Poli-301:

The Political Science Discipline

TODAY'S AGENDA

1. Causal diagrams: DAGs
2. Controlling, adjusting
3. The Four Elemental Confounds

Last week

We want to know if X **causes** Y using data

Some correlations are causal; many aren't

How can we tell?

Experiments

Experiments are the “gold standard” for research

We randomly assign X, so we know differences in Y are unlikely to be driven by other factors

Rats and hyperglycemia

Can this new drug lower blood-sugar levels?

```
A tibble: 100 x 3
    rat drug      blood_sugar
  <int> <fct>        <I<dbl>>
1     1 drug           -2.97
2     2 placebo        -0.823
3     3 placebo         1.26
4     4 placebo        -1.22
5     5 drug           -2.53
6     6 drug           -1.94
7     7 drug           -0.934
8     8 drug           -1.83
9     9 placebo         0.785
0    10 drug           -1.91
```

```
rat_experiment$drug
 [1] drug    placebo placebo placebo drug    drug    drug    drug    placebo
[10] drug    placebo drug    placebo placebo drug    placebo placebo drug
[19] placebo drug    placebo placebo placebo drug    placebo placebo drug
[28] placebo drug    drug    placebo drug    placebo drug    drug    drug
[37] drug    placebo drug    drug    drug    placebo placebo drug    placebo
[46] drug    drug    drug    drug    placebo placebo placebo drug    drug
[55] drug    placebo placebo drug    placebo drug    placebo drug    placebo
[64] placebo drug    drug    drug    drug    drug    placebo drug    drug
[73] placebo placebo drug    drug    drug    drug    placebo drug    drug
[82] placebo drug    placebo placebo drug    drug    placebo placebo placebo
[91] drug    drug    placebo placebo placebo placebo drug    drug    placebo
100] placebo
evels: placebo drug
```

```
lm(blood_sugar ~ drug, data = rat_experiment)
```

```
  term      estimate
  <chr>        <dbl>
1 intercept    0.043
2 drugdrug    -1.90
```

People and media exposure

Is biased TV news making the American public more extreme?

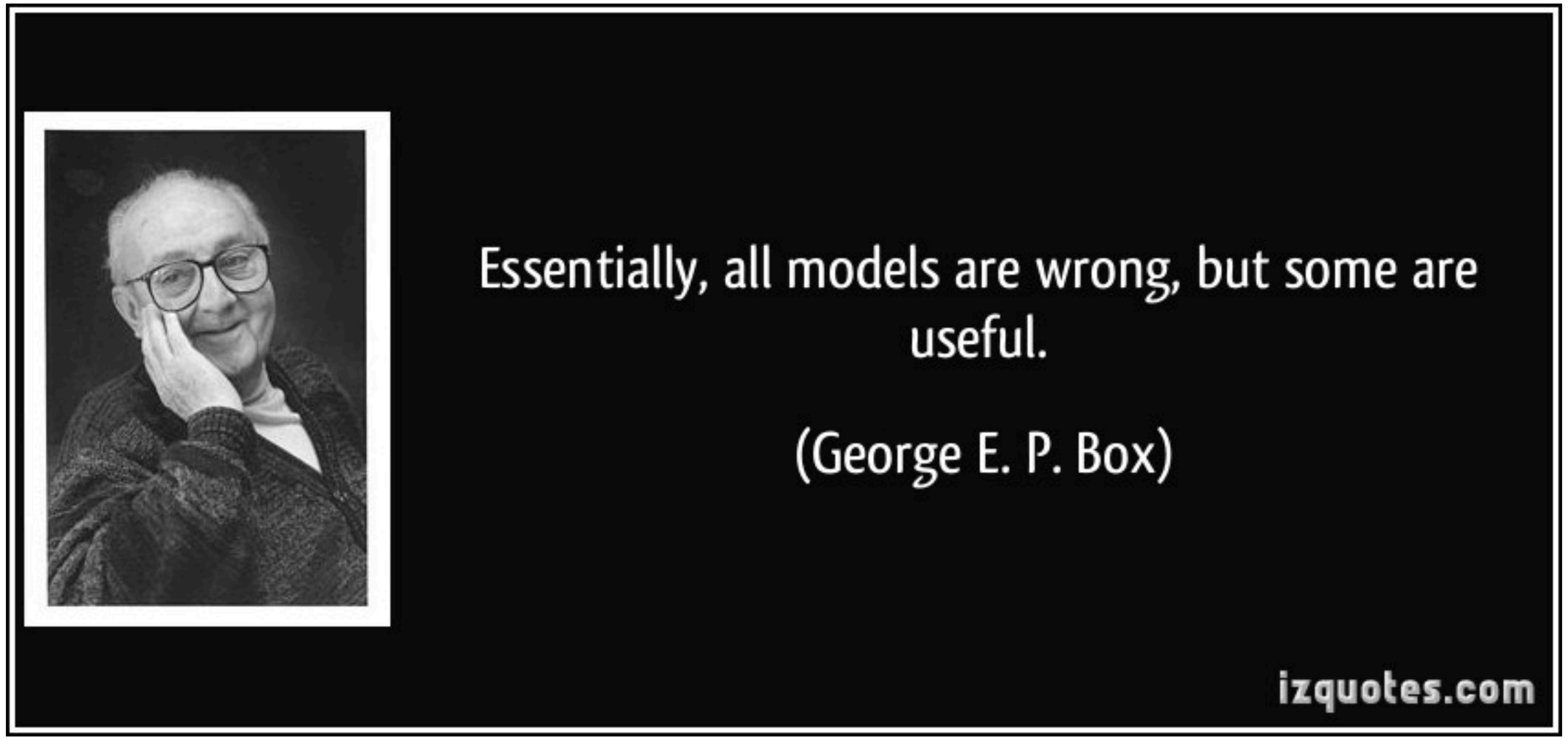
```
   person    media liberal
   <chr>     <fct> <I<dbl>>
 1 Guisel    Other        3
 2 Nieta     MSNBC        6
 3 Laquetha  MSNBC        9
 4 Jyrese    Other        2
 5 Lafiamma  Other        6
 6 Ahmirah   Other        7
 7 Eissa     MSNBC        4
 8 Shrenik   MSNBC        9
 9 Rossalind Other        6
10 Kaliyah   MSNBC        4
```

```
lm(liberal ~ media, data = media_bias)
```

```
  term        estimate
  <chr>          <dbl>
1 intercept       4.02
2 mediaMSNBC      2.03
```

How is this different from rats?

Causal models



We need **a model** to get at causality

A model is our idea of the **data-generating process**

The model tells us how to find a causal effect (if possible)

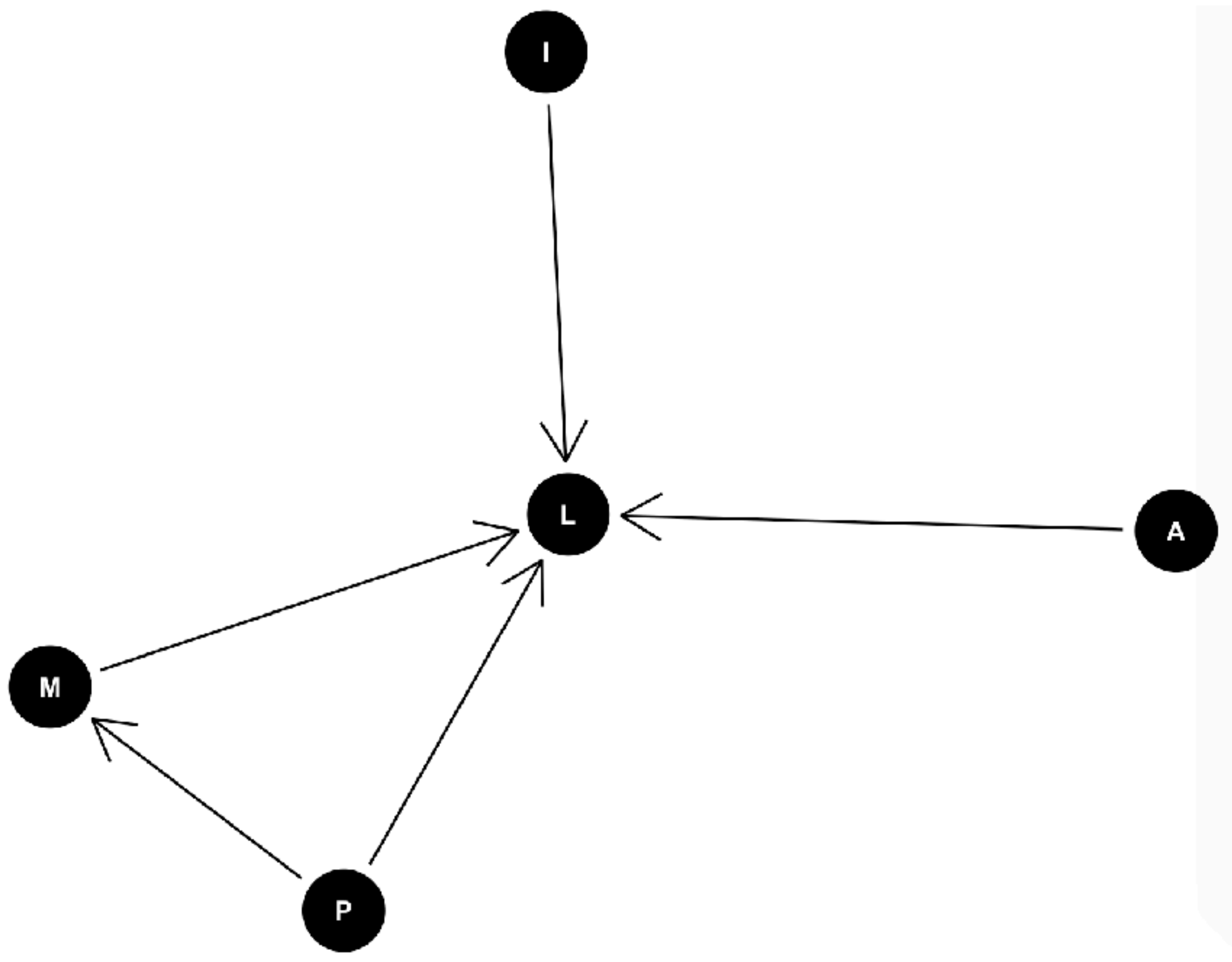
DAGs

Directed acyclical graphs

Directed = arrows

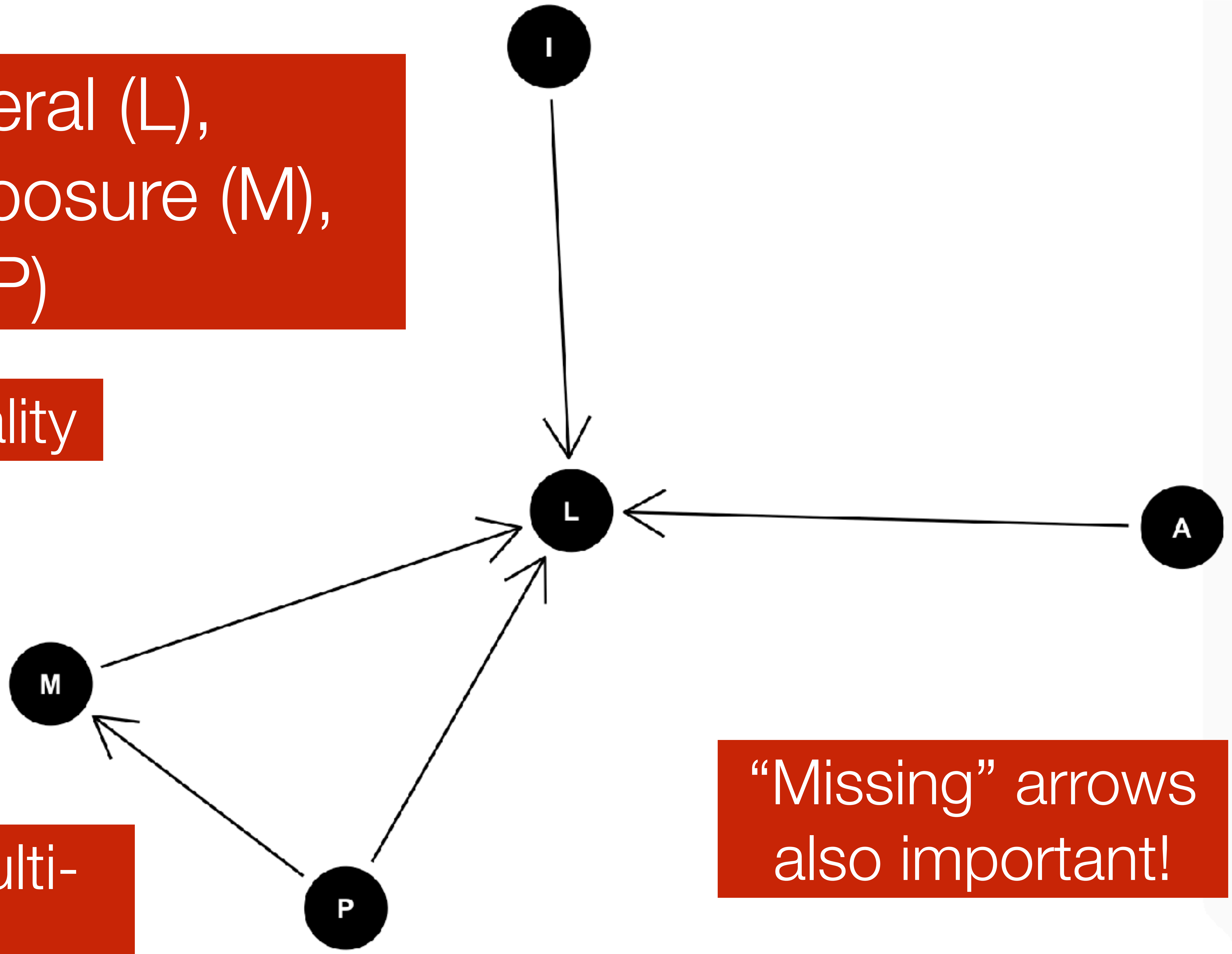
Acyclical = no loops

Graphs = nodes and edges



Income (I), Liberal (L),
Age (A), Media Exposure (M),
Parents (P)

—> = "flow" of causality



"flow" can also be multi-step

"Missing" arrows also important!

