C S
C U

An Introduction to Causal Inference

Cornell Statistical Consulting Unit
https://cscu.cornell.edu

Matt Thomas

mthomas@cornell.edu

10/17/2022

C S
C U

Cornell Statistical Consulting Unit

http://www.cscu.cornell.edu/

Webinar guidelines

- Please use the Q & A to ask questions.
- Please do not use the raise your hand option.
- The workshop will be recorded and be available to the Cornell community in about a week
- I'll email slides and examples to everyone here (including R code, but this is not necessary to follow the talk)

Let's Start With Pasta

Enjoy that pasta salad: Noodles linked to lower BMI

By Susan Scutti, CNN
Updated 6:12 AM ET, Tue July 5, 2016

Italian researchers say pasta isn't fattening

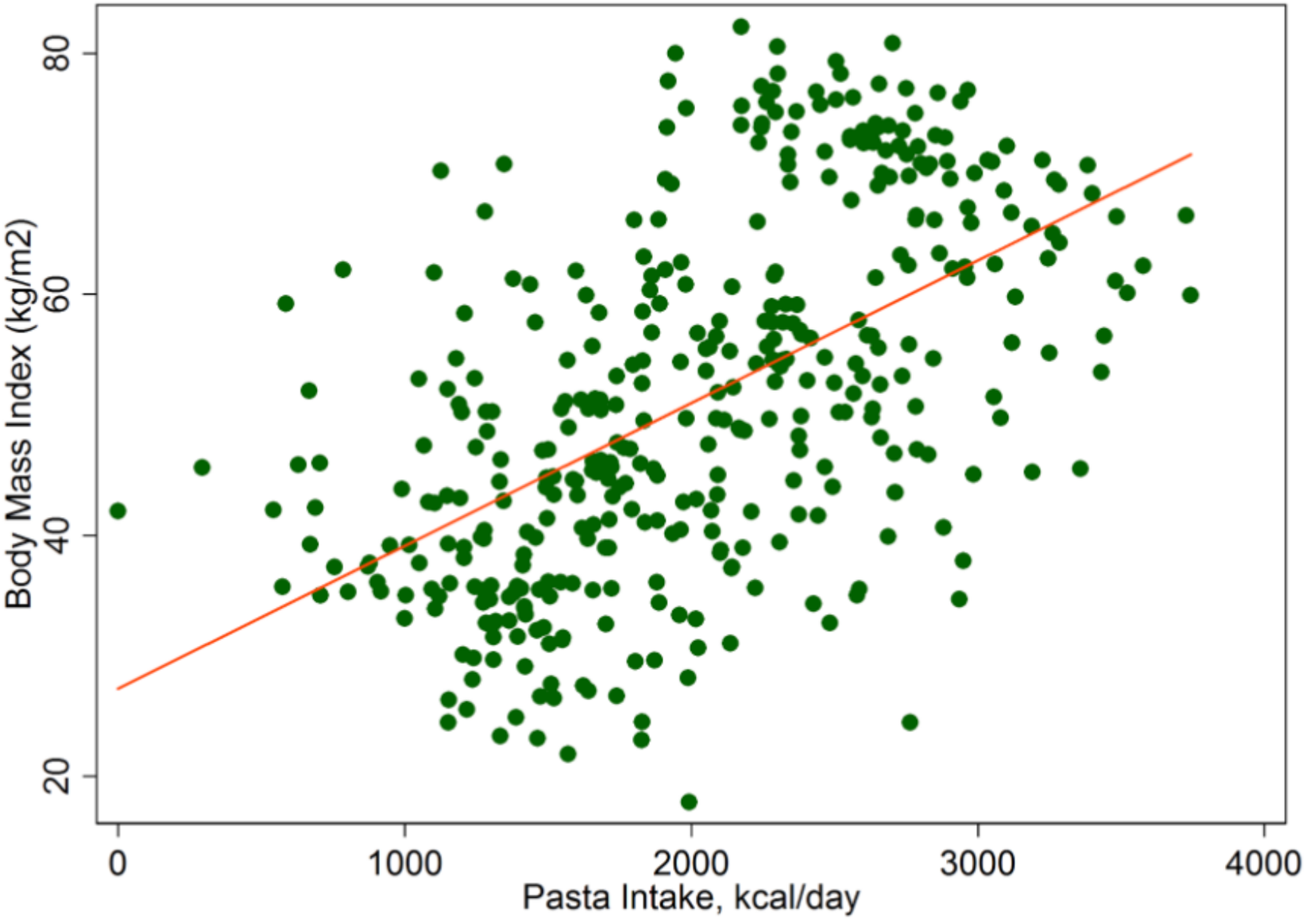
"In popular views, pasta is often considered not adequate when you want to lose weight," said researcher Licia Iacoviello.

By Brooks Hays | July 4, 2016 at 4:40 PM

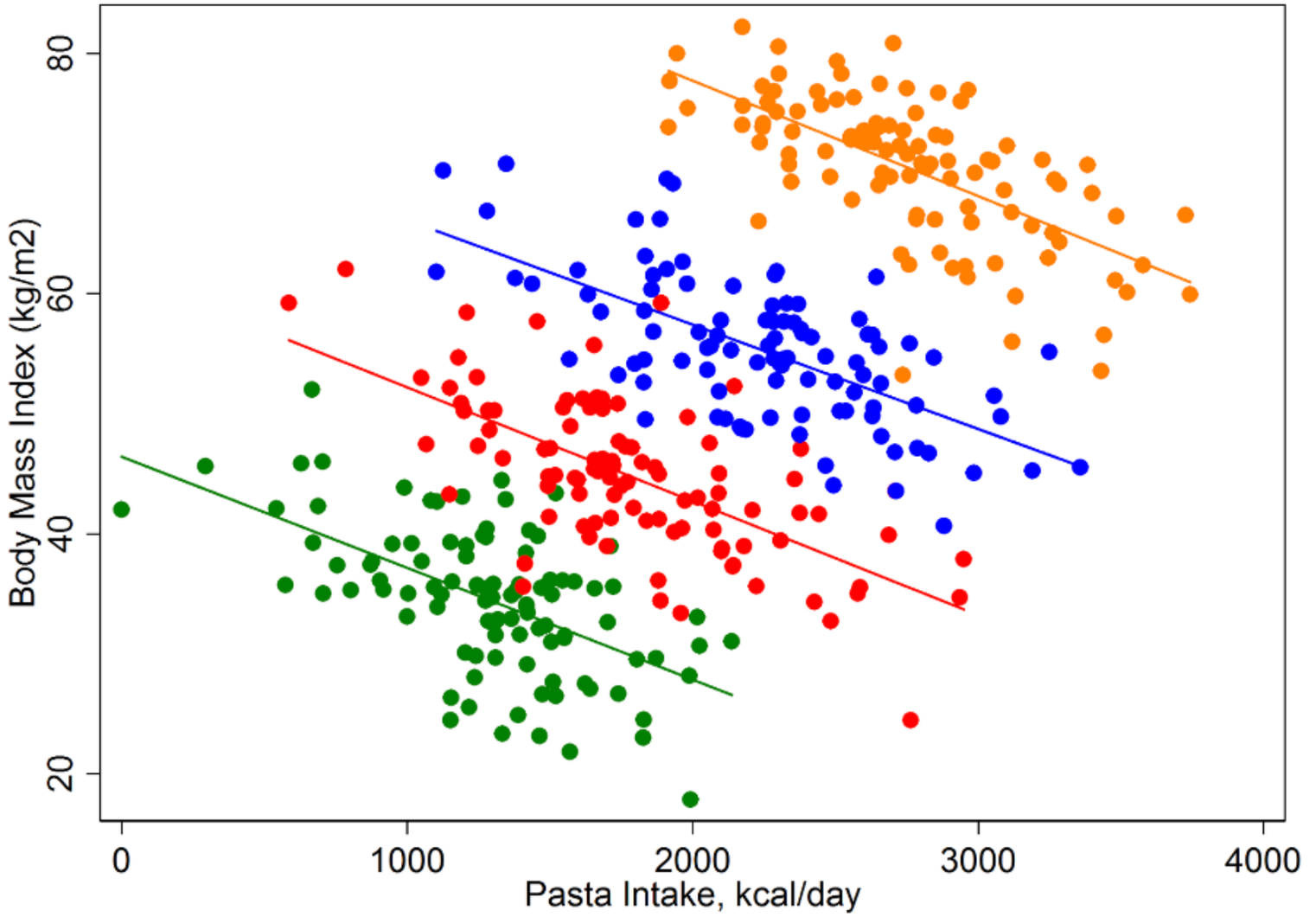
Pasta leads to lower BMI

Moderation is key but feel free to lose the guilty conscience next time you eat Italian

Reid Binion
Published: July 7, 2016, 8:18 pm | Updated: July 7, 2016, 8:23 pm


Body Mass Index (kg/m2)
20
40
60
80
0
1000
2000
3000
4000
Pasta Intake, kcal/day

Weight categories: Simpson's paradox



Taking a step back

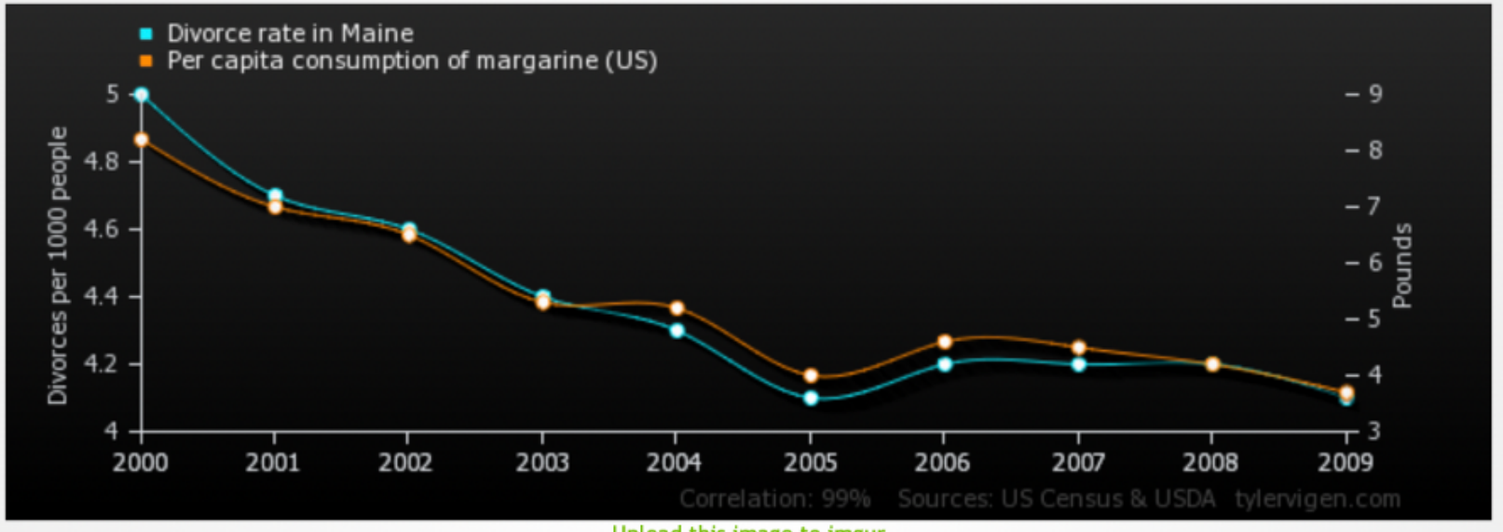
What does it really mean for A to *cause* B?

