Statistical Causality

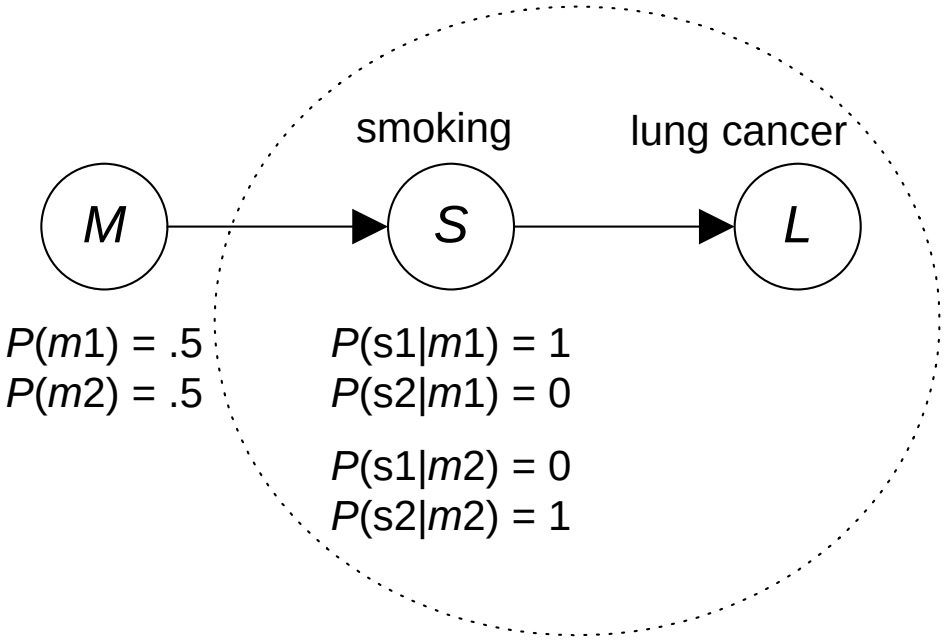
by

Rich Neapolitan

http://orion.neiu.edu/~reneapol/renpag1.htm

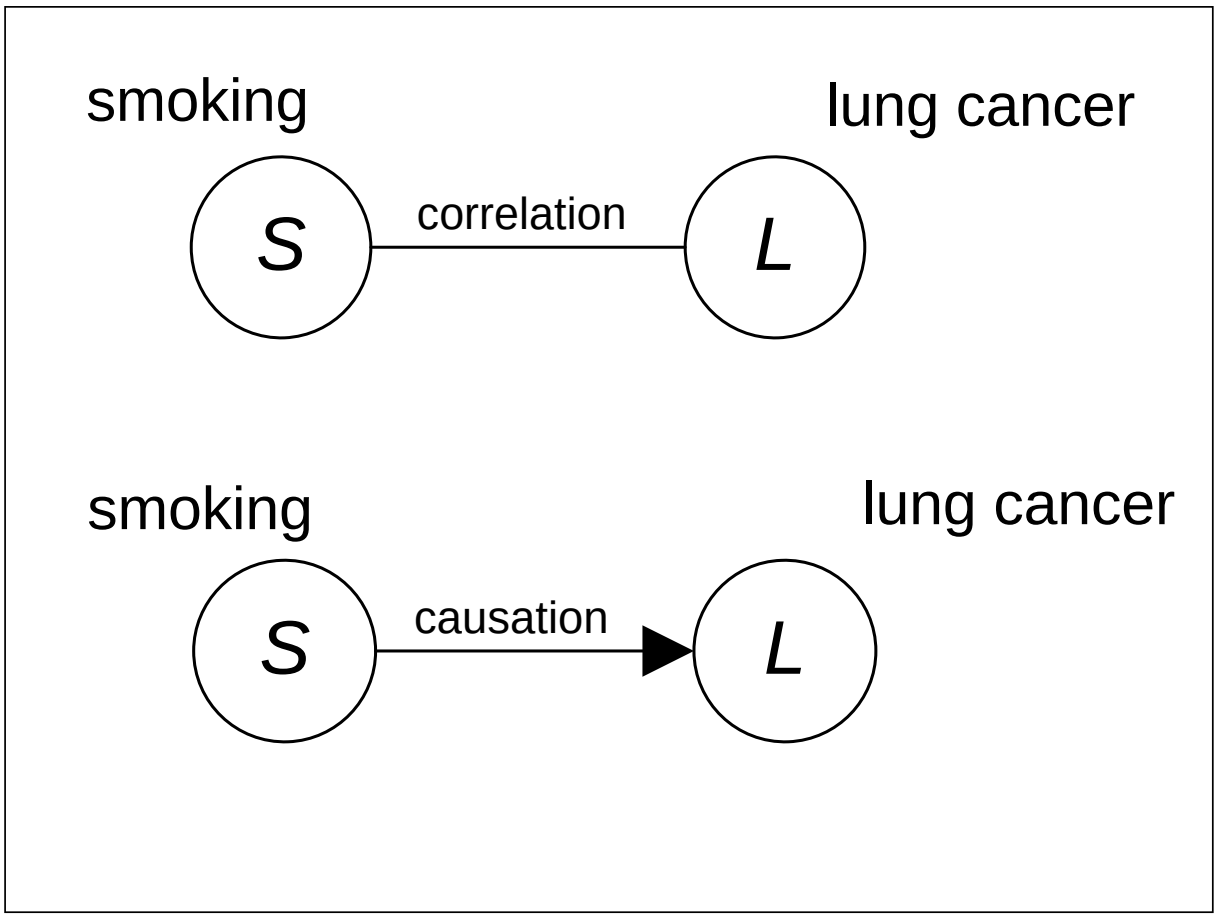
- The notion of causality discussed here is that forwarded in the following texts:
  - [Pearl, 1988]
  - [Neapolitan, 1990]
  - [Spirtes et al., 1993, 2000]
  - [Pearl, 2000]
  - [Neapolitan, 2004]
- It concerns variables influencing other variables
  - Does smoking cause lung cancer?
- It does not concern ‘token’ causality.
  - Did the gopher running into my golf ball cause it to go in the hole?

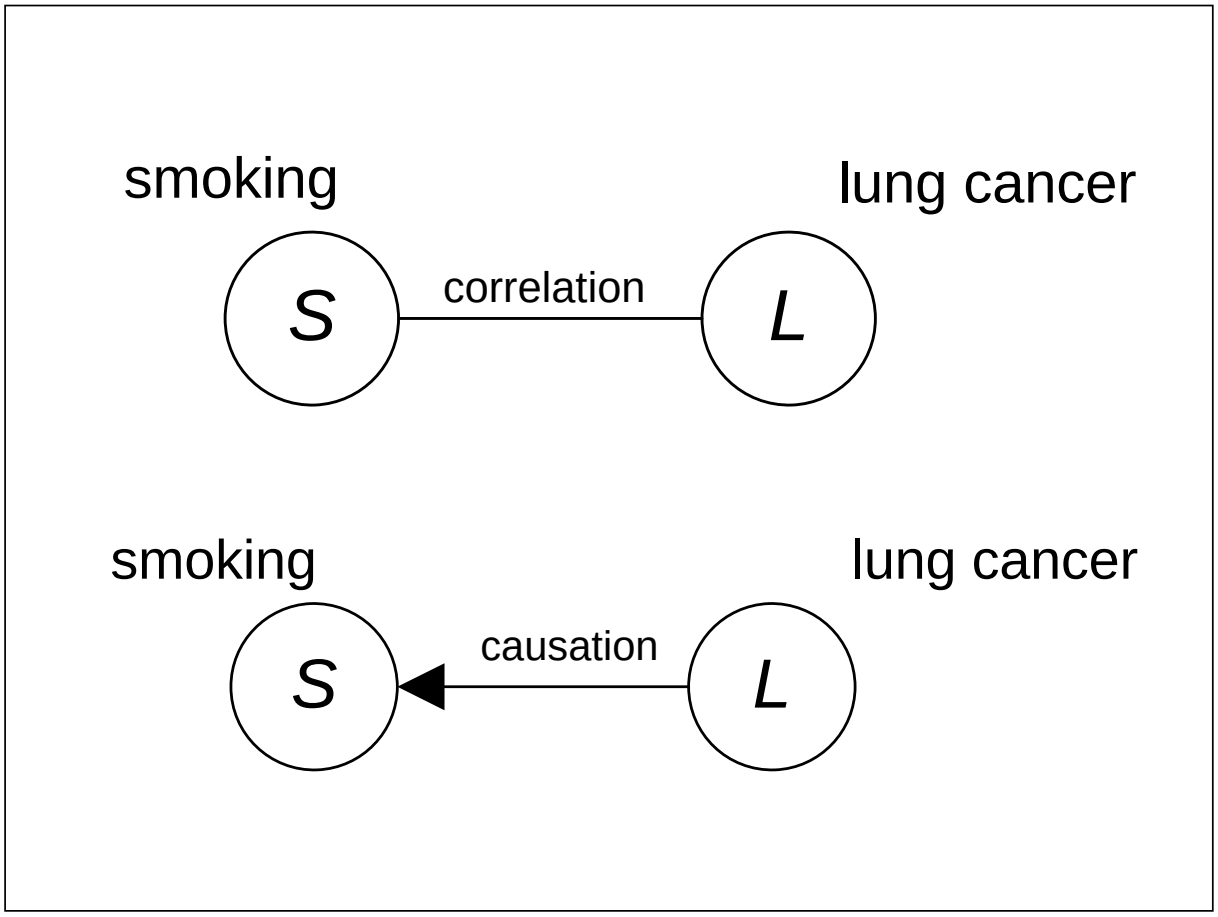
A common way to learn (perhaps define) causation is via manipulation experiments (non-passive data).

