How do we build a model?

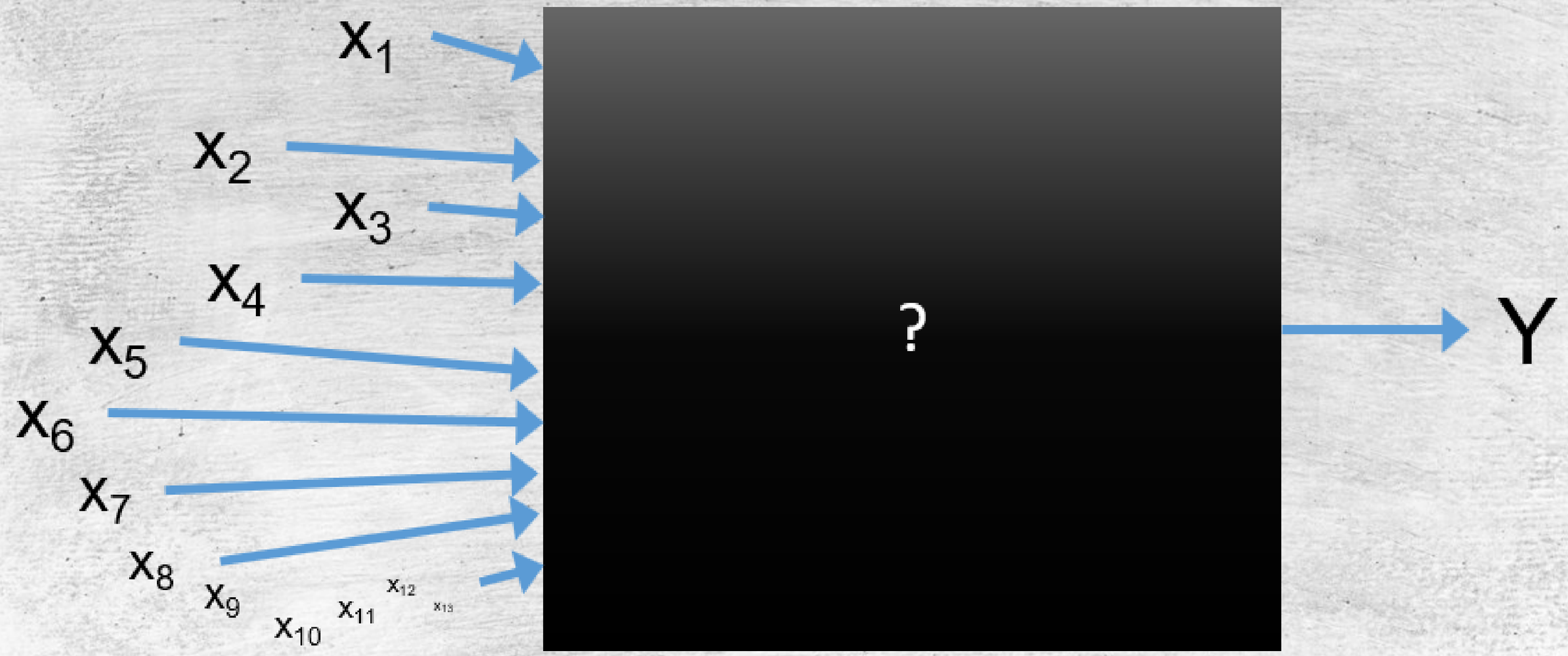
Variable selection and causality

Julien Martin
University of Ottawa

2024-02-05

Why to we build models?

Data generating process



Interpreting relationships in data

- When we say there's a relationship between two variables... how do we interpret that?
- What precisely do we mean?
- What do we want to do with this information?

Distinguishing goals of data analysis

Descriptive: Document or quantify observed relationships between inputs and outputs.

- Does not not necessarily tell us about the true DGP.
- Can often inspire questions for further research.

Causal: Learn about causal relationships.

- Try to understand how the box works (the true DGP)
- When you change one factor, how does it change the result?

Predictive: Be able to guess the value of one variable from other information

- DGP doesn't matter, create your own box..
- Helps us know what's likely to happen in a new situation.

