Data, Data Storage, Data Collection

Lecture 4: Practical Data Collection

Romain Pascual

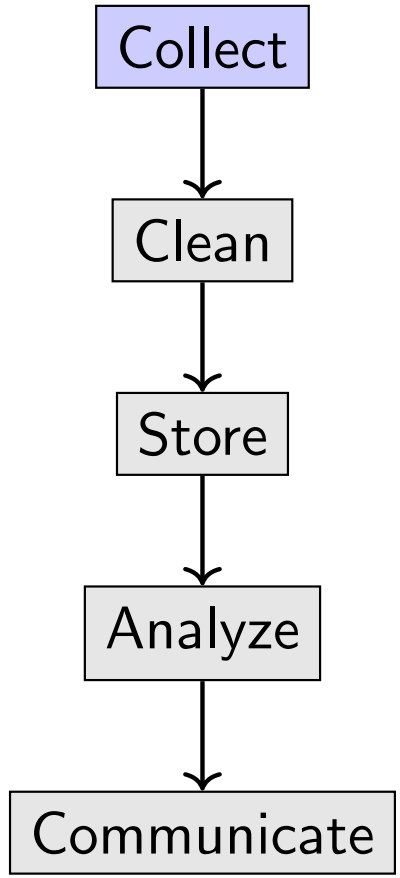
MICS, CentraleSupélec, Université Paris-Saclay

Recap

Recap: 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

Within the lifecycle



Introduction

Quick Poll

Study Aim:

Why do people skip breakfast?

Which method would you choose? Why?

- **A. Survey** Ask 1000 people directly
- **B. Logs** Analyze breakfast sales at cafeterias and bakeries
- **C. Interviews** In-depth conversations with volunteers
- **D. Fitness trackers** Infer eating habits indirectly

Session Objectives

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

- Compare and select **data collection methods** based on study aims and data types
- Recognize common **biases** in collection and strategies to mitigate them
- Appreciate the importance of **provenance** and **documentation** for later reuse

Why Methods Matters

Data can come from experiments, surveys, sensors, APIs, or logs. Each method shapes the **kind of information** we obtain.

- What is feasible to measure?
- How reliable and representative is it?
- What risks of bias or error are introduced?

The choice of the method is inseparable from the **question** and the **quality of the conclusions**.

Data Collection Methods

Overview of Data Collection Methods

Methods depend on the **study aim** and **type of data required**.

Experiments & Simulations

When to use: Test causal relationships

How to collect: Manipulate variables and record outputs

Overview of Data Collection Methods

Methods depend on the **study aim** and **type of data required**.

Experiments & Simulations

When to use: Test causal relationships

How to collect: Manipulate variables and record outputs

Surveys

When to use: Opinions, characteristics

How to collect: Structured questionnaires to sample

Overview of Data Collection Methods

Methods depend on the **study aim** and **type of data required**.

Experiments & Simulations

When to use: Test causal relationships
How to collect: Manipulate variables and record outputs

Surveys

When to use: Opinions, characteristics
How to collect: Structured questionnaires to sample

Sensors & Logs

When to use: Quantitative, continuous data
How to collect: Automated devices or software logs

Overview of Data Collection Methods

Methods depend on the **study aim** and **type of data required**.

Experiments & Simulations

When to use: Test causal relationships
How to collect: Manipulate variables and record outputs

Surveys

When to use: Opinions, characteristics
How to collect: Structured questionnaires to sample

Sensors & Logs

When to use: Quantitative, continuous data
How to collect: Automated devices or software logs

APIs & Secondary Data

When to use: Use existing external data
How to collect: Requests, data feeds

Experiments

Purpose: Test causal links, validate hypotheses

Definition (Experiment)

A study in which the researcher:

- Manipulates one or more **independent variables** (X)
- Observes the effect on **dependent variables** (Y)
- Controls or randomizes other factors to reduce confounding

Modeled as:

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

Goal: Estimate how changes in X cause changes in Y, accounting for noise ϵ.

Examples:

- Does caffeine improve reaction time?
- Do fertilizers influence crop yield?

Confounding Variables

Definition (Confounding Variable)

A factor that affects both the independent (X) and dependent variables (Y), potentially misleading the experiment.

Example: (Studying caffeine and reaction time)

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

Solution: randomize subjects, control confounders, or include them in the analysis.

And document them!

Causal vs Correlation

Definition (Correlation)

A statistical relationship between two variables: they change together.

Definition (Causation)

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

Causal vs Correlation: Example

Example

Ice cream sales and shark attacks both increase in summer.

Causal vs Correlation: Example

Example

Ice cream sales and shark attacks both increase in summer.

This is only a correlation and the causal link comes from summer temperature.

Determining causation vs correlation is one of the central challenges in data science and experimental research. There are some methods (mostly from statistics) that can help (but its beyond the goal of this lecture).

Designing an Experiment

Steps to analyze a factor:

1. Form **treatment group(s)** (receives the intervention) and **control group** (baseline)
2. Randomly assign subjects to reduce confounding
3. Measure outcomes
4. Use statistical tests to assess significance:
   - Continuous data: Student's t-test, ANOVA
   - Categorical data: Chi-square test

Example (Studying caffeine and reaction time)

- Control (no coffee): 0.55, 0.60, 0.57, 0.58, 0.56 s
- Treatment (coffee): 0.50, 0.52, 0.51, 0.49, 0.50 s

If you do the math (t-test), you find out that the difference is much bigger than what you would expect by random chance.

Experiments: Advantages and Challenges

- ✓ **Causal inference**: estimate effect of variables
- ✓ **Reproducibility**: same inputs produce same outputs
- ✓ **Isolation**: control specific variables
- ✓ **Mechanistic understanding**: explore how variables interact

- ✗ **Ethical constraints**: some manipulations may be unethical
- ✗ **External validity**: may not generalize beyond experimental setting
- ✗ **Practical limits**: large parameter sweeps require resources
- ✗ **Sample size / measurement error**: too small or noisy $\rightarrow$ low power

Some experiments are too costly, slow, or unethical

Some experiments are too costly, slow, or unethical

Simulations can mimic experimental settings!

Simulation

Purpose: Explore and predict system behavior when real experiments are impractical.

Definition (Simulation)

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

Modeled as:

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

Goal: Estimate how changes in X and assumptions θ (model parameters) influence Y, accounting for randomness ϵ.

⚠ f is no longer the reality but a model approximation

Examples:

- Airplane crash tests (instead of building many real planes)
- Queueing in a supermarket or call center.

