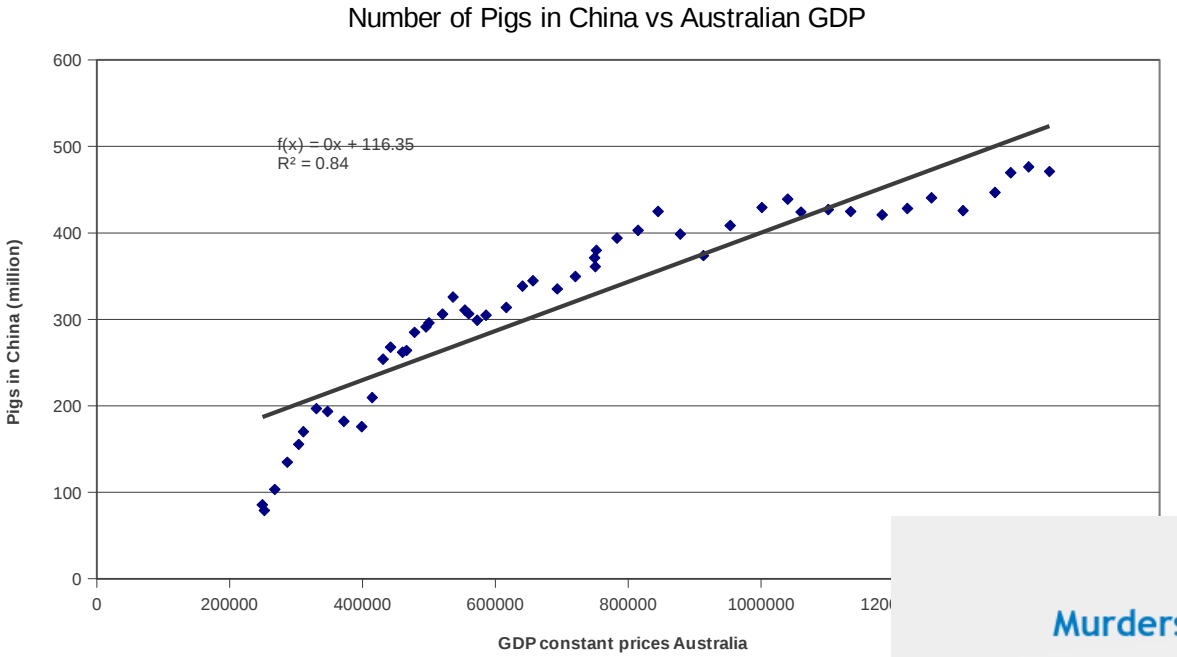
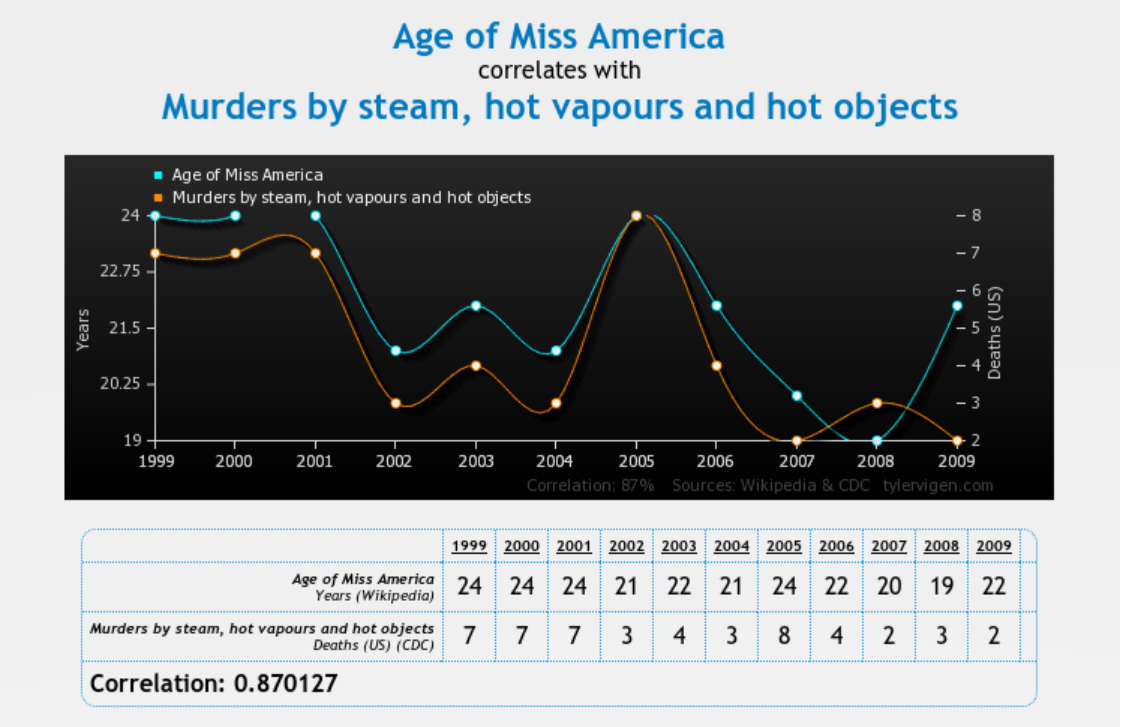
Correlation is not causation

Ways things can go wrong





	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Age of Miss America *Years (Wikipedia)*	24	24	24	21	22	21	24	22	20	19	22
Murders by steam, hot vapours and hot objects *Deaths (US) (CDC)*	7	7	7	3	4	3	8	4	2	3	2

Correlation: 0.870127

Image source: www.tylervigen.com/

We care about causality

Chocolate 'may help keep people slim'

COMMENTS (251)

By Michelle Roberts
Health reporter, BBC News

People who eat chocolate regularly tend to be thinner, new research suggests.

The findings come from a study of nearly 1,000 US people that looked at diet, calorie intake and body mass index (BMI) - a measure of obesity.

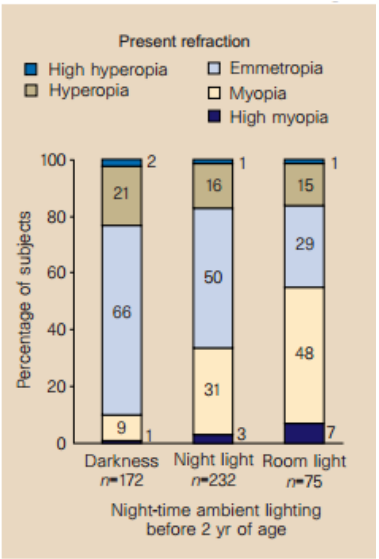
It found those who ate chocolate a few times a week were, on average, slimmer than those who ate it occasionally.

Even though chocolate is loaded with calories, it contains ingredients that may favour weight loss rather than fat synthesis, scientists believe.

Chocolate contains antioxidants but is also high in fat and sugar

Related Stories

http://www.bbc.com/news/health-17511011



Sleeping with the light on is associated with short-sightedness in kids.

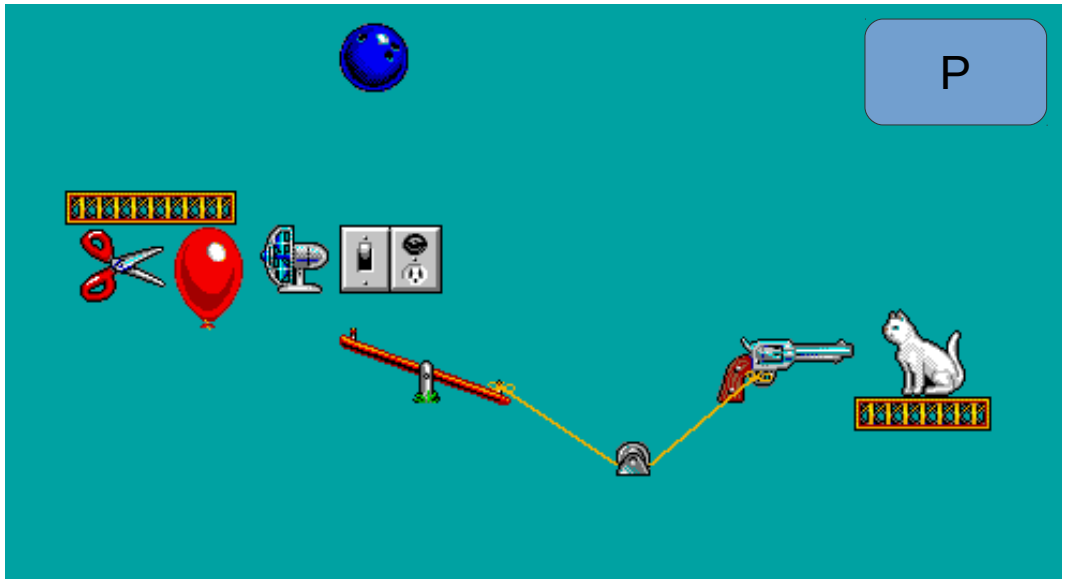
Quinn, Graham E., et al. "Myopia and ambient lighting at night." Nature 399.6732 (1999): 113-114

Childhood obesity partly caused by strict parenting, say scientists

Parents who struck a balance between being strict and kind were less likely to bring up obese children

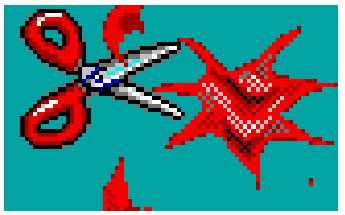
http://www.independent.co.uk/life-style/health-and-families/health-news/childhood-obesity-partly-caused-by-strict-parenting-say-scientists-9206147.html

Machine Learning



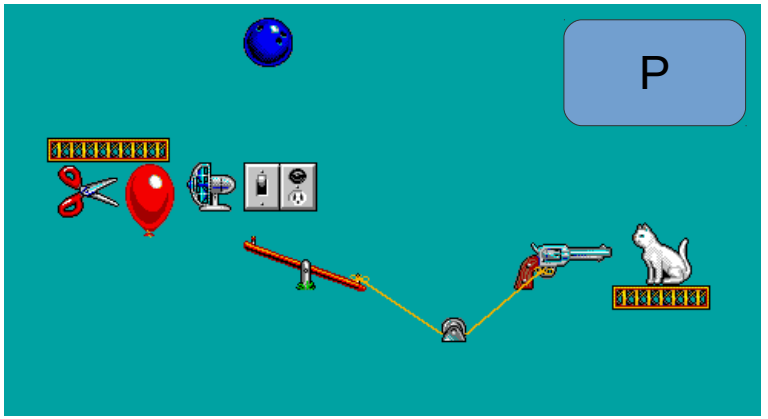
Use data generated by a process P to make a model that can predict some variable(s) in terms of others

Data sampled from P

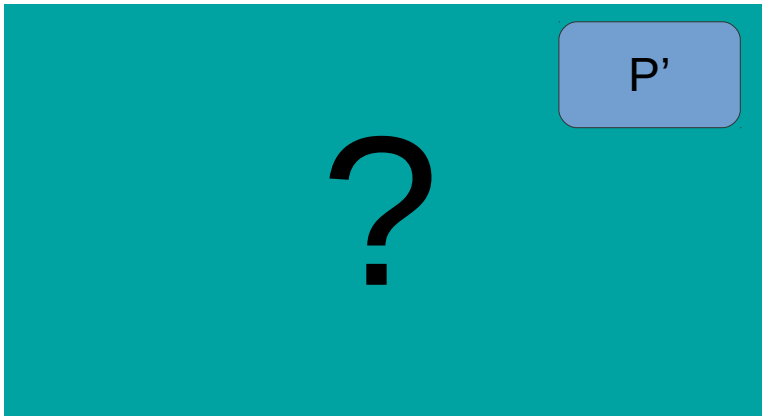


?

Causal inference



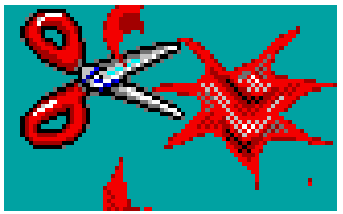
Do something



Data sampled from P





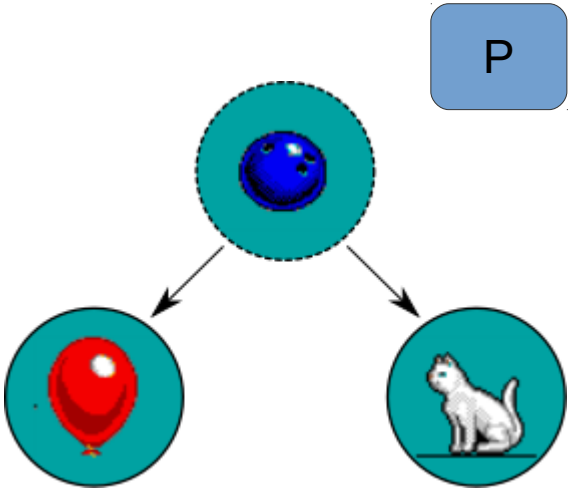


?

