Causality

Data Analytics and Visualization with R
Session 4

Viktoriia Semenova

University of Mannheim

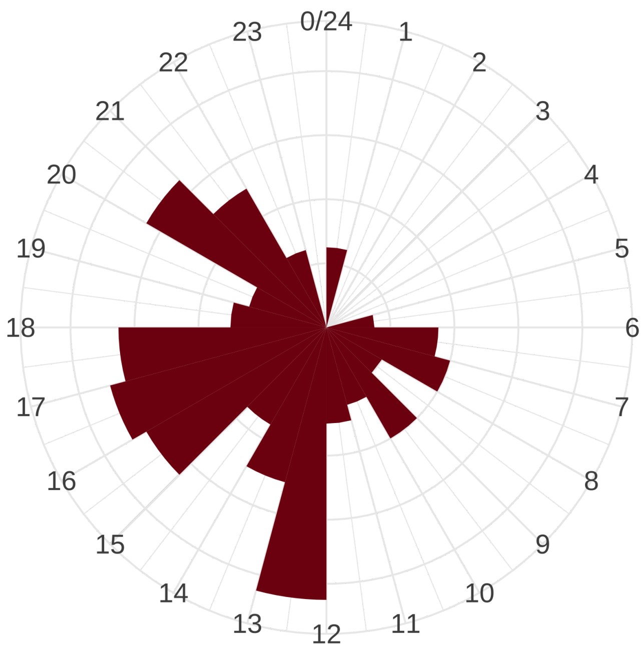
Spring 2023

🔗 r4da.live

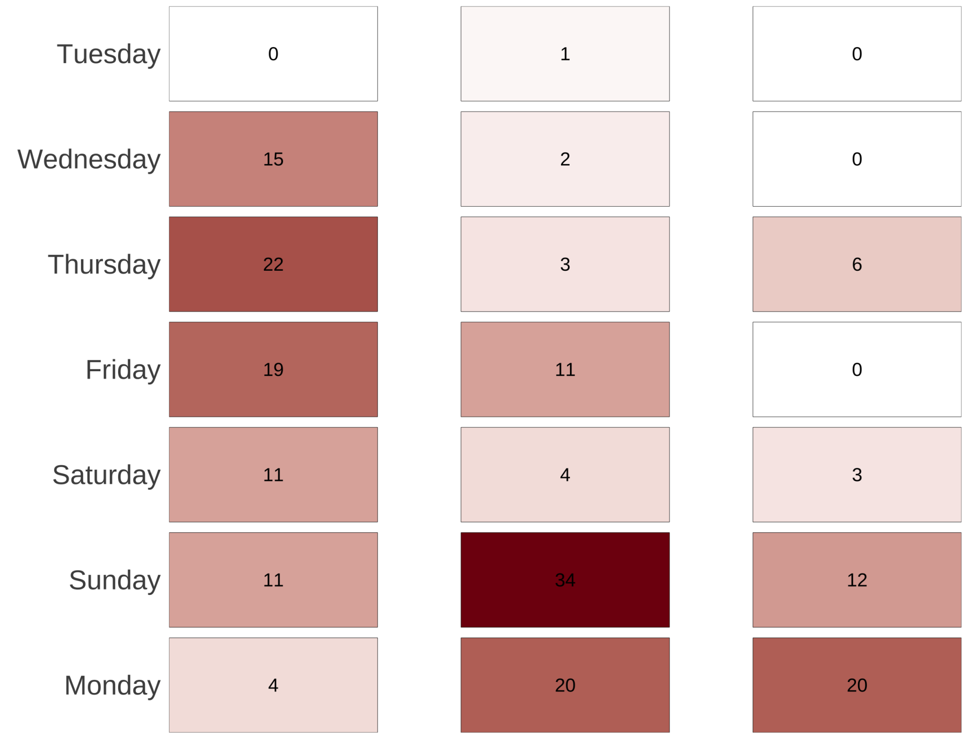
Warm Up

Your GitHub Stats 🤓

Commits to Problem Set 3 by Time of Day



Commits to Problem Sets by Week Day

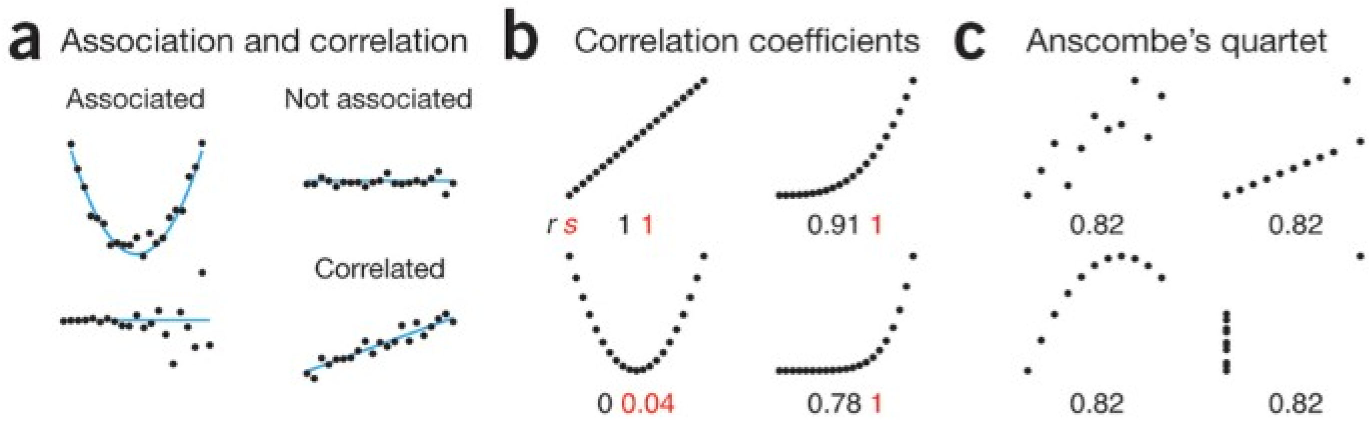


Quiz: Which of these statements are correct?

1. Regression line represents a conditional mean of the explanatory variable X given the value of the outcome variable Y.
2. Extreme values of correlation coefficient (i.e. close to -1 or 1) imply that there is a large substantive effect of X on Y.
3. Correlation between X and Y implies there is a causal relationship between them.
4. Causal relationship between X and Y implies there is a correlation between them.
5. Causal relationship between X and Y implies there is an association between them.

04:00

Association vs. Correlation



Correlation is a type of association and measures increasing or decreasing trends quantified using correlation coefficients.

Causality

Data Generating Process

- An unknown process in the real world that “generates” the data we are interested in
- In social sciences, DGP is often not very precise
- Our understanding of DGP comes from the theory and subject knowledge

Causality

- A variable X is a cause of a variable Y if Y in any way relies on X for its value.... X is a cause of Y if Y listens to X and decides its value in response to what it hears (Pearl, Glymour, and Jewell 2016, 5–6)
- This incorporates:
  - association between X and Y
  - time ordering: cause precedes outcome
