ETSI TR 104 180 V1.1.1 (2026-05)

TECHNICAL REPORT

Data Solutions (DATA);
Development and identification of Data Quality Metrics

Reference
DTR/DATA-00104180

Keywords
AI, data, quality, trust

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI IPR online database.

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Data Solutions (DATA).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Executive summary

The present document establishes a standardized framework for data quality assessment to ensure reliability in digital and AI ecosystems. It defines 18 Data Quality Metrics, providing formal definitions and mathematical formulas for quantitative measurement across intrinsic and contextual dimensions.

Through empirical case studies on Industrial IoT and Demographic data, the present document validates the applicability of these metrics while demonstrating that effective assessment relies heavily on domain-specific contexts and metadata availability.

Based on these findings, the present document recommends two key future standardization activities:

- Transitioning the definitions and descriptive examples into a normative TS to ensure interoperability.
- Developing standardized Test Methodologies and Test Formulas for automated and reproducible verification of data.

1 Scope

The present document introduces quality metrics for data for future standardization. It focuses on identifying key properties of data quality including trustworthiness (e.g. completeness, accuracy, bias, reliability, redundancy, source reliability and integrity) and formulating standardized methods to assess these properties.

The present document refers to developed standard specifications related to data quality and collaborates with standards bodies & groups working on data, security and AI trustworthiness (e.g. ETSI TC SAI, JTC 21).

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long-term validity.

The following referenced documents may be useful in implementing an ETSI deliverable or add to the reader's understanding, but are not required for conformance to the present document.

[i.1] ISO/IEC 25012:2008: "Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model".

[i.2] Hitzler P, Zaveri A, Rula A, et al.: "Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework". Semantic Web. 2015;7(1):63-93. doi:10.3233/SW-150175.

[i.3] Recommendation ITU-T E.800: "Definitions of terms related to quality of service".

[i.4] ISO/IEC 23988:2007: "Information technology — A code of practice for the use of information technology (IT) in the delivery of assessments".

[i.5] ISO/IEC 25010:2023: "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model".

[i.6] ISO/IEC 16350:2015: "Information technology — Systems and software engineering — Application management".

[i.7] Recommendation ITU-T Y.3605: "Big data - Reference architecture".

[i.8] IEEE 610.12-1990™: "IEEE Standard Glossary of Software Engineering Terminology".

[i.9] ISO 19157-1:2023: "Geographic information — Data quality — Part 1: General requirements".

[i.10] IEEE 100™: "The Authoritative Dictionary of IEEE Standards Terms".

[i.11] Recommendation ITU-T X.1252: "Baseline identity management terms and definitions".

[i.12] ISO 8000-61:2016: "Data quality — Part 61: Data quality management: Process reference model".

[i.13] Recommendation ITU-T Y.3600: "Big data – Cloud computing based requirements and capabilities".

[i.14] ISO/IEC 11179-1: "Information technology — Metadata registries (MDR) — Part 1: Framework".

[i.15] ISO/IEC 5259-4: "Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework".

[i.16] Recommendation ITU-T X.1601: "Security framework for cloud computing".

[i.17] NIST SP 800-53: "Security and Privacy Controls for Information Systems and Organizations".

[i.18] ISO/IEC 20889:2018: "Privacy enhancing data de-identification terminology and classification of techniques".

[i.19] ISO/IEC 29100:2024: "Information technology — Security techniques — Privacy framework".

[i.20] ISO/IEC TR 24027:2021: "Information technology — Artificial Intelligence (AI) — Bias in AI Systems and AI aided decision making".

[i.21] ISO/IEC 27001:2022: "Information security, cybersecurity and privacy protection — Information security management systems — Requirements".

[i.22] IEEE 7003:2024™: "IEEE Standard for Algorithmic Bias Considerations".

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AI	Artificial Intelligence
BRCR	Business Rule Consistency Rate
CCR	Calculation Consistency Rate
CDE	Critical Data Element
CS	Consistency Score
CV	Coefficient of Variation
CVR	Consistency Violation Rate
DC	Duplicate Count
DIR	Disparate Impact Ratio
DRR	Duplicate Record Rate
DSR	Data Synchronization Rate
FCR	Format Consistency Rate
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
IIoT	Industrial Internet of Things
IoT	Internet of Things
LCR	Logical Consistency Rate
MAE	Mean Absolute Error
MDAR	Master Data Alignment Rate
ML	Machine Learning
MSE	Mean Squared Error
NIST	National Institute of Standards and Technology
PII	Personal Identifiable Information

PIPA	Personal Information Protection Act
POS	Point of Sale
RAID	Redundant Array of Independent Disks
RICR	Referential Integrity Consistency Rate
RMSE	Root Mean Squared Error
RRR	Record Redundancy Ratio
SCE	Source Consistency Error
SLA	Service Level Agreement
TCR	Timeliness Compliance Rate
TS	Temporal Stability
US	Uniqueness Score

4 General Principles of Data Quality Metrics

This clause describes fundamental principles applicable to the definition and assessment of data quality metrics within the scope of the present document. These principles guide the development and application of data quality metrics to ensure they are meaningful, consistent, and actionable across different data contexts:

- Purpose: Metrics are defined and interpreted according to specific data usage scenarios, business needs or regulatory requirements. Relevance is context-dependent.
- Measurable: Every metric is accompanied by a clear, repeatable measurement approach, preferably with a quantitative formula or method to ensure objective assessment.
- Scalable: Metrics are applicable at various levels of granularity (attribute, record, dataset) and scalable to large or high-velocity data environments.
- Interpretable: Results are understandable to both technical and non-technical stakeholders, and indicate not only a score but also potential corrective actions.
- Metadata-aware: Many metrics rely on external information such as business rules, data definitions, and lineage. Dependencies on metadata are explicitly acknowledged.
- Ethical: Metrics related to bias, privacy and fairness need to be considered in the present document.
- Evolvable: The set of metrics and their definitions are better to be periodically reviewed and refined based on practical experience, technological changes, and evolving standards.
- AI/ML Training dataset: Metrics for data used in AI and ML can be considered with additional dimensions such as label quality, representational bias, as these directly impact model performance and fairness.

5 Data Quality Metrics

5.1 Completeness

5.1.1 Definition

Completeness refers to the degree to which all required or expected data values are present and available for a given purpose. It quantifies the absence of missing information, ensuring that a dataset contains all the necessary attributes and records to adequately represent the phenomena it describes or to fulfil the requirements of a specific task (e.g. model training, validation, or inference). In the context of a feature set for an AI/ML model, completeness ensures that all relevant features for each observation are populated. For a time-series dataset, it pertains to the presence of data points at expected intervals.

- ISO/IEC 25012:2008 [i.1]: Completeness refers to the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.
- Hitzler P et al. [i.2]: Completeness refers to the degree to which all required information is present in a particular dataset.

5.1.2 Measurement Formula

The completeness of a dataset can be measured at various granularities (e.g. attribute-level, record-level, or dataset-level). A common approach to quantify completeness for an attribute or a set of required attributes within a dataset is as follows:

- Let D be a dataset with N records (rows) and M attributes (columns).
- Let A_j be the j^{th} attribute in the dataset, where $j \in \{1,2,...,M\}$.
- Let I_{ij} be an indicator variable such that:
  - I_{ij}=1 if the value for record i and attribute j is present (not null/missing).
  - I_{ij}=0 if the value for record i and attribute j is missing (null/empty).

Attribute Completeness (C_A): The completeness of a specific attribute A_j is given by the ratio of present values for that attribute to the total number of records:

$$C_{Aj} = \frac{\sum_{i=1}^{N} I_{ij}}{N}$$

Record Completeness (C_R): The completeness of a specific record R_i (considering a set of K required attributes within that record) is given by the ratio of present values for the required attributes in that record to the total number of required attributes:

$$C_{R_i} = \frac{\sum_{j \in Required\_Attributes} I_{ij}}{K}$$

Dataset Completeness (C_D): The overall completeness of the dataset, considering all N records and M attributes (or a subset of M_{req} required attributes), can be defined as the average attribute completeness or the total proportion of non-missing values:

$$C_D = \frac{\sum_{j=1}^{M_{req}} \sum_{i=1}^{N} I_{ij}}{N \times M_{req}}$$

Where M_{req} is the number of attributes deemed "required" for the dataset's intended use.

5.1.3 Example

Consider a dataset collected for a predictive maintenance model for industrial machinery, specifically focusing on pump performance. The dataset aims to predict potential failures based on operational parameters. Assume the following attributes are deemed required for each operational log record: `Timestamp`, `PumpID`, `Temperature_C`, `Vibration_mm_s`, `Pressure_kPa`.

A dataset snippet consisting of five operational log records for a single pump (`PumpID: P001`) is presented below.

Table 5.1.3-1: Example of a dataset for illustrating the Completeness property

Index	Timestamp	PumpID	Temperature_C	Vibration_mm_s	Pressure_kPa
1	2024-06-25 10:00:00	P001	45,2	1,5	500
2	2024-06-25 10:05:00	P001	45,8	1,7	505
3	2024-06-25 10:10:00	P001	46,1	NaN	510
4	2024-06-25 10:15:00	P001	NaN	1,6	508
5	2024-06-25 10:20:00	P001	47,0	1,8	NaN

Here, N=5 records and M_{req}=5 required attributes.

Attribute Completeness Example:

1) Completeness of `Vibration_mm_s`:
   - Missing in Record 3
   - Present in 4 out of 5 records
   - $C_{Vibration\_mm\_s}$ = (4/5) × 100 % = 80 %
2) Completeness of `Temperature_C`:
   - Missing in Record 4
   - Present in 4 out of 5 records
   - $C_{Temperature\_C}$ = (4/5) × 100 % = 80 %

Record Completeness Example:

1) Completeness of Record 3 (Required attributes: 5):
   - Vibration_mm_s is missing
   - Present values: 4 out of 5 records
   - C_{R3} = (4/5) × 100 % = 80 %
2) Completeness of Record 4 (Required attributes: 5):
   - Temperature_C: is missing
   - Present values: 4 out of 5 records
   - C_{R4} = (4/5) × 100 % = 80 %

Overall Dataset Completeness Example (overall for required attributes):

- Total expected values ($N \times M_{req}$) = 5 × 5 = 25
- Total missing values: 3 (one in `Temperature_C`, one in `Vibration_mm_s` and one in `Pressure_kPa`)
- Total present values = 25 - 3 = 22
- C_D = 22/25 × 100 % = 88 %

5.2 Reliability

5.2.1 Definition

Reliability refers to the degree to which data can be trusted to be consistent, stable, and free from unexpected variations over time and across different sources. It measures the dependability of the data generation and collection processes, ensuring that the data produced is accurate, reproducible, and fit for its intended use under stated conditions. In the context of AI/ML systems, reliable data ensures that models are trained and evaluated on a consistent foundation, reducing the risk of performance degradation due to data drift or source inconsistencies.

- Recommendation ITU-T E.800 [i.3]: Reliability refers to the probability that an item can perform a required function under stated conditions for a given time interval.
- ISO/IEC 23988:2007 [i.4]: Consistency with which an assessment measures.
- ISO/IEC 25010:2023 [i.5]: Capability of a product to perform specified functions under specified conditions for a specified period of time without interruptions and failures.
- ISO/IEC 16350:2015 [i.6]: Degree to which an object or an object's services provide agreed or expected functionality during a defined time period under specified conditions.