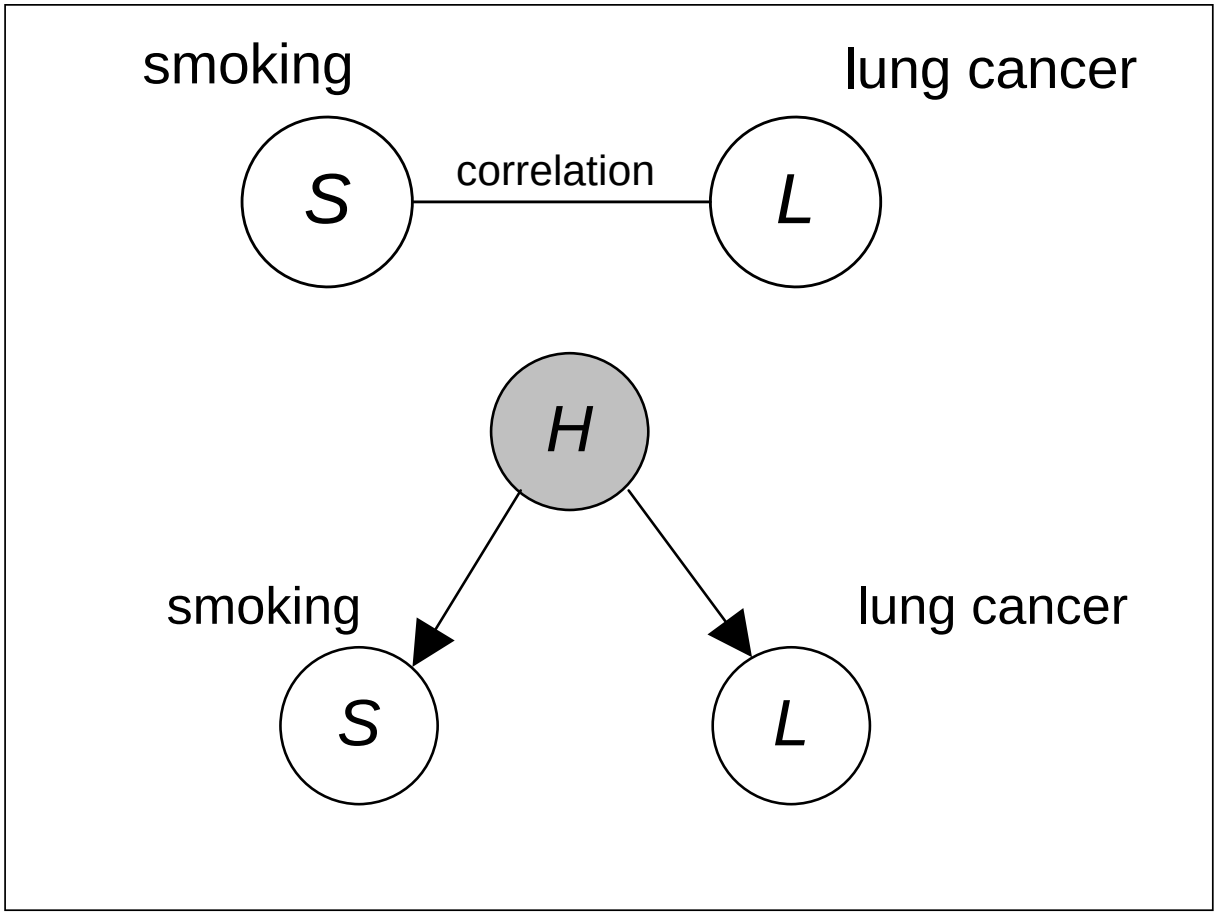


We do not really want to manipulate people and make them smoke.

Can we learn something about causal influences from passive data?


smoking
lung cancer
S
correlation
L
smoking
lung cancer
S
causation
L


smoking
lung cancer
S
correlation
L
smoking
lung cancer
S
causation
L



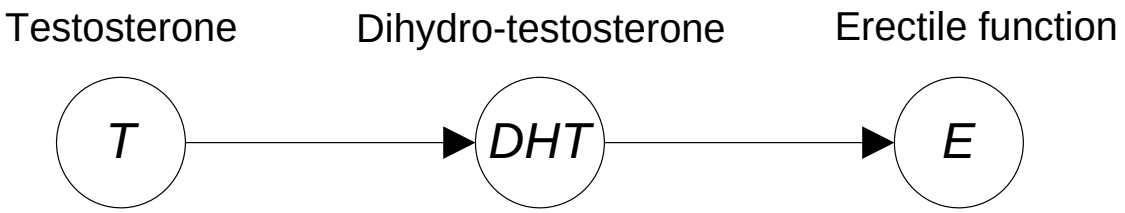
It is difficult to learn causal influences when we have data on only two variables.

We will show that we can sometimes learn causal influences if we have data on at least four variables.

Causal Graphs

A causal graph is a directed graph containing a set of causally related random variables *V* such that for every *X*, *Y* in *V* there is an edge from *X* to *Y* if and only if *X* is a **direct cause** of *Y*.

Study in rats by Lugg et al. [1995]:



The Causal Markov Assumption:

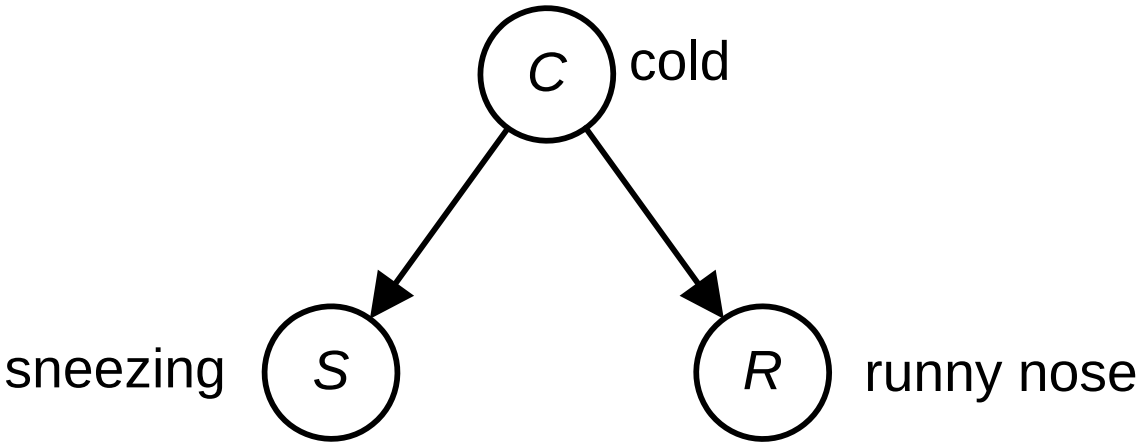
If we assume the observed probability distribution *P* of a set of observed random variables *V* satisfies the Markov condition with the causal DAG *G* containing the variables,

1. We say we are making the **causal Markov assumption**.
2. We call (*G*,*P*) a causal network.

A probability distribution *P* satisfies the **Markov condition** with a DAG *G* if the probability of each variable/node in the DAG is independent of its nondescendents conditional on its parents.

