

Data Quality - Notes

Nils Rechberger

2026-02-09

Lecture 01: Introduction

Relevance of Data Quality

A key advance in data quality emerged in the 1920s through R. A. Fisher's work in experimental design, which introduced randomization and replication to estimate error, bias, and precision.

Definitions

Data

Data are abstract representations of selected features of real-world entities, events, and concepts, expressed and understood through clearly definable conventions (Sebastian-Coleman, 2013).

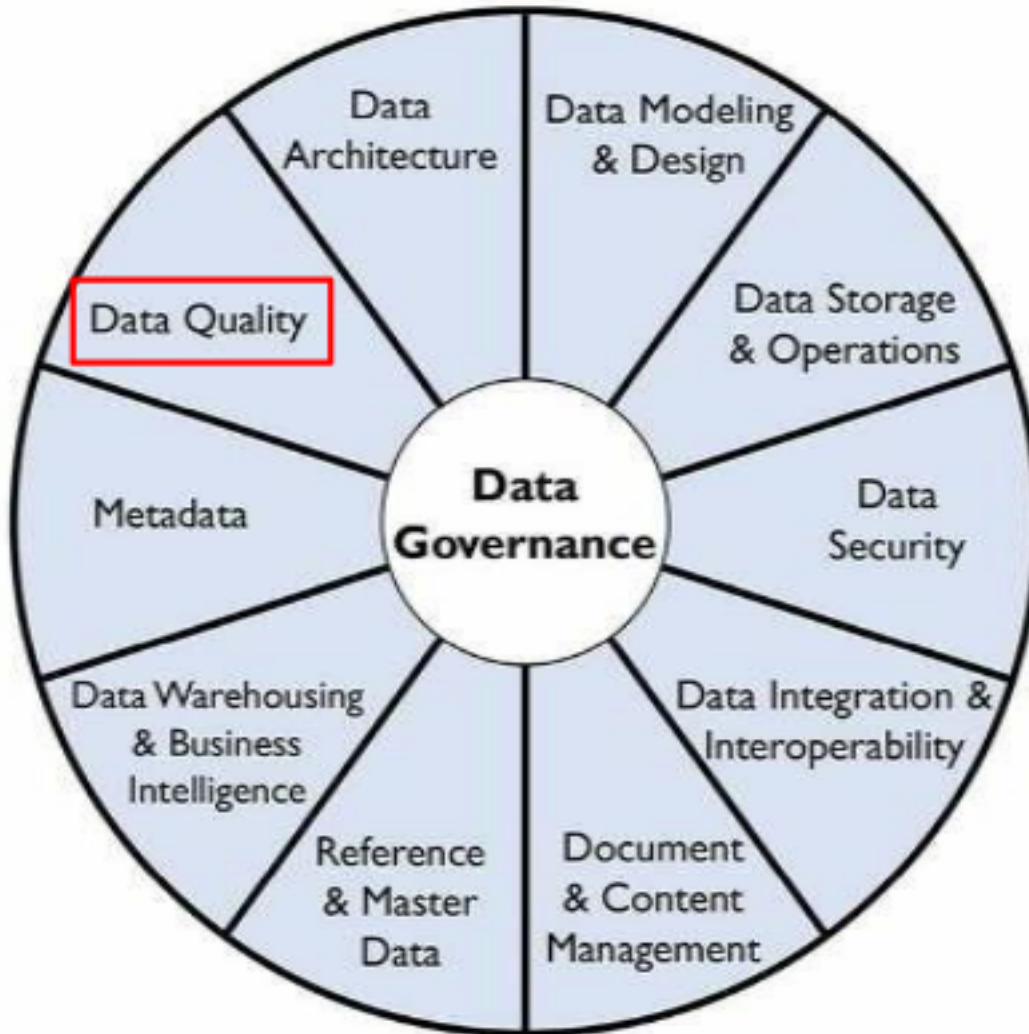
Data Quality

- Contextual data quality: The quality of data is defined by two related factors: how well it meets the expectations of data consumers (how well it is able to serve the purposes of its intended use).
- Intrinsic data quality: How well it represents the objects, events, and concepts it is created to represent.

Data Management

DQ as part of Data Management

Data Governance sits at the center (the hub) because it provides oversight, direction, policies, and coordination. Data quality is but one element of effective data management.



