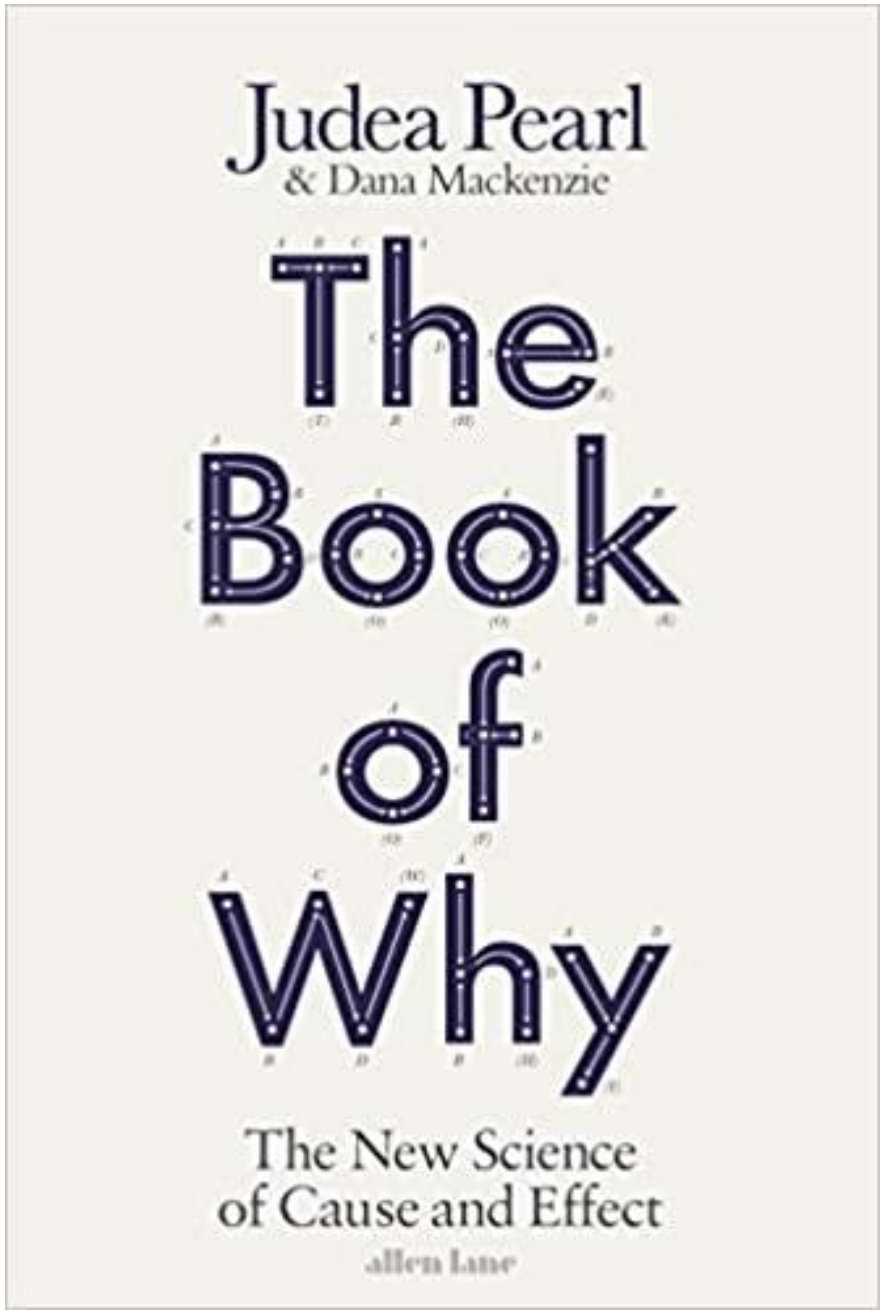


Correlation is NOT Causation

Towards Causality and Causal Inference

Prabakaran Chandran

03-March-2022

Any Connection with today’s Session?

Is there any relationships between these characters and Causation ?

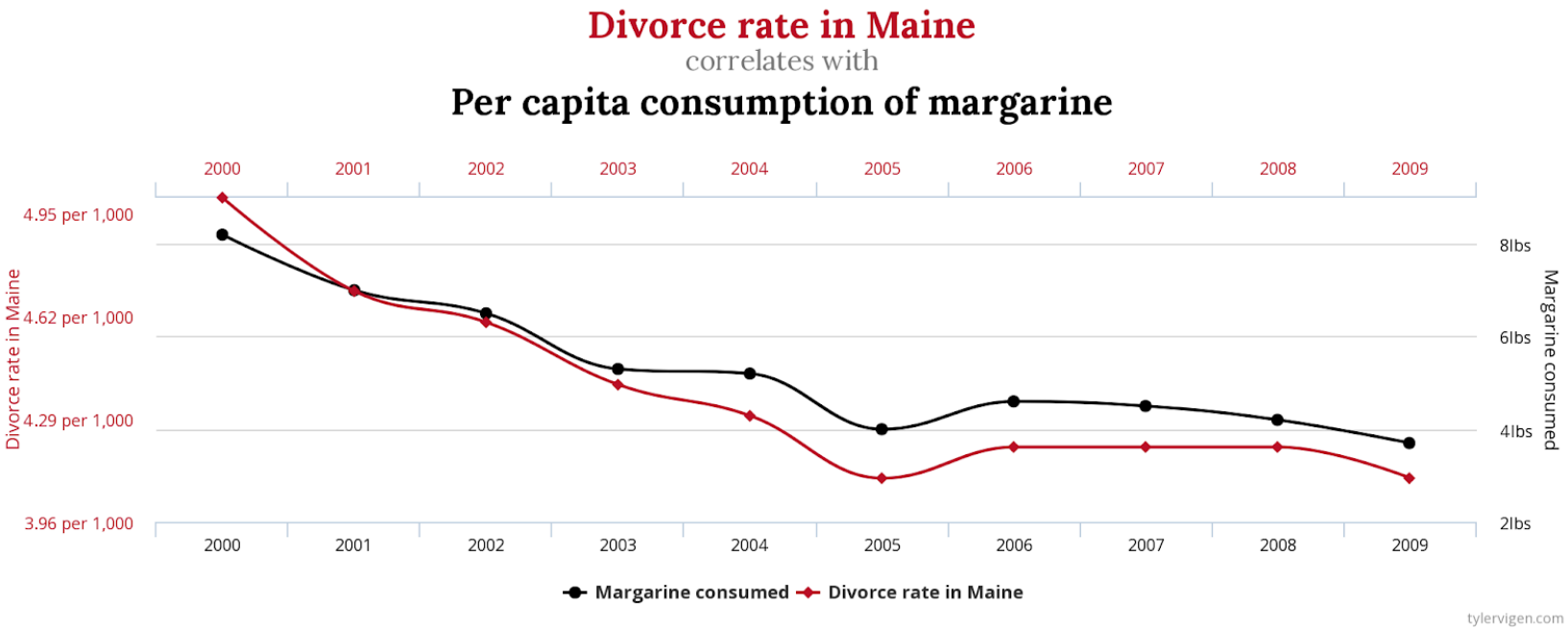
Our Mind either intentionally or unintentionally, probably tried to make a connection.

