



Welcome to this session: Causal Inference

The session will start shortly...

Questions? Drop them in the chat.
We'll have dedicated moderators
answering questions.





What is Safeguarding?

Safeguarding refers to actions and measures aimed at protecting the human rights of adults, particularly vulnerable individuals, from abuse, neglect, and harm.



**To report a safeguarding concern reach out to us via email:
safeguarding@hyperiondev.com**

Live Lecture Housekeeping:

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly.
- No question is daft or silly - ask them!
- For all non-academic questions, please submit a query:
www.hyperiondev.com/support
- To report a safeguarding concern reach out to us via email:
safeguarding@hyperiondev.com
- If you are hearing impaired, please kindly use your computer's function through Google chrome to enable captions.



Learning Outcomes

- ❖ **Discuss** the concept of **Causal Inference**
- ❖ **Identify** and **Apply Causal Inference Methods**
- ❖ **Assess Causal Relationships** in Data
- ❖ **Explain challenges** in Causal Inference



Which of the following best describes the difference between correlation and causation?

- A. Correlation means one variable causes another, while causation means they just move together
- B. Causation means one variable influences another, while correlation only shows they move together
- C. Correlation and causation are the same
- D. Correlation always implies causation



Which of the following methods is most commonly used to determine causality?

- A. Randomized controlled trials (RCTs)
- B. Observing two variables move together
- C. Running a simple regression analysis
- D. Using only historical data to infer causality

Lecture Overview

- Introduction
- Causal graphs
- Key Techniques
- Real world Applications
- Challenges





Introduction to Causal Inference

Definition

- According to wikipedia, **causal inference** is the process of determining the **independent, actual effect** of a particular phenomenon that is a component of a larger system.
- Causal inference is the scientific process of determining **cause-and-effect relationships** between variables.
- It goes beyond correlation to establish whether one variable directly influences another.
- Without a rigorous approach, it is easy to make incorrect causal claims.

Correlation vs Causation

- One of the most fundamental misconceptions in data science is mistaking **correlation** for **causation**.
- Correlation simply measures the **relationship between two variables**
- Causation implies that **one variable directly influences the other**.
- **Example:** Ice cream sales and drowning incidents are correlated, but ice cream does not cause drowning. The underlying confounder is temperature. Hotter weather increases both ice cream sales and swimming activities.

Correlation vs Causation

CORRELATION WITH CAUSATION

Hey, Cee! Wanna go to the gym with me?

Why???



So we can get in shape!

YAY!!



CORRELATION WITHOUT CAUSATION

Hey, Cee! Wanna go to the gym with me?

Why???



So we can get more food!



Correlation vs Causation

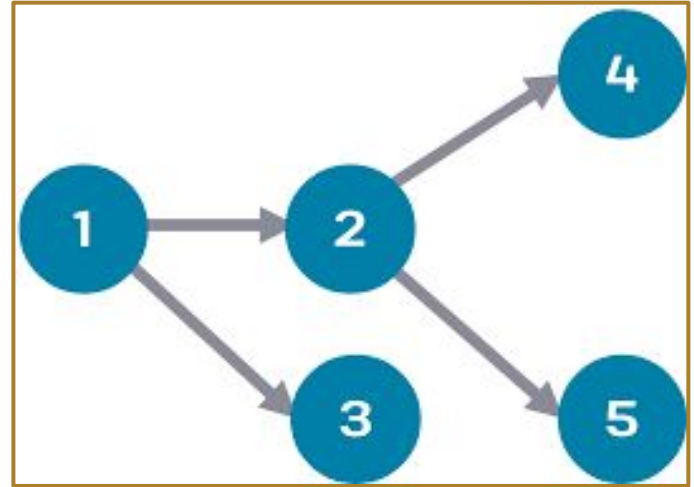
- In data-driven decision-making, understanding causality helps:
 - **Predict Outcomes More Accurately**: Helps in making better policy or business decisions.
 - **Avoid Spurious Relationships**: Identifies real causal effects rather than coincidental correlations.
 - **Optimize Interventions**: Determines what actions will yield the desired outcomes.



