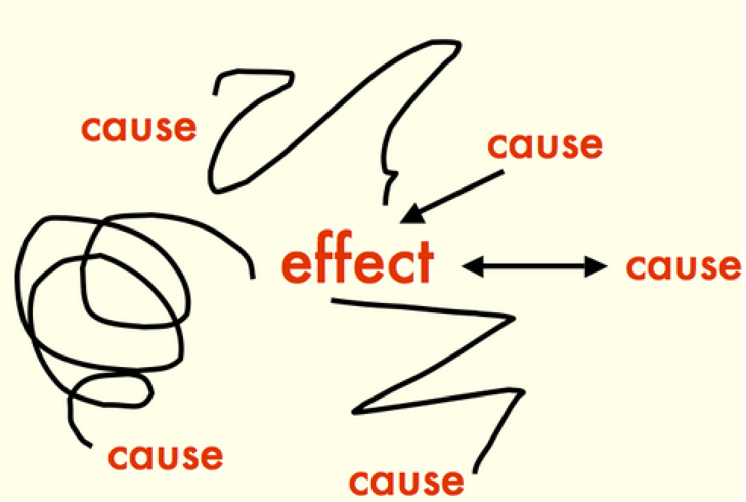
Causal inference with graphical models – in small and big data

Outline



Association is not causation

How adjustment can help or harm

Counterfactuals

- individual-level causal effect
- average causal effect

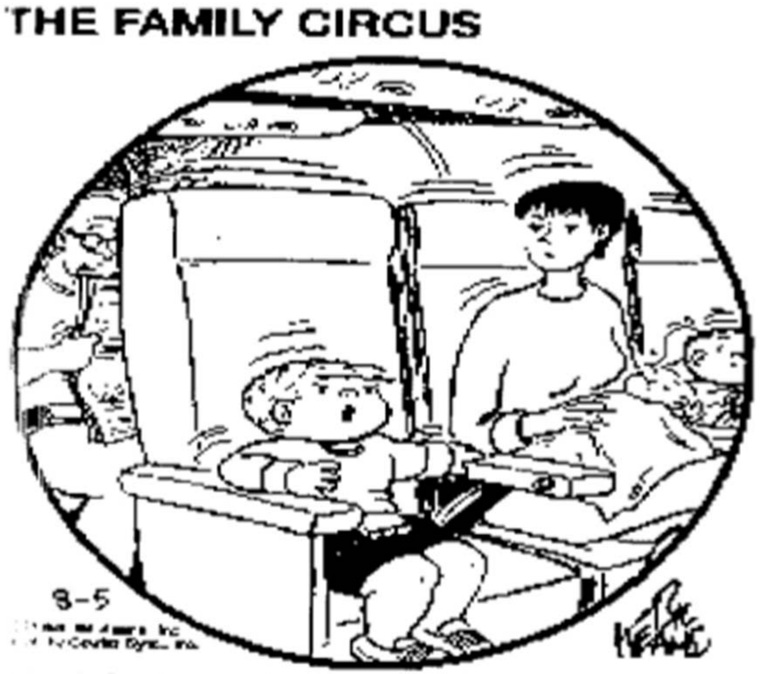
Causal graphs

- Graph structure, joint distribution, conditional independencies
- how to estimate a causal effect without bias: back-door criterion
- how to predict effect of interventions: do-Calculus

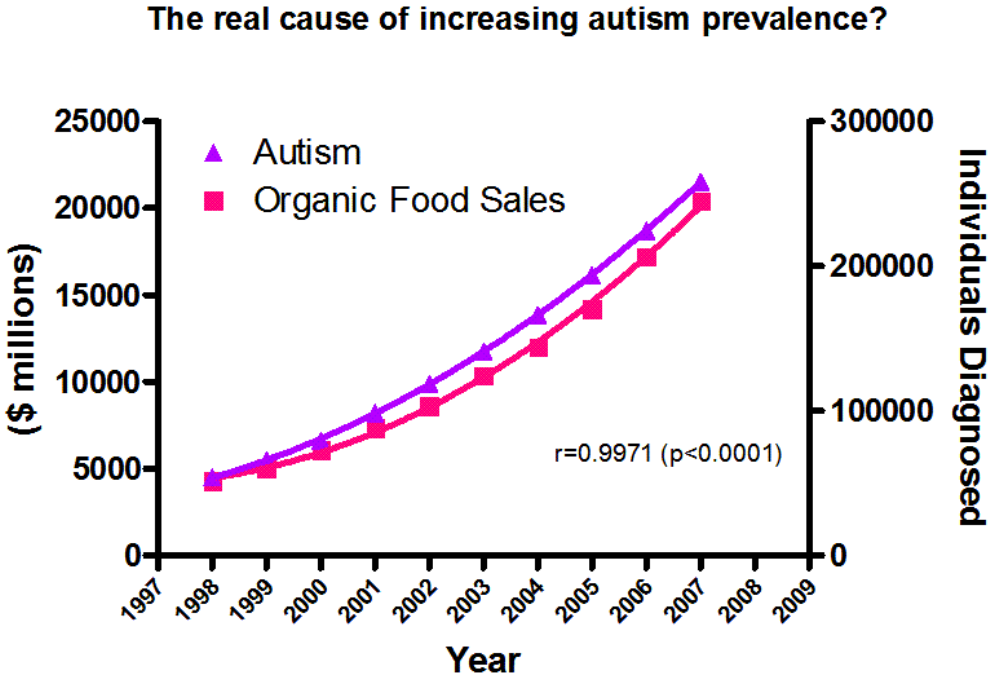
R: - fitting causal graphs
- estimating bounds for causal effects

Outlook, summary, and discussion

Association does not imply causation



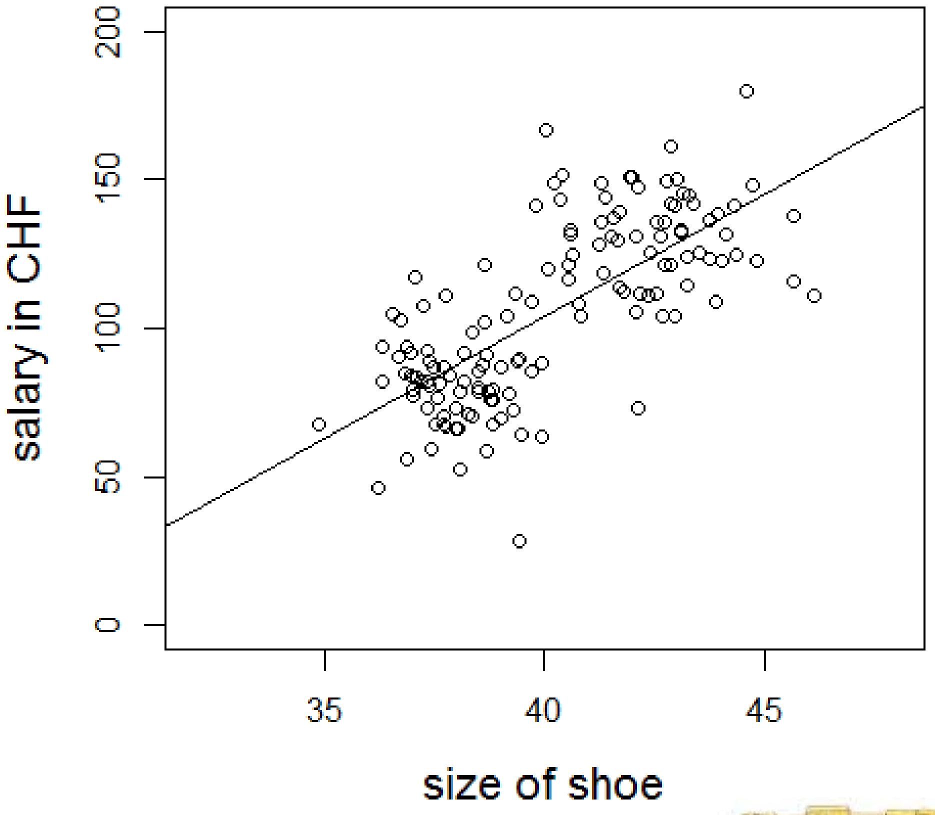
"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

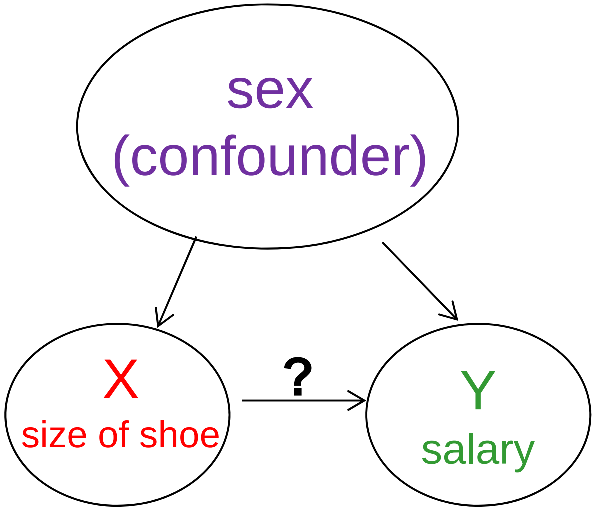


Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Spec
Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Spec
Education Under Part B of the Individuals
with Disabilities Education Act

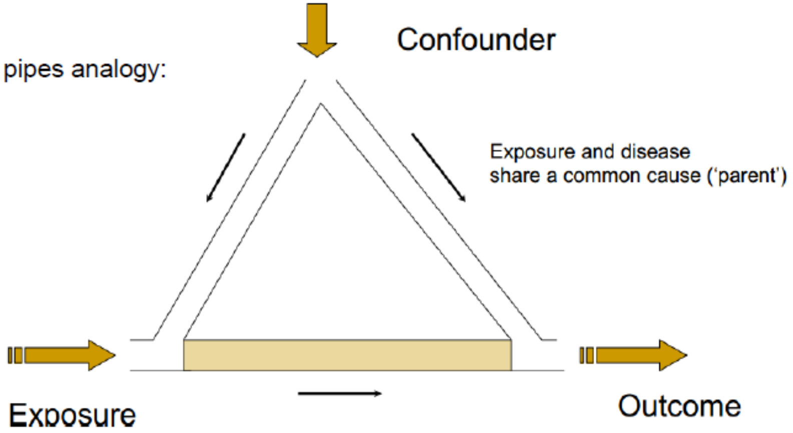
Taken from: http://stats.stackexchange.com/questions/36/examples-for-teaching-correlation-does-not-mean-causation

