

An Introduction to Causality  
Mathematical Association of Cambodia

Irving Gómez Méndez

January, 2022



The examples throughout this talk were taken from the references presented at the end.

# Philosophy and Some History

## Behavioral Modernity and Causality

Behavioral modernity is a suite of behavioral and cognitive traits that distinguishes current homo sapiens from other anatomically modern humans, hominins, and primates. Most scholars agree that modern human behavior can be characterized by **abstract thinking, planning depth**, symbolic behavior (e.g., art, ornamentation), music and dance, exploitation of large game, and blade technology, among others.

Very early in our evolution, **we realized that the world is not** made up only of dry facts (what we might call **data** today); rather, these facts are glued together by **an intricate web of cause-effect relationships**. Causal explanations, not dry facts, make up the bulk of our knowledge.

## Attempting to Define Causality

In his *Treatise of Human Nature*, David Hume defines the cause-effect relationship entirely as a product of our own memory and experience:

*Thus we remember to have seen that species of object we call flame, and to have felt that species of sensation we call heat (...) Without any further ceremony, **we call the one cause and the other effect, and infer the existence of the one from the other.***

If we observe a cause — for example, Bobby throws a ball toward a window — most of us can predict the effect (the ball will probably break the window).

## Temporal Relationship is not Sufficient

During the 1960's an advisory committee was created in US to study the possibility that cigarette smoking was a causative agent in lung cancer. The committee listed (among others) the **temporal relationship (the effect should follow the cause)** as a sufficient criterion.

However, temporal relation has some exceptions — for example, a rooster crow does not cause the sun to rise, even though it always precedes the sun.

# Fixing the Definition of Causality

David Hume proposed the following definition of causation:

*We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, **if the first object had not been, the second never had existed.***

David Lewis pointed out that Hume really gave two definitions, not one.

- ▶ **Regularity:** The cause is regularly followed by the effect. However a rooster crow does not cause the sun to rise, even though it always precedes the sun.
- ▶ **Counterfactual:** (“if the first object had not been...”). Lewis argued that the counterfactual definition aligns more closely with human intuition. We have been making judgments like this since we were children:

*We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it.*



**Counterfactual** reasoning, which **deals with what-ifs**, might strike some readers as unscientific. Indeed, empirical observation can never confirm or refute the answers to such questions since, per definition, we cannot observe counterfactuals.

Yet our minds make very reliable and reproducible judgments all the time about what might be or might have been. We all understand, for instance, that had the rooster been silent this morning, the sun would have risen just as well.

Counterfactuals are the building blocks of moral behavior as well as scientific thought. The ability to reflect on one's past actions and envision alternative scenarios is the basis of free will and social responsibility.

## 426 BC Malian Gulf Tsunami



The 426 BC Malian Gulf tsunami **was caused** by one of a series of earthquakes which affected the course of the Peloponnesian War. It devastated the coasts of the Malian and Euboean Gulfs, Greece, in the summer of 426 BC.

**Thucydides** inquired into its causes, and **concluded that the tsunami must have been caused by an earthquake**. He was thus historically the first known to correctly interpret the cause of a tsunami as a preceding geological event. **Herodotus**, in contrast, had **attributed** the Potidaea tsunami **to the divine wrath of Poseidon**.

*About the same time that these earthquakes were so common, the sea at Orobiae, in Euboea, retiring from the then line of coast, returned in a huge wave and invaded a great part of the town, and retreated leaving some of it still under water; so that what was once land is now sea; such of the inhabitants perishing as could not run up to the higher ground in time.... **The cause, in my opinion, of this phenomenon must be sought in the earthquake.** At the point where its shock has been the most violent the sea is driven back, and suddenly recoiling with redoubled force, causes the inundation. **Without an earthquake I do not see how such an accident could happen.***

# The Three Levels of Causal Reasoning and Simple Examples

# The Ladder of Causation

We can distinguish three “levels” of causal inference.

