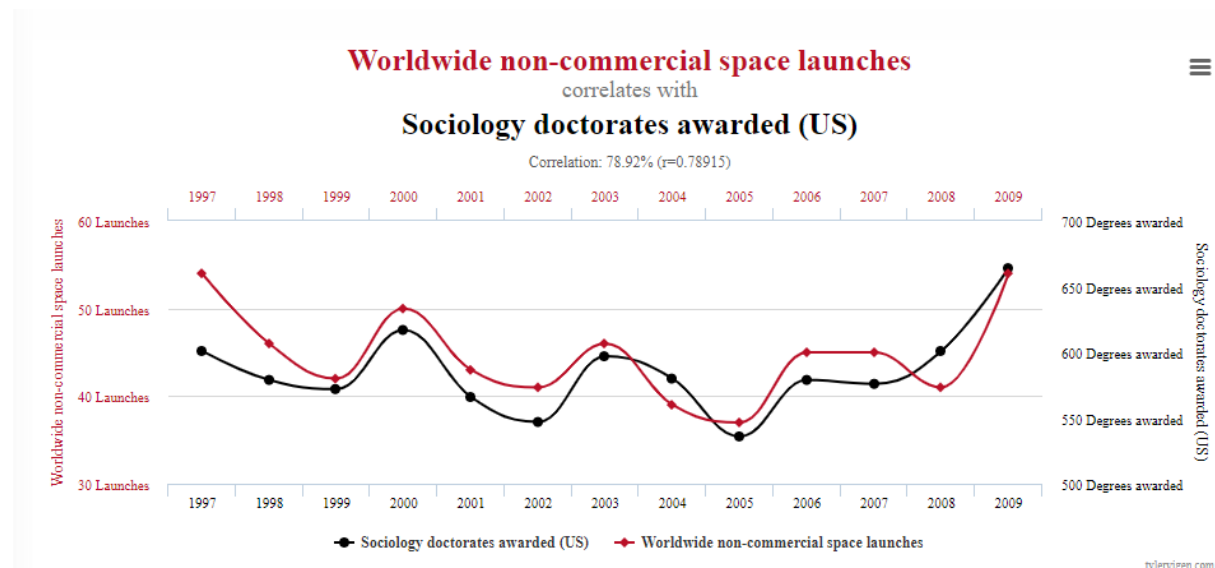
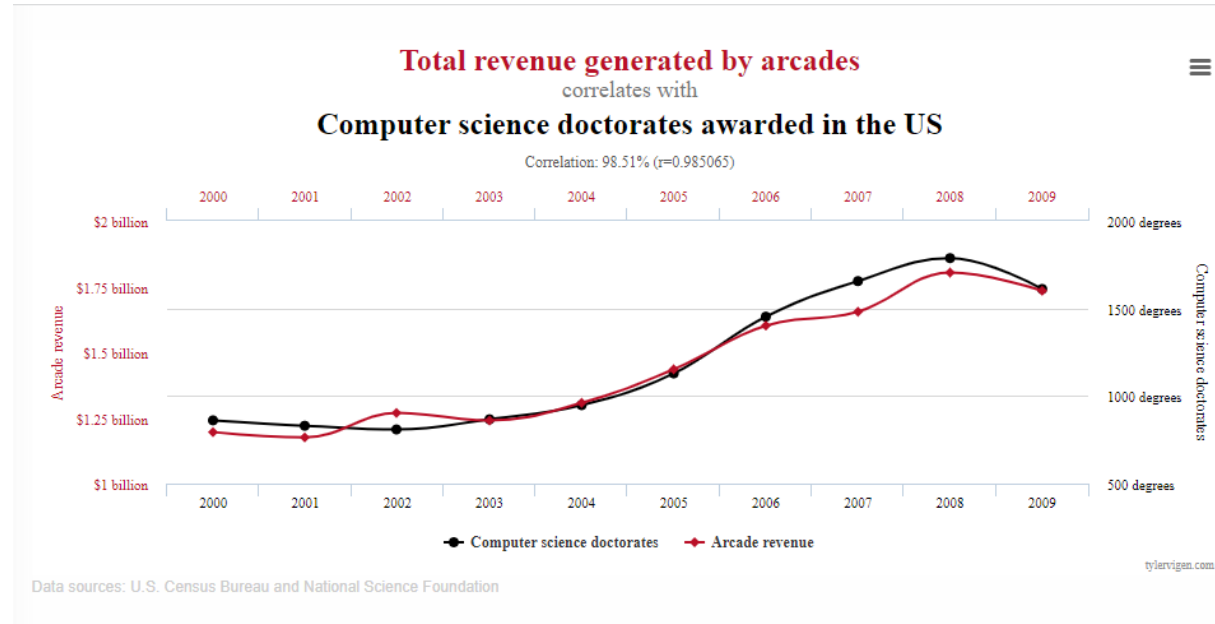


# Lecture 38: Causality

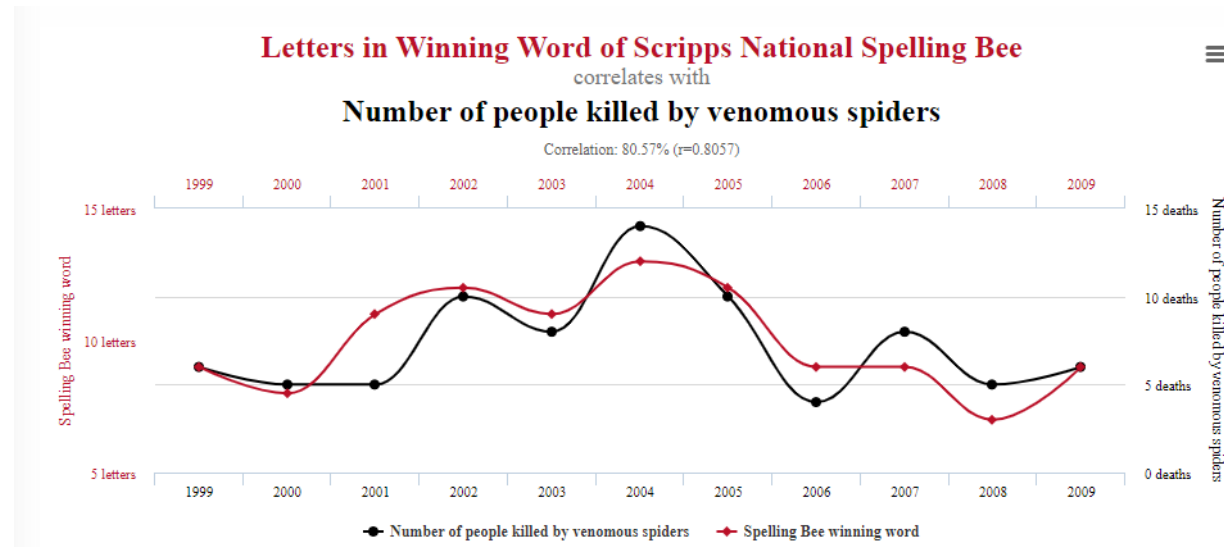
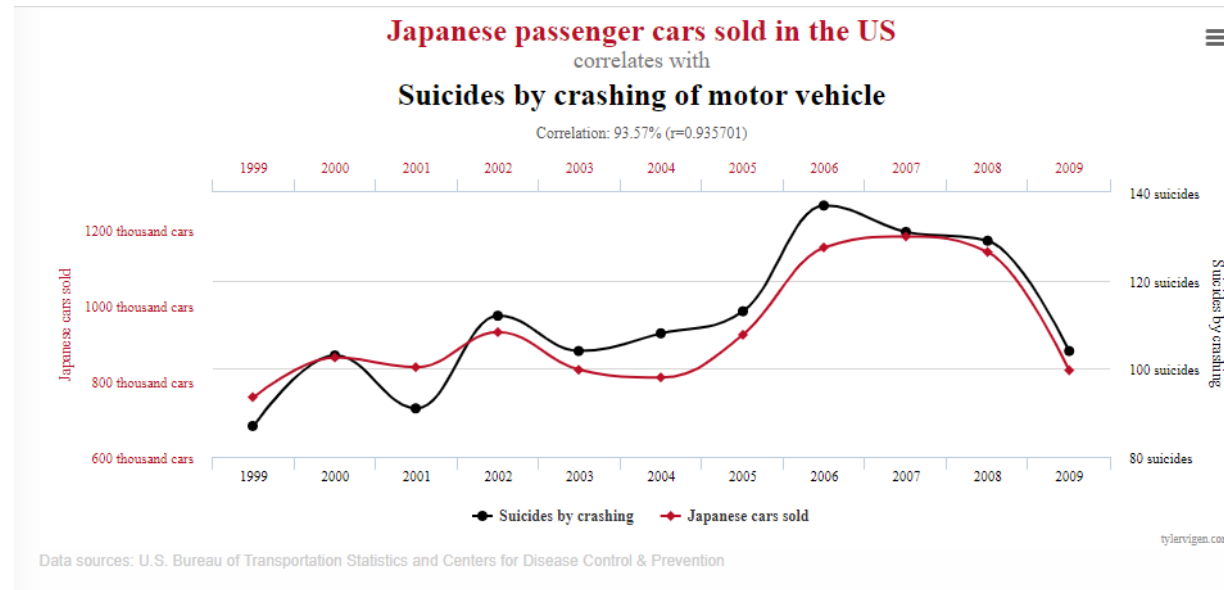
Instructor: Sergei V. Kalinin

