Data, Data Storage, Data Collection

Lecture 3: Generating and Measuring Data

Romain Pascual

MICS, CentraleSupélec, Université Paris-Saclay

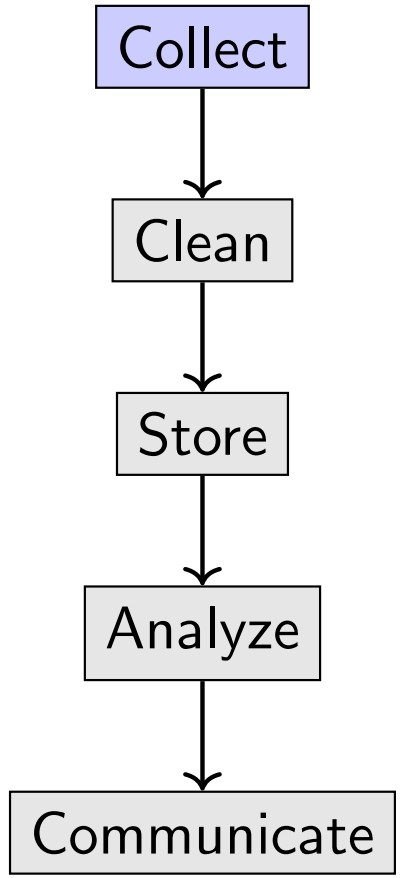
Recap

Recap: 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**

Introduction

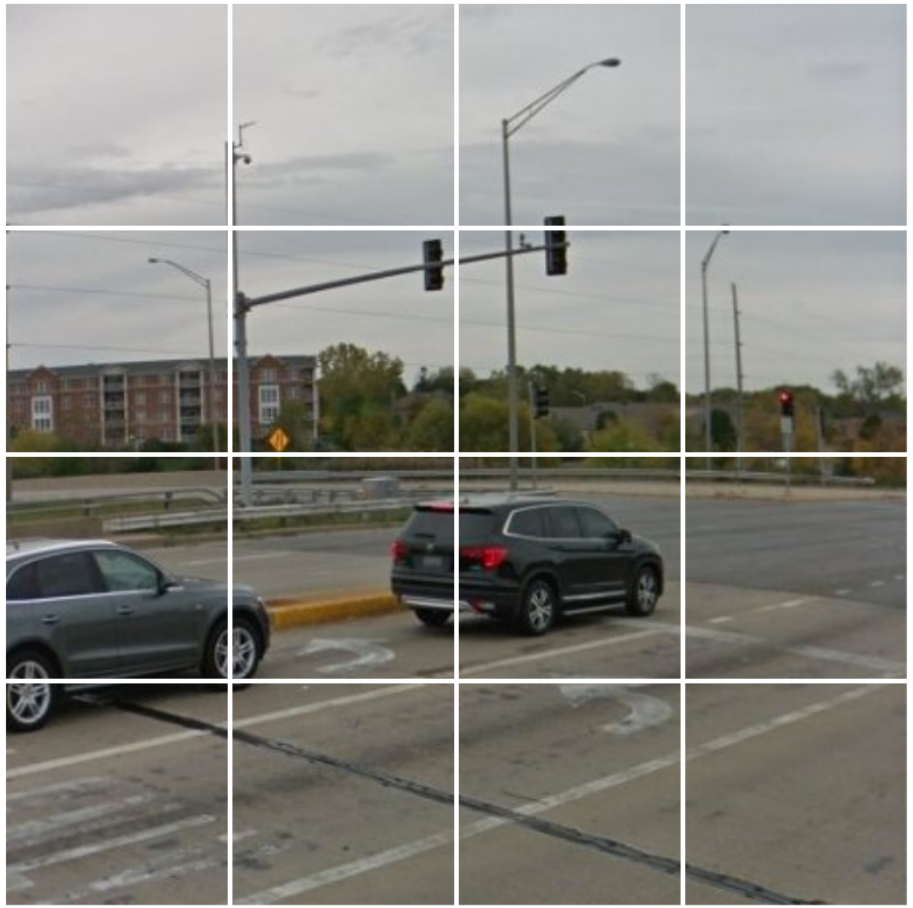
Within The Lifecycle



Select all squares with

traffic lights

If there are none, click skip



SKIP

Session Objectives

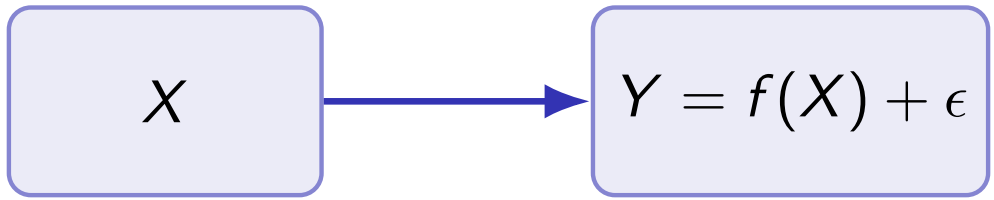
By the end of this session, you should be able to:

- Distinguish observational, experimental, and simulated data, and say what each lets you see and what it hides
- Use Monte Carlo simulation to estimate a quantity, reusing the sampling framework from Session 2
- Decompose measurement error into bias and variance (MSE)
- Explain, using an example, why sampling rate limits what a continuous signal can reveal
- Situate representation-side and measurement-side errors within the Total Survey Error framework

Producing an Observation

From Phenomenon to Data

Let X denote the phenomenon or quantity we care about.



f is the process (instrument, protocol, model) used to turn X into a recorded value, and ϵ is whatever f leaves out or gets wrong.

Two separate questions follow:

1. Which X do we actually get to observe?
2. How well does Y approximate the true value?

Some Ways to Generate Data

How do we get access to X in the first place?

Observational

Find it

Experimental

Make it

Simulated

Model it

Each choice determines what conclusions the resulting data can support.

Observational Data

Find an existing phenomenon, and record it as it occurs.

Definition (Observational data)

X exists, embedded in some ongoing process. We record $Y = f(X) + \epsilon$.

Target population, access frame, and sample can all diverge.

