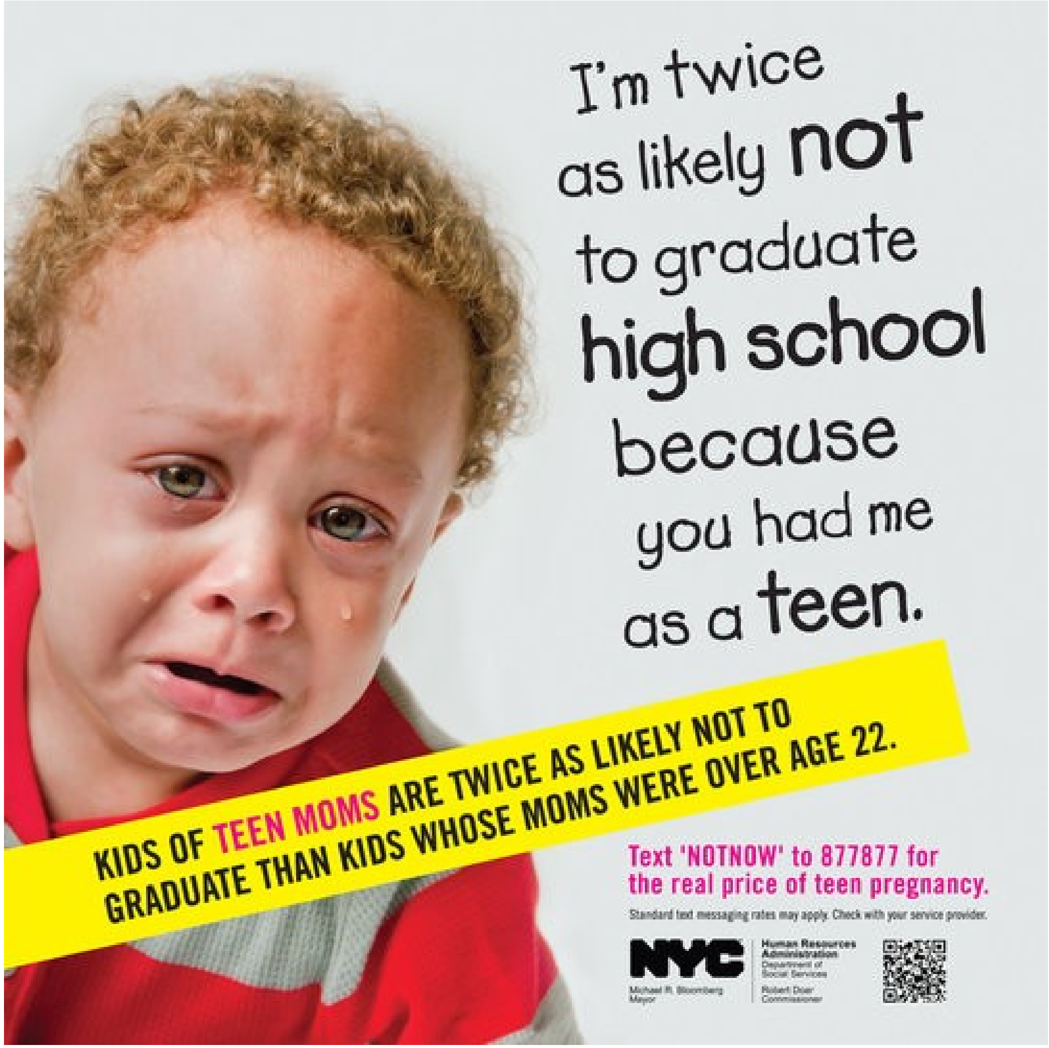
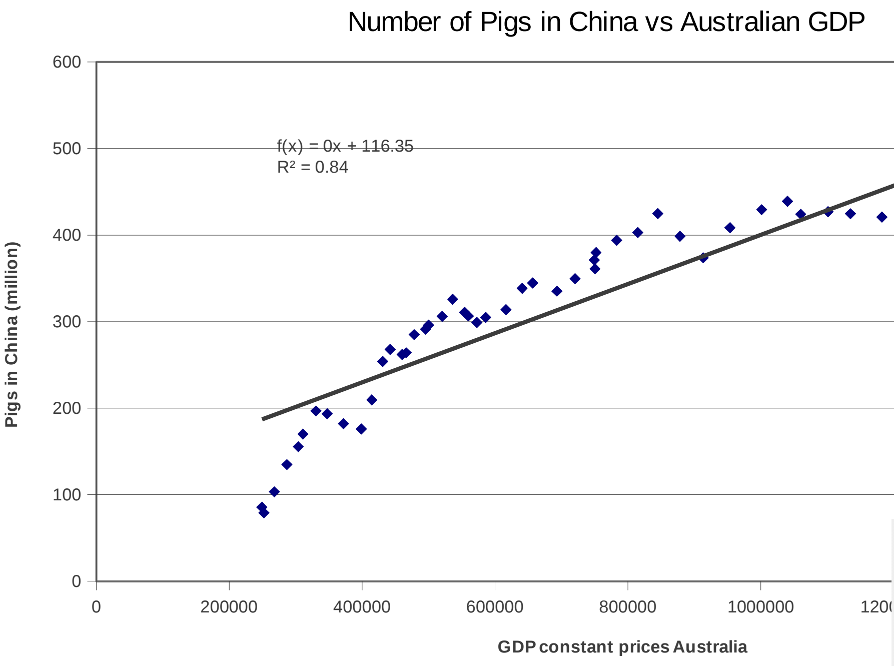
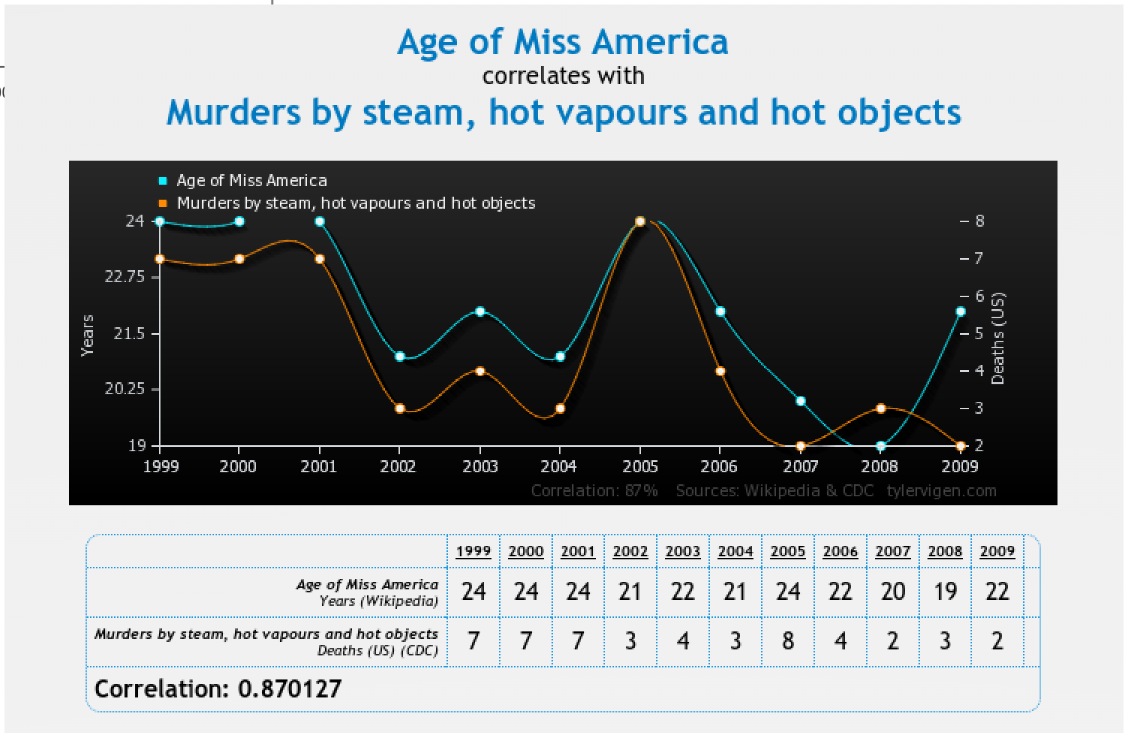
Causal Inference in Machine Learning


