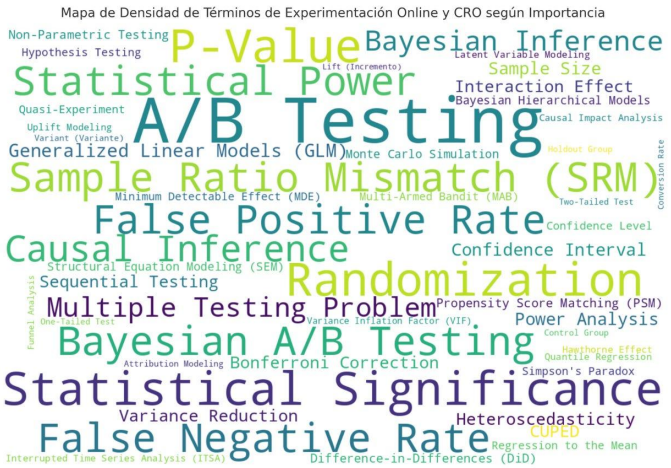
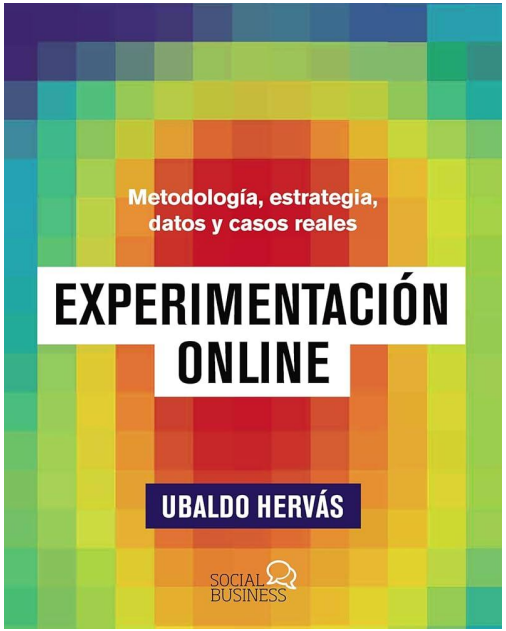
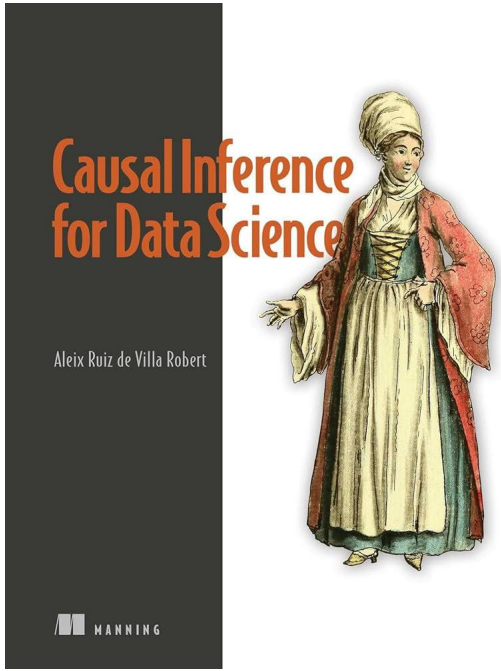
Causal Inference:
a brief introduction.

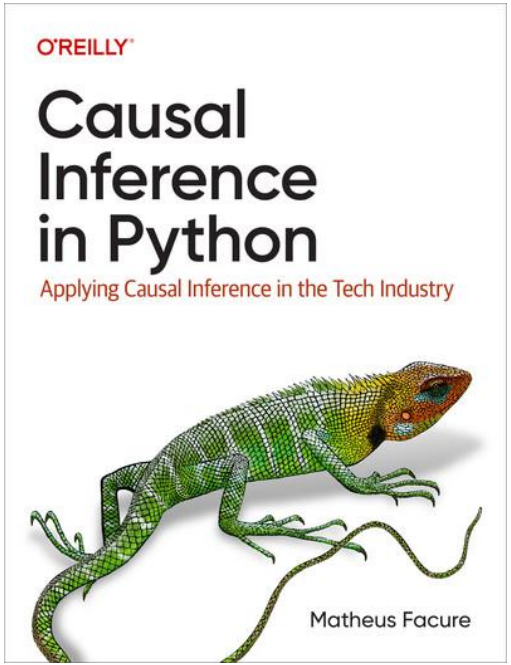
@Ubaldo Hervás

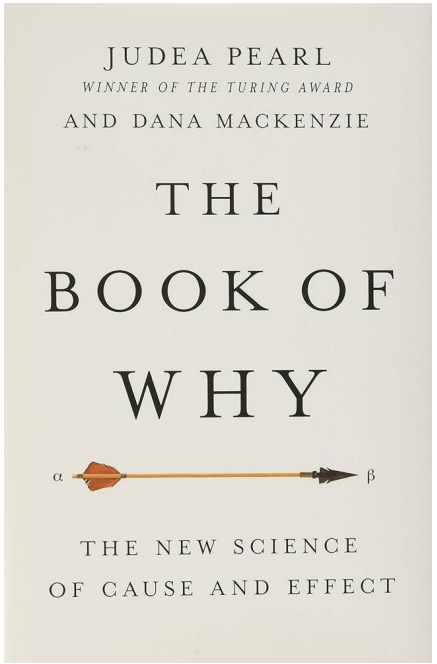

Mapa de Densidad de Términos de Experimentación Online y CRO según Importancia
Non-Parametric Testing
P-Value
Bayesian Inference
Hypothesis Testing
Latent Variable Modeling
Lift (Incremento)
Sample Size
Statistical Power
Interaction Effect
Bayesian Hierarchical Models
Quasi-Experiment
A/B Testing
Causal Impact Analysis
Uplift Modeling
Variant (Variante)
Generalized Linear Models (GLM)
Monte Carlo Simulation
Holdout Group
Conversion Rate
Sample Ratio Mismatch (SRM)
Minimum Detectable Effect (MDE)
Multi-Armed Bandit (MAB)
Two-Tailed Test
False Positive Rate
Confidence Level
Causal Inference
Confidence Interval
Structural Equation Modeling (SEM)
Randomization
Sequential Testing
Funnel Analysis
Multiple Testing Problem
Propensity Score Matching (PSM)
One-Tailed Test
Variance Inflation Factor (VIF)
Power Analysis
Bayesian A/B Testing
Control Group
Hawthorne Effect
Quantile Regression
Attribution Modeling
Bonferroni Correction
Simpson's Paradox
Statistical Significance
Variance Reduction
Heteroscedasticity
False Negative Rate
CUPED
Regression to the Mean
Interrupted Time Series Analysis (ITSA)
Difference-in-Differences (DiD)


Analítica de datos con Python para marketing digital
Desarrollo de KPIs en negocios *online*
Joseba Ruiz Diez
Ubaldo Hervás Olvera
IA
Σ
ANAYA


Metodología, estrategia, datos y casos reales
EXPERIMENTACIÓN ONLINE
UBALDO HERVÁS
SOCIAL BUSINESS


Causal Inference for Data Science
Aleix Ruiz de Villa Robert
MANNING


O'REILLY
Causal Inference in Python
Applying Causal Inference in the Tech Industry
Matheus Facure


JUDEA PEARL
WINNER OF THE TURING AWARD
AND DANA MACKENZIE
THE BOOK OF WHY
α
β
THE NEW SCIENCE OF CAUSE AND EFFECT

The menu:

Theory (45 min):

1. What is causation?
2. RCTs
3. DAGs

Practice (45 min):

4. How can we start?
5. Five steps process
6. A real case (hotels in Madrid, minibar margin and difference in differences) + coding!
7. Conclusions!

What is causation?

“We may define a cause to be an object, followed by another, [...] where if the first object had not been, the second never had existed”

David Hume

Two kind of dependence:

Temporal:

“The cause precedes the effect in time.”

Two kind of dependence:

Temporal:

“The cause precedes the effect in time.”

Counterfactual:

“If the cause had not occurred, the effect would no have occurred either.”

Causal Quantities of interest:

T —> Treatment Variables (binary):

T = 1 —> The user did something
T = 0 —> The user did not something

Causal Quantities of interest:

T —> Treatment Variables (binary):

T = 1 —> The user did something
T = 0 —> The user did not something

Y —> Observed Outcome Variable:

Y = 0 —> The user did not achieve something (eg a purchase)
Y = 1 —> The user did achieve something (eg a purchase)

The question here is...

Y(1) = What the outcome variable Y would be if you did something.

Y(0) = What the outcome variable Y would be if you didn't.

