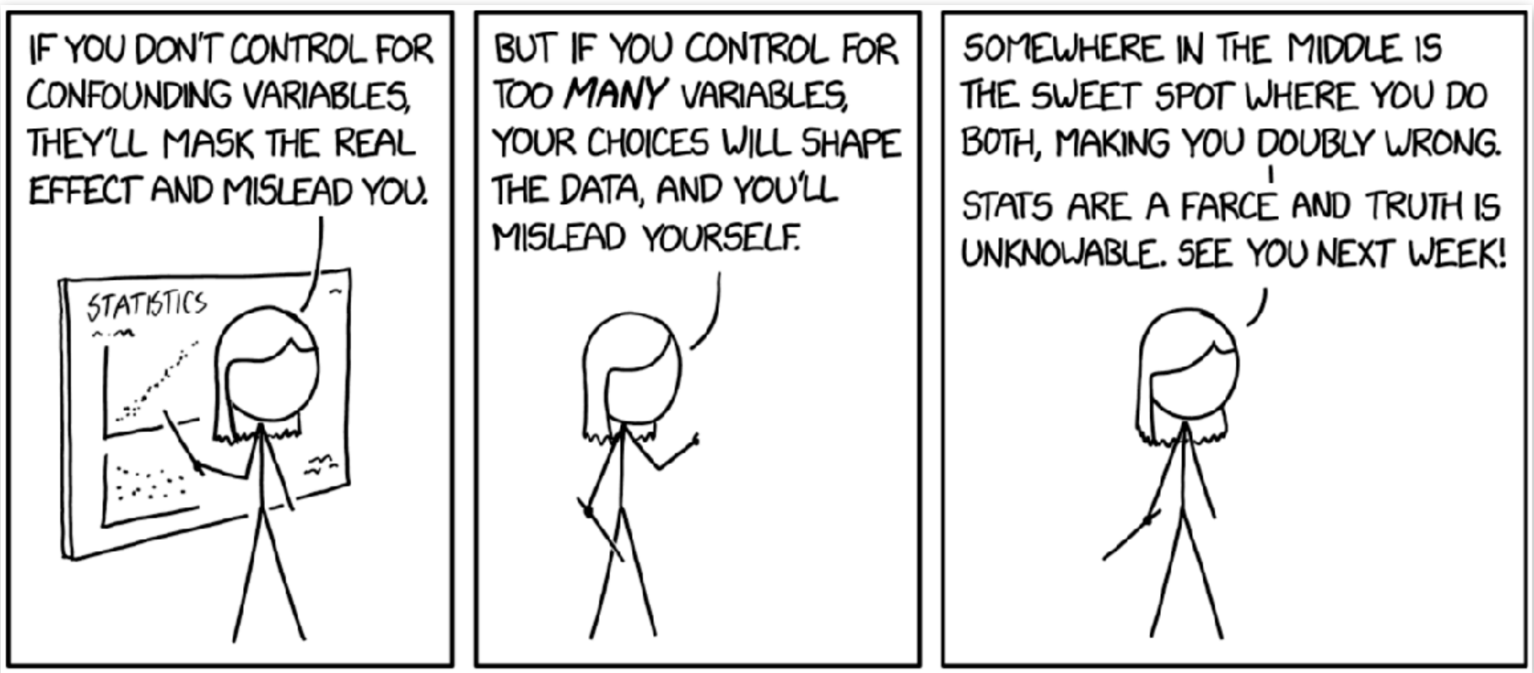
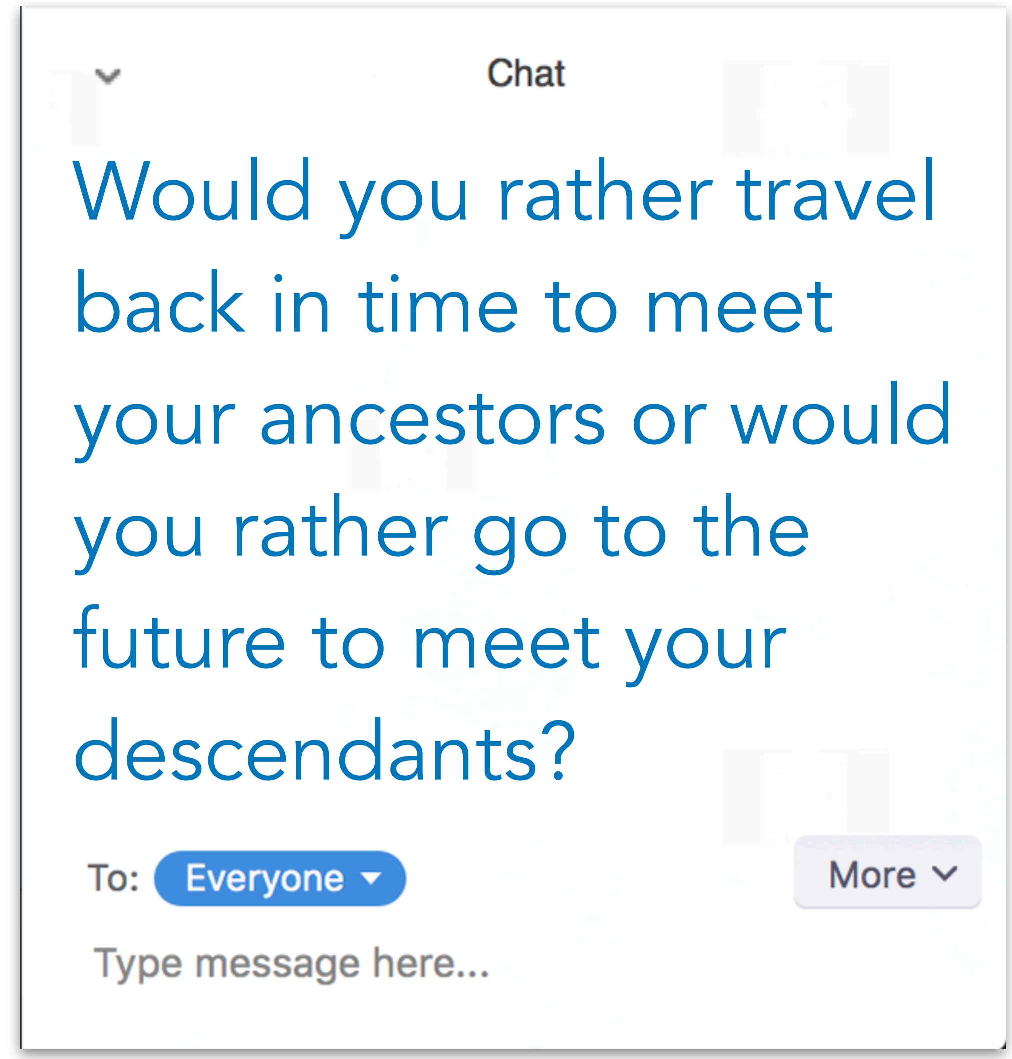
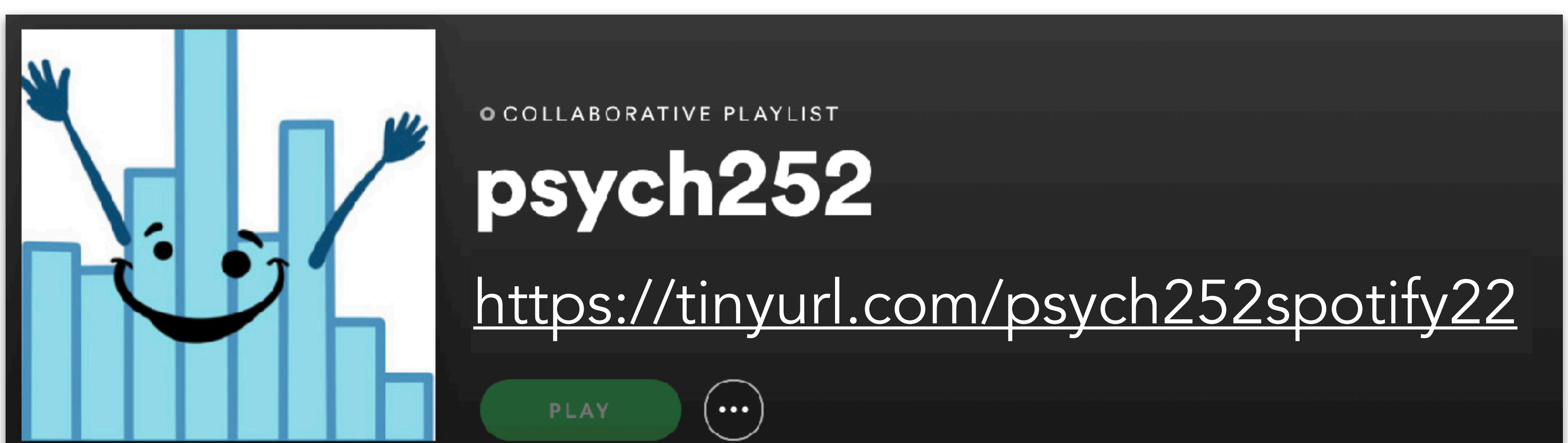
Causation





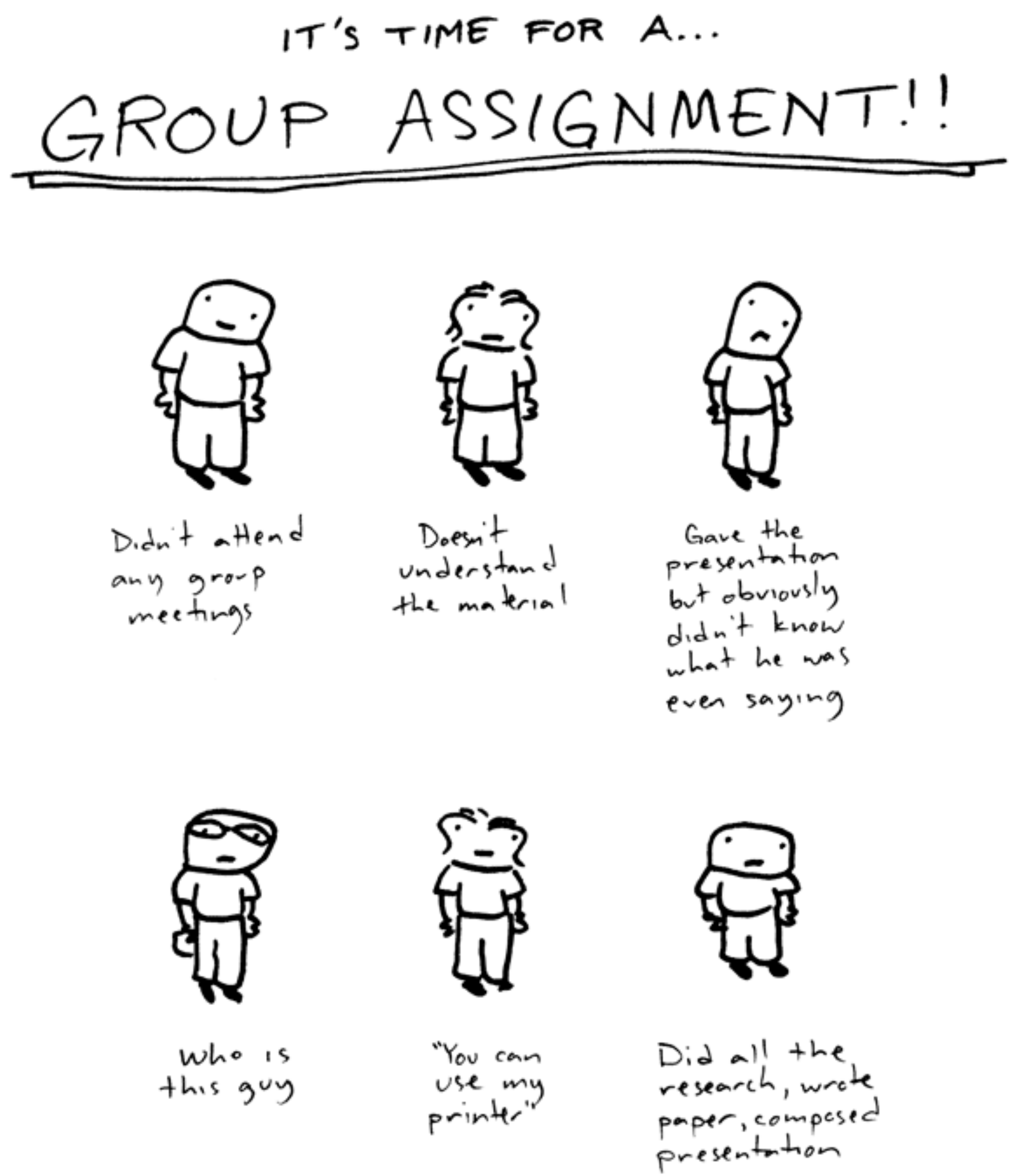


02/11/2022

Logistics

Project proposal

Project proposal

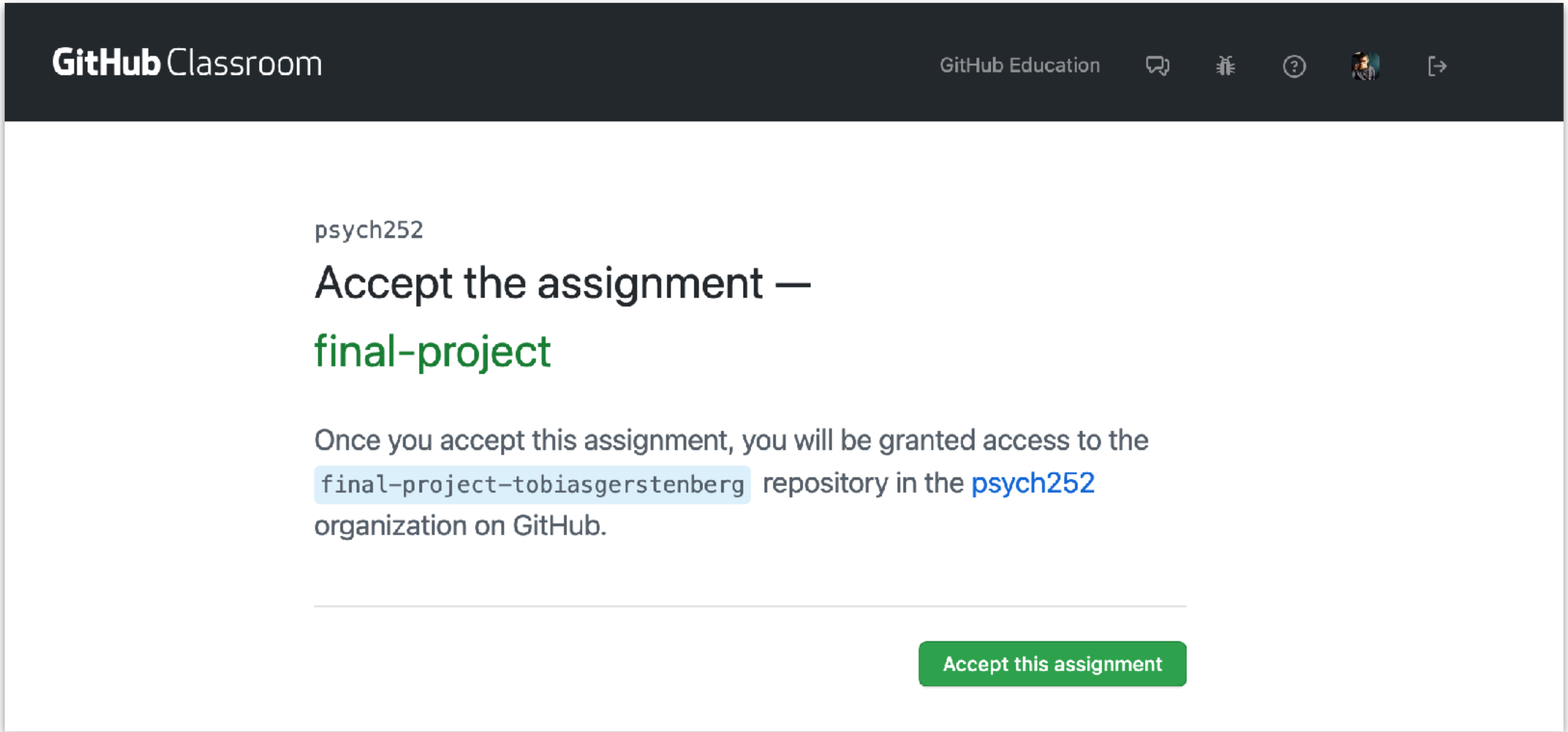


Toothpaste For Dinner.com

Members of teams will all the get same grade!

maximum 3 team members

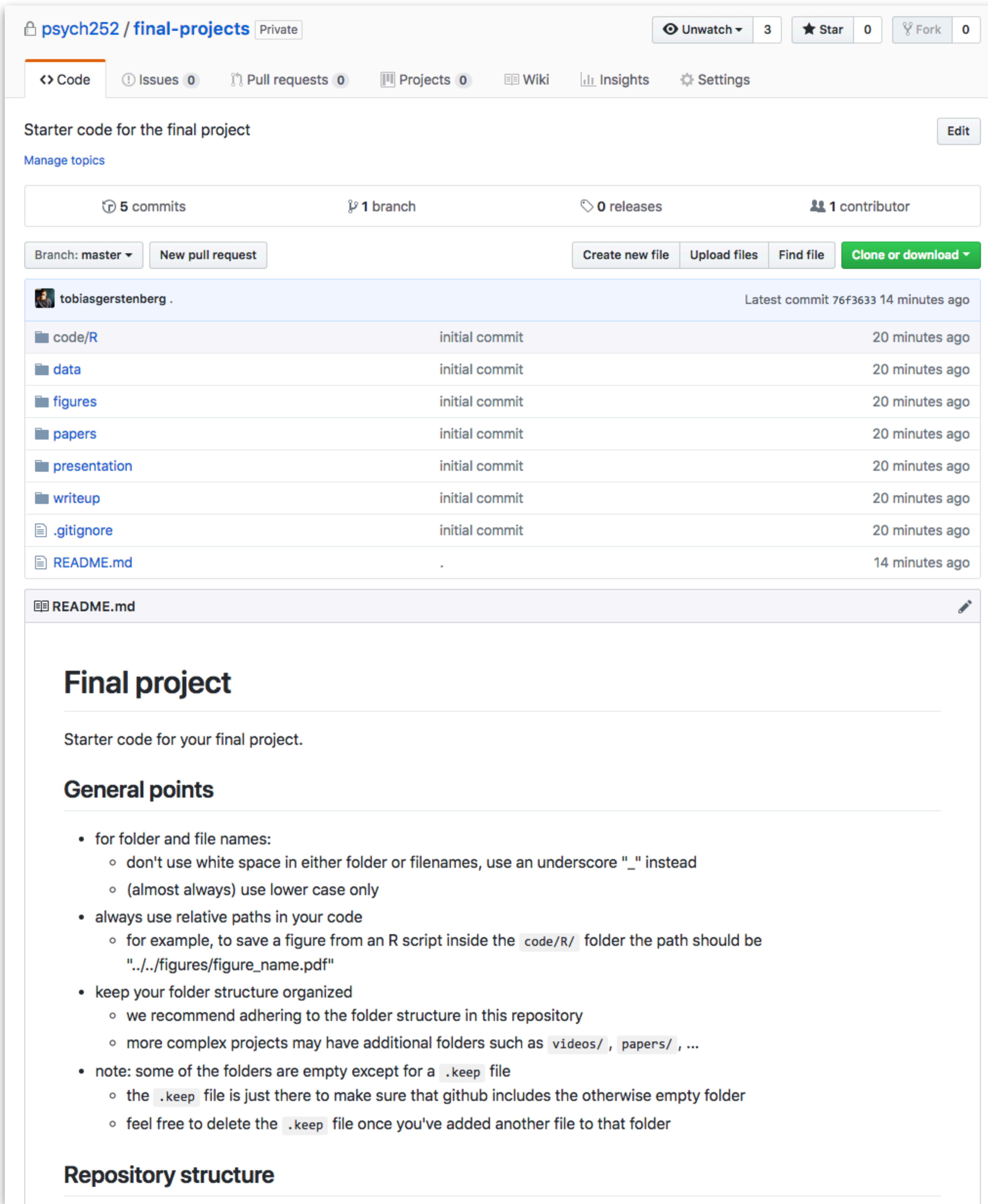
Project proposal



if you work as a team, have one person accept the assignment and the others can then join that repository

https://classroom.github.com/a/g3qRXbG_

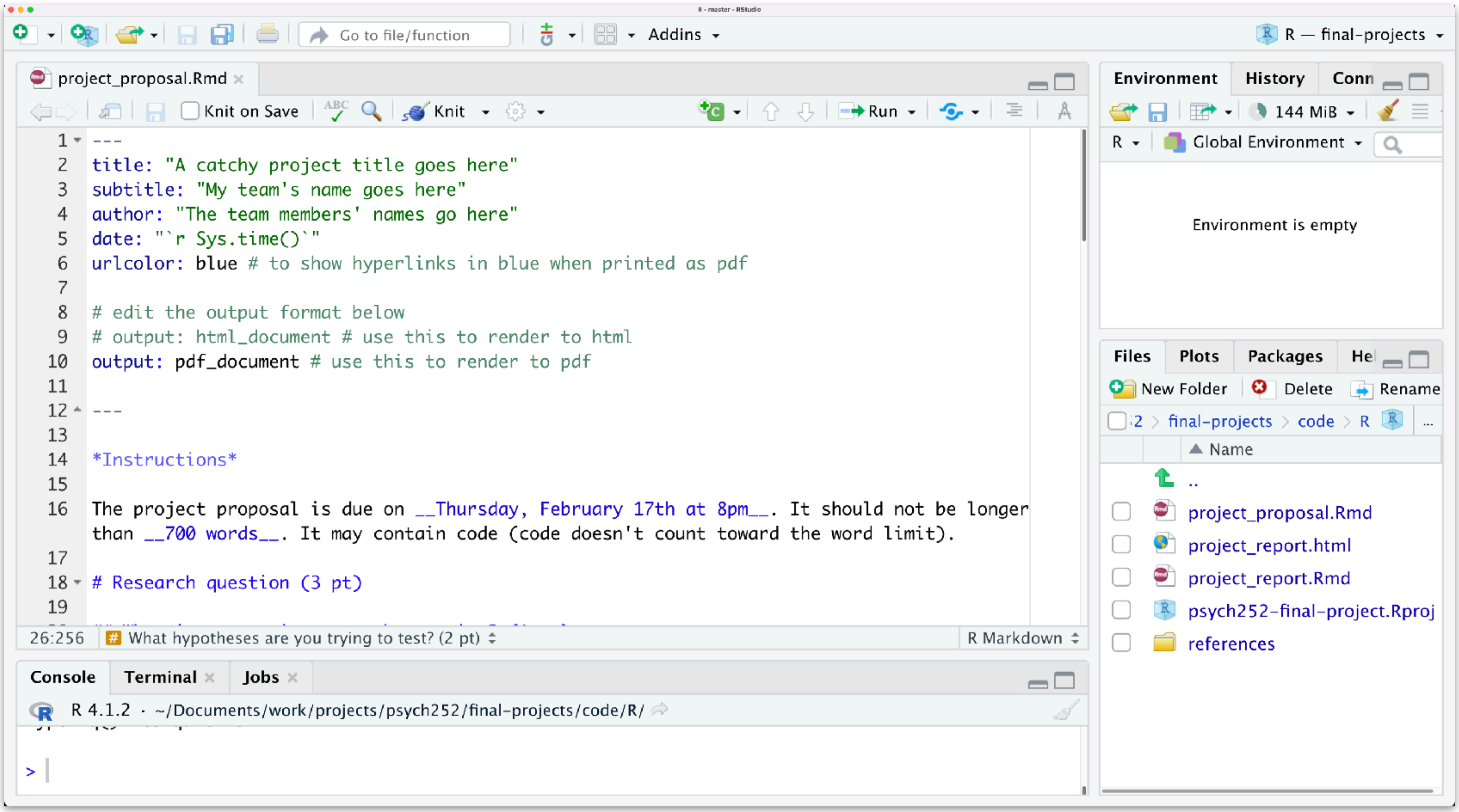
Project proposal



- each team will have their own private github repository
- all work on your final project should happen within this repository
- you can get **github** help in section
- post on Ed Discussion in case you experience any problems getting set up (use **final_project** tag)

Project proposal

RMarkdown template

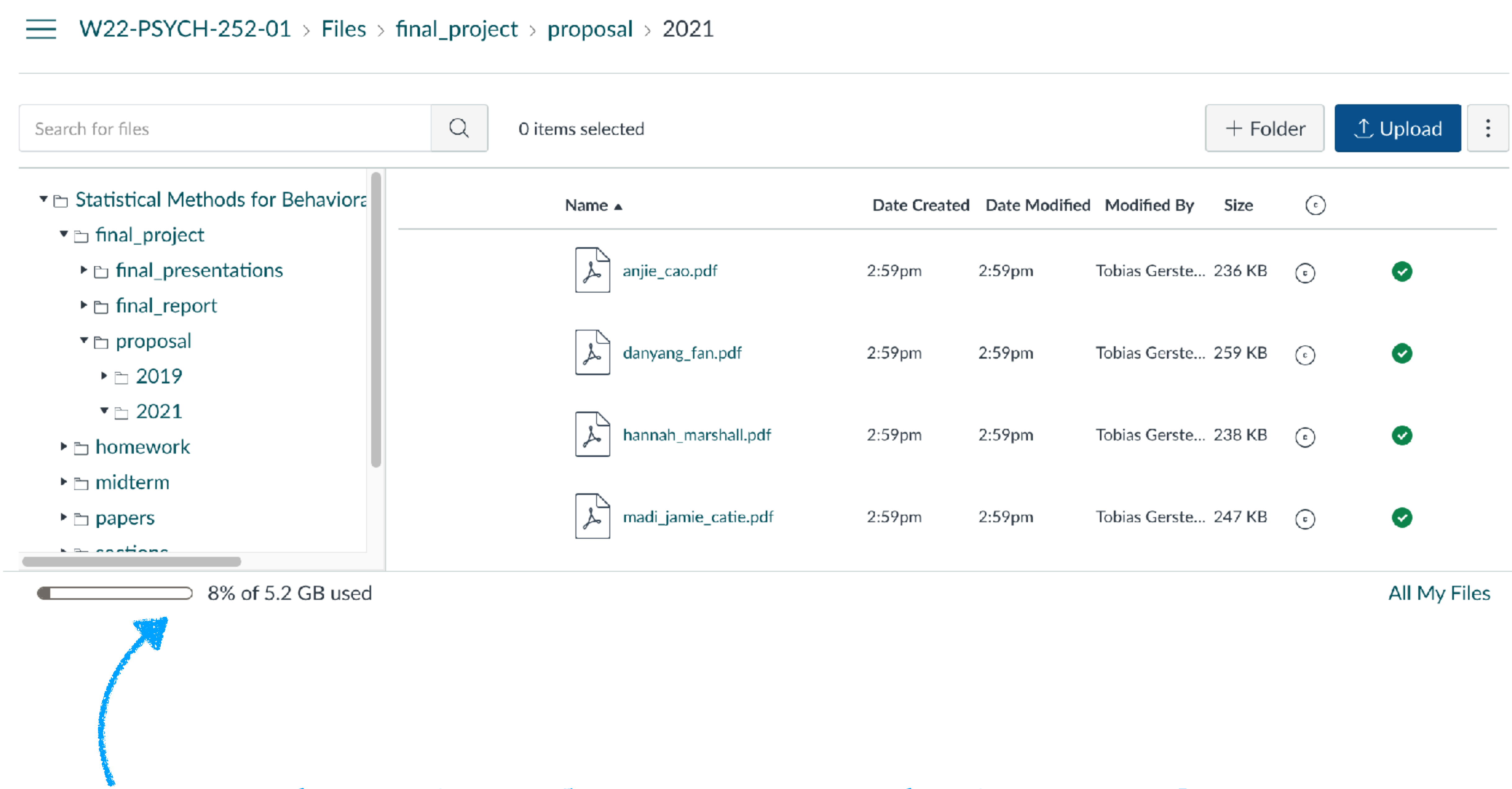


Upload the pdf to canvas

Project proposal

The project proposal is due on
Thursday, February 17th at 8pm

Project proposal



examples of project proposals from prior years

Plan for today

- Quick recap
- Model comparison
  - Cross-validation in action
  - AIC and BIC
- Causation vs. correlation
- Controlling for variables
- Mediation
- Moderation

Quick recap

Quick recap: Power analysis via simulation

Power simulation recipe

- assume:
  - α, *n*, effect size
- simulate a large number of data sets of size *n* with the specified effect size
- for each data set, run a statistical test to calculate the p-value
- determine the probability of rejecting the H_0 (given that H_1 is true)

13

Let's simulate ...

```r
# make reproducible
set.seed(1)

# number of simulations
n_simulations = 5

# run simulation
expand_grid(n = seq(10, 40, 2),
            simulation = 1:n_simulations,
            p = 0.75) %>%
  mutate(index = 1:n(),
         .before = n) %>%
  group_by(index, n, p, simulation) %>%
  mutate(response = rbinom(n = 1,
                           size = n,
                           prob = p),
         p.value = binom.test(x = response,
                              n = n,
                              p = 0.5,
                              alternative = "two.sided")$p.value) %>%
  group_by(n, p) %>%
  summarize(power = sum(p.value < 0.05) / n())
```

n	p	power
10	0.75	0.2
12	0.75	0.2
14	0.75	0.4
16	0.75	0.2
18	0.75	0.6
20	0.75	0.8
22	0.75	0.6
24	0.75	0.4
26	0.75	0.6
28	0.75	0.8
30	0.75	0.8

18

RStudio time

```r
map()
unnest()
```

21

Quick recap: Model comparison

The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
3. Calculate the proportional reduction of error (PRE) in our sample
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

23

Any problems with our approach?

sometimes it doesn't work ...

Model C	Model A
$\text{balance}_i = \beta_0 + \epsilon_i$	$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \beta_2 \cdot \text{age}_i + \epsilon_i$
$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \epsilon_i$	$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \beta_2 \cdot \text{age}_i + \epsilon_i$
$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \epsilon_i$	$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \epsilon_i$
$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{student}_i + \epsilon_i$	$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{degree}_i + \epsilon_i$

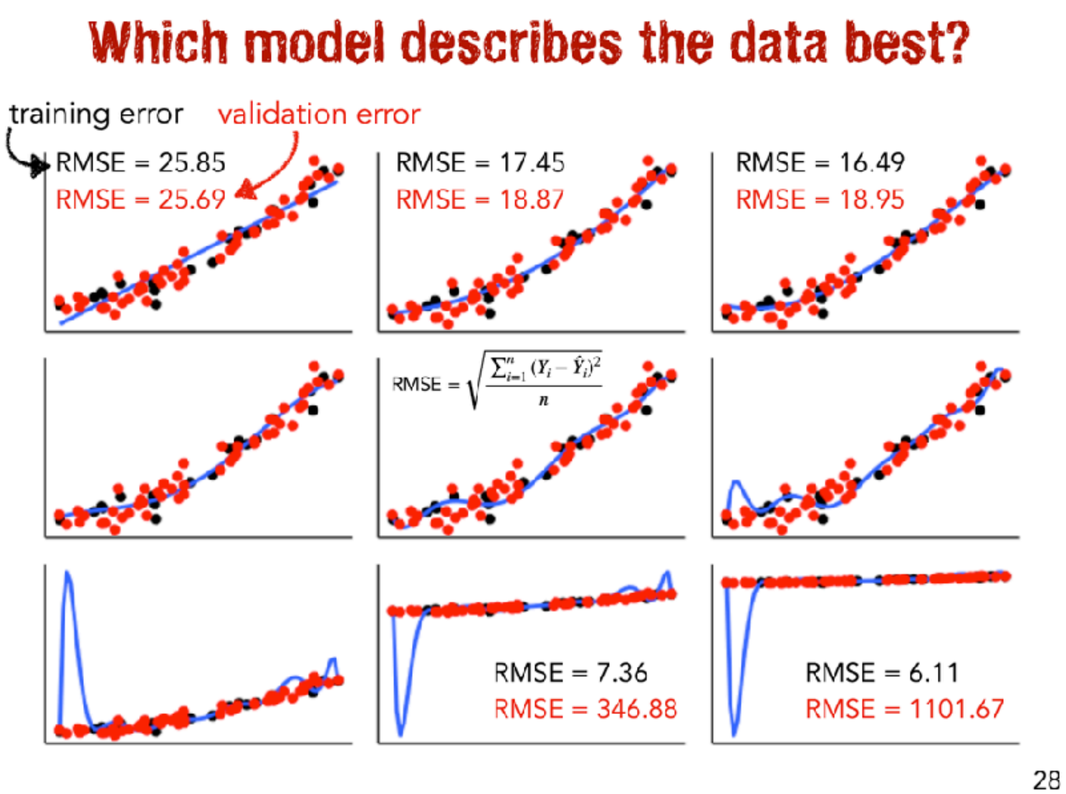
24

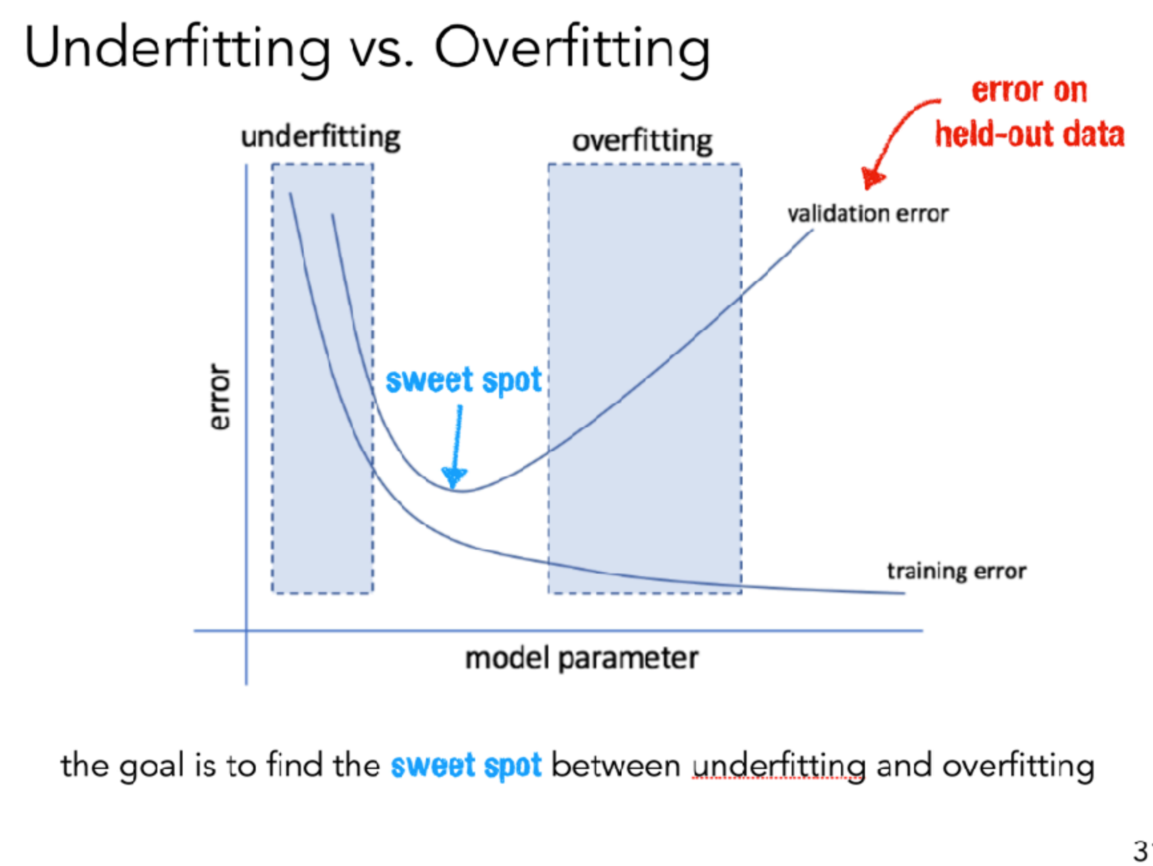
Tools for model comparison

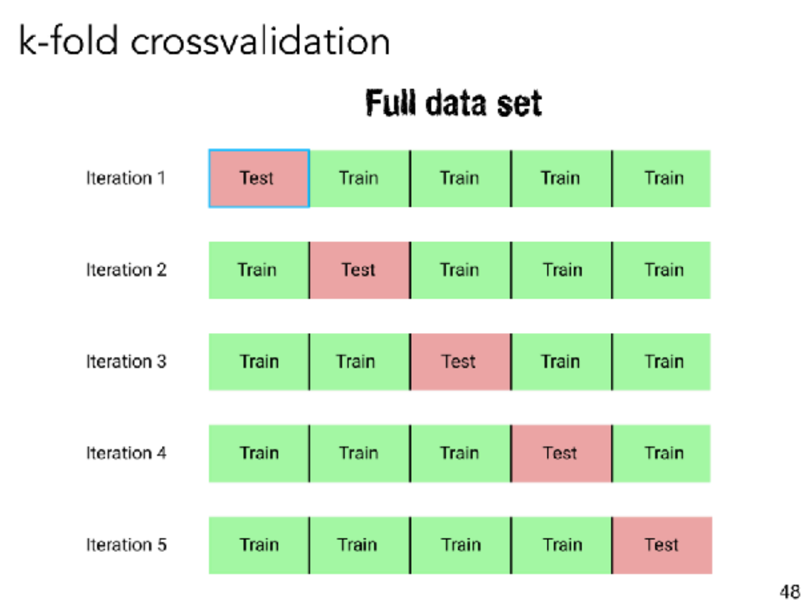
- **anova()**: compare a compact model with an augmented model via the F-test
  - problem: only works for nested models (where the augmented model contains all the predictors of the compact model and more)
- **What if we want to compare models that aren't nested?**
  - Cross-validation
  - AIC and BIC
  - Bayesian data analysis (we'll get there soon)

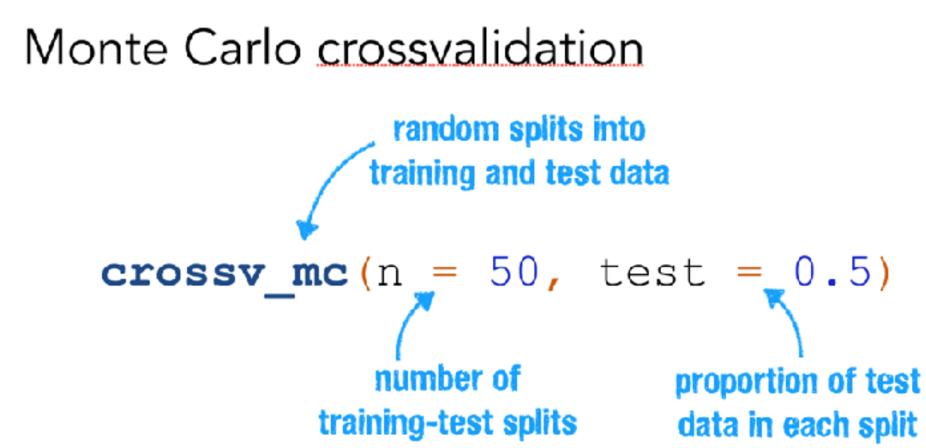
25

Quick recap: Cross-validation









Model comparison

Cross-validation in action

RStudio®

time

AIC and BIC

AIC and BIC

- AIC = Akaike Information Criterion
- BIC = Bayesian Information Criterion

not that much Bayesian about it ...

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

?