# Correlation and causation



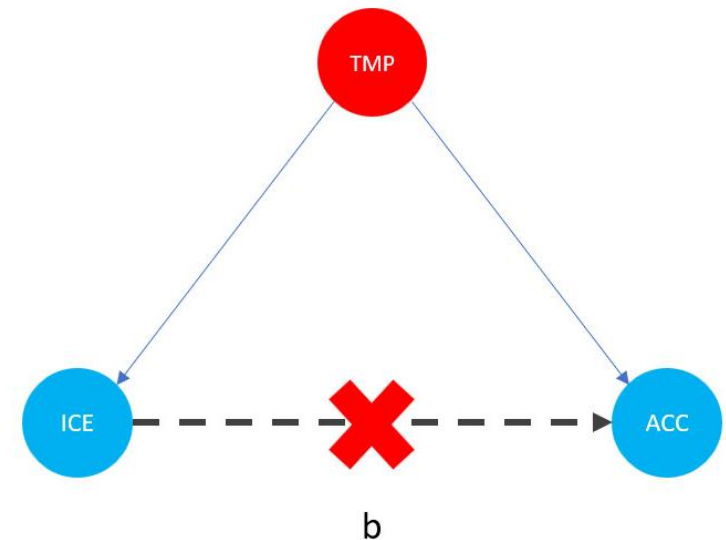
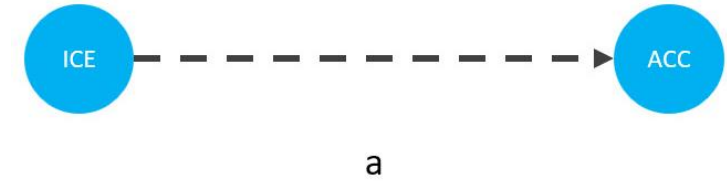
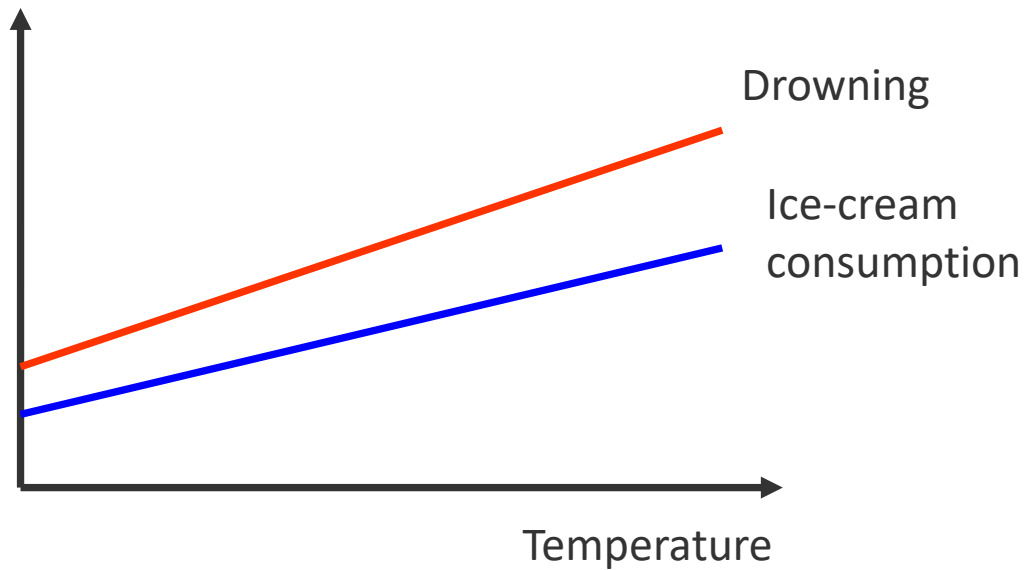
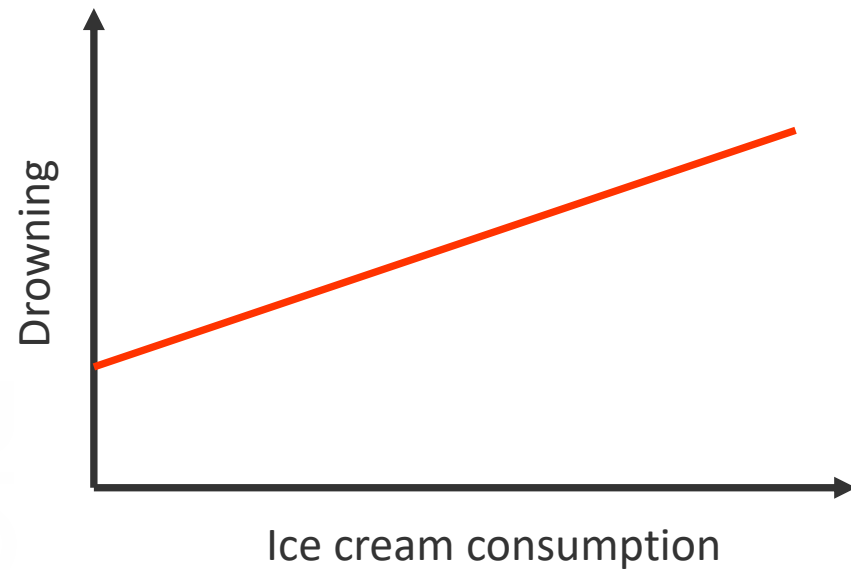
<https://www.tylervigen.com/spurious-correlations>

# Correlation and causation



<https://www.tylervigen.com/spurious-correlations>

# Ice-cream and drowning



# Treatment effects

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

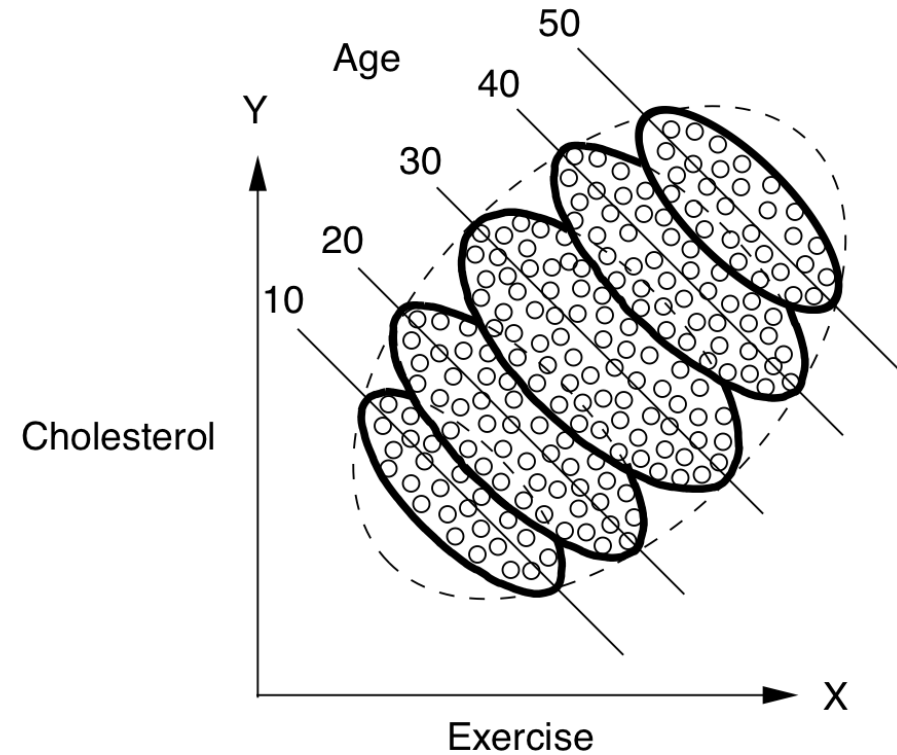
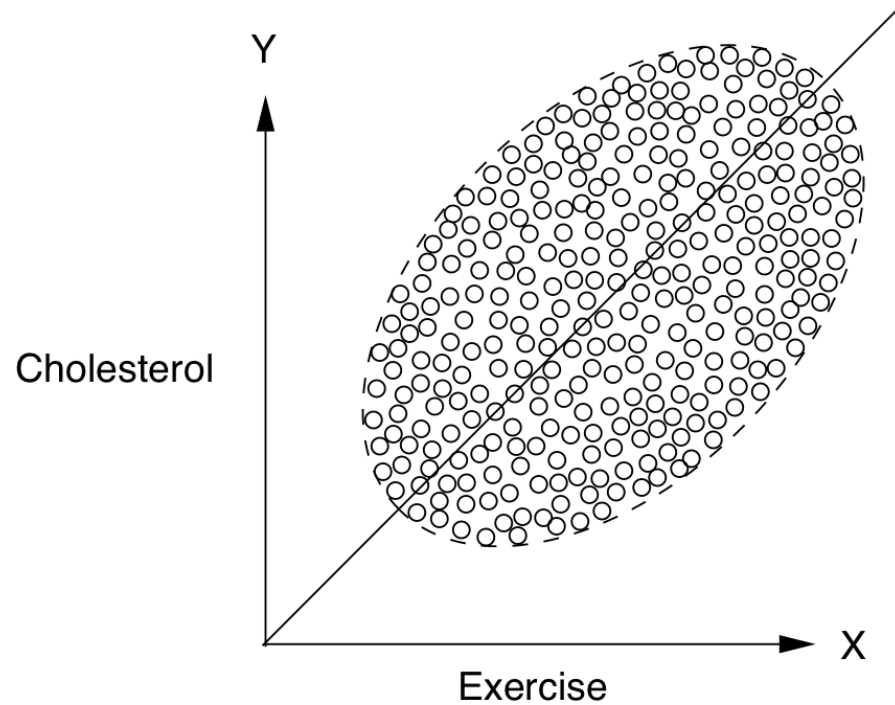
- $\tau_i$  is the treatment effect for person  $i$
- $Y_i(1)$  is the outcome for person  $i$  when they received the treatment T
- $Y_i(0)$  is the outcome for the same person  $i$  given they did not receive the treatment

## **Very fundamental approach in:**

- Marketing
- Medicine
- And so on

But... How can the same person be treated and not treated?