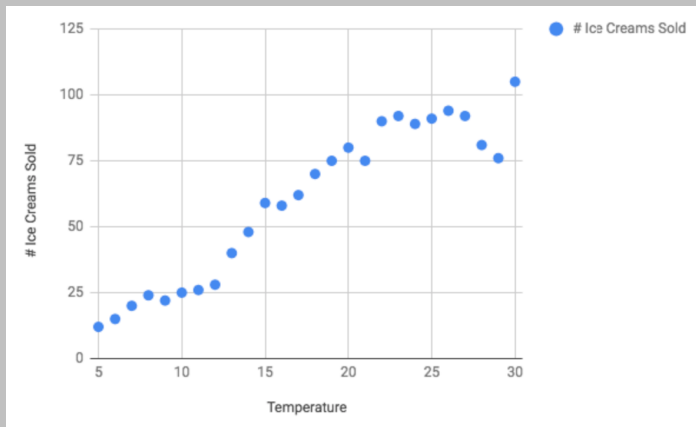
1. At the most basic level is **association**, which corresponds to the activity of **seeing**. Seeing or observing, entails detection of regularities in our environment and is shared by many animals as well as early humans before the Cognitive Revolution. At this level, we merely observe that a set of variables are statistically related.

- 
2. The second is the **intervention** level. The activity of **doing** corresponds to this level. **Doing, entails predicting the effect(s) of deliberate alterations** of the environment and choosing among these alterations to produce a desired outcome. Only a small handful of species have demonstrated elements of this skill.

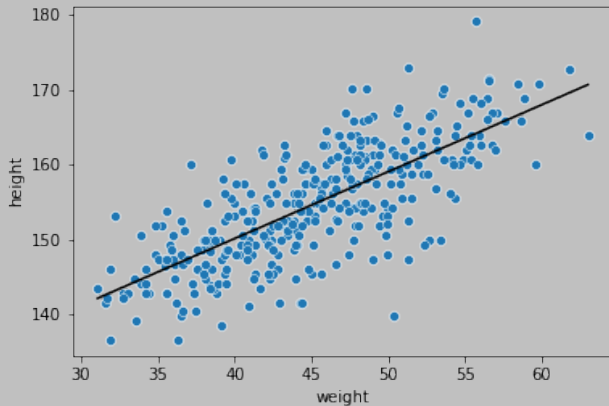
3. At the highest level are **counterfactuals**, which correspond to the activity of **imaginig**. Imagining and causal relations is almost self-evident. It is useless to ask for the causes of things unless you can imagine their consequences.



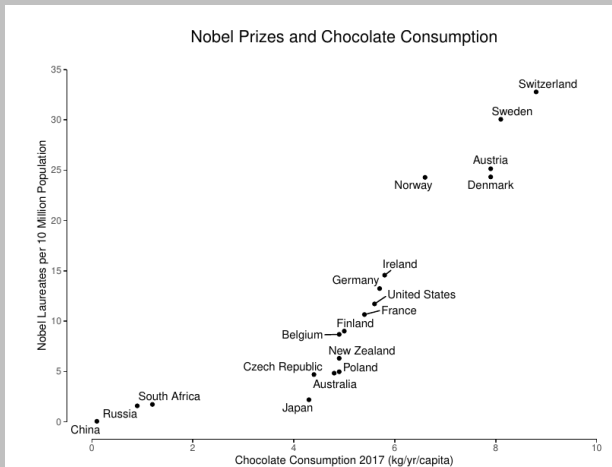
When it is hot you're gonna sell more icecream!



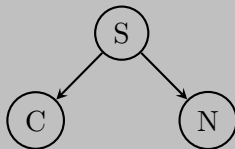
# You can predict height from weight



# Chocolate Produces Nobel Prize Winners! Or not?



While chocolate consumption could cause an increase in Nobel laureates, more plausibly, unobserved variables such as socio-economic status or quality of the education system might cause an increase in both chocolate consumption and Nobel laureates, thus rendering their correlation spurious, that is, non-causal.



S: Socio-economic status.

C: Chocolate consumption.

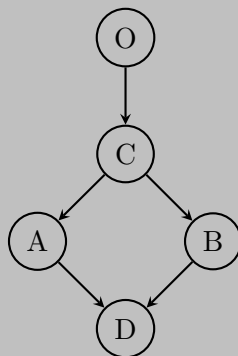
N: Nobel laureates.

Plenty of weird correlations can be found at  
<https://tylervigen.com/spurious-correlations>.

The rooster's crow is highly correlated with the sunrise; yet it does not cause the sunrise.

Moral: Data do not understand causes and effects; humans do.