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

$\hat{L}$ = maximized value of the likelihood function of the model

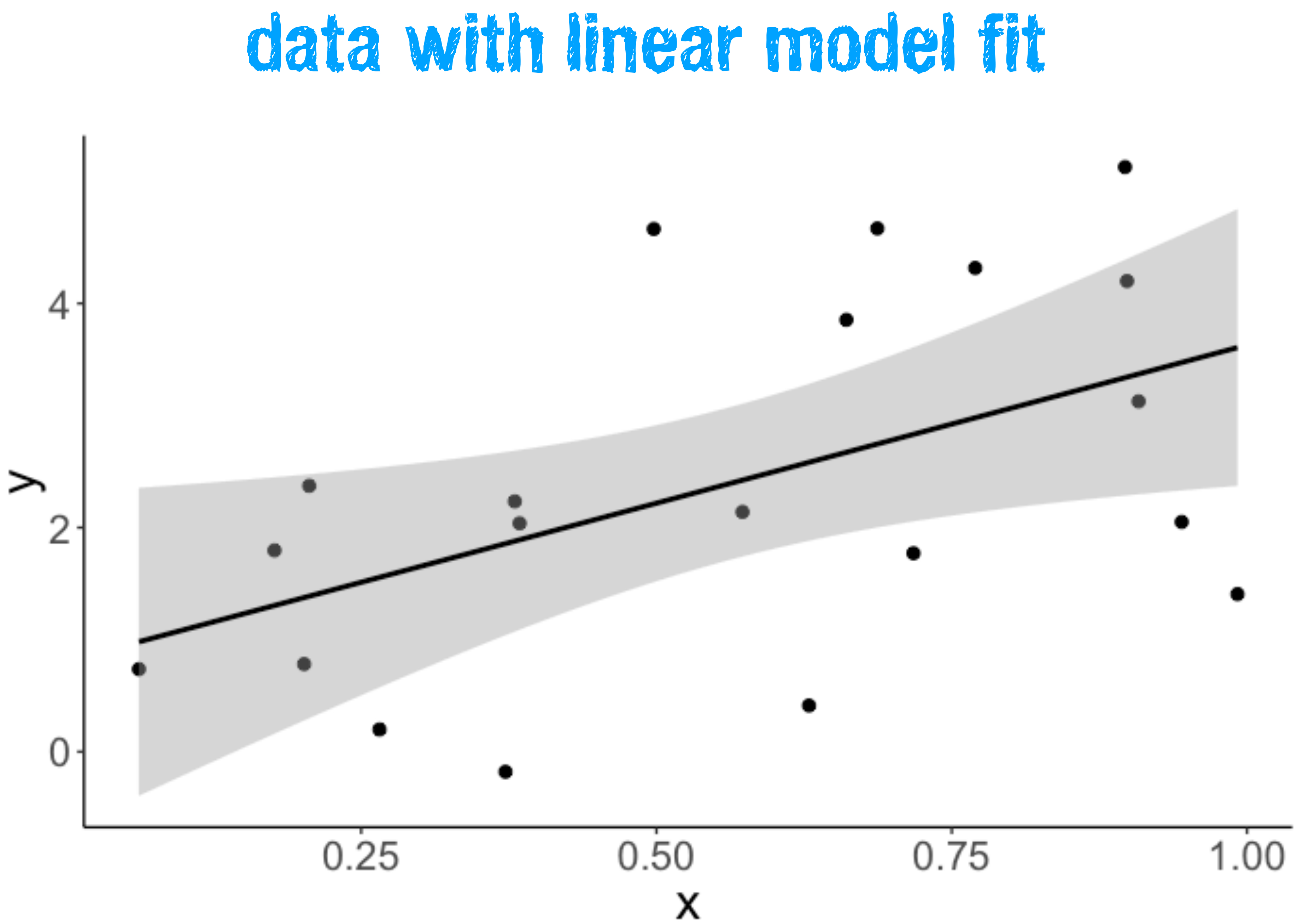
k = number of parameters in the model

n = number of observations

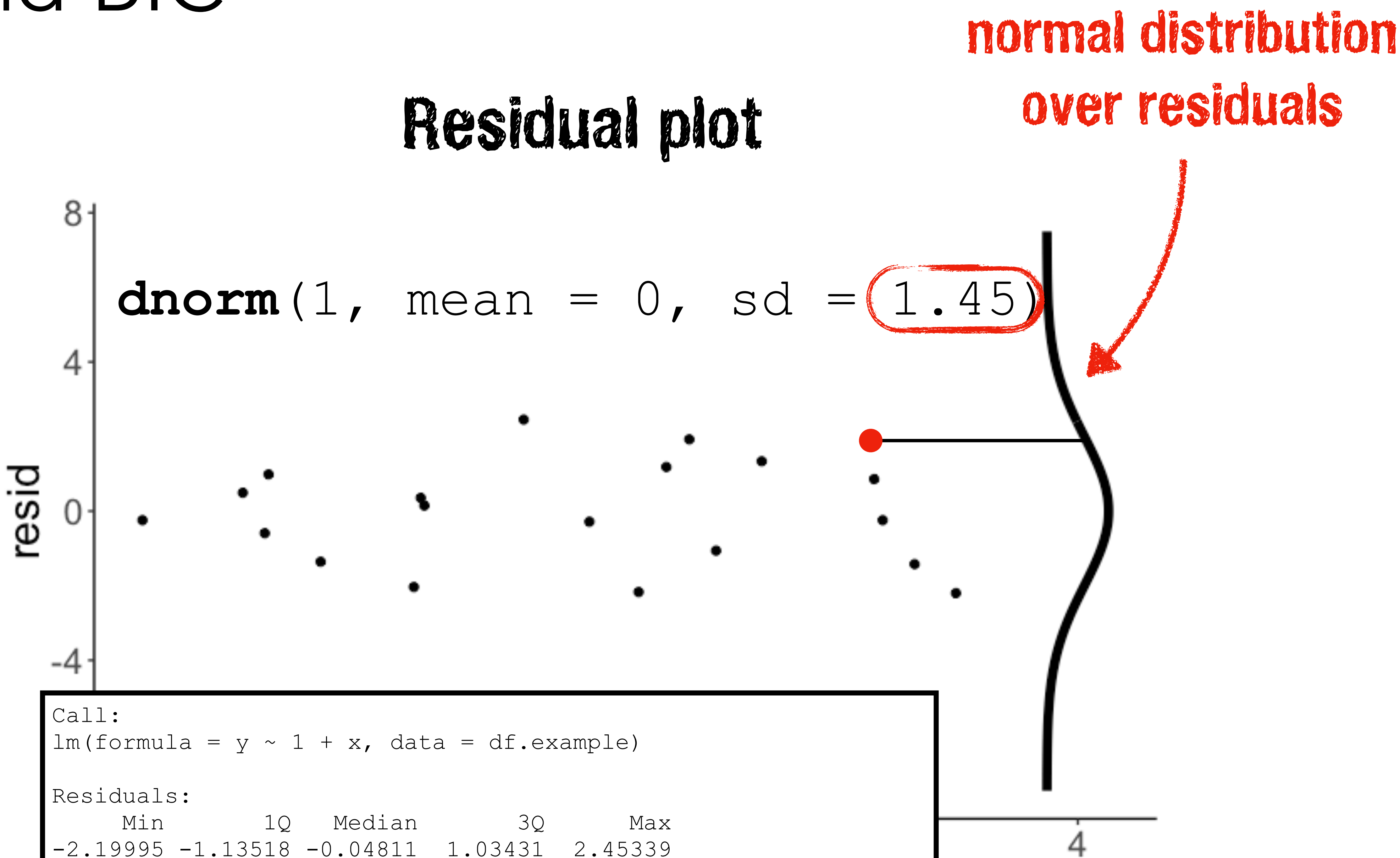
AIC and BIC

- How do we get the likelihood of our model?
  - in a linear regression, minimizing least squares is equivalent to maximizing the likelihood of the data given the model
- Assumptions of the linear model:
  - residuals are normally distributed with:
    - `mean = 0` and `sd = sigma`
  - calculate overall likelihood by computing the likelihood of each residual, and then multiplying

AIC and BIC

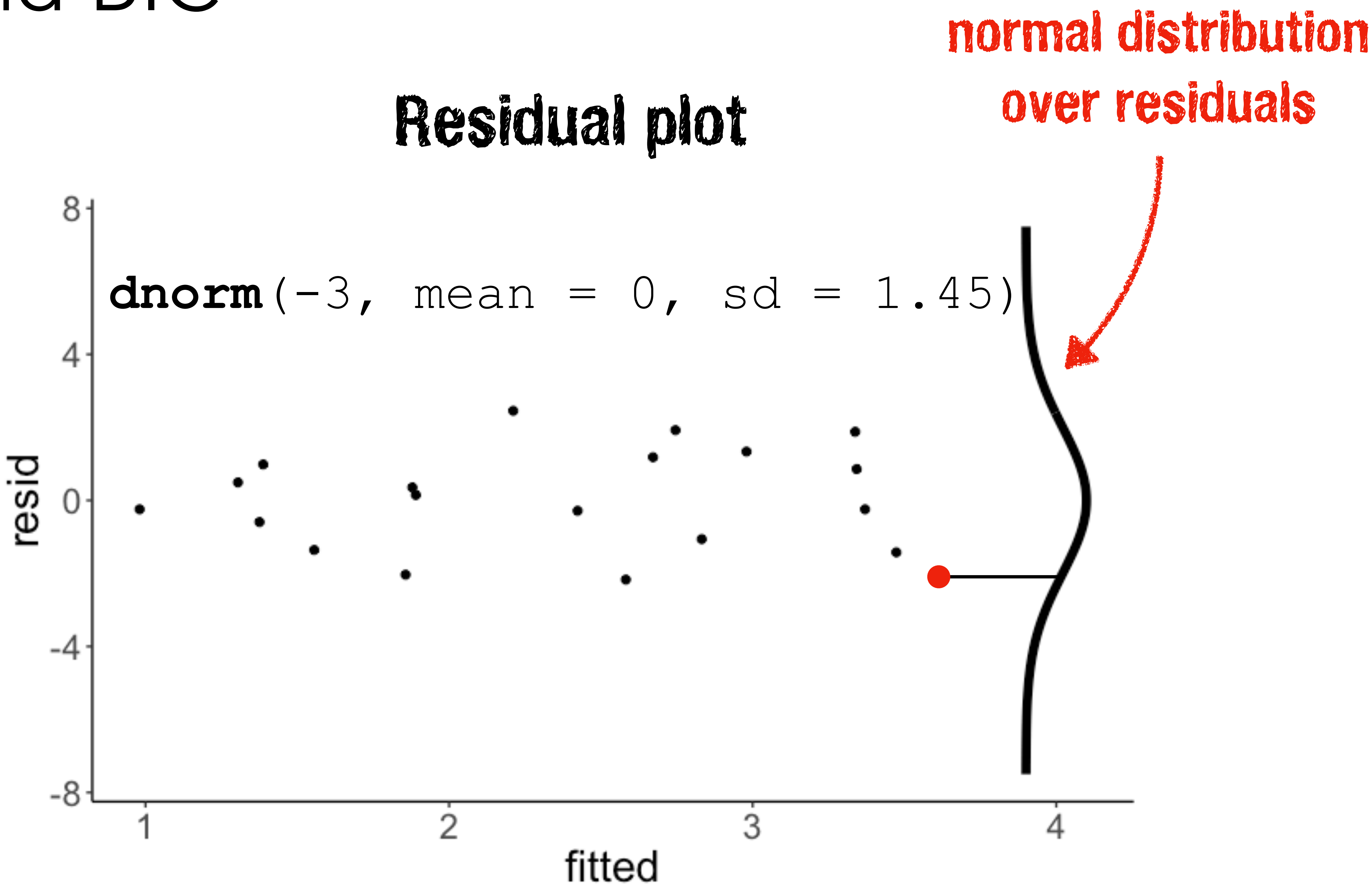


AIC and BIC




```
Call:
lm(formula = y ~ 1 + x, data = df.example)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19995 -1.13518 -0.04811  1.03431  2.45339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8062     0.7157   1.126   0.2748
x             2.8232     1.1374   2.482   0.0232 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.449 on 18 degrees of freedom
Multiple R-squared:  0.255, Adjusted R-squared:  0.2136
F-statistic: 6.161 on 1 and 18 DF,  p-value: 0.02315
```

AIC and BIC



since the data points are independent, we can calculate the overall likelihood by multiplying the likelihood of each observation

AIC and BIC

```
dnorm(1.88, mean = 0, sd = 1.45) = 0.12
```

```r
# generate some data
df.like = tibble(
  x = runif(20, min = 0, max = 1),
  y = 1 + 3 * x + rnorm(20, sd = 2)
)

# fit the model
fit = lm(formula = y ~ x,
         data = df.like)

# model summary
fit %>%
  glance()
```

x	y	fitted	resid	likelihood
0.90	5.22	3.34	1.88	0.12
0.27	0.20	1.56	-1.36	0.18
0.37	-0.18	1.86	-2.04	0.10
0.57	2.14	2.42	-0.28	0.27
0.91	3.13	3.37	-0.24	0.27
0.20	0.78	1.38	-0.59	0.25
0.90	4.20	3.34	0.86	0.23
0.94	2.05	3.47	-1.42	0.17
0.66	3.85	2.67	1.18	0.20
0.63	0.41	2.58	-2.17	0.09

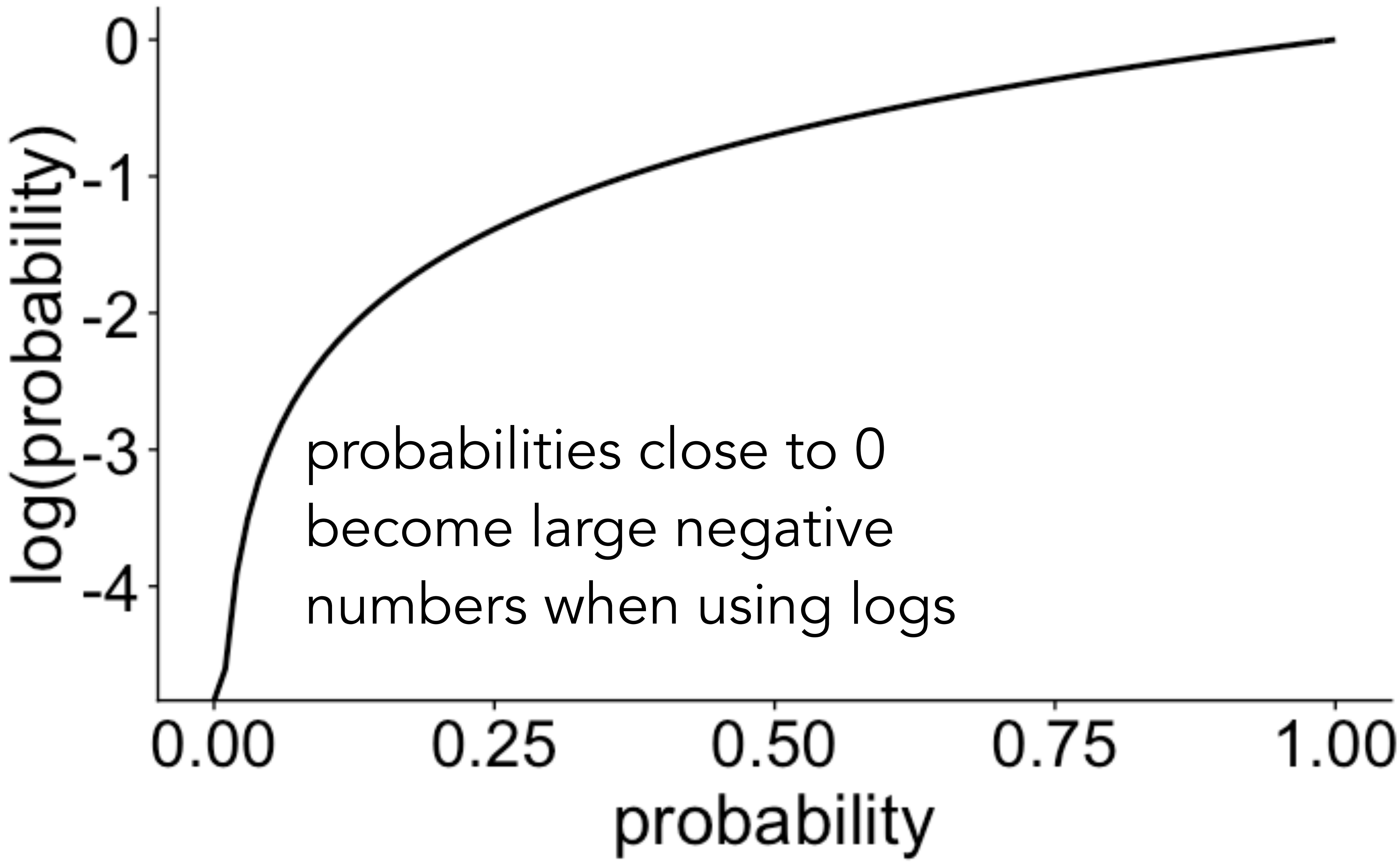
inferred standard deviation of the error

$$\sum_{i=1}^{n} \ln(\text{likelihood})$$

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.25	0.21	1.45	6.16	0.02	2	-34.74	75.47	78.46	37.77	18

$$e \sim \mathcal{N}(\text{mean} = 0, \text{sd} = 1.45)$$

`log()` is your friend!



Clue guide to probability

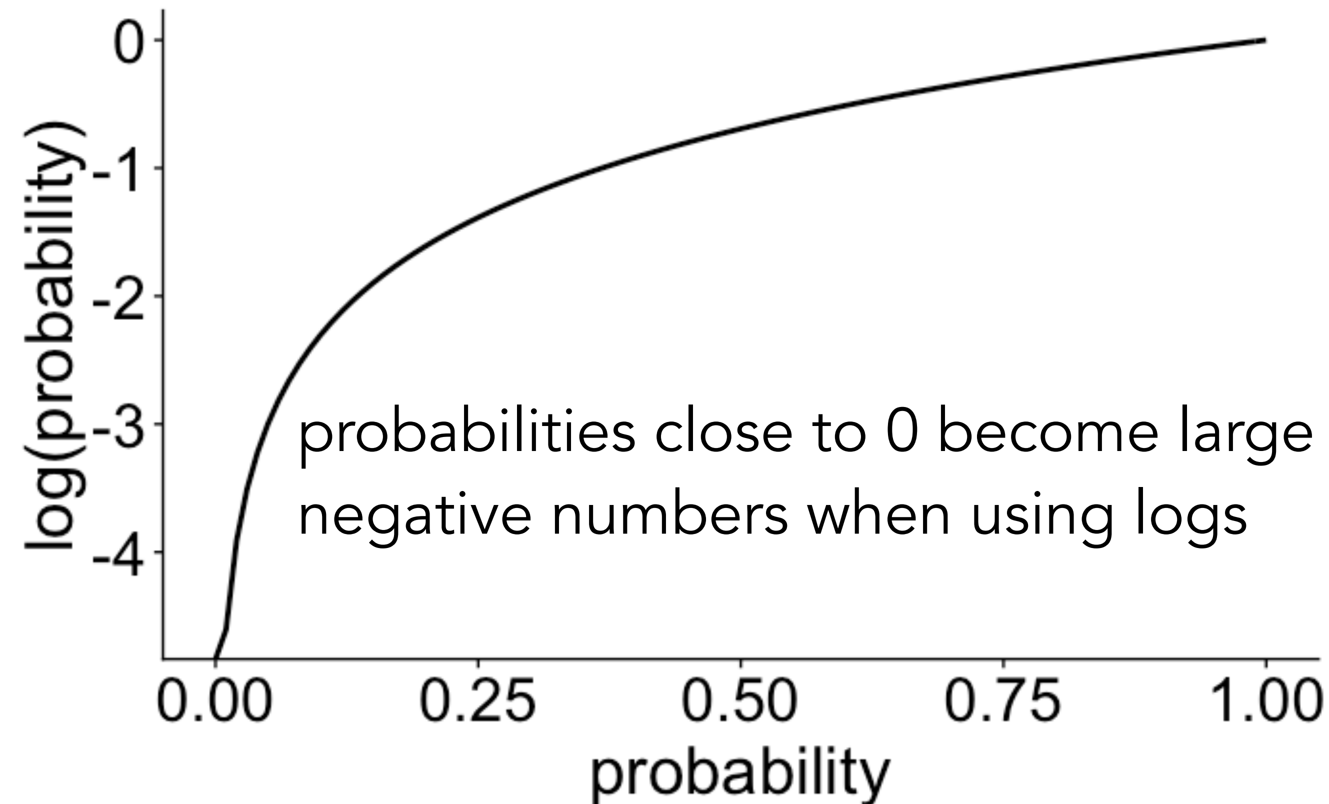
- *joint probability*:
- if A and B are independent then
- **Definition**: $p(A, B) = p(A) \cdot p(B)$

- $p(\text{Prof Plum}, \text{candle stick}) =$
  $p(\text{Prof Plum}) \cdot p(\text{candle stick}) =$
  $$\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Who?

What?

`log()` is your friend!



multiplying probabilities

$$0.01 \cdot 0.01 \cdot 0.01 \cdot 0.01 = 0.00000001$$

number becomes extremely small quickly

take log()

```
log(0.01)= -4.60517
```

number becomes large but that's ok

summing logs

$$(-4.60517) + (-4.60517) + (-4.60517) + (-4.60517) = -18.42068$$

transform back into probability

```
exp(-18.42068)= 0.00000001
```

often not necessary since we just use logLikelihood

AIC and BIC

dnorm(1.88, mean = 0, sd = 1.45) = 0.12

```r
# generate some data
df.like = tibble(
  x = runif(20, min = 0, max = 1),
  y = 1 + 3 * x + rnorm(20, sd = 2)
)

# fit the model
fit = lm(formula = y ~ x,
         data = df.like)

# model summary
fit %>%
  glance()
```

x	y	fitted	resid	likelihood
0.90	5.22	3.34	1.88	0.12
0.27	0.20	1.56	-1.36	0.18
0.37	-0.18	1.86	-2.04	0.10
0.57	2.14	2.42	-0.28	0.27
0.91	3.13	3.37	-0.24	0.27
0.20	0.78	1.38	-0.59	0.25
0.90	4.20	3.34	0.86	0.23
0.94	2.05	3.47	-1.42	0.17
0.66	3.85	2.67	1.18	0.20
0.63	0.41	2.58	-2.17	0.09

inferred standard deviation of the error

$$\sum_{i=1}^{n} \ln(\text{likelihood})$$

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.25	0.21	1.45	6.16	0.02	2	-34.74	75.47	78.46	37.77	18

$$e \sim \mathcal{N}(\text{mean} = 0, \text{sd} = 1.45)$$

AIC and BIC

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

$\ln(\hat{L})$ = maximized value of the likelihood function of the model **-34.74**

k = number of parameters in the model **3**

n = number of observations **20**

the `sd` of the normal distribution modeling the residuals counts as a parameter

```
lm(formula = y ~ 1 + x, data = df.example)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.25	0.21	1.45	6.16	0.02	2	-34.74	75.47	78.46	37.77	18

AIC and BIC

$$\text{AIC} = 2k - 2\ln(\hat{L}) = 2 \cdot 3 - 2 \cdot (-34.74) = 75.47$$

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L}) = \ln(20) \cdot 3 - 2 \cdot (-34.74) = 78.46$$

$\ln(\hat{L})$ = maximized value of the likelihood function of the model **-34.74**

k = number of parameters in the model **3**

n = number of observations **20**

the `sd` of the normal distribution modeling the residuals counts as a parameter

```
lm(formula = y ~ 1 + x, data = df.example)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.25	0.21	1.45	6.16	0.02	2	-34.74	75.47	78.46	37.77	18

AIC and BIC

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

ΔBIC	Evidence against higher BIC
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very Strong

- for both AIC and BIC, *lower* is better!
- neither provide a test of a model in the sense of testing a null hypothesis
  - AIC or BIC tell us nothing about the absolute quality of a model, only the quality relative to other models
- BIC generally penalizes free parameters more strongly than AIC (though it depends on the size of *n*)

What shall I use when?

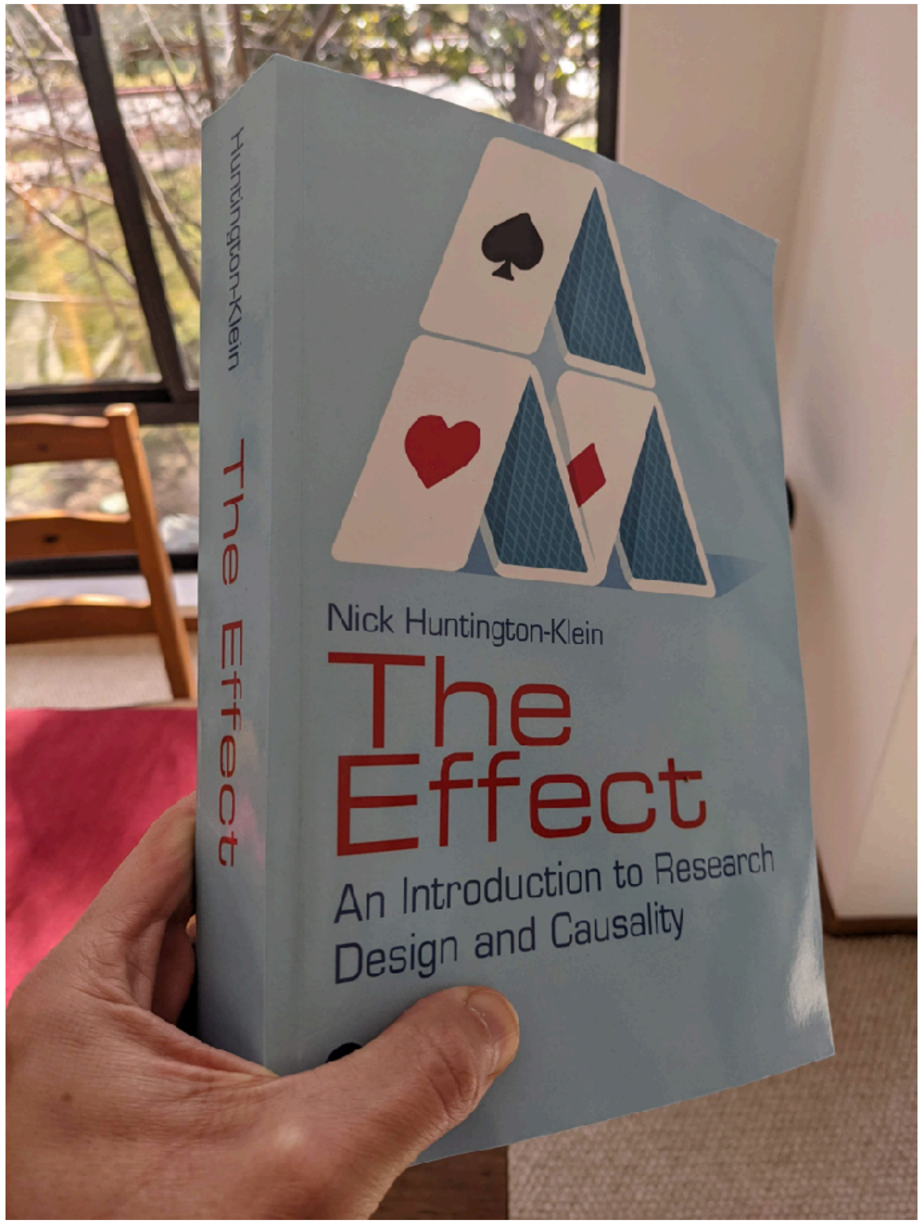
- Use it all!
- ideally, the different measures provide converging evidence

Table 2
Summary of the model results. Values for *r* and RMSE indicate means (with 5% and 95% quantiles in parentheses) based on 100 split-half cross-validation runs. BIC scores are based on running the models on the full data set.

Model	*r*	RMSE	BIC
Difference & pivotality	.86 (.66, .95)	10.56 (6.17, 17.21)	158.59
Difference	.70 (.30, .90)	26.92 (16.4, 40.6)	209.74
Pivotality	.63 (.41, .77)	14.23 (11.39, 17.54)	199.53
Optimality	.66 (.42, .84)	14.55 (10.54, 17.91)	199.47

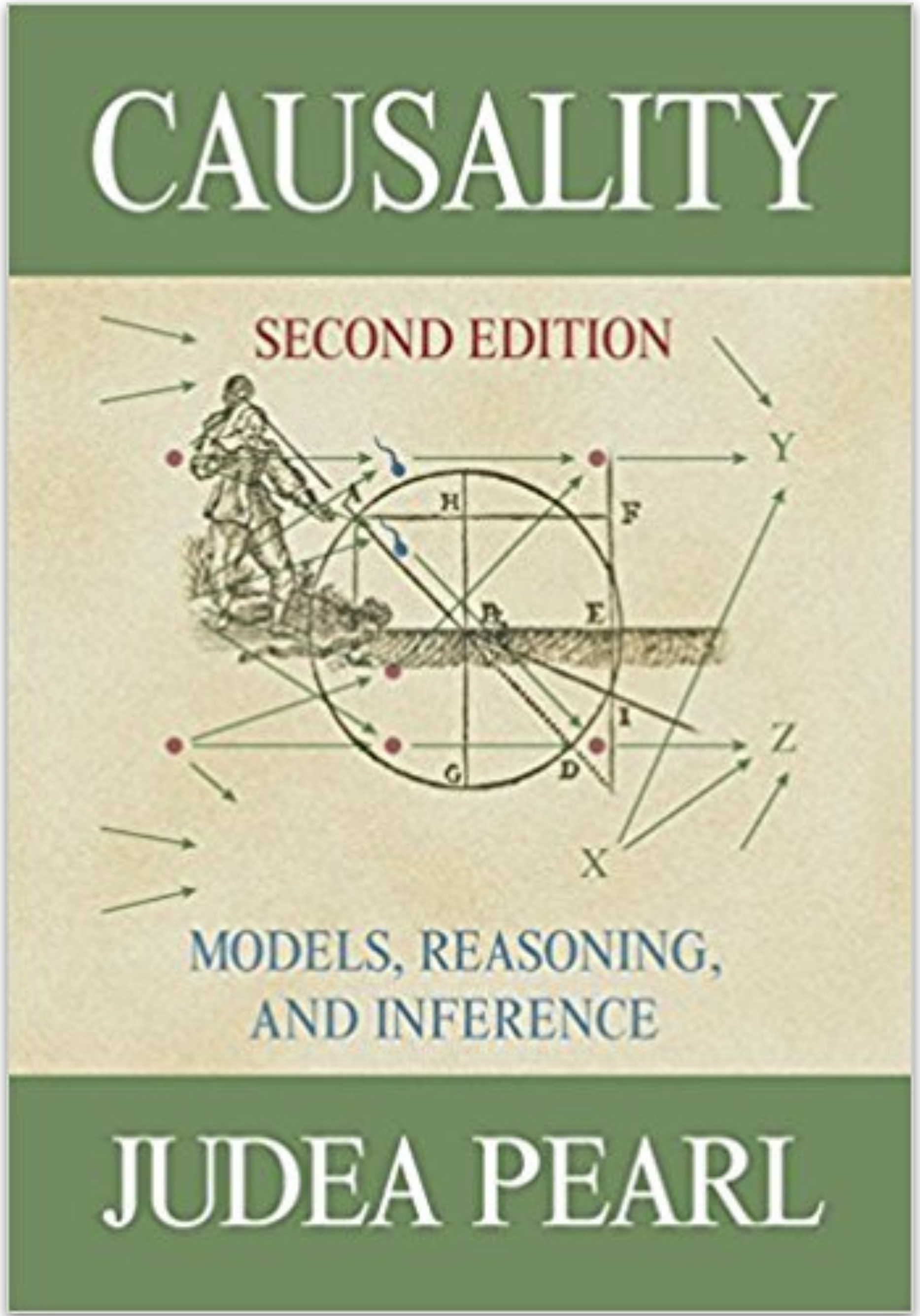
Note: BIC = Bayesian Information Criterion (lower values indicate better model performance).

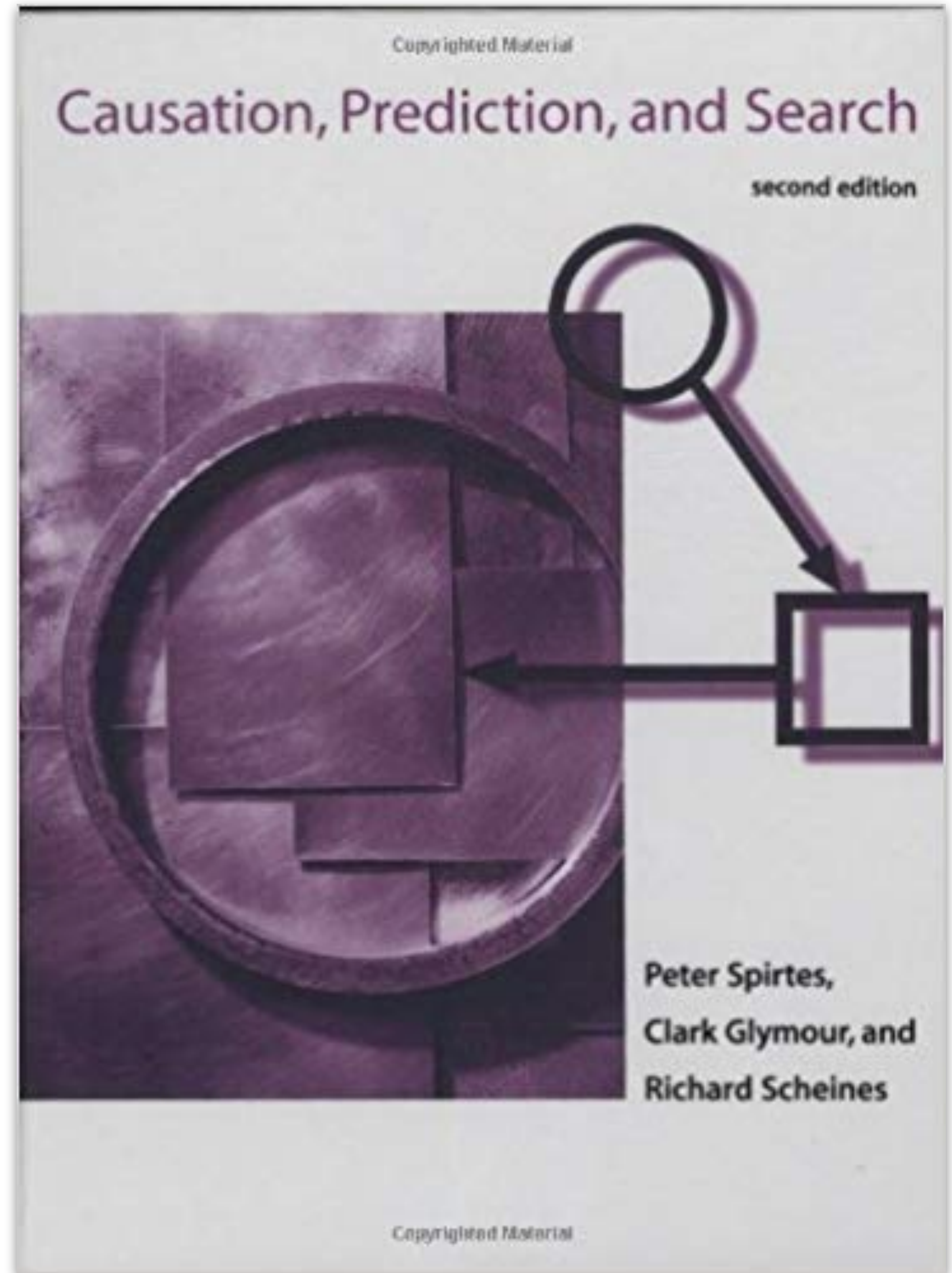
Causation vs. correlation



made with RMarkdown

https://theeffectbook.net/





Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge, England: Cambridge University Press.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. The MIT Press.

Breakout room

Suppose there is a robust, statistically significant, and long-term correlation between the color of cars and the annual rate at which they are involved in accidents.

To be concrete, assume that red cars, in particular, are involved in accidents year after year at a higher rate than cars of any other color. When you go to buy a new car, should you avoid the color red in your quest to remain safe on the road?

