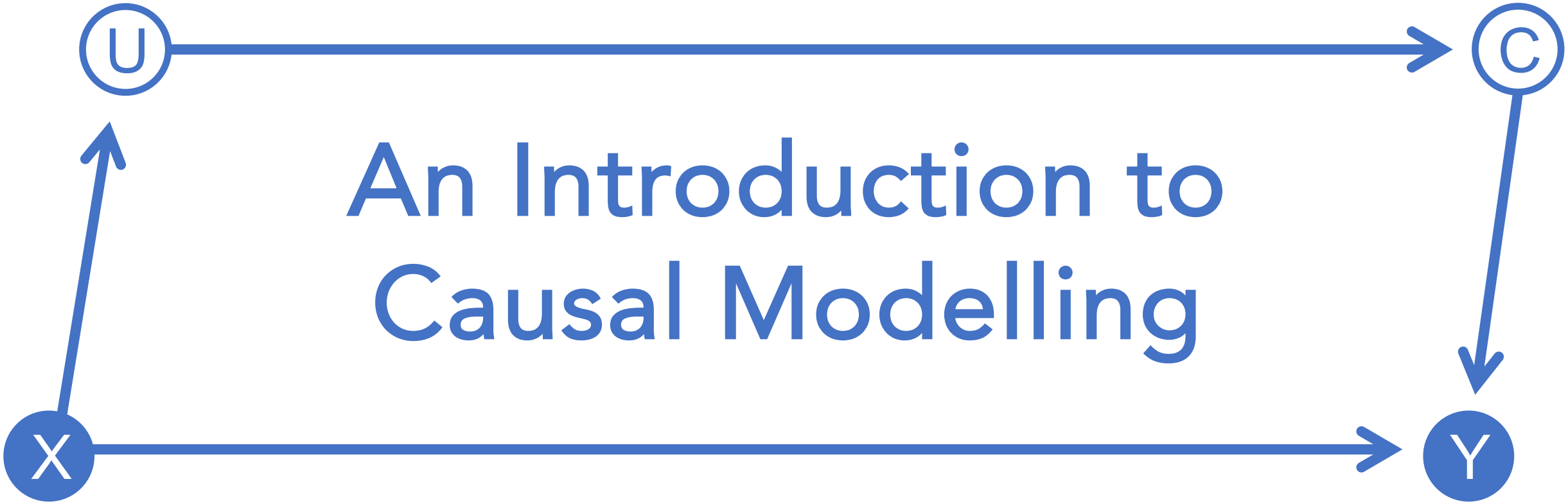
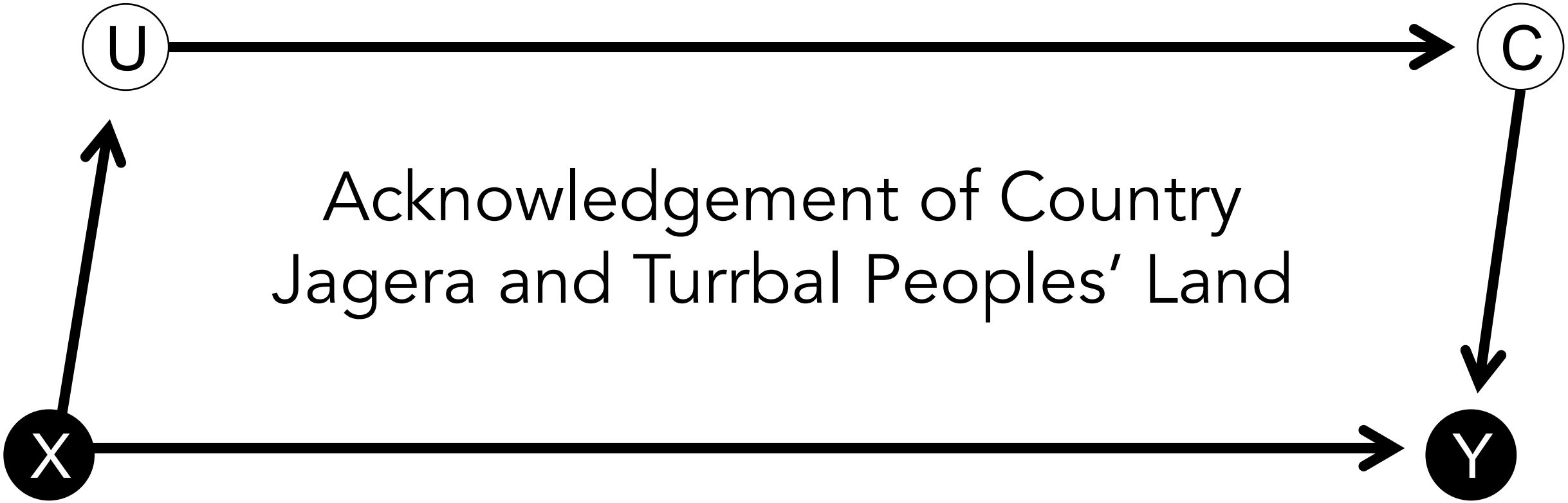


An Introduction to Causal Modelling

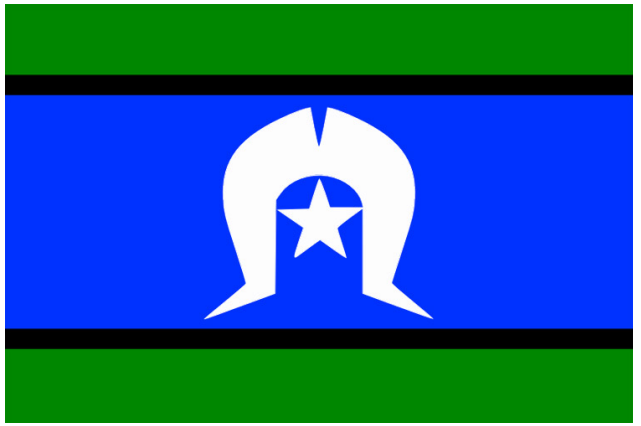
Kevin Bairos-Novak

CodeR-TSV Coding Club, 11am, 25 Oct 2021

Riginos Lab, 4pm, 3 Nov 2023



Acknowledgement of Country
Jagera and Turrbal Peoples' Land

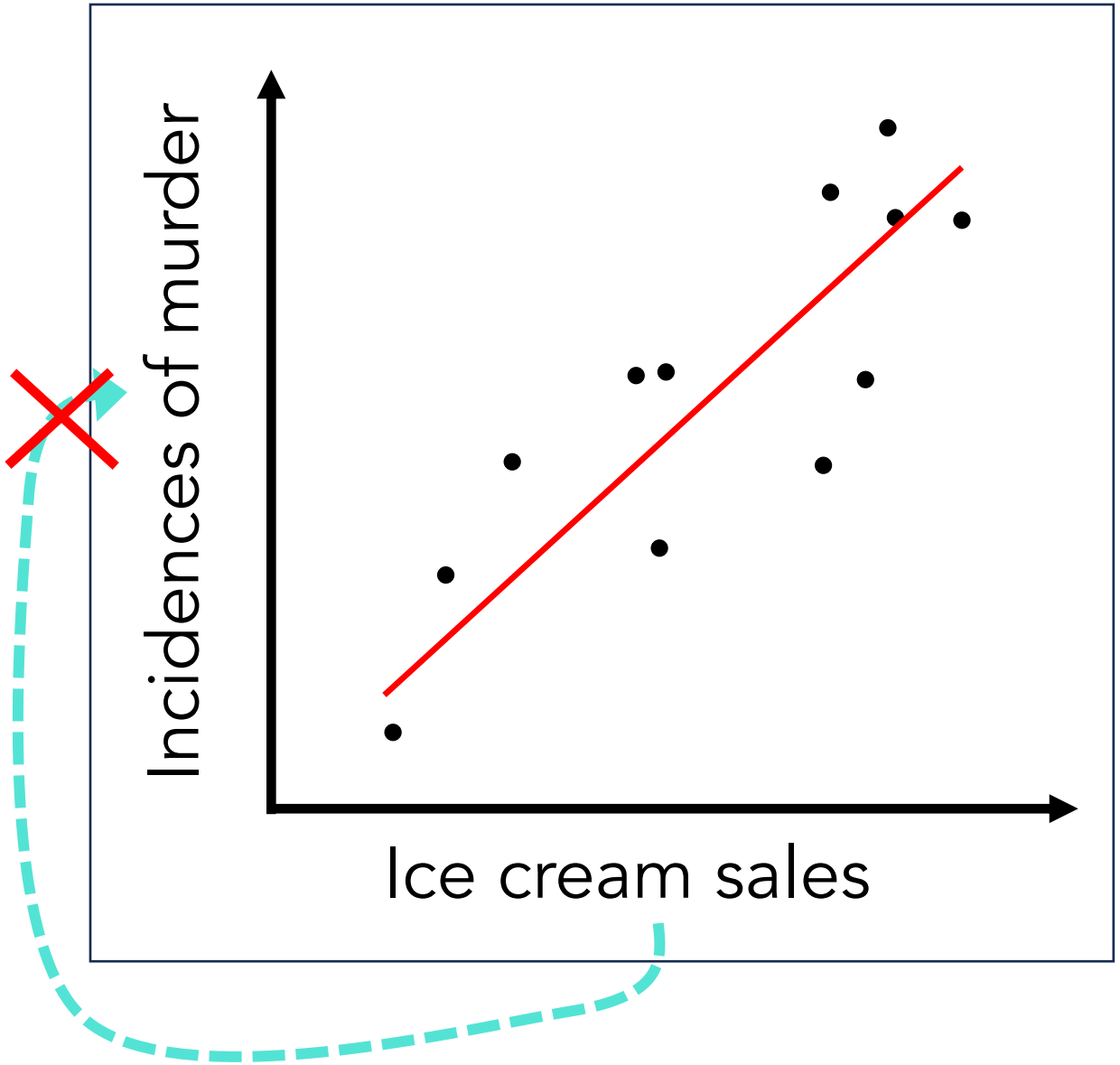



FREE YOUR MIND.
THE MATRIX

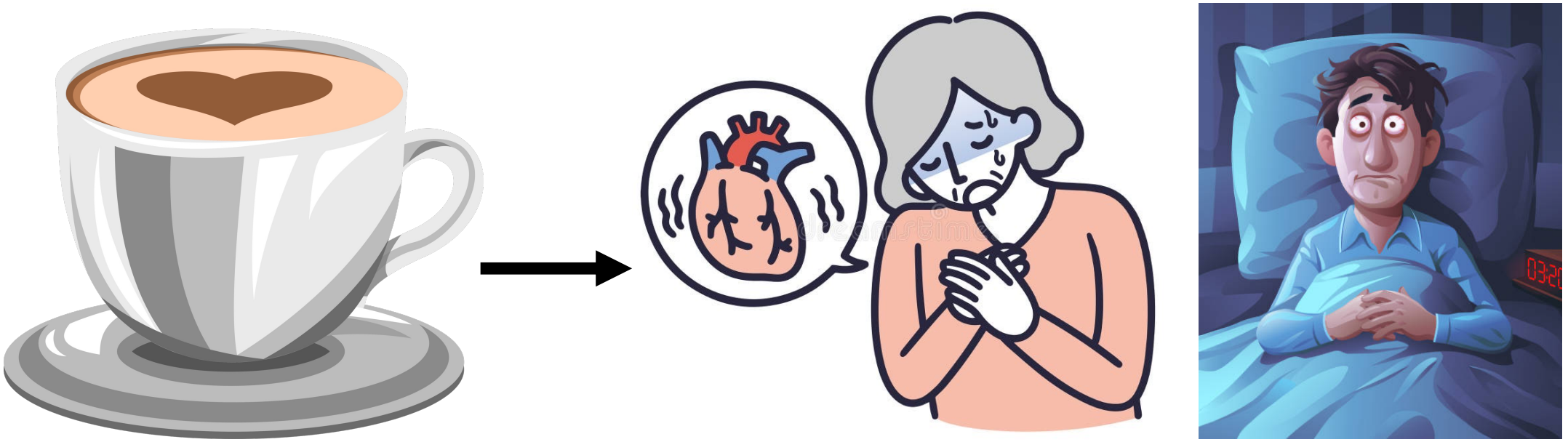

Chapter 1: We are living in a *lie*, Neo…

“Correlation does not equal causation”

Karl Pearson, inventor of the Pearson’s correlation coefficient (*r*) and longingly staring at data



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

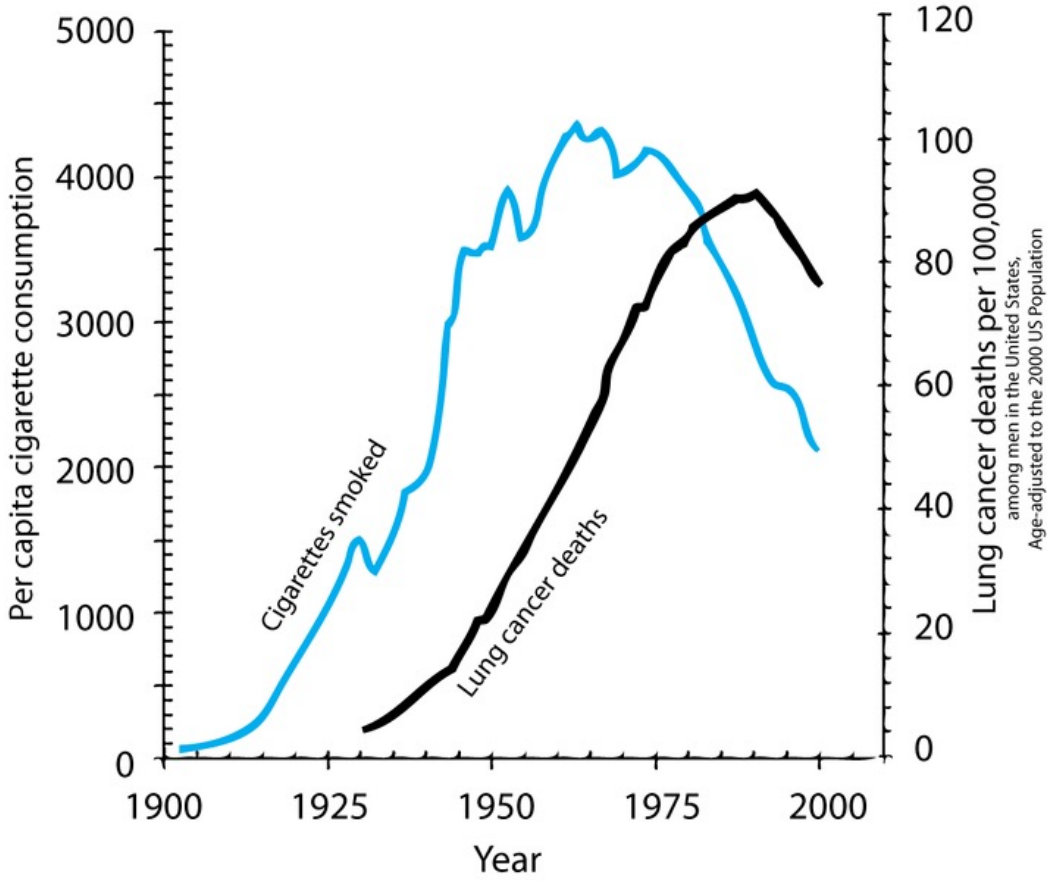


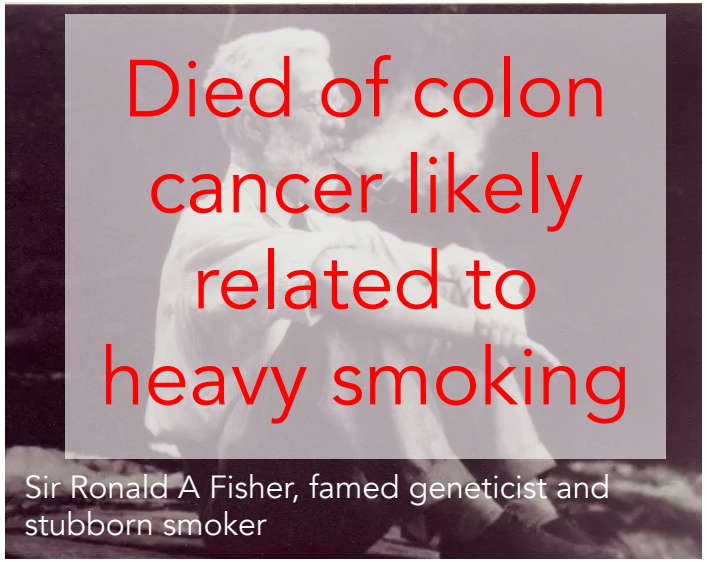
Test by manipulating number of coffees consumed!
Time to fall asleep (Y) ~ # coffees (X)

But what if we can’t manipulate coffee?

Does smoking cause cancer?

Strong correlation, but not causal?





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

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

Chapter 2: Déjà vu?
Three ‘glitches’ in the statistical matrix

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

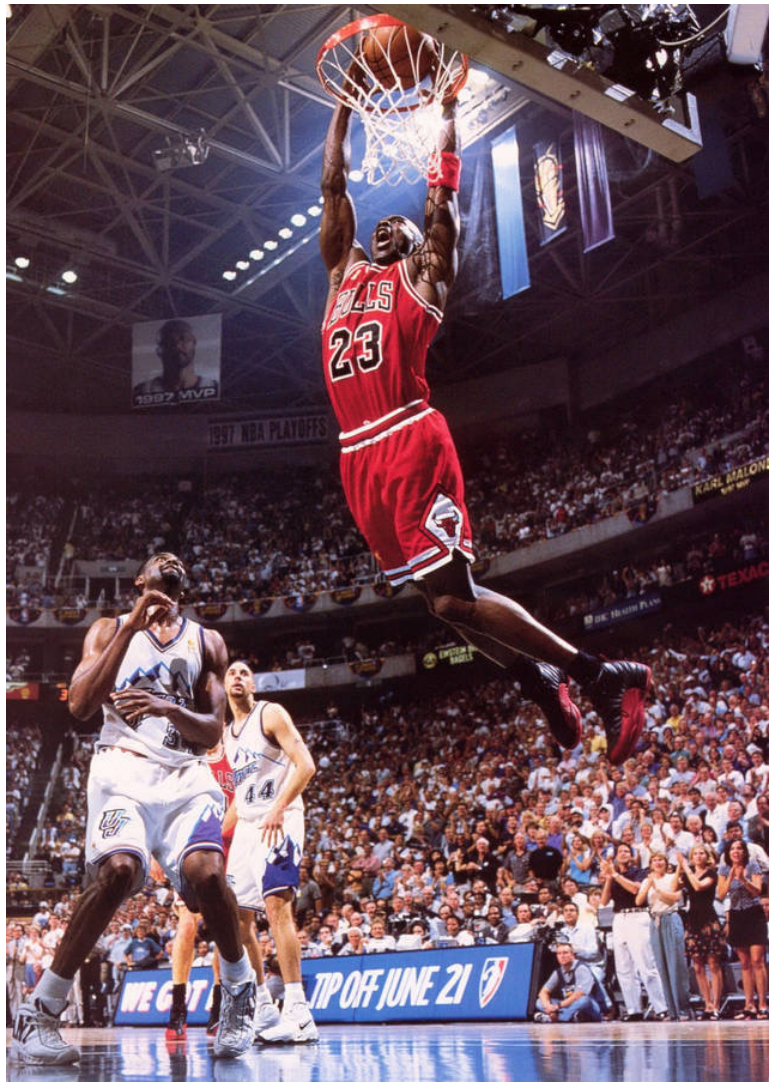
Test score (Y) ~ Sleep (X_1) + Alertness (X_2)

