Identifying and Managing Data Quality Requirements: A Design Science Study in the Field of Automated Driving - Artifact Package

Shameer Kumar Pradhan, Hans-Martin Heyn, Eric Knauss

1 Introduction

This document acts as an artifact package for the article titled ***Identifying and Managing Data Quality Requirements: A Design Science Study in the Field of Automated Driving***. It lists and explains the components of the artifact developed in the study. It covers the following components.

- Section 2 - Data Quality Workflow
- Section 3 - List of Data Quality Challenges
- Section 4 - List of Data Quality Attributes
- Section 5 - Solution Candidates

2 Data Quality Workflow

This component presents a step-by-step workflow for assessing and managing data quality and requirements. It includes six steps, as shown in Fig. 1. Most of the steps can be performed in parallel, as depicted by the dotted line in Fig. 1. Loops indicate that the steps can be done iteratively.

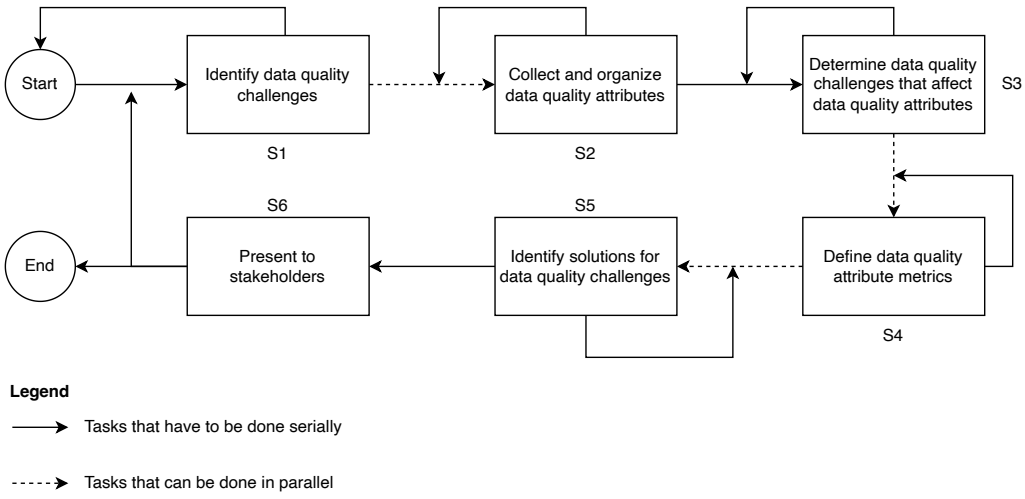


Figure 1: *Data Quality Workflow* Framework Component

1. **Identify data quality challenges**

   Challenges concerning data quality can be identified from several sources. Primary sources of data collection, such as interviews, field studies, and surveys, can be utilized to identify the challenges. Research papers and books can be used as second-hand sources as well. Furthermore, the collected challenges can be divided into different categories. In this study, 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, and interviews. Data quality attributes can also be elicited from interviews. A single attribute can also represent differently phrased data quality attributes. E.g., *understandability* and *ease of understanding attributes* can be represented by the same attribute.

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

   Data quality challenges and quality attributes can be associated with each other after their identification. The association, 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 those 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**

   Metrics to measure data quality attributes are formulated in this step. The metrics help to put a quantitative value on 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. E.g., the *degree of accuracy* can be calculated as a ratio of the number of correctly labeled data records and the total number of data records. The formulae are mostly dependent on the context of the application.

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

   A way of improving data quality attribute metrics, and thus, improving 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). Several sources, such as research papers, technical reports, and books, can identify solutions. Teams can also brainstorm and devise new solution candidates for the challenges. An effective way to validate solution candidates is to implement them as tests in part of a system.

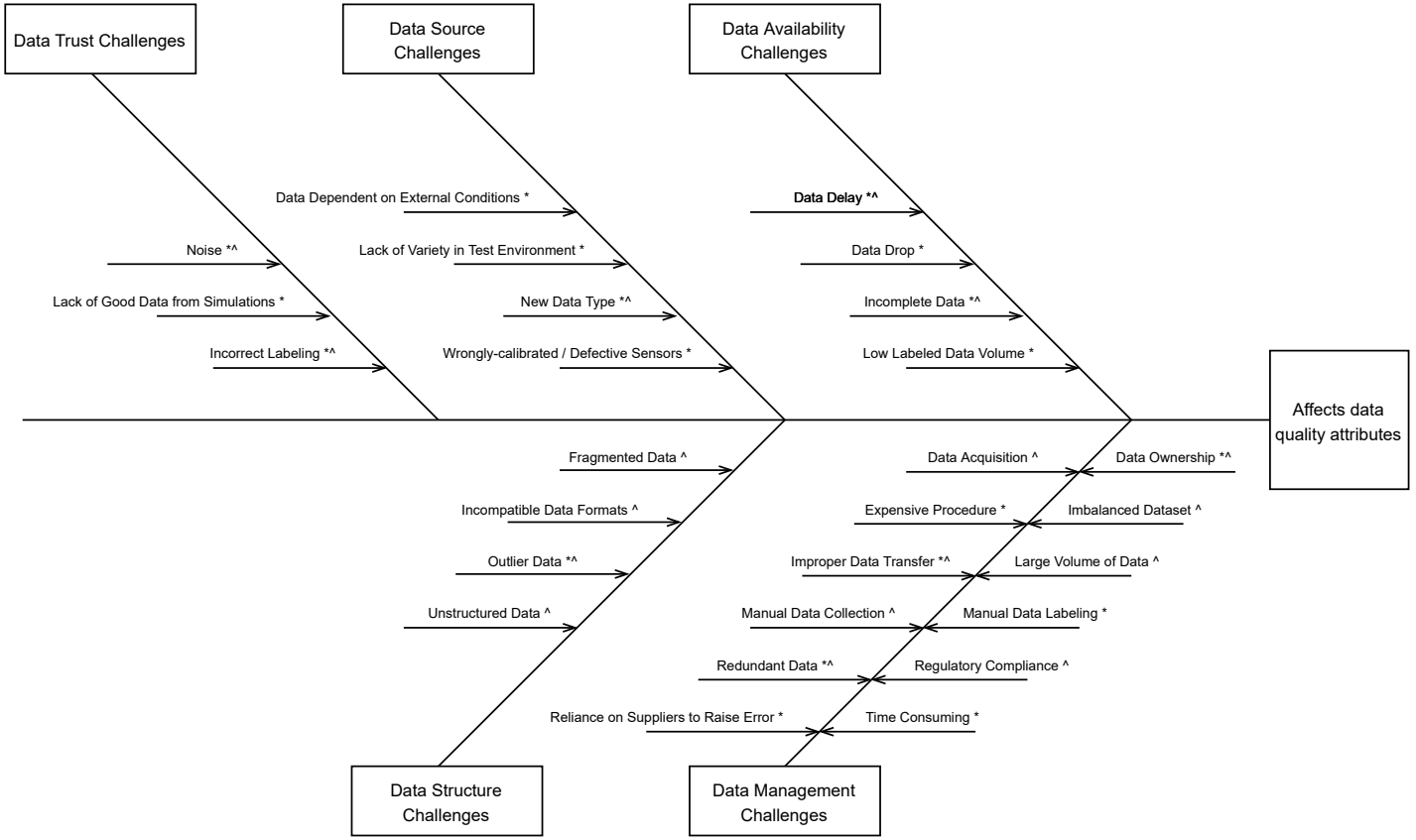
6. **Present to stakeholders**

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

3 List of Data Quality Challenges

Table 1: Definition of Challenge Sets and List of Challenges

Challenge Set	Definition	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 data management and 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 by the lack of transparency in the data and 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 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 the "Directly affects AI Functions" field, the response from the survey in the third cycle is provided. The response is in the format of the number of "yes" and "no" answered by the survey participants.

3.1 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. E.g., a delay can occur in data transmission from sensor to long-term storage, sensor to deep learning functions, and long-term storage to deep learning functions. Similarly, there can also be a delay in receiving a signal sent out by a sensor.
Directly affects AI Functions: 1 "Yes", 3 "No"
Challenge Score: Survey 1 - 1.583 (Rank 22/31), Survey 2 - 1.000 (Rank 24/25)

Name: Data Drop
Reference: Interviewee D
Description: Some data cycles are dropped now and then, which causes tracking of data to be difficult and disrupts the management and processing of data. Such disruption, in turn, will hinder the training of deep learning models. E.g., dropping three frames in a 30-second clip would mean losing 0.7 seconds, thus causing 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),
Description: This challenge is similar to data drop as 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 a drop in bits.
Directly affects AI functions: 3 "Yes", 1 "No"
Challenge score: Survey 1 - 3.333 (Rank 3/31), Survey 2 - 3.250 (Rank 5/25)

Name: Low Labeled Data Volume
Reference: Interviewee C
Description: Most of the time, in the training dataset, the volume of the labeled data is significantly lesser than the volume of the unlabeled data. Since a large volume of data is unlabeled, the unlabeled data is useless, and the deep learning models cannot be adequately trained. E.g., if only 30% of the traffic signs in a scene are labeled, it would be "more difficult for the neural network to learn traffic signs since there are quite a lot of traffic signs among the negative samples."
Directly affects AI functions: 4 "Yes", 0 "No"
Challenge score: Survey 1 - 4.333 (Rank 1/31), Survey 2 - 3.750 (Rank 1/25*)