  - nonspuriousness: there is plausible relationship
- **Causal effect** is the change in variable Y that would result from a change in variable X

Example: Boston Commuters Experiment

Causal Effects & Counterfactuals

- Two **potential outcomes:**
  - $Y_i(1)$: would commuter i report pro-immigration attitudes if exposed to Spanish-speakers ($T = 1$)?
  - $Y_i(0)$: would commuter i report pro-immigration attitudes if **not** exposed to Spanish-speakers ($T = 0$)?
- **Causal effect:** $Y_i(1) - Y_i(0)$ (aka **treatment effect**)
  - $Y_i(1) - Y_i(0) = 0$: exposure to Spanish-speakers has no impact on attitudes
  - $Y_i(1) - Y_i(0) = +1$: exposure to Spanish-speakers leads to pro-immigration attitudes
  - $Y_i(1) - Y_i(0) = -1$: exposure to Spanish-speakers leads to anti-immigration attitudes

Potential Outcomes

	Attitude if Treated	Attitude if Control
Jack	Pro-immigration	Anti-immigration

More formally:

	$Y_i(1)$	$Y_i(0)$
Jack	1	0

Fundamental Problem of Causal Inference

			Causal Effect
	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
Jack	1	0	1

- We cannot observe $Y_i(1) - Y_i(0)$ in real life though:
  - We only observe one of the two potential outcomes $Y_i(1)$ or $Y_i(0)$
  - To infer causal effect, we need to infer the missing *counterfactuals*

Multiple Units

	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
Jack	1	0	1
Dan	0	0	0
Anne	1	0	1
Yao	0	0	0
Judy	0	1	-1

- Individual treatment effects: value of $Y_i(1) - Y_i(0)$ for each i
- **Average treatment effect:** mean of all the individual causal effects $ATE = \frac{1+0+1+0+(-1)}{5} = 0.2$

Back to Real World...