Types of Simulations

Deterministic

Same inputs $\rightarrow$ same outputs

Example: Physics equations for projectile motion

Stochastic

Random inputs $\rightarrow$ random outputs

Example: Monte Carlo simulation for risk assessment

Agent-based

Rules for individual entities and measure emergent system behavior

Example: Epidemic spread, crowd movement

Simulation: Advantages and Challenges

- ✓ Explore many scenarios quickly ("what-if" cases)
- ✓ No ethical constraints
- ✓ Useful for complex systems, e.g., if exact solutions are missing
- ✓ Can guide real experiments

- ✗ Quality depend on model assumptions
- ✗ Needs validation against real data
- ✗ Large simulations may require computational resources
- ✗ Cannot fully replace real-world experiments

Survey

Purpose: Collect information on opinions, behaviors, or characteristics of a population.

Definition (Survey)

A systematic method to gather data from individuals using questions or prompts, often through questionnaires or interviews.

Modeled as:

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

Goal: Estimate population characteristics and relationships between variables X based on responses Y, accounting for individual variability ϵ.

Examples:

- Online questionnaire about students' study habits.
- National health survey measuring smoking prevalence.

Practical Guidelines for Surveys

When designing or using surveys:

- **Keep it short:** Long surveys reduce response rates
- **Use clear language:** Avoid jargon
- **Be specific:** Ask one thing at a time
- **Pilot test:** Try on a small group to detect issues
- **Document:** Record when, how, and to whom the survey was given

✗ Do you like this course?
✓ On a scale of 1–5, how would you rate your satisfaction with this course?

For more further guidance, ask your favorite social scientist

Example of a Survey Question

Do you have any specific food restriction?

Example of a Survey Question

Do you have any specific food restriction?

- ☐ Vegetarian
- ☐ Vegan
- ☐ No fish
- ☐ No meat
- ☐ With meat but without pork
- ☐ With fish
- ☐ With meat and fish

Example of a Survey Question

Do you have any specific food restriction?

- ☐ Vegetarian
- ☐ Vegan
- ☐ No fish
- ☐ No meat
- ☐ With meat but without pork
- ☐ With fish
- ☐ With meat and fish

What do you think of this question design?

Survey (Advantages and Challenges)

- ✓ **Scalability**: reach large and diverse populations
- ✓ **Standardization**: same questions to all respondents
- ✓ **Flexibility**: applicable to many domains

- ✗ **Bias**: the sample might not be representative of the population
- ✗ **Question design**: Poorly worded questions can lead to misleading results
- ✗ **Survey fatigue**: Long surveys may lead to incomplete responses

Automated Data Collection

Purpose: Collect quantitative data from physical devices or digital systems

Definition (Automated Data Collection)

A systematic method to gather data by measuring a quantity of interest (X) and converting it into recorded signals (Y).

Modeled as:

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

Goal: Obtain reliable and interpretable values Y from raw measurements X, accounting for processing f and noise ϵ.

Sensors and Logs

Definition (Sensors)

Physical devices recording measurements (X) from the world.

Examples:

- Thermometer: $f = Id$ and $Y \approx X$
- Pedometer: f gives the number of steps from acceleration X

Definition (Logs)

Digital devices recording events (X) in software or hardware.

Examples:

- Application logs: f aggregates button clicks X into usage statistics Y
- Database transaction logs: f counts queries from transaction events X

Automated Data Collection (Advantages and Challenges)

- ✓ **Automation**: no active user input required
- ✓ **High-frequency**: Continuous or near-continuous streams
- ✓ **Precision**: fine temporal/spatial resolution
- ✓ **Scalable**: large amounts of data captured continuously
- ✓ **Multimodal integration**: Combine multiple sources

- ✗ **Noise-prone**: faulty sensors, device drift, missing values
- ✗ **Context-sensitive**: logs and measurements often lack interpretation of events
- ✗ **Privacy concerns**: sensitive personal or behavioral traces (for logs)

Continuous vs Batch Processing of Automated Data

Automated data can be handled in two main ways:

Batch (stored) processing:

- Suitable for large-scale analytics, historical trends, and complex computations
- Storage and retrieval strategies are important

Continuous (streaming) processing:

- Useful for real-time monitoring, alerts, and control
- Requires fast processing and often temporary storage

Streaming (Online) Algorithms

Definition (Streaming (Online) Algorithm)

An algorithm that processes data sequentially as it arrives, without storing the entire dataset.

No knowledge of future and cannot store all the past.

- **Single pass:** input items are processed one by one
- **Memory efficient:** uses limited memory, cannot store all X
- **Approximate:** results are often estimates, but updated continuously
- **Real-time:** suitable for continuous or high-frequency data (sensors, logs)

Goal: Maintain useful statistics or models Y while new data X_t keeps arriving.

Streaming Counter

Problem

Count the number of elements seen so far in a stream (X^ω).

X^ω are infinite words with values in X.

Algorithm 1: Naive Exact Counter

$n \leftarrow 0$;
foreach element in the stream do
  $n \leftarrow n + 1$;
return n;

Uses $\mathcal{O}(\log n)$ bits to store n exactly.

Can we do better?

Can we do better?

Result

There is no exact solution in $o(\log n)$ space.

Can we do better?

Result

There is no exact solution in $o(\log n)$ space.

But we can look for approximate answers!

Morris Counter

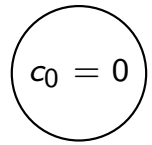
Goal: Estimate n with very little memory, but allow only an approximation of the answer.

Algorithm 2: Morris's Streaming Counter

$c \leftarrow 0$;
foreach element in the stream do
  $c \leftarrow c + 1$ with probability 2^{-c};
return $2^c - 1$;

Memory usage: $c \approx \log n$, so $\mathcal{O}(\log \log n)$ space.

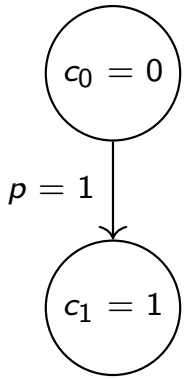
Morris Counter: Processing a Stream



Stream of 5 elements: x_1, x_2, x_3, x_4, x_5

Element	Counter c	Estimate $2^c - 1$

Morris Counter: Processing a Stream



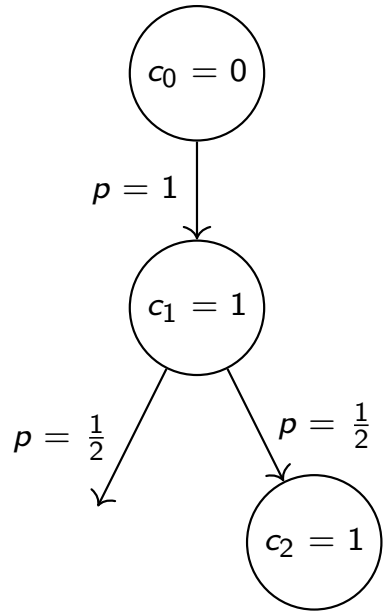
Stream of 5 elements: x_1, x_2, x_3, x_4, x_5

