DAGs and potential outcomes

February 12, 2020

PMAP 8521: Program Evaluation for Public Service
Andrew Young School of Policy Studies
Spring 2020

Fill out your reading report on iCollege!

Plan for today

Paths, doors, and adjustment

***do*()ing observational causal inference**

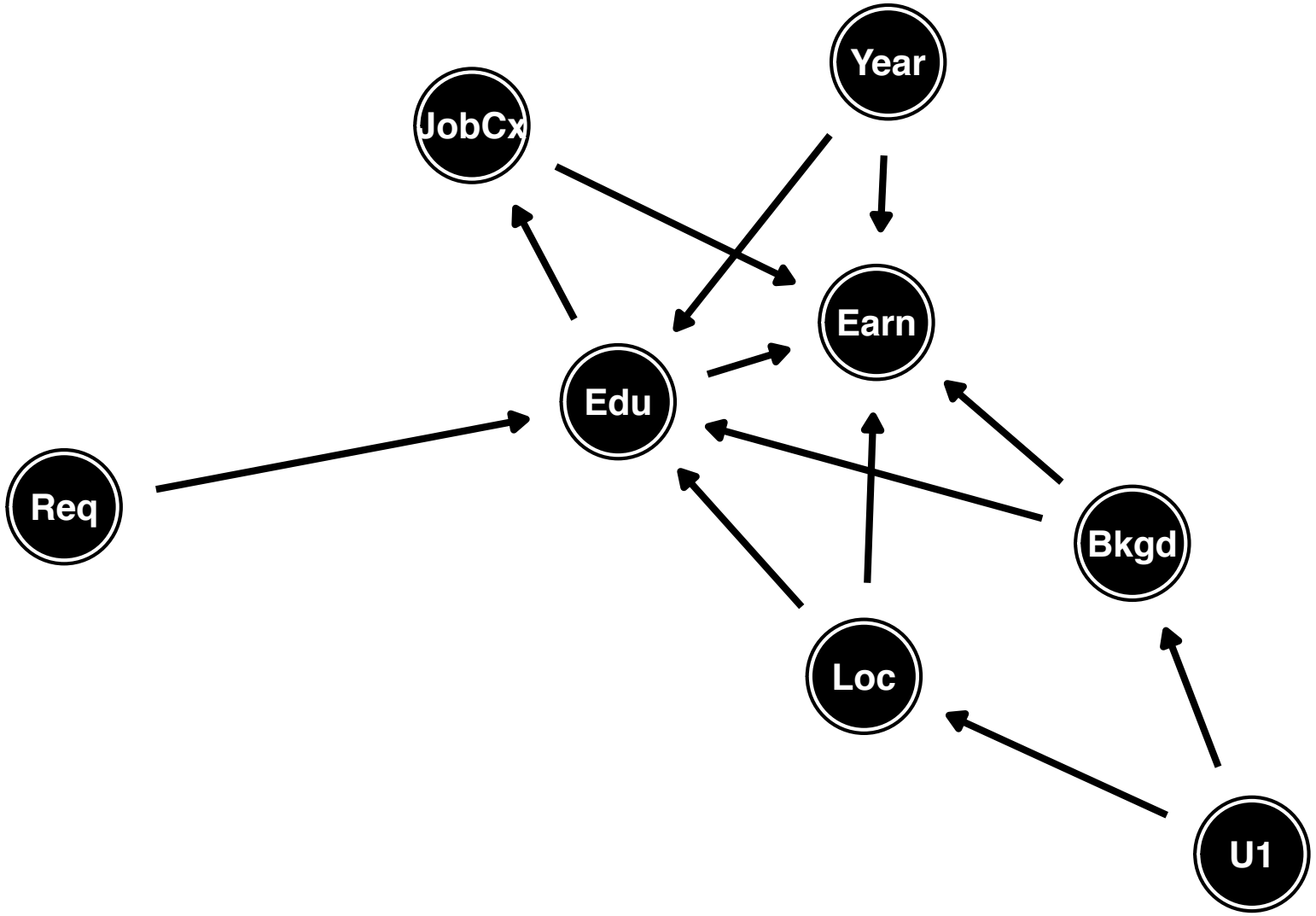
Potential outcomes

Paths, doors, and adjustment

Causal identification

All these nodes are related; there's correlation between them all

We care about Edu → Earn, but what do we do with all the other nodes?



Causal identification

A causal effect is "identified" if the association between treatment and outcome is properly stripped and isolated

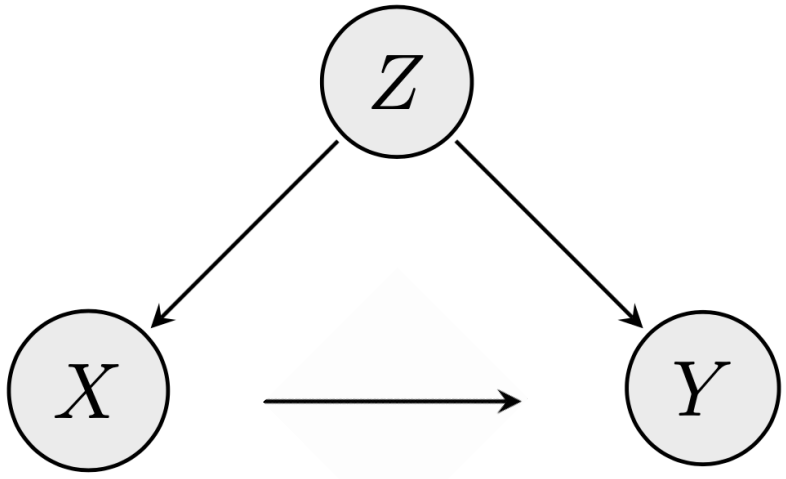
Paths and associations

Arrows in a DAG transmit associations

You can redirect and control those paths by “adjusting” or “conditioning”

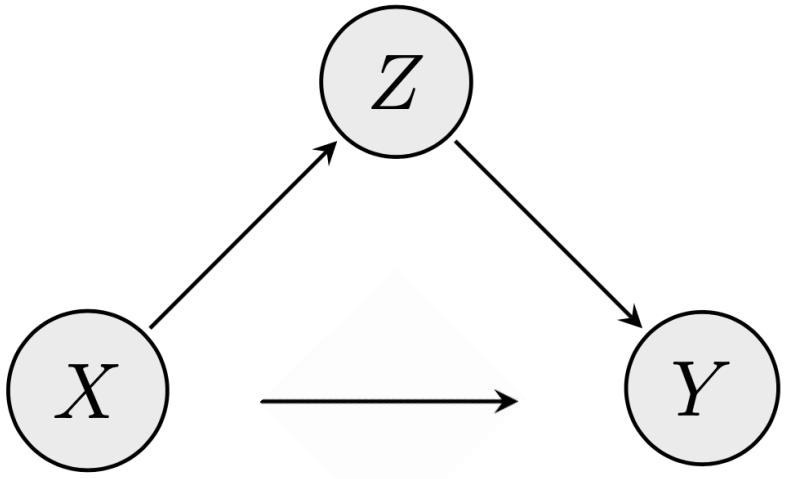
Three types of associations

Confounding



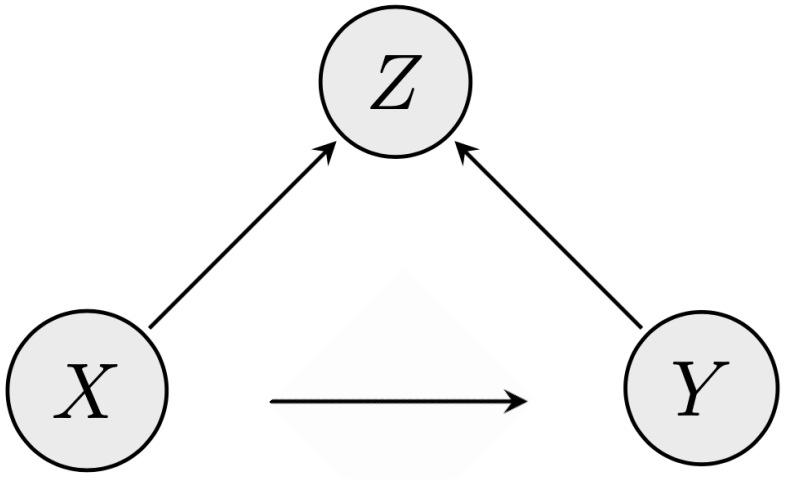
Common cause

Causation



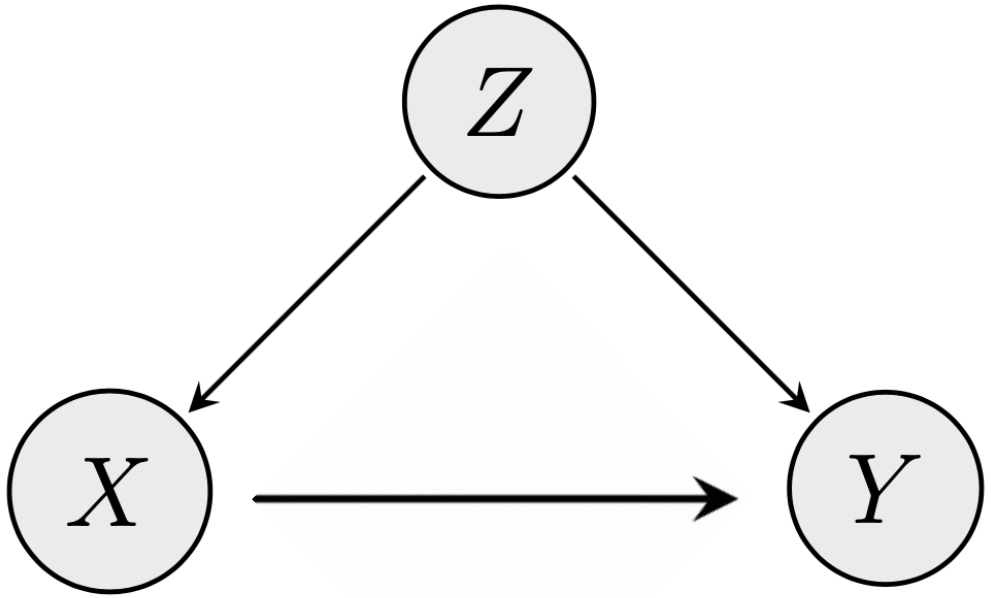
Mediation

Collision



Selection / Endogeneity

Confounding

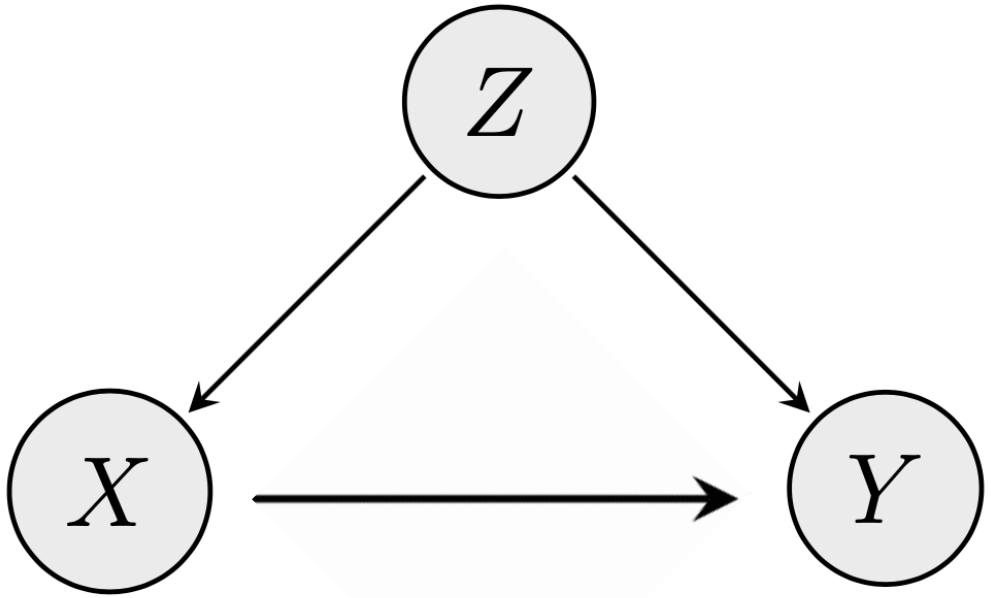


X causes Y

But Z causes both X and Y

Z confounds X → Y association

Paths



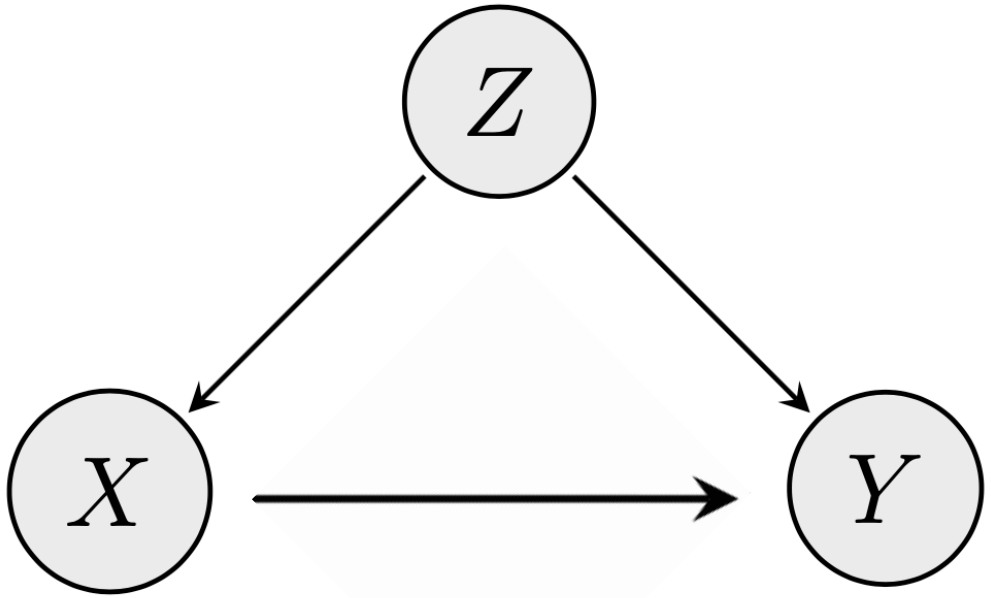
Paths between X and Y?

X → Y

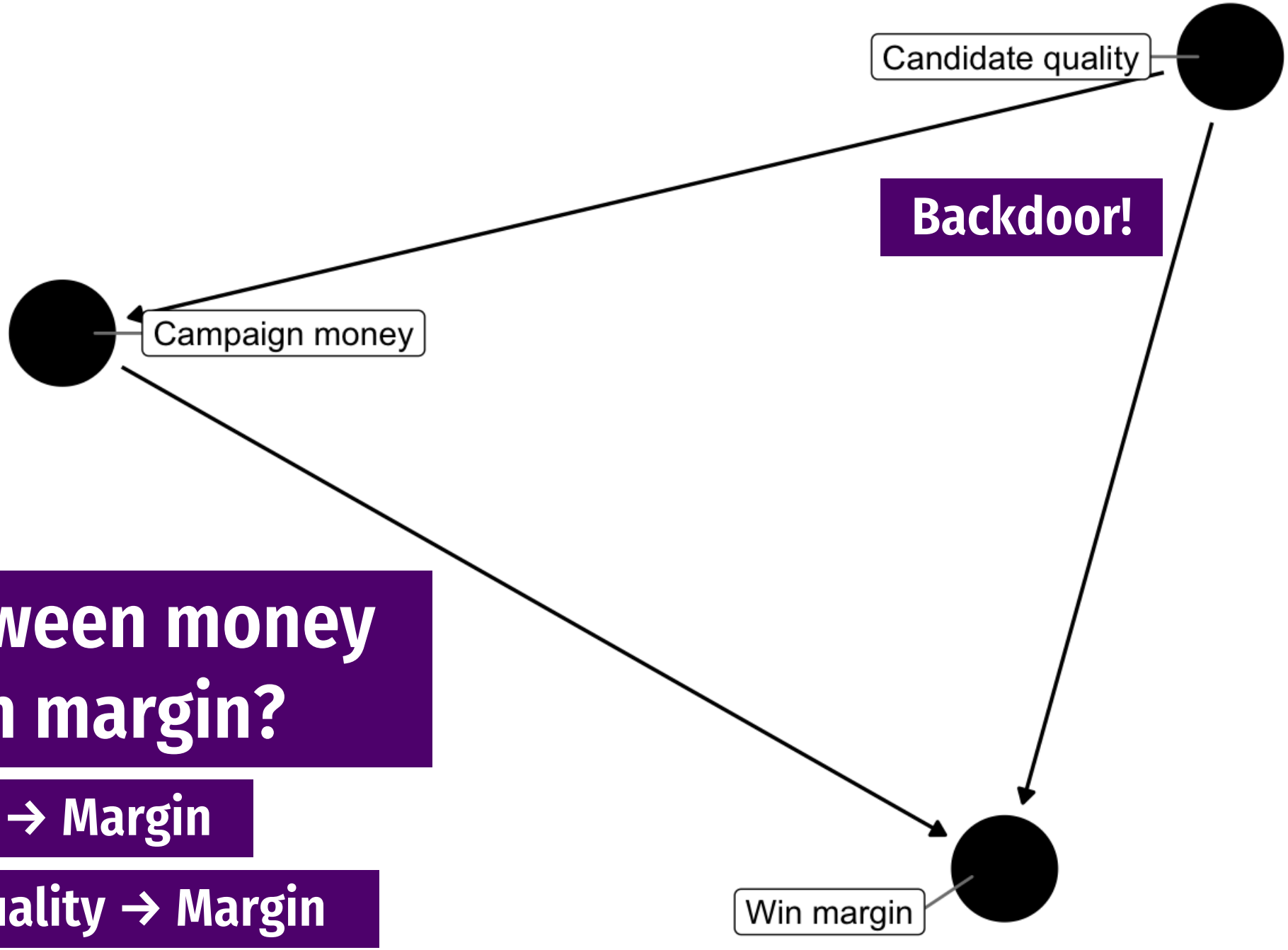
X ← Z → Y

Z is a backdoor

d-connection



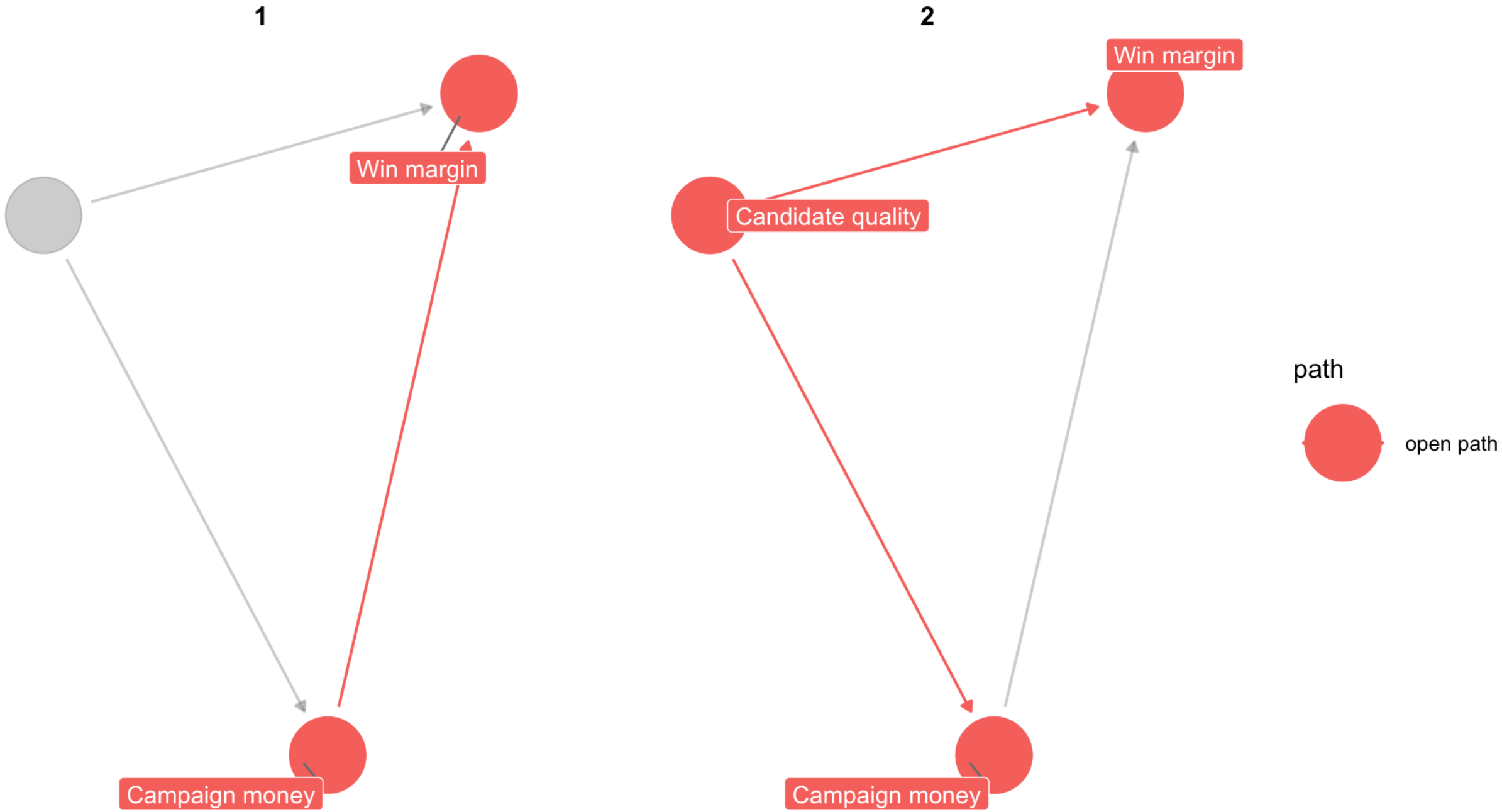
X and Y are "*d*-connected" because information can pass through Z



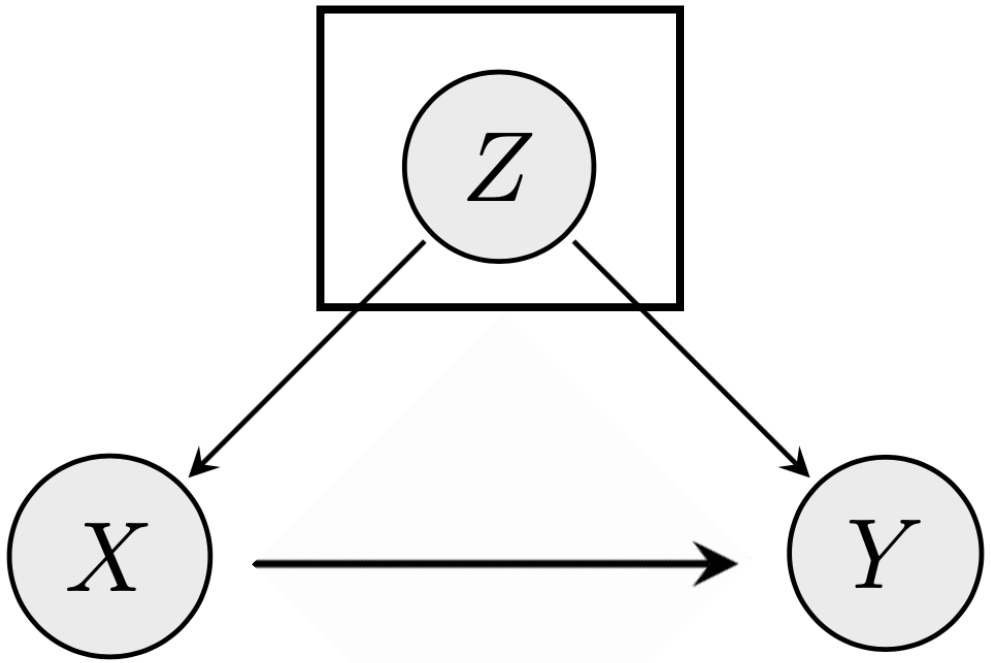
Paths between money and win margin?