## Firing Squad Example



O: Court Order, C: Captain, A, B: Soldiers, D: Prisoner status.

# Association

- ▶ If the prisoner is dead, does that mean the court order was given? Yes.
- ▶ Suppose we find out that A fired. What does that tell us about B? A would not have fired if the captain hadn't signaled, so B must have fired as well. Note that in this case A and B are perfectly correlated even though A does not cause B.

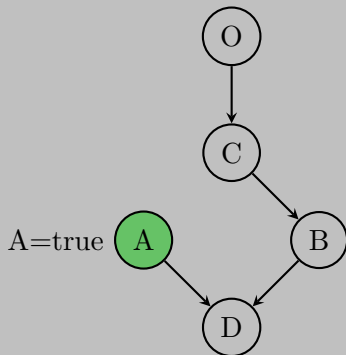
## Intervention

- ▶ What if soldier A decides on his own initiative to fire, without waiting for the captain's command? Will the prisoner be dead or alive?

If you're just using the rules of logic, as computers typically do, the question is meaningless. If we want our computer to understand causation, we have to teach it how to break the rules. We have to teach it the difference between merely observing an event and making it happen.



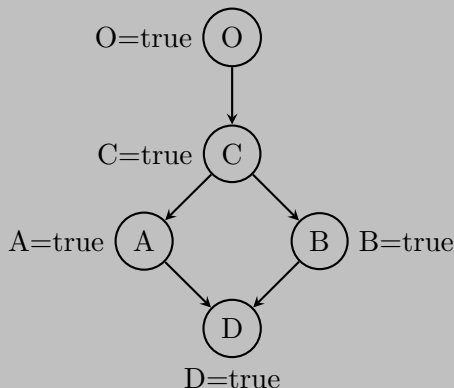
Making an event happen means that you emancipate it from all other influences. Thus we erase all the arrows leading into the intervened variable (A), set that variable manually to its prescribed value (true) and continue the analysis by ordinary logic. This is because after *doing* the intervention all that is left for us to do is to *see* its effect.



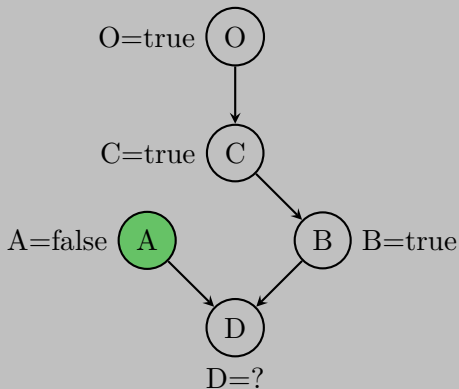
## Counterfactual

Suppose the prisoner is lying dead on the ground. From this we can conclude that A shot, B shot, the captain gave the signal, and the court gave the order. But

- ▶ what if A had decided not to shoot? Would the prisoner be alive?



This question requires us to compare the real world with a fictitious and contradictory world where A did not shoot. In the fictitious world, the arrow leading into A is erased. Instead A is set to false, leaving its past history the same as it was in the real world.



We conclude that the prisoner would be dead in the fictitious world as well, because B's shot would have killed him. So A's courageous change of heart would not have saved his life. Undoubtedly this is one reason firing squads exist: they guarantee that the court's order will be carried out and also lift some of the burden of responsibility from the individual shooters, who can say with a (somewhat) clean conscience that their actions did not cause the prisoner's death as "he would have died anyway."

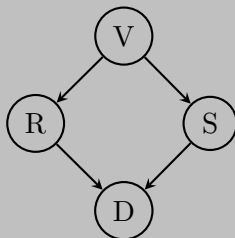
It may seem as if we are going to a lot of trouble to answer toy questions whose answer was obvious anyway. I completely agree! Causal reasoning is easy for you because you are human, and you were once a three-year-old, and you had a marvelous three-year-old brain that understood causation better than any animal or computer.

Assume that you are a reporter collecting records of execution scenes day after day. Your data might look something like:

O	C	A	B	D
true	true	true	true	true
true	true	true	true	true
false	false	false	false	false

There is no way that this kind of data, in the absence of an understanding the causal relation, will enable you to predict the results of persuading marksman A not to shoot, no matter how much data you collect.

## Vaccines (do not) Kill



V: Vaccine, R: Reaction, S: Disease, D: Death.