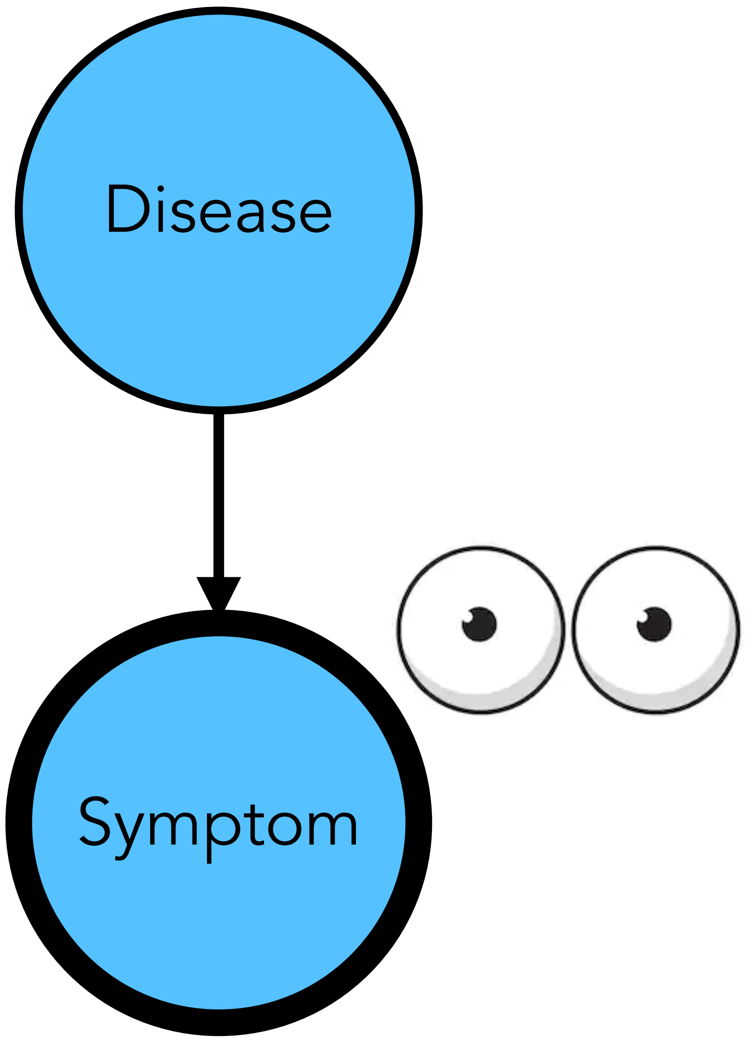
https://bostonreview.net/articles/tim-maudlin-you-are-smarter-your-data/

Hidden

Seeing vs. doing

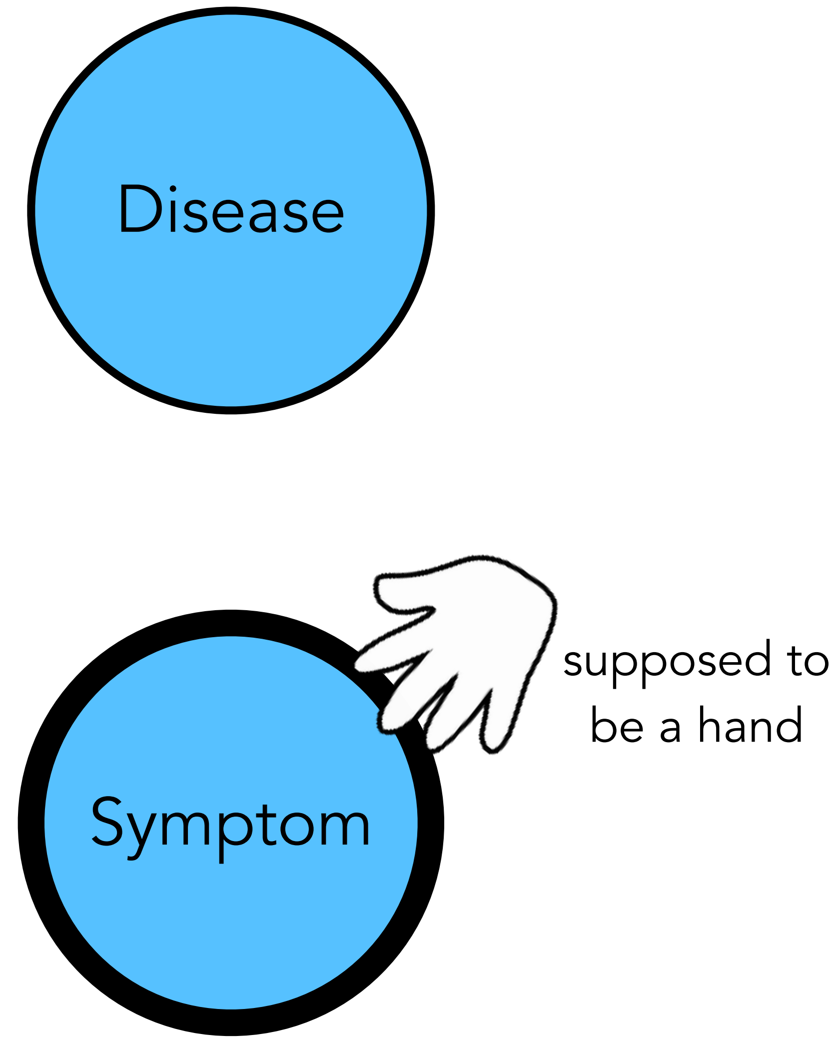
Observation vs. Intervention

seeing



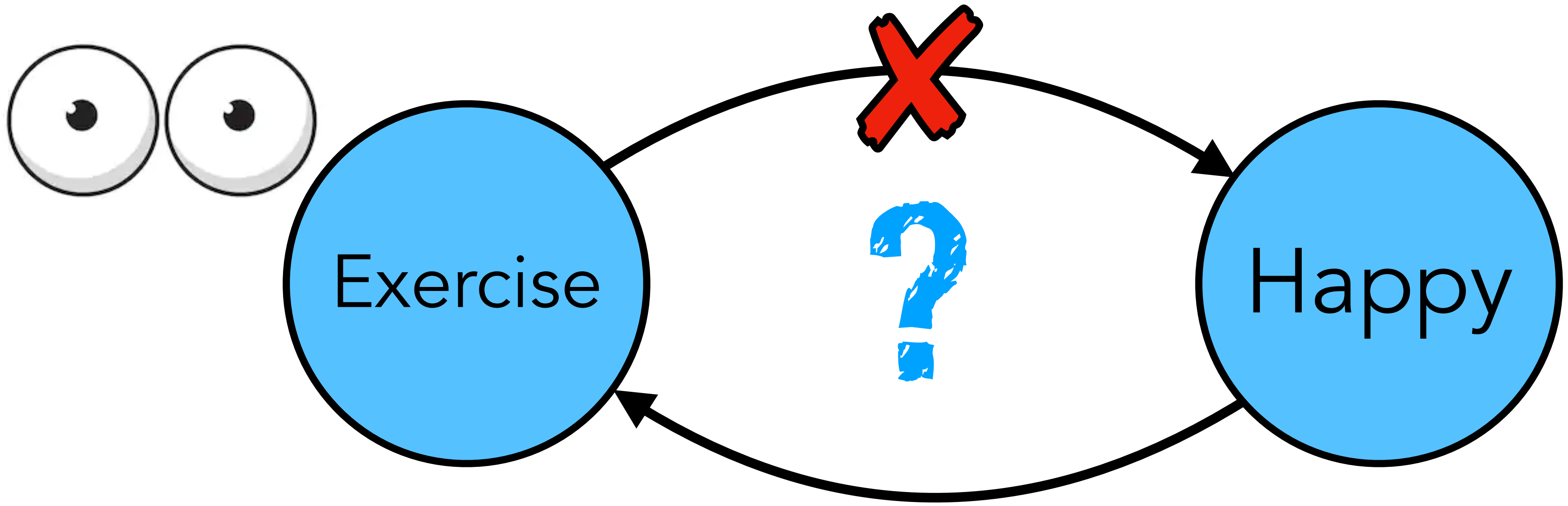
$$p(D \mid S) > p(D)$$

doing

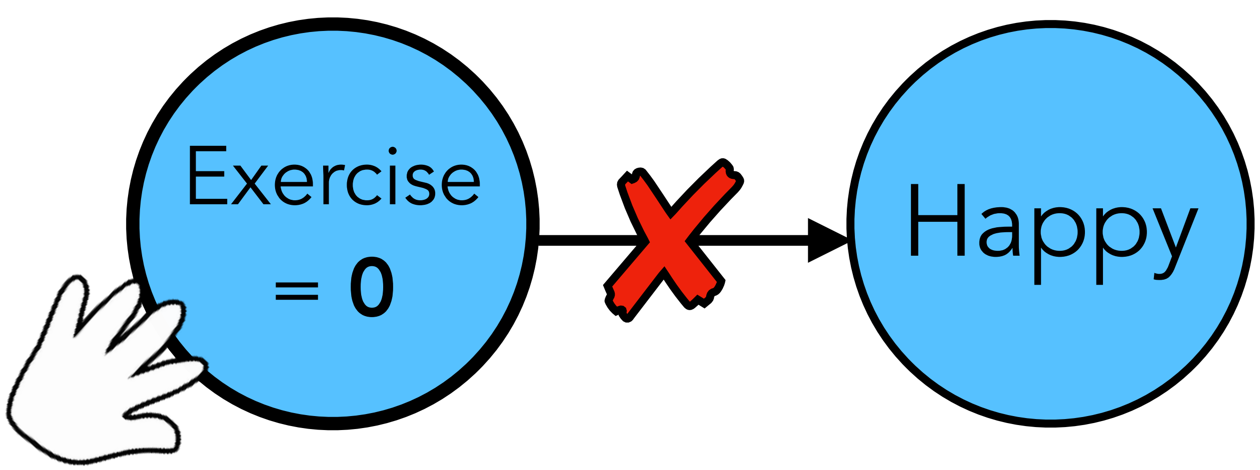


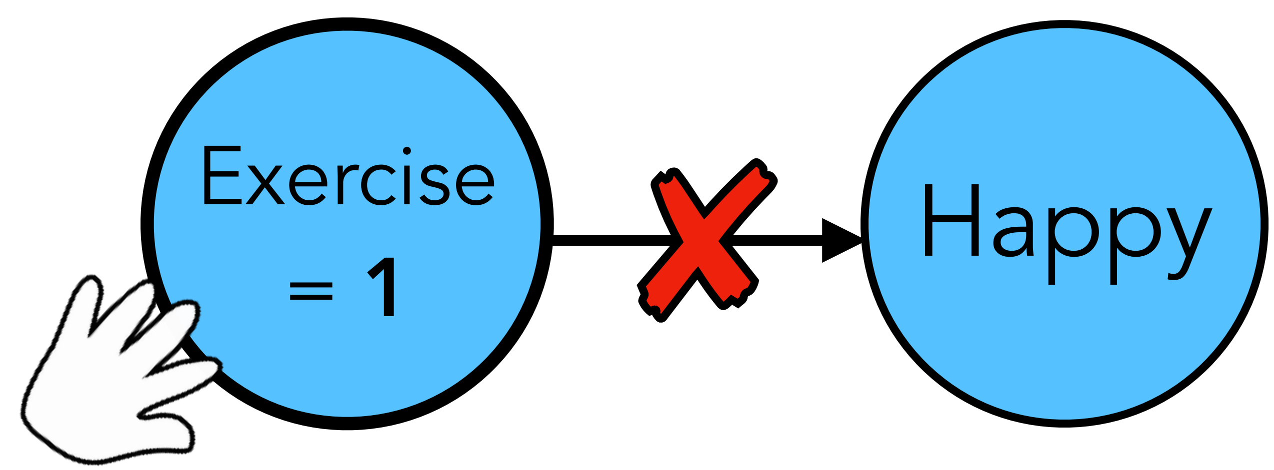
$$p(D \mid \mathrm{do}(S)) = p(D)$$

Inferring causal structure through intervention



Experiment 1

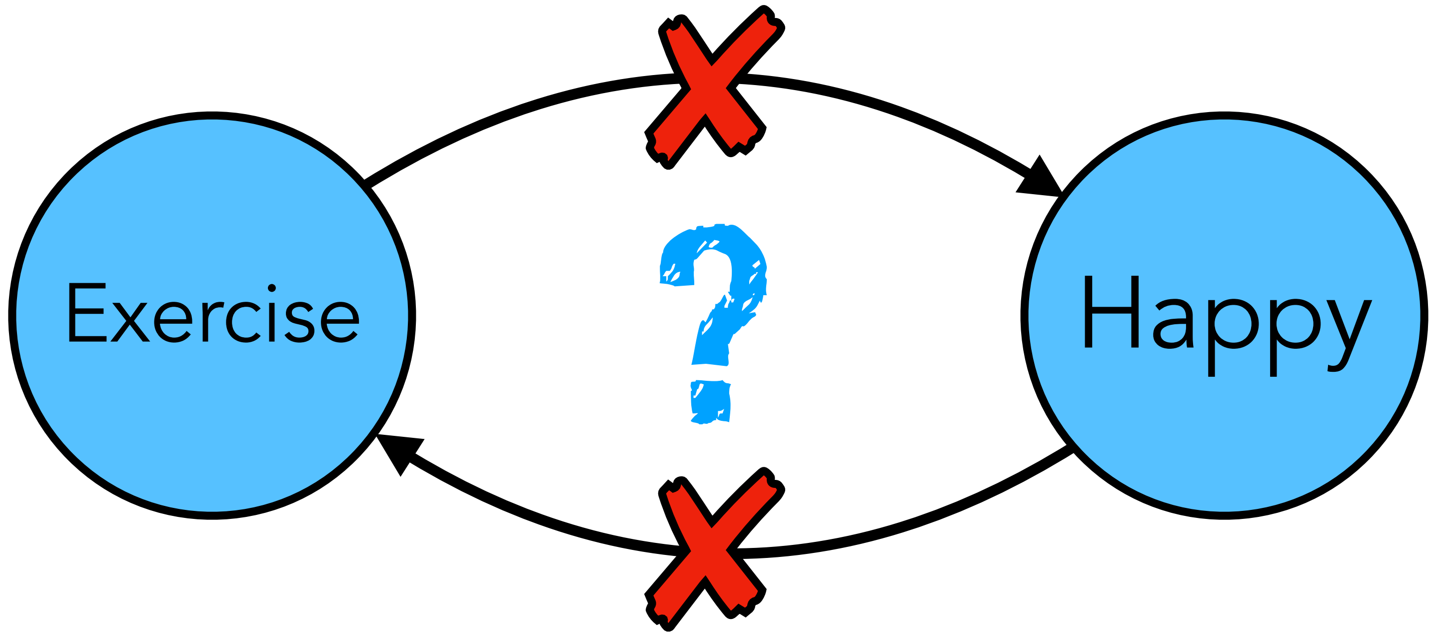




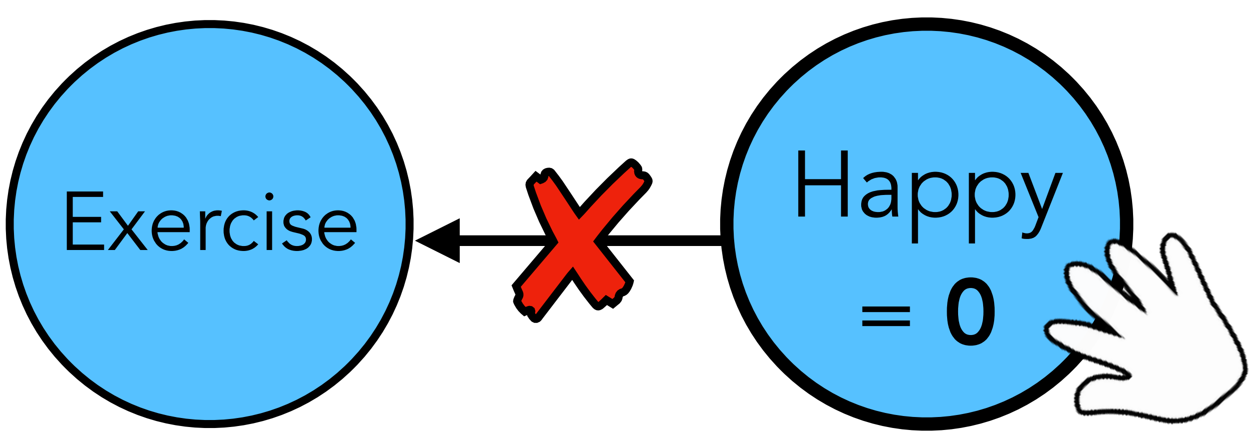
$p(\text{Happy} \mid \text{do}(\text{Exercise} = 0)) = 0.3$

$p(\text{Happy} \mid \text{do}(\text{Exercise} = 1)) = 0.3$

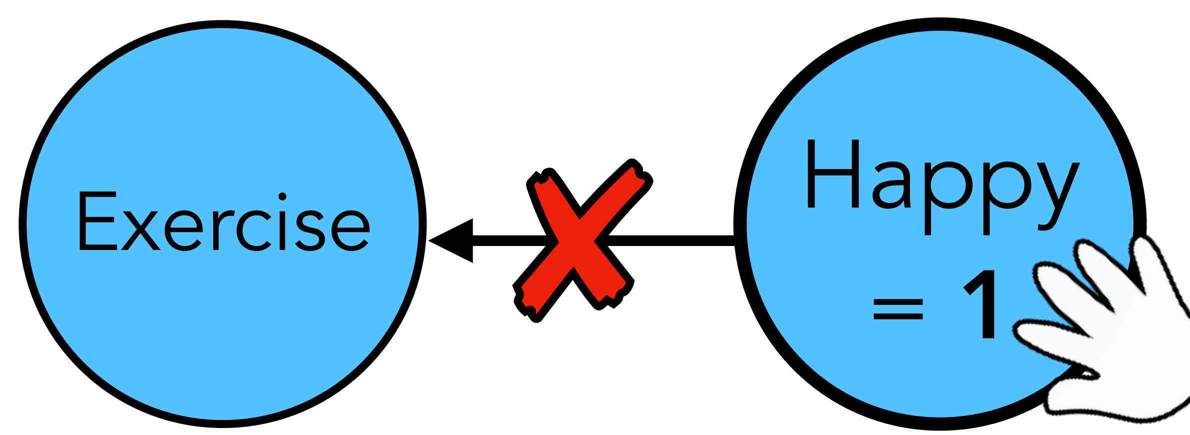
Inferring causal structure through intervention



Experiment 2

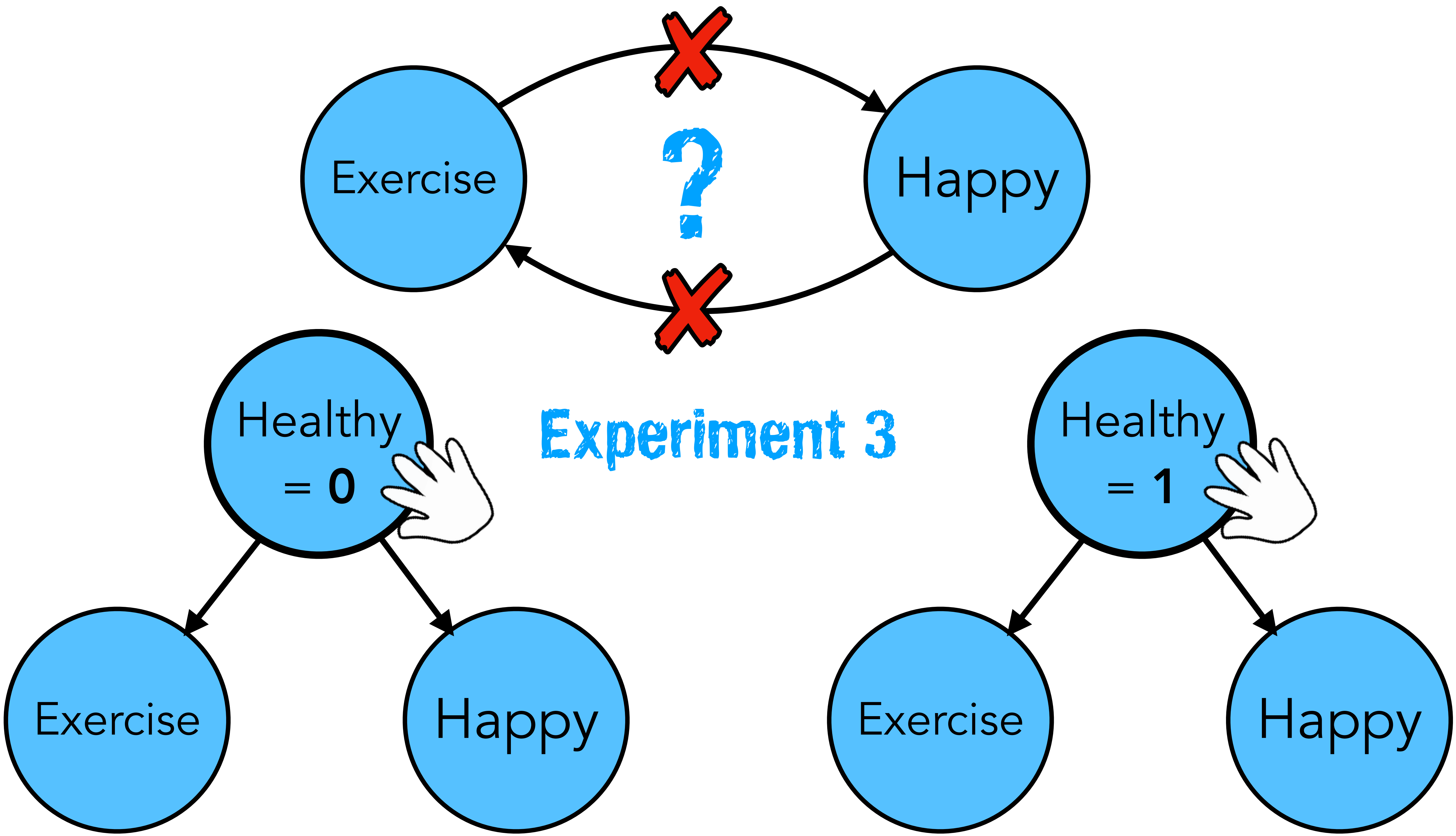


$p(\text{Exercise} \mid \text{do}(\text{Happy} = 0)) = 0.1$



$p(\text{Exercise} \mid \text{do}(\text{Happy} = 1)) = 0.1$

Inferring causal structure through intervention



$p(\text{Happy} \,|\, \text{do}(\text{Healthy} = 0)) = 0.05$

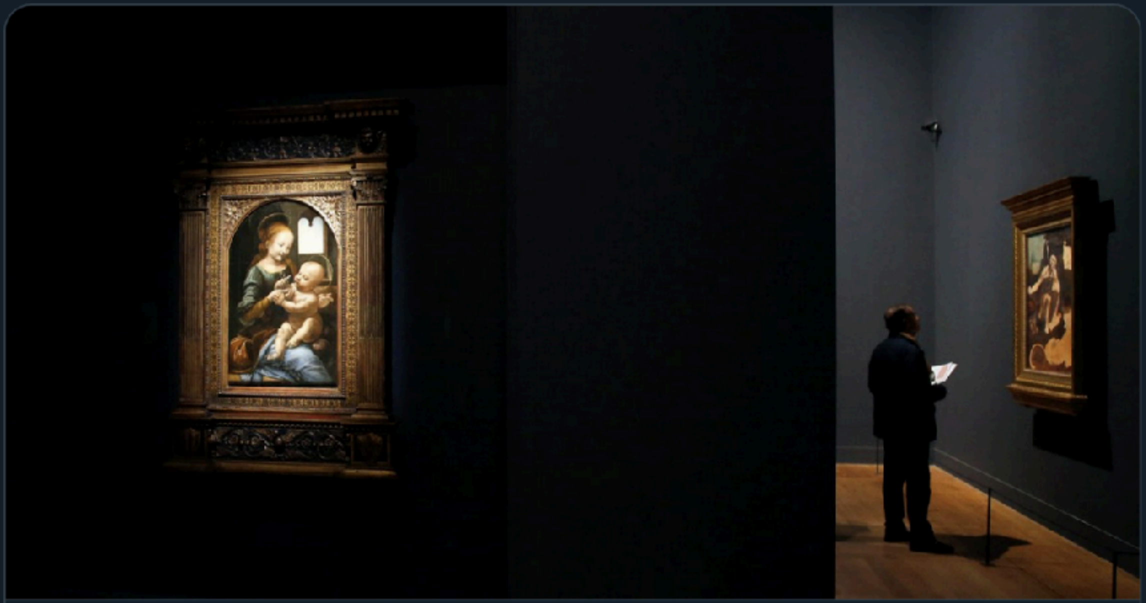
$p(\text{Exercise} \,|\, \text{do}(\text{Healthy} = 0)) = 0.1$

$p(\text{Happy} \,|\, \text{do}(\text{Healthy} = 0)) = 0.5$

$p(\text{Exercise} \,|\, \text{do}(\text{Healthy} = 0)) = 0.75$

NYT Health
@NYTHealth

Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't.



Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

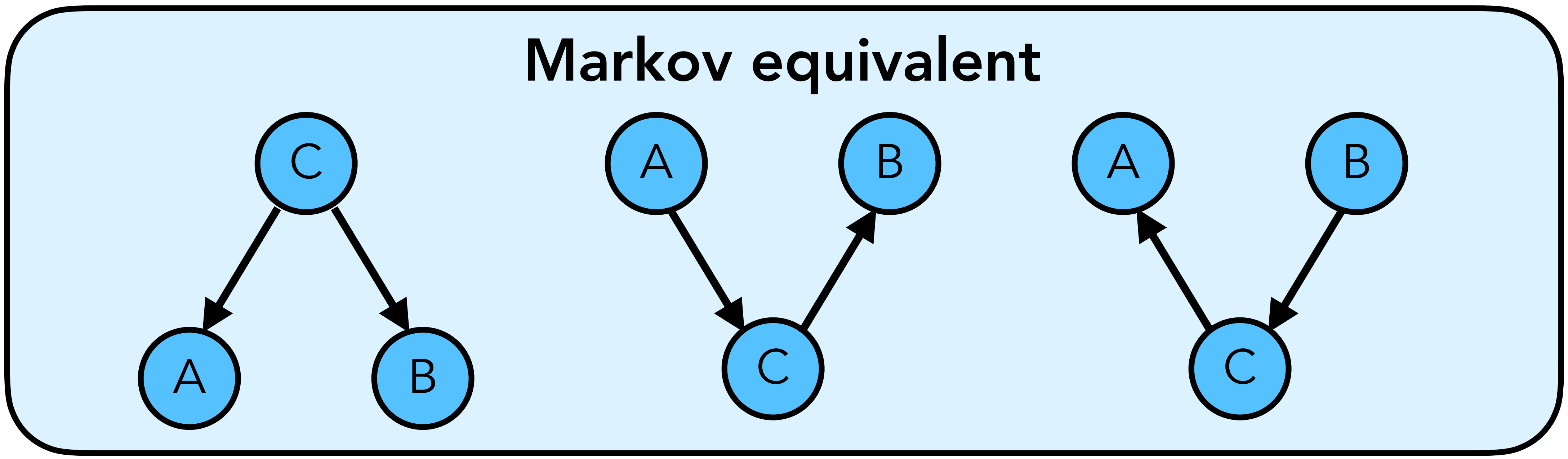
nytimes.com

9:19 AM · Dec 22, 2019 · SocialFlow

336 Retweets 1.3K Likes

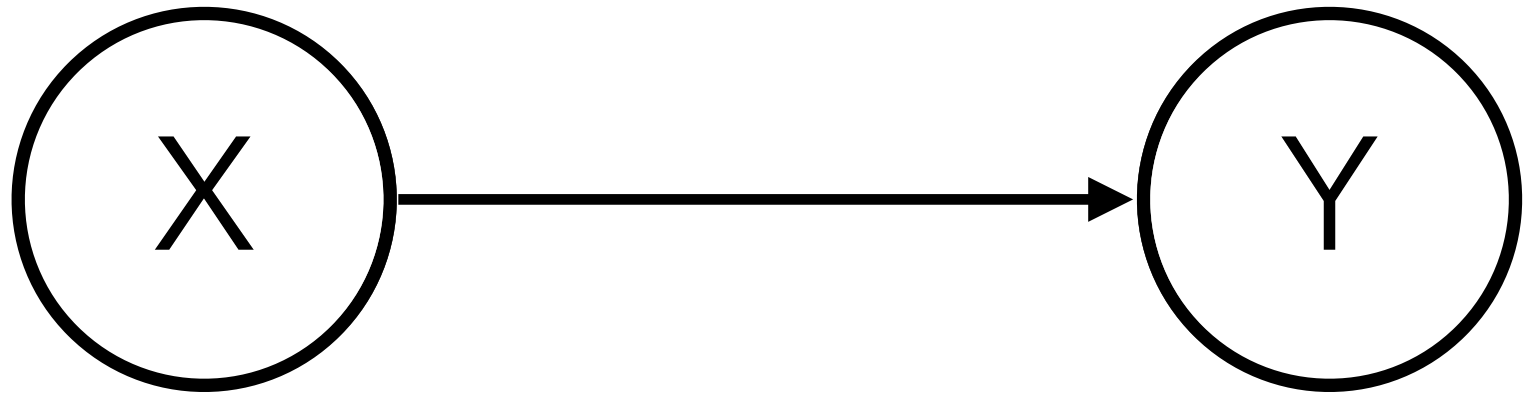
Important take home message

- correlation is not causation
- correlation (= probabilistic dependence) suggests that there is *some* causal relationship
- but we don't know which one it is

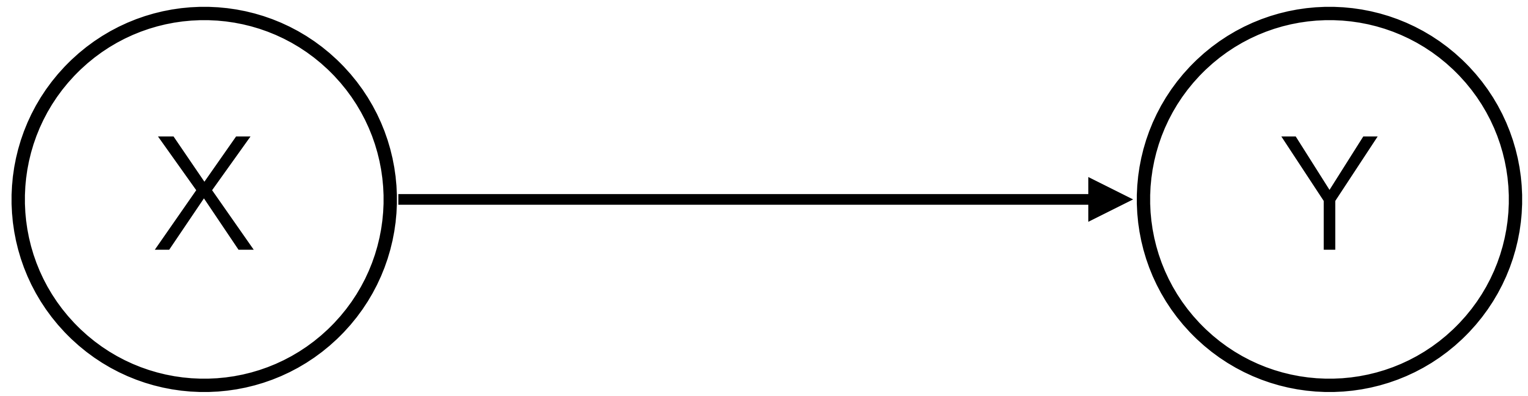


- **causal interventions** / experiments can reveal the underlying causal structure

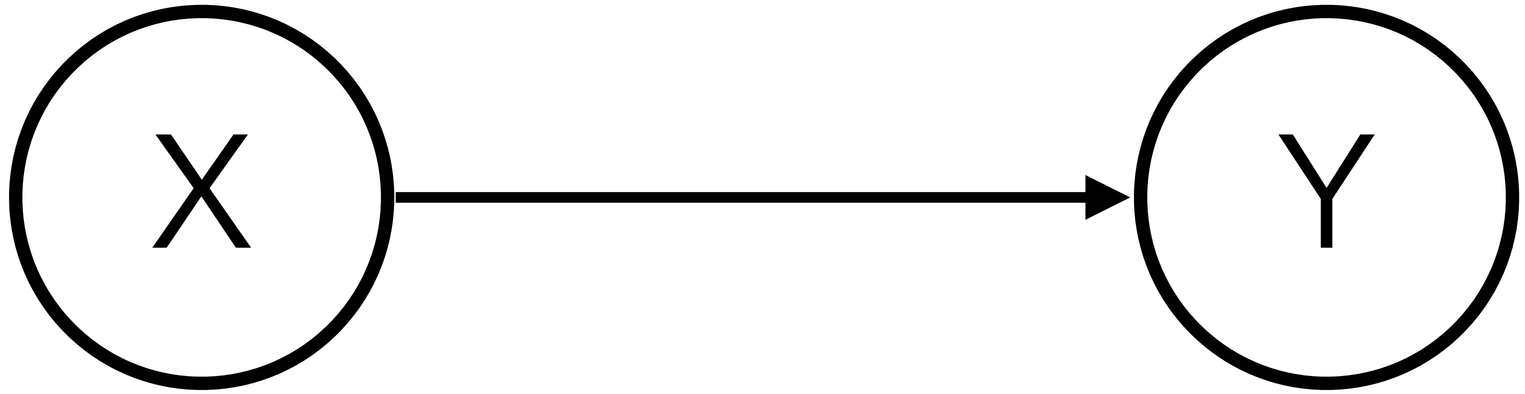
Observation, intervention, counterfactual



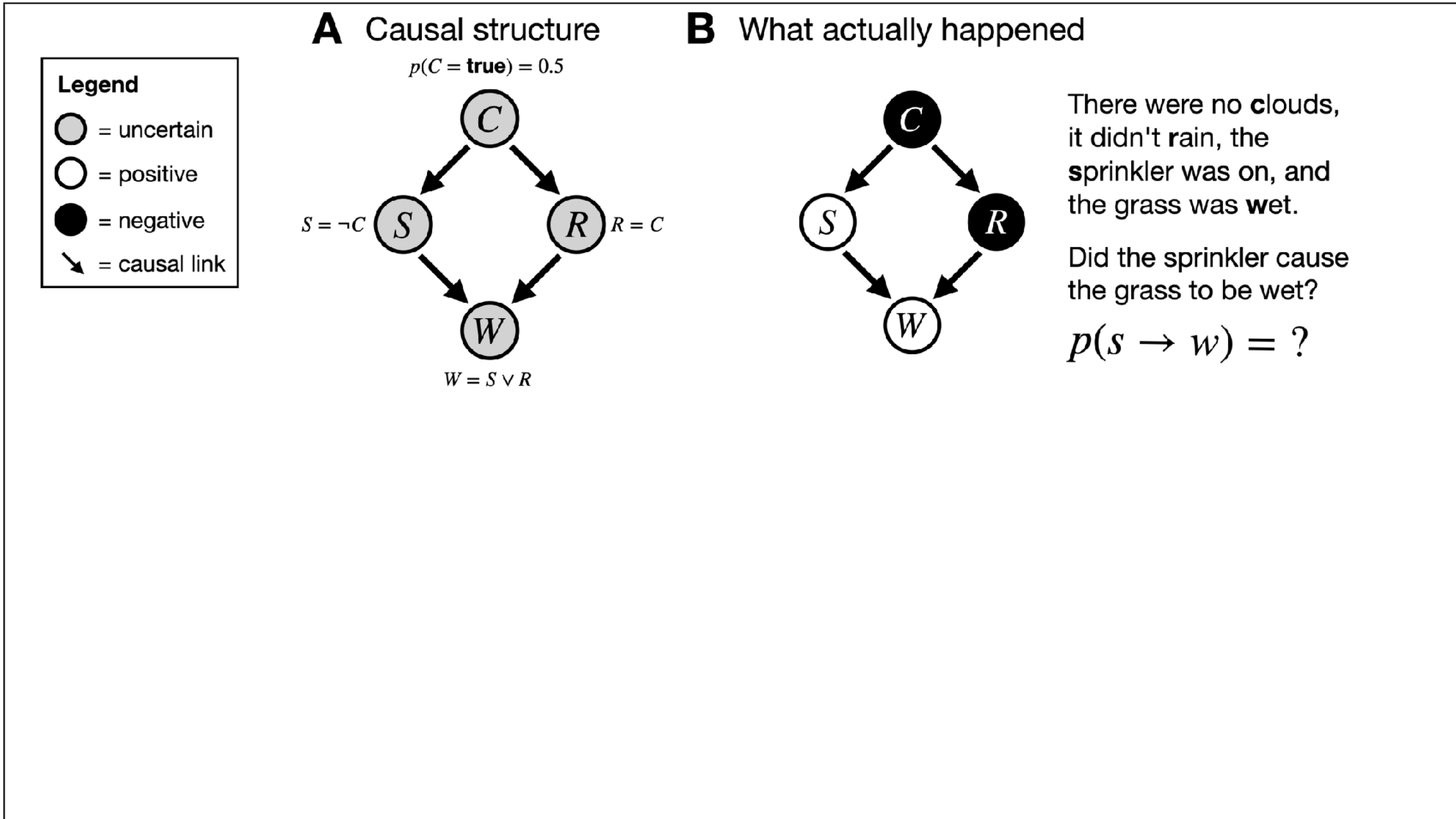
Level	Concept	Expression	Activity	Question	Example
I	Observation/ Prediction	$p(y\|x)$	Seeing	How does x change my belief in y?	Would the grass be wet if we *found* the sprinkler off?



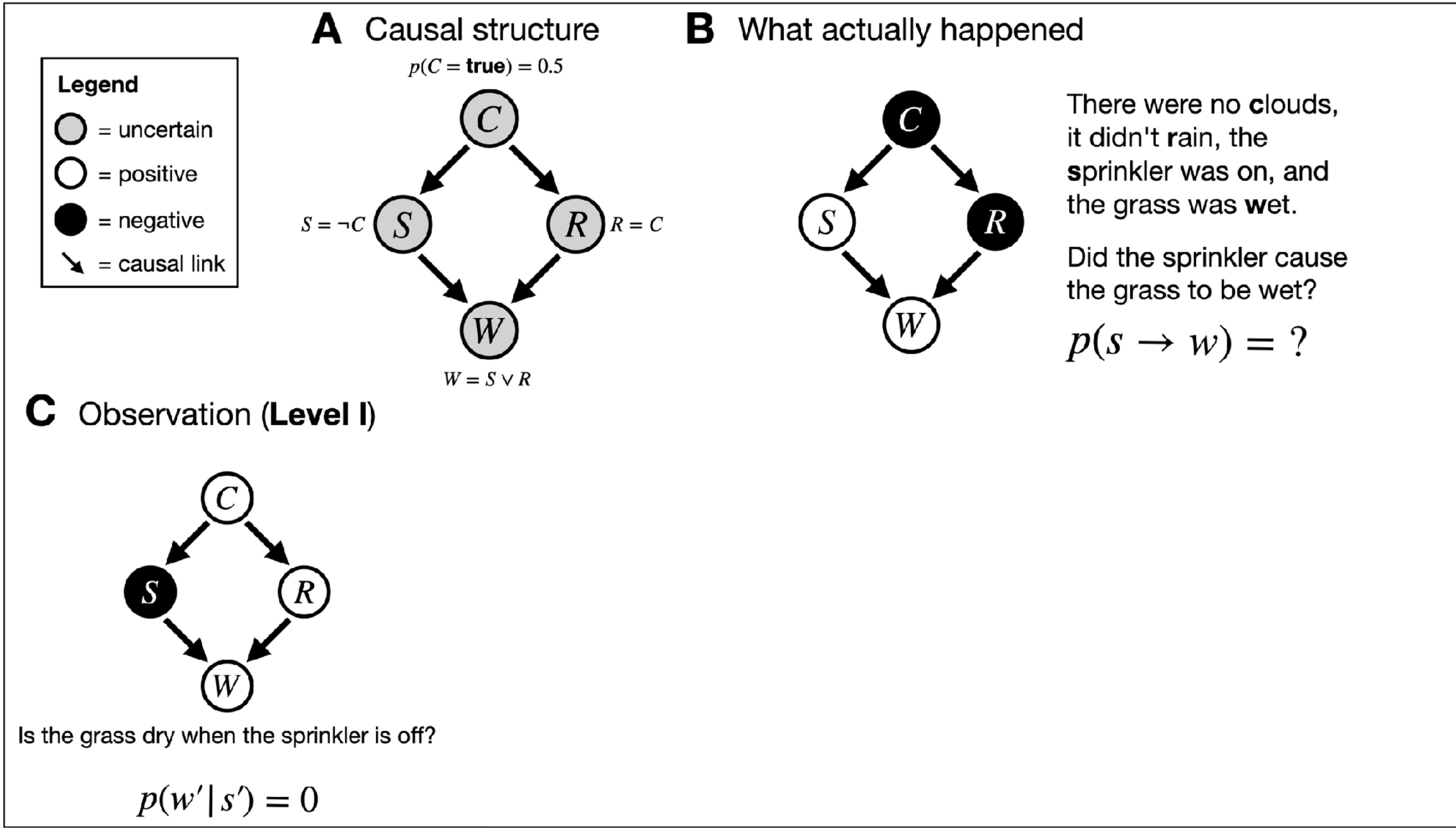
Level	Concept	Expression	Activity	Question	Example
I	Observation/ Prediction	$p(y\|x)$	Seeing	How does x change my belief in y?	Would the grass be wet if we *found* the sprinkler off?
II	Intervention/ Hypothetical	$p(y\|\mathrm{do}(x))$	Doing	Would y happen if I did x?	Would the the grass be wet if *made sure* that the sprinkler was off?

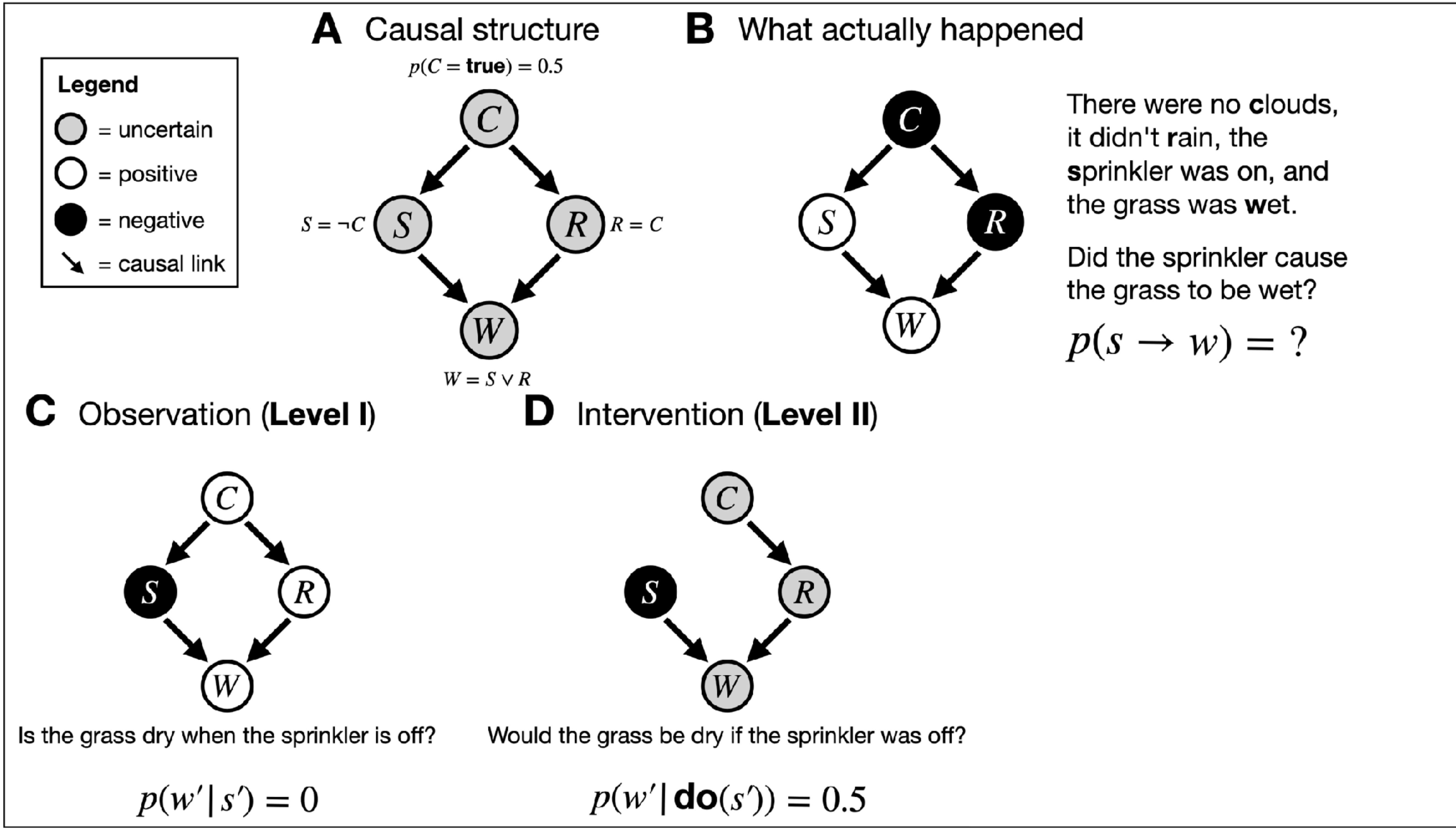


Level	Concept	Expression	Activity	Question	Example
I	Observation/ Prediction	$p(y\|x)$	Seeing	How does x change my belief in y?	Would the grass be wet if we *found* the sprinkler off?
II	Intervention/ Hypothetical	$p(y\|\text{do}(x))$	Doing	Would y happen if I did x?	Would the the grass be wet if *made sure* that the sprinkler was off?
III	Counterfactual	$p(y_x\|x', y')$	Explaining	Would y have happened instead of y', if I had done x instead of x'?	Would the grass have been wet if we *had made sure* that the sprinkler was off, given that the grass is wet and the sprinkler on?