Difference of Focus

Description:

- Focus on showing relationships among a few variables.
- **Give up goal of correctly modeling the true DGP**

Prediction:

- Focus on predicting given observed data by any possible means.
- **Give up goal of correctly modeling the true DGP**

Causal inference:

- Focus on determining the true direct effect of a treatment variable
- Give up goal of understanding causal effects of any other factors

Impact of selection bias

Description

- NO. Only want to infer patterns from observed data.

When there are a lot of people wearing shorts, there often is an ice cream truck

Prediction:

- NO. Only want to infer patterns from observed data.

Given how many people are wearing shorts, will an ice cream truck show up?

Causal inference:

- YES. Want to infer the result of active intervention. Must eliminate selection bias to estimate the treatment effect.

If someone chooses to wear shorts, will it make an ice cream truck show up?

Difference Interpretation of β_n

Description:

- β_n represents an association between X_{n_i} and Y_i.
- Only a statement about the data, not about the reasons behind the pattern.

Prediction: Model does not need to be interpretable.

- Coefficients β_n are informative only of predictive power, not causal effects.
- Model can be treated as a black box.

Causal inference:

- β_1 is a causal effect of x_1 under stated assumptions (of the identification strategy).
- Many coefficients generally lack interpretability.

Consequences

Consequences

Discerning which type of goal you have is critical for:

- **Interpreting results**: Mistaking one goal for another can lead your audience to make very bad decisions.
- **Choosing methods**: Distinct approaches are required to achieve different goals.

Consequences for models

Models for prediction and causal inference differ with respect to the following:

1. The covariates that should be considered for inclusion in (and possibly exclusion from) the model.
2. How a suitable set of covariates to include in the model is determined.
3. Which covariates are ultimately selected, and what functional form (i.e. parameterization) they take.
4. How the model is evaluated.
5. How the model is interpreted.

Consequences for methods ?

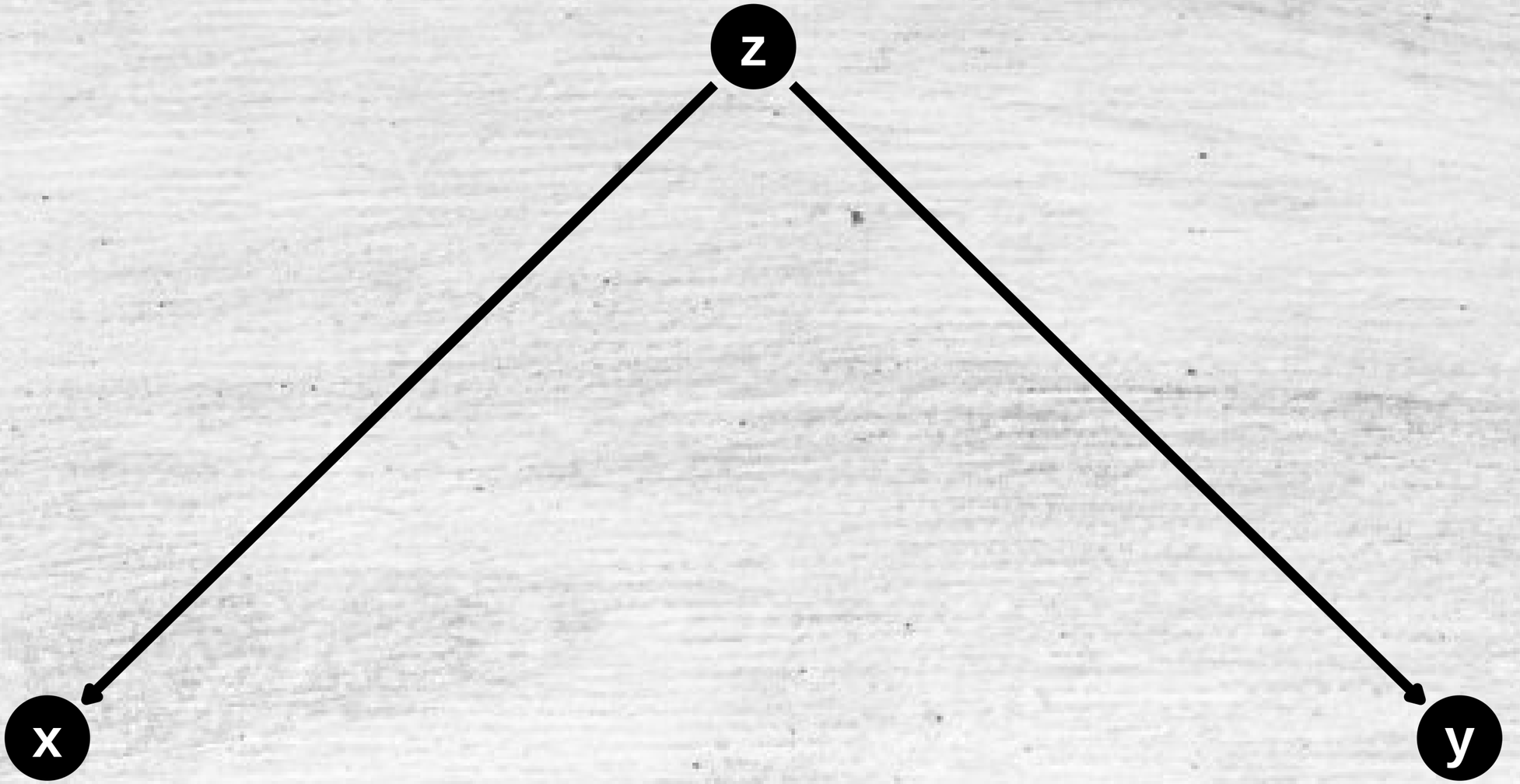
What methods should we use for each goal?

