# Causal diagrams (DAGs)

**HPDS Book Club | Week 3**

Oana Enache

# This week’s papers

- [Rohrer] Rohrer JM. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Adv Methods Pract Psychol Sci*. 2018;1(1):27-42. doi:10.1177/2515245917745629
- [Tennant et al] Tennant PWG, Murray EJ, Arnold KF, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*. 2021;50(2):620-632. doi:10.1093/ije/dyaa213

# Some other helpful resources

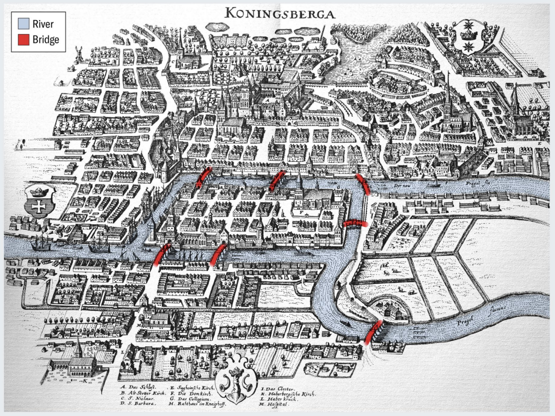
- Barrett M, McGowan LD, Gerke T. Causal Inference in R. March 28, 2025. Accessed April 7, 2025. https://www.r-causal.org/
    - Specifically the “Expressing causal questions as DAGs” chapter
- Westreich D. *Epidemiology by Design: A Causal Approach to the Health Sciences*. Oxford University Press; 2019. doi:10.1093/oso/9780190665760.001.0001

# Roadmap for today

1. Review of (some) key terms: graph theory
2. What’s a causal DAG?
3. DAGs & bias
4. Current use of DAGs + suggestions for improvement
5. Other questions/discussion

# Review of key terms: Graph theory

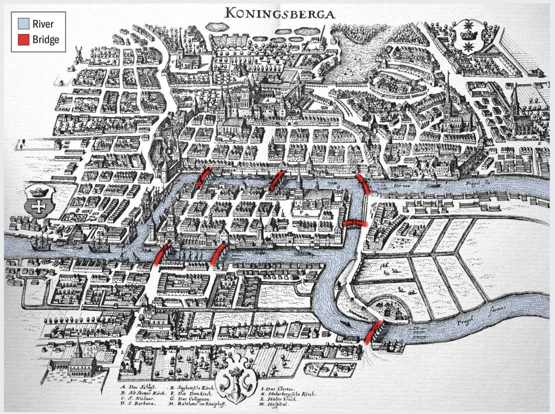
# What is a graph?



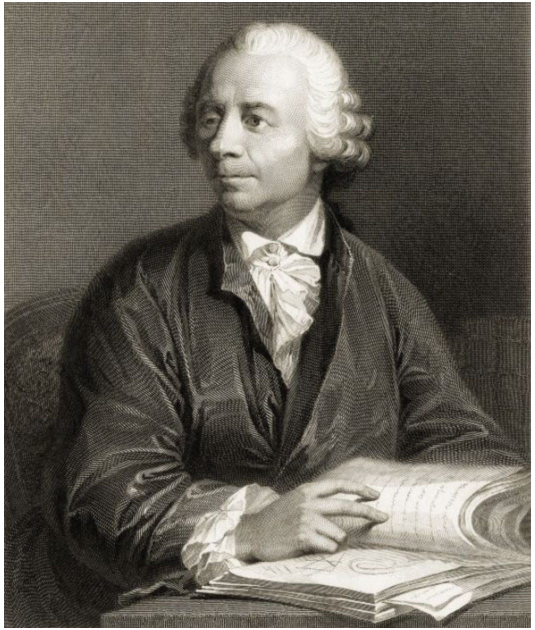
The Prussian city of Königsberg, now Kaliningrad in Russia, in the Middle Ages. The 18th-century city map is a digitally enhanced reproduction that was modified to highlight the river and bridges. Credit: Alamy Stock Photo (*map*); Amanda Montañez (*highlighting*)

Barabasi AL. *Network Science*. Cambridge University Press; 2016.
Images: https://www.scientificamerican.com/article/how-the-seven-bridges-of-koenigsberg-spawned-new-math/

# What is a graph?



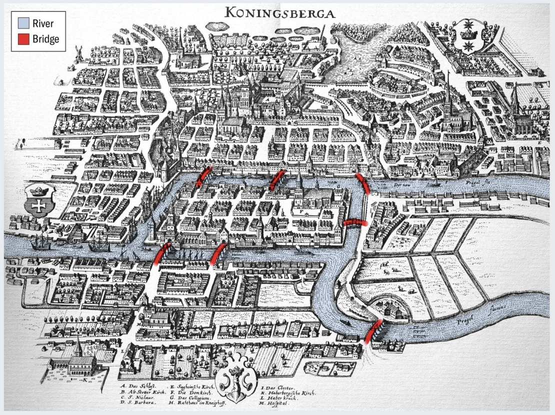
The Prussian city of Königsberg, now Kaliningrad in Russia, in the Middle Ages. The 18th-century city map is a digitally enhanced reproduction that was modified to highlight the river and bridges. Credit: Alamy Stock Photo (*map*); Amanda Montañez (*highlighting*)



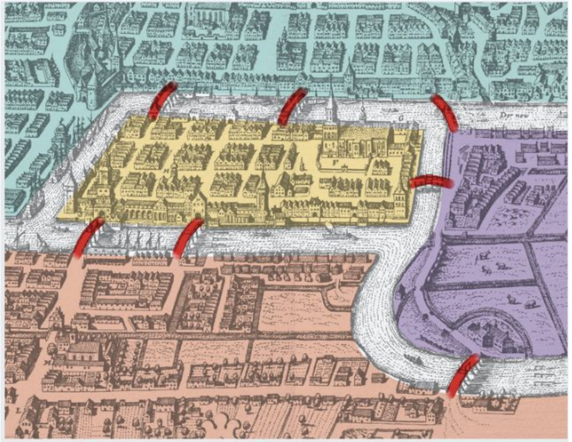
Barabasi AL. *Network Science*. Cambridge University Press; 2016.

Images: https://www.scientificamerican.com/article/how-the-seven-bridges-of-koenigsberg-spawned-new-math/

# What is a graph?

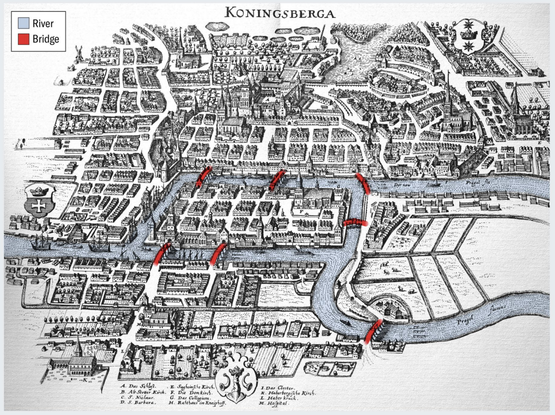


The Prussian city of Königsberg, now Kaliningrad in Russia, in the Middle Ages. The 18th-century city map is a digitally enhanced reproduction that was modified to highlight the river and bridges. Credit: Alamy Stock Photo (*map*); Amanda Montañez (*highlighting*)