- Coverage gap: who is even reachable?
- Nonresponse gap: who actually responds?

Observational Data: Examples

Observational data reaches us through different mechanisms.

Survey

f = a person's self-report

- Study-habits questionnaire

Each is a different way of implementing f.

Observational Data: Examples

Observational data reaches us through different mechanisms.

Survey

f = a person's self-report

- Study-habits questionnaire

Sensor

f = a physical instrument

- Pedometer: number of steps from acceleration

Each is a different way of implementing f.

Observational Data: Examples

Observational data reaches us through different mechanisms.

Survey

f = a person's self-report

- Study-habits questionnaire

Sensor

f = a physical instrument

- Pedometer: number of steps from acceleration

Log

f = an automatic digital trace

- App logs: button clicks aggregated into usage statistics

Each is a different way of implementing f.

Observational Data: Advantages and Challenges

Advantages

- **Real-world validity**: reflects behaviour as it naturally occurs
- **Feasibility**: works as long as an instrument can record the measure

Challenges

- **Selection bias**: nothing about X was controlled
- **No control**: hides whether some phenomenon caused others

So far, we have **found** data occurring in the world.
What if we want to test hypotheses and explore phenomena that have not occurred?

So far, we have **found** data occurring in the world.
What if we want to test hypotheses and explore phenomena that have not occurred?

Experimental data can stand in instead.

Causation vs Correlation

So far, we have only **watched** X and Y move together. Is that enough to say one causes the other?

Example

Ice cream and shark attacks both increase in summer.

Causation vs Correlation

So far, we have only **watched** X and Y move together. Is that enough to say one causes the other?

Example

Ice cream and shark attacks both increase in summer.

Definition (Correlation)

A statistical relationship between two variables: they change together.

Definition (Causation)

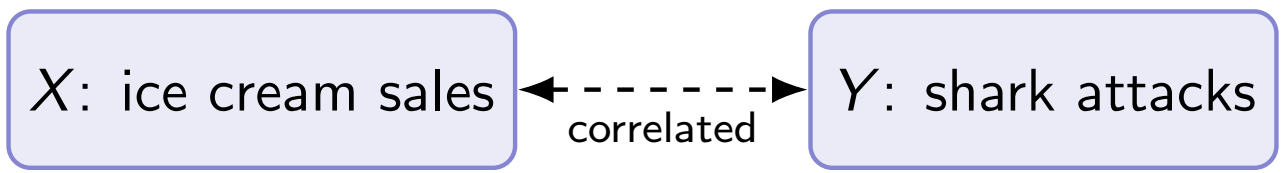
A change in one variable (**cause**) produces a change in another (**effect**).

Confounding Variables

Definition (Confounding variable)

A hidden factor that affects both X and Y at once, making them move together without either causing the other.

Ice cream and shark attacks

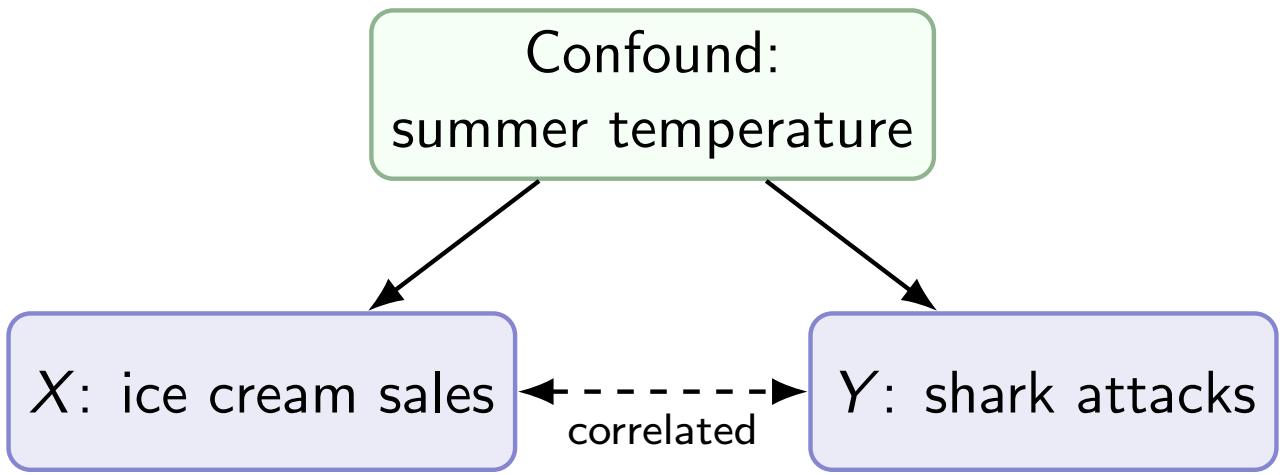


Confounding Variables

Definition (Confounding variable)

A hidden factor that affects both X and Y at once, making them move together without either causing the other.

Ice cream and shark attacks



What if, instead of watching X occur, we chose X ourselves?

Experimental Data

Create the conditions, then observe the outcome.

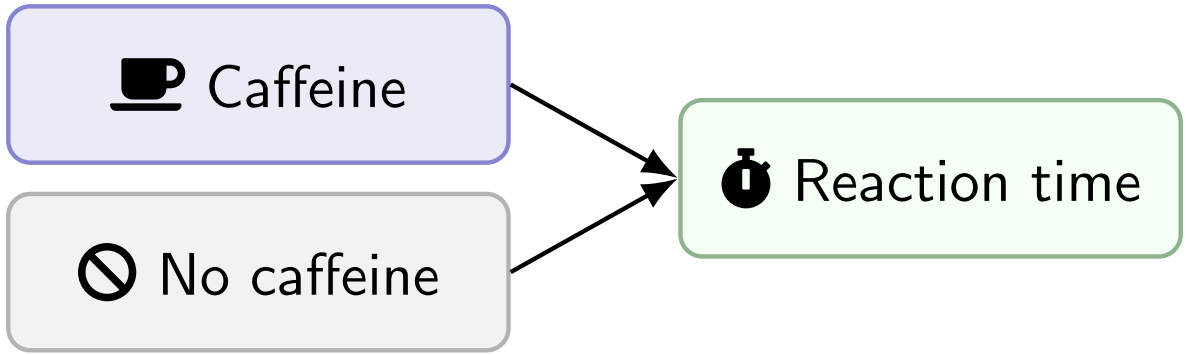
Definition (Experimental data)

The researcher chooses which X occurs (**independent variable**), then observes the outcome Y (**dependent variable**).

