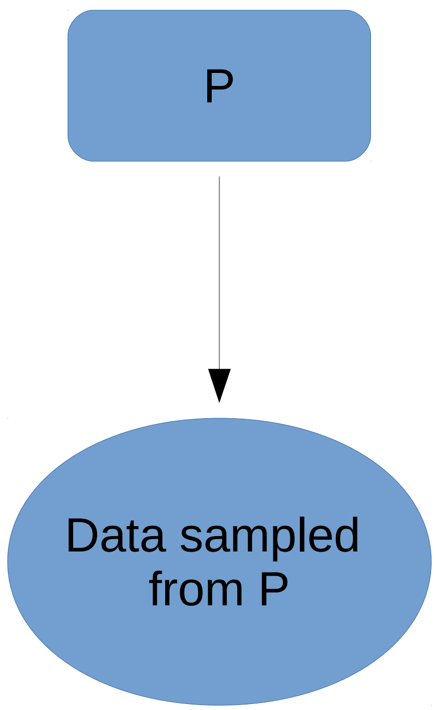
Causal Inference in Machine Learning

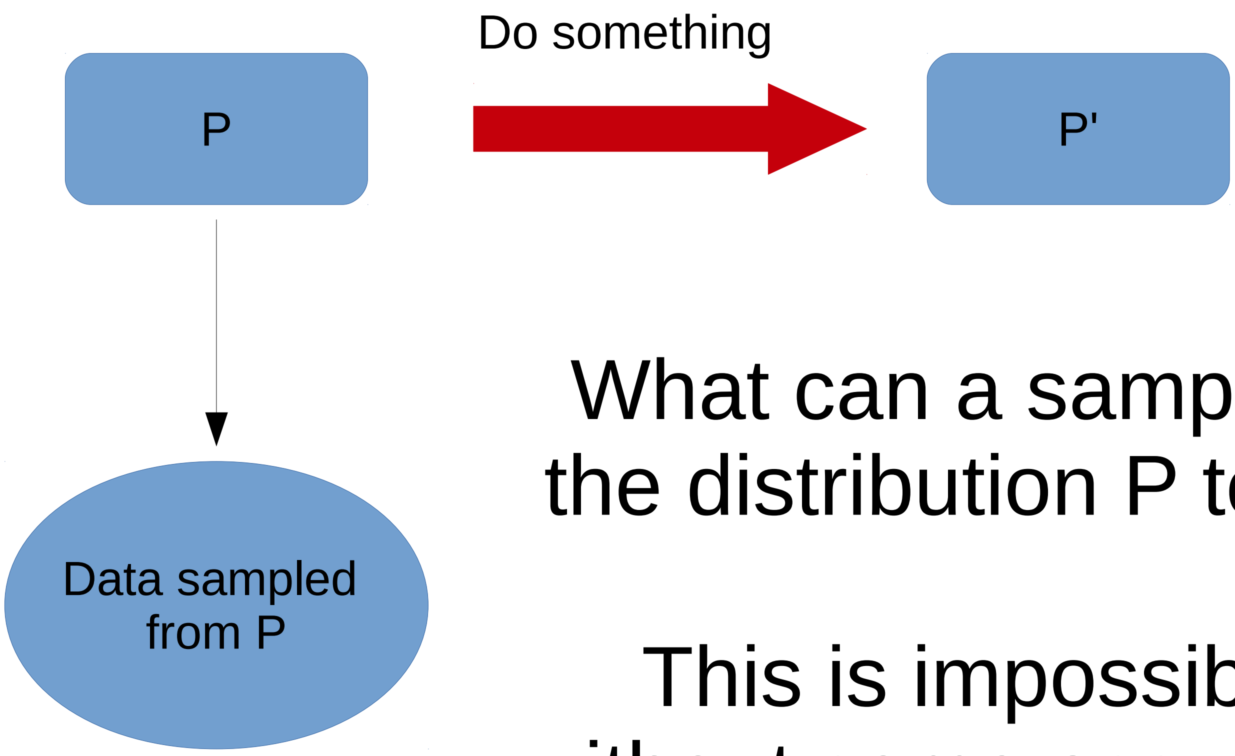
Finnian Lattimore (finnlattimore@gmail.com)

Machine Learning/Statistics



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

Causal inference



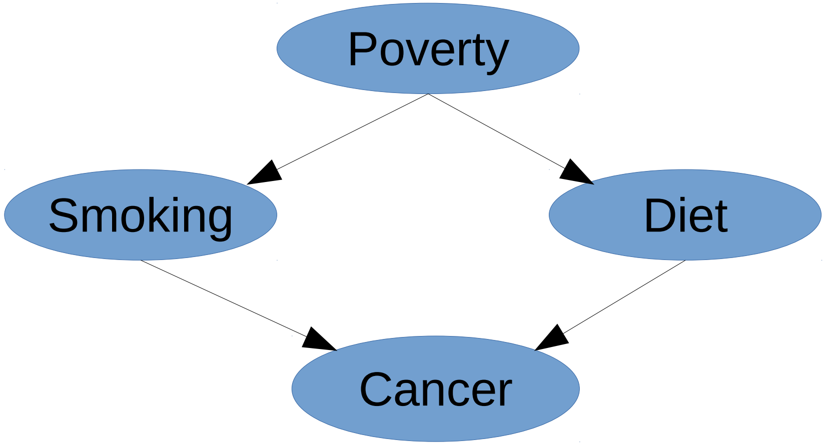
What can a sample of data from the distribution P tell us about P'?

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)

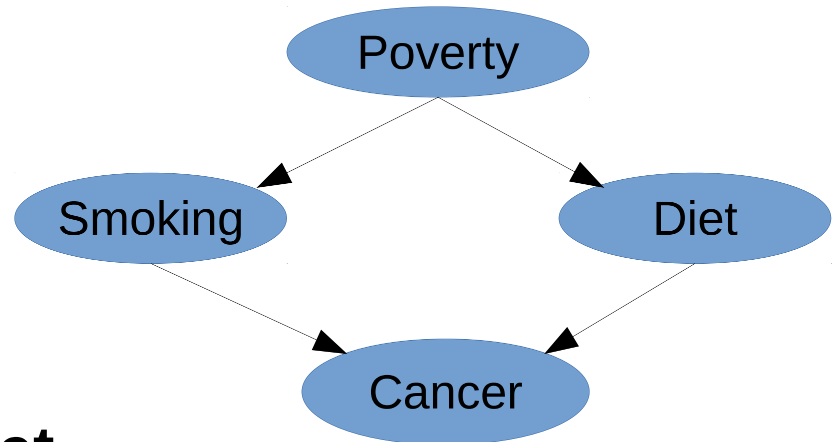


Absent links imply the factorisation of the full distribution can be simplified.