Some possible interpretations

- When A happens, B happens
- When A happens, B is more/less likely to happen

Sprious correlations

Divorce rate in Maine
correlates with
Per capita consumption of margarine (US)



Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Some possible interpretations

- When A happens, B happens
- When A happens, B is more/less likely to happen
- B relies on A (at least in part) to determine its value

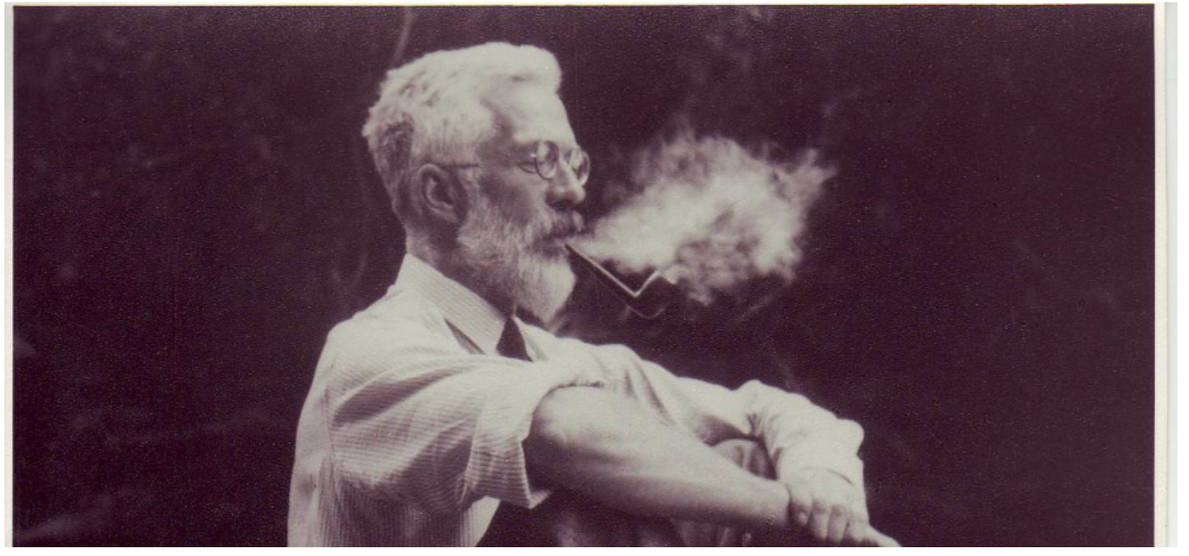
PRICEONOMICS

In Data We Trust CONTENT BOOTCAMP DATA STUDIO TRACKER DATA VISUALIZATION

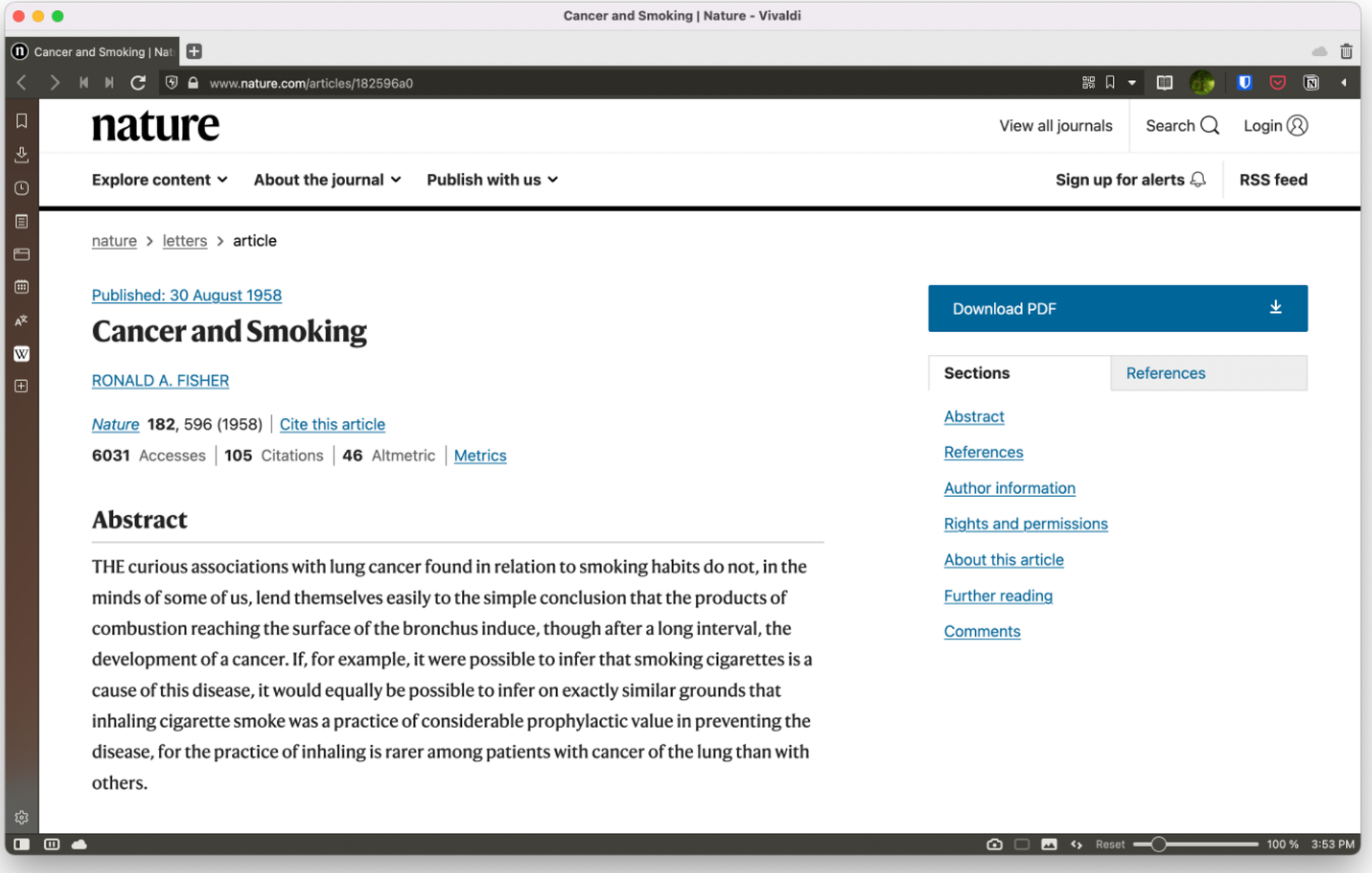
Why the Father of Modern Statistics Didn't Believe Smoking Caused Cancer

By Ben Christopher

Share Tweet



https://www.york.ac.uk/depts/maths/histstat/smoking.htm
https://priceonomics.com/why-the-father-of-modern-statistics-didnt-believe/


Cancer and Smoking | Nature - Vivaldi
Cancer and Smoking | Nat
www.nature.com/articles/182596a0
nature
View all journals
Search
Login
Explore content
About the journal
Publish with us
Sign up for alerts
RSS feed
nature › letters › article
Published: 30 August 1958
Cancer and Smoking
RONALD A. FISHER
Nature 182, 596 (1958) | Cite this article
6031 Accesses | 105 Citations | 46 Altmetric | Metrics
Abstract
THE curious associations with lung cancer found in relation to smoking habits do not, in the minds of some of us, lend themselves easily to the simple conclusion that the products of combustion reaching the surface of the bronchus induce, though after a long interval, the development of a cancer. If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others.
Download PDF
Sections
References
Abstract
References
Author information
Rights and permissions
About this article
Further reading
Comments
Reset
100 %
3:53 PM

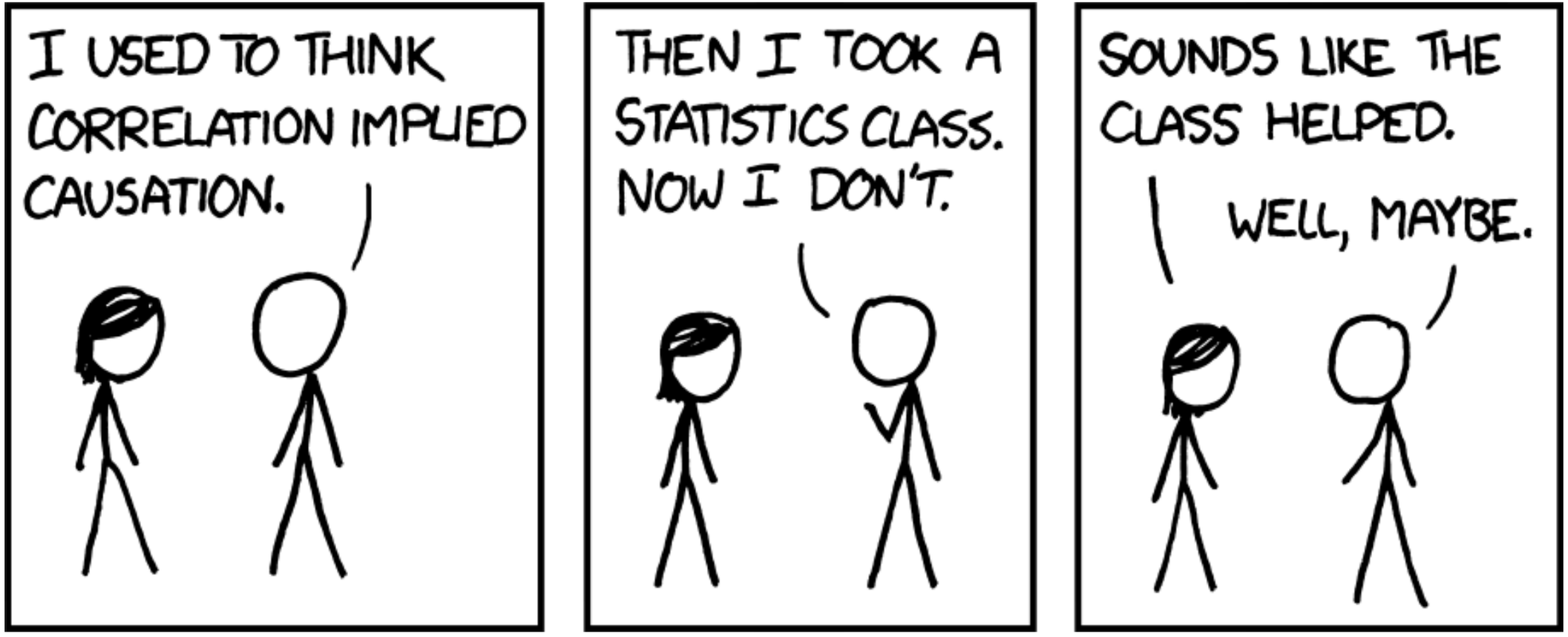
A "Toy" Example

- Suppose you're interested in the effects of taking vitamins on blood pressure
- You collect some (observational) data, and find that people who take vitamins every day have, on average, lower blood pressure
- What's the story here?