Element	Counter c	Estimate $2^c - 1$
x_1	$0 \to 1$	1

Morris Counter: Processing a Stream

Stream of 5 elements: x_1, x_2, x_3, x_4, x_5

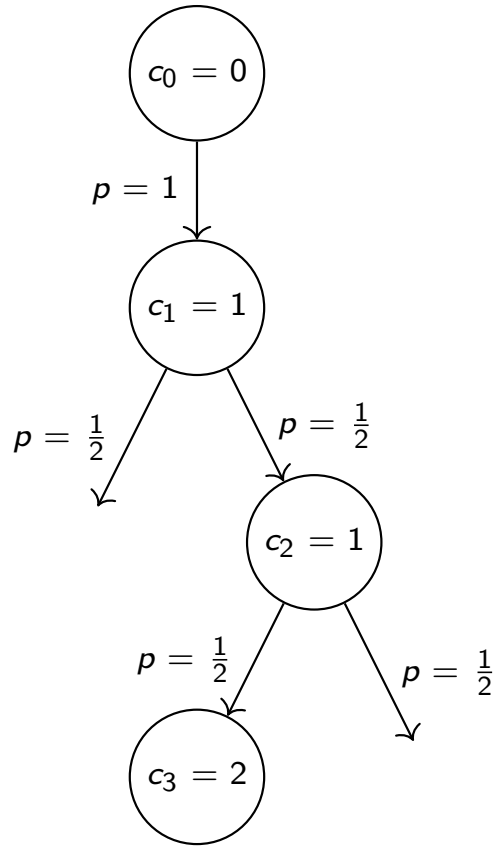
Element	Counter c	Estimate $2^c - 1$
x_1	$0 \to 1$	1
x_2	$1 \to 1$	1



Morris Counter: Processing a Stream

Stream of 5 elements: x_1, x_2, x_3, x_4, x_5

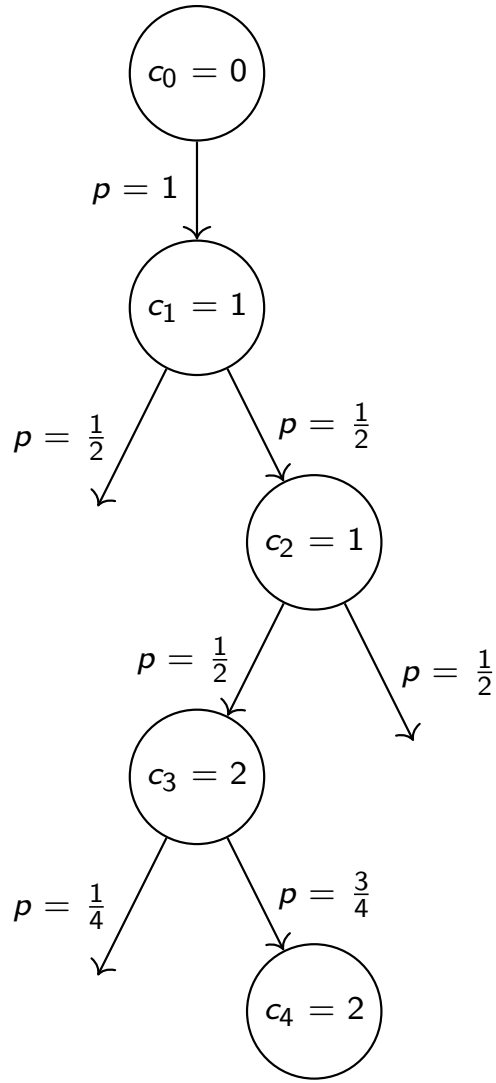
Element	Counter c	Estimate $2^c - 1$
x_1	$0 \to 1$	1
x_2	$1 \to 1$	1
x_3	$1 \to 2$	3



Morris Counter: Processing a Stream

Stream of 5 elements: x_1, x_2, x_3, x_4, x_5

Element	Counter c	Estimate $2^c - 1$
x_1	$0 \to 1$	1
x_2	$1 \to 1$	1
x_3	$1 \to 2$	3
x_4	$2 \to 2$	3

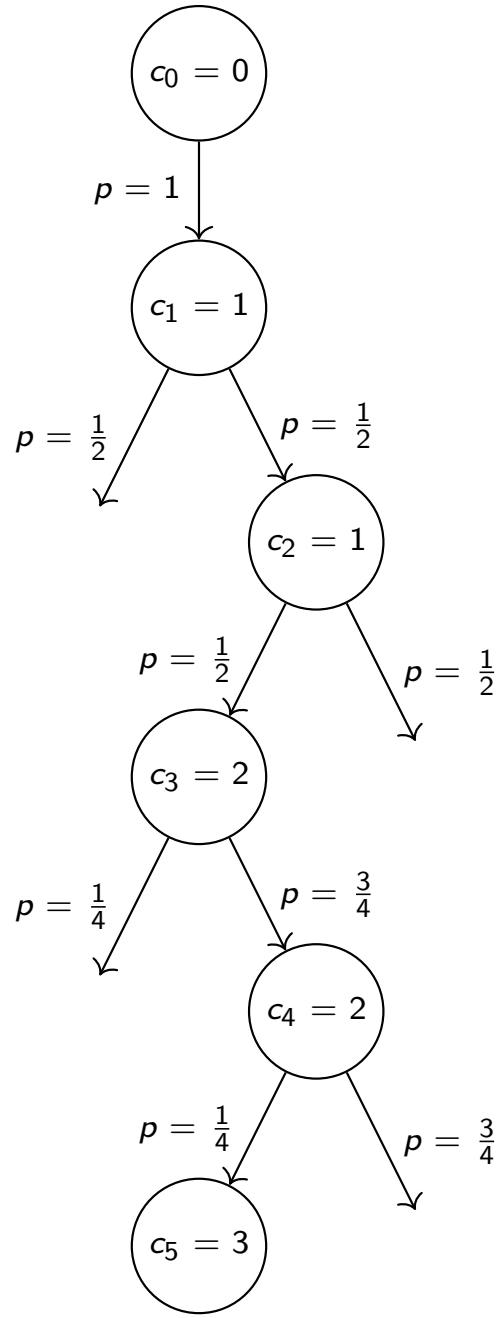


Morris Counter: Processing a Stream

Stream of 5 elements: x_1, x_2, x_3, x_4, x_5

Element	Counter c	Estimate $2^c - 1$
x_1	$0 \to 1$	1
x_2	$1 \to 1$	1
x_3	$1 \to 2$	3
x_4	$2 \to 2$	3
x_5	$2 \to 3$	7

The final estimate $2^c - 1 = 7$ approximates the true count $n = 5$.



Morris Counter: What Does it Return?

Let c_n be the value of the counter after n stream items, then $\mathbb{E}[2^{c_n}] = n + 1$.