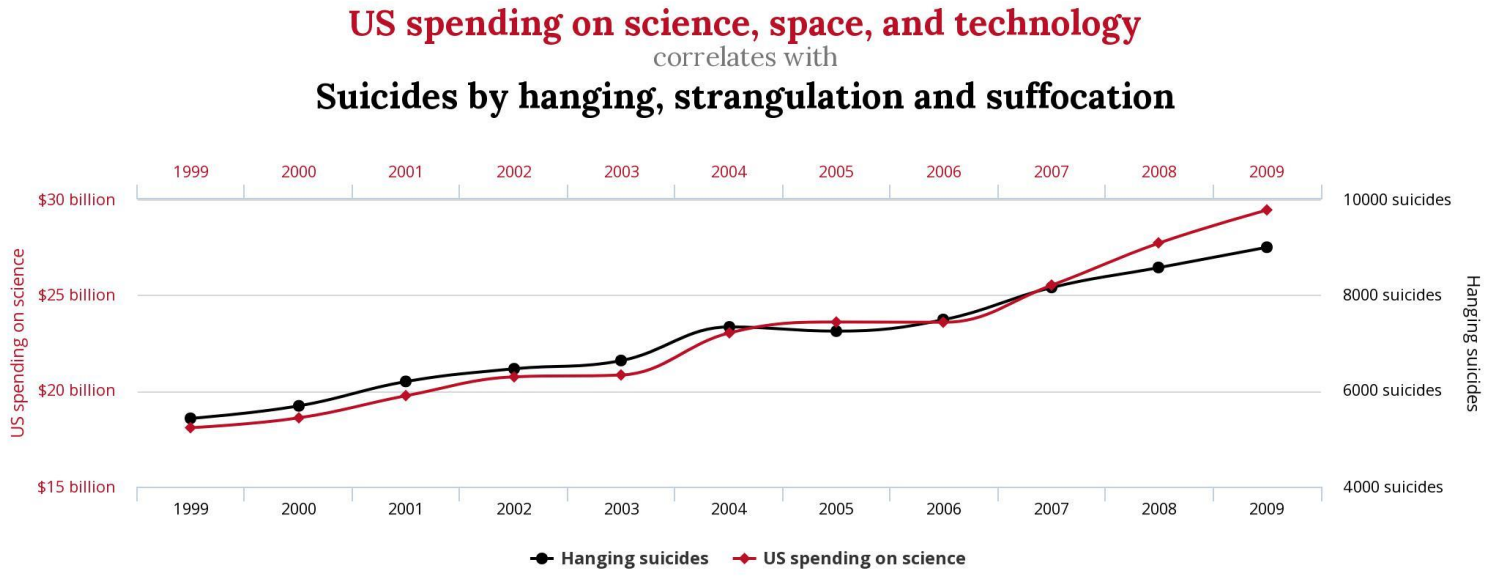
This leads to a Question Why this Character ?
Few Derivatives :

- What is the cause?
- What is the reason for this?
- What's next?
- Where is this going?

Unfortunately, these questions cannot be answered with just statistical inference

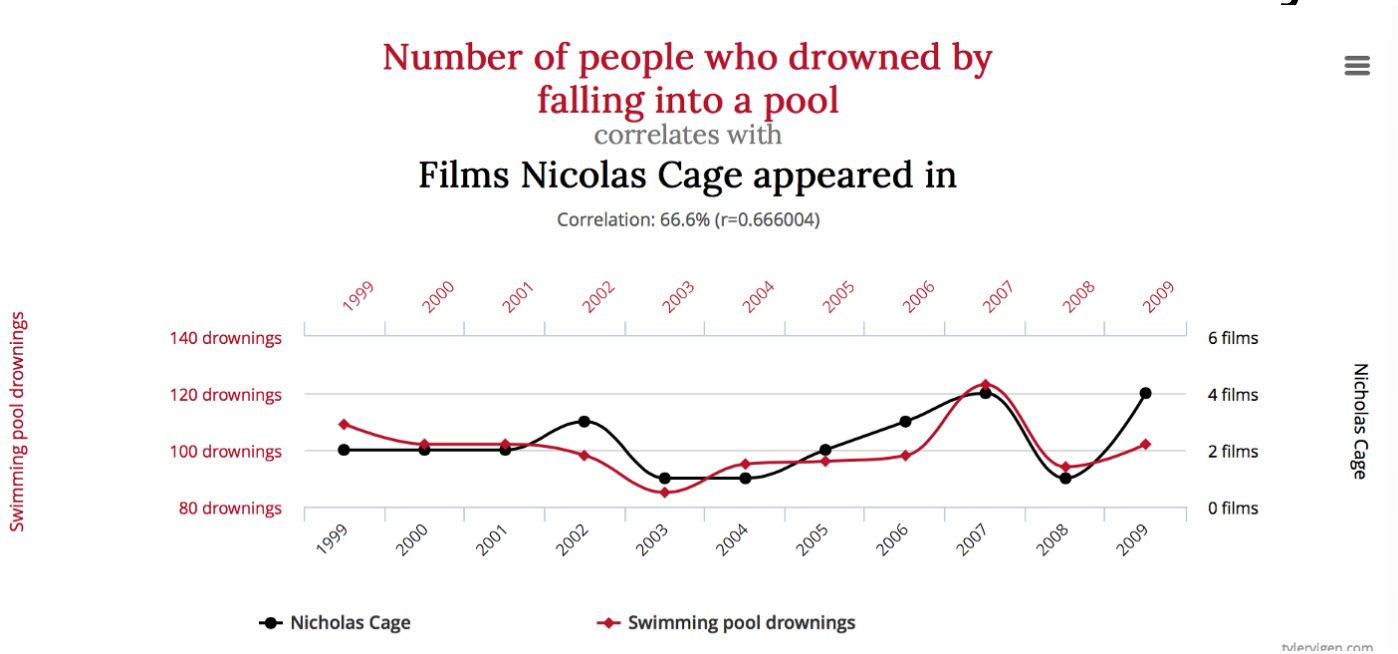
Can we infer these situations just with correlation