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

Intervention in Causal DAGs

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

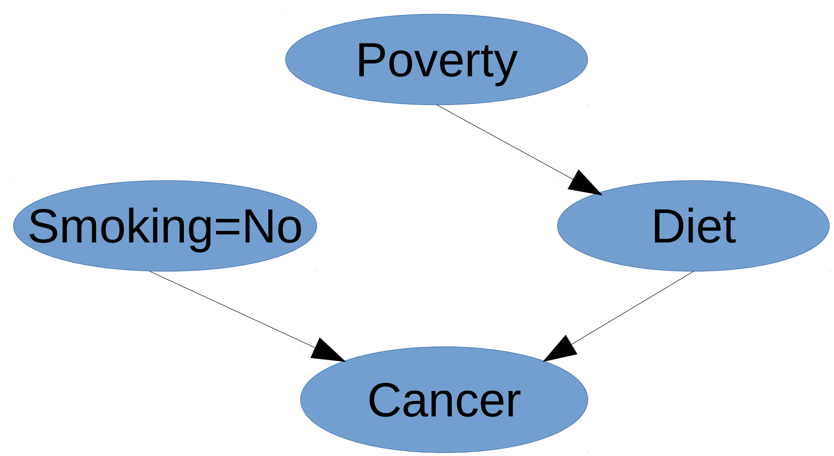


Truncated product formula

Drop from terms for intervened on variables from the factorization

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

$$do(Smoking=No)$$



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

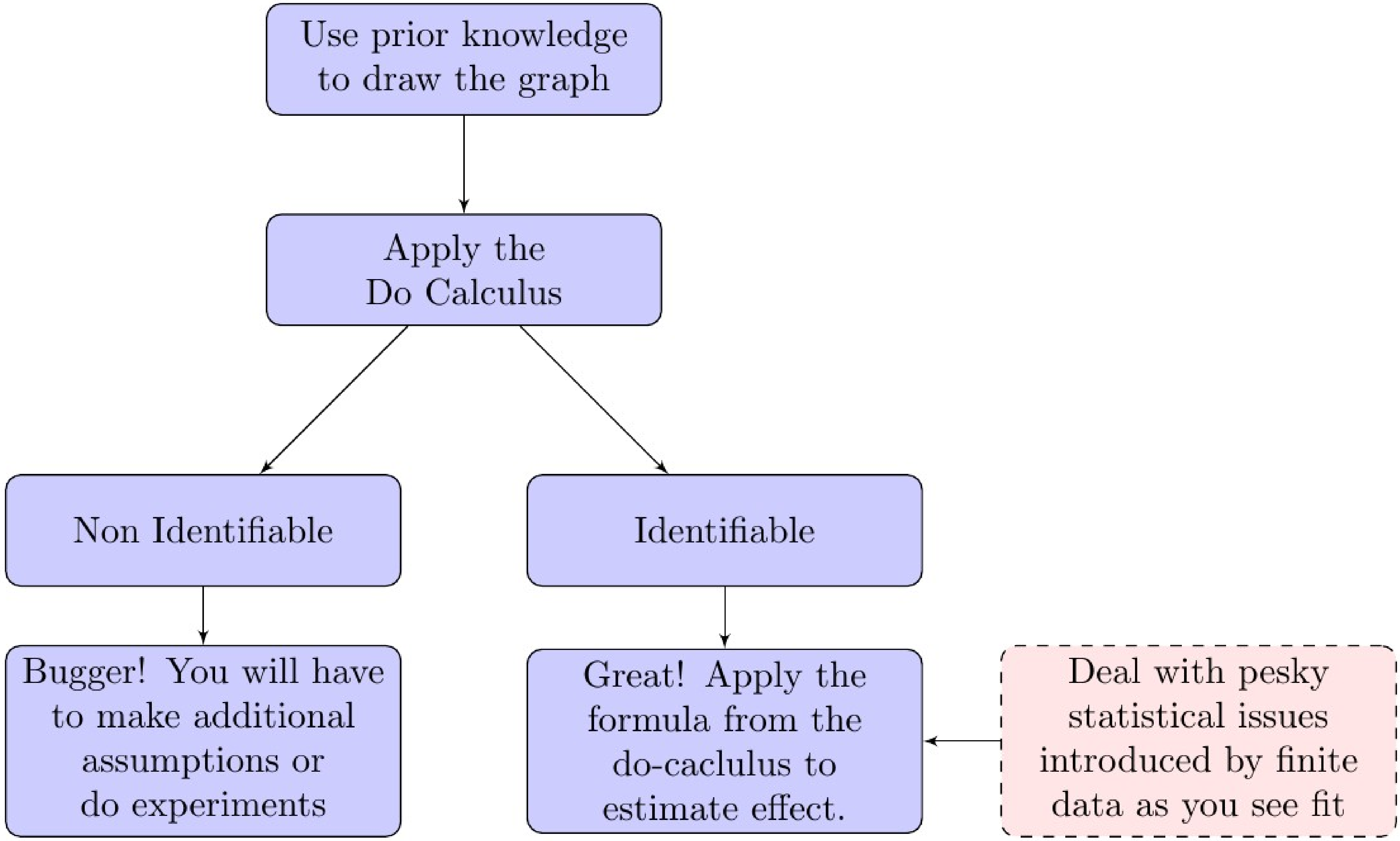
Causal Inference

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)

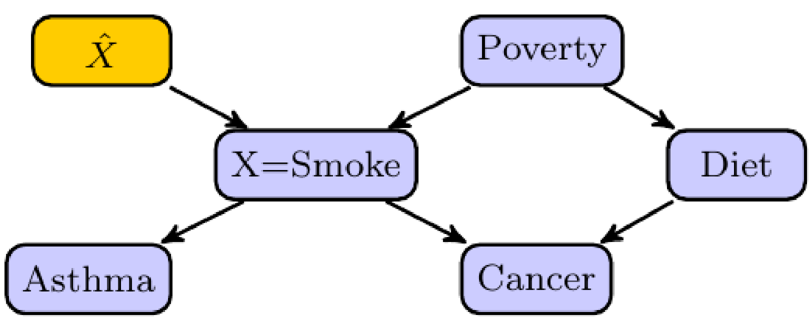
A recipe for causal inference from observational data



