Reminders:

- **Tim's office hours 11:20-11:50am PT today**
- **Final project presentations in person G01/G10 on Monday, June 3, 10:30am-12:30pm** (check Ed for links, schedule, details)
  - Great opportunity to learn about each other's projects. Lots of Q&A!
  - Attendance is mandatory
  - Active participation rewarded with extra credit
- **Upload your deliverables on Gradescope by Sunday 23:59pm PT – no late periods so that we can prepare the session and give you feedback and grades quickly**
  - Project Report
  - Presentation Video (and slides PDF)
    - 6 minutes (no credit if longer – we need to be fair across groups)
  - Metadata (primarily dataset info)

Causal Inference I

Introduction to Counterfactual Reasoning

CSE547 Machine Learning for Big Data

Tim Althoff

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Plan for today: Introduction to Counterfactual Reasoning

- When is prediction / big data not enough?
- What is causality?
- Potential Outcomes Framework (Rubin)
  - How can we define and compute causal effects?
- Unobserved Confounds & Simpson's Paradox
  - Why we should always worry about confounding in decision making?
- Structural Causal Model Framework (Pearl)
  - How can we make our assumptions explicit?
  - Given our assumptions, is causal inference feasible? Can we *identify* a causal effect?

Plan for Thursday

- We will learn today that more efficient estimators are needed.

- Effective estimation of causal effects
  - Conditioning on Key Variables
  - Matching
  - Stratification
  - Weighting
  - Regression
  - Sensitivity Analyses

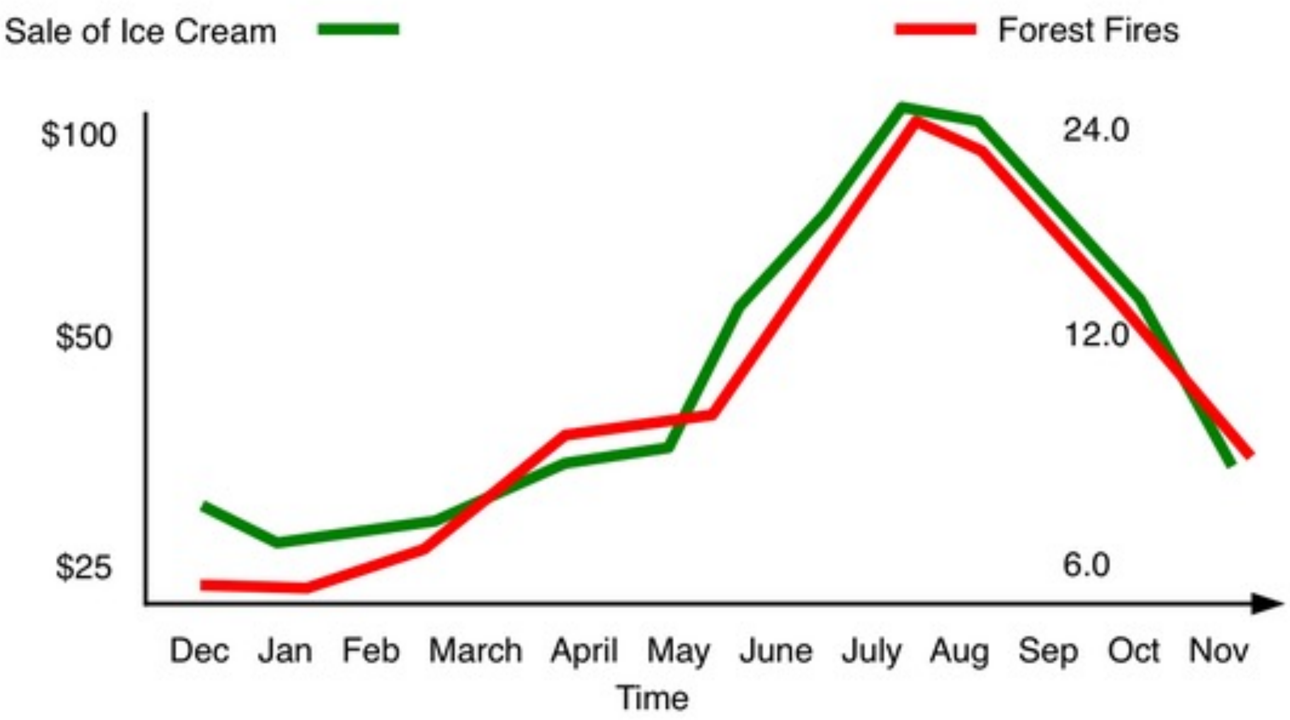
When is prediction / big data not enough?

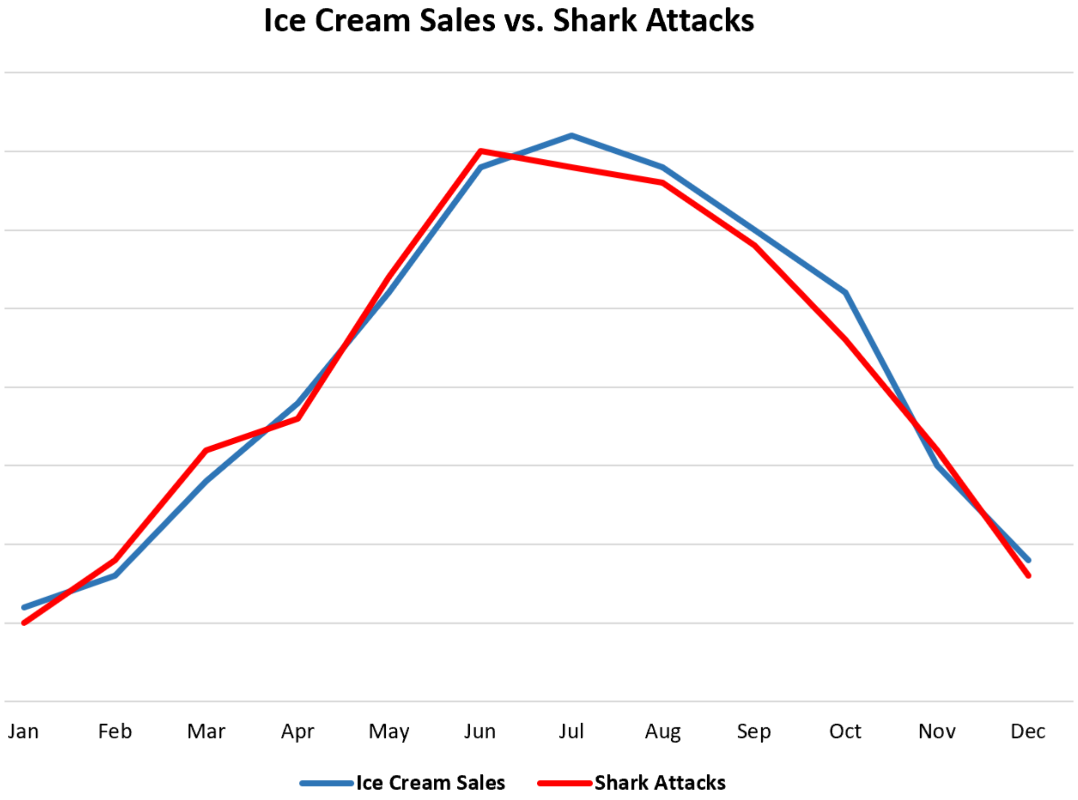
Prediction is everywhere!

- Recommender Systems: Predicting future rating/consumption
- Social Networks: Link prediction
- Course projects …

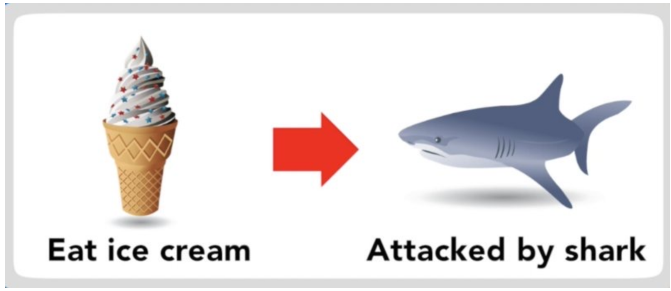
- We have increasing amounts of data and increasingly accurate predictions! Why do we need causal inference? When is more data not enough?