DAMA-DMBOK, 2017

Figure 1: Data Management

Role of Data Quality Managers

- Develop a governed approach to make data fit for purpose based on data consumers requirements.
- Define standards, requirements, and specifications for data quality controls as part of the data lifecycle.
- Define and implement processes to measure, monitor, and report on data quality levels.
- Identify and advocate for opportunities to improve the quality of data, through process and system improvements.

Data Quality Dimensions

- Correctness/accuracy: The data accurately describe the entity in question
- Completeness: No missing records/field values
- Conformity/validity: Correct types, value ranges, etc.
- Consistency: No contradictions between different data sets / tables

Note: Breaking the issue down into dimensions helps to quantify issues.

i Note

Most sources agree that the concept of data quality has several dimensions. But they don't agree on what these dimensions are.

Lecture 02: Data Quality Dimensions

In order to measure a broad concept such as data quality, it needs to be broken down into measurable, actionable dimensions. Each dimension captures one measurable aspect of data quality. Measuring data quality serves to analyze/understand/resolve or minimize data quality problems.

Note: Some of the dimensions are objective/quantitative or subjective/qualitative.

Classification of Dimensions

Objective vs. subjective measurement

- Objective: Verifiable by rules, counts, constraints, timestamps, or comparisons. Independent of user perception.
- Subjective: Depends on human judgment, expectations, or context of use.

Data value (intrinsic) vs. usage-dependent (contextual)

- Data value: Intrinsic quality of the dataset itself. Affects correctness, analytical validity, and downstream results even if usage is unknown. E.g. completeness, consistency
- Usage: Quality emerges only in interaction with users, systems, or tasks. Context-dependent.

Popular Dimensions

Quantitative and intrinsic dimensions

- Completeness: No missing records/field values
- Correctness/accuracy: The data accurately describe the entity in question
- Conformance/validity: Correct types, value ranges, etc.

Quantitative and contextual dimension

- Consistency: No contradictions between different data sets / tables
- Currency: the data are up to date

Completeness

Completeness is the most fundamental dimension. It simply measures whether data are present or absent. Analysis may be limited to critical/mandatory attributes. Measurement:

- at the field level (per column)
- at the record level (how many rows have any missing fields)
- at the table level (how many records are missing)

Note: The completeness can be calculated as a percentage at the field, record and table levels.

i Note

Not all missingness indicates a data quality problem. To record (and therefore measure) qualified missingness in a database, one can use special sentinel values (e.g., -9999 for numeric attributes, NA for text attributes)

Correctness/accuracy

It measures the extent to which data are the true representation of reality. Measurement:

- at the field
- at record level

Conformance / Validity

Conformity (or validity) means compliance of the data with a set of standards, e.g., in terms of:

- Type (“Name” is a string; “Number of children” is an integer)
- Range (“Contractual work hours” is between 0 and 50)
- Format (“Date” in ISO format YYYY-MM-DD, “Email” me{at}example.com)
- Value from a reference list (“CH”, “DE”, ...)

Note: But even fully conformant data can be 100% inaccurate.

Role of Representation

Besides conformance/validity, another prerequisite of correctness/accuracy is correct representation.

Caution

- **December 13, 1941** -> invalid
- **1941-13-12** -> valid entry, but wrong representation
- **1941-12-13** -> valid, correct representation correct/ accurate (provided the information is intrinsically accurate)

Role of Metadata

Metadata are data about data. Without metadata, it can be difficult to make sense of data. High quality metadata is important for representational effectiveness. Standard database metadata includes:

- Table/ column names and definitions
- Data types and domains of values
- Whether columns can have NULL values
- Rules regarding data relationships
- Cardinality (Number of unique values compared to the total rows, e.g. gender, boolean flags)

Lecture 03: Exploratory Data Analysis

Data Preparation

The process of manipulating data before analysis is referred to as data preparation. The meaning of the term data preparation depends on Point of view, Roles, Intended data use.

Point of view

The meaning of data preparation varies depending on one's position in the hierarchy.

Note: We take a statistical view and focus on the EXPLORE/TRANSFORM level.

