Data, Data Storage, Data Collection

Lecture 3: Foundations of Data Collection

Romain Pascual

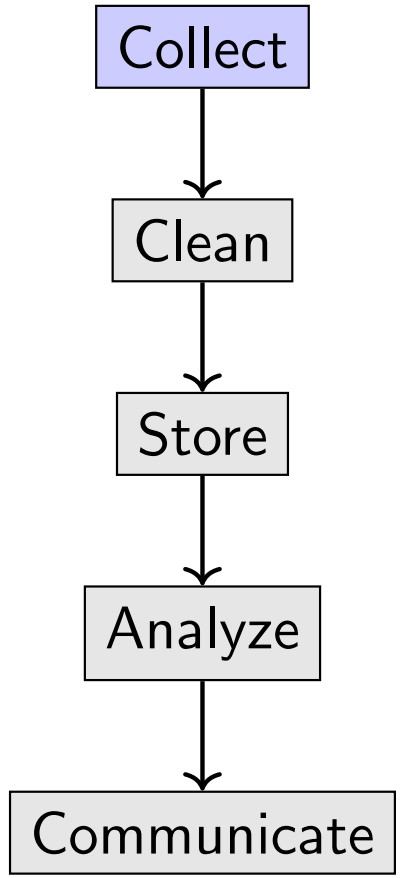
MICS, CentraleSupélec, Université Paris-Saclay

Recap

Recap: 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. Communication is as critical as analysis

Within the lifecycle

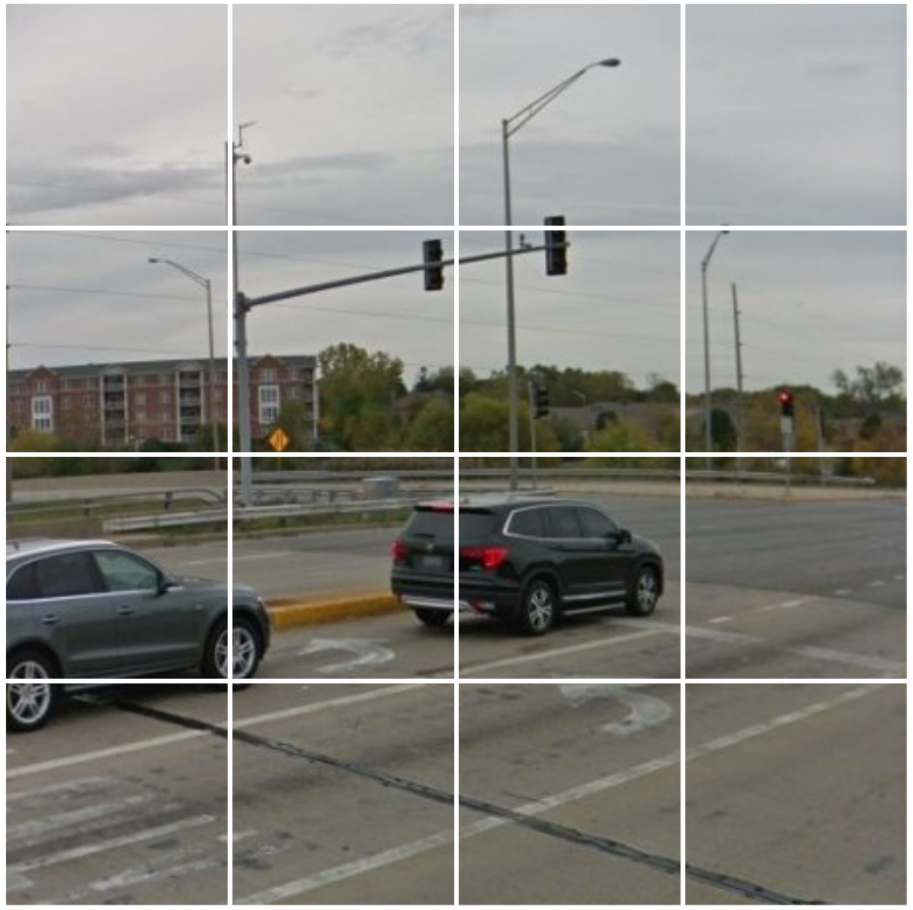


Introduction

Select all squares with

traffic lights

If there are none, click skip



SKIP

Session Objectives

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

- Understand the role of data collection within the data lifecycle
- Understand that the research question drives the collection process
- Differentiate between the various types of data and their use
- Define the scope of the collected data

Data Collection within the Data Lifecycle

Data collection is the first step of bringing raw data into a system. It ensures that you will have the raw material needed.