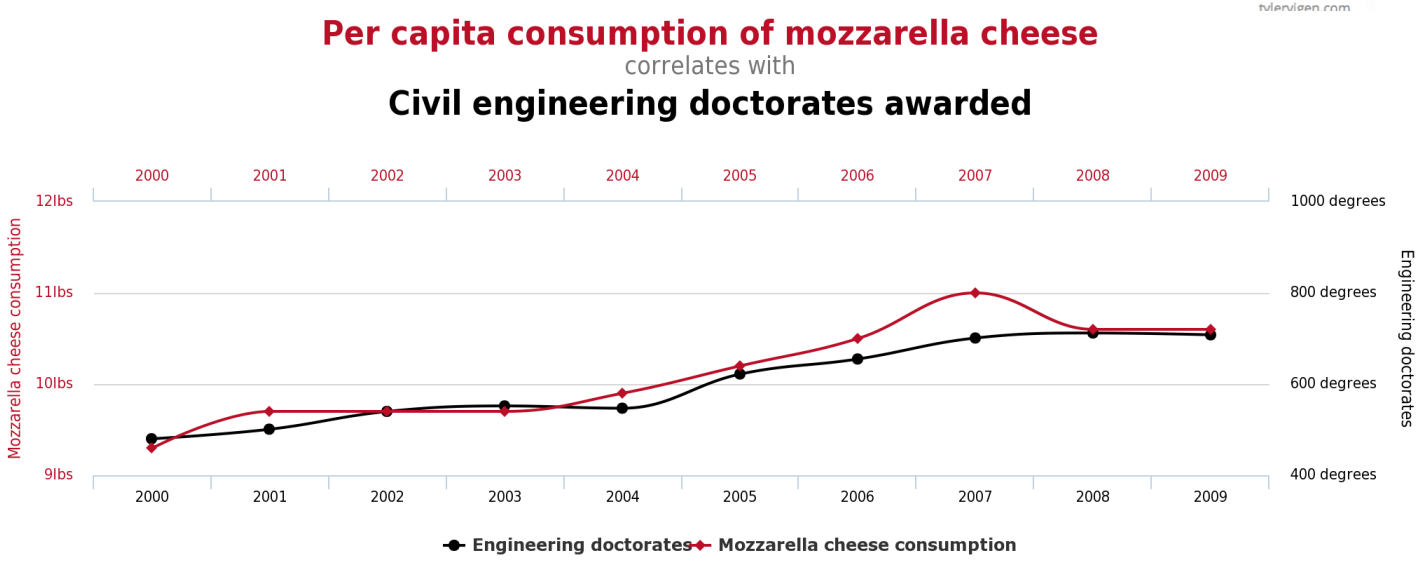




Can we say , Suicides causing US Spending on Science and Tech ? Or Other way around ?

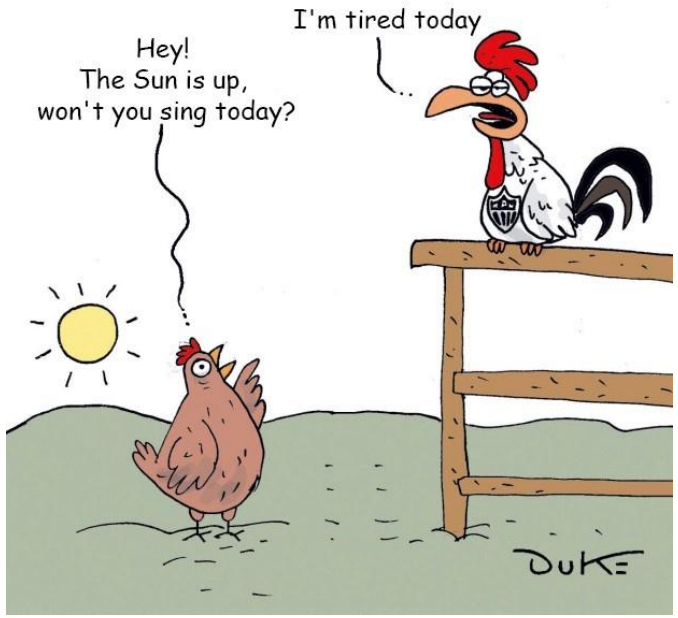
Can we infer these situations just with correlation





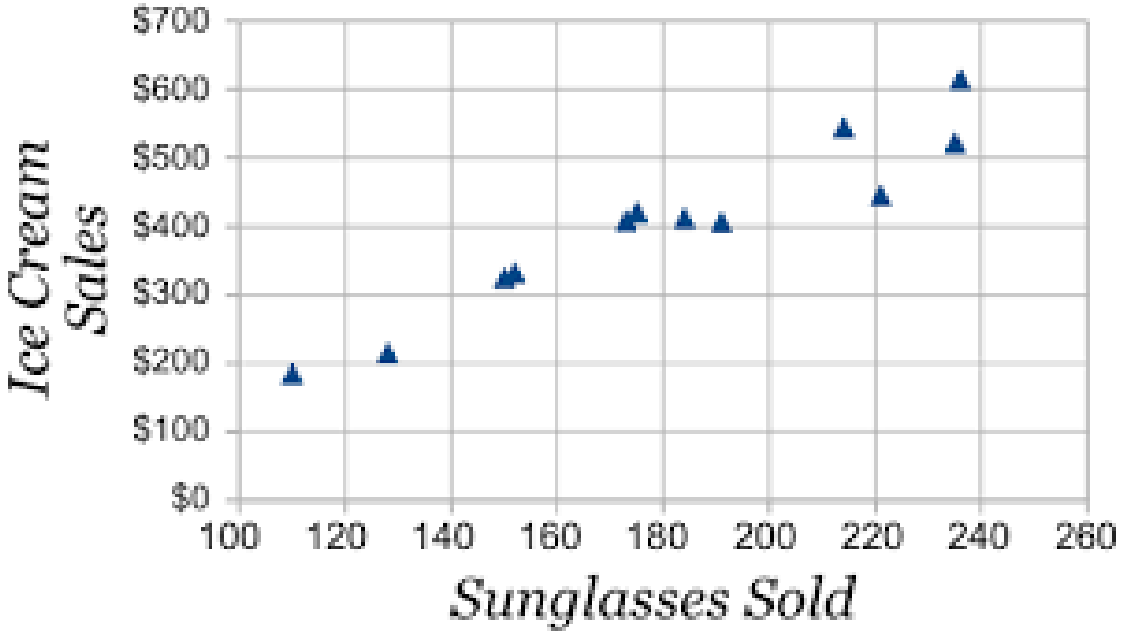
We could understand one Event is not causing another event , It is just the Same kind Data Shape

