proposed fields can be added to the existing templates after further validation. However, they are not added in this particular version of the framework.

**Table 4.15:** List of Proposed Fields for Artifact Components in Survey

| Participant | Artifact Component | Proposed New Field | Reasoning |
|---|---|---|---|
| Participant 3 | *List of Challenges* | Methods for reducing the impact of the challenge on AI functions, models | |

**Table 4.15:** List of Proposed Fields for Artifact Components in Survey 2

| Participant | Artifact Component | Proposed New Field | Reasoning |
|---|---|---|---|
| Participant 4 | *List of Challenges* | Scope | Does the challenge affect only specific use-cases or models (and which ones), or is this a general issue? |
| Participant 4 | *List of Data Quality Attributes* | Relevance | To indicate the importance of the attribute |
| Participant 4 | *List of Data Quality Attribute Metrics* | Relevance | To indicate the importance of the metric |
| Participant 4 | *Potential Solutions* | Constraints | Specific constraints that might be required to apply the solution (e.g. processing power, statistical parameters, etc.) |

Likewise, the assumptions regarding the challenges that affect the AI models were validated via survey as well. The challenges were presented to the participants and asked to answer if the challenges affect the AI model. They were given a binary choice of a "yes" and a "no." Seven of the challenges were unanimously agreed by the participants that the challenges directly affect AI models. This can be seen by the weighted average score of 1 in Appendix A.33. Some of such challenges are *Low Labeled Data Volume*, *Imbalanced Dataset*, *Incorrect Labeling*, etc.

Three challenges were validated to affect AI model directly with a weighted average of 0.75. This means that more than half but not all of the participants agreed that these challenges affect the AI models. For eight challenges, only half of the participants agreed that the challenges directly affect AI model. This is shown in Appendix A.33 by a weighted average of 0.5. Seven challenges were validated to affect AI model with the weighted average of 0.25. This denotes less than half but all of the participants agreed that the challenges affect the AI models. For one challenge (*Expensive Procedure*) all of the participants agreed that it does not affect the AI model.

Nineteen data quality challenges were presented for validation regarding their association with data quality attributes. The number of participants for the survey was four but one of them has not provided response for all of the questions.

A total of 131 data quality challenge-attribute associations were presented in the survey. There was unanimity between all the participants that the associations

between 43 data quality challenge-attribute were valid. Some of such associations are *Data Delay* challenge and *Latency* attribute, *Incorrect Labeling* challenge and *Accuracy* attribute, *Outlier Data* challenge and *Fitness* attribute, etc.

Fifteen challenge-attribute associations had a weighted average of 0.75 and 39 challenge - attribute association had a weighted average of 0.66. This means that more than half but not all the participants agreed that 54 challenge-attribute association are valid (i.e., all of the participants agreed that 54 attributes are affected by the challenges). For 10 challenge-attribute association, half of the participants agreed that the associations were valid; the other half did not. Thirteen challenge-attribute associations had a weighted average of 0.33 and five challenge-attribute association had a weighted average of 0.25. This means that less than half but not all the participants agreed that 18 challenge-attribute association are valid. Similarly, for 6 challenge-attribute associations, none of the participants regarded them as valid. Such associations are *Data Delay* challenge - *Portability* attribute, *Data Dependent on External Conditions* challenge - *Accuracy* attribute, *Imbalanced Dataset* challenge - *Efficiency* attribute, *Incomplete Data* challenge - *Flexibility* attribute, *Regulatory Compliance* challenge - *Release policy* attribute, and *Unstructured Data* challenge - *Credibility* attribute.

This survey also included questions to identify severity of the challenges. The questions for this part of the survey asked participants to rank each of the challenges of respective challenge sets in order of severity. They were also asked to provide a Likert scale value for the challenge sets. The challenges are listed in descending order based on the challenge score.

As survey was conducted in two cycles, the results from them can be compared. During first cycle survey, the top three challenges based on the scores given by the participants are *Low Labeled Data Volume*, *Lack of Variety in Test Environment* and *Incomplete Data*. For second cycle survey, *Low labeled Data Volume* remains he most pressing challenge. However, *Incorrect Labeling* also rose to the top position as it has the same score (3.750) as *Low labeled Data Volume*. Two challenges - *Wrongly-calibrated / Defective Sensors* and *Lack of Variety in Test Environment* are in the third position with a score of 3.625.

The bottom three challenges in the first cycle survey, in the order from the third last to the last, are *Fragmented Data*, *Fast Increasing Data*, and *Data Acquisition*. However, *Fast Increasing Data* is removed from the final version of the artifact. This is depicted by the gray row in Appendix A.3. In second cycle survey, *Data Acquisition*, *Data Delay*, and *Redundant Data* ranked as the bottom three.

The challenges that were removed from the final version of the artifact are *Uncertain Data Quality Identification*, *Wrong Metadata*, *Reliance on a Single Data Source*, *Data Mix-up*, *Fake Data*, and *Fast Increasing Data*. They were removed because of them being vague and/or identified as not being a challenge during interviews. However, two new challenges were added for the second cycle survey. This is also shown by gray rows in Appendix A.5. The newly added challenges that were not part

of the first cycle survey are *Time Consuming* and *Expensive Procedure*. *Regulatory Compliance* challenge was not part of the second cycle survey due to technical error and hence, does not carry any score.

# 5

# Discussion

In this chapter, the entirety of this study is evaluated and discussed. The researchers' opinions are provided and future work is prescribed. Implications to research and academia are provided in Section 5.1. Similarly, implications to practitioners are provided in Section 5.2. In Section 5.3, validity and ethical considerations of this study are presented.

## 5.1 Implication to Research

The study conducted for this thesis can have implications on research in the field of data quality. The thesis provides associations between data quality challenges and data quality attributes. A comprehensive list of such associations have not been previously provided by a research. The artifact developed for this thesis can provide a building block for researchers to further research in the field of data quality. The templates devised for each of the artifact components can be improved through a large-scale expert analysis. The artifact is also useful for researchers as it provides a step-by-step workflow to manage data quality requirements.

Researchers could use *List of Challenges* artifact component to study the challenges in detail. They can home in on intricacies of individual challenges and identify sub-challenges. They could identify other attributes and incorporate them in *List of Data Quality Attributes* artifact component.

Furthermore, as a future work, researchers can study the artifact components proposed in this thesis and improve them. They could employ expertise from more people than the experts that were interviewed and/or surveyed during the iterations of this thesis. As the artifact developed in this thesis is general in nature, not only can researchers improve it for the field of AD, but they can upgrade it to encompass other fields as well. For instance, they could consult with experts in healthcare domain and improve the artifact such that it could identify data quality challenges, data quality attributes, and potential solutions in that field.

In the same way, researchers can also use the *List of Data Quality Attribute Metrics* artifact component presented in this thesis in order to develop novel metrics that could provide more accurate and effective numerical representation of the at-

tributes. In addition to improving the potential solutions mentioned in Section 4.2.5, researchers can devise solutions for those challenges that have not been provided a potential solution for in this thesis.

Researchers could implement the artifact components developed in this thesis in preferably more than one real-world application and evaluate the effectiveness of each of those components. They could use the results obtained during such implementation to enhance the components so that they better suit the context of use.

## 5.2 Implication to Practitioners

The research work conducted in the thesis can impact practitioners as well. Practitioners can be any party implementing the artifact in a real-world scenario. For instance, they could be requirements manager, software engineer, data quality specialist, etc. in a company or an organization. First, as for the researchers, the *Data Quality Workflow* artifact component can also be used by practitioners in order to have a consistent workflow of data quality management.

Similarly, the artifact presents a list of metrics and formula to calculate those metrics (Section 4.2.4). Practitioners can record data in order to derive values for the metrics. The metrics can help practitioners to adapt and make changes to their processes if needed.

The artifact also provides potential solutions to data quality challenges. Although it is not an exhaustive list of solutions, it provides a starting point for practitioners to improve the listed solutions and devise novel ones.

The study conducted for this thesis can also be utilized by practicing researchers from the industry as well. Since industry also can include a research and development department, the researchers in such department can use the artifact in their study of data quality for the system they are developing.

A comparison can be made between the framework proposed in this study and the OpenMDM framework described in the Section 2.1. A difference between the two frameworks is that OpenMDM provides a workflow management of measurement data whereas the DQAMF provides a workflow for overall management of data quality. Similarly, OpenMDM is a Eclipse IDE based tool. Whereas, DQAMF could be employed in a programming language-neutral and IDE-neutral fashion. However, as stated below, the framework can be converted into a software tool, which could be dependent on certain things.

A future work relating with the artifact could be adoption of it as an automated tool. Data can be passed through a pipeline in this tool and checked for quality. With it, the quality of the data could be assessed. Then, quality information can be presented to appropriate stakeholders using different medium and visualization techniques.

## 5.3 Validity and Ethical Considerations

### 5.3.1 Internal Validity

Internal validity is concerned with how different variables affect the result of an experiment. The researchers can have their own biases regarding the research topic. For instance, bias can be during data collection, conducting interviews, and so on. To mitigate this, interviews with many experts were conducted to gain a wider perspective. The researchers were in constant communication with supervisors and industry partners to reduce such biases. Similarly, both the researchers did the coding separately using the same coding technique and then combined the independent codes during "code combination meetings." Brainstorming process was performed during those meetings. The outcome of these meetings was a single final set of codes. This was done in both the iterations regardless of thematic analysis techniques. However, since most of the interviews were conducted with the experts from the case company, that could have led to its own bias to be formed. Furthermore, all of the experts were from ADAS and AD domains. There were no experts from other domains that were interviewed. There were some challenges that were not elicited from the experts, but were identified through literature review. Those challenges include Unstructured Data, Incompatible Data Type, Fragmented Data, Regulatory Compliance, Manual Data Collection, Large Volume of Data, Imbalanced Dataset, and Data Acquisition. However, these challenges were validated in survey and focus group by the experts and hence are part of challenges. Similarly, a predefined set of questions were used for the interviews, which could have limited the discussion during interview sessions. To solve the issue, at the end of each interview, the interviewees were asked if any questions that should have been asked were missed.

### 5.3.2 Construct Validity

Construct validity is "the extent to which the measurements used, often questionnaires, actually test the hypothesis or theory they are measuring" (Ginty 2013). The thesis was briefly explained to each interviewee before interviews so as to mitigate any doubts about the purpose of the study. It focused the discussion on the study of data quality and did not let the discussion to be deviated. The potential solutions devised during the second iteration of study were presented and explained to the interviewees. The motivation for this presentation was to validate the thought process behind the solutions and to find if there are any problems with the solutions. If there were any problems, ways to fix them were also tried elicited from the interviewees. While the solutions were validated, the assumptions that were made about the data quality challenge themselves were also validated. However, the order of the questions that were asked might not be optimum. Efforts were made to reduce ambiguity in the questions as much as possible. However, there could have still been confusion regarding the questions because of reasons such as gaps in communication. There is always scope for improvement in terms of the order of questions and their clarity.

### 5.3.3 External Validity

External validity pertains to the generalizability of the research. Although the outcome of a DSR is supposed to be based on a particular context and not necessarily need to be generalizable, this thesis study has tried to make it as generalizable as possible. It was done by reviewing literature that discussed data quality in different domains. The literature review regarding data quality and data quality attributes were not focused on autonomous vehicles but data quality and data quality frameworks in general. Another threat to validity could be the case company providing the researchers access to those experts that would provide only that information that bolster the viewpoint of the case company. To mitigate this threat, experts from other organizations such as Trafikverket[1] and CEVT[2] were also interviewed. Additionally, a focus group session was conducted which allowed open discussion regarding the topic of interest. Furthermore, artifact components of this thesis are generalizable as well since they do not explicitly mention autonomous domain and hence can be applicable in any field.

### 5.3.4 Reliability

Reliability concerns with the replicability of an experiment, i.e., future experiments that are designed in the same fashion as the first experiment should produce same results as the first experiment. To facilitate this, interview questions for both iterations are provided in Appendix A.1.2. The different versions of interview questions are also provided so that researchers can see how the research questions evolved based on the response from the participants. The interview questions help researchers to ask similar questions in the future. However, the responses by experts might be different despite them being from the same domain and having similar years of experience. This is because they could have different backgrounds and experiences over the course of their careers or simply because they can have different perspectives. Because of this, the responses any future researchers might elicit do not necessarily have to be similar to the results of this thesis. Future technologies could also influence the response of the experts involved in future experiments. If researchers conduct experiments in the future with only a few number of participants, then the result might differ from the result this study. However, if the data is collected from a large number of participants (i.e., a large sample size), then the result of that experiment could converge with the results of this study.

### 5.3.5 Conclusion Validity

Conclusion validity deals with the reasonability of the results of an experiment. Since a focus group session and a survey was conducted to evaluate the artifacts developed in the thesis, it can be stated that the conclusion of this study is valid. However, the researchers of this study have not validated the conclusion with other domain experts like from healthcare, aerospace, law enforcement, and so on. The artifacts have not been implemented in real world context. So, there is scope for

[1] `https://www.trafikverket.se`
[2] `https://www.cevt.se`

future study in terms of real-world implementation of the artifact developed in this thesis.

### 5.3.6 Informed Consent

A standardized consent form was created in which information about the interviewee was filled up. The questions asked in the interviews are presented in Appendix A.1.1. The interviewees were presented this form prior to the interviews. Consent was taken from the interviewees for audio recording of the interview session. Similarly, consent was taken to be able to use the interview data for future research and publication. The interviewees were also informed that the results of the thesis would be shared with them after the completion of the study. Likewise, the survey participants were presented a brief explanation about the thesis study so that they understand the motive behind the thesis.

### 5.3.7 Confidentiality and Anonymity

The interview participants were ensured that their personal data would be kept confidential and would only be used for the purpose of the study. Any personal data that could be used in publication would be anonymized. For example, the participants' names are anonymized in Table 3.1 and 3.2. The survey itself was anonymous as no personal information was solicited. Hence, there was not any agreement regarding confidentiality of survey data. There was also an understanding with the case company that the results of the study would be first shared with them before publication to a larger audience.