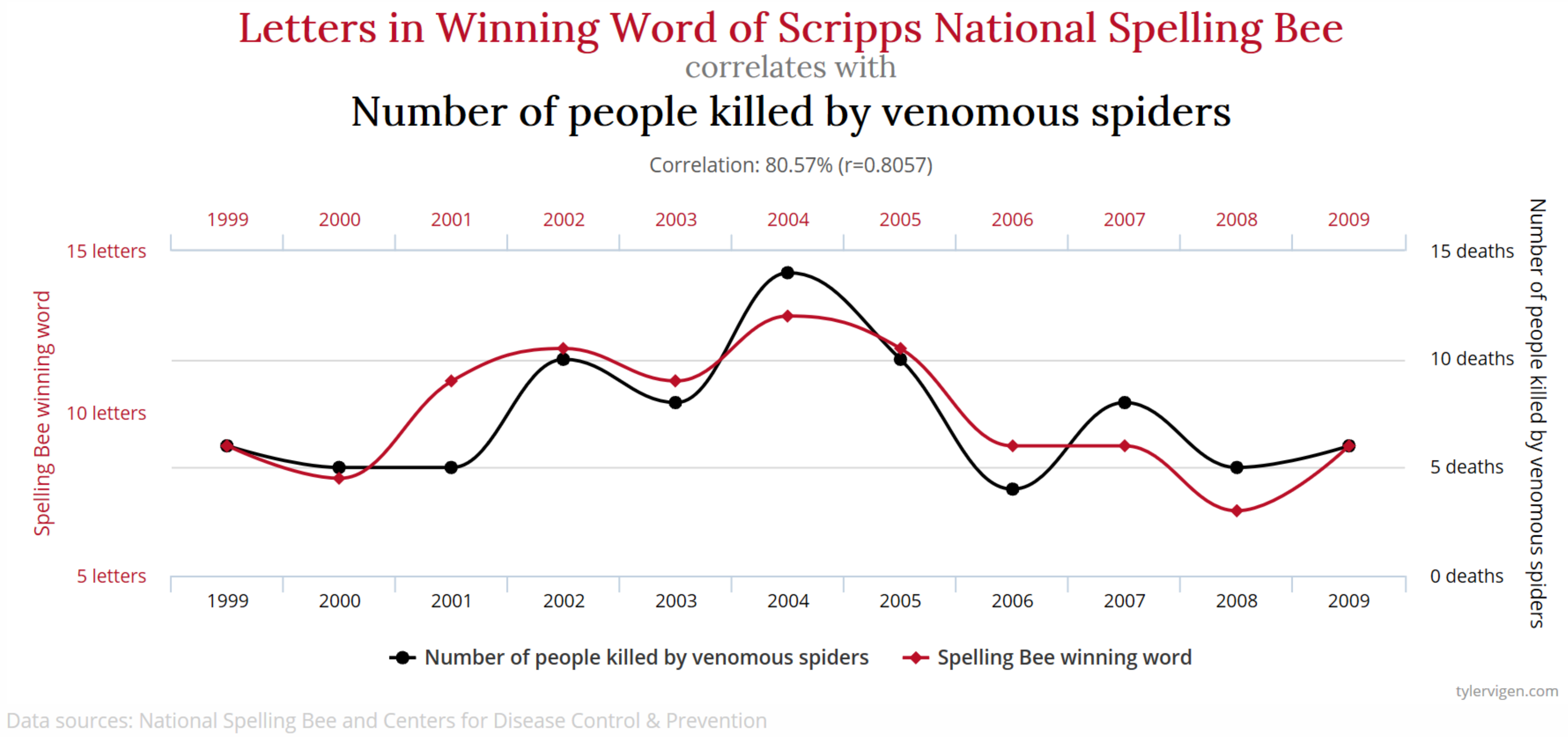
Do prediction models guide decision-making?





Do we need to regulate ice cream sales?





[Source: http://www.tylervigen.com/spurious-correlations](http://www.tylervigen.com/spurious-correlations)

Takeaways

- If it is a spurious correlation without underlying causal connection, we likely won't be able to predict the future well!

- But wait! Wouldn't we be able to predict shark attacks well next summer based on our ice cream sales model?

- Changes in the environment & interventions may cause your predictive model to fail
  - What if we move to southern hemisphere? July is winter now.
- Typically assume that train/test sets are drawn from same distribution in supervised learning - No such guarantee in real life!

- Problematic: Acting on a prediction changes distribution!
  - Echo chamber: Recommend political news – if we start recommending only certain articles, we will see less clicks on other articles in the future, even if someone might have read them if recommended.
  - Incl. critical domains: healthcare or adversarial scenarios

What if I have a ton of data?

Big data to the rescue?

- “Look at how much data I had…”
- ”How could I be wrong? I used 3 billion data points!”
- “This is just noise. All the problems will cancel out…”

- Beware! You do need to worry about bias **and** variance!
- **More data does not help you reduce bias!**

- **This week: Sources of bias, how to model it, and what to do about it**

The Reasonable Uneffectiveness of Big Data

- Play on “The Unreasonable Effectiveness of Data”
  - By Alon Halevy, Peter Norvig, and Fernando Pereira at Google
  - Simple models + Lots of data work very well

- Now consider context of causal inference
  - Common threats to causal inference are independent of sample size (more details later)
  - When we cannot observe counterfactuals, observing more data will not help us! (formal definition coming later)

Big Data does not address…

...common threats to causal inference, including:

1. Construct validity
   - E.g. measurement error
2. Internal Validity
   - E.g. confounding
3. External Validity
   - E.g. selection effects

Challenge 1: Construct Validity

- Def: Are you measuring what you think you are measuring?
  - Especially important operationalization of theoretical construct / new “sensor”
    (e.g. social media, linguistic construct such as empathy)
  - Measurement error (e.g. drift in accelerometer sensor)
- How to demonstrate?
  - Convergent validity: Simultaneous measures of same construct correlate
  - Discriminant validity: Doesn't measure what it shouldn’t

Big Data typically means little control over how anything was measured

Challenge 2: Internal Validity

- Def: Soundness of research design
  - Are you able to appropriately answer your research question with the right level of evidence? (e.g., correlational, causal)
- What potential selection effects / confounding are there?
  - Is data missing non-randomly?
  - Could measurement be biased across key groups? (e.g. phone steps count for women vs. pockets)
  - Does population change across multiple analyses (complicating comparisons)?

Internal Validity (cont.)

- How robust are findings across different analytic choices along the way?
  - How robust are results with respect to inclusion/exclusion of outliers?
- How many hypotheses are being tested?
  - May need to control false discovery rate (cf. Bonferroni and Benjamini-Hochberg correction)
- Are distributional / parametric assumptions valid?
  - Consider non-parametric models incl. bootstrapping

Big Data typically means observational data, convenience samples, and no pre-registration

Challenge 3: External Validity

- Def: Can findings be generalized to other situations and to other people?
- How biased is the study population?
  - Ex: “Internet Explorer users”
  - Ex: “Chrome latest beta users”
  - Ex: “Smartphone owner + health app installed”
  - Convenience samples can be WEIRD, especially motivated, lack key groups of interest, …

Big Data typically means more data, but more of the same!

Recap: Prediction is insufficient for choosing interventions, more data may not help!

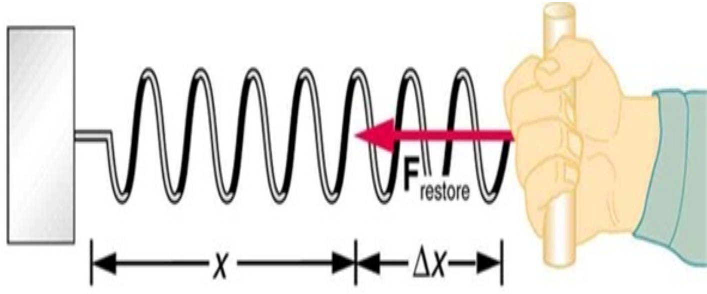
Do they lead us to the right decision? Not necessarily

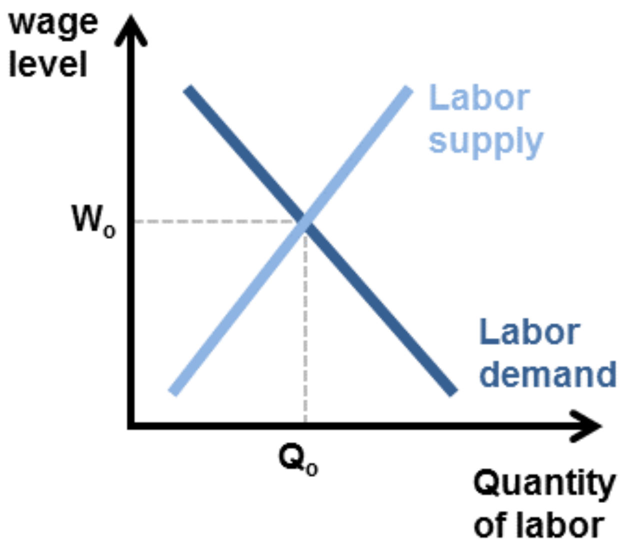
- Unclear, predictive algorithms provide no insight on effects of decisions

What is causality?

Cause and Effect

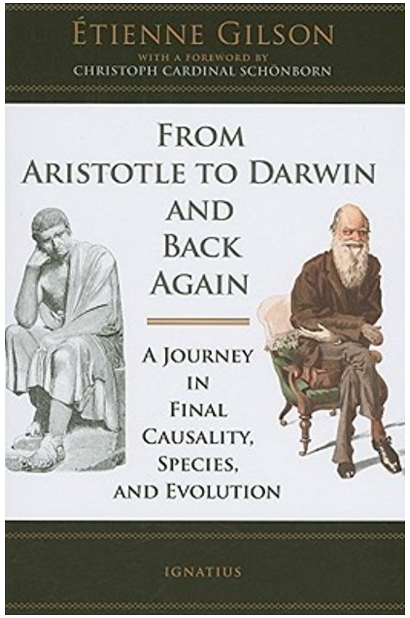
- Questions of cause and effect common in biomedical and social sciences
- Such questions form the basis of almost all scientific inquiry
  - Medicine: drug trials, effect of a drug
  - Social sciences: effect of a certain policy
  - Genetics: effect of genes on disease

