Selection bias

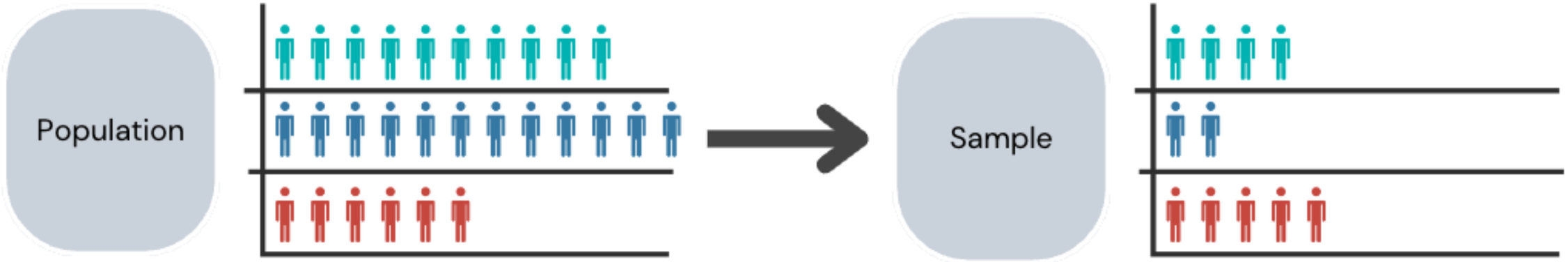
CONQUERING DATA BIAS

Konstantinos Kattidis
Data Analytics Lead

datacamp

What is selection bias?

> It's the bias introduced when the data for analysis is **selected** in a way that **systematically favors** certain individuals, groups, or characteristics

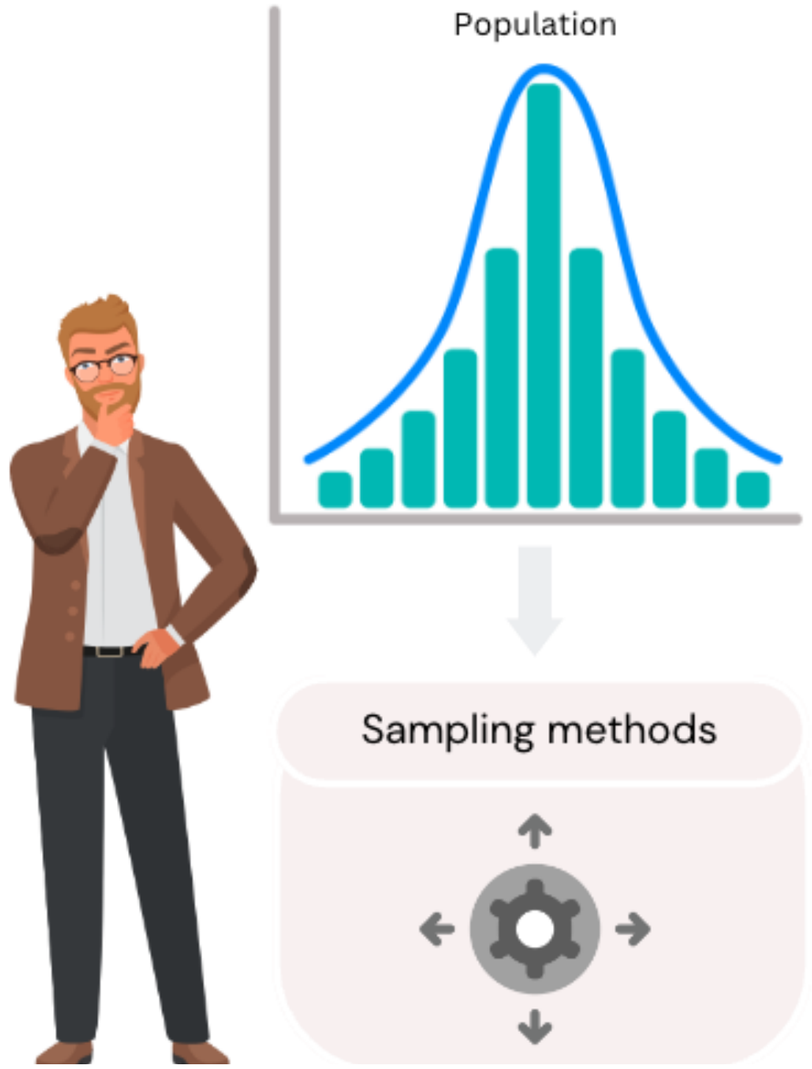


Hence, the sample obtained is **not representative of the population** intended to be analyzed

Let's delve into the **five common types** of selection bias

1. Sampling bias

- Sampling bias occurs when the **sampling method** is not fair or random
- It originates from the **approach** we choose to obtain our sample which can make it hard or impossible to apply the findings to the whole population
- For example:
  - An e-commerce platform analyzes customer satisfaction using convenience sampling
  - The findings may not reflect the sentiments of the entire customer base