# 6

# Conclusion

This study based on DSR tries to approach the challenges associated with data quality attributes in distributed deep learning systems in the context of autonomous driving iteratively. Furthermore, the solution is designed as a set of artifacts which are then evaluated with different validation techniques such as focus group and survey.

*Data Quality Workflow* component which is comprised of 6 steps can act as a guide in proper management of data quality. Each step clearly defines the process to be followed to ensure data quality.

The set of 13 broad themes that were identified categorized different codes based on their relevancy. The researchers were also able to identify a set of 27 data quality challenges which had the potential to affect AI models. Challenge Sets and Challenge Score further helped in determination of severity of challenges. Based on the survey, the study was able to identify that *Low Labeled Data Volume*, *Lack of Variety in Test Environment*, and *Incorrect Labeling* are three of the most pressing challenges.

All components of the framework and their associated templates help in better comprehension of the challenges, attributes, metrics, and solutions. The templates act as a single point of reference to industry practitioners and academic researchers by enabling them to refer and make decisions to improve data quality and define data quality requirements for their systems. The metrics identified and the formula derived could be beneficial in improving the degree of data quality. This would highly improve the performance of the AI systems in any domain to make better predictions and mitigate the risks that are caused due to bad data quality. Similarly, by providing associations between data quality attributes and data quality challenges, related stakeholders can tailor their system requirements accordingly. With this information they can proactively try to mitigate the challenges to improve respective data quality attributes.

The assumptions that the researchers had were validated. More than 75% of the assumptions regarding association between data quality challenges and data quality attributes were validated by at least half of the experts. Only around 4% of the associations were invalidated by the experts.

This research could utilize further study to improve the framework developed and further validation from implementation in real world application perspective. This could help evolve the framework developed and make it more generalizable in broader use cases and domains.

# Bibliography


Alhojailan, M. I. (2012), 'Thematic analysis: A critical review of its process and evaluation', *West East Journal of Social Sciences* **1**(1), 39–47. Available: https://fac.ksu.edu.sa/sites/default/files/ta_thematic_analysis_dr_mohammed_alhojailan.pdf.

Anscombe, F. J. (1960), 'Rejection of outliers', *Technometrics* **2**(2), 123–146. DOI: https://doi.org/10.2307/1266540.

Azeroual, O. & Abuosba, M. (2017), 'Improving the data quality in the research information systems', *International Journal of Computer Science and Information Security* **15**(11), 82–86. Available: https://arxiv.org/pdf/1901.07388.pdf.

Batini, C., Barone, D., Mastrella, M., Maurino, A. & Ruffini, C. (2007), A framework and a methodology for data quality assessment and monitoring, *in* 'Proceedings of the 12th International Conference on Information Quality', MIT, Cambridge, MA, USA, pp. 333–346. Available: http://mitiq.mit.edu/iciq/pdf/a%20framework%20and%20a%20methodology%20for%20data%20quality%20assessment%20and%20monitoring.pdf.

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. (2009), 'Methodologies for data quality assessment and improvement', *ACM Computing Surveys* **41**(3), 1–52. DOI: https://doi.org/10.1145/1541880.1541883.

Bobrowski, M., Marré, M. & Yankelevich, D. (1970), 'A software engineering view of data quality'. Available: https://www.researchgate.net/publication/2598252_A_Software_Engineering_View_of_Data_Quality.

Borrego-Carazo, J., Castells-Rufas, D., Biempica, E. & Carrabina, J. (2020), 'Resource-constrained machine learning for adas: A systematic review', *IEEE Access* **8**, 40573–40598. DOI: https://doi.org/10.1109/ACCESS.2020.2976513.

Béland, S., Jolani, S., Pichette, F. & Renaud, J.-S. (2018), 'Impact of simple substitution methods for missing data on classical test theory difficulty and discrimination', *The Quantitative Methods for Psychology* **14**(3), 180–192. DOI: https://doi.org/10.20982/tqmp.14.3.p180.

Cai, L. & Zhu, Y. (2015), 'The challenges of data quality and data quality assessment in the big data era', *Data Science Journal* **14**(2), 1–10. DOI: https://doi.org/10.5334/dsj-2015-002.

CDDQ (2017), 'List of conformed dimensions of data quality'. Available: https://dimensionsofdataquality.com/alldimensions.

Chen, M., Ebert, D., Laramee, R., van Liere, R., Ma, K.-L., Ribarsky, W., Scheuermann, G. & Silver, D. (2009), 'Data, information, and knowledge in visualization', *Computer Graphics and Applications, IEEE* **29**(1), 12 – 19. DOI: https://doi.org/10.1109/MCG.2009.6.

Chen, X.-W. & Lin, X. (2014), 'Big data deep learning: Challenges and perspectives', *Access, IEEE* **2**, 514–525. DOI: https://doi.org/10.1109/ACCESS.2014.2325029.

Chhabra, G., Vashisht, V. & Ranjan, J. (2019), 'A review on missing data value estimation using imputation algorithm', *Journal of Advanced Research in Dynamical and Control Systems* **1**, 312–318. Available: https://www.jardcs.org/abstract.php?id=1782.

Corrales, D. C., Espino, A. I. L. & Corrales, J. C. (2016), 'A systematic review of data quality issues in knowledge discovery tasks', *Revista Ingenierías Universidad de Medellín* **15**(28), 125–150. DOI: https://doi.org/10.22395/rium.v15n28a7.

Cox, B. E., Mcintosh, K., Reason, R. & Terenzini, P. (2014), 'Working with missing data in higher education research: A primer in real-world example', *The Review of Higher Education* **37**(3), 377–402. DOI: https://doi.org/10.1353/rhe.2014.0026.

Cox III, E. P. (1980), 'The optimal number of response alternatives for a scale: A review', *Journal of Marketing Research* **17**(4), 407–422. DOI: https://doi.org/10.2307/3150495.

Cruz-Sandoval, D., Beltran-Marquez, J., Garcia-Constantino, M., Gonzalez-Jasso, L., Favela, J., López-Nava, I., Cleland, I., Ennis, A., Hernández-Cruz, N., Rafferty, J., Synnott, J. & Nugent, C. (2019), 'Semi-automated data labeling for activity recognition in pervasive healthcare', *Sensors* **19**(4), 3035. DOI: https://doi.org/10.3390/s19143035.

DeCastro-García, N., Castañeda, A. L. M., Rodríguez, M. F. & Carriegos, M. V. (2018), 'On detecting and removing superficial redundancy in vector databases', *Mathematical Problems in Engineering* **2018**(3702808). DOI: https://doi.org/10.1155/2018/3702808.

Denscombe, M. (1998), *The Good Research Guide: For Small-Scale Social Research Projects*, 2 edn, Open University Press, Buckingham, England, United Kingdom.

Earley, S. & Henderson, D. (2017), *DAMA-DMBOK: Data management body of knowledge*, 2nd edn, Technics Publications, LLC, Denville, NJ, United States.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. & Dean, J. (2019), 'A guide to deep learning in healthcare', *Nature Medicine* **25**. DOI: https://doi.org/10.1038/s41591-018-0316-z.

Eur (2020), European Statistical System handbook for quality and metadata reports, Standard, Eurostat, Brussels, Belgium. Available: https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf.

Evans, J. R. & Mathur, A. (2005), 'The value of online surveys', *Internet Research* **15**(2), 195–219. DOI: https://doi.org/10.1108/10662240510590360.

Farooq, M. B. & de Villiers, C. (2017), 'Telephonic qualitative research interviews: When to consider them and how to do them', *Meditari Accountancy Research* **25**(2), 291 – 316. DOI: https://doi.org/10.1108/MEDAR-10-2016-0083.

Fletcher, F. (1998), A framework for addressing data quality in distributed computing systems, *in* 'Proceedings of the 1998 International Conference on Information Quality', MIT, Cambridge, MA, USA. Available: http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%201998/Papers/AFrameworkForAddressDQinDistributedCompu.pdf.

Fox, C., Levitin, A. & Redman, T. (1994), 'The notion of data and its quality dimensions', *Information Processing Management* **30**(1), 9–19. DOI: https://doi.org/10.1016/0306-4573(94)90020-5.

Freitas, H., Oliveira, M., Jenkins, M. & Popjoy, O. (2021), 'The focus group a qualitative research method'. Available: http://gianti.ea.ufrgs.br/files/artigos/1998/1998_079_ISRC.pdf.

Gacenga, F., Cater-Steel, A., Toleman, M. & Tan, W.-G. (2012), 'A proposal and evaluation of a design method in design science research', *The Electronic Journal of Business Research Methods* **10**, 89–100. Available: http://www.ejbrm.com/issue/download.html?idArticle=281.

Gao, J., Xie, C. & Tao, C. (2016), Big data validation and quality assurance – issuses, challenges, and needs, *in* '2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)', Institute of Electrical and Electronics Engineers, New Jersey, United States, pp. 433–441. DOI: https://doi.org/10.1109/SOSE.2016.63.

Gawankar, S., Kamble, S. S. & Raut, R. (2015), 'Performance measurement using balance score card and its applications: A review', *Journal of Supply Chain Management Systems* **4**(3). DOI: https://doi.org/10.21863/jscms/2015.4.3.009.

Gentleman, J. F. & Wilk, M. B. (1975), 'Detecting outliers. ii. supplementing the direct analysis of residuals', *Biometrics* **31**(2), 387–410. DOI: https://doi.org/10.2307/2529428.

Gibbs, G. R. (2007), *Analyzing Qualitative Data*, SAGE Publications, California, United States. DOI: https://doi.org/10.4135/9781849208574.

Ginty, A. T. (2013), *Construct Validity*, Springer New York, New York, NY, pp. 487–487. DOI: https://doi.org/10.1007/978-1-4419-1005-9_861.

Guda, R., Mohanraj, V., Kameshwar Rao, J. V. & Chandan Kumar, N. A. (2018), *ADAS$^{DL}$—Innovative Approach for ADAS Application Using Deep Learning: Pro-*

*ceedings of Fifth International Conference INDIA 2018 Volume 1*, Vol. 862, pp. 497–504. DOI: https://doi.org/10.1007/978-981-13-3329-3_46.

Gupta, S. & Gupta, A. (2019), 'Dealing with noise problem in machine learning data-sets: A systematic review', *Procedia Computer Science* **161**, 466–474. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia. DOI: https://doi.org/10.1016/j.procs.2019.11.146.

Heinrich, B., Hristova, D., Klier, M., Schiller, A. & Szubartowicz, M. (2018), 'Requirements for data quality metrics', *Journal of Data and Information Quality* **9**(2), 1–32. DOI: https://doi.org/10.1145/3148238.

Heravizadeh, M., Mendling, J. & Rosemann, M. (2008), Dimensions of business processes quality (qobp), Vol. 17, pp. 80–91. DOI: https://doi.org/10.1007/978-3-642-00328-8_8.

Houari, R., Bounceur, A., Tari, A. K. & Kecha, M. T. (2016), Handling missing data problems with sampling methods, *in* 'Proceedings - 2014 International Conference on Advanced Networking Distributed Systems and Applications, INDS 2014', Institute of Electrical and Electronics Engineers, New Jersey, United States, pp. 99–104. DOI: https://doi.org/10.1109/INDS.2014.25.

Hu, H., Wen, Y., Chua, T.-S. & Li, X. (2014), 'Toward scalable systems for big data analytics: A technology tutorial', *IEEE Access* **2**, 652–687. DOI: https://doi.org/10.1109/ACCESS.2014.2332453.

Huisman, M. (2000), 'Imputation of missing item responses: Some simple techniques', *Quality and Quantity* **34**, 331–351. DOI: https://doi.org/10.1023/A:1004782230065.

Illing, J. (2013), *Thinking About Research: Theoretical Perspectives, Ethics and Scholarship*, pp. 329–347. DOI: https://doi.org/10.1002/9781118472361.ch24.

International Data Corporation (2020), 'Idc's global storagesphere forecast shows continued strong growth in the world's installed base of storage capacity'. Available: https://www.idc.com/getdoc.jsp?containerId=prUS46303920.

Isaac, S. & Michael, W. B. (1997), *Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences*, 3 edn, Educational and Industrial Testing Services, San Diego, California, United States.

ISO (2019), Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model, Standard, International Organization for Standardization, Geneva, CH. Available: https://www.iso.org/standard/35736.html.

Jakobsen, J. C., Gluud, C., Wetterslev, J. & Winkel, P. (2017), 'When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts', *BMC Medical Research Methodology* **17**(162). DOI: https://doi.org/10.1186/s12874-017-0442-1.

Juddoo, S. (2015), Overview of data quality challenges in the context of big data, *in* '2015 International Conference on Computing, Communication and Security (ICCCS)', Institute of Electrical and Electronics Engineers, New Jersey, United States, pp. 1–9. DOI: https://doi.org/10.1109/CCCS.2015.7374131.

Kahn, M. G., Brown, J. S., Chun, A. T., Davidson, B. N., Meeker, D., Ryan, P. B., Schilling, L. M., Weiskopf, N. G., Williams, A. E. & Zozus, M. N. (2015), 'Transparent reporting of data quality in distributed data networks', *eGEMs (Generating Evidence Methods to improve patient outcomes)* **3**(1). DOI: https://doi.org/10.13063/2327-9214.1052.

Kang, H. (2013), 'The prevention and handling of the missing data', *Korean Journal of Anesthesiology* **64**(5), 402–406. DOI: https://doi.org/10.4097/kjae.2013.64.5.402.

Kelley, K., Clark, B., Brown, V. & Sitzia, J. (2003), 'Good practice in the conduct and reporting of survey research', *International Journal for Quality in Health Care* **15**(3), 261–266. DOI: https://doi.org/10.1093/intqhc/mzg031.

Knauss, E. (2020), 'Constructive master's thesis work in industry: Guidelines for applying design science research'. Available: https://arxiv.org/abs/2012.04966.

Knight, S.-A. & Burn, J. (2005), 'Developing a framework for assessing information quality on the world wide web', *Informing Science Journal* **8**. DOI: https://doi.org/10.28945/493.

Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. (2005), 'Handling imbalanced datasets: A review', *GESTS International Transactions on Computer Science and Engineering* **30**, 25–36. Available: https://www.researchgate.net/publication/228084509_Handling_imbalanced_datasets_A_review.

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**(4). DOI: https://doi.org/10.2196/medinform.5359.

Kukkala, V. K., Pasricha, S., Tunnell, J. & Bradley, T. H. (2018), 'Advanced driver-assistance systems: A path toward autonomous vehicles', *IEEE Consumer Electronics Magazine* **7**(5), 18–25. DOI: https://doi.org/10.1109/MCE.2018.2828440.

Kwak, S. K. & Kim, J. H. (2017), 'Statistical data preparation: Management of missing values and outliers', *Korean Journal of Anesthesiology* **70**(4), 407–411. DOI: https://doi.org/10.4097/kjae.2017.70.4.407.

Labra, O., Castro, C., Wright, R. & Chamblas, I. (2019), *Thematic Analysis in Social Work: A Case Study*, IntechOpen, London, United Kingdom, pp. 1–20. DOI: https://doi.org/10.5772/intechopen.89464.

Laudon, K. C. & Laudon, J. P. (2009), *Essentials of Management Information Systems*, 8 edn, Pearson Prentice Hall, New York, United States.

Liew, A. (2007), 'Understanding data, information, knowledge and their inter-relationships', *Journal of Knowledge Management Practice* **7**. Available: https://www.researchgate.net/publication/224937037_Understanding_Data_Information_Knowledge_And_Their_Inter-Relationships.

Linneberg, M. S. & Korsgaard, S. (2019), 'Coding qualitative data: a synthesis guiding the novice', *Qualitative Research Journal* **19**(3), 259–270. DOI: https://doi.org/10.1108/QRJ-12-2018-0012.

Majid, M. A. A., Othman, M., Mohamad, S. F., Lim, S. A. H. & Yusof, A. (2017), 'Piloting for interviews in qualitative research: Operationalization and lessons learnt', *International Journal of Academic Research in Business and Social Sciences* **7**, 1073–1080. DOI: https://doi.org/10.6007/IJARBSS/v7-i4/2916.

McGilvray, D. (2008), *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

McGrath, C., Palmgren, P. J. & Liljedahl, M. (2018), 'Twelve tips for conducting qualitative research interviews', *Medical Teacher* **41**(9), 1002–1006. DOI: https://doi.org/10.1080/0142159X.2018.1497149.

Memon, M. A., Soomro, S., Jumani, A. K. & Kartio, M. A. (2017), 'Big data analytics and its applications', *Annals of Emerging Technologies in Computing (AETiC)* **1**(1), 45–54. DOI: https://doi.org/10.33166/AETiC.2017.01.006.

Michener, W. K. (2015), 'Ten simple rules for creating a good data management plan', *PLoS computational biology* **11**(10), e1004525. DOI: https://doi.org/10.1371/journal.pcbi.1004525.

Muniasamy, A., Tabassam, S., Hussain, M. A., Sultana, H., Muniasamy, V. & Bhatnagar, R. (2019), Deep learning for predictive analytics in healthcare, Vol. 921. DOI: https://doi.org/10.1007/978-3-030-14118-9_4.

Namatevs, I., Sudars, K. & Polaka, I. (2019), 'Automatic data labeling by neural networks for the counting of objects in videos', *Procedia Computer Science* **149**, 151–158. ICTE in Transportation and Logistics 2018 (ICTE 2018). DOI: https://doi.org/10.1016/j.procs.2019.01.118.

Noumir, Z., Honeine, P. & Richard, C. (2012), On simple one-class classification methods, pp. 2022–2026. DOI: https://doi.org/10.1109/ISIT.2012.6283685.

Nowell, L. S., Norris, J. M., White, D. E. & Moules, N. J. (2017), 'Thematic analysis: Striving to meet the trustworthiness criteria', *International Journal of Qualitative Methods* **16**(1). DOI: https://doi.org/10.1177/1609406917733847.

Osborne, J. W. & Overbay, A. (2004), 'The power of outliers (and why researchers should always check for them)', *Practical Assessment, Research, and Evaluation (PARE)* **9**(6). Available: https://core.ac.uk/download/pdf/239583894.pdf.

Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V. & Bragge, J. (2006), 'The design science research process: A model for producing and present-

ing information systems research', *Proceedings of First International Conference on Design Science Research in Information Systems and Technology DESRIST* . Available: https://arxiv.org/abs/2006.02763.

Peralta, V. (2006), Data quality evaluation in data integration systems. Available: https://www.researchgate.net/publication/30515880_Data_Quality_Evaluation_in_Data_Integration_Systems.

Pew Research Center (n.d.), 'Questionnaire design'. Available: https://www.pewresearch.org/methods/u-s-survey-research/questionnaire-design/.

Pipino, L., Lee, Y. & Wang, R. (2003), 'Data quality assessment', *Communications of the ACM* **45**(4), 211–218. DOI: https://doi.org/10.1145/505248.506010.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Presa Reyes, M., Shyu, M.-L., Chen, S.-C. & Iyengar, S. (2018), 'A survey on deep learning: Algorithms, techniques, and applications', *ACM Computing Surveys* **51**, 1–36. DOI: https://doi.org/10.1145/3234150.

Prasad, H. K., Faruquie, T. A., Joshi, S., Chaturvedi, S., Subramaniam, L. & Mohania, M. (2011), Data cleansing techniques for large enterprise datasets, *in* '2011 Annual SRII Global Conference', Vol. 1, Institute of Electrical and Electronics Engineers, New Jersey, United States, pp. 135–144. DOI: https://doi.org/10.1109/SRII.2011.26.

Rai, A. (2017), 'Editor's comments: avoiding type iii errors: formulating is research problems that matter', *Management Information Systems Quarterly* **41**, 3–7. Available: https://misq.org/misq/downloads/download/editorial/659/.

Raj, A., Bosch, J., Olsson, H., Arpteg, A. & Brinne, B. (2019), 'Data management challenges for deep learning'. DOI: https://doi.org/10.13140/RG.2.2.16447.25764.

Rezvan, P. H., Lee, K. J. & Simpson, J. A. (2015), 'The rise of multiple imputation: a review of the reporting and implementation of the method in medical research', *BMC Medical Research Methodology* **15**(30), 30. DOI: https://doi.org/10.1186/s12874-015-0022-1.

Rogério, R. & Hirama, K. (2015), 'Characterizing big data management', *Issues in Informing Science and Information Technology* **12**, 165–180. Available: http://iisit.org/Vol12/IISITv12p165-180Rossi1921.pdf.

Roh, Y., Heo, G. & Whang, S. (2019), 'A survey on data collection for machine learning: A big data - ai integration perspective', *IEEE Transactions on Knowledge and Data Engineering* **33**(4), 1–1. DOI: https://doi.org/10.1109/TKDE.2019.2946162.

Rousseeuw, P. J. & Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, John Wiley  Sons, Inc., Hoboken, New Jersey, United States. DOI: https://doi.org/10.1002/0471725382.

Seale, C. (2017), *Researching Society and Culture*, SAGE Publications, California, United States.

Sessions, V. & Valtorta, M. (2006), The effects of data quality on machine learning algorithms, *in* 'Proceedings of the 11th International Conference on Information Quality', MIT, Cambridge, MA, USA, pp. 485–498. Available: http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202006/papers/The%20Effects%20of%20Data%20Quality%20on%20Machine%20Learning%20Algorithms.pdf.

Shrestha, A. & Mahmood, A. (2019), 'Review of deep learning algorithms and architectures', *IEEE Access* **7**, 53040–53065. DOI: https://doi.org/10.1109/ACCESS.2019.2912200.

Sidi, F., Panahy, P. H. S., 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. DOI: https://doi.org/10.1109/InfRKM.2012.6204995.

Sun, C. & Wang, J. (n.d.), 'Study on the data quality management and the data quality control – a case study of the earth system science data sharing project', *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **38**(2). Available: https://www.isprs.org/proceedings/xxxviii/part2/papers/97_paper.pdf.

Tavakoli, A. S., Jackson, K., Moneyham, L., Phillips, K. D., Murdaugh, C. & Meding, G. (2006), 'Data management plans: Stages, components, and activities', *Applications and Applied Mathematics (AAM): An International Journal* **1**(2), 141–151. Available: https://www.pvamu.edu/aam/wp-content/uploads/sites/49/05_tavakoli_r9-052206-vol.-1_issue_2_12-30-2011.pdf.

The Open Measured Data Management Working Group (2021), 'Open-measured-data-management-wg'.

Vaishnavi, V. & Kuechler, B. (2004), 'Design science research in information systems', *Association for Information Systems* . Available: https://www.researchgate.net/publication/235720414_Design_Science_Research_in_Information_Systems.

Vaismoradi, M., Jones, J., Turunen, H. & Snelgrove, S. (2016), 'Theme development in qualitative content analysis and thematic analysis', *Journal of Nursing Education and Practice* **6**(5), 100–110. DOI: https://doi.org/10.5430/jnep.v6n5p100.

vom Brocke, J. & Maedche, A. (2019), 'The dsr grid: six core dimensions for effectively planning and communicating design science research projects', *Electron Markets* **29**, 379–385. DOI: https://doi.org/10.1007/s12525-019-00358-7.

Walters, T. (2016), 'Using thematic analysis in tourism research', *Tourism Analysis* **21**(1), 107–116. DOI: https://doi.org/10.3727/108354216X14537459509017.

Wang, R. Y. & Strong, D. M. (1996), 'Beyond accuracy: What data quality means

to data consumers', *Journal of Management Information Systems* **12**(4), 5–33. Available: https://www.jstor.org/stable/40398176.

Xu, W. & Zammit, K. (2020), 'Applying thematic analysis to education: A hybrid approach to interpreting data in practitioner research', *International Journal of Qualitative Methods* **19**. DOI: https://doi.org/10.1177/1609406920918810.

Zizlavsky, O. (2013), 'The balanced scorecard: Innovative performance measurement and management control system', *Journal of Technology Management and Innovation* **9**(3), 210–222. DOI: https://doi.org/10.4067/S0718-27242014000300016.

Ziębiński, A., Cupek, R., Grzechca, D. & Chruszczyk, L. (2017), Review of advanced driver assistance systems (adas), Vol. 1906, p. 120002. DOI: https://doi.org/10.1063/1.5012394.

# A

# Appendix

## A.1 Interview

### A.1.1 Interview Standardized Consent Form

This section presents standardized consent form presented to the interviews prior to the interviews in the first and second iteration. Table A.1 presents the fields of the form, description of those fields, and sample data that could be filled in those fields.

**Table A.1:** Interview Standardized Consent Form Template

| Rank | Challenge Set | Challenge |
|---|---|---|
| Interview number | The identifier for the order of the interview with a particular interviewee. | *1-1* |
| Interviewee | Name of the interviewee | *John Doe* |
| Position | Current position held by the interviewee | *Product Owner* |
| Team | The team that the interviewee belongs to in his/her organization | *Research* |
| Company | Name of the company in which the interviewee is employed at | *Veoneer Sweden* |
| Academic background | Academic degree held by the interviewee and the name of his/her field | *PhD in Electrical Engineering* |
| Experience | Number of years of work experience the interviewee has in his/her field | *15 years* |
| Date | Date on which the interview is conducted | *March 17, 2021* |
| Length (in minutes) | Number of minutes the interview was conducted | *80 minutes* |
| Consent for audio recording | Boolean value denoting if consent was given by the interviewee for audio recording of the interview session | *Yes* |
| Consent for publication | Boolean value denoting if consent was given by the interviewee for publication of data from the interview | *Yes* |

### A.1.2 Interview Questions

This section presents the questions of the interviews conducted in both the first and second iterations of this study. There are three versions of the Iteration 1 question set. Although there are no vast difference between the questions in these versions, there are slight modifications based on the feedback received during the interviews.

#### A.1.2.1 Iteration 1 Interview Question Set - Version 1

**General Questions**

1. What is your role in the company?
2. What does your team focus on?
3. What kind of system are you working on in your team?

**Data Questions**

4. What are the constituent components of the system that you are working on?
5. How are the volume, variety, and velocity of the data in the system you work on and how does it impact the quality of data?
6. What are the key data gathered in your field?
7. How are the data collected used in your field?
8. How does the data affect the behavior of the system you are working on?
9. What do you think are appropriate metrics for each of the data you mentioned in Q.7?
10. Why are the metrics you mentioned earlier are appropriate for the data?

**Data Quality Procedure Questions**

11. In your current context, how would you make data quality measurable?
12. What's the current procedure Veoneer follows to ensure data quality?
    (a) If yes, what are the drawbacks of the current procedure if there is any to ensure data quality?
13. (If they mentioned documentation in Q.12) How can documentation procedure be improved?
14. (If they do not mentioned documentation in Q.12) How are data quality requirements documented in Veoneer?

15. What could be the potential ways in which data quality assessment procedure can be improved?

16. What are some of the challenges that you have faced or you think exist in handling data and assessing the quality of it?

**Data and Safety Questions**

17. What is the relationship between safety and the data collected?

**Wrap-up Questions**

18. To summarize, what do you think is the most challenging aspect of handling data?

19. In the future, how do you envision the data quality requirements are dealt with in Veoneer?

20. Do you think we forgot to ask you about something that we should know about?

#### A.1.2.2 Iteration 1 Interview Question Set - Version 2

**General Questions**

1. What is your role in the company?

2. What does your team focus on?

3. What kind of system are you working on in your team? What are the goals of the system?

4. What are the constituent components of the system you are working?

**Data Questions**

5. Can you tell me what you understand by 'data' in your system? What is typical data in your field?

6. Among these data that you mentioned, what is the most important data in the field?

7. How would you set requirements on data?

8. Who is responsible for setting the data requirements?

9. In your opinion, how would you evaluate the importance of each person or group responsible to set the data requirements?

10. How would you characterize good data quality in your system? Can you give

examples of it?

