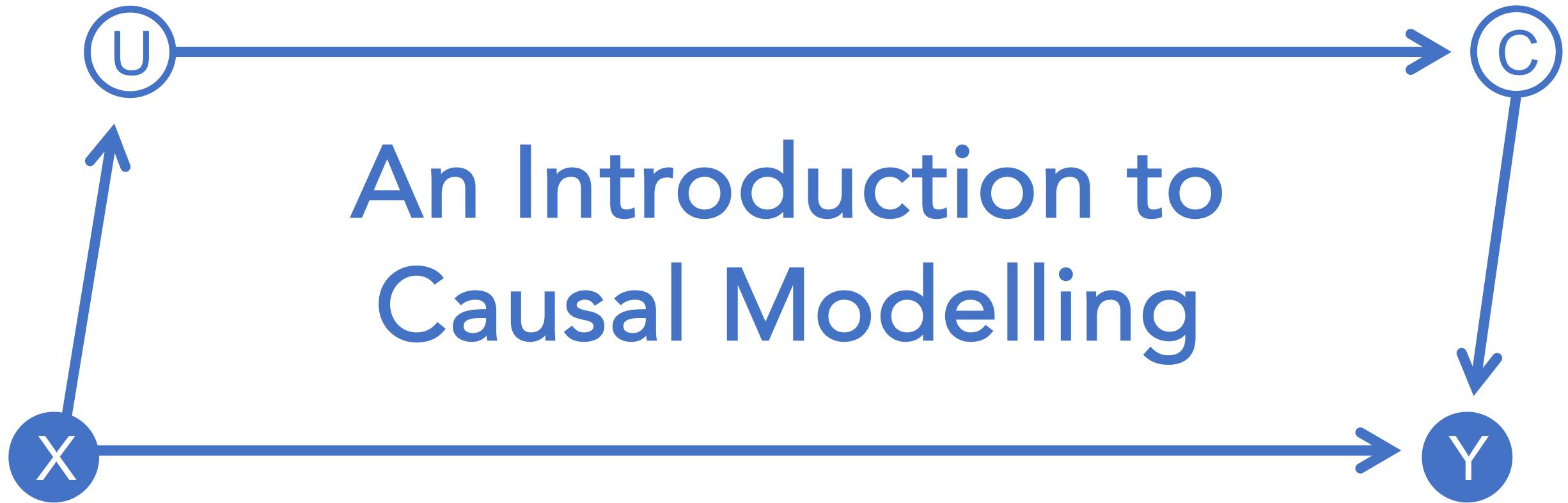
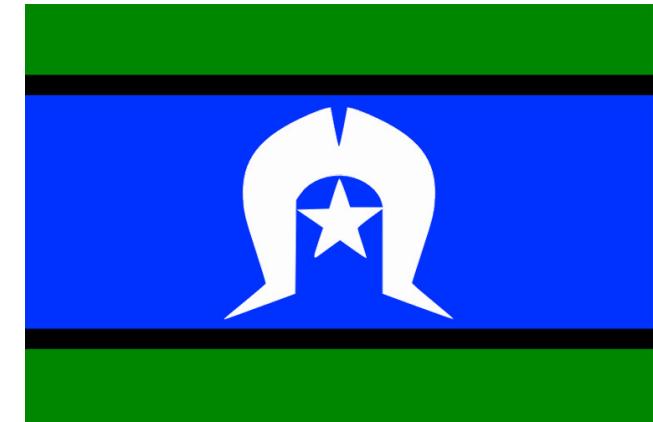
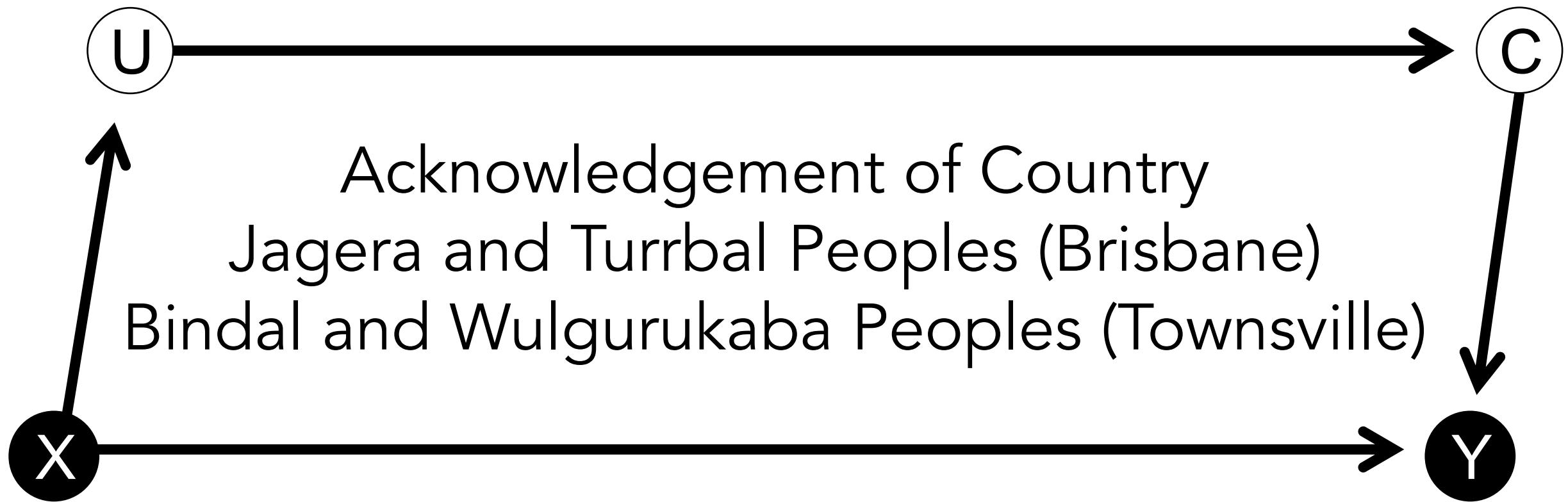


An Introduction to Causal Modelling



Kevin Bairois-Novak
UQ Coding Club
28 Jan 2025



Wake up, Coding Club...

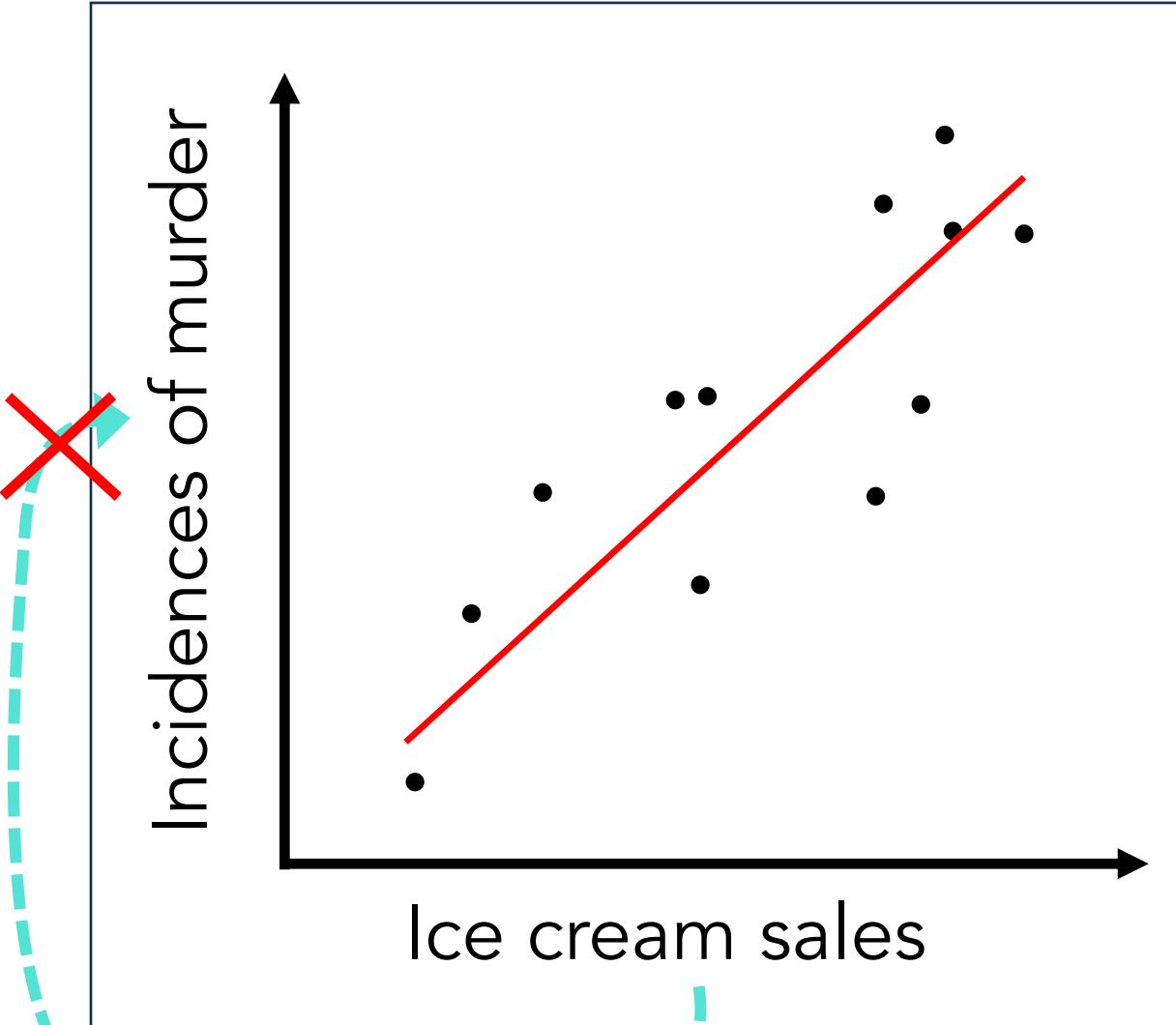
The Scientific Method_{TM} has you...

Follow your best intuition.

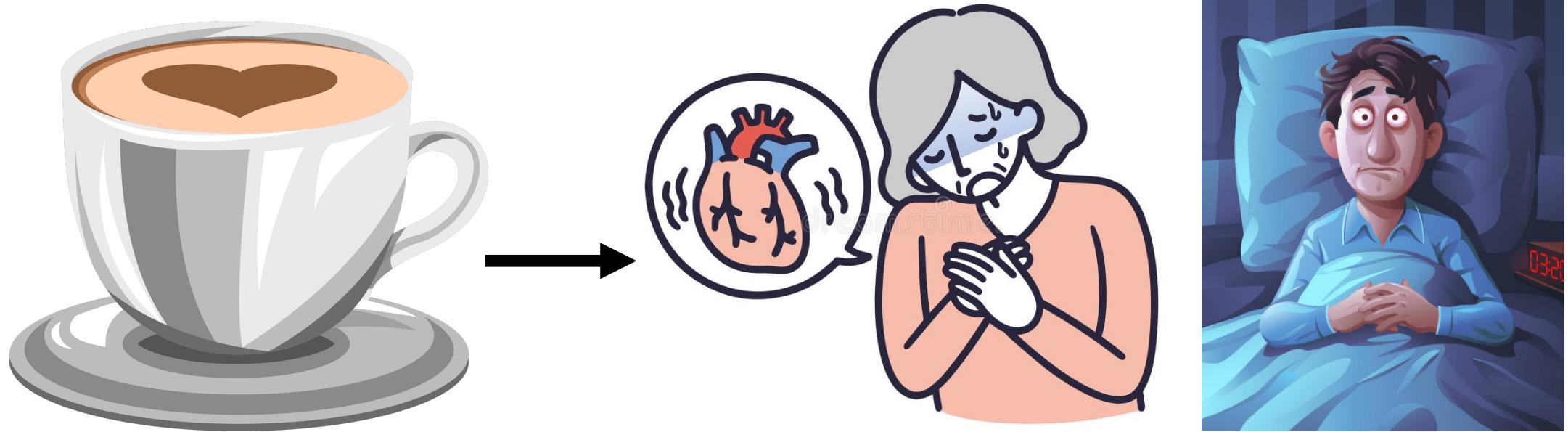
Knock, knock...

Chapter 1: The lie we live

“Correlation does not equal causation”



“Correlation does not equal causation”
...except when it does?



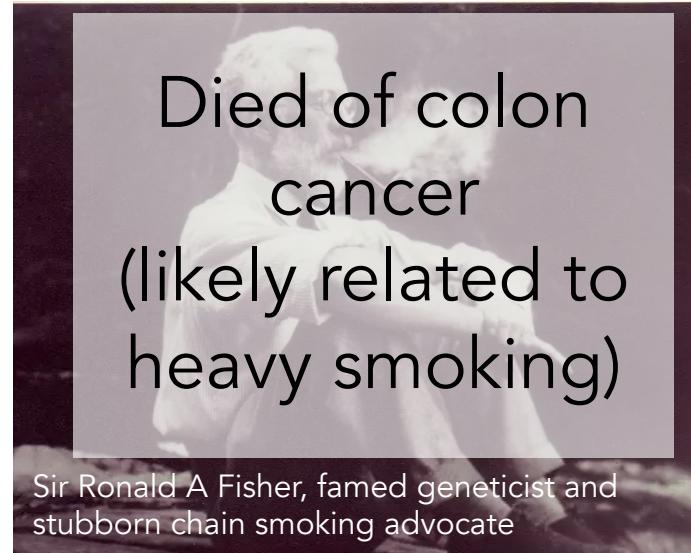
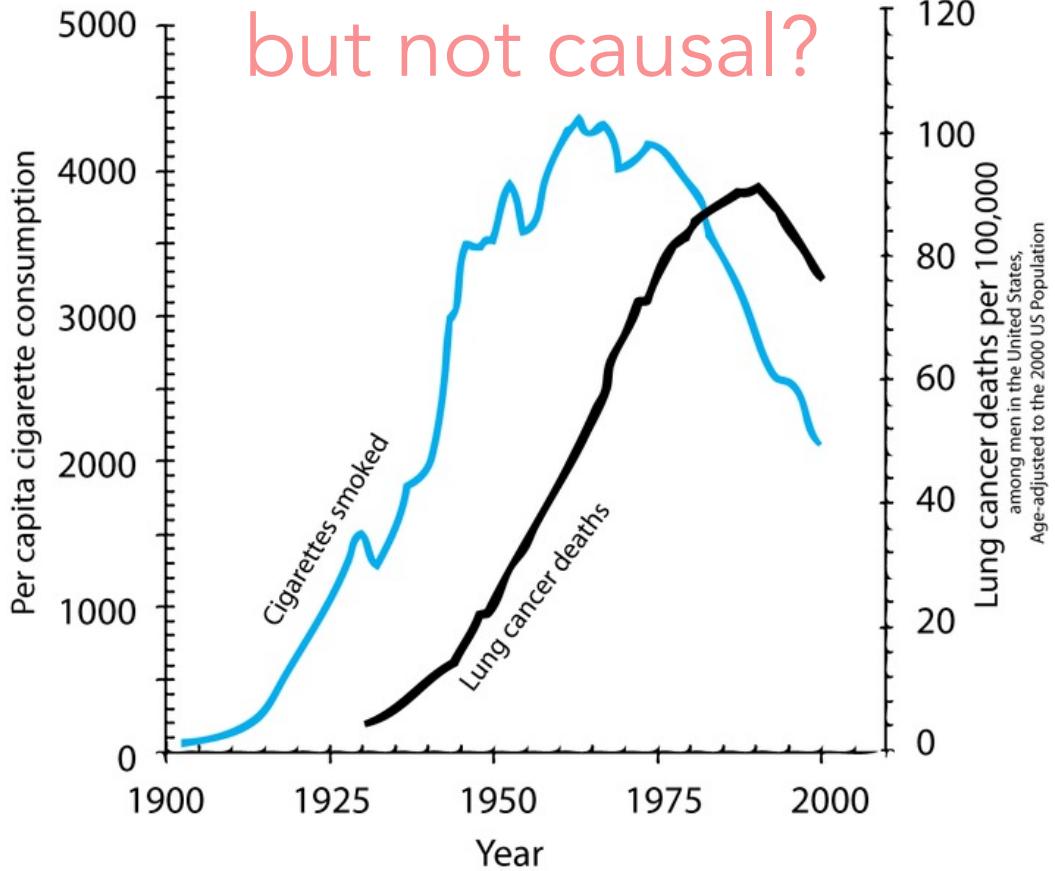
Test by manipulating number of coffees consumed!
Time to fall asleep (Y) ~ # coffees in arvo (X)

But what if we can't (ethically) manipulate?



Does smoking cause cancer?

Strong correlation, but not causal?



Sir Ronald A Fisher, famed geneticist and stubborn chain smoking advocate



~~Test by forcing people to smoke/not smoke!~~

~~Cancer incidence (Y) ~
Cigarette consumption (X)~~

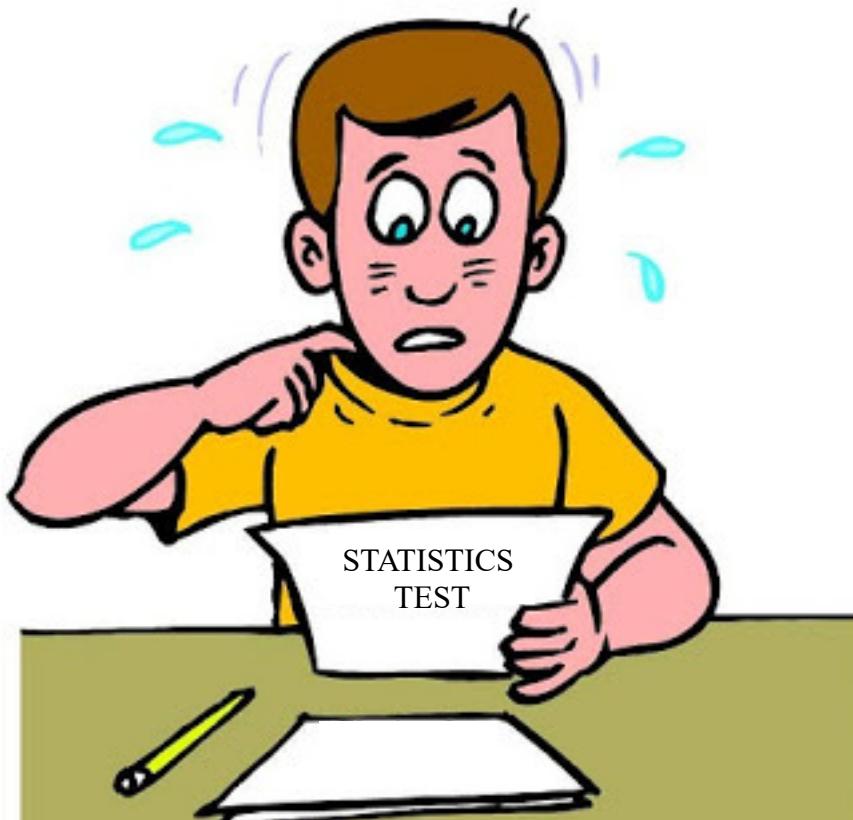




Chapter 2: We are living in a *lie*, Neo...

Example 1: Does sleep level affect student test results?

	Sleep	Alert	Test_score
	<dbl>	<dbl>	<dbl>
1	-0.626	0.509	-0.378
2	0.184	1.30	-0.627



Test score (Y) ~ Sleep (X₁) + Alertness (X₂)