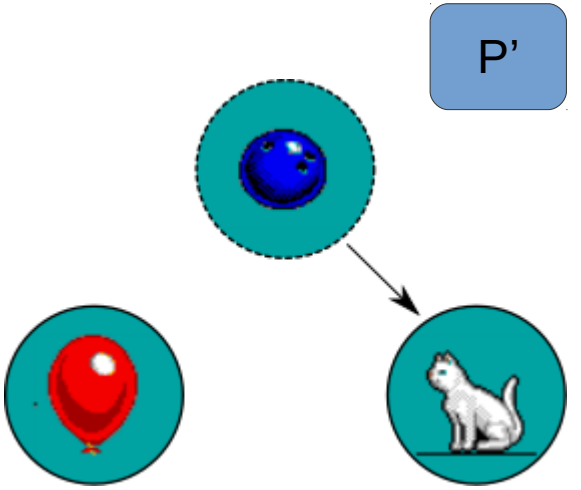
What can the data from P tell us about P'?

Causal bayesian networks

- Just consider changes to P that fix the value of some variable(s)
- A causes B if changing the system such that A is forced to some value, do(A=a), can change the distribution of B.
- A directly causes B, (A → B) if changing A still changes B, after holding constant all other variables that are causes of B and consequences of A.



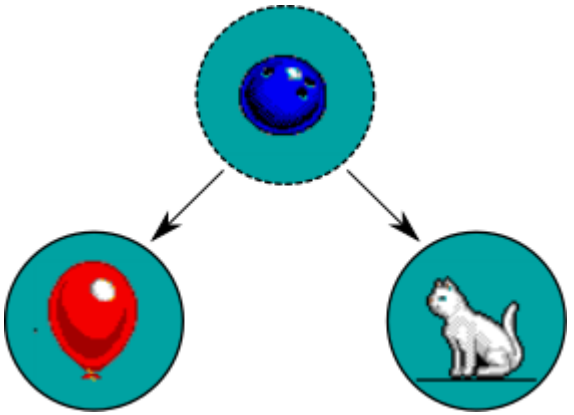
Causal network for original system

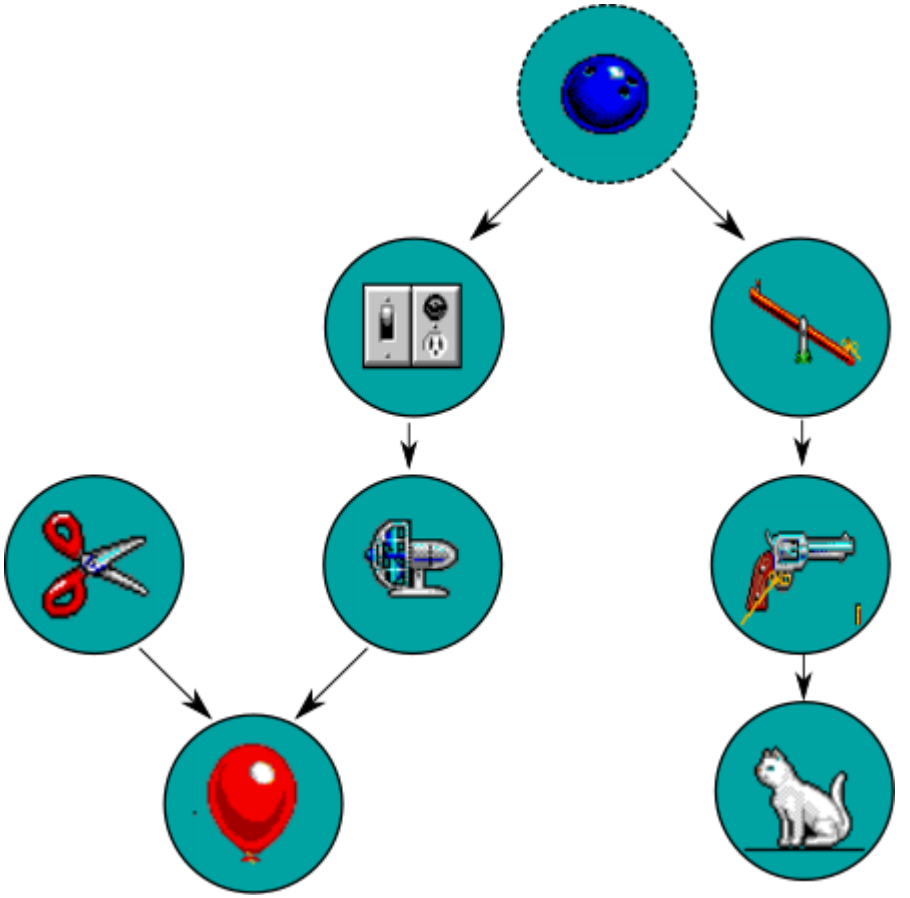


Causal network after do(Balloon = ?)

Causal bayesian networks

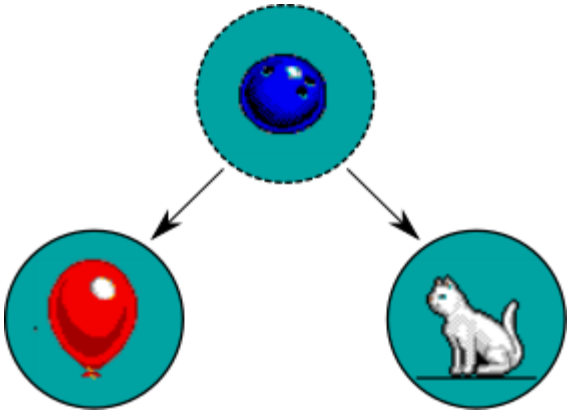
- A directly causes B, (A → B) if changing A still changes B, after holding constant all other variables that are causes of B and consequences of A. (This depends on what variables we consider)





Causal bayesian networks

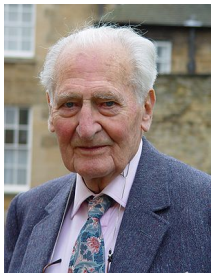
Causal bayesian networks have same factorisation properties as regular bayesian networks.

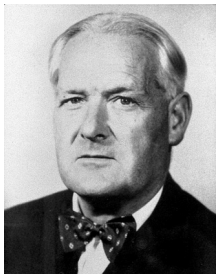


$$P(Ball, Balloon, Cat) = P(Ball)P(Balloon|Ball)P(Cat|Ball)$$

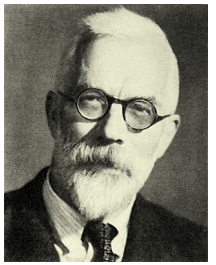
The key extra thing they give you is a way to go from the original network, to the network given an intervention on any variable (just delete all the links into the set variable).

A real life example





Smoking → Cancer



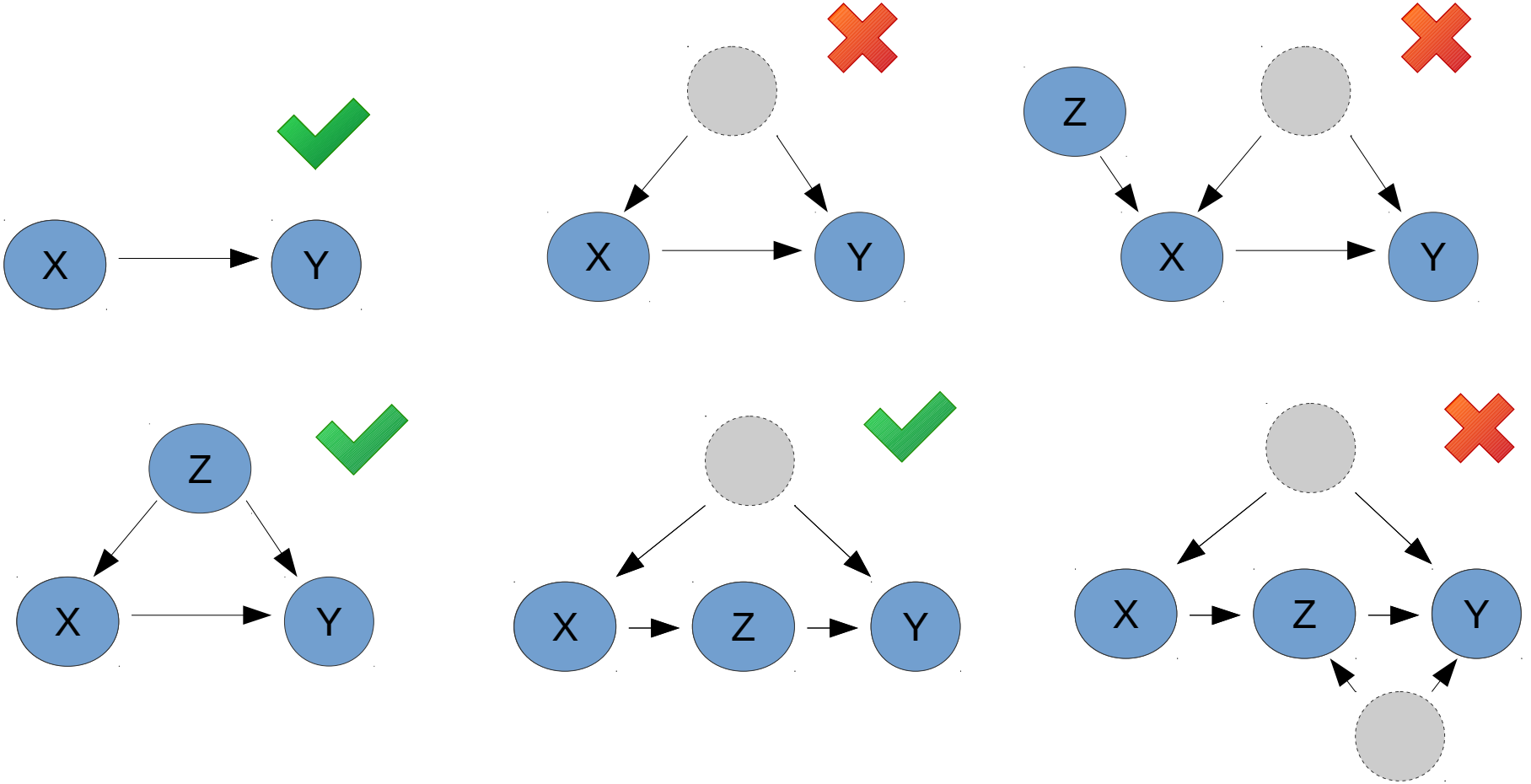
Genetics → Smoking, Genetics → Cancer

Genetics → Smoking → Cancer

heavy smokers 30 times more likely to die of lung cancer than non-smokers, but does smoking cause lung cancer?

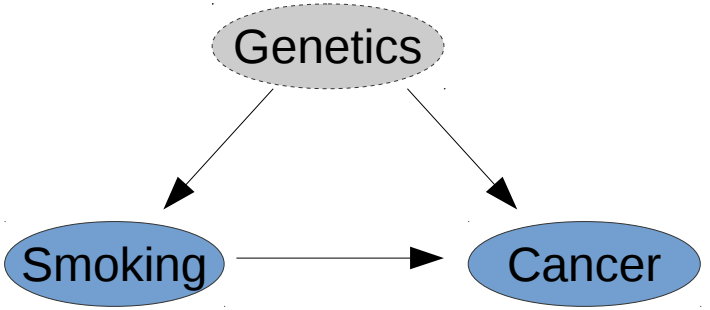
The do calculus

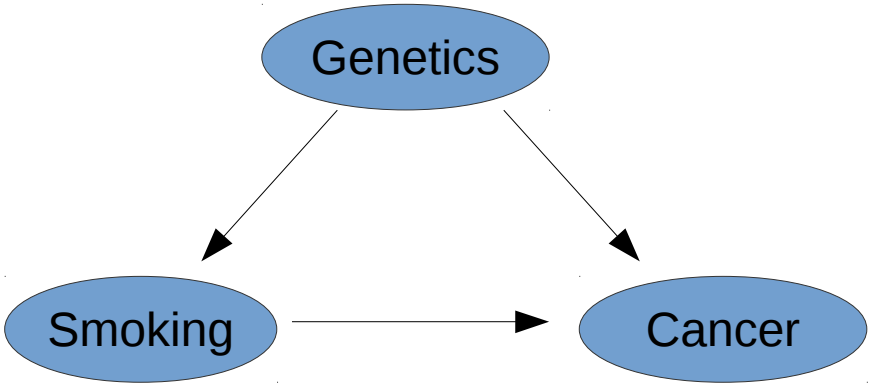
When can we estimate the effect of intervening on X on Y (from observational data)?



