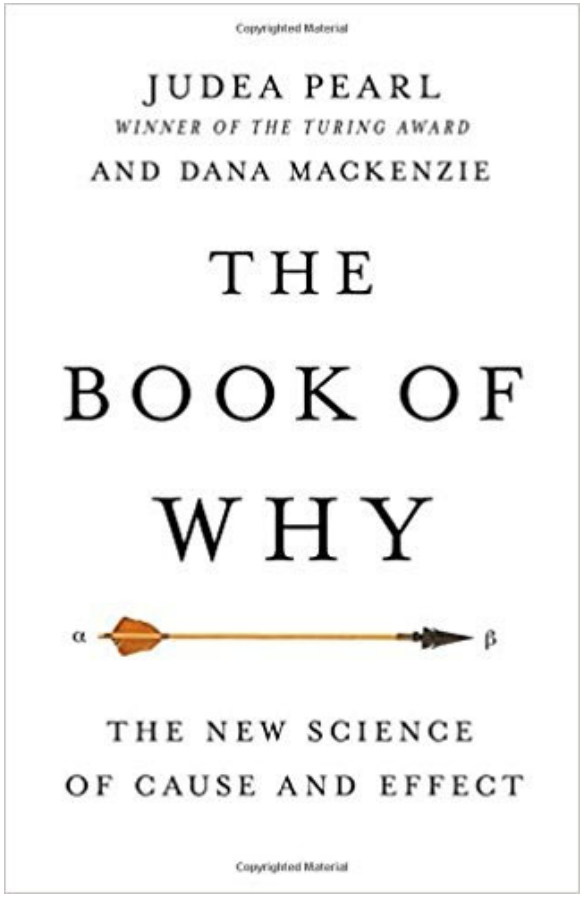
Causality Workshop 2018

The book of WHY

University of Zurich[UZH]



published in May 2018

current amazon bestseller #1
in the category “statistics”
(followed by Elements of Statistical Learning)

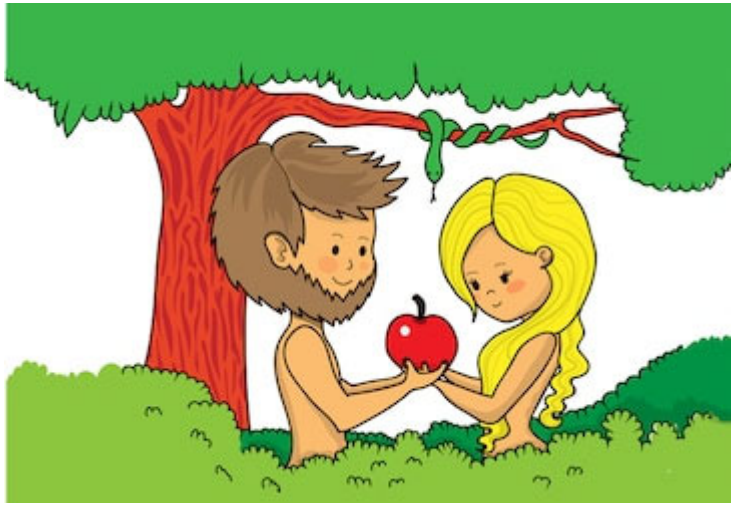
Pearl received the Turing Award 2011

Beate Sick

Topics of today

- Humans and scientists want/need to understand the “WHY”
- Correlation: birth of statistics – end of causal thinking?
- Regression to the mean
- Pearl’s ladder of causation
- Can our statistical and ML/DL models “only do curve fitting” ?
- Historic anecdotes in statistics and ML seen through a causal lens

Humans conscious rises the question of WHY?



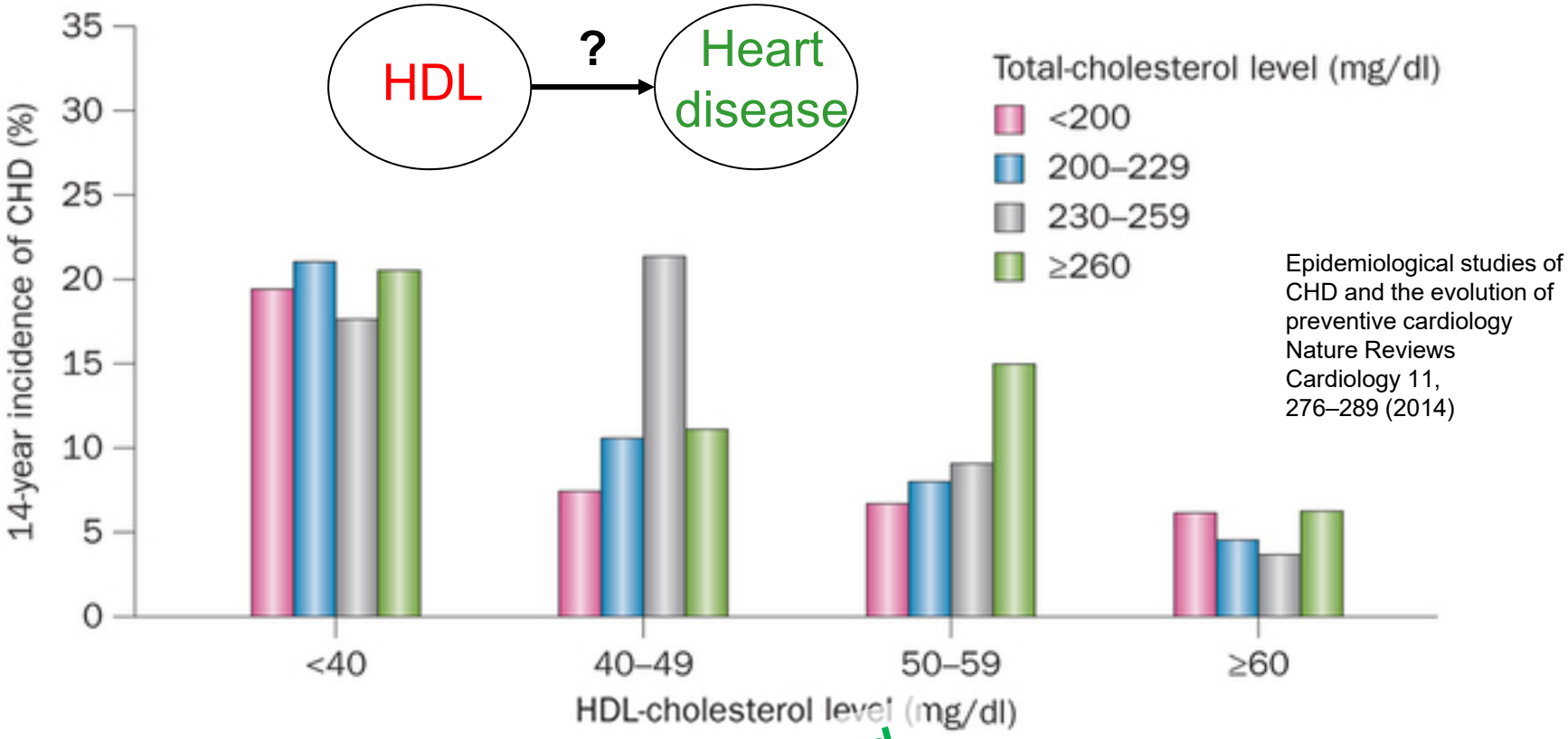
God asks for WHAT

“Have you eaten from the tree which I forbade you?”

Adam answers with WHY

“The woman you gave me for a companion, she gave me fruit from the tree and I ate.”

For intervention planning we need to understand the WHY



Epidemiological studies of CHD and the evolution of preventive cardiology Nature Reviews Cardiology 11, 276–289 (2014)

HDL gives a strong negative association with heart disease in cross-sectional studies and is the strongest predictor of future events in prospective studies.

Prediction worked

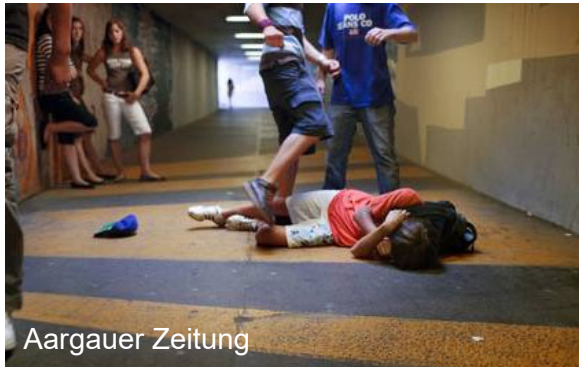
Roche tested the effect of drug "dalcetrapib" in phase III on 15'000 patients which proved to boost HDL ("good cholesterol") but failed to prevent heart diseases. Roche stopped the failed trial on May 2012 and immediately lost $5billion of its market capitalization.

Planned intervention failed

We need to understand causality to plan intervention

Do violent video games cause violence among young people?

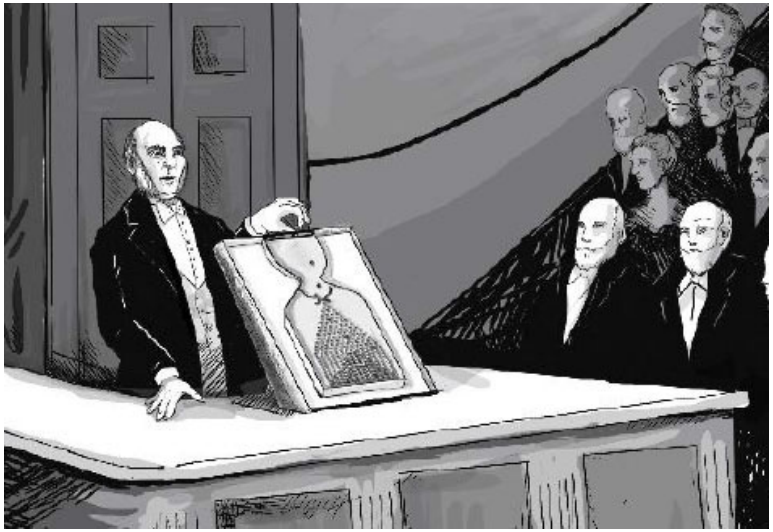
Then ban them!



Does unconditional basic income crank up economy?

Then launch it!

Galton on the search for causality



Galton in 1877 at the Friday Evening Discourse at the Royal Institution of Great Britain in London.

Francis Galton (first cousin of Charles Darwin) was interested to **explain** how traits like “intelligence” or “height” is passed from generation to generation.

Galton presented the “quincunx” (Galton nailboard) as causal model for the inheritance.

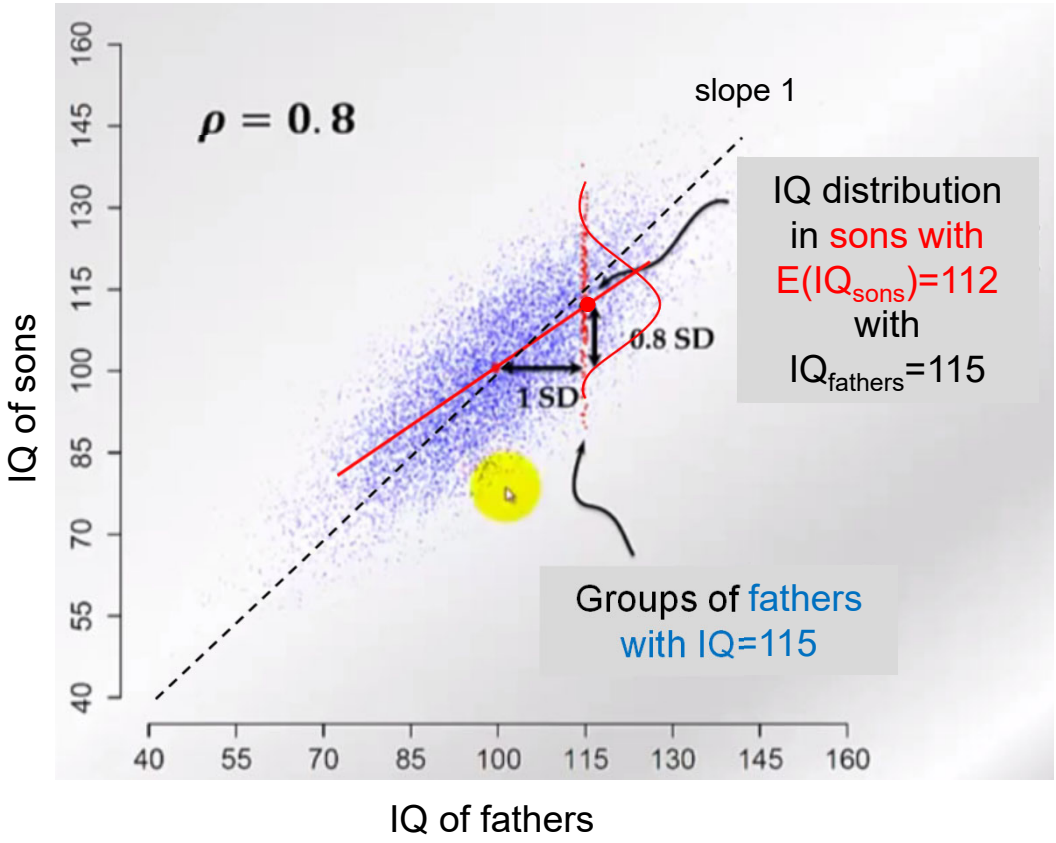
Balls “inherit” their position in the quincunx in the same way that humans inherit their stature or intelligence.

