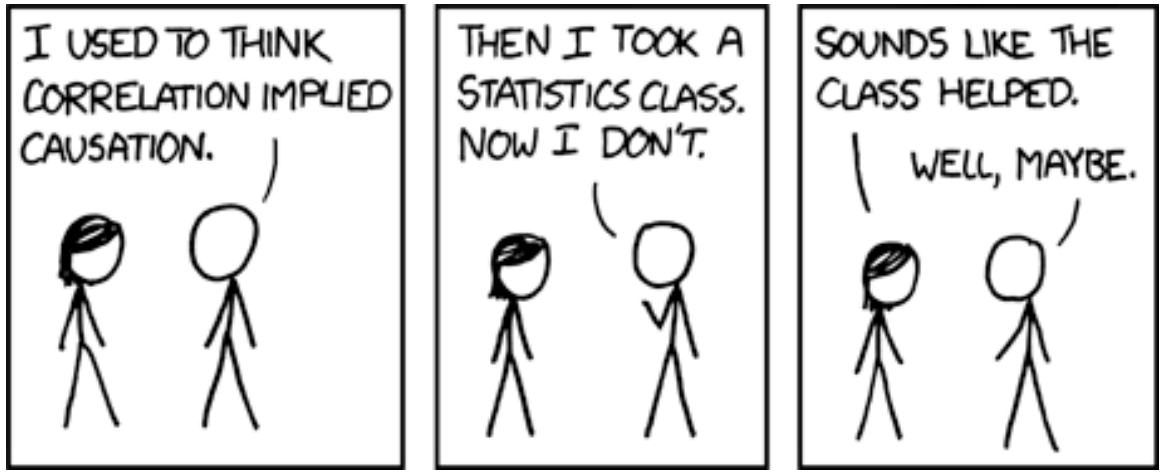
An introduction to causal graphs & Causal graphs for missing data

Julia M. Rohrer
Leipzig University
www.juliarohrer.com
@dingdingpeng.the100.ci
www.the100.ci

Slides: Resources

Causal Inference 101



- »Option 1: Run an experiment
- »Option 2: Give up
  - » include covariates, maybe?
  - » or maybe use longitudinal data???

Comic: https://xkcd.com/552/

It's bad for non-experimental research

1. virtually all interesting research questions concern causality
2. observational data is not admissible for causal inference

1. + 2. = ???

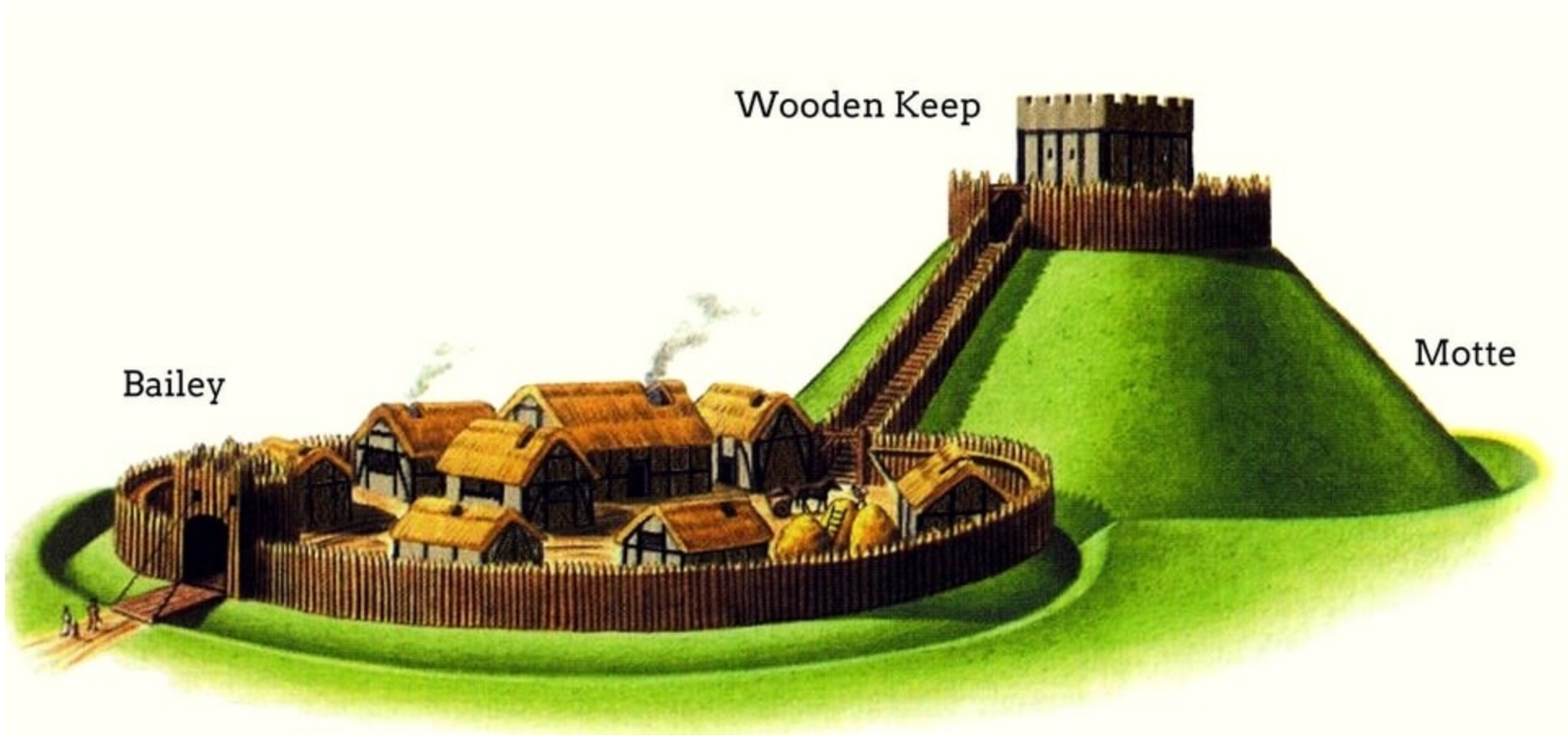
Grosz, Rohrer, & Thoemmes (2020)

The non-experimental psych workaround

- introduction: relies on a causal reading of the literature
- methods & results: implicitly (but never explicitly) causal inference-y
  - X predicts Y…even after accounting for…
  - X is a risk factor for Y
  - longitudinal associations
- discussion: only makes sense in terms of causality
- **IMPORTANT:** add a paragraph that your study was only observational and no causal conclusions are warranted
  - future longitudinal or experimental studies will surely fix this problem

Grosz, Rohrer, & Thoemmes (2020)

Causal inference Motte-and-Bailey



X has a causal effect on Y.