11. What would data with subpar quality look like in your system?
12. What are the typical metrics you apply to the data? Why are these the appropriate metrics for the data?
13. How does the data affect the behavior of the system you are working on?
14. In your current context, how do you make data quality measurable?
15. Do you have a procedure that you follow in your team to ensure data quality?
    (a) If yes, can you describe the procedure?
        i. How are data quality requirements elicited and set in your team?
        ii. How do you segregate good and bad data?
    (b) If yes, what are the drawbacks and challenges of the current procedure?
    (c) If yes, how do you document the data quality requirements?
    (d) If yes, what could be the potential ways in which data quality assessment procedure can be improved?
    (e) If not, how would you make data quality measurable? Can you describe an ideal procedure to assess data quality in your system?

**Data and Safety Questions**

16. What is the relationship between the goal of your system and the data collected?

**Wrap-up Questions**

17. In the future, how do you envision the data quality requirements are dealt with in your team?
18. Do you think we forgot to ask you about something that we should know about?

#### A.1.2.3 Iteration 1 Interview Question Set - Version 3

**General Questions**

1. What is your role in the company?
2. What does your team focus on?

3. What kind of system are you working on in your team? What are the goals of the system?

4. What are the constituent components of the system you are working?

**Data Questions**

5. Can you tell me what you understand by 'data' in your system? What is typical data in your field?

6. Among these data that you mentioned, what is the most important data in the field?

7. How are data quality requirements elicited and set in your team?

8. Who is responsible for setting the data requirements?

9. In your opinion, how would you evaluate the importance of each person or group responsible to set the data requirements?

10. How would you characterize good data quality in your system? Can you give examples of it?

11. What would data with subpar quality look like in your system?

12. What are the typical metrics you apply to the data? Why are these the appropriate metrics for the data?

13. How does the data affect the behavior of the system you are working on?

14. In your current context, how do you make data quality measurable?

15. Do you have a procedure that you follow in your team to ensure data quality?

    (a) If yes, can you describe the procedure?

    (b) If yes, what are the drawbacks and challenges of the current procedure?

    (c) If yes, how do you document the data quality requirements?

    (d) If yes, what could be the potential ways in which data quality assessment procedure can be improved?

    (e) If not, can you describe an ideal procedure to assess data quality in your system?

16. How do you segregate good and bad data?

**Data and Safety Questions**

17. What is the relationship between the goal of your system and the data collected?

**Wrap-up Questions**

18. In the future, how do you envision the data quality requirements are dealt with in your team?

19. Do you think we forgot to ask you about something that we should know about?

The questions asked in Iteration 2 interviews were more open-ended. The questions pertain to the potential solutions of the data quality challenges. The following is a general framing of the questions for individual potential solutions.

#### A.1.2.4 Iteration 2 Interview Question Set

**Potential Solutions Questions**

1. What is the information needed to determine the solution?

2. When should you decide to determine the information for this solution?

3. Is the solution useful for first-party or third-party data?

4. How to use this solution? What is the process to identify and solve this challenge?

5. What constraints do you see in practical implementation of this solution, if any?

6. Is this solution currently being used in your organization?

7. Does this solution actually help solve the challenge or not?

8. What is your opinion regarding the flow of the solution?

9. Is this just a theoretical solution or a practically applicable one as well?

10. If not, what can be other potential solutions for the challenge?

11. In your opinion, what kind of terms between the seller and the buyer help solve this challenge?

12. What kind of penalty and tolerance are applicable for this challenge?

## A.2 Initial Challenges

Figure A.1 depicts a collection of challenges identified through interviews and literature review in the first iteration. They are divided into *System Challenges* and *Data Challenges*.

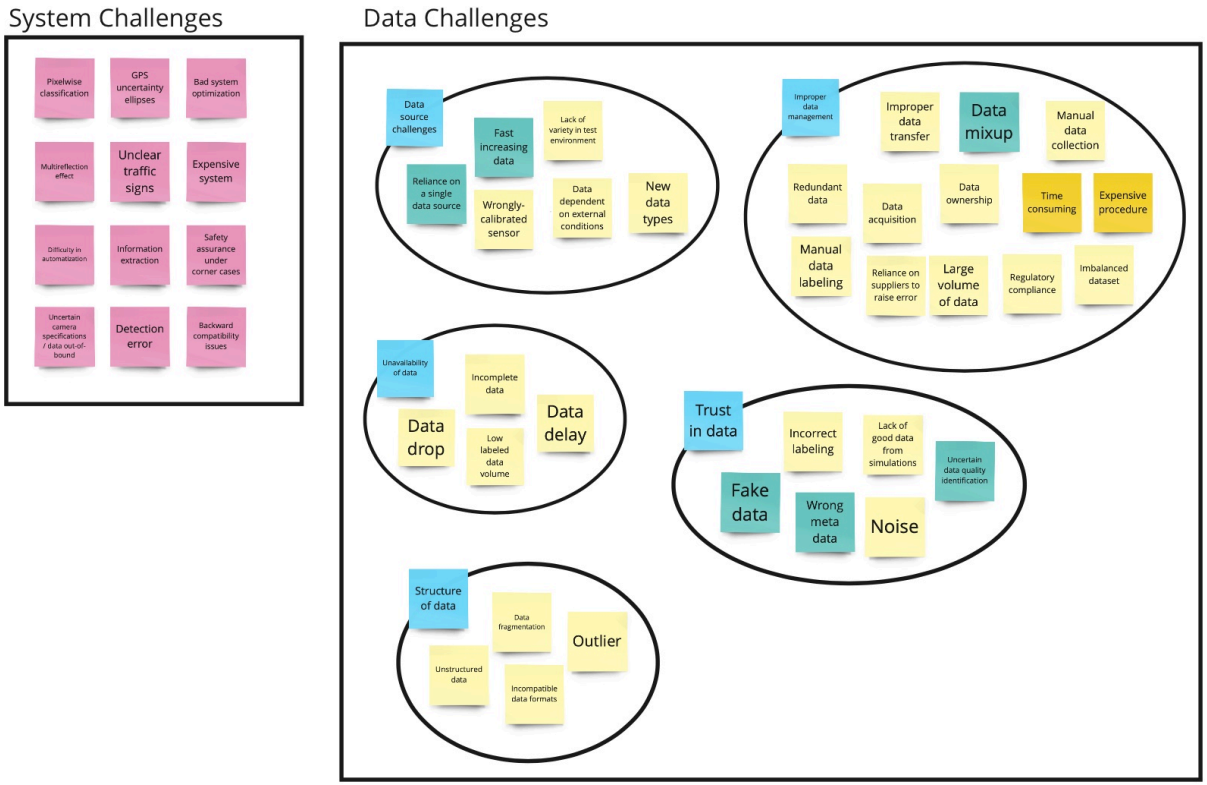



**Figure A.1:** Initial Challenges

## A.3 Survey Questionnaires

Note: The options that the survey participants can select from are provided inside square brackets ([ ]) and are separated by slash (/). For example, *[Yes/No]* means a survey participant can provide either a "Yes" or a "No" response.

### A.3.1 Survey 1 Questionnaire

Survey 1 Questionnaire was sent during the first iteration of this study. Below is the set of questions presented in the questionnaire form.

Note: Mandatory questions are denoted with an asterisk (*).

1. Please rank the following 'Data source' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*
    - Wrongly-calibrated / defective sensor generates incorrect data
    - Reliance on a single data source (e.g. only depending on a single radar to collect data, or depending on a single type of sensor)
    - Fast increasing data overwhelms the ML algorithms that need to process the data
    - Data dependent on external conditions can be of low quality sometimes
    - New data types from various sources make data integration difficult
    - Lack of variety in test environment causes AI to be poorly trained for situations that it has not been trained for
    - Noise (unwanted data that is mixed with valuable data)
2. Please rank the following 'Data Availability' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*
    - Low labeled data volume (i.e. the amount of data that is labeled is lesser than the the amount of data that is unlabeled)
    - Data delay (i.e. there is a delay in transmission of data from the source to the destination, e.g. data store to function, sensor to sensor)
    - Data drop (e.g. dropping of a required attribute of data, dropping a chunk of data during transmission)
    - Incomplete data (e.g. missing attribute of data, missing chunk of data)

3. Please rank the following 'Data management' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Improper data transfer (e.g. mismanagement of the way data is transferred between source and destination)
    - Redundant data
    - Large volume of data makes it difficult to assess the quality of the data
    - Reliance on suppliers to raise error (i.e. there is no internal process to raise an error, and raising error is dependent on the suppliers doing so)
    - Data ownership (i.e. who is the legal owner of the data and can use it without any prior approval?)
    - Data acquisition (i.e. purchase of data from third-party and processes associated with it)
    - Imbalanced dataset (e.g. data from a single location or single weather type dominates other locations or weathers)
    - Regulatory compliance (i.e. what are the rules and regulations for handling data and assuring data quality)
    - Manual data collection
    - Manual data labeling

4. Please rank the following 'Trust in data' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Fake data (e.g. data is maliciously manipulated)
    - Incorrect labeling
    - Wrong metadata
    - Lack of good data from simulations (i.e. data from simulations are relatively of lesser quality than data from real-world collection)
    - Uncertain data quality identification

5. Please rank the following 'Data structure' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*

    - Data fragmentation (i.e. data required by a function is located in different

places)

- Outlier data (i.e. data is out of bounds of acceptable range)
- Incompatible data formats
- Unstructured data
- Data mixup (e.g. rows and columns are mixed together)

6. Please rate these challenges (1 being the least pressing and 6 being the most pressing)*
   - 'Data source' challenges (related to question 1) [1/2/3/4/5/6]
   - 'Availability of data' challenges (related to question 2) [1/2/3/4/5/6]
   - 'Data management' challenges (related to question 3) [1/2/3/4/5/6]
   - 'Trust in data' challenges (related to question 4) [1/2/3/4/5/6]
   - 'Data structure' challenges (related to question 5) [1/2/3/4/5/6]

### A.3.2 Survey 2 Questionnaire

Survey 1 Questionnaire was sent during the third iteration of this study.Below is the set of questions presented in the questionnaire form.

Note: Mandatory questions are denoted with an asterisk (*).

**Artifact**

In this section, we are trying to validate the structure of the artifacts.

1. The following is the template for 'List of challenges' artifact. This artifact attempts to present each of the data quality challenges identified using different methods.*

   Name - The name of the challenge (e.g., Data Delay)
   Sources - How did we identify this challenge? (e.g., interviews, literature review)
   Description - Description of the challenge (e.g., what is it? why is it a challenge?)
   Whether the challenge directly affects AI function - Boolean value stating whether this challenge directly affects AI functions, models
   Challenge score - Ranking of the challenge in terms of 'pressing'-ness (calculated using a formula developed during the thesis work)

In your opinion, are these appropriate fields for the template for 'List of challenges' artifact?

- Name [Yes/No]
- Sources [Yes/No]
- Description [Yes/No]
- Whether the challenge directly affects AI function [Yes/No]
- Challenge score [Yes/No]

2. For the above template for 'List of challenges' artifact, what other field/s do you think can be added so that the artifact provides more information than it does right now?

3. The following is the template for 'List of Data quality attributes' artifact. This artifact attempts to present a list of data quality attributes and whether they are affected by data quality challenges.*

   Name - The name of the data quality attribute (e.g., Accuracy)
   Sources - How did we identify this data quality attribute? (e.g., interviews, literature review)
   Definition - Definition of the data quality attribute (e.g., what is it?)
   Which challenges affect the data quality attribute? - A list of challenges that affect the data quality attribute

   In your opinion, are these appropriate fields for the template for 'List of Data quality attributes' artifact?

   - Name [Yes/No]
   - Sources [Yes/No]
   - Definition [Yes/No]
   - Which challenges affect the data quality attribute? [Yes/No]

4. For the above template for 'List of Data quality attributes' artifact, what other field/s do you think can be added so that the artifact provides more information than it does right now?

5. The following is the template for 'List of Data quality attribute metrics' artifact. This artifact attempts to present a list of data quality attribute metrics.

   Data quality attribute - The name of the data
   quality attribute (e.g., Accuracy)

Metric - The name of the data quality attribute metric (e.g., Degree of accuracy)
Formula - Formula of how the metric is calculated

In your opinion, are these appropriate fields for the template for 'List of Data quality attribute metrics' artifact?*

- Data quality attribute [Yes/No]
- Metric [Yes/No]
- Formula [Yes/No]

6. For the above template for 'List of Data quality attribute metrics' artifact, what other field/s do you think can be added so that the artifact provides more information than it does right now?

7. The following is the template for 'Potential solutions' artifact. This artifact attempts to present and explain a list of potential solutions for data quality challenges.

   Name - The name of the solution (e.g., Continuous data processing)
   Requirements specifications - The information that should specified before the implementation fo the solution (e.g., Determine an acceptable range of time for data arrival)
   Implementation details - How the solution is implemented?

   In your opinion, are these appropriate fields for the template for 'Potential solutions' artifact?*

   - Name [Yes/No]
   - Requirements specifications [Yes/No]
   - Implementation details [Yes/No]

8. For the above template for 'Potential solutions' artifact, what other field/s do you think can be added so that the artifact provides more information than it does right now?

**Challenges Affecting AI models**

In this section, we are trying to validate the structure of the artifacts.

