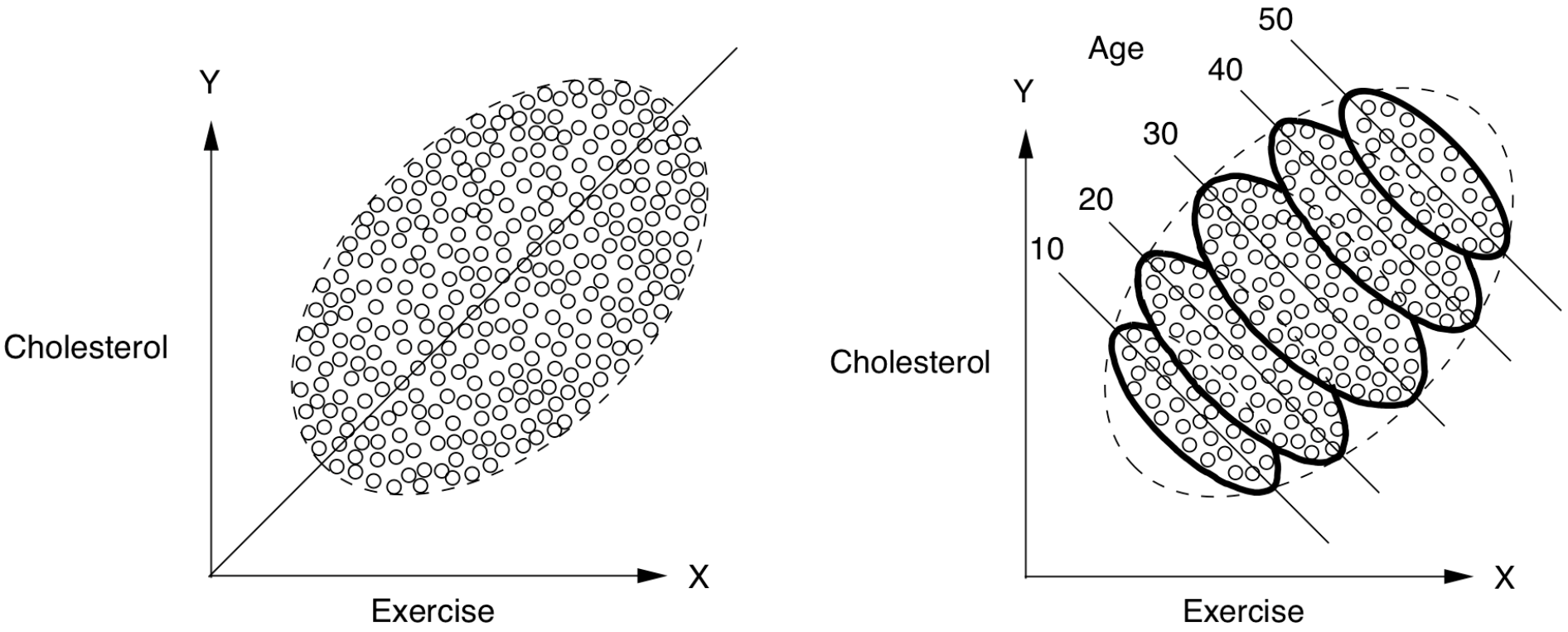
THE UNIVERSITY of TENNESSEE KNOXVILLE

Lecture 35: Causality

Sergei V. Kalinin

Simpson paradox

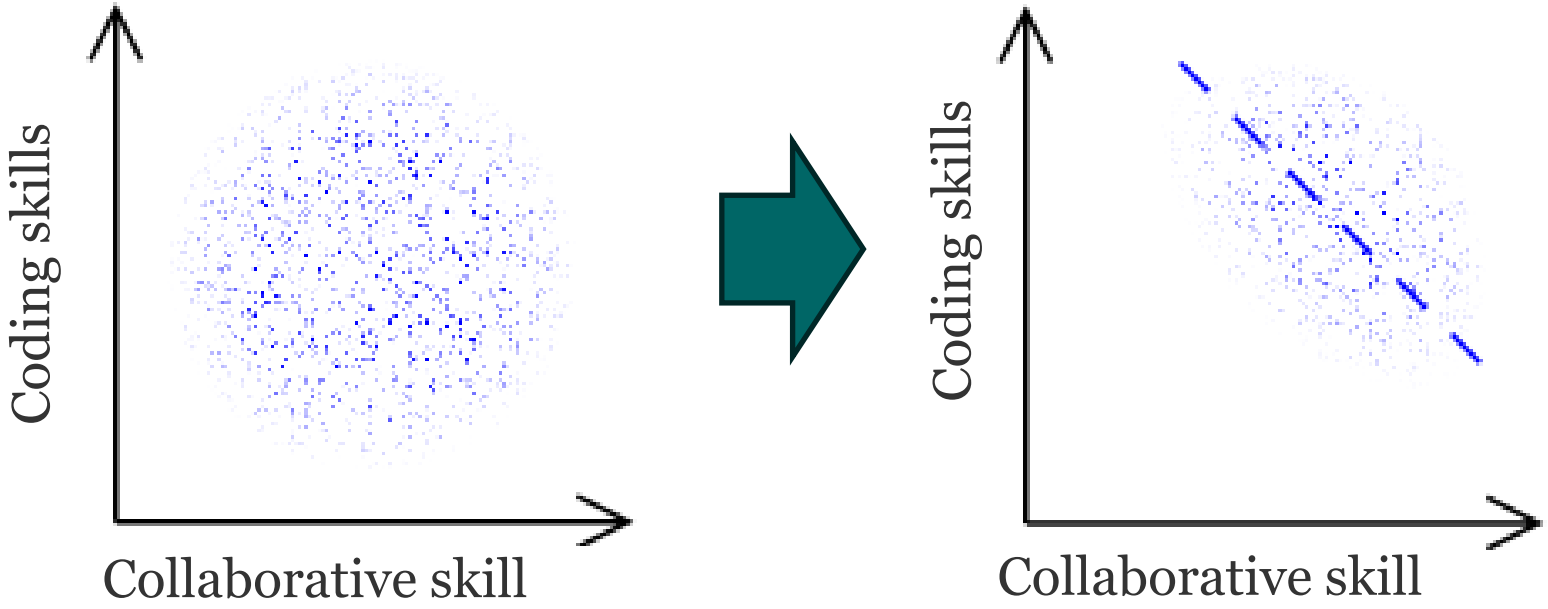


Exercise is helpful in every age group but harmful for a typical person.