Base case ($n = 0$). $c_0 = 0$, so $2^{c_0} = 1$ and $\mathbb{E}[2^{c_0}] = 1 = 0 + 1$.

Inductive step (assume $\mathbb{E}[2^{c_n}] = n + 1$). Consider one more element ($n \to n + 1$). By the formula of total probabilities:

$$\mathbb{E}[2^{c_{n+1}}] = \sum_{i=0}^{\infty} \mathbb{P}(c_n = i)\mathbb{E}[2^{c_{n+1}} | c_n = i]$$

Morris Counter: Conditional Expectation

If $c_n = i$, then:

$$c_{n+1} = \begin{cases} c_n + 1 & \text{with probability } 2^{-i} \\ c_n & \text{with probability } 1 - 2^{-i} \end{cases}$$

Morris Counter: Conditional Expectation

If $c_n = i$, then:

$$c_{n+1} = \begin{cases} c_n + 1 & \text{with probability } 2^{-i} \\ c_n & \text{with probability } 1 - 2^{-i} \end{cases}$$

Conditional expectation:

$$\mathbb{E}\left[2^{c_{n+1}} \mid c_n = i\right] = \underbrace{2^i(1 - 2^{-i})}_{\textit{unchanged}} + \underbrace{2^{i+1} \cdot 2^{-i}}_{\textit{incremented}} = 2^i + 1.$$

Plugging it in the formula, we obtain

$$\mathbb{E}[2^{c_{n+1}}] = \underbrace{\sum_{i=0}^{\infty} \mathbb{P}(c_n = i) 2^i}_{=\mathbb{E}[2^{c_n}]} + \underbrace{\sum_{i=0}^{\infty} \mathbb{P}(c_n = i)}_{=1}.$$

Morris Counter: Conclusion

Thus,

$$\mathbb{E}[2^{c_{n+1}}] = \mathbb{E}[2^{c_n}] + 1$$

By hypothesis $\mathbb{E}[2^{c_n}] = n + 1$, and by induction

$$\mathbb{E}[2^{c_{n+1}}] = (n + 1) + 1$$

Hence, $2^{c_n} - 1$ is an **unbiased** estimator of n.

Streaming Algorithms in Data Collection

Sensors and logs can generate massive streams of data that are impossible to store entirely.

Streaming algorithms process this data on the fly, producing approximate statistics.

Applications:

- Monitoring IoT devices, sensors, or telemetry
- Real-time analytics without massive storage

What is actually collected?

- Not the full raw stream X_t
- Instead, the algorithm maintains a compact state c_t
- The collected data is either the internal state c_t or the derived statistic $Y_t = f(c_t)$ (e.g., $2^{c_t} - 1$)

APIs[1]

Purpose: Reuse existing data through programmatic access.

Definition (API-based Data Collection)

A method to gather data by querying a digital system to obtain responses Y derived from query variables X.

Modeled as:

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

Goal: Obtain data Y from API query parameters X, accounting for processing f and noise or incompleteness ϵ.

Examples:

- Weather API: X is the location and $f(X)$ is the temperature.
- Stock prices from financial APIs.

You have seen this already last year in the cloud computing lecture.

[1] Application Programming Interfaces

API (Advantages and Challenges)

- ✓ **Cost-efficient**: no need to collect primary data yourself
- ✓ **Scalable**: access to large datasets with repeated calls
- ✓ **Reproducibility**: standardized sources and formats
- ✓ **Programmatic automation**: Can be integrated into pipelines

- ✗ **Access limits**: rate limits, paywalls
- ✗ **Partial view**: only exposed data are accessible, may differ from full internal state
- ✗ **Opacity**: data might have been collected for other purposes
- ✗ **Dependency on provider**: API changes can break pipelines

Bias

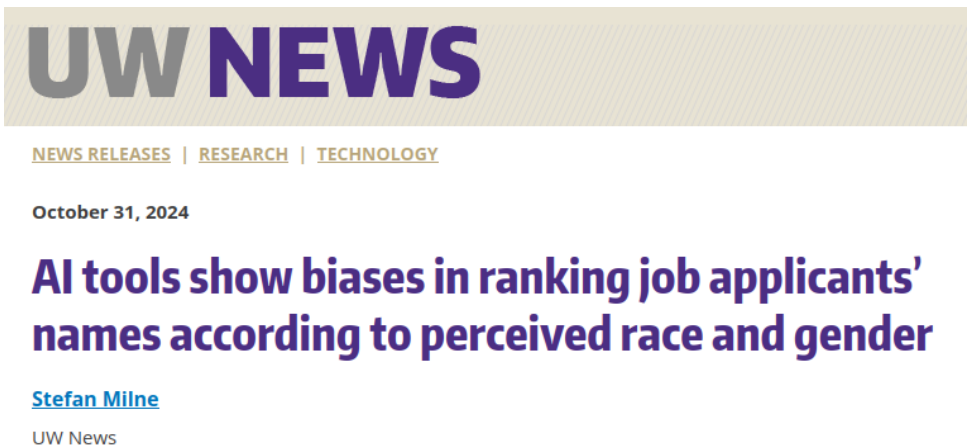
What is Bias?

Definition (Bias)

Systematic deviation from the true characteristics of the population.

Why it matters:

- Incorrect conclusions (e.g., flawed policies, ineffective treatments)
- Perpetuate inequalities (e.g., hiring algorithms favoring certain groups)