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

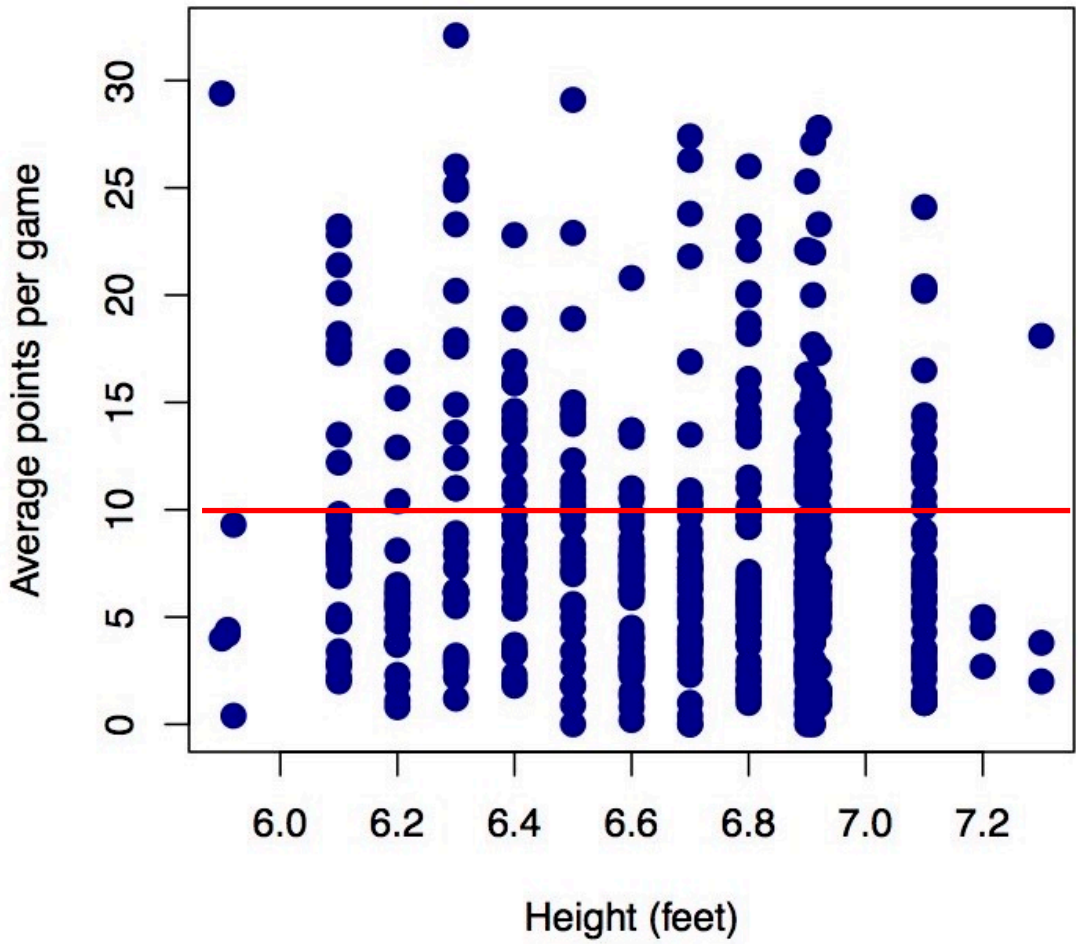
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.016811   0.032880   0.511    0.609
sleep       0.004877   0.045431   0.107    0.915
alertness   1.010407   0.031905  31.670   <2e-16 ***
---
```

No, but alertness does!

Example 2: Does basketball player height affect scoring?

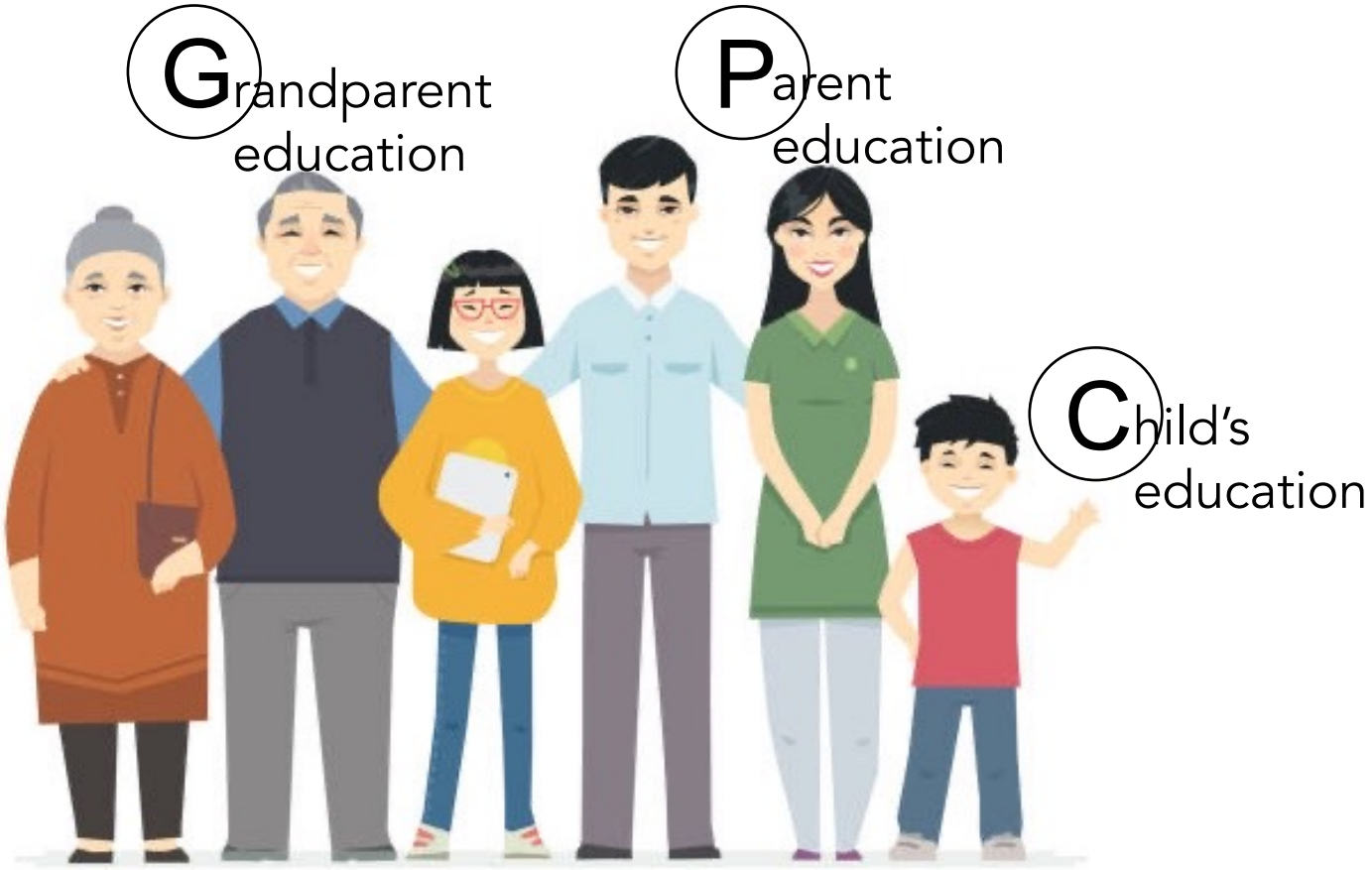


NBA player scoring (Y) ~ NBA player height (X)



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 ~ P

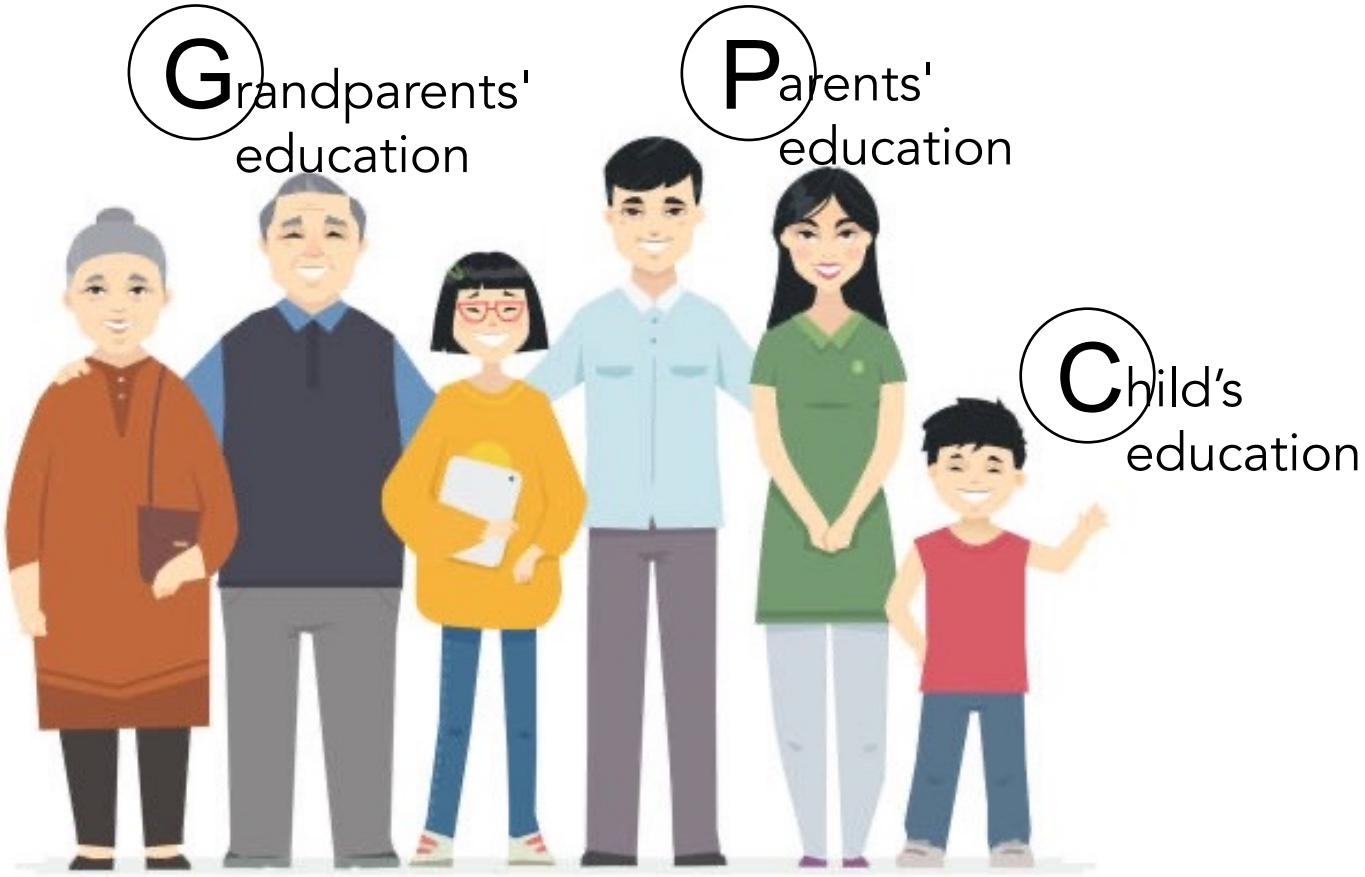
Model 2: C ~ G

Model 3: C ~ P + G

Model 4: C ~ P x G

	G	P	C
	<dbl>	<dbl>	<dbl>
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?

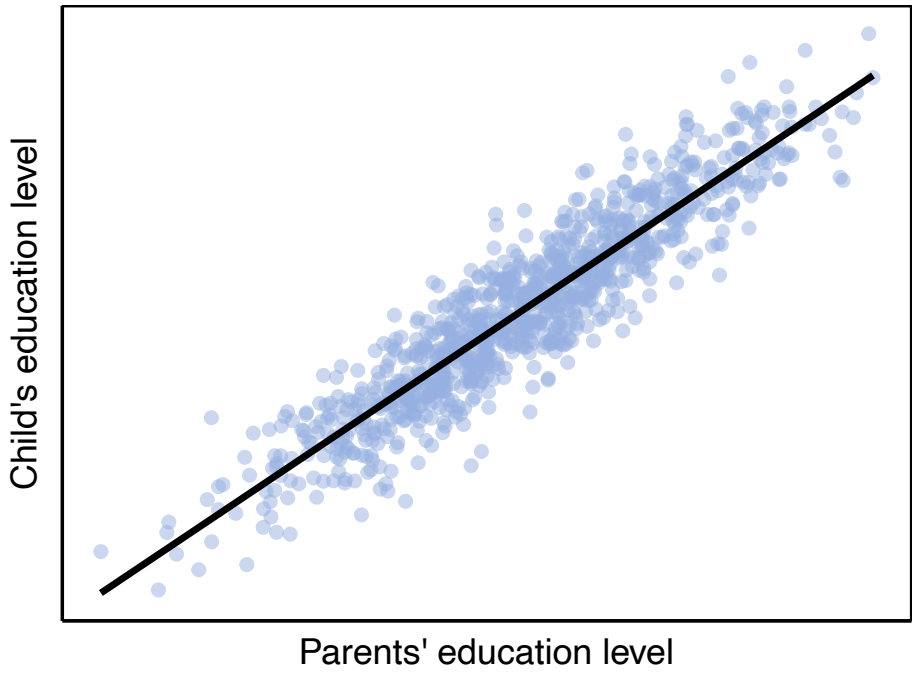


Model 1: C ~ P

Model 2: C ~ G

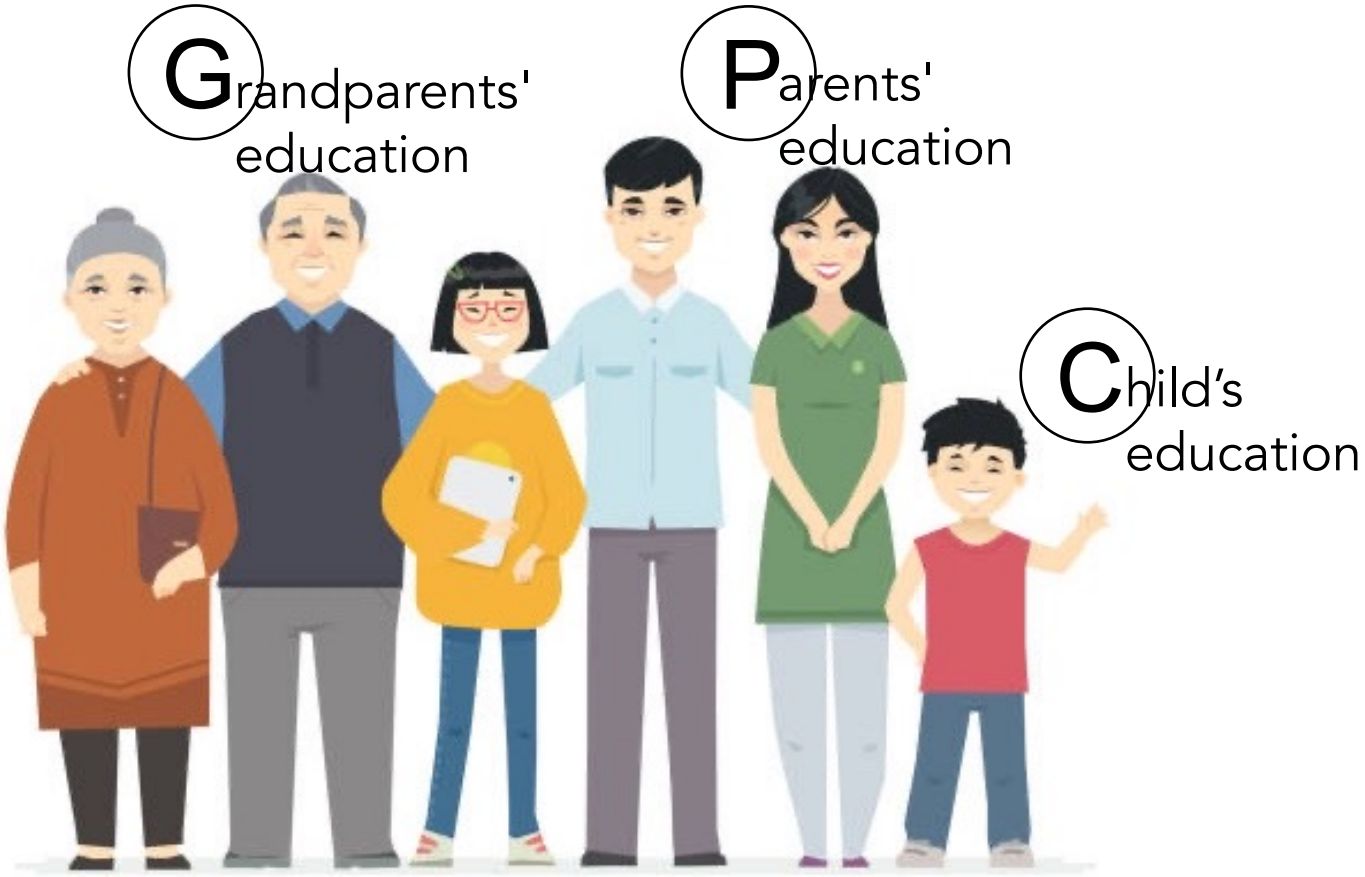
Model 3: C ~ P + G

Model 4: C ~ P x G



Parents positively affect child's education level

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

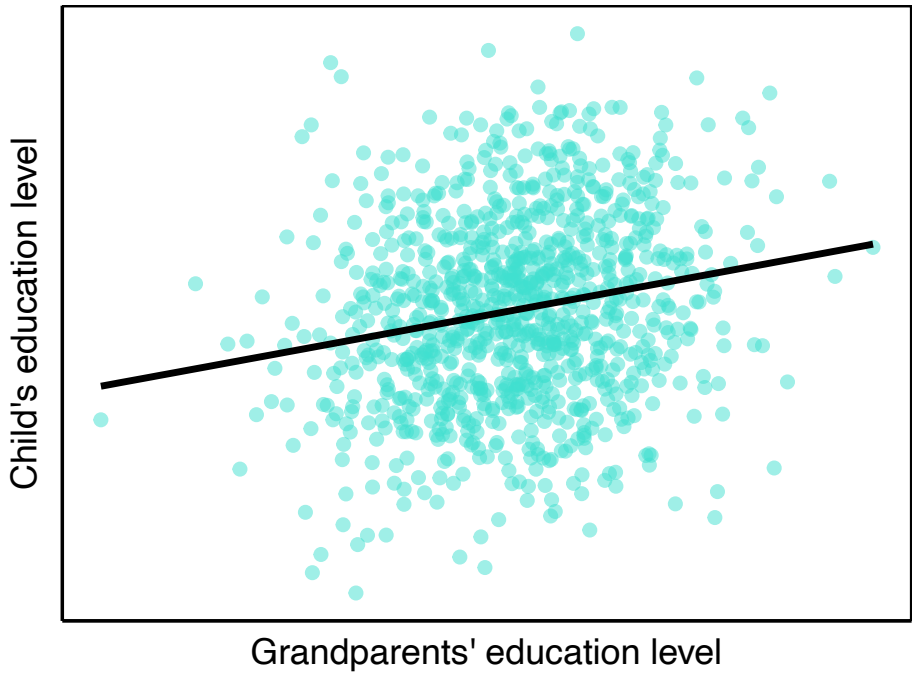


Model 1: C ~ P

Model 2: C ~ G

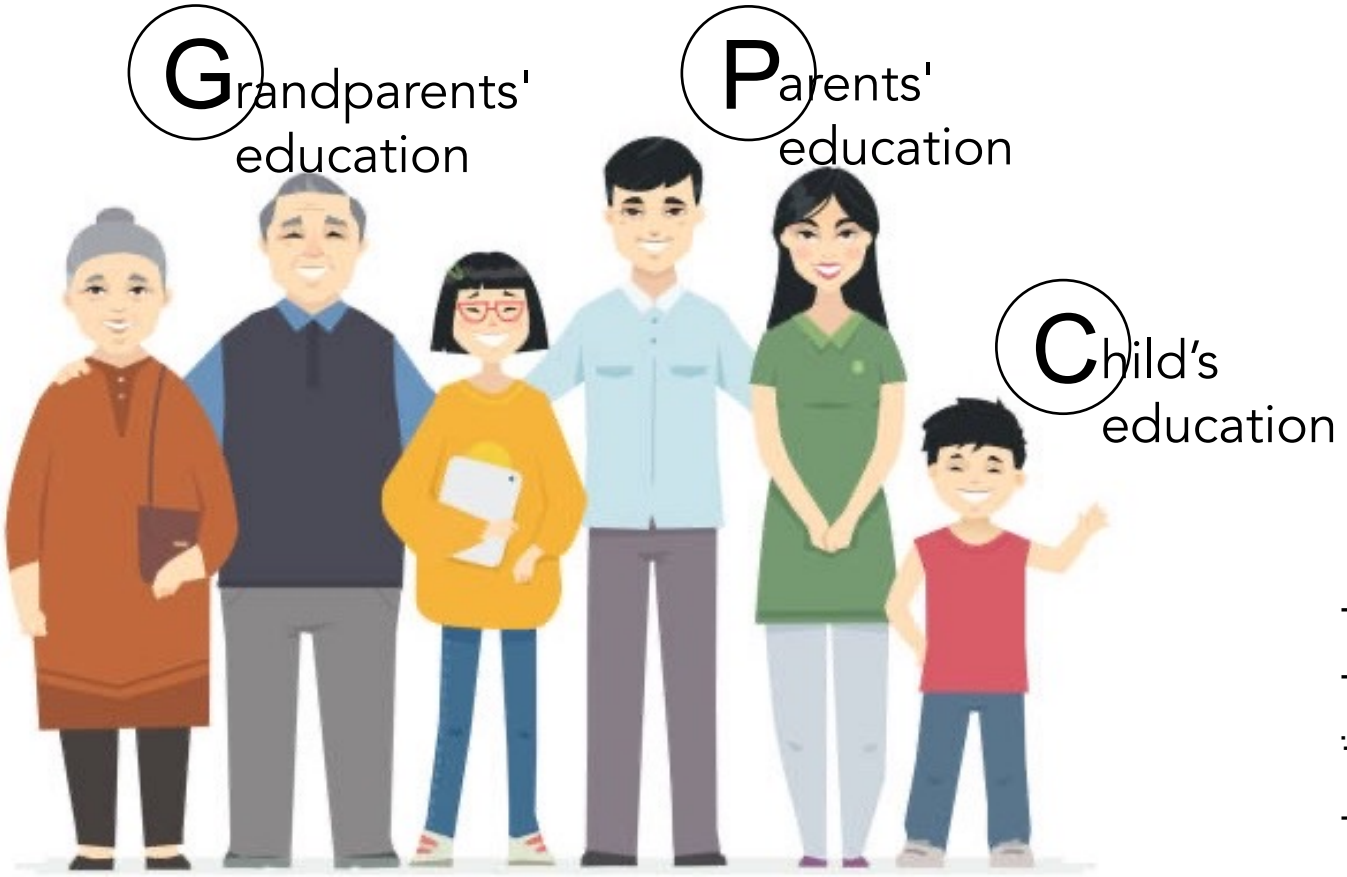
Model 3: C ~ P + G

Model 4: C ~ P x G



Grandparents positively affect child's education level

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

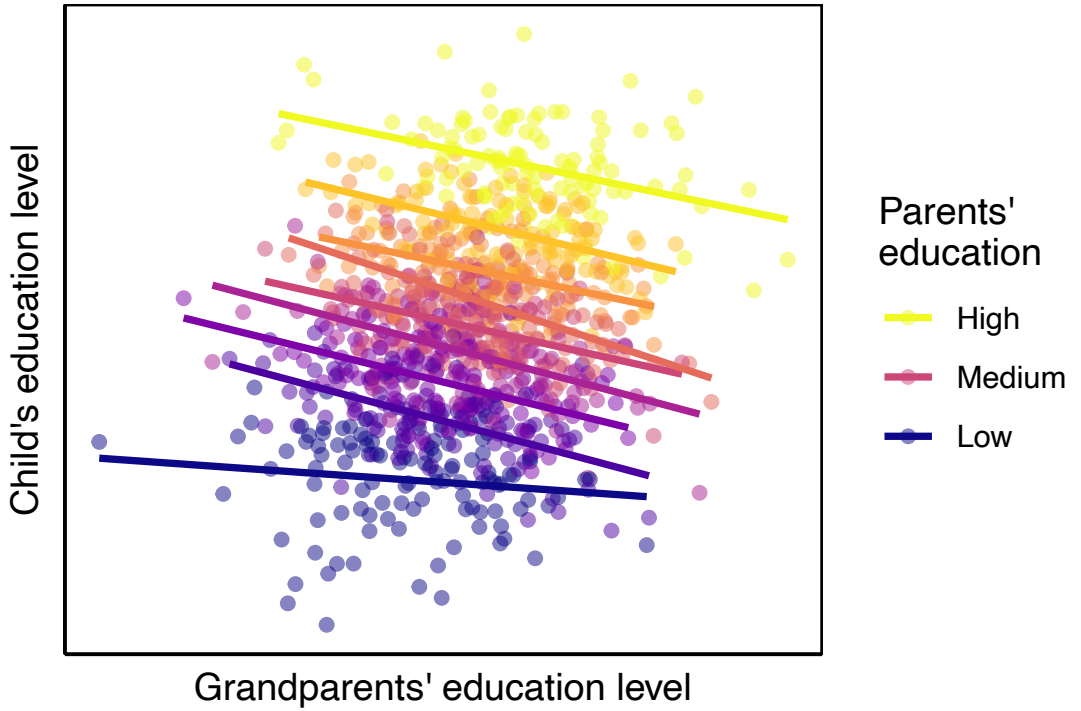


Model 1: C ~ P

Model 2: C ~ G

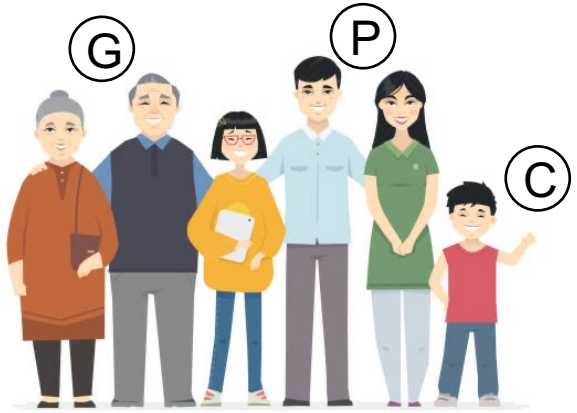
Model 3: C ~ P + G

Model 4: C ~ P x G



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

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 Baysian 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