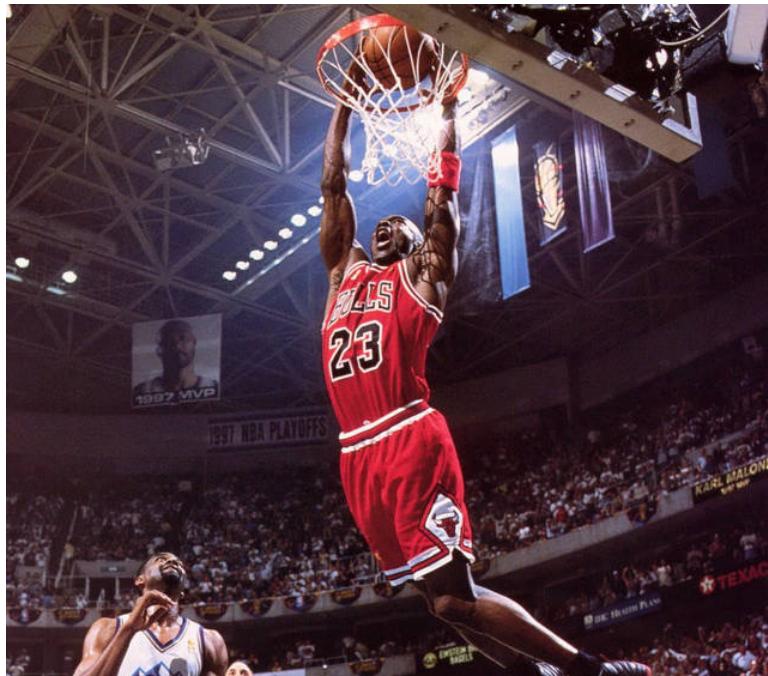
```
mod1 <- lm(test_result ~ sleep + alertness,  
            data = sleep_dataset)  
summary(mod1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.016811	0.032880	0.511	0.609
sleep	0.004877	0.045431	0.107	0.915
alertness	1.010407	0.031905	31.670	<2e-16 ***

No, but alertness does!

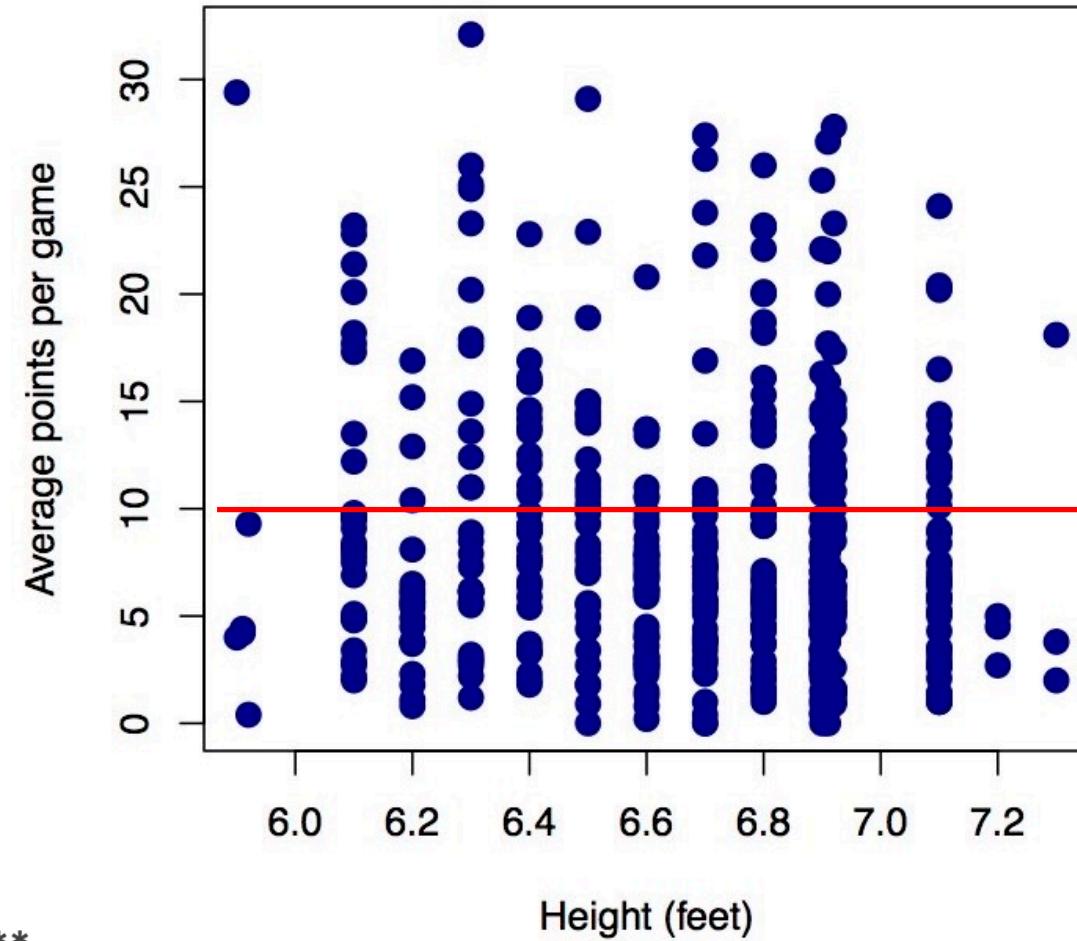
Example 2: Does basketball player height affect scoring?



Call:
`lm(formula = Scoring ~ Height, data = .)`

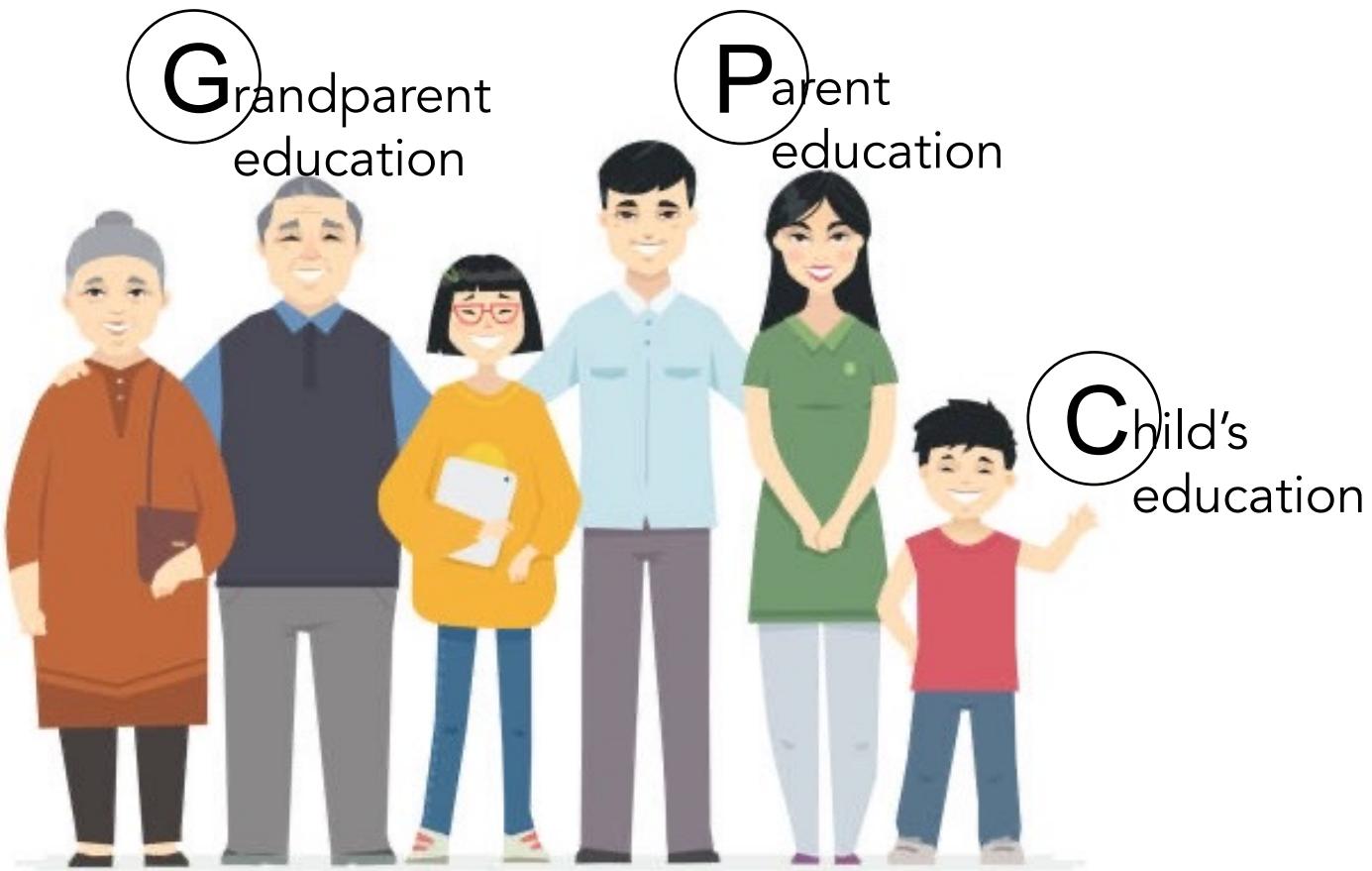
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.81281	0.23751	11.368	< 2e-16 ***
Height	0.05523	0.11003	0.502	0.617

NBA player scoring (Y) ~ NBA player height (X)



No relationship between height and scoring ability!

Example 3: Does your (grand)parent's education level affect your own?



Four options for models:

Model 1: $C \sim P$

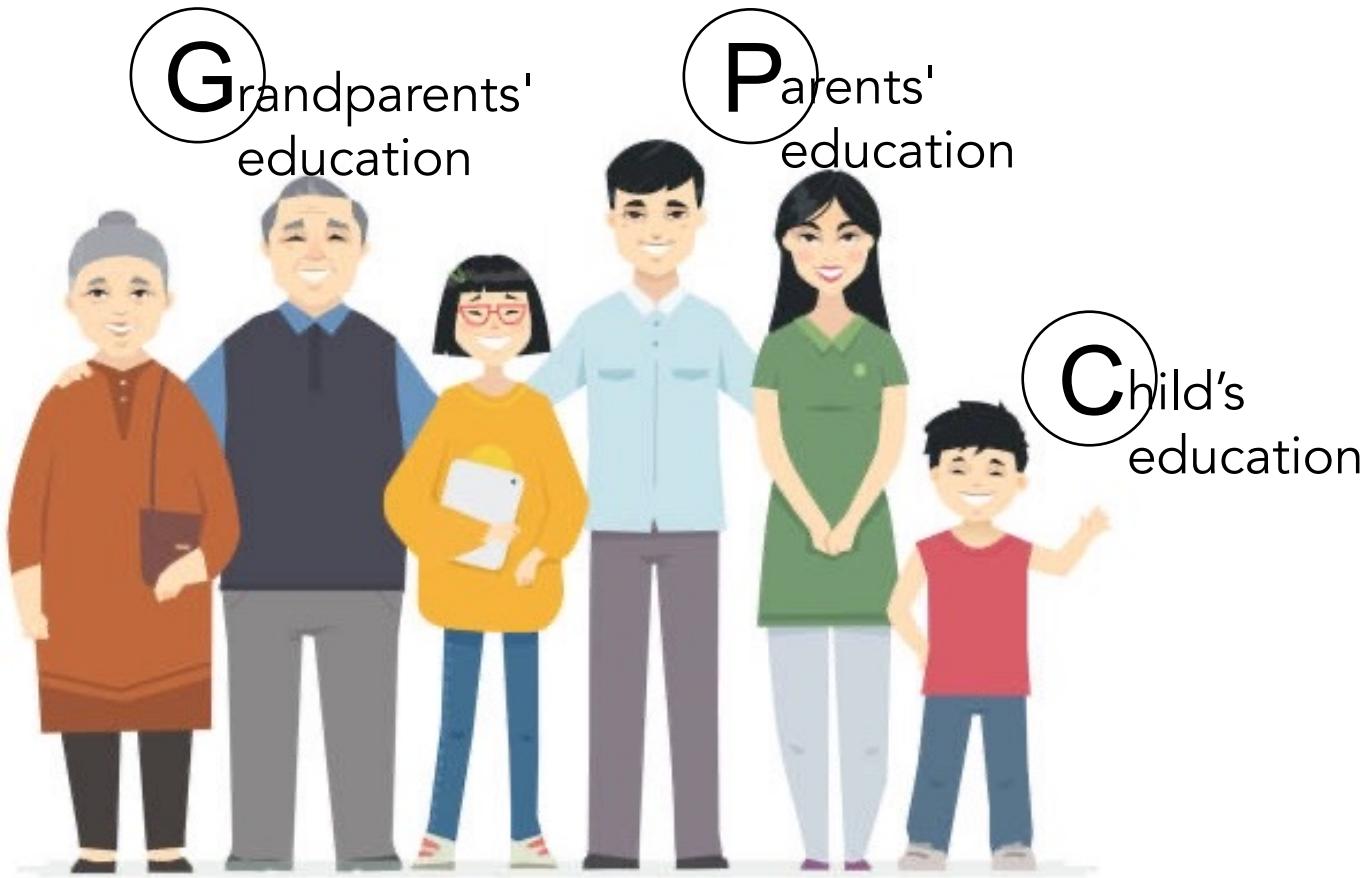
Model 2: $C \sim G$

Model 3: $C \sim P + G$

Model 4: $C \sim P \times G$

	G	P	C
	<i><dbl></i>	<i><dbl></i>	<i><dbl></i>
1	0.538	-0.347	0.582
2	1.26	4.06	8.67
3	-0.643	-0.392	-0.689

Example 3: Does your (grand)parent's education level affect your own?



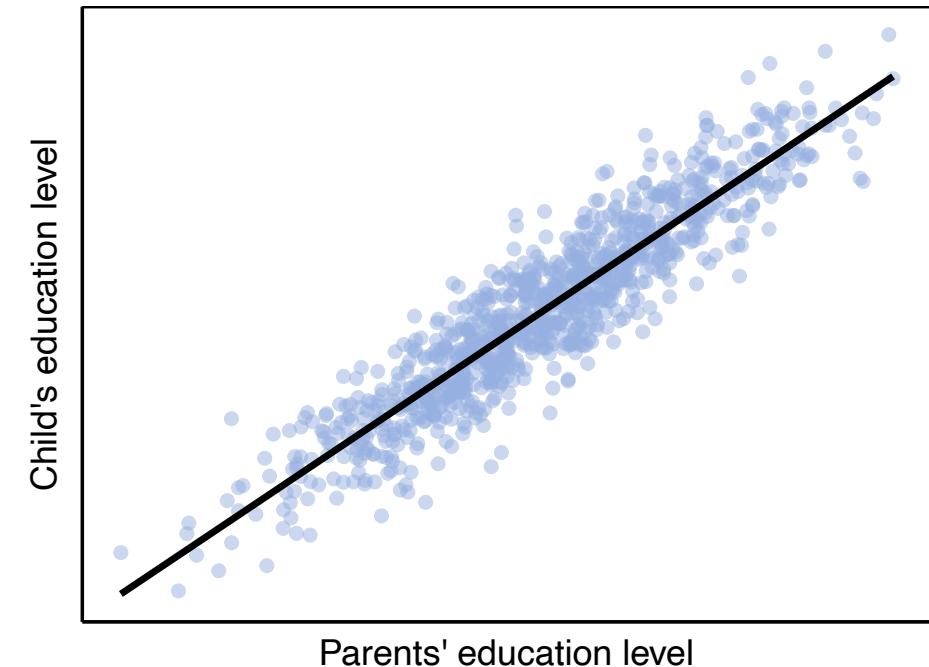
Parents positively affect child's education level

Model 1: $C \sim P$

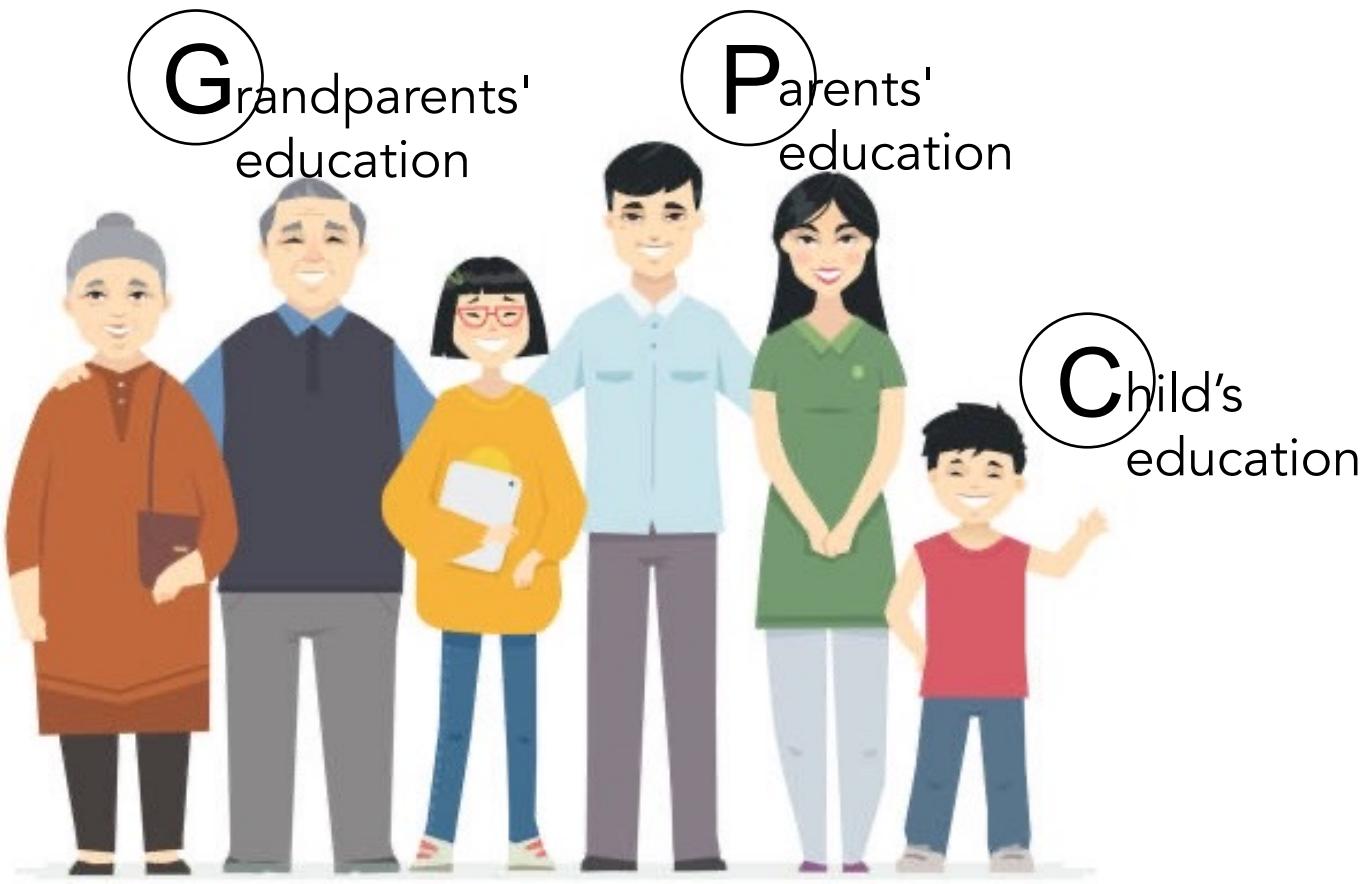
Model 2: $C \sim G$

Model 3: $C \sim P + G$

Model 4: $C \sim P \times G$



Example 3: Does your (grand)parent's education level affect your own?



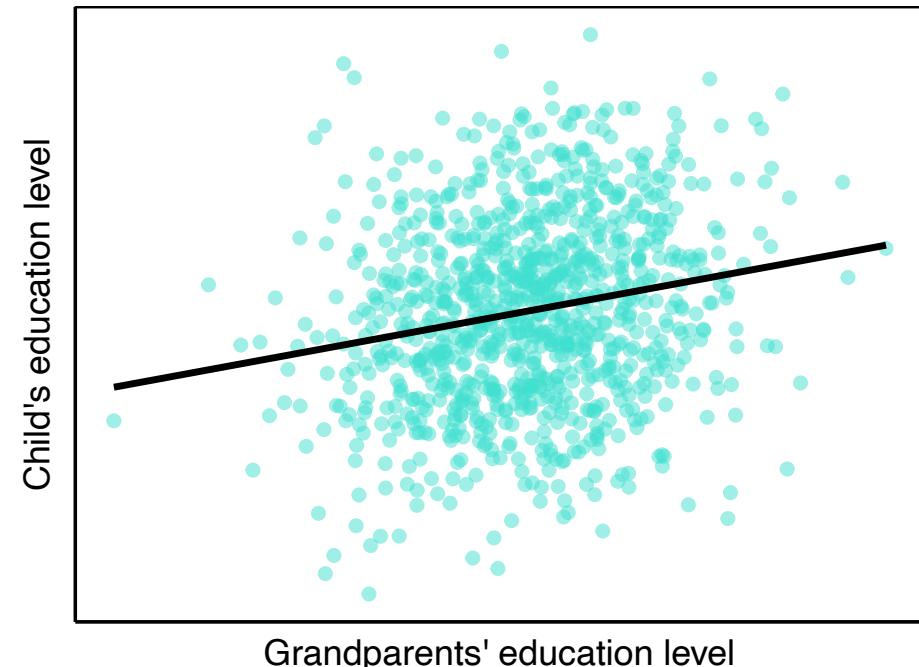
Grandparents positively affect
child's education level

Model 1: $C \sim P$

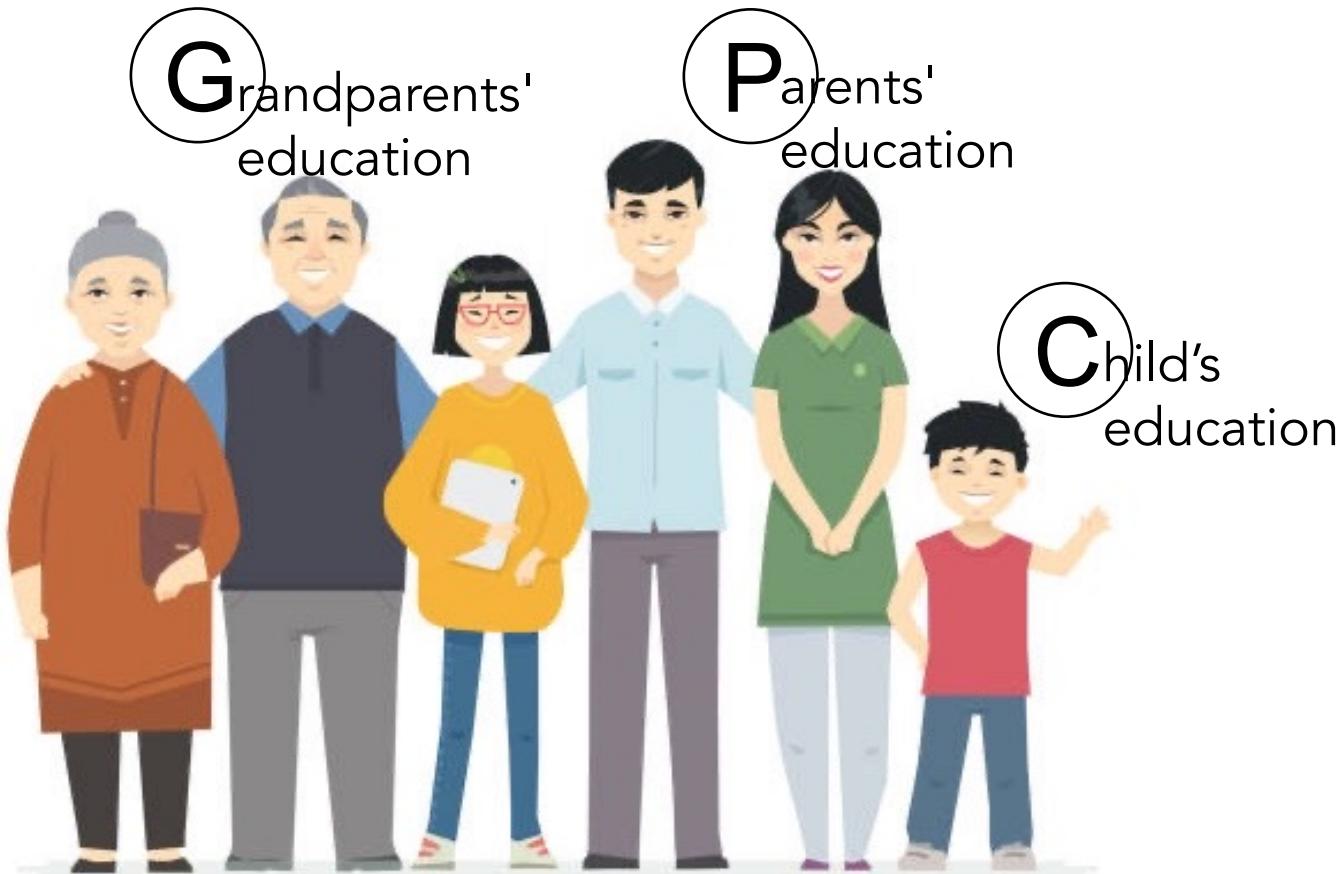
Model 2: $C \sim G$

Model 3: $C \sim P + G$

Model 4: $C \sim P \times G$



Example 3: Does your (grand)parent's education level affect your own?



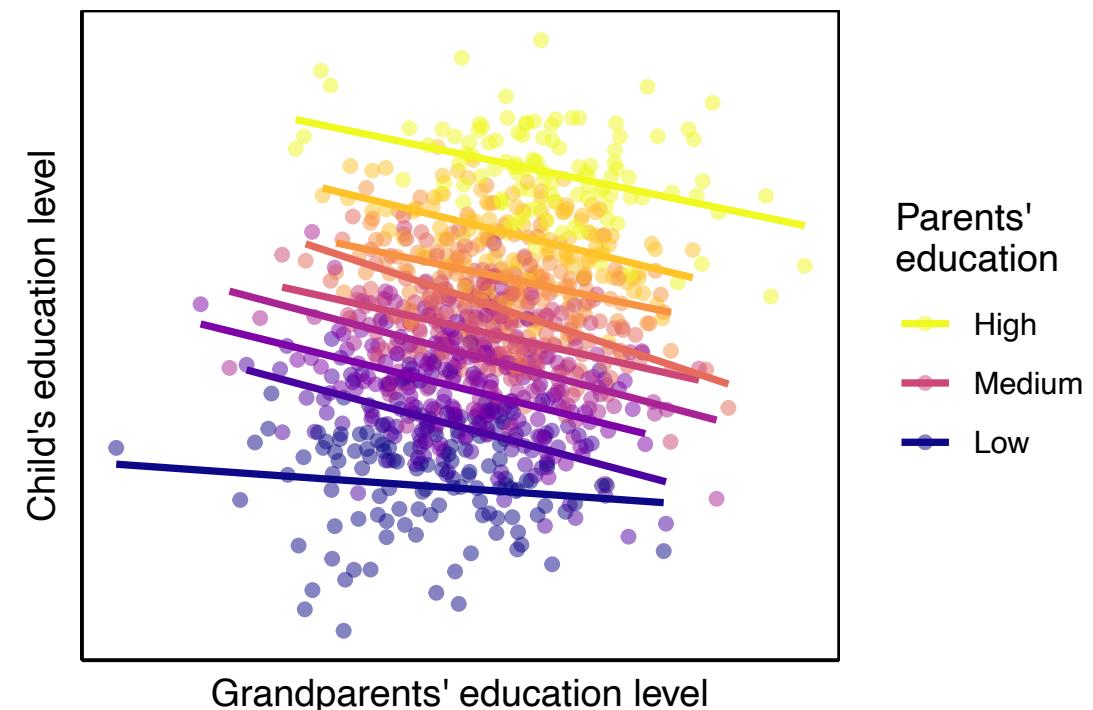
Grandparents **negatively** affect
child's education level, after
accounting for parents!

Model 1: $C \sim P$

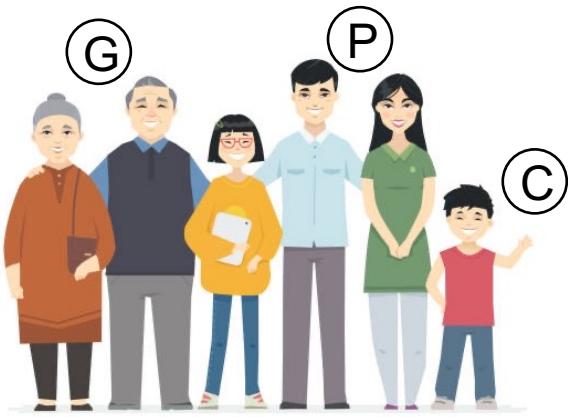
Model 2: $C \sim G$

Model 3: $C \sim P + G$

Model 4: $C \sim P \times G$



Example 3: Does your (grand)parent's education level affect your own?



But which model is best?

"Use model selection with AIC(c)!"
–Classical frequentists

"Use Bayesian stats and LOO/w_{aic}!"
–Bayes nerds

Model 1: C ~ P

Model 2: C ~ G

Model 3: C ~ P + G

Model 4: C ~ P x G

	df	AICc	ΔAICc
grandparent_plus_parent_mod	4	4243.016	0.000000
grandparent_x_parent_mod	5	4244.497	1.480975
parent_only_mod	3	4613.918	370.902240
grandparent_only_mod	3	6489.732	2246.716541

Conclusion: If your grandparents have PhDs, bad news for you!!