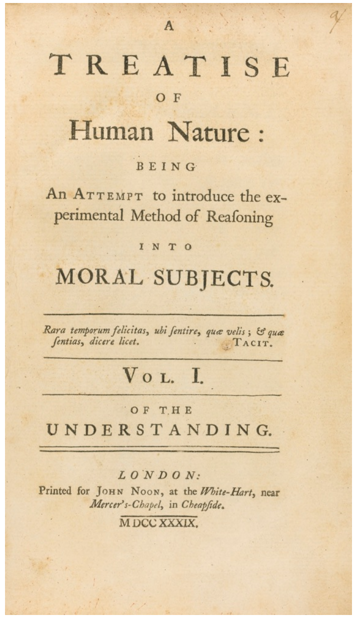


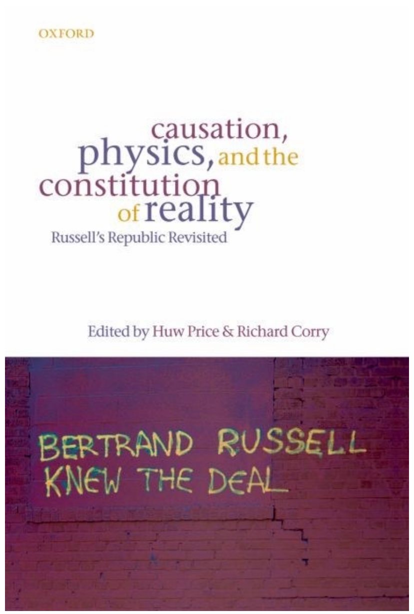


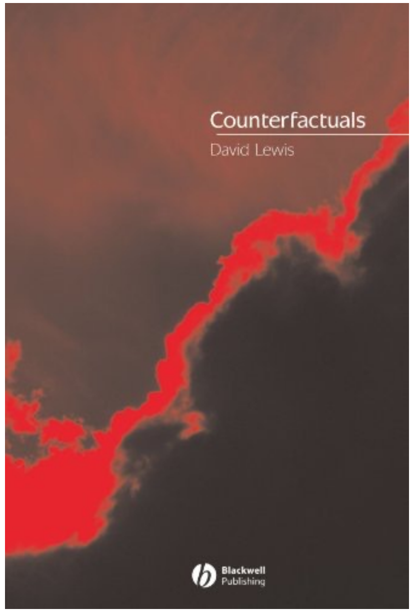
- **So what is causality?**
- **What does it mean to *cause* something?**

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 $P(y \mid x)$	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?

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 $P(y \mid x)$	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 $P(y \mid do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?

The Three Layer Causal Hierarchy

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

Level	Typical Activity	Typical Question	Examples
1. Association $P(y \mid x)$	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 $P(y \mid do(x), z)$	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 $P(y_x \mid x', y')$	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 if 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.

T is often binary, but can be categorical, ordinal, continuous.

Called the “interventionist” definition of causality.

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

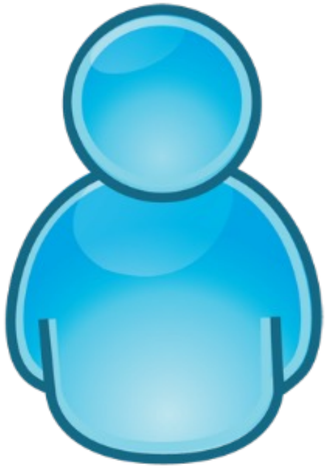
Keeping everything else constant:
Imagine a *counterfactual* world

"What-if" questions
Reason about a world that does not exist.

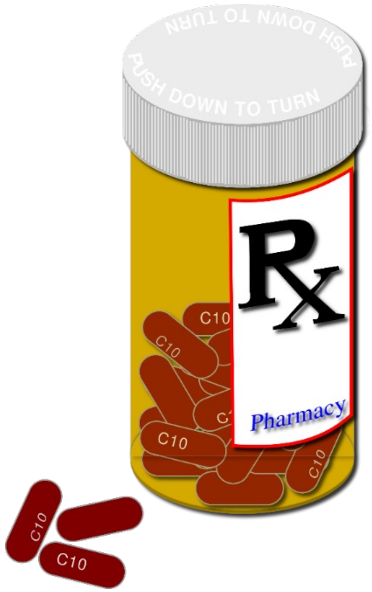
- What if a system intervention was not done?
- What if an algorithm had been changed?
- What if we give a drug to a patient?

Potential Outcomes Framework

Potential Outcomes framework

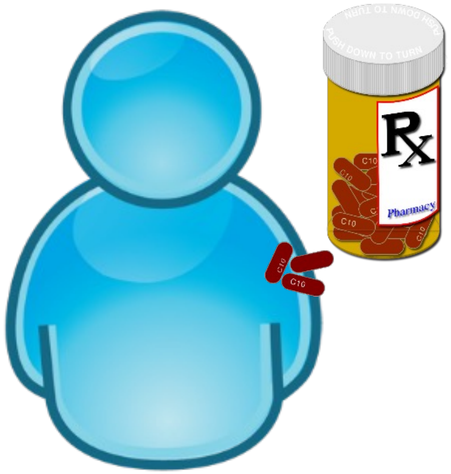


Alice



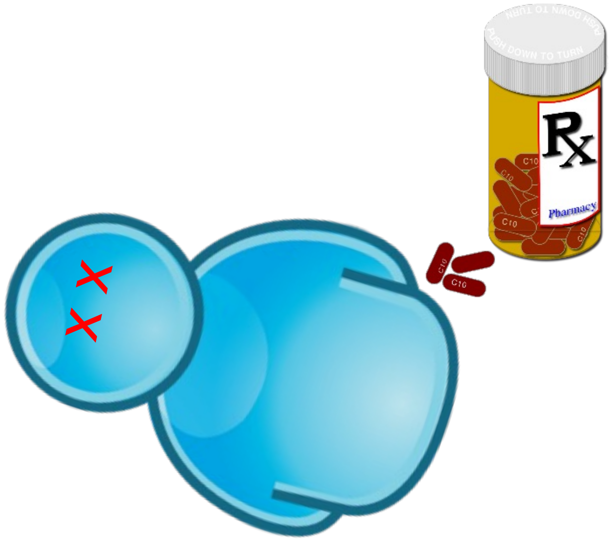
Treatment

Potential Outcomes framework



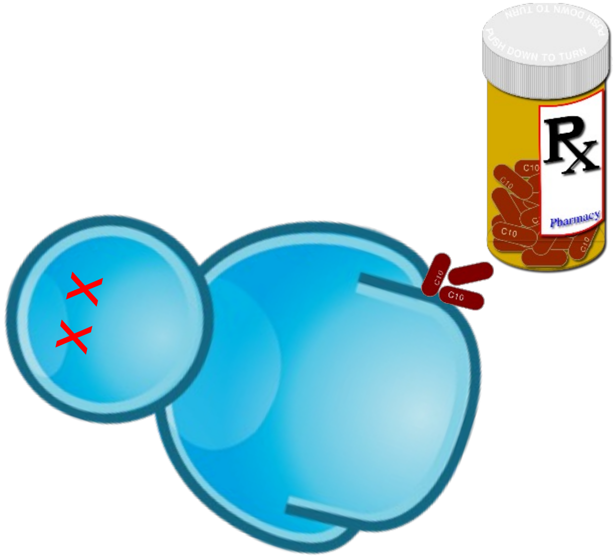
Alice

Potential Outcomes framework

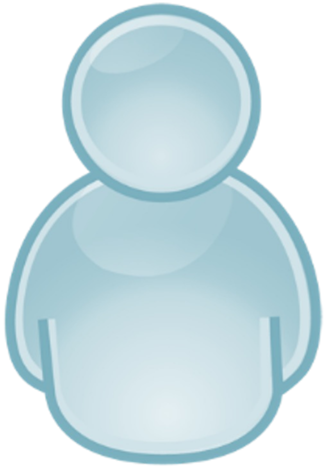


Alice

Potential Outcomes framework: Introduce a counterfactual quantity



$$Y_{T=1}$$



$$Y_{T=0}$$

Causal effect of treatment =

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

Average Treatment Effect (ATE)

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

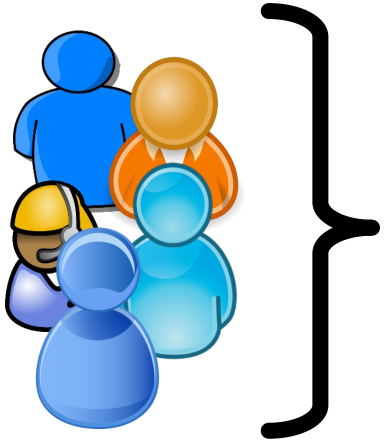
Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	0.3
P2	0	0.8	0.6
P3	1	0.3	0.2
P4	0	0.3	0.1
P5	1	0.5	0.5
P6	0	0.6	0.5
P7	0	0.3	0.1

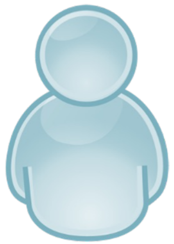
Causal effect: $E[Y_{t=1} - Y_{t=0}]$

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

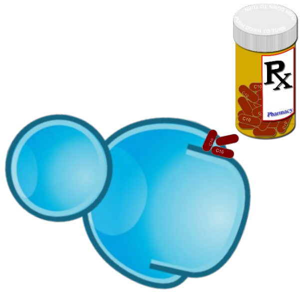
Fundamental problem: Counterfactual outcome is not observed

- Causal inference is really a “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.

Simple: sample mean difference gives you unbiased ATE estimate

Experiments are not always possible!

In many cases, we cannot randomize / intervene / A-B test (cf. offline evaluation).

