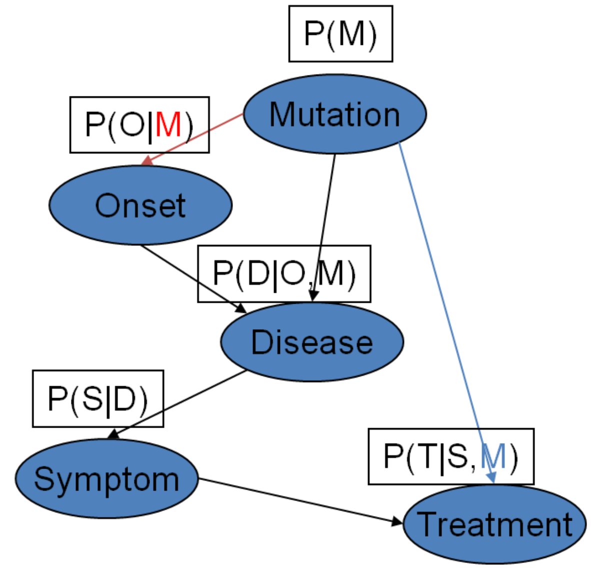
Causal Bayesian networks

Peter Antal
antal@mit.bme.hu

Outline

- Can we represent exactly (in)dependencies by a BN?
- Can we interpret/learn
  - edges as causal relations
    - with no hidden variables?
    - in the presence of hidden variables?
  - local models as autonomous mechanisms?
- Can we infer the effect of interventions?
- Can we quantify the consequences of interventions?

Bayesian networks: interpretations



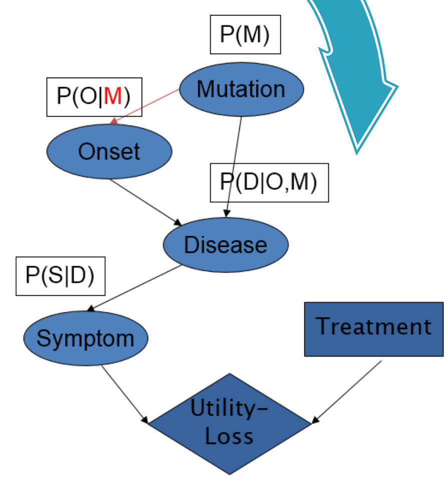
1. Causal model

3. Concise representation of joint distributions

$$P(M,O,D,S,T) = P(M)P(O \mid M)P(D \mid O,M)P(S \mid D)P(T \mid S,M)$$

$$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$$

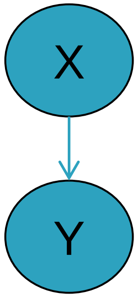
2. Graphical representation of (in)dependencies



4. Decision network

Motivation: from observational inference…

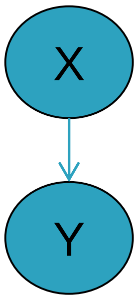
- In a Bayesian network, any query can be answered corresponding to passive observations: $p(Q=q|E=e)$.
  - What is the (conditional) probability of $Q=q$ given that $E=e$.
  - Note that Q can preceed temporally E.



- Specification: $p(X)$, $p(Y|X)$
- Joint distribution: $p(X,Y)$
- Inferences: $p(X)$, $p(Y)$, $p(Y|X)$, $p(X|Y)$

Motivation: to interventional inference…

- Perfect intervention: do(X=x) as set X to x.
- What is the relation of p(Q=q|E=e) and p(Q=q|do(E=e))?



- Specification: p(X), p(Y|X)
- Joint distribution: p(X,Y)
- Inferences:
  - p(Y|X=x)=p(Y|do(X=x))
  - p(X|Y=y)≠p(X|do(Y=y))

- What is a formal knowledge representation of a causal model?
- What is the formal inference method?

Motivation: and to counterfactual inference

- Imagery observations and interventions:
  - We observed X=x, but imagine that x' would have been observed: denoted as X'=x'.
  - We set X=x, but imagine that x' would have been set: denoted as do(X'=x').
- What is the relation of
  - Observational p(Q=q|E=e, X=x')
  - Interventional p(Q=q|E=e, do(X=x'))
  - Counterfactual p(Q'=q'|Q=q, E=e, do(X=x), do(X'=x'))

- O: What is the probability that the patient recovers if he takes the drug *x'*.
- I:What is the probability that the patient recovers if we prescribe* the drug *x'*.
- C: Given that the patient did not recovered for the drug x, what would have been the probability that patient recovers if we had prescribed* the drug *x'*, instead of *x*.

- *: Assume that the patient is fully compliant.
- **" expected to neither he will.

