
CAUSAL
DESIGN

Letter | Published: 20 March 2019

RETRACTED ARTICLE: Complex societies precede moralizing gods throughout world history

Harvey Whitehouse, Pieter François, Patrick E. Savage, Thomas E. Currie, Kevin C. Feeney, Enrico Cioni, Rosalind Purcell, Robert M. Ross, Jennifer Larson, John Baines, Barend ter Haar, Alan Covey & Peter Turchin

Nature **568**, 226–229 (2019) | Cite this article

This article was retracted on 07 July 2021

Matters Arising to this article was published on 07 July 2021

Letter | Published: 20 March 2019

RETRACTED ARTICLE: Complex societies precede moralizing gods throughout world history

Harvey Whitehouse, Pieter François, Patrick E. Savage, Thomas E. Currie, Kevin C. Feeney, Enrico Cioni, Rosalind Purcell, Robert M. Ross, Jennifer Larson, John Baines, Barend ter Haar, Alan Covey & Peter Turchin

Nature **568**, 226–229 (2019) | Cite this article

This article was retracted on 07 July 2021

Matters Ar

Matters Arising | Published: 07 July 2021

Treatment of missing data determined conclusions regarding moralizing gods

Bret Beheim, Quentin D. Atkinson, Joseph Bulbulia, Will Gervais, Russell D. Gray, Joseph Henrich, Martin Lang, M. Willis Monroe, Michael Muthukrishna, Ara Norenzayan, Benjamin Grant Purzycki, Azim Shariff, Edward Slingerland, Rachel Spicer & Aiyana K. Willard

Nature **595**, E29–E34 (2021) | Cite this article

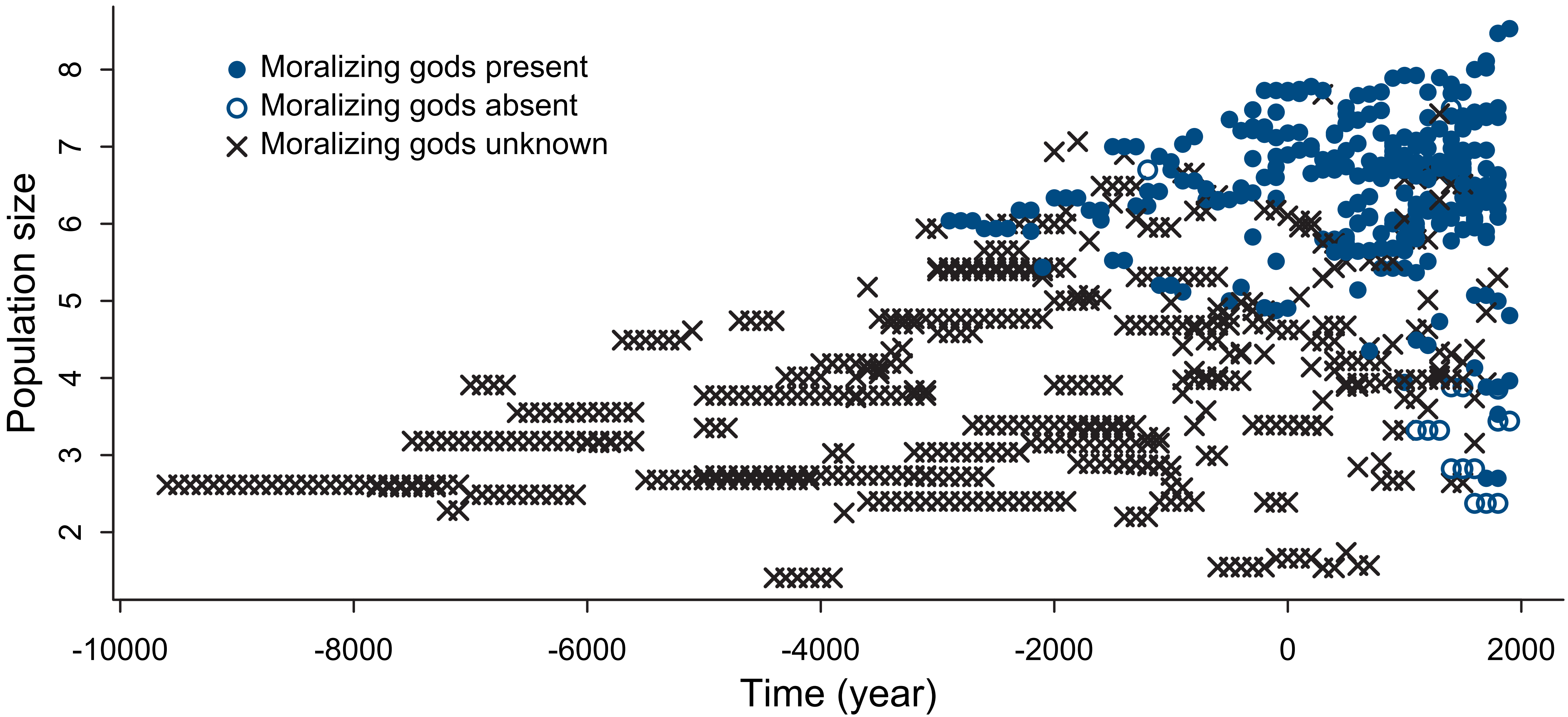
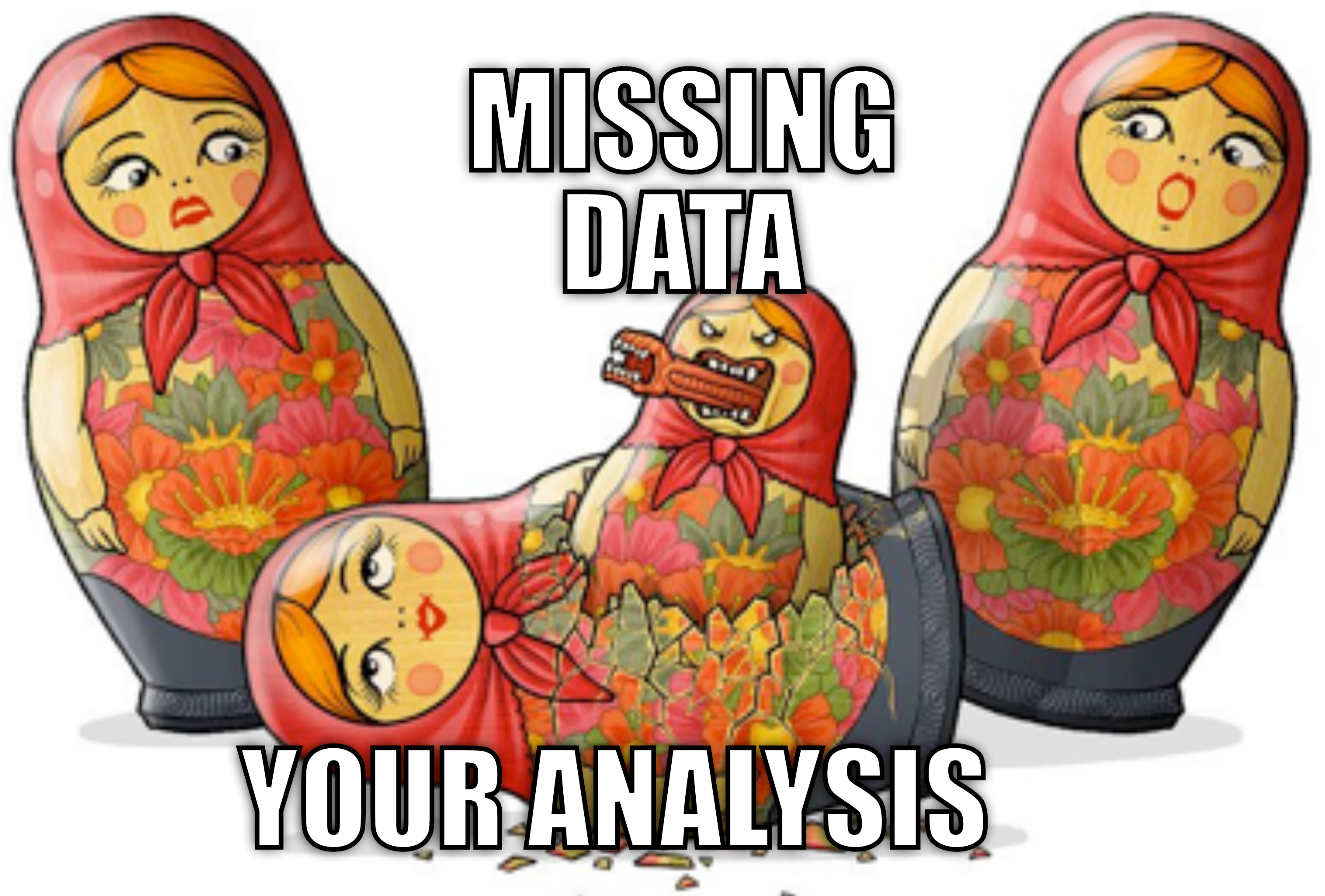


FIGURE 15.7. Missing values in the `Moralizing_gods` data. The blue points, both open and filled, are observed values for the presence of beliefs about moralizing gods. The x symbols are unknowns, the missing values.

McElreath 2020 *Statistical Rethinking*, page 514

Missing methods

- A circus of **cause-blind methods** for missing values
  - drop all cases with missing values
  - replace missing values with means
  - "multiple imputation"
  - Bayesian imputation
  - prayer



Causal Design

- Causal inference **requires** a model outside of the statistical model
- Step 1: Make a causal model
- Step 2: Use the model to design data collection and statistical procedures

Making Models

- Fundamental component of a causal model: Function that determines how some variables are influenced by others.

- In simplest case, can represent these functions with **arrows** that connect variables

$$X \longleftarrow Z \longrightarrow Y$$

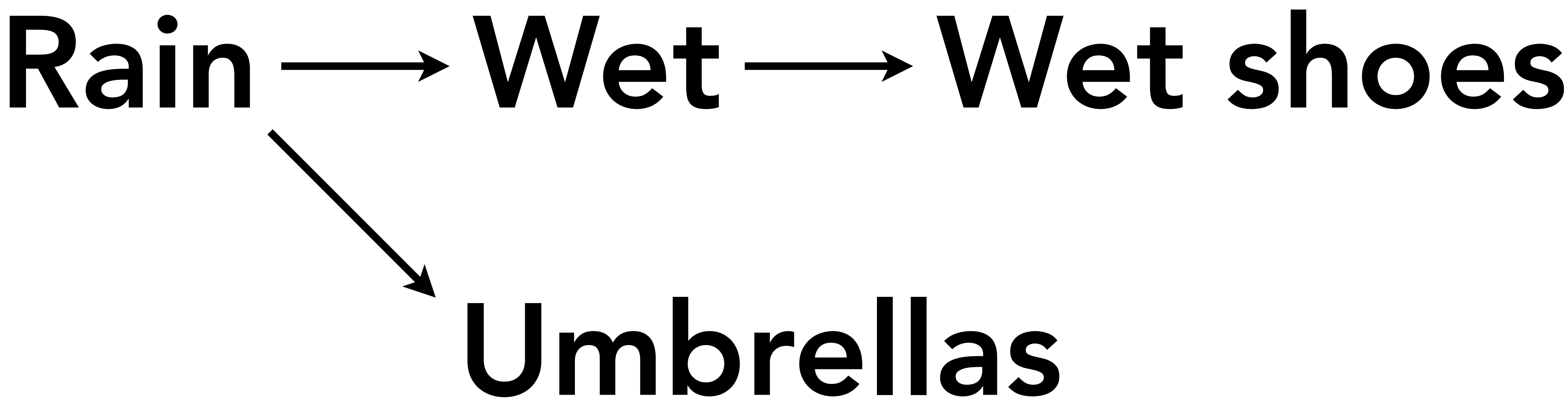
- There is no method for making causal models other than **science**. There is no method to science other than **honest anarchy**.

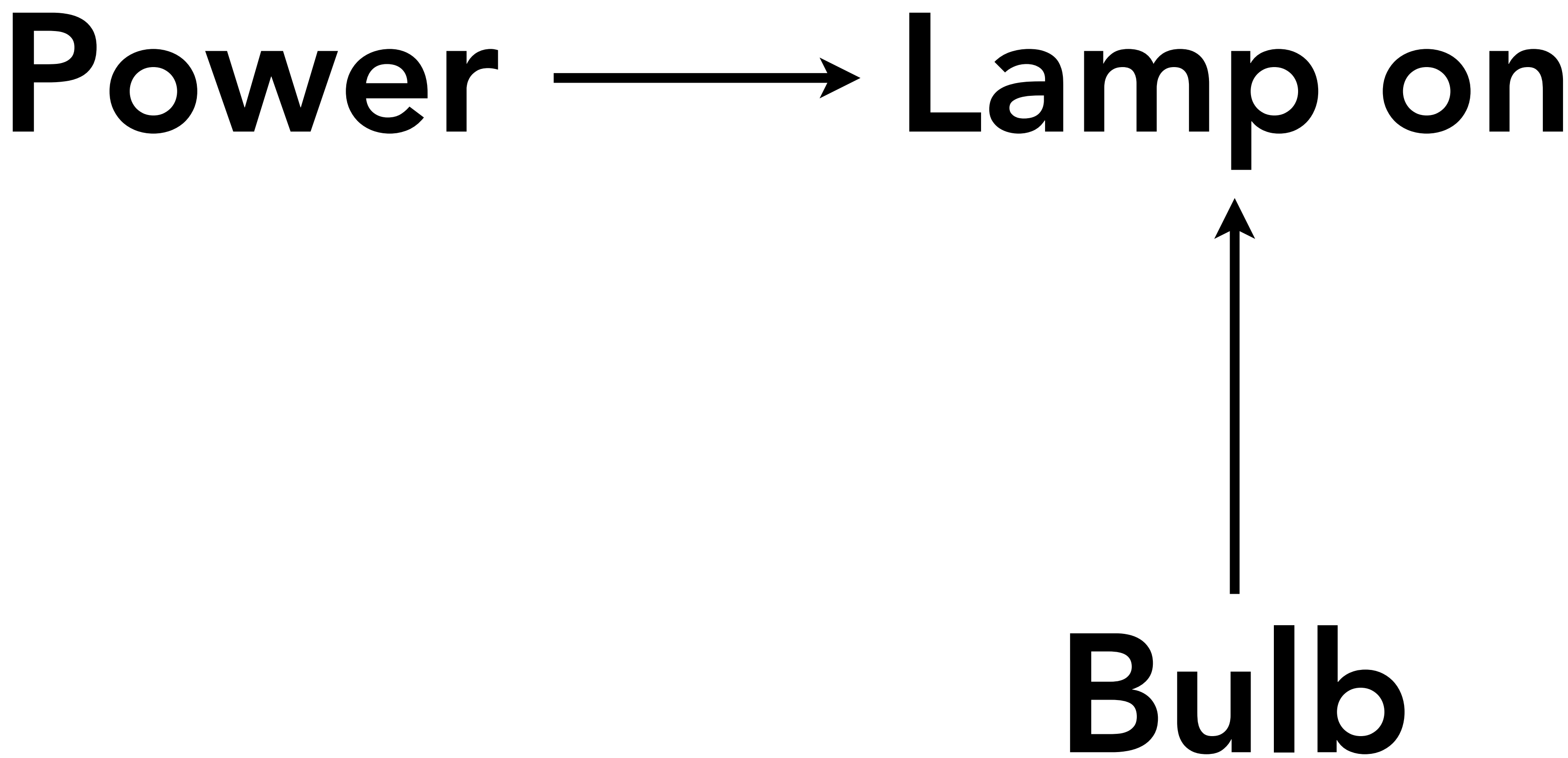
Directed Acyclic Graphs

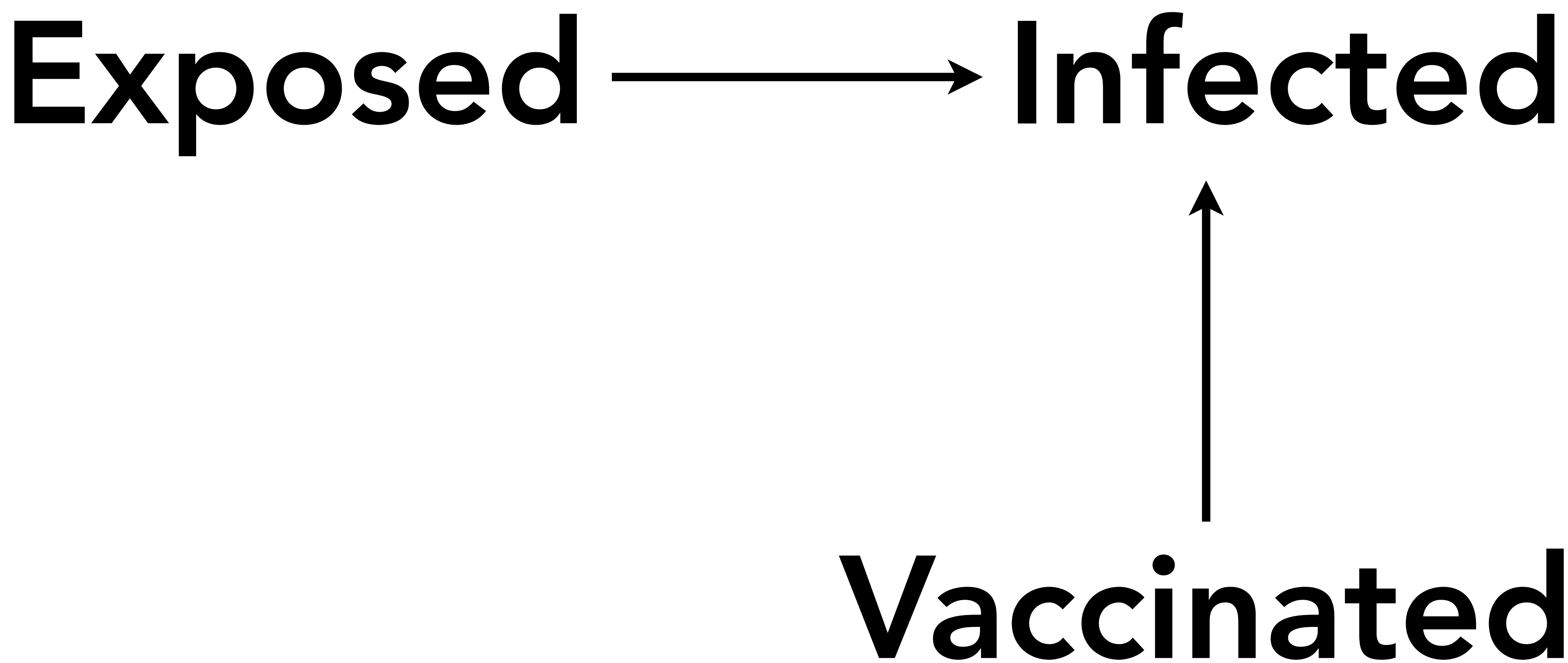
- Common heuristic form of causal model is Directed Acyclic Graph (DAG)
- DAGs leave how variables influence one another anonymous; only say which variables influence which
- Examples…

Rain **Wet**

Rain $\longrightarrow$ **Wet**


Rain
Wet
Wet shoes
Umbrellas


Power
Lamp on
Bulb


Exposed
Infected
Vaccinated

Two Moms / Peer Bias

- **Two Moms**
  - Variables: family sizes M and D, birth orders B1 and B2.
  - How would you connect these variables?
- **Peer Bias**
  - Variables: field E, category X, outcome Y
  - How would you connect these variables?

