Learning Bayesian Networks and Causal Discovery

Marek J. Drużdżel

Politechnika Białostocka
Wydział Informatyki

[*m.druzdzel@pb.edu.pl*](mailto:m.druzdzel@pb.edu.pl)
[*http://www.wi.pb.edu.pl/~druzdzel/*](http://www.wi.pb.edu.pl/~druzdzel/)

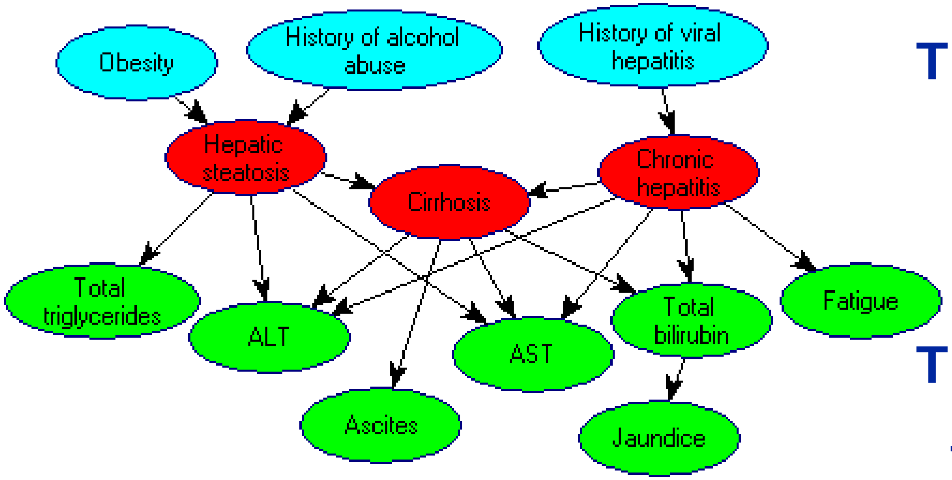
Overview

- Motivation
- Constraint-based learning
- Bayesian learning
- Example
- Software demo
- Concluding remarks

(Essentially, a handful of slides interleaved with software demos.)

Bayesian networks

A Bayesian network (also referred to as belief network, probabilistic network, or causal network) is an acyclic directed graph (DAG) consisting of:



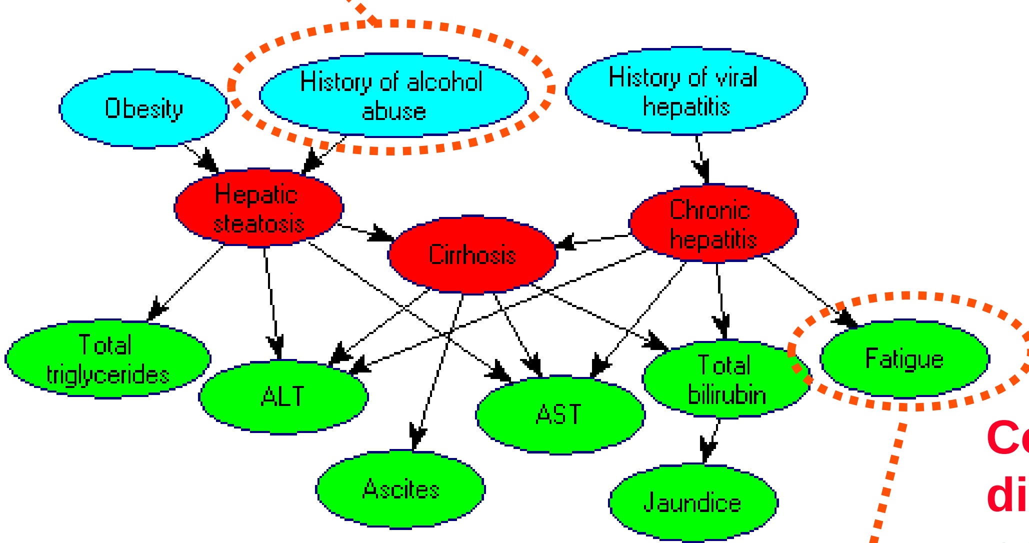
The qualitative part, encoding a domain's variables (nodes) and the probabilistic (usually causal) influences among them (arcs).

The quantitative part, encoding the joint probability distribution over these variables.

Bayesian networks: Numerical parameters

present	0.15
absent	0.85

Prior probability distribution tables for nodes without predecessors (*History of viral hepatitis*, *History of alcohol abuse, Obesity*)

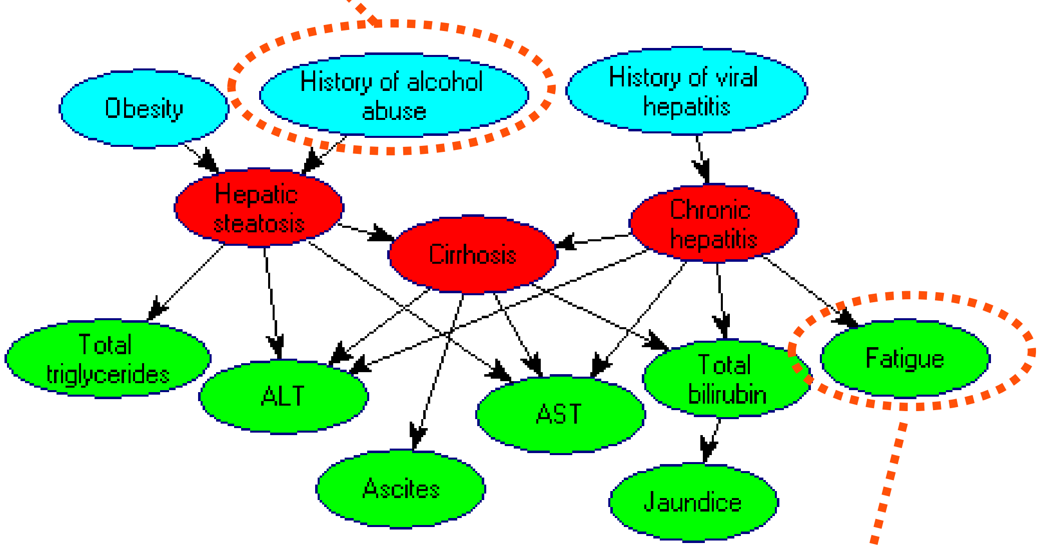


Conditional probability distributions tables for nodes with predecessors (*Fatigue*, *Jaundice*, ...)

Chronic hepatitis	present	absent
present	0.6	0.2
absent	0.4	0.8

What do the numbers come from?

present	0.15
absent	0.85

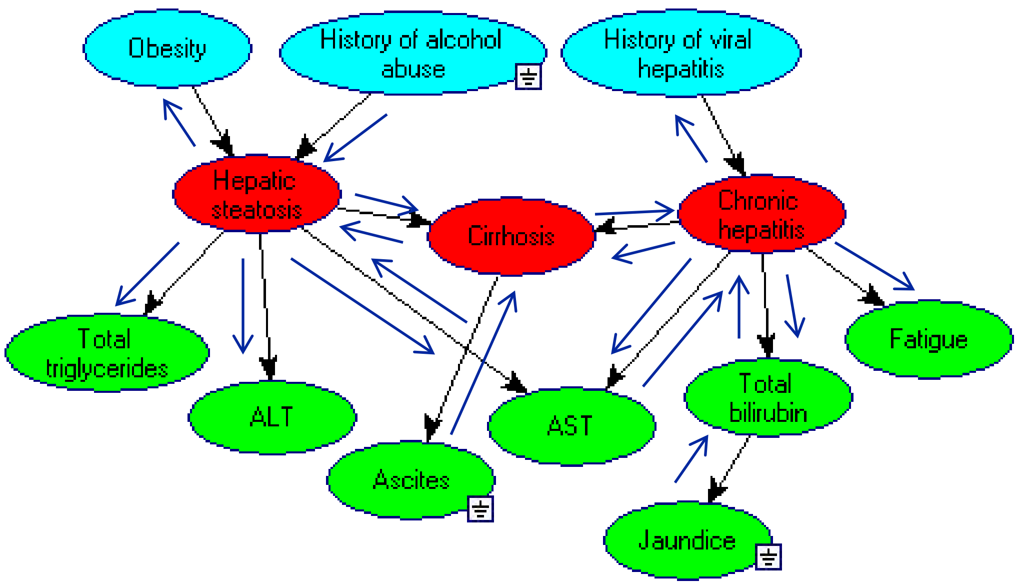


Chronic hepatitis	present	absent
present	0.6	0.2
absent	0.4	0.8

- **Textbooks**
- **Literature**
- **Expert opinion**
- **Databases**

Reasoning in Bayesian networks

The most important type of reasoning in Bayesian networks is updating the probability of a hypothesis (e.g., a diagnosis) given new evidence (e.g., medical findings, test results).



Example:

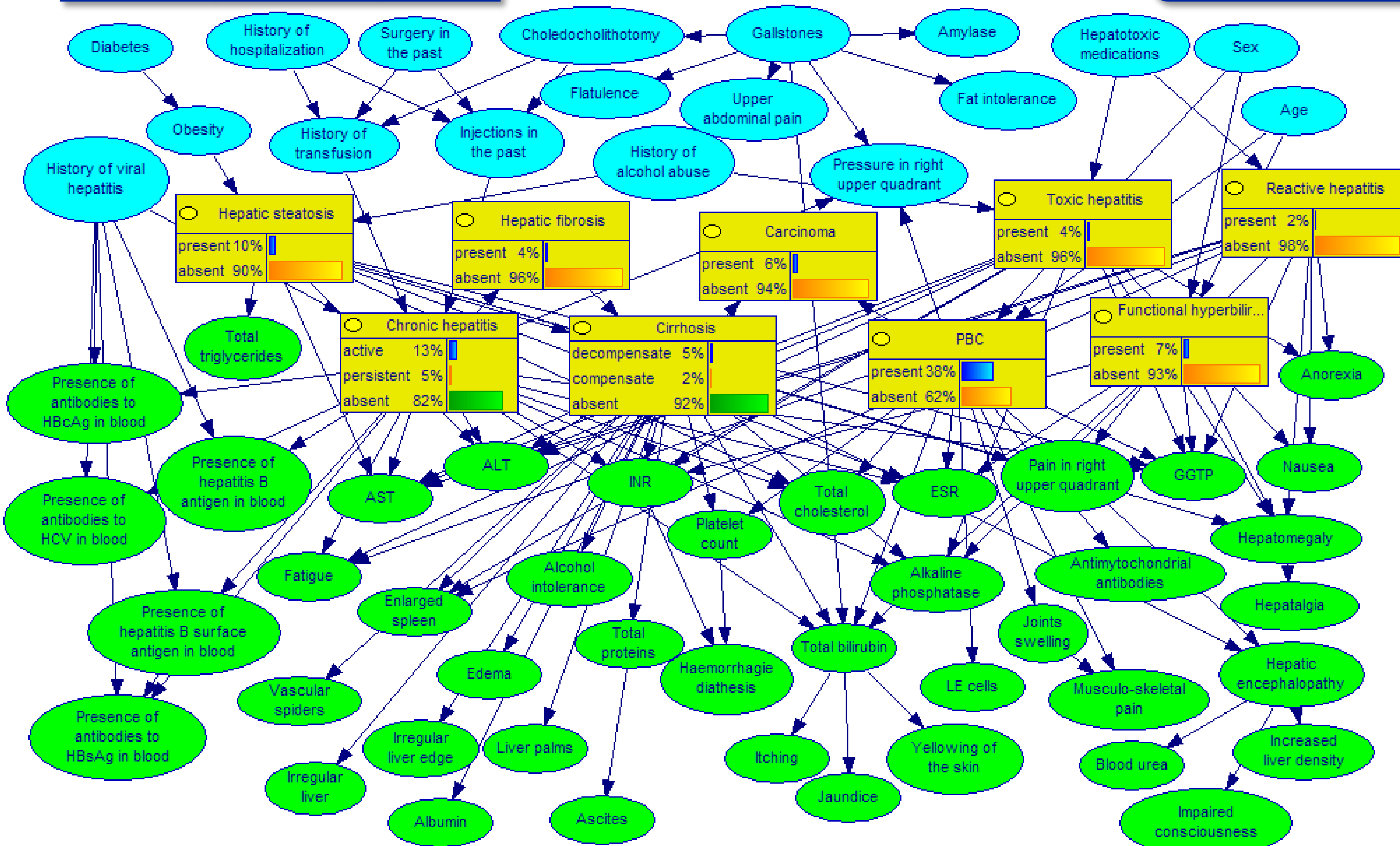
What is the probability of Chronic Hepatitis in an alcoholic patient with *jaundice* and *ascites*?

Which disease is most likely?

Which tests should we perform next?

P(Hepatitis | *alcoholism*=present, *jaundice*=present, *ascites*=present)?

Example: Hepar II



70 variables; 2,139 numerical parameters (instead of over $2^{70} \approx 10^{21}$!)

Learning Bayesian networks from data

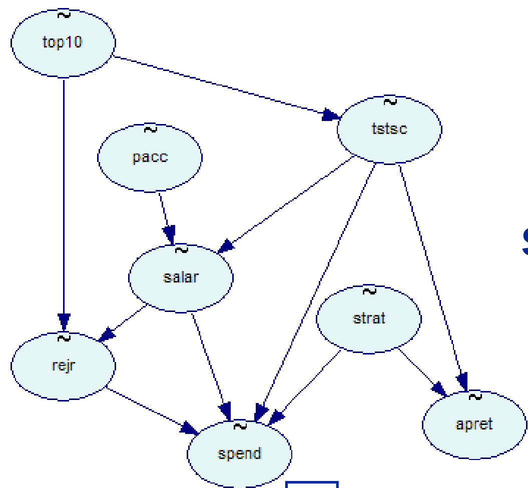
There exist algorithms with a capability to analyze data, discover causal patterns in them, and build models based on these data.

Retention.txt

spend	apret	top10	rejr	tstsc	pacc	strat	salar
9855	52.5	15	29.474	65.063	36.887	12	60800
10527	64.25	36	22.309	71.063	30.97	12.8	63900
7904	37.75	26	25.853	60.75	41.985	20.3	57800
6601	57	23	11.296	67.188	40.289	17	51200
7251	62	17	22.635	56.25	46.78	18.1	48000
6967	66.75	40	9.718	65.625	53.103	18	57700
8489	70.333	20	15.444	59.875	50.46	13.5	44000
9554	85.25	79	44.225	74.688	40.137	17.1	70100
15287	65.25	42	26.913	70.75	28.276	14.4	71738
7057	55.25	17	24.379	59.063	44.251	21.2	58200
16848	77.75	48	26.69	75.938	27.187	9.2	63000
18211	91	87	76.681	80.625	51.164	12.8	74400
21561	69.25	58	44.702	76.25	26.689	9.2	75400
20667	65	68	22.995	75.625	28.038	11	66200
10684	61.75	26	8.774	66	33.99	9.5	52900
11738	74.25	32	25.449	66.875	27.701	12	63400
10107	74	43	11.315	71	29.096	16.2	66200
7817	65.75	36	33.709	64.25	52.548	17.7	54600
7050	26	11	0	55.313	55.651	18.8	59500
9082	83.5	73	64.668	77.375	43.185	13.6	66700
11706	60	56	16.937	73.75	39.479	12.7	62100
7643	49.25	23	36.635	62.813	39.302	18.7	57700
25734	90	77	67.758	80.938	44.133	10	80200
20155	86	84	69.31	79.688	48.766	17.6	74000
29852	94.5	84	75.009	81.313	51.363	10.6	74100
7980	68.5	34	9.122	63.875	35.294	16.3	53100

Row 1 of 170

data



structure

Success	0.2
Failure	0.8

Success	Success	Failure
Good	0.4	0.1
Moderate	0.4	0.3
Poor	0.2	0.6

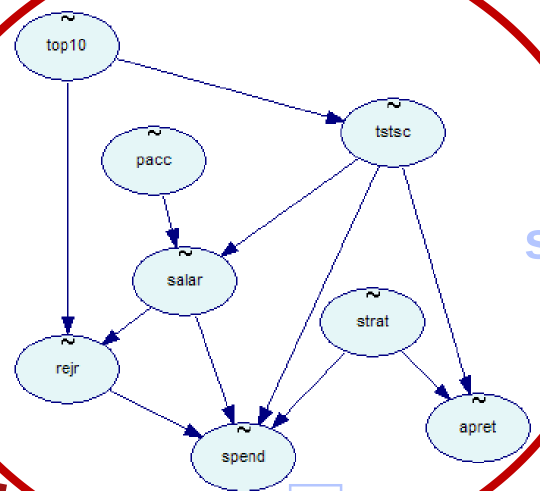
numerical parameters

Learning Bayesian networks from data

The problem of learning:
Given a set of variables (a.k.a. attributes) X and a data set D of simultaneous values of variables in X

spend	apret	top10	rejr	tstsc	pacc	strat	salar
9855	52.5	15	29.474	65.063	36.887	12	60800
10527	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12.8	63900
7904	37.75	26	25.853	60.75	41.985	20.3	57800
6601	57	23	11.296	67.188	40.289	17	51200
7251	62	17	22.635	56.25	46.78	18.1	48000
6967	66.75	40	9.718	65.625	53.103	18	57700
8489	70.333	20	15.444	59.875	50.46	13.5	44000
9554	85.25	79	44.225	74.688	40.137	17.1	70100
15287	65.25	42	26.913	70.75	28.276	14.4	71738
7057	55.25	17	24.379	59.063	44.251	21.2	58200
16848	77.75	48	26.69	75.938	27.187	9.2	63000
18211	91	87	76.681	80.625	51.164	12.8	74400
21561	69.25	58	44.702	76.25	26.689	9.2	75400
20667	65	68	22.995	75.625	28.038	11	66200
10684	61.75	26	8.774	66	33.99	9.5	52900
11738	74.25	32	25.449	66.875	27.701	12	63400
10107	74	43	11.315	71	29.096	16.2	66200
7817	65.75	36	33.709	64.25	52.548	17.7	54600
7050	26	11	0	55.313	55.651	18.8	59500
9082	83.5	73	64.668	77.375	43.185	13.6	66700
11706	60	56	16.937	73.75	39.479	12.7	62100

bility to analyze data, discover models based on these data.



structure

Success	0.2
Failure	0.8

Success	Success	Failure
Good	0.4	0.1
Moderate	0.4	0.3
Poor		

numerical parameters

data

Obtain insight into causal connections among the variables in X (for the purpose of understanding and prediction of the effects of manipulation)

Learn the joint probability distribution over the variables in X

Goal 1 (insight): Why are we interested in causality?

Reason 1: Causality allows us to predict the effects of manipulation.

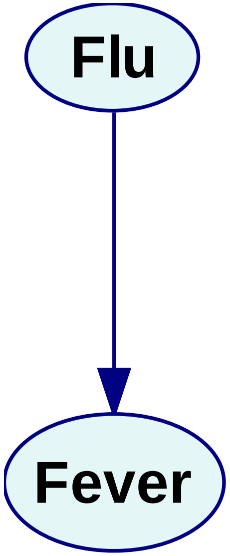
Reason 2: People (and that includes experts) think in causal terms, so it is easier to build causal models.

Given (1), is (2) really surprising?

Causality and probability

Causality and probability are closely related and their relation should be made clear in statistics.

Probabilistic dependence is considered a necessary condition for establishing causation (is it sufficient?).



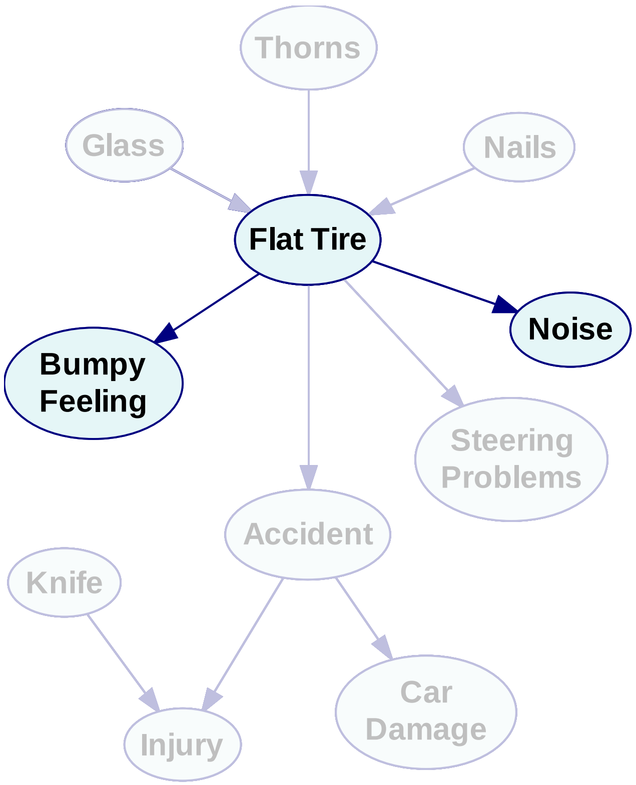
Flu and fever are correlated **because** flu may cause fever.
A cause can cause an effect but it does not have to. Causal connections result in probabilistic dependencies (or correlations in linear case).