I'm twice as likely not to graduate high school because you had me as a teen.
KIDS OF TEEN MOMS ARE TWICE AS LIKELY NOT TO GRADUATE THAN KIDS WHOSE MOMS WERE OVER AGE 22.
Text 'NOTNOW' to 877877 for the real price of teen pregnancy.
Standard text messaging rates may apply. Check with your service provider.
NYC
Michael R. Bloomberg
Mayor
Human Resources Administration
Department of Social Services
Robert Doar
Commissioner

Finnian Lattimore (finnlattimore@gmail.com)

Ways things can go wrong





Age of Miss America

correlates with

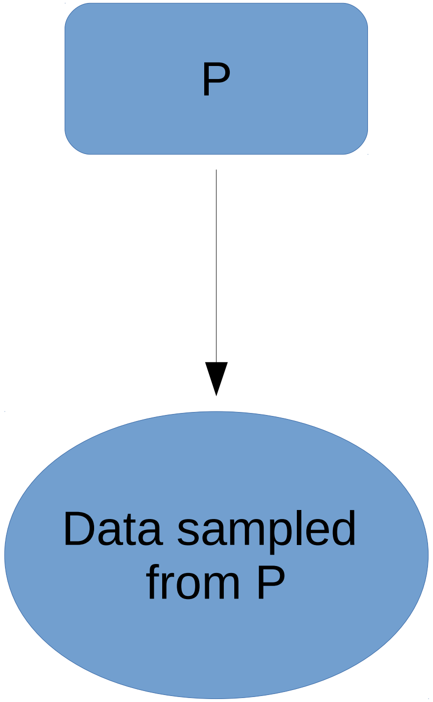
Murders by steam, hot vapours and hot objects

