Announcements:

- **Zoom (check Ed for links and schedule), 10:00am-1:00pm**
- Great opportunity to learn about each other's projects
- Attendance is mandatory
- Participation rewarded with extra credit
- **Upload your deliverables on Gradescope by Sunday 23:59pm – no late periods**
  - Project Report
  - Presentation Video (and slides PDF if possible)
  - Metadata (e.g. dataset info)

Causal Inference I

Introduction to Counterfactual Reasoning

CS547 Machine Learning for Big Data

Tim Althoff

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Announcement

- Course evaluation is out
  - https://uw.iasystem.org/survey/224031
  - Also see link on Ed (pinned)
  - Please fill out the form before June 7. Thanks!!!
- We appreciate your feedback!

Overview of this week's lectures

- Overview of causal inference and counterfactual reasoning
- Slides based on KDD 2018 Tutorial by Emre Kıcıman and Amit Sharma: http://causalinference.gitlab.io/kdd-tutorial/
- Additional resources
  - 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)

Plan for today: Introduction to Counterfactual Reasoning

- When is prediction / big data not enough?
- What is causality?
- Potential Outcomes Framework
- Unobserved Confounds & Simpson's Paradox
- Structural Causal Model Framework

When is prediction / big data not enough?

Prediction is everywhere!

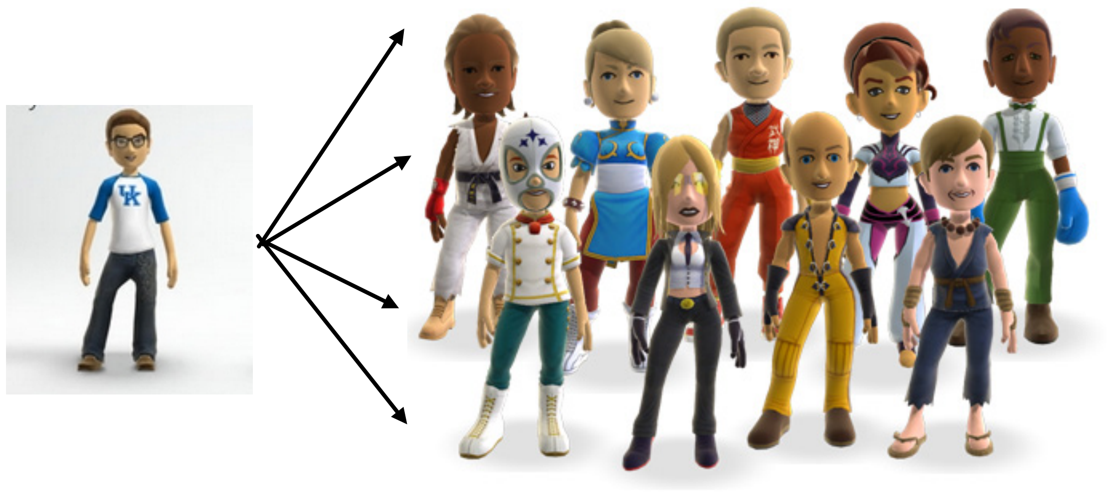
- Recommender Systems
- Social Networks
- …

- We have increasing amounts of data and highly accurate predictions! Why do we need causal inference?

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 -> f(items in social feed) + ϵ

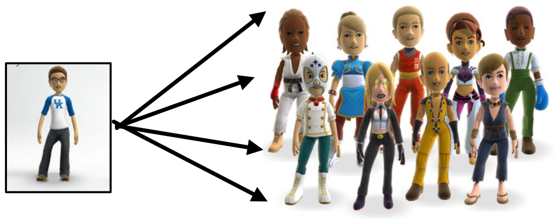
Highly predictive model.

Does it mean that feeds are influencing us significantly?

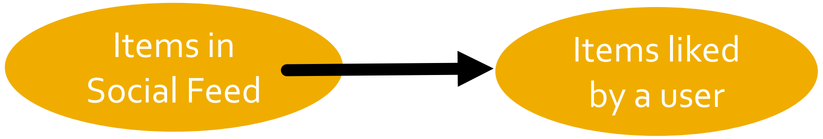
From prediction to decision-making

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

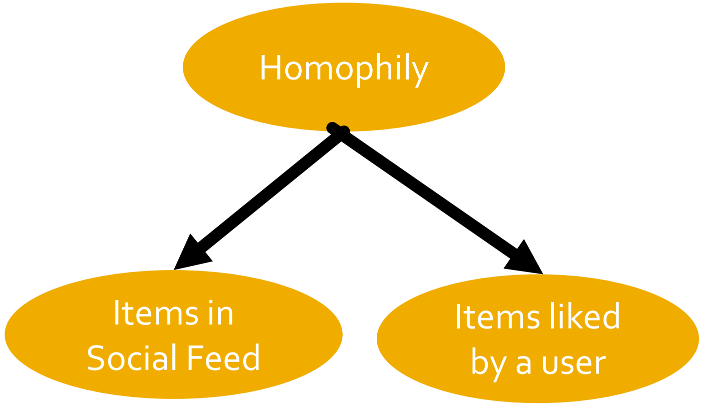
Maybe, maybe not (!)