5.2.2 Measurement Formula

Reliability can be assessed through various lenses, such as source consistency, temporal stability, and reproducibility. A common approach involves measuring the variance or error rate from an expected baseline or across multiple sources.

Let *S* be a data source producing a stream of values for a specific attribute over time.

Let X_i be the value of the attribute at time *i*.

Let E_i be the expected or reference value for the attribute at time *i* (which could be a known ground truth, a value from a trusted source, or a predicted value from a stable model).

Source Consistency Error (SCE) for a single source over *N* observations can be measured as the Mean Absolute Error (MAE).

$$SCE_{MAE} = \frac{1}{N}\sum_{i=1}^{N}|X_i - E_i|$$

Temporal Stability (TS) can be measured for a source by calculating the Coefficient of Variation (CV) for a specific attribute over a defined time window, indicating how much the data fluctuates relative to its average value:

$$TS_{CV} = \frac{\sigma\chi}{\mu\chi} \times 100\ \%$$

Where:

- $\sigma\chi$ is the standard deviation of the attribute values over the time period.
- $\mu\chi$ is the mean of the attribute values over the time period.

A lower CV indicates higher temporal stability.

5.2.3 Example

Consider two sensors (Sensor A and Sensor B) measuring the temperature of a critical machine component every hour. The expected temperature, based on a calibrated laboratory standard, is also recorded.

Table 5.2.3-1: Example of a dataset for illustrating the Reliability property

Index	Time	Expected Temp (°C)	Sensor A (°C)	Sensor B (°C)
1	10:00	75,0	75,2	74,8
2	11:00	75,0	76,1	75,0
3	12:00	76,0	76,0	75,5
4	13:00	76,0	77,0	76,0
5	14:00	75,5	75,4	75,8

Source Consistency for Sensor A:

$$\boldsymbol{SCE_{MAE}} = \frac{1}{5}(|75{,}2 - 75{,}0| + |76{,}1 - 75{,}0| + |76{,}0 - 76{,}0| + |77{,}0 - 76{,}0| + |75{,}4 - 75{,}5|) = 0{,}48\ (°C)$$

Temporal Stability for Sensor A:

- Mean $\mu\chi$ = (75,2 + 76,1 + 76,0 + 77,0 + 75,4) / 5 = 75,94 °C
- Std Dev $\sigma\chi$ = 0,75 °C
- TS_{CV} = (0,75/75,94) × 100 % = 0,99 %

5.3 Redundancy

5.3.1 Definition

Redundancy refers to the degree to which data is duplicated or unnecessarily repeated within a dataset or across a system. While some redundancy can be intentional for fault tolerance (e.g. backup systems, RAID storage), excessive or uncontrolled redundancy in operational datasets can lead to inefficiencies in storage, processing, and analysis. It can also introduce inconsistencies if redundant copies are not properly synchronized and managed. This metric helps identify opportunities for data deduplication and normalization.

- Recommendation ITU-T Y.3605 [i.7]: Data redundancy refers to the repeated occurrence of the same data in the system.
- ISO 8000-61 [i.12]: Describes redundancy as "the presence of duplicate data elements that can be eliminated without loss of information".
- ISO/IEC 11179-1 [i.14]: Defines data redundancy as "the representation of data more than once within a specified scope".

5.3.2 Measurement Formula

Redundancy is often measured as the proportion of duplicate records or data elements within a defined context.

Let *D* be a dataset with *N* total records.

Let N_{unique} be the number of unique records in *D*, determined by a defined set of key attributes.

Record Redundancy Ratio (RRR) is the proportion of the dataset that consists of duplicate records:

$$RRR = \frac{N - N_{unique}}{N} \times 100\ \%$$

Alternatively, Duplicate Count (DC) is the absolute number of duplicate records:

$$DC = N - N_{unique}$$

For attribute-level redundancy, the focus is on the storage of derivable data. For example, storing both "Date of Birth" and "Age" is redundant as one can be derived from the other. This is often identified through schema analysis and data lineage rather than a single universal formula.

5.3.3 Example

Consider a customer contact list dataset where records are considered duplicates if they have the same EmailAddress.

Table 5.3.3-1: Example of a dataset for illustrating the Redundancy property

RecordID	Name	EmailAddress	PhoneNumber
1	John Doe	John.doe@email.com	555-0100
2	John Doe	John.doe@email.com	555-0101
3	Jane Smith	Jane.smith@email.com	555-0200
4	Alice Johnson	alice@email.com	555-0300
5	John Doe	John.doe@email.com	555-0100

- Total Records (*N*): 5
- Unique Records (N_{unique}, based on EmailAddress): 3 (john.doe@email.com, jane.smith@email.com, alice@email.com)
- Record Redundancy Ratio (RRR): $\frac{(5-3)}{5} \times 100\ \% = 40\ \%$
- Duplicate Count (DC): 5-3 = 2 records (RecordID 2 and 5 are duplicates of RecordID 1, though with slight variations in PhoneNumber for RecordID 2).

5.4 Accuracy

5.4.1 Definition

Accuracy refers to the degree to which data matches the true, real-world values or states it is intended to represent. It quantifies the absence of errors and deviations, ensuring that data values are correct and faithful to the actual entities or events they model.

- Recommendation ITU-T Y.3605 [i.7]: Data redundancy refers to the repeated occurrence of the same data in the system.
- ISO 19157-1:2023 [i.9]: closeness of agreement between a test result or measurement result and the true value.
- ISO/IEC 25012:2008 [i.1]: The degree to which data has attributes that correctly represent the true value of the intended attributes of a concept or event in a specific context of use.
- IEEE 610.12-1990 [i.8]:
  1) A qualitative assessment of correctness, or freedom from error.
  2) A quantitative measure of the magnitude of error.

5.4.2 Measurement Formula

The accuracy of a dataset or a specific attribute can be measured using the following common approaches:

- Let N be the total number of values being assessed.
- Let $N_{correct}$ be the number of values that match the true, authoritative reference values.

Error Rate: The proportion of values that are incorrect:

$$Error\ Rate = \frac{N - N_{correct}}{N} \times 100\ \%$$

Accuracy Rate: The proportion of values that are correct:

$$Accuracy\ Rate = \frac{N_{correct}}{N} \times 100\ \%$$

For numerical data, the magnitude of error can be quantified using these statistical measures:

- Let $V_{\{data,i\}}$ be a data value.
- Let $V_{\{true,i\}}$ be the corresponding true value.

Mean Absolute Error (MAE): The average magnitude of the errors:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |V_{data,i} - V_{true,i}|$$

Mean Squared Error (MSE): The average of the squares of the errors (more sensitive to large errors):

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (V_{data,i} - V_{true,i})^2$$

Root Mean Squared Error (RMSE): The square root of the MSE, in the same units as the original data:

$$RMSE = \sqrt{MSE}$$

5.4.3 Example

Scenario: Product Database in an E-commerce System

An e-commerce company is auditing the accuracy of product weights in its database. The true values are established by weighing a sample of products from the warehouse using a calibrated scale.

An auditor randomly selects 10 products from the database and records their listed weight. The products are then physically weighed.

Table 5.4.3-1: Example of a dataset for illustrating the Accuracy property

ProductID	Listed Weight (kg)	True Weight (kg)	Is Accurate?	Absolute Error (kg)
P-1001	1,5	1,5	Yes	0,0
P-1002	2,3	2,1	No	0,2
P-1003	0,5	0,5	Yes	0,0
P-1004	8,7	8,7	Yes	0,0
P-1005	1,2	1,0	No	0,2
P-1006	15,0	15,5	No	0,5
P-1007	3,4	3,4	Yes	0,0
P-1008	0,8	0,8	Yes	0,0
P-1009	5,5	5,0	No	0,5
P-1010	2,0	2,0	Yes	0,0

Calculation of Accuracy and Error Rates:

- Total values assessed N: 10
- Number of correct values $N_{correct}$: 6
- Accuracy Rate: $^{6}/_{10} \times 100\,\% = 60\,\%$
- Error Rate: $^{4}/_{10} \times 100\,\% = 40\,\%$

Calculation of Numerical Error Metrics:

- Mean Absolute Error (MAE):

  Sum of Absolute Errors = $0{,}2 + 0{,}2 + 0{,}5 + 0{,}5 = 1{,}4\ kg$

  $MAE = {}^{1{,}4}/_{10} = 0{,}14\ kg$

- Root Mean Squared Error (RMSE):

  Sum of Squared Error = $0{,}2^2 + 0{,}2^2 + 0{,}5^2 + 0{,}5^2 = 0{,}04 + 0{,}04 + 0{,}25 + 0{,}25 = 0{,}58$

$$MSE = {}^{0,58}/_{10} = 0,058$$

$$RMSE = \sqrt{0,058} \approx 0,24\ kg$$

5.5 Timeliness

5.5.1 Definition

Timeliness refers to the degree to which data is available for use within the required timeframe or by a specified deadline. It measures the delay between the occurrence of a real-world event and the point when the corresponding data becomes accessible and usable in the system. Timely data is critical for operational efficiency, real-time decision-making, and maintaining the relevance of information. Data that is accurate and complete but delivered too late may lose its value and utility, potentially leading to missed opportunities or ineffective responses.

- ISO/IEC 25012:2008 [i.1]: Defines timeliness as "the degree to which data is sufficiently up-to-date for the task at hand" and "the extent to which the age of the data is appropriate for the intended use".
- IEEE 100 [i.10]: Defines timeliness as "the degree to which data is current and available when needed".
- ISO 8000-61 [i.12]: Describes timeliness as "the availability of data at the time required for the intended use".
- Recommendation ITU-T Y.3600 [i.13]: Defines timeliness in big data contexts as "the degree to which data is delivered within the time constraints required by the application".

5.5.2 Measurement Formula

Timeliness can be quantified by measuring the time delay between key events in the data lifecycle and comparing it against a defined Service Level Agreement (SLA) or business requirement.

Let $T_{available}$ be the timestamp when data becomes available in the target system.

Let T_{event} be the timestamp when the real-world event occurred or when the data was initially generated.

Data Latency (L): The absolute time delay for a single data point or record:

$$L = T_{available} - T_{event}$$

Average Latency (L_{avg}): The mean latency across N records over a specific period:

$$L_{avg} = {}^{1}/_{N} \sum_{i=1}^{N} L_i$$

Timeliness Compliance Rate (TCR): The proportion of data records that meet the specified latency SLA:

- Let T_{event} be the number of records where $L \leq SLA_{\text{threshold}}$.

$$\text{TCR} = \frac{N_{\text{on-time}}}{N} \times 100\ \%$$

Data Freshness (F): For datasets that require periodic updates, freshness measures the time elapsed since the last update relative to the expected update frequency:

- Let T_{current} be the current time.
- Let $T_{\text{last\_update}}$ be the timestamp of the most recent update to the dataset.
- Let FI be the expected update frequency interval (e.g. 24 hours).

$$F = \frac{T_{\text{current}} - T_{\text{last\_update}}}{FI}$$

A value of $F \leq 1$ indicates the data is fresh according to its update schedule.

5.5.3 Example

Scenario: Financial Transaction Monitoring System

A bank is expected to monitor transactions for fraudulent activity in near real-time. The compliance policy requires that 99 % of transactions are required to be available for analysis within 2 minutes of the transaction being authorized at the point of sale.

