

Data, Data Storage, Data Collection

Lecture 5: Data Quality

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

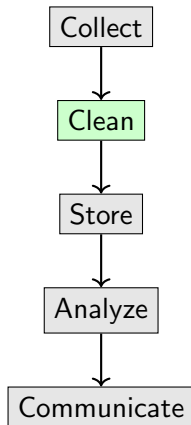
Recap

Recap: Programmatic Data Collection

- ① **HTTP** is the common mechanism behind downloading, calling an API, and scraping a page
- ② **APIs** provide a structured interface for programs.
- ③ **Web scraping** extracts data from web representations such as HTML.
- ④ **Crawling** extends scraping across linked pages.
- ⑤ **Reproducibility** requires documenting how, when and what you collected, even for automated collection.

Introduction

Within The Lifecycle



Session Objectives

By the end of this session, you should be able to:

- Explain what makes data **fit for an intended use**, and why quality is always contextual
- **Diagnose** common quality problems
- Distinguish problems checkable from the data alone from those that need outside context
- Choose and **justify a response**
- Explain how each response **affects** downstream analysis
- Document a cleaning decision so it can be audited

Why Data Quality Matters?

- Models and algorithms are only as good as the data they use.
- Biased or incorrect data → biased or incorrect conclusions.
- Automation amplifies errors

“Garbage In, Garbage Out (GIGO)”

NASA (1999)

Mars Climate Orbiter crashed due to metric vs. imperial unit mismatch.

Excel Gene Names

Auto-conversion corrupted research (e.g., "SEPT2" → "September 2").

 So we need to clean the data

Not just a technical step but a **critical thinking** exercise.

Published Scientific Data From Experiment Or Study

How much cleaning do you expect for this data?

Published Scientific Data From Experiment Or Study

How much cleaning do you expect for this data?

- Usually **clean, well-documented**, and with a **simple structure**.
- Designed for **reproducibility and sharing**.

→ Typically ready for analysis after little to no wrangling.

Administrative Data From Organizations

How much cleaning do you expect for this data?

Administrative Data From Organizations

How much cleaning do you expect for this data?

- May be clean but requires **inside knowledge** of the source.
- Often repurposed for **secondary analysis**.
- May need **transformation or combining tables**.

→ Often requires moderate cleaning, especially if used for a new purpose.

Data From An HTTP API

How much cleaning do you expect for this data?

Data From An HTTP API

How much cleaning do you expect for this data?

- Usually **structured** and consistent, by design.
- But you depend on someone else's schema and versioning.
- Requires **validation**: checking the pipeline is still receiving what you expect

→ Often requires little cleaning once the pipeline is validated.

Scraped Data From A Webpage

How much cleaning do you expect for this data?

Scraped Data From A Webpage

How much cleaning do you expect for this data?

- No schema is guaranteed: structure is inferred from HTML not meant for you.
- A page redesign can silently break your selectors.
- Formatting is inconsistent across pages, sections, or over time.

→ Typically requires **extensive** cleaning, and ongoing maintenance as the source changes.

Data Quality

Data Quality

Definition (Data Quality)

Data quality is the **suitability of data for its intended use**.

Is this a problem?

```
temperature = -20°C
```

Data Quality

Definition (Data Quality)

Data quality is the **suitability of data for its intended use**.

Is this a problem?

temperature = -20°C

- Recording a summer day in Paris? **Maybe**
- Recording a freezer? **Maybe not**

Data Quality

Definition (Data Quality)

Data quality is the **suitability of data for its intended use**.

Is this a problem?

temperature = -20°C

- Recording a summer day in Paris? **Maybe**
- Recording a freezer? **Maybe not**



A data-quality issue is a problem **relative to a context**
— not an absolute property of the value.

Two Kinds of Problems

Checkable from the data

- Missing values
- Invalid values
- Inconsistent representations
- Duplicates

A predicate on a record (or pair) that holds or not.

Two Kinds of Problems

Checkable from the data

- Missing values
- Invalid values
- Inconsistent representations
- Duplicates

A predicate on a record (or pair) that holds or not.

Needs context

- Suspicious / extreme values
- Measurement errors (Session 3)
- Valid, but inappropriate for the question

Requires context outside the record itself.

 A diagnostic can tell us that something is unusual — not necessarily that it is wrong.

Data Quality Is Not Only About Tables

A collection of photographs



Corrupted
file



Resolution
too low



Wrong
orientation



Missing
metadata



Wrong
observation



Duplicate
image

The question remains the same:

Is the data **suitable for its intended use?**

A Problem Can Also Be About Scope

Suppose we want to study:

“What images contain bicycles?”

Our dataset contains:

- 10,000 perfectly valid images;
- correct file formats;
- no missing files;
- no duplicates.

But all images were collected from a website about **motorcycles**.



