Data, Data Storage, Data Collection

Lecture 2: Representation in Data Collection

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

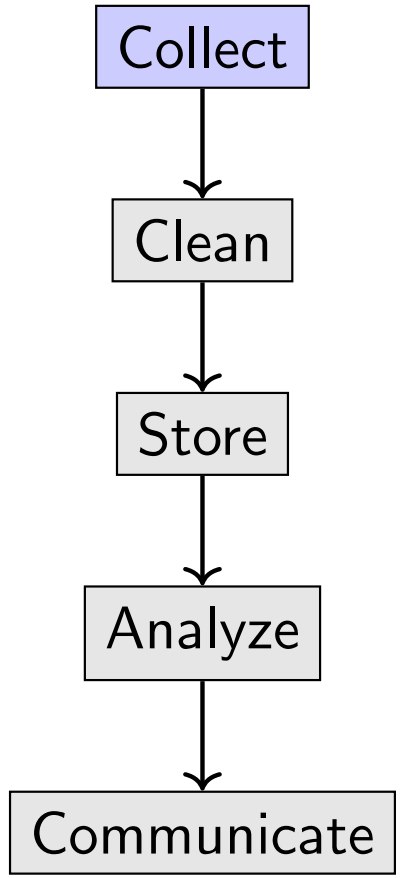
Recap

Recap: Data and its Lifecycle

1. Data in context gives information that put into action gives knowledge
2. Data always has a **context**, including ethical considerations
3. Data can be represented with different structures
4. The data lifecycle consists of several steps and is inherently **iterative**

Introduction

Within The Lifecycle

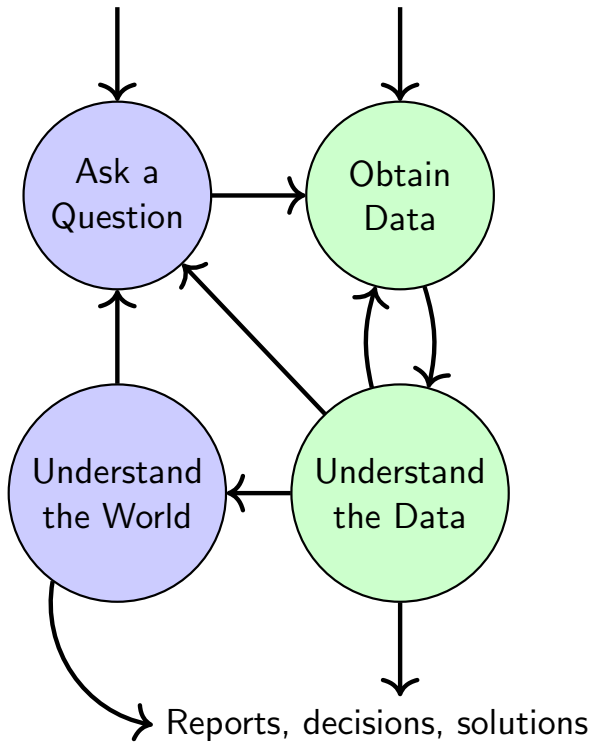


Data Collection

Definition (Data collection)

Data collection is the systematic process of gathering observations or measurements and placing them in a dataset.

1. What is the aim of the study?
2. What data is required?
3. How will the data be collected?



Data Collection

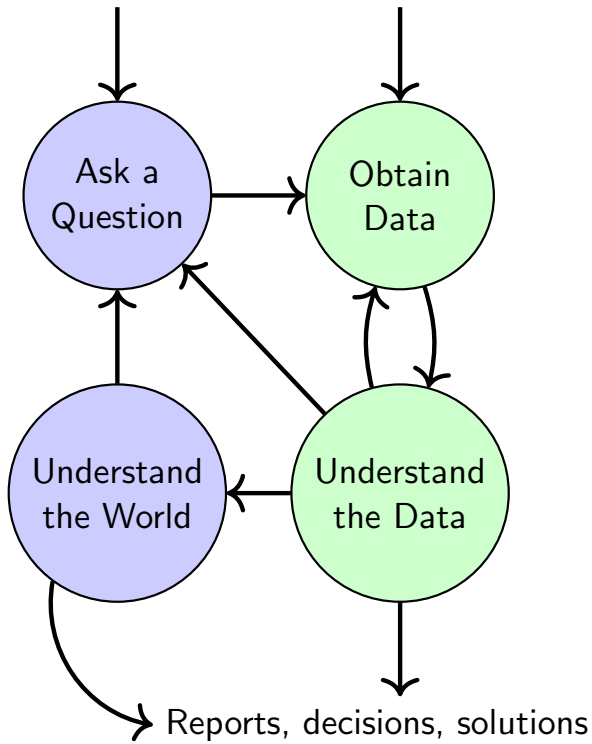
Definition (Data collection)

Data collection is the systematic process of gathering observations or measurements and placing them in a dataset.

1. What is the aim of the study?
2. What data is required?
3. How will the data be collected?

Collect **useful** data **representative** of the phenomenon.

⚠ Data quality influences the validity of the answer.



What Data Would You Collect?

Example ()

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

Example ()

Understand how patients **describe their pain**.

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**.

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

Session Objectives

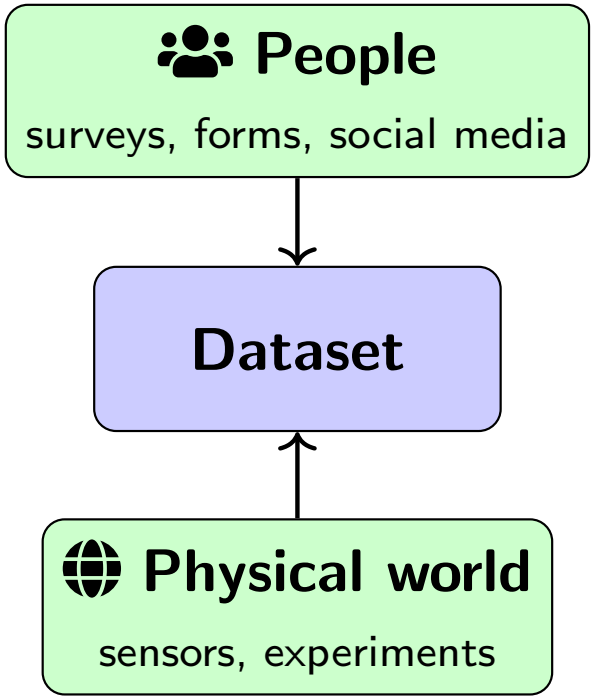
By the end of this session, you should be able to:

- Characterize a dataset's provenance and identify what it does and does not tell you
- Assess whether a dataset's scope fits a given research question
- Identify some common bias in a scenario
- Compute a sample size or confidence interval for a simple proportion