```
Call:  
lm(formula = C ~ G + P, data = .)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-6.3770 -1.4236  0.0484  1.4384  6.6580  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -0.09936   0.06632 -1.498   0.134  
G             -1.70282   0.07160 -23.783 <2e-16 ***  
P              2.67736   0.02937  91.150 <2e-16 ***
```

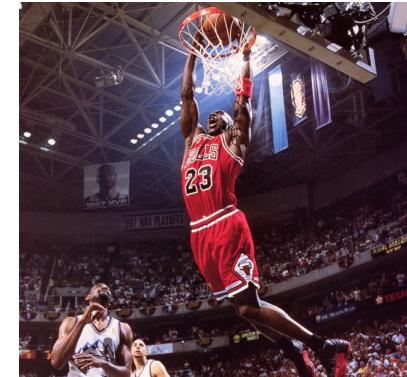
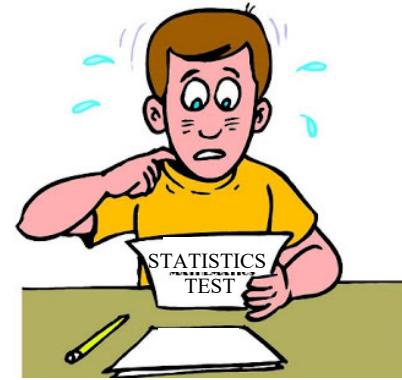
Chapter 3: Two choices



Modern Science/Statistics

(Pearson & Fisher's legacy)

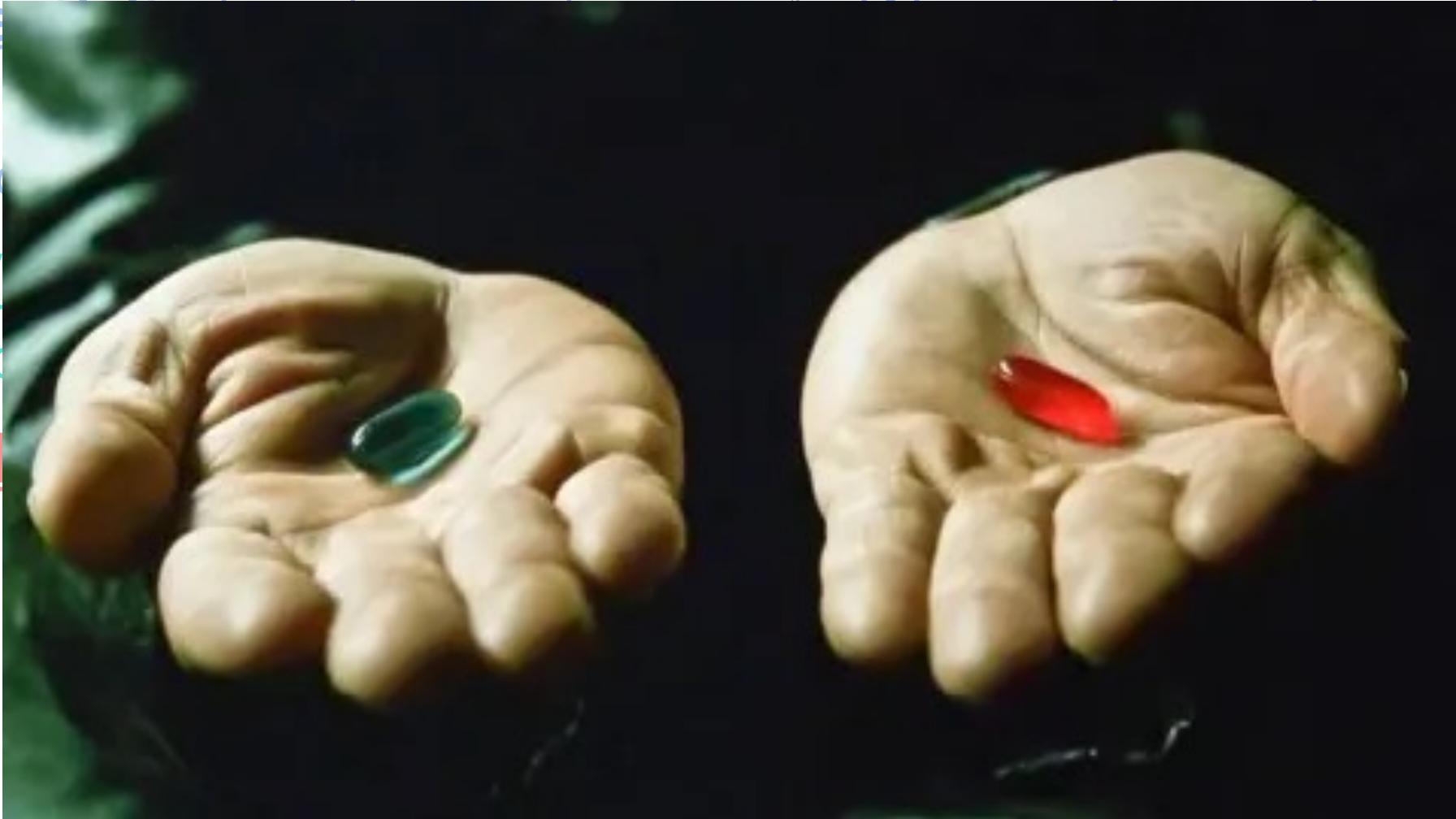
- Objectivity in science: do not include any of our own personal assumptions
 - Can't attribute causation without experimentation!
 - Model selection decides which x-variables to include in our model!
- Despite our best pre-conceived hypotheses, **sleep has no effect on test scores!**
- To truly get at causality, we must force the NBA to recruit shorter players with no scoring ability
- (Note: I am available for recruitment to the NBA on: +6144875309)
- **Grandparents' education level is BAD** for our own! (or at least negatively correlated)



Modern Science/Statistics

(Pearson & Fisher's legacy)

- Objective statistics
not individual
personal
expertise
- Can't
without
information
- Model
which
includes

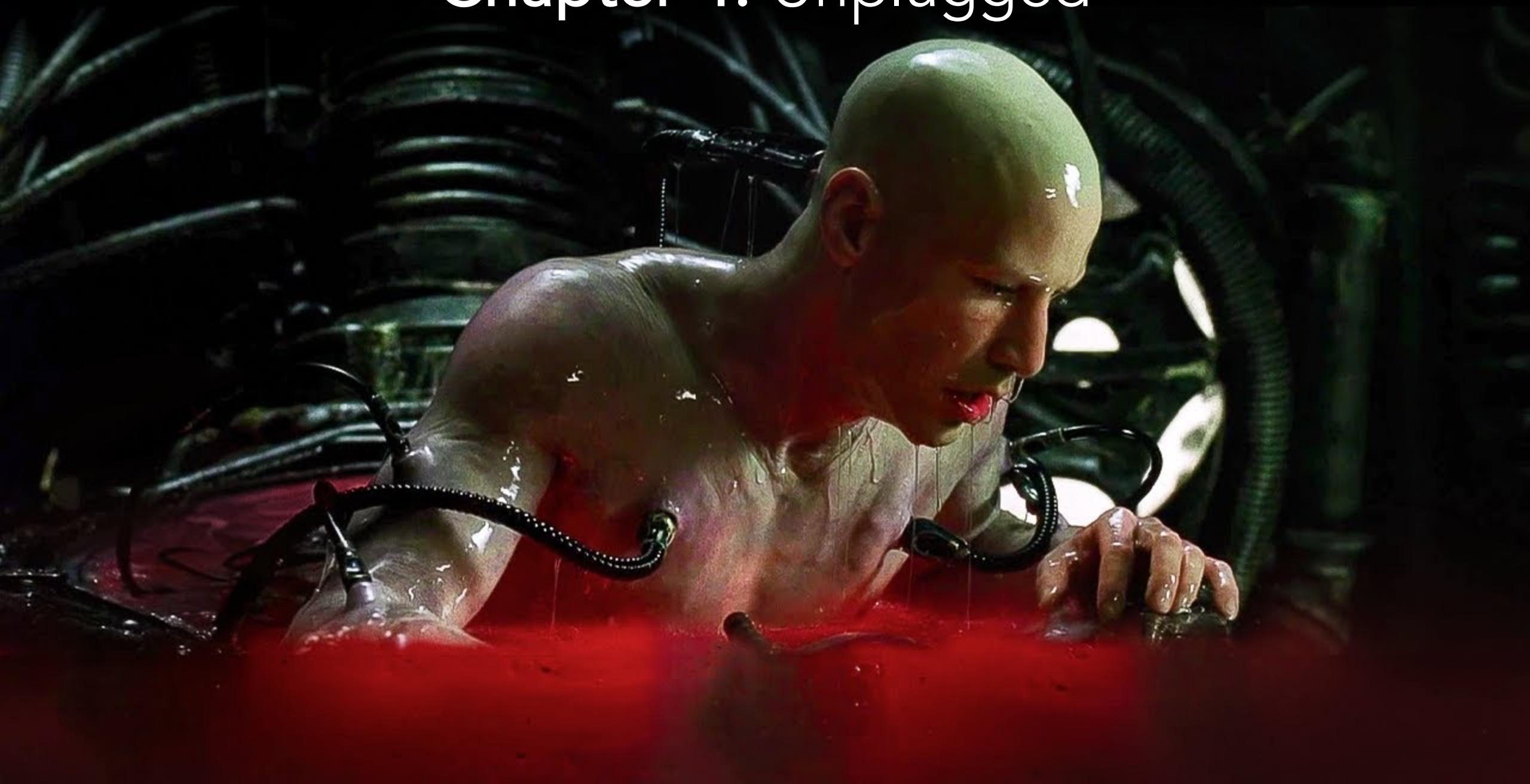


Causal Inference

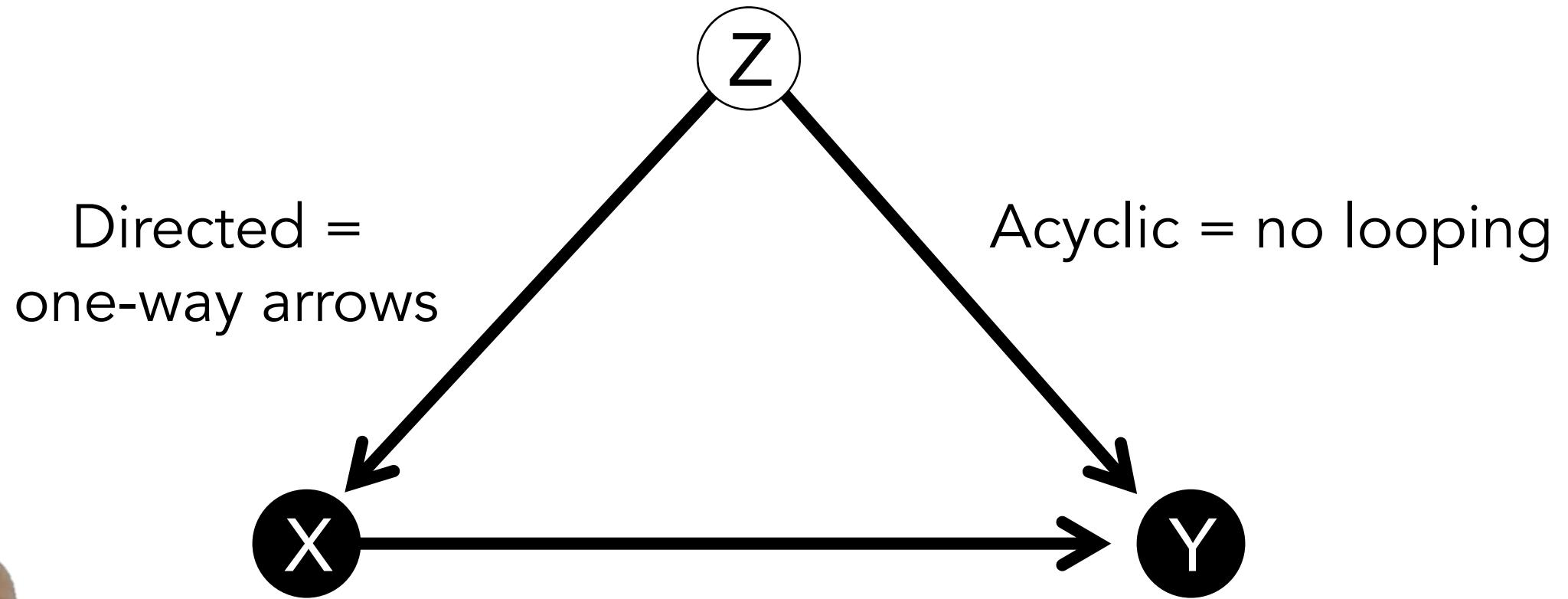
(Sewall Wright and Judea Pearl)

in inform
ables to
' upon!

Chapter 4: Unplugged



Constructing DAGs (Directed Acyclical Graphs)

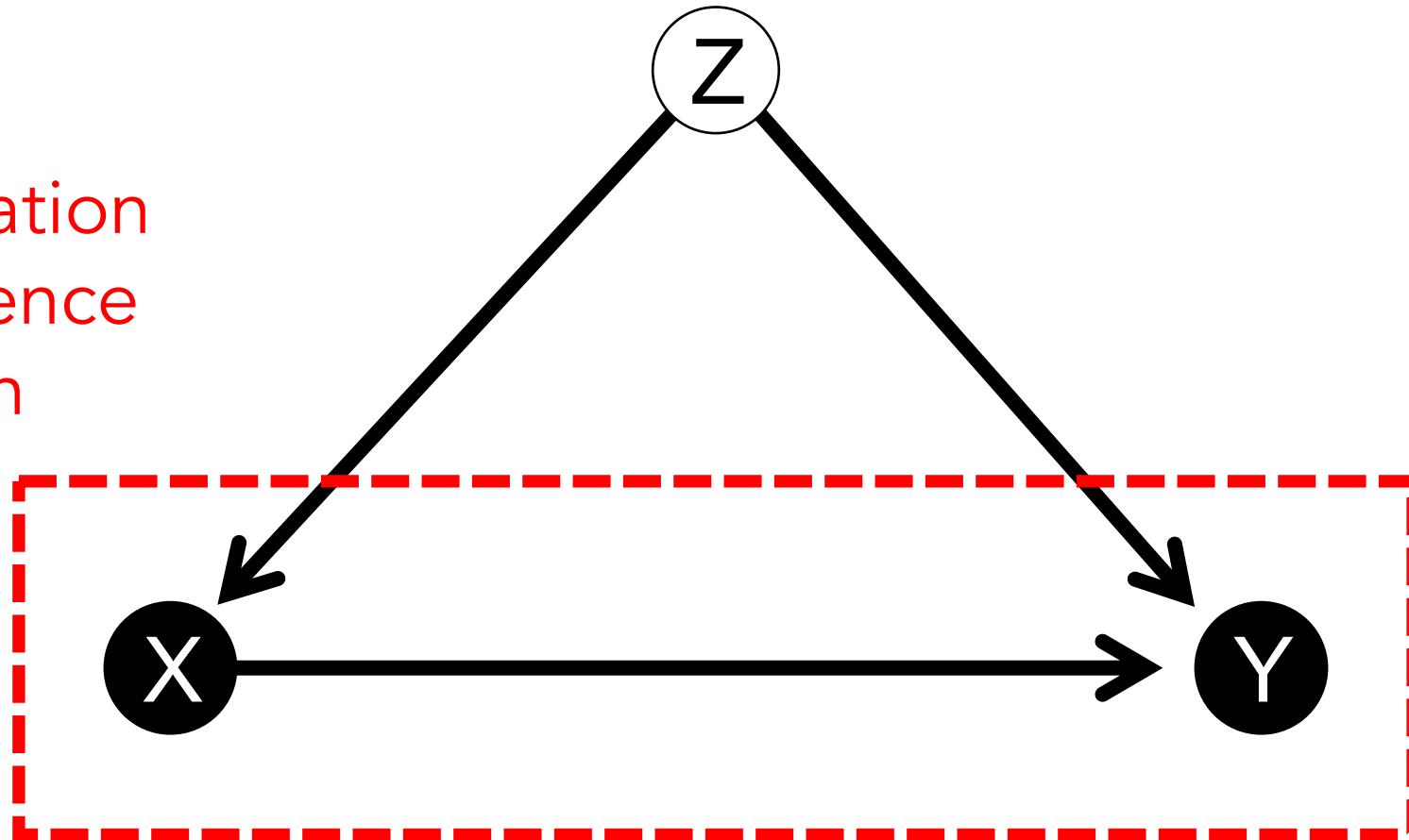


Graph = a late 19th century
abbreviation of "graphic formula"



Good experimental design can help eliminate unknown confounds

Randomization
Independence
Replication
Blocking
Controls



When we can't manipulate, we can make use of our causal assumptions!

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

$$X \rightarrow Z \rightarrow Y$$

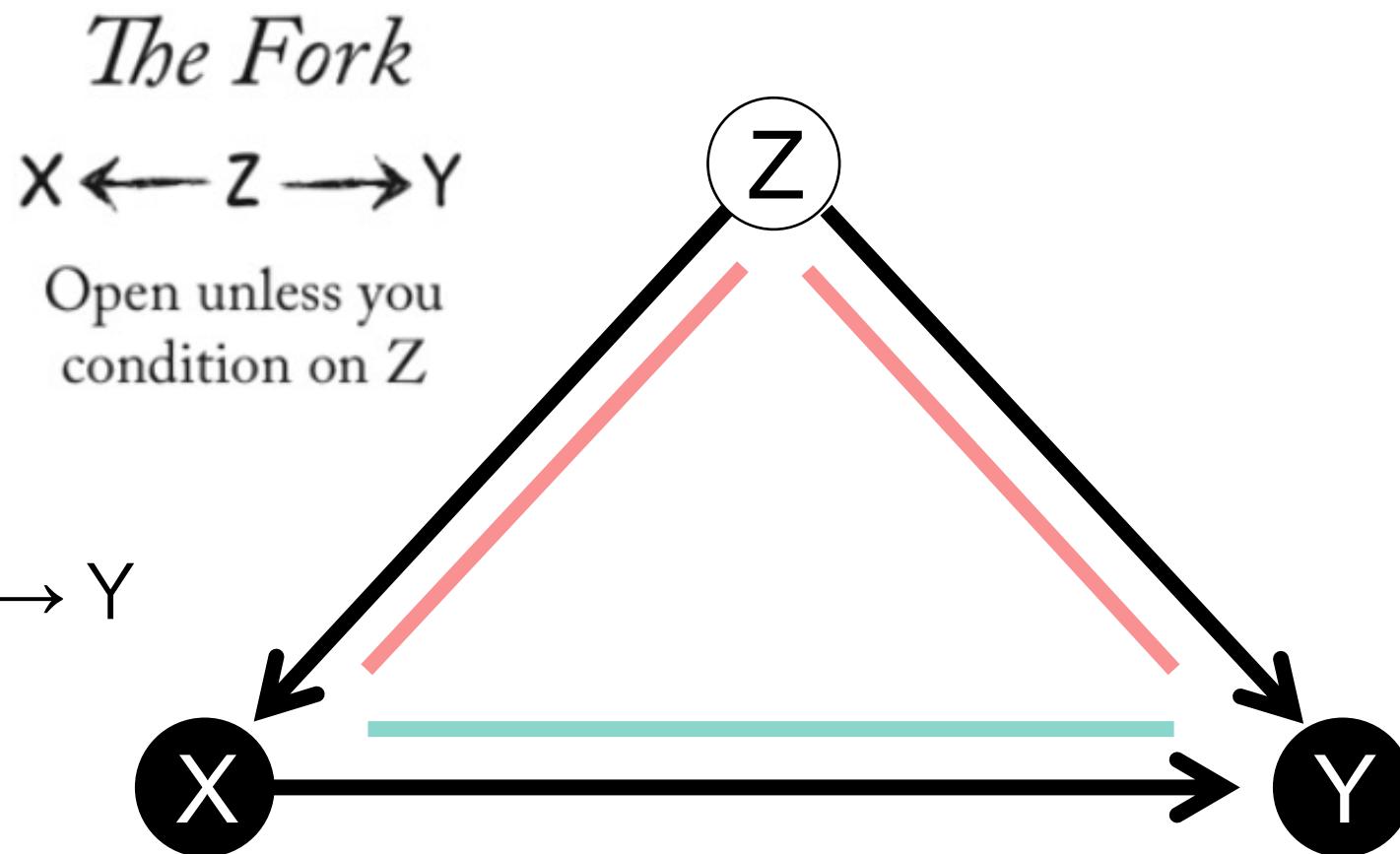
↓
A



When we can't manipulate, we can make use of our causal assumptions!

Two paths:

- Path A: $X \rightarrow Y$
- Path B: $X \leftarrow Z \rightarrow Y$



When we can't manipulate, we can make use of our causal assumptions!