Is exercise helpful or not?

From J. Pearl presentation, Why - 19

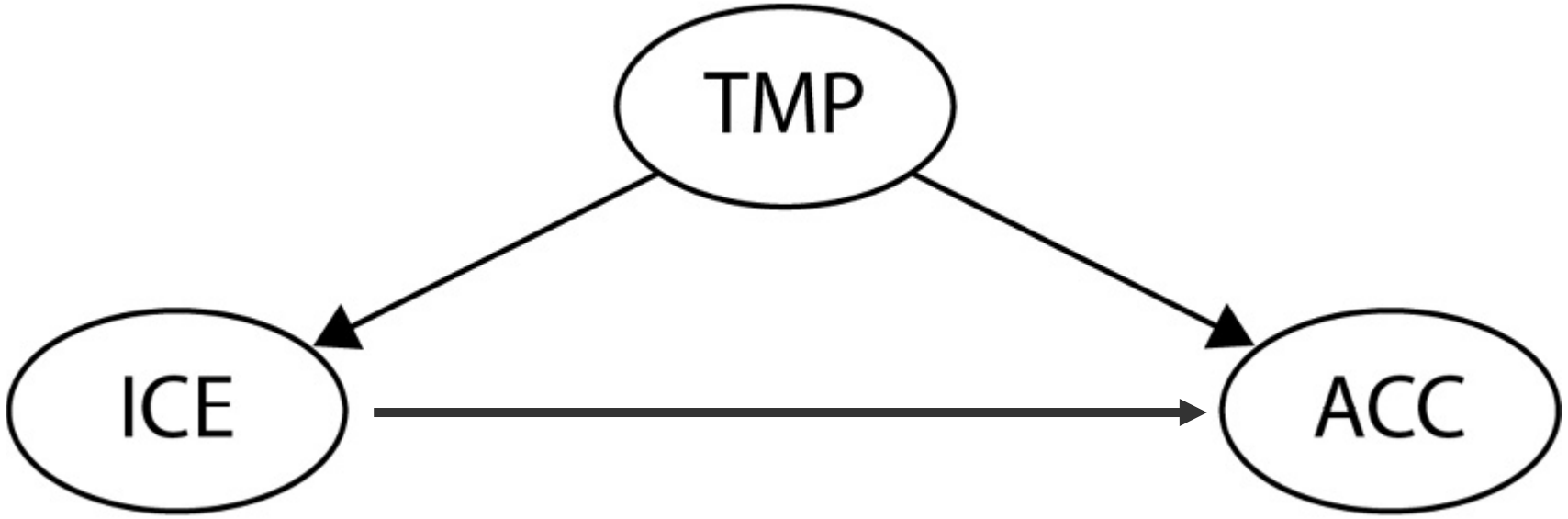
Colliders and Berkson paradox



- 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.

https://en.wikipedia.org/wiki/Berkson%27s_paradox From A. Molak, Causal Inference and Discovery in Python

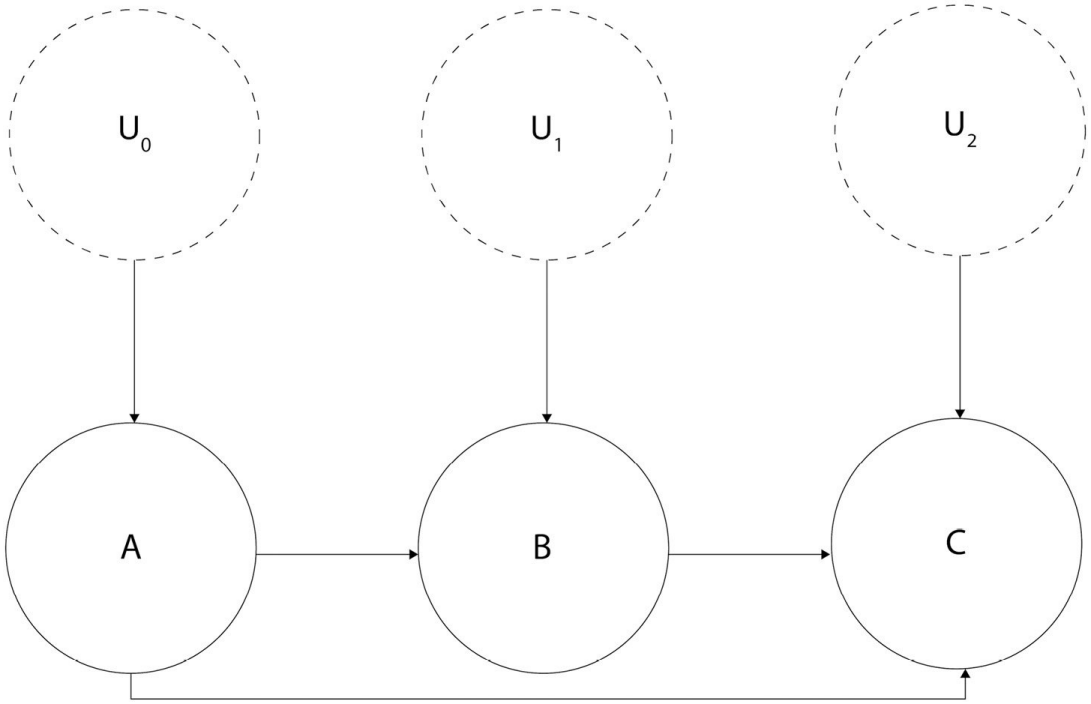
How can we even approach such problems?



- 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!