A toy example: shoe size and salary

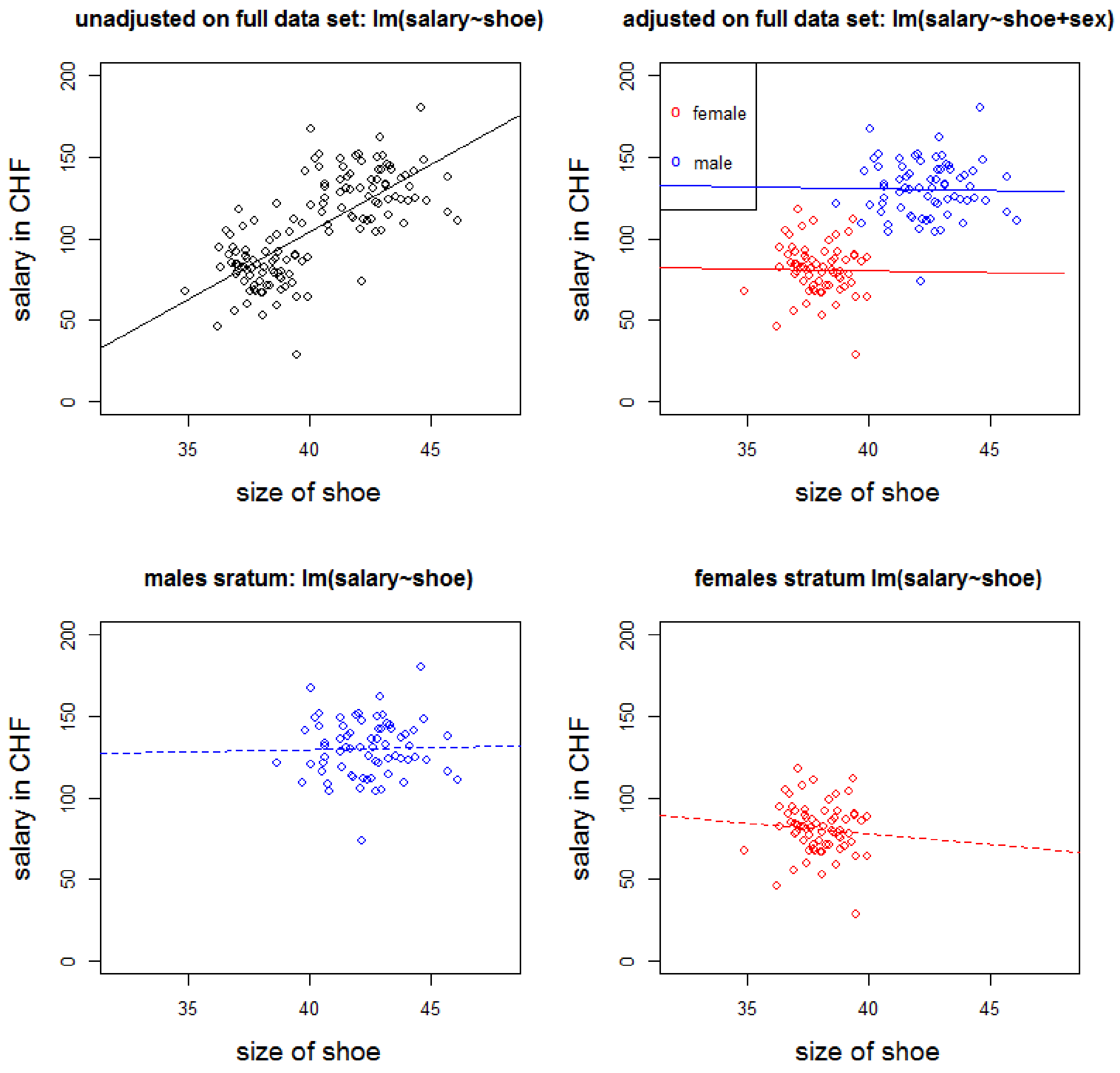




Water pipes analogy:



Adjustment methods can work well
A toy example: effect of shoe size on salary



Adjusting via multiple regression without interaction leads only similar results as a stratified analysis if the interaction is not significant -> parallel regression lines are «feature of model»

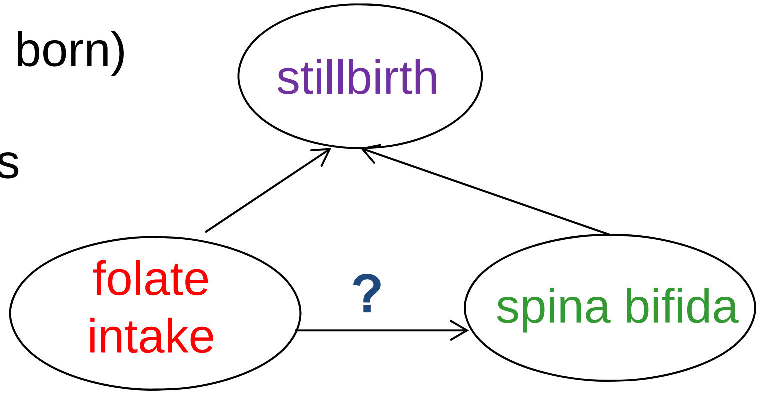
Stratified analysis -> different models for male and females are possible, but here not necessary.

Looking into adjustment methods
A real example: effect of folate intake on spina bifida

RQ: Is the lack of folate a cause for spina bifida?

Folate intake is associated with stillbirth (child dead born)

Spina bifida (open back) is associated with stillbirths



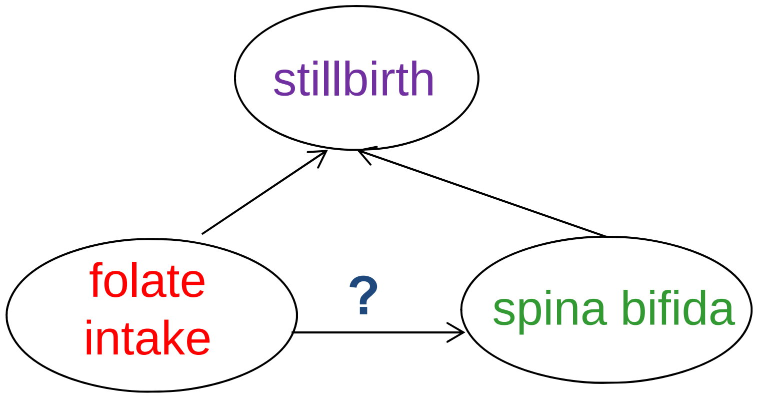
What if we adjust for stillbirths?

An adjusted or stratified analysis of observational data does not show an association of folate intake an spina bifida.

Data from a randomized clinical trial would show an association between folate intake and spina bifida.

What went wrong?

Looking into adjustment methods
A real example: effect of folate intake on spina bifida



Stillbirth is a common effect of both, lack of folate and spina bifida.

-> in a randomized clinical trial, where women are randomized to low or high folate intake, the two treatment groups are not equal in the rate of stillbirths

An adjusted or stratified analysis measures the association (folate – spina bifida) between groups with equal rates of stillbirths.

It is a pitfall to assume that the two randomized group are equal with respect to all possible co-variables.

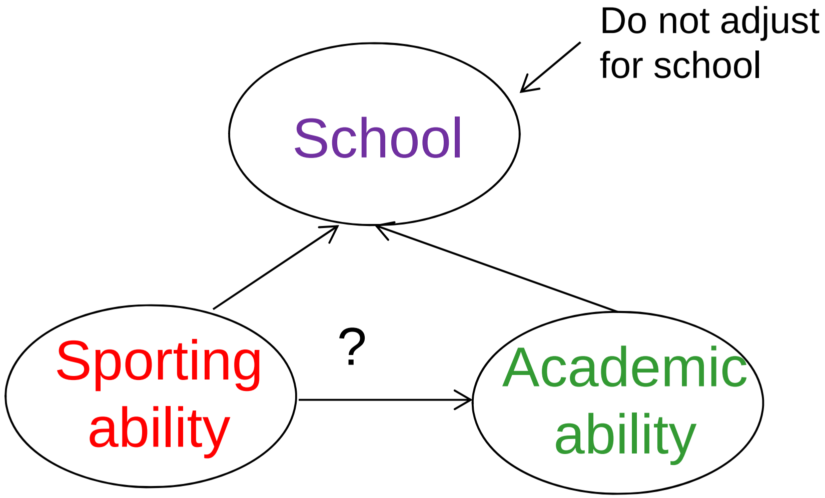
Looking into adjustment methods
Never adjust for a common effect: a toy example

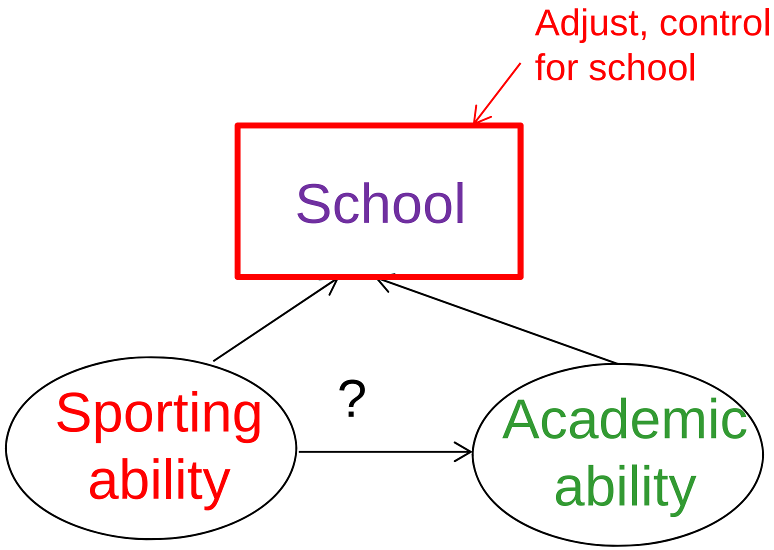
A school accepts pupils who are either good at sport, or good academically, or both