Your Turn

Give me an outcome that matters

We'll build a DAG

DAGs

DAGs encode everything we know about some **process**

We can see all of our **assumptions** and how they fit together

For example: the effect of Parents on Liberal happens directly and indirectly (thru Media)

Identification

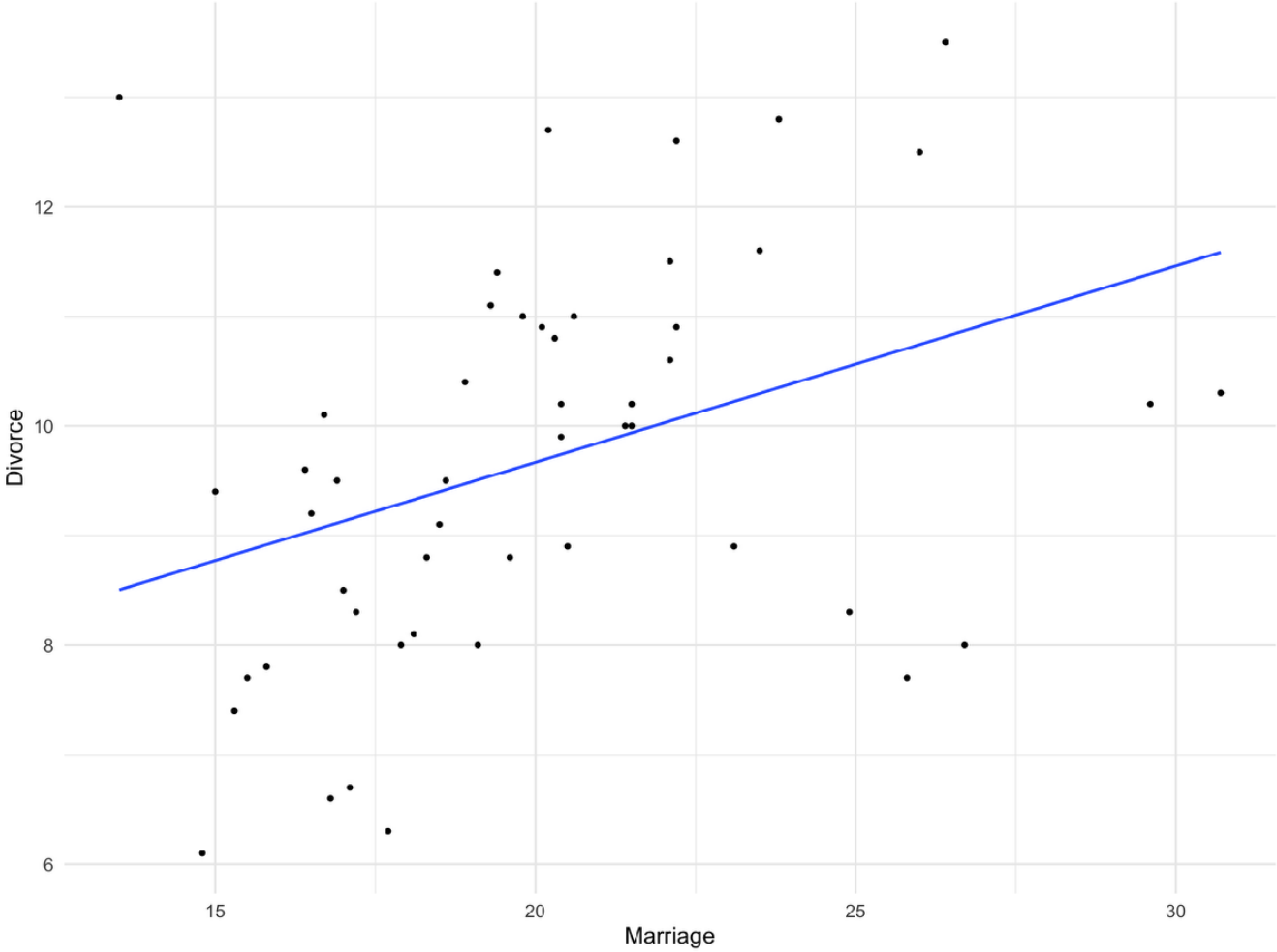
We can also use DAGs to isolate the effect of one variable on another (e.g., Media —> Liberal)

This process is called "identification"

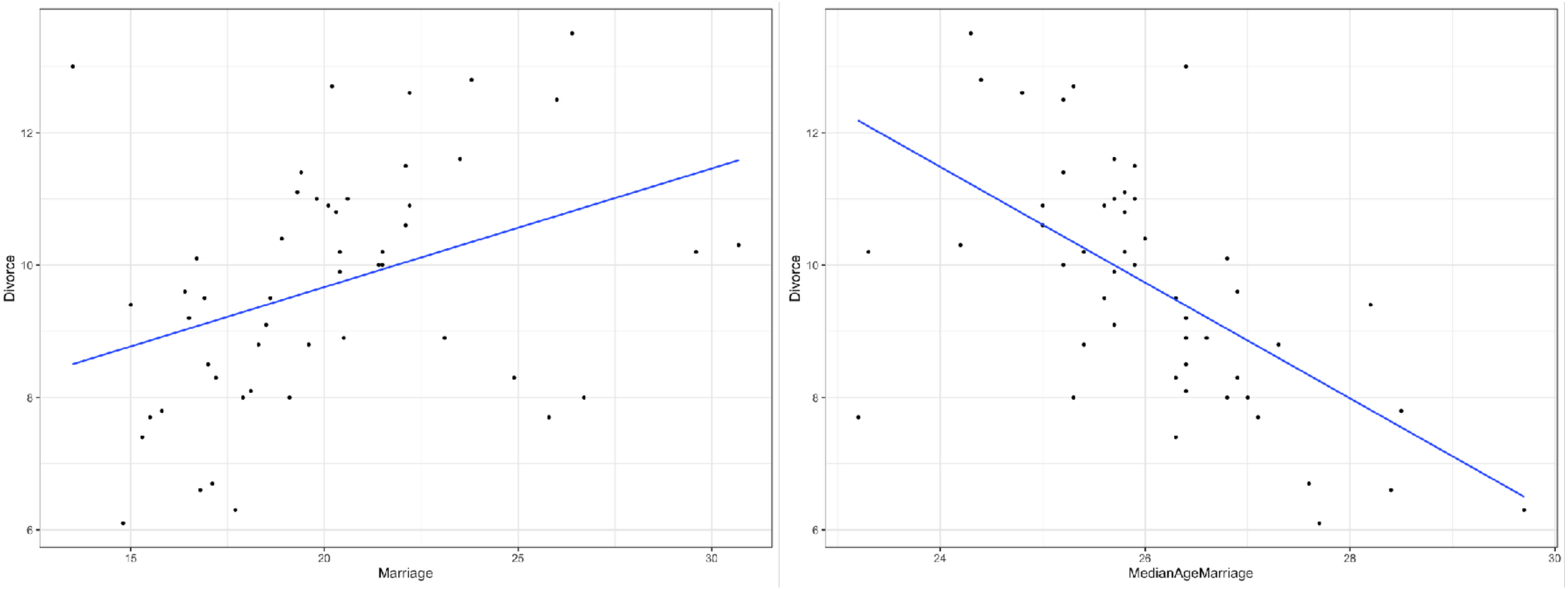
We say we want to "*identify the effect of X on Y*"

Does marriage cause divorce?

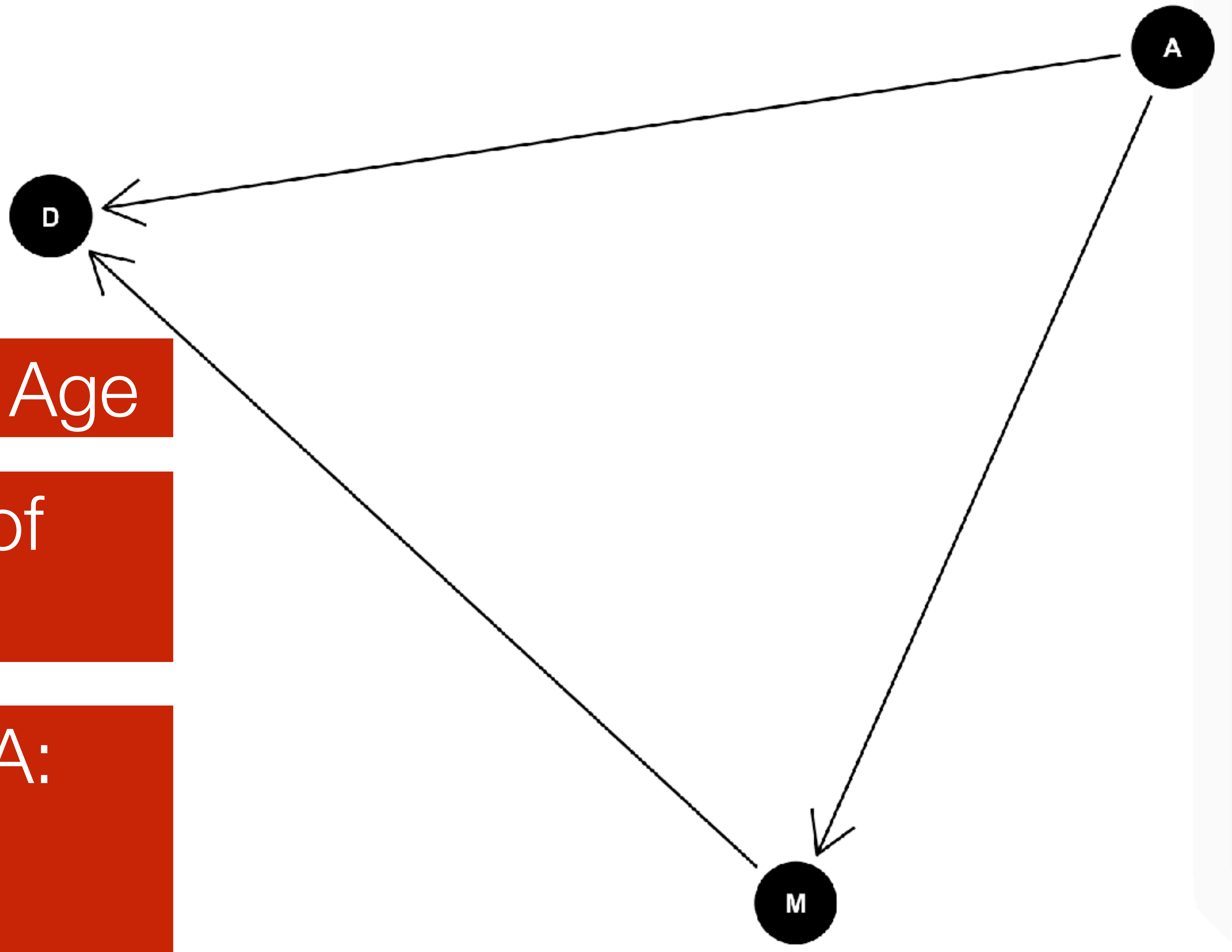
	Loc	Marriage	Divorce
1	AL	20.2	12.7
2	AK	26.0	12.5
3	AZ	20.3	10.8
4	AR	26.4	13.5
5	CA	19.1	8.0
6	CO	23.5	11.6
7	CT	17.1	6.7
8	DE	23.1	8.9
9	DC	17.7	6.3
0	FL	17.0	8.5
1	GA	22.1	11.5
2	HI	24.9	8.3
3	ID	25.8	7.7
4	IL	17.9	8.0
5	IN	19.8	11.0
6	IA	21.5	10.2
7	KS	22.1	10.6
8	KY	22.2	12.6
9	LA	20.6	11.0
0	ME	13.5	13.0



Spurious association



The DAG



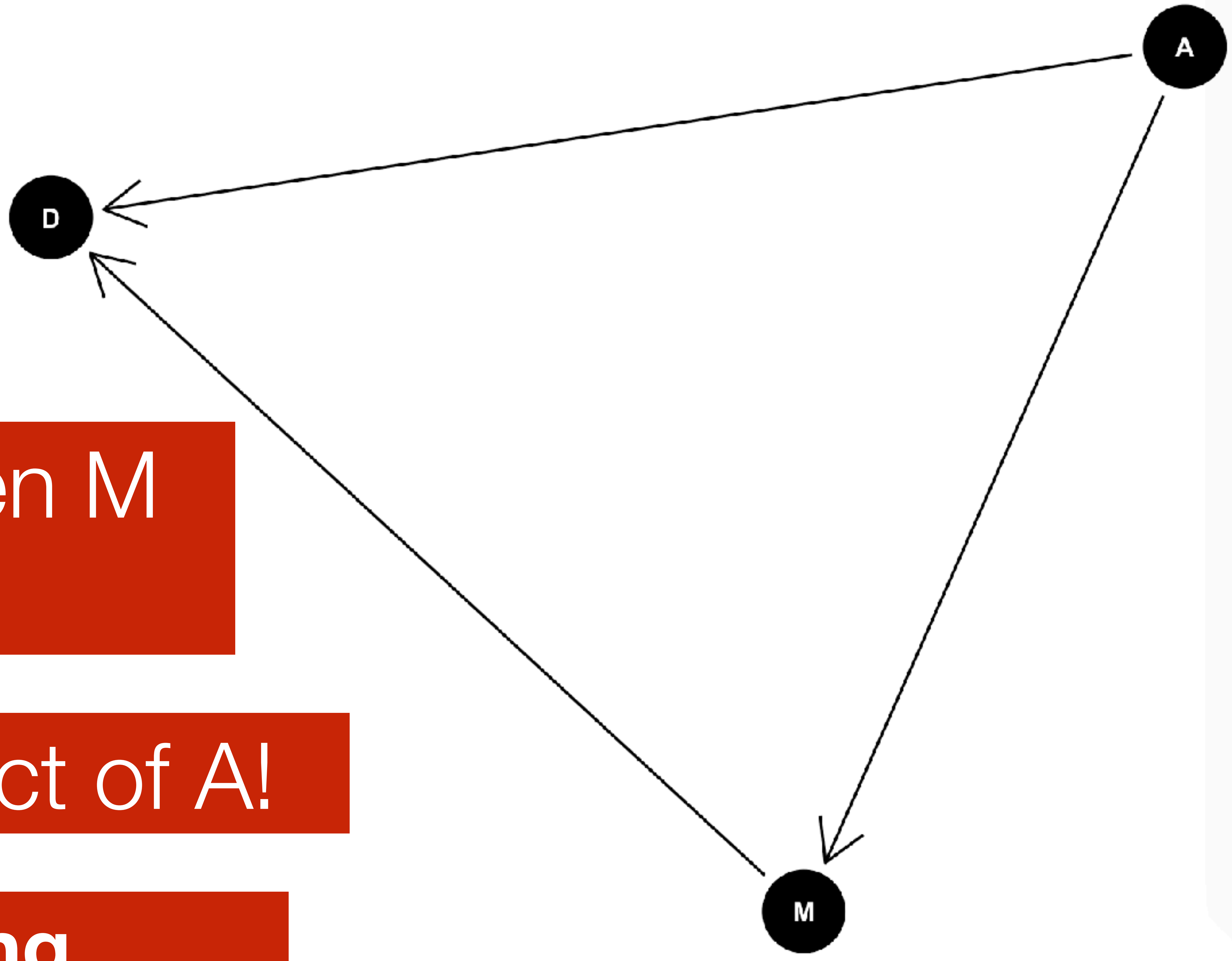
Marriage is a function of Age

Divorce is a function of marriage and age

Two causal paths for A:
A —> M —> D
A —> D

Does marriage cause divorce?