- **Practicality:** Exposure to treatment may be hard to manipulate
  - Ex: Environmental effects (air pollution)
- **Ethical concerns:** Known negative effects
  - Ex: Smoking with known negative effect (today known; more on Thu)
  - Extreme Ex: Is suicide contagious?
- **Efficiency:** Experimental science is expensive and takes time
  - Ex: Studying impact on mortality 10 years later
- …

Experiments are not always possible!

In many cases, we cannot randomize / intervene / A-B test (cf. offline evaluation).

- **Practicality:** Exposure to treatment m... rd to manipulate
  - Ex: Environmental effects (...
- **Ethical concern...**
  - Ex: Smok... ...own; more on Thu)
  - Extreme E...
- **Efficiency:** E... ...l science is expensive and takes time
  - Ex: Studying impact on mortality 10 years later
- ...

What can we do when an experiment is not possible?
More on Thursday!

What causal effects might you want to estimate?

- So far: ATE – Average Treatment Effect
  - $E[Y_{T=1} \quad - Y_{T=0}]$: This is average causal effect across entire population
- ATE could be different on treated vs untreated group
  - Often we care about particular populations!
  - Ex: Hip Surgery -> Walking Ability
    - Doctors are not interested in effect of hip surgery on healthy population. What does it change for someone who has difficulty walking?
- ATT – Average Treatment Effect **on the Treated**
  - $E[\, Y_{T=1} - Y_{T=0} \mid \mathbf{T=1}]$

Recap: Potential Outcomes Framework

- **Potential outcomes** reasons about causal effects by comparing outcome of treatment to outcome of no-treatment
- **The Fundamental Problem of Causal Inference:** For any individual, we cannot observe both treatment and no-treatment.
- **Randomized experiments** are one elegant solution, but not always possible
  - We'll discuss other solutions on Thursday that rely only on observational data

Unobserved Confounds / Simpson's Paradox

Unobserved Confounds

- Which treatment should a doctor recommend for kidney stones?
- **Simpson's paradox:** After accounting for the confounder (stone size) the best choice reverses.
- Critical for decision making

Treatment A	Treatment B
78% (273/350)	**83%** (289/350)

Charig et al., BMJ 1986

Questions for you ☺

- What explains this “paradox”? Concretely, why did treatment B look so effective?
- What could researchers have done to fix this at the time?

	Treatment A	Treatment B
Small stones	**93%** (81/87)	87% (234/270)
Large stones	**73%** (192/263)	69% (55/80)
Both	78% (273/350)	**83%** (289/350)

Charig et al., BMJ 1986

Recap: Unobserved Confounds

- Unobserved confounds are a threat to causal reasoning and to decision making

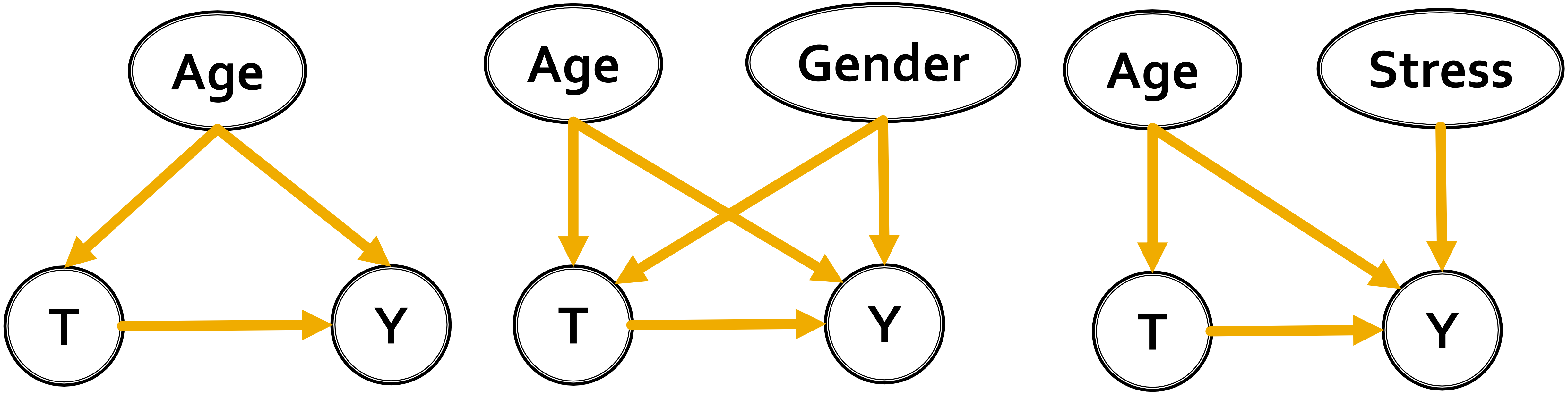
Structural Causal Model Framework

The real world is complicated

- We observed that accounting for kidney stone size was critical
- Many other factors might influence the observed outcome
  - How do they affect treatment and outcome?
  - Which ones to include?

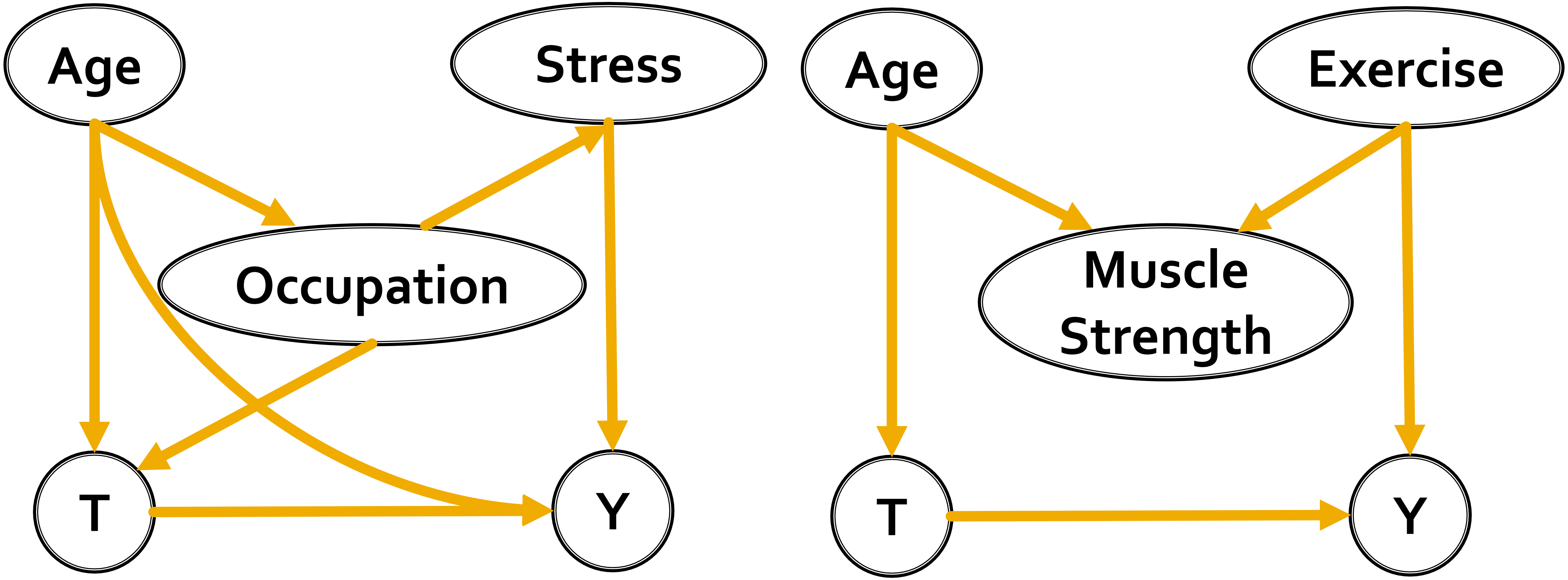
- How to we formalize all of our **assumptions?**
  - Causal inference cannot be done with data alone. It requires making assumptions about the world.
- How to **identify** the causal effect in such cases?
  - The task of causal **identification** is to determine an expression, the causal estimand, that expresses our target value as a function of the observable correlational relationships in our system.
  - We can use graphical model framework to answer this

Which variables to condition on?