9. Does the following challenges affect AI models directly or not?*

   - Data delay (i.e. there is a delay in transmission of data from the source to

the destination, e.g. data store to function, sensor to sensor) [Yes/No]

- Data drop (e.g. dropping of a required attribute of data, dropping a chunk of data during transmission) [Yes/No]
- Incomplete data (e.g. missing attribute of data, missing chunk of data) [Yes/No]
- Low labeled data volume (i.e. the amount of data that is labeled is lesser than the the amount of data that is unlabeled) [Yes/No]
- Data acquisition (i.e. purchase of data from third-party and processes associated with it) [Yes/No]
- Improper data transfer (e.g. mismanagement of the way data is transferred between source and destination) [Yes/No]
- Imbalanced dataset (e.g. data from a single location or single weather type dominates other locations or weathers) [Yes/No]
- Redundant data [Yes/No]
- Manual data collection [Yes/No]
- Manual data labeling [Yes/No]
- Expensive procedure [Yes/No]
- Reliance on suppliers to raise error (i.e. there is no internal process to raise an error, and raising error is dependent on the suppliers doing so) [Yes/No]
- Large volume of data makes it difficult to assess the quality of the data [Yes/No]
- Time consuming [Yes/No]
- Data ownership (i.e. who is the legal owner of the data and can use it without any prior approval?) [Yes/No]
- Wrongly-calibrated / defective sensor generates incorrect data [Yes/No]
- New data types from various sources make data integration difficult [Yes/No]
- Lack of variety in test environment causes AI to be poorly trained for situations that it has not been trained for [Yes/No]
- Data dependent on external conditions can be of low quality sometimes [Yes/No]

- Data fragmentation (i.e. data required by a function is located in different places) [Yes/No]
- Incompatible data formats [Yes/No]
- Outlier data (i.e. data is out of bounds of acceptable range) [Yes/No]
- Unstructured data [Yes/No]
- Noise (unwanted data that is mixed with valuable data) [Yes/No]
- Lack of good data from simulations (i.e. data from simulations are relatively of lesser quality than data from real-world collection) [Yes/No]
- Incorrect labeling [Yes/No]

**Data Quality Challenges**

The questions in this section attempt to understand the 'pressing'-ness of data quality challenges.

10. Please rank the following 'Data Availability' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*
    - Data delay (i.e. there is a delay in transmission of data from the source to the destination, e.g. data store to function, sensor to sensor)
    - Data drop (e.g. dropping of a required attribute of data, dropping a chunk of data during transmission)
    - Incomplete data (e.g. missing attribute of data, missing chunk of data)
    - Low labeled data volume (i.e. the amount of data that is labeled is lesser than the the amount of data that is unlabeled)
11. Please rank the following 'Data management' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*
    - Data acquisition (i.e. purchase of data from third-party and processes associated with it)
    - Improper data transfer (e.g. mismanagement of the way data is transferred between source and destination)
    - Imbalanced dataset (e.g. data from a single location or single weather type dominates other locations or weathers)
    - Redundant data

- Manual data collection and labeling
- Expensive procedure
- Reliance on suppliers to raise error (i.e. there is no internal process to raise an error, and raising error is dependent on the suppliers doing so)
- Large volume of data makes it difficult to assess the quality of the data
- Time consuming
- Data ownership (i.e. who is the legal owner of the data and can use it without any prior approval?)

12. Please rank the following 'Data source' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)
    - Wrongly-calibrated / defective sensor generates incorrect data
    - New data types from various sources make data integration difficult
    - Lack of variety in test environment causes AI to be poorly trained for situations that it has not been trained for
    - Data dependent on external conditions can be of low quality sometimes
13. Please rank the following 'Data structure' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*
    - Data fragmentation (i.e. data required by a function is located in different places)
    - Incompatible data formats
    - Outlier data (i.e. data is out of bounds of acceptable range)
    - Unstructured data
14. Please rank the following 'Data Trust' challenges from the most pressing challenge to the least pressing challenge. (Most pressing challenge should be at the top of the list and the least pressing one should be at the bottom)*
    - Noise (unwanted data that is mixed with valuable data)
    - Lack of good data from simulations (i.e. data from simulations are relatively of lesser quality than data from real-world collection)
    - Incorrect labeling

15. Please rate these challenge sets (1 being the least pressing and 6 being the most pressing)

    - 'Data Availability' challenges (related to question 1) [1/2/3/4/5/6]
    - 'Data Management' challenges (related to question 2) [1/2/3/4/5/6]
    - 'Data Source' challenges (related to question 3)) [1/2/3/4/5/6]
    - 'Data Structure' challenges (related to question 4) [1/2/3/4/5/6]
    - 'Data Trust' challenges (related to question 5) [1/2/3/4/5/6]

Table A.2 provides a list of data quality attributes that were presented in Survey 2 questionnaire for validation of association between data quality challenges and attributes. It also provides definitions provided in the survey questionnaire for respective data quality attributes.

**Table A.2:** List of Data Quality Attributes and Their Definitions Provided in the Survey 2 Questionnaire

| Data Quality Attribute | Definition of the Attribute Provided in the Survey Questionnaire |
|---|---|
| Accessibility | The extent to which data are available or easily and quickly retrievable |
| Access security | The extent to which access to data can be restricted and hence kept secure |
| Accuracy | The extent to which data are correct, reliable, and certified free of error |
| Auditability | It means that auditors can fairly evaluate data accuracy and integrity within rational time and manpower limits during the data use phase |
| Availability | The degree to which data can be consulted or retrieved by data consumers or processes |
| Completeness | Refers to whether all required data is present |
| Compliance | 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 |
| Confidentiality | 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 |
| Consistency | Measures whether or not data is equivalent across systems or location of storage |
| Contact | Individual or organisational contact points for the data or metadata, including information on how to reach the contact points |
| Correctness | Every set of data stored represents a real world situation |
| Cost effectiveness | The extent to which the cost of collecting appropriate data is reasonable |

**Table A.2:** List of Data Quality Attributes and Their Definitions Provided in the Survey 2 Questionnaire

| Data Quality Attribute | Definition of the Attribute Provided in the Survey Questionnaire |
|---|---|
| Credibility | The extent to which data are trusted or highly regarded in terms of their source or content |
| Currency | The measure of whether data values are the most up-to-date version of the information |
| Ease of operation | The extent to which data are easily managed and manipulated (i.e., updated, moved, aggregated, reproduced, customized) |
| Efficiency | 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 |
| Fitness | 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. |
| Flexibility | The extent to which data are expandable, adaptable, and easily applied to other needs |
| Frequency of dissemination | The time interval at which the statistics are disseminated over a given time period |
| Institutional mandate | 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 |
| Interpretability | The extent to which data are in an appropriate language and units and the data definitions are clear |
| Latency | The time between when the data was created and when it was made available for use |
| Lineage | Lineage measures whether factual documentation exists about where data came from, how it was transformed, where it went and end-to-end graphical illustration |
| Portability | 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 |
| Objectivity | The extent to which data are unbiased (unprejudiced) and impartial |
| Reasonability | Asks whether a data pattern meets expectations |
| Release policy | Rules for disseminating statistical data to all interested parties |
| Relevance | The extent to which data are applicable and helpful for the task at hand |
| Reliability | Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value |
| Representational consistency | The extent to which data are always presented in the same format and are compatible with previous data |

**Table A.2:** List of Data Quality Attributes and Their Definitions Provided in the Survey 2 Questionnaire

| Data Quality Attribute | Definition of the Attribute Provided in the Survey Questionnaire |
|---|---|
| Structure | It refers to the level of difficulty in transforming semi-structured or unstructured data to structured data through technology |
| Timeliness | Length of time between data availability and the event or phenomenon the data describe |
| Traceability | The extent to which data are well documented, verifiable, and easily attributed to a source |
| Understandability | 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 |
| Uniqueness | No entity exists more than once within the dataset |
| Usability | Is it understandable, simple, relevant, accessible, maintainable and at the right level of precision? |
| Usefulness | Extent to which information is applicable and helpful for the task at hand |
| Validity | Refers to whether data values are consistent with a defined domain of values |
| Variety | The extent to which data are available from several differing data sources |

**Data Availability Challenges - Data Quality Attributes**

The questions in this section attempt to understand if certain data availability challenges affect data quality attributes or not.

16. Does Data Delay (i.e. there is a delay in transmission of data from the source to the destination, e.g. data store to function, sensor to sensor) affect any of the following data quality attributes? (For e.g. read the question as in "is completeness of data affected by data delay?")

    - Accessibility [Yes/No]
    - Availability [Yes/No]
    - Completeness [Yes/No]
    - Currency [Yes/No]
    - Efficiency [Yes/No]
    - Latency [Yes/No]
    - Portability [Yes/No]

- Timeliness [Yes/No]
- Usefulness [Yes/No]

17. Does Data Drop (e.g. dropping of a required attribute of data, dropping a chunk of data during transmission) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by data drop?")

    - Accessibility [Yes/No]
    - Accuracy [Yes/No]
    - Availability [Yes/No]
    - Completeness [Yes/No]
    - Consistency [Yes/No]
    - Currency [Yes/No]
    - Efficiency [Yes/No]
    - Fitness [Yes/No]
    - Flexibility [Yes/No]
    - Objectivity [Yes/No]
    - Portability [Yes/No]
    - Reasonability [Yes/No]
    - Reliability [Yes/No]
    - Timeliness [Yes/No]
    - Usefulness [Yes/No]

18. Does Incomplete Data (e.g. missing attribute of data, missing chunk of data) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by incomplete data?")

    - Accuracy [Yes/No]
    - Availability [Yes/No]
    - Completeness [Yes/No]
    - Consistency [Yes/No]

- Correctness [Yes/No]
- Credibility [Yes/No]
- Currency [Yes/No]
- Efficiency [Yes/No]
- Fitness [Yes/No]
- Flexibility [Yes/No]
- Objectivity [Yes/No]
- Reasonability [Yes/No]
- Reliability [Yes/No]
- Understandability [Yes/No]
- Usability [Yes/No]
- Usefulness [Yes/No]

19. Does Low labeled data volume (i.e. the amount of data that is labeled is lesser than the the amount of data that is unlabeled) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by low labeled data volume?")

    - Accuracy [Yes/No]
    - Availability [Yes/No]
    - Correctness [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]
    - Usability [Yes/No]
    - Usefulness [Yes/No]
    - Validity [Yes/No]

**Data Management Challenges - Data Quality Attributes**

The questions in this section attempt to understand if certain data management challenges affect data quality attributes or not.

20. Does Data acquisition (i.e. purchase of data from third-party and processes

associated with it) affect any of the following data quality attributes? (For e.g. read the question as in "is accessibility of data affected by data acquisition?")

- Accessibility [Yes/No]
- Availability [Yes/No]
- Cost effectiveness [Yes/No]
- Ease of operation [Yes/No]
- Lineage [Yes/No]
- Traceability [Yes/No]

21. Does Data ownership (i.e. who is the legal owner of the data and can use it without any prior approval?) affect any of the following data quality attributes? (For e.g. read the question as in "is accessibility of data affected by data ownership?")

    - Accessibility [Yes/No]
    - Auditability [Yes/No]
    - Compliance [Yes/No]
    - Confidentiality [Yes/No]
    - Ease of operation [Yes/No]
    - Lineage [Yes/No]
    - Traceability [Yes/No]

22. Does Imbalanced dataset (e.g. data from a single location or single weather type dominates other locations or weathers) affect any of the following data quality attributes? (For e.g. read the question as in "is correctness of data affected by imbalanced dataset?")

    - Correctness [Yes/No]
    - Efficiency [Yes/No]
    - Fitness [Yes/No]
    - Usability [Yes/No]
    - Usefulness [Yes/No]

23. Does Redundant data affect any of the following data quality attributes? (For

e.g. read the question as in "is accuracy of data affected by redundant data?")

- Accuracy [Yes/No]
- Objectivity [Yes/No]
- Uniqueness [Yes/No]
- Usability [Yes/No]

24. Does Improper data transfer (e.g. mismanagement of the way data is transferred between source and destination) affect any of the following data quality attributes? (For e.g. read the question as in "is completeness of data affected by improper data transfer?")

    - Completeness [Yes/No]
    - Consistency [Yes/No]
    - Correctness [Yes/No]
    - Currency [Yes/No]
    - Ease of operation [Yes/No]
    - Portability [Yes/No]
    - Reliability [Yes/No]

25. Does Manual Data Collection affect any of the following data quality attributes? (For e.g. read the question as in "is accessibility of data affected by manual data collection?")

    - Accessibility [Yes/No]
    - Cost effectiveness [Yes/No]
    - Ease of operation [Yes/No]
    - Timeliness [Yes/No]

26. Does Manual Data Labeling affect any of the following data quality attributes? (For e.g. read the question as in "is timeliness of data affected by manual data labeling?")

    - Cost effectiveness [Yes/No]
    - Ease of operation [Yes/No]
    - Timeliness [Yes/No]

27. Does Regulatory Compliance affect any of the following data quality attributes? (For e.g. read the question as in "is confidentiality of data affected by regulatory compliance?")

    - Access security [Yes/No]
    - Compliance [Yes/No]
    - Contact [Yes/No]
    - Confidentiality [Yes/No]
    - Frequency of dissemination [Yes/No]
    - Institutional mandate [Yes/No]
    - Lineage [Yes/No]
    - Portability [Yes/No]
    - Release policy [Yes/No]
    - Traceability [Yes/No]

**Data Source Challenges - Data Quality Attributes**

The questions in this section attempt to understand if certain data source challenges affect data quality attributes or not.

28. Does New data types from various sources affect any of the following data quality attributes? (For e.g. read the question as in "is relevance of data affected by new data types from various sources?")

    - Relevance [Yes/No]

29. Does Data dependent on external conditions affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by data dependent on external conditions?")

    - Accuracy [Yes/No]
    - Accessibility [Yes/No]
    - Correctness [Yes/No]

**Data Structure Challenges - Data Quality Attributes**

The questions in this section attempt to understand if certain data structure challenges affect data quality attributes or not.

30. Does Incompatible data formats affect any of the following data quality attributes? (For e.g. read the question as in "is consistency of data affected by incompatible data formats challenge?")

    - Consistency [Yes/No]
    - Interpretability [Yes/No]
    - Validity [Yes/No]

31. Does Outlier data (i.e. data is out of bounds of acceptable range) affect any of the following data quality attributes? (For e.g. read the question as in "is consistency of data affected by outlier data challenge?")

    - Accuracy [Yes/No]
    - Correctness [Yes/No]
    - Credibility [Yes/No]
    - Efficiency [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]

32. Does Unstructured data (i.e. data is out of bounds of acceptable range) affect any of the following data quality attributes? (For e.g. read the question as in "is consistency of data affected by outlier data challenge?")

