

# Causal Inference and Counterfactual Reasoning

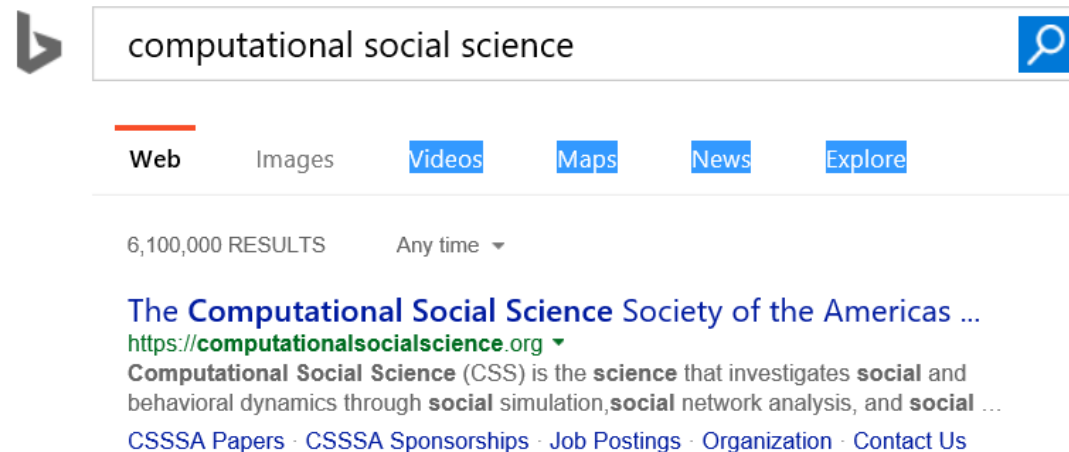
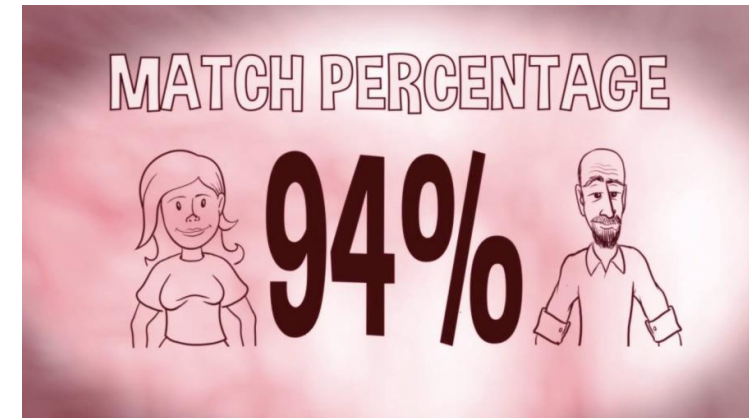
Emre Kıcıman and Amit Sharma

emrek@microsoft.com, amshar@microsoft.com

[Causal Inference and Counterfactual Reasoning](#) at Microsoft Research

# Predictive systems are impacting our lives

Customers Who Bought This Item Also Bought









# Why should we care about causality?

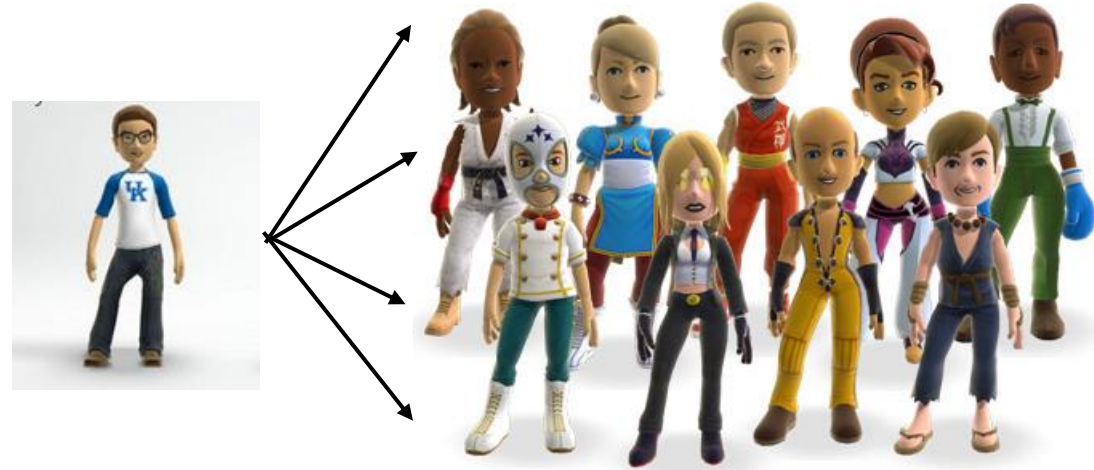
We have increasing amounts of data and highly accurate predictions.

How is causal inference useful?

1) Do prediction models guide decision-making?

# From data to prediction

Can we predict a user's future activity based on exposure to their social feed?



Use the social feed to predict a user's future activity.

- Future Activity  $\rightarrow f(\text{items in social feed}) + \epsilon$

Highly predictive model.

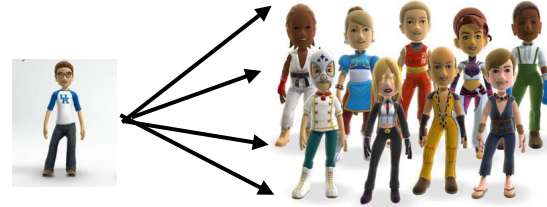
Does it mean that feeds are influencing us significantly?



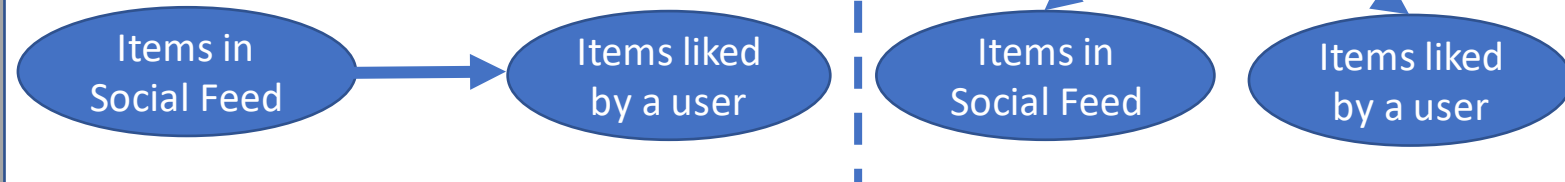
# From prediction to decision-making

Would changing what people see in the feed affect what a user likes?

Maybe, maybe not (!)



Predictability due to  
**feed influence**

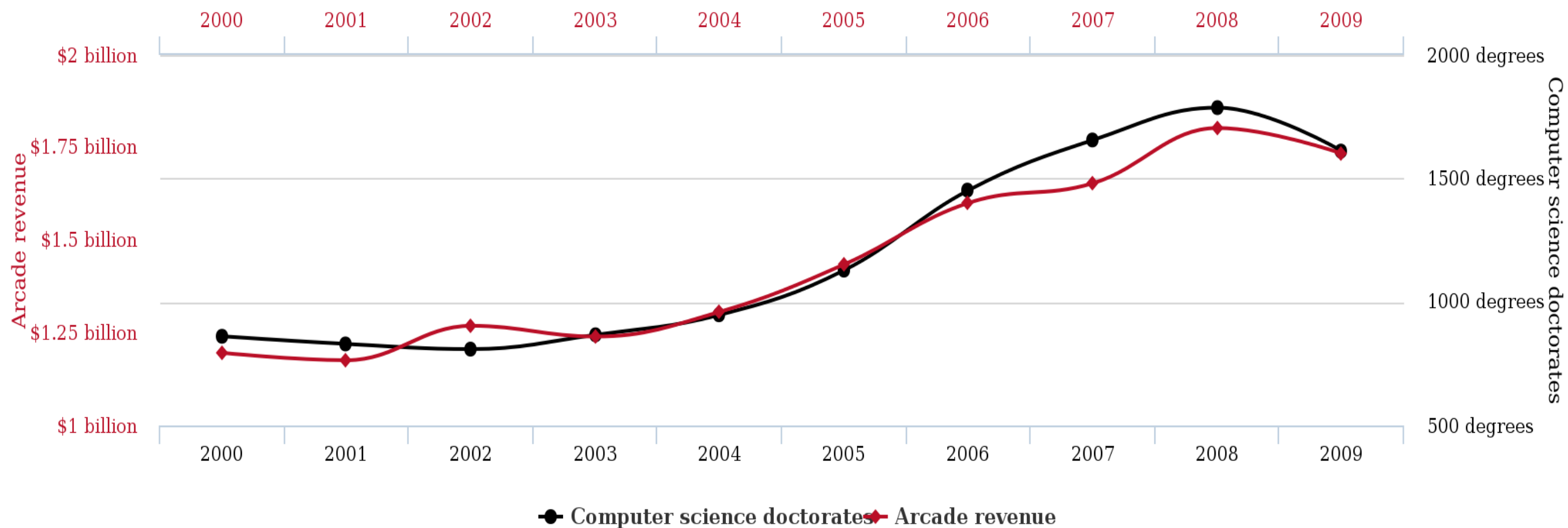


Friends' activity can predict a person's activity with high accuracy.  
But that tells us *nothing* about the effect of the social feed.



2) Will the predictions be robust tomorrow, or in new contexts?