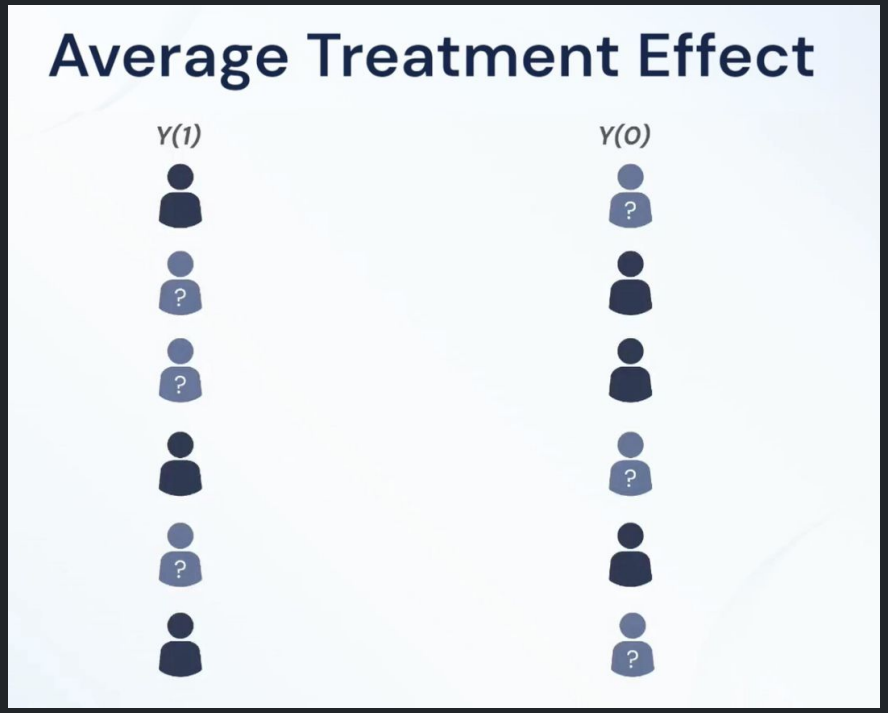
Y(1) - Y(0) = The result of our research

The fundamental problem of causal inference:

You can only observe one of these potential outcomes [Y(1) or Y(0)].
If the user does something (a click) you will observe Y(1).

You will “never” be able to exactly calculate this quantity. We will recreate an approximation.

The fundamental problem of causal inference:



We can never observe both worlds.

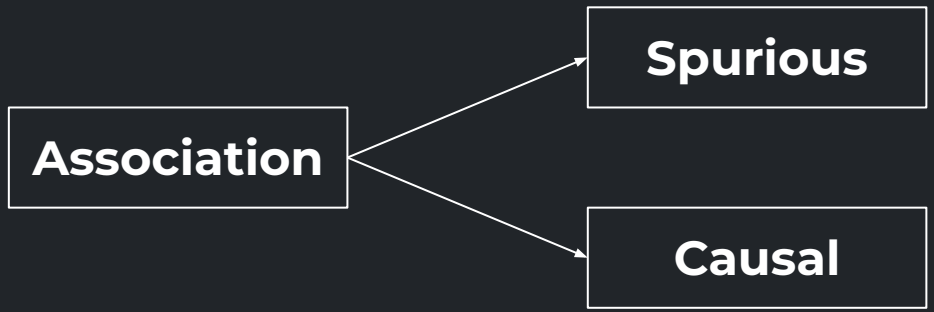
For each user, we only see **Y(1)** *or* **Y(0)**. Never both.

Causal inference replaces the missing outcome with its expected value, **allowing us to approximate the effect of the treatment across comparable groups.**

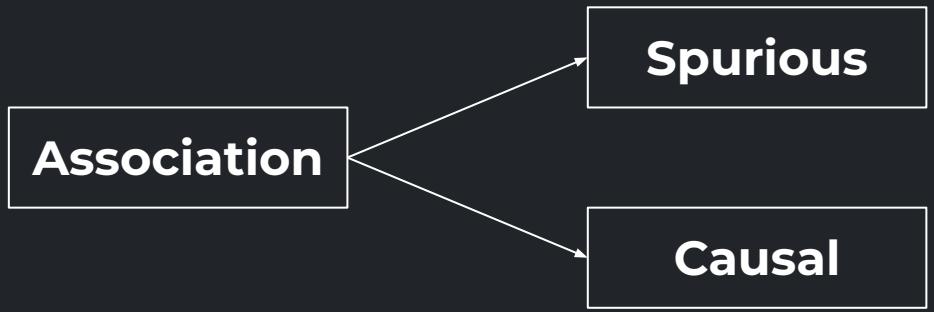
RCT: Randomized Controlled Trials

@Ubaldo Hervás

What is the relationship between association and causation?



What is the relationship between association and causation?



Association = Causation + Bias

What is the relationship between association and causation?

If Bias = 0, then causation equals association and thus:

Association = Causation

What is the relationship between association and causation?

If Bias = 0, then causation equals association and thus:

Association = Causation

$$E[Y(1)] - E[Y(0)] = E[Y \mid T=1] - E[Y \mid T=0]$$

If we eliminate bias we can use observational data to estimate causal effects.

What about RCTs?

We do, mostly:

E[Y|T=1] - E[Y|T=0]

What about RCTs?

We do, mostly:

E[Y|T=1] - E[Y|T=0] 

With randomization we achieve covariance balance.

It means that the distribution of covariates X is the same across treatment groups.

What about RCTs?

We do, mostly:

E[Y|T=1] - E[Y|T=0] ⟶

With randomization we achieve covariance balance.

It means that the distribution of covariates X is the same across treatment groups.

If done well, association = causation with RCTs

Most important assumptions (RCTs):

1. Exchangeability / ignorability: states that the potential outcomes (Y1 and Y0) are independent of Treatment (T).

Most important assumptions (RCTs):

2. Consistency: states that if the treatment is T = t, the observed outcome Y is the potential outcome under treatment T = t.

Example: Website A or B. Y=1 (user is satisfied). Y=0 user is not satisfied. But... The AB test works better on Mark's laptop and it's slower on Windows laptop. **It introduces bias.**

Most important assumptions (RCTs):

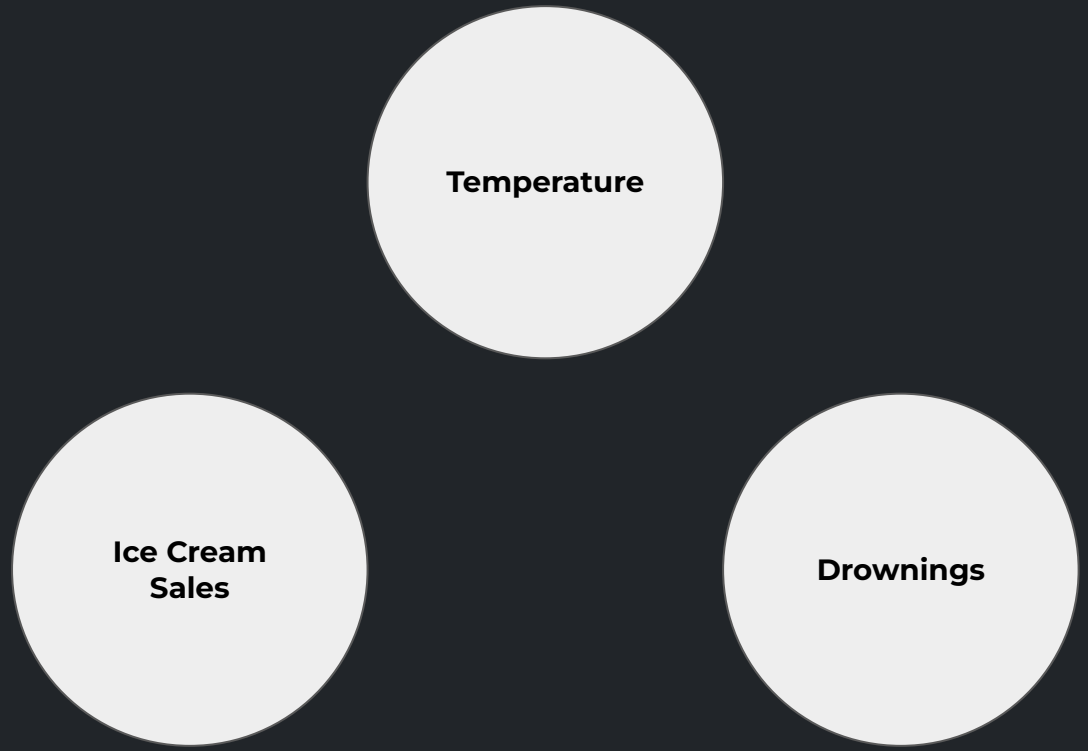
3. No interference: states that the treatment assigned to one unit does not affect the potential outcomes of other units.