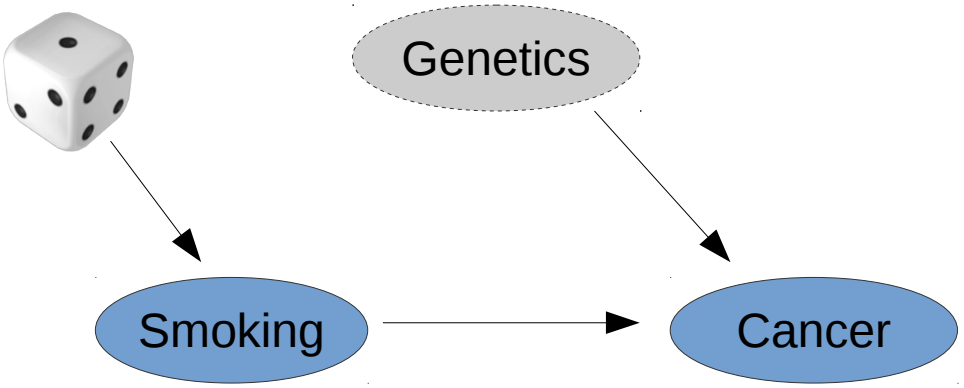
Estimating causal effects

How can we solve this smoking problem?





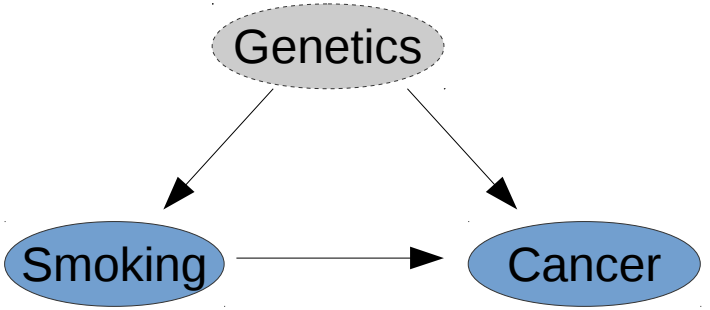
Observational (adjust for genetics)

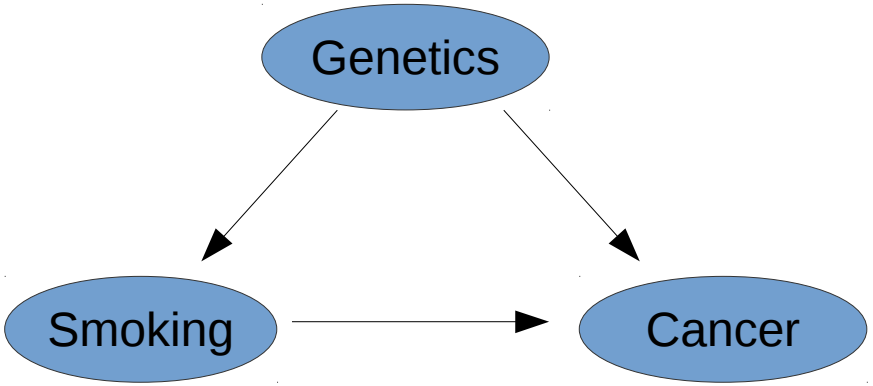


Randomised experiment

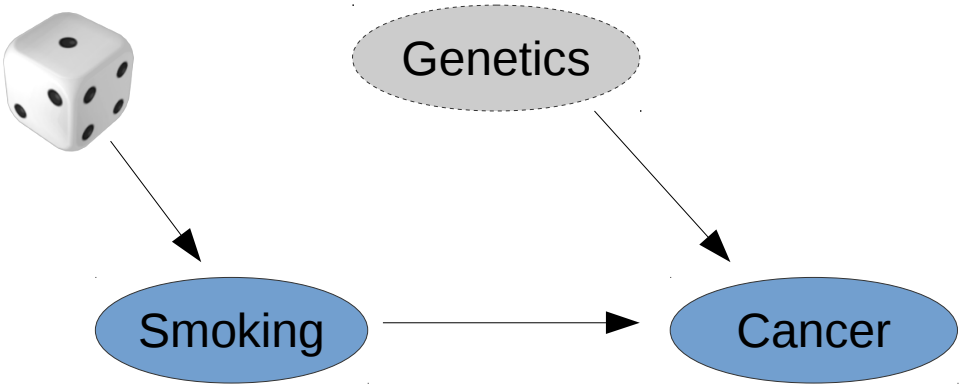
Estimating causal effects

How can we solve this smoking problem?





Observational (adjust for genetics)



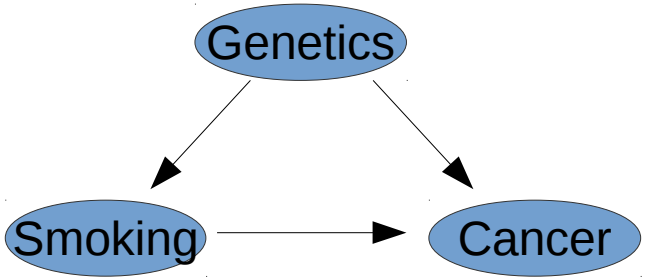
Randomised experiment

Experiments

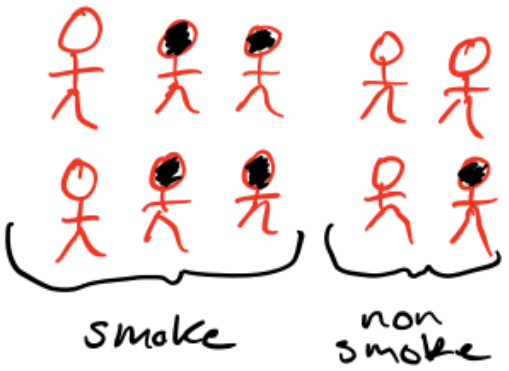
Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials

"We think that everyone might benefit if the most radical protagonists of evidence based medicine organised and participated in a double blind, randomised, placebo controlled, crossover trial of the parachute."

Adjusting for confounding



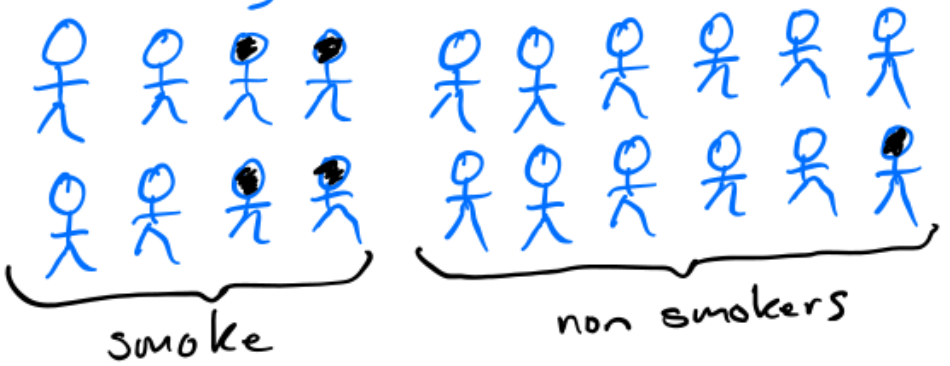
type A genetics
($\frac{1}{3}$ of people)



$P(\text{cancer} \mid \text{smoke}, A) = \frac{4}{6}$

$P(\text{cancer} \mid \text{no smoke}, A) = \frac{1}{4}$

type B genetics
($\frac{2}{3}$ of people)

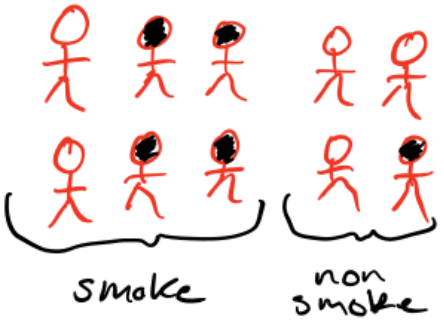


$P(\text{cancer} \mid \text{smoke}, B) = \frac{4}{8}$

$P(\text{cancer} \mid \text{no smoke}, B) = \frac{1}{12}$

Adjusting for confounding variables

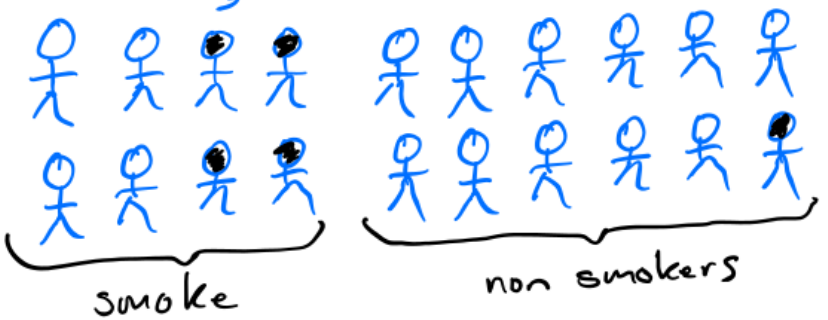
type A genetics
($\frac{1}{3}$ of people)



$P(\text{cancer} \mid \text{smoke}, A) = \frac{4}{6}$

$P(\text{cancer} \mid \text{no smoke}, A) = \frac{1}{4}$

type B genetics
($\frac{2}{3}$ of people)



$P(\text{cancer} \mid \text{smoke}, B) = \frac{4}{8}$

$P(\text{cancer} \mid \text{no smoke}, B) = \frac{1}{12}$

Adjusted:

$P(\text{cancer} \mid do(\text{smoke})) = \frac{4}{6} \cdot \frac{1}{3} + \frac{4}{8} \cdot \frac{2}{3} = .56$

$P(\text{cancer} \mid do(\text{not smoke})) = \frac{1}{4} \cdot \frac{1}{3} + \frac{1}{12} \cdot \frac{2}{3} = .14$, ratio = 4

Without Adjusting

$P(\text{cancer} \mid \text{smoke}) = \frac{8}{14}$

$P(\text{cancer} \mid \text{not smoke}) = \frac{2}{16}$, ratio = 4.6

Which variables do you adjust for?

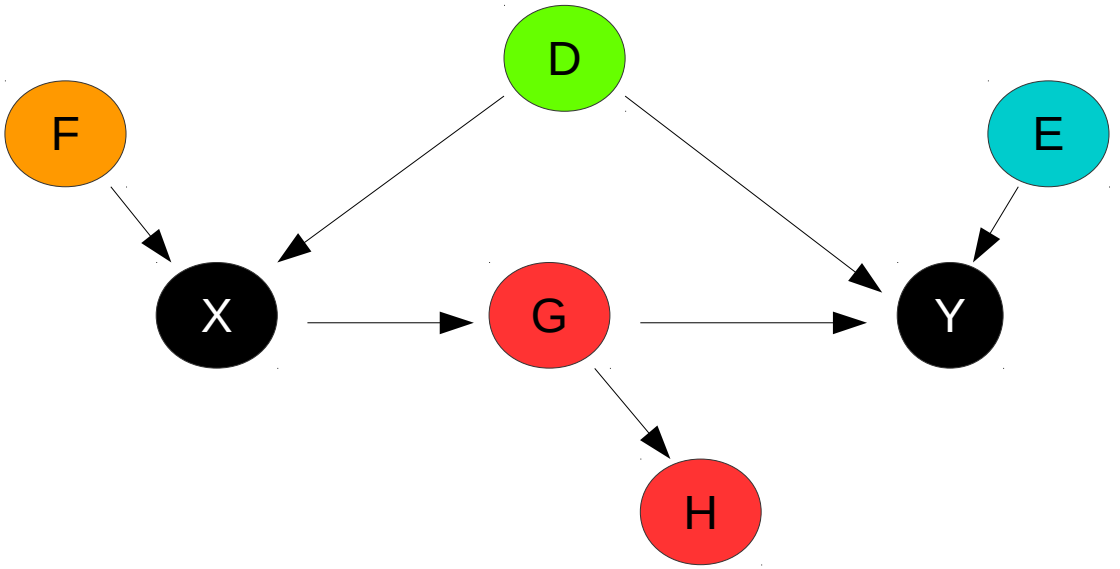
Should you adjust?

