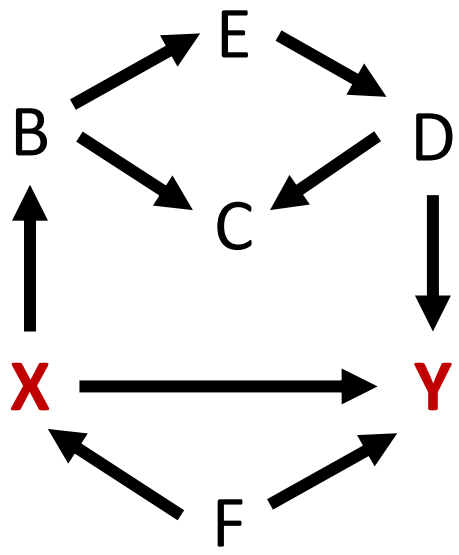
Introduction to causal inference for observational studies in Ecology

Séminaire ECODIV – Nov. 2023



David Bauman
UMR AMAP, Univ. Montpellier, CNRS, CIRAD, INRAE, **IRD**
david.bauman@ird.fr
@davbauman

Causal inference for observational data in Ecology

- Most questions asked in Ecology

  « *How does variation in X influences variation in Y?* » X → Y

- Ecology increasingly relies on broad-scale observational data (climate change, invasive species...)

- Training to think of causation often missing

- Predictive models: Model selection (AIC, machine learning...)

- Predictive models for causal inference = 

- Common stat. adjustments can bias causal estimates (overcontrol bias, confounding, collider bias)

Structural Causal Models

ECOLOGY LETTERS

Predictive models aren't for causal inference

Suchinta Arif | M. Aaron MacNeil

REVIEW

ECOLOGICAL MONOGRAPHS
ECOLOGICAL SOCIETY OF AMERICA

Applying the structural causal model framework for observational causal inference in ecology

Suchinta Arif | M. Aaron MacNeil

Front Ecol Environ 2022; doi:10.1002/fee.2530

Toward an improved understanding of causation in the ecological sciences

Ethan T Addicott[1*], Eli P Fenichel[1], Mark A Bradford[1,2], Malin L Pinsky[3], and Stephen A Wood[1,4]

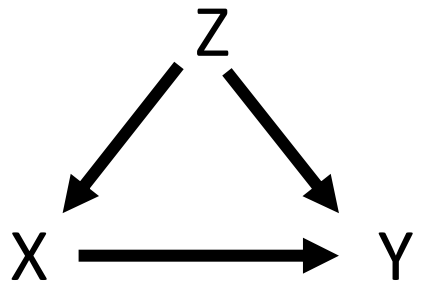
Structural Causal Models and DAGs

- **Structural Causal Models (SCMs)**: Strong features of different causal methods to create theory of causation and framework for causal inference.
  → **Cause-effect relations from observational data** without randomised controlled experiments

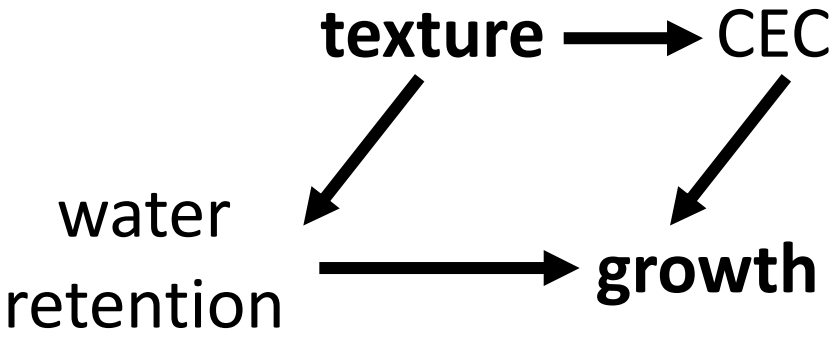
- **Causal inference**: Process → establish an association of causality b/ an element and its effects.

*« **Predicting the effect of an intervention** »*

- Directed Acyclic Graphs (**DAGs**): Graph of causal assumptions about data-generating processes

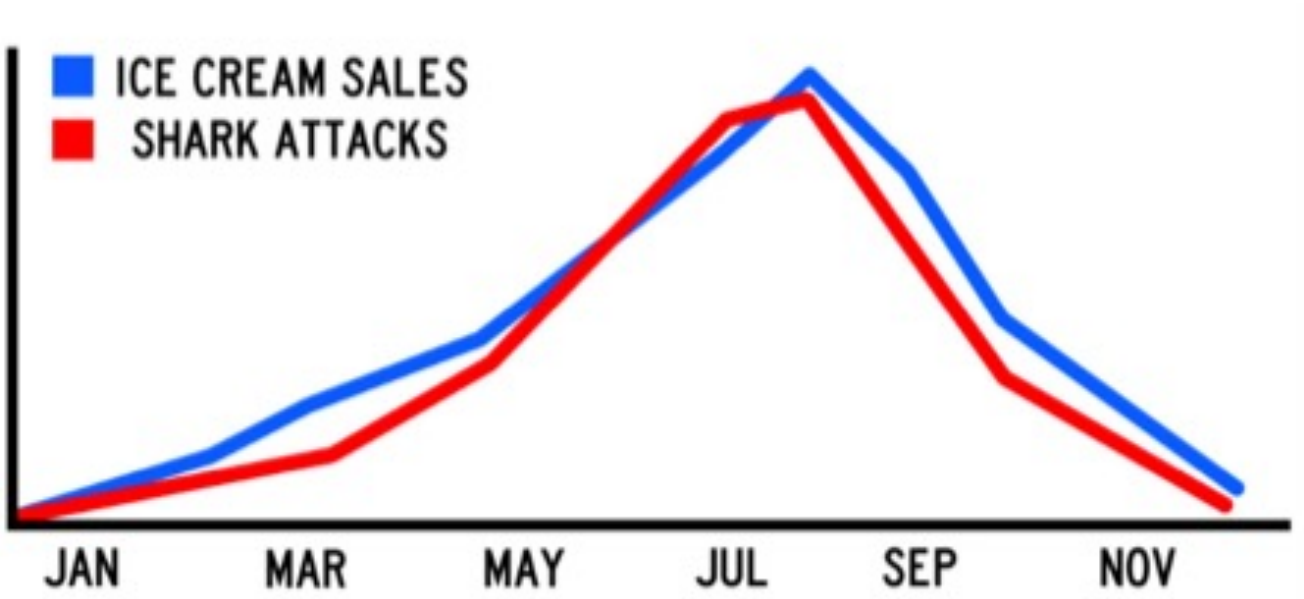


- X causes Y
- Z causes Y, directly, and through its effect on X



Association *vs* causation

Rain → Umbrella



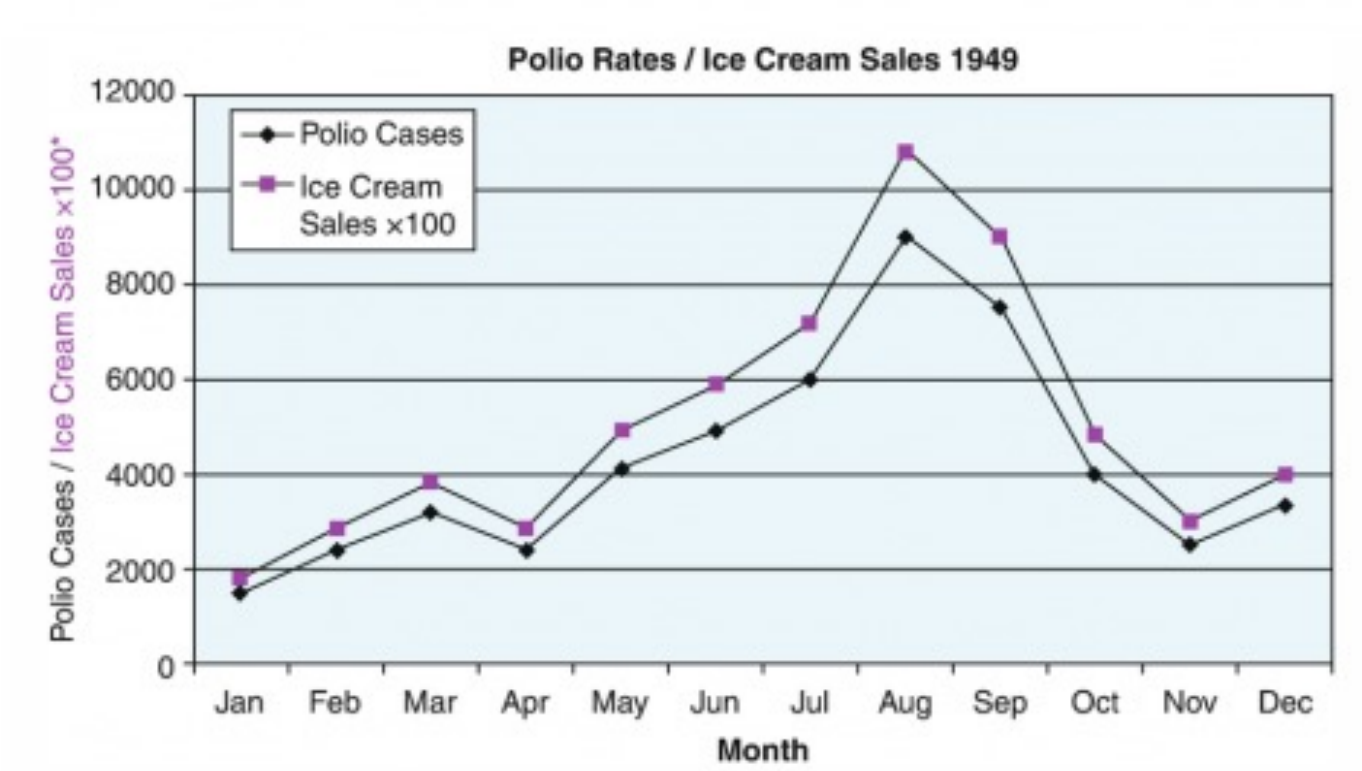
temperature → ice cream sales

temperature → # people in the see → shark attacks

ice cream sales ⇢ shark attacks

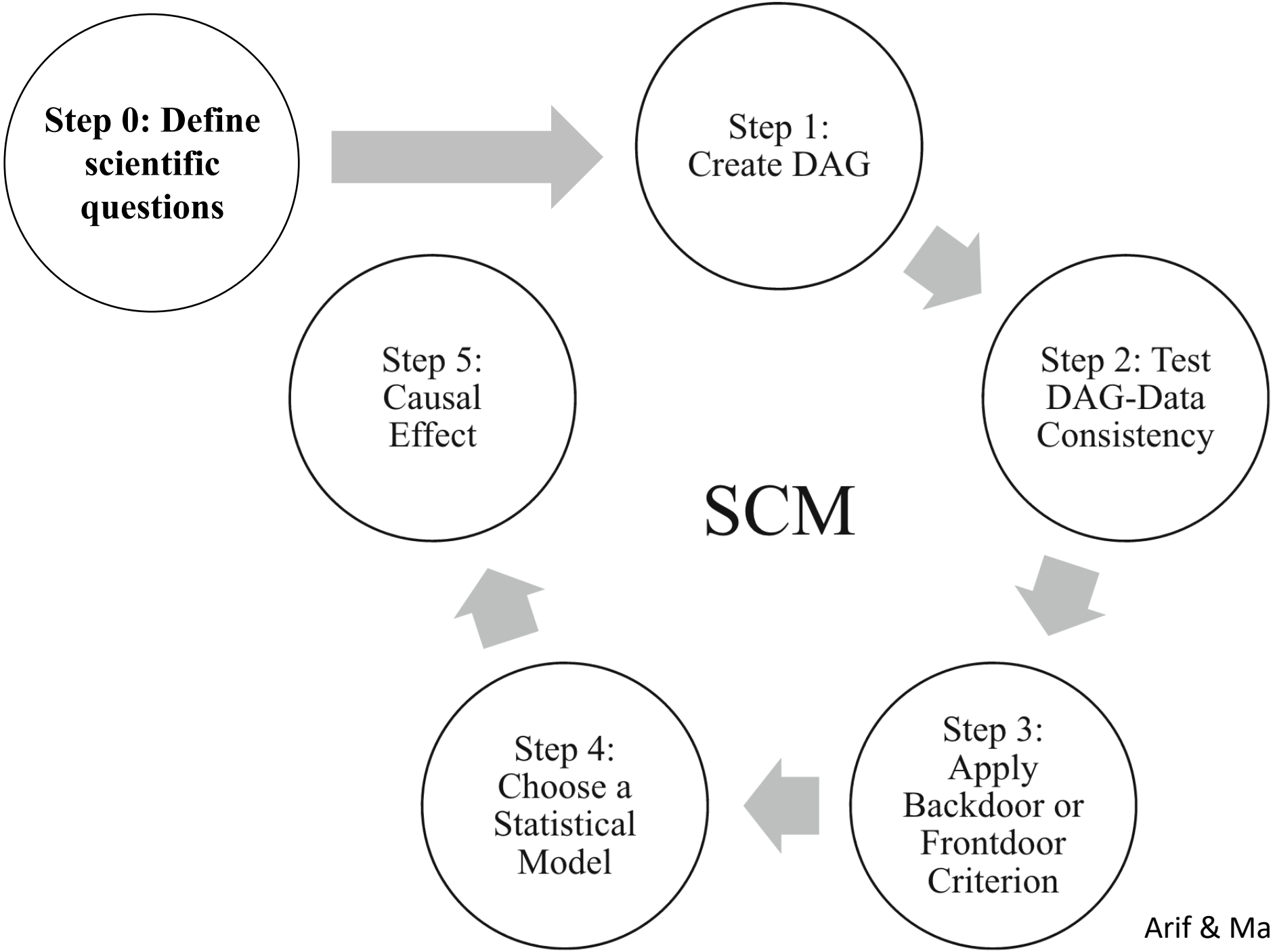
Association *vs* causation

Rain ⟶ Umbrella



The causes are not in the data

Statistical models only estimate *associations*

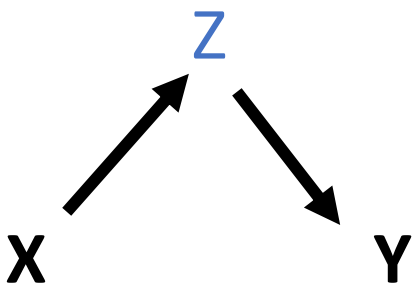


Arif & MacNeil 2022

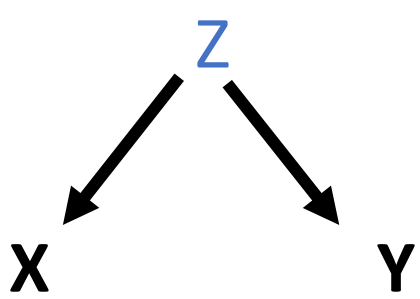
The elemental confounds

Question of interest (= *estimand*):
X → Y?