Causal graphs

Acyclic directed graphs (hence, no time and no dynamic reasoning) representing a snapshot of the world at a given time.
Nodes are random variables and arcs are direct causal dependencies between them.

Causal connections result in *correlation* (in general *probabilistic dependence*).

- **glass on the road will be correlated with flat tire**
- **glass on the road will be correlated with noise**
- **bumpy feeling will be correlated with noise**



Causal Markov condition

An axiomatic condition describing the relationship between causality and probability.

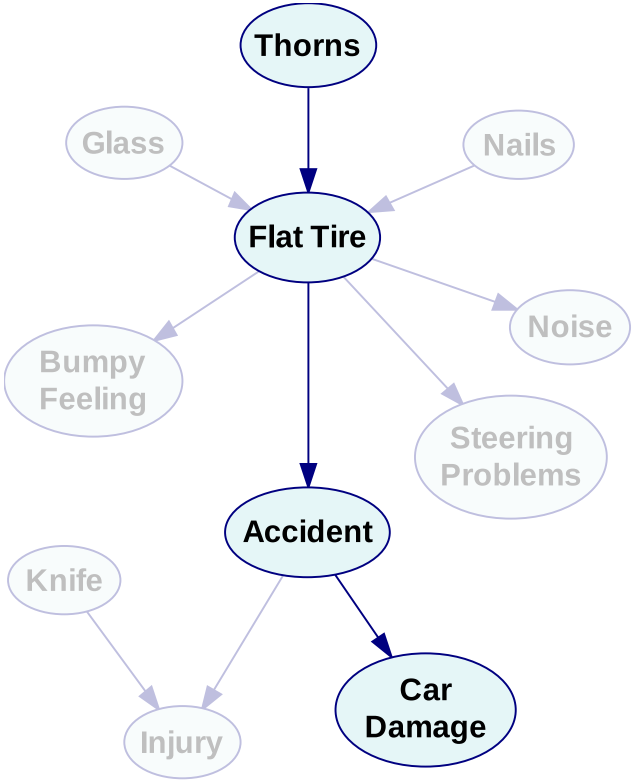
A variable in a causal graph is probabilistically independent of its non-descendants given its immediate predecessors.

Axiomatic, but used by almost everybody in practice and no convincing counter examples to it have been shown so far (at least outside the quantum world).

Markov condition: Implications

Variables A and B are probabilistically dependent if there exists a directed active path from A to B or from B to A:

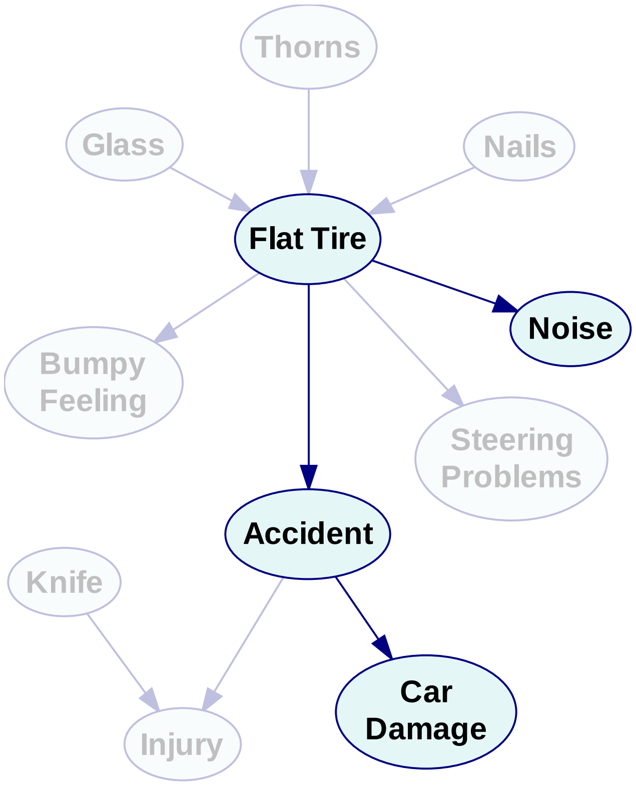
Thorns on the road are correlated with car damage because there is a directed path from thorns to car damage.



Markov condition: Implications

Variables A and B are probabilistically dependent if there exists a C such that there exists a directed active path from C to A and there exists a directed active path from C to B:

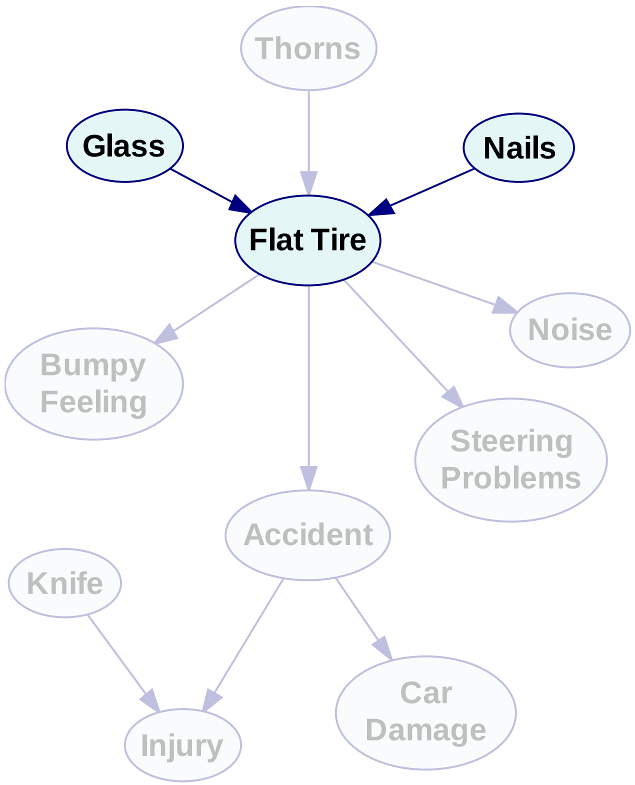
Car damage is correlated with noise because there is a directed path from flat tire to both (flat tire is a common cause of both).



Markov condition: Implications

Variables A and B are probabilistically dependent if there exists a D such that D is observed (conditioned upon) and there exists a C such that A is dependent on C and there exists a directed active path from C to D and there exists an E such that B is dependent on E and there exists a directed active path from E to D:

Nails on the road are correlated with glass on the road given flat tire because there is a directed path from glass on the road to flat tire and from nails on the road to flat tire and flat tire is observed (conditioned upon).



Markov condition: Summary of implications

Variables A and B are probabilistically dependent if:

- there exists a directed active path from A to B or there exists a directed active path from B to A
- there exists a C such that there exists a directed active path from C to A and there exists a directed active path from C to B
- there exists a D such that D is observed (conditioned upon) and there exists a C such that A is dependent on C and there exists a directed active path from C to D and there exists an E such that B is dependent on E and there exists a directed active path from E to D

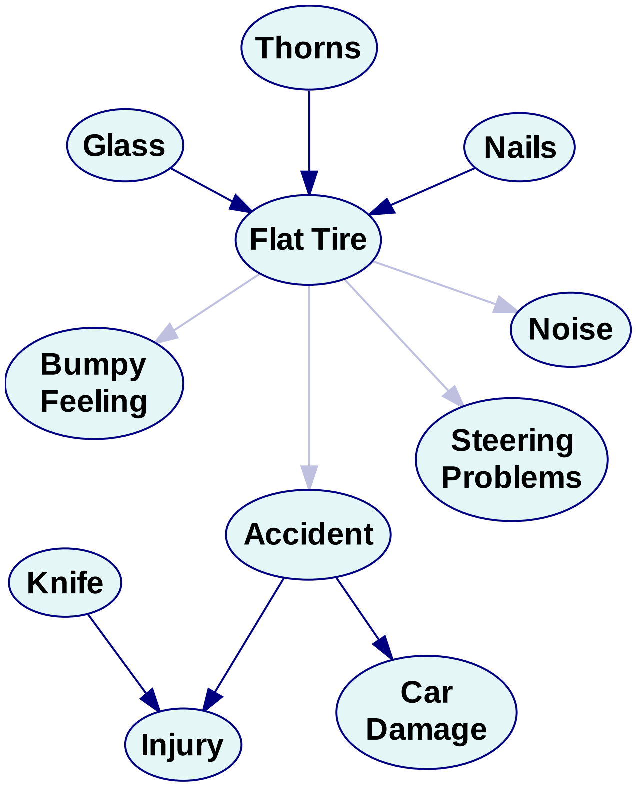
Markov condition: Conditional independence

Once we know all direct causes of an event E, the causes and effects of those causes do not tell anything new about E and its successors.

(also known as "screening off")

E.g.,

- Glass and thorns on the road are independent of noise, bumpy feeling, and steering problems conditioned on flat tire.
- Noise, bumpy feeling, and steering problems become independent conditioned on flat tire.

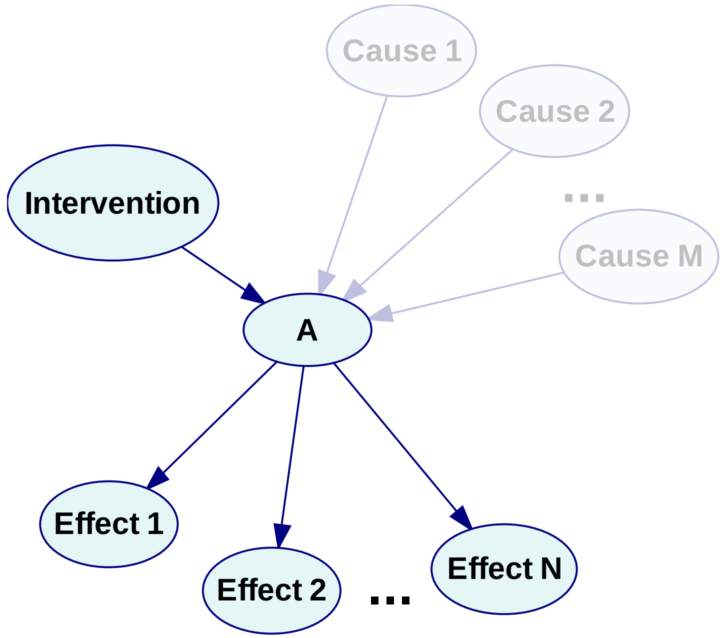


Intervention

Manipulation theorem [Spirtes, Glymour & Scheines 1993]:

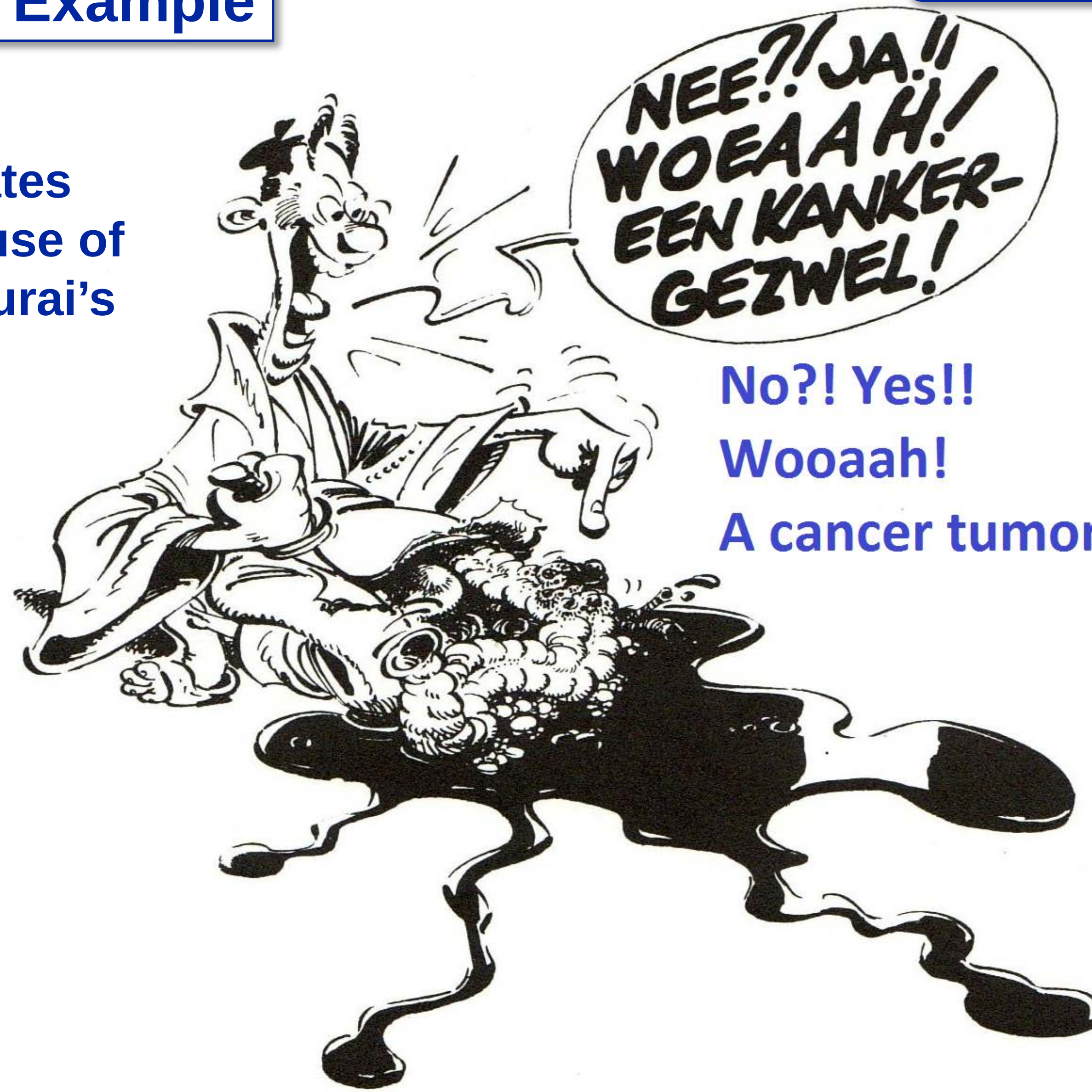
Given an external intervention on a variable A in a causal graph, we can derive the posterior probability distribution over the entire graph by simply modifying the conditional probability distribution of A.

If this intervention is strong enough to set A to a specific value, we can view this intervention as the only cause of A and reflect this by removing all edges that are coming into A. Nothing else in the graph needs to be modified.



Intervention: Example

Suicide eliminates cancer as a cause of this brave samurai's death.



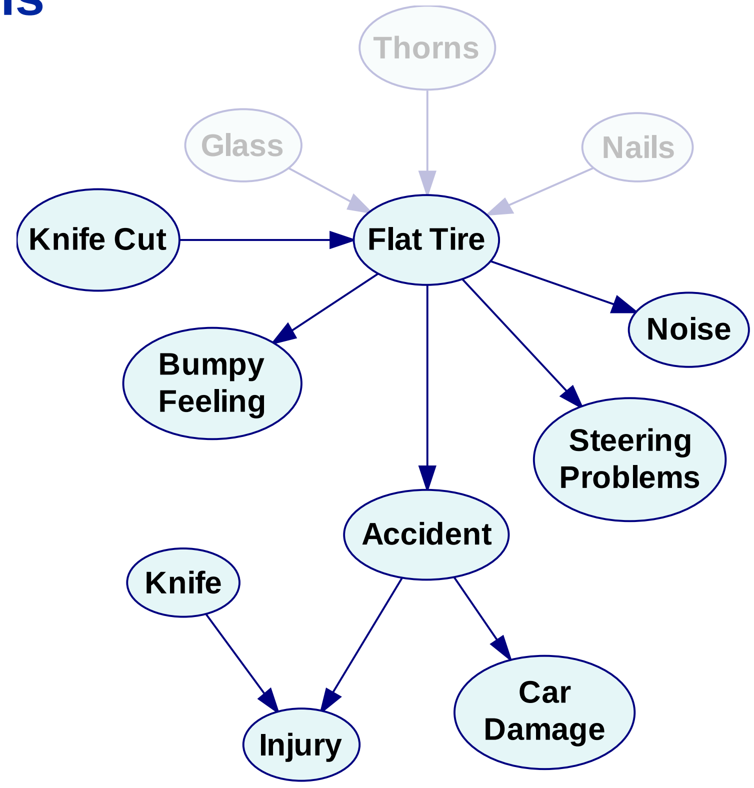
No?! Yes!!

Wooaah!

A cancer tumor!

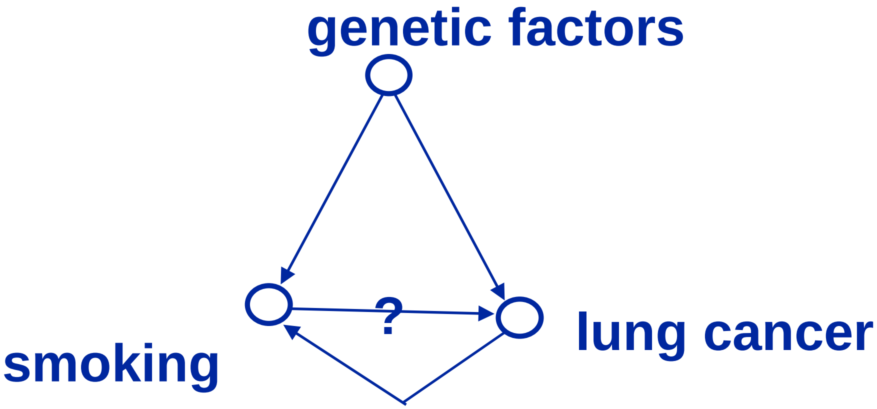
Intervention: Example

Making the tire flat with a knife makes glass, thorns, nails, and what-have-you irrelevant to flat tire. The knife is the only cause of flat tire.



Selection bias

Observing correlation is in general not enough to establish causality.



- **If we do not randomize, we run the danger that there are common causes between smoking and lung cancer (for example genetic factors).**
- **These common causes will make smoking and lung cancer dependent.**
- **It may, in fact, also be the case that lung cancer causes smoking.**
- **This will also make them dependent without smoking causing lung cancer.**

Experimentation

Empirical research is usually concerned with testing causal hypotheses.

Smoking and lung cancer are correlated.

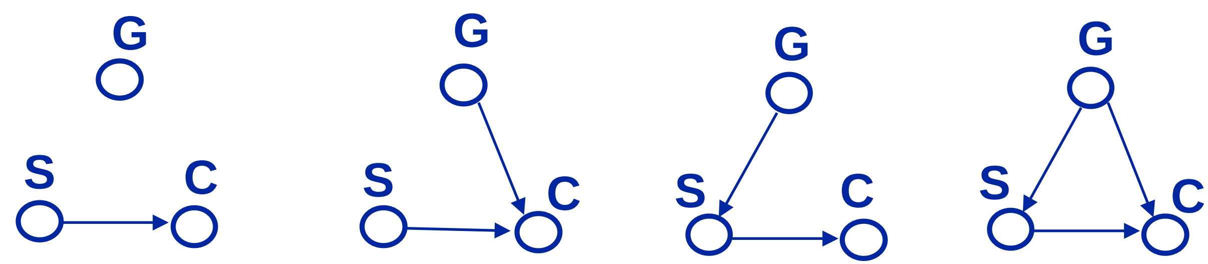
Can we reduce the incidence of lung cancer by reducing smoking?
In other words: Is smoking a cause of lung cancer?

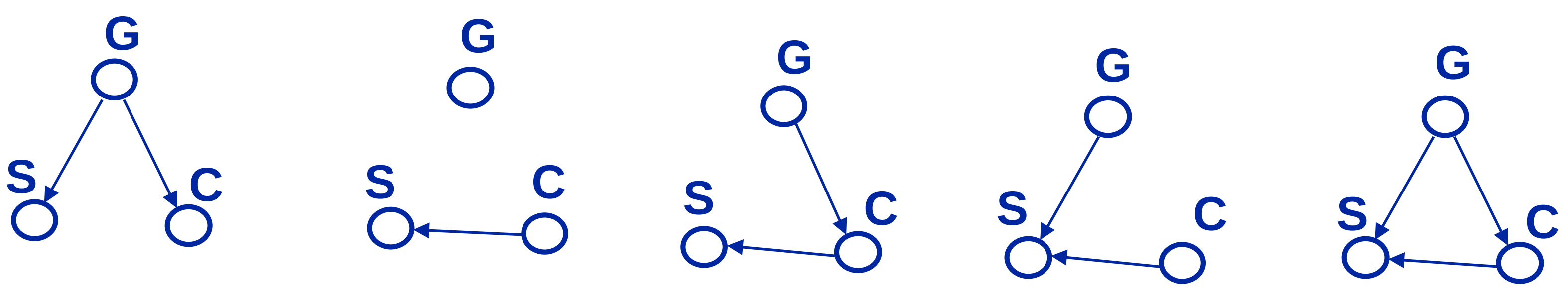
Each of the following causal structures is compatible with the observed correlation:

G = genetic factors

S = smoking

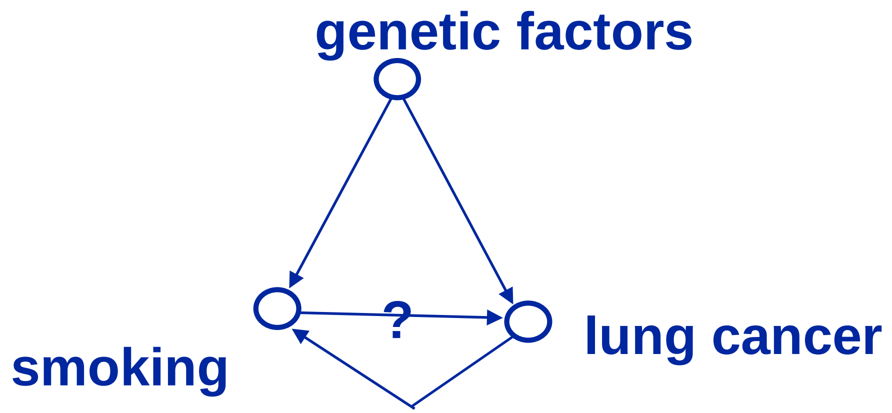
C = lung cancer





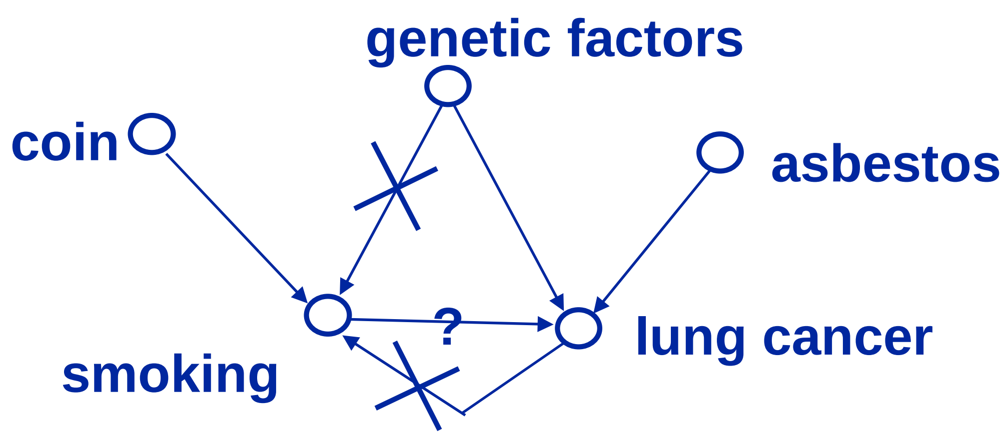
Selection bias

Observing correlation is in general not enough to establish causality.



- If we do not randomize, we run the danger that there are common causes between smoking and lung cancer (for example genetic factors).
- These common causes will make smoking and lung cancer dependent.
- It may, in fact, also be the case that lung cancer causes smoking.
- This will also make them dependent without smoking causing lung cancer.

Experimentation



- In a randomized experiment, coin becomes the only cause of smoking.
- Smoking and lung cancer will be dependent only if there is a causal influence from smoking to lung cancer.
- If Pr(C|S) $\neq$ Pr(C|~S) then smoking is a cause of lung cancer.
- Asbestos will simply cause variability in lung cancer (add noise to the observations).

But, can we really experiment in this domain?

Science by observation

"... Correlation between smoking and lung cancer means as much as correlation between apple imports and raise of divorce ..."

Sir Ronald A. Fisher, a prominent statistician, father of experimental design

"... George Bush taking credit for the end of the cold war is like a rooster taking credit for the daybreak ..."

Vice-president Al Gore towards vice –president Dan Quayle during their first (vice) presidential debate, Fall 1992

Science by observation

- **Experimentation is not always possible.**
- **We can do quite a lot by just observing.**
- **Assumptions are crucial in both experimentation and observation, although they are usually stronger in the latter.**
- **New methods in causal discovery: squeezing data to the limits**

Approaches to learning Bayesian networks

Constraint search-based learning

Search the data for independence relations to give us a clue about the causal relations [Spirtes, Glymour, Scheines 1993].

Bayesian learning

Search over the space of models and score each model using the posterior probability of the model given the data [Cooper & Herskovitz 1992; many others].

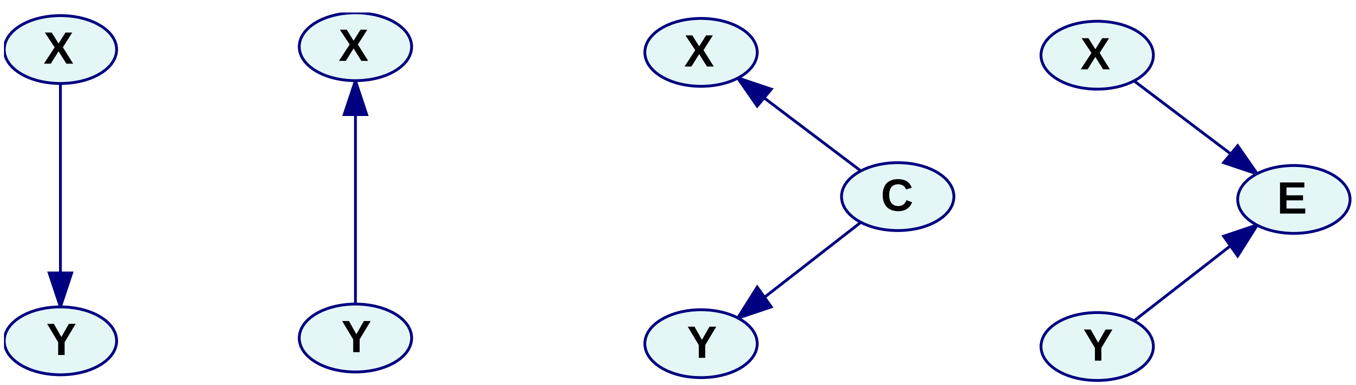
Constraint search-based learning

Constraint search-based learning

"Correlation does not imply causation"

True but only in limited settings (e.g., two variables) and typically abused by authors of college textbooks ☺.

If x and y are dependent, we can indeed simplify the causal picture to four simplified cases:

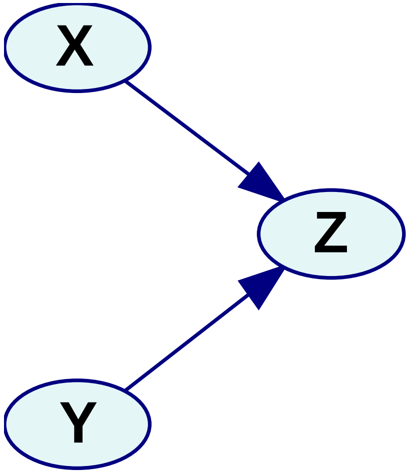


Constraint search-based learning

Not necessarily true in case of three variables:

x and z are dependent
y and z are dependent
x and y are independent
x and y are dependent given z

We can establish causality! $\Rightarrow$



Foundations of constrain-based search causal discovery

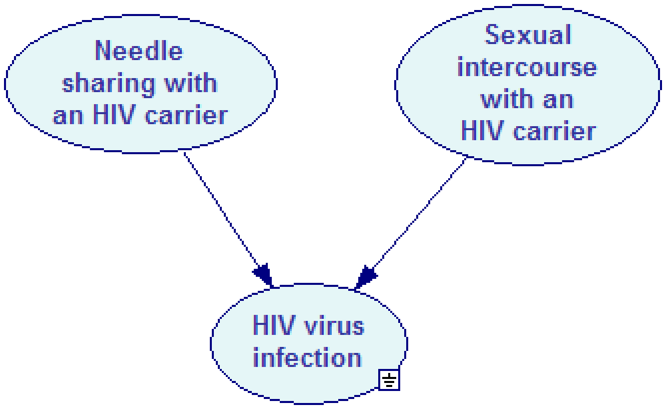
- **Markov Condition:**
  ***d*-separation $\Rightarrow$ independence in data.**
- **Faithfulness Condition:**
  ***d*-separation $\Leftarrow$ independence in data.**

The causal graph determines what is independent.

All independences in the data are structural, i.e., are consequences of Markov condition.

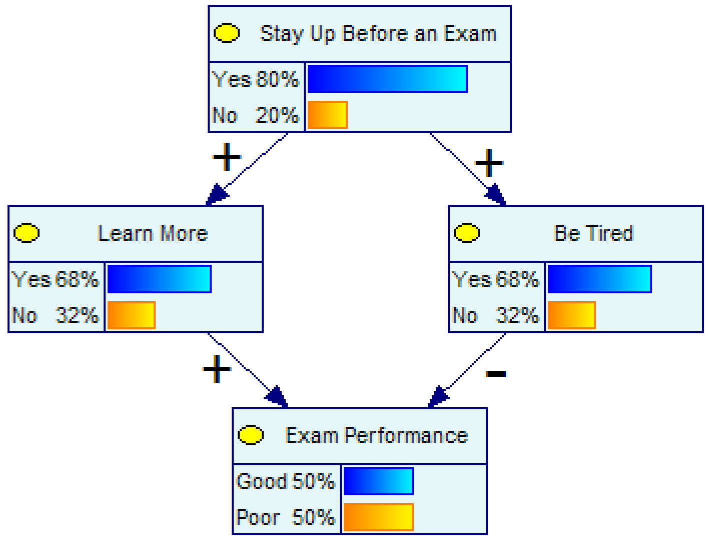
Violations of faithfulness condition

Faithfulness assumption is more controversial. While every scientist makes it in practice, it does not need to hold.



Given that HIV virus infection has not taken place, needle sharing is independent from intercourse.