Brainstorming

What might prevent us from making causal claims from collected data?

Observational studies vs experimental studies



Experiments

- Why not just run an experiment?
  - Ethics
  - Practicality
  - Cost
  - Sometimes not even possible – e.g. fairness studies, weather studies like forest fires

Hill Criteria[1]

Interventions vs Conditioning

- Intervening means you are (at least hypothetically) intervening, or setting a variable to a value
- Conditioning means we are restricting what we're looking at to a specific group/outcome

Claims and Reality

- In reality, we might want to be able to do an intervention / experiment, but we can't
- We often make watered-down causal claims because of this[1]

Let’s look at actual code examples

Data are houses in Saratoga Springs

If you remember one thing:

Controlling for *everything* without thinking about it is dangerous

(in the previous examples, adding extra predictors improved the model, we'll see examples in a bit where the opposite is true)

Causal Inference as a collection of tools

- *Directed acyclic graphs (DAGs)*
- Do-calculus
- Propensity scores
- Matching

C S
C U

DAGs

Directed

Acyclic

Graphs

Consider a study

Suppose we want to study the relationship between smoking and FEV

What variables might we want to include?

Actual study variables

Table 1. Description of variables in this study.

Variable	Description
AGE	Subject age (years)
FEV	Forced expiratory volume (L)
HEIGHT	Subject height (inches)
SEX	Biological sex: Female (0), Male (1)
SMOKE	Has the subject ever smoked? No (0), Yes (1)

How could we set up a DAG for these? (Take a second to draw one for yourself)

Table 1. Description of variables in this study.

Variable	Description
AGE	Subject age (years)
FEV	Forced expiratory volume (L)
HEIGHT	Subject height (inches)
SEX	Biological sex: Female (0), Male (1)
SMOKE	Has the subject ever smoked? No (0), Yes (1)

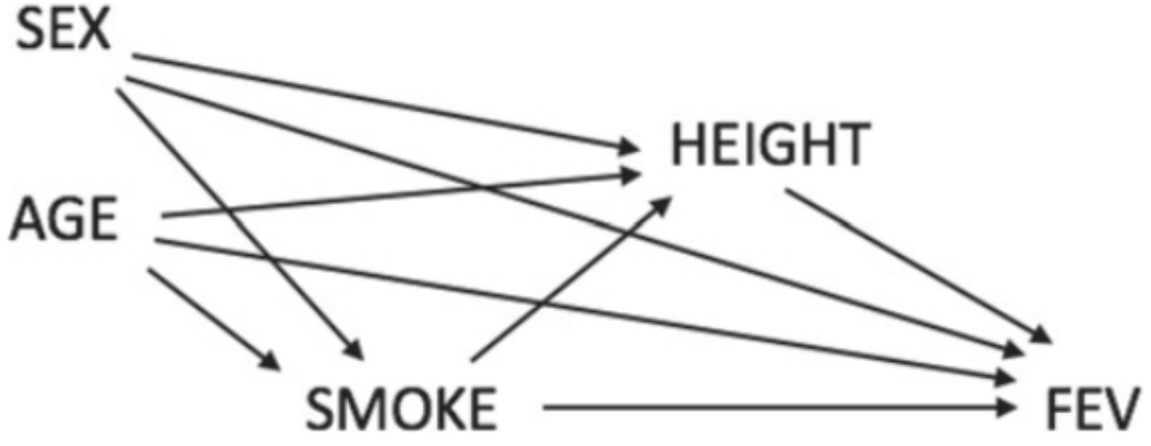
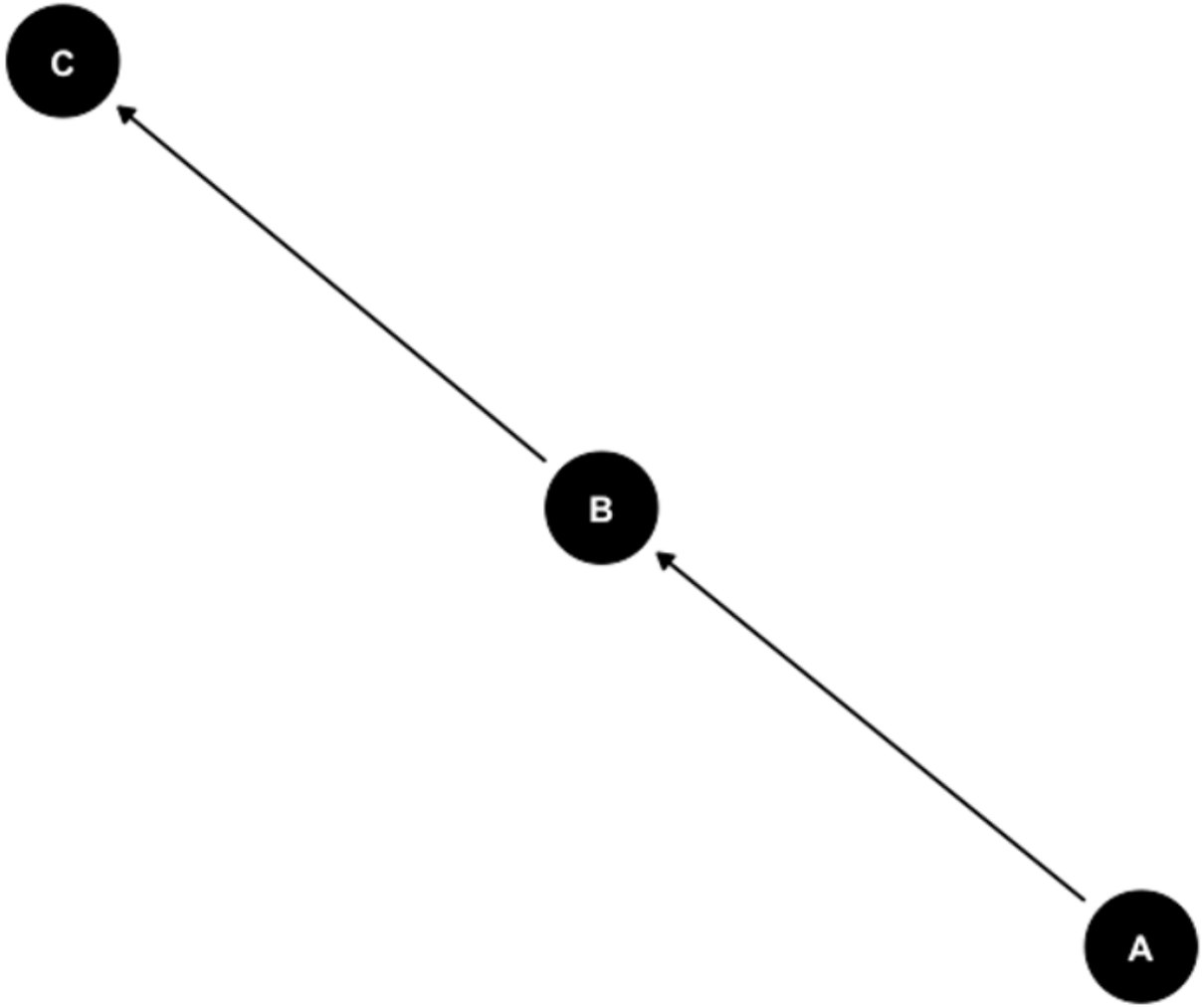


Figure 6. A causal diagram depicting relationships between variables in this study.