Yes definitely

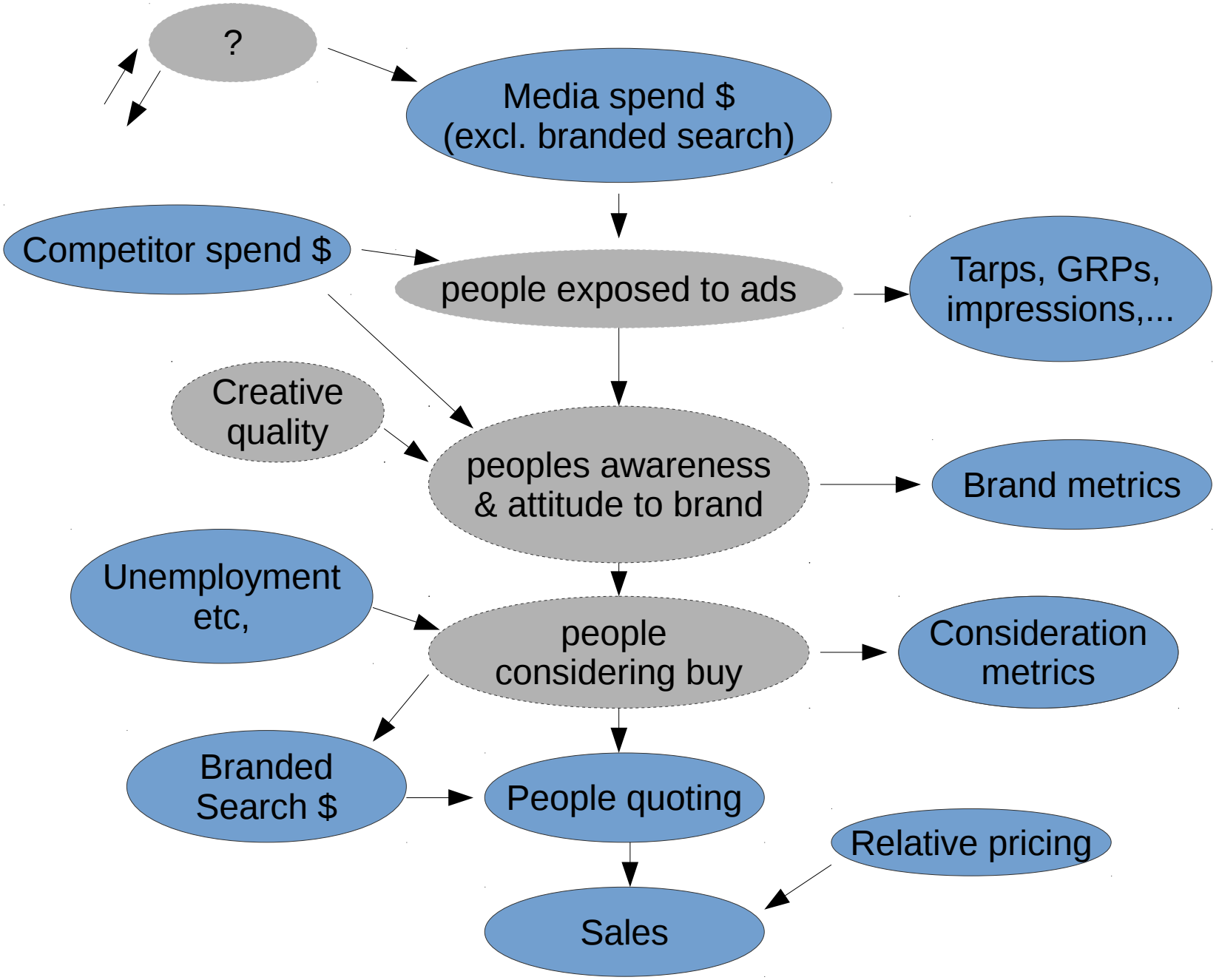
Probably

Probably not

Definitely not



Marketing mixed model example



Regression is causal inference?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

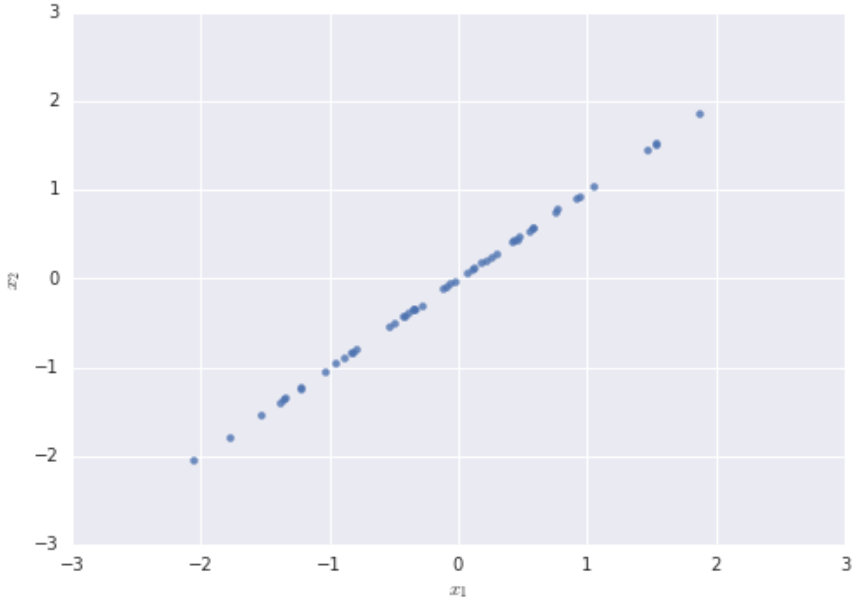
The coefficient β_1 says how much Y increases when increases by 1, holding all the other X variables fixed.

If the causal graph is such that the all the other variables is a valid set to adjust for, the regression coefficient(s) have a causal interpretation.

$$\mathbf{E}[Y|do(X_i = x + 1)] - \mathbf{E}[Y|do(X_i = x)] = \beta_i$$

Co-linearity

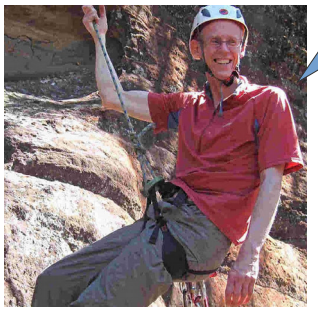
You are building a model, and two columns are highly correlated. Does that matter?



Machine learner

No

Economist



Yes, it can be a huge problem

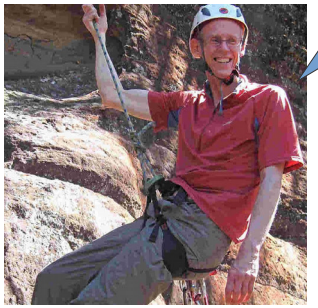
Co-linearity

You are building a model, and two columns are highly correlated. Does that matter?

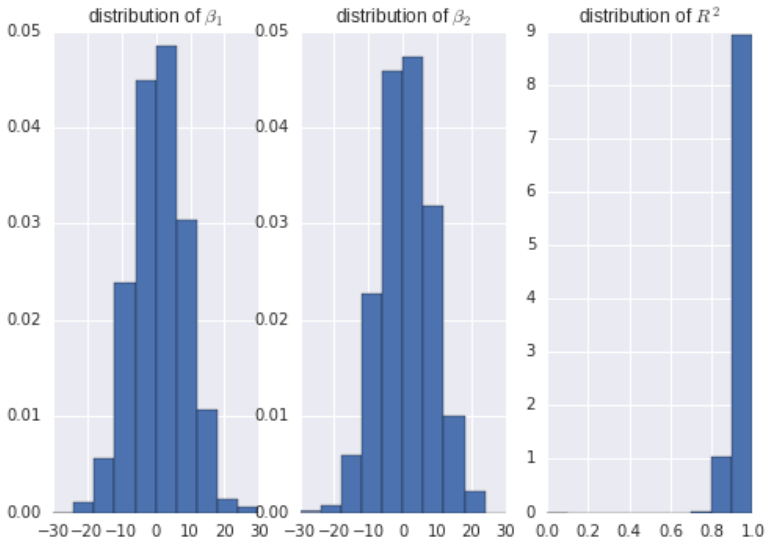
Machine learner

No

Economist



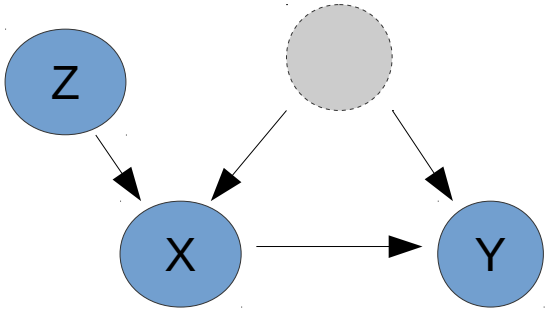
Yes, it can be a huge problem



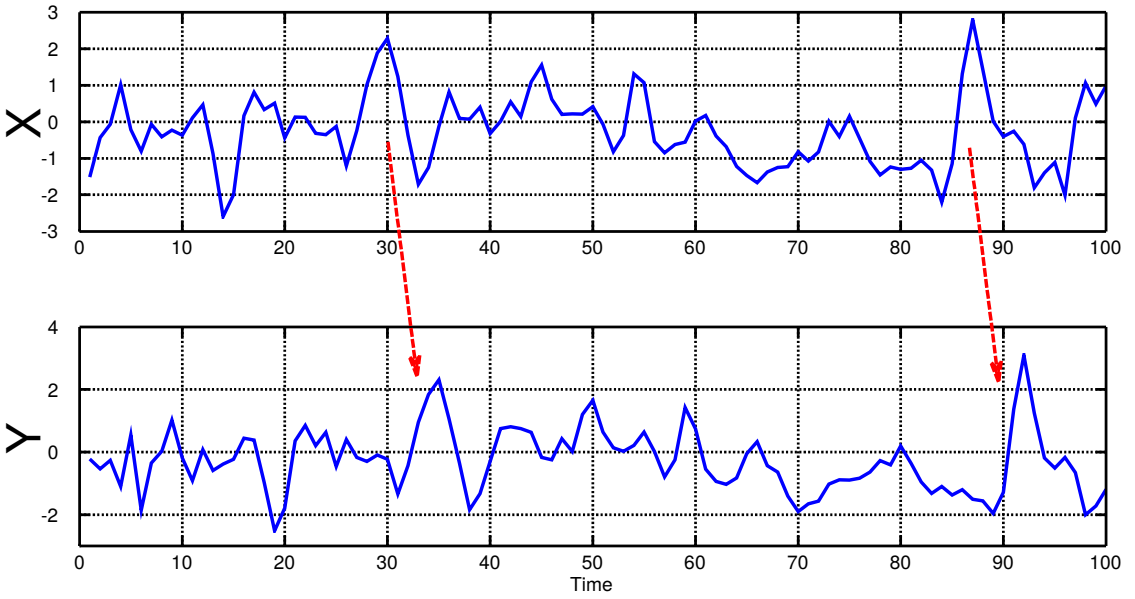
$$y = x_1 + x_2 + \epsilon \sim N(0, .5)$$