It can involve **creation** of data, such as running a survey or installing sensors, or **acquisition** of data, such as retrieving logs or calling an API.

It is the foundation for data quality: the reliability of all subsequent analysis depends on the scientific rigor of this stage.

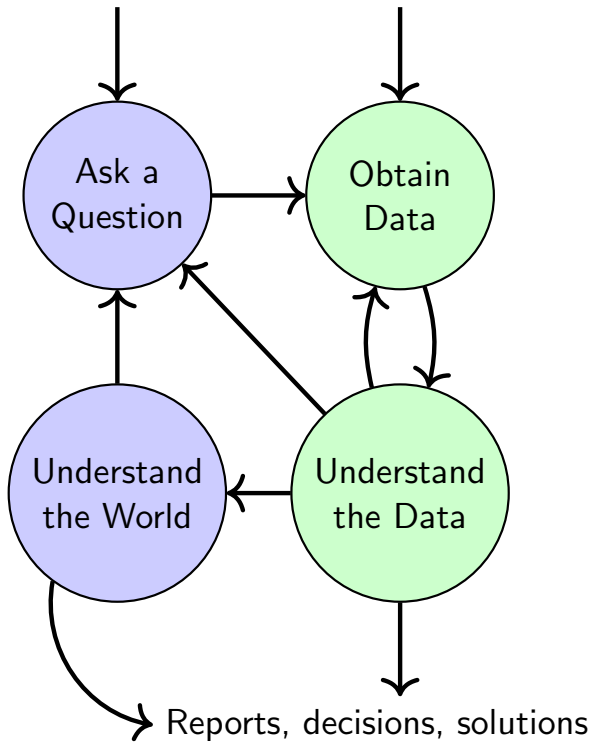
Collecting Data?

We use data to answer a question or test a hypothesis.

Data collection is **not neutral**:

- What you measure depends on the question you want to answer.
- The quality of the data influences the validity and accuracy of the answer to the question.

We want to collect **useful** data that is representative of the studied phenomenon.



Data Collection

Data do not magically appear.

Definition (Data collection)

Data collection is the systematic process of gathering observations or measurements and to place them in a database.

Three guiding questions structure the process:

- What is the aim of the study?
- What data is required?
- How will the data be collected (, stored and processed)?

Types of Data

Study Aim

A study begins with a practical or theoretical **question**.

The **aim** specifies what we want to learn. This directly determines the **data** we collect.

Examples: What Data Would You Collect?

Example

Estimate the **average time** students spend commuting to university.

Example

Understand how patients **describe their pain**.

Examples: What Data Would You Collect?

