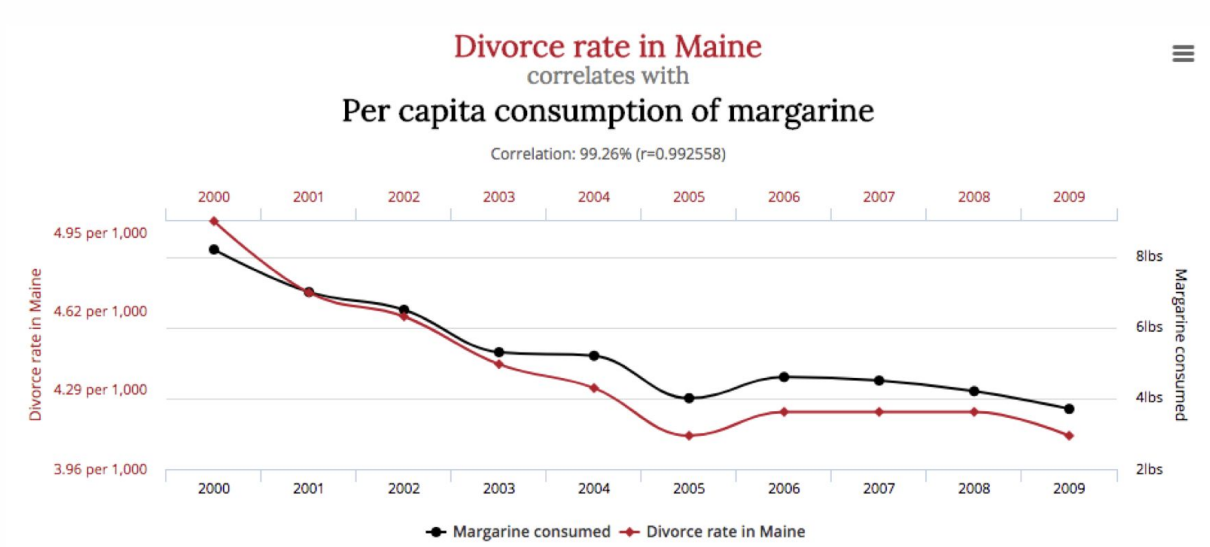
Introduction to **Causal Inference**:
Unveiling the **"Why"** in Healthcare **Data**