# Total revenue generated by arcades correlates with Computer science doctorates awarded in the US

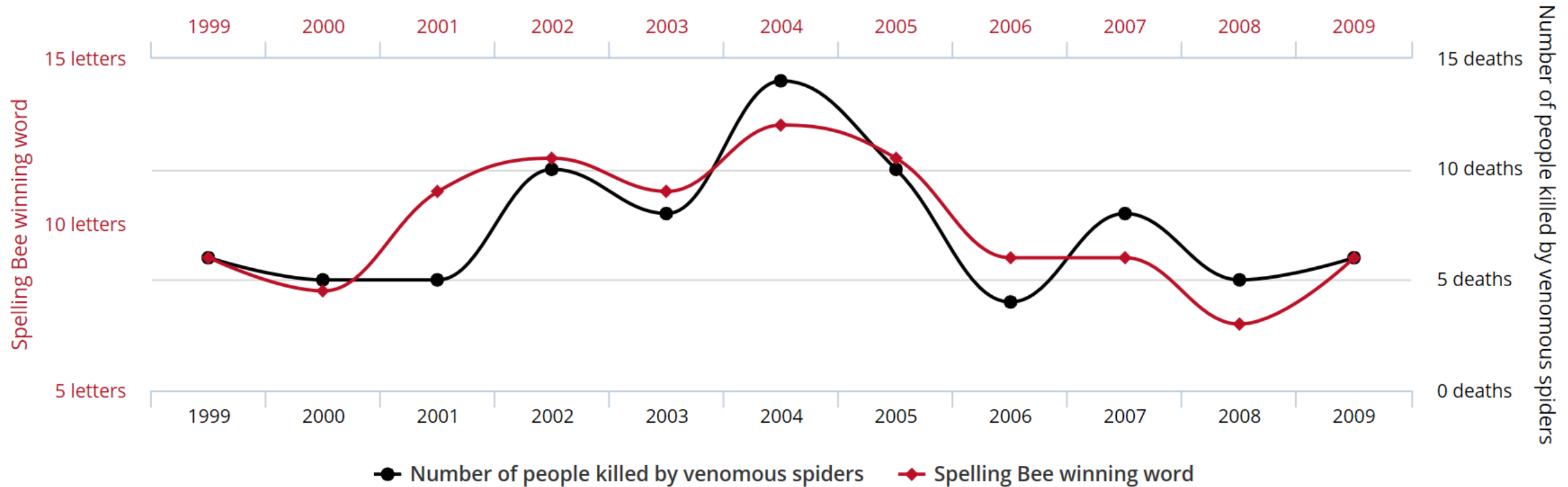


# Letters in Winning Word of Scripps National Spelling Bee

correlates with

## Number of people killed by venomous spiders

Correlation: 80.57% ( $r=0.8057$ )



tylervigen.com

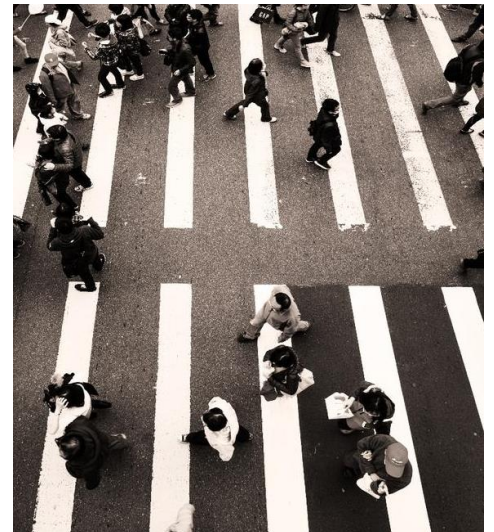
Data sources: National Spelling Bee and Centers for Disease Control & Prevention

3) What if the prediction accuracy is really high?



# Interventions change the environment

- Train/test from same distribution in supervised learning
- No such guarantee in real life!
- Problematic: Acting on a prediction changes distribution!
  - Incl. critical domains: healthcare or adversarial scenarios.
- Connections to covariate shift, domain adaptation [Mansour et al. 2009, Ben-David 2007].



# Recap: Prediction is insufficient for choosing interventions

How often do they lead us to the right decision?

- Unclear, predictive algorithms provide no insight on effects of decisions

Will the predictions be robust tomorrow, or in new contexts?

- Correlations can change
- Causal mechanisms more robust

What if the prediction accuracy is really high? Does that help?

- Active interventions change correlations

PART I. Introduction to Counterfactual Reasoning

PART II. Methods for Causal Inference

PART III. Large-scale and Network Data

PART IV. Broader Landscape

# PART I. Introduction to Counterfactual Reasoning

What is causality?

Potential Outcomes Framework

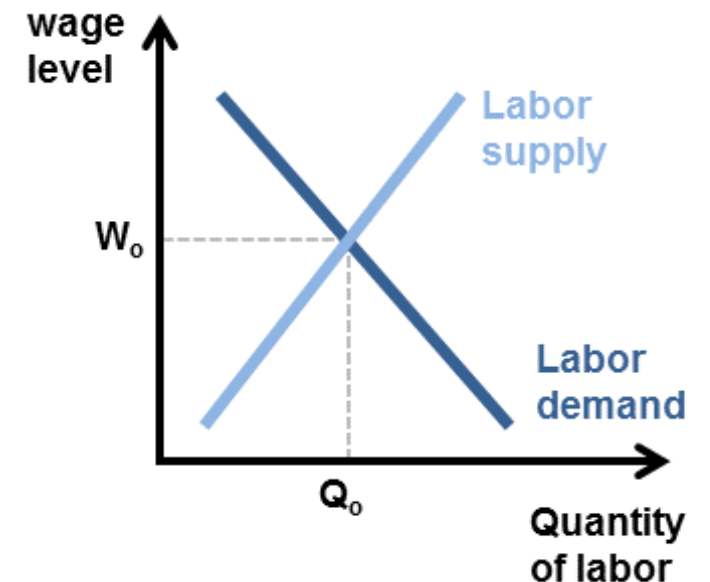
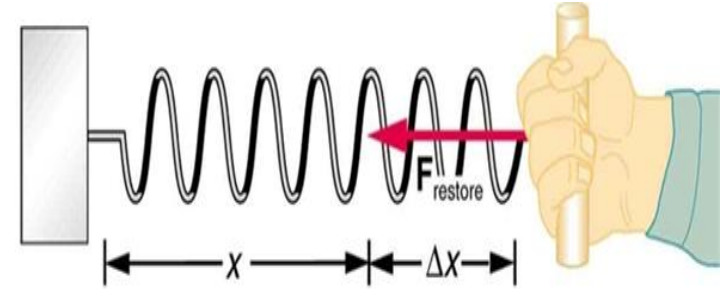
Unobserved Confounds /  
Simpson's Paradox

Structural Causal Model  
Framework



# Cause and Effect

- Questions of cause and effect common in biomedical and social sciences
- Such questions form the basis of almost all scientific inquiry
  - Medicine: drug trials, effect of a drug
  - Social sciences: effect of a certain policy
  - Genetics: effect of genes on disease
- **So what is causality?**
- **What does it mean to *cause* something?**



# A big scholarly debate, from Aristotle to Russell





# What is causality?

- A fundamental question
- Surprisingly, until very recently---maybe the last 30+ years---we have not had a mathematical language of causation. We have not had an arithmetic for representing causal relationships.

*“More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history”*

--Gary King, Harvard University



# The Three Layer Causal Hierarchy

Pearl, Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution, arXiv:1801.04016v1. 11 Jan 2018

Level	Typical Activity	Typical Question	Examples
1. Association $P(y   x)$	Seeing	What is? How would seeing $X$ change my belief in $Y$ ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y   do(x), z)$	Doing, Intervening	What if? What if I do $X$ ?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x   x', y')$	Imagining, Retrospection	Why? Was it $X$ that caused $Y$ ? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What I had not been smoking the past 2 years?

# A practical definition

**Definition:** T causes Y iff  
changing T leads to a change in Y,  
*keeping everything else constant.*

The **causal effect** is the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

\**Interventionist* definition [<http://plato.stanford.edu/entries/causation-mani/>]

# Keeping everything else constant: Imagine a *counterfactual* world

“What-if” questions

Reason about a world that does not exist.



- What if a system intervention was not done?
- What if an algorithm was changed?
- What if I gave a drug to a patient?

# PART I. Introduction to Counterfactual Reasoning

What is causality?

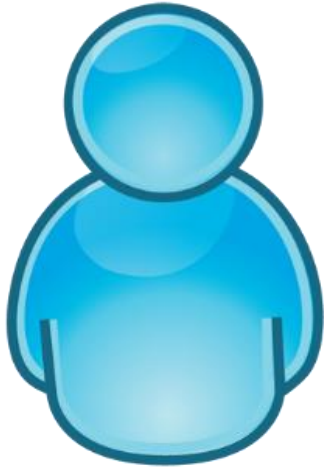
Potential Outcomes Framework

Unobserved Confounds /  
Simpson's Paradox

Structural Causal Model  
Framework



# Potential Outcomes framework

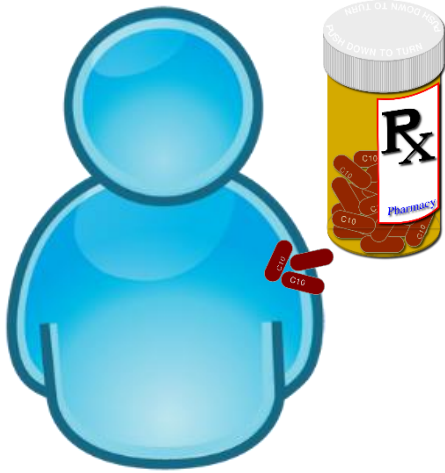


Alice



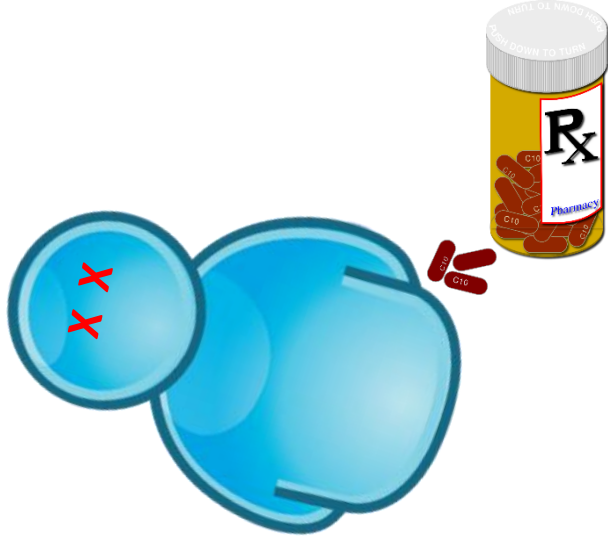
Treatment

# Potential Outcomes framework



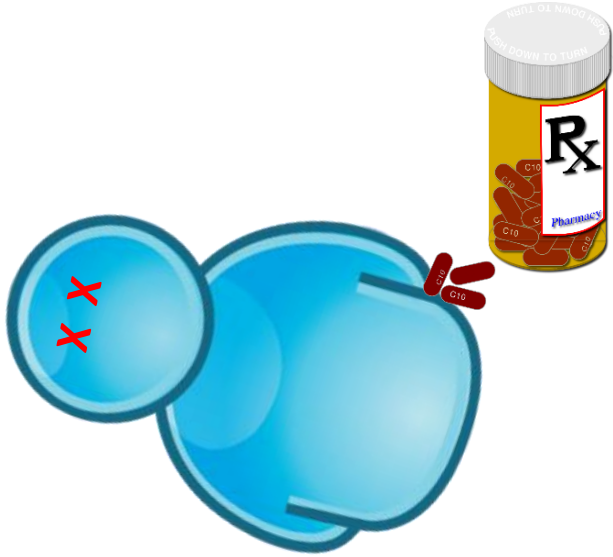
Alice

# Potential Outcomes framework



Alice

# Potential Outcomes framework: Introduce a counterfactual quantity



$Y_{T=1}$



$Y_{T=0}$



Causal effect of treatment =

$$E[Y_{T=1} - Y_{T=0}]$$

Causal inference is the problem of estimating the counterfactual  $Y_{t=\sim t}$

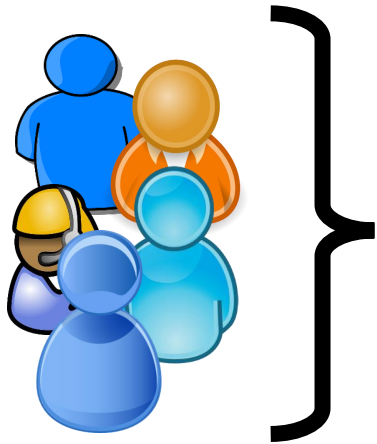
Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	0.3
P2	0	0.8	0.6
P3	1	0.3	0.2
P4	0	0.3	0.1
P5	1	0.5	0.5
P6	0	0.6	0.5
P7	0	0.3	0.1

Causal effect:  $E[Y_{t=1} - Y_{t=0}]$

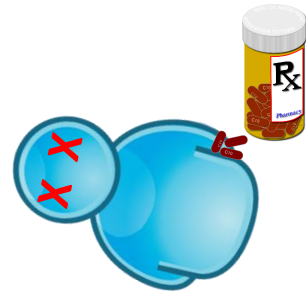
**Fundamental problem of causal inference:** For any person, observe only one: either  $Y_{t=1}$  or  $Y_{t=0}$

# Fundamental problem: counterfactual outcome is not observed

- “Missing data” problem
- Estimate missing data values using various methods
- $Y_{T=0}$  now becomes an estimated quantity, based on outcomes of other people who did not receive treatment



$$\hat{Y}^{T=0}$$



$$Y^{T=1}$$



# Randomized Experiments are the “gold standard”

One way to estimate counterfactual



# Cost: Possibly risky, unethical

Unethical to deny useful treatment or administer risky treatment.