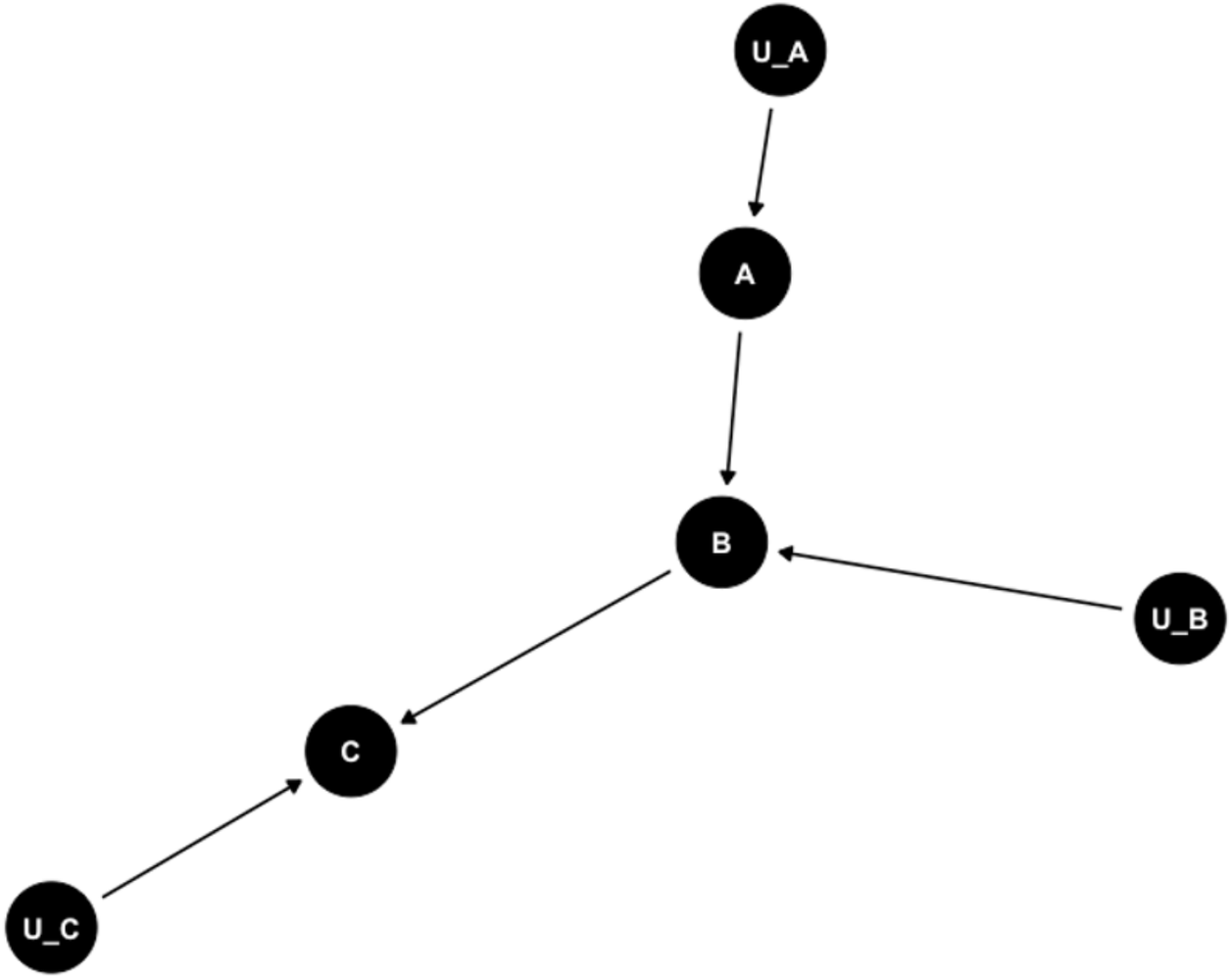
Types of Relations in DAGs

- Chains
- Forks
- Colliders

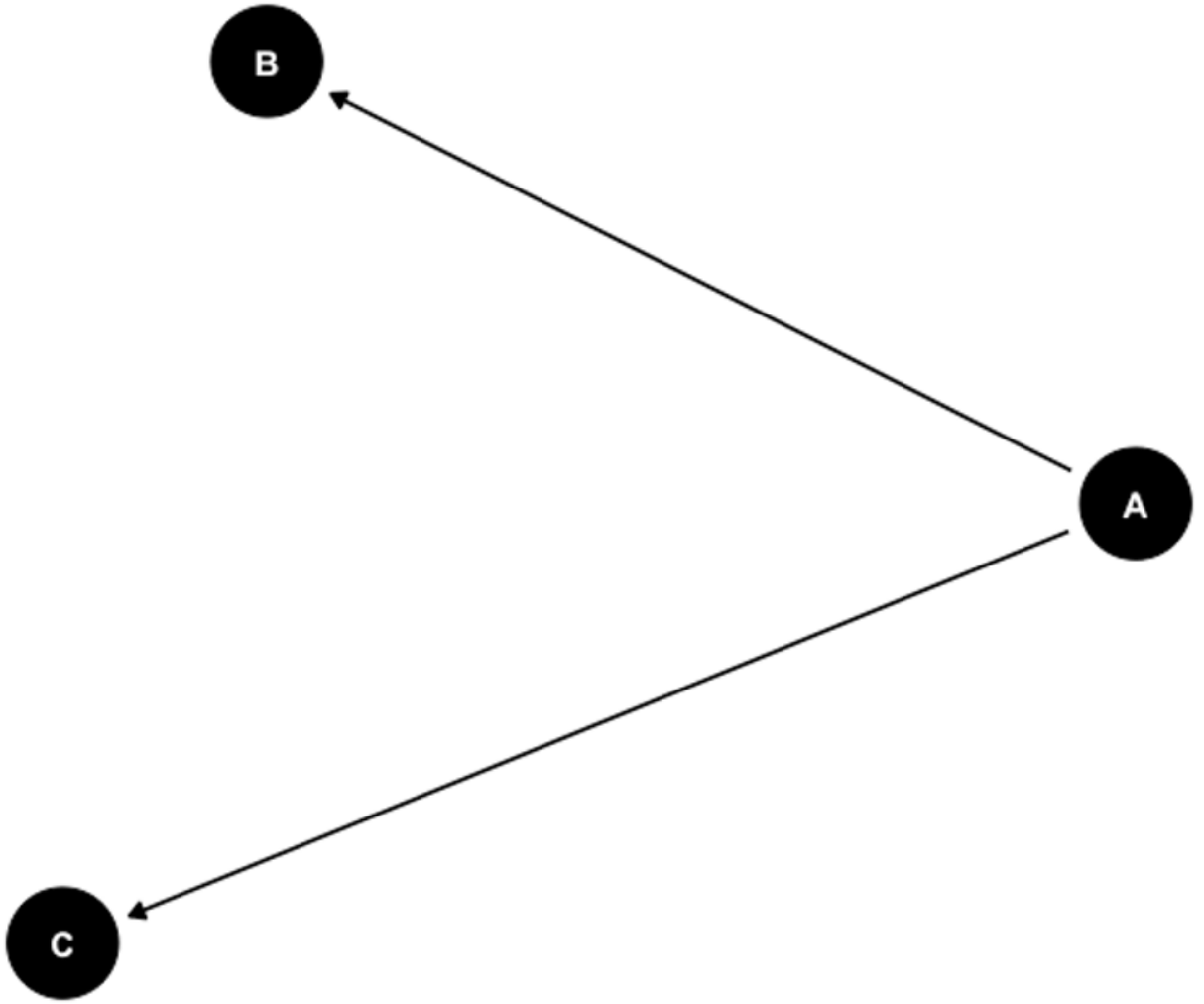
Chains



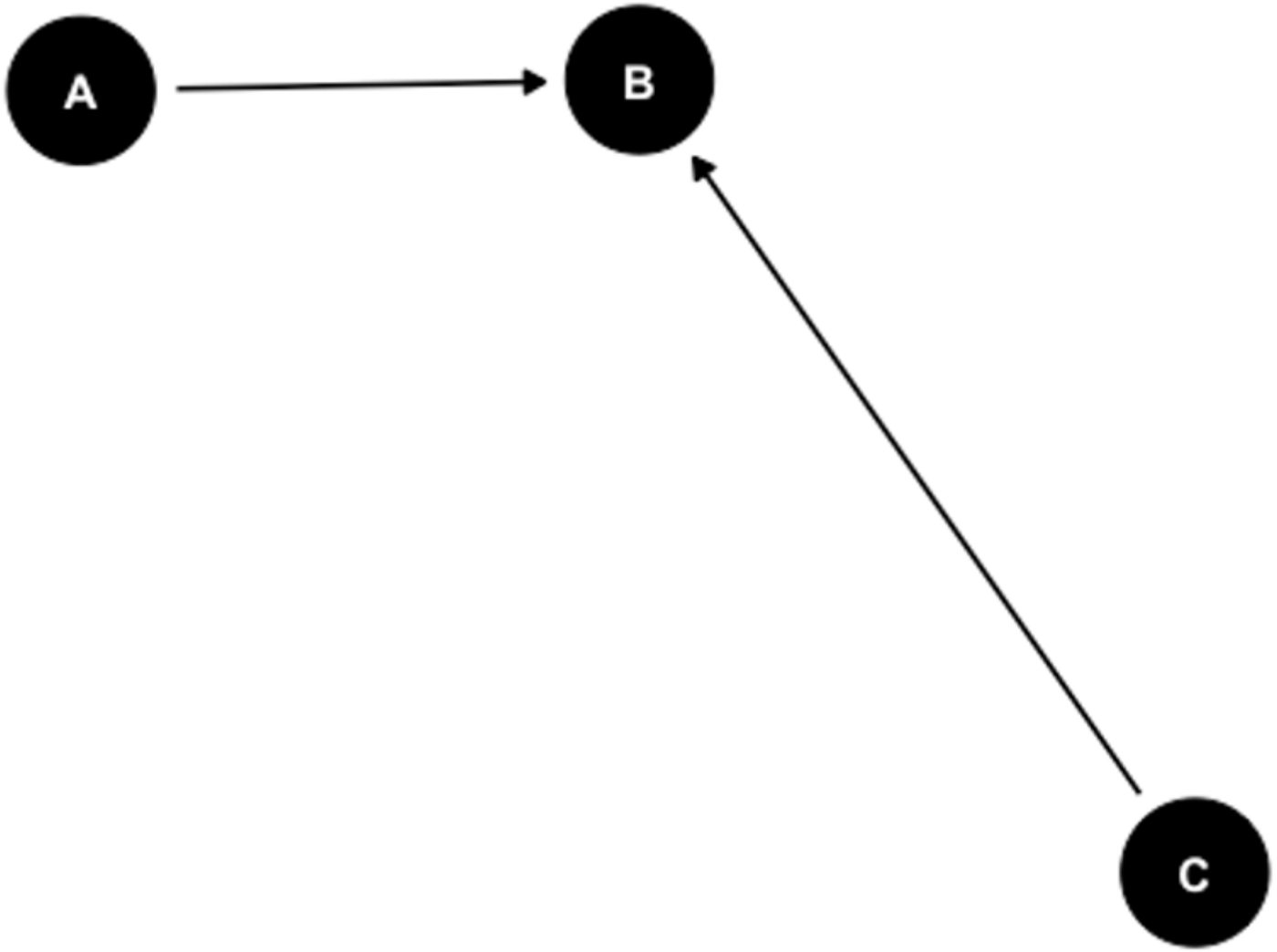
Chains



Forks

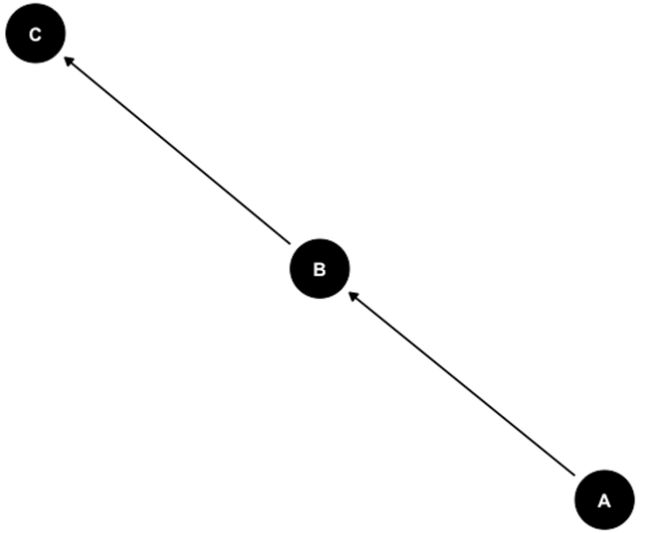


Colliders

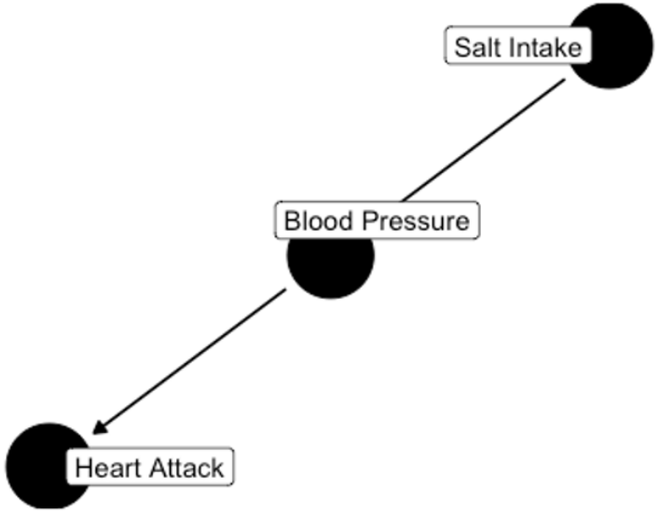


Why Care About This?