The files may be technically perfect.

The dataset is still unsuitable for the question.

Assessing Data Quality

How Do We Assess Data Quality? (ISO 25012 Standard)



	Characteristic	Metric
✓	Validity	% satisfying constraint
1/3	Completeness	% non-null
⚖️	Consistency	% agreeing records
👤	Uniqueness	# key violations

⚠️ Thresholds are **not universal** but depend on the intended use.

Tooling can help the automation (ydata-profiling, pandera or Great Expectations) → **Profiling** for the unknown unknowns.

Completeness

Completeness asks whether the information we need is present.



$$\text{Completeness}(X) = \frac{\# \text{non-missing values}}{\# \text{expected values}}$$

Example

300 missing temperatures on 10 000 observations gives 97% completeness.



97% is not a verdict. Is it enough for **this** analysis?

Why Would Data Be Missing?

Data can be missing for many reasons:



Sensor Failure



Connection Failure



Missing Answers



Human Error



Privacy Concerns



Data Corruption

Missing data is often not random: it reflects underlying issues in collection or context (or storage).



Missingness can even be information (Survivorship bias)

Sentinel Values

Definition (Sentinel Value)

A special symbol value indicating missing data.

Example: return value of an algorithm indicates “no solution”.

Problem: Find the index of a value in an array, return -1 if the value does not exist.

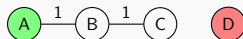
Algorithm: Linear search

[3, 7, 1, 5, 9]

- For 5 $\rightarrow$ 3
- For 2 $\rightarrow -1$

Problem: Find the value of the shortest path from s to t in a graph.

Algorithm: Dijkstra's



- From A to C $\rightarrow$ 2
- From A to D $\rightarrow \infty$

The exact values and their meanings **depend on the context**.

Sentinel Values Behave Differently

Equality testing

```
1 >>> None == None           # Base Python
```

Sentinel Values Behave Differently

Equality testing

```
1 >>> None == None           # Base Python
2 True
3 >>> np.nan == np.nan       # Numpy's Not A Number
```

Sentinel Values Behave Differently

Equality testing

```
1 >>> None == None           # Base Python
2 True
3 >>> np.nan == np.nan       # Numpy's Not A Number
4 False
5 >>> pd.NaT == pd.NaT       # Panda's Not A Time
```

Sentinel Values Behave Differently

Equality testing

```
1 >>> None == None           # Base Python
2 True
3 >>> np.nan == np.nan       # Numpy's Not A Number
4 False
5 >>> pd.NaT == pd.NaT       # Panda's Not A Time
6 False
7 >>> pd.NA == pd.NA         # Panda's Not Available
```

Sentinel Values Behave Differently

Equality testing

```
1 >>> None == None           # Base Python
2 True
3 >>> np.nan == np.nan       # Numpy's Not A Number
4 False
5 >>> pd.NaT == pd.NaT       # Panda's Not A Time
6 False
7 >>> pd.NA == pd.NA         # Panda's Not Available
8 <NA>
```

Sentinel Values Behave Differently

Equality testing

```
1 >>> None == None           # Base Python
2 True
3 >>> np.nan == np.nan       # Numpy's Not A Number
4 False
5 >>> pd.NaT == pd.NaT       # Panda's Not A Time
6 False
7 >>> pd.NA == pd.NA         # Panda's Not Available
8 <NA>
```

Formally: NA uses a three-values logic (Kleene's strong logic of indeterminacy)

Pandas documentation (NA semantics)

“Experimental: the behaviour of NA can still change without warning”.

Validity

Validity asks whether values satisfy known rules or constraints.

- age must be between 0 and 120;
- probability must be between 0 and 1;
- a temperature recorded in Kelvin cannot be negative.

Valid, but wrong

age = 42 is syntactically valid, in range ... but the person is 24.

 **Validity is not accuracy.** Detecting this needs another source, domain knowledge, or manual verification.

Consistency

Sometimes the problem only appears when we compare pieces of data.

Field	Value
Date of birth	2004-03-12
Date of event	1998-07-21

Or two sources report:

	Source A	Source B
Paris	21.4	21.4
London	18.2	18.1
Berlin	17.5	25.7

Consistency requires defining what agreement should mean.

Uniqueness

Sometimes each record looks perfectly reasonable.

Temperature_Records

ID	Date	Temperature
1842	2026-09-20	21.4
1843	2026-09-20	19.8
1842	2026-09-20	21.4

If **ID** is supposed to uniquely identify an observation:

uniqueness violation = 1

 Duplicate detection requires knowing what should be unique.

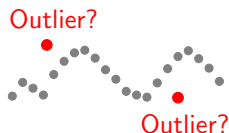
Uniqueness is enforced by UNIQUE/PRIMARY KEY in SQL.