-> School acceptance is associated with sporting and academic abilities

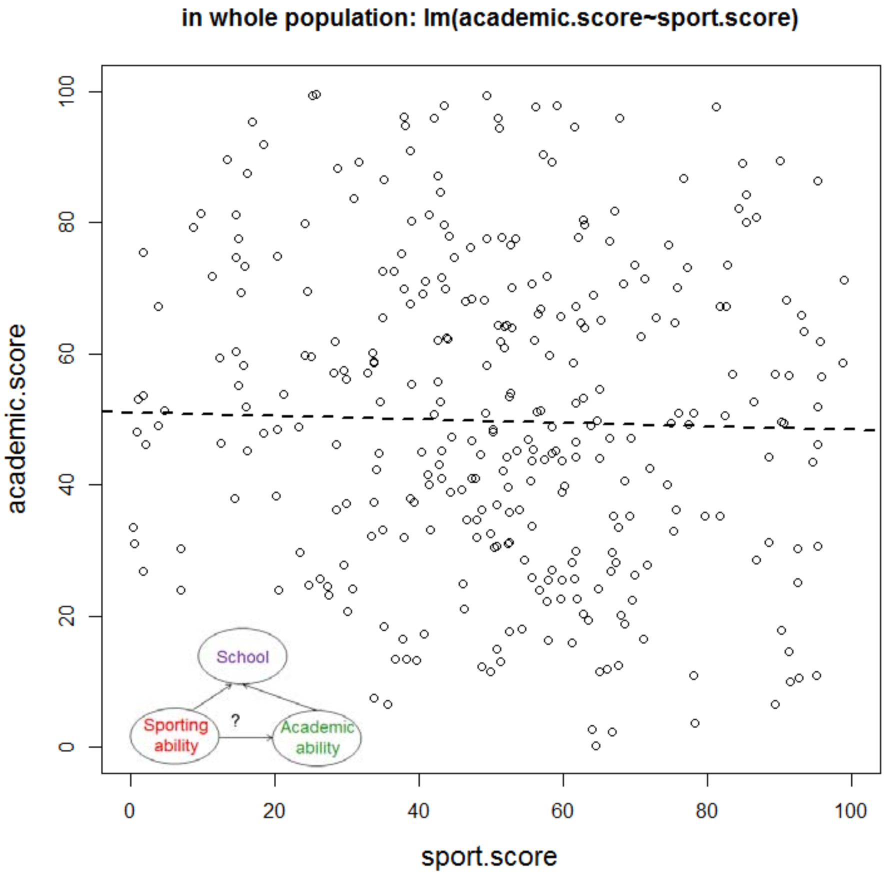
Suppose: in Population sport and academic skills are independent

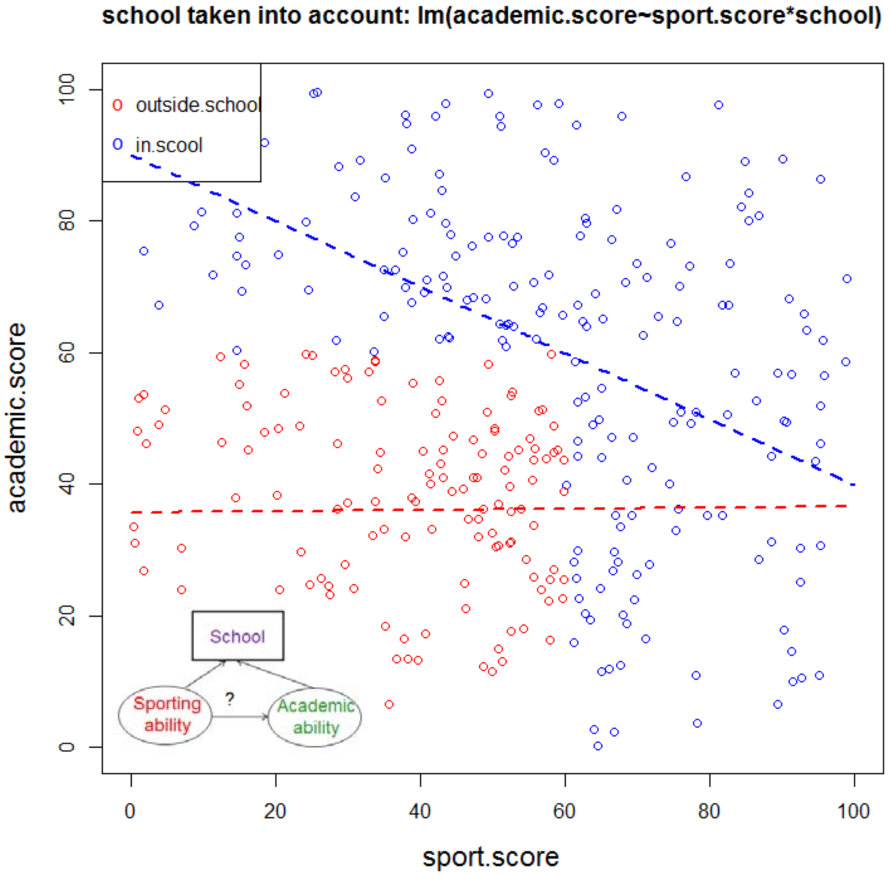
What happens if we “adjust” for the factor “accepted in school”?





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



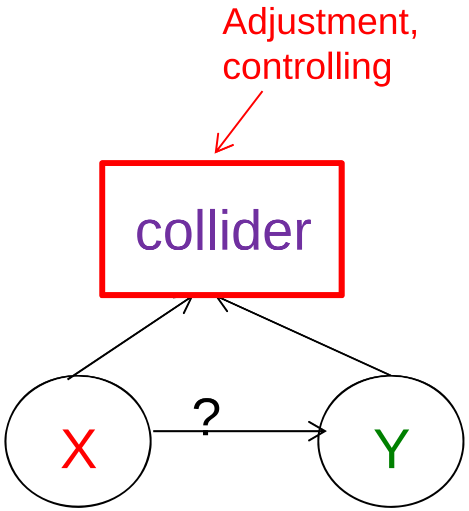


In the population there is no association between sport score and academic score, but by controlling □ for the school-variable we created a spurious association.

Looking into adjustment methods
Never adjust for a common effect

Adjusting for a collider-covariate, that is a common effect of X (treatment) and Y (outcome)

- Can hide an existing effect of X on Y
- Can lead to a spurious association of X and Y



Judea Pearl (2009):
"… in the bulk of the statistical literature before 2000, causal claims rarely appear in the mathematics. …
For example, the assumption that a **covariate not be affected by a treatment**, a **necessary assumption for the control of confounding** (Cox, 1958, p. 48), is expressed in plain English, not in a mathematical expression."

Introduction to causal analysis of observational data

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

$$\left.\begin{array}{c}\text{data}\\ \text{causal assumptions}\end{array}\right\} \Rightarrow \text{estimate causal effects}$$

A mathematical concept of causation must be able to (Pearl 2009):

- represent causal questions in some mathematical language
- provide a precise language for communicating underlying assumption
- provide a systematic way of answering causal questions
  and labeling others as "unanswerable"
- provide a method of determining what assumptions or new measurements would be needed to answer the "unanswerable" questions.

Counterfactuals and potential outcomes

What if??

How would the world look like if Dino's would have survived?

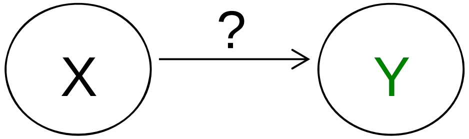
Would he live longer if he would always eat an apple instead of a cake?

Would we have earned more if we had doubled the price?

The unobserved outcome is called counterfactual.

Counterfactuals and potential outcomes
Individual-level causal effect

〈그림 4〉



causal effect of X $\in$ {A,B} on Y $\in$ {0,1}:

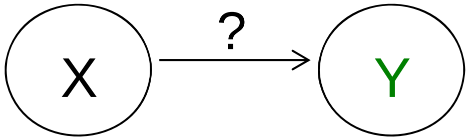
$$Y_A - Y_B =$$

$$\big(Y \mid do(X = A)\big) - \big(Y \mid do(X = B)\big)$$

The individual-level causal effect would be most interesting. However, in reality, we never observe both Y_A and Y_B on the same individual. Each individual gets treatment A or treatment B and we observe the outcome under the received treatment.

Knowing the counterfactuals we could determine causal effects.

Population-level causal effect



In reality we are only able to estimate the mean population-level causal effect e.g. by a randomized intervention study (RCT) where the treatment group A and B are exchangeable.

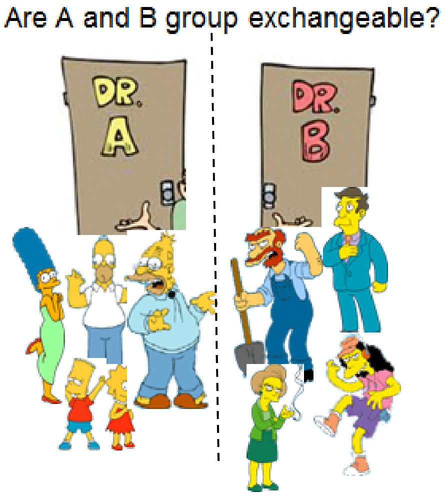
population-level causal effect of X ∈ {A,B} on Y ∈ {0,1} =

$$E(Y_A) - E(Y_B) = P(Y=1|do(X=A)) - P(Y=1|do(X=B))$$

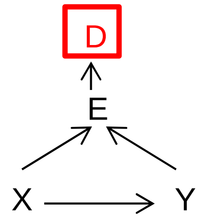
Externally (@random) set the treatment to A or B

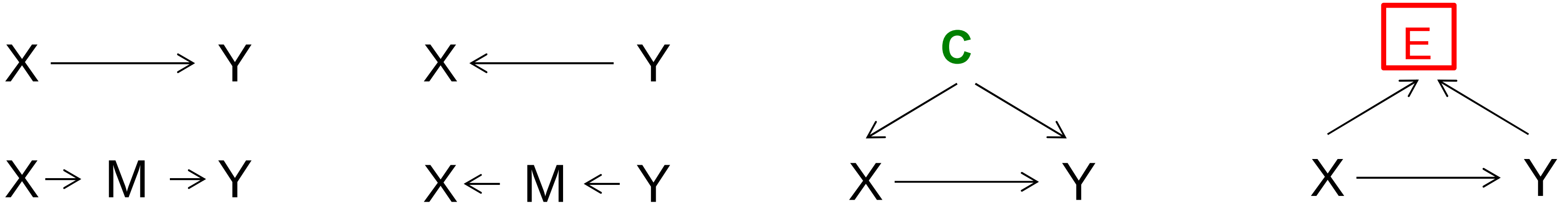
$$P(Y=1|X=B) \overset{\text{in general}}{\neq} P(Y=1|do(X=B))$$

We observe the probability of outcome 1 among the individuals who have chosen the treatment B



Association in observational data
Building blocks of causal graphs





X and Y will appear as associated if:

- X causes Y (directly or via a Mediator M)
- Y causes X (directly or via a Mediator M)