- Say we have a chain:
  - C depends on B, which depends on A
  - This means C (probably) depends on A
- What if we condition on B?



Chain Example



- Condition on Blood Pressure, so we look at people with only a specific blood pressure
- Heart Attack and Salt Intake are *conditionally independent* given Blood Pressure

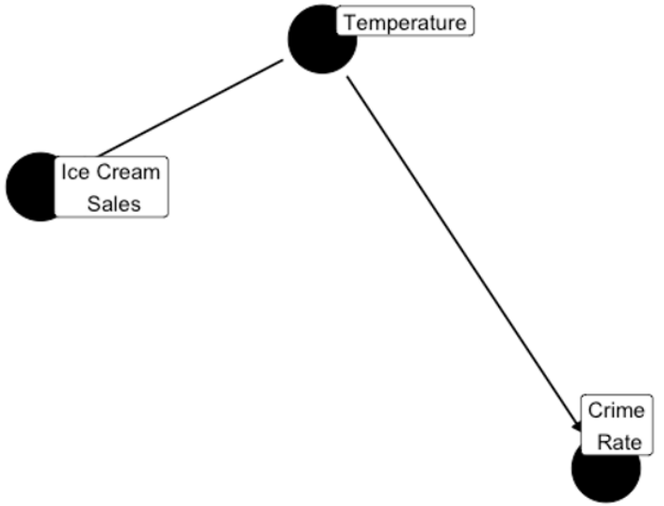
Let's simulate this

X – Learning

C – Knowing

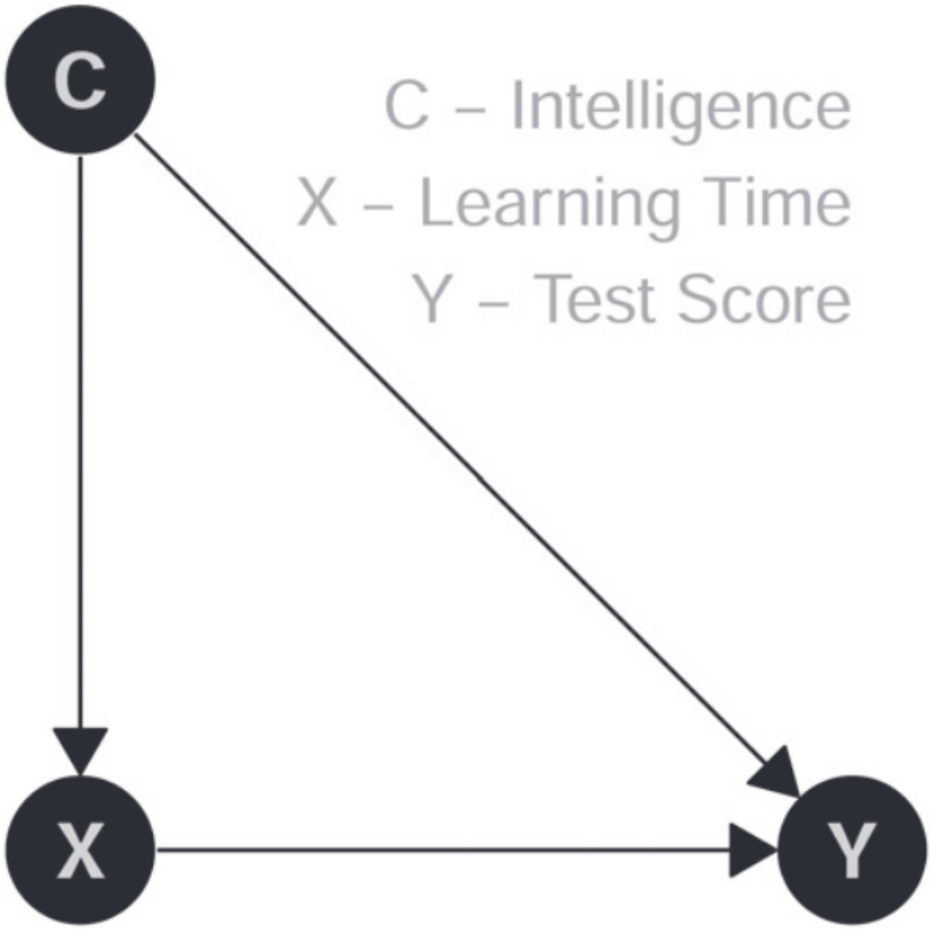
Y – Understanding

Fork Example



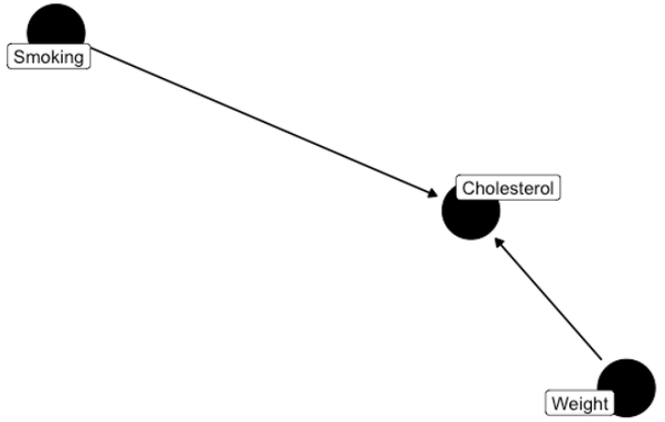
- Ice cream sales and crime rate are correlated
- What if we condition on temperature?

Fork simulation



(This is Simpson’s paradox)

Collider Example

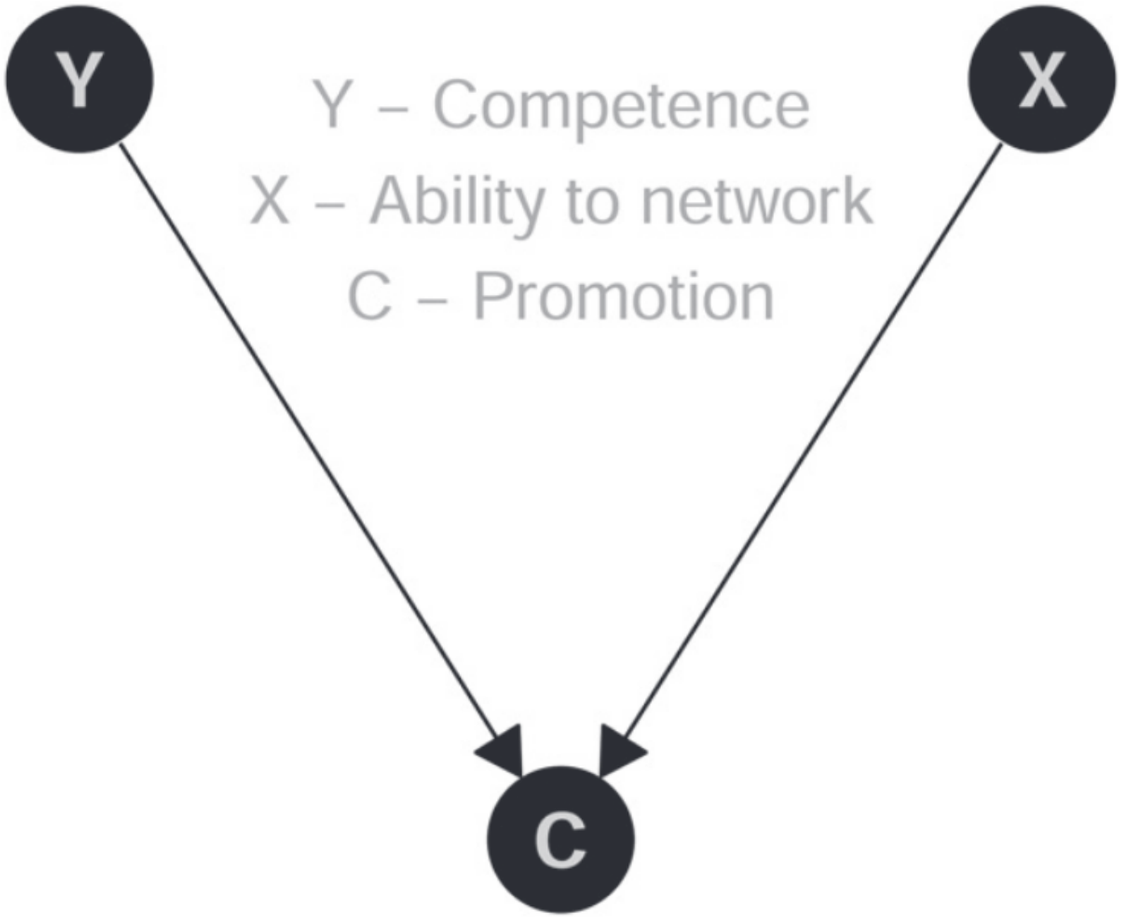


Using this diagram, are weight and smoking independent conditioned on (or given) Cholesterol?

Another Collider Example

- Getting the flu and chicken pox are independent
- Both lead to a fever
- Are they independent if we condition on having a fever?