Understanding the Data Source

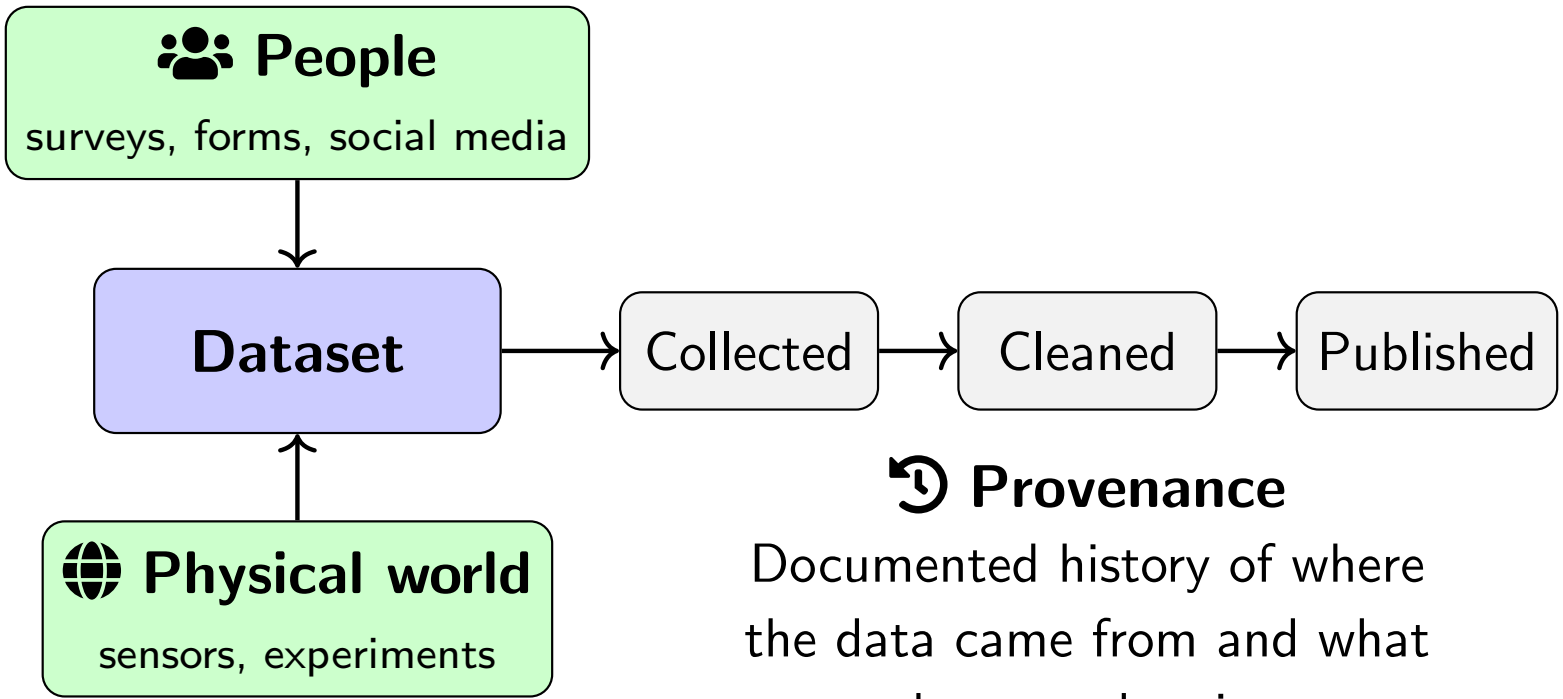
Where Does the Data Come From?

Before using a dataset, ask **how it came to exist**.



Where Does the Data Come From?

Before using a dataset, ask **how it came to exist**.

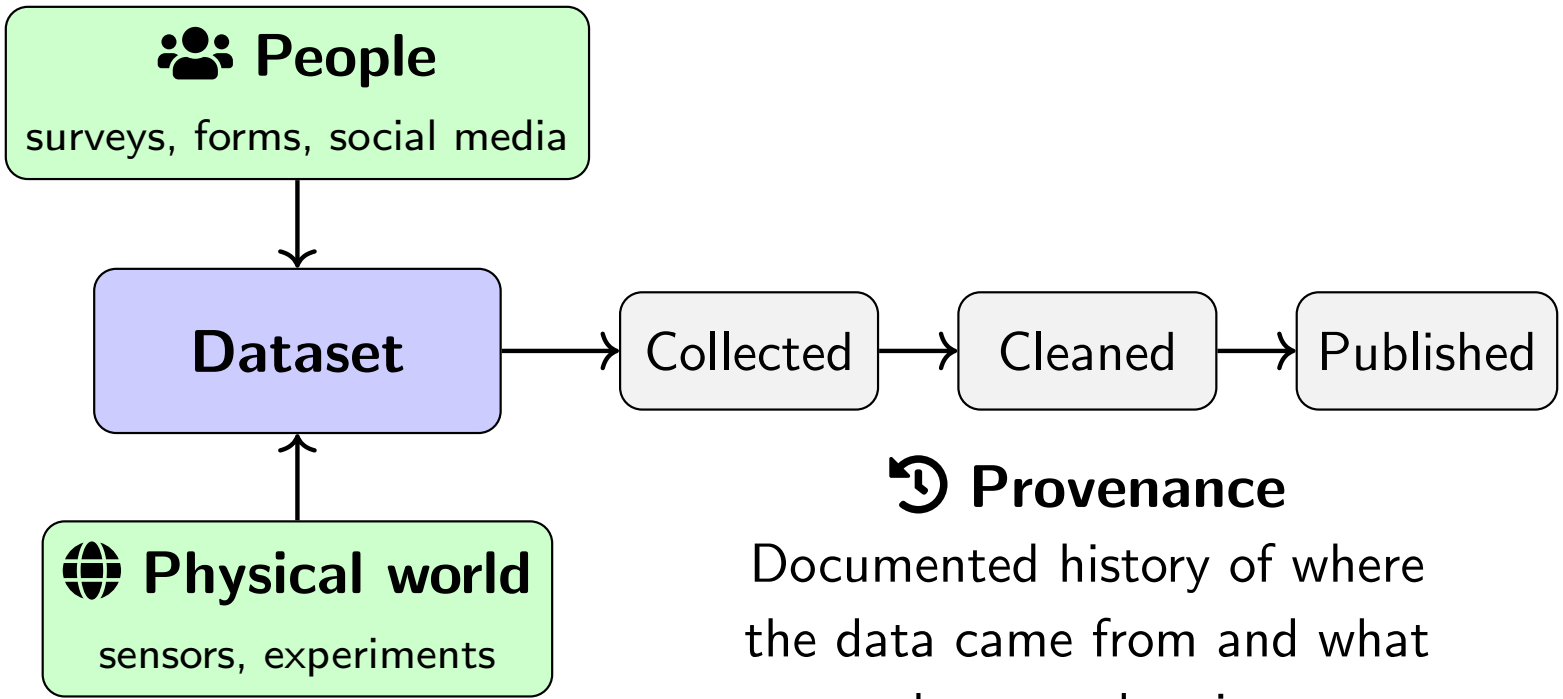


Provenance

Documented history of where the data came from and what happened to it.

Where Does the Data Come From?

Before using a dataset, ask **how it came to exist**.



Provenance

Documented history of where the data came from and what happened to it.

Accessible data is not reusable data.

Check the applicable license or terms of use.

Why Was the Data Collected?

What was the data collected for?

- **Primary data**: collected specifically for the study at hand.
- **Secondary data**: collected for another purpose, and reused.

How was the data generated?

- **Designed data**: intentionally collected for a purpose.
- **Organic data**: generated as a byproduct of an activity.

Why do these distinctions matter?

Why Was the Data Collected?

What was the data collected for?

- **Primary data**: collected specifically for the study at hand.
- **Secondary data**: collected for another purpose, and reused.

How was the data generated?

- **Designed data**: intentionally collected for a purpose.
- **Organic data**: generated as a byproduct of an activity.

Why do these distinctions matter?