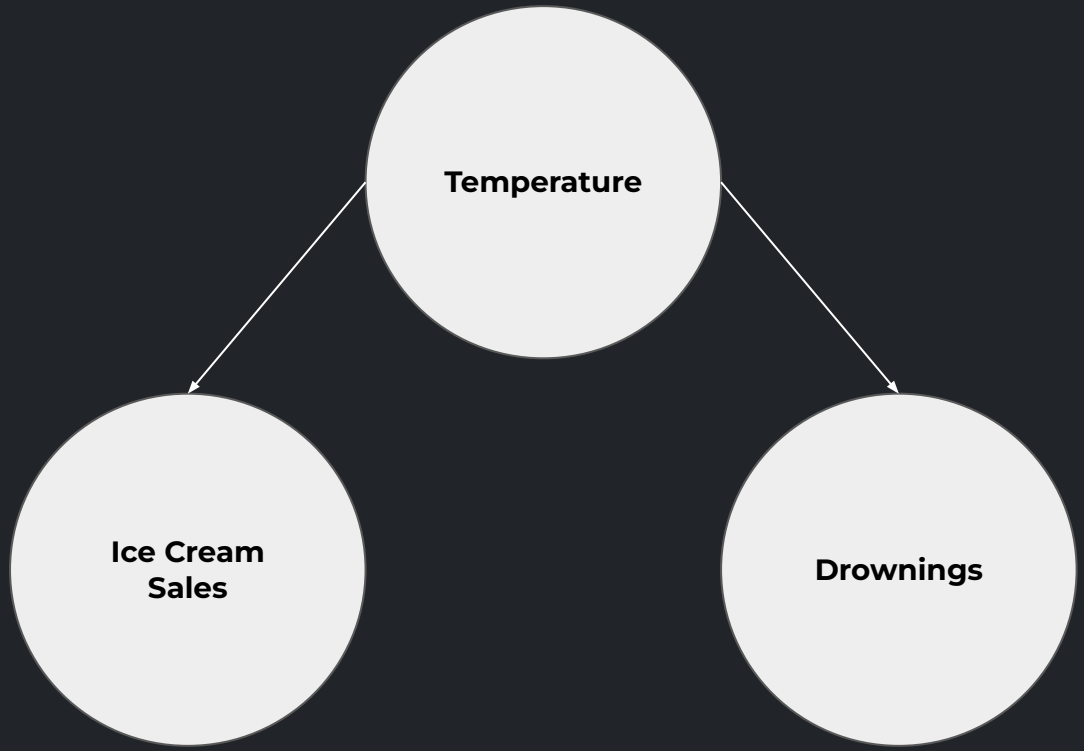
If we can not run
AB tests,
then we will use
observational data

DAGs: Directed Acyclic Graphs

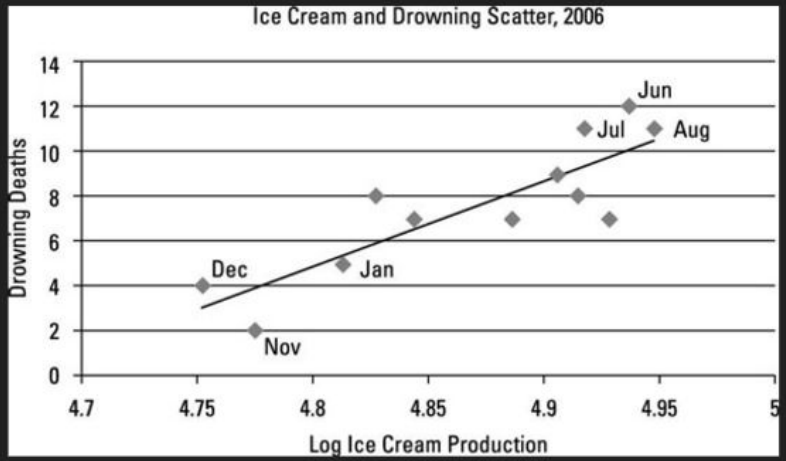
@Ubaldo Hervás

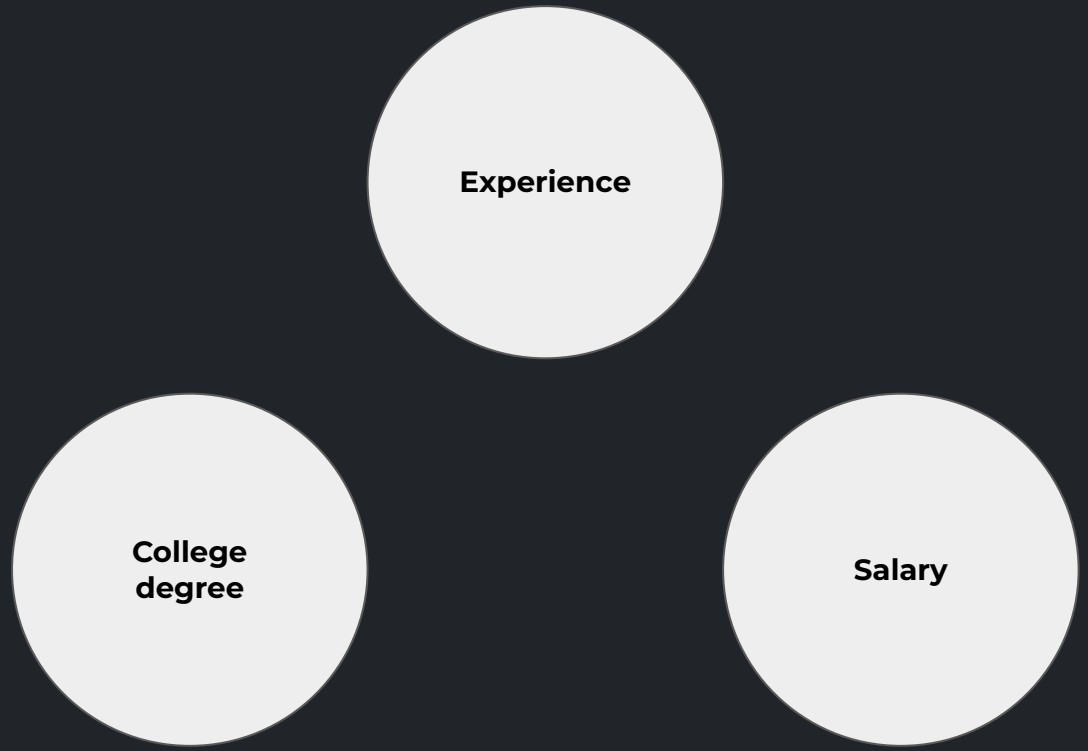
What is the causal relationship between these variables?