Causal Graphs: Directed Acyclic Graphs (DAGs)

Understanding DAGs

A Directed Acyclic Graph (DAG) is a graphical representation of causal relationships between variables. Nodes represent variables, and directed edges (arrows) indicate causal effects.

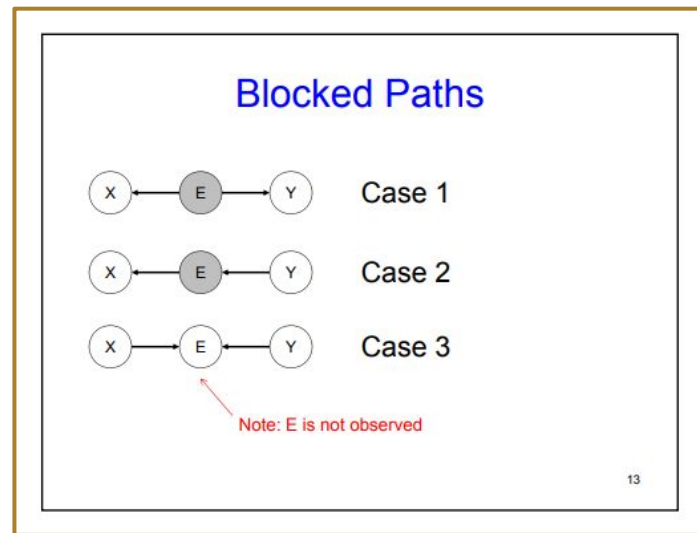


Example of a DAG

- Suppose we want to determine the effect of **Exercise (X)** on **Weight Loss (Y)**.
- A potential confounder is **Diet (Z)**.
- The DAG might look like:
 - $X \rightarrow Y$
 - $Z \rightarrow X$
 - $Z \rightarrow Y$
- Here, **Diet (Z)** affects both **Exercise (X)** and **Weight Loss (Y)**, meaning failing to account for it may lead to incorrect conclusions about Exercise's effect on Weight Loss.

D-separation and Blocking Paths

- **D-separation** is a method to determine if two variables are independent given another variable.
- **Blocking paths:** If all backdoor paths (non-causal paths) between X and Y are blocked by controlling for confounders, we can estimate the causal effect accurately.

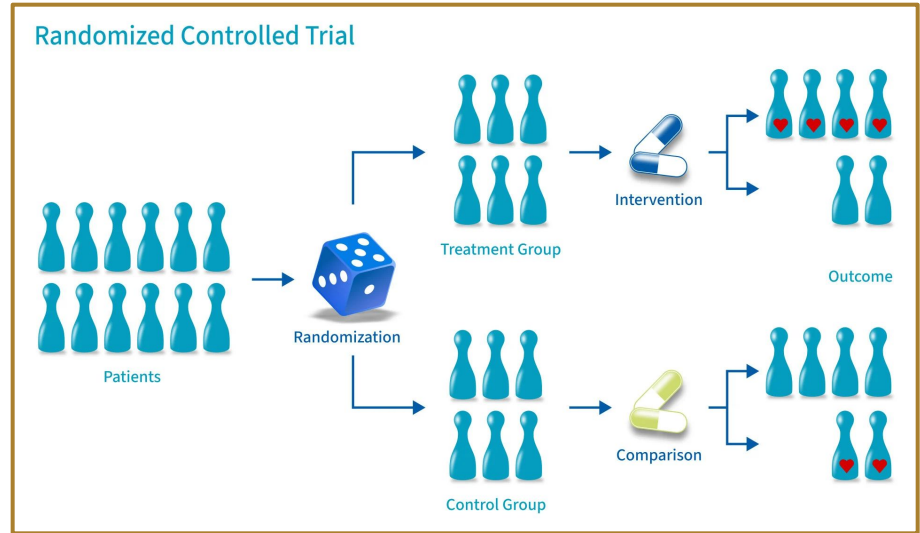




Key Techniques for Causal Inference

Randomized Controlled Trials (RCTs)

The gold standard for causal inference is **randomized experiments**, where subjects are randomly assigned to treatment and control groups, ensuring that differences in outcomes are due to the intervention and not confounding factors.