1. ***Descriptive analysis***
    - Exploratory analysis and regression.
2. ***Causal inference***
    - Path analysis
    - Structural equation modelling
    - Graph theory
3. ***Prediction***
    - Statistical learning / machine learning.
    - AIC and any kind of model selection

How to figure it out?

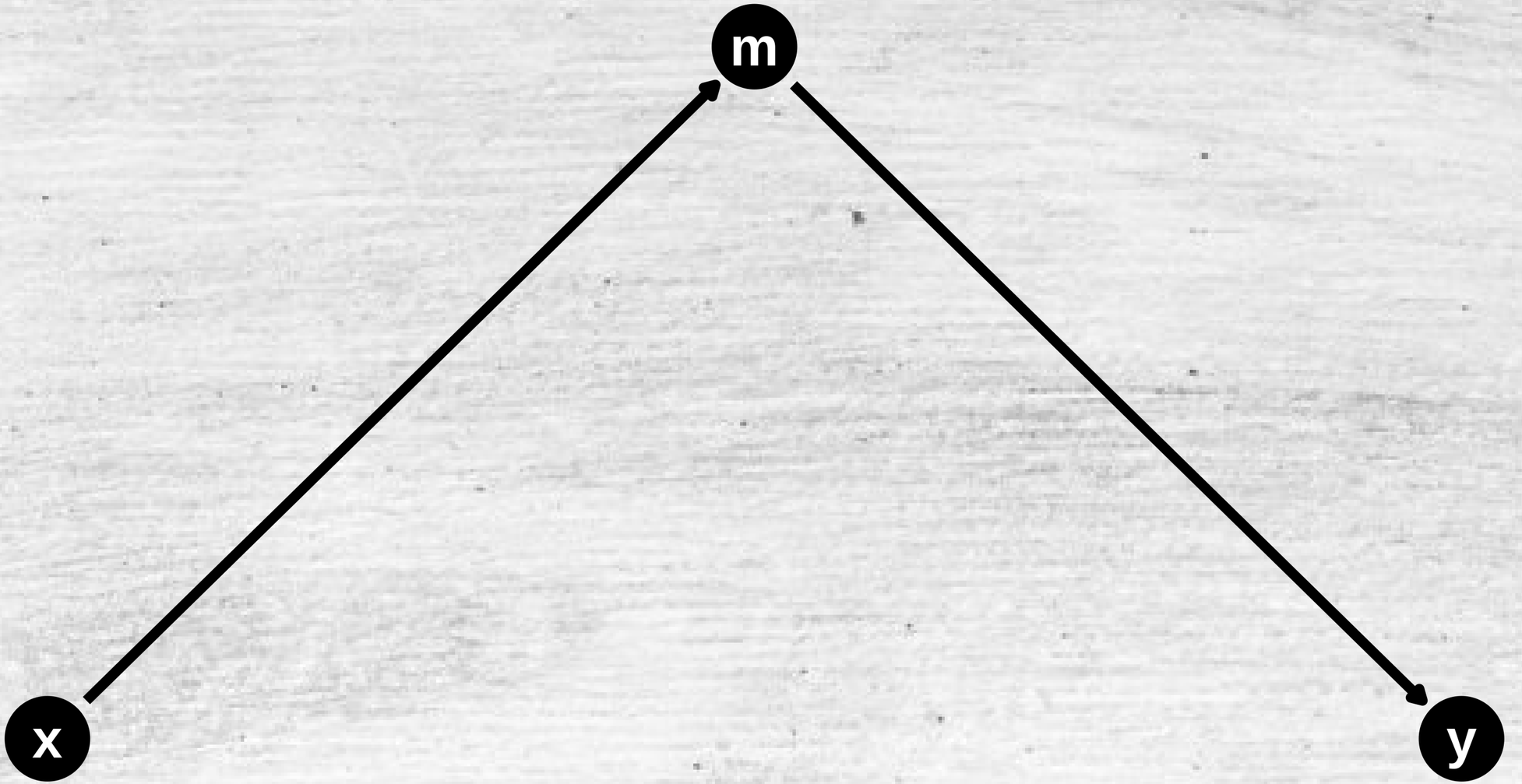
Confounder

Relation · Adjusted