# Simpson paradox



Exercise is helpful in every age group but harmful for a typical person.

Is exercise helpful or not?

# Simpson paradox

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

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%

- 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?

# Berkeley discrimination lawsuit

In the early 1970s, the University of California, Berkeley was sued for gender discrimination over admission to graduate school. Of the 8,442 male applicants for the fall of 1973, 44 percent were admitted, but only 35 percent of the 4,351 female applicants were accepted

**Table 1: Data From Six Largest Departments of 1973 Berkeley Discrimination Case**

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
<b>A</b>	825	62%	108	82%
<b>B</b>	560	63%	25	68%
<b>C</b>	325	37%	593	34%
<b>D</b>	417	33%	375	35%
<b>E</b>	191	28%	393	24%
<b>F</b>	272	6%	341	7%

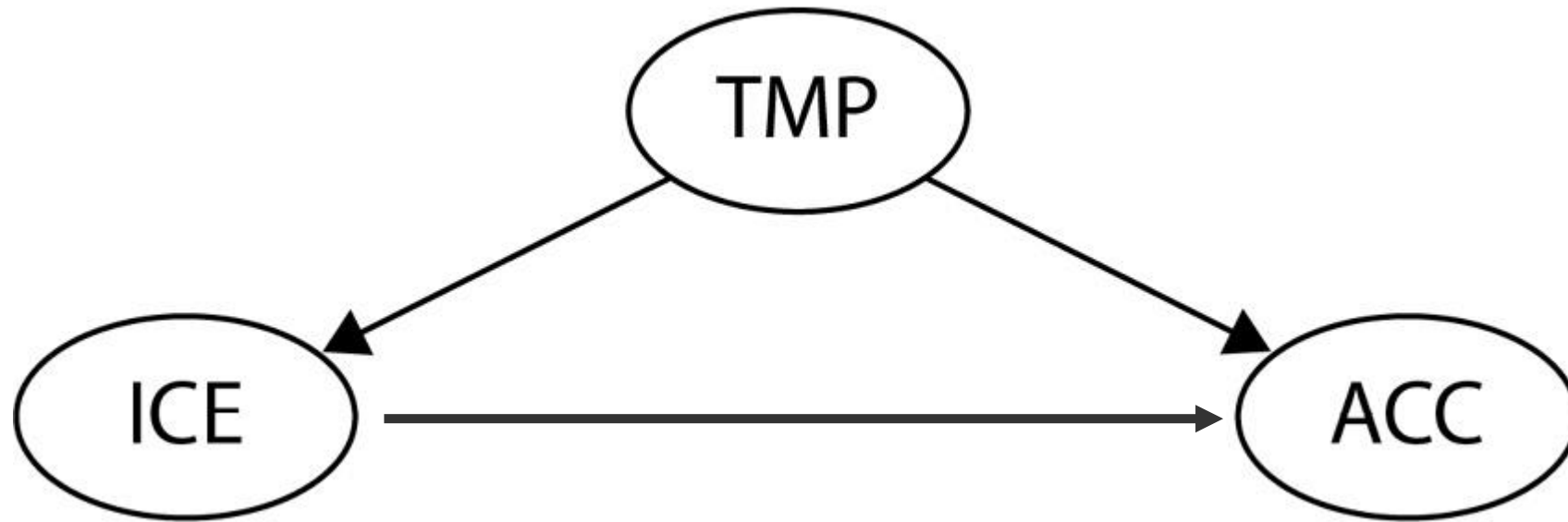
Source: Bickel, Hammel, and O'Connell (1975); table accessed via Wikipedia at [https://en.wikipedia.org/wiki/Simpson%27s\\_paradox](https://en.wikipedia.org/wiki/Simpson%27s_paradox)

In the Berkeley case, the “paradox” occurred because women disproportionately applied to departments with low acceptance rates, as shown in the table above, while men disproportionately applied to departments with high acceptance rates.

<https://www.brookings.edu/articles/when-average-isnt-good-enough-simpsons-paradox-in-education-and-earnings/>

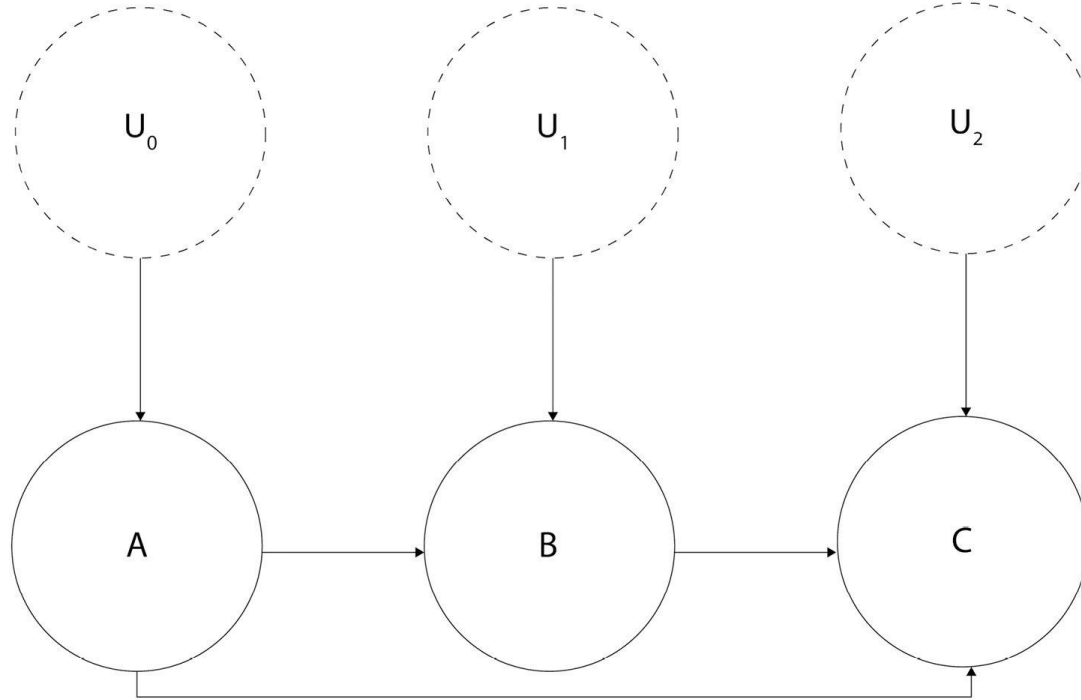


# How can we even approach such problems?



- Observations give us correlations between temperature, ice cream consumption, and accident rate
- What we need to know is the causal links between these characteristics. Does change in ice cream consumption affect temperature or accident rate?
- But we cannot make an experiment!

# Causal graphs



$$A := f_A(U_0)$$

$$B := f_B(A, U_1)$$

$$C := f_C(A, B, U_2)$$

- Here,  $:=$  is an **assignment operator**, also known as a **walrus operator**. We use it 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.
- And  $f_A$ ,  $f_B$ ,  $f_C$  represent arbitrary functions (they can be as simple as a summation or as complex as you want).