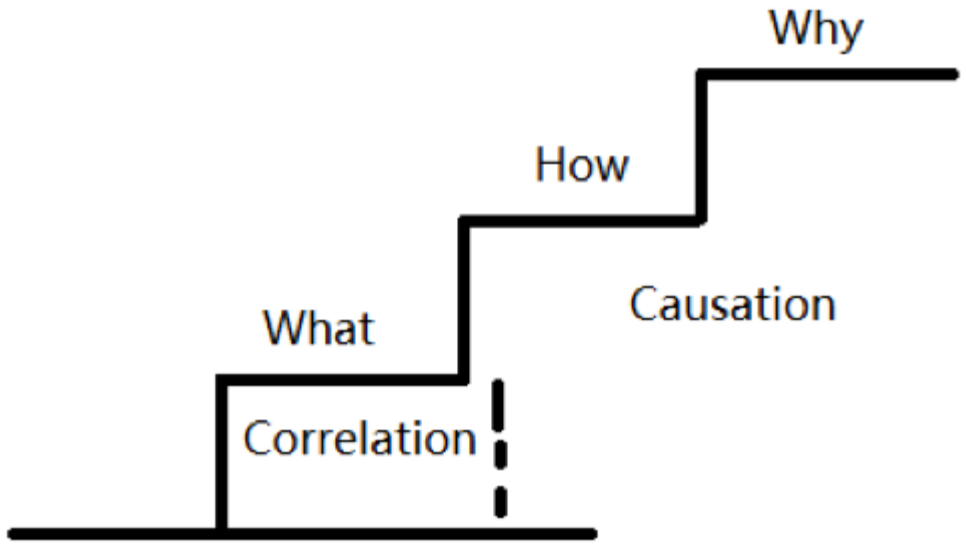
Even statistical significance is not sufficient to conclude One event is impacting another one!



Correlation – Association – A Small Recap

- statistical measure of the relationship between two variables
- The measure is best used in variables that demonstrate a linear relationship between each other
- The correlation coefficient is a value that indicates the strength of the relationship between variables
- The coefficient can take any values from -1 to 1.