A couple of economist's tricks

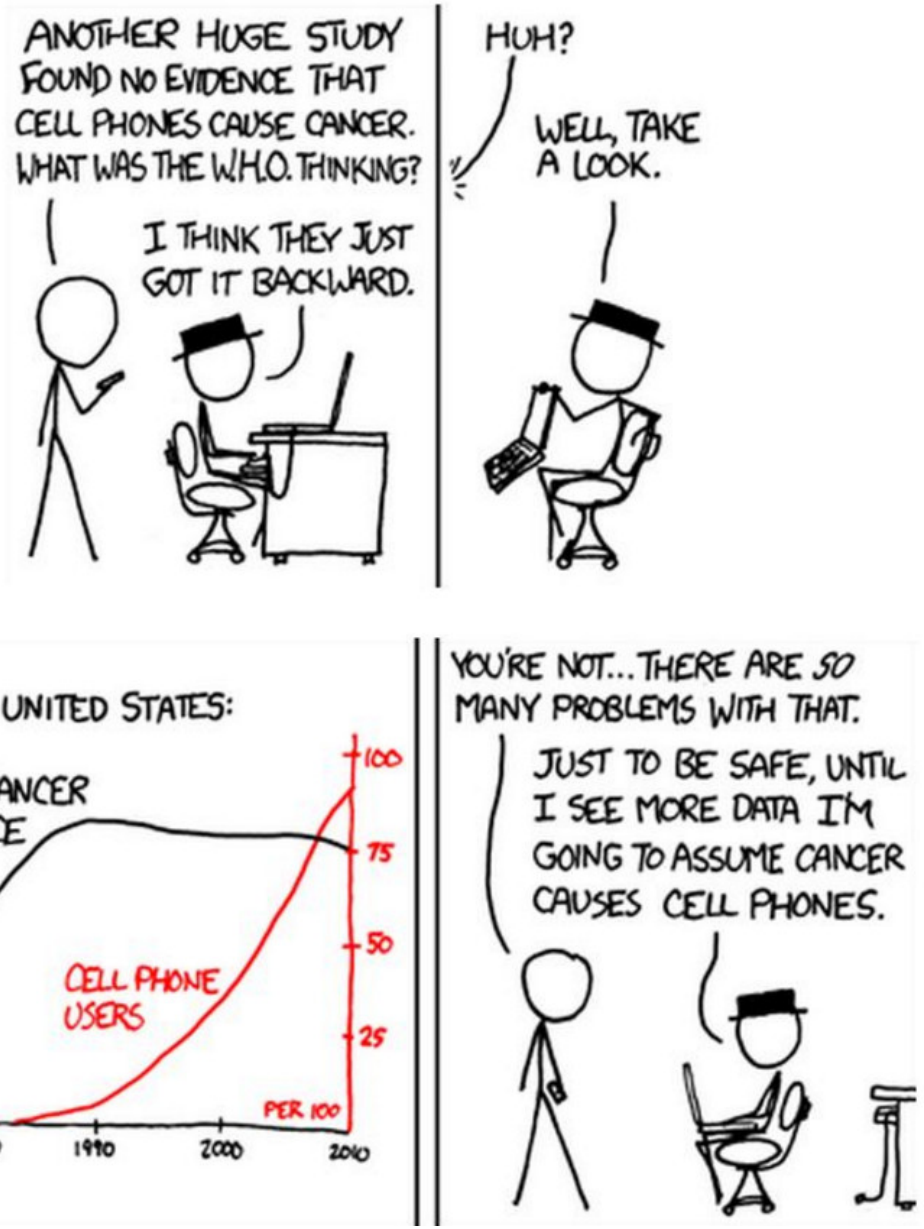
Instrumental variables



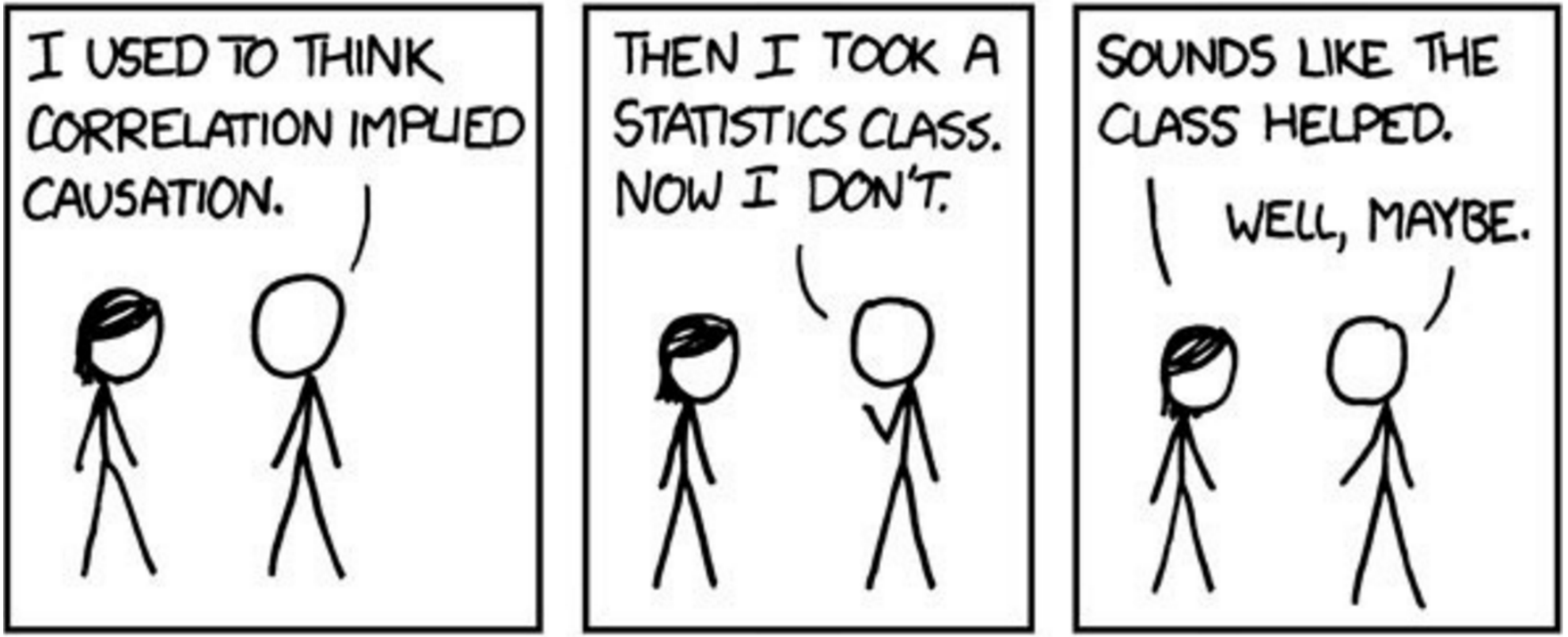
Granger causality



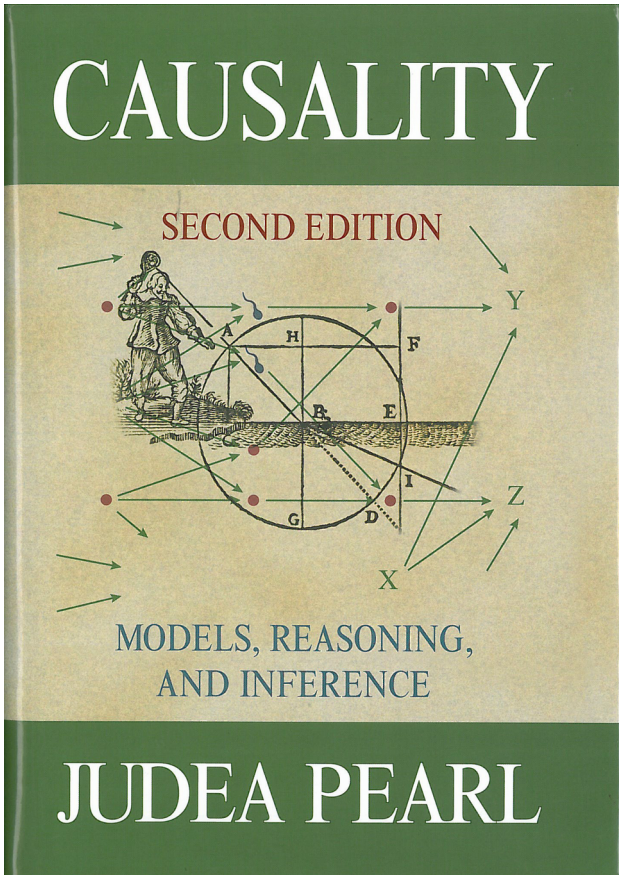
Granger causality

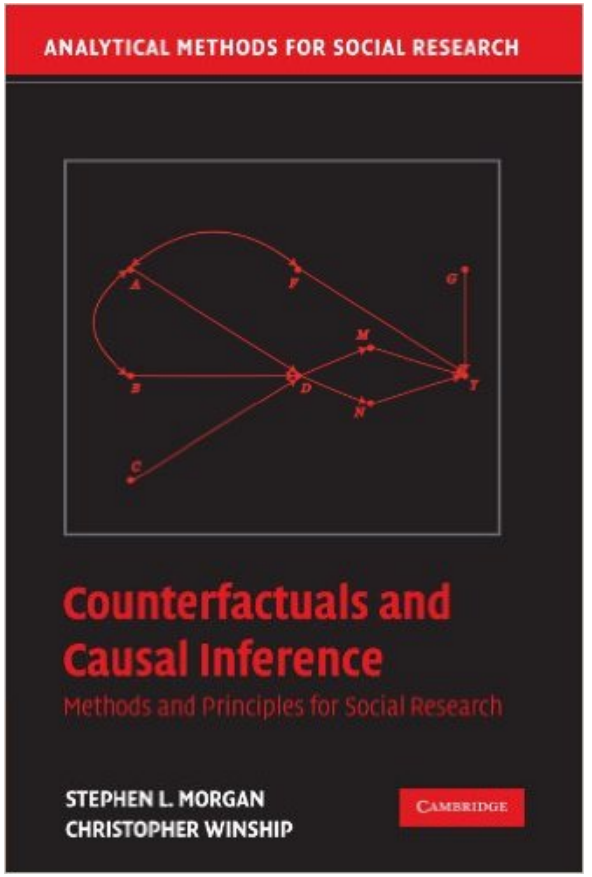


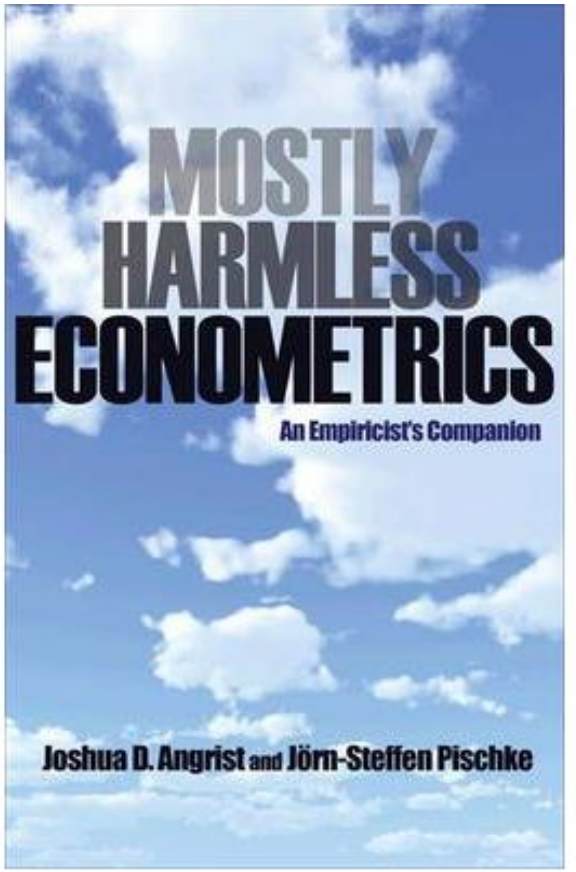
To conclude ...


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

That wasn't enough?







What if you don’t know the graph?

This is called causal discovery. Amazingly, you can still do something. There are two key approaches

1) Finding graphs that have the imply the same conditional independences that you observe in the distribution of your data

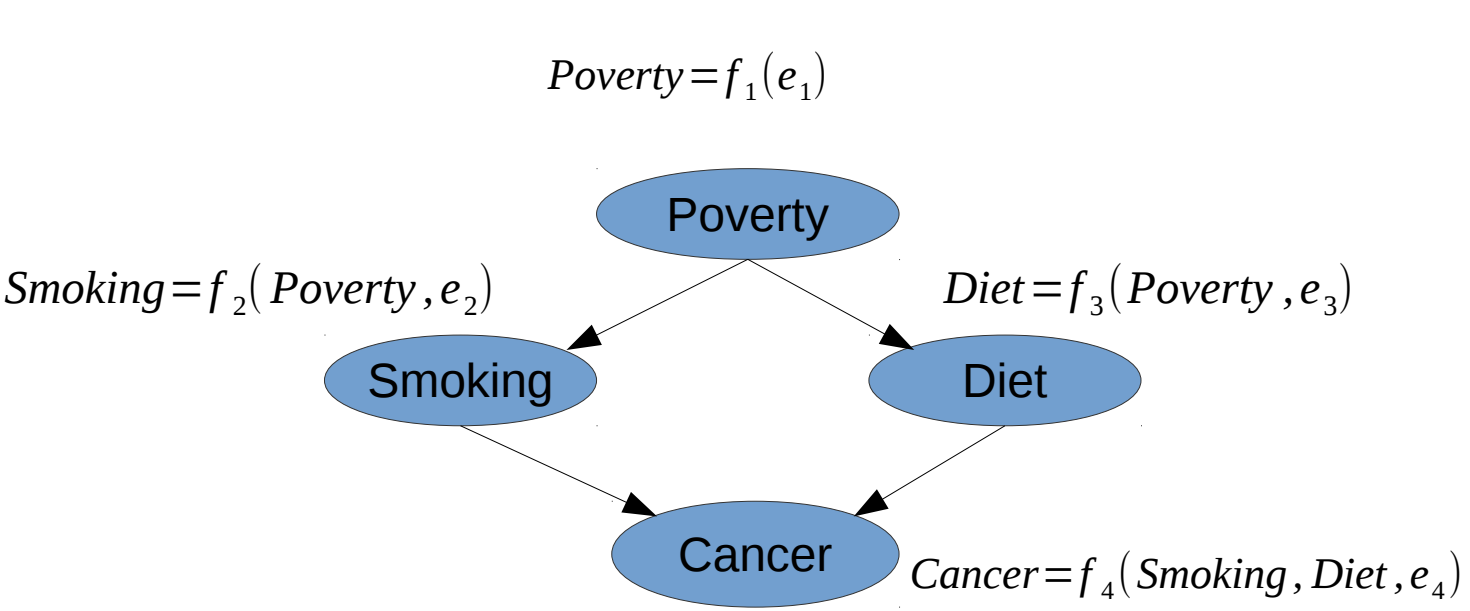
2) Looking for structures such that the residuals are consistent with independence of mechanism and input.