Money → Margin

Money ← Quality → Margin


1
Win margin
Campaign money
2
Win margin
Candidate quality
Campaign money
path
open path

Closing doors

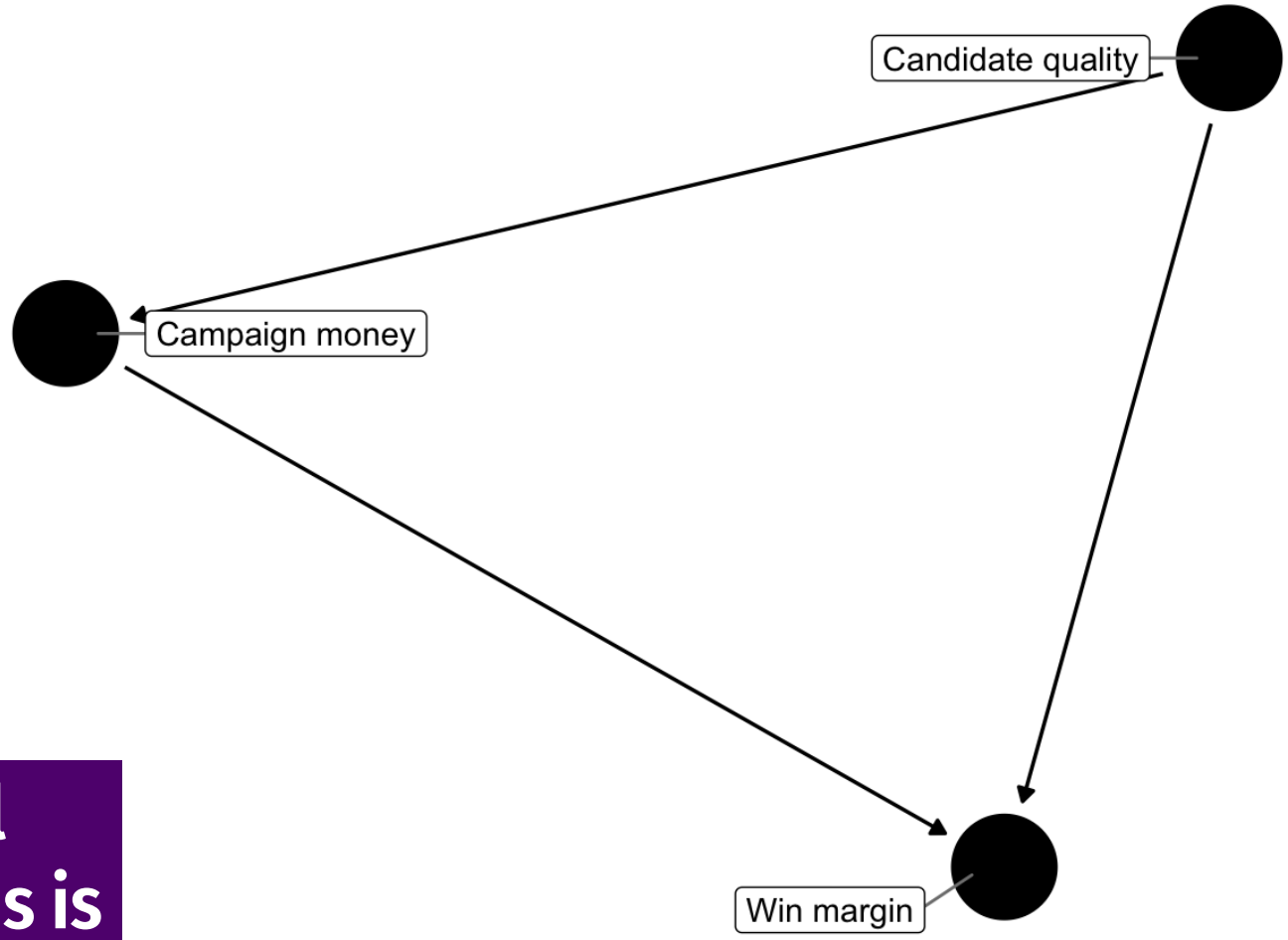


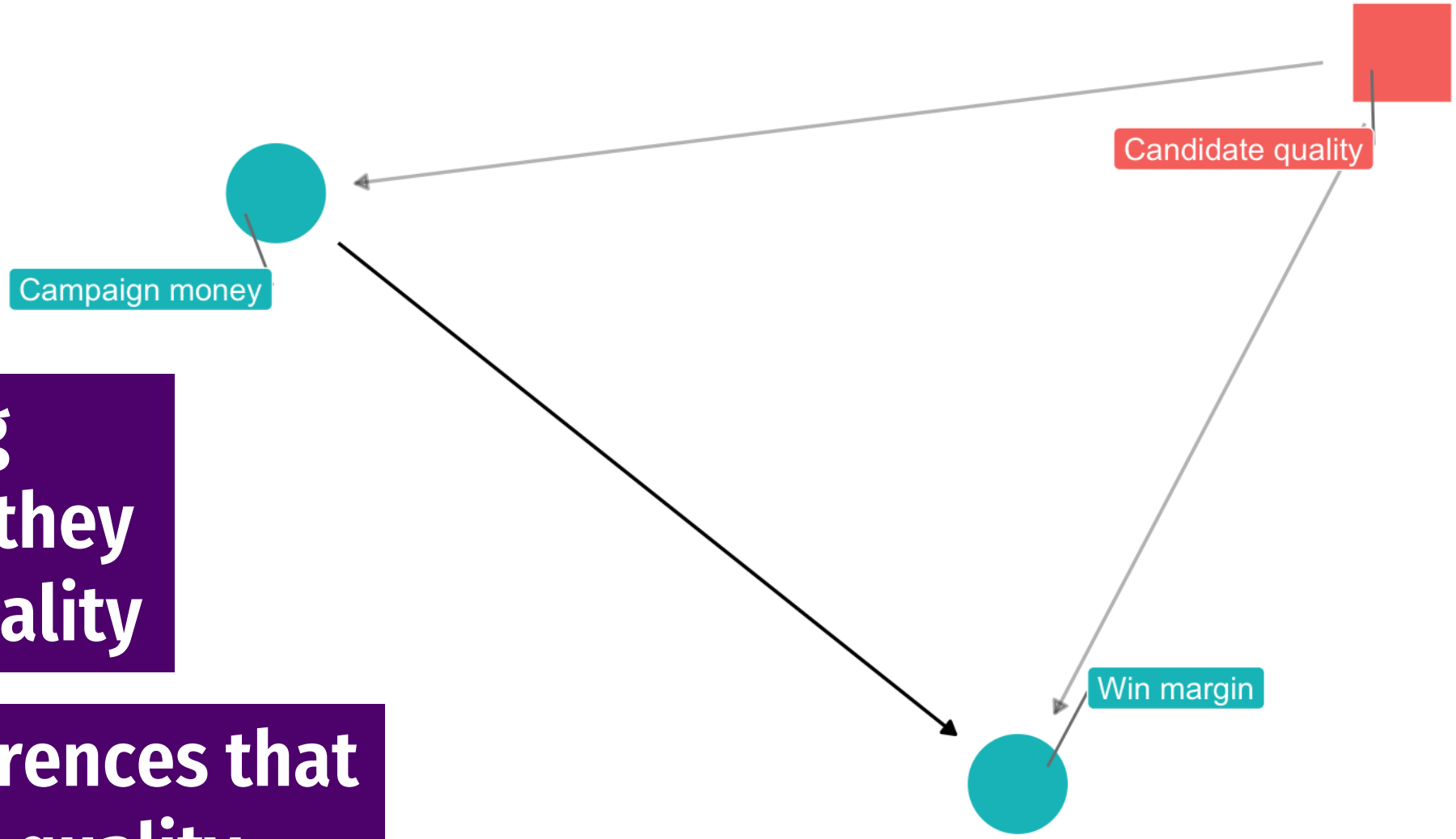
Close the backdoor by adjusting for Z

Find what part of X (campaign money) is explained by Q (quality), subtract it out. This creates the residual part of X.

Find what part of Y (the win margin) is explained by Q (quality), subtract it out. This creates the residual part of Y.

Find relationship between residual part of X and residual part of Y. This is the causal effect.





We're comparing candidates as if they had the same quality

We remove differences that are predicted by quality

Holding quality constant

How to adjust?

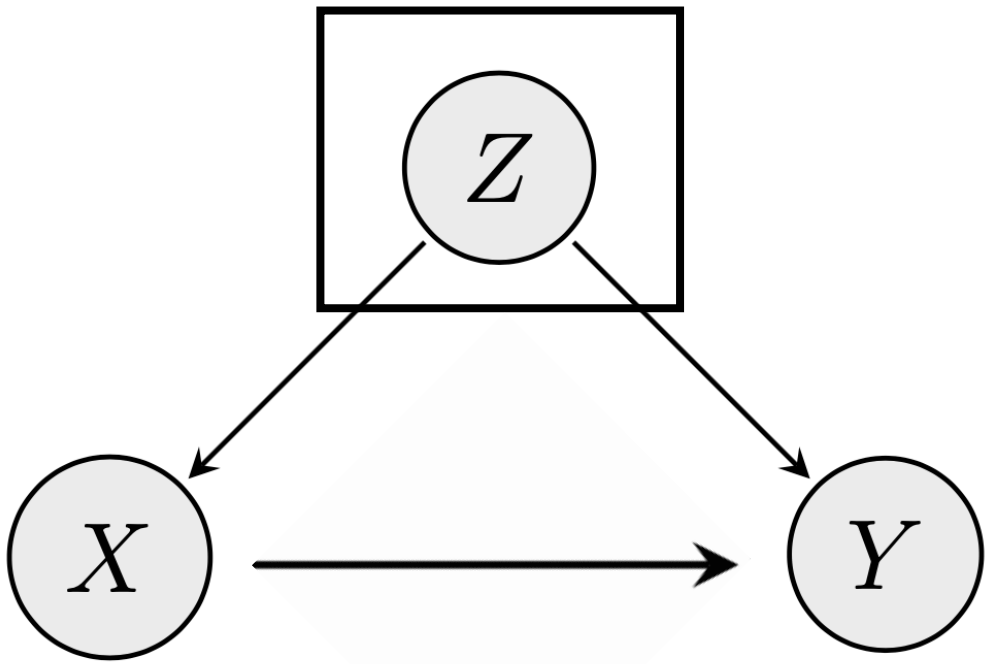
Include term in regression

$$\text{Win margin} = \beta_0 + \beta_1 \text{Campaign money} + \beta_2 \text{Candidate quality} + \epsilon$$

$$\text{Win margin} = \alpha + \beta\ \text{Campaign money} + \gamma\ \text{Candidate quality} + \epsilon$$

Matching **Stratifying** **Inverse probability weighting**

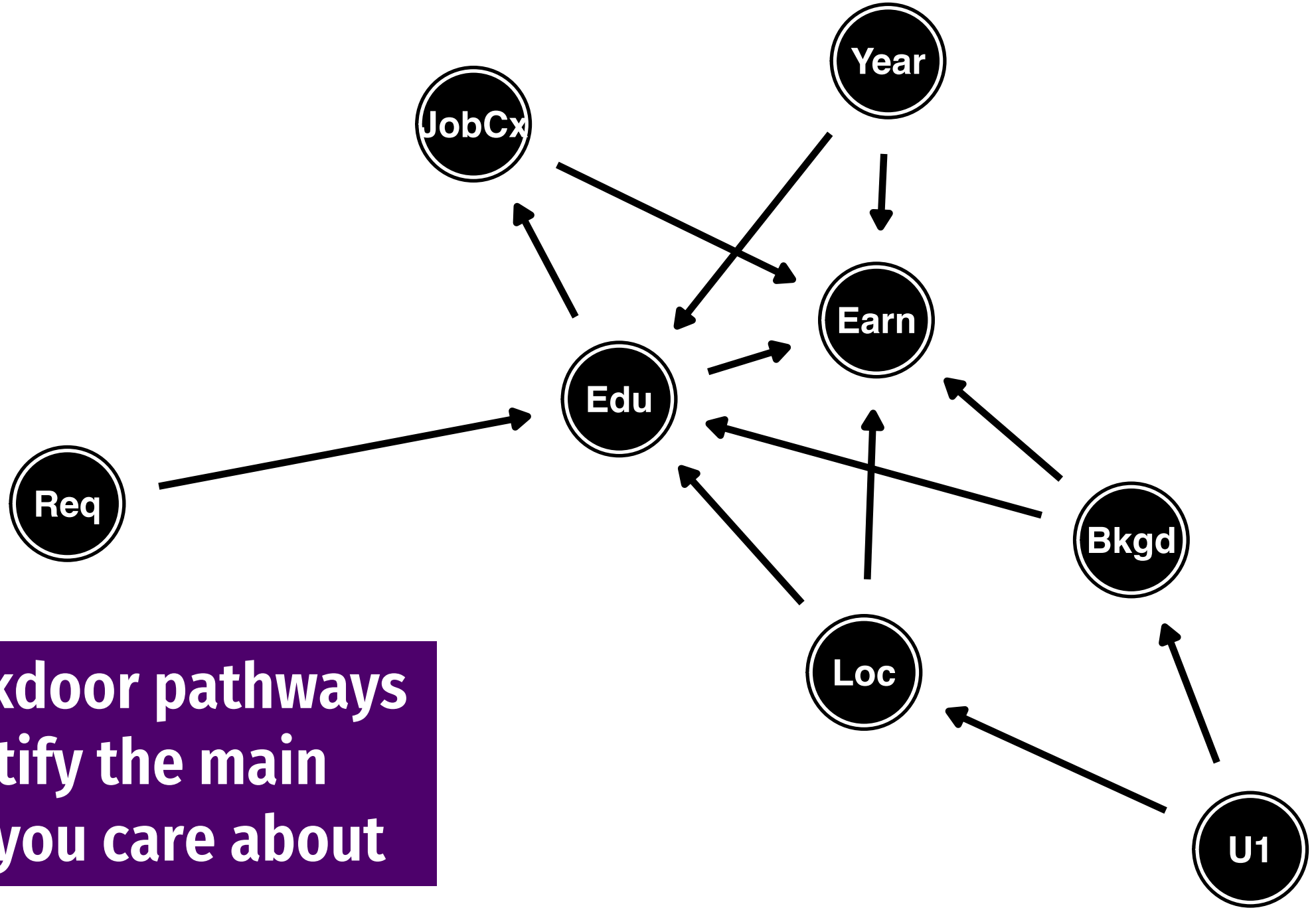
d-separation



If we control for Z, X and Y are now "*d*-separated" and association is isolated!