    - Credibility [Yes/No]
    - Efficiency [Yes/No]
    - Representational consistency [Yes/No]
    - Structure [Yes/No]
    - Usability [Yes/No]
    - Validity [Yes/No]

**Data Trust Challenges - Data Quality Attributes**

The questions in this section attempt to understand if certain data trust challenges affect data quality attributes or not.

33. Does Lack of good data from simulations (i.e. data from simulations are relatively of lesser quality than data from real-world collection) affect any of the following data quality attributes? (For e.g. read the question as in "is

accuracy of data affected by lack of good data from simulations?")

- Accuracy [Yes/No]
- Credibility [Yes/No]
- Fitness [Yes/No]
- Objectivity [Yes/No]
- Usefulness [Yes/No]
- Variety [Yes/No]

34. Does Incorrect labeling affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by incorrect labeling?")

    - Accuracy [Yes/No]
    - Correctness [Yes/No]
    - Credibility [Yes/No]
    - Efficiency [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]
    - Reliability [Yes/No]
    - Usability [Yes/No]
    - Usefulness [Yes/No]
    - Validity [Yes/No]

35. Does Noise (unwanted data that is mixed with valuable data) affect any of the following data quality attributes? (For e.g. read the question as in "is accuracy of data affected by noise?")

    - Accuracy [Yes/No]
    - Correctness [Yes/No]
    - Fitness [Yes/No]
    - Objectivity [Yes/No]
    - Usefulness [Yes/No]

## A.4 Challenge Score

### A.4.1 Survey 1

Following is the ranking of data quality challenges given by participants in Survey 1 during the first iteration. Here, *A*-*F* are the six survey participants, *Total Score* is the sum of the product of rankings, and *Challenge Score* is the final normalized *Challenge Score*. The data is presented in descending order of the *Challenge Score*.

**Note:** Rows with gray background are the challenges removed from the final version of the artifact.

**Table A.3:** Survey 1 - Challenge Score

| Rank | Challenge Set | Challenge | A | B | C | D | E | F | Total Score | Challenge Score |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Data Availability | Low Labeled Data Volume | 4 | 1 | 4 | 4 | 4 | 4 | 104 | 4.333 |
| 2 | Data Source | Lack of Variety in Test Environment | 6 | 5 | 6 | 7 | 6 | 7 | 159 | 3.786 |
| 3 | Data Availability | Incomplete Data | 3 | 4 | 2 | 3 | 2 | 3 | 80 | 3.333 |
| 4 | Data Source | Data Dependent on External Conditions | 4 | 7 | 7 | 5 | 7 | 2 | 136 | 3.238 |
| 5 | Data Management | Manual Data Labeling | 10 | 3 | 8 | 10 | 9 | 9 | 193 | 3.217 |
| 6 | Data Management | Imbalanced Dataset | 7 | 10 | 1 | 9 | 10 | 10 | 178 | 2.967 |
| 7 | Data Availability | Data Drop | 2 | 3 | 3 | 1 | 3 | 2 | 68 | 2.833 |
| 8 | Data Trust | Incorrect Labeling | 5 | 1 | 4 | 5 | 5 | 4 | 80 | 2.667 |
| 9 | Data Source | Wrongly-Calibrated / Defective Sensors | 3 | 4 | 2 | 6 | 4 | 6 | 111 | 2.643 |
| 10 | Data Trust | Uncertain Data Quality Identification | 4 | 5 | 5 | 2 | 4 | 2 | 78 | 2.600 |
| 11 | Data Management | Reliance on Suppliers to Raise Error | 6 | 5 | 10 | 5 | 7 | 7 | 155 | 2.583 |

**Table A.3:** Survey 1 - Challenge Score

| Rank | Challenge Set | Challenge | A | B | C | D | E | F | Total Score | Challenge Score |
|---|---|---|---|---|---|---|---|---|---|---|
| 12 | Data Management | Manual Data Collection | 9 | 2 | 7 | 6 | 8 | 4 | 150 | 2.500 |
| 13 | Data Source | Noise | 5 | 6 | 4 | 3 | 1 | 5 | 103 | 2.452 |
| 14 | Data Source | New Data Type | 7 | 3 | 5 | 2 | 5 | 1 | 101 | 2.405 |
| 15 | Data Management | Large Volume of Data | 8 | 7 | 5 | 7 | 3 | 8 | 135 | 2.250 |
| 16 | Data Trust | Wrong Metadata | 2 | 4 | 3 | 4 | 2 | 3 | 61 | 2.033 |
| 17 | Data Management | Regulatory Compliance | 4 | 8 | 6 | 2 | 6 | 2 | 112 | 1.867 |
| 18 | Data Structure | Unstructured Data | 5 | 5 | 5 | 3 | 2 | 5 | 55 | 1.833 |
| 19 | Data Management | Data Ownership | 5 | 9 | 9 | 3 | 2 | 1 | 109 | 1.817 |
| 20 | Data Trust | Lack of Good Data from Simulation | 3 | 3 | 2 | 3 | 3 | 1 | 54 | 1.800 |
| 21 | Data Management | Improper Data Transfer | 1 | 4 | 4 | 8 | 4 | 6 | 100 | 1.667 |
| 22 | Data Availability | Data Delay | 1 | 2 | 1 | 2 | 1 | 1 | 38 | 1.583 |
| 23 | Data Source | Reliance on Single Data Source | 1 | 1 | 3 | 4 | 3 | 4 | 66 | 1.571 |
| 24 | Data Structure | Data Mix-up | 1 | 1 | 3 | 1 | 5 | 3 | 47 | 1.567 |
| 25 | Data Structure | Outlier Data | 2 | 3 | 1 | 4 | 4 | 1 | 43 | 1.433 |
| 26 | Data Trust | Fake Data | 1 | 2 | 1 | 1 | 1 | 5 | 42 | 1.400 |
| 27 | Data Management | Redundant Data | 2 | 6 | 3 | 1 | 5 | 5 | 82 | 1.367 |
| 28 | Data Structure | Incompatible Data Formats | 4 | 4 | 4 | 4 | 1 | 2 | 40 | 1.333 |
| 28 | Data Structure | Data Fragmentation | 3 | 2 | 2 | 2 | 3 | 4 | 40 | 1.333 |
| 30 | Data Source | Fast Increasing Data | 2 | 2 | 1 | 1 | 2 | 3 | 52 | 1.238 |
| 31 | Data Management | Data Acquisition | 3 | 1 | 2 | 4 | 1 | 3 | 51 | 0.850 |

**Note:** In Table A.3, *Expensive Procedure* and *Time Consuming* challenges are not included as they were identified during the second iteration.

Table A.4 presents the values of Likert scale selected for each challenge set by the survey participants. Here, *A*-*F* are the six survey participants. The data is presented in alphabetical order of the challenge set.

**Table A.4:** Survey 1 - Ranking of Challenge Sets

| Challenge Set | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 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 |

### A.4.2 Survey 2

Following is the ranking of data quality challenges given by participants in Survey 2 during the third iteration. Here, *A-D* are the four survey participants, *Total Score* is the sum of the product of rankings, and *Challenge Score* is the final normalized *Challenge Score*. The data is presented in descending order of the *Challenge Score*.

**Note:** Rows with gray background are the challenges added during the second iteration of the study.

**Table A.5:** Survey 2 - Challenge Score

| Rank | Challenge Set | Challenge | A | B | C | D | Total Score | Challenge Score |
|---|---|---|---|---|---|---|---|---|
| 1 | Data Availability | Low Labeled Data Volume | 3 | 4 | 4 | 4 | 60 | 3.750 |
| 1 | Data Trust | Incorrect Labeling | 3 | 3 | 2 | 3 | 45 | 3.750 |
| 3 | Data Source | Wrongly-Calibrated / Defective Sensors | 4 | 2 | 4 | 3 | 58 | 3.625 |
| 3 | Data Source | Lack of Variety in Test Environment | 3 | 4 | 3 | 4 | 58 | 3.625 |
| 5 | Data Availability | Incomplete Data | 4 | 3 | 3 | 3 | 52 | 3.250 |
| 6 | Data Management | Imbalanced Dataset | 8 | 6 | 9 | 10 | 121 | 3.025 |
| 7 | Data Source | Noise | 2 | 1 | 3 | 1 | 35 | 2.917 |
| 8 | Data Management | Large Volume of Data | 4 | 8 | 10 | 4 | 106 | 2.650 |
| 9 | Data Structure | Outlier Data | 4 | 3 | 2 | 3 | 40 | 2.500 |
| 10 | Data Structure | Incompatible Data Formats | 3 | 2 | 4 | 1 | 39 | 2.438 |
| 11 | Data Management | Manual Data Collection and Labeling | 6 | 7 | 7 | 5 | 97 | 2.425 |
| 12 | Data Management | Data Ownership | 2 | 10 | 6 | 9 | 96 | 2.400 |
| 13 | Data Management | Time Consuming | 9 | 9 | 2 | 6 | 94 | 2.350 |
| 14 | Data Management | Reliance on Suppliers to Raise Error | 7 | 5 | 5 | 8 | 89 | 2.225 |
| 15 | Data Availability | Data Drop | 2 | 2 | 2 | 2 | 32 | 2.000 |
| 15 | Data Structure | Data Fragmentation | 1 | 4 | 1 | 4 | 32 | 2.000 |
| 17 | Data Management | Improper Data Transfer | 5 | 3 | 8 | 3 | 78 | 1.950 |

**Table A.5:** Survey 2 - Challenge Score

| Rank | Challenge Set | Challenge | A | B | C | D | Total Score | Challenge Score |
|---|---|---|---|---|---|---|---|---|
| 18 | Data Source | Data Dependent on External Conditions | 1 | 3 | 2 | 2 | 31 | 1.937 |
| 19 | Data Management | Expensive Procedure | 10 | 2 | 3 | 7 | 77 | 1.925 |
| 20 | Data Trust | Lack of Good Data from Simulation | 1 | 2 | 1 | 2 | 22 | 1.833 |
| 21 | Data Structure | Unstructured Data | 2 | 1 | 3 | 2 | 29 | 1.813 |
| 22 | Data Source | New Data Type | 2 | 1 | 1 | 1 | 23 | 1.438 |
| 23 | Data Management | Data Acquisition | 1 | 4 | 4 | 2 | 44 | 1.100 |
| 24 | Data Availability | Data Delay | 1 | 1 | 1 | 1 | 16 | 1.000 |
| 25 | Data Management | Redundant Data | 3 | 1 | 1 | 1 | 23 | 0.575 |
| 26 | Data Management | Regulatory Compliance | | | | | | |

**Note:**

(1) 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 the purpose of ranking. They are still regarded as separate challenges in the *List of Challenges* artifact component.

(2) Due to technical error, *Regulatory Compliance* was not included in the second iteration survey. Hence, the calculation of *Challenge Score* ranking disregards it. This is only for purpose of calculation of the *Challenge Score*; the challenge is still included in the *List of Challenges* artifact component.

Table A.6 presents the values of Likert scale selected for each challenge set by the survey participants. Here, *A-D* are the four survey participants. The data is presented in alphabetical order of the challenge set.

**Table A.6:** Survey 2 - Ranking of Challenge Sets

| Challenge Set | A | B | C | D |
|---|---|---|---|---|
| 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 |

## A.5 Themes and Codes

**Note:** One of the themes — *Challenges* — is not mentioned in Table A.7 as the codes in that themes represent the challenges listed in the *List of Challenges* artifact component.

**Table A.7:** List of Identified Themes and Codes Associated with them

| Themes | Codes |
|---|---|
| Applications | Adaptive Cruise Control, Advanced Driving Assistance Systems, Acceleration Request Management, Analog to Digital Conversion, Anomaly detection, ASP Software, Automatic Emergency Braking, Base Software for ECU, Braking Request Management, Computer Vision, Intrusion detection, Pattern detection, Vehicle-to-Anything, Pedestrian detection, Object detection, Lane departure assist, Headlight function, Object classification, Lane detection, Vehicle detection, Level2 automation, Traffic sign detection |
| Current Procedures | Data collection, Data management, Data review process, Simulated validation, Data validation, Data security, Data access process, Calibration checking process, Data storage, Error reporting process, Data processing, Lack of segregation of good and bad data, Goal-based data collection, Data control process, Data quality level, Bad data removal, Requirement areas, Data blacklist, Lack of automation, Data pipeline, Metadata representing, purpose-stability of data, Data pipeline, Written documentation, Model-based predictive system, Virtual consent control system, Team-dependent procedures, Monitoring, Requirements management, Feature development, Approve projects, Peer review, Release, Testing, System validation, Function pre-development, Platform development, Resource allocation. |
| Data Assumptions | Trust in labeled data, Trust in bounding box, Trust in sensor calibration |

**Table A.7:** List of Identified Themes and Codes Associated with them

| Themes | Codes |
|---|---|
| Data Types | Time for impact, Number of pixels, Weather data, Size of object, Camera data, Video stream, Distance to object, Interior sound level, Geographic position data, Function performance data, Vehicle type, Frame rate, Velocity, Wheel speed, Pedestrian direction, Driver's mood, Timestamp, Angle, Position, Metadata, Weight of vehicle, Driver's behavior, Temporal data, Spatial data, Temporal-spatial data, Data variation(Road types variation, Temporal variation, Weather variation, Geographic variation, Lighting variation, Vehicular variation), Vehicle component status(Brake stability status, Wiper status, High beam status, Seat belt status, Turn indicator status, Engine status, Battery status) |
| Extra Info | ISO standard, Predictable event, Real-world driving, Legality of GPS data collection, Tier 1 stakeholders, Cost vs time trade-off, Inability to disclose, Original equipment manufacturers, Customers, Vehicle dynamics model, Simulated driving, Iterative process, Load distribution vs balance, Opinion of machine learning engineers vs product owner, GDPR, Data labeling(Traffic sign labeling, Pedestrian labeling, Automatic label generation, Semi-automatic label generation, Manual labeling) |