By choosing X ourselves, we can break the link between a confound and X, since we no longer rely on X occurring naturally.

⚠ This only works if X is chosen **the right way**. Choosing it carelessly can bring the same problem right back.

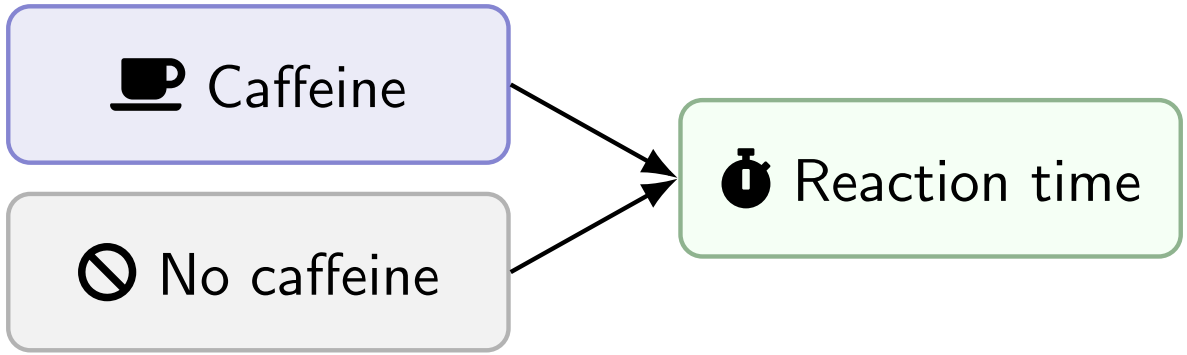
Experimental Data: Example



Studying caffeine and reaction time

- Independent variable: caffeine intake
- Dependent variable: reaction time

Experimental Data: Example

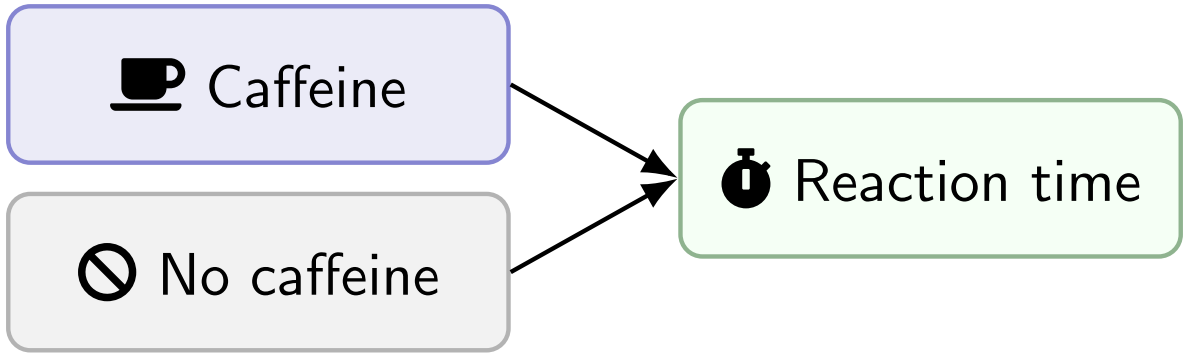


Suppose volunteers are simply asked whether they would like coffee before the test.

Studying caffeine and reaction time

- Independent variable: caffeine intake
- Dependent variable: reaction time

Experimental Data: Example



Suppose volunteers are simply asked whether they would like coffee before the test.

Studying caffeine and reaction time

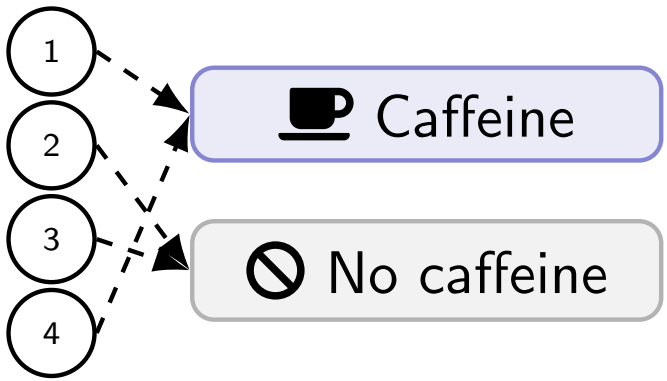
- Independent variable: caffeine intake
- Dependent variable: reaction time
- Confounders: sleep quality, age, or time of day

Same failure as ice cream and sharks.

⚠ Choosing X ourselves did not, by itself, solve the issue.

Random Assignment

Assign each participant's condition **at random**, rather than letting them choose or self-select.