```
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 ***
```

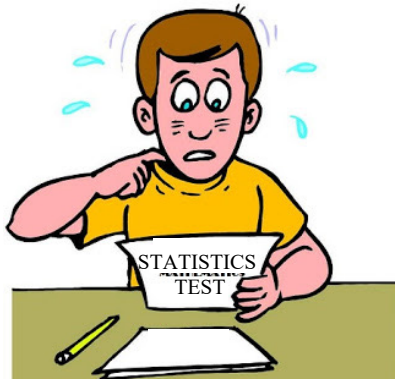
Conclusion: If your grandparents are illiterate, this is good for your own education success!

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)

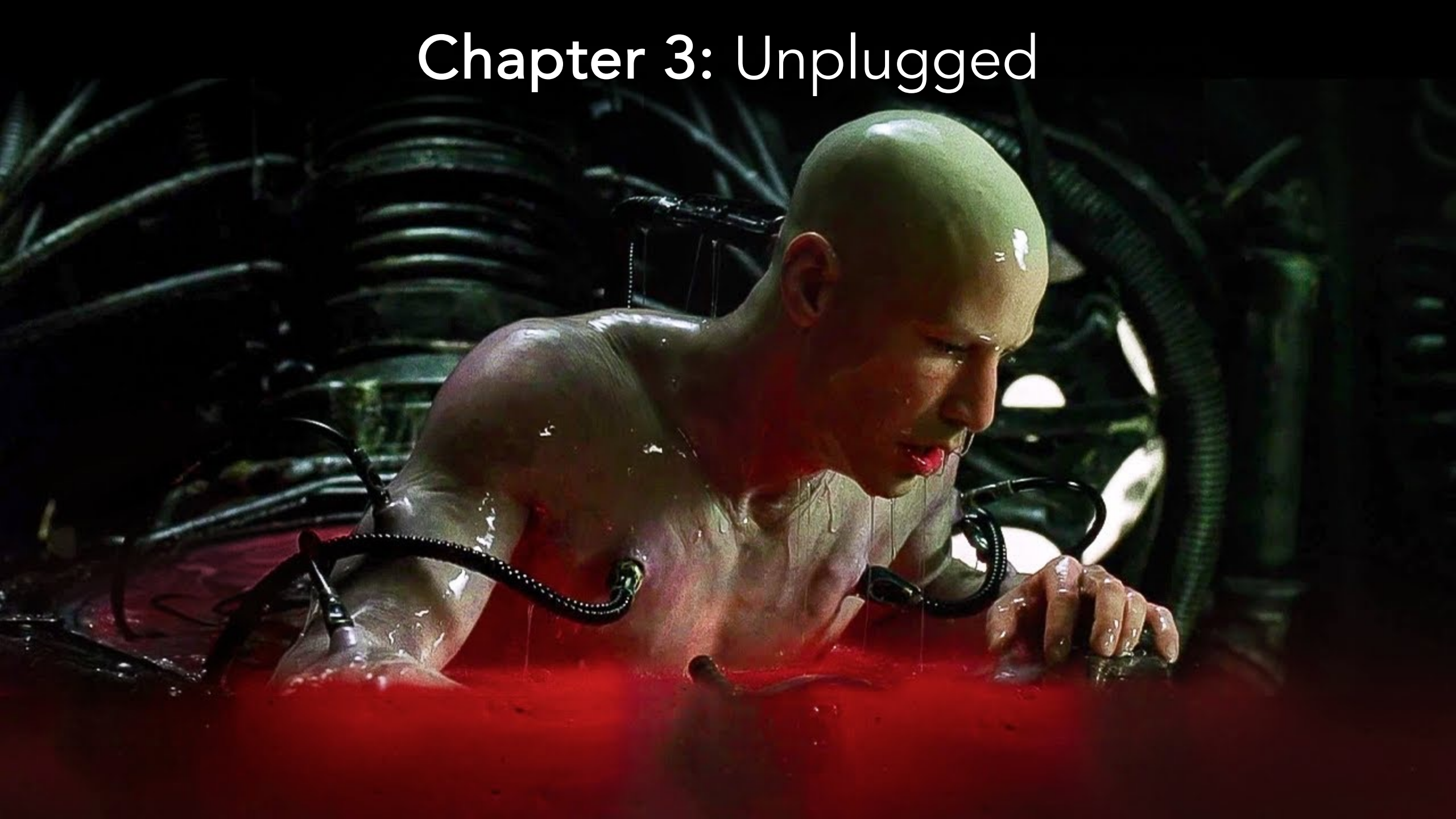
Modern Science/Statistics
(Pearson & Fisher's legacy)

- **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!

Causal Inference
(Sewall Wright and Judea Pearl)

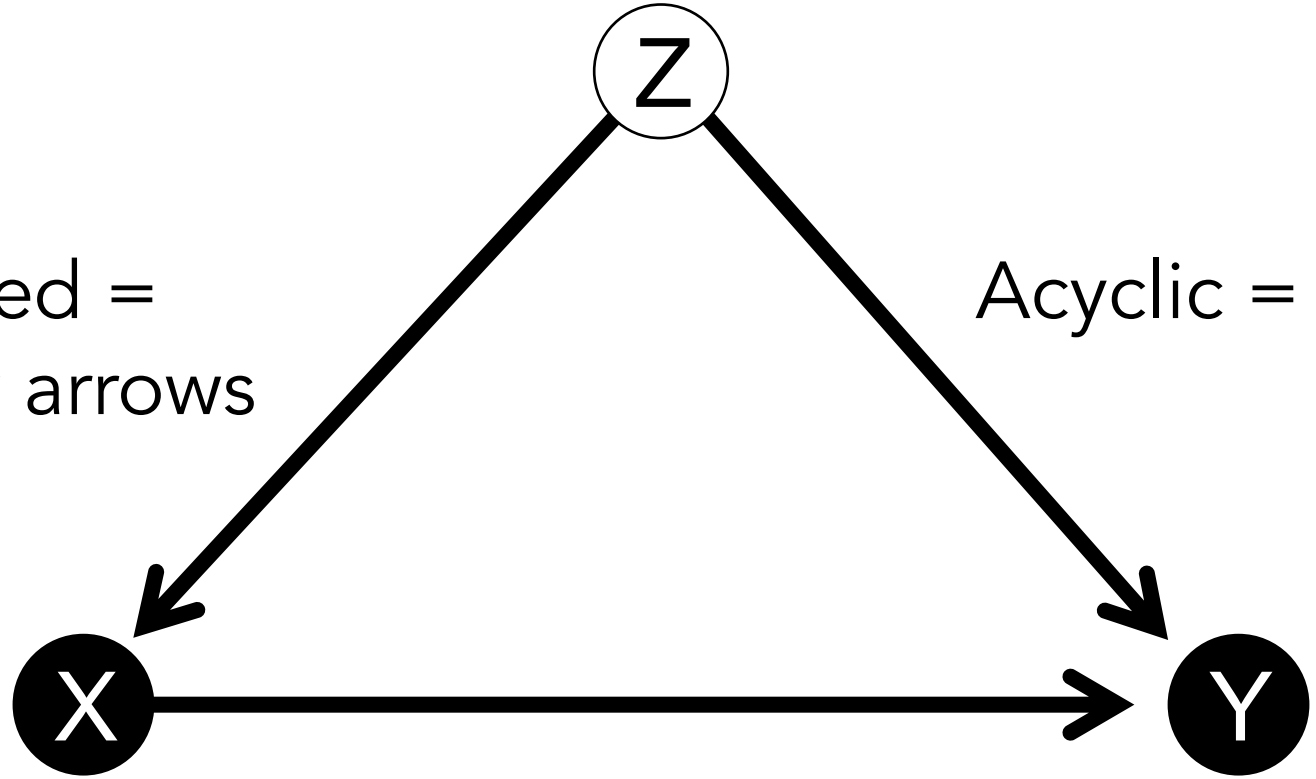
- We need to make **subjective** decisions using our expertise
- Actually, we can…
- **Causal diagrams** can inform us about which variables to include or 'condition' upon!

Chapter 3: Unplugged



Constructing DAGs (Directed Acyclical Graphs)

Directed =
one-way arrows

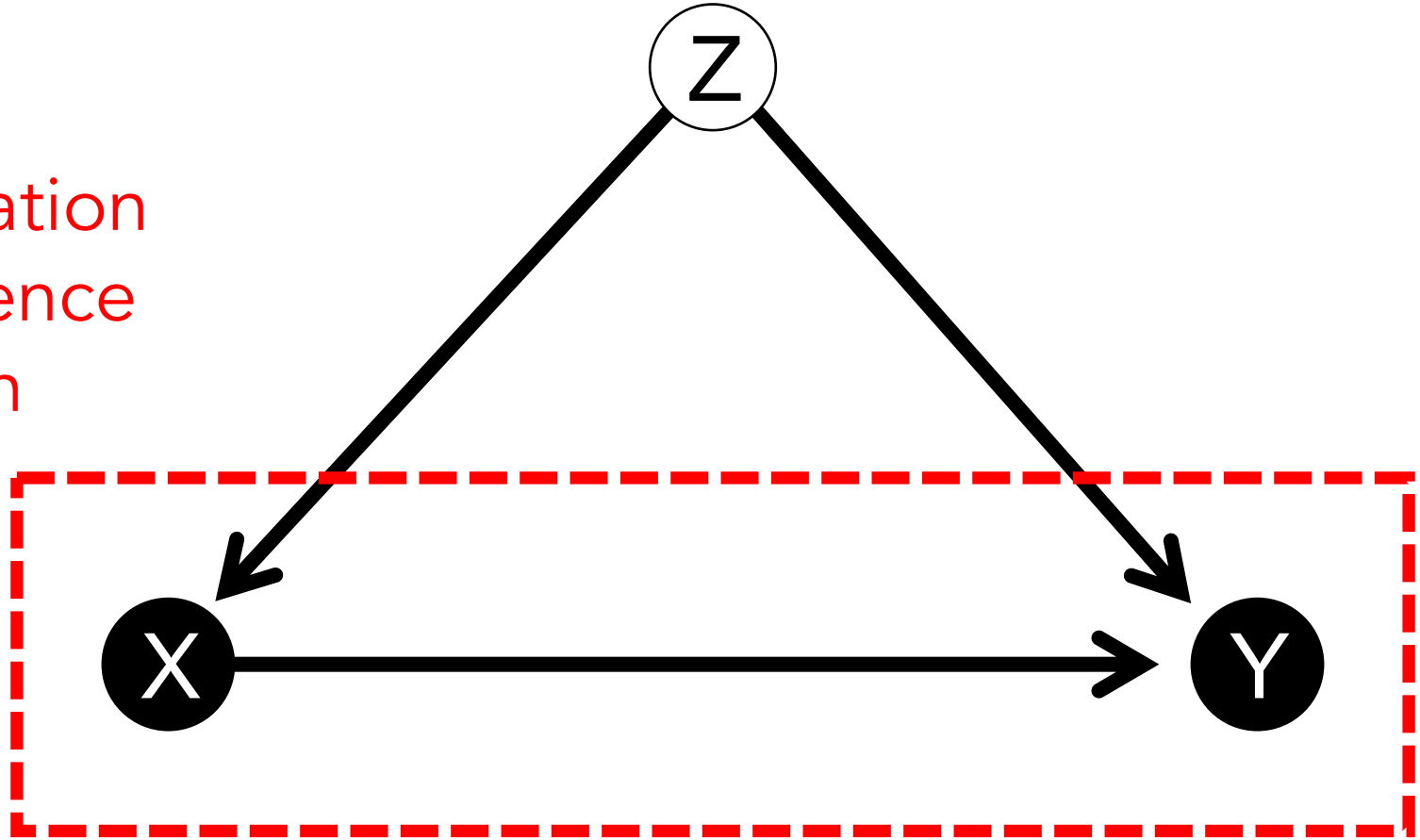


Acyclic = no looping

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

$X \rightarrow Z \rightarrow Y$

$Z \downarrow A$

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

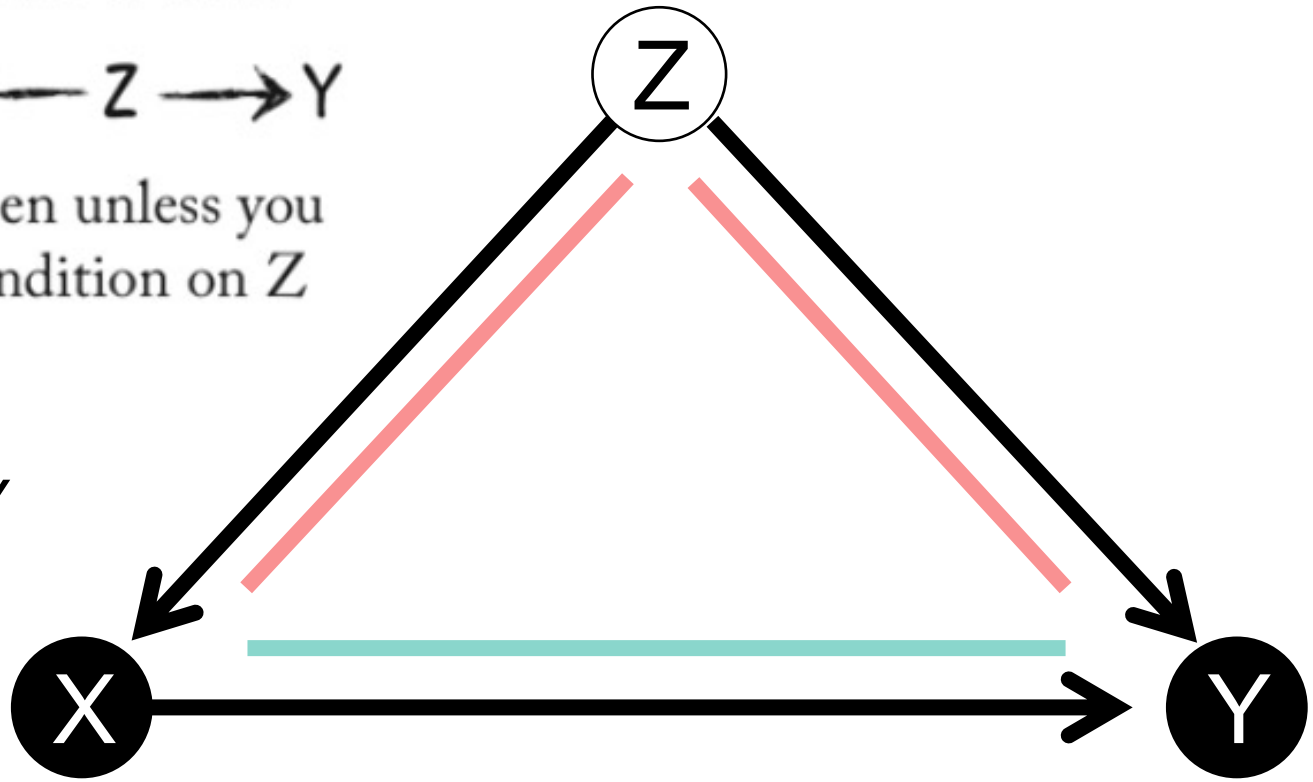
The Fork

X ← Z → Y

Open unless you condition on Z

Two paths:

- Path A: X → Y
- Path B: X ← Z → Y



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

Not ideal:
Y ~ X
(information can flow through variable Z)

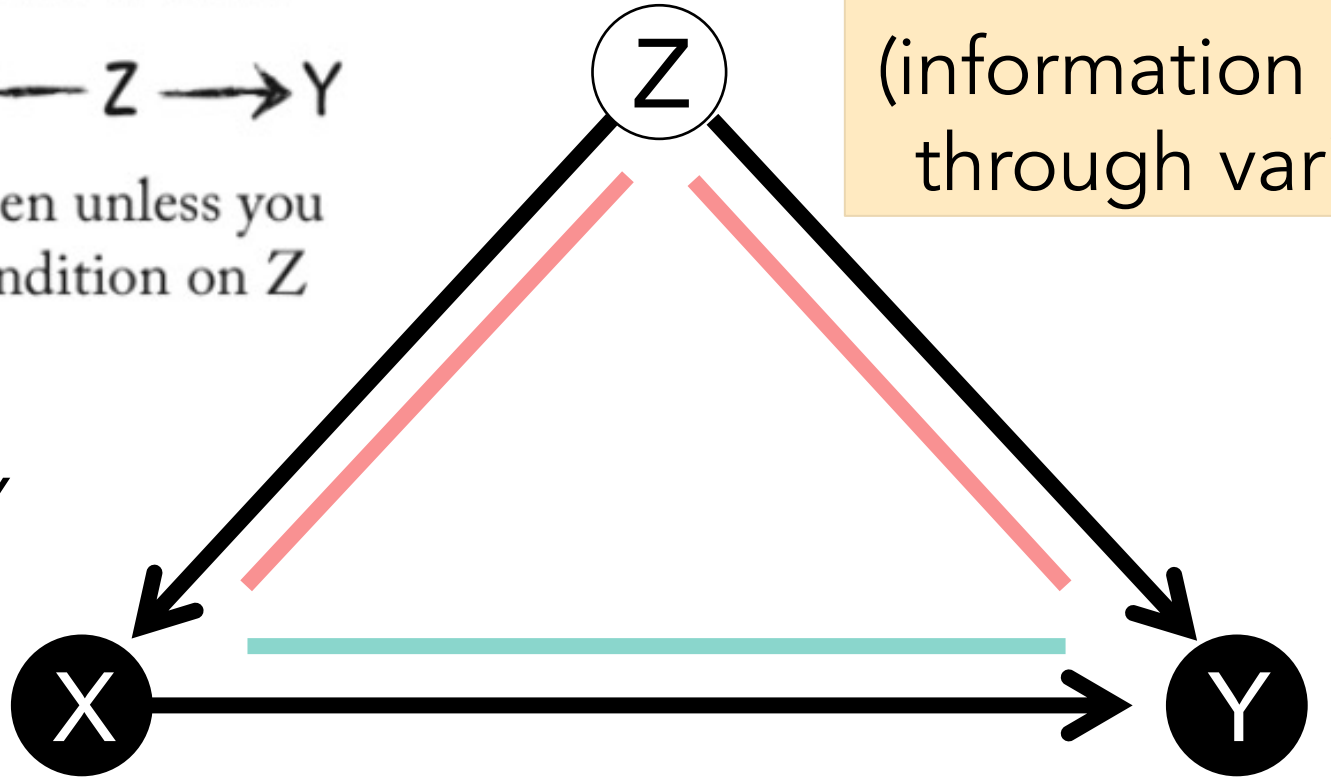
The Fork

$X \leftarrow Z \rightarrow Y$

Open unless you condition on Z

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!

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

The Fork

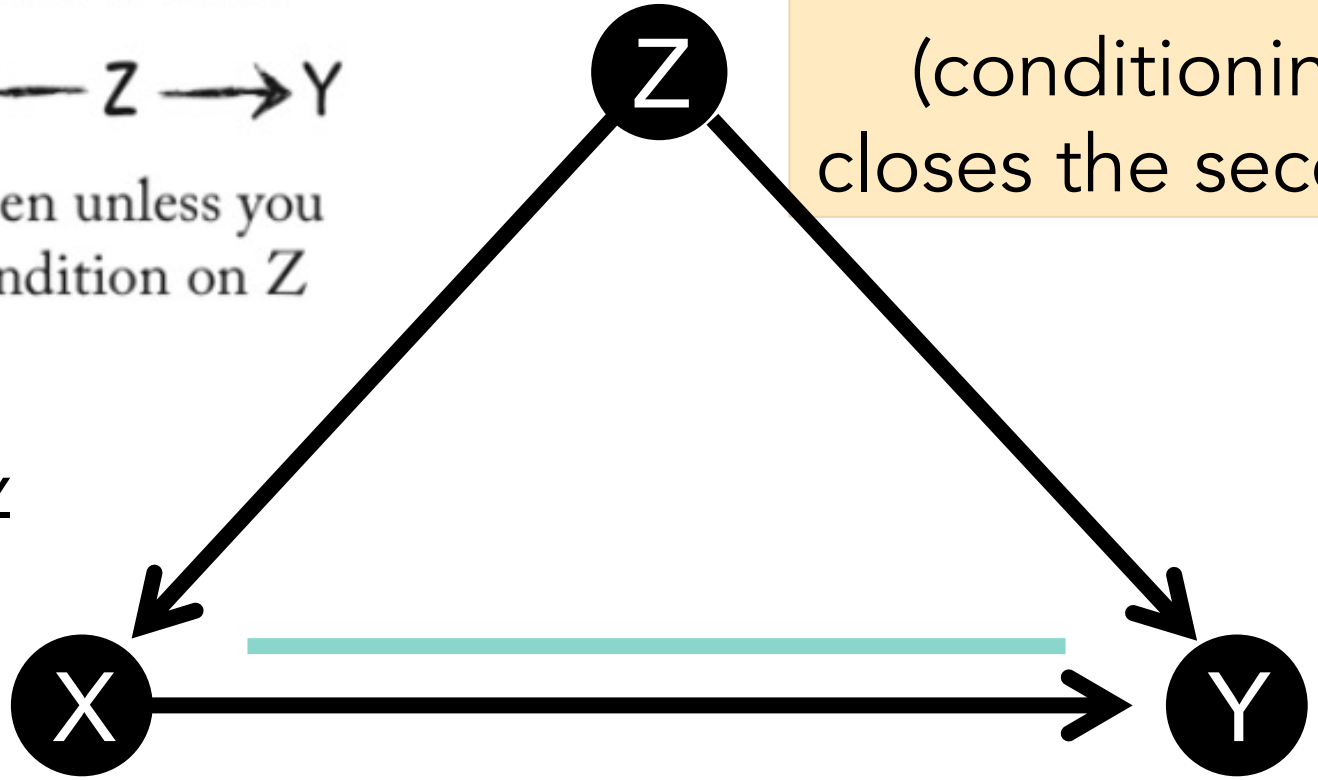
X ← Z → Y

Open unless you condition on Z

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

Two paths:

- Path A: X → Y
- ~~Path B: X ← Z → Y~~



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

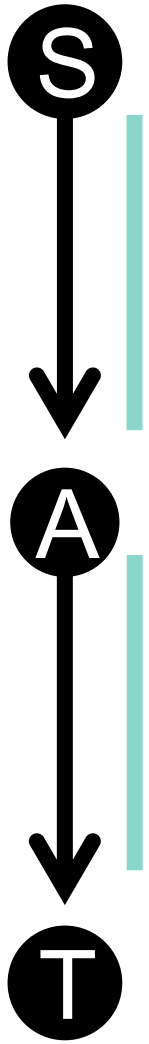
Chapter 4: Test your skills!

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



The Pipe

$X \rightarrow Z \rightarrow Y$

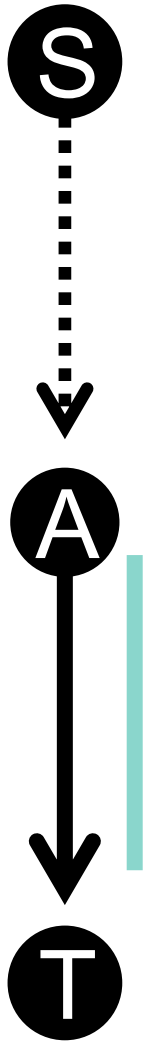
Open unless you condition on Z

One path:

- $S \rightarrow A \rightarrow T$

Example 1: Does sleep level affect student test results?

	Sleep	Alert	Test_score
	<dbl>	<dbl>	<dbl>
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:
T ~ S + A

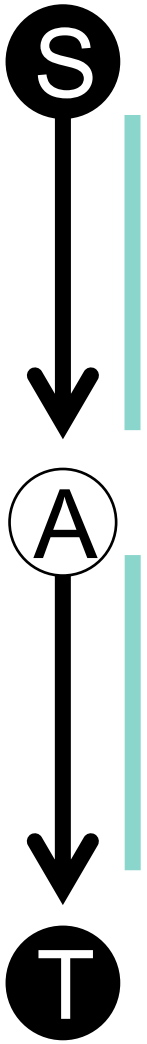
Including alertness blocks the effect of sleep on test score!

One path:

- S → A → T

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