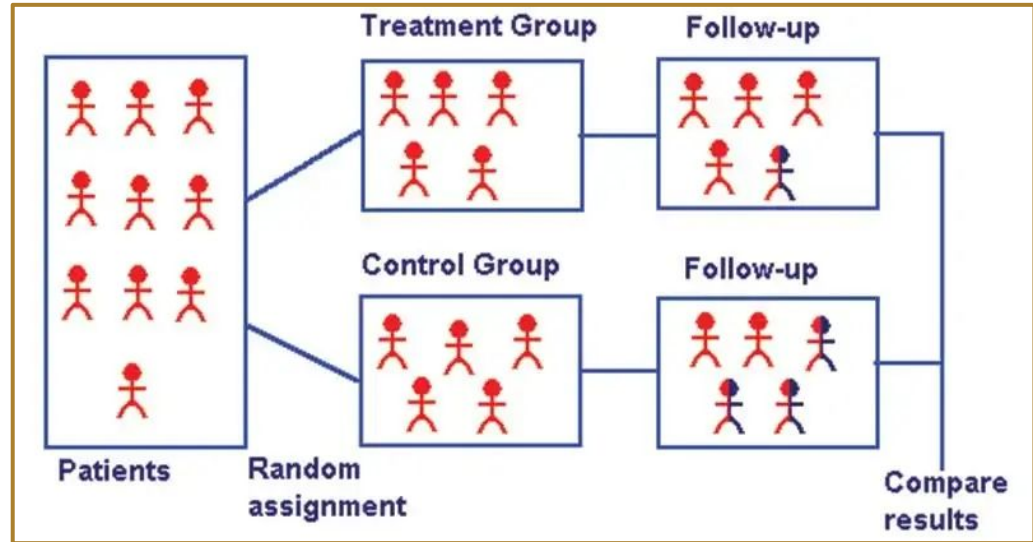


Example

A pharmaceutical company tests a new drug.

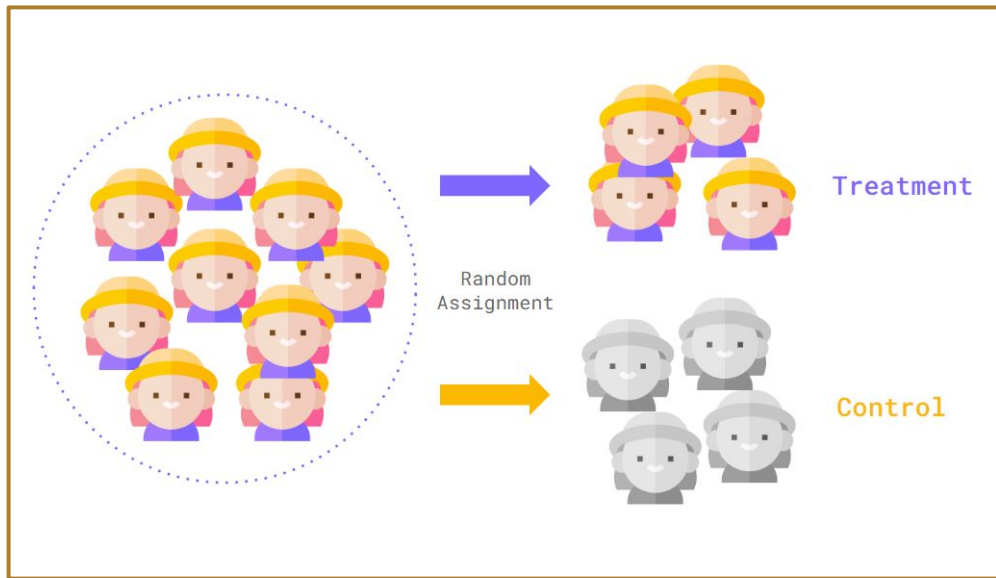
Patients are randomly assigned to receive either the drug or a placebo.

Randomization ensures that differences in outcomes are due to the treatment rather than confounding factors.



