# Data, Data Storage, Data Collection

## Lecture 2: Data Lifecycle

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

# Recap

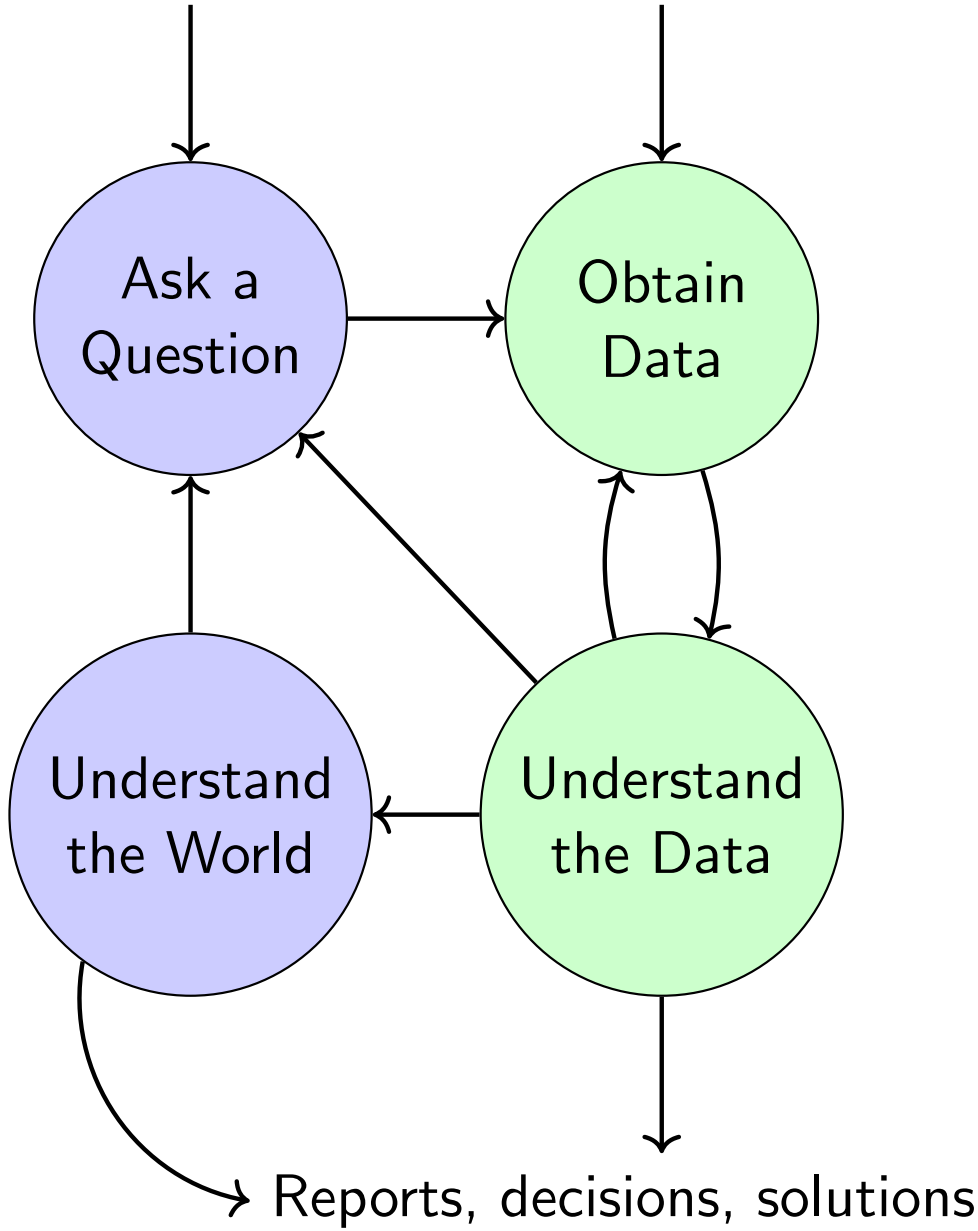
**What is data?**
Data are raw, context-free elements. With context and meaning, data they become **information**. Used for decision-making, information becomes **knowledge**.