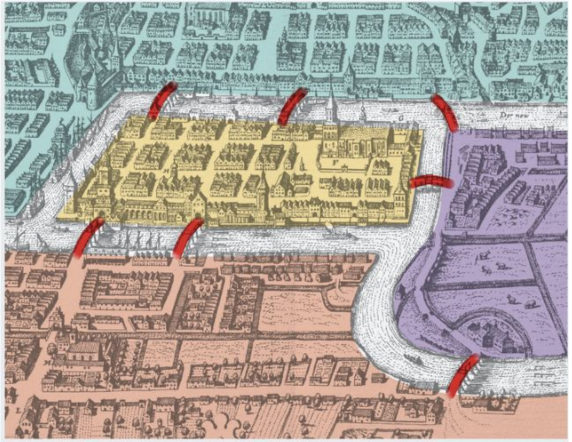


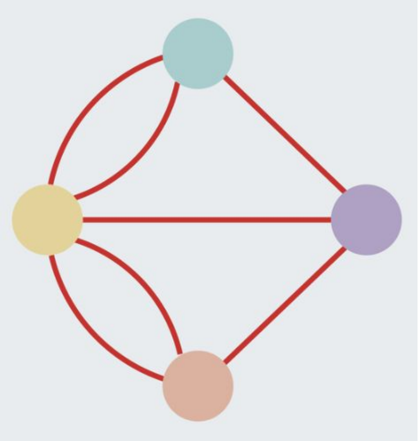
Barabasi AL. *Network Science*. Cambridge University Press; 2016.
Images: https://www.scientificamerican.com/article/how-the-seven-bridges-of-koenigsberg-spawned-new-math/

# What is a graph?



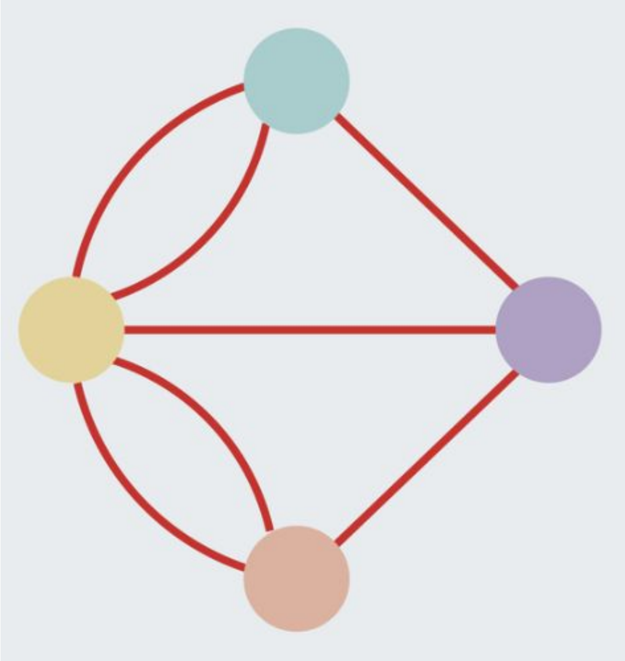
The Prussian city of Königsberg, now Kaliningrad in Russia, in the Middle Ages. The 18th-century city map is a digitally enhanced reproduction that was modified to highlight the river and bridges. Credit: Alamy Stock Photo (*map*); Amanda Montañez (*highlighting*)





Barabasi AL. *Network Science*. Cambridge University Press; 2016.
Images: https://www.scientificamerican.com/article/how-the-seven-bridges-of-koenigsberg-spawned-new-math/

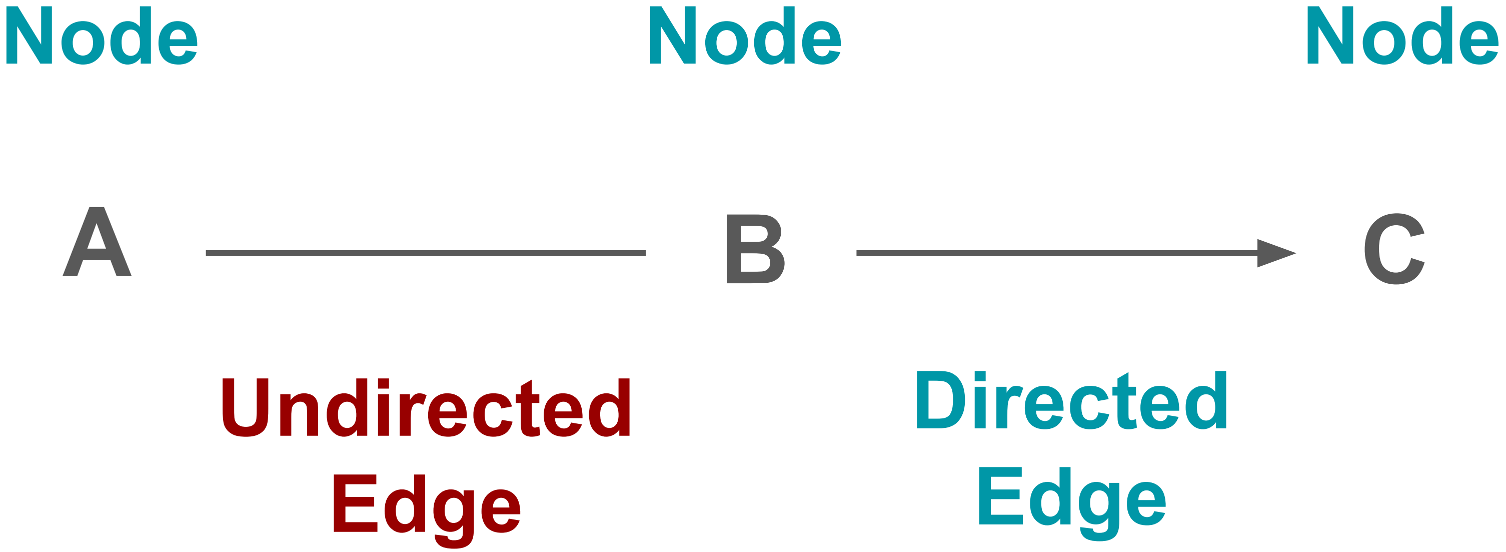
# What is a graph?



A data structure consisting of
a set of **nodes** (aka: vertex, point, variable)
and a set of **edges**

Barabasi AL. *Network Science*. Cambridge University Press; 2016.
Images: https://www.scientificamerican.com/article/how-the-seven-bridges-of-koenigsberg-spawned-new-math/