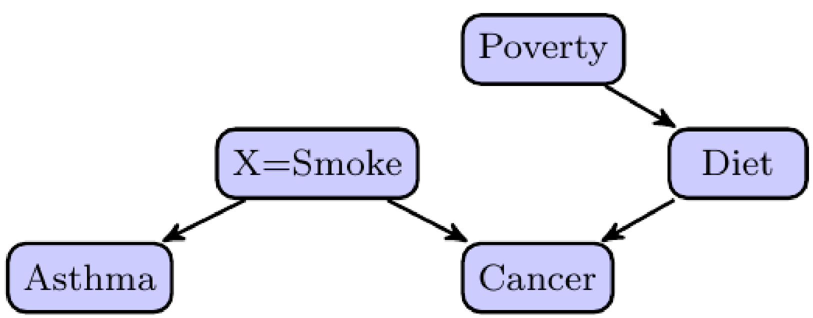
The Do Calculus (simplified)

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

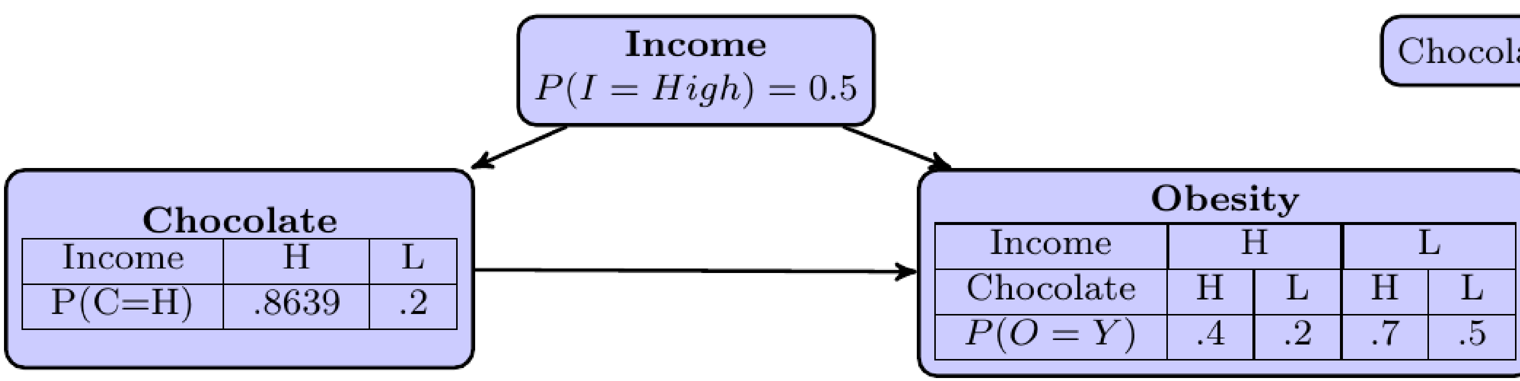
1) We assume our distribution P was generated by some (unknown) causal DAG over our observed variables (causal sufficiency)

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

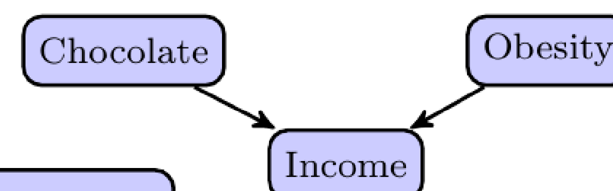
3) Finding the causal structure equates to finding the graph(s) that imply exactly the set of conditional independence relations as are observed in P.

An example violating faithfulness

(a) True causal graph generating P



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



$$P(O)=P(O|C=H)=P(O|C=L)=.46$$

Independence based Algorithms

A few of the many independence based causal discovery algorithms

Alg.	**Method**	**Scales (num.vars)**	**$\sim$ Vars**	**Latent**	**Reference**
IC/SGS	Constraint based	Exponential	10	No	Pearl(2000)/Sprites(2000)
PC	Constraint based	Worst case exponential, polynomial for sparse graphs	5000	No	Sprites(2000)
FCI	Constraint based	Worst case exponential, polynomial variant FCI+ for sparse graphs	30	Yes	Sprites(2000)
RFCI	Constraint based	?	500	Yes	Colombo(2012)
GES	Search & Score	Worst case exponential	50	No	Chickering(2002)
MMHC	Hybrid	?	5000	No	Tsamardinos(2006)

- Constraint based methods perform sequential conditional independence tests and eliminate inconsistent graphs.
- Search and Score methods search over the space of graphs and score them according to how well they fit the independences given a complexity penalising prior.