$$X \perp\!\!\!\perp Y \mid Z$$

X is independent of Y, given Z



Block backdoor pathways to identify the main pathway you care about

All paths

Education → Earnings

Education → Job connections → Earnings

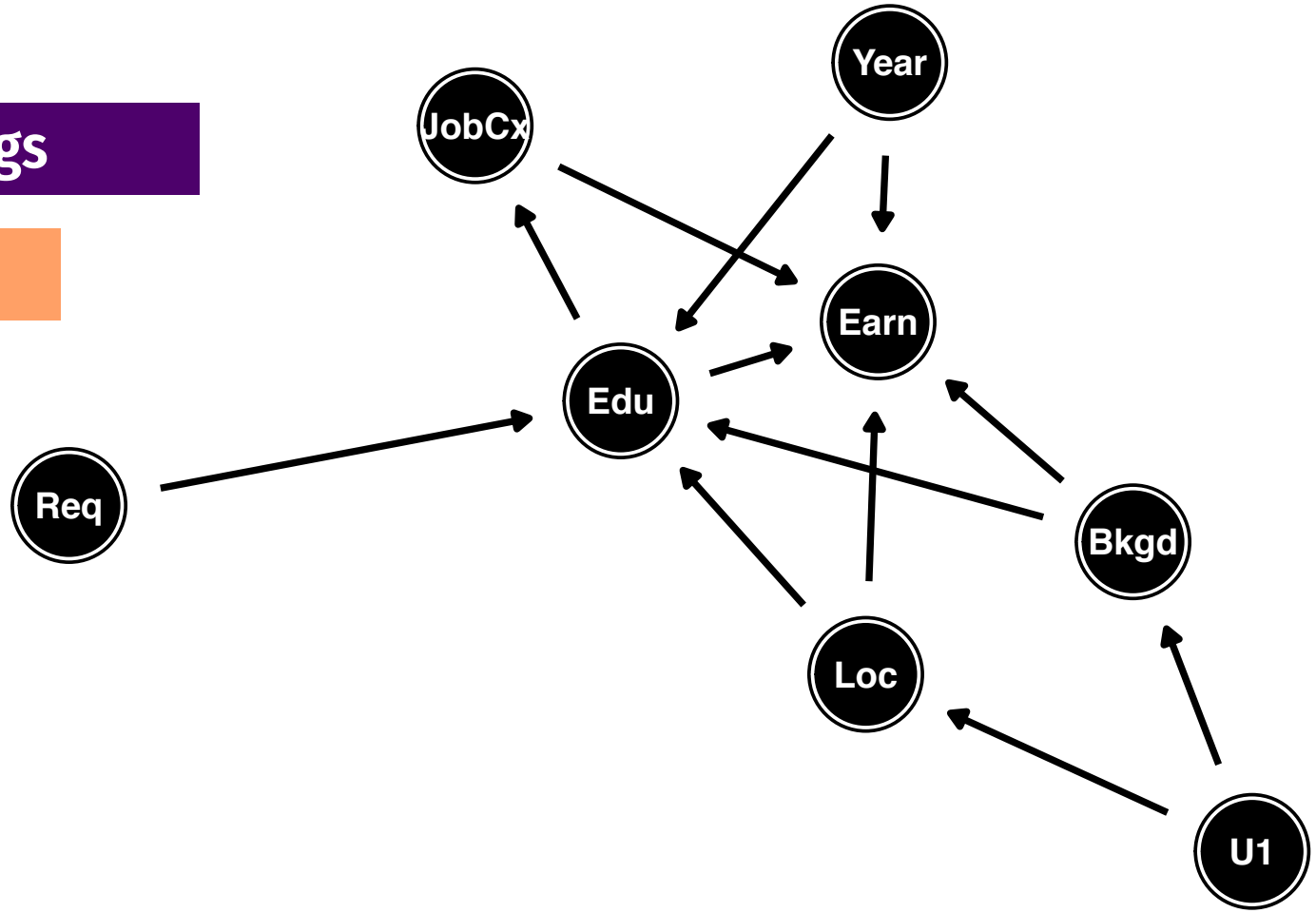
Education ← Background → Earnings

Education ← Background ← U1 → Location → Earnings

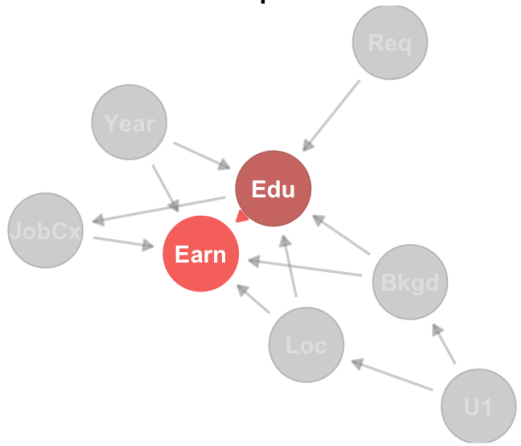
Education ← Location → Earnings

Education ← Location ← U1 → Background → Earnings

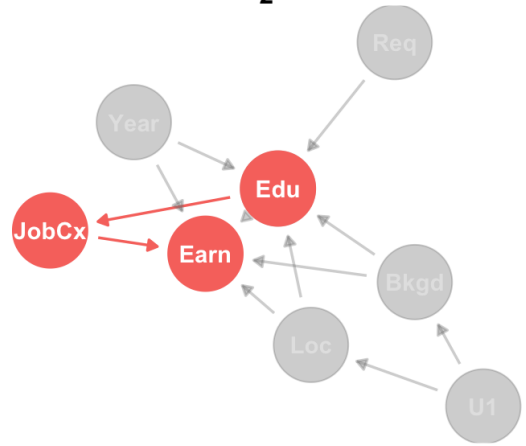
Education ← Year → Earnings



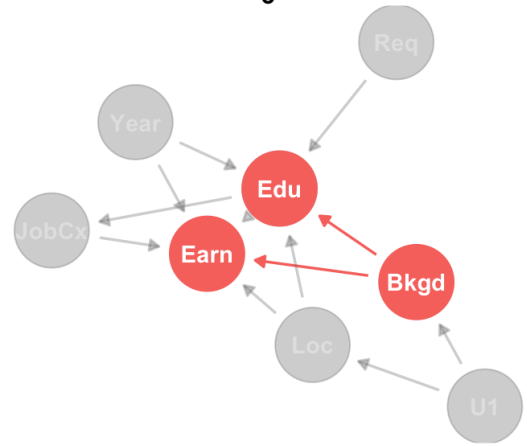
1



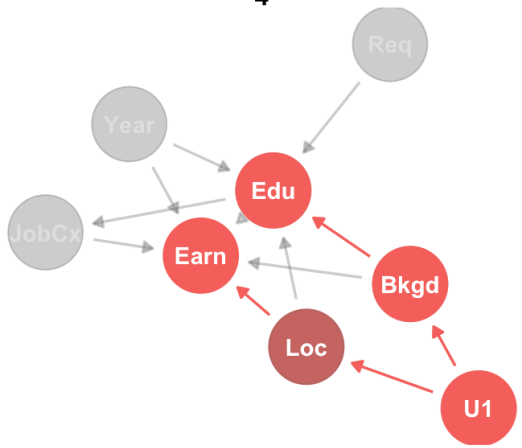
2



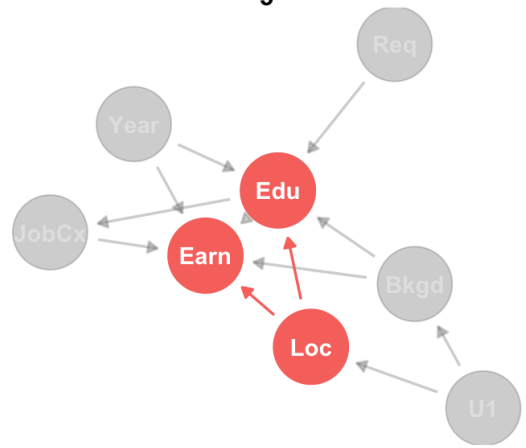
3



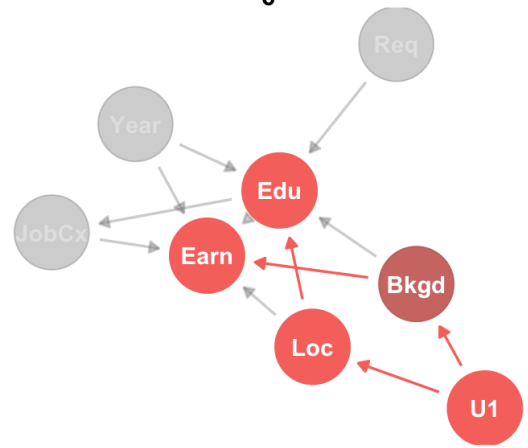
4



5



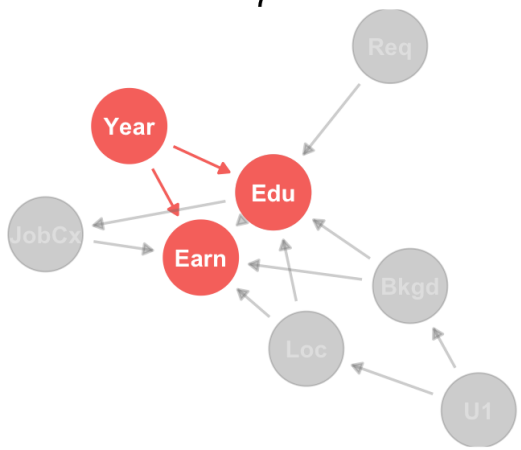
6



path

a open path

7



Closing doors

Education → Earnings

Education → Job connections → Earnings

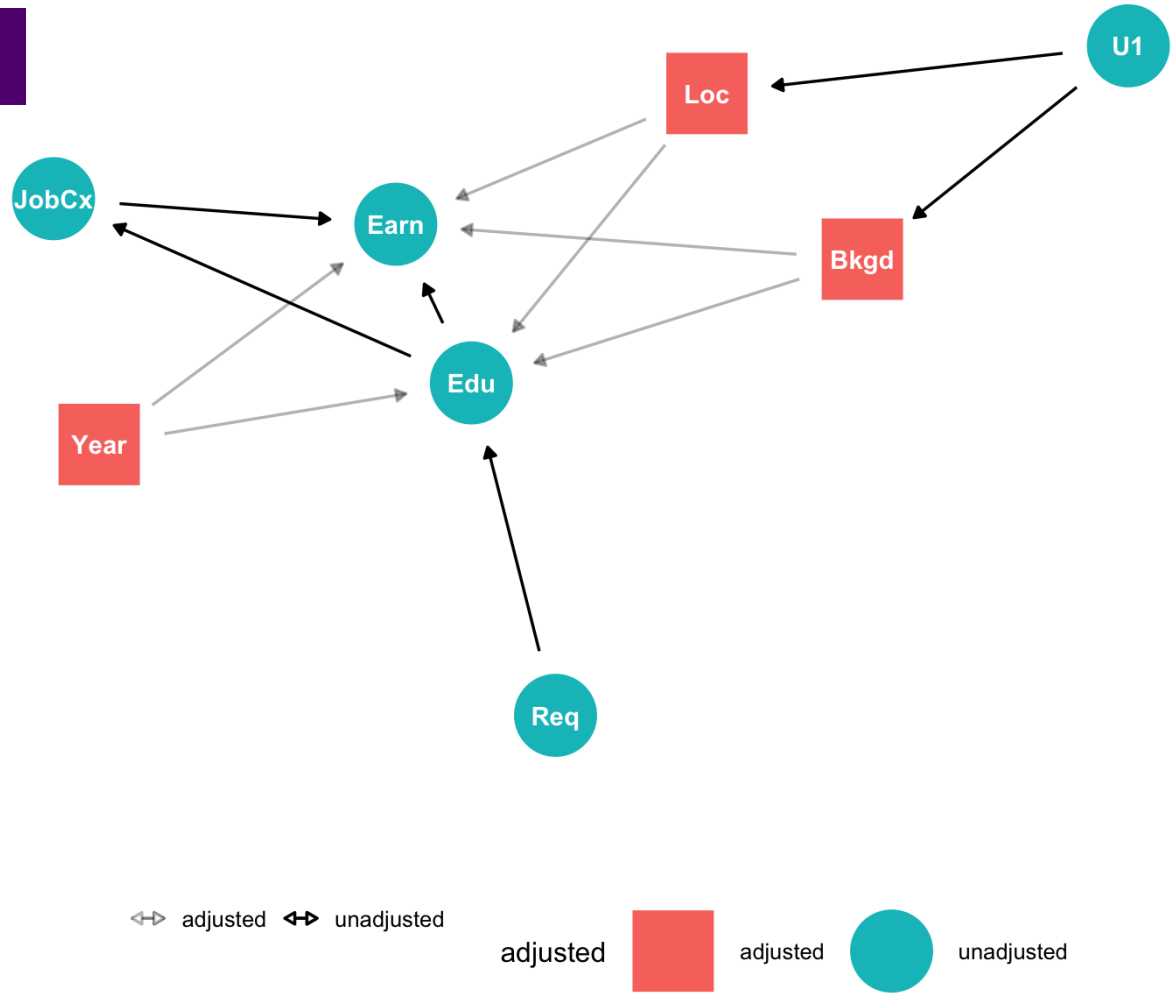
Education ← Background → Earnings

Education ← Background ← U1 → Location → Earnings

Education ← Location → Earnings

Education ← Location ← U1 → Background → Earnings

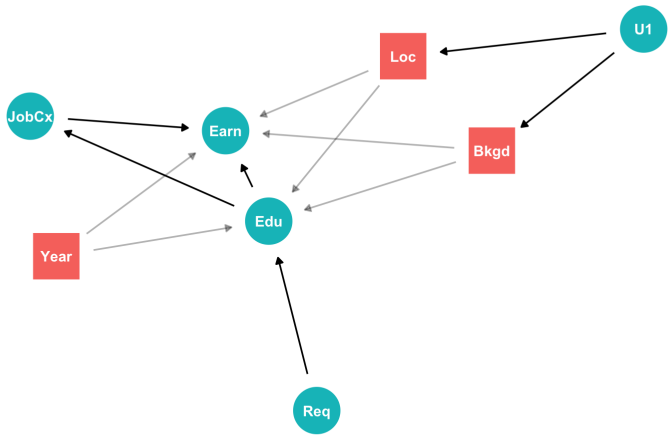
Education ← Year → Earnings



Closing doors

$$\text{Earnings} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Location} + \beta_3 \text{Background} + \beta_4 \text{Year} + \epsilon$$

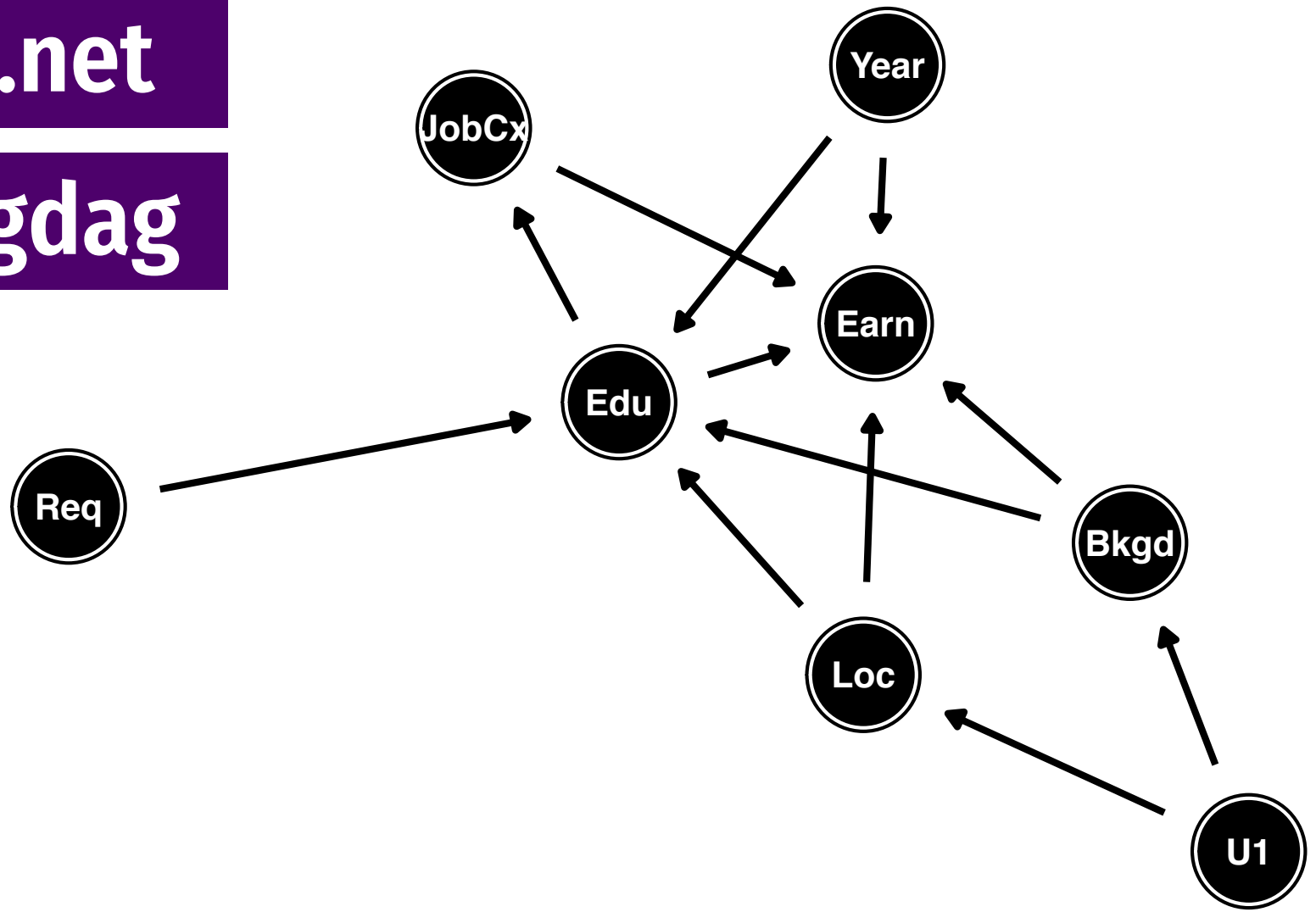
$$\text{Earnings} = \alpha + \beta \text{Education} + \gamma_1 \text{Location} + \gamma_2 \text{Background} + \gamma_3 \text{Year} + \epsilon$$



Let the computer do this!

dagitty.net

R and ggdag

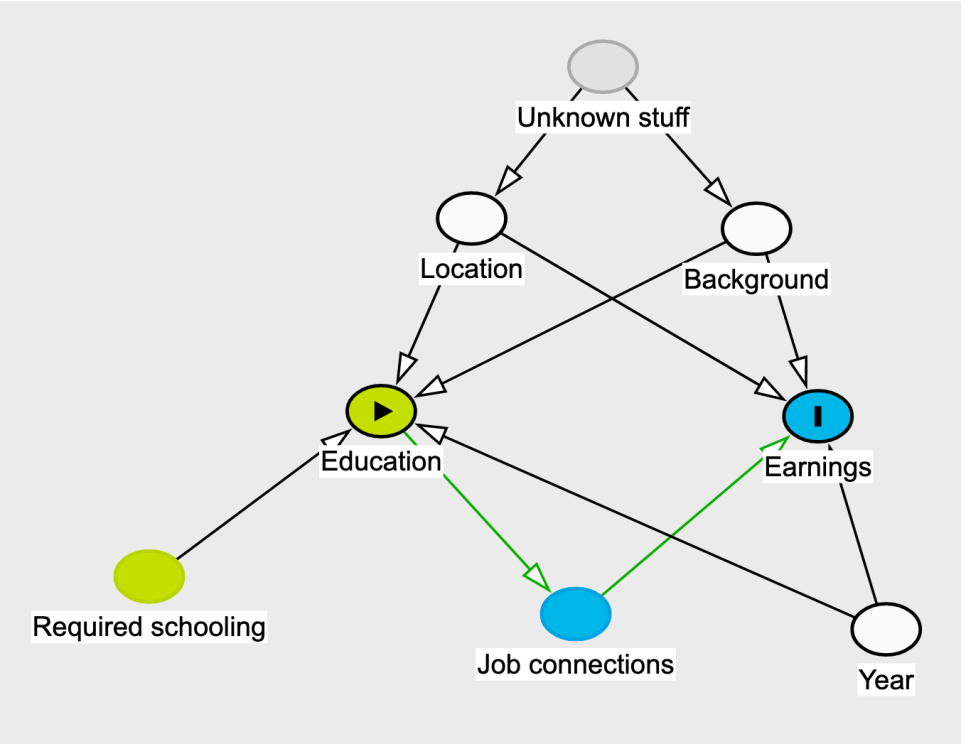


How do you know if this is right?

Once you start closing doors, you can test the implications of the model and see if they're right in your data

$X \perp\!\!\!\perp Y \mid Z$

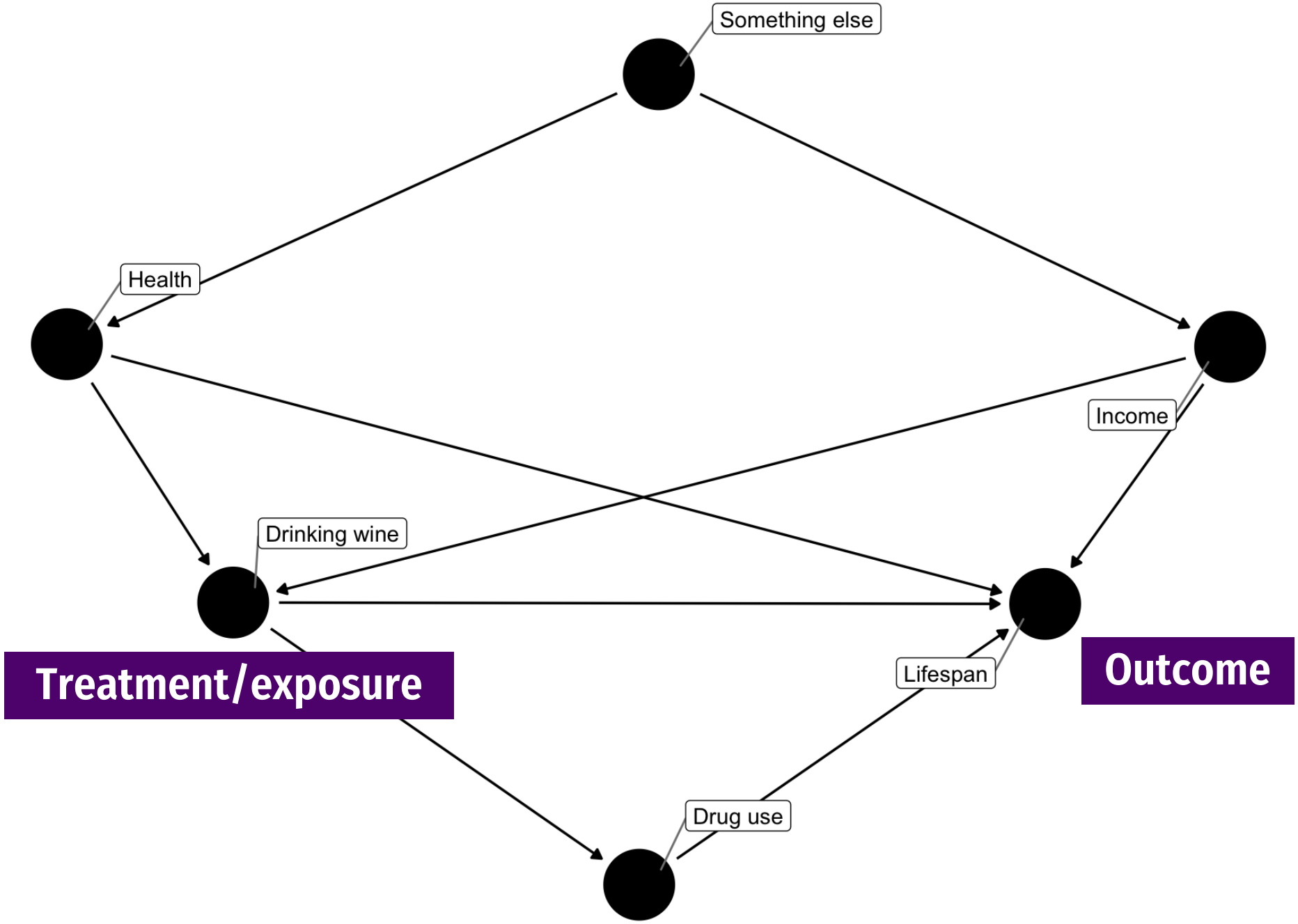
X is independent of Y, given Z



Testable implications

The model implies the following conditional independences:

- Education ⊥ Earnings | Background, Job connections, Location, Year
- Required schooling ⊥ Job connections | Education
- Required schooling ⊥ Year
- Required schooling ⊥ Earnings | Background, Job connections, Location, Year
- Required schooling ⊥ Earnings | Background, Education, Location, Year
- Required schooling ⊥ Background
- Required schooling ⊥ Location
- Job connections ⊥ Year | Education
- Job connections ⊥ Background | Education
- Job connections ⊥ Location | Education
- Year ⊥ Background
- Year ⊥ Location


Something else
Health
Income
Drinking wine
Treatment/exposure
Lifespan
Outcome
Drug use