Data: SLA Threshold ($L \leq SLA_{\text{threshold}}$): 2 minutes (120 seconds)

Sample of 5 transaction records:

Table 5.5.3-1: En example of a dataset for illustrating the Timeliness property

Transaction ID	Event Time (Authorization)	Available Time (in Monitoring System)	Latency (seconds)	Meet SLA
TX-1001	2025-07-25 10:00:05 CET	2025-07-25 10:00:55 CET	50	Yes
TX-1002	2025-07-25 10:01:10 CET	2025-07-25 10:02:45 CET	95	Yes
TX-1003	2025-07-25 10:02:00 CET	2025-07-25 10:04:15 CET	135	No
TX-1004	2025-07-25 10:03:22 CET	2025-07-25 10:03:55 CET	33	Yes
TX-1005	2025-07-25 10:04:30 CET	2025-07-25 10:07:10 CET	160	No

Calculation of Timeliness Metrics:

1) Average Latency (L_{avg}):

$$L_{avg} = \frac{50+95+135+33+160}{5} = \frac{473}{5} = 94{,}6 \text{ seconds}$$

2) Timeliness Compliance Rate (TCR):

   - Total records N: 5
   - Records meeting SLA (N_{on_time}): 3 (TX-1001, TX-1002, TX-1004)
   - $TCR = \frac{N_{\text{on-time}}}{N} \times 100\,\% = \frac{3}{5} \times 100\,\% = 60\,\%$

Usage and Impact:

- Fraud Detection: Transactions TX-1003 and TX-1005 were available for analysis with delays of 135 and 160 seconds, respectively. A fraudster could have completed multiple fraudulent transactions in this time window before the system flagged the activity, leading to potential financial loss.
- Regulatory Compliance: The bank is failing its internal compliance policy of 99 % TCR. This could result in regulatory fines and reputational damage.

5.6 Consistency

5.6.1 Definition

Consistency refers to the degree to which data is uniform, coherent, and free from contradictions across different datasets, systems, or within the same dataset over time. It ensures that data follows the same rules, formats, and definitions wherever it appears. Inconsistent data can lead to conflicting reports, erroneous analytics, and flawed decision-making as different parts of an organization may be using different versions of the "truth".

Consistency can be measured at three distinct levels:

- Intra-dataset Consistency: Within a single dataset (format, logic, calculation rules)
- Inter-dataset Consistency: Across multiple datasets within the same system (referential integrity, business rules)

- Cross-system Consistency: Between different systems or applications (data synchronization, master data alignment)

- ISO/IEC 25012:2008 [i.1]: Defines consistency as "the degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It includes the absence of difference when comparing two or more representations of a thing against a definition".
- IEEE 100 [i.10]: Describes data consistency as "the degree to which a set of data satisfies stated constraints on the relationships between data values".
- Recommendation ITU-T X.1252 [i.11]: Addresses consistency in the context of identity management, emphasizing "the extent to which identity information is presented in the same format and is coherent across different systems and applications".
- ISO 8000-61 [i.12]: Defines data consistency as "the property of data that is coherent and either does not conflict with other data or has any conflicts resolved".

5.6.2 Measurement Formula

Consistency can be measured at different levels using appropriate violation detection methods.

Let *N* be the total number of records, data elements, or constraints being checked.

Overall Consistency Violation Rate (CVR):

CVR = (Number of consistency violations detected) / (Total number of checks performed) × 100 %

Overall Consistency Score (CS):

CS = 100 % - CVR

Level-specific Consistency Measures:

1) Intra-Dataset Consistency:
   - Format Consistency Rate (FCR):

     FCR = (Number of values adhering to defined format) / (Total number of values checked) × 100 %
   - Logical Consistency Rate (LCR):

     LCR = (Number of records satisfying logical rules) / (Total records checked) × 100 %
   - Calculation Consistency Rate (CCR):

     CCR = (Number of correct calculations) / (Total calculations checked) × 100 %
2) Inter-Dataset Consistency:
   - Referential Integrity Consistency Rate (RICR):

     RICR = (Number of valid foreign key relationships) / (Total foreign key relationships) × 100 %
   - Business Rule Consistency Rate (BRCR):

     BRCR = (Number of business rule compliant records) / (Total records checked) × 100 %
3) Cross-System Consistency:
   - Data Synchronization Rate (DSR):

     DSR = (Number of matching values across systems) / (Total values compared) × 100 %
   - Master Data Alignment Rate (MDAR):

MDAR = (Number of aligned master data records) / (Total master data records) × 100 %

5.6.3 Example

Scenario: Multi-level Consistency Assessment in Retail Operations

a) Intra-Dataset Consistency Check (Single Table)

Product table

Table 5.6.3-1: Example of a dataset for illustrating the intra-dataset consistency check

Production ID	ProductName	Category	Price (Euro)	DiscountPrice	StockQuantity	Notes
P-001	Laptop ABC	Electronics	€ 999	€ 899	50	
P-002	Mouse XYZ	Electronics	€ 25	€ 30	100	Discount violation
P-003	Keyboard Pro	Electronics	€ 75	€ 65	75	
P-004	Monitor Ultra	ELECTRONICS	€ 300	€ 280	25	Format inconsistency

Intra-Dataset Consistency Results:

- Format Consistency: 3/4 = 75 % (Category format inconsistency)
- Logical Consistency: 3/4 = 75 % (Discount price > Regular price)
- Calculation Consistency: 4/4 = 100 % (All calculations valid)
- Overall Intra-Dataset Consistency = (75 % + 75 % + 100 %) / 3 = 83,3 %

b) Inter-Dataset Consistency Check (Within Same Dataset)

Table 5.6.3-2: Example of a orders table dataset for illustrating the inter-dataset consistency check

OrderID	CustomerID	ProductID	OrderDate	Quantity	Notes
ORD-101	CUST-001	P-001	2025-01-15	2	
ORD-102	CUST-002	P-005	2025-01-16	1	Referential
ORD-103	CUST-003	P-002	2025-01-17	5	

Table 5.6.3-3: Example of a customer table dataset for illustrating the intra-dataset consistency check

CustomerID	CustomerName	CustomerType	Notes
CUST-001	John Smith	Premium	
CUST-002	Jane Doe	Standard	
CUST-004	Mike Brown	Premium	Orphan record

Inter-Dataset Consistency Results:

- Referential Integrity: 2/3 = 66,7 % (One invalid product reference)
- Orphan Record Rate: 1/3 = 33,3 % (One orphan customer record)
- Overall Inter-Dataset Consistency = (66,7 % + 66,7 %) / 2 = 66,7 %

c) Cross-System Consistency Check (Between Different Datasets)

POS system vs Warehouse System Price Comparison

Table 5.6.3-4: Example of a dataset for illustrating the cross-system consistency check

ProductID	POS_Price	WMS_Price	Status
P-001	€ 999	€ 999	Consistent
P-002	€ 25	€ 28	Inconsistent
P-003	€ 75	€ 75	Consistent
P-004	€ 300	€ 300	Consistent

Cross-System Consistency Results:

- Data Synchronization Rate: 3/4 = 75 %
- Overall Cross-System Consistency = 75 %

Comprehensive Consistency Assessment:

- Total checks performed: 8 different consistency metrics
- Weighted Overall Consistency Score = (83,3 % + 66,7 % + 75 %) / 3 = 75 %

5.7 Integrity

5.7.1 Definition

Integrity refers to the accuracy, consistency, and reliability of data throughout its lifecycle. It ensures that data remains unaltered, complete, and trustworthy, free from corruption, unauthorized changes, or errors. In the context of a dataset, integrity encompasses aspects like completeness (no missing values), accuracy (values are correct and valid), consistency (data adheres to defined rules or formats), and timeliness (data is up-to-date). High data integrity means the dataset can be relied upon for analysis, decision-making, or machine learning without introducing biases or errors.

- ISO/IEC 25012:2008 [i.1]: Defines integrity as "the degree to which data is complete and accurate and has not been modified except in an authorized manner".
- ISO/IEC 5259-4 [i.15]: Specifies data integrity as "the property of data that maintains complete and accurate lineage from source to consumption".

5.7.2 Measurement Formula

A common way to quantify data integrity in a dataset is through an Integrity Score, which can be calculated as a composite metric based on key dimensions such as completeness, accuracy, and consistency. One simplified formula is:

$$Integrity\ Score = \left(\frac{Completeness + Accuracy + Consistency}{3}\right) \times 100\ \%$$

Where:

- $Completeness\ Score = \frac{Number\ of\ non-missing\ values}{Total\ expected\ values}$
- $Accuracy\ Score = \frac{Number\ of\ valid\ values\ (e.g.,within\ expected\ ranges\ or\ formats)}{Total\ values}$
- $Consistency\ Score = \frac{Number\ of\ records\ adhering\ to\ business\ rules\ (e.g.,no\ duplicates,referential\ integrity)}{Total\ records}$

This formula assumes equal weighting for each dimension; in practice, weights can be adjusted based on domain-specific needs. Scores range from 0 % (poor integrity) to 100 % (perfect integrity).

5.7.3 Example

Consider a simple dataset representing customer records in a CSV-like format. The dataset has 10 records with columns:

- CustomerID (unique integer);
- Name (string);
- Age (integer, ranging between 18-100);
- Email (valid email format); and
- PurchaseDate (date in YYYY-MM-DD format).

Table 5.7.3-1: Example of a dataset for illustrating the Integrity property

CustomerID	Name	Age	Email	PurchaseDate
1	John Doe	35	john@example.com	2023-05-15
2	Jane Smith		jane@example.com	2023-06-20
3	Bob Johnson	28	bob@invalid	2023-07-10
4	Alice Lee	45	alice@example.com	2023-08-05
5	John Doe	35	john@example.com	2023-05-15
6	Eve Adams	150	eve@example.com	2023-09-12
7	Mike Brown	22	mike@example.com	2023-10-01
8	Sara White			2023-11-18
9	Tom Green	31	tom@example.com	2023-12-22
10	Lily Black	40	lily@invalidcom	invalid-date

Example of Measuring Integrity:

The Integrity Score formula is applied to this dataset in a step by step manner:

1) Total Records: 10
2) Total Values: 10 records × 5 columns = 50 expected values
   - Completeness Score:
     - Missing values: Row 2 (Age missing), Row 8 (Age and Email missing) → 3 missing values.
     - Non-missing values: 50 - 3 = 47.
     - Completeness Score = $^{47}/_{50} = 0{,}94$.
   - Accuracy Score:
     - Invalid values:
       - Age: Row 6 (150 > 100) → 1 invalid.
       - Email: Row 3 (invalid format), Row 10 (invalid format), Row 8 (missing, but counted as invalid here) → 3 invalid.
       - PurchaseDate: Row 10 (invalid format) → 1 invalid.
     - Total invalid values: 5 (considering only populated fields; missing are handled in completeness).
     - Valid values: 50 - 3 missing - 5 invalid = 42 (but adjust for accuracy on populated: 47 populated, 5 invalid → 42 valid).
     - Accuracy Score = $^{42}/_{47} \approx 0{,}89$ (focusing on populated values for accuracy).