UW NEWS

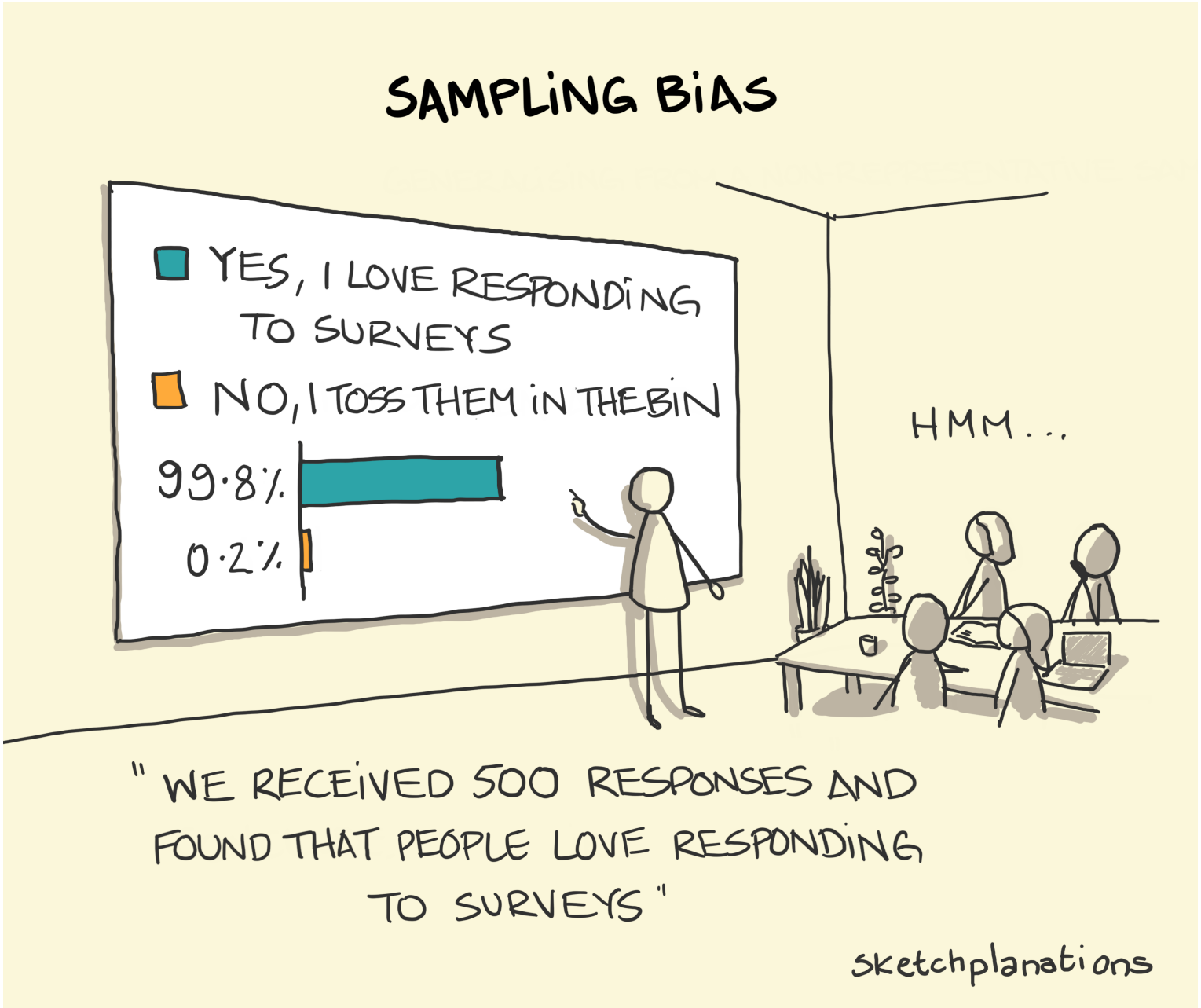
NEWS RELEASES | RESEARCH | TECHNOLOGY

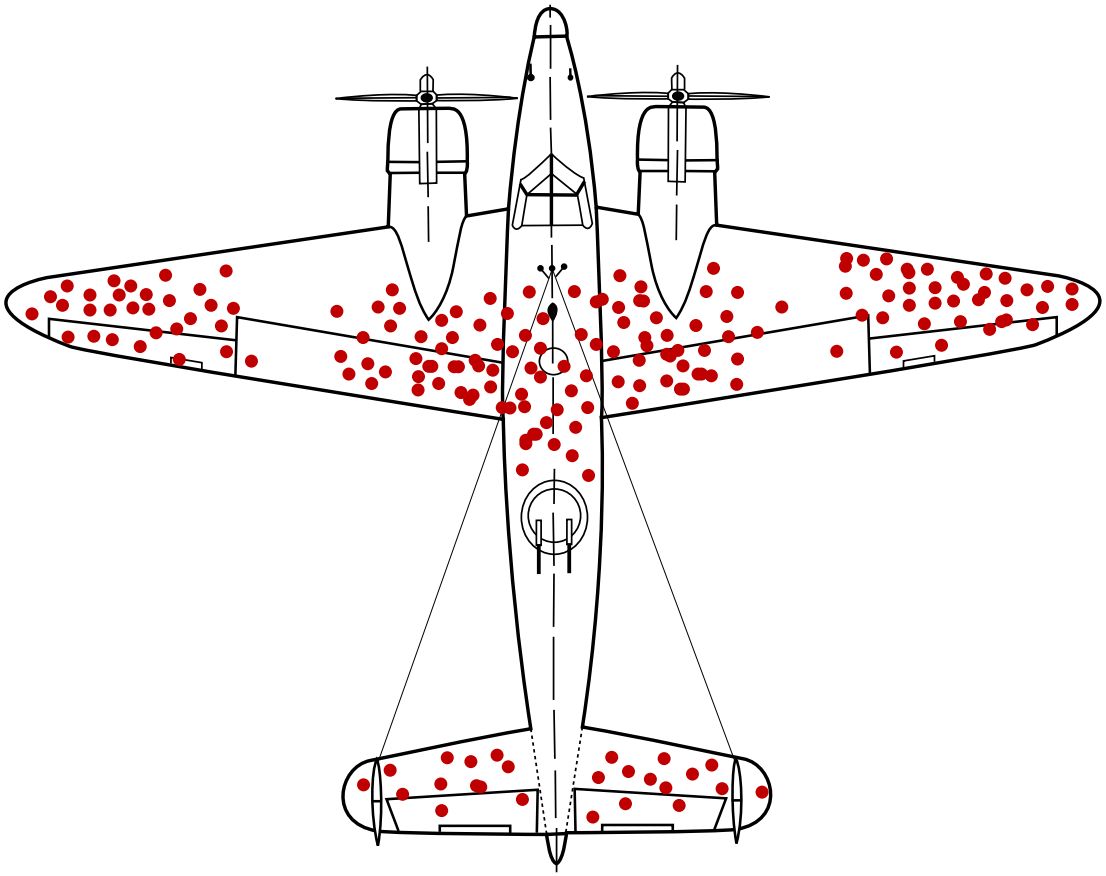
October 31, 2024

AI tools show biases in ranking job applicants' names according to perceived race and gender