The stability of the observed spread of traits in a population over many generations contradicted the model and puzzled Galton for years.

Image credits: “The Book of Why”

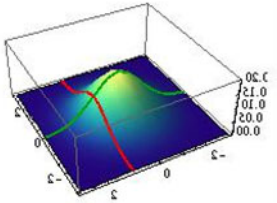
Galton's discovery of the regression line

Remark: Correlation of IQs of parents and children is only 0.42 https://en.wikipedia.org/wiki/Heritability_of_IQ



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$
$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \mathrm{cov}(X1,X2) \\ \mathrm{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$

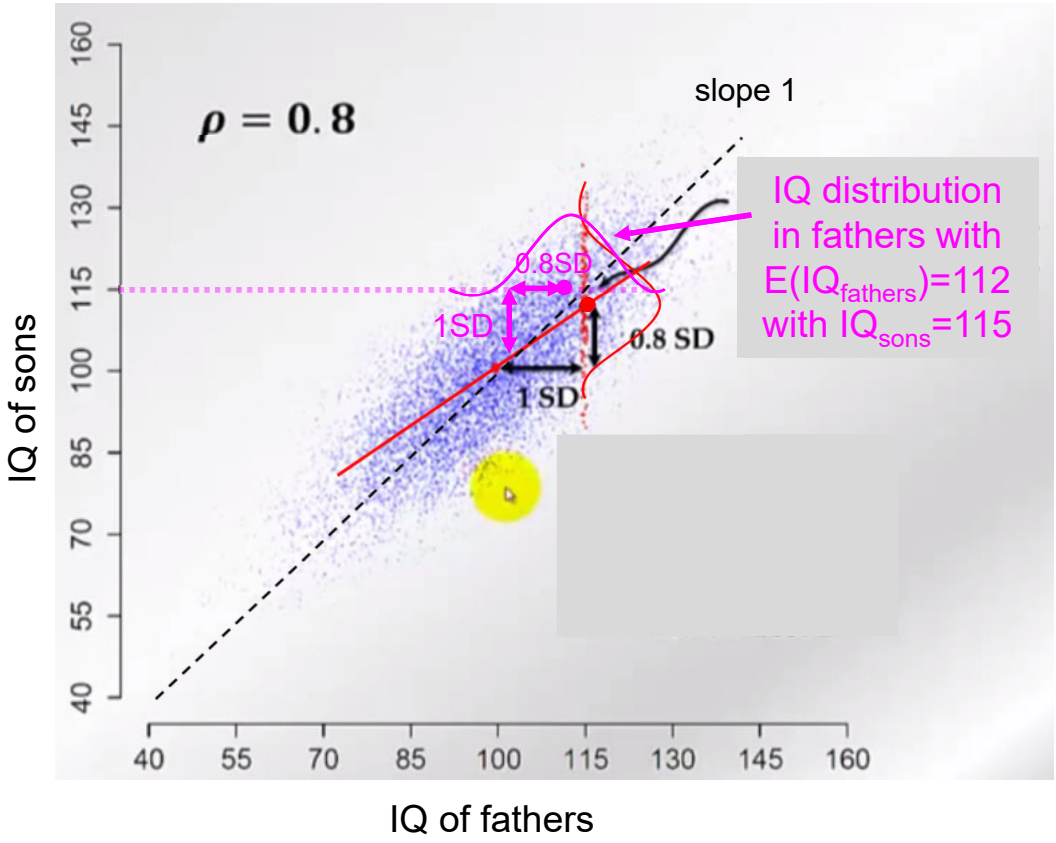


For each group of father with fixed IQ, the mean IQ of their sons is closer to the overall mean IQ (100) -> Galton aimed for a causal explanation.

All these predicted $E(IQ_{son})$ fall on a "regression line" with slope<1.

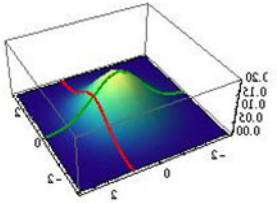
Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

Galton's discovery of the regression to the mean phenomena



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$
$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

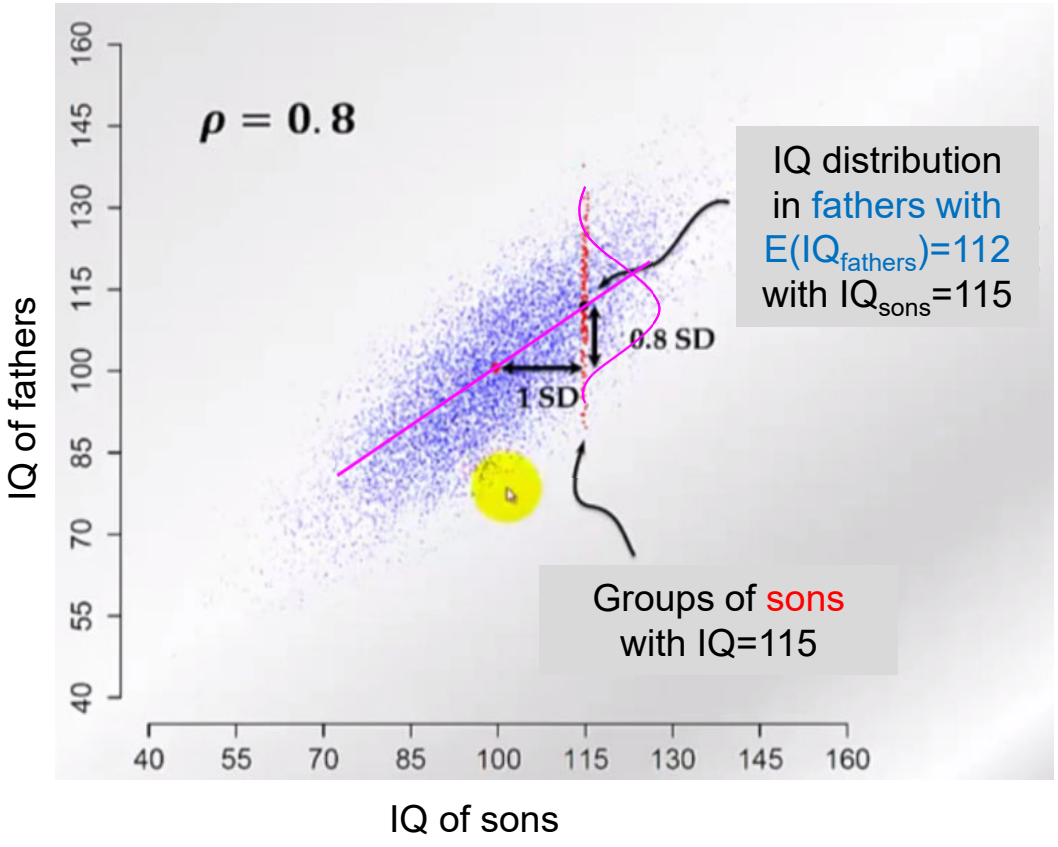
$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \text{cov}(X1,X2) \\ \text{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$



Also the mean of all fathers who have a son with IQ=115 is only 112.

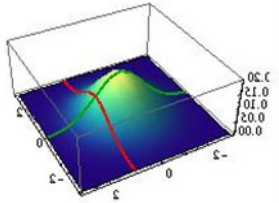
Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

Galton's discovery of the regression to the mean phenomena



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$
$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \text{cov}(X1,X2) \\ \text{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$

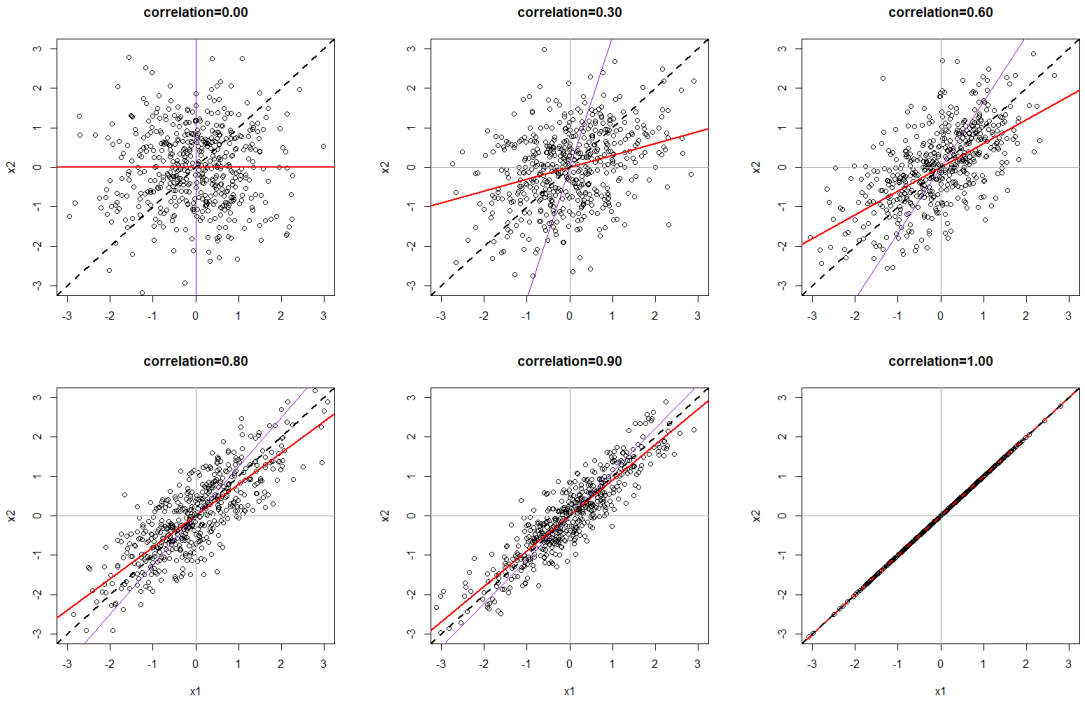


After switching the role of sons's IQ and father's IQ, we again see that E(IQ$_{fathers}$) fall on the regression line with the same slope <1.

There is no causality in this plot -> causal thinking seemed unreasonable.

Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

Pearson's mathematical definition of correlation unmasks "regression to the mean" as statistical phenomena



The correlation c of a bivariate Normal distributed pair of random variables are given by the slope of the regression line after standardization!

c quantifies strength of linear relationship and is only 1 in case of deterministic relationship.