Example

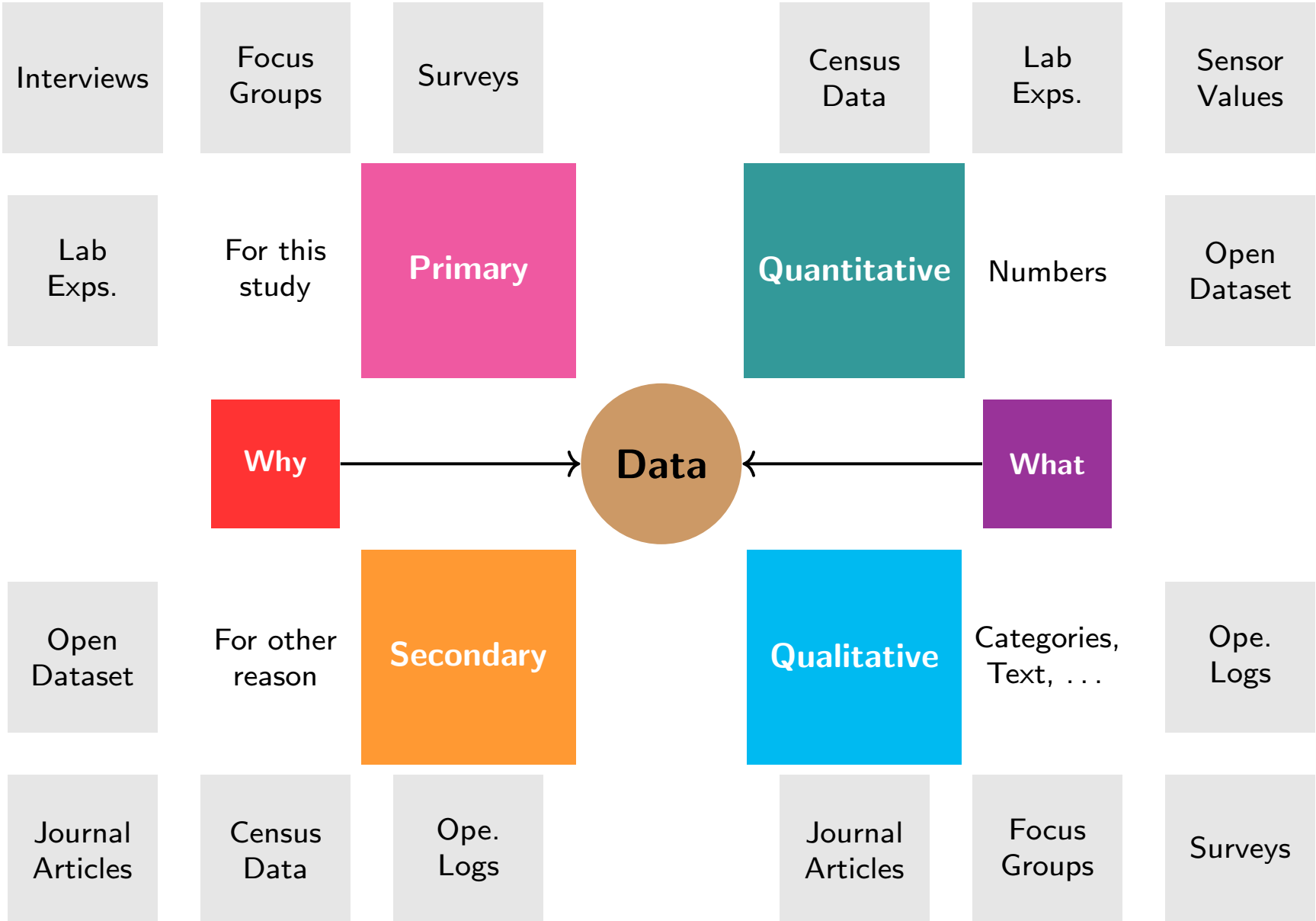
Estimate the **average time** students spend commuting to university.

Possible data: travel-time measurements

Example

Understand how patients **describe their pain**.

Possible data: interview transcripts


Interviews
Focus Groups
Surveys
Census Data
Lab Exps.
Sensor Values
Lab Exps.
For this study
Primary
Quantitative
Numbers
Open Dataset
Why
Data
What
Open Dataset
For other reason
Secondary
Qualitative
Categories, Text, . . .
Ope. Logs
Journal Articles
Census Data
Ope. Logs
Journal Articles
Focus Groups
Surveys

Quantitative Data

Definition (Quantitative data)

Numerical measures about the individuals.

Example: Travel-time measurements to estimate the average time students spend commuting to university.

Data quality: Measurement errors.

Qualitative Data

Definition (Qualitative data)

Attributes or characteristics about the individuals.

Example: Interview transcripts describing patient pain.

Data quality: Misclassification.

Primary Data

Definition (Primary data)

Data collected **specifically for the study at hand**.

Example: Wearable devices monitoring heart rate in a clinical trial.

✓ Tailored to research questions; full control over the collection protocol.

✗ Time-consuming, potentially expensive, sometimes limited in scale.