3.2 Data Management Challenges

Name: Data Acquisition
Reference: Kruse et al. (2016)
Description: Data can be acquired in many ways, such as real-world collection, simulations, and third-party purchases. 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 data collection, transmission, and storage 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 data owner and the need for necessary consent to use the data. E.g., data generated by a vehicle in operation is owned by the vehicle owner. Suppose the vehicle manufacturer wishes to use data for improving the deep learning models. In that case, it has to identify and execute the appropriate steps it needs to take to use that data legally.
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: Data collection, data labeling, simulations, and data management are expensive procedures. The cost increases, especially when tasks have to be done manually or by a specialized system.
Directly affects AI functions: 0 "Yes", 4 "No"
Challenge score: Survey 1 - Not Available (challenge added in second cycle only), Survey 2 - 1.925 (Rank 19/25)

Name: Imbalanced Dataset
Reference: Corrales et al. (2016)
Description: An imbalanced dataset or unevenly represented data leads to bias in the neural networks. If a dataset contains records that skew towards particular label(s), the training will be skewed toward those labels, thus leading to the possibility of biased decisions.
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: Suppose proper standards and methods are not followed when transmitting data between the sensors, from sensors to data storage, deep learning functions, and to-and-fro data storage and deep learning functions. In that case, it could lead to 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 time frame and cost. As the number 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. E.g. in ADAS, vehicles have to be driven in different situations and environments to gather data. Such a process 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. Manual labeling of data requires an adequately trained and enough workforce. E.g., manually labeling a car as a car and a truck as a truck in every frame is challenging.
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)
Description: This challenge causes a similar problem as that caused by an imbalanced dataset. If there is duplicate data, neural networks will be biased towards the more prevalent labels than those that are not, thus leading to biased decisions 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 comply. Tasks such as identifying the regulations and 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, refer *Data Package*)

Name: Reliance on Suppliers to Raise Error
Reference: Interviewee B
Description: Function developers rely on the sensors' suppliers to detect erroneous data and raise such errors. There are 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. Furthermore, it takes even more time if they are done manually. 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 cycle only), Survey 2 - 2.350 (Rank 13/25)

3.3 Data Source Challenges

Name: Data Dependent on External Conditions
Reference: Interviewee B
Description: Sometimes data could be affected by external conditions. E.g., during bad weather like heavy snow, heavy rain, sandstorms, and when it is very dark, the camera or the sensor would not be much of use, and the data (pictures and video streams) captured by the camera would be of low quality. Such inability of the camera or the sensor would affect the performance of deep learning algorithms if trained with the low-quality data gathered during such weather conditions.
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 a lack of variety data from different environmental sources across the globe in

which the vehicle has to be tested. E.g., if the car is only tested in Sweden, there might be no test data on the vehicle's 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 collected from various data sources can be of different types, thus increasing the difficulty in integrating and processing the data in a short time frame. An example source of new data types can be a camera that collects a large amount of data every single moment. Examples of such new data are color, different vehicle sizes, and traffic signals.
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, European Commission. Statistical Office of the European Union. (2020)
Description: Wrongly calibrated or defective sensors cause the value of the data collected to differ from the actual value. Using that data to train AI models will not be effective since the training is done with the wrong data. When the AI model in the future encounter actual data, it will not perform as it should as it was trained with the 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*)

3.4 Data Structure Challenges

Name Fragmented Data (this challenge 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 a solution to a different problem unrelated to the field of this study.)
Reference: International Organization for Standardization (2008)
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, which 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 adequately trained as they would be expecting data in a specific format. There might be a need for 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 the rest of the data is considered 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, it should be detected as an actual intrusion and not as an anomaly in data if it happens.
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 primarily work well with structured data. Unstructured data is useless, and training the algorithms with such data would lead to poor performance of the deep learning applications. 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)

3.5 Data Trust Challenges

Name: Incorrect Labeling
Reference: Interviewee C, Interviewee D, Interviewee E, Corrales et al. (2016)
Description: If a dataset contains incorrect labels, it will affect the training. Neural networks do not have a way to segregate data in terms of the correctness of labels. Incorrectly labeled data is useless as using such data could make the neural networks function in unintended manners. E.g., as an incorrect label, a traffic light can be labeled as a standard 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: A technique of data collection is simulation. 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 a 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: Noise refers to data that is irrelevant and meaningless. There are various kinds of noise that affect the performance of the deep learning algorithms, reducing the application's accuracy and predictive capability. Therefore it is essential to identify and handle the data noise. Data with a low signal-to-noise ratio is an example of 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 List of Data Quality Attributes

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

Table 2: 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), International Organization for Standardization (2008), European Commission. Statistical Office of the European Union. (2020), DQMatters.com (2017)	*The conditions and modalities by which users can access, use and interpret data.* (European Commission. Statistical Office of the European Union. 2020), *The extent to which data are available or easily and quickly retrievable.* (Wang & Strong 1996)	Data Acquisition (0.8, 0.66), Data Delay (0.5, 0.5), Data Dependent on External Conditions (1, 1), Data Drop (0.6, 0.5), Data Ownership (, 1), Manual Data Collection (0.2, 0.66)

Table 2: 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. (1998), Sidi et al. (2012), Wang & Strong (1996), International Organization for Standardization (2008)	*The degree to which data values correctly represents real-world entities.* (Data Management Association et al. 2017), *The extent to which data are correct, reliable, and certified free of error.* (Wang & Strong 1996), *Accuracy of data is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure.* (European Commission. Statistical Office of the European Union. 2020)	Data Dependent on External Conditions (0.6, 0), Data Drop (0.8, 1), Incomplete Data (1, 1), Incorrect Labeling (1, 1), Lack of Good Data from Simulations (0.8, 0.66), Low Labeled Data Volume (0.8, 1), Noise (1, 0.66), Outlier Data (0.4, 0.66), Redundant Data (0.4, 0.33)
Amount of Data	Bobrowski et al. (1998), Wang & Strong (1996)	*The number of facts stored.* (Bobrowski et al. 1998), *The extent to which the quantity or volume of available data is appropriate.* (Wang & Strong 1996)	NA
Appropriate Amount of Data	Sidi et al. (2012), Wang & Strong (1996)	*The extent to which the quantity or volume of available data is appropriate.* (Wang & Strong 1996)	NA
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

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Availability	Interviewees, Cai & Zhu (2015), Sidi et al. (2012), International Organization for Standardization (2008)	*The degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use.* (International Organization for Standardization 2008)	Data Acquisition (1, 0.66), Data Delay (0.25, 0.5), Data Drop (1, 0.75), Incomplete Data (0.6, 0.75), Low Labeled Data Volume (0.2, 0.25)
Believability / Credibility / Reputation	Sidi et al. (2012), Wang & Strong (1996), Cai & Zhu (2015), International Organization for Standardization (2008)	*The degree to which data has attributes that are regarded as true and believable by users in a specific context of use. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, commitments).* (International Organization for Standardization 2008), *The extent to which data are trusted or highly regarded in terms of their source or content.* (Wang & Strong 1996)	Incomplete Data (1, 1), Incorrect Labeling (1, 1), Lack of Good Data from Simulations (0.6, 1), Outlier Data (0.2, 1), Unstructured Data (, 0)
Clarity / Interpretability / Unambiguous	Bobrowski et al. (1998), Sidi et al. (2012), Wang & Strong (1996), European Commission. Statistical Office of the European Union. (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	European Commission. Statistical Office of the European Union. (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.* European Commission. Statistical Office of the European Union. (2020)	NA
Comment	European Commission. Statistical Office of the European Union. (2020)	*Supplementary descriptive text which can be attached to data or metadata.* (European Commission. Statistical Office of the European Union. 2020)	NA

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Completeness	Interviewees, Cai & Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012), Wang & Strong (1996)	*Refers to whether all required data is present.* (Data Management Association et al. 2017), *The extent to which data are of sufficient breadth, depth, and scope for the task at hand.* (Wang & Strong 1996)	Data Delay (0, 0.25), Data Drop (0.8, 1), Improper Data Transfer (0.6, 1), Incomplete Data (1, 1)
Compliance	International Organization for Standardization (2008)	*The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.* International Organization for Standardization (2008)	Data Ownership (, 1), Regulatory Compliance (, 1)
Conciseness / Concise Representation	Bobrowski et al. (1998), 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	European Commission. Statistical Office of the European Union. (2020), International Organization for Standardization (2008)	*A property of data indicating the extent to which their unauthorised disclosure could be prejudicial or harmful to the interest of the source or other relevant parties.* (European Commission. Statistical Office of the European Union. 2020) *The degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use.* (International Organization for Standardization 2008)	Data Ownership (, 0.66), Regulatory Compliance (, 0.66)