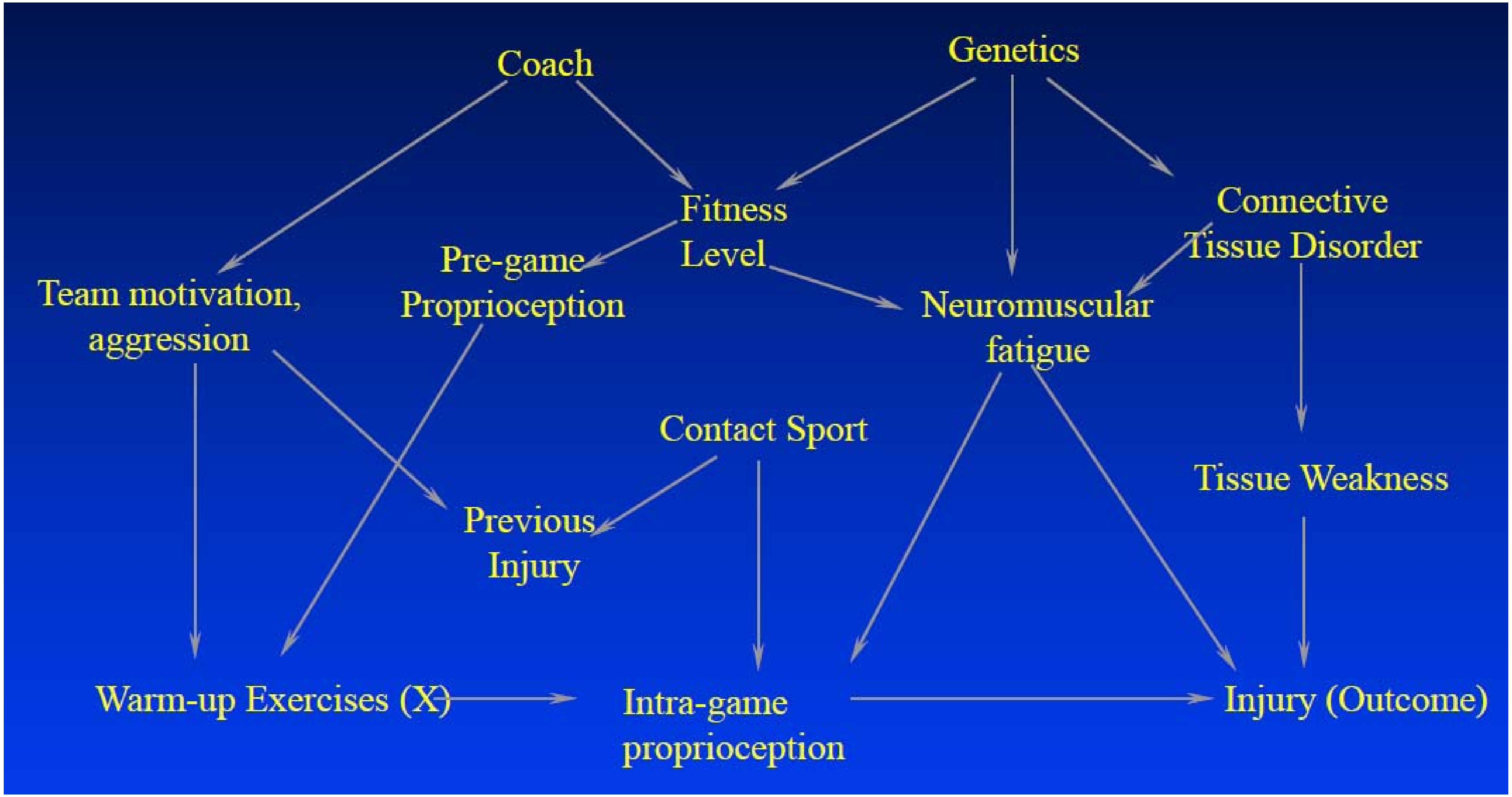
$\Rightarrow$ assocation $\triangleq$ causal effects

- X and Y have a common cause C for which we do not adjust
- X and Y have a common effect E for which do adjust

$\Rightarrow$ spurious assocation

Causal graphs can get complicated

For which set of covariates should we adjust (in a regression model) to estimate an unbiased effect of X (Warm-up Exercises) on the outcome (Injury)?

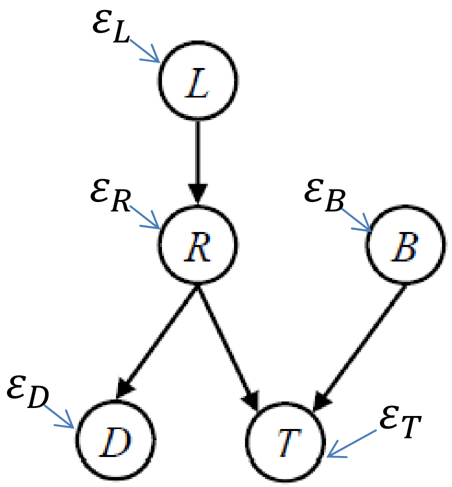


Taken from Pearl

A causal graph model

In a causal graph model:

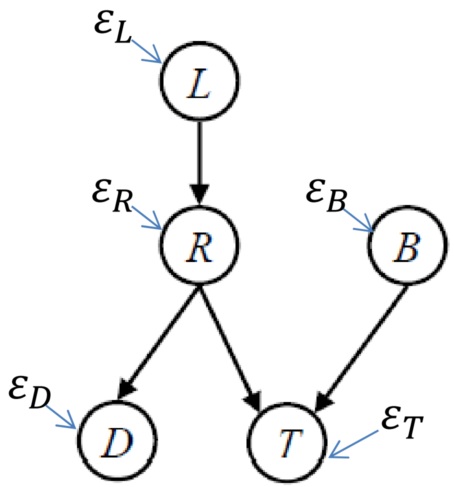
- Each vertex represents a random variable (observable or unobserved)
- Edges represent conditional dependencies
- An arrow indicates the direction of causation



the error terms ε are often called disturbances and are assumed to be jointly independent and often omitted in the DAG and the model.

A DAG as representation of a causal model

In a directed acyclic graph (DAG) all edges are directed an it is not possible to trace a cycle when following the arrowheads. A graphical model can be seen as map of dependence structures with arrows indicating the direction of potential causation. The underlying joint (faithful[1]) probability distribution which can be directly read of the (Markovian[2]) graph as **factorized product** of conditional probabilities:



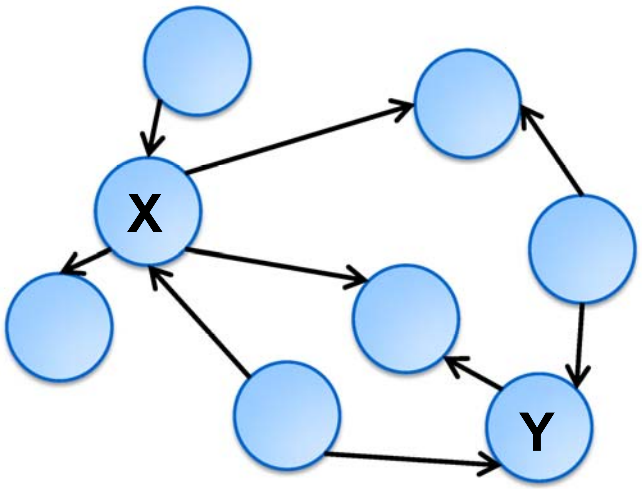
$$P(L,R,D,T,B) = \prod_{x\in\{L,R,D,T,B\}} P(\mathrm{x}\,|\,\mathrm{parent}(x))$$

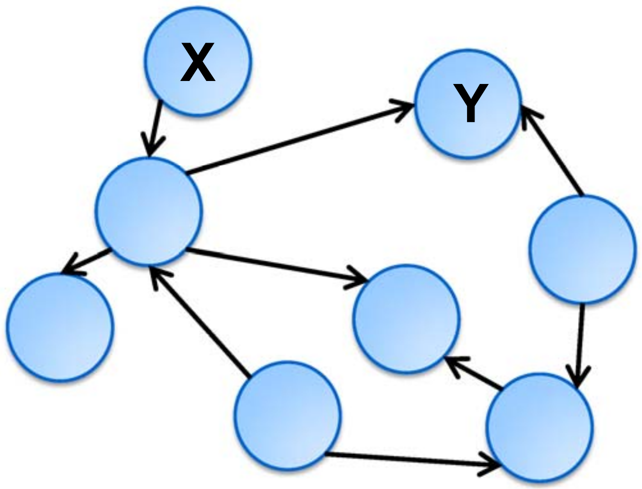
$$= P(L)\cdot P(R\,|\,L)\cdot P(\mathrm{D}\,|\,\mathrm{R})\cdot P(\mathrm{T}\,|\,\mathrm{R,B})\cdot P(\mathrm{B})$$

[1]: in a faithful distribution the conditional independencies perfectly matches the d-separation relations of the DAG – this is true for the vast majority of all distributions (Meek 1995) and therefor no strong restriction.
[2]: based on the assumption that all the error terms are jointly independent – these errors or disturbances are often omitted in the DAG and the model.

A DAG can represent a graphical causal model or causal Bayesian net (belief network).

Causal graphs helps us to judge identifiability of causal effects





Given the causal structure depicted in the DAG:

- Is it possible to identify the effect of X on Y?
  (causal effect from X on Y is transported paths starting with an arrow leaving X)

- If yes, for which set of covariates should we adjust to estimate an unbiased effect of X on Y (given by the coefficient if a linear model is assumed)? Y ~ X + ? + … + ?

Back-door criterion

To judge if it is it possible to identify the causal effect of X on Y we proceed as follows:

- Remove all arrows starting from X
  a path starting from X transports causes from X on Y
- Identify all open (active) back-door paths
  backdoor paths may introduce association which is not due to the causal effect of X on Y
- Determine whether a set S of covariates is sufficient to block all backdoor paths
  a open backdoor paths can be blocked by controlling/adjusting for a variable within the path

$X \perp Y \mid \mathrm{S}$: X and Y are independent if conditioned on S

: X and Y are d-separated by controlling the variables specified in set S

: all backdoor paths between X and Y are blocked if conditioning on S

: the causal effect of X on Y can be estimated when adjusting for variables in S

Identifying open and blocked paths

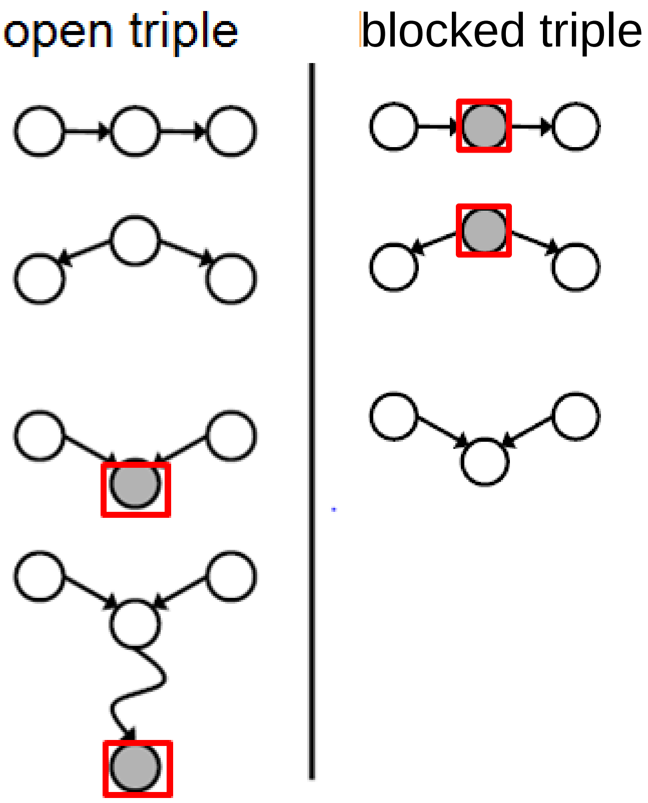
A path is a connection between X an Y – direction of the edges are ignored. Identify all (potential backdoor) paths between X and Y after removing all arrows starting from X.