```
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 ***
```

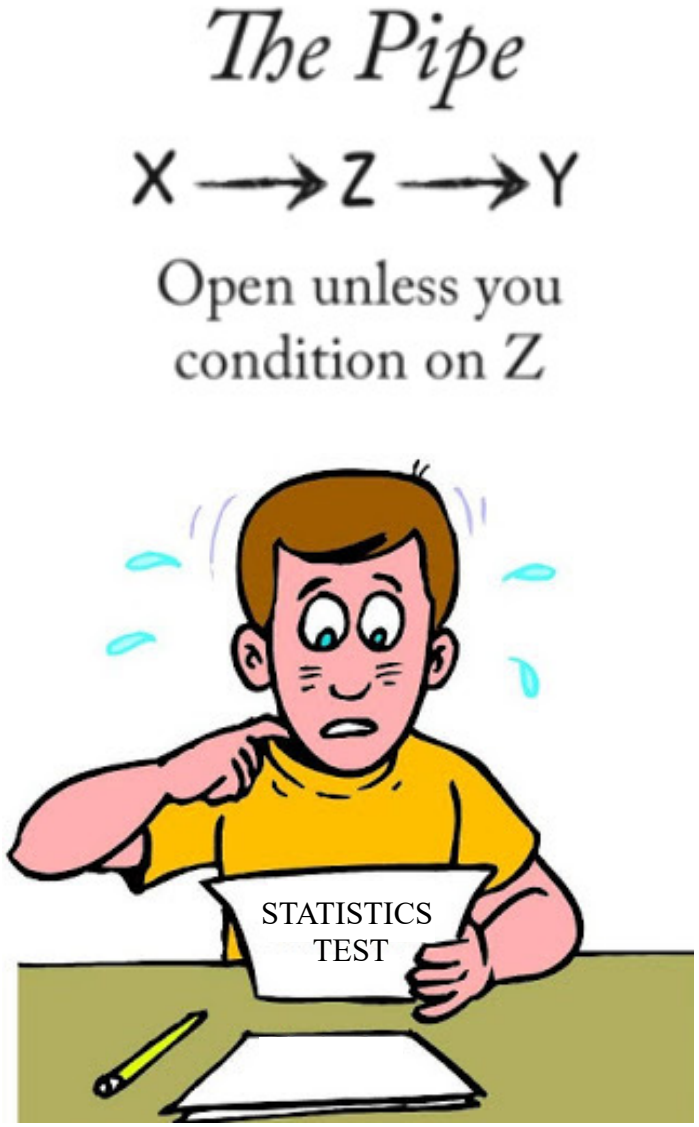
The Pipe

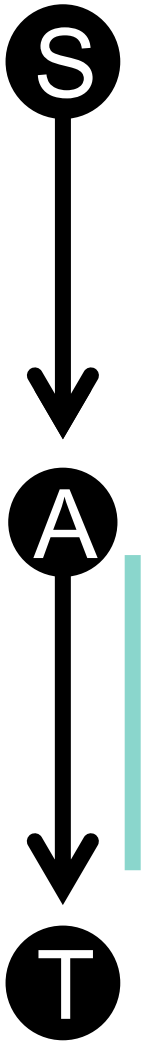
X → Z → Y

Open unless you condition on Z

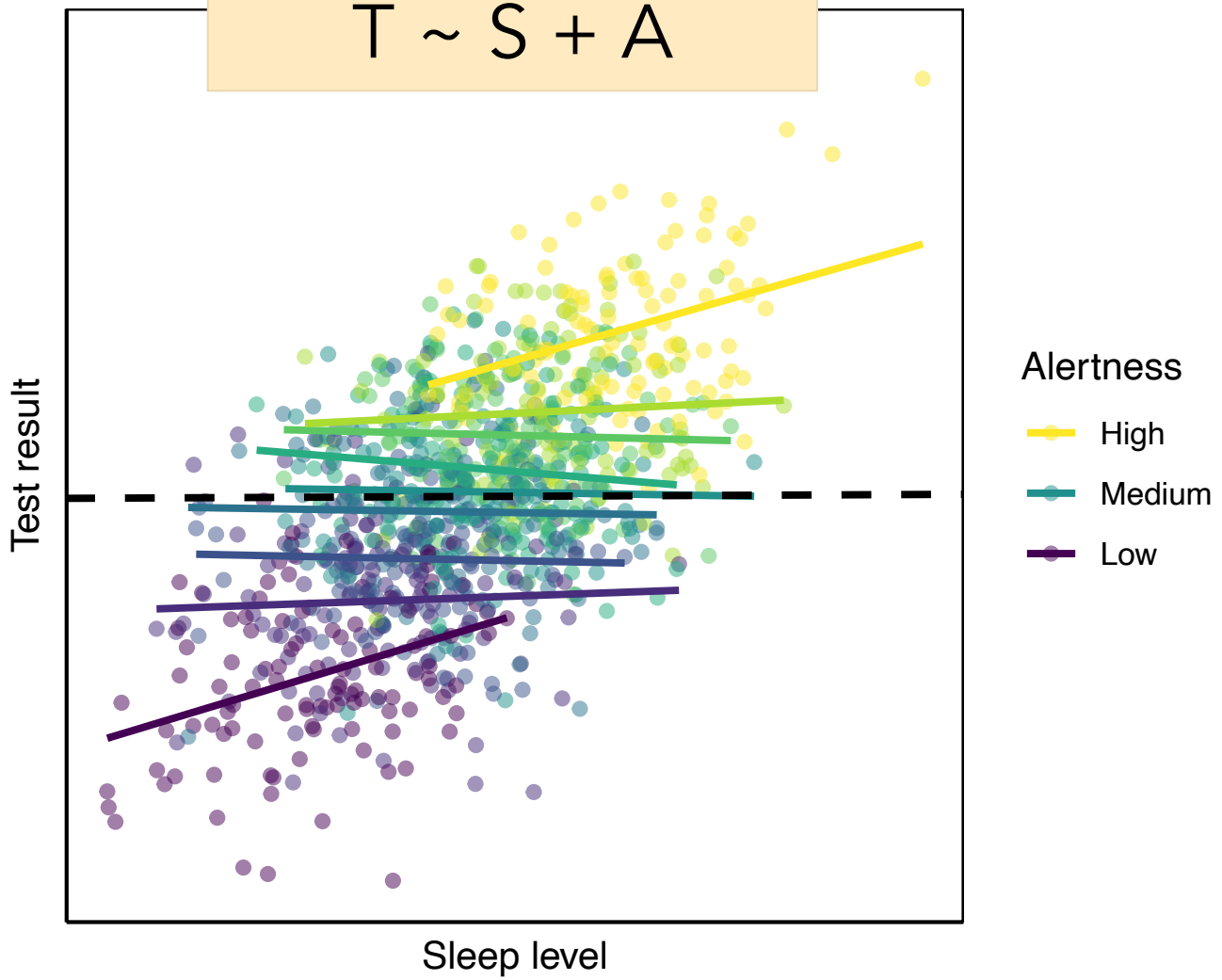
Ideal model:
T ~ S

Example 1: Does sleep level affect student test results?

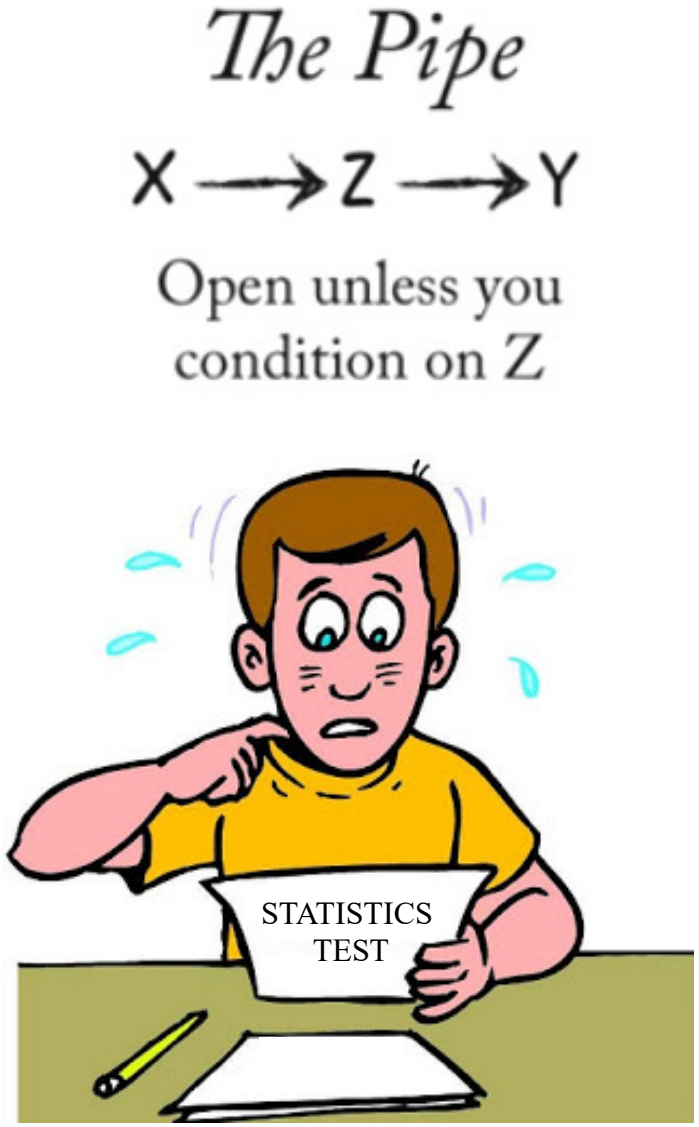


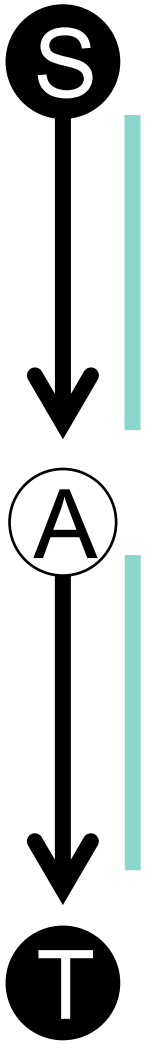


Not ideal model:
T ~ S + A

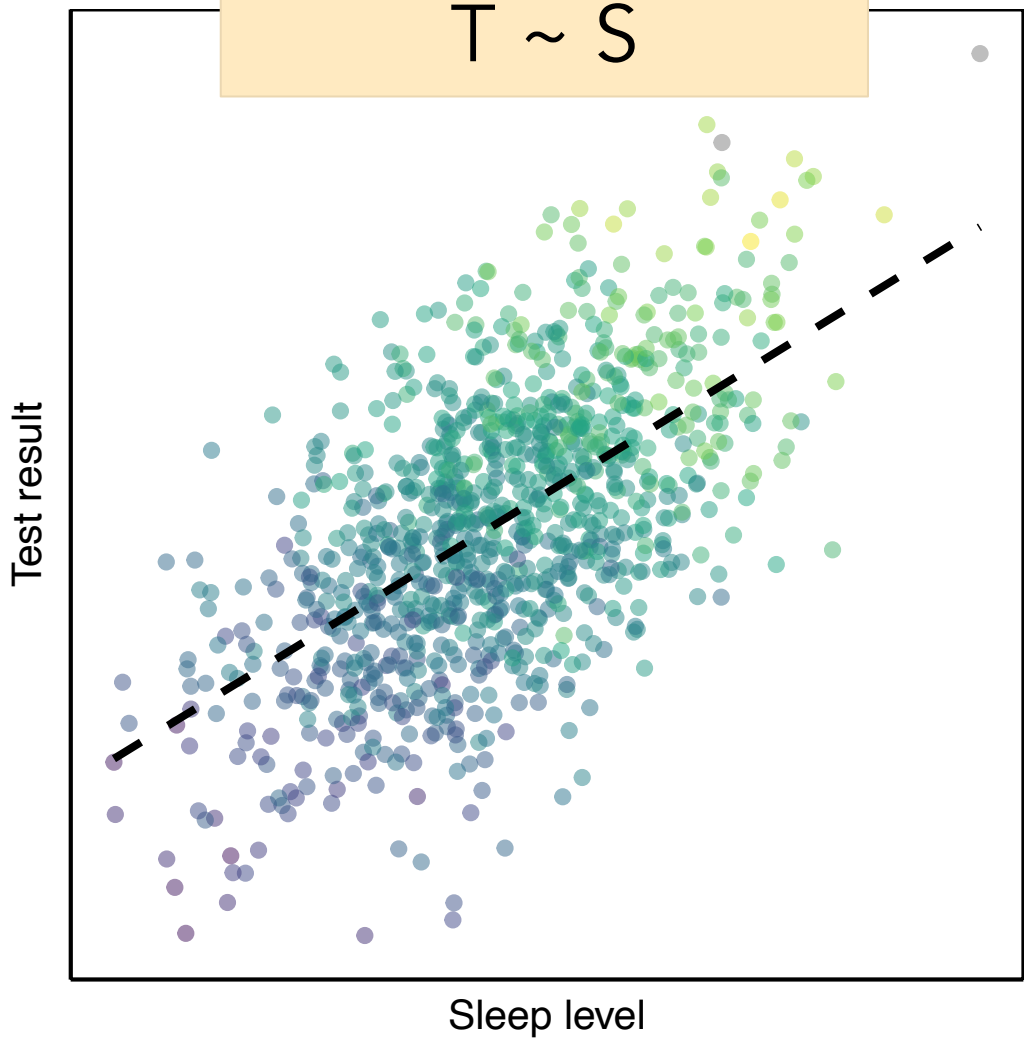


Example 1: Does sleep level affect student test results?

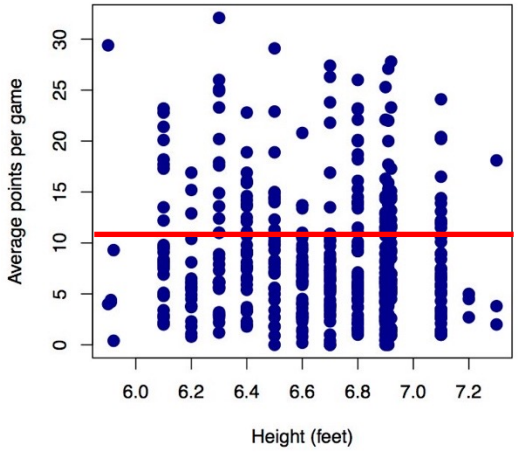




Ideal model:
T ~ S



Example 2: Does basketball player height affect scoring?



Two paths:

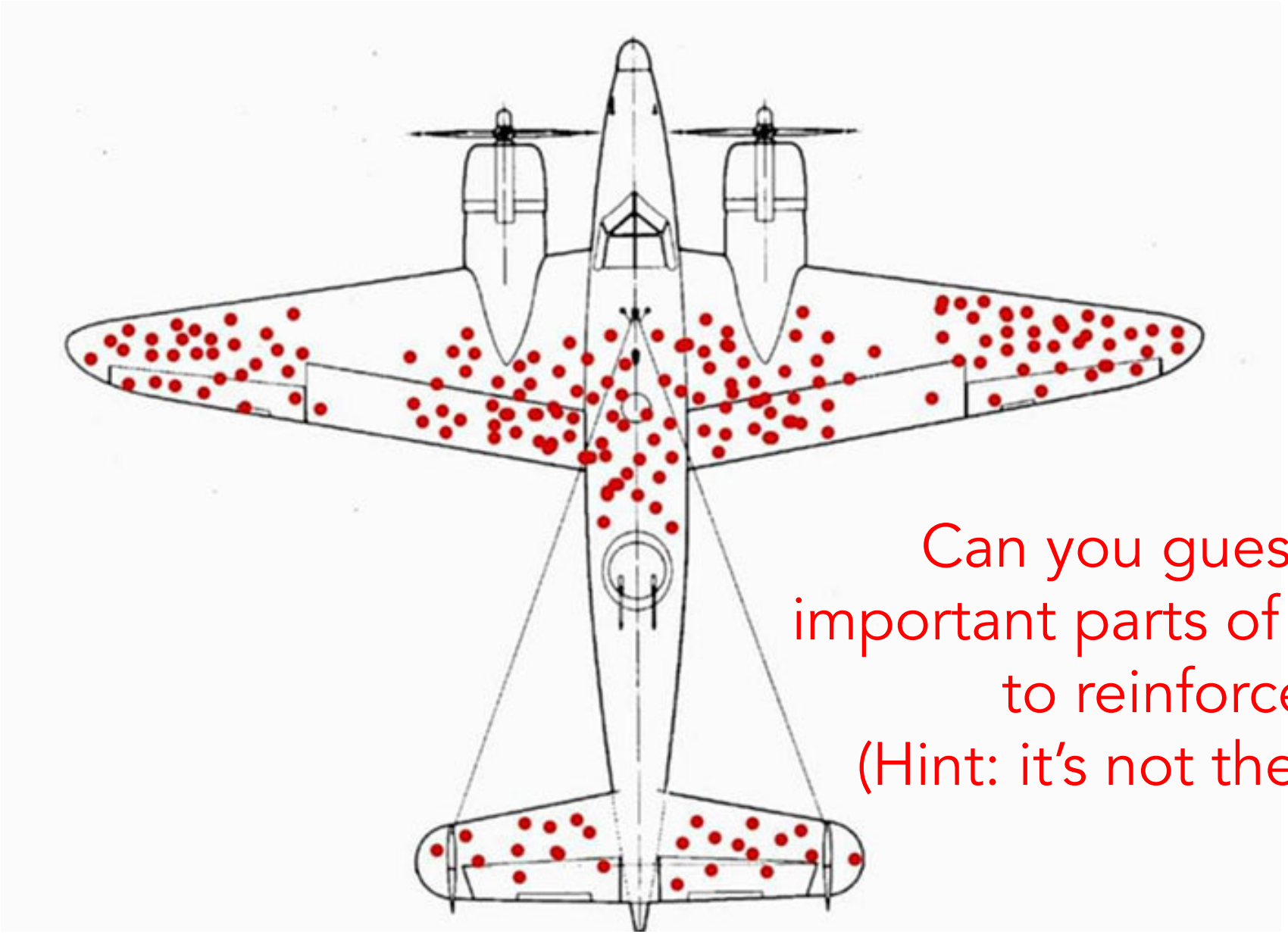
- Path A: H → S
- Path B: H → B ← S

The Collider

X → Z ← Y

Closed until you condition on Z

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

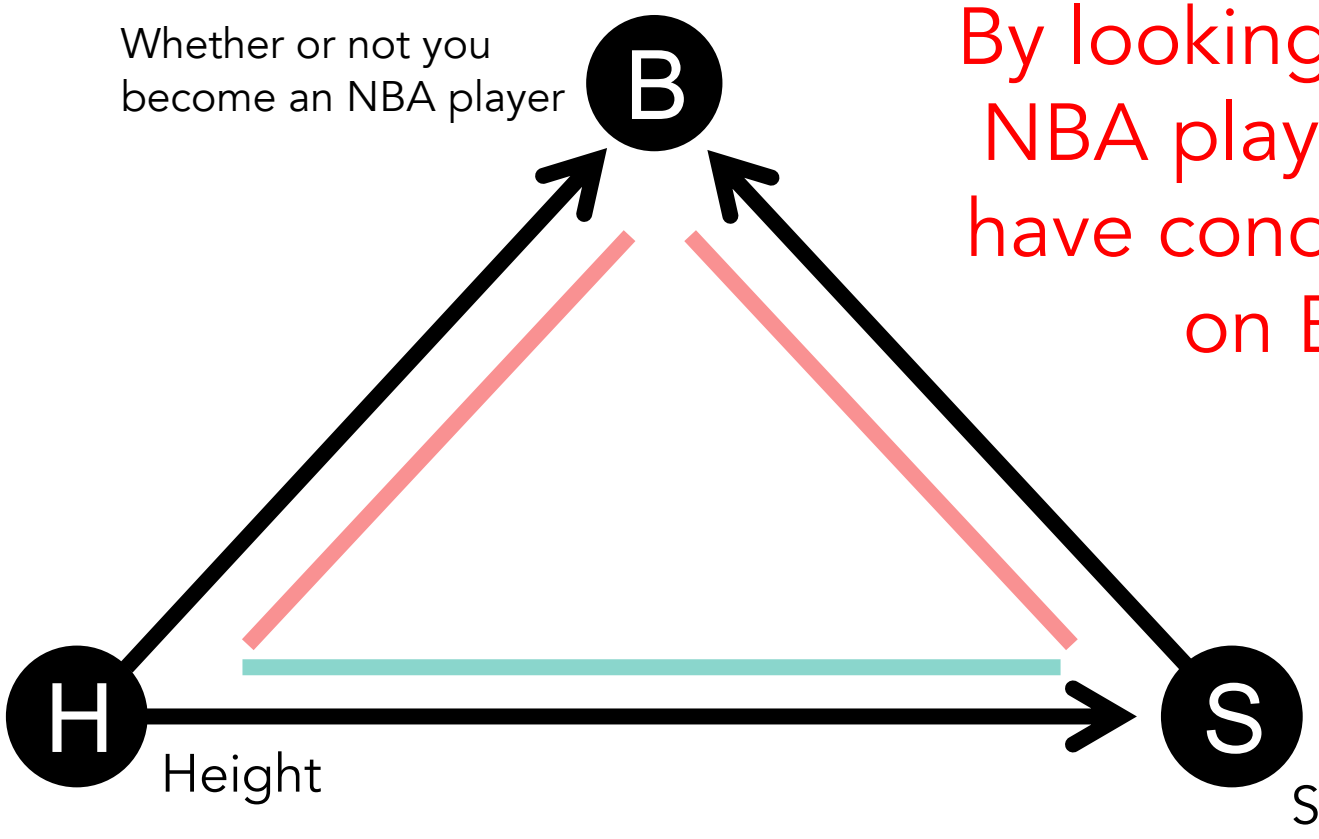
X → Z ← Y

Closed until you condition on Z

Two paths:

- Path A: H → S
- Path B: H → B ← S

Not ideal:
S ~ H + B



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

Example 2: Does basketball player height affect scoring?

The Collider

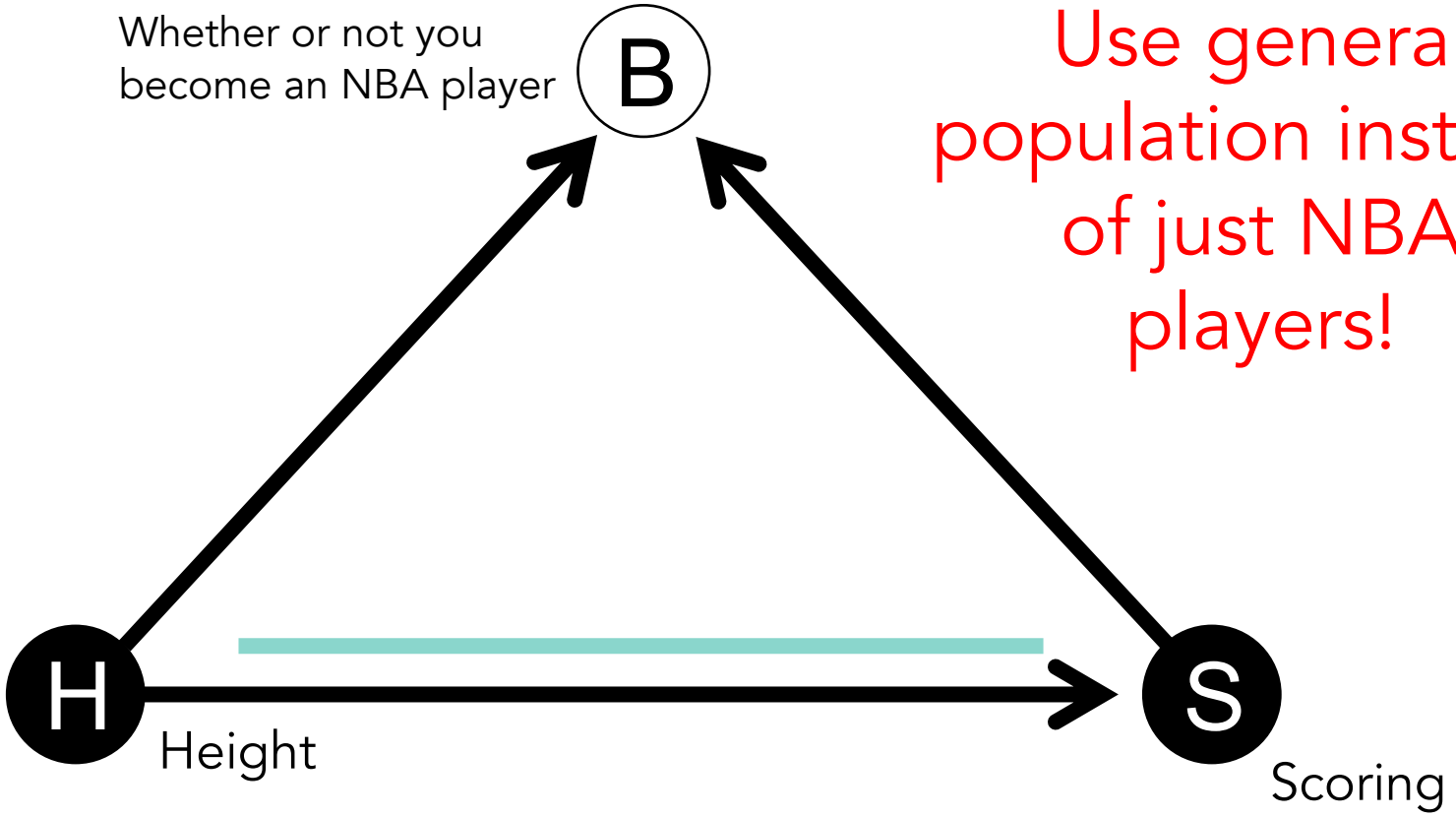
X → Z ← Y

Closed until you condition on Z

Two paths:

- Path A: H → S
- ~~Path B: H → B ← S~~

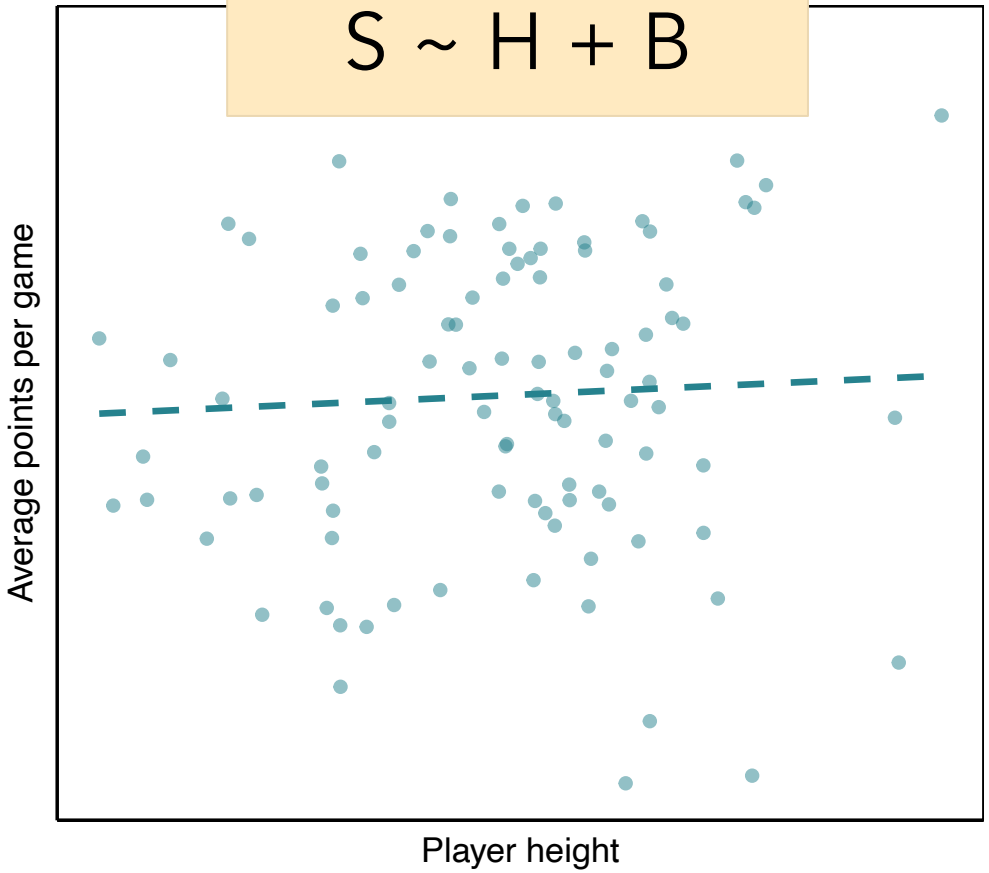
Ideal:
S ~ H



Use general population instead of just NBA players!

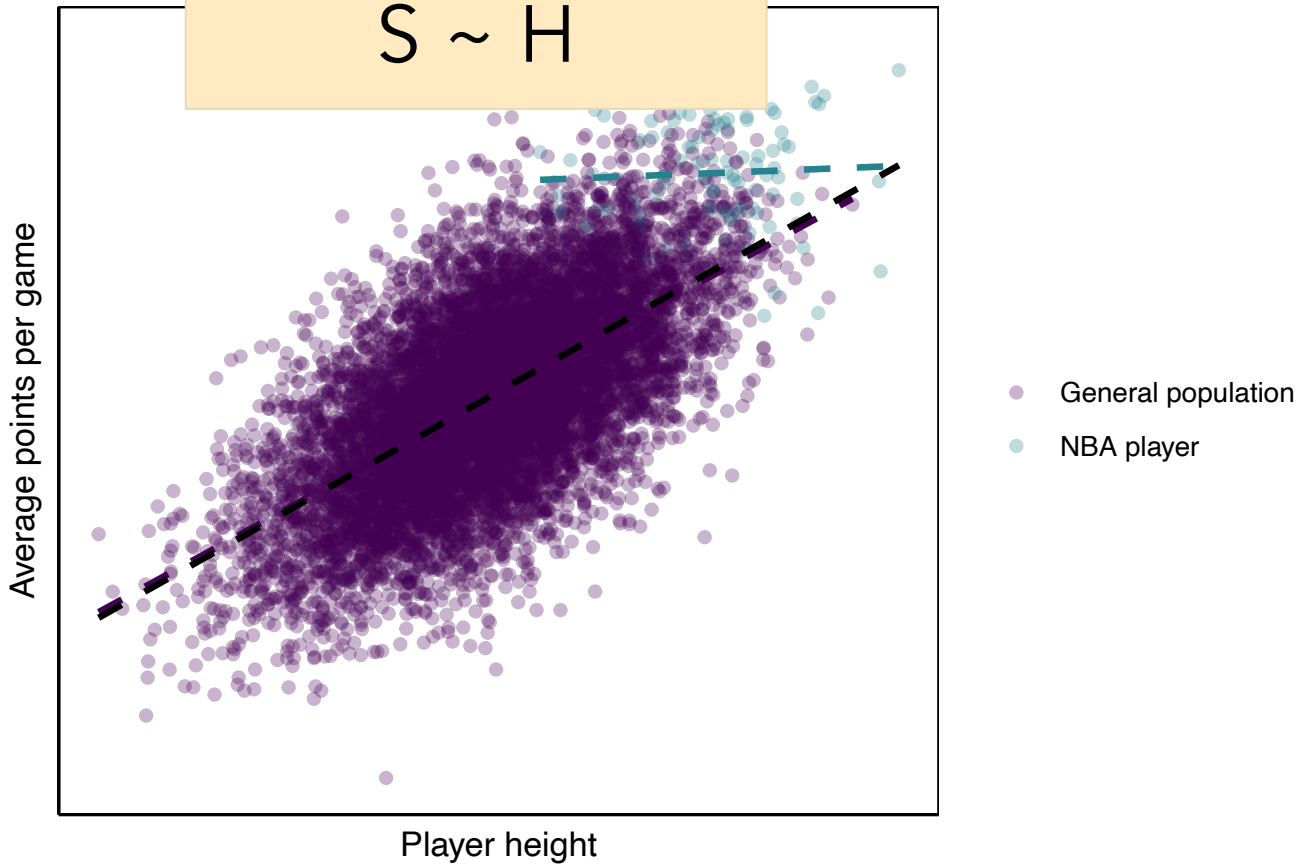
Example 2: Does basketball player height affect scoring?

Not ideal:
S ~ H + B



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

Ideal:
S ~ H



nature

CAREER COLUMN | 27 September 2021

Beware survivorship bias in advice on science careers

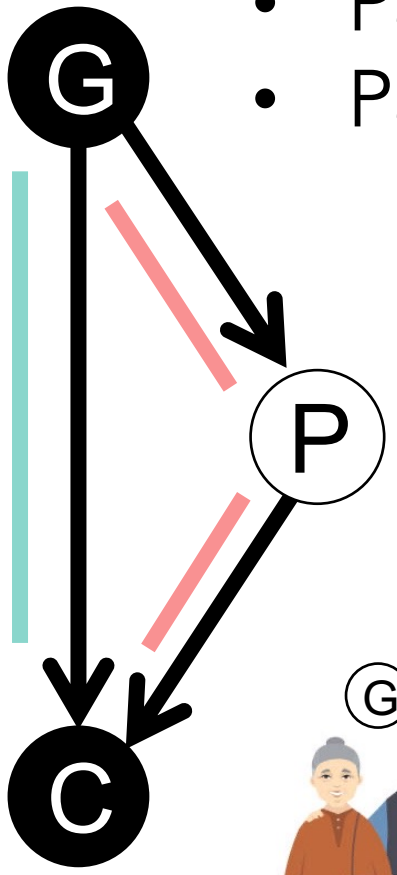
For objective careers advice, talk to those who left science as well as those who stayed.

Dave Hemprich-Bennett , Dani Rabaiotti & Emma Kennedy

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

Two paths:

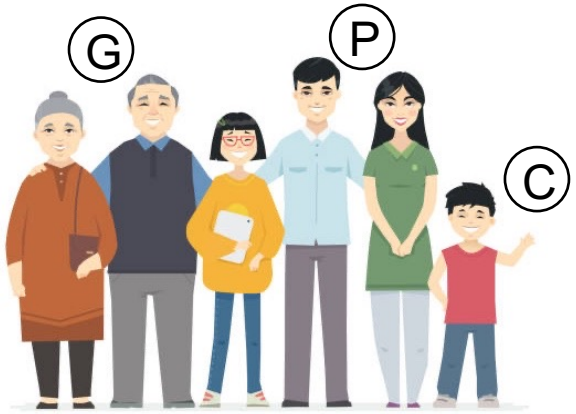
- Path A: G → C
- Path B: G → P → C



The Pipe

X → Z → Y

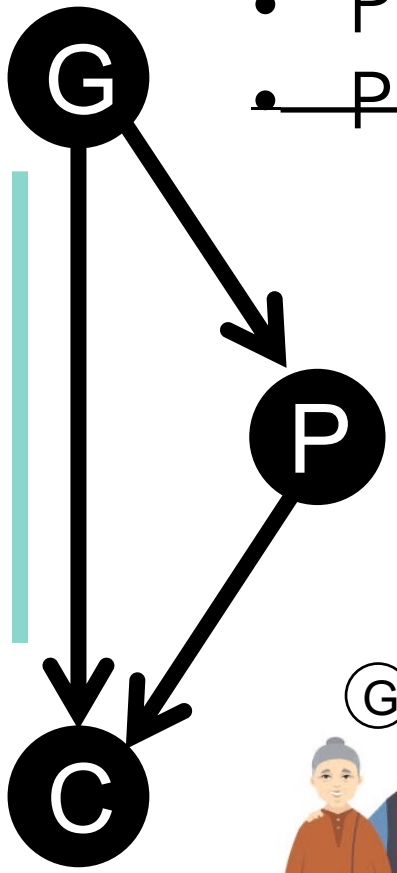
Open unless you condition on Z



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

Two paths:

- Path A: G → C
- ~~Path B: G → P → C~~



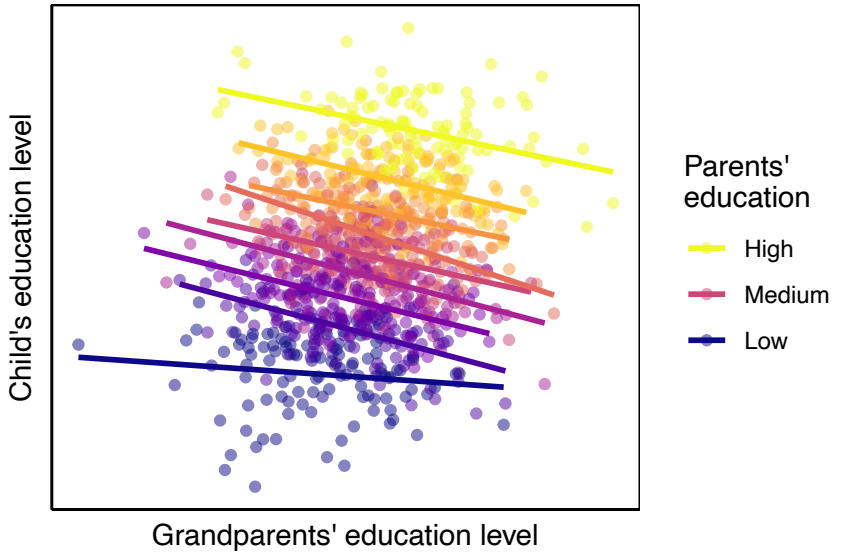
The Pipe

X → Z → Y

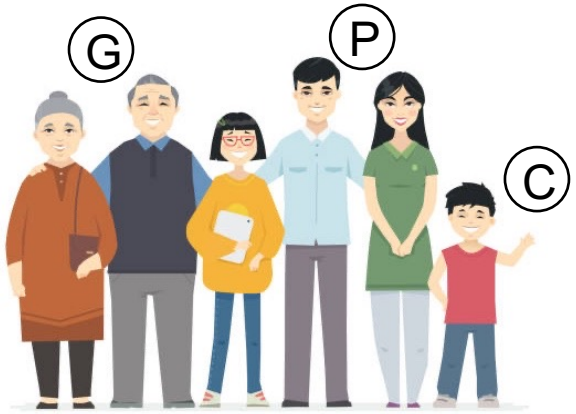
Open unless you condition on Z

Ideal model

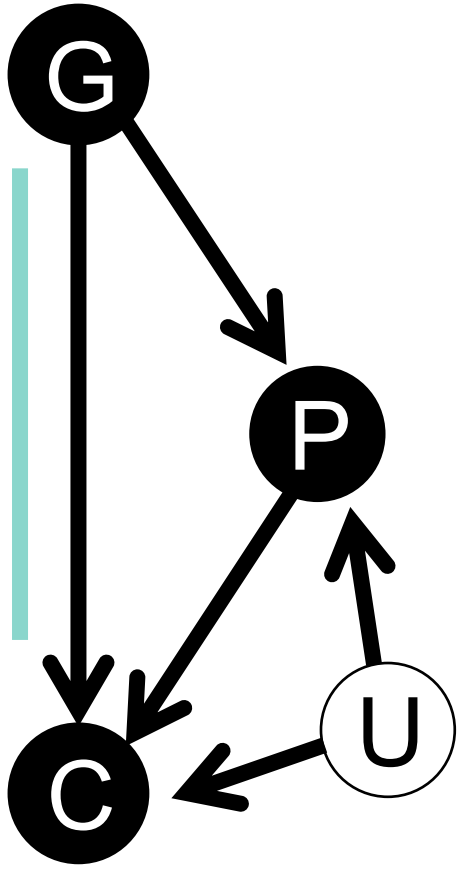
C ~ G + P

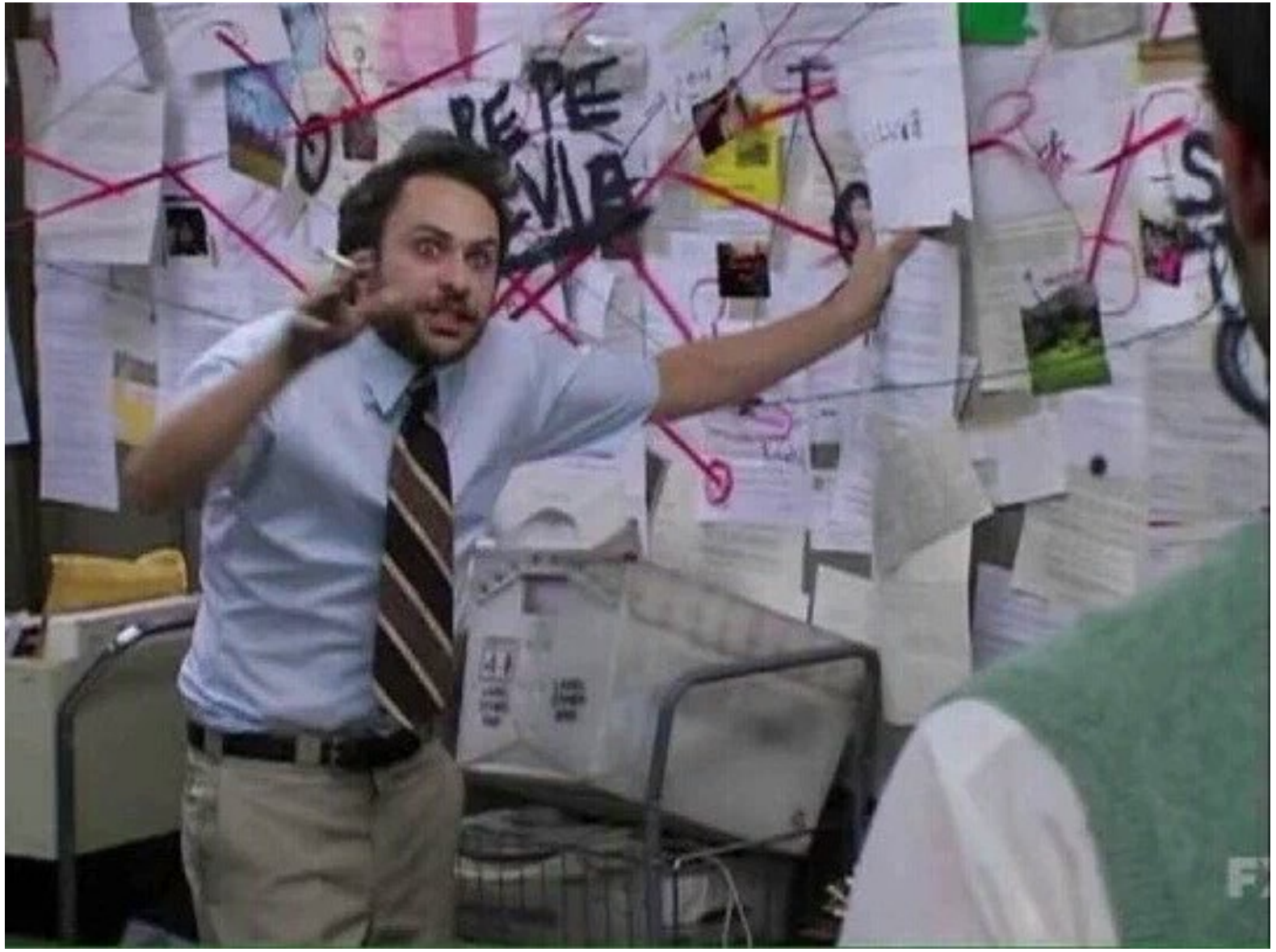



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09936    0.06632  -1.498    0.134
G           -1.70282    0.07160 -23.783   <2e-16 ***
P            2.67736    0.02937  91.150   <2e-16 ***
```