Random assignment

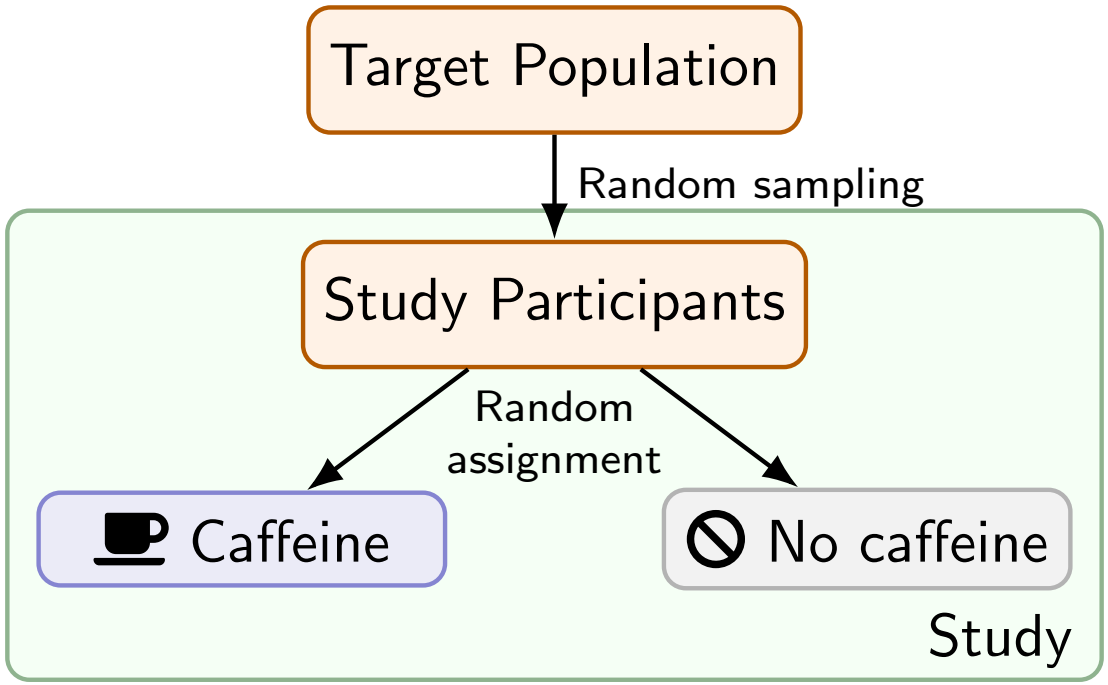
Randomly assigning X to participants spreads confounders evenly across groups, **on average**.

Difference in reaction time can no longer be explained away by who happened to be better rested.

Document confounders anyway: randomization balances them on average, not in every single sample.

Random Sampling vs Random Assignment

Random assignment fixes the comparison **within** the study.



⚠ Randomizing the **condition** does not make the **participants** representative — an experiment still has a target population, an access frame, and a sample.

Experimental Data: Advantages and Challenges

Advantages

- **Causal inference**: random assignment neutralizes confounding by design
- **Replicability**: the design can be repeated to verify results
- **Isolation**: control specific variables one at a time

Challenges

- **External validity**: fixes comparability, not who was recruited
- **Ethical constraints**: some manipulations may be unethical
- **Practical limits**: cost and feasibility restrict achievable sample size

Some experiments are too costly, slow, or unethical

Some experiments are too costly, slow, or unethical

Simulated data can stand in instead.

Simulated Data

Use a model to generate synthetic observations.

Definition (Simulation)

A study where a model is used to imitate a system and generate synthetic outputs (Y) from chosen inputs (X).

Modeled as:

$$Y = f(X, \theta) + \epsilon$$

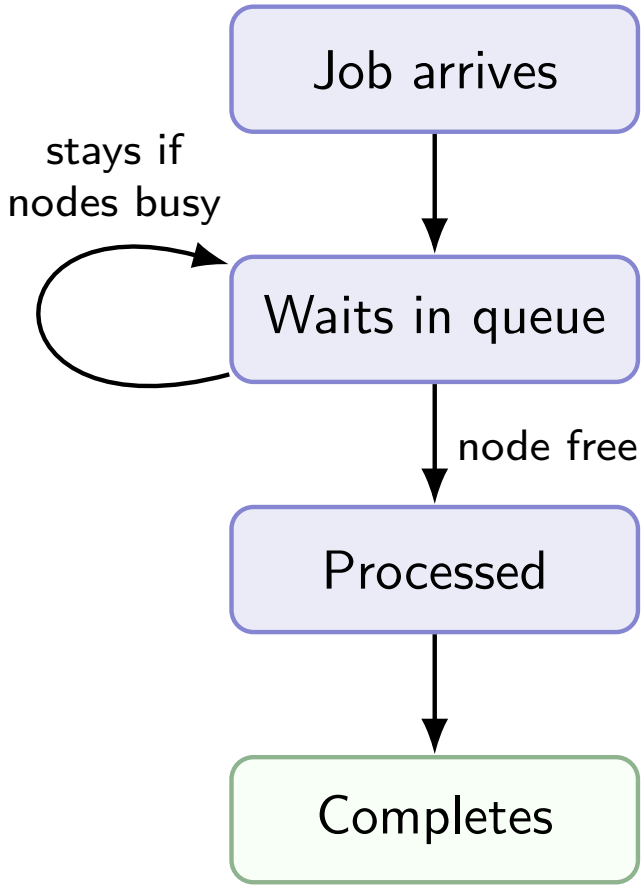
where θ are the model's parameters or assumptions.

⚠ f is no longer reality but a **model approximation** of it.

Simulated Data: Example

Stress-testing a compute cluster

Simulated jobs arrive at random, queue if all nodes are busy, and get processed



Running many jobs through the model, many times, traces out a distribution of wait times.

⚠ There is no real cluster to test yet: only a model of the system we want to understand.

Monte Carlo Simulation

A simulation lets us generate synthetic draws directly. As with any sample, the question remains: **how many draws do we need to trust the result?**

Suppose our model defines a probability distribution $\mathbb{P}$ and we generate independent observations

$$Z_1, \ldots, Z_n \sim \mathbb{P}.$$

Suppose we want to estimate the probability of an event $\mathcal{A}$. For each simulation, record whether the event occurred:

$$X_i = \begin{cases} 1 & \text{if } Z_i \in \mathcal{A}, \\ 0 & \text{otherwise.} \end{cases}$$

This is exactly the same setup as in Session 2, with "draws from a model" instead of "draws from a population."