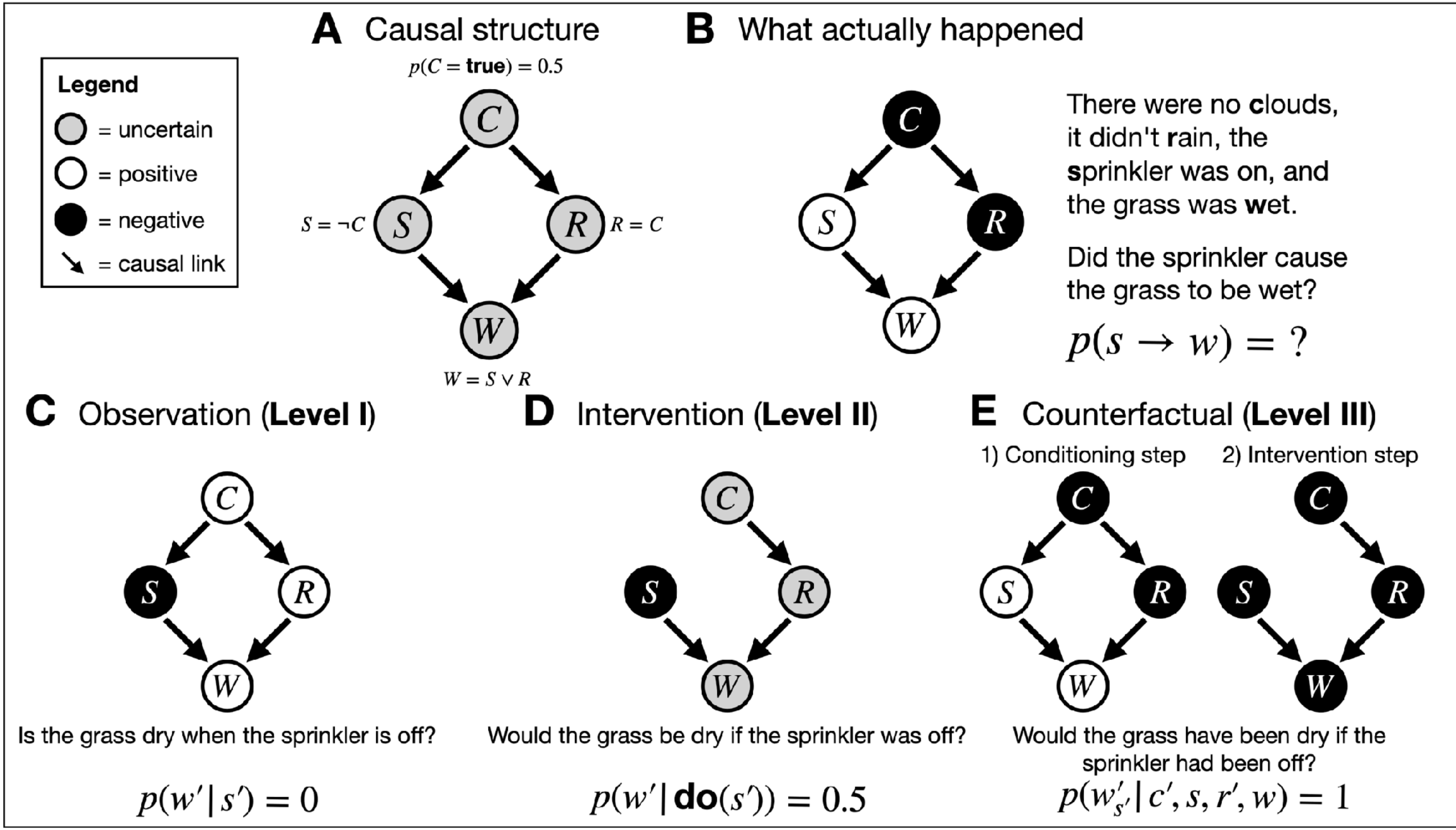


Legend
= uncertain
= positive
= negative
= causal link
A Causal structure
$p(C = \textbf{true}) = 0.5$
C
S
R
W
$S = \neg C$
$R = C$
$W = S \vee R$
B What actually happened
C
S
R
W
There were no **c**louds, it didn't **r**ain, the **s**prinkler was on, and the grass was **w**et.
Did the sprinkler cause the grass to be wet?
$p(s \rightarrow w) = ?$


Legend
= uncertain
= positive
= negative
= causal link
A Causal structure
$p(C = \textbf{true}) = 0.5$
$S = \neg C$
$R = C$
$W = S \vee R$
B What actually happened
There were no clouds, it didn't rain, the sprinkler was on, and the grass was wet.
Did the sprinkler cause the grass to be wet?
$p(s \rightarrow w) = ?$
C Observation (Level I)
Is the grass dry when the sprinkler is off?
$p(w' \mid s') = 0$



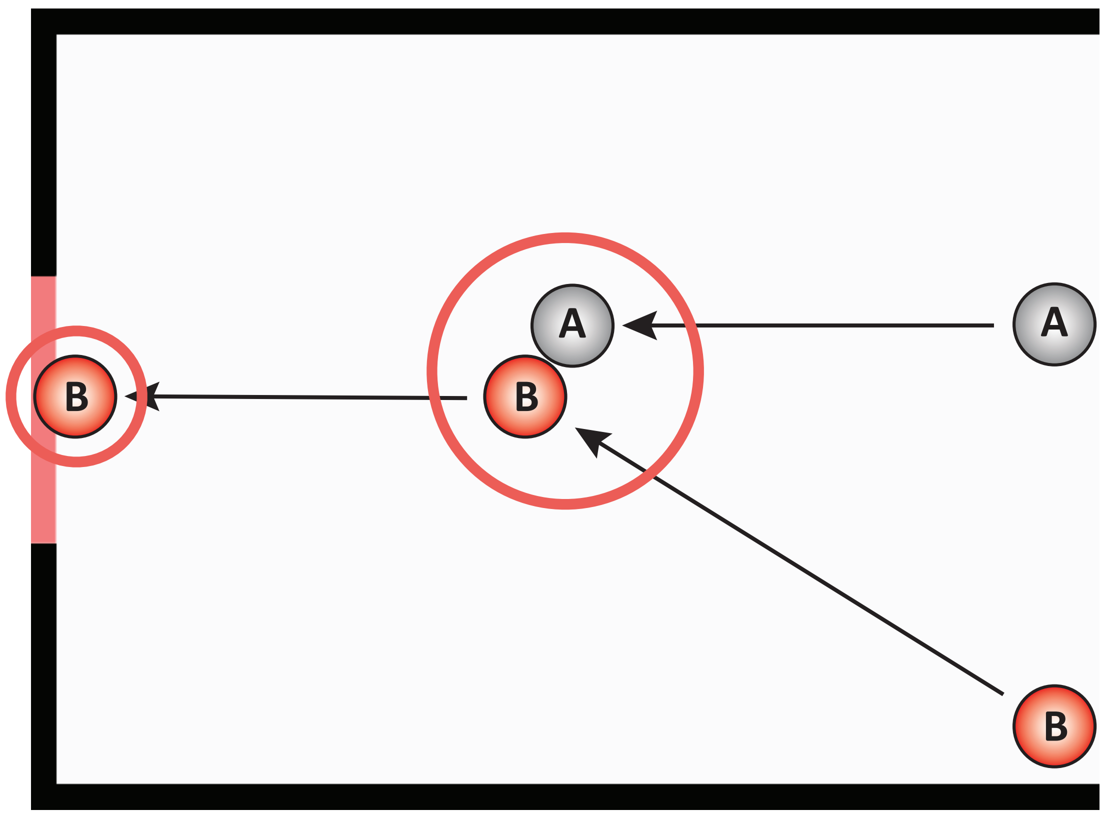
A Causal structure
B What actually happened
Legend
= uncertain
= positive
= negative
= causal link
p(C = true) = 0.5
C
S = ¬C
S
R
R = C
W
W = S ∨ R
There were no clouds, it didn't rain, the sprinkler was on, and the grass was wet.
Did the sprinkler cause the grass to be wet?
p(s → w) = ?
C Observation (Level I)
D Intervention (Level II)
Is the grass dry when the sprinkler is off?
p(w′ | s′) = 0
Would the grass be dry if the sprinkler was off?
p(w′ | do(s′)) = 0.5


A Causal structure
B What actually happened
Legend
= uncertain
= positive
= negative
= causal link
$p(C = \textbf{true}) = 0.5$
$S = \neg C$
$R = C$
$W = S \vee R$
There were no **c**louds, it didn't **r**ain, the **s**prinkler was on, and the grass was **w**et.
Did the sprinkler cause the grass to be wet?
$p(s \rightarrow w) = ?$
C Observation (**Level I**)
D Intervention (**Level II**)
E Counterfactual (**Level III**)
1) Conditioning step
2) Intervention step
Is the grass dry when the sprinkler is off?
$p(w' \mid s') = 0$
Would the grass be dry if the sprinkler was off?
$p(w' \mid \textbf{do}(s')) = 0.5$
Would the grass have been dry if the sprinkler had been off?
$p(w'_{s'} \mid c', s, r', w) = 1$

Did A cause B to go through the gate?

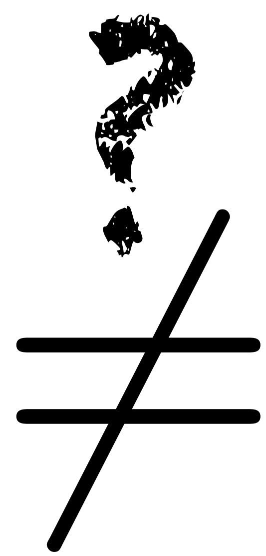
Counterfactual Simulation Model

What happened?

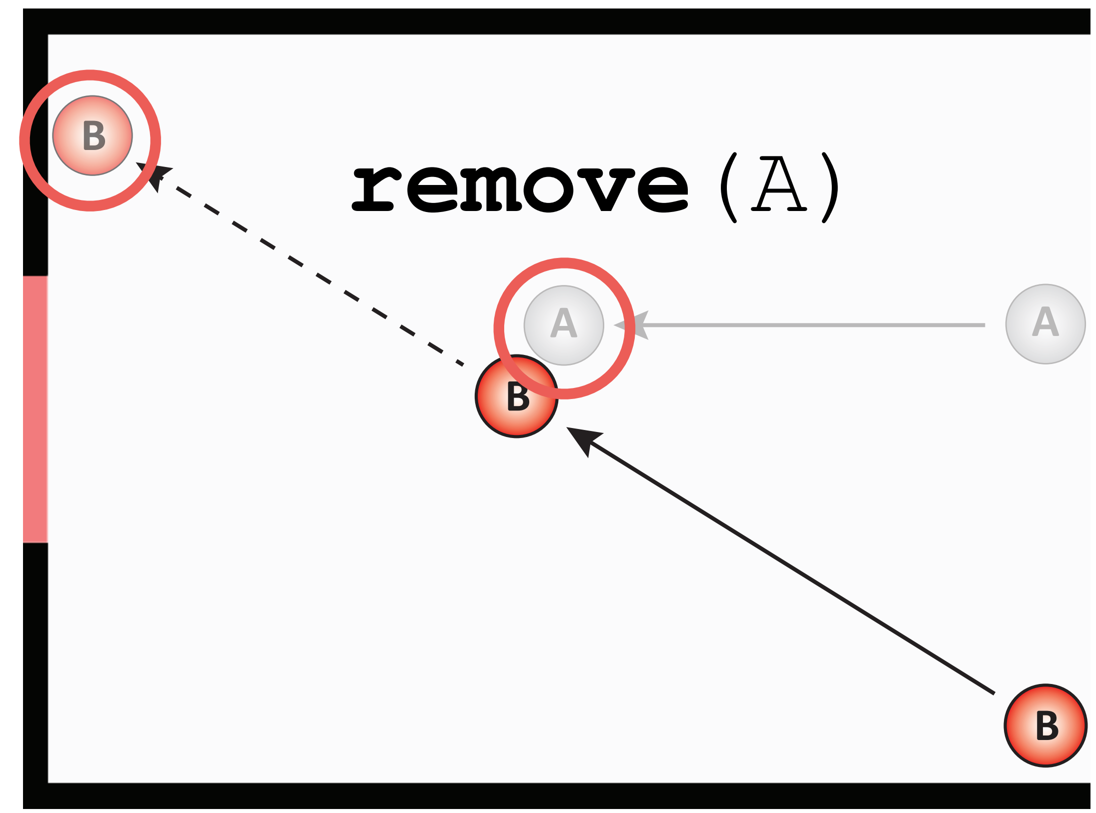


Actual situation

B went through the gate



What would have happened?



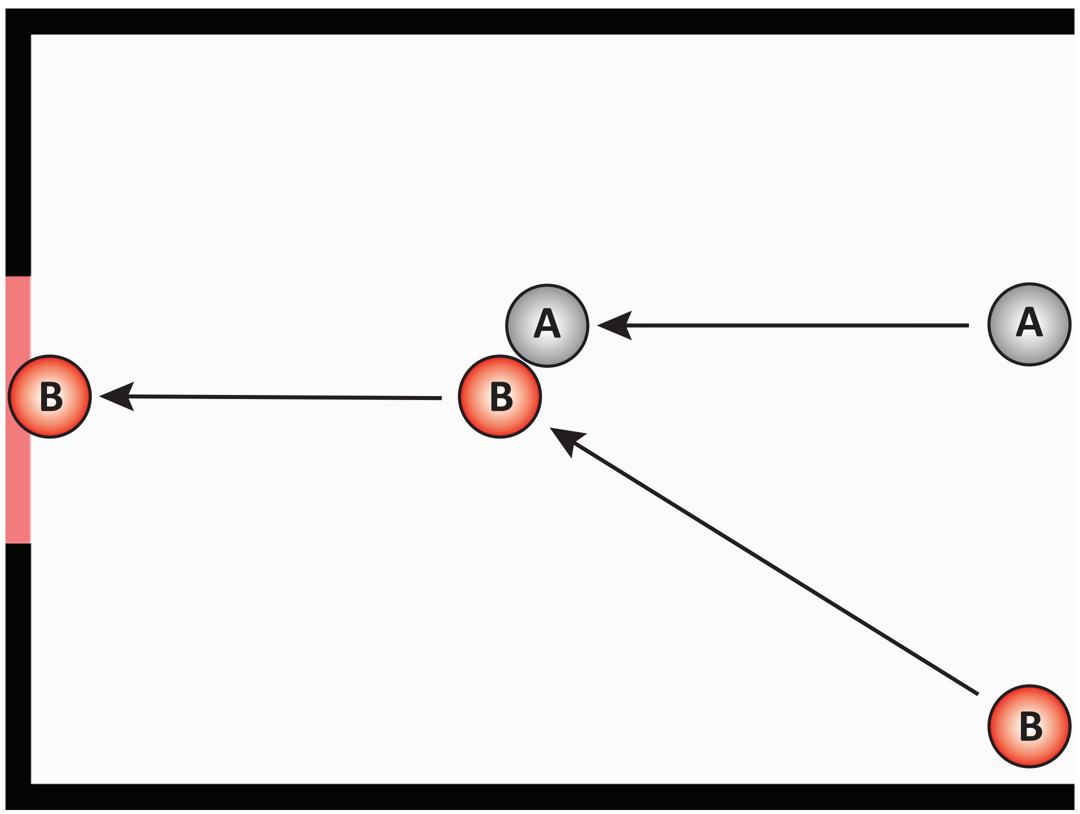
Counterfactual situation

B would have missed the gate

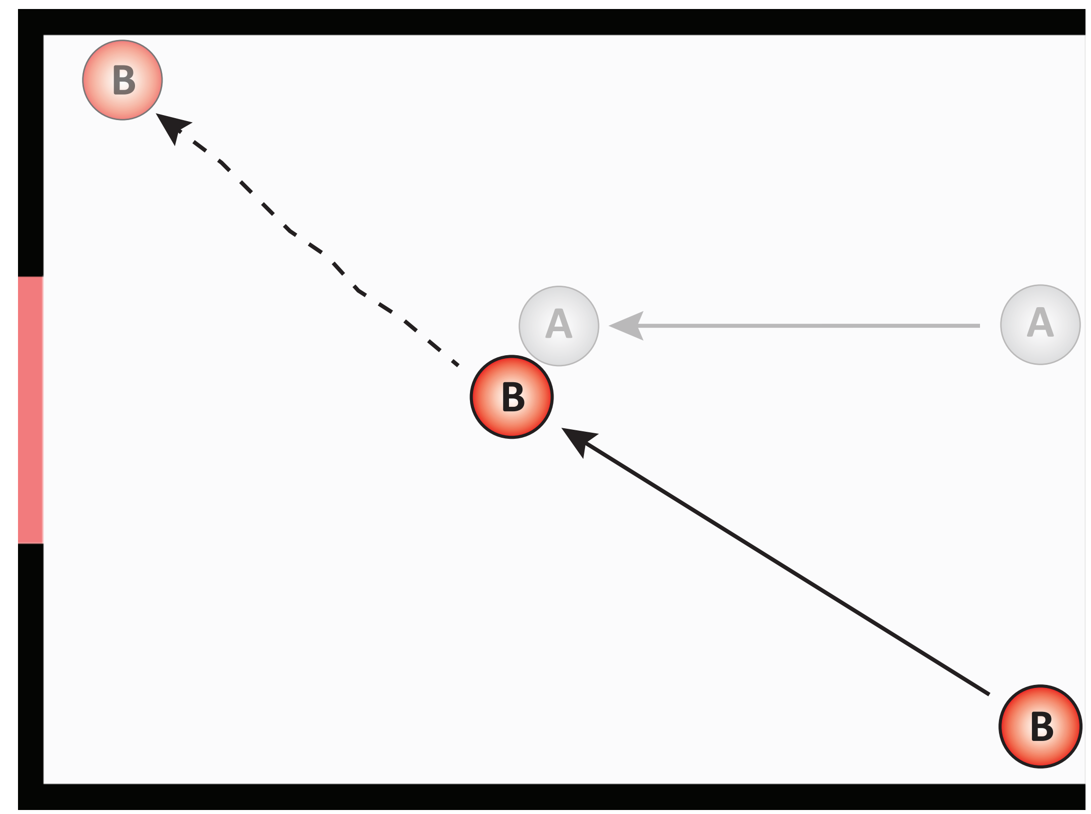
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Quantitative predictions

What happened?



What would have happened?



Actual situation

B went through the gate

?

≠

Counterfactual situation

B would have missed the gate ✓

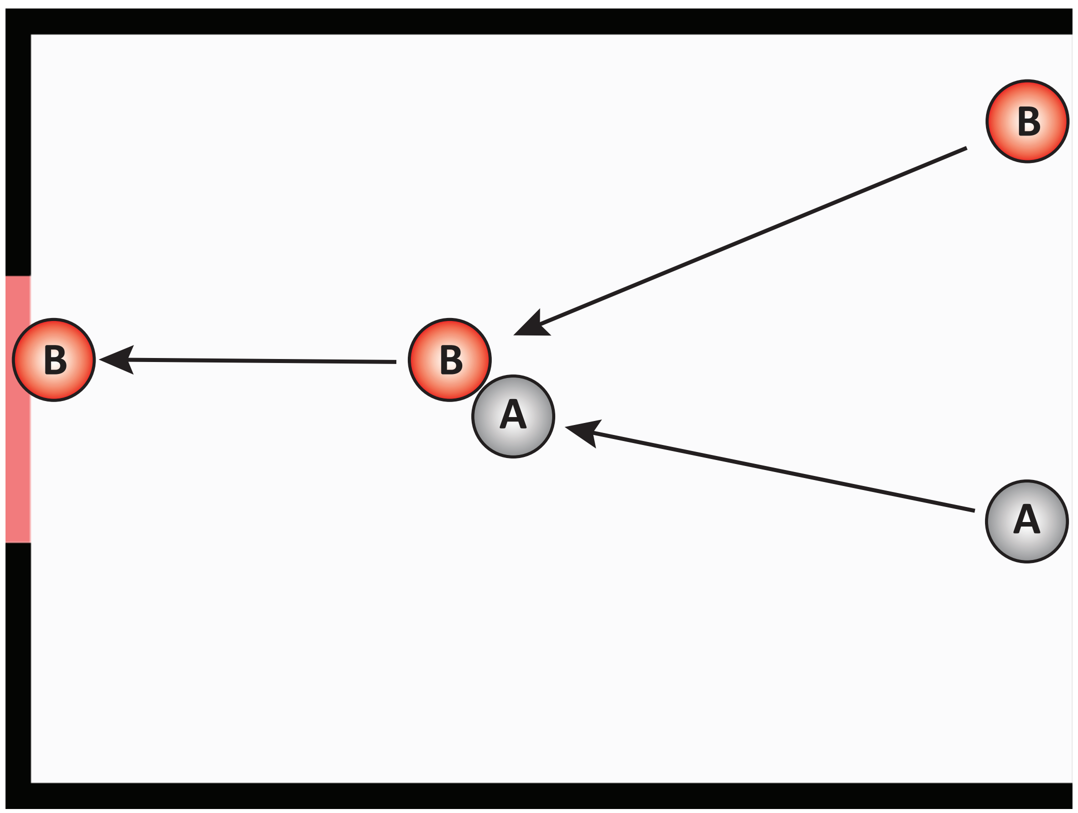
B would have missed the gate ✓

B would have missed the gate ✓

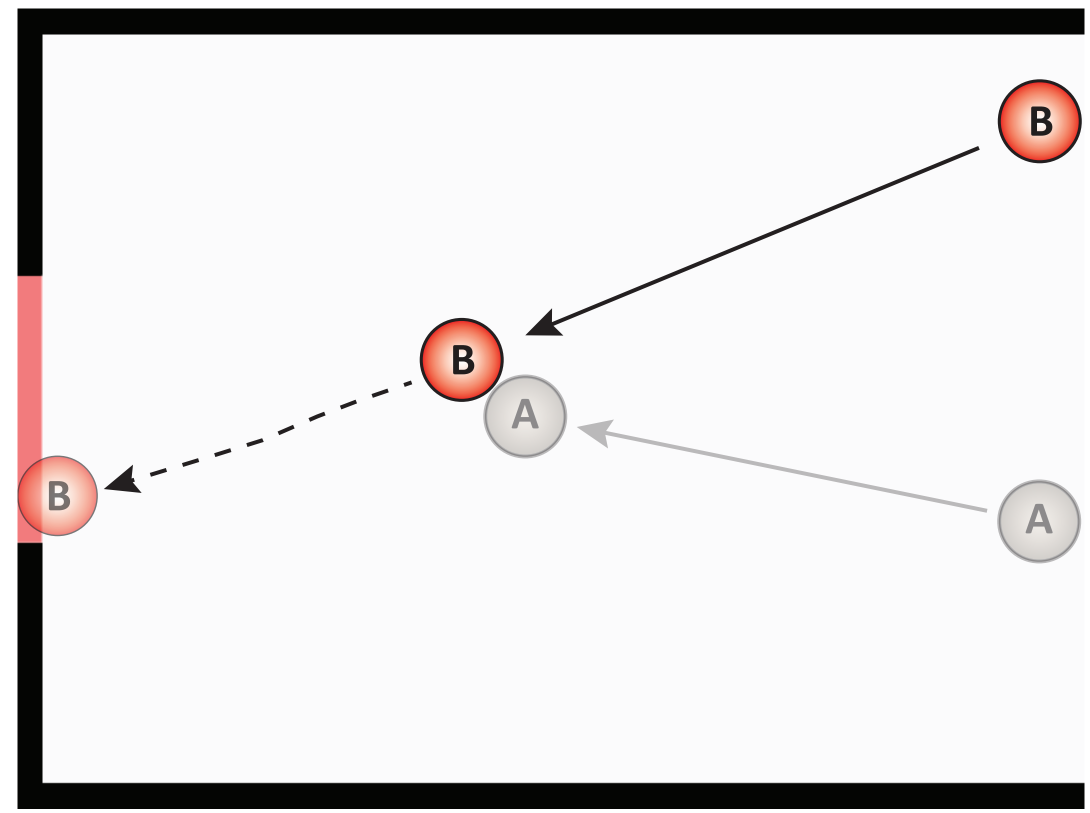
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

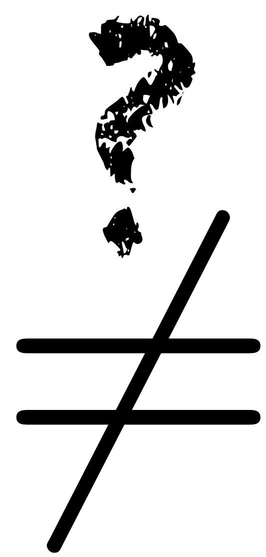
Quantitative predictions

What happened?



What would have happened?





Actual situation

B went through the gate

Counterfactual situation

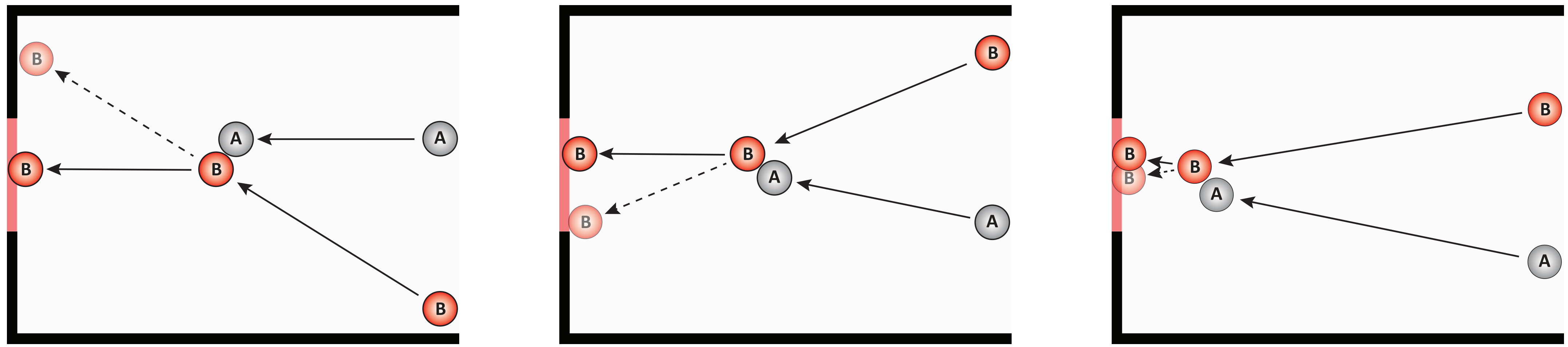
B would have **missed** the gate ✓

B would have **gone through** gate ✗

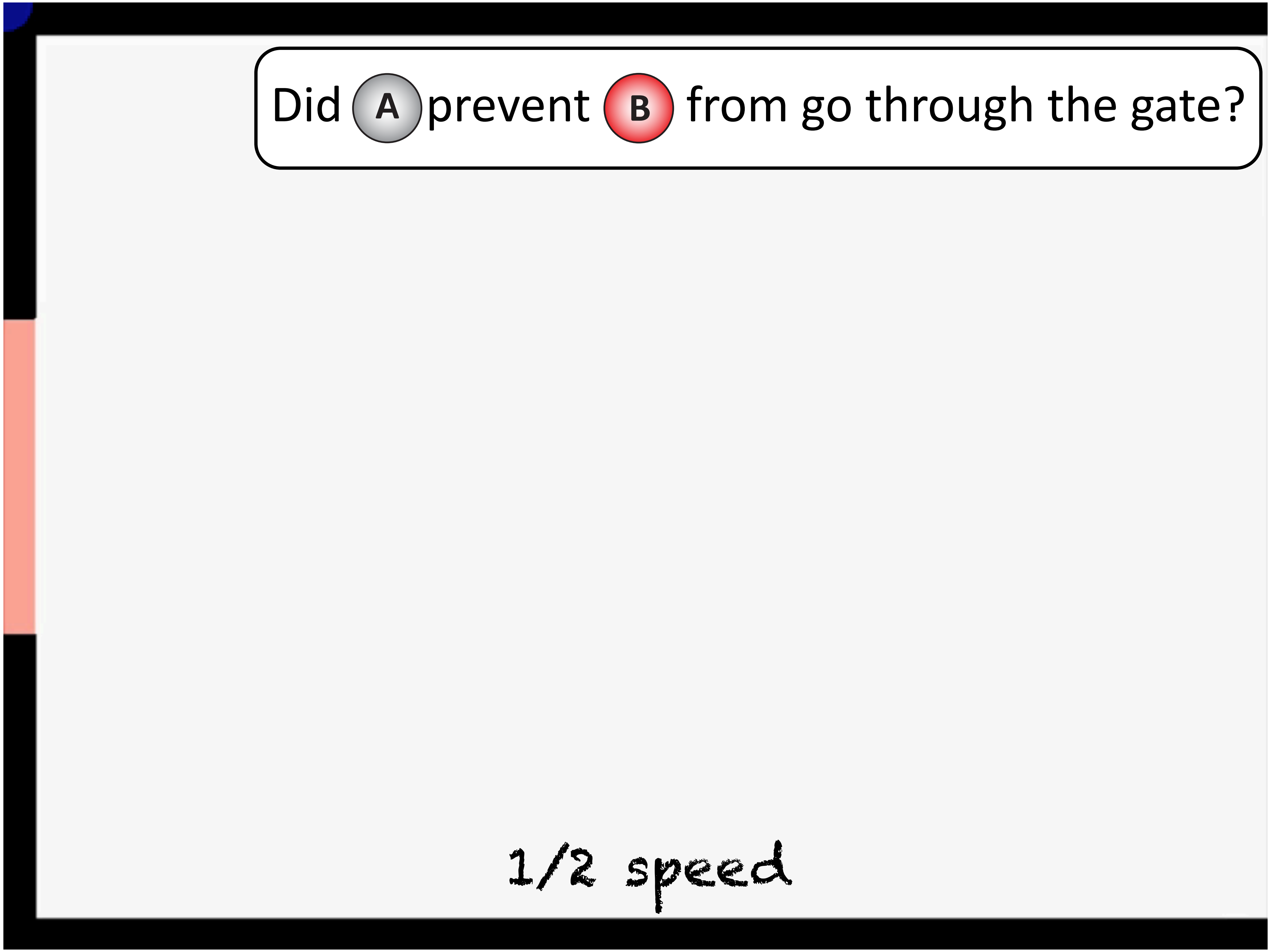
B would have **gone through** gate ✗

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*

Quantitative predictions



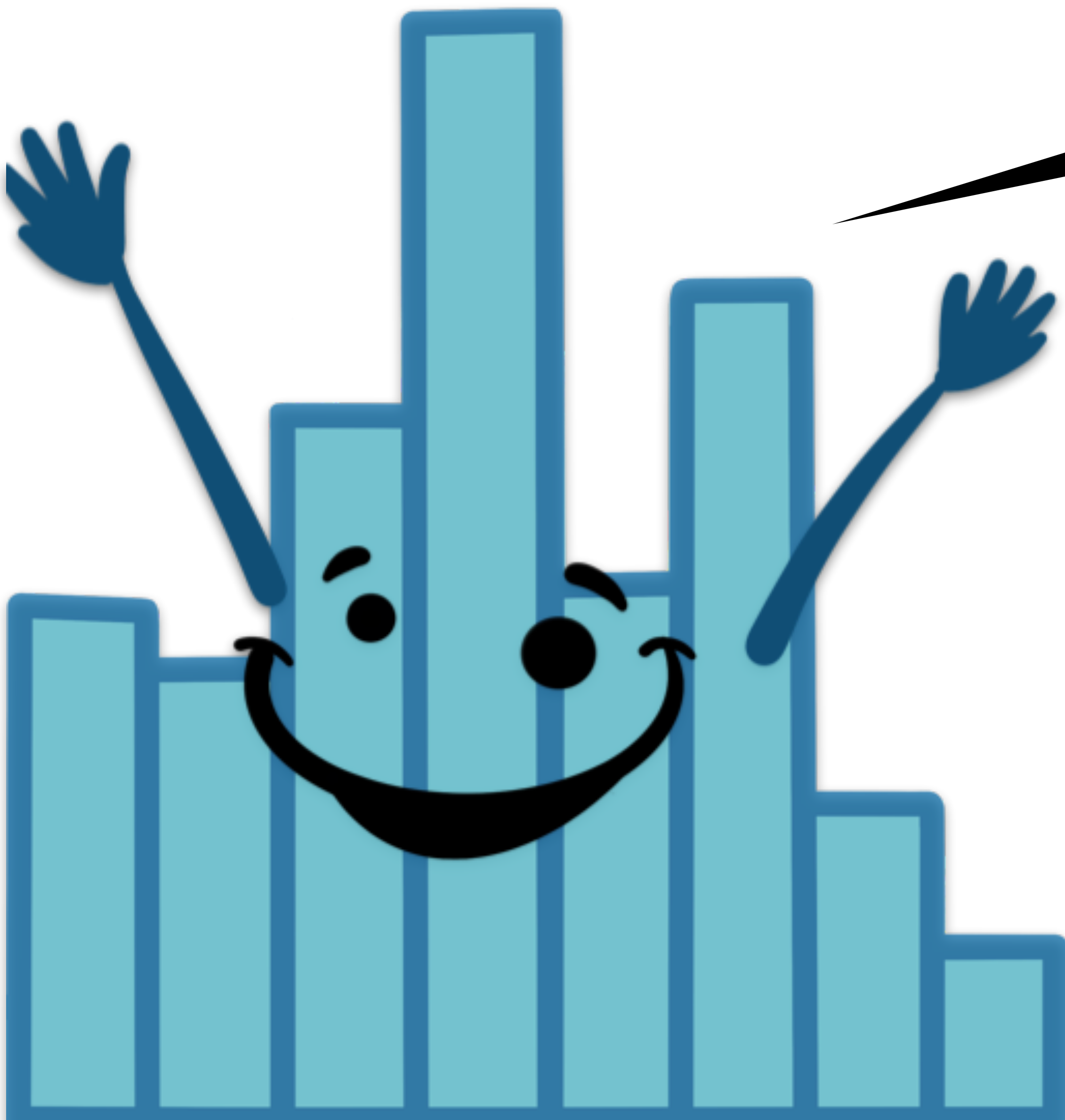
Gerstenberg, Goodman, Lagnado, & Tenenbaum (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*



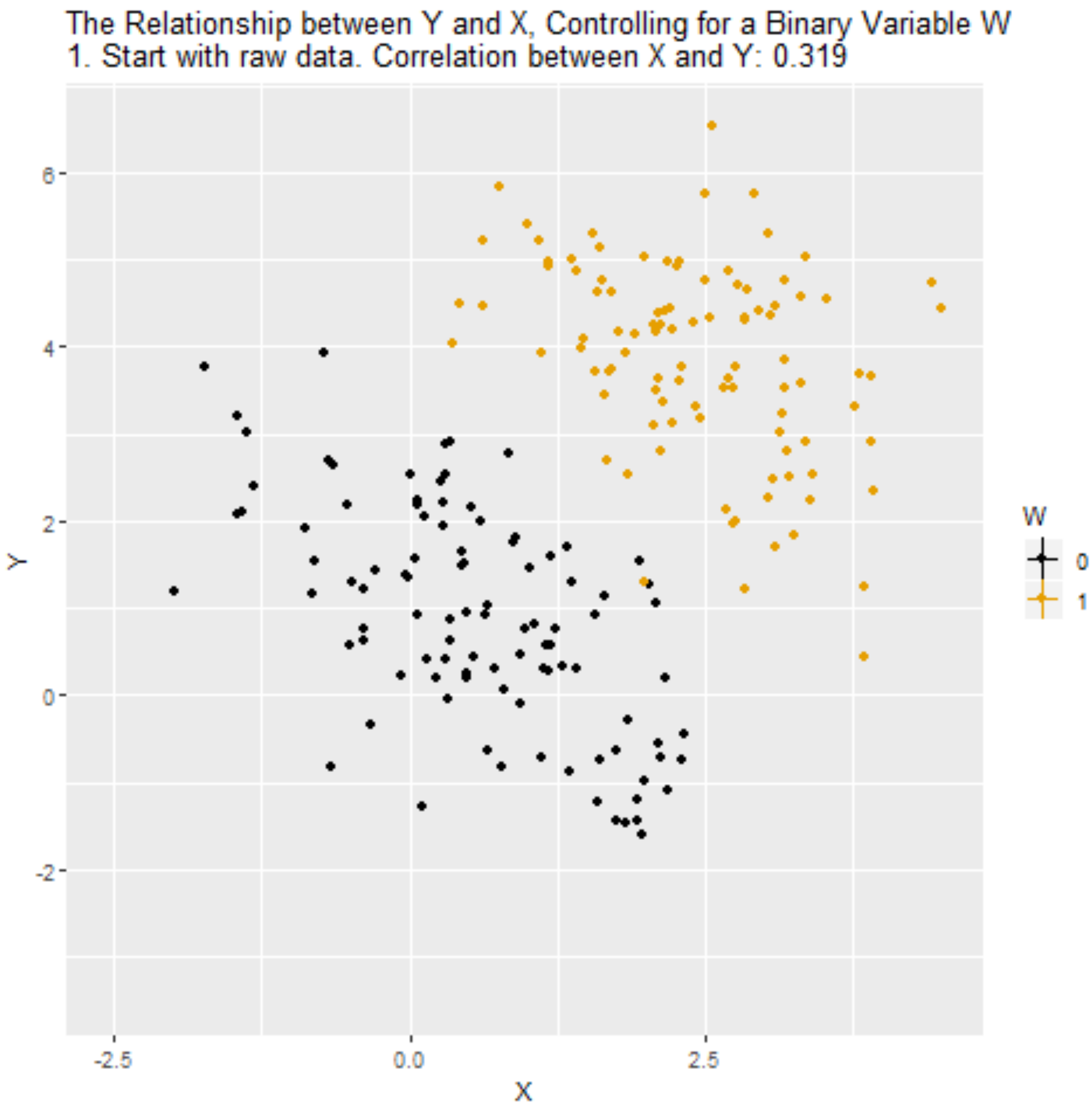
Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum (2017) Eye-tracking causality. *Psychological Science*

02:00

stretch break!

