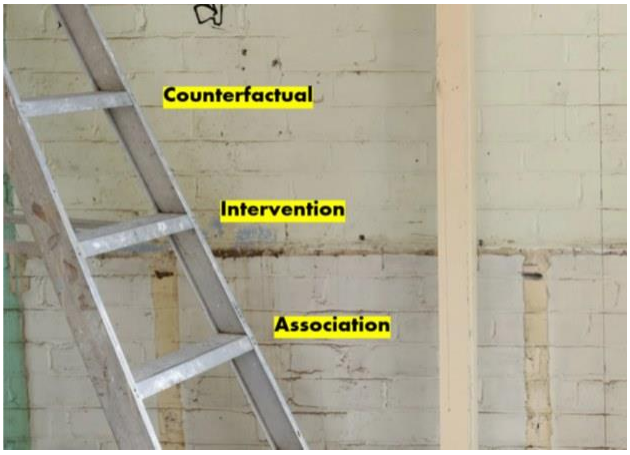
# Ladder of Causation

## The Ladder of Causation

The Ladder of Causation, introduced by Judea Pearl (Pearl, Mackenzie, 2019), is a helpful metaphor for understanding distinct levels of relationships between variables – from simple associations to counterfactual reasoning.

Pearl's ladder has three rungs. Each rung is related to different activity and offers answers to different types of causal questions. Each rung comes with a distinct set of mathematical tools.




Rung one of the ladder represents association. The activity that is related to this level is observing. Using association, we can answer questions about how seeing one thing changes our beliefs about another thing – for instance, how observing a successful space launch by SpaceX changes our belief that SpaceX stock price will go up.

Rung two represents intervention. Remember the babies from the previous chapter? The action related to rung two is doing or intervening. Just like babies throwing their toys around to learn about the laws of physics, we can intervene on one variable to check how it influences some other variable. Interventions can help us answer questions about what will happen to one thing if we change another thing – for instance, if I go to bed earlier, will I have more energy the following morning?

Rung three represents counterfactual reasoning. Activities associated with rung three are imagining and understanding. Counterfactuals are useful to answer questions about what would have happened if we had done something differently. For instance, would I have made it to the office on time if I took the train rather than the car?

| Rung | Action | Question |
|---|---|---|
| Association (1) | *Observing* | How does observing X change my belief in Y? |
| Intervention (2) | *Doing* | What will happen to Y if I do X? |
| Counterfactual (3) | *Imagining* | If I had done X, what would Y be? |

## An Example About Drug Effectiveness

Imagine that you're a doctor and you consider prescribing drug D to one of your patients. First, you might recall hearing other doctors saying that D helped their patients. It seems that in the sample of doctors you heard talking about D, there is an association between their patients taking the drug and getting better. That's rung one. We are skeptical about the rung one evidence because it might just be the case that these doctors only treated patients with certain characteristics (maybe just mild cases or only patients of a certain age). To overcome the limitation of rung one, you decide to read articles based on randomized clinical trials.

These trials were based on interventions (rung two) and – assuming that they were properly designed – they can be used to determine the relative efficacy of the treatment. Unfortunately, they cannot tell us whether a patient would be better off if they had taken the treatment earlier, or which of two available treatments with similar relative efficacy would have worked better for this particular patient. To answer this type of question, we need rung three.

## Conditional Probability

Conditional probability is the probability of one event, given that another event has occurred.

A mathematical symbol that we use to express conditional probability is | (known as a pipe or vertical bar). We read P ( X | Y ) as a probability of X given Y.

This notation is a bit simplified (or abused if you will). What we usually mean by P ( X | Y ) is P ( X = x | Y = y ), the probability that the variable X takes the value x , given that the variable Y takes the value y.

This notation can also be extended to continuous cases, where we want to work with probability densities – for example, P ( 0 < X < 0.25 | Y > 0.5 ).

## Associations

We'll demonstrate how to quantify associational relationships using conditional probability.