- Consistency Score:
  - Inconsistencies: Duplicates (Row 5 is exact duplicate of Row 1) → 1 inconsistent record.
  - Referential/business rules: Assume no other referential checks, but duplicates violate uniqueness for CustomerID (though IDs are unique, full row duplicate suggests error).
  - Consistent records: 10 - 1 = 9.
  - Consistency Score = $^{9}/_{10} = 0{,}90$.
- Integrity Score:

$$Integrity\ Score = \left(\frac{0{,}94+0{,}89+0{,}90}{3}\right) \times 100\,\% \approx \left(\frac{2{,}73}{3}\right) \times 100\,\% = 91\,\%$$

5.8 Uniqueness

5.8.1 Definition

Uniqueness refers to the degree to which all entities or records in a dataset are represented without duplication, ensuring each distinct real-world object is captured only once. It measures the absence of unnecessary redundancy within a single data source.

- ISO/IEC 11179-1 [i.14]: Defines uniqueness as "the property of an attribute value being different from all other values of that attribute in the specified scope".
- ISO 8000-61 [i.12]: Describes uniqueness as "the extent to which data about an entity is recorded without duplication".
- IEEE 100 [i.10]: Defines data uniqueness as "the property of data elements being distinct and not repeated".

5.8.2 Measurement Formula

Let D be a dataset with N total records.

Let N_{unique} be the number of unique records in D, determined by a defined set of key attributes.

Duplicate Record Rate (DRR):

$$DRR = \frac{N - N_{unique}}{N} \times 100\,\%$$

Uniqueness Score (US):

$$US = 100\,\% - DRR$$

5.8.3 Example

Customer Master Database Assessment:

Table 5.8.3-1: Example of a dataset for illustrating the Uniqueness property

CustomerID	NationalID	CustomerName	Email	Note
C001	901010-1234567	Cheolsoo Park	cheolsoo@example.com	
C002	901010-1234567	Cheolsoo Park	cheolsoo@example.com	Exact duplicate
C003	850505-2345678	Younghee Lee	lee@email.com	
C004	950505-1456789	Gildong Hong	hong@example.com	
C005	850505-2345678	Minsoo Park	minsoo@example.com	Partial duplicate

- Total Records (N): 5
- Unique Records (N_{unique}, NationalID basis): 3
- Duplicate Record Rate (DRR): ${}^{(5-3)}/_{5} \times 100\,\% = 40\,\%$
- Uniqueness Score (US): $100\,\% - 40\,\% = 60\,\%$

The dataset shows significant duplication (40 %), indicating potential issues with data entry processes and requiring deduplication measures.

5.9 Precision

5.9.1 Definition

Precision refers to the level of detail and exactness in the data values. It is a measure of the data's granularity and its consistency in format and scale. High precision means the data is recorded with enough detail to distinguish subtle differences and that the values adhere to expected formats (e.g. the correct number of decimal places, or consistent units). In a scientific or engineering context, precision often relates to the closeness of repeated measurements to each other, indicating the reliability of the measurement process itself, regardless of whether the measurements are correct (accuracy).

- ISO 8000-61 [i.12]: Defines precision in data quality as "the proportion of identified non-conforming data that is truly non-conforming".
- IEEE 100 [i.10]: Describes precision as "the degree to which repeated measurements or identifications show the same results".
- ISO/IEC 25012:2008 [i.1]: References precision in the context of data quality measurement accuracy.

5.9.2 Measurement Formula

Precision is typically assessed by examining the adherence of data values to established format standards (e.g. number of significant digits, unit consistency) or by measuring the dispersion of values in a series of repeated observations.

Let A_j be a quantitative attribute in the dataset.

Let S_j be the defined standard or required format for attribute A_j (e.g. "having exactly 3 decimal places" or "measured in kPa").

Let I_{ij} be an indicator variable such that:

$I_{ij} = 1$ if the value for record *i* and attribute *j* adheres to the standard S_j.

$I_{ij} = 0$ if the value for record *i* and attribute *j* violates the standard S_j.

Attribute Precision (CP_j): The precision of a specific attribute A_j is given by the ratio of present values that adhere to the established precision standard S_j to the total number of non-missing values for that attribute:

$$CP_j = \frac{\sum_{i=1}^{N_{present}} I_{ij}}{N_{present}}$$

where $N_{present}$ is the count of non-missing values for attribute A_j.

Alternative Measurement (for Repeated Measures): In cases where multiple measurements are taken for the same phenomenon (e.g. by different sensors or repeated tests), precision can be measured by the Coefficient of Variation (CV), which quantifies the data dispersion relative to the mean.

$$CV = \frac{Standard\ Deviation}{Mean}$$

A lower CV indicates higher precision (less variation among the measurements).

5.9.3 Example

Consider a dataset used for predicting pump performance in industrial machinery, where the attribute Pressure should be recorded in Kilopascals (kPa) with exactly one decimal place for sufficient precision.

Table 5.9.3-1: Example of a dataset for illustrating the Precision property

Index	Timestamp	Pressure (kPa)	Adherence to Standard (1 Decimal Place)?
1	10:00:00	500,1	Yes
2	10:05:00	505,00	No (2 decimal places)
3	10:10:00	510	No (0 decimal places)
4	10:15:00	508,4	Yes
5	10:20:00	512,9	Yes

Calculation for Pressure Attribute Precision:

- Total non-missing records ($N_{present}$) = 5
- Values adhering to the 1 decimal place standard = 3 (Index 1, 4, 5)
- Pressure Attribute Precision ($CP_{pressure}$) = 3/5=0,6 or 60 %

In this example, the pressure attribute has a precision score of 60 % based on the defined format standard. The lack of precision in 40 % of the data could lead to errors if the predictive model is highly sensitive to the scale and format of its input features.

5.10 Availability

5.10.1 Definition

Availability refers to the degree to which data and the data systems are accessible for users and applications when needed. It is a critical operational metric that focuses on the accessibility, discoverability, and recoverability of the data. High availability means the data can be reliably retrieved and used on demand, minimizing downtime or delays caused by system failures, maintenance, or complex access protocols. It is often measured in the context of system uptime and data freshness, especially in real-time or mission-critical applications.

- ISO/IEC 25012:2008 [i.1]: Defines availability as "the degree to which data is present and available for use" and "the extent to which data can be retrieved by authorized users and systems".
- IEEE 100 [i.10]: Describes availability as "the degree to which a system or component is operational and accessible when required for use".
- Recommendation ITU-T X.1601 [i.16]: Defines availability in cloud computing as "the property of being accessible and usable upon demand by an authorized entity".
- NIST SP 800-53 [i.17]: Defines availability as "ensuring timely and reliable access to and use of information".

5.10.2 Measurement Formula

Availability is most commonly measured as a ratio of time the data or system is operational and accessible to the total required operational time.

Let T_{up} be the time period during which the data system is fully operational and accessible.

Let T_{total} be the total time period for which the system is expected to be available (e.g. a month or year).

System Availability (CA): The availability of the data system is given by the ratio of the system's operational time to the total time period:

$$CA = \frac{T_{up}}{T_{total}} \times 100\,\%$$

Data Availability (CD): For a specific dataset, availability can sometimes be measured by the percentage of time that the latest required data is present and accessible within the system.

$$CD = \frac{Number\ of\ time\ intervals\ the\ required\ data\ was\ present}{Total\ number\ of\ expected\ time\ intervals}$$

5.10.3 Example

Consider a real-time data feed (e.g. stock market data or sensor readings) that is expected to be available 24 hours a day, 7 days a week (720 hours in a 30-day month).

Scenario: During a 30-day monitoring period (T_{total}=720 hours), the data system experienced:

1) A hardware failure that caused 4 hours of downtime.
2) A scheduled maintenance window that took 2 hours.

System Availability (CA):

- Total time expected to be up (T_{total}) = 720 hours
- Total downtime = 4 hours (failure)+2 hours (maintenance) = 6 hours
- Operational time (T_{up}) = 720−6 = 714 hours
- System Availability (CA) = 714 / 720 ×100 % ≈ 99,17 %

This availability score is often referred to as "three nines" (99,9 %) or similar "nines" levels, where higher percentages indicate fewer service interruptions.

Data Availability (CD):

- Total time expected time intervals = 1 hour = 60 mins
- Time intervals during which price updates were missing = 5 mins
- Time intervals the required data was present = 60 - 5 = 55 mins
- Data Availability (CD) = $^{55}/_{60} \times 100\,\% \approx 91{,}67\,\%$

This example shows that while the system itself was highly available, the data feed itself was only 91,67 % available, which could lead to missed trades or inaccurate analysis in real-time system.

5.11 Coverage

5.11.1 Definition

Coverage refers to the degree to which data represents the complete set of all relevant entities, events, or phenomena within the defined domain or population of interest. It measures how well the dataset captures the entire scope of the target domain, ensuring that no significant segments are missing or underrepresented. High coverage ensures that all relevant entities, domains, or required categories are present in the dataset, preventing significant analytical blind spots or selection bias. It is often evaluated against a reference dataset or a well-defined set of business rules.

- ISO/IEC 25012:2008 [i.1]: Defines coverage as "the extent to which data represents the complete set of all relevant instances of the concept it describes".
- ISO 8000-61 [i.12]: Describes coverage as "the degree to which data encompasses the entire population or domain of interest".
- IEEE 100 [i.10]: Defines data coverage as "the completeness with which a dataset represents the target population".

5.11.2 Measurement Formula

Coverage is typically measured as the proportion of the target universe that is represented in the dataset.

Let $N_{present}$ be the number of entities, domains, or categories covered in the dataset (*D*).

Let N_{target} be the total size of the required or defined target universe (*T*).

Coverage (CC): The coverage of the dataset relative to the target universe is calculated as:

$$CC = \frac{N_{present}}{N_{target}} \times 100\ \%$$

This metric is often applied to categorical attributes (e.g. product categories, geographic regions) to ensure all expected values are present.

5.11.3 Example

Scenario #1: Consider a retail company's sales dataset that is expected to include sales records from all 10 major geographic regions of a country. The sales dataset currently contains records from 8 distinct regions. Two regions (Northwest and Southeast) have not yet been successfully integrated into the reporting system.

Calculation for Geographic Coverage:

- Total required target universe (N_{target}) = 10 major regions.
- Number of regions currently present in the dataset (N_{target}) = 8 regions.
- Coverage (CC):

$$CC = \frac{8}{10} \times 100\ \%$$

In this case, the sales dataset has 80 % geographic coverage, meaning any analysis performed will omit 20 % of the target market, potentially skewing performance metrics.

Scenario #2: National Retail Sales Data Coverage Assessment

Target Domain: All retail transactions in a country for Q1 2024:

- Total retail stores in country: 50 000
- Stores represented in dataset: 42 500
- Required time period: 90 days (January 1 to March 31, 2024)
- Days covered in dataset: 85 days
- Required attributes: 15 (sales amount, product category, location, etc.)
- Attributes with sufficient coverage: 13

$$CC_{entity} = \frac{42{,}500}{50{,}000} \times 100\ \% = 85\ \%$$

$$CC_{temporal} = \frac{85}{90} \times 100\,\% = 94{,}4\,\%$$

$$CC_{attribute} = \frac{13}{15} \times 100\,\% = 86{,}7\,\%$$

The dataset shows good temporal coverage but has gaps in entity coverage (15 % of stores missing) and attribute completeness.

5.12 Lineage

5.12.1 Definition

Lineage refers to the degree to which the origin, movement, transformation, and processing history of data is completely documented and traceable throughout its lifecycle. It provides comprehensive visibility into data provenance and transformation path. It is a qualitative or structural data quality dimension crucial for trust, auditability, and reproducibility. Complete and accurate lineage allows users to trace a data element back to its source, understand how it was calculated (e.g. which formula or algorithm was applied), and verify its compliance with regulatory standards.