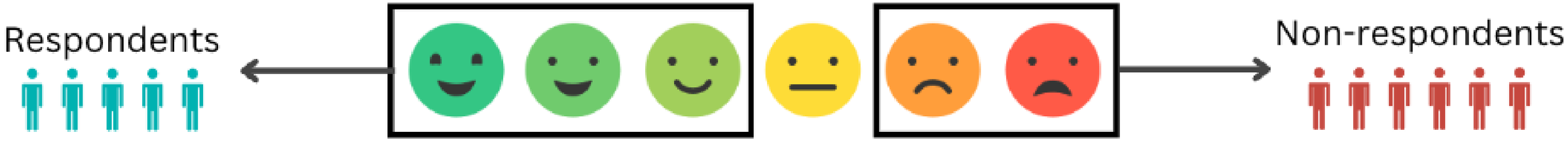


2. Undercoverage bias

- Consider a market research study targeting online consumers, excluding individuals without internet access
- Undercoverage bias highlights the **inadequate representation of certain groups** within the chosen sample
- It is distinguished from sampling bias by its focus on the **representation of specific groups** rather than the randomness or fairness of the sampling method itself

3. Non-response bias

> Non-response bias arises when individuals who choose not to participate in a survey or study differ systematically from those who do participate



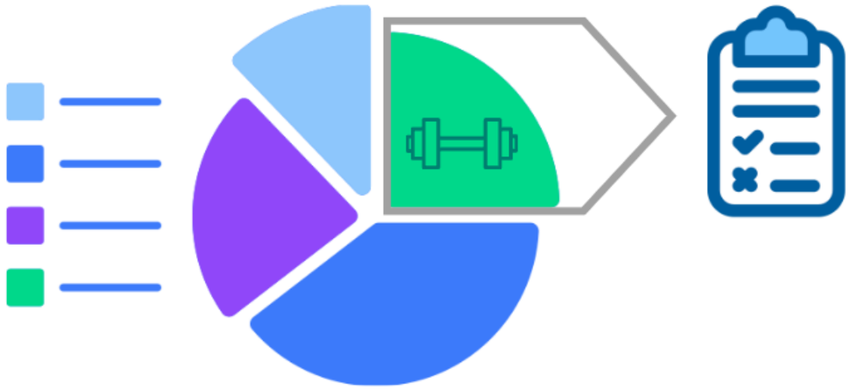
- In a survey assessing employee satisfaction **dissatisfied employees are less likely to participate**
- This leads to to an overly optimistic view of employee morale

4. Self-selection bias

- Self-selection bias occurs when individuals **choose to participate** in a study or provide feedback
- For example:
  - Customers self-select to participate in a satisfaction survey
  - When their views do not represent the broader customer base
  - This skews the overall perception

5. Survivorship bias

> It occurs when only **successful entities** are included in the analysis



For example:

- Analyzing successful product launches without considering the ones that failed
- This would lead to biased insights, overlooking critical factors that contribute to failure

Creating a cohesive understanding

- It's not uncommon for multiple biases to interact, complicating analyses
- For example, a customer satisfaction survey may exhibit both self-selection bias and non-response bias

Let's practice!

CONQUERING DATA BIAS

Historical bias

CONQUERING DATA BIAS

Konstantinos Kattidis
Data Analytics Lead

datacamp

Understanding historical bias

- You're an analyst investigating a series of thefts in a neighborhood
- You notice inconsistencies such as shifts in crime patterns, demographics and law enforcement practices
- Historical bias occurs when past events do not provide an accurate reflection of present circumstances
- Two common types: **technological bias** and **contextual bias**

Technological bias

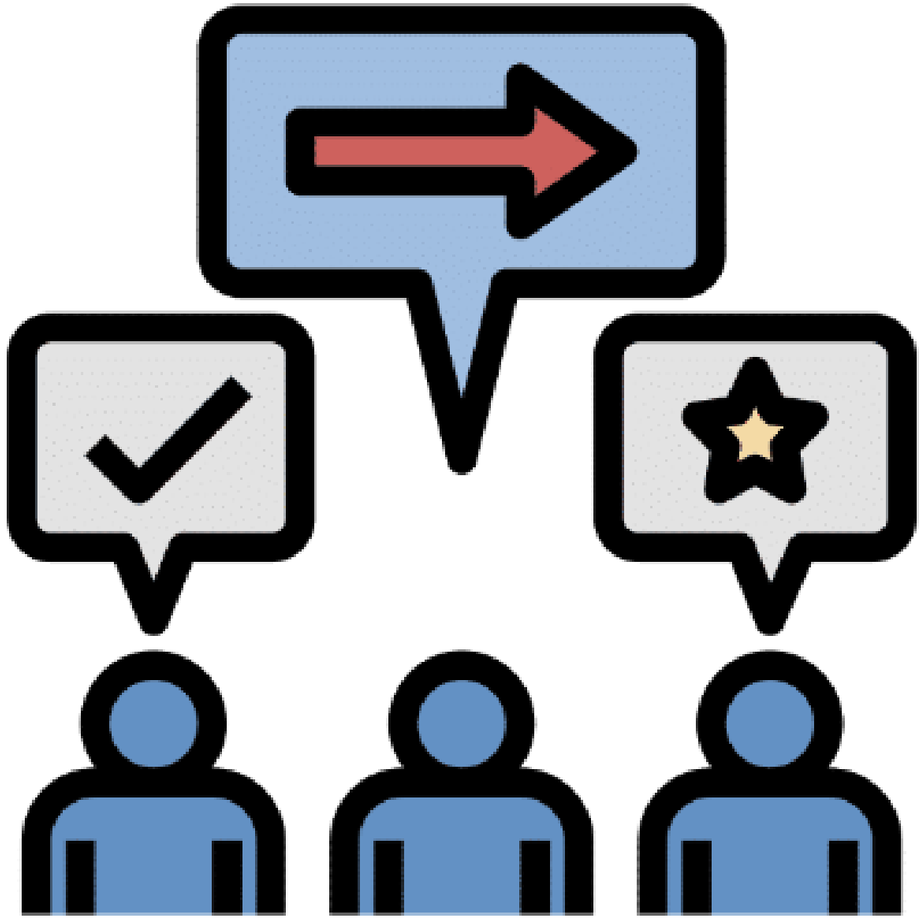
> It arises when advancements render historical data outdated or less relevant