After standardization of the RV:

$$X1 \sim N\left(\mu_1 = 0,\ \sigma_1^2 = 1^2\right)$$
$$X2 \sim N\left(\mu_2 = 0,\ \sigma_2^2 = 1^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{X1}^2 = 1 & c \\ c & \sigma_{X2}^2 = 1 \end{pmatrix}\right)$$

Regression line equation:

$$\hat{X}_2 = E\left(X_2 \mid X_1\right) = \beta_0 + \beta_1 \cdot X_1$$

$$\beta_1 = c \cdot \frac{\sigma_2}{\sigma_1} \overset{\text{stand.}}{=} c$$

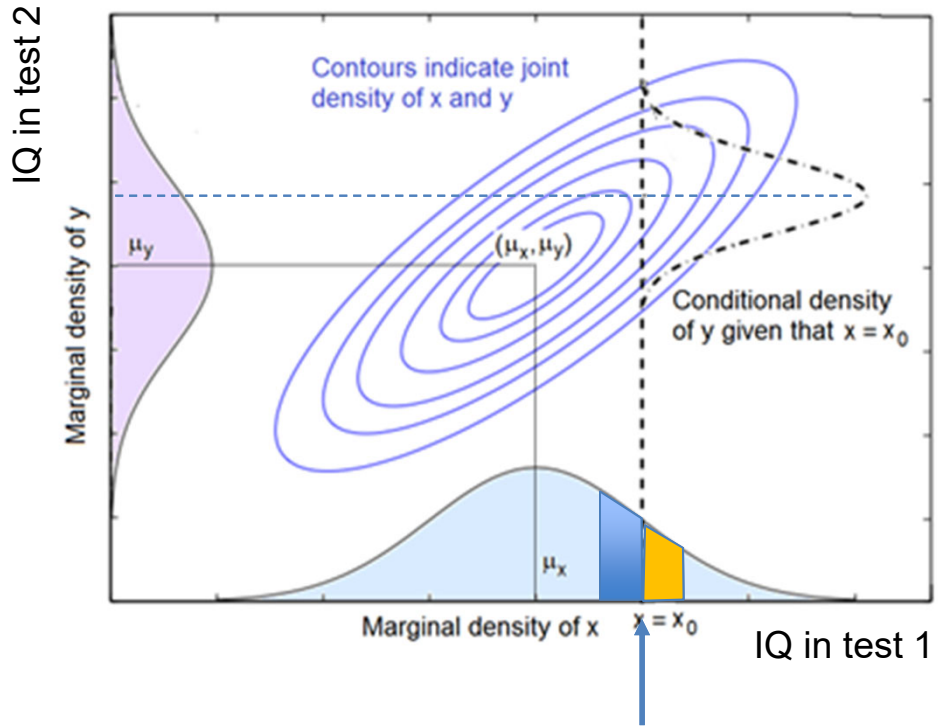
β_1 quantifies regression to the mean

$$c = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)\cdot(x_{i2} - \bar{x}_2)}{\mathrm{sd}(x_1)\cdot\mathrm{sd}(x_2)}$$

Intuitive explanation of "regression to the mean"

IQ test result (at both time points) = true IQ + luck or bad luck

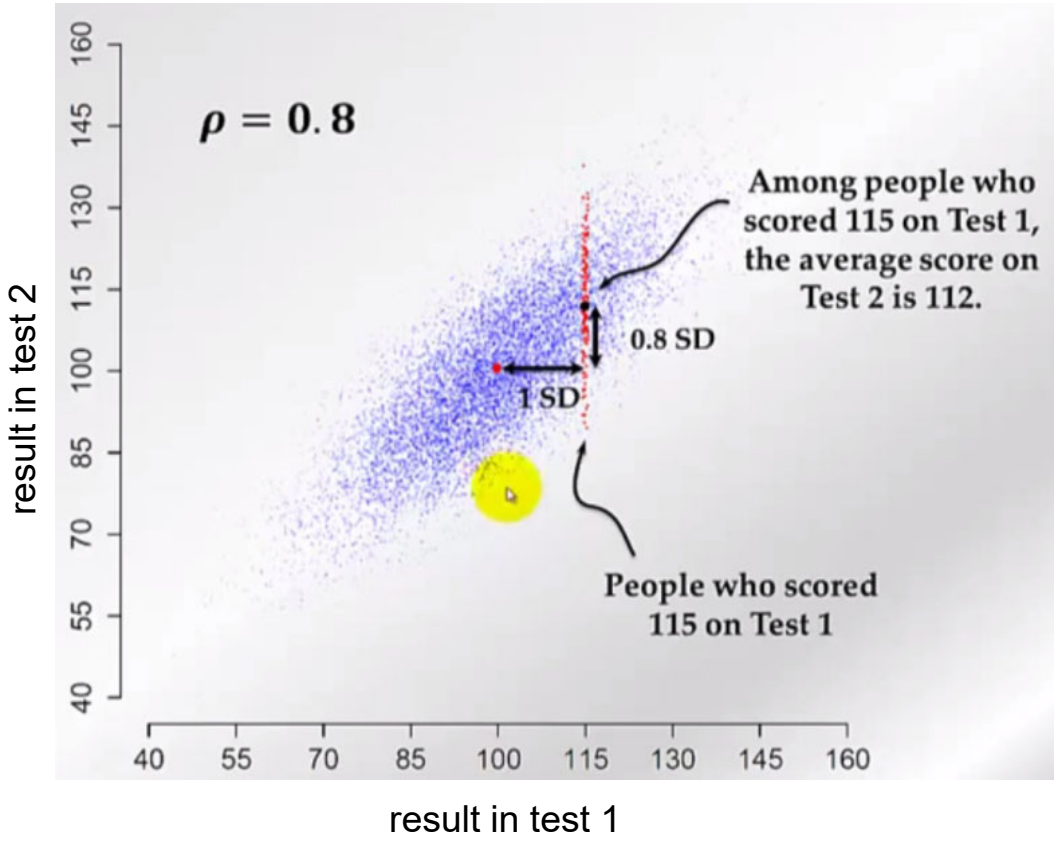
Not reproducible in second test



To get this test result, a person might

- have truly this high IQ (this are some people)
- have a lower true IQ (**many people** have a lower IQ) but **had luck**
- have a higher true IQ (**fewer people** have a higher IQ) but **had bad luck**

Regression to the mean occurs in all test-retest situations

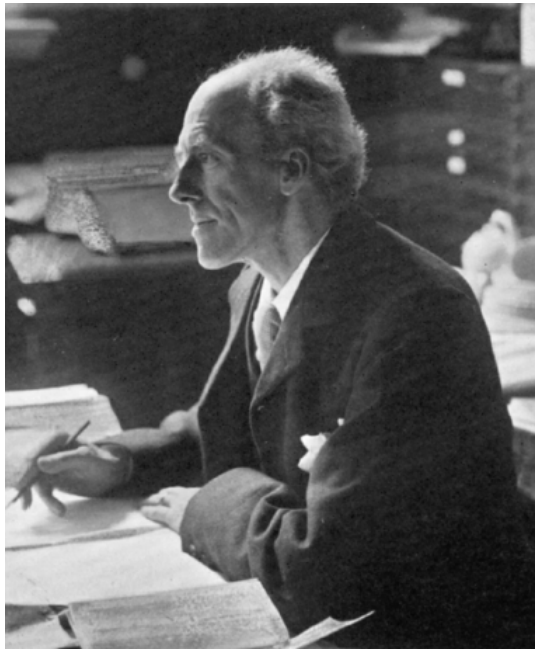


Retesting a extreme group (w/o intervention in between) in a second test leads in average to a results that are closer to the overall-mean -> to assess experimentally the effect of an intervention also a control group is needed!

With the correlation statistics was born and abandoned causality as "unscientific"

"the ultimate scientific statement of description of the relation between two things can always be thrown back upon… a contingency table [or correlation]."

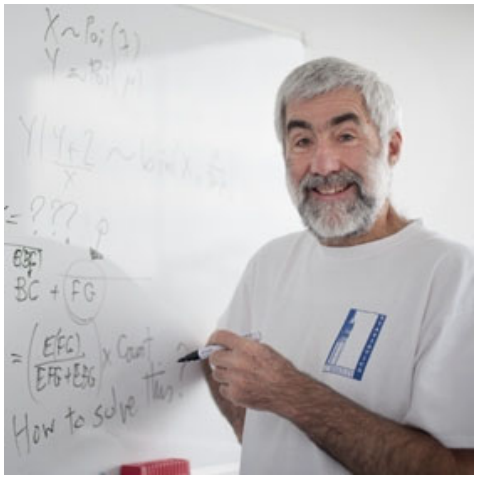
Karl Pearson (1895-1936), The Grammar of Science



Pearl's rephrasing of Pearson's statment:
"**data is all there is to science**".

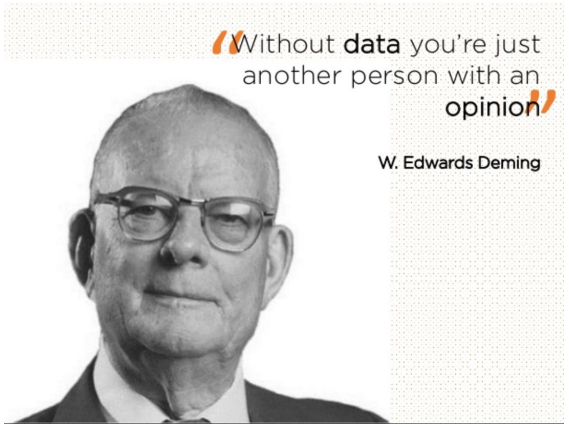
However, Pearson himself wrote several papers about "spurious correlation" vs "organic correlation" (meaning organic=causal?) and started the culture of "think: 'caused by', but say: 'associated with' "…

Quotes of data scientists



"Considerations of causality should be treated as they have always been in statistics: preferably not at all."

Terry Speed, president of the Biometric Society 1994



In God we trust. All others must bring data.

W. Edwards Deming (1900-1993), statistician and father of the total quality management

See also http://bigdata-madesimple.com/30-tweetable-quotes-data-science/

Pearl's statements

Observing [and statistics and AI] entails detection of regularities

We developed [AI] tools that enabled machines to reason with uncertainty [Bayesian networks].. then I left the field of AI

Mathematics has not developed the asymmetric language required to capture our understanding that if *X* causes *Y* .

As much as I look into what's being done with deep learning, I see they're all stuck there on the level of associations. Curve fitting.

The book of Why
https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/

Probabilistic versus causal reasoning

Traditional statistics, machine learning, Bayesian networks

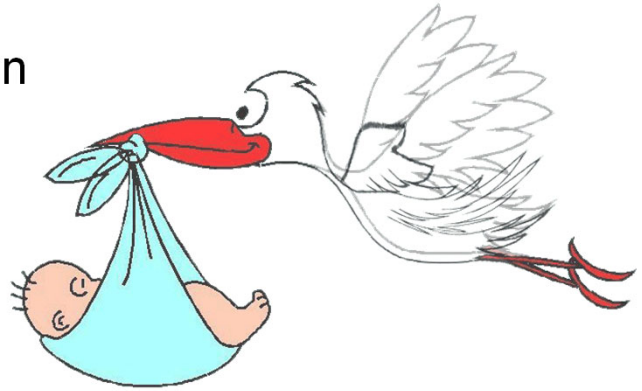
- About associations (are stork population and human birth number per year are associated?)
- The dream is a models for the joined distribution of the data
- Conditional distribution are modeled by regression or classification

  (if we observe a certain number of storks, what is our best estimate of human birth rate?)

Causal models

- About causation (do storks do affect human birth rate?)
- The dream is a models for the data generation
- Predict results of interventions

  (if we change the number of storks, what will happen with the human birth rate?)



Pearl's ladder of causality

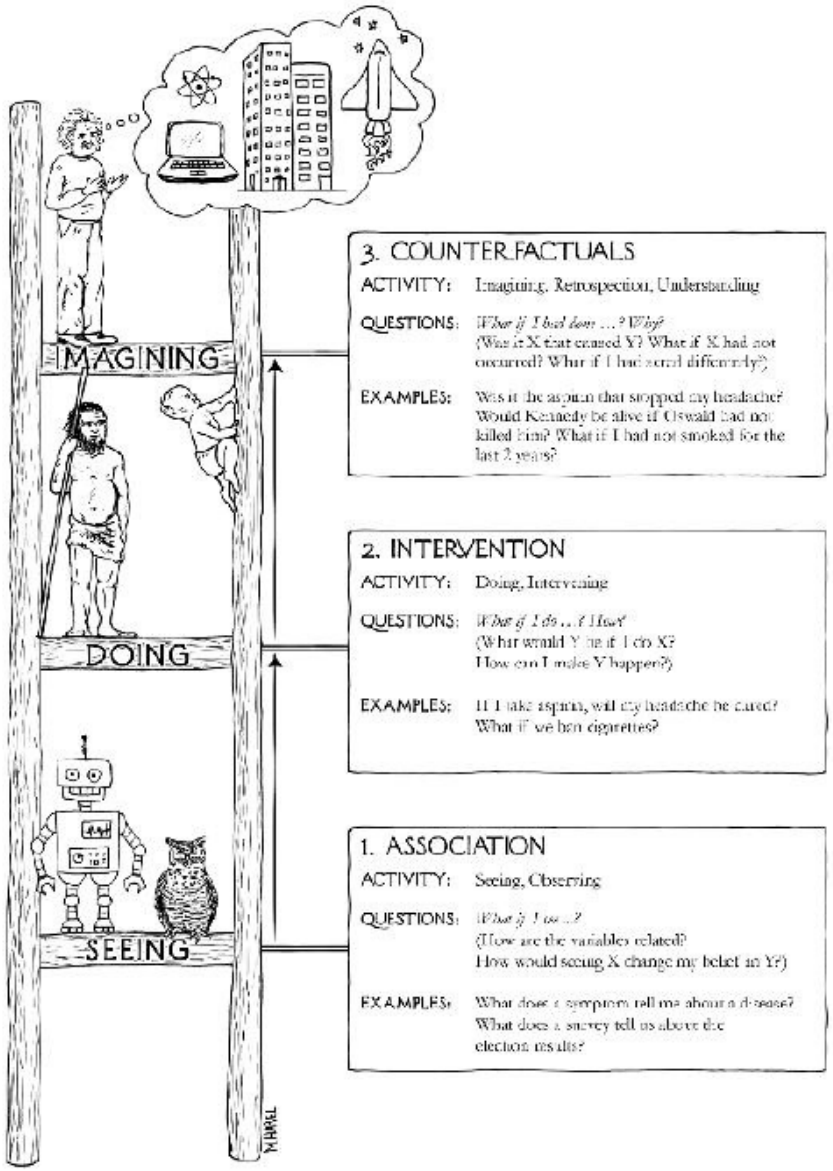
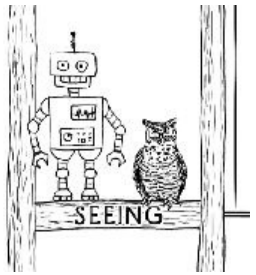


FIGURE 1.2. The Ladder of Causation, with representative organisms at each level. Most animals, as well as present-day learning machines, are on the first

Image credits: "The Book of Why"

Regression Model
What can they tell us?