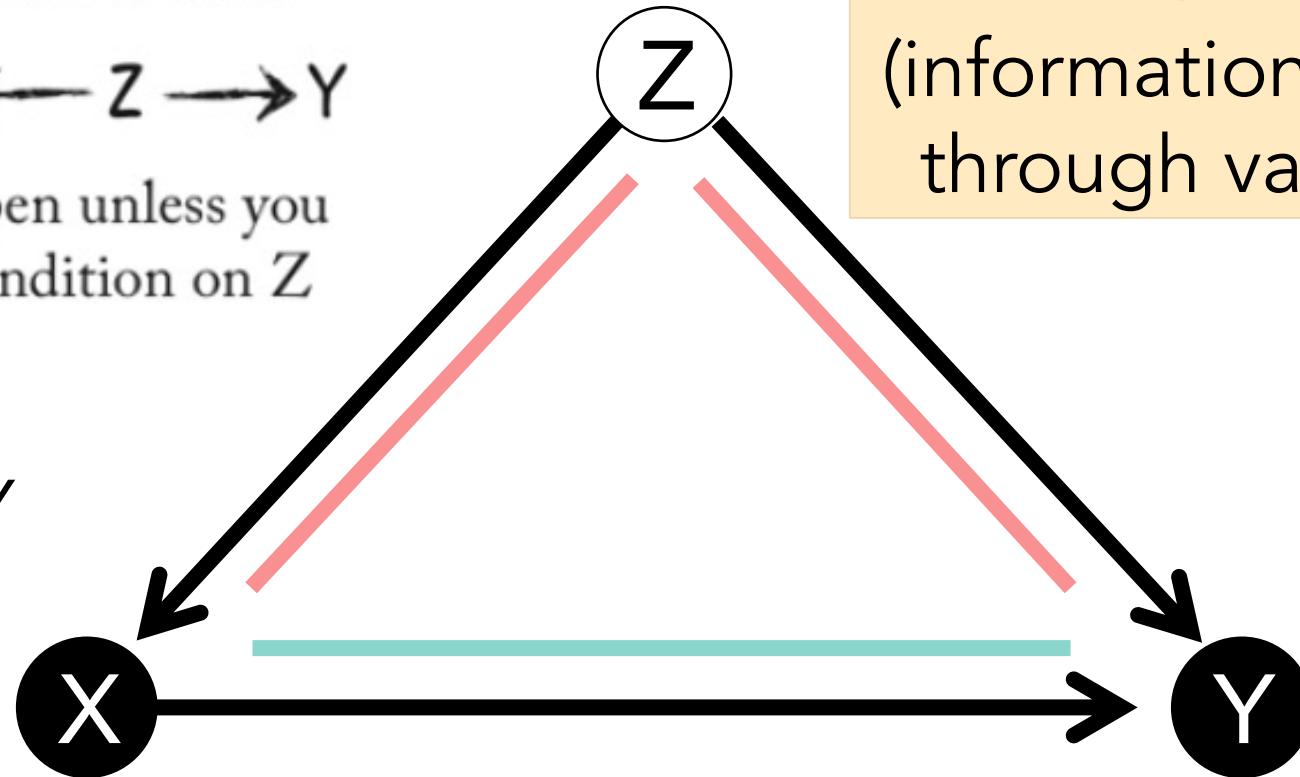
Two paths:

- Path A: $X \rightarrow Y$
- Path B: $X \leftarrow Z \rightarrow Y$



The Fork
 $X \leftarrow Z \rightarrow Y$

Open unless you
condition on Z



There is a backdoor path through Z
that must be closed!

When we can't manipulate, we can make use of our causal assumptions!

The Fork

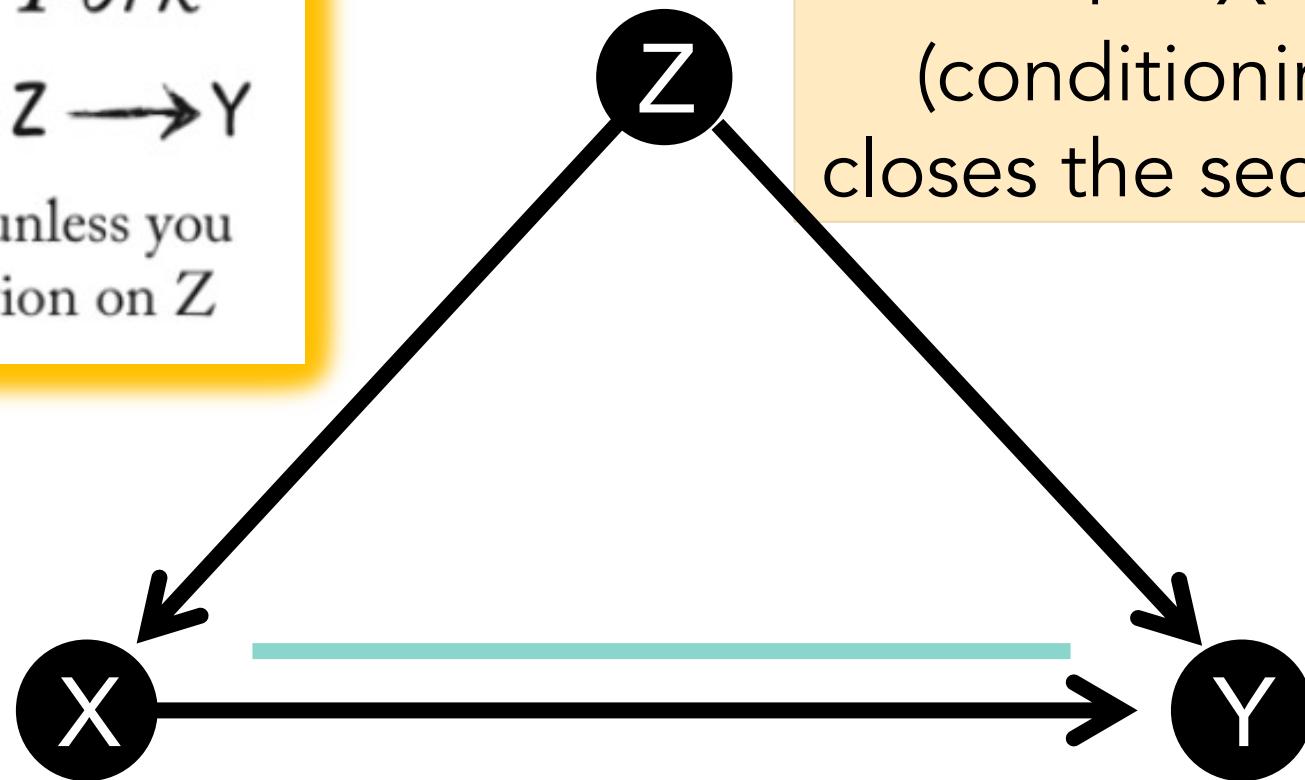
$$X \leftarrow Z \rightarrow Y$$

Open unless you condition on Z

Ideal statistical model:
 $Y \sim X + Z$
(conditioning on Z
closes the second path)

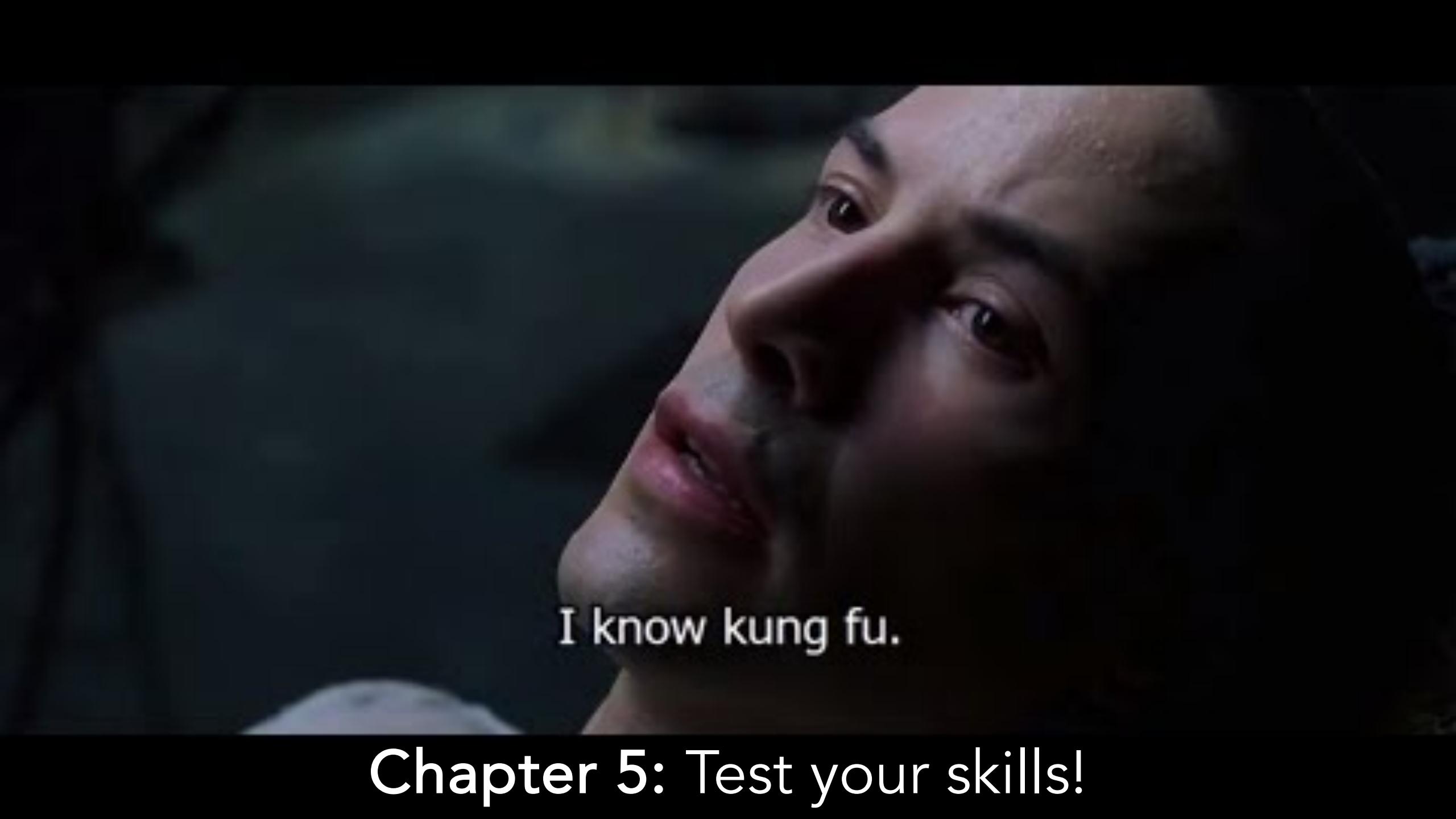
Two paths:

- Path A: $X \rightarrow Y$
- Path B: $X \leftarrow Z \rightarrow Y$



There is a backdoor path through Z
that must be closed!



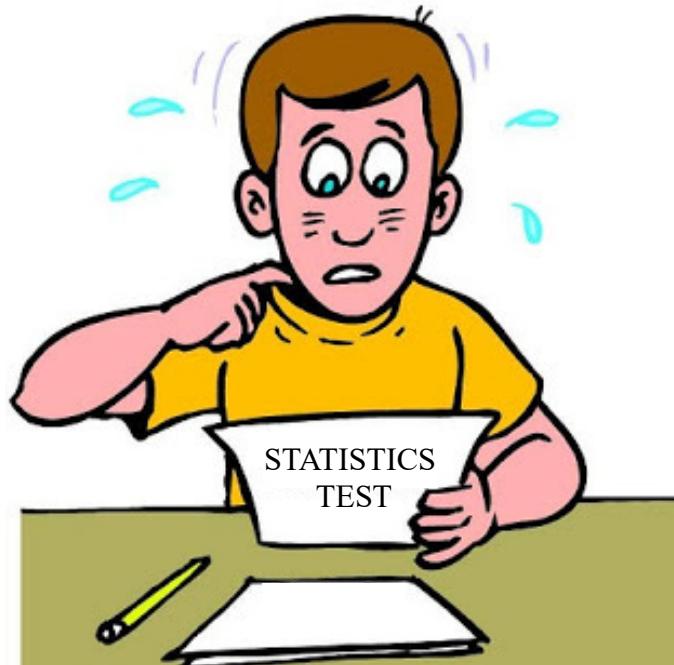


I know kung fu.

Chapter 5: Test your skills!

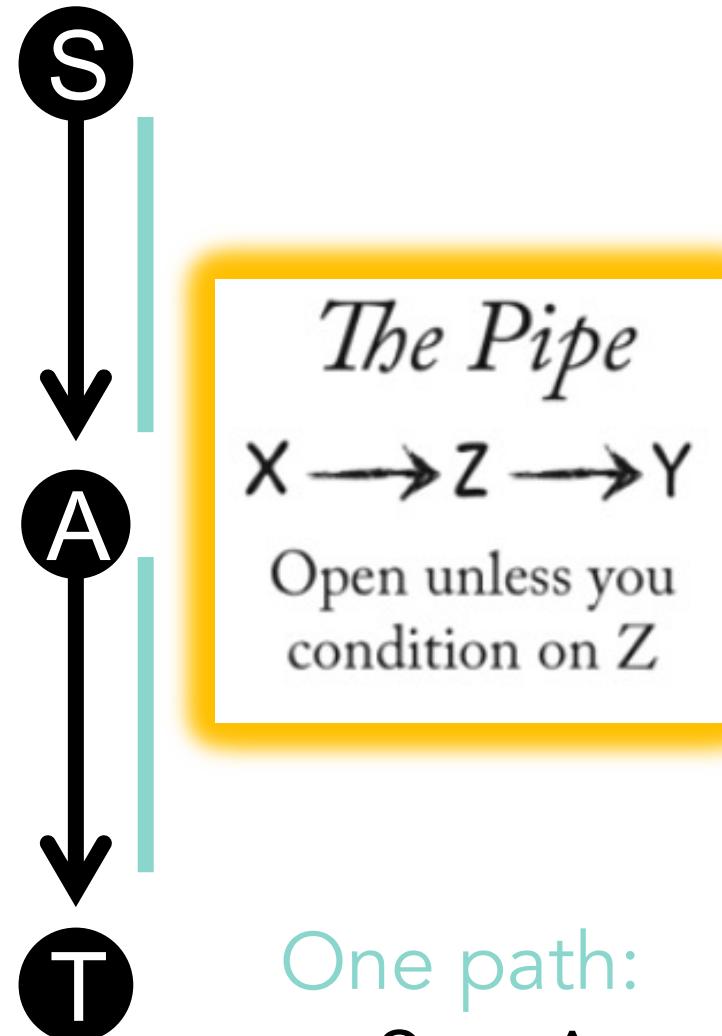
Example 1: Does sleep level affect student test results?

	Sleep	Alert	Test_score
	<dbl>	<dbl>	<dbl>
1	-0.626	0.509	-0.378
2	0.184	1.30	-0.627



Example 1: Does sleep level affect student test results?

	Sleep	Alert	Test_score
1	-0.626	0.509	-0.378
2	0.184	1.30	-0.627

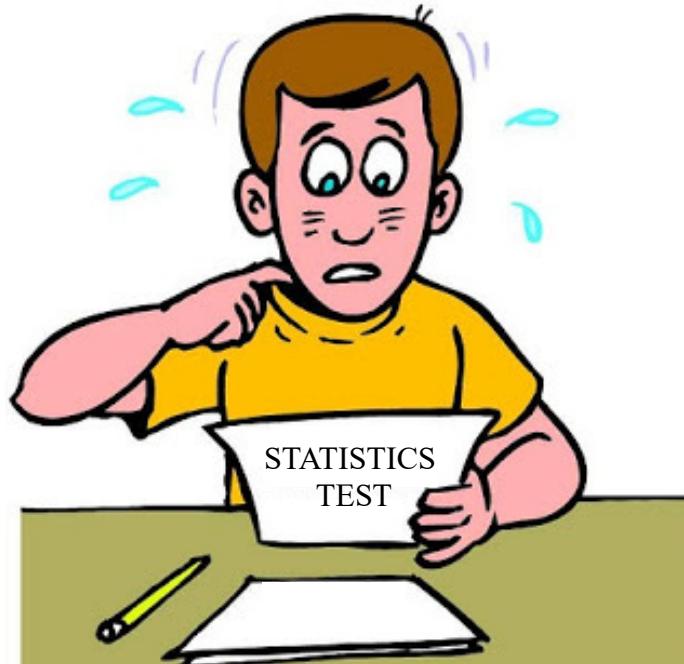


The Pipe
 $X \rightarrow Z \rightarrow Y$
Open unless you condition on Z

One path:
• $S \rightarrow A \rightarrow T$

Example 1: Does sleep level affect student test results?

	Sleep	Alert	Test_score
1	-0.626	0.509	-0.378
2	0.184	1.30	-0.627



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.016811	0.032880	0.511	0.609
sleep	0.004877	0.045431	0.107	0.915
alertness	1.010407	0.031905	31.670	<2e-16 ***

The Pipe

$x \rightarrow z \rightarrow y$

Open unless you
condition on Z

Not ideal model:
Test score $\sim S + A$

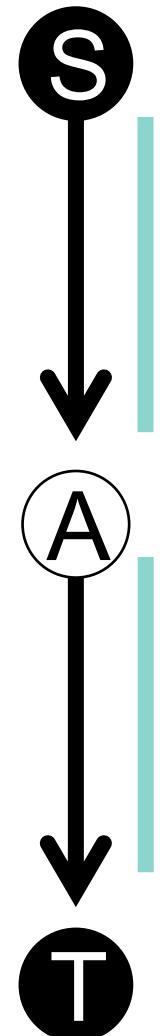
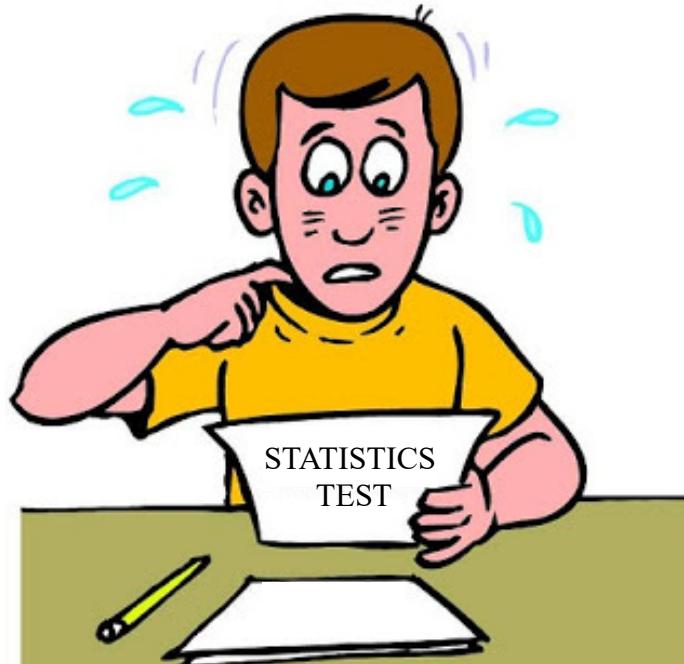
Including alertness blocks the
effect of sleep on test score!

One path:

- $S \rightarrow A \rightarrow T$

Example 1: Does sleep level affect student test results?

	Sleep	Alert	Test_score
1	-0.626	0.509	-0.378
2	0.184	1.30	-0.627



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.000305	0.046821	-0.007	0.995
sleep	1.055623	0.045261	23.323	<2e-16 ***