Example

A training program is introduced to improve student performance. Students are randomly assigned to receive the program (treatment group) or not (control group). By comparing their average scores, researchers can quantify the program's impact.





BREAK

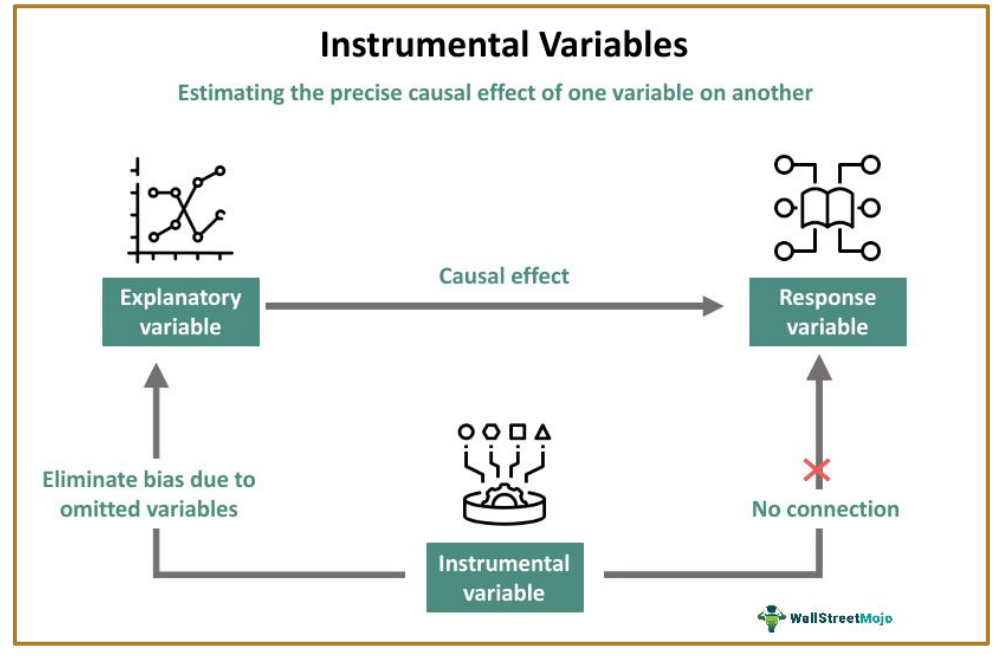


Limitations of RTCs

- Expensive and time-consuming.
- Ethical concerns (e.g., withholding treatment from some patients).
- Not always feasible in real-world settings.
- RCTs are powerful but challenging to implement, especially in business settings where customer behavior cannot always be controlled.

Instrumental Variables (IV)

When randomization is not possible, instrumental variables help estimate causal effects by introducing a variable that influences the treatment but not the outcome directly.



Instrumental Variables (IV)

- Example:

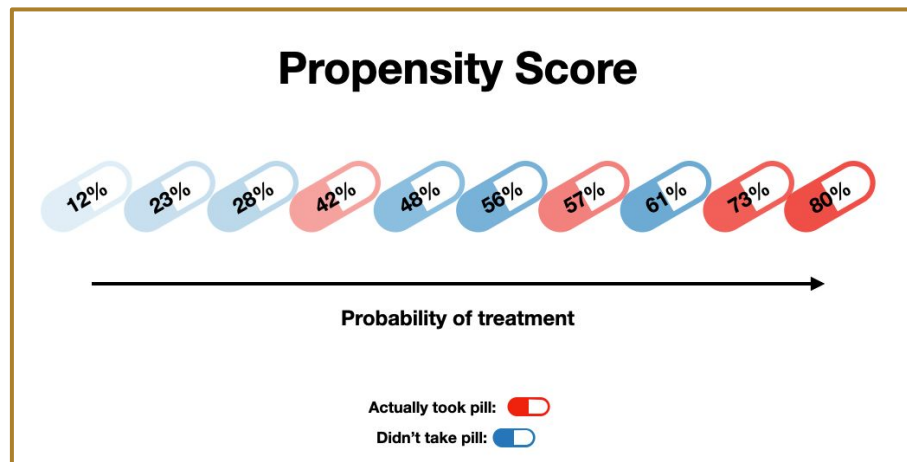
- To measure the effect of education on income, we can use proximity to a college as an instrument. It influences education but does not directly affect income beyond its impact on education.