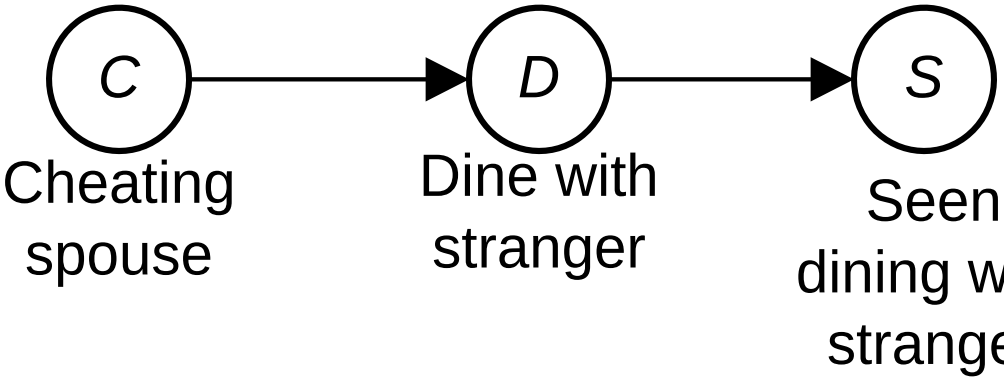
Examples

In these examples, I use a capital letter (e.g. *C*) to denote both a binary variable and one value of the variable.



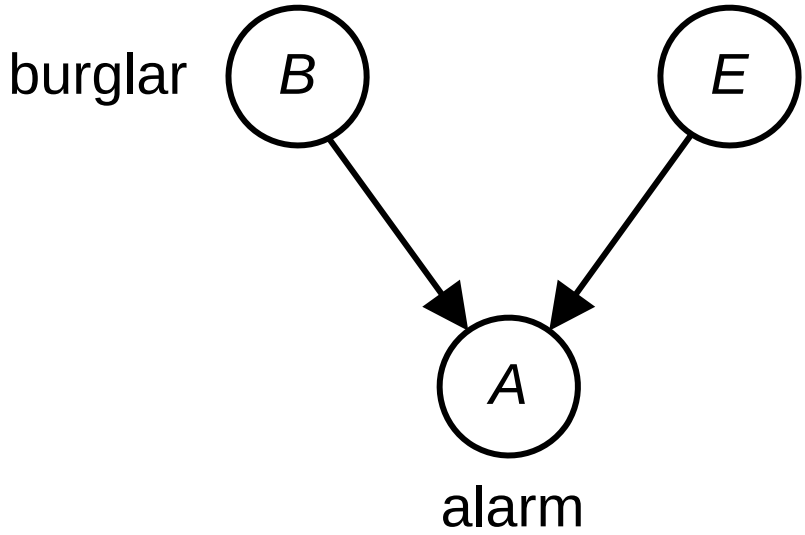
$P(S \mid R) > P(S)$ $\quad\quad$ $\neg I(S,R)$

$P(S \mid R,C) = P(S \mid C)$ $\quad\quad$ $I(S,R \mid C)$



$P(C \mid S) > P(C)$ $\quad\quad$ $\neg I(C,S)$

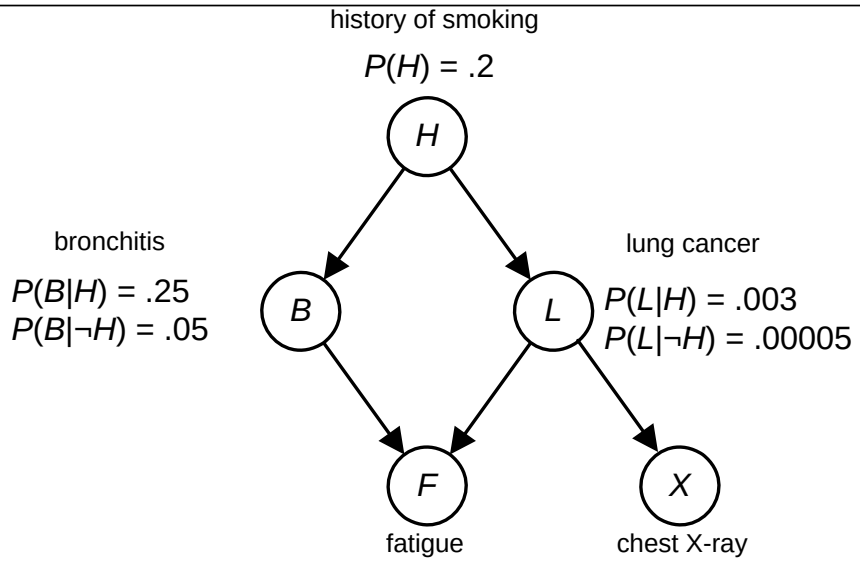
$P(C \mid S,D) = P(C \mid D)$ $\quad\quad$ $I(C,S \mid D)$



$P(B \mid E) = P(B)$ $\quad I(B,E)$

$P(B \mid E,A) < P(B \mid A)$ $\quad \neg I(B,E \mid A)$

E 'discounts' *B*.



$P(B \mid L) > P(B)$

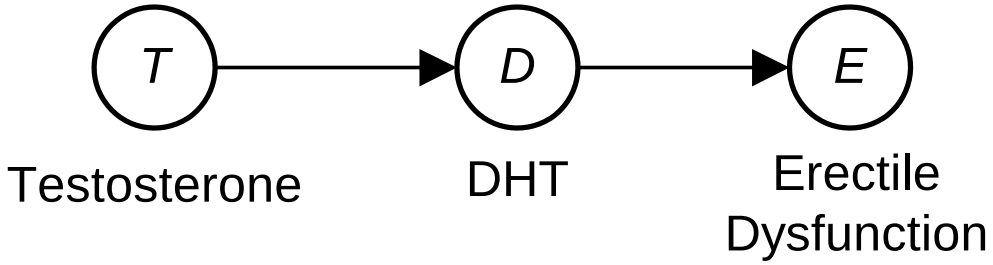
$P(B \mid X) > P(B)$

$P(B \mid L,H) = P(B \mid H)$

$P(B \mid X,H) = P(B \mid H)$

$I(B,\{L,X\} \mid H)$

Experimental evidence for the Causal Markov Assumption:

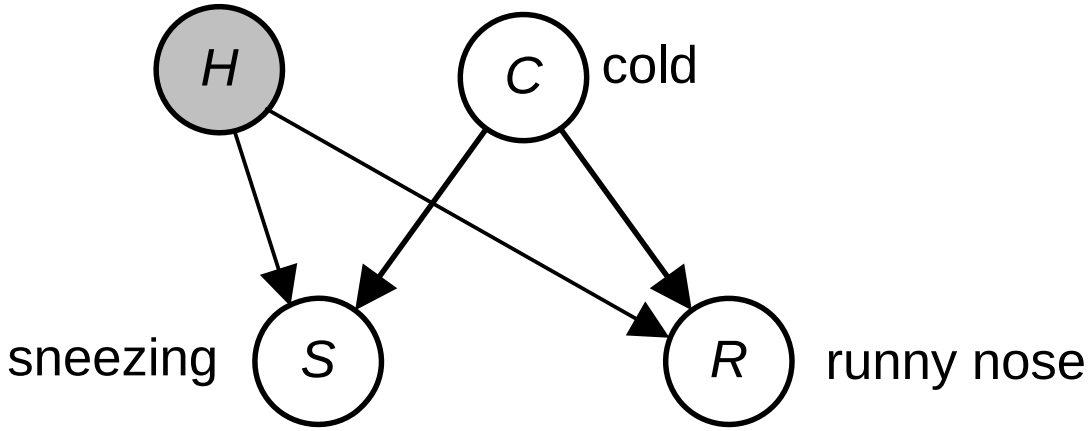


$$I(E,T \mid D)$$

Study in rats by Lugg et al. [1995]

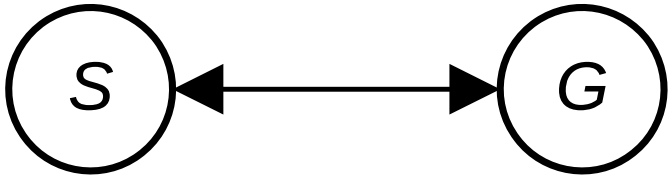
Exceptions to the Causal Markov Assumption