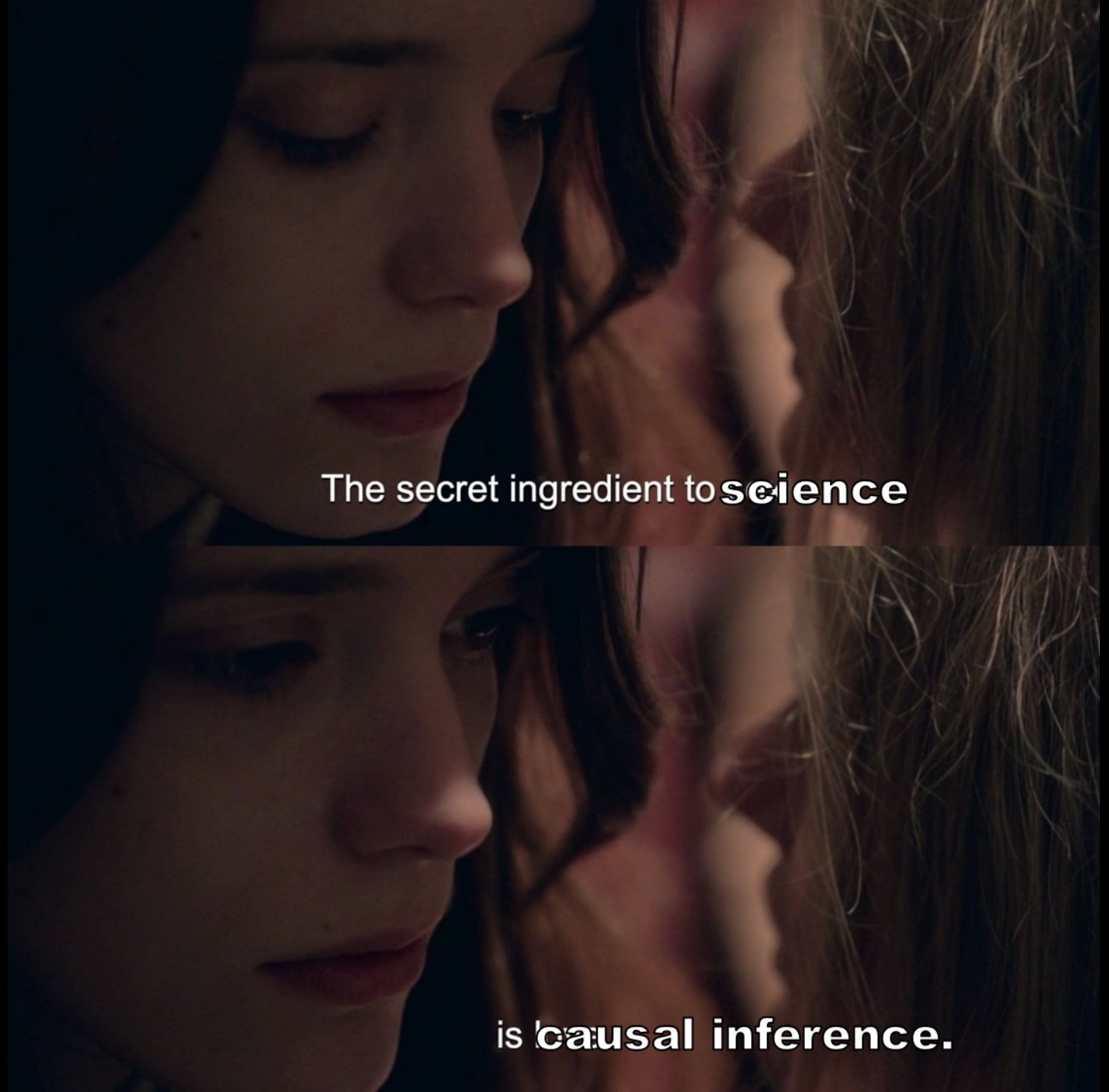
X „predicts“ Y.

[Grosz, Rohrer, & Thoemmes (2020)]

It's bad for experimental researchers

- Causal inference issues that remain despite randomization
  - **Missing data**
  - Mediation analysis
  - Moderation analysis
  - Generalization across populations, settings, independent and dependent variables
  - Measurement
  - …


TRESKO


The secret ingredient to science
is causal inference.

Causal inference issues

Identification

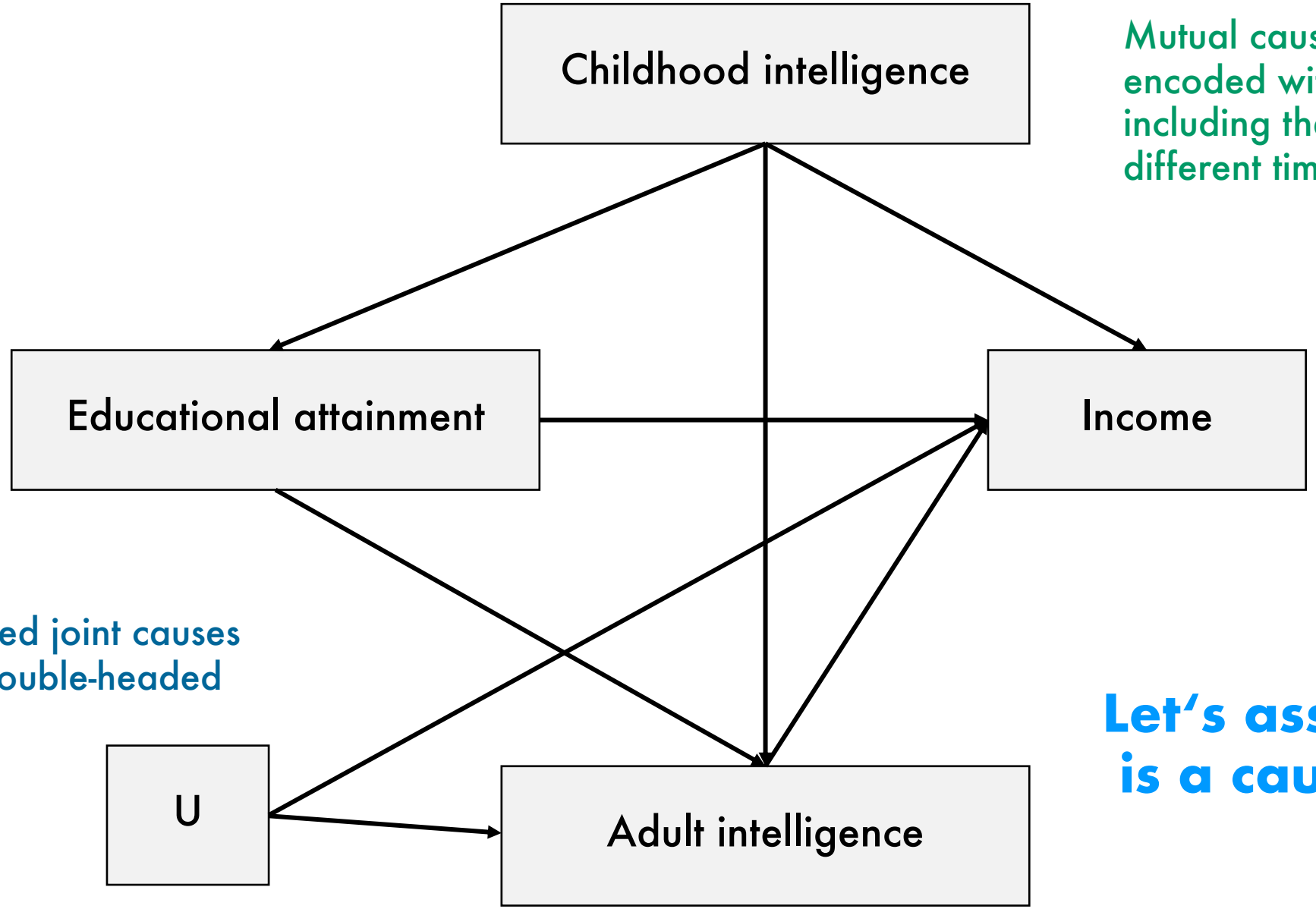
Given your assumptions, if your sample was infinitely large—would you be able to estimate the causal effect of interest?