**Where does data come from?**
Sources include sensors, surveys, logs, APIs, and human input. The origin of data strongly influences its trustworthiness and applicability.

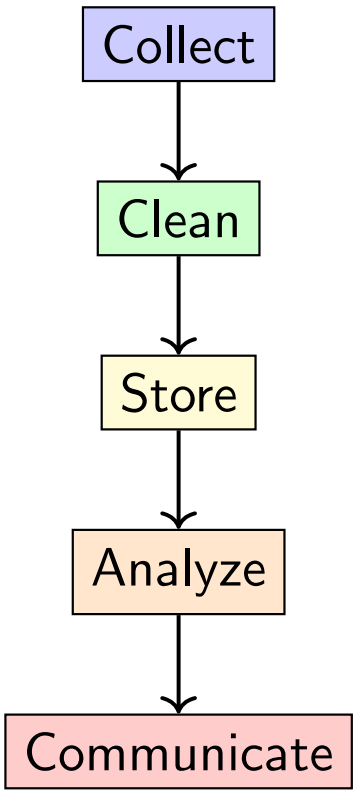
**How is data represented?**
**Unstructured** (text, media), **semi-structured** (JSON, XML), and **structured** (tables, databases).


Understanding data's **nature, sources, representations, and lifecycle** is the foundation of modern data science.

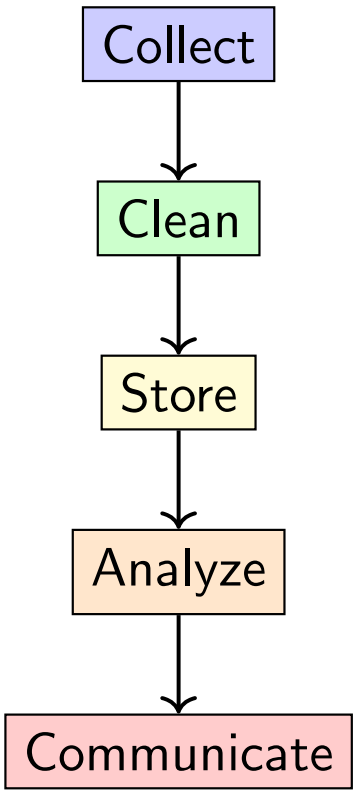
# Data Lifecycle


Ask a Question
Obtain Data
Understand the World
Understand the Data
Reports, decisions, solutions

# The Data Lifecycle

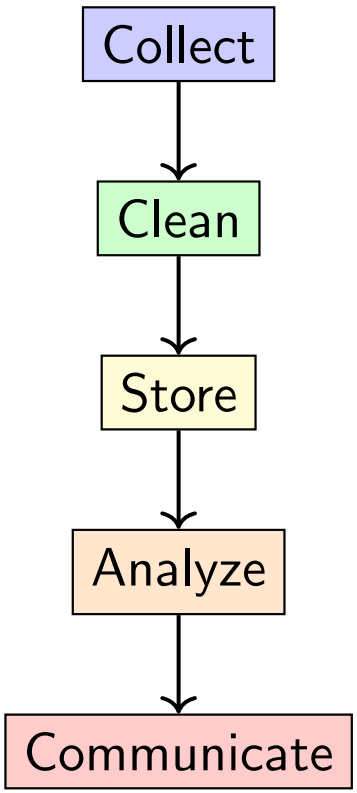


# The Data Lifecycle




Which step do you think takes the most time in a real project?