- A product analyst wants to understand the evolution of user behavior over time using user engagement data
- During this period, desktop computers were the primary means of internet access
- The patterns of user behavior may not accurately reflect interactions with mobile devices

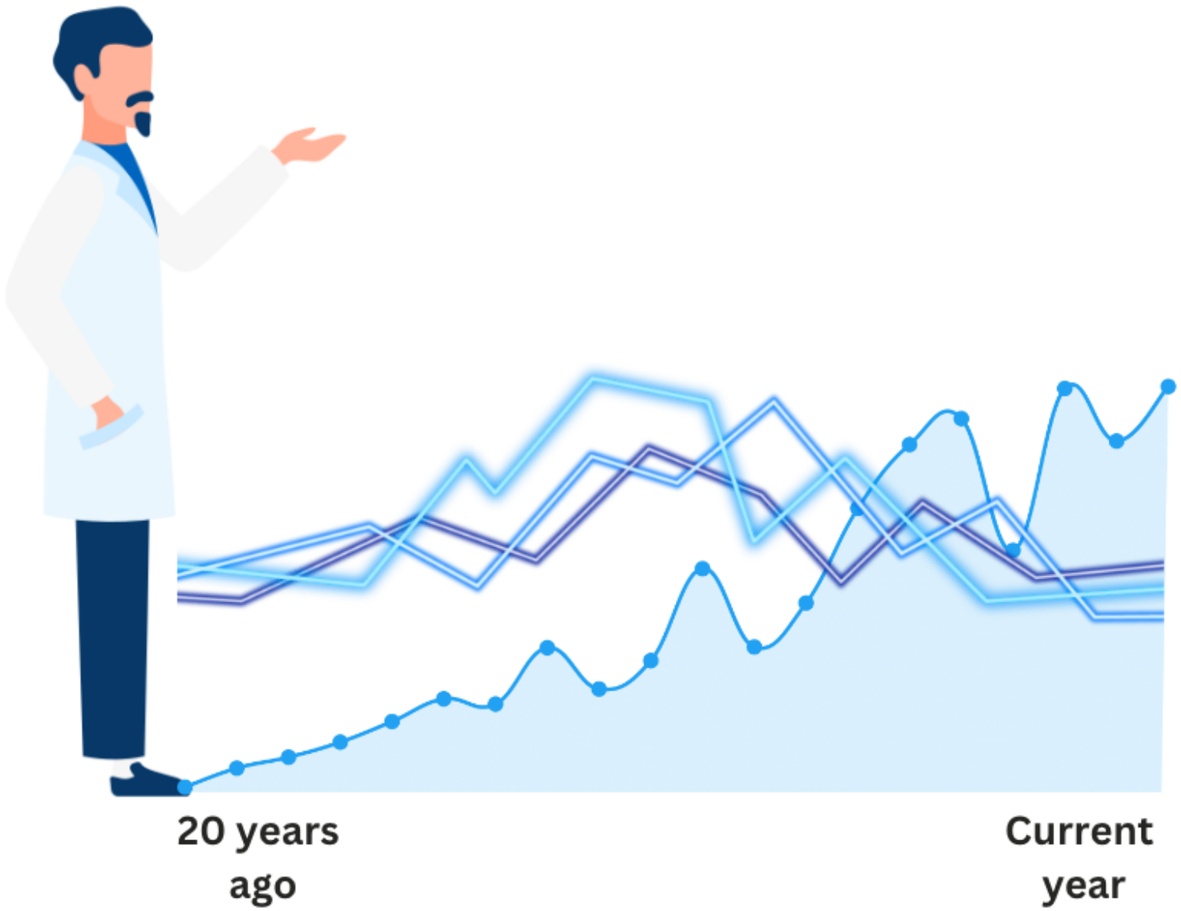
Contextual bias

> It refers to changes in societal norms, preferences, cultural values or economic conditions that influence historical data

- A marketing analyst is analyzing customer data to inform advertising campaigns for a clothing brand
- Accounting for evolving cultural norms and values
- The need for contextual understanding in decision-making is essential