Imagine that you run an internet bookstore. What is the probability that a person will buy book A, given that they bought book B? This question can be answered using the following conditional probability query:

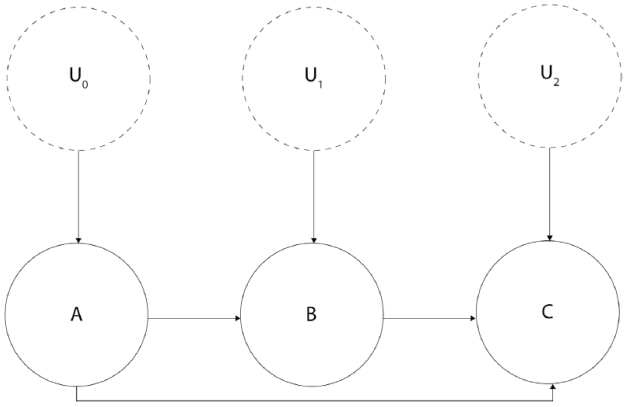
P (book A| book B )

Note that the preceding formula does not give us any information on the causal relationship between both events. We don't know whether buying book A caused the customer to buy book B, buying book B caused them to buy book A, or there is another (unobserved) event that caused both. We only get information about non-causal association between these events.

## Structural Causal Models (SCMs)

Structural causal models (SCMs) are a simple yet powerful tool to encode causal relationships between variables.

You might be surprised that we are discussing a causal model in the section on association. Didn't we just say that association is usually not enough to address causal questions? That's true. The reason why we're introducing an SCM now is that we will use it as our data-generating process. After generating the data, we will pretend to forget what the SCM was. This way, we'll mimic a frequent real-world scenario where the true data-generating process is unknown, and the only thing we have is observational data.




Circles or nodes in the preceding figure represent variables. Lines with arrows or edges represent relationships between variables.

As you can see, there are two types of variables (marked with dashed versus regular lines). Arrows at the end of the lines represent the direction of the relationship.

Nodes A, B, and C are marked with solid lines. They represent the observed variables in our model. We call this type of variable **endogenous**. Endogenous variables are always children of at least one other variable in a model.

The other type of nodes (UX nodes) are marked with dashed lines. We call these variables exogenous, and they are represented by root nodes in the graph (they are not descendants of any other variable. **Exogenous** variables are also called noise variables.

Note that most causal inference and causal discovery methods require that noise variables are uncorrelated with each other (otherwise, they become unobserved confounders). This is one of the major difficulties in real-world causal inference, as sometimes, it's very hard to be sure that we have met this assumption.

## Functional Representation

Let's return to the SCM from Figure 2.2. We'll define the functional relationships in this model in the following way:

A ≔ f A ( U 0 )

B ≔ f B (A, U 1 )

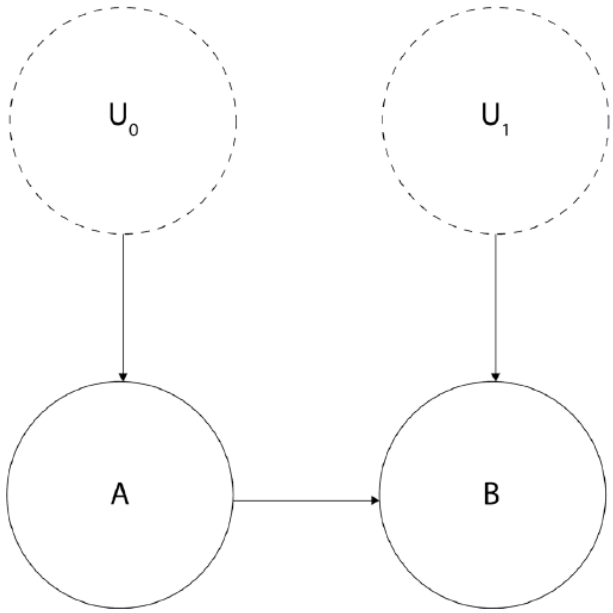
C ≔ f C ( A, B, U 2 )

A , B , C, and U X represent the nodes in Figure 2.1, and ≔ is an assignment operator, also known as a walrus operator. We use it here to emphasize that the relationship that we're describing is directional (or asymmetric), as opposed to the regular equal sign that suggests a symmetric relation.

Finally, f A , f B , f C represent arbitrary functions (they can be as simple as a summation or as complex as you want).

## Coding Exercise

First, let's define an SCM that can generate data with a non-zero probability of buying book A, given we bought book B. There are many possible SCMs that could generate such data. Figure 2.3 presents the model we have chosen for this section:




To precisely define causal relations that drive our SCM, let's write a set of equations:

U 0 ~ U ( 0, 1 )

U 1 ~ N (0, 1)

A : = 1 { U 0 >.61}

B ≔ 1 { (A+ .5* U 1 ) > .2}

In the preceding formulas, U 0 is a continuous random variable uniformly distributed between 0 and 1. U 1 is a normally distributed random variable, with a mean value of 0 and a standard deviation of 1. A and B are binary variables, and 1 {f} is an indicator function.