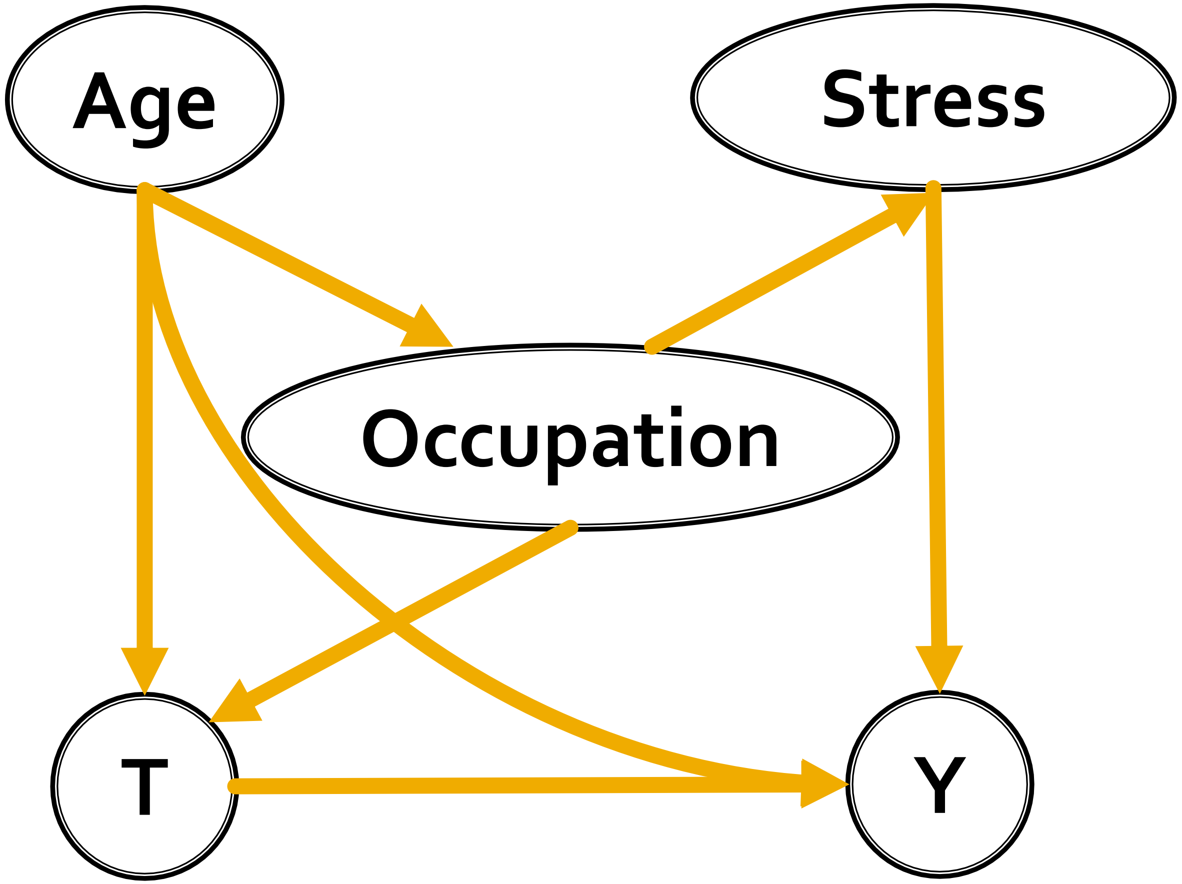
$\boldsymbol{X} = \{Age\}$ $\boldsymbol{X} = \{Age, Gender\}$ $\boldsymbol{X} = \{Age\}$

What about these?



X = ? X = ?

Structural Causal Model: A framework for expressing complex causal relationships



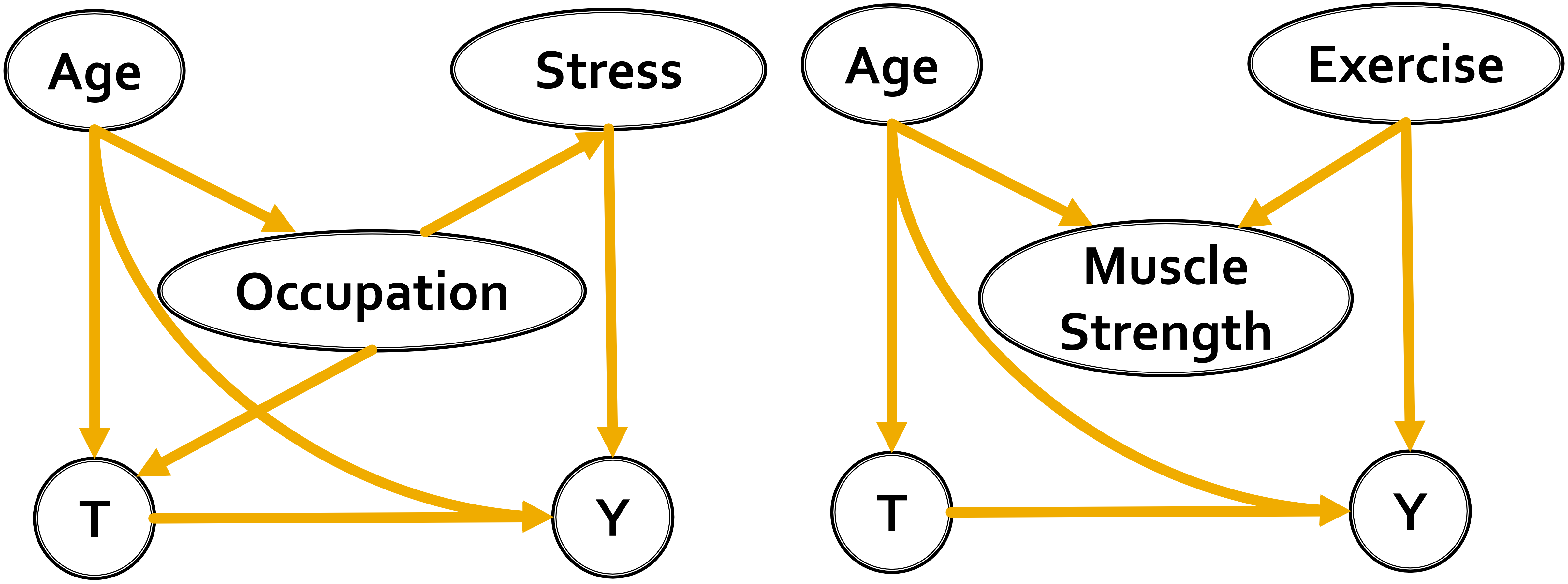
Structural Equation Models with Random Errors
u's are "error variables" or "exogenous variables"

$$Occupation = h(Age, u_o)$$
$$Stress = k(Occupation, u_s)$$

$$T = g(Age, Occupation, u_t)$$
$$Y = f(T, Age, Stress, u_y)$$

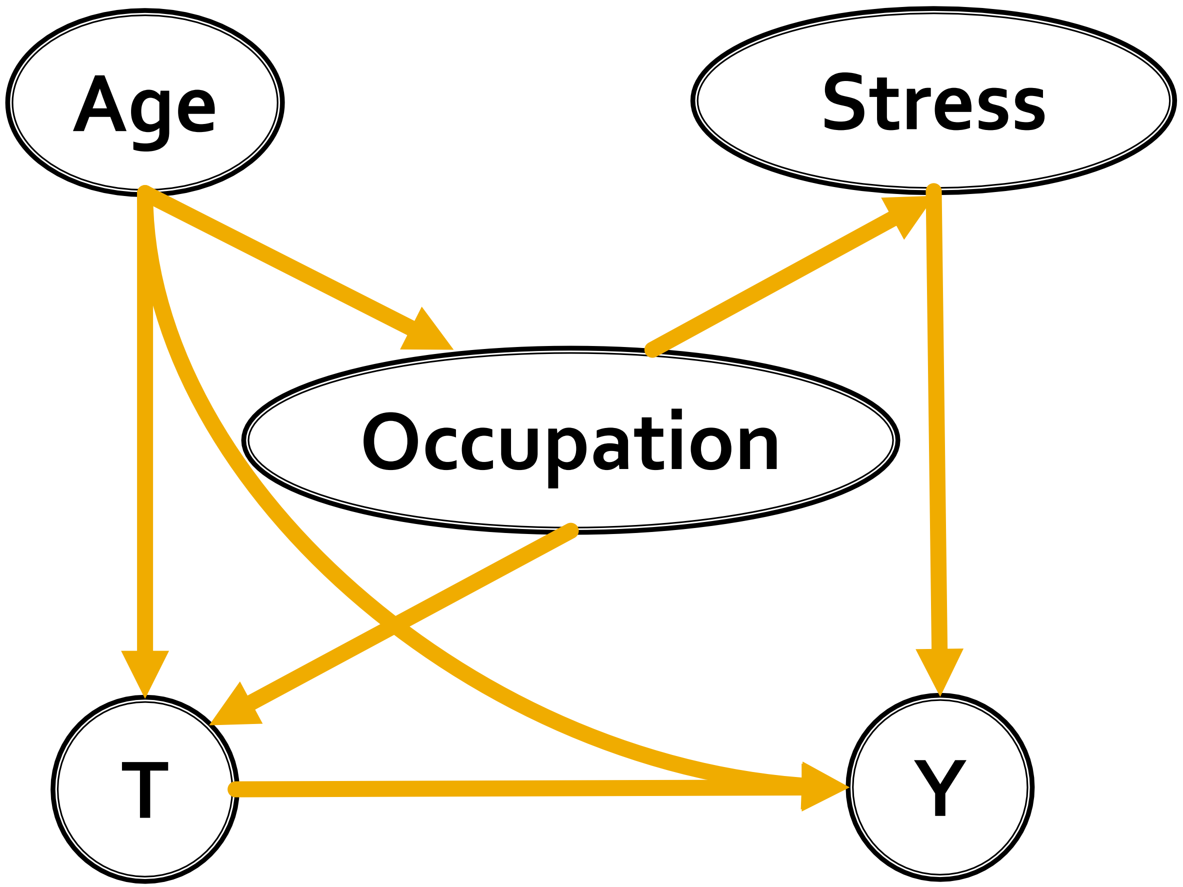
Edges represent *direct* causes.
Directed paths represent *indirect* causes.

Structural Causal Model makes assumptions explicit



The graph encodes all causal assumptions.

Important: Assumptions are the edges that are *missing*



Assumption 1: Occupation does not directly affect outcome Y.
Assumption 2: Age does not directly affect stress.
Assumption 3: Stress does not directly affect Occupation.
Assumption 4: Treatment does not directly affect stress.