**Table A.7:** List of Identified Themes and Codes Associated with them

| Themes | Codes |
|---|---|
| Goals | Timely activate safety systems, Validate trained algorithms, Identify improvements, Uninterrupted data collection, Proper function switch off, Well-calibrated sensor, Cover all scenarios, Functional safety, Manage-measure-improve, Evolving goals, Evaluate data source, Development of presently-unknown future functionality, Continuous data collection, Application safety, Crash avoidance, Pixelwise separation of image, Track system development, Identify problems with data, Improve data quality, Get closer to reference data, Correct data labeling, Calculate time for impact, Cost reduction, Assist driver, Appropriate sensor selection, Balanced sensor performance, Purpose-suitability of data, Less manual labeling, Data reuse, Synchronized data, Semantic segmentation, Correct system behavior, Functional correctness, Verify missed corner cases, Timely communication of information, Team-dependent requirements, Maximize performance(Maximize vision system performance), Train deep learning algorithms(CNN), Learn appropriate metrics, Project dependent requirements |
| Hardware Components | Inertia measurement unit, Pressure sensor, Internal sensors, Reference sensors, Detection sensors, Hydraulic systems, Breakout box, Radar, electronic Control Unit, GPS, External sensors, Optical cables, Automotive Ethernet, Single-sensor system, Sampling frequency, Multi-sensor system, Camera(Report uncertain data, Record video, Steriovision camera, Monovision camera), Actuators (Brakes), Lidar (Range, Distance) |
| Impact of Low Data Quality | Underfitting, Overfitting, Unwanted algorithm behaviors(Algorithm reset), Improper algorithm training, False positive, Issue low data quality warning, Low precision, Biased data, Untimely system activation, Aggressive system behavior, Low confidence level, False negative |

**Table A.7:** List of Identified Themes and Codes Associated with them

| Themes | Codes |
|---|---|
| Metrics | Standard deviation, Usability, Timeliness of data, Percentage error, Error rate, Tracking duration, Uniformity, Variance, Age of data, Frequency, Data loss rate, Volume, Velocity of data, Availability, Error margin, Accuracy, Mean of data, Update rate, System size, Elasticity, Scalability, Correctness of data, Estimated bias, Signal-to-noise ratio, Signal strength, Completeness, Resiliency |
| Nature of Data | Corner case data, Generic data, Validation data, Pre-processed data, Data from simulation, Outlier, Reference data, Difficult-to-obtain data, Difficult-to-label data, Training data, Rare data, First-party data, Representative data, Raw data, Computed data, Digitally-convertible data, Algorithm-dependent data, Third-party data |
| Solutions | Correct labeling, Heuristics(Identify commonly-occurring field), Data visualization, MMI scorecard, Marking bad data, Automated quality analysis, Collection of many varieties of data, Bounding box, Guiding principles, Data contract, Error reporting, Couple data quality with physical behavior like SNR, Use of requirements engineering tracing tool, Regressive testing, Integrate multiple data frames, Automated data analysis, Simulated testing, Review data, Software-in-the-loop testing, Simplified reporting, Sensor redundancy, Infrastructure to handle large data, Quick feedback loop, Predictive modeling, Scenario based documentation |
| Team Structure | Dedicated specialist, Technical sales, Data collection team, Agile team, Feature tech lead, Function developer, Technical lead, Manager, Data management team, Tester, Applied researcher, Product owner, Non-static team, Data scientist, Machine learning engineer |

## A.6 Focus Group Data

### A.6.1 Challenge Ranking

Tables A.8, A.9, A.10, A.11, and A.12 present the data quality challenges and the number of participants who selected respective challenges as their preferred ranking (e.g., the number of items a particular challenge is selected as the 1st (i.e., top-most pressing), 2nd, and so on).

**Table A.8:** Focus Group - *Data Availability* Challenges Ranking

| Challenge | 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|
| Data Delay | 0 | 1 | 1 | 3 |
| Data Drop | 0 | 1 | 3 | 1 |
| Incomplete Data | 3 | 2 | 0 | 0 |
| Low Labeled Data Volume | 2 | 1 | 1 | 1 |

**Table A.9:** Focus Group - *Data Management* Challenges Ranking

| Challenge | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th | 11th | 12th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data Acquisition | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Data Ownership | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| Expensive Procedure | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| Imbalanced Dataset | 2 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Improper Data Transfer | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 |
| Large Volume of Data | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Manual Data Collection | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Manual data Labeling | 1 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| Redundant Data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Regulatory Compliance | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| Reliance on Suppliers to Raise Error | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Time Consuming | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |

**Table A.10:** Focus Group - *Data Source* Challenges Ranking

| Challenge | 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|
| Data Dependent on External Conditions | 1 | 2 | 1 | 0 |
| Lack of Variety in Test Environment | 3 | 1 | 0 | 0 |
| New Data Types | 0 | 1 | 3 | 0 |
| Wrongly-calibrated / Defective Sensor | 0 | 0 | 0 | 4 |

**Table A.11:** Focus Group - *Data Structure* Challenges Ranking

| Challenge | 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|
| Data Fragmentation | 0 | 0 | 2 | 2 |
| Incompatible Data Formats | 0 | 3 | 0 | 1 |
| Outlier Data | 2 | 0 | 1 | 1 |
| Unstructured Data | 2 | 1 | 1 | 0 |

**Table A.12:** Focus Group - *Data Trust* Challenges Ranking

| Challenge | 1st | 2nd | 3rd |
|---|---|---|---|
| Incorrect Labeling | 3 | 0 | 1 |
| Lack of Good Data from Simulations | 1 | 3 | 0 |
| Noise | 0 | 1 | 3 |

Table A.13 presents the number of times the challenge sets are given a certain value in a Likert scale.

**Table A.13:** Focus Group - Challenge Set Ranking

| Challenge Set | 1 | 2 | 3 | 4 | 5 | 6 | Weighted Average |
|---|---|---|---|---|---|---|---|
| Data Availability | 0 | 1 | 1 | 1 | 1 | 0 | 3.50 |
| Data Management | 0 | 1 | 0 | 1 | 2 | 0 | 4.00 |
| Data Source | 0 | 0 | 2 | 1 | 0 | 1 | 4.00 |
| Data Structure | 0 | 1 | 3 | 0 | 0 | 0 | 2.75 |
| Data Trust | 0 | 0 | 0 | 0 | 1 | 3 | 5.75 |

### A.6.2 Data Quality Challenge - Attribute Association

Tables between Table A.14 and Table A.28 present the validation results from the focus group session. It includes the number of participants who selected *"yes, the challenge affects the attribute"* and *"no, the challenge does not affect the attribute."*

**Data Availability Challenges**

**Table A.14:** *Data Delay* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accessibility | 2 | 2 | 0.5 |
| Availability | 1 | 3 | 0.25 |
| Completeness | 0 | 4 | 0 |
| Currency | 4 | 0 | 1 |
| Efficiency | 3 | 1 | 0.75 |
| Latency | 4 | 0 | 1 |
| Portability | 0 | 4 | 0 |
| Timeliness | 4 | 0 | 1 |
| Usefulness | 2 | 2 | 0.5 |

**Table A.15:** *Data Drop* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accessibility | 3 | 2 | 0.6 |
| Accuracy | 4 | 1 | 0.8 |
| Availability | 5 | 0 | 1 |
| Completeness | 4 | 1 | 0.8 |
| Consistency | 4 | 1 | 0.8 |
| Currency | 2 | 3 | 0.4 |
| Efficiency | 3 | 2 | 0.6 |
| Fitness | 2 | 3 | 0.4 |
| Flexibility | 1 | 4 | 0.2 |
| Objectivity | 1 | 4 | 0.2 |
| Portability | 1 | 4 | 0.2 |
| Reasonability | 2 | 3 | 0.4 |
| Reliability | 4 | 1 | 0.8 |
| Timeliness | 3 | 2 | 0.6 |
| Usefulness | 2 | 3 | 0.4 |

**Table A.16:** *Incomplete Data* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accuracy | 5 | 0 | 1 |
| Availability | 3 | 2 | 0.6 |
| Completeness | 5 | 0 | 1 |
| Consistency | 3 | 2 | 0.6 |
| Correctness | 3 | 2 | 0.6 |
| Credibility | 5 | 0 | 1 |
| Currency | 0 | 5 | 0 |
| Efficiency | 1 | 4 | 0.2 |
| Fitness | 4 | 1 | 0.8 |
| Flexibility | 3 | 2 | 0.6 |

**Table A.16:** *Incomplete Data* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Objectivity | 1 | 4 | 0.2 |
| Reasonability | 4 | 1 | 0.8 |
| Reliability | 4 | 1 | 0.8 |
| Understandability | 4 | 1 | 0.8 |
| Usability | 5 | 0 | 1 |
| Usefulness | 4 | 1 | 0.8 |

**Table A.17:** *Low Labeled Data Volume* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accuracy | 4 | 1 | 0.8 |
| Availability | 1 | 4 | 0.2 |
| Correctness | 3 | 2 | 0.6 |
| Fitness | 4 | 1 | 0.8 |
| Objectivity | 3 | 2 | 0.6 |
| Usability | 5 | 0 | 1 |
| Usefulness | 5 | 0 | 1 |
| Validity | 5 | 0 | 1 |

### Data Management Challenges

**Table A.18:** *Data Acquisition* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accessibility | 4 | 1 | 0.8 |
| Availability | 5 | 0 | 1 |
| Cost Effectiveness | 5 | 0 | 1 |
| Ease of Operation | 2 | 3 | 0.4 |
| Lineage | 5 | 0 | 1 |
| Traceability | 4 | 1 | 0.8 |

**Table A.19:** *Imbalanced Dataset* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Correctness | 5 | 0 | 1 |
| Efficiency | 1 | 4 | 0.2 |
| Fitness | 5 | 0 | 1 |
| Usability | 5 | 0 | 1 |
| Usefulness | 5 | 0 | 1 |

**Table A.20:** *Improper Data Transfer* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Completeness | 3 | 2 | 0.6 |
| Consistency | 3 | 2 | 0.6 |
| Correctness | 3 | 2 | 0.6 |
| Currency | 2 | 3 | 0.4 |
| Ease of Operation | 4 | 1 | 0.8 |
| Portability | 4 | 1 | 0.8 |
| Reliability | 4 | 1 | 0.8 |

**Table A.21:** *Manual Data Collection* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accessibility | 1 | 4 | 0.2 |
| Cost Effectiveness | 5 | 0 | 1 |
| Ease of Operation | 3 | 2 | 0.6 |
| Timeliness | 1 | 4 | 0.2 |

**Table A.22:** *Manual Data Labeling* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Cost Effectiveness | 5 | 0 | 1 |
| Ease of Operation | 4 | 1 | 0.8 |
| Timeliness | 4 | 1 | 0.8 |

**Table A.23:** *Redundant Data* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accuracy | 2 | 3 | 0.4 |
| Objectivity | 1 | 4 | 0.2 |
| Usability | 3 | 2 | 0.6 |
| Usefulness | 3 | 2 | 0.6 |

### Data Source Challenges

**Table A.24:** *Data Dependent on External Conditions* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accessibility | 5 | 0 | 1 |
| Accuracy | 3 | 2 | 0.6 |
| Correctness | 3 | 2 | 0.6 |

### Data Structure Challenges

**Table A.25:** *Outlier Data* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accuracy | 2 | 3 | 0.4 |
| Correctness | 0 | 5 | 0 |
| Credibility | 1 | 4 | 0.2 |
| Efficiency | 1 | 4 | 0.2 |
| Fitness | 1 | 4 | 0.2 |
| Objectivity | 1 | 4 | 0.2 |

### Data Trust Challenges

**Table A.26:** *Incorrect Labeling* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accuracy | 5 | 0 | 1 |
| Correctness | 5 | 0 | 1 |
| Credibility | 5 | 0 | 1 |
| Efficiency | 1 | 4 | 0.2 |
| Fitness | 5 | 0 | 1 |
| Objectivity | 3 | 2 | 0.6 |
| Reliability | 5 | 0 | 1 |
| Usability | 5 | 0 | 1 |
| Usefulness | 5 | 0 | 1 |
| Validity | 5 | 0 | 1 |

**Table A.27:** *Lack of Good Data from Simulations* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accuracy | 4 | 1 | 0.8 |
| Credibility | 3 | 2 | 0.6 |
| Fitness | 4 | 1 | 0.8 |
| Objectivity | 4 | 1 | 0.8 |
| Usefulness | 4 | 1 | 0.8 |
| Variety | 2 | 3 | 0.4 |

**Table A.28:** *Noise* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Accuracy | 5 | 0 | 1 |
| Correctness | 3 | 2 | 0.6 |
| Fitness | 3 | 2 | 0.6 |

**Table A.28:** *Noise* Challenge and Attributes Associated with it

| Attribute | Yes | No | Weighted Average |
|---|---|---|---|
| Objectivity | 1 | 4 | 0.2 |
| Usefulness | 3 | 2 | 0.6 |

## A.7 Survey 2 Data

### A.7.1 Template Fields

**Table A.29:** Template Fields Validation Result for *List of Challenges* Artifact Component

| Field | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|
| Name | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Sources | Yes | Yes | Yes | No | 3 | 1 | 0.75 |
| Description | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Whether the challenge directly affects AI function | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Challenge Score | Yes | Yes | Yes | Yes | 4 | 0 | 1 |

**Table A.30:** Template Fields Validation Result for *List of Data Quality Attributes* Artifact Component

| Field | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|
| Name | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Sources | No | Yes | Yes | No | 2 | 2 | 0.5 |
| Definition | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Which challenges affect the data quality attribute? | Yes | Yes | Yes | Yes | 4 | 0 | 1 |

**Table A.31:** Template Fields Validation Result for *List of Data Quality Attribute Metrics* Artifact Component

| Field | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|
| Data quality attribute | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Metric | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Formula | Yes | Yes | Yes | Yes | 4 | 0 | 1 |

**Table A.32:** Template Fields Validation Result for *Potential Solutions* Artifact Component

| Field | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|
| Name | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Requirements specifications | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Implementation details | Yes | Yes | Yes | Yes | 4 | 0 | 1 |