# Types of edges



Node

Node

Node

A

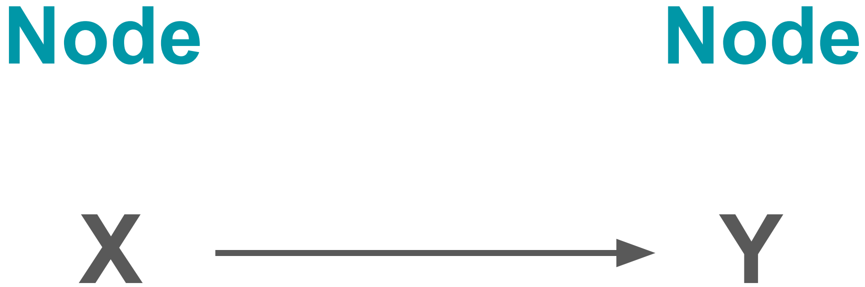
B

C

Undirected Edge

Directed Edge

Aka: arrow, arc

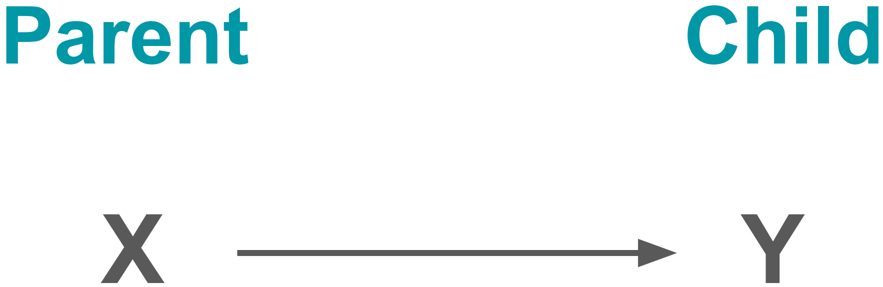


Node

Node

X

Y

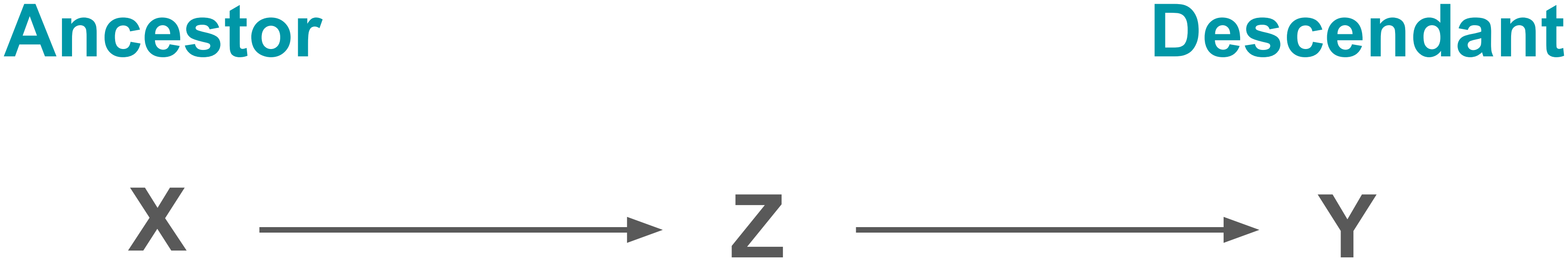


Parent

Child

X

Y



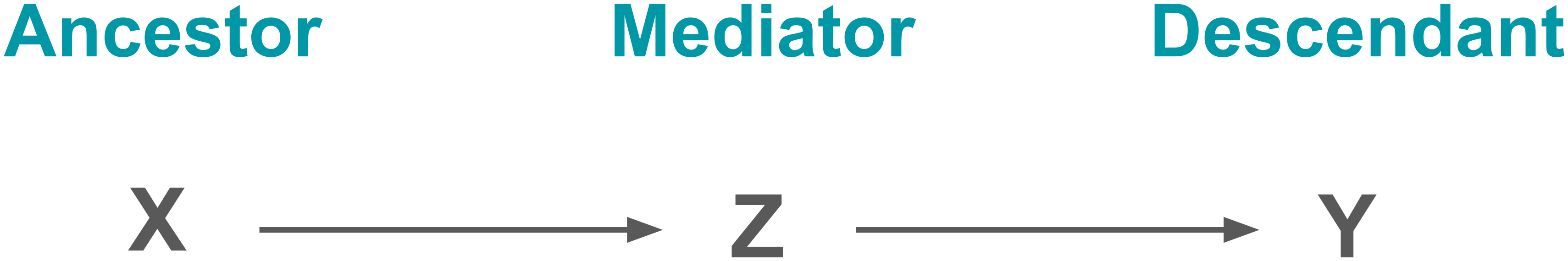
Ancestor

Descendant

X

Z

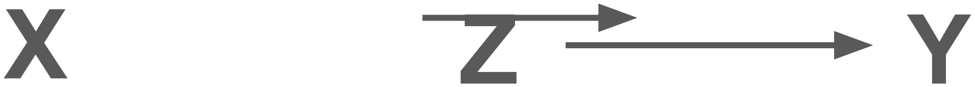
Y



Ancestor Mediator Descendant

X Z Y

## Chain

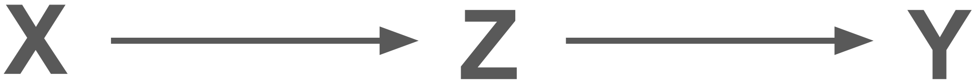


X

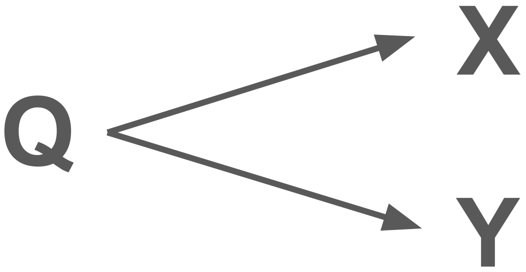
Z

Y

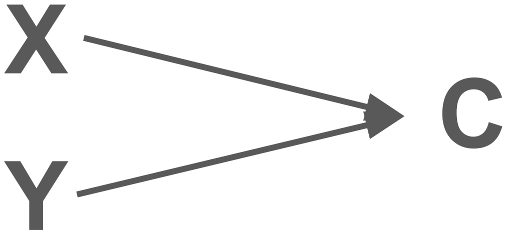
Chain



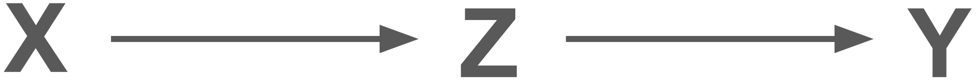
Fork



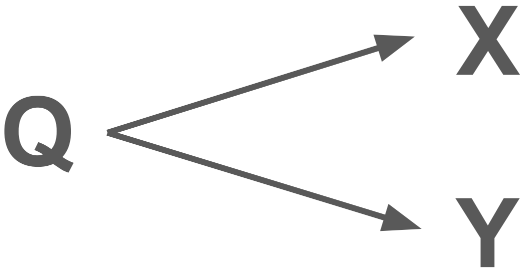
Inverse fork



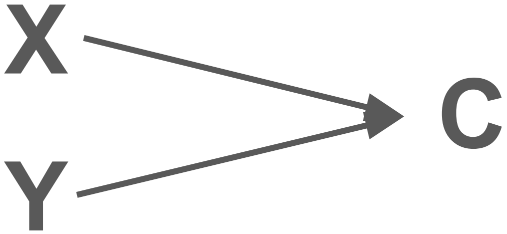
Chain



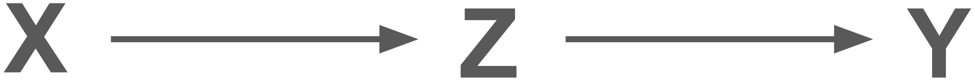
Fork



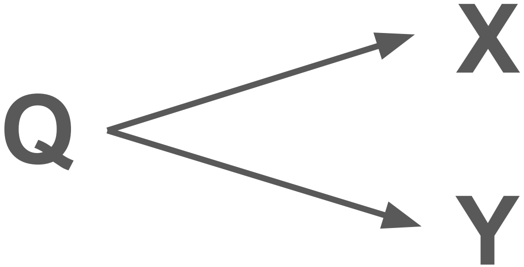
## Collider



## Chain

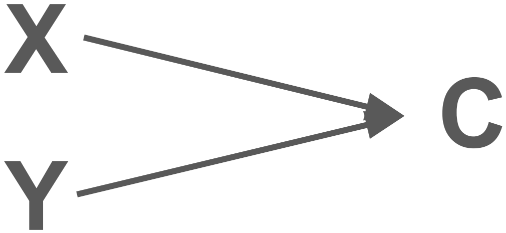


## Fork

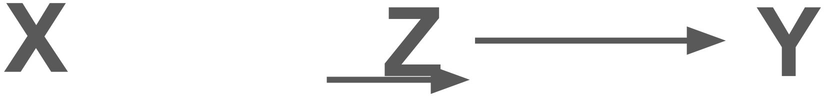


# Paths

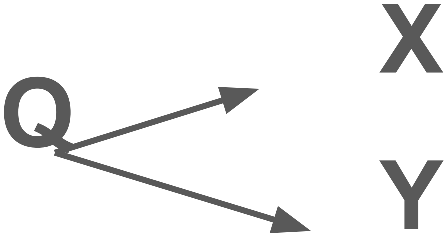
Collider



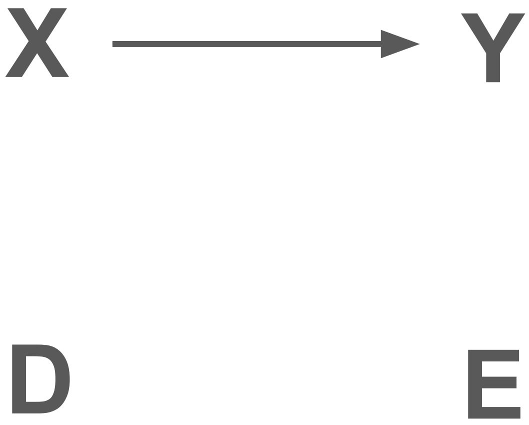
Chain



Fork



# Paths & causal relationships



Assumed causal relationship

No causal relationship

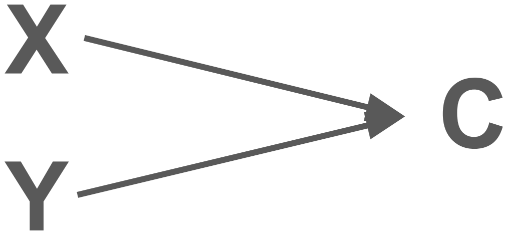
# Open or closed paths

- When a path transmits association between two nodes, the path between those nodes is **open**
- When a path does not transmit association between two nodes, the path between those nodes is **closed**

# Your turn! Which paths between X & Y are open?

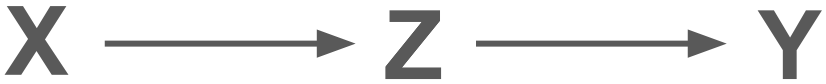
?