Challenges in a complex domain

The domain is defined by the joint distribution

$$P(X_1,\ldots, X_n|\text{Structure},\text{parameters})$$

1. Representation of parameteres

   „small number of parameters”

2. Representation of independencies

   „what is relevant for diagnosis”

3. Representation of causal relations

   „what is the effect of a treatment”

4. Representation of possible worlds

?

quantitave

qualitative

passive
(observational)

Active
(interventional)

Imagery
(counterfactual)

Decision theory
probability theory+utility theory

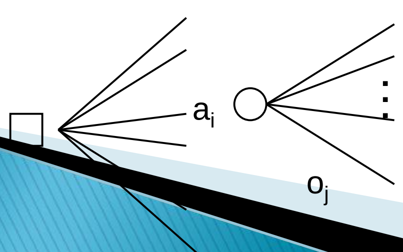
- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$
    - QALY, micromort

$$EU(a_i) = \sum_j U(o_j \mid a_i) p(o_j \mid a_i)$$

  - Maximum Expected Utility Principle (MEU)
    - Best action is the one with maximum expected utility

$$a^* = \arg\max_i EU(a_i)$$

Actions a_i (which experiment) | Outcomes (e.g. dataset) | Probabilities | Utilities, costs | Expected utilities



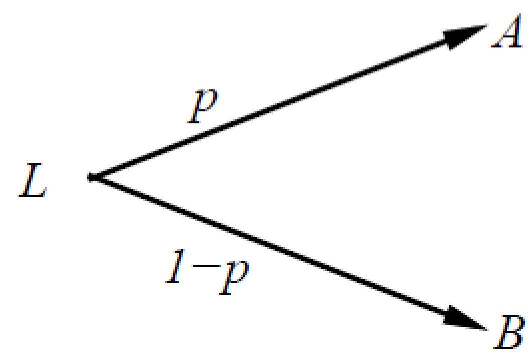
$P(o_j|a_i)$ $\vdots$

$U(o_j)$, $C(a_i)$ $\vdots$

$$EU(a_i) = \sum P(o_j|a_i)U(o_j)$$

Preferences

An agent chooses among prizes (A, B, etc.) and lotteries, i.e., situations with uncertain prizes



Lottery $L = [p, A;\ (1-p), B]$

Notation:

$A \succ B$	A preferred to B
$A \sim B$	indifference between A and B
$A \succsim B$	B not preferred to A

Rational preferences

Idea: preferences of a rational agent must obey constraints.

Rational preferences $\Rightarrow$

behavior describable as maximization of expected utility

Constraints:

Orderability

$$(A \succ B) \vee (B \succ A) \vee (A \sim B)$$

Transitivity

$$(A \succ B) \wedge (B \succ C) \Rightarrow (A \succ C)$$

Continuity

$$A \succ B \succ C \Rightarrow \exists p \; [p, A; \; 1-p, C] \sim B$$

Substitutability

$$A \sim B \Rightarrow [p, A; \; 1-p, C] \sim [p, B; 1-p, C]$$

Monotonicity

$$A \succ B \Rightarrow (p \geq q \Leftrightarrow [p, A; \; 1-p, B] \succsim [q, A; \; 1-q, B])$$

An irrational preference

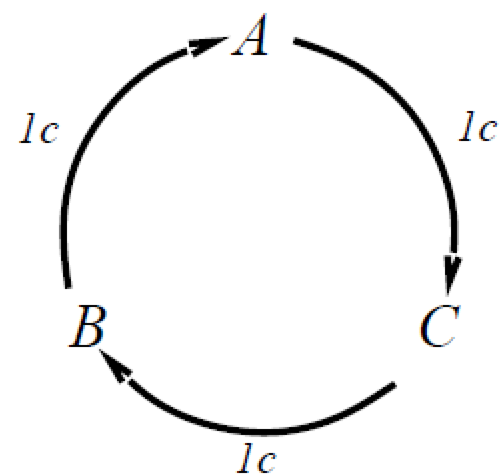
Violating the constraints leads to self-evident irrationality

For example: an agent with intransitive preferences can be induced to give away all its money

If $B \succ C$, then an agent who has C would pay (say) 1 cent to get B

If $A \succ B$, then an agent who has B would pay (say) 1 cent to get A

If $C \succ A$, then an agent who has A would pay (say) 1 cent to get C



Maximizing expected utility

Theorem (Ramsey, 1931; von Neumann and Morgenstern, 1944):
Given preferences satisfying the constraints
there exists a real-valued function U such that

$$U(A) \geq U(B) \quad \Leftrightarrow \quad A \succsim B$$
$$U([p_1, S_1; \ \ldots \ ; \ p_n, S_n]) = \Sigma_i \ p_i U(S_i)$$

MEU principle:

Choose the action that maximizes expected utility

Note: an agent can be entirely rational (consistent with MEU) without ever representing or manipulating utilities and probabilities

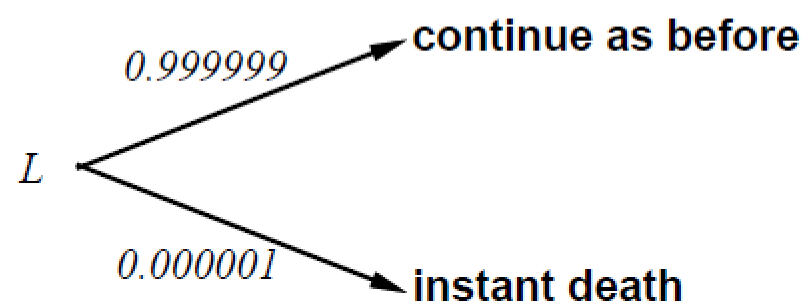
E.g., a lookup table for perfect tictactoe

Utilities

Utilities map states to real numbers. Which numbers?