On the first rung of the ladder
Pure regression can only model associations



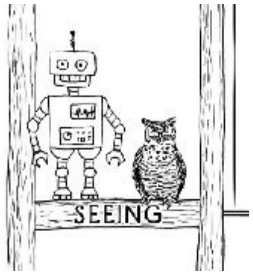
$$(\mathrm{Y}_i|\mathbf{X}_i) \sim \mathrm{N}(\mathbf{X}^t{}_i\beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1},\ \sigma^2)$$

Usual interpretation:
The coefficient β_k gives the change of the outcome y, given the explanatory variable x_k is increased by one unit and all other variables are held constant.

But: How can we increase just one predictor and hold the others constant?

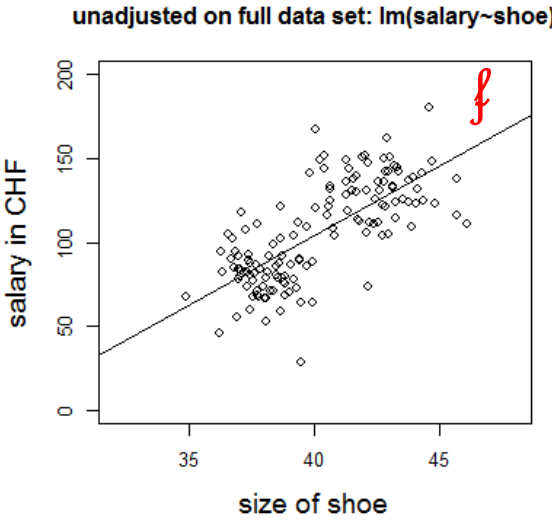
Interpretation for biostatistical problems:
β_k is the amount the outcome would change had the participant shown a covariate x_k increased by one unit – all other do not change ;-)

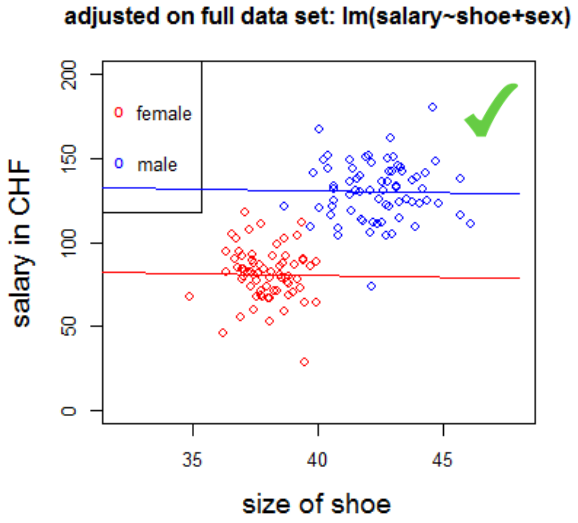
How we work with rung-1 regression or ML models

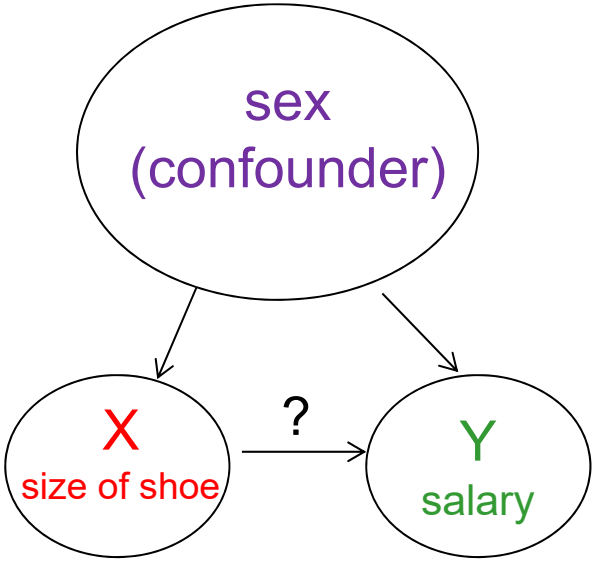


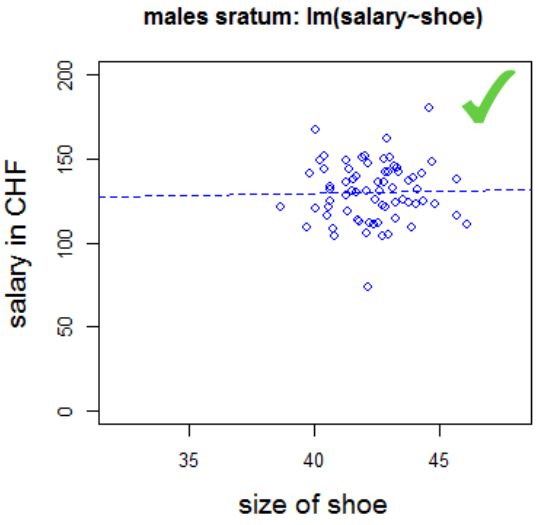
xkcd.com

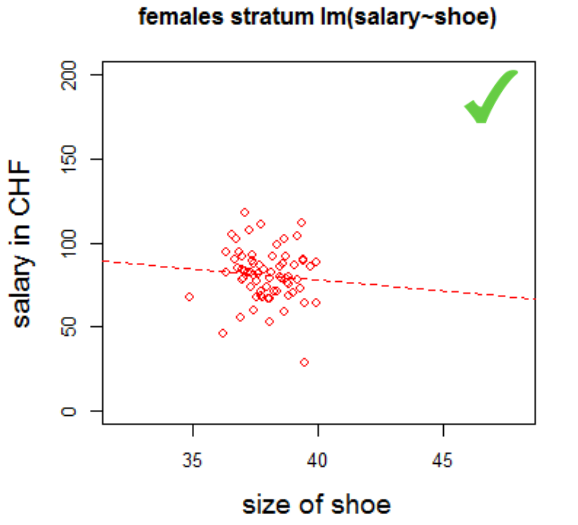
Confounder can introduce spurious association: Adjustment methods can work well (toy example)











Stratified analysis -> different models for male and females

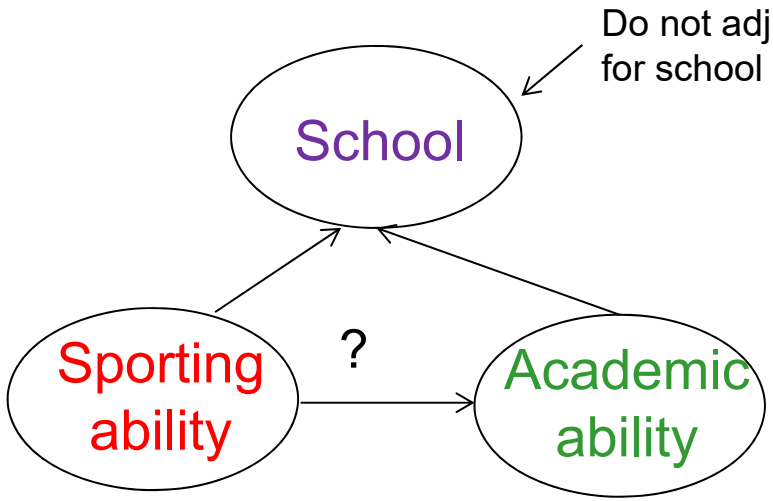
Looking into adjustment methods
Never adjust for a common effect: a toy example

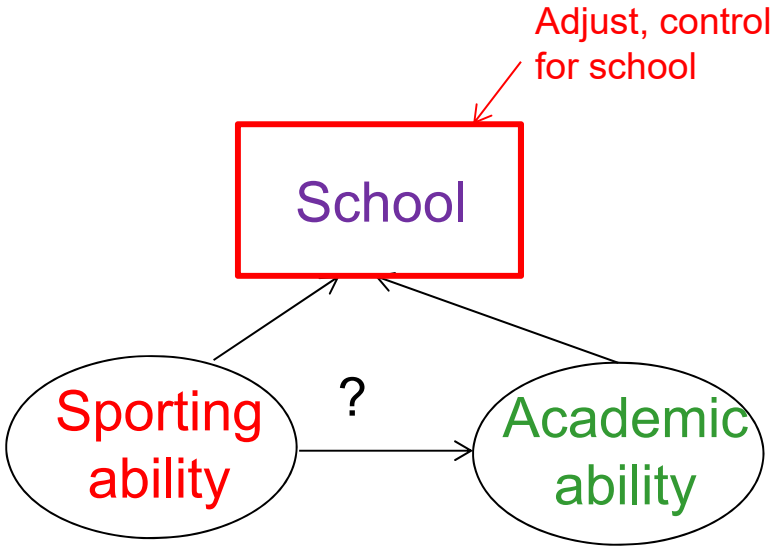
A school accepts pupils who are either good at sport, or good academically, or both

-> School acceptance is associated with sporting and academic abilities

Suppose: in Population sport and academic skills are independent

What happens if we “adjust” for the factor “accepted in school”?