**Collider**



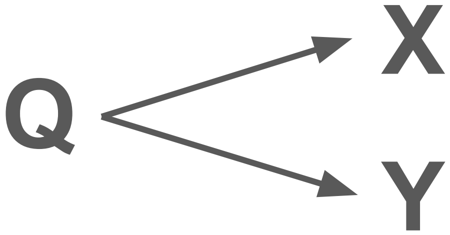
?

**Chain**



?

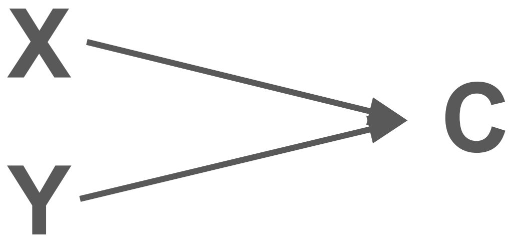
**Fork**



# Your turn! Which paths between X & Y are open?

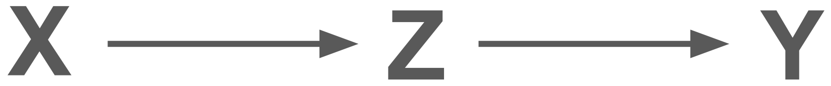
Closed

Collider



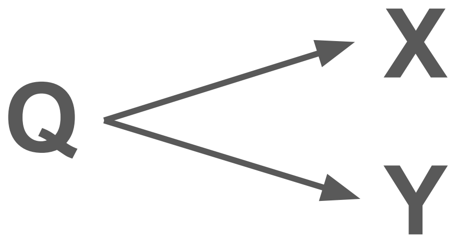
Open

Chain



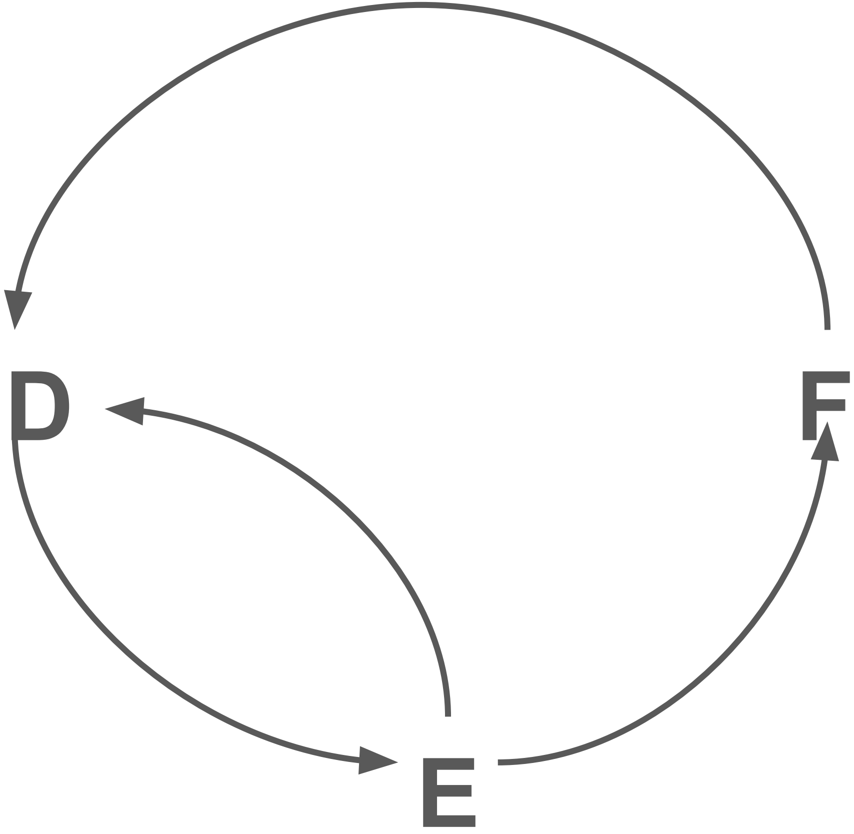
Open

Fork



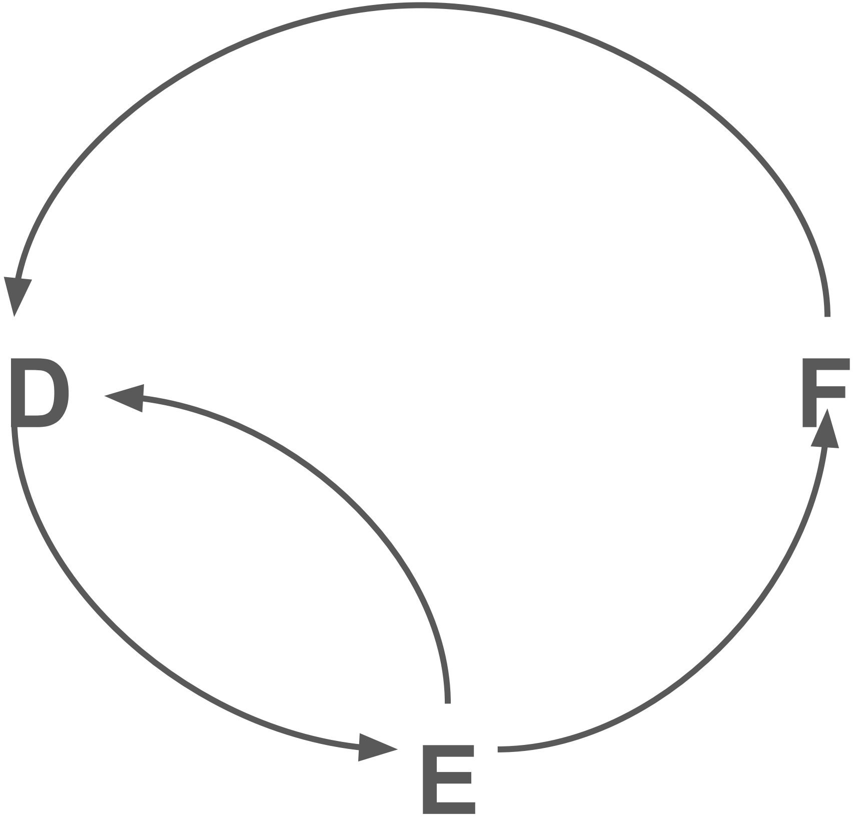
# Paths

Cylic



# Paths

Directed
Cylic

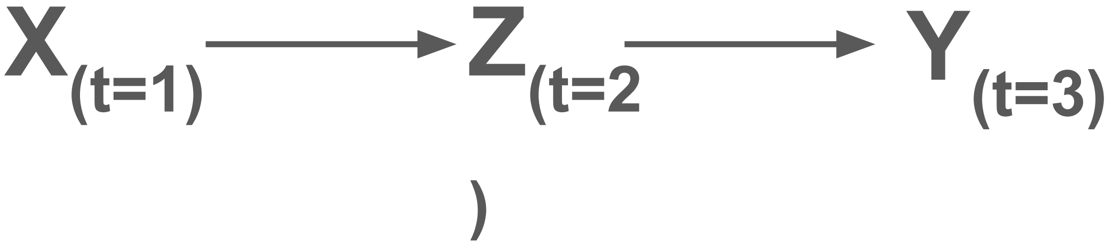


# Causal Directed Acyclic Graphs (DAGs)

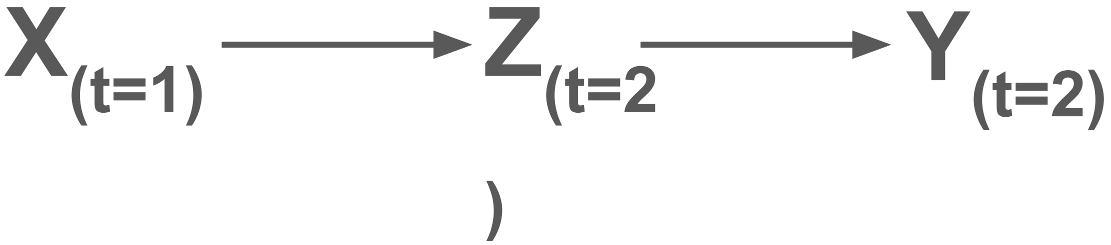
# Causal DAGs

- (Rohrer) “DAGs provide visual representations of causal assumptions… [and] offer an intuitive approach for thinking about causal structures”
- (Tennant et al) “DAGs are a non-parametric diagrammatic representation of the assumed data-generating process for a set of variables (and measurements thereof) in a specific context.”