Infeasible or costly in other situations.

What can we do when an experiment is not possible?  
Coming soon in Section 2

# Recap: Potential Outcomes Framework

- Potential outcomes reasons about causal effects by comparing outcome of treatment to outcome of no-treatment
- For any individual, we cannot observe both treatment and no-treatment.
- Randomized experiments are one solution
- We'll discuss others in tutorial Section 2

# PART I. Introduction to Counterfactual Reasoning

What is causality?

Potential Outcomes Framework

Unobserved Confounds /  
Simpson's Paradox

Structural Causal Model  
Framework

# Example: Auditing the effect of an algorithm

System changes algorithm from A to B at some point.

Is the new algorithm B better?

Say a feature that provides information or discount for a financial product.



Algorithm A

Success  
Rate= $\rho$



?

Algorithm B

# New algorithm increases overall success rate

Two algorithms, A (old) and B (new) running on the system.

From system logs, collect data for 1000 sessions for each.

Measure Success Rate (SR).

Old Algorithm (A)	New Algorithm (B)
50/1000 ( <b>5%</b> )	54/1000 ( <b>5.4%</b> )

New algorithm is better?



# Unobserved Confounds

What if there are unobserved features of audience that matter?



Old Algorithm (A)	New Algorithm (B)	Low-income Users
10/400 ( <b>2.5%</b> )	4/200 ( <b>2%</b> )	

Old Algorithm (A)	New Algorithm (B)	High-income Users
40/600 ( <b>6.6%</b> )	50/800 ( <b>6.2%</b> )	

The Simpson's paradox: New algorithm is better overall, but worse for each subgroup

	Old algorithm (A)	New Algorithm (B)
CTR for Low-Activity users	10/400 (2.5%)	4/200 (2%)
CTR for High-Activity users	40/600 (6.6%)	50/800 (6.2%)
<b>Total CTR</b>	<b>50/1000 (5%)</b>	<b>54/1000 (5.4%)</b>

So, which is better?

# From metrics to decision-making

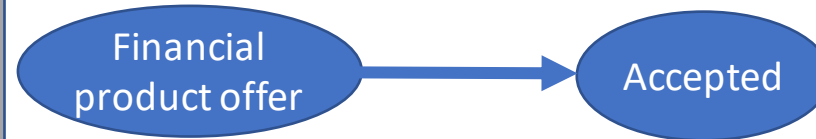
Did the change to new Algorithm increase success rate for the system?

Answer (as usual):

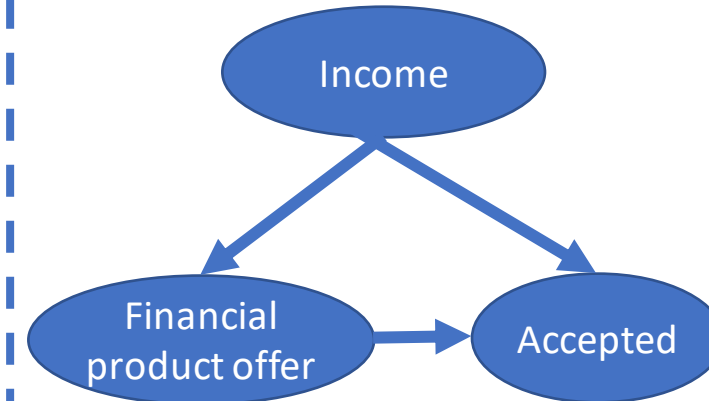
Maybe, maybe not (!)



Higher success rate due to  
**new algorithm**



Higher success rate due to  
**selection effects**

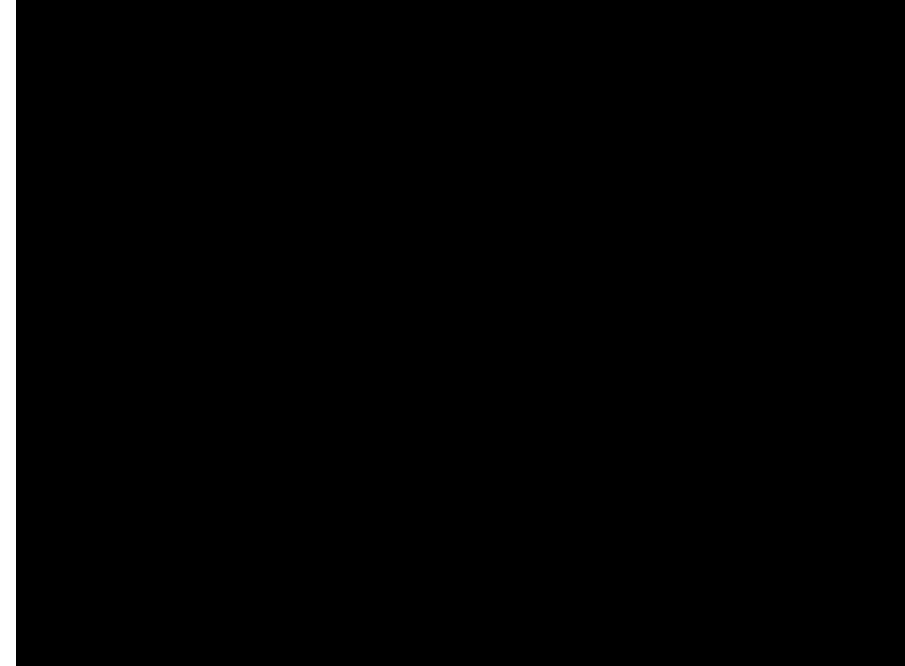
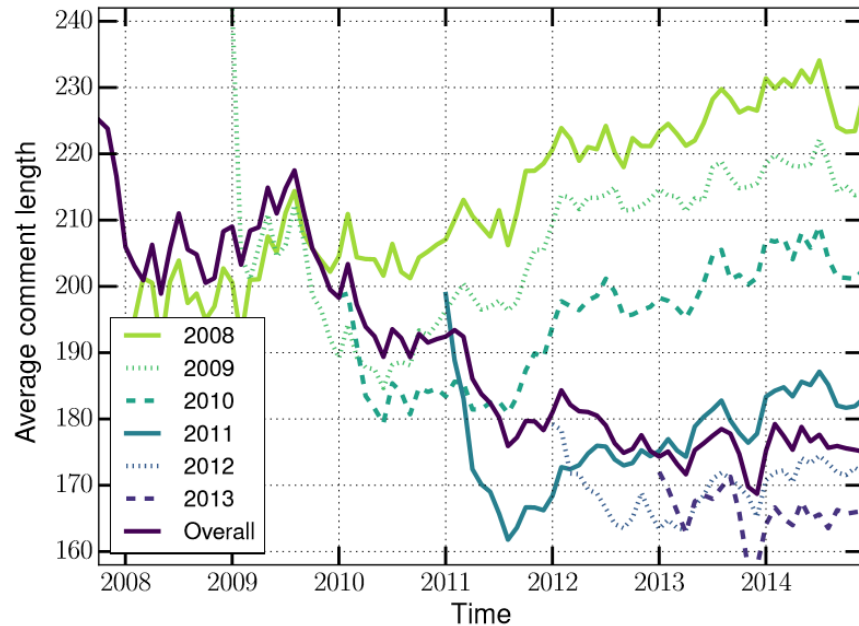


E.g., Algorithm B is shown at a different time than A.

There could be other hidden causal variations.

Not just theory. Differences in interpretations can attract lawsuits (UC Berkeley admissions, 1973)

# Example: Simpson's paradox in Reddit



Average comment length decreases over time.

Making sense of  
such data can be  
too complex.

**D'oh!**



Not Simpson's Paradox

# Recap: Unobserved Confounds

- Unobserved confounds are a threat to causal reasoning



# PART I. Introduction to Counterfactual Reasoning

What is causality?