- Key Assumptions:

- **Relevance:** The instrument affects the treatment.
- **Exogeneity:** The instrument is not related to the outcome, except through the treatment.

Propensity Score Matching (PSM)

When randomization is not feasible, PSM attempts to match treated and untreated units based on their likelihood of receiving the treatment.



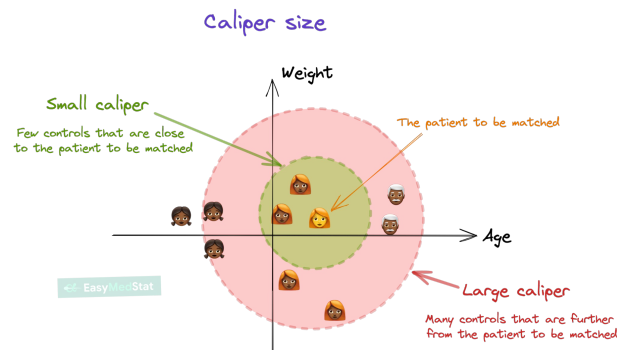
Propensity Score Matching (PSM)

- **Example:**

- In healthcare, comparing patients who received a new therapy with similar patients who did not.
- The propensity score is the probability of receiving treatment based on observed characteristics.

- **Limitations:**

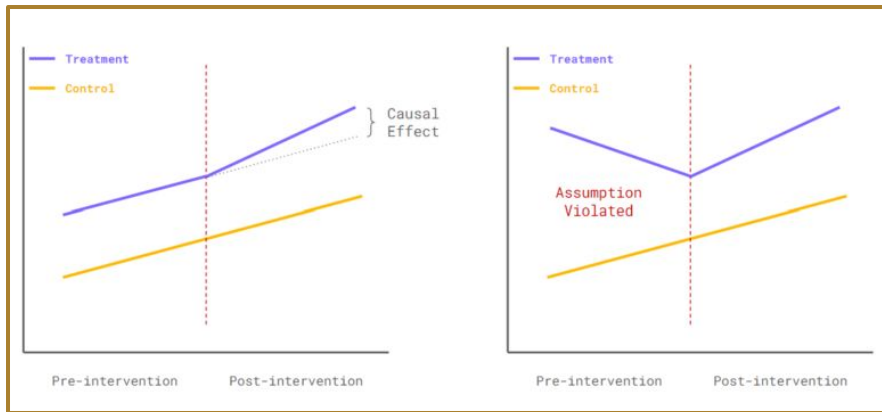
- Can only account for observable confounders.
- Matching quality depends on data availability.



Difference-in-Differences (DiD)

Agricultural Productivity

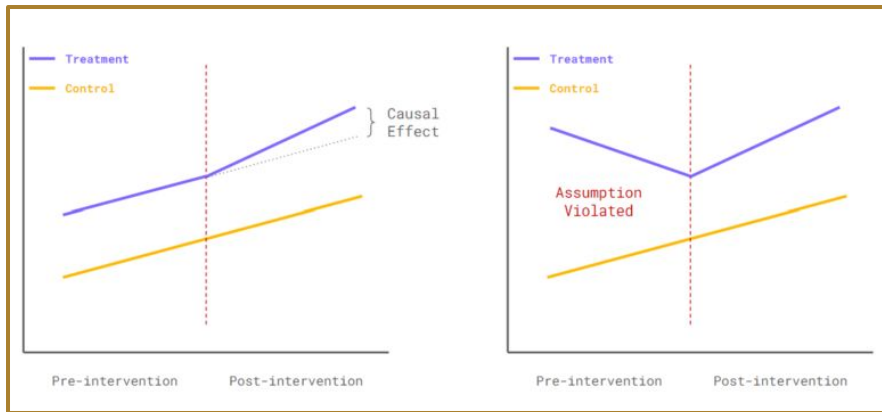
A new farming technique is introduced in one region (treatment) but not another (control). By comparing the productivity change in both regions before and after implementation, researchers estimate the technique's impact.