	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

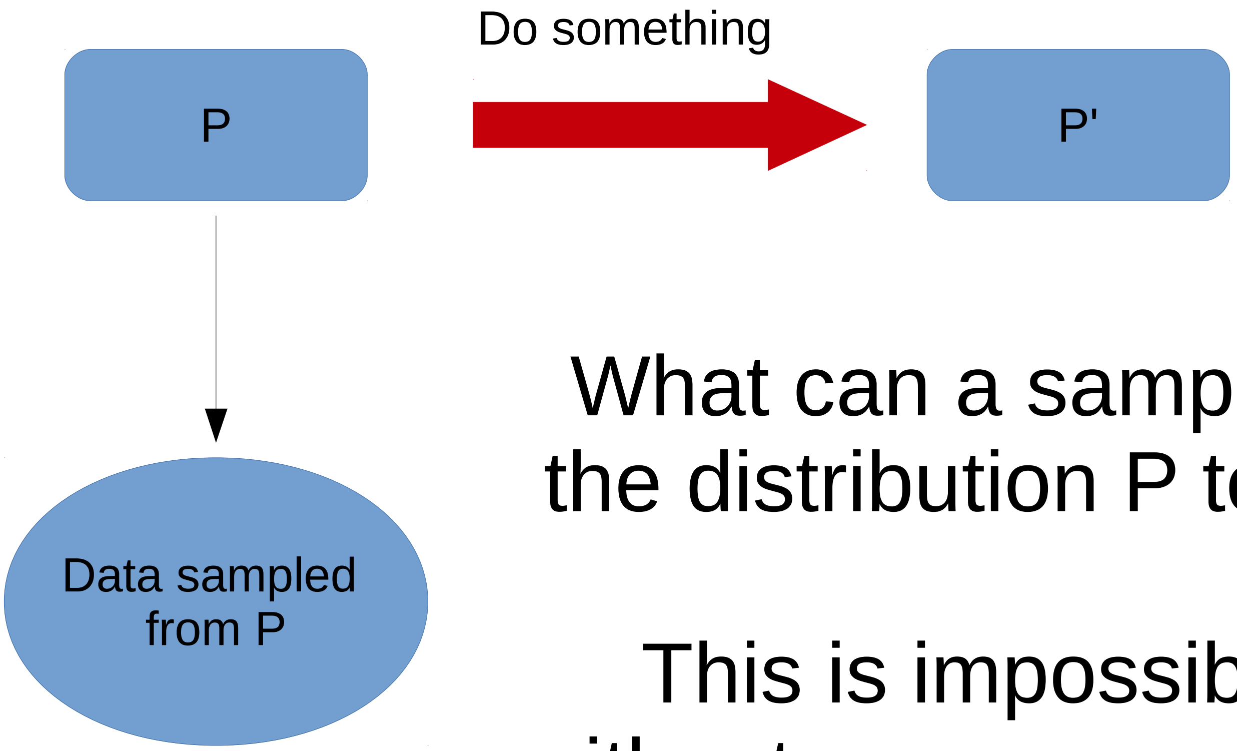
Image source: www.tylervigen.com/

Machine Learning/Statistics



What can we learn about the distribution P from a sample of data drawn from it?

Causal inference



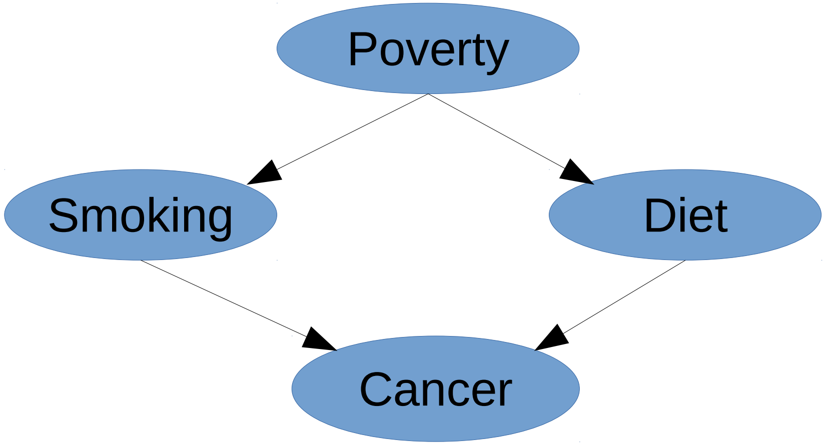
What can a sample of data from the distribution P tell us about P'?

This is impossible to answer without some assumptions on how 'do something' changes P

Causal bayesian networks (causal DAGs)

A bayesian network where A → B is defined to mean A causes B

=> Variables are independent of their non-effects given their direct causes (Causal Markov Property)



Absent links imply the factorisation of the full distribution can be simplified.