Monte Carlo: From Simulations to an Estimate

Since

$$X_i \sim \text{Bernoulli}(\Pr(\mathcal{A})),$$

and

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \Pr(\mathcal{A}) \quad \text{as } n \to \infty,$$

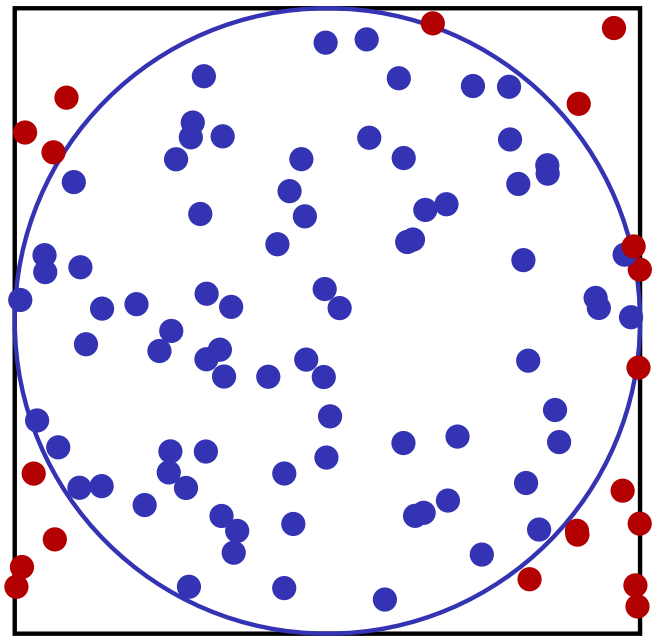
everything derived for $\hat{p}$ — $\mathbb{E}[\hat{p}]$, $\text{Var}(\hat{p})$, $SE(\hat{p})$, confidence intervals — applies unchanged.

Monte Carlo

Generate n independent realizations from a model, and use their empirical behaviour to estimate a quantity of interest with error $\propto 1/\sqrt{n}$.

Example: Estimating π

1. Generate points uniformly in $[-1, 1] \times [-1, 1]$ (area 4)
2. Check if $x^2 + y^2 \leq 1$ (area π)
3. $X_i = 1$ if point i lands in the disk



Then $\hat{p} = \frac{1}{n} \sum X_i$ estimates $\pi/4$, so

$$\pi \approx 4\hat{p}.$$

More details in the lab.

Simulated Data: Advantages and Challenges

Advantages

- Explore many scenarios quickly ("what-if" cases)
- No ethical constraints on the manipulated conditions
- Useful when exact real-world solutions are unavailable

Challenges

- Quality depends entirely on the model's assumptions
- Needs validation against real data
- Large simulations may require significant compute

We now have an observation.

We now have an observation.

Is it a good one?

Measurement Error

Operationalization

Definition (Operationalization)

Operationalization is the choice of how a phenomenon (X) will be represented and measured in data.

Does commuting affect student stress?

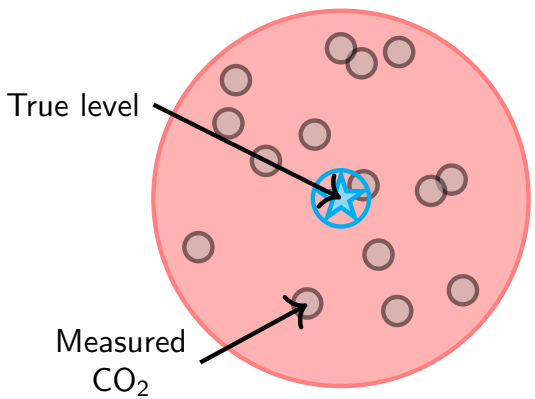
- Phenomenon: **stress**
- Possible operationalizations: questionnaire score, cortisol level, sleep duration, heart rate, absenteeism, ...
- Observation: e.g. a questionnaire score of 17/40

Different operationalizations capture different aspects of the same phenomenon.

⚠ A measurement can be precise and unbiased **for what it measures**, yet still be a poor representation of the phenomenon.

From Scope to Measurement

Session 2 extended the scope diagram to a single measured quantity:



- **Target** shrank to one unknown true value
- **Frame** became the instrument's accuracy
- **Sample** were the measurements taken

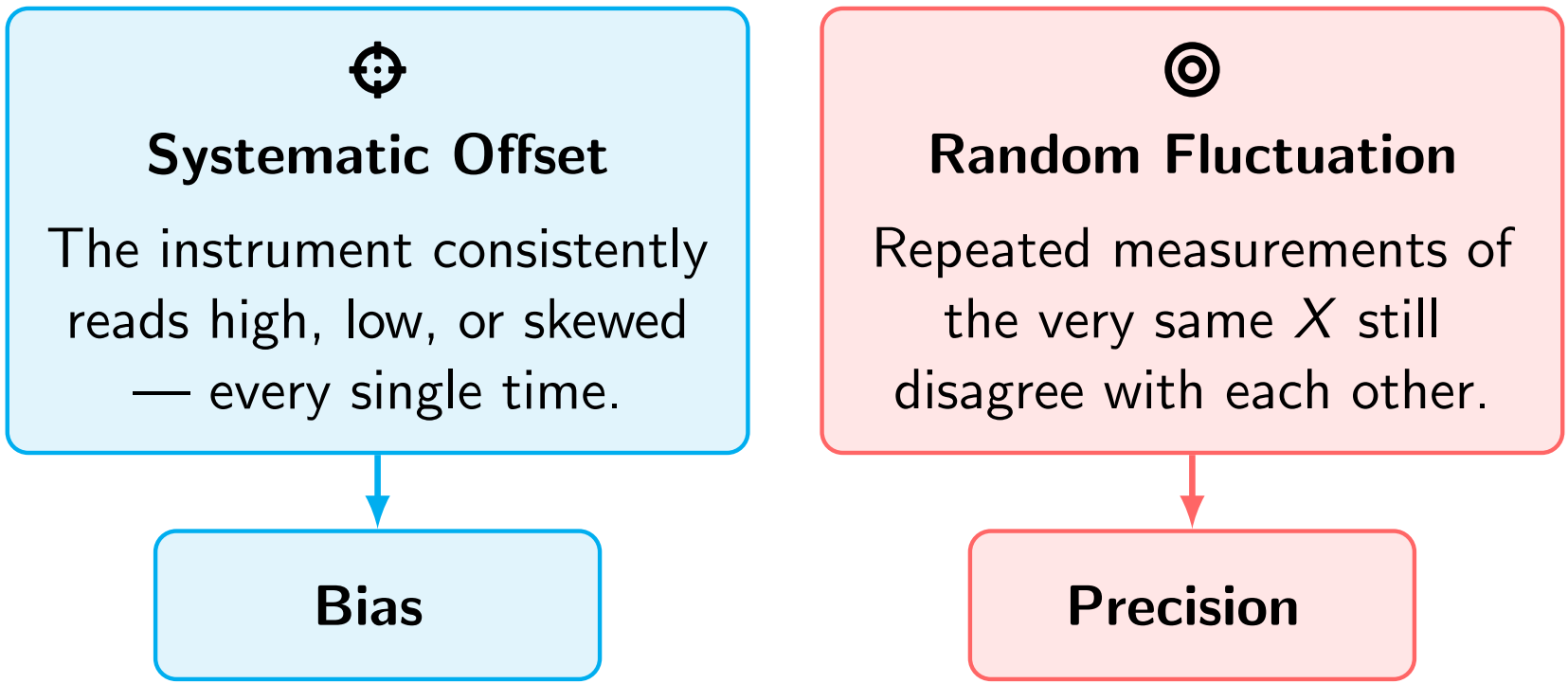
Like a dartboard: the instrument throws darts, the true value is the unseen bullseye.