Estimation

Actually estimating the effect with the data you got + your assumptions

Elwert (2013)

Directed Acyclic Graphs



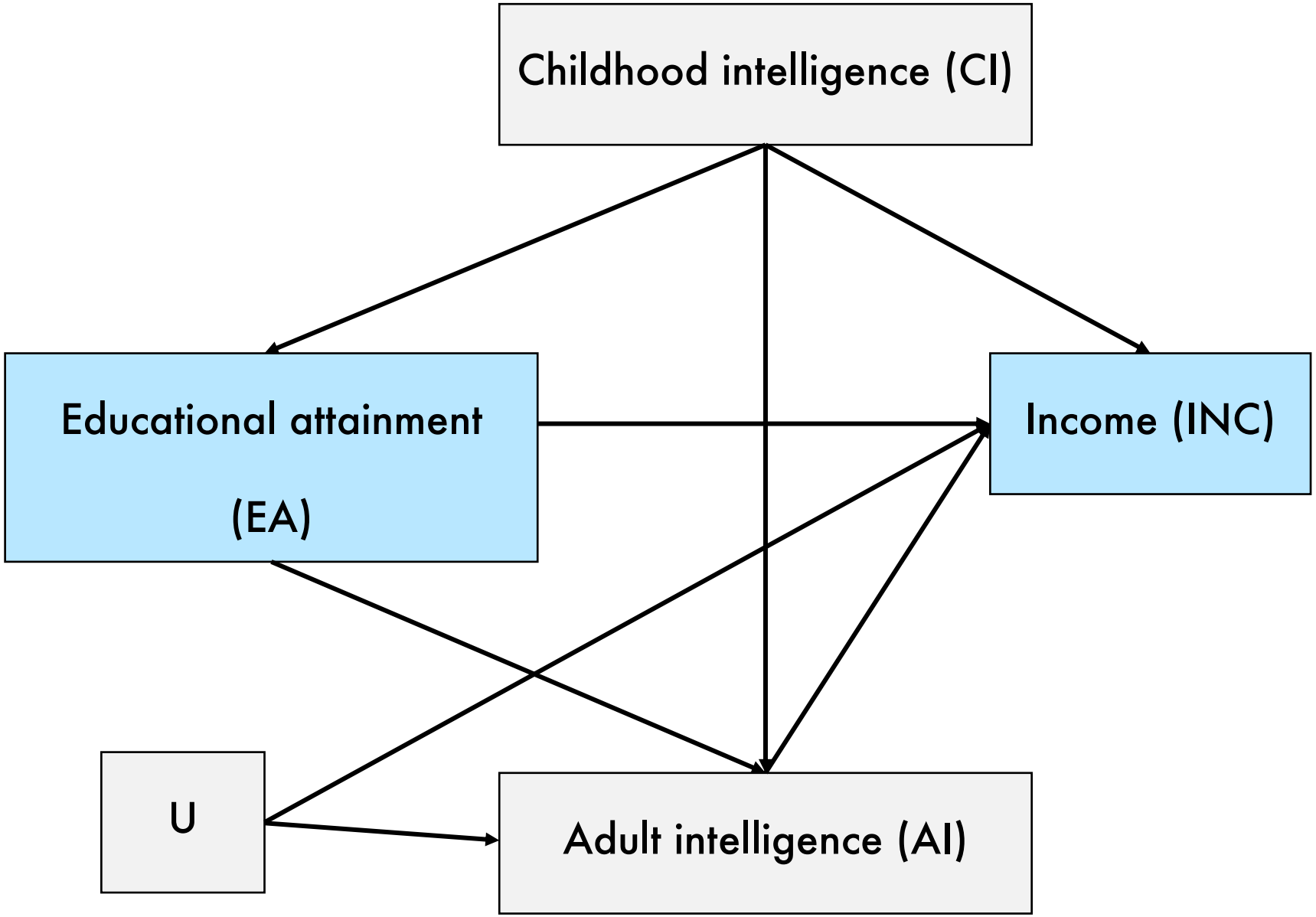
Mutual causation can be encoded without cycles by including the same variable at different time points

Unobserved joint causes replace double-headed arrows

Let's assume this is a causal DAG

Rohrer (2018)

Backdoor Criterion



EA -> INC
EA <- CI -> INC

EA <- CI -> AI -> INC
EA <- CI -> AI <- U -> INC

EA -> AI -> CI -> INC
EA -> AI -> INC
EA -> AI <- U -> INC

Which of these paths lead to associations between EA and INC?

Are these associations causal or non-causal?

→ keep the causal ones, block the non-causal ones

3 Fundamental structures

- Chains: X → M → Y
  - transmits a causal association from X to Y
  - conditioning on (control for) M blocks transmission

Elwert (2013), Rohrer (2018)

What does it mean to condition on a variable?

- Any means of introducing information about the variable into the analysis (using it as a control variable, a covariate…)
- For example
  - Statistical adjustment in an ANCOVA, regression, SEM…
  - Using the variable for matching, weighting
  - Stratification by the variable
  - Only including participants with a certain level of a variable

3 Fundamental structures

- Chains: X → M → Y
  - transmits a causal association from X to Y
  - conditioning on (control for) M blocks transmission
    - if you're interested in the (total) effect of X on Y, you usually don't want to do that, because the chain is part of the causal effect of interest → control for mediator leads to overcontrol bias
  - aka mediation

Elwert (2013), Rohrer (2018)