Correlation – Association – A Small Recap

Population Correlation Coefficient

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where,

$\sigma_x, \sigma_y \rightarrow$ Population Standard Deviation

$\sigma_{xy} \rightarrow$ Population Covariance

$\bar{X}, \bar{Y} \rightarrow$ Population Mean

Sample Correction, coefficient between x and y

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where,

$S_x, S_y \rightarrow$ Sample Standard Deviation

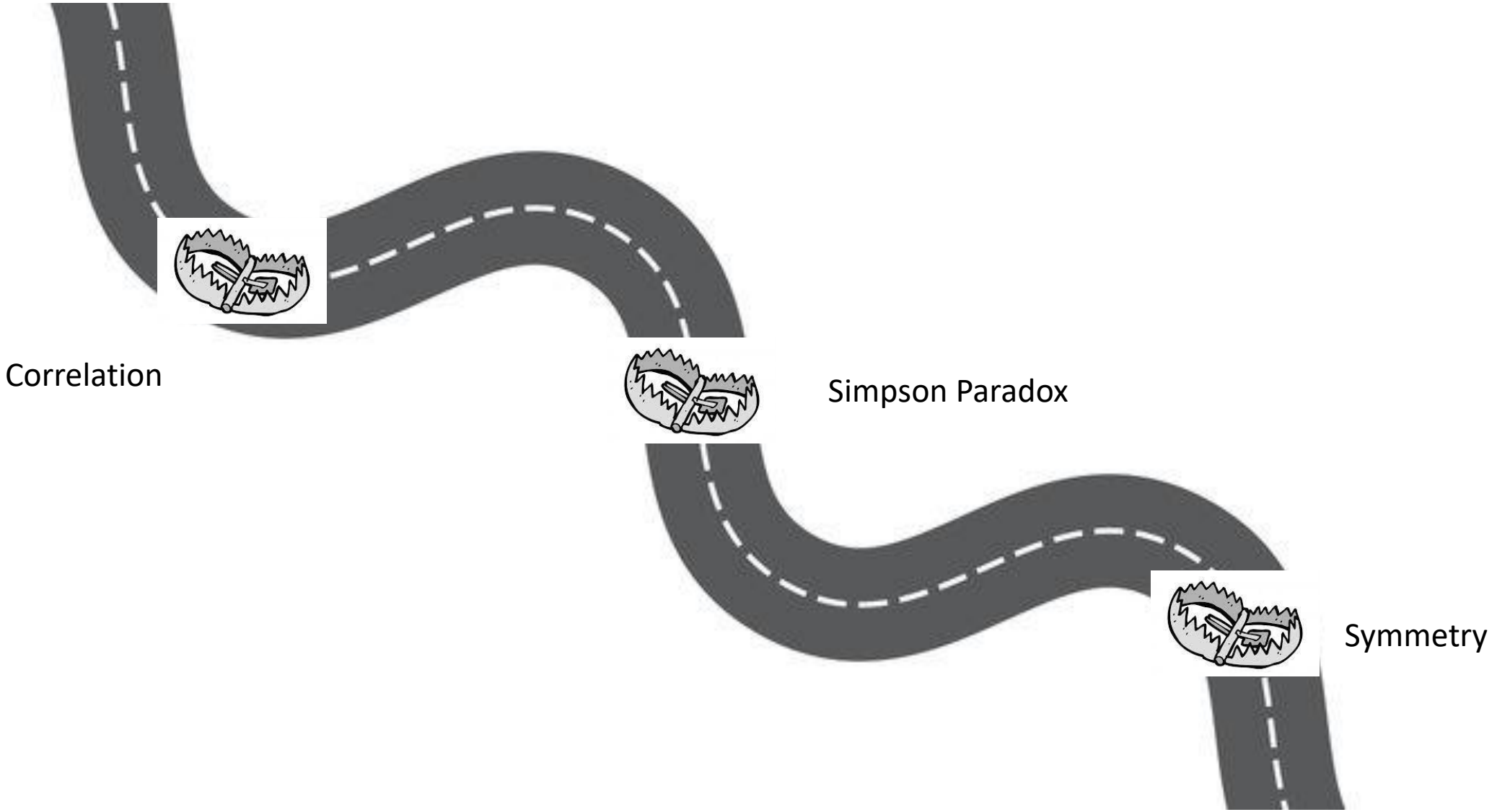
$S_{xy} \rightarrow$ Sample Covariance

$\bar{X}, \bar{Y} \rightarrow$ SampleMean

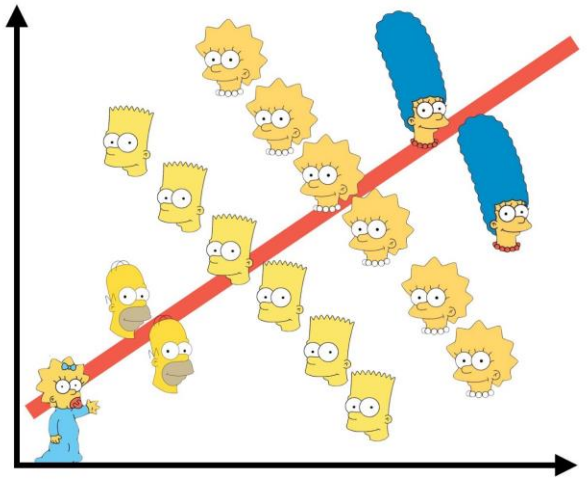
many traditional ML methodologies, from linear regression to deep learning, do not consider causality and instead only model correlation between datapoints. They may identify that there is a relationship between variables without defining what this relationship is or how they influence each other.

This can have a drastic impact on the model's suggested intervention, diluting the effectiveness of interventions or even producing entirely irrelevant recommendations. For example, a non-causal model aiming to mitigate drought may recognise that there is a relationship between rising drought and rising ice cream sales, but may spuriously conclude that banning ice cream would mitigate drought.

Few Statistical Traps that We should be Aware of!



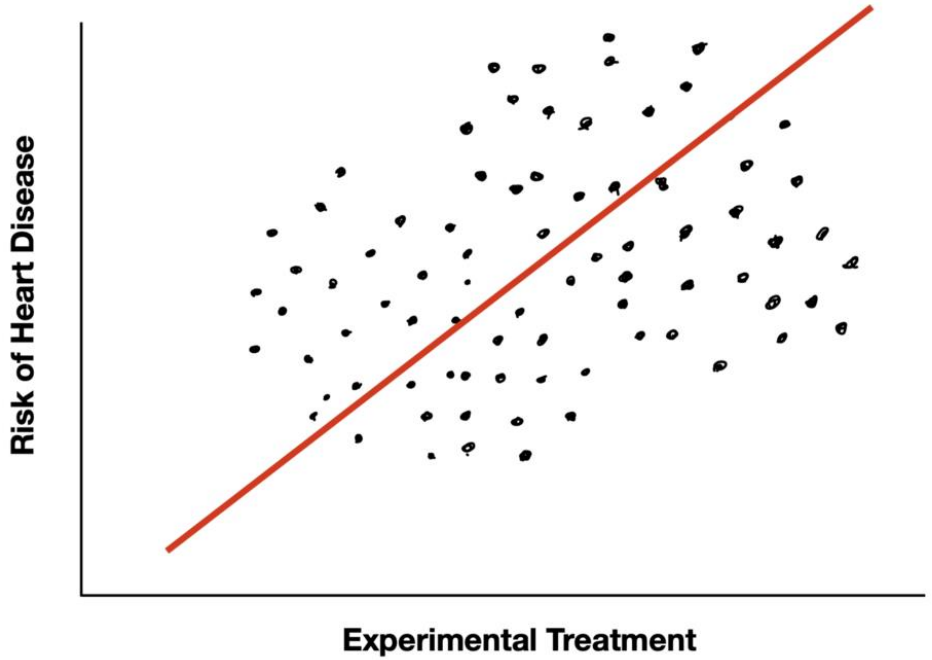
Trap - Simpson Paradox

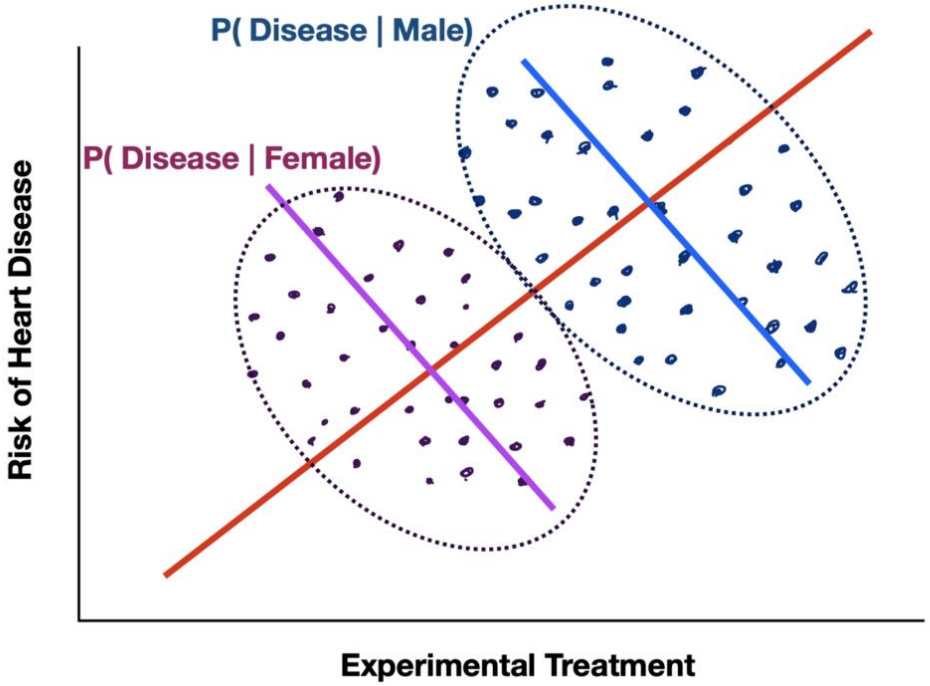


Simpson's Paradox which is **when the same data give contradictory conclusions depending on how you look at them**.