The notation for the indicator function might look complicated, but the idea behind it is very simple. The indicator function returns 1 when the condition in the curly braces is met and returns 0 otherwise. For instance, let's take the following function:

X = 1 {Z>0}

If Z > 0 then X = 1 , otherwise X = 0

[See the coding lab in Google Golab]

## What Are Interventions?

The idea of intervention is very simple. We change one thing in the world and observe whether and how this change affects another thing in the world. This is the essence of scientific experiments.

To describe interventions mathematically, we use a special do -operator. We usually express it in mathematical notation in the following way:

$P(Y = 1 \mid do(X = 0))$

The preceding formula states that the probability of $Y = 1$ , given that we set X to 0.

The fact that we need to change X 's value is critical here, and it highlights the inherent difference between ***intervening*** and ***conditioning*** (conditioning is the operation that we used to obtain conditional probabilities in the previous section).

Conditioning only modifies our ***view*** of the data, while intervening affects the distribution by actively setting one (or more) variable(s) to a fixed value (or a distribution). This is very important – intervention changes the system, but conditioning does not.

## The Graph Saga – Parents, Children, and More

When we talk about graphs, we often use terms such as parents, children, descendants, and ancestors.

- We say that the node X is a parent of the node Y and that Y is a child of X when there's a direct arrow from X to Y.
- If there's also an arrow from Y to Z, we say that Z is a grandchild of X and that X is a grandparent of Z.
- Every child of X, all its children and their children, their children's children, and so on are descendants of X, which is their ancestor.

## Change The World

When we intervene in a system and fix a value or alter the distribution of some variable – let's call it X – one of three things can happen:

- The change in X will influence the values of its descendants (assuming X has descendants and excluding special cases where X 's influence is canceled – for example, $f(x) = x - x$ )

- X will become independent of its ancestors (assuming that X has ancestors)
- Both situations will take place (assuming that X has descendants and ascendants, excluding special cases)

Note that none of these would happen if we conditioned on X , because conditioning does not change the value of any of the variables – it does not change the system.

Let's translate interventions into code. We will use the following SCM for this purpose:

U 0 ~ N (0, 1)

U 1 ~ N (0, 1)

A ≔ U 0

B ≔ 5A + U 1

The graphical representation of this model is identical to the one in Figure 2.3. Its functional assignments are different though, and – importantly – we set A and B to be continuous variables (as opposed to the model in the previous section, where A and B were binary; note that this is a new example, and the only thing it shares with the bookstore example is the structure of the graph).

[See the coding lab in Google Golab]

## Correlation and Causation

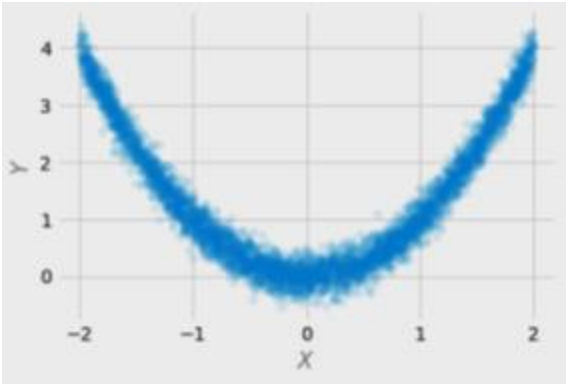
You might have heard the phrase that correlation is not causation. That's approximately true.

How about the opposite statement? Is causation correlation?

Figure 2.4 presents the data generated according to the following set of structural equations:

X ≔ $\mathcal{U}$ (− 2, 2)

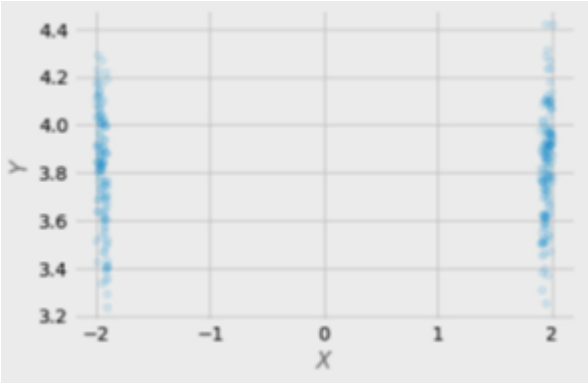
Y : = X 2 + 0.2 × $\mathcal{N}$ ( 0, 1 )




Although from the structural point of view, there's a clear causal link between X and Y, the correlation coefficient for this dataset is essentially equal to 0. The reason for this is that the relationship between X and Y is not monotonic, and popular correlation metrics such as Pearson's r or Spearman's rho cannot capture non-monotonic relationships. This leads us to an important realization that a lack of traditional correlation does not imply independence between variables.