	Vaccinated	Non-Vaccinated	Total
	990,000 (99%)	10,000 (1%)	1'000,000
Reaction	9,900 (1%)	0	9,900
Disease	0	200 (2%)	200
Dead	99 (1%)	40 (20%)	139

In summary, more people died from vaccination (99) than from the disease (40).

## Do not Vaccinate Kills

We now ask the counterfactual question: “What if we had set the vaccination rate to zero?”

	Vaccinated	Non-Vaccinated	Total
	0 (0%)	1'000,000 (100%)	1'000,000
Reaction	0 (1%)	0	0
Disease	0	20,000 (2%)	20,000
Dead	0 (1%)	4,000 (20%)	4,000

Comparing the counterfactual world with the real world, we see that not vaccinating would have cost the lives of 3,861 people.



# Probability and Directed Acyclic Graphs

## Relation between Probability and Causality

The idea of causes and effects is much more fundamental than the idea of probability. We begin learning causes and effects before we understand language and before we know any mathematics.

Understanding the meaning of “cause” has been the focus of a long tradition of philosophers, they have tried to define causation in terms of probability, using the notion of “probability raising”:  $X$  causes  $Y$  if  $X$  raises the probability of  $Y$ .

## Defining Causation as Raising a Probability

We say, for example, “Reckless driving causes accidents” or “You will fail this course because of your laziness,” knowing quite well that the antecedents merely tend to make the consequences more likely, not absolutely certain.

But this increase may come about for other reasons, including  $Y$  being a cause of  $X$  or some other variable ( $Z$ ) being the cause of both of them (eating more chocolate will not make you win the Nobel). That’s the catch!

## Conciliating Causation with Probability

The proper way to rescue the probability-raising idea is with the do-operator. While  $\mathbb{P}(Y|X = x)$  denotes the observational distribution, which corresponds of seeing,  $\mathbb{P}(Y|do(X = x))$  corresponds to the interventional distribution, which corresponds to the process of doing. Thus  $\mathbb{P}(Y|do(X = x))$  describes the causal effect of  $X$  on  $Y$ , which can be calculated using causal directed acyclic graphs (DAGs).

## Chain / Mediator

There are three basic types of junctions, with the help of which we can characterize any pattern of arrows in the network.

- ▶  $A \rightarrow B \rightarrow C$ . This junction is the simplest example of a “chain,” or of mediation, in which  $B$  is thought as the mechanism or “mediator” that transmits the effect of  $A$  to  $C$ . A familiar example is  $\text{Fire} \rightarrow \text{Smoke} \rightarrow \text{Alarm}$ . Likewise, we say that Fire and Alarm are conditionally independent, given the value of Smoke ( $A \perp\!\!\!\perp C | B$ ).

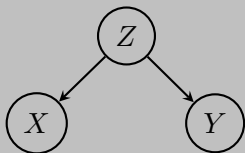
## Fork / Confounder

- ▶  $A \leftarrow B \rightarrow C$ . This kind of junction is called a “fork,” and  $B$  is often called a common cause or confounder of  $A$  and  $C$ .

A good example (due to David Freedman) is  $\text{Shoe Size} \leftarrow \text{Age of Child} \rightarrow \text{Reading Ability}$ . Children with larger shoes tend to read at a higher level. But the relationship is not one of cause and effect. Giving a child larger shoes won't make him read better! Instead, both variables are explained by a third, which is the child's age. Older children have larger shoes, and they also are more advanced readers.

We can eliminate this spurious correlation by conditioning on the child's age. For instance, if we look only at seven-year-olds, we expect to see no relationship between shoe size and reading ability. As in the case of chain junctions,  $A$  and  $C$  are conditionally independent, given  $B$  ( $A \perp\!\!\!\perp C|B$ ).