# DAGs can show longitudinal relationships

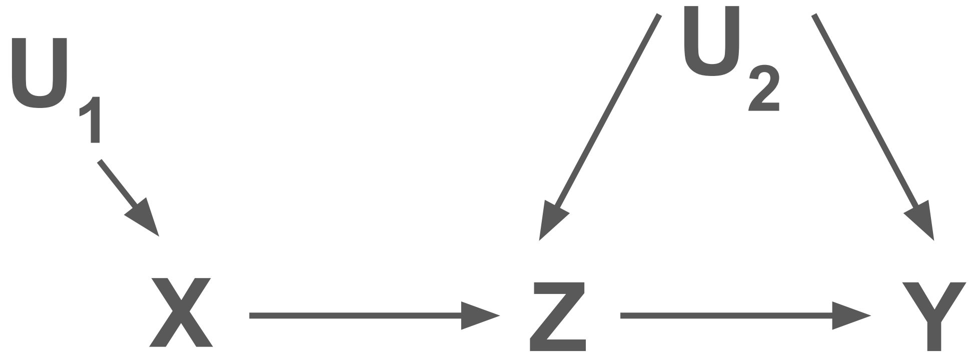


Three separate time points 1, 2, 3



Two time points; assumes Z precedes Y at time 2

# DAGs can include measured & unmeasured variables



# Causal DAGs and bias*

*__Bias__: “Any difference between the true causal effect and the expected value of the causal effect estimated in our data, where by ‘expected’ we mean ‘not due to chance alone’” (Westreich)

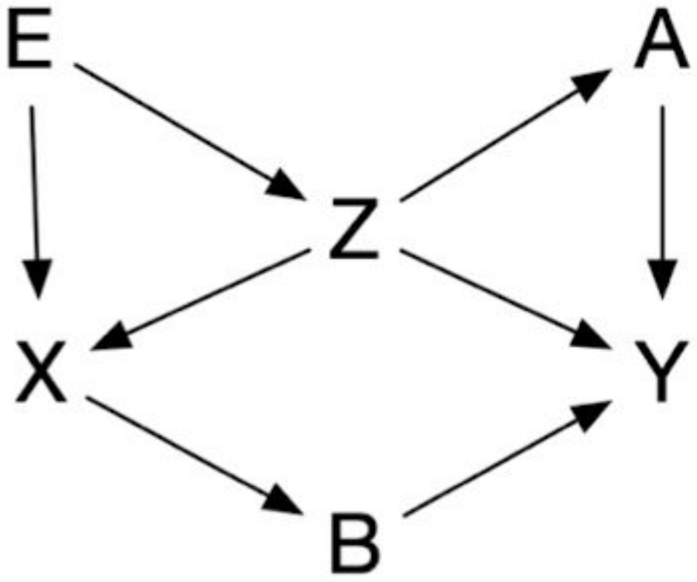
# Causal DAGs can help identify sources of bias

1. Confounding
2. Collider bias (aka: endogenous selection bias, selection bias, collider-stratification bias)

# Backdoor paths

For an ancestor and descendant of interest, a backdoor path:

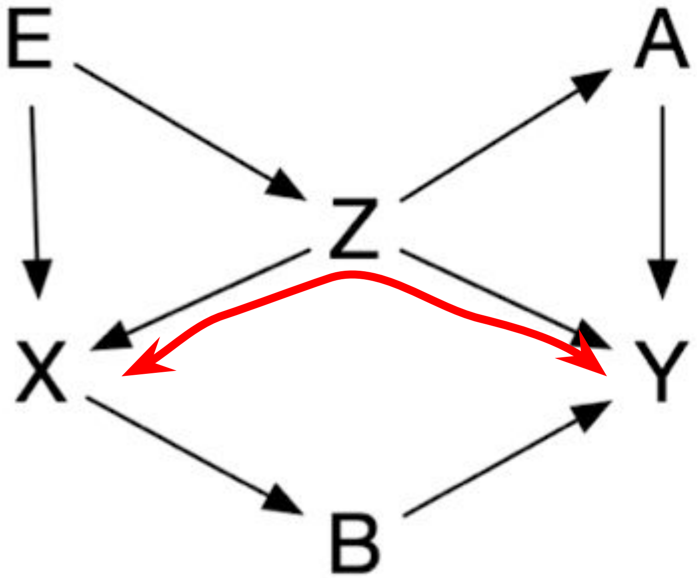
1. Begins with an arrow into the ancestor
2. Ends with an arrow into the descendant



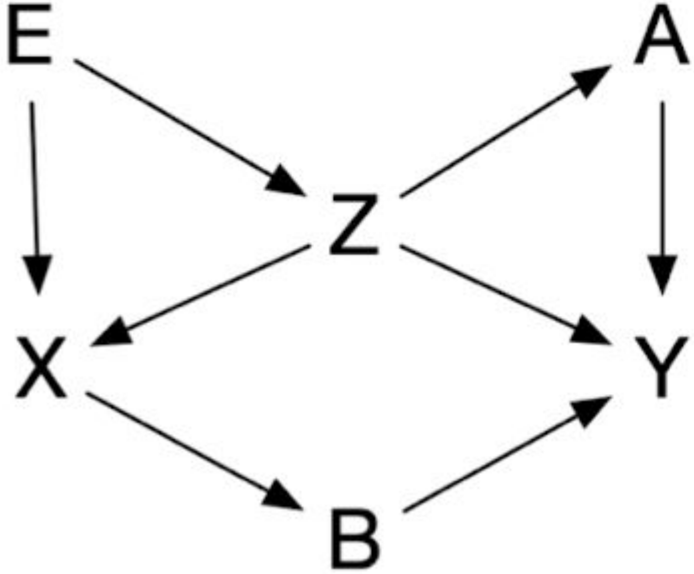
# Backdoor paths

For an ancestor and descendant of interest, a backdoor path:

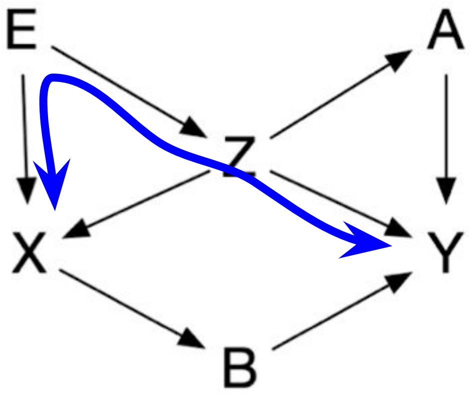
1. Begins with an arrow into the ancestor
2. Ends with an arrow into the descendant