Role and intended data use

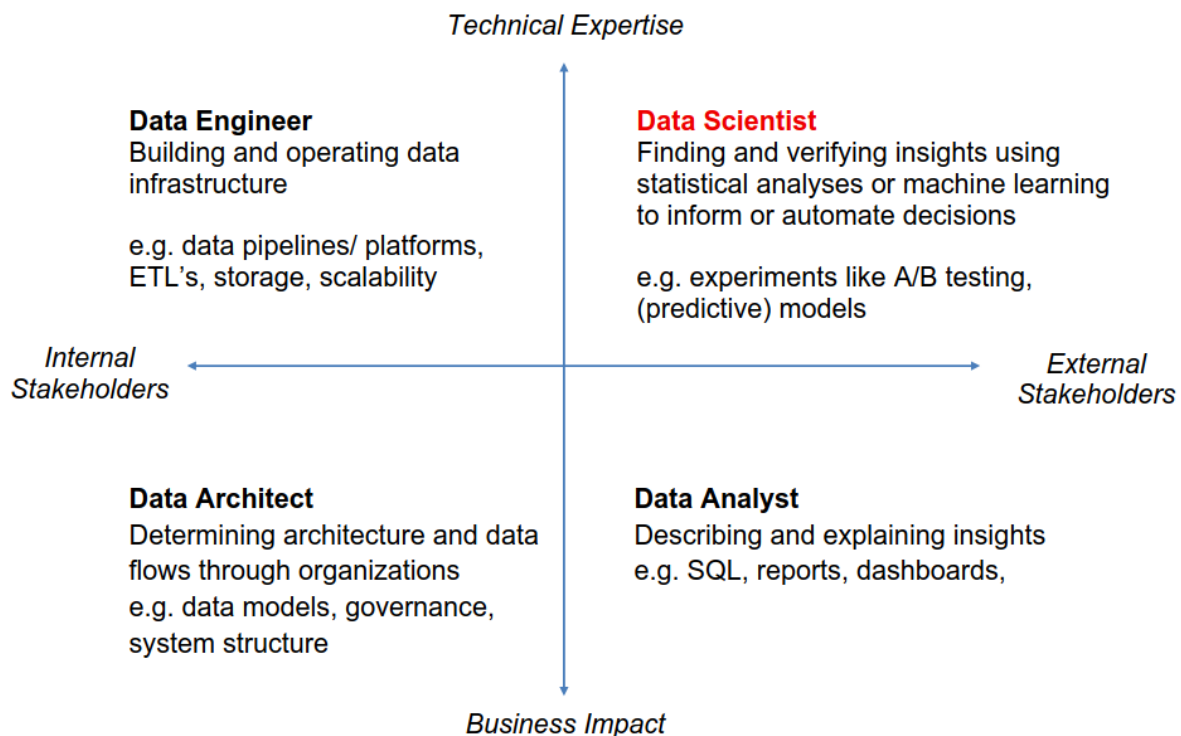


Figure 2: Role and intended data use

From Data to Understanding

The process from data to understanding is often described to as a cycle

“Data Wrangling is the ability to take a messy, unrefined source of data and wrangle it into something useful. It’s the art of using computer programming to extract raw data and creating clear and actionable bits of information for your analysis. Data wrangling is the entire front end of the analytic process and requires numerous tasks that can be categorized within the get, clean, and transform components.” (Bradley Boehmke)

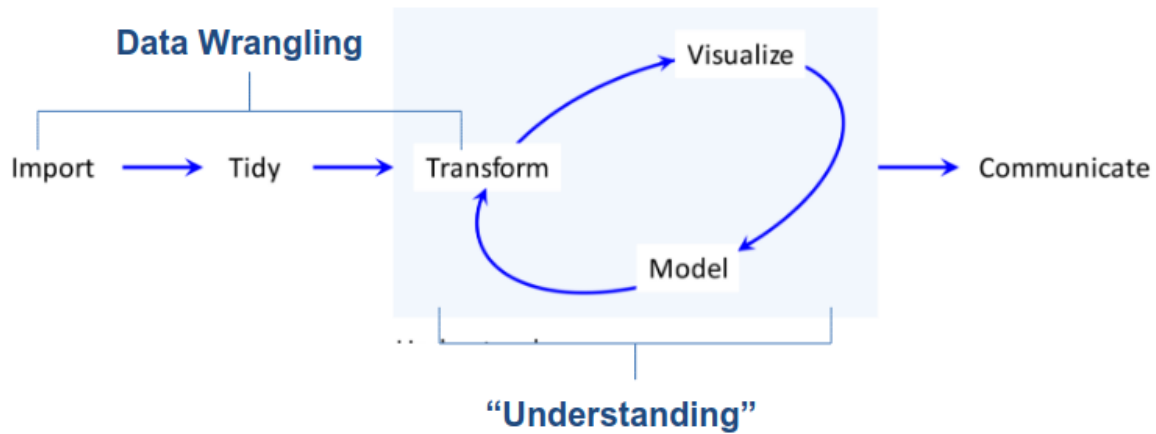


Figure 3: Data understanding cycle

Exploratory Data Analysis (EDA)

💡 Tip

Always plot your data first.