3 Fundamental structures

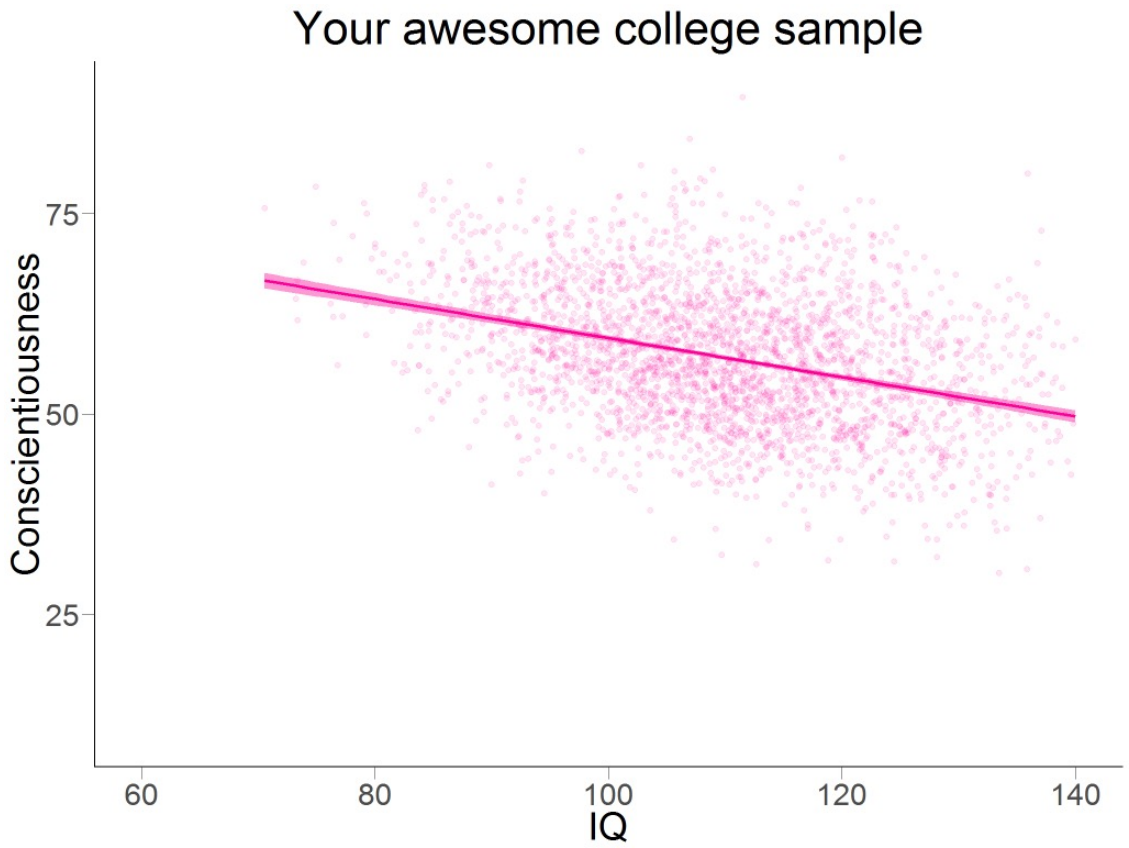
- Forks: X ← C → Y
  - induces a non-causal association between X and Y
  - conditioning on C removes the non-causal association
    - you usually want to get rid of these, so conditioning on confounders is usually the way to go
  - aka confounding

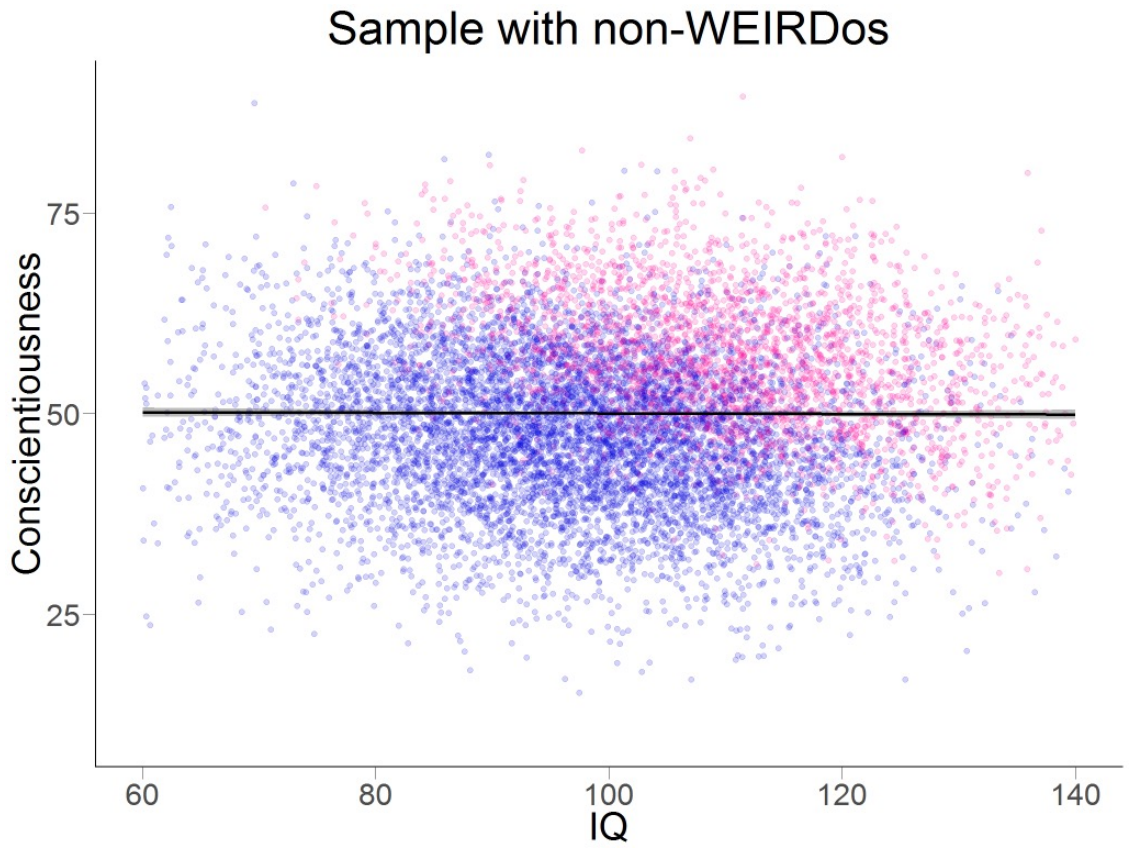
Elwert (2013), Rohrer (2018)

Inverted forks X → Z ← Y

- does not transmit any association
  - e.g., sickness → attendance of this talk ← interest in causal inference
- conditioning on Z *opens* transmission of a non-causal association
  - e.g., conditioning on attendance
  - among the people who attend this talk, those who are sick probably are particularly interested in causal inference
- in other words, third-variable control *introduces* a new spurious association
- aka collider bias