1. Hidden common causes



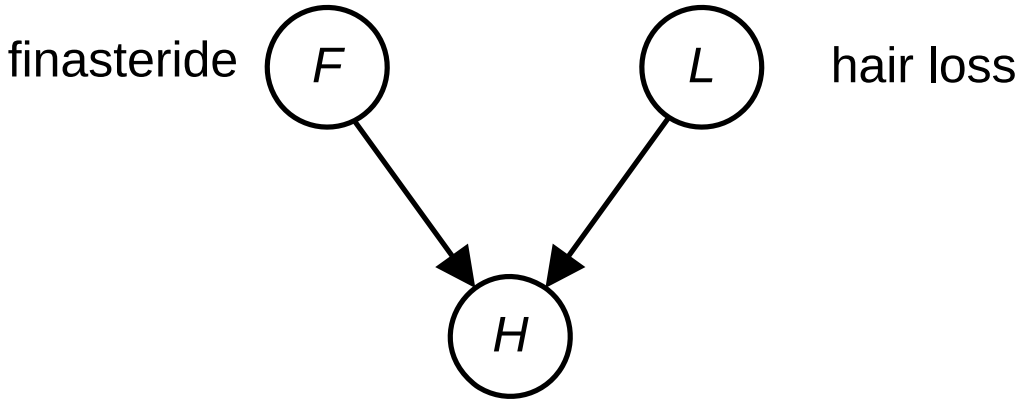
$\neg I(S,R|C)$

2. Causal feedback



Studying Good Grades

3. Selection bias

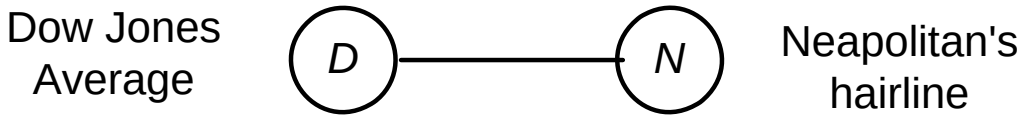


- We obtain data on only *F* and *L*.
- Everyone in our sample suffers from hypertension.
- Finasteride discounts hair loss.
- ¬*I*(*F*,*L*) in our observed distribution.

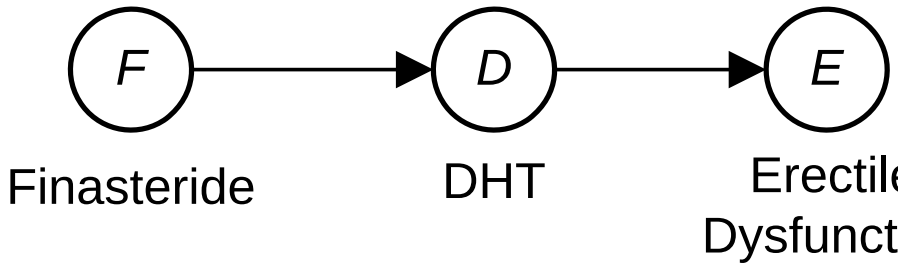
4. The entities in the population are units of time

causal DAG

statistical correlation



- Perhaps the condition that is most violated is that there can be no hidden common causes.
- We will come back to this.



There is a conditional independency that is not entailed by the Markov condition.

$I(E,F)$

Causal Faithfulness Assumption

If we assume

1. the observed probability distribution *P* of a set of observed random variables *V* satisfies the Markov condition with the causal DAG *G* containing the variables,
2. All conditional independencies in the observed distribution *P* are entailed by the **Markov condition** in *G,*

we say we are making the **causal faithfulness assumption**.

Exceptions to the Causal Faithfulness Assumption:

- All the exceptions to the causal Markov assumption.
- 'Unusual' causal relationships as in the finasteride example.

Learning Causal Influences Under the Causal Faithfulness Assumption

- In what follows we assume we have a large amount of data on the variables of interest.
- We learn conditional independencies from these data.

Example: From the data on the right we learn

$P(Y=1|X=1) = P(Y=1|X=2)$

$I(X,Y)$.

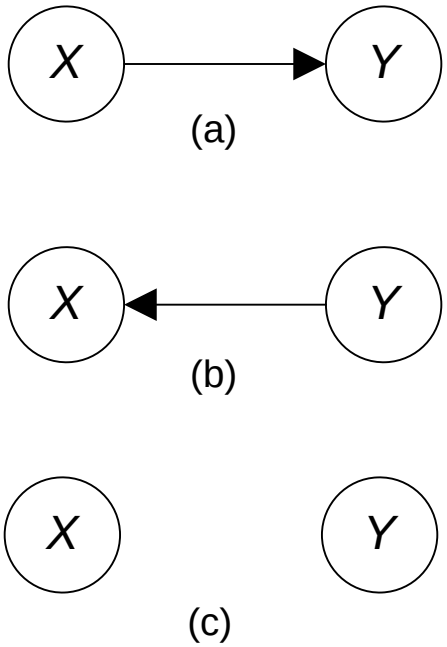
X	Y
1	1
1	1
1	2
1	2
2	1
2	1
2	2
2	2

- How much data do we need?
- We will come back to this.
- In what follows we will assume we have learned the conditional independencies for certain.

Example 1. Suppose $V = \{X, Y\}$ and our set of conditional independencies is

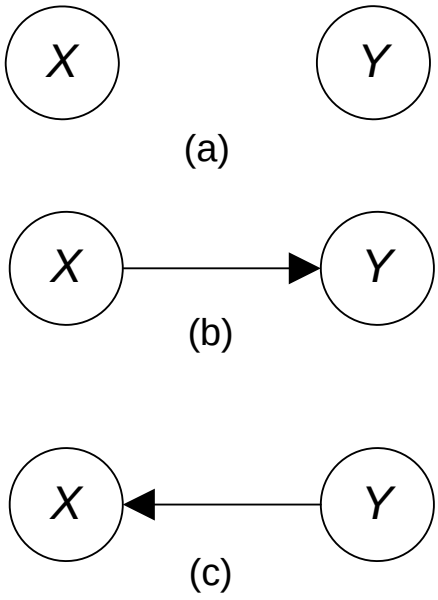
$$\{I(X, Y)\}.$$

- Then we cannot have the causal DAG in (a) or in (b).
- The Markov condition, applied to these DAGs, does not entail that X and Y are independent, and this independency is present.
- So we must have the causal DAG in (c).



Example 2. Suppose $V = \{X, Y\}$ and our set of conditional independencies is the empty set.

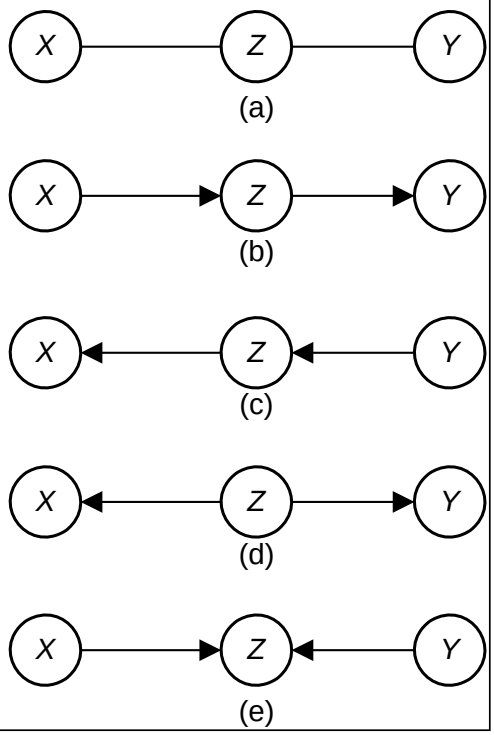
- Then we cannot have the causal DAG in (a).
- The Markov condition, applied to that DAG, entails that X and Y are independent, and this independency is not present.
- So we must have the causal DAG in (b) or in (c).



Example 3. Suppose $V = \{X, Y, Z\}$ and our set of conditional independencies is

$$\{I(X, Y)\}.$$

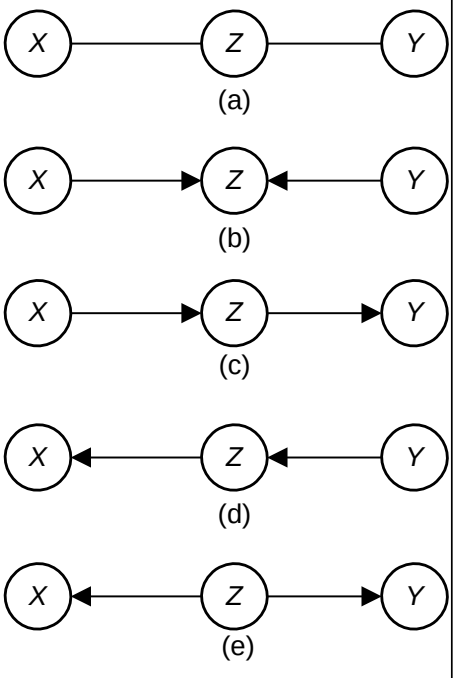
- There can be no edge between X and Y in the causal DAG owing to the reason in Example 1.
- There must be edges between X and Z and between Y and Z owing to the reason in Example 2.
- So we must have the links in (a).
- We cannot have the causal DAG in (b) or in (c) or in (d).
- The reason is that Markov condition entails $I(X, Y|Z)$, and this conditional independency is not present.
- Furthermore, the Markov condition does not entail $I(X, Y)$.
- So we must have the causal DAG in (e).