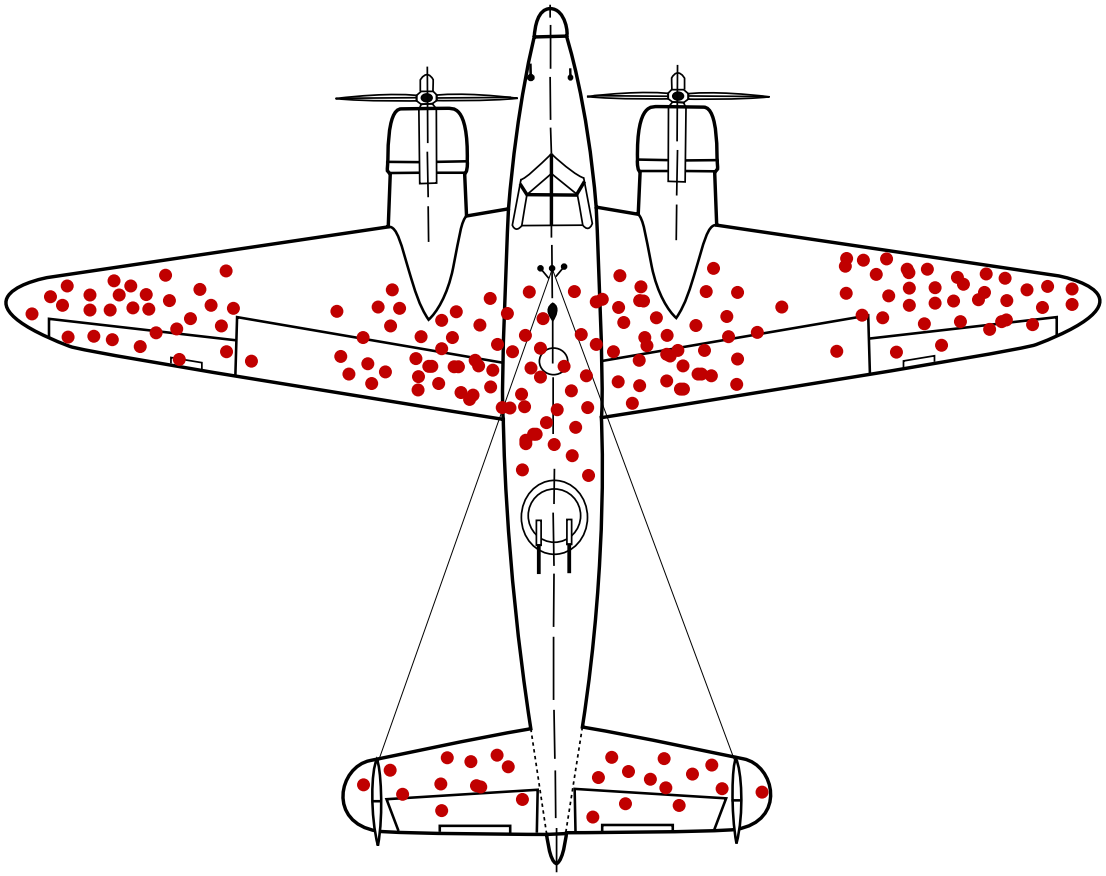
Stefan Milne

UW News

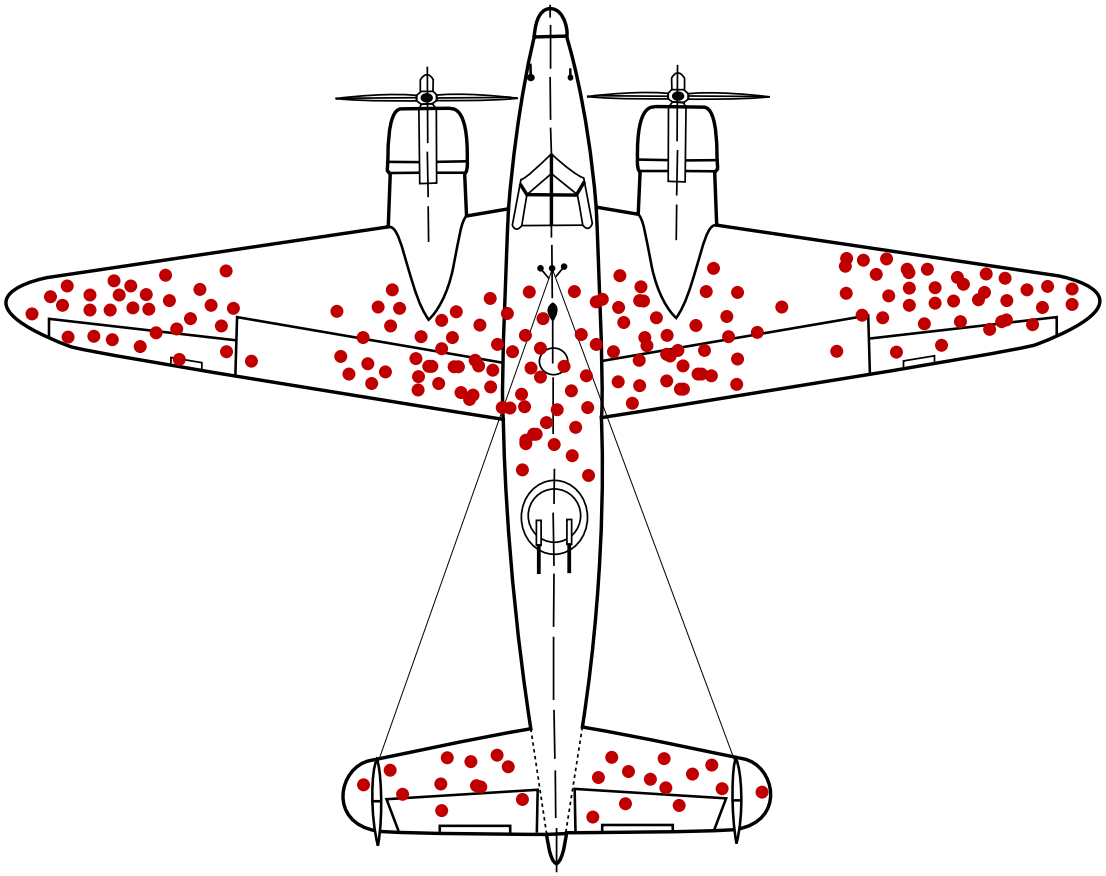

SAMPLING BIAS
YES, I LOVE RESPONDING TO SURVEYS
NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM...
" WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS "
sketchplanations



Survivorship Bias



Abraham Wald



Abraham Wald

Can you think of modern examples?

THE AVAILABILITY HEURISTIC†

THINGS THAT COME TO MIND EASILY WE THINK OF AS MORE COMMON AND MORE IMPORTANT

	EASY TO RECALL	HARD TO RECALL	CAN LEAD TO
	PLANE CRASHES	SAFE FLIGHTS	FEAR OF FLYING
	COLD SPELL	GRADUAL WARMING	IGNORING CLIMATE CHANGE
	WINNING TICKETS	NORMAL TICKETS	BUYING MORE

. . . AT THE EXPENSE OF THINGS THAT MAY BE MORE COMMON BUT THAT DON'T EASILY COME TO MIND

† AND BIAS

sketchplanations

Common Biases in Data Collection

Non-response Bias Participants who do not respond may differ from those who do (e.g., unhappy customers skip surveys).

Common Biases in Data Collection

Non-response Bias Participants who do not respond may differ from those who do (e.g., unhappy customers skip surveys).

Response Bias Participants answer in a way they think is "desirable" (e.g., overreporting healthy behaviors).

Common Biases in Data Collection

Non-response Bias Participants who do not respond may differ from those who do (e.g., unhappy customers skip surveys).