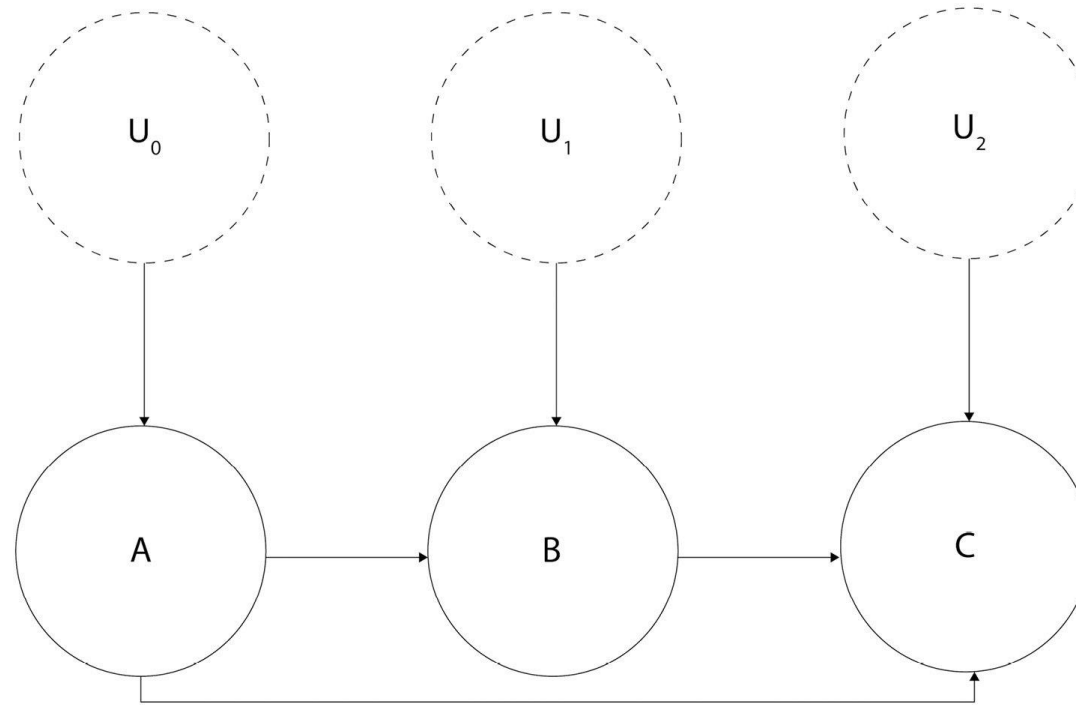
# Do-operator

**Conditioning:**  $P(X = x | Y = y)$

**Intervention:**  $P(Y = 1 | do(X = 0))$

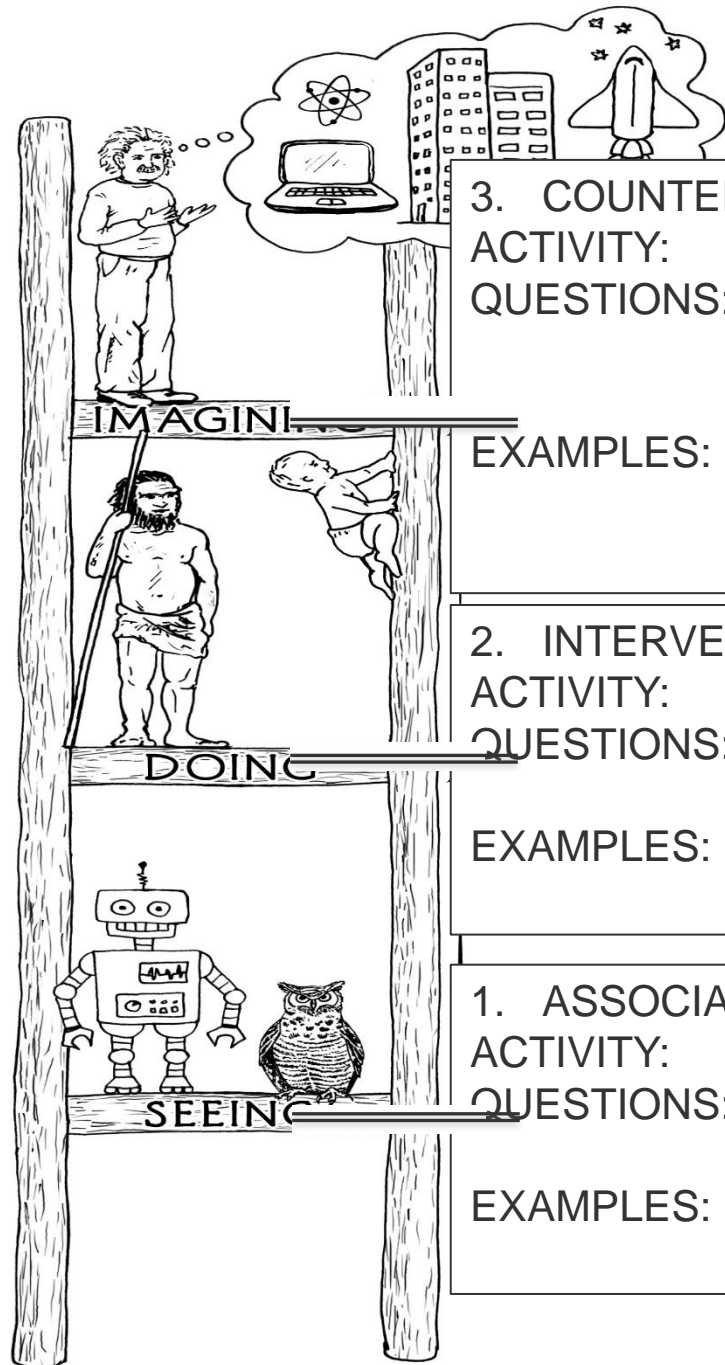
- Conditioning only modifies our *view* of the data, while intervening affects the distribution by *actively* setting one (or more) variable(s) to a *fixed value* (or a distribution).
- This is very important – intervention *changes* the system, but conditioning *does not*.
- You might ask, what does it mean that *intervention changes the system*? Great question!

# Properties of do - operator



- The change in  $B$  will influence the values of its descendants
- $B$  will become independent of its ancestors

# Ladder of causation



## 3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done . . . ? Why?*

(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES:

Was it the aspirin that stopped my headache?

Would Kennedy be alive if Oswald had not killed him?

What if I had not smoked the last 2 years?

## 2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do . . . ? How?*

(What would Y be if I do X?)

EXAMPLES:

If I take aspirin, will my headache be cured?

What if we ban cigarettes?

## 1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see . . . ?*

(How would seeing X change my belief in Y?)

EXAMPLES:

What does a symptom tell me about a disease?

What does a survey tell us about the election results?

# Invention of counterfactuals

- About 70,000 years ago, Sapiens from East Africa spread into the Arabian peninsula, and from there they quickly overran the entire Eurasian landmass, wiping out the native population.
- What was the Sapiens' secret of success?
- The ability to imagine things that do not really exist.



# How can we learn causality

- **Causal discovery** and **causal structure learning** are umbrella terms for various kinds of methods used to uncover causal structure from observational or interventional data.
- **Expert knowledge** is a term covering various types of knowledge that can help define or disambiguate causal relations between two or more variables. Depending on the context, expert knowledge might refer to knowledge from randomized controlled trials, laws of physics, a broad scope of experiences in a given area, and more.
- **Combining causal discovery and expert knowledge:** Some causal discovery algorithms allow us to easily incorporate expert knowledge as a priority. This means that we can either *freeze* certain edges in the graph or *suggest* the existence or direction of these edges.

# Independence and conditional independence

- Notation for independence involves the symbol,  $\perp\!\!\!\perp$  (usually called *double up tack*), whose form visually encodes the notion of orthogonality.
- We can express the fact that X and Y are independent in the following way:

$$P(X, Y) = P(X)P(Y) \quad X \perp\!\!\!\perp Y$$

- The concept of independence plays a vital role in statistics and causality.
- Its generalization – **conditional independence** – is even more important. We say that X and Y are conditionally independent given Z, when X does not give us any new information about Y assuming that we observed Z.

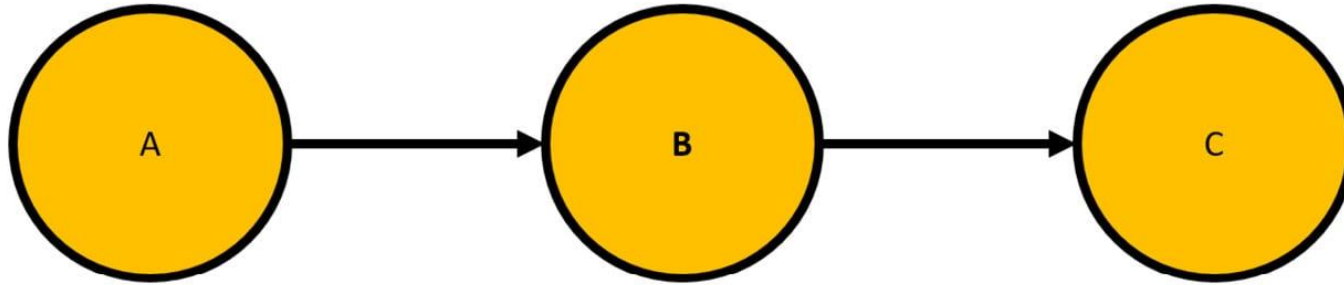
$$P(X, Y|Z) = P(X|Z)P(Y|Z) \quad X \perp\!\!\!\perp Y|Z$$



# Conditional and unconditional independence

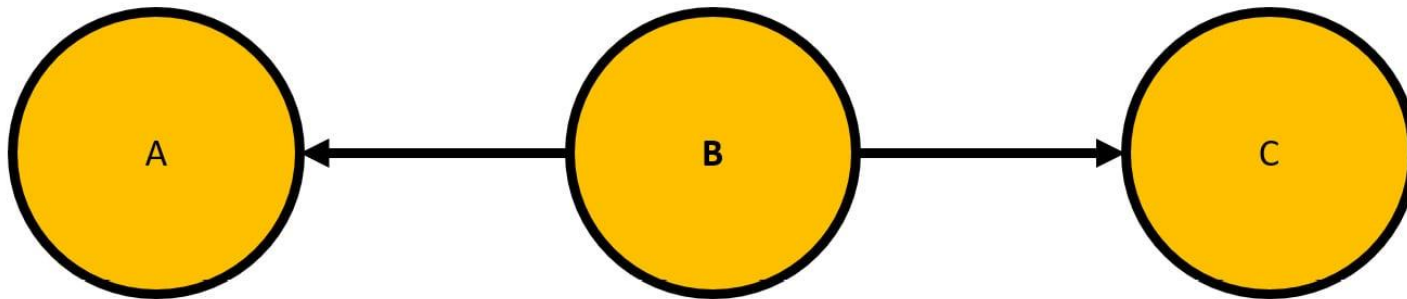
- We say that two nodes are *unconditionally* (or marginally) *independent* in the graph when there's *no open path* that connects them *directly* or *indirectly*.
- We say that two nodes,  $X$  and  $Y$ , are *conditionally independent* given (a set of) node(s)  $Z$  when  $Z$  blocks *all open paths* that connect  $X$  and  $Y$ .