Predictability due to
feed influence

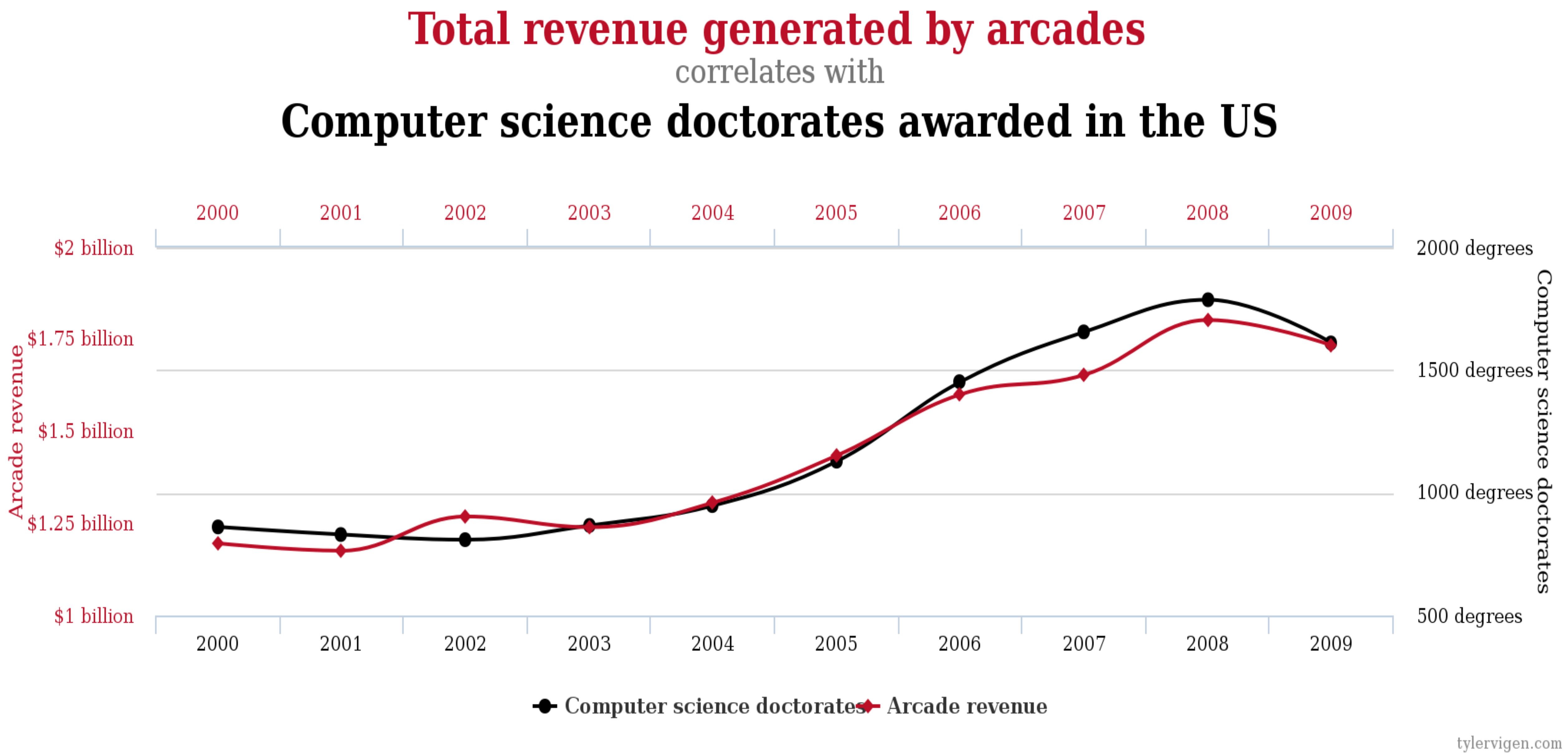


Predictability due to
homophily

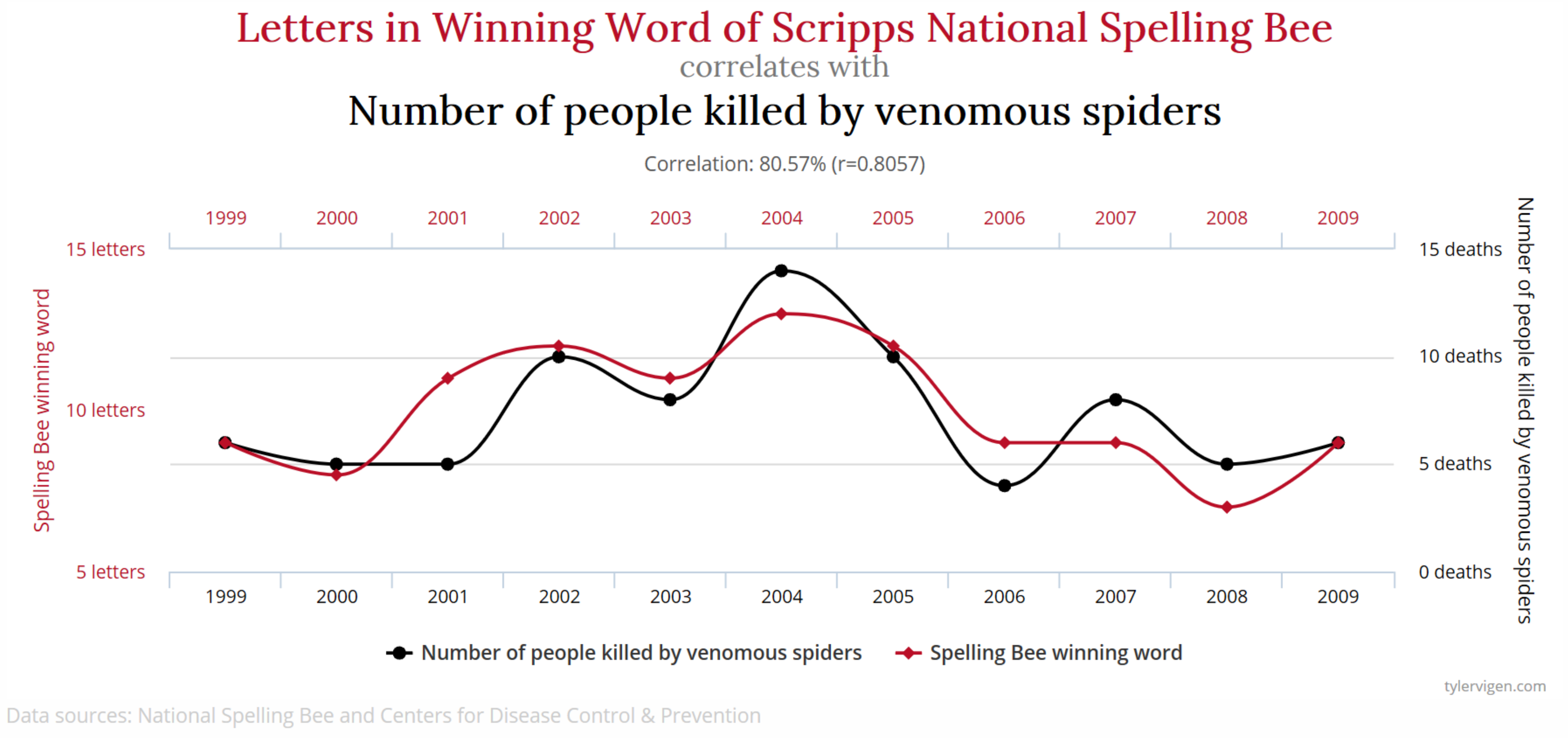


Friends' activity can predict a person's activity with high accuracy.
But that tells us *nothing* about the effect of the social feed.

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



http://www.tylervigen.com/spurious-correlations

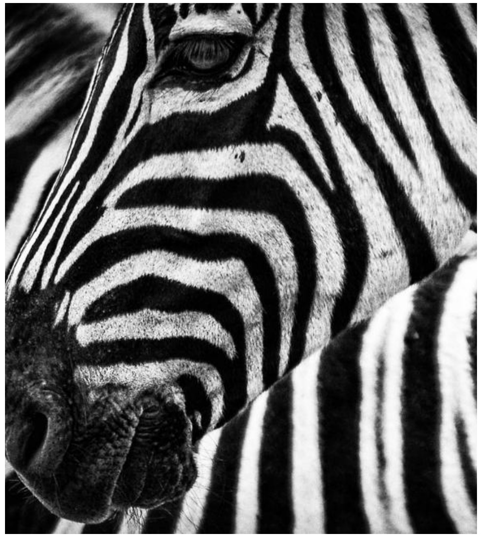


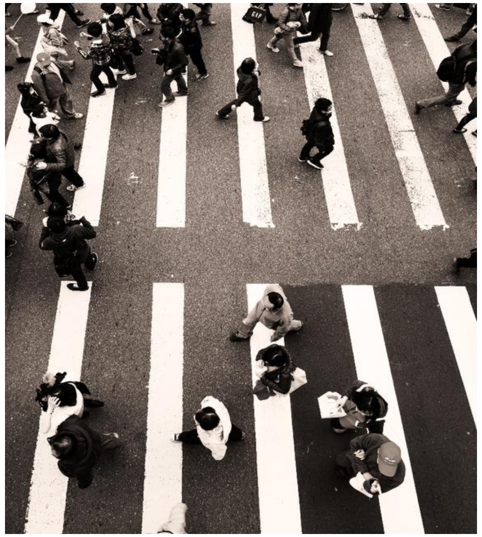
Letters in Winning Word of Scripps National Spelling Bee
correlates with
Number of people killed by venomous spiders
Correlation: 80.57% (r=0.8057)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
15 letters
10 letters
5 letters
Spelling Bee winning word
15 deaths
10 deaths
5 deaths
0 deaths
Number of people killed by venomous spiders
Number of people killed by venomous spiders
Spelling Bee winning word
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].





4) 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 need to worry about bias and variance!
- **More data does not help you reduce bias!**
- **Today: Sources of bias, how to model it, & what to do about it**

The Reasonable Uneffectiveness of Big Data

- “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
  - Measurement error, confounding, and selection bias common threats to causal inference, are independent of sample size
  - When we can’t observe counterfactuals, observing more data will not help us!

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 proxy)
- 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
- What potential selection effects / confounding are there?
  - Is data missing non-randomly?
  - Could measurement be biased across key groups?
  - Does population change across multiple analyses (complicating comparisons)?

Internal Validity (cont.)

- How robust are findings across different 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
- Are distributional / parametric assumptions valid?
  - Consider non-parametric models and 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!

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 are more robust

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

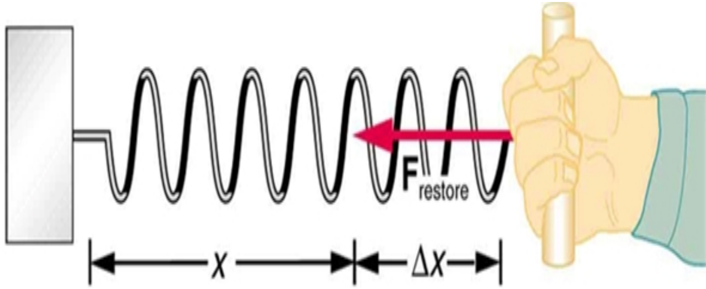
- Active interventions change correlations

Does Big Data save us?