At first glance, we might conclude this is a terrible treatment. The more someone takes the pill or engages in the prescribed behavior the worse their risk of heart disease gets. But now suppose we look at two subpopulations of study participants as shown in the figure below.

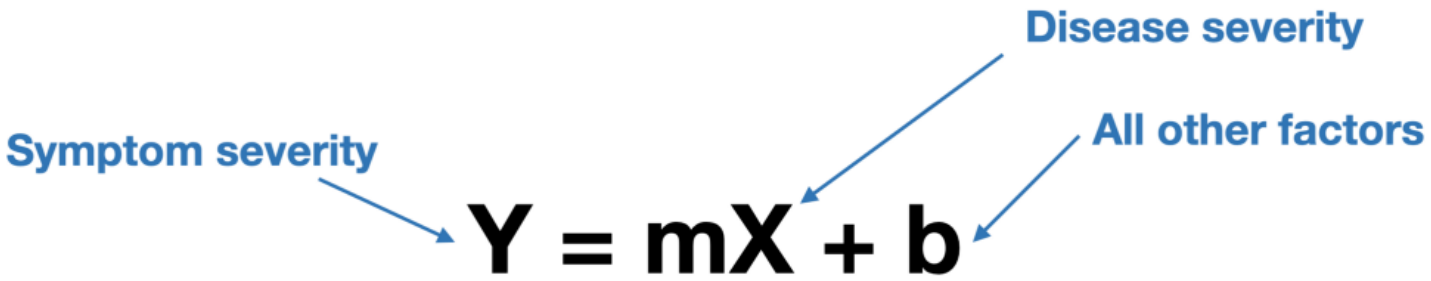
This paradox is summarized nicely by quote from Judea Pearl, **"we have a treatment that is good for man, good for a woman, but bad for a person"**





Trap - Symmetry

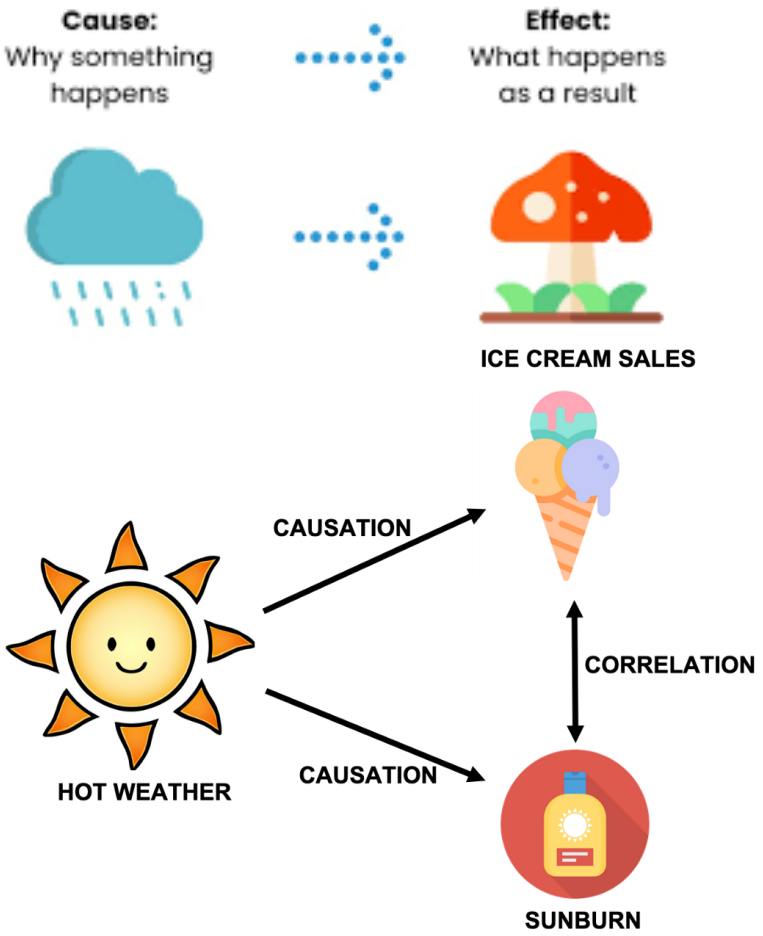
- Most of the Statistical Inference concepts are based on Linear Algebra
- The left-hand side of an equation equals the right-hand side (that's the point of algebra). The equal sign implies symmetry.
- causality is fundamentally asymmetric i.e. causes lead to effects and not the other way around.
- correlation between X and Y == correlation between Y and X.
- we need a different way that can help us to build asymmetric relationships to represent causality.



In Real word Causation , Relationships are asymmetric

What is Causality?

Causality means that there is a clear cause-effect relationship between two variables. Therefore, there is causation, when action A causes outcome B. It is a combination of action and reaction.