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Consistency / Uniformity	Cai & Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012), International Organization for Standardization (2008), Data Management Association et al. (2017), DQMatters.com (2017)	*Can refer to ensuring that data values are consistently represented within a dataset and between datasets, and consistently associated across datasets.* (Data Management Association et al. 2017), *Measures whether or not data is equivalent across systems or location of storage.* (DQMatters.com 2017)	Data Drop (0.8, 1), Improper Data Transfer (0.6, 0.66), Incompatible Data Formats (, 0.66), Incomplete Data (0.6, 1)
Consistency and Synchronization	Data Management Association et al. (2017), DQMatters.com (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	European Commission. Statistical Office of the European Union. (2020)	*Individual or organisational contact points for the data or metadata, including information on how to reach the contact points.* (European Commission. Statistical Office of the European Union. 2020)	Regulatory Compliance (, 0.66)
Correctness	Interviewees, Bobrowski et al. (1998)	*Every set of data stored represents a real world situation.* (Bobrowski et al. 1998)	Data Dependent on External Conditions (0.6, 0.33), Imbalanced Dataset (1, 0.66), Improper Data Transfer (0.6, 0.66), Incomplete Data (0.6,), Incorrect Labeling (1, 1), Low Labeled Data Volume (0.6, 0.75), Noise (0.6, 0.66), Outlier Data (0, 0.33)
Cost and Burden	European Commission. Statistical Office of the European Union. (2020)	*Cost associated with the collection and production of a statistical product and burden on respondents.* (European Commission. Statistical Office of the European Union. 2020)	see *Cost Effectiveness* for the challenges affecting this attribute

Table 2: 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), Data Management Association et al. (2017), DQMatters.com (2017), International Organization for Standardization (2008)	*The measure of whether data values are the most up-to-date version of the information.* (Data Management Association et al. 2017), *The degree to which data has attributes that are of the right age in a specific context of use.* (International Organization for Standardization 2008)	Data Delay (1, 0.75), Data Drop (0.4, 0.25), Improper Data Transfer (0.4, 1), Incomplete Data (0, 0.75)
Data Coverage	Sidi et al. (2012)	*A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest.* (Sidi et al. 2012)	NA
Data Decay	Sidi et al. (2012)	*A measure of the rate of negative change to data.* (Sidi et al. 2012)	NA
Data Revision	European Commission. Statistical Office of the European Union. (2020)	*Any change in a value of a statistic released to the public.* (European Commission. Statistical Office of the European Union. 2020)	NA
Data Specification	Sidi et al. (2012)	*A measure of the existence, completeness, quality and documentation of data standards, data models, business rules, meta data, and reference data.* (Sidi et al. 2012)	NA
Definition / Documentation	Cai & Zhu (2015)	*It consists of data specification, which includes data name, definition, ranges of valid values, standard formats, business rules, etc. Normative data definition improves the degree of data usage.* (Cai & Zhu 2015)	NA
Duplication	Sidi et al. (2012)	*A measure of unwanted duplication existing within or across systems for a particular field, record, or data set.* (Sidi et al. 2012)	NA

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Ease of Manipulation	Pipino et al. (2002), Sidi et al. (2012)	*The extent to which data is easy to manipulate and apply to different same format.* (Pipino et al. 2002)	NA
Ease of Operation	Wang & Strong (1996)	*The extent to which data are easily managed and manipulated (i.e., updated, moved, aggregated, reproduced, customized).* (Wang & Strong 1996)	Data Acquisition (0.4, 0.33), Data Ownership (, 0.66), Improper Data Transfer (0.8, 0.66), Manual Data Collection (0.6, 0.66), Manual Data Labeling (0.8,)
Ease of Use and Maintainability	Sidi et al. (2012)	*A measure of the degree to which data can be accessed and used and the degree to which data can be updated, maintained, and managed.* (McGilvray 2008), (Sidi et al. 2012)	NA
Effectiveness	Sidi et al. (2012)	*It is the capability of the function to enable users to achieve specified goals with accuracy and completeness in a specified context of use.* (Batini et al. 2009), (Sidi et al. 2012)	NA
Efficiency	Sidi et al. (2012), International Organization for Standardization (2008)	*The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.* (International Organization for Standardization 2008)	Data Delay (0.75, 0.5), Data Drop (0.6, 0.5), Imbalanced Dataset (0.2, 0), Incomplete Data (0.2, 0.75), Incorrect Labeling (0.2, 0.66), Outlier Data (0.2, 0.33), Unstructured Data (, 0.66)
Fitness	Cai & 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. 2002), (Sidi et al. 2012)	NA

Table 2: 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	European Commission. Statistical Office of the European Union. (2020)	*The time interval at which the statistics are disseminated over a given time period.* (European Commission. Statistical Office of the European Union. 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	European Commission. Statistical Office of the European Union. (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.* (European Commission. Statistical Office of the European Union. 2020)	Regulatory Compliance (, 1)
Integrity	Cai & Zhu (2015), Sidi et al. (2012), DQMatters.com (2017)	*Measures the structural or relational quality of datasets.* (DQMatters.com 2017)	NA
Integrity or Coherence	See *Integrity* and *Coherence*		
Latency	Data Management Association et al. (2017)	*The time between when the data was created and when it was made available for use.* (Data Management Association et al. 2017)	Data Delay (1, 1)
Learnability	Sidi et al. (2012)	*It means the capability of the function to enable to user to learn it.* (Heravizadeh et al. 2009), (Sidi et al. 2012)	NA
Lineage	DQMatters.com (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.* (DQMatters.com 2017)	Data Acquisition (1, 1), Data Ownership (, 0.66), Regulatory Compliance (, 0.66)

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
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	European Commission. Statistical Office of the European Union. (2020)	*The date on which the metadata element was inserted or modified in the database.* (European Commission. Statistical Office of the European Union. 2020)	NA
Navigation	Sidi et al. (2012)	*Extent to which data are easily found and linked to.* (Knight & Burn 2005), (Sidi et al. 2012)	NA
Objectivity	Bobrowski et al. (1998), Sidi et al. (2012), Wang & 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	International Organization for Standardization (2008)	*The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another (while) preserving the existing quality in a specific context of use.* (International Organization for Standardization 2008)	Data Delay (0, 0), Data Drop (0.2, 0.33), Improper Data Transfer (0.8, 0.66), Regulatory Compliance (, 0.66)
Precision	Bobrowski et al. (1998), International Organization for Standardization (2008)	*The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.* (International Organization for Standardization 2008)	NA

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
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	European Commission. Statistical Office of the European Union. (2020)	*Time lag between the actual delivery of the data and the target date when it should have been delivered.* (European Commission. Statistical Office of the European Union. 2020)	NA
Quality Management	European Commission. Statistical Office of the European Union. (2020)	*Systems and frameworks in place within an organisation to manage the quality of statistical products and processes.* (European Commission. Statistical Office of the European Union. 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	Data Management Association et al. (2017)	*Asks whether a data pattern meets expectations.* (Data Management Association et al. 2017)	Data Drop (0.4, 0.5), Incomplete Data (0.8, 0.5)
Recoverability	International Organization for Standardization (2008)	*The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.* (International Organization for Standardization 2008)	NA
Reference Period	European Commission. Statistical Office of the European Union. (2020)	*The period of time or point in time to which the measured observation is intended to refer.* (European Commission. Statistical Office of the European Union. 2020)	NA
Release Policy	European Commission. Statistical Office of the European Union. (2020)	*Rules for disseminating statistical data to all interested parties.* (European Commission. Statistical Office of the European Union. 2020)	Regulatory Compliance (, 0)

Table 2: 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. (1998), Sidi et al. (2012), Wang & Strong (1996), European Commission. Statistical Office of the European Union. (2020)	*The extent to which data are applicable and helpful for the task at hand.* (Wang & Strong 1996), *The degree to which statistical information meet current and potential needs of the users.* (European Commission. Statistical Office of the European Union. 2020)	New Data Type (, 0.33)
Reliability	Cai & Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012)	*Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value.* (European Commission. Statistical Office of the European Union. 2020)	Data Drop (0.8, 1), Improper Data Transfer (0.8, 0.66), Incomplete Data (0.8, 1), Incorrect Labeling (1, 1)
Representation	DQMatters.com (2017)	*Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).* (DQMatters.com 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. 2009), (Sidi et al. 2012)	NA
Security	Sidi et al. (2012)	*Extent to which access to information is restricted appropriately to maintain its security.* (Wang & Strong 1996), (Sidi et al. 2012)	NA
Statistical Presentation	European Commission. Statistical Office of the European Union. (2020)	*Description of the disseminated data which can be displayed to users as tables, graphs or maps.* (European Commission. Statistical Office of the European Union. 2020)	NA

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Statistical Processing	European Commission. Statistical Office of the European Union. (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 (European Commission. Statistical Office of the European Union. 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. (1998), Sidi et al. (2012), Wang & Strong (1996), Data Management Association et al. (2017), DQMatters.com (2017)	*Length of time between data availability and the event or phenomenon the data describe.* (European Commission. Statistical Office of the European Union. 2020), *The extent to which the age of the data is appropriate for the task at hand.* (Wang & Strong 1996)	Data Delay (1, 0.75), Data Drop (0.6, 0.25), Manual Data Collection (0.2, 0.66), Manual Data Labeling (, 0.8)
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), International Organization for Standardization (2008)	*The extent to which data are well documented, verifiable, and easily attributed to a source.* (Wang & Strong 1996), *The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.* (International Organization for Standardization 2008)	Data Acquisition (0.8, 1), Data Ownership (, 0.66), Regulatory Compliance (, 0.66)