Example 4. Suppose $V = \{X, Y, Z\}$ and our set of conditional independencies is

$$\{I(X, Y \mid Z)\}.$$

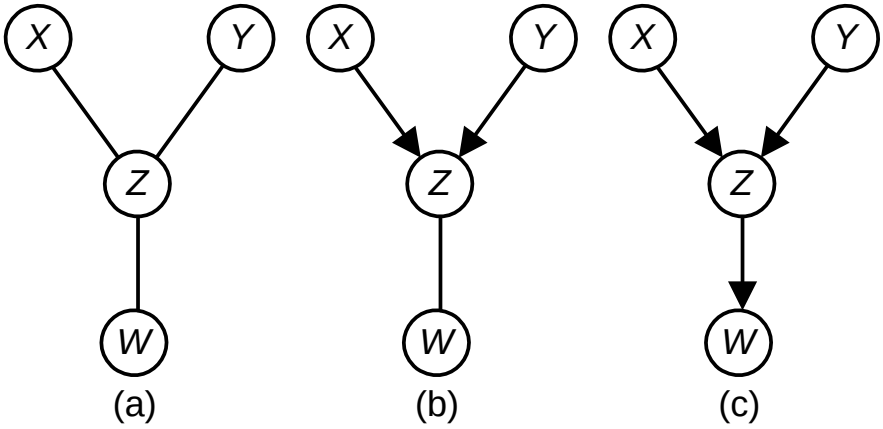
- Then, as before, the only edges in the causal DAG must be between X and Z and between Y and Z.
- So we must have the links in (a).
- We can't have the causal DAG in (b).
- The Markov condition, applied to that DAG, entails $I(X,Y)$, and this conditional independency is not present.
- Furthermore, the Markov Condition does not entail $I(X, Y \mid Z)$.
- So we must have the causal DAG in (c) or in (d) or in (e).



Theorem: If (*G*,*P*) satisfies the faithfulness condition, then there is an edge between *X* and *Y* if and only if *X* and *Y* are not conditionally independent given any set of variables.

Example 5. Suppose $V = \{X, Y, Z, W\}$ and our set of conditional independencies is

$$\{I(X, Y), \quad I(W, \{X, Y\} \mid Z)\}.$$



- Due to the previous theorem, the links must be those in (a).
- We must have the arrow directions in (b) because $I(X, Y)$.
- Therefore, we must have the arrow directions in the (c) because we do not have $I(W, X)$.

How much data do we need?

- In the limit with probability 1 we will learn the correct DAG.
- [Oz et al., 2006] obtain bounds on the number of records needed to be confident we will not learn a particular wrong DAG.
- As far as I know, there are no bounds on the number of records needed to be confident we will not learn any wrong DAG.

Empirical Results

[Dai et al., 1997]:

# of Nodes	# Data Items Needed to Learn Correct DAG
2	10
3	200
4 (undirected cycle in one)	1000-5000
5	1000-2000

Stronger links (greater dependencies) are easier to learn.

Conflicting Empirical Results

- [Cooper and Herskovits, 1992] correctly learned the 37 node Alarm Network from 3000 data items, which were generated using the network.
- [Neapolitan and Morris, 2003] obtained goofy results when trying to learn an 8 node network from 9640 records.
  - I.e., they learned that whether an individual in the military holds the military responsible for a racial incident has a causal effect on the individual's race.

Relaxing the Assumption of No Hidden Common Causes

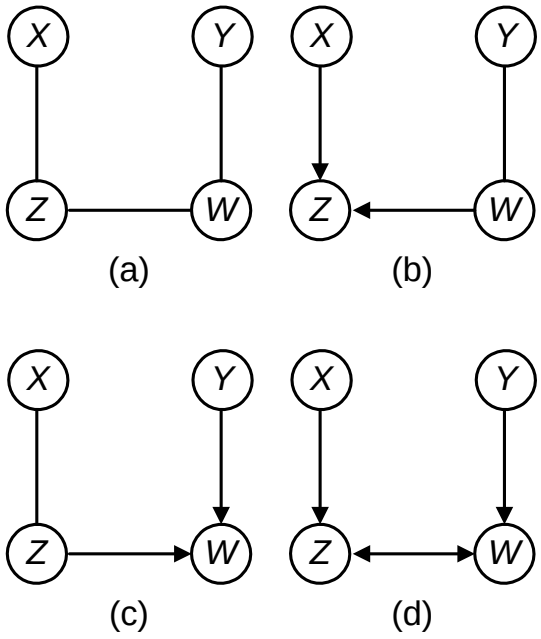
- It seems the main exception to the causal Markov (and therefore causal faithfulness) assumption is the presence of hidden common causes.
- In most domains it does seem reasonable to assume that there can be no hidden common causes.
- It is still possible to learn some causal influences if we relax this assumption.
- The causal embedded faithfulness assumption allows for hidden common causes.

- Sometimes the data tells us that there must be a hidden common cause.
- In such cases we know the causal faithfulness assumption is not warranted.

Example 6. Suppose $V = \{X, Y, Z, W\}$ and our set of conditional independencies is

$$\{I(X, \{Y, W\}), \quad I(Y, \{X, Z\}).$$

- We must have the links in (a).
- We must have the arrow directions in (b) because $I(X,\{Y,W\})$,
- We must have the arrow directions in (c) because $I(Y,\{X,Z\})$.
- We end up with the graph in (d).
- There is no DAG faithful to this probability distribution.
- The causal faithfulness assumption is not warranted.



How could this happen?

- Suppose the causal DAG in (a) satisfies the causal faithfulness assumption.
- Then we do not have

$$I(Z,W).$$

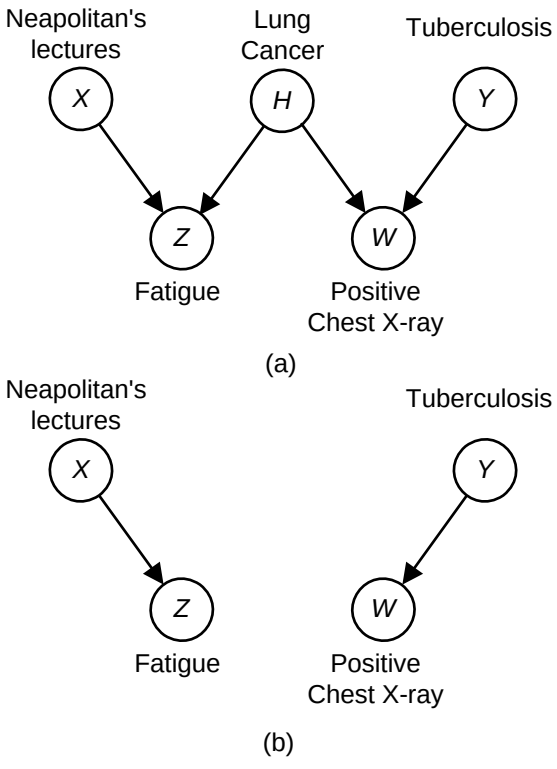
- Suppose further we only observe

$$V = \{X, Y, Z, W\}.$$

- The causal DAG containing the **observed variables** is then the one in (b). This causal DAG entails

$$I(Z,W).$$

- Since the **observed distribution** does not satisfy the Markov condition with the causal DAG containing the **observed variables**, the causal Markov assumption is not warranted.
- Problem is that there is a hidden common cause H.



Causal Embedded Faithfulness Assumption

- If we assume the observed distribution *P* of the variables is **embedded faithfully** in a causal DAG containing the variables, we say we are making the **causal embedded faithfulness assumption**.
- Suppose we have a probability distribution *P* of the variables in *V*, *V* is a subset of *W*, and *G* is a DAG whose set of nodes is *W*.

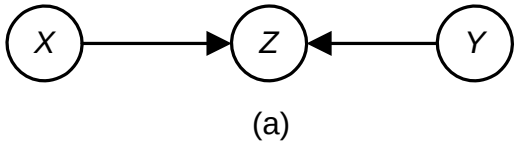
  Then *P* is **embedded faithfully** in *W* if all and only the conditional independencies in *P* are entailed by the Markov condition applied to *W* and restricted to the nodes in *V*.

Basically, when we make the causal embedded faithfulness assumption, we are assuming that there is a causal DAG with which the observed distribution is faithful, but that DAG might contain hidden variables.