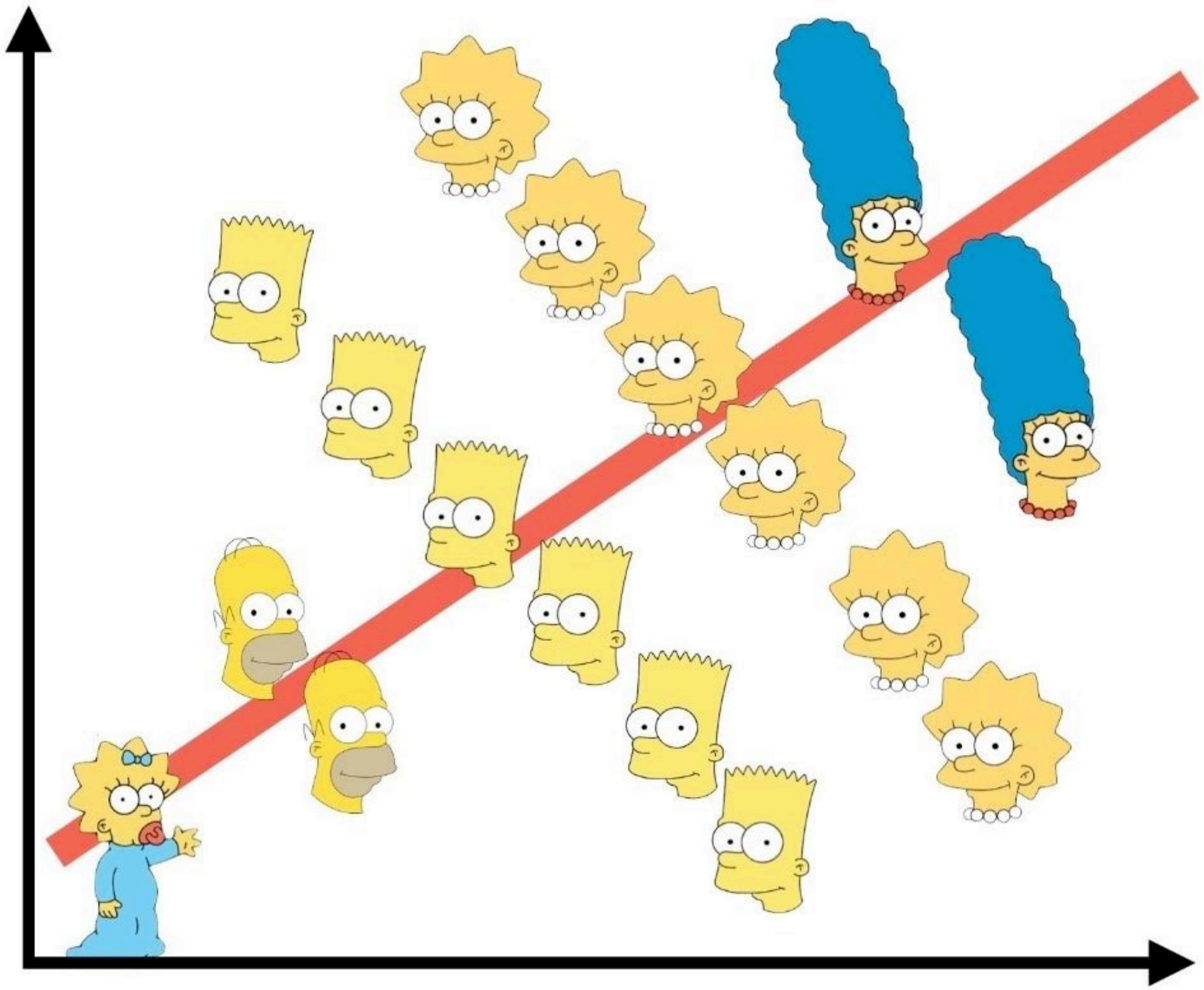


Controlling for variables



https://nickchk.com/causalgraphs.html

Simpson's paradox



- when the relationship between two variables changes strongly after conditioning on another variable
- interesting real world cases
- **google** it!

What does controlling for variables mean?

we are not actually "**controlling**" the variable

instead, we are taking the variable into consideration when making predictions

the hope is that we get a better estimate of the parameter that we are interested in by taking into account other factors

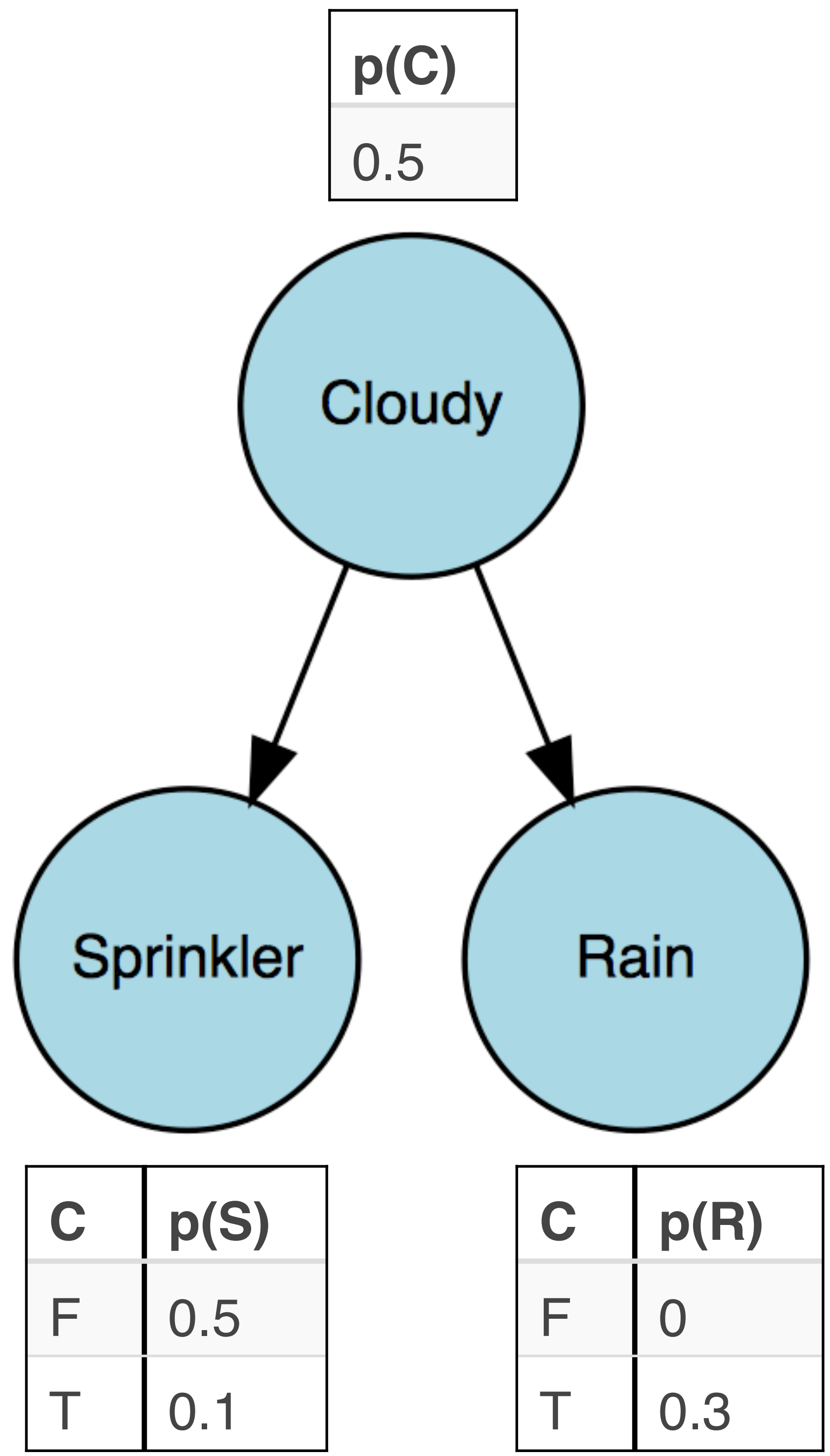
Patterns of inference: **Common cause**

$$p(S\,|\,R) = p(S)$$

or

$$p(S\,|\,R) \neq p(S)$$

?



Patterns of inference: **Common cause**

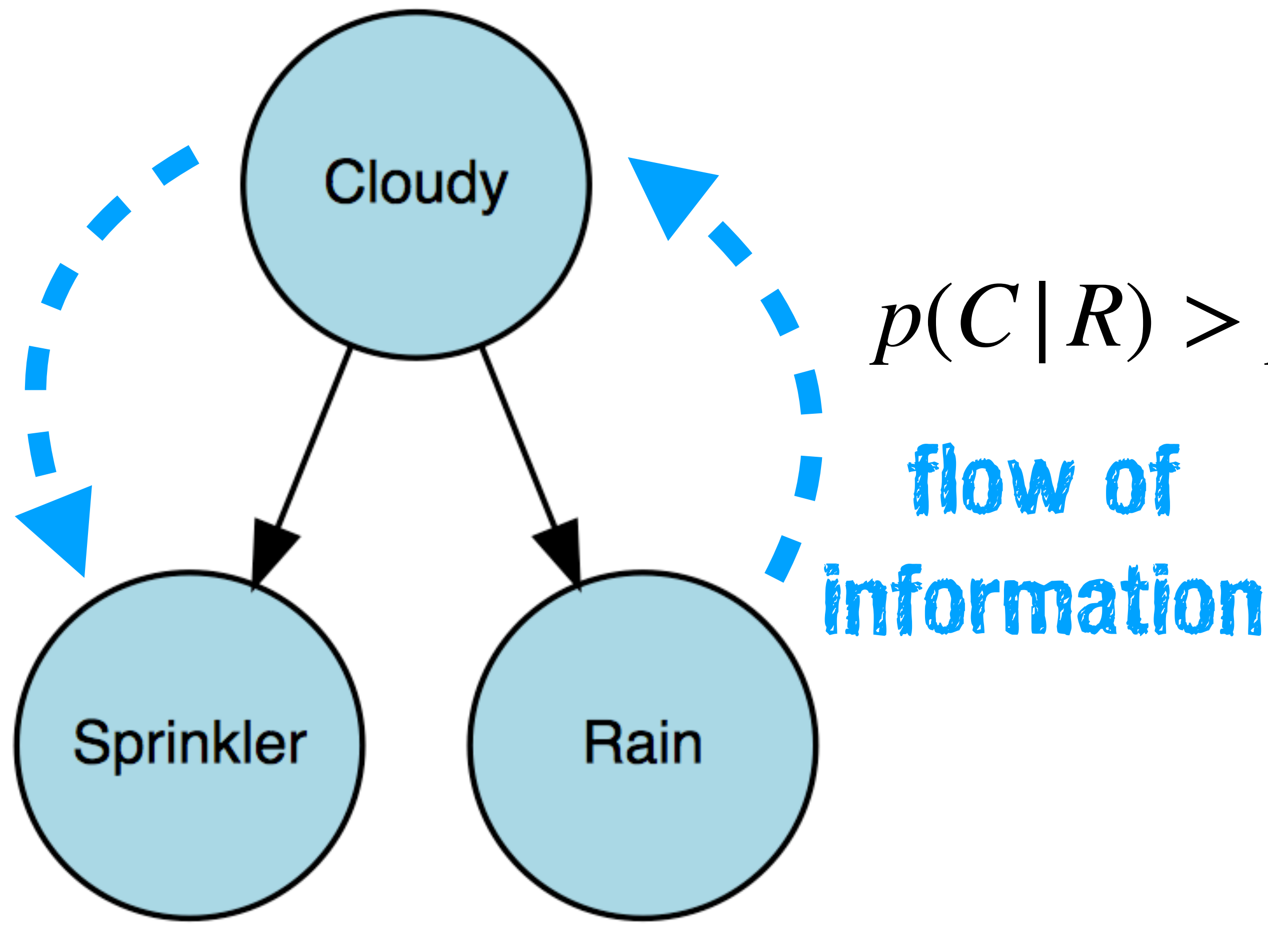
- effects of a common cause are *unconditionally dependent*

$$p(S \mid R) \neq p(S)$$

p(C)
0.5

$p(S \mid C) < p(S)$

$p(C \mid R) > p(C)$



C	p(S)
F	0.5
T	0.1

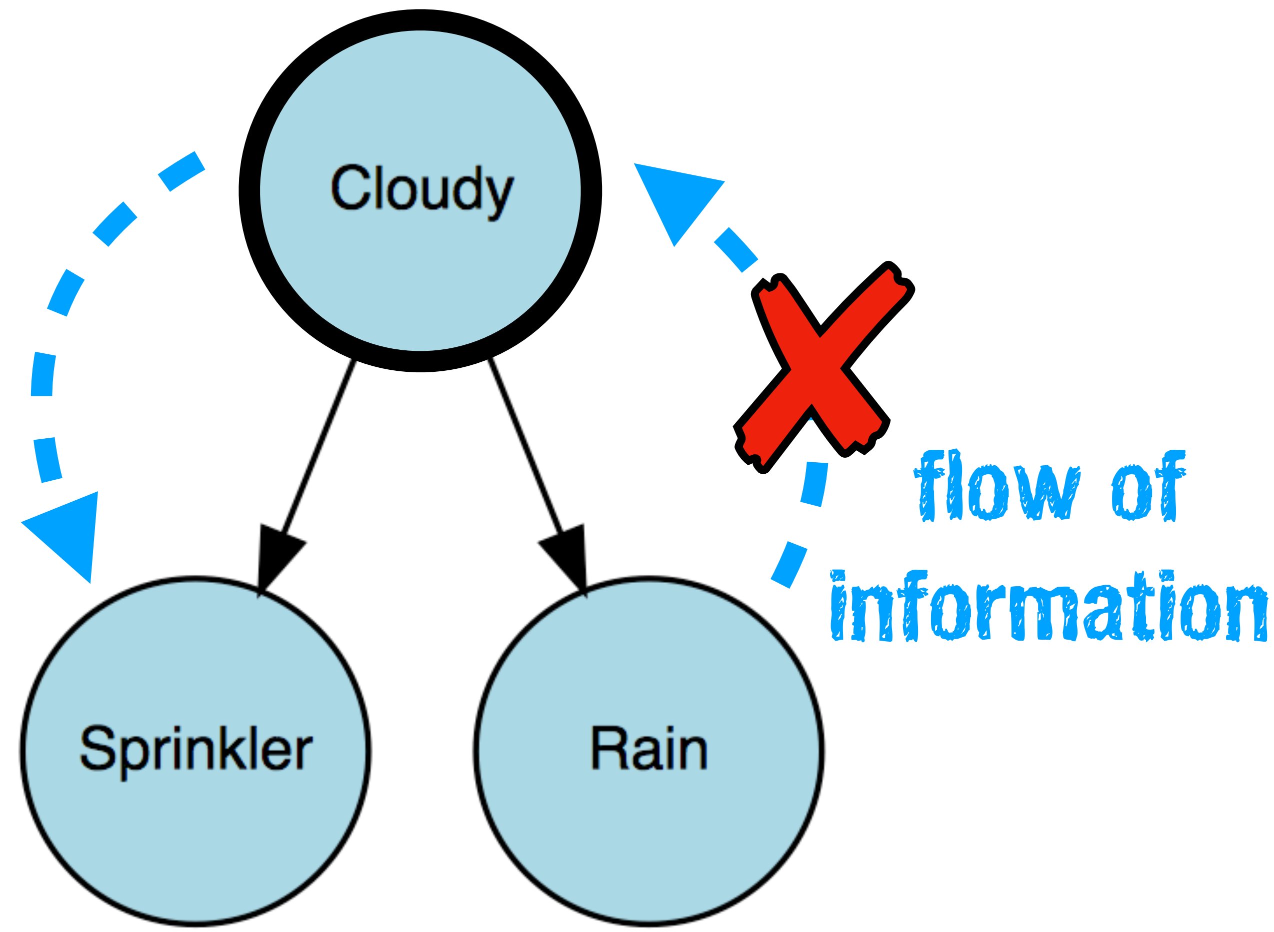
C	p(R)
F	0
T	0.3

Patterns of inference: **Common cause**

- effects of a common cause are *conditionally independent given the cause*

$$p(S \mid R, C) = p(S \mid C)$$

p(C)
1



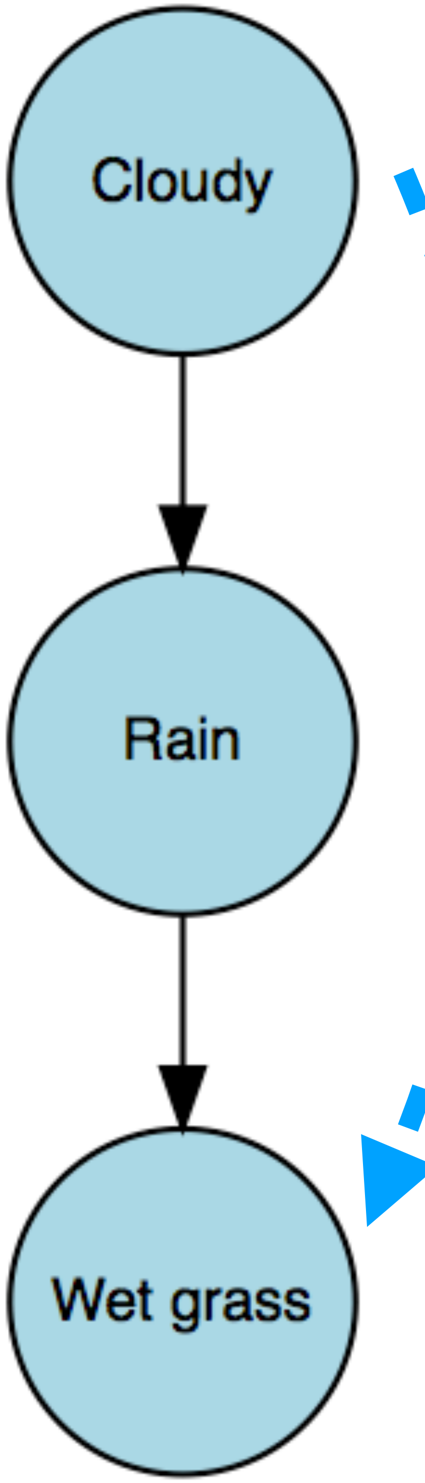
C	p(S)
F	0.5
T	0.1

C	p(R)
F	0
T	0.3

Patterns of inference: **Causal chain**

- cause and effect in a causal chain are *unconditionally dependent*

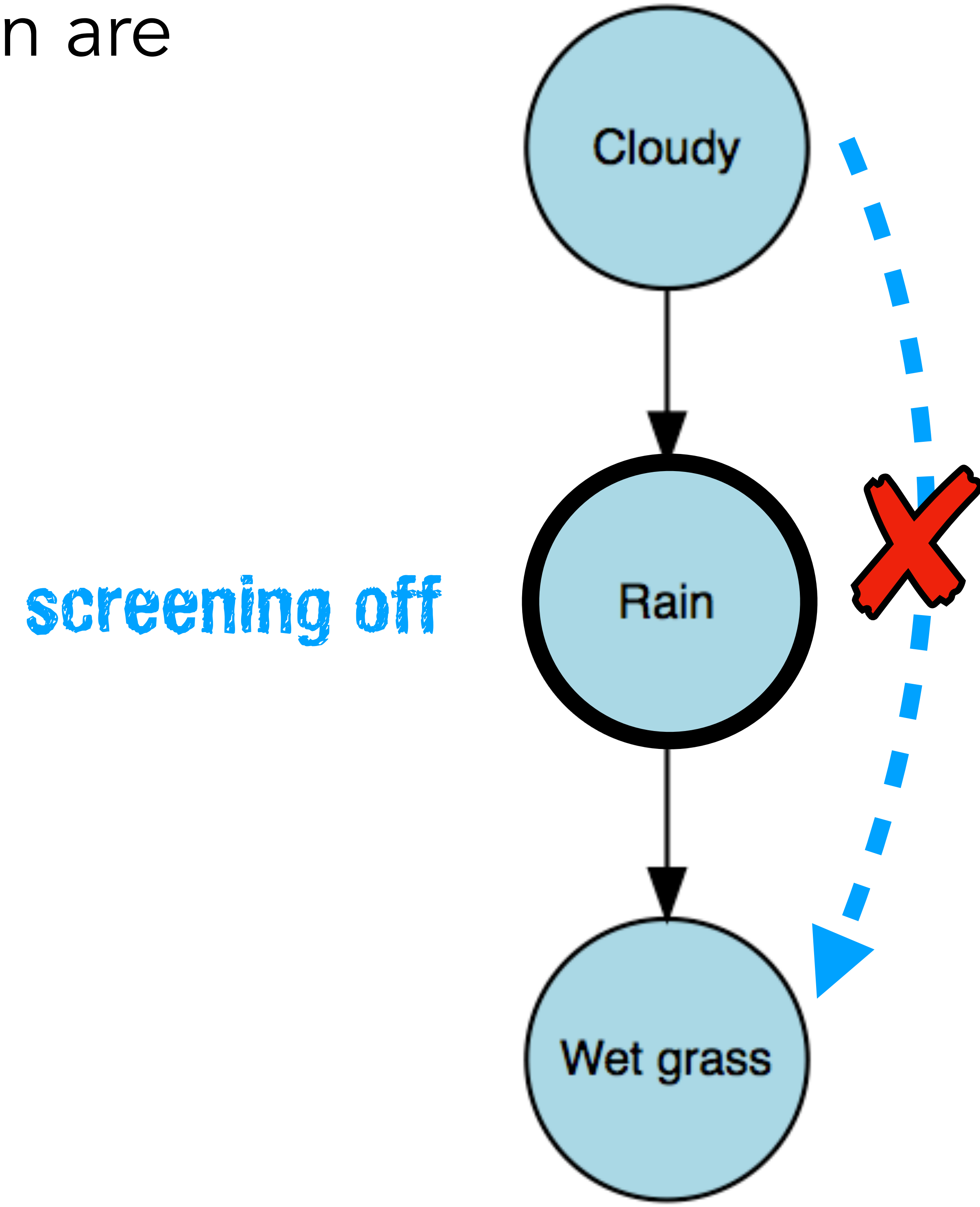
$$p(W \mid C) \neq p(W)$$



Patterns of inference: **Causal chain**

- cause and effect in a causal chain are *conditionally independent*

$$p(W \mid C, R) = p(W \mid R)$$

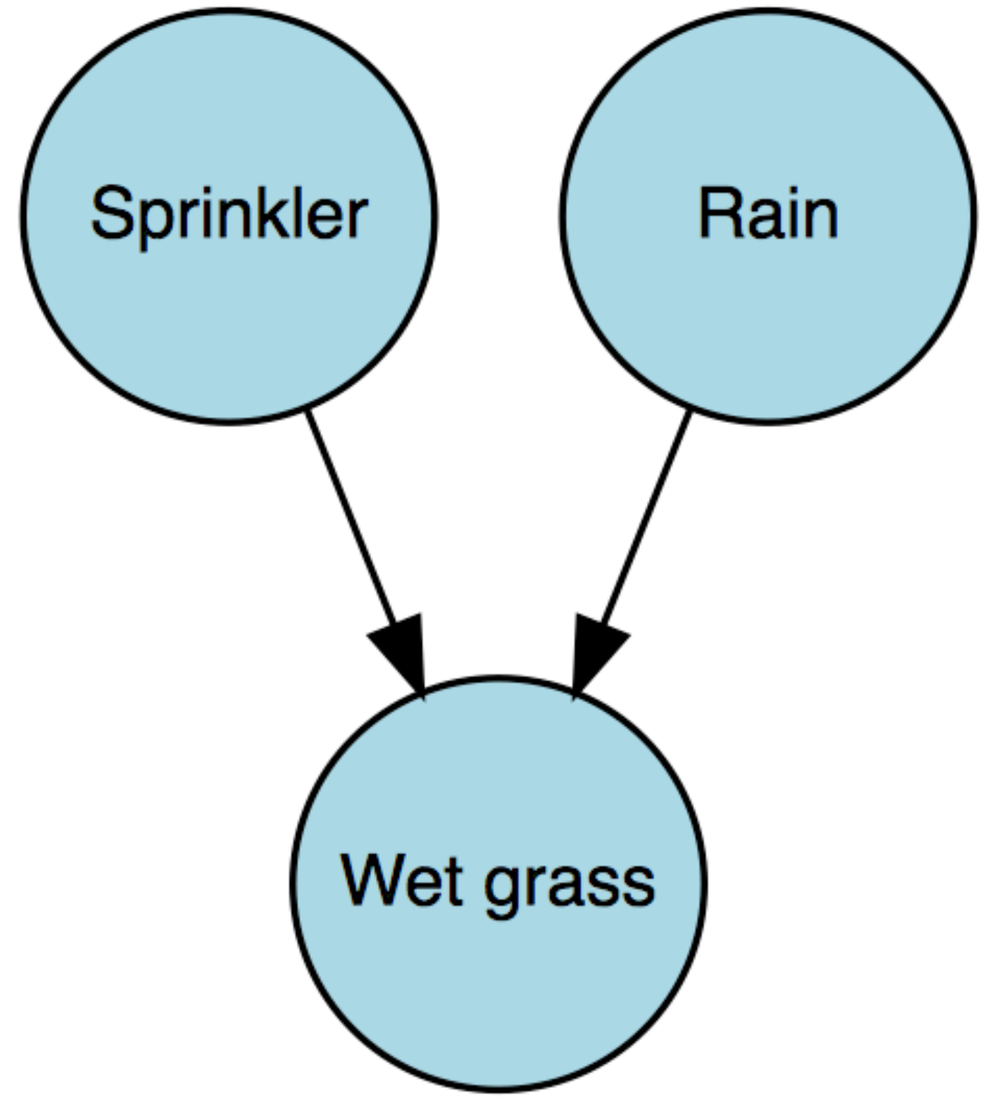


Patterns of inference: **Common effect**

- two causes of a common effect are *unconditionally independent*

$$p(S \mid R) = p(S)$$

(e.g. Sprinkler is set by a timer)

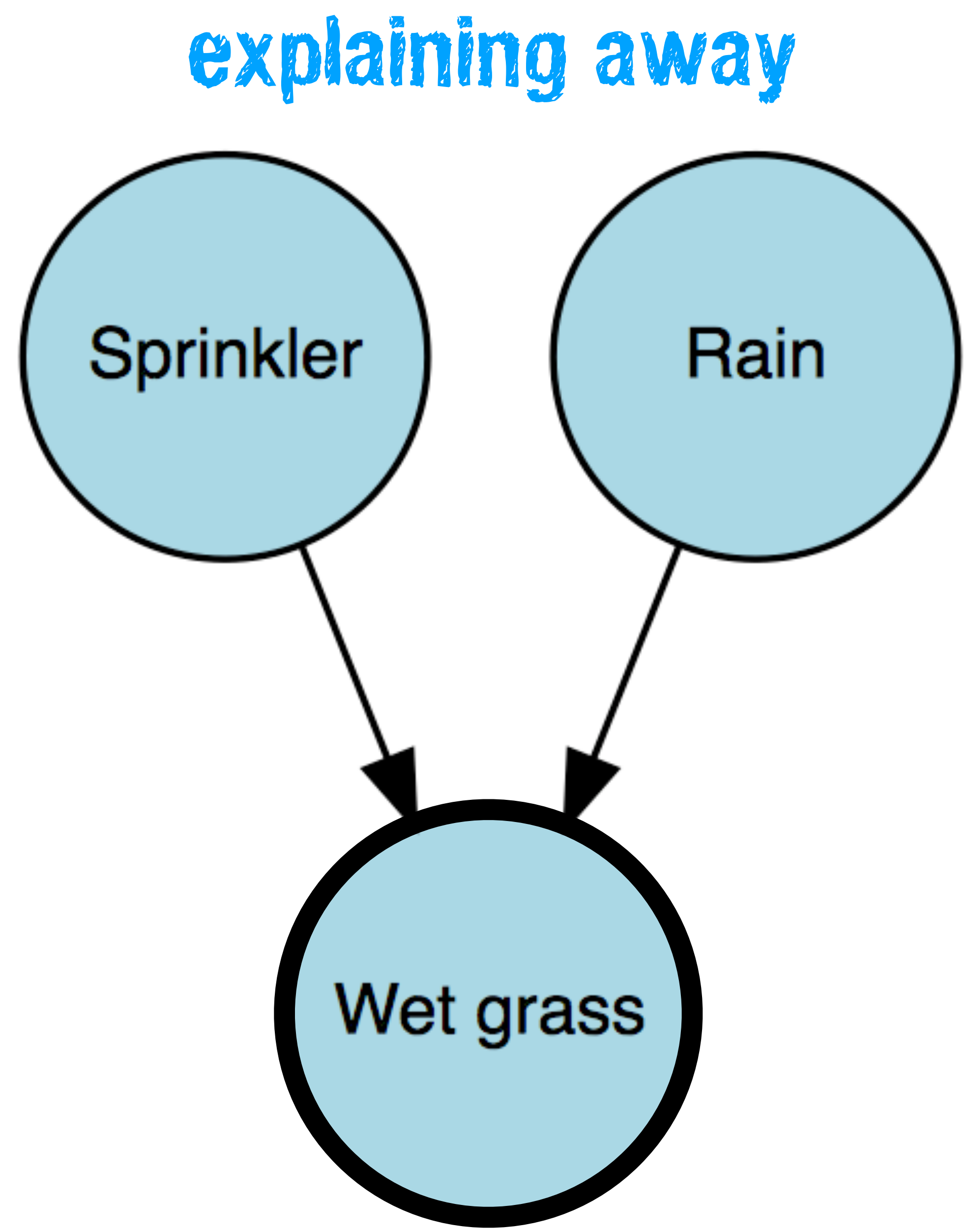


Patterns of inference: **Common effect**

- two causes of a common effect are *conditionally dependent given the effect*

$$p(S \mid R, W) \neq p(S \mid W)$$

- intuitively: both causes compete to explain the effect



Note: The pattern of inference depends on the structural form which captures how `Sprinkler` and `Rain` jointly affect `Wet grass`. Explaining away holds for the commonly used noisy-or integration function.

When should I control for variables?

recent advances in graphical models have produced a way to help distinguish good from bad controls

***d*-separation**

directional

decide from a causal graph whether a set of variables `X` is independent of another set `Y`, given a third set `Z`

Goal: we want a precise (and unbiased) estimate of the predictive relationship between `X` and `Y`

we want to block all other paths from X to Y

http://bayes.cs.ucla.edu/BOOK-09/ch11-1-2-final.pdf

When should I control for variables?

How can I tell whether two variables are independent?

Recipe for independence

1. Draw the ancestral graph
2. "Moralize" the graph by "marrying" the parents
3. "Disorient" the graph by replacing arrows with edges
4. Delete the givens and their edges
5. Read the answer off the graph

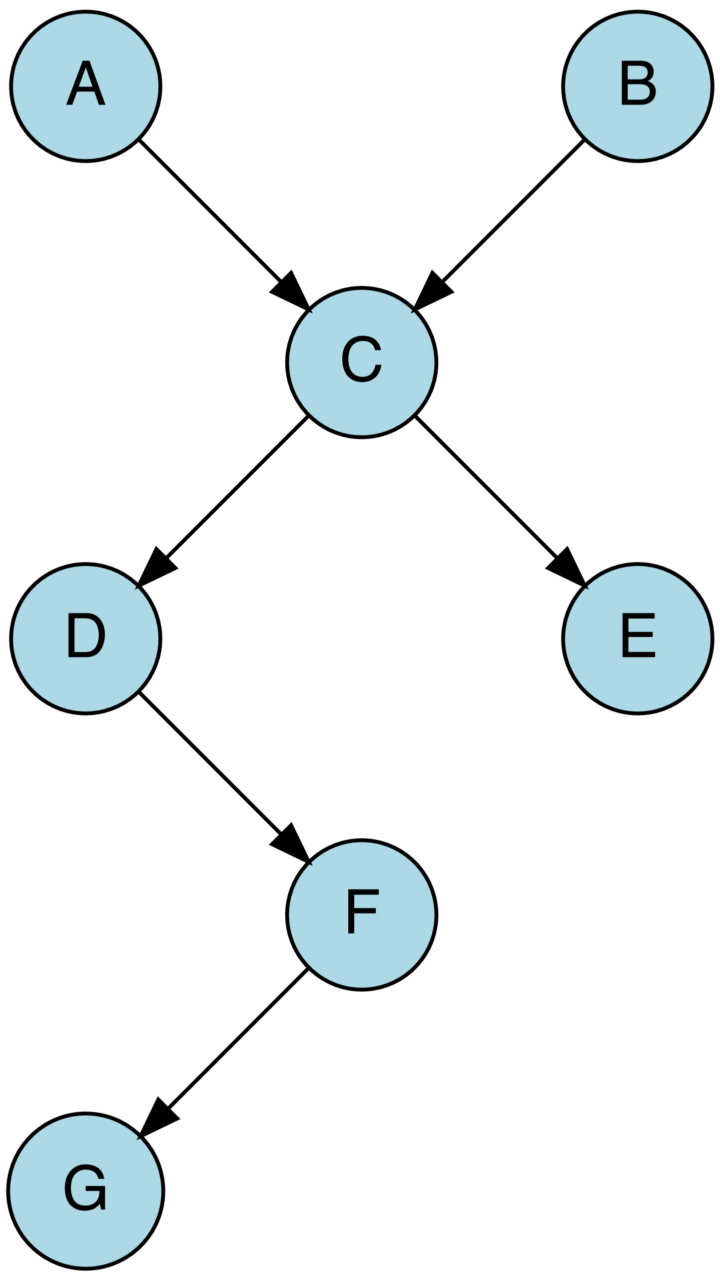
- if variables are **disconnected** they are independent
- if variables are connected (have a path between them) they are not guaranteed to be independent

http://web.mit.edu/jmn/www/6.034/d-separation.pdf

When should I control for variables?

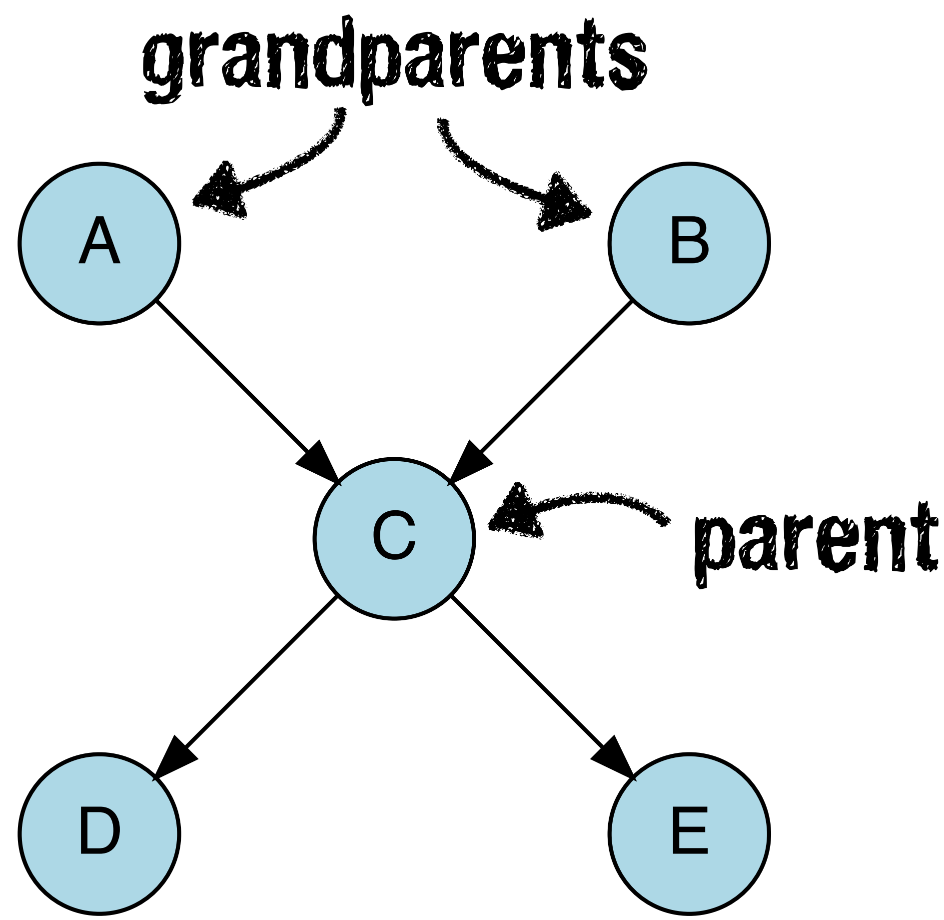
Are D and E independent?

$$p(D \mid E) = p(D) \text{ ?}$$



1. Draw the ancestral graph

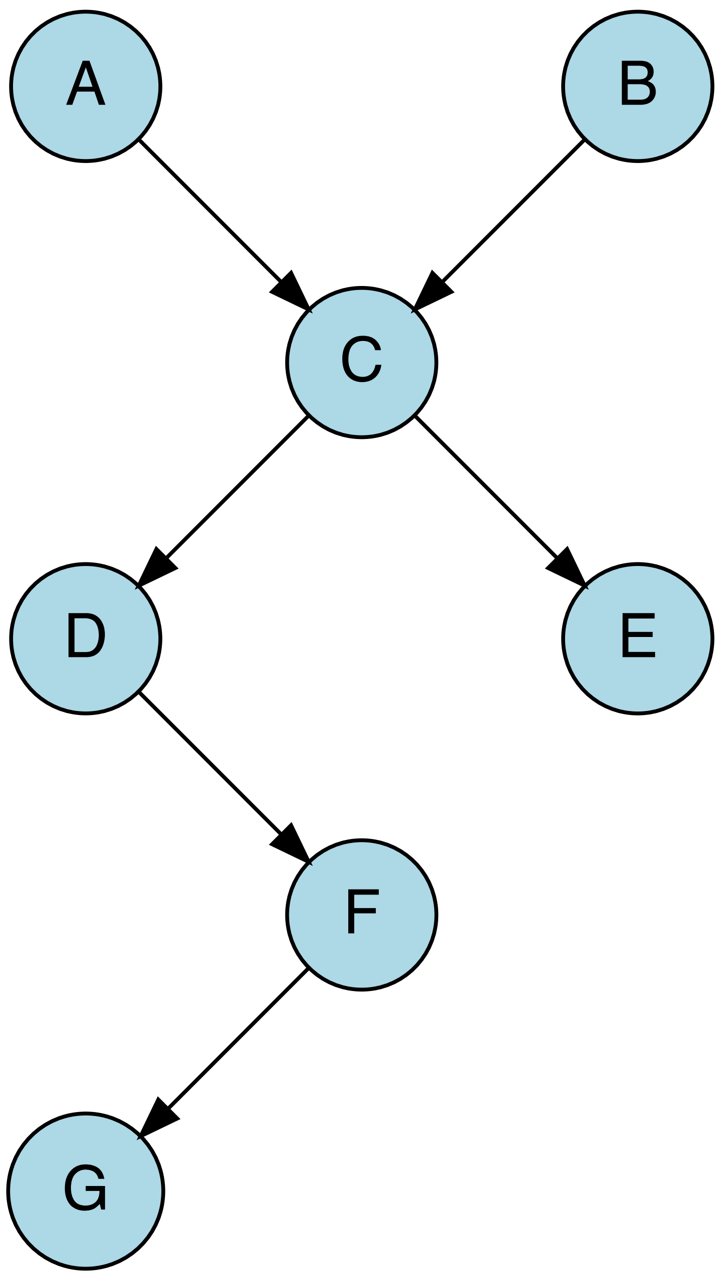
Construct the "ancestral graph" of all variables mentioned in the probability expression. This is a reduced version of the original net, consisting only of the variables mentioned and all of their ancestors (parents, parents' parents, etc.)



When should I control for variables?

Are D and E independent?

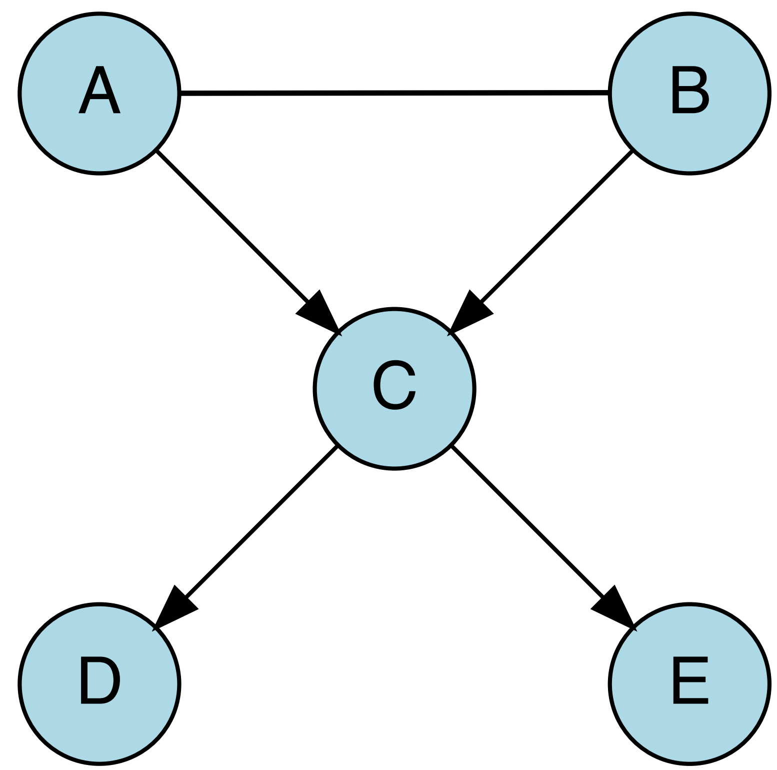
$$p(D\,|\,E) = p(D)\ ?$$



2. "Moralize" the graph

let's get married!