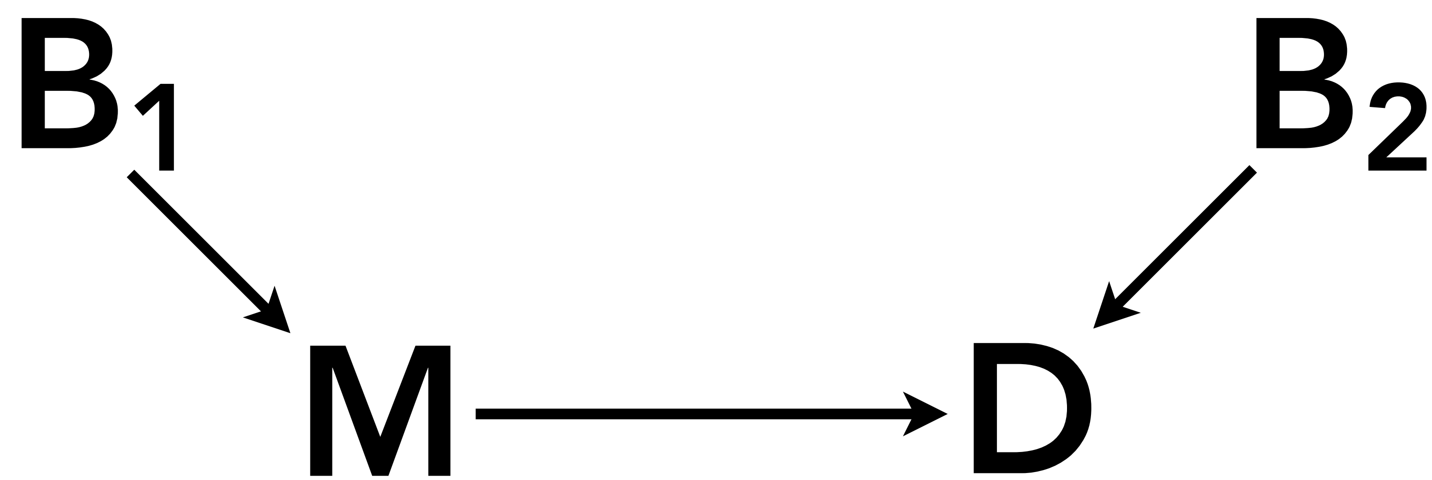
M D B_1 B_2

TWO MOMS

M D B_1 B_2

$$\begin{tikzcd}
B_1 \arrow[dr] & \\
& M
\end{tikzcd}$$

M D B_1 B_2



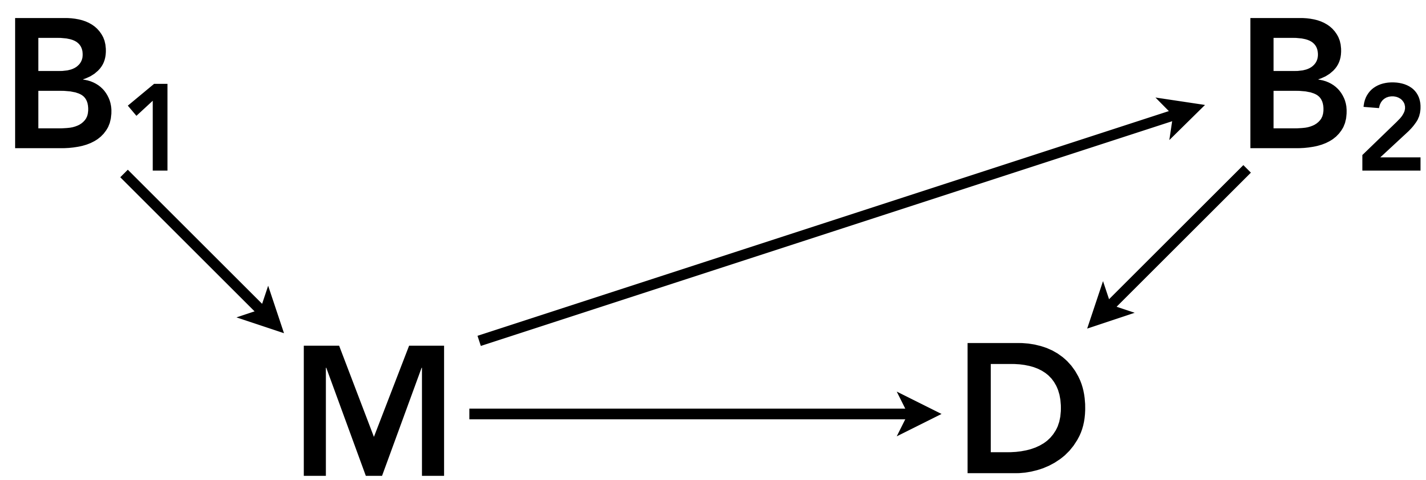
$M \qquad D \qquad B_1 \qquad B_2$

$$\begin{tikzcd}
B_1 \arrow[dr] & & & B_2 \arrow[dl] \\
& M \arrow[r] & D &
\end{tikzcd}$$

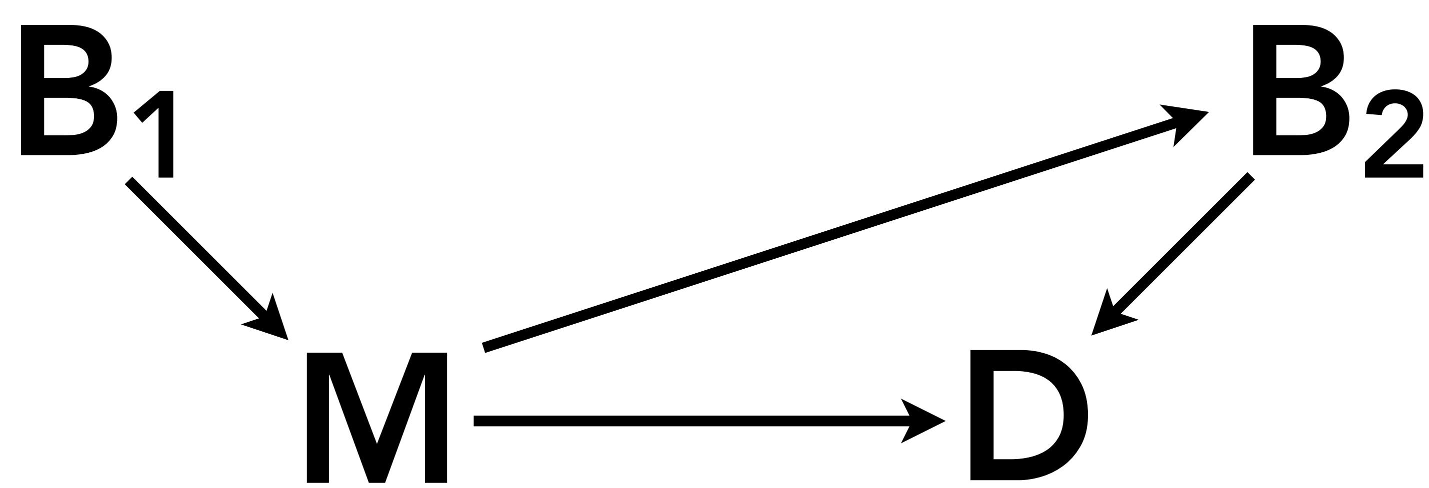
$M \qquad D \qquad B_1 \qquad B_2$

$$\begin{tikzcd}
B_1 \arrow[dr] & & & B_2 \arrow[dl] \\
& M \arrow[r] & D &
\end{tikzcd}$$