Secondary Data

Definition (Secondary data)

Data that already exist, **collected for other purposes**.

Example: Census data to analyze commuting patterns.

✓ Fast to obtain, cost-effective, often large-scale.

✗ Consider who collected it and why as data may not contain what you need It will save you fruitless analysis and prevent inappropriate conclusions.

Example: City Air Quality Study

Study aim: Estimate city air quality.

Which type of data would you collect? Quantitative or qualitative? Primary or secondary?

Example: City Air Quality Study

Study aim: Estimate city air quality.

Which type of data would you collect? Quantitative or qualitative? Primary or secondary?

- **Quantitative:** Numerical measurements such as CO2 levels, because we need precise, measurable indicators of pollution.

Example: City Air Quality Study

Study aim: Estimate city air quality.

Which type of data would you collect? Quantitative or qualitative? Primary or secondary?

- **Quantitative:** Numerical measurements such as CO2 levels, because we need precise, measurable indicators of pollution.
- **Primary:** Direct measurements from environmental sensors provide up-to-date, local data tailored to the study.
- **Secondary:** Government monitoring reports or historical datasets allow broader coverage and longitudinal analysis.

Example: Understanding Student Stress

Study aim: Understand factors affecting student stress.

Which type of data would you collect? Quantitative or qualitative? Primary or secondary?

Example: Understanding Student Stress

Study aim: Understand factors affecting student stress.

Which type of data would you collect? Quantitative or qualitative? Primary or secondary?

- **Qualitative:** Interviews, survey responses, diaries capture subjective experiences that numbers alone cannot convey. Quantitative measures (e.g., stress scores) could complement qualitative insights, but qualitative primary data is critical to understand personal perceptions.

Example: Understanding Student Stress

Study aim: Understand factors affecting student stress.

Which type of data would you collect? Quantitative or qualitative? Primary or secondary?

- **Qualitative:** Interviews, survey responses, diaries capture subjective experiences that numbers alone cannot convey. Quantitative measures (e.g., stress scores) could complement qualitative insights, but qualitative primary data is critical to understand personal perceptions.
- **Primary:** Collected directly from students to ensure relevance to the research question and control over question framing.

Combining Types of Data

Key idea

Integrating various (types of) data provides a richer understanding.

Qualitative data can explain trends observed in **quantitative** data.

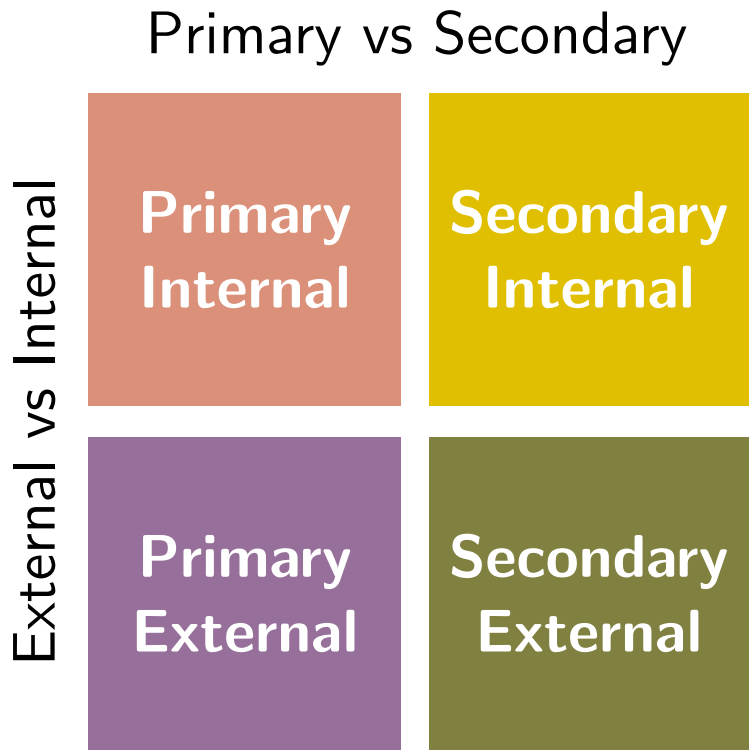
Combining **primary** and **secondary** data balances novelty and efficiency.