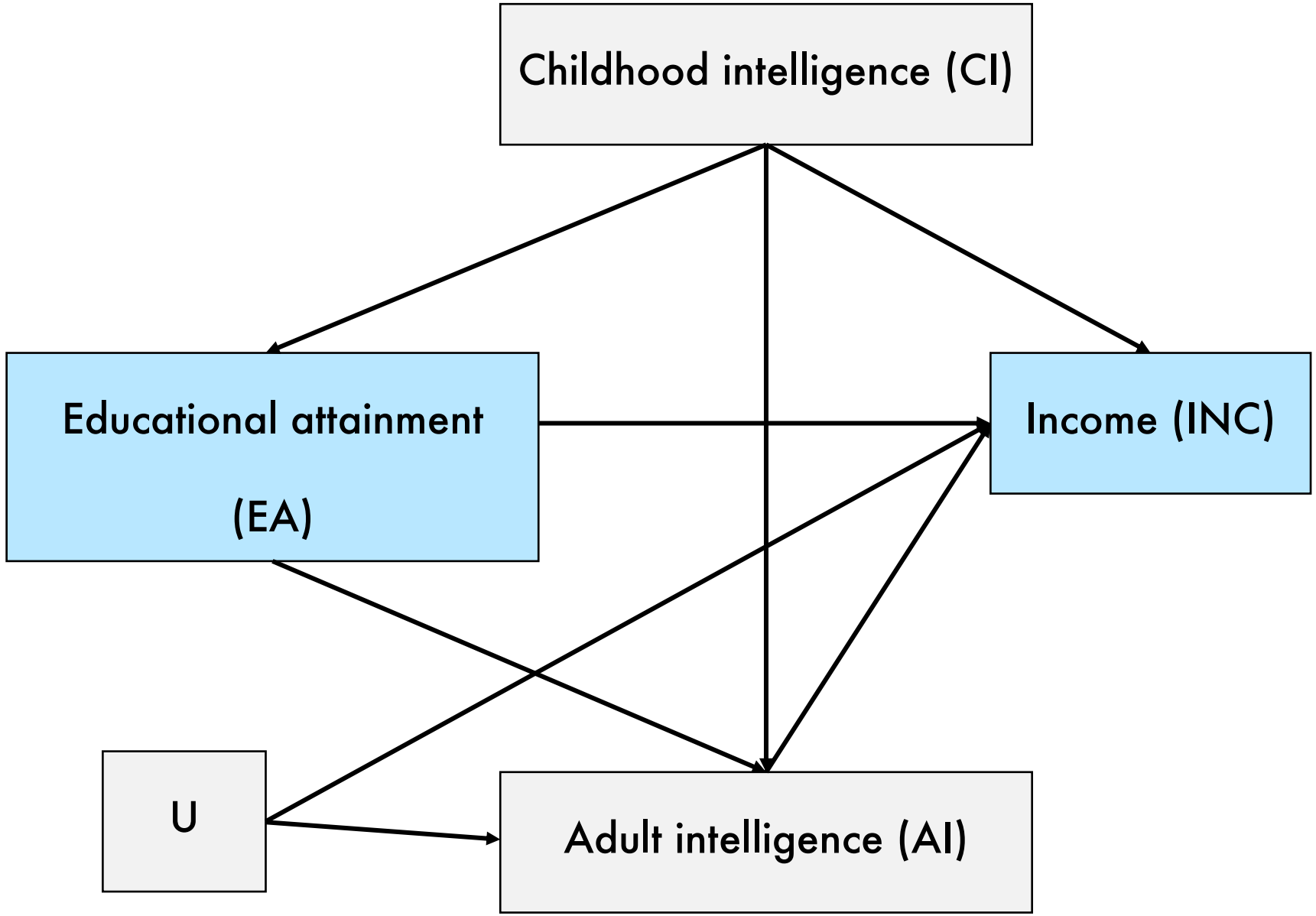
Conscientiousness → College ← IQ





Rohrer (2017)

	Association between X and Y when not conditioning on anything	Association between X and Y when conditioning on the variable in the middle
Chain: X → Mediator → Y	**Causal association**	No association (overcontrol bias)
Fork: X ← Confounder → Y	Non-causal association	**No association**
Inverted fork: X → Collider ← Y	**No association**	Non-causal association (collider bias)



EA -> INC
EA <- CI -> INC

EA <- CI -> AI -> INC
EA <- CI -> AI <- U -> INC

EA -> AI -> INC
EA -> AI <- U -> INC

Paths

- Paths that transmit non-causal associations and need to be blocked (backdoor paths)
  - EA <- CI -> INC
  - EA <- CI -> AI -> INC

Confounder that introduces the non-causal association

Conditioning on the confounder closes both backdoor paths

Mediator passes on the non-causal association

Conditioning on the mediator AI would close this one backdoor path (but there are other reasons why we wouldn't want to do that)

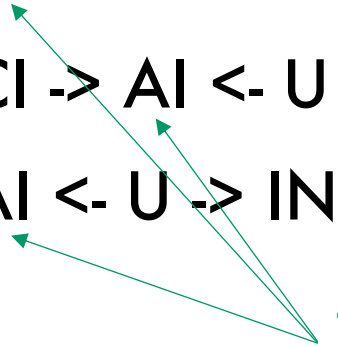
Paths

- Paths that transmit causal associations and should not be blocked
  - EA -> INC
  - EA -> AI -> INC

Arrows all flow from cause to outcome – all fine, don't control away the good stuff

Paths

- Paths that are blocked thanks to a collider…
- …but would lead to non-causal associations *if* the collider was conditioned on
  - EA -> AI <- CI -> INC
  - EA <- CI -> AI <- U -> INC
  - EA -> AI <- U -> INC



Two arrows pointing into node → collider → this path doesn't do anything, you're all good
UNLESS you condition on the collider

Paths

» What happens if you condition on the collider?