	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
Jack	?	0	?
Dan	0	?	?
Anne	1	?	?
Yao	0	?	?
Judy	?	1	?

Randomized Experiment as a Solution

- Each unit's treatment assignment is determined by chance
- Randomization ensures balance between treatment and control group:
  - they are identical *on average*
  - we shouldn't see large differences between treatment and control group on *pretreatment* variable

ATE vs. Difference-in-Means

We want to estimate the average causal effects over all units:

$$\text{Average Treatment Effect} = \frac{\sum_{i=1}^{n}(Y_i(1) - Y_i(0))}{n}$$

But we can only estimate instead:

$$\text{Difference in means} = \overline{Y}_i(1) - \overline{Y}_i(0)$$

This is a pretty good estimate of ATE if randomization worked!

Casual Diagrams

Directed Acyclic Graphs (DAGs)

Nodes: variables in the DGP

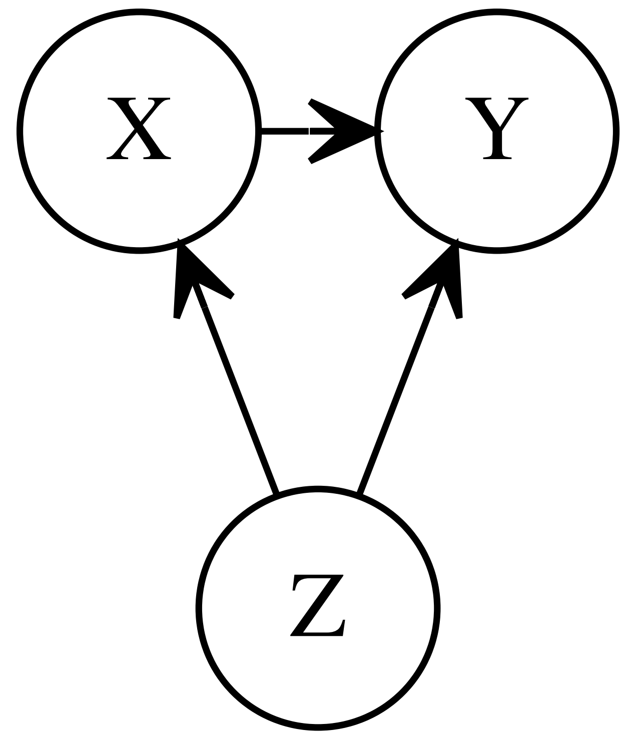
Arrows: causal relationships in the DGP (associations)

Direction: from the cause variable to the caused variable

Directed: Each **node** has an arrow that points to another node

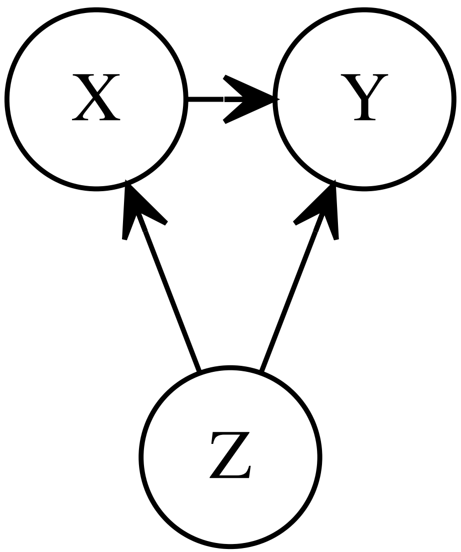
Acyclic: You can't cycle back to a node (and arrows only have one direction)

Graph: Well…it is a graph.