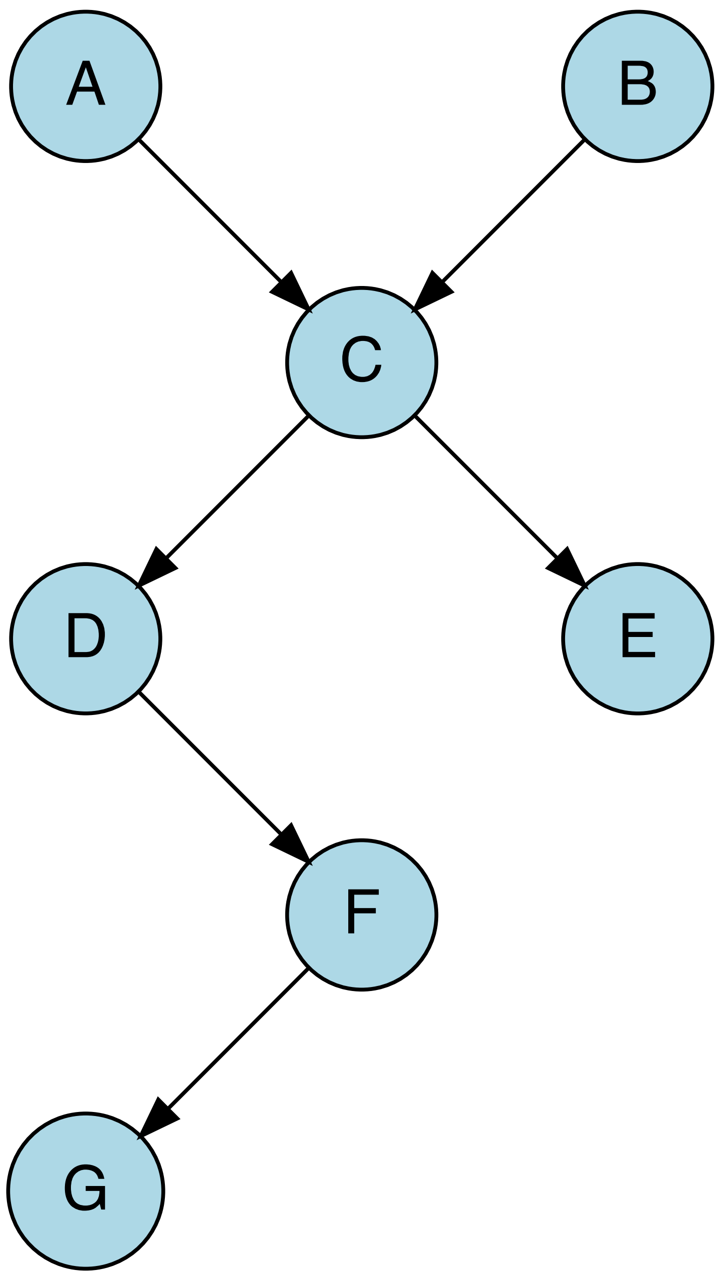
For each pair of variables with a common child, draw an undirected edge (line) between them. (If a variable has more than two parents, draw lines between every pair of parents.)



When should I control for variables?

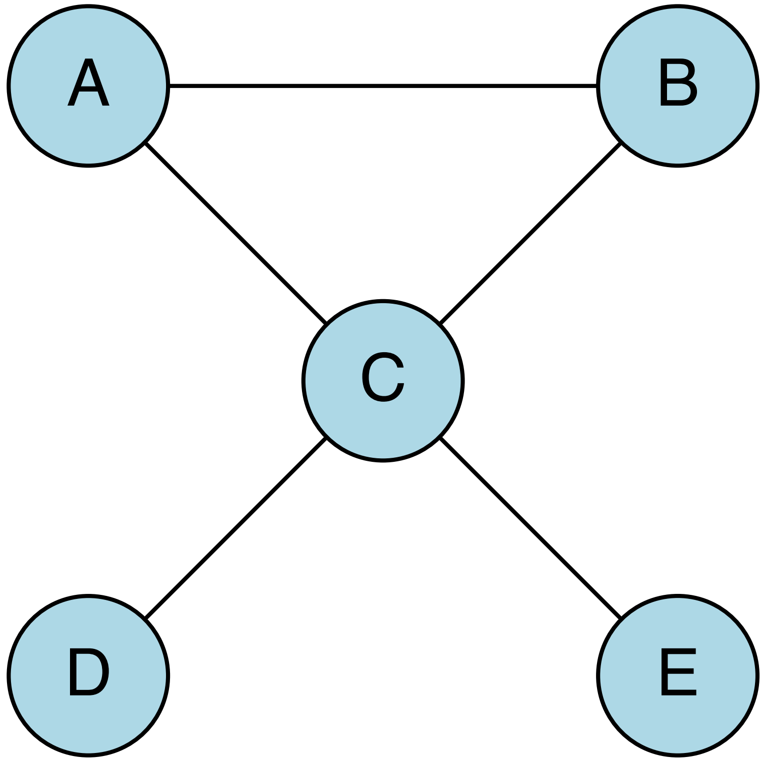
Are D and E independent?

$$p(D \mid E) = p(D) \text{ ?}$$



3. "Disorient" the graph

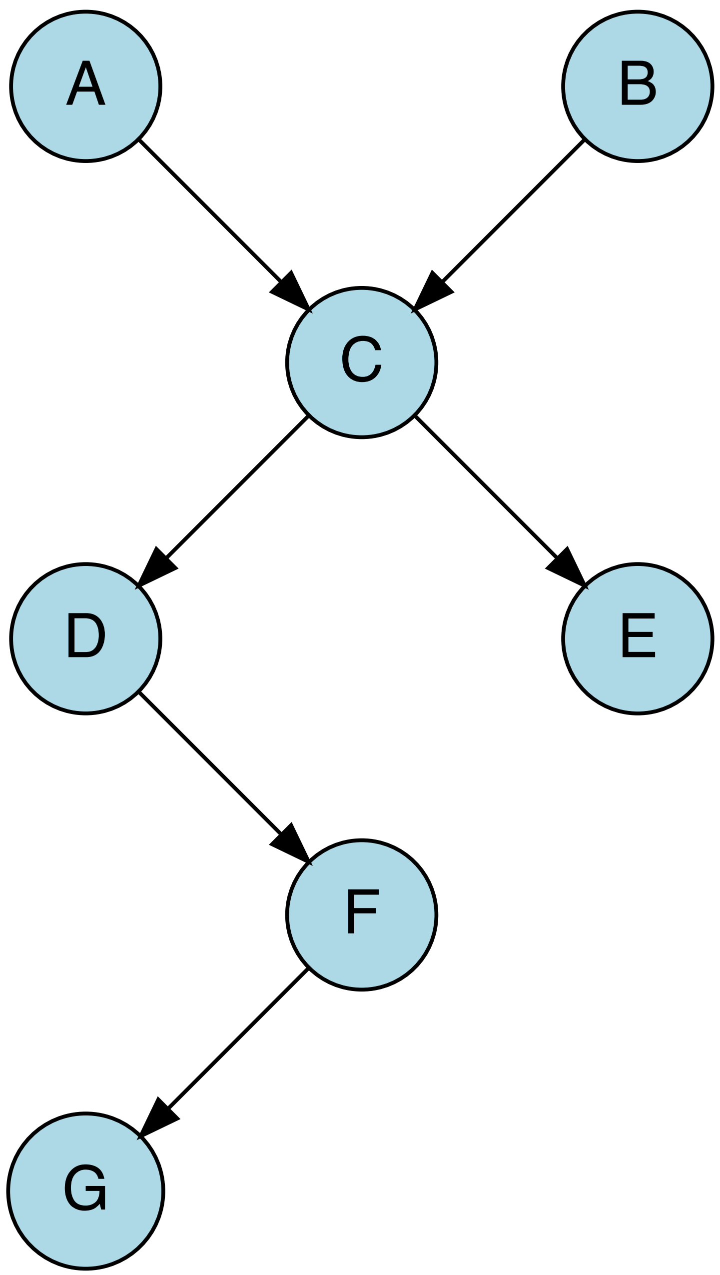
Replace arrows arrows with lines



When should I control for variables?

Are D and E independent?

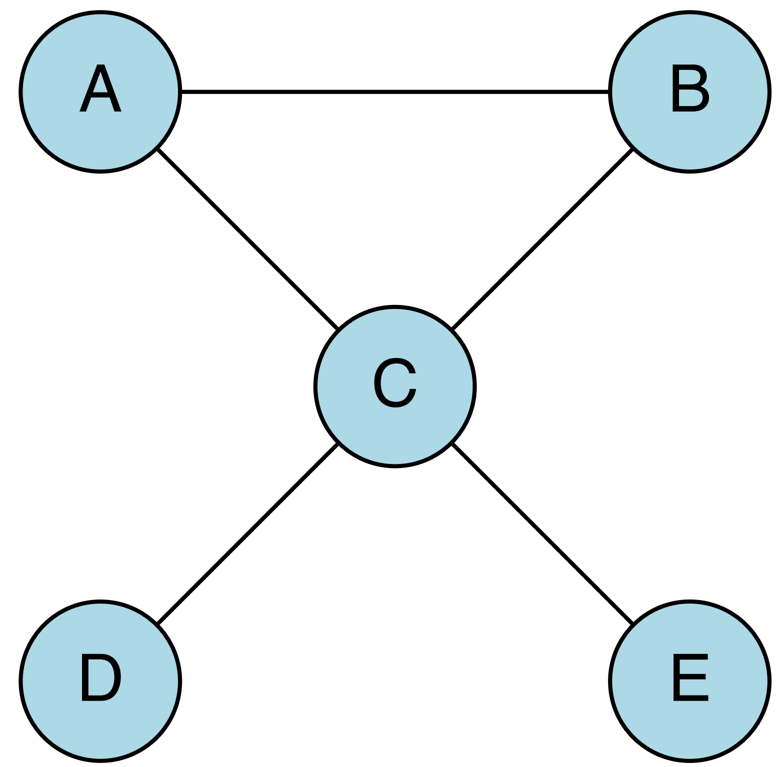
$$p(D \mid E) = p(D) \text{ ?}$$



4. Delete the givens

Remove the variables that we condition on, as well as their edges

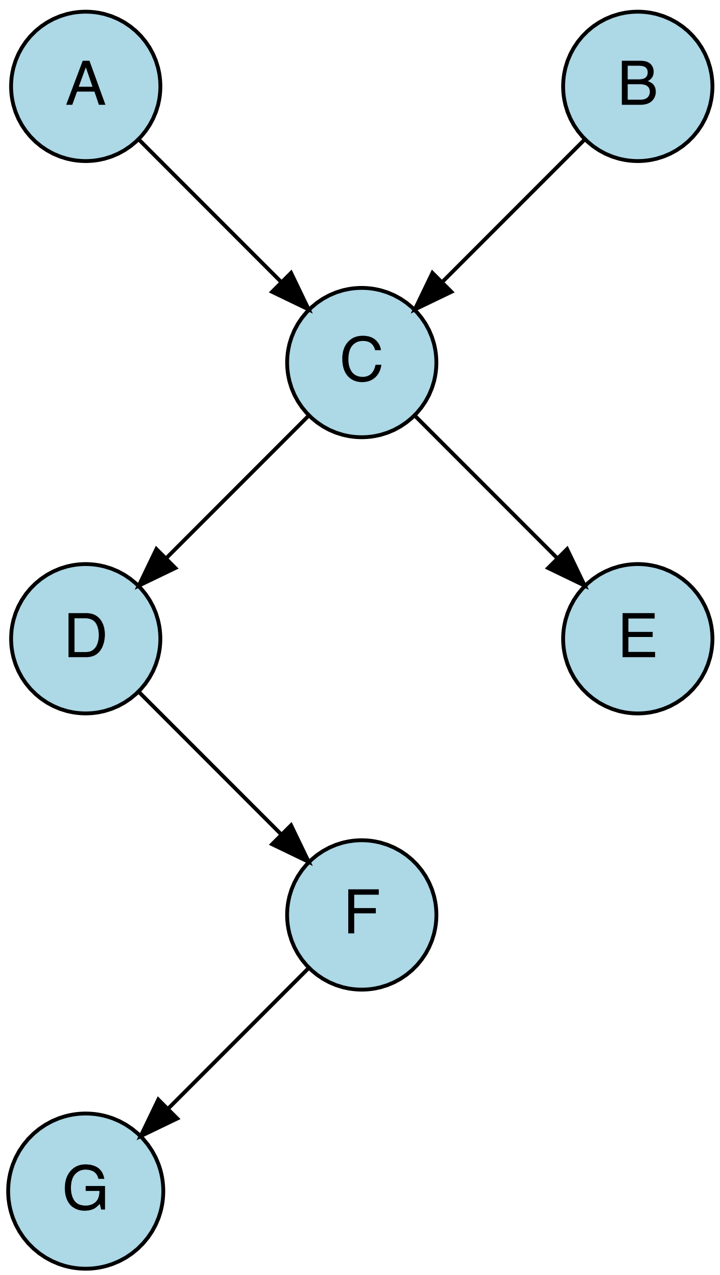
we didn't condition on anything, so there is nothing to delete



When should I control for variables?

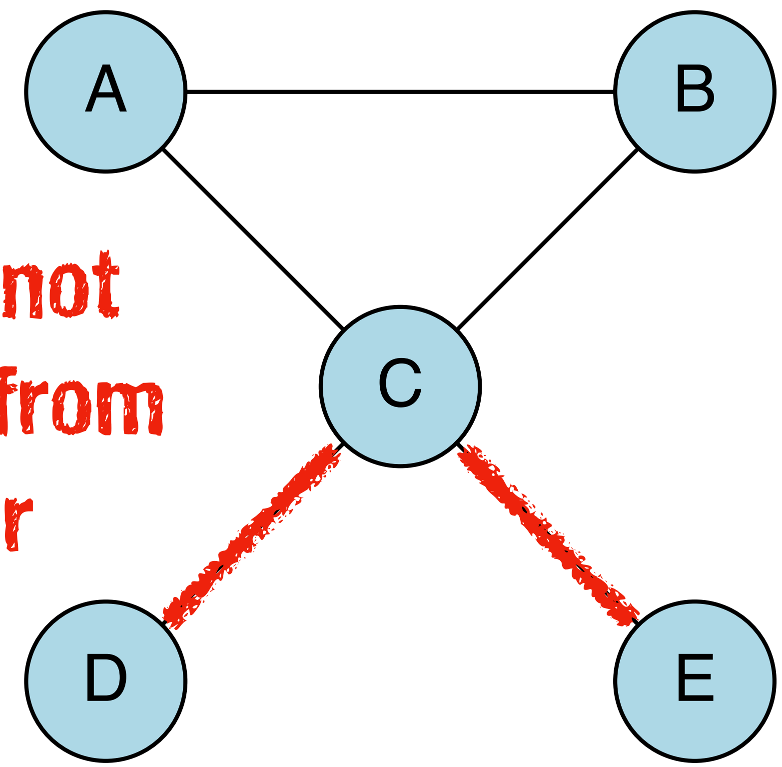
Are D and E independent?

$$p(D \mid E) = p(D) \textbf{ ?}$$



5. Read answer off the graph

- if variables are **disconnected** they are independent
- if variables are connected (have a path between them) they are not guaranteed to be independent



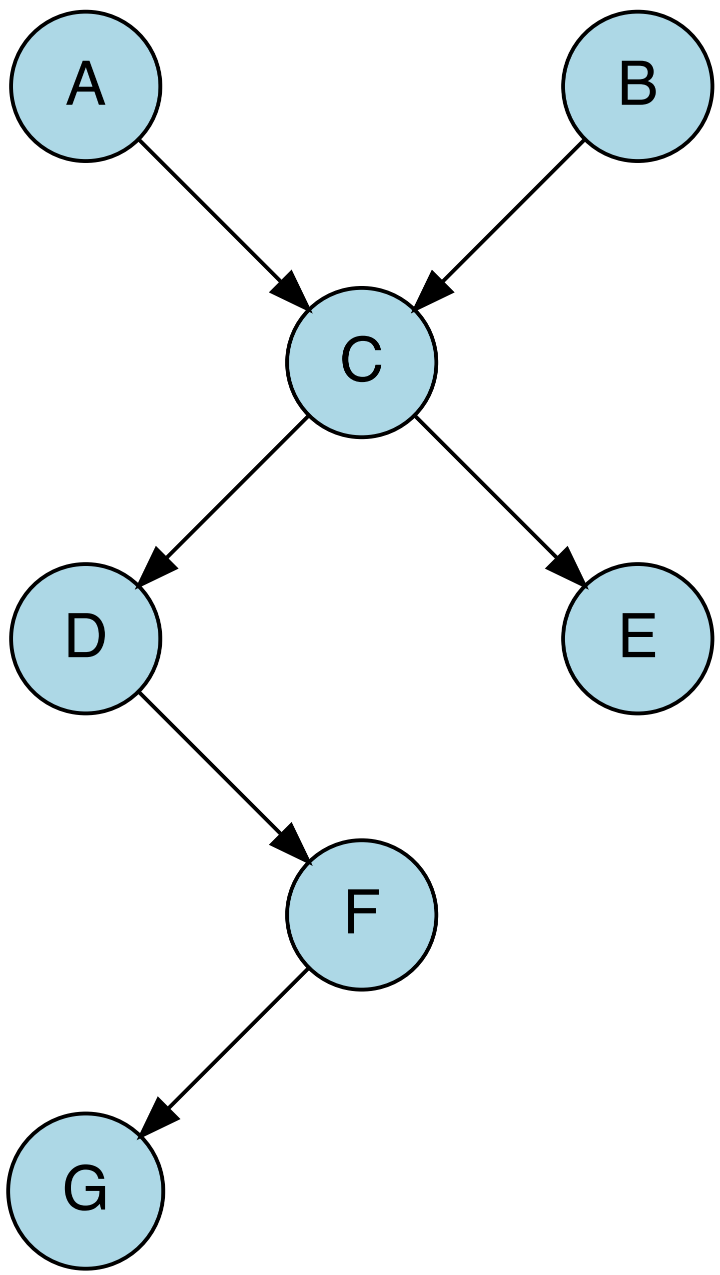
D and E are not independent from each other

they are connected via at least one path

When should I control for variables?

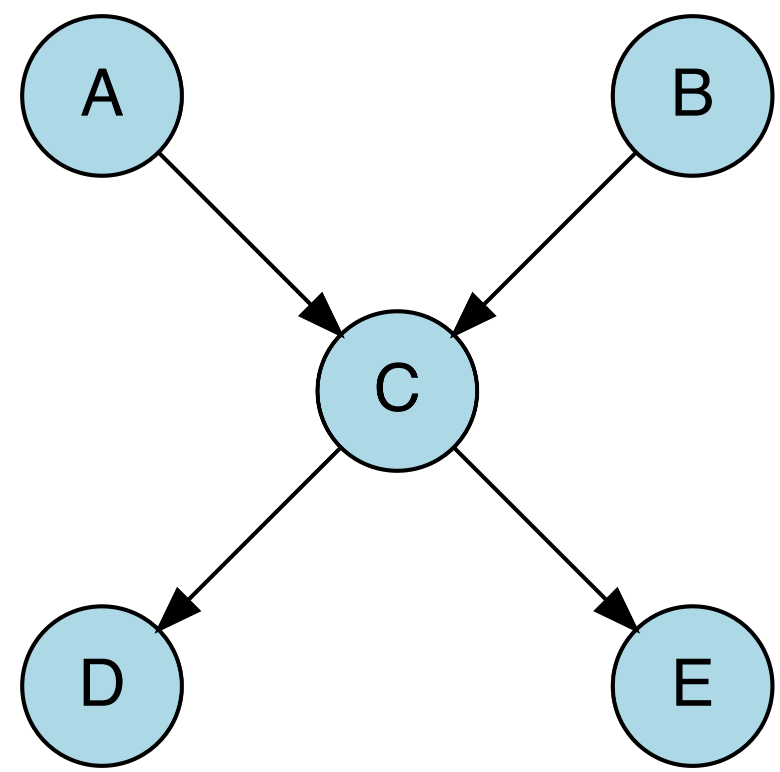
Are D and E independent, given C?

$$p(D \mid E, C) = p(D \mid C) \ ?$$



1. Draw the ancestral graph

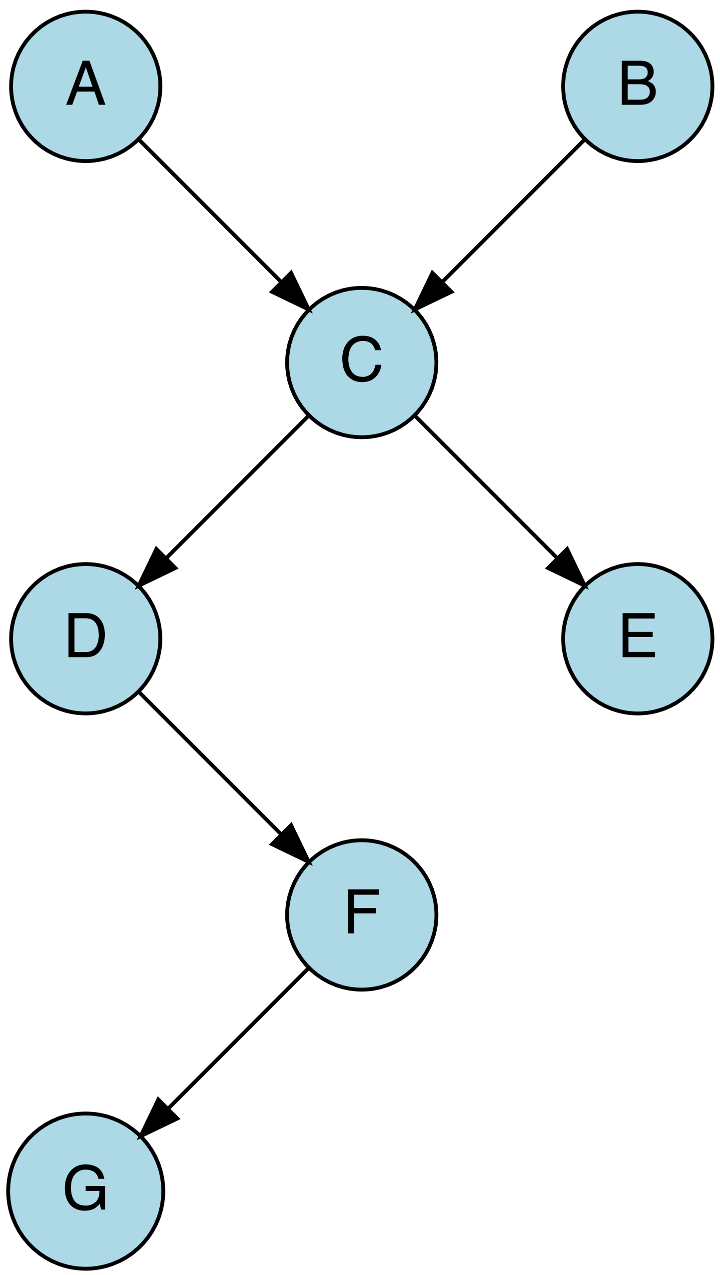
Construct the "ancestral graph" of all variables mentioned in the probability expression. This is a reduced version of the original net, consisting only of the variables mentioned and all of their ancestors (parents, parents' parents, etc.)



When should I control for variables?

Are D and E independent, given C?

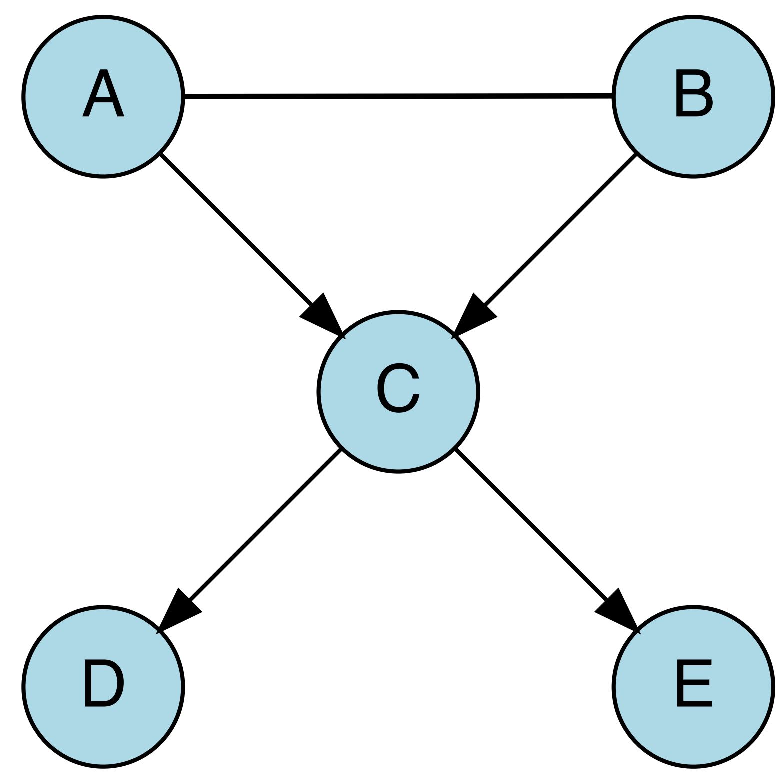
$$p(D \mid E, C) = p(D \mid C) \ ?$$



2. "Moralize" the graph

let's get married!

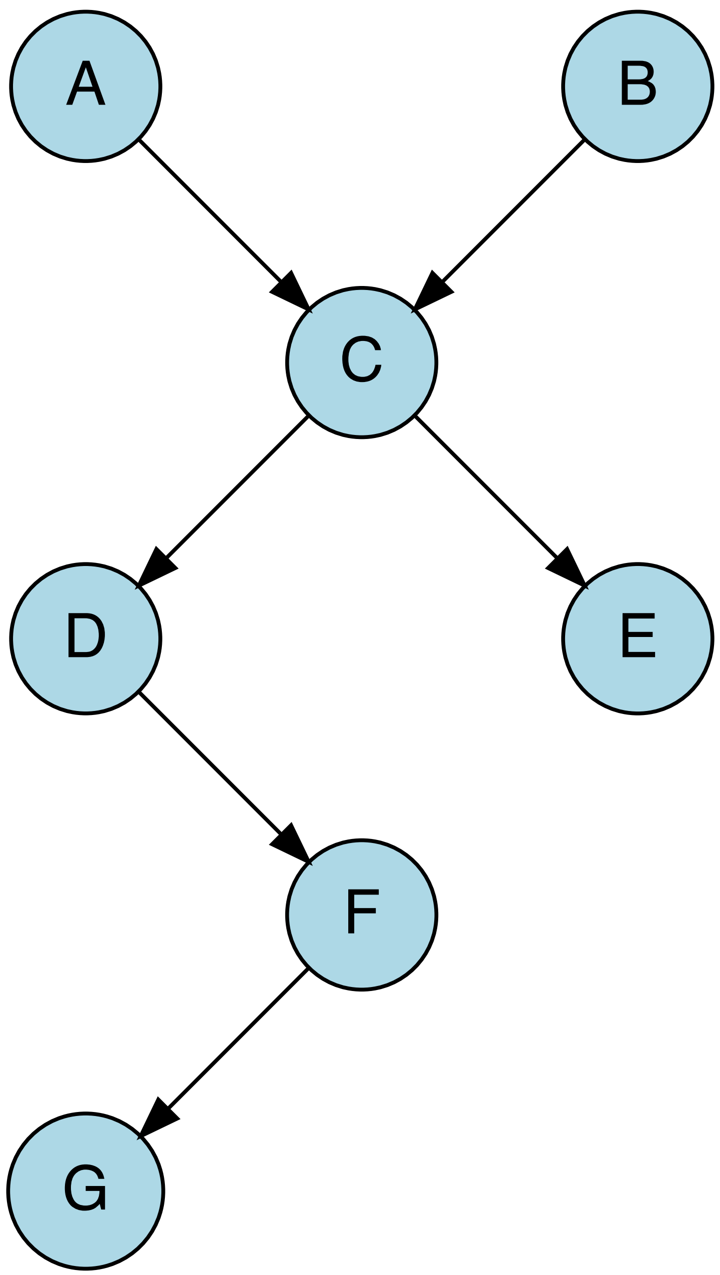
For each pair of variables with a common child, draw an undirected edge (line) between them. (If a variable has more than two parents, draw lines between every pair of parents.)



When should I control for variables?

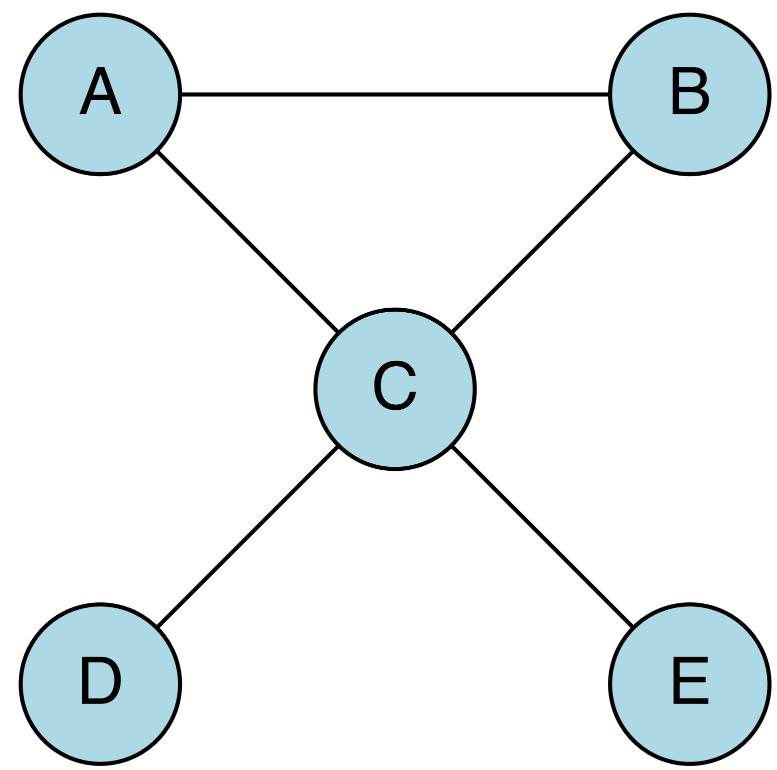
Are D and E independent, given C?

$$p(D \mid E, C) = p(D \mid C) \;?$$



3. "Disorient" the graph

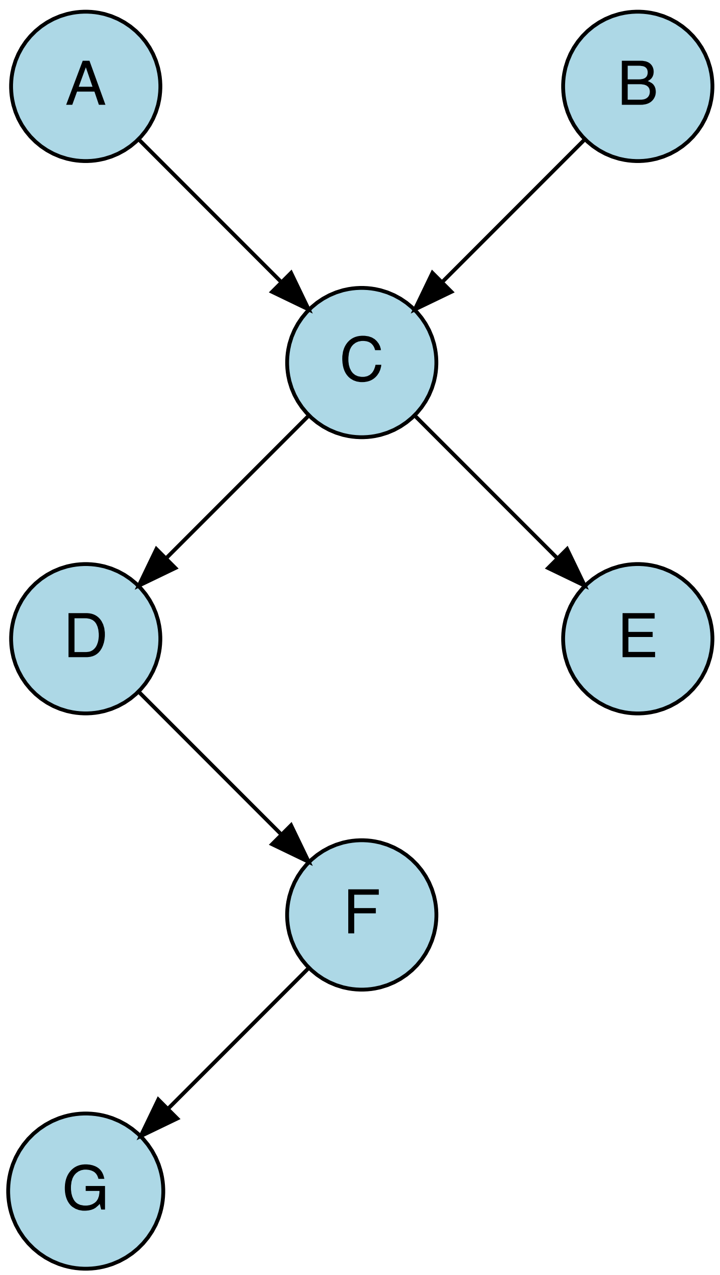
Replace arrows arrows with lines



When should I control for variables?

Are D and E independent, given C?

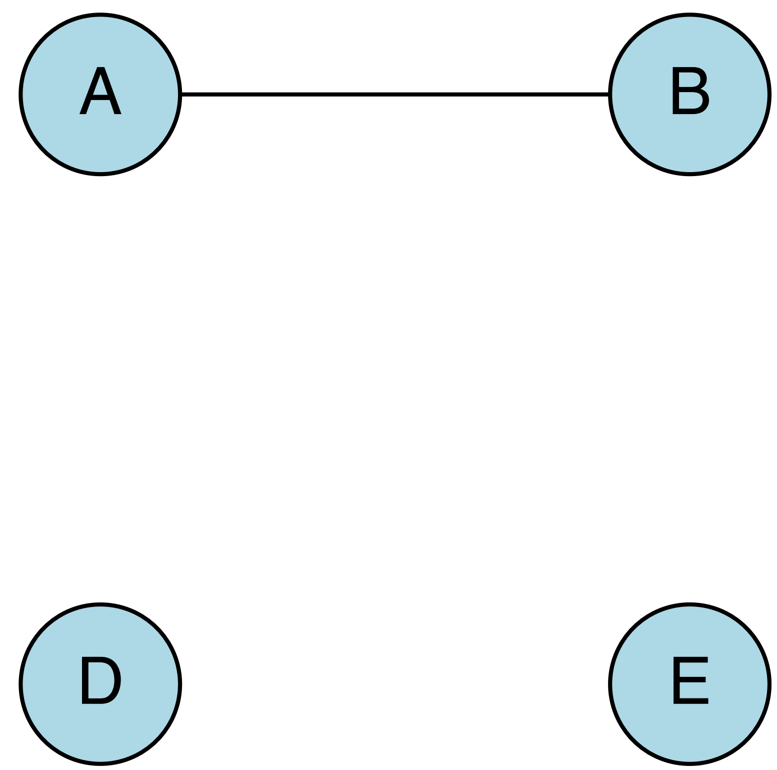
$$p(D \mid E, C) = p(D \mid C) \text{ ?}$$



4. Delete the givens

Remove the variables that we condition on, as well as their edges

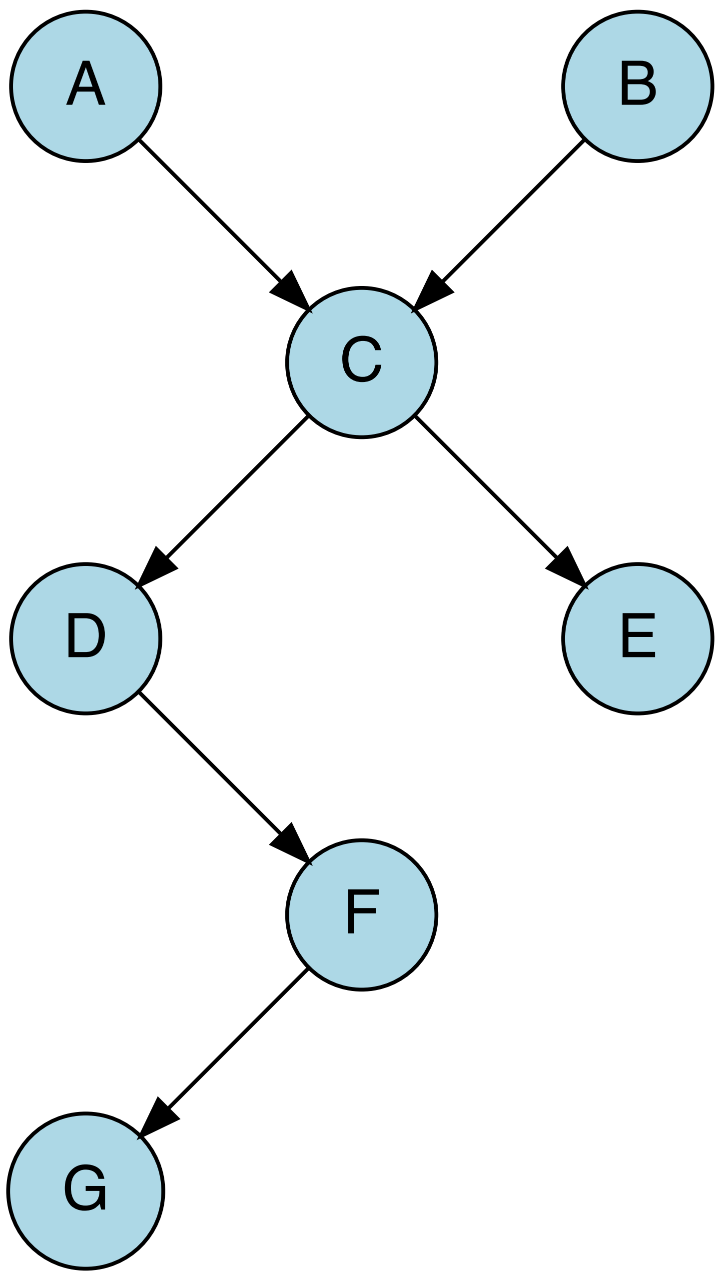
we conditioned on C!



When should I control for variables?

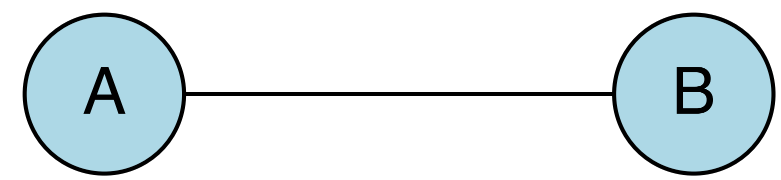
Are D and E independent, given C?

$$p(D \mid E, C) = p(D \mid C) \text{ ?}$$



5. Read answer off the graph

- if variables are **disconnected** they are independent
- if variables are connected (have a path between them) they are not guaranteed to be independent



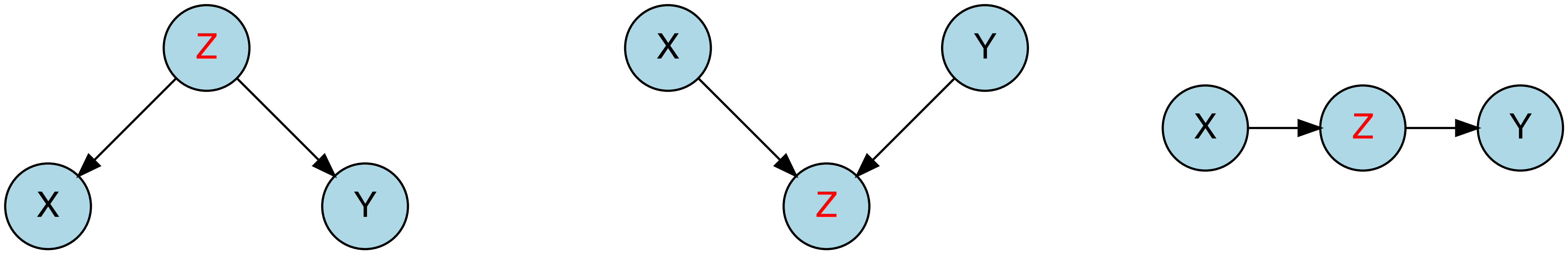
D and E are independent from each other conditioned on C

they aren't connected via a path

So what?