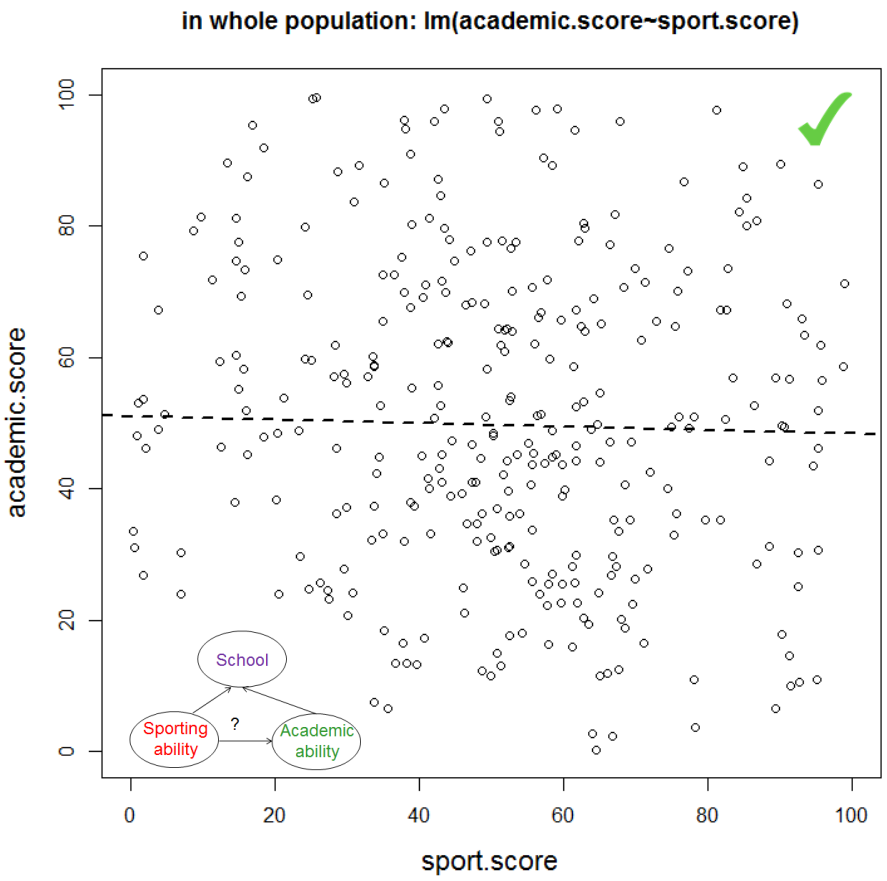


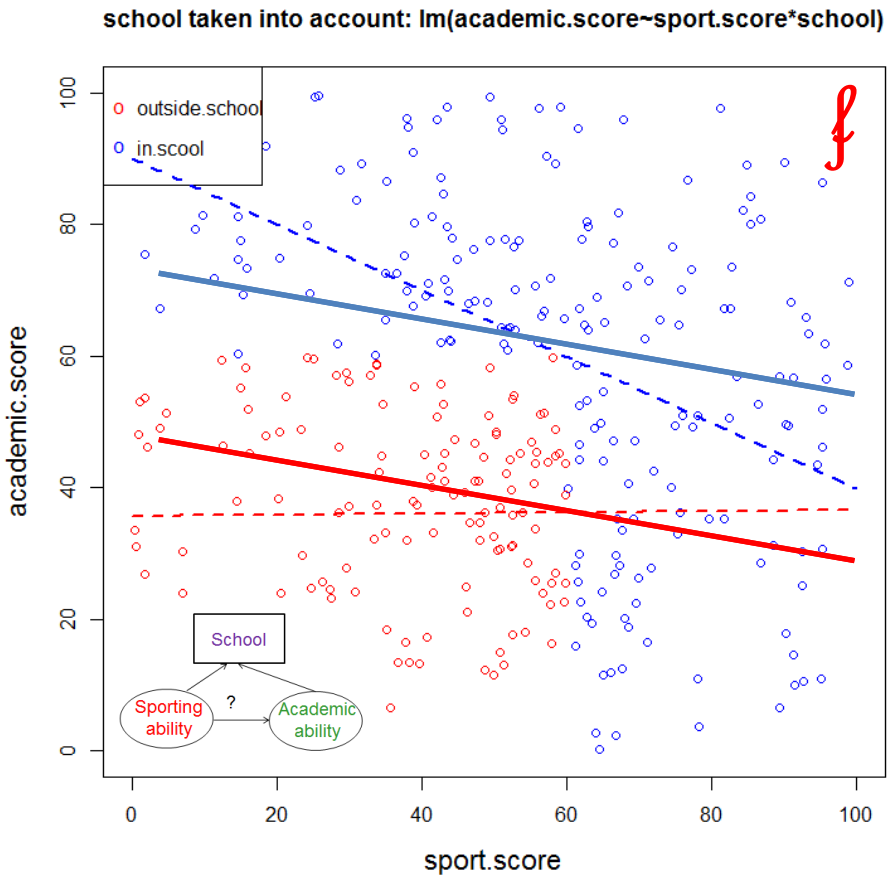



```
m1=lm(academic ~ sport, data=dat)
```

```
m2=lm(academic ~ sport + school, data=dat)
```

Adjusting for associated variables can work out badly
A toy example: effect of sport on academic abilities






```
m1=lm(academic ~ sport, data=dat)
```

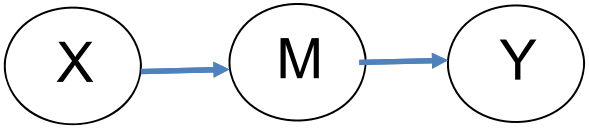
```
m2=lm(academic ~ sport + school, data=dat)
```

In the population there is no association between sport score and academic score, but by controlling for the school-variable we created a spurious association.

Looking into adjustment methods
Never adjust for mediator

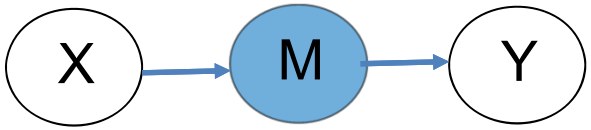
Toy example: a treatment X makes an enzyme M working which reduces pain Y

Not adjusting for M




```
Y ~ X
```

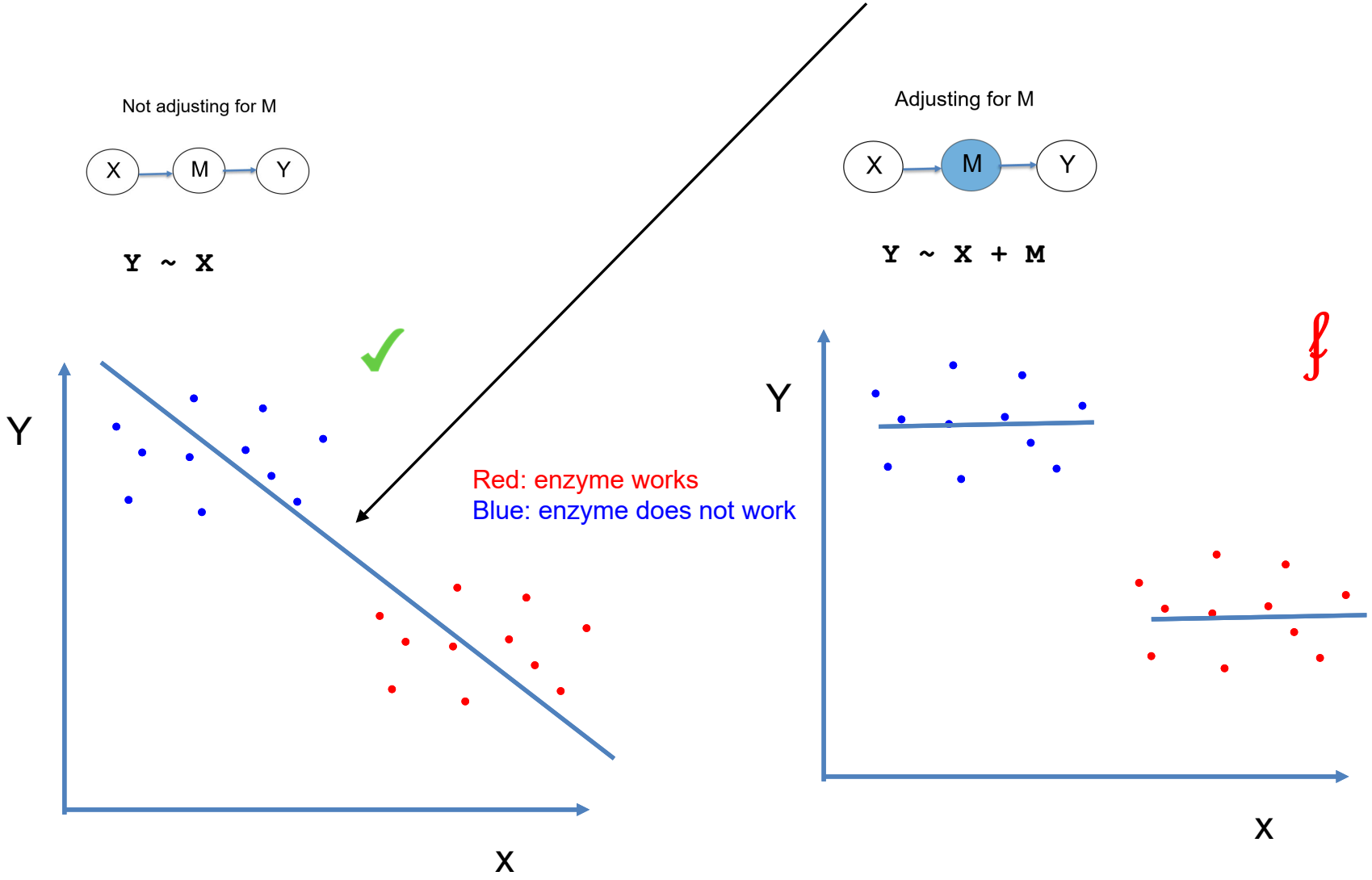
Adjusting for M




```
Y ~ X + M
```

Do not adjust for a mediator

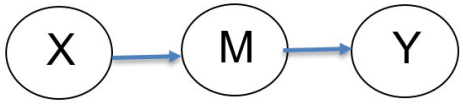
Truth: because of treatment the enzyme starts working and pain Y is reduced!



A third variable is associated with X and Y
To adjust or not to adjust – that is the question

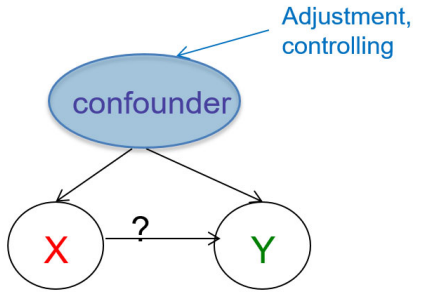
Do not adjusting for a mediator!

```
y ~ x
```



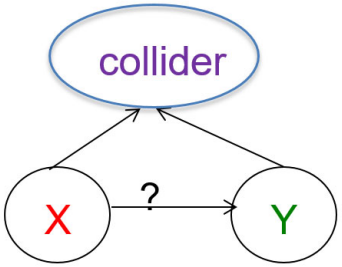
Adjust for a confounder!

```
y ~ x + c
```



Do not adjust for a collider!

```
y ~ x
```



Can and should we try to learn about causal relationships?

If yes – what and how can we learn?

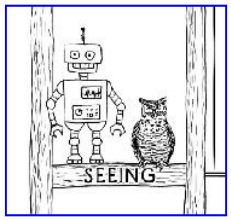
Ascending the second rung: go from "seeing" to "doing"

Research question:

What is the distribution of the blood pressure if people do not drink coffee?

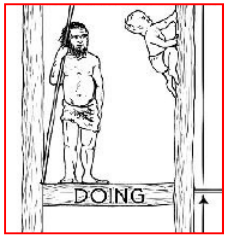
Conditioning / Seeing:
Filter - restrict on non-coffee drinker

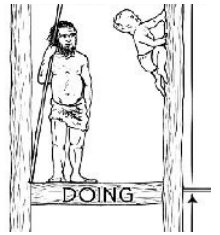
$$P(\mathrm{BP} \mid \mathrm{coffee} = 0)$$

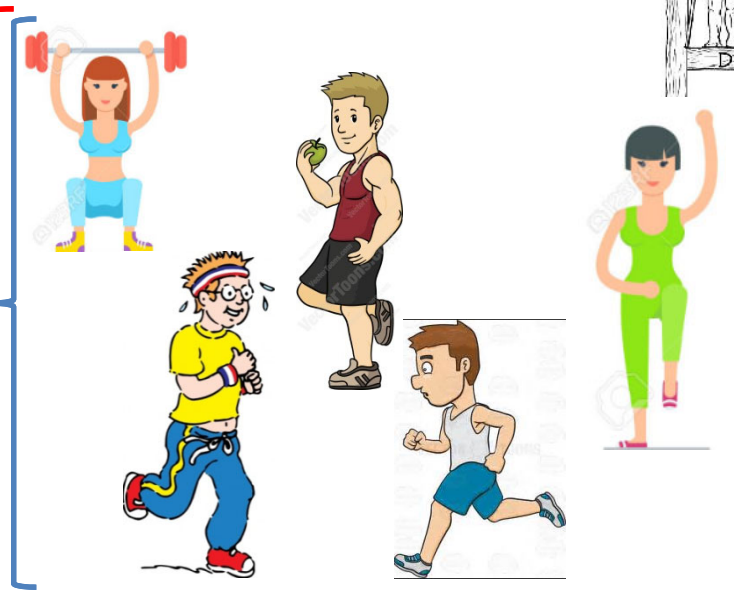


"Do"-Operator:
Full population, after intervention that prohibits coffee consume

$$P(\mathrm{BP} \mid \mathrm{do}(\mathrm{coffee} = 0))$$



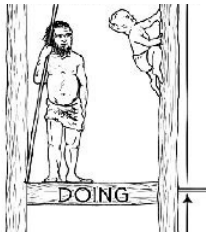


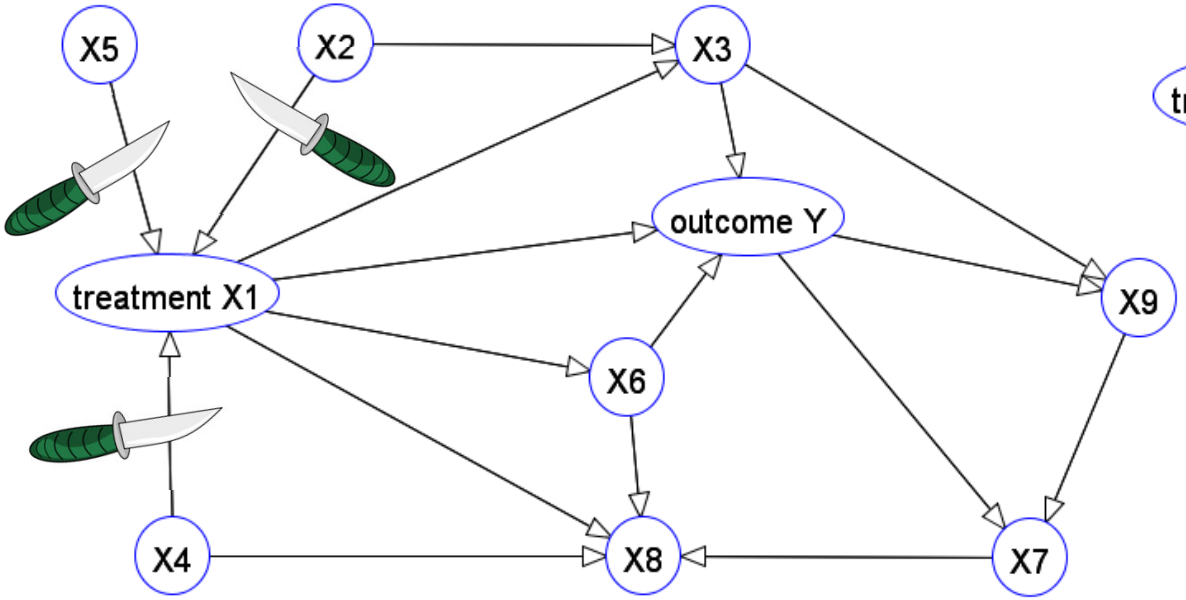


non-coffee drinker by choice

coffee drinker by choice

On the second "doing" rung of the ladder
Assessing the intervention effect by a RCT