Temperature
Ice Cream Sales
Drownings


Temperature
Ice Cream Sales
Drownings

Ice Cream Sales VS Drowning Deaths

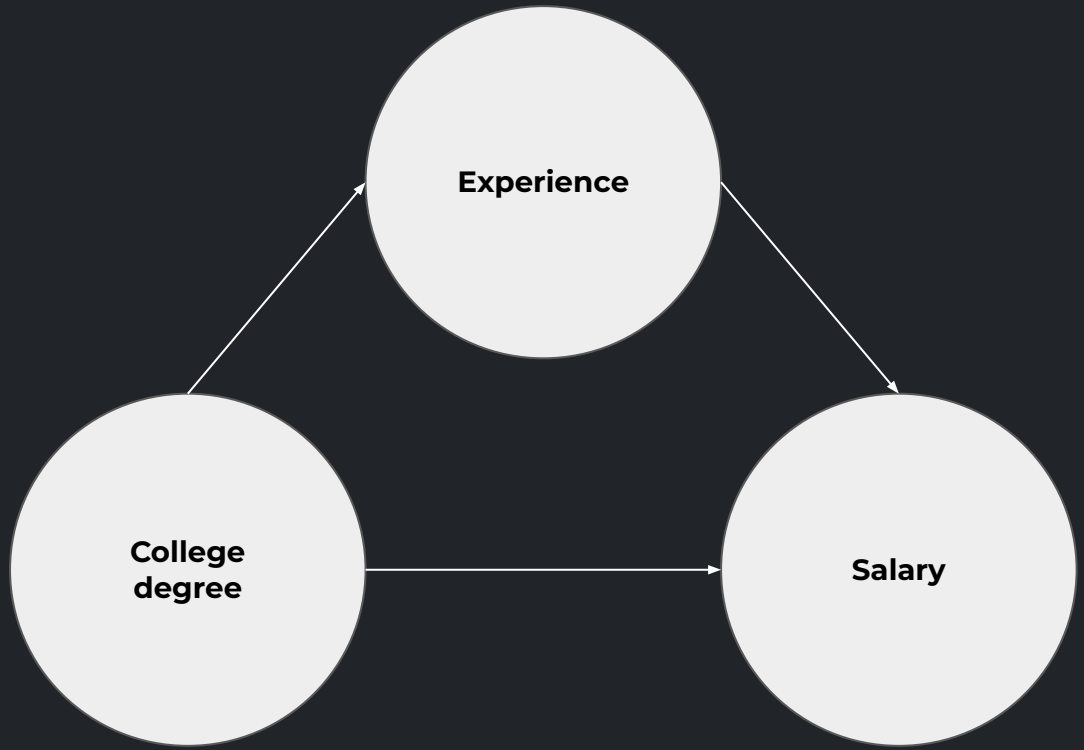



Experience
College degree
Salary

Experience

College degree

Salary


Experience
College degree
Salary

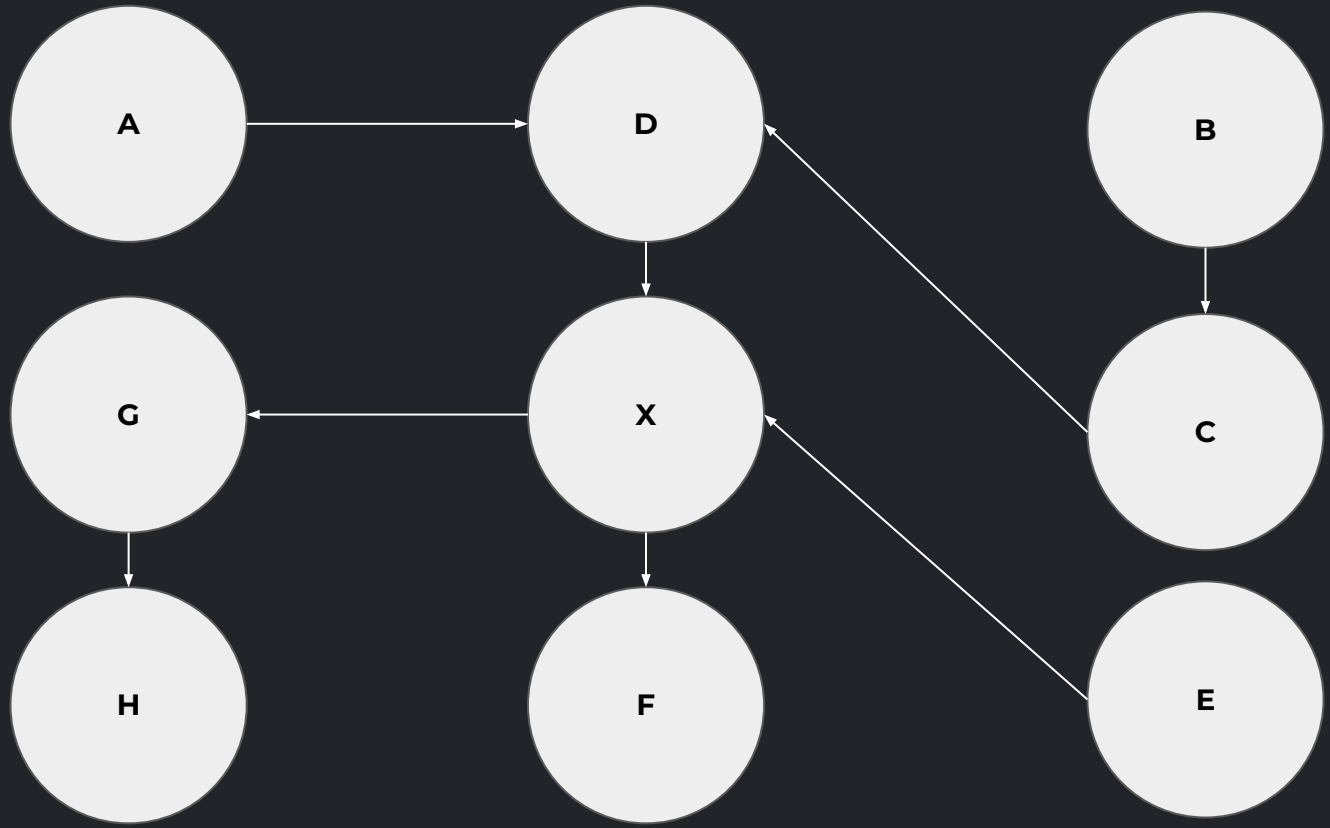
The industry knowledge matters to establish relationships

The harder the industry or hypothesis we are working on, the more we ask the experts.

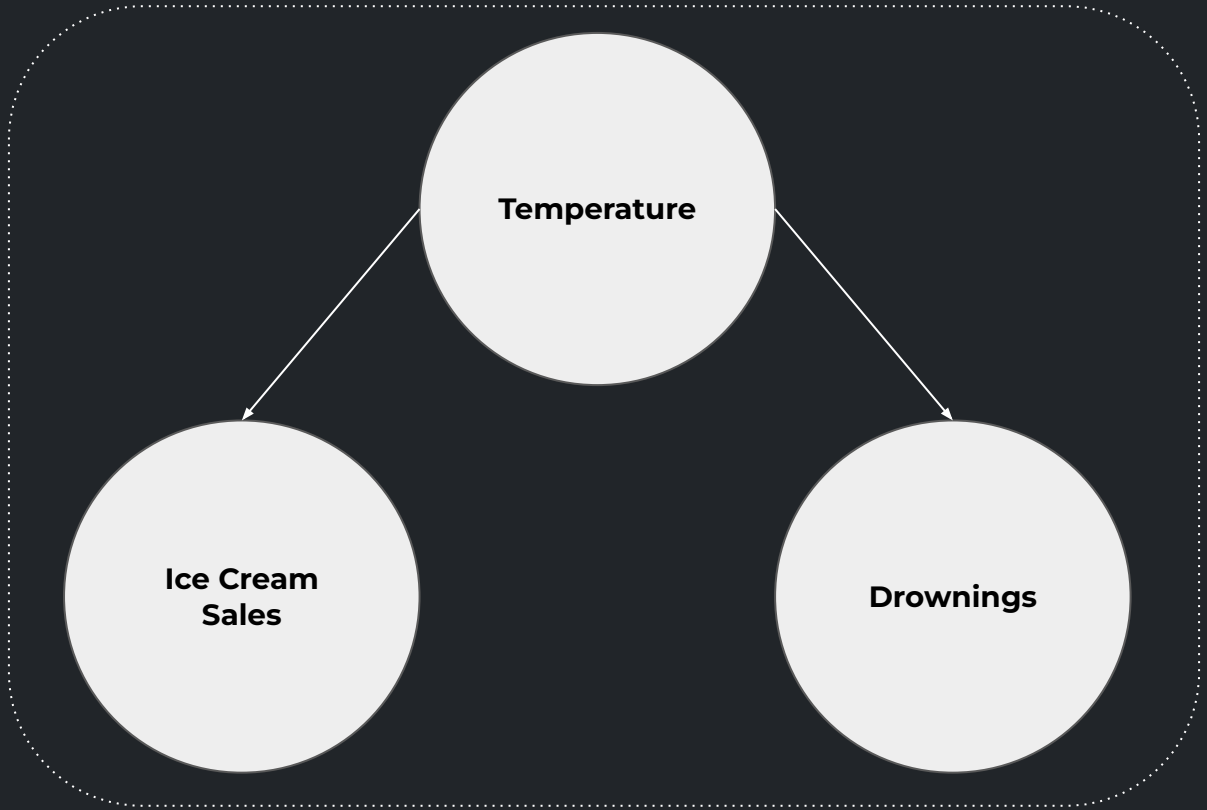
The industry knowledge matters to establish relationships

The harder the industry or hypothesis we are working on, the more we ask the experts.

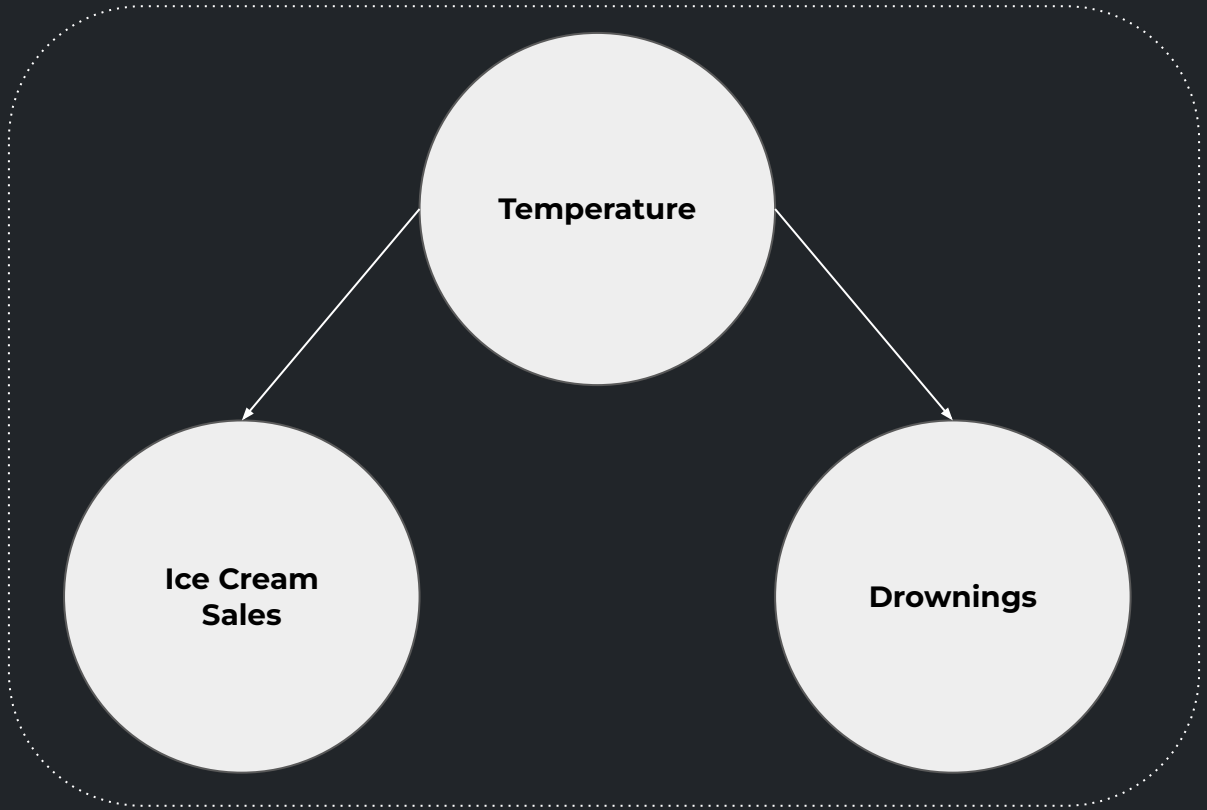
When we are running a causal inference project, we need to ask the experts because we are not able to establish theoretically causal relationships between variables.