We now revise the terminology and properly name the **instrument's** half of the story.

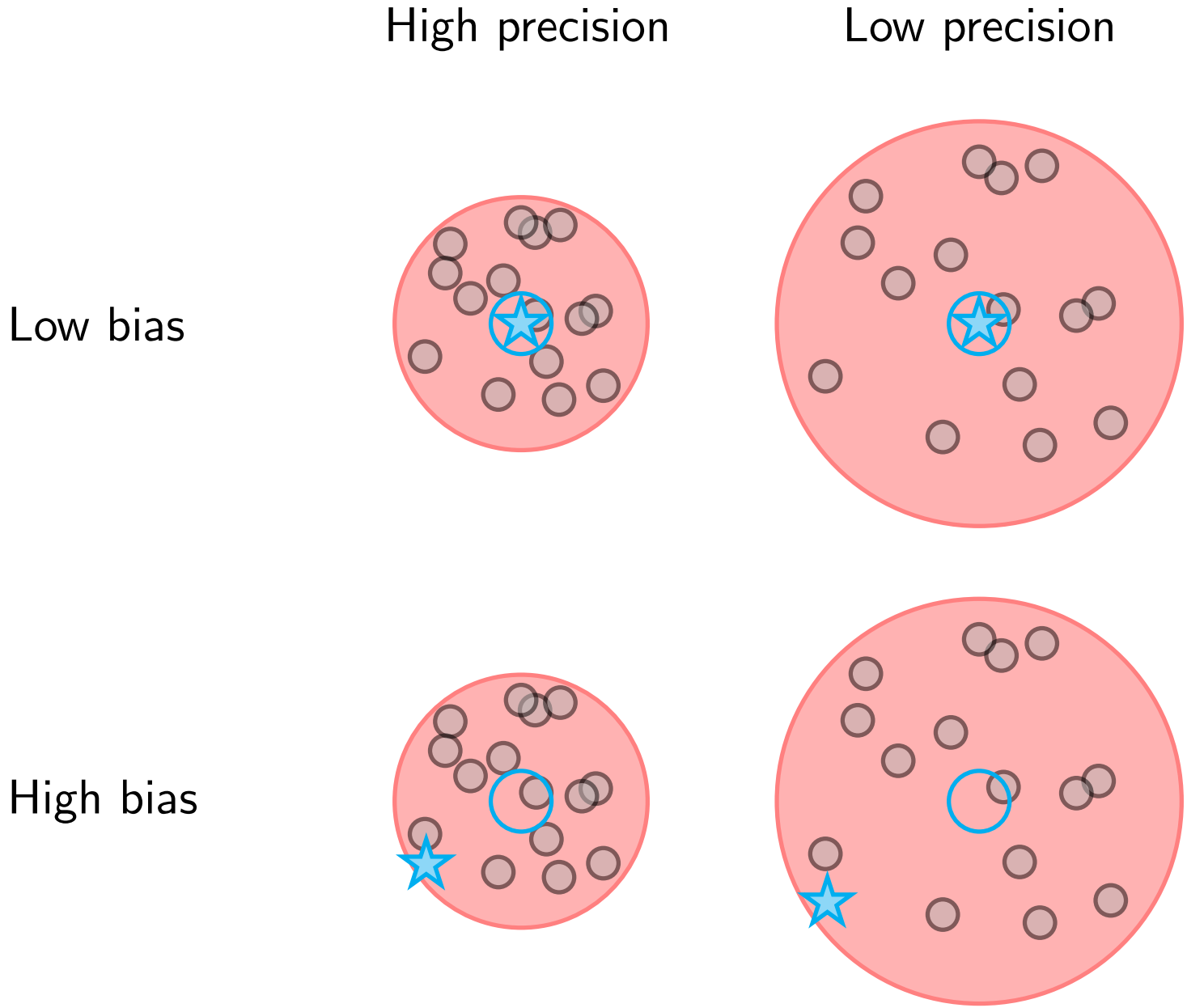
Bias and Precision

Even once we have fixed what to measure, **observing** something and **successfully measuring it** are not the same thing:

$$Y = f(X) + \epsilon$$



⚠ A different failure mode from sampling error: the problem lives in f, not in which X was obtained.


High precision
Low precision
Low bias
High bias

From Bias and Precision to MSE

We can formalize the dartboard using the **Mean Squared Error** of an estimator $\hat{\theta}$ of a true value θ — here, θ plays the role of X, and $\hat{\theta}$ the role of Y:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right].$$

Decomposing around $\mathbb{E}[\hat{\theta}]$:

$$\mathrm{MSE}(\hat{\theta}) = \underbrace{\left(\mathbb{E}[\hat{\theta}] - \theta\right)^2}_{\mathrm{Bias}^2} + \underbrace{\mathrm{Var}(\hat{\theta})}_{\mathrm{Variance}}.$$

$$\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Variance}$$

This decomposition applies to **any** measurement, regardless of how it was produced.

MSE Across the Three Modes

Observational

CO_2 sensor: a miscalibration is bias; instrument noise is variance.