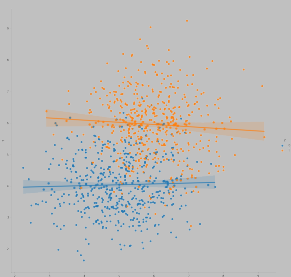
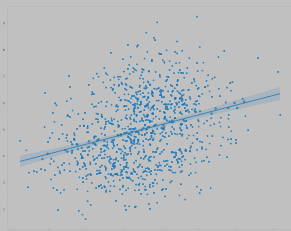
## Another Example of Confounder



$$Z \sim \text{Ber}(p).$$

$$X \sim \begin{cases} \mathcal{N}(5, 1) & \text{if } Z = 0, \\ \mathcal{N}(6, 1) & \text{if } Z = 1. \end{cases}$$

$$Y \sim \begin{cases} \mathcal{N}(4, 1) & \text{if } Z = 0, \\ \mathcal{N}(6, 1) & \text{if } Z = 1. \end{cases}$$





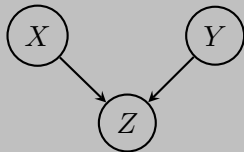
## Collider

- ▶  $A \rightarrow B \leftarrow C$ . This junction is called a “collider.” Felix Elwert and Chris Winship have illustrated this junction using three features of Hollywood actors:  $Talent \rightarrow Celebrity \leftarrow Beauty$ . Here we are asserting that both talent and beauty contribute to an actor’s success, but beauty and talent are completely unrelated to one another in the general population. If  $A$  and  $C$  are independent to begin with, conditioning on  $B$  will make them dependent ( $A \not\perp C | B$ ).

For example, if we look only at famous actors (in other words, we observe the variable  $\text{Celebrity} = 1$ ), we will see a negative correlation between talent and beauty: finding out that a celebrity is unattractive increases our belief that he or she is talented.

This correlation is sometimes called collider bias or the “explain-away” effect. For simplicity, suppose that you don’t need both talent and beauty to be a celebrity; one is sufficient. Then if Celebrity A is a particularly good actor, that “explains away” his success, and he doesn’t need to be any more beautiful than the average person. On the other hand, if Celebrity B is a really bad actor, then the only way to explain his success is his good looks. So, given the outcome  $\text{Celebrity} = 1$ , talent and beauty are inversely related—even though they are not related in the population as a whole.

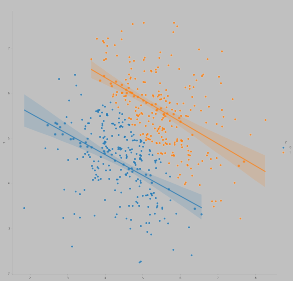
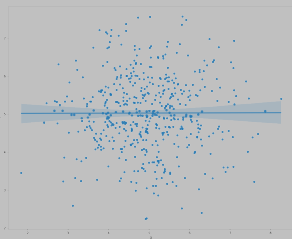
## Another Example of Collider



$$X \sim \mathcal{N}(5, 1).$$

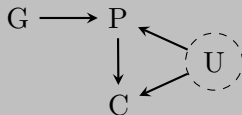
$$Y \sim \mathcal{N}(5, 1).$$

$$Z = \begin{cases} 1 & \text{if } 5X + 5Y > 50, \\ 0 & \text{otherwise.} \end{cases}$$



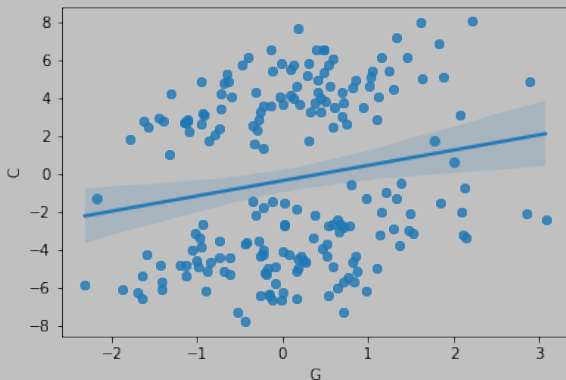
## The Haunted DAG

Unfortunately some biases might arrived due to unobserved variables. The next DAG represents the effect of the education of grandparents  $G$  and parents  $P$  into children  $C$ . Moreover, we assume that there are some unobserved variables  $U$  that influence both  $P$  and  $C$ .