Standard approach to assessment of human utilities:
compare a given state A to a standard lottery L_p that has
"best possible prize" $u_\top$ with probability p
"worst possible catastrophe" $u_\bot$ with probability $(1-p)$
adjust lottery probability p until $A \sim L_p$

pay \$30 ~



Utility scales

Normalized utilities: $u_\top = 1.0$, $u_\perp = 0.0$

Micromorts: one-millionth chance of death
  useful for Russian roulette, paying to reduce product risks, etc.

QALYs: quality-adjusted life years
  useful for medical decisions involving substantial risk

Note: behavior is **invariant** w.r.t. +ve linear transformation

$$U'(x) = k_1 U(x) + k_2 \quad \text{where } k_1 > 0$$

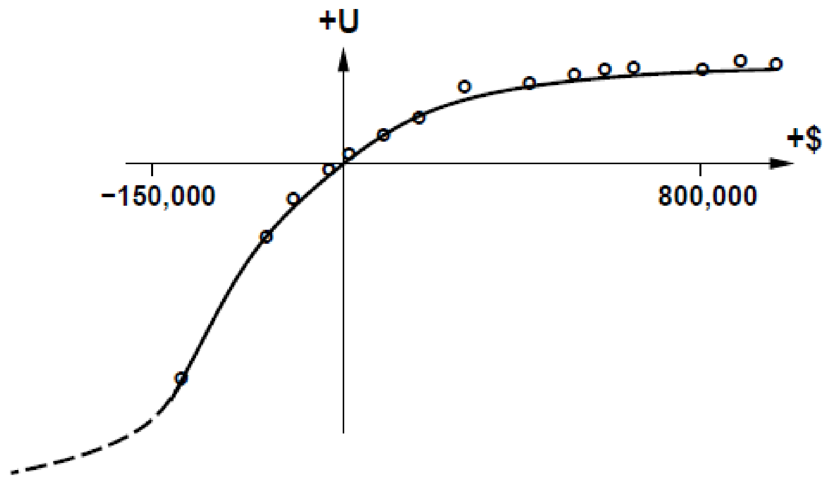
With deterministic prizes only (no lottery choices), only
ordinal utility can be determined, i.e., total order on prizes

Money

Money does **not** behave as a utility function. Given a lottery L with expected monetary value $EMV(L)$,
usually $U(L) < U(EMV(L))$, i.e., people are risk-averse.

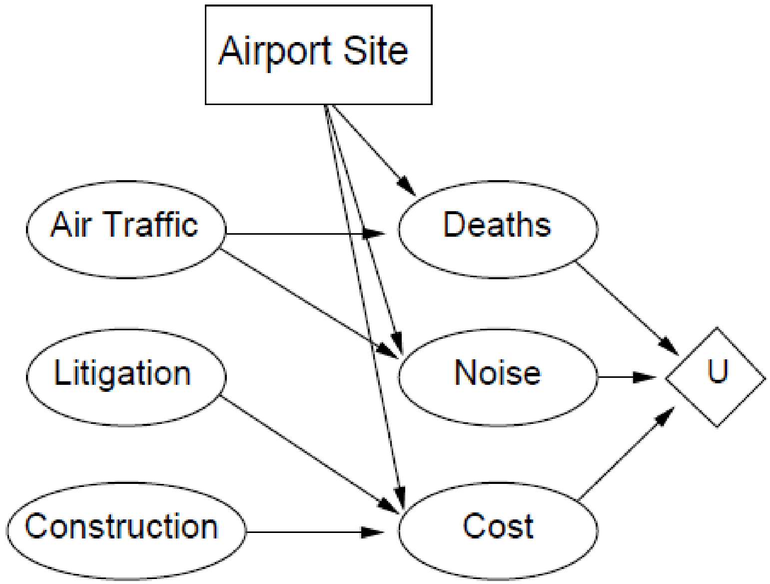
Utility curve: for what probability p am I indifferent between a prize x and a lottery $[p, \$M;\ (1-p), \$0]$ for large M?

Typical empirical data, extrapolated with risk-prone behavior:



Decision networks (DNs)

Add action nodes and utility nodes to belief networks
to enable rational decision making



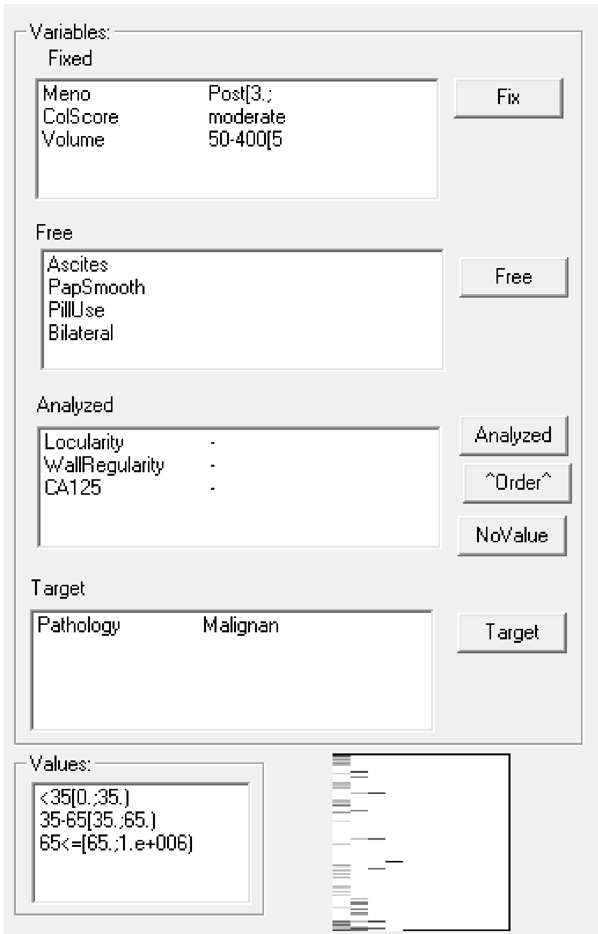
Algorithm:

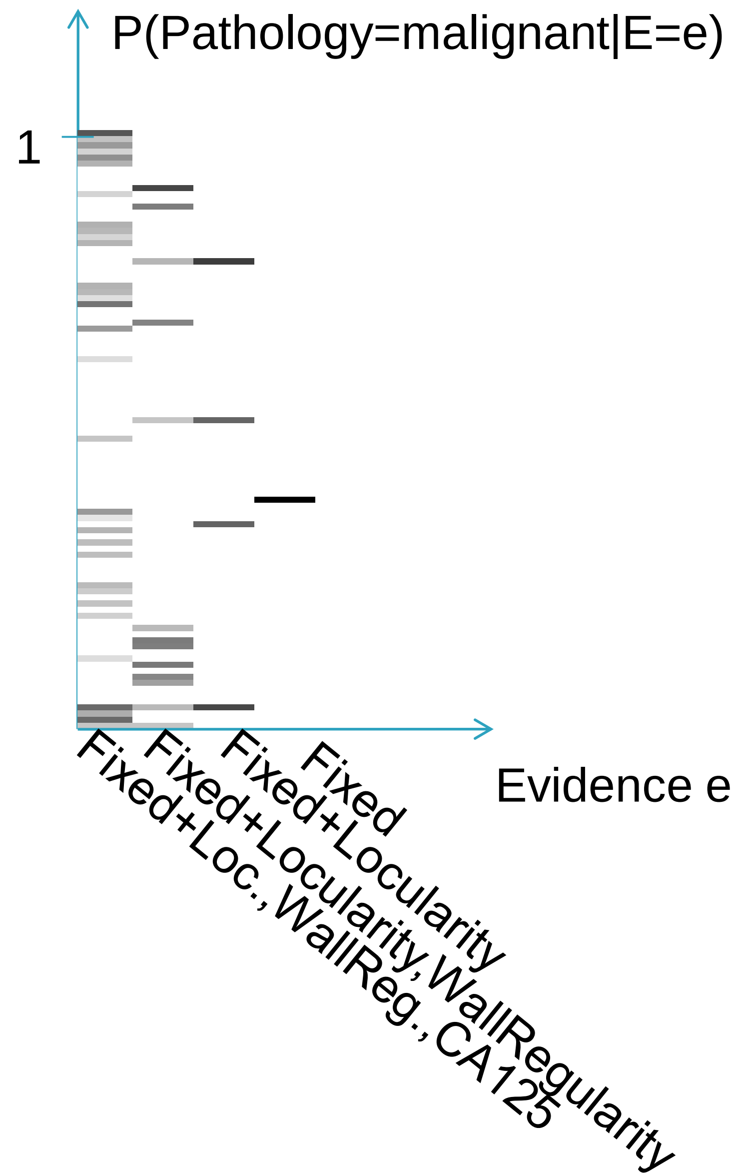
For each value of action node

compute expected value of utility node given action, evidence

Return MEU action

Sensitivity of the inference





Value of (perfect) Information

Current evidence E, current best action α
Possible action outcomes S_i, potential new evidence E_j

$$EU(\alpha|E) = \max_a \Sigma_i\, U(S_i)\; P(S_i|E, a)$$

Suppose we knew $E_j = e_{jk}$, then we would choose $\alpha_{e_{jk}}$ s.t.

$$EU(\alpha_{e_{jk}}|E, E_j = e_{jk}) = \max_a \Sigma_i\, U(S_i)\; P(S_i|E, a, E_j = e_{jk})$$

E_j is a random variable whose value is *currently* unknown
$\Rightarrow$ must compute expected gain over all possible values:

$$VPI_E(E_j) = \left(\Sigma_k\, P(E_j = e_{jk}|E) EU(\alpha_{e_{jk}}|E, E_j = e_{jk})\right) - EU(\alpha|E$$