Difference-in-Differences (DiD)

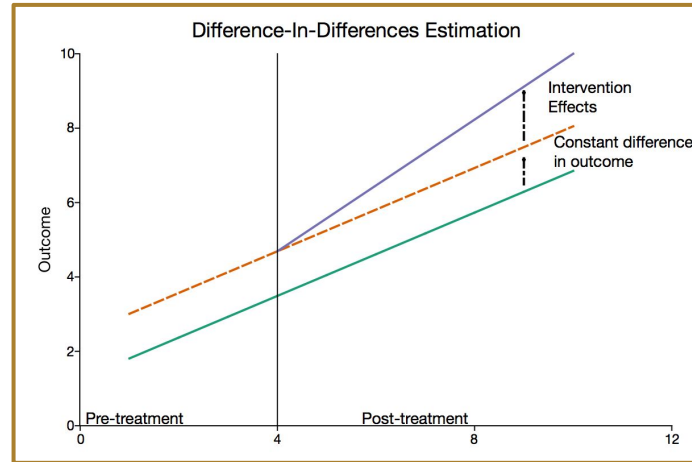
Key Assumption: Parallel Trends

DiD relies on the assumption that both groups would have followed similar trends in the absence of treatment. If violated, results can be biased.



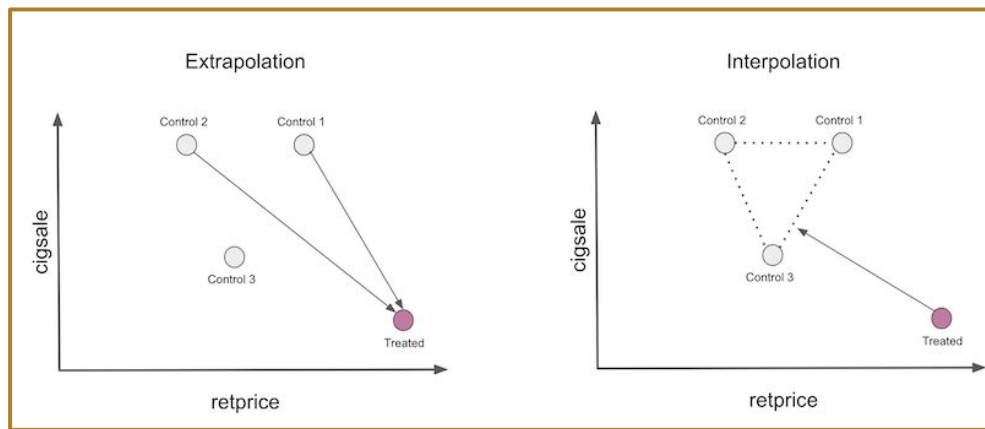
Difference-in-Differences (DiD)

A quasi-experimental approach used when randomization is not possible. This method compares changes over time between a treatment and a control group.



Synthetic Control (SC)

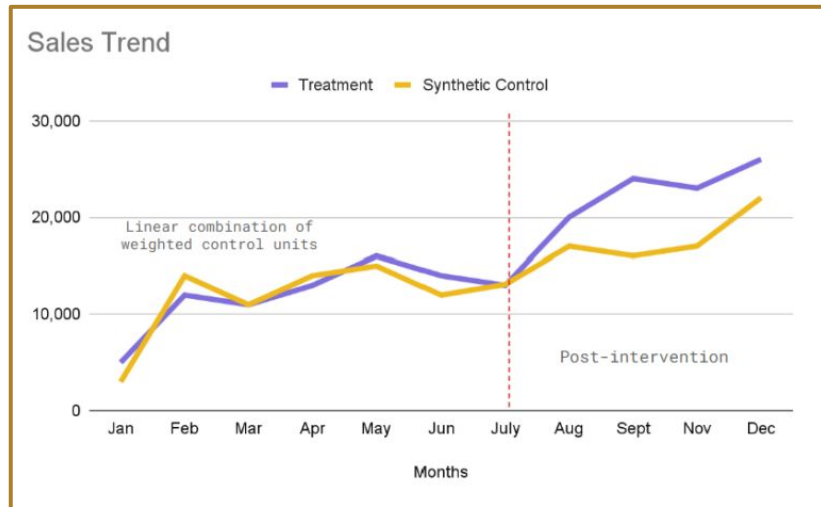
SC is used when a single treated unit (e.g., a state or a company) undergoes an intervention. Instead of using a traditional control group, SC creates a weighted combination of multiple control units to serve as a synthetic counterfactual.