# The Data Lifecycle

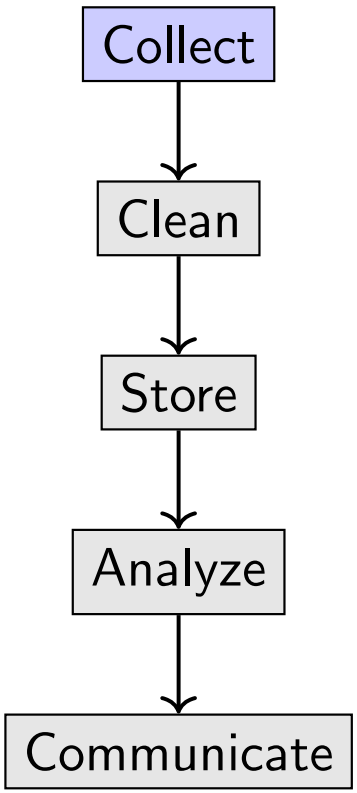



Which step do you think takes the most time in a real project?

**Cleaning and Preprocessing** (up to 80% of project time)

Gil Press. "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says". In: **Forbes** (2016)

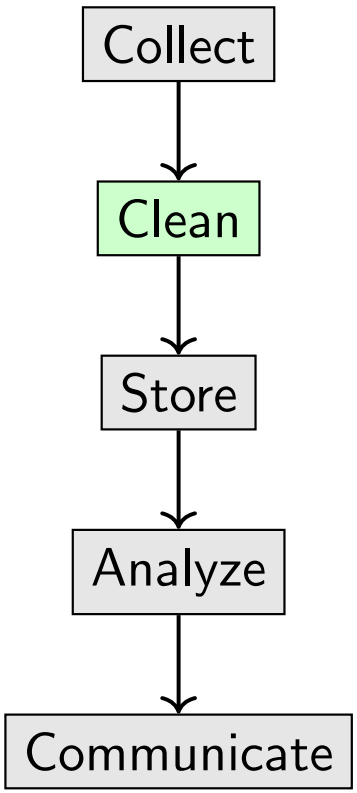
# Lifecycle Step – Collection




Sources of data: sensors, surveys, logs, APIs

Implication for data quality: representativity, missing data, reliability

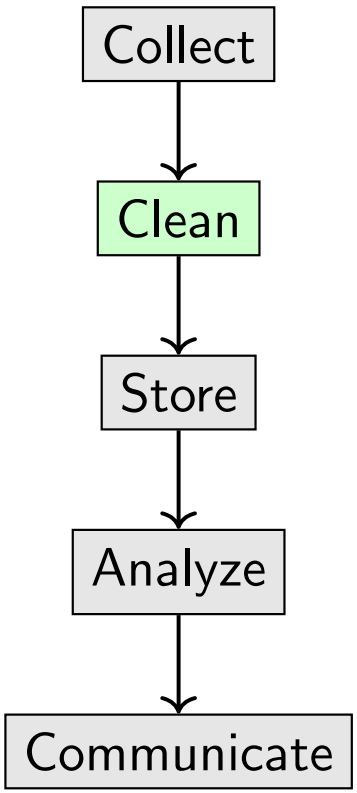
# Lifecycle Step – Cleaning




How would you fix the following?

| ID | Age | Email |
|---|---|---|
| 1 | 25 | test@example.com |
| 2 | ? | user@sample.com |
| 3 | 30 | (missing) |
| 4 | 29 | user2.fr |

# Lifecycle Step – Cleaning



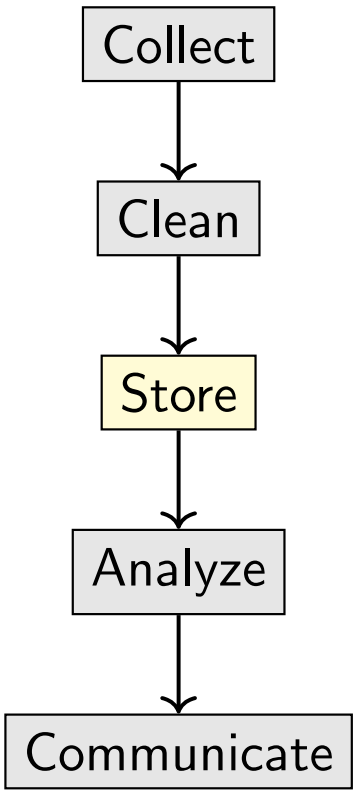

How would you fix the following?