Raw Approach:

- Examine the data before applying a specific probability model
- Focus on descriptive analysis over hypothesis testing
- Methods of graphical data analysis

Iterative cycle:

- Generate questions about your data.
- Search for answers by transforming, visualizing, and modeling your data.

Steps in the EDA Process

Identify data attributes. Determine measurement scales

		Possible statements	Example
Categorical → non-metric; qualitative	Nominal	Equality (=) Inequality (≠)	Sex male = 0, female = 1 female = female, male = male , male ≠ female
	Ordinal	In addition: Relation smaller (<) Relation larger (>)	Attitude to health Very bad = 1 to very good = 5 (=> 1 < 2 < 3 < 4 < 5) <i>Very good</i> is neither five times better than <i>very bad</i> nor does <i>very good</i> have a distance of 4 to <i>very bad</i>
Metric → metric, cardinal scaled; quantitative (discrete or continuous)	Interval	In addition: Comparison of differences	Body temperature in °C Difference 40° - 35° = 5° = difference 35° - 30° But a temperature of 10° is not twice as warm as 5°. Fahrenheit scale: 10° C = 50° F, 5° C = 41° F
	Ratio	In addition: Comparison of ratios	Height: 200 cm is twice as tall as 100 cm. There is a true zero point in this scale: 0 cm

Figure 4: Hierarchy of Scales

Univariate data analysis. Recognize basic properties of the data (e.g. boxplot)

Goal: Recognize basic properties of the data.

Description of central tendency:

- Arithmetic mean → Very sensitive to outliers
- Median → Splits data in half (25% smaller, 75% larger) → More robust to outliers than mean

Description of variability/ dispersion:

- Empirical variance
- Empirical standard deviation
- 1st quartile/ 25th percentile → 25% smaller, 75% larger.
- 3rd quartile/ 75th percentile → 75% smaller, 25% larger.
- Interquartile range (IQR) = 3rd quartile – 1st quartile → More robust than standard deviation

Bivariate & Multivariate data analysis. Recognize interactions in the data

Goal: Using bivariate & multivariate data analysis to recognize interactions in the data by using Plots and hypothesis tests.

Typical graphical procedures: - Scatterplot - Mosaic Plot

Typical tests: - Pearson Chi-Square test

Following EDA steps

- Detect aberrant & missing values. Analysis & adjustment
- Outlier detection. Analysis & cleaning
- Create derived variables (index, etc.) & Variable transformation

Lecture 04: Missing Values

Why do we care about missing data?

Naively, one might think that missing data just means there is less data to analyze, so long as there are enough data to begin with. Unfortunately, this is only true in rare circumstances. In many cases, missing data may actually bias the results (give systematically wrong answers).

Note: Missing data is the rule, not the exception.

Reasons for missing data

At the source

- Survey respondents refuse to answer certain questions
- Answers are invalid or cannot be decoded – for example, in paper questionnaires.
- In longitudinal studies, parts of the sample do not participate at all measurement time points

- Sensor failure

Causes during data processing

- typing errors during coding
- Reading errors in digitalizing (for example, scanning of paper questionnaires)

Note: The more technical reasons are often the most benign, because they may occur randomly and therefore don't lead to bias.