z
x
y

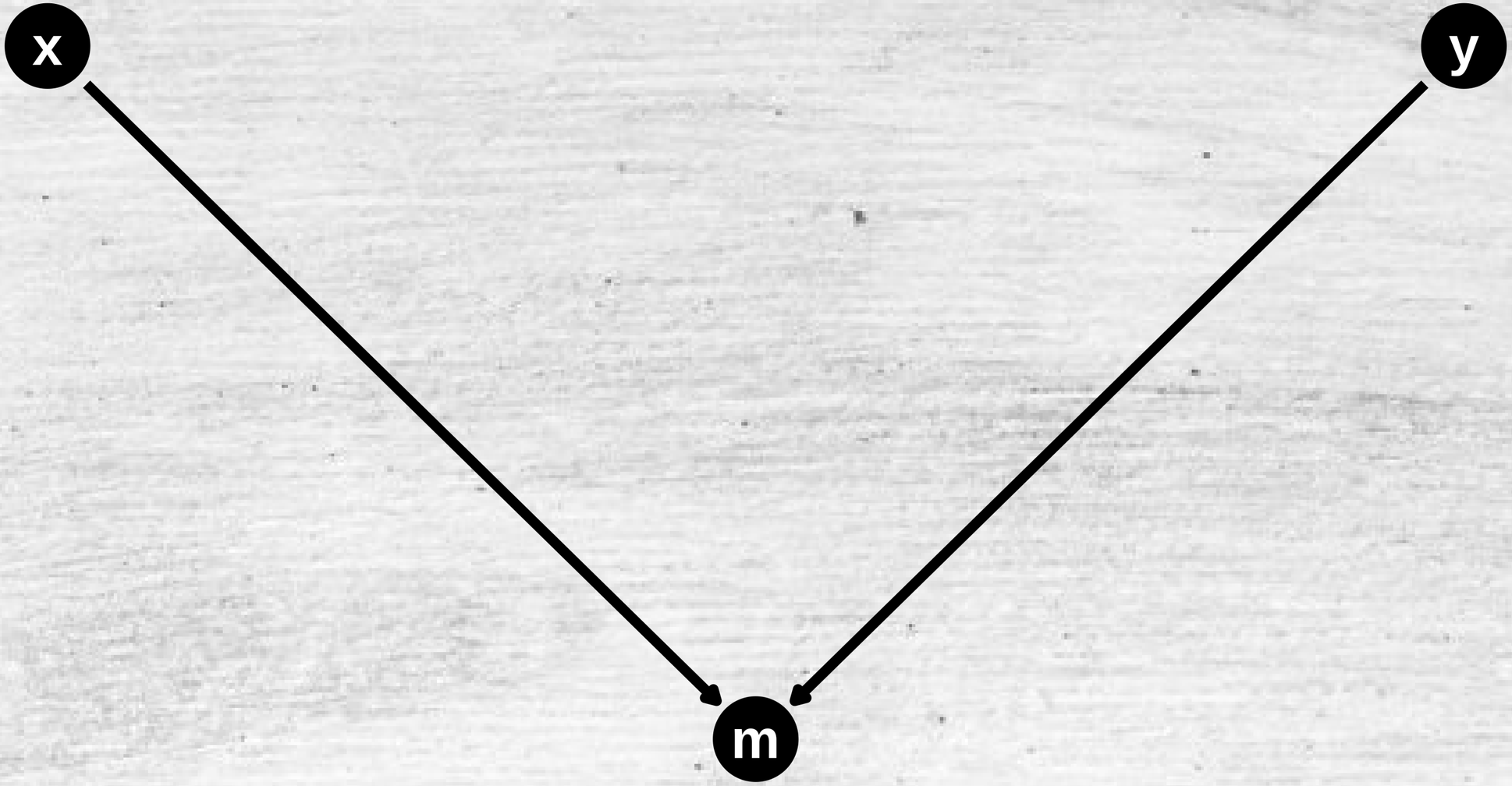
Mediator

Relation · Adjusted


m
x
y

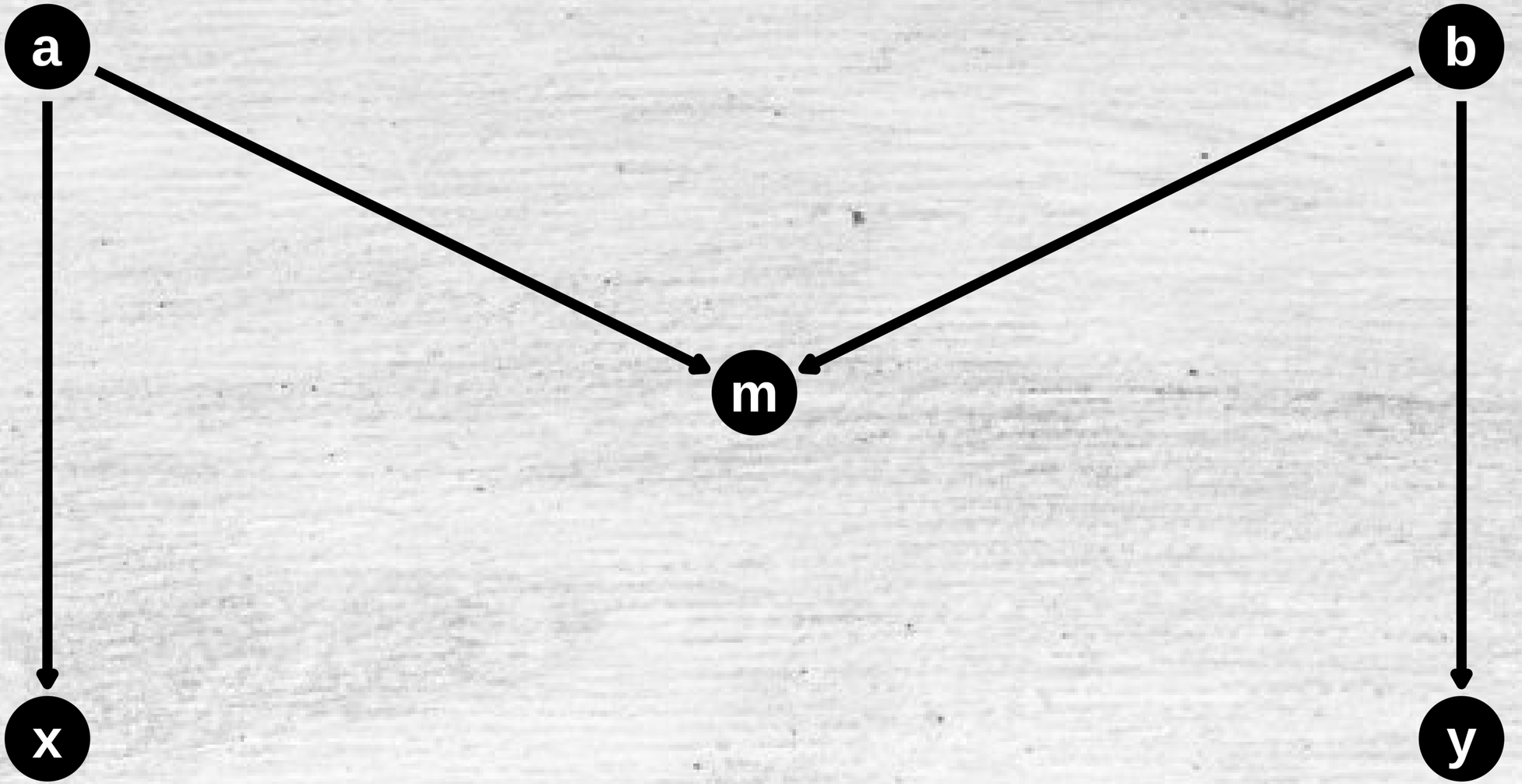
Collider

Relation · Adjusted


x
y
m

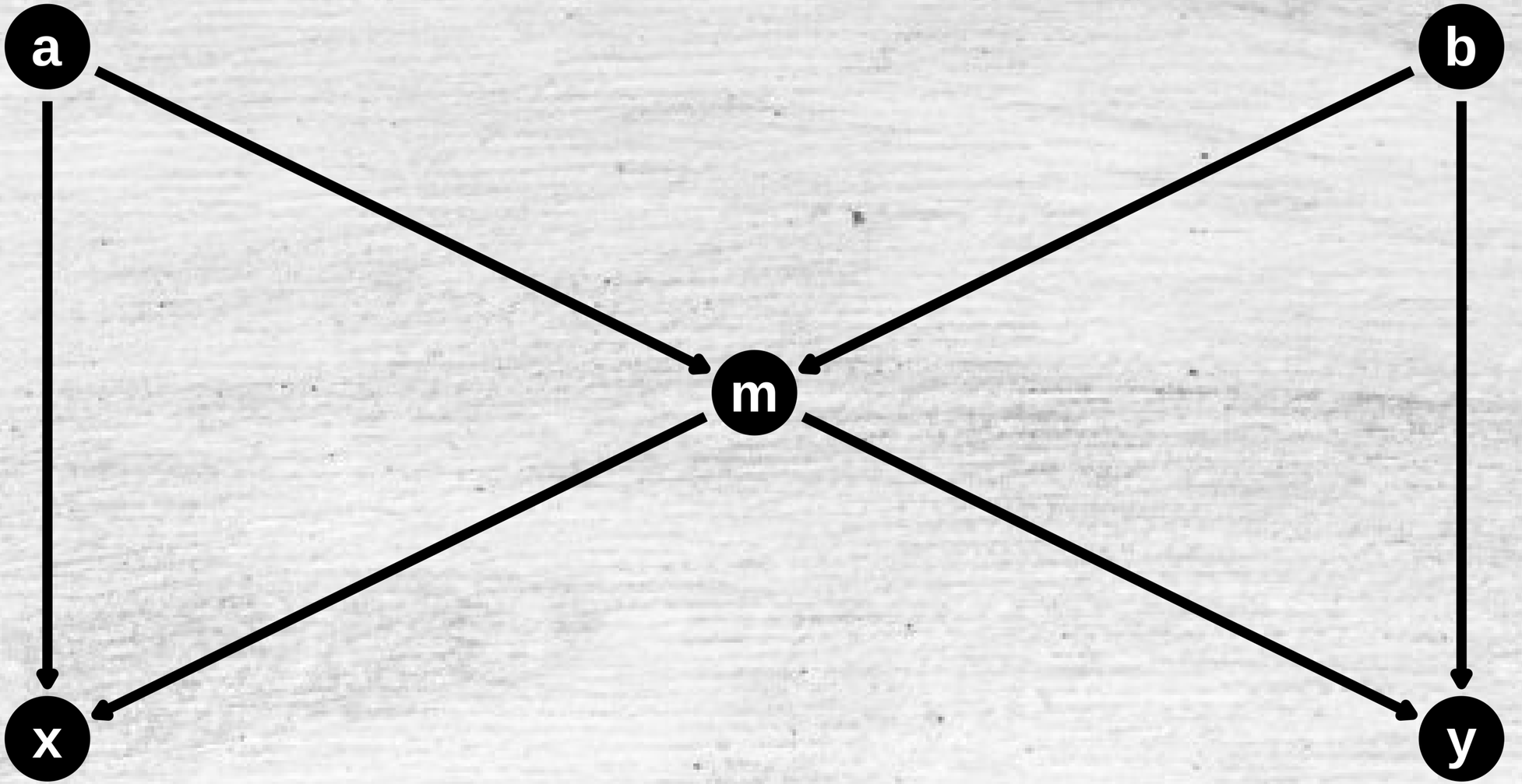