Properties of VPI

Nonnegative—in **expectation**, not **post hoc**

$$\forall j, E \;\; VPI_E(E_j) \geq 0$$

Nonadditive—consider, e.g., obtaining E_j twice

$$VPI_E(E_j, E_k) \neq VPI_E(E_j) + VPI_E(E_k)$$

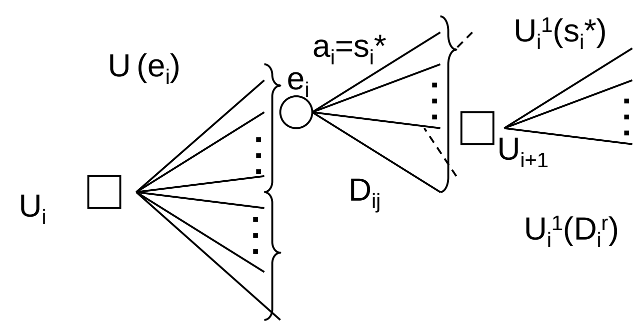
Order-independent

$$VPI_E(E_j, E_k) = VPI_E(E_j) + VPI_{E,E_j}(E_k) = VPI_E(E_k) + VPI_{E,E_k}(E_j)$$

Note: when more than one piece of evidence can be gathered, maximizing VPI for each to select one is not always optimal
$\Rightarrow$ evidence-gathering becomes a **sequential** decision problem

Extensions

- Bayesian learning
  - Predictive inference
  - Parametric inference
- Value of further information
- Sequential decisions
  - Optimal stopping (secretary problem)
  - Multiarmed bandit problem
  - Markov decision problem, reinforcement learning
  - ….learning a causal model and losses



Principles of causality

- strong association,
- X precedes temporally Y,
- plausible explanation without alternative explanations based on confounding,
- necessity (generally: if cause is removed, effect is decreased or actually: y would not have been occurred with that much probability if x had not been present),
- sufficiency (generally: if exposure to cause is increased, effect is increased or actually: y would have been occurred with larger probability if x had been present).

- Autonomous, transportable mechanism.

- The probabilistic definition of causation formalizes many, but for example not the counterfactual aspects.

Conditional independence

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z: $P(X;Y|z)=P(Y|z)\,P(X|z)$ for all z with $P(z)>0$.

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

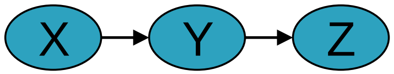
Other notations: $D_P(X;Y|Z)$ =def= $\neg\, I_P(X;Y|Z)$

Contextual independence: for not all z.

The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$



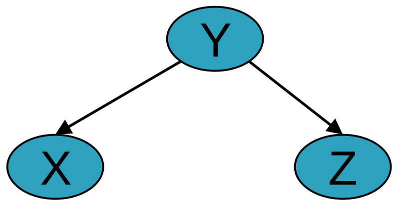
If P(X,Y,Z) is a Markov chain, then

$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$

Normally/almost always: $D(X;Z)$

Exceptionally: $I(X;Z)$

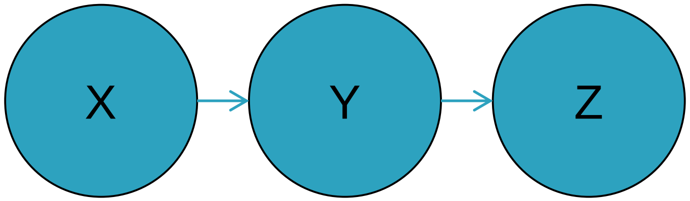
The independence map of a N-BN



If P(Y,X,Z) is a naive Bayesian network, then
M_P={D(X;Y), D(Y;Z), I(X;Z|Y)}
Normally/almost always: D(X;Z)
Exceptionally: I(X;Z)

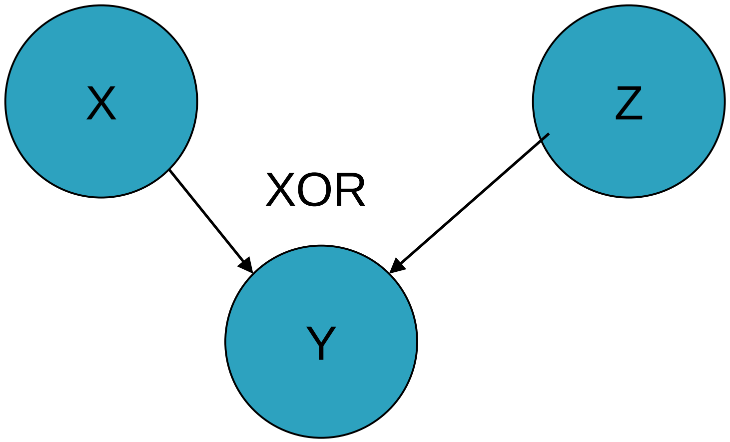
Parametrically encoded intransitivity of dependencies

- In the first order Markov chain below, despite the dependency of X–Y and Y–Z, X and Z can be independent (assuming non-binary Y).

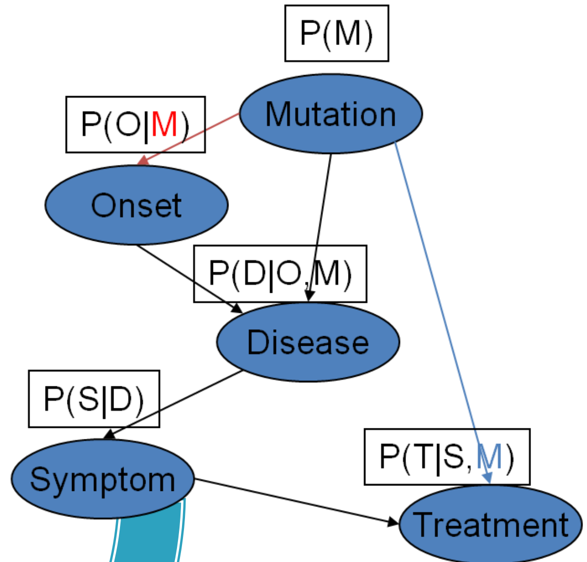


Parametrically encoded pairwise in dependencies

- Pairwise independence does not imply multivariate independence!