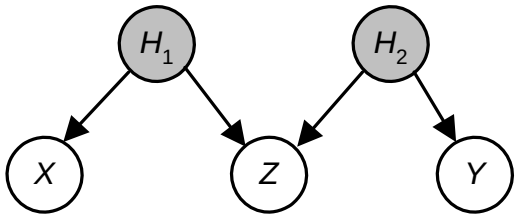
Learning Causal Influences Under the Causal Embedded Faithfulness Assumption

Example 3 revisited. Suppose $V = \{X, Y, Z\}$ and our set of conditional independencies is

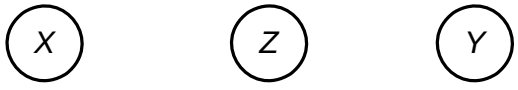
$$\{I(X, Y)\}.$$



(a)



(b)

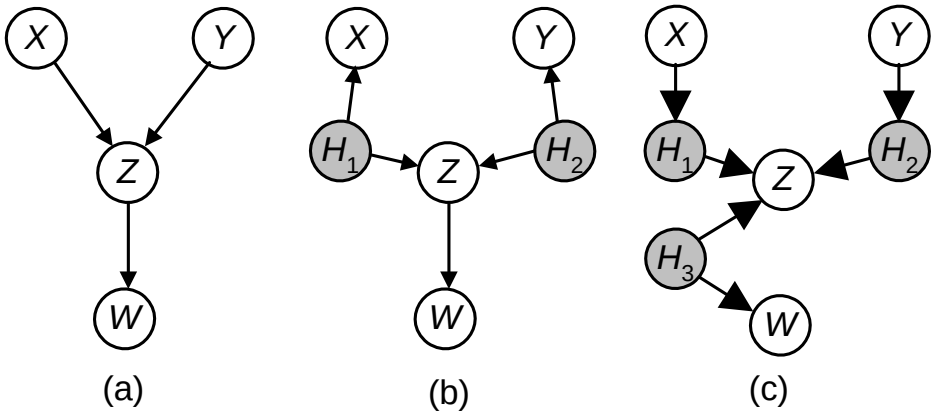


(c)

- The probability distribution is embedded faithfully in the DAG in (a) and in (b).
- Note: d-separation implies $I(X, Y)$ for the DAG in (b).
- So the causal relationships among the observed variables may be those shown on the bottom.
- We cannot learn any causal influences.

Example 5 revisited. Suppose $V = \{X, Y, Z, W\}$ and our set of conditional independencies is

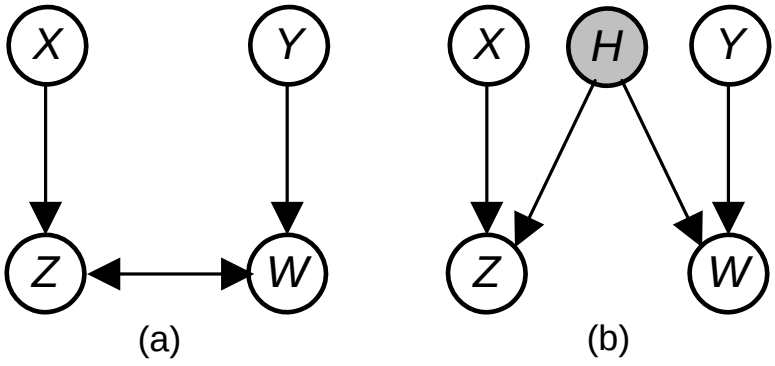
$$\{I(X, Y), \quad I(W, \{X, Y\} \mid Z)\}.$$



- P is embedded faithfully in the DAG in (a) and in (b).
- P is not embedded faithfully in the DAG in (c). That DAG entails $I(W,\{X,Y\})$.
- We conclude Z causes W.
- So we can learn causes from data on four variables.

Example 6 revisited. Suppose $V = \{X, Y, Z, W\}$ and our set of conditional independencies is

$$\{I(X, \{Y, W\}), \quad I(Y, \{X, Z\}).$$



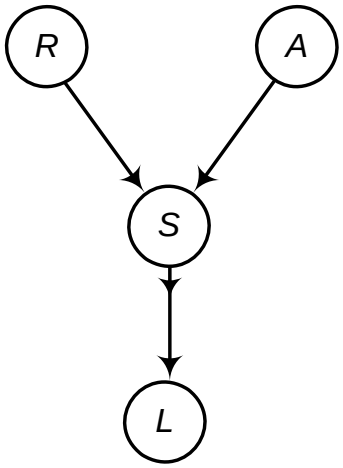
- Recall we obtained the graph in (a) when we tried to find a faithful DAG.
- P is embedded faithfully in the DAG in (b).
- We conclude Z and W have a hidden common cause.

Example 7. Suppose we have these variables:

R: Parent's smoking history
A: Alcohol consumption
S: Smoking behavior
L: Lung cancer,

and we learn the following from data:

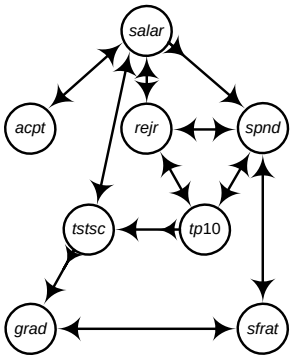
$$\{I(R, A), \quad I(L, \{R, A\} \mid S)\}$$



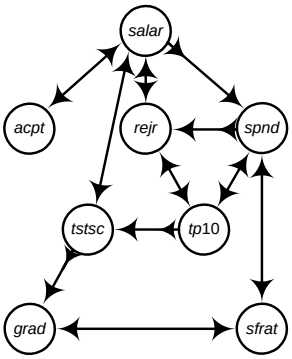
- We conclude the graph on the right.
- The one-headed arrow means *R* causes *S* or they have a hidden common cause.
- The two-headed arrow means *S* causes *L*.

Example 8. University Student Retention (Druzdzel and Glymour, 1999)

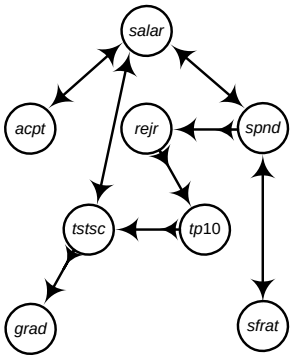
Variable	**What the Variable Represents**
grad	Fraction of students who graduate from the institution
rejr	Fraction of applicants who are not offered admission
tstsc	Average standardized score of incoming students
tp10	Fraction of incoming students in the top 10% of high school class
acpt	Fraction of students who accept the institution's admission offer
spnd	Average educational and general expenses per student
sfrat	Student/faculty ratio
salar	Average faculty salary



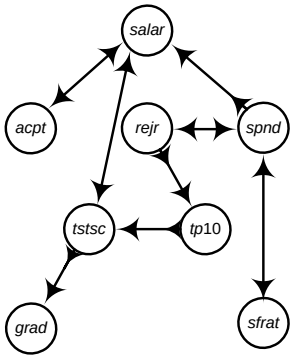
alpha = .2



alpha = .1



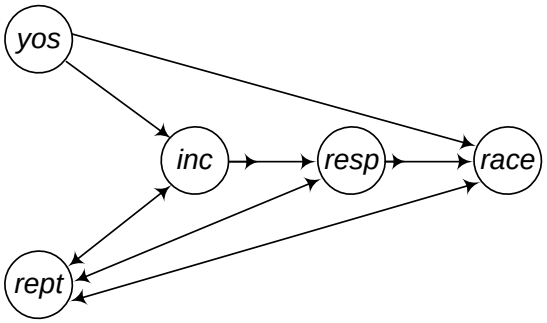
alpha = .05

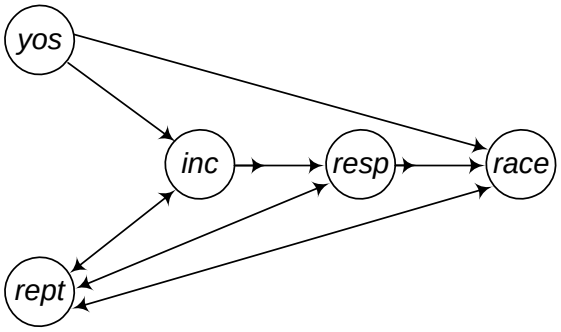


alpha = .01

Example 9. Racial Incidents in the U.S. Military (Neapolitan and Morris, 2003)

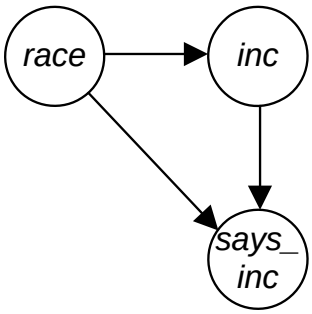
Variable	What the Variable Represents
race	Respondent's race/ethnicity
yos	Respondent's years of military service
inc	Whether respondent replied he/she experienced a racial incident
rept	Whether incident was reported to military personnel
resp	Whether respondent held the military responsible for the incident





- The causal influence of *resp* on *race* would not be learned if there was a direct causal influence of *race* on *inc*.
- It seems suspicious that *race* does not have a causal influence on *inc*.
- These are probabilistic relationships among responses; they are not necessarily the probabilistic relationships among the actual events.
- A problem with using survey responses to represent occurrences in nature is that subjects may not respond accurately.