It is really, really hard to get good results (or know if you have got good results)

Structural equation models (SEMs)

- Represent each variable as a deterministic function of its direct causes and noise
- Noise terms must be mutually independent

$$x_i = f_i(Parents_i, e_i) \quad \text{where} \quad \{f_1 \ldots f_n\}\ deterministic\ functions$$
$$\{e_1 \ldots e_n\}\ mutually\ independent$$



If the set of equations does not create a cycle then the Causal Markov property holds and the SEM is a causal bayesian network

Counterfactuals

- Statements about what would have happened in some alternate reality where some specified thing were different.
- A medical trial:

  for an individual, i, $\begin{cases} y_i^1 = outcome\ if\ treated \\ y_i^0 = outcome\ if\ not\ treated \end{cases}$

  we only get to measure one of these.
- Let Y^1 be a random variable, $P(Y^1)$ is the distribution of outcome, Y that would occur if everyone was treated.
- The causal effect is defined as $E[P(Y^1) - P(Y^0)]$
- We can measure $\begin{cases} P(Y|X=0) = P(Y^0|X=0) \\ P(Y|X=1) = P(Y^1|X=1) \end{cases}$

$If\ (X \perp Y^0)\ \&\ (X \perp Y^1)$ ← Ignoreability assumption

$$E[P(Y^1) - P(Y^0)] = E[P(Y|X=1)] - E[P(Y|X=0)]$$

Counterfactuals

Counterfactural queries form a larger set than interventional queries

group	placebo	treatment	probability of group
1	die	die	$\alpha = P(Y^0 = 0, Y^1 = 0)$
2	die	recover	$\beta = P(Y^0 = 0, Y^1 = 1)$
3	recover	die	$\gamma = P(Y^0 = 1, Y^1 = 0)$
4	recover	recover	$\delta = P(Y^0 = 1, Y^1 = 1)$

What is the probability that a patient, who was not treated and died, would have recovered if they had been treated? We know they are in either group 1 or 2 since they died without treatment, so the answer is $\frac{\beta}{\alpha+\beta}$

Can we identify that from $P(Y^0)$ and $P(Y^1)$?

How do the models relate?

- Do type queries can be phrased in terms of counterfactuals

$$Y^0 \equiv P(Y|do(X=0))$$
$$Y^1 \equiv P(Y|do(X=1))$$

- For do type queries, the ignoreability assumption is true if the 'back door criterion' holds.
- Causal bayesian networks don't support full counterfactuals as they only contain information on interventional *distributions*
- SEMs do define counterfactuals $Y = f(X, e) \Rightarrow Y^0 = f(0, e)$
- There is a subtle difference between the assumptions implied by the SEM models and what is actually needed for identification of do-queries (see Richardson & Robins 2013). This makes a difference for identifiability of non-interventional queries.

The Do Calculus: calculating causal effects in a (partially) known network

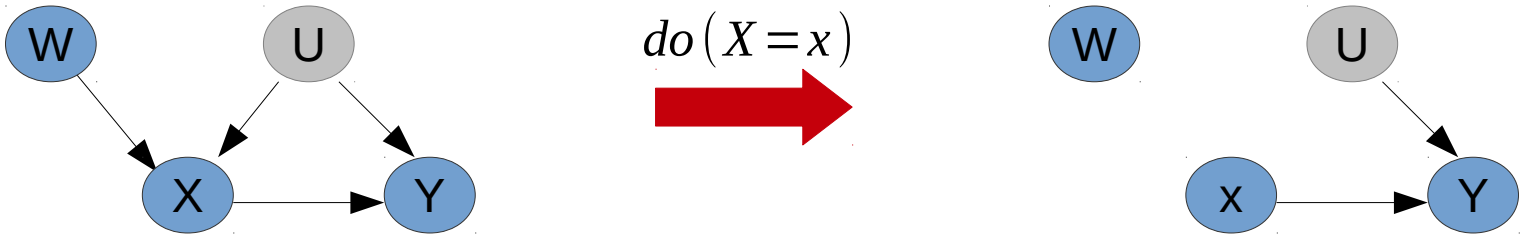
- The do-calculus rules result from d-separation in a causal DAG
- A causal effect is non-parametrically identifiable if and only if the interventional query can be reduced to an observational one via repeat application of the three rules (see Shpitser&Pearl 2012 for algorithm)

Rule 1: D-separation still applies

The graph $G_{\overline{X}}$ that results from the intervention do(X=x) is still a bayesian network and d-separation applies.

$$(Y \perp W | X)_{G_{\overline{X}}} \Rightarrow P(y|do(x), w) = P(y|do(x))$$

$$(Y \perp W | X, Z)_{G_{\overline{X}}} \Rightarrow P(y|do(x), z, w) = P(y|do(x), z)$$



Rule 2: intervention and observation

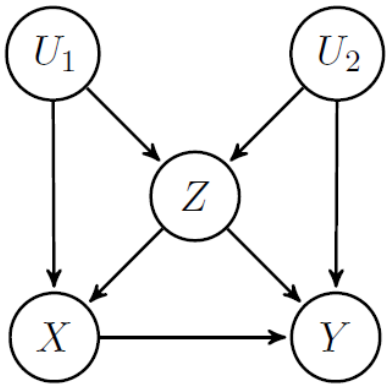
If the target, y, is independent of *how* x is determined intervention is the same as observation.

$$\left(Y \perp \hat{X} \middle| X\right)_{G^+} \Rightarrow P\left(y \middle| do\left(x\right)\right) = P\left(y \middle| x\right)$$

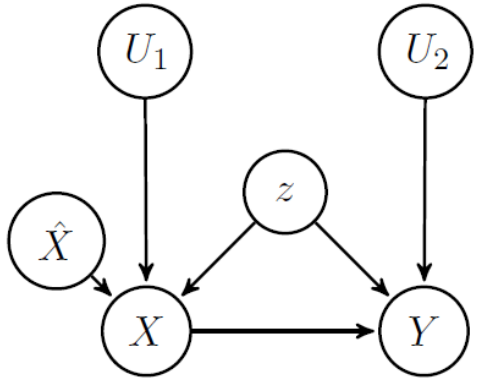
$$\left(Y \perp \hat{X} \middle| X, Z, W\right)_{G^+_{\overline{Z}}}$$

$$\Rightarrow P\left(y \middle| do\left(z\right), do\left(x\right), w\right) = P\left(y \middle| do\left(z\right), x, w\right)$$

(a) original network, G



(b) augmented network after the intervention $do(Z = z)$, $G^{\dagger}_{\overline{Z}}$



Rule 3: Sometimes intervention changes nothing

$$\left(Y \perp \hat{X}\right)_{G^+}$$

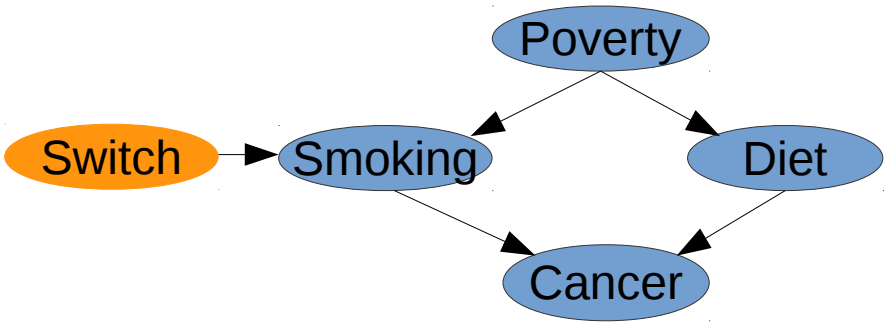
$$\Rightarrow P\left(y|do\left(x\right)\right)=P\left(y\right)$$

$$\left(Y \perp \hat{X}|,Z,W\right)_{G^+_{\bar{Z}}}$$

$$\Rightarrow P\left(y|do\left(z\right),do\left(x\right),w\right)=P\left(y|do\left(z\right),w\right)$$

Smoking switch is d-separated from diet. There is no direct causal path from smoking to diet, so the intervention doesn't change anything.

$$P\left(Diet|do\left(Smoke=N\right)\right)=P\left(Diet\right)$$



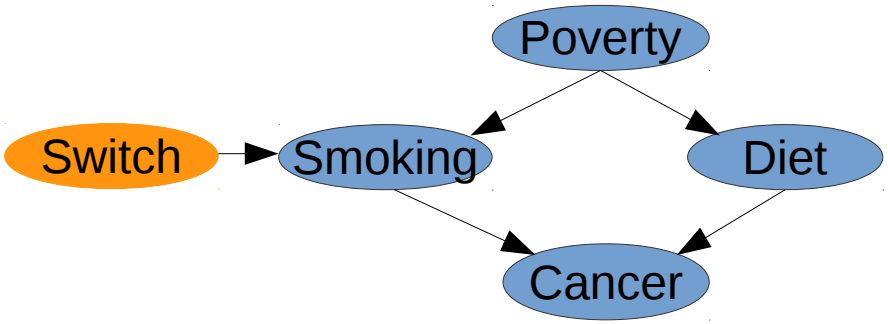
The back door criterion

Just a special case of rule 2. If Z blocks all back door paths from X to Y then the causal effect of X on Y is obtained by adjusting for Z.

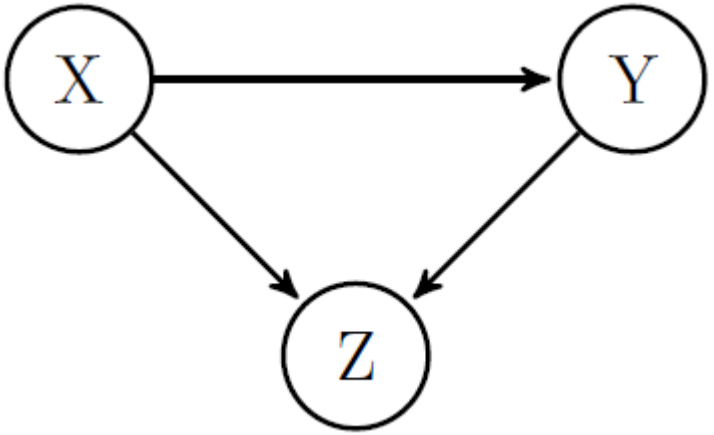
$$\left(Y \perp \hat{X} \middle| X, Z\right)_{G^+}$$

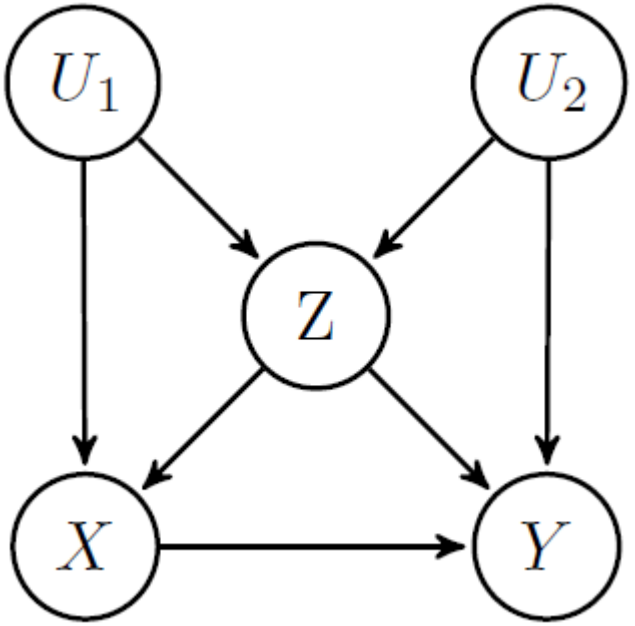
$$\Rightarrow P\left(y \middle| do\left(x\right), z\right) = P\left(y \middle|, x, z\right)$$

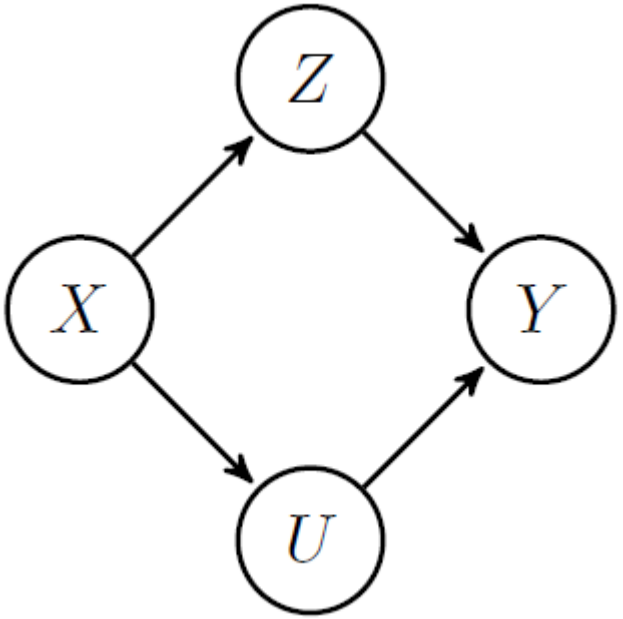
$$\Rightarrow P\left(y \middle| do\left(x\right)\right) = \sum_z P\left(y \middle|, x, z\right) P\left(z\right)$$