Control
what is measured

Efficiency
time, cost, scale

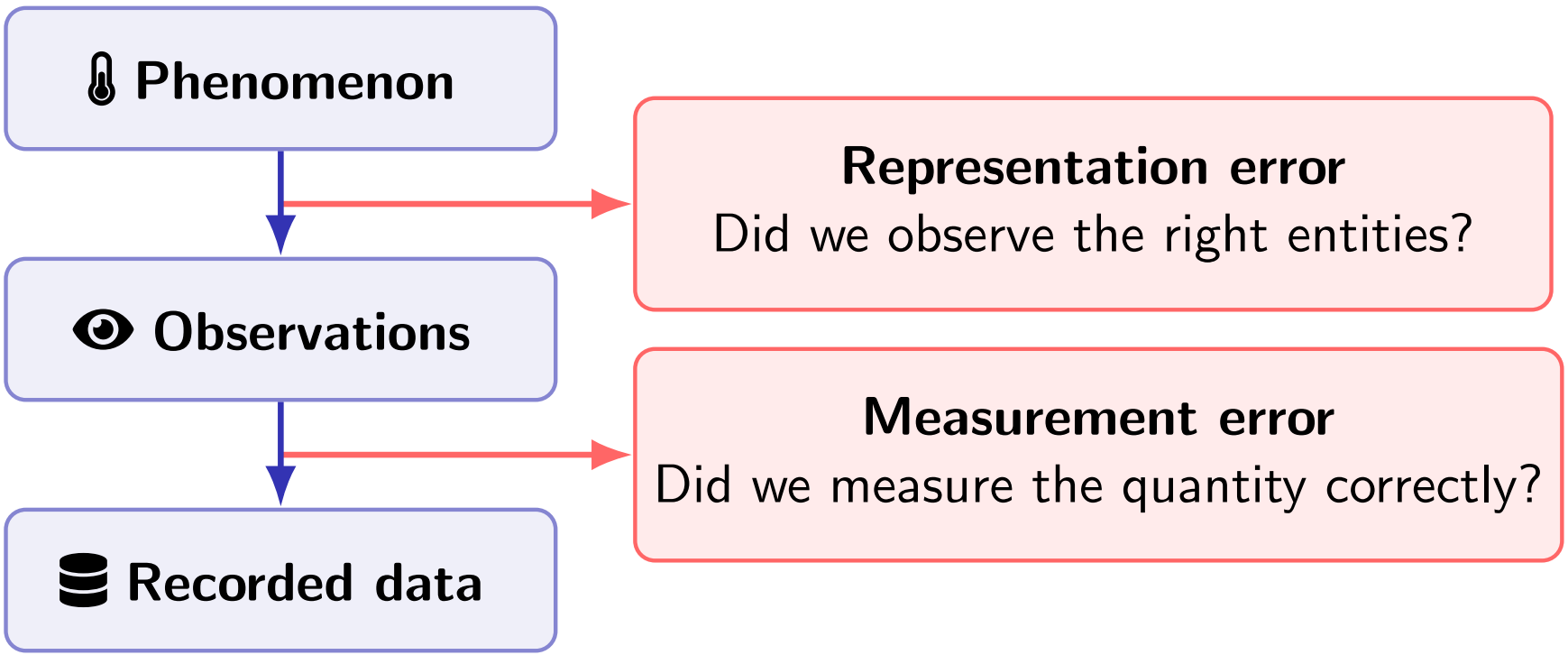
Relevance
to the study

Messiness
missing, errors, inconsistencies

Realism
real-world behavior

What Does the Dataset Represent?

A dataset is not the phenomenon itself: getting trustworthy evidence can go wrong in essentially two different ways.



Total Survey Error: Full picture by the end of the next session.

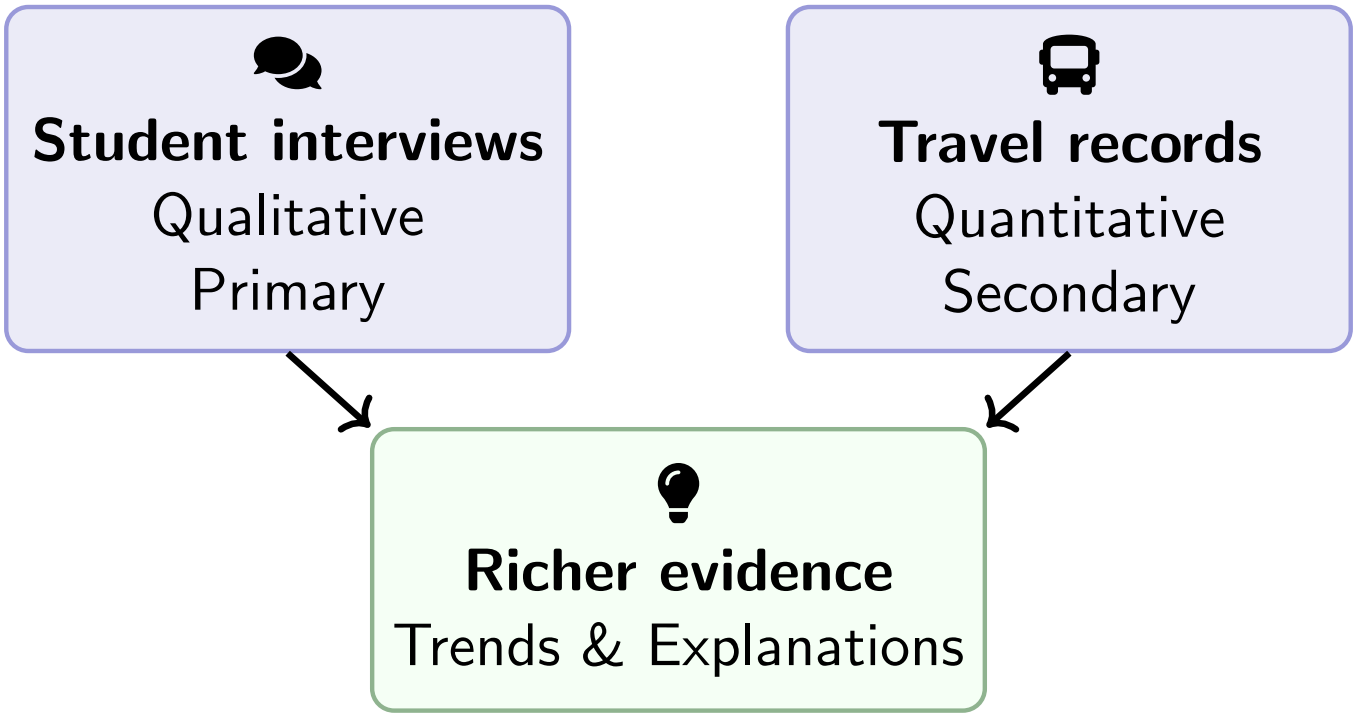
Combining Data (Commuting Patterns Example)

Combining Data (Commuting Patterns Example)