Wine → Lifespan

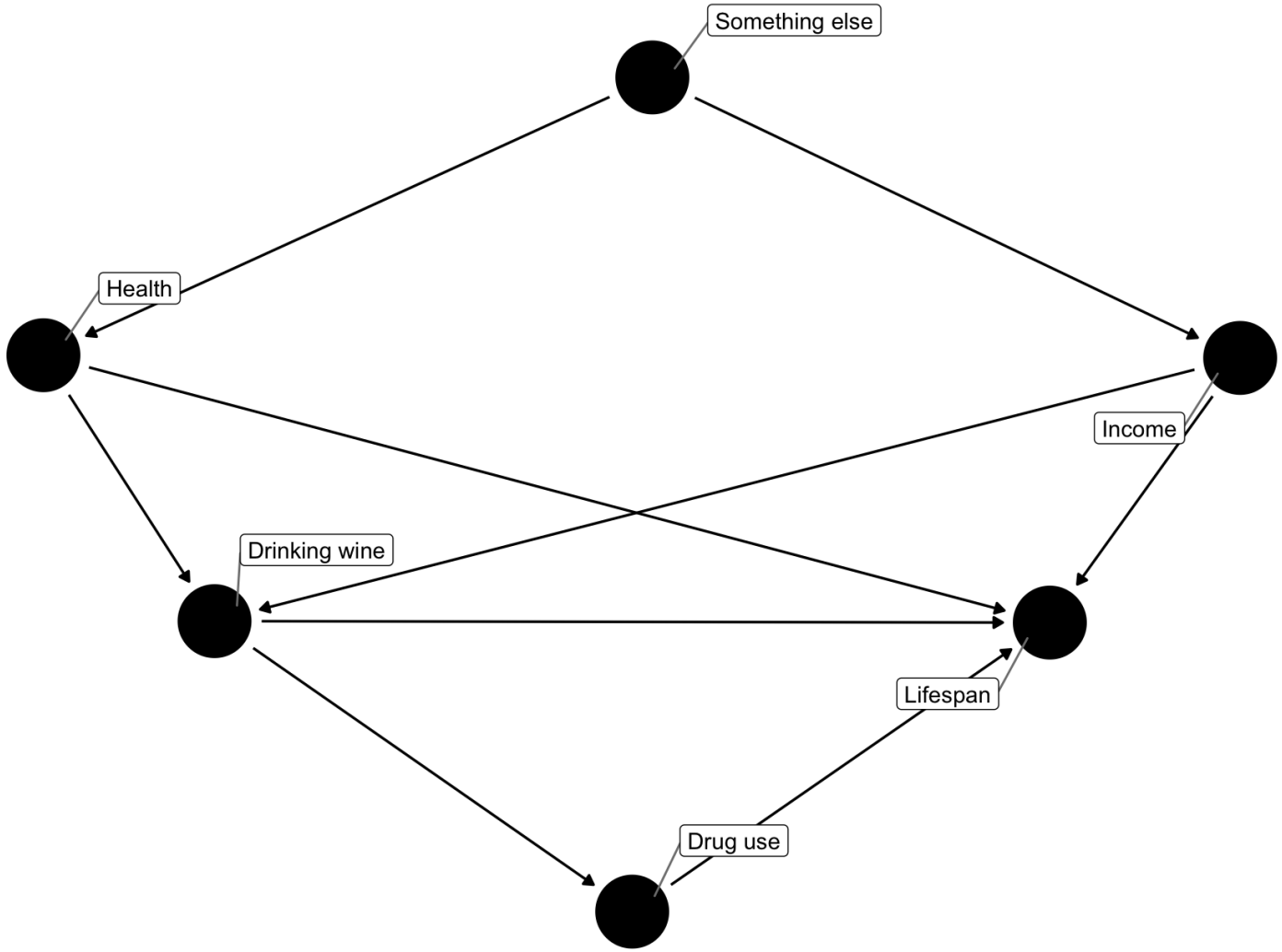
Wine → Drugs→ Lifespan

Wine ← Health → Lifespan

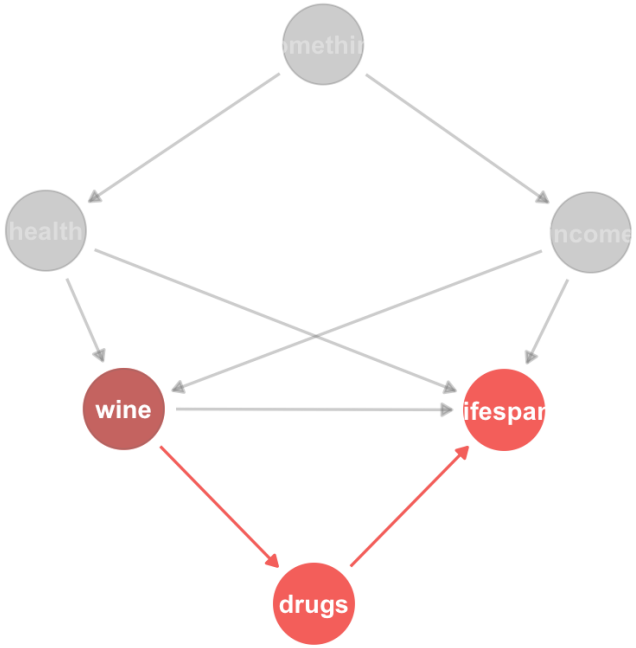
Wine ← Health ← Something → Income → Lifespan

Wine ← Income → Lifespan

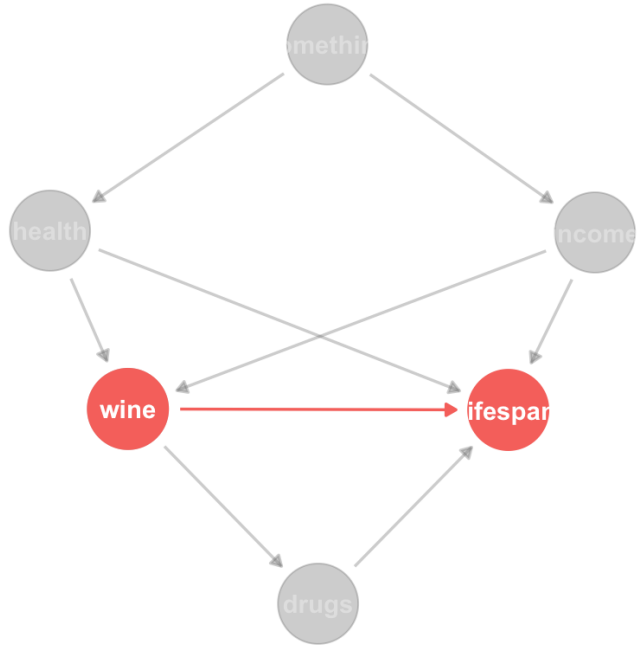
Wine ← Income ← Something → Health → Lifespan



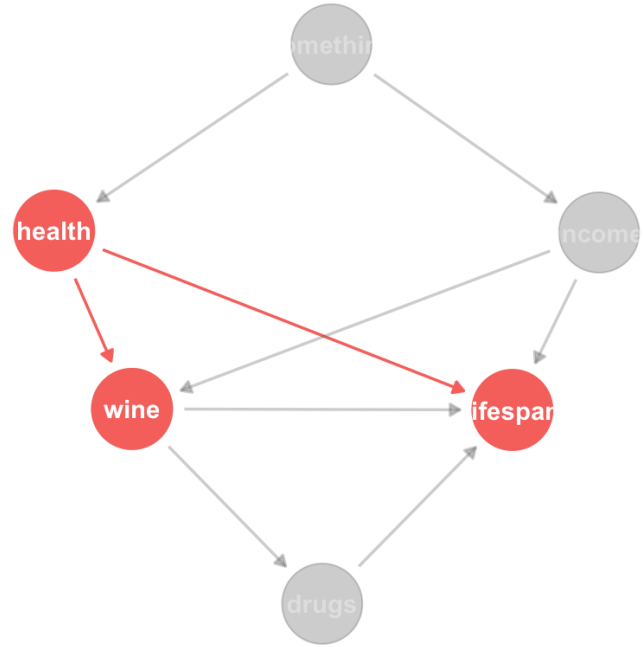
1


wine
drugs

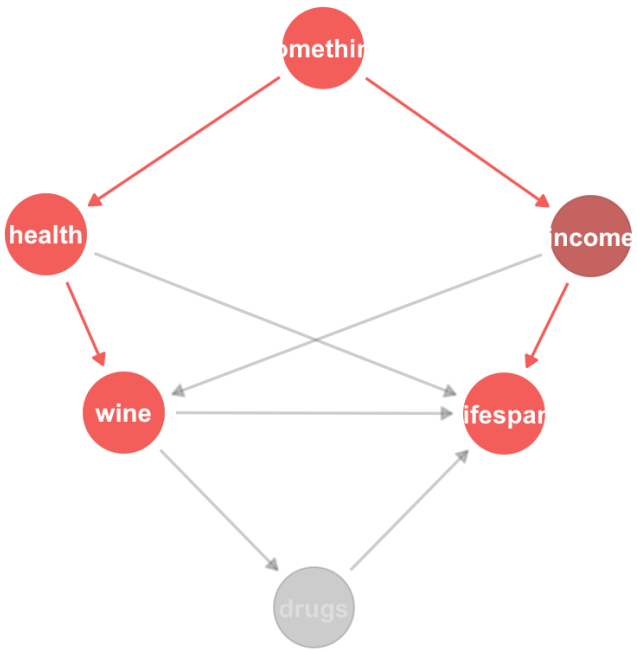
2


wine
drugs

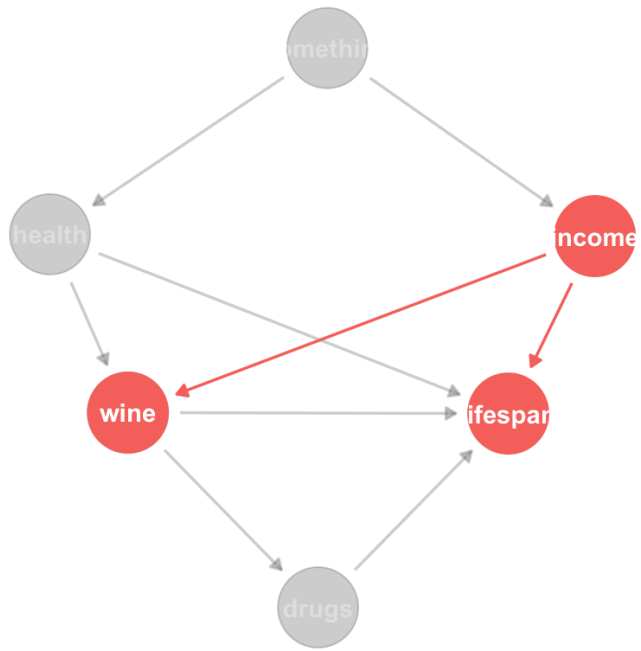
3


health
wine
drugs

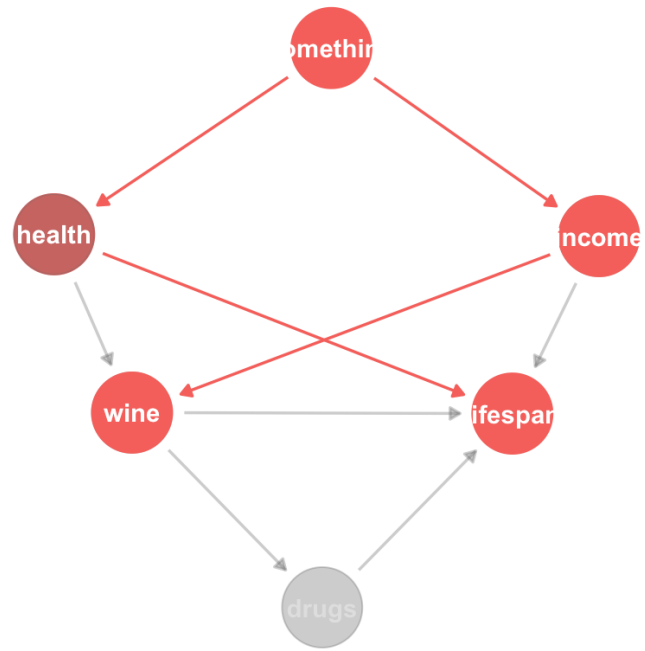
4

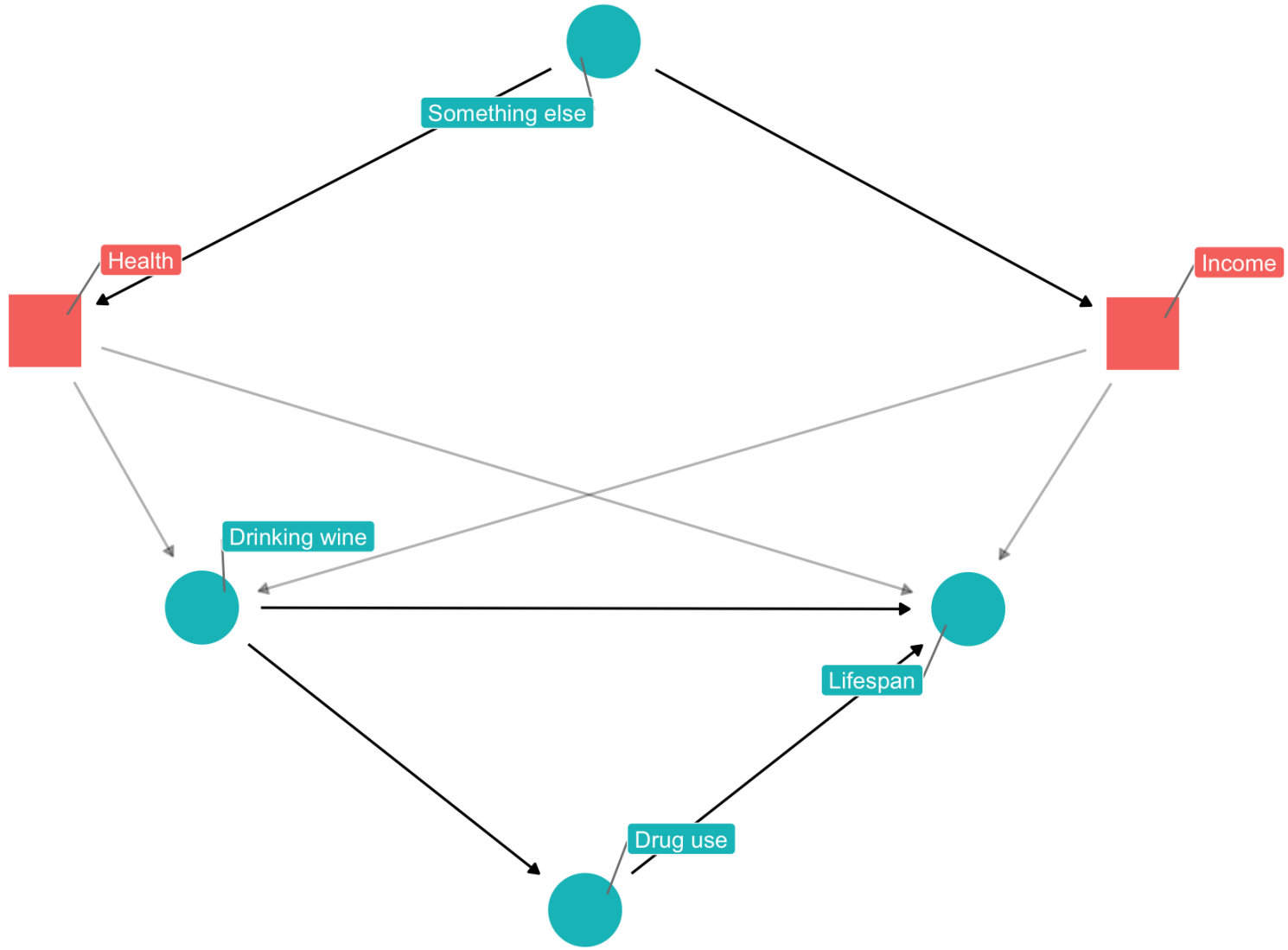

health
wine
drugs

5


wine
drugs

6


health
wine
drugs



adjusted unadjusted

adjusted adjusted unadjusted

Your turn

Go to andhs.co/nyt and read the article

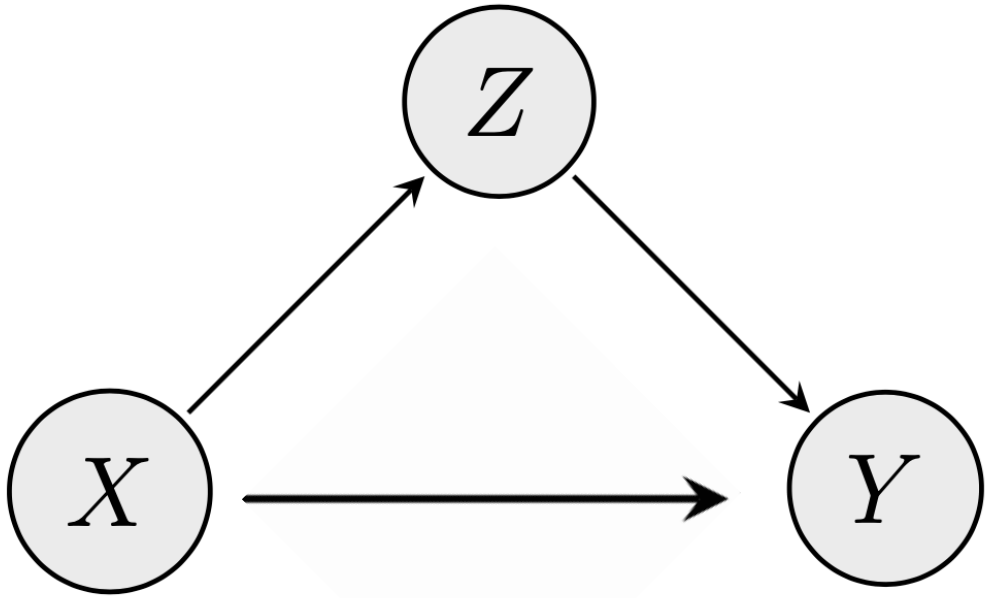
Pick one of the causal claims in the article

(There are a lot! Look for words like “improve”, “affect”, and “reduces)

Draw a diagram for that causal claim

Determine what needs to be adjusted to identify the effect

Causation

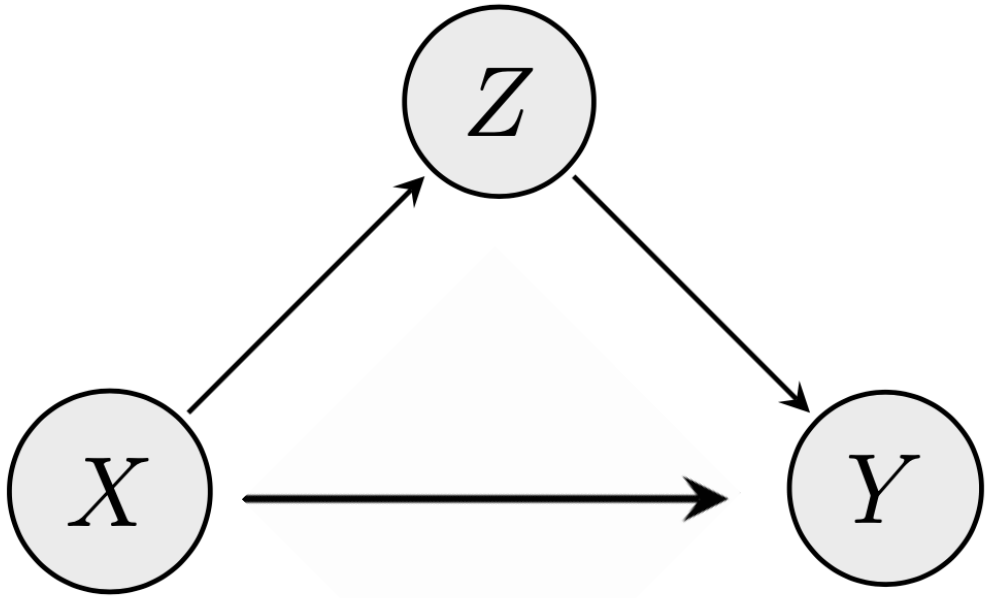


X causes Y

X causes Z which causes Y

Should you control for Z?

Causation

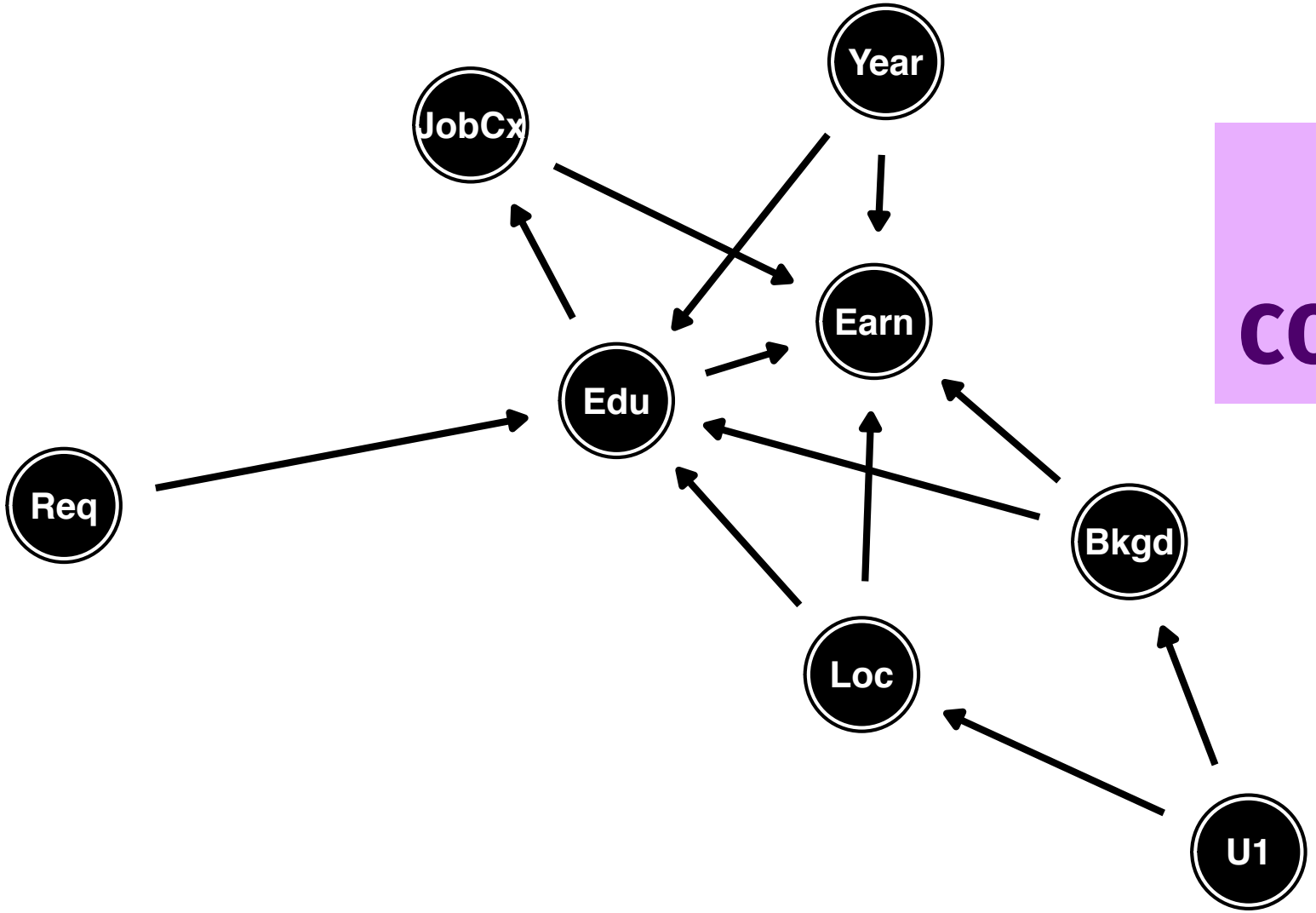


Should you control for Z?