Potential Outcomes Framework

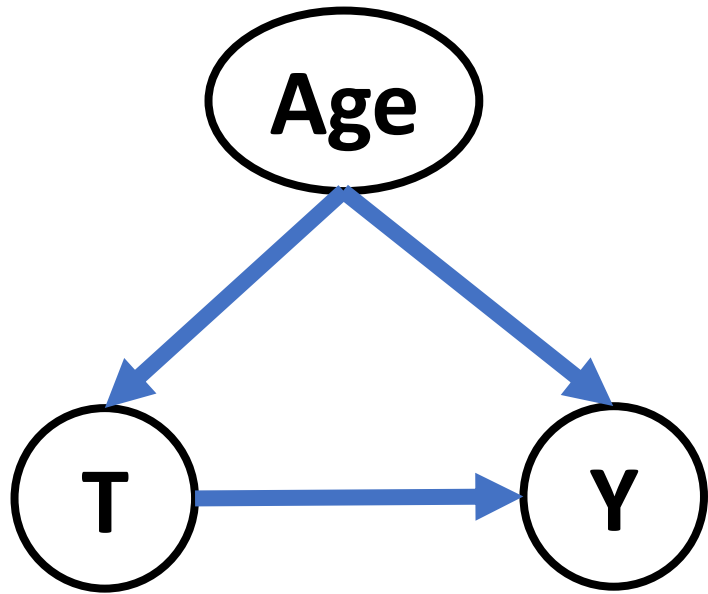
Unobserved Confounds /  
Simpson's Paradox

Structural Causal Model  
Framework

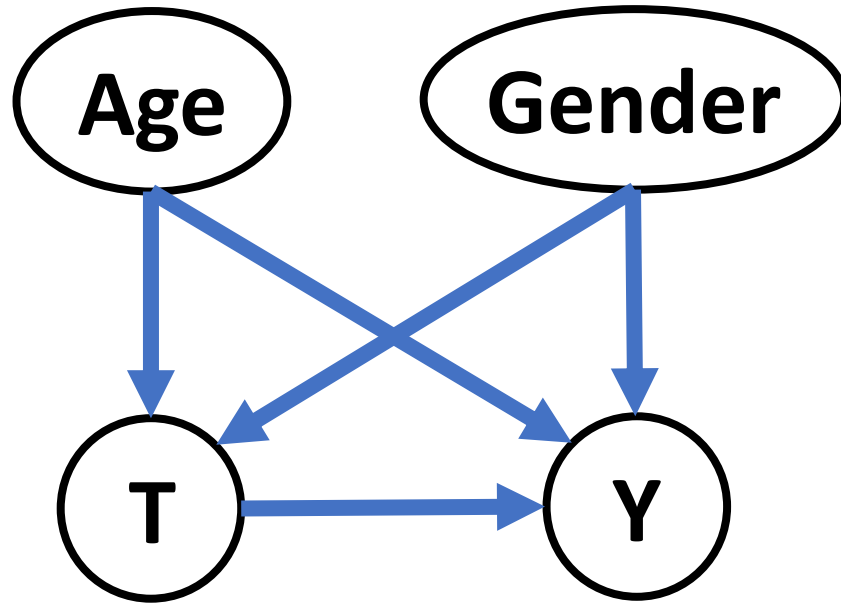
# Real world is complicated

- People may have inter-related characteristics
  - How are these characteristics associated with each other?
- Other factors can influence the observed outcome
  - How do they affect treatment and outcome?
  - Which ones to include?
- How to identify the causal effect in such cases?
- When is it possible to find a causal effect?
  - We can use graphical model framework to answer this

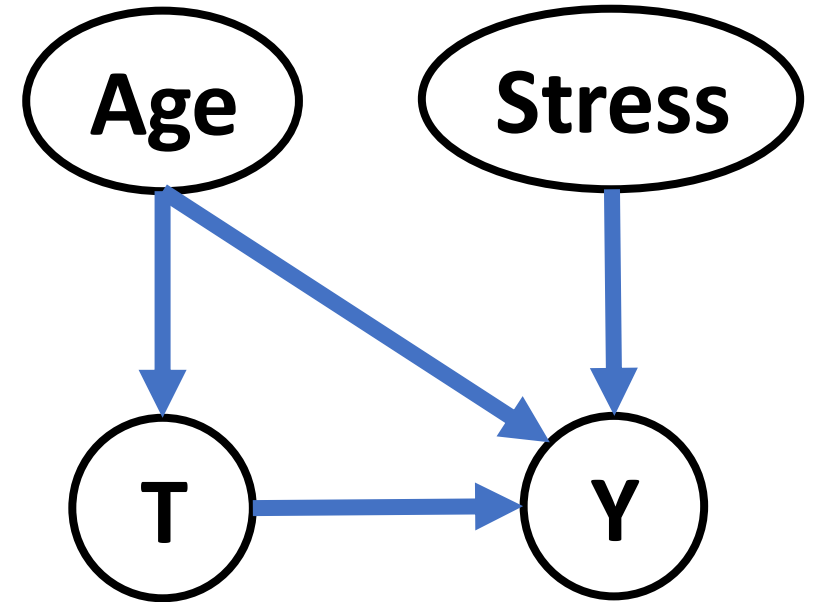
Which variables to condition on?