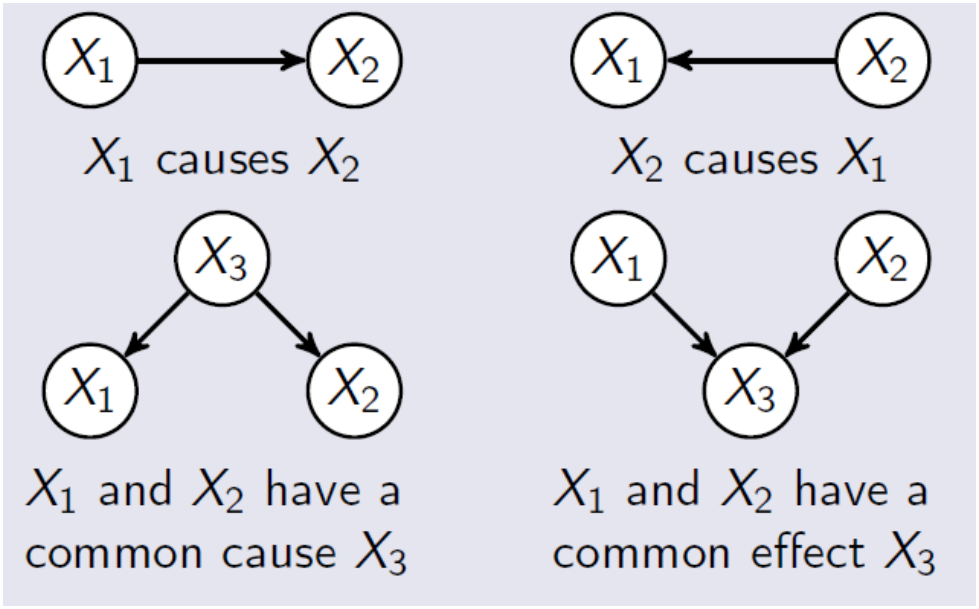
RCT through the lens of a causal graphical model

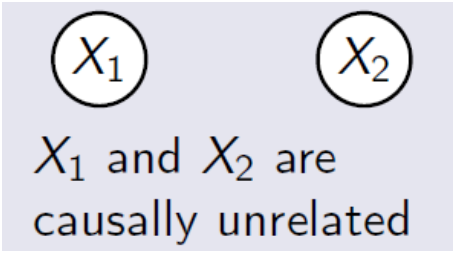
Since the treatment is assigned randomly to both treatment groups are exchangeable. Hence observed differences of the outcome in both groups is due to the treatment.

-> Model after collecting data from a RT: $outcome \sim treatment$

From Bayesian networks to causal graphical models

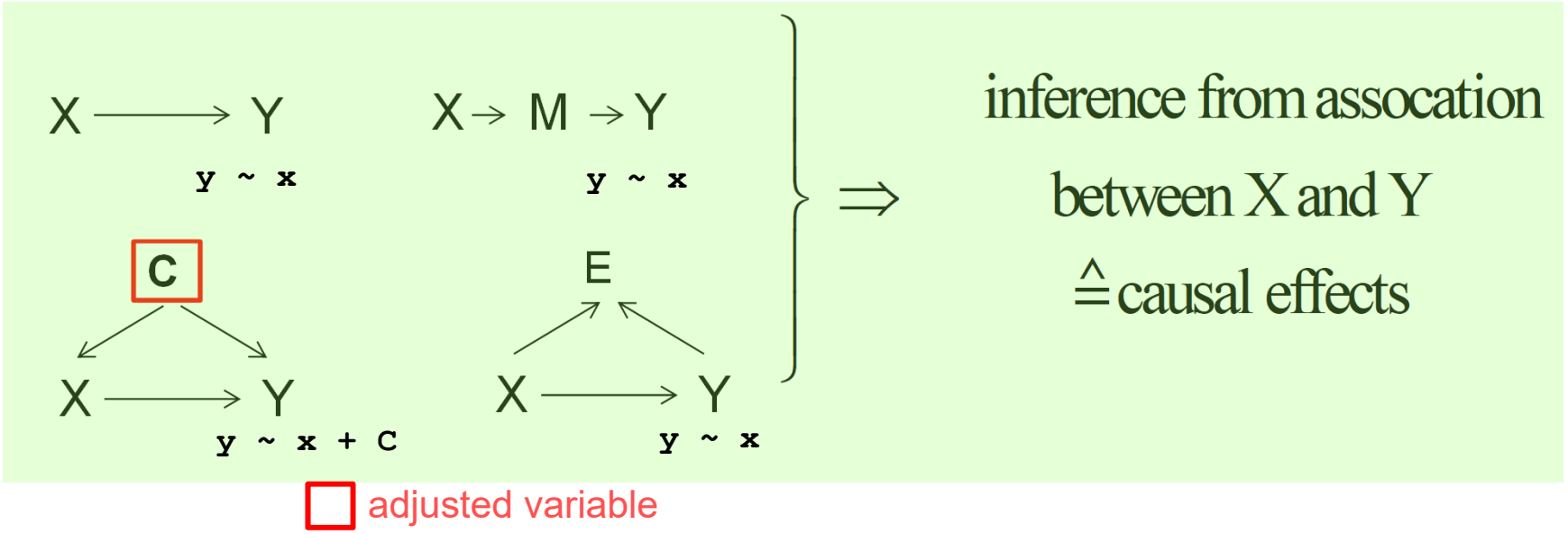
A causal BN is a DAG about causal relationships where again nodes are variables, but a directed edge represents a potential causal effect.

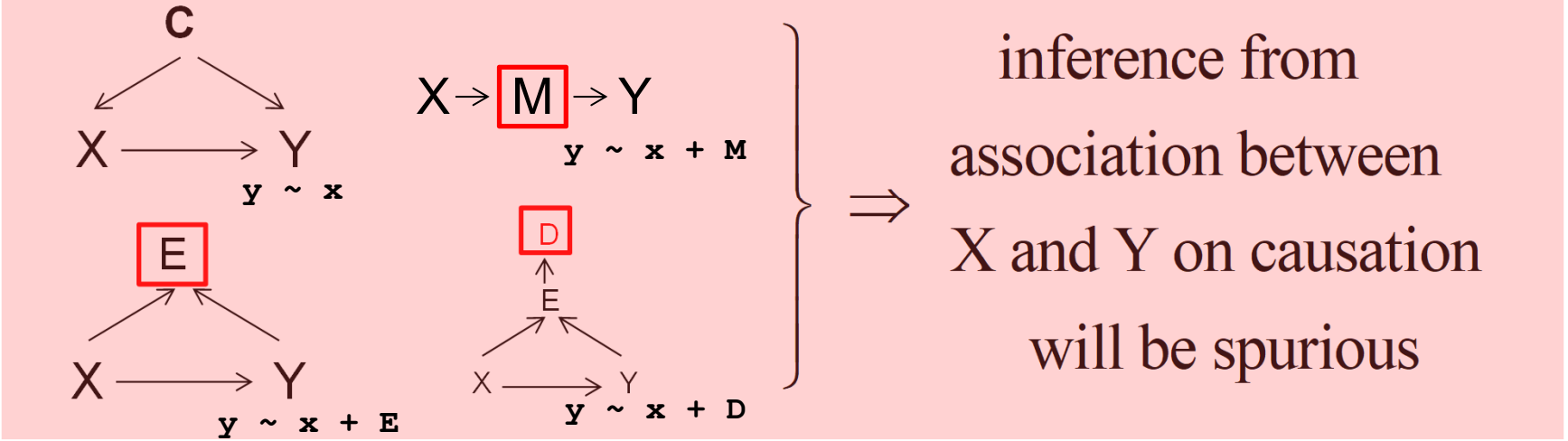




Causal effects can only be transported along the direction of arrows!

Building blocks of causal model



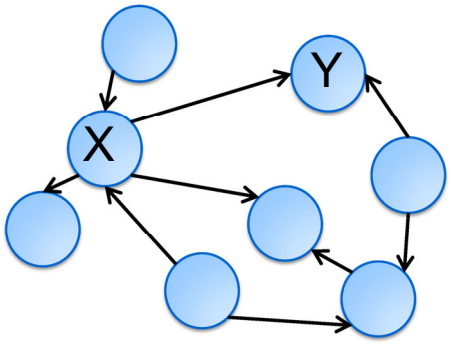


Can we do causal/intervential inference from observational data?

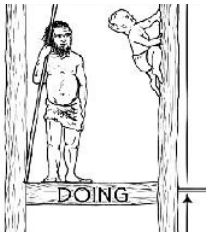
The very short answer: No!

Principle be Cartwright (1989): **No causes in – no causes out!**

observational data



Backdoor criterion or frontdoor criterion or 3 Rules of do-Calculus

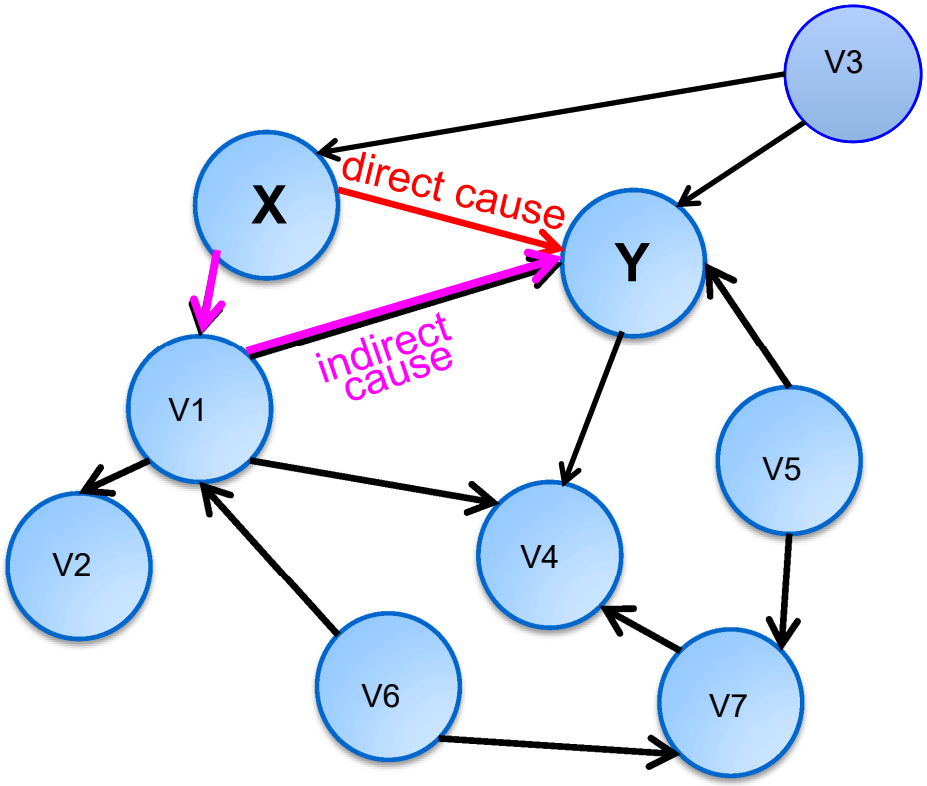


$$P'(y \mid do(X = x_0))$$

$$=$$

Expression without do (!!) which only uses information from observed JPD P

What is a causal path?