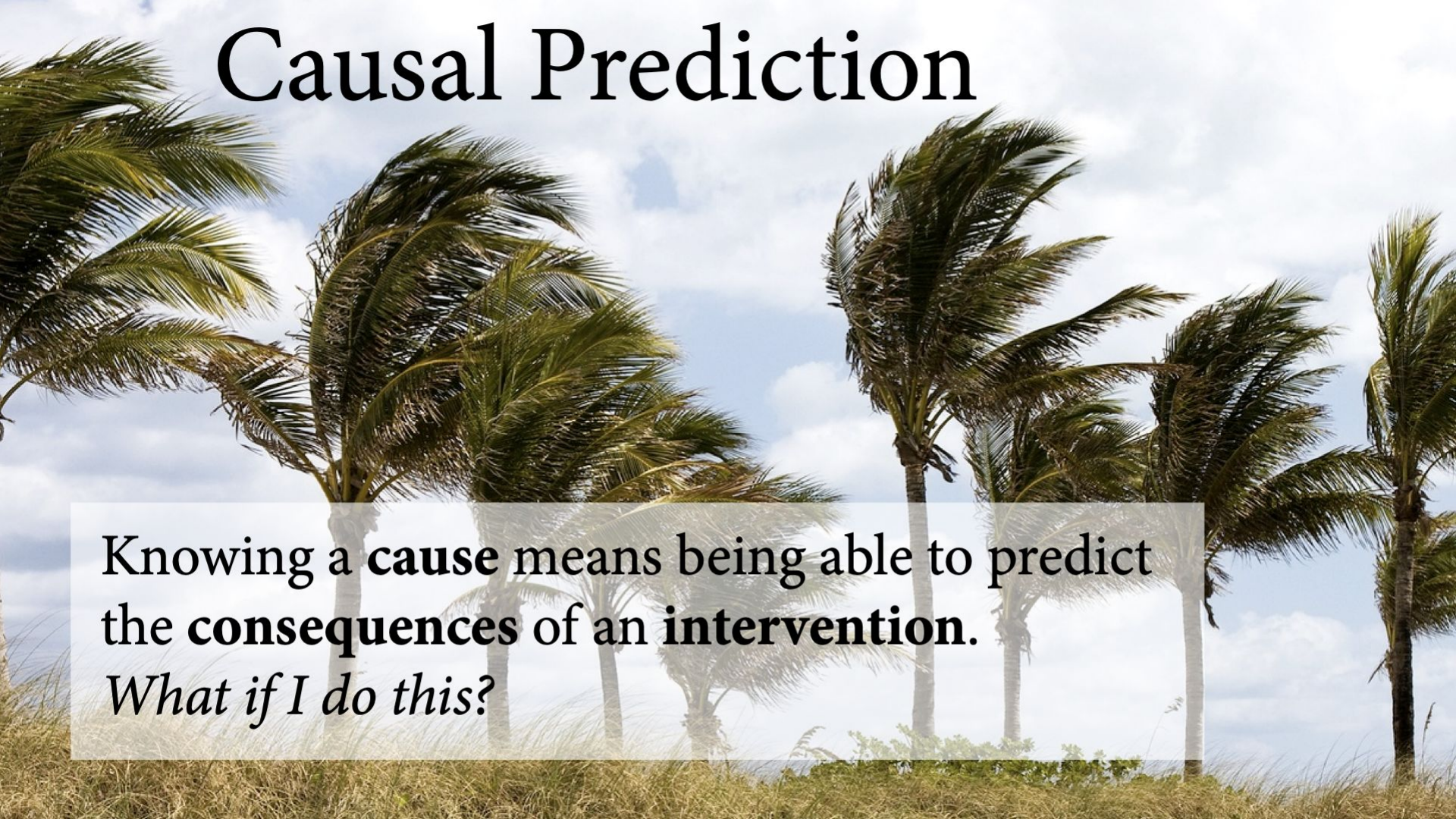
1. What is Causality

Correlation ≠ Causation



https://tylervigen.com/spurious-correlations

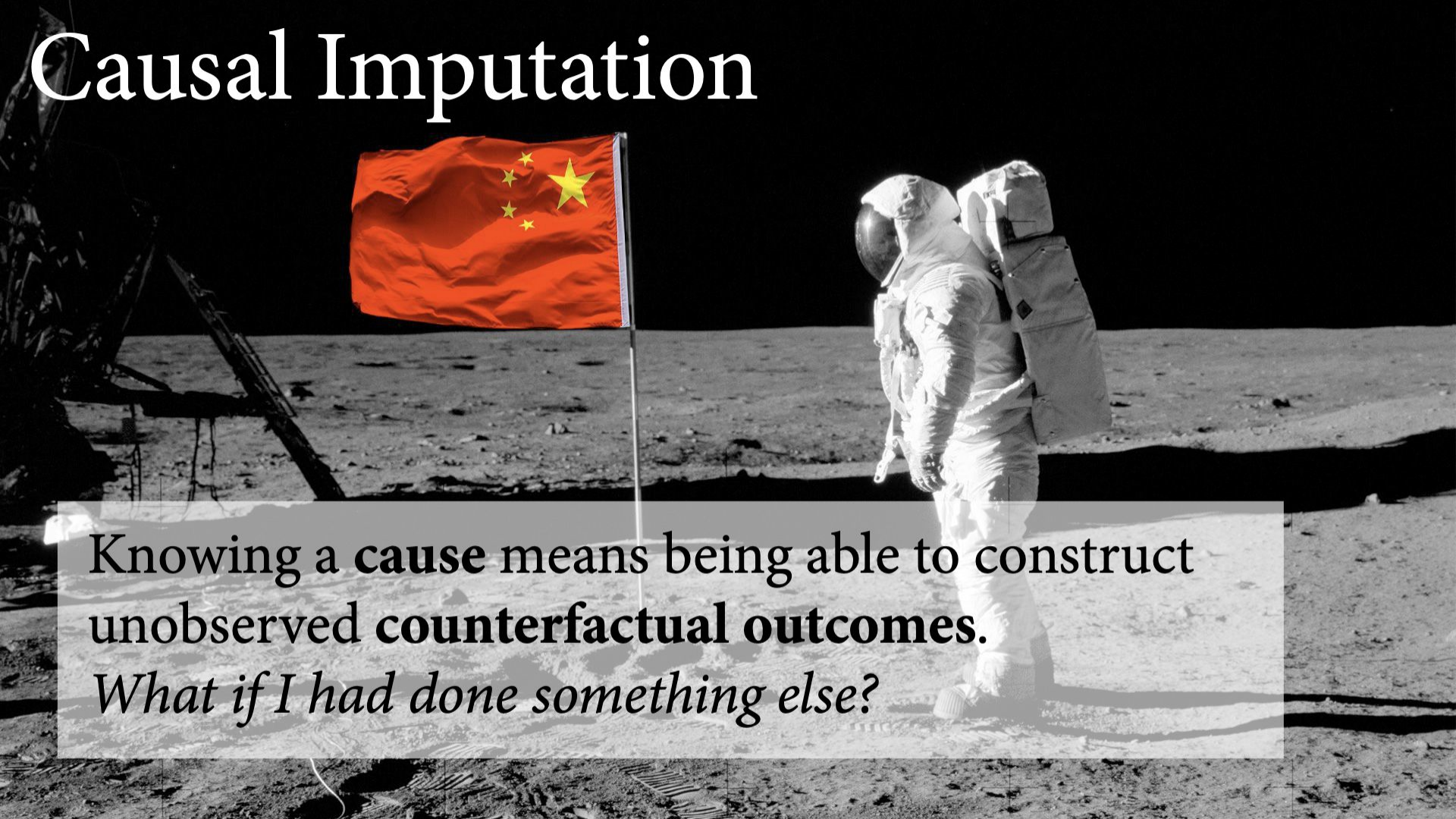
Causal Prediction



Knowing a **cause** means being able to predict the **consequences** of an **intervention**.

What if I do this?

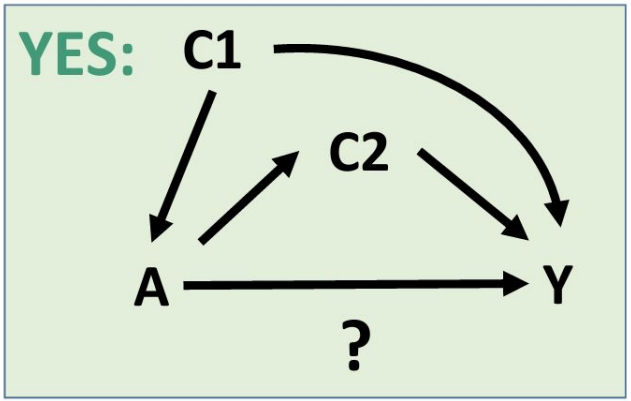
Causal Imputation

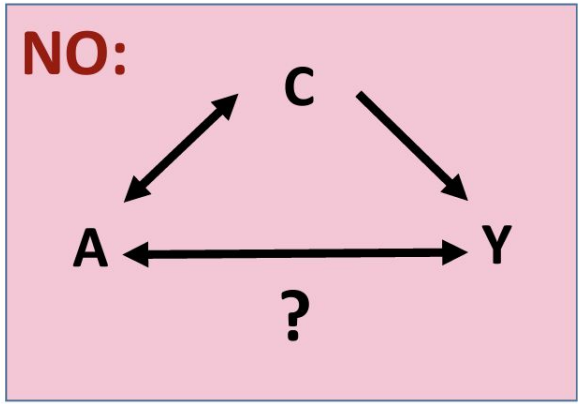


Knowing a **cause** means being able to construct unobserved **counterfactual outcomes**.
What if I had done something else?

Representing Causality

- Directed: point from cause to effect
  - Causal effects cannot be bidirectional
- Acyclic: no directed path can form a closed loop



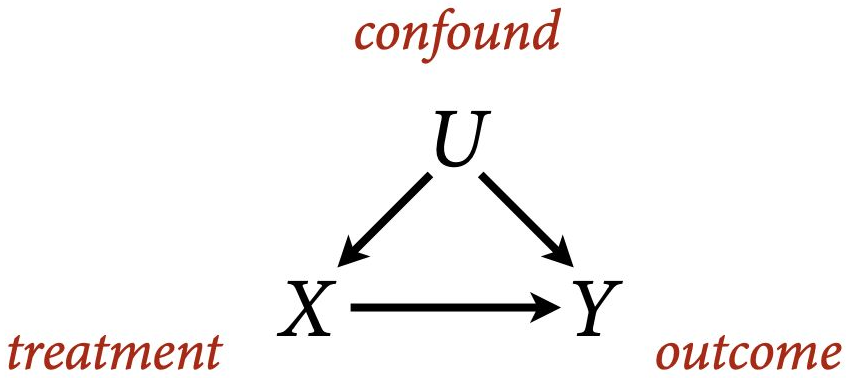


S Talebi, Answering causal questions, Towards Data Science, https://towardsdatascience.com/causal-inference-962ae97cefda
Pearl J, The book of why (2018)

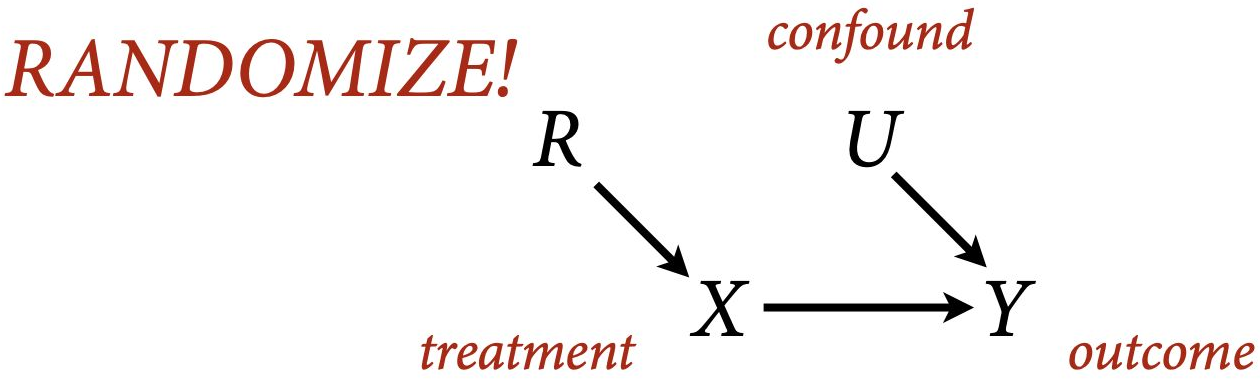
~~Symmetry~~

S Talebi, An introduction to Causal inference, Towards Data Science, https://towardsdatascience.com/causality-an-introduction-f8a3f6ac4c4a

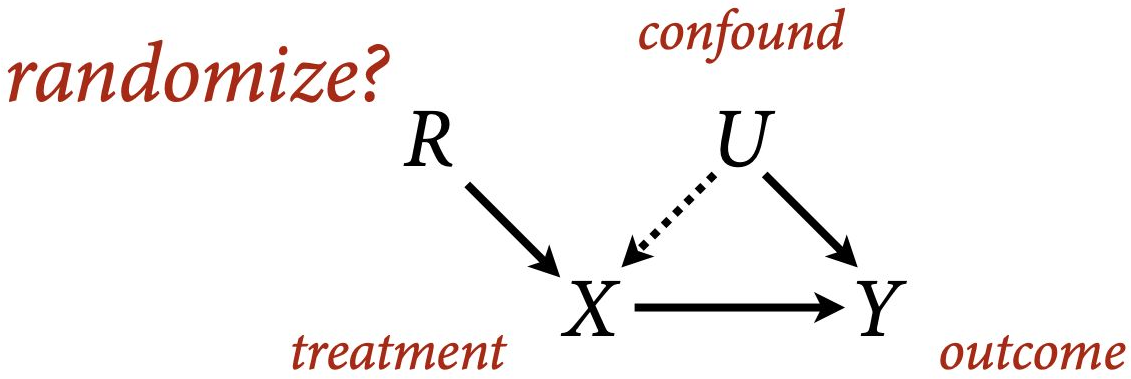
Representing Causality



Representing Causality

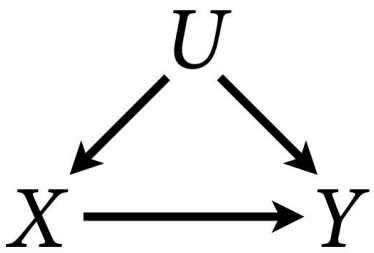


Representing Causality

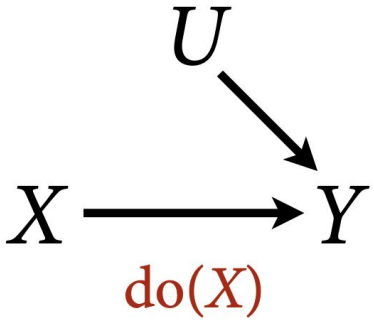


Investigating Causality

Without randomization



With randomization



In an experiment, we cut causes of the treatment
We ***randomize*** (or **mimic** at least)
P(Y|do(X)) = P(Y|?)