A
D
B
G
X
C
H
F
E

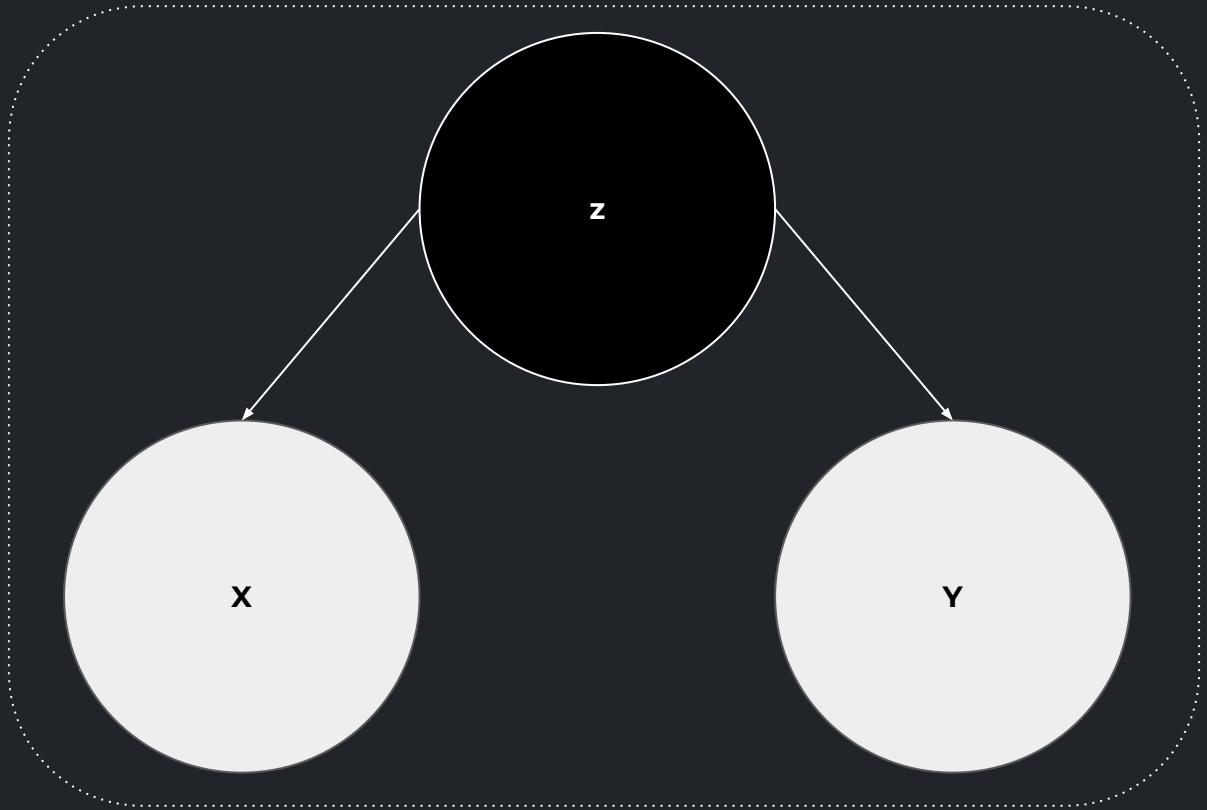
Spurious Association



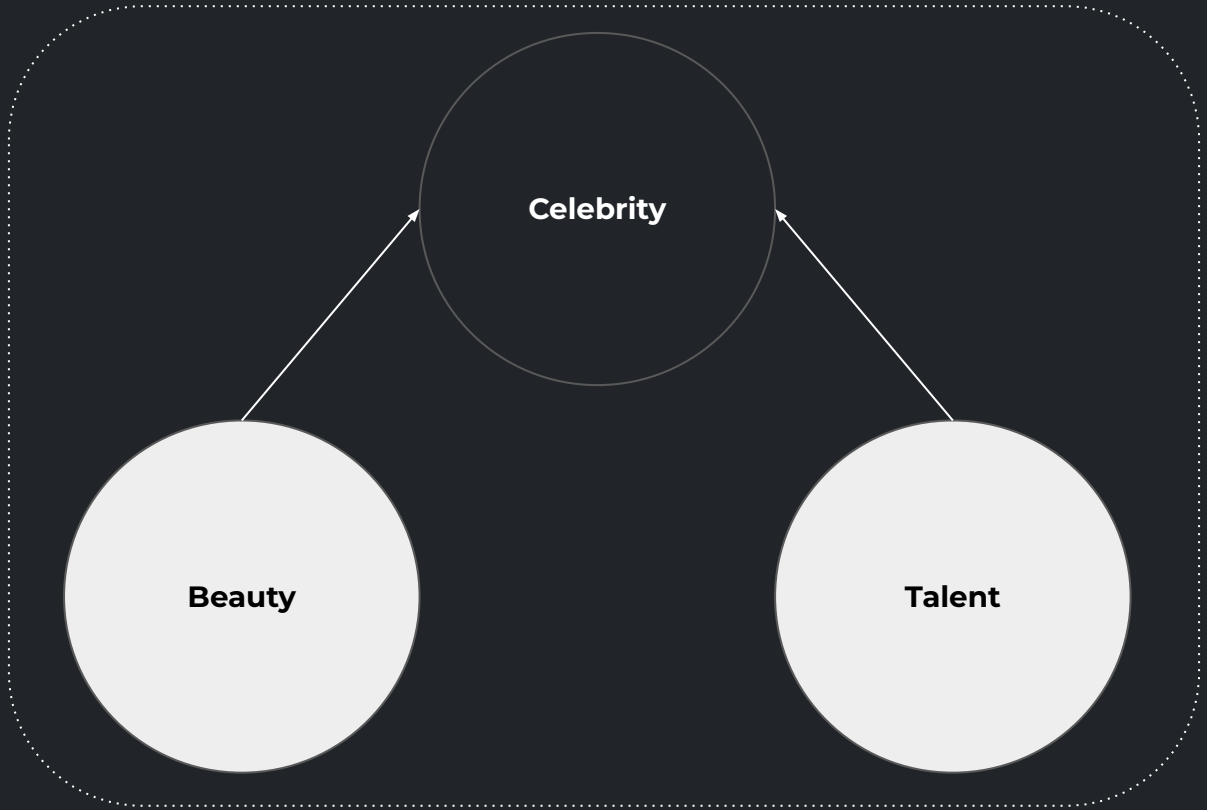
Temperature is a confounder



Confounder (z)



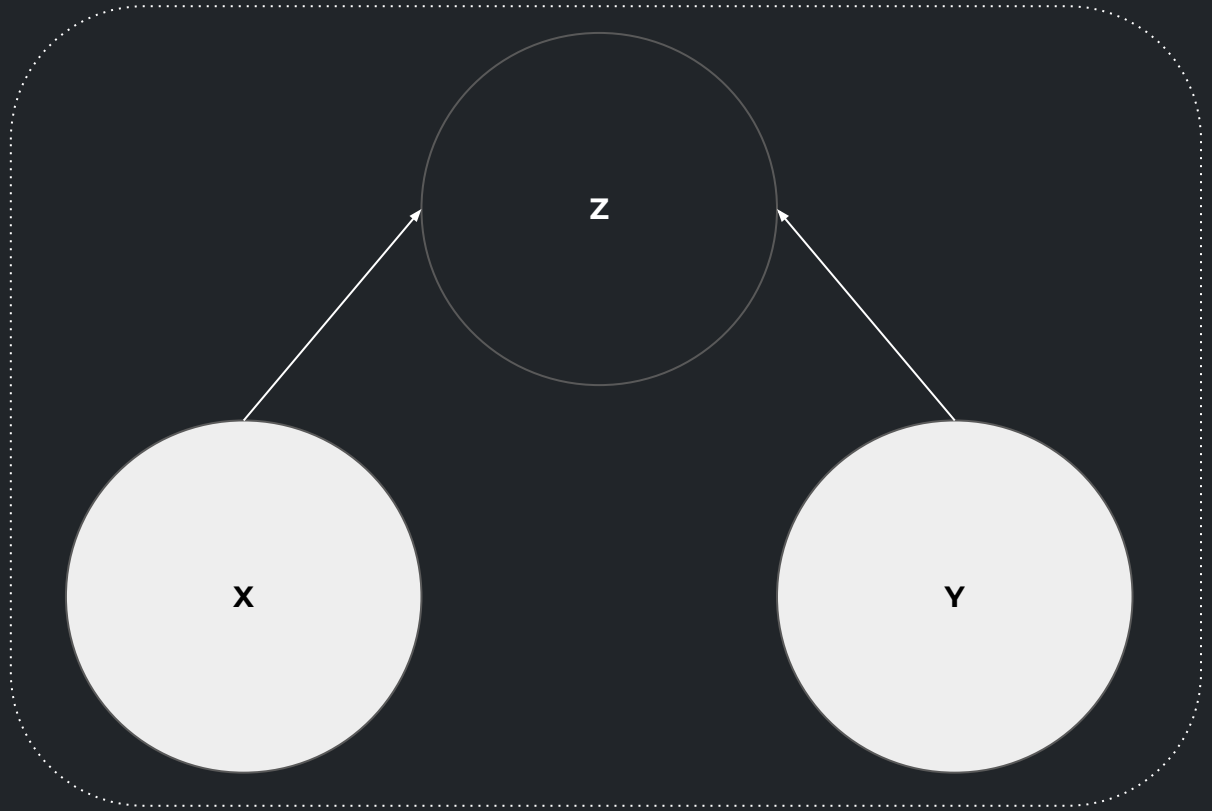
Collider (celebrity)



Colliders

The collider structure shows why we cannot arbitrarily include variables in our model.