M-bias

Relation · Adjusted


a
b
m
x
y

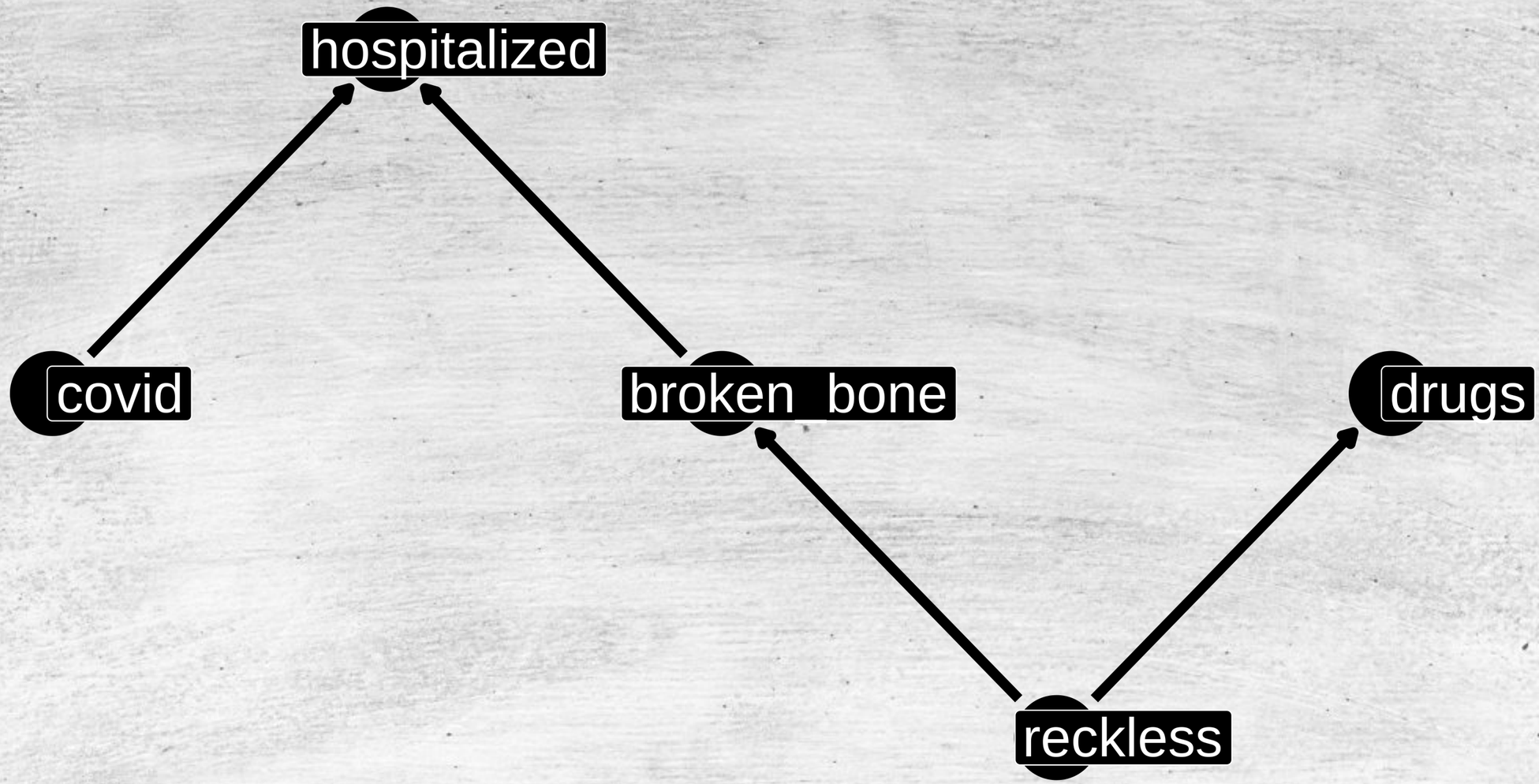
Butterfly bias

Relation	{a, m}	{b, m}	{a, b, m}


a
b
m
x
y

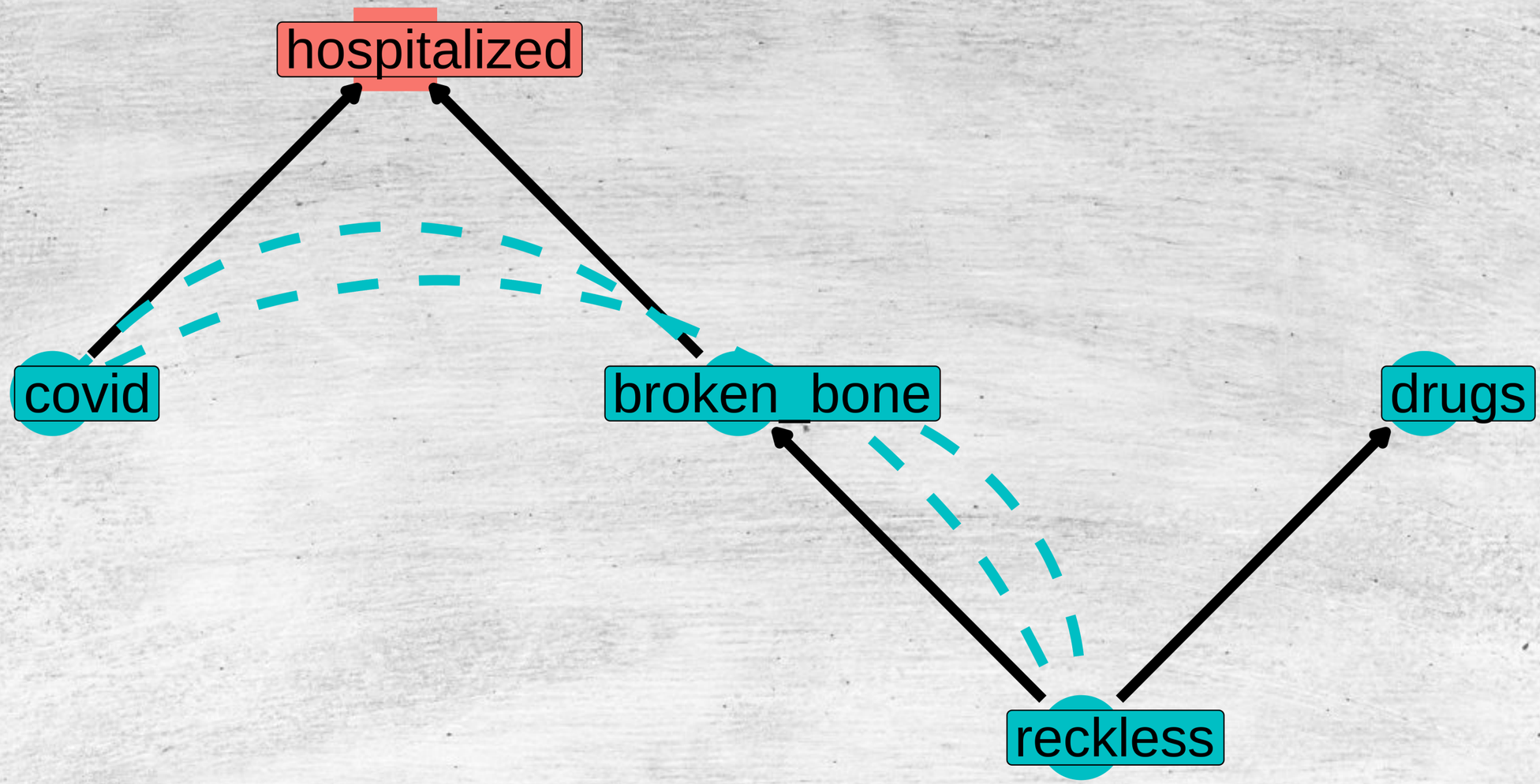
Selection bias

Case



Selection bias

Controlled



More complexity

Code | Plot

```
my_dag <- dagify(y ~ x + a + b,
  x ~ a + b,
  a ~ d,
  exposure = "x",
  outcome = "y"
)
my_dag %>%
  ggplot(aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_dag_point() +
  geom_dag_edges(edge_width = 2) +
  geom_dag_text(size = 50) +
  theme_dag()
```

Happy modelling