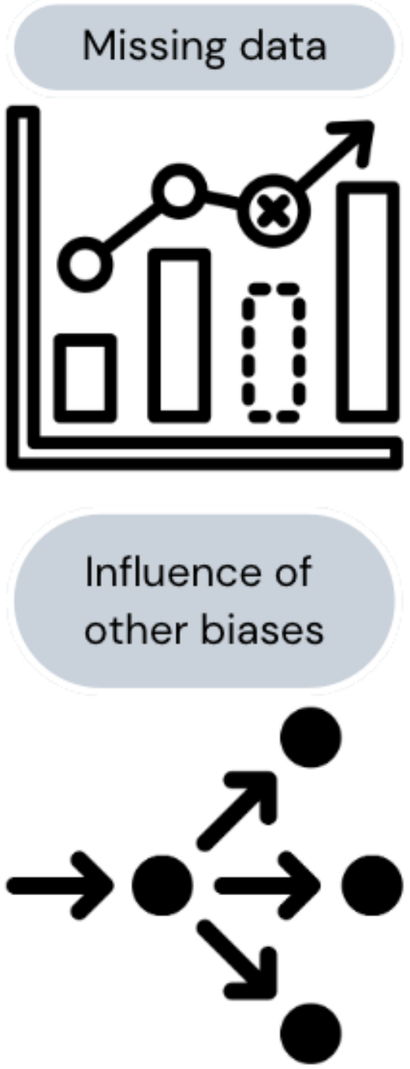
Longitudinal studies



- Longitudinal studies, which track subjects over an extended period, are particularly susceptible to historical bias
- For example, accounting for the climate change in long-term climate studies
- Failing to account for this shift can introduce historical bias in climate analyses

Additional factors causing historical bias

1. Limitations in historical data sources:
   - Incomplete data or underrepresented groups can lead to biased interpretations of past events
   - Historical events may have led to certain data being lost or inaccessible
2. Other bias types
   - Other bias types can contribute to historical bias
   - For example, survivorship bias can lead to a distorted historical perspective



Impact on data interpretation

- Historical bias significantly influences how data is interpreted
- Overlooking historical context may lead to inaccurate assumptions about trends

Let's practice!

CONQUERING DATA BIAS

Measurement bias

CONQUERING DATA BIAS

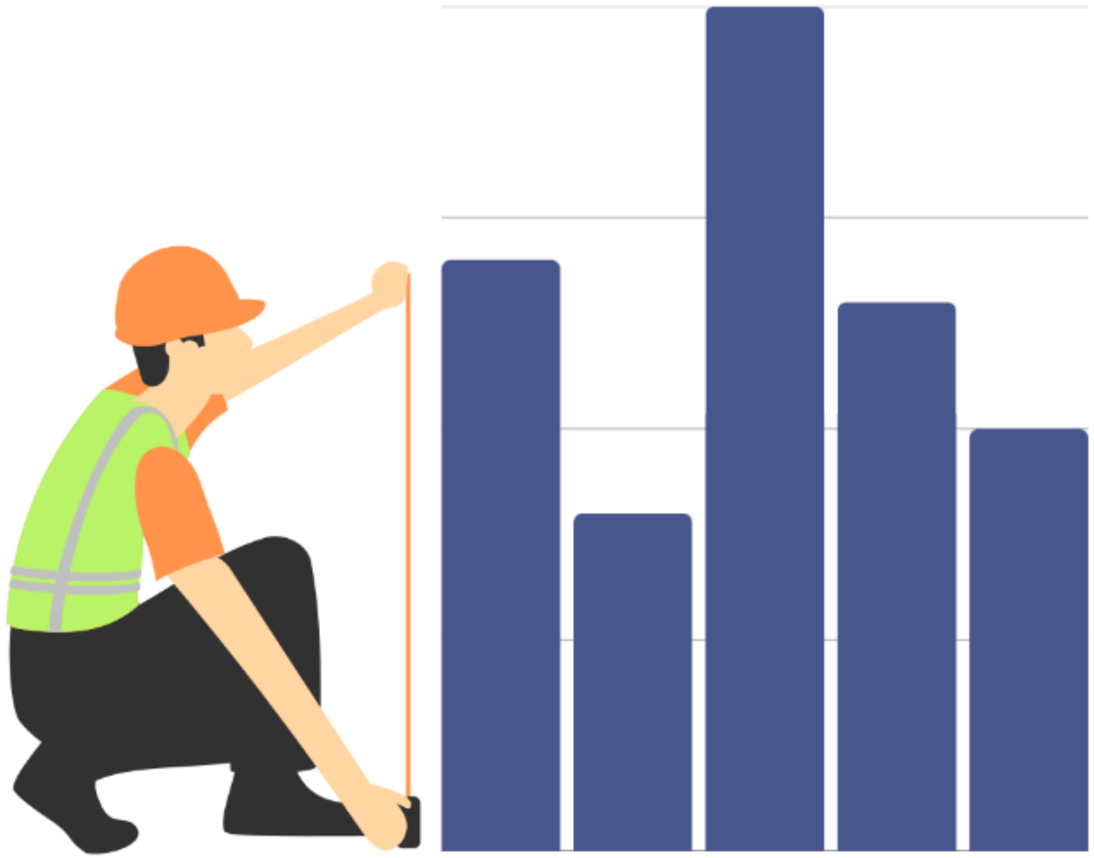
Konstantinos Kattidis
Data Analytics Lead