Why? Conditioning on a collider can actually ***induce bias.***



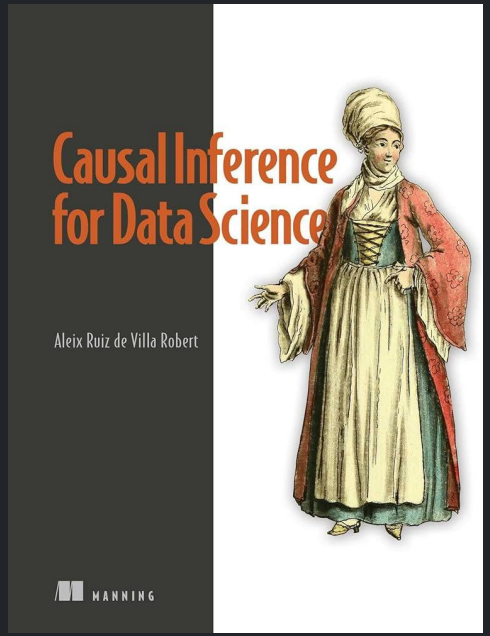
In general X and Y are independent, but they become dependent when we control for z.

Z = collider.

How can we start?

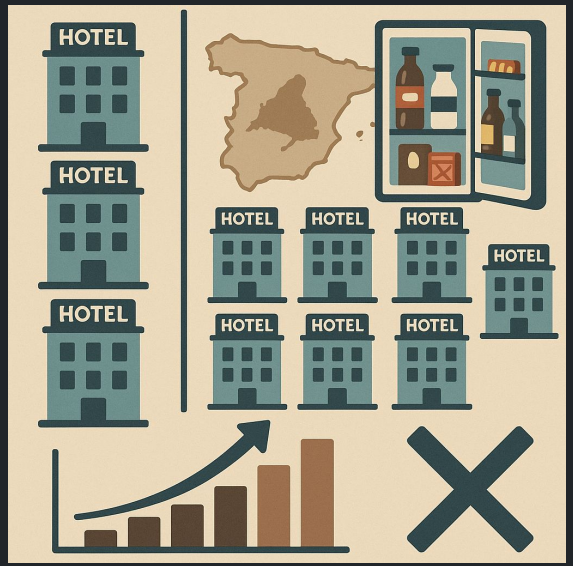
Five steps process

1. Determine whether you have experimental or observational data.
2. Understand your problem (context, variables, ...) as thoroughly as possible.
3. Create a model that describes how variables are related to each other.
4. Share your model, goal, and assumptions with the other people involved in the analysis, to get their input.
5. Apply causal inference techniques to your data and your model to answer your causal questions:
   a. State what objectives can be answered in your model.
   b. Estimate causal effects and confidence intervals using the appropriate formulas.



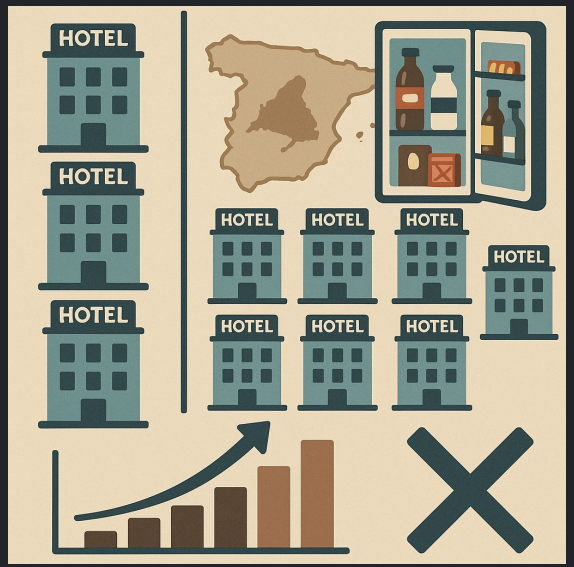
1. Understanding the problem

- Goal and context:
  - Evaluate whether upgrading minibar items hotels increases margin despite lacking random assignment across locations.
  - Only in Madrid (12 hotels) against 88 others hotels all around Spain.



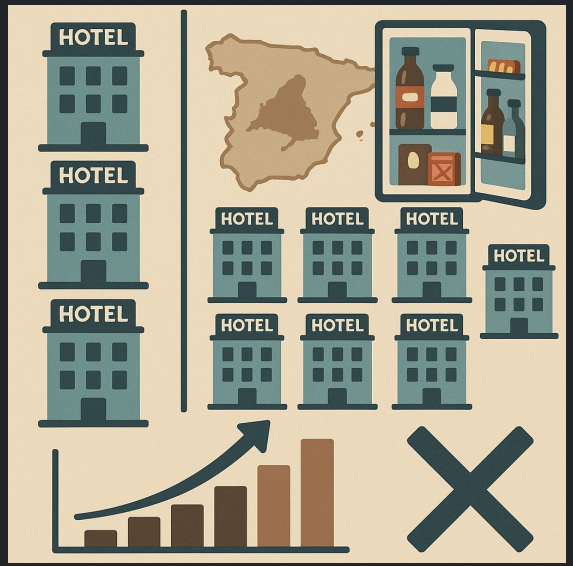
1. Understanding the problem

- Goal and context:
  - Evaluate whether upgrading minibar items hotels increases margin despite lacking random assignment across locations.
  - Only in Madrid (12 hotels) against 88 others hotels all around Spain.
  - Hyppo (Highest Paid Person's Opinion) attacks again!

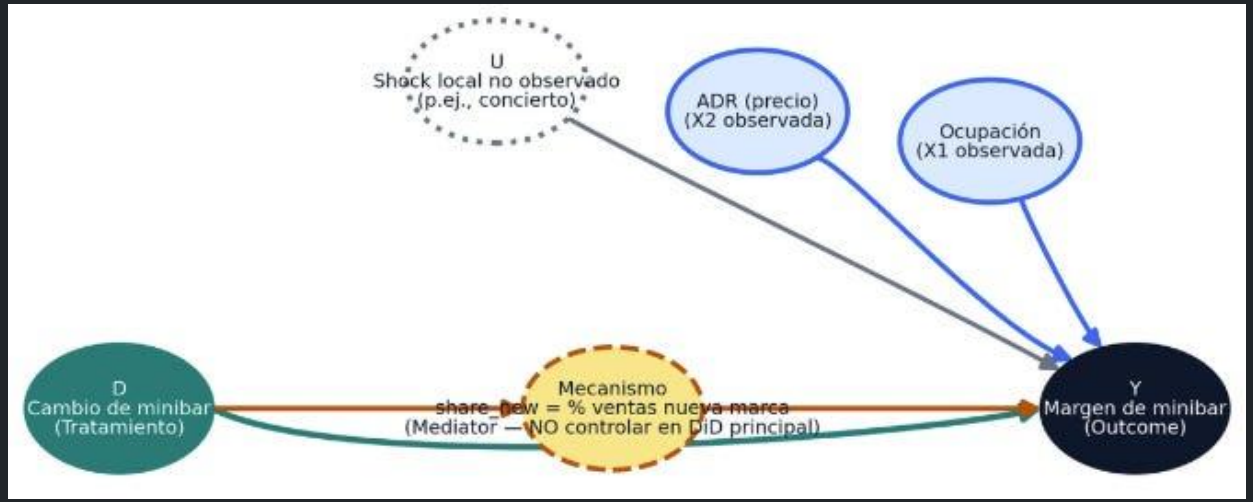


1. Understanding the problem

- Change applied only in Madrid
- No random assignment is possible (Hippo)
- We need alternatives to A/B testing