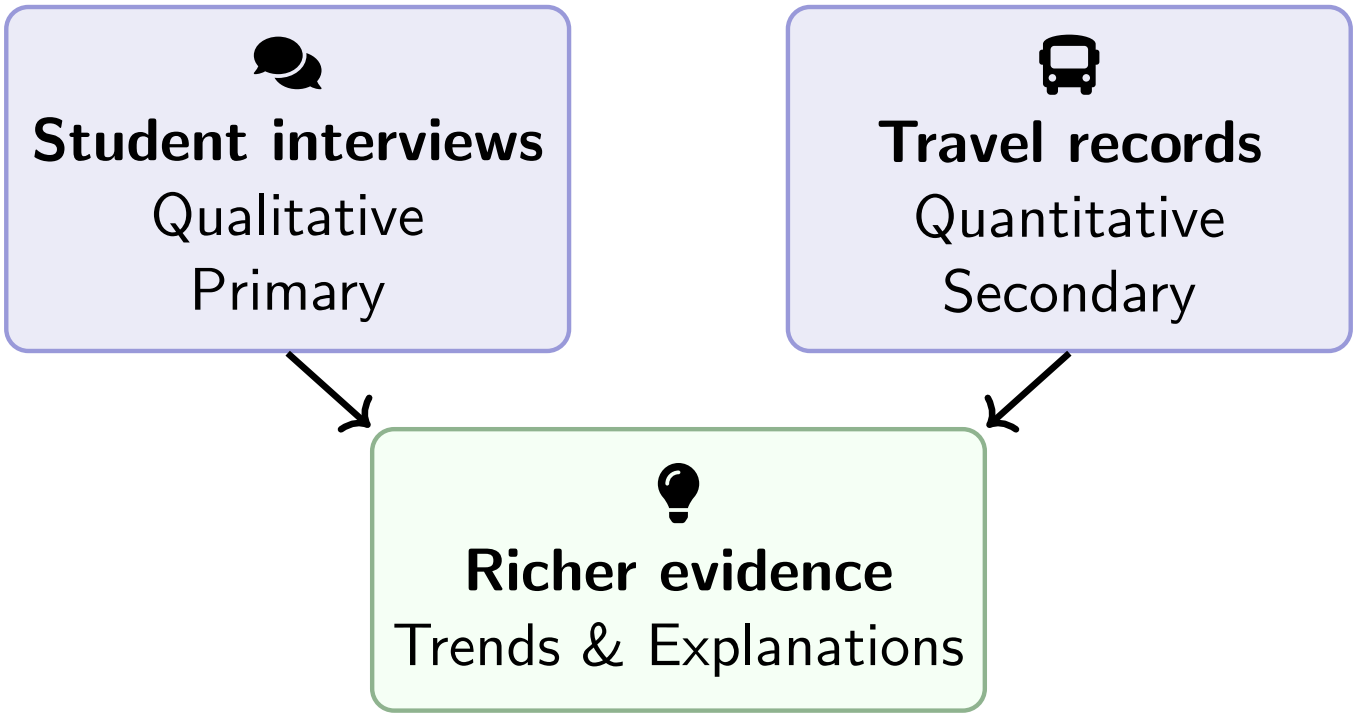
Student interviews
Qualitative
Primary

Travel records
Quantitative
Secondary

Combining Data (Commuting Patterns Example)



Combining Data (Commuting Patterns Example)



Different sources can compensate for each other's blind spots.

Before We Trust the Dataset

So far, we have asked:

1. Where did the data come from?
2. Why was it collected?
3. What does it actually represent?

Before We Trust the Dataset

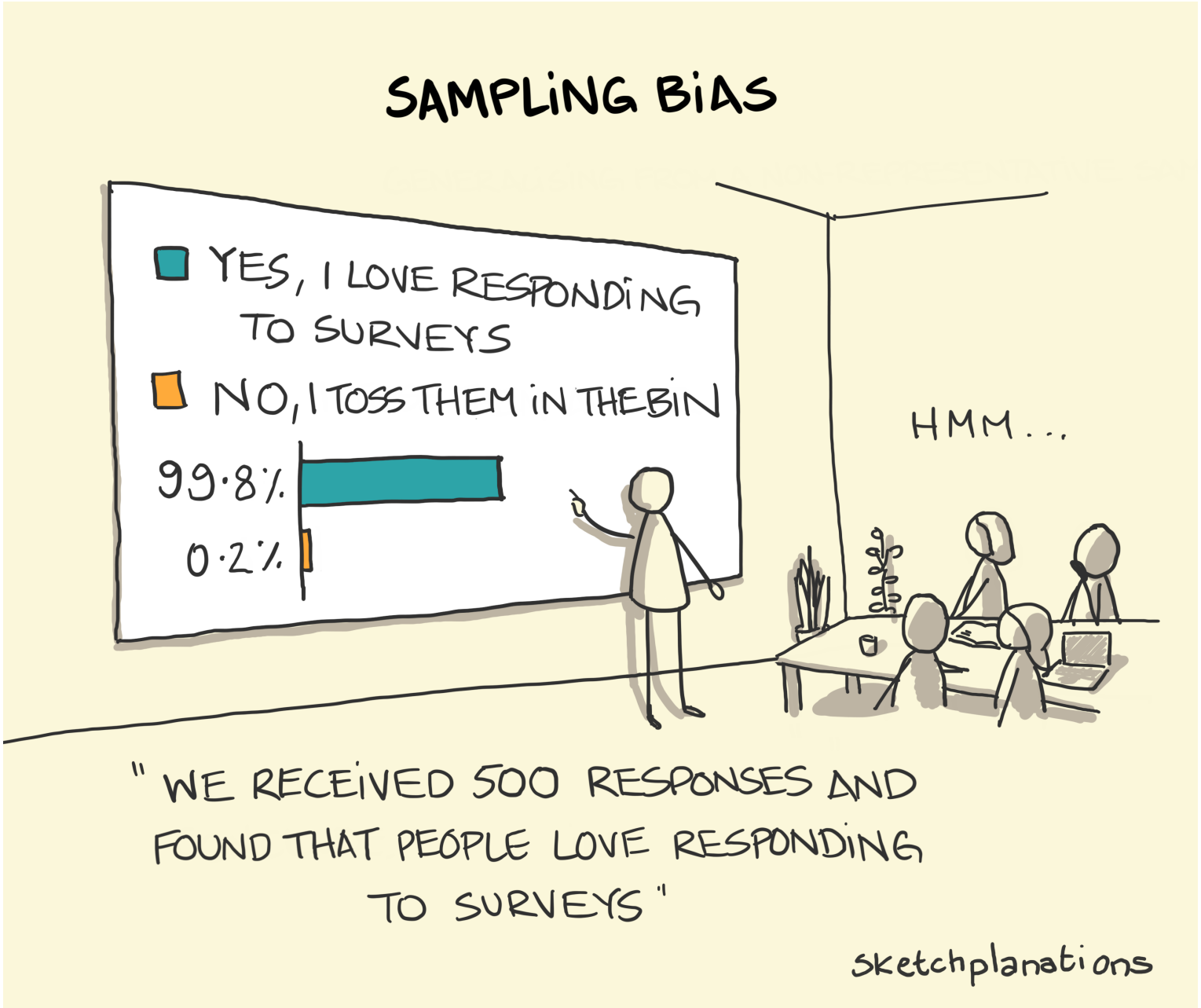
So far, we have asked:

1. Where did the data come from?
2. Why was it collected?
3. What does it actually represent?

But there is another question:

Does it cover the people, objects, or phenomena we want to study?

Scope of the Data


SAMPLING BIAS
YES, I LOVE RESPONDING TO SURVEYS
NO, I TOSS THEM IN THE BIN
99·8%
0·2%
HMM...
" WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS "
sketchplanations

Scope of the Data

Research question

What proportion of students at the university commute for more than one hour?

Before collecting data, we need to specify:

1. **Who** do we want to make a statement about?
2. **Who or what** can we actually observe?
3. **Which observations** will we collect?

Scope of the Data

Research question

What proportion of students at the university commute for more than one hour?