- More data doesn't necessarily help.
- Consider construct, internal and external validity when answering questions through data.

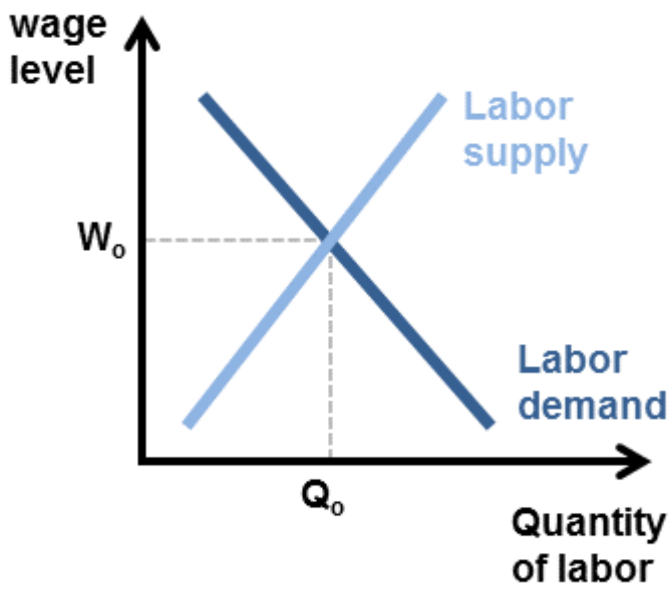
What is causality?

Cause and Effect

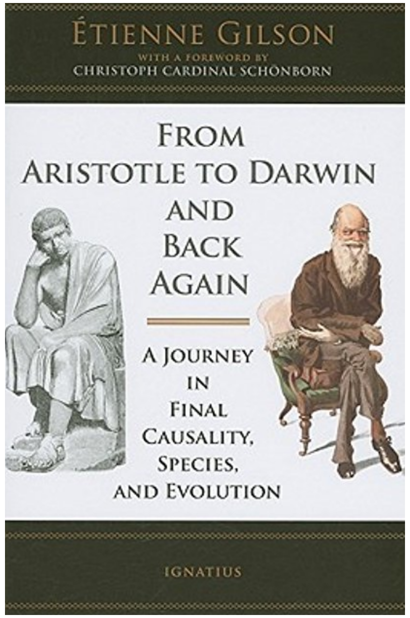


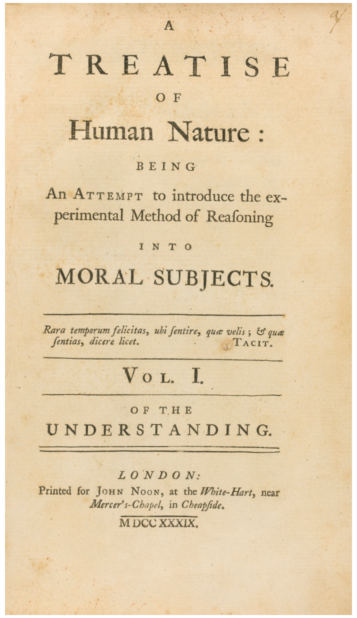
- Questions of cause and effect common in biomedical and social sciences
- Such questions form the basis of almost all scientific inquiry
  - Medicine: drug trials, effect of a drug
  - Social sciences: effect of a certain policy
  - Genetics: effect of genes on disease

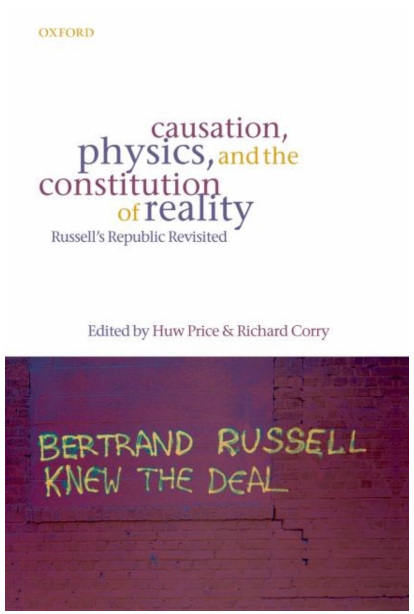
- **So what is causality?**
- **What does it mean to *cause* something?**

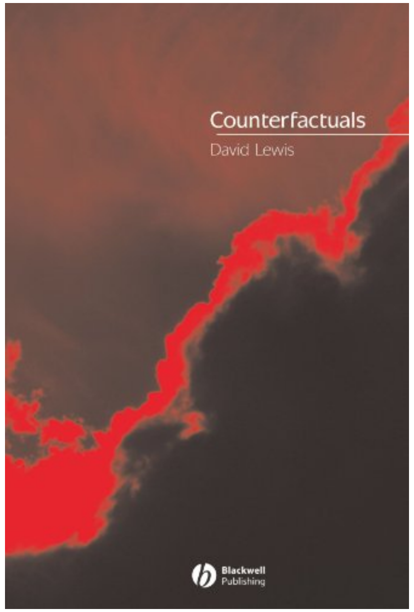


A big scholarly debate, from Aristotle to Russell









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. 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?
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.

Called the "interventionist" interpretation of causality.

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

Keeping everything else constant:
Imagine a *counterfactual* world

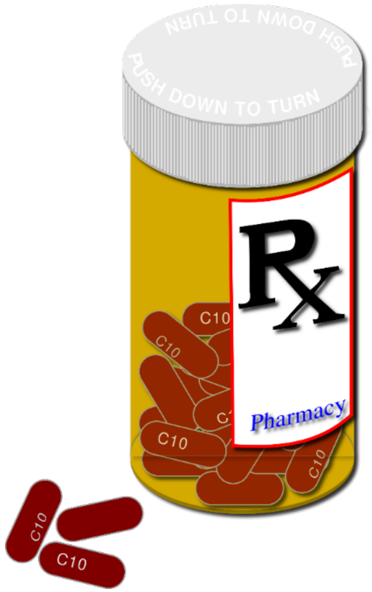
"What-if" questions
Reason about a world that does not exist.