do-calculus

For DAGs, **rules for finding P(Y|do(X))**

known as do-calculus

do-calculus says **what is possible** to say

before picking functions

Justifies **graphical analysis**

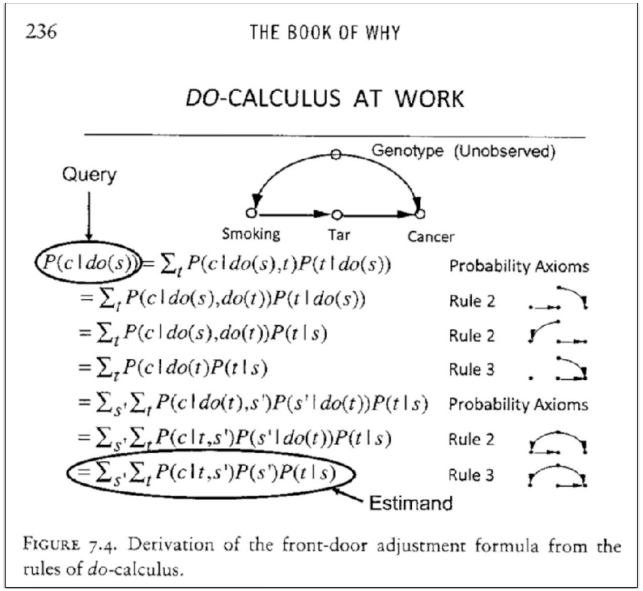
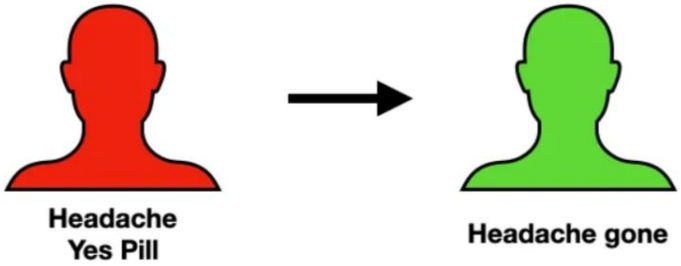


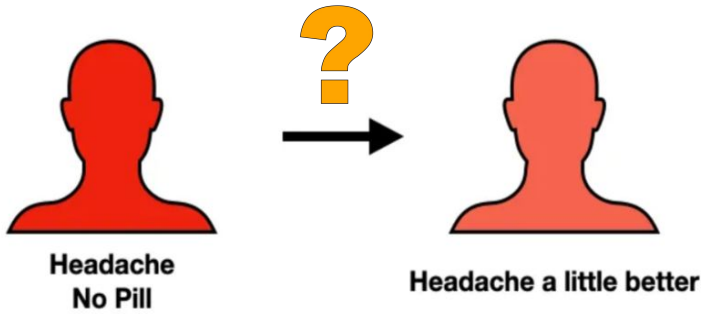
FIGURE 7.4. Derivation of the front-door adjustment formula from the rules of *do*-calculus.

Outcomes Framework



Scenario 1- Reality

What actually happened



Scenario 2- Counterfactual

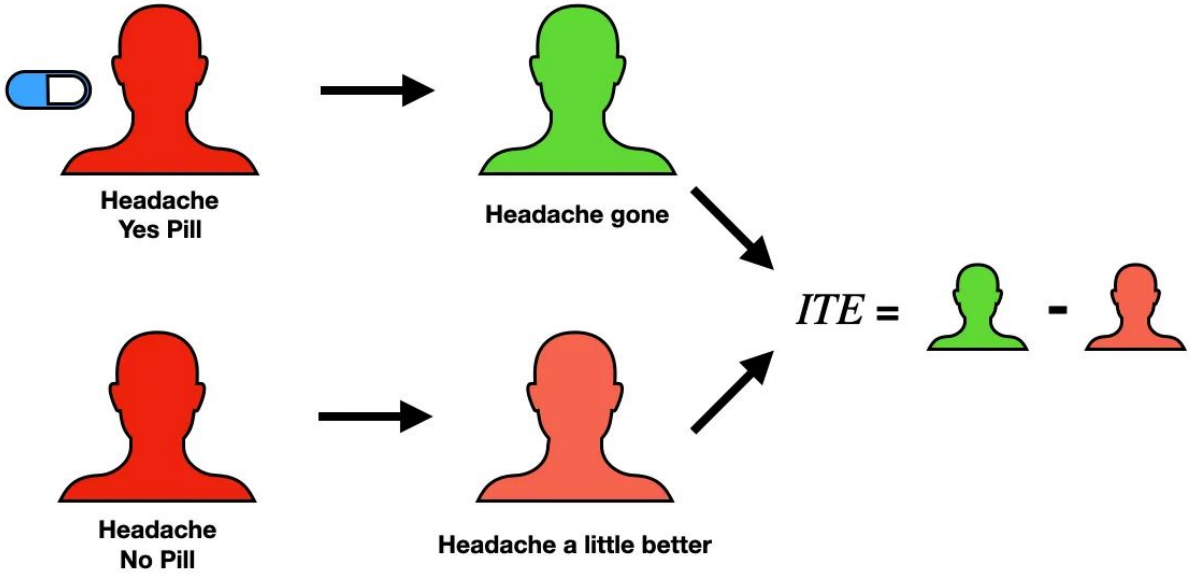
What if this hadn't have happened?

S Talebi, Causal Discovery, Towards Data Science, https://towardsdatascience.com/causal-discovery-6858f9af6dcb

Treatment Effects

Individual Treatment Effect (ITE)

Impact of Treatment for a particular individual



S Talebi, Causal Discovery, Towards Data Science, https://towardsdatascience.com/causal-discovery-6858f9af6dcb

Treatment Effects

Individual Treatment Effect (ITE)

Impact of Treatment for a particular individual

Average Treatment Effect (ATE)

Expected impact of treatment for a population

What if we look at the average of each group?

subject	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
Joe	?	5	?
Mary	-10	?	?
Sally	?	10	?
Bob	-20	?	?
Mean	-15	7.5	?

Average treatment effect: $\bar{Y}(1) - \bar{Y}(0) = -22.5$

P English, Causal inference primer, (2016), https://github.com/paul-english/causal-inference-notes/blob/master/causal_inference_primer.pdf

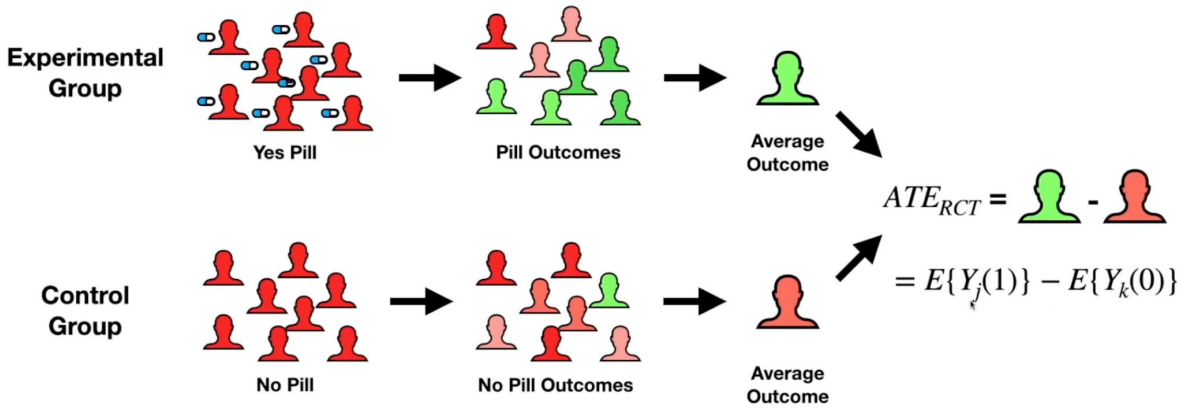
Treatment Effects

Individual Treatment Effect (ITE)

Impact of Treatment for a particular individual

Average Treatment Effect (ATE)

Expected impact of treatment for a population



P English, Causal inference primer, (2016), https://github.com/paul-english/causal-inference-notes/blob/master/causal_inference_primer.pdf