- ISO/IEC 11179-1 [i.14]: (In the context of metadata) Specifies provenance as "information about the origin, creation, or derivation of data".
- ISO 8000-61 [i.12]: Describes lineage as "the recorded history of data from its origin through its transformations to its current state".

5.12.2 Measurement Formula

Lineage is generally not quantifiable with a single numerical formula in the same way as metrics like Completeness or Uniqueness. Instead, it is measured by the completeness and accuracy of the metadata that documents the data flow.

Let $N_{data\_element}$ be the total number of Critical Data Elements (CDEs) in the system.

Let $N_{documented}$ be the number of CDEs for which a complete, traceable lineage path (source, transformations and current state) is formally documented in a metadata repository.

Lineage Score (CL): The score is a measure of the documentation completeness:

$$CL = \frac{N_{documented}}{N_{data_{element}}} \times 100\,\%$$

For operational assessment, a key metric is the ability to trace any given data element back to its origin within a defined timeframe.

5.12.3 Example

Consider a financial firm that uses a complex model to derive a Risk Score for each client:

- Scenario: The firm defines the Risk Score as a Critical Data Element (CDE) requiring full lineage documentation.
- Required Documentation: Source systems for raw inputs (e.g. Credit History, Income), the specific version of the calculation model (Transformation 1), and the final database location.
- Audit Result: An audit finds that 95 out of 100 CDEs have complete and accurate lineage documentation. The remaining 5 CDEs reference outdated model versions.

Calculation for Lineage Documentation Completeness:

- Total critical data elements ($N_{data\_element}$) = 100
- Number of elements with documented lineage ($N_{documented}$) = 95

- Lineage Score (CL):

$$CL = \frac{95}{100} \times 100\,\% = 95\,\%$$

The 5 % gap indicates a risk where the calculation or origin of certain critical data values cannot be reliably verified or audited.

5.13 Anonymity

5.13.1 Definition

Anonymity is the state of data in which individual subjects cannot be identified or re-identified by the data recipient, even by combining data from external sources. It is a critical metric for privacy and regulatory compliance (e.g. GDPR, PIPA, HIPAA). Achieving anonymity involves applying various techniques like masking, generalization, perturbation, or aggregation to personal or sensitive data. The measurement focuses on the risk of re-identification.

- ISO/IEC 20889:2018 [i.18]: Defines De-identified data as "data that does not identify an individual and where the risk of re-identification is minimal, considering all relevant factors".
- ISO/IEC 29100:2024 [i.19]: Defines Anonymization as the "process of irreversibly transforming personal identifiable information (PII) such that the individual concerned cannot be identified, directly or indirectly".
- ISO/IEC TR 24027:2021 [i.20]: Defines data anonymity as "the property that data does not relate to an identified or identifiable natural person".

5.13.2 Measurement Formula

Anonymity is quantified using metrics that measure the size of the "anonymity set" or the probability of re-identification. The most widely used concept is k-Anonymity, where k represents the minimum number of individuals whose records share the same combination of identifying attribute values (called quasi-identifiers).

Let k be the anonymity parameter. A dataset satisfies k-anonymity if, for every combination of values of quasi-identifiers, there are at least k records that share those values.

Re-identification Risk (R): The risk to any individual within the anonymity set of size k is inversely proportional to k:

$$R = \frac{1}{k}$$

The goal is to ensure that for all records, the calculated R is below an acceptable privacy threshold (e.g. $R \leq 0{,}05$, which implies $k \geq 20$).

5.13.3 Example

Consider a health survey dataset where the quasi-identifiers are Age (generalized) and Zip Code (generalized to the first two digits). The target anonymity level is k=3.

Dataset Snippet (After Generalization)

Table 5.13.3-1: Example of a dataset for illustrating the Anonymity property

ID	Generalized Age	Generalized Zip	Medical Condition
1	30-39	10xx	Flu
2	30-39	10xx	Cold
3	30-39	10xx	Allergy
4	40-49	20xx	Flu
5	40-49	20xx	Flu
6	40-49	30xx	Cold

Calculation for k-Anonymity:

- Group 1 (30-39, 10xxx) has 3 records. ($k \geq 3$)
- Group 2 (40-49, 20xxx) has 2 records. ($k < 3$)
- Group 3 (40-49, 30xxx) has 1 record. ($k < 3$)

Since Group 3 has only 1 record, the dataset does not satisfy 3-anonymity (k=1). The lowest k value in the dataset is k=1.

Re-identification Risk (R) for Group 3:

$$R = \frac{1}{k} = \frac{1}{1} = 1{,}0 \; or \; 100\,\%$$

This 100 % risk indicates that an external attacker knowing an individual's Age (40-49) and Zip Code (30xxx) could uniquely identify that person and their medical condition (Cold). Further generalization or suppression is required to achieve the target k-anonymity level.

The previous example assumed a target anonymity level of k=3. The result is now calculated for the scenario where both Group 2 and Group 3 have 2 records.

Updated Dataset Snippet (After Generalization)

Table 5.13.3-2: Example of an updated dataset for illustrating the Anonymity property

ID	Generalized Age	Generalized Zip	Medical Condition
1	30-39	10xx	Flu
2	30-39	10xx	Cold
3	30-39	10xx	Allergy
4	40-49	20xx	Flu
5	40-49	20xx	Flu
6	40-49	30xx	Cold
7	40-49	30xx	Fever

The overall k value for the dataset is determined by the smallest group size across all equivalence classes:

$$Data\ Set\ k = \min\left(k_{Group\ 1}, k_{Group\ 2}, k_{Group\ 3}\right) = \min(3, 2, 2) = 2$$

The dataset now satisfies 2-anonymity, but it still fails to meet the target of 3-anonymity.

With k=2, the re-identification risk is calculated as:

$$R = \frac{1}{k} = \frac{1}{2} = 0{,}5 \; or \; 50\,\%$$

Interpretation: For any individual in Group 2 or Group 3, an external attacker knowing their quasi-identifiers (e.g. Age 40-49, Zip Code 30xxx) can narrow down their sensitive attribute (Medical Condition) to one of only two possibilities. This means there is a 50 % chance of correctly identifying their specific medical condition.

Achieving k=2 significantly reduces the risk from the k=1 (100 % risk) scenario, but further generalization or record suppression would be needed to reach the target k=3 level.

5.14 Label Quality

5.14.1 Definition

Label Quality (or **Annotation Quality**) is defined as the degree to which the labels or annotations assigned to data instances accurately and consistently reflect the true characteristics, classes, or values of the real-world entities they are intended to categorize. In the context of machine learning and AI, it measures the fidelity of the "ground truth" used for model training and evaluation.

- ISO/IEC TR 24027:2021 [i.20]: Emphasizes that data bias often originates from data sources, collection processes, and labelling practices, requiring scrutiny to identify where systematic errors or disparities are introduced.

5.14.2 Measurement Formula

Label quality is commonly measured by calculating the agreement of the assigned labels against an established Gold Standard (expert-validated ground truth).

Let N_{total} be the total number of data instances (labels) assessed. Let $N_{correct}$ labels be the number of data instances whose labels match the gold standard:

$$Label\ Accuracy\ Rate = \frac{N_{correct\ labels}}{N_{total}} \times 100\ \%$$

5.14.3 Example

Consider a quality audit on an object detection dataset used for autonomous vehicles, where a label represents a bounding box and class.

Table 5.14.3-1: Example of a dataset for illustrating the Label Quality property

Label ID	Assigned Class	Gold Standard Class	Bounding Box IoU (see note) (≥ 0,9 required)	Correct?
BB-001	Pedestrian	Pedestrian	0,95	O
BB-002	Car	Car	0,92	O
BB-003	Truck	Car	0,94	X (Classification error)
BB-004	Pedestrian	Pedestrian	0,88	X (IoU Error)
BB-005	Car	Car	0,97	O
BB-006	Bus	Bus	0,91	O
BB-007	Truck	Truck	0,96	O
BB-008	Car	Car	0,93	O
BB-009	Pedestrian	Pedestrian	0,90	O
BB-010	Bus	Truck	0,95	X (Classification Error)

NOTE: IntersectionoverUnion (IoU) measures the spatial accuracy of the bounding box:

- Total Labels Assessed (N_{total}): 10
- Correct Labels Count ($N_{correct\ labels}$): 7 (BB-003, BB-004, and BB-010 are incorrect)

$$\boldsymbol{Label\ Accuracy\ Rate} = \frac{\mathbf{7}}{\mathbf{10}} \times \mathbf{100}\ \% = \mathbf{70}\ \%$$

5.15 Traceability

5.15.1 Definition

Traceability is the degree to which it is possible to trace the origin, movement, transformations, and usage of a data element throughout its lifecycle. It ensures that a clear data lineage is available, allowing auditors and users to understand precisely how a final data value was created and derived from its initial inputs, thereby supporting accountability and compliance.

- ISO/IEC 25012:2008 [i.1]: 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.

5.15.2 Measurement Formula

Traceability is measured by assessing the completeness and accessibility of the data lineage metadata for critical data attributes.

Let $A_{critical}$ be the set of critical data attributes. Let N_A be the total number of critical data attributes. Let $N_{lineage\ available}$ be the number of critical data attributes for which complete lineage information (source system, transformation logic, and processing history) is documented and accessible:

$$Traceability\ Coverage = \frac{N_{lineage\ available}}{N_A} \times 100\ \%$$

5.15.3 Example

A financial institution tracks five key metrics derived from various systems. The requirement is for 100 % lineage documentation.

Table 5.15.3-1: Example of a dataset for illustrating the Traceability property

Metric ID	Metric Name	Source System Documented?	Transformation Logic Documented?	Change Log Available?	Full Lineage Available?
M-001	Total Daily Trade Volume	O	O	O	O
M-002	Customer Credit Score	O	O	X (Missing last 6 months)	X
M-003	Liquidity Reserve Ratio	O	O	O	O
M-004	Market Volatility Index	O	X (Proprietary logic not logged)	O	X
M-005	Interbank Rate Forecast	O	O	O	O

- Total Critical Attributes (N_A): 5
- Full Lineage Available Count ($N_{lineage\ available}$): 3 (M-001, M-003, M-005)

$$Traceability\ Coverage = \frac{3}{5} \times 100\ \% = 60\ \%$$

5.16 Confidentiality

5.16.1 Definition

Confidentiality is the characteristic of data that describes the degree to which information is protected from unauthorized access, disclosure, or use. As a system-dependent characteristic, it ensures that only approved individuals, entities, or systems can view or process sensitive data in a specified context of use.

- ISO/IEC 25012:2008 [i.1]: 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/IEC 27001:2022 [i.21]: While not a direct definition, this standard establishes controls that directly enforce confidentiality requirements, such as access control and cryptography.

5.16.2 Measurement Formula

Confidentiality is typically measured by auditing the compliance of sensitive data records with mandatory security and privacy controls (e.g. encryption, access restriction or masking).

Let $R_{confidential}$ be the total number of records containing sensitive information (e.g. Personal Identifiable Information). Let $R_{protected}$ be the number of records where all defined confidentiality controls are correctly applied and enforced:

$$Confidential\ Data\ Compliance\ Rate = \frac{R_{protected}}{R_{confidential}} \times 100\ \%$$

5.16.3 Example

A Human Resources (HR) database with 10 records is required to ensure that the Salary column is masked (or encrypted) in the development environment.