Patterns of inference

We want to estimate the (causal) relationship between `X` and `Y`



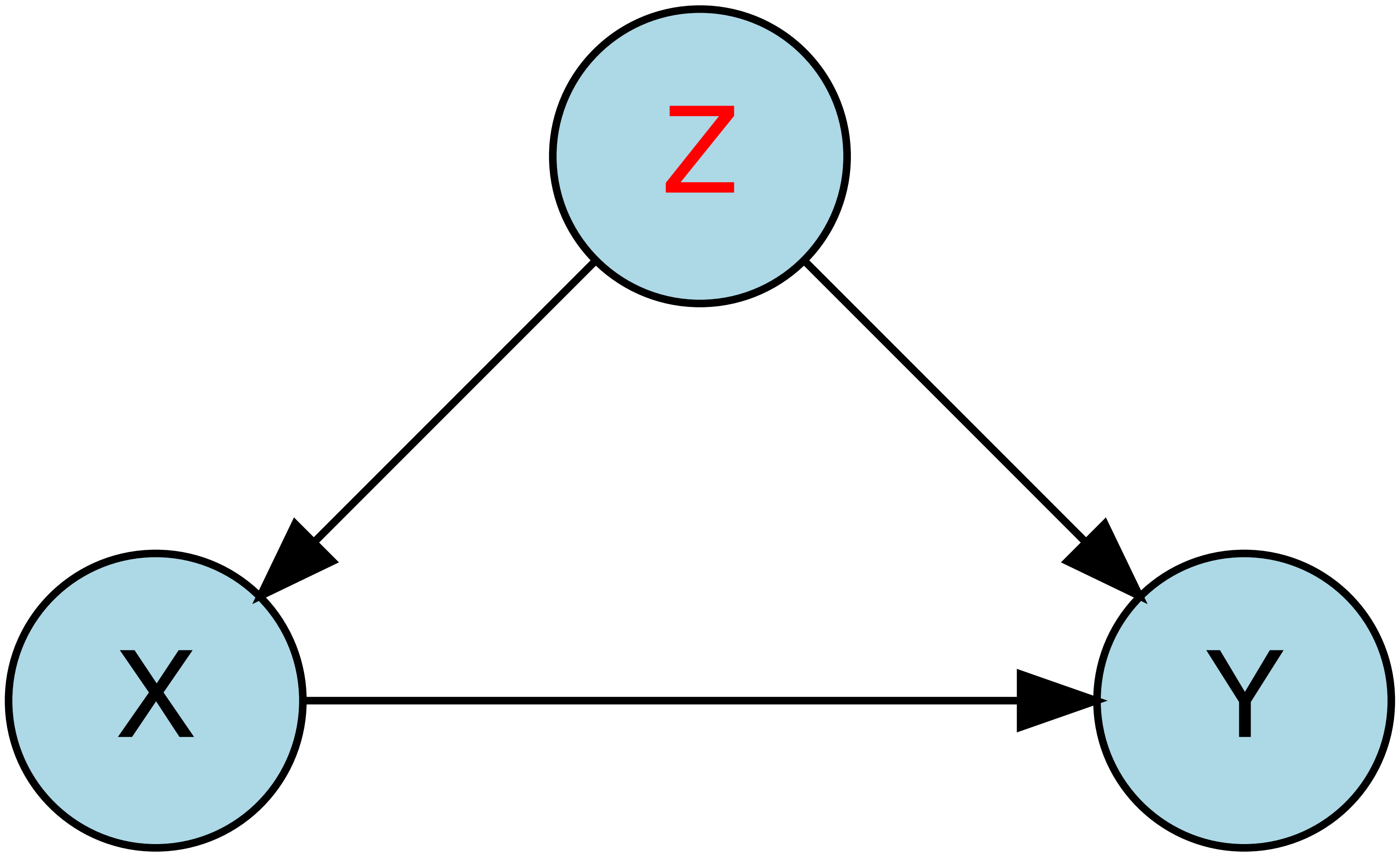
common cause | common effect | causal chain

by controlling for `Z` we hope to get a better estimate of the relationship between `X` and `Y`

d-separation helps us tell apart **good controls** from **bad controls**

When should I control for variables?

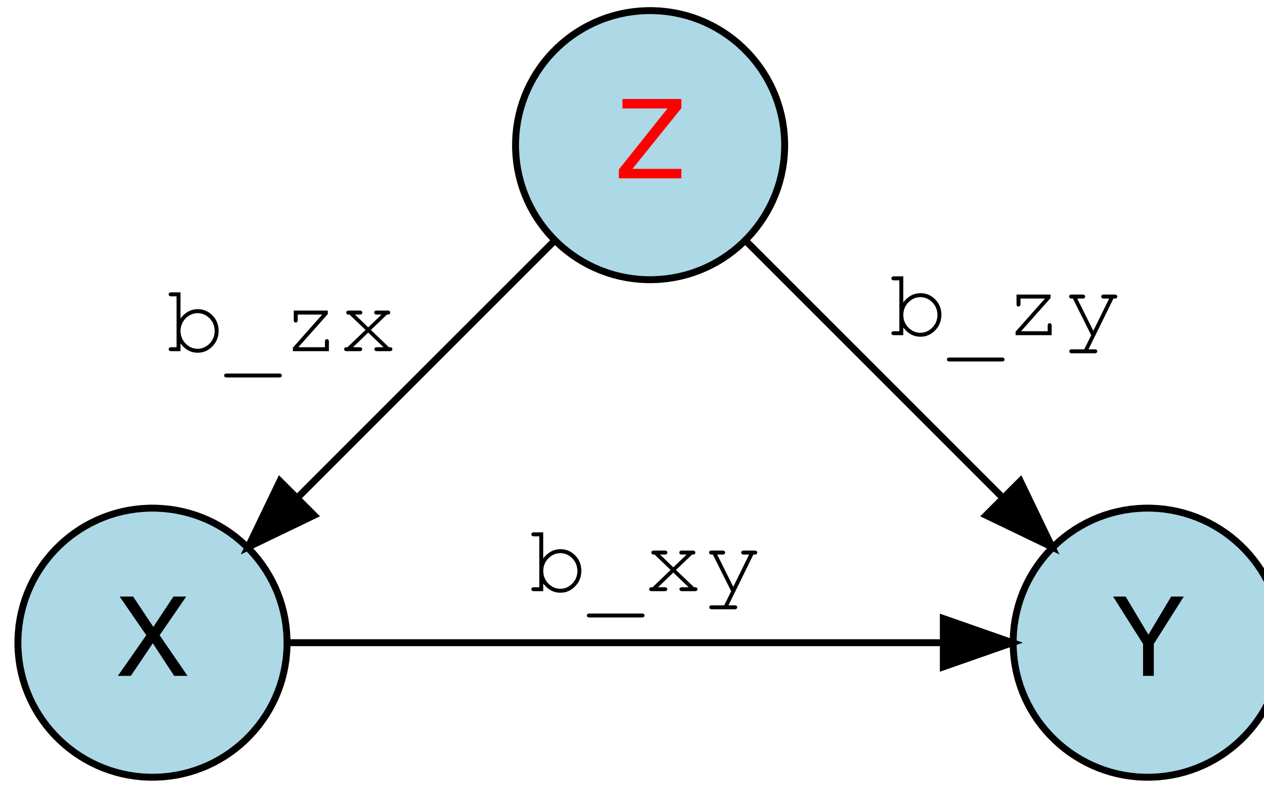
I want to estimate the effect that X has on Y



Is Z a **good** or a **bad** control here?

When should I control for variables?

```r
set.seed(1)

n = 1000
b_zx = 2
b_xy = 2
b_zy = 2
sd = 1

fun_error = function(n, sd){
  rnorm(n = n,
        mean = 0,
        sd = sd)
}

df = tibble(z = fun_error(n, sd),
            x = b_zx * z + fun_error(n, sd),
            y = b_zy * z + b_xy * x + fun_error(n, sd))
```



$$Y = b_0 + b_1 \cdot X + e$$

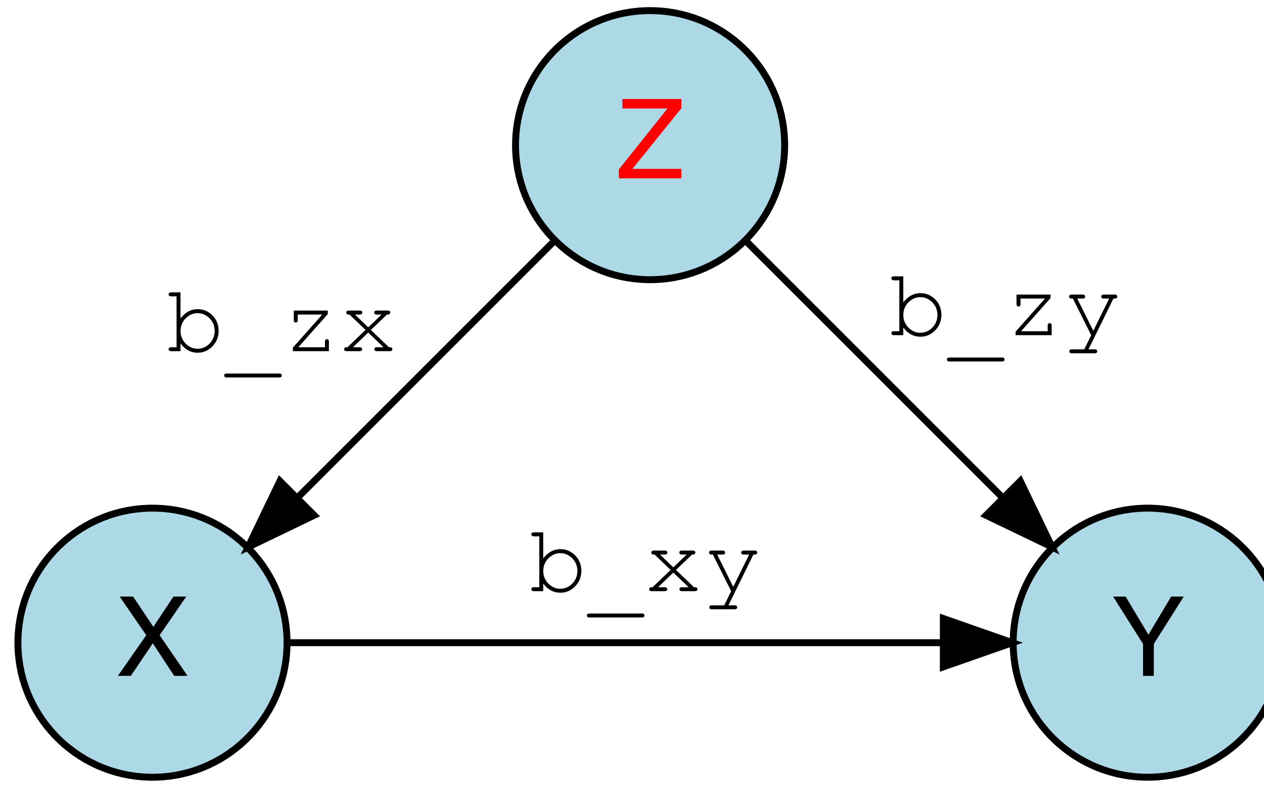
```r
# without control
lm(formula = y ~ x,
   data = df) %>%
  summary()
```

overestimating X's effect on Y

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6011 -0.9270 -0.0506  0.9711  4.0454 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.02449    0.04389   0.558    0.577    
x            2.82092    0.01890 149.225   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.388 on 998 degrees of freedom
Multiple R-squared:  0.9571,	Adjusted R-squared:  0.9571 
F-statistic: 2.227e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

When should I control for variables?

```r
set.seed(1)

n = 1000
b_zx = 2
b_xy = 2
b_zy = 2
sd = 1

fun_error = function(n, sd){
  rnorm(n = n,
        mean = 0,
        sd = sd)
}

df = tibble(z = fun_error(n, sd),
            x = b_zx * z + fun_error(n, sd),
            y = b_zy * z + b_xy * x + fun_error(n, sd))
```



$$Y = b_0 + b_1 \cdot X + b_2 \cdot Z + e$$

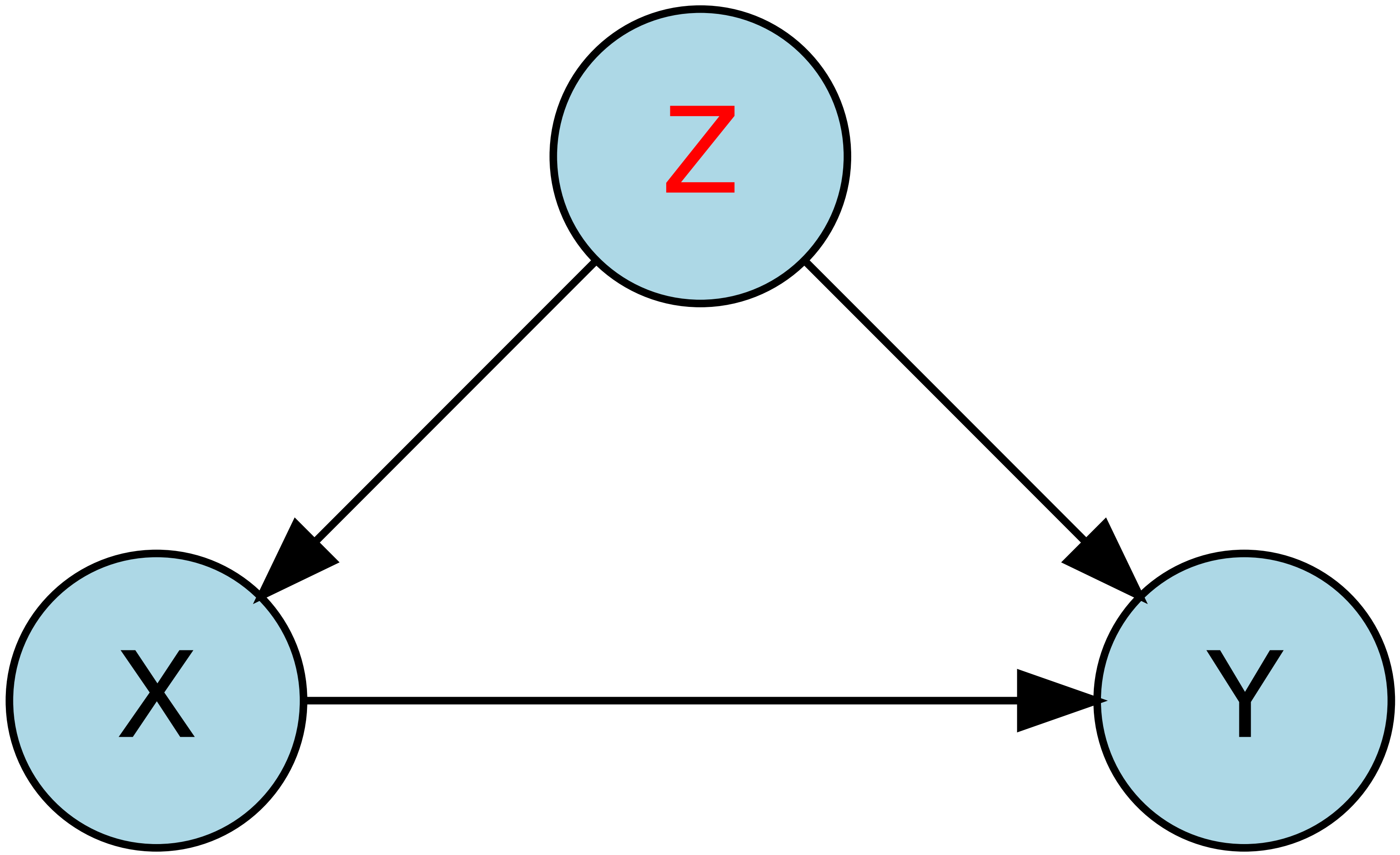
```r
# with control
lm(formula = y ~ x + z,
   data = df) %>%
  summary()
```

accurate estimate of X's effect on Y

```
Call:
lm(formula = y ~ x + z, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6151 -0.6564 -0.0223  0.6815  2.8132

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01624    0.03260   0.498    0.618
x            2.02202    0.03135  64.489   <2e-16 ***
z            2.00501    0.07036  28.497   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.031 on 997 degrees of freedom
Multiple R-squared:  0.9764,	Adjusted R-squared:  0.9763
F-statistic: 2.059e+04 on 2 and 997 DF,  p-value: < 2.2e-16
```

When should I control for variables?

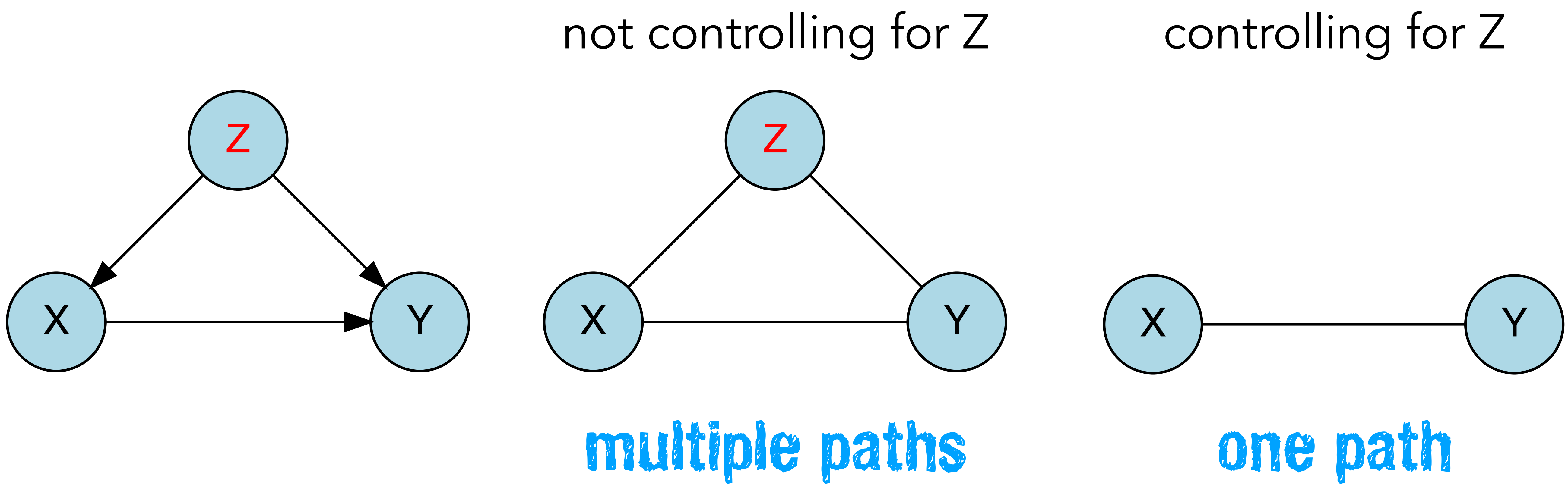
I want to estimate the effect that X has on Y



Z is a **good** control here!

When should I control for variables?

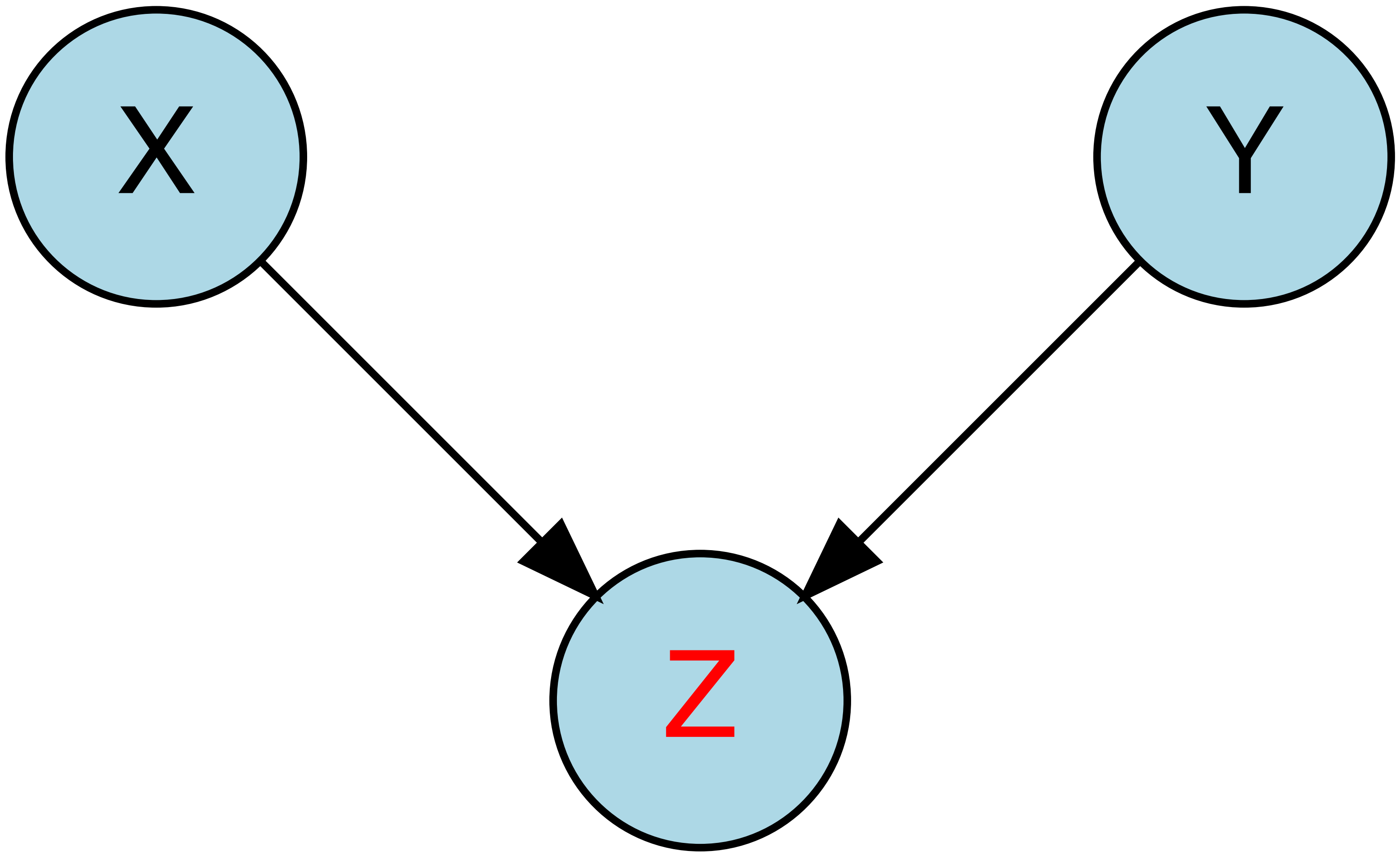
I want to estimate the effect that X has on Y



Z is a **good** control here!

When should I control for variables?

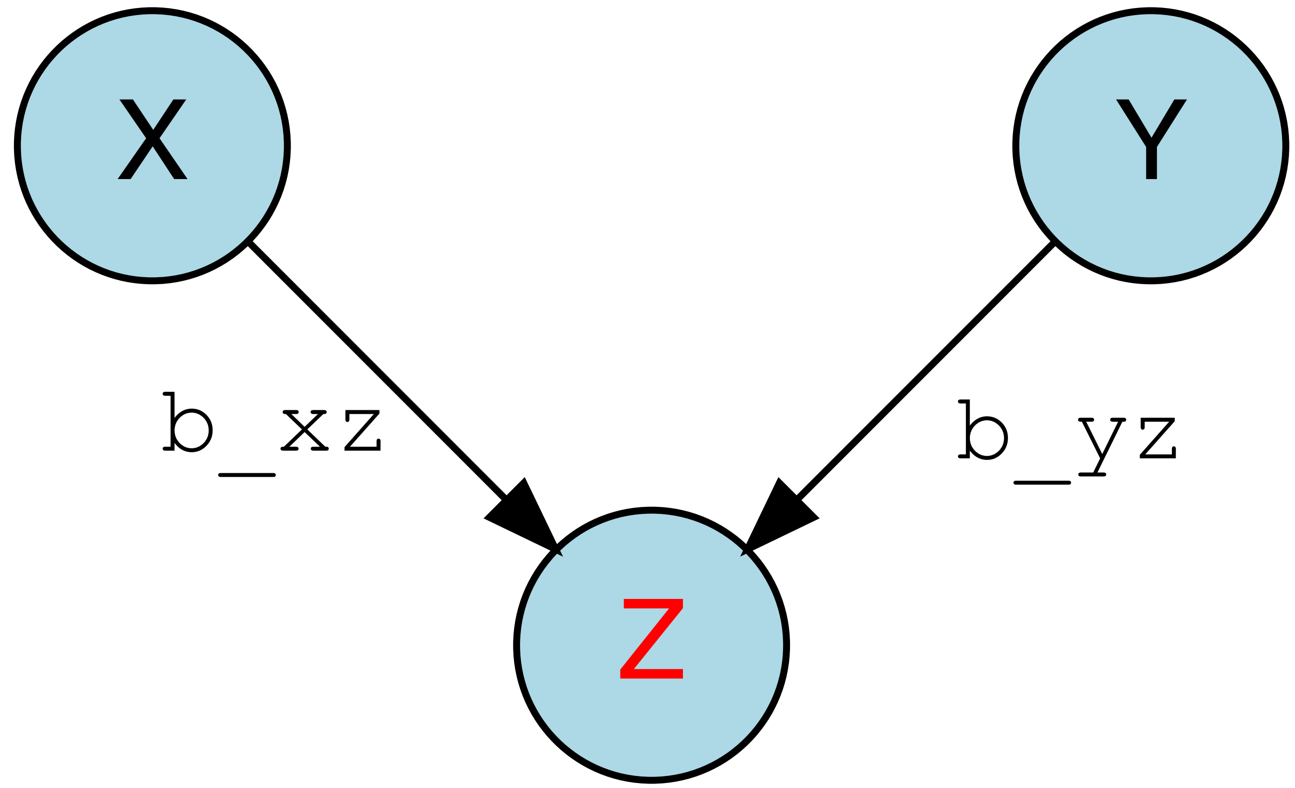
I want to estimate the effect that X has on Y



Is Z a **good** or a **bad** control here?

When should I control for variables?

```r
set.seed(1)
n = 1000
b_xz = 2
b_yz = 2
sd = 1

fun_error = function(n, sd){
  rnorm(n = n,
        mean = 0,
        sd = sd)
}

df = tibble(x = fun_error(n, sd),
            y = fun_error(n, sd),
            z = x * b_xz + y * b_yz + fun_error(n, sd))
```



accurate estimate of X's effect on Y

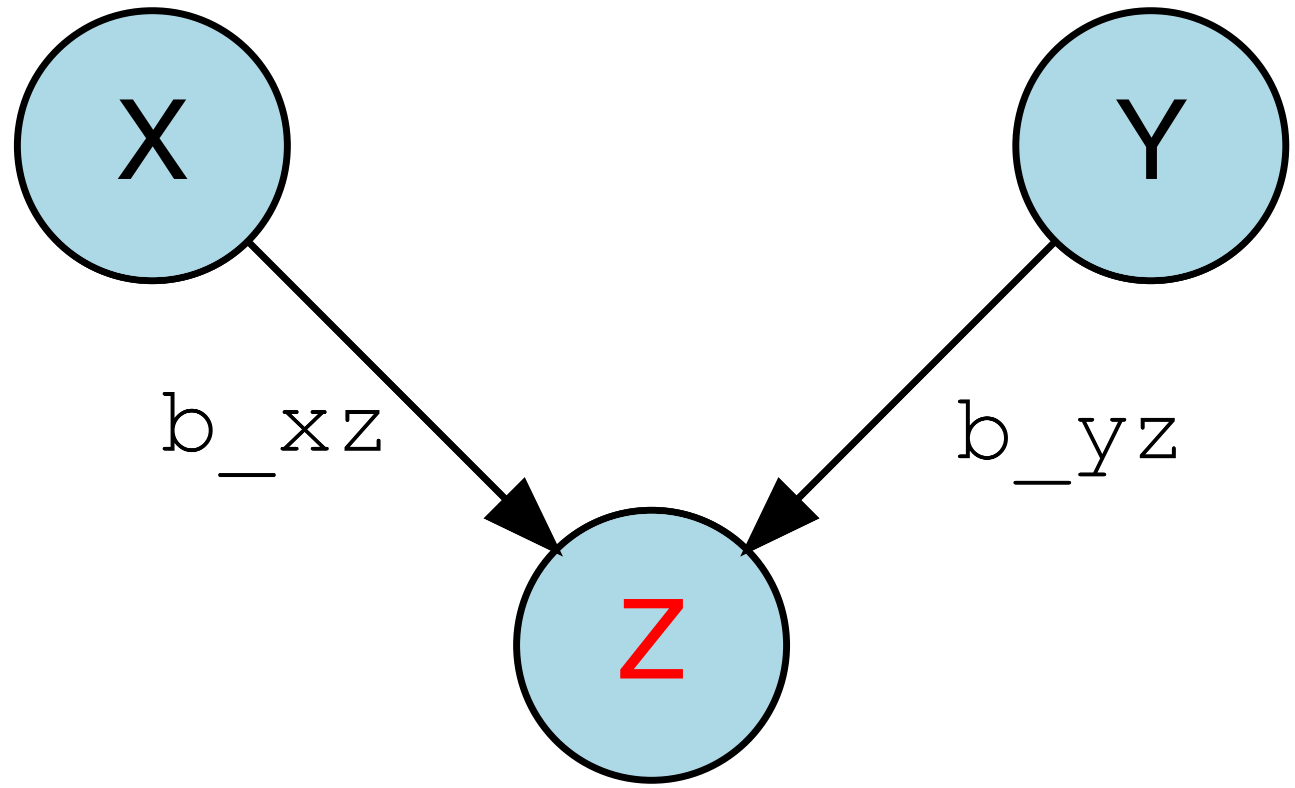
$$Y = b_0 + b_1 \cdot X + e$$

```r
# without control
lm(formula = y ~ x,
   data = df) %>%
  summary()
```

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2484 -0.6720 -0.0138  0.7554  3.6443

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.016187   0.032905  -0.492    0.623
x            0.006433   0.031809   0.202    0.840

Residual standard error: 1.04 on 998 degrees of freedom
Multiple R-squared:  4.098e-05, Adjusted R-squared:  -0.000961
F-statistic: 0.0409 on 1 and 998 DF,  p-value: 0.8398
```

When should I control for variables?

```r
set.seed(1)
n = 1000
b_xz = 2
b_yz = 2
sd = 1

fun_error = function(n, sd){
  rnorm(n = n,
        mean = 0,
        sd = sd)
}

df = tibble(x = fun_error(n, sd),
            y = fun_error(n, sd),
            z = x * b_xz + y * b_yz + fun_error(n, sd))
```



inaccurate estimate of X's effect on Y

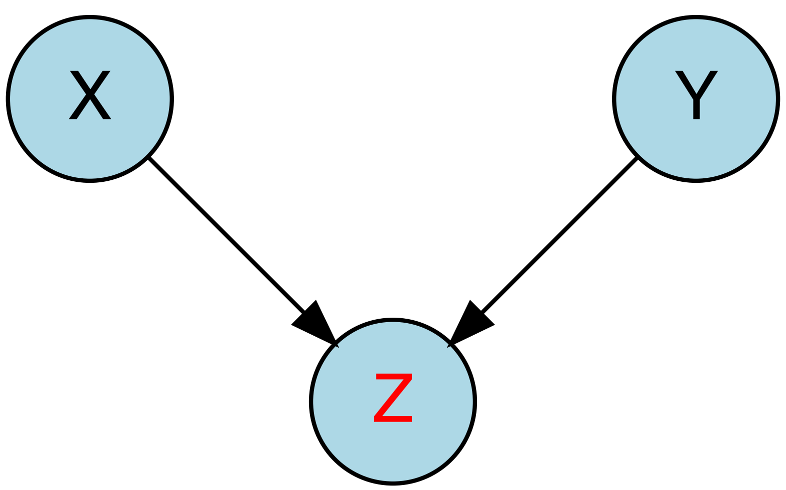
$$Y = b_0 + b_1 \cdot X + b_2 \cdot Z + e$$