Before collecting data, we need to specify:

1. **Who** do we want to make a statement about?
2. **Who or what** can we actually observe?
3. **Which observations** will we collect?

⚠ **Who we want to study** need not be the same as **who we can observe**.

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

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, an event, a company, . . .

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, an event, a company, . . .

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, an event, a company, . . .

Census Data

With census data: Perfect Scope

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

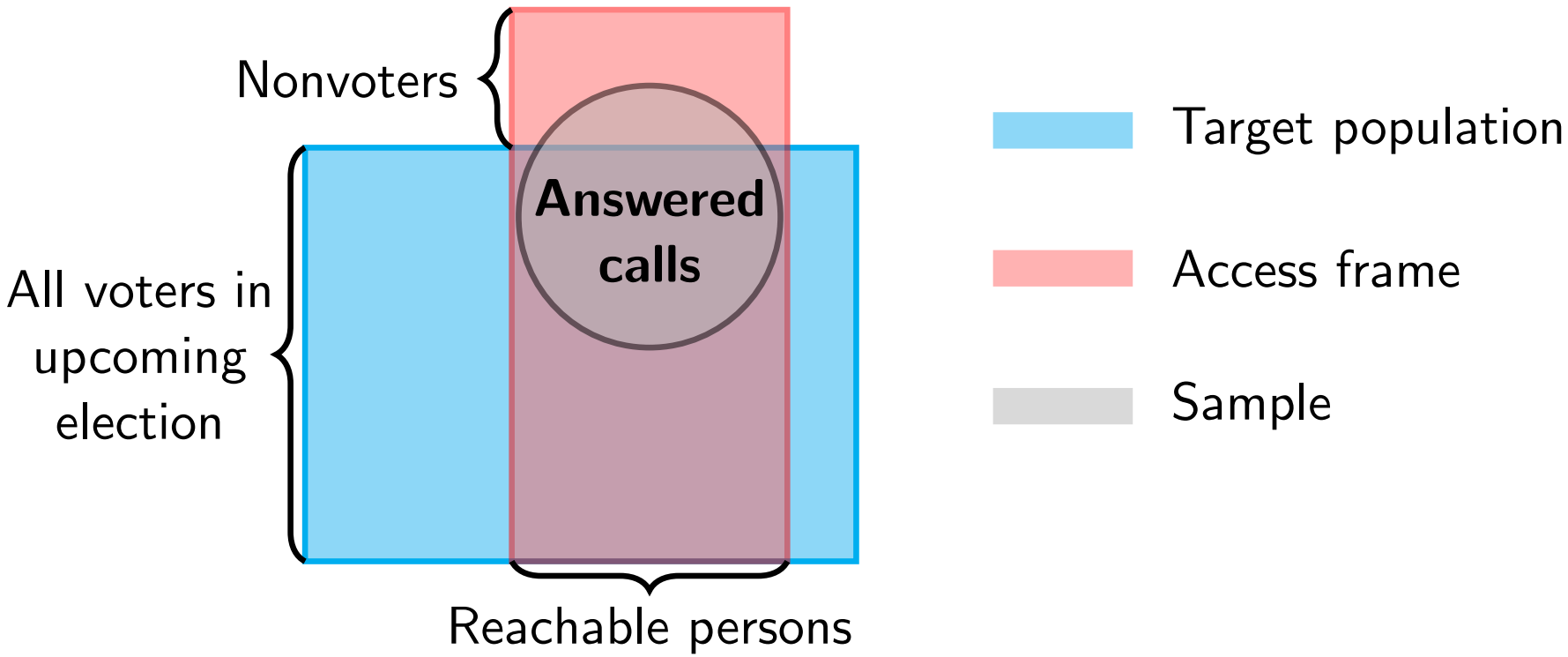
⚠ This is the ideal case. In practice, it rarely holds ...