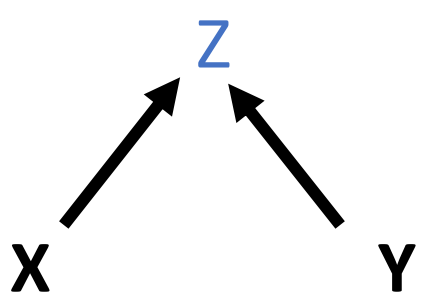
The Pipe



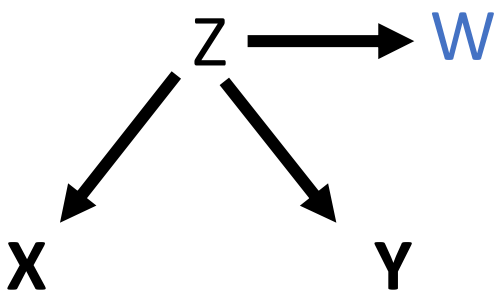
The Fork



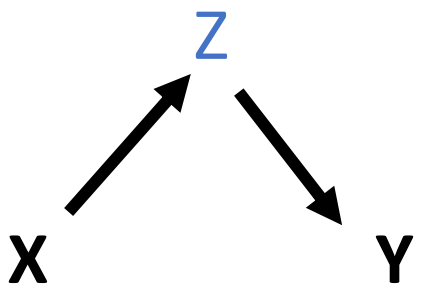
The Collider



The Descendant



The Pipe



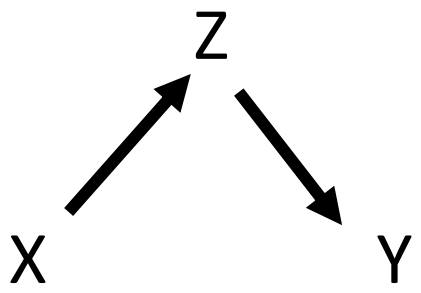
- Z is a mediator of the effect of X on Y

- Controlling for Z blocks the effect of X on Y

- Information flows through the pipe
  ➔ **Open by default**
  ➔ Must be kept open
  Do not control for a mediator (Z)
  → *overcontrol bias*

```
> x <- rnorm(n, 0, 1)
> z <- rbinom(n = n, size = 1,
+             p = inv_logit(x)) # x --> z
> y <- rnorm(n, z, 0.5)         # z --> y
>
> # no control of the mediator:
> round(cor(x, y), 3)
[1] 0.288
> # control of the mediator:
> round(cor(x[z == 0], y[z == 0]), 3)
[1] 0.045
> round(cor(x[z == 1], y[z == 1]), 3)
[1] -0.019
```

The Pipe

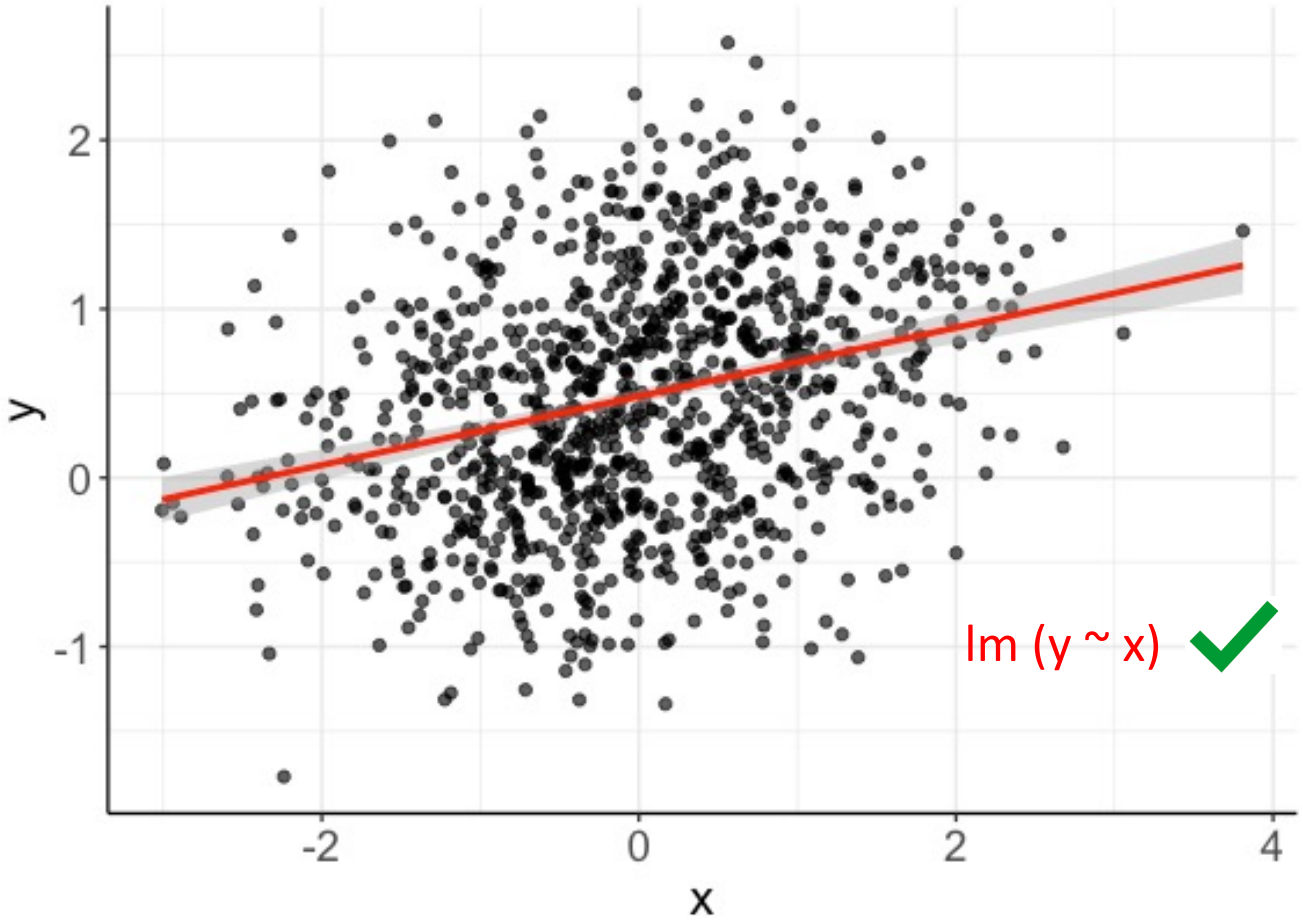


- Z is a mediator of the effect of X on Y

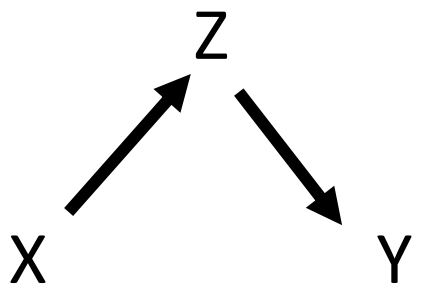
- Controlling for Z blocks the effect of X on Y

- Information flows through the pipe
  ➔ **Open by default**
  ➔ Must be kept open
  Do not control for a mediator (Z)
  ➔ *overcontrol bias*

```
> x <- rnorm(n, 0, 1)
> z <- rbinom(n = n, size = 1,
+             p = inv_logit(x)) # x --> z
> y <- rnorm(n, z, 0.5)         # z --> y
>
> # no control of the mediator:
> round(cor(x, y), 3)
```



The Pipe

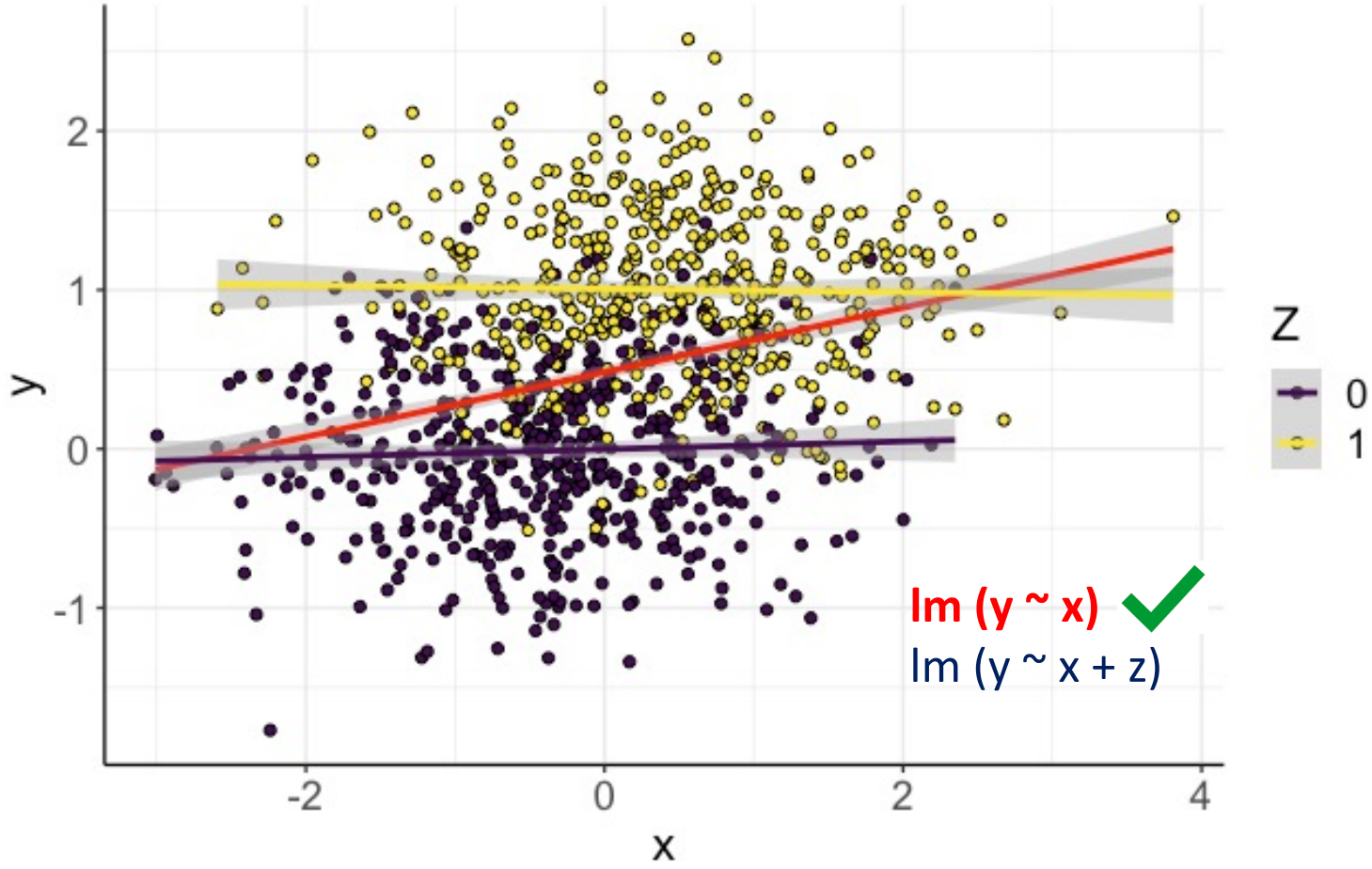


- Z is a mediator of the effect of X on Y

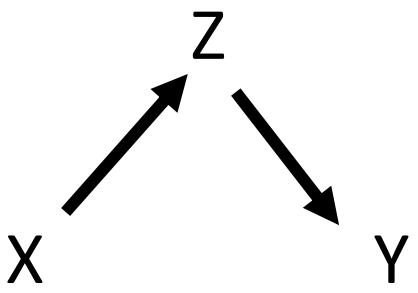
- Controlling for Z blocks the effect of X on Y

- Information flows through the pipe
  ➔ **Open by default**
  ➔ Must be kept open
  Do not control for a mediator (Z)
  ➔ *overcontrol bias*

```
> x <- rnorm(n, 0, 1)
> z <- rbinom(n = n, size = 1,
+             p = inv_logit(x)) # x --> z
> y <- rnorm(n, z, 0.5)         # z --> y
>
> # no control of the mediator:
> round(cor(x, y), 3)
```

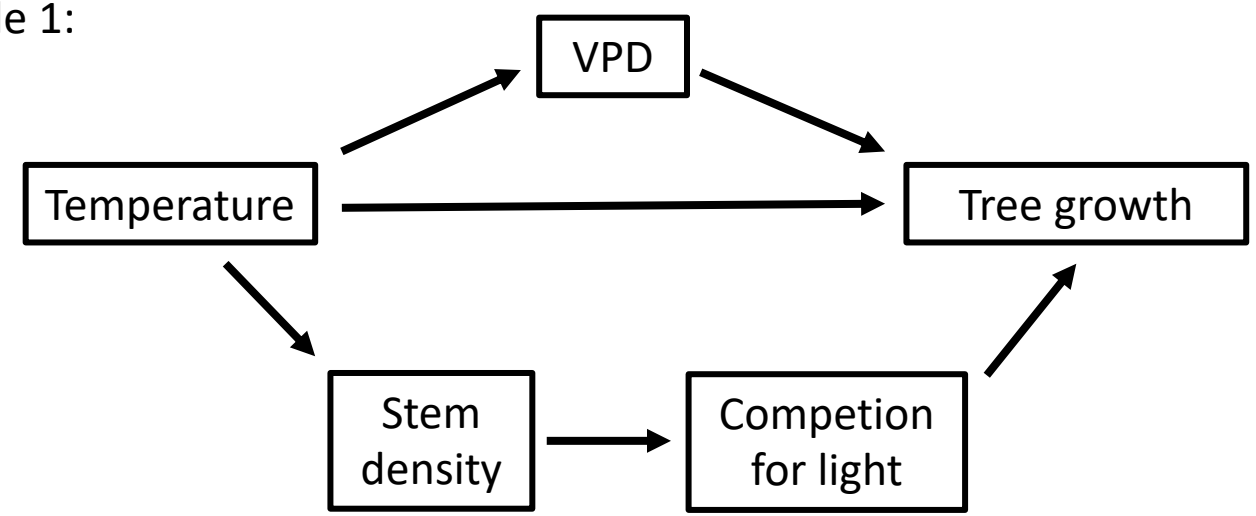


The Pipe

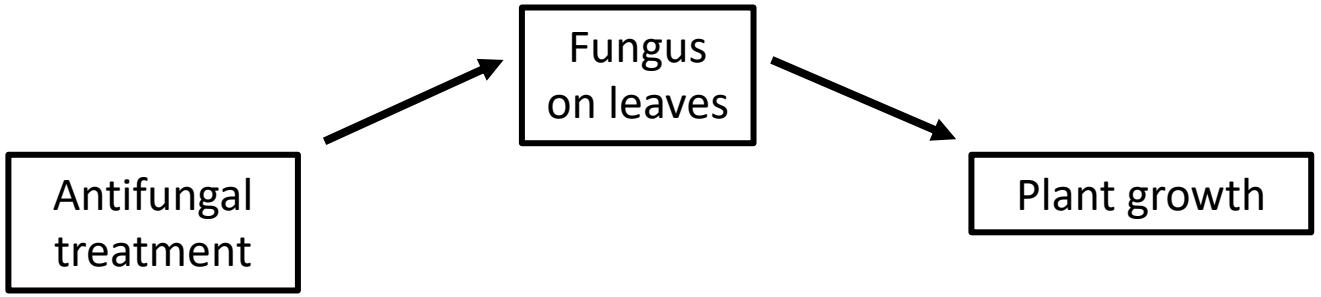


- Overcontrol bias can affect experimental approaches as much as observational ones