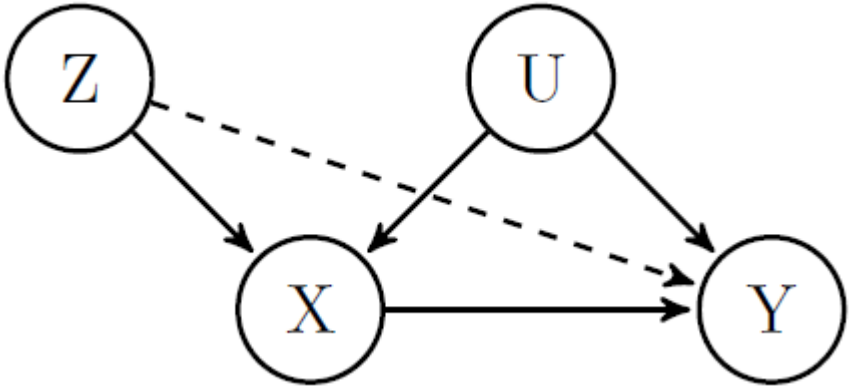


Causal estimation by 'adjusting'









Causal Discovery

Independence based methods

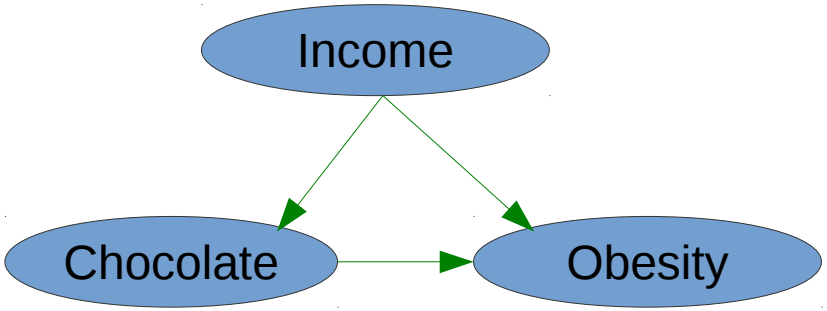
1) We assume our distribution P was generated by some (unknown) causal DAG over our observed variables (causal sufficiency)

2) We assume that all the conditional independences in P are implied by d-separation in the true causal network (***faithfulness***)

3) Finding the causal structure equates to finding perfect maps for P

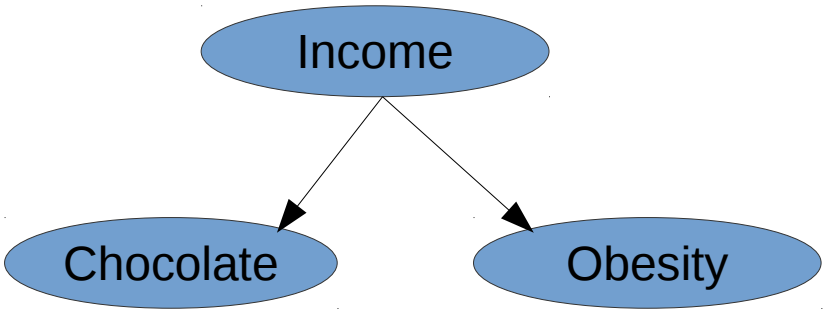
True causal structure generating P

I	High	Low
P(I)	0.5	0.5



I	High		Low	
C	Y	N	Y	N
P(O=Y)	.4	.2	.7	.5

Perfect map for P



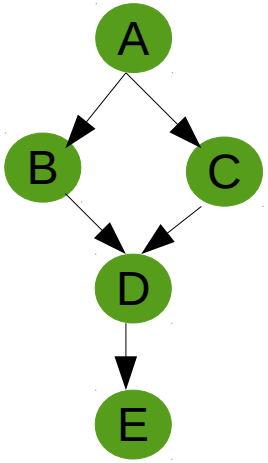
The IC (or SGS) algorithm

1. *for all pairs of variables* a, b *search for a set* S_{ab} *such that* $a \perp b | S_{ab}$.
   If there is no such set, then draw an undirected link between them.

2. *for all pairs of non−linked nodes with a common neighbour,* c*,:*
   If $c \notin S_{ab}$ *direct links towards* c

3. *Orient any undirected edges so as to avoid creating cycles or additional v-structures*

True Causal Model

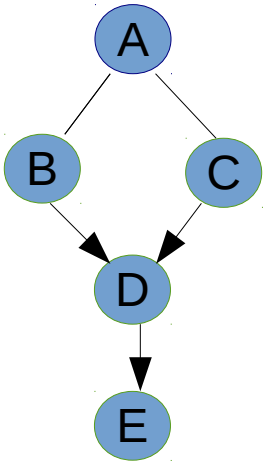


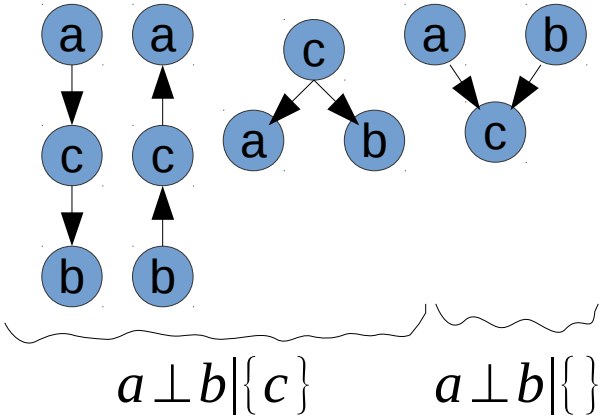
$B \perp C | A$

$D \perp A | B, C$

$E \perp A, B, C | D$

Inferred output





$a \perp b | \{c\}$ $a \perp b | \{\}$

The PC algorithm

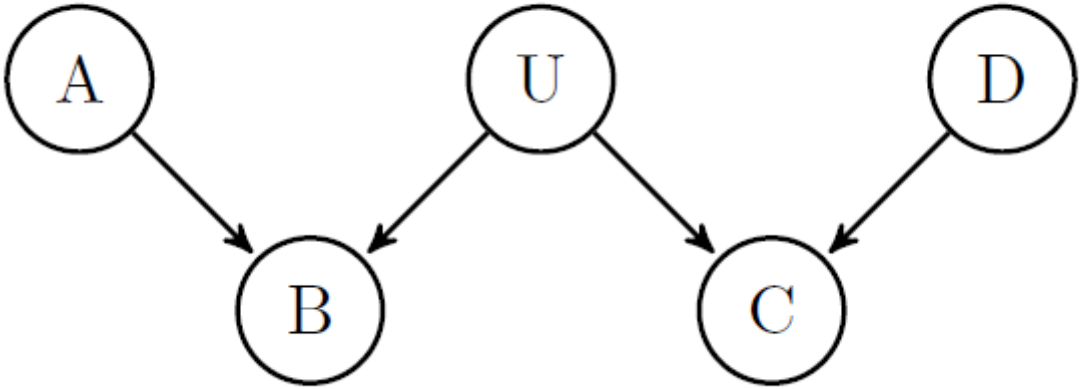
A more tractable version of the SGS algorithm

2. **for** each link $a - b$:
   $n = 0$
   $\boldsymbol{A}_{a,b} = \{A_1 ... A_j\}$ be the set of nodes adjacent to a and/or b
   while a and b are connected and $n < j$:
      if any subset of size n of $\boldsymbol{A}$ makes a and b conditionally independent:
         delete the link
      $n = n + 1$

Latent variables

Problems:

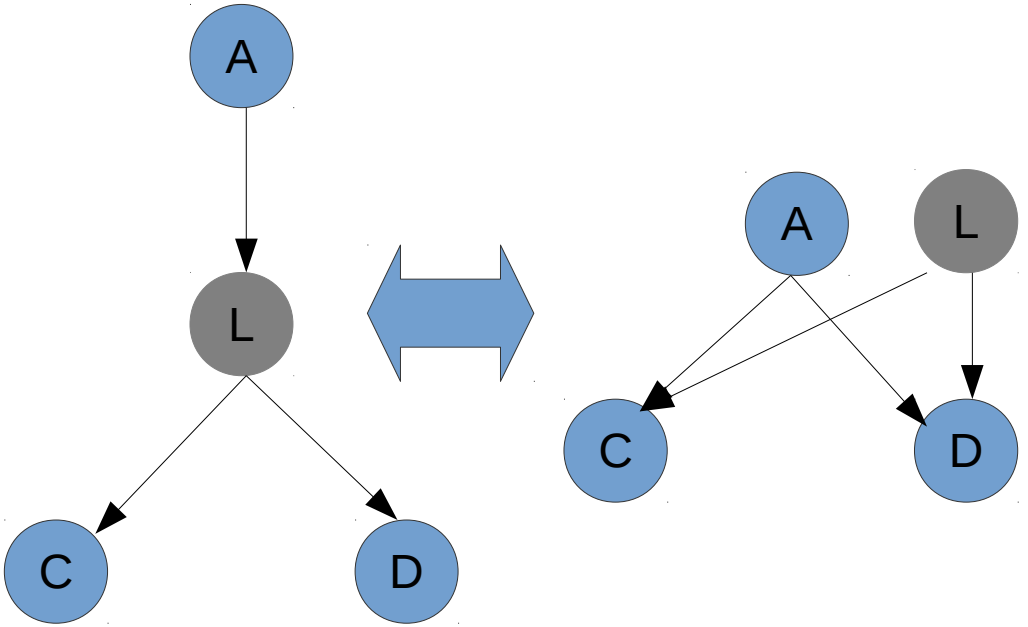
- Infinite search space
- The space of causal DAGs is not closed under marginalization



Latent variables

Theorem (Verma 1993):

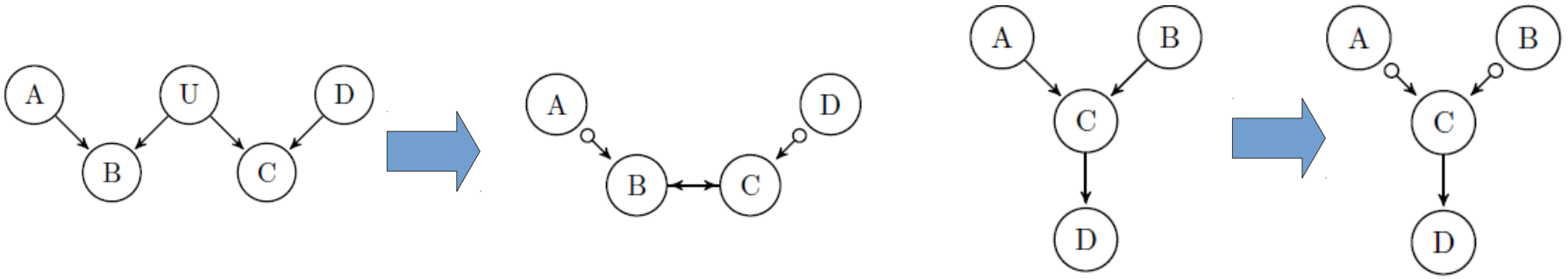
for any latent structure there is an equivalent structure such that every latent variable is a root node with exactly 2 children.



FCI algorithm

Outputs a Markov equivalence class of Maximal Ancestral Graphs

1. $X \rightarrow Y$, meaning X causes Y
2. $X \leftrightarrow Y$, meaning there is a latent variable that causes X and Y.
3. $X \circ\!\!\rightarrow Y$, either X causes Y or a latent variable causes both.
4. $X \circ\!\!-\!\!\circ Y$, either X causes Y or Y causes X or a latent variable causes both.



- Discovers all aspects of causal structure identifiable from conditional independence relations (Zhang 2008)
- Can be made to require a worst case polynomial (rather than exponential) number of conditional independence tests for sparse graphs (Claassen et al 2013).