| ID | Age | Email |
|---|---|---|
| 1 | 25 | test@example.com |
| 2 | ? | user@sample.com |
| 3 | 30 | (missing) |
| 4 | 29 | user2.fr |

Handling missing values, errors, duplicates

Data formatting / standardization

# Lifecycle Step – Storage

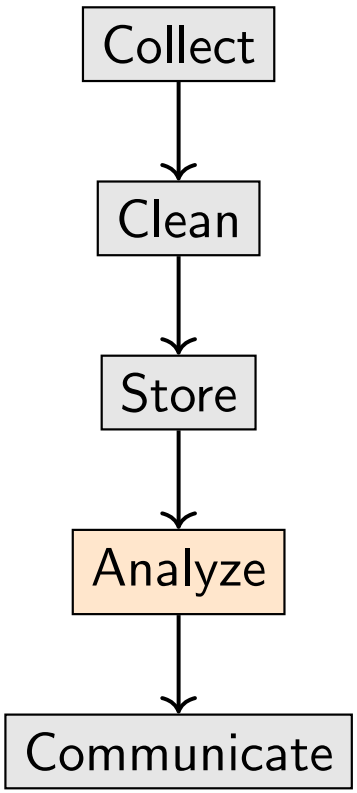



From files (CSV, JSON) to databases (SQL, NoSQL)

Tradeoffs between scalability, cost, accessibility

Short-term vs long-term storage

# Lifecycle Step – Analysis

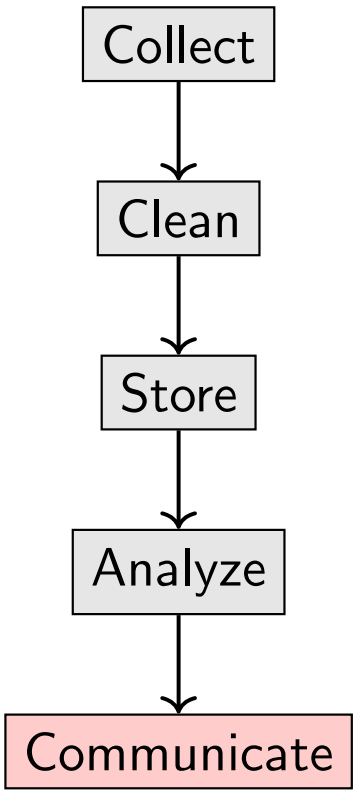



Descriptive vs inferential analysis

Tools: database queries, statistics, ML (see resp. lectures)