Justifying Controls

$$Y \sim X$$

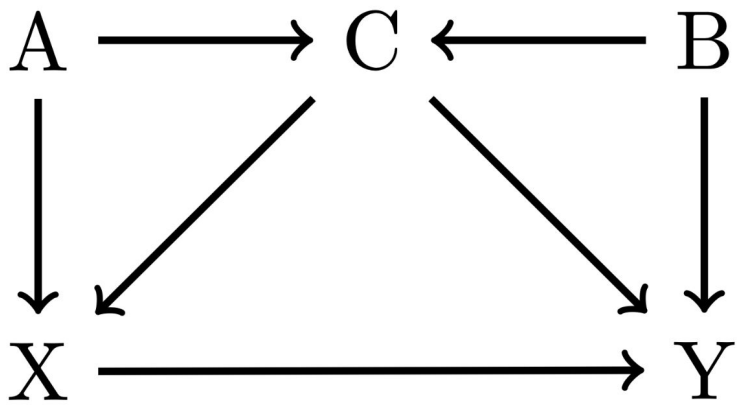
$$Y \sim X + A$$

$$Y \sim X + A + B$$

$$Y \sim X + C$$

$$Y \sim X + A + C$$

$$Y \sim X + B + C$$

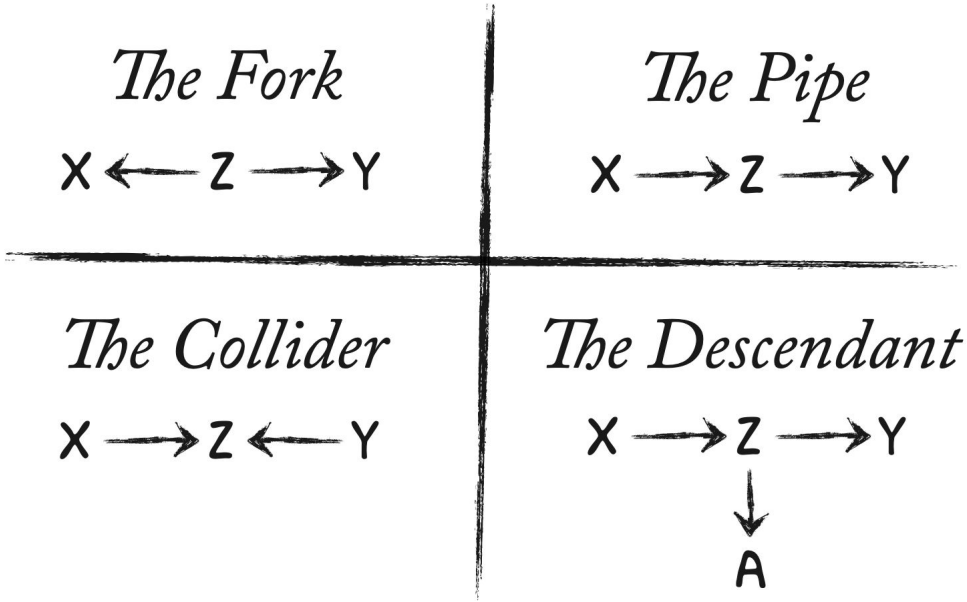


2. Good and Bad Controls

Good and Bad Controls

Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

Representing Causality

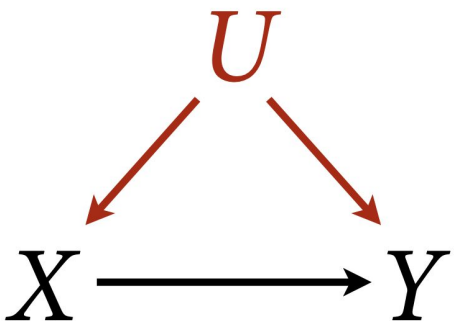


Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

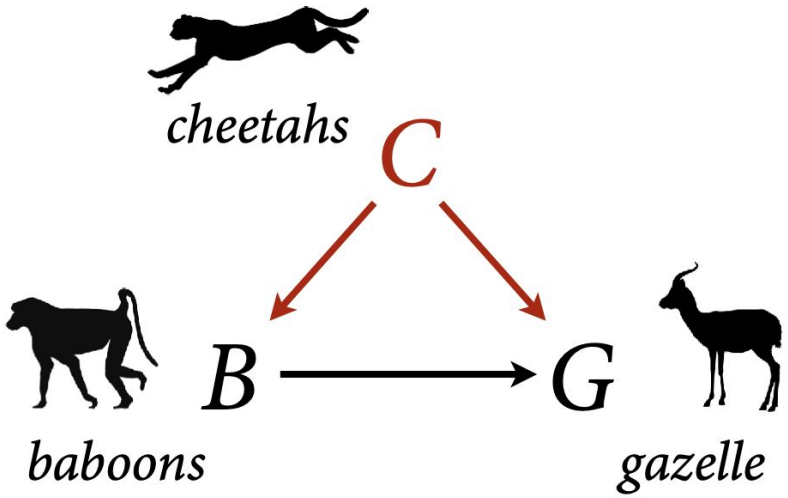
Forks

Causal effect of X on Y is ***not*** (in general) the coefficient relating X to Y

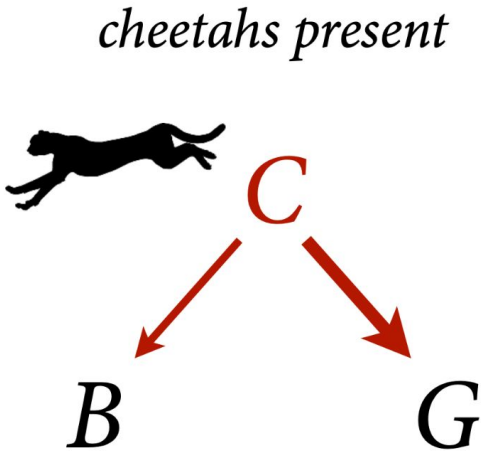
But the **distribution of Y when we change X**, averaged over the ***distributions of the control*** variables (U)



Forks



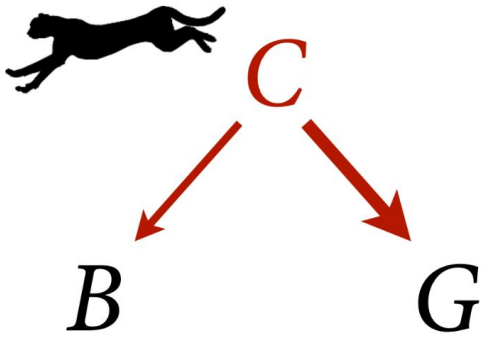
Forks



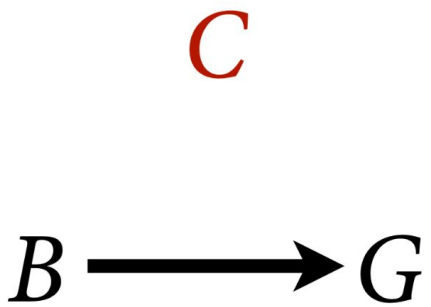
Causal effect of baboons depends on distribution of cheetahs

Forks

cheetahs present



cheetahs absent

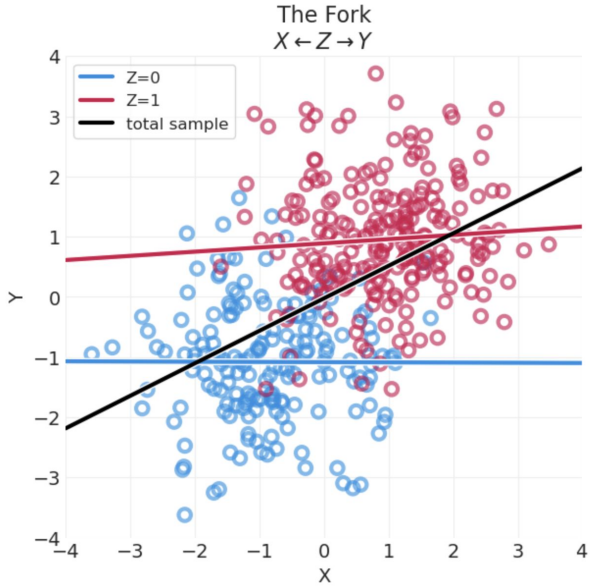


Causal effect of baboons depends on distribution of cheetahs

Forks

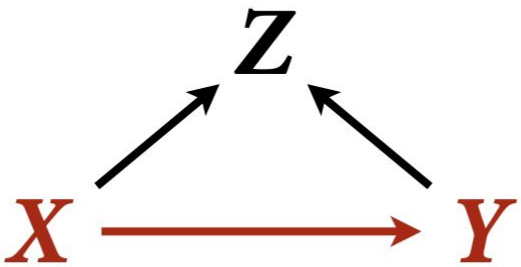
Black line suggests that knowing X tells you something about Y

But when we **block by Z** then this correlation again ***disappears***



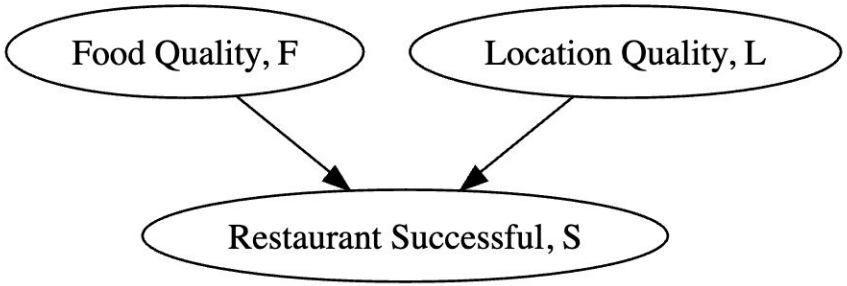
Colliders

Do not touch the collider!