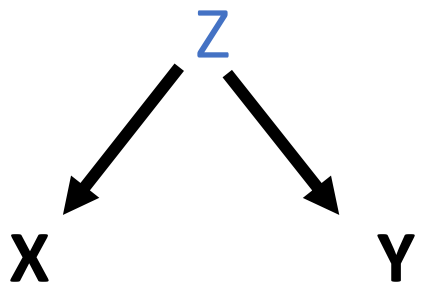
Example 1:



Example 2:



The Fork



- Z (confounder) is a common cause of X and Y

- The way from X to Y is not causal

→ Manipulating X would not change Y

- Information flows through the fork

  ➔ **Forks are open by default**

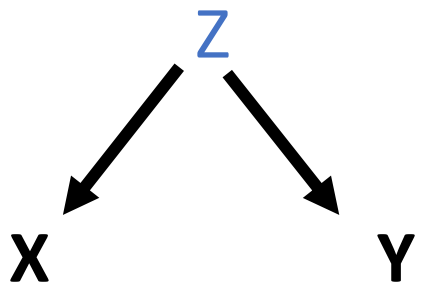
  ➔ Must be closed

  Confounders must be controlled for

  → *confounding bias*

```
> n <- 1000
> z <- rbinom(n, 1, 0.5) # z has no parent
> x <- rnorm(n, z, 0.5)  # z --> x
> y <- rnorm(n, z, 0.3)  # z --> y
> # Without conditioning on the common cause z:
> round(cor(x, y), 3)
[1] 0.597
> # After stratifying by the common cause z:
> round(cor(x[z == 0], y[z == 0]), 3)
[1] 0.031
> round(cor(x[z == 1], y[z == 1]), 3)
[1] 0.032
```

The Fork



- Z (confounder) is a common cause of X and Y

- The way from X to Y is not causal

→ Manipulating X would not change Y

- Information flows through the fork

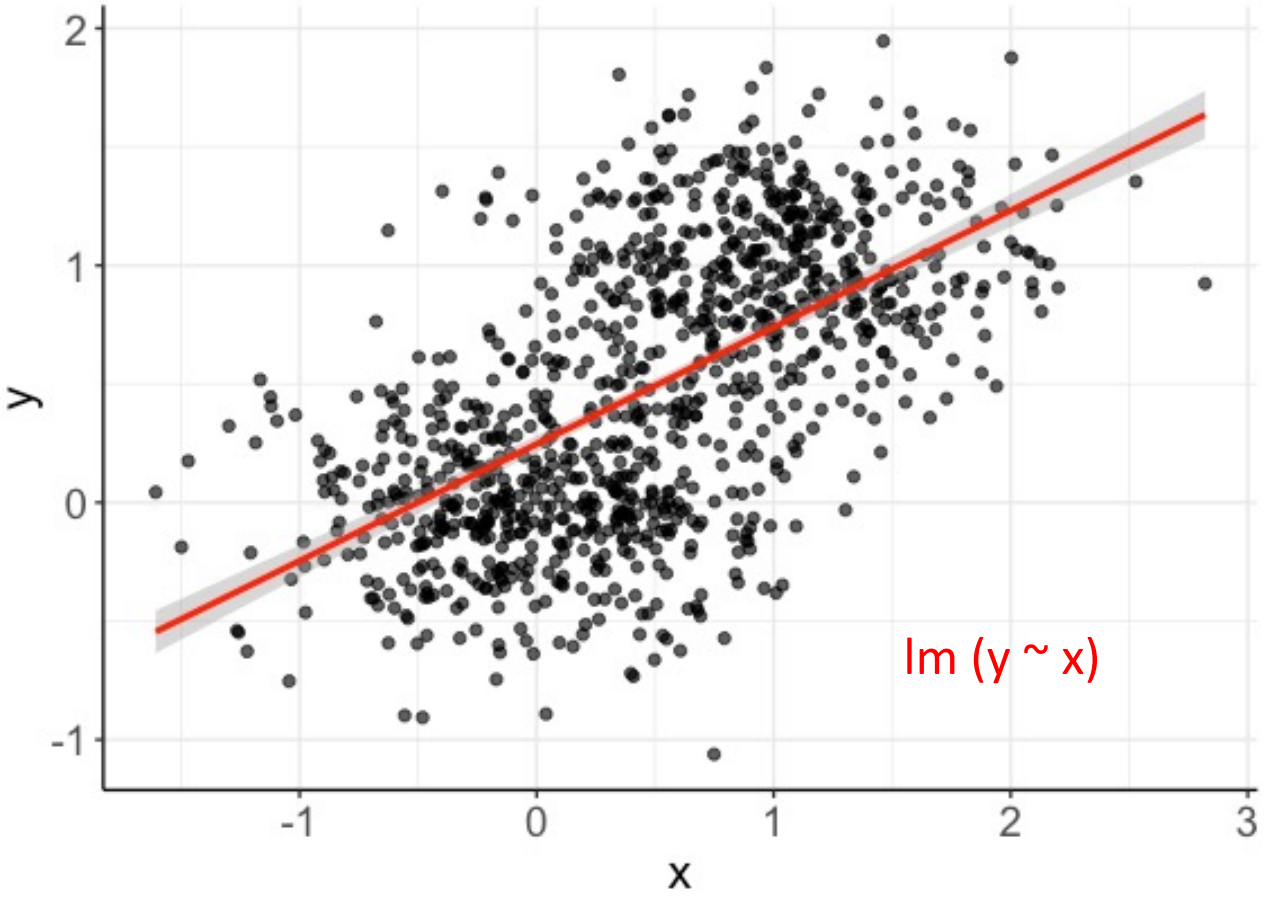
  ➔ **Forks are open by default**

  ➔ Must be closed

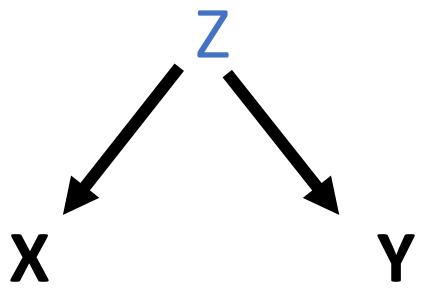
  Confounders must be controlled for

  → *confounding bias*

```
> n <- 1000
> z <- rbinom(n, 1, 0.5) # z has no parent
> x <- rnorm(n, z, 0.5)  # z --> x
> y <- rnorm(n, z, 0.3)  # z --> y
> # Without conditioning on the common cause z:
> round(cor(x, y), 3)
[1] 0.597
```



The Fork



- Z (confounder) is a common cause of X and Y

- The way from X to Y is not causal

→ Manipulating X would not change Y

- Information flows through the fork

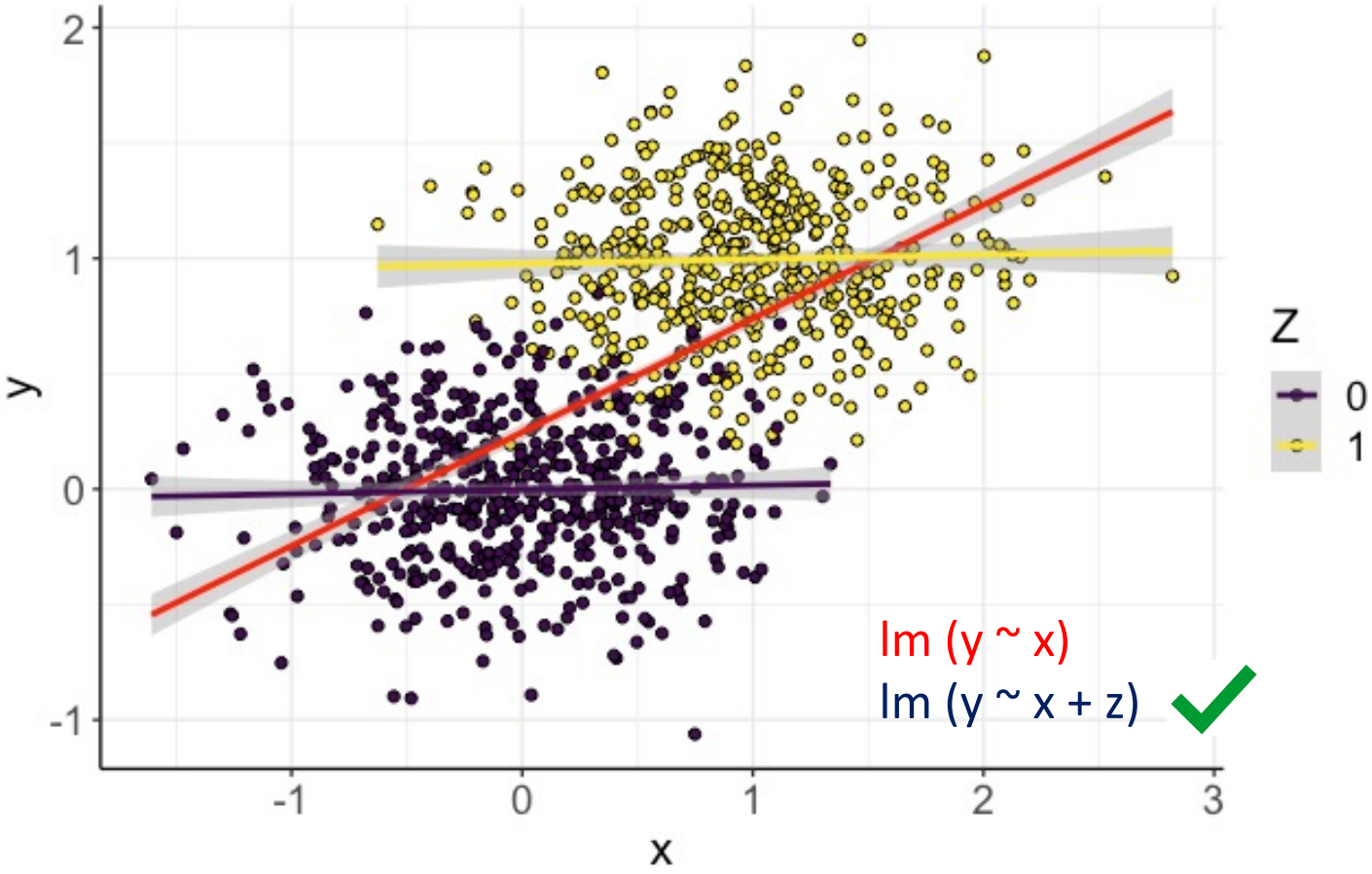
  ➔ **Forks are open by default**

  ➔ Must be closed

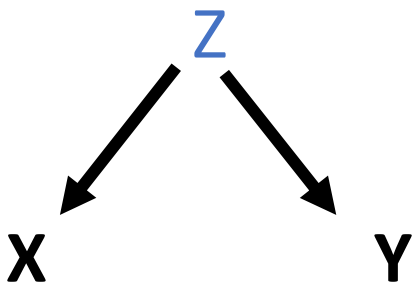
  Confounders must be controlled for

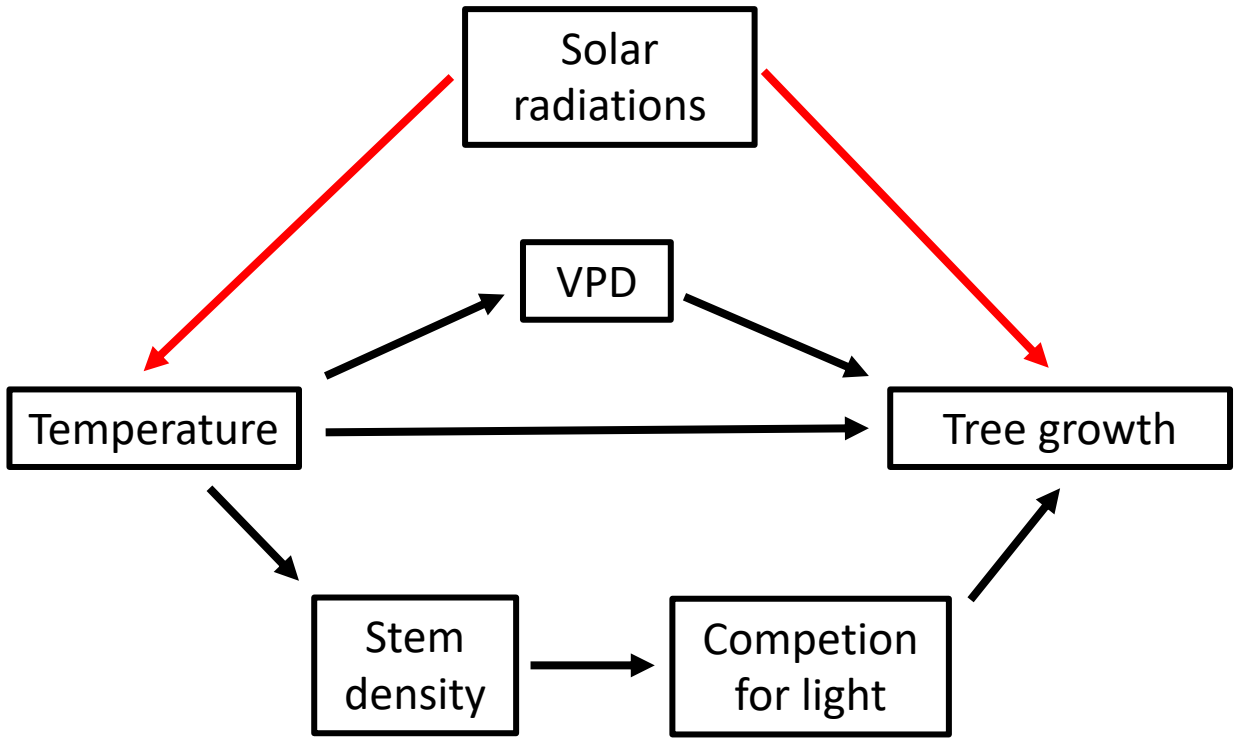
  → *confounding bias*

```
> n <- 1000
> z <- rbinom(n, 1, 0.5) # z has no parent
> x <- rnorm(n, z, 0.5)  # z --> x
> y <- rnorm(n, z, 0.3)  # z --> y
> # Without conditioning on the common cause z:
> round(cor(x, y), 3)
[1] 0.597
```

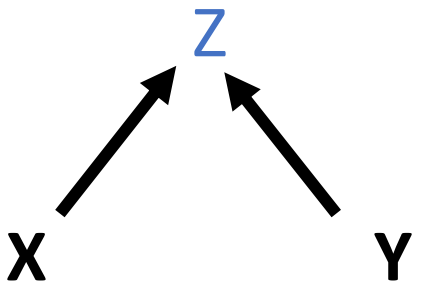


The Fork


Z
X
Y


Solar radiations
VPD
Temperature
Tree growth
Stem density
Competion for light

The Collider



- Z is a collider between X and Y

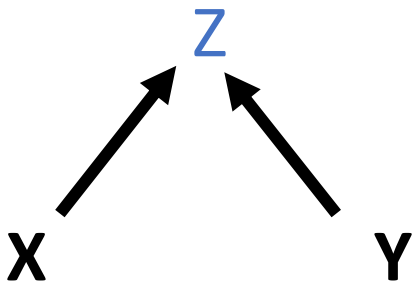
- The way from X to Y is not causal
  → Changing X would not affect Y

- Information does not flow through the collider
  ➔ **The path is closed** by default...
  ... and must stay closed
  Never condition on a collider