Understanding measurement bias

> It occurs when the measurement procedure introduces distortions or misleading outcomes

Common bias types include:

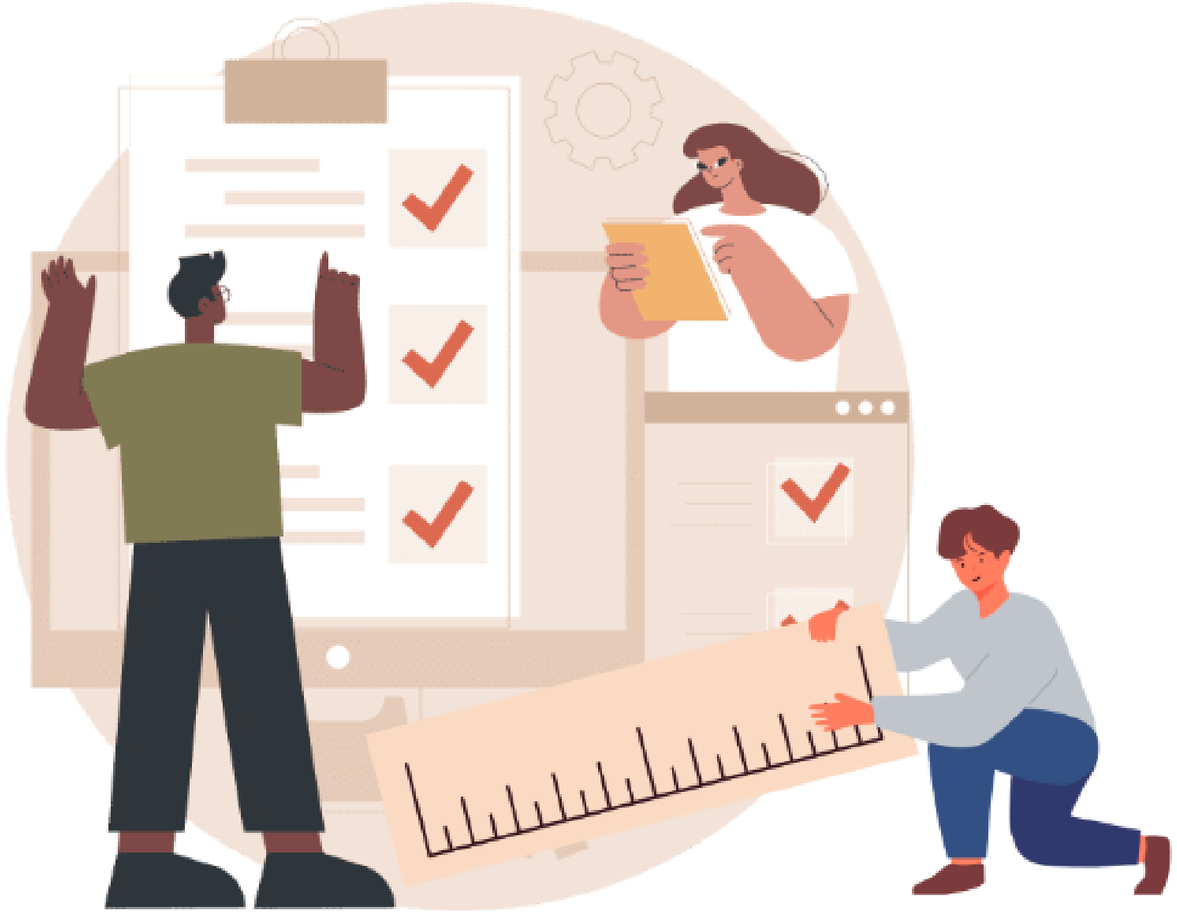
- Instrument bias
- Observer bias
- Recall bias
- Social desirability bias



Instrument bias

> It occurs when the tools used to measure variables, such as surveys or analytics software, introduce inaccuracies

- For example:
  - An analytics software failing to accurately capture certain user behaviors or attributes
  - Poorly designed survey questions causing biased respondents answers