..and so on.

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

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?

$$\boldsymbol{P(Y = 1 | T = 1, T = 0)}$$

Non-sensical.

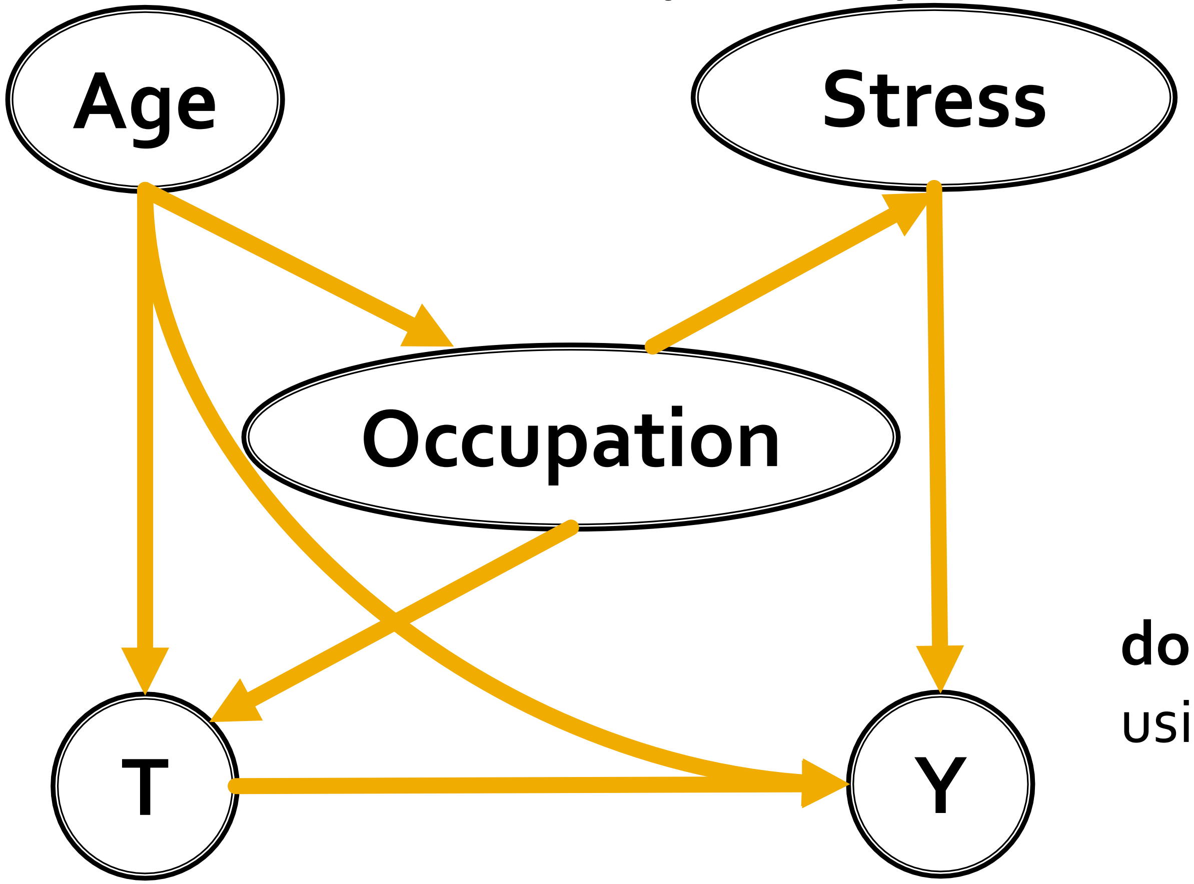
Can write it as:

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

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



E.g.,

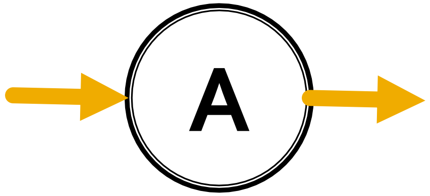
$$P(Y|do(T))$$
$$= \cdots do\text{−}calculus\ rules\ \ldots$$
$$= \sum_{\boldsymbol{x}} P(Y|T, \boldsymbol{X} = x)\, P(\boldsymbol{X} = \boldsymbol{x})$$

do-calculus is complete: If we cannot identify using do-calculus, causal effect is unidentifiable.

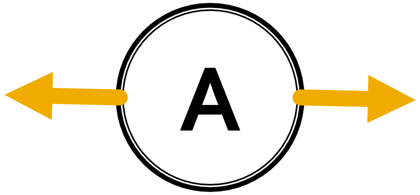
Advanced Topic: Back-door criterion

Three kinds of node-edges

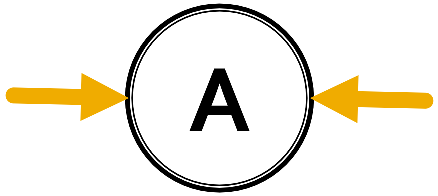
Path is "blocked"



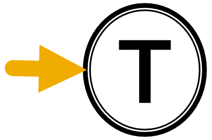
If conditioned on A

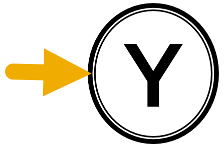


If conditioned on A



If **not** conditioned on A

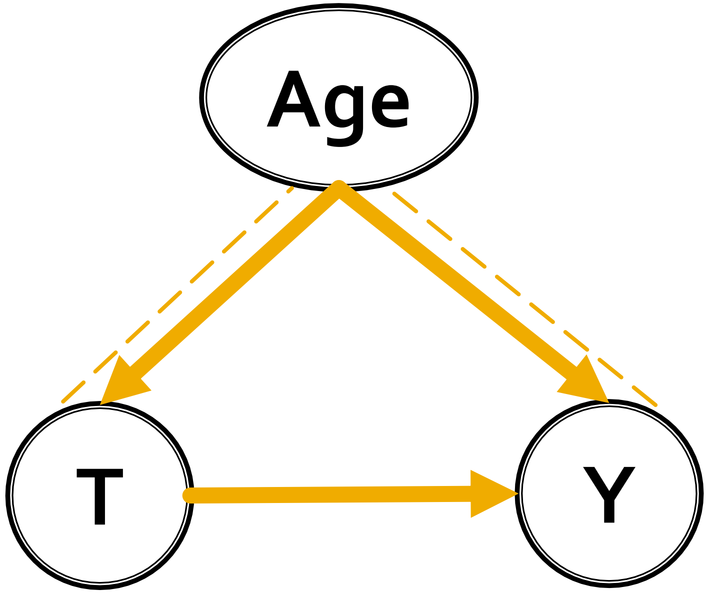
"Back-door" path: Any undirected path that starts with 

and ends with 

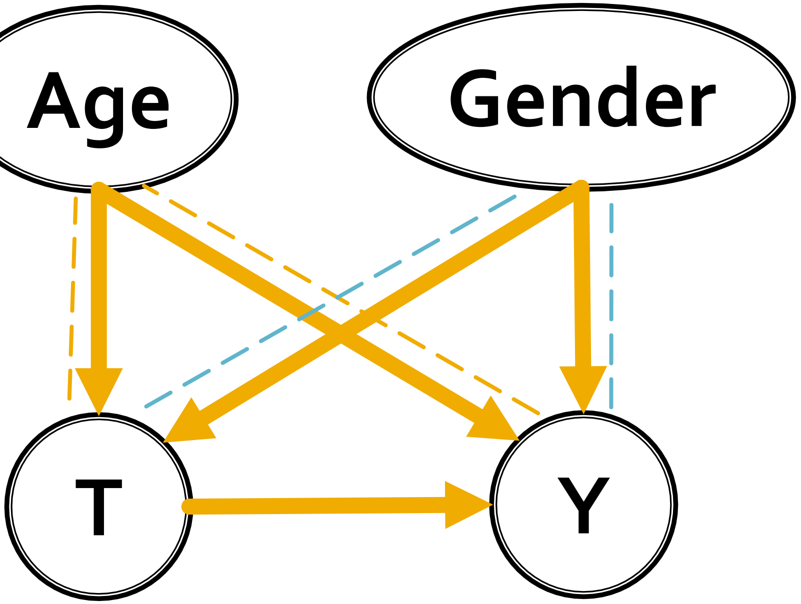
Back-door criterion: If conditioning on X blocks all back-door paths between treatment T and outcome Y, and X does not include any descendants of T, then

$$\boldsymbol{P(Y|do(T)) = \sum_{x} P(Y|T, X = x)P(X = x)}$$

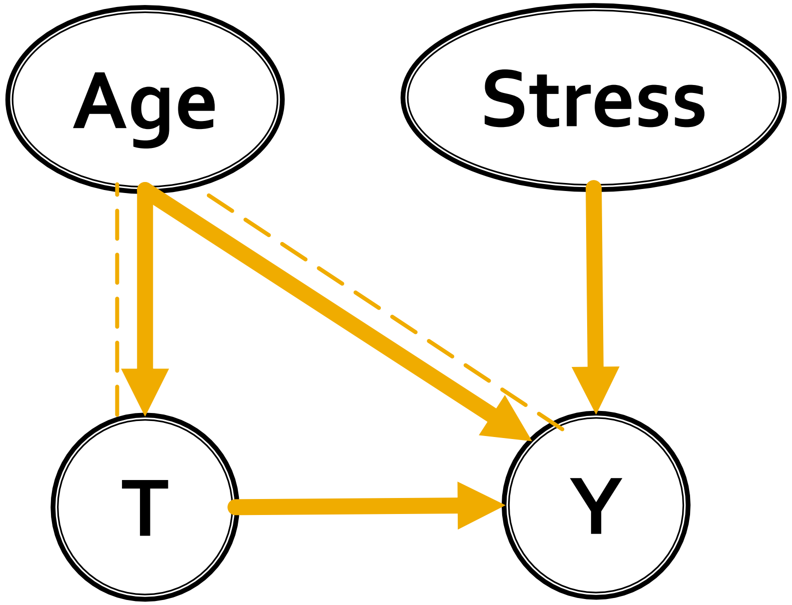
Back-door criterion provides a precise way to find the set of variables to condition on