That one weird third variable problem nobody ever mentions: Conditioning on a collider

https://www.the100.ci/2017/03/14/that-one-weird-third-variable-problem-nobody-ever-mentions-conditioning-on-a-collider/

The Collider



- Z is a collider between X and Y

- The way from X to Y is not causal
  → Changing X would not affect Y

- Information does not flow through the collider
  ➔ **The path is closed** by default...
  ... and must stay closed
  Never condition on a collider

```
> n <- 1000
> x <- rnorm(n, 0, 1)
> y <- rnorm(n, 0, 1)
> # z <- rnorm(n, x + y, 1)
> z <- rbinom(n, 1, ifelse(x + y > 0, 0.9, 0.1))
```

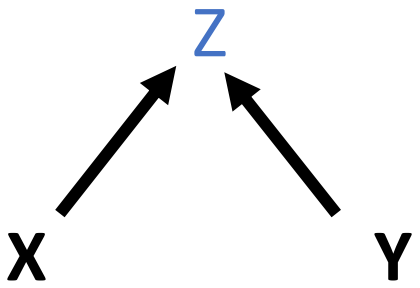
lm (y ~ x)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02047    0.03154  -0.649    0.516
x           -0.02799    0.03161  -0.885    0.376
```

lm (y ~ x + z)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.55748    0.04068 -13.703   <2e-16 ***
x           -0.25461    0.03030  -8.404   <2e-16 ***
z            1.08124    0.06045  17.887   <2e-16 ***
```

The Collider

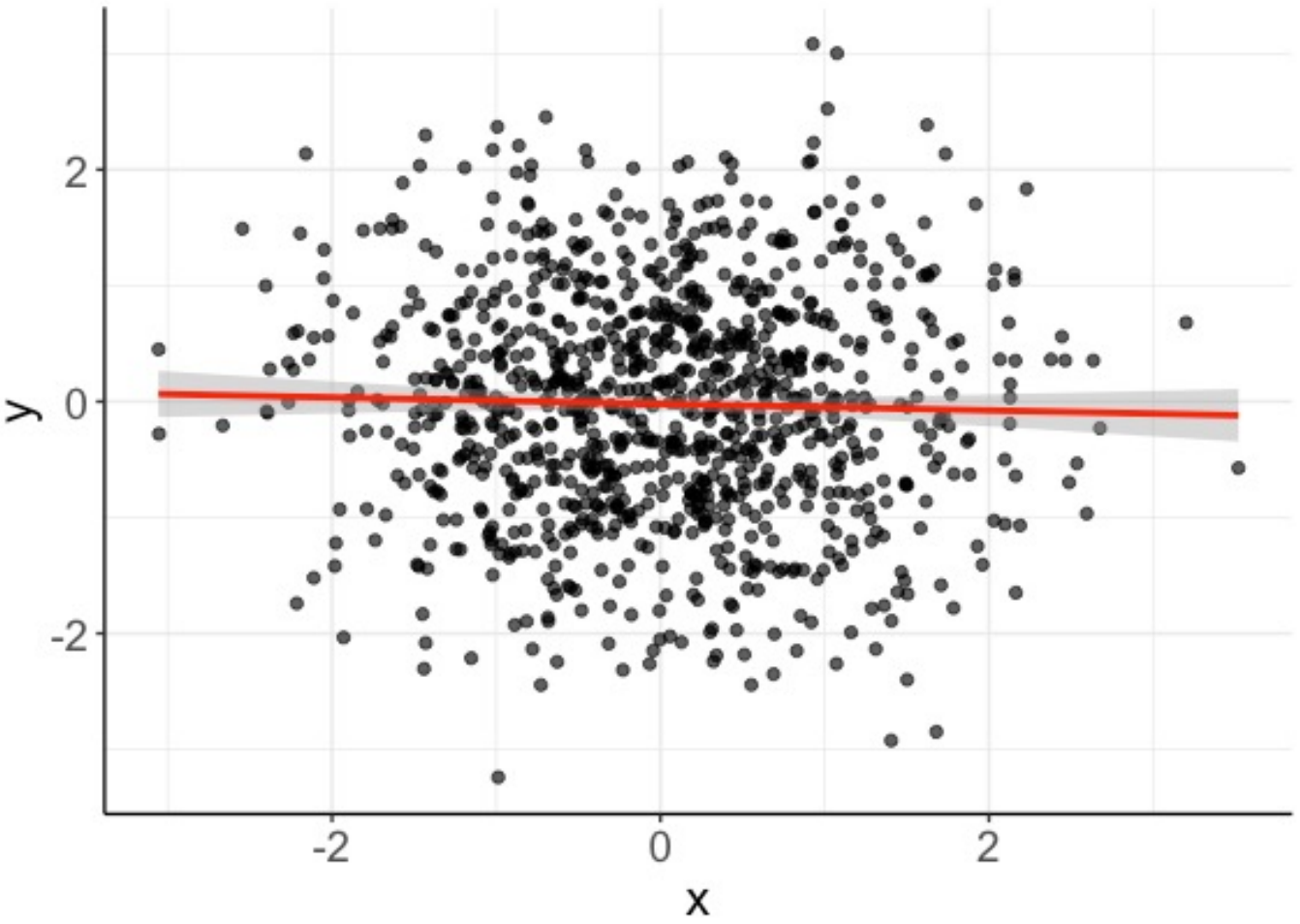


- Z is a collider between X and Y

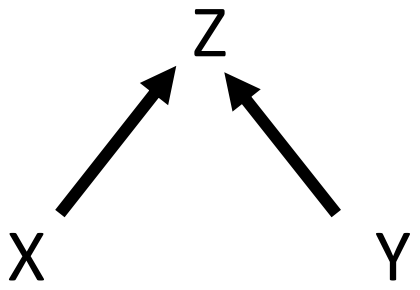
- The way from X to Y is not causal
  → Changing X would not affect Y

- Information does **not** flow through the collider
  ➔ **The path is closed** by default...
  ... and must stay closed
  Never condition on a collider

```
> n <- 1000
> x <- rnorm(n, 0, 1)
> y <- rnorm(n, 0, 1)
> # z <- rnorm(n, x + y, 1)
> z <- rbinom(n, 1, ifelse(x + y > 0, 0.9, 0.1))
```



The Collider

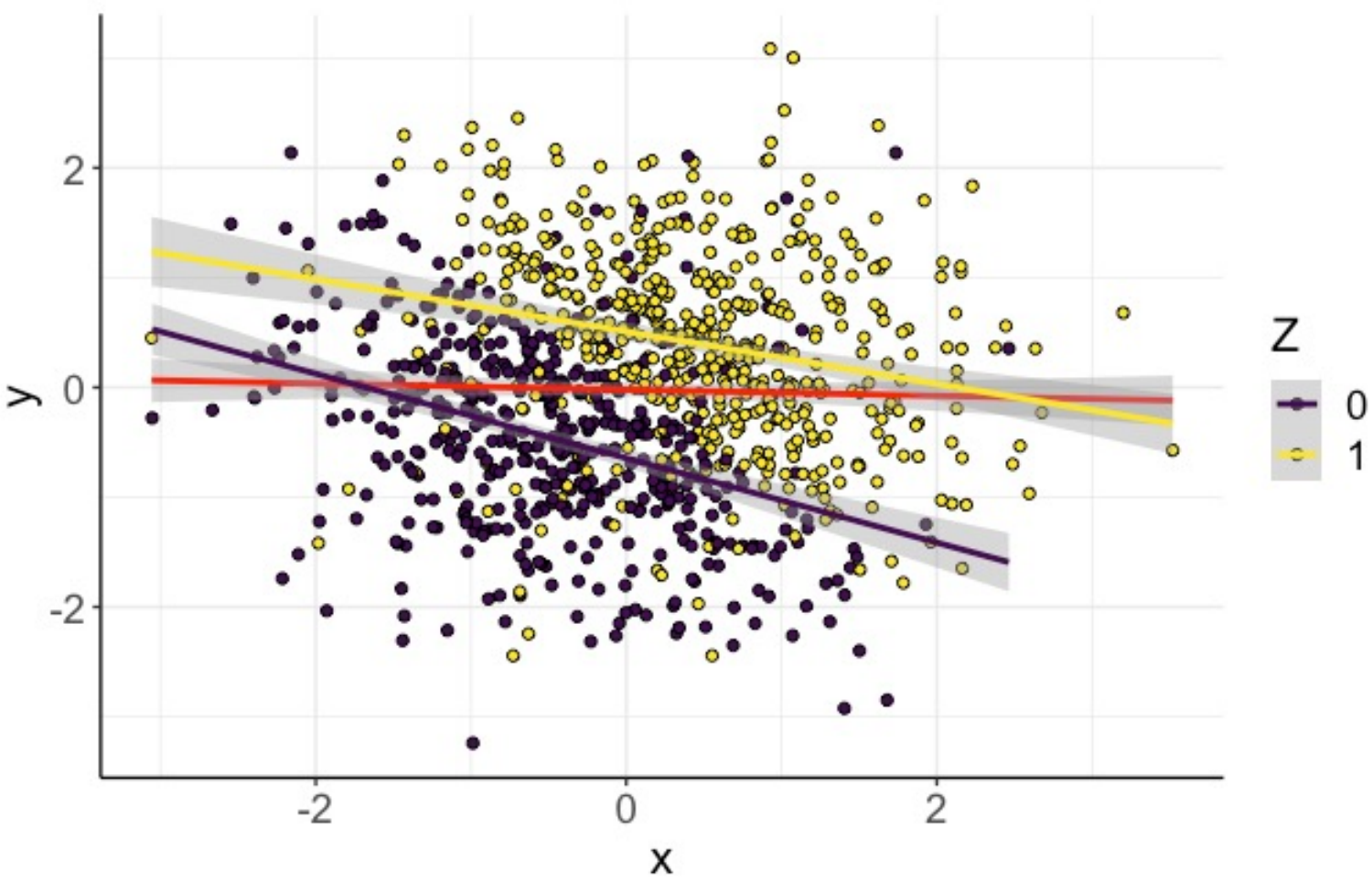


- Z is a collider between X and Y

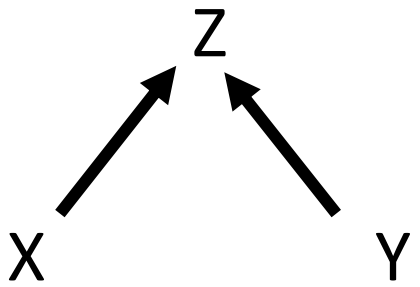
- The way from X to Y is not causal
  → Changing X would not affect Y

- Information does **not** flow through the collider
  ➔ **The path is closed** by default...
  ... and must stay closed!
  Never control for a collider

```
> n <- 1000
> x <- rnorm(n, 0, 1)
> y <- rnorm(n, 0, 1)
> # z <- rnorm(n, x + y, 1)
> z <- rbinom(n, 1, ifelse(x + y > 0, 0.9, 0.1))
```



The Collider

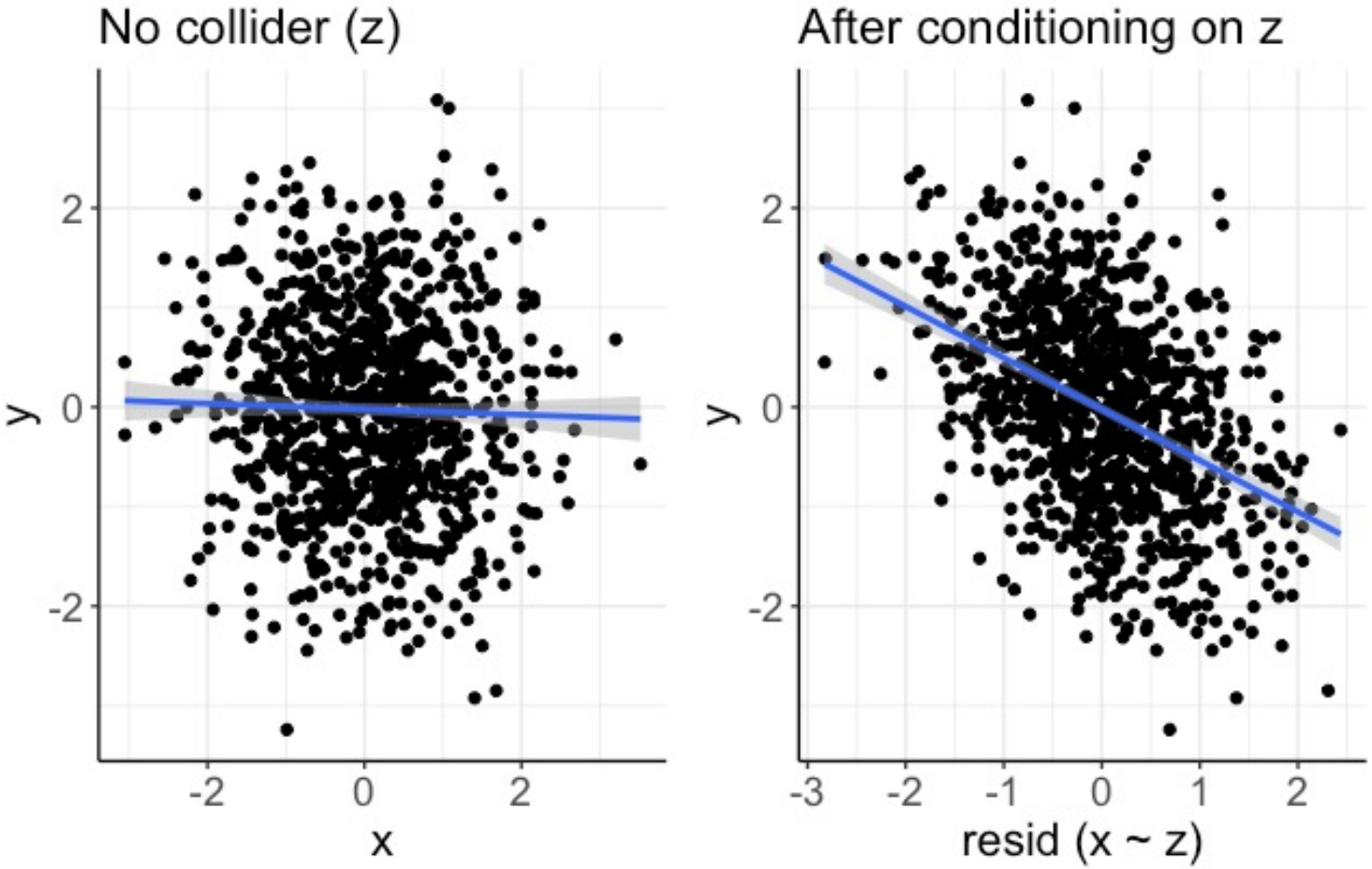


- Z is a collider between X and Y

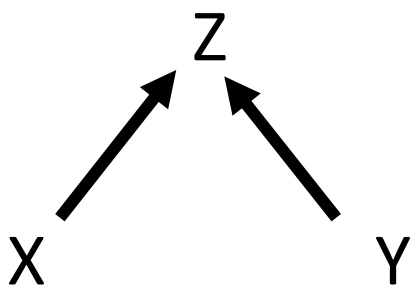
- The way from X to Y is not causal
  → Changing X would not affect Y