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

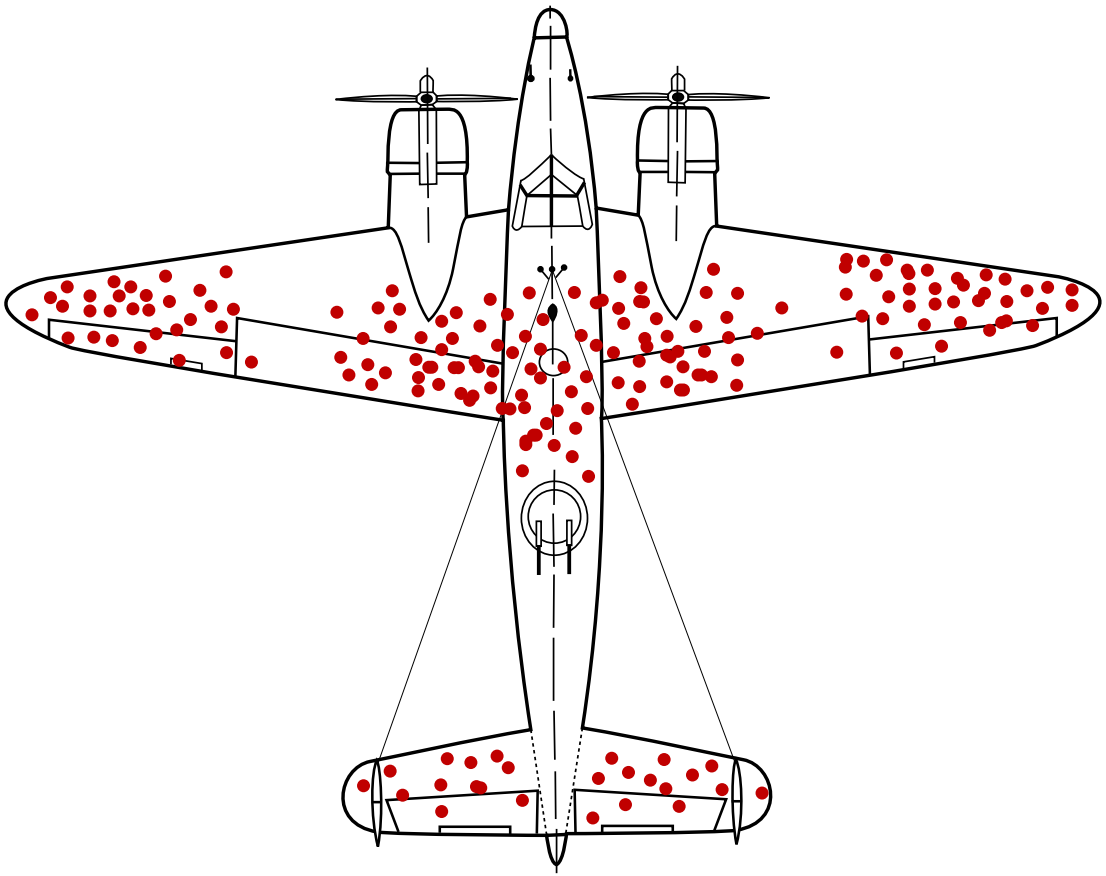
Illustration: Polls

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

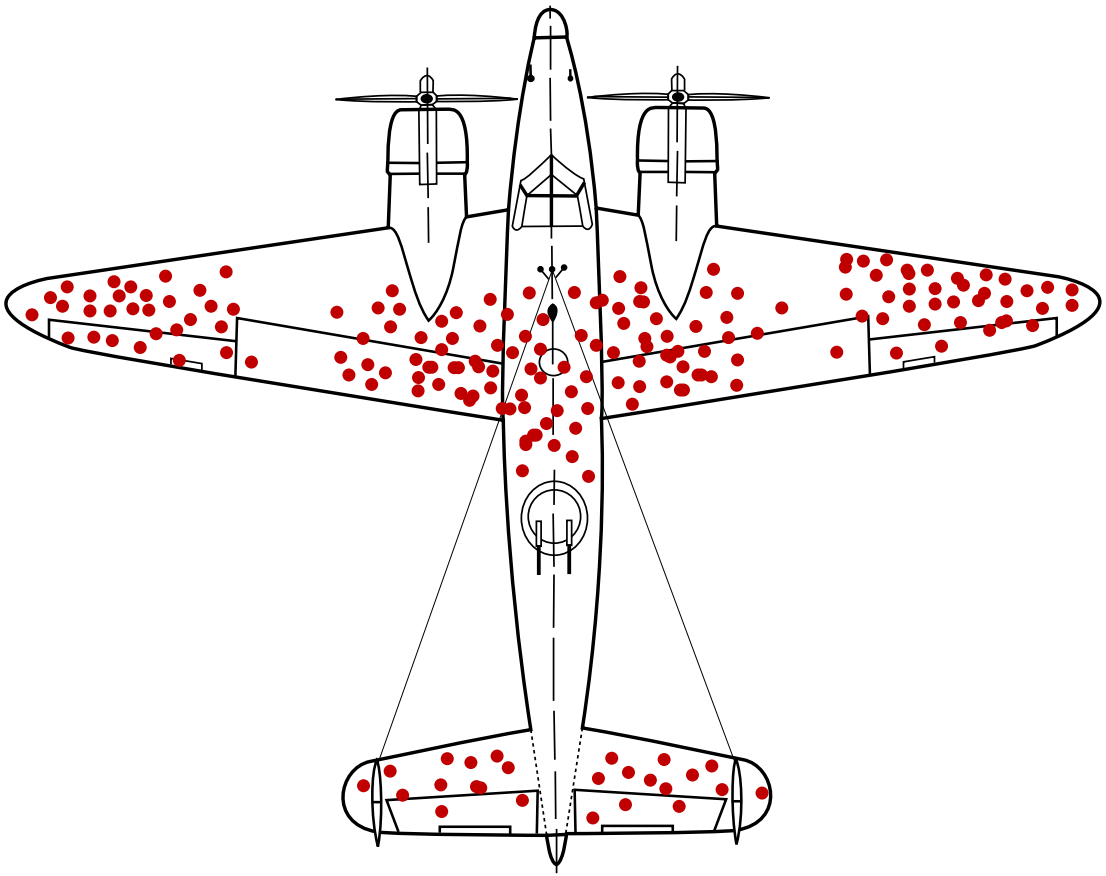


An extreme coverage issue. . .



Survivorship Bias

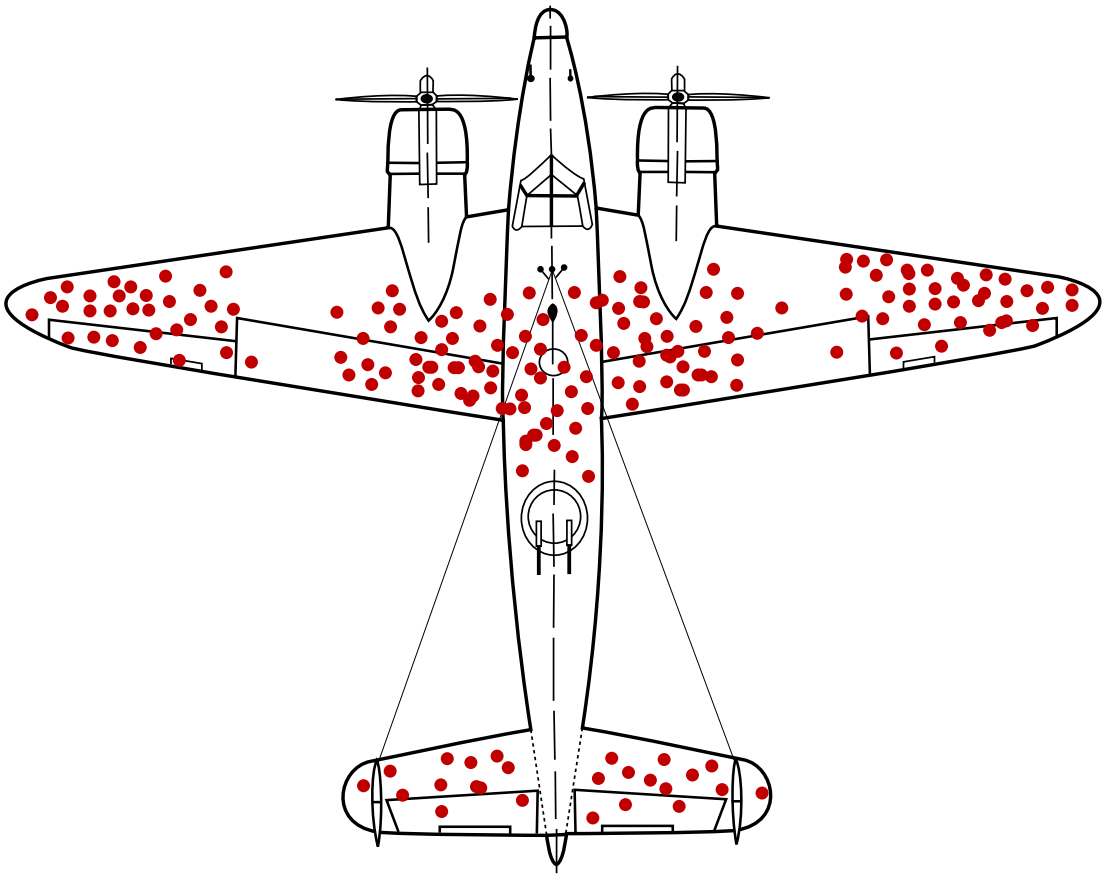
An extreme coverage issue. . .



Abraham Wald

Survivorship Bias

An extreme coverage issue. . .



Abraham Wald

Can you think of modern examples?