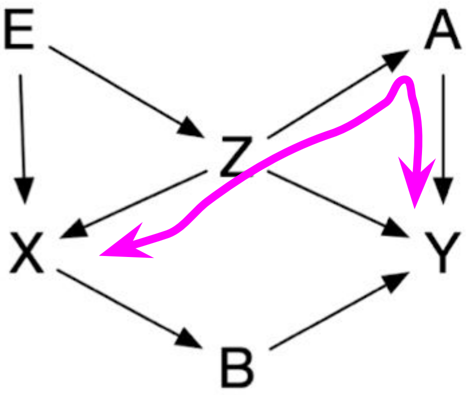


# Your turn! What are some other backdoor paths between X & Y?



# Your turn! What are some other backdoor paths between X & Y?



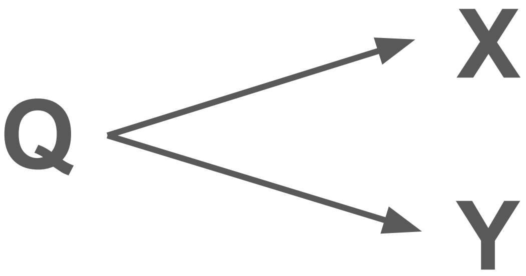


# Backdoor paths

- If an open backdoor path exists between an ancestor and a descendant, then the measure of association will be subject to **bias** and/or **confounding**
- This is also important for the **exchangeability** assumption of causal inference, because for exchangeability of potential outcomes to hold we need no open non-causal pathways

# DAGs and confounding

- (Rohrer) Confounding is:
  - “The central problem of observational data”
  - “A common cause that lurks behind the potential cause of interest…and the outcome of interest”
- DAGs can help us:
  - Identify confounders
  - Determine how confounding could be addressed



# Identifying confounders with DAGs

**0)**

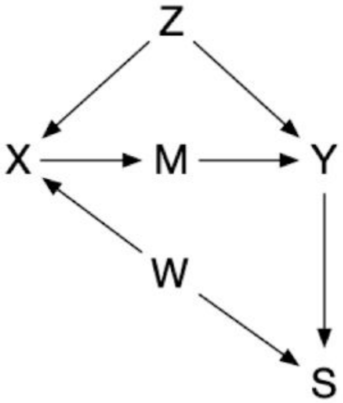
Include all variables relevant to the causal effect of interest

# Identifying confounders with DAGs

**0)**

Include all variables relevant to the causal effect of interest

**1)**

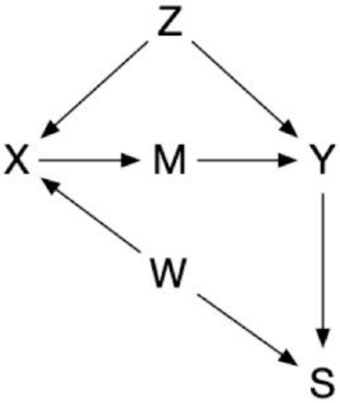


# Identifying confounders with DAGs

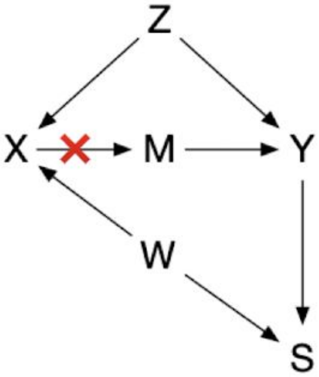
**0)**

Include all variables relevant to the causal effect of interest

**1)**



**2)**

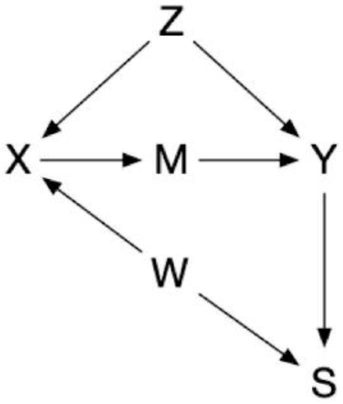


# Identifying confounders with DAGs

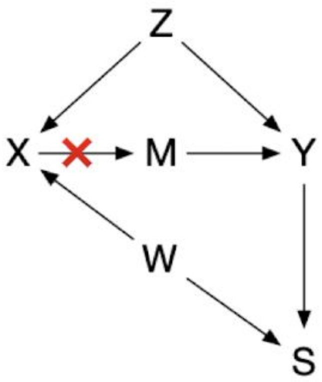
**0)**

Include all variables relevant to the causal effect of interest

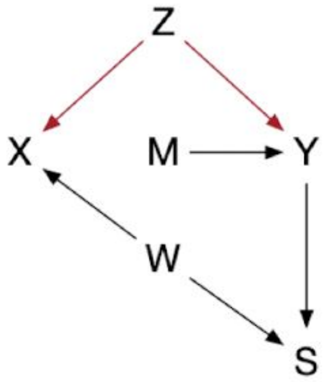
**1)**



**2)**



**3)**

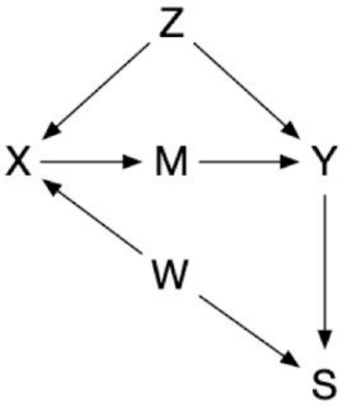


# Identifying confounders with DAGs

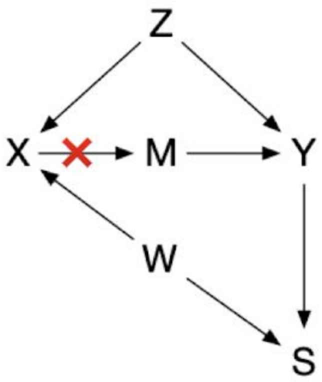
**0)**

Include all variables relevant to the causal effect of interest

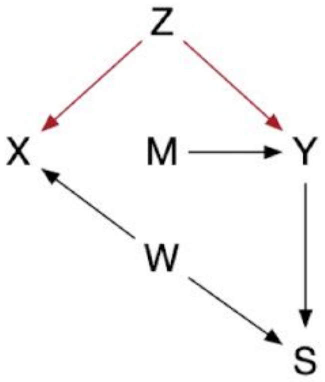
**1)**



**2)**



**3)**



**4)**

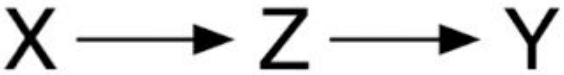
**Z is a confounder here!**

# Approaches for blocking backdoor paths

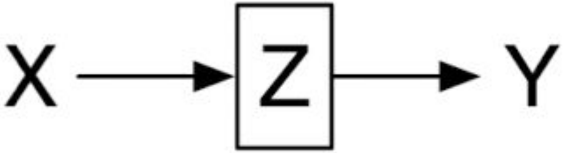
- Blocking is sometimes referred to as “conditioning” or “adjustment”
- We care about this because blocking all backdoor paths between an ancestor and descendant of interest gives us the true causal effect between these variables (assuming the DAG captures the true underlying set of causal relationships)
- Common approaches:
  - Stratified analyses
  - Multiple regression
  - Matching