From A. Molak, Causal Inference and Discovery in Python

Causal graphs



$$A := f_A(U_0)$$

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

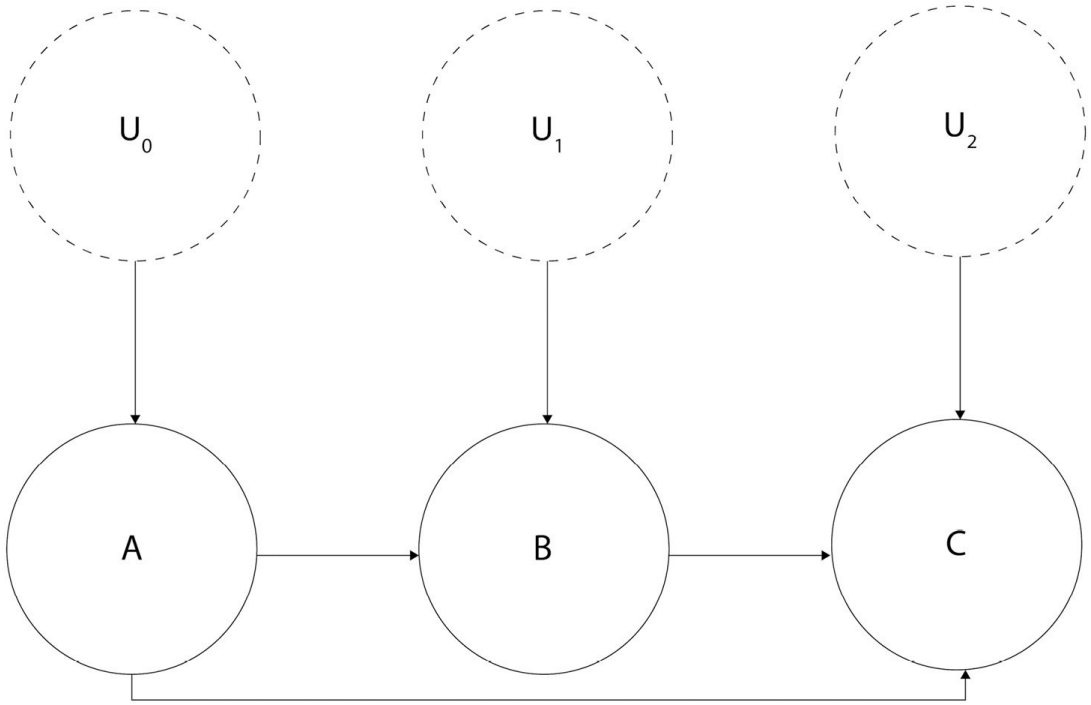
$$C := f_C(A, B, U_2)$$

- 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).

From A. Molak, Causal Inference and Discovery in Python

Properties of do - operator



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

From A. Molak, Causal Inference and Discovery in Python

How can we learn causality

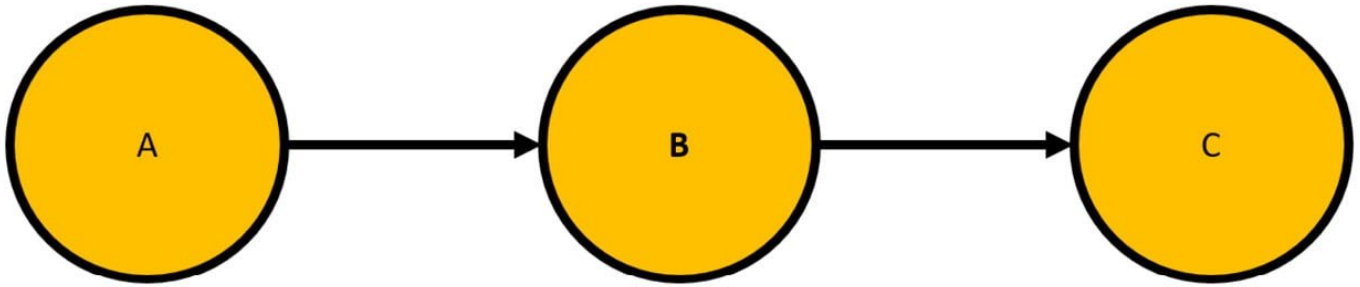
- **Causal discovery** and **causal structure learning** are umbrella terms for various kinds of methods used to uncover causal structure from observational or interventional data.

- **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.

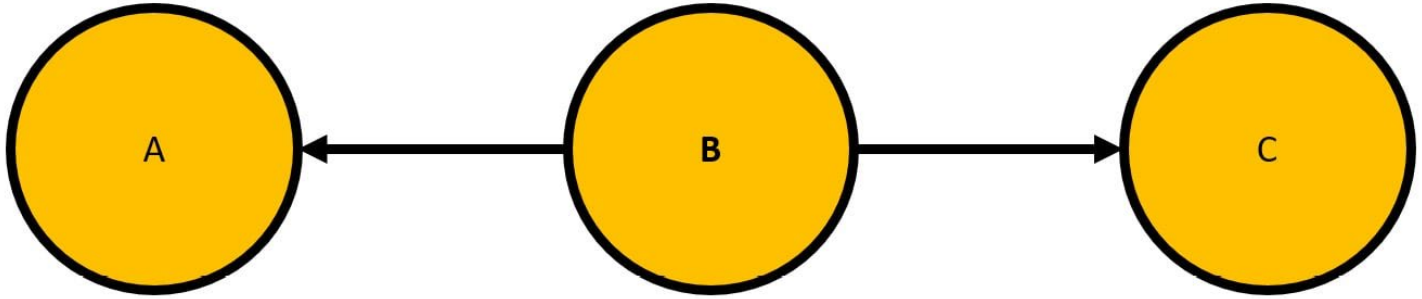
From A. Molak, Causal Inference and Discovery in Python

