Introduction to Causality

Aristotle's Four Causes

Aristotle's Four Causes describe why a thing is the way it is.

1. Material Cause:

- Definition: Refers to what something is made of.

- Example: For a statue, the material cause would be the marble or bronze from which it is carved.

2. Formal Cause:

- Definition: Refers to the design, pattern, or archetype of something. It's the "blueprint" or form that entities embody.

- Example: For the statue, the formal cause would be the shape or design that the sculptor envisions and subsequently gives to the material.

3. Efficient Cause:

- Definition: Refers to the process or agent that brings something into being, i.e., the cause of the change.

- Example: For the statue, the efficient cause would be the sculptor, who carves and shapes the marble or bronze into the intended form.

4. Final Cause:

- Definition: Refers to the purpose or function of something – why it exists or why it was made.

- Example: For the statue, the final cause might be to honor a deity, commemorate an event, or beautify a space.

Together, these causes provide a comprehensive explanation for why a thing is as it is, from its basic material to its ultimate purpose.

Hume's Causality Theory

David Hume, a famous 18th-century Scottish philosopher, proposed a more unified framework for cause-effect relationships. Hume starts with an observation that we never observe cause-effect relationships in the world. The only thing we experience is that some events are conjoined.

1. We only observe how the movement or appearance of object A precedes the movement or appearance of object B.

2. If we experience such a succession a sufficient number of times, we'll develop a feeling of expectation.

3. This feeling of expectation is the essence of our concept of causality (it's not about the world; it's about a feeling we develop).

Machine Learning Works on Associations

Hume's Causality theory has a high resemblance to a very powerful idea in psychology called ***conditioning***. Conditioning is a form of learning. There are multiple types of conditioning, but they all rely on a common foundation – namely, ***association*** (hence the name for this type of learning – associative learning). In any type of conditioning, we take some event or object (usually called stimulus) and associate it with some behavior or reaction.

Most classic machine learning algorithms also work on the basis of association. When we're training a neural network in a supervised fashion, we're trying to find a function that maps input to the output. To do it efficiently, we need to figure out which elements of the input are useful for predicting the output. And, in most cases, association is just good enough for this purpose.

How Do Babies Interact with the World?

Have you ever seen a parent trying to convince their child to stop throwing around a toy? Some parents tend to interpret this type of behavior as rude, destructive, or aggressive, but babies often have a different set of motivations. They are running systematic experiments that allow them to learn the laws of physics and the rules of social interactions.

Infants as young as 11 months prefer to perform experiments with objects that display unpredictable properties (for example, can pass through a wall) than with objects that behave predictably (Stahl & Feigenson, 2015). This preference allows them to efficiently build models of the world.

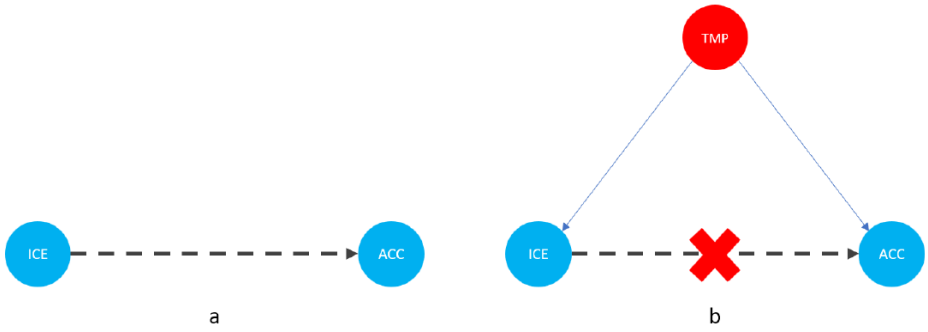
Confounding – Relationships That Are Not Real

A confounding variable influences two or more other variables and produces a spurious association between them. From a purely statistical point of view, such associations are indistinguishable from the ones produced by a causal mechanism.

Imagine you work at a research institute and you're trying to understand the causes of people drowning. Your organization provides you with a huge database of socioeconomic variables. You decide to run a regression model over a large set of these variables to predict the number of drownings per day in your area of interest. When you check the results, it turns out that the biggest coefficient you obtained is for daily regional ice cream sales. Interesting! Ice cream usually contains large amounts of sugar, so maybe sugar affects people's attention or physical performance while they are in the water.

This hypothesis might make sense, but before we move forward, let's ask some questions. How about other variables that we did not include in the model? Did we add enough predictors to the model to describe all relevant aspects of the problem? What if we added too many of them? Could adding just one variable to the model completely change the outcome?