The Pipe

$X \rightarrow Z \rightarrow Y$

Open unless you condition on Z

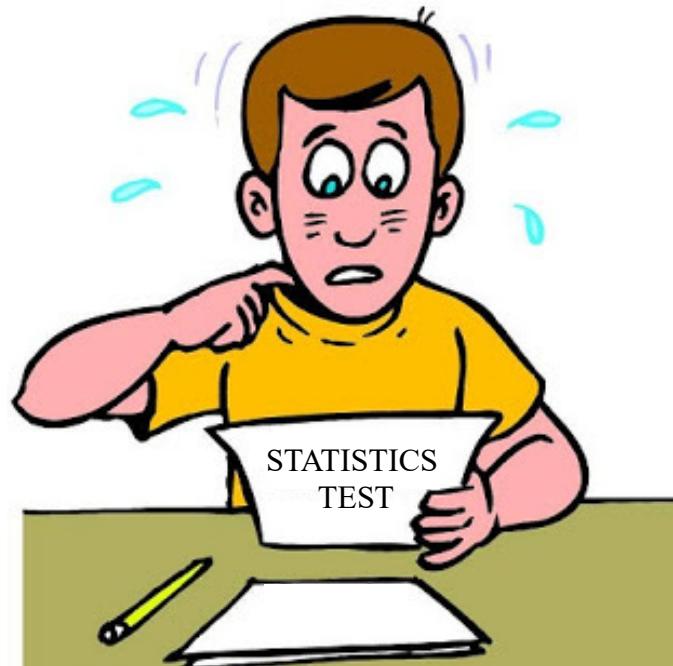
Ideal model:
Test score ~ Sleep

Example 1: Does sleep level affect student test results?

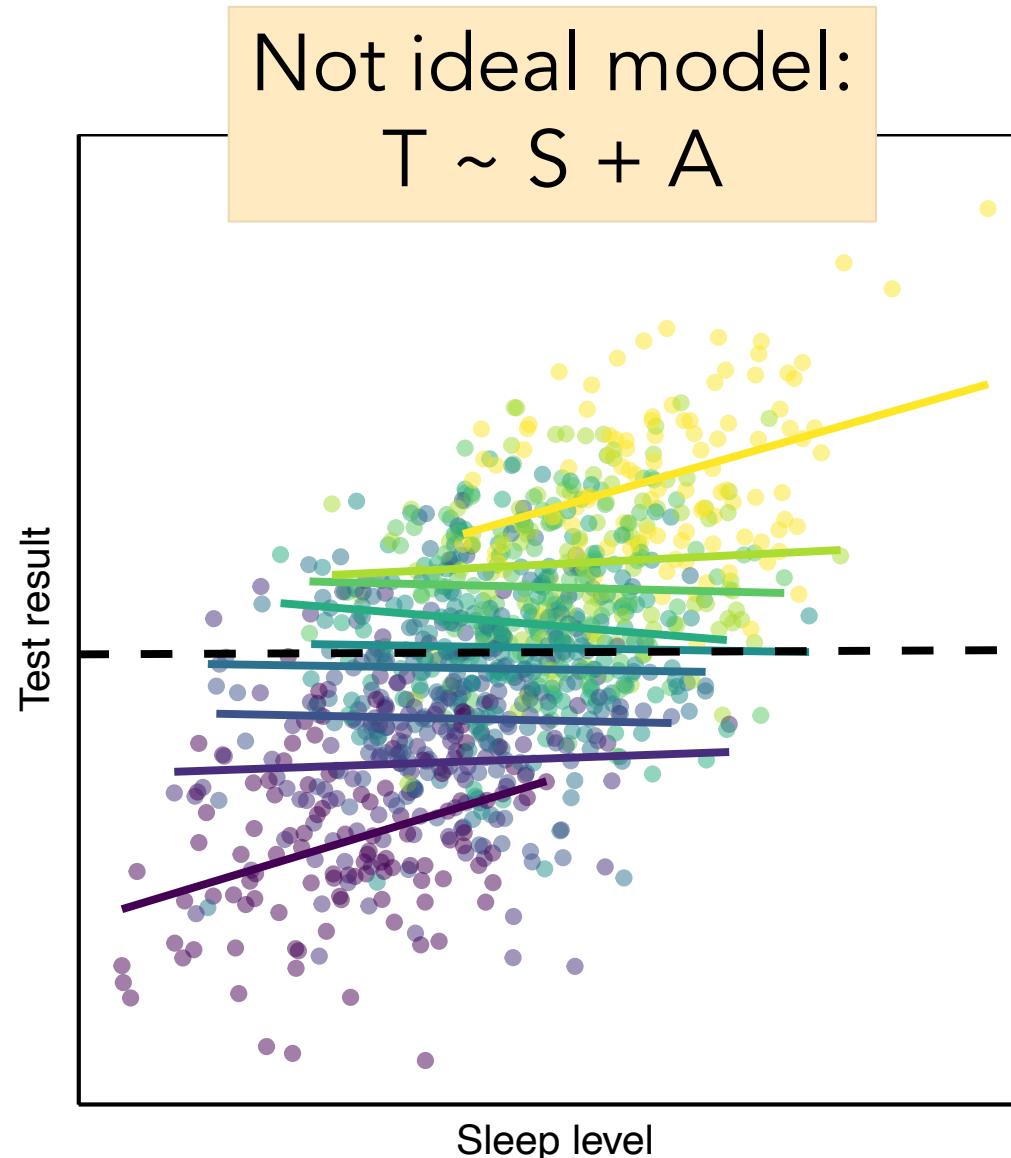
The Pipe

$X \rightarrow Z \rightarrow Y$

Open unless you
condition on Z



Not ideal model:
 $T \sim S + A$

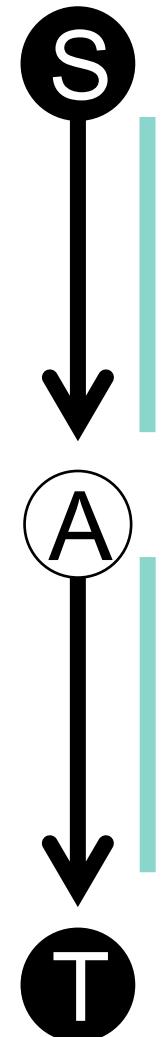


Example 1: Does sleep level affect student test results?

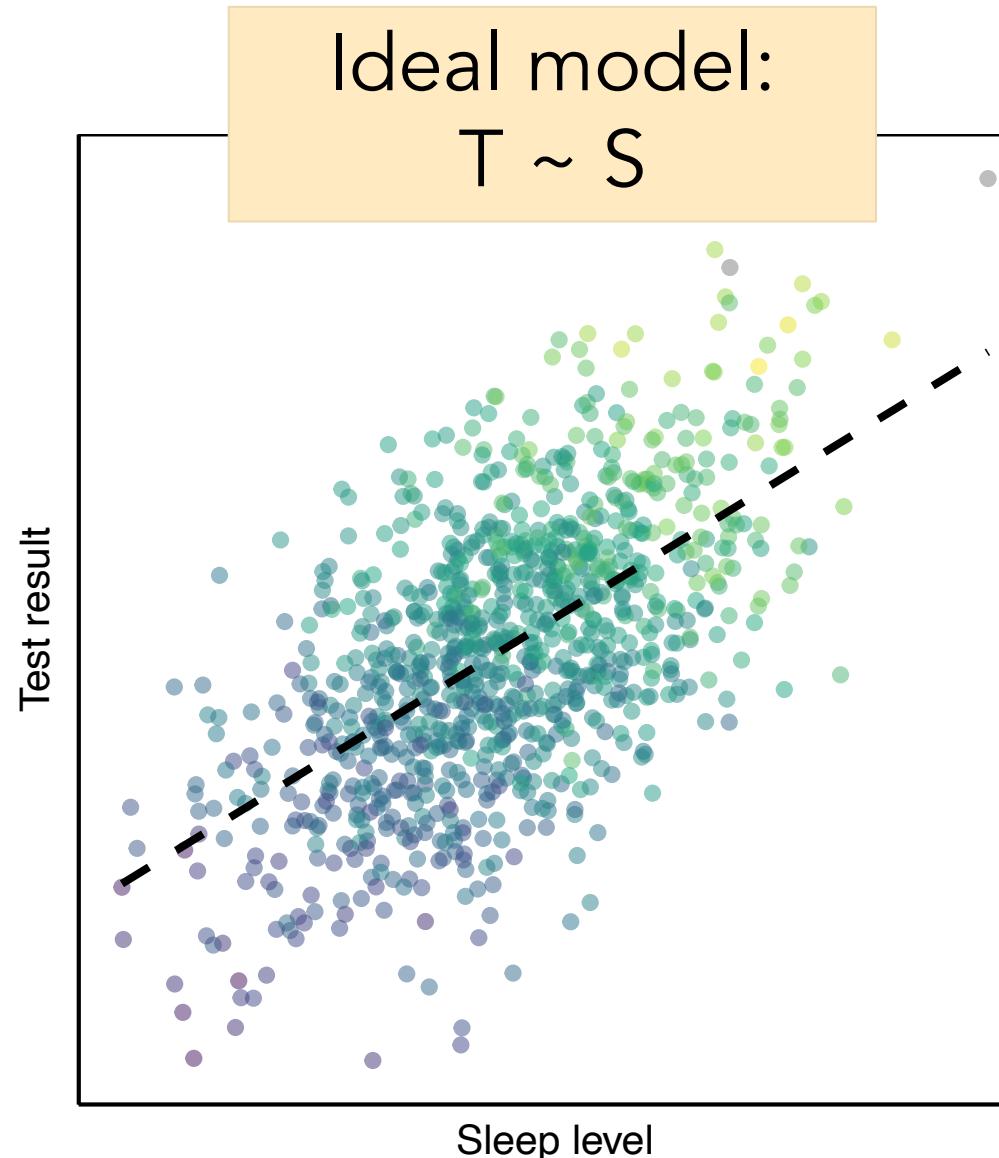
The Pipe

$X \rightarrow Z \rightarrow Y$

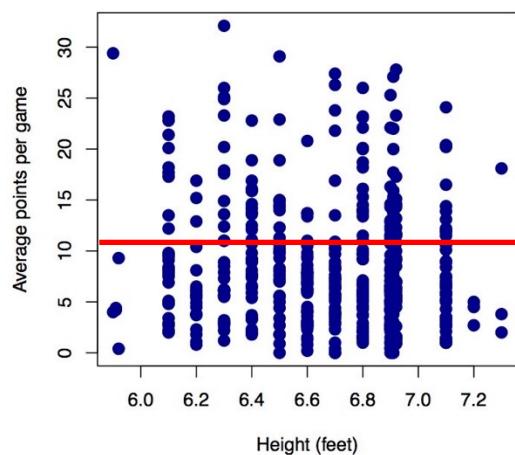
Open unless you
condition on Z



Ideal model:
 $T \sim S$



Example 2: Does basketball player height affect scoring?



Two paths:

- Path A: $H \rightarrow S$
- Path B: $H \rightarrow B \leftarrow S$

The Collider

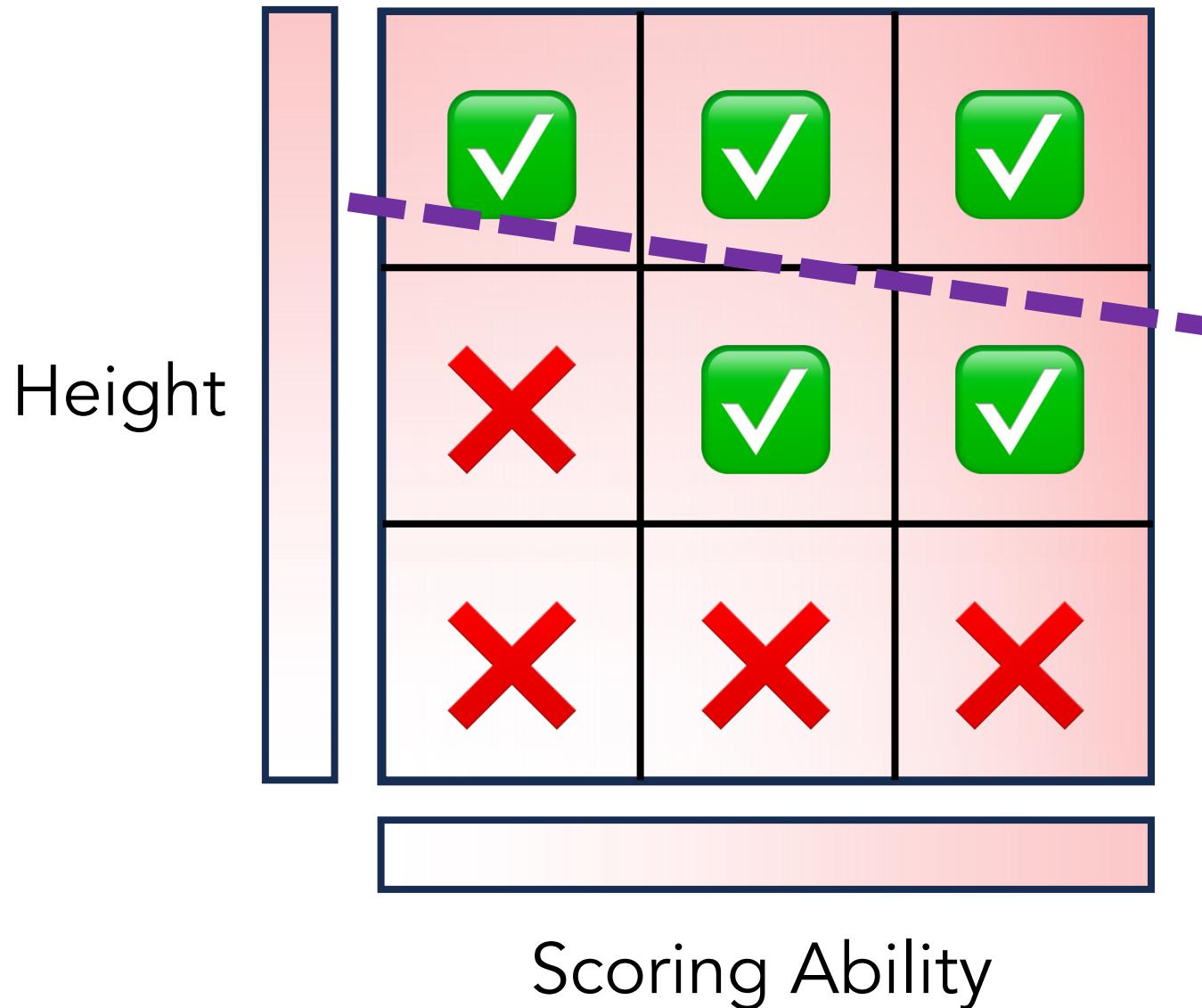
$$X \rightarrow Z \leftarrow Y$$

Closed until you
condition on Z

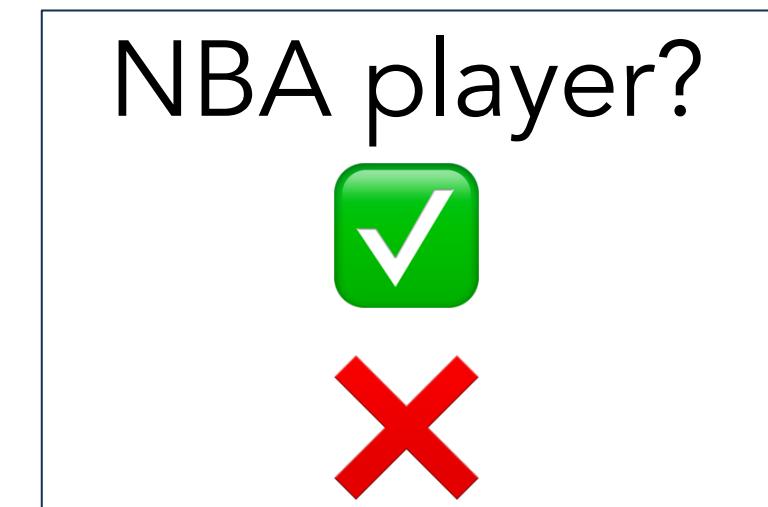
Whether or not you
become an NBA player



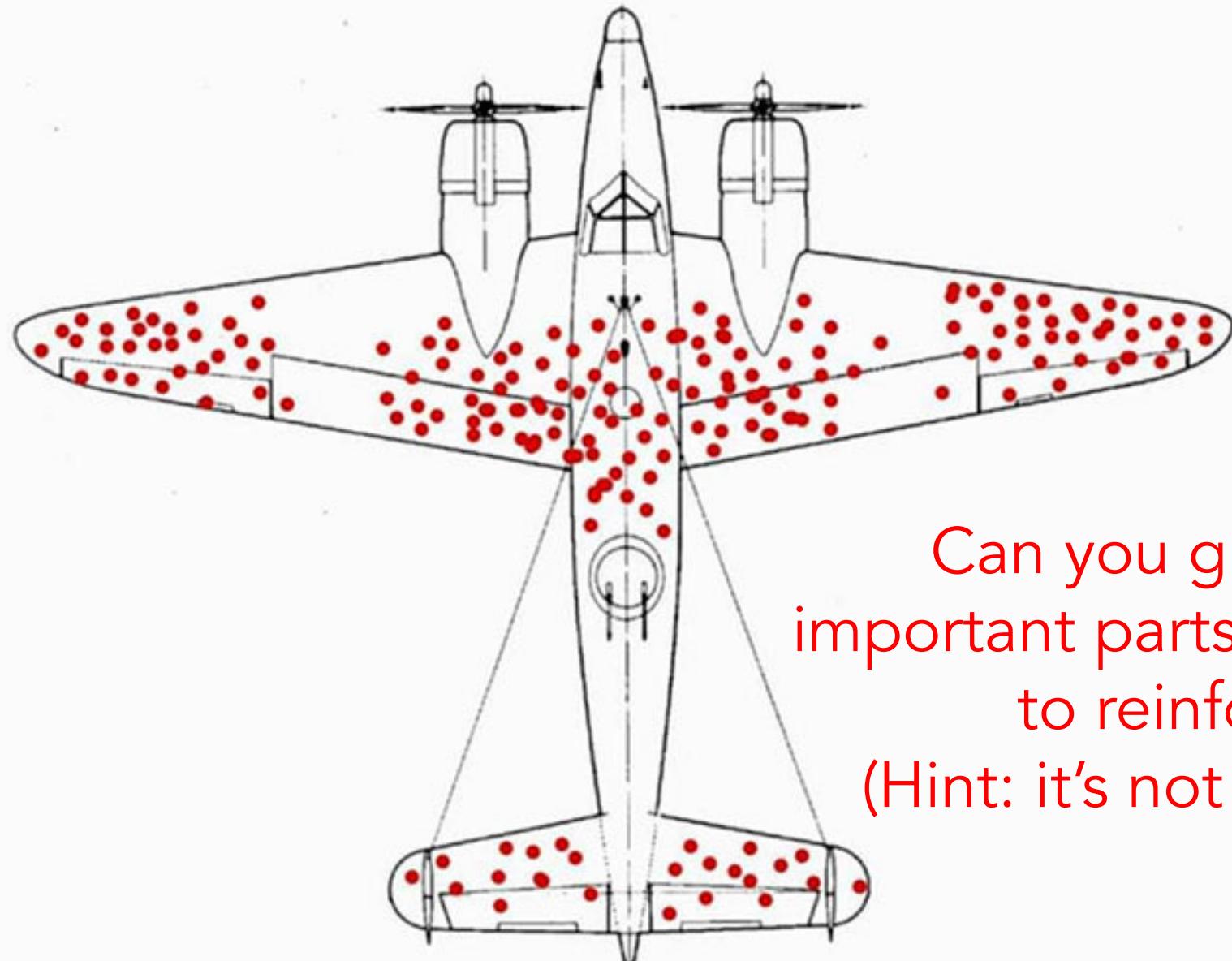
Population



Simpson's paradox
Survivorship bias



Planes that came back from WWII tended to be shot on the wings



Can you guess the
important parts of the plane
to reinforce?
(Hint: it's not the wings)

Example 2: Does basketball player height affect scoring?



The Collider

$$X \rightarrow Z \leftarrow Y$$

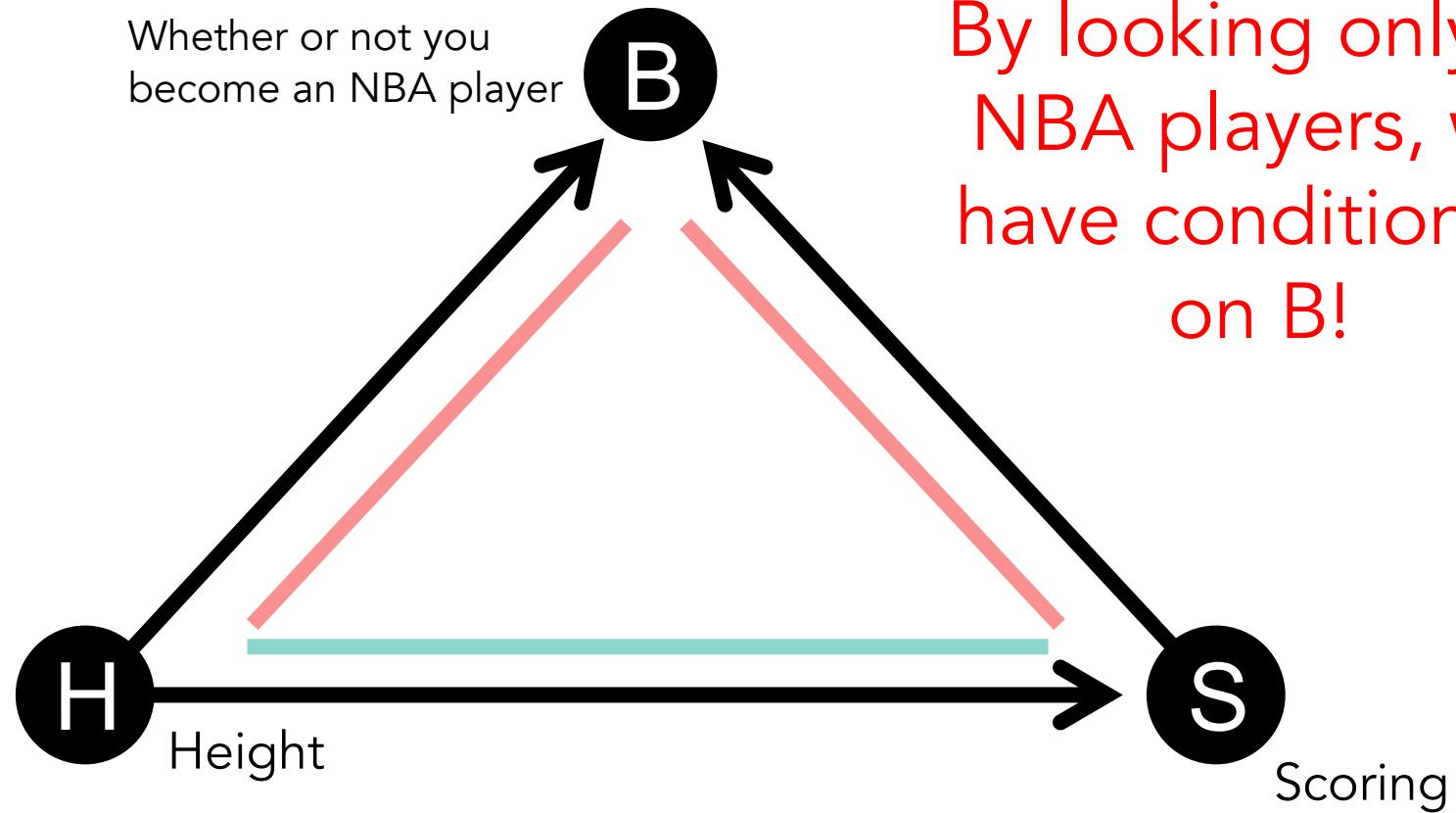
Closed until you
condition on Z

Two paths:

- Path A: H → S
 - Path B: H → B ← S

Not ideal:
 $S \sim H + B$

By looking only at
NBA players, we
have conditioned
on B!



Example 2: Does basketball player height affect scoring?



The Collider

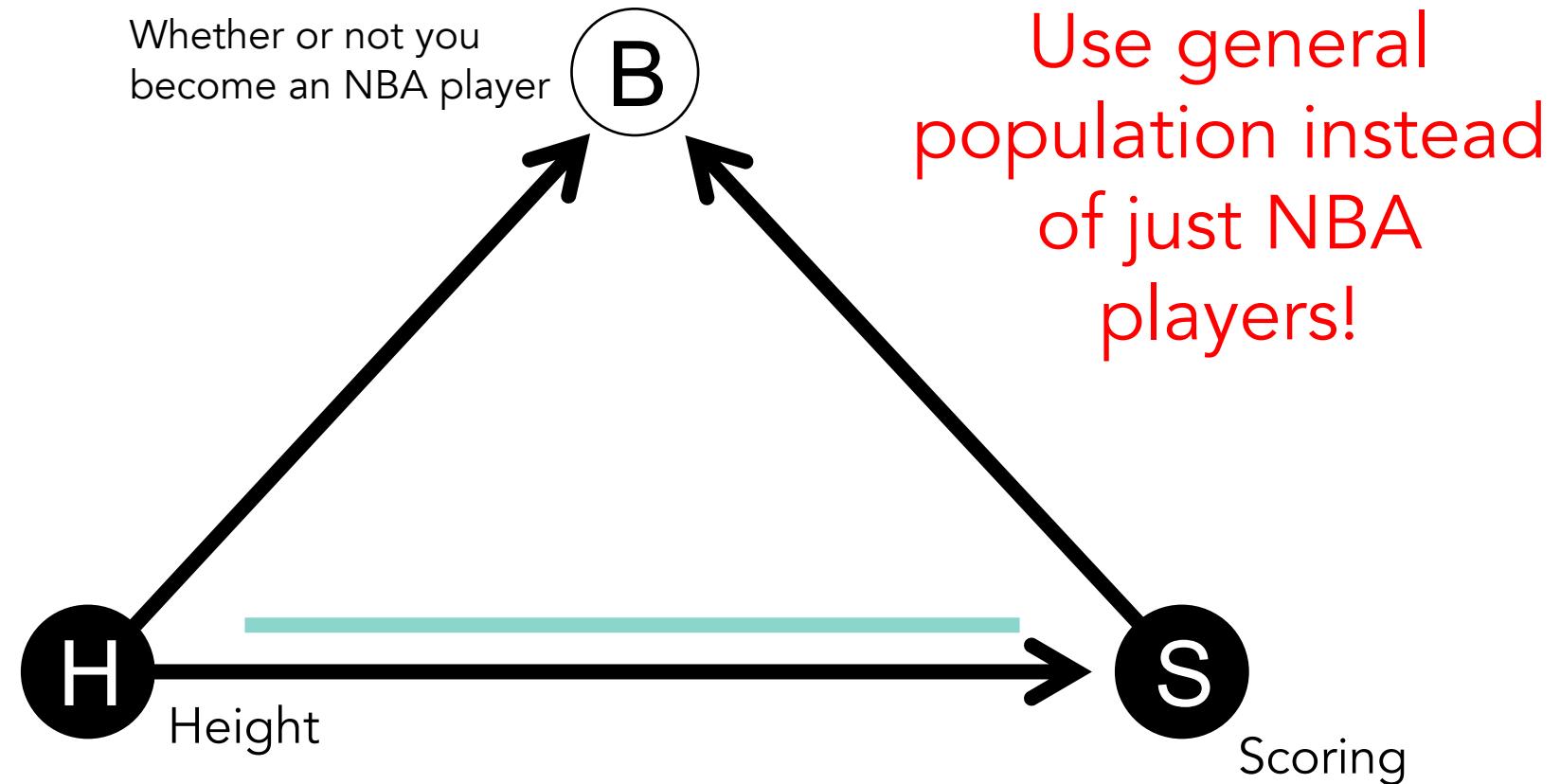
$$X \rightarrow Z \leftarrow Y$$

Closed until you
condition on Z

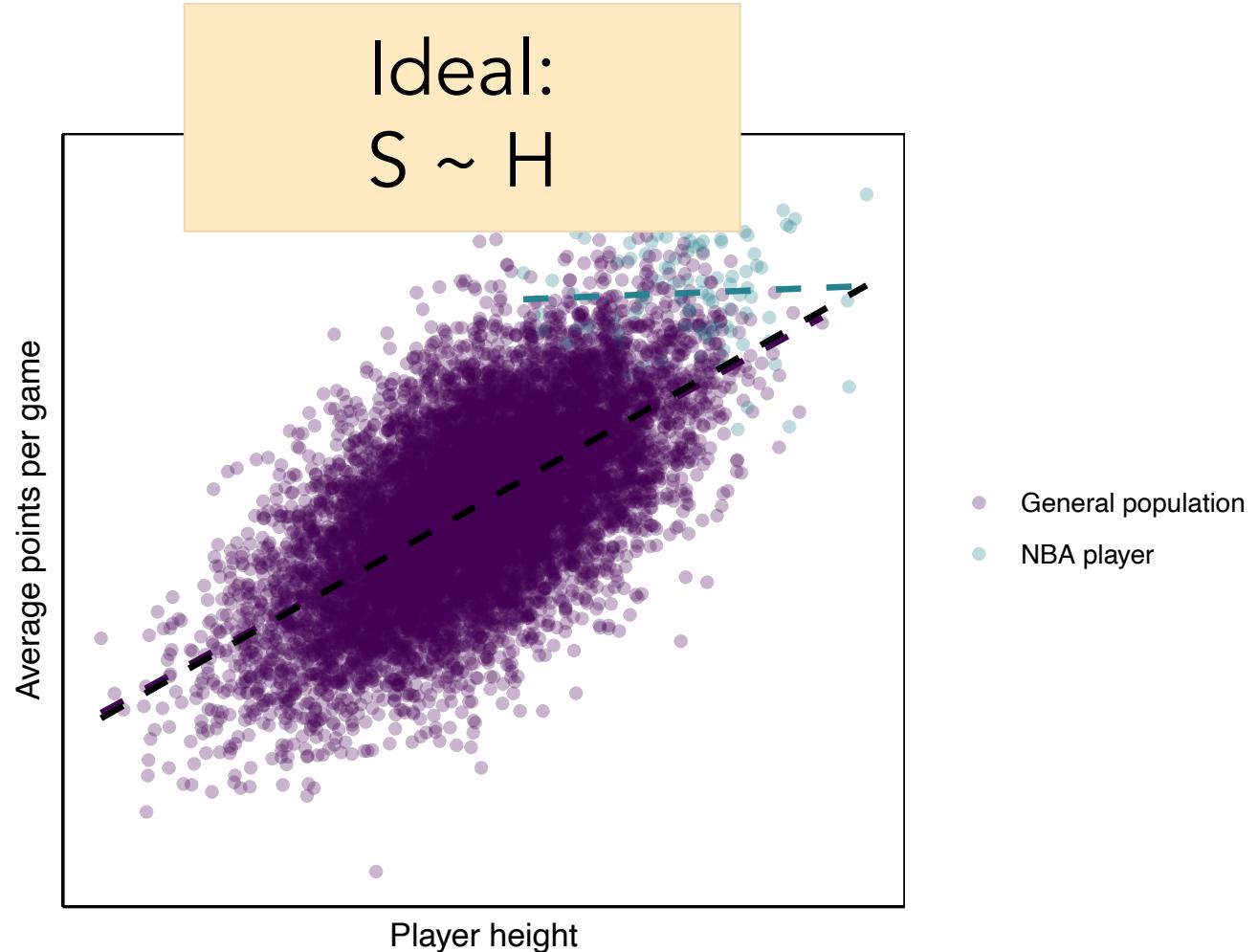
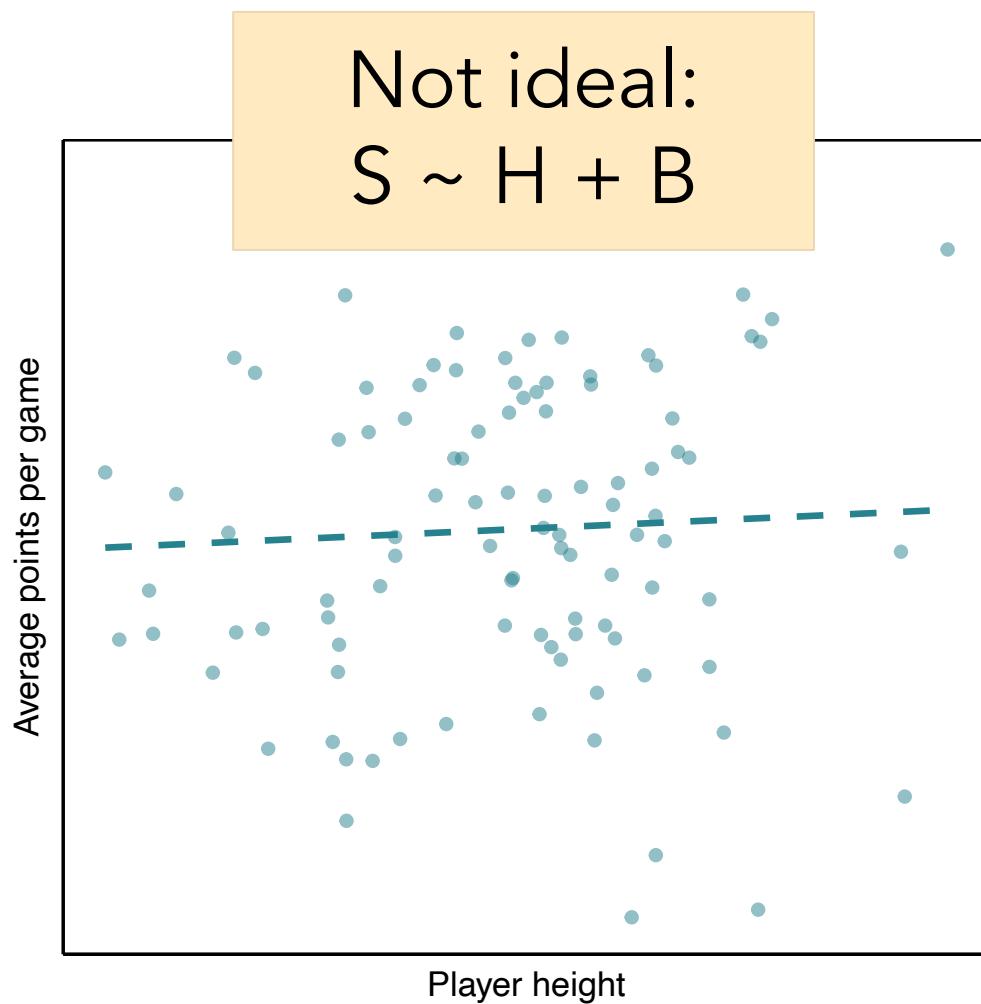
Two paths:

- Path A: $H \rightarrow S$
- Path B: $H \rightarrow B \leftarrow S$

Ideal:
 $S \sim H$



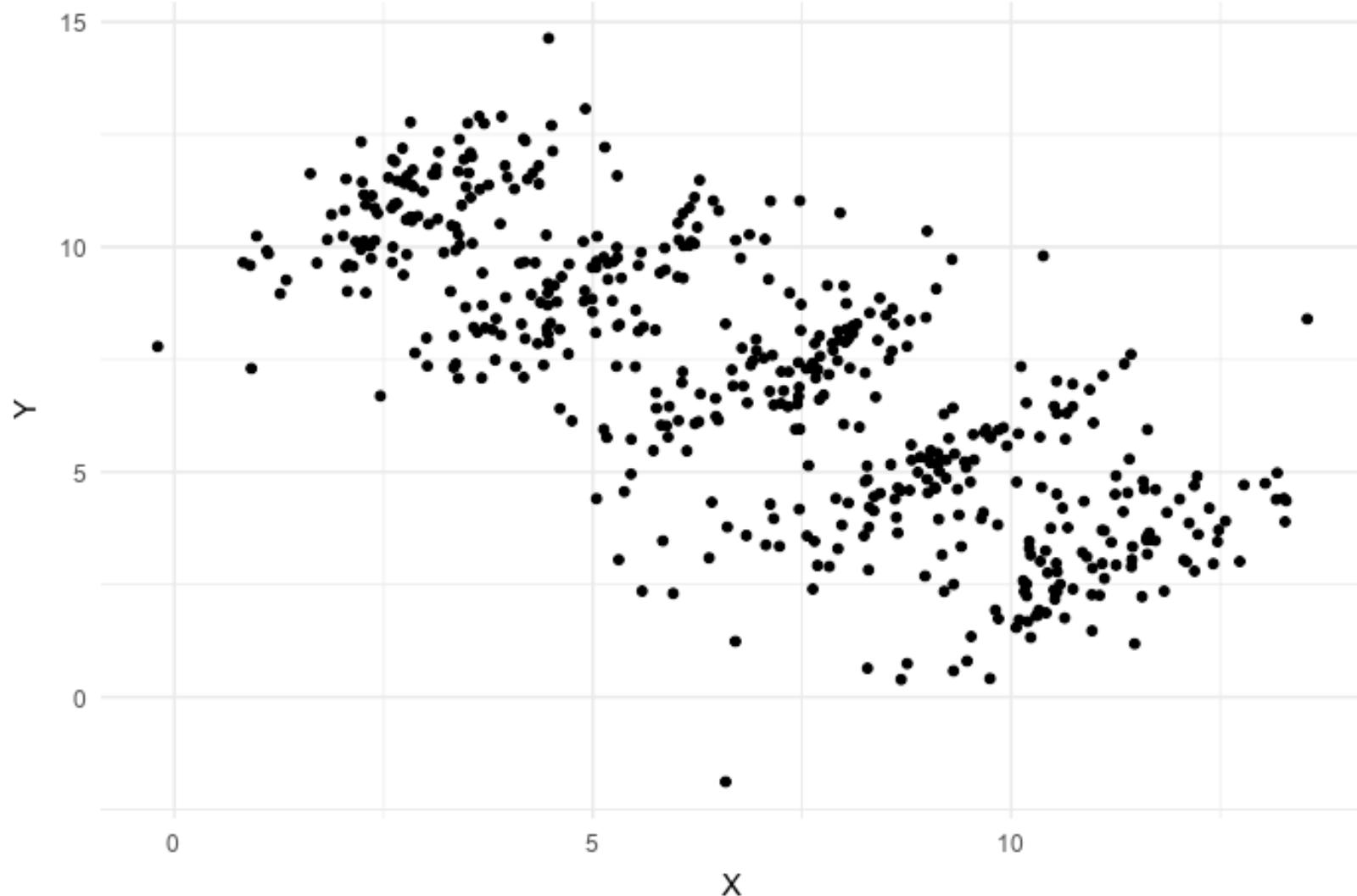
Example 2: Does basketball player height affect scoring?



No relationship between height
and scoring ability *when only
looking at NBA players!*