Chains and forks



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

Chain 🔗

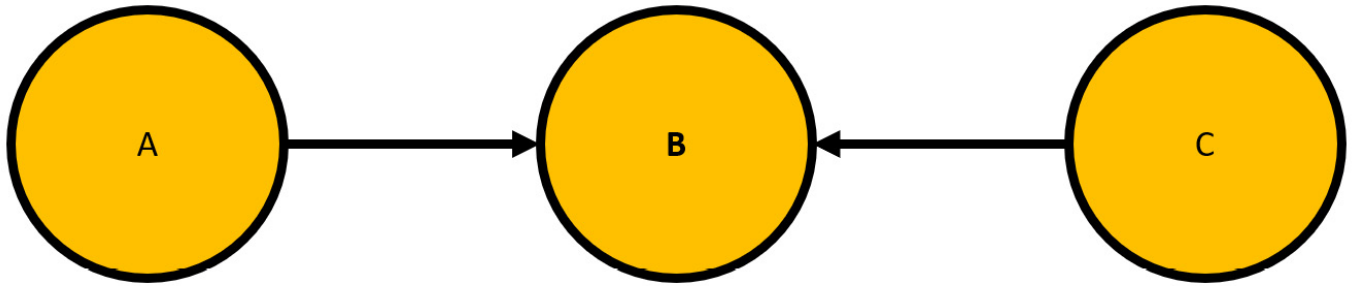


$$A \perp\!\!\!\perp_G C \mid B$$

Fork 🍴

From A. Molak, Causal Inference and Discovery in Python

Colliders



$$A \perp\!\!\!\perp C$$

$$A \not\perp\!\!\!\perp C \mid B$$

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

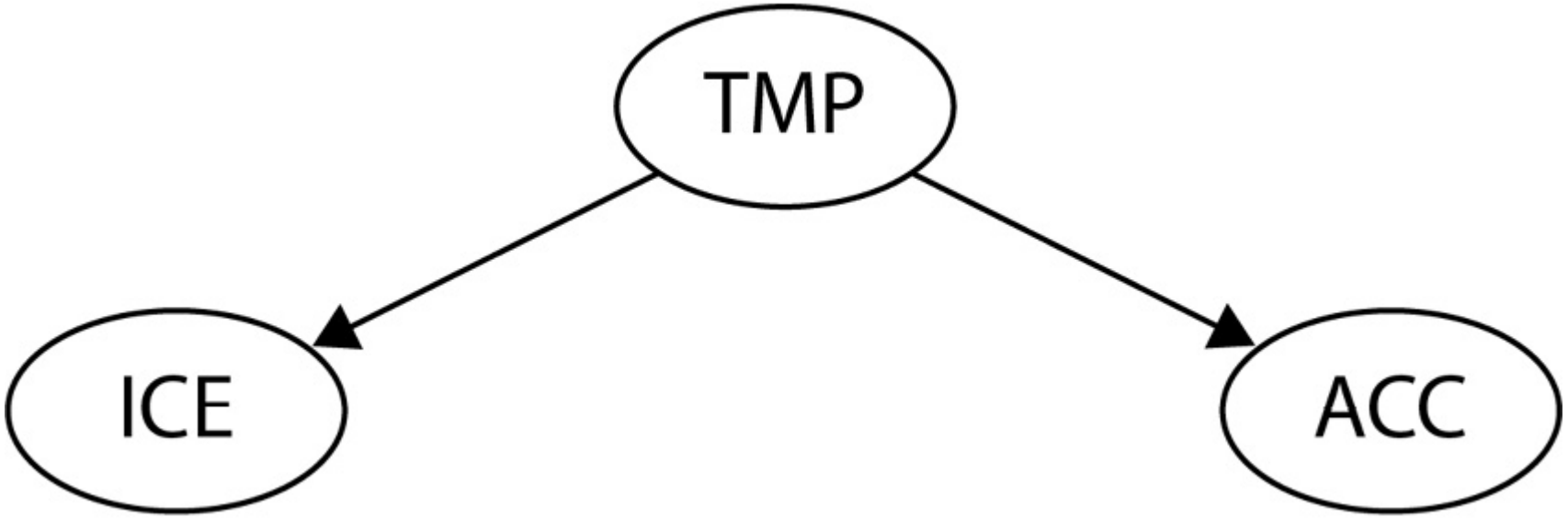
Although A and C are unconditionally independent (there's no correlation between them as they randomly and independently generate integers), they become correlated when we observe C !