# Visual representation of adjusted path

Chain



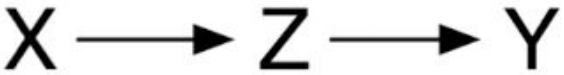
unblocked



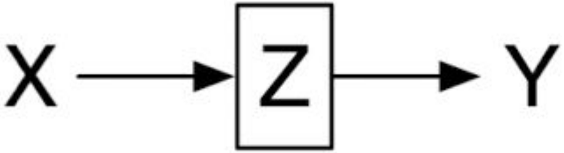
blocked

# Watch out for colliders!

## Chain

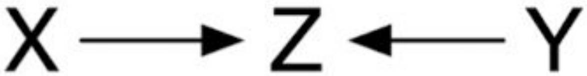


unblocked

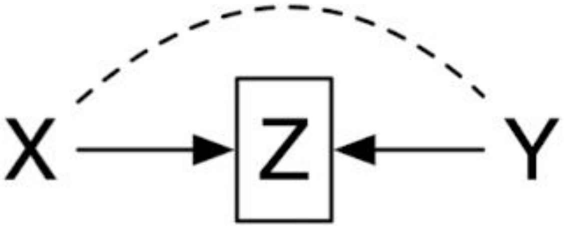


blocked

## Collider



blocked



unblocked

# Colliders and selection bias

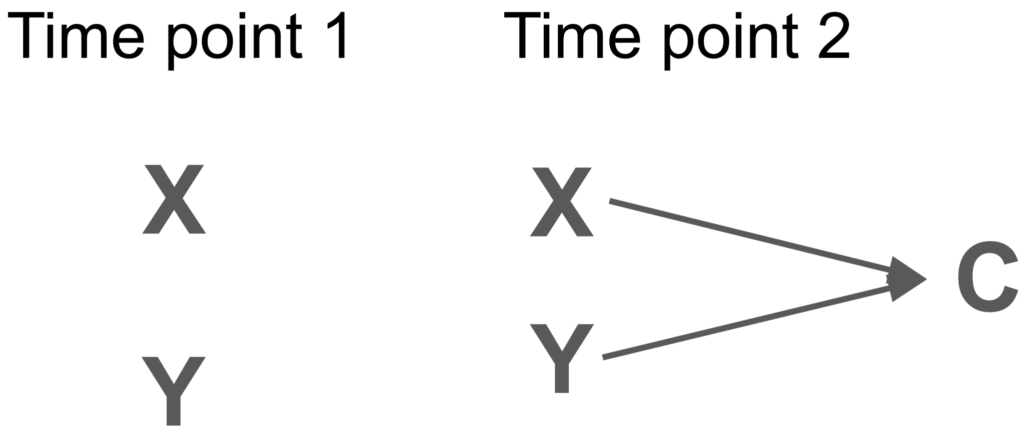
Time point 1

X

Y

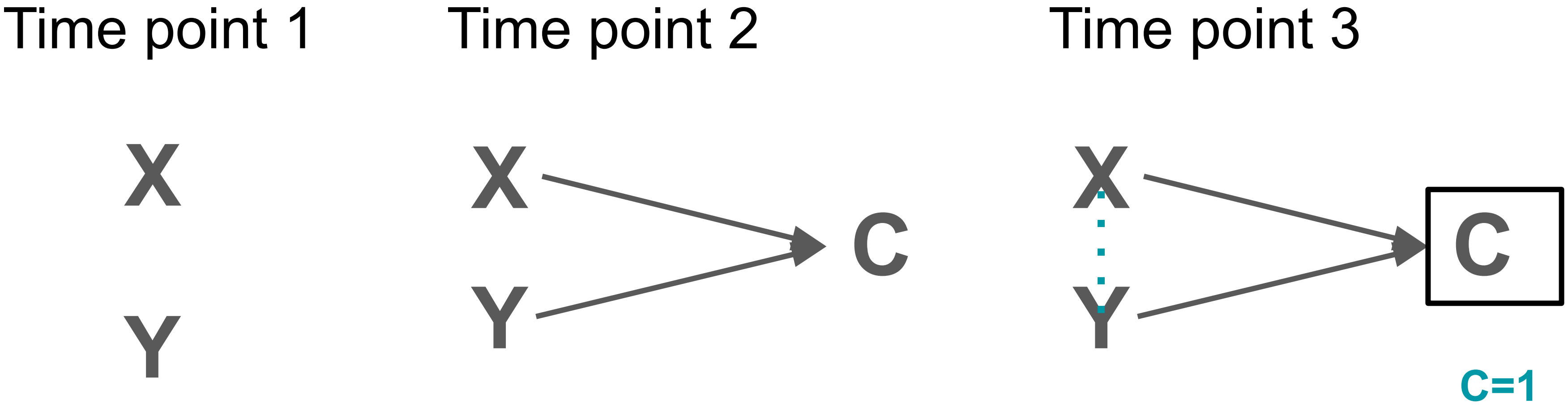
Adapted from Fig. 4.7 in “Expressing causal questions as DAGs”, Causal Inference in R

# Colliders and selection bias



Adapted from Fig. 4.7 in “Expressing causal questions as DAGs”, Causal Inference in R

# Colliders and selection bias



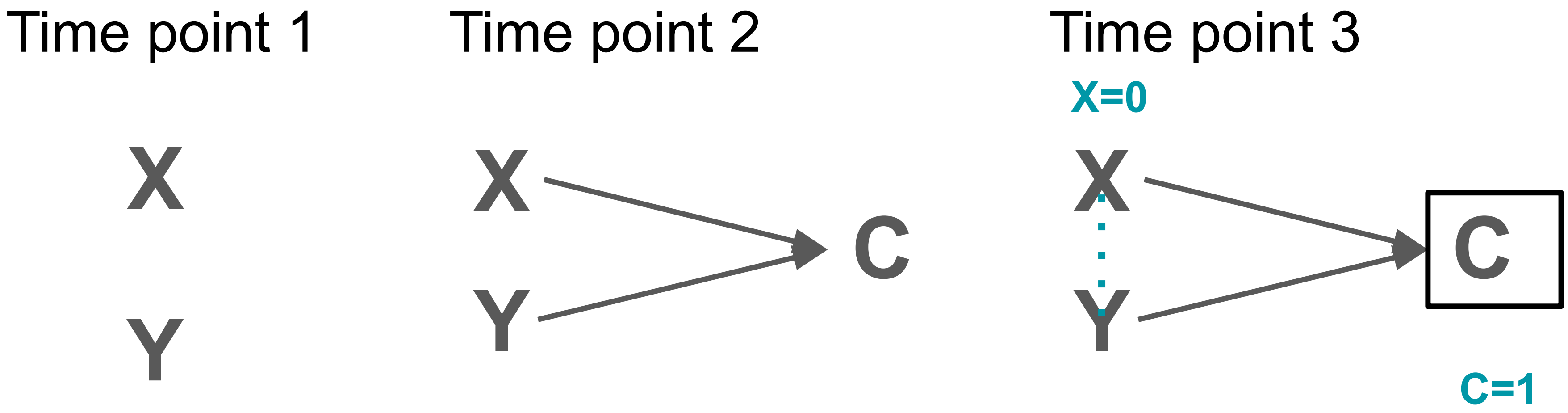
Time point 1

Time point 2

Time point 3

Adapted from Fig. 4.7 in “Expressing causal questions as DAGs”, Causal Inference in R