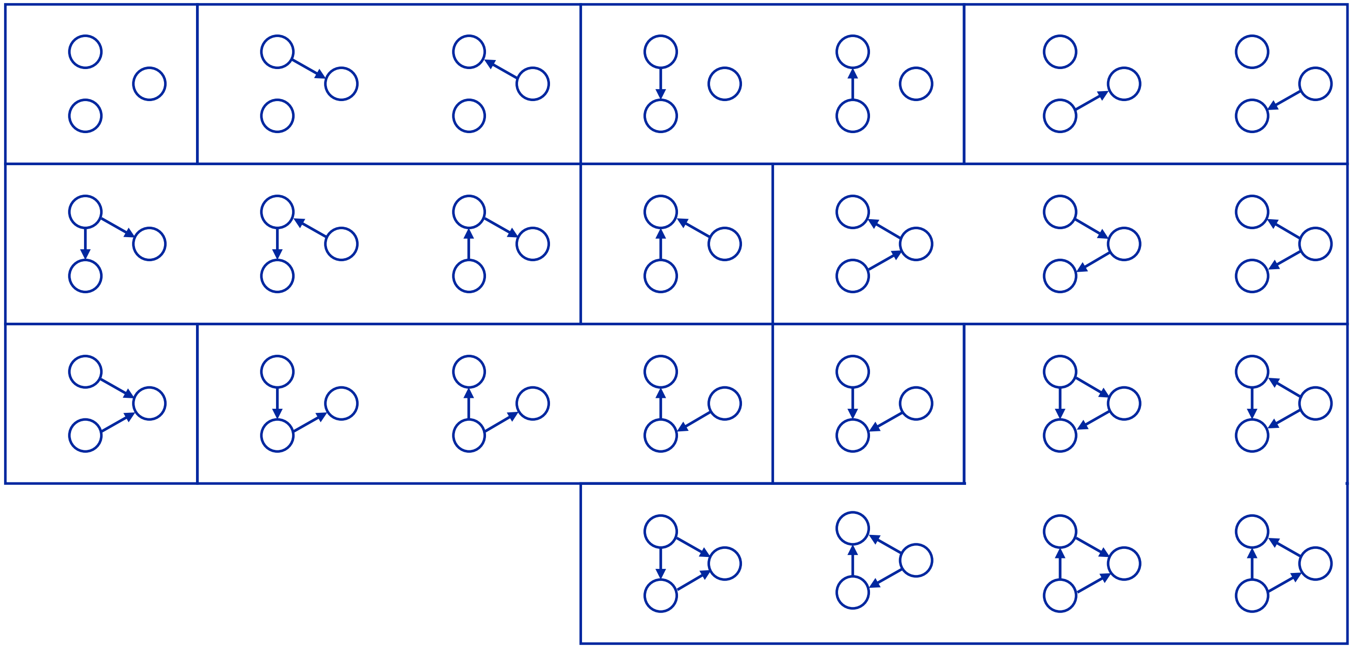
Violations of faithfulness condition



The effect of staying up late before the exam on the exam performance may happen to be zero:
being tired may cancel out the effect of more knowledge.
But is it likely?

Constraint search-based learning

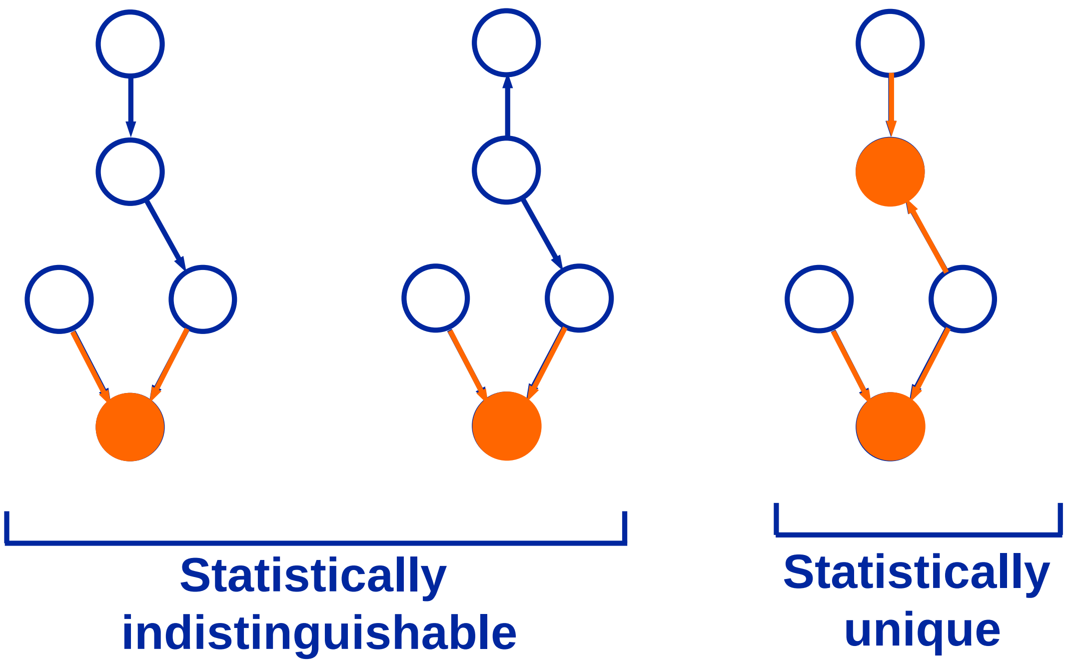
All possible networks …



… can be divided into equivalence classes

Equivalence criterion

Two graphs are statistically indistinguishable (belong to the same equivalence class) iff they have the same adjacencies and the same "v-structures".



Constraint search-based learning

Principles:

- Search for independencies among variables in the database.
- Use the *independencies* in the data to infer (lack of) *causal links* among the variables (given some basic assumptions).

Theorems useful in search

Theorem 1

There is no edge between X and Y if and only if X and Y are independent given *any* subset (including the null set) of the other variables.

Theorem 2

If X—Y — Z, X and Z are not adjacent, and X and Z are independent given some set W, then X→Y←Z if and only if W does *not* contain Y.

PC algorithm

Input:

a set of conditional independencies

Output:

a “pattern” which represents a Markov equivalence class of causally sufficient causal models.

PC algorithm (sketch)

Step 0:

Begin with a complete undirected graph.

Step 1 (Find adjacencies):

For each pair of variables <X,Y> if X and Y are independent given some subset of the other variables, remove the X–Y edge.

Step 2: (Find v-structures):

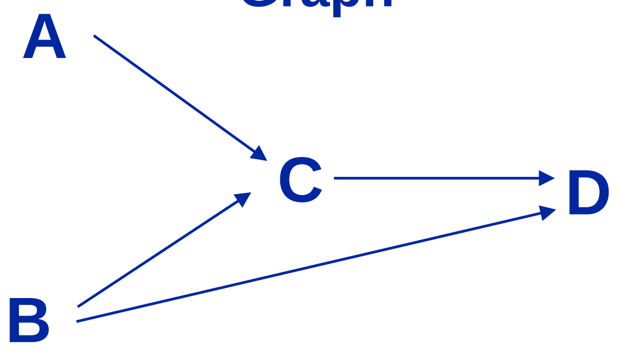
For each triple X–Y–Z, with no edge between X and Z, if X and Z are independent given some set not containing Y, then orient X–Y–Z as X→Y←Z.

Step 3 (Avoid new v-structures and cycles):

- if X→Y—Z, but there is no edge between X and Z, then orient Y–Z as Y→Z.
- if X—Z, and there is already a directed path from X to Z, then orient X — Z as X→Z.

PC algorithm: Example

Causal Graph

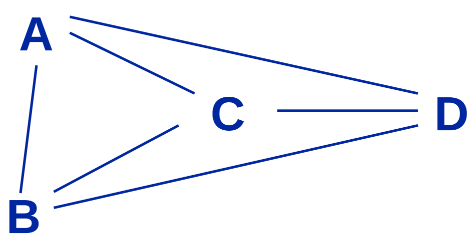


Independencies entailed by the Markov condition:

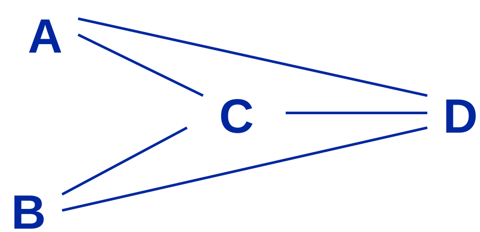
$$A \perp B$$

$$A \perp D \mid B,C$$

(0) Begin with

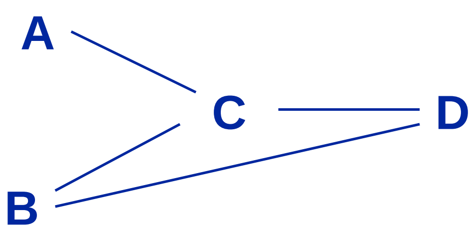


(1) From $A \perp B$, remove A—B

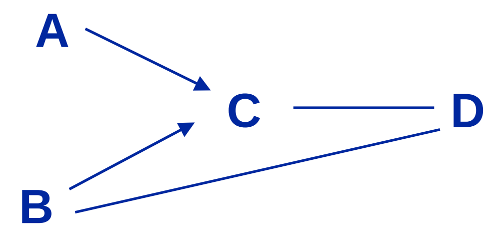


PC algorithm: Example

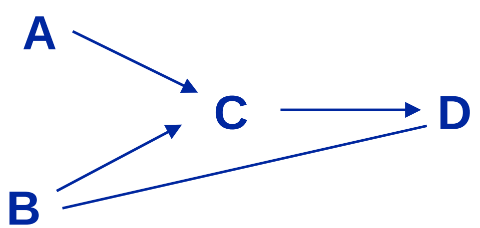
(1) From $A \perp D \mid B,C$, remove A—D



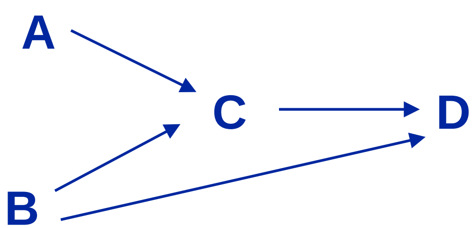
(2) From $A \perp B$, orient A–C–B as $A \rightarrow C \leftarrow B$



(3) Avoid a new v-structure ($A \rightarrow C \leftarrow D$), Orient C –D as $C \rightarrow D$.



(3) Avoid a cycle ($B \rightarrow C \rightarrow D \rightarrow B$), Orient B –D as $B \rightarrow D$.