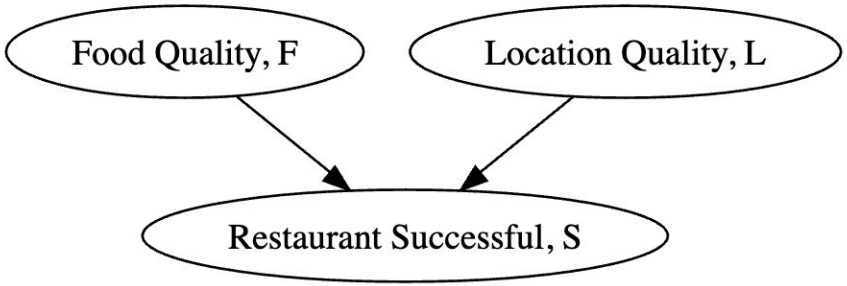
Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

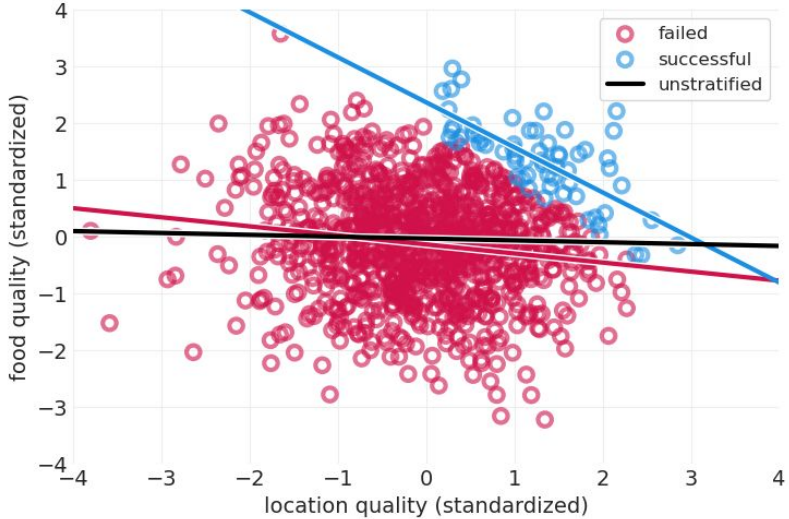
Colliders



Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

Colliders

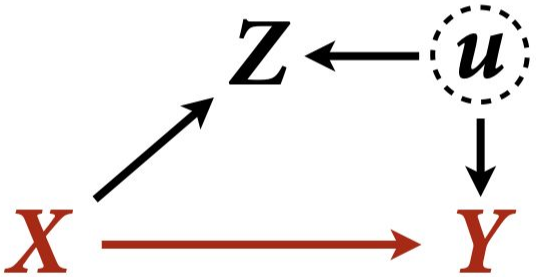




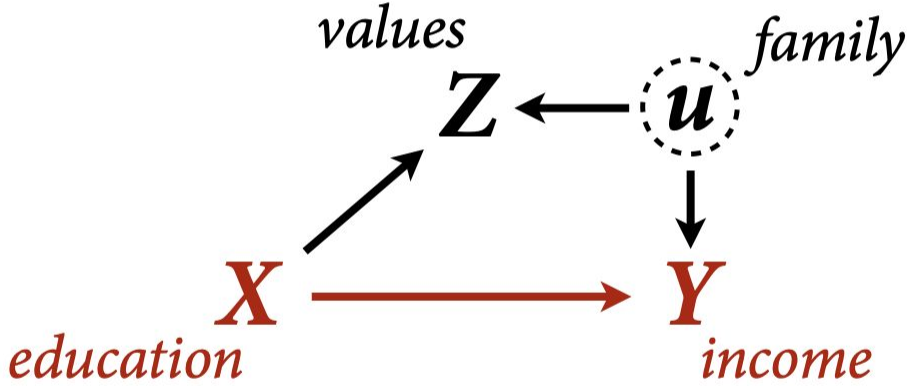
Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

Colliders

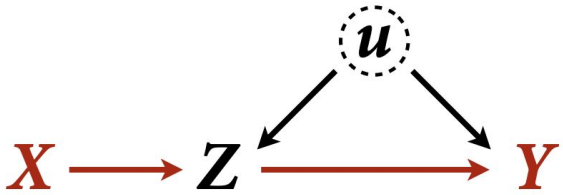
Colliders not always so obvious



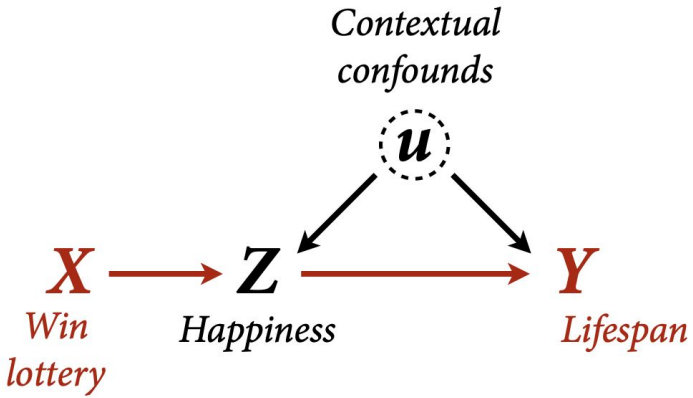
Colliders



Pipes



Pipes

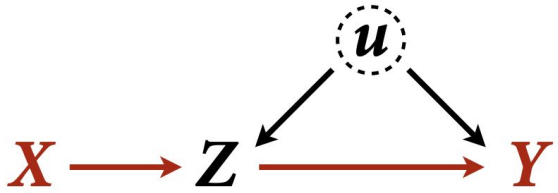

Contextual
confounds
u
X
Win
lottery
Z
Happiness
Y
Lifespan

Pipes

$$X \rightarrow Z \rightarrow Y$$

$$X \rightarrow Z \leftarrow u \rightarrow Y$$

No backdoor, no need to control for Z

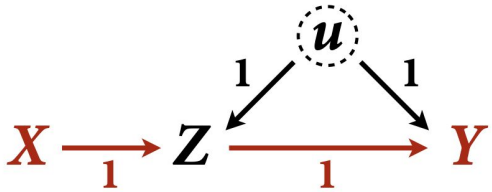


Pipes

```
f <- function(n=100,bXZ=1,bZY=1) {
    X <- rnorm(n)
    u <- rnorm(n)
    Z <- rnorm(n, bXZ*X + u)
    Y <- rnorm(n, bZY*Z + u )
    bX <- coef( lm(Y ~ X) )['X']
    bXZ <- coef( lm(Y ~ X + Z) )['X']
    return( c(bX,bXZ) )
}

sim <- mcreplicate( 1e4 , f() , mc.cores=8 )

dens( sim[1,] , lwd=3 , xlab="posterior mean" )
dens( sim[2,] , lwd=3 , col=2 , add=TRUE )
```

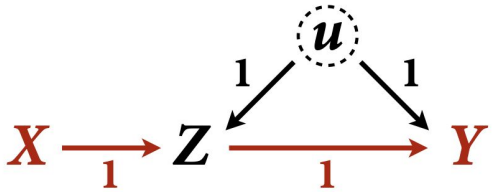


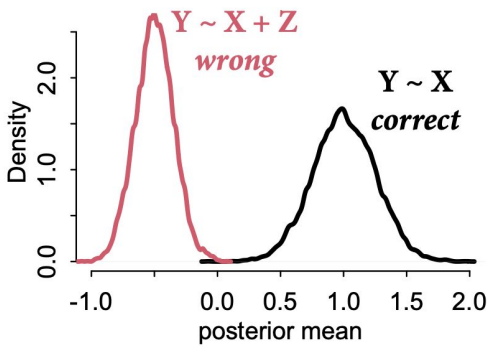
Pipes

```
f <- function(n=100,bXZ=1,bZY=1) {
    X <- rnorm(n)
    u <- rnorm(n)
    Z <- rnorm(n, bXZ*X + u)
    Y <- rnorm(n, bZY*Z + u )
    bX <- coef( lm(Y ~ X) )['X']
    bXZ <- coef( lm(Y ~ X + Z) )['X']
    return( c(bX,bXZ) )
}

sim <- mcreplicate( 1e4 , f() , mc.cores=8 )

dens( sim[1,] , lwd=3 , xlab="posterior mean" )
dens( sim[2,] , lwd=3 , col=2 , add=TRUE )
```