Simulated collider



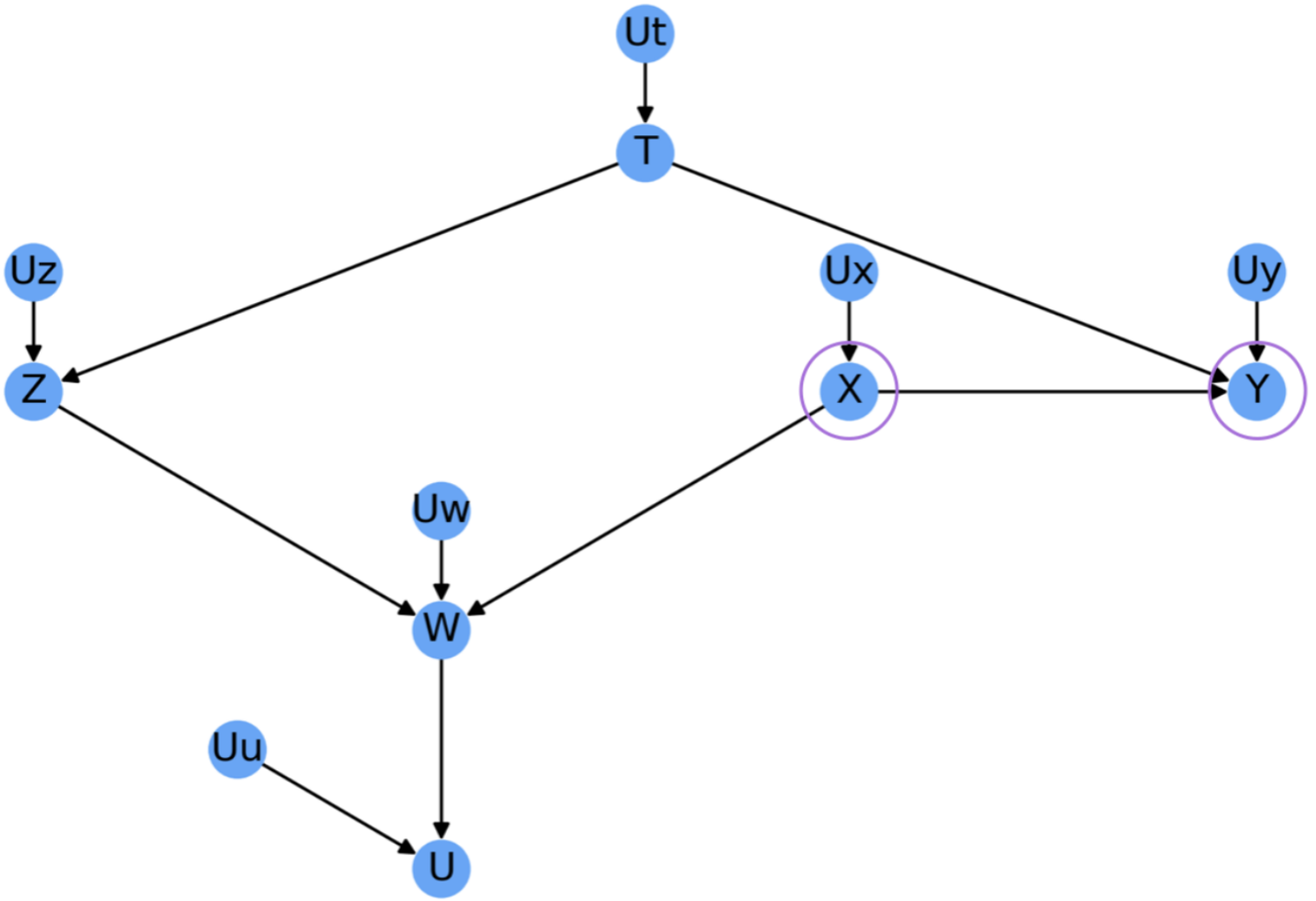
(Berkson’s paradox)

- These independences lead to testable theories, in that we can test the graph
- We could also create a set of all graphs which are compatible with the data

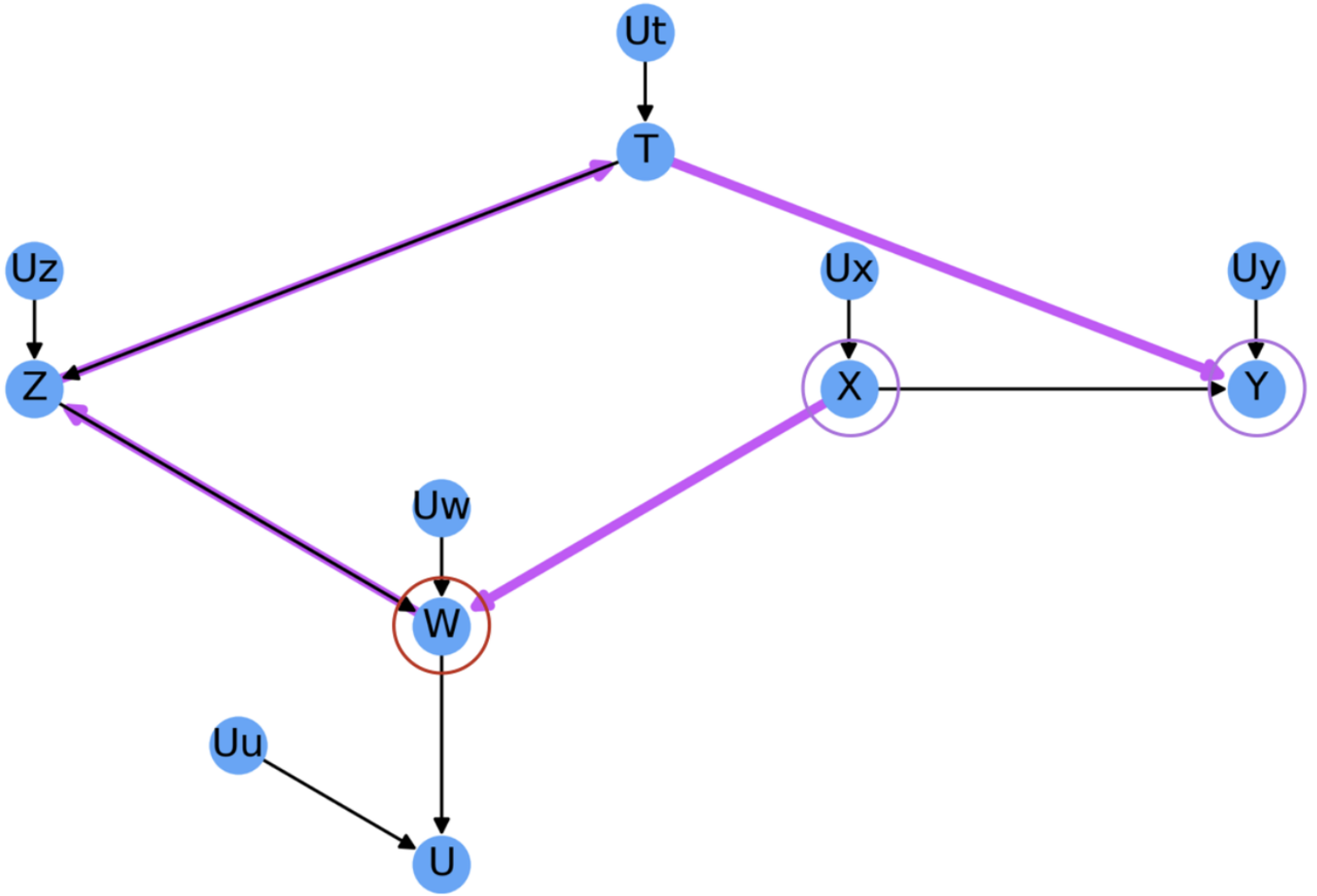
Backdoor Criteria

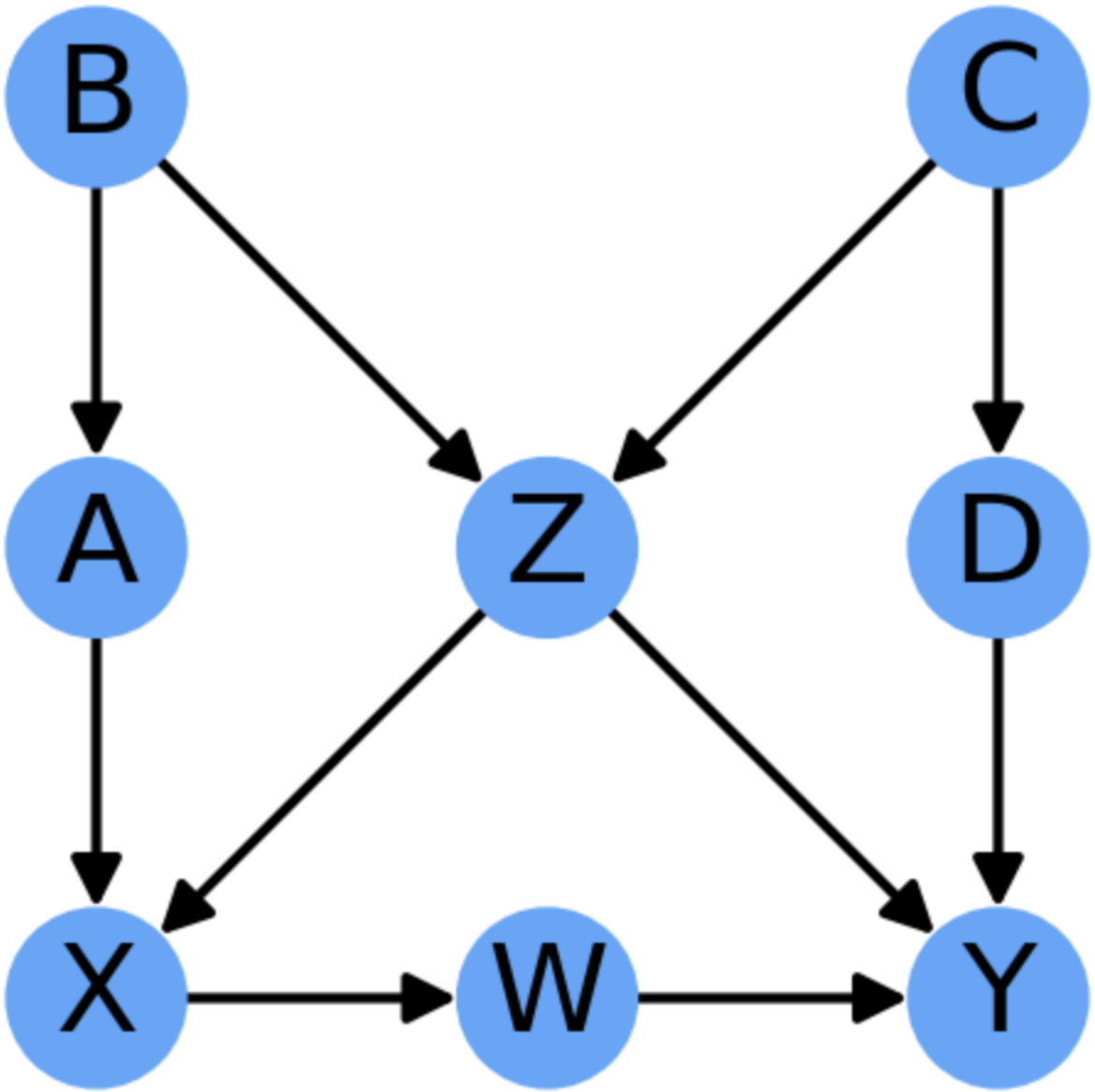
To figure out what we should condition on: From X to Y: Find a set so that it contains no descendants of X, and blocks every path between X and Y that contains an arrow into X

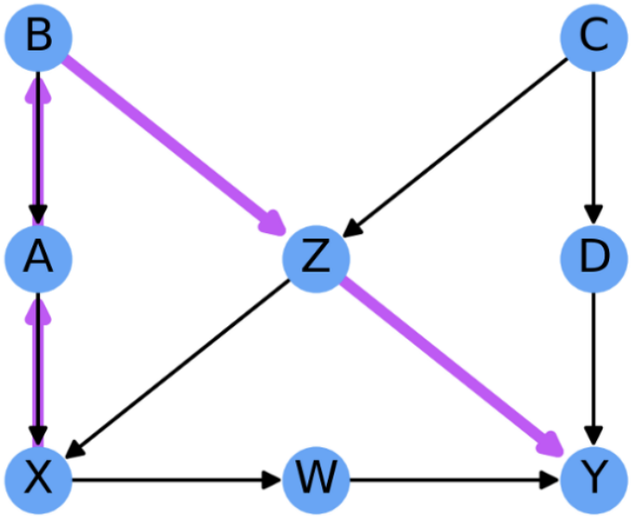
Examples - Effect of X on Y?