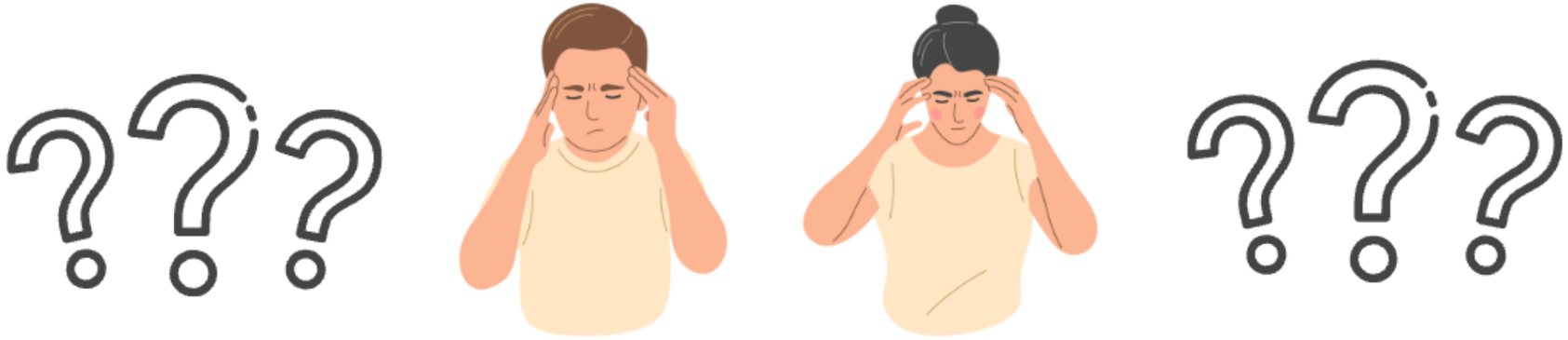


Observer bias

> It's the systematic difference between what is observed due to variation in observers, and the true value

- It arises when data collectors bring their own biases or expectations into the analysis
- For example:
  - Researchers observing classroom behavior but have preconceived notions about what constitutes "engagement"
  - Personal biases in performance evaluations

Recall bias



> It occurs when participants inaccurately remember past events or experiences, affecting data reliability

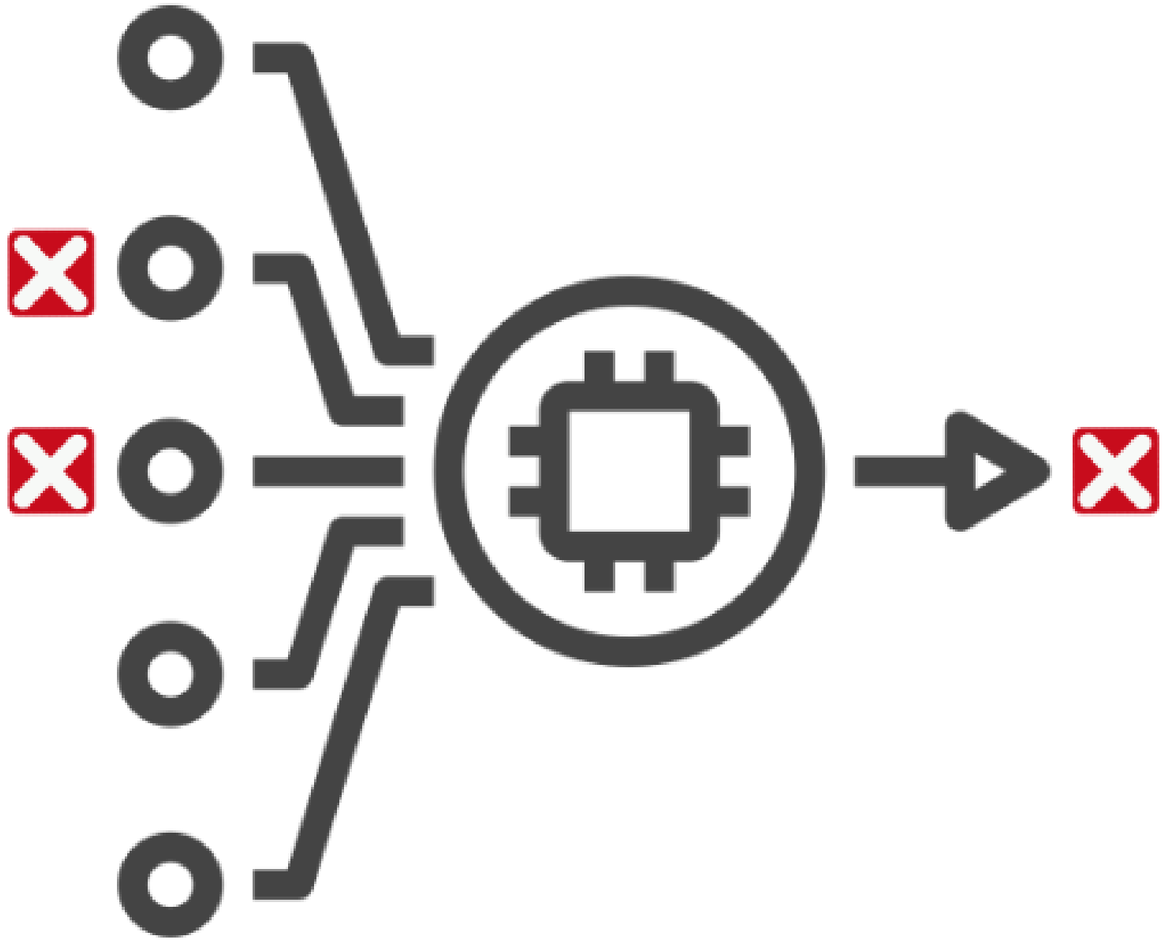
- It is common when customers provide feedback on past experiences or preferences
- It can be due to various factors such as memory decay, or external influences

Social desirability bias

> It occurs when respondents provide answers they believe are socially acceptable or desirable, rather than truthful responses

- In customer satisfaction surveys, respondents may exaggerate positive experiences to avoid appearing critical or negative
- In employee feedback surveys, respondents may inflate their ratings to maintain a favorable image within the organization

Impact of measurement bias



- It adheres to the principle of "garbage in, garbage out"
- Inaccuracies in measurement methods leads to flawed data inputs
- It inevitably results in unreliable outputs and flawed decision-making

Let's practice!