# Chains and forks



$$A \perp\!\!\!\perp_G C | B$$

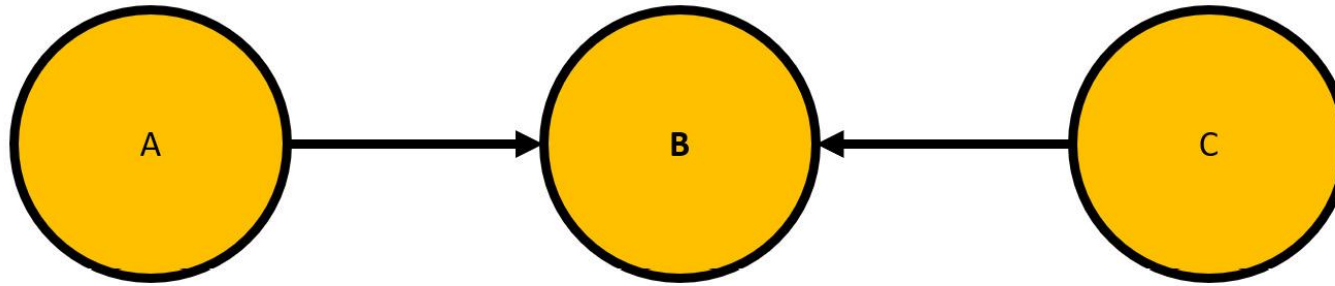
Chain 



$$A \perp\!\!\!\perp_G C | B$$

Fork 

# Colliders



$$\begin{array}{l} A \perp\!\!\!\perp C \\ A \not\perp\!\!\!\perp C \mid B \end{array}$$

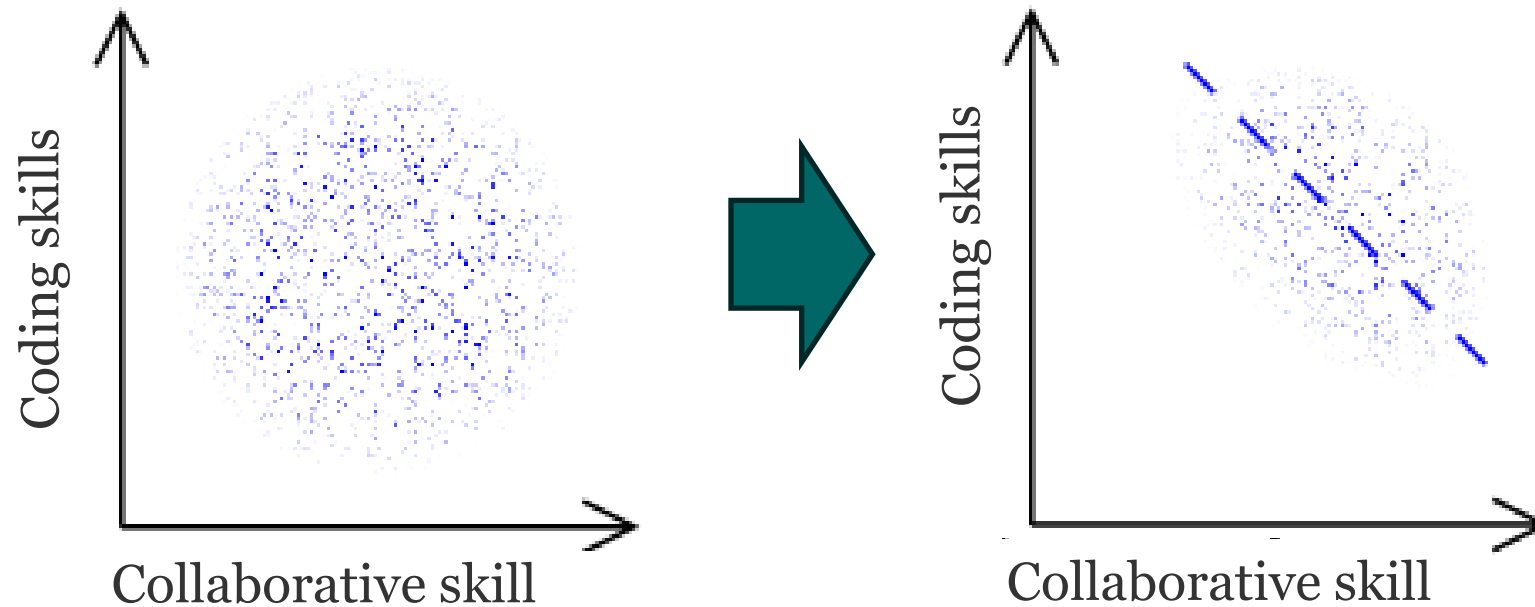
Collider 

Imagine that both A and C randomly generate integers between 1 and 3. Let's also say that B is a sum of A and C. Now, let's take a look at values of A and C when the value of B is 4. The following are the combinations of A and C that lead to  $B = 4$ :

- $A = 1, C = 3$
- $A = 2, C = 2$
- $A = 3, C = 1$

Although A and C are unconditionally independent (there's no correlation between them as they randomly and independently generate integers), they become correlated when we observe C !

# Colliders and Berkson paradox



Many companies might hire people based on their skills and their personality traits. Imagine that company *X* quantifies a person's coding skills on a scale from one to five. They do the same for the candidate's ability to cooperate and hire everyone who gets a total score of at least seven. Assuming that coding skills and ability to cooperate are independent in the population (which doesn't have to be true in reality), you'll observe that in company *X*, people who are better coders are less likely to cooperate on average, and those who are more likely to cooperate have fewer coding skills. You could conclude that being non-cooperative is related to being a better coder, yet this conclusion would be incorrect in the general population.

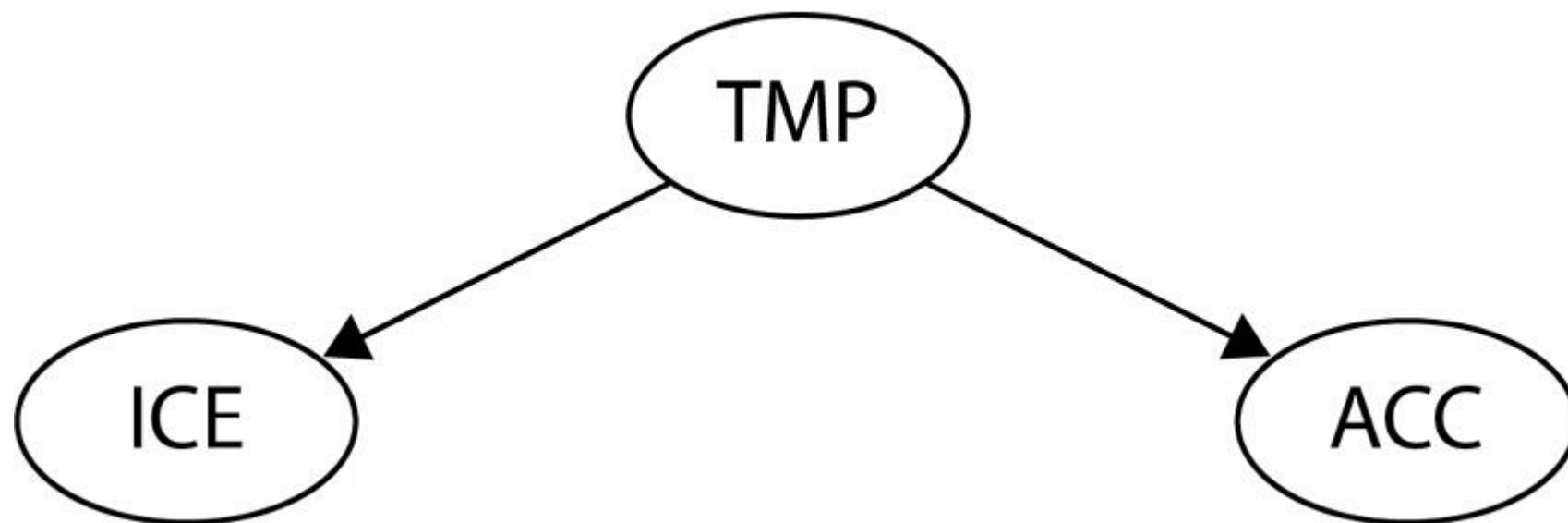
# Estimator, estimate, and estimand

- 1. Estimand:** The quantity or parameter that is intended to be estimated. It represents the true value of the parameter in the population. Suppose you're interested in the average height of adult men in a particular country. The actual average height of all adult men in that country is the estimand.
- 2. Estimate:** The approximation or value obtained from the data to estimate the estimand. This is derived from a sample and used to infer information about the population. For example, from a sample of 1,000 adult men in the same country, you calculate an average height of 5 feet 9 inches. This value (5 feet 9 inches) is your estimate of the average height (the estimand).
- 3. Estimator:** A rule, formula, or algorithm by which you derive the estimate from the data. It is a function of the sample data and is used to produce an estimate of the estimand. Here, the formula for calculating the mean (average) from a set of numbers is an estimator. When you apply this formula to your sample data, you obtain the estimate.

# Estimator, estimate, and estimand

- **Estimand:** What you want to know (the actual, often unknown, value).
- **Estimate:** What you got from your sample data.
- **Estimator:** How you got it (the method or formula used).