A number of tools for more general independence testing exist. For instance, information-theoretic metrics such as the **maximal information coefficient (MIC)** work for non-linear, non-monotonic data out of the box. The same goes for the **Hilbert-Schmidt independence criterion (HSIC).**

Another scenario where you might see no correlation although causation is present is when your sampling does not cover the entire support of relevant variables.




This data is a result of exactly the same process as the data presented in Figure 2.4. The only difference is that we sampled according to the following condition:

X sample = 1.9 < X < − 1.9

It’s virtually impossible to estimate the true relationship between X and Y from this data, even with sophisticated tools.

Situations such as these can happen in real life. Everything from faulty sensors to selection bias in medical studies can remove a substantial amount of information from observational data and push us toward wrong conclusions.

## What are Counterfactuals?

Have you ever wondered where you would be today if you had chosen something different in your life? Moved to another city 10 years ago? Studied art? Dated another person? Taken a motorcycle trip in Hawaii? Answering these types of questions requires us to create alternative worlds, worlds that we have never observed. If you’ve ever tried doing this for yourself, you already know intuitively what counterfactuals are.

Counterfactuals can be thought of as **hypothetical or simulated** interventions that assume a particular state of the world (note that interventions do not require any assumptions about the state of the world).

For instance, answering a counterfactual question such as “Would John have bought the chocolate last Friday had he seen the ad last week?” requires us to know many things about John (and his environment) in the past. On the other hand, we don’t need to know anything about John’s life or environment from last week in order to perform an intervention (show John an ad and see whether he buys).

Contrary to interventions, counterfactuals can never be observed.

One idea that emphasizes the fundamental difference between counterfactuals and interventions is that interventional queries can be computed as the expected value of counterfactual queries over the population (Huszár, 2019).

## The Counterfactual Wonderland

Imagine you had a coffee this morning and now you feel bad in your stomach. Would you feel the same or better if you hadn’t had your coffee?

Note that we cannot answer this question with interventions. A randomized experiment would only allow us to answer questions such as “What is the probability that people similar to you react to coffee

the way you reacted, given similar circumstances?" or "What is the probability that you'll feel bad after drinking a coffee in the morning on a day similar to today, given similar circumstances?"

Counterfactuals are trying to answer a different question: "Given an alternative world that is identical to ours and only differs in the fact that you did not drink coffee this morning (plus the necessary consequences of not drinking it), what is the probability that you'd feel bad in your stomach?"

Let's try to formalize it. We'll denote the fact that you drank coffee this morning by $X = 1$ and the fact that you now feel bad as $Y_{X=1} = 1$ . The $X=1$ subscript informs us that the outcome, $Y$ , happened in the world where you had your coffee in the morning ( $X = 1$ ). The quantity we want to estimate, therefore, is the following:

$$P ( Y_{X=0} = 1 \mid X = 1, Y_{X=1} = 1 )$$

We read this as the probability that you'd feel bad if you hadn't had your coffee, given you had your coffee and you feel bad.

Let's unpack it:

- $P ( Y_{X=0} = 1 )$ stands for the probability that you'd feel bad (in the alternative world) if you hadn't had your coffee
- $X = 1$ denotes that you had your coffee in the real world
- $Y_{X=1} = 1$ says that you had your coffee and you felt bad (in the real world)

Note that everything on the right side of the conditioning bar comes from the actual observation. The expression on the left side of the conditional bar refers to the alternative, hypothetical world.

Many people feel uncomfortable seeing this notation for the first time. The fact that we're conditioning on $X = 1$ to estimate the quantity in the world where $X = 0$ seems pretty counterintuitive. On a deeper level, this makes sense though. This notation makes it virtually impossible to reduce counterfactuals to do expressions. This reflects the inherent relationship between interventions and counterfactuals that we discussed earlier in this section.

## Computing Counterfactuals

The basic idea behind computing counterfactuals is simple in theory, but the goal is not always easily achievable in practice. This is because computing counterfactuals requires that we have full knowledge of functions that relate to relevant variables in the SCM and full knowledge about the values of all relevant exogenous variables in a system.