Synthetic Control (SC)

Marketing Campaign Impact

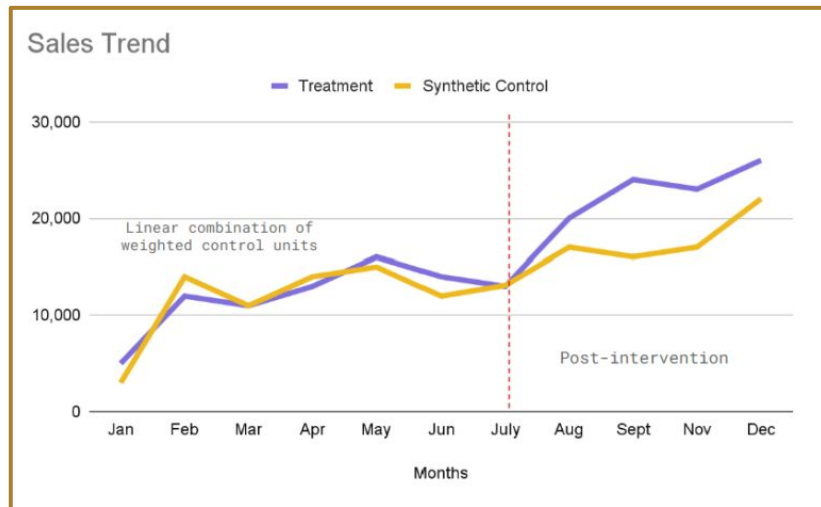
A company runs a major marketing campaign in Region A. To assess the impact, researchers construct a synthetic version of Region A using data from other similar regions. The difference in sales between actual and synthetic Region A estimates the campaign's true effect.



Synthetic Control (SC)

Limitations

SC requires rich historical data and careful selection of control units to ensure validity.





Real-World Applications

Real World Applications

- **Healthcare**

- **Example:** Determining if a new drug reduces heart attacks.
- **Approach:** Use RCTs or instrumental variables if randomization is infeasible.

- **Policy Evaluation**

- **Example:** Does increasing the minimum wage reduce employment?
- **Approach:** Use difference-in-differences (DiD) by comparing employment trends before and after a policy change in treated and untreated regions.

Real World Applications

- Marketing

- **Example:** Measuring the impact of an online ad campaign on sales.
- **Approach:** Use A/B testing (randomized experiments) or synthetic control methods to estimate the ad's true effect.



Challenges in Causal Inference

Confounding Variables

- **Confounders** are variables that affect both the treatment and the outcome, leading to biased causal estimates.
- **Solution:** Use DAGs to identify and control for confounders via matching, regression, or instrumental variables.