No!

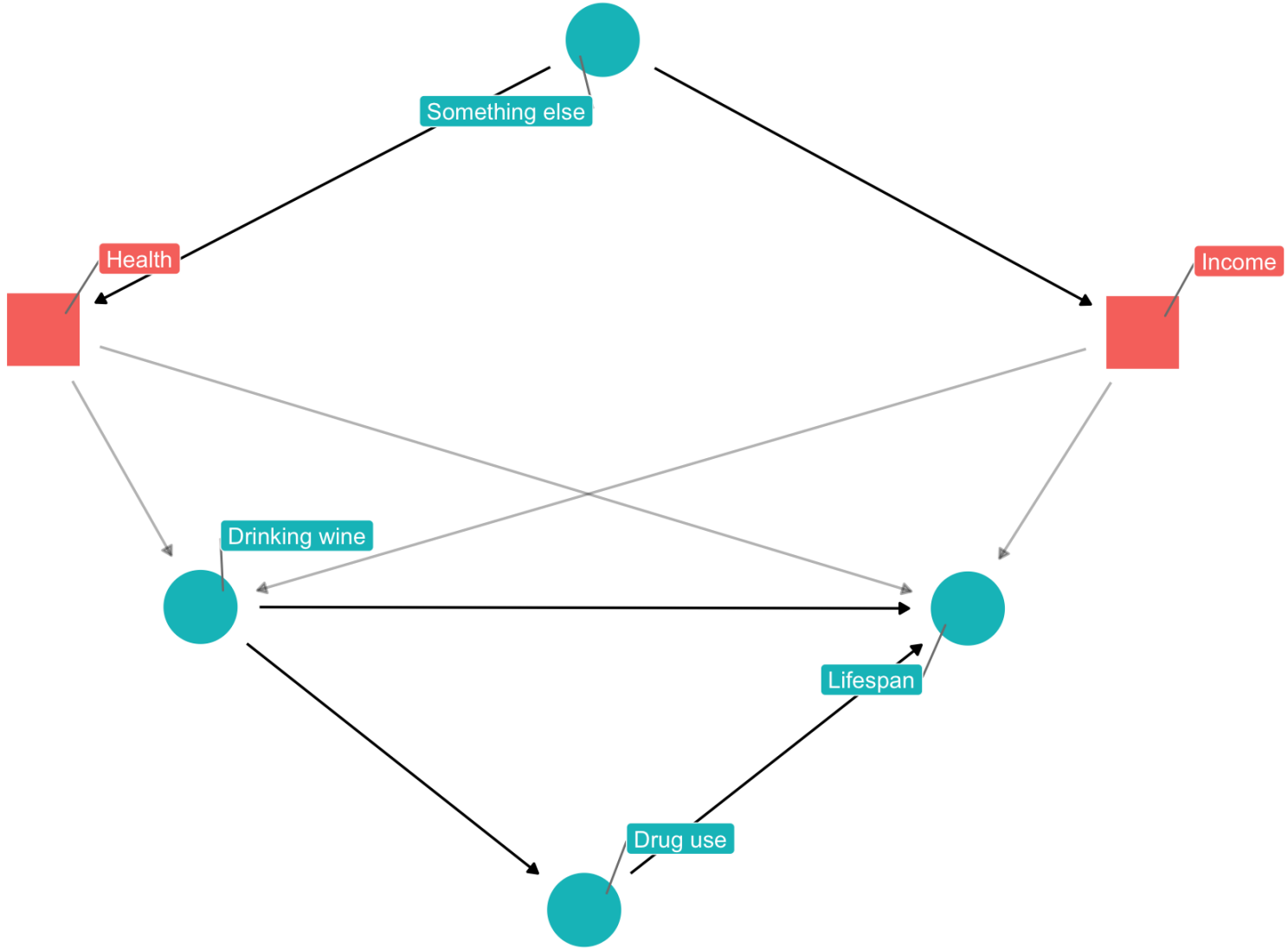
Overcontrolling

Causation and overcontrolling



Should you control for JobCx?

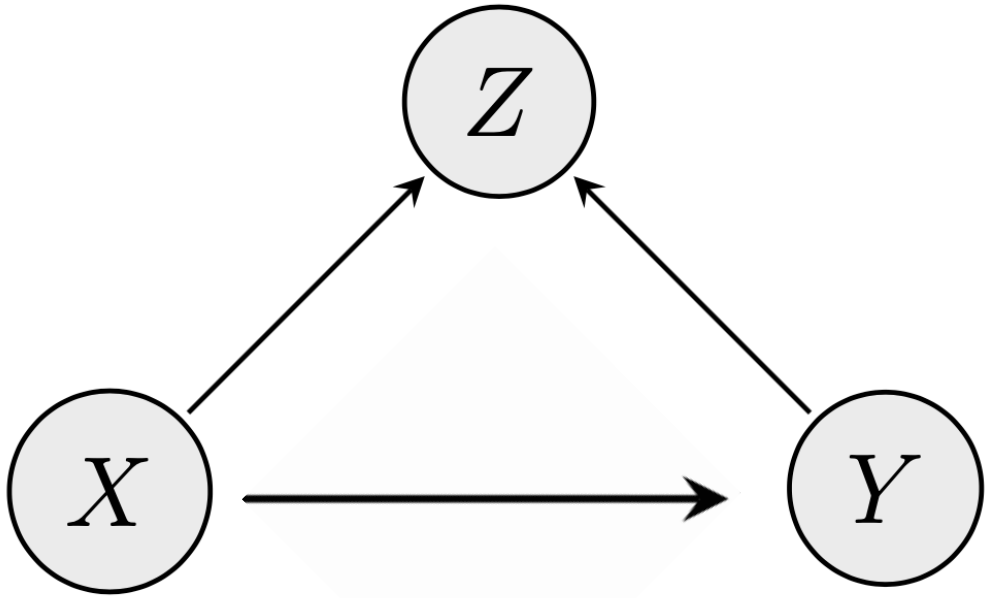
What would happen if we controlled for drug use?



adjusted unadjusted

adjusted adjusted unadjusted

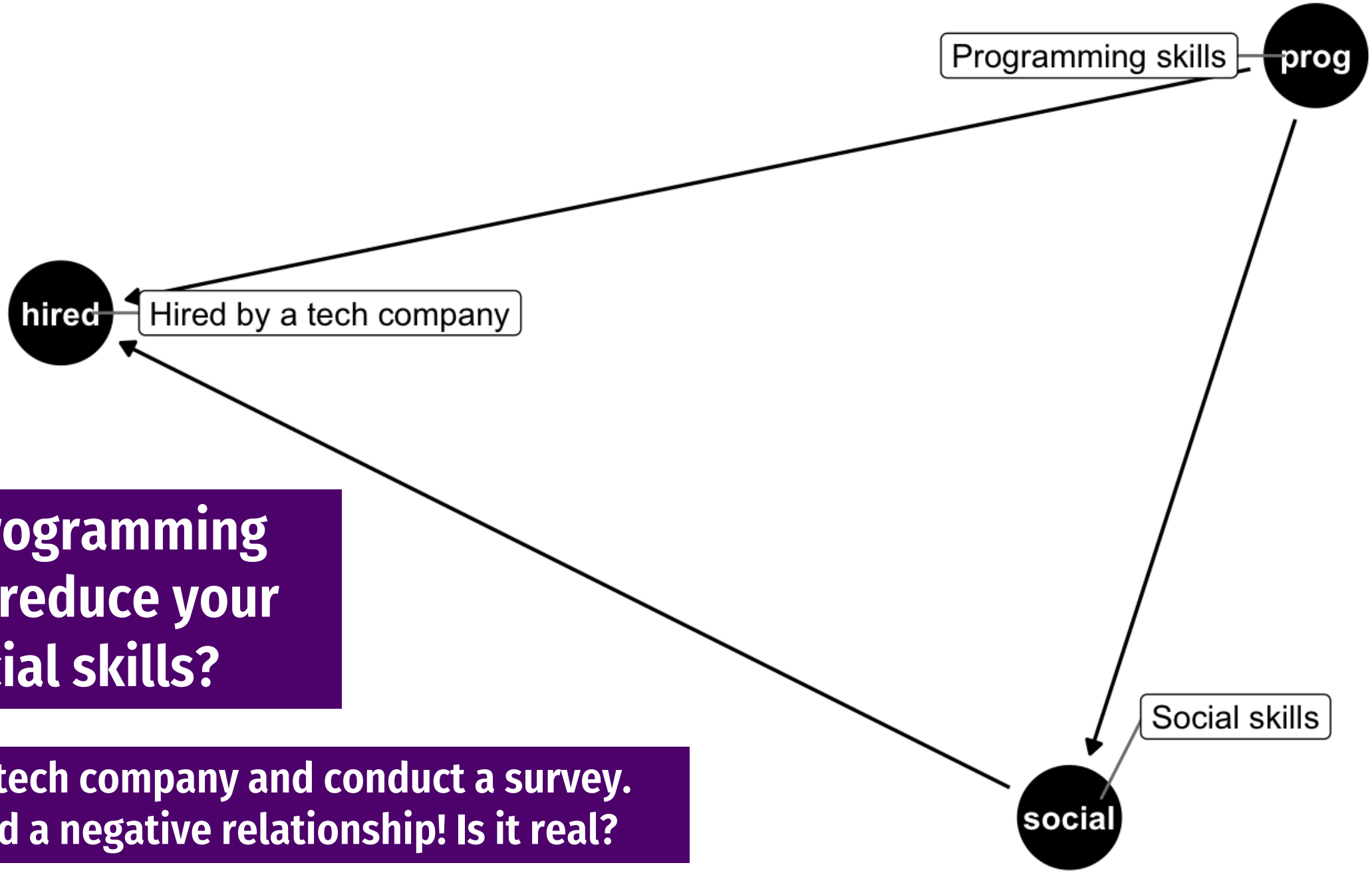
Colliders



X causes Z

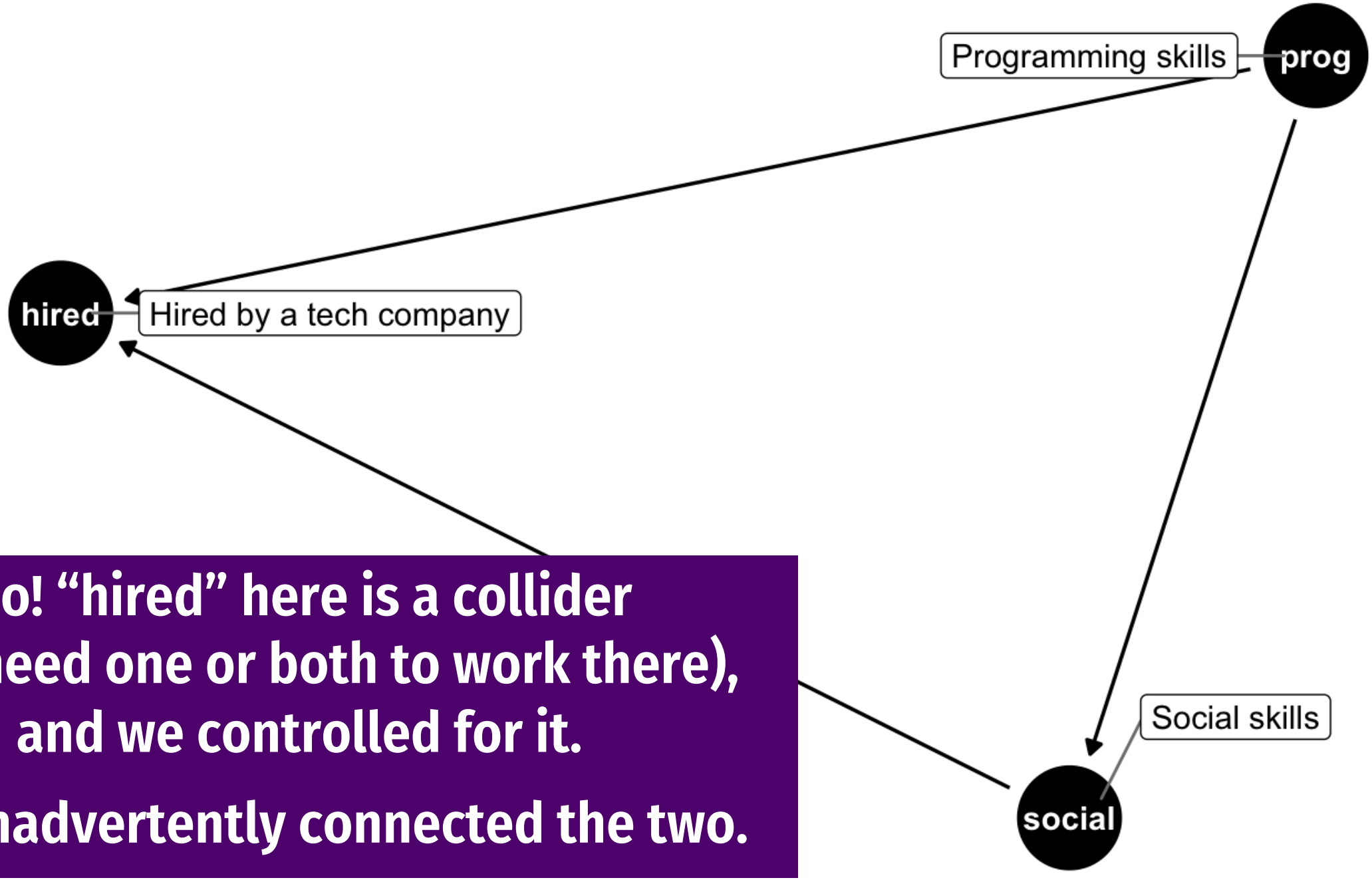
Y causes Z

Should you control for Z?



Do programming skills reduce your social skills?

Go to a tech company and conduct a survey. You find a negative relationship! Is it real?



No! “hired” here is a collider (you need one or both to work there), and we controlled for it.

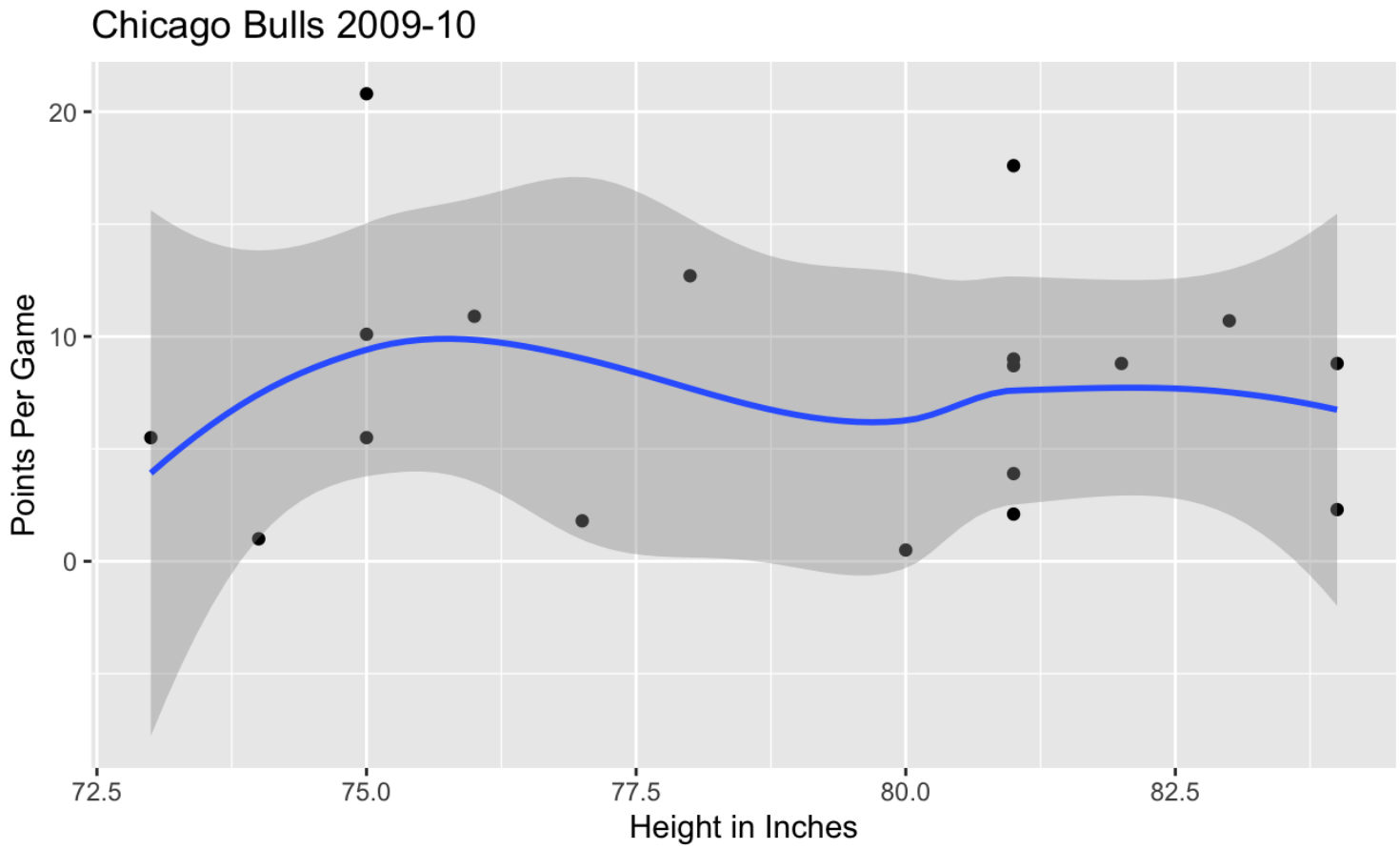
That inadvertently connected the two.

Colliders can create fake causal effects

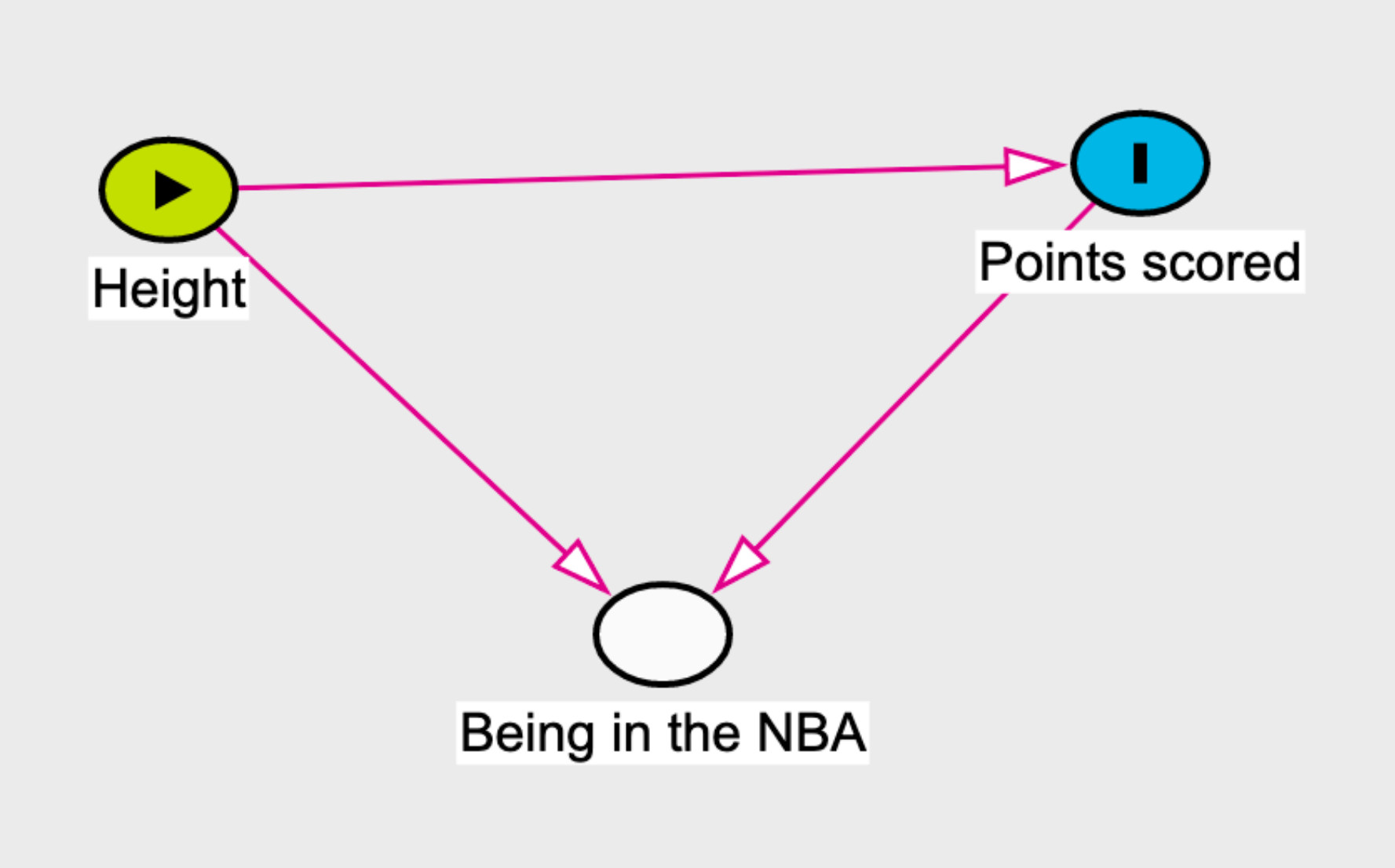
Colliders can hide real causal effects

Height is unrelated to basketball skill!