Scope in Natural Measurements

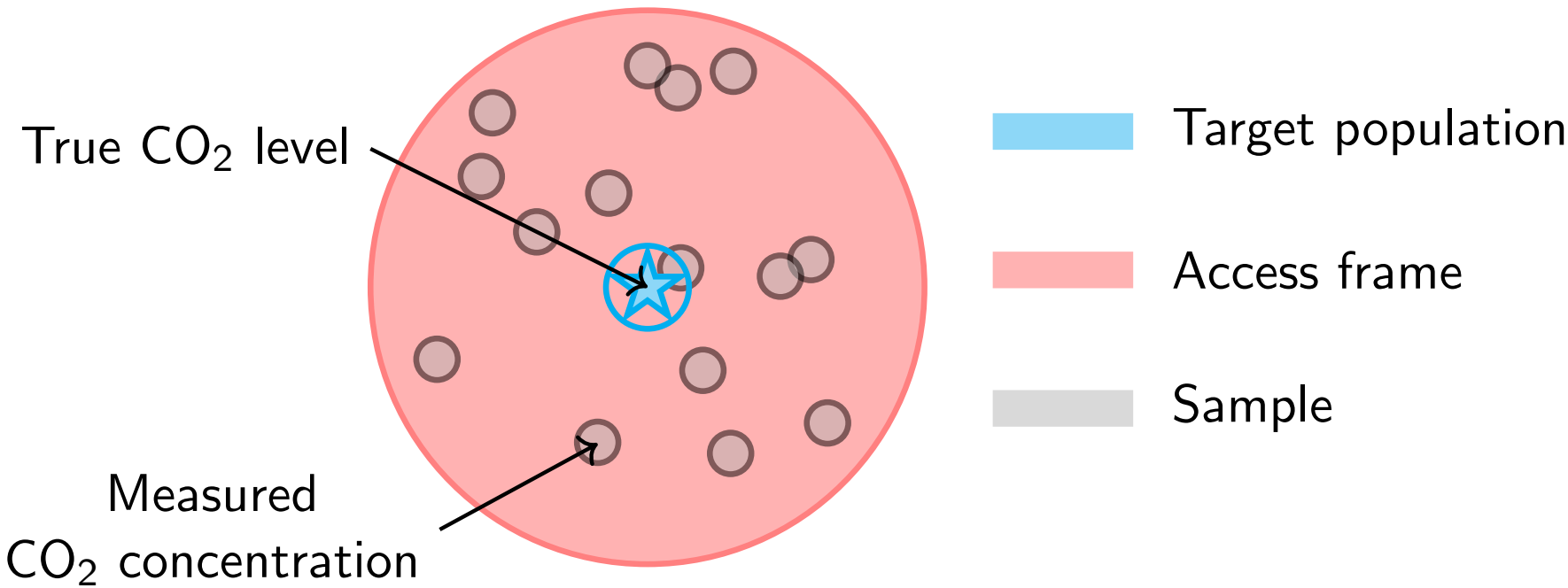
When measuring a natural phenomenon, the exact, unknown quantity of interest is called the **parameter**.

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

Illustration: CO_2 Concentration

You want to find CO_2 levels at a given location

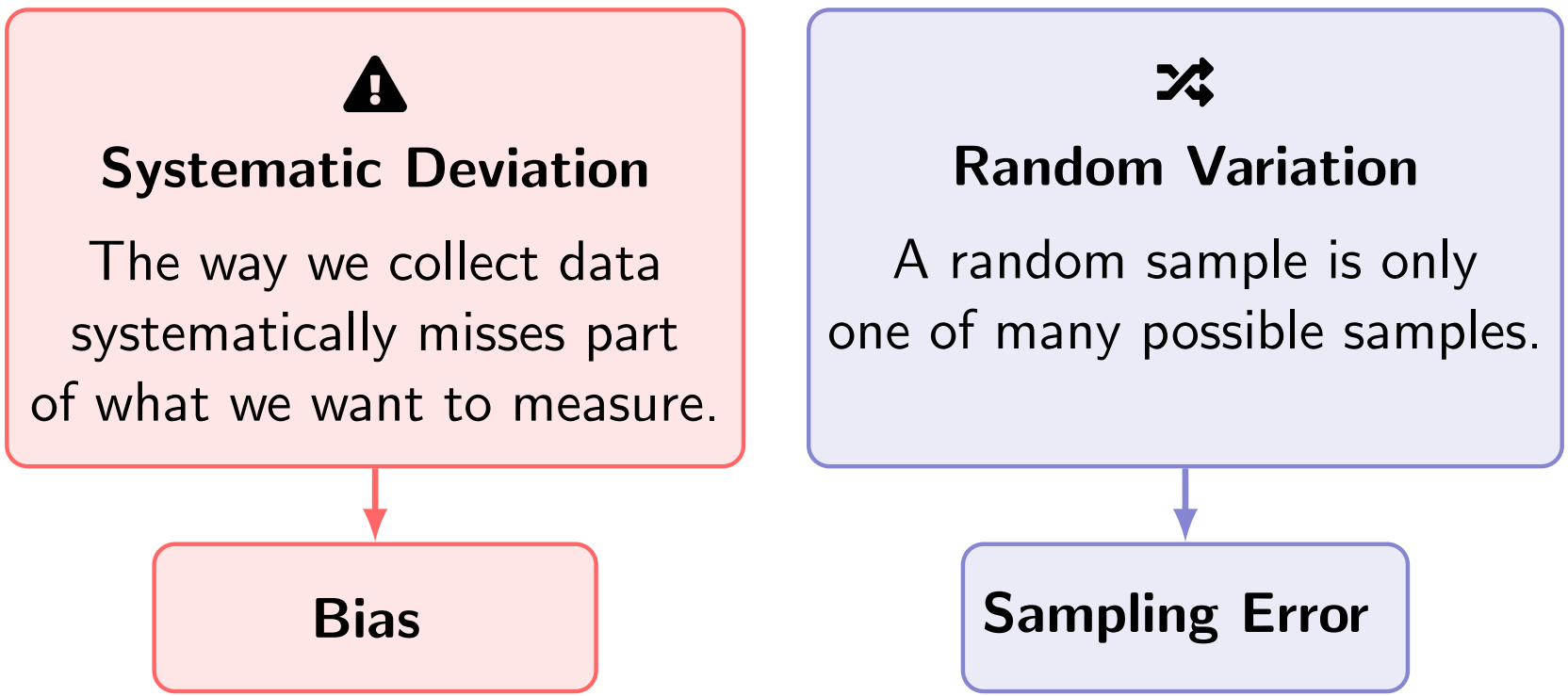
Instruments report averages at given frequency. True concentration is unknown.



Two Sources of Uncertainty

Suppose the target population and access frame coincide, and we select observations randomly.

⚠ Different random samples still give different answers.



How **large** can this random variation be?

Sampling

Suppose we want to estimate the proportion of students who commute for more than one hour.

Do we really need to ask everyone?

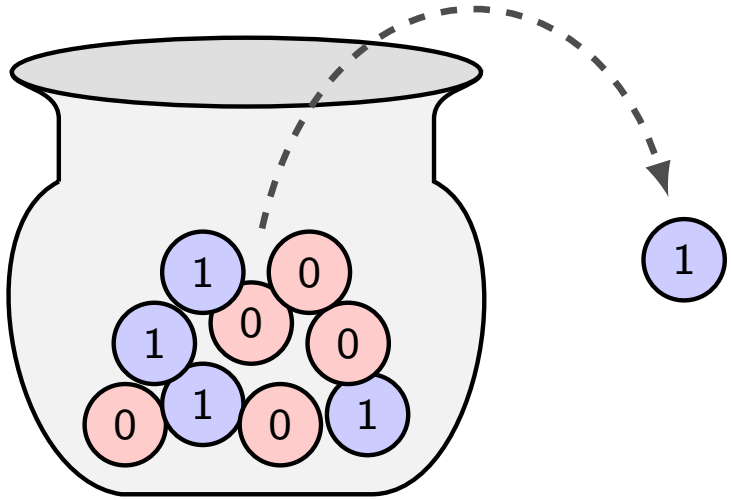
Suppose we want to estimate the proportion of students who commute for more than one hour.

Do we really need to ask everyone?

If the target population contains thousands of students, we could instead select a sample.

How much can we learn from a sample?