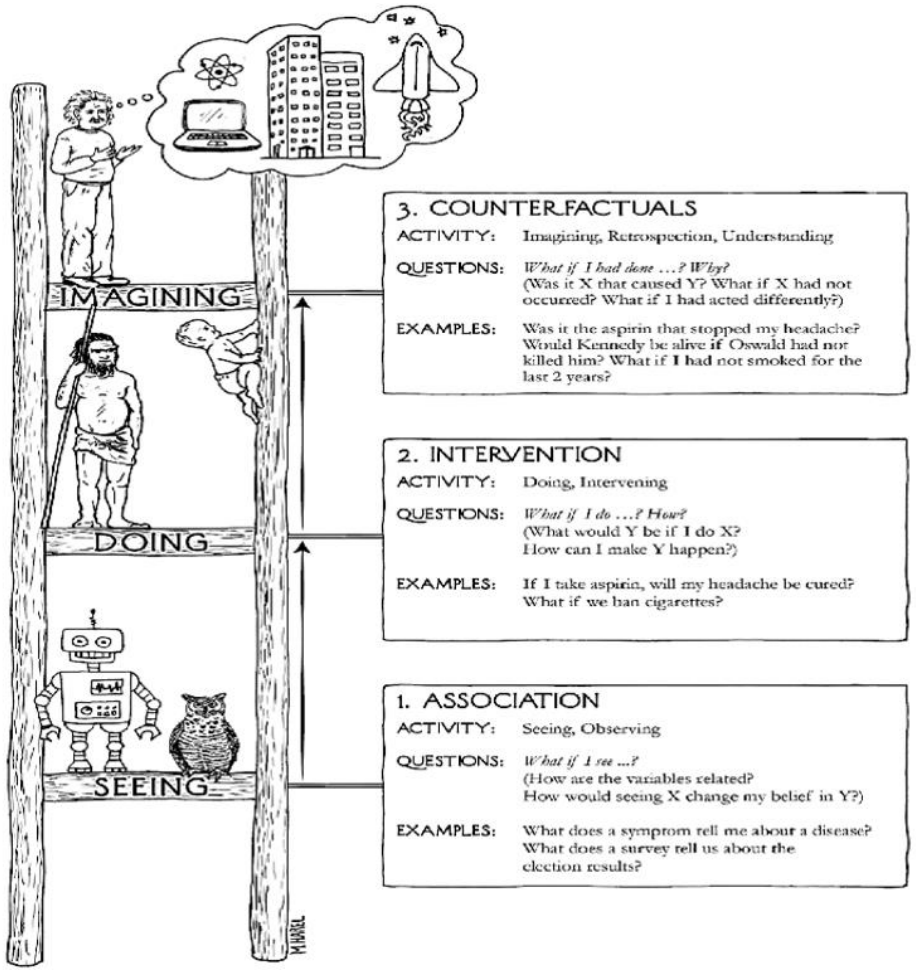


The Ladder of Causality

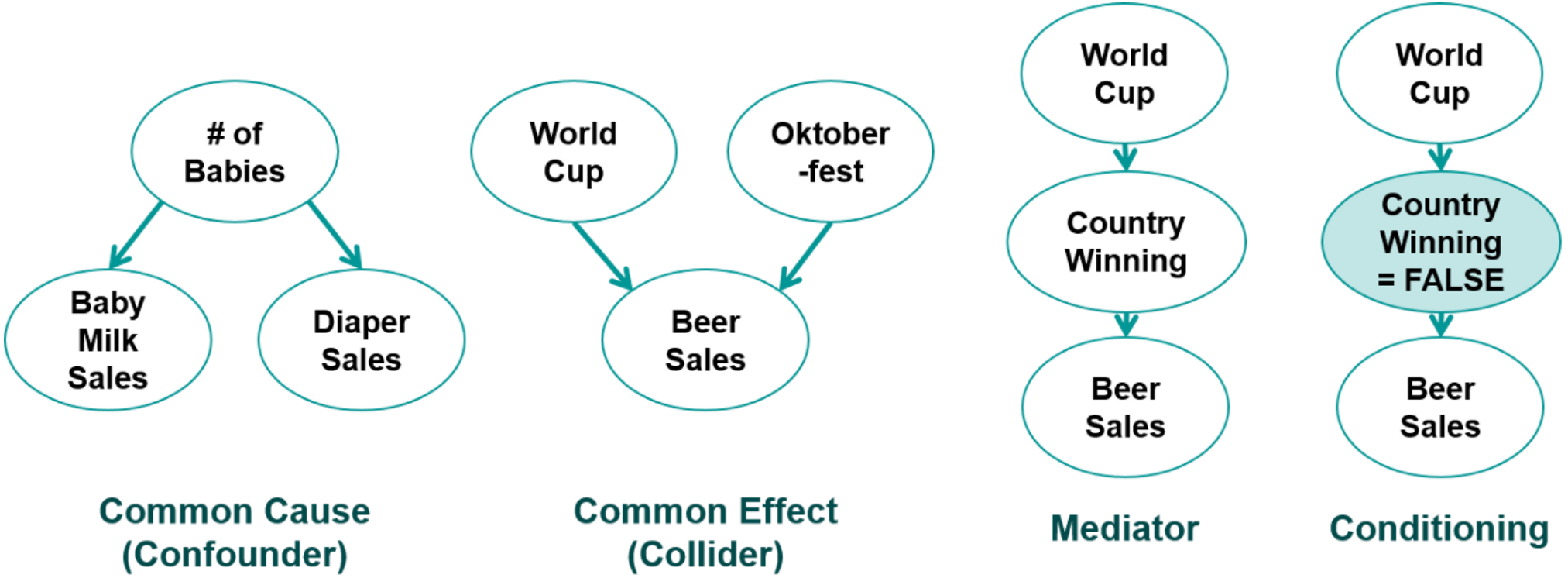
"Actual" Causality

"Causality-in-mean"

Statistics



What is Causation ? What is Cause and Effect ?



To go from correlation to causation, we need to remove all possible confounders.
If we control for all confounders (and account for random chance), and we still observe an association, we can say that there is causation.

1. Randomized Control Trails and A/B Testing Can be helpful in these scenarios
2. There are really expensive tests

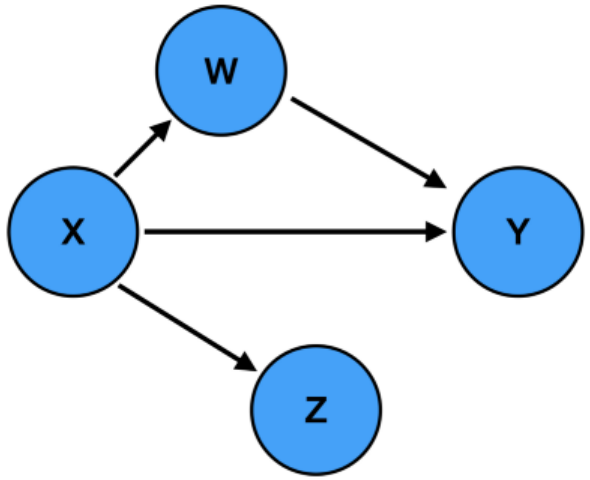
Mathematical Representation of Causality

Causality is represented mathematically via Structural Causal Models (SCMs) The two key elements of SCMs

1. graph - Directed Acyclic Graphs (DAG)
2. a set of equations - Structural Equation Model (SEM).