Table 5.16.3-1: Example of a dataset for illustrating the Confidentiality property

Employee ID	SSN (Masked)	Salary (Masked/Encrypted)	Masking Policy Met?	Protected?
E-001	XXX-XX-1234	Encrypted Value	O	O
E-002	XXX-XX-5678	Encrypted Value	O	O
E-003	XXX-XX-9012	75 000,00	X (Plaintext Error)	X
E-004	XXX-XX-3456	Encrypted Value	O	O
E-005	XXX-XX-2783	Encrypted Value	O	O
E-006	XXX-XX-3894	Encrypted Value	O	O
E-007	XXX-XX-2784	Encrypted Value	O	O
E-009	XXX-XX-9873	Encrypted Value	O	O
E-010	XXX-XX-0123	Encrypted Value	O	O

Assume the audit found 1 record (E-003) with a masking failure among the 10 total records:

- Total Sensitive Records ($R_{confidential}$): 10
- Protected Records Count ($R_{protected}$): 9 (E-003 failed)

$$Confidential\ Data\ Compliance\ Rate = \frac{9}{10} \times 100\,\% = 90\,\%$$

5.17 Measurement Bias

5.17.1 Definition

Measurement Bias characterizes systematic non-random errors or prejudice in data collection, processing, or labelling - leading to a consistent, systematic deviation of observed values from their true or expected values. It quantifies the direction and magnitude of the error across the dataset.

- IEEE 7003:2024 [i.22]: Defines "unjustified bias" as differential treatment or impact on individuals or groups for which no valid operational justification exists, and "inappropriate bias" as bias that is legally or morally unacceptable in the social context where the system operates.

5.17.2 Measurement Formula

Measurement Bias can be quantified using statistical measures that compare the observed data ($Observed_i$) against the true or expected value ($Expected_i$), where *n* is the number of observations:

$$Mean\ Bias = \frac{1}{n}\sum_{i=1}^{n}(Observed_i - Expected_i)$$

$$Root\ Mean\ Sqare\ Bias = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Observed_i - Expected_i)^2}$$

5.17.3 Example

Scenario: Auditing the Systematic Bias of an automated sensor network (Observed) against a set of highly accurate, professionally calibrated sensors (Expected/True) used for air quality monitoring.

Sample Dataset: Sensor Reading Bias Audit (PM2.5 concentration in $\mu g/m^3$).

Table 5.17.3-1: Example of a dataset for illustrating the Measurement Bias property

Observation (i)	Observed ($Observed_i$)	Expected ($Expected_i$)	Difference ($Observed_i - Expected_i$)
1	15,2	16,0	-0,8
2	14,5	15,0	-0,5
3	14,8	15,5	-0,7
4	15,0	15,8	-0,8
5	14,9	15,2	-0,3
Sum			-3,1

- Total Observation: 5
- Sum of Difference: -3,1

Calculation of Mean Bias:

$$Mean\ Bias = \frac{-3,1}{5} = -0,62\ \mu g/m^3$$

The Mean Bias of $-0,62\ \mu g/m^3$ indicates a systematic tendency for the automated sensor to underestimate the true concentration.

5.18 Representation Bias

5.18.1 Definition

Representational Bias is the presence of a systematic disparity in a dataset or AI system that leads to disproportionate representation, inaccurate outcomes, or unfair treatment for individuals based on their demographic or protected group attributes. It fundamentally undermines the Fairness and Representativeness of the data when applied to different population segments.

- ISO/IEC 24027:2021 [i.20]: Defines bias as "a systematic disparity of treatment, judgment, or outcomes towards certain groups of individuals or attributes". The standard provides guidance for assessing and mitigating such bias in AI systems.

5.18.2 Measurement Formula

Representational Bias is commonly measured using the Disparate Impact Ratio (DIR), which assesses the proportional difference in favourable outcomes between a protected group and a reference group.

Let $P_{protected\ group}$ be the rate of a favourable outcome (e.g. selection, approval) for the protected group. Let $P_{reference\ group}$ be the rate of a favourable outcome for the reference (non-protected) group:

$$Disparate\ Impact\ Ratio\ (DIR) = \frac{P_{protected\ group}}{P_{reference\ group}}$$

A DIR below 0,8 (or 80 %) is often considered evidence of potential adverse impact (bias) against the protected group.

5.18.3 Example

Scenario: An algorithmic tool screens 200 job applicants, divided equally into a Reference Group and a Protected Group, to determine who receives an interview (the favourable outcome).

Sample Dataset: Interview Selection Outcome.

Table 5.18.3-1: Example of a dataset for illustrating the Representation Bias property

Group	Total Applicants (N_{total})	Selected for Interview ($N_{selected}$)	Selection Rate (P)
Reference Group (Zip A)	100	80	80/100 = 80 %
Reference Group (Zip B)	100	56	56/100 = 56 %

- Reference Group Rate ($P_{reference\ group}$): 80 %
- Protected Group Rate ($P_{protected\ group}$): 56 %

Calculation of Disparate Impact Ratio (DIR):

$$Disparate\ Impact\ Ratio\ (DIR) = \frac{56\ \%}{80\ \%} = 0{,}70$$

Since the DIR of 0,70 is less than 0,80, this indicates the system exhibits Representational Bias against the Protected Group (Zip B).

6 Standard Methods for Data Quality Assessment

6.1 General

Theoretical definitions and mathematical formulas for data quality may remain abstract concepts. To apply these to a real-world data ecosystem - where vast amounts of data are generated, transmitted, and stored in real-time - concrete and empirical assessment practices are required. To demonstrate the effectiveness of the assessment and explore its applicability across various industries, this clause conducts case studies on two representative public datasets with distinct characteristics:

- **Industrial Sensor Data (Time-series):** Representative of the Industrial Internet of Things (IIoT) environment.
- **General Demographic Data (Structured):** Representative of social data with human characteristics.

The datasets selected for evaluation contain tens of thousands of records. Each case study selects and intensively analyses 10 to 12 attributes that are most critical to the respective domain out of the properties defined in the previous clause.

6.2 Case Study 1: Industrial Sensor Data

6.2.1 Dataset Overview and Selection Background

The first case study evaluates the quality of time-series sensor data, which is critical in manufacturing, energy, and aerospace industries. Such data is generated at high frequency and reflects physical phenomena; thus, it is highly susceptible to noise, drift, and missing values. Consequently, this evaluation focuses on verifying physical characteristics such as accuracy, reliability, precision, and timeliness:

- **Dataset Name:** NASA Turbofan Jet Engine Degradation Simulation Data Set (C-MAPSS).
- **Data Source:** NASA Prognostics Data Repository.
- **Access Link:** https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/ (Select "6. Turbofan Engine Degradation Simulation Data Set").
- **Data Characteristics:**
  - This dataset consists of multivariate time-series data collected from multiple aircraft gas turbine engines.
  - It records the process of each engine starting from a normal state, gradually degrading and eventually reaching failure.

- The data comprises 21 sensor measurements (temperature, pressure, speed, etc.) and 3 operational settings.

- **Scale:** Based on the train_FD001.txt file which is part of the publicly available C-MAPSS dataset, it includes 20 631 data points (rows) for 100 engines, sufficiently satisfying the standard requirement of 100 or more samples.

- **Sample dataset:**

Table 6.2.1-1: Sample dataset excerpted from the NASA C-MAPSS train_FD001.txt file

Unit	Cycle	Op_set1	Op_set2	Op_set3	S_1	S_2	S_3
1	1	-0,0007	-0,0004	100,0	518,67	641,82	1 589,70
1	2	0,0019	-0,0003	100,0	518,67	642,15	1 591,82
1	3	-0,0043	0,0003	100,0	518,67	642,35	1 587,99
1	4	0,0007	0,0000	100,0	518,67	642,37	1 582,79
1	5	-0,0019	-0,0002	100,0	518,67	642,37	1 582,85
1	6	-0,0043	-0,0001	100,0	518,67	642,10	1 584,47
1	7	0,0010	0,0001	100,0	518,67	642,48	1 592,32
1	8	-0,0034	0,0003	100,0	518,67	642,56	1 582,96
1	9	0,0008	0,0001	100,0	518,67	642,12	1 590,98
1	10	-0,0033	0,0001	100,0	518,67	641,71	1 591,24
1	11	0,0018	-0,0003	100,0	518,67	642,28	1 581,75
1	12	0,0016	0,0002	100,0	518,67	642,06	1 583,41

- 1st Column: Unit - Engine ID, Engines 1 to 100
- 2nd Column: Cycle - Time in Cycles, Number of engine runs (Time-series index)
- 3rd Column: op_set1 - Operational Setting 1, Operating conditions such as flight altitude
- 4th Column: op_set2 - Operational Setting 2, Operational conditions such as Mach number (speed)
- 5th Column: op_set3 - Operational Setting 3, Operational conditions such as Throttle Resolver Angle
- 6th Column: S_1 - Total Temperature at Fan Inlet, Fan Inlet Temperature
- 7th Column: S_2 - Total Temperature at LPC Outlet, Low-Pressure Compressor (LPC) Outlet Temperature
- 8th Column: S_3 - Total Temperature at HPC Outlet, High-Pressure Compressor (HPC) Outlet Temperature

6.2.2 Target Metrics

Considering the characteristics of industrial data, the following 10 metrics are evaluated:

1) **Accuracy:** Are sensor values within a physically valid range?
2) **Reliability:** Is the sensor signal stable over time?
3) **Timeliness:** Is the data available without delay for real-time analysis?
4) **Precision:** Do measurement values comply with the required decimal units or scale?
5) **Availability:** Are there any gaps in the time-series data?
6) **Completeness:** Do null values exist due to temporary sensor errors?
7) **Measurement Bias:** Does a specific sensor systematically output higher or lower values?
8) **Traceability:** Are the data origin and transformation logic documented?
9) **Redundancy:** Do unnecessarily duplicated sensors or derived variables exist?
10) **Integrity:** Does the data violate physical laws or correlations?

6.2.3 Detailed Assessment

6.2.3.1 Accuracy Assessment

- **Definition and Formula:**

  Accuracy indicates how closely data matches the true value of the real world. In cases where the "true value" of physical sensor data is difficult to ascertain, the Error Rate is calculated based on statistical outlier detection techniques or physical thresholds:

$$Error\ Rate = \frac{N - N_{correct}}{N} \times 100\ \%$$

- **Evaluation Execution:**

  - **Target:** Sensor 2 (T2, Total Temperature at Fan Inlet) of the NASA C-MAPSS data.
  - **Assumption:** Based on domain knowledge, the T2 sensor is required to maintain a range of 640 - 645 °C under the given engine operating conditions.
  - **Sample Data:** Data for Engine 1, cycles 1 - 192.
  - **Analysis:** For the 192 data points, the mean (μ) was 642,45, and the standard deviation ($\boldsymbol{\sigma}$) was 0,5.
  - **Criterion:** Values outside the range of $\mu \pm 3\sigma$ (640,95 ~ 643,95) are considered potential errors.
  - **Result:** All data points were confirmed to be within the range.