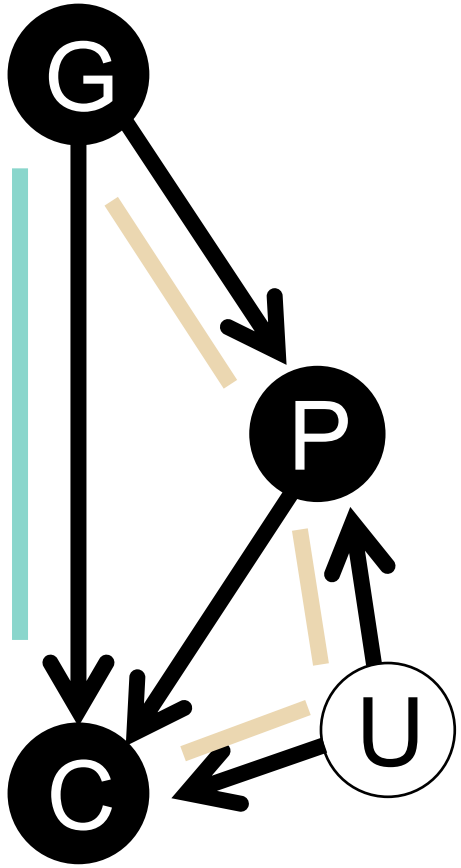


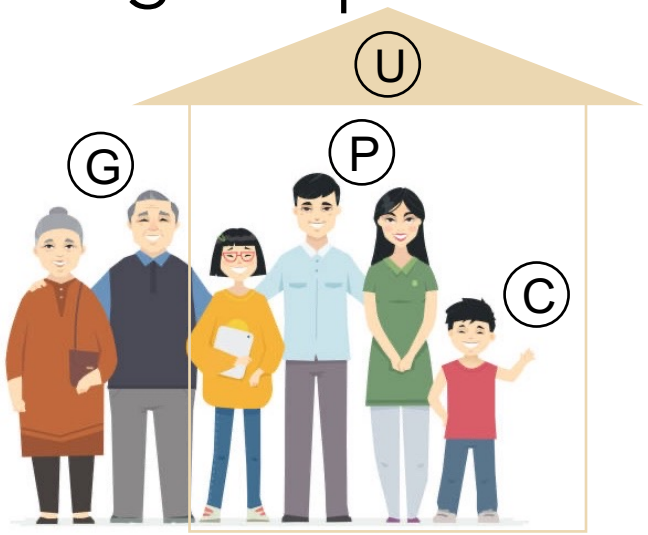
Wait – it's still a negative relationship??


G
P
U
C



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





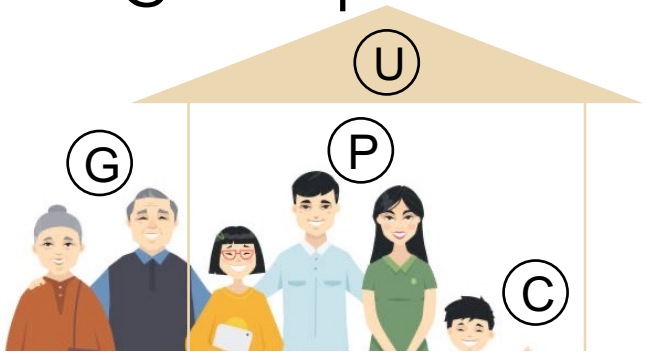
- Path A: G → C
- ~~Path B: G → P → C~~
- Path C: G → P ←U → C

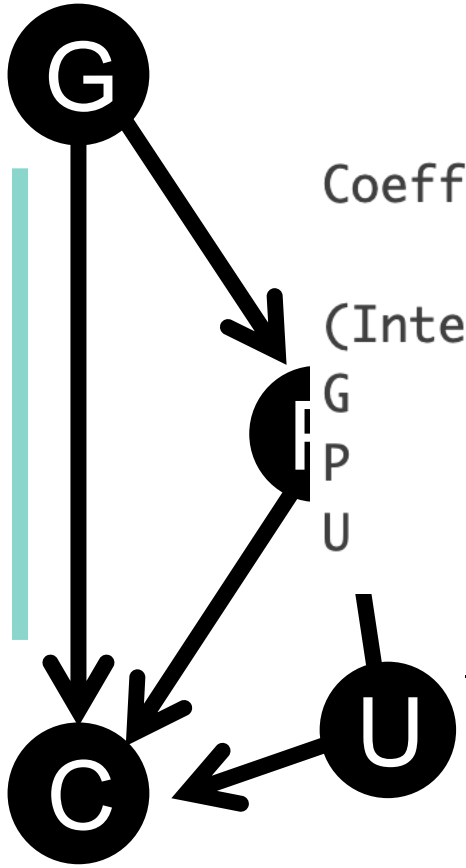
The Collider

X → Z ← Y

Closed until you condition on Z

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





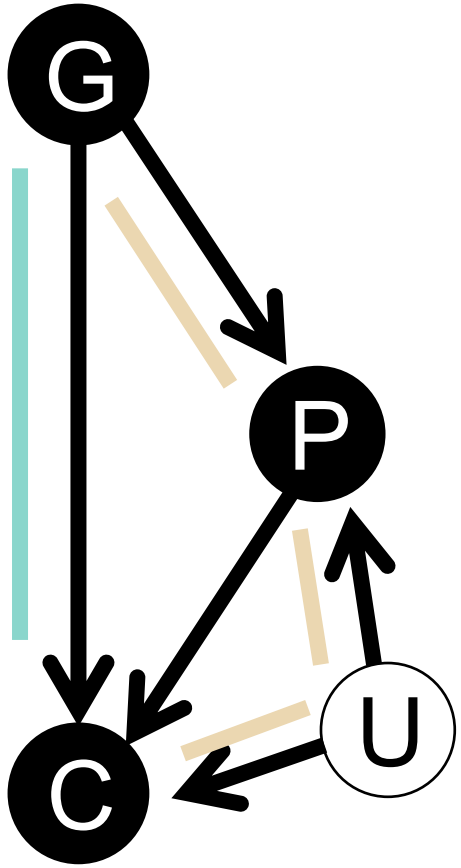

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02382    0.03135    0.76    0.448
G            1.00520    0.04384   22.93   <2e-16 ***
P            1.04283    0.03066   34.02   <2e-16 ***
U            3.85112    0.06858   56.16   <2e-16 ***
```

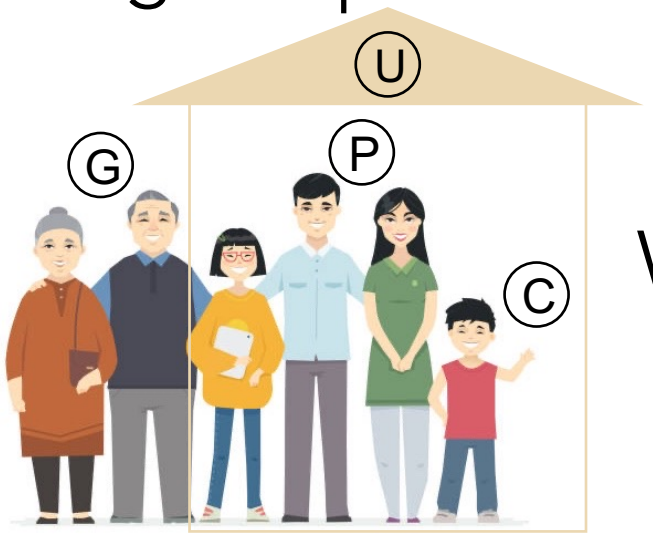
- ~~Path C: G → P ← U → C~~

Ideal model
C ~ P + C + U

```
Coefficients:
             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 **<u>can't</u>** control for *U*?!

Three paths:

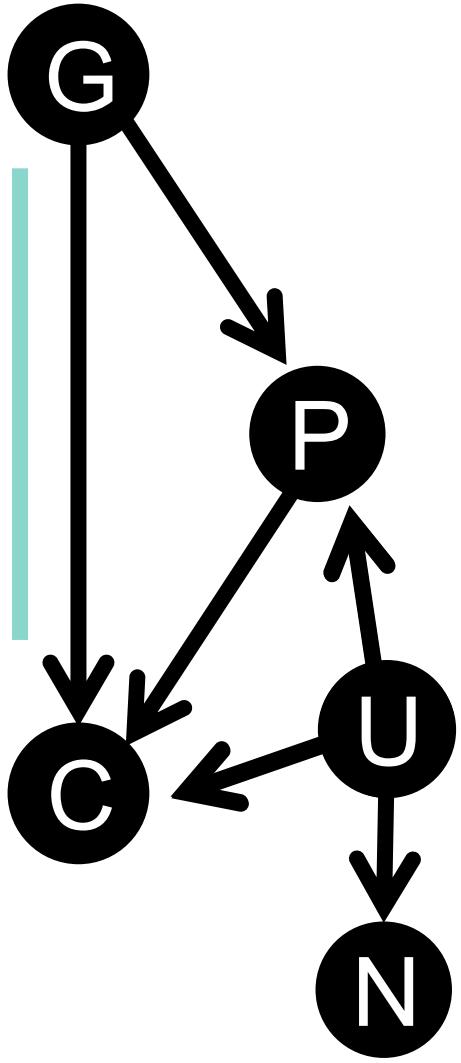
- Path A: G → C
- ~~Path B: G → P → C~~
- Path C: G → P ←U → C

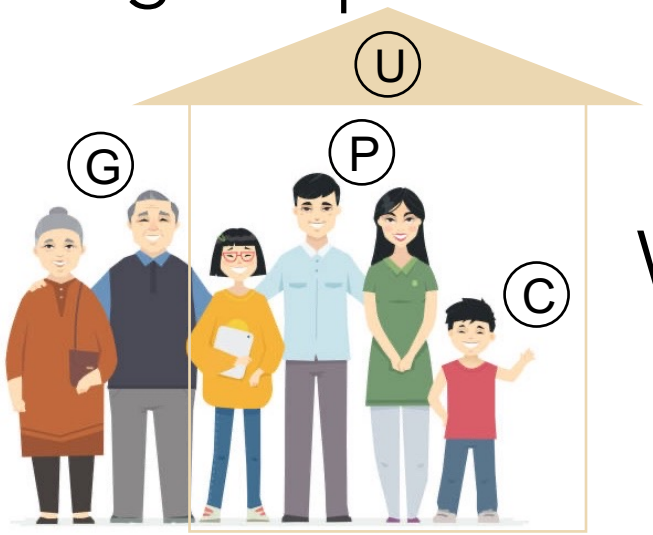
The Collider

X → Z ← Y

Closed until you condition on Z

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





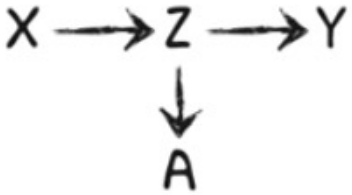
Three paths:

- Path A: G → C
- ~~Path B: G → P → C~~
- ~~Path C: G → P ← U → C~~

What if we **can't** control for *U*?!