Example: A public health study may integrate patient interviews (qualitative primary data) with hospital admission statistics (quantitative secondary data).

Primary/Secondary vs Internal/External

Two guiding questions :

- **1.** Why was the data collected?
  Primary (for this study)
  vs **Secondary** (reused)
- **2.** Who generated and controls the data?
  Internal (inside the organization)
  vs **External** (outside)



	Primary	Secondary
Internal	We design and run a survey for our own students (for this study).	
External		

	Primary	Secondary
Internal	We design and run a survey for our own students (for this study).	Our university already ran a survey last year, we reuse the results for another purpose.
External		

	Primary	Secondary
Internal	We design and run a survey for our own students (for this study).	Our university already ran a survey last year, we reuse the results for another purpose.
External	Another university runs the same survey specifically for our collaborative project, and shares the results.	

	Primary	**Secondary**
Internal	We design and run a survey for our own students (for this study).	Our university already ran a survey last year, we reuse the results for another purpose.
External	Another university runs the same survey specifically for our collaborative project, and shares the results.	The ministry of education has already collected and published survey data; we download and use it.

How is Data Generated?

Data is not just **primary** or **secondary**, it is also defined by **how it is created**.

Two key categories:

- Data **designed** for a purpose (e.g., experiments).
- Data **organically** generated (e.g., logs, social media).

Definition (Designed Data)

Data **intentionally collected** for a specific purpose.

- Surveys
- Clinical trials
- A/B tests
- Sensor networks

✓ Controlled, structured, aligned with research goals.

✗ Costly, time-consuming, potential artificiality.

Definition (Organic Data)

Data **naturally generated** as a byproduct of activities.

- Web logs
- Social media posts
- Purchase histories
- GPS traces

✓ Low cost, large scale, real-world behavior.

✗ Noisy, biased, lacks context.

Why Does This Matter?

- **Designed data** lets you **control** what you measure
  but may miss real-world complexity.
- **Organic data** reflects **real behavior**
  but is often messy and biased.

Combining both can balance control and realism, but requires addressing their distinct and limitations.

Next: How to **collect** data that is actually **useful** for answering our question?

Key Properties of Data: A Recap

Category	Properties and Descriptions
Origin	**Internal/External:** Was the data collected inside or outside the organization? **Open/Proprietary:** Is the data freely accessible?
Generation	**Machine/Human:** Was the data generated by automated systems or human input? **Designed/Organic:** Was the data created for a purpose or arose from activities?
Collection	**Primary/Secondary:** Was the data collected firsthand or obtained from existing sources?
Type	**Quantitative/Qualitative:** Is the data measurable or descriptive?

Why it matters: These properties guide how we collect, analyze, and use data.

Scope of the Data

Connecting Question and Data

Once the research question is defined and data is collected (or even before collection), we need to ensure a good connection between the question and the data.

Understanding the **scope of the data** helps us check whether the data can actually answer the question, even before cleaning or analysis.

Scope of the Data

Scope involves understanding:

- 1. The population that we want to study
- 2. The means to access the population and the topic of the investigation
- 3. The tools or instruments used for measurement
- 4. Additional protocols used in the process

The context of the study also includes time and space boundaries, if temporal or spatial relations matter.

Target, Frame and Sample

We can use three concepts to help us understand the scope:

Definition (Target)

The **target population** is the collection of elements (or units) that we ultimately intend to draw conclusions about.

A **unit** can be a person, a company, a sensor, ...

Target, Frame and Sample

We can use three concepts to help us understand the scope:

Definition (Target)

The **target population** is the collection of elements (or units) that we ultimately intend to draw conclusions about.

Definition (Frame)

The **access frame** is the collection of elements (or units) that are reachable for measurement and observation.

A **unit** can be a person, a company, a sensor, ...

Target, Frame and Sample

We can use three concepts to help us understand the scope:

Definition (Target)

The **target population** is the collection of elements (or units) that we ultimately intend to draw conclusions about.

Definition (Frame)

The **access frame** is the collection of elements (or units) that are reachable for measurement and observation.