Patterns: Output of the PC algorithm

PC algorithm outputs a ‘pattern’, a kind of graph containing directed (→), bi-directional (↔), and undirected (—) edges which represents a Markov equivalence class of Models

- A directed edge A→B in the ‘pattern’ indicates that there is an edge oriented A→B in every graph in the Markov equivalence class
- A bi-directional edge A↔B in the ‘pattern’ indicates that there is an edge between A and B in every graph in the Markov equivalence class, although its direction is impossible to establish based on the data
- An undirected edge A—B in the ‘pattern’, indicates that there is an edge between A and B in every graph in the Markov equivalence class, although its direction is impossible to establish based on the data; there is a possible common cause between these variables in every graph in the Markov equivalence class

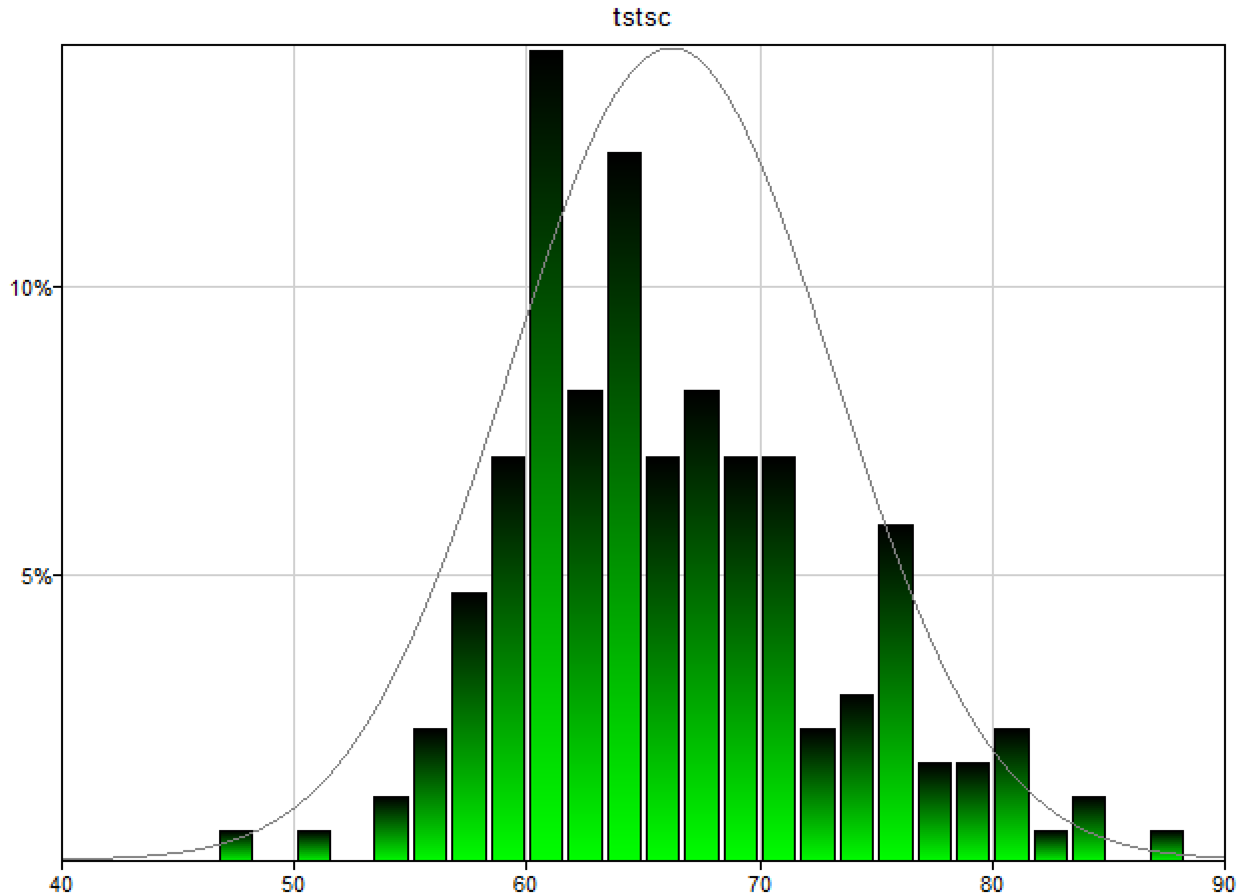
Dealing with errors in independence tests: Search with a varying value of statistical significance

- Independence tests performed in the first phase of the algorithm may result in Type I and Type II errors.
- It is a good practice to vary the level of statistical significance α, from very low to very high values.
- Graphs found with low values of α will be sparse. One can trust existence of arcs (low value of α, hard to reject null hypothesis H_0 that variables are independent; when H_0 still gets rejected, it means that the dependence was strong/robust).
- Graphs found with high values of α will be dense. One can trust absence of arcs (high value of α, easy to reject H_0 that variables are independent; when H_0 still does not get rejected, it means that the independence was strong/robust).

Continuous data

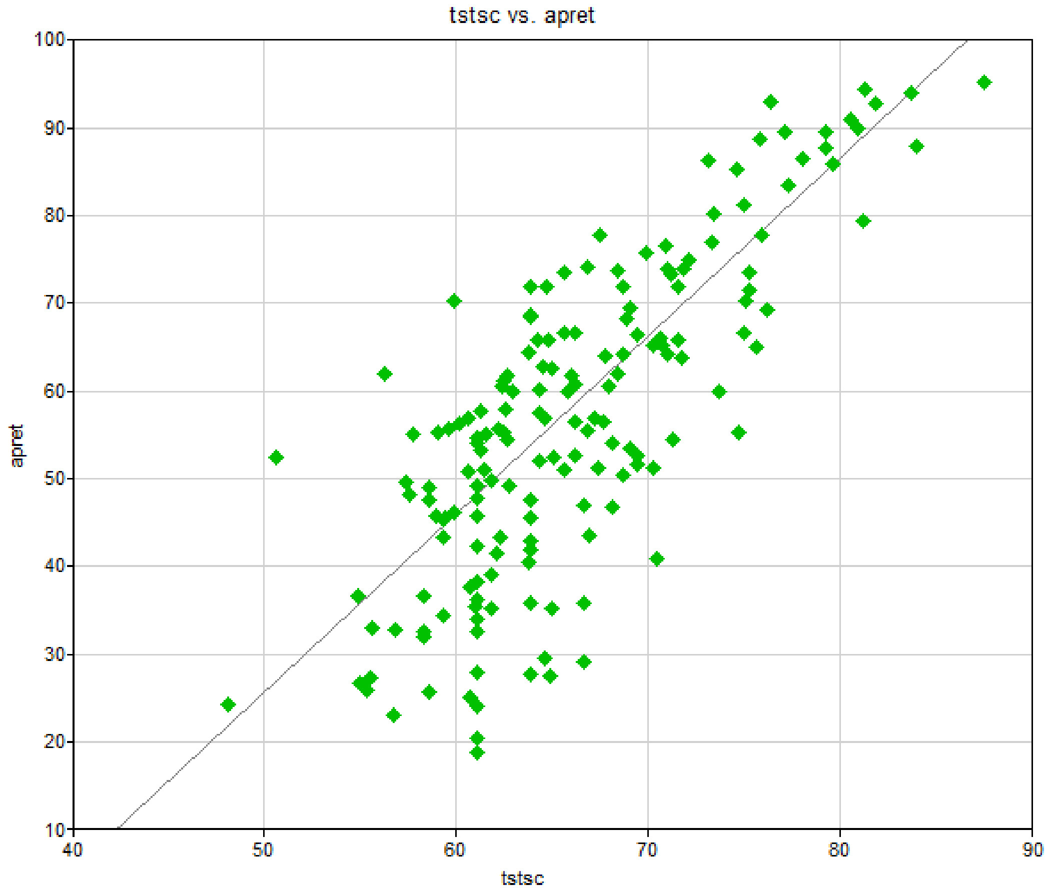
- Causal discovery is independent of the actual distribution of the data.
- The only thing that we need is a test of (conditional) independence.
- No problem with discrete data.
- In continuous case, we have a test of (conditional) independence (partial correlation test) when the data comes from multi-variate Normal distribution.
- Need to make the assumption that the data is multi-variate Normal.
- The discovery algorithm turns out to be very robust to this assumption [Voortman & Druzdzel, 2008].

Normality



Multi-variate normality is equivalent to two conditions:
(1) Normal marginals and (2) linear relationships

Linearity



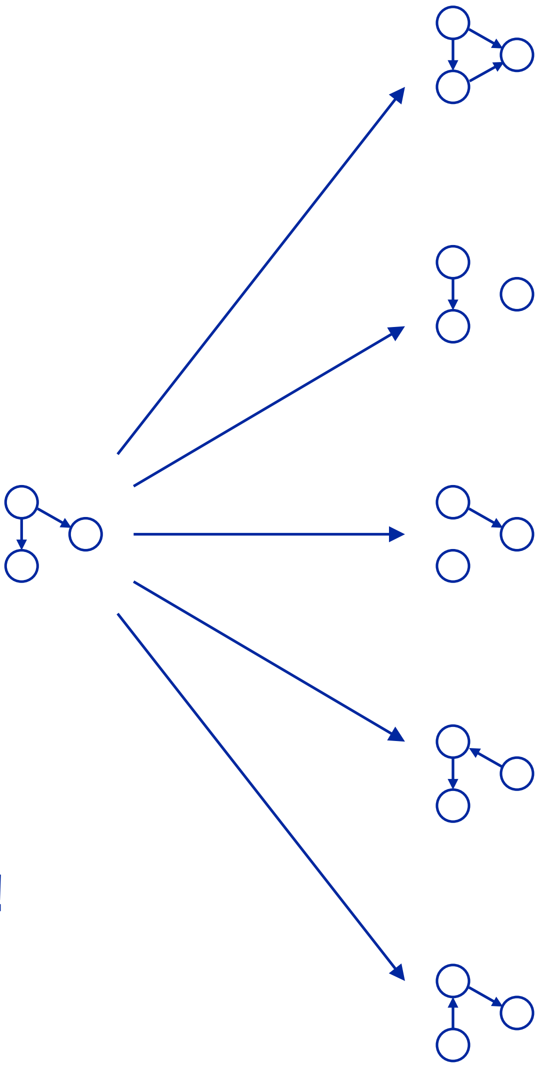
Multi-variate normality is equivalent to two conditions:

(1) Normal marginals and (2) linear relationships

Bayesian search learning

Elements of a search procedure

- ***A representation for the current state*** **(a network structure.)**
- ***A scoring function for each state*** **(the posterior probability).**
- ***A set of search operators.***
  - **AddArc(X,Y)**
  - **DelArc(X,Y)**
  - **RevArc(X,Y)**
- ***A search heuristic*** **(e.g., greedy search).**
- **The size of the search space for n variables is almost $3^\wedge C^n{}_2$ possible graphs!**