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

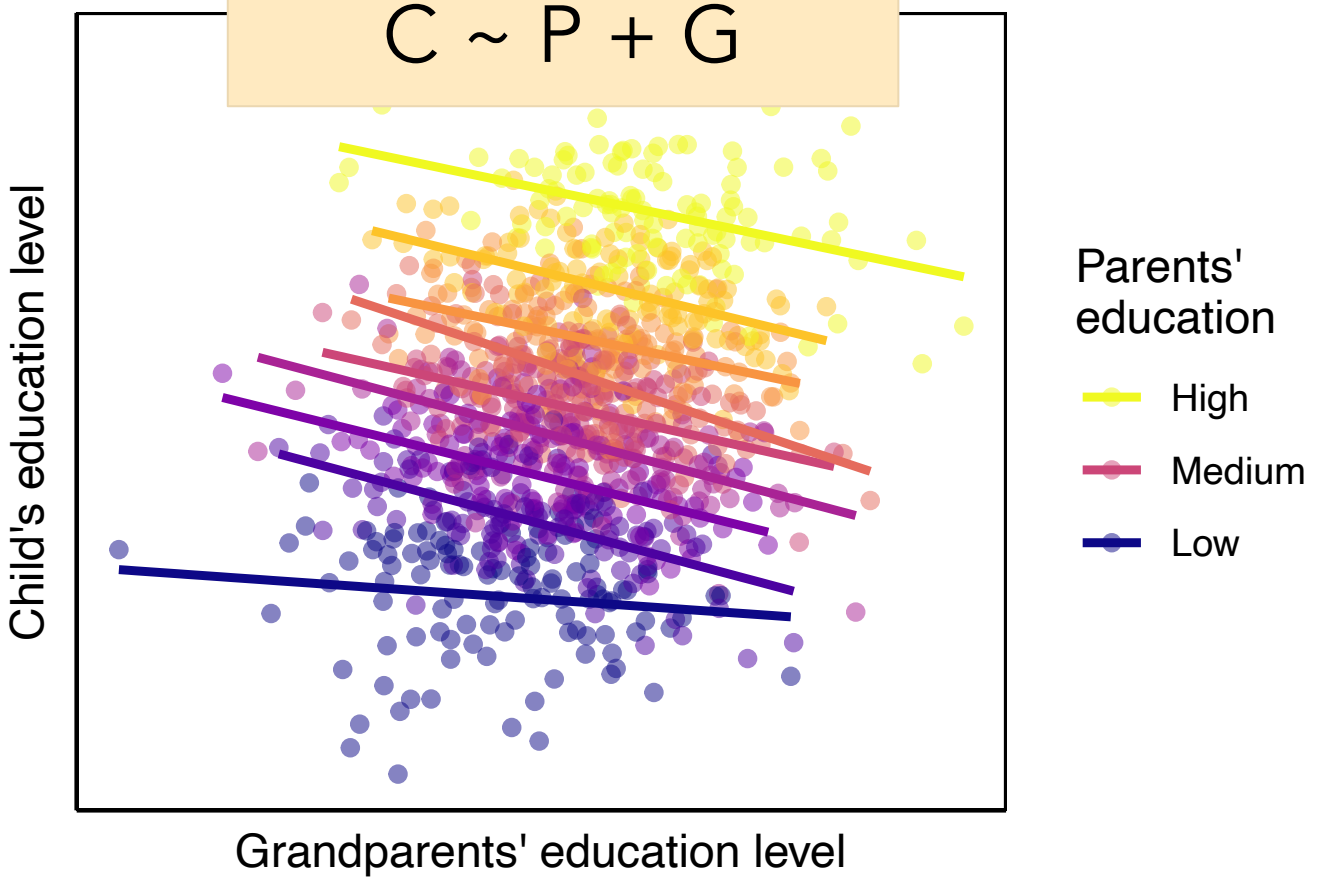
The Descendant



Conditioning on A is
like conditioning on Z

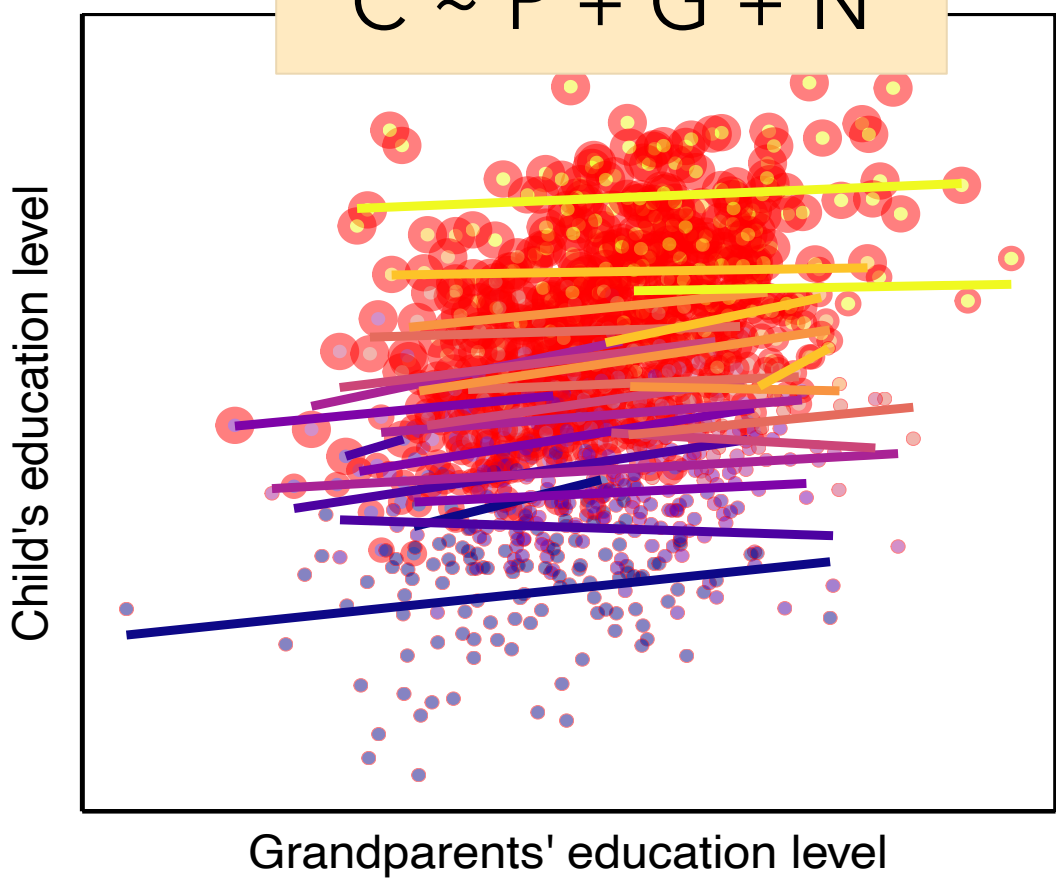
Example 3: Does your grandparent's education level affect your own?

Not ideal model:
C ~ P + G



Parents' education
High
Medium
Low

Ideal model
C ~ P + G + N

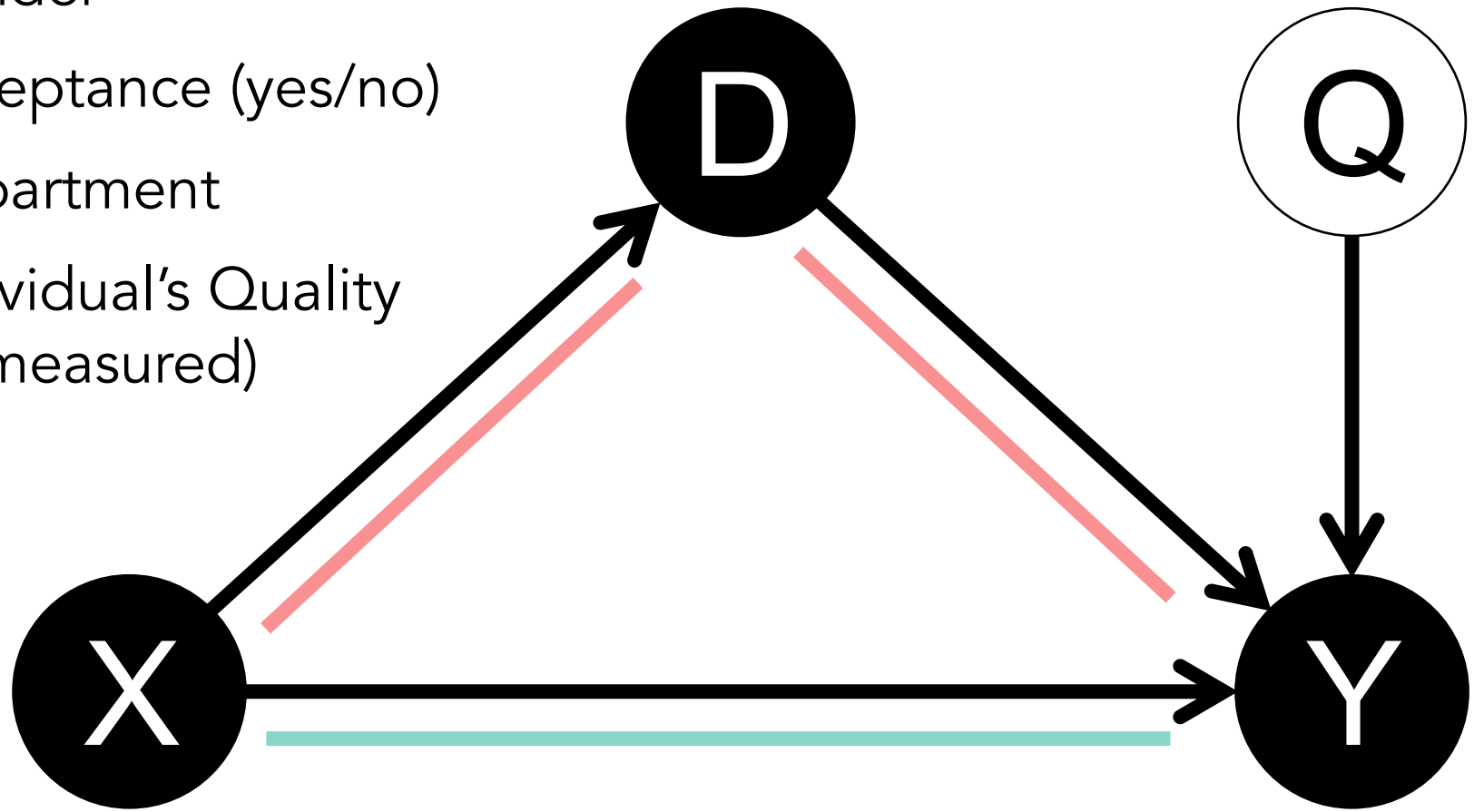


Chapter 5: Many paths…

Hidden paths are EVERYWHERE in the real world!

e.g., UC Berkley gender bias

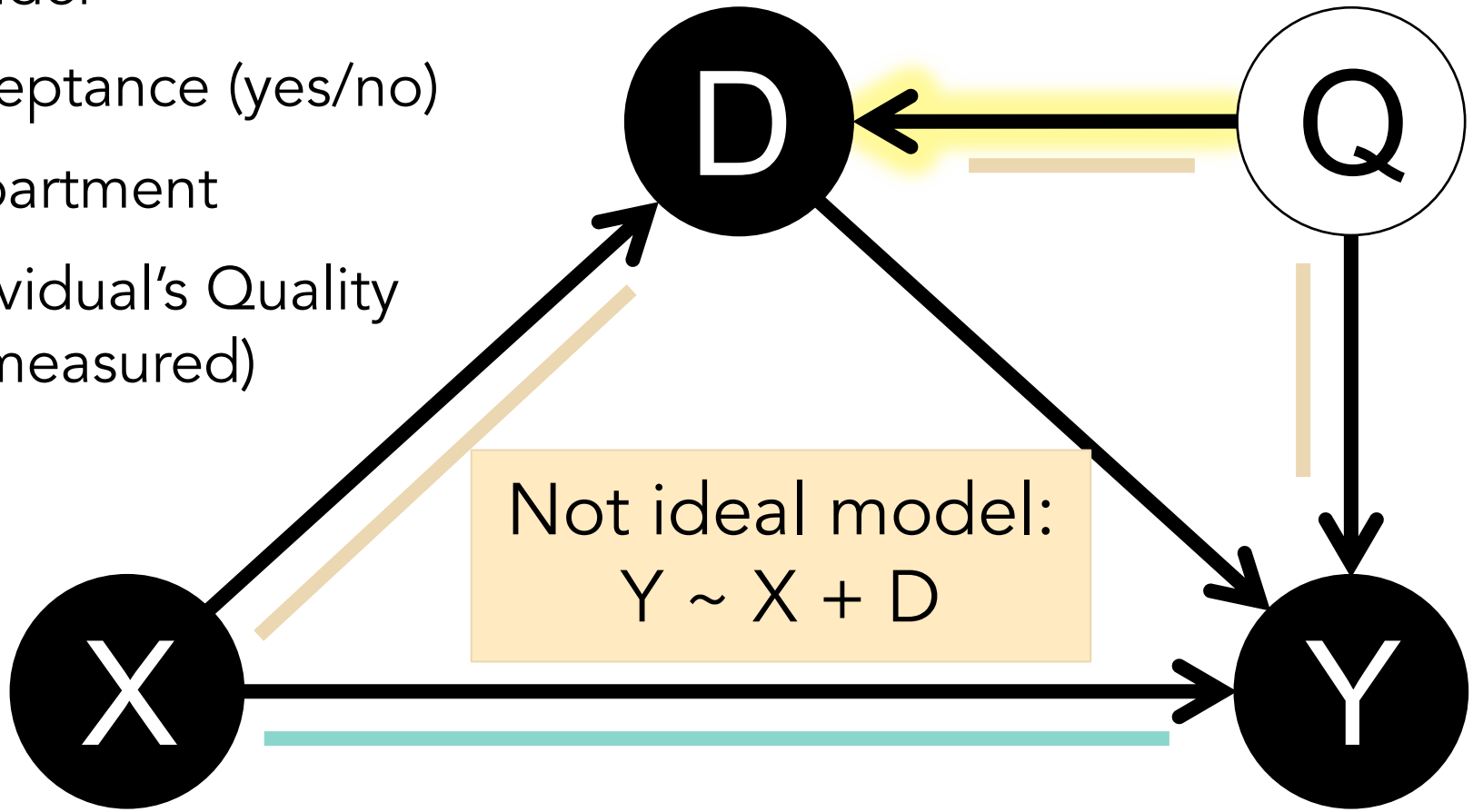
- X Gender
- Y Acceptance (yes/no)
- D Department
- Q Individual's Quality (unmeasured)



Hidden paths are EVERYWHERE in the real world!

e.g., UC Berkley gender bias

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



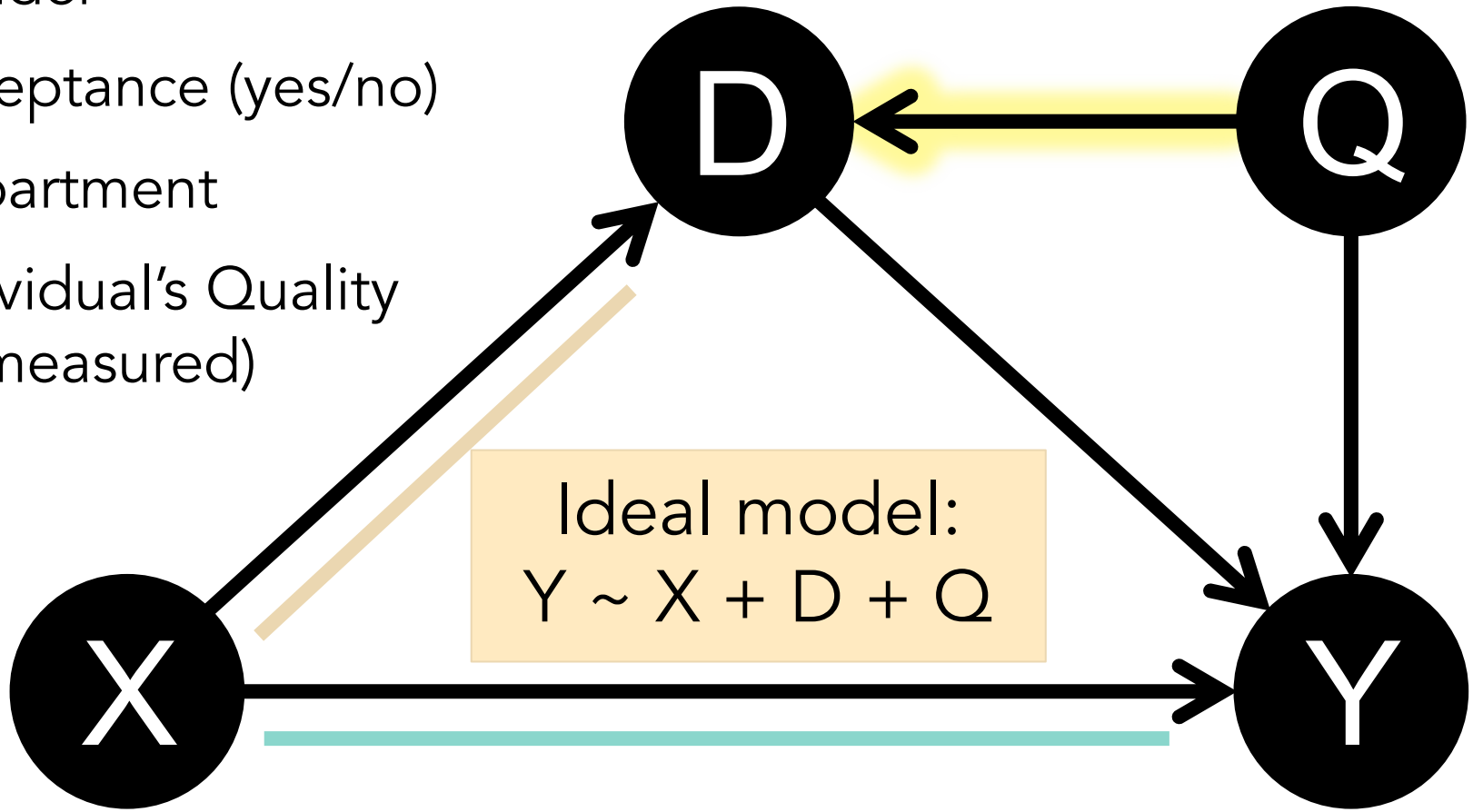
IF women are of higher quality, but this means they select better departments…

-> False gender bias created by conditioning on Department

Hidden paths are EVERYWHERE in the real world!

e.g., UC Berkley gender bias

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



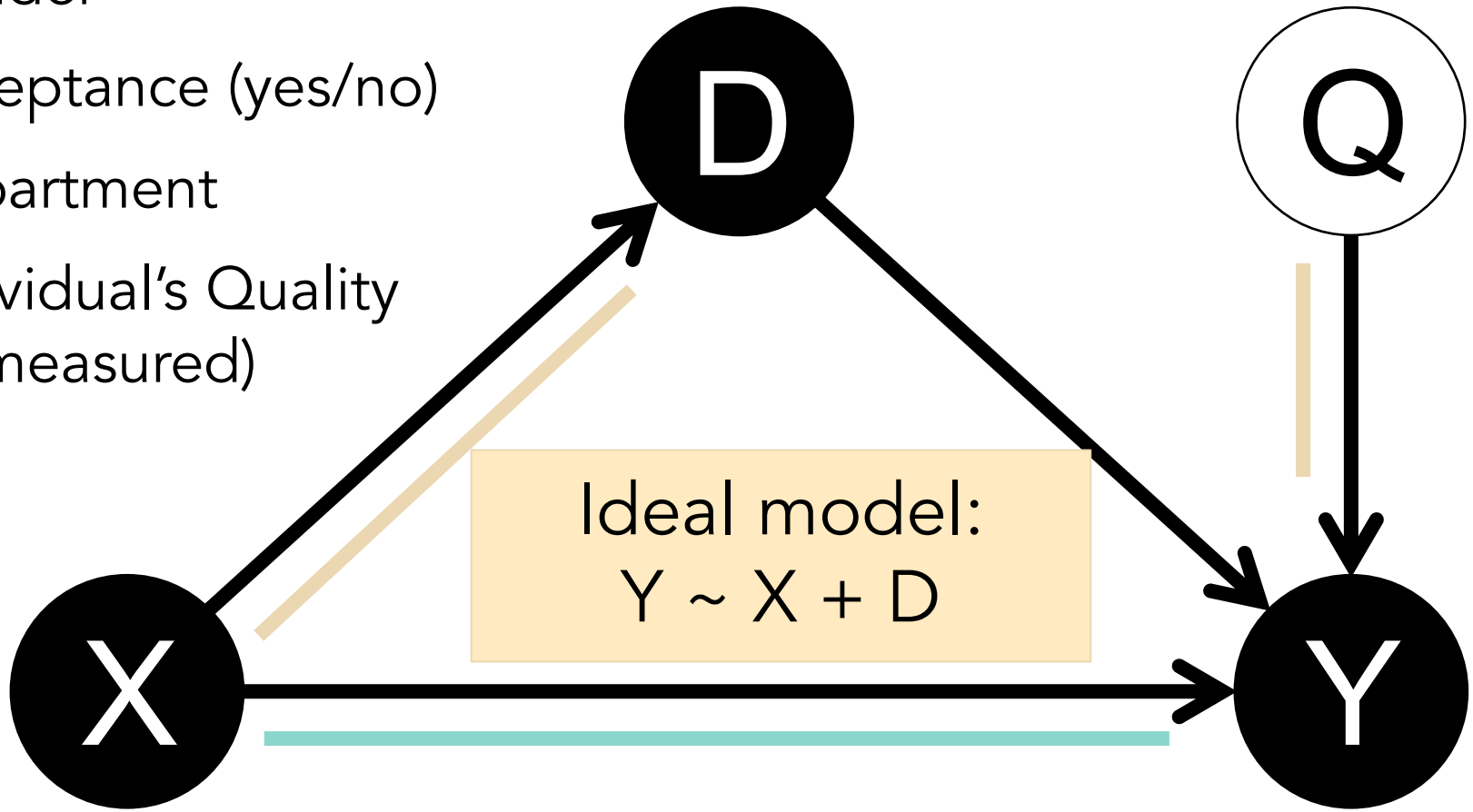
IF women are of higher quality, but this means they select better departments…