CONQUERING DATA BIAS

Mitigating bias in data collection

CONQUERING DATA BIAS

Konstantinos Kattidis
Data Analytics Lead

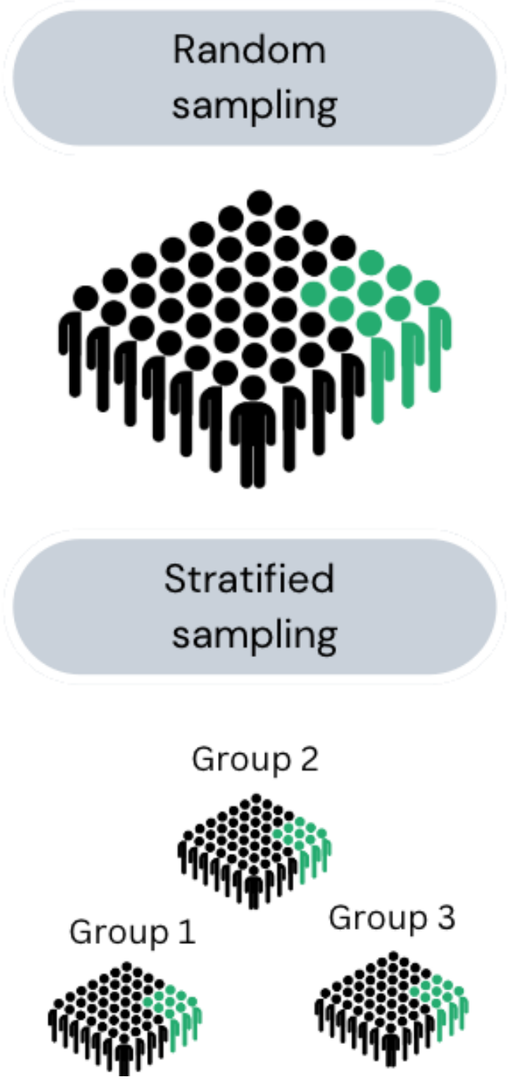
datacamp

Identifying bias in data collection

- Selection bias, historical bias and measurement bias
- Understanding these biases creates **awareness**, enabling data experts to proactively **identify** them and take action

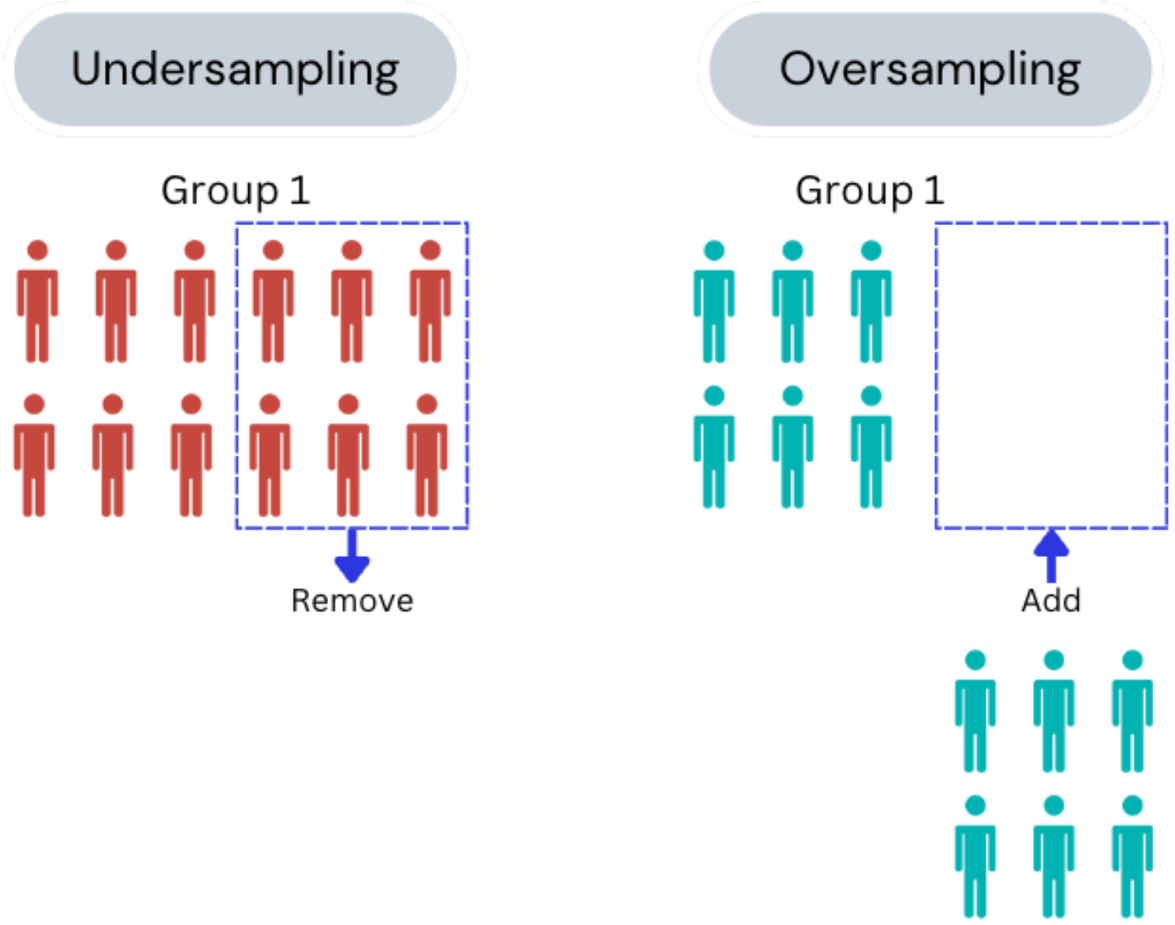
- **Sensitivity analysis** involves exploring how different assumptions, alternative subgroups, or weighting strategies affect the analysis results
- **External validation** compares data against independent sources to check for consistency and accuracy