Definition (Sample)

The **sample** is the collection of elements (or units) that are reached for measurement and observation.

A **unit** can be a person, a company, a sensor, ...

Access Frame vs Target Population

Ideally:

$$\text{Access Frame} = \text{Target Population}$$

Reality:

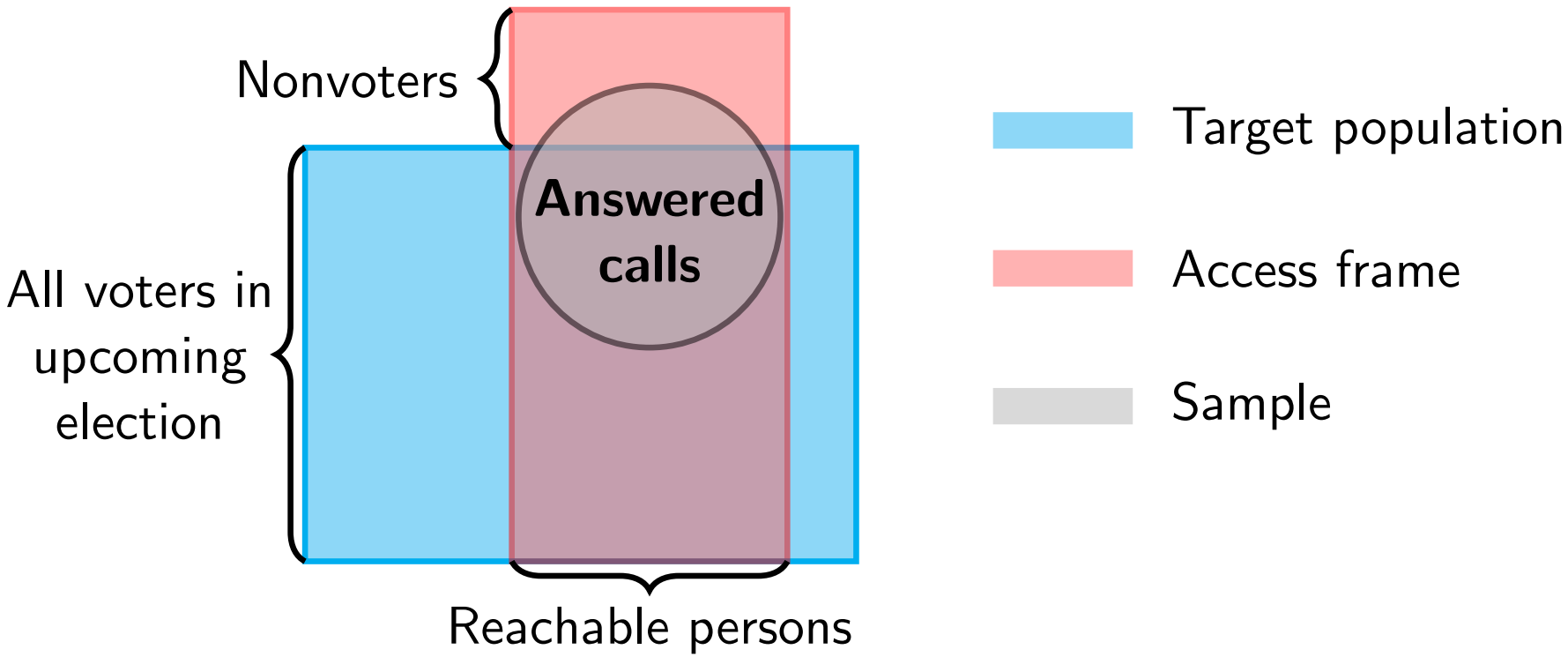
- Only a subset of the target population may be accessible
- The frame may include units outside the target population

Implications: Generalizations are only valid for the accessible units and understanding the gap helps identify potential issues and improve the study design.

Illustration: Polls

You run a poll by calling people to find out their vote intentions.

- Calling a non-voter $\rightarrow$ in frame, but not in population
- Someone never answering calls $\rightarrow$ in target population, but not in frame



Practical Insight

Following up on non-respondents or hard-to-reach units can be more effective than increasing sample size indiscriminately.