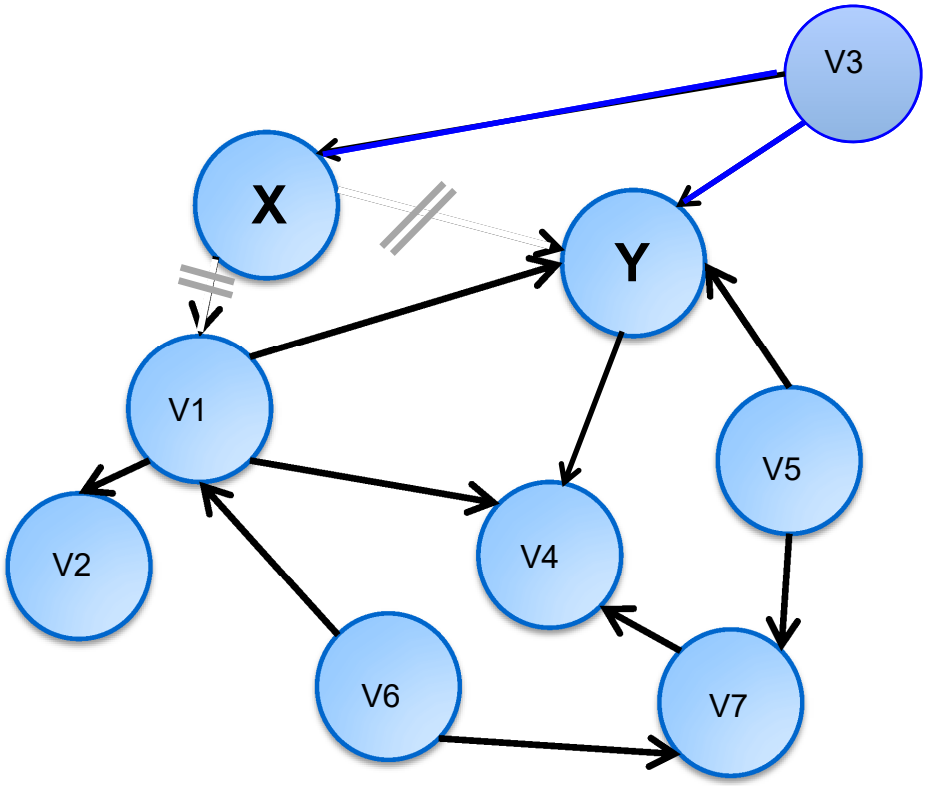


In a causal path from X to Y is a directed path from X to Y

→ if follow the arrows in a causal path we get from X to Y.

→ We have 2 causal paths transporting direct and indirect causes

What is a backdoor path?



First we ignore (delete) all arrows starting from X

A backdoor path from X to Y starts with an arrow pointing into X: $X \leftarrow \cdots Y$

→ Any path (regardless of the arrow directions) that still connects X and Y.

Pearl's backdoor criterion for causal graphical models

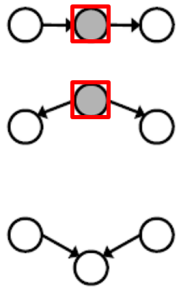
Goal: Close all backdoor paths connecting X and Y.

- Determine a set S of "de-confounder" variable closing all backdoor paths by controlling for these variables.
- S must not contain any descendent of X.
  (This ensures that we do not block a causal path from X to Y)

- S can be used for covariate adjustment to estimate the total causal effect of X on Y

A path is blocked if **1 single** triple-segment is blocked!

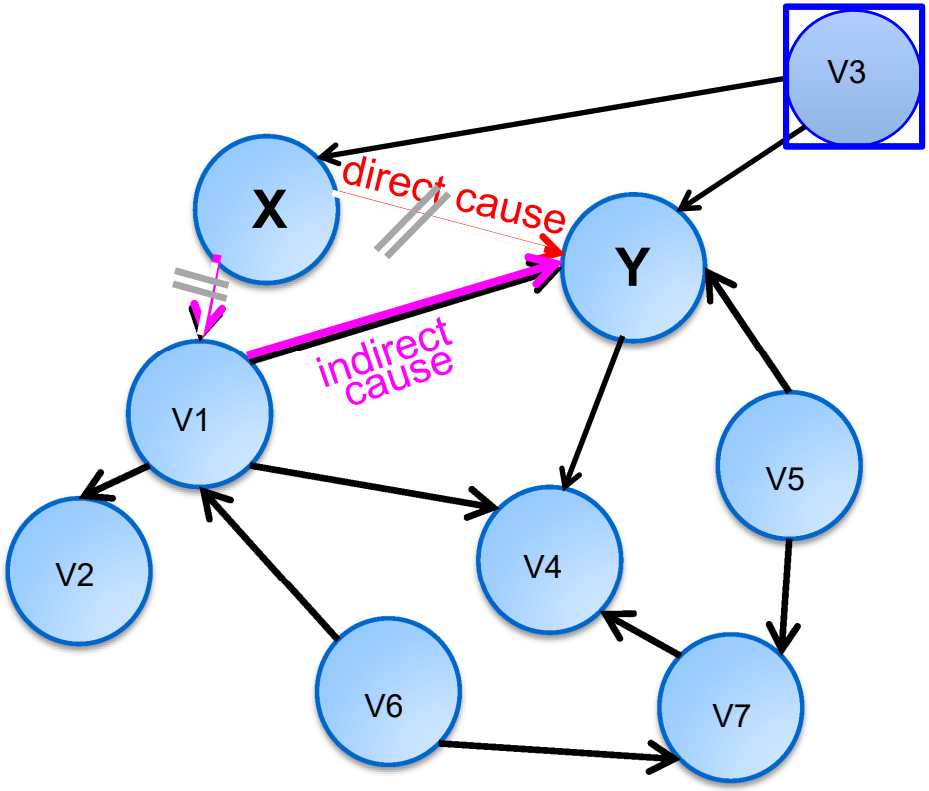
blocked triple



controlled variable
observed variable

Control for a variable
=
using the variable in the regression model

Has X an causal influence on Y?
Are all backdoor paths closed?



To close all backdoor paths we must adjust for this confounder.

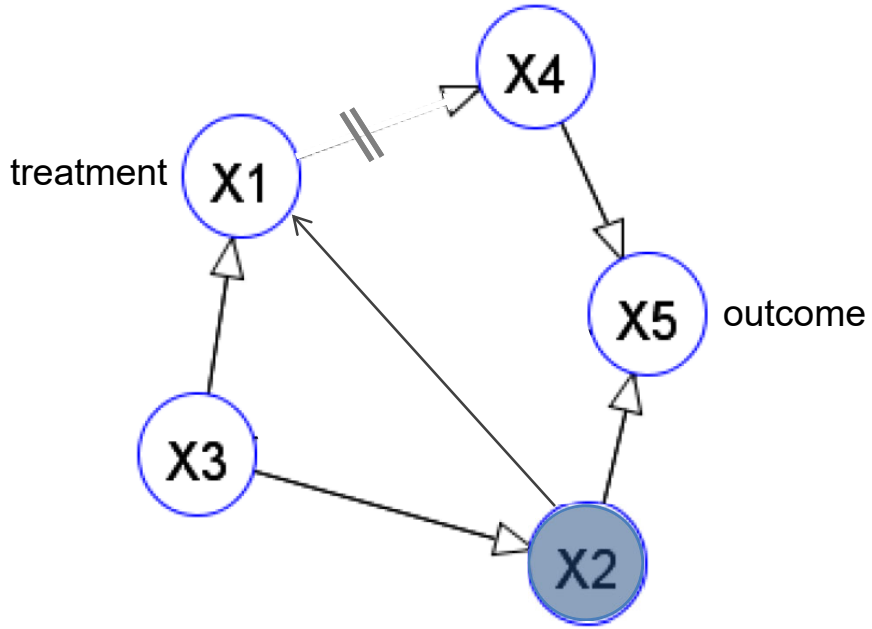
```
y ~ x + v3
```

Use the back door criterion to check a model

RQ: Has X_1 ("treatment") a causal effect on X_5 ("outcome")?

Is the proposed model appropriate to Interpret the estimated β_1 causally?

$$X_5 \sim X_1 + X_2$$



Are all back door paths (BDP) closed?

Yes, since all BDP go through the confounder X_2 and we control for X_2 by using it as covariable and thereby closing the BDP.

→ The estimated β_1 can be interpreted causally, given the graphical model is correct.

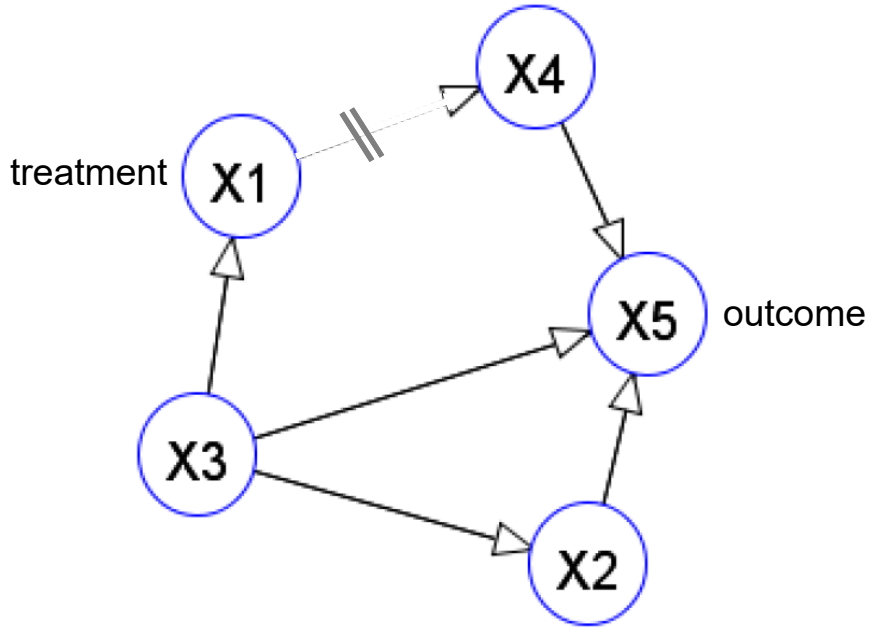
DIYS time ☺

Use the back door criterion to check a model

RQ: Has X_1 ("treatment") a causal effect on X_5 ("outcome")?

Is the proposed model appropriate to Interpret the estimated β_1 causally?

$$X_5 \sim X_1 \qquad \not\perp$$



Are all back door paths (BDP) closed?

No, since the BDP X1-X3-X5 goes through an uncontrolled confounder X_3 and is therefor open.

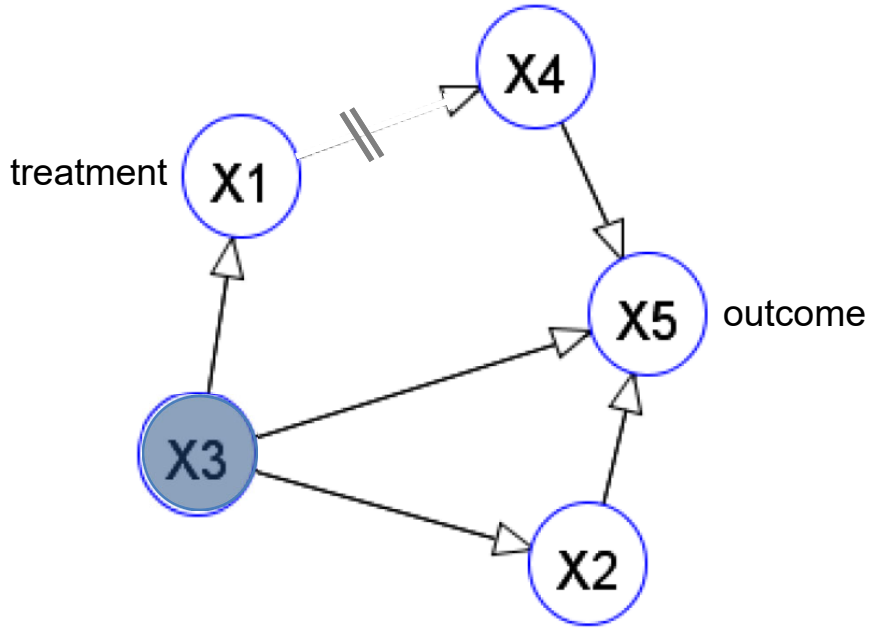
→ The estimated β_1 must not be interpreted causally, given the graphical model is correct.

Use the back door criterion to check a model

RQ: Has X_1 ("treatment") a causal effect on X_5 ("outcome")?

Is the proposed model appropriate to Interpret the estimated β_1 causally?

$$X_5 \sim X_1 + X_3$$ ✓



Are all back door paths (BDP) closed?

Yes, since all BDP go through the confounder X_3 and we control for X_3 by using it as covariable and thereby closing the BDP.