Posterior probability score

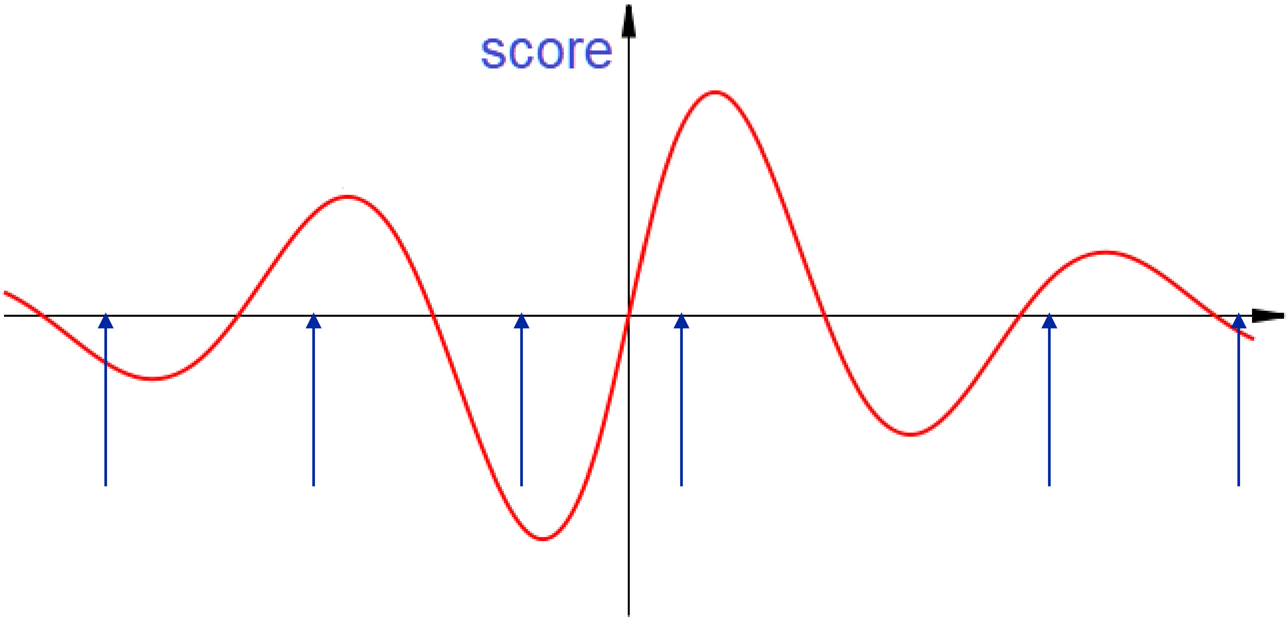
$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D)} \propto P(D \mid S)P(S)$$

"Marginal likelihood" P(D|S):

- **Given a database**
- **Assuming Dirichlet priors over parameters**

$$P(D \mid S) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

Dealing with local maxima: Restarts



Starting from a variety of different points (in this case, a variety of different graphs) increases the probability of finding the graph with a maximum score.

Constraint-based learning: Open problems

Pros:

- Efficient, $O(n^2)$ for sparse graphs.
- Hidden variables can be discovered in a modest way.
- “Older” technology, many researchers do not seem to be aware of it.

Cons:

- Discrete independence tests are computationally intensive

  $\Rightarrow$ heuristic independence tests?
- Missing data is difficult to deal with

  $\Rightarrow$ Bayesian independence test?

Bayesian learning: Open problems

Pros:

- Missing data and hidden variables are easy to deal with (in principle).
- More flexible means of specifying prior knowledge.
- Many open research questions!

Cons:

- Essentially intractable.
- Search heuristics (most efficient) typically lead to local maxima.
- Monte-Carlo techniques (more accurate) are very slow for most interesting problems.

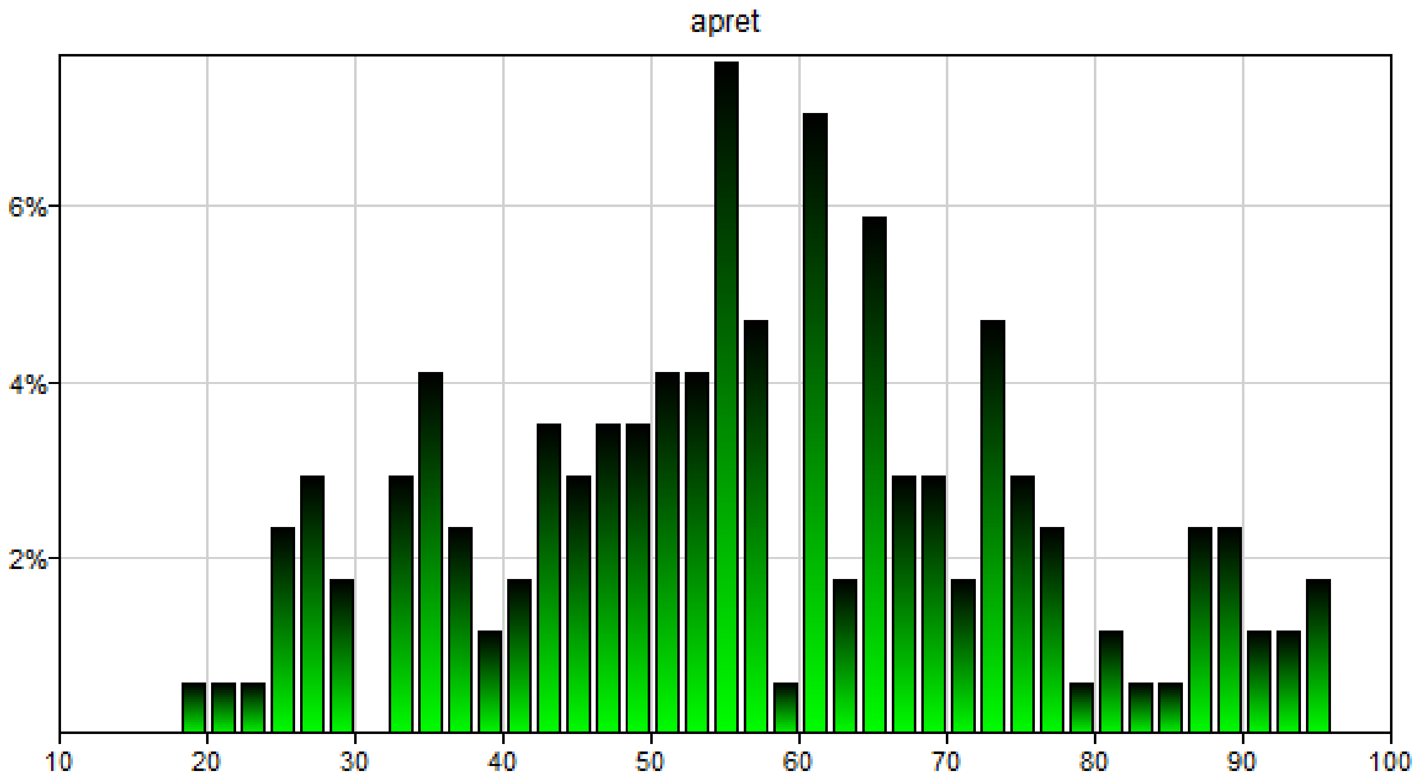
Example application

- Student retention in US colleges.
- Large problem for US colleges.
- Correctly predicted that the main causal factor in low student retention is the quality of incoming students.

  [Druzdzel & Glymour, 1994]

Example: What causes low student retention?

- Some US colleges lose over 80% of their incoming (undergraduate) students within the first year.
- Below a histogram of the 1994 retention rates of 170 US national colleges.



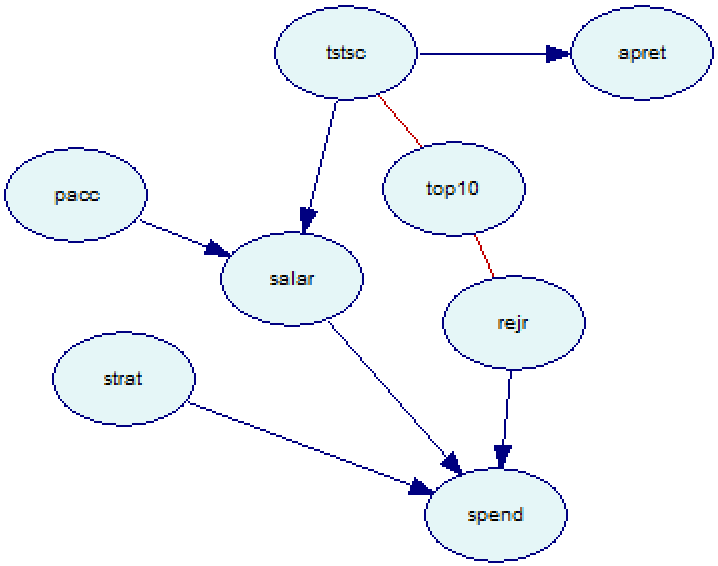
Example: What causes low student retention?

Everything seems to be correlated with everything.
What would you suggest causes low student retention?

	spend	apret	top10	rejr	tstsc	pacc	strat	salar
spend	-							
apret	0.601231	-						
top10	0.675656	0.642464	-					
rejr	0.633544	0.514958	0.643163	-				
tstsc	0.71491	0.782183	0.798807	0.628601	-			
pacc	-0.23673	-0.302834	-0.207505	-0.0715207	-0.164223	-		
strat	-0.561755	-0.458311	-0.247857	-0.283617	-0.465226	0.131858	-	
salar	0.711838	0.635852	0.637648	0.606777	0.715472	-0.37524	-0.347673	-

Example: What causes low student retention?

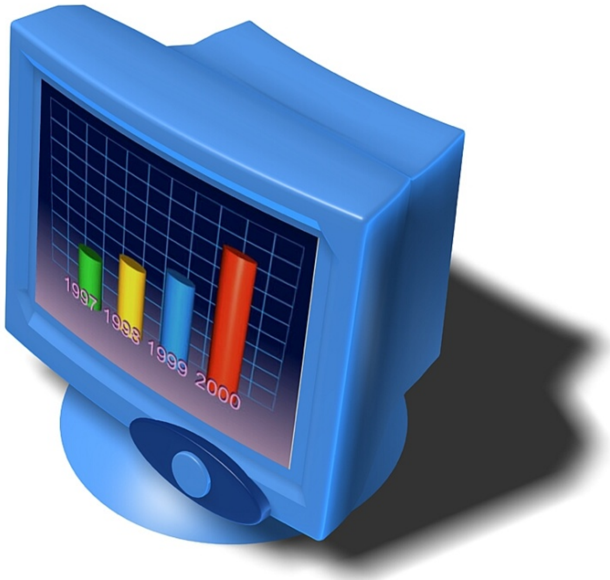
- It turns out that every model that we obtain by means of a learning procedure has a direct link between test scores and high school standing (measures of the quality of incoming students) and retention.
- This finding has been confirmed by a real-world experiment.



Some challenges

Scaling up -- especially Monte Carlo techniques.
Practically dealing with hidden variables -- unsupervised classification.
Applying these techniques to real data and real problems.
Hybrid techniques: Constraint-based + Bayesian (e.g., Dash & Druzdzel, 1999).
Learning causal graphs in time-dependent domains (Dash & Druzdzel, 2002).
Learning causal graphs and causal manipulation (Dash & Druzdzel, 2002).
Learning dynamic causal graphs from time series data (Voortman, Dash & Druzdzel 2010)

The rest



Concluding remarks

- **Observation is a valid scientific method**
- **Observation allows often to restrict the class of possible causal structures that could have generated the data.**
- **Learning Bayesian networks/causal graphs is very exciting: It is a different and powerful way of doing science.**
- **There is a rich assortment of unsolved problems in causal discovery / learning Bayesian networks, both practical and theoretical.**
- **Learning has been an active area of my research (GeNIe, https://www.bayesfusion.com/, is a product of this work).**