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
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. (1998)	*Each piece of data has a unique meaning.* (Bobrowski et al. 1998)	NA
Understandability / Ease of Understanding	Sidi et al. (2012), Wang & Strong (1996), International Organization for Standardization (2008)	*The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in (an) appropriate languages, symbols and units in a specific context of use.* (International Organization for Standardization 2008)	Incomplete Data (0.8, 0.75)
Uniqueness	Data Management Association et al. (2017)	*No entity exists more than once within the dataset.* (Data Management Association et al. 2017)	Redundant Data (, 1)
Unit of Measure	European Commission. Statistical Office of the European Union. (2020)	*The unit in which the data values are measured.* (European Commission. Statistical Office of the European Union. 2020)	NA
Usability	Interviewees, Cai & Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012)	*Extent to which information is clear and easily used.* (Sidi et al. 2012)	Imbalanced Dataset (1, 1), Incomplete Data (1, 0.5), Incorrect Labeling (1, 1), Low Labeled Data Volume (1, 0.75), Redundant Data (0.6, 0.33), Unstructured Data (, 0.33)
Usefulness	Sidi et al. (2012)	*Extent to which information is applicable and helpful for the task at hand* (Wang & Strong 1996), (Sidi et al. 2012)	Data Delay (0.5, 0.75), Data Drop (0.4, 0.75), Imbalanced Dataset (1, 1), Incomplete Data (0.8, 0.5), Incorrect Labeling (1, 1), Low Labeled Data Volume (1, 0.75), Lack of Good Data from Simulations (0.8, 0.66), Noise (0.6, 0.66), Redundant Data (0.6,)
Validity	Data Management Association et al. (2017), DQMatters.com (2017)	*Refers to whether data values are consistent with a defined domain of values.* (Data Management Association et al. 2017)	Incorrect Labeling (1, 1), Incompatible Data Format (, 0.66), Low Labeled Data Volume (1, 0.75), Unstructured Data (, 0.3)

Table 2: List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
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
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	Data Management Association et al. (2017)	*Remain current for a short period.* (Data Management Association et al. 2017)	NA

Table 3: List of Data Quality Attribute Metrics

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

Table 3: List of Data Quality Attribute Metrics

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

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

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

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

35. Representation
36. Resiliency
37. Safety
38. Scalability
39. Security
40. Statistical Presentation
41. Statistical Processing
42. Structure
43. Traceability
44. Unambiguous
45. Understandability / Ease of Understanding
46. Unit of Measure
47. Usability
48. Validity
49. Value Added

5 Solution Candidates

Table 4: List of Solution Candidates and Mitigated Challenges

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

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, the requirement mentioned above specification should be satisfied.
- Check if there is a field for auto-increasing sequential numbers in the dataset.
- If there exists a field for auto-increasing sequential number, check if each record in the dataset has a value in the auto-increasing sequential number field.
- If the above condition is true, check if the auto-increasing sequential number field is in sequence and no values are missing.
- If each of the above conditions is not satisfied, notify the appropriate stakeholders.
- Finally, if the auto-increasing sequential number field is in sequence and no values are missing, start processing the data.

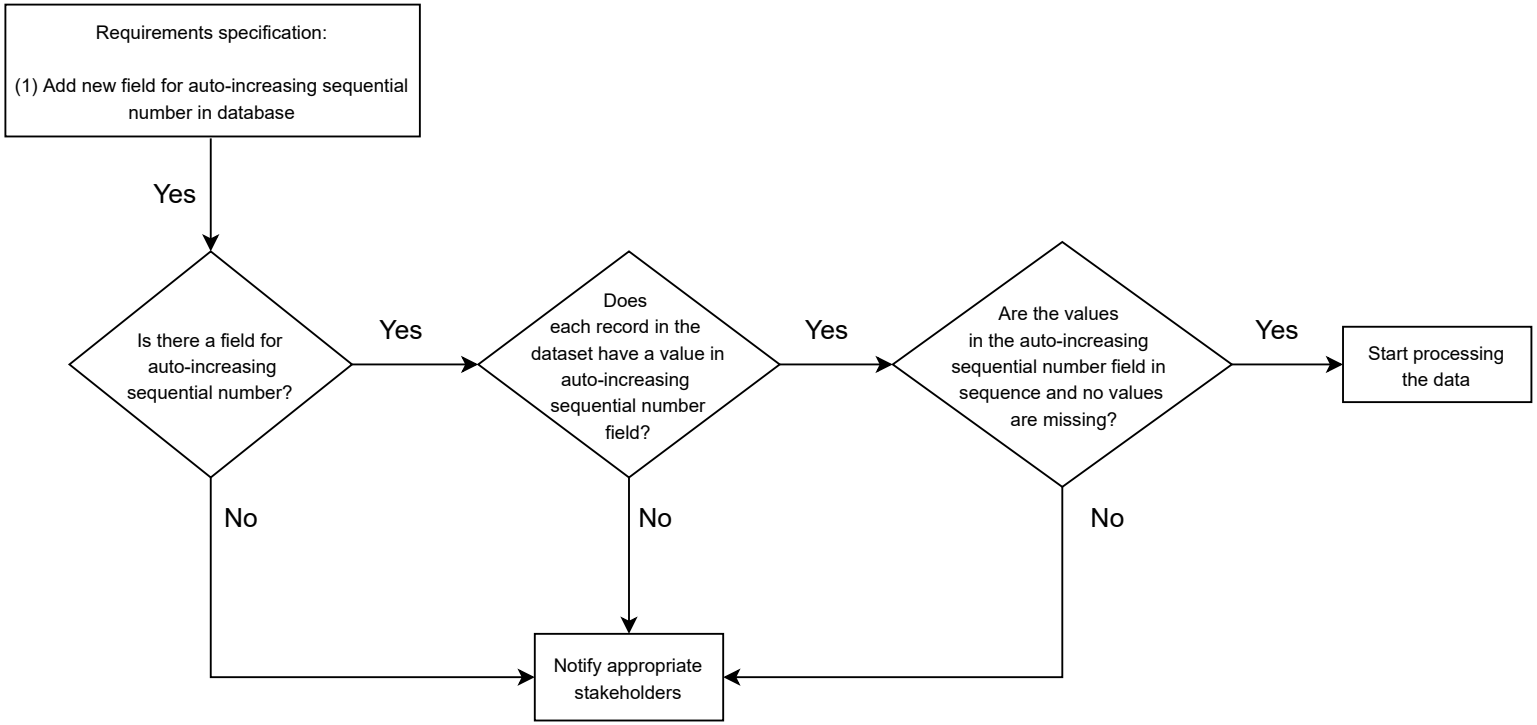


Figure 3: Flowchart for *Auto Increasing Sequential Number* Solution

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 a bounding box around each object manually. In this solution, the labeler needs to label only a few frames instead of labeling each object in each frame. Then, the pixels related to 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 must be manually labeled. Again, keep tracking their pixel until they are visible in the succeeding frames.
- Cruz-Sandoval et al. (2019) describes two approaches to semi-automated data labeling in the healthcare sector.
  - First, they propose using gesture recognition with smartwatches to label data, which is done using sensors such as an accelerometer and gyroscope in a smartwatch. This approach works by recognizing a limited set of pre-determined "discrete" gestures. They use the 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. This experiment determined that "the system can reduce the burden of online self-labeling by using automatic gesture recognition."
  - Secondly, they study intelligent microphones to label "audible home activities." They use an Intelligent System for Sound Annotation (ISSA) that includes a sound detector, an audio classifier, and voice assistant functions to label audio. It uses already-trained models for audio classification. In an experiment to test this approach, they collected ten samples of 2.5 to 5.2 seconds in a range of 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 the ImageNet dataset for object detection and MS COCO dataset for image captioning. They state that a "faster and more accurate convolution" can be formulated by dividing an image into multiple grids. 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.