...among NBA players

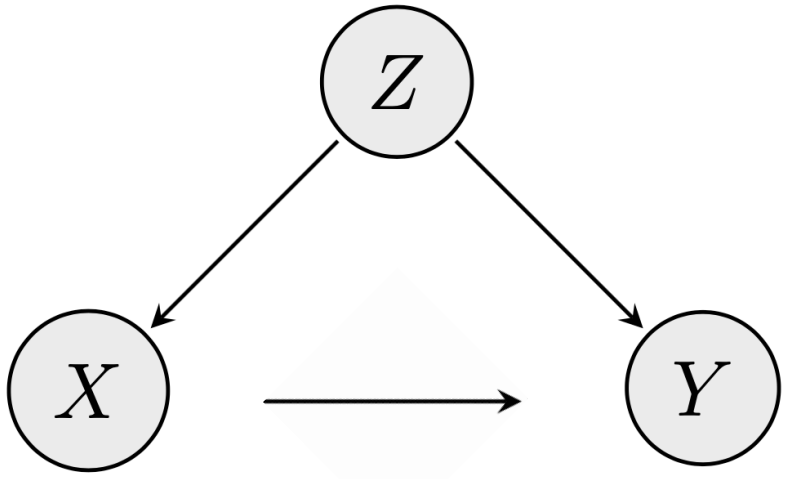


Colliders and selection bias



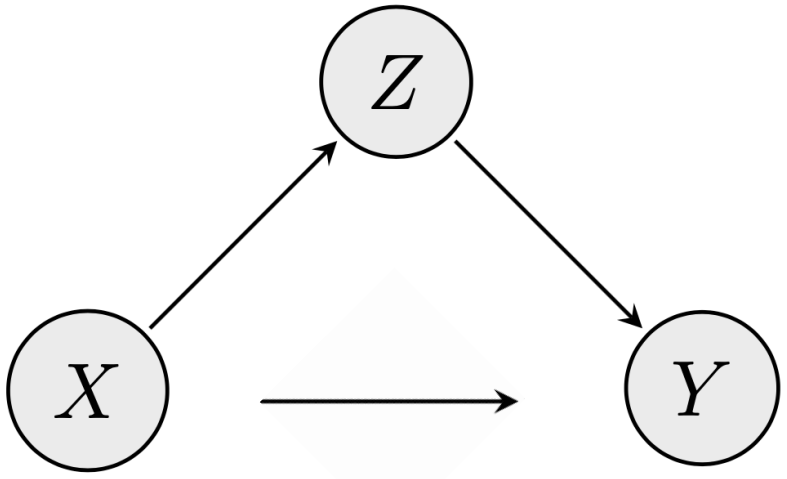
Three types of associations

Confounding



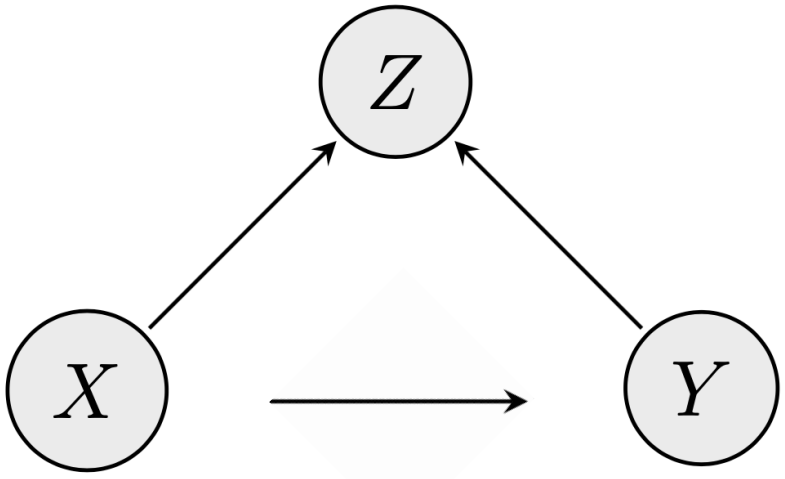
Common cause

Causation



Mediation

Collision



Selection / Endogeneity

Does SES cause vaping?

Green *et al. BMC Public Health* (2020) 20:183
https://doi.org/10.1186/s12889-020-8270-3

BMC Public Health

RESEARCH ARTICLE Ope

Socioeconomic patterning of vaping by smoking status among UK adults and youth

Michael J. Green[1*], Linsay Gray[1], Helen Sweeting[1] and Michaela Benzeval[2]

Abstract

Background: Smoking contributes significantly to socioeconomic health inequalities. Vaping has capture
interest as a less harmful alternative to smoking, but may be harmful relative to non-smoking. Examining

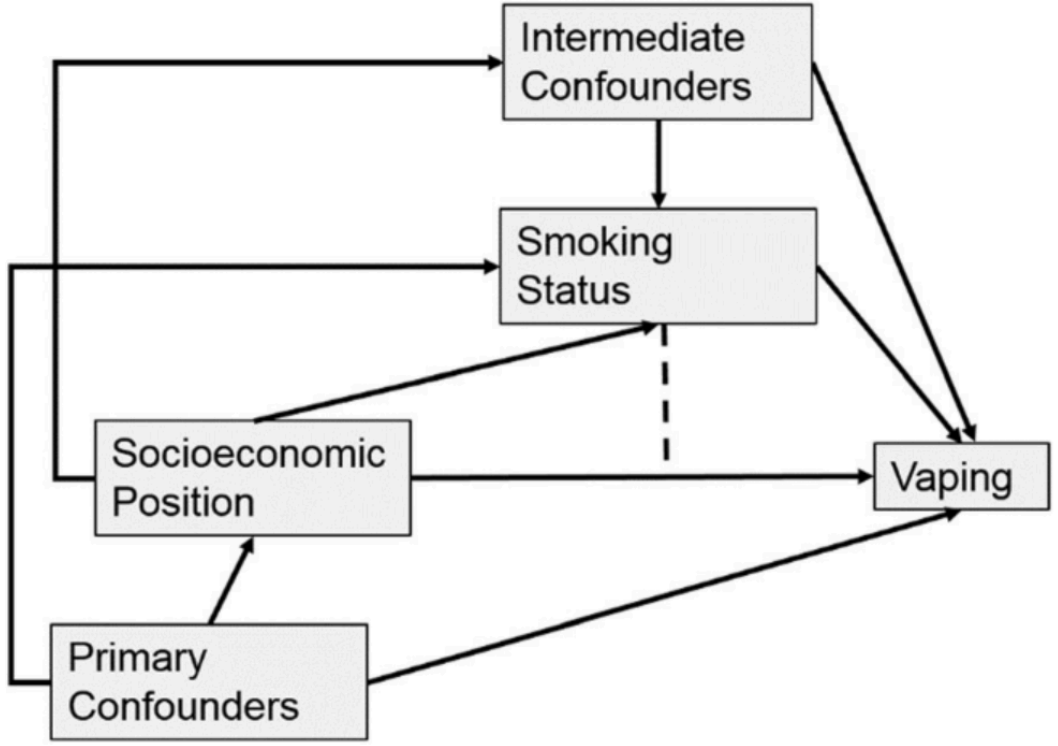
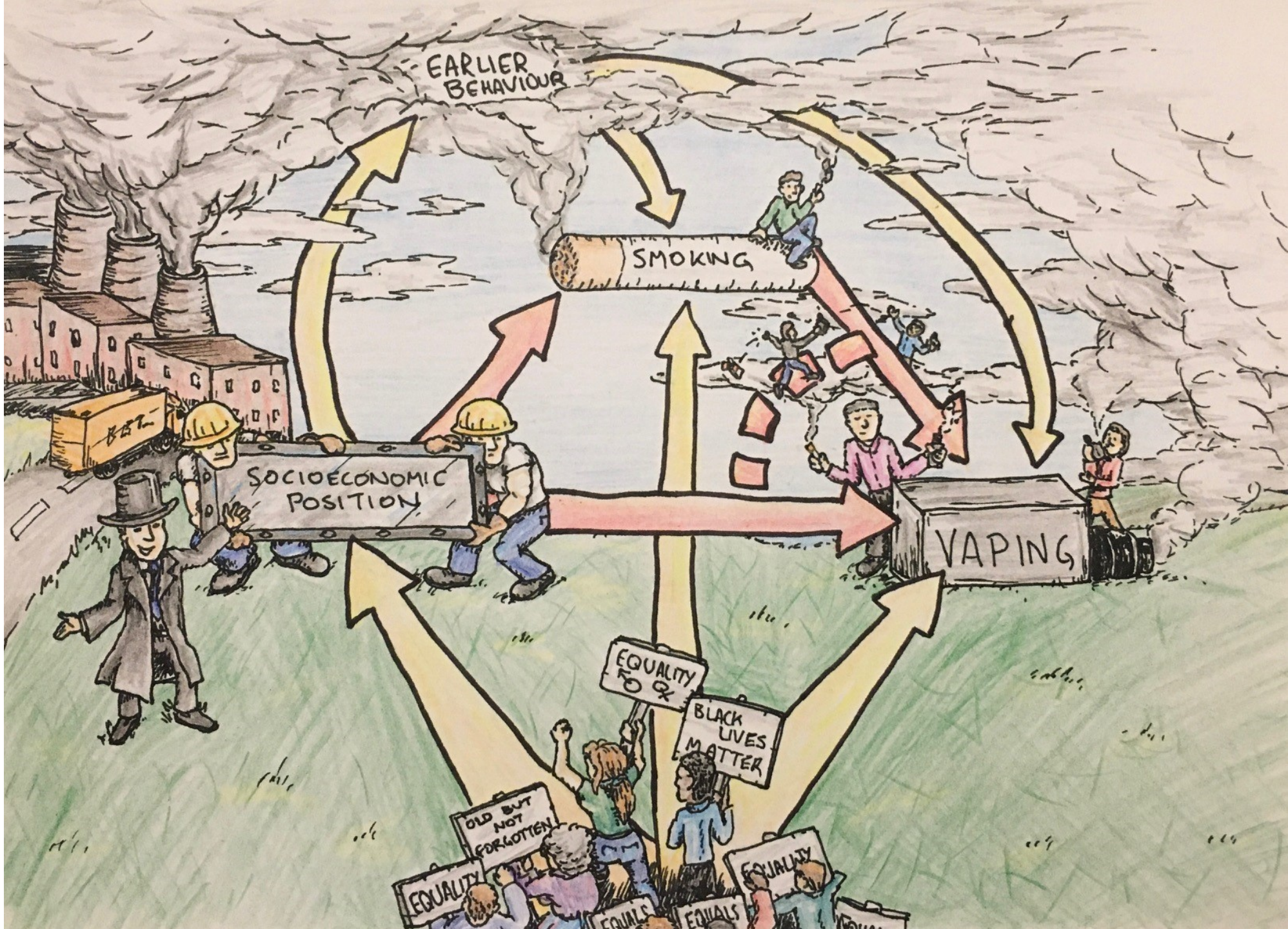


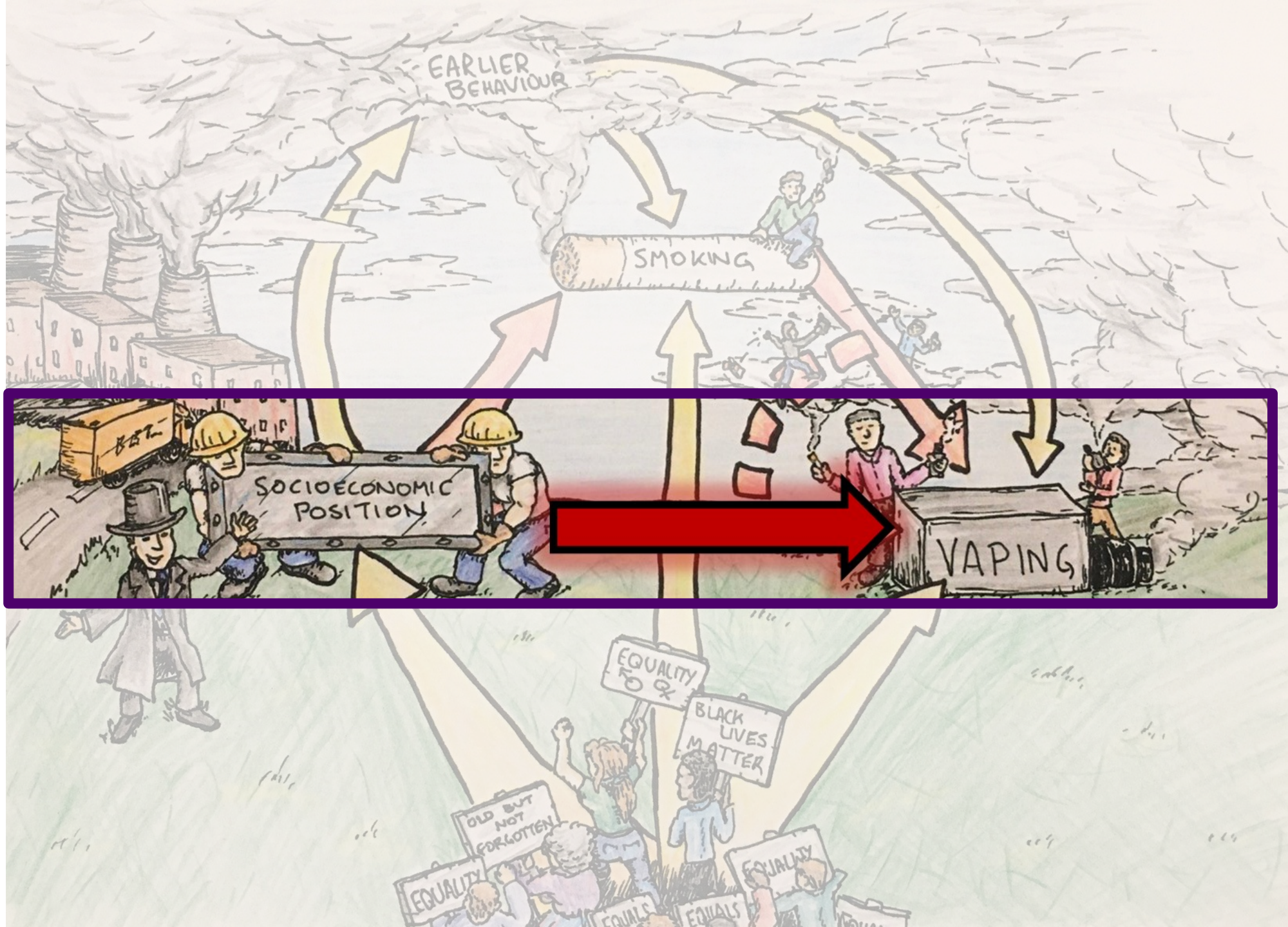
Fig. 1 Causal Diagram for analyses of SEP, smoking and vaping


EARLIER BEHAVIOUR
SMOKING
SOCIOECONOMIC POSITION
VAPING
EQUALITY TO 0X
BLACK LIVES MATTER
OLD BUT NOT FORGOTTEN
EQUALITY
EQUALITY

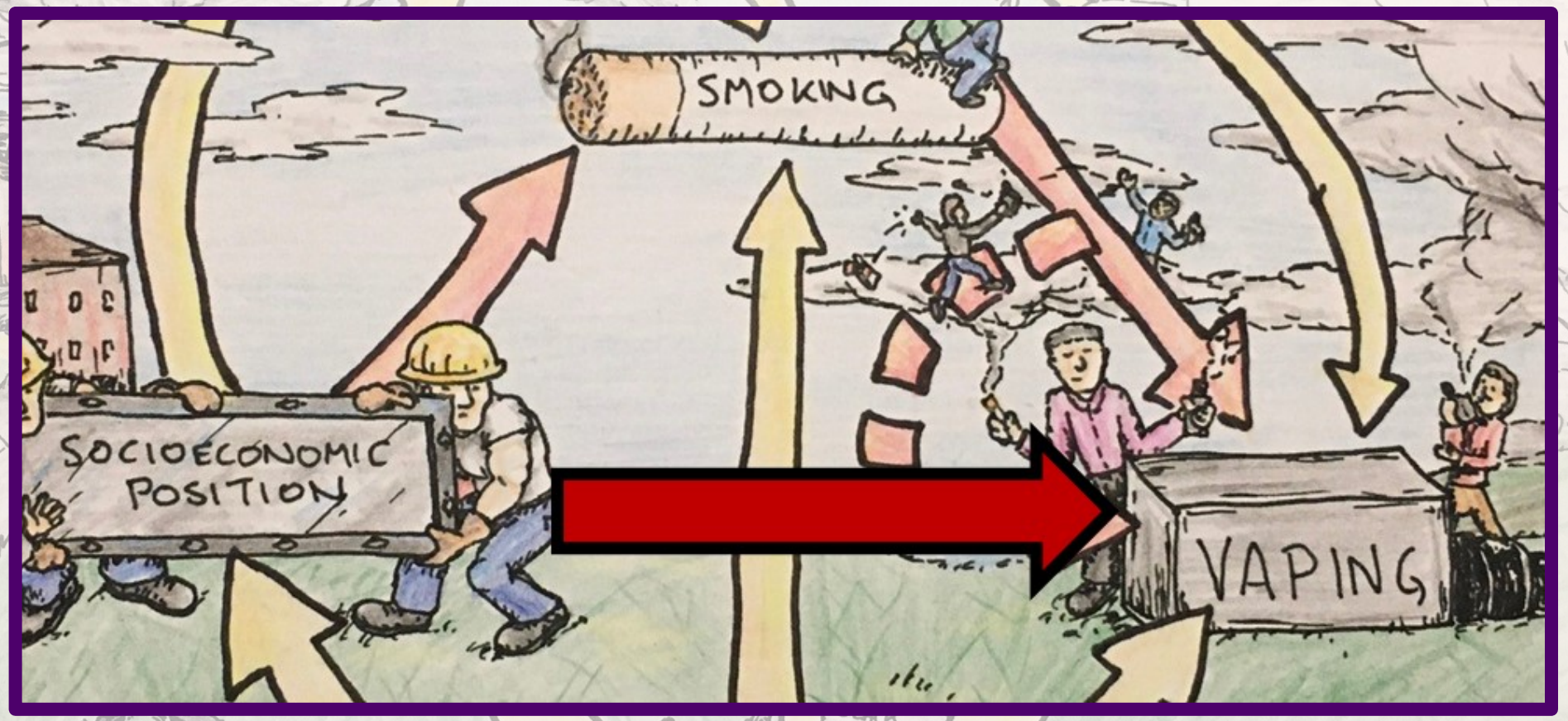
Smoking → socioeconomic inequalities in health

Vaping = alternative to smoking

So what is effect of SES on vaping?

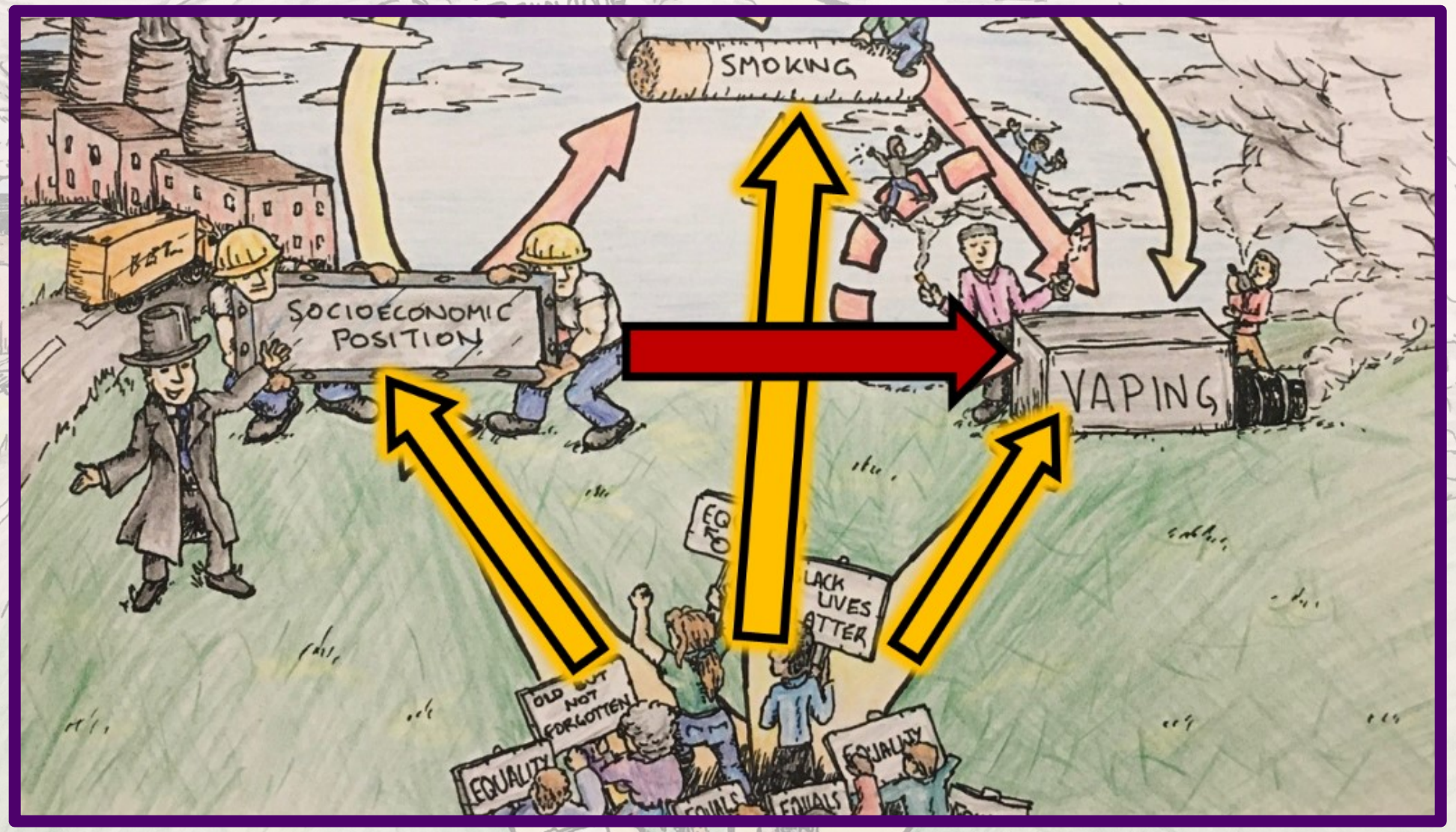


Adjust for smoking, not because it's a backdoor, but because vaping by never-smokers is different than vaping by current/former smokers