$$P(Po,S,D,C)=P(Po)P(S|Po)P(D|Po,S)P(C|Po,S,D)=P(Po)P(S|Po)P(D|Po)P(C|S,D)$$

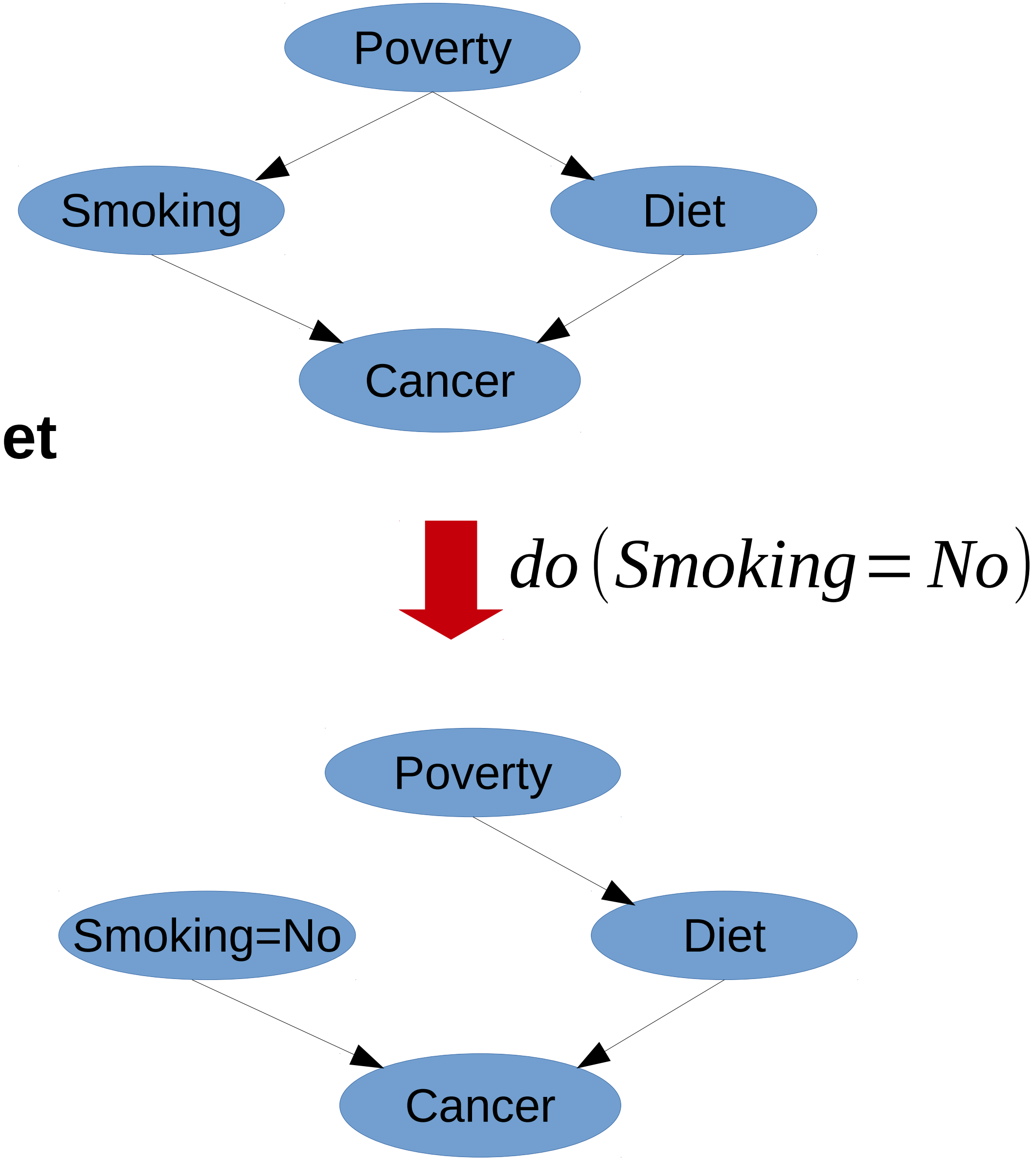
Intervention in Causal DAGs

Truncated product formula

Drop from terms for intervened on variables from the factorization

A causal DAG represents the set of all possible interventional distributions over its variables

$$P(Po,S,D,C)=P(Po)P(S|Po)P(D|Po)P(C|S,D)$$



$$do(Smoking=No)$$

$$P(Po,D,C|do(S=no))=P(Po)P(D|Po)P(C|S=no,D)$$

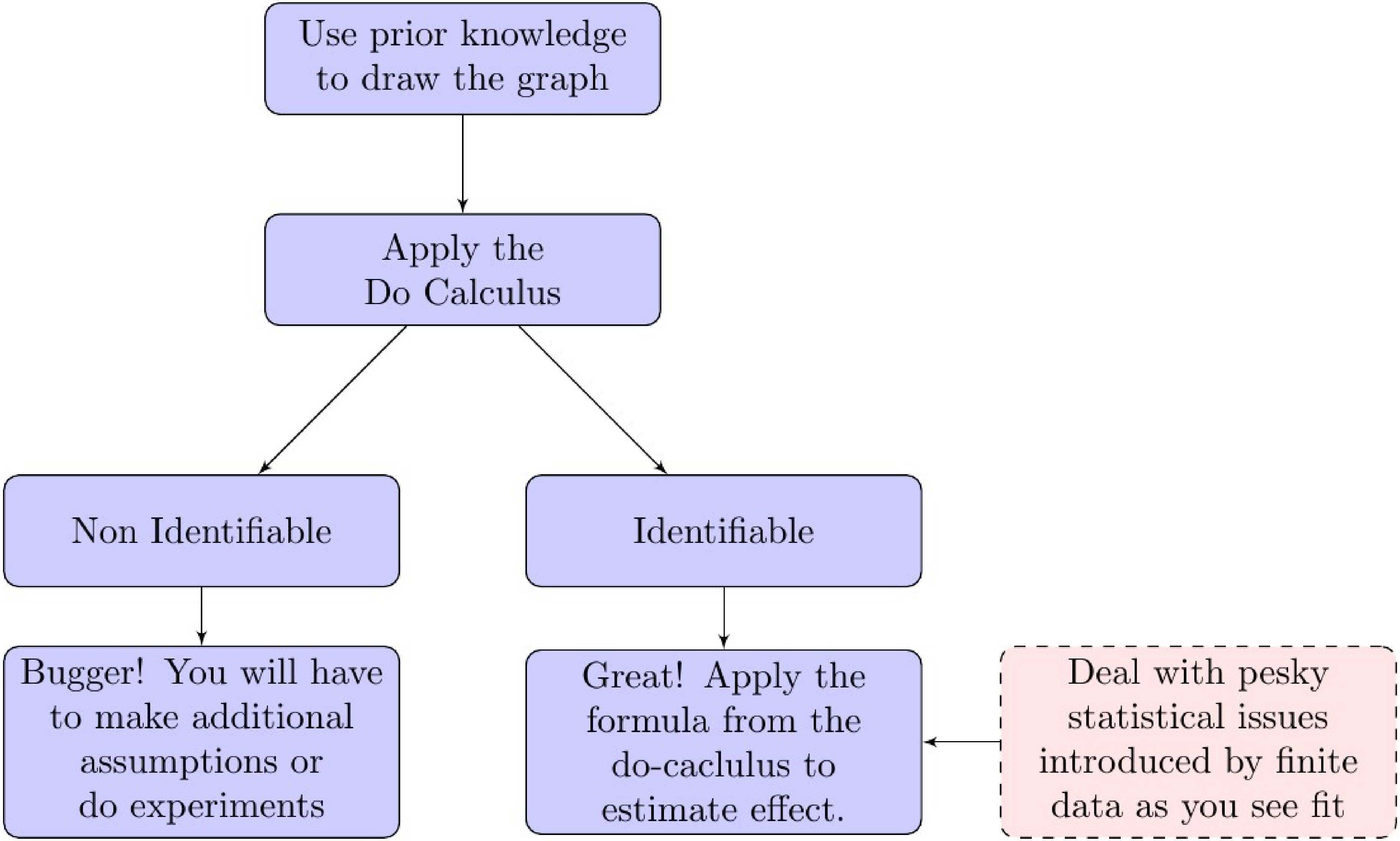
Causal Inference

Problem: Given a graph with known structure, predict the outcome of an intervention based on observational data.