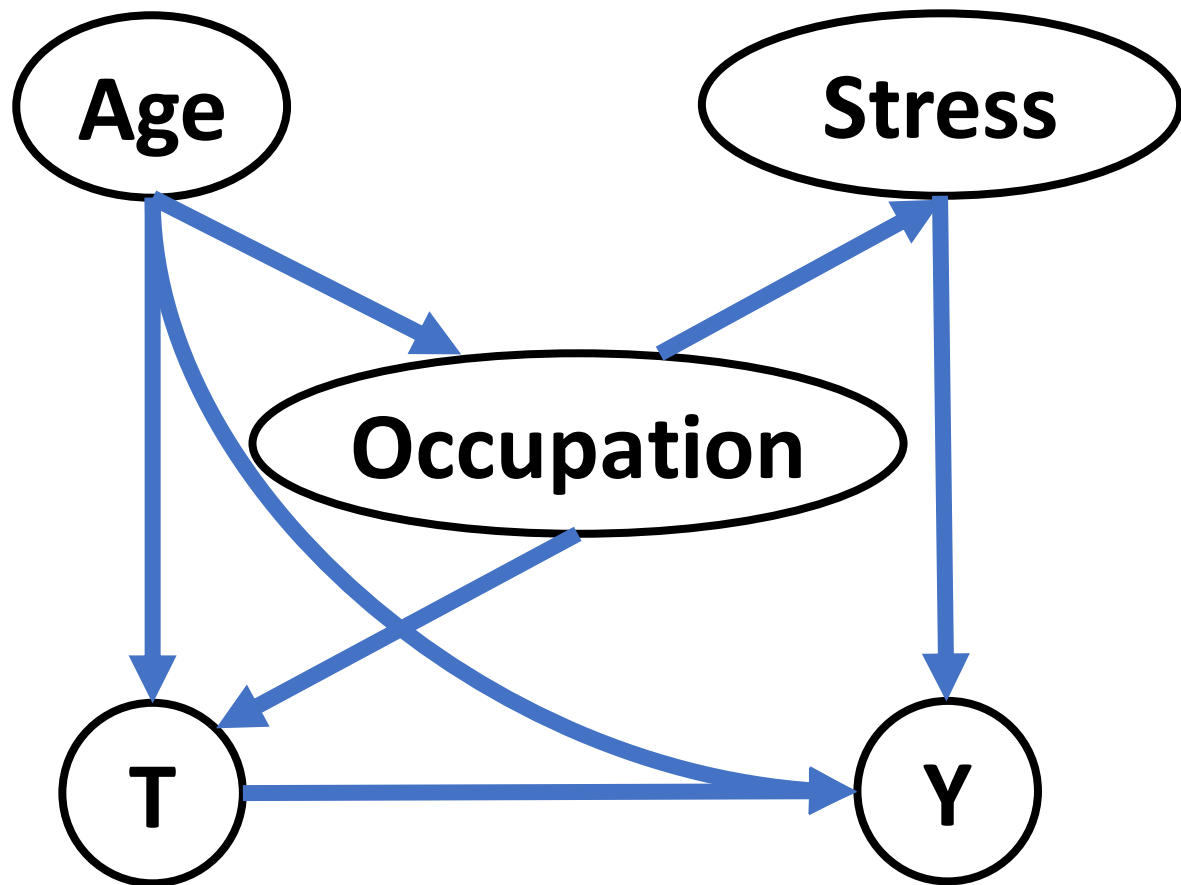
$X = \{Age\}$



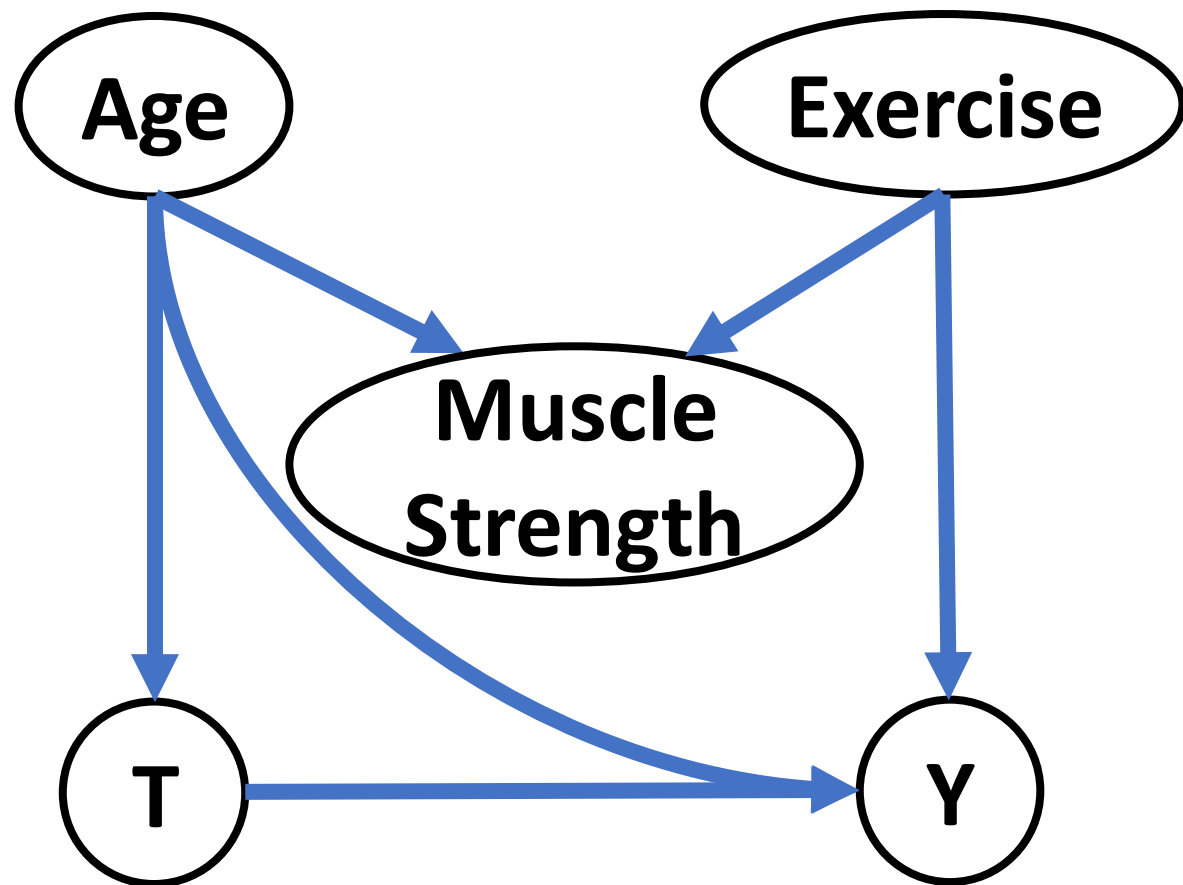
$X = \{Age, Gender\}$



$X = \{Age\}$

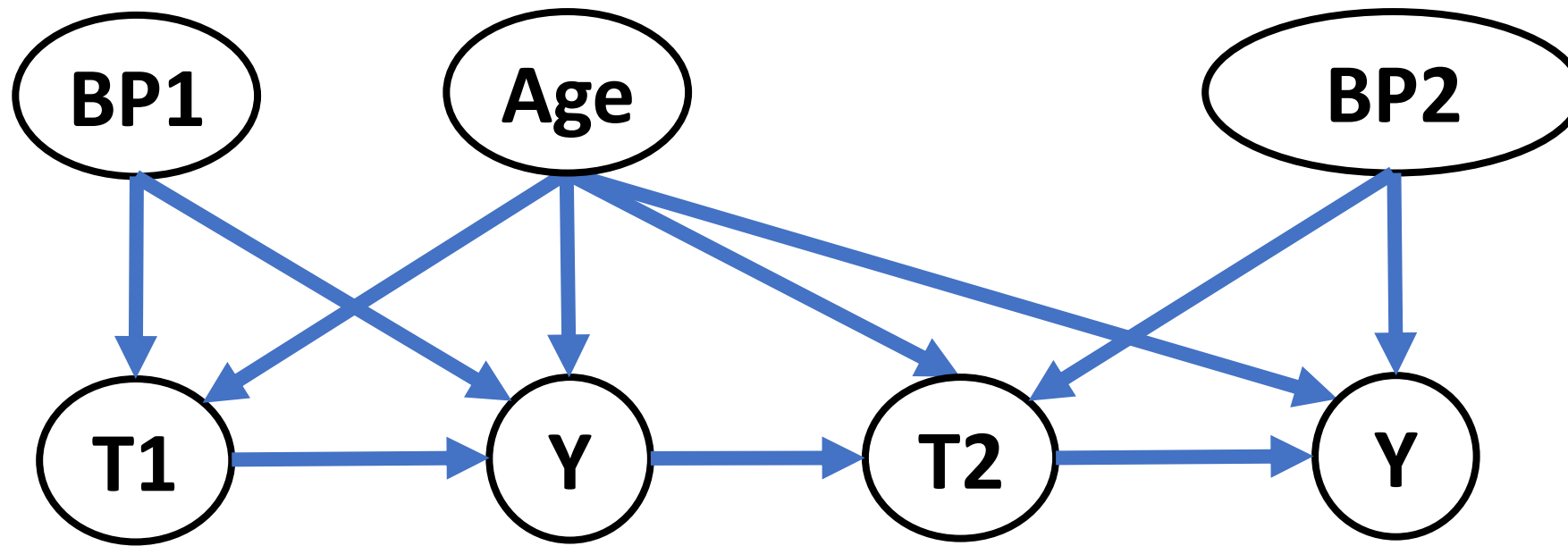


$X = ?$



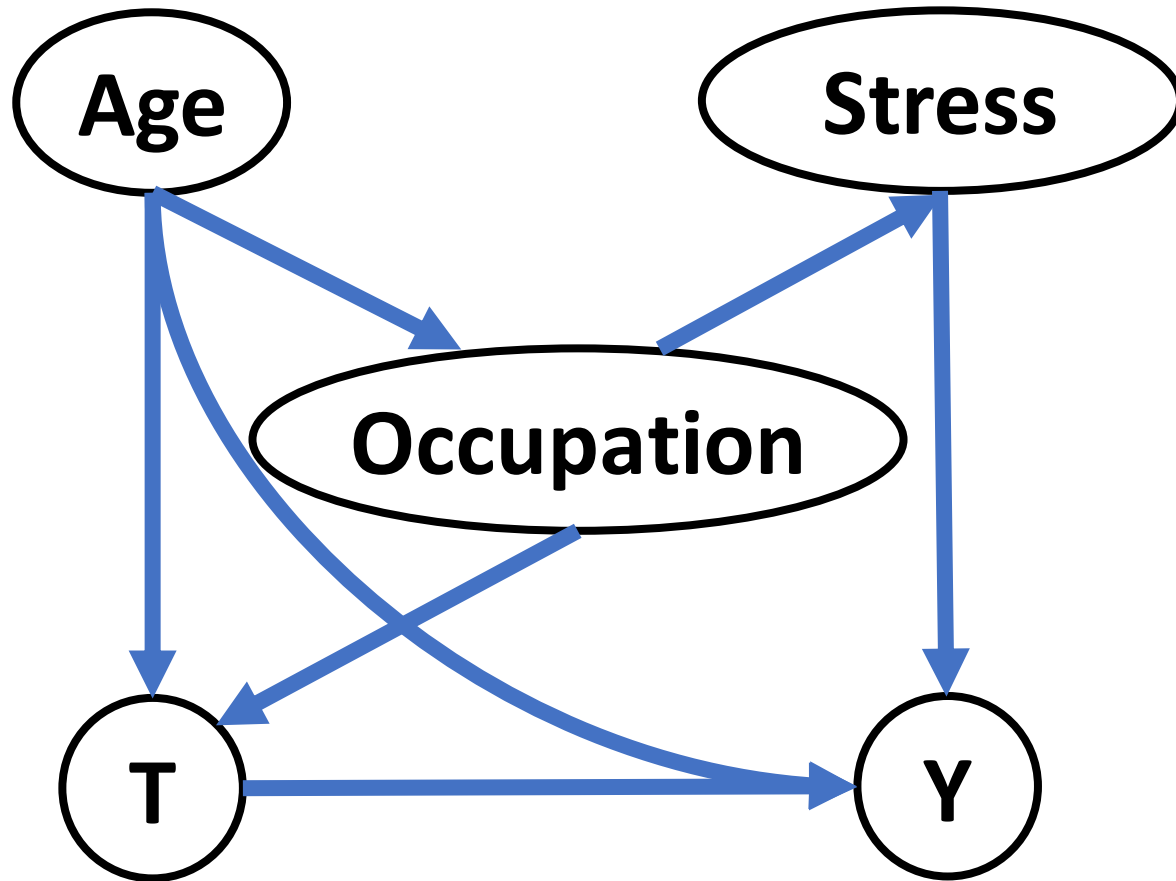
$X = ?$

Another example: Repeated treatment (!)



**How to reason about causal effects in such cases?**

# Structural Causal Model: A framework for expressing complex causal relationships



$$\begin{aligned} \text{Occupation} &= h(\text{Age}, u_o) \\ \text{Stress} &= k(\text{Occupation}, u_s) \end{aligned}$$

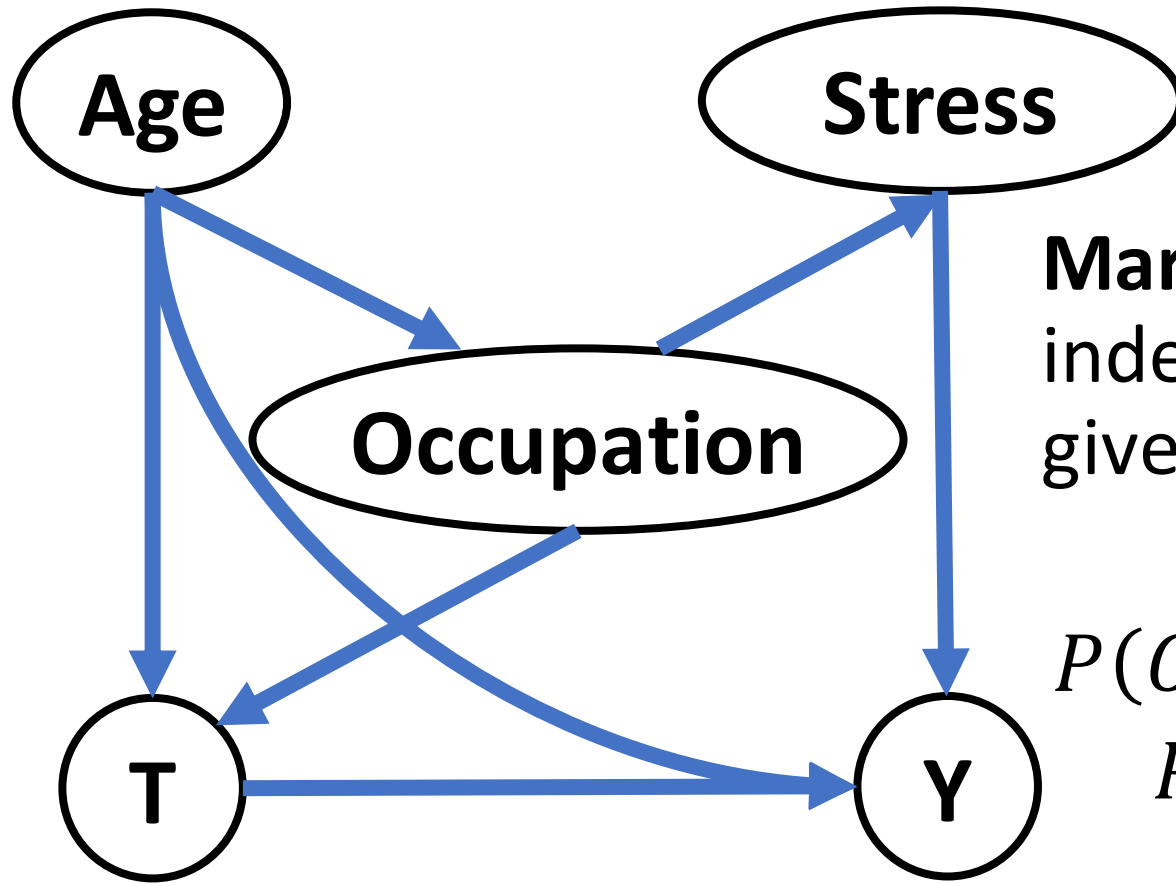
$$\begin{aligned} T &= g(\text{Age}, \text{Occupation}, u_t) \\ Y &= f(T, \text{Age}, \text{Stress}, u_y) \end{aligned}$$

Edges represent *direct* causes.

Directed paths represent *indirect* causes.