Simpson's Paradox

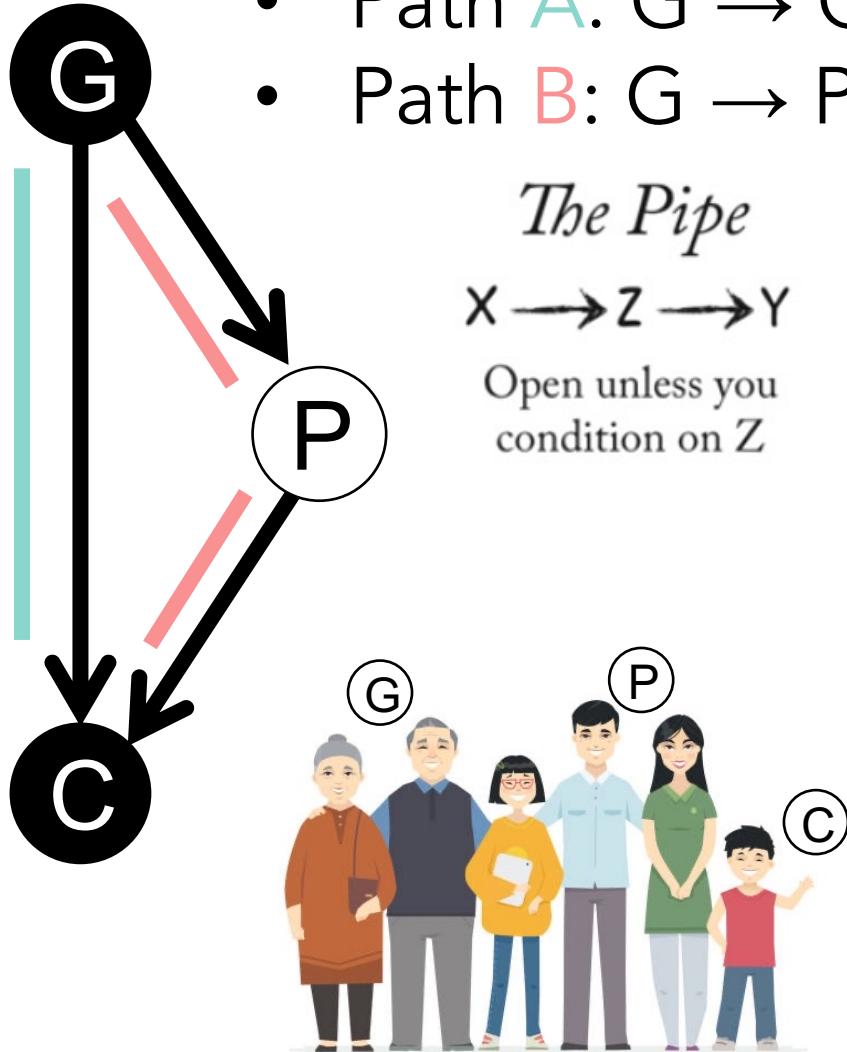
Korrelation:



Example 3: Does your grandparent's education level affect your own?

Two paths:

- Path A: G → C
- Path B: G → P → C

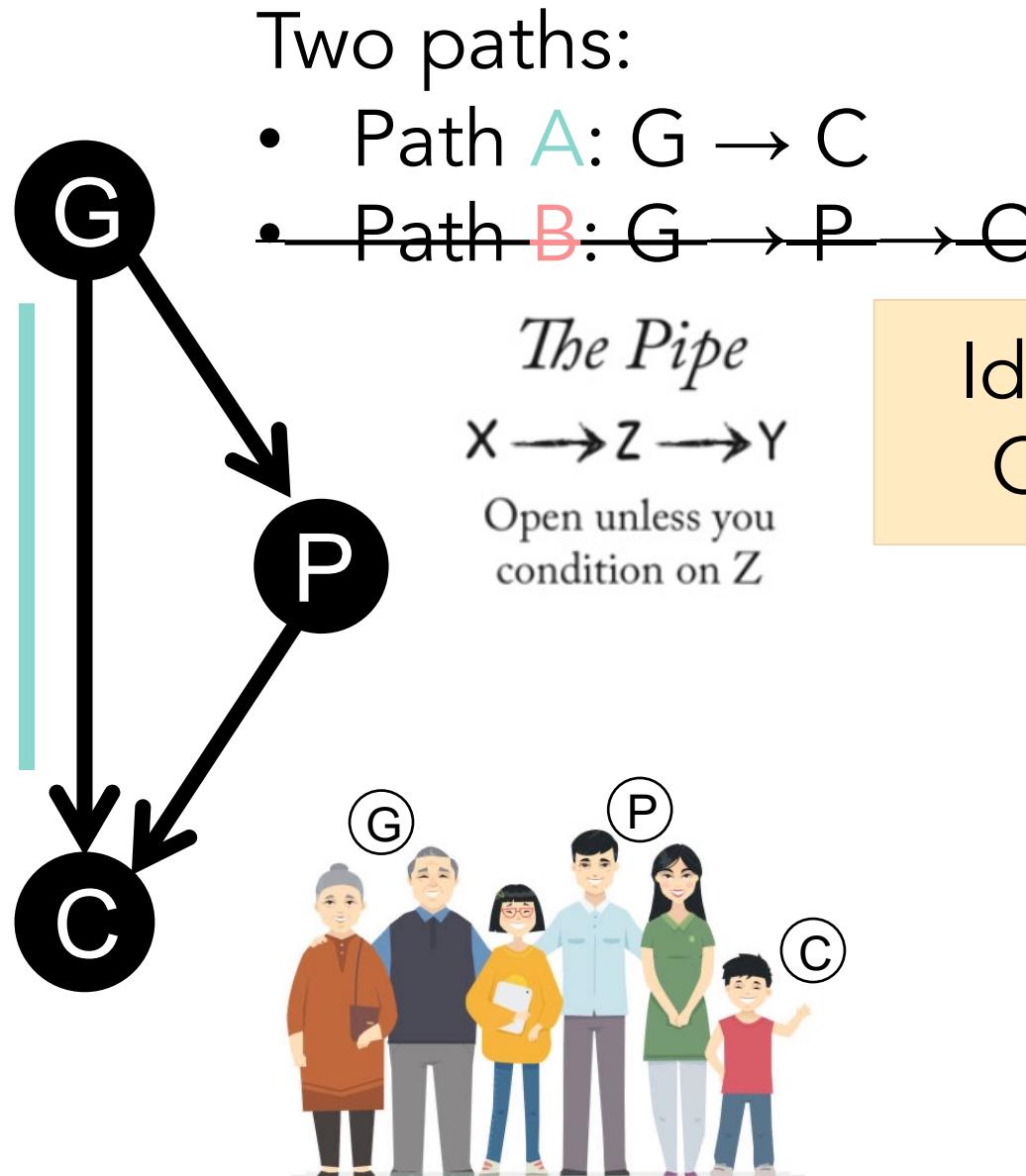


The Pipe

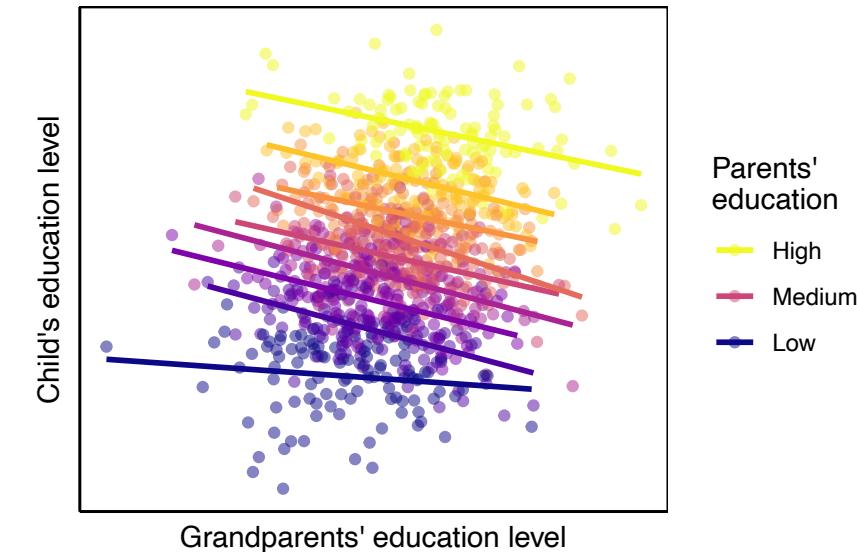
X → Z → Y

Open unless you
condition on Z

Example 3: Does your grandparent's education level affect your own?



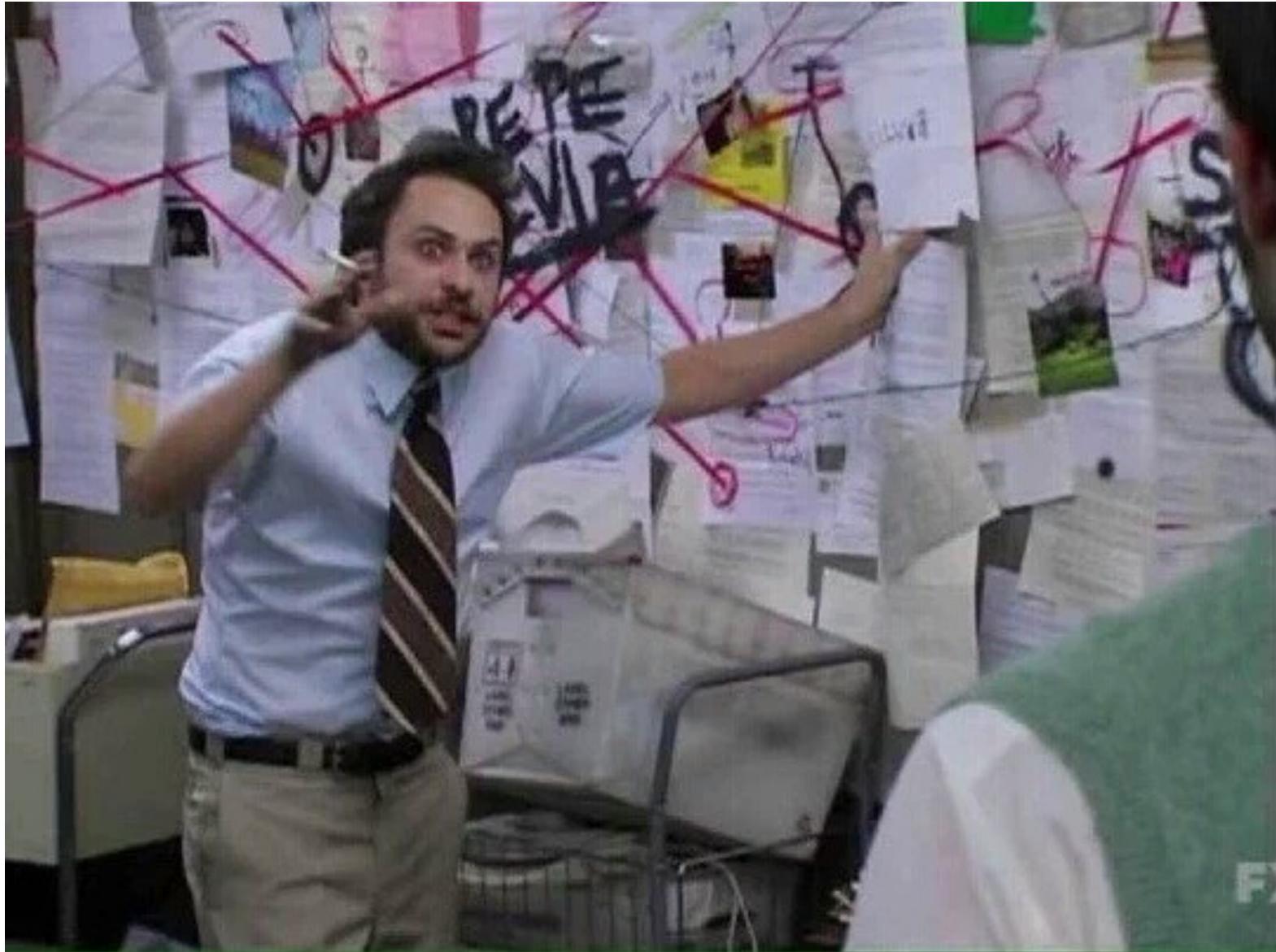
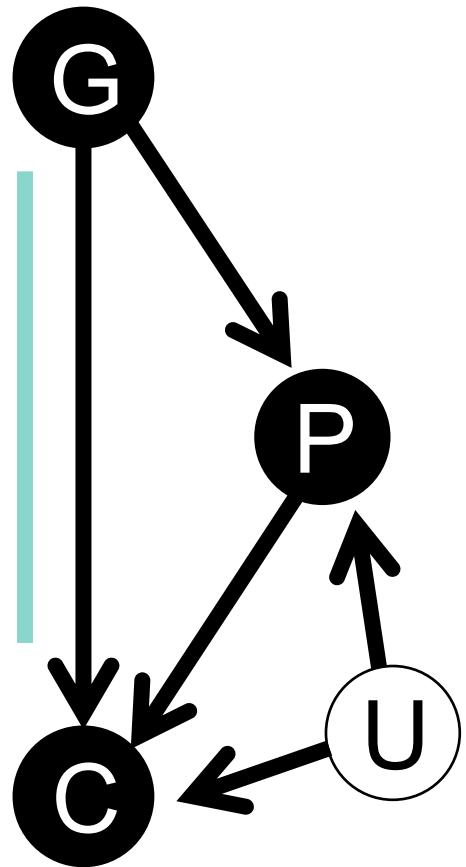
Ideal model
 $C \sim G + P$



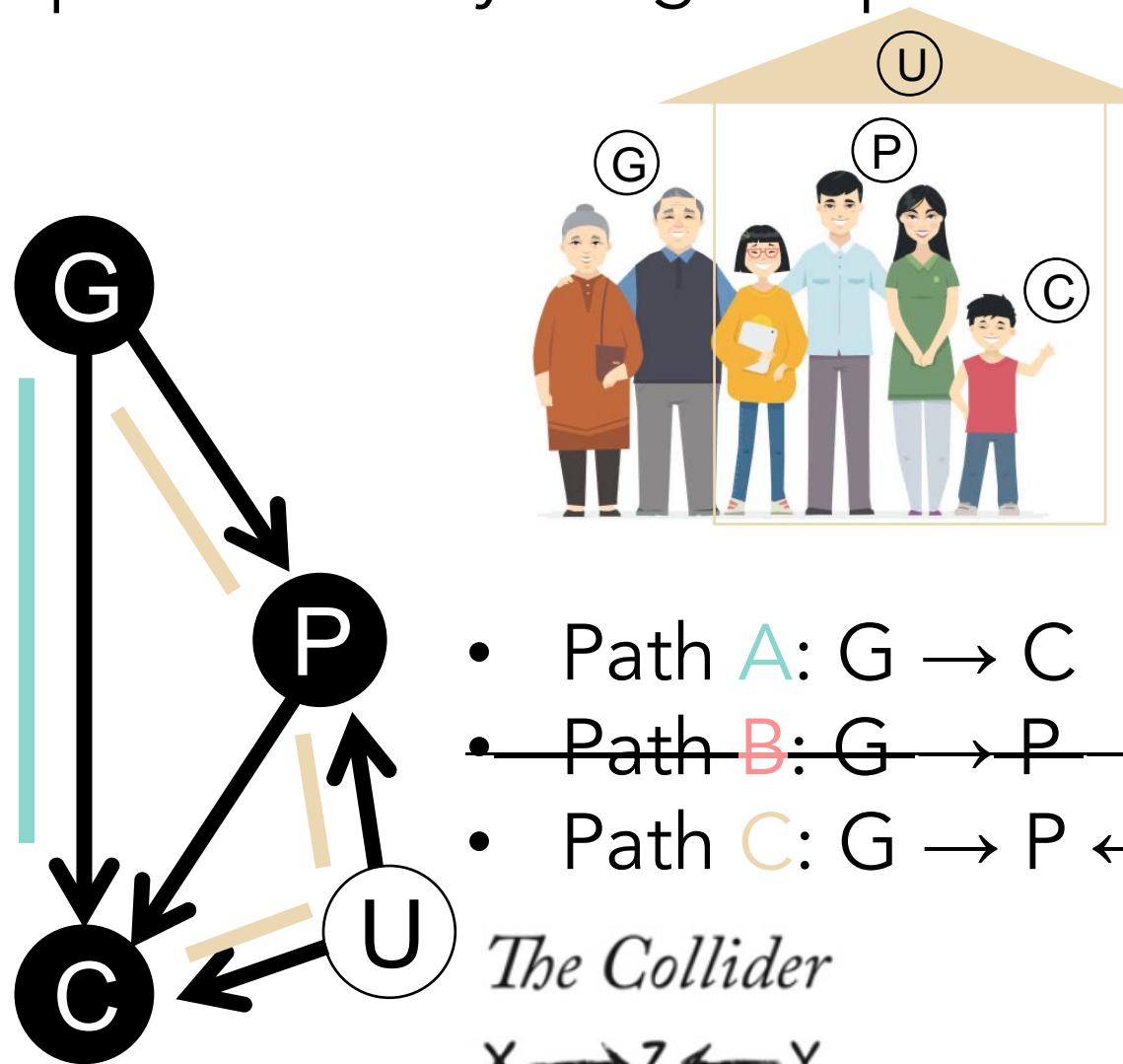
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.09936	0.06632	-1.498	0.134
G	-1.70282	0.07160	-23.783	<2e-16 ***
P	2.67736	0.02937	91.150	<2e-16 ***

Wait – it's still a negative relationship??

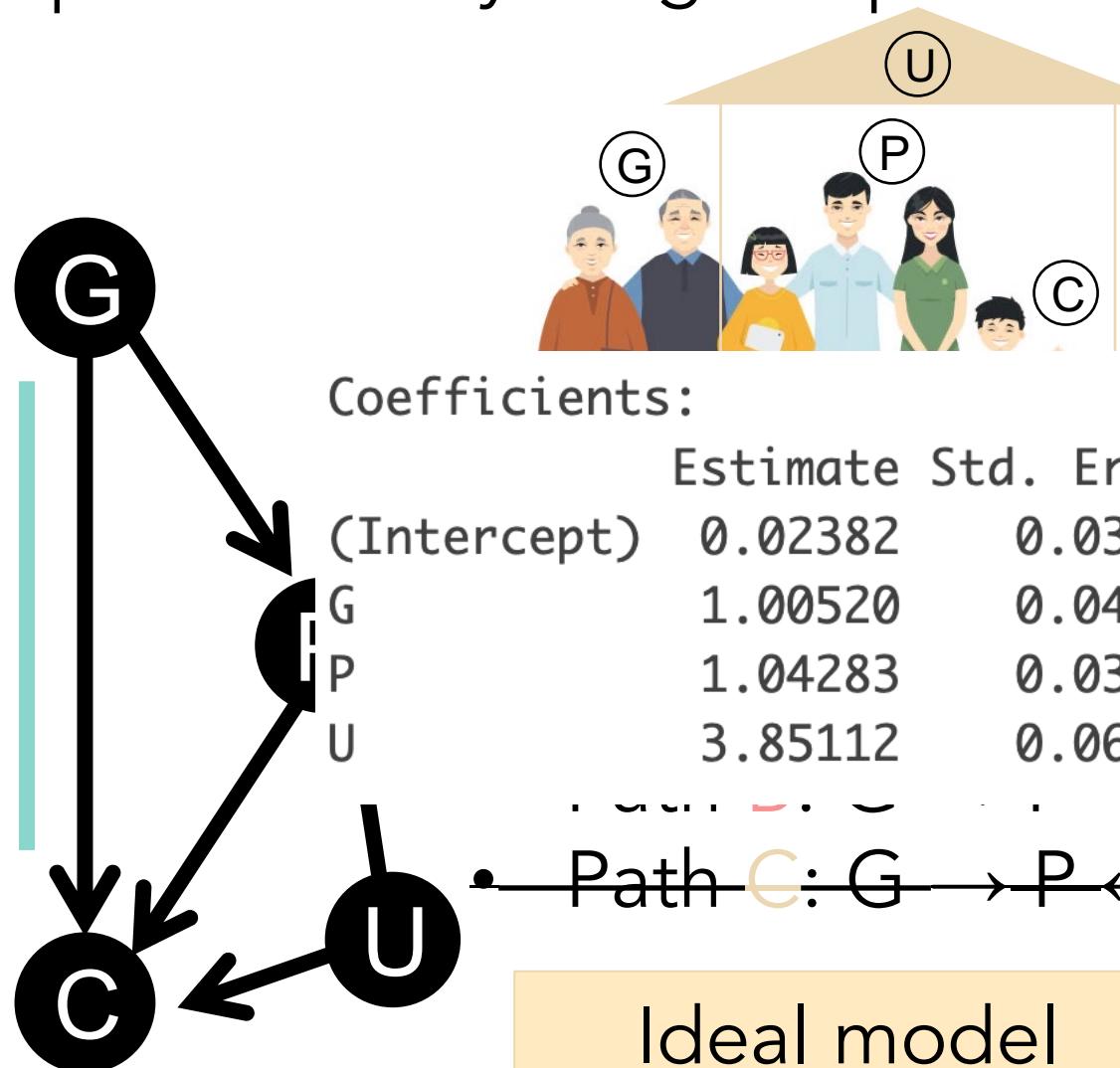


Example 3: Does your grandparent's education level affect your own?



Closed until you
condition on Z

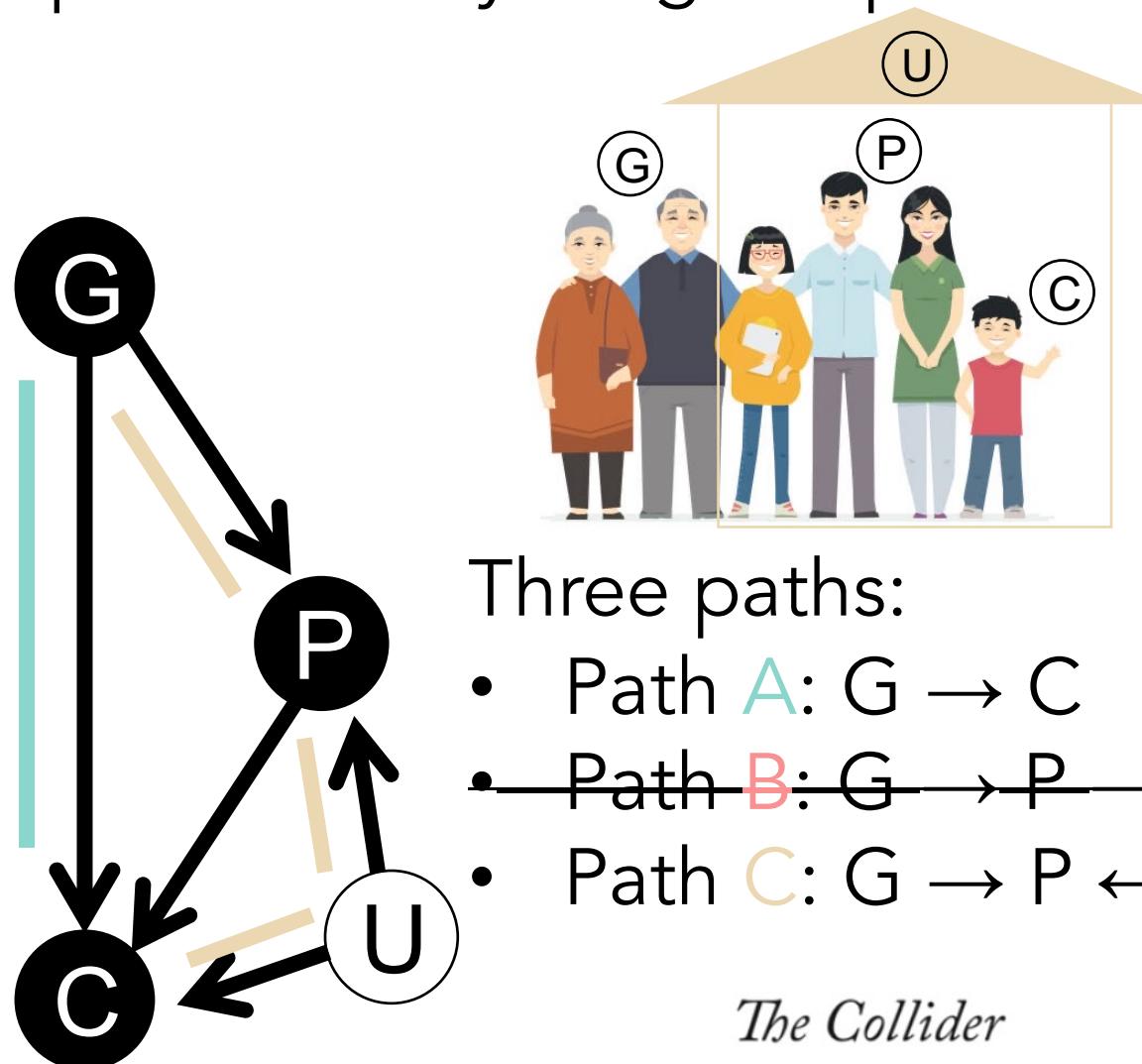
Example 3: Does your grandparent's education level affect your own?



Ideal model
 $C \sim P + C + U$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.02862	0.03217	0.890	0.374
G	0.03016	0.04438	0.679	0.497
P	0.99272	0.03178	31.241	<2e-16 ***
U	4.04779	0.06930	58.412	<2e-16 ***

Example 3: Does your grandparent's education level affect your own?



What if we can't control for U !?

Three paths:

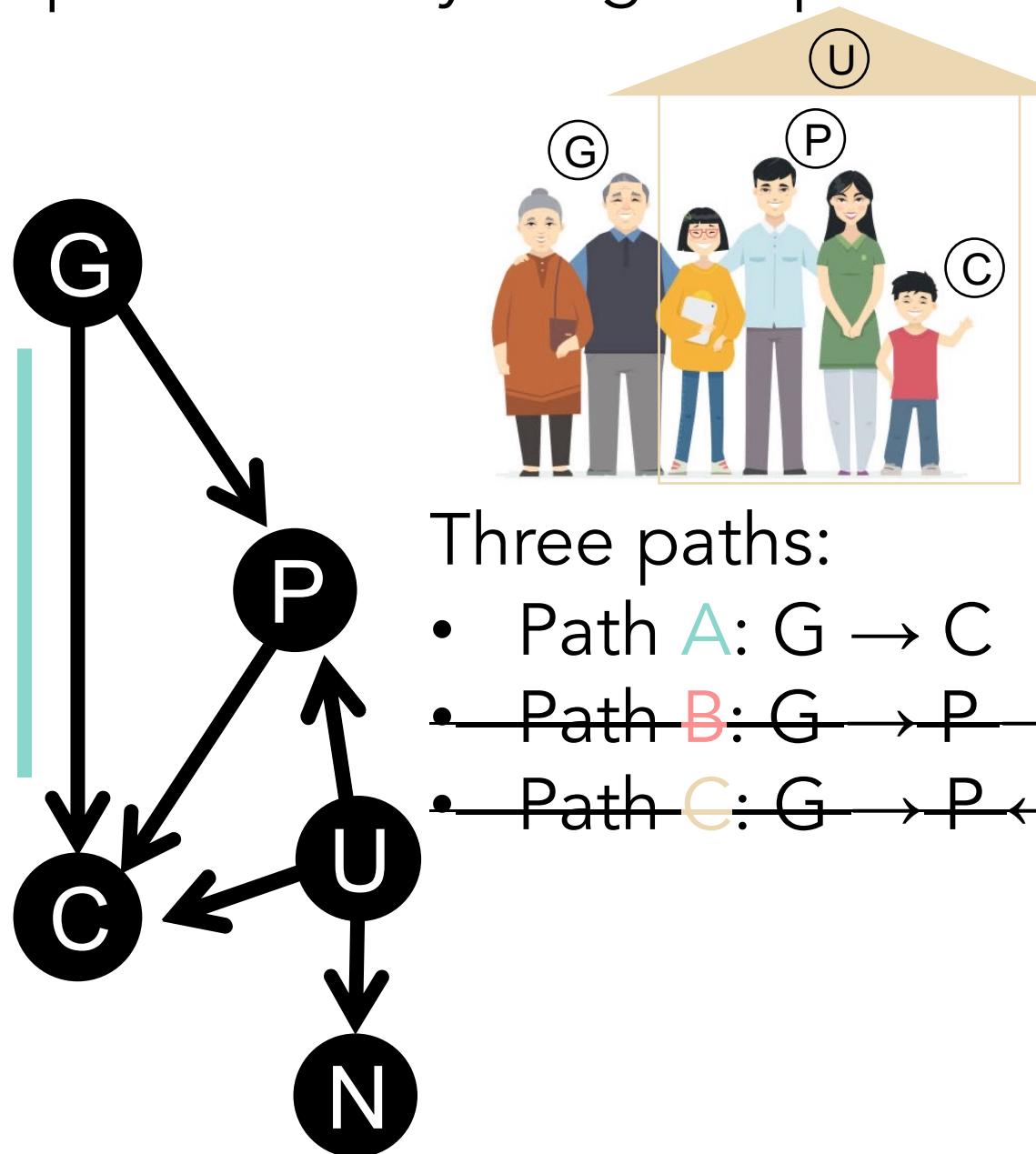
- Path A: $G \rightarrow C$
- Path B: $G \rightarrow P \rightarrow C$
- Path C: $G \rightarrow P \leftarrow U \rightarrow C$

The Collider

$X \rightarrow Z \leftarrow Y$

Closed until you
condition on Z

Example 3: Does your grandparent's education level affect your own?



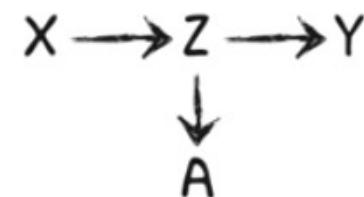
Three paths:

- Path A: $G \rightarrow C$
- Path B: $G \rightarrow P \rightarrow C$
- Path C: $G \rightarrow P \leftarrow U \rightarrow C$

What if we can't control for U !?

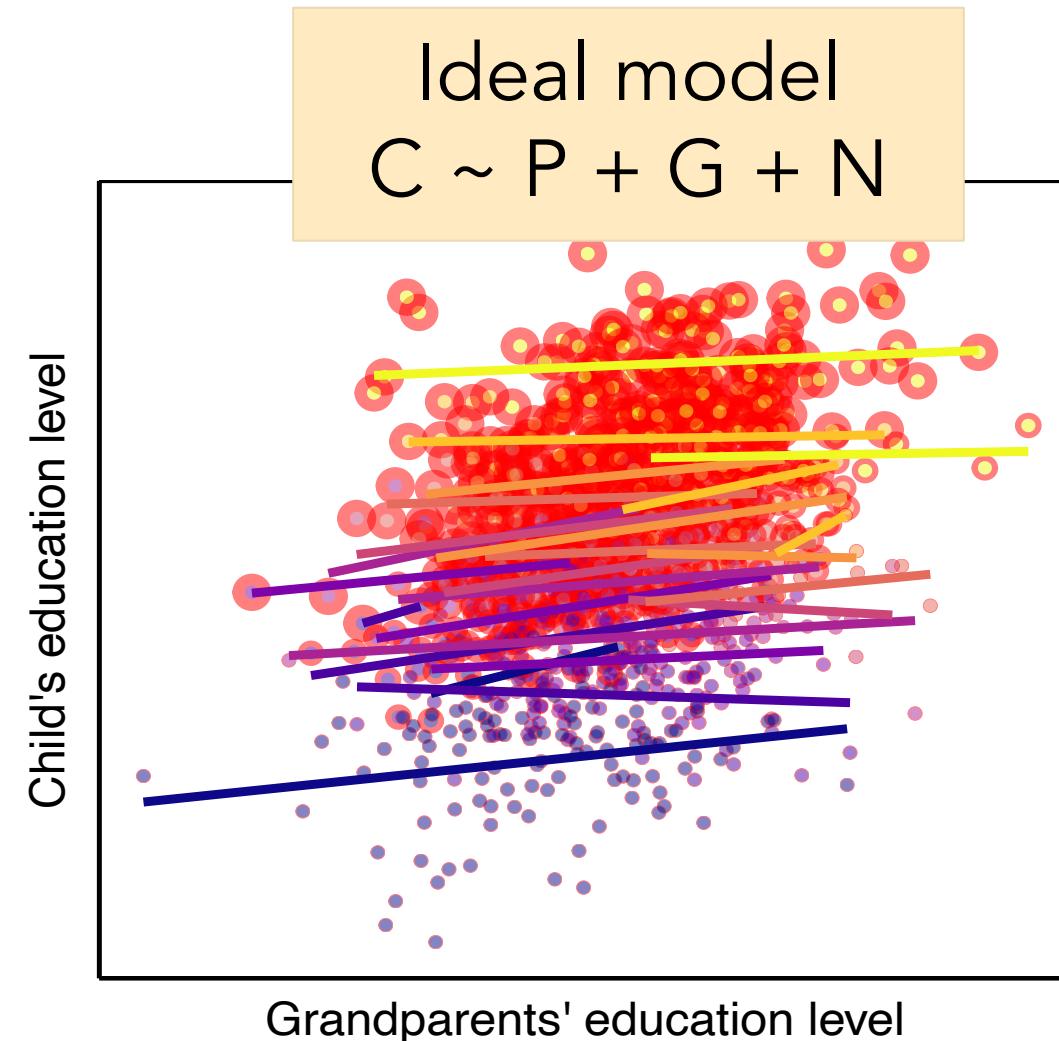
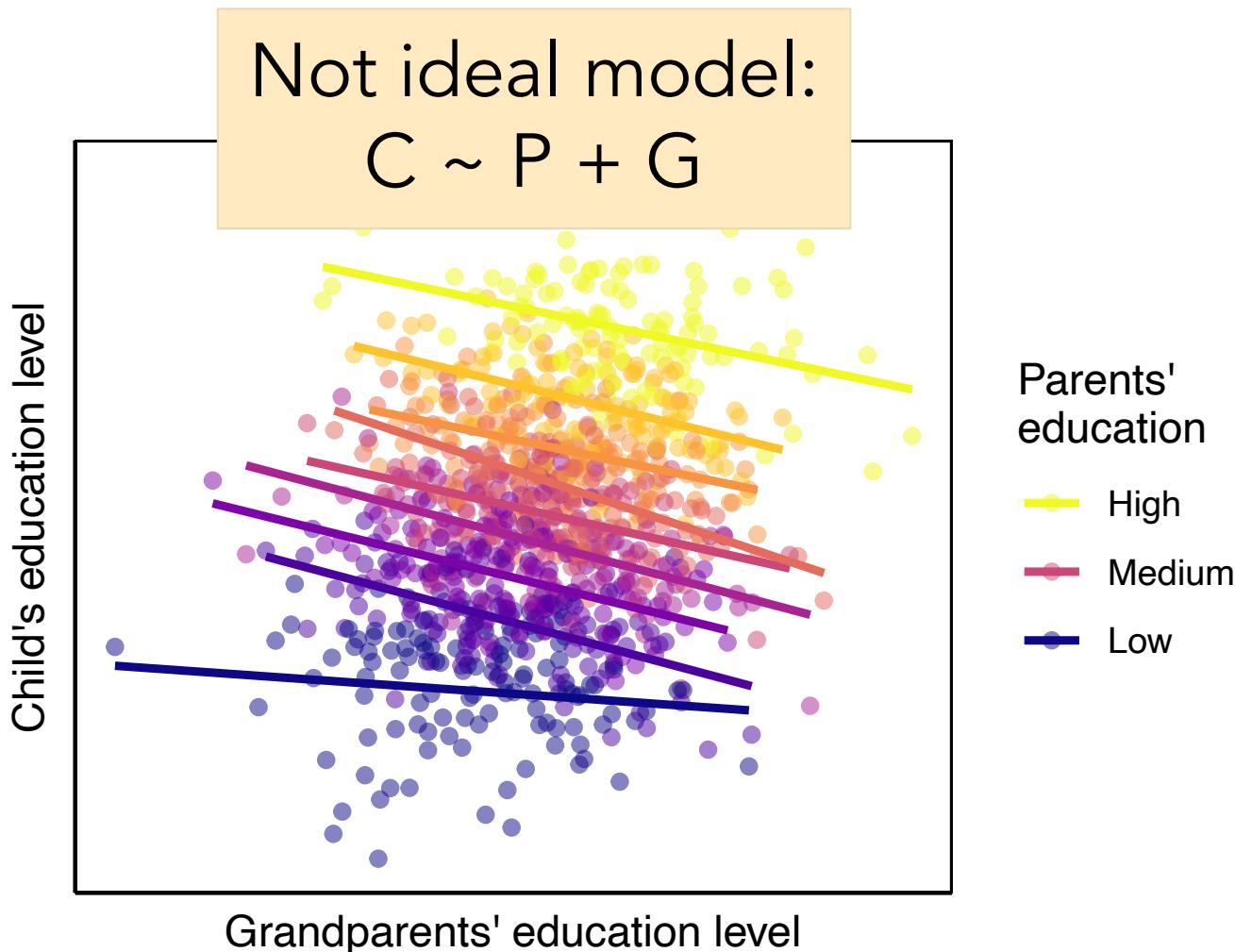
Use a *descendant* of U
e.g., control for
neighbourhood quality, N

The Descendant



Conditioning on A is
like conditioning on Z

Example 3: Does your grandparent's education level affect your own?