Random and stratified sampling



- Selecting an appropriate sampling technique is important
- **Random sampling** involves selecting individuals or data points from a population randomly
- **Stratified sampling** divides the population into subgroups and then selects samples from each subgroup

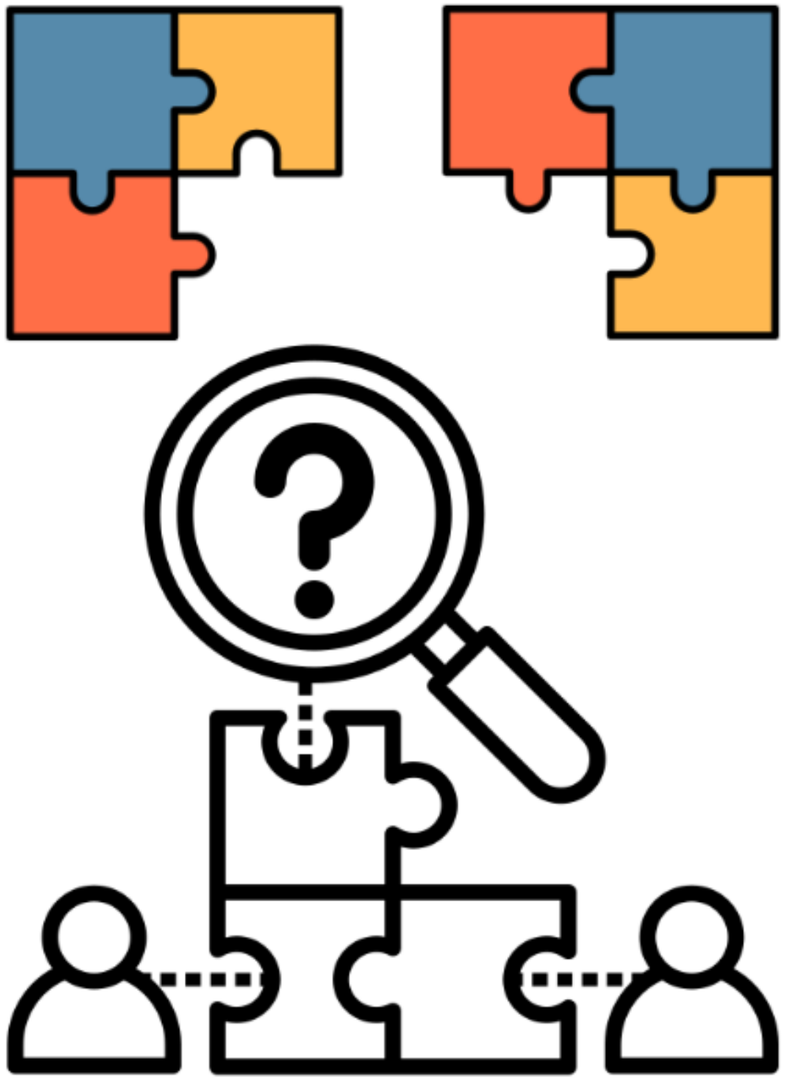
Balancing subgroup representation



- **Oversampling** involves deliberately increasing the representation of certain groups or classes in a dataset to balance the distribution
- **Undersampling** involves reducing the representation of overrepresented groups to achieve a more balanced dataset
- **Weighting** involves assigning different weights to observations based on their importance, compensating for any imbalances in the sample distribution

Data augmentation

- To address historical bias, this technique **enriches the dataset with additional data points**
- The aim is to cover underrepresented periods or events
- It includes:
  - Filling data gaps
  - Diversifying perspectives
  - Updating and correcting errors



Data measurement practices

- Standardization of measurement tools and protocols
- Training and calibration of data collectors
- Pilot testing can be used to assess the accuracy and consistency of data collection procedures
- Regular quality assurance checks and automation of processes can further enhance data quality

Continuous monitoring and adjustment

- Continuous monitoring and adjustment are essential to address emerging biases
- Regular reviews of data quality metrics
- Bias assessments
- These enable immediate identification of biases

Let's practice!

CONQUERING DATA BIAS