From A. Molak, Causal Inference and Discovery in Python

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

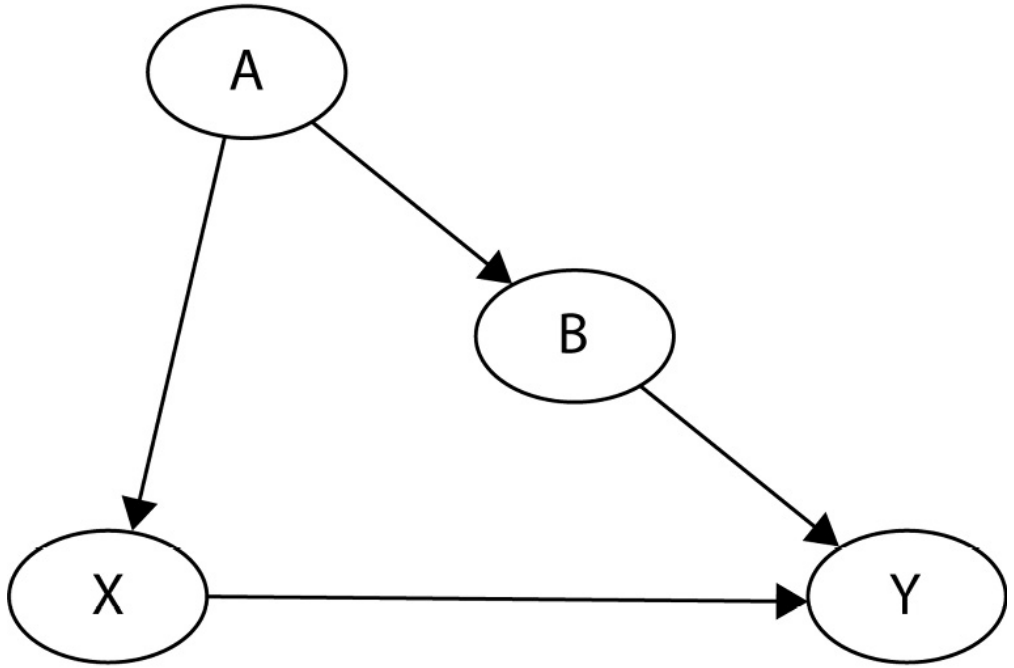


$$ACC \sim ICE + TMP$$

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

From A. Molak, Causal Inference and Discovery in Python

Back door criterion



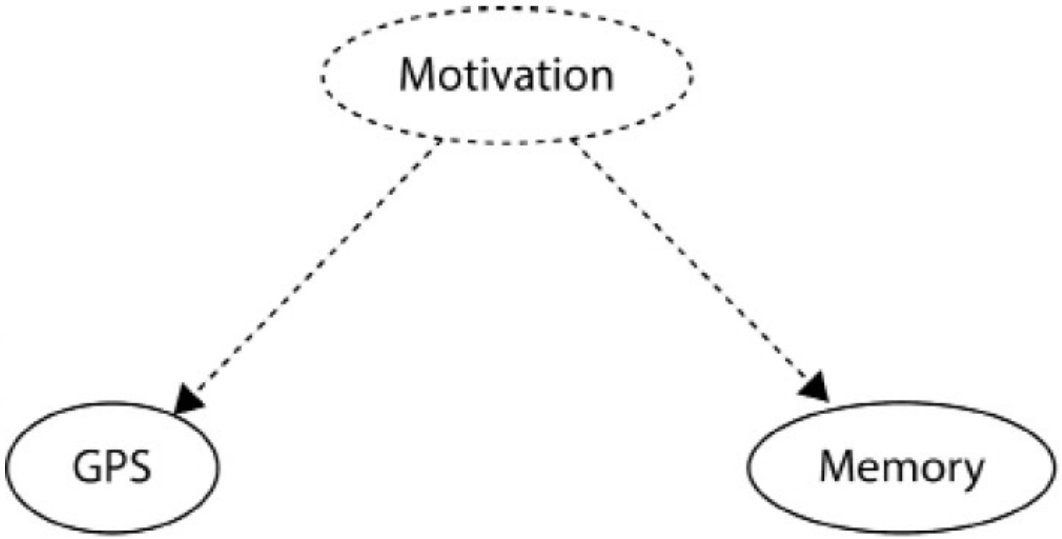
$$P(Y = y|do(X = x)) = \sum_a P(Y = y|X = x, A = a)P(A = a)$$

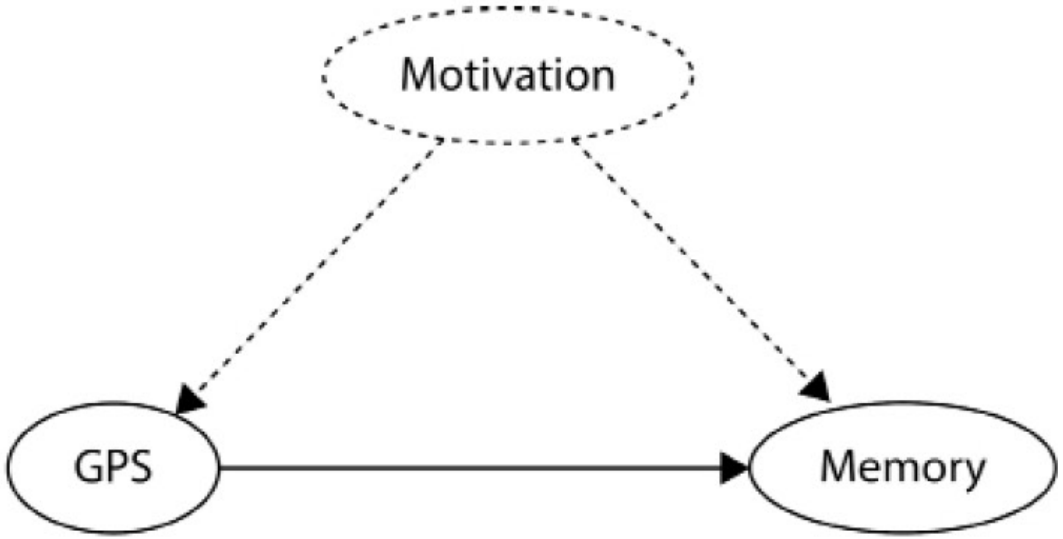
$$= \sum_b P(Y = y|X = x, B = b)P(B = b)$$

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

From A. Molak, Causal Inference and Discovery in Python

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

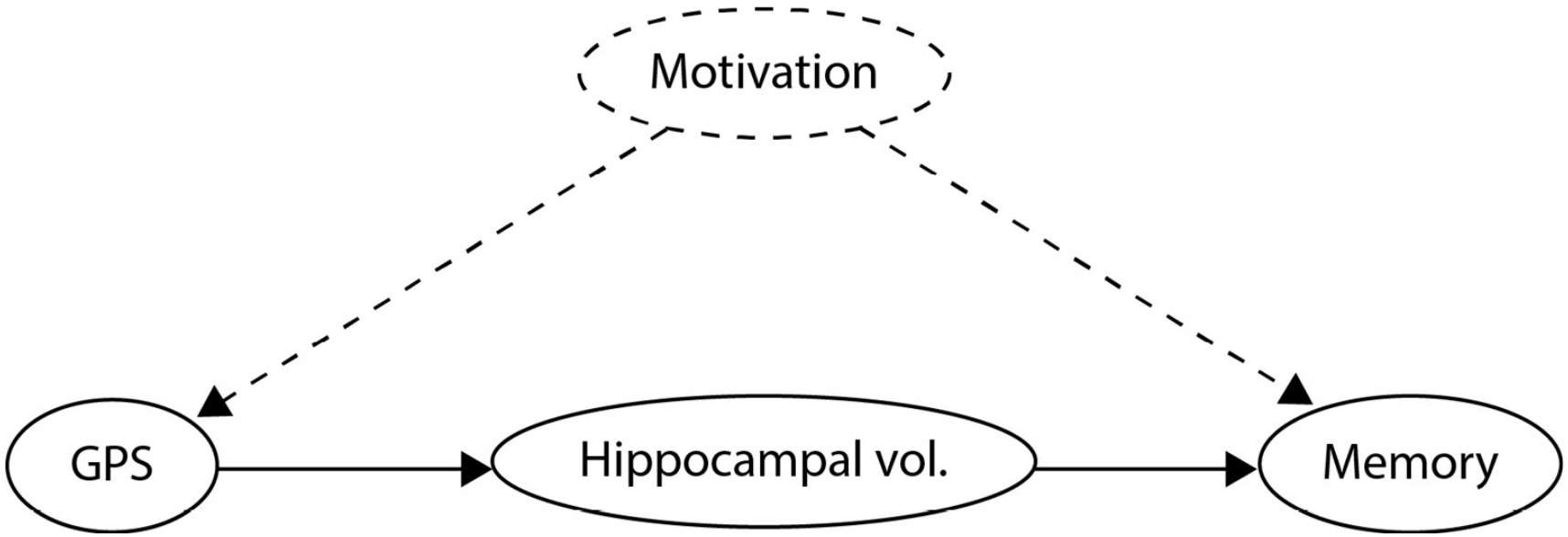
From A. Molak, Causal Inference and Discovery in Python

Mediation

- The influence of one variable X on another Y is *mediated* by a third variable, Z (or a set of variables, **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*.