The actual causal relationships may be as follows:



inc: Whether there really was an incident.
says-inc: The survey response.

- It could be that races which experience higher rates of harassment are less likely to report racial incidents.
- So the causal effect of *race* on *says_inc* through *inc* is negated by the direct influence of *race* on *says_inc*.
- This would be case in which embedded faithfulness is violated similar to the Finasteride/DHT/ED example.
- Stangor et al. [2002] found minority members are less likely to report discrimination publicly.

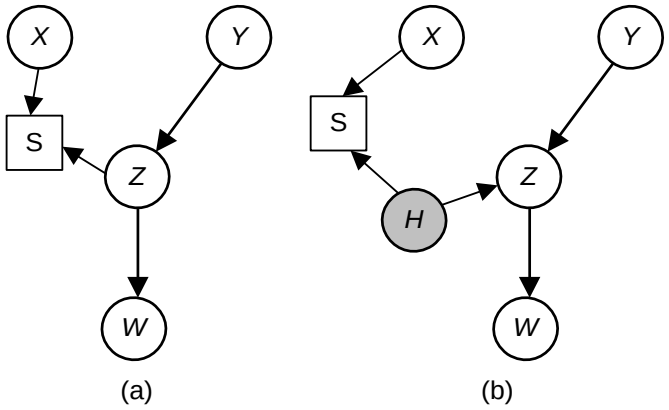
The Causal Embedded Faithfulness Assumption with Selection Bias:

If we assume the probability distribution P of the observed variables is embedded faithfully in a causal DAG containing the variables, but that possibly selection bias is present when we sample, we say we are making the **causal embedded faithfulness assumption**.

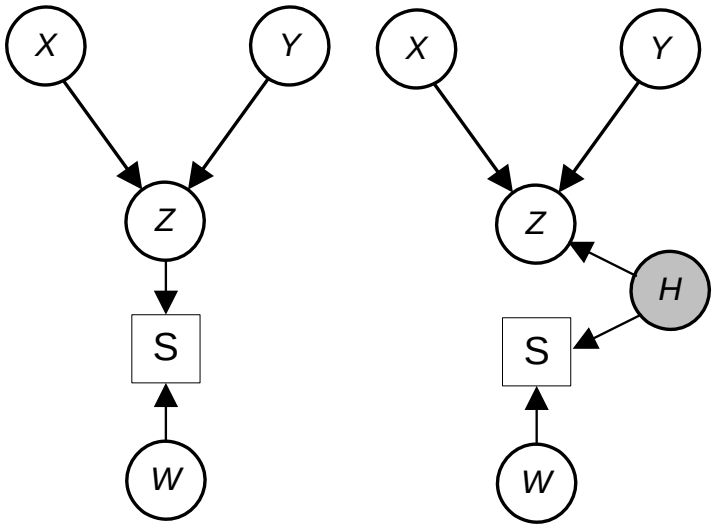
Example 5 (revisited). Suppose $V = \{X, Y, Z, W\}$ and the set of conditional independencies in the observed distribution P_{obs} is

$$\{I(X, Y), \quad I(W, \{X, Y\} \mid Z)\}.$$

Suppose further selection bias may be present.



- The causal DAG could not be the one in (a) because we would not have $I(X,Y)$ in P_{obs}.
- The causal DAG could be the one in (b).

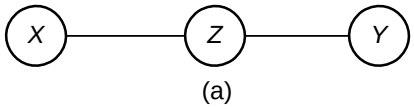


- The causal DAG could not be the one in (a) because we would not have $I(X,Y)$ in P_{obs}.
- The causal DAG could not be the one in (b) because we would then have $I(W,\{X,Y\})$ in P_{obs}.
- So we can still conclude Z has a causal influence on W.

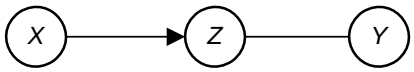
Causal Learning with Temporal Information

Example 4 (revisited). Suppose $V = \{X, Y, Z\}$ and our set of conditional independencies is

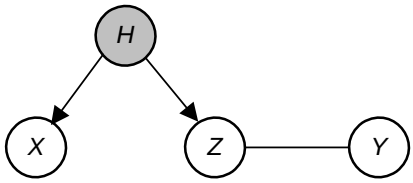
$$\{I(X, Y \mid Z)\}.$$



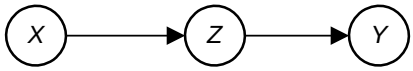
(a)



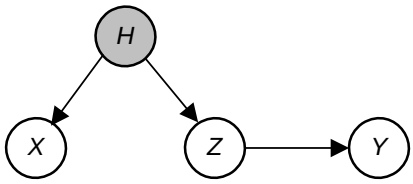
(b)



(c)



(d)



(e)

- Assuming embedded faithfulness, we must have the links in (a).
- Suppose there can be no causal path from X to Z (e.g. we know X precedes Z in time).
- Then the link between X and Z must be either that in (b) or in (c).
- So the causal DAG must be either that in (d) or in (e).
- With temporal information we can learn causes from data on only three variables.

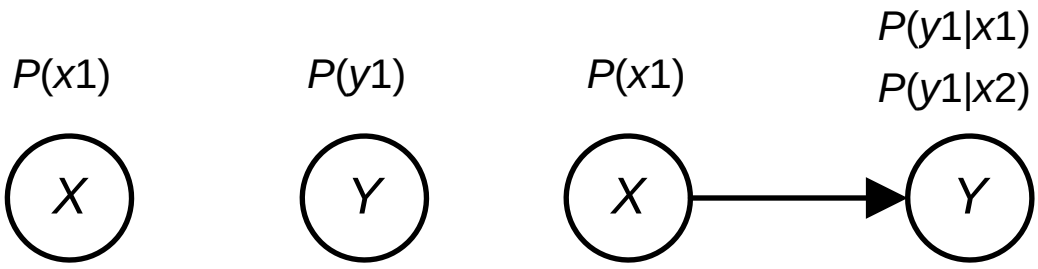
Learning Causes From Data on Two Variables

- One way to learn DAG structure from data is by scoring DAGs.
- The Bayesian Score:
  - Choose DAG that maximizes $P(data|\text{DAG})$.
- For certain classes of models, a smallest DAG model that includes the probability distribution will (with high probability) be chosen when the data set is large.

Example 1. Suppose $V = \{X, Y\}$, both variables are binary, and our set of conditional independencies is

$$\{I(X, Y)\}.$$

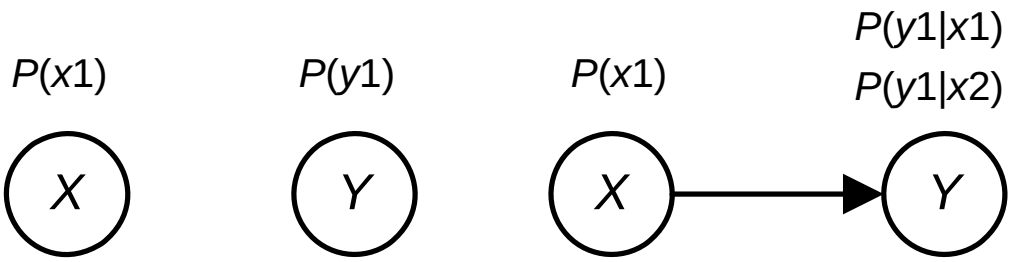
Possible DAG models are as follows:



- Both models include P.
- The model on the left will be chosen because it is smaller.

Example 2. Suppose $V = \{X, Y\}$, both variables are binary, and our set of conditional independencies is the empty set.

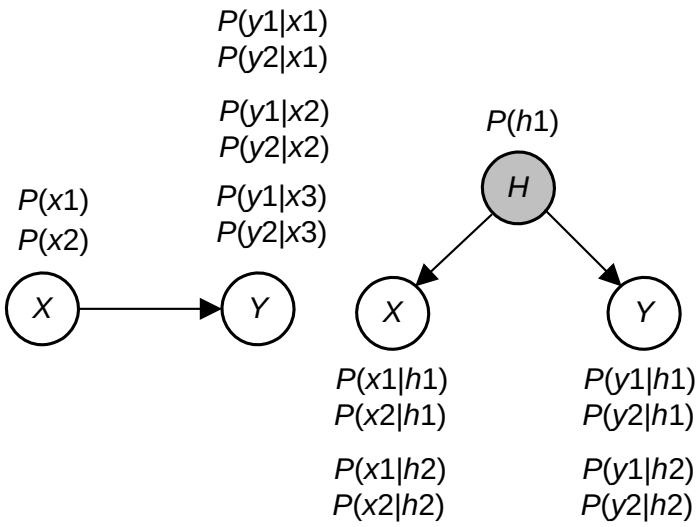
Possible DAG models are as follows:



- Only the model on the right includes P.
- So the model on the right will be chosen.