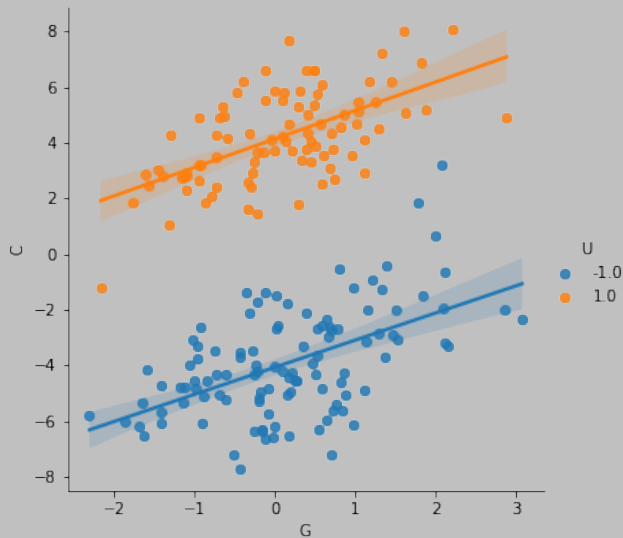


So now  $P$  is a common consequence of  $G$  and  $U$ , so conditioning on  $P$  would create a collider bias.

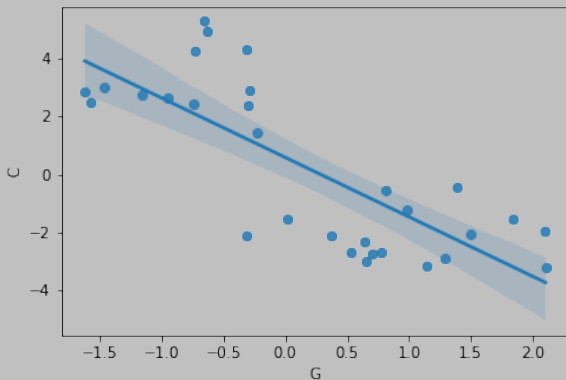
The next figure shows the result when we regressed  $C$  in function of  $G$  without conditioning on  $P$  neither  $U$ . where we can see the total effect of  $G$  into  $C$ .



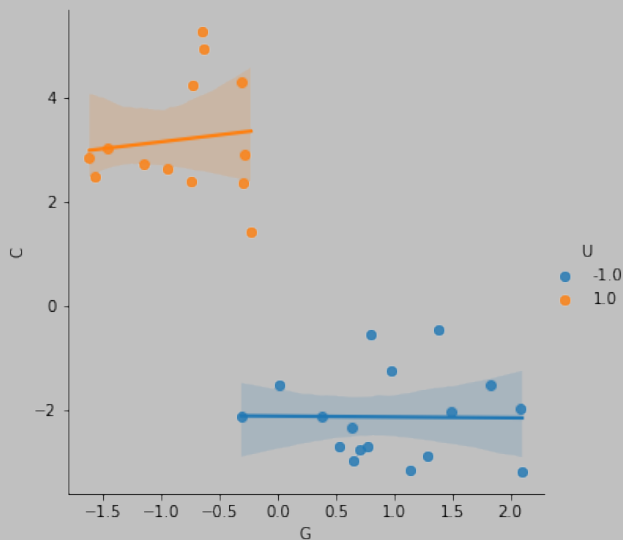
If we condition on  $U$  we still get the total effect of  $G$  into  $C$ .



But, when we controlled by  $P$  the problems arrive! In the next image we select the values of  $P$  between the 0.45 and 0.6 quantiles. We now see that **the apparent effect is now negative!** This phenomenon is known as the Simpson's paradox



If we condition on both  $P$  and  $U$  we can get the direct effect of  $G$  into  $C$ .





## Simpson's Paradox and *do*-Calculus

## Simpson's Paradox

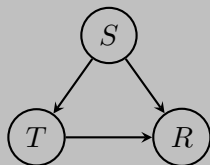
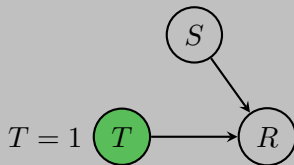
Suppose that you observe 700 patients who either *choose* to take treatment or not

	Treatment
Men	81 out of 87 recovered (93%)
Women	192 out of 263 recovered (73%)
Men & Women	273 out of 350 recovered (78%)

	No Treatment
Men	234 out of 270 recovered (87%)
Women	55 out of 80 recovered (69%)
Men & Women	289 out of 350 recovered (83%)

Should a doctor prescribe the treatment or not?

To answer the question we need to compute the causal effect that treatment has on recovery.