From A. Molak, Causal Inference and Discovery in Python

Front door criterion

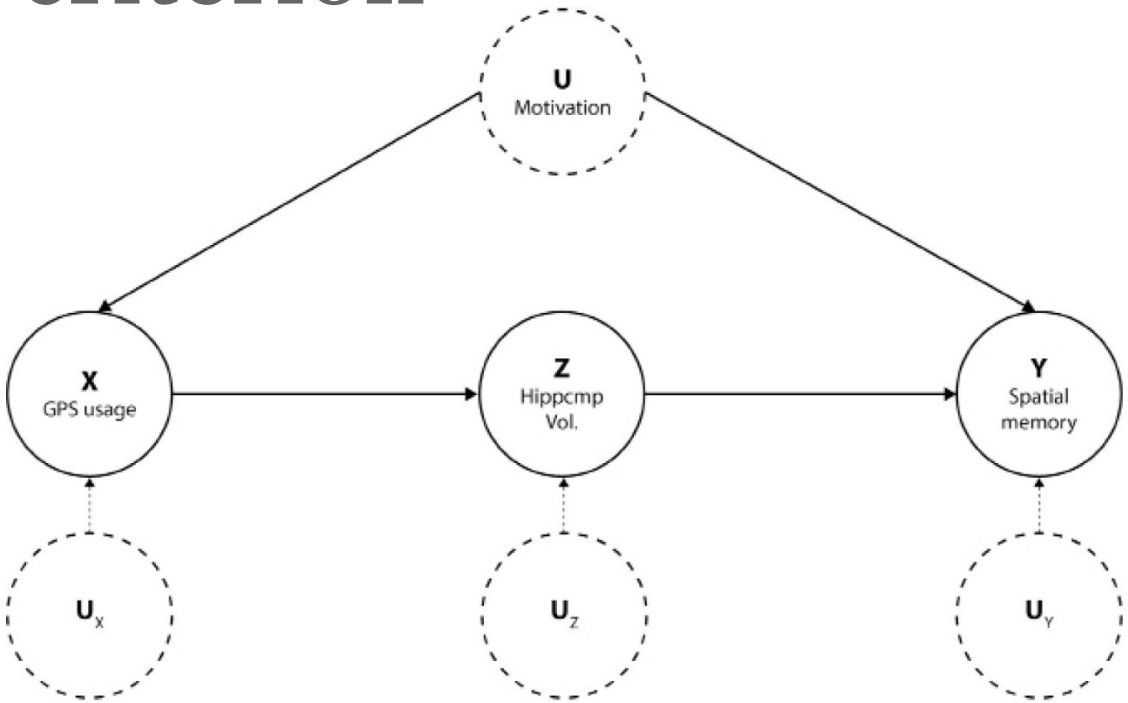


- We assume that hippocampal volume fully mediates the effects of GPS usage on a decline in spatial memory.
- The second important assumption we make is that motivation can only affect *hippocampal volume indirectly through GPS usage*.

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!).

From A. Molak, Causal Inference and Discovery in Python

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

From A. Molak, Causal Inference and Discovery in Python

Do-calculus

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

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

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

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

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

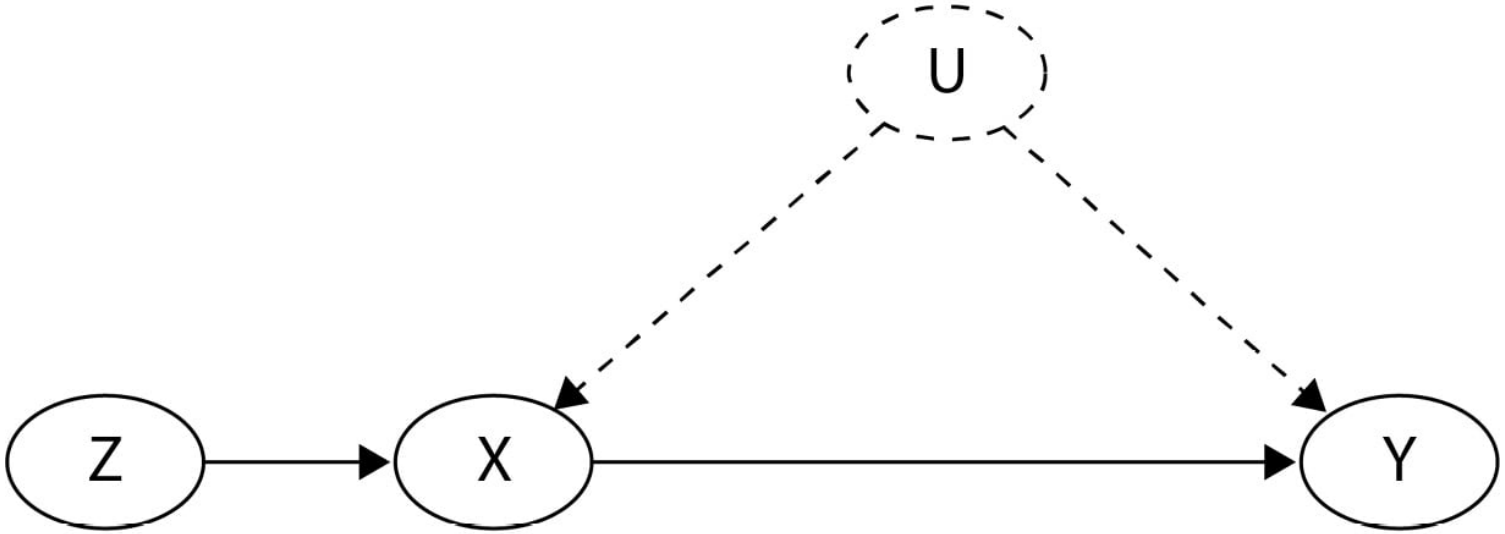
$$P(Y = y|do(X = x), do(Z = z), W = w) = P(Y = y|do(X = x), W = w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X},\overline{Z(W)}}}$$

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

From A. Molak, Causal Inference and Discovery in Python

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

From A. Molak, Causal Inference and Discovery in Python

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.