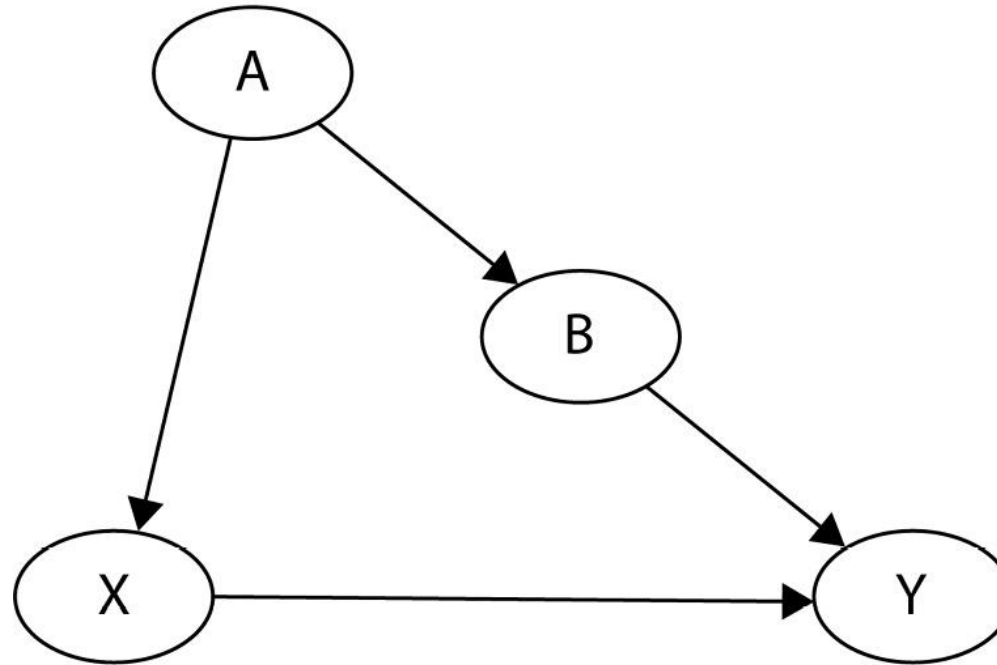
# Adjustment



$$ACC \sim ICE + TMP$$

$$P(ACC|do(ICE)) = \sum_{tmp} P(ACC|ICE, TMP)P(TMP)$$

# Back door criterion

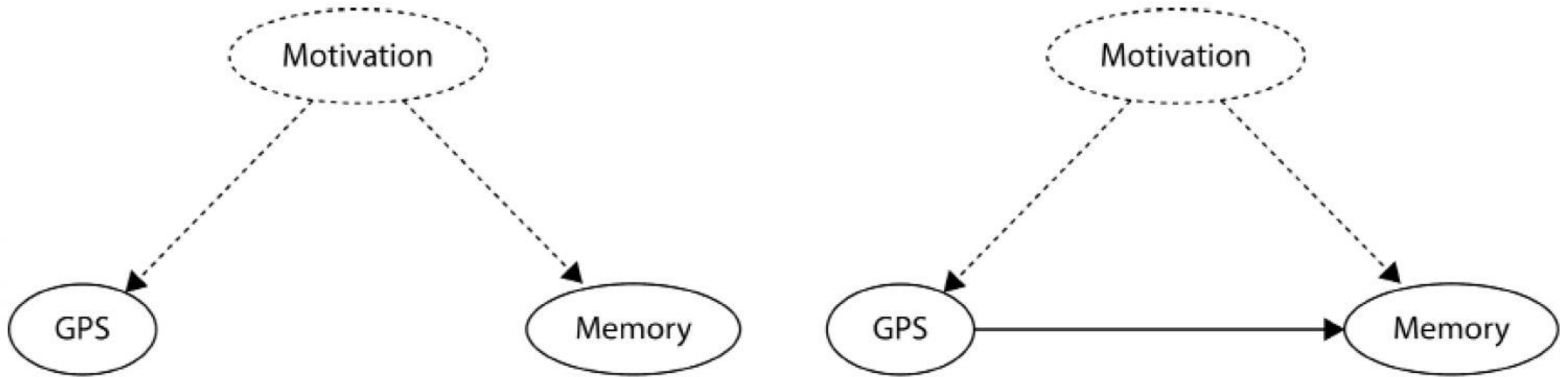


$$\begin{aligned}P(Y = y | do(X = x)) &= \sum_a P(Y = y | X = x, A = a) P(A = a) \\ &= \sum_b P(Y = y | X = x, B = b) P(B = b)\end{aligned}$$

We can estimate effect even if one of A, B is unobserved!



# Front door criterion and mediation



**Observation:** People that use GPS more have less good memory

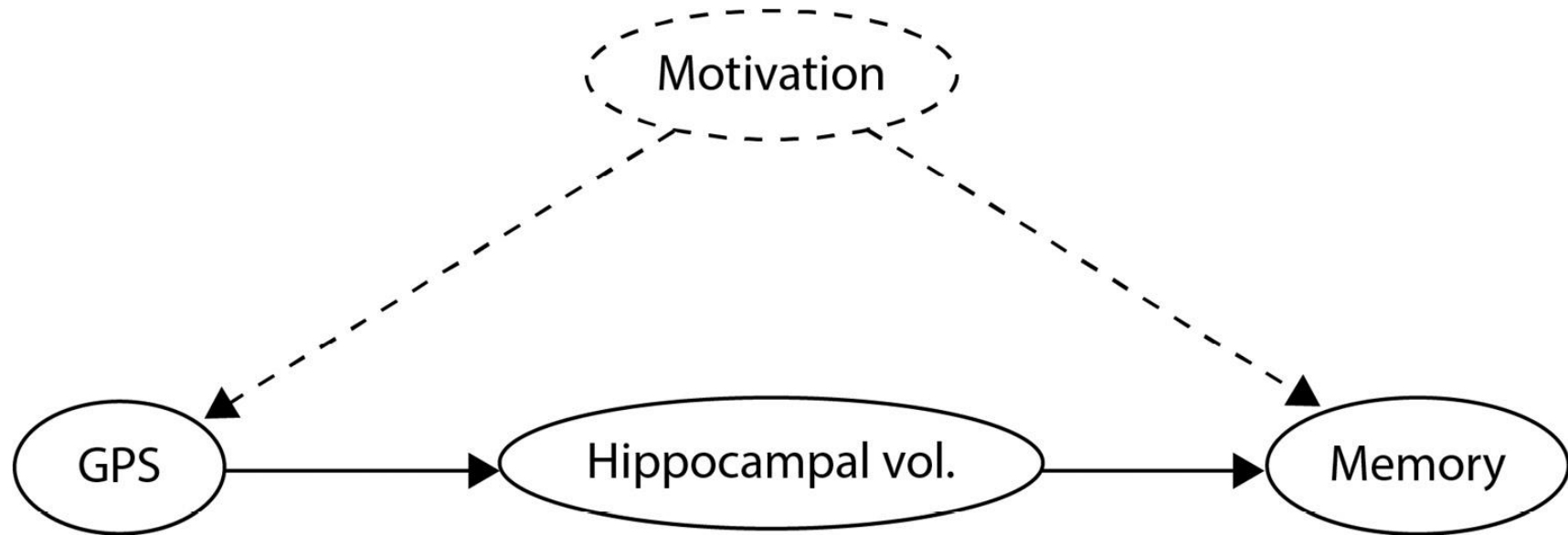
**Hypothesis 1:** Usage of GPS precludes memory development

**Hypothesis 2:** There is a common (unobserved) factor that affects both GPS usage and memory

# Mediation

- The influence of one variable  $X$  on another  $Y$  is *mediated* by a third variable,  $Z$  (or a set of variables,  $\mathbf{Z}$ ), when at least one path from  $X$  to  $Y$  goes through  $Z$ .
- $Z$  *fully mediates* the relationship between  $X$  and  $Y$  when the only path from  $X$  to  $Y$  goes through  $Z$ .
- If there are paths from  $X$  to  $Y$  that do not pass through  $Z$ , the mediation is *partial*.

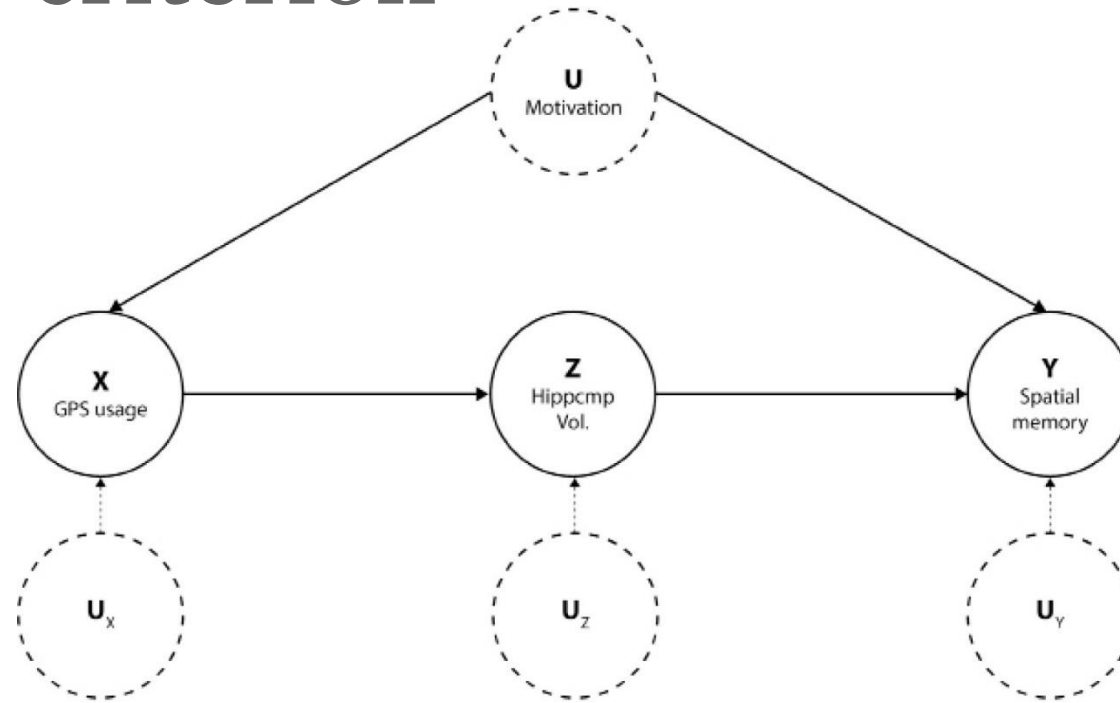
# Front door criterion



- We assume that hippocampal volume fully mediates the effects of GPS usage on a decline in spatial memory.
- The second important assumption we make is that motivation can only affect *hippocampal volume indirectly through GPS usage*.