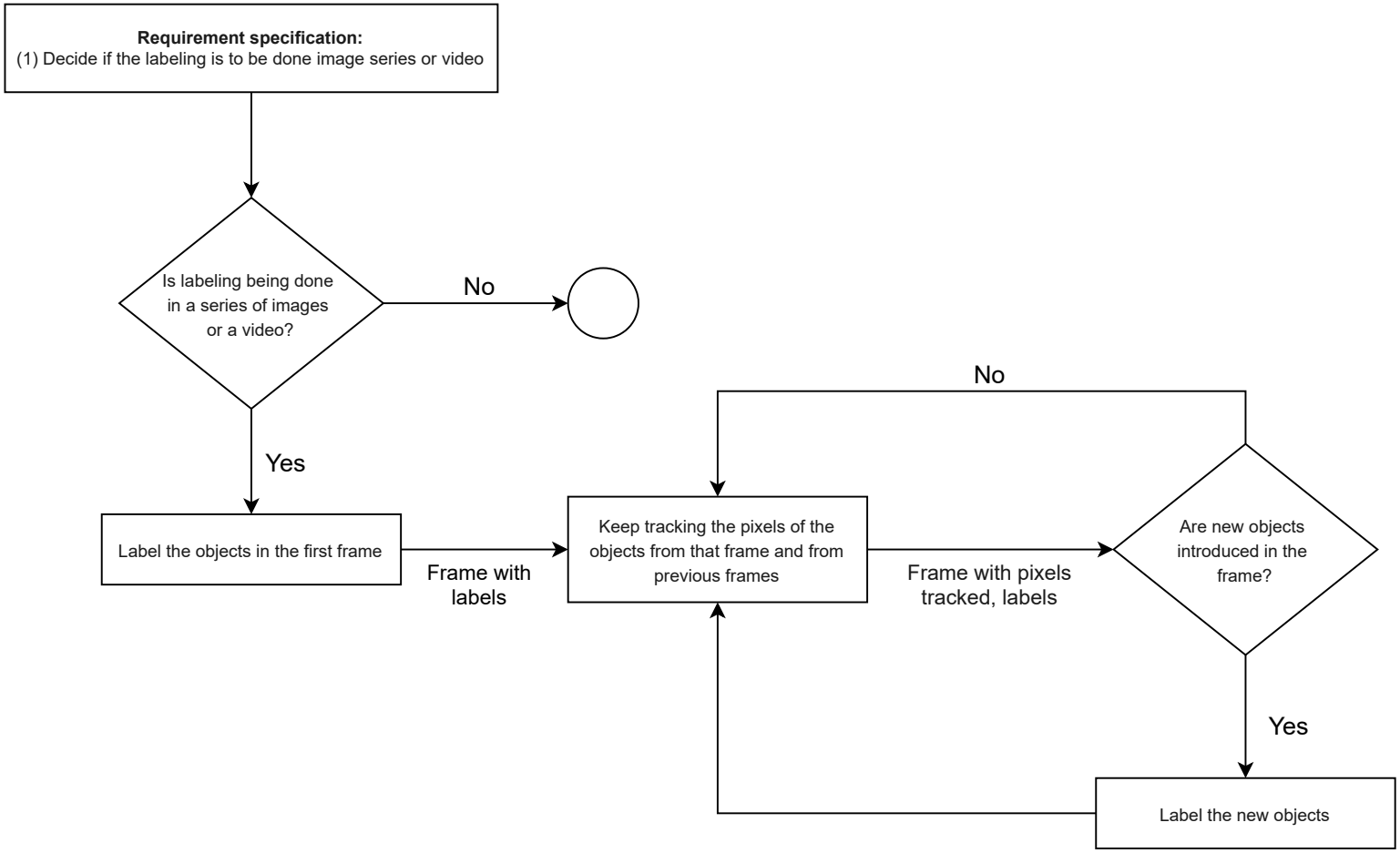


Figure 4: Flowchart for *Automated Labeling* Solution

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

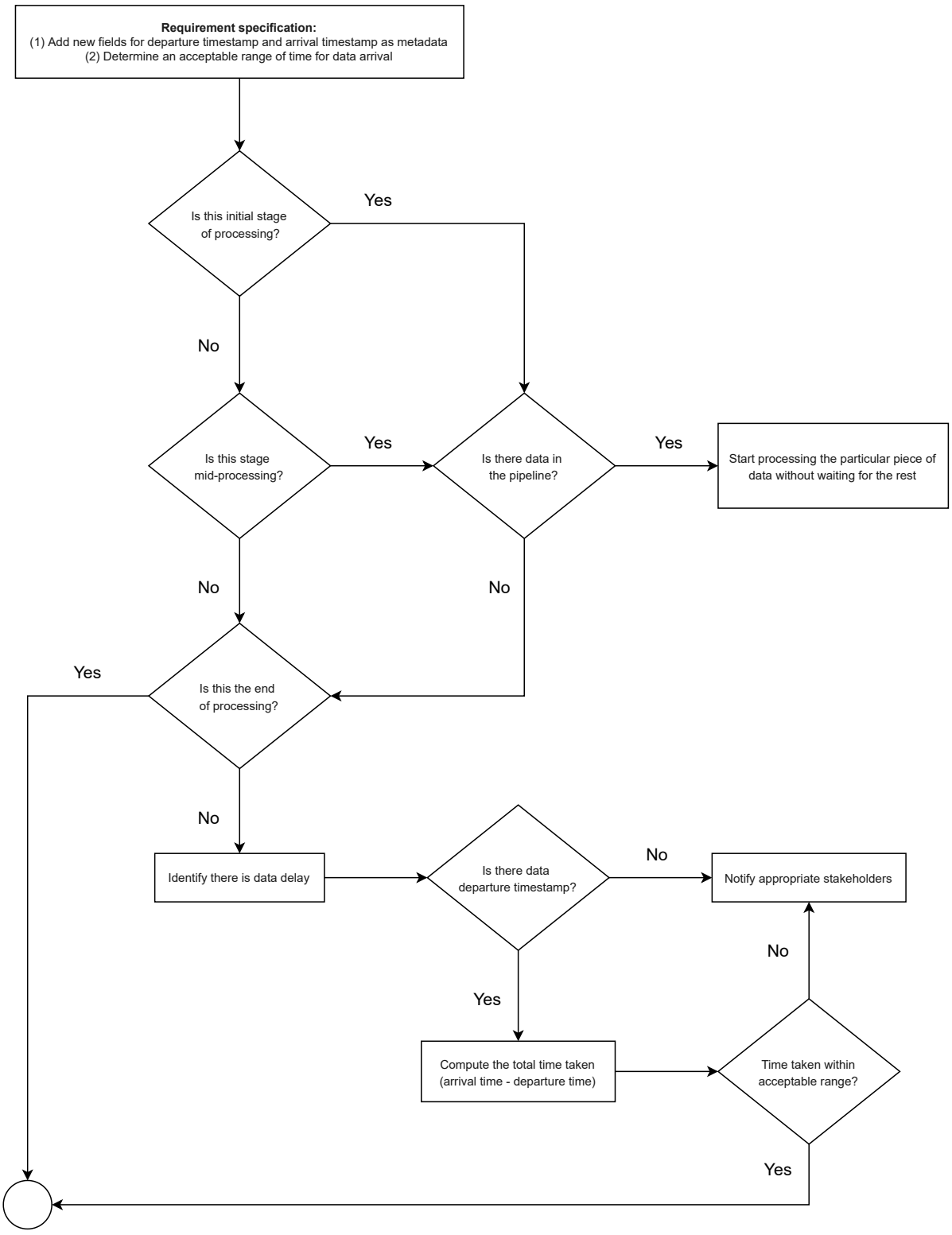


Figure 5: Flowchart for *Continuous Data Processing* Solution

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 the 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 data repository does not exist, check if the system behaves correctly.
  - 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.

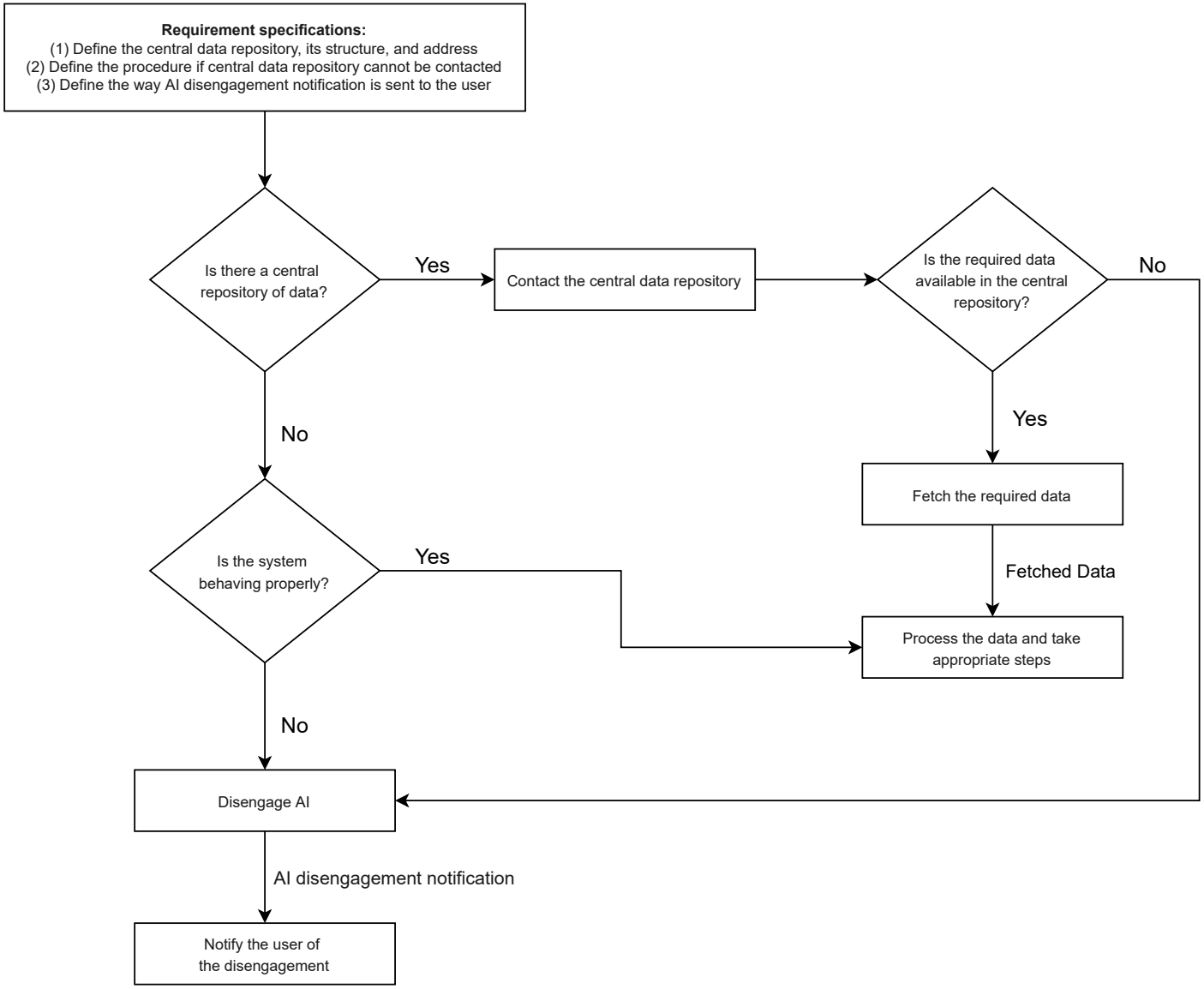


Figure 6: Flowchart for *Corroboration of Data with Central Data Repository* Solution