# Structural Causal Model: A framework for expressing complex causal relationships

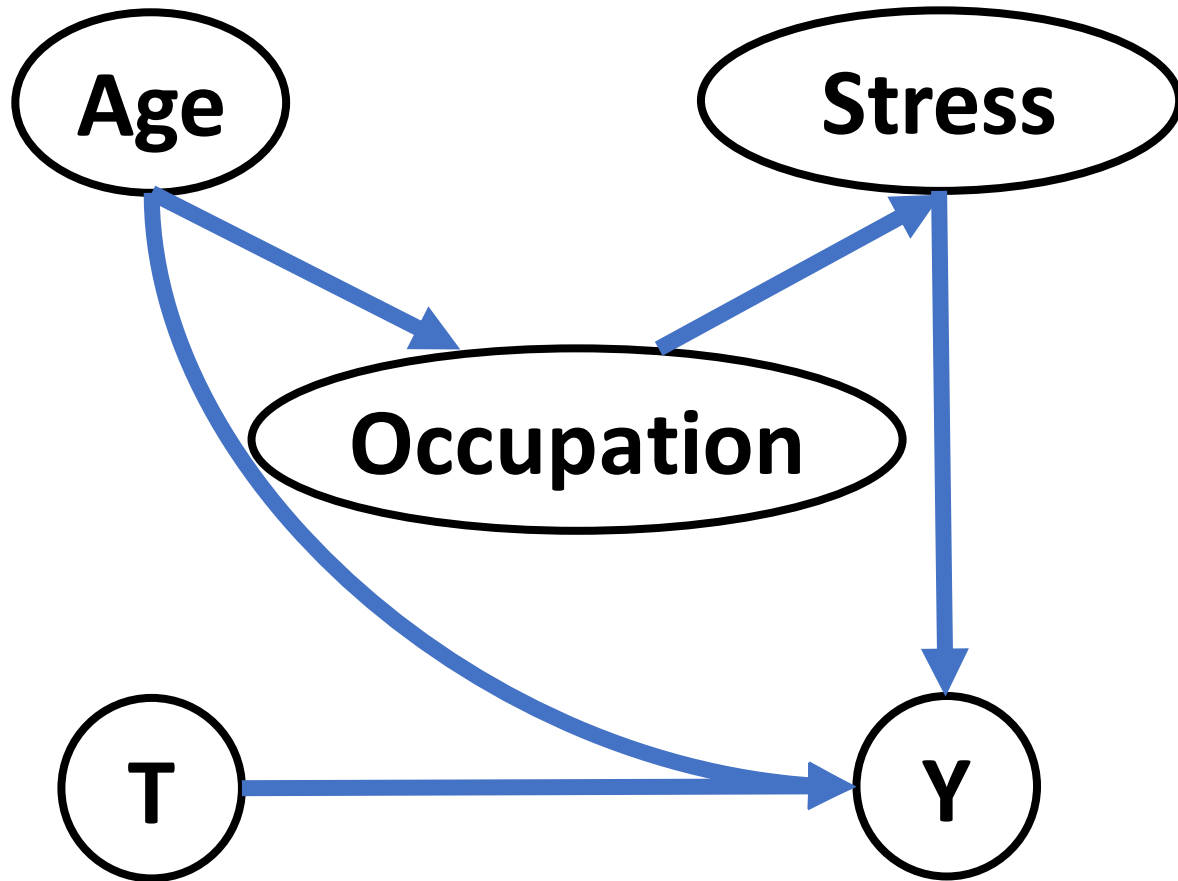


**Markov assumption:** A node is independent of all its non-descendants given its parents.

$$P(Occ. | Age, Stress) = P(Occ. | Age)$$
$$P(T | Occ., Stress) = P(T | Occ.)$$

$$P(G) = P(Age)P(Occ. | Age)P(Stress | Occ.)P(T | Age, Occ.)P(Y | T, Age, Stress)$$

Structural Causal Model: Causal effect is represented by the intervention distribution



**Counterfactual (Intervention) world:**

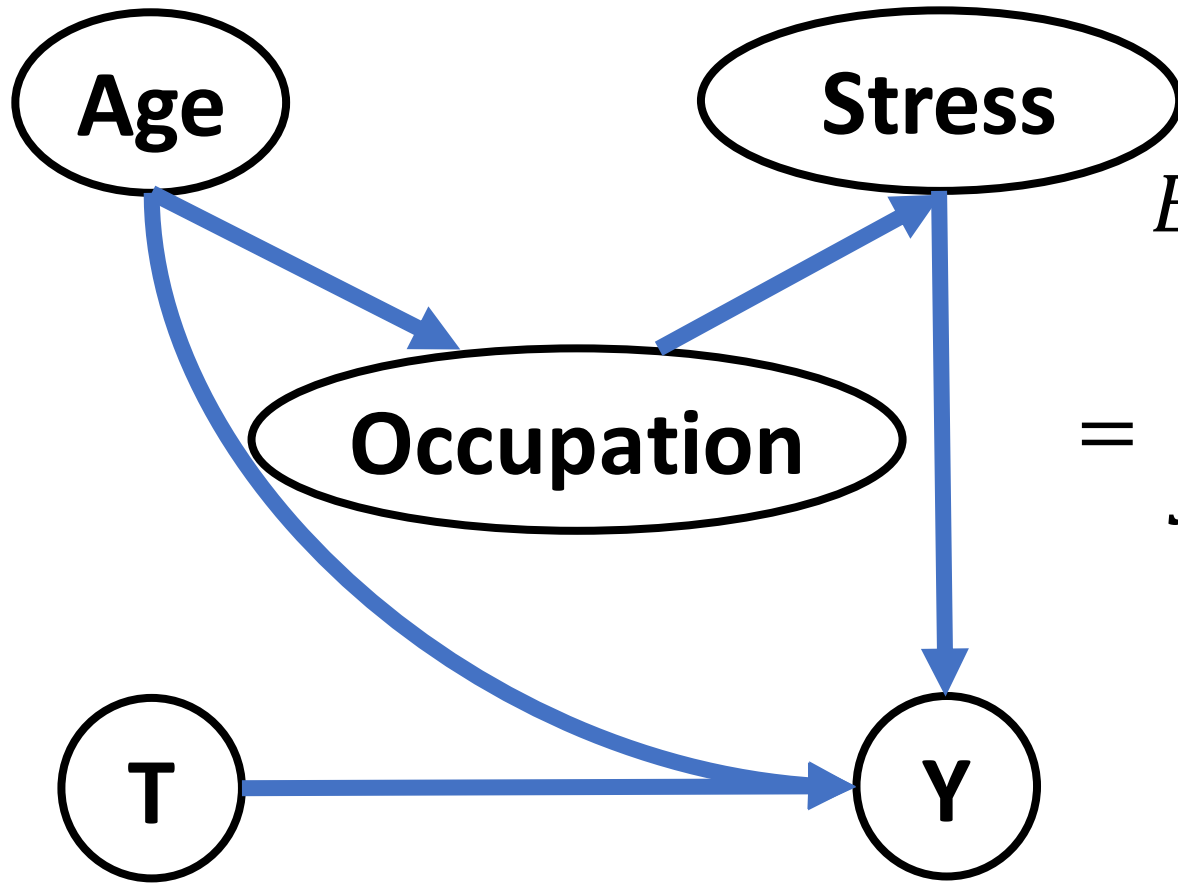
All edges to Treatment T removed, *keeping everything else the same.*

Observed correlation =  $P(Y|T)$

Causal Effect =  $P^*(Y|T)$

$$P^*(\Phi) = P(\text{Age})P(\text{Occ.}|\text{Age})P(\text{Stress}|\text{Occ.})P^*(T|\text{Age}, \text{Occ.})P(Y|T, \text{Age}, \text{Stress})$$

Structural Causal Model: Causal effect is represented by the intervention distribution

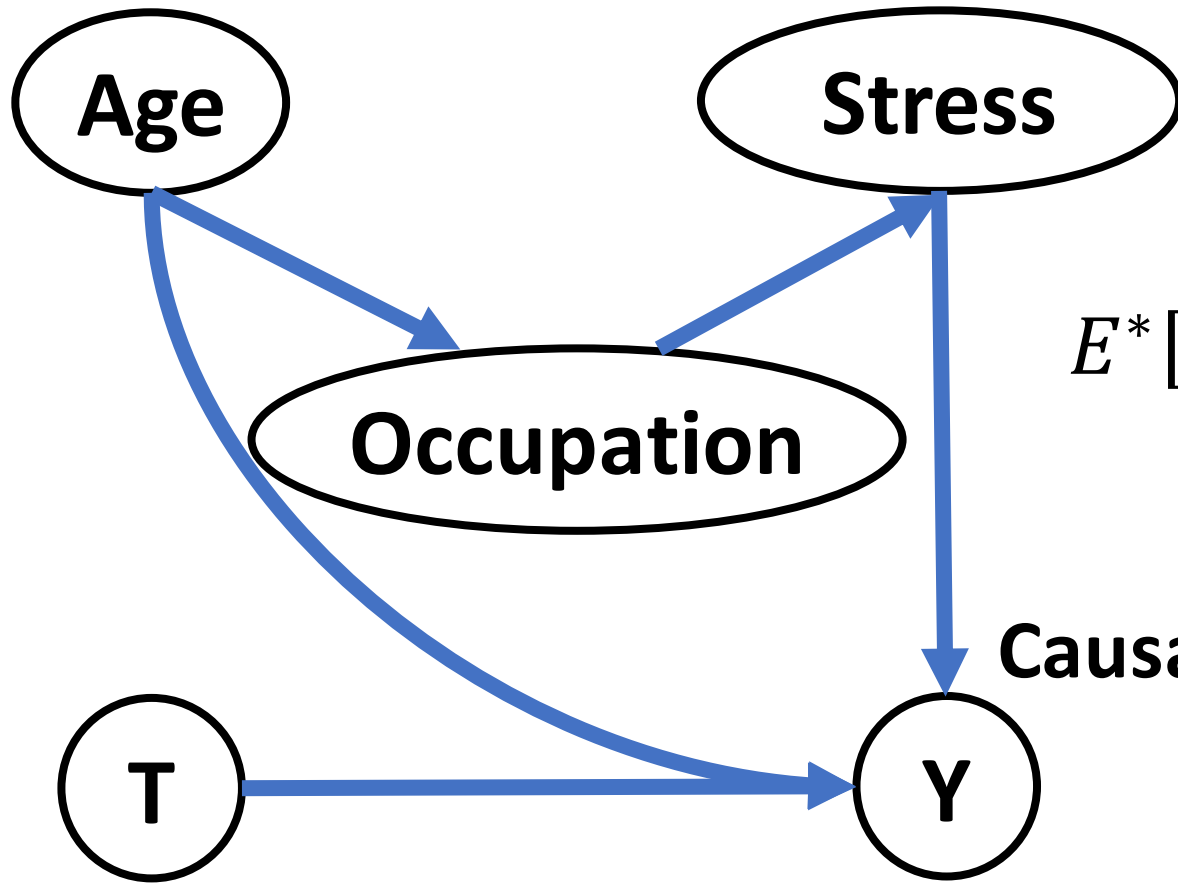


$$\begin{aligned}
 E^*[Y] &= E_{\phi \sim P^*(\Phi)}[y] = \int_{\phi} y P^*(\phi) \\
 &= \int_{\phi} y \frac{P(\phi)}{P(\phi)} P^*(\phi) = \int_{\phi} y \frac{P^*(\phi)}{P(\phi)} P(\phi) \\
 &= \int_{\phi} y \left[ \frac{P^*(T|Age, Occ.)}{P(T|Age, Occ.)} \right] P(\phi)
 \end{aligned}$$