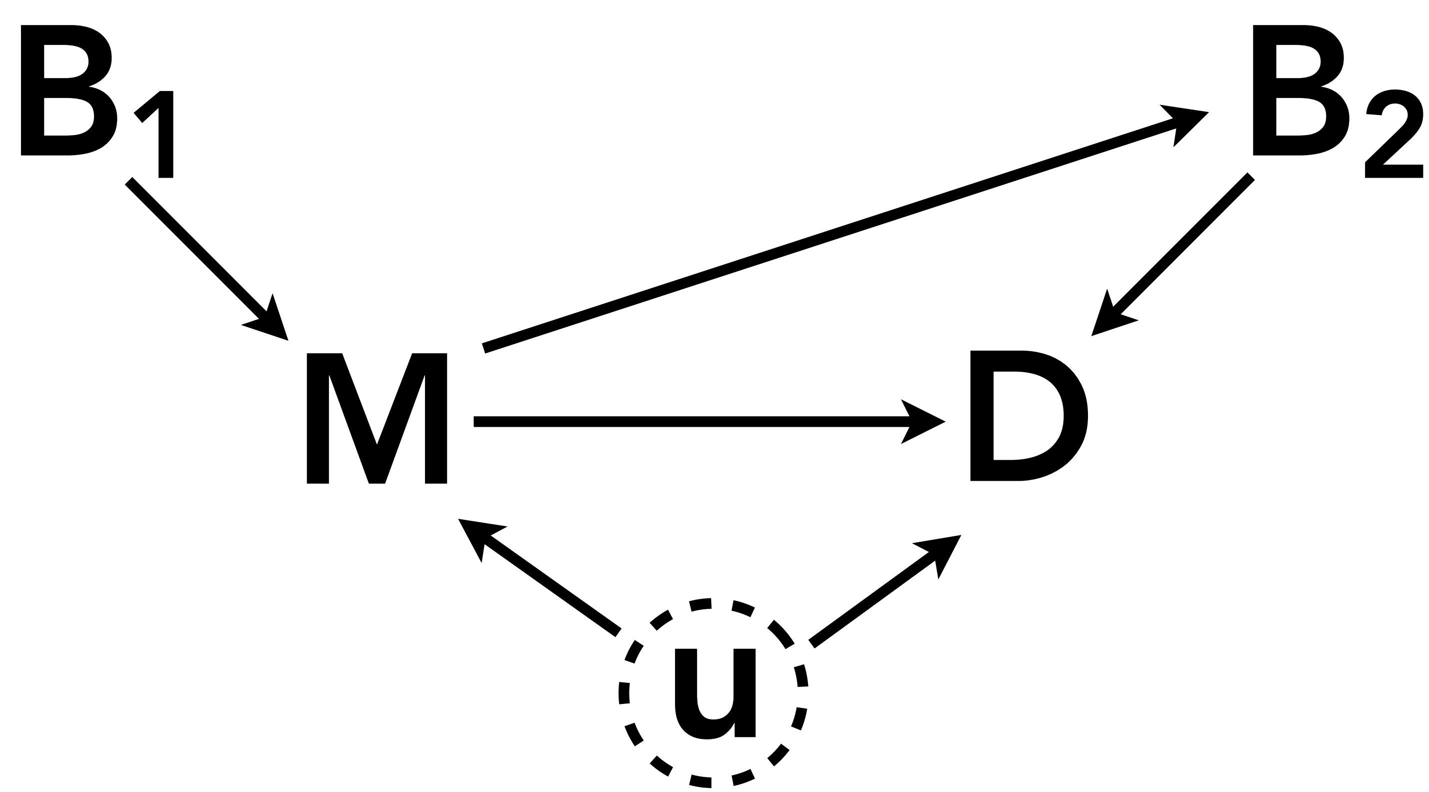
$M \quad D \quad B_1 \quad B_2$



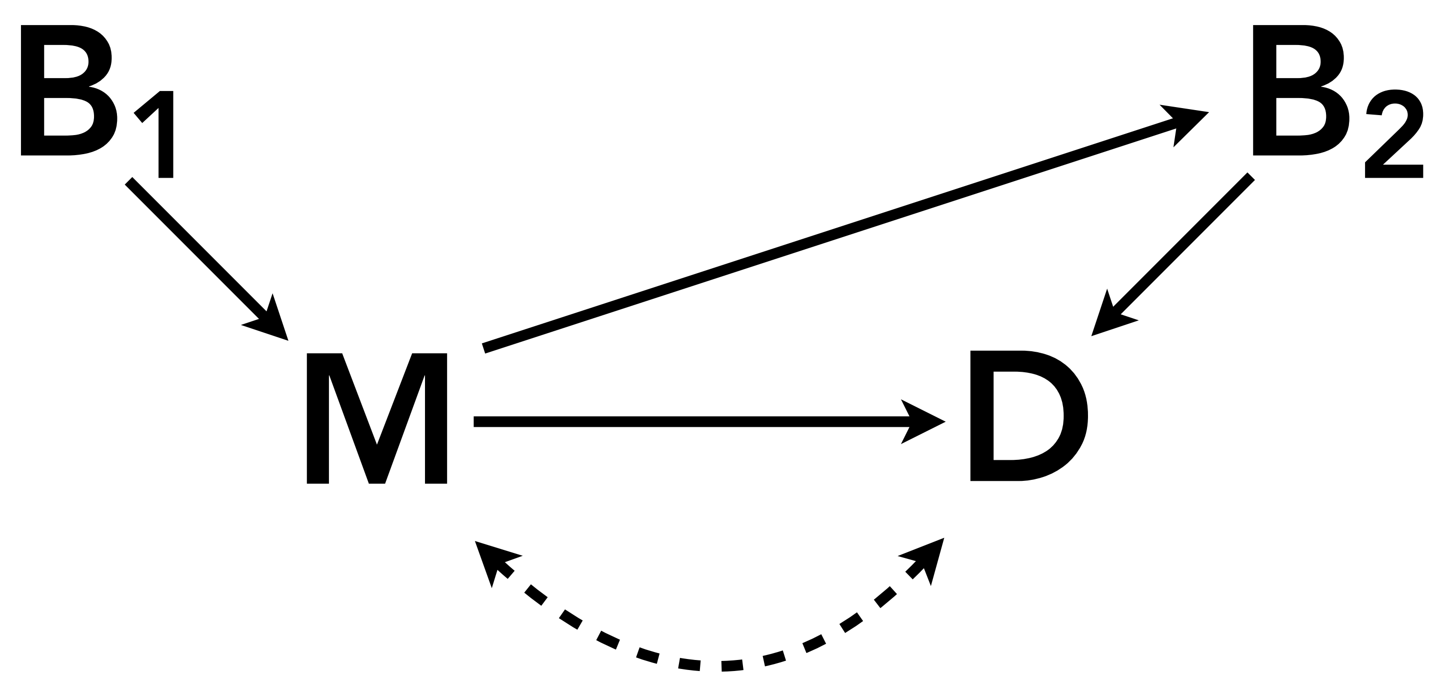
Now consider unobserved variables



Now consider unobserved variables



Now consider unobserved variables



X E Y

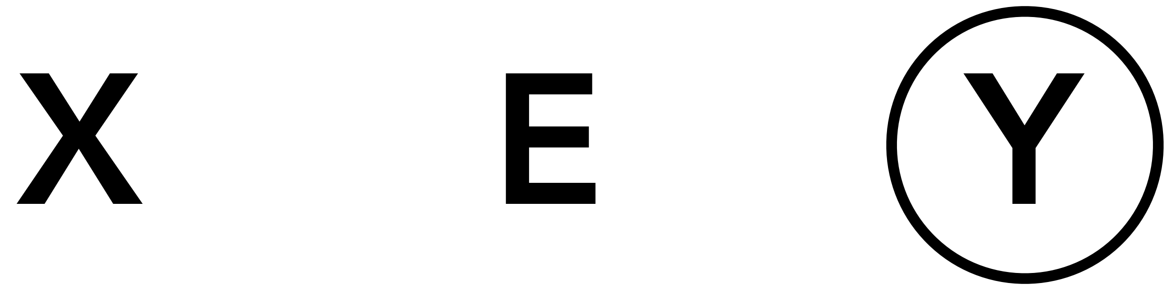
PEER BIAS

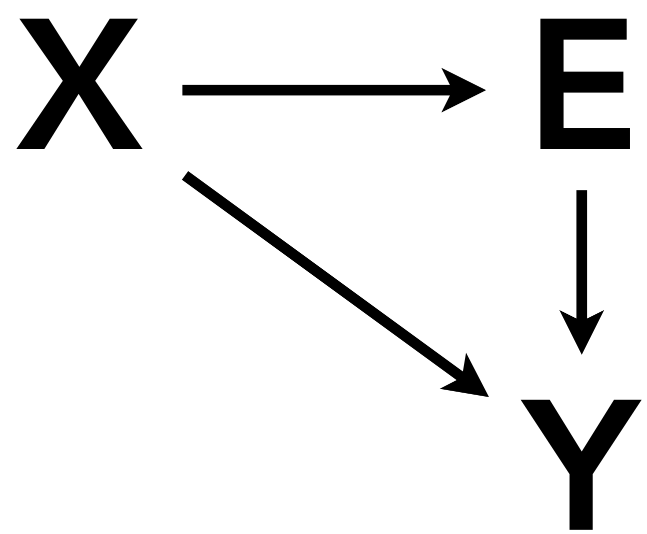
X E Y

X

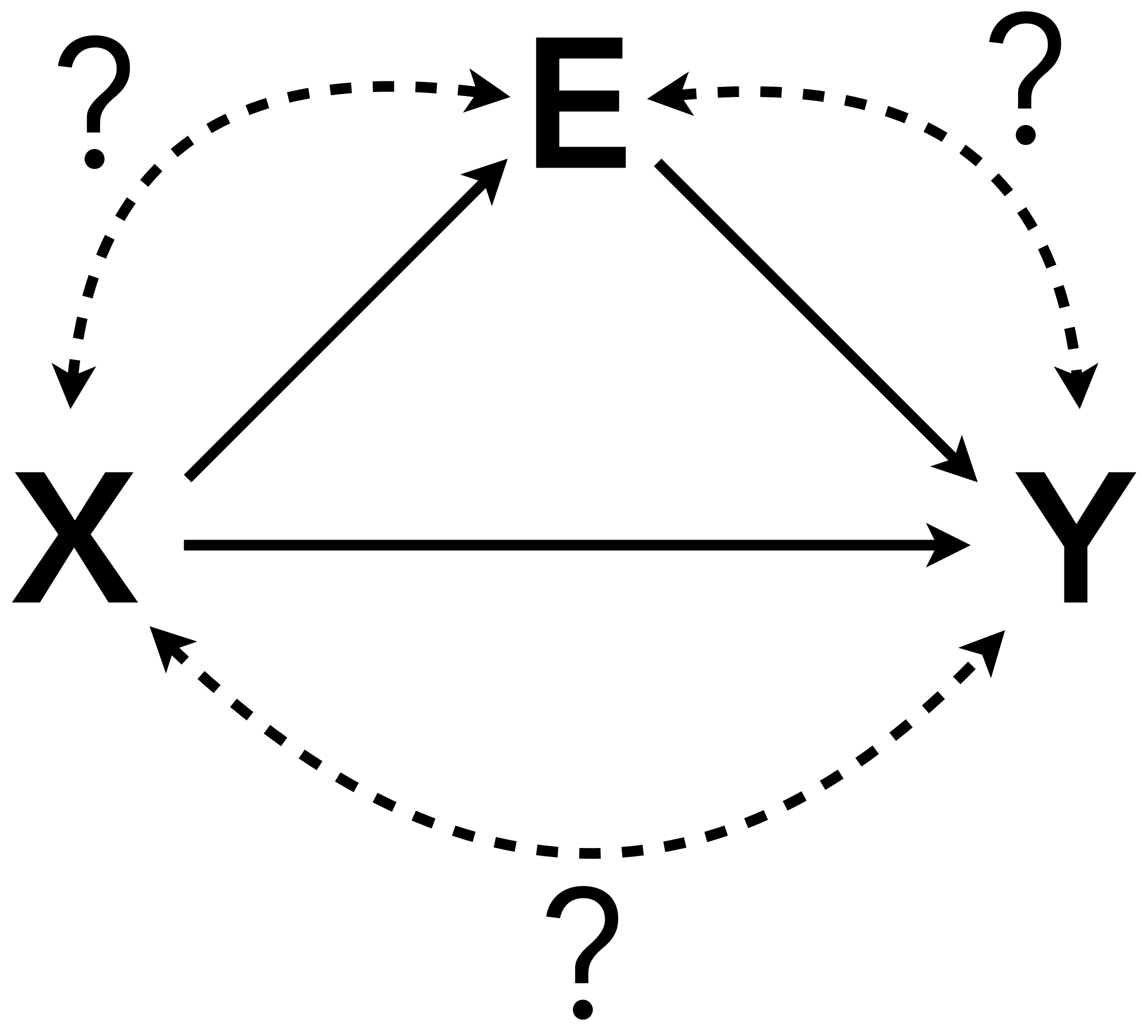
X E Y

$$X \longrightarrow E$$

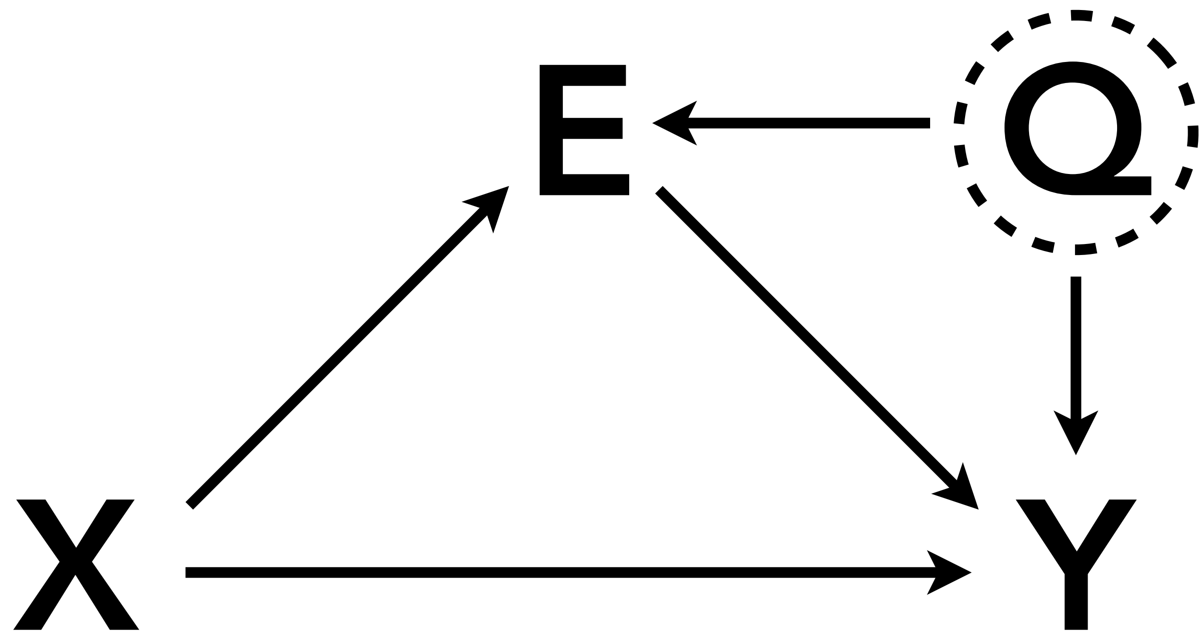

X
E
Y


X
E
Y

Now consider unobserved variables



Now consider unobserved variables



Analyzing DAGs

- Causal models can be complicated
- But rules for analyzing them are simple
- Analysis allows:
  - Testing some assumptions
  - Derivation of correct statistical procedures
- All of this requires understanding **d-separation**

Elemental paths

Fork	$X \longleftarrow Z \longrightarrow Y$
Pipe	$X \longrightarrow Z \longrightarrow Y$
Collider	$X \longrightarrow Z \longleftarrow Y$

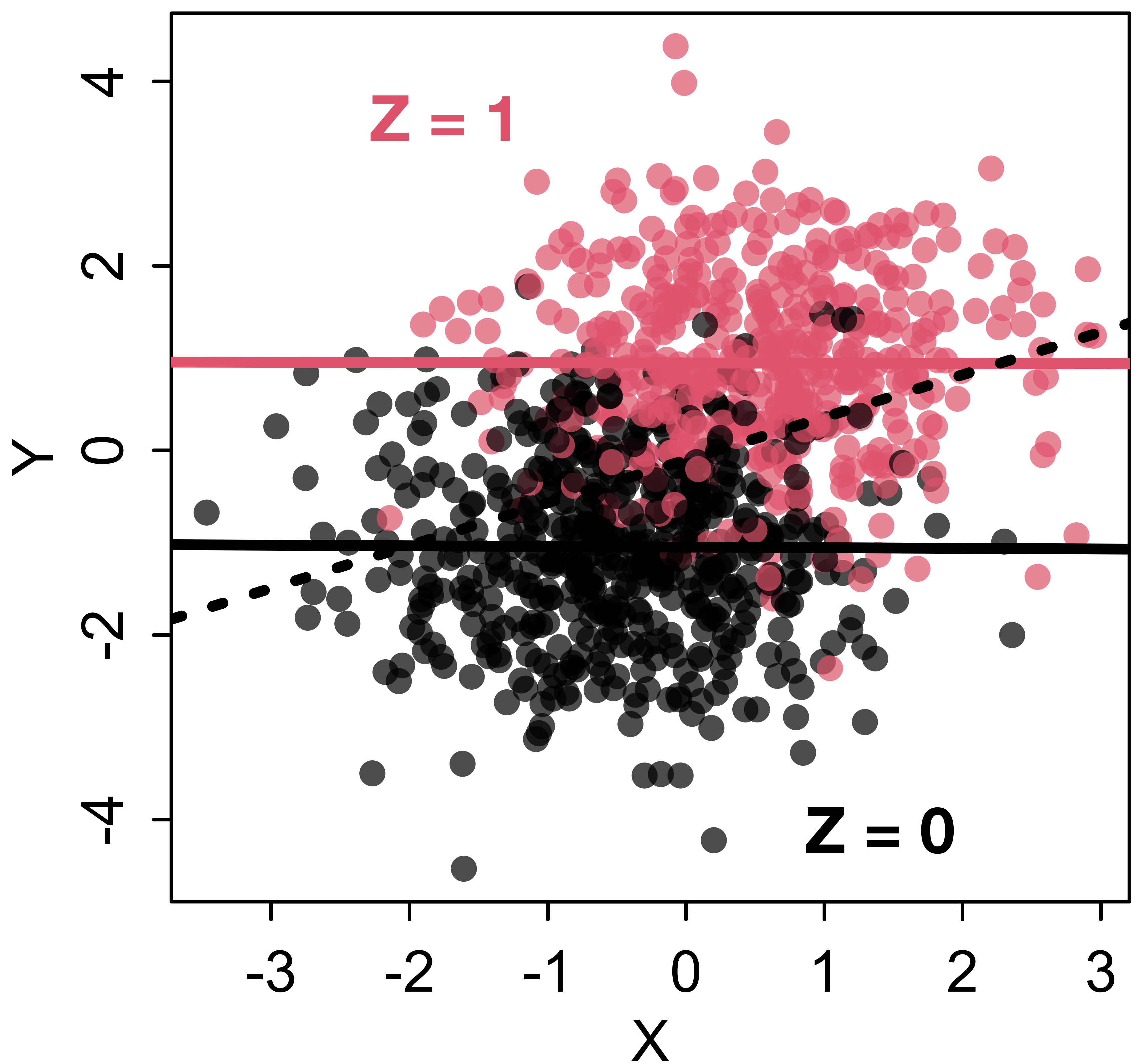
Pipe $X \longrightarrow Z \longrightarrow Y$

Ignoring **Z**, **X** and **Y** associated

Stratified by **Z**, **X** and **Y** NOT associated

Pipe

$X \longrightarrow Z \longrightarrow Y$



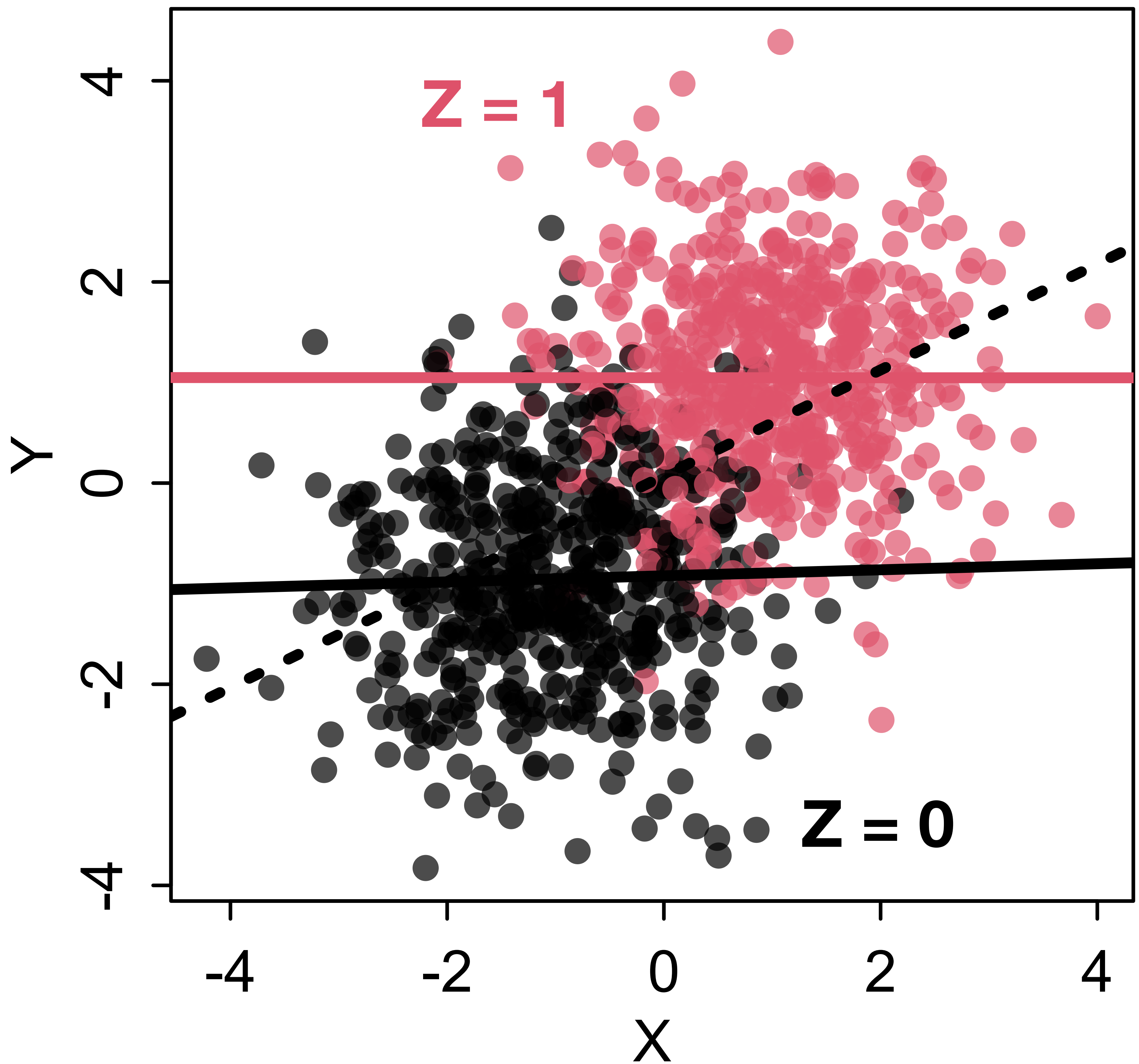
Fork $\mathbf{X} \longleftarrow \mathbf{Z} \longrightarrow \mathbf{Y}$

Ignoring **Z**, **X** and **Y** associated

Stratified by **Z**, **X** and **Y** NOT associated

Fork

$X \longleftarrow Z \longrightarrow Y$



Fork and **Pipe** look the same in the data alone.

Fork $\quad \mathbf{X} \longleftarrow \mathbf{Z} \longrightarrow \mathbf{Y}$

Pipe $\quad \mathbf{X} \longrightarrow \mathbf{Z} \longrightarrow \mathbf{Y}$

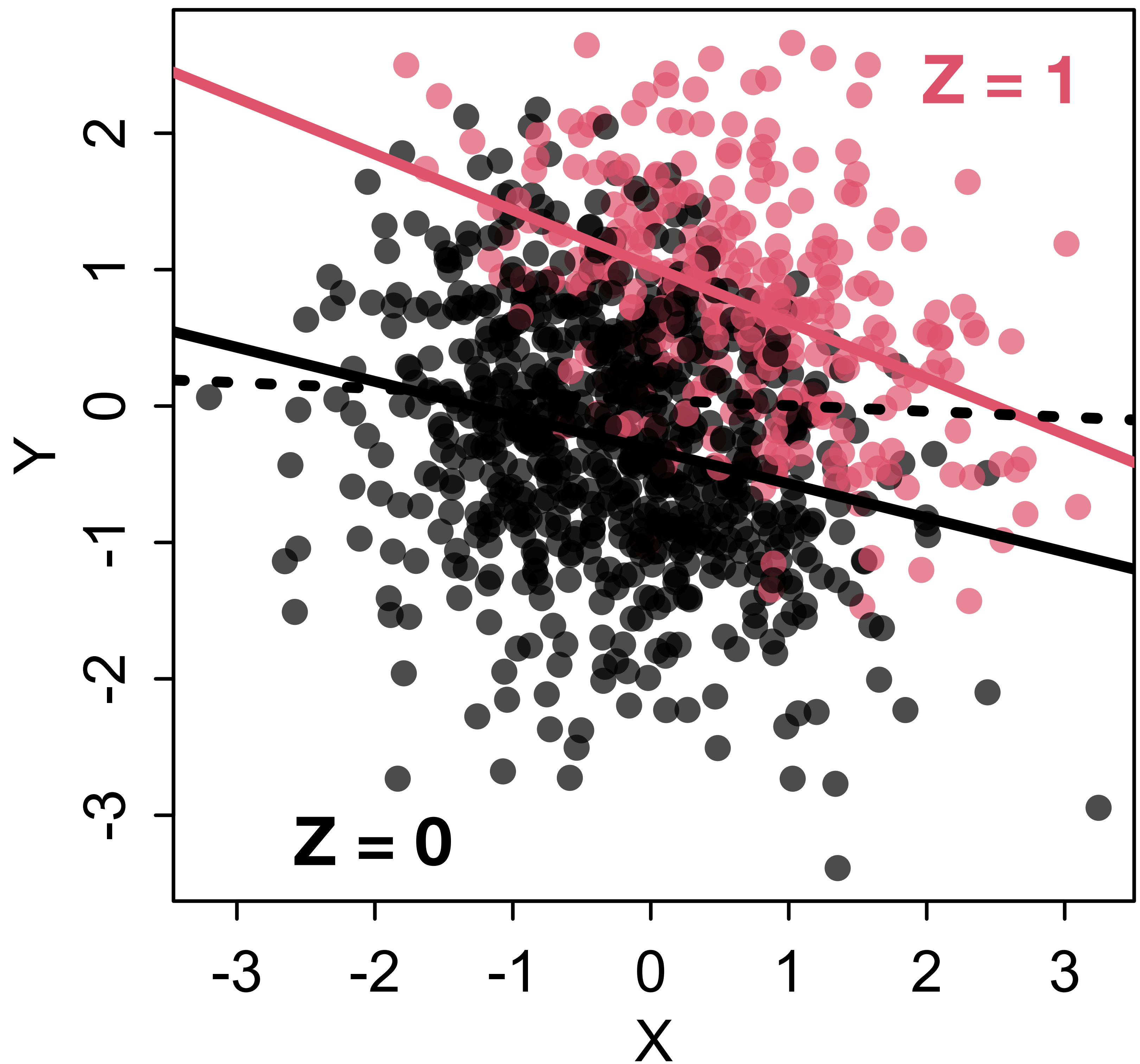
Collider $\mathbf{X} \longrightarrow \mathbf{Z} \longleftarrow \mathbf{Y}$

Ignoring **Z**, **X** and **Y** NOT associated

Stratified by **Z**, **X** and **Y** associated

Collider

$X \longrightarrow Z \longleftarrow Y$



Understanding Colliders

$$X \longrightarrow Z \longleftarrow Y$$

- Why does learning (conditioning on) the outcome **Z** induce an association among the causes **X** and **Y**?
- An association indicates **mutual information**: if I learn **X** then I also learn something about **Y**
- For any given value of **Z**, learning **X** tells us what **Y** might have been

Location $\longrightarrow$ \$ $\longleftarrow$ Food

- A restaurant in a **good location** can $, even if it has bad food
- A restaurant with **great food** can $, even if it has a bad location
- Conditioning on $, **negative association** between location and quality of food