⚠ Data Sciences is sometimes (often) reduced to this step

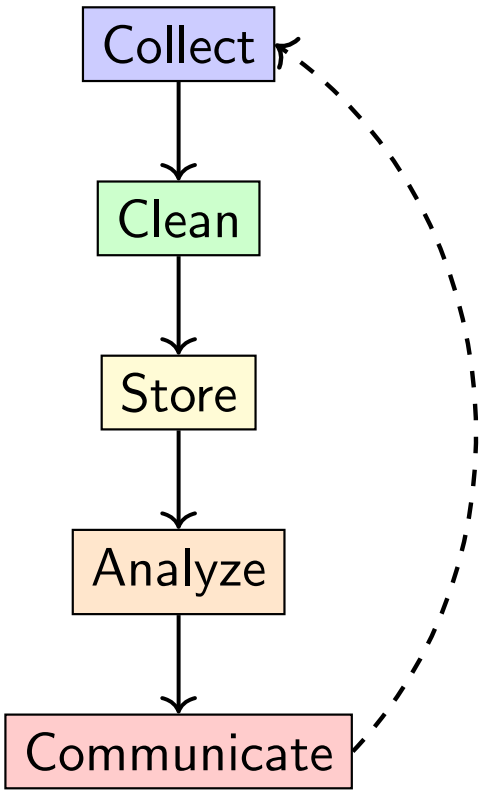
# Lifecycle Step – Communicating




Often forgotten, but crucial: without it, all previous steps are wasted

At this point we have **information**, not just data

# The Lifecycle is a Loop



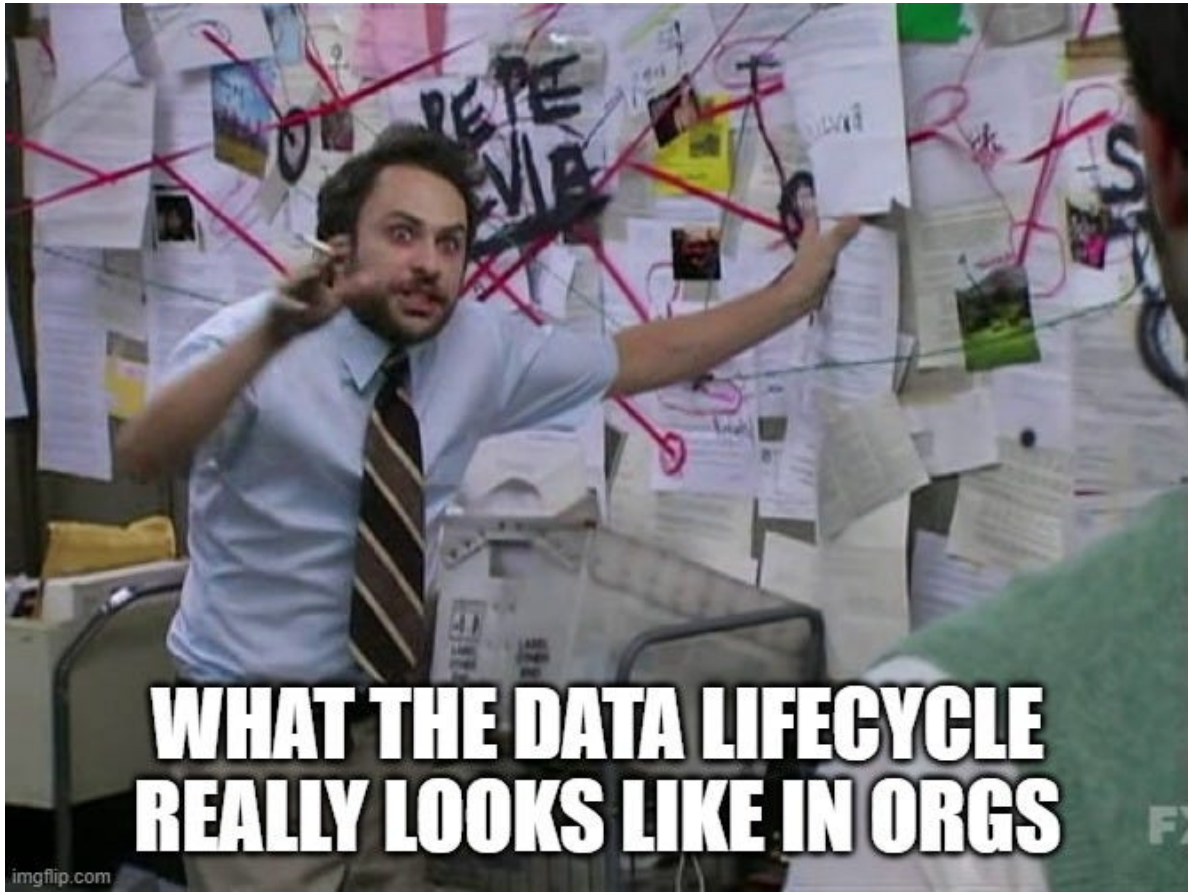

The process is **iterative**

Communicating results may highlight new data needs.

**Remark:** In reality, each of these steps include sub-activities like ingestion, integration, or deletion

# Takeaways: Data and its Lifecycle

1. Always start with a clear question or objective
2. The data lifecycle is **iterative** and **interconnected**
3. Cleaning and preprocessing are often the most time-consuming task
4. Without communication, the work is wasted




**Data lifecycle in the wild:**

Master the process, embrace the chaos,
and always document your steps!