Beyond conditional independence

$X \rightarrow Y$ vs $Y \rightarrow X$

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

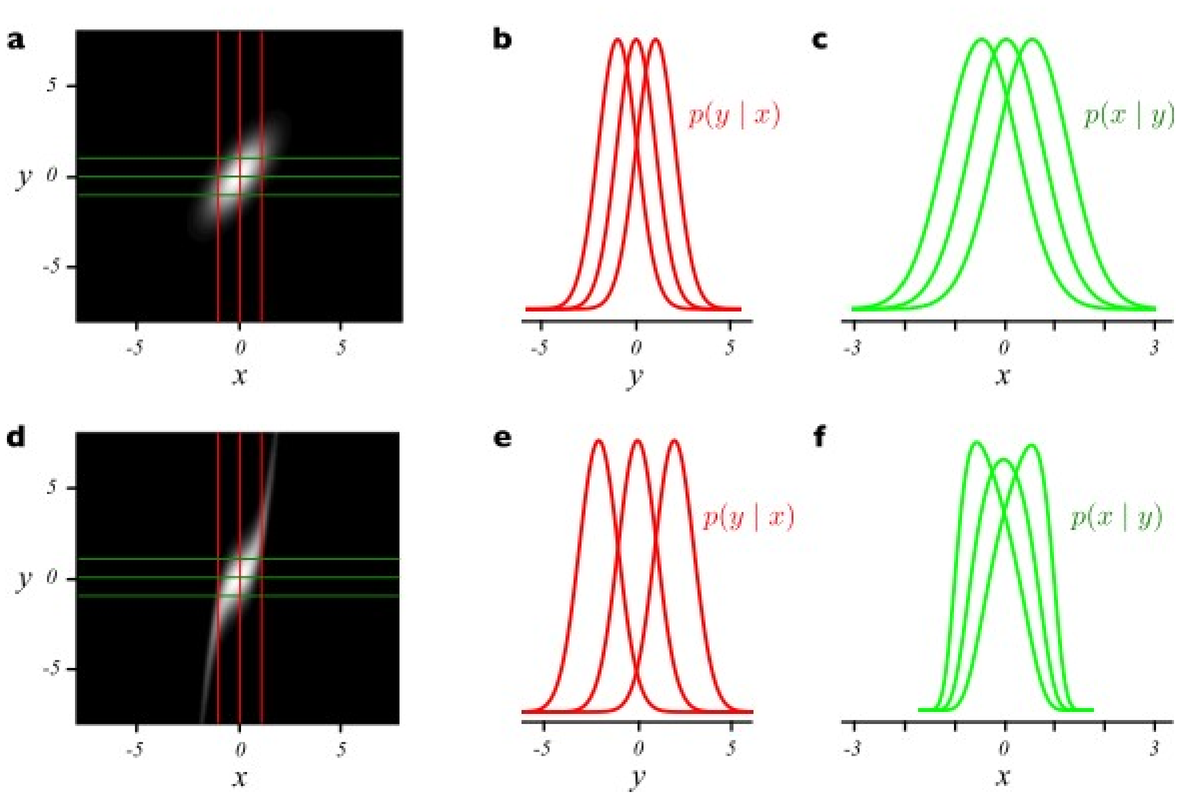


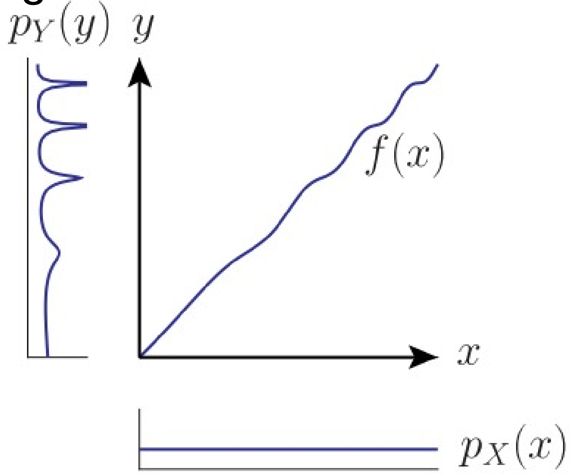
Figure 1, (Hoyer et al 2009)

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)

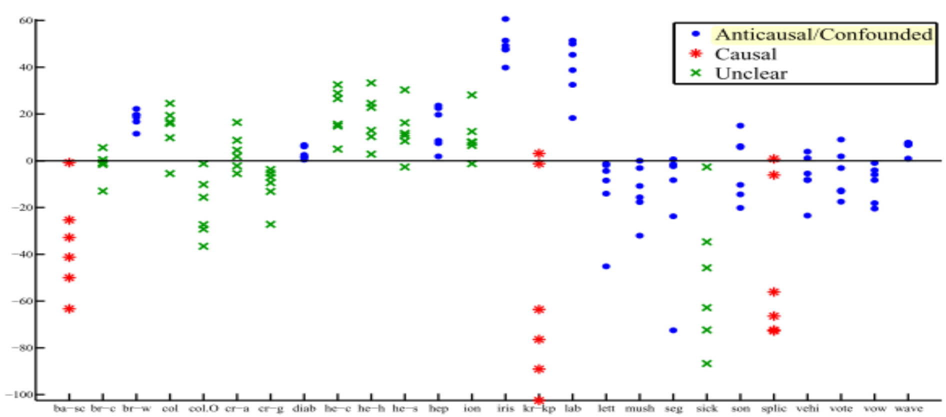


Figure 6, Janzing & Peters 2012

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 $i = 1...M$

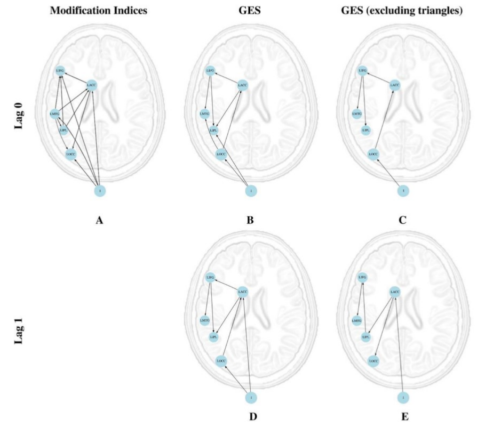
Construct a feature vector that approximates $\mu(P(X)), \mu(P(Y)), \mu(P(X, Y))$

Apply a standard classification algorithm

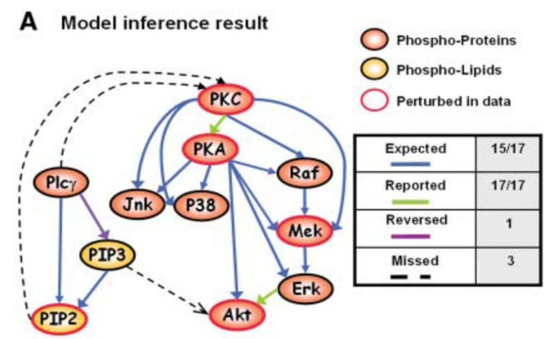
See Lopez-Paz et al 2014

Applications

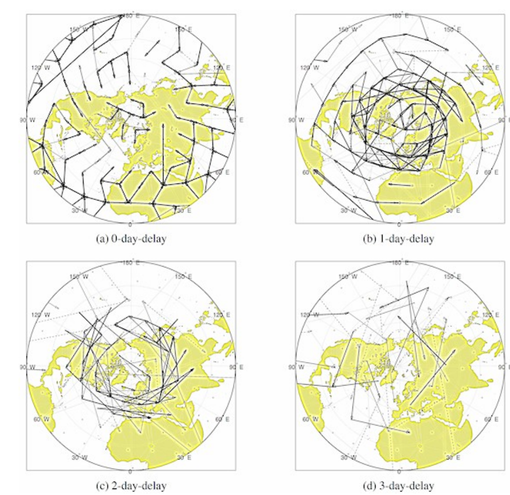
- FMRI
  - Smith, S., et al. Neuroimage (2011)
  - Ramsey, J., et al. Neuroimage (2010)
  - Iyer, Swathi P., et al. Neuroimage (2013)
- Protein signalling
  - Sachs, K., et al. Science (2005)
- Climate modelling
  - Ebert-Uphoff, et al. Geophysical Research Letters (2012)
- Genomics
  - Gene expression Taruttis, F., et al Bioinformatics (2015)
  - Genome wide association, Alekseyenko, A. et al., Biology direct (2011)
  - Molecular interactions, Statnikov, A., et al. BMC genomics (2012)
- Medicine
  - Borsboom, D.,et al, Annual review of clinical psychology (2013)
  - Ruzzano, L., et al , Journal of autism and developmental disorders (2015)



Ramsey, J., et al. (2010)