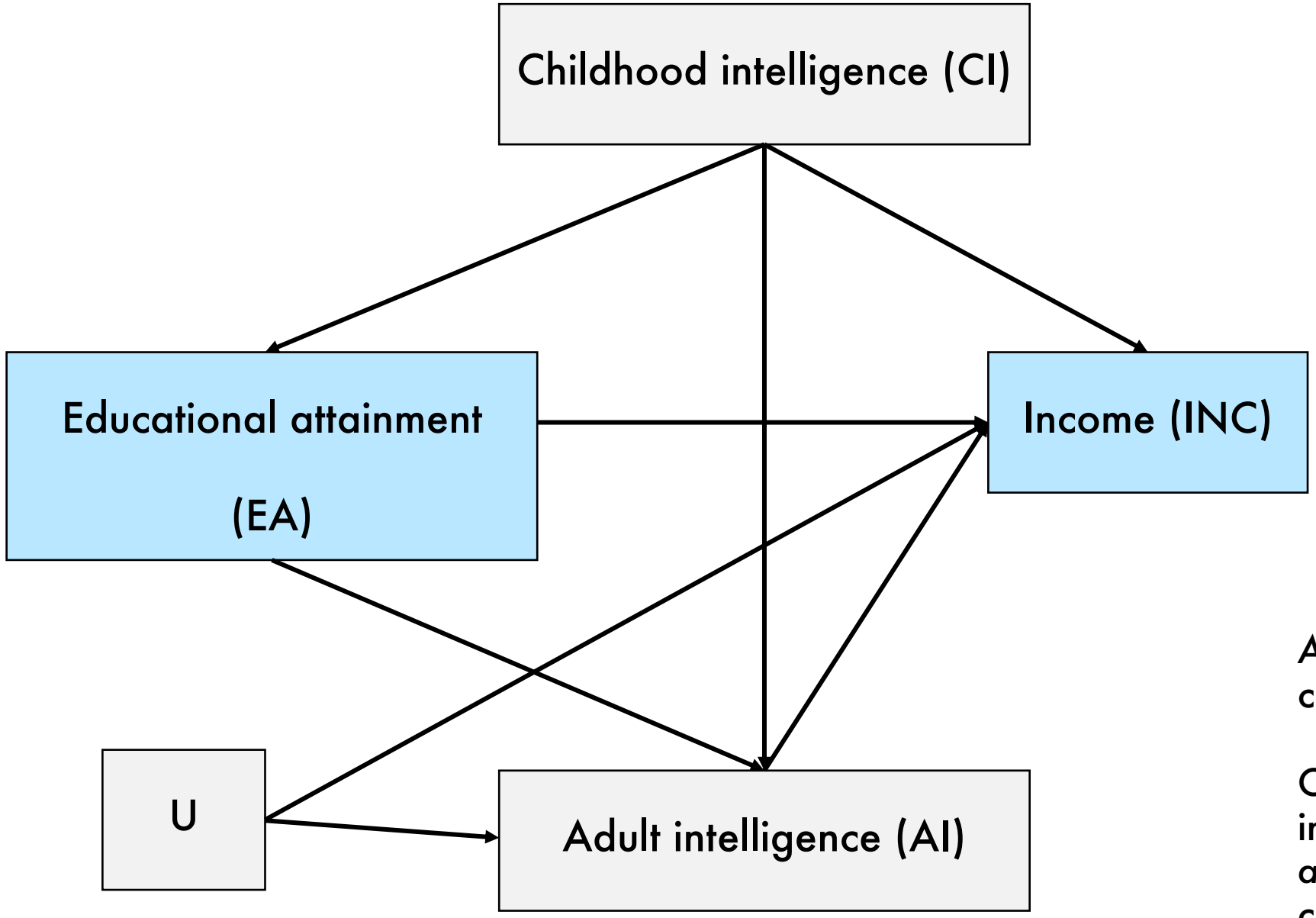
EA -> AI <- CI -> INC

EA <- CI -> AI <- U -> INC

New backdoor paths that could be fixed by conditioning on CI

EA -> AI <- U -> INC

New backdoor paths that can only be fixed if U is measured and conditioned on



Assuming that this is the correct causal structure:

Conditioning on childhood intelligence is both sufficient *and* necessary to identify the causal effect of interest

Causal estimation

- different ways to condition on a variable, for example
  - stratification, sub-group analysis
  - regression adjustment
  - weighting & matching approaches
- if your causal identification strategy was wrong, no amount of estimation can rescue you
- if your causal identification strategy was right, things can still go wrong
  - Insufficient control due to misspecified models, measurement error in covariates…

Causal graphs for missing data

Missing data

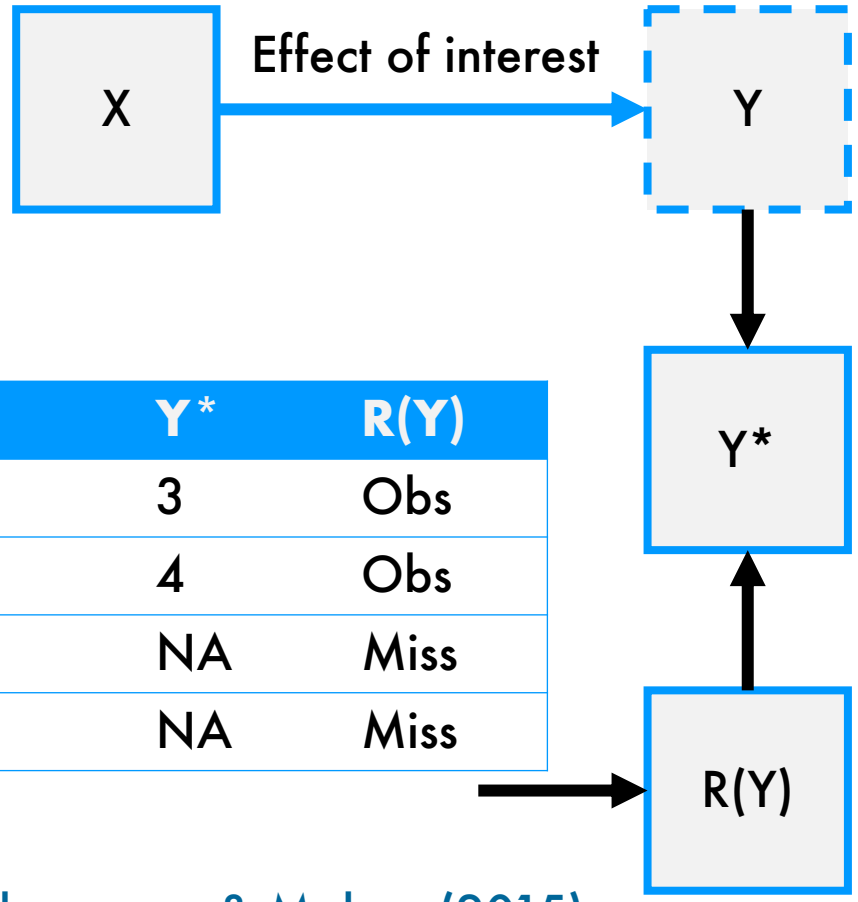
- Not all participants always provide all necessary data
- Often presented as a purely statistical/predictive problem
  - how can we predict the missing values as well as possible?
- Somewhat confusing terminology
  - MCAR, MAR, NMAR (or MNAR)

Heeeeey, MCARena!

Missing data

- Causal angle (Thoemmes & Mohan, 2015)
  - Focus on *why* the data are missing
  - Which variables need to be accounted for to recover an unbiased estimate of some effect of interest
  - Sometimes, using a variable that predicts the outcome for imputation purposes can make things *worse*

Missing Completely at Random



Y	Y*	R(Y)
3	3	Obs
4	4	Obs
3	NA	Miss
4	NA	Miss

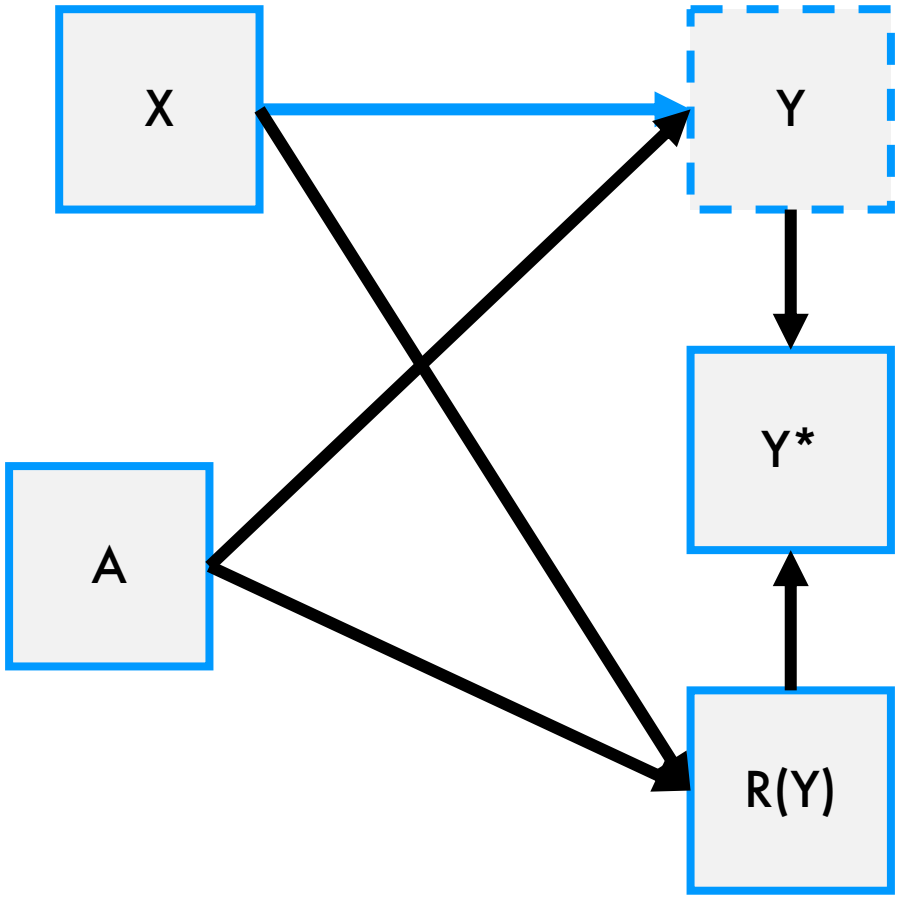
Thoemmes & Mohan (2015)