# Colliders and selection bias



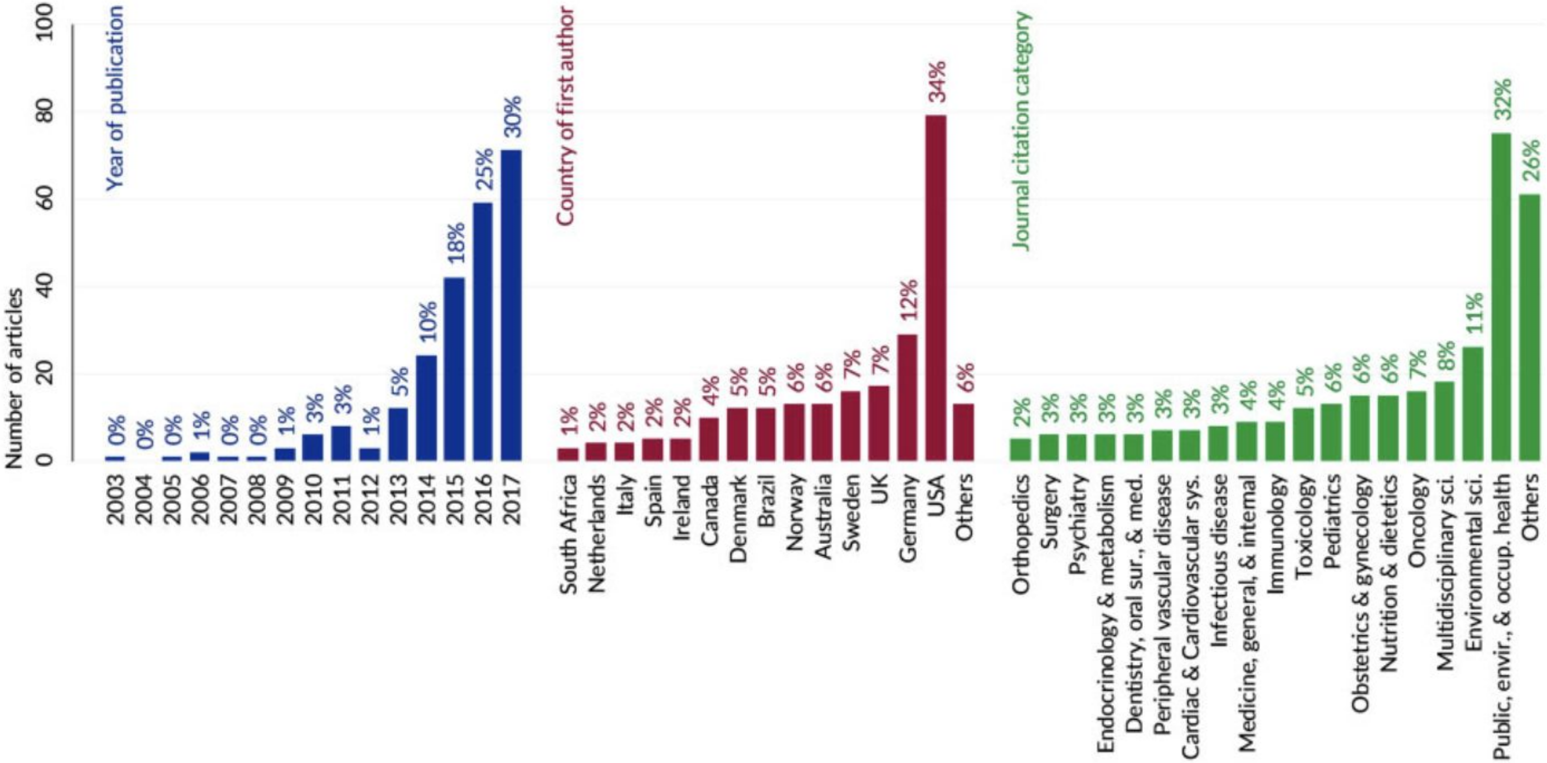
Adapted from Fig. 4.7 in “Expressing causal questions as DAGs”, *Causal Inference in R*

# How do you choose what to adjust?

- Adjustment set: The set(s) of variables/nodes we need to adjust for
- Theory vs. practice
  - Theory: If you have perfect data + a perfect DAG modeled correctly, it doesn’t matter
  - In practice: It depends and is usually a judgement call!
- Some considerations:
  - Type of node/variable we want to adjust for
  - Effect of interest
  - What adjustment sets are available
    - How well are adjustment set variables measured?
  - Temporality

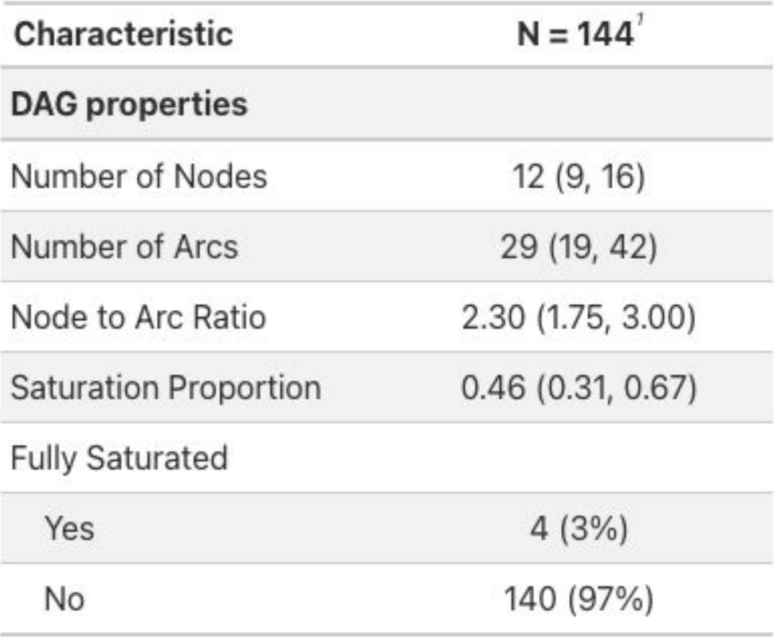
# Current use of DAGs in applied health research and suggestions for improvement

# Use of DAGs is growing overall

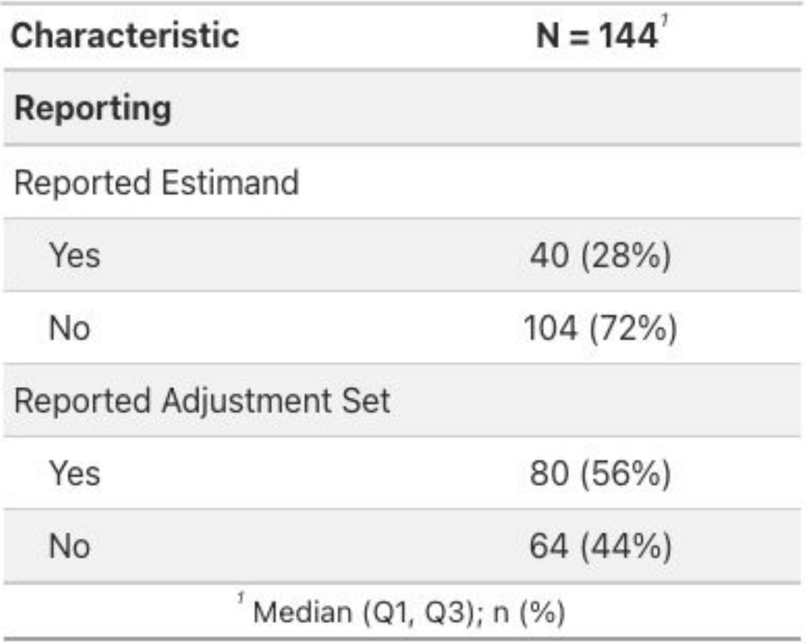


Adapted from
Tennant et al

# Inclusion and reporting of DAGs varies



| Characteristic | N = 144[1] |
|---|---|
| **DAG properties** | |
| Number of Nodes | 12 (9, 16) |
| Number of Arcs | 29 (19, 42) |
| Node to Arc Ratio | 2.30 (1.75, 3.00) |
| Saturation Proportion | 0.46 (0.31, 0.67) |
| Fully Saturated | |
| Yes | 4 (3%) |
| No | 140 (97%) |



| Characteristic | N = 144[1] |
|---|---|
| **Reporting** | |
| Reported Estimand | |
| Yes | 40 (28%) |
| No | 104 (72%) |
| Reported Adjustment Set | |
| Yes | 80 (56%) |
| No | 64 (44%) |

[1] Median (Q1, Q3); n (%)

Adapted from Table 4.1 in “Expressing causal questions as DAGs”, *Causal Inference in R*
And Tennant et al

# Tennant et al recommendations

1. Focal relationship(s) and estimand(s) of interest should be stated in study aims
2. DAG(s) for each focal relationship and estimand of interest should be made available
3. DAGs should include all relevant variables, including those w/o direct measurement
4. Variables should be visually arranged so arrows go in the same direction
5. By default an arrow should be assumed to exist between two variables
6. DAG-implied adjustment set(s) for every estimand should be explicitly stated
7. Estimate(s) from unmodified DAG-implied adjustment set(s) should be reported
8. Alternative adjustment set(s) should be justified, with estimates reported separately

# Malcolm et al recommendations

1. Build a DAG before conducting a study, and then iterate early and often
2. Explicitly consider/state important question-specific considerations
    a. Clearly define target estimand
    b. Consider how your DAG may vary over {time, space, populations} as relevant
    c. What confounders are relevant and if/when they occur
3. Order nodes by time
4. Consider the data collection process in addition to the causal structure of your question
5. Include unmeasured variables
6. Saturate your DAG and then prune
7. Include instrumental and precision variables
8. Focus on the causal structure, and then consider measurement bias
9. Pick adjustment sets most likely to be successful
10. Use robustness checks

# Other questions/topics to discuss?