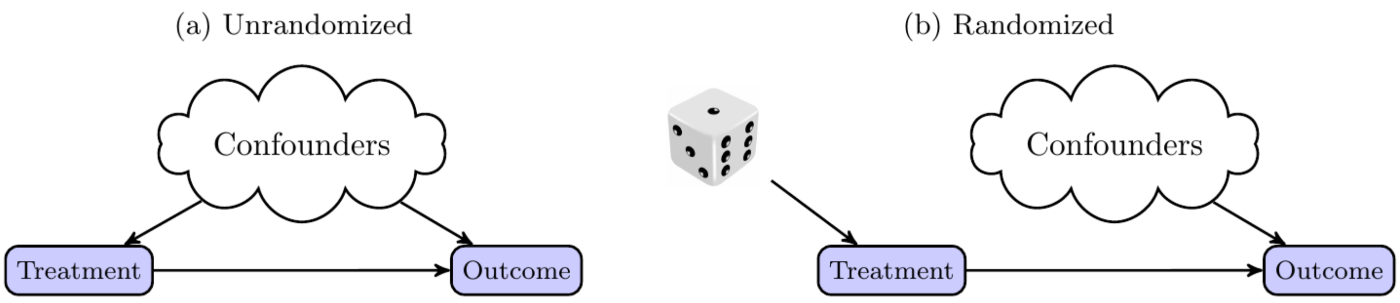
Sachs, K.,et al.(2005)



Ebert-Uphoff, et al. (2012)

Causal Inference and Bandits

Randomized trials considered gold standard for determining causality

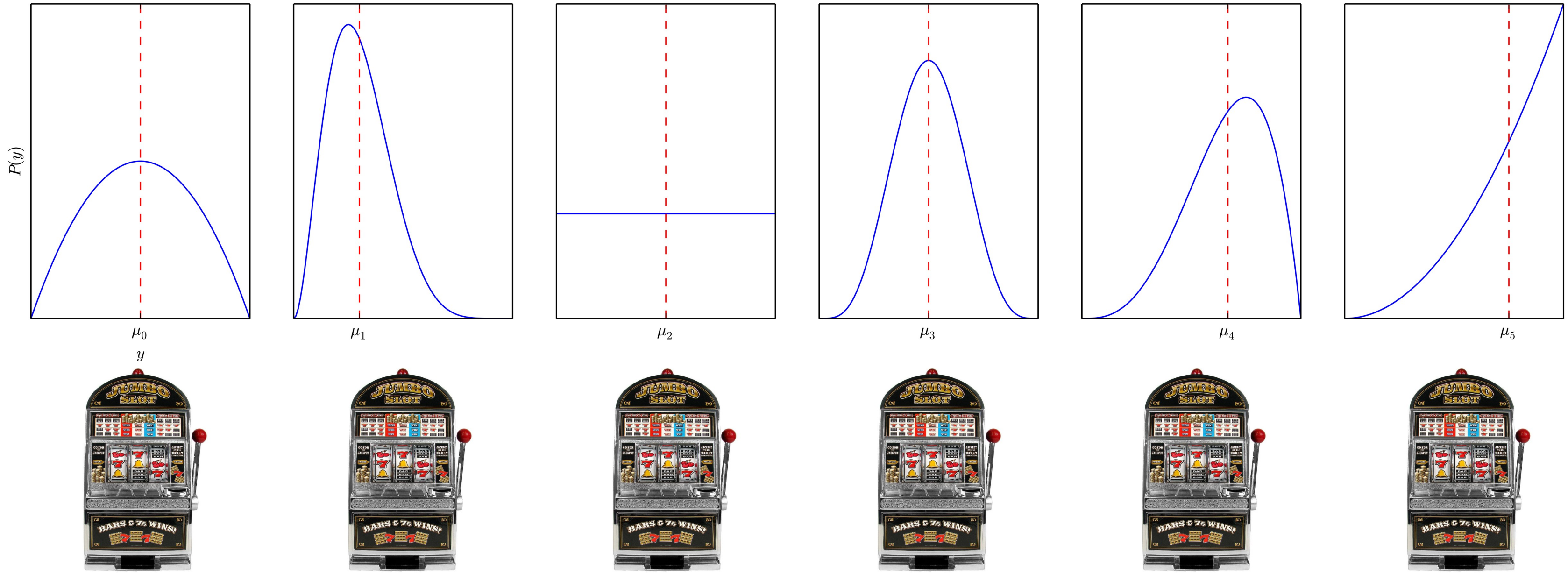


Bandits algorithms can be seen as an improvement on randomized trials that leverage the sequential nature of the decision process.

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

Classic Multi-armed Bandits

Multiple actions (arms). Each associated with an unknown but fixed distribution over reward, y.



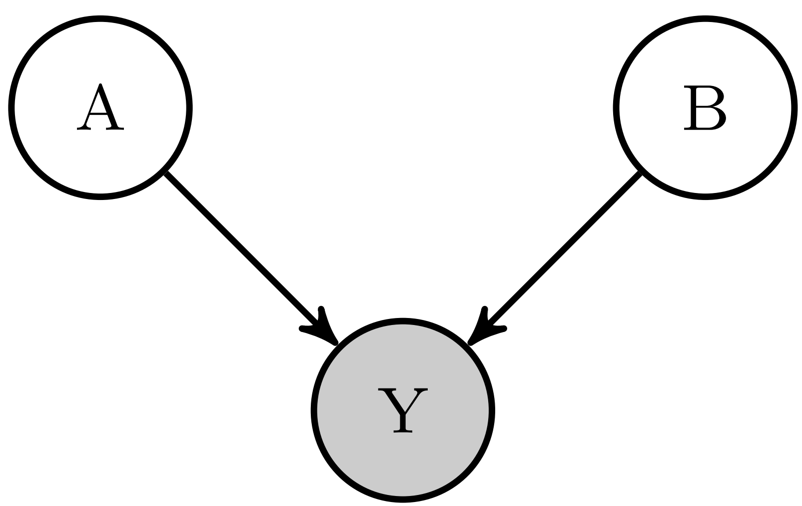
Measure algorithm performance in terms of (psuedo)-regret.

$$R_T = T\mu^* - \sum_{t=1}^{T} E[\mu_{i_t}]$$

We are learning if the regret is sublinear in T. Optimal algorithms get $R_T = O(\sqrt{TK})$

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.