-> False gender bias created by conditioning on Department

Hidden paths are EVERYWHERE in the real world!

e.g., UC Berkley gender bias

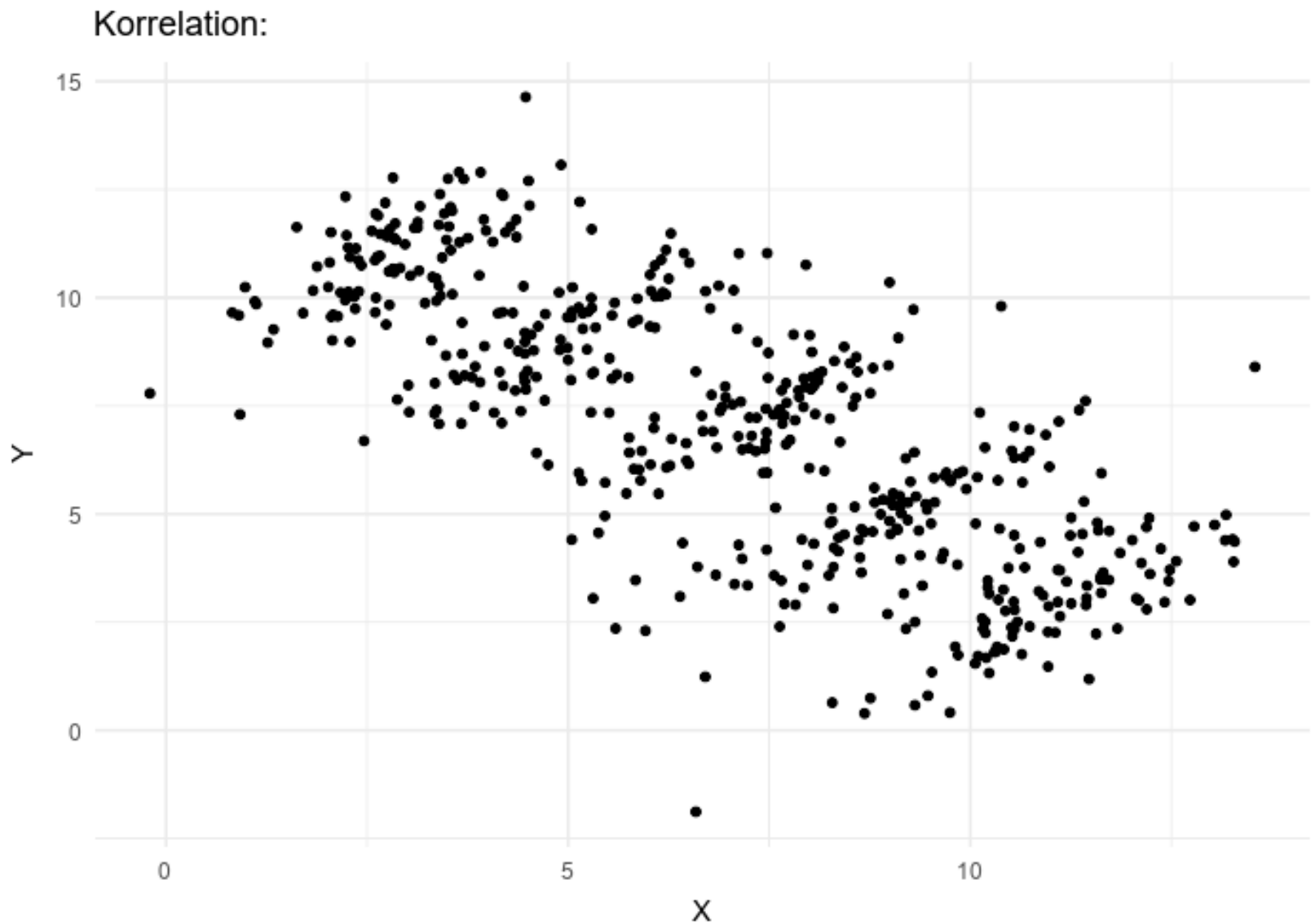
- X Gender
- Y Acceptance (yes/no)
- D Department
- Q Individual's Quality (unmeasured)



IF women are of higher quality, but this means they select better departments…

-> False gender bias created by conditioning on Department

Simpson's Paradox



“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

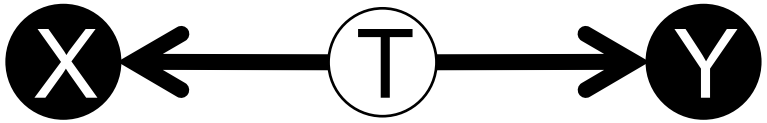
I cover the latest trends in science, tech and healthcare.

Follow

The Fork

X ← Z → Y

Open unless you condition on Z



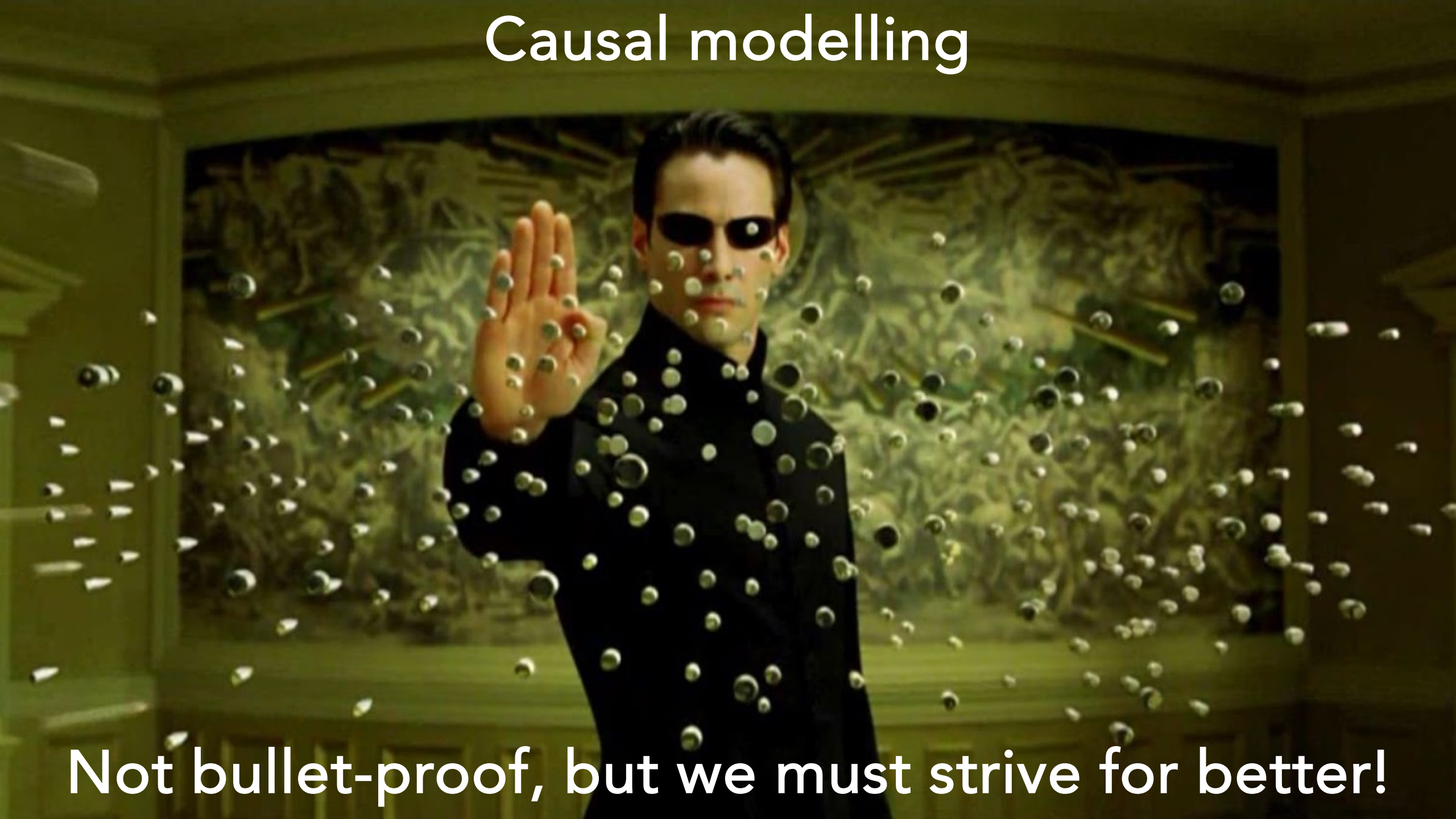
Take-aways:

- 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!

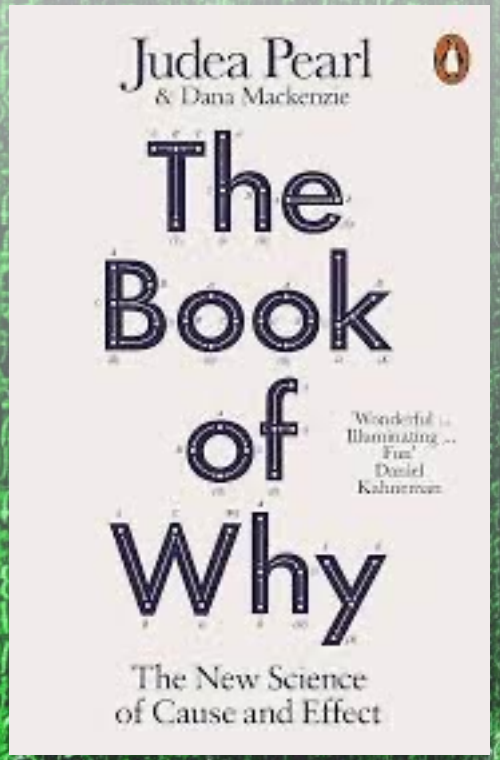
- When a confound cannot be known, turn to **path analysis**
  - SEMs: Structural Equation Modelling
  - Bayesian Networks
  - Estimate the influence of unknown/latent parameters

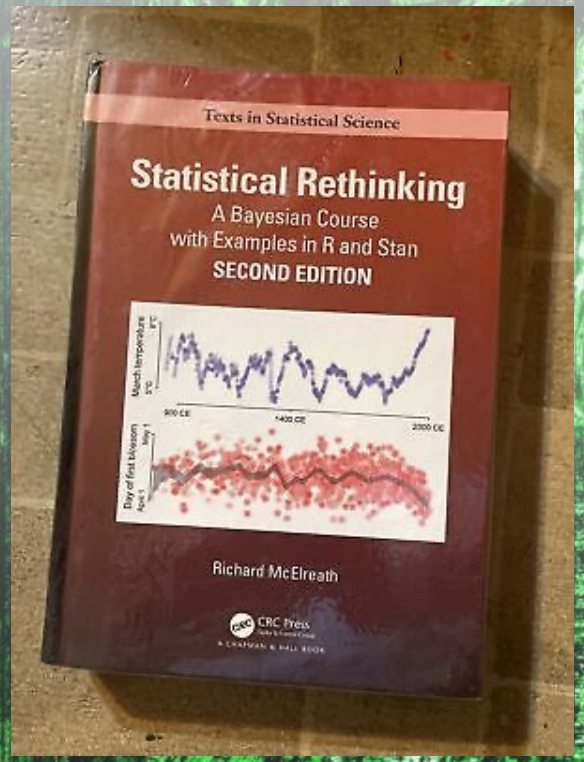
Causal modelling

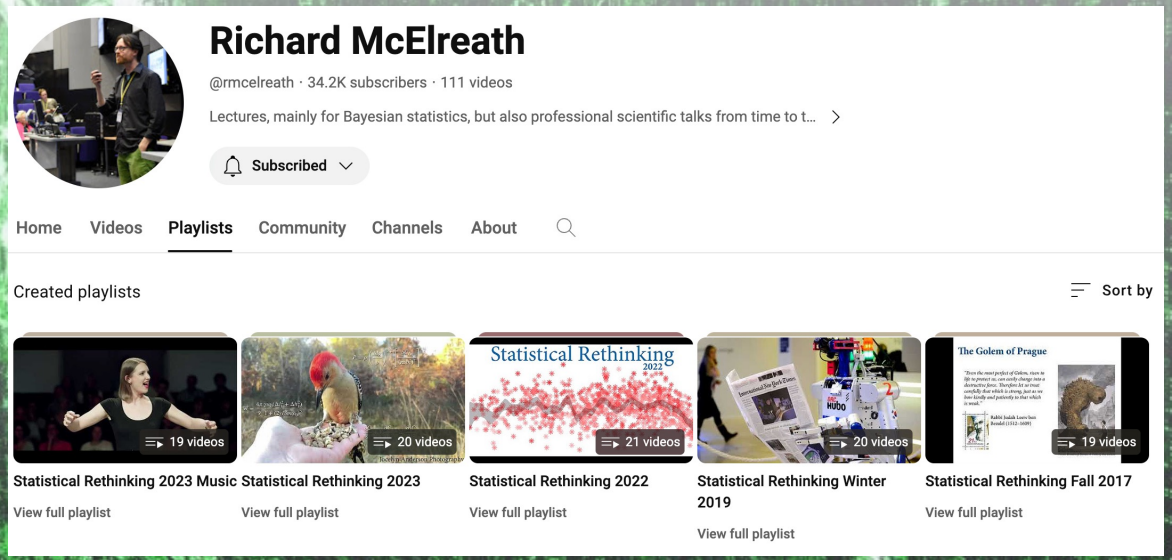


Not bullet-proof, but we must strive for better!

Thanks for listening!



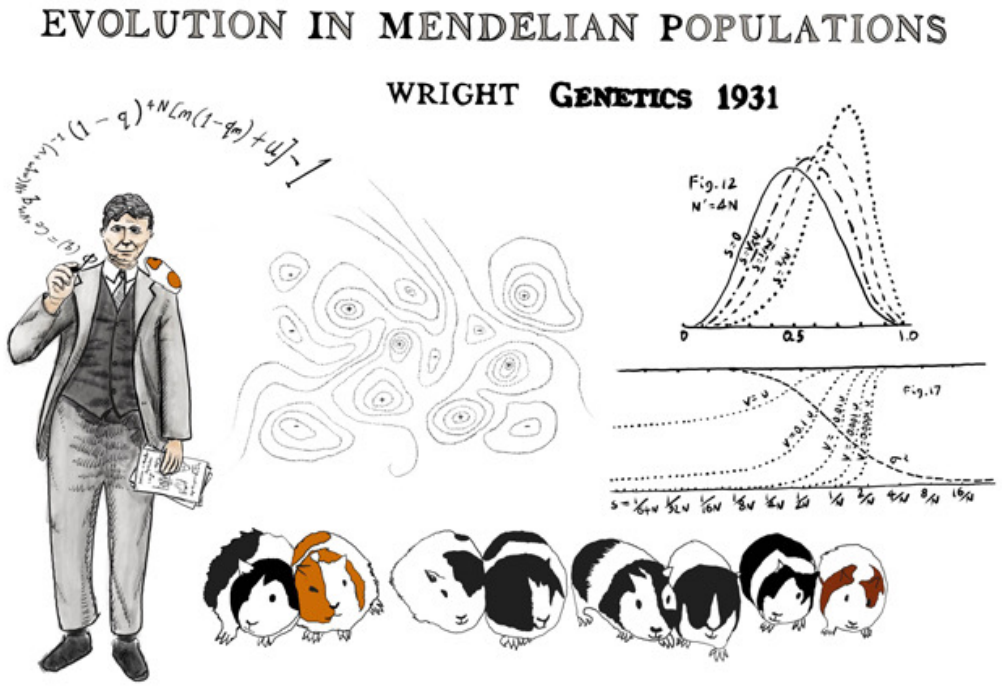


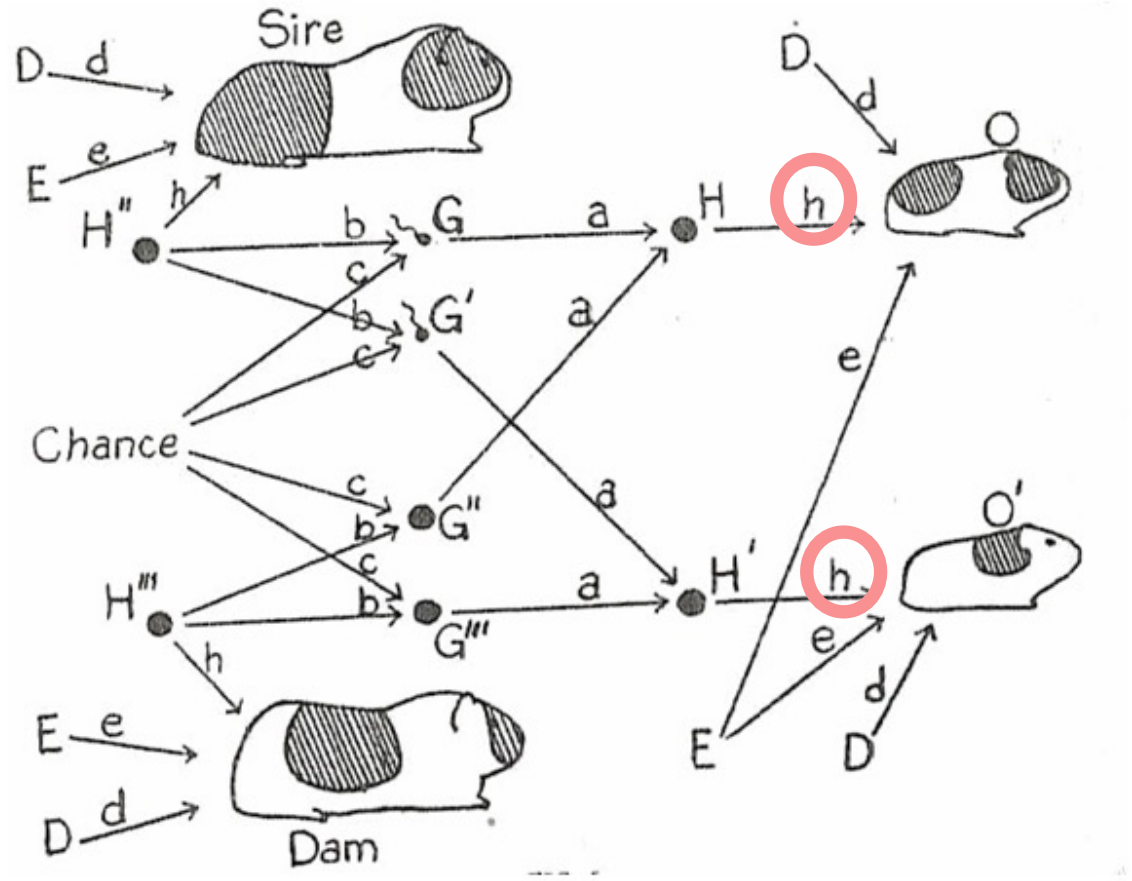



BAYES
FREQUENTISM


CAUSAL
INFERENCE

Sewall Wright's Path Analysis





- Path analysis allowed accurate estimates of trait heritability in the 1930s!
- Wright's work went ignored for years!