Bayesian networks: the three facets



1. Causal model

$$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$$

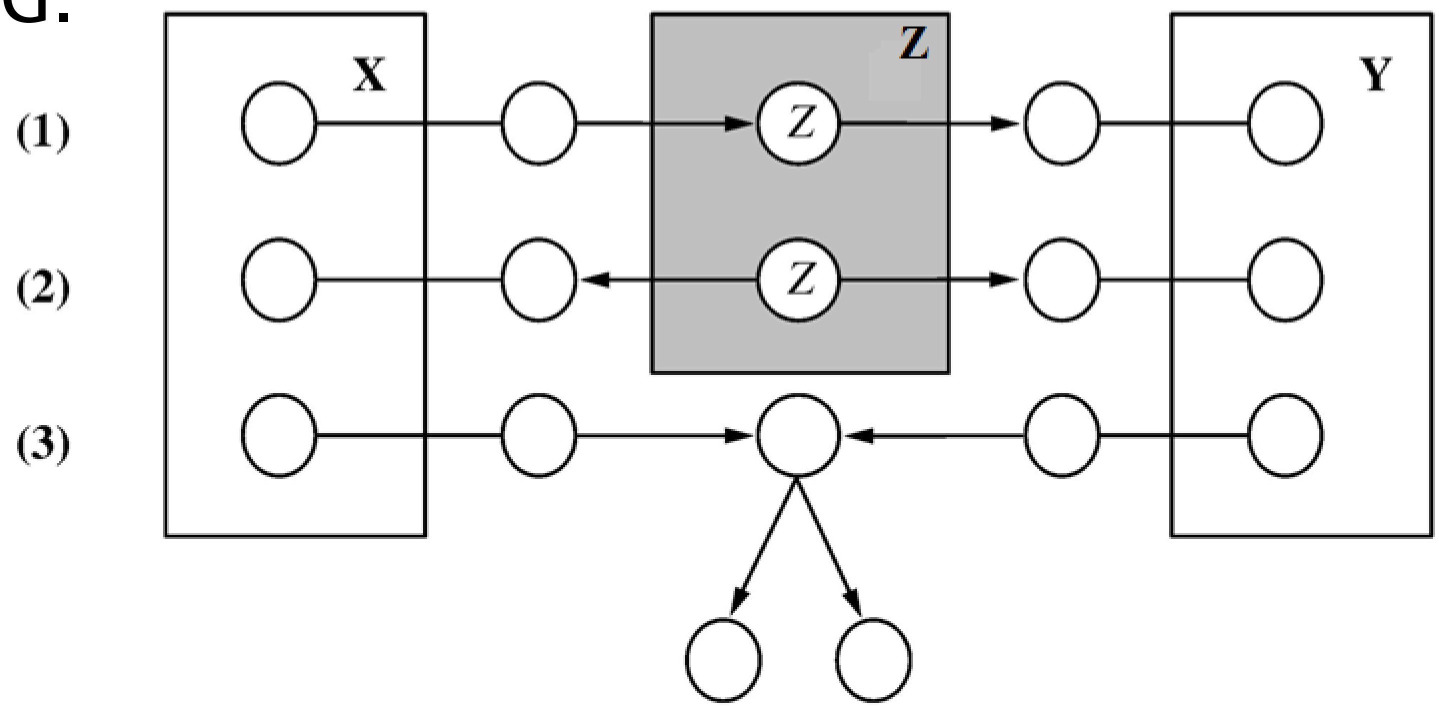
2. Graphical representation of (in)dependencies

3. Concise representation of joint distributions

$$P(M,O,D,S,T)=$$
$$P(M)P(O\mid M)P(D\mid O,M)P(S\mid D)P(T\mid S,M)$$

Inferring independencies from structure: d-separation

$I_G(X;Y|Z)$ denotes that X is d-separated (directed separated) from Y by Z in directed graph G.



d-separation and the global Markov condition

Definition 7 *A distribution $P(X_1, \ldots, X_n)$ obeys the global Markov condition w.r.t. DAG G, if*

$$\forall\, X, Y, Z \subseteq U \; (X \perp\!\!\!\perp Y | Z)_G \Rightarrow (X \perp\!\!\!\perp Y | Z)_P, \qquad (9)$$

where $(X \perp\!\!\!\perp Y | Z)_G$ denotes that X and Y are d-separated by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows

1. *either path p contains a node n in Z with non-converging arrows (i.e. $\rightarrow n \rightarrow$ or $\leftarrow n \rightarrow$),*
2. *or path p contains a node n not in Z with converging arrows (i.e. $\rightarrow n \leftarrow$) and none of its descendants of n is in Z.*