Decompose each path in its triple-segments

Classify each triple segment as open or blocked

A path is open if all triple-segments are open

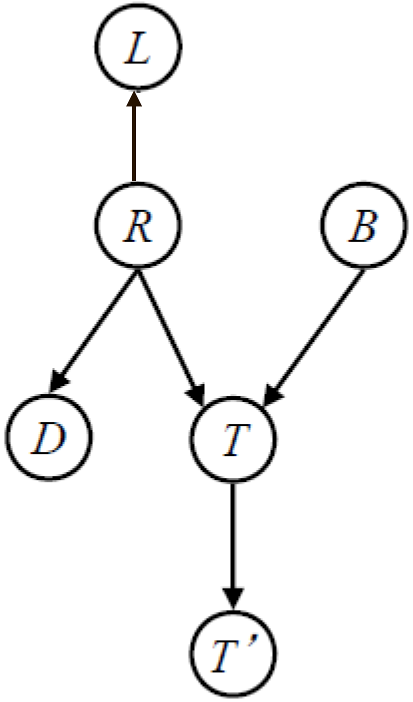
A path is blocked if a **single** triple-segment is blocked



□ controlled variable

Using the backdoor path criterion in an example

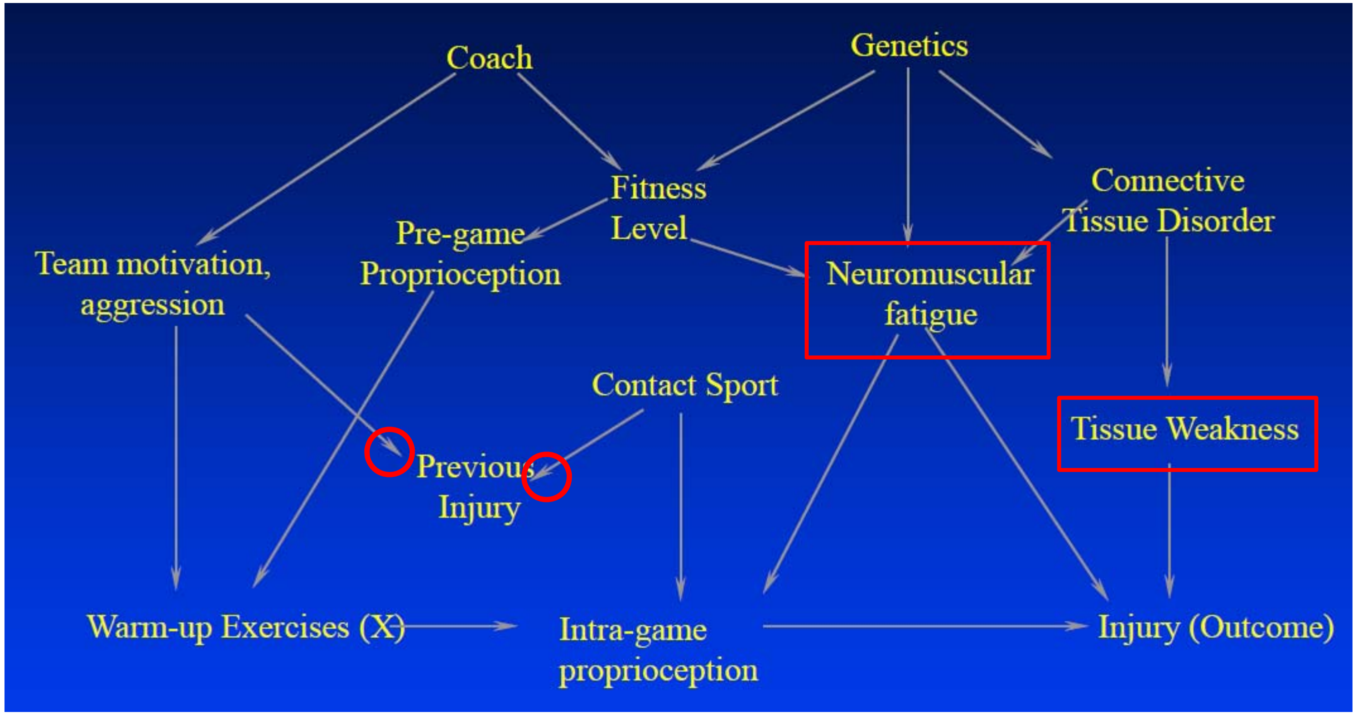
$L \perp\!\!\!\perp T' \mid T$	*Yes*
$L \perp\!\!\!\perp B$	*Yes*
$L \perp\!\!\!\perp B \mid T$	No
$L \perp\!\!\!\perp B \mid T'$	No
$L \perp\!\!\!\perp B \mid T, R$	*Yes*



(Arrows starting from L have been removed.)

Using the backdoor path criterion in a complex example

Can we estimate the effect of X on the outcome without bias, if we adjust for “Neuromuscular fatigue” and “Tissue Weakness””?



Taken from Pearl

If there is no open backdoor path from X to the Outcome, meaning X is disconnected from the Outcome (X and Y are d-separated), there is no confounding!

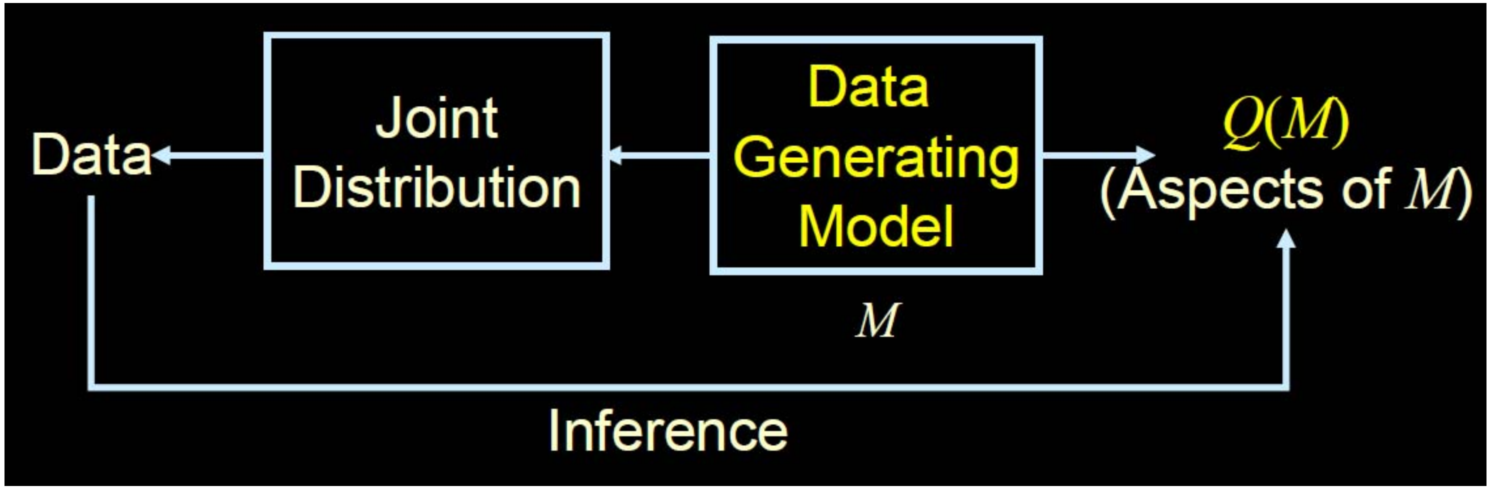
Yes, the causal effect of X on the outcome is identifiable, meaning we can estimate it without bias from the corresponding regression model (assuming no interactions).

From joint distribution towards the structural model paradigm

Joint distribution can help to answer questions like:

How likely fails a dutch female student on subject A if she fails on subject B?

It would be nice to use directly the data generating model !

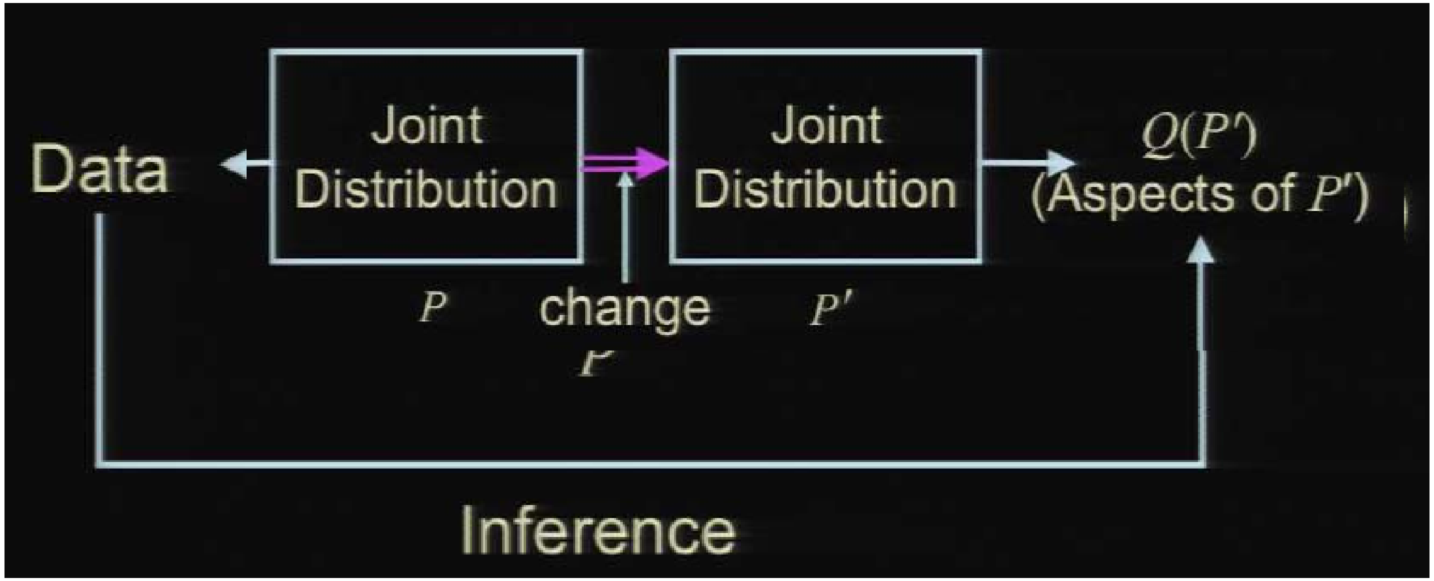


M – Invariant strategy (model, mechanism, recipe, physical law, protocol) by which Nature assigns values to variables in the analysis.