Solution: Use the Do Calculus

- 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)

A recipe for causal inference from observational data



Use prior knowledge to draw the graph

Apply the Do Calculus

Non Identifiable

Identifiable

Bugger! You will have to make additional assumptions or do experiments

Great! Apply the formula from the do-caclulus to estimate effect.

Deal with pesky statistical issues introduced by finite data as you see fit

The Do Calculus (simplified)

1. D-separation still applies after intervention.

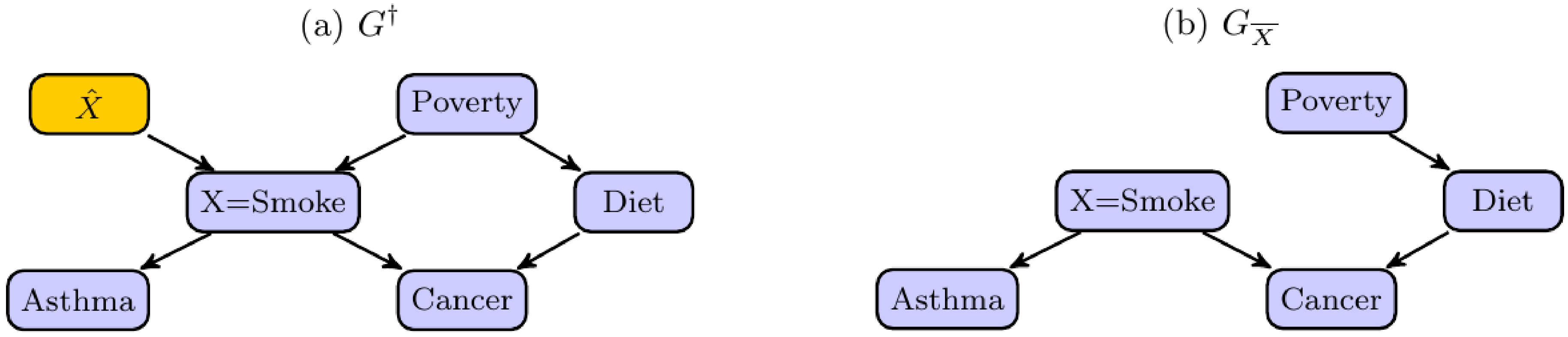
   $(Cancer \perp\!\!\!\perp Asthma | Smoke)_{G_{\overline{X}}} \implies P(Cancer | do(Smoke), Asthma) = P(Cancer | do(Smoke))$

2. If there are no backdoor paths from X to Y then intervention≡observation.

   $(\hat{X} \perp\!\!\!\perp Cancer | X, Poverty)_{G^\dagger} \implies P(Cancer | do(Smoke), Poverty) = P(Cancer | Smoke, Poverty)$

3. If there are only backdoor paths from X to Y then intervention doesn't change $P(Y)$.

   $(\hat{X} \perp\!\!\!\perp Diet)_{G^\dagger} \implies P(Diet | do(Smoke)) = P(Diet)$



(a) $G^\dagger$

(b) $G_{\overline{X}}$

Causal Discovery
when you don't know the graph

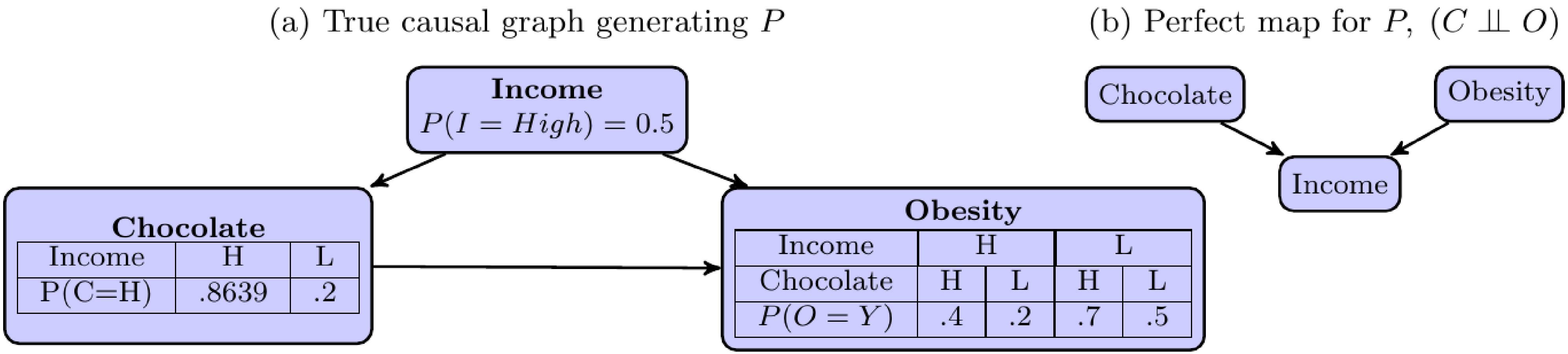
Independence based methods

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 the graph(s) that imply exactly the set of conditional independence relations as are observed in P.