Representation of independencies

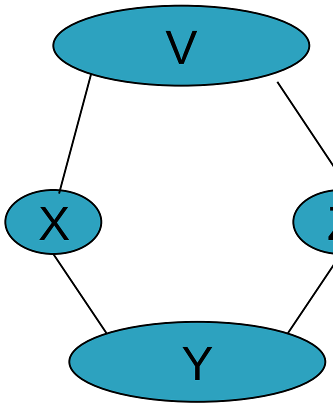
D-separation provides a sound and complete, computationally efficient algorithm to read off an (in)dependency model consisting the independencies that are valid in all distributions Markov relative to G, that is $\forall\ X, Y, Z \subseteq V$

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow ((X \perp\!\!\!\perp Y|Z)_P \text{ in all P Markov relative to G}). \tag{10}$$

For certain distributions exact representation is not possible by Bayesian networks, e.g.:

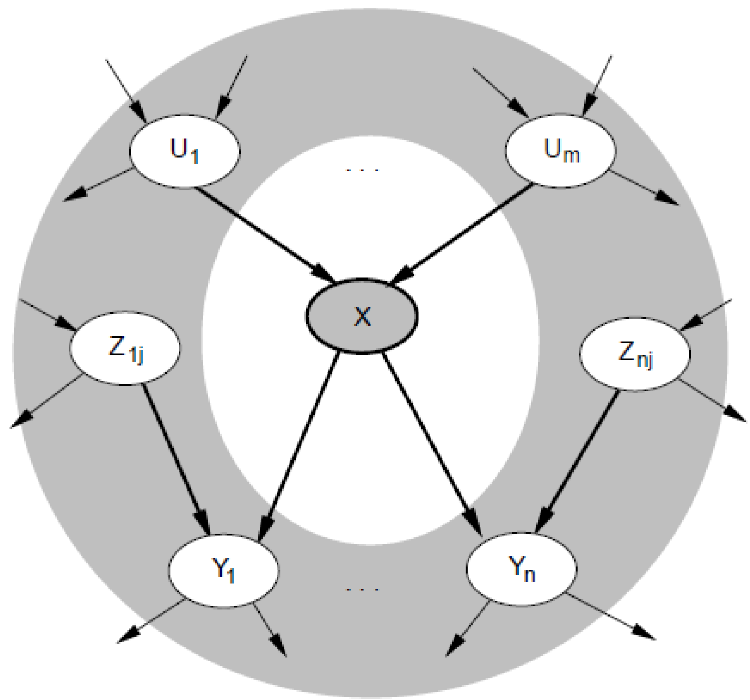
1. Intransitive Markov chain: X➔Y➔Z
2. Pure multivariate cause: {X,Z}➔Y
3. Diamond structure:

P(X,Y,Z,V) with M_P={D(X;Z), D(X;Y), D(V;X), D(V;Z), I(V;Y|{X,Z}), I(X;Z|{V,Y}).. }.



Markov blanket (and boundary)

Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



A Bayesian network definition

A directed acyclic graph (DAG) G is a Bayesian network of distribution $P(U)$ iff $P(U)$ obeys the global Markov condition with respect to G and G is minimal (i.e. no edges can be omitted without violating this property).

A practical definition

Definition 9 *A Bayesian network model* M *of domain with variables* U *consists of a structure* G *and parameters* $\boldsymbol{\theta}$. *The structure* G *is a* **DAG** *such that each node represents a variable and* *local probabilistic models* $p(X_i|pa(X_i))$ *are attached to each node w.r.t. the structure* G, *that is they describe the stochastic dependency of variable* X_i *on its parents* $pa(X_i)$. *As the conditionals are frequently from a certain parametric family, the conditional for* X_i *is parameterized by* $\boldsymbol{\theta_i}$, *and* $\boldsymbol{\theta}$ *denotes the overall parameterezation of the model.*

Markov conditions

Definition 4 *A distribution $P(X_1, \ldots, X_n)$ is* Markov relative *to DAG G or factorizes w.r.t G, if*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)), \tag{6}$$

where $Pa(X_i)$ denotes the parents of X_i in G.

Definition 5 *A distribution $P(X_1, \ldots, X_n)$ obeys the* ordered Markov *condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_{\pi(i)} \perp\!\!\!\perp \{X_{\pi(1)}, \ldots X_{\pi(i-1)}\} / Pa(X_{\pi(i)}) | Pa(X_{\pi(i)}))_P, \tag{7}$$

where $\pi()$ is some ancestral ordering w.r.t. G (i.e. compatible with arrows in G).

Definition 6 *A distribution $P(X_1, \ldots, X_n)$ obeys the* local (or parental) Markov *condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_i \perp\!\!\!\perp \mathrm{Nondescendants(X_i)} | \mathrm{Pa(X_i)})_{\mathrm{P}}, \tag{8}$$

where $\mathrm{Nondescendants(X_i)}$ denotes the nondescendants of X_i in G.