How to deal with changes or interventions?

From observational data we can e.g. estimate the joint distributions $P(y,x_1,\ldots,x_p)$

We cannot estimate how the joint distribution will change upon intervention.



Taken from Pearl

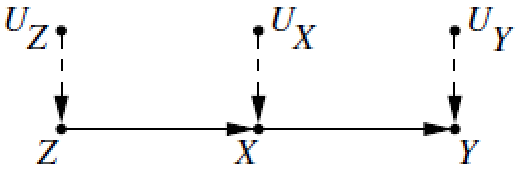
How would our sales change if we double the price?

How would the cancer rate be if we ban smoking?

-> to deal with intervention causal graphical model / causal Bayesian networks are needed

How to use causal graph to deal with interventions?

Example:
X: represents a treatment variable,
Y: a response variable, and
Z: covariate that affects the amount of treatment received
U: jointly independent disturbances (errors)



pre-intervention graph

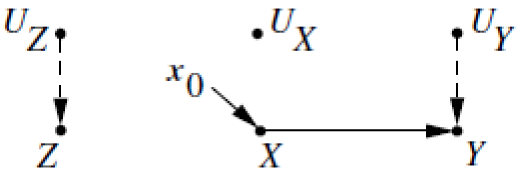
$$P(z, y, x) = P(z) \cdot P(x \mid z) \cdot P(y \mid x)$$

Structural equations:

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= f_X(z, u_X) \\ y &= f_Y(x, u_Y) \end{aligned}$$

Intervention:
The treatment X is set to the value x_0 for all individuals in the population.

Question:
Can the controlled (post-intervention) distribution, P*(Y = y | do(x)), be estimated from data collected from the pre-intervention distribution P(z, x, y)?



post-intervention graph

$$P^*(y \mid do(X = x_0)) = ?$$

graphs taken from Pearl

How to use causal graphs to deal with interventions?

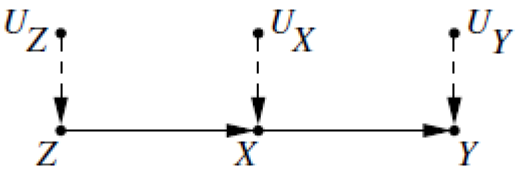
For a (Markovian) model M represented by a DAG, the distribution P* generated by an intervention do(X = x_0) on a set X of pre-interventional observed variables is given by **the truncated factorization** of the pre-interventional distribution P :

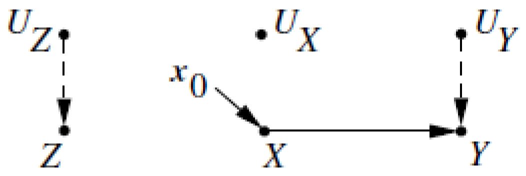
Pearl

G formula:

$$P^*(\mathrm{v}_1,\mathrm{v}_2,\ldots,\mathrm{v}_k \mid \mathrm{do}(\mathrm{x}_0)) = \prod_{i|V_i \notin X} \mathrm{P}\big(\mathrm{v}_i \mid \mathrm{parent}(\mathrm{v}_i)\big)$$

This instructs us to remove from the full factorization all factors associated with the intervened variables X, since they have with probability 1 the imposed value.





pre-intervention graph

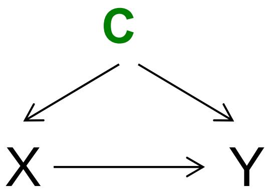
post-intervention graph

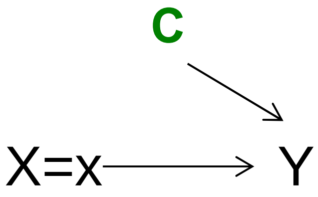
$$P(\mathrm{z},\mathrm{y},\mathrm{x}) = \mathrm{P}(\mathrm{z}) \cdot \mathrm{P}(\mathrm{x} \mid \mathrm{z}) \cdot \mathrm{P}(\mathrm{y} \mid \mathrm{x})$$

$$P^*(\mathrm{y} \mid \mathrm{do}(\mathrm{X} = \mathrm{x}_0)) = \mathrm{P}(\mathrm{y} \mid \mathrm{X} = \mathrm{x}_0)$$

Remark: from a causal graph or structural model we can determine counterfactuals or predict effects after interventions!

How to use causal graphs to deal with interventions?





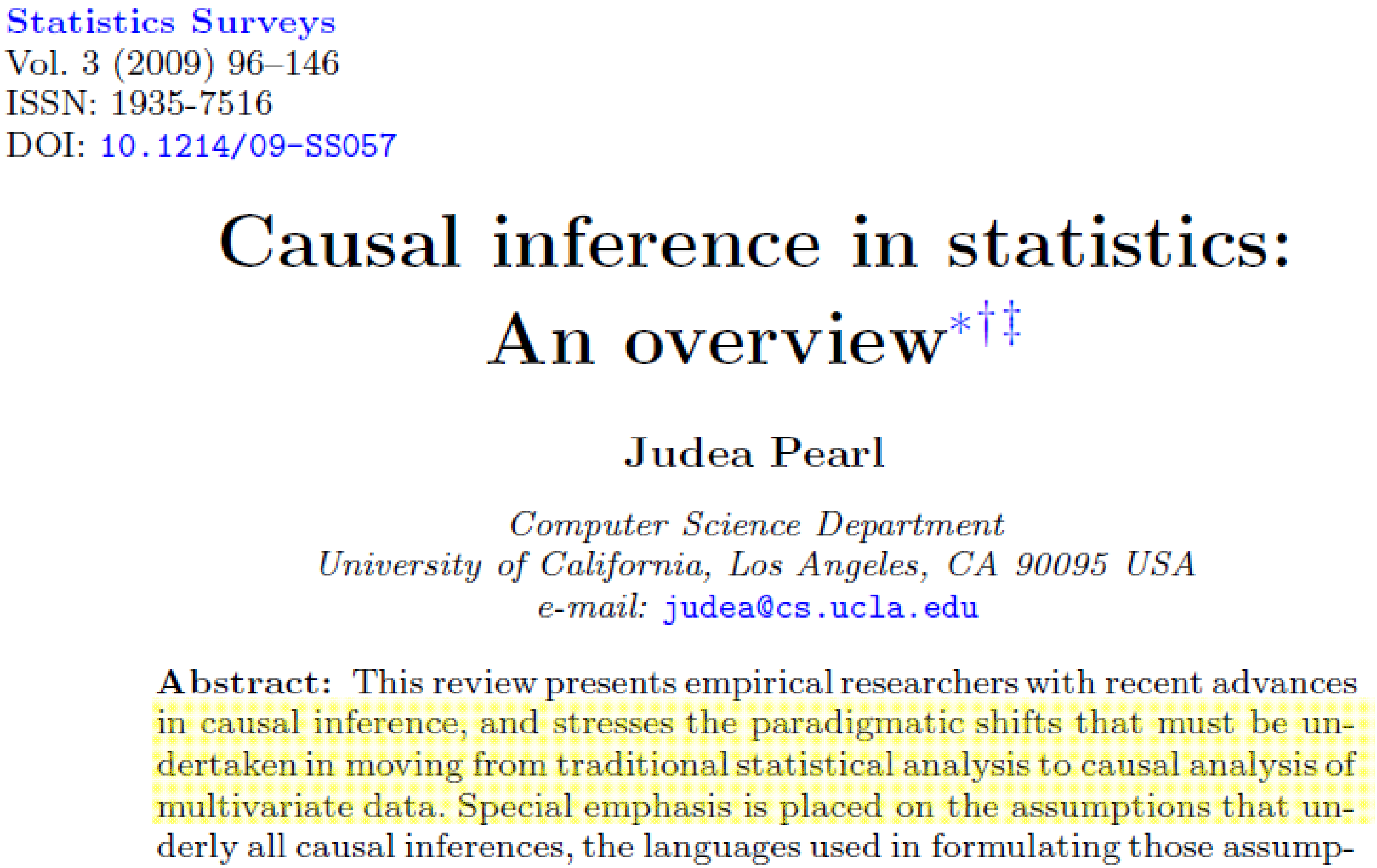
$P(\mathrm{x},\mathrm{y},\mathrm{c}) = \mathrm{P}(\mathrm{c})\cdot\mathrm{P}(\mathrm{x}\,|\,\mathrm{c})\cdot\mathrm{P}(\mathrm{y}\,|\,\mathrm{x},\mathrm{c})$

$P(\mathrm{y},\mathrm{c}\,|\,\mathrm{do}(\mathrm{X}=\mathrm{x}) = \mathrm{P}(\mathrm{c})\cdot\mathrm{P}(\mathrm{y}\,|\,\mathrm{x},\mathrm{c})$

e.g. population-level causal effect of X $\in \{\mathrm{A},\mathrm{B}\}$ on Y $\in \{0,1\}$ =

$$E(Y_A) - E(Y_B) = P\big(Y = 1\,|\,do(X = A)\big) - P\big(Y = 1\,|\,do(X = B)\big)$$

How to learn more about causal inference?