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

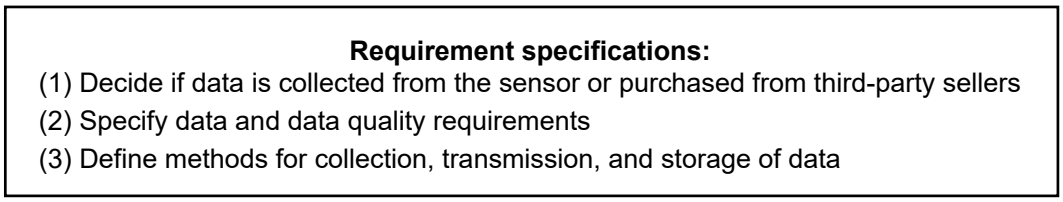


Figure 7: Requirement Specifications for *Data Acquisition* Solution Task

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 the filters should check,
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 from third-party sellers.
- If the data is collected through sensors, send it through the filters assigned for sensor-collected data.
- If the data comes from third-party sensors, send it through the filters assigned for purchased data.
- Check if there is an error in the filtration process after the data goes through either of the filters,
- If there is an error in the filtration process, notify appropriate stakeholders. Else, start processing the remaining data.

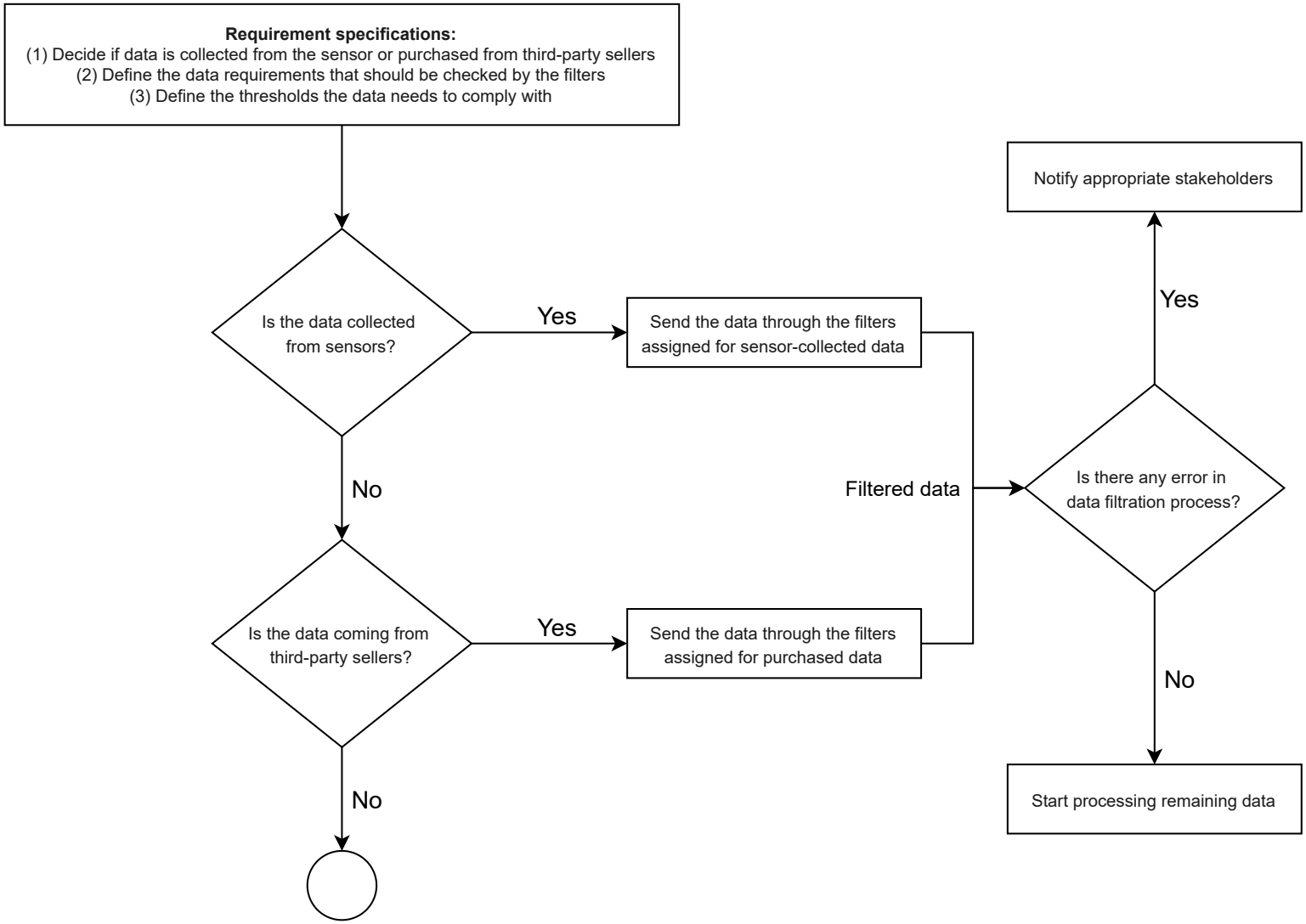


Figure 8: Flowchart for *Data Filter* Solution

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. (2006), in their article, discuss *Data level* methods and *Algorithm level* methods to handle imbalanced dataset.

  Data level methods have various sampling techniques to handle imbalanced datasets. These techniques are

  1. *Undersampling* randomly eliminates the 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 comprise three techniques to handle imbalanced datasets. They are

  1. In *Threshold method*, a score is produced by neural networks that determine 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 do not belong to the same class of the training set (Noumir et al. 2012).
  3. *Cost-sensitive learning* is an approach to incorporating cost in decision making to improve classifier performance, which is done to define fixed and unequal misclassification costs between classes.

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, the requirement specifications mentioned above should be completed, which is done to identify mandatory fields for the 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. 2014).
  - In listwise deletion, the record with missing data is deleted entirely. The remaining data is used for analysis. 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 methods 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). 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 9: Corrected Item Mean Substitution (CIM)

Here, PM is the participant's mean, and IM is the 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 the last observed carried forward method, the last observed value replaces a missing value in the particular field from the same object (Kang 2013). E.g., suppose a value of 50.0 is measured in the previous observation. The value is recorded as the value of the current observation if the current observation value is missing.
  - "Plausible" values calculated from an imputation model replace the missing values in the multiple imputation method. The replacement is done multiple times. Then, a statistical analysis of interest is performed in each dataset independently. Then, a "single MI estimate" is derived by combining independent estimates (Hayati Rezvan et al. 2015).
- Cox et al. (2014) uses several 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 consider the "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.