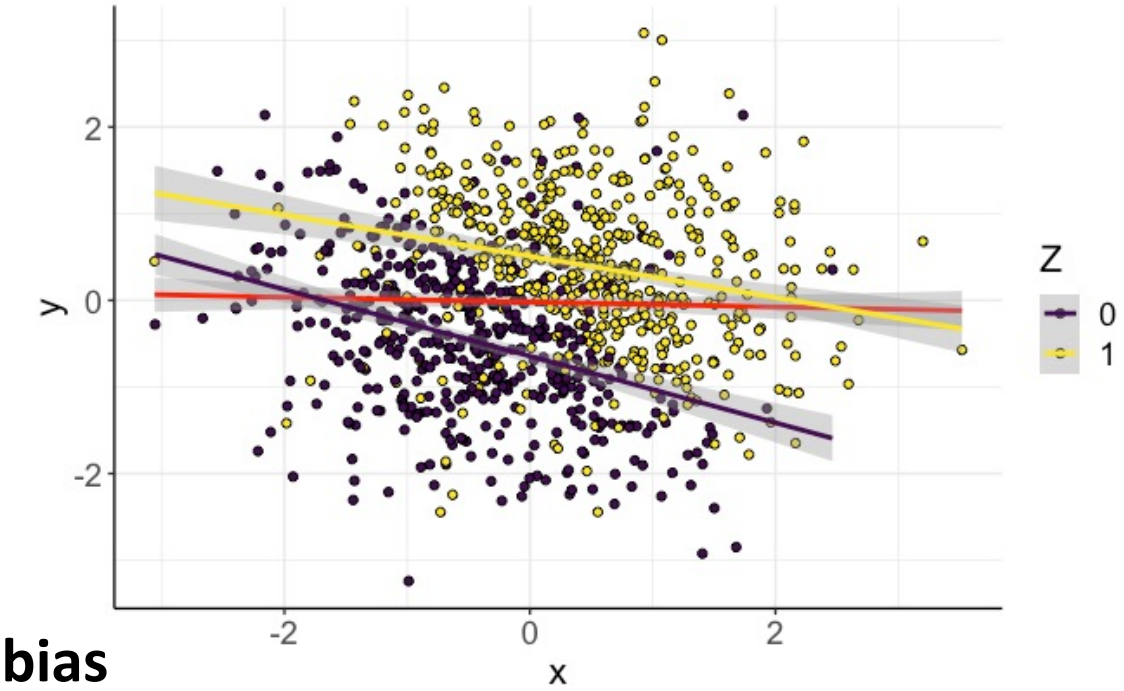
- Information does **not** flow through the collider
  ➔ **The path is closed** by default...
  ... and must stay closed
  Never condition on a collider

```r
# The collider:
# *************
# x -> z
# y -> z
# Both x and y cause z. x and y are not associated.
set.seed(3)
n <- 1000
x <- rnorm(n, 0, 1)
y <- rnorm(n, 0, 1)
z <- rnorm(n, x + y, 1)
```



The Collider





- Z as « selected / rejected » condition
  → rejected cases are not observable → **selection bias**

- Examples:

Participating in survey

depression — health status

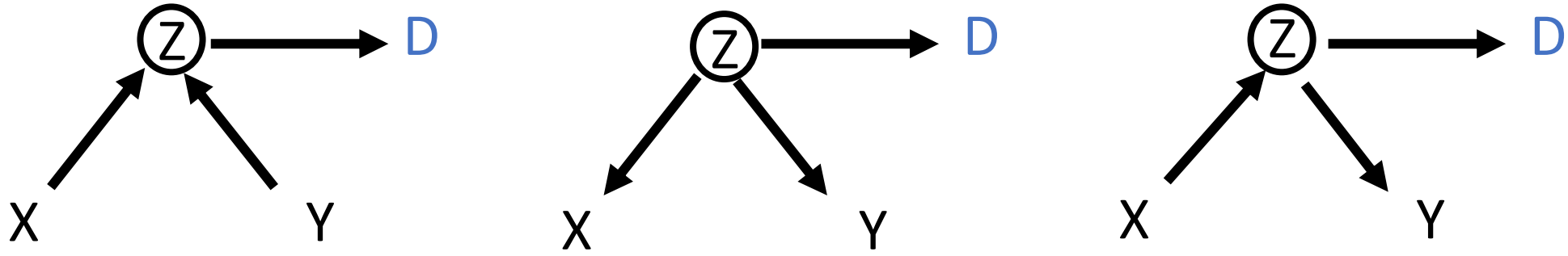
univ. student

IQ — perseverance

recommend paper

rigour — match my bias

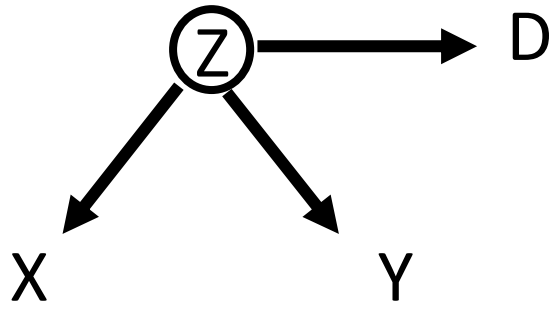
The Descendant

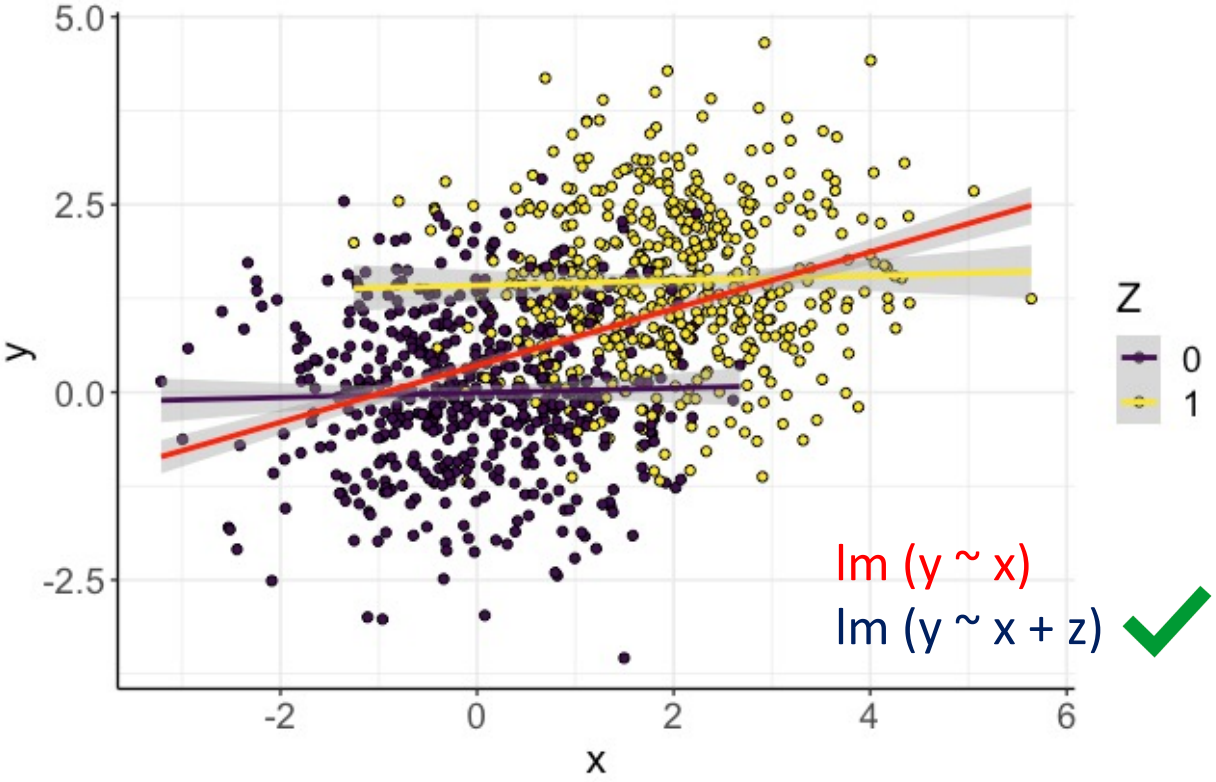


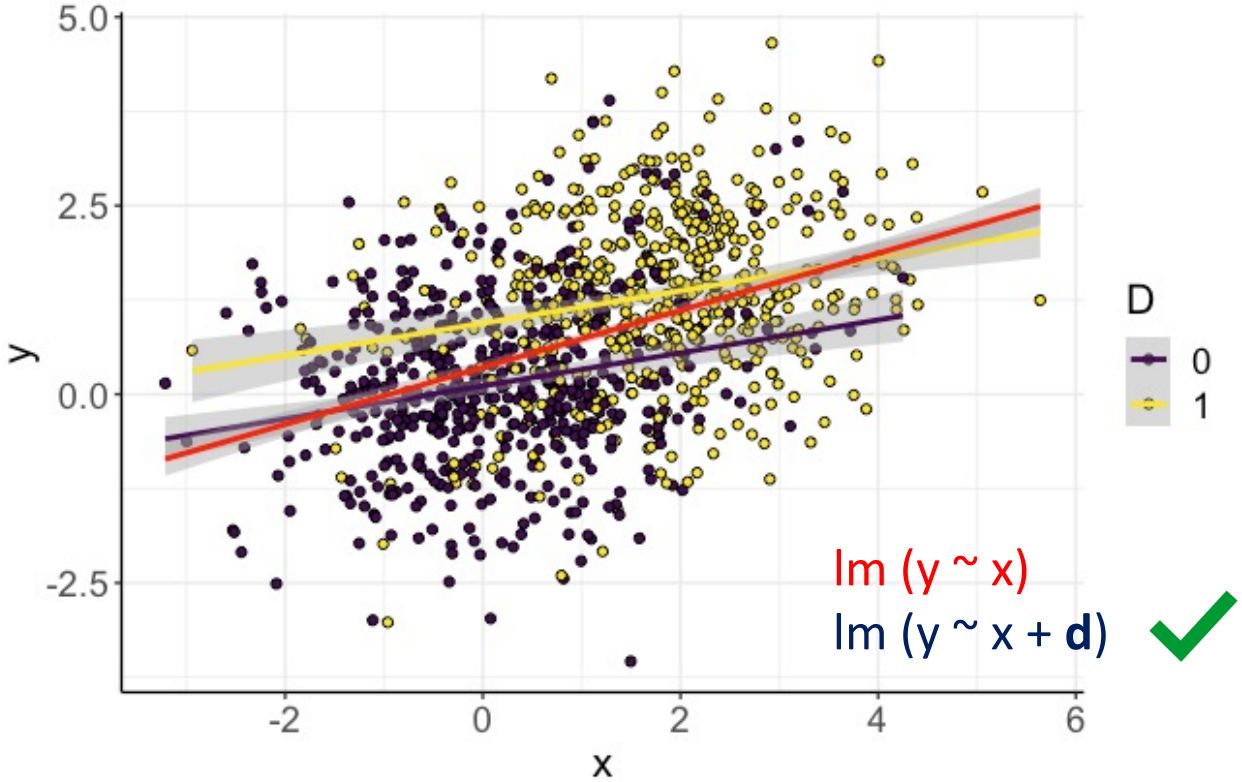
- D is a descendant of Z
- Conditioning on D = noisy version of conditioning on Z

➔ Apply to D the same rules as you would for the parent (Z)

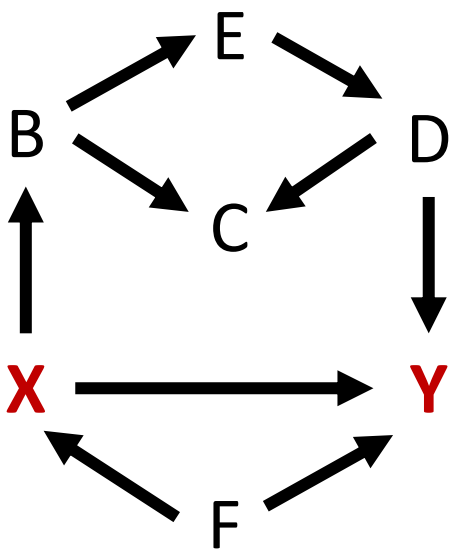
The Descendant

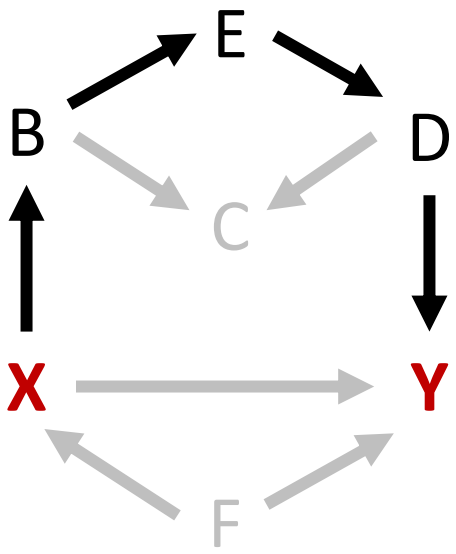




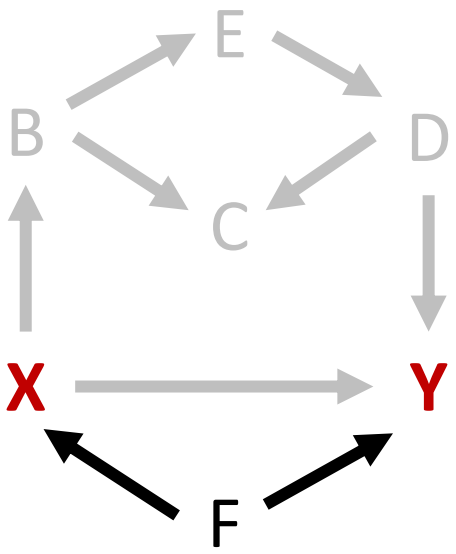


DAGs are combination of elemental confounds

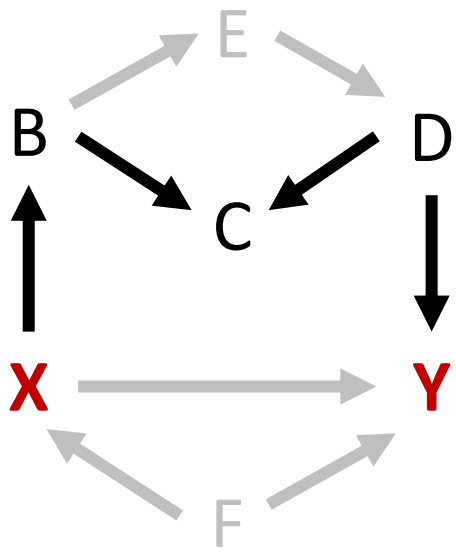




Pipes



Forks



Colliders

From a causal model...
... to a statistical model

The Backdoor Criterion

- Causal effect p(Y | do(X))