- **Metric Calculation:**

  - N = 192
  - $N_{correct}$ = 192
  - Error Rate = 0 %
  - Accuracy = 100 %

6.2.3.2 Reliability Assessment

- **Definition and Formula:**

  Reliability measures whether data is consistent and stable over time. In time-series data, Temporal Stability is typically measured using the Coefficient of Variation (CV):

$$Temporal\ Stability\ (CV) = \frac{\sigma}{\mu} \times 100\ \%$$

- **Evaluation Execution:**

  - **Target:** Sensor 14 (Physical Core Speed) data during the initial 50 cycles when the engine state is "Healthy."
  - **Assumption:** The engine has not degraded during this period, so sensor values should be constant.
  - **Measurement:** Mean (μ) = 8 050,5 rpm; Standard Deviation (σ) = 4,1 rpm.
  - **Calculation:**

$$Temporal\ Stability\ (CV) = \frac{4,1}{8\ 050,5} \times 100\ \% \approx 0,051\ \%$$

- **Result and Interpretation:**

  The CV value is measured at a very low 0,051 %, indicating very high data reliability. If this value exceeded 1 %, it would indicate an unreliable state due to sensor vibration or electrical noise. However, a CV of 0 (complete invariance) implies a potential "stuck-at fault," so a "minimum variability" criterion should also be considered during reliability assessment.

6.2.3.3 Timeliness Assessment

- **Definition and Formula:**

  Timeliness measures the latency between the event occurrence time (T_{event}) and the data availability time ($T_{available}$):

$$L = T_{available} - T_{event}$$

- **Evaluation Execution (Simulation):**

  - Since C-MAPSS is a static file, a streaming environment simulation is performed. A scenario is set where a virtual IoT gateway collects data and loads it onto a cloud server.
  - SLA Requirement: Data is required to reach the analysis server within 100 ms of generation for the real-time prognosis system.
  - Scenario**:** Network latency follows a normal distribution with a mean of 50 ms and a standard deviation of 20 ms. The test is performed on 1 000 samples.

- **Result:**

  - Packets exceeding 100 ms: 6 (0,6 %).
  - Timeliness Compliance Rate (TCR): 99,4 %.

- **Metric Calculation:**

$$TCR = \frac{1\,000-6}{1\,000} \times 100\,\% = 99{,}4\,\%$$

- **Insight:**

  Even if the average latency meets the SLA, the fact that 0,6 % of data arrives late means real-time alarms could be missed 6 times out of 1 000.

6.2.3.4 Precision Assessment

- **Definition and Formula:**

  Precision refers to the level of detail represented by the data, typically evaluated by the consistency of significant figures or decimal places.

$$CP_j = \frac{\sum I_{ij}}{N_{present}}$$

  Where I_{ij} is 1 if the data value adheres to the standard format - e.g. 2 decimal places - and 0 otherwise.

- **Evaluation Execution:**

  - **Target:** Sensor 7 (Total Pressure at HPC outlet).
  - **Assumption:** System design specifications require recording to two decimal places.
  - **Data Sample:** 553,33, 554,12, 553,89...
  - **Result:** All 20 631 records maintain exactly two decimal places.
  - **Precision = 100 %**

6.2.3.5 Availability Assessment

- **Definition and Formula:**

  Availability indicates the degree to which a system or data is accessible when needed. For time-series data, it checks whether data exists at expected time intervals:

$$C_A = \frac{Actual\ Records}{Expected\ Records} \times 100\,\%$$

- **Evaluation Execution:**

  - **Target:** Engine 1 failed at cycle 192. Therefore, cycle numbers (`time_cycle`) is required to be continuous from 1 to 192 without omission.
  - **Validation:** Calculate the difference (δ) in the `time_cycle` column. For all *t*, $Cycle_{t+1} - Cycle_t = 1$.
  - **Result:** The difference is constant at 1 for all intervals. No gaps found.
  - **Availability:** 100 %.

6.2.3.6 Measurement Bias Assessment

- **Definition and Formula:**

  Measurement bias quantifies the degree to which observations systematically deviate from true or expected values:

$$Mean\ Bias = \frac{1}{N}\sum(Observed_i - Expected_i)$$

- **Evaluation Execution:**

  - **Assumption:** Engines of the same model should show similar performance in a healthy state (initial 10 cycles).
  - **Baseline:** The average value of Sensor 4 (LPT Outlet Temperature) across all 100 engines is set as the Expected value (assumed 1 400,0).
  - **Target:** Engine 5 average (Observed): 1 402,5.
  - **Calculation:** 1 402,5 - 1 400,0 = +2,5

- **Result:** A positive bias of +2,5 units exists.

6.2.3.7 Completeness Assessment

- **Definition and Formula:**

$$C_D = \frac{\sum I_{ij}}{N \times M}$$

  This measures the ratio of missing values (NaN, Null) within the dataset.

- **Evaluation Execution:**

  - **Target:** Full inspection of 26 columns across all 20 631 records.
  - **Result:** 0 missing values.
  - **Completeness Score:** 100 %.

6.2.3.8 Traceability Assessment

- **Definition and Formula:**

  Evaluates whether the data generation and transformation history is managed as metadata. A checklist-based scoring method is applied rather than quantitative figures:

$$Score = \frac{N_{documented}}{N_{required}} \times 100\,\%$$

- **Evaluation Execution:**
- **Audit:** Review the readme.txt file provided with the dataset.
- **Requirements:**

  1) Data generation model (C-MAPSS version);

  2) Input variable definitions, 3) Failure mode descriptions, 4) Unit definitions.

- **Result:** Items 1 - 3 are described in detail, but explicit definitions for the physical units of some sensors (e.g. Kpa, psi, Rad/s) are ambiguous.
- **Score:** 3/4 = 75 %.

6.2.3.9 Redundancy Assessment

- **Definition and Formula:**

$$RRR = \frac{N - N_{unique}}{N} \times 100\,\%$$

- **Evaluation Execution:**

  - **Target:** sensor_1, sensor_18, sensor_19 columns.
  - **Observation:** The values of these sensors are 100 % identical or invariant (standard deviation 0) across all records. This is because the simulation settings fixed these conditions as constants.
  - **Result:** From an information theory perspective, these columns are 100 % redundant and contain zero information.

6.2.3.10 Integrity Assessment

- **Definition and formula:**

  According to the definition in clause 5.7.2, Integrity is quantified as a composite metric derived from three key dimensions: Completeness, Accuracy, and Consistency. The formula assumes equal weighting for each dimension:

$$Integrity\ Score = \left(\frac{Completeness + Accuracy + Consistency}{3}\right) \times 100\,\%$$

- **Evaluation Execution:**

  To calculate the Integrity Score for the NASA C-MAPSS (FD001) dataset, the results from the individual metric assessments performed in previous sections are utilized:

  - Completeness Score (100 %):
    - Definition: Ratio of non-missing values to total expected values.
    - Result: As analysed in Section 7 (Completeness), the dataset contains 20 631 records with 0 null values across all 26 columns.
    - Value: 1.0 (100 %)

- Accuracy Score (100 %):
  - Definition: Ratio of valid values (within expected ranges/formats) to total values.
  - Result: As analysed in Section 1 (Accuracy) and Section 4 (Precision), all sensor readings fall within the physically valid ranges defined for the simulation (e.g. Sensor T24 within 640-645°R) and adhere to the correct data types.
  - Value: 1.0 (100 %).
- Consistency Score (100 %):
  - Definition: Ratio of records adhering to business rules (e.g. no duplicates, referential integrity).
  - Rule applied: The combination of `unit_number` and `time_in_cycles` is required to be unique (Primary Key Constraint) and sequential.
  - Validation: A total of 20 631 records were verified to contain no duplicate rows, and the time cycles for each engine were confirmed to be strictly sequential without conflict.
  - Value: 1.0 (100 %):

$$Integrity\ Score = \left(\frac{100\ \% + 100\ \% + 100\ \%}{3}\right) = 100\ \%$$

- Final Score: 100 %:
  - Interpretation: The Integrity Score is perfect, which is characteristic of high-quality synthetic/simulation data.

6.3 Case Study 2: General Demographic Data

6.3.1 Dataset Overview and Selection Background

The second case study deals with data regarding people, specifically demographic data. This is widely used in marketing, policy-making, social science research, and as training data for hiring/credit scoring algorithms, which are directly linked to recent AI ethics issues. In such data, "representativeness", "fairness" and "privacy" become key quality elements rather than physical accuracy.

- **Dataset Name:** UCI Adult Data Set (Census Income)
- **Data Source:** UCI Machine Learning Repository
- **Access Link:** https://archive.ics.uci.edu/dataset/2/adult

  (Alternative site: https://www.kaggle.com/datasets/sagnikpatra/uci-adult-census-data-dataset)
- **Data Characteristics:**
  - Extracted from the 1994 US Census database.
  - Used for classification problems to predict whether an individual's annual income exceeds $50 000 ($50K).
  - **Attributes:** 14 variables including age, occupation, education level, marital status, race, sex, hours per week, native country, etc.
- **Scale:** Contains 32 561 training records and 16 281 test records, making it suitable for large-scale data quality assessment.
- **Description of attributes:**
  - Age - continuous feature
  - Workclass - continuous feature

- fnlwgt - final weight of object, continuous feature
- Education - categorical feature
- Education_Num - number of years of education, continuous feature
- Marital_Status - categorical feature
- Occupation - categorical feature
- Relationship - categorical feature
- Race - categorical feature
- Sex - categorical feature
- Capital_Gain - continuous feature
- Capital_Loss - continuous feature
- Hours_per_week - continuous feature
- Country - categorical feature
- Target - earnings level, categorical (binary) feature.

- **Sample dataset:**

Table 6.3.1-1: Example of a customer dataset excerpted from the UCI Adult Data Set

Age	Workclass	fnlwgt	Education	Education_Num	Marital_Status	Occupation	Relationship	Sex	Capital_Gain	Capital_Loss	Hours_per_week	Country	Target
39	State-gov	77 516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	Male	2 174	0	40	United States	≤ 50K
50	Self-emp-not-inc	83 311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Male	0	0	13	United States	≤ 50K
38	Private	215 646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	Male	0	0	40	United States	≤ 50K
53	Private	234 721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Male	0	0	40	United States	≤ 50K
28	Private	338 409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Female	0	0	40	Cuba	≤ 50K
37	Private	284 582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	Female	0	0	40	United States	≤ 50K
49	Private	160 187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Female	0	0	16	Jamaica	≤ 50K
52	Self-emp-not-inc	209 642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	Male	0	0	45	United States	> 50K
31	Private	45 781	Masters	14	Never-married	Prof-specialty	Not-in-family	Female	14 084	0	50	United States	> 50K
42	Private	159 449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Male	5 178	0	40	United States	> 50K
37	Private	280 464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Male	0	0	80	United States	> 50K
30	State-gov	141 297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Male	0	0	40	India	> 50K
23	Private	122 272	Bachelors	13	Never-married	Adm-clerical	Own-child	Female	0	0	30	United States	≤ 50K
32	Private	205 019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Male	0	0	50	United States	≤ 50K
40	Private	121 772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Male	0	0	40		> 50K
34	Private	245 487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Male	0	0	45	Mexico	≤ 50K
25	Self-emp-not-inc	176 756	HS-grad	9	Never-married	Farming-fishing	Own-child	Male	0	0	35	United States	≤ 50K
32	Private	186 824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	Male	0	0	40	United States	≤ 50K
38	Private	28 887	11th	7	Married-civ-spouse	Sales	Husband	Male	0	0	50	United States	≤ 50K

6.3.2 Target Metrics

The following 8 metrics were selected to align with the characteristics of social/demographic data:

- **Completeness:** Analysis of missing patterns due to survey refusal, etc.
- **Uniqueness:** Do duplicate individual records exist?
- **Consistency:** Are there logical contradictions between attributes (e.g. age vs. pension receipt)?
- **Anonymity:** Is there a risk of re-identification?
- **Representation Bias:** Are specific races or genders under/over-represented?
- **Label Quality:** Is the income classification label (> 50K) accurately assigned?
- **Confidentiality:** Is sensitive Personal Information (PII) encrypted or masked?
- **Coverage:** Does it sufficiently cover the population (entire US population)?