Bayesian network definitions

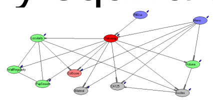
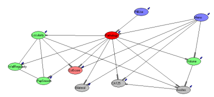
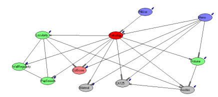
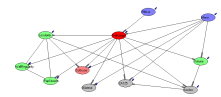
Theorem 1 *Let $P(U)$ a probability distribution and* **G** *a DAG, then the conditions above (repeated below) are equivalent:*

- **F** P *is Markov relative* G *or* P *factorizes w.r.t* G,
- **O** P *obeys the ordered Markov condition w.r.t.* G,
- **L** P *obeys the local Markov condition w.r.t.* G,
- **G** P *obeys the global Markov condition w.r.t.* G.

Definition 8 *A directed acyclic graph* **(DAG)** G *is a* Bayesian network *of distribution* $P(U)$ *iff the variables are represented with nodes in* G *and* (G, P) *satisfies any of the conditions* F, O, L, G *such that* G *is minimal (i.e. no edge(s) can be omitted without violating a condition* F, O, L, G*).*

Observational equivalence of causal models

Observationally equivalent causal models:

= = = = d-separation

Independence model:

$$P(X_1,\dots, X_n)$$

$$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\dots, I_{P,K}(X_K;Y_K|Z_K)\}$$

Different causal models can have the same independence map!

Typically causal models cannot be identified from passive observations, they are ***observationally equivalent***.

Association vs. Causation: Markov chain

Causal models:

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \qquad X_1 \leftarrow X_2 \leftarrow X_3 \leftarrow X_4$$

Markov chain

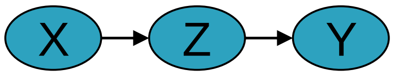
$P(X_1,\ldots)$

$M_P=\{I(X_{i+1};X_{i-1}|X_i)\}$

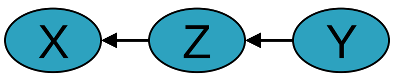
„first order Markov property”

The building block of causality: v–structure (arrow of time)

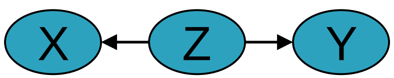
$p(X), p(Z|X), p(Y|Z)$



$p(X|Z), p(Z|Y), p(Y)$

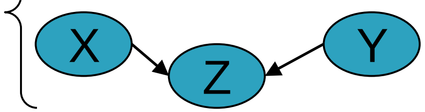


$p(X|Z), p(Z), p(Y|Z)$



“transitive” M ≠ „intransitive” M

$p(X), p(Z|X,Y), p(Y)$



„v-structure”

$M_P=\{D(X;Z),\ D(Z;Y),\ D(X,Y),\ I(X;Y|Z)\}$

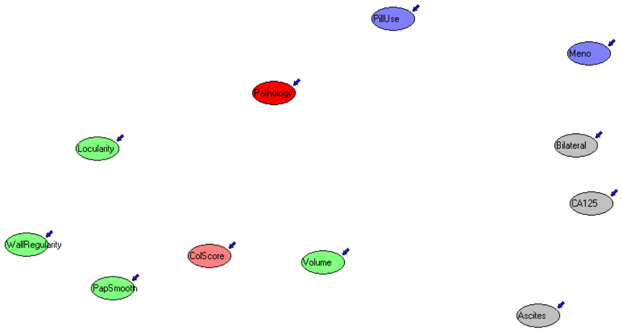
$M_P=\{D(X;Z),\ D(Y;Z),\ I(X;Y),\ D(X;Y|Z)\ \}$

Often (confounding): present knowledge renders (otherwise dependent) future states conditionally independent.

Ever(?): present knowledge renders (otherwise independent) future states conditionally dependent.

Observational equivalence: total independence

„Causal” model:



One-to-one relation

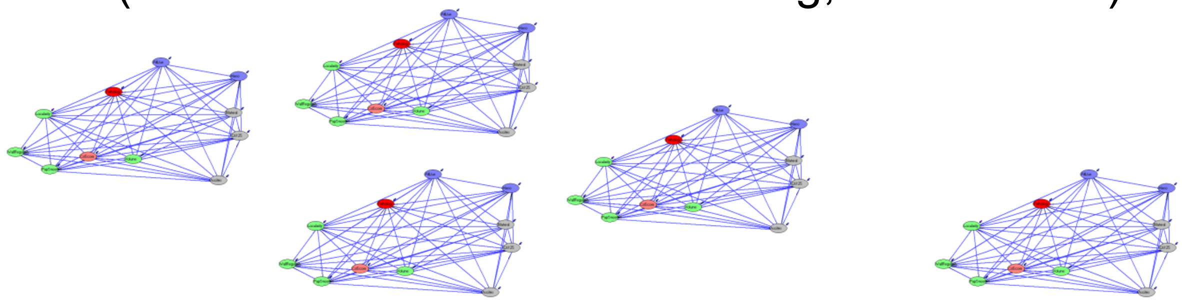
Dependency map:

$P(X_1,\ldots, X_n)$

$M_P=\{I_{P,1}(X_1;X_2),\ldots\}$

Observational equivalence: full dependence

„Causal” models (there is a DAG for each ordering, i.e. n! DAGs):



One-to-many relation

Dependency map:

$P(X_1,\ldots, X_n)$

$M_P=\{D_{P,1}(X_1;X_2),\ldots\}$

Observational equivalence of causal models

Definition 11 *Two DAGs G_1, G_2 are* observationally equivalent *, if they imply the same set of independence relations (i.e