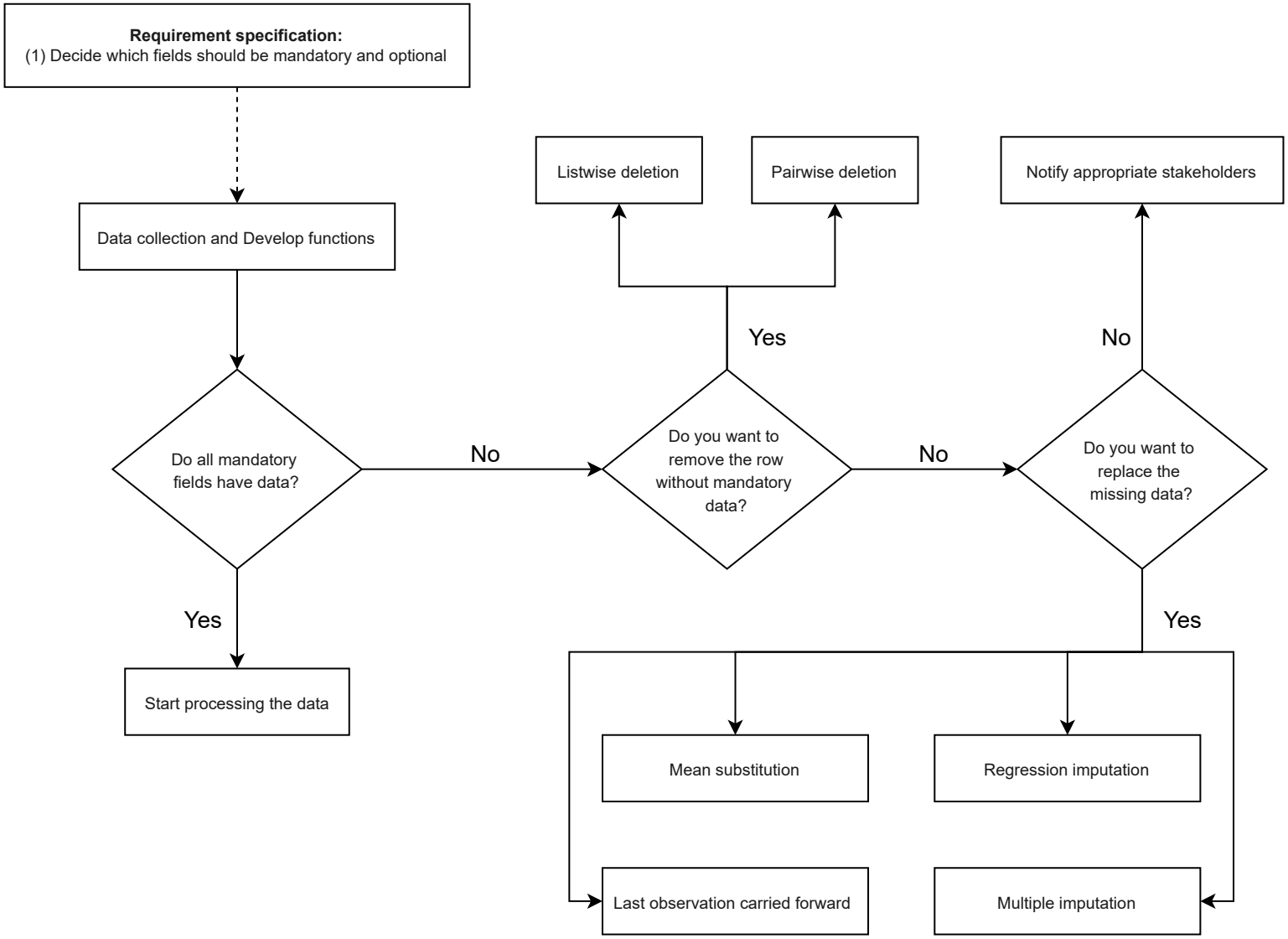


Figure 10: Flowchart for *Identify Mandatory and Optional Fields* Solution

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

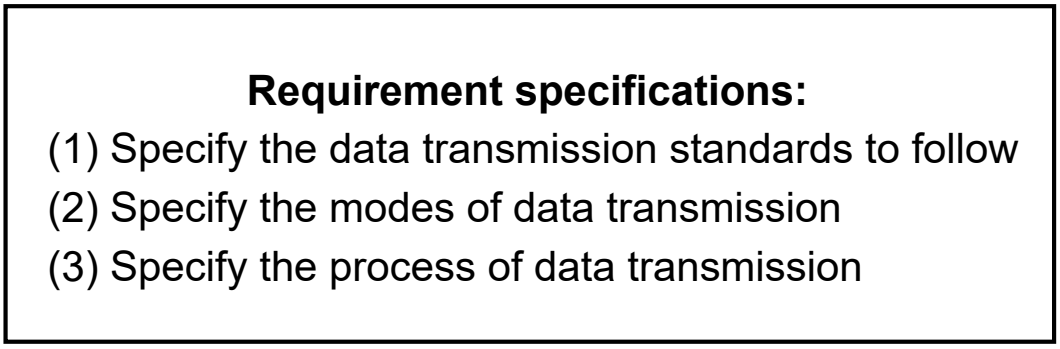


Figure 11: Requirement Specifications for *Improper Data Transfer* Solution Task

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).
  - 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 several 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).

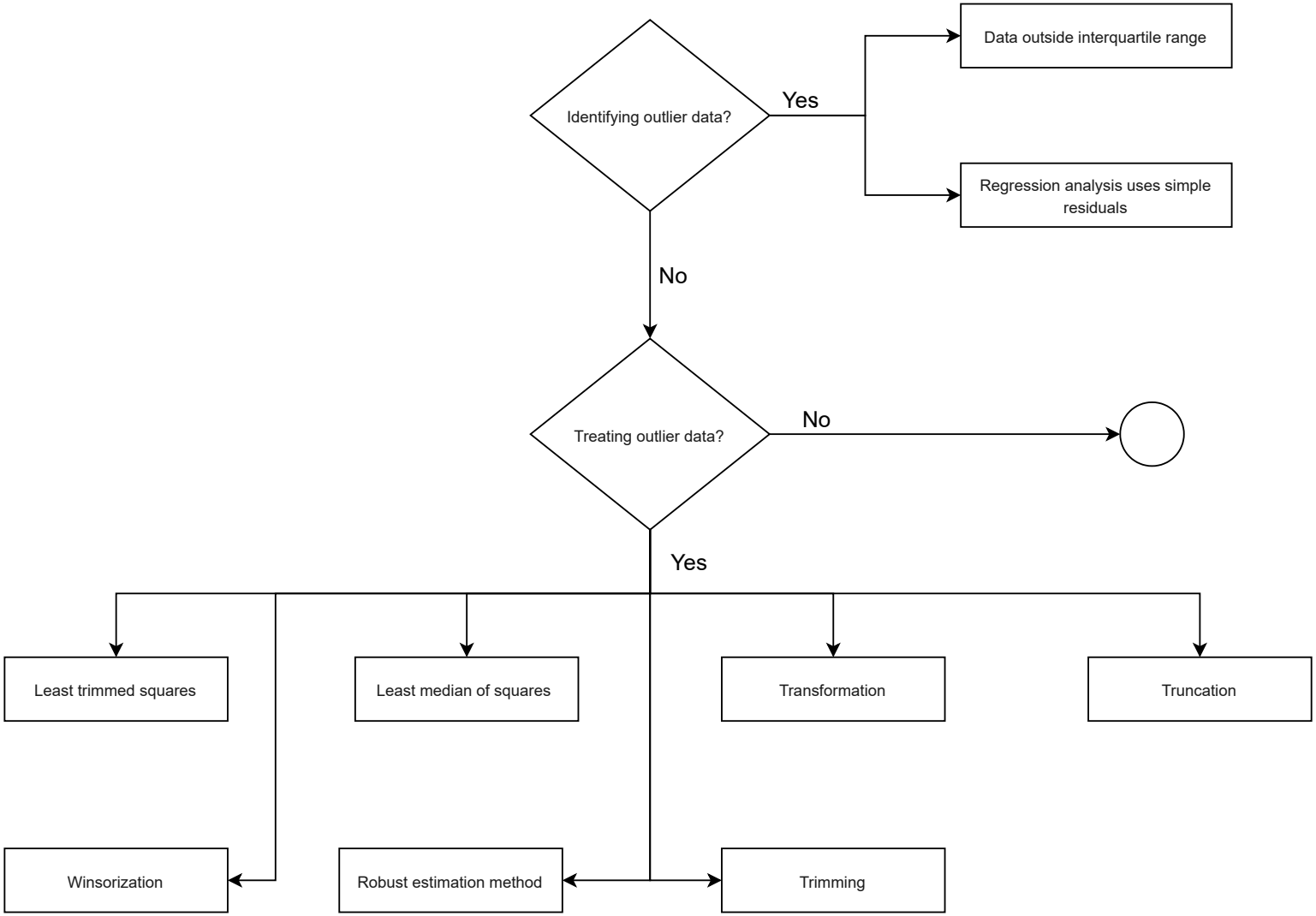


Figure 12: Flowchart for *Outlier Techniques* Solution

5.11 Pair-wise Attribute Algorithm

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

Implementation Details:

- To handle noisy data, firstly, the 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 their article discuss that either ignore the noisy data, filter the noisy data, or alter the noisy data. Removing the erroneous attribute values from the data set is called filtering. Alteration of noisy data is also called polishing or data scrubbing, or relabeling.

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. The second is to erase the redundant variables. Finally, the graphs are generated by variable recognition and eliminating the redundancies in the later stage.

5.13 Test Environments

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

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

Implementation Details:

- Check if the data should be collected from a real-world 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 be collected from desert, highway, mountain, rural, snowy, and urban locations.
- In terms of temporal context, data can be collected during the daytime, dusk, and nighttime.
- Vehicular data using different types of vehicles 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 weather, data can be collected in weather situations like cloudy, drizzle, fog, hurricane, overcast, partly cloudy, rainy, sandstorm, snowy, stormy, and sunny.