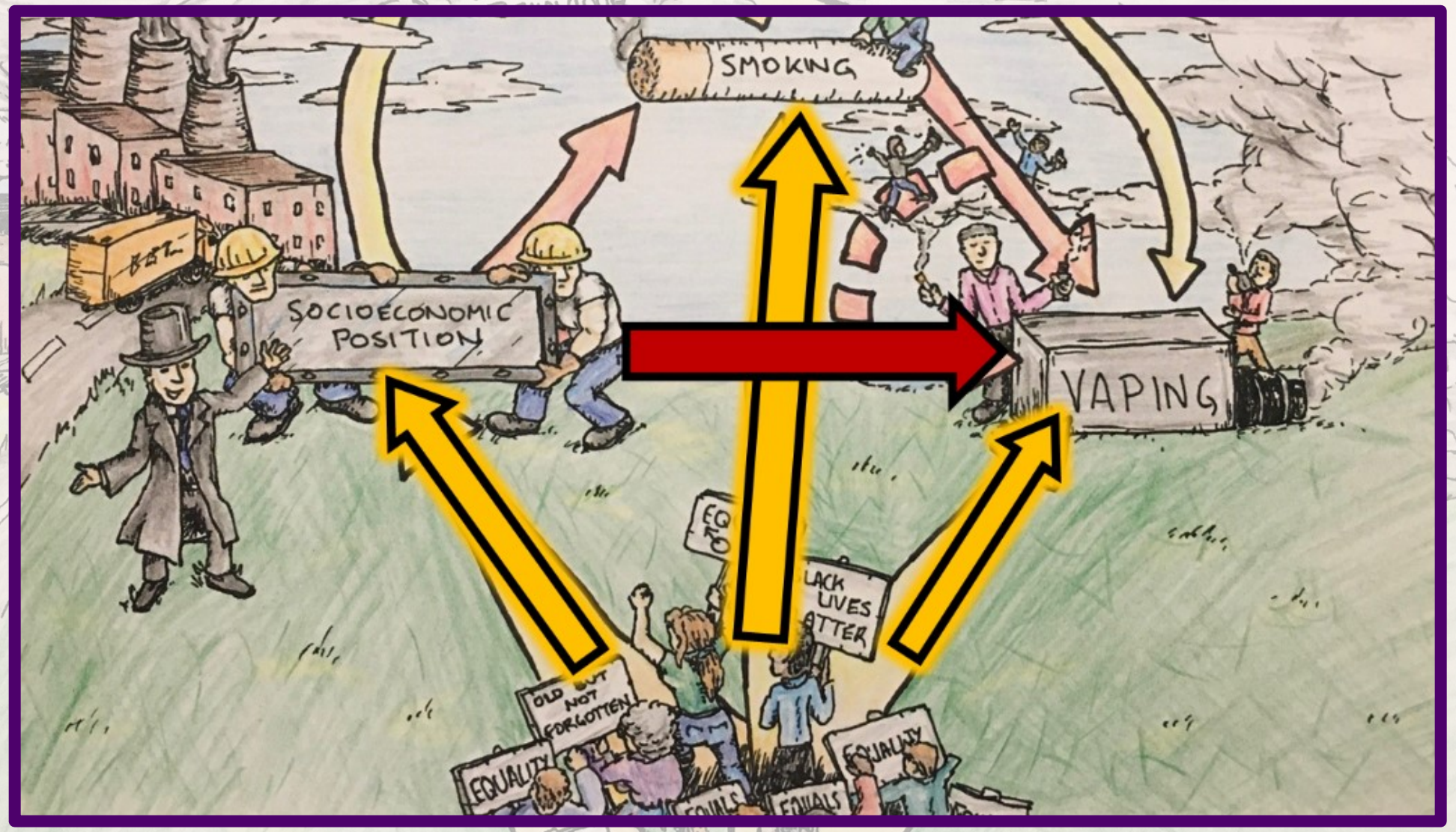
Adjust for other confounders because they're backdoors

Age, gender, race, ethnicity, etc.



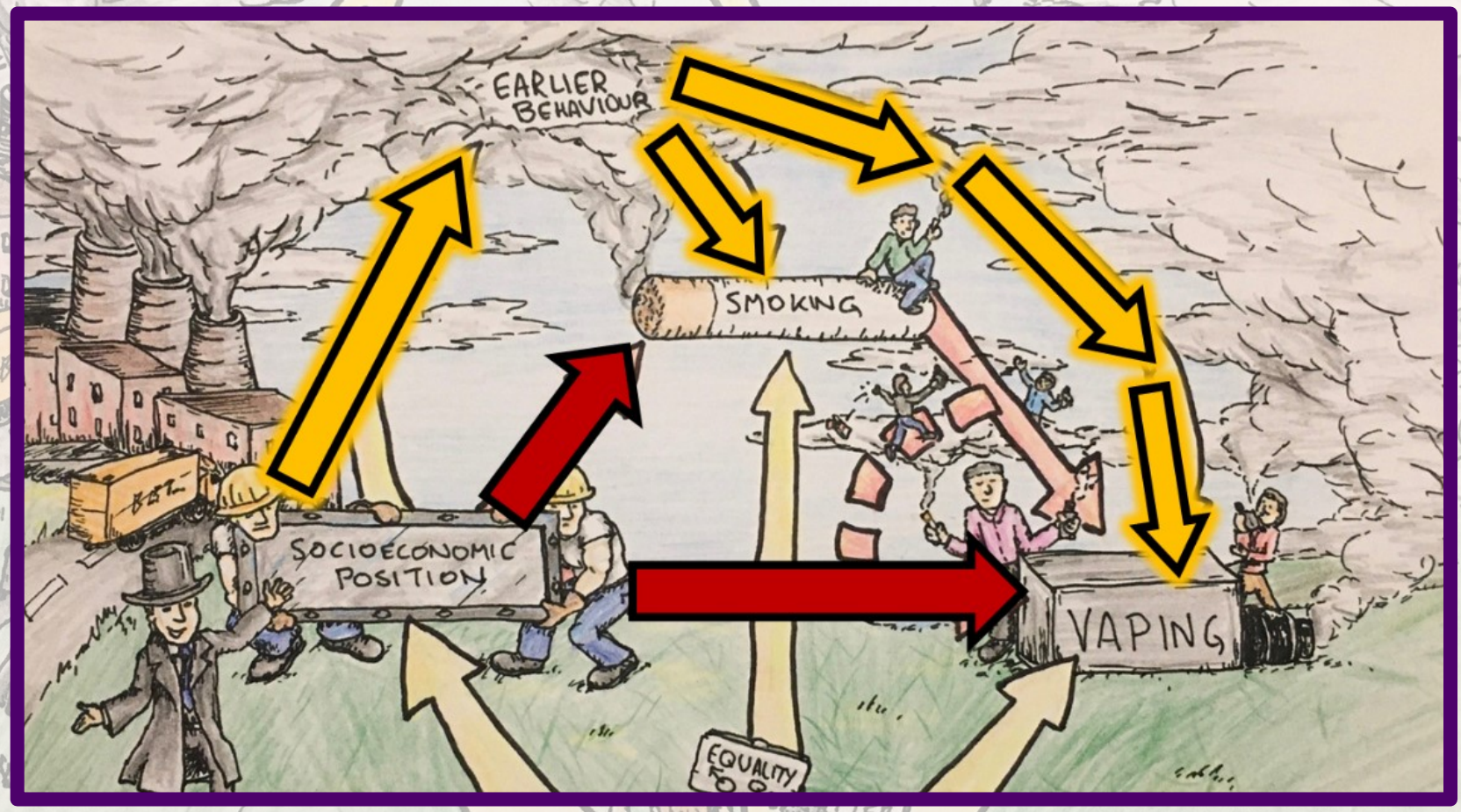
But smoking can also be caused by SES

Including smoking introduces collider bias!



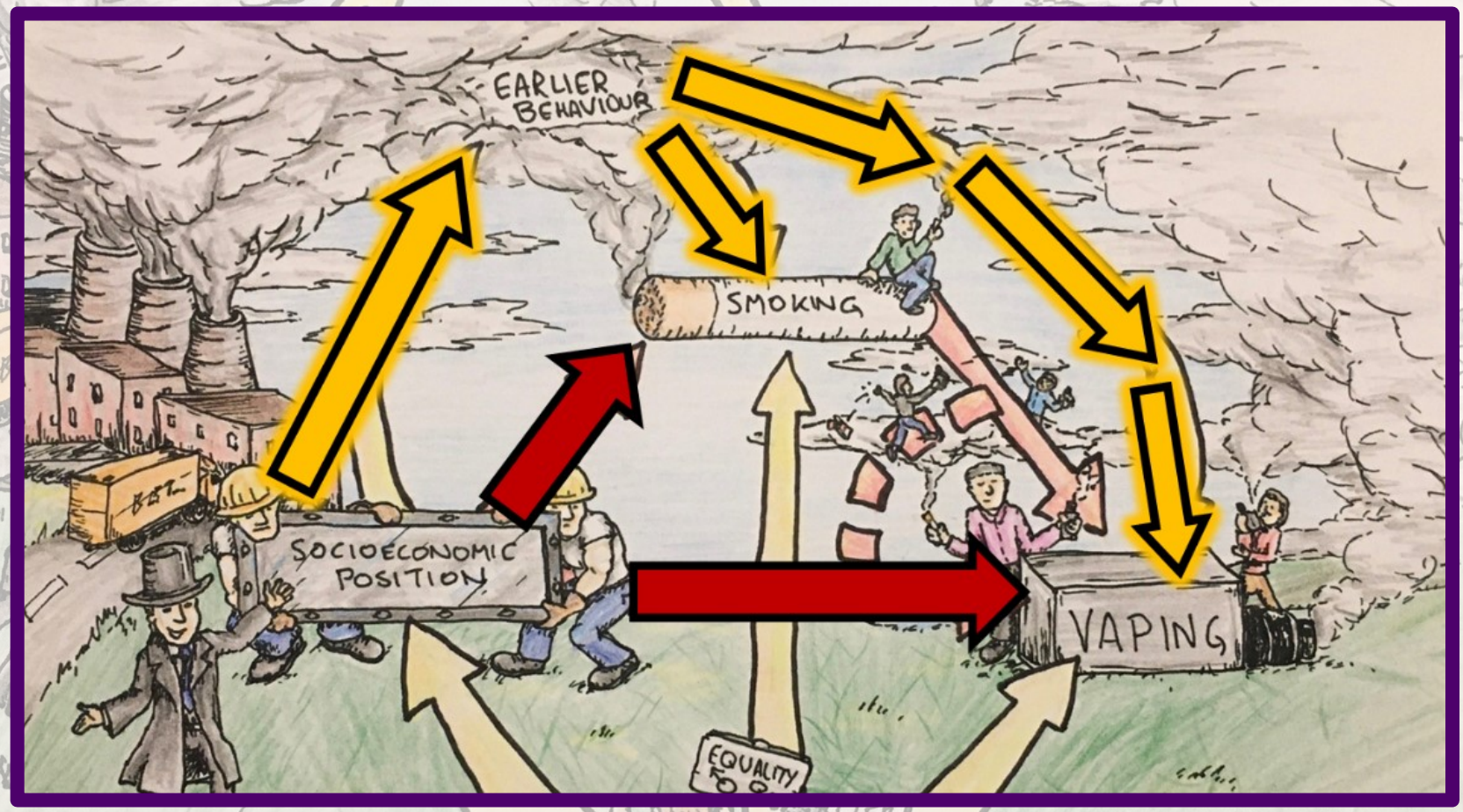
There are other confounders between SES and vaping, like earlier behavior

Adjust for earlier behavior to fix the collider that comes from adjusting for smoking

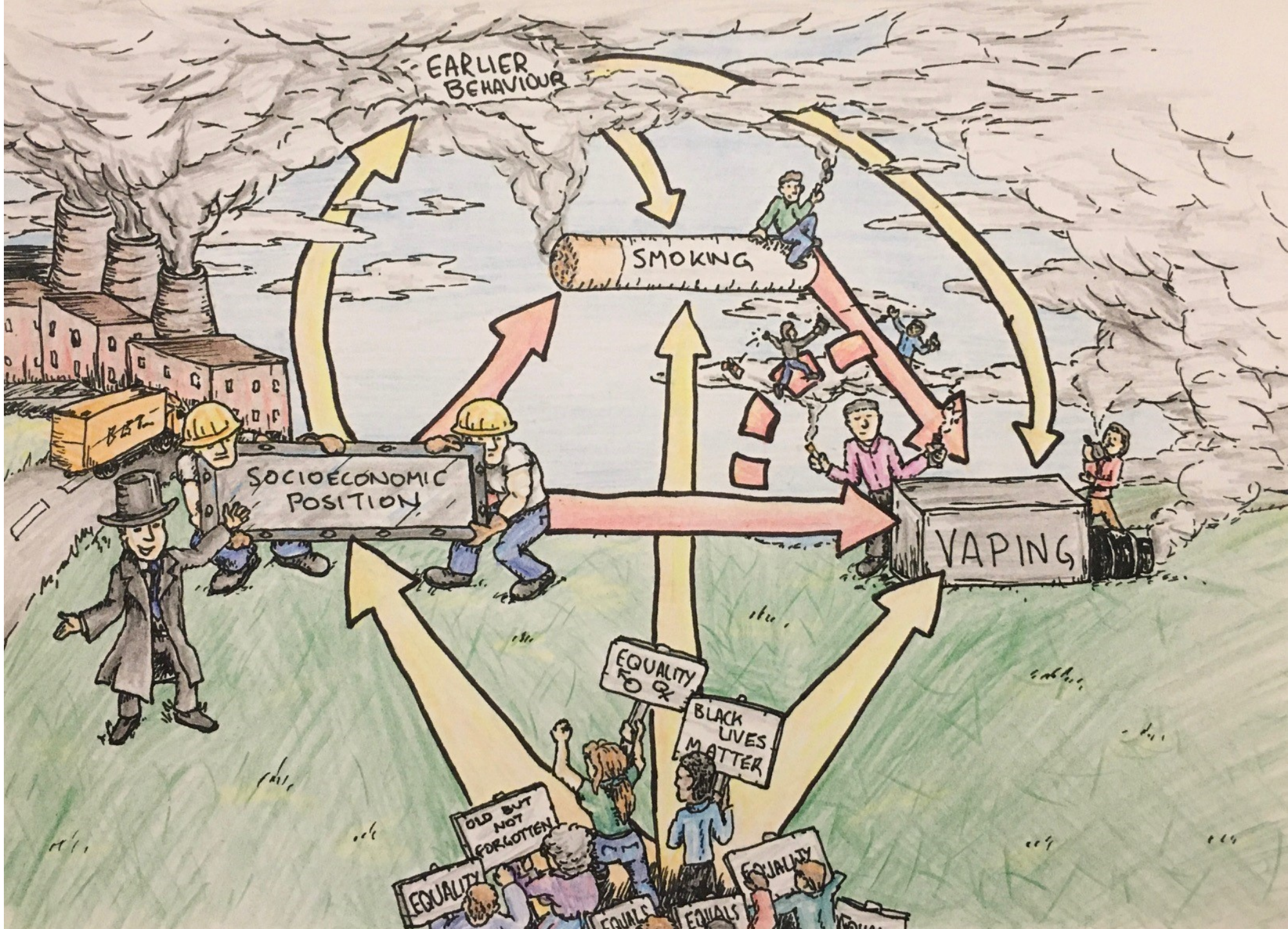


But adjusting for earlier behavior creates a bad control! It takes away some of the SES → vaping effect

oh no

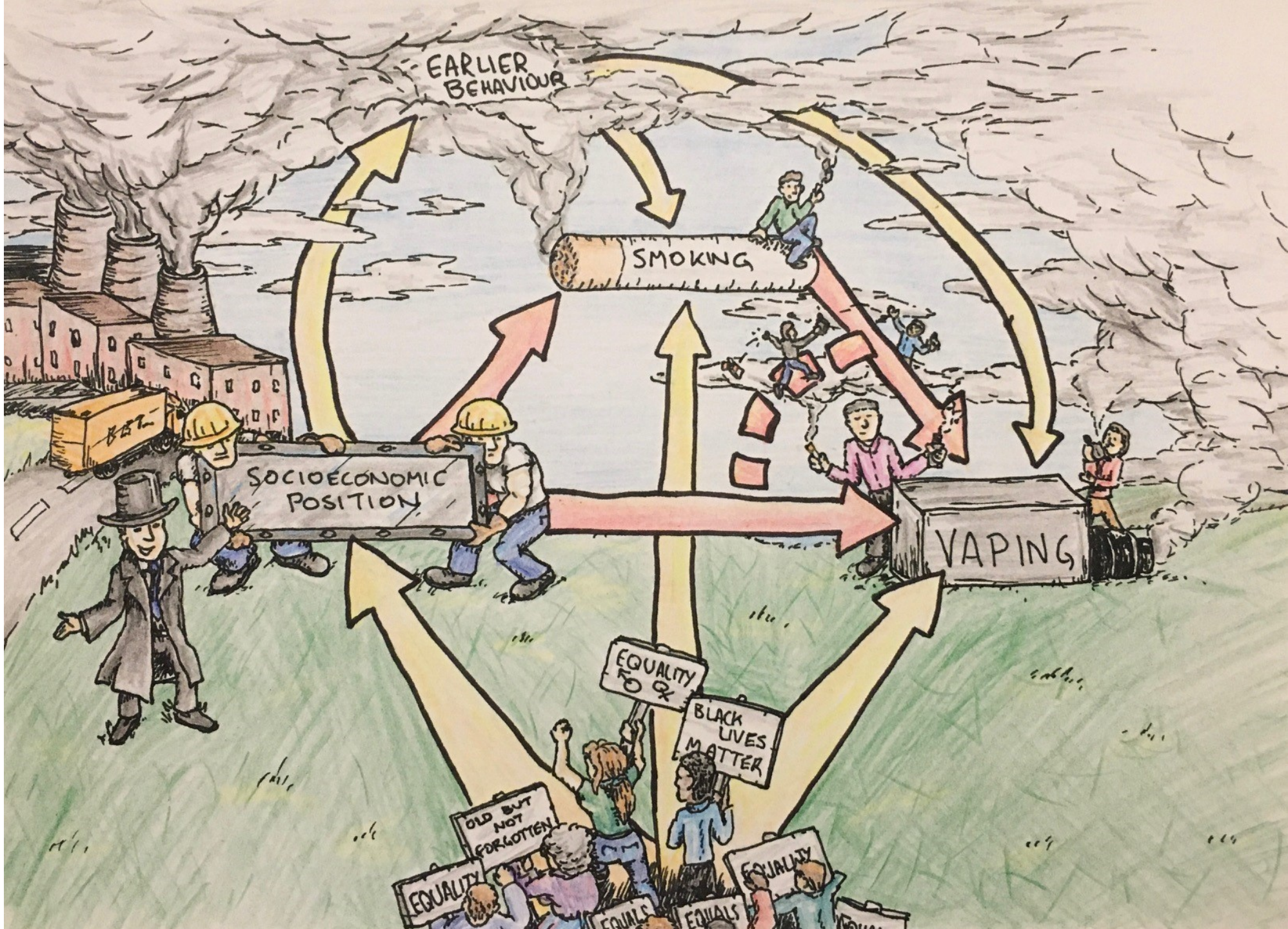


They figure out how to control for earlier behavior *and* smoking without causing a collider effect *or* causing bad controls



Main findings: Low SES causes more vaping among never-smoking youth + former-smoking adults