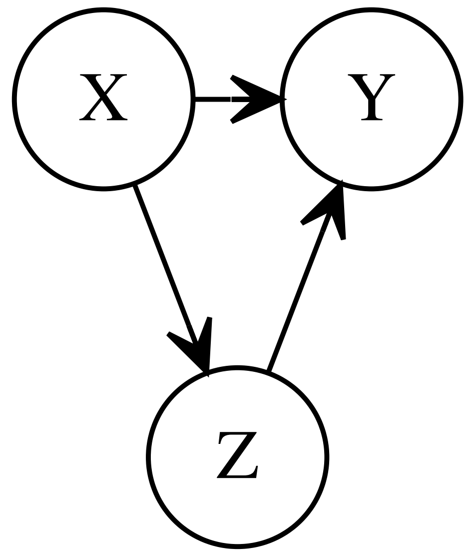
Major Types of Association

Confounding (Fork)



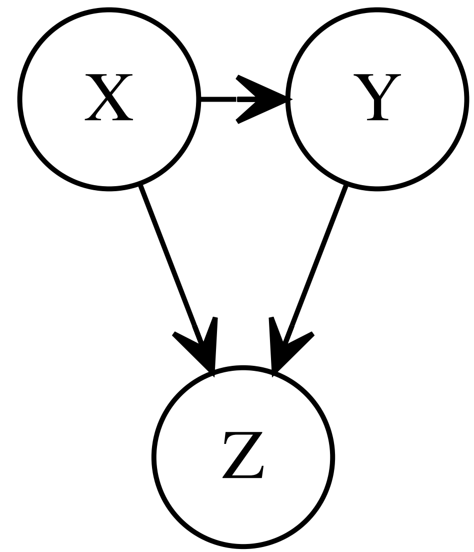
Common cause

Causation (Chain)



Mediation

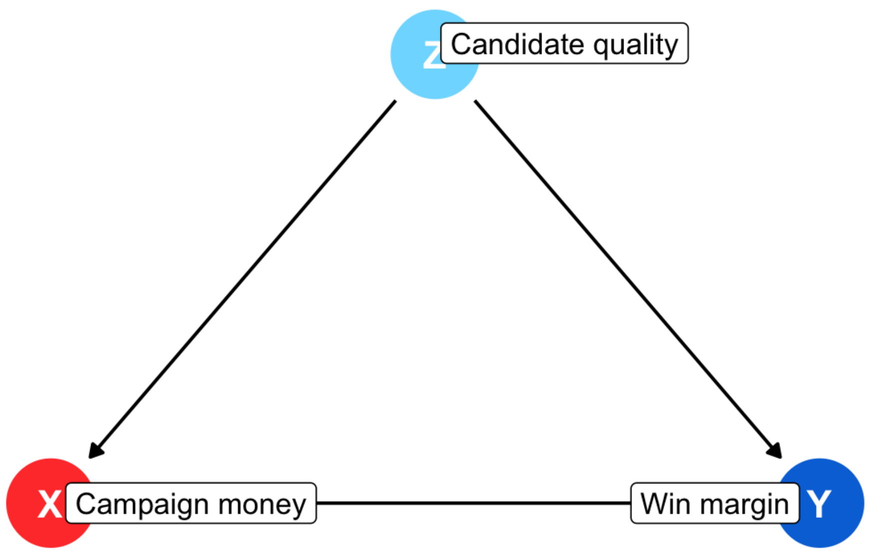
Collision (Inverted Fork)