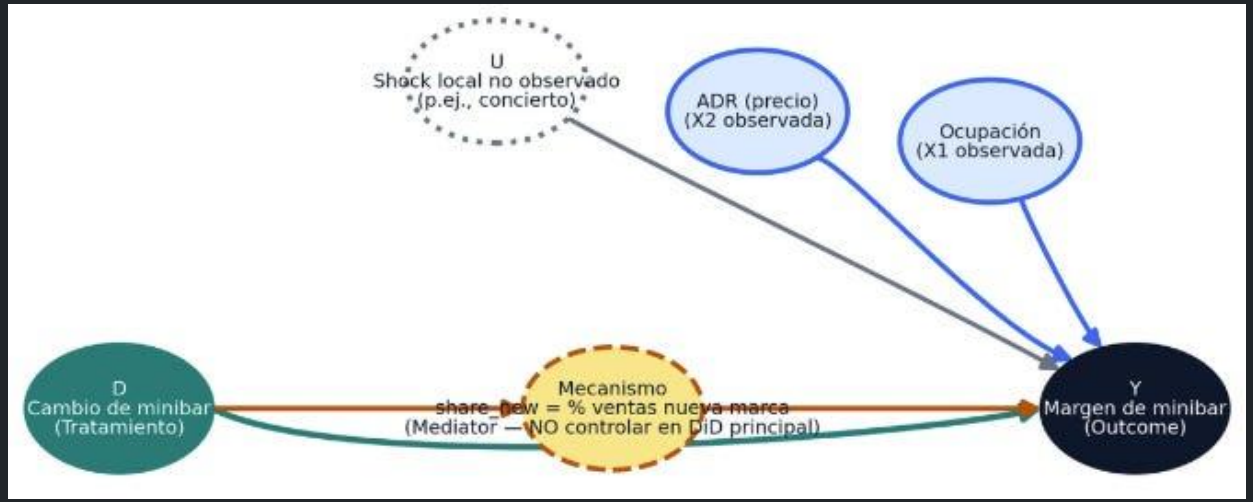


1. What other factors could influence “Margin”?



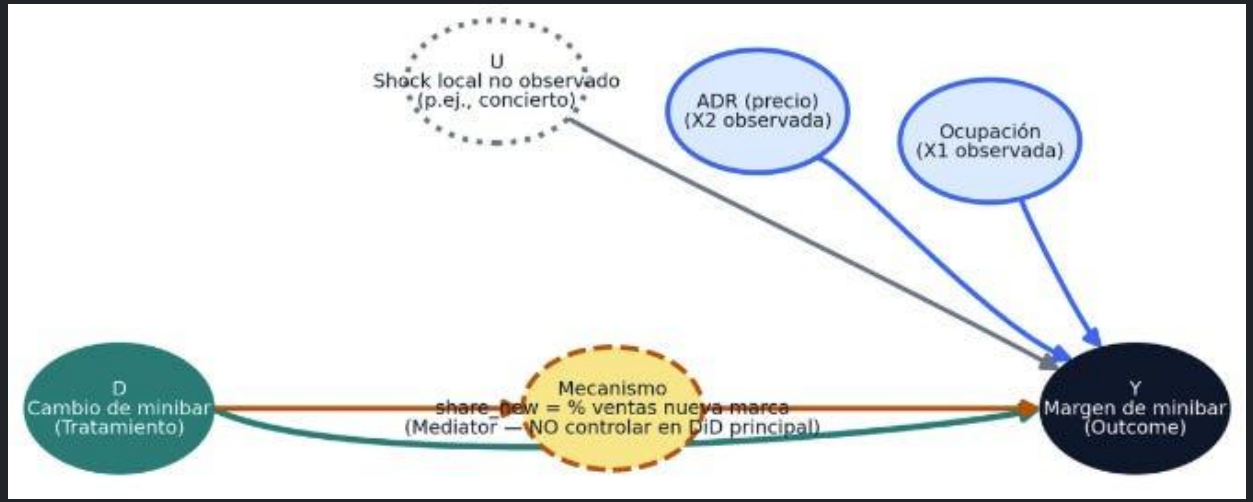
- **Occupancy** and **ADR** (average daily room revenue) causally influence margin.

1. What other factors could influence “Margin”?



- **Occupancy** and **ADR** (average daily room revenue) causally influence margin.
- **Minibar change** affects margin via mediator.
    - % sales new brand in the minibar.

1. What other factors could influence “Margin”?



- **What about “Unobserved”?**

Missing relevant variables (specially if the are confounders) is Achilles’ heel of Causal Inference.

1. What other factors could influence “Margin”?

Prev Sci. Author manuscript; available in PMC: 2014 Dec 1.

Published in final edited form as: Prev Sci. 2013 Dec;14(6):570–580. doi: 10.1007/s11121-012-0339-5

An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-Experimental Prevention Research

Weiwei Liu [1,✉], S Janet Kuramoto [2], Elizabeth A Stuart [3]

Author information Copyright and License information

PMCID: PMC3800481 NIHMSID: NIHMS470690 PMID: 23408282

Sensitivity Analysis for Unobserved Confounding

How to know the unknowable in observational studies

Ugur Yildirim

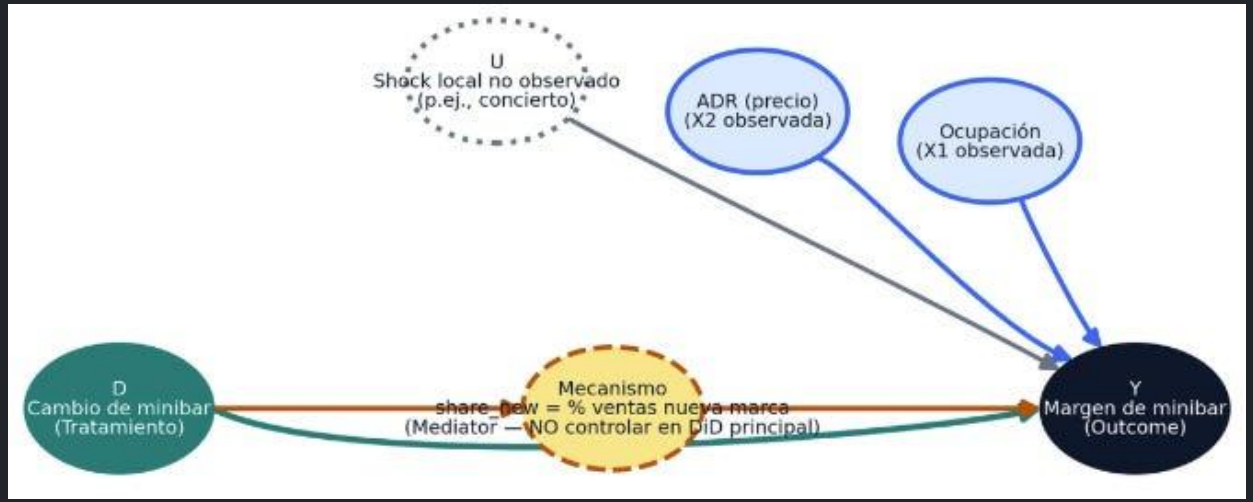
Feb 13, 2024 • 12 min read

- **What about “Unobserved”?**

In real-world scenarios, you may not have access to all of them.

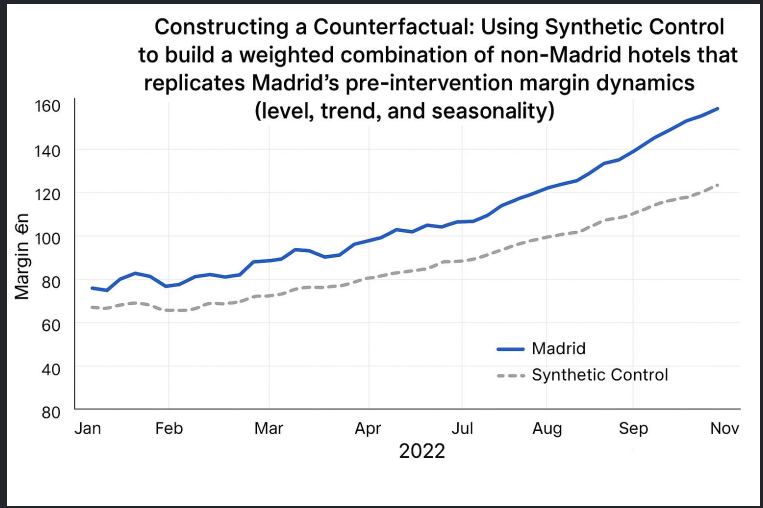
There are tools to analyze the sensitivity of your results to unmeasured confounders: sensitivity analysis.

2. Ask the expert and improve your model



3. How did we deal with this problem?

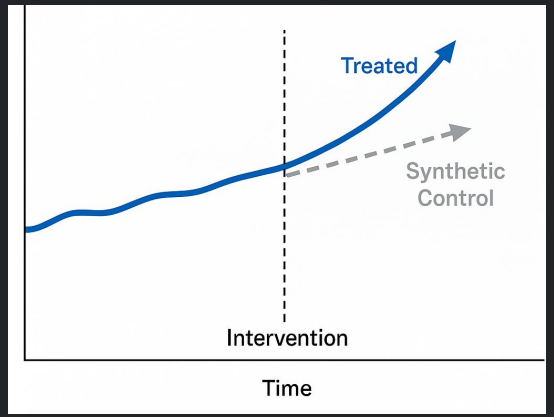
- **Constructing a counterfactual:** Using Synthetic Control to build a weighted combination of non-Madrid hotels that replicates Madrid's pre-intervention margin dynamics (level, trend, and seasonality).



3. How did we deal with this problem?

- **Constructing a counterfactual:** Using Synthetic Control to build a weighted combination of non-Madrid hotels that replicates Madrid's pre-intervention margin dynamics (level, trend, and seasonality).

- **We don't assume parallel trends:** We construct them via Synthetic Control and then statistically verify them through pre-trend tests, event-study diagnostics, and robustness checks.



3. How did we deal with this problem?

- **Constructing a counterfactual:** Using Synthetic Control to build a weighted combination of non-Madrid hotels that replicates Madrid's pre-intervention margin dynamics (level, trend, and seasonality).

- **We don't assume parallel trends:** We construct them via Synthetic Control and then statistically verify them through pre-trend tests, event-study diagnostics, and robustness checks.

- **Improve precision with valid covariates:** Include occupancy and ADR only as precision controls, after verifying they are not affected by the treatment to reduce noise without introducing bias.

3. How did we deal with this problem?

- **Constructing a counterfactual:** Using Synthetic Control to build a weighted combination of non-Madrid hotels that replicates Madrid's pre-intervention margin dynamics (level, trend, and seasonality).