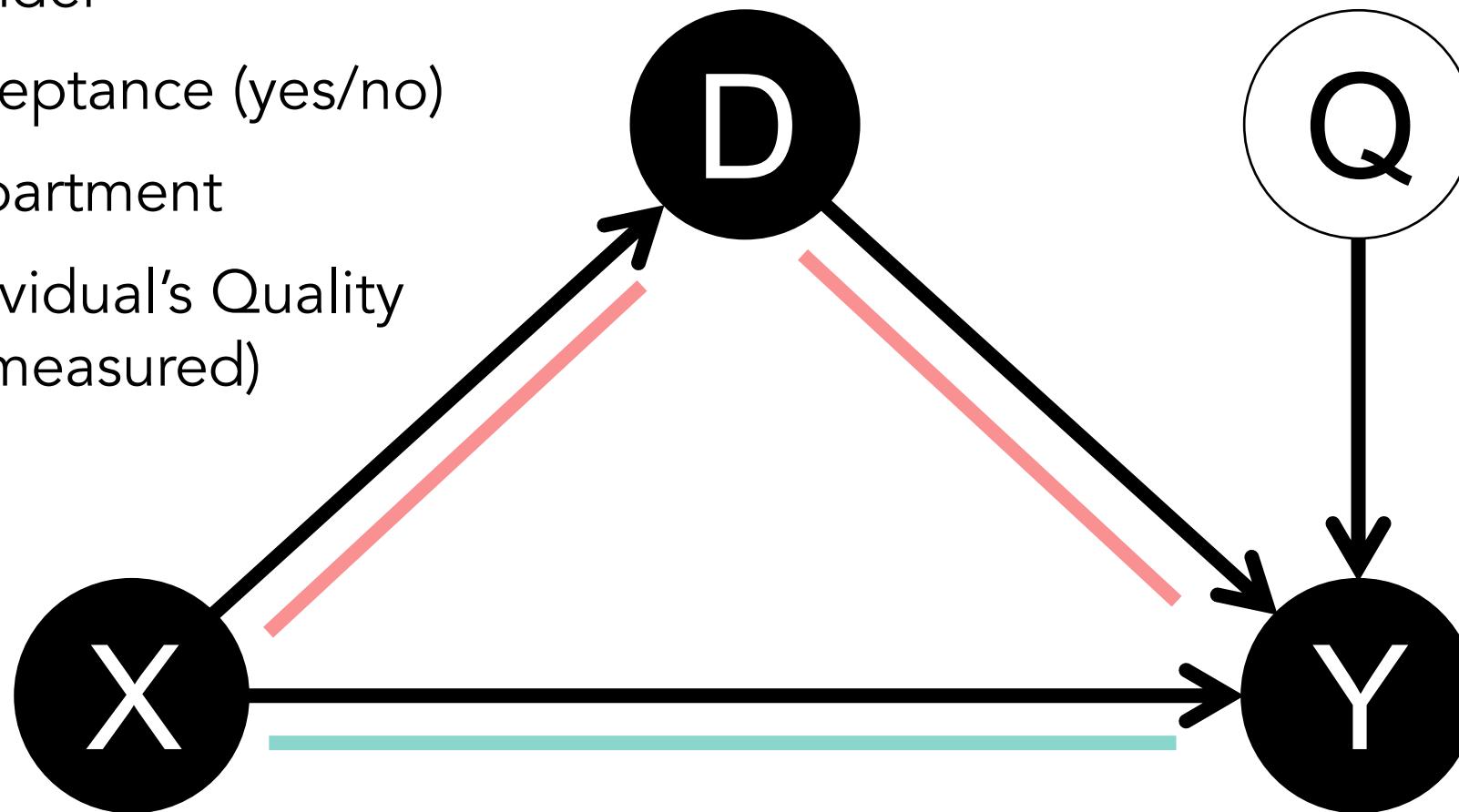


Chapter 6: Many confounds... (Skip??)

Hidden paths are EVERYWHERE in the real world!

e.g., UC Berkeley gender bias

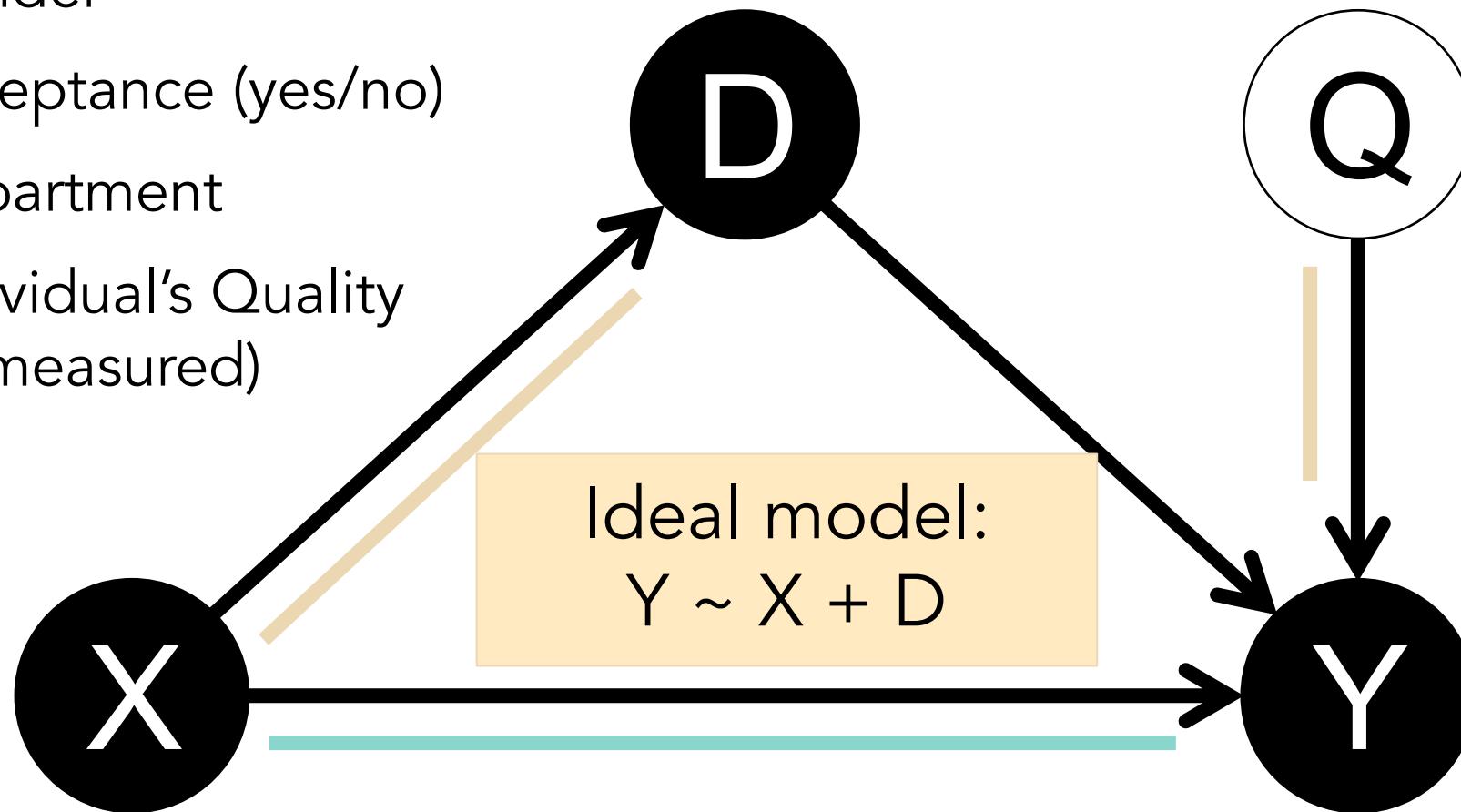
- (X) Gender
- (Y) Acceptance (yes/no)
- (D) Department
- (Q) Individual's Quality (unmeasured)



Hidden paths are EVERYWHERE in the real world!

e.g., UC Berkeley gender bias

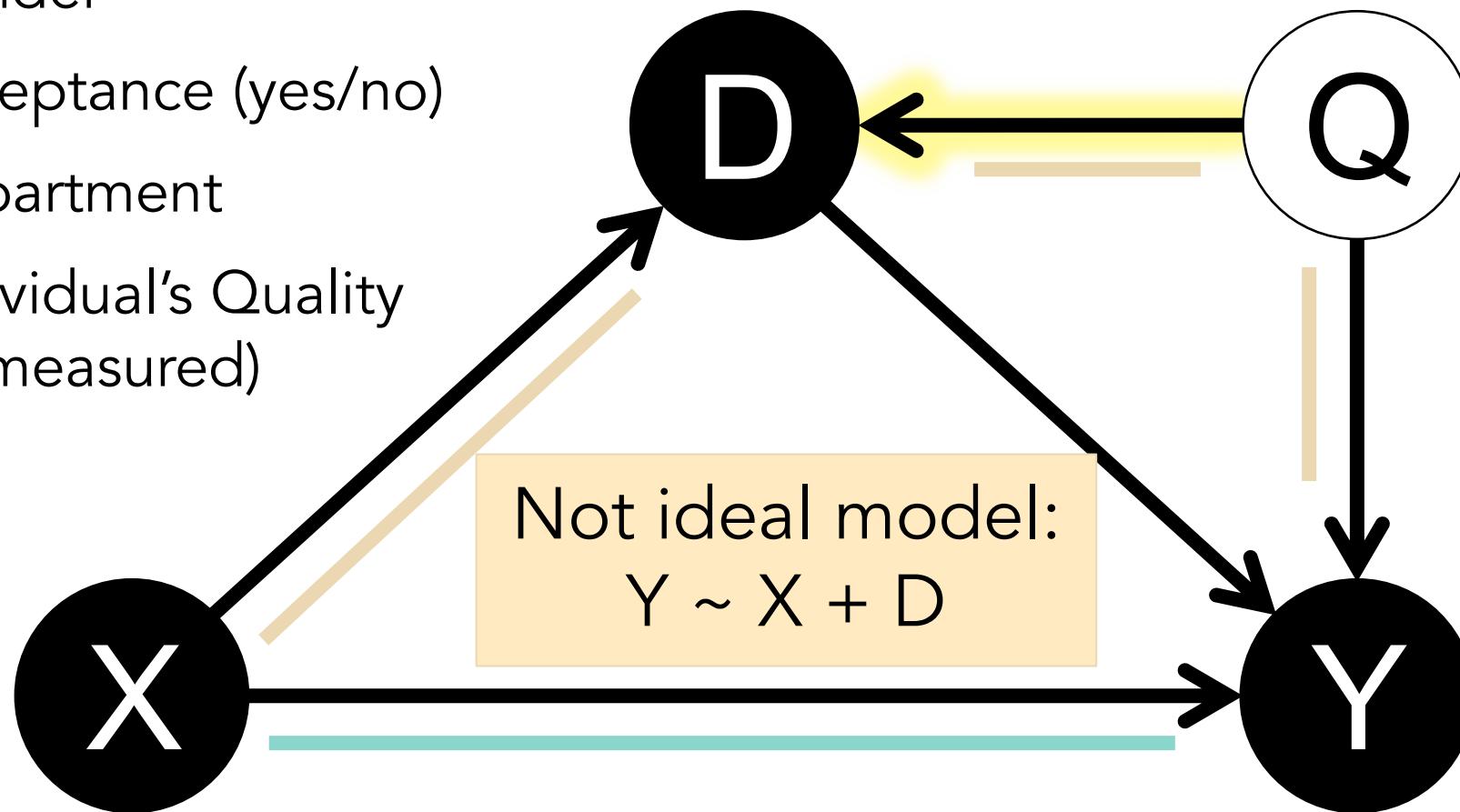
- (X) Gender
- (Y) Acceptance (yes/no)
- (D) Department
- (Q) Individual's Quality (unmeasured)



Hidden paths are EVERYWHERE in the real world!

e.g., UC Berkeley gender bias

- (X) Gender
- (Y) Acceptance (yes/no)
- (D) Department
- (Q) Individual's Quality (unmeasured)



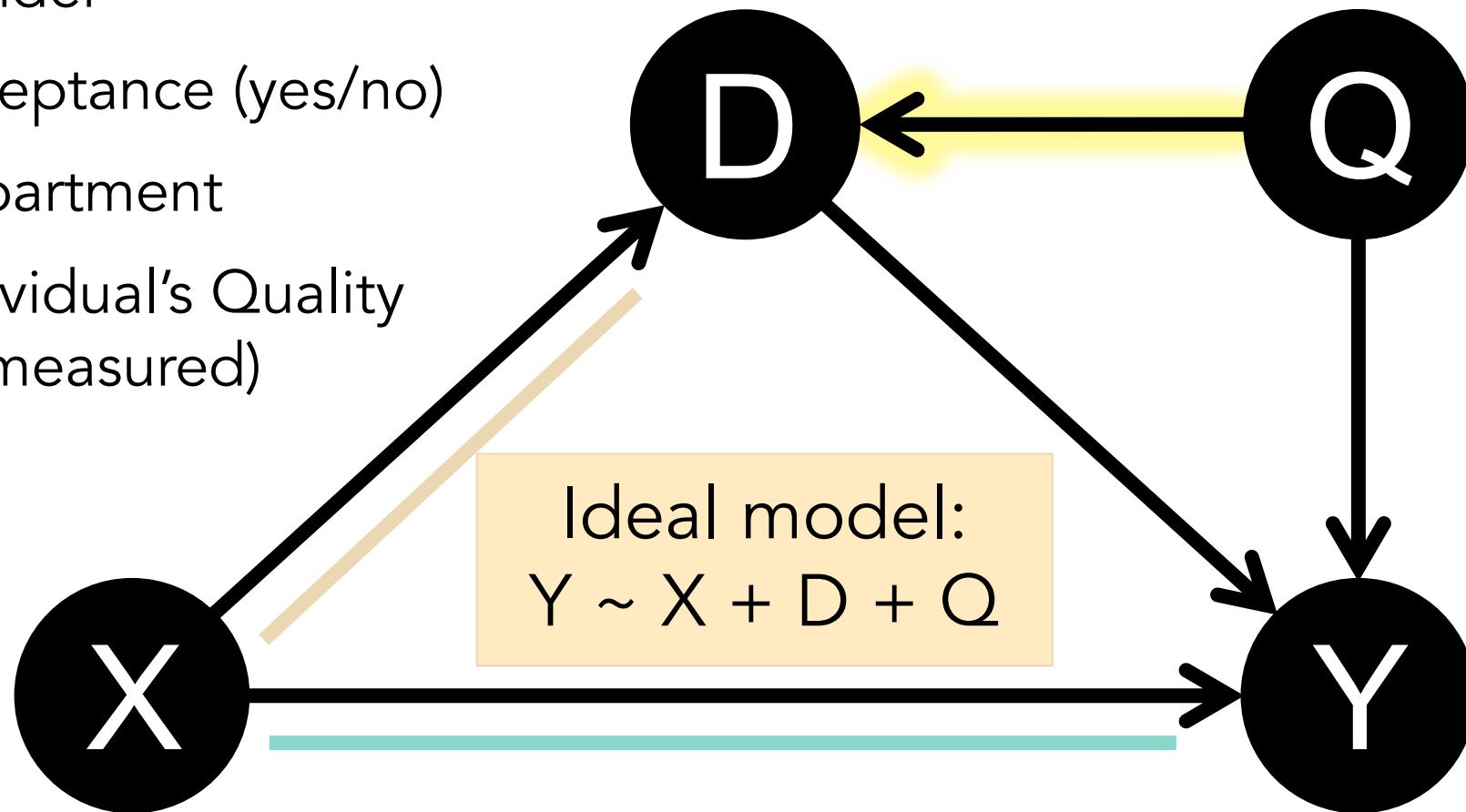
IF quality of individual affects likelihood to apply...

-> False gender bias created by conditioning on Department

Hidden paths are EVERYWHERE in the real world!

e.g., UC Berkeley gender bias

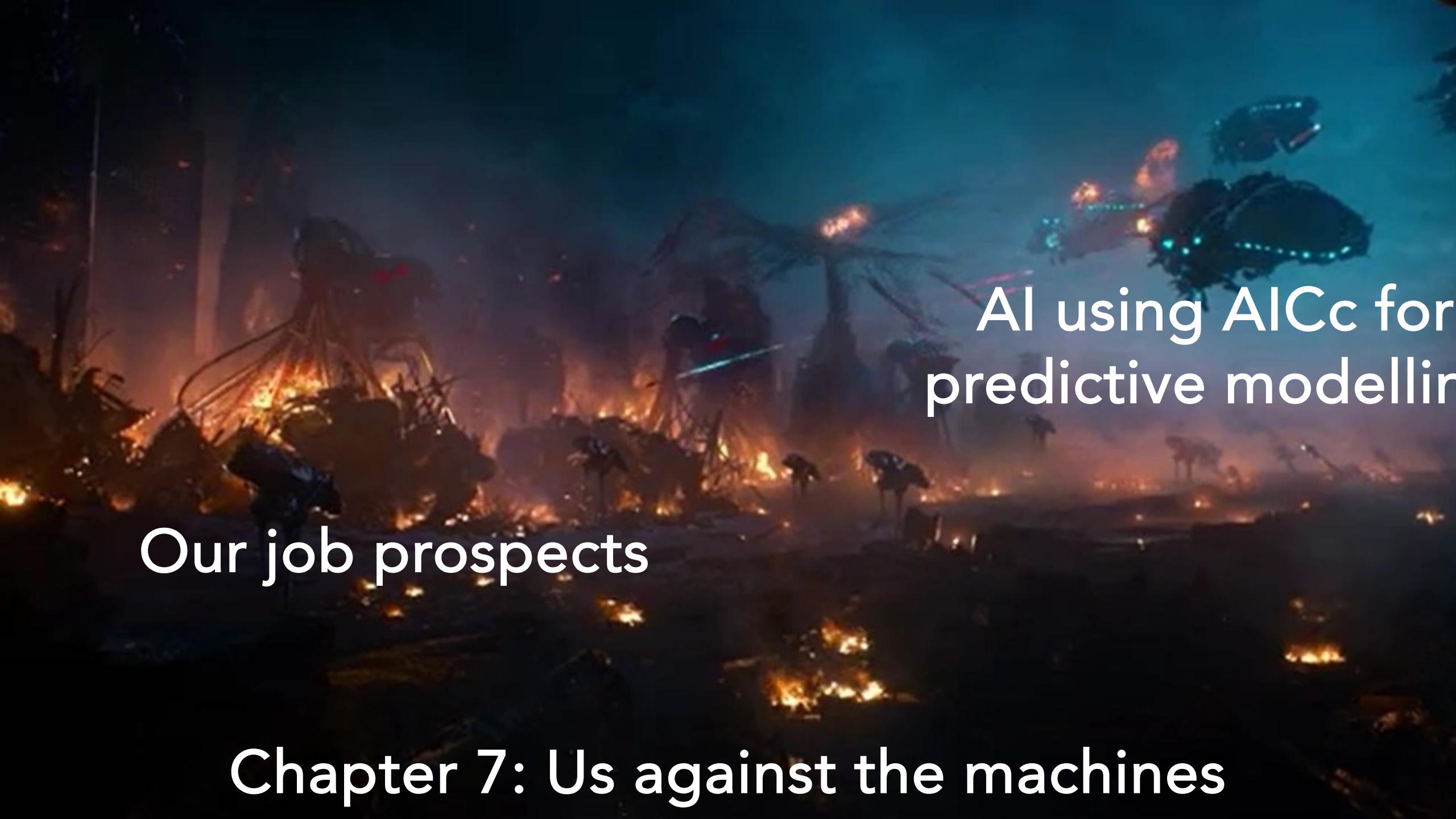
- (X) Gender
- (Y) Acceptance (yes/no)
- (D) Department
- (Q) Individual's Quality (unmeasured)



Must condition
on individual's
Quality as well!

But how do we
measure???

-> False gender
bias created by
conditioning on
Department

The background of the slide is a dark, atmospheric scene from a science fiction movie. It depicts a landscape engulfed in flames and smoke, with several large, dark, winged robotic creatures flying through the air. The overall mood is one of destruction and chaos.

Our job prospects

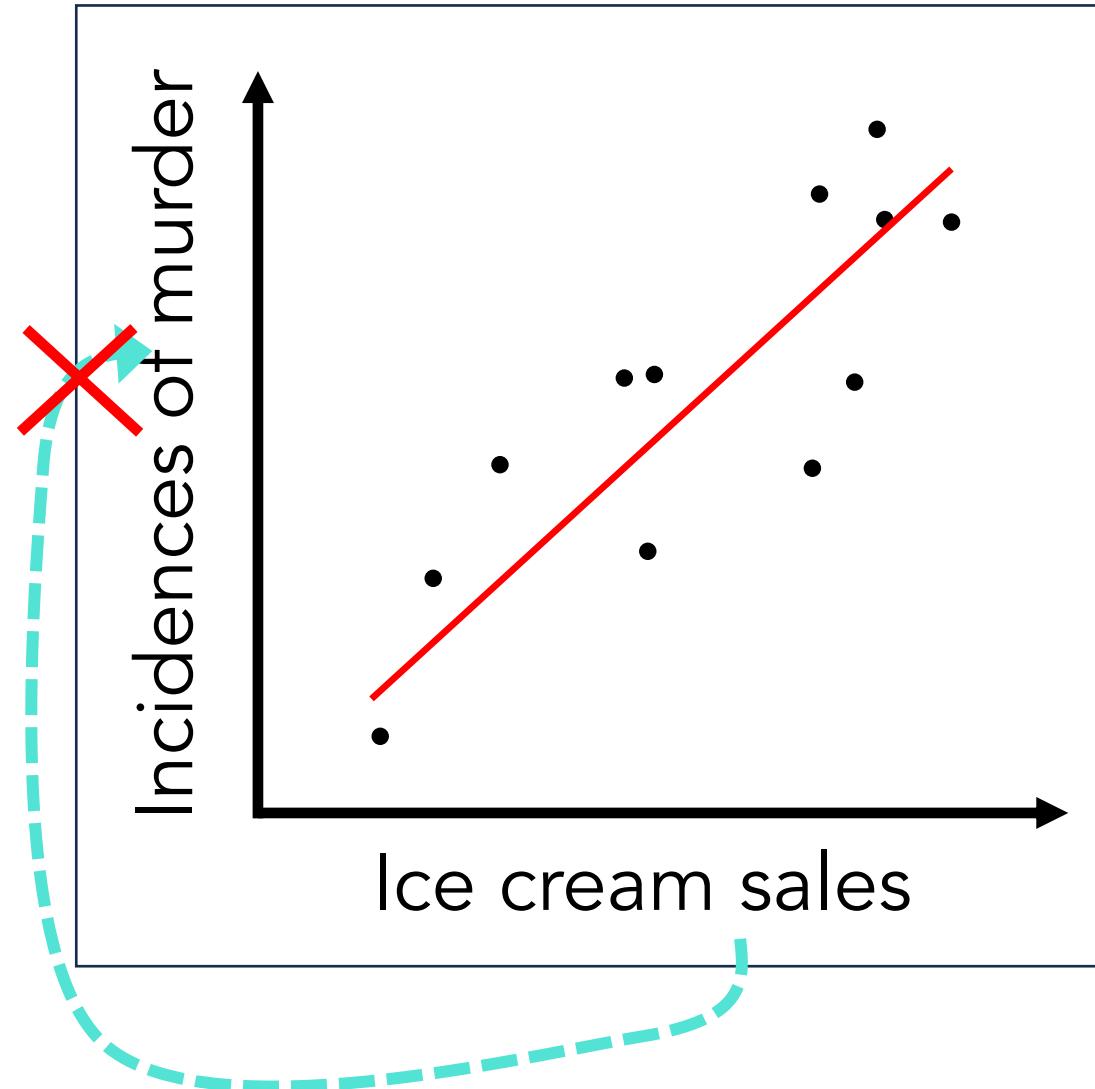
AI using AICc for
predictive modellin

Chapter 7: Us against the machines

Don't give up hope!

- Use experimental design and randomization to exclude confounds' influence(s)
- Use DAGs to form causal assumptions about your system!
 - Use R packages *ggdag* + *dagitty* to plot + determine best 'adjustment set' (i.e. x-variables to choose)
 - Can also be used to test against other alternative DAGs!
- When a confound cannot be known, turn to path analysis
 - SEMs: Structural Equation Modelling
 - Bayesian Networks
 - Estimate the influence of unknown/latent parameters