The Urn Model

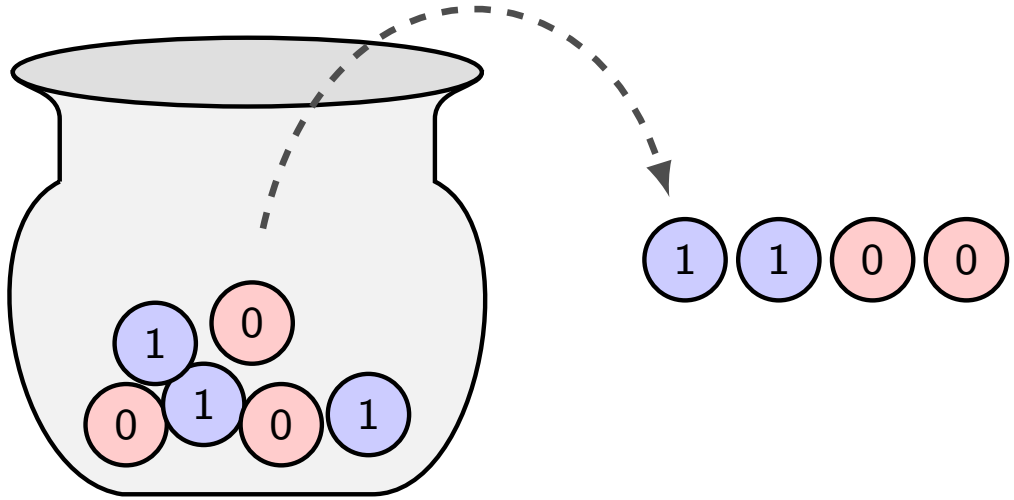


A simple way to think about sampling is to imagine that the population is an urn containing one marble for each individual.

Each marble is labelled, e.g.

$$x_j = \begin{cases} 1 & \text{if student } j \text{ commutes for more than one hour,} \\ 0 & \text{otherwise.} \end{cases}$$

Sampling as a Random Experiment



Draw n out of N marbles, without replacement.

Definition (Simple Random Sample)

A sample of n individuals from a population of N where every subset of n individuals has the same probability of being selected.

⚠ The population is fixed. The **sample** is random.

Population Parameter

Suppose the population contains N students.

Each student j has a fixed value

$$x_j = \begin{cases} 1 & \text{if student } j \text{ commutes for more than one hour,} \\ 0 & \text{otherwise.} \end{cases}$$

The **population proportion** is

$$p = \frac{1}{N} \sum_{j=1}^{N} x_j$$

p is fixed, but unknown.

From Sample to Estimator

Now select n students at random and let J_i be the population index selected at draw i. The observed value is $X_i = x_{J_i}$.

The **sample proportion** is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Example

In a sample of 100 students, 37 commute more than one hour:

$$\hat{p} = 37/100 = 0.37.$$

A Simplified Sampling Model

Urn Model: Finite population

Fixed values $x_1, \ldots, x_N$

Random sample drawn **without replacement**

↓ for the calculations below

Simplified probability model

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$

Under the iid model:

$$\mathbb{E}[X_i] = p$$

$$\text{Var}(X_i) = p(1-p)$$

⚠ Under literal sampling without replacement, the X_i are **not independent**.

Another Sample, Another Answer

Suppose the true proportion is

$$p = 0.40.$$

We repeatedly take random samples of 100 students.

Sample	1	2	3	4	5	6
$\hat{p}$	0.37	0.42	0.35	0.41	0.44	0.38

Another Sample, Another Answer

Suppose the true proportion is

$$p = 0.40.$$

We repeatedly take random samples of 100 students.

Sample	1	2	3	4	5	6
$\hat{p}$	0.37	0.42	0.35	0.41	0.44	0.38

$\hat{p}$ is itself a random variable — its distribution across repeated samples is the **sampling distribution**.

Is the Estimator Correct on Average?

Recall

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Using linearity of expectation,

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} p = p.$$

Over repeated samples, the sample proportion $\hat{p}$ is centered on p.

How Much Does the Estimator Vary?

Different samples give different $\hat{p}$. How much does $\hat{p}$ vary?

Under the iid model,

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{p(1-p)}{n}.$$

The variance of the estimator decreases $\propto 1/n$.

Standard Error

It is often more intuitive to express sampling variability on the same scale as the estimator.

$$\mathrm{SE}(\hat{p}) = \sqrt{\mathrm{Var}(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$$

But p is unknown.

We therefore estimate the standard error using $\hat{p}$:

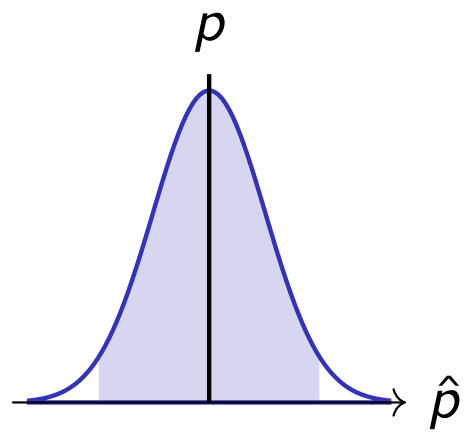
$$\mathrm{SE}(\hat{p}) \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The standard error decreases proportionally to $1/\sqrt{n}$.

The Sampling Distribution Becomes Predictable

For sufficiently large n, the Central Limit Theorem tells us that the sampling distribution of $\hat{p}$ is approximately normal.

$$\hat{p} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$



For large samples, the random estimator $\hat{p}$ has a predictable distribution around the true value p.

From Sampling Variability to an Interval

Approximately 95% of a normal distribution lies within 1.96 standard deviations of its mean:

$$\Pr\left(|\hat{p} - p| \leq 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95.$$

If p were known,

$$\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

contains p 95% of the time

plug in $\hat{p}$ →

As $\hat{p}$ is known,

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is used as an estimate

The result is the familiar 95% confidence interval.

Example: Estimating the Commuting Proportion

Suppose a sample of $n = 100$ students gives

$$\hat{p} = 0.37.$$

The estimated standard error is

$$\text{SE}(\hat{p}) \approx \sqrt{\frac{0.37(1 - 0.37)}{100}} \approx 0.048.$$

Therefore,

$$0.37 \pm 1.96 \times 0.048 \approx [0.28, 0.46].$$

Interpretation

The data are compatible with a population proportion somewhere around between 28% and 46%.

How Large Should a Sample Be?

Suppose we want a margin of error of at most e. From the confidence interval,

$$e \approx 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Solving for n gives

$$n \approx \frac{1.96^2\, p(1-p)}{e^2}.$$

Since $p(1-p) \leq 1/4$ (with $p = 0.5$) we obtain the conservative sample size:

$$n \approx \frac{1.96^2}{4e^2}.$$

Why Polls Use Around 1000 People

For a margin of error of $e = 0.03$,

$$n \approx \frac{1.96^2}{4 \times 0.03^2} \approx 1067.$$

About 1000 responses gives a margin of error of about 3 points at 95% confidence

A larger sample reduces sampling variability: $\text{SE}(\hat{p}) \propto \frac{1}{\sqrt{n}}$.

Why Polls Use Around 1000 People

For a margin of error of $e = 0.03$,

$$n \approx \frac{1.96^2}{4 \times 0.03^2} \approx 1067.$$

About 1000 responses gives a margin of error of about 3 points at 95% confidence

A larger sample reduces sampling variability: $\text{SE}(\hat{p}) \propto \frac{1}{\sqrt{n}}$.

⚠ More data does not fix systematic problems:

Coverage

Who could enter the sample?

Nonresponse

Who actually responded?

Measurement

Is the measure correct?

Conclusion

Takeaways: Representation in 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: each has strengths and limitations.
4. **Scope** matters: target population, access frame, and sample are rarely identical
5. Sampling introduces unavoidable random variation, which can be quantified
6. **More data does not fix systematic problems**

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