Selection / endogeneity

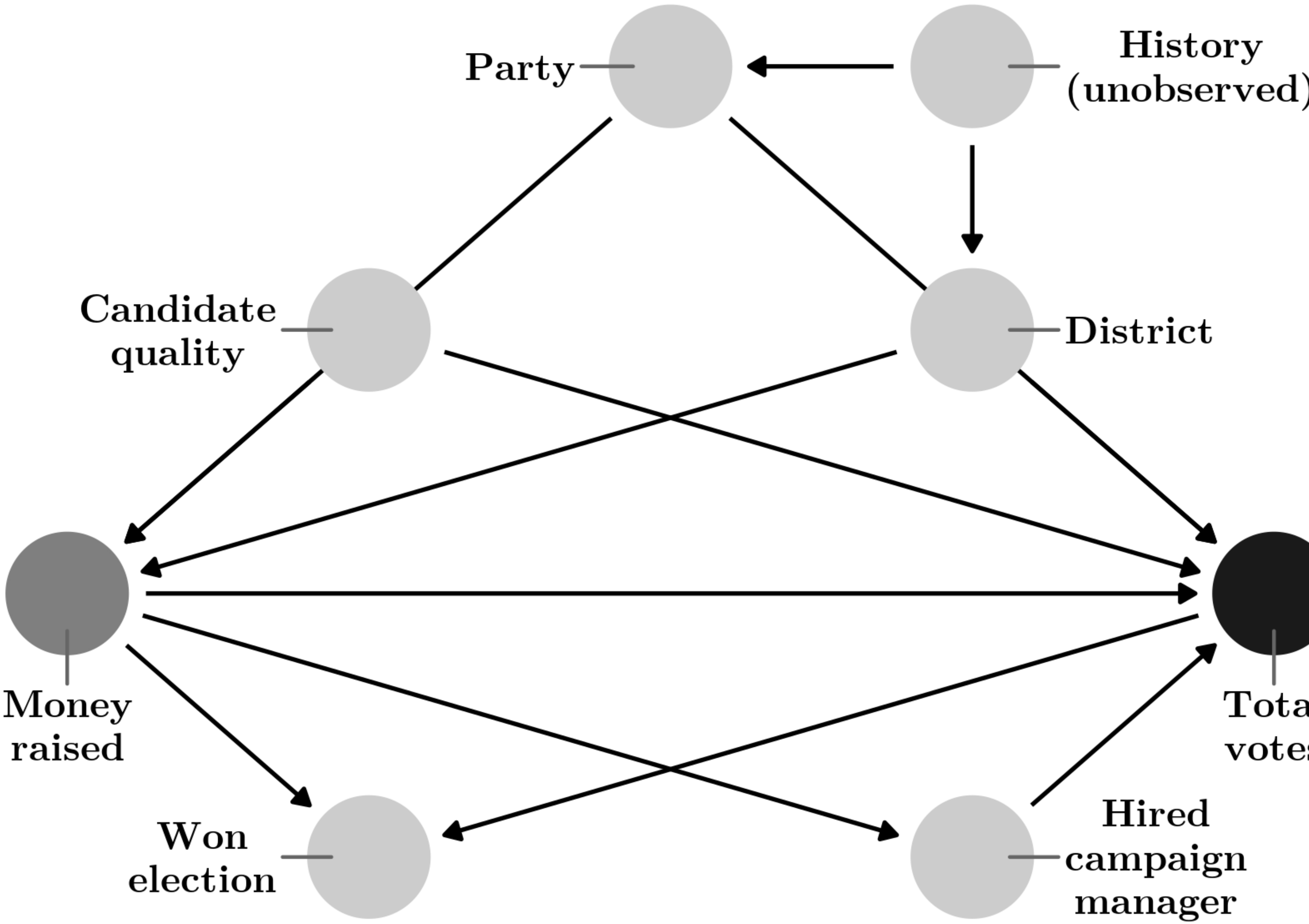
Confounding

Effect of money on elections



1. Find the part of campaign money that is explained by quality, remove it.
2. Find the part of win margin that is explained by quality, remove it.
3. Find the relationship between the residual part of money and residual part of win margin. This is the *causal effect*.