Outliers

Definition (Outliers)

Outliers are data points that differ significantly from others.

⚠ Not necessarily irrelevant:
possible interesting phenomena.



Detection methods

- Vary and are often case-specific or based on domain knowledge.
- Statistical tests and visualization are a good starting point.

Detecting Outliers: Z-Scores

A common standardized measure is:

$$z_i = \frac{x_i - \bar{x}}{s}$$

A common heuristic is to **investigate** observations with $|z_i| > 3$.

Detecting Outliers: Z-Scores

A common standardized measure is:

$$z_i = \frac{x_i - \bar{x}}{s}$$

A common heuristic is to **investigate** observations with $|z_i| > 3$.

But $\bar{x}$ and s are themselves affected by extreme observations.

10, 10, 11, 10, 12, 11, 100

The outlier increases **both** the mean and the standard deviation — which can make its own z-score **less** extreme.

 A z-score is a **diagnostic**, not a deletion rule — and it can be distorted by the very problem it is trying to detect.

A More Robust Alternative: IQR



The **interquartile** range is:

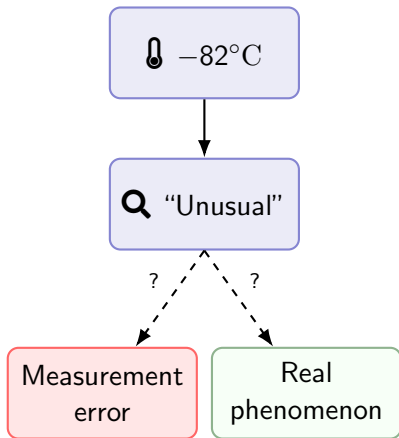
$$IQR = Q_3 - Q_1$$

Outliers as observations $x \leq Q_1 - 1.5 IQR$ or $x \geq Q_3 + 1.5 IQR$.

Why is this more robust?

- Q_1 and Q_3 depend on the ordering of the data;
- a small number of extreme observations has limited influence;

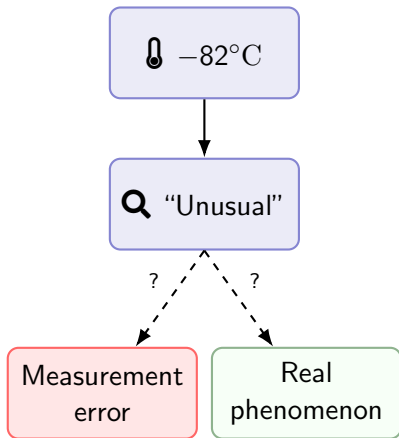
 But again: **flagged** does not mean **wrong**.



A weather station reports -82°C .

An outlier detector says "this is unusual".

Location? Instrument? Units?
Time of year? Plausibility?

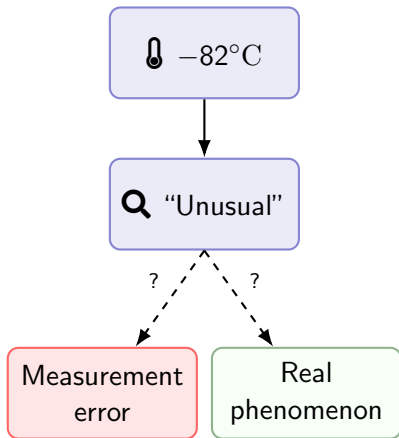


A weather station reports -82°C .

An outlier detector says "this is unusual".

Diagnosing a problem tells you **that something deserves a closer look.**

Location? Instrument? Units?
Time of year? Plausibility?



Location? Instrument? Units?
Time of year? Plausibility?

A weather station reports -82°C .

An outlier detector says "this is unusual".

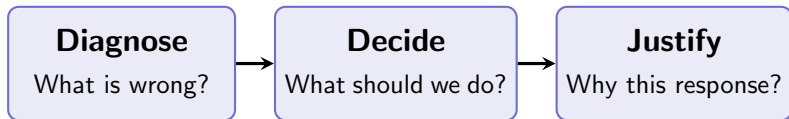
Diagnosing a problem tells you **that something deserves a closer look**.

It does not tell you **what to do about it**.

Handling Data Quality Issues

From Diagnosis to Decision

⚠ Cleaning is not about making data look nice.



Finding a problem does not determine the response.

Strategies for Handling Data Quality Issues

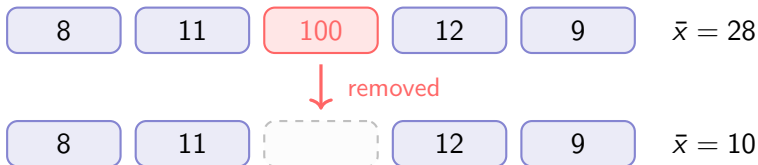
-  **Remove** Drop the record.
-  **Correct** Fix the value when the correct one is known or derivable.
-  **Impute** Replace with an estimate or fill.
-  **Flag** Add a marker column; let downstream analysis decide.
-  **Retain** Leave as is.