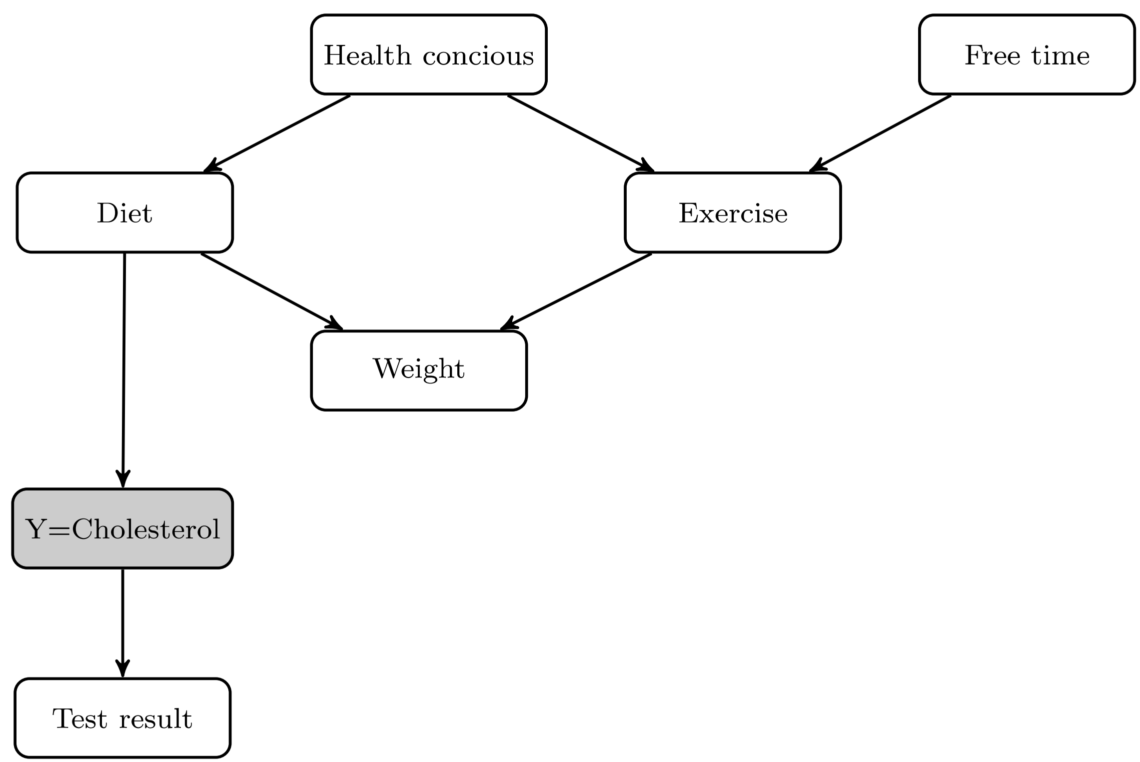


Actions =

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()

Feedback on reward node only

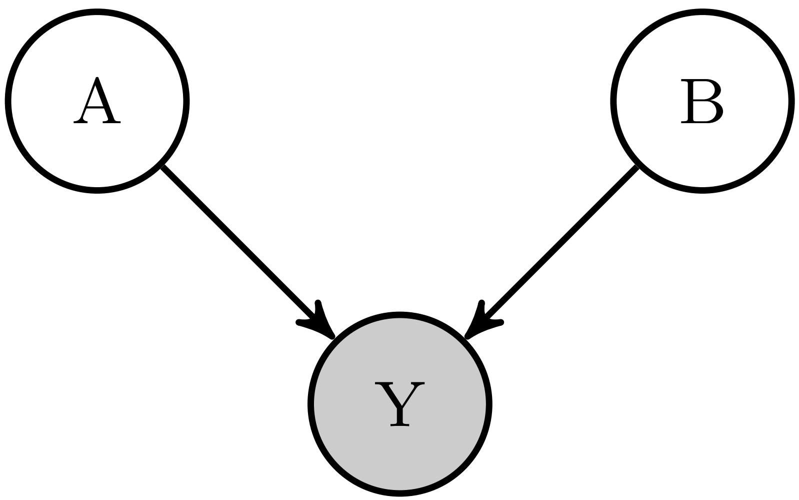
- We can rule out some actions immediately based on the graph structure
- Then run a standard bandit algorithm on remaining actions



Network adapted from Koller(2009)

Feedback on additional nodes

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

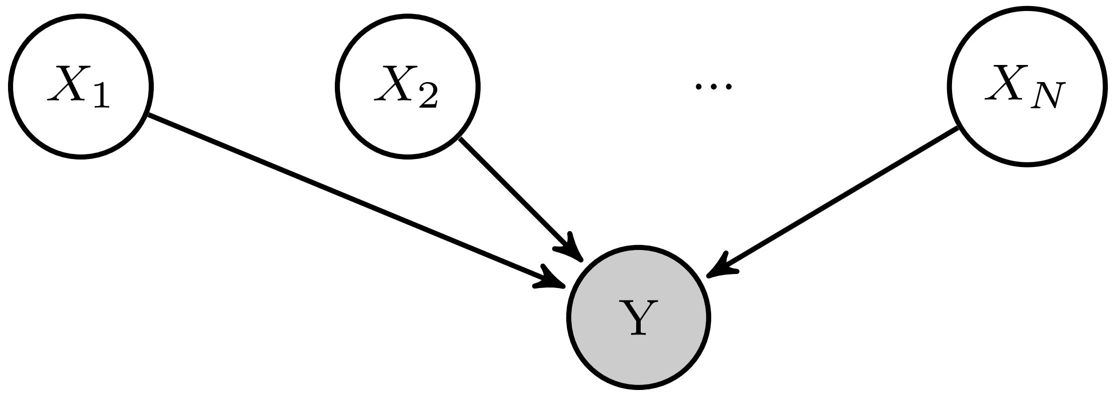


$$\text{Actions} = \begin{array}{|c|}\hline \text{do(A=0,B=0)} \\ \text{do(A=0,B=1)} \\ \text{do(A=1,B=0)} \\ \text{do(A=1,B=1)} \\ \hline \text{do(A=0)} \\ \text{do(A=1)} \\ \text{do(B=0)} \\ \text{do(B=1)} \\ \text{do()} \\ \hline \end{array}$$

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

Bernoulli-bandit with causal structure

Assumed causal structure



Let $q \in [0,1]^N$ be a fixed vector where $q_i = P(X_i = 1)$. In each time-step t upto a known end point T:

1. The learner chooses an $I_t \in \{0, \ldots, N\}$ and $J_t \in \{0, 1\}$, setting $X_{I_t,t} = J_t$. Selecting $I_t = 0$ corresponds to taking the nothing action $do()$ and just observing.
2. For $i \neq I_t$, $X_{i,t} \sim Bernoulli(q_i)$
3. The learner observes $X_t = [X_{1,t} ... X_{N,t}]$
4. The learner receives reward $Y_t \sim \text{Bernoulli}(r(X_t))$, where $r : \{0,1\}^N \to [0,1]$ is unknown and arbitrary.