→ Distribution of Y when intervening on X

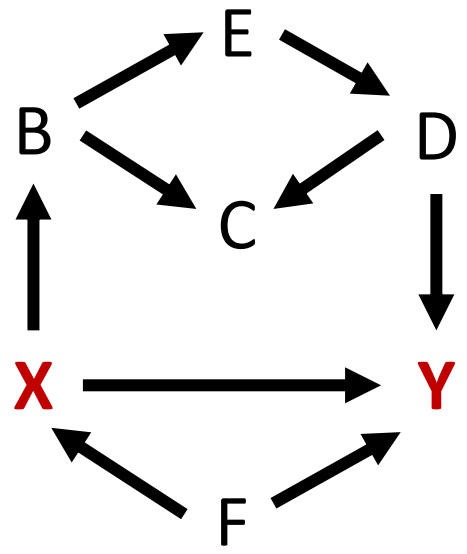
1) Define all paths relating X and Y

2) Differentiate the *causal* from the *non-causal* paths

3) Close (or keep closed) the non-causal paths

Minimal set of control variables to estimate X → Y: **F**

Conditioning on anything else would bias the causal estimate



Causal paths

X → Y

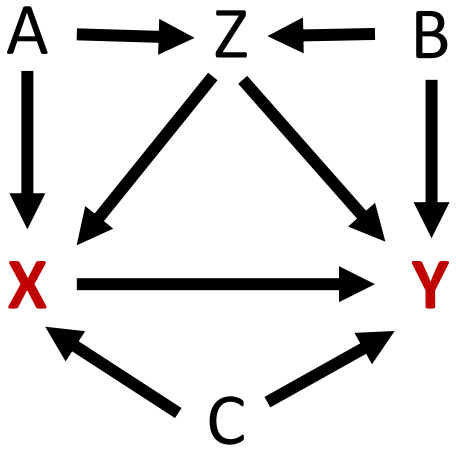
X → B → E → D → Y

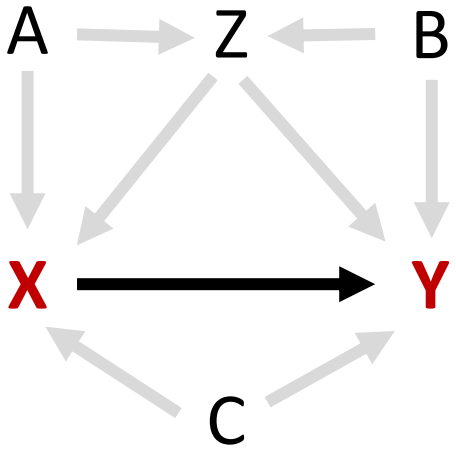
Non causal paths

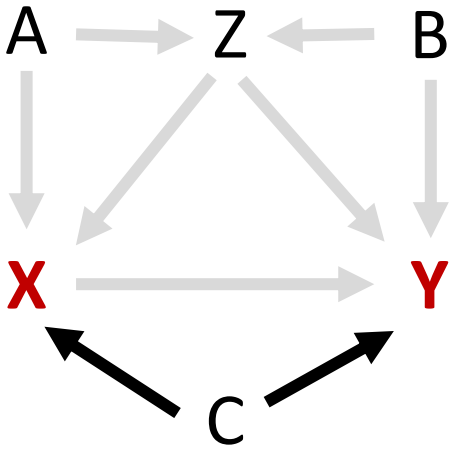
X ← F → Y

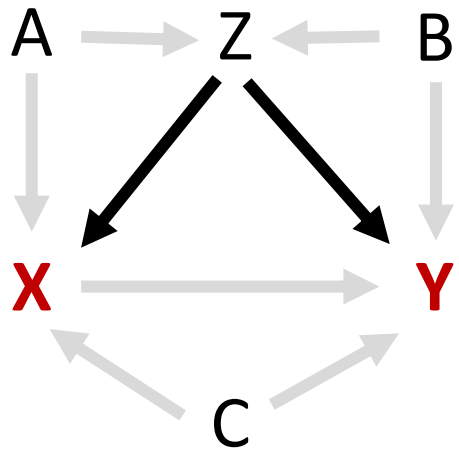
X → B → C ← D → Y

From a causal model...
... to a statistical model

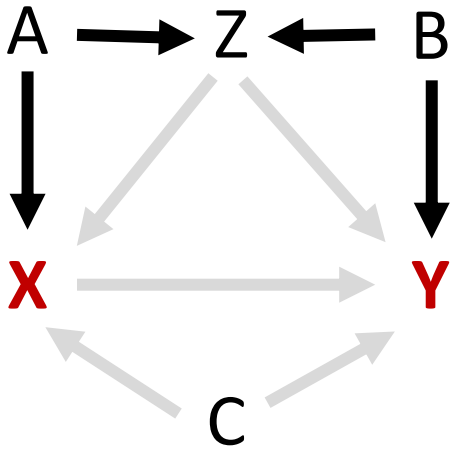


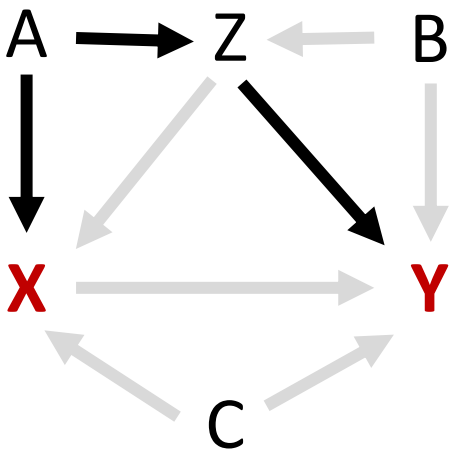


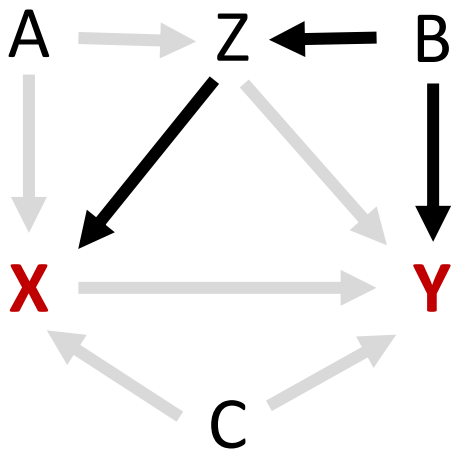




We control for:
C,
Z,
A or B,
A or Z,
B or Z

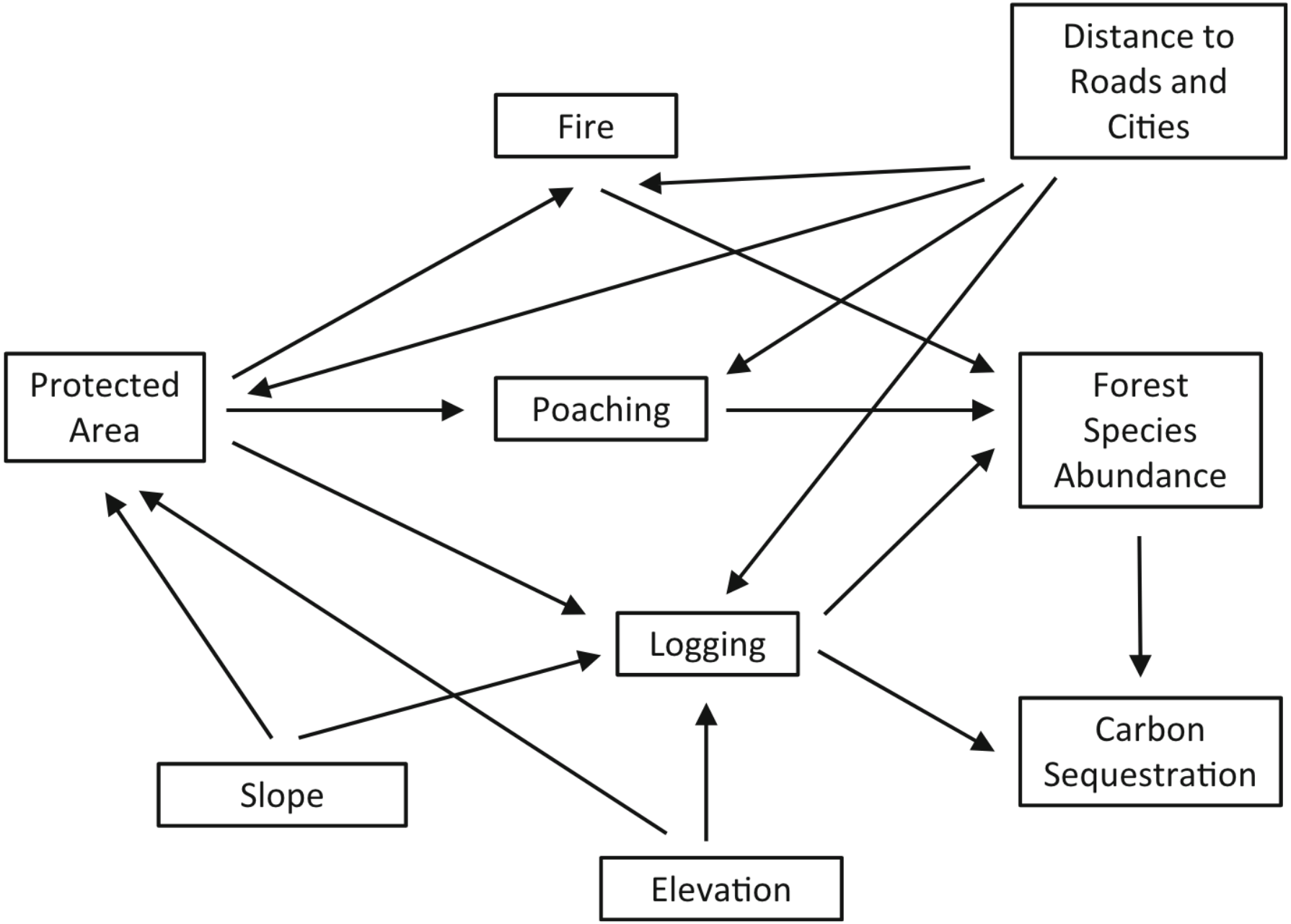






➔ C, Z, B

➔ Y ~ f(X + C + Z + B)