Pipes

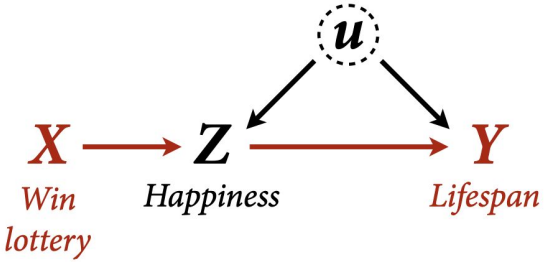
$X \rightarrow Z \rightarrow Y$
$X \rightarrow Z \leftarrow u \rightarrow Y$

No backdoor, no need to control for Z

Controlling for Z biases treatment estimate X

Controlling for Z opens biasing path through u

Can estimate effect of X; Cannot estimate mediation effect Z



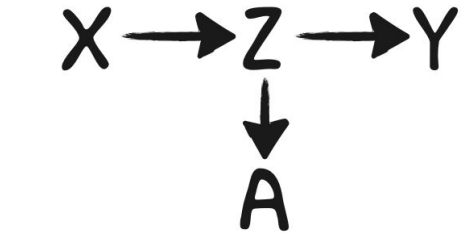
Descendants

X and Y are causally associated through Z

A holds information about Z

Once stratified by A, X and Y **less associated**

Many measurements are **proxies** of what we want to measure



A is a "descendant"

Representing Causality

The Fork

X and *Y* associated
unless stratify by *Z*

The Pipe

X and *Y* associated
unless stratify by *Z*

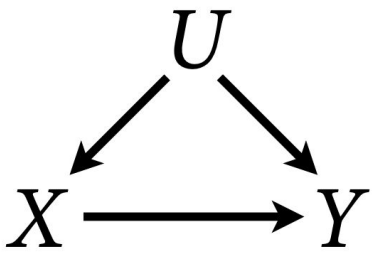
The Collider

X and *Y* not associated
unless stratify by *Z*

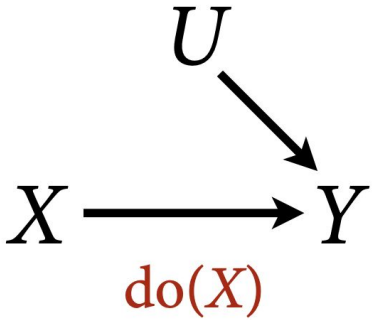
3. Investigating Causality

Investigating Causality

Without randomization

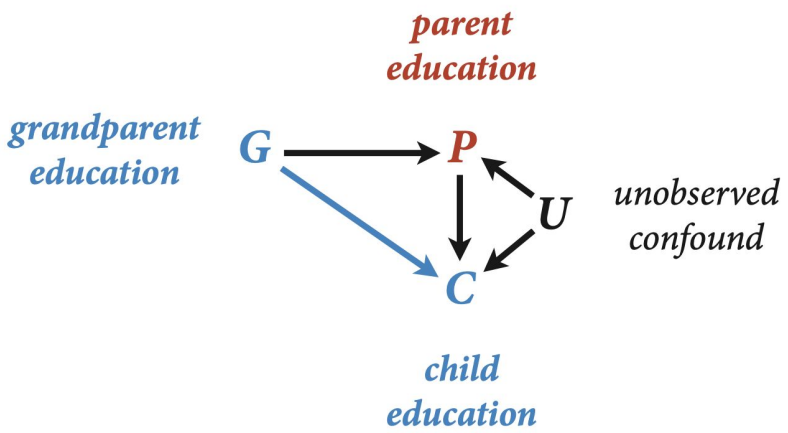


With randomization



In an experiment, we cut causes of the treatment
We ***randomize*** (or **mimic** at least)
P(Y|do(X)) = P(Y|?)

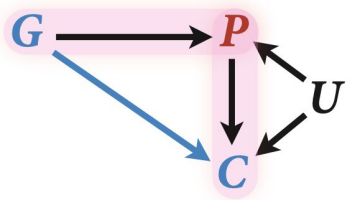
Identifiability



Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

Identifiability

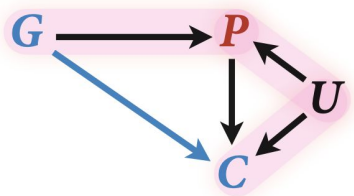
Pipe: G -> P -> C



P is a mediator

Identifiability

Pipe: G -> P -> C

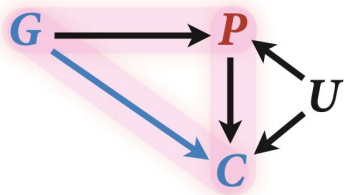


P is a collider

Fork: C <- U -> P

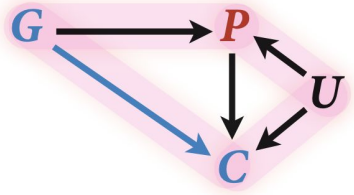
Identifiability

Can estimate **total** effect of G on C



$$C_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_G G_i$$

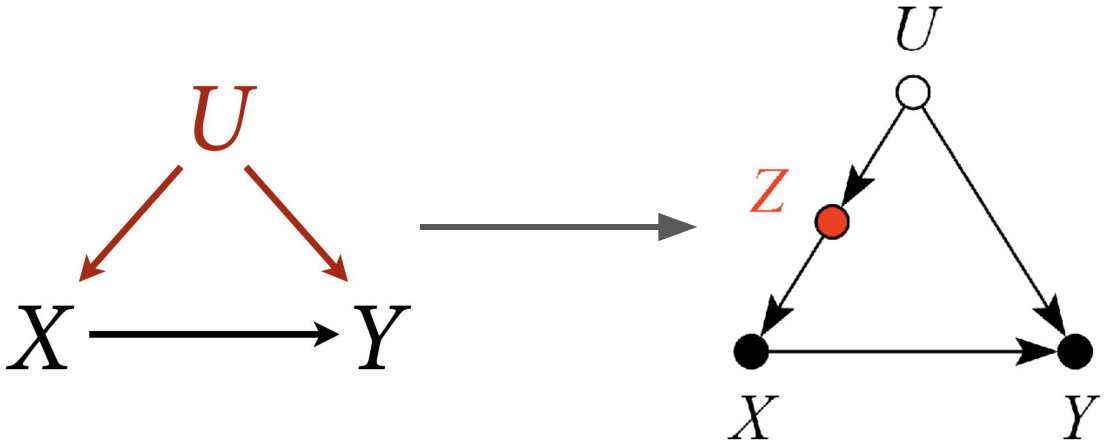
Cannot estimate **direct** effect



$$C_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_G G_i + \beta_P P_i$$

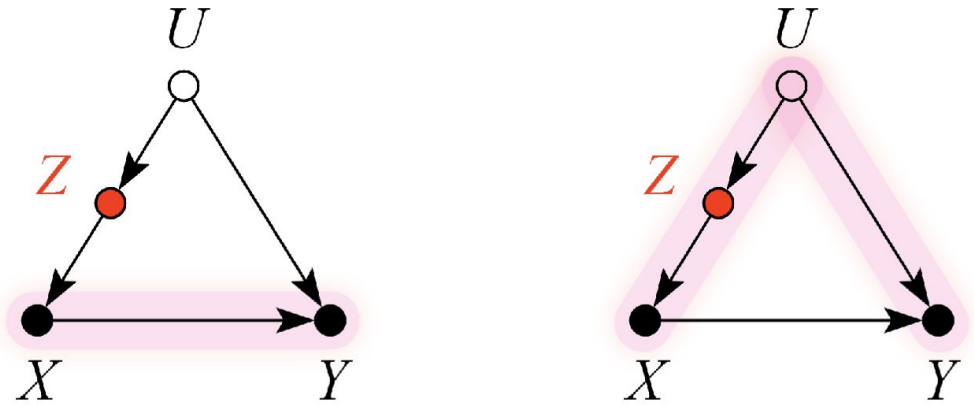
Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