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

Potential Outcomes Framework

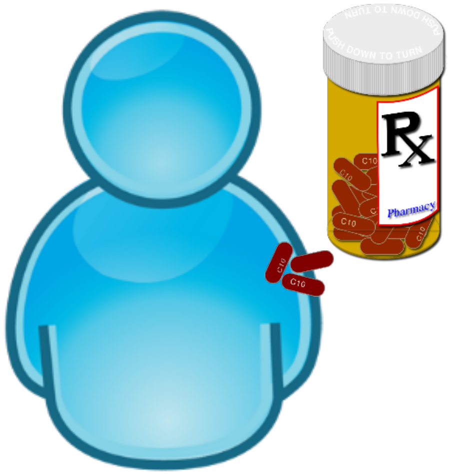
Potential Outcomes framework

Alice



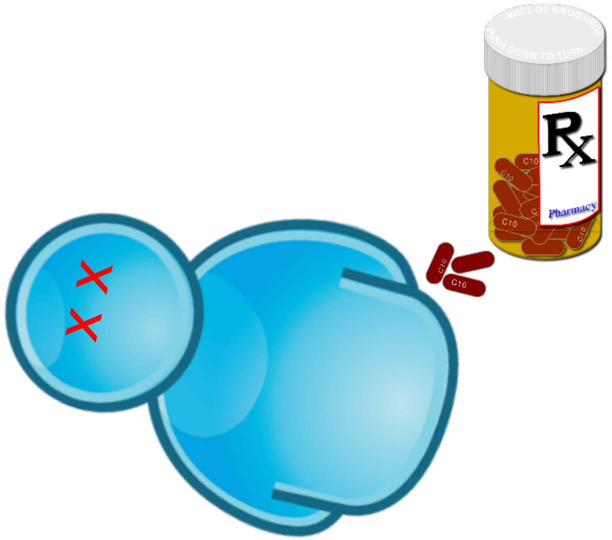
Treatment

Potential Outcomes framework



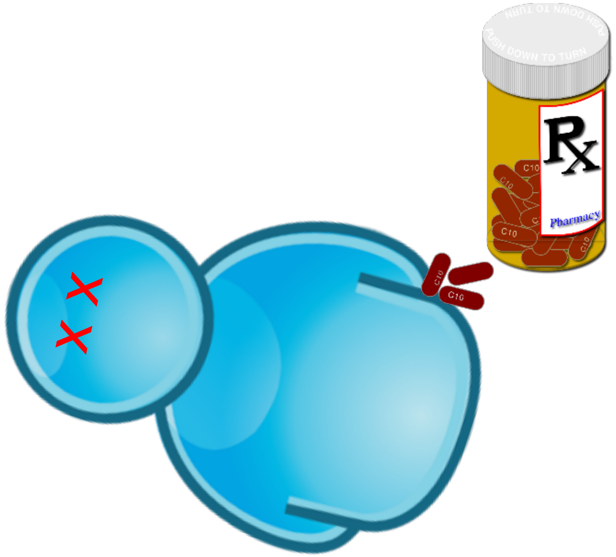
Alice

Potential Outcomes framework

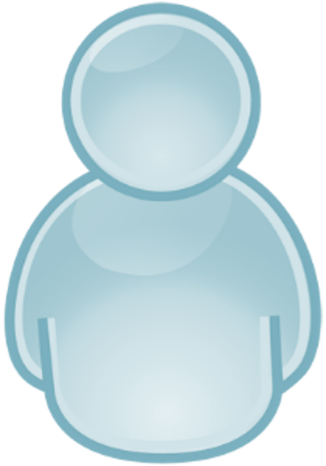


Alice

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 $Y_{t=\sim t}$

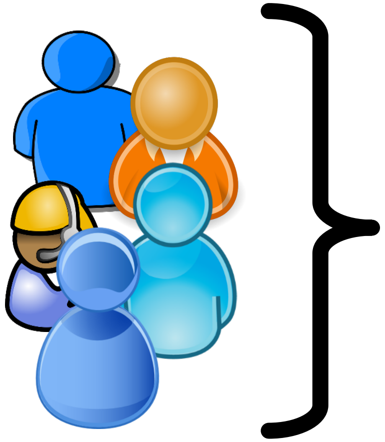
Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	0.3
P2	0	0.8	0.6
P3	1	0.3	0.2
P4	0	0.3	0.1
P5	1	0.5	0.5
P6	0	0.6	0.5
P7	0	0.3	0.1

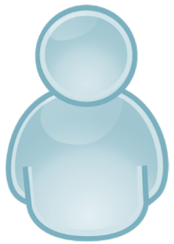
Causal effect: $E[Y_{t=1} - Y_{t=0}]$

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

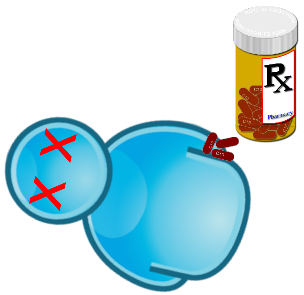
Fundamental problem: counterfactual outcome is not observed

- "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





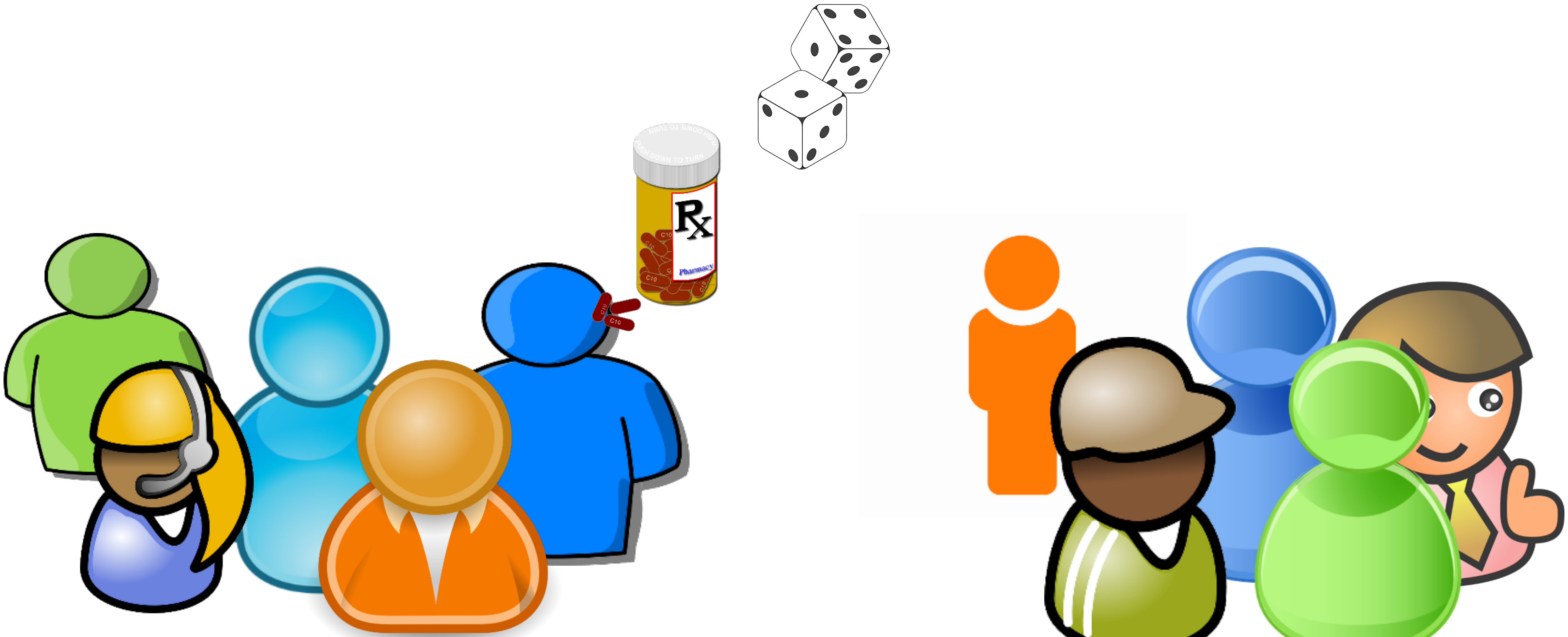
$$\hat{Y}^{T=0}$$



$$Y^{T=1}$$

Randomized Experiments are the "gold standard"

One way to estimate counterfactual



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: 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 may be hard to manipulate
  - Ex: Environmental effects (ai
- **Ethical concerns:** K
  - Ex: I
- **Efficie** xpensive and ta
  - Ex: Stu mpact 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?

- Before: 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
  - Ex: Special Job Training -> Average Annual Earning
    - Not everyone needs that job training – Policymakers may be interested only in effect on low income population.
  - 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?
  - Often we care about particular populations!

- ATT – Average Treatment Effect **on the Treated**
  - $E[Y_{T=1} \quad - Y_{T=0}$ | **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

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

Recap: Unobserved Confounds

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

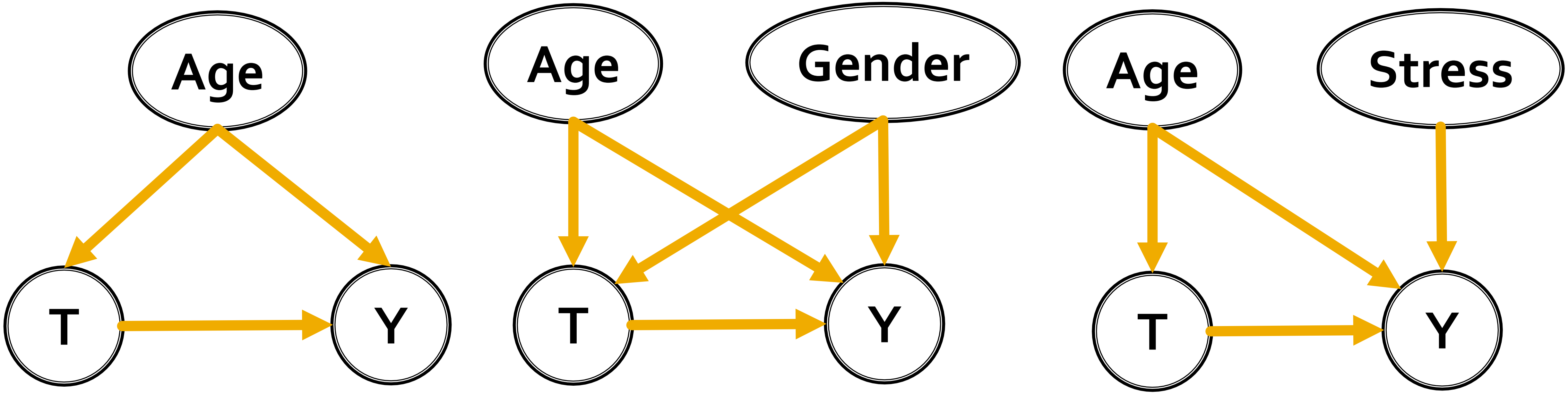
Structural Causal Model Framework

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?