If motivation would be able to influence hippocampal volume *directly*, front-door would be of no help. Luckily enough, the assumption that motivation cannot directly change the volume of the hippocampus seems reasonable (though perhaps you could argue against it!).

# Front door criterion



$$P(Y = y | do(X = x)) = \sum_z P(Z = z | X = x) \sum_{x'} P(Y = y | X = x', Z = z) P(X = x')$$

- Fit a model,  $Z \sim X$
- Fit a model,  $Y \sim Z + X$
- Multiply the coefficients from model 1 and model 2

# Do-calculus

- *Rule 1:* When an observation can be ignored:

$$P(Y = y | do(X = x), Z = z, W = w) = P(Y = y | do(X = x), W = w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_X}$$

- *Rule 2:* When intervention can be treated as an observation:

$$P(Y = y | do(X = x), do(Z = z), W = w) = P(Y = y | do(X = x), Z = z, W = w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_{XZ}}$$

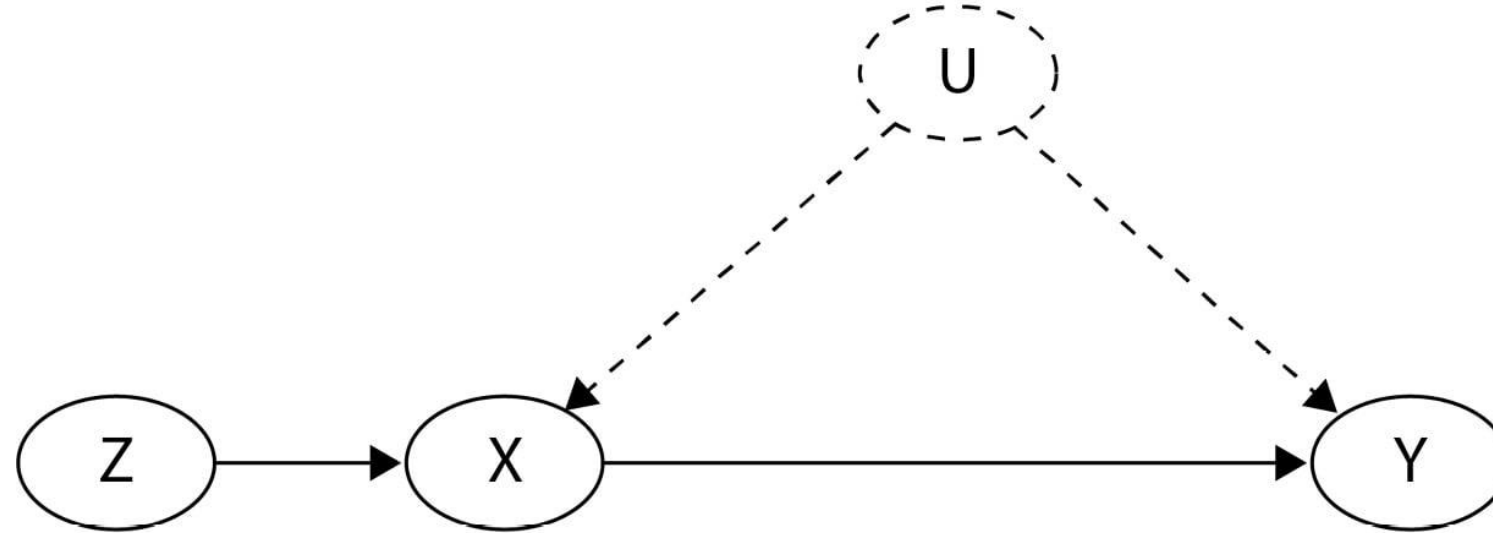
- *Rule 3:* When intervention can be ignored:

$$P(Y = y | do(X = x), do(Z = z), W = w) = P(Y = y | do(X = x), W = w) \text{ if } (Y \perp\!\!\!\perp Z | X, W)_{G_{XZ(W)}}$$

Given a DAG  $G$ , we can say that  $G_{\bar{X}}$  is a modification of  $G$ , where we removed all the *incoming* edges to the node  $X$ . We will call  $G_{\bar{X}}$  a modification of  $G$ , where we removed all the *outgoing* edges from the node  $X$ . For example, will denote a DAG,  $G_{\bar{X}\bar{Z}}$ , where we removed all the incoming edges to the node  $X$  and all the outgoing edges from the node  $Z$ .

- **Good news:** do-calculus exists and is complete
- **Not so good news:** it can be quite incomprehensible and takes a while to learn
- **Super good news:** now there are codes (DoWhy) that allow us to apply it

# Instrumental Variables



We're interested in estimating the causal effect of X on Y .

- Cannot use the back-door criterion here because U is unobserved.
- Cannot use the front-door criterion because there's no mediator between X and Y .

Instrumental Variables: require a special variable called an *instrument*, Z, to be present in a graph. An *instrument* needs to meet the following three conditions:

- The instrument, Z, is associated with X
- The instrument, Z, doesn't affect Y in any way except through X
- There are no common causes of Z and Y

We want to study the effect of education (years of schooling) on earnings. However, the level of education might be influenced by many factors like family background, which also affect earnings. This correlation between the unobserved factors (like family background) and education can bias the results if you simply run a linear regression of earnings on education.

We need an instrument that is correlated with education but does not directly affect earnings except through education. Let's say we choose "proximity to college" as instrument.

### **Two-Stage Least Squares (2SLS) Regression:**

1. Regress the potentially endogenous variable (education) on the instrument (proximity to college). This predicts the values of education that are not influenced by the unobserved confounders.

$$\text{Education} = \alpha + \beta * \text{ProximityToCollege} + \varepsilon$$

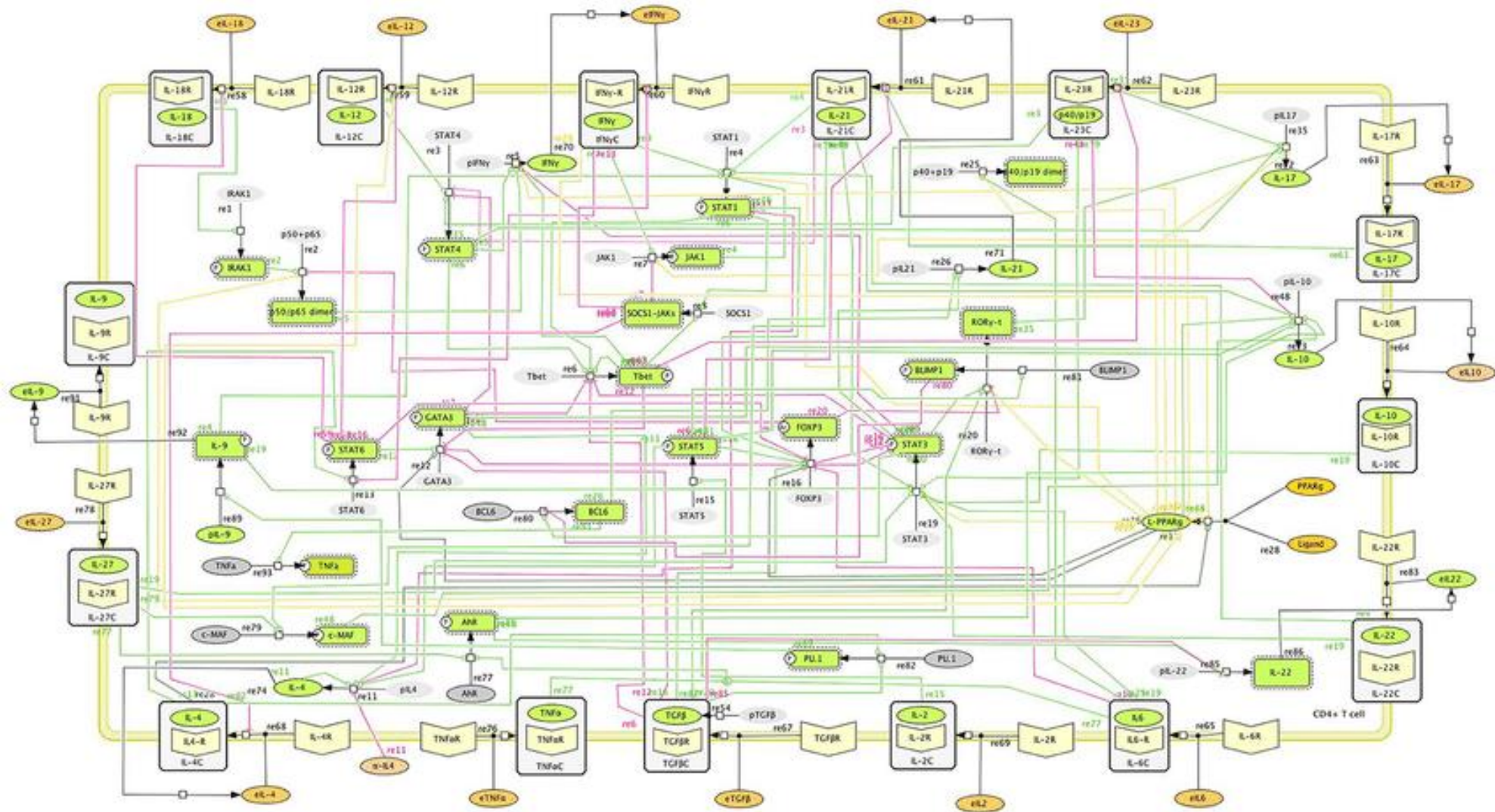
2. Regress the outcome (earnings) on the predicted values of education from the first stage

$$\text{Earnings} = \gamma + \delta * \text{PredictedEducation} + \zeta$$

3. The coefficient  $\delta$  on PredictedEducation in the second stage gives the estimated causal effect of education on earnings. This helps to isolate the variation in education that is independent of the unobserved confounders that also affect earnings.