Crucial question to recover the effect of interest:
Can we we separate Y and R(Y)?

Here:
Y → Y* ← R(Y) blocked because of the collider

Nothing *needs* to be done fo an unbiased estimate (complete cases analysis would work)

Missing at Random

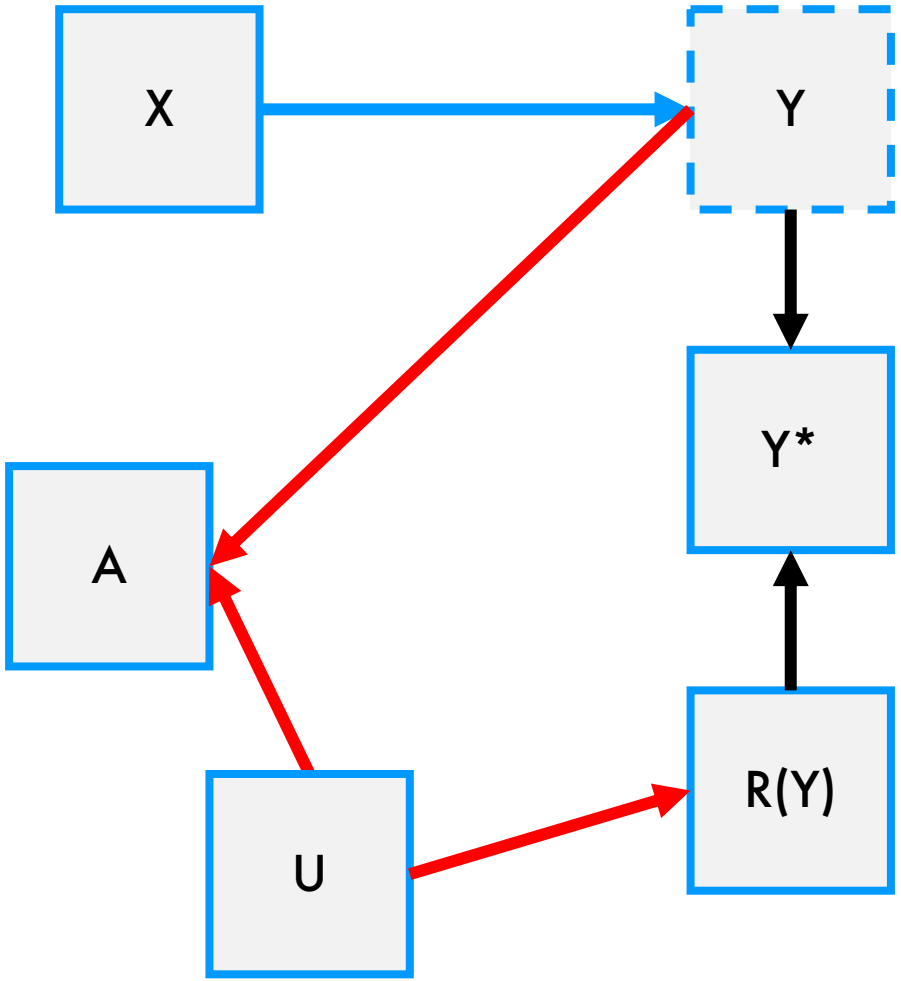


Two open paths between Y and R(Y):
Y ← X → R(Y)
Y ← A → R(Y)

Block these! e.g.
- multiple imputation
- FIML estimation
- inverse probability weighting

using both X and A

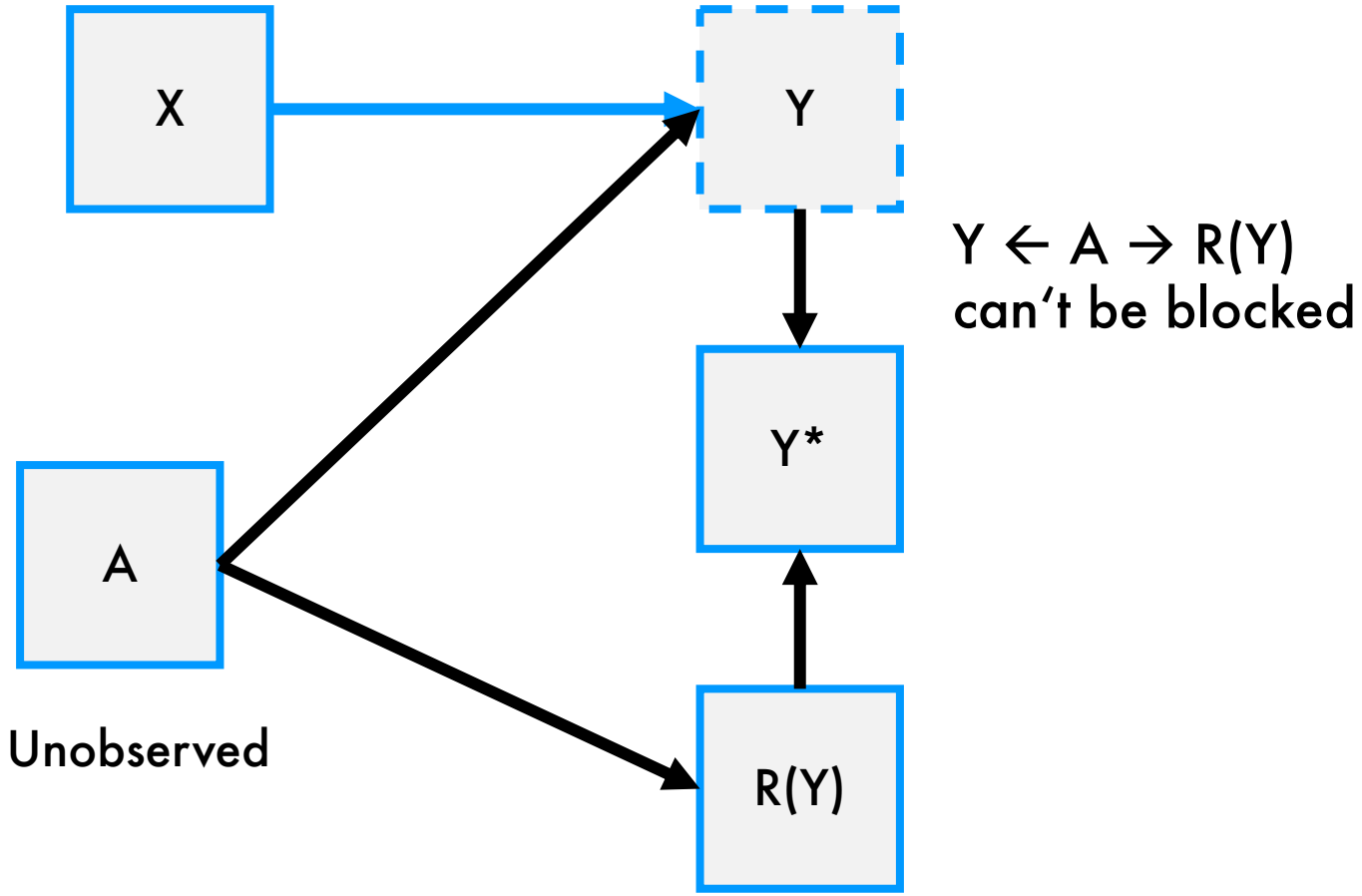
Missingness is a Causal Inference Problem