An example violating faithfulness



(a) True causal graph generating P

(b) Perfect map for P, $(C \perp\!\!\!\perp O)$

Income	H	L
P(C=H)	.8639	.2

Income	H		L	
Chocolate	H	L	H	L
$P(O = Y)$	.4	.2	.7	.5

Independence based Algorithms

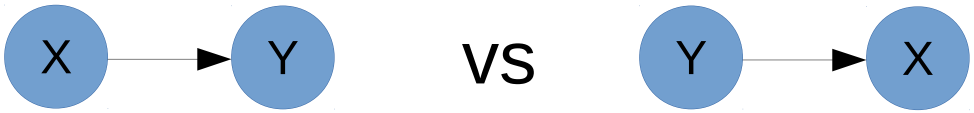
Constraint based

- IC/SGS algorithm Sprites 2000/Pearl 2000
- PC
- FCI
- RFCI

Search and Score

- GES

Beyond conditional independence



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

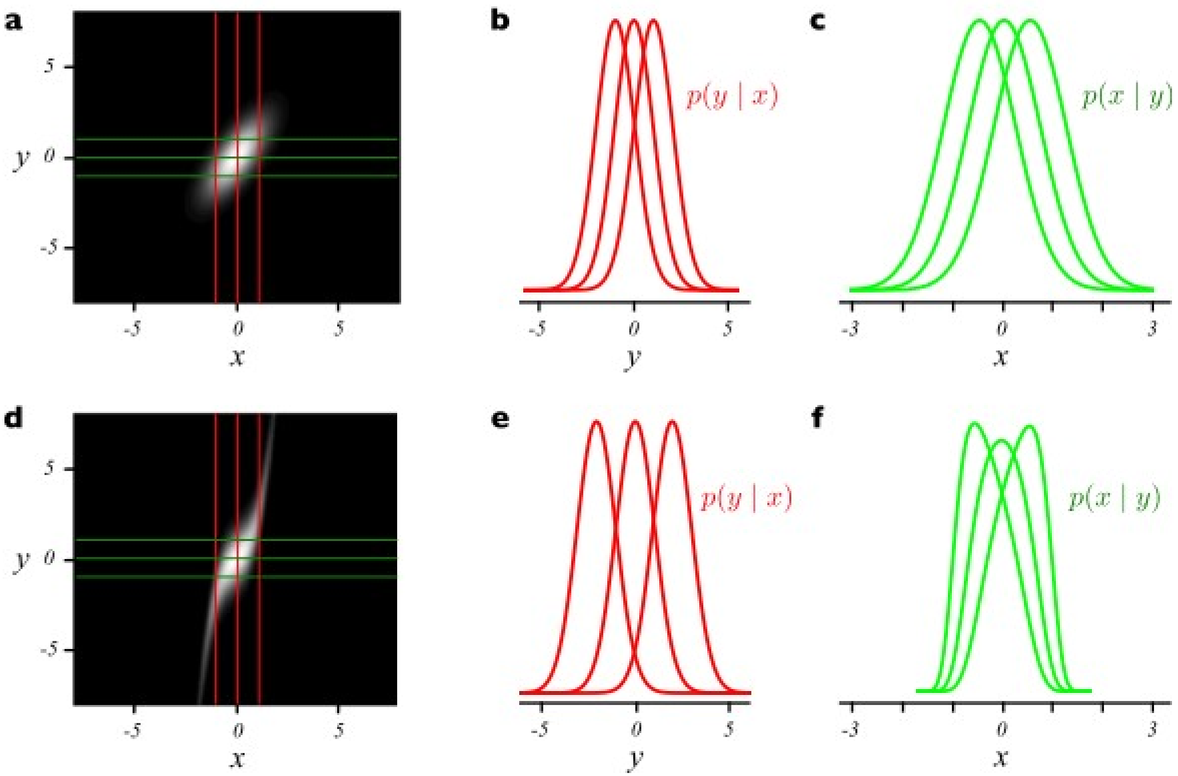


Figure 1, (Hoyer et al 2009)

Can be extended to post-non-linear additive noise, $y = h(f(x) + e)$, (Zhang et al 2009)
Can be extended beyond bi-variate graphs. (Peters et al 2014)

More asymmetries of cause and effect

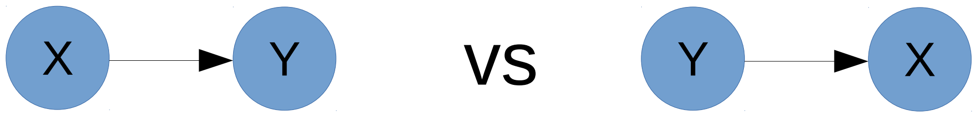
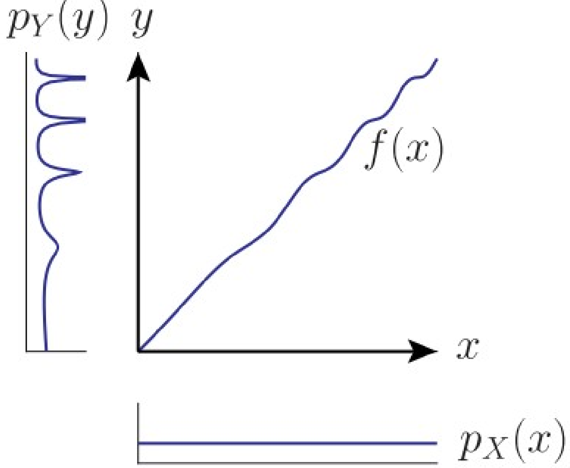

X → Y vs Y → X

Figure 1: Daniusis et al 2012



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

We expect P(Y|X) to be related to P(Y) but not to P(X)

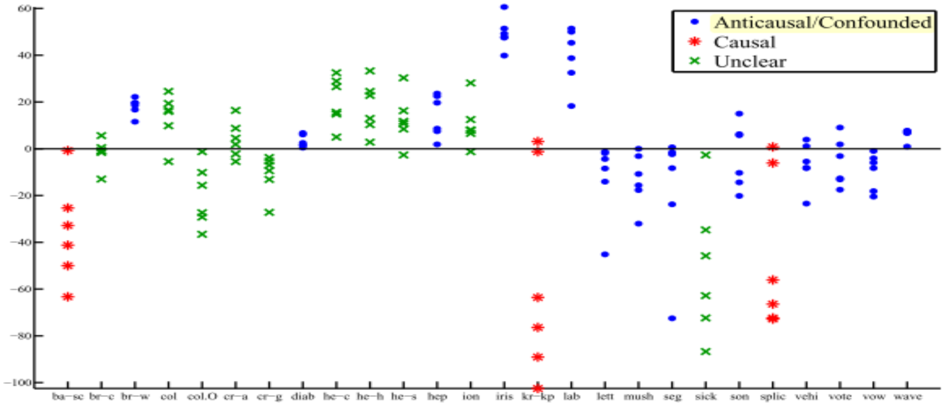


Figure 6, Janzing & Peters 2012

Semi-supervised learning supplements data sampled from P(X,Y) with additional points from P(X) with the goal of learning P(Y|X). If X → Y the additional data should not help.