$\boldsymbol{X} = \{Age\}$

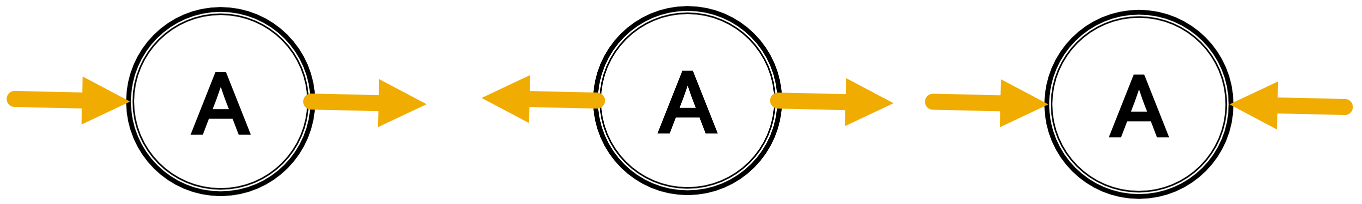


$\boldsymbol{X} = \{Age, Gender\}$



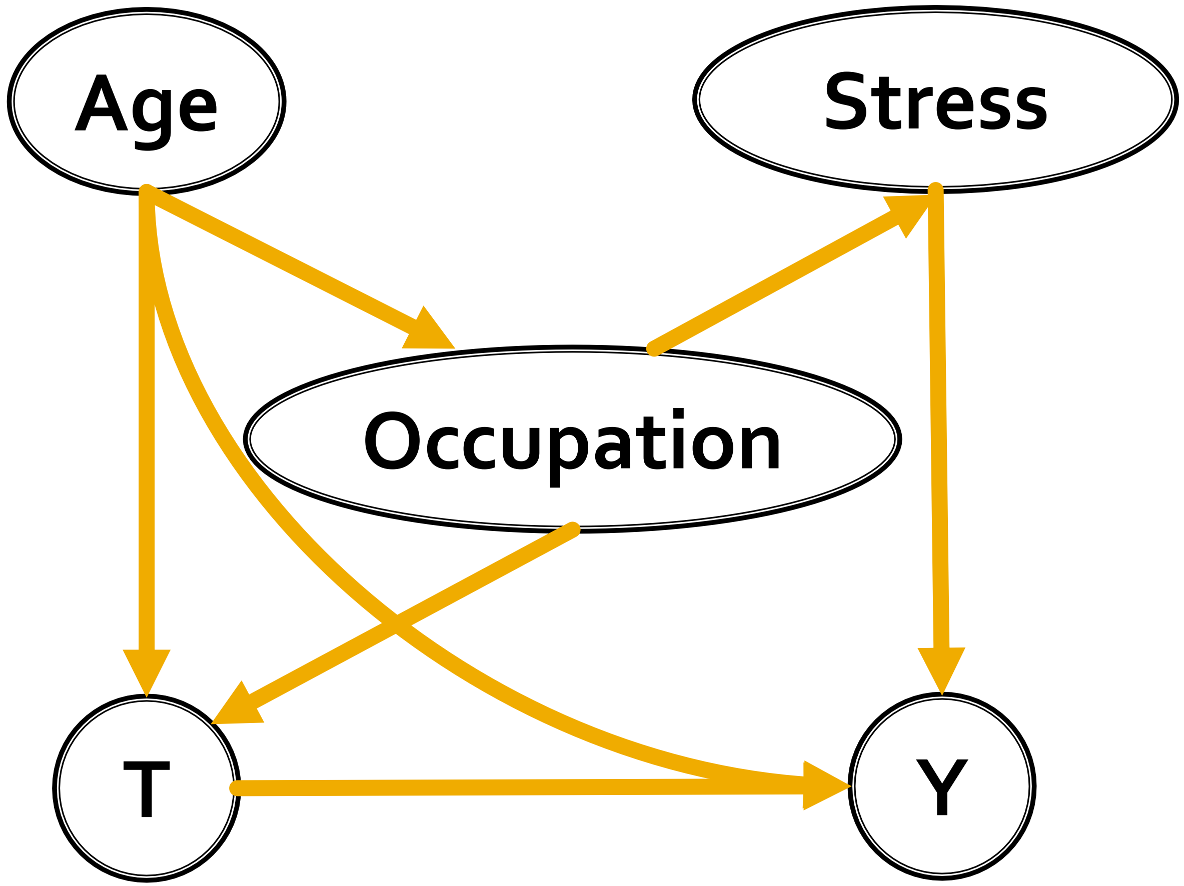
$\boldsymbol{X} = \{Age\}$

Now it's your turn!

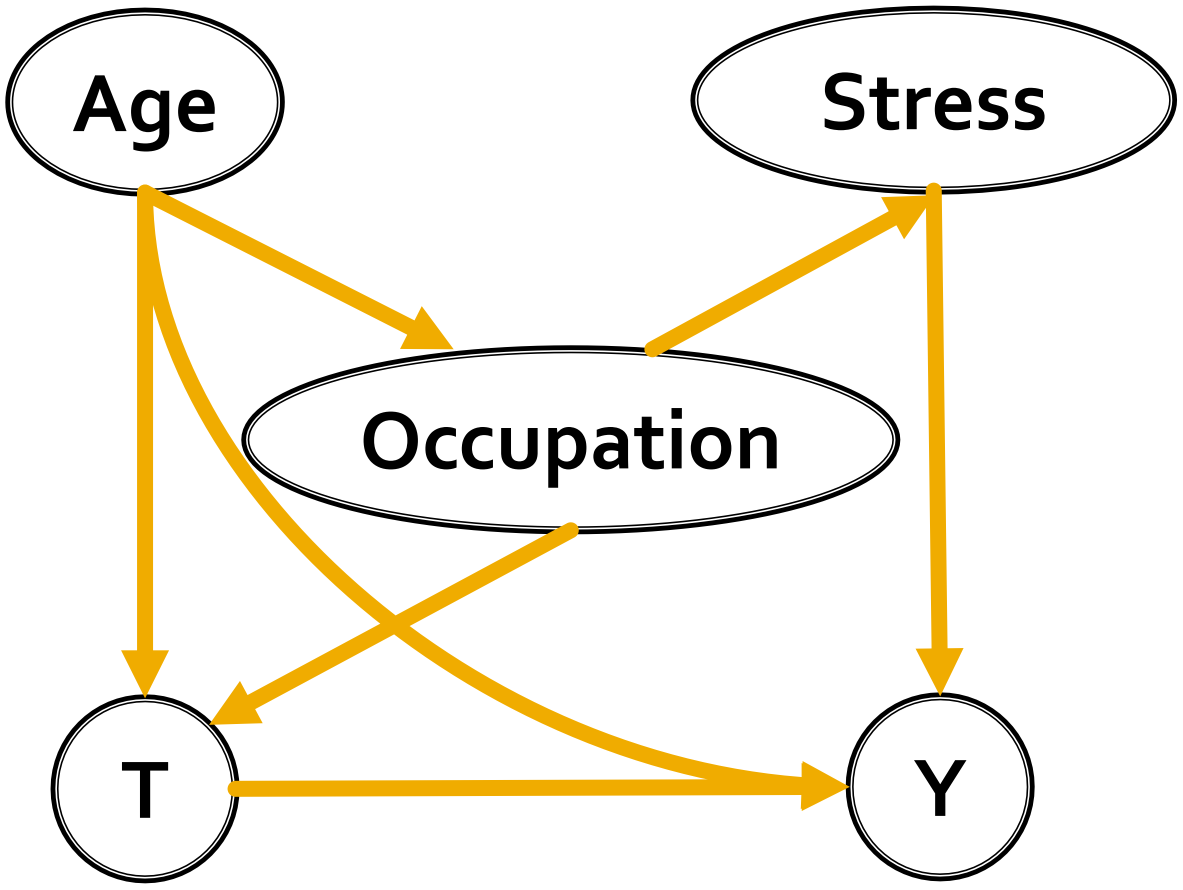


If conditioned on A | If conditioned on A | If **not** conditioned on A

- Find a partner close to you.
  - If watching recording: I invite you to pause now and try it out.
- **Q:** What variables do we need to condition on to block all backdoor paths?
- **Too easy?** Find all such sets