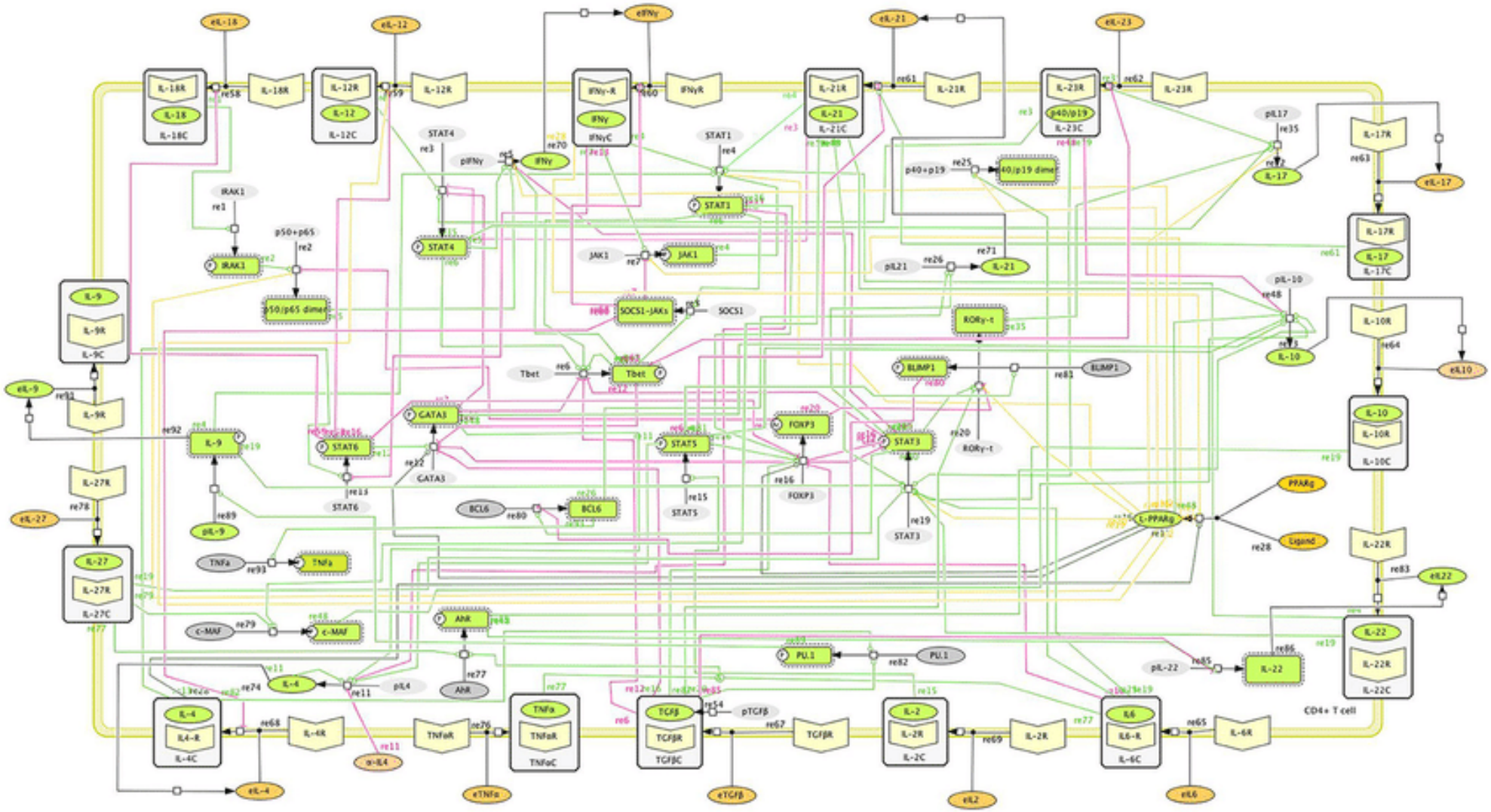
$$\text{Education} = \alpha + \beta * \text{ProximityToCollege} + \varepsilon$$

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

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

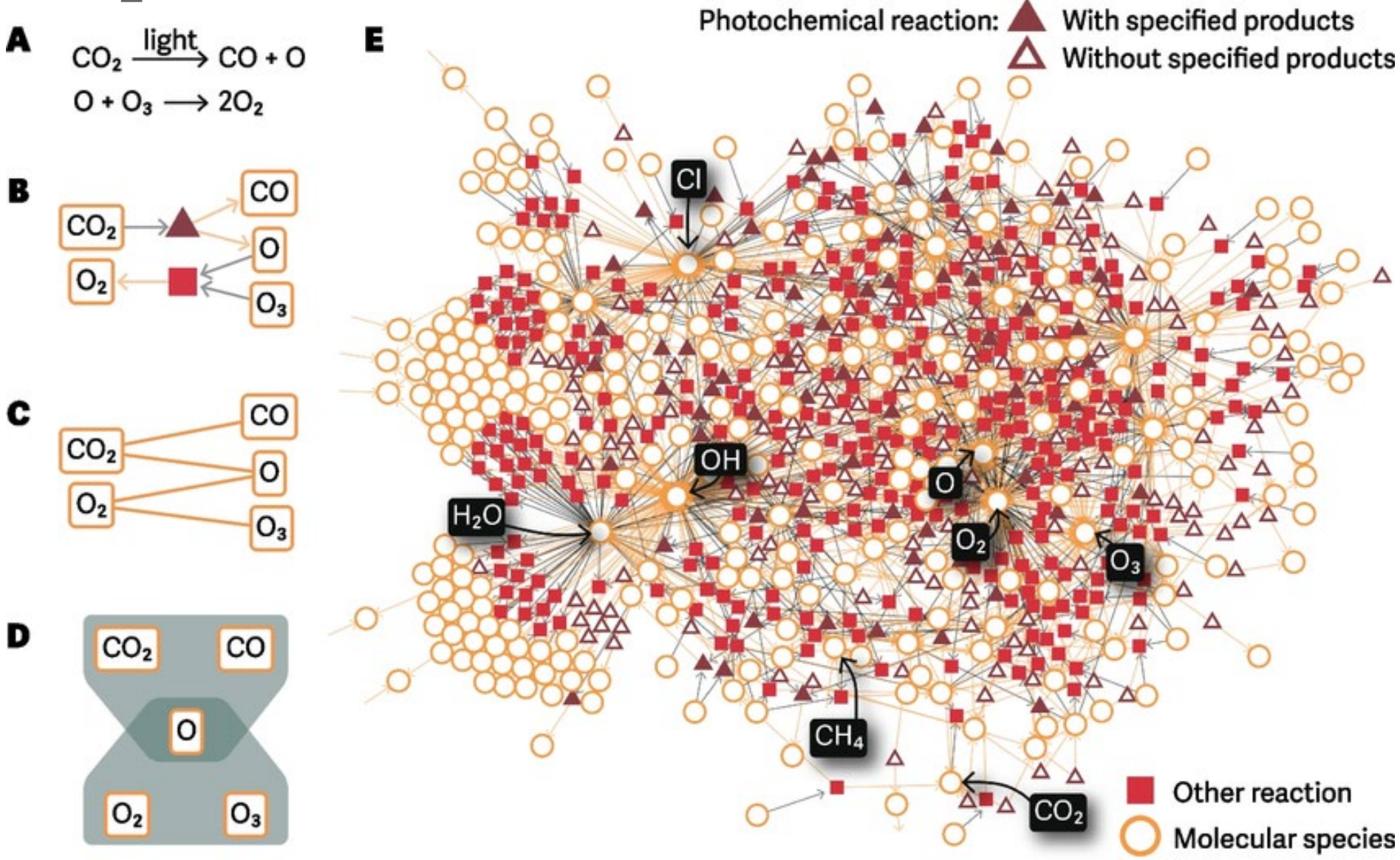
3. The coefficient δ on PredictedEducation in the second stage gives the estimated causal effect of education on earnings. This helps to isolate the variation in education that is independent of the unobserved confounders that also affect earnings.