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 $X \rightarrow Y$ or $Y \rightarrow X$ in dataset i.

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

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
```

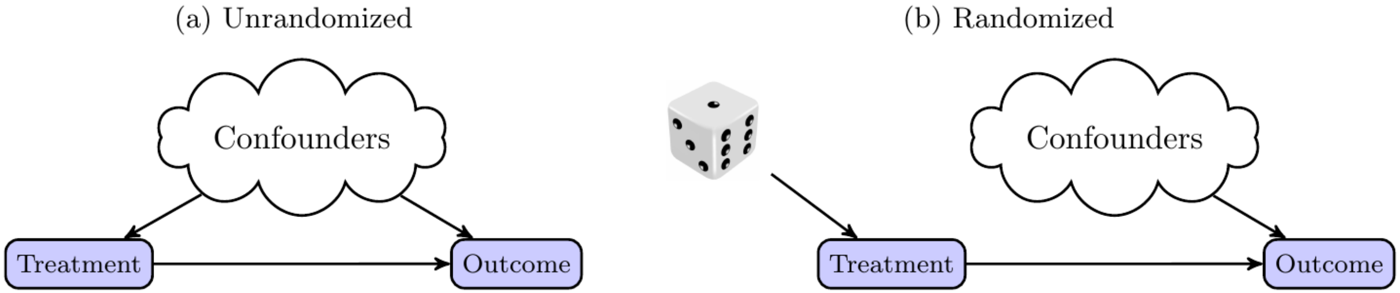
See Lopez-Paz et al 2014

Applications

- Some links to research that have applied some of these methods...A first place to look would be follow up papers from that symposium on causal inference.

Causal Inference and Bandits

Randomized trials considered gold standard for determining causality



Bandits algorithms can be seen as an improvement on randomized trials that leverges the sequential nature of the decision process

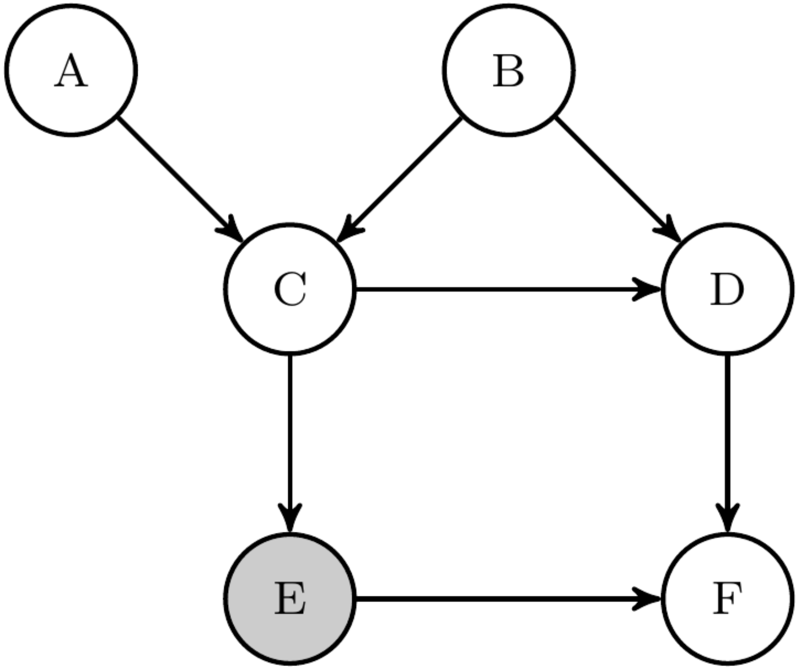
Can we incorporate ideas from causal inference into the bandit framework?
What problems would this be useful for?

Establishing a link between causal graphs and bandits

- Each possible assignment of variables to values that we can make is an action (or bandit arm)
- Reward is value of a single specified node in the graph after the action is chosen – cost of actions.
- Problem takes on characteristics of different bandit problems depending on what you get to see before you select an action what feedback you get afterward