The problem



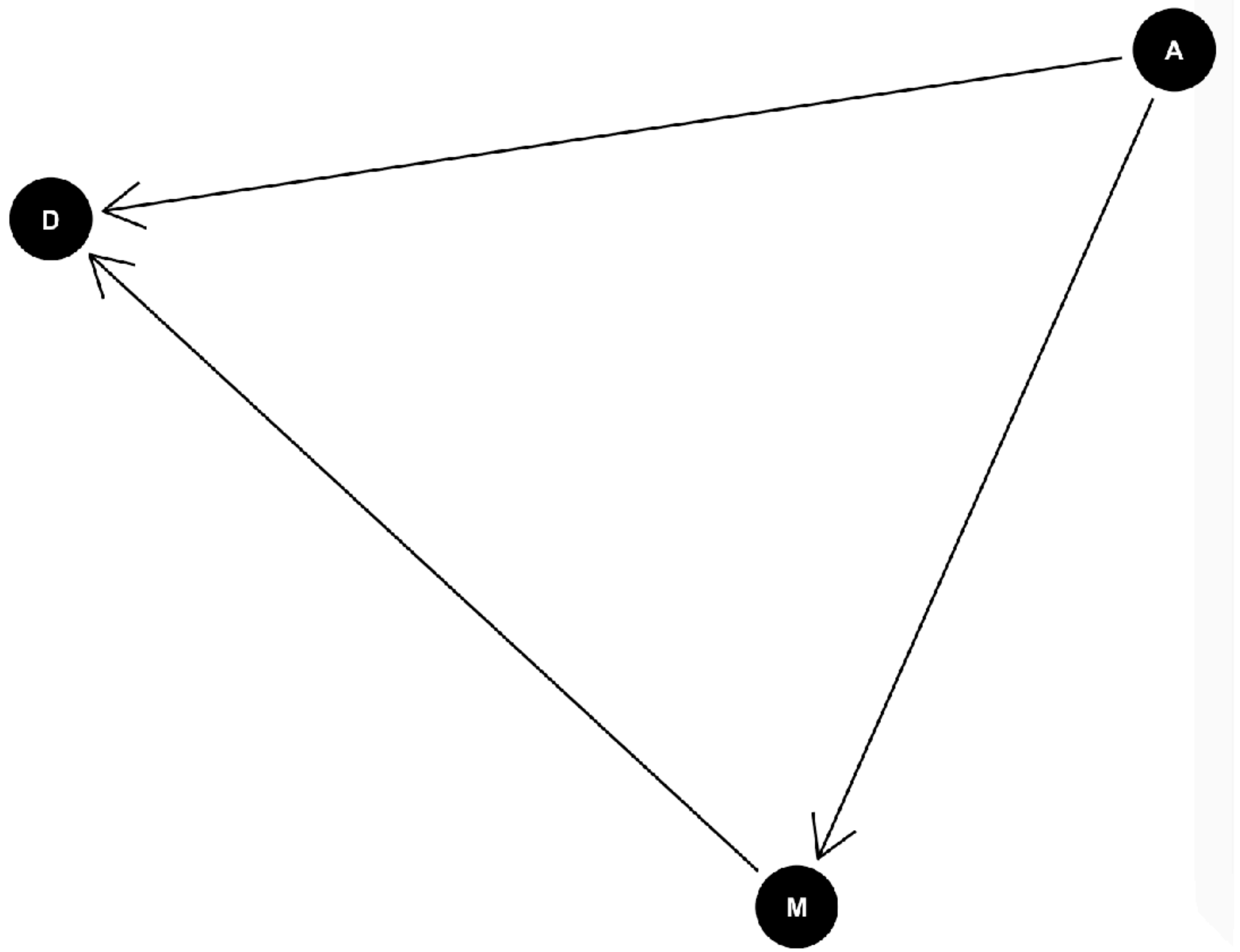
Part of correlation between M and D is caused by A

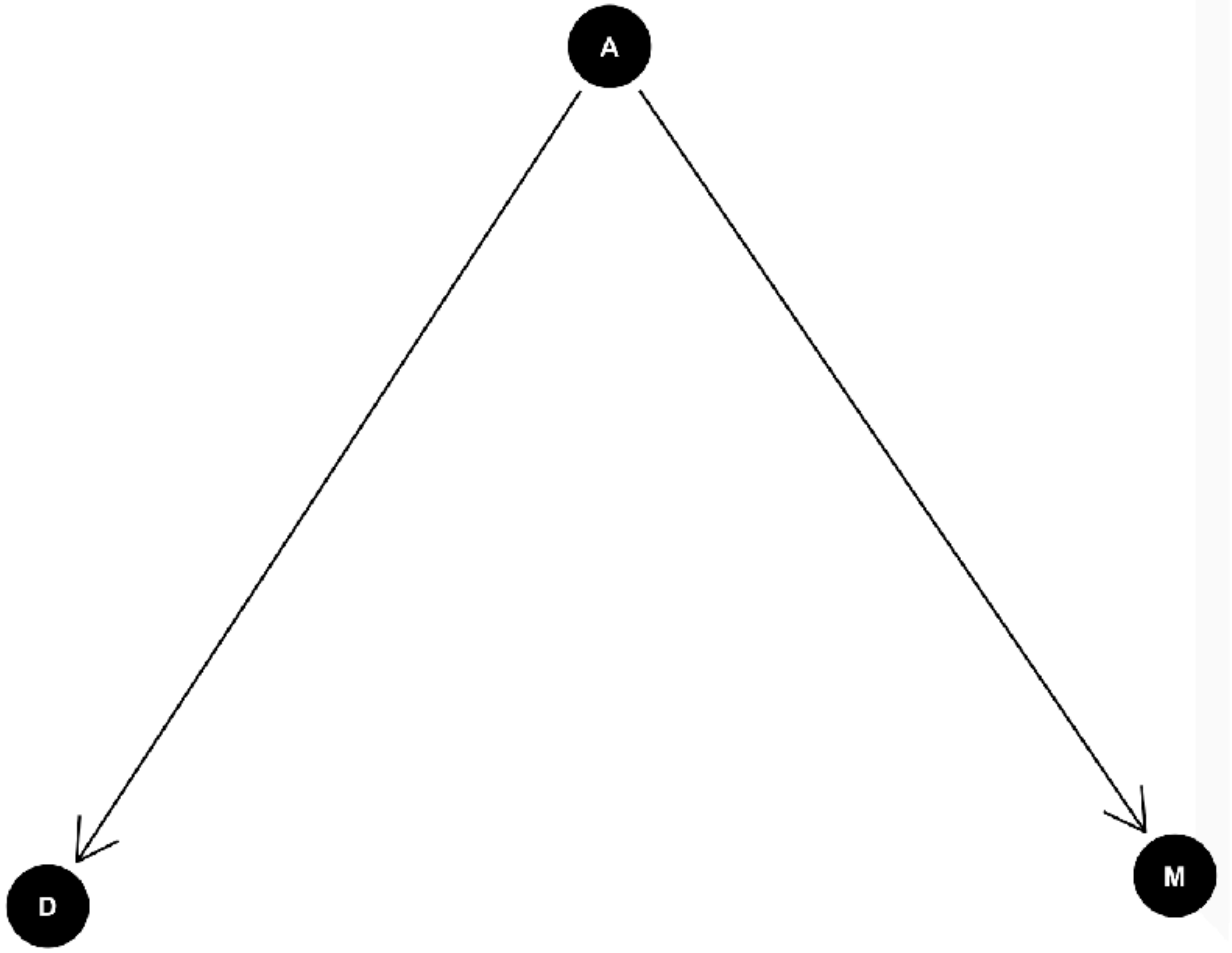
`lm(D~M)` will include effect of A!