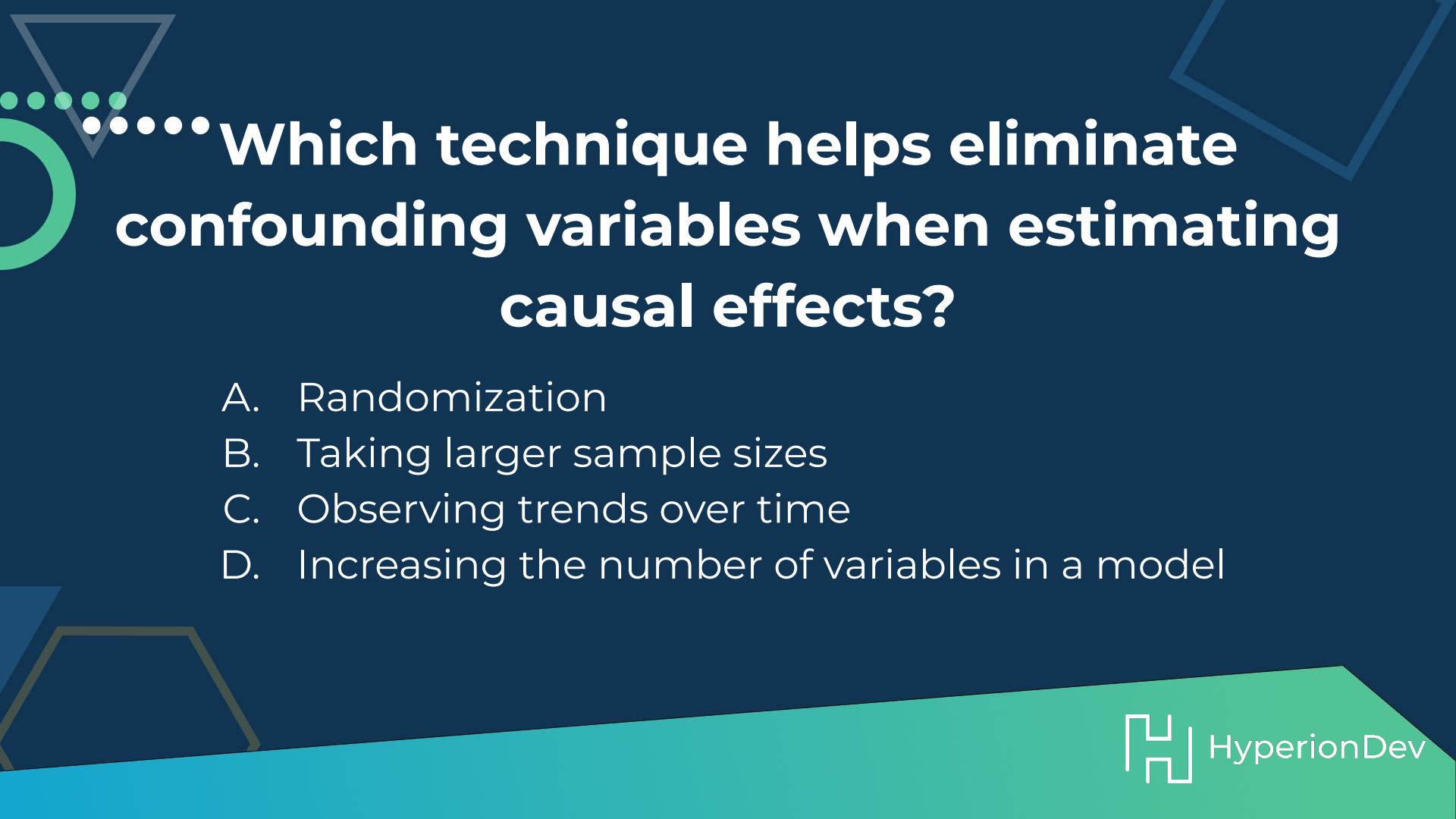
Omitted Variable Bias

- Occurs when an important variable is left out of the analysis, leading to incorrect causal conclusions.
- **Example:** Studying the effect of education on income without considering cognitive ability.
- **Solution:** Collect better data and use proxy variables if direct measurement is unavailable.



Sample Selection Bias

- When the sample is not representative of the entire population, results may not generalize.
- **Example:** Studying the impact of an executive MBA program on salary by only looking at enrolled students (ignoring those who couldn't afford or qualify).
- **Solution:** Use techniques like Heckman correction or apply reweighting methods to adjust for selection bias.



••••• Which technique helps eliminate confounding variables when estimating causal effects?

- A. Randomization
- B. Taking larger sample sizes
- C. Observing trends over time
- D. Increasing the number of variables in a model

Q & A SECTION

**Please use this time to ask
any questions relating to the
topic, should you have any.**

**Thank you
for attending**



HyperionDev