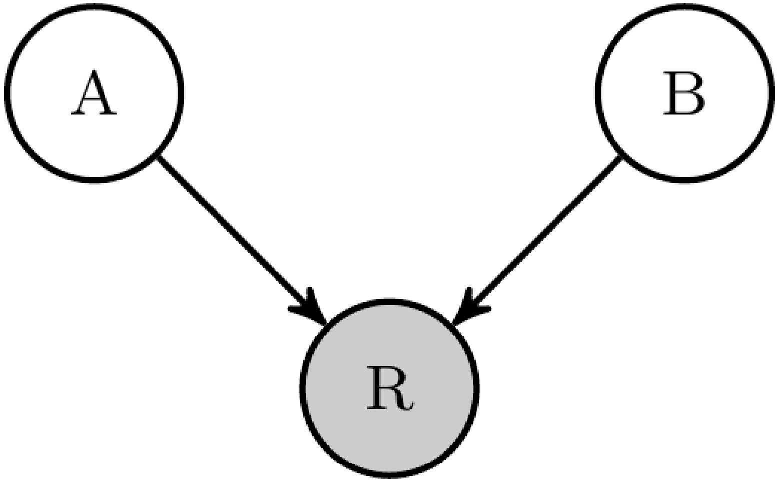
Feedback on reward node only

- We can rule out some actions immediately based on the graph structure
- Then run a standard bandit algorithm on remaining actions



Feedback on additional nodes

- Can give us some, but not always full, information on actions that were not selected.



Actions =

do(A=0,B=0)
do(A=0,B=1)
do(A=1,B=0)
do(A=1,B=1)
do(A=0)
do(A=1)
do(B=0)
do(B=1)
do()

$$\begin{aligned} P(R|do(A=1)) &= P(R|A=1) \\ &= P(R|A=1, do(B=0))P(B=0) + P(R|A=1, do(B=1))P(B=1) \end{aligned}$$

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).

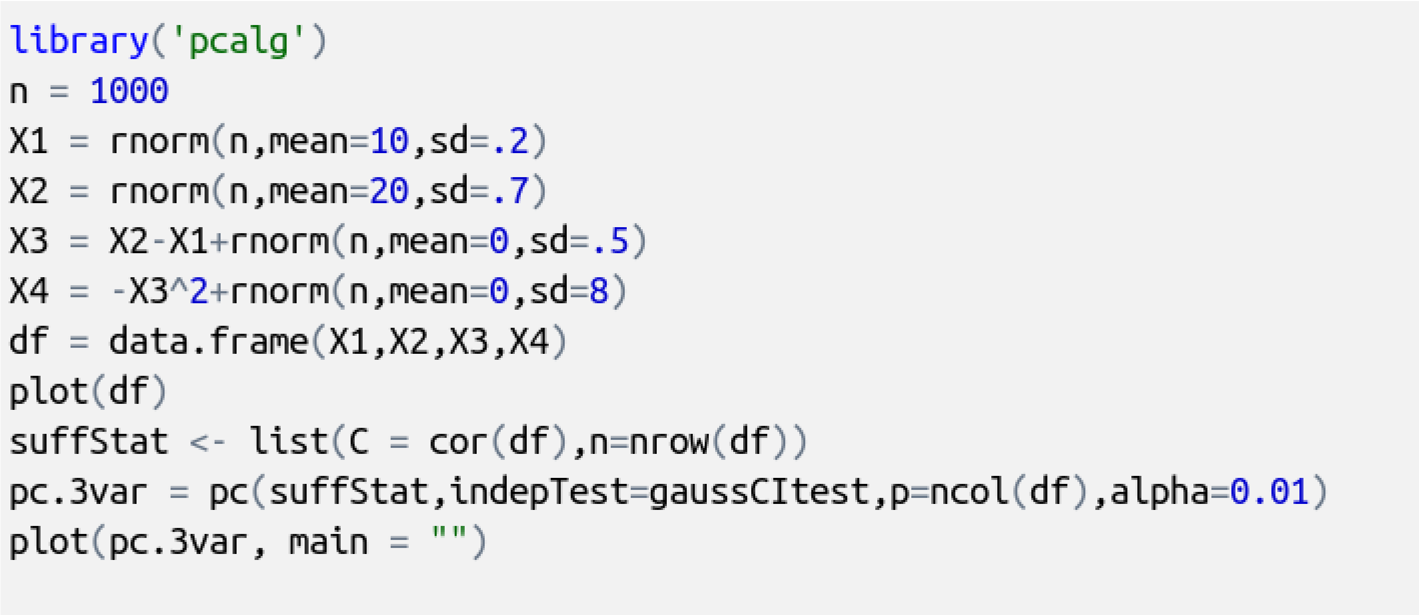
Dominik Janzing and Jonas Peters. On causal and anticausal learning JMLR. , 2012.

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

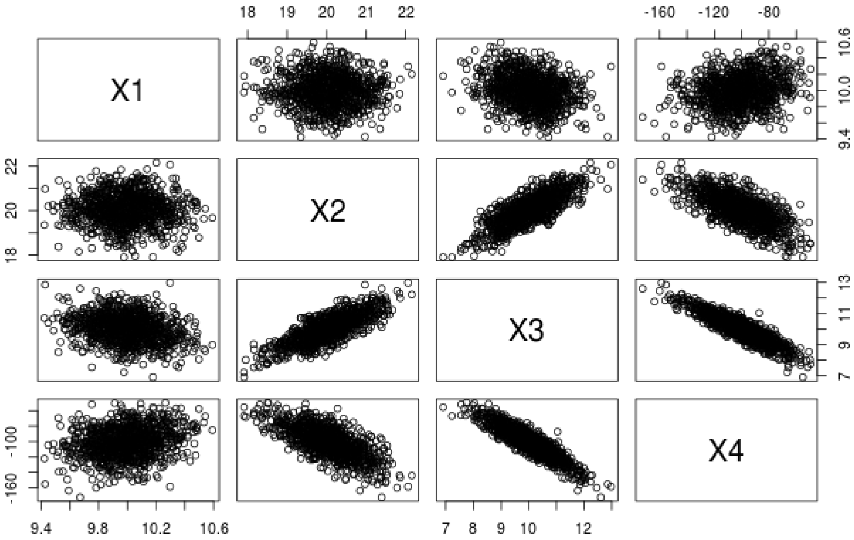
TS Richardson and JM Robins. Single world intervention graphs (SWIGs): a unication of the counterfactual and graphical approaches to causality. Cent. Stat. . . . , (128), 2013.

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

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 = "")
```


X1
X2
X3
X4
18 19 20 21 22
-160 -120 -80
10.6 10.0 9.4
22 20 18
13 11 9 7
-100 -160
9.4 9.8 10.2 10.6
7 8 9 10 12