The causal structure gives us:

$$\begin{aligned} P(Y|do(X_i = j)) =& P(Y|X_i = j) \\ =& P(Y|do(X_a = 1), X_i = j)q_a + P(Y|do(X_a = 0), X_i = j)(1 - q_a) \end{aligned}$$

At each timestep, observing will reveal the reward for half the arms.

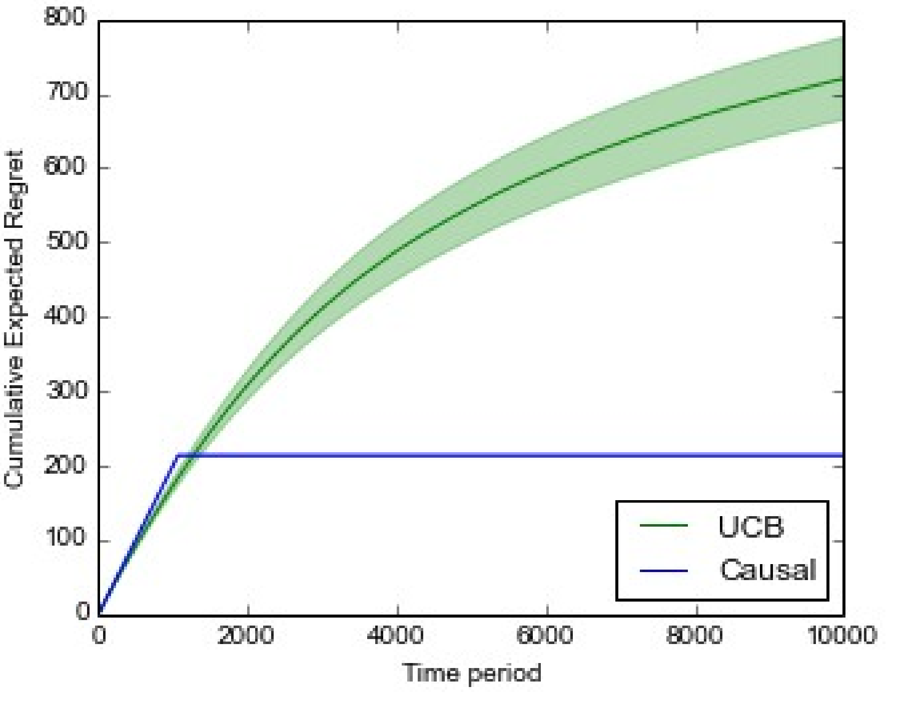
Balanced q, $q_i = \frac{1}{2} \; \forall i$

We can use an explore-exploit style algorithm,

- Observe for some number of timesteps, h
- Then pick arm with highest estimated mean for remaining time-steps
- Optimise h to minimise regret.

$$R_T(causal) = O(T^{2/3}(log(KT))^{1/3} \text{ vs } R_T(classic) = O(\sqrt{KT})$$

We expect to do better if $K >> T^{1/3}$



Comparison of the UCB and causal-explore-exploit for K=20 and T=10000. Note, K ~T^{1/3}.

Arbitrary q

Goal, quantify how unbalanced an arbitrary q is and get a regret bound in terms of that.

- Need to trade of observing vs explicitly playing low probability arms.
- Spend half our exploration time h doing each.
- Assume $q_i \in [0, \frac{1}{2}]$ and order variables such that $q_1 \leq q_2 \leq ... \leq q_N$
- Let $m \in [2, N] = i : q_i > \frac{1}{i}$
- Divide the arms into low probability $\{(i, 1) : q_i < m\}$ and frequent $\{(i, 0)\forall i \cup (i, 1) : q_i > m\}$
- Divide the $h/2$ explicit play budget between the m low probability arms, giving $\frac{h}{2m}$ samples each.
- For the frequent arms, we expect $\sim q_i \frac{h}{2} \geq \frac{h}{2m}$ samples from the observe phase.

Example: q = [0.03 0.12 0.32 0.33 0.33 0.35 0.41 0.45 0.49 0.49], m = 4

$$R_T = O(T^{2/3} m^{1/3} (log(KT))^{1/3})$$

Summary and future

- Lower bounds (is the algorithm above roughly as good as it gets?)
- Return to case where we get feedback only on reward node, but consider if the graph is not (fully) known. Can we learn it and use it to eliminate actions simultaneously.
- Does NICTA have any data/problems where we should try out causal inference techniques?

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.

Sachs, Karen, et al. "Causal protein-signaling networks derived from multiparameter single-cell data." Science 308.5721 (2005): 523-529.

Ebert-Uphoff, Imme, and Yi Deng. "A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer." Geophysical Research Letters 39.19 (2012).

Taruttis, Franziska, Rainer Spang, and Julia C. Engelmann. "A statistical approach to virtual cellular experiments: improved causal discovery using accumulation IDA (aIDA)." Bioinformatics (2015): btv461.

Statnikov, Alexander, et al. "New methods for separating causes from effects in genomics data." BMC genomics 13.Suppl 8 (2012): S22.

Smith, Stephen M., et al. "Network modelling methods for FMRI." Neuroimage 54.2 (2011): 875-891.

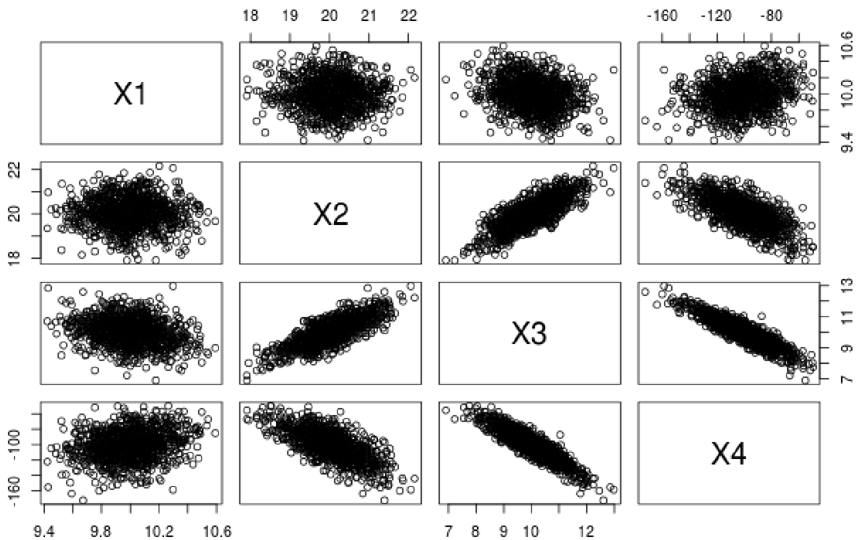
Ramsey, Joseph D., et al. "Six problems for causal inference from fMRI." Neuroimage 49.2 (2010): 1545-1558