Backdoor Criterion

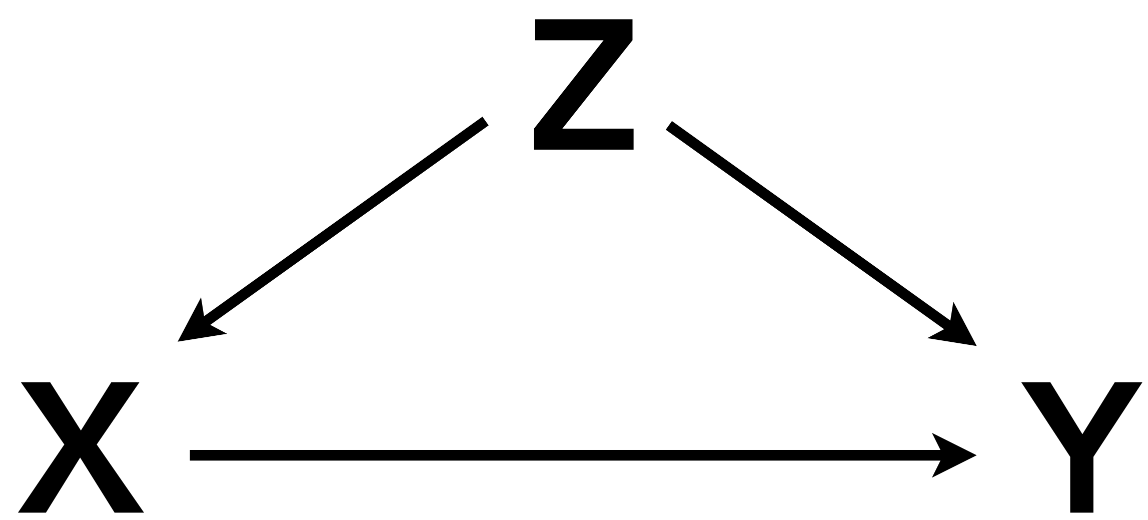
- These three relations used to derive **implications** of DAG
  - (1) design statistical procedures that estimate desired causal relationship
  - (2) design tests of the causal model
- Let’s start with (1) and apply something called the **backdoor criterion**

Backdoor Criterion

- A valid causal estimate is available, if it is possible to **condition** on variables such that all **backdoor paths** are closed

- **Backdoor path**: A non-causal path that enters the cause rather than exits it

BACK DOOR $\longrightarrow$ X $\longrightarrow$ FRONT DOOR


Z
X
Y

Backdoor Criterion

- Identify all paths connecting **X** and **Y**
- The paths with arrows entering **X** are possible **backdoor** paths that contaminate causal inference
- Causes do not flow against arrows but **ASSOCIATION DOES**
- For each backdoor path, find a way to close it by **conditioning** (or not) on relevant variables

Fork

$X \longleftarrow Z \longrightarrow Y$ **Z** closes

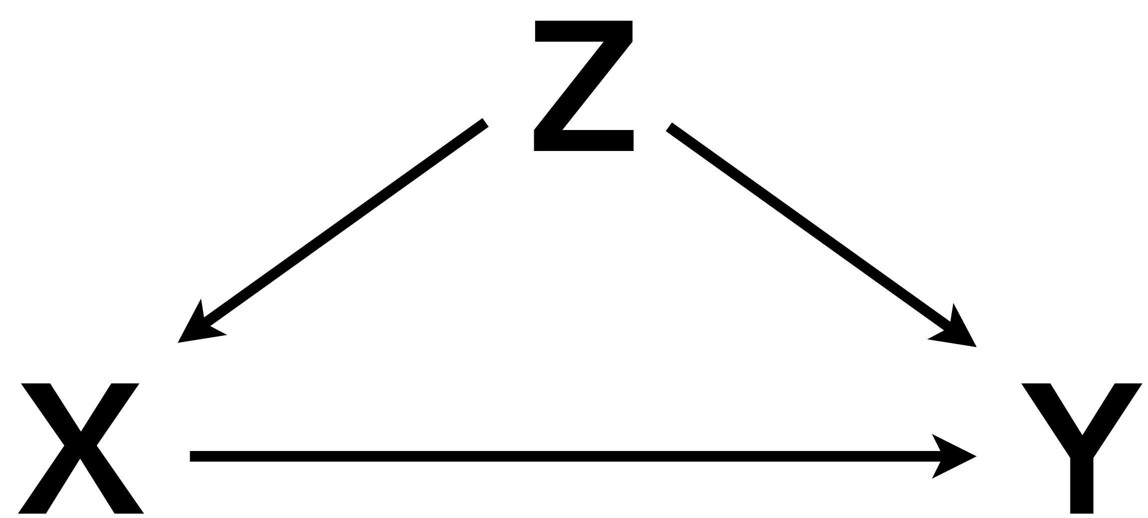
Pipe

$X \longrightarrow Z \longrightarrow Y$ **Z** closes

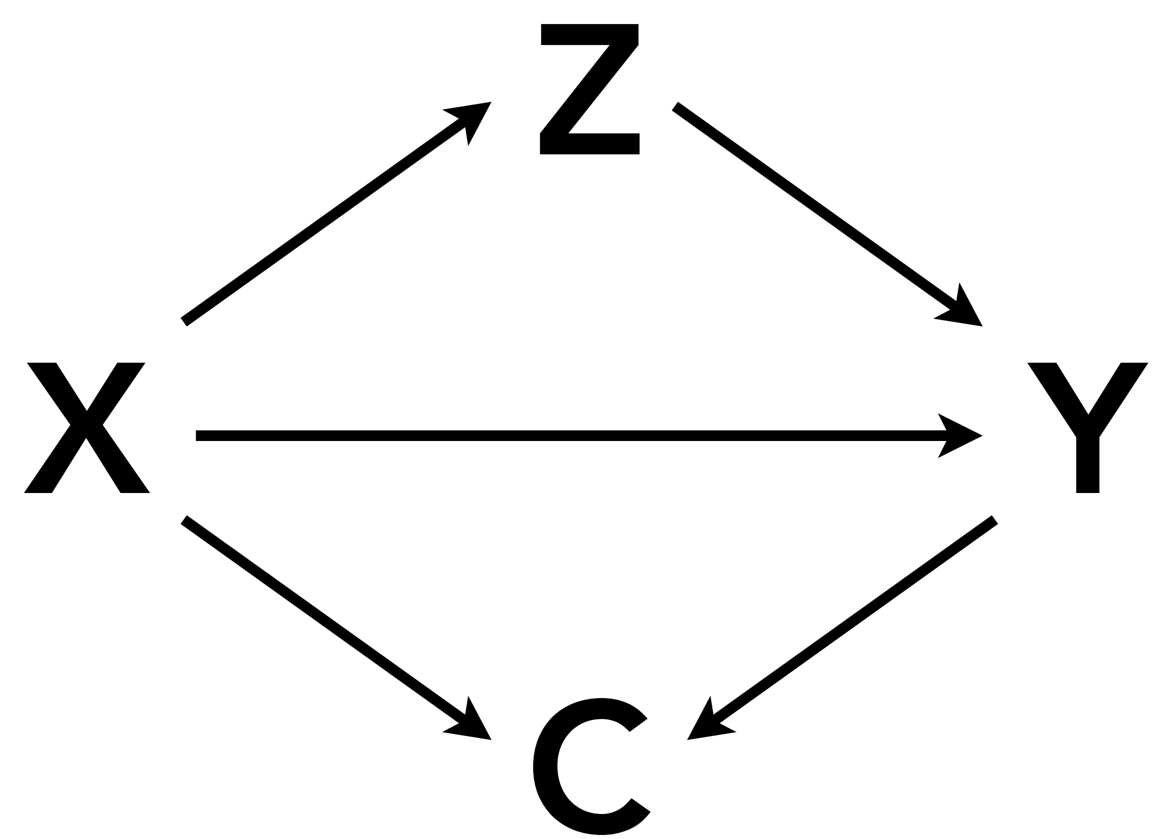
Collider

$X \longrightarrow Z \longleftarrow Y$ **Z** opens

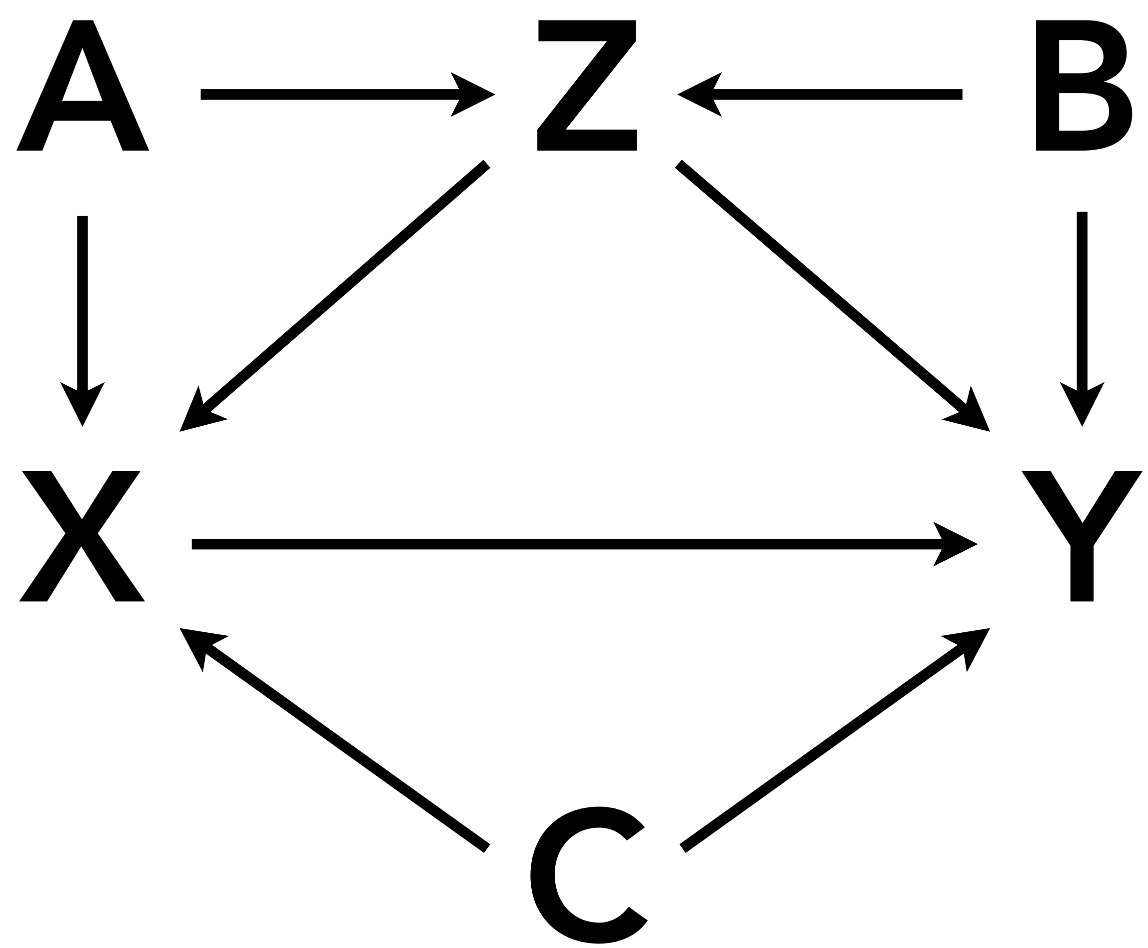
List all the paths connecting **X** and **Y**. Which need to be closed to estimate effect of **X** on **Y**?



List all the paths connecting **X** and **Y**. Which need to be closed to estimate effect of **X** on **Y**?

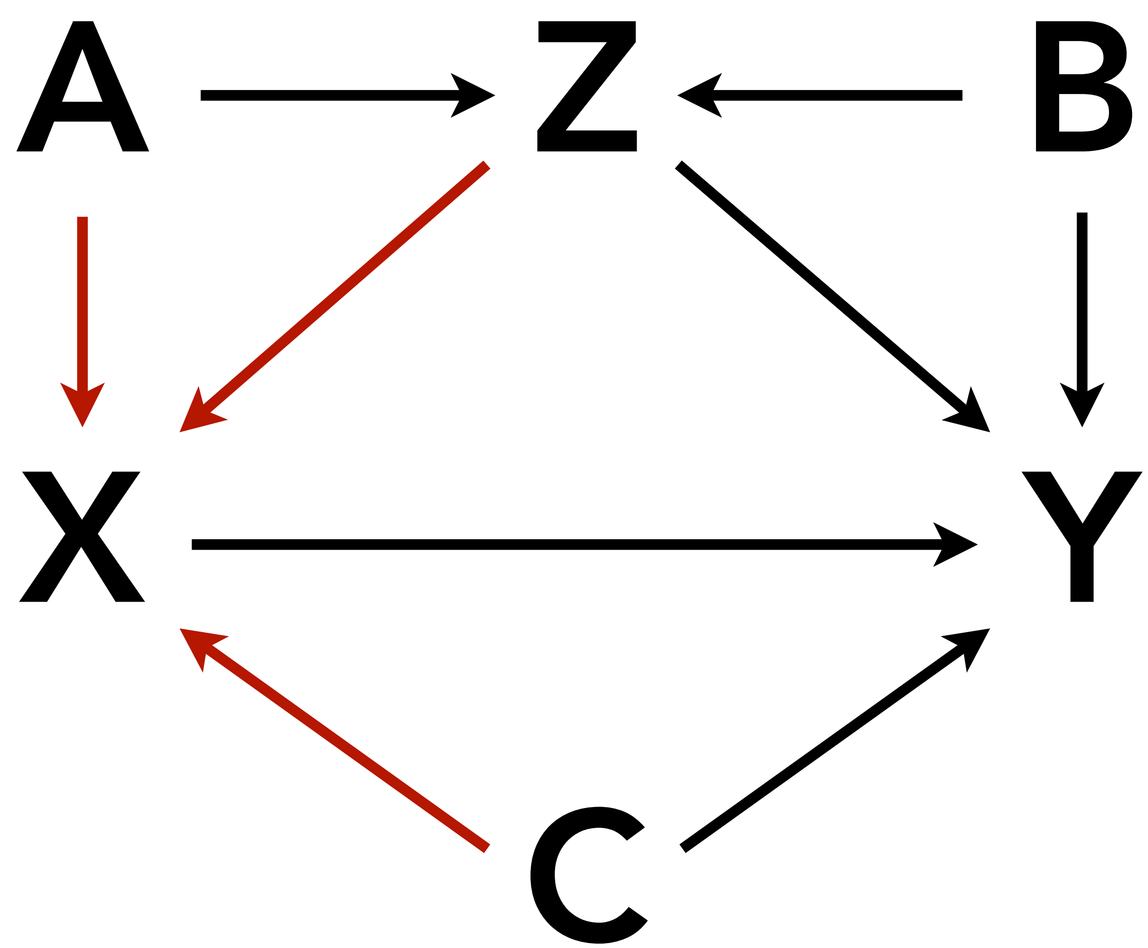


List all the paths connecting **X** and **Y**. Which need to be closed to estimate effect of **X** on **Y**?