DAG:  $\mathcal{G}$ DAG:  $\mathcal{G}'$ 

$$\begin{aligned}
 \mathbb{P}(R = 1 | do(T = 1)) &= \mathbb{P}^{\mathcal{G}'}(R = 1 | T = 1) \\
 &= \mathbb{P}^{\mathcal{G}'}(R = 1 | S = 1, T = 1) \mathbb{P}^{\mathcal{G}'}(S = 1 | T = 1) \\
 &\quad + \mathbb{P}^{\mathcal{G}'}(R = 1 | S = 0, T = 1) \mathbb{P}^{\mathcal{G}'}(S = 0 | T = 1)
 \end{aligned}$$

## Back-Door Criterion

$$\begin{aligned}
 \mathbb{P}(R = 1|do(T = 1)) &= \mathbb{P}(R = 1|S = 1, T = 1)\mathbb{P}(S = 1) \\
 &\quad + \mathbb{P}(R = 1|S = 0, T = 1)\mathbb{P}(S = 0) \\
 &= \frac{81}{87} \left( \frac{87 + 270}{700} \right) + \frac{192}{263} \left( \frac{263 + 80}{700} \right) \\
 &\approx 0.833
 \end{aligned}$$

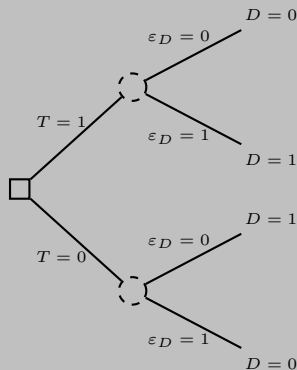
$$\begin{aligned}
 \mathbb{P}(R = 1|do(T = 0)) &= \mathbb{P}(R = 1|S = 1, T = 0)\mathbb{P}(S = 1) \\
 &\quad + \mathbb{P}(R = 1|S = 0, T = 0)\mathbb{P}(S = 0) \\
 &= \frac{234}{270} \left( \frac{87 + 270}{700} \right) + \frac{55}{80} \left( \frac{263 + 80}{700} \right) \\
 &\approx 0.779
 \end{aligned}$$

On average 5.4% more patients would recover if they were given the treatment.

# Potential Outcomes and Structural Causal Models

## Structural Causal Models (SCM)

Suppose that there is a treatment for a deadly disease. For 99% of the population, the treatment ( $T = 1$ ) works and they get cured ( $D = 0$ ). The remaining 1% has a rare condition ( $\varepsilon_D = 1$ ) which makes them immune to the disease, but the treatment becomes fatal.



In this case we have the causal graph:  $T \rightarrow D \leftarrow \varepsilon_D$  and the *structural causal equation (SCE)*:

$$D := T\varepsilon_D + (1 - T)(1 - \varepsilon_D),$$

both determines the SCM  $\mathfrak{C}$ .

A patient comes to the hospital and dies ( $D = 1$ ) after the medic administers the treatment. What would have happened had the doctor not administer the treatment?

To answer this question, we first use the the observed information and the SCE to get the value of  $\varepsilon_D$ :

$$(\varepsilon_D|T = 1, D = 1) = 1.$$

In terms of the *do* operator the counterfactual question implies to know the value of  $D$  given  $do(T = 0)$  and the observed data  $T = 1, D = 1$ . That is

$$(D|T = 1, D = 1; do(T = 0)) = 0(1) + 1(1 - 0)(1 - 1) = 0.$$

So, we can conclude that the person would be alived if he/she would not have received the treatment.



However, note that

$$\mathbb{P}^{\mathcal{C}}(D = 0 | do(T = 1)) = \mathbb{P}^{\mathcal{C}}(\varepsilon_D = 0) = 0.99,$$

and

$$\mathbb{P}^{\mathcal{C}}(D = 0 | do(T = 0)) = \mathbb{P}^{\mathcal{C}}(\varepsilon_D = 1) = 0.01.$$

Therefore, if the medic didn't know the value of  $\varepsilon_D$  we can not say that he/she was negligent.

## Potential Outcomes

Suppose that we have collected some data on the existing salaries, letting  $X$  represent years of experience,  $D$  represent education, and  $S$  represent salary. We're also assuming, for simplicity, just three levels of education: 0 = high school degree, 1 = college degree, 2 = graduate degree. Thus  $S_0(u)$ , represents the salary of individual  $u$  if  $u$  were a high school graduate but not a college graduate, and  $S_1(u)$  represents  $u$ 's salary if  $u$  were a college graduate.