Statistics Surveys
Vol. 3 (2009) 96–146
ISSN: 1935-7516
DOI: 10.1214/09-SS057

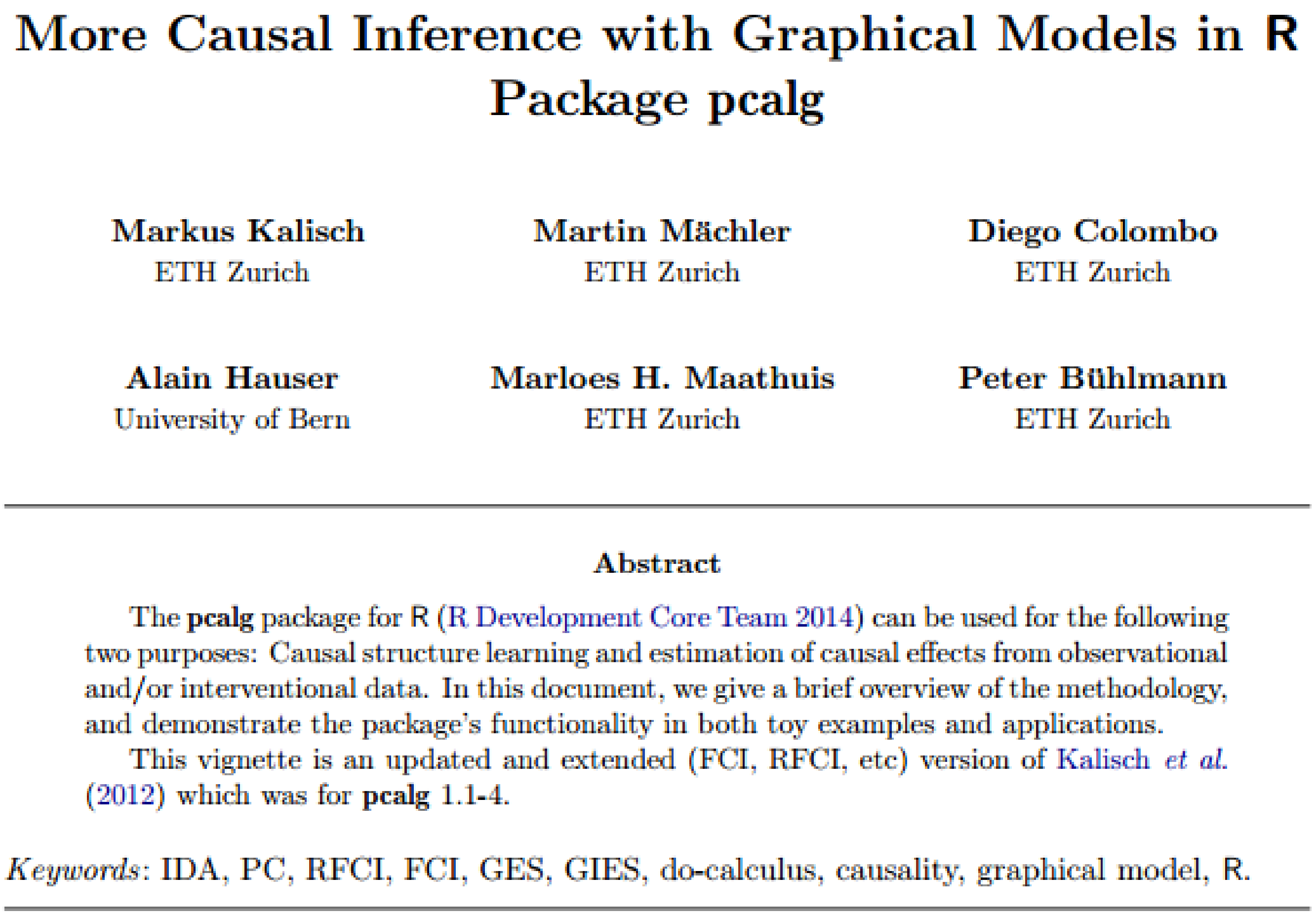
Causal inference in statistics: An overview[*†‡]

Judea Pearl

Computer Science Department
University of California, Los Angeles, CA 90095 USA
e-mail: judea@cs.ucla.edu

Abstract: This review presents empirical researchers with recent advances in causal inference, and stresses the paradigmatic shifts that must be undertaken in moving from traditional statistical analysis to causal analysis of multivariate data. Special emphasis is placed on the assumptions that underly all causal inferences, the languages used in formulating those assump-

How to do causal inference in R?



More Causal Inference with Graphical Models in R Package pcalg

Markus Kalisch
ETH Zurich

Martin Mächler
ETH Zurich

Diego Colombo
ETH Zurich

Alain Hauser
University of Bern

Marloes H. Maathuis
ETH Zurich

Peter Bühlmann
ETH Zurich

Abstract

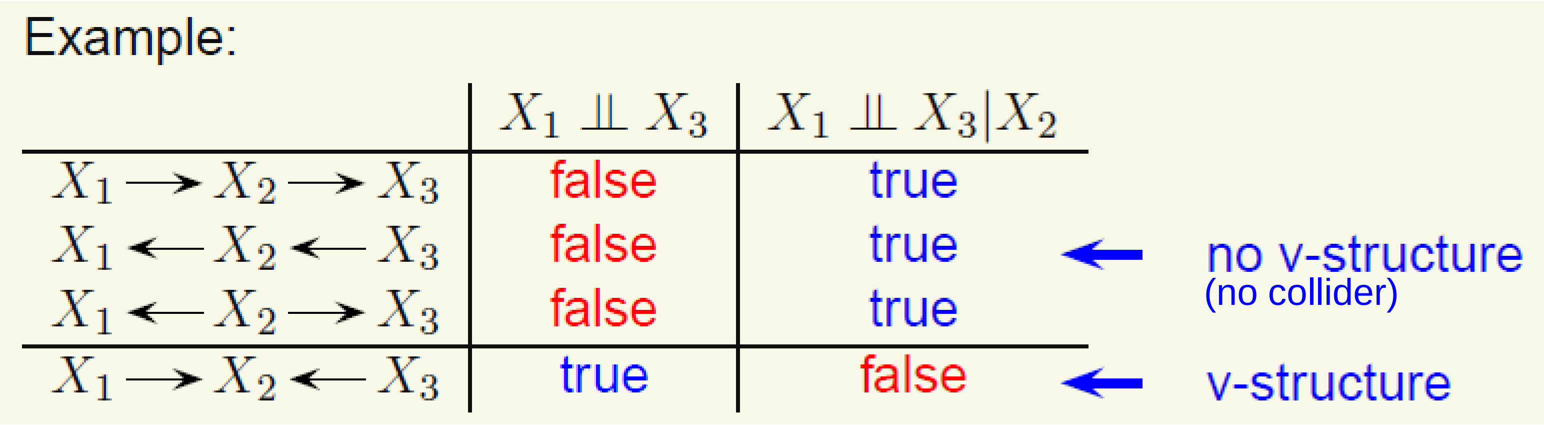
The **pcalg** package for R (R Development Core Team 2014) can be used for the following two purposes: Causal structure learning and estimation of causal effects from observational and/or interventional data. In this document, we give a brief overview of the methodology, and demonstrate the package's functionality in both toy examples and applications.

This vignette is an updated and extended (FCI, RFCI, etc) version of Kalisch *et al.* (2012) which was for **pcalg** 1.1-4.

Keywords: IDA, PC, RFCI, FCI, GES, GIES, do-calculus, causality, graphical model, R.

Basics of the pcalg R package for causal inference

Main idea when estimating causal graphs from data:
A DAG encodes conditional independence relationships.
So first determine all conditional independence relationships in the observational distribution. However, several DAGs can encode the same conditional independence relationships. They are Markov equivalent.
All DAGs in a Markov equivalence class have the same edges and the same v-structures (colliders) (Verma and Pearl, 1990).



Example:

	$X_1 \perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	
$X_1 \rightarrow X_2 \rightarrow X_3$	false	true	
$X_1 \leftarrow X_2 \leftarrow X_3$	false	true	← no v-structure (no collider)
$X_1 \leftarrow X_2 \rightarrow X_3$	false	true	
$X_1 \rightarrow X_2 \leftarrow X_3$	true	false	← v-structure

Taken from M.Maathuis

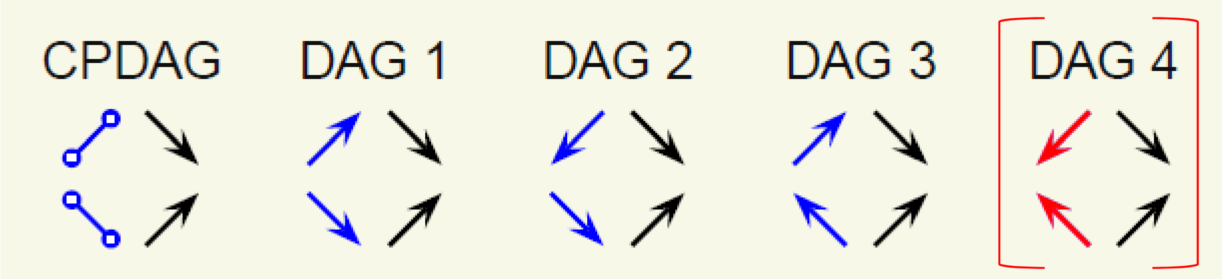
Knowledge about many conditional independence relationships may allow to determine the direction of arrows in the graphical model.

Basics of the pcalg R package for causal inference

- This means that the skeleton of a DAG is determined uniquely by conditional independence relationships
- But the directions of the edges are generally not uniquely determined -> CPDAG only shows consistent arrows

PC-algorithm of Peter Spirtes and Clark Glymour

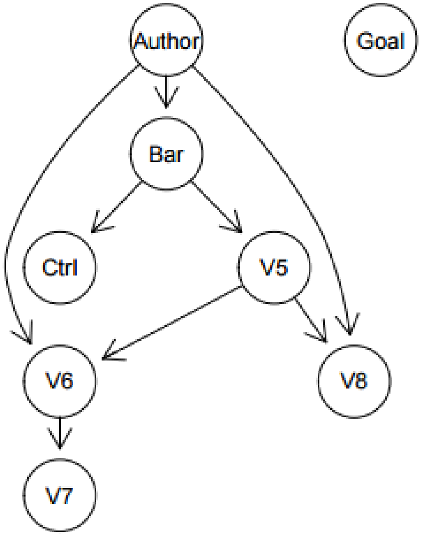
- Determine the skeleton (edges)
- Determine the v-structures (collider)
- Direct as many of the remaining edges as possible



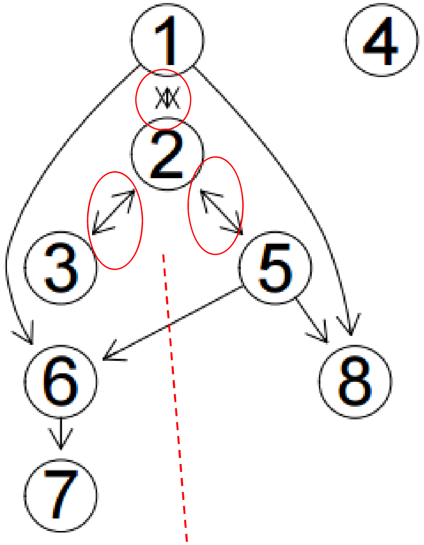
Taken from M.Maathuis

Validation on simulated data

The simulated data set contains p = 8 continuous variables with Gaussian noise and n = 5000 observations. The “true” causal relation with known effects strengths was predefined and used for the simulation.