10 MINS

The effect of randomization is to remove all arrows into treatment



The effect of randomization is to remove all arrows into treatment

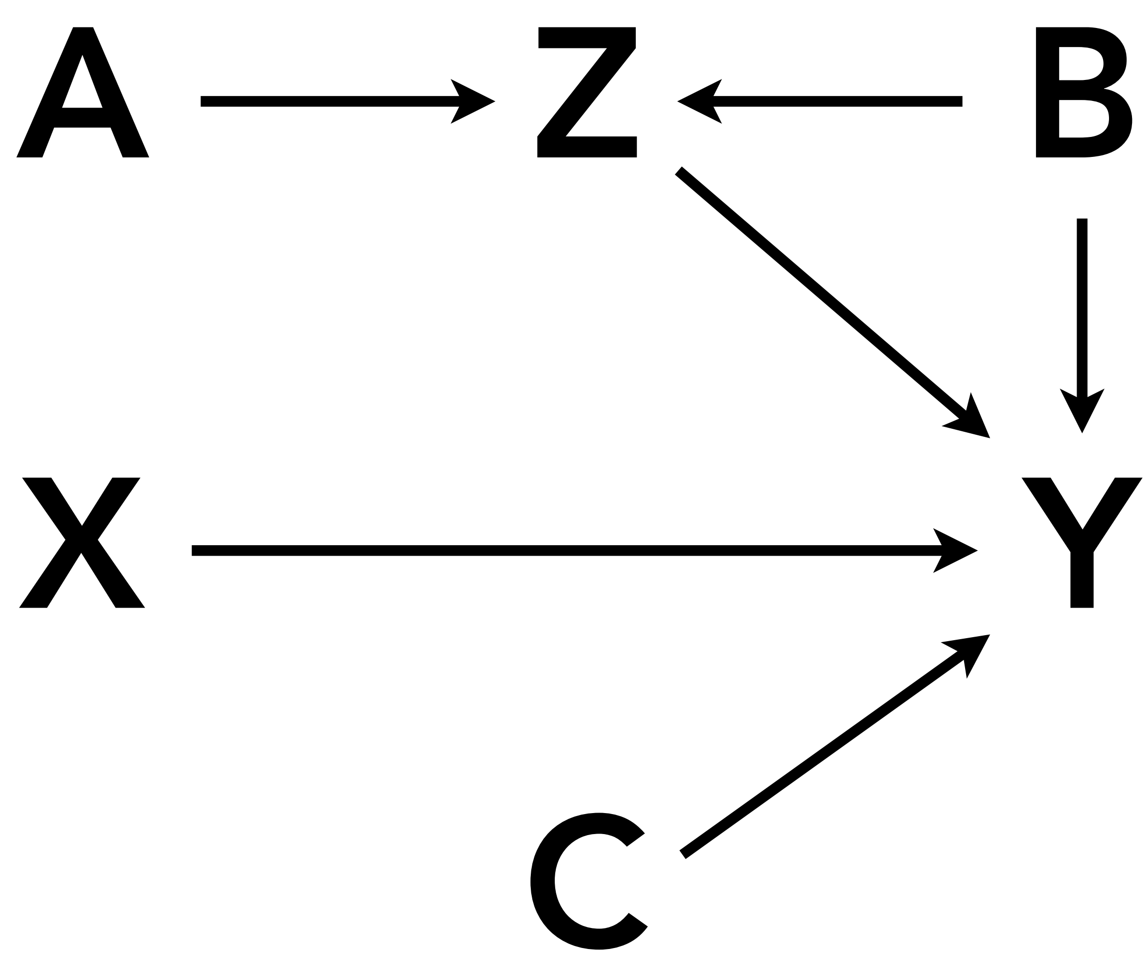
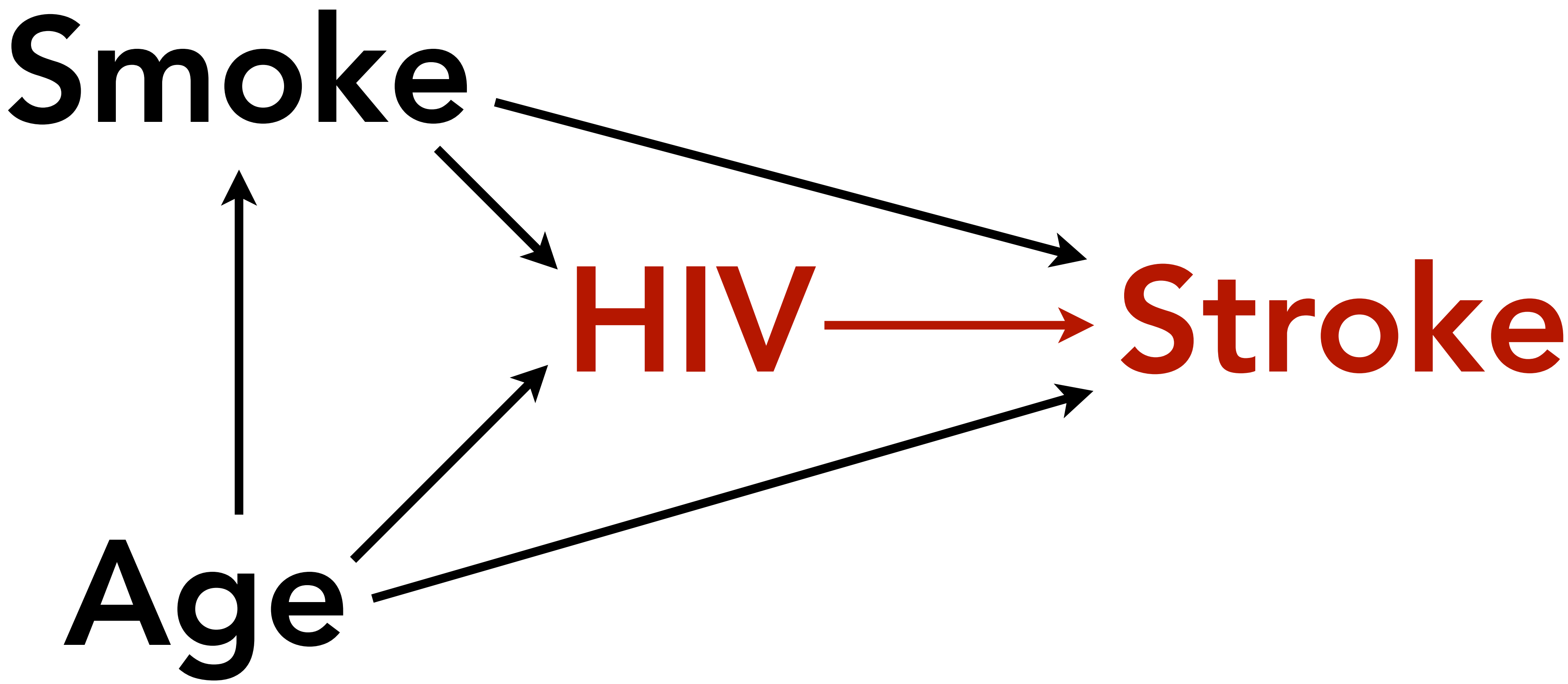
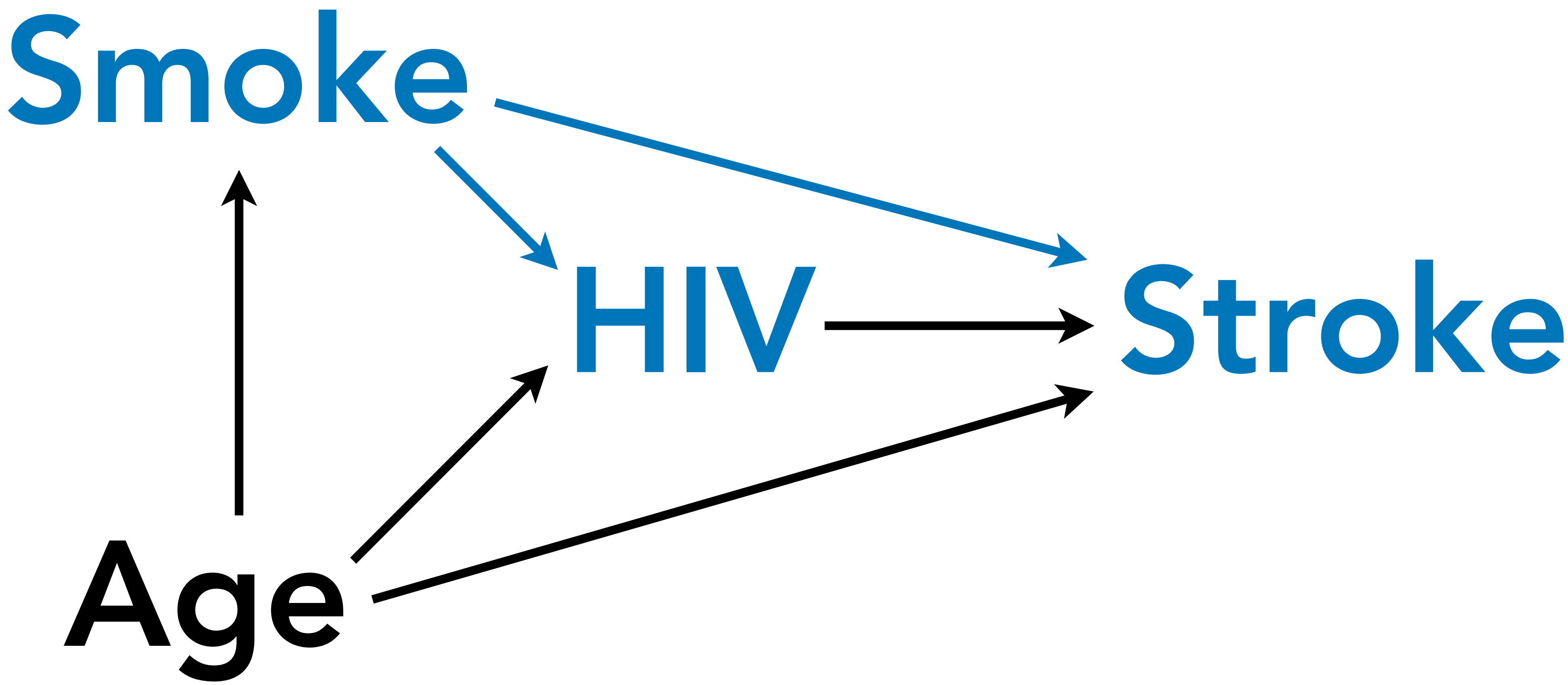


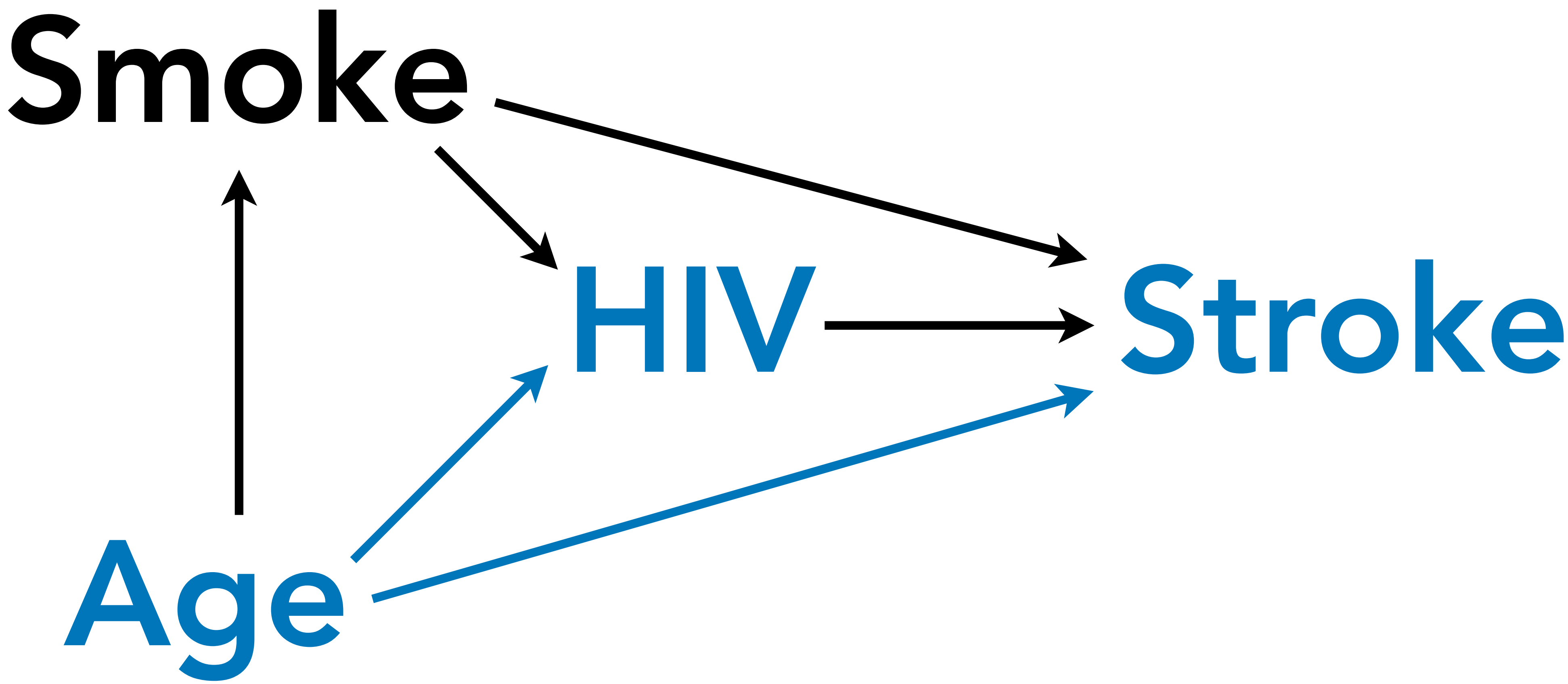

TABLE
TWO
FALLACY



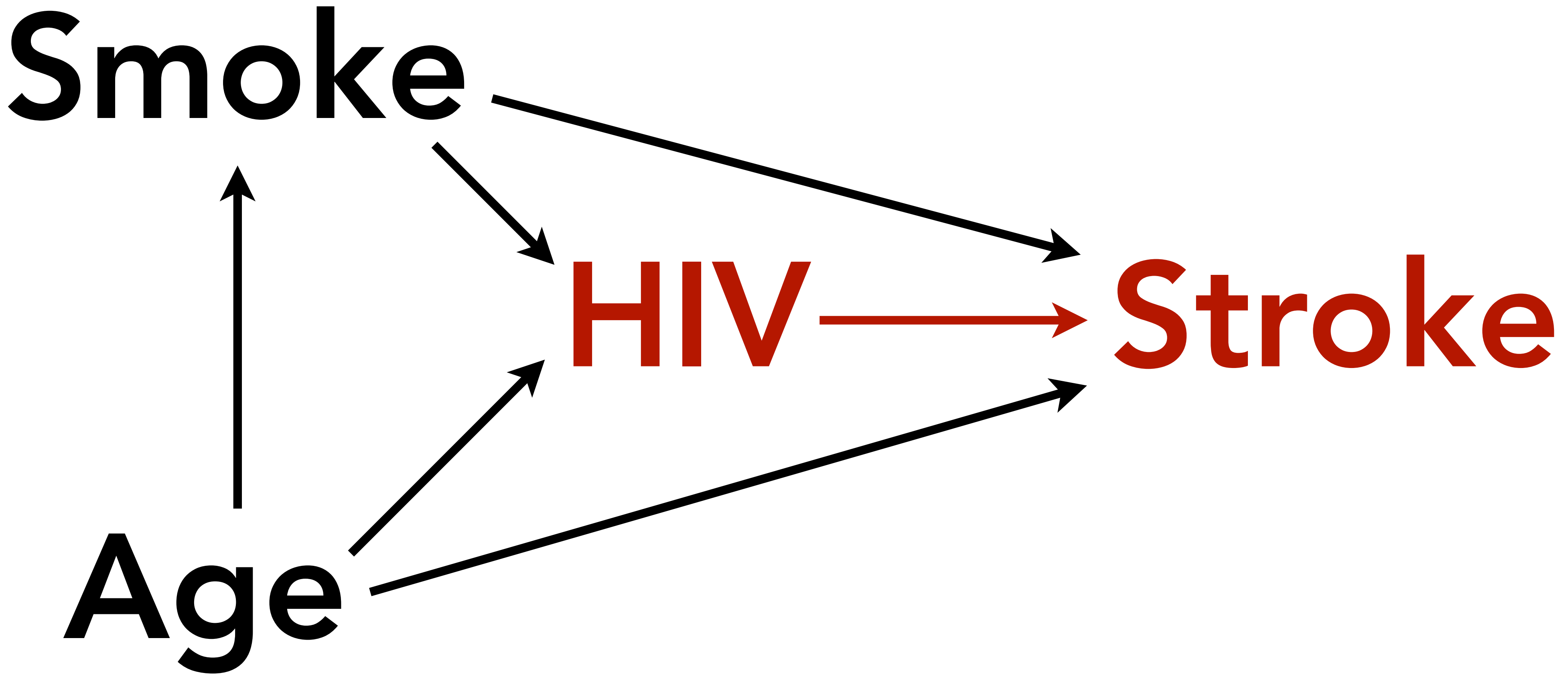
Westreich & Greenland 2013 The Table 2 Fallacy



Westreich & Greenland 2013 The Table 2 Fallacy



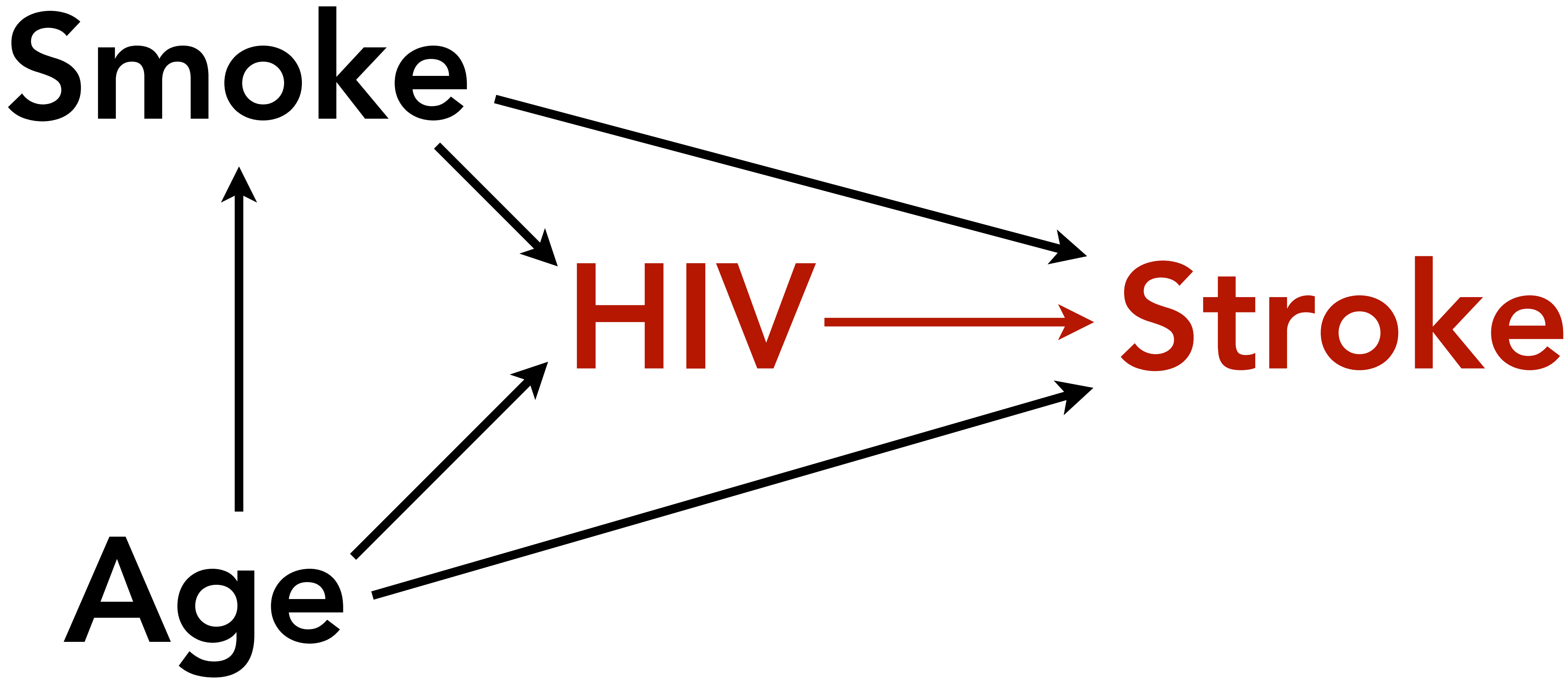
Westreich & Greenland 2013 The Table 2 Fallacy



Backdoor paths:

HIV ← Age → Stroke

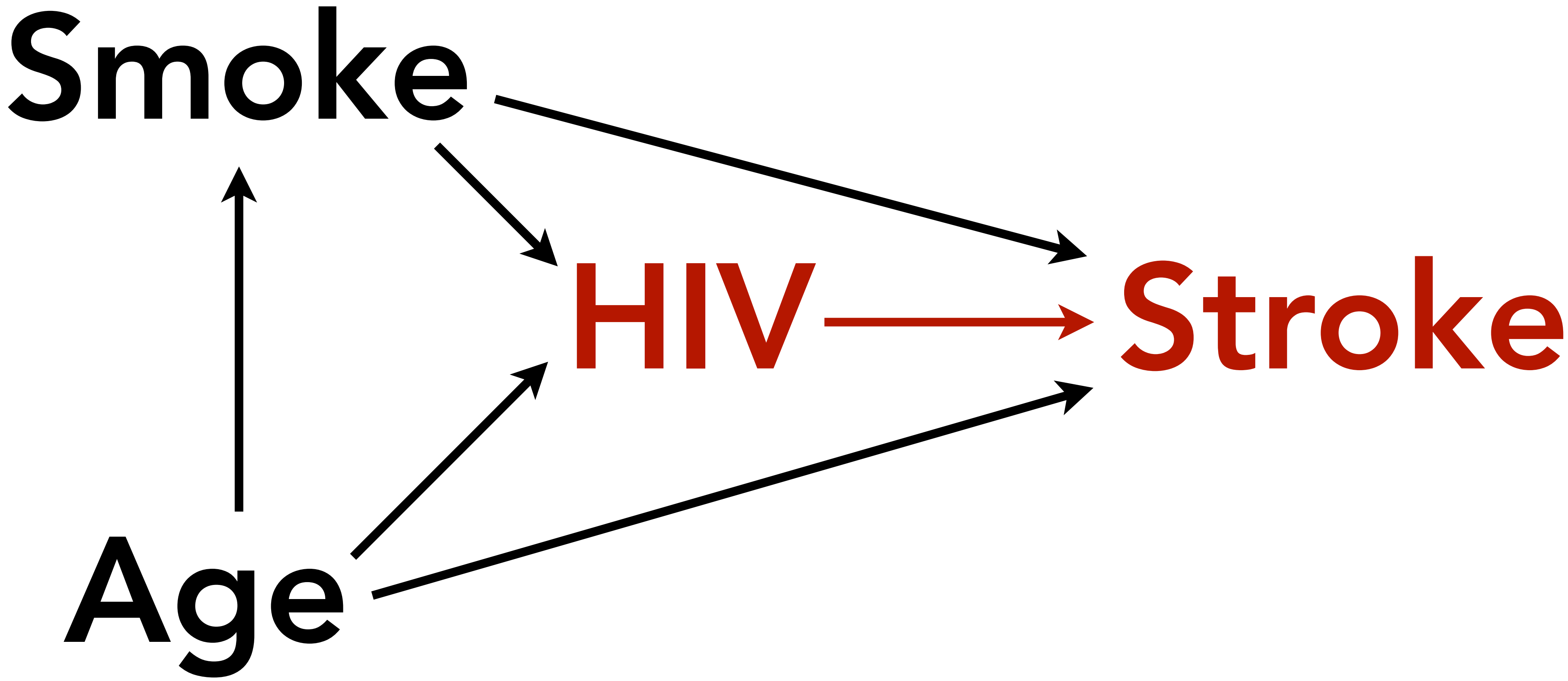
HIV ← Smoke → Stroke



Regression:

Stroke ~ HIV + Age + Smoke

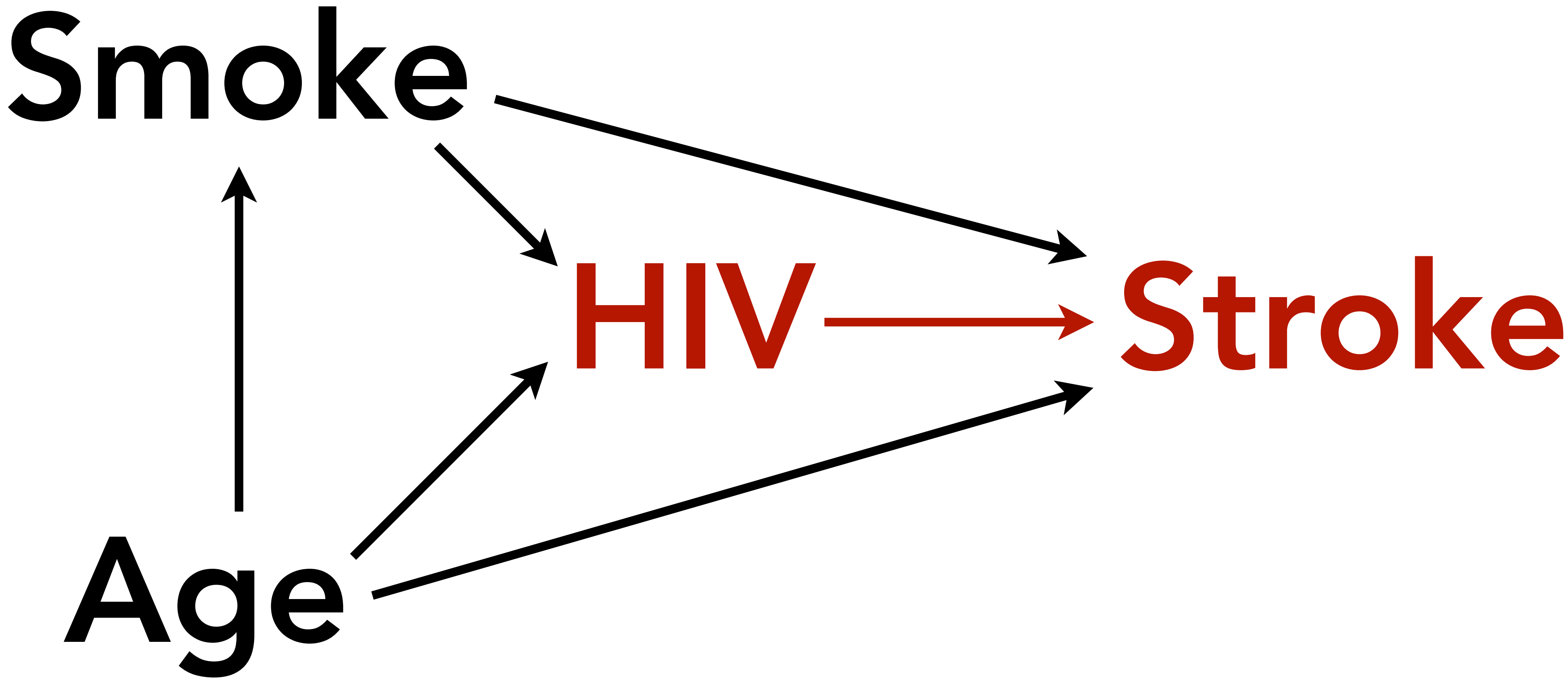
How to interpret coefficients?



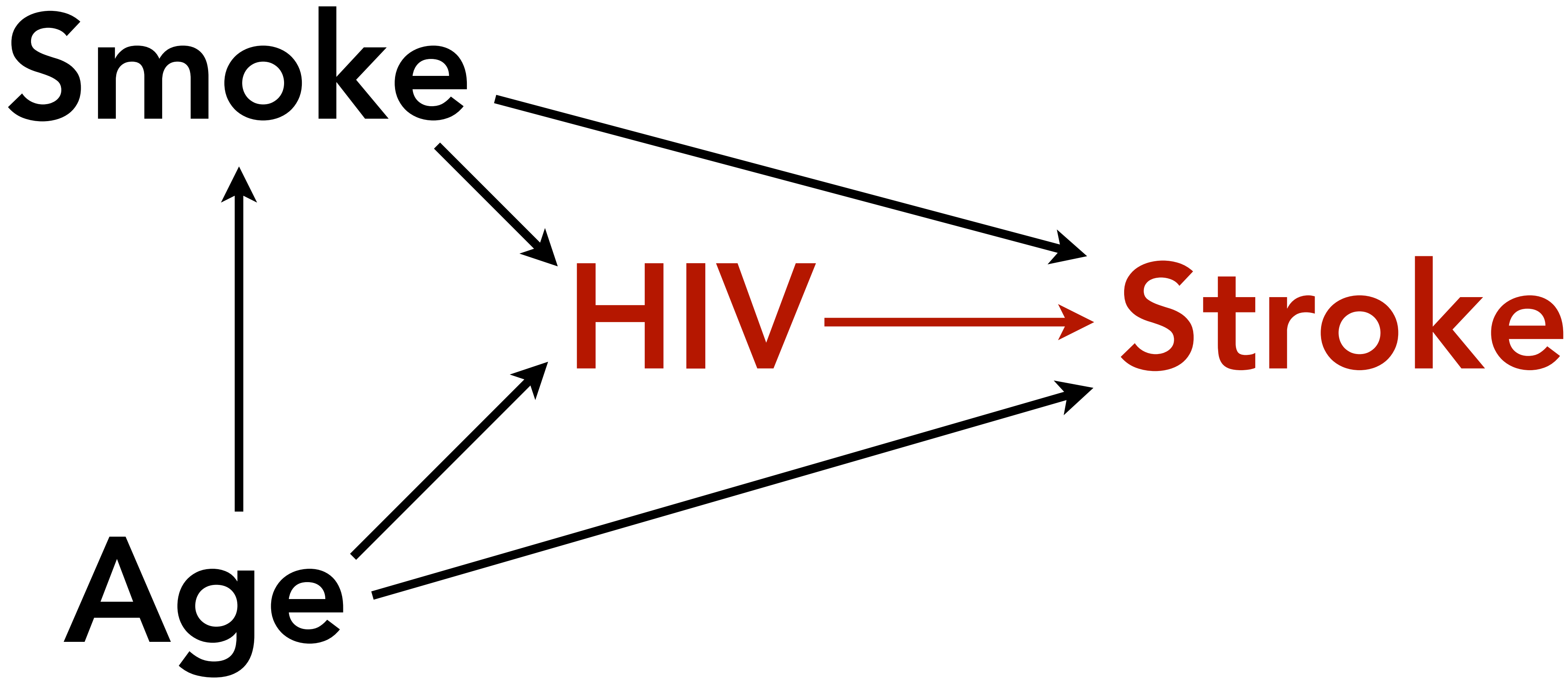
Coefficient on **HIV**:

The causal effect of HIV on Stroke

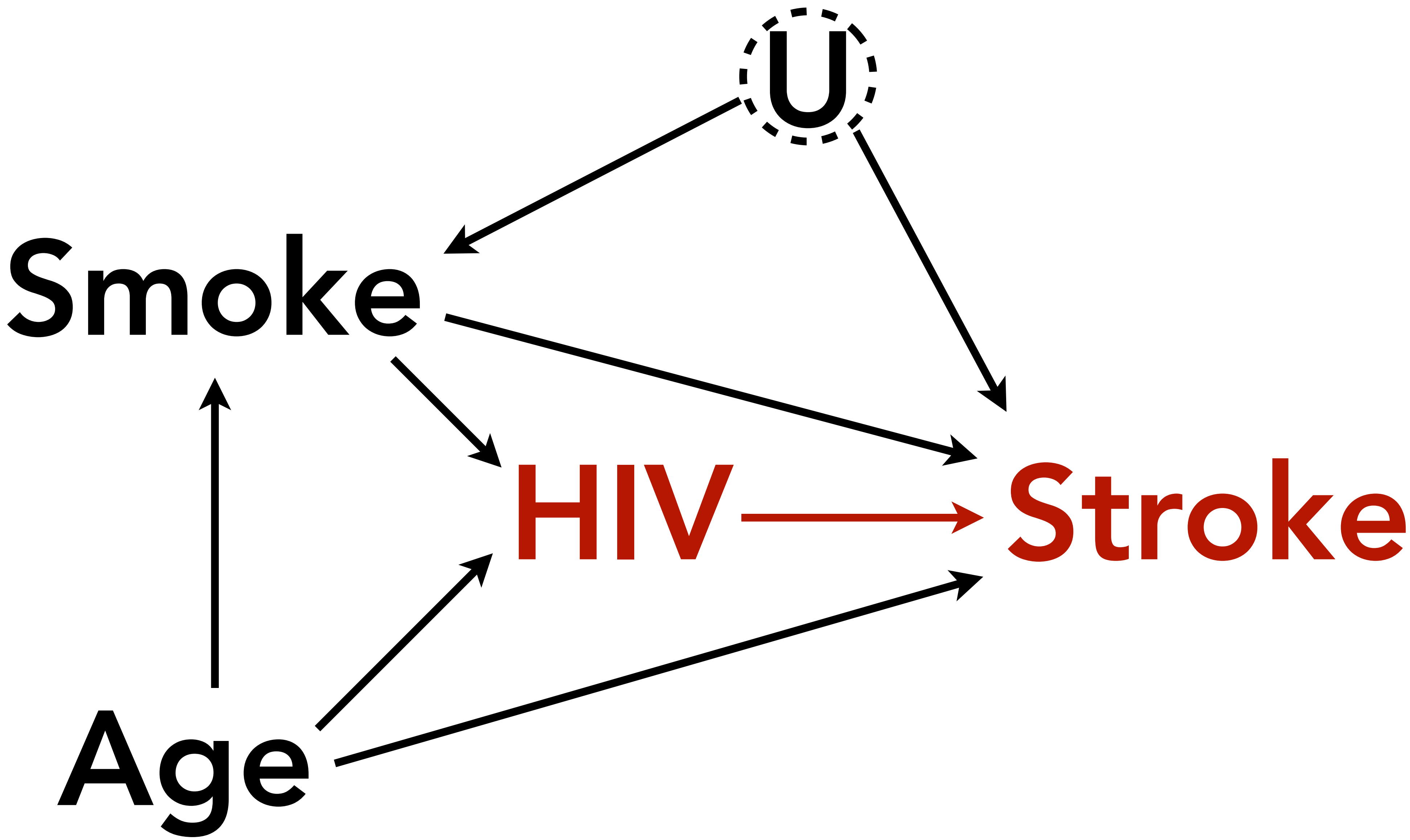
(Holding other factors constant)



Coefficient on **Age**:
The **partial** effect of Age, after removing contributions through HIV and Smoking.



Coefficient on **Smoke**:
The **partial** effect of Smoking, after removing contributions through HIV.



Now what is the valid interpretation of coefficients?

Table Two Fallacy

- Do not **present** or **interpret** individual coefficients as causal effects.
- Interpretation depends upon the **causal model**, not just the **statistical model**
- Statistical model designed to identify X –>Y will not also identify causal effect of all other variables
- Each causal query requires bespoke statistical model

Westreich & Greenland 2013 The Table 2 Fallacy


BAD
CONTROLS

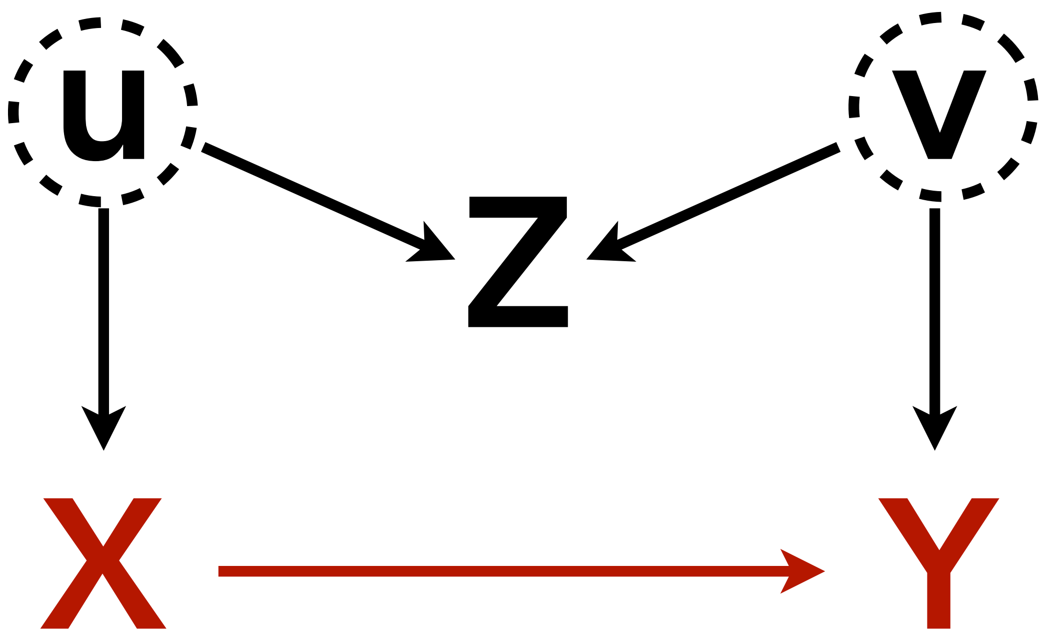
Good and Bad Controls

- "Control" variable: Variable introduced to an analysis in order to enable a causal estimate; not of interest itself
- Common heuristics for choosing control variables
  - Anything in the spreadsheet YOLO!
  - Any variables not highly **collinear**
  - Any **pre-treatment** measurement (baseline)

$$X \longrightarrow Y$$

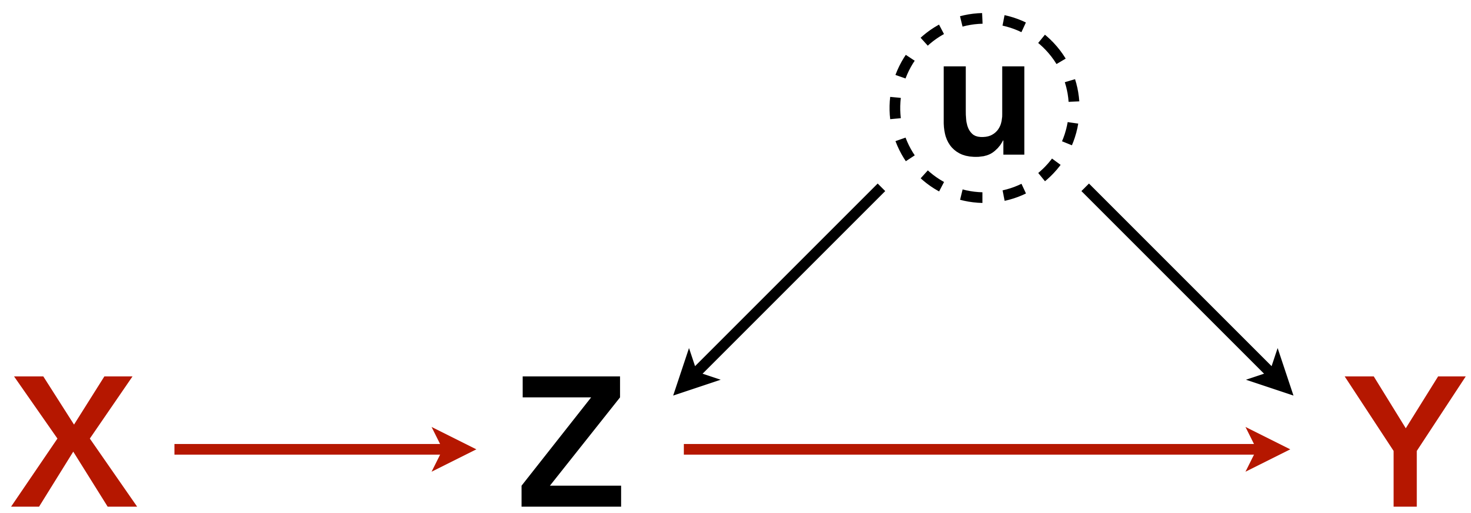
Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

M-bias



Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

Post-treatment bias



Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

Posttreatment bias is common

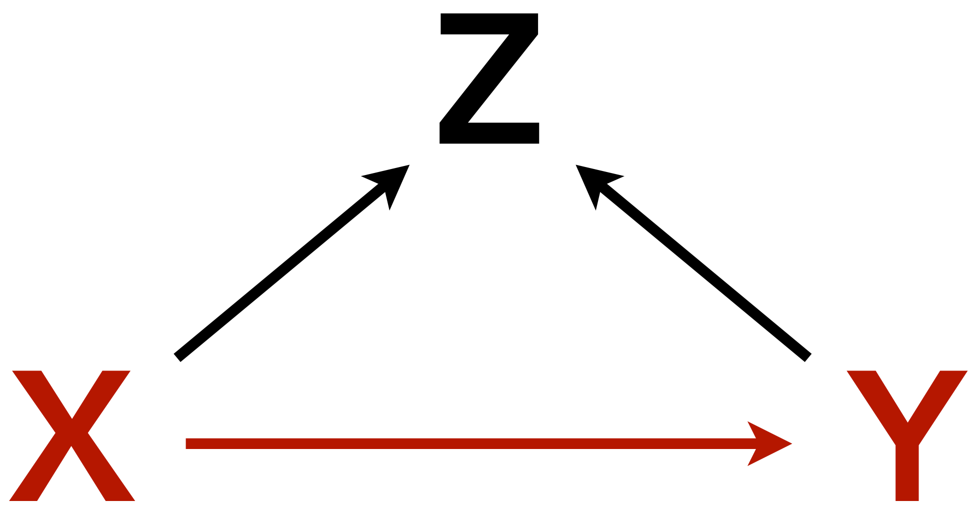
TABLE 1 Posttreatment Conditioning in Experimental Studies

Category	Prevalence
Engages in posttreatment conditioning	46.7%
Controls for/interacts with a posttreatment variable	21.3%
Drops cases based on posttreatment criteria	14.7%
Both types of posttreatment conditioning present	10.7%
No conditioning on posttreatment variables	52.0%
Insufficient information to code	1.3%

Note: The sample consists of 2012–14 articles in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* including a survey, field, laboratory, or lab-in-the-field experiment (n = 75).