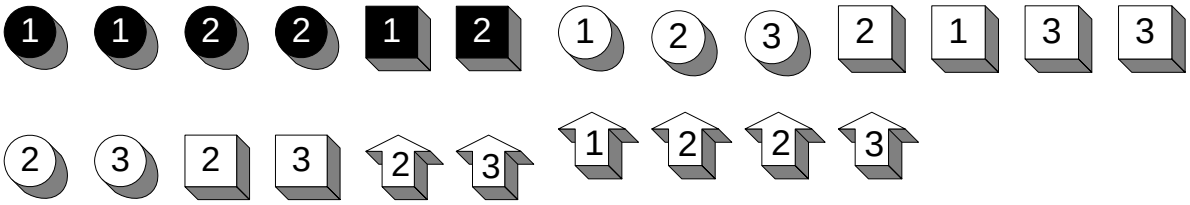
Example 3. Suppose $V = \{X, Y\}$, both variables have space size 3, and our set of conditional independencies is the empty set.

Consider these two models.



Although the hidden variable model appears larger, it is actually smaller because it has fewer effective parameters.

- The following is an intuitive explanation.
- Clearly the model without the hidden variables includes any probability distribution which the hidden variable model includes.
- However, the hidden variable model only includes distributions which can be represented by urn problems in which there is a division of the objects which renders *X* and *Y* independent.



Value and shape are independent given color.

There is no apparent division of the objects into two groups which renders value and shape independent.

- In the space consisting of all possible assignments of values to the parameters in the DAG model, the subset, whose distributions are included in the hidden variable model, has measure zero.
- Consider this experiment:
  1. Randomly generate one of the models.
  2. Randomly assign parameter values to the model chosen.
  3. Generate a large amount of data.
  4. Score the models using the data.
- When the DAG model is chosen, almost certainly the probability distribution will not be included in the hidden variable model. So the DAG model will with high probability score higher.
- When the hidden variable model is generated, with high probability it will score higher because it is smaller.

Application to Causal Learning

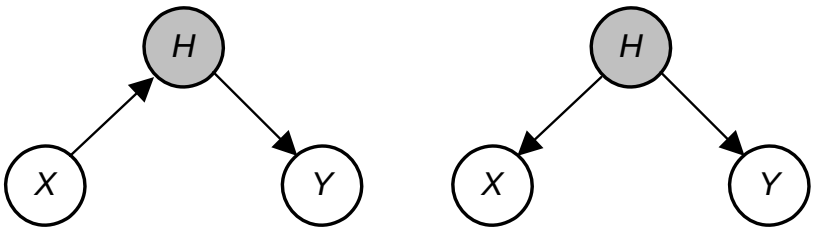
Suppose the entire population is distributed as follows:

• Case	• Sex	• Height	• Wage ($)
• 1	• F	• 64	• 30,000
• 2	• F	• 64	• 30,000
• 3	• F	• 64	• 40,000
• 4	• F	• 64	• 40,000
• 5	• F	• 68	• 30,000
• 6	• F	• 68	• 40,000
• 7	• M	• 64	• 40,000
• 8	• M	• 64	• 50,000
• 9	• M	• 68	• 40,000
• 10	• M	• 68	• 50,000
• 11	• M	• 70	• 40,000
• 12	• M	• 70	• 50,000

black/F white/M circle/64 square/68
arrow/70 1/30,000 2/40,000 3/50,000

- Suppose we only observe and collect data on height and wage.
- Sex is then a hidden variable in the sense that it renders the observed variables independent.
- If we only look for correlation, we would find height and wage are correlated and perhaps conclude height has a causal effect on wage.
- If we score both the DAG model and the hidden variable model, the hidden variable model will most probably win. We can conclude that possibly there is a hidden common cause.

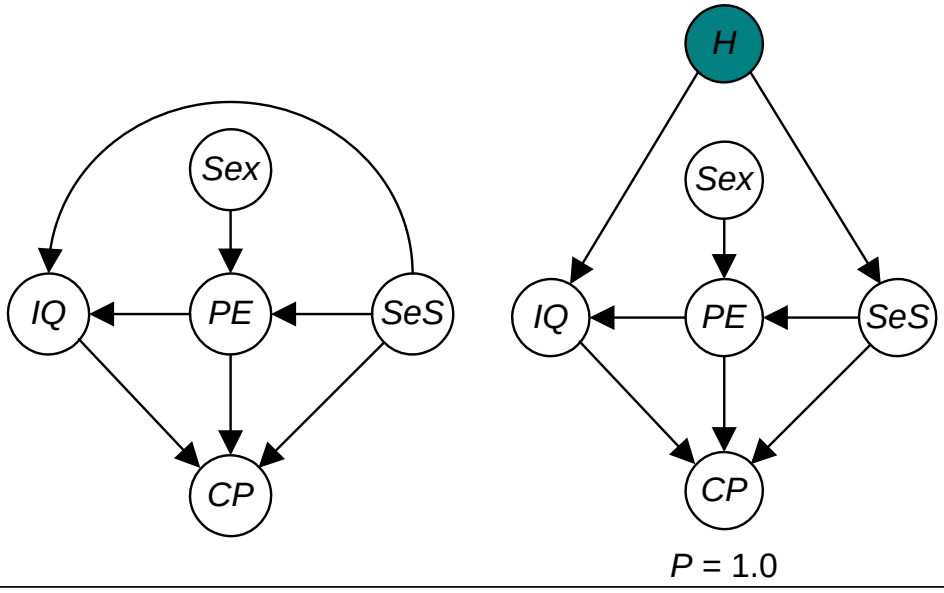
Some Caveats



1. The hidden variable models above have the same score. So we may have a hidden intermediate cause instead of a hidden common cause.
2. In real applications features like height and wage are continuous.
   - We create discrete values by imposing cutoff points.
   - With different cutoff points we may not create a division which renders wage and height independent.

3. Similar caution must be used if the DAG model wins and we want to conclude there is no hidden common cause.
   - Different cutoff points may result in a division of the objects which renders them independent, which means the hidden variable model would win.
   - If there is a hidden common cause, it may be modeled better with a hidden variable that has a larger space.
   - Clearly, if we increase the space size of *H* sufficiently the hidden variable model will have the same size as the DAG model.

- At one time researchers were excited about the possibilities.
- Learning college attendance influences [Heckerman et al., 1999].



P = 1.0

References

Cooper, G.F., and E. Herskovits [1992], "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Machine Learning*, 9.

Dai, H., K. Korb, C. Wallace, and X. Wu [1997], "A Study of Causal Discovery With Weak Links and Small Samples," *Proceedings of IJCAI-97*, Nagoya, Japan.

Druzdzel, M.J., and C. Glymour [1999], "Causal Inference from Databases: Why Universities Lose Students," in Glymour, C., and G.F. Cooper (Eds.): *Computation, Causation, and Discovery*, AAAI Press, Menlo Park, CA.

Heckerman, D., C. Meek, and G.F. Cooper [1999], "A Bayesian Approach to Causal Discovery," in Glymour, C., and G.F. Cooper (Eds.): *Computation, Causation, and Discovery*, AAAI Press, Menlo Park, CA.

Lugg, J.A., Raifer J., and C.N.F. González [1995], "Dehydrotestosterone is the Active Androgen in the Maintenance of Nitric Oxide-Mediated Penile Erection in the Rat," *Endocrinology*, 136(4).

Neapolitan, R.E. [1990], *Probabilistic Reasoning in Expert Systems*, Wiley, New York.

Neapolitan, R. E. [2004], *Learning Bayesian Networks*, Prentice Hall, Upper Saddle River, NJ.

Neapolitan, R.E., and S. Morris [2003], "Probabilistic Modeling Using Bayesian Networks," in D. Kaplan (Eds.): *Handbook of Quantitative Methodology in the Social Sciences*, Sage, Thousand Oaks, CA.

Pearl, J. [1988], *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA.

Pearl, J. [2000], *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, UK.

Spirtes, P., C. Glymour, and R. Scheines [1993, 2000], *Causation, Prediction, and Search*, Springer-Verlag, New York, 1993; 2nd ed.: MIT Press, Cambridge, MA.

Zuk, O., S. Margel, and E. Domany [2006], "On the Number of Samples Needed to Learn the Correct Structure of a Bayesian Network," *Proceedings of UAI 2006*.