Experimental

Reaction-time trials vary trial to trial (variance); a slow stopwatch adds bias.

Simulated

Monte Carlo estimate $\hat{p}$: variance shrinks with n; estimator bias is typically negligible.

Same decomposition, three different sources of randomness.

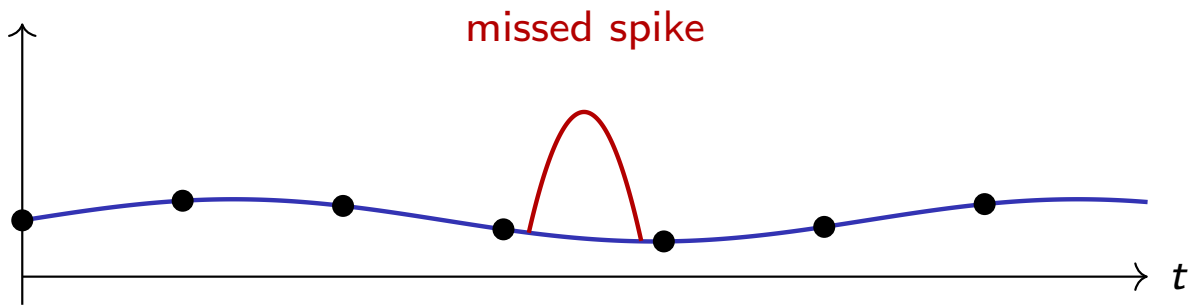
⚠ A simulation's $\hat{p}$ can be unbiased **as an estimator of the model's own probability** — and still be a poor stand-in for reality, if the model itself is a bad operationalization of X.

Sampling a Continuous Signal

So far, error has been about **value**. But error can also come from **when** we look.

Recall CO_2 concentration: the true value $X(t)$ varies continuously, but an instrument only ever reports it at fixed intervals:

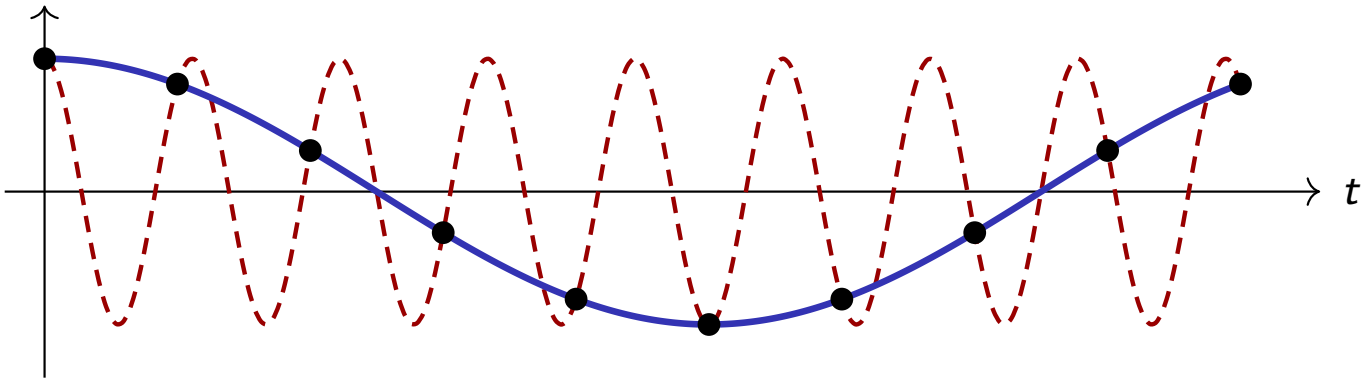
$$X(t) \longrightarrow X(0), X(\Delta t), X(2\Delta t), \ldots$$



Whatever happens **between** two readings is simply never recorded.

Example

Consider a 9 Hz signal sampled at 10 Hz



- - - true 9 Hz signal — apparent 1 Hz signal • samples at 10 Hz

By 2π-periodicity and symmetry, sampling $\cos(2\pi \cdot 9t)$ at $t_n = n/10$ gives

$$\cos(2\pi \cdot 9n/10) = \cos(2\pi n/10)$$

It produces **exactly** the same samples as a 1 Hz signal.

Aliasing

Theorem (Nyquist–Shannon sampling theorem)

If a continuous-time signal has no frequencies above B Hz, it can be reconstructed exactly from uniform samples taken at a rate

$$f_s > 2B.$$

Definition (Aliasing)