Low SES doesn't cause vaping among never-smoking or current-smoking adults



do()ing observational causal inference

Structural models

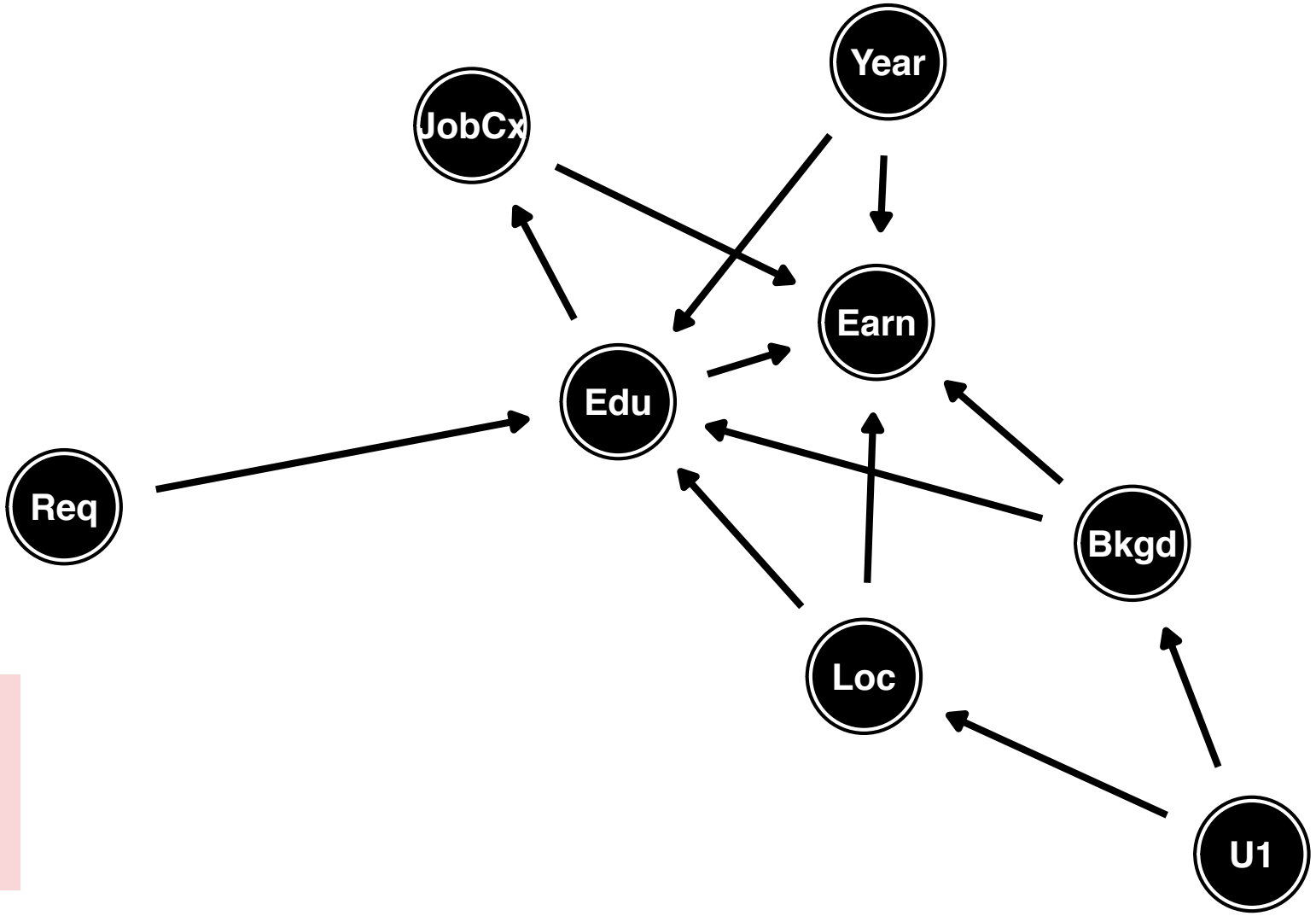
$Loc = f_L(U1)$

$Bkgd = f_B(U1)$

$JobCx = f_J(Edu)$

$Edu = f_{Edu}(Req, Loc, Year)$

$Earn = f_{Earn}(Edu, Year, Bkgd, Loc, JobCx)$



Interventions

do-operator

Marking an intervention in a DAG

$$P(Y \mid do(X = x))$$

P = probability distribution, or effect

Y = outcome; X = treatment;
x = specific value of treatment

Interventions

$$P(Y \mid do(X = x))$$

$P(\text{Firm growth} \mid do(\text{Government R\&D funding}))$

$P(\text{Wages} \mid do(\text{College}))$

$P(\text{Air quality} \mid do(\text{Carbon tax}))$

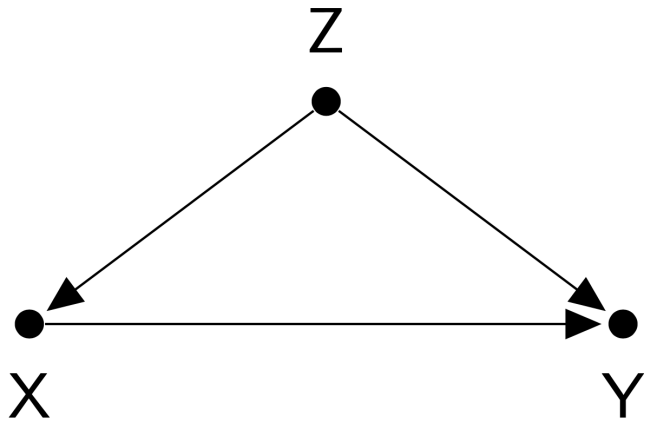
$P(\text{Juvenile delinquency} \mid do(\text{Truancy program}))$

$P(\text{Malaria infection rate} \mid do(\text{Mosquito net}))$

Interventions

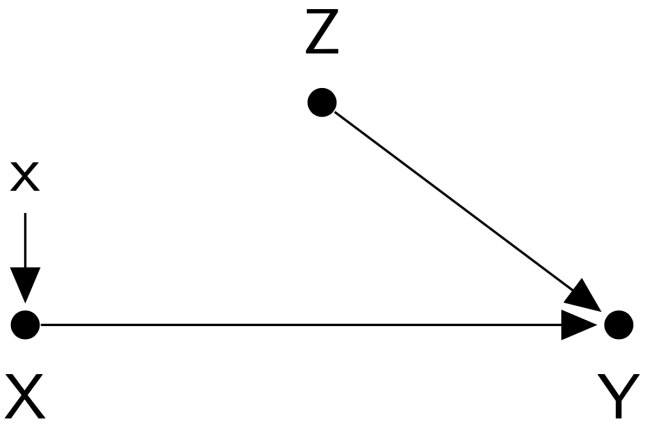
When you *do*() X, remove all arrows into it

Observational DAG



$Z = f_Z(\cdot)$ $X = f_X(Z)$ $Y = f_Y(X, Z)$

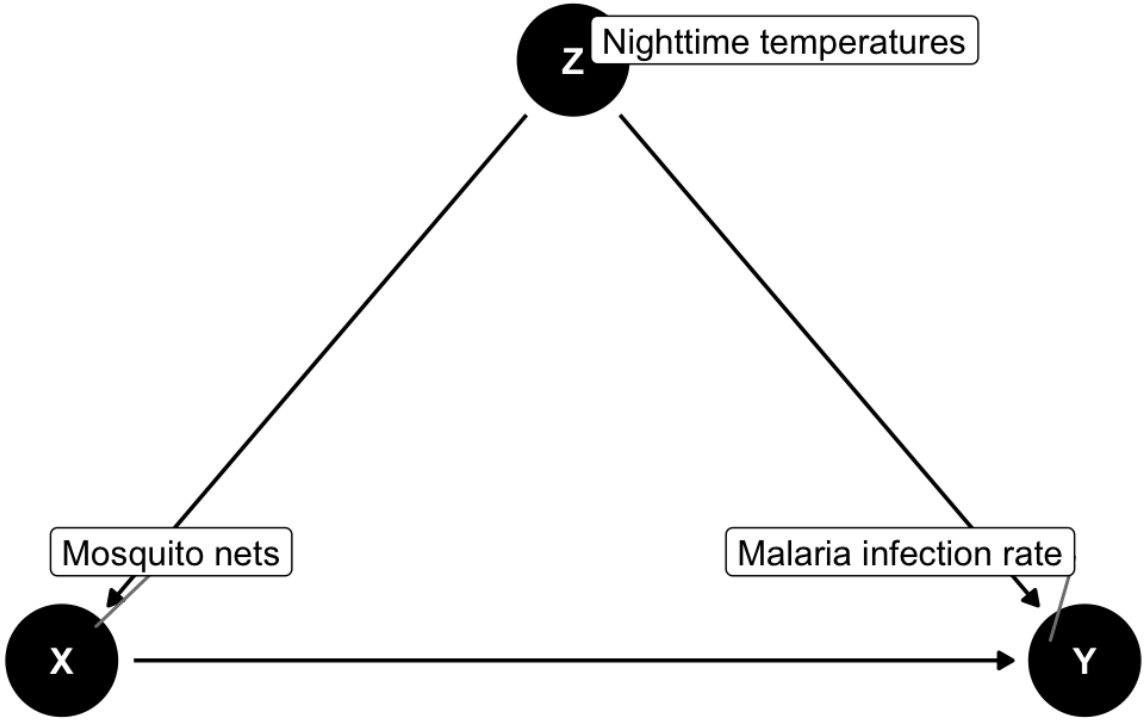
Interventional DAG



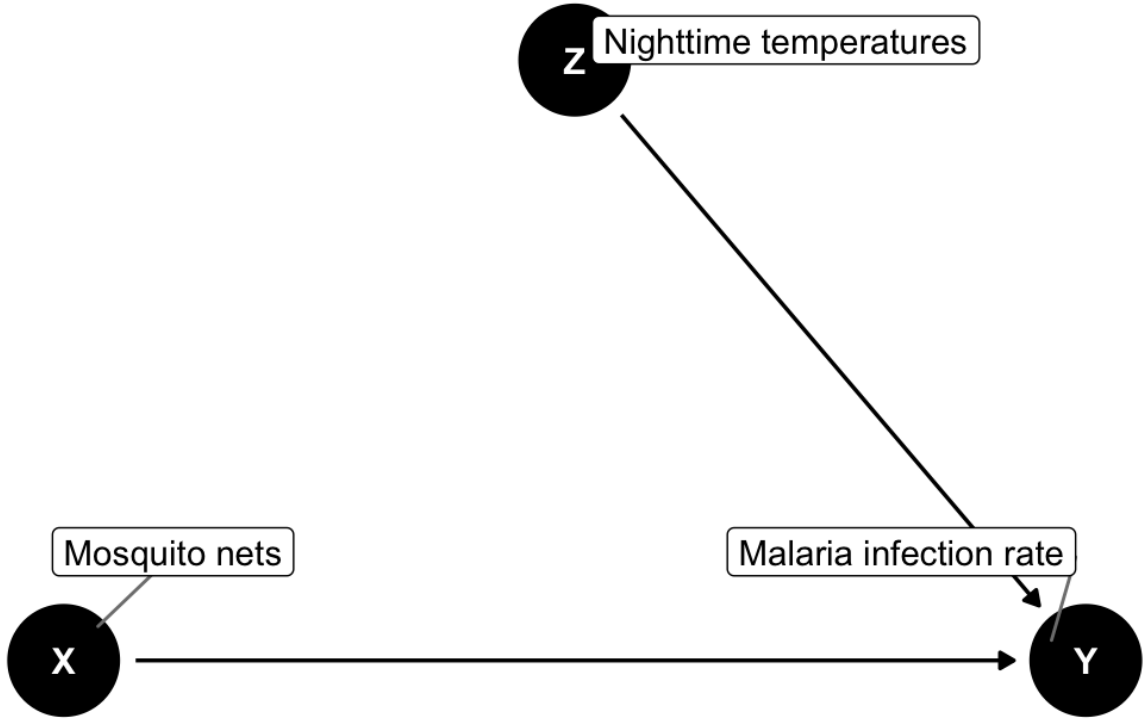
$Z = f_Z(\cdot)$ $X = x$ $Y = f_Y(X, Z)$

Interventions

P(Malaria infection rate| *do*(Mosquito net))



Observational



Experimental

Un *do*()ing things

We want to know P(Y | *do*(X)),
or P(Malaria rate | *do*(Mosquito net)),
but all we have is observational data X, Y, and Z

$$P(Y \mid do(X)) \neq P(Y \mid X)$$

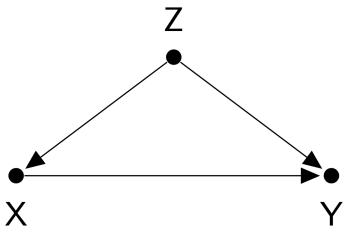
Correlation isn't causation!

We need to transform P(Y | *do*(X)) into something that is "*do*-free" and only uses observed data

Un *do*()ing things

Backdoor adjustment

Matching, regression, stratifying, inverse probability weighting

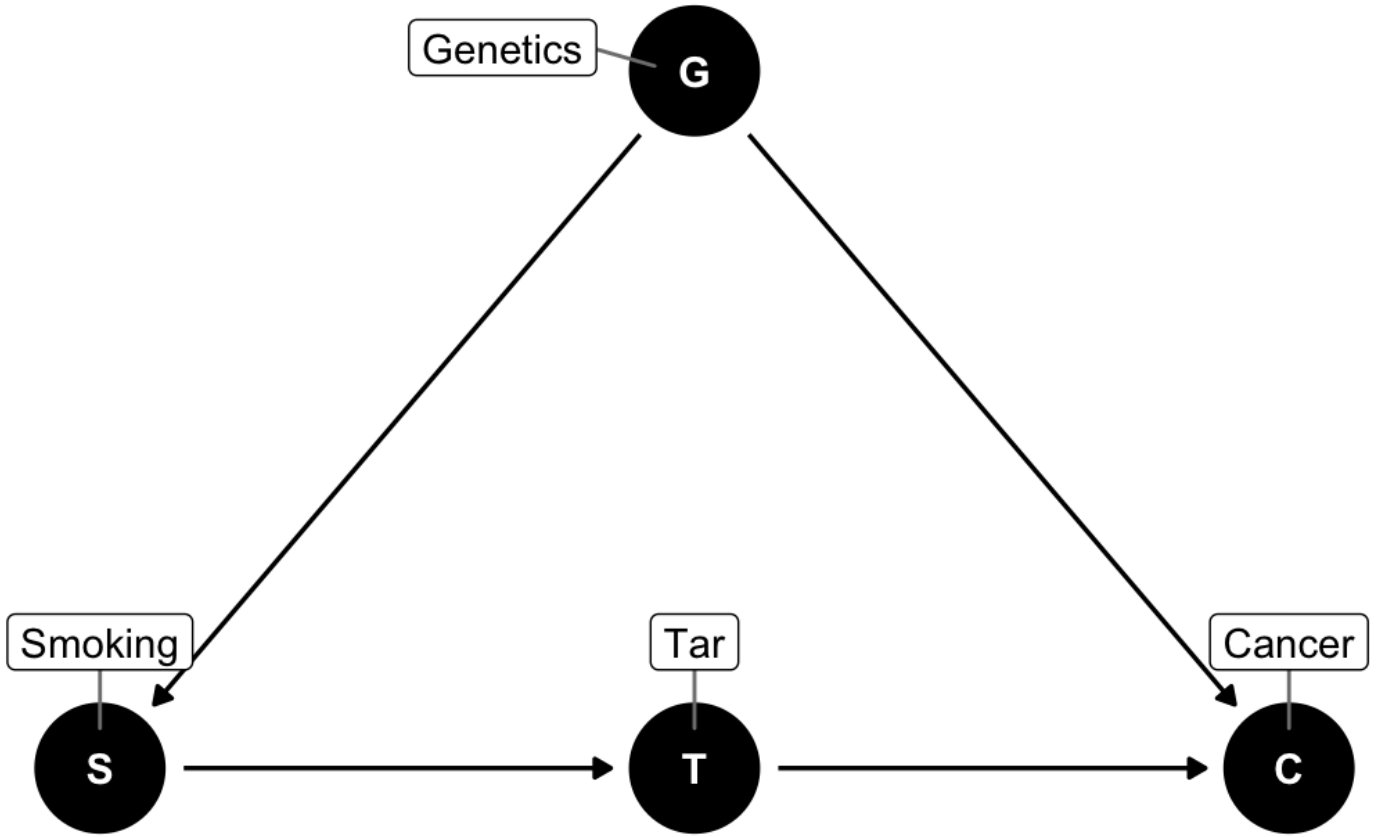


$$P(Y|do(X)) = \sum_{Z} P(Y|X,Z) \times P(Z)$$

Frontdoor adjustment

Do-calculus

Frontdoor adjustment



S → T is *d*-separated; T → C is *d*-separated; combine the effects for S → C

Do-calculus

A set of three rules that let you manipulate a DAG in special ways to remove *do*() expressions

The do-calculus Let G be a CGM, $G_{\overline{T}}$ represent G post-intervention (i.e with all links into T removed) and $G_{\underline{T}}$ represent G with all links *out of* T removed. Let $do(t)$ represent intervening to set a single variable T to t,

Rule 1: $\mathbb{P}(y|do(t), z, w) = \mathbb{P}(y|do(t), z)$ if $Y \perp\!\!\!\perp W|(Z, T)$ in $G_{\overline{T}}$

Rule 2: $\mathbb{P}(y|do(t), z) = \mathbb{P}(y|t, z)$ if $Y \perp\!\!\!\perp T|Z$ in $G_{\underline{T}}$

Rule 3: $\mathbb{P}(y|do(t), z) = \mathbb{P}(y|z)$ if $Y \perp\!\!\!\perp T|Z$ in $G_{\overline{T}}$, and Z is not a decedent of T.

WAAAAAY beyond the scope of this class!
Just know it exists and computer algorithms can do it for you

Moral of the story

If you can transform *do*() expressions to *do*-free versions, you can legally make causal inferences from observational data

Backdoor adjustment

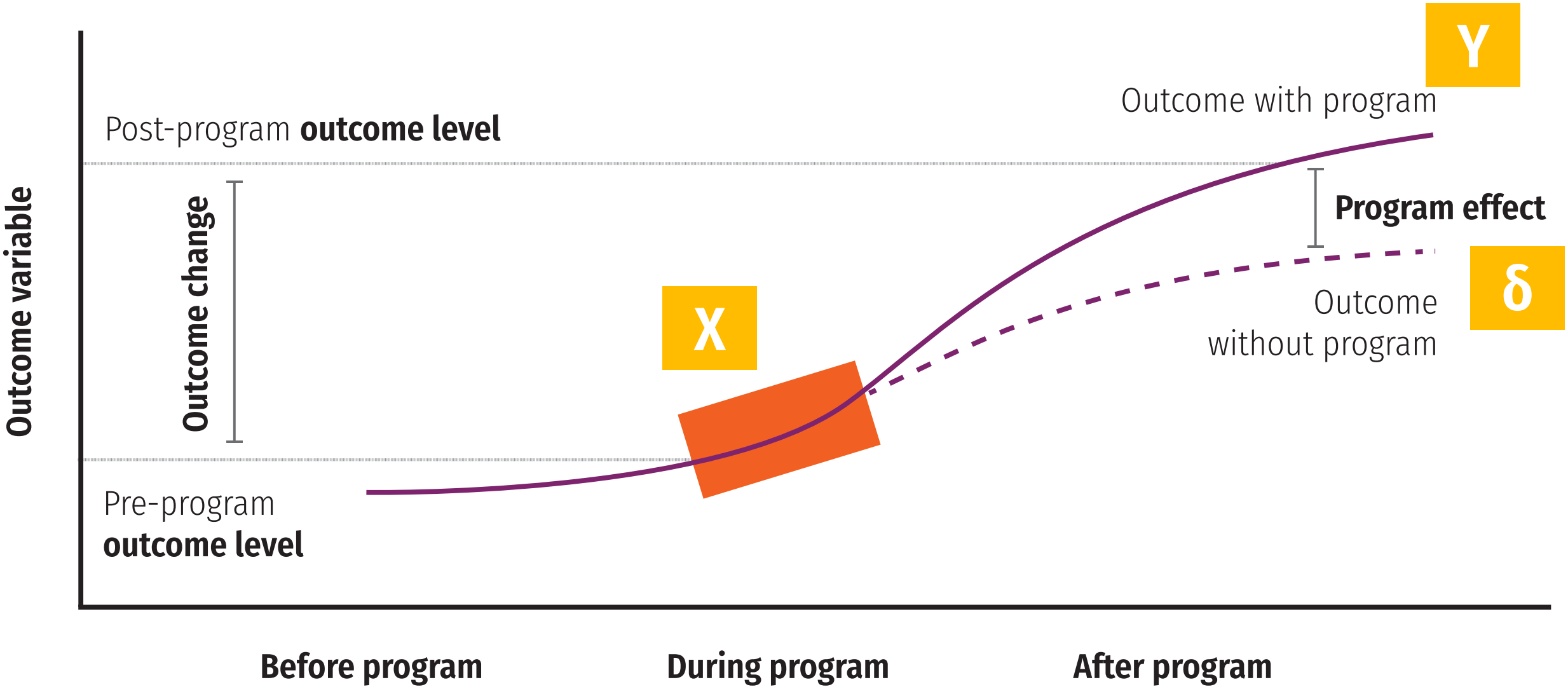
Frontdoor adjustment

***Do*-calculus**

Calculating adjustment sets with R

Potential outcomes

Program effect



Some equation translations

P = probability distribution

$$\delta = P(Y|do(X))$$

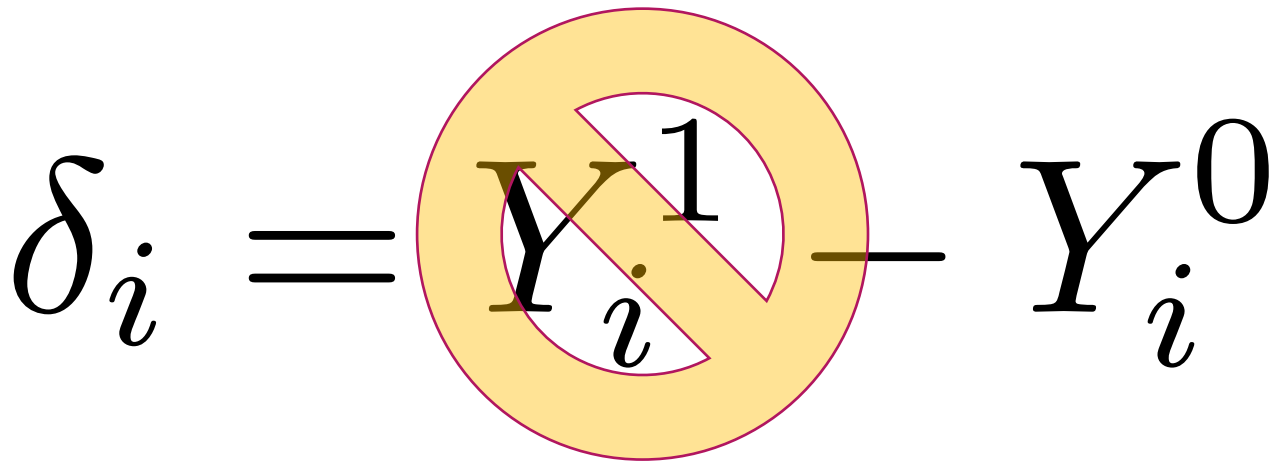
E = expected value, or average

$$\delta = E(Y|do(X)) - E(Y|!do(X))$$

$$\delta = (Y|X = 1) - (Y|X = 0)$$

$$\delta = Y_1 - Y_0$$

Fundamental problem of causal inference



Individual-level effects are impossible to observe!

No individual counterfactuals!

Average treatment effect (ATE)

Solution: Use averages instead

$$ATE = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$$

Difference between average/expected value when program is on vs. expected value when program is off

$$\delta = (\bar{Y}|P = 1) - (\bar{Y}|P = 0)$$