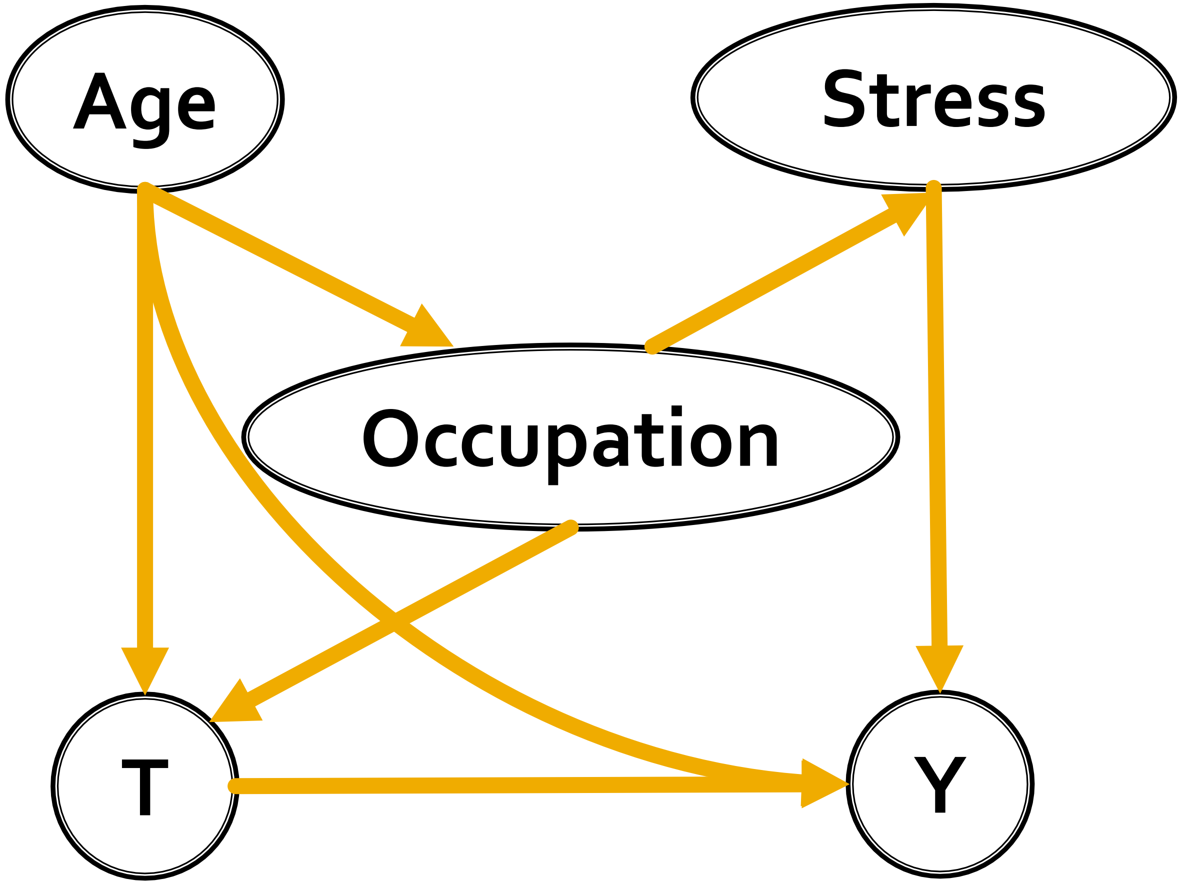
- When is it possible to find a causal effect?
  - We can use graphical model framework to answer this

Which variables to condition on?

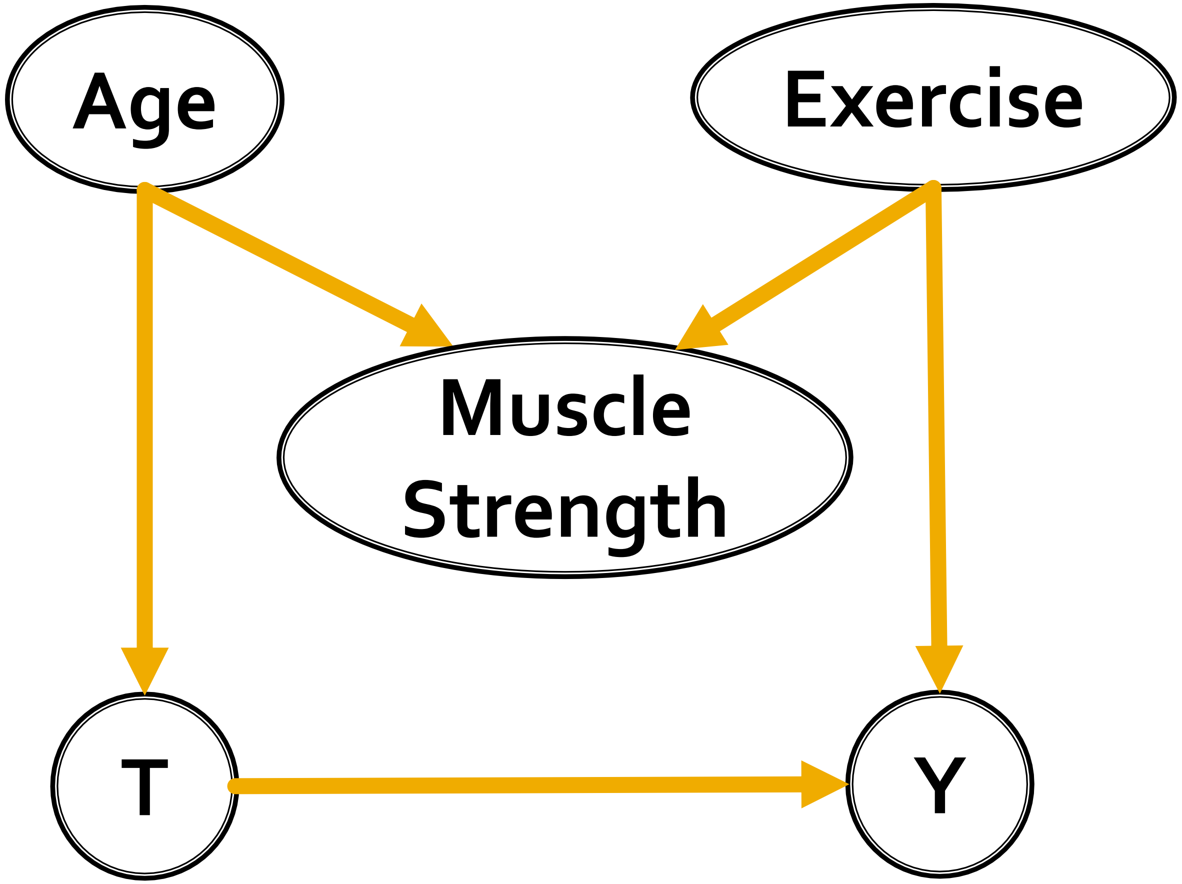


$\boldsymbol{X} = \{Age\}$ $\boldsymbol{X} = \{Age, Gender\}$ $\boldsymbol{X} = \{Age\}$

What about these?