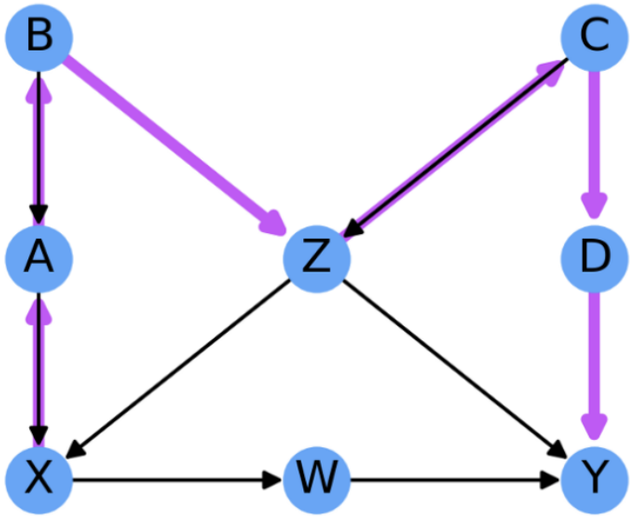


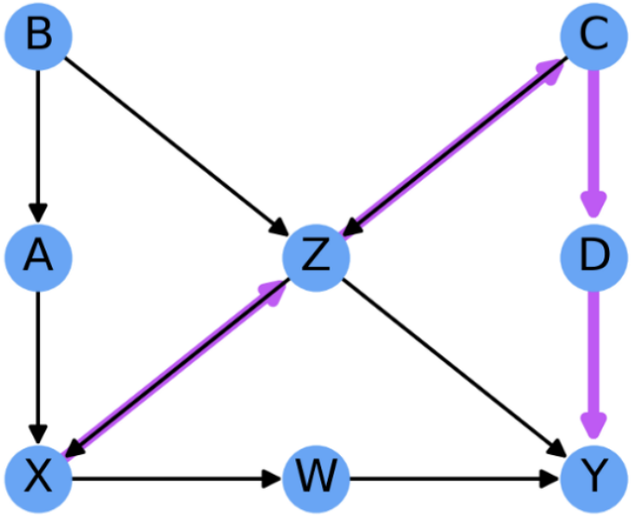
What if we want to condition on W?




B
C
A
Z
D
X
W
Y


B
C
A
Z
D
X
W
Y


B
C
A
Z
D
X
W
Y


B
C
A
Z
D
X
W
Y

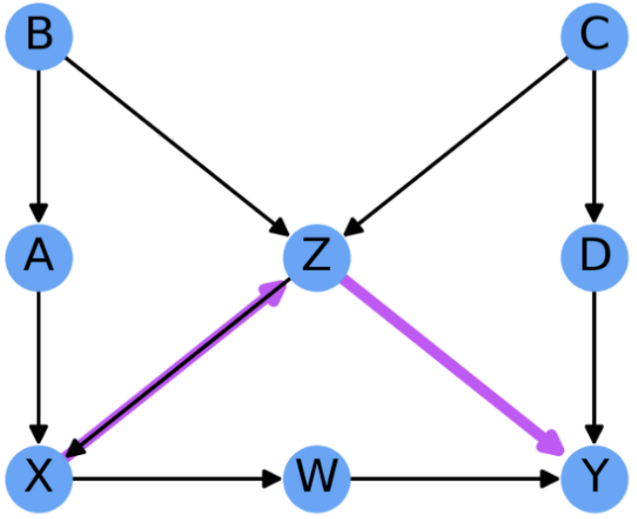

B
C
A
Z
D
X
W
Y

Table 1. Description of variables in this study.

Variable	Description
AGE	Subject age (years)
FEV	Forced expiratory volume (L)
HEIGHT	Subject height (inches)
SEX	Biological sex: Female (0), Male (1)
SMOKE	Has the subject ever smoked? No (0), Yes (1)