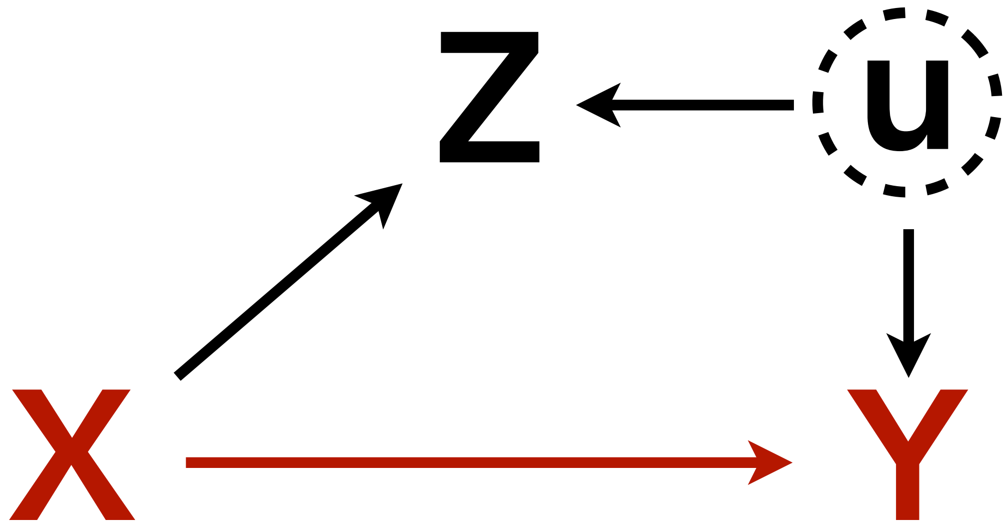
Montgomery et al 2018 How Conditioning on Posttreatment Variables Can Ruin Your Experiment

Selection bias



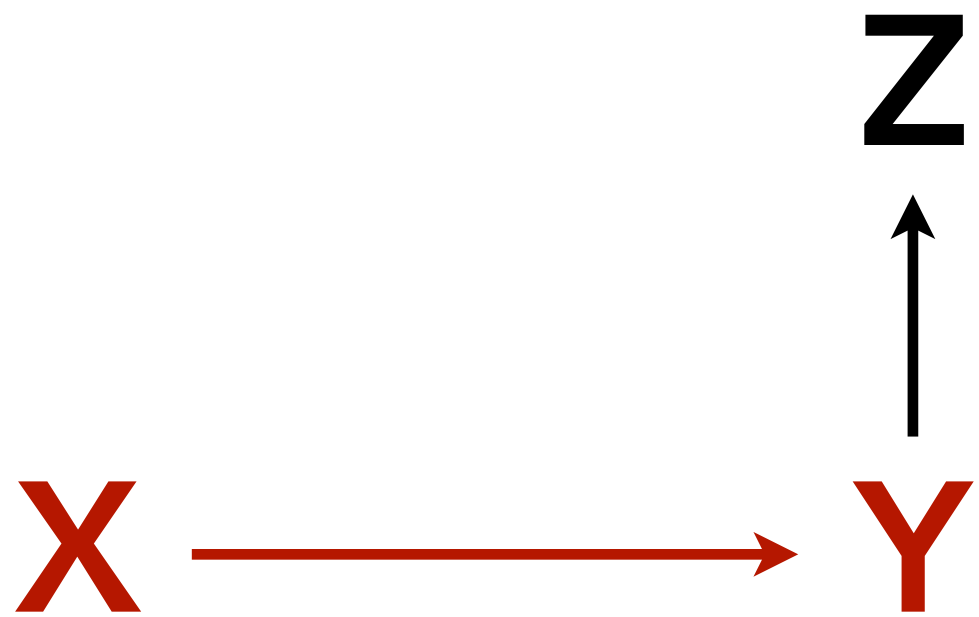
Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

Selection bias



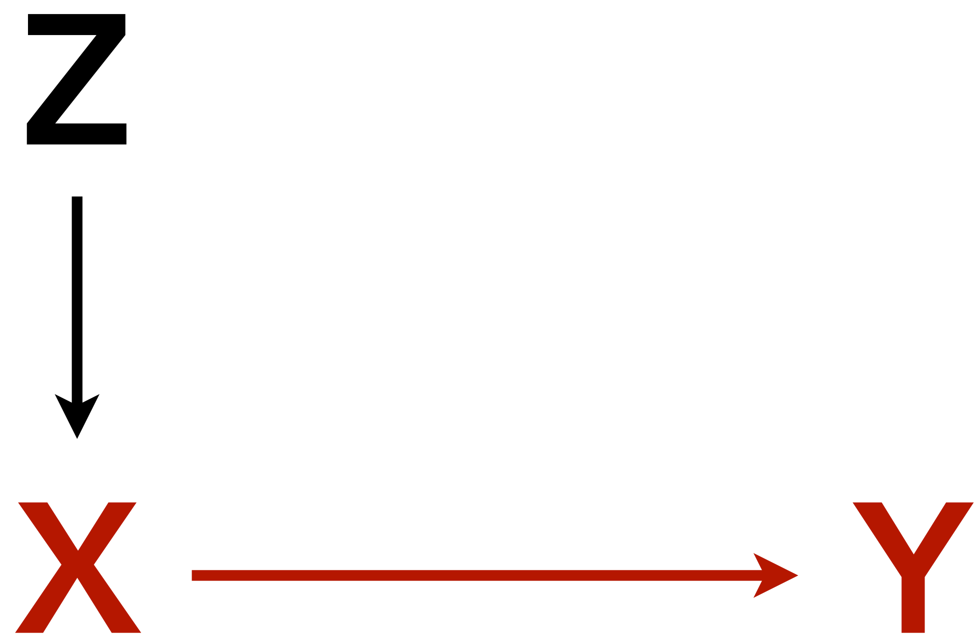
Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

Case-control bias



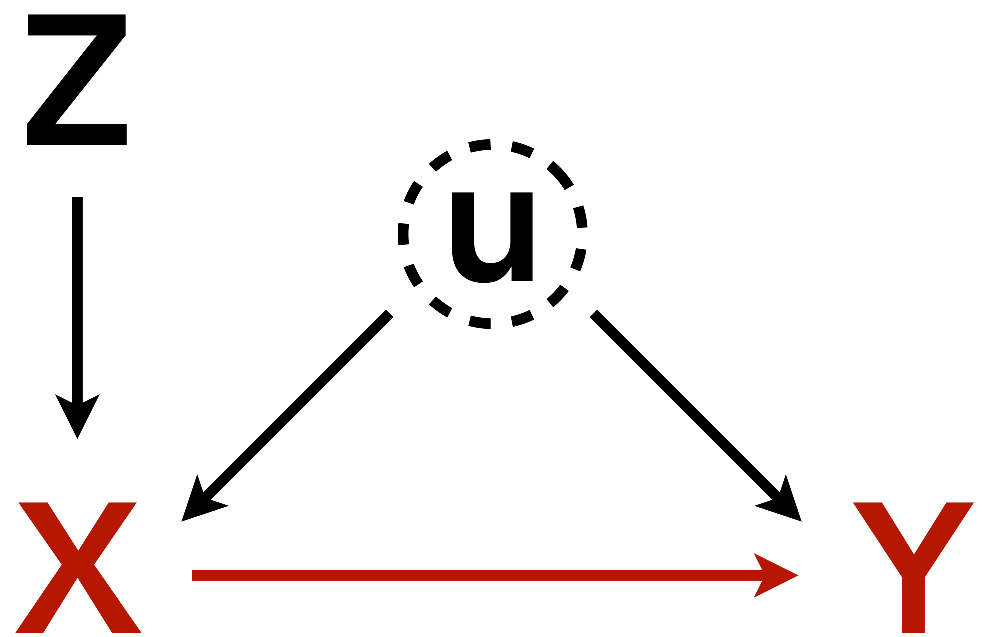
Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

Precision parasite



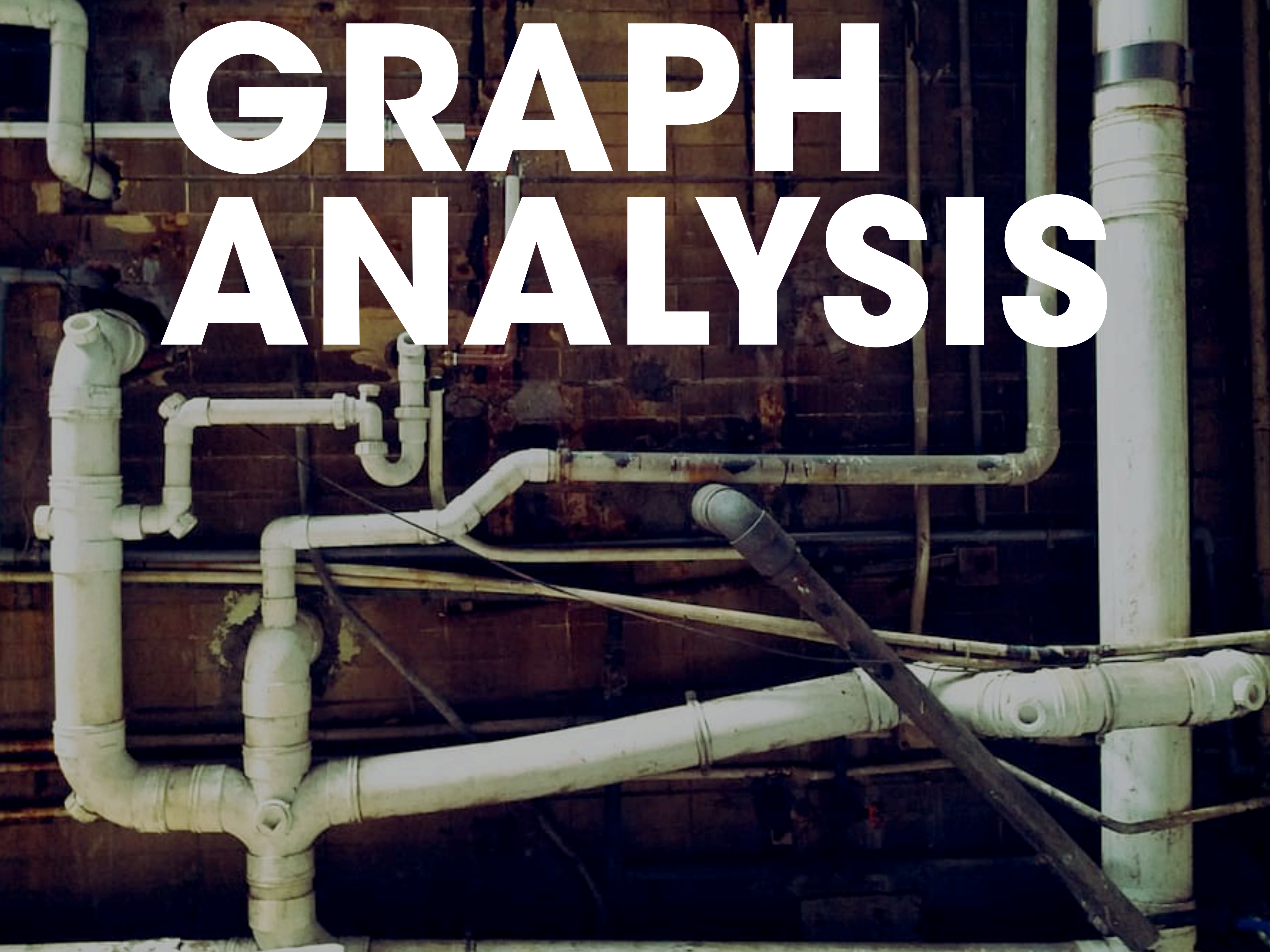
Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

Bias amplification

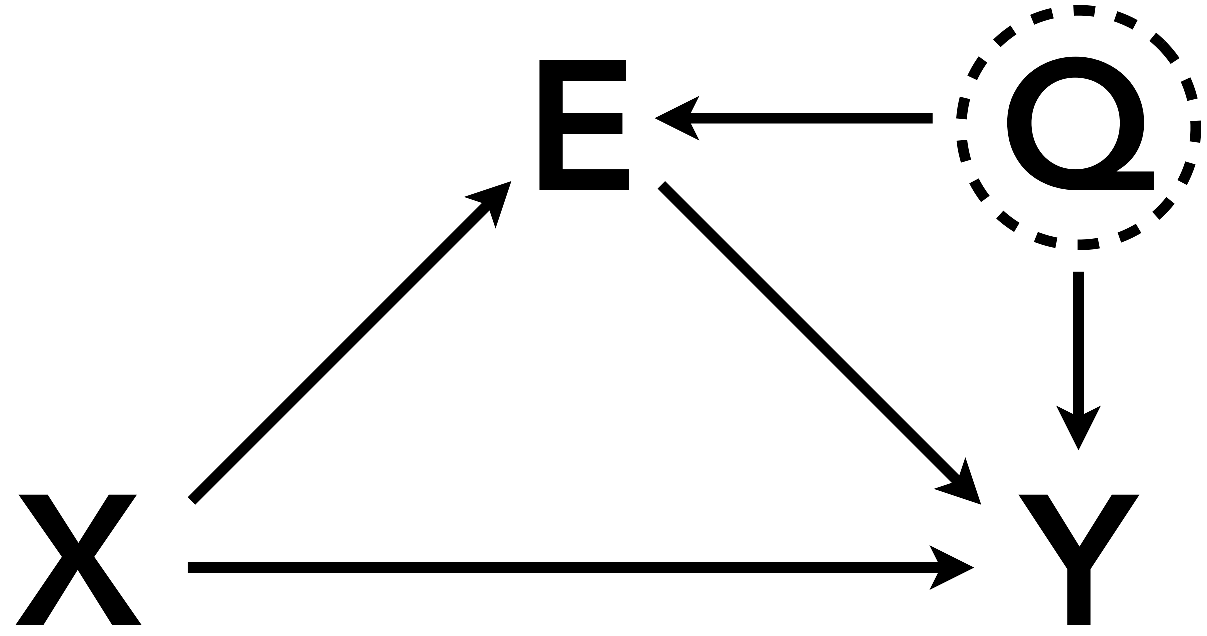


Cinelli, Forney, Pearl 2021 A Crash Course in Good and Bad Controls

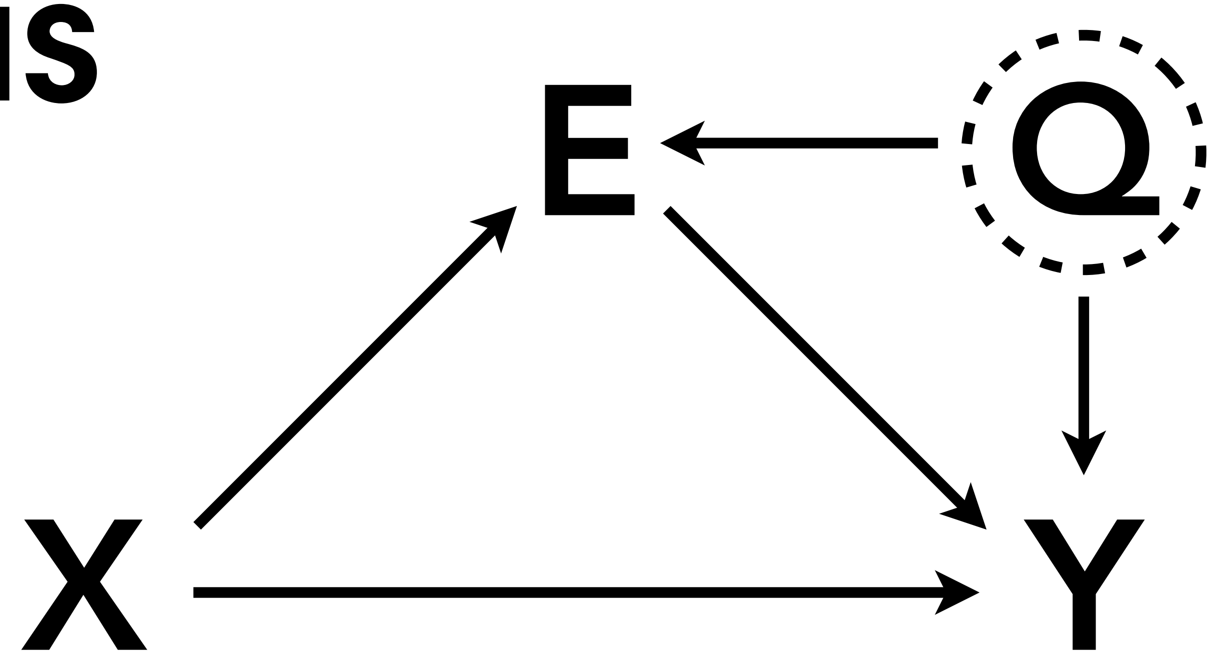
GRAPH ANALYSIS



Peer Bias



Peer Bias

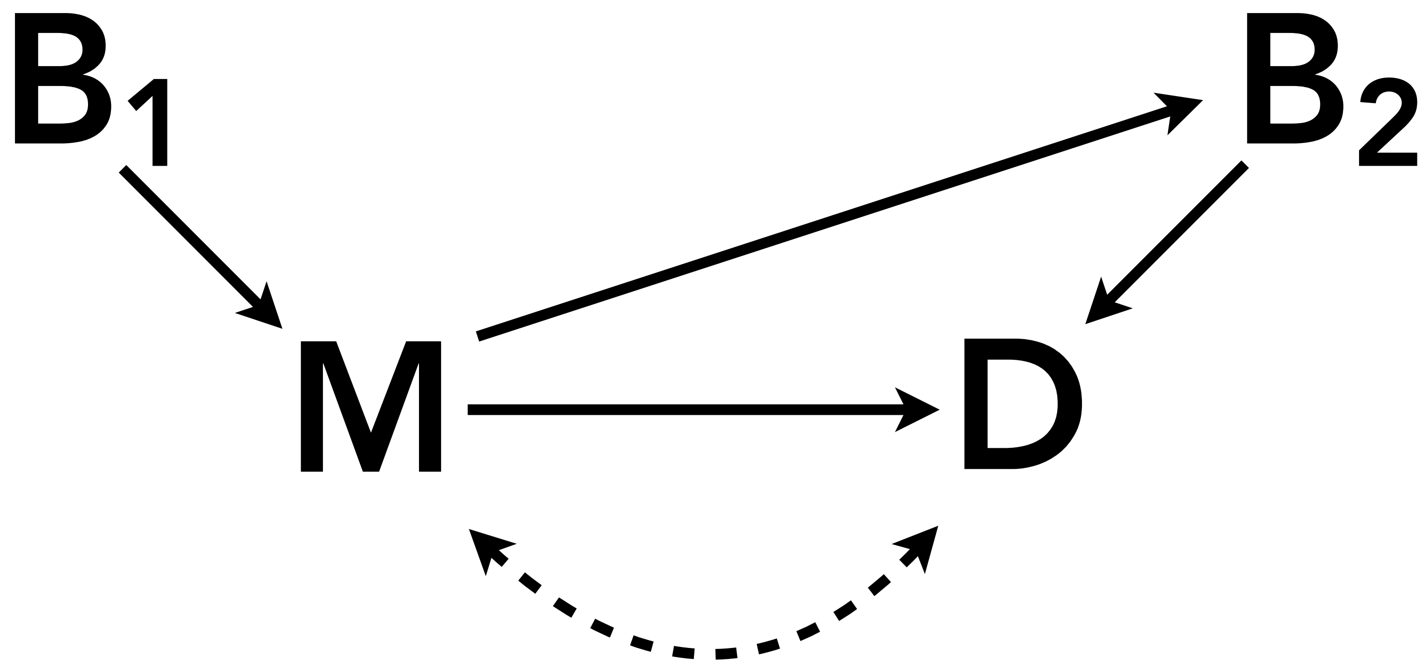


- Controlling for **E** opens **non-causal** (not **backdoor**!) path through **Q**
- We can estimate total causal effect of **X** on **Y**, but this is not what we want
- We cannot (yet) estimate direct causal effect