$$P^*(\Phi)$$

$$= P(Age)P(Occ.|Age)P(Stress|Occ.)P^*(T|Age, Occ.)P(Y|T, Age, Stress)$$

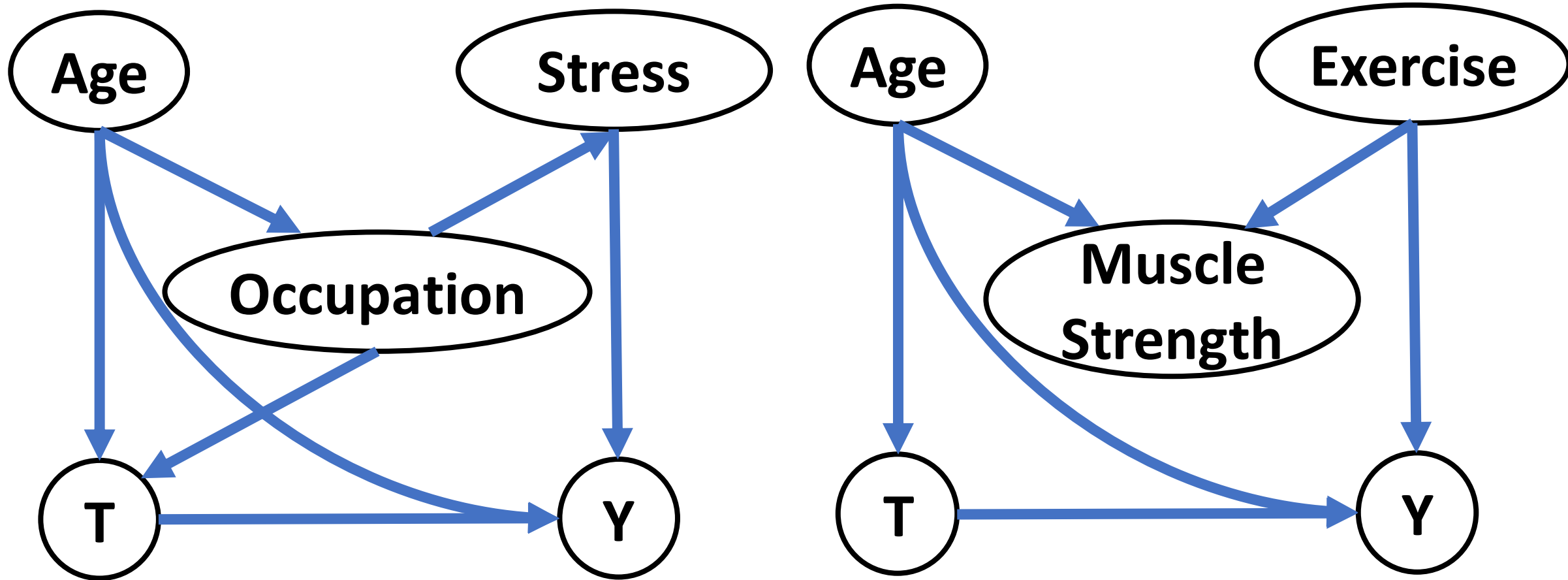
Structural Causal Model: Causal effect is represented by the intervention distribution



$$E^*[Y] = \int_{\phi} y \left[ \frac{P^*(T|Age, Occ.)}{P(T|Age, Occ.)} \right] P(\phi)$$

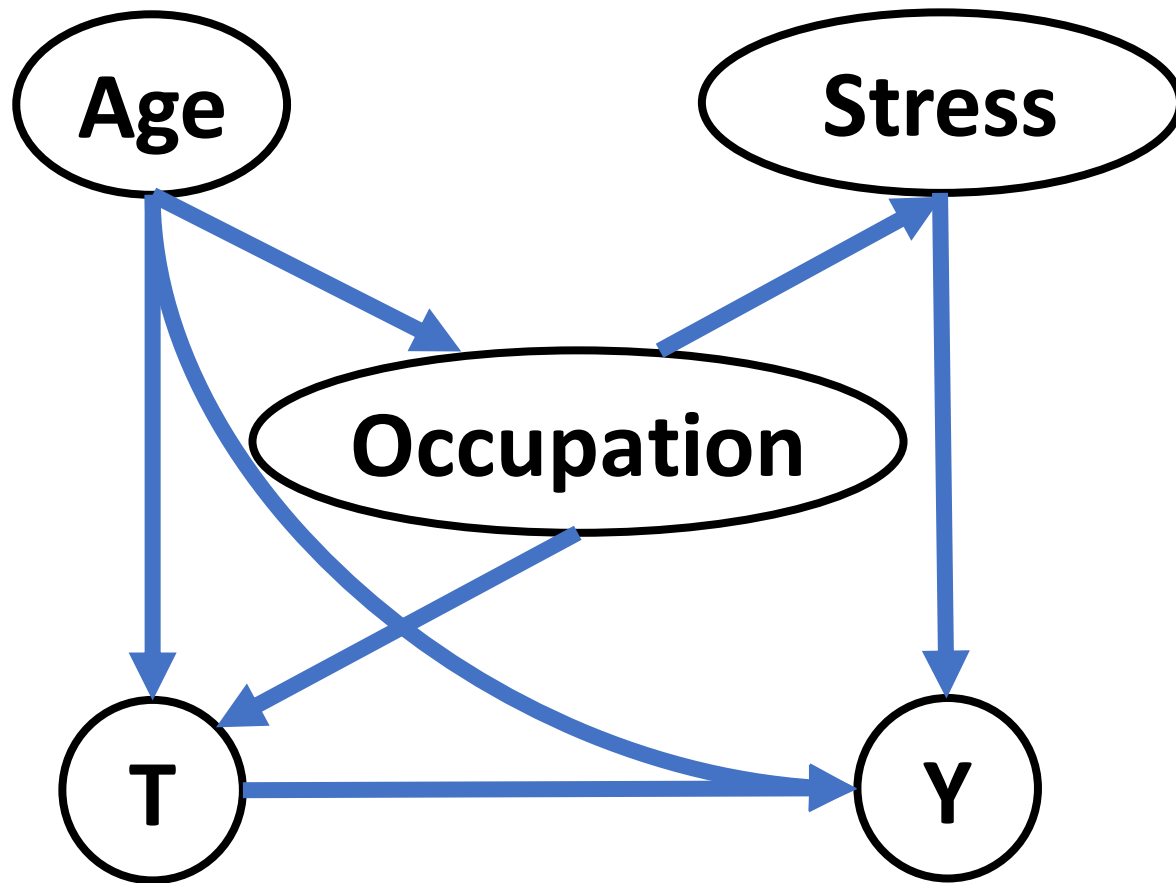
**Causal Effect:**  $E^*[Y|T = 1] - E^*[Y|T = 0]$

Structural Causal Model makes assumptions explicit



The graph encodes all causal assumptions.

Important: Assumptions are the edges that are *missing*



**Assumption 1:** Occupation does affect outcome Y.

**Assumption 2:** Age does not affect stress.

**Assumption 3:** Stress does not affect Occupation.

**Assumption 4:** Treatment does not affect stress.

*..and so on.*

**Condition for validity:** The graph reflects all relevant causal processes.

# Important: SCM and Potential Outcome frameworks are equivalent

## Potential Outcomes

$$E[Y_{T=1}] - E[Y_{T=0}]$$

## Structural Causal Model

$$E^*[Y|T = 1] - E^*[Y|T = 0]$$

If we denote  $E[Y_T] \leftarrow E^*[Y|T]$ , then the formulations are equivalent.

More formally, a theorem in one framework is a theorem in another.



# Key Benefit (1) of SCM: Provides a language for expressing counterfactuals

*If a person was given treatment, what is the probability that he would be cured if he was not given treatment?*

$$P(Y = 1|T = 1, T = 0)$$

**Non-sensical.**

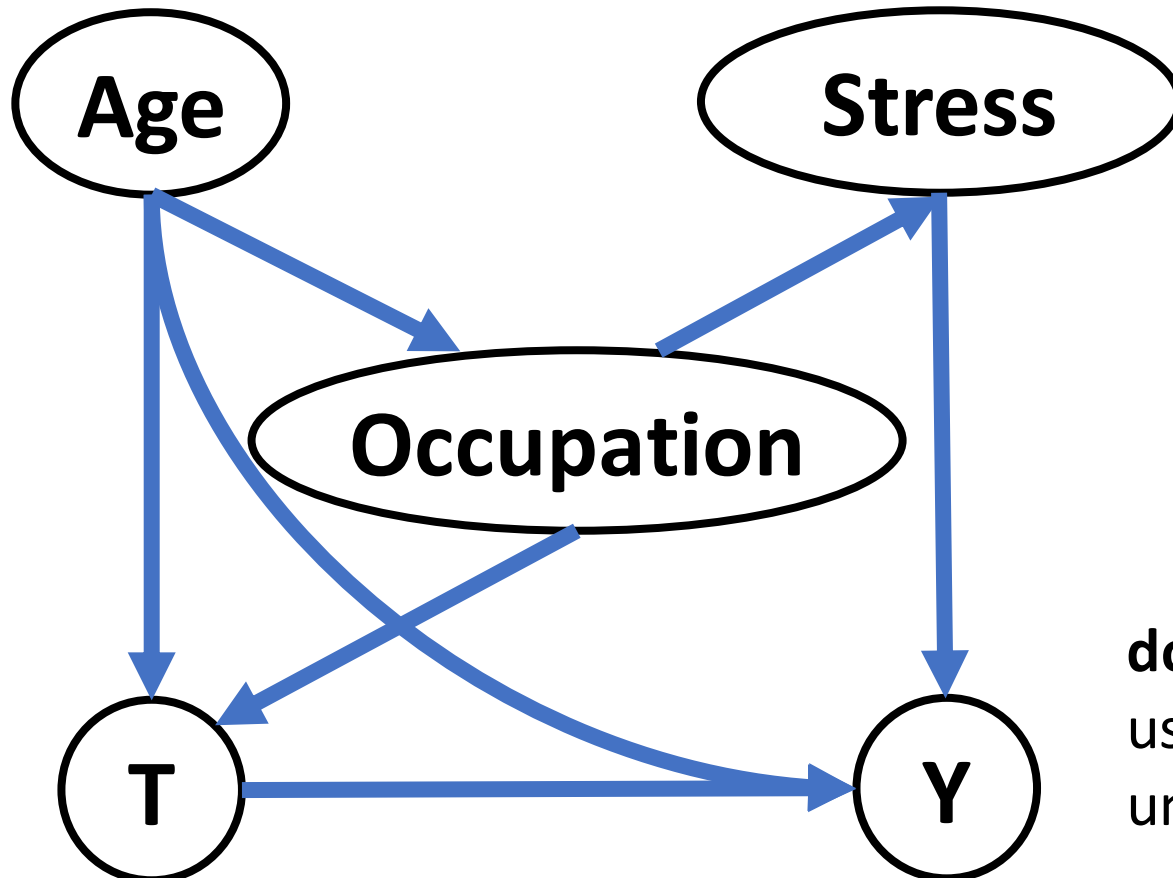
*Can write it as:*

$$P(Y_{T=0} = 1|T = 1), \text{ or } \\ P(Y = 1|T = 1, do(T = 0))$$

$P(Y|do(T))$  avoids confusion with  $P(Y|T)$

# Key Benefit 2 of SCM: Provides a mechanistic way of identifying causal effect

**do-calculus:** A rule-based calculus that can help identify any counterfactual quantity.



E.g.,

$$P(Y|do(T))$$

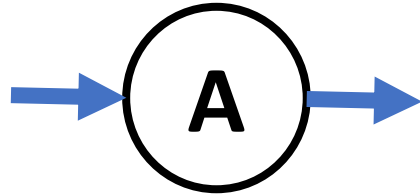
$= \dots do\text{-calculus rules} \dots$

$$= \sum_{Age, Stress} P(Y|T, Age, Stress) P(Age, Stress)$$

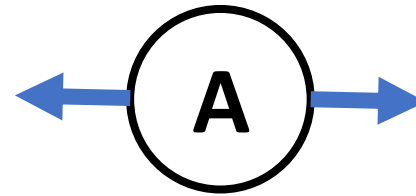
**do-calculus is complete:** If we cannot identify using do-calculus, causal effect cannot be identified.