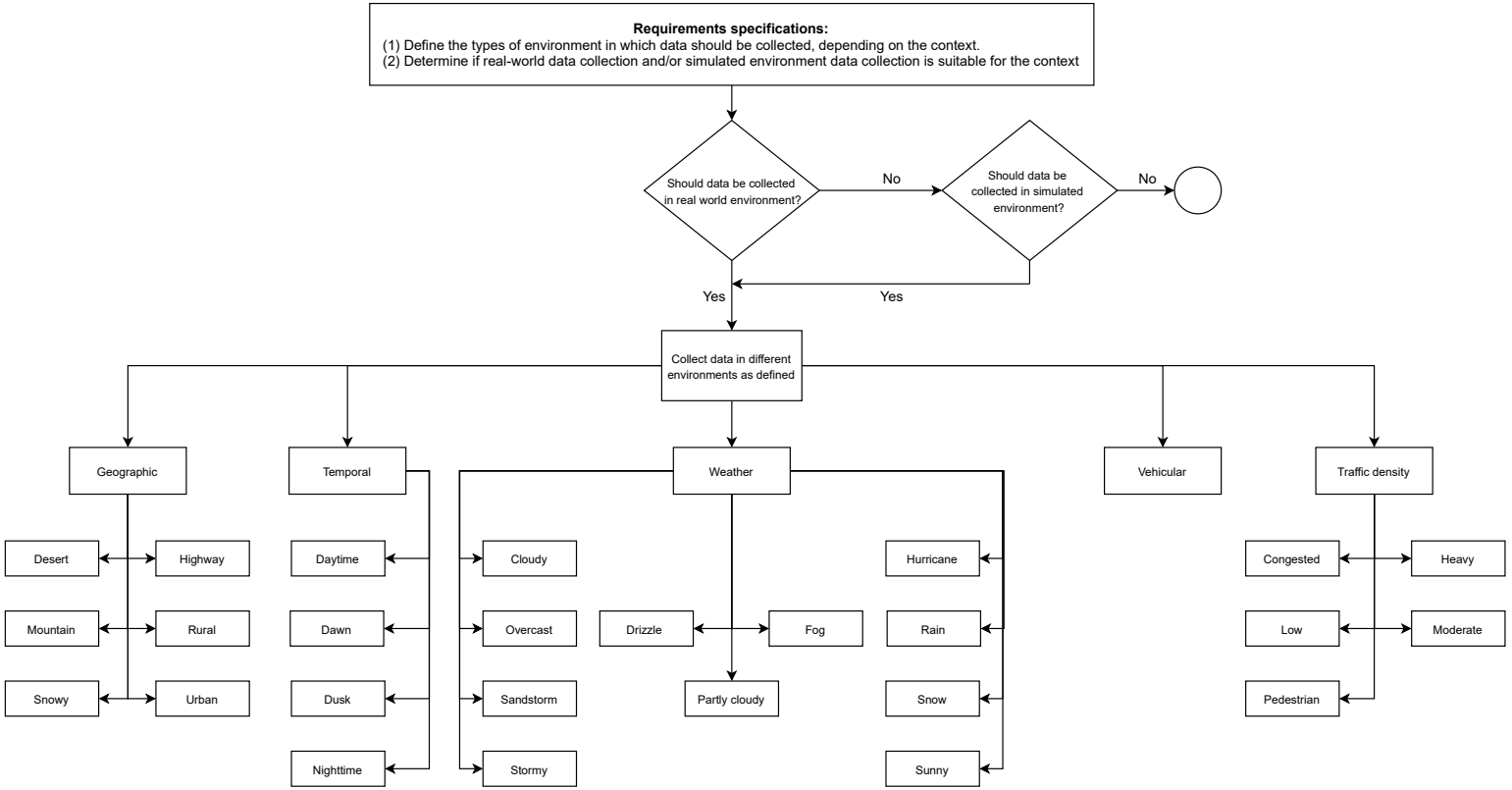


Figure 13: Flowchart for *Test Environments* Solution

References

Anscombe, F. J. (1960), 'Rejection of Outliers', *Technometrics* **2**(2), 123–146. https://doi.org/10.1080/00401706.1960.10489888.

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. (2009), 'Methodologies for data quality assessment and improvement', *ACM Computing Surveys* **41**(3), 16:1–16:52. https://doi.org/10.1145/1541880.1541883.

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

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. 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**(0), 2. https://doi.org/10.5334/dsj-2015-002.

Chhabra, G., Vashisht, V. & Ranjan, J. (2019), 'A Review on Missing Data Value Estimation Using Imputation Algorithm', *Journal of Advanced Research in Dynamic and Control Systems* **11**(07-Special Issue), 312–318. https://www.jardcs.org/abstract.php?id=1782.

Corrales, D. C., Ledezma, A. I. & Corrales, J. C. (2016), 'A systematic review of data quality issues in knowledge discovery tasks', *Revista Ingenierías Universidad de Medellín* **15**(28), 125–149. https://doi.org/10.22395/rium.v15n28a7.

Cox, B. E., McIntosh, K., Reason, R. D. & Terenzini, P. T. (2014), 'Working with Missing Data in Higher Education Research: A Primer and Real-World Example', *The Review of Higher Education* **37**(3), 377–402. https://doi.org/10.1353/rhe.2014.0026.

Cruz-Sandoval, D., Beltran-Marquez, J., Garcia-Constantino, M., Gonzalez-Jasso, L. A., Favela, J., Lopez-Nava, I. H., Cleland, I., Ennis, A., Hernandez-Cruz, N., Rafferty, J., Synnott, J. & Nugent, C. (2019), 'Semi-Automated Data Labeling for Activity Recognition in Pervasive Healthcare', *Sensors* **19**(14), 3035. https://doi.org/10.3390/s19143035.

Data Management Association, Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E. & Smith, E. (2017), *DAMA-DMBOK: data management body of knowledge*, Technics Publications, LLC, Denville, NJ, United States. OCLC: 1048611774.

DeCastro-García, N., Muñoz Castañeda, Á. L., Fernández Rodríguez, M. & Carriegos, M. V. (2018), 'On Detecting and Removing Superficial Redundancy in Vector Databases', *Mathematical Problems in Engineering* **2018**, e3702808. https://doi.org/10.1155/2018/3702808.

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

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

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

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. https://doi.org/10.1109/SOSE.2016.63.

Gentleman, J. F. & Wilk, M. B. (1975), 'Detecting Outliers. II. Supplementing the Direct Analysis of Residuals', *Biometrics* **31**(2), 387–410. https://doi.org/10.2307/2529428.

Gupta, S. & Gupta, A. (2019), 'Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review', *Procedia Computer Science* **161**, 466–474. https://doi.org/10.1016/j.procs.2019.11.146.

Hayati Rezvan, P., 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). https://doi.org/10.1186/s12874-015-0022-1.

Heravizadeh, M., Mendling, J. & Rosemann, M. (2009), Dimensions of Business Processes Quality (QoBP), *in* D. Ardagna, M. Mecella & J. Yang, eds, 'Business Process Management Workshops', Springer, Berlin, Heidelberg, pp. 80–91. https://doi.org/10.1007/978-3-642-00328-8_8.

Houari, R., Bounceur, A., Tari, A. K. & Kecha, M. T. (2014), Handling Missing Data Problems with Sampling Methods, *in* '2014 International Conference on Advanced Networking Distributed Systems and Applications', pp. 99–104. https://doi.org/10.1109/INDS.2014.25.

Huisman, M. (2000), 'Imputation of Missing Item Responses: Some Simple Techniques', *Quality and Quantity* **34**(4), 331–351. https://doi.org/10.1023/A:1004782230065.

International Organization for Standardization (2008), ISO/IEC 25012:2008, Technical report, International Organization for Standardization, Geneva, Switzerland. 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). 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. https://doi.org/10.1109/CCCS.2015.7374131.

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

Knight, S.-A. & Burn, J. (2005), 'Developing a Framework for Assessing Information Quality on the World Wide Web', *Informing Science: The International Journal of an Emerging Transdiscipline* **8**, 159–172. https://doi.org/10.28945/493.

Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. (2006), 'Handling imbalanced datasets: A review', *GESTS International Transactions on Computer Science and Engineering* **30**, 13.

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), e5359. https://doi.org/10.2196/medinform.5359.

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

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

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. https://doi.org/10.1016/j.procs.2019.01.118.

Noumir, Z., Honeine, P. & Richard, C. (2012), On simple one-class classification methods, *in* '2012 IEEE International Symposium on Information Theory Proceedings', pp. 2022–2026. https://doi.org/10.1109/ISIT.2012.6283685.

Osborne, J. & Overbay, A. (2004), 'The power of outliers (and why researchers should ALWAYS check for them)', *Practical Assessment, Research, and Evaluation* **9**(6). https://doi.org/10.7275/qf69-7k43.

Peralta, V. (2006), Data Quality Evaluation in Data Integration Systems, phdthesis, Université de Versailles-Saint Quentin en Yvelines ; Université de la République d'Uruguay. https://tel.archives-ouvertes.fr/tel-00325139/.

Pipino, L. L., Lee, Y. W. & Wang, R. Y. (2002), 'Data quality assessment', *Communications of the ACM* **45**(4), 211–218. https://doi.org/10.1145/505248.506010.

Rousseeuw, P. J. & Leroy, A. M. (1987), Outlier Diagnostics, *in* 'Robust Regression and Outlier Detection', John Wiley & Sons, Ltd, pp. 216–247. https://doi.org/10.1002/0471725382.ch6.

Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H. & Mustapha, A. (2012), Data quality: A survey of data quality dimensions, *in* '2012 International Conference on Information Retrieval Knowledge Management', pp. 300–304. https://doi.org/10.1109/InfRKM.2012.6204995.

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. https://www.jstor.org/stable/40398176.