Biochemical reaction networks



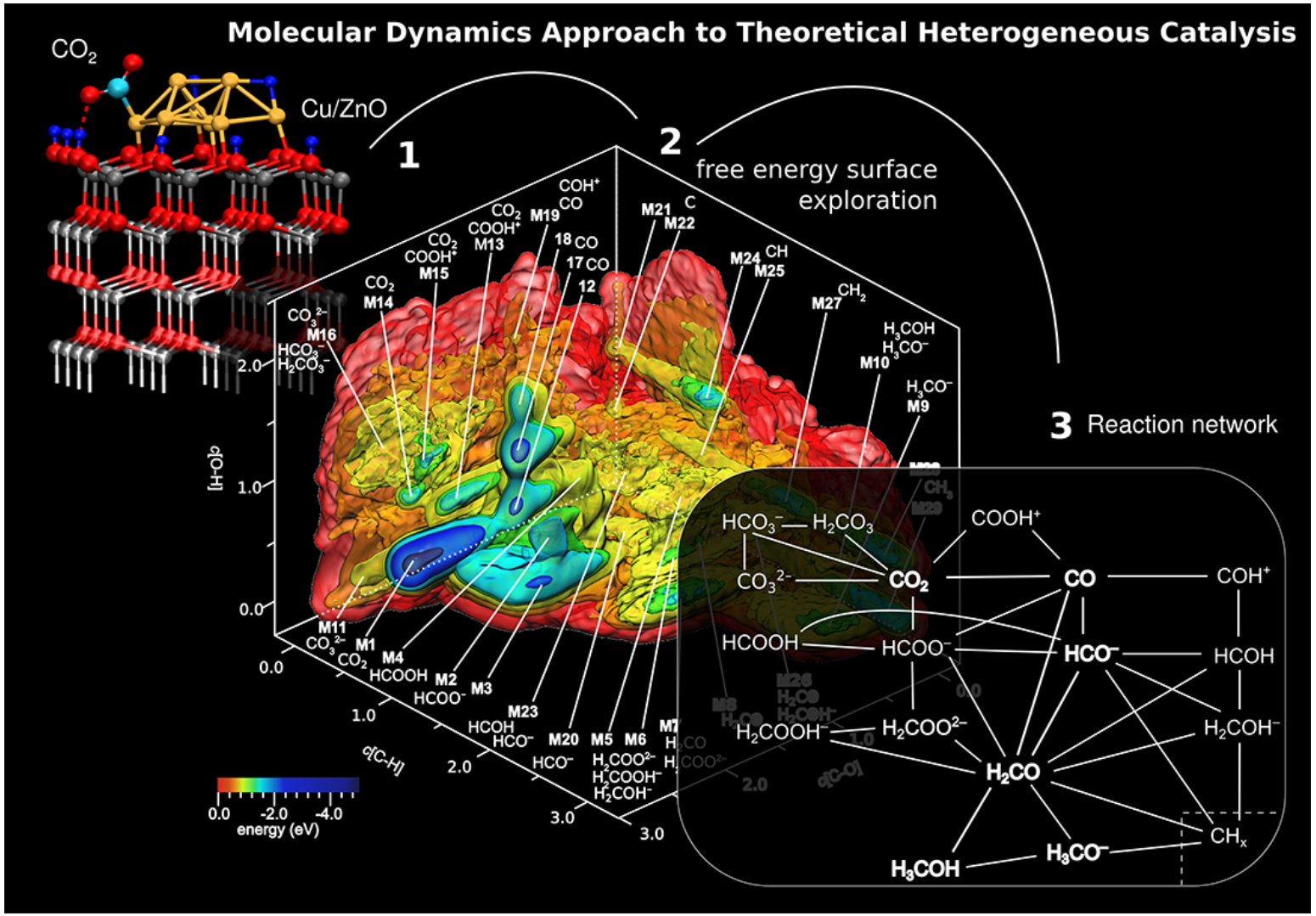
https://www.researchgate.net/figure/Main-intracellular-differentiation-pathways-of-a-single-CD4-T-cell-Systems-Biology_fig2_267753905

Atmospheric reaction networks



https://www.researchgate.net/figure/Network-structure-of-Earths-atmospheric-reaction-system-Panel-A-shows-a-minimal_fig4_368305699

Catalysis



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