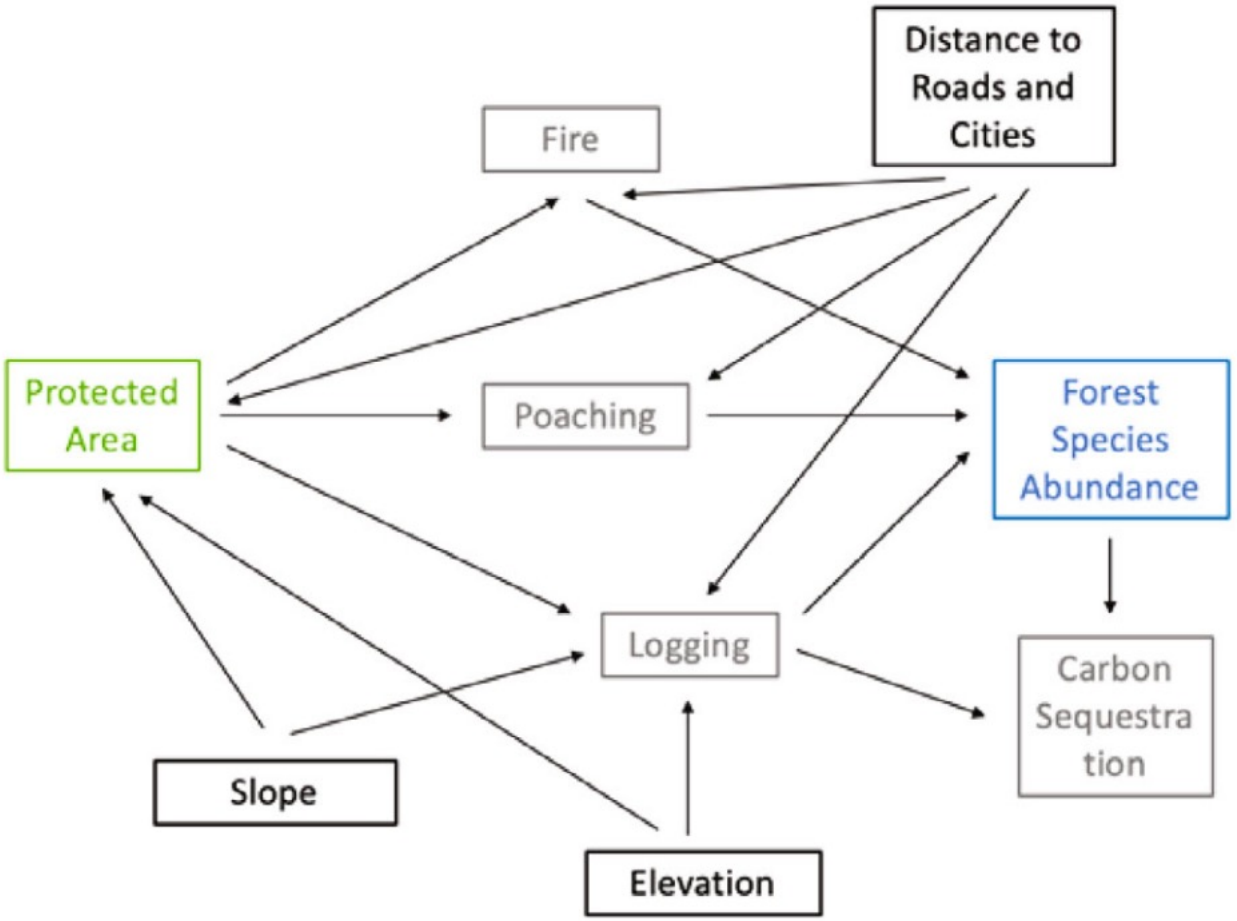


Arif & MacNeil 2022

Protected Area Model

Known: **1.98**

Estimated: **1.98 ± 0.03 (SE)**



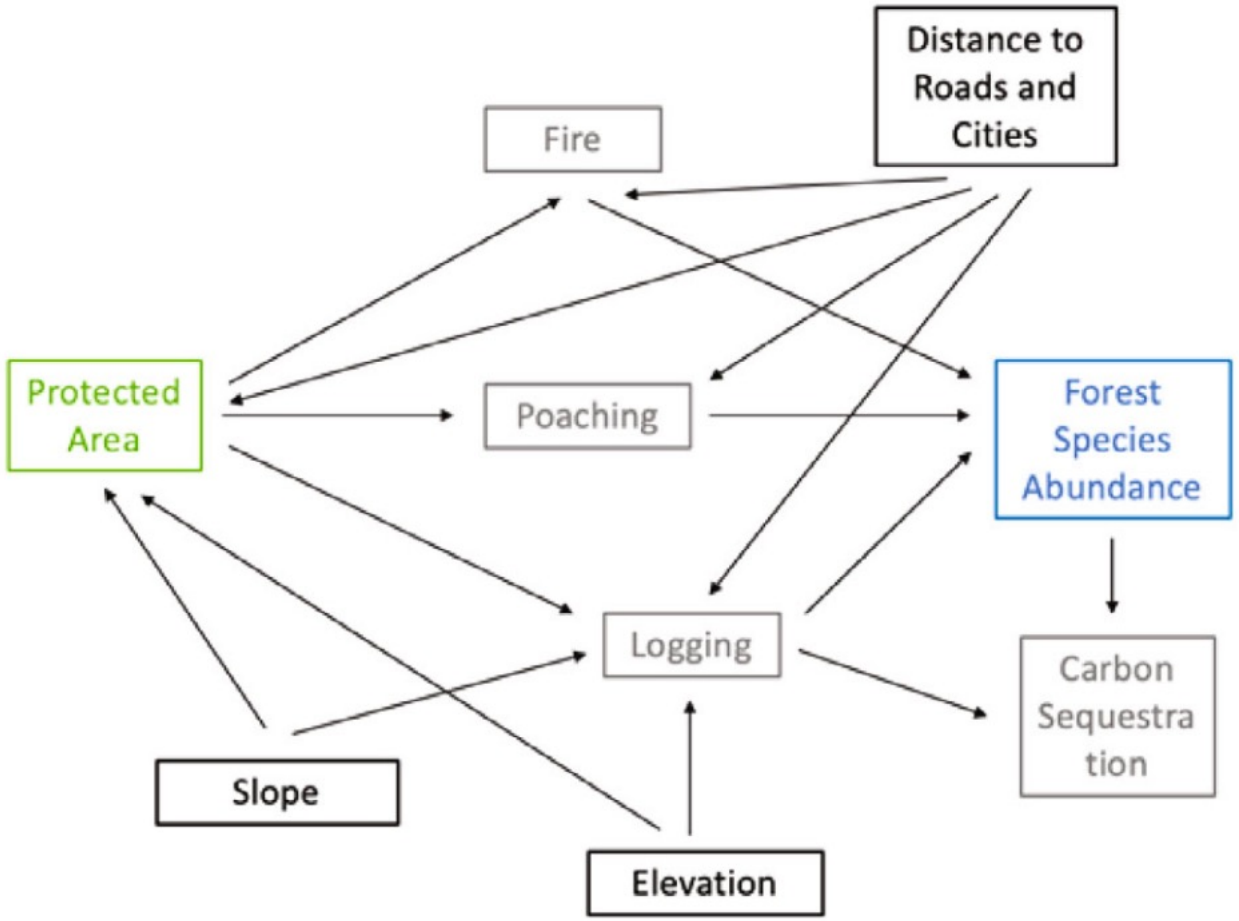
AIC: 35347

Arif & MacNeil 2022

Protected Area Model

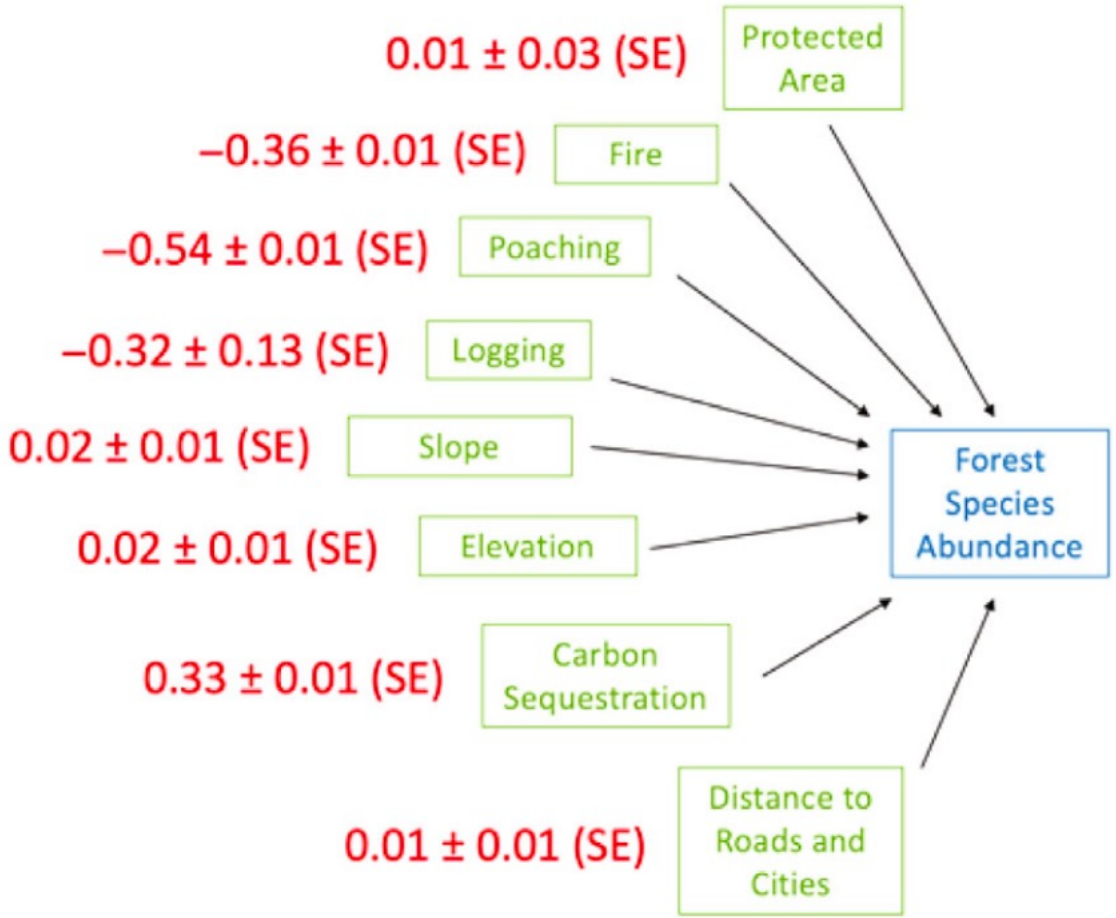
Known: **1.98**

Estimated: **1.98 ± 0.03 (SE)**



AIC: 35347

Causal Salad Model



AIC: 26978

Arif & MacNeil 2022

The Table 2 Fallacy

American Journal of Epidemiology

Vol. 177, No. 4
DOI: 10.1093/aje/kws412
Advance Access publication:
January 30, 2013

Commentary

The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients

Daniel Westreich* and Sander Greenland

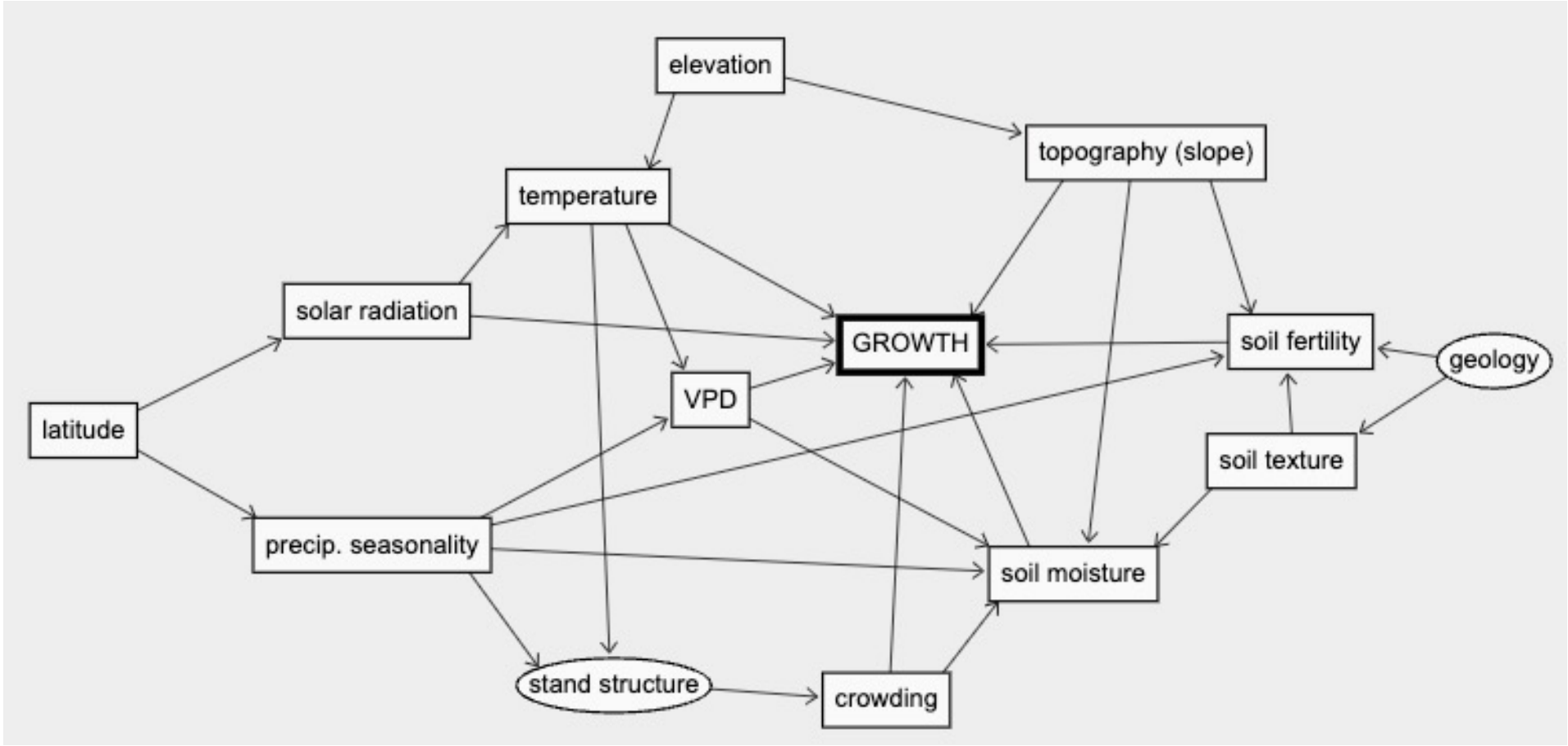
➔ Carefully think of what to present in a result Table or Figure

Coefficients of different covariates can be:

- total causal effects
- direct or partial causal effects

➔ Only present comparable types of effects
➔ Don't present control variables or only partially unconfounded effects

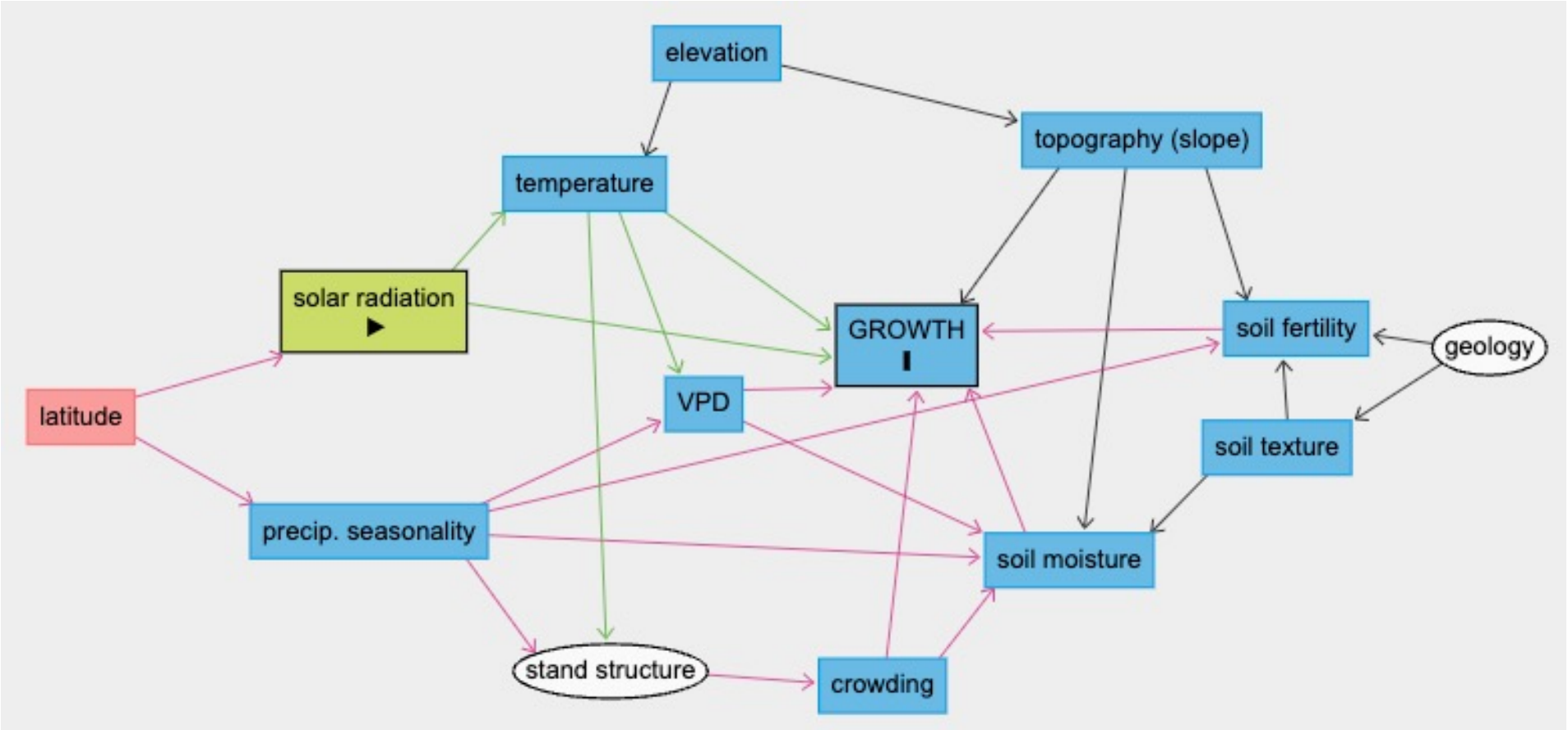
The Table 2 Fallacy




```
dag {
bb="0,0,1,1"
"precip. seasonality" [pos="0.204,0.501"]
"soil fertility" [pos="0.370,0.417"]
"soil moisture" [pos="0.329,0.513"]
"soil texture" [pos="0.366,0.466"]
"solar radiation" [pos="0.204,0.404"]
"stand structure" [latent,pos="0.247,0.560"]
"topography (slope)" [pos="0.342,0.338"]
GROWTH [pos="0.300,0.418"]
VPD [pos="0.264,0.441"]
crowding [pos="0.296,0.566"]
elevation [pos="0.266,0.302"]
geology [latent,pos="0.404,0.425"]
latitude [pos="0.152,0.454"]
temperature [pos="0.242,0.356"]
"precip. seasonality" -> "soil fertility"
"precip. seasonality" -> "soil moisture"
"precip. seasonality" -> "stand structure"
"precip. seasonality" -> VPD
"soil fertility" -> GROWTH
"soil moisture" -> GROWTH
"soil texture" -> "soil fertility"
"soil texture" -> "soil moisture"
"solar radiation" -> GROWTH
"solar radiation" -> temperature
"stand structure" -> crowding
"topography (slope)" -> "soil fertility"
"topography (slope)" -> "soil moisture"
"topography (slope)" -> GROWTH
VPD -> "soil moisture"
VPD -> GROWTH
crowding -> "soil moisture"
crowding -> GROWTH
elevation -> "topography (slope)"
elevation -> temperature
geology -> "soil fertility"
geology -> "soil texture"
latitude -> "precip. seasonality"
latitude -> "solar radiation"
temperature -> "stand structure"
temperature -> GROWTH
temperature -> VPD
}
```