Response Bias Participants answer in a way they think is "desirable" (e.g., overreporting healthy behaviors).

Selection Bias Certain groups are overrepresented (e.g., only highly motivated individuals respond).

Common Biases in Data Collection

Non-response Bias Participants who do not respond may differ from those who do (e.g., unhappy customers skip surveys).

Response Bias Participants answer in a way they think is "desirable" (e.g., overreporting healthy behaviors).

Selection Bias Certain groups are overrepresented (e.g., only highly motivated individuals respond).

Undercoverage Bias Some groups are excluded from the sample (e.g., only surveying students who attend lectures).

Bias in Computational Data

Bias is not just a human problem: it also affects machine-generated data.

Sensors Sensor drift or faulty calibration leads to systematic errors.

Bias in Computational Data

Bias is not just a human problem: it also affects machine-generated data.

Sensors Sensor drift or faulty calibration leads to systematic errors.

Logs Missing or truncated events (e.g., failed transactions not logged).

Bias in Computational Data

Bias is not just a human problem: it also affects machine-generated data.

Sensors Sensor drift or faulty calibration leads to systematic errors.

Logs Missing or truncated events (e.g., failed transactions not logged).

APIs Provider filters or rate limits exclude certain data.

Bias in Computational Data

Bias is not just a human problem: it also affects machine-generated data.

Sensors	Sensor drift or faulty calibration leads to systematic errors.
Logs	Missing or truncated events (e.g., failed transactions not logged).
APIs	Provider filters or rate limits exclude certain data.
Simulations	Underrepresented states (e.g., rare events not modeled).

Strategies to Reduce Bias

- Use multiple sampling methods to ensure representation
- Randomize selection or sampling
- Calibrate and validate measurement instruments (sensors, logs, APIs)
- Document limitations and check for missing data
- Compare with external or real-world data where possible

Provenance & Documentation

The FAIR Data Principles

Mark D. Wilkinson et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: **Scientific Data** 3.1 (Mar. 2016). ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18

Motivation:

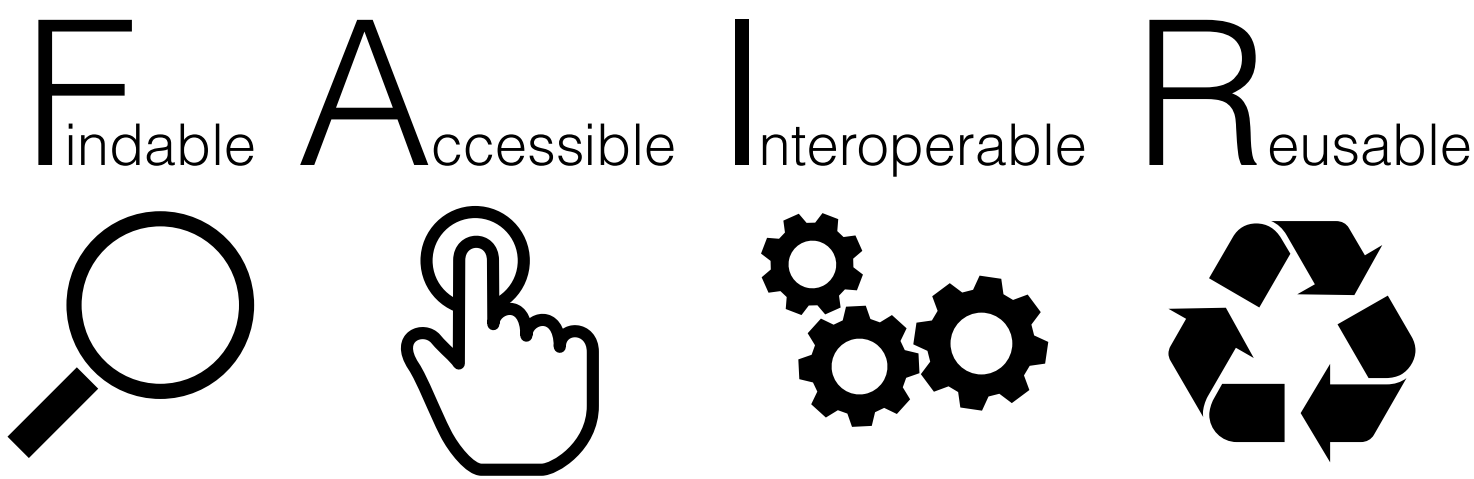
- Digital data abundant but hard to find, access, combine, or reuse
- Need for reproducible, interoperable, trustworthy data

FAIR principles guide good data management to make datasets usable by both computers and humans. It is about both the **data itself** and its **metadata**.

The FAIR Principles: Making Data Usable

Data should be:

- **Findable:** Easy to locate
- **Accessible:** Retrievable with standard tools
- **Interoperable:** Usable with other datasets
- **Reusable:** Well-documented for future use



How to Make Data FAIR

Findable, **A**ccessible, **I**nteroperable, **R**eusable

Which is the principle associated with each advice?

- Use **standard formats** (CSV, JSON)
- Include a **clear license** (e.g., Creative Commons)
- Assign a **DOI** (e.g., via `https://zenodo.org/`)
- Allow downloading over **standard protocols** (e.g., HTTP, FTP)
- Document with **metadata standards** (e.g., Dublin Core)

How to Make Data FAIR

Findable, **A**ccessible, **I**nteroperable, **R**eusable

Which is the principle associated with each advice?

(I) Use **standard formats** (CSV, JSON)

- Include a **clear license** (e.g., Creative Commons)
- Assign a **DOI** (e.g., via `https://zenodo.org/`)
- Allow downloading over **standard protocols** (e.g., HTTP, FTP)
- Document with **metadata standards** (e.g., Dublin Core)

How to Make Data FAIR

Findable, **A**ccessible, **I**nteroperable, **R**eusable

Which is the principle associated with each advice?

(I) Use **standard formats** (CSV, JSON)

(R) Include a **clear license** (e.g., Creative Commons)

- Assign a **DOI** (e.g., via `https://zenodo.org/`)
- Allow downloading over **standard protocols** (e.g., HTTP, FTP)
- Document with **metadata standards** (e.g., Dublin Core)

How to Make Data FAIR

Findable, **A**ccessible, **I**nteroperable, **R**eusable

Which is the principle associated with each advice?

(I) Use **standard formats** (CSV, JSON)

(R) Include a **clear license** (e.g., Creative Commons)

(F) Assign a **DOI** (e.g., via `https://zenodo.org/`)

- Allow downloading over **standard protocols** (e.g., HTTP, FTP)
- Document with **metadata standards** (e.g., Dublin Core)

How to Make Data FAIR

Findable, **A**ccessible, **I**nteroperable, **R**eusable

Which is the principle associated with each advice?