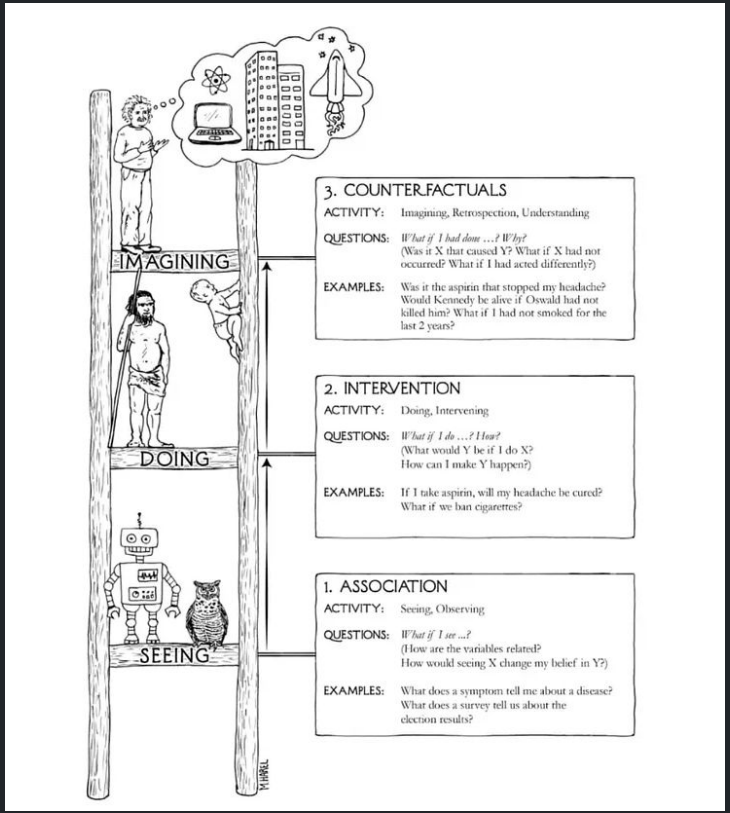
- **We don't assume parallel trends:** We construct them via Synthetic Control and then statistically verify them through pre-trend tests, event-study diagnostics, and robustness checks.

- **Improve precision with valid covariates:** Include occupancy and ADR only as precision controls, after verifying they are not affected by the treatment to reduce noise without introducing bias.

- **Stress-test the causal claim:** Validate assumptions with pre-trend checks, event-study plots, placebo tests, donut windows, and leave-one-city-out analyses to ensure robustness and rule out false causal signals.

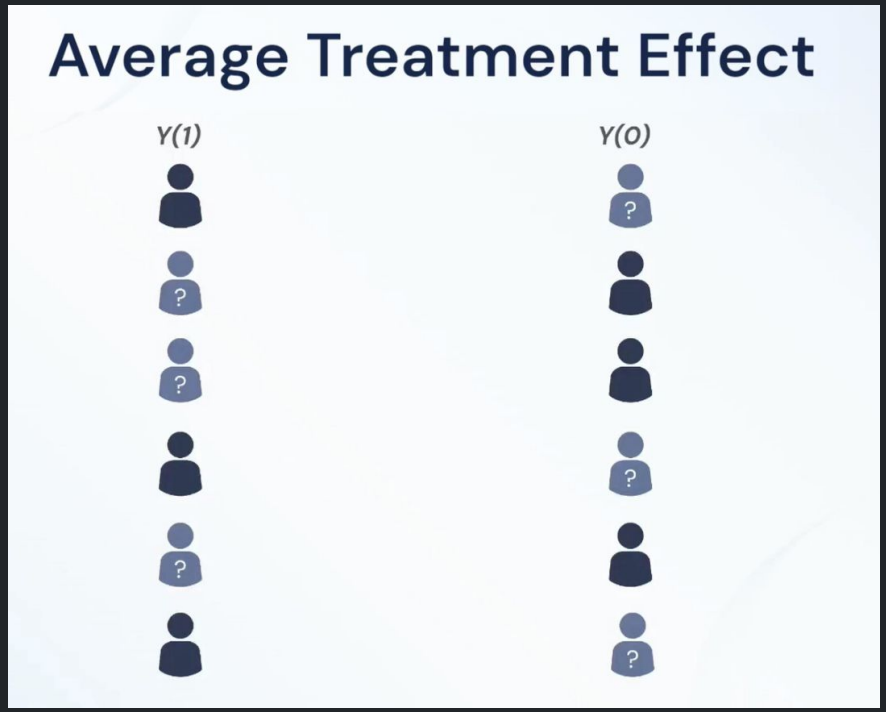
Let’s code!

Conclusions



Judea Pearl (The book of why)

The fundamental problem of causal inference:



We can never observe both worlds.

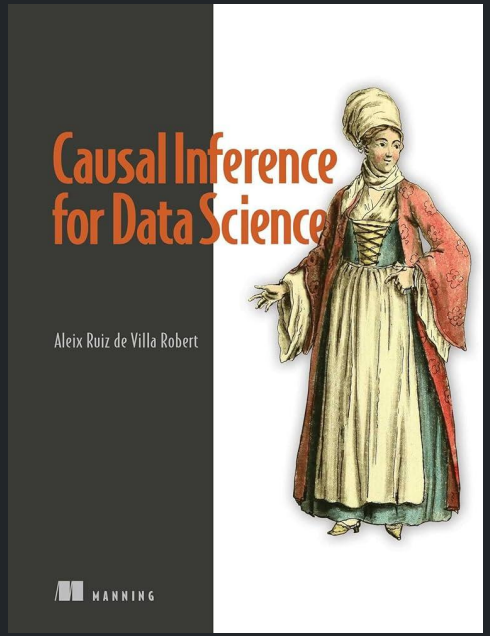
Experience

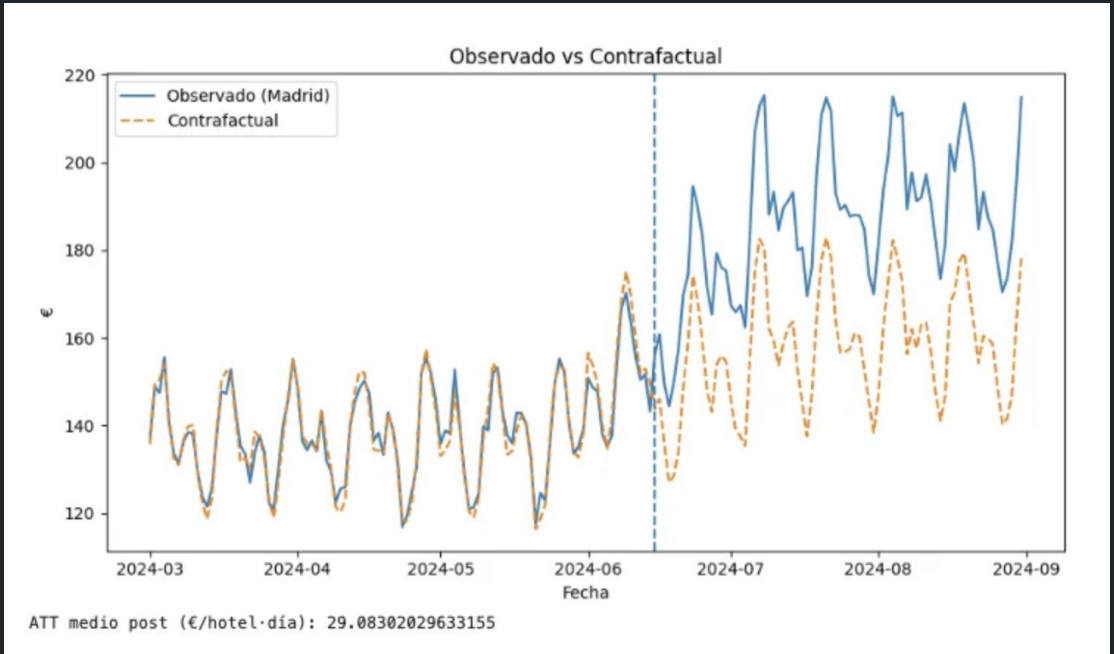
College degree

Salary

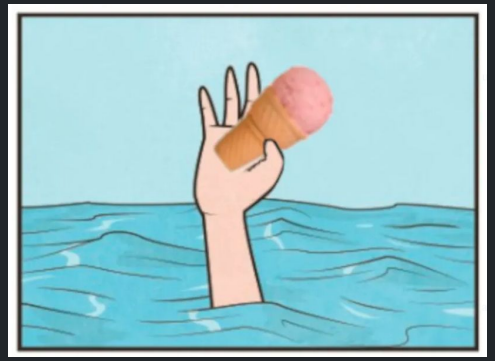
Five steps process

1. Determine whether you have experimental or observational data.
2. Understand your problem (context, variables, ...) as thoroughly as possible.
3. Create a model that describes how variables are related to each other.
4. Share your model, goal, and assumptions with the other people involved in the analysis, to get their input.
5. Apply causal inference techniques to your data and your model to answer your causal questions:
   a. State what objectives can be answered in your model.
   b. Estimate causal effects and confidence intervals using the appropriate formulas.




Observado vs Contrafactual
Observado (Madrid)
Contrafactual
220
200
180
160
140
120
€
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
Fecha
ATT medio post (€/hotel·día): 29.08302029633155

May the ice creams be with you



@Ubaldo Hervás