“Correlation does not equal causation”
but if we understand the system well enough,
we can test causality!



Forbes

FORBES > INNOVATION

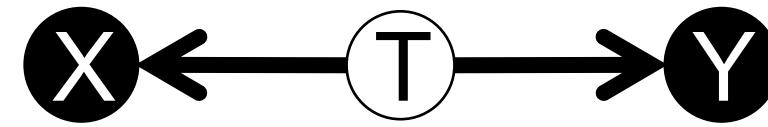
BREAKING

Here's Why Warm Weather Causes More Violent Crimes —From Mass Shootings To Aggravated Assault

Arianna Johnson Forbes Staff
I cover the latest trends in science, tech and healthcare.

Follow

The Fork
 $X \leftarrow Z \rightarrow Y$
Open unless you condition on Z

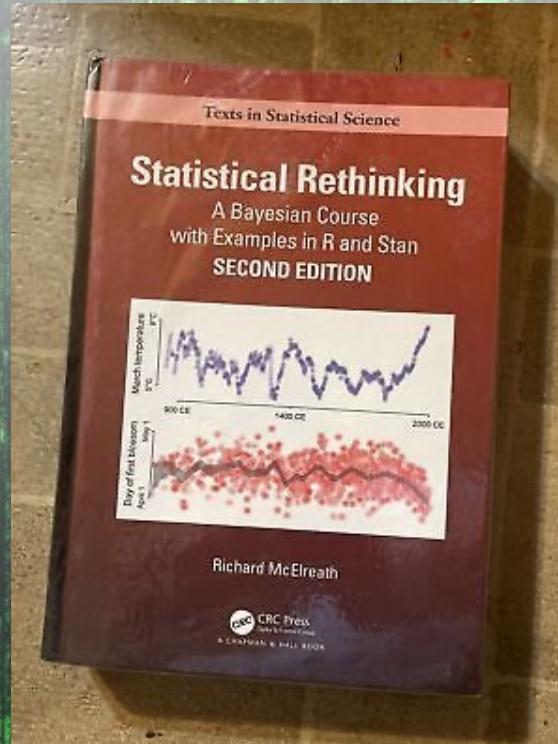
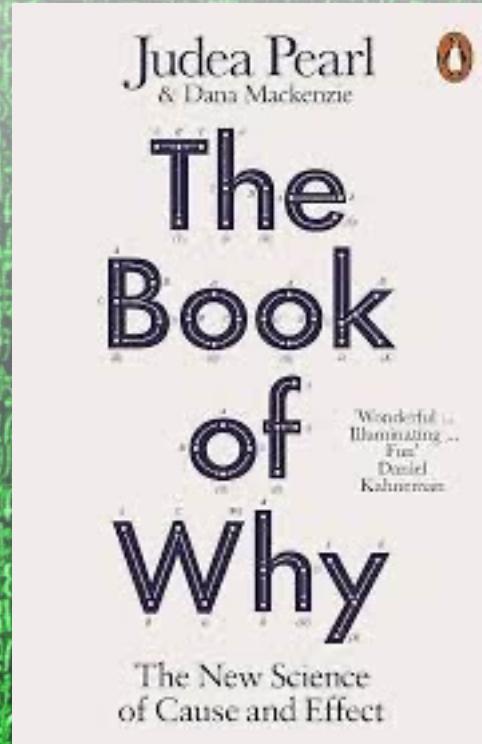


Causal modelling



Not bullet-proof, but we must strive for better!

Thanks for listening!



The image is a screenshot of a YouTube channel page for 'Richard McElreath'. The channel has 34.2K subscribers and 111 videos. The bio reads: 'Lectures, mainly for Bayesian statistics, but also professional scientific talks from time to time...' with a link. The page shows a circular profile picture of the speaker, a 'Subscribed' button, and navigation links for Home, Videos, Playlists, Community, Channels, and About. Below this, a section titled 'Created playlists' lists several playlists: 'Statistical Rethinking 2023 Music' (19 videos), 'Statistical Rethinking 2023' (20 videos), 'Statistical Rethinking 2022' (21 videos), 'Statistical Rethinking Winter 2019' (20 videos), and 'Statistical Rethinking Fall 2017' (19 videos). Each playlist thumbnail includes a video preview and the number of videos.

<https://github.com/kevolve/Causality-Presentation>

BAYES

FREQUENTISM



A dramatic scene from a Godzilla movie. On the left, a large black Godzilla is shown in profile, its mouth open as if roaring. It appears to be standing on or near a burning building. On the right, a smaller, yellowish-orange Godzilla is shown from behind, also appearing to be in motion or attacking. The background is a bright, cloudy sky at sunset or sunrise, with a warm glow. In the foreground, there are dark silhouettes of other buildings and debris, suggesting a destroyed city.

CAUSAL INFERENCE

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

$$X \rightarrow Z \rightarrow Y$$

↓
A



Instrument

Risk Factor