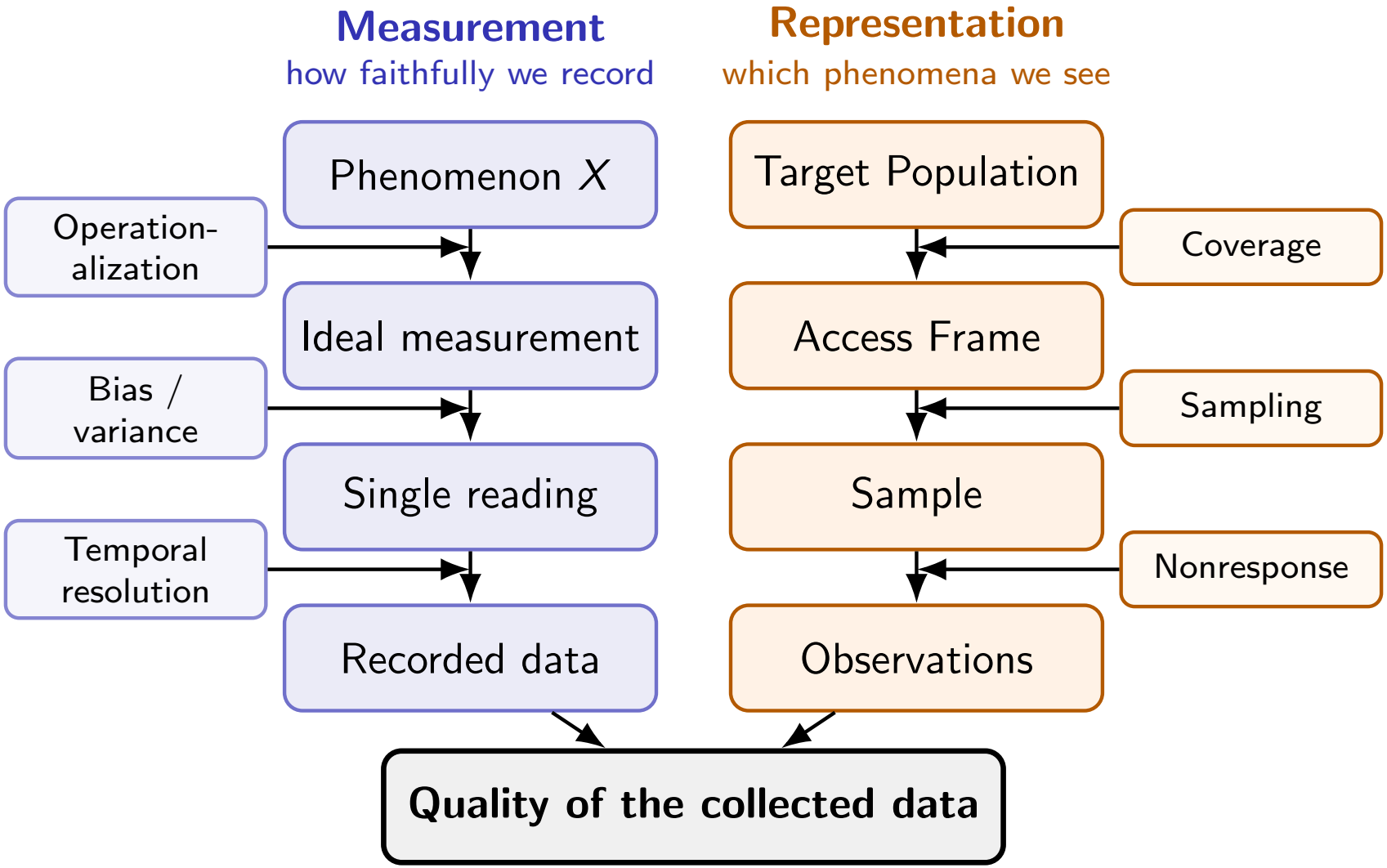
If $f_s \leq 2B$, different signals can produce the **same** sequence of samples. This ambiguity is **aliasing**

Exactly what happened in the 9 Hz/10 Hz example.

⚠ Aliasing is not bias or random noise: even a perfectly accurate instrument can misrepresent a signal if it is sampled too slowly.

Conclusion

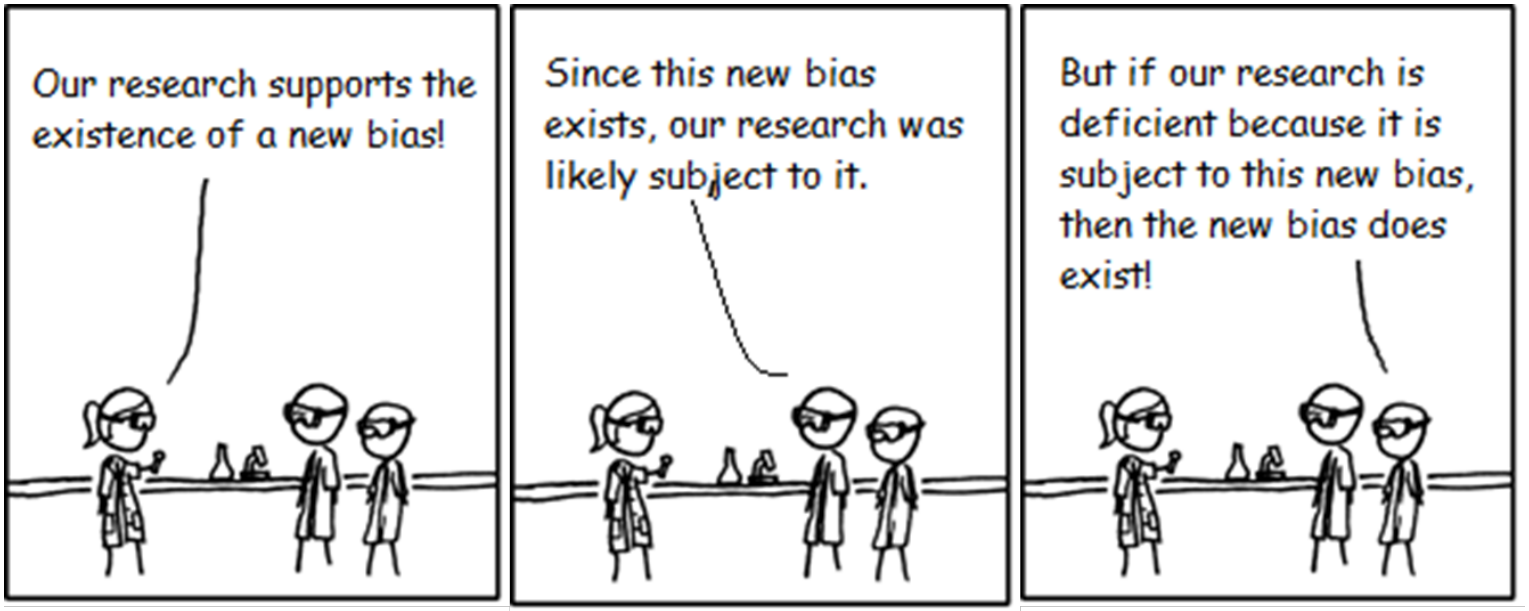
Two Sides of Data Collection Error



Related to the **Total Survey Error** framework: survey error arises both from representation and measurement processes.

Takeaways: Generating and Measuring Data

1. Data can be observational, experimental, or simulated, each granting a different relationship to the phenomenon
2. Whichever mode, we only ever approximate the phenomenon.
3. Measurement error decomposes into bias and variance (MSE)
4. Sampling too slowly in time hides or disguises fast variation (aliasing)
5. Data collection error has two sides: representation (Session 2) and measurement (this session)



Choosing **how** to observe is half the battle.
Measuring it **faithfully** is the other half.