Understanding where the sample lives within the **access frame** and **target population** helps evaluate representativeness.

Scope in Natural Measurements

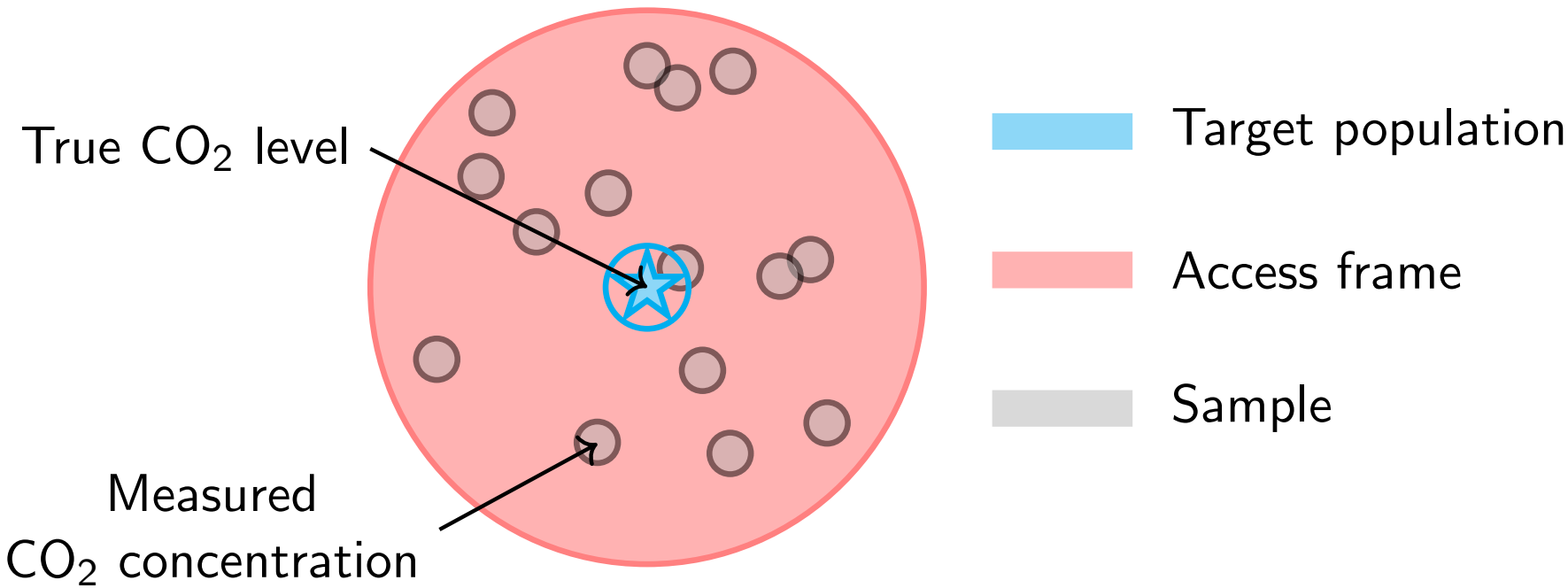
When measuring a natural phenomenon, the exact, unknown quantity of interest is called the **parameter**.

Target	the true value (a point)
Frame	determined by instrument accuracy
Sample	collection of measurements

Example: CO_2 Concentration

You want to find CO_2 levels at a given location

Instruments report averages at given frequency. True concentration is unknown.



Measurement Accuracy

Census data:

$$\text{Access Frame} = \text{Population} = \text{Sample} \rightarrow \text{Perfect Scope}$$

With perfect instruments, measurements also match target exactly.