This is called **confounding**

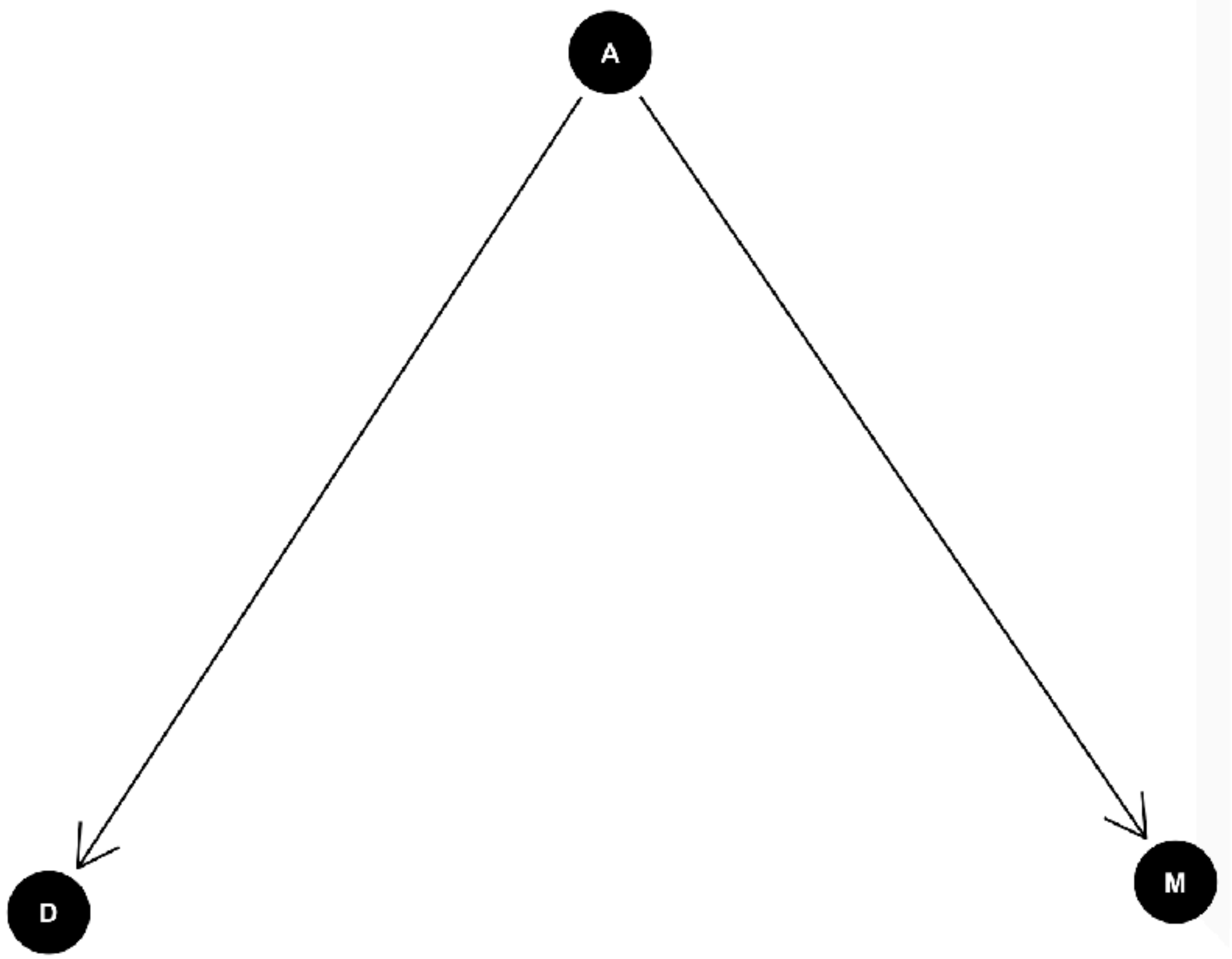
The problem

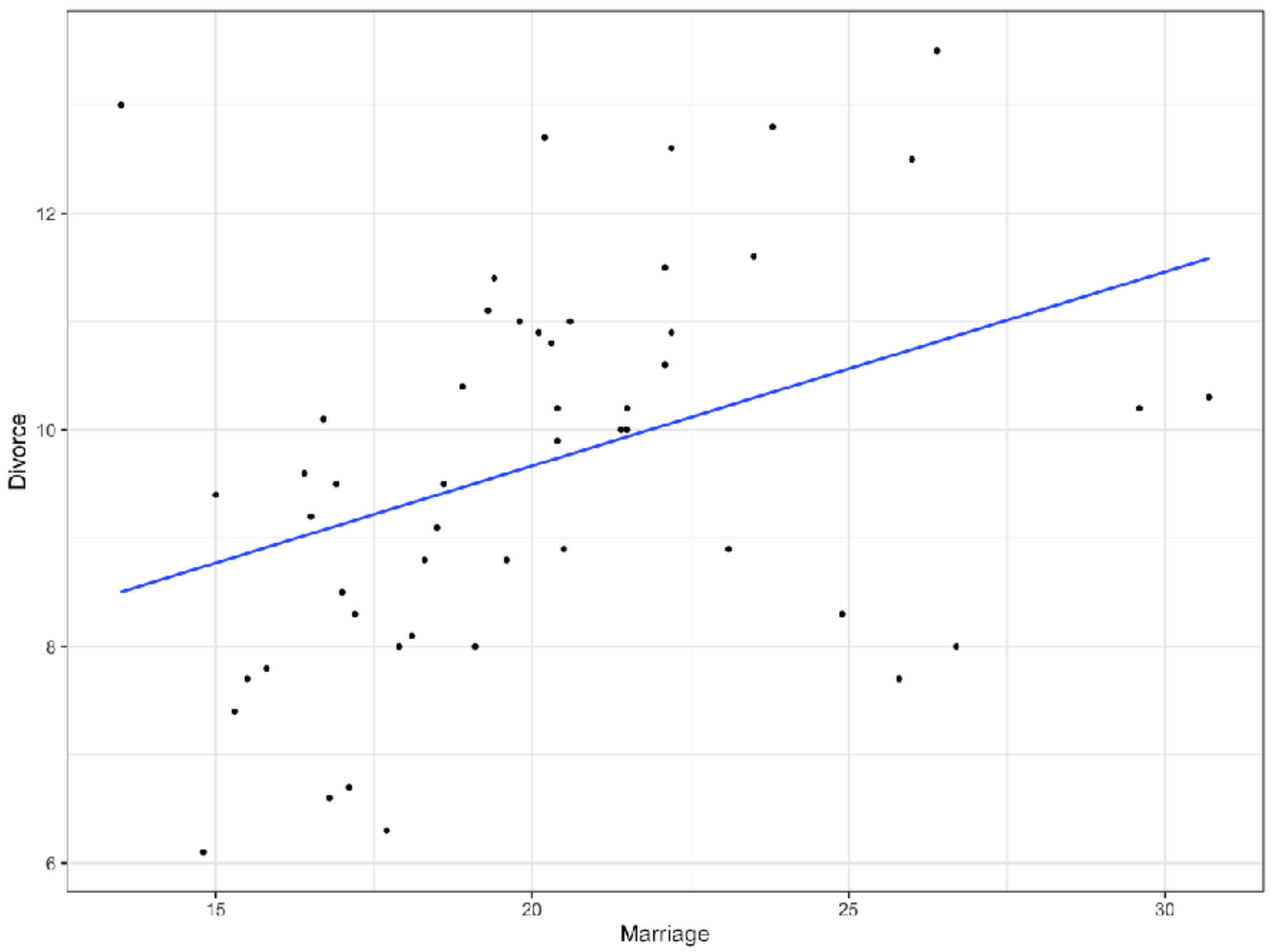
In fact, maybe there's no M —> D





The problem



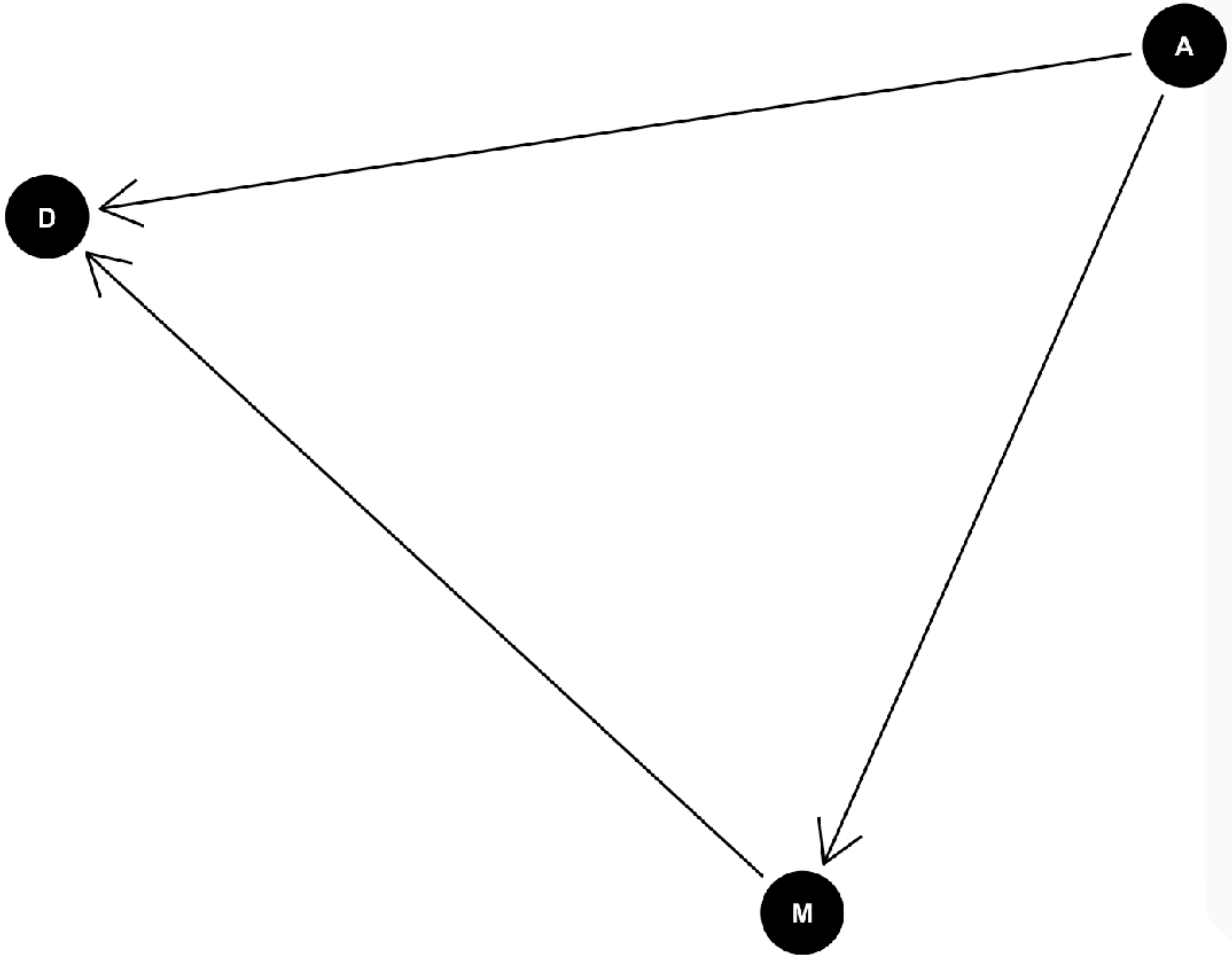


This DAG produces correlation between M and D!

The solution

We want to **identify,** or know, what part (if any) of M —> D is **not** a result of A

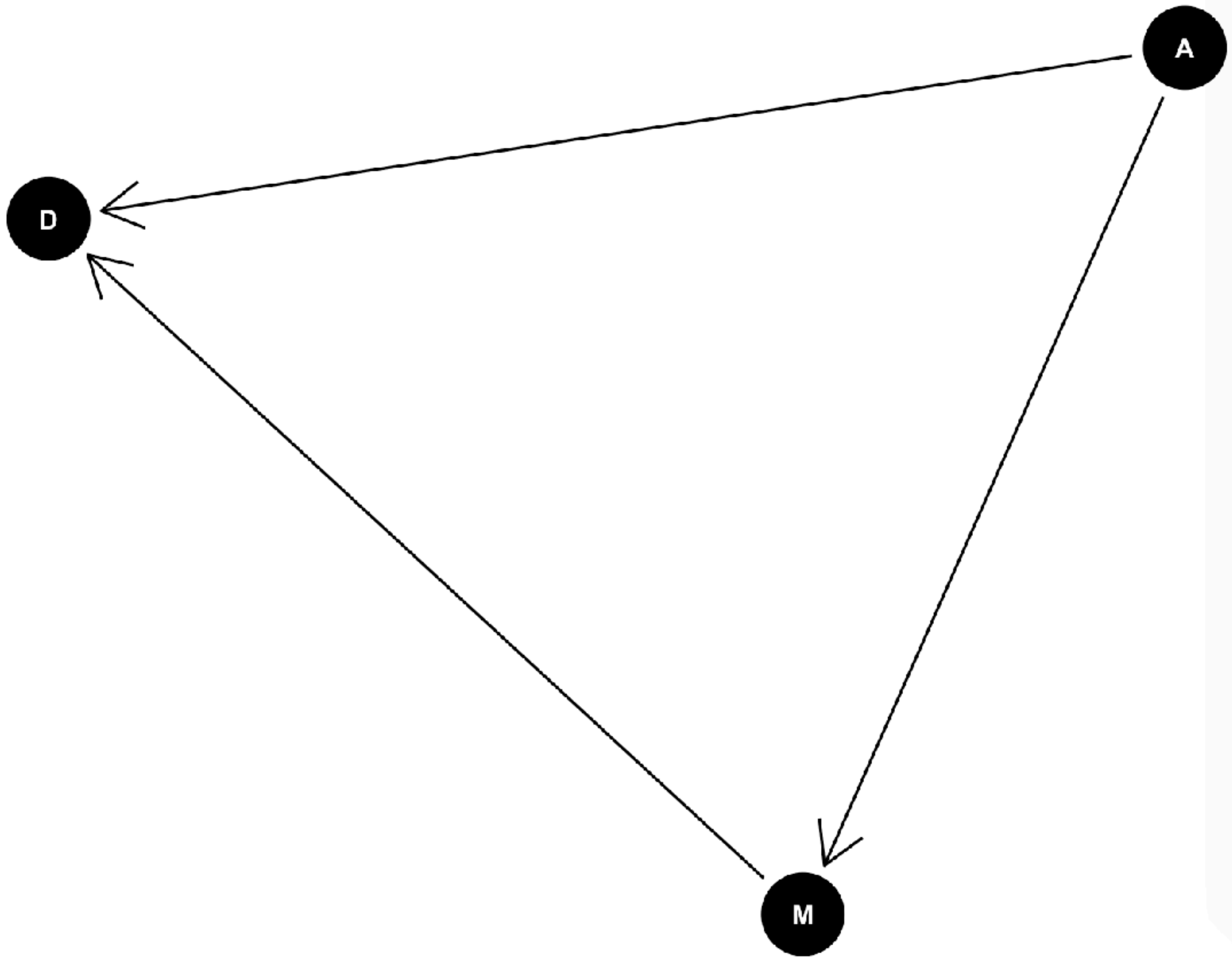
We want to **block** flow from A to D and M



Controlling

This is also known as **controlling** or **adjusting**

This means looking at the relationship between M and D ***controlling for A***



What is "controlling"?

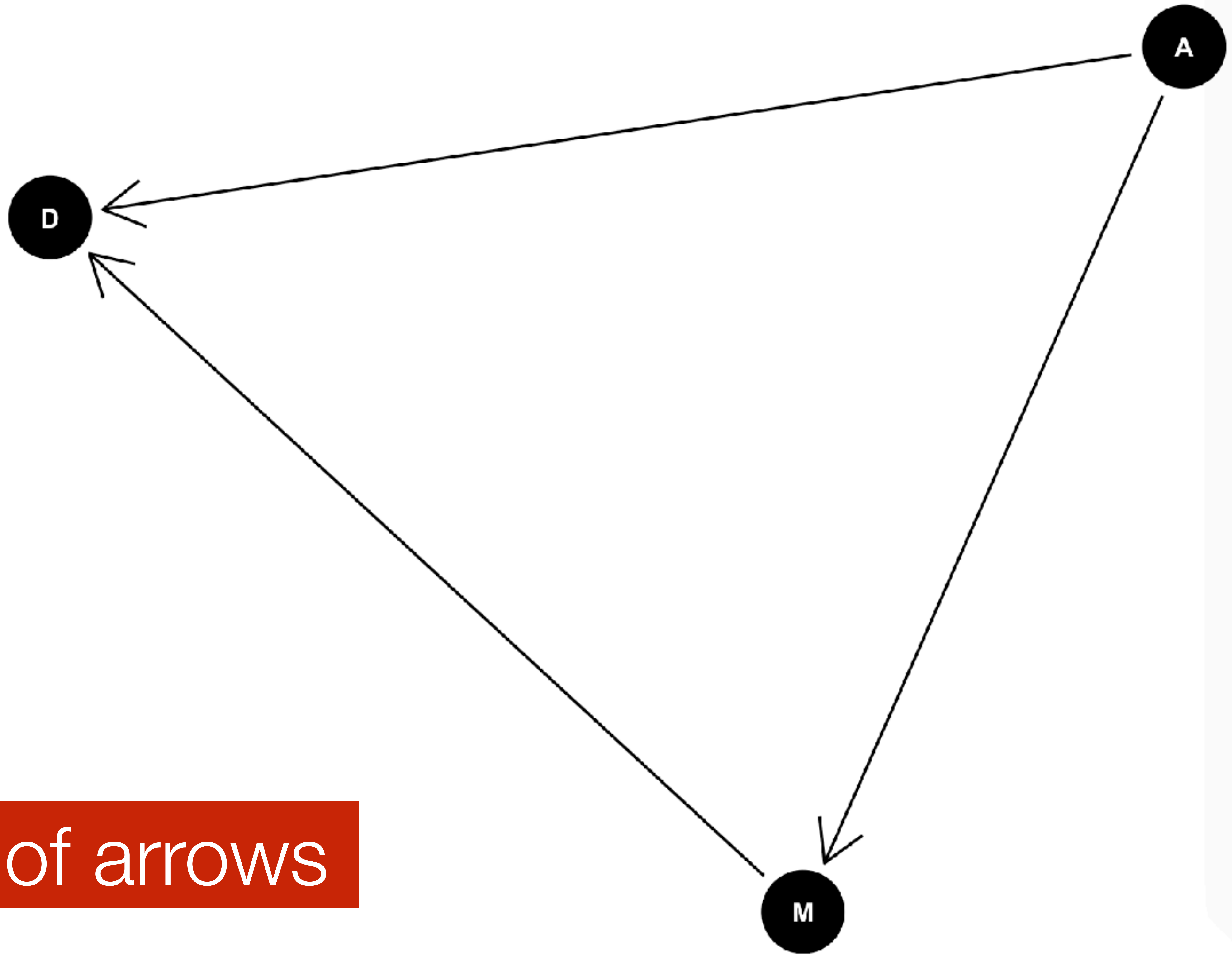
More on this next week, but basically:

Look at part of correlation between M and D *that isn't explained by A*

Look at relationship between M and D *in states with similar A levels*

Make an *apples-to-apples* comparison that mimics **experiments**

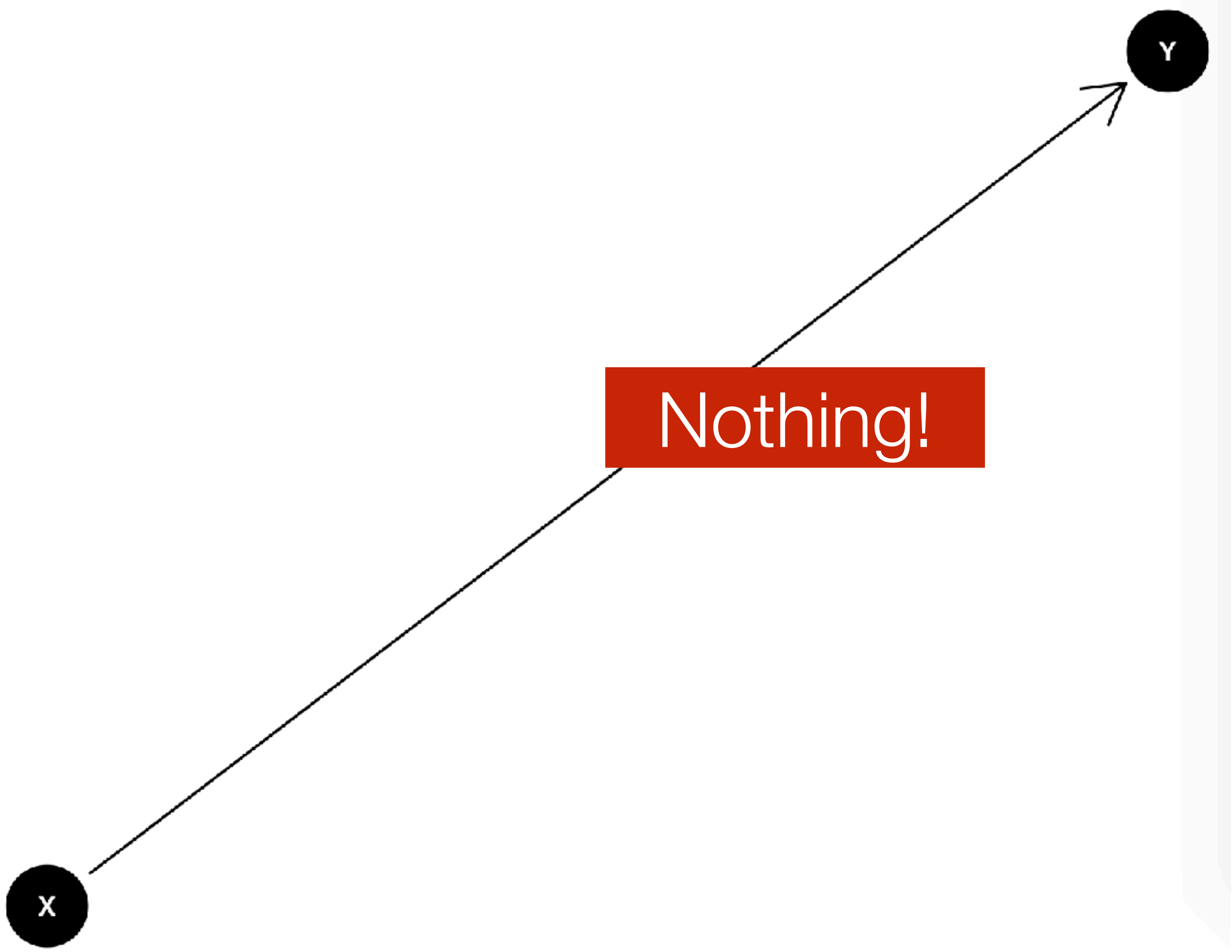
An analogy



Causality “flows” in direction of arrows

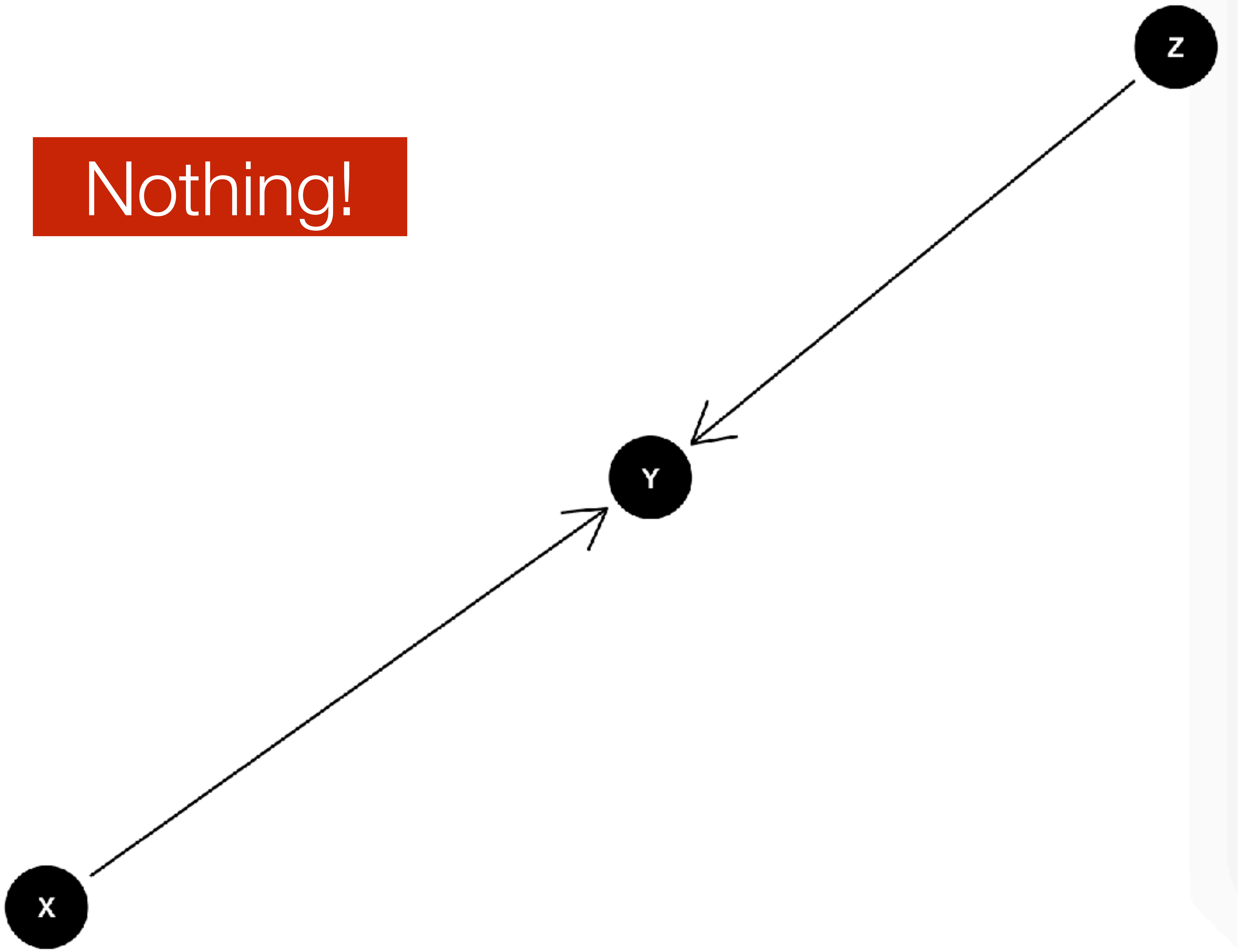
We want to shut off flow of A to learn about M —> D

What do we need to close?



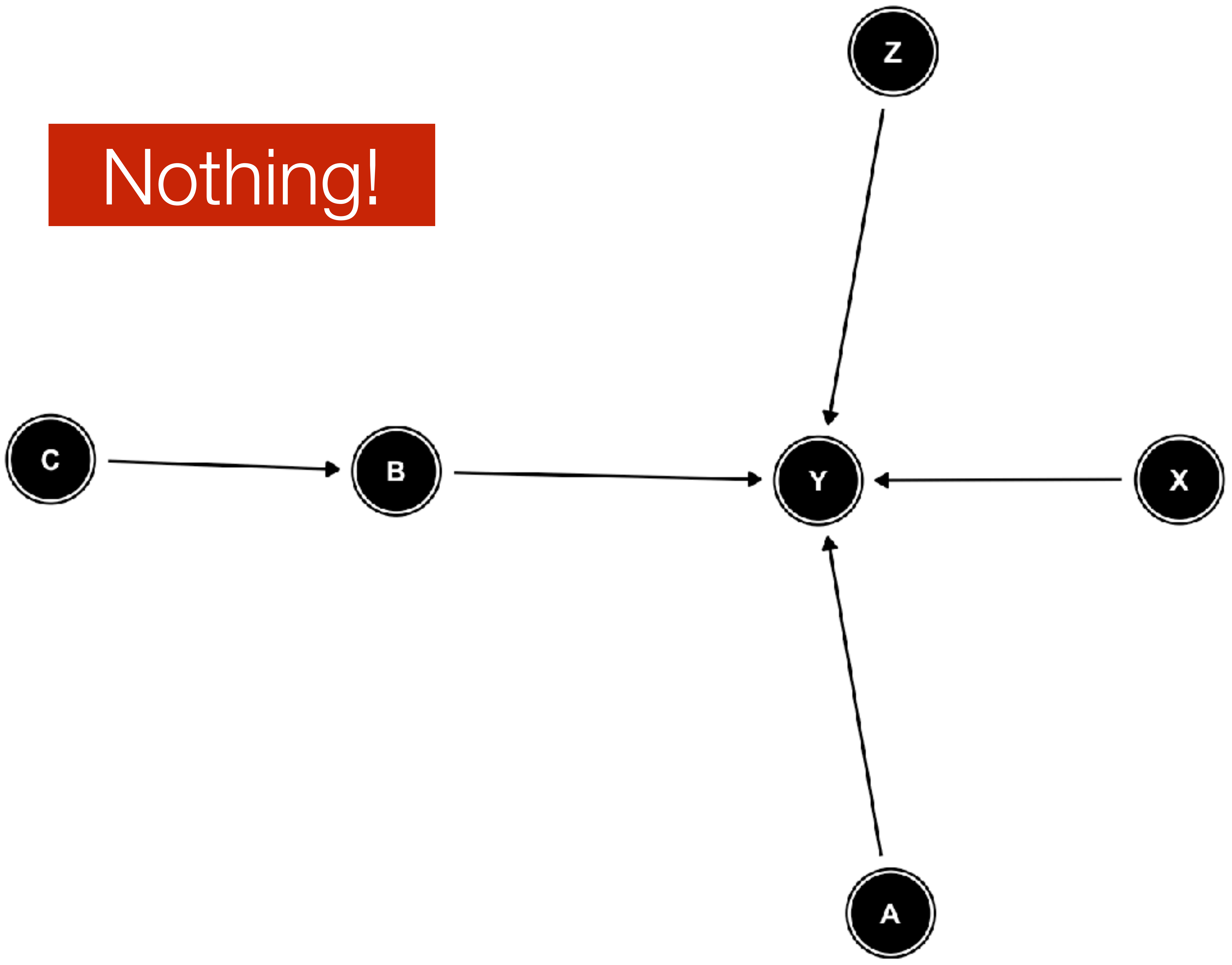
What do we need to close?

Nothing!



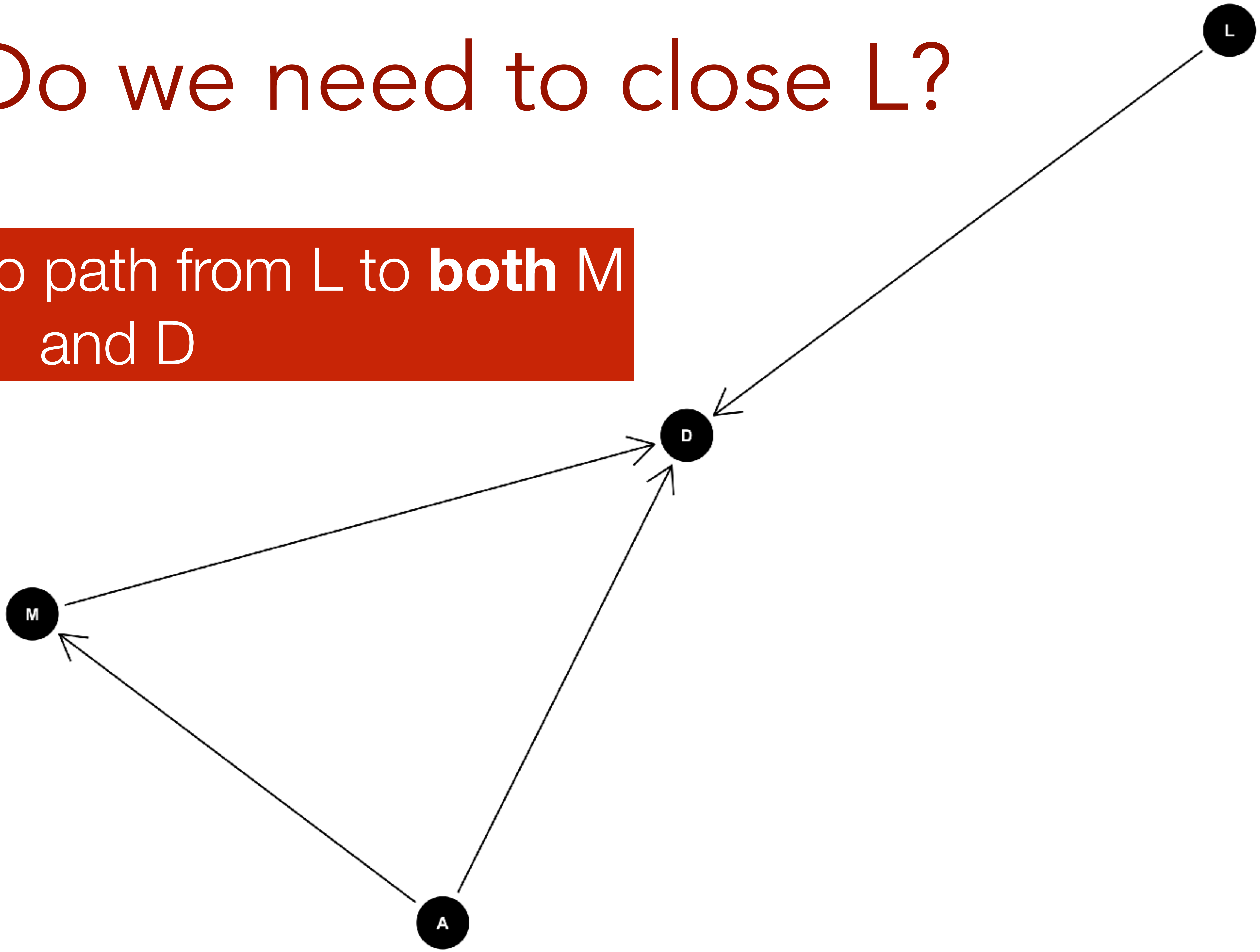
What do we need to close?