(I) Use **standard formats** (CSV, JSON)

(R) Include a **clear license** (e.g., Creative Commons)

(F) Assign a **DOI** (e.g., via `https://zenodo.org/`)

(A) Allow downloading over **standard protocols** (e.g., HTTP, FTP)

- Document with **metadata standards** (e.g., Dublin Core)

How to Make Data FAIR

Findable, **A**ccessible, **I**nteroperable, **R**eusable

Which is the principle associated with each advice?

(I) Use **standard formats** (CSV, JSON)

(R) Include a **clear license** (e.g., Creative Commons)

(F) Assign a **DOI** (e.g., via `https://zenodo.org/`)

(A) Allow downloading over **standard protocols** (e.g., HTTP, FTP)

(R) Document with **metadata standards** (e.g., Dublin Core)

What to Document: The Essentials

1. **Who, When, Where:** Who collected the data? When and where was it collected?
2. **How:** What tools/methods were used? (e.g., survey questions, sensor types, API versions)
3. **What:** What do the columns/variables mean? (e.g., units, data types, encoding)
4. **Why:** What was the purpose of collecting this data?

Example (Survey Documentation)

1. **Who:** 3rd-year students at Bachelor AIDAMS
2. **How:** Online questionnaire via Google Forms
3. **What:** `StudyHours`: weekly hours spent studying (integer)
4. **Why:** Analyze the influence of study time on exam scores

Conclusion

Takeaways: Data Collection Methods

1. Choose the right method based on your study aim and the type of data required.
2. **Experiments & Simulations:** Use to test causal relationships and explore scenarios.
3. **Surveys:** Use to collect opinions, behaviors, or characteristics.
4. **Automated Collection** Use for quantitative, continuous data collection.
5. **APIs & Secondary Data:** Use existing external data for efficiency and scalability.

"Choosing a collection method is like picking the right lens — it shapes what you will see."

Takeaways: Bias in Data Collection

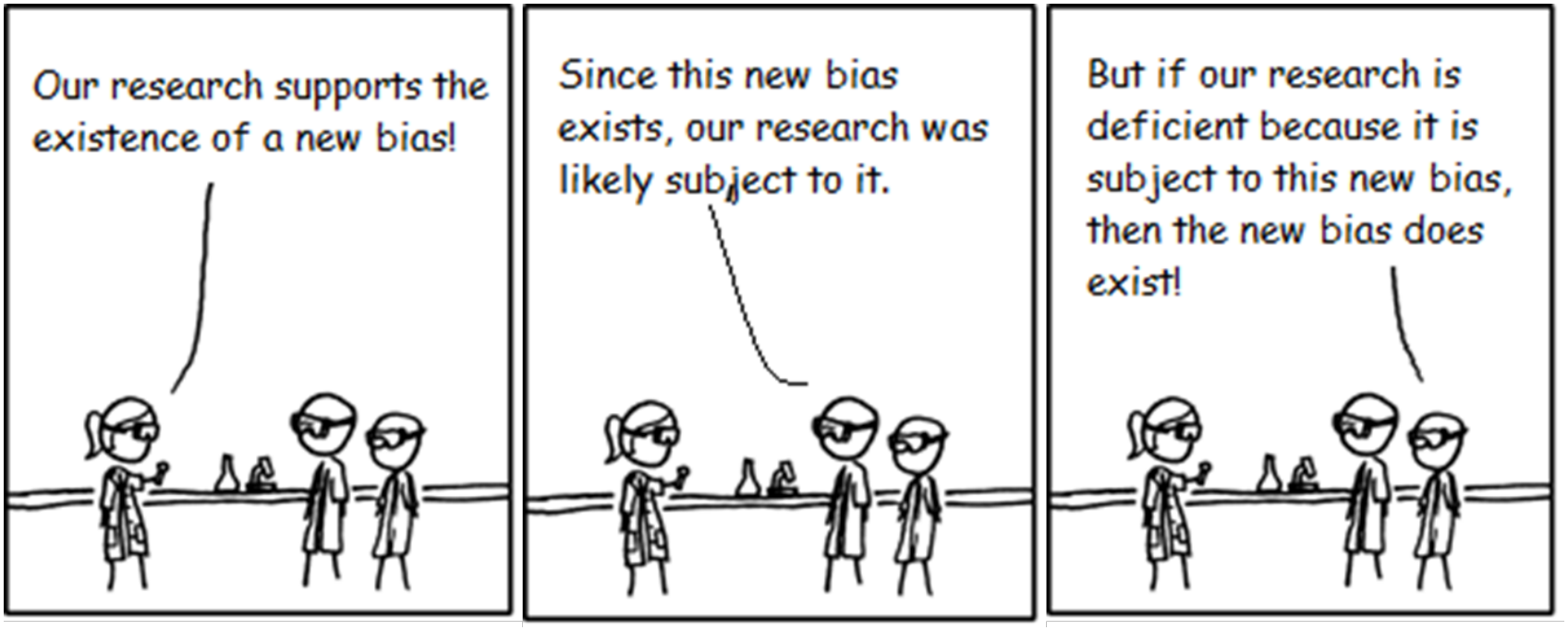
1. Bias is **systematic deviation** from the true population characteristics.
2. Common biases: non-response, undercoverage, selection, response, and computational bias.
3. Mitigation strategies: diversify samples, randomize, pilot test.
4. Always ask: **Who or what is missing from this data? Why?**
5. If cannot mitigate them, at least document them!

"Bias in data collection is like a shadow – it follows you everywhere, but you can shine a light on it!"

Takeaways: Provenance & Documentation

1. FAIR principles make data **Findable**, **Accessible**, **Interoperable**, and **Reusable**.
2. Always document **who, how, what, and why** for your data.
3. Use tools like DOIs, standard formats, and licenses to improve reusability.
4. Poor documentation = wasted time, incorrect conclusions, or ethical risks.

"Data without documentation is like a book without a title – hard to find, use, or trust!"


Our research supports the existence of a new bias!
Since this new bias exists, our research was likely subject to it.
But if our research is deficient because it is subject to this new bias, then the new bias does exist!

T-test Computation

Compute whether coffee improves reaction time:

$$\bar{X}_c = 0.566, \quad \bar{X}_t = 0.504$$

$$s_c^2 = 0.00013, \quad s_t^2 = 0.00006$$

$$t = \frac{\bar{X}_c - \bar{X}_t}{\sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}}} = \frac{0.566 - 0.504}{\sqrt{\frac{0.00013}{5} + \frac{0.00006}{5}}} = \frac{0.062}{0.0061} \approx 10.16$$

- Degrees of freedom: $n_c + n_t - 2 = 8$
- Critical t-value ($\alpha = 0.05$, two-tailed) ≈ 2.306
- Conclusion: $t = 10.16 > 2.306 \rightarrow$ significant difference