Beyond conditional independence

Additive noise: $y = f(x) + e$

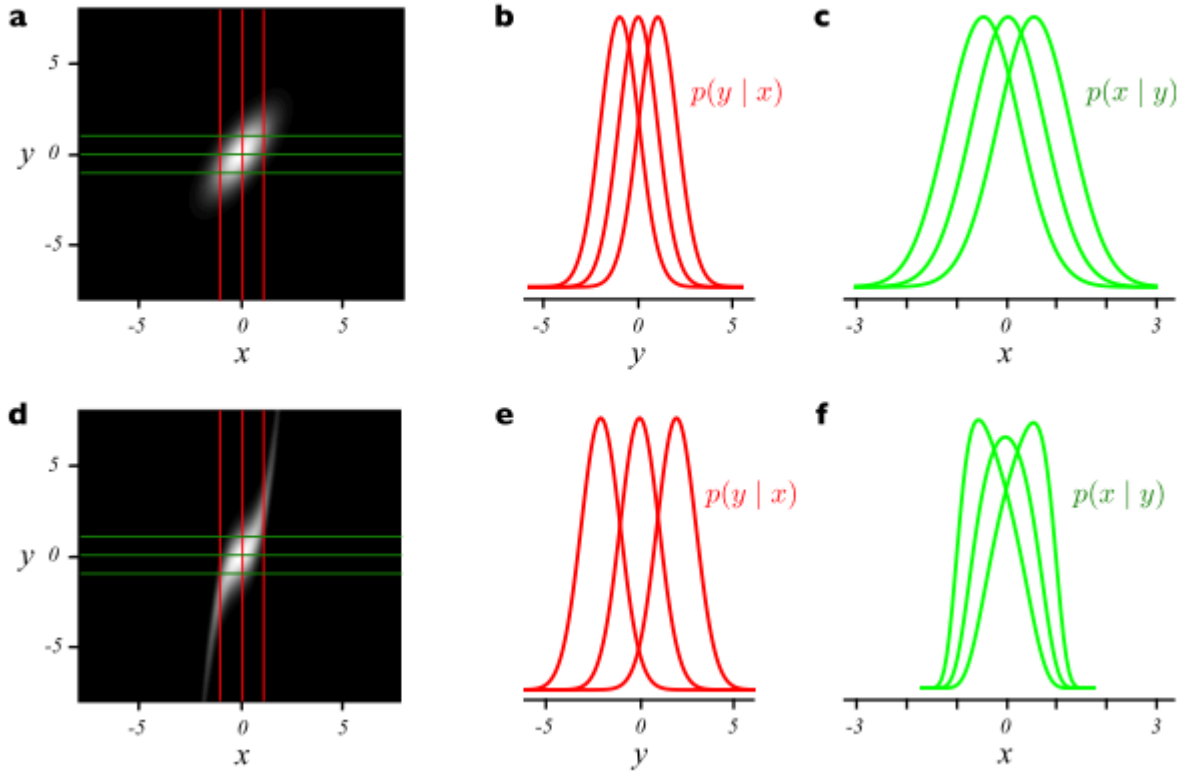


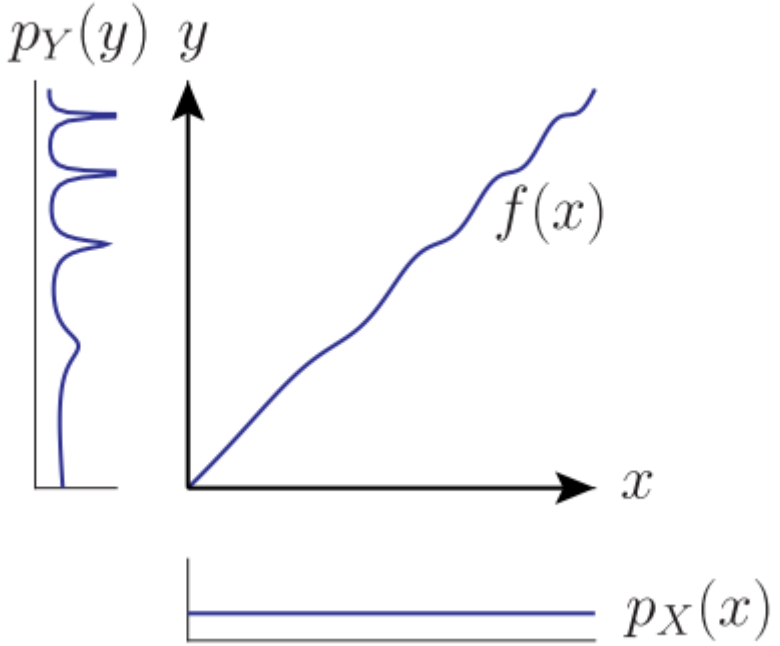
Figure 1, (Hoyer et al 2009)

Can be extended to post-non-linear additive noise $y = h(f(x) + e)$

(Zhang et al 2009)

More asymmetries of cause and effect

Figure 1: Daniusis et al 2012



Independence of function and input:
If X → Y and we have a functional causal model y = f(x,e) then the input distribution P(X) and function f represent independent mechanisms. Changing the input distribution does not modify the function itself.

Causal-Anticausal

- If X → Y, what is the relationship between P(X) and P(Y|X)?
- What about P(Y) and P(X|Y)?

In semi supervised learning we are given training points from P(X,Y) and additional points sampled from P(X). The goal is to learn P(Y|X). The extra points from P(X) will not help if X → Y.

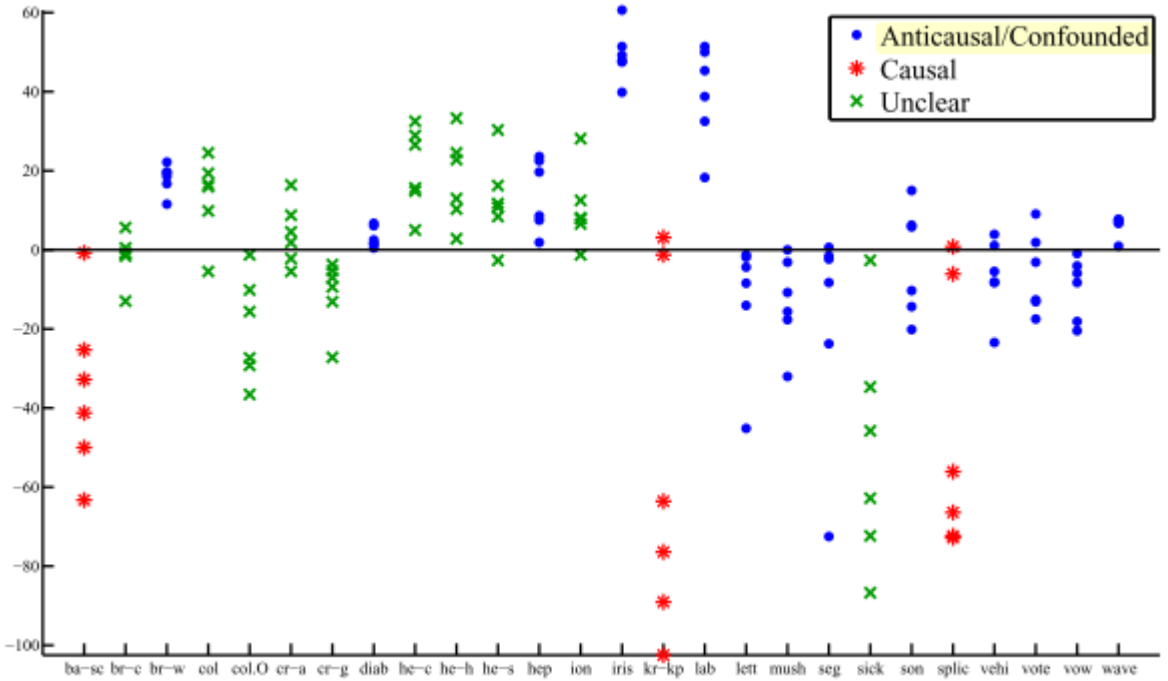


Figure 6, Janzing & Peters 2012

Learning what causality looks like

Suppose we had M different causal pairs data sets.

$$D = \{\{x_j, y_j\}_{j=1}^{N_i}, l_i\}_{i=1}^{M}$$

Where l_i is a binary label that indicates if $X \to Y$ or $Y \to X$ in dataset i.

We expect there to be differences in the relationships between $P(X)$ $P(Y)$ and $P(Y|X)$ for $X \to Y$ and $Y \to X$

Let μ be a kernel mean embedding that maps a distribution P into some Hilbert space.

For each data set $i = 1...M$

Construct a feature vector that approximates $\mu(P(X)), \mu(P(Y)), \mu(P(X, Y))$

Apply a standard classification algorithm

See Lopez-Paz et al 2014

Now with more than two variables!

Bi-variate additive noise model $Y = f(X) + e$

If the triple: $(P(X), P(e), f)$ satisfies a certain condition then model is identifiable

To extend to multivariate case we need to satisfy the condition on the conditional distributions

For each variable j

Pick a single parent, i, and fix all the others

For each subset S that contains the remaining parents, and no descendants of j

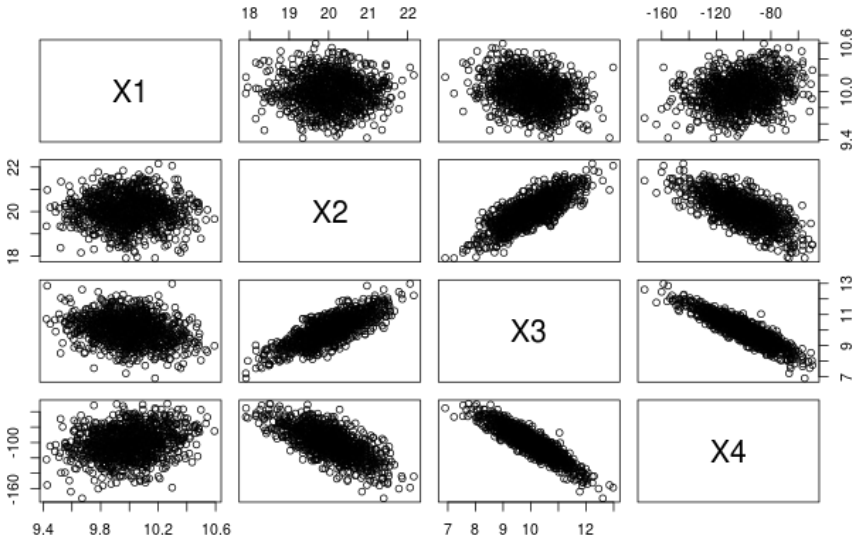
The triple $(P(X|S=s), P(e_j), f^{ji}(X_i))$ must satisfy the condition for at least one s.

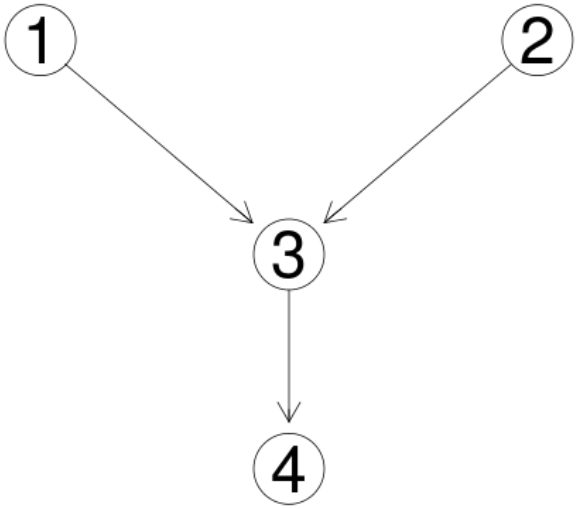
If we can come up with a condition that guarantees identifiability for the bi-variate case, we can extend that result to get the conditions under which the multivariate case is identifiable.

See Peters et al 2014

Causal structure learning in R (pcalg)

```
library('pcalg')
n = 1000
X1 = rnorm(n,mean=10,sd=.2)
X2 = rnorm(n,mean=20,sd=.7)
X3 = X2-X1+rnorm(n,mean=0,sd=.5)
X4 = -X3^2+rnorm(n,mean=0,sd=8)
df = data.frame(X1,X2,X3,X4)
plot(df)
suffStat <- list(C = cor(df),n=nrow(df))
pc.3var = pc(suffStat,indepTest=gaussCItest,p=ncol(df),alpha=0.01)
plot(pc.3var, main = "")
```





References

Pearl, J. (2000). *Causality: models, reasoning and inference*

Tom Claassen, J Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. arXiv Prepr. ArXiv1309.6824, 2013.

PO Hoyer, Dominik Janzing, and JM Mooij. Nonlinear causal discovery with additive noise models. Adv. Neural . . . , 2009.

Kun Zhang and A Hyv¨arinen. On the identiability of the post-nonlinear causal model.Proc. Twenty-Fifth Conf. . . . , 2009.

P Daniusis, Dominik Janzing, and Joris Mooij. Inferring deterministic causal relations.arXiv Prepr. arXiv . . . , pages 2-9, 2012.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artif. Intell., 172(16-17):1873{1896, November 2008

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. (chapters 3 & 21)

Verma 1993 *Graphical aspects of causal models Technical* Report. UCLA

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*.

Maathuis, Marloes H., et al. (2010) *Predicting causal effects in large-scale systems from observational data.* Nature Methods 7.4 : 247-248.

Kalisch, Markus, et al. (2012) Causal inference using graphical models with the R package pcalg. Journal of Statistical Software 47.11 : 1-26.

Shpitser, Ilya, and Judea Pearl. "Identification of conditional interventional distributions." arXiv preprint arXiv:1206.6876 (2012).

Dominik Janzing and Jonas Peters. On causal and anticausal learning JMLR. , 2012.

David Lopez-Paz, Krikamol Muandet, and Benjamin Recht. The Randomized Causation Coefficient. September 2014.

TS Richardson and JM Robins. Single world intervention graphs (SWIGs): a unication of the counterfactual and graphical approaches to causality. Cent. Stat. . . . , (128), 2013.

Jonas Peters, J Mooij, Dominik Janzing, and B Sch¨olkopf. Causal discovery with continuous additive noise models. J. Mach. Learn. Res. 2014.