Nothing!

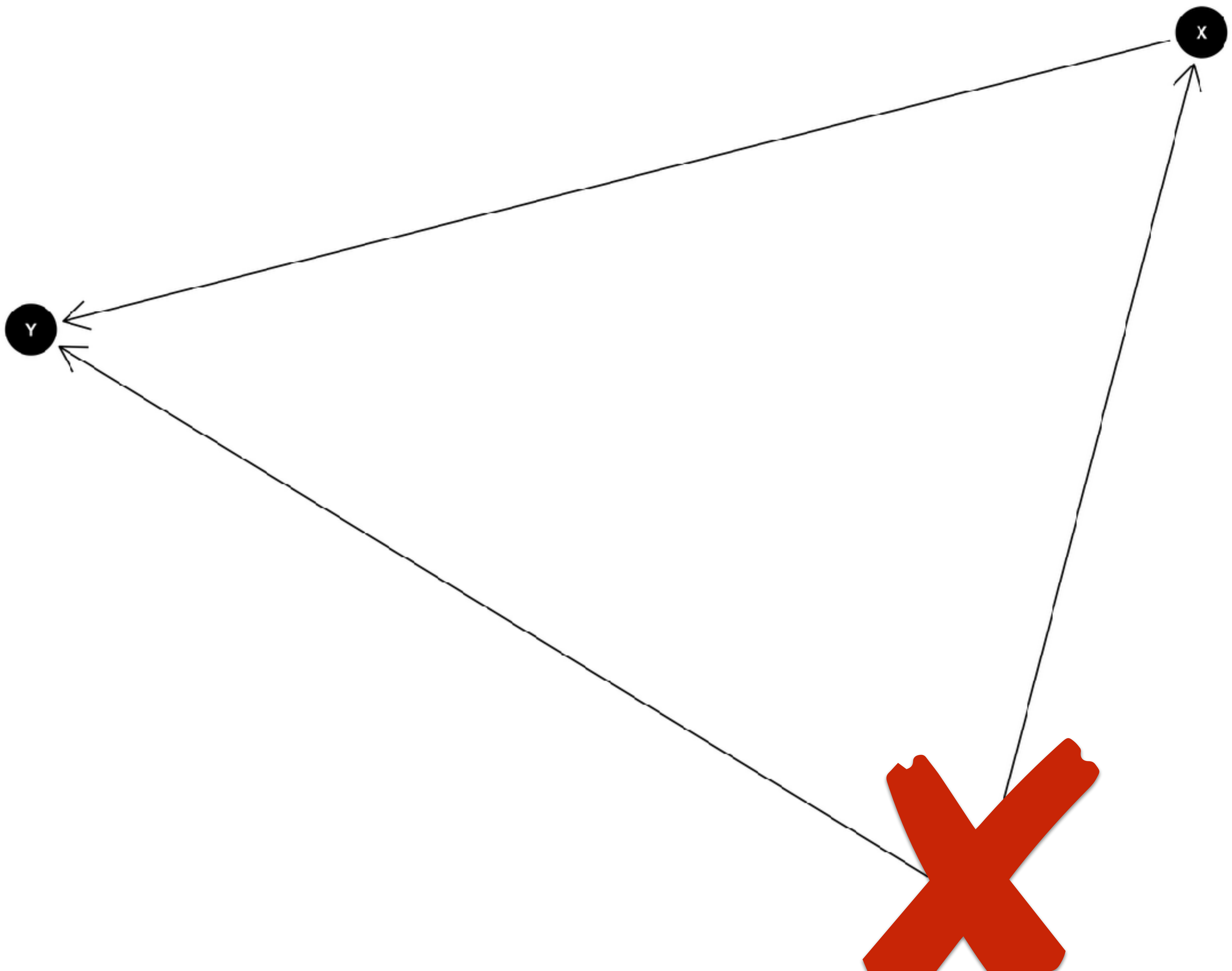


Do we need to close L?

No! Because no path from L to **both** M and D



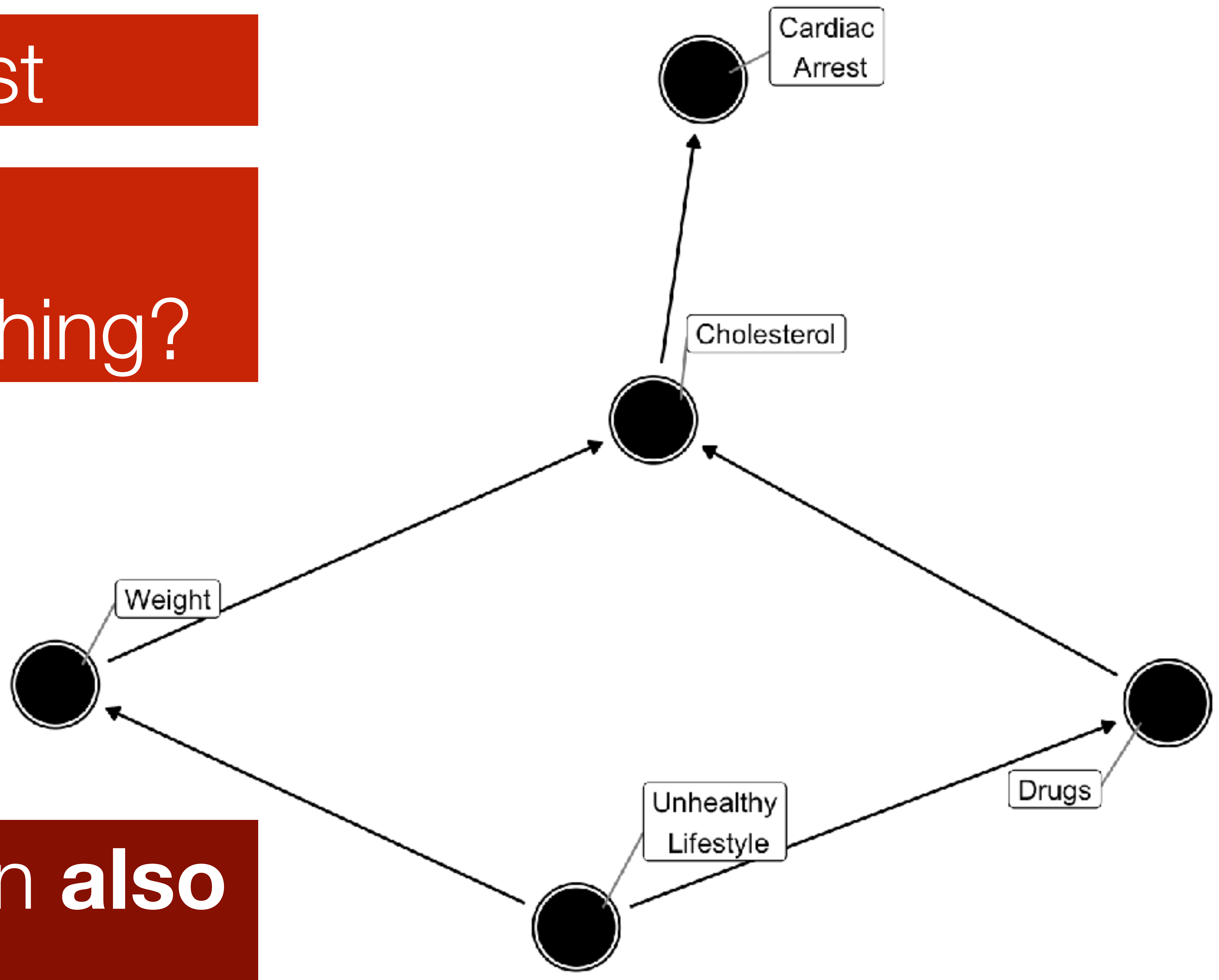
What do we need to close?



Why not close everything?

Drug use —> Cardiac arrest

This looks complicated;
Why don't we just block everything?



“Closing off” the wrong thing can **also** confound our results!

Ye Olde Causal Alchemy

The Four Elemental Confounds

The Fork

$$X \leftarrow Z \rightarrow Y$$

The Pipe

$$X \rightarrow Z \rightarrow Y$$

The Collider

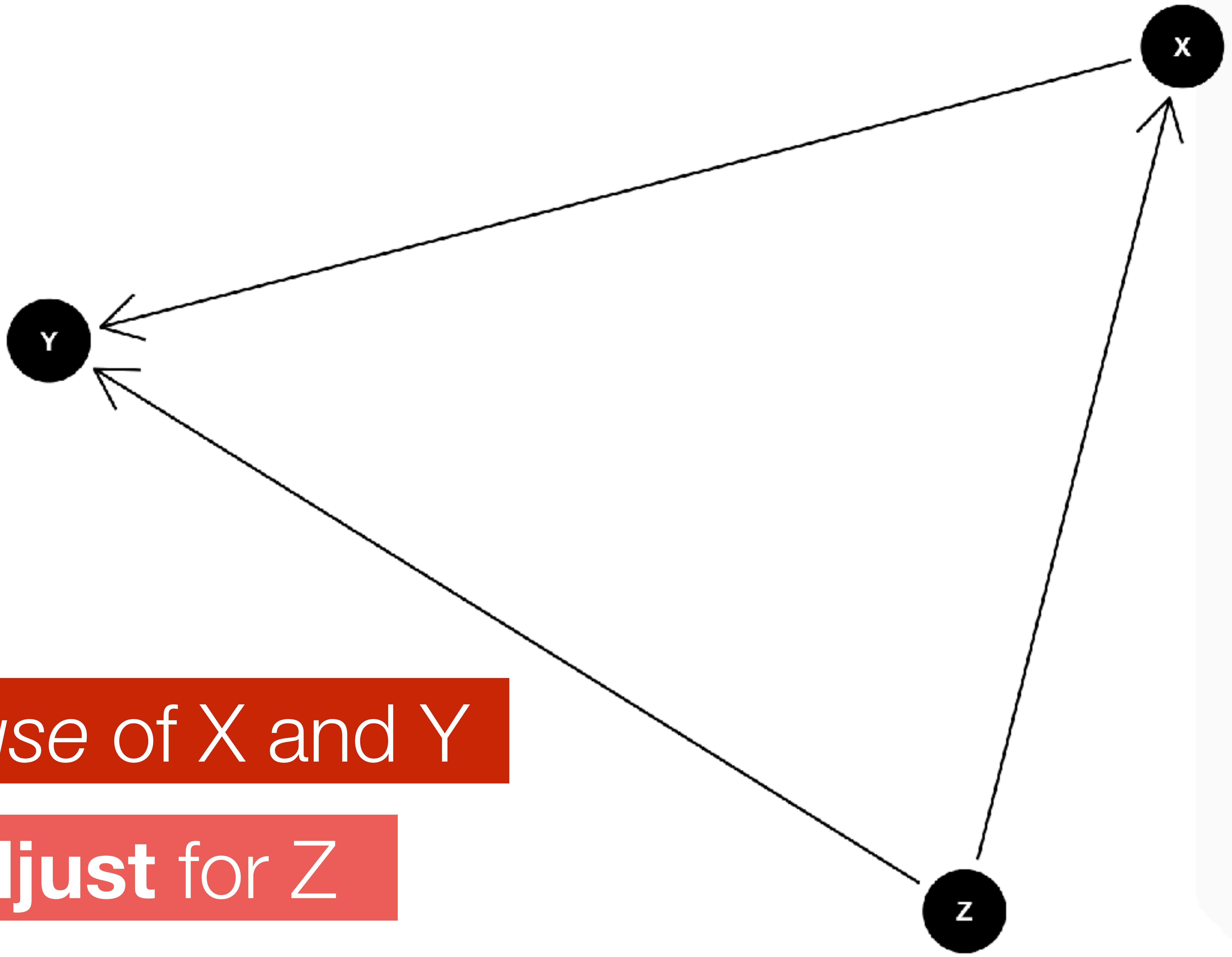
$$X \rightarrow Z \leftarrow Y$$

The Descendant

$$\begin{array}{ccccc} X & \rightarrow & Z & \rightarrow & Y \\ & & \downarrow & & \\ & & A & & \end{array}$$