Great tool for DAGs, backdoor criterion, etc: https://dagitty.net/dags.html#

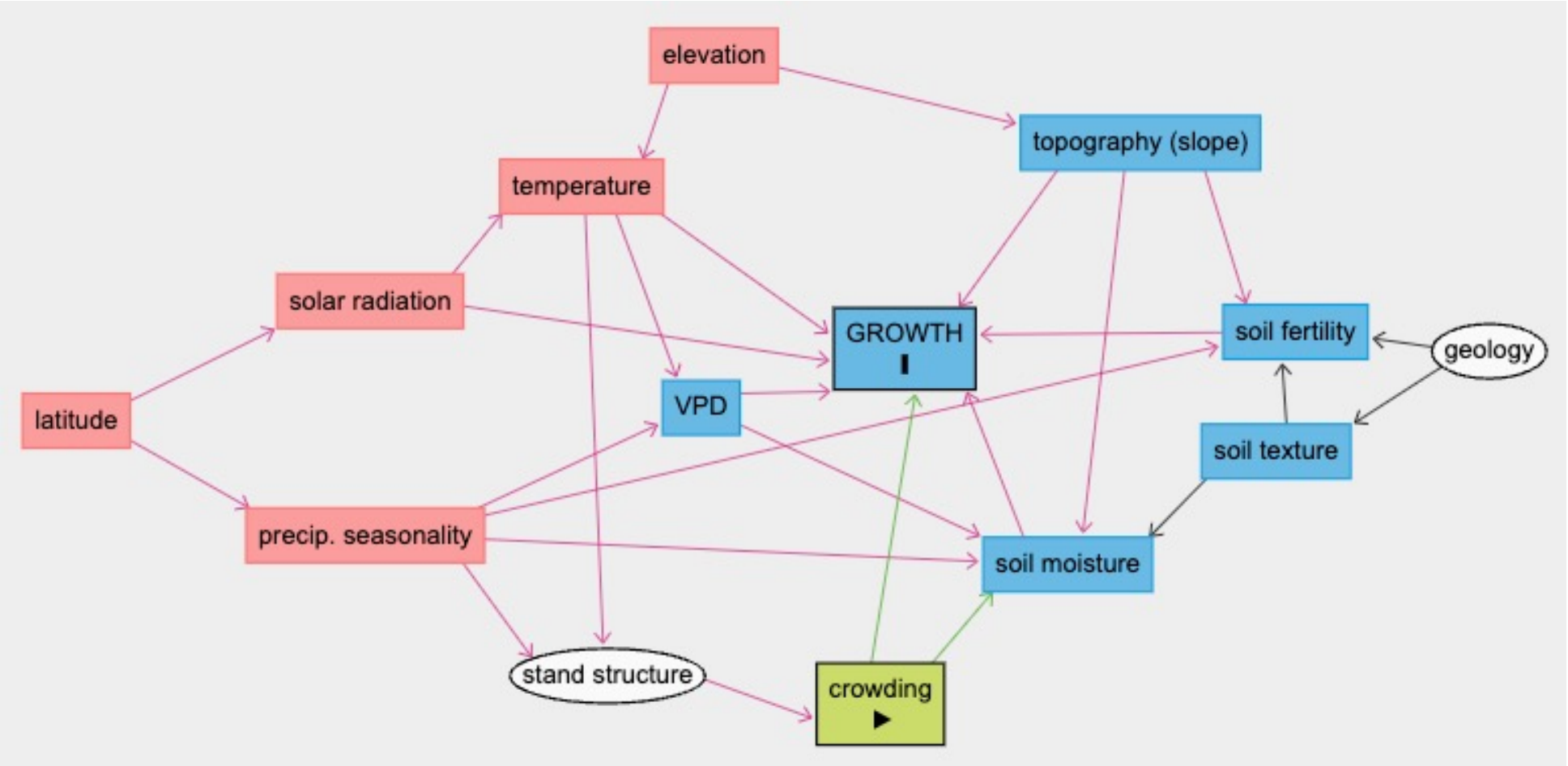
Total causal effect of solar radiation



Minimal set of control variables:

- latitude
- precip. seasonality

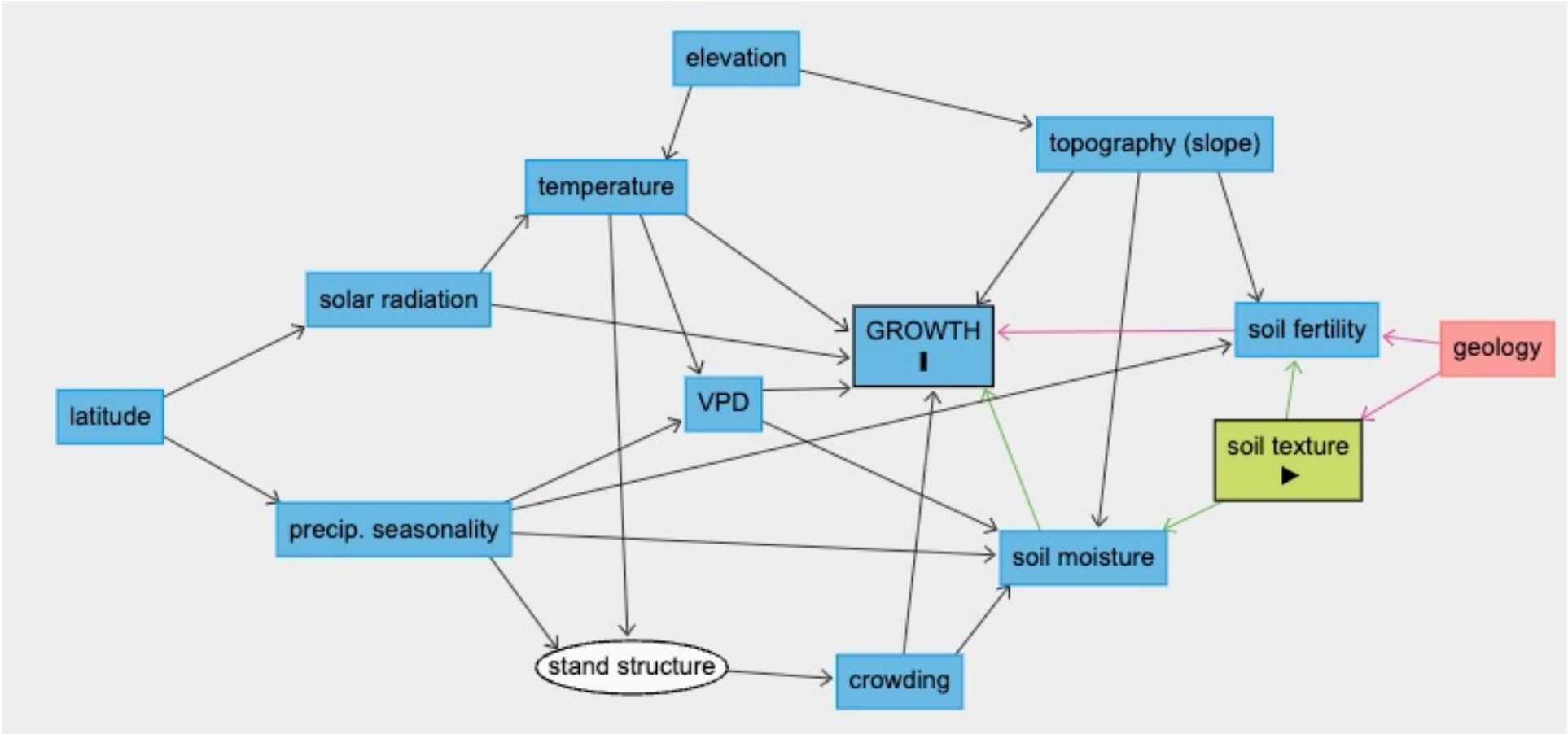
Total causal effect of neighbourhood crowding (competition, natural ennemies)



Minimal set of control variables:

- temperature + precip. seasonality

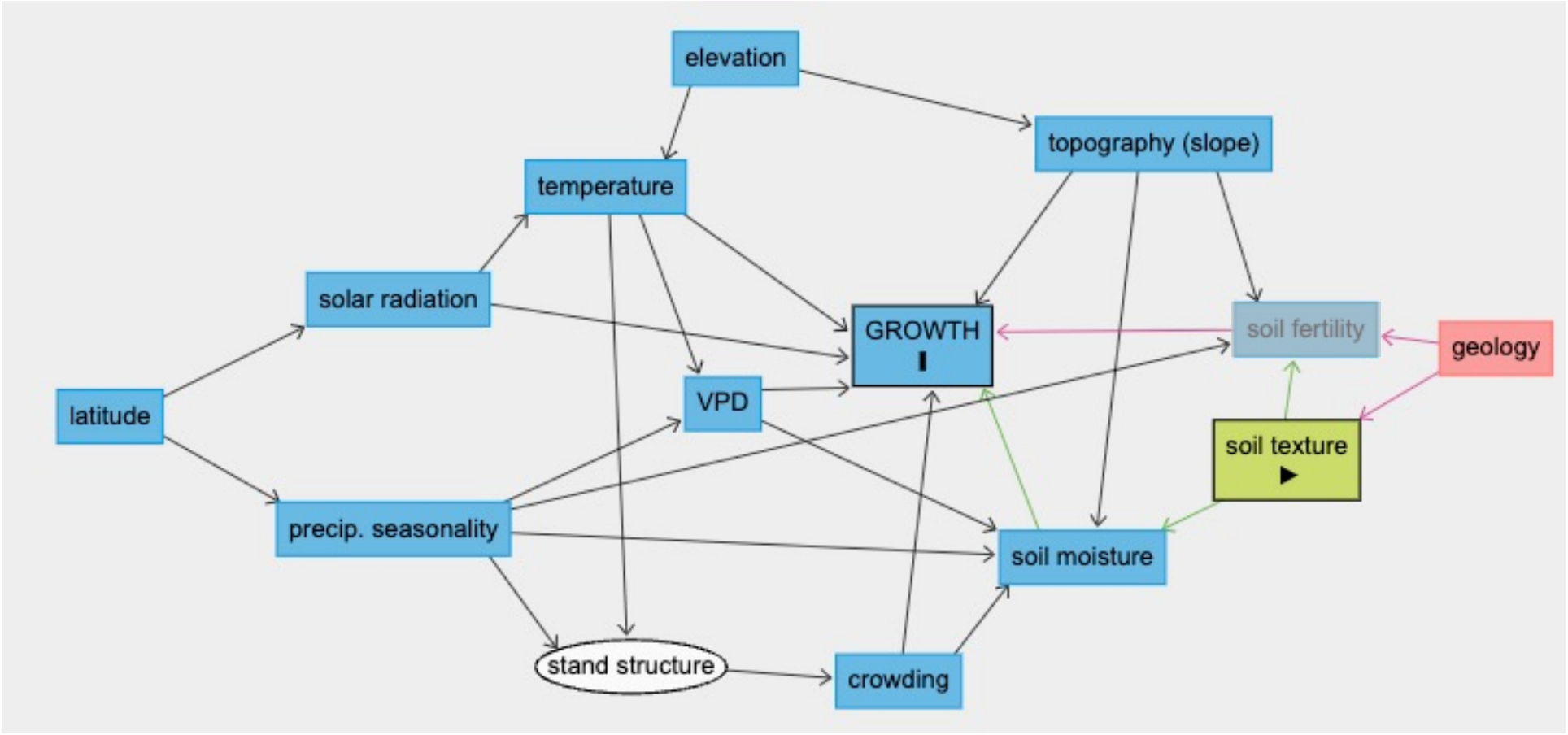
Total causal effect of soil texture (e.g. clay content)



Minimal set:

- geology

Total causal effect of soil texture (e.g. clay content)



Minimal set:

- geology

Conditioning on fertility (e.g. GROWTH ~ texture + fertility) blocks a pipe of causation, *AND* opens several non-causal paths (collider bias), through ‘topography (slope)’

TAKE-HOME MESSAGES

- **« Science before statistics »**

- **Multiple regression can help as much as it can hurt our inference**

  ↓ controlling for counfounders

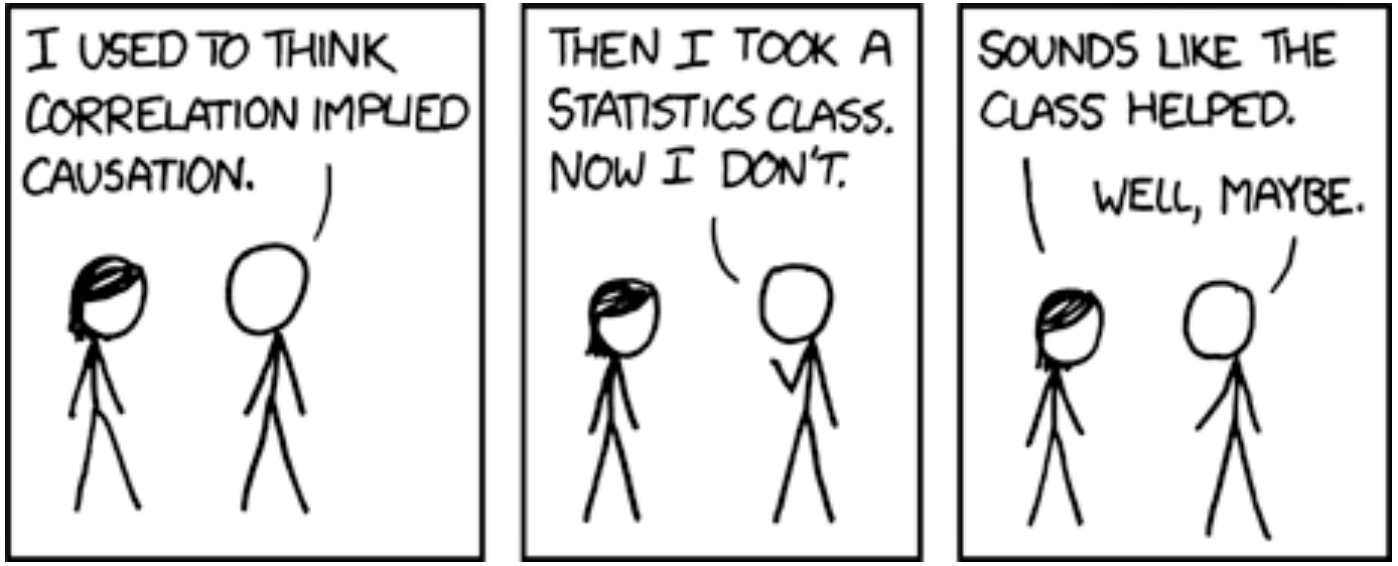
  ↓ - overcontrol bias
  - collider bias

- **The causes are not in the data** → No causal inference without explicit causal model

- **Structural causal model:** principled framework to use field knowledge to:

  * use analytical solutions (do-calculus and backdoor criterion) to define sets of necessary controls

  * make our scientific assumptions explicit
    → dare **sharing subjectivity** behind model output interpretation
      → foster **cummulative science**
        → tackle **reproducibility crisis**



Recommended starting points

Arif and MacNeil 2022 (https://doi.org/10.1002/ecm.1554) – SCM for observational causal inference in Ecology
Cinelli et al. 2020 (https://doi.org/10.2139/ssrn.3689437) – A Crash Course in Good and Bad Controls
Westreich and Greenland 2013 (https://doi.org/10.1093/aje/kws412) – The Table 2 Fallacy

Richard McElreath's teaching material:
Statistical Rethinking, 2nd Edition, CRC Press 💚
Science before statistics: Causal Inference: https://youtu.be/KNPYUVmY3NM
Statistical Rethinking 2023 – 05 Elemental Confounds: https://www.youtube.com/watch?v=mBEA7PKDmiY
Statistical Rethinking 2023 – 06 Good and Bad Controls: https://www.youtube.com/watch?v=uanZZLlzKHw

Great tool for DAGs, backdoor criterion, etc: https://dagitty.net/dags.html#