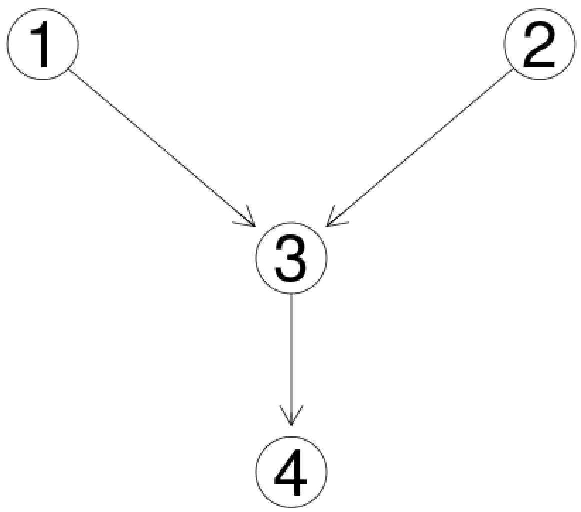
Iyer, Swathi P., et al. "Inferring functional connectivity in MRI using Bayesian network structure learning with a modified PC algorithm." Neuroimage 75 (2013): 165-175.

Ruzzano, Laura, Denny Borsboom, and Hilde M. Geurts. "Repetitive Behaviors in Autism and Obsessive–Compulsive Disorder: New Perspectives from a Network Analysis." Journal of autism and developmental disorders 45.1 (2015): 192-202.

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 = "")
```





Causal Inference in Machine Learning

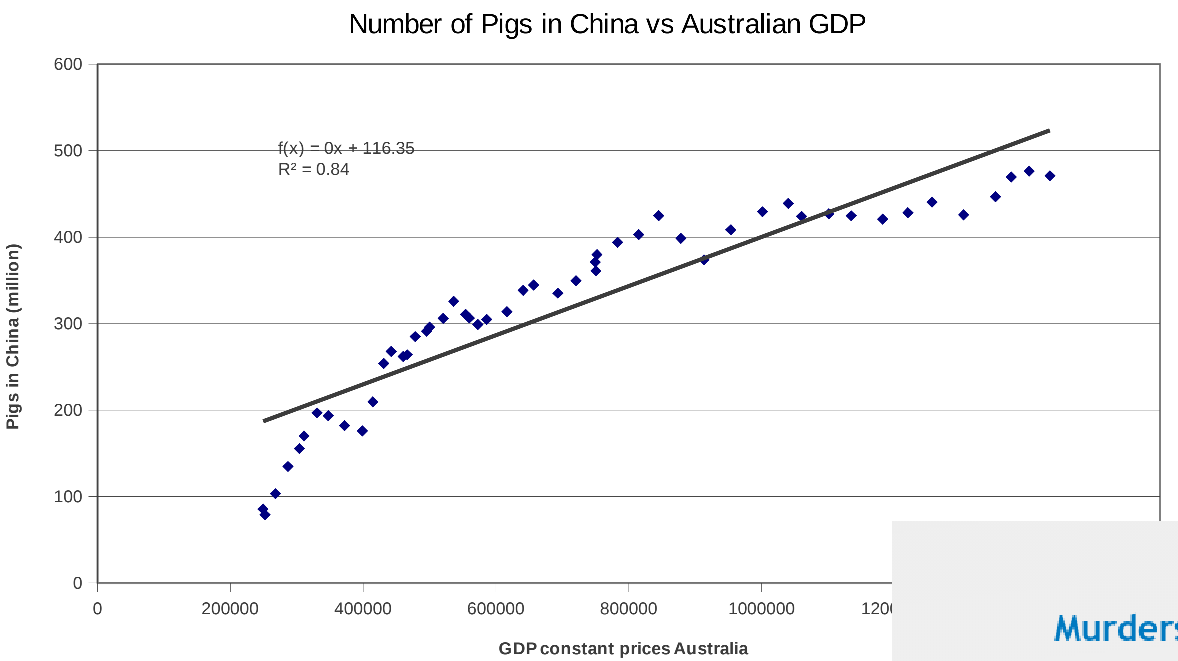
Abstract

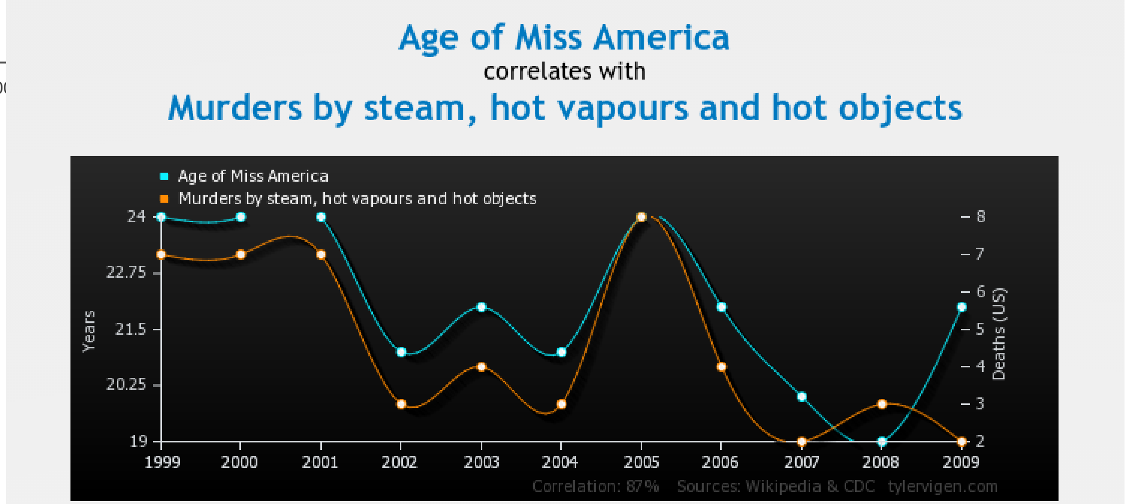
Inferring causal relationships is central to many problems involving decision making or predicting the outcome of an intervention. The past two decades has seen substantial progress formalising frameworks and developing algorithms for causal inference, particularly utilising graphical models.

Bandit algorithms, an example of reinforcement learning, present an alternative approach to decision making that takes account of the sequential nature of many such problems.

I will present a review of the key ideas in causal inference and discovery, discuss how we might start to merge them with the bandit framework and present some preliminary results demonstrating that we can incorporate causal assumptions to improve the performance of bandit algorithms.

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/