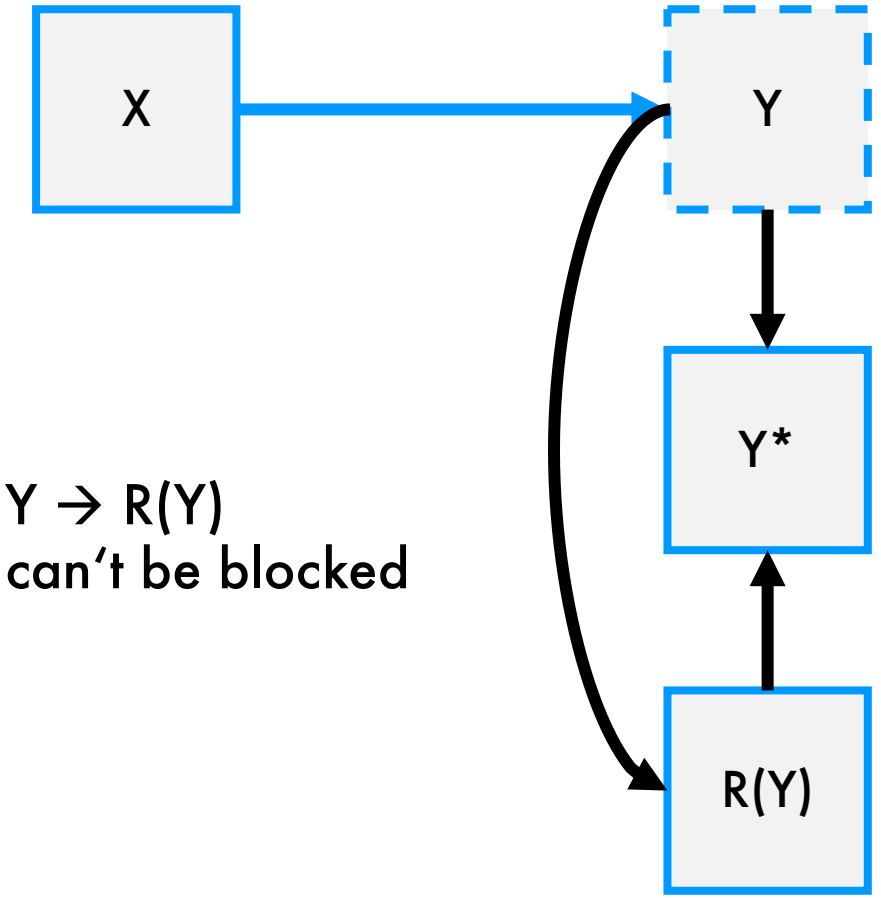
$Y \rightarrow A \leftarrow U \leftarrow R(Y)$
is blocked because A is a collider

Introducing information about A (e.g., by using it as an auxiliary variable) will *bias* estimates (unless U is also included)

Not Missing at Random



$Y \leftarrow A \rightarrow R(Y)$
can't be blocked



$Y \rightarrow R(Y)$
can't be blocked

DAGitty — draw and analyze causal diagrams

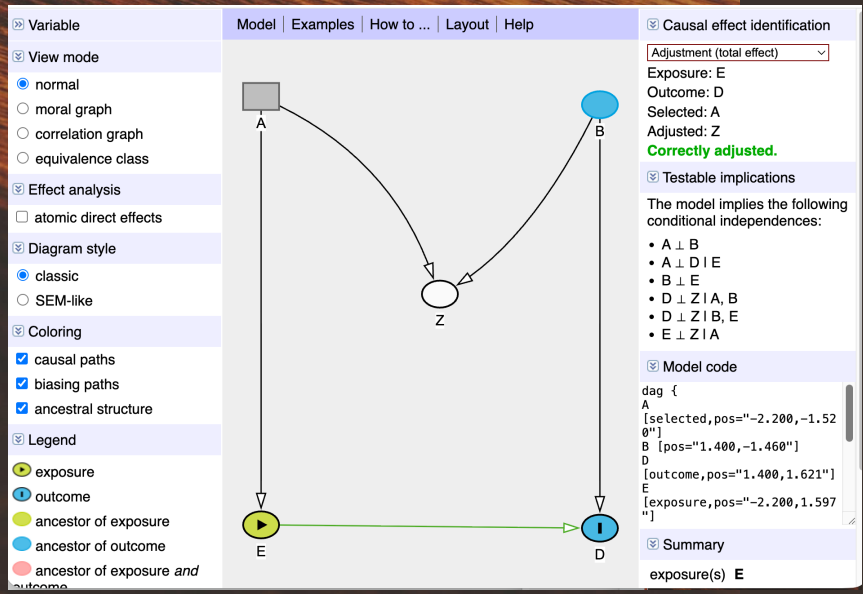
DAGitty is a browser-based environment for creating, editing, and analyzing causal diagrams (also known as directed acyclic graphs or causal Bayesian networks). The focus is on the use of causal diagrams for minimizing bias in empirical studies in epidemiology and other disciplines. For background information, see the "learn" page.

Launch	Download	Learn	Code
Launch DAGitty online in your browser.	Download DAGitty's source for offline use.	Learn more about DAGs and DAGitty.	The R package "dagitty" is available on CRAN or github.

DAGitty is developed and maintained by Johannes Textor (Institute for Computing and Information Sciences, Radboud University, and Medical BioSciences department, Radboudumc, Nijmegen, The Netherlands).

Many algorithms implemented in DAGitty were developed in close collaboration with Maciej Liśkiewicz and Benito van der Zander, University of Lübeck, Germany (see literature references below).

DAGitty development happens on github. You can download all source code from there and also get involved.



Thank you for your attention!

Julia M. Rohrer
Leipzig University
www.juliarohrer.com
@dingdingpeng.the100.ci
www.the100.ci

Slides: Resources