### A.7.2 Survey Result - Challenges Directly Affecting AI Models

**Table A.33:** List of Challenges Directly Affecting AI Models

| Challenge | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|
| Data Acquisition | Yes | Yes | No | No | 2 | 2 | 0.5 |
| Data Delay | No | Yes | No | No | 1 | 3 | 0.25 |
| Data Dependent on External Conditions | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Data Drop | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Data Ownership | No | Yes | No | No | 1 | 3 | 0.25 |

**Table A.33:** List of Challenges Directly Affecting AI Models

| Challenge | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|
| Expensive Procedure | No | No | No | No | 0 | 4 | 0 |
| Fragmented Data | No | No | No | Yes | 1 | 3 | 0.25 |
| Imbalanced Dataset | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Improper Data Transfer | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Incompatible Data Formats | No | No | Yes | No | 1 | 3 | 0.25 |
| Incomplete Data | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Incorrect Labeling | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Lack of Good Data from Simulations | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Lack of Variety in Test Environment | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Large Volume of Data | No | Yes | Yes | No | 2 | 2 | 0.5 |
| Low Labeled Data Volume | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Manual Data Collection | Yes | Yes | No | No | 2 | 2 | 0.5 |
| Manual Data Labeling | Yes | Yes | No | No | 2 | 2 | 0.5 |
| New Data Types | Yes | Yes | No | No | 2 | 2 | 0.5 |
| Noise | Yes | No | Yes | No | 2 | 2 | 0.5 |
| Outlier Data | Yes | No | Yes | No | 2 | 2 | 0.5 |
| Redundant Data | Yes | No | Yes | No | 2 | 2 | 0.5 |
| Reliance on Suppliers to Raise Error | No | Yes | No | No | 1 | 3 | 0.25 |
| Time Consuming | No | Yes | No | No | 1 | 3 | 0.25 |
| Unstructured Data | No | No | Yes | No | 1 | 3 | 0.25 |
| Wrongly-Calibrated / Defective Sensor | Yes | Yes | Yes | Yes | 4 | 0 | 1 |

### A.7.3 Data Quality Challenge - Data Quality Attribute Association Survey Results

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Data Acquisition | Accessibility | Yes | | Yes | No | 2 | 1 | 0.66 |
| Data Acquisition | Availability | Yes | | Yes | No | 2 | 1 | 0.66 |
| Data Acquisition | Cost effectiveness | Yes | | No | Yes | 2 | 1 | 0.66 |
| Data Acquisition | Ease of operation | Yes | | No | No | 1 | 2 | 0.33 |
| Data Acquisition | Lineage | Yes | | Yes | Yes | 3 | 0 | 1 |
| Data Acquisition | Traceability | Yes | | Yes | Yes | 3 | 0 | 1 |
| | | | | | | | | |
| Data Delay | Accessibility | Yes | No | Yes | No | 2 | 2 | 0.5 |
| Data Delay | Availability | Yes | No | Yes | No | 2 | 2 | 0.5 |
| Data Delay | Completeness | No | No | Yes | No | 1 | 3 | 0.25 |
| Data Delay | Currency | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Data Delay | Efficiency | No | No | Yes | Yes | 2 | 2 | 0.5 |
| Data Delay | Latency | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Data Delay | Portability | No | No | No | No | 0 | 4 | 0 |
| Data Delay | Timeliness | Yes | Yes | Yes | No | 3 | 1 | 0.75 |
| Data Delay | Usefulness | Yes | Yes | Yes | No | 3 | 1 | 0.75 |
| | | | | | | | | |
| Data Dependent on External Conditions | Accessibility | Yes | | Yes | Yes | 3 | 0 | 1 |
| Data Dependent on External Conditions | Accuracy | No | | No | No | 0 | 3 | 0 |

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Data Dependent on External Conditions | Correctness | No | | No | Yes | 1 | 2 | 0.33 |
| | | | | | | | | |
| Data Drop | Accessibility | Yes | No | Yes | No | 2 | 2 | 0.5 |
| Data Drop | Accuracy | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Data Drop | Availability | Yes | Yes | Yes | No | 3 | 1 | 0.75 |
| Data Drop | Completeness | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Data Drop | Consistency | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Data Drop | Currency | Yes | No | No | No | 1 | 3 | 0.25 |
| Data Drop | Efficiency | Yes | No | Yes | No | 2 | 2 | 0.5 |
| Data Drop | Fitness | No | Yes | Yes | No | 2 | 2 | 0.5 |
| Data Drop | Flexibility | No | No | Yes | No | 1 | 3 | 0.25 |
| Data Drop | Objectivity | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Data Drop | Portability | No | No | Yes | | 1 | 2 | 0.33 |
| Data Drop | Reasonability | No | Yes | Yes | No | 2 | 2 | 0.5 |
| Data Drop | Reliability | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Data Drop | Timeliness | Yes | No | No | No | 1 | 3 | 0.25 |
| Data Drop | Usefulness | Yes | Yes | Yes | No | 3 | 1 | 0.75 |
| | | | | | | | | |
| Data Ownership | Accessibility | Yes | | Yes | Yes | 3 | 0 | 1 |
| Data Ownership | Auditability | Yes | | No | No | 1 | 2 | 0.33 |
| Data Ownership | Compliance | Yes | | Yes | Yes | 3 | 0 | 1 |
| Data Ownership | Confidentiality | Yes | | Yes | No | 2 | 1 | 0.66 |
| Data Ownership | Ease of operation | Yes | | No | Yes | 2 | 1 | 0.66 |

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Data Ownership | Lineage | Yes | | Yes | No | 2 | 1 | 0.66 |
| Data Ownership | Traceability | Yes | | Yes | No | 2 | 1 | 0.66 |
| | | | | | | | | |
| Imbalanced Dataset | Correctness | Yes | | No | Yes | 2 | 1 | 0.66 |
| Imbalanced Dataset | Efficiency | No | | No | No | 0 | 3 | 0 |
| Imbalanced Dataset | Fitness | Yes | | No | Yes | 2 | 1 | 0.66 |
| Imbalanced Dataset | Usability | Yes | | Yes | Yes | 3 | 0 | 1 |
| Imbalanced Dataset | Usefulness | Yes | | Yes | Yes | 3 | 0 | 1 |
| | | | | | | | | |
| Improper Data Transfer | Completeness | Yes | | Yes | Yes | 3 | 0 | 1 |
| Improper Data Transfer | Consistency | Yes | | Yes | No | 2 | 1 | 0.66 |
| Improper Data Transfer | Correctness | Yes | | Yes | No | 2 | 1 | 0.66 |
| Improper Data Transfer | Currency | Yes | | Yes | Yes | 3 | 0 | 1 |
| Improper Data Transfer | Ease of operation | Yes | | Yes | No | 2 | 1 | 0.66 |

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Improper Data Transfer | Portability | Yes | | Yes | No | 2 | 1 | 0.66 |
| Improper Data Transfer | Reliability | Yes | | Yes | No | 2 | 1 | 0.66 |
| | | | | | | | | |
| Incompatible Data Formats | Consistency | Yes | | Yes | No | 2 | 1 | 0.66 |
| Incompatible Data Formats | Interpretability | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incompatible Data Formats | Validity | Yes | | Yes | No | 2 | 1 | 0.66 |
| | | | | | | | | |
| Incomplete Data | Accuracy | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Incomplete Data | Availability | Yes | Yes | Yes | No | 3 | 1 | 0.75 |
| Incomplete Data | Completeness | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Incomplete Data | Consistency | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Incomplete Data | Correctness | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Incomplete Data | Credibility | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Incomplete Data | Currency | Yes | Yes | Yes | No | 3 | 1 | 0.75 |
| Incomplete Data | Efficiency | Yes | Yes | Yes | No | 3 | 1 | 0.75 |
| Incomplete Data | Fitness | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Incomplete Data | Flexibility | No | No | No | No | 0 | 4 | 0 |
| Incomplete Data | Objectivity | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Incomplete Data | Reasonability | No | Yes | Yes | No | 2 | 2 | 0.5 |
| Incomplete Data | Reliability | Yes | Yes | Yes | Yes | 4 | 0 | 1 |

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Incomplete Data | Understandability | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Incomplete Data | Usability | Yes | Yes | No | No | 2 | 2 | 0.5 |
| Incomplete Data | Usefulness | No | Yes | Yes | No | 2 | 2 | 0.5 |
| | | | | | | | | |
| Incorrect Labeling | Accuracy | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incorrect Labeling | Correctness | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incorrect Labeling | Credibility | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incorrect Labeling | Efficiency | Yes | | Yes | No | 2 | 1 | 0.66 |
| Incorrect Labeling | Fitness | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incorrect Labeling | Objectivity | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incorrect Labeling | Reliability | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incorrect Labeling | Usability | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incorrect Labeling | Usefulness | Yes | | Yes | Yes | 3 | 0 | 1 |
| Incorrect Labeling | Validity | Yes | | Yes | Yes | 3 | 0 | 1 |
| | | | | | | | | |
| Lack of Good Data from Simulations | Accuracy | Yes | | Yes | No | 2 | 1 | 0.66 |
| Lack of Good Data from Simulations | Credibility | Yes | | Yes | Yes | 3 | 0 | 1 |
| Lack of Good Data from Simulations | Fitness | Yes | | Yes | No | 2 | 1 | 0.66 |
| Lack of Good Data from Simulations | Objectivity | Yes | | Yes | Yes | 3 | 0 | 1 |
| Lack of Good Data from Simulations | Usefulness | Yes | | Yes | No | 2 | 1 | 0.66 |

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Lack of Good Data from Simulations | Variety | Yes | | Yes | Yes | 3 | 0 | 1 |
| | | | | | | | | |
| Low Labeled Data Volume | Accuracy | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Low Labeled Data Volume | Availability | No | Yes | No | No | 1 | 3 | 0.25 |
| Low Labeled Data Volume | Correctness | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Low Labeled Data Volume | Fitness | Yes | Yes | Yes | Yes | 4 | 0 | 1 |
| Low Labeled Data Volume | Objectivity | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Low Labeled Data Volume | Usability | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Low Labeled Data Volume | Usefulness | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| Low Labeled Data Volume | Validity | No | Yes | Yes | Yes | 3 | 1 | 0.75 |
| | | | | | | | | |
| Manual Data Collection | Accessibility | Yes | | No | Yes | 2 | 1 | 0.66 |
| Manual Data Collection | Cost effectiveness | Yes | | No | Yes | 2 | 1 | 0.66 |

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Manual Data Collection | Ease of operation | Yes | | No | Yes | 2 | 1 | 0.66 |
| Manual Data Collection | Timeliness | Yes | | No | Yes | 2 | 1 | 0.66 |
| | | | | | | | | |
| New Data Types from Various Sources | Relevance | No | | Yes | No | 1 | 2 | 0.33 |
| | | | | | | | | |
| Noise | Accuracy | Yes | | Yes | No | 2 | 1 | 0.66 |
| Noise | Correctness | Yes | | Yes | No | 2 | 1 | 0.66 |
| Noise | Fitness | Yes | | Yes | No | 2 | 1 | 0.66 |
| Noise | Objectivity | Yes | | Yes | No | 2 | 1 | 0.66 |
| Noise | Usefulness | Yes | | Yes | No | 2 | 1 | 0.66 |
| | | | | | | | | |
| Outlier Data | Accuracy | Yes | | Yes | No | 2 | 1 | 0.66 |
| Outlier Data | Correctness | Yes | | No | No | 1 | 2 | 0.33 |
| Outlier Data | Credibility | Yes | | Yes | Yes | 3 | 0 | 1 |
| Outlier Data | Efficiency | Yes | | No | No | 1 | 2 | 0.33 |
| Outlier Data | Fitness | Yes | | Yes | Yes | 3 | 0 | 1 |
| Outlier Data | Objectivity | Yes | | No | No | 1 | 2 | 0.33 |
| | | | | | | | | |
| Redundant Data | Accuracy | Yes | | No | No | 1 | 2 | 0.33 |
| Redundant Data | Objectivity | Yes | | No | No | 1 | 2 | 0.33 |
| Redundant Data | Uniqueness | Yes | | Yes | Yes | 3 | 0 | 1 |

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Redundant Data | Usability | Yes | | No | No | 1 | 2 | 0.33 |
| | | | | | | | | |
| Regulatory Compliance | Access security | Yes | | Yes | No | 2 | 1 | 0.66 |
| Regulatory Compliance | Compliance | Yes | | Yes | Yes | 3 | 0 | 1 |
| Regulatory Compliance | Confidentiality | Yes | | Yes | No | 2 | 1 | 0.66 |
| Regulatory Compliance | Contact | Yes | | Yes | No | 2 | 1 | 0.66 |
| Regulatory Compliance | Frequency of dissemination | Yes | | Yes | No | 2 | 1 | 0.66 |
| Regulatory Compliance | Institutional mandate | Yes | | Yes | Yes | 3 | 0 | 1 |
| Regulatory Compliance | Lineage | Yes | | Yes | No | 2 | 1 | 0.66 |
| Regulatory Compliance | Portability | Yes | | Yes | No | 2 | 1 | 0.66 |
| Regulatory Compliance | Release policy | Yes | | Yes | Yes | 3 | 0 | 0 |
| Regulatory Compliance | Traceability | Yes | | Yes | No | 2 | 1 | 0.66 |
| | | | | | | | | |
| Unstructured Data | Credibility | No | | No | No | 0 | 3 | 0 |
| Unstructured Data | Efficiency | Yes | | No | Yes | 2 | 1 | 0.66 |

**Table A.34:** List of Data Quality Challenge - Attribute Association Survey Validation Results

| Challenge | Attribute | A | B | C | D | No. of "Yes" | No. of "No" | Weighted Average |
|---|---|---|---|---|---|---|---|---|
| Unstructured Data | Representational consistency | Yes | | Yes | Yes | 3 | 0 | 1 |
| Unstructured Data | Structure | Yes | | Yes | No | 2 | 1 | 0.66 |
| Unstructured Data | Usability | Yes | | No | No | 1 | 2 | 0.33 |
| Unstructured Data | Validity | Yes | | No | No | 1 | 2 | 0.33 |