6.3.3 Detailed Assessment

6.3.3.1 Completeness Assessment

- **Definition and Formula:**

  Completeness measures the proportion of stored data against the potential for 100 % data availability. In the context of demographic data, it specifically assesses the presence of missing values or null placeholders:

  $$C_D = \left(1 - \frac{\sum Missing\ Values}{N \times M}\right) \times 100\ \%$$

  Where N is the number of records and M is the number of attributes.

- **Evaluation Execution:**

  - **Context:** The UCI Adult dataset marks missing values with a question mark (?) rather than NaN. Therefore, a pre-processing step is required to be added to the standard assessment method.
  - **Target:** `workclass`, `occupation`, `native-country`.
  - **Measurement:**
    - Total records (*N*): 32 561
    - Number of missing values ('?') in workclass: 1 836 (5,6 %)
    - Number of missing values ('?') in occupation: 1 843 (5,6 %)
    - Number of missing values ('?') in native-country: 583 (1,8 %)

- **Metric Calculation**

  - Total cells: 32 561 × 14 = 455 854
  - Total missing cells: 1 836 + 1 843 + 583 = 4 262
  - $C_D = 1 - (4\ 262/455\ 854) \approx 99{,}1\ \%$

6.3.3.2 Uniqueness Assessment

- **Definition and Formula:**

  Uniqueness ensures that no entity exists more than once within the dataset. It is calculated by comparing the number of unique records to the total number of records:

$$U = \frac{N_{unique}}{N} \times 100\,\%$$

- **Evaluation Execution:**

  - **Context**: This dataset has no unique ID (Primary Key). Therefore, records where all attribute values are identical are considered duplicates.
  - **Measurement:** Out of 32 561 records, 24 pairs were found where all columns matched perfectly.

- **Metric Calculation:**

$$US = 100\,\% - \left(\frac{24}{32\,561} \times 100\,\%\right) = 99{,}93\,\%$$

6.3.3.3 Consistency Assessment

- **Definition and Formula:**

  Consistency checks whether the data satisfies defined business rules or logical constraints between attributes:

$$C_S = \frac{N - N_{violation}}{N} \times 100\,\%$$

- **Evaluation Execution:**

  - **Method:** Verify compliance with Business Rules.
  - **Rule 1:** A person with marital-status 'Never-married' cannot have a relationship of 'Husband' or 'Wife'. → Result: 0 violations.
  - **Rule 2:** Persons with age under 17 are required to have workclass 'Never-worked' or be absent from the dataset (Child Labor Laws). → Result: Minimum age in dataset is 17. No violations.

- **Metric Calculation:** Consistency Score 100 %.

6.3.3.4 Anonymity Assessment

- **Definition and Formula:**

  Anonymity evaluates the risk of re-identification using the k-anonymity model. It ensures that any combination of quasi-identifiers maps to at least individuals:

$$R = \frac{1}{min(GroupSize)}$$

  A value of 1 indicates a unique individual is identifiable; lower values indicated better anonymity.

- **Evaluation Execution:**

  - **Quasi-Identifiers**: `age`, `race`, `sex`, `native-country`.
  - **Analysis:** Group data by the combination of these attributes and count.
  - **Finding:** Combinations such as `{Age: 90, Race: 'Asian-Pac-Islander', Sex:'Male', Country: 'South Korea'}` may exist as a single individual within the dataset.
  - **Result:** Minimum Group Size = 1 (k = 1).

- **Metric Calculation:** R = 1/1 = 100 %. Very high risk of re-identification.

6.3.3.5 Representation Bias Assessment

- **Definition and Formula:**

  Representation Bias measures the Disparate Impact Ratio (DIR) to determine if a specific demographic group receives outcomes systematically different from a reference group:

$$DIR = \frac{P(Positive\ Outcome \mid Protected\ Group)}{P(Positive\ Outcome \mid Reference\ Group)}$$

- **Evaluation Execution:**

  - Protected Group: `Female`.
  - Reference Group: `Male`.
  - Positive Outcome: `Annual income` → 50K.
  - Data Aggregation:
    - Ratio of high income among Males: approximately 30,6 % (6 662 / 21,790).
    - Ratio of high income among Females: approximately 10,9 % (1 179 / 10,771).

- **Metric Calculation:**

$$DIR = \frac{10{,}9\,\%}{30{,}6\,\%} \approx 0{,}36$$

6.3.3.6 Label Quality Assessment

- **Definition and Formula:**

  Label Quality assesses the accuracy and logical consistency of the target variable (labels) used for training machine learning models:

$$LQ = \frac{N_{consistent}}{N_{checked}} \times 100\,\%$$

- **Evaluation Execution:**

  - **Target:** Verify the accuracy of the `income` label.
  - **Logic:** Individuals with `capital-gain` > $50 000 are required to have a total income > 50K, even if wage income is 0.
  - **Analysis:** Extracted 153 records with `capital-gain` > 50 000.
  - **Result:** All 153 records are labelled > 50K.

- **Metric Calculation:** Label Consistency 100 %.

6.3.3.7 Confidentiality Assessment

- **Definition and Formula:**

  Confidentiality measures the extent to which sensitive information is protected through encryption or masking:

$$Conf = \frac{N_{protected}}{N_{sensitive}} \times 100\,\%$$

- **Evaluation Execution:**

  - **Target**: Check encryption of sensitive info within the dataset.
  - **Items:** `fnlwgt` (Final Weight - theoretically allows reverse tracking of census blocks), `relationship` (family relations).

- **Status:** All text and numbers are exposed in Plaintext.

- **Metric Calculation:** Confidentiality Compliance 0 %.

6.3.3.8 Coverage Assessment

- **Definition and Formula:**

  Coverage is typically measured as the proportion of the target universe that is represented in the dataset. It is specifically applied to categorical attributes to ensure all expected values are present:

$$CG = \frac{N_{present}}{N_{target}} \times 100\ \%$$

  Where $N_{present}$ is the number of entities or categories covered in the dataset, and N_{target} is the total size of the required target universe.

- **Evaluation Execution:**

  - **Target Universe (N_{target})**: Assumed to be the 5 major categories used in the 1990 census (White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other).

  - **Present Categories ($N_{present}$):** The dataset contains records for 'White', 'Black', 'Asian-Pac-Islander', 'Amer-Indian-Eskimo' and 'Other'.

- **Metric Calculation:**

$$CG = \frac{5}{5} \times 100 = 100\ \%$$

6.4 Comparative Analysis and Insights

6.4.1 Feasibility of Quantitative Measurement

The case studies demonstrate that the mathematical formulas defined in the present document can be effectively applied to real-world datasets to yield concrete, quantitative quality scores:

- Precision and Reliability: For the Industrial Sensor data, metrics such as Reliability (Temporal Stability) were successfully calculated using statistical variance (CV ≈ 0,051 %), proving the formula's effectiveness for time-series data.

- Bias and Fairness: For the Demographic data, Representation Bias was quantifiable using the Disparate Impact Ratio (DIR ≈ 0,36), providing a clear numerical indicator of social fairness issues.

6.4.2 Applicability and Constraints across Domains

While the standardized formulas are mathematically valid, their applicability is not universal and depends heavily on the dataset's nature and context:

- Available Metrics: Metrics like Timeliness and Reliability are critical and easily measurable in dynamic IoT environments. In contrast, these metrics are often not applicable (N/A) to static, snapshot-based datasets like the Census Income dataset, where data freshness is less critical than representational coverage.

- Unattainable Metrics: Accuracy was calculated as 100 % for the industrial simulation data because the physical bounds were known. However, for demographic data, calculating true Accuracy is often impossible without a "ground truth" verification source (e.g. verifying a person's actual income against tax records), which is rarely available in public datasets.

6.4.3 Dependency on External Information and Metadata

A key insight from the assessment is that data quality cannot always be measured solely by inspecting the dataset itself. Many metrics require additional external information or metadata to be computable:

- Requirements for Accuracy: To measure accuracy, a Ground Truth or defined Physical Threshold is required. In Case Study 1, domain knowledge of the engine's temperature range (640 - 645 °C) was a prerequisite for the calculation.
- Requirements for Timeliness: Calculating latency requires access to system logs that record T_{event} (generation time) and $T_{available}$ (arrival time). This information is often outside the dataset payload itself.
- Requirements for Traceability: This metric relies entirely on the existence of a Metadata Registry or documentation (e.g. readme.txt) rather than the data values.
- Requirements for Completeness: Measuring completeness effectively requires distinguishing between "structural missingness" (valid nulls) and actual data loss, which requires business logic definitions.

6.4.4 Domain-Specific Weighting

The assessment highlights that the "quality" of a dataset is relative to its intended use:

- Industrial Domain: Prioritizes Physical Integrity. High weights should be assigned to Reliability, Timeliness, and Accuracy to ensure operational safety.
- Social/AI Domain: Prioritizes Ethical Integrity. High weights should be assigned to Representation Bias, Anonymity, and Fairness to prevent algorithmic discrimination and privacy violations.

Table 6.4.4.-1 summarizes the comparative feasibility observed in the case studies.

Table 6.4.4-1: Comparison of metric applicability between Industrial IoT and Social Demographic domains

Metric Category	Industrial IoT Data (Time-Series)	Social Demographic Data (Static)
Physical/Temporal	Highly Applicable (Timeliness, Reliability)	Low Applicability
Integrity/Logic	Applicable (Range checks)	Applicable (Consistency rules)
Ethical/Social	Low Applicability	Highly Applicable (Bias, Anonymity)
Metadata	Applicable (Traceability)	Applicable (Lineage)

7 Conclusion

The present document has defined a comprehensive set of 18 Data Quality Metrics and demonstrated their application through case studies with real dataset. The analysis confirms that while standard formulas provide a necessary baseline for quantification, the assessment of data quality is not a mechanical process but a context-aware discipline.

The key conclusions and recommendations for future work are as follows:

- Necessity of Contextual Data: Effective quality measurement requires more than just the raw data. It necessitates a "Quality Context" that includes metadata, domain-specific thresholds, business rules, and system logs. Future data governance frameworks need to standardize the storage of this auxiliary information to enable automated quality measurement.
- Transition to Technical Specification (TS): The definitions and descriptive measurement method examples developed in the present document serve as the foundational framework for a normative TS. Future standard work is required to formalize these definitions and descriptive measurement guidelines into a TS, ensuring consistent interpretation and application across diverse data ecosystems.

- Development of Test Methodologies and Test Formular: To enable practical execution, the specific formulas for each metric identified in the present document need to be developed into standardized Test Methodologies and Test Formular. Future work can focus on defining concrete testing procedures and executable test libraries that implement these formulas, thereby facilitating automated, reproducible measurement and verification of data quality scores.
- Usage of the developed metrics: The metrics developed in the present document enable data owns to perform self-verification of their dataset quality and facilitate self-reporting. This capability allows data owner to transparently signal reliability and foster trust in AI ecosystems.

History

Version	Date	Status
V1.1.1	May 2026	Publication