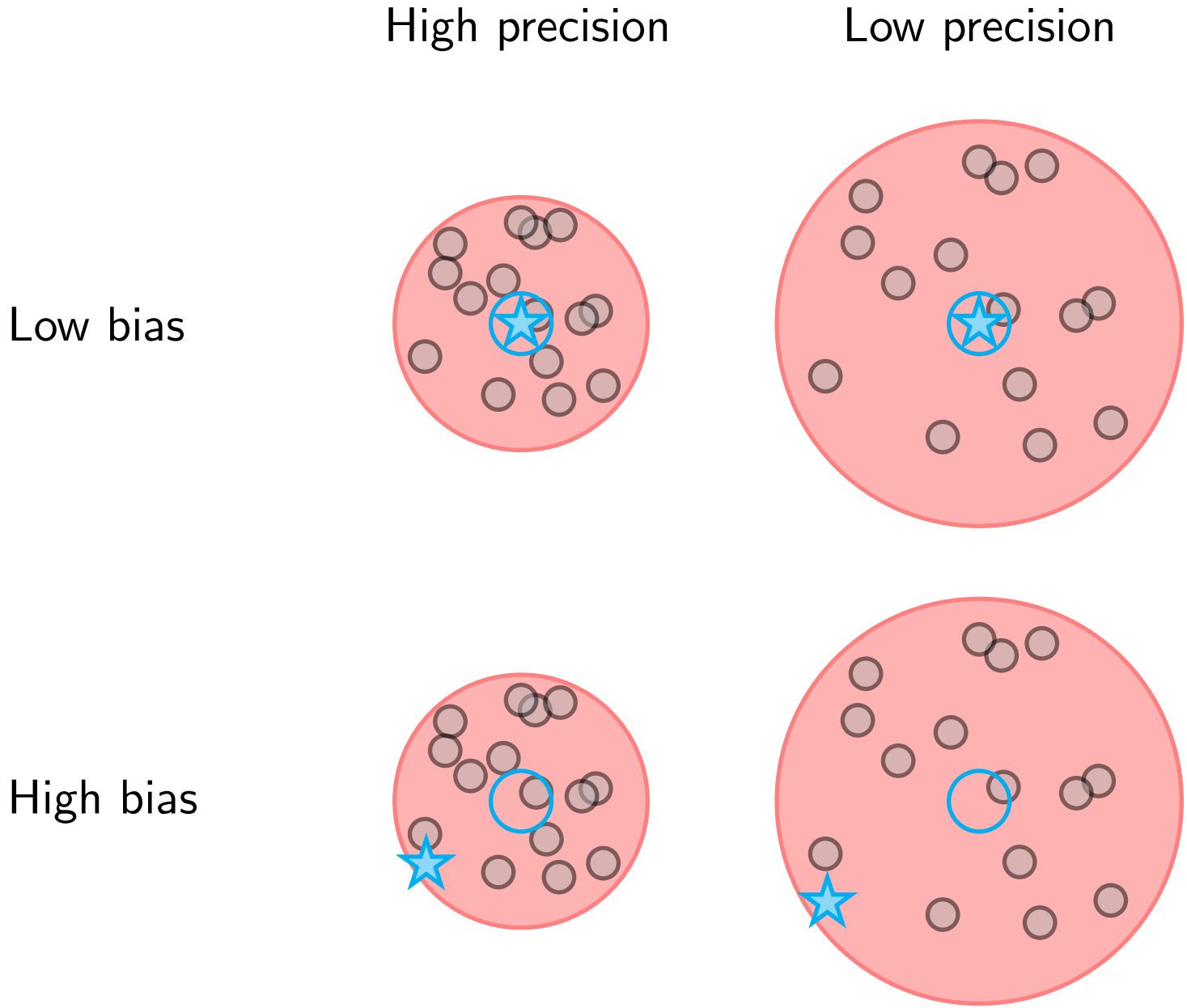
In practice, we need to quantity the accuracy of the measurements to know how to generalize the findings.

Accuracy

Accuracy can be divided into

Bias describing the distance between the average measurement and the unknown value.

Precision describing the dispersion of the measurements.


High precision
Low precision
Low bias
High bias

Conclusion

Takeaways: Foundations of Data Collection

1. Data collection is the **first step** of the data lifecycle
2. Always align the data you collect with the research question
3. Differentiate **types** of data (qualitative/quantitative, primary/secondary), each has strengths and limitations.
4. Scope matters: target population, access frame, and sample are rarely identical

Collecting data is **never neutral**
Choose wisely, measure carefully, and always question what your data **really** represents.