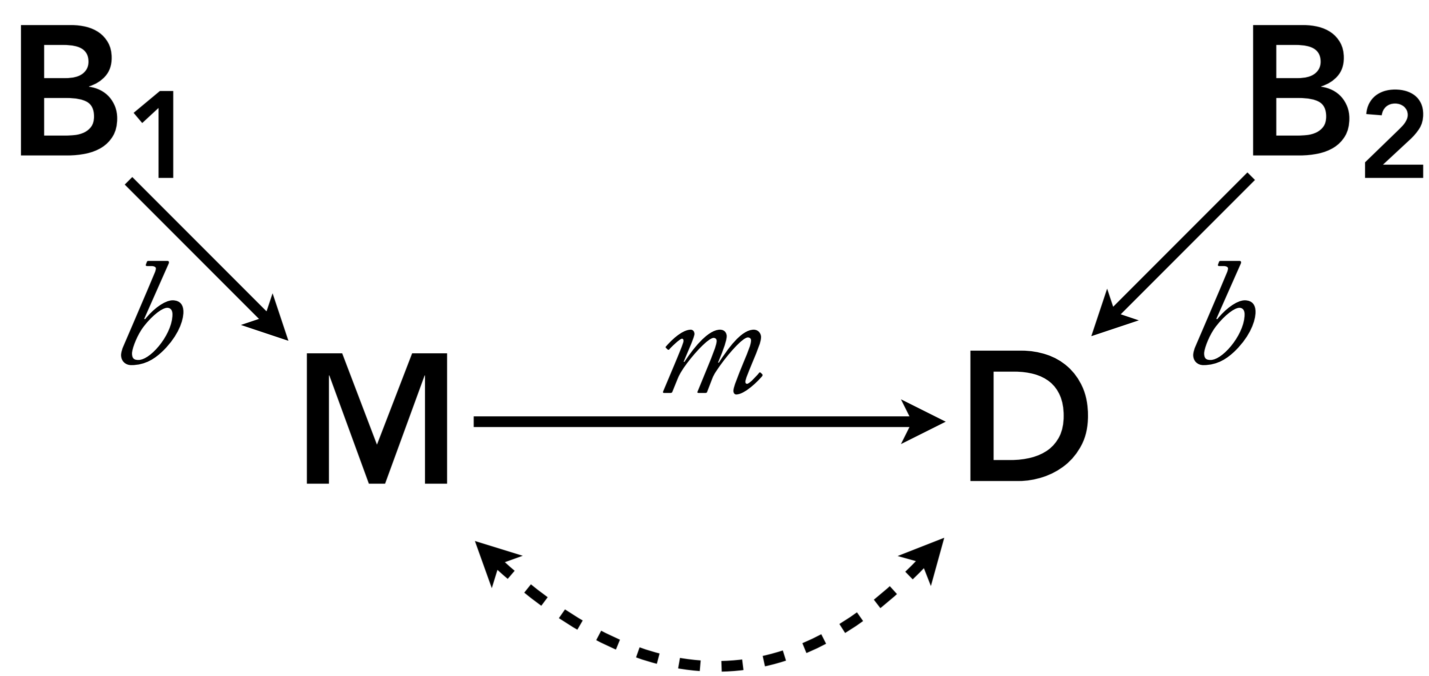
Graph Analysis

- More than backdoors
- Given a **structural causal model**, possible to logically derive a valid causal estimator (if one exists)
- Sometimes result looks like a multiple regression
- Often it does not
- We can use this strategy to solve our Two Moms problem

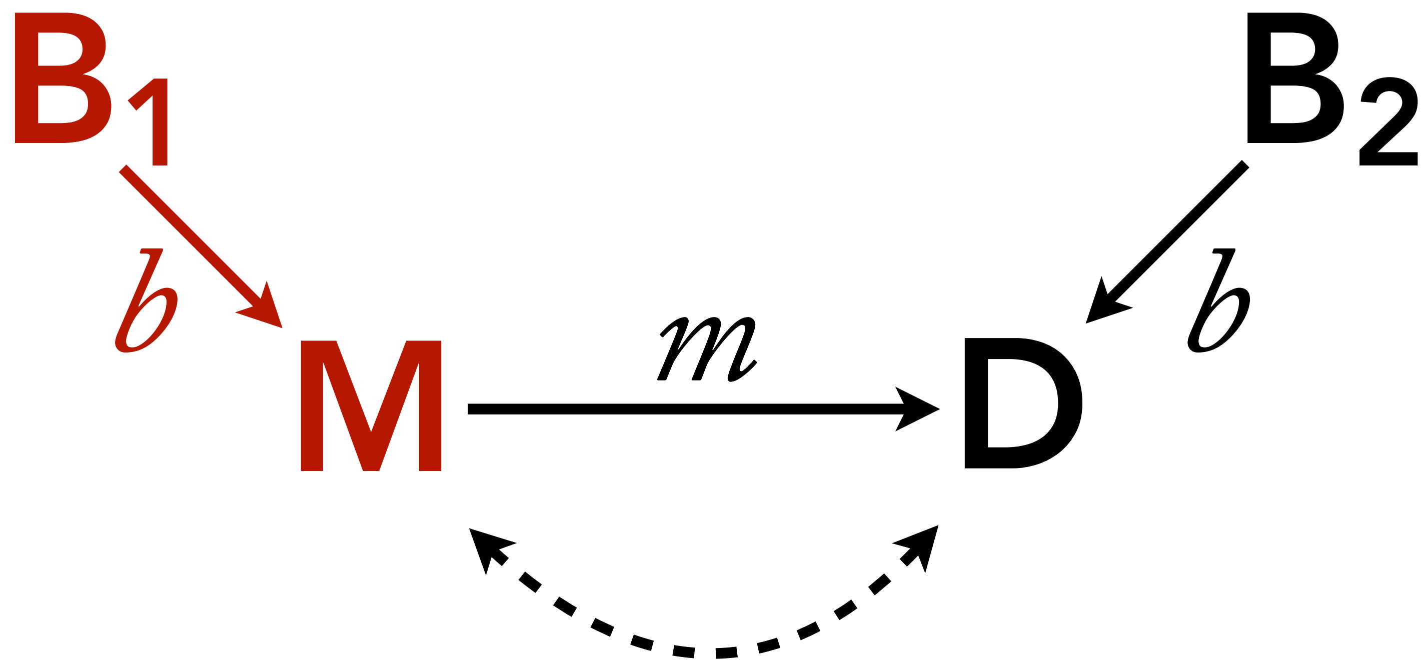
Two Moms



Two Moms

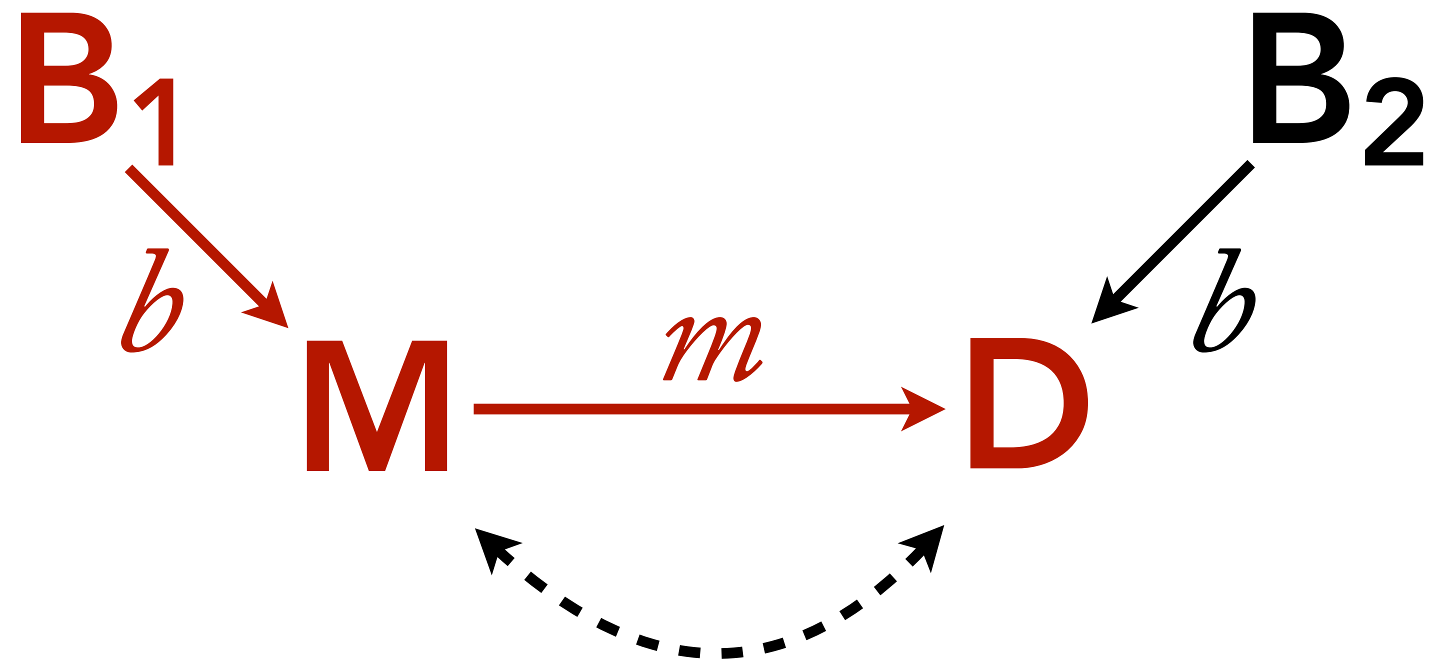


Two Moms

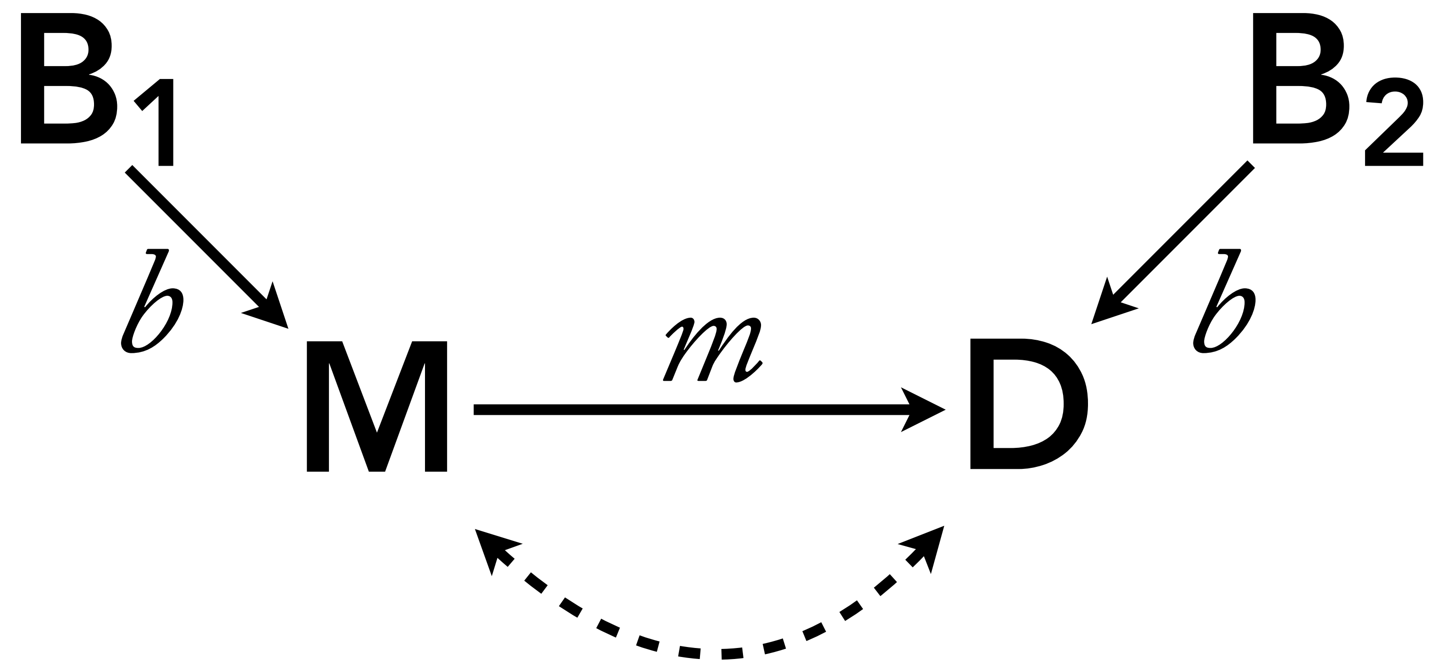


$$\mathrm{cov}(\mathrm{B}_1, \mathrm{M}) = b\ \mathrm{var}(\mathrm{B}_1)$$

Two Moms

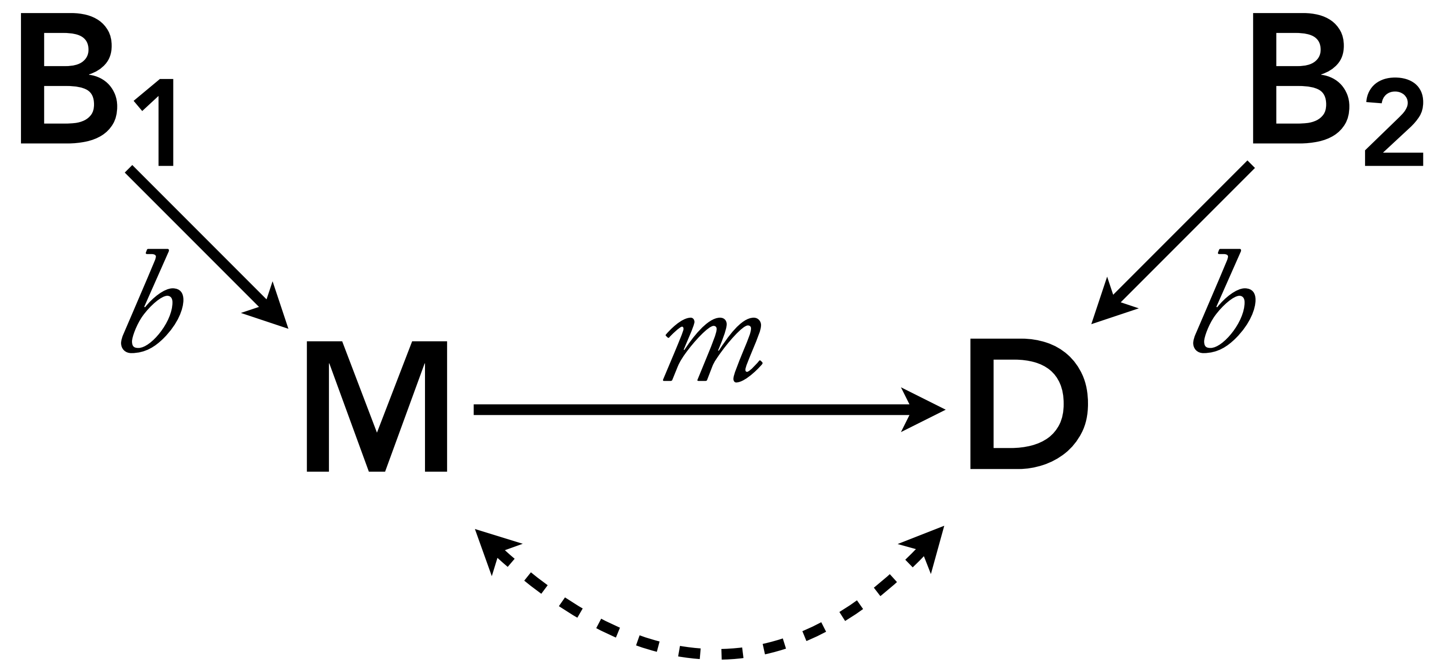


$$\mathrm{cov}(B_1,D) = b\; m\; \mathrm{var}(B_1)$$



$$\mathrm{cov}(B_1, M) = b\,\mathrm{var}(B_1)$$

$$\mathrm{cov}(B_1, D) = b\,m\,\mathrm{var}(B_1)$$



$$\mathrm{cov}(\mathrm{B}_1,\mathrm{M}) = b\ \mathrm{var}(\mathrm{B}_1)$$

$$\mathrm{cov}(\mathrm{B}_1,\mathrm{D}) = b\ m\ \mathrm{var}(\mathrm{B}_1)$$

$$m = \mathrm{cov}(\mathrm{B}_1,\mathrm{D})/\mathrm{cov}(\mathrm{B}_1,\mathrm{M})$$

Generative Implications

- Not all causal models are DAGs or linear equations (SEMs)
- But all causal models are **generative**
- Generative models have **causal implications**
  - differential equations, agent-based models, etc
- Different methods of analysis, but all can be analyzed