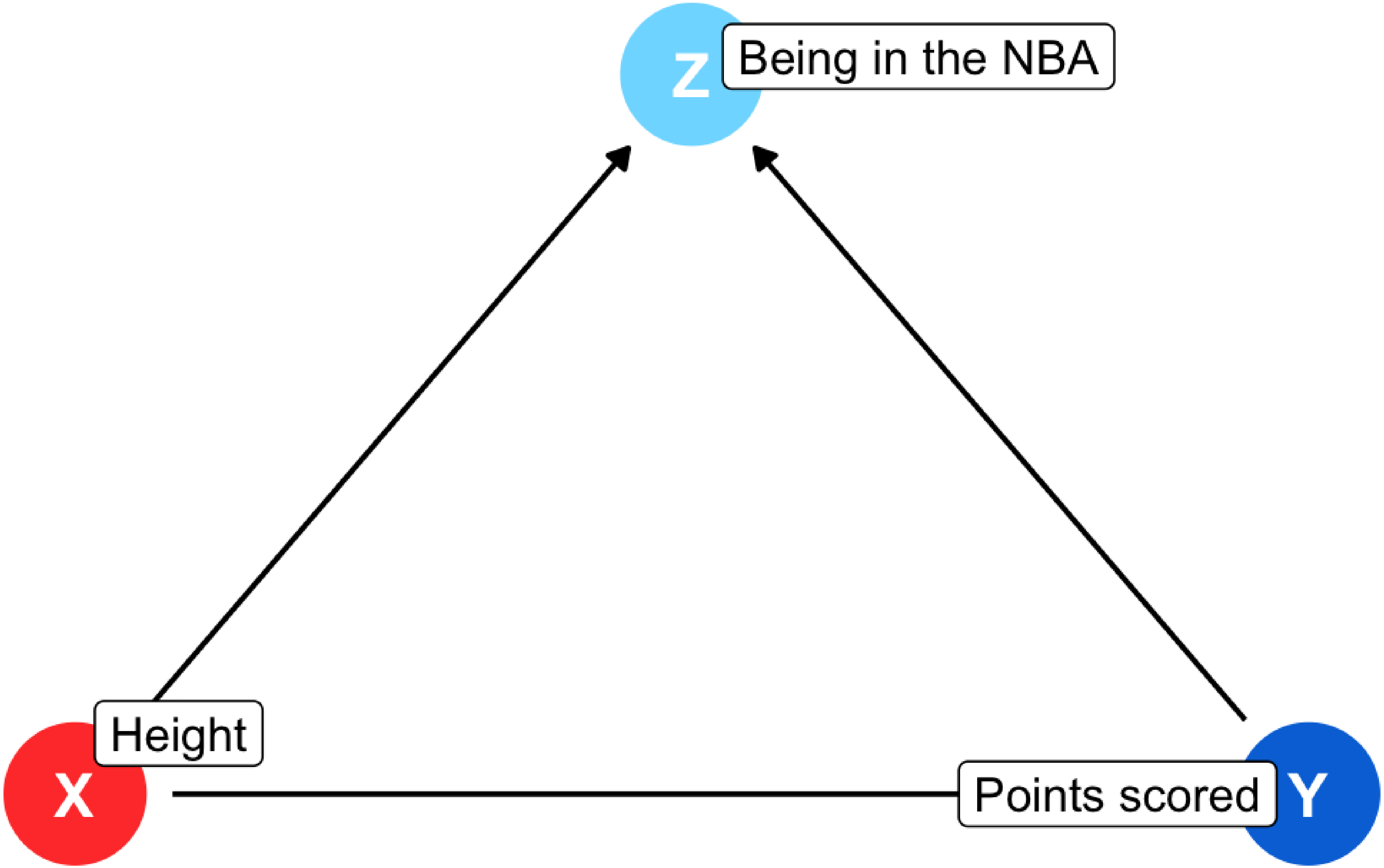
Campaign Example



Party
History
(unobserved)
Candidate
quality
District
Money
raised
Total
votes
Won
election
Hired
campaign
manager

Collider

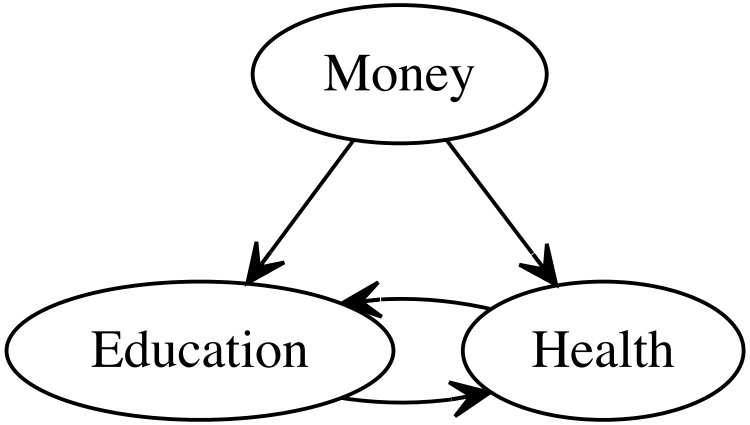
Height is unrelated to basketball skill… among NBA players



Colliders can create fake causal effects

Causal Identification

- DAGs help us with the process of identification
- Causal effect is *identified* if the association between treatment and outcome is properly stripped and isolated
- Identification implies that:
  - All alternative stories are ruled out
  - We have enough information to answer a specific causal inference question
- Sometimes we cannot identify the effect with our data alone


Money
Education
Health

Studying Example