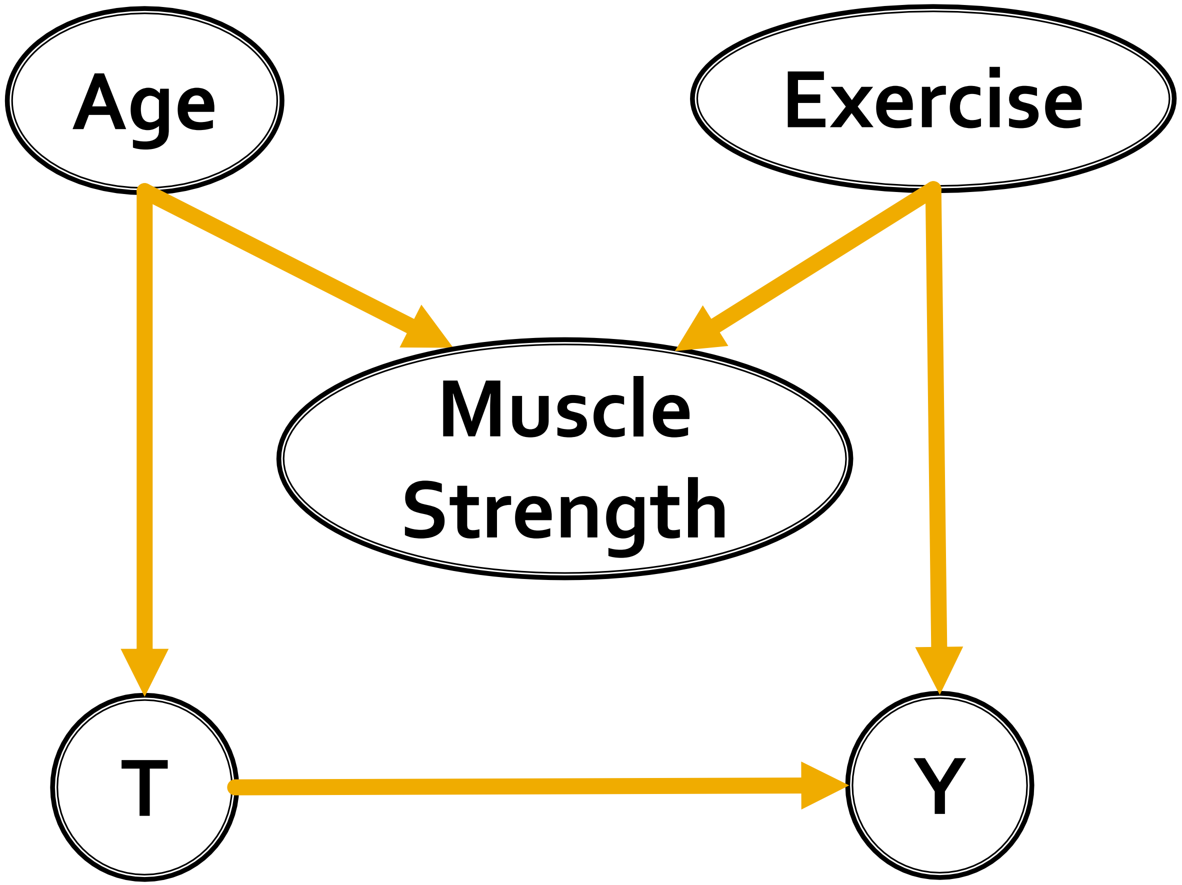


Correct Answer



X = {Age, Stress}
X = {Age, Occupation}
X = {Age, Stress, Occupation}

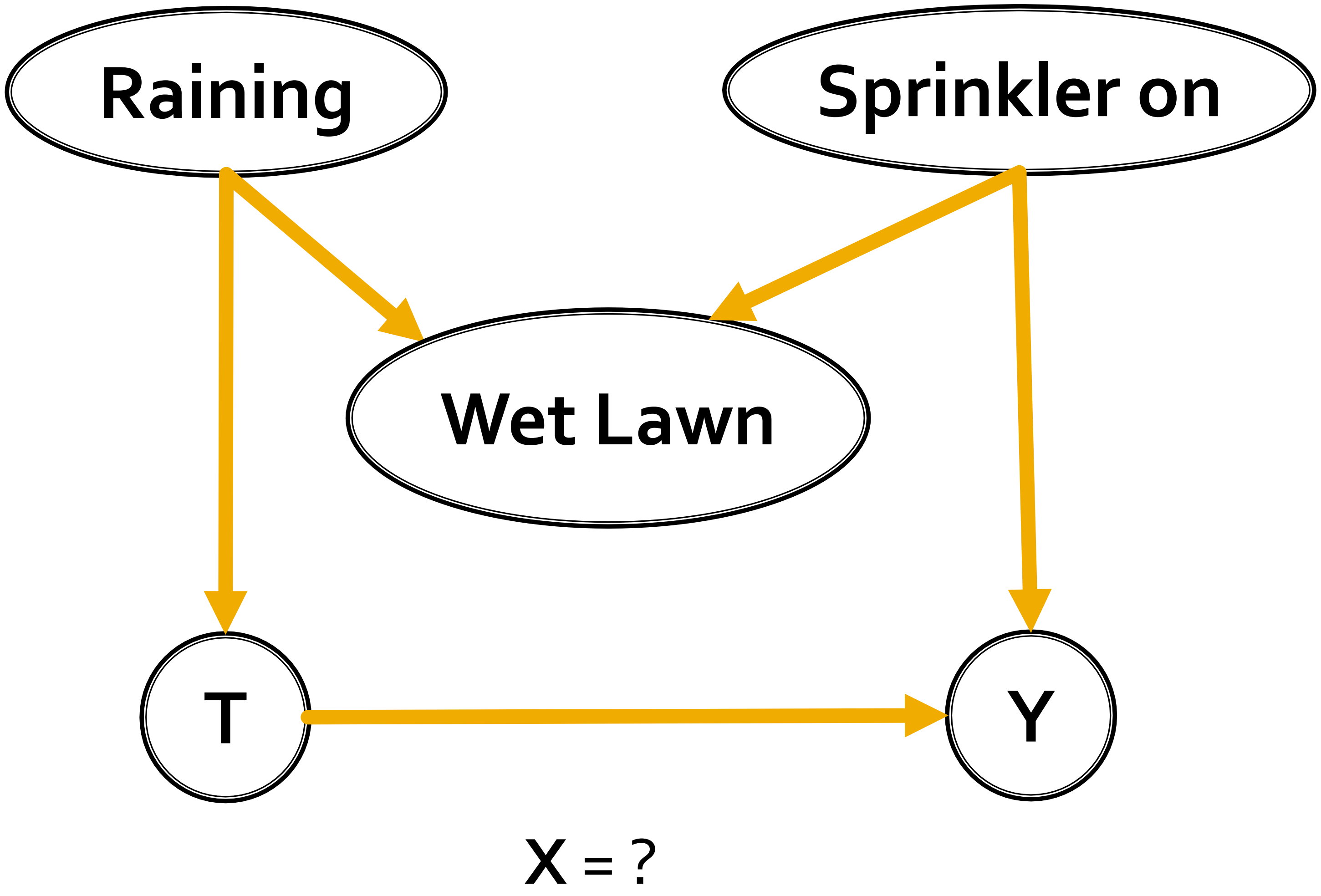
Next example:



X = {} – Muscle Strength is a collider!
X = {MuscleStrength, Exercise}
X = {MuscleStrength, Age}
X = {MuscleStrength, Age, Exercise}

It is important to know about colliders! Controlling for more variables can even hurt!

Special Case: Colliders



Q: Why are colliders special and we should not condition on them to block the backdoor path?

A: If you know that the lawn is wet, but sprinkler is not on. Now you know it must be raining.

Conditioning on collider induces correlation between parent nodes.

Why is computing conditional probability not enough?

- **Back-door criterion:** If conditioning on X blocks all back-door paths between treatment T and outcome Y, and X does not include any descendants of T, then

$$\boldsymbol{P(Y|do(T)) = \sum_{x} P(Y|T, X = x)P(X = x)}$$

- **This is a correct, unbiased estimator. But it can be so inefficient, essentially requiring "infinite" data, that it loses all practical utility.**
  - This happens when **X** is high-dimensional and/or incl. continuous variables
  - Example: Say **X** is lots of sentences in the English Language and we want to identify the causal effect of my advising on how well my students are doing?
  - Example: **X** is your entire genome. How often will you observe this?
  - Example: **X** is continuous dosage amount of a drug

- More on Thu about effective estimation techniques
  - We will carefully make **bias variance trade-offs** that allow for more efficient estimation

Let’s do the course evaluation right now

- https://uw.iasystem.org/survey/290650
  - Also available from Ed link.
- We take your feedback very seriously. It makes a difference!
- Consider sharing what you liked about the course. What is working well? What should we keep?
- Thank you for participating!

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

Both PO & SCM 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 of today:

- Causality is important for decision-making and study of effects
- Big Data does not necessarily address threats to causal inference
- Potential Outcomes Framework gives practical method for estimating causal effects. Translates causal inference into counterfactual estimation
- Unobserved confounds are a critical challenge – what might we miss?
- Structural Causal Model Framework gives language for expressing and reasoning about causal relationships: Causal **Identification**
- On Thursday: Methods for efficient **Estimation** of causal treatment effects in observational data

Additional Resources

- Slides based in part on KDD 2018 Tutorial by Emre Kıcıman and Amit Sharma: http://causalinference.gitlab.io/kdd-tutorial/

- Courses
  - UW Econ 488: Causal Inference
  - UW Stat 566: Causal Modeling

- Books
  - Pearl. Book of Why
  - Rosenbaum. Design of Observational Studies
  - Kiciman & Sharma. https://causalinference.gitlab.io/ (free, in-progress)

Please give us feedback ☺

https://bit.ly/CSE547feedback2024