$u$	$X$	$D$	$S_0(u)$	$S_1(u)$	$S_2(u)$
Alice	6	0	81,000	?	?
Bert	9	1	?	92,500	?
Caroline	9	2	?	?	97,000
David	8	1	?	91,000	?
Ernest	12	1	?	100,000	?
Frances	13	0	97,500	?	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

A typical counterfactual question we might want to ask is “What would Alice’s salary be if she had a college degree?” In other words, what is  $S_1(\text{Alice})$ ?

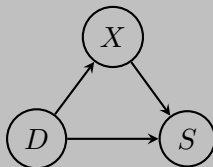
We decide to model  $S = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon$  and estimate  $\beta_0, \beta_1$  and  $\beta_2$ . Thus, the model might look like this:

$$\mathbb{E}[S|X, D] = 65,000 + 2,500X + 5,000D.$$

Accordingly, a regression analyst would claim, our estimate of Alice's salary, if she had a college degree, is

$$65,000 + 2,500 \times 6 + 5,000 \times 1 = 85,000.$$

However, if Alice had college degree, she would have not been able to use that time to gain experience compared to what she now has.



We can use the same statistical methods as before to find the best-fitting linear equation. The result would look just like before, with one small difference:

$$S = 65,000 + 2,500X + 5,000D + U_S.$$

We must also have a structural equation for  $X$  that might look like this:

$$X = 10 - 4D + U_X.$$

Alice's data:  $S_0(\text{Alice}) = 81,000$ ,  $E(\text{Alice}) = 0$ ,  $X(\text{Alice}) = 6$ .

Then,

$$U_X(\text{Alice}) = 6 - 10 = -4$$

and

$$U_S(\text{Alice}) = 81,000 - 65,000 - 2,500 \times 6 = 1,000$$

We use now the do-operator, so we erase the arrows pointing to the variable that is being set to a counterfactual value (Education) and set Alice's Education to a college degree ( $D = 1$ ). In this example, this step is trivial, because there are no arrows pointing to Education and hence no arrows to erase.

$$X_{D=1}(Alice) = 10 - 4 - 4 = 2$$

and the potential salary of Alice would be

$$\begin{aligned} S_{D=1}(Alice) &= 65,000 + 2,500 \times 2 + 5,000 \times 1 + 1,000 \\ &= 76,000 \end{aligned}$$

## Other Ways to Learn Causation



## There are other Ways to Learn Causation

The gold standard for modeling natural phenomena is a set of coupled differential equations modeling physical mechanisms responsible for the time evolution. This allows us to predict the future behavior of a physical system, reason about the effect of interventions, and predict statistical dependencies between variables that are generated by coupled time evolution.

Consider the coupled set of differential equations






$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^p,$$

with initial value  $\mathbf{x}(t_0) = \mathbf{x}_0$  .

If we formally write this in terms of infinitesimal differentials  $dt$  and  $d\mathbf{x} = \mathbf{x}(t + dt) - \mathbf{x}(t)$ , we get:  $\mathbf{x}(t + dt) = \mathbf{x}(t) + dt \cdot f(\mathbf{x}(t))$ .

While a differential equation is a rather comprehensive description of a system, a statistical model can be viewed as a much more superficial one. It often does not refer to dynamic processes; instead, it tells us how some of the variables allow prediction of others as long as experimental conditions do not change however, its strength is that it can often be learned from observational data, while a differential equation usually requires an intelligent human to come up with it.

**Causal modeling** lies in between these two extremes. Like models in physics, it aims to provide understanding and predict the effect of interventions. However, causal discovery and learning try to arrive at such models in a **data-driven way, replacing expert knowledge with weak and generic assumptions.**

-  Dablander, Fabian (2020). “An introduction to causal inference”. In.
-  McElreath, Richard (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.
-  Pearl, Judea and Dana Mackenzie (2018). *The book of why: the new science of cause and effect*. Basic books.
-  Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.
-  Schölkopf, Bernhard et al. (2021). “Toward causal representation learning”. In: *Proceedings of the IEEE* 109.5, pp. 612–634.

Thank you!

gomendez.irving@gmail.mx

<https://irvinggomez.com>