The Fork

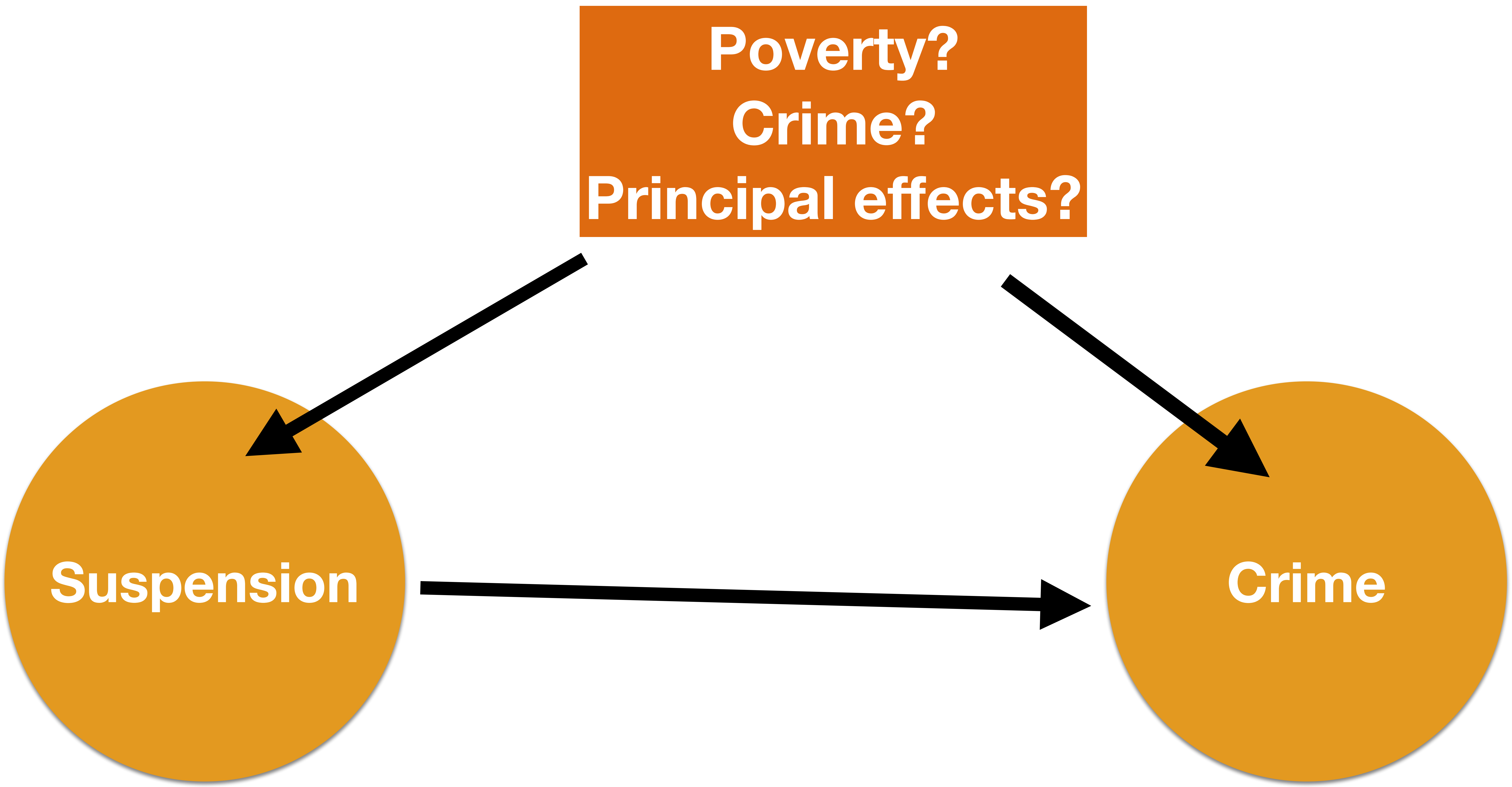
$X \leftarrow Z \rightarrow Y$



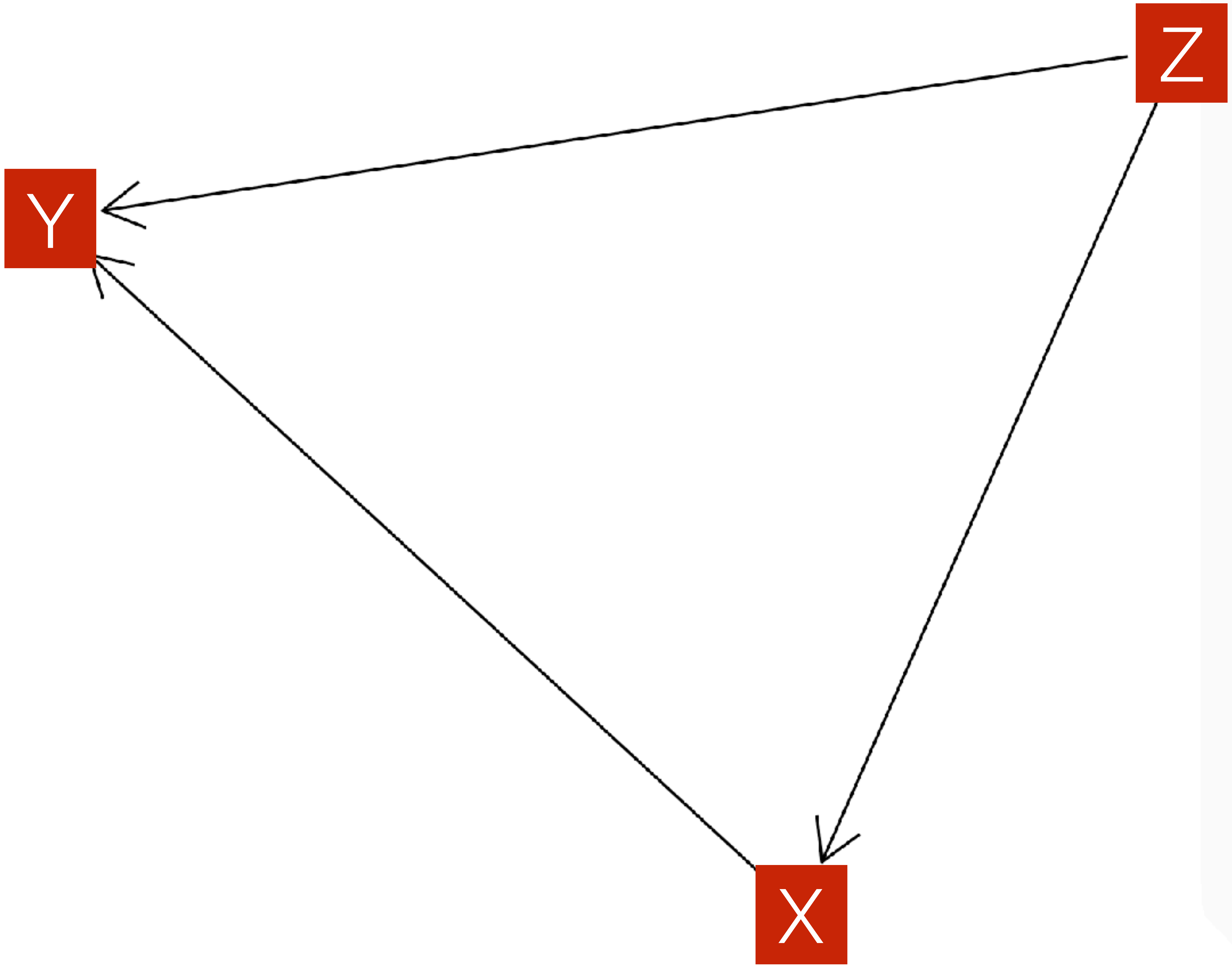
Z is a *common cause* of X and Y

Control or **adjust** for Z

Does School Suspension Work?

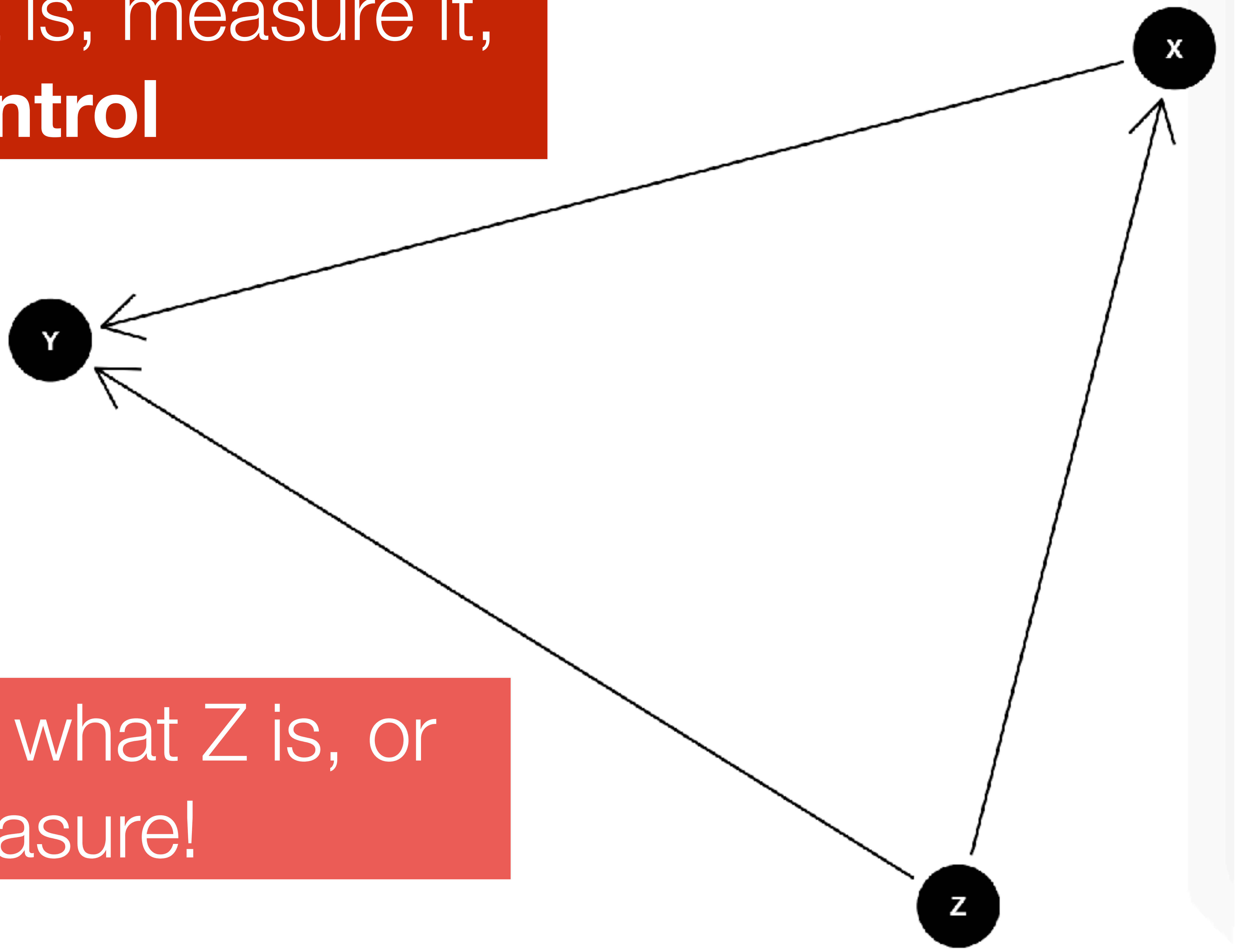


Other examples of forks?



Dealing with forks

Figure out what Z is, measure it, and **control**



Not easy to know what Z is, or how to measure!

The (other) pipe

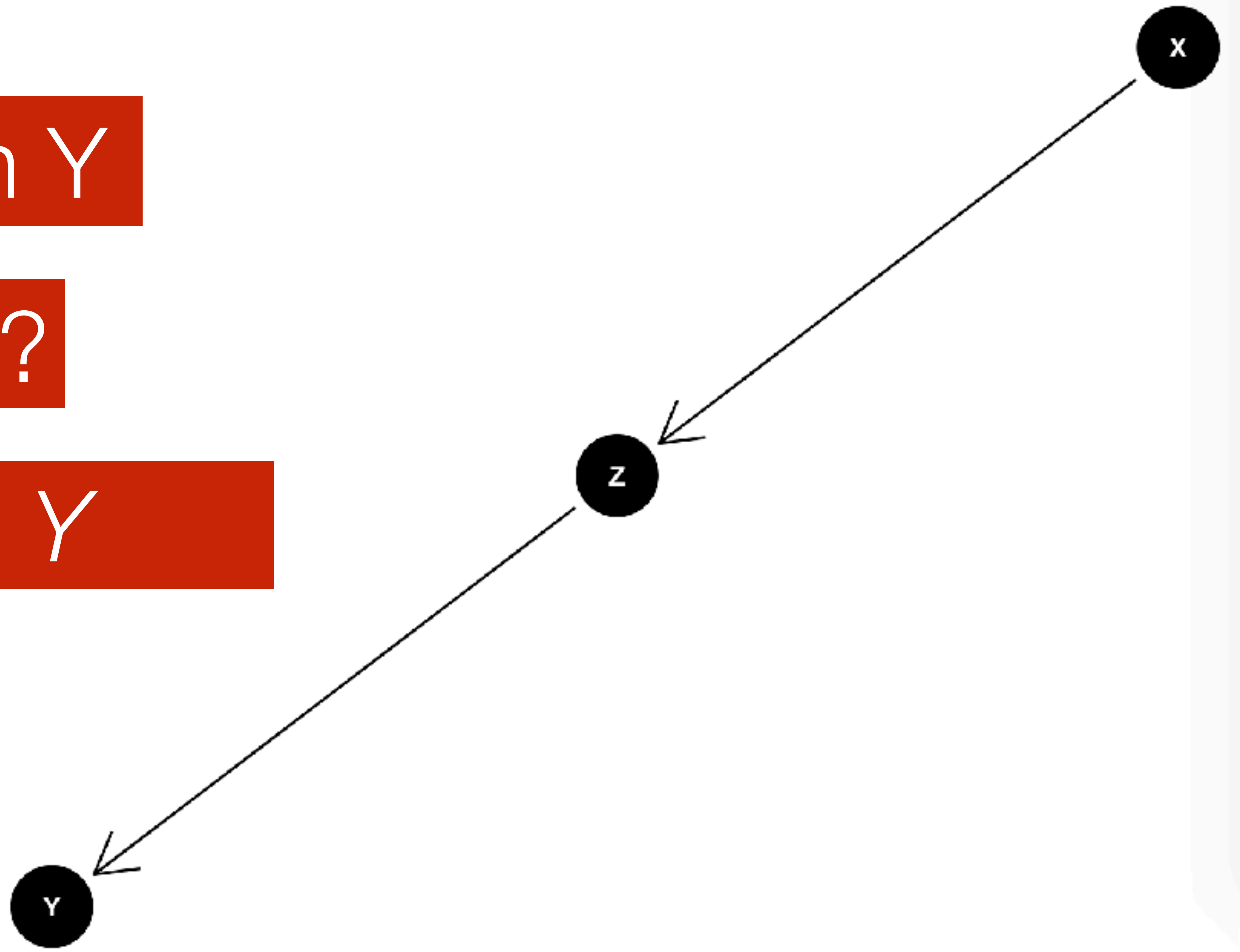
$X \longrightarrow Z \longrightarrow Y$ X causes Z causes Y

Z **mediates** the effect of X on Y

What happens if we block Z?