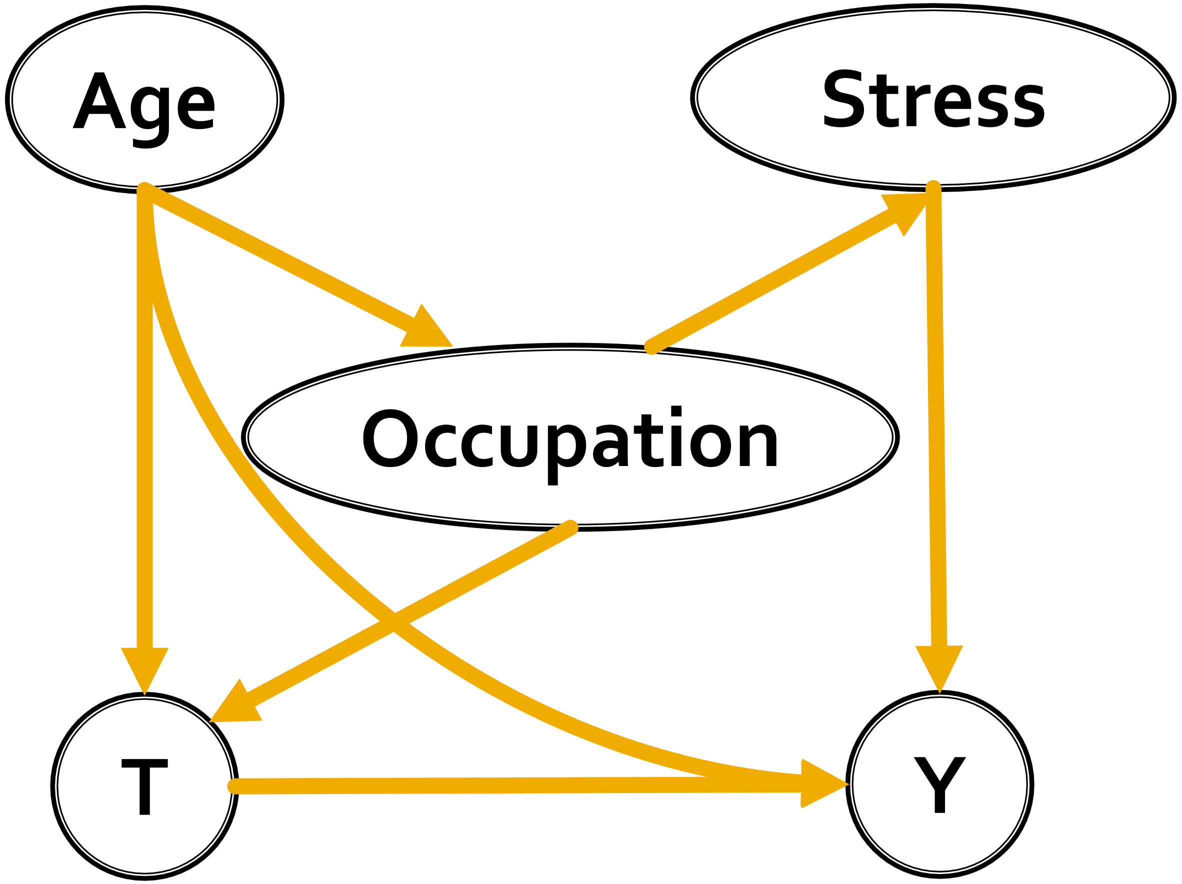


$$X = ?$$



$$X = ?$$

Structural Causal Model: A framework for expressing complex causal relationships



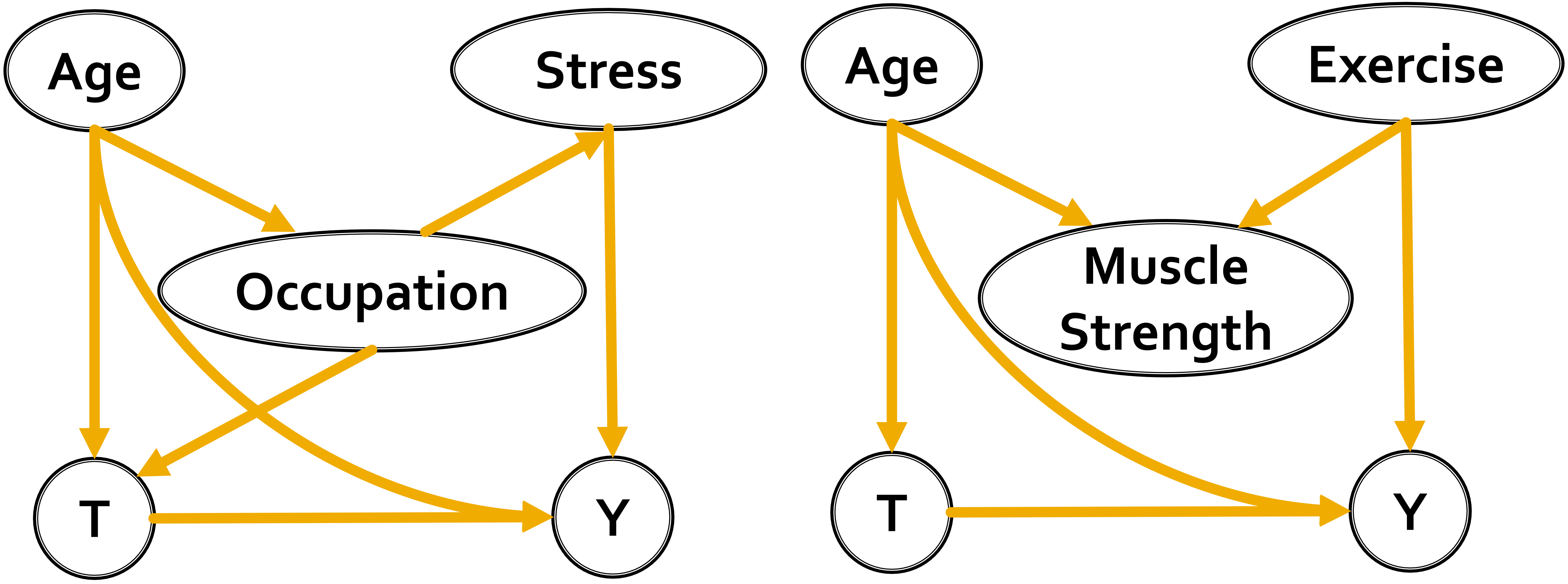
Structural Equation Models with Random Errors
u's are "error variables" or "exogenous variables"

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

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

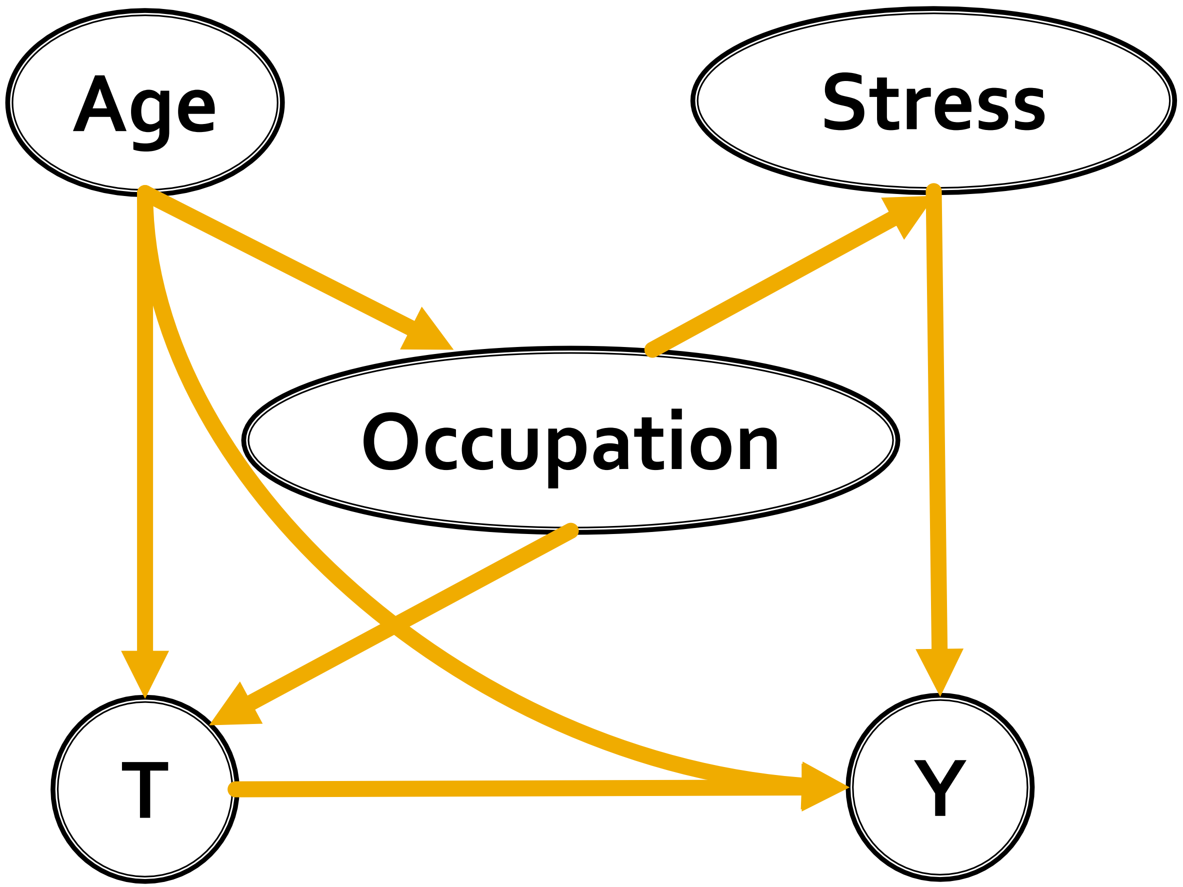
Edges represent *direct* causes.
Directed paths represent *indirect* causes.