Backdoor Criterion



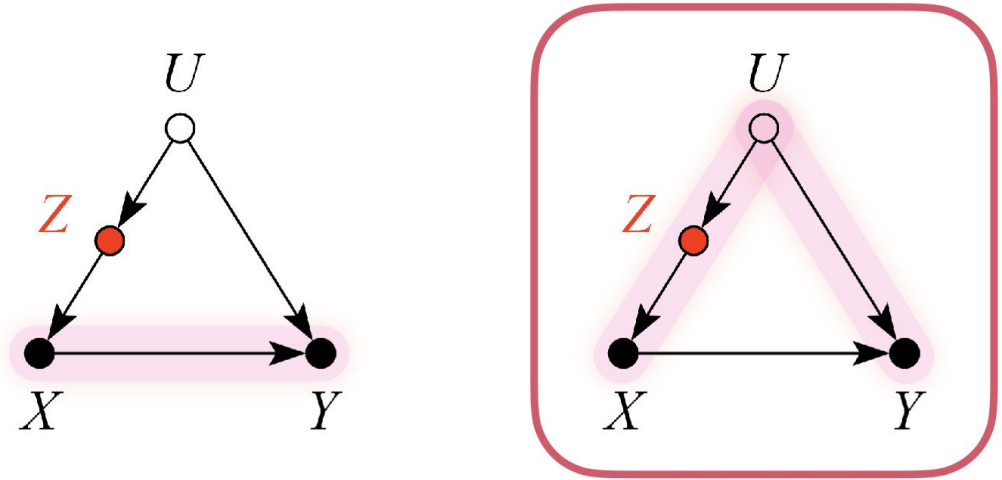
Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

(1) Identify all paths connecting treatment (X) to outcome (Y)



Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

(2) Paths with arrows entering X are backdoor paths (confounding paths)

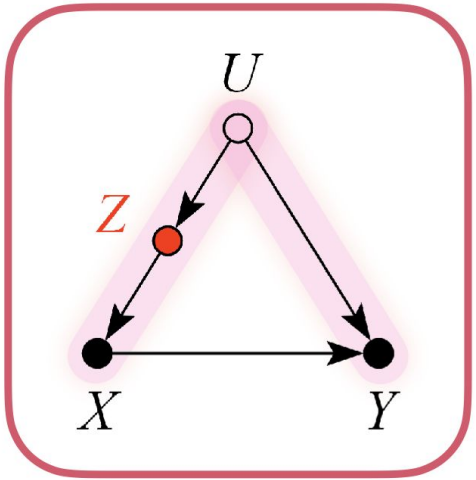


Richard McElreath, Statistical rethinking series, https://github.com/rmcelreath/stat_rethinking_2022#statistical-rethinking-2022-edition

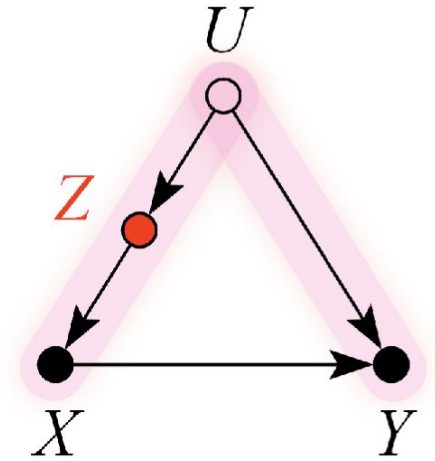
(3) Find a set of control variables that close/block all backdoor paths