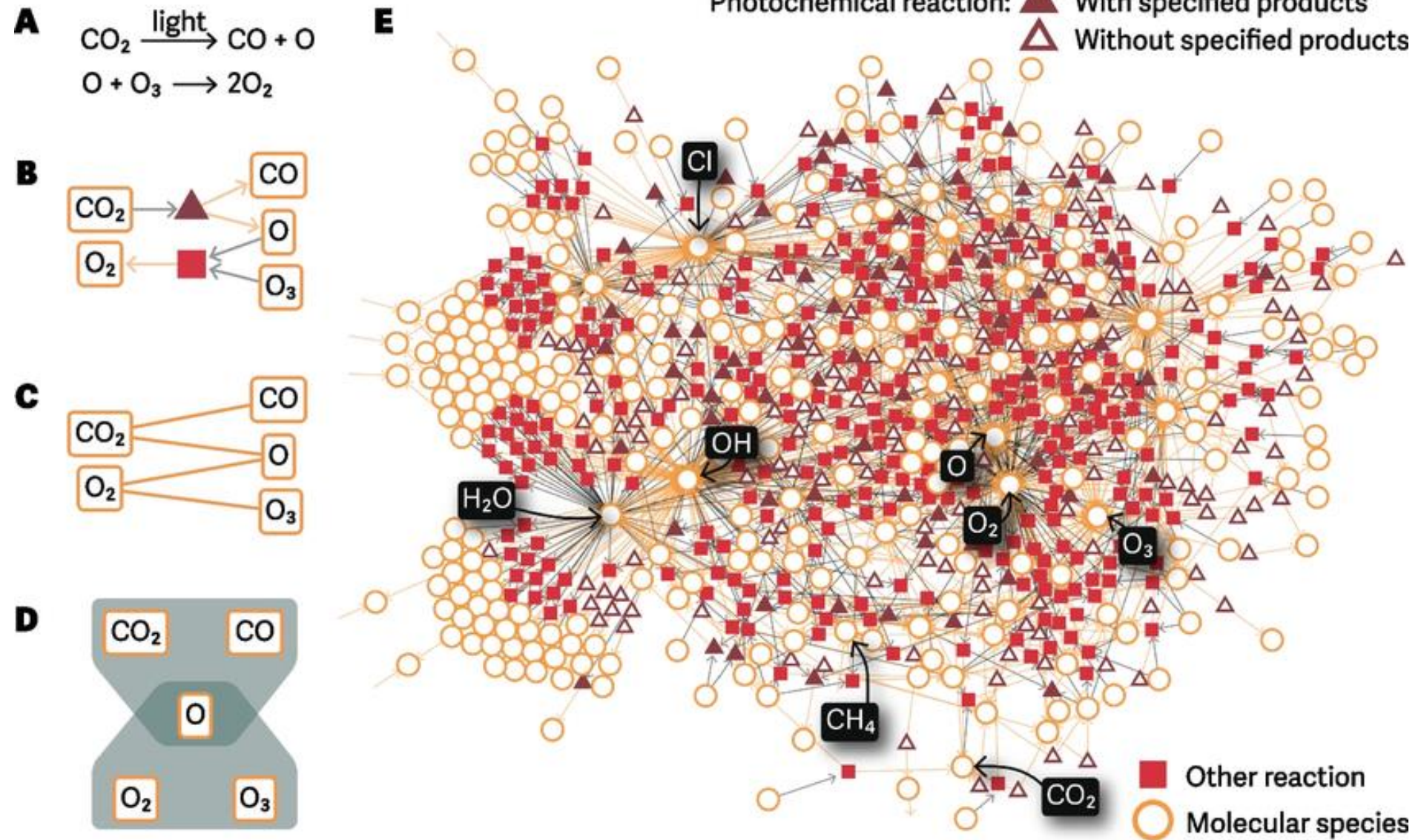
# Biochemical reaction networks



[https://www.researchgate.net/figure/Main-intracellular-differentiation-pathways-of-a-single-CD4-T-cell-Systems-Biology\\_fig2\\_267753905](https://www.researchgate.net/figure/Main-intracellular-differentiation-pathways-of-a-single-CD4-T-cell-Systems-Biology_fig2_267753905)

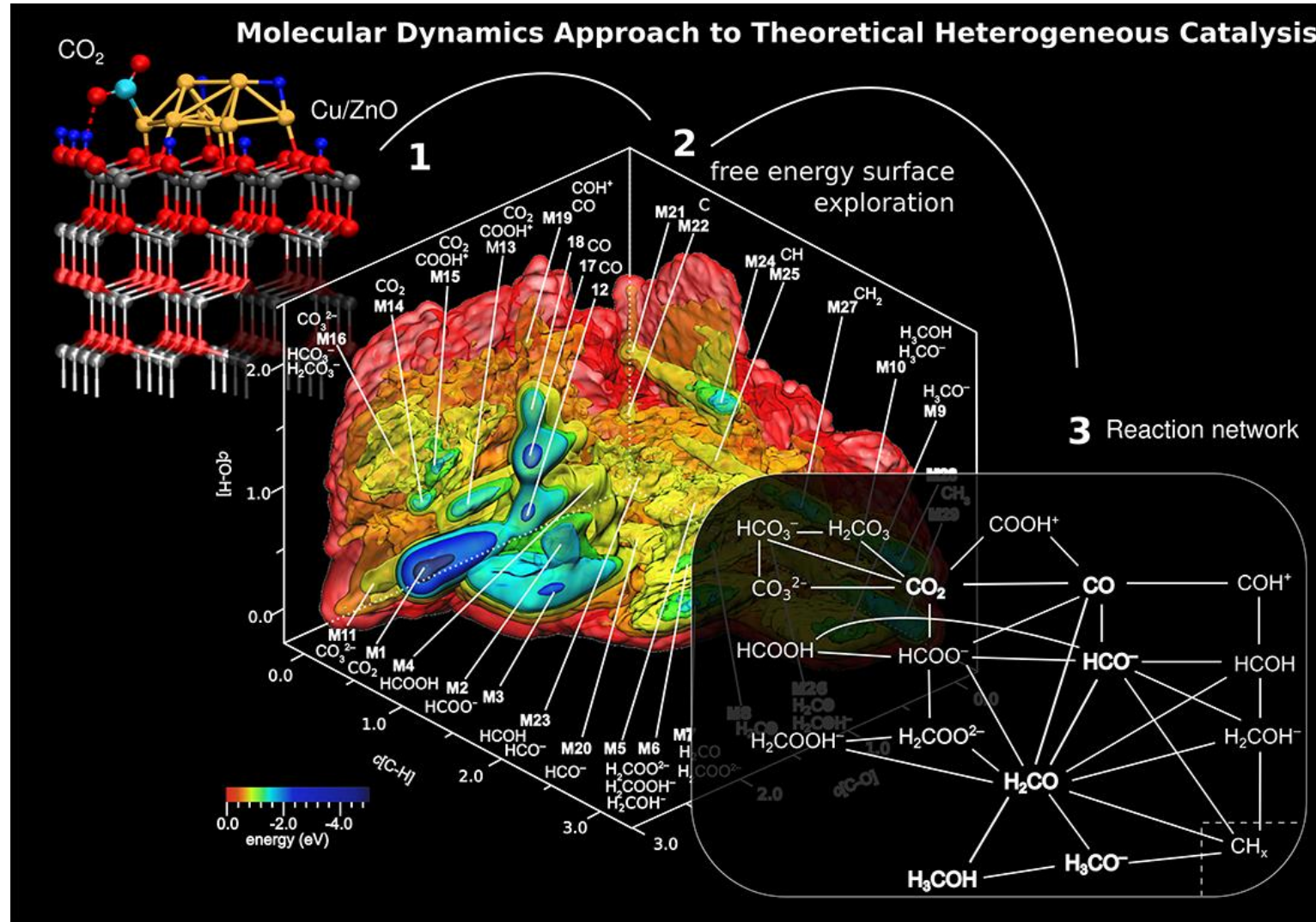


# Atmospheric reaction networks



[https://www.researchgate.net/figure/Network-structure-of-Earths-atmospheric-reaction-system-Panel-A-shows-a-minimal\\_fig4\\_368305699](https://www.researchgate.net/figure/Network-structure-of-Earths-atmospheric-reaction-system-Panel-A-shows-a-minimal_fig4_368305699)

<https://www.gauss-centre.eu/results/materials-science-and-chemistry/theoretical-heterogeneous-catalysis-from-advanced-ab-initio-molecular-dynamics-simulations>



# Why causal inference is difficult?

- unmeasured confounders
- measurement error, or discretization of data
- mixtures of different causal structures in the sample
- feedback
- reversibility
- the existence of a number of models that fit the data equally well
- an enormous search space
- low power of tests of independence conditional on large sets of variables
- selection bias
- missing values
- sampling error
- complicated and dense causal relations among sets of variables,
- complicated probability distributions





# Counterfactuals

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, Pearl (2009) describes a two-model scenario where the average (interventional) causal effect for a drug is equal to 0. This is true for both models.

The difference is that in one model, no patient is affected by the drug, while in the second model, all patients are affected by the drug – a pretty striking difference. As you can expect, counterfactual outcomes for both models differ (for more details, check out Pearl (2009, pp. 33–38)).

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). This fact further underlines the asymmetry between the two and, at the same time, reveals the deep, beautiful structure of Pearlian causal formalism.

# Computing counterfactuals

Judea Pearl and colleagues (Pearl, Glymour and Jewell, 2016) 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

# Causal ML vs. active learning

- Bottou et al. (2013) proposed 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 rule