Forms of non-response

- Unit non-response: subjects refuse to participate or are systematically unreachable
- Item non-response: subjects refuse to answer specific questions

Forms of item non-response (Little and Rubin, 2020)

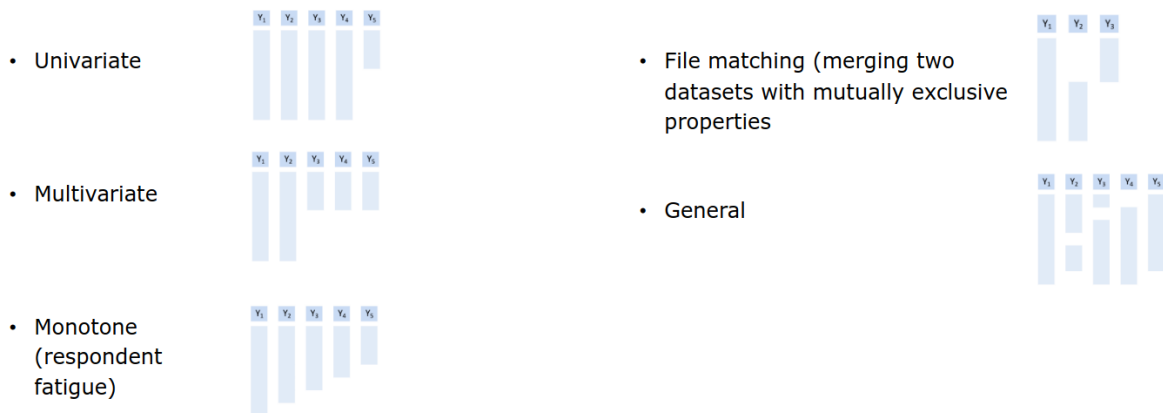


Figure 5: Forms of item non-response (Little and Rubin, 2020)

Missing data mechanism

The missing data mechanism describes the underlying process that led to the missing data. It is crucial for determining which methods are adequate for dealing with the problem, if any.

Missing Completely at Random (MCAR)

The probability of a value being missing does not depend on any observed or missing data. I.e., missing values are completely randomly distributed across all cases (persons, etc.) Cases with missing values do not differ systematically from cases without missing values.

Missing at Random (MAR)

The probability of a value being missing may depend on observed data, but not on missing data. The occurrence of a missing value occurs conditionally at random and can be explained by the values in other variables. Cases with missing values may differ systematically from cases without missing values, but in a way that can be modeled.

Missing not at Random (MNAR)

The probability of a value being missing depends on missing data. Values are systematically missing but no information is available to model their absence. There is no adequate statistical procedure to avoid bias.

Dealing with Missing Data

Listwise deletion / Complete case analysis

Delete all rows that have a missing value. Advantage: simple. Disadvantages: Wasteful. Imagine a dataset with 10 attributes, each with 10% missings. Then $P(\text{complete row}) = 0.9^{10} = 35\%$.

Pairwise deletion / Complete case analysis

Delete all rows that have a missing value. Advantage: simple & less wasteful. Results in different sample sizes for different model.

Single Imputation

Generally, imputation refers to the practice of replacing missing values with values constructed in some way. Specifically, single imputation means that each missing value is replaced by one value.

Mean Imputation

NAs are replaced by the mean of each variable. Advantages: Simple. Disadvantages: Not unbiased for regression or correlation.

Regression Imputation

For each target variable to be imputed, run a regression on some or all other variables and use it to predict the missing values. Advantages: Under MCAR and MAR, unbiased for regression. Disadvantages: Not unbiased for correlation.

Stochastic Imputation

For each target variable to be imputed, randomly draw a value from some distribution. Advantages: Under MCAR and MAR, unbiased even for correlation. Disadvantages: Standard errors too small.

Overview of Imputation Methods

	Mean	Unbiased Reg Weight	Correlation	Standard Error
Listwise deletion	MCAR	MCAR	MCAR	Too large
Pairwise deletion	MCAR	MCAR	MCAR	Complicated
Mean imputation	MCAR	–	–	Too small
Regression imp	MAR	MAR	–	Too small
Stochastic imp	MAR	MAR	MAR	Too small

Figure 6: Overview of Simple Methods

Multiple Imputation

In multiple imputation, instead of applying stochastic imputation once, it is applied m times (e.g., $m=5$). This creates 5 complete datasets. The data sets are analyzed separately, and the results are combined (pooled) according to certain rules. Advantages: Only method that can yield correct standard errors, even under MAR. Disadvantages: Still cannot handle MNAR (nothing can, only more data).

Note: In ML, we're often not interested in standard errors.

Little's MCAR Test

It essentially tests if the means / covariances are the same under each missingness pattern.

- H_0 : The data are MCAR
- H_A : They are not

! Important

The test has low power: even if the test doesn't reject, MCAR is not guaranteed.