```r
# with control
lm(formula = y ~ x + z,
   data = df) %>%
  summary()
```

```
Call:
lm(formula = y ~ x + z, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.35547 -0.30016  0.00298  0.31119  1.73408 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.009608   0.014477  -0.664    0.507    
x           -0.816164   0.018936 -43.102   <2e-16 ***
z            0.398921   0.006186  64.489   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4578 on 997 degrees of freedom
Multiple R-squared:  0.8066,	Adjusted R-squared:  0.8062 
F-statistic:  2079 on 2 and 997 DF,  p-value: < 2.2e-16
```

When should I control for variables?

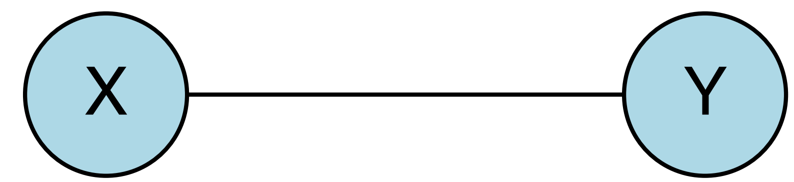
I want to estimate the effect that X has on Y



not controlling for Z

no path

controlling for Z



one path

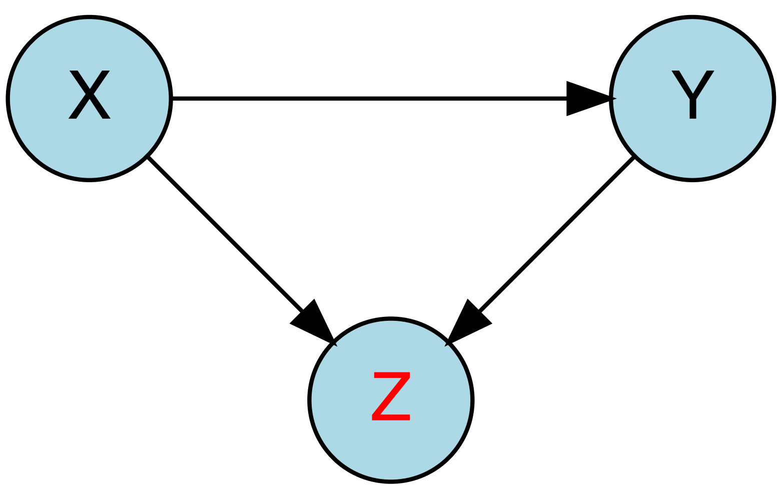
Z is a **bad** control here!

When should I control for variables?

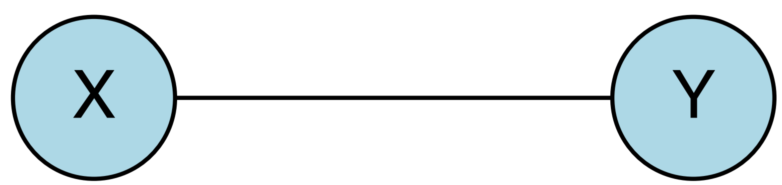
- checking for **d-separation** tells us whether or not variables are (conditionally) independent
- it also tells us whether paths of dependence "open up", or get "closed down"
- the graphical procedure doesn't necessarily reveal whether the dependence between variables changes: it reveals the **structure** of dependence but not the **strength**
- you can always double check via running simulations in R

When should I control for variables?

I want to estimate the effect that X has on Y

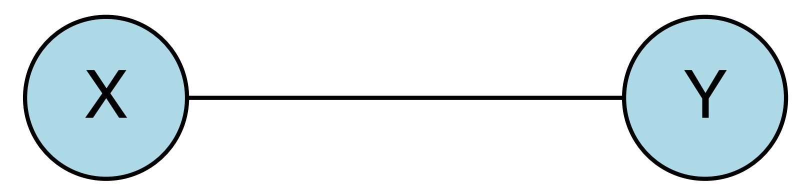


not controlling for Z



one path

controlling for Z



change in dependence

not revealed in the undirected graph

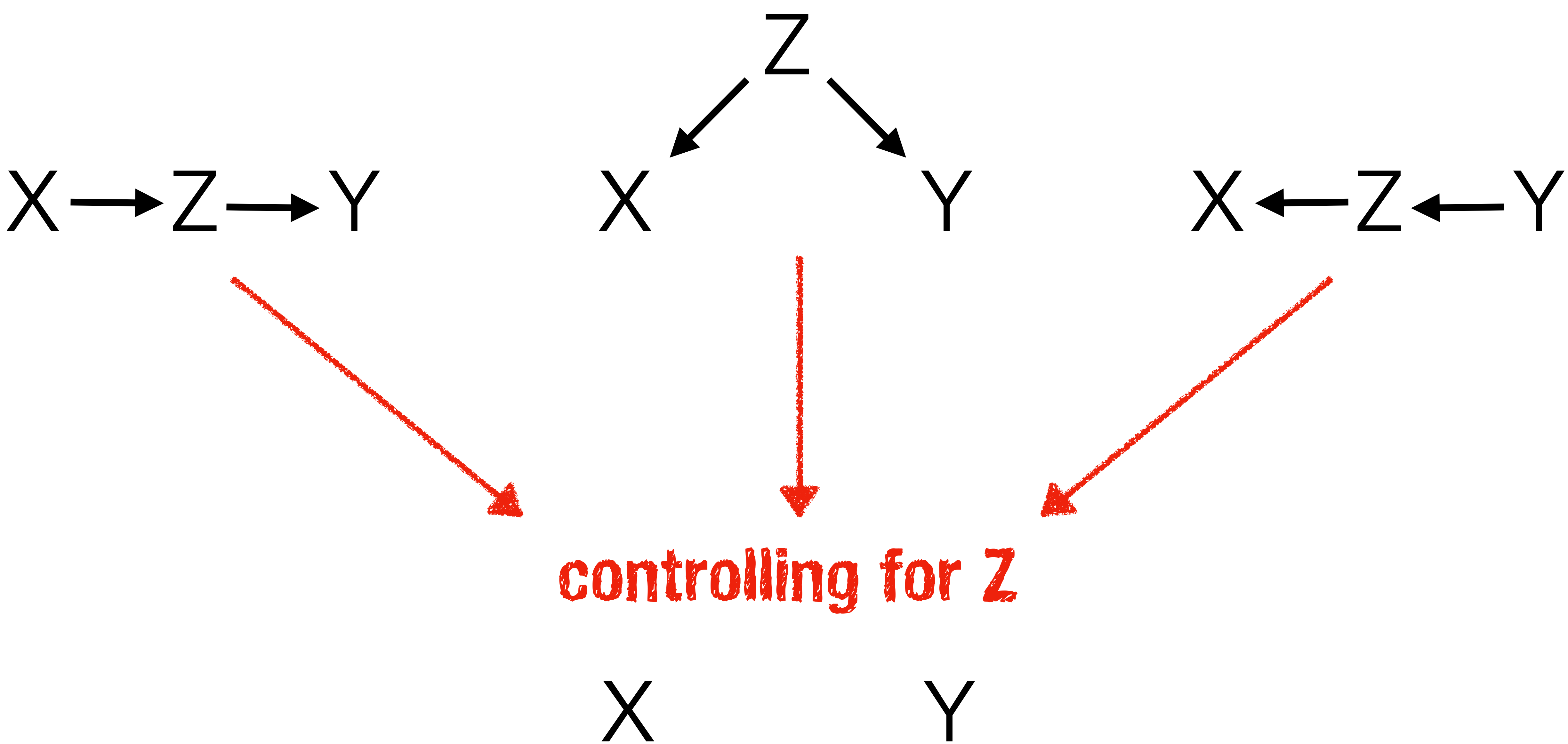
Z is a **bad** control here!

When should I control for variables?

- **good controls** reduce additional paths from `X` to `Y` apart from the direct path we are interested in estimating
- **bad controls** introduce additional paths (or change existing ones) that lead to a biased estimate of the direct path between `X` and `Y`

When should I control for variables?

Problem: We don't know the ground truth ...



we need to manipulate X experimentally to tell these apart*

* sort of (see next slide)

When should I control for variables?

- **causal discovery** is a very active field

what causal claims can we make from observational data?

Identifiability of Gaussian structural equation models with equal error variances

Jonas Peters*
Seminar for Statistics
ETH Zurich
Switzerland

Peter Bühlmann*
Seminar for Statistics
ETH Zurich
Switzerland

October 29, 2018

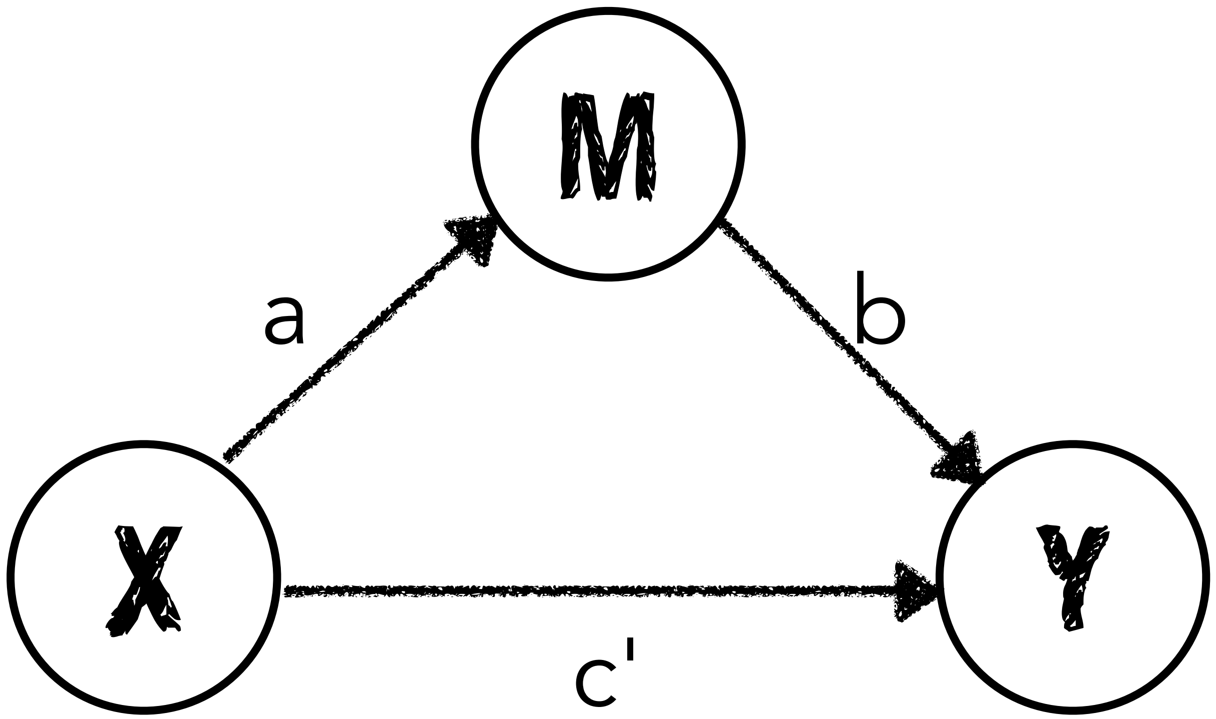
causal model is fully identifiable if all noise variables have the same variances, and all variables are observed

beyond the scope of our class ...

Peters & Buhlmann (2018) Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*

Mediation

Definition

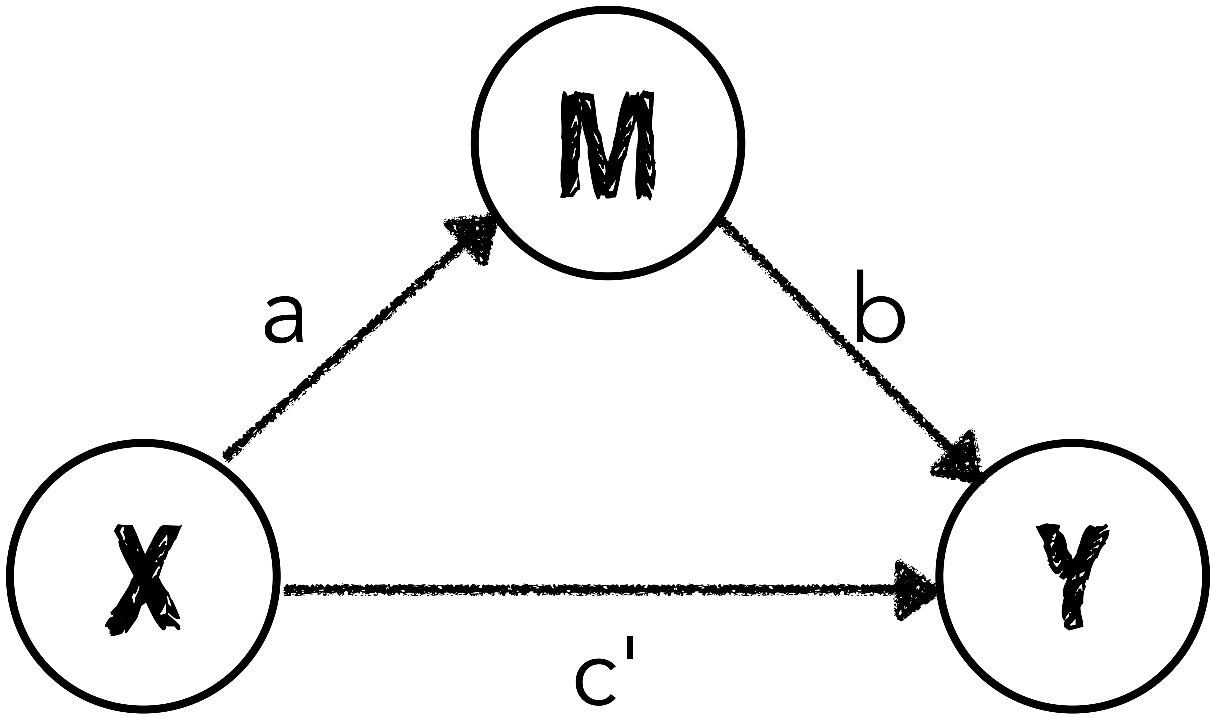


Rather than a direct causal relationship between **X** and **Y**, a mediation model proposes that **X** influences the mediator variable **M**, which in turn influences **Y**. Thus, the mediator variable serves to clarify the nature of the relationship between **X** and **Y**.

Adapted from Wikipedia

https://en.wikipedia.org/wiki/Mediation_(statistics)

Example



X = grades in Psych 252

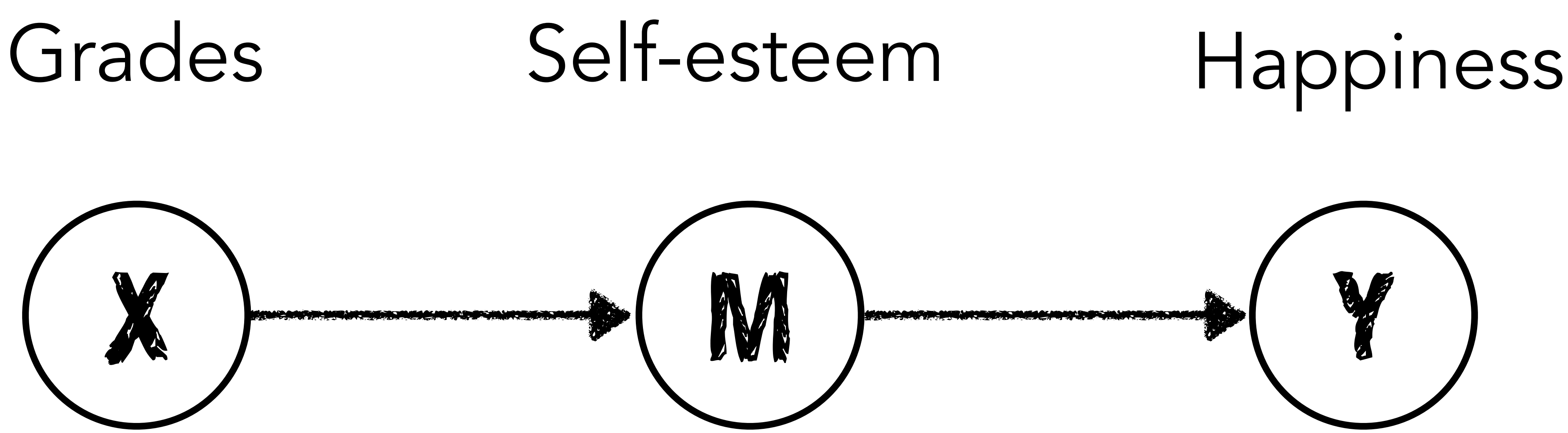
M = feelings of self-esteem

Y = happiness

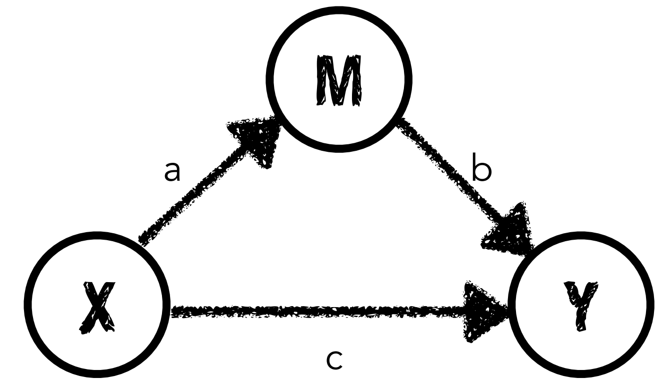
Is the relationship between grades in Psych 252 and happiness mediated by feelings of self-esteem?

Simulate a mediation analysis

```
# number of participants
n = 100

# generate data
df.mediation = tibble(
  x = rnorm(n, 75, 7),              # grades
  m = 0.7 * x + rnorm(n, 0, 5),     # self-esteem
  y = 0.4 * m + rnorm(n, 0, 5)      # happiness
)
```



Bootstrapping



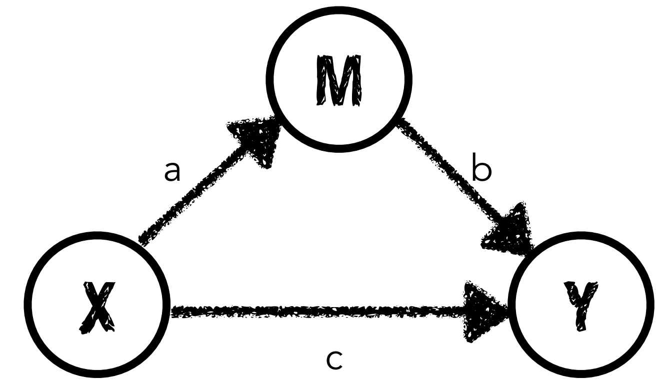

```r
library("mediation")

# bootstrapped mediation
fit.mediation = mediate(model.m = fit.m_x,
                        model.y = fit.y_mx,
                        treat = "x",
                        mediator = "m",
                        boot = T)

# summarize results
fit.mediation %>% summary()
```

$$\hat{m} = b_0 + b_1 \cdot x$$

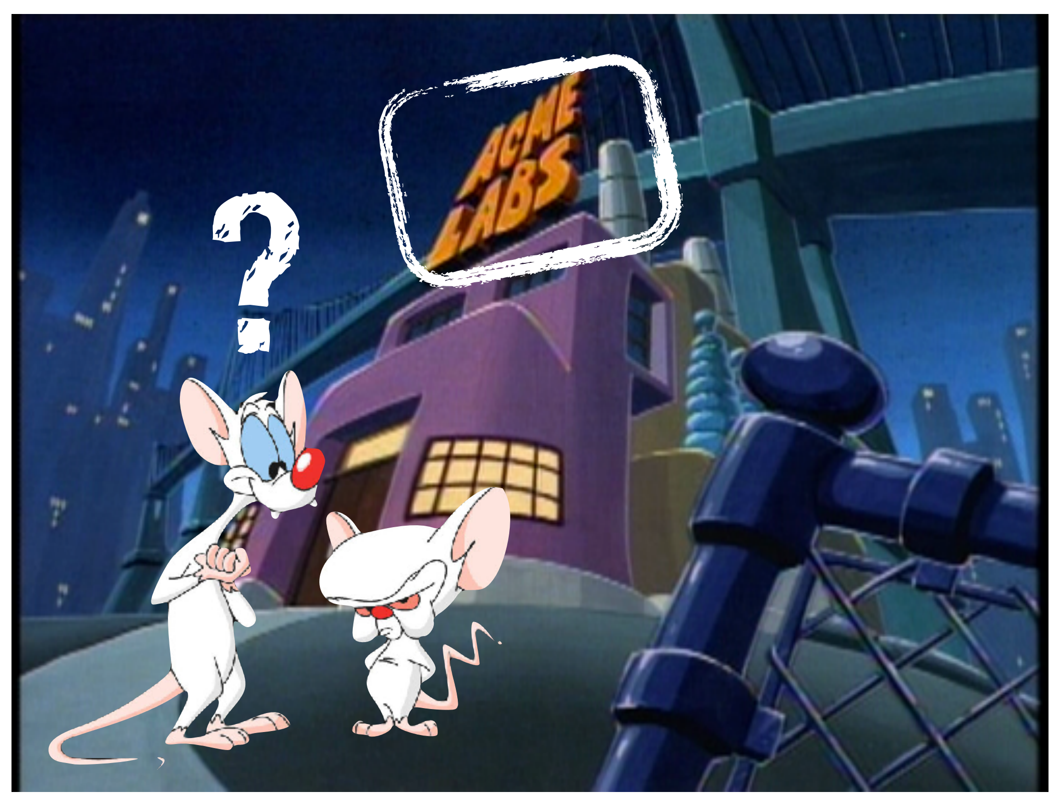
$$\hat{y} = b_0 + b_1 \cdot m + b_2 \cdot x$$

```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value    
ACME            0.28078      0.14059         0.42  <2e-16 ***
ADE            -0.11179     -0.29276         0.10   0.272    
Total Effect    0.16899     -0.00415         0.34   0.064 .  
Prop. Mediated  1.66151     -3.22476        11.46   0.064 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 100


Simulations: 1000
```

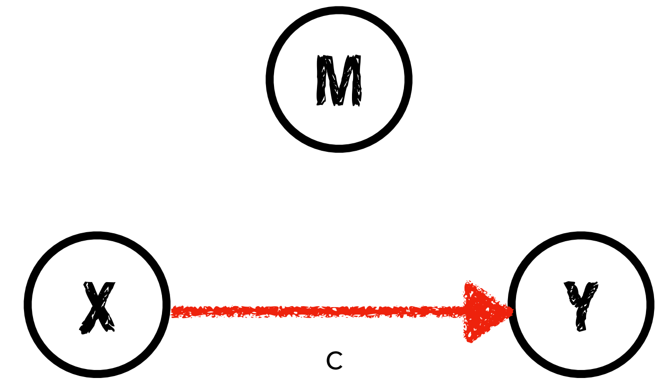
2. Bootstrapping




```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

                 Estimate 95% CI Lower 95% CI Upper p-value
ACME              0.28078      0.14059         0.42  <2e-16 ***
ADE              -0.11179     -0.29276         0.10   0.272
Total Effect      0.16899     -0.00415         0.34   0.064 .
Prop. Mediated    1.66151     -3.22476        11.46   0.064 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 100


Simulations: 1000
```



2. Bootstrapping

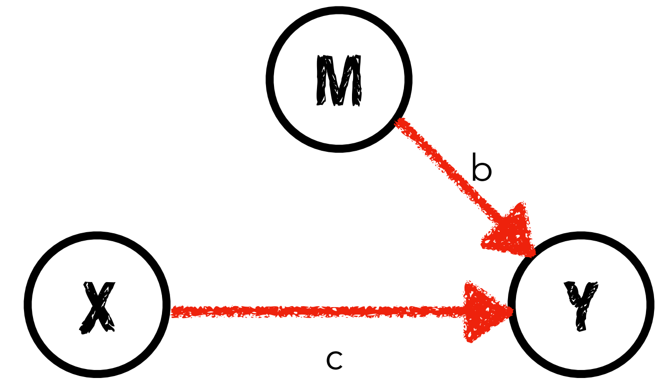



```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value
ACME            0.28078      0.14059         0.42  <2e-16 ***
ADE            -0.11179     -0.29276         0.10   0.272
Total Effect    0.16899     -0.00415         0.34   0.064 .
Prop. Mediated  1.66151     -3.22476        11.46   0.064 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 100


Simulations: 1000
```

$$\hat{y} = b_0 + b_1 \cdot x$$

```
Call:
lm(formula = y ~ 1 + x, data = df.mediation)

Residuals:
    Min      1Q  Median      3Q     Max
-10.917  -3.738  -0.259   2.910  12.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.78300    6.16002   1.426   0.1571
x            0.16899    0.08116   2.082   0.0399 *
```

2. Bootstrapping




```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value    
ACME            0.28078      0.14059         0.42  <2e-16 ***
ADE            -0.11179     -0.29276         0.10   0.272    
Total Effect    0.16899     -0.00415         0.34   0.064 .  
Prop. Mediated  1.66151     -3.22476        11.46   0.064 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 100 


Simulations: 1000 
```

$$\hat{y} = b_0 + b_1 \cdot m + b_2 \cdot x$$ **ADE:** Average direct effect

```
Call:
lm(formula = y ~ 1 + m + x, data = df.mediation)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.3651 -3.3037 -0.6222  3.1068 10.3991 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.80952    5.68297   1.374    0.173    
m            0.42381    0.09899   4.281 4.37e-05 ***
x           -0.11179    0.09949  -1.124    0.264    
```

2. Bootstrapping

```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value
ACME            0.28078      0.14059         0.42  <2e-16 ***
ADE            -0.11179     -0.29276         0.10   0.272
Total Effect    0.16899     -0.00415         0.34   0.064 .
Prop. Mediated  1.66151     -3.22476        11.46   0.064 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 100


Simulations: 1000
```

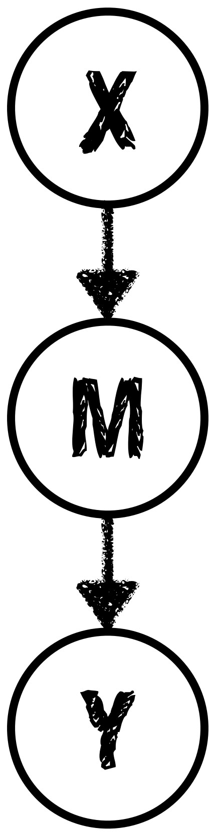
ACME: Average causal mediation effect

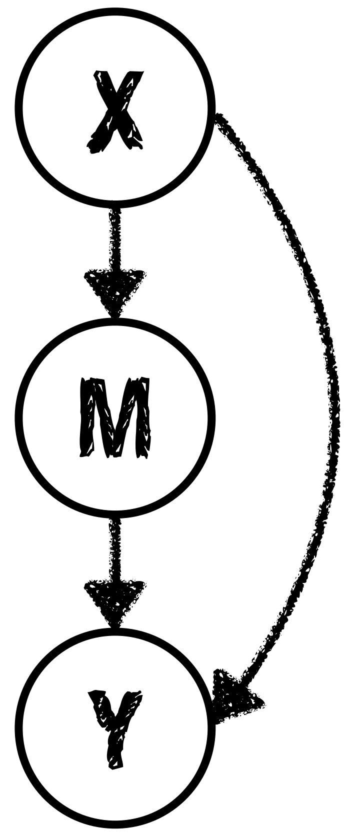
ACME = Total effect - ADE

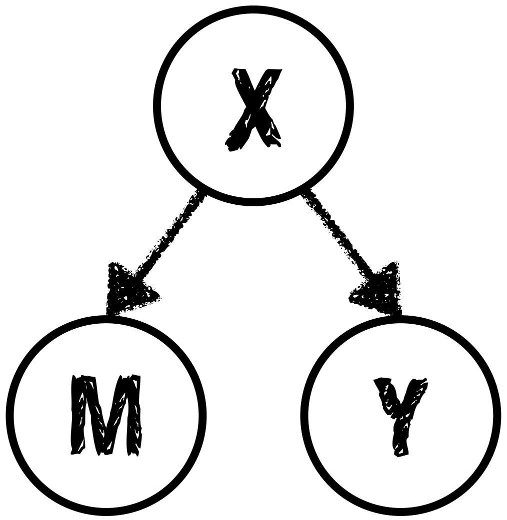
ADE: Average direct effect

Underlying causal model

Full mediation **Partial mediation** **No mediation**





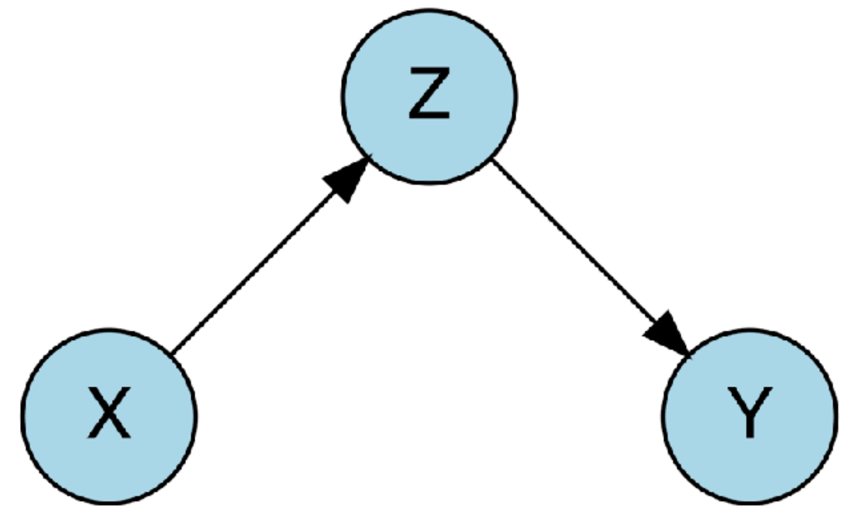


Full mediation: When the effect of **X** on **Y** completely disappears, **M** fully mediates between **X** and **Y**.

Partial mediation: When the effect of **X** on **Y** still exists, but in a smaller magnitude, **M** partially mediates between **X** and **Y**.

Beware of mediation analyses!!!

```r
set.seed(1)

n = 100 # number of observations

# causal chain
df.causal_chain = tibble(x = rnorm(n, 0, 1),
                         z = 2 * x + rnorm(n, 0, 1),
                         y = 2 * z + rnorm(n, 0, 1))
```

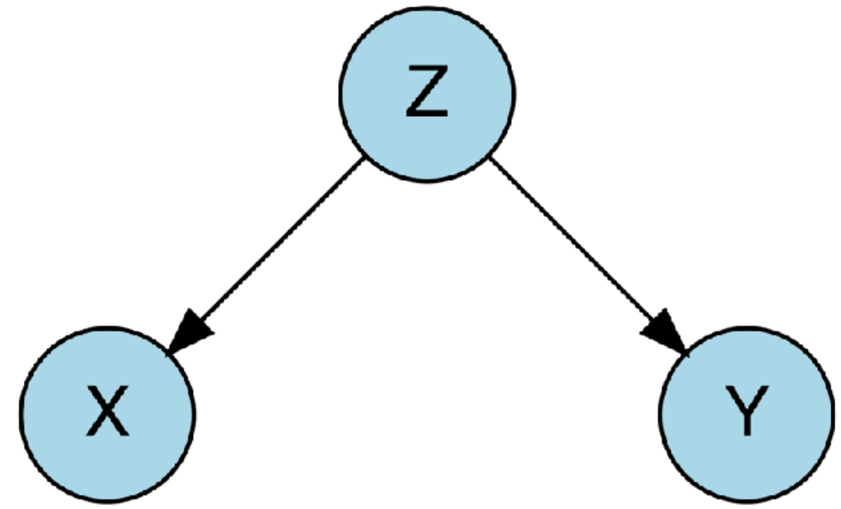



```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value    
ACME             0.8287       0.6234         1.05  <2e-16 ***
ADE             -0.0535      -0.2548         0.15    0.55    
Total Effect     0.7752       0.6391         0.90  <2e-16 ***
Prop. Mediated   1.0690       0.8131         1.35  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 100


Simulations: 1000
```

nice mediation result!

Beware of mediation analyses!!!

```r
set.seed(1)

n = 100 # number of observations

# common cause
df.common_cause = tibble(z = rnorm(n, 0, 1),
                         x = 2 * z + rnorm(n, 0, 1),
                         y = 2 * z + rnorm(n, 0, 1))
```




```
Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value    
ACME             0.8287       0.6065         1.04  <2e-16 ***
ADE             -0.0535      -0.2675         0.16    0.56    
Total Effect     0.7752       0.6353         0.90  <2e-16 ***
Prop. Mediated   1.0690       0.8134         1.37  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 100


Simulations: 1000
```

(not) nice mediation result!

Limitations

- correlational analysis
  - we need theories / experiments to tease apart causes and effects to properly map our variables onto the diagram

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Reports

What mediation analysis can (not) do☆

Klaus Fiedler [a,*], Malte Schott [a], Thorsten Meiser [b]

[a] University of Heidelberg, Germany
[b] University of Mannheim, Germany

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47(6), 1231-1236.

Limitations

many-to-one mapping

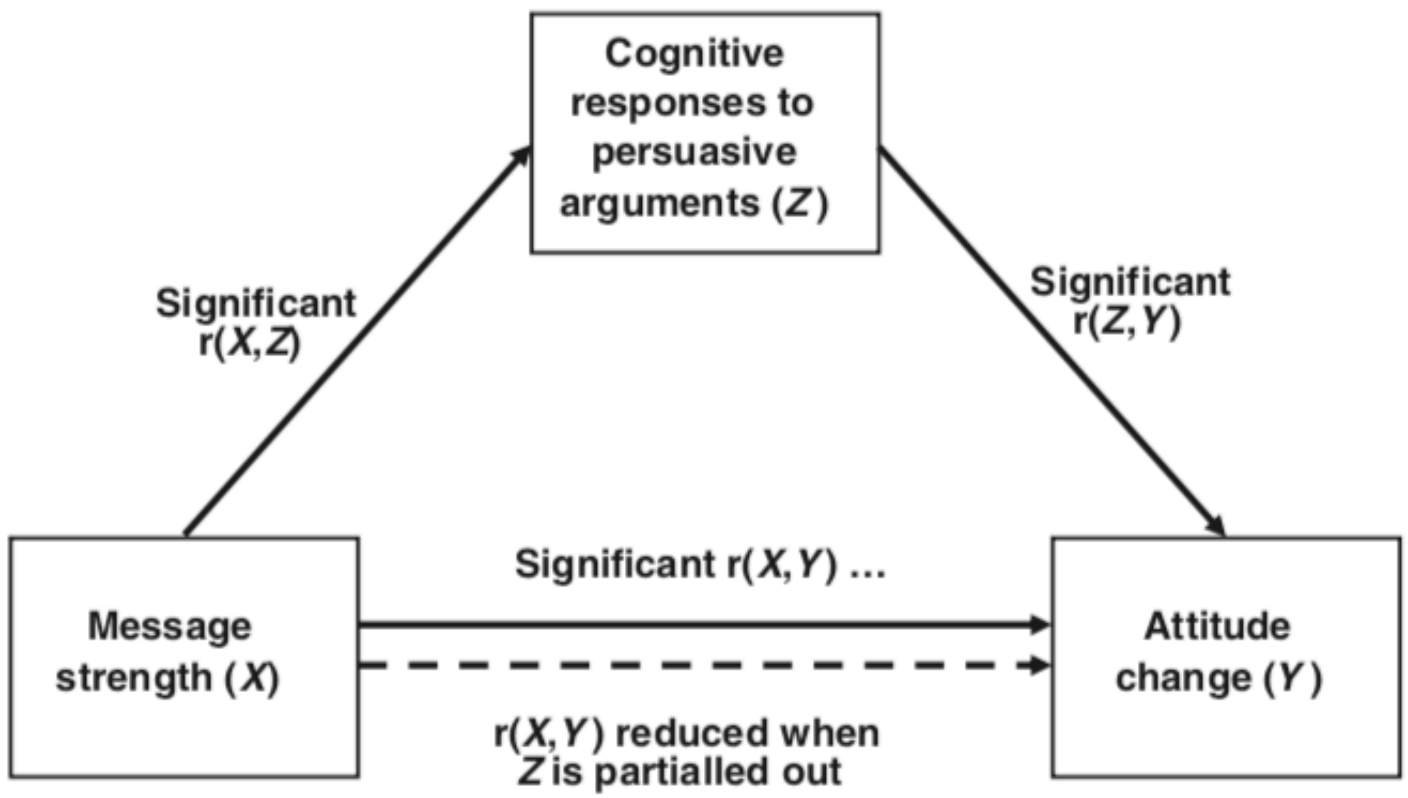


Fig. 1. Illustration of a mediation analysis in the context of persuasion research: The impact of message strength (*X*) on attitude change (*Y*) is supposed to be mediated by the recipients' cognitive responses (*Z*).

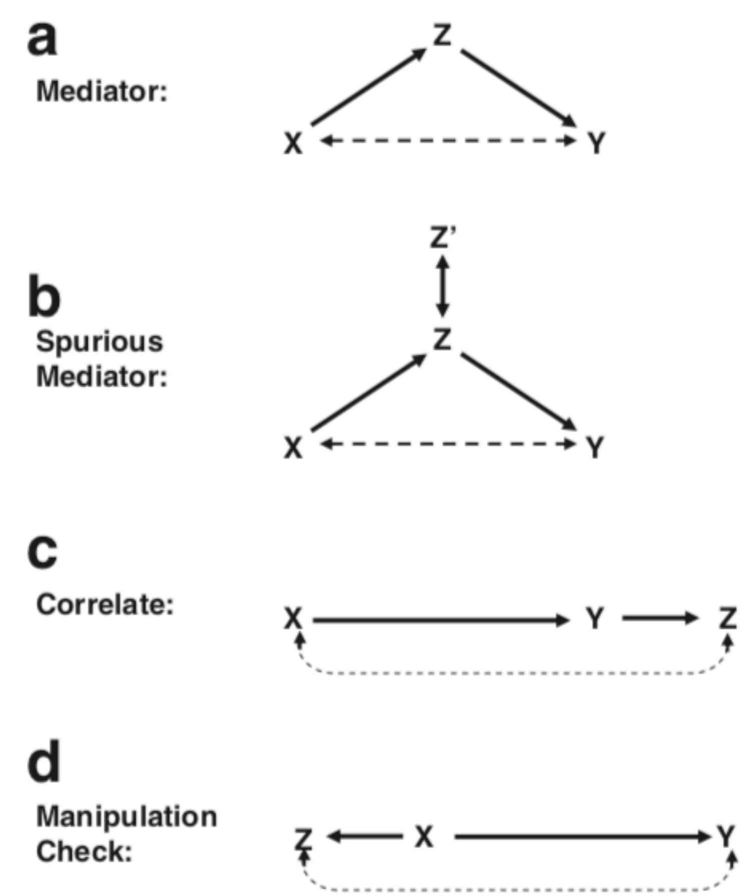


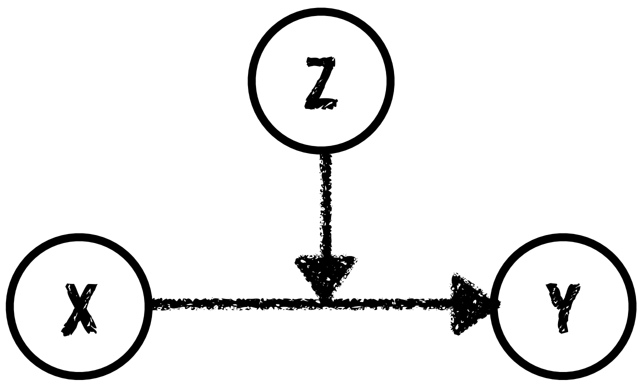
Fig. 2. Different causal models for the interpretation of the same intercorrelations between three variables, *X*, *Y*, and *Z*.

only experiments allow us to tell apart possible causal structures

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47(6), 1231-1236.

Moderation

Definition

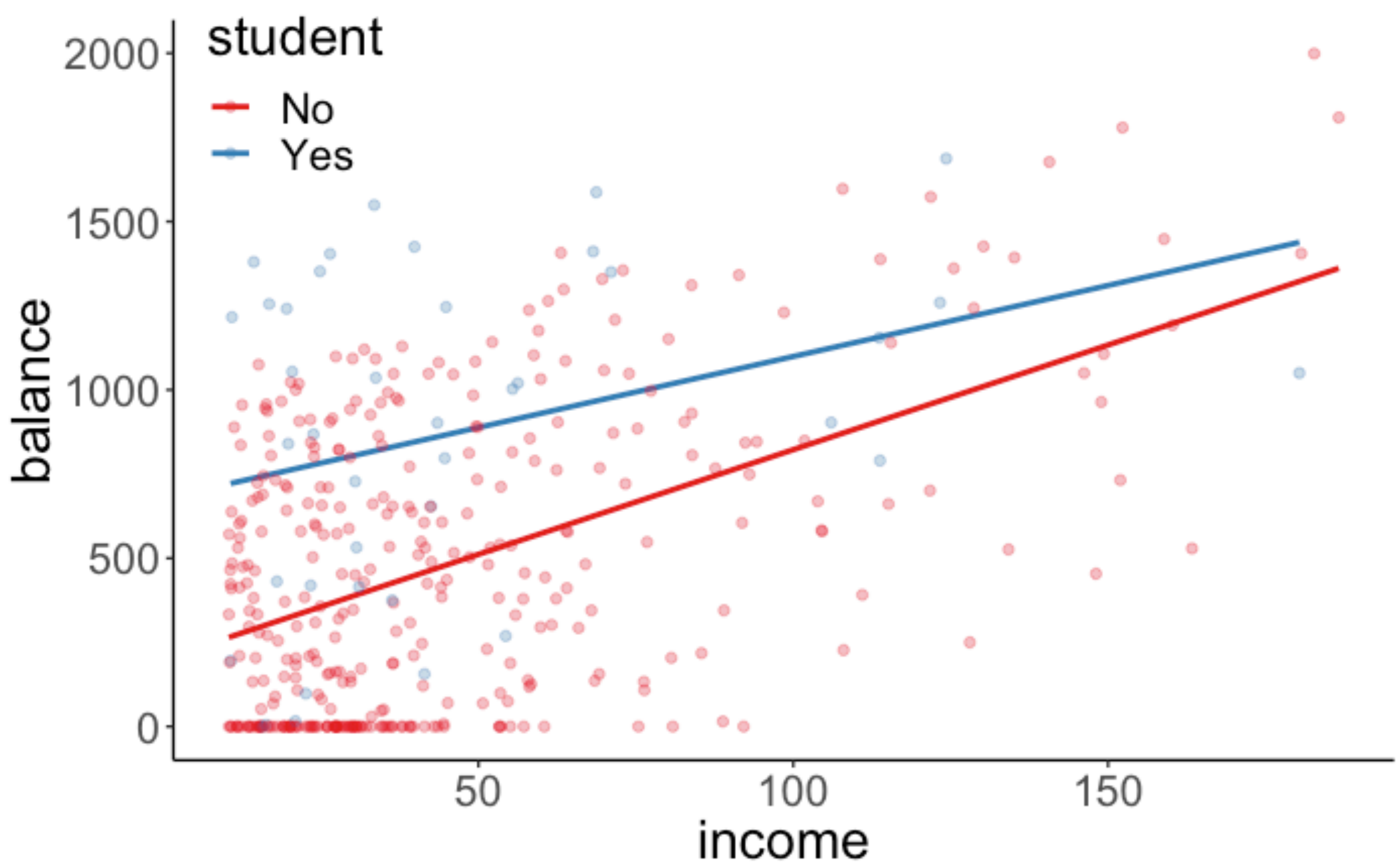


Moderation means that the effect of a predictor depends on the value of another.

Here, the nature of the relationship between **X** and **Y** depends on **Z**.

Have we come across moderation already?

Relationship between credit card balance, income, and whether the person is a student.



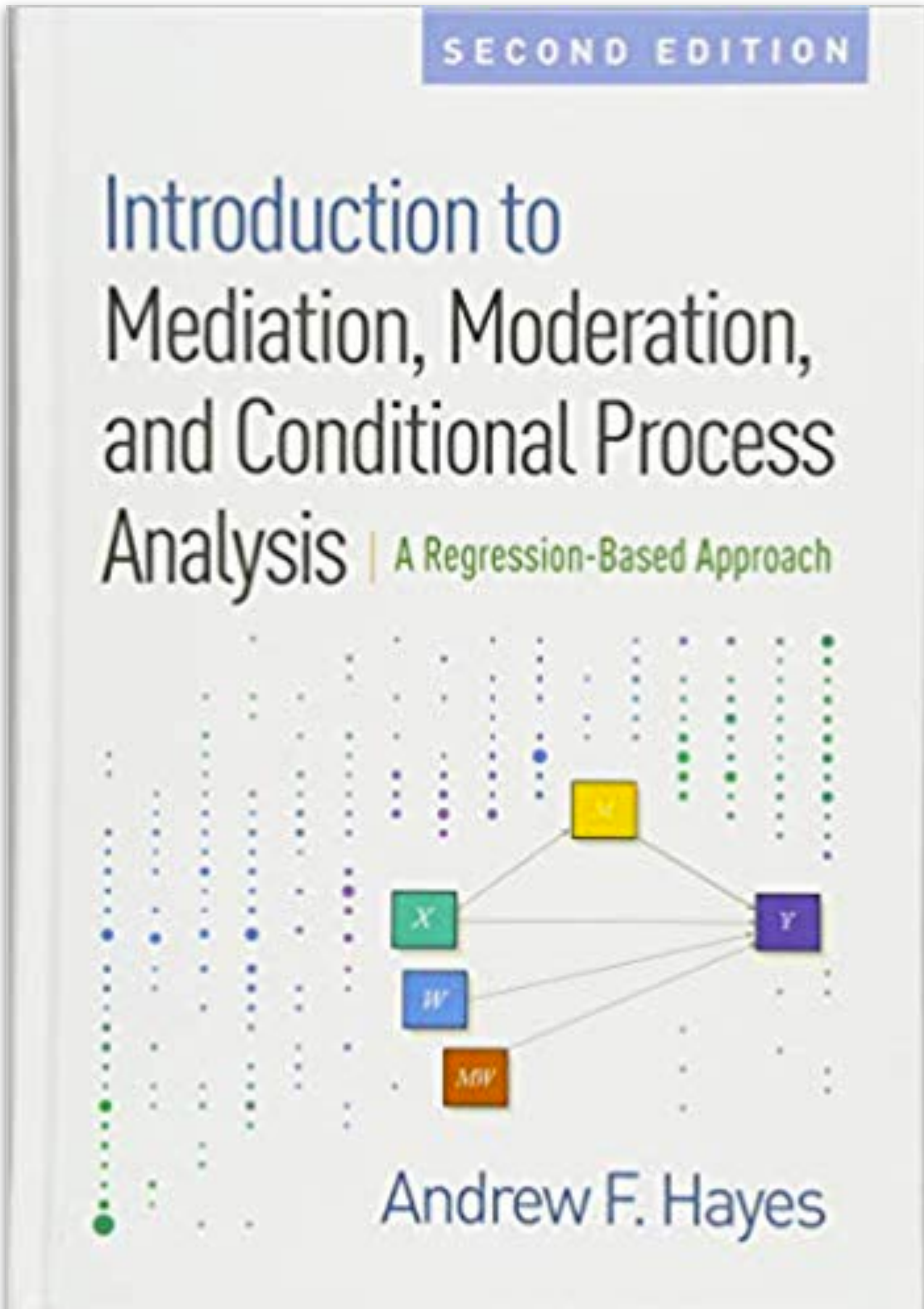
$$\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot \text{student}_i - 2.00 \cdot (\text{income}_i \times \text{student}_i)$$

if student = "No" $\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i$

if student = "Yes"

$$\begin{aligned}\widehat{\text{balance}}_i &= 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot 1 - 2.00 \cdot (\text{income}_i \times 1) \\ &= 677.3 + 6.22 \cdot \text{income}_i - 2.00 \cdot \text{income}_i \\ &= 677.3 + 4.22 \cdot \text{income}_i\end{aligned}$$

Learn more about mediation and moderation



Recoded with `brms` by Solomon Kurz here: https://bookdown.org/connect/#/apps/1523/access

Summary

- Quick recap
- Model comparison
  - Cross-validation in action
  - AIC and BIC
- Causation vs. correlation
- Controlling for variables
- Mediation
- Moderation

Feedback

How was the pace of today's class?

much too slow	a little too slow	just right	a little too fast	much too fast

How happy were you with today's class overall?

What did you like about today's class? What could be improved next time?

Thank you!