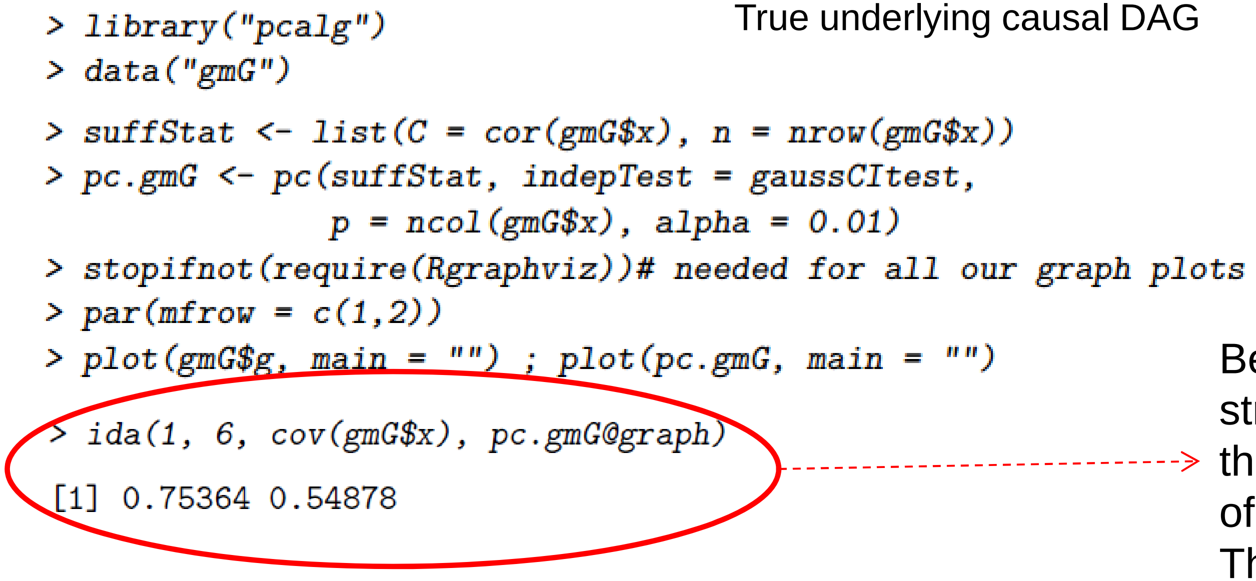


True underlying causal DAG



Estimated causal structure

Figure is taken from the pcalg vignette



```
> library("pcalg")
> data("gmG")
> suffStat <- list(C = cor(gmG$x), n = nrow(gmG$x))
> pc.gmG <- pc(suffStat, indepTest = gaussCItest,
               p = ncol(gmG$x), alpha = 0.01)
> stopifnot(require(Rgraphviz))# needed for all our graph plots
> par(mfrow = c(1,2))
> plot(gmG$g, main = "") ; plot(pc.gmG, main = "")
> ida(1, 6, cov(gmG$x), pc.gmG@graph)
[1] 0.75364 0.54878
```

Because of the estimated structure has undirected edges, the estimate of the causal effect of V1 on V6 is **not unique**.
The “true” causal effect is 0.52.

Validation on real and quite big data

With n=63 samples of observational data measuring the expression of p=5361 genes in yeast, the ida-function of the pcalg-package was used to identify the largest intervention effects between all pairs of genes.

Predict the effect of the knock-down of different genes by using the do-calculus and the causal graphs.

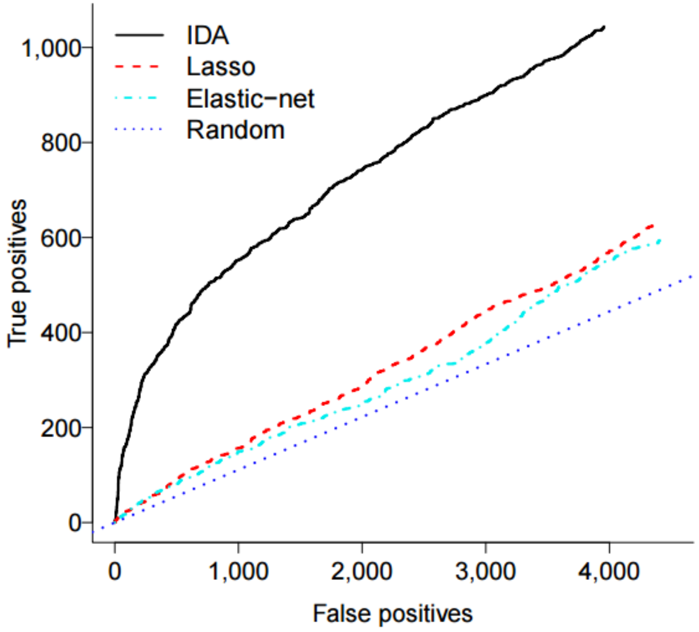


Figure is taken from the pcalg vignette

The largest 10% of the causal effects found in experiments among yeast genes

The 10% strongest predicted effects are compared to the 10% largest observed gene expression changes measured in intervention studies – here in 234 single-gene deletion mutant strains compared to the wild type yeast.
A ROC curve was used to visualize, how well the ranking of the predicted causal effects identify effects in the experimental data set.

Outlook

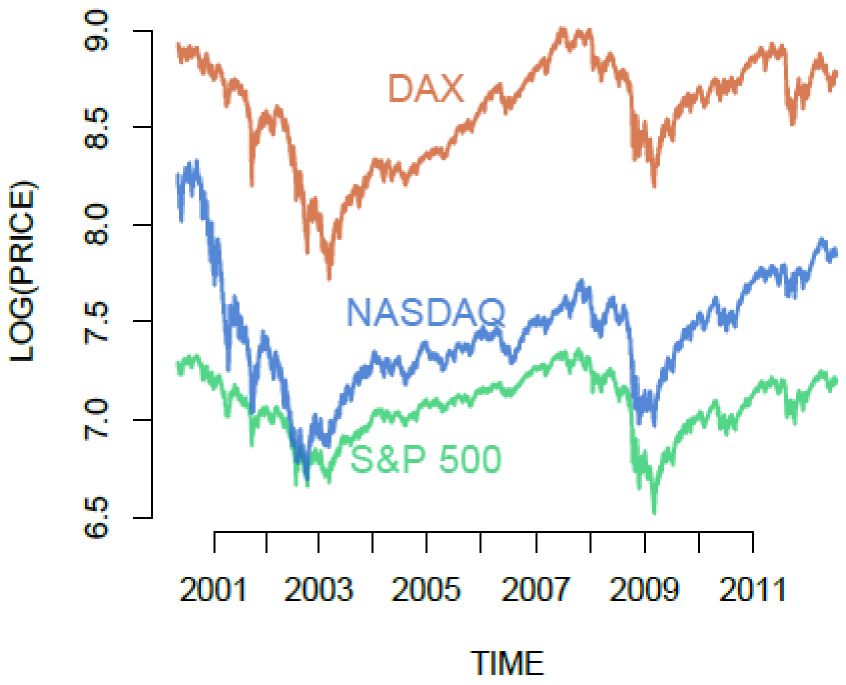
Hidden (or latent) variables:
Factors influencing two or more measured variables may not themselves be measured.
Selection variables:
Variables influencing whether a unit is included in the data sample.
Feedback loops:
In practice we often have feedback loops among variables (no DAGs)

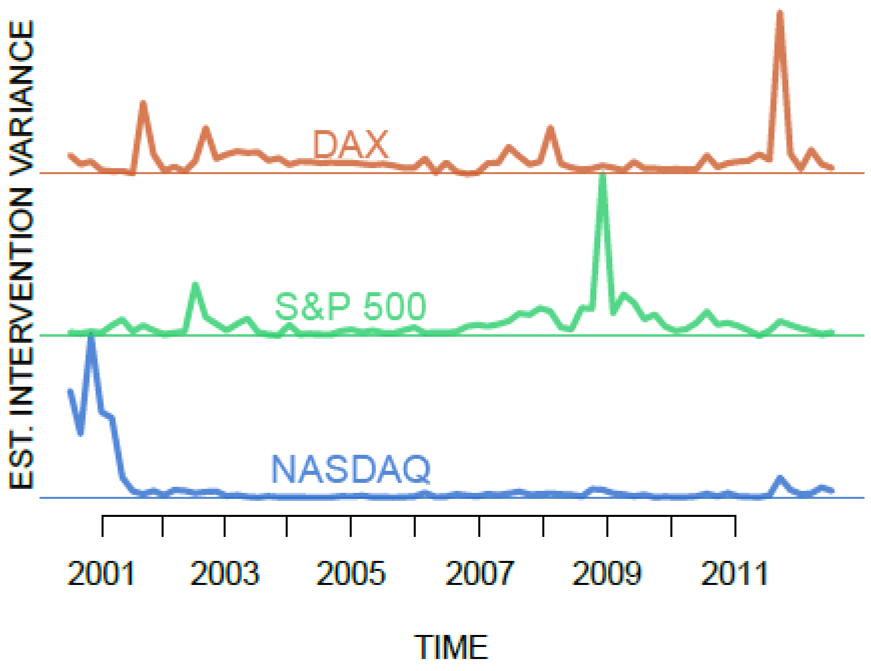
Methods are currently developed and implemented allowing for causal inference
for continuous variables and some further restrictions (relations, errors)

- on observational data without hidden or selection variables, no feedbacks
- on observational data with hidden but no selection variables, no feedbacks
- mixture of interventional and optional observational data without hidden or selection variables, no feedbacks
- mixture of interventional and optional observational data with hidden variables and feedback loops

Cyclic Graphs modling stock development with hidden variables

Environment is clearly changing over time which is modeled by hidden variables





From the fitted causal model the origins of the following three major down-turns of the markets were identified:

- Technology is the epicenter of the dot-com crash in 2001 (NASDAQ as proxy)
- American equities during the financial crisis in 2008 (proxy is S&P 500)
- European instruments (DAX as best proxy) during the August 2011 downturn

backShift: Learning causal cyclic graphs from unknown shift interventions
Dominik Rothenhäusler, Christina Heinze, Jonas Peters, Nicolai Meinshausen *(Submitted on 8 Jun 2015)*

Summary

Advantages when using methods of causal inference:

- **cause-effect relationships between variables can be tackled**
  - with transparent assumptions
  - with systematic mathematical methods
- **Causal graphs allow to**
  - visualize (assumed or fitted) underlying causal relationships
  - represent the joint probability distribution
  - investigate if a causal effect can be estimated without bias
  - identify sets of covariates for which one should adjust
  - avoid misleading stratification or adjustment with respect to colliders
  - formulate testable hypothesis which could be used to falsify the assumptions
  - test consistency between data and models
  - determine counterfactuals and predict effect of interventions
  - much more which has not been discussed in this talk!