Structural Causal Model makes assumptions explicit



The graph encodes all causal assumptions.

Important: Assumptions are the edges that are *missing*



Assumption 1: Occupation does 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.

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.

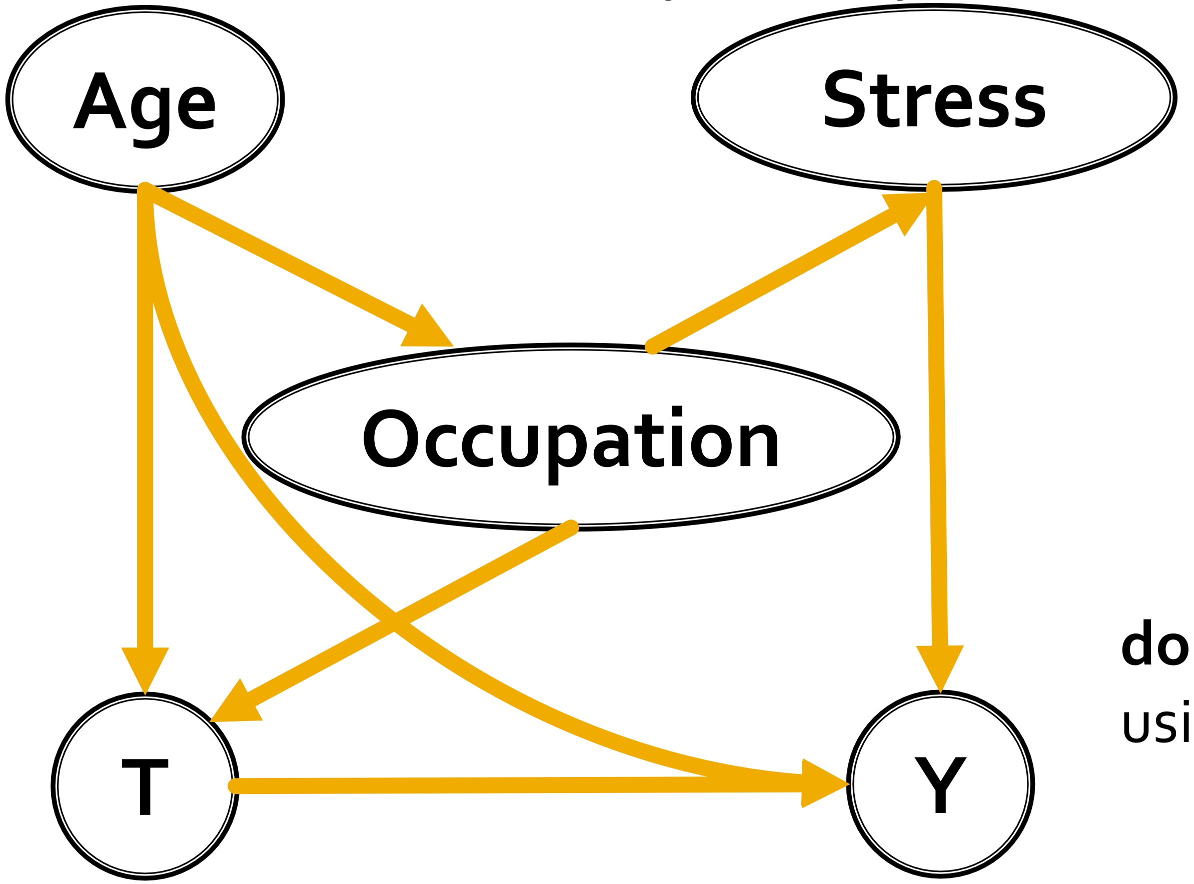
Can write it as:

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

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

$$P(Y|do(T))$$
$$= \cdots do{-}calculus\ rules\ \ \ldots$$

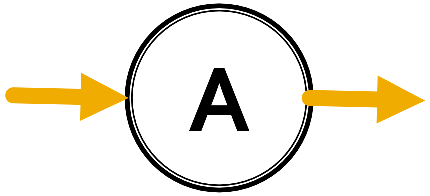
$$= \sum_{Age,Stress} P(Y|T, Age, Stress)\, P(Age, Stress)$$

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

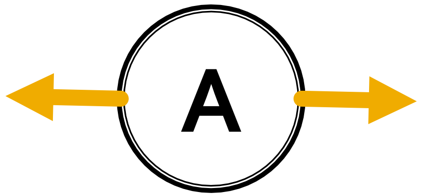
Advanced Topic: Back-door criterion

Three kinds of node-edges

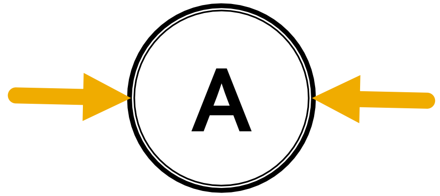
Path is "blocked"



If conditioned on X



If conditioned on X



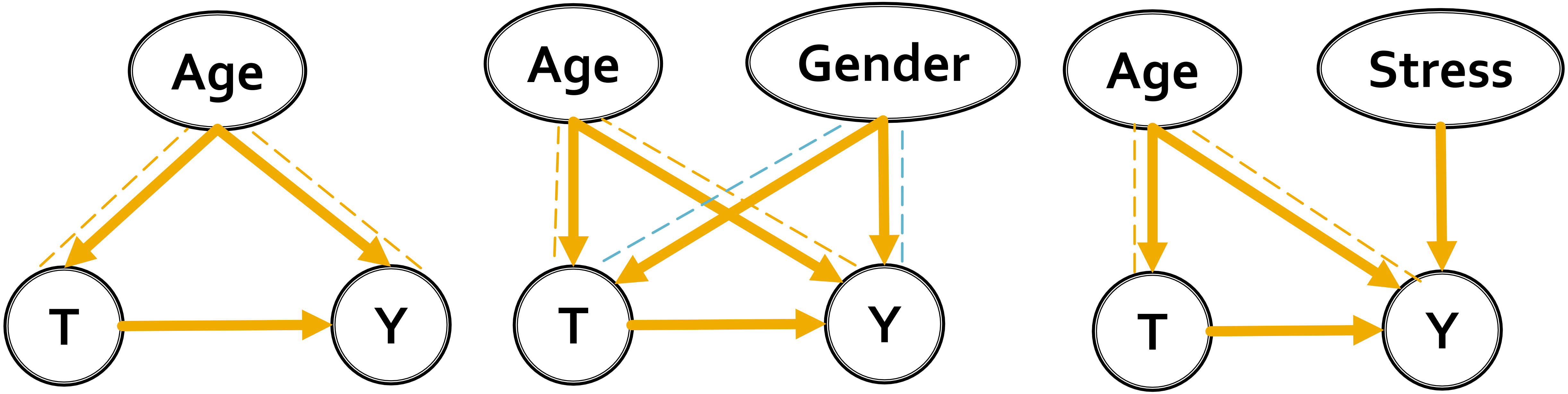
If **not** conditioned on X

"Back-door" path: Any undirected path that starts with $\to$ T and ends with $\to$ Y