These are **decisions**, not a fixed pipeline: the same issue can justify different responses under different questions.

Remove

Remove an observation not contributing to the analysis.

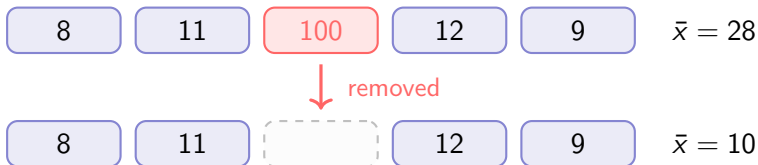
- a duplicated observation;
- a record outside the defined study population;
- a corrupted observation that cannot be recovered.



Remove

Remove an observation not contributing to the analysis.

- a duplicated observation;
- a record outside the defined study population;
- a corrupted observation that cannot be recovered.



Removal changes n , shifting the mean, the variance, and the uncertainty of any estimate built on the data.

 If the removed records are **not** random, removal isn't just less precise: it shifts the estimate itself (cf. Session 2).

Correct a value when we have evidence about the intended value.

temperature	city
21.4	Paris
291.5	London

Suppose the second value was mistakenly recorded as Celsius.

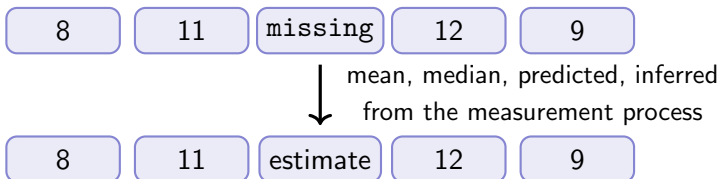
If the original measurement was 291.5 K, we can convert:

$$291.5 - 273.15 = 18.35^{\circ}\text{C}$$

 Correction requires evidence.

Changing a suspicious value to something “more reasonable” is not correction.

Imputation replaces missing information with an estimate.



Mean imputation replaces values with the observed mean $\bar{x}_{\text{obs}}$:

- the overall mean unchanged
- the imputed values form a **zero-variance subgroup** sitting exactly at the mean

From the law of total variance (Session 2): $\text{Var}_{\text{after}} = \frac{n_{\text{obs}}}{n} \text{Var}_{\text{obs}}$.



Mean imputation systematically **shrinks the variance**.

Flag

Sometimes the best solution is to keep the record and mark it.

temperature	quality_flag
21.4	OK
-82.0	CHECK

This allows later analysis to:

- include the observation;
- exclude it;
- compare analyses with and without it;
- investigate it further.

Flagging preserves information while making uncertainty explicit.

Sometimes the correct response is:

Do nothing.

An observation can be:

- unusual;
- statistically extreme;
- inconvenient for a model;

and still be valid evidence.

For example, an unusually large value might be exactly the phenomenon we want to study.



“Unusual” is not a sufficient reason to remove data.

Different Decisions, Different Consequences

	Effect	Potential consequence
 Remove	Decreases n	Selection, bias
 Correct	Changes values	Depends on evidence
 Impute	Adds estimates	Reduced variance, uncertainty
 Flag	Preserves data	Analysis must account for flag
 Retain	Changes nothing	Unusual values remain influential



There is no universally “safe” cleaning operation.

Documenting Decisions: The Audit Table

Every decision made today is **recorded**, not just applied.

Field	Issue	Check	Response	Justification
age	invalid (< 0)	range check	remove	unrecoverable
city	inconsistent	format check	correct	known mapping
income	missing (12%)	completeness	flag	may matter

The goal is not only a clean dataset.
The goal is a defensible cleaning process.

Conclusion

Ethical Dimension of Data Cleaning

Cleaning is **not neutral**:

-  Removing, imputing, or transforming data can introduce or hide variability and bias.
-  Always question how a cleaning decision affects the interpretation of the data.
-  Be especially cautious with social or behavioral datasets, where biases carry real-world consequences.

 A **defensible** dataset is not the same as a **comfortable** one.

Takeaways: Data Quality

- 1 **Quality is contextual:** relative to an intended use, not an absolute property.
- 2 **Assess it systematically:** characteristic, metric, and threshold.
- 3 **Diagnosis is not a verdict:** unusual is not the same as wrong.
- 4 **Every response has a cost:** remove, correct, impute, flag, retain change the evidence differently. None is universally safe.
- 5 **Document the decision,** not just the dataset.

SPENDING A MONTH CLEANING DATA



**TOPPING THE RESULTS
OF A YEAR OF FINE TUNING**

imgflip.com