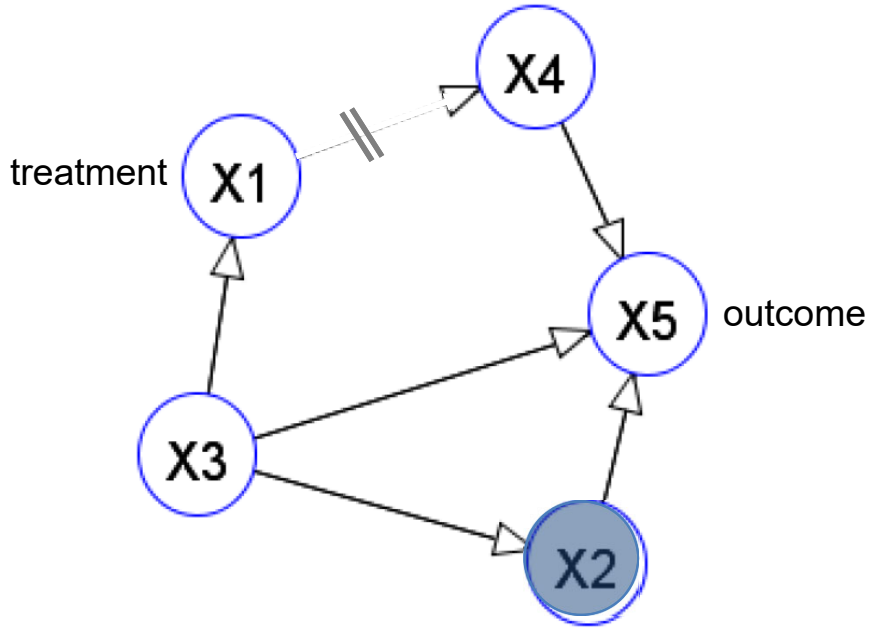
→ The estimated β_1 can be interpreted causally, given the graphical model is correct.

Use the back door criterion to check a model

RQ: Has X_1 ("treatment") a causal effect on X_5 ("outcome")?

Is the proposed model appropriate to Interpret the estimated β_1 causally?

$$X_5 \sim X_1 + X_2$$



Are all back door paths (BDP) closed?

No, since the BDP X_1-X_3-X_5 goes through an uncontrolled confounder and is therefor open.

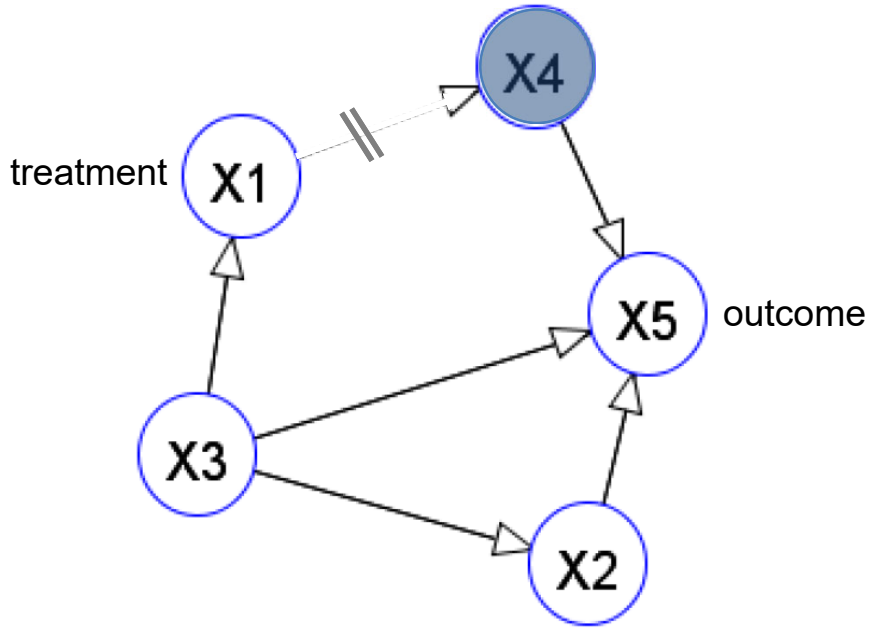
→ The estimated β_1 must not be interpreted causally, given the graphical model is correct.

Use the back door criterion to check a model

RQ: Has X_1 ("treatment") a causal effect on X_5 ("outcome")?

Is the proposed model appropriate to
Interpret the estimated β_1 causally?

$$X_5 \sim X_1 + X_4$$ ∉

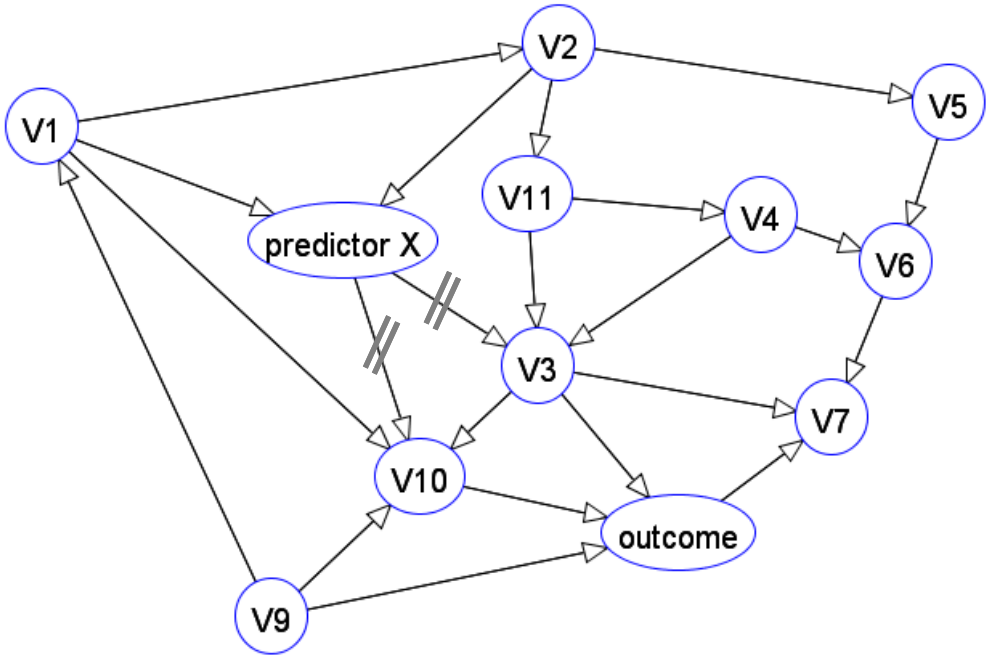


Are all back door paths (BDP) closed?

X_4 is a descendent of X_1
(mediator on causal path)
You must not use X_4 as covariable!!!

Use backdoor criterion to do regression properly for causal inference

What is the intervention effect of the predictor X on the outcome?

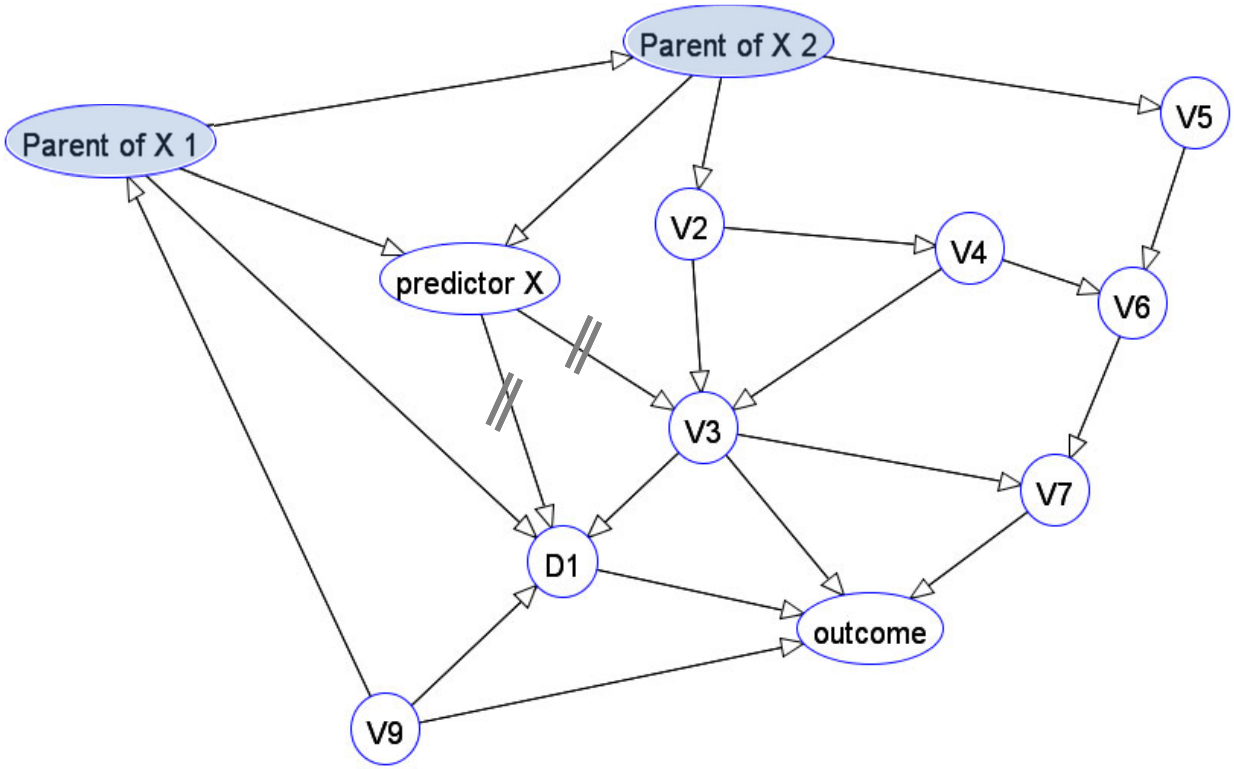


Regression can be used to asses the causal effect of the predictor X if we adjust with a **set S_B** of covariates V_i (e.g. parents of X) which would be sufficient to **close all backdoor paths** from intervention X to the outcome Y (several valid S_B might exist):

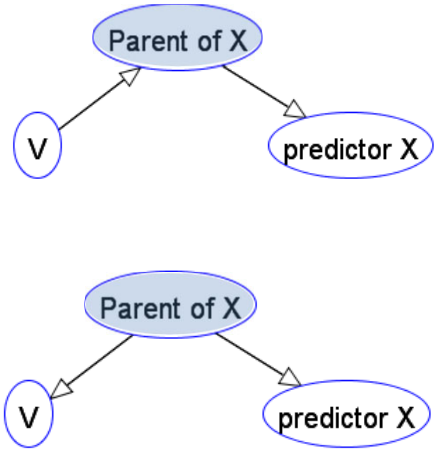
$$\text{outcome} \sim \text{predictor} + \sum_{V_i \in S_B} V_i$$

Special case of the backdoor criterion: intervention parents

All backdoor paths are closed if we control for the parents of the intervention variable X!



A controlled parent blocks the backdoor path either as controlled mediator or controlled confounder.



$$\text{outcome} \sim \text{predictor} + \sum \text{parents(predictor)}$$

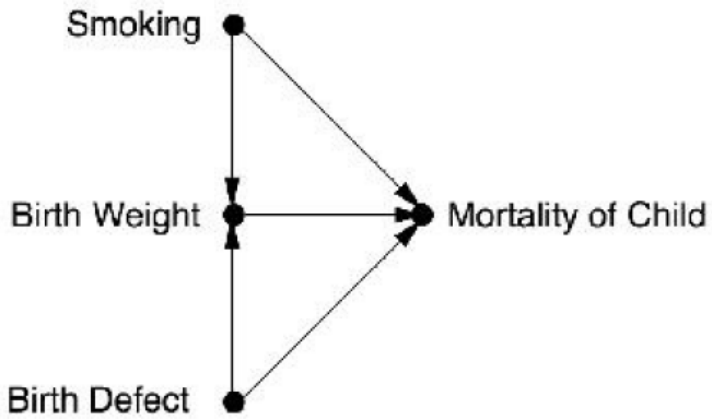
Historic anecdotes of of (non-) causal thinking

Are smoking mothers for underweighted newborns beneficial?

Since 1960 data on newborns showed consistently that low-birth-weight babies of smoking mothers had a better survival rate than those of nonsmokers.

This paradox was discussed for 40 years!

An article by Tyler VanderWeele in the 2014 issue of the *International Journal of Epidemiology* nails the explanation perfectly and contains a causal diagram:



Association is due to a collider bias caused by conditioning on low birth weight.

Image credits: “The Book of Why”

Any questions