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

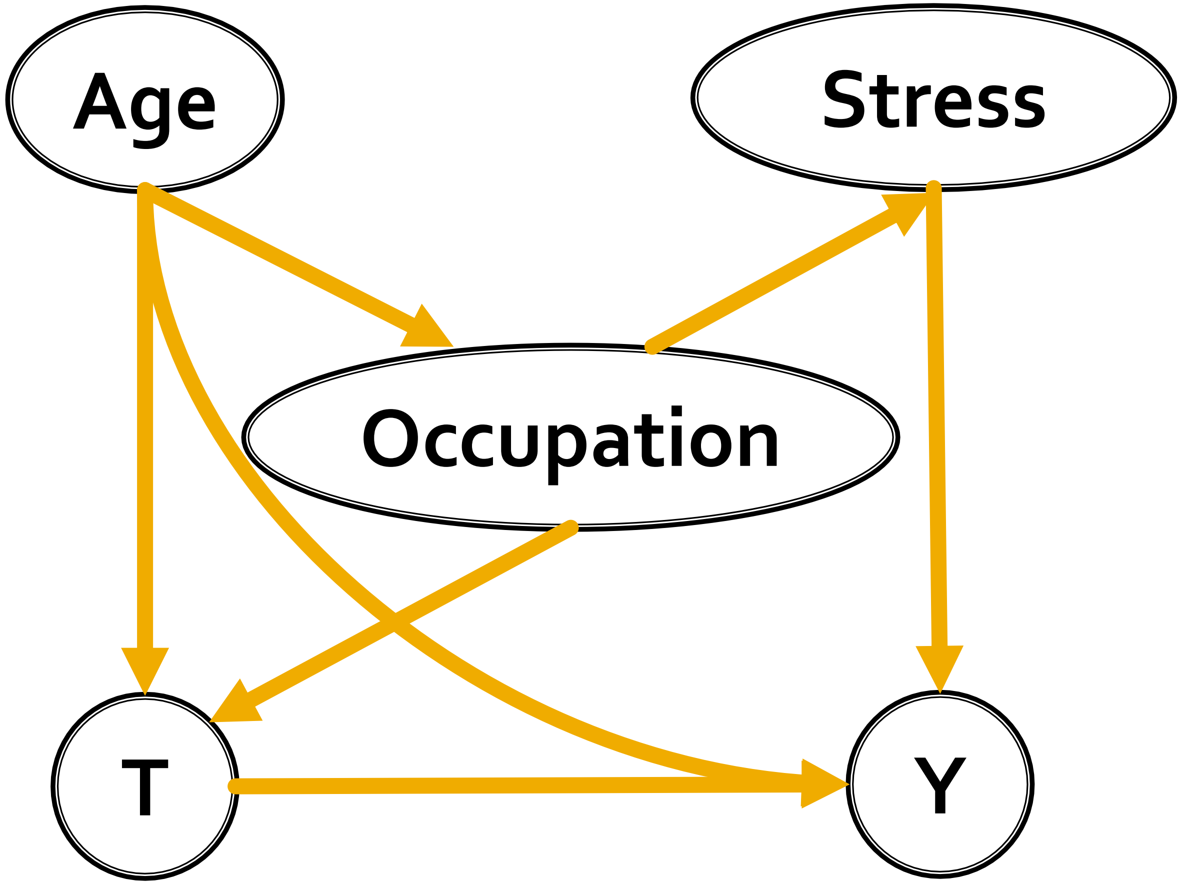
$$\boldsymbol{P(Y|do(T)) = \sum_{x} P(Y|T, X = x)P(X = x)}$$

Let us return to our examples

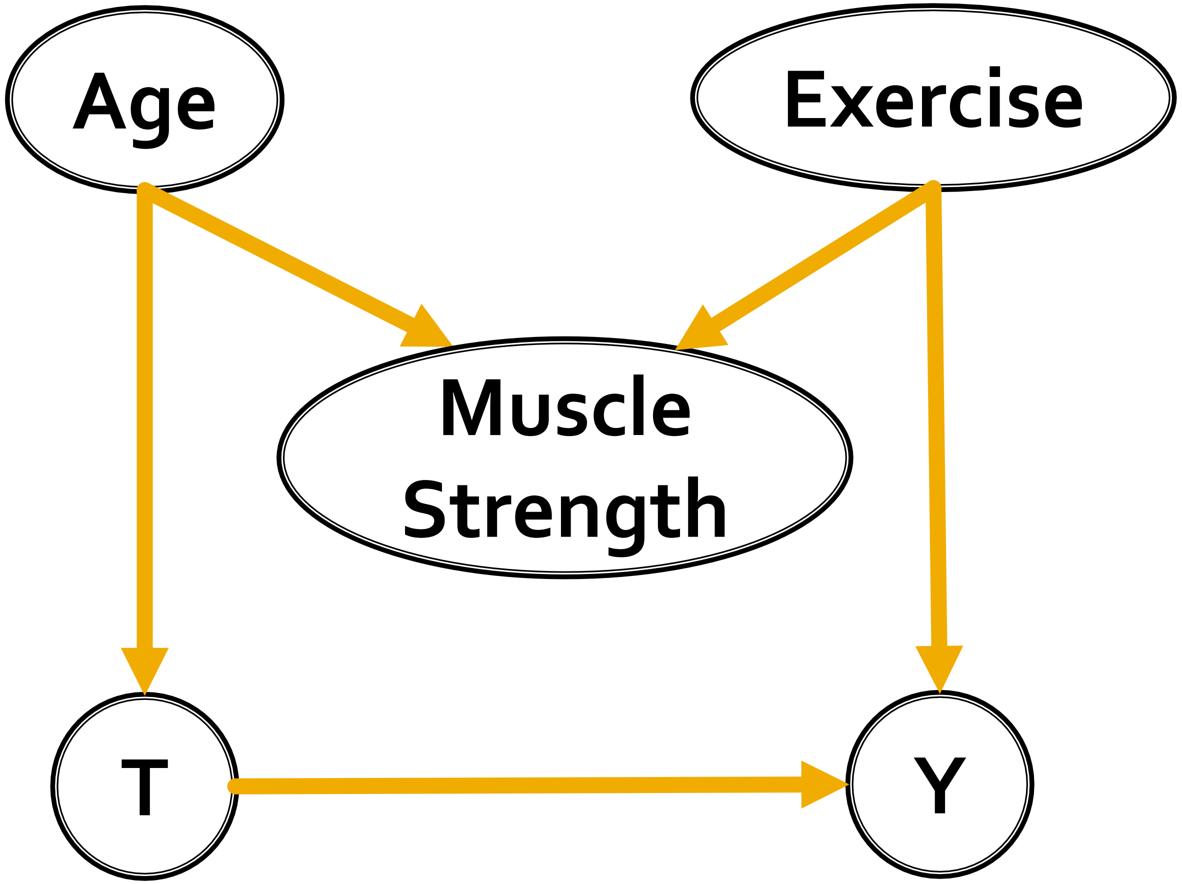


$\boldsymbol{X} = \{Age\}$ $\boldsymbol{X} = \{Age, Gender\}$ $\boldsymbol{X} = \{Age\}$

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



$$\boldsymbol{X} = \{Age, Stress\}$$
$$\boldsymbol{X} = \{Age, Occupation\}$$



$$\boldsymbol{X} = \{\quad\}$$
$$\boldsymbol{X} = \{MuscleStrength, Exercise\}$$

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: 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 of today:

- Causality is important for decision-making and study of effects
- Big Data does not necessarily address threats to causal infernce
- 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
- On Thursday: Methods for causal inference in observational data

Let's do the course evaluation

- https://uw.iasystem.org/survey/224031

- Your feedback makes a difference!
- Thank you!