We remove the effect of X on Y

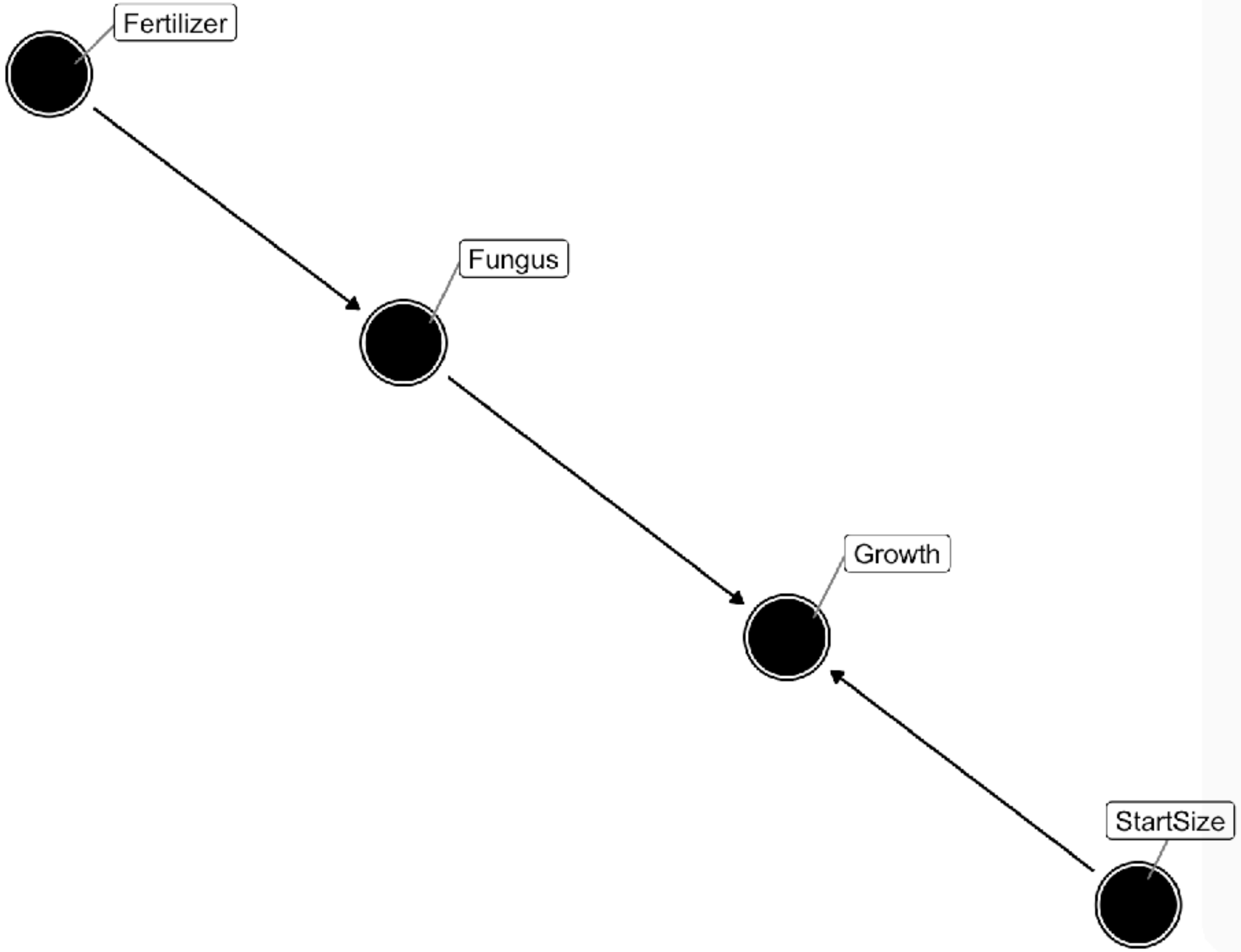
Leave Z alone



Making plants grow

What's the effect of our fertilizer on plant growth?

Conditioning on Fungus cuts off effect of Fertilizer

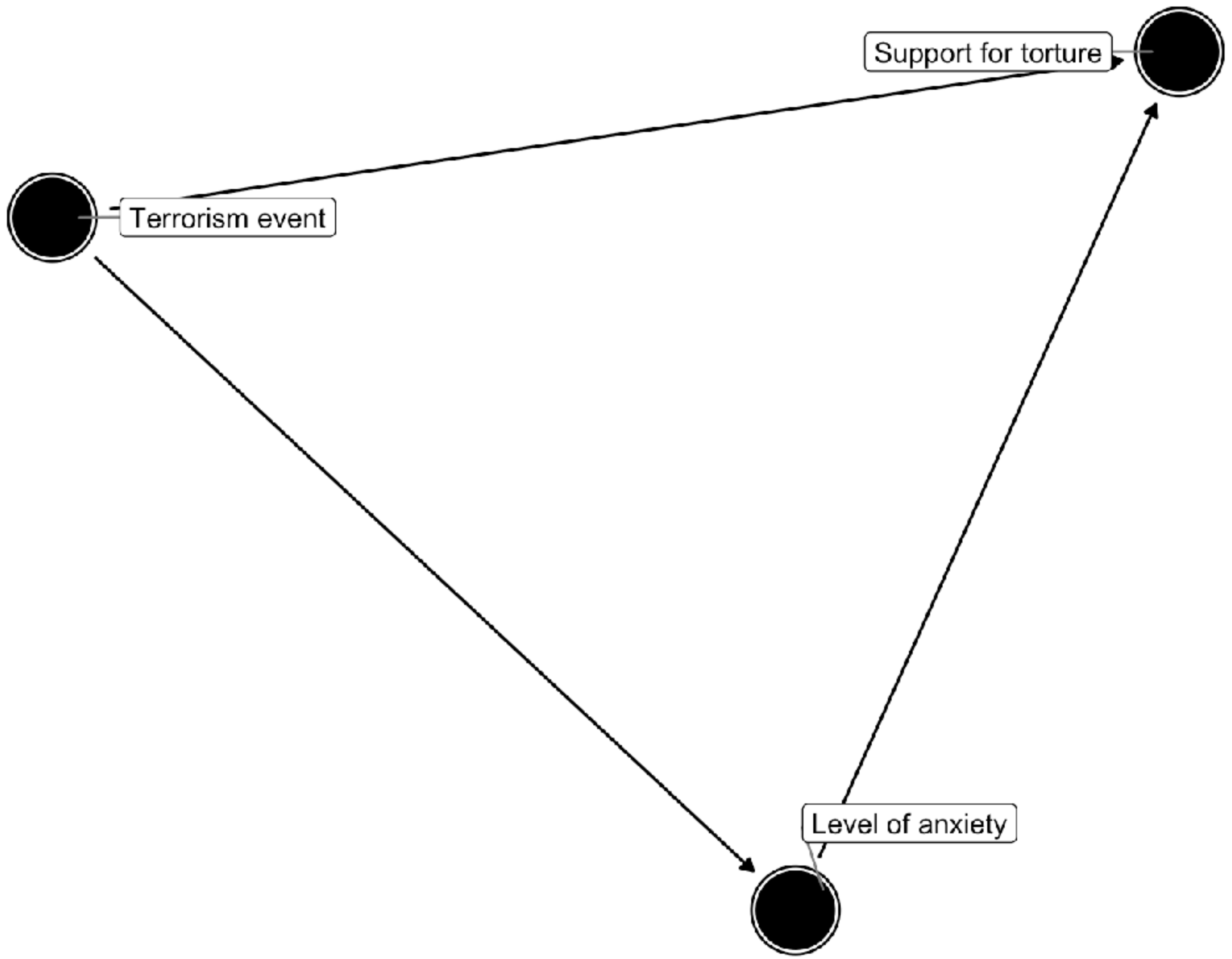


Terror example

What's the effect of being exposed to terrorism on support for harsh anti-terror policies?

Two causal flows:
T —> A —> S
T —> S

Conditioning on Anxiety controls away the *indirect effect*

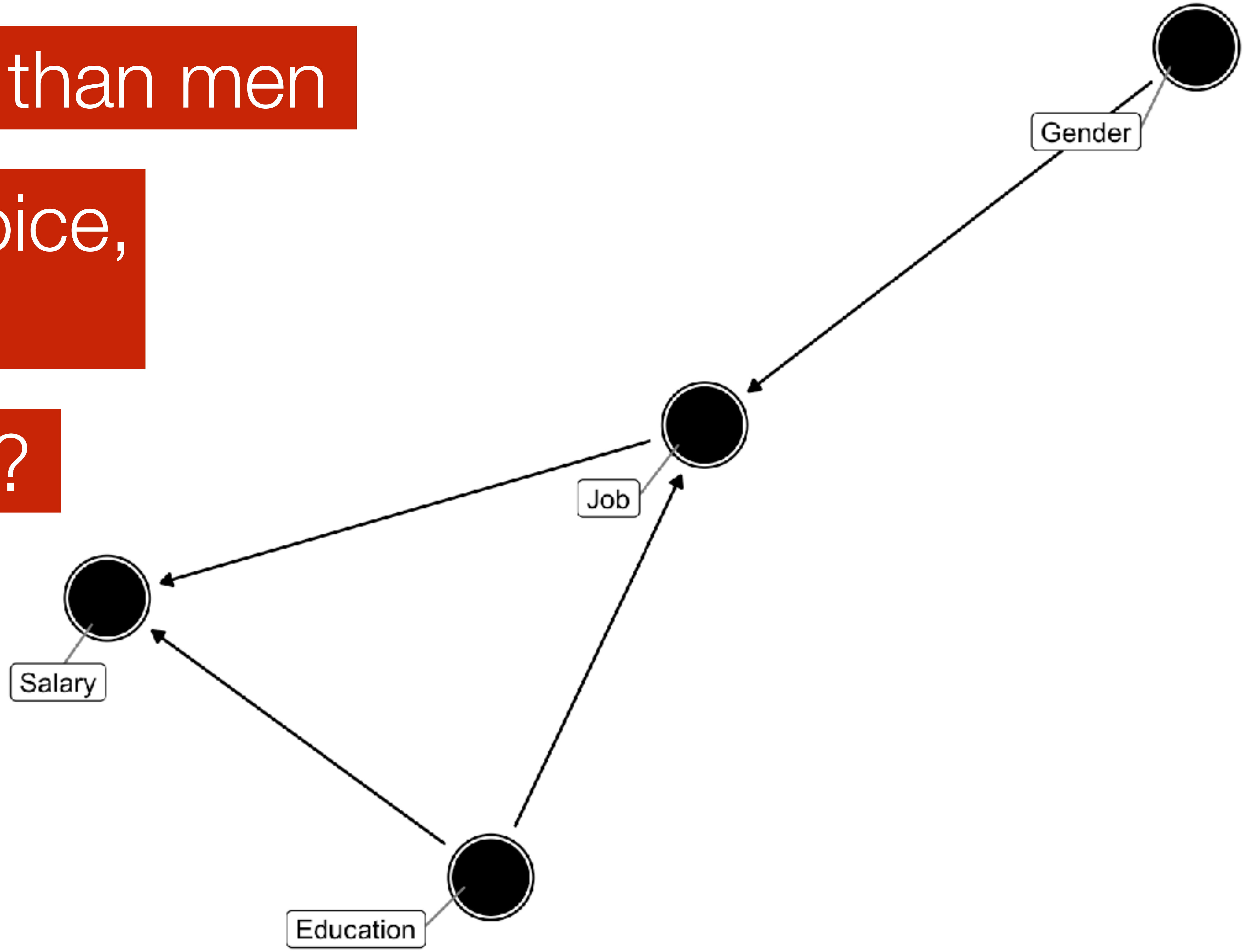


The Wage Gap

Women make X amount less than men

But *controlling* for career choice, wage gap —> 0

What if DAG looks like this?



Dealing with pipes

Avoid controlling for Z!

Why is this even a problem?

Bad statistics: we are so afraid of *forks* we simply control for everything

Sometimes we are **unaware** we are doing it

"But wait, did you control for…"

Need DAGs!

Harvard discrimination case

Professor Arcidiacono omitted intended career, staff interview indicator, and parental occupation from his model. [Oct. 25 Tr. 145:21–148:12]. The Court prefers a model that includes these variables because they play a role in the admissions process. [Oct. 26 Tr. 8:25–9:21, 10:17–11:6; Oct. 31 Tr. 9:3–7]. Further, these variables correlate with race and therefore create a significant potential for omitted variable bias if excluded. [Oct. 31 Tr. 10:16–18, 11:15–12:21, 21:19–22:14; DX677; DX681; DD10 at 54]. Professor Arcidiacono excluded these variables primarily because of data issues, including unexplained year-to-year fluctuations in the distribution of parental occupation and intended career categorizations. [Oct. 25 Tr. 145:21–148:12; DD10 at 50–52, 56]. As examples, numerous parents who were categorized as having low-skill jobs for the class of 2014 would likely have been categorized as being self-employed for the class of 2015, and there is a substantial decrease in the number of parents categorized as unemployed among applicants to the class of 2017 versus the class of 2018. [Oct. 25 Tr. 146:4–147:9; DD10 at 51–52]. Although the data for parental education and intended career are not as consistent year-to-year as would be ideal, including the variables is preferable because their exclusion results in omitted variable bias that exaggerates the effect of race that is implied by the models. [Oct. 30 Tr. 146:18–147:6; DX695; DD10 at 35]. Professor Card's model deals

The Descendant

Boring version of a pipe

Even if you control for Z, if you control for A and A is caused by Z you are in trouble

Need to be careful

Recap

DAGs are causal models of a process we are interested in

DAGs tell us what to **control** (and what to avoid!) in order to get at X —> Y

Fork = control for Z
Pipe = (mostly) don't control for Z

Thursday = colliders