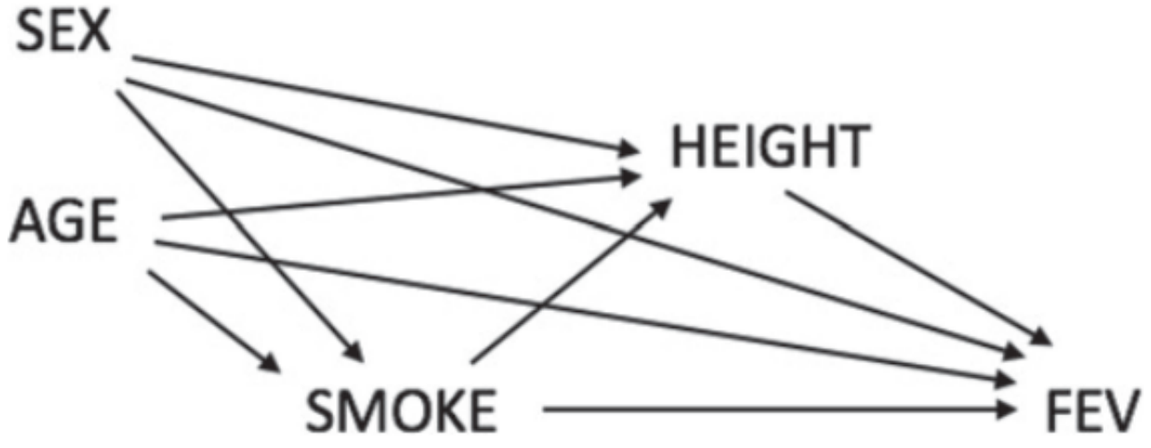
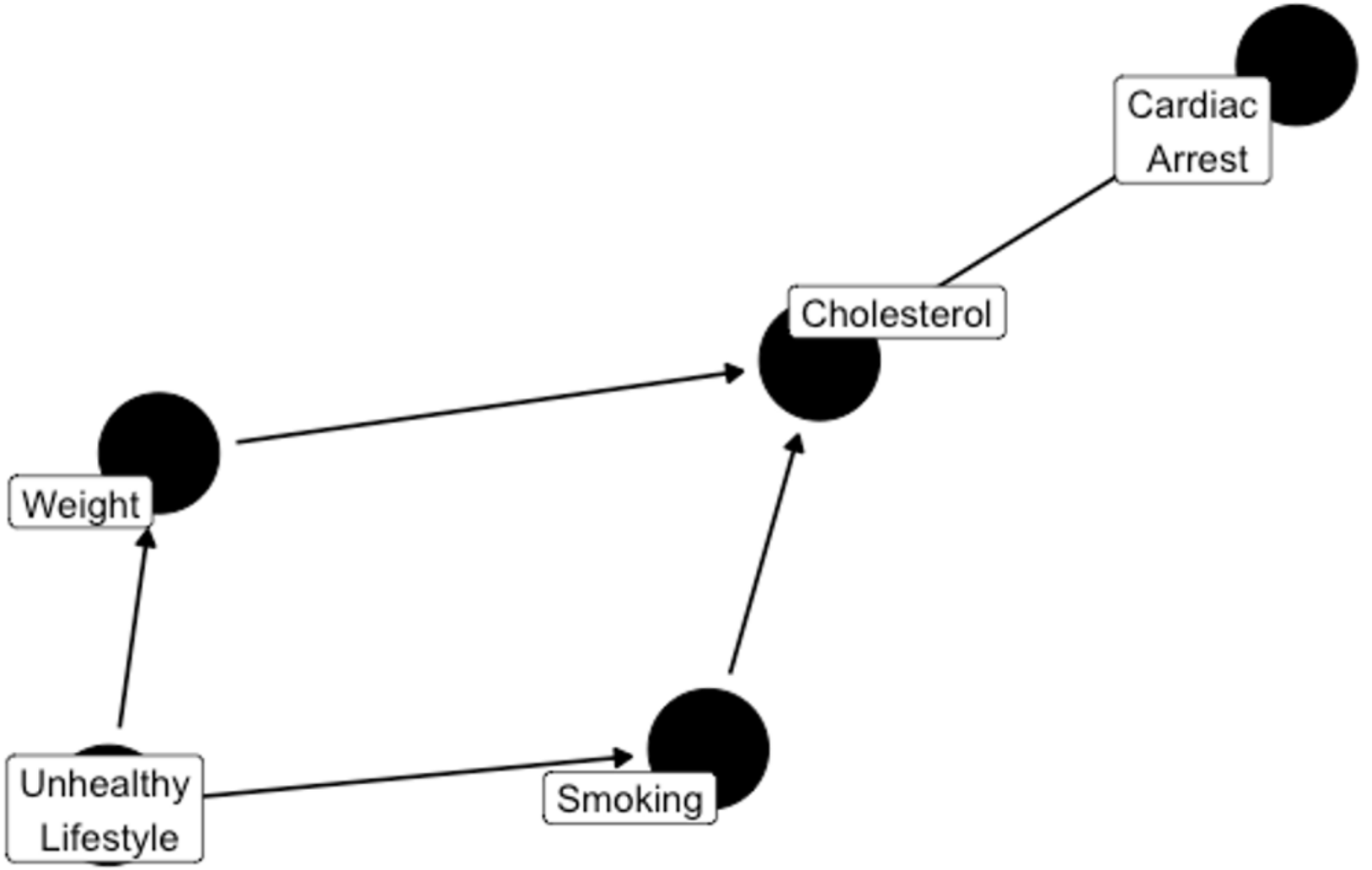
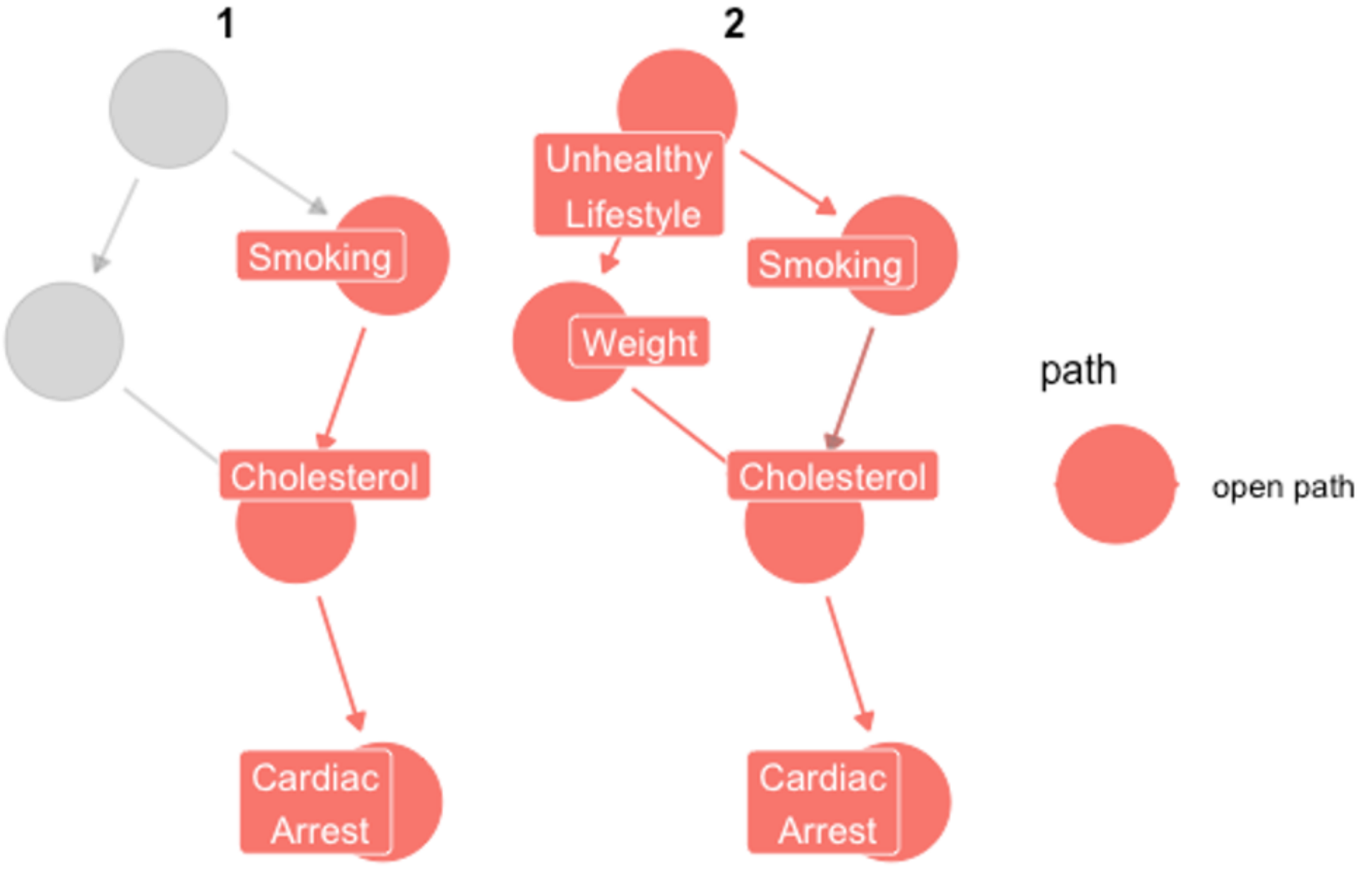
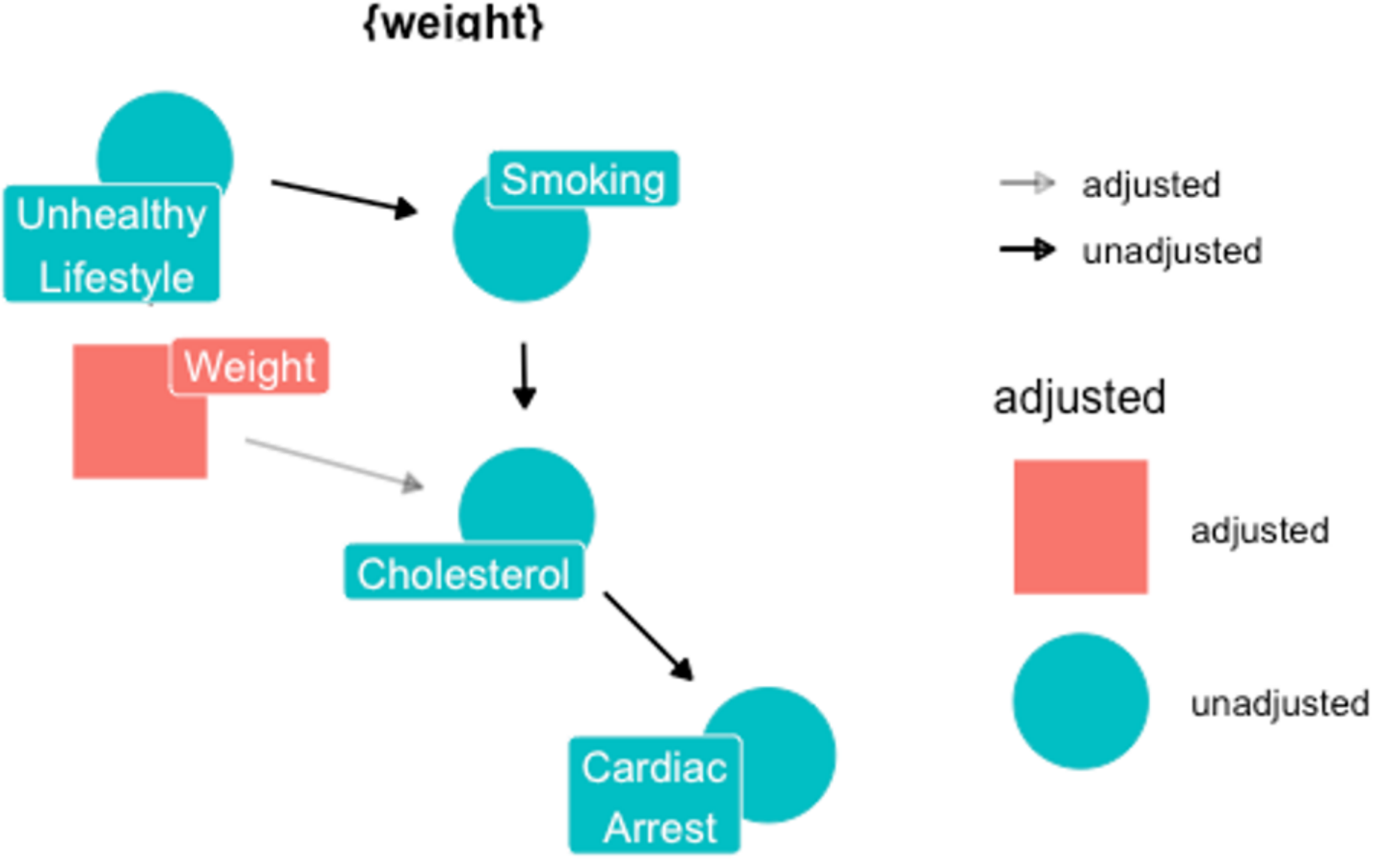


Figure 6. A causal diagram depicting relationships between variables in this study.

Smoking -> Cardiac Arrest




1
Smoking
Cholesterol
Cardiac Arrest
2
Unhealthy Lifestyle
Smoking
Weight
Cholesterol
Cardiac Arrest
path
open path


{weight}
Unhealthy Lifestyle
Smoking
Weight
Cholesterol
Cardiac Arrest
adjusted
unadjusted
adjusted
adjusted
unadjusted

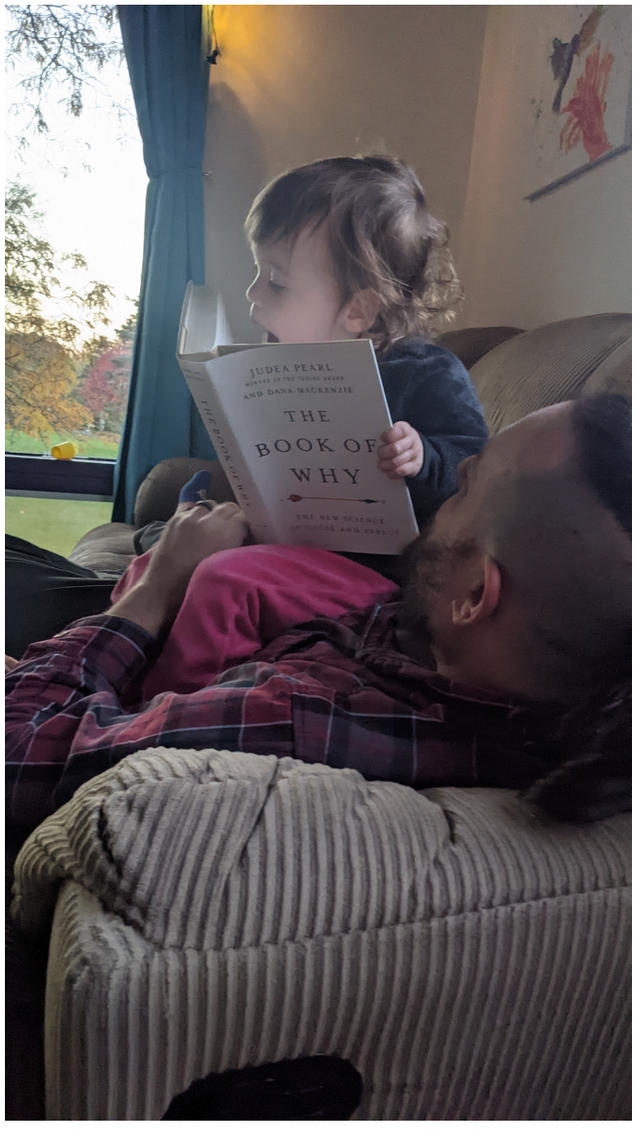
Moral of the Story

Conditioning on everything can lead to unbiased estimates, or worse, even the wrong conclusions (like Simpson's paradox)

Conclusion: Making Causal Inferences on the Basis of Correlational Data Is Very Hard

(Rohrer, 2018)

Pearl and Mackenzie, The Book of Why



C S
C U

Questions?