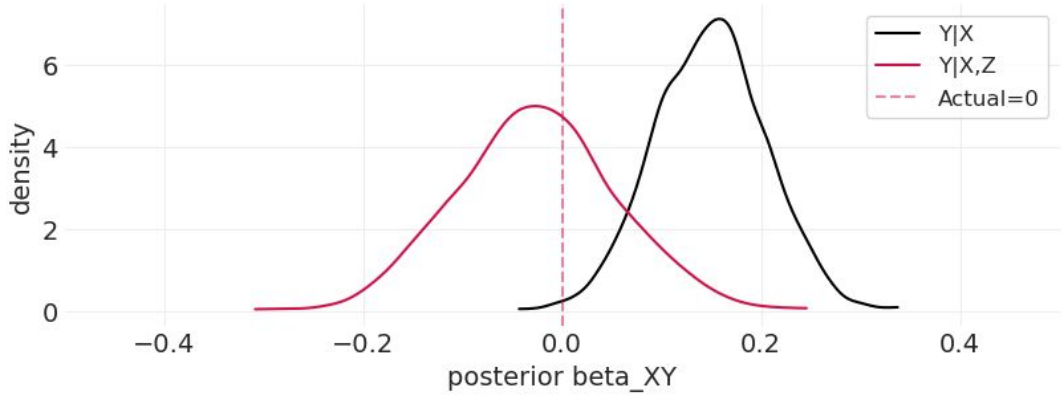
Block the pipe: X ⫫ U | Z

Z "knows" all of the association between X,Y that is due to U

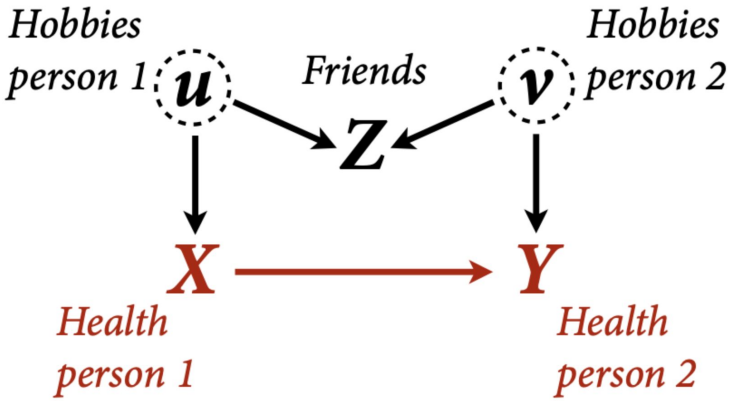


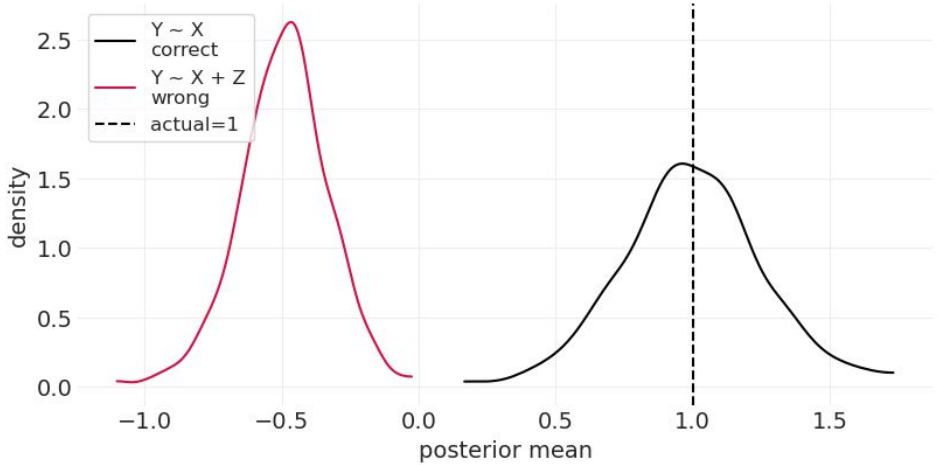
Good Controls



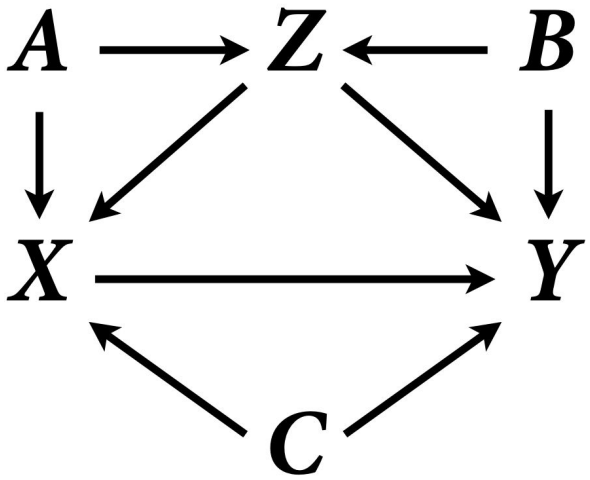


Bad Controls



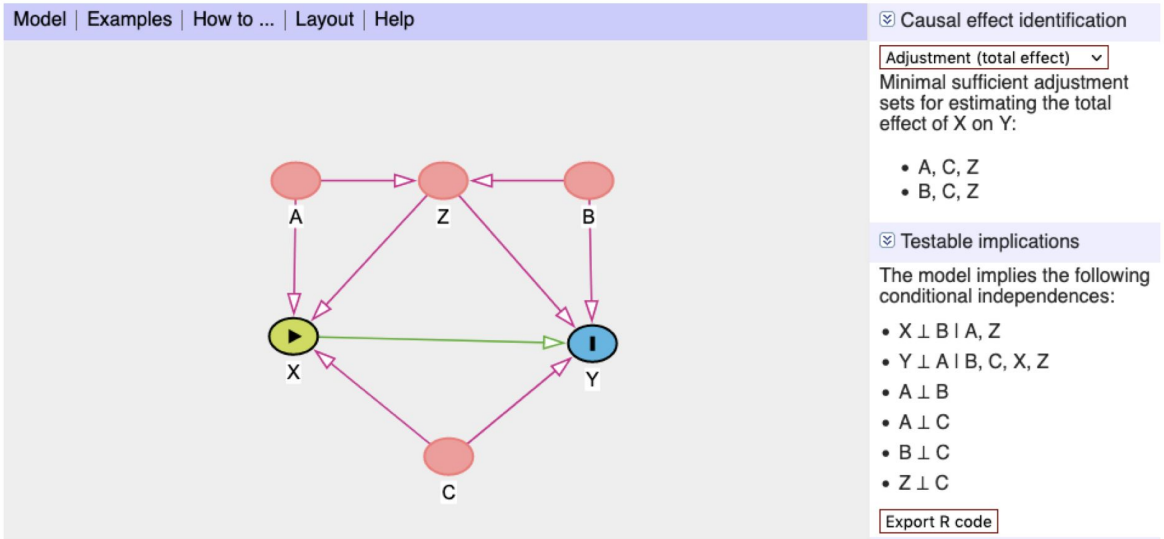


Complex Example



Minimum adjustment set

www.dagitty.net



4. Practical

Notebooks

MIT Intro to Causal Inference 2023

Credit

- The best resource for getting started is Richard McElreath's outstanding lecture series Statistical Rethinking 2023.
- The full notebooks have been converted to Python/PyMC 5 by Dustin Stanbury and are available here
- This repository is a consolidation of the notebooks from the above two sources, with some additional notes and exercises.
- The slides for the accompanying lectures are available here

Walkthrough:
https://github.com/Gallifantjack/mit_causal_inference_intro/blob/dev/causal_inference_workshop.ipynb

Real World Problems #1

Article | Open Access | Published: 12 November 2020

Collider bias undermines our understanding of COVID-19 disease risk and severity

Gareth J. Griffith, Tim T. Morris, Matthew J. Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C. Sharp, Jonathan Sterne, Tom M. Palmer, George Davey Smith, Kate Tilling, Luisa Zuccolo, Neil M. Davies & Gibran Hemani

Nature Communications **11**, Article number: 5749 (2020) | Cite this article

Real World Problems #1

UK Biobank example of collider bias in Covid-19 test data

About this document

This document forms part of the analysis used in the paper:

Collider bias undermines our understanding of COVID-19 disease risk and severity. Gareth Griffith, Tim T Morris, Matt Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C Sharp, Tom M Palmer, George Davey Smith, Kate Tilling, Luisa Zuccolo, Neil M Davies, Gibran Hemani

It is hosted at https://github.com/MRCIEU/ukbb-covid-collider.

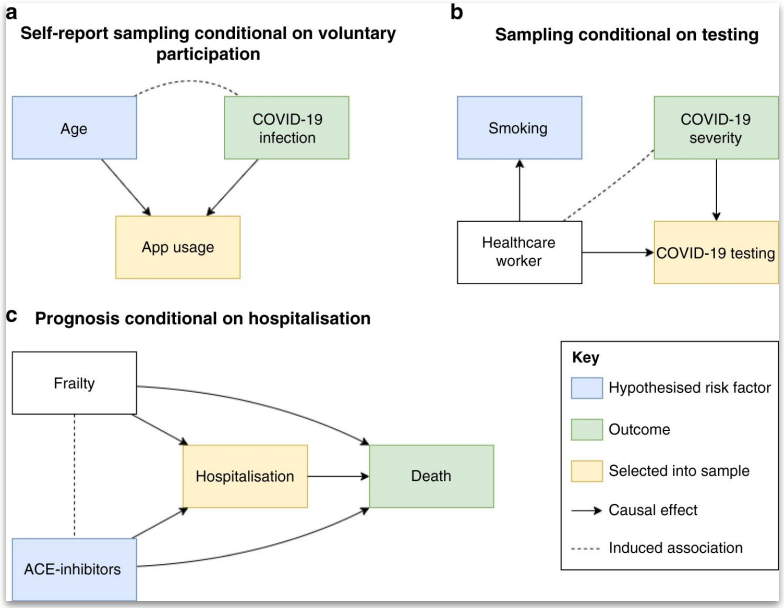
Here we show a set of analyses to illustrate collider bias induced by non-random testing of Covid-19 status amongst the UK Biobank participants, and some approaches to adjust for the bias. The methods are described in further detail in Griffith et al. (2020).

The following variables from the UK biobank phenotype data are used:

- `34-0.0` - Year of birth (converted into age for this analysis)
- `31-0.0` - Sex (male = 1, female = 0)
- `23104-0.0` - Body mass index (BMI)

Also, the linked Covid-19 freeze from `2020-06-05` is used to identify which individuals have been tested and tested positive.

In the analysis that follows, we will be estimating the association between testing positive for Covid-19 and the risk factors age, sex and BMI. The key concern with such an analysis is that we only observe test results among individuals who have received a test. SARS-CoV-2 infection and the risk factors themselves will influence the likelihood of receiving a test, which could induce spurious associations among them when we condition on receiving a test. We will explore inverse probability weighting and sensitivity analyses to address the potential collider bias.



Walkthrough:
https://mrcieu.github.io/ukbb-covid-collider/

Real World Problems #2

CAUSAL THINKING FOR DECISION MAKING ON EHR: WHY AND HOW

Matthieu Doutreligne[1,2,*], Tristan Struja[3,4], Judith Abecassis[1], Claire Morgand[5], Leo Anthony Celi[3,6,7], and Gaël Varoquaux[1]

[1]Inria, Soda, Saclay, France
[2]Mission Data, Haute Autorité de Santé, Saint-Denis, France
[3]Medical University Clinic, Division of Endocrinology, Diabetes & Metabolism, Kantonsspital Aarau, Aarau, Switzerland
[4]Massachusetts Institute of Technology, Institute for Medical Engineering and Science, Cambridge, MA, USA
[5]Agence Régionale de Santé Ile-de-France, France
[6]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
[7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[*]Corresponding author: m.doutreligne@has-sante.fr

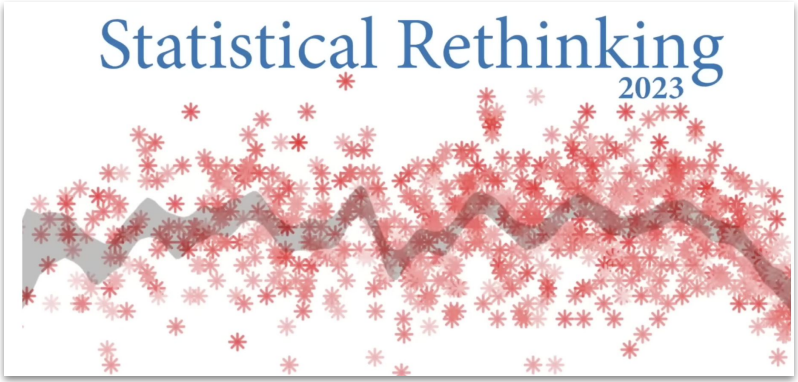
September 8, 2023

Walkthrough:
https://github.com/soda-inria/causal_ehr_mimic/

The less you understand the world, the easier it is to make a decision

Nassim Nicholas Taleb

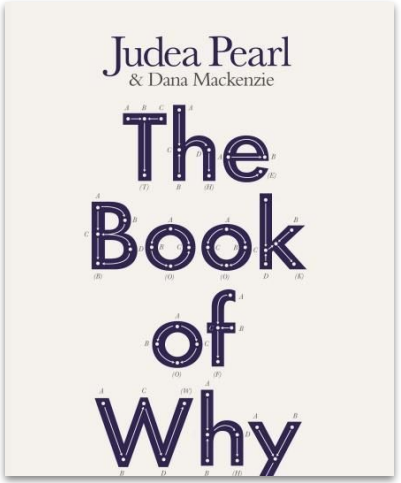
Next Steps



Statistical Rethinking

Richard McElreath

Youtube



The Book of Why

Judea Pearl