# Advanced Topic: Back-door criterion

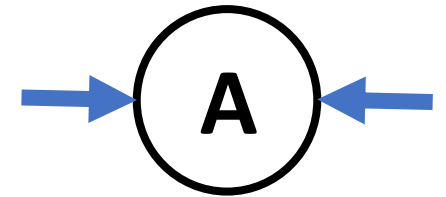
Three kinds of  
node-edges



If conditioned on X




If conditioned on X



If not conditioned on X

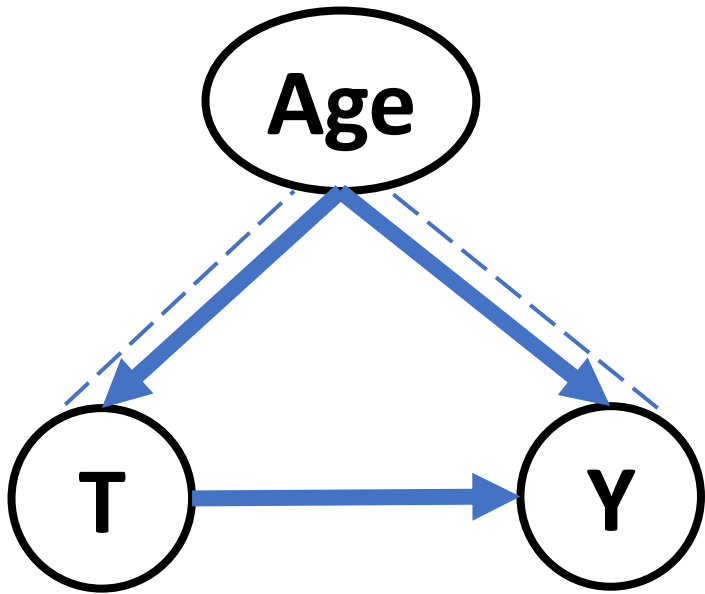
**Path is  
“blocked”**

**“Back-door” path:** Any undirected path that starts with  **T** and ends with  **Y**

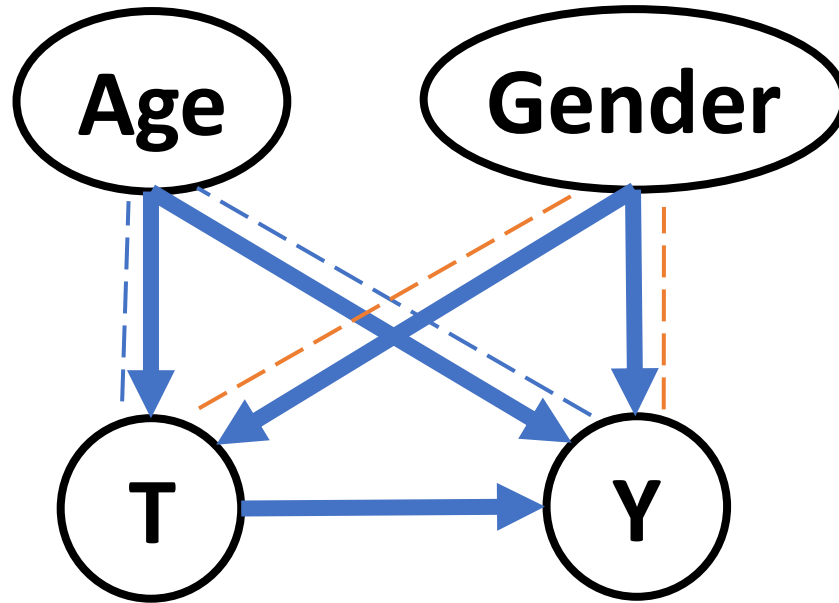
**Back-door criterion:** If conditioning on X blocks all back-door paths between treatment T and outcome Y, then

$$P(Y|do(T)) = \sum_x P(Y|T, X = x)P(X = x)$$

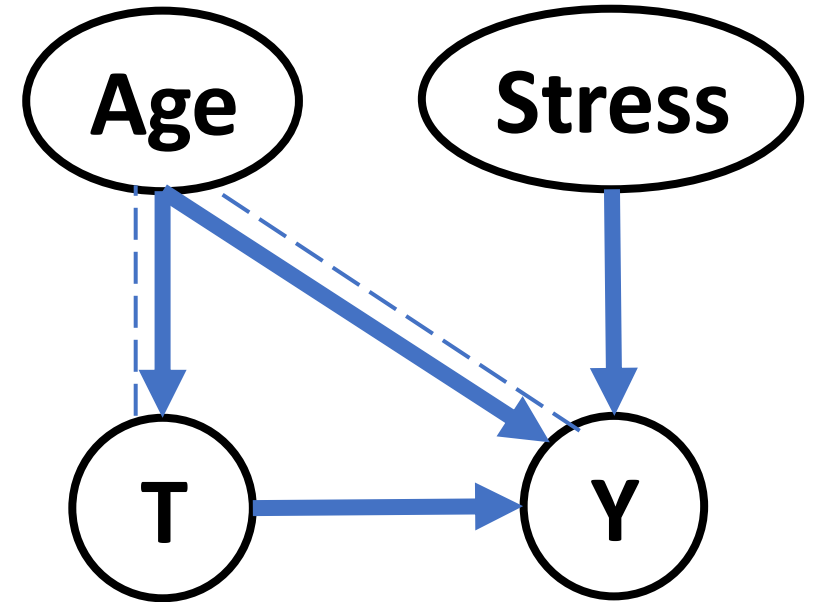
Let us return to our examples



$$X = \{Age\}$$

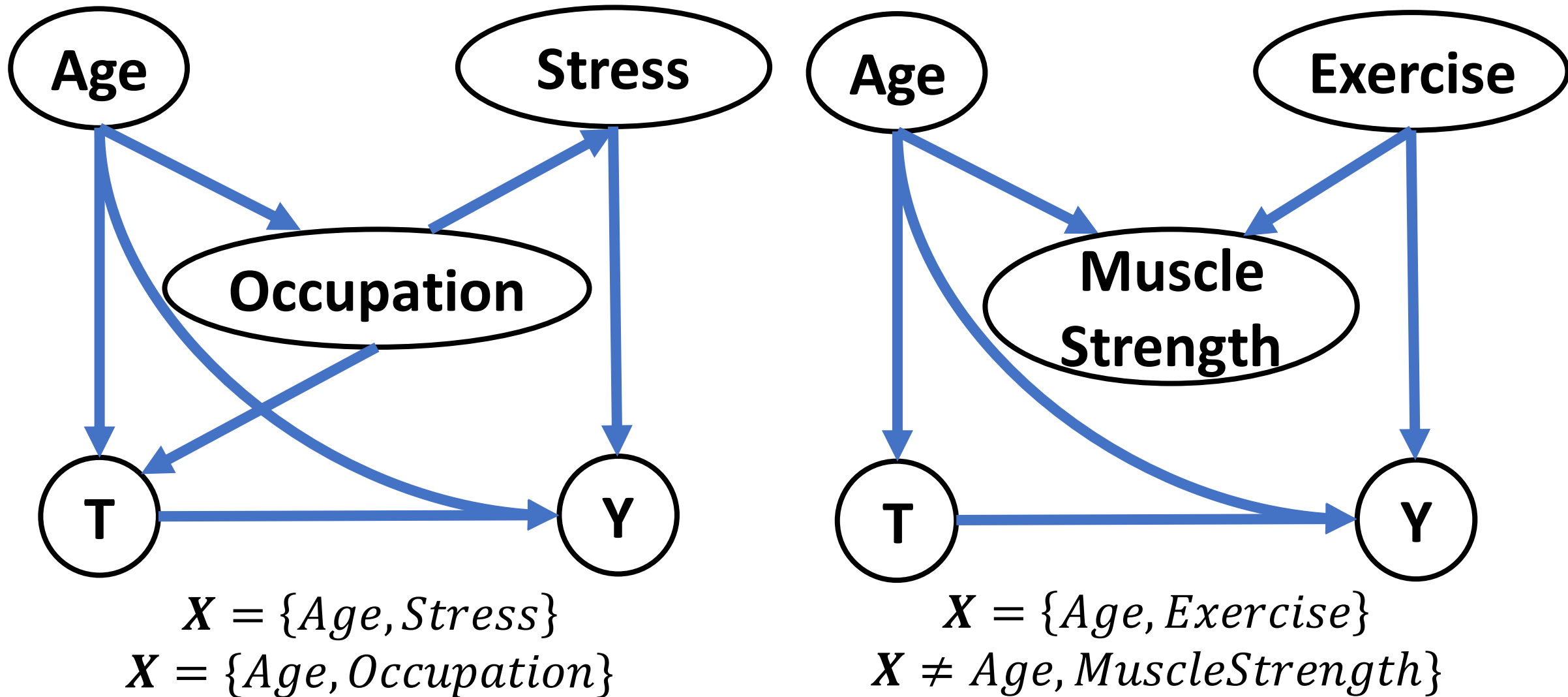


$$X = \{Age, Gender\}$$



$$X = \{Age\}$$

Back-door criterion provides a precise way to find variables to condition to



# Both frameworks have merits

**Use structural causal model and do-calculus for**

modeling the problem

making assumptions explicit

identifying the causal effect

**Use potential outcomes-based methods for**

estimating the causal effect

# Recap: Structural Causal Models

- Allow us to make causal assumptions explicit
  - Assumptions are the missing edges!
- Provide language for expressing counterfactuals
- Well-defined mechanisms for reasoning about causal relationships
  - E.g., Backdoor criterion

# Recap: Section 1 - Introduction

- **Causality** is important for decision-making and study of effects
- **Potential Outcomes Framework** gives practical method for estimating causal effects
  - Translates causal inference into counterfactual estimation
- **Unobserved confounds** are a critical challenge
- **Structural Causal Model Framework** gives language for expressing and reasoning about causal relationships



PART I. Introduction to Counterfactual Reasoning

PART II. Methods for Causal Inference

PART III. Large-scale and Network Data

PART IV. Broader Landscape