Let me introduce you to daily average temperature – our confounder. Higher daily temperature makes people more likely to buy ice cream and more likely to go swimming. When there are more people swimming, there are also more accidents. Let's try to visualize this relationship:



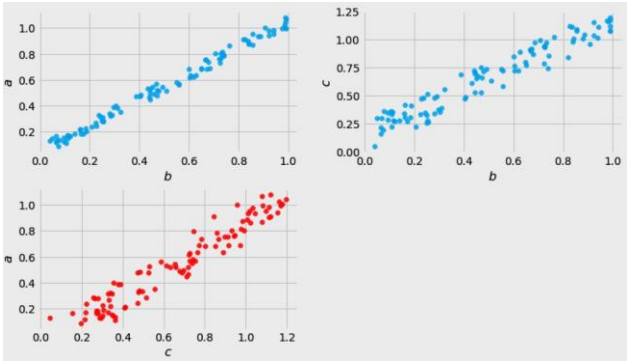
Confounding Is a Strictly Causal Concept

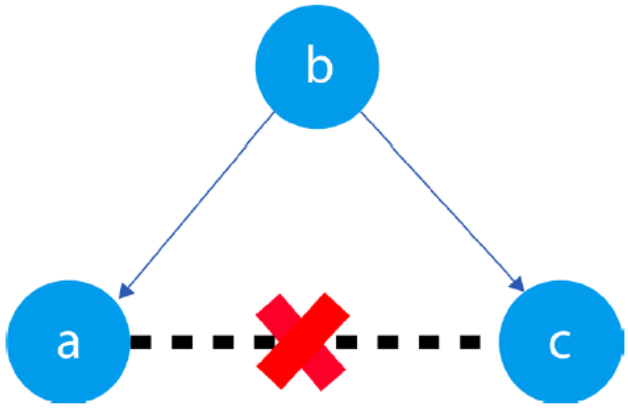
What does it mean? It means that we're not able to say much about confounding using purely statistical language (note that this means that Hume's definition as we presented it here cannot capture it).

In Figure 1.2, blue points signify a causal relationship while red points signify a spurious relationship, and variables a, b, and c are related in the following way:

• b causes a and c

• a and c are causally independent





A Marketer's Dilemma

Imagine you are a tech-savvy marketer and you want to effectively allocate your direct marketing budget. How would you approach this task? When allocating the budget for a direct marketing campaign, we'd like to understand what return we can expect if we spend a certain amount of money on a given person. In other words, we're interested in estimating the effect of our actions on some customer outcomes (Gutierrez, Gérardy, 2017). Perhaps we could use supervised learning to solve this problem? To answer this question, let's take a closer look at what exactly we want to predict.

We're interested in understanding how a given person would react to our content. Let's encode it in a formula:

$$\tau_i = Y_i(1) - Y_i(0)$$

In the preceding formula, the following applies:

• τ i is the treatment effect for person i

• Y i (1) is the outcome for person i when they received the treatment T (in our example, they received marketing content from us)

• Y i (0) is the outcome for the same person i given they did not receive the treatment T

What the formula says is that we want to take the person i 's outcome Y i when this person does not receive treatment T and subtract it from the same person's outcome when they receive treatment T.

An interesting thing here is that to solve this equation, we need to know what person i 's response is under treatment and under no treatment. In reality, we can never observe the same person under two mutually exclusive conditions at the same time. To solve the equation in the preceding formula, we need counterfactuals.

Counterfactuals are estimates of how the world would look if we changed the value of one or more variables, holding everything else constant. Because counterfactuals cannot be observed, the true causal effect τ is unknown. This is one of the reasons why classic machine learning cannot solve this problem for us.

Let's Play Doctor

Let's take another example. Imagine you're a doctor. One of your patients, Jennifer, has a rare disease, D.

Additionally, she was diagnosed with a high risk of developing a blood clot. You study the information on the two most popular drugs for D. Both drugs have virtually identical effectiveness on D, but you're not sure which drug will be safer for Jennifer, given her diagnosis. You look into the research data presented in Table 1.1.

Drug	**A**		**B**	
Blood clot	Yes	No	Yes	No
Total	27	95	23	99
Percentage	22%	78%	19%	81%

The numbers in Table 1.1 represent the number of patients diagnosed with disease D who were administered drug A or drug B. Row 2 (Blood clot) gives us information on whether a blood clot was found in patients or not. Note that the percentage scores are rounded. Based on this data, which drug would you choose? The answer seems pretty obvious. 81% of patients who received drug B did not develop blood clots. The same was true for only 78% of patients who received drug A. The risk of developing a blood clot is around 3% lower for patients receiving drug B compared to patients receiving drug A.

This looks like a fair answer, but you feel skeptical. You know that blood clots can be very risky and you want to dig deeper. You find more fine-grained data that takes the patient's gender into account. Let's look at Table 1.2.

Drug	**A**		**B**	
Blood clot	Yes	No	Yes	No
Female	24	56	17	25
Male	3	39	6	74
Total	27	95	23	99
Percentage	22%	78%	18%	82%
Percentage (F)	30%	70%	40%	60%
Percentage (M)	7%	93%	7.5%	92.5%

Something strange has happened here. We have the same numbers as before and drug B is still preferable for all patients, but it seems that drug A works better for females and for males!

What we've just experienced is called **Simpson's paradox** (also known as the Yule-Simpson effect). Simpson's paradox appears when data partitioning (which we can achieve by controlling for the additional variable(s) in the regression setting) significantly changes the outcome of the analysis. In the real world, there are usually many ways to partition your data. You might ask: okay, so how do I know which partitioning is the correct one?

We could try to answer this question from a pure machine learning point of view: perform cross-validated feature selection and pick the variables that contribute significantly to the outcome. This solution is good enough in some settings. For instance, it will work well when we only care about making predictions (rather than decisions) and we know that our production data will be independent and identically distributed; in other words, our production data needs to have a distribution that is virtually identical (or at least similar enough) to our training and validation data. If we want more than this, we'll need some sort of a (causal) world model.