Fortunately, if we know structural equations, we can compute noise variables describing the subject in the **abduction step**.

## Computing Counterfactuals Step by Step

Judea Pearl and colleagues proposed a three-step framework for computing counterfactuals:

- **Abduction**: Using evidence to calculate values of exogenous variables
- **Modification** (originally called an action): Replacing the structural equation for the treatment with a counterfactual value
- **Prediction**: Using the modified SCM to compute the new value of the outcome under the counterfactual

Let's take our coffee example. Let T denote our treatment – drinking coffee, while U will characterize you fully as an individual in our simplified world.

U = 1 stands for coffee sensitivity, while U = 0 stands for a lack thereof.

Additionally, let's assume that we know the causal mechanism for reaction to coffee. The mechanism is defined by the following SCM:

$T := t$

$Y := TU + (T - 1)(U - 1)$

Great! We know the outcome under the actual treatment, $Y_{T=1} = 1$ (you drank the coffee and you felt bad), but we don't know your characteristics ( U ). Can we do something about it?

It turns out that our model allows us to unambiguously deduct the value of U (that we'll denote as u ), given we know the values of Y and T.

Let’s solve for u by transforming our structural equation for Y :

$u = (T + Y - 1) / (2\,T - 1)$

Now, let’s assign the values for T and Y :

$u = (1 + 1 - 1) / (2\,1 - 1) = 1$

The value we obtained for U reveals that you’re a coffee-sensitive person.

This step (solving for the values of exogenous variables U ) is called **abduction**.

Now, we have all the elements necessary to compute counterfactual outcomes at our disposal – a causal model and knowledge about your personal characteristics.

We’re ready for the next step, **modification**. We will fix the value of our treatment at the counterfactual of interest, ( T = 0 ):

$T := 0$

$Y := 0U + ( 0 - 1 ) ( U - 1 )$

Finally, we’re ready for the last step, prediction. To make a prediction, we need to substitute U with the value(s) of your personal characteristic(s) that we computed before:

$Y := 0 * 1 + (0 - 1) (1 - 1) = 0$

And here is the answer – you wouldn’t feel bad if you hadn’t had your coffee!

The last thing we’ll do before concluding this section is to implement our counterfactual computation in Python:

[See the coding lab in Google Golab]

## Causality and Reinforcement Learning

For many people, the first family of machine learning methods that come to mind when thinking about causality is reinforcement learning (RL).

In the classic formulation of RL, an agent interacts with the environment. This suggests that an RL agent can make interventions in the environment. Intuitively, this possibility moves RL from an associative rung one to an interventional rung two.

Bottou et al. (2013) amplify this intuition by proposing that causal models can be reduced to multiarmed bandit problems – in other words, that RL bandit algorithms are special cases of rung two causal models.

Although the idea that all RL is causal might seem intuitive at first, the reality is more nuanced. It turns out that even for certain bandit problems, the results might not be optimal if we do not model causality explicitly (Lee and Bareinboim, 2018).

Moreover, model-based RL algorithms can suffer from confounding. This includes models such as the famous DeepMind's MuZero (Schrittwieser et al., 2019), which was able to master games such as Chess, Go, and many others without explicitly knowing the rules.

## Causality and Semi-Supervised and Unsupervised Learning

Peters et al. (2017, pp. 72–74) proposed that semi-supervised methods can be used for causal discovery, leveraging the information-theoretic asymmetry between conditional and unconditional probabilities. Based on a similar idea, Sgouritsa et al. (2015) proposed unsupervised inverse regression as a method to discover which causal direction between two variables is the correct one. Wu et al. (2022) explored the links between causality and semi-supervised learning in the context of domain adaptation, while Vowels et al. (2021) used unsupervised neural embeddings for the causal discovery of video representations of dynamical systems.

Furthermore, data augmentation in semi-supervised learning can be seen as a form of representation disentanglement, making the learned representations less confounded. As deconfounding is only partial in this case, these models cannot be considered fully causal, but perhaps we could consider putting these partially deconfounded representations somewhere between rungs one and two of the Ladder of Causation (Berrevoets, 2023).

In summary, the relationship between different branches of contemporary machine learning and causality is nuanced. That said, most broadly adopted machine learning models operate on rung one, not having a causal world model. This also applies to large language models such as GPT-3, GPT-4, LlaMA, or LaMDA, and other popular generative models such as DALL-E 2. Models such as GPT-4 can sometimes correctly answer causal or counterfactual queries, yet their general performance suggests that these abilities do not always generalize well.