The goal of causal inference is to **answer questions based on the causal structure** of the problem.

Directed Acyclic Graphs (DAGs)



Structural Equation Models (SEMs)

$$W := f_1(X)$$

$$Z := f_2(X)$$

$$Y := f_3(X, W)$$

The causal connections of a system are often unknown.

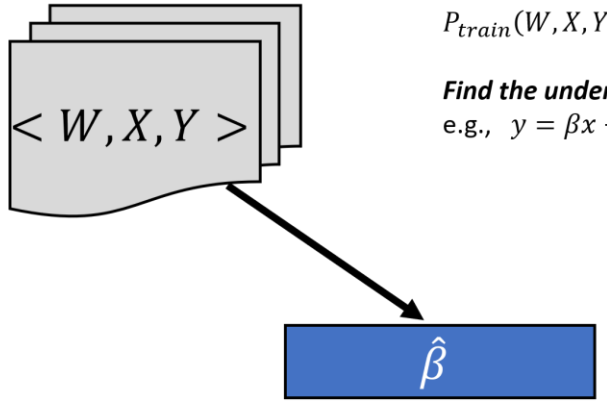
Causal discovery aims to uncover causal structure from observational data. causal discovery is an **inverse problem**.

It's like predicting the shape of ice cube based on the puddle it left on the kitchen floor

Where as Causal Inference assumes a defined Structure

Causal Inference using doWhy

Causal Inference

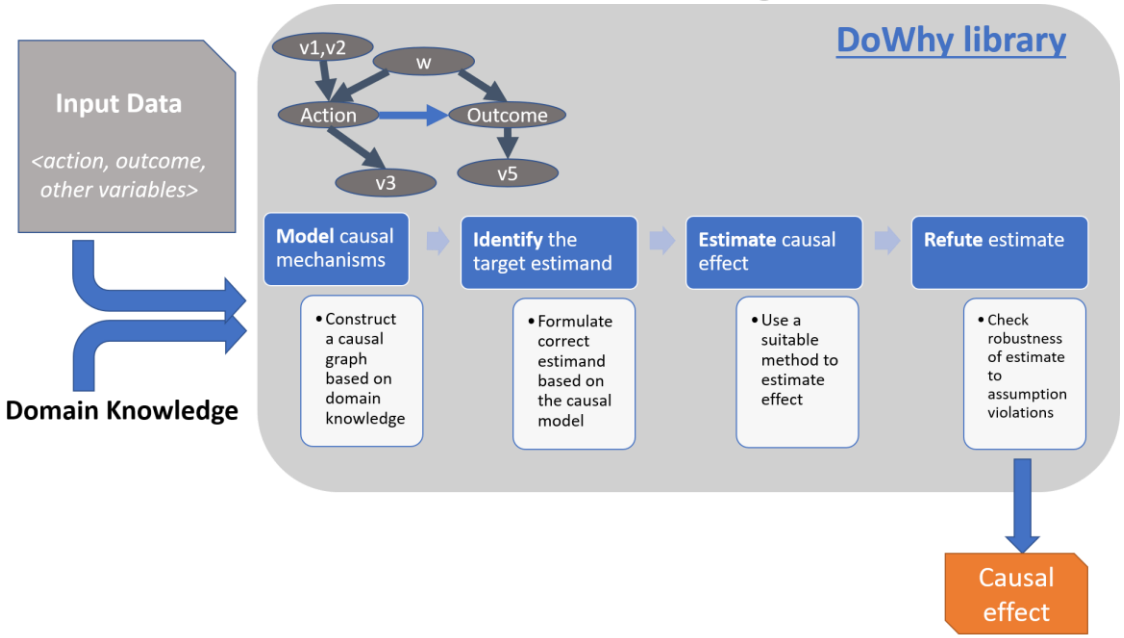


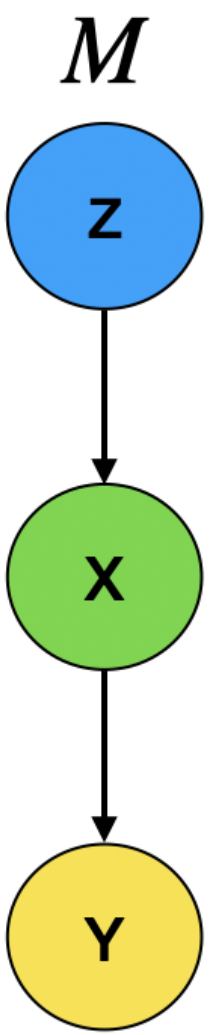
$P_{train}(W, X, Y) \neq P_{test}(W, X, Y)$

Find the underlying generative model

e.g., $y = \beta x + f(w) + \epsilon$

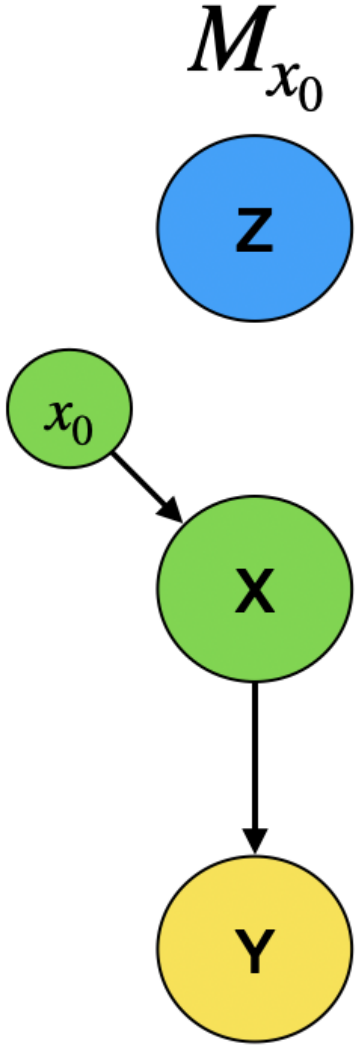
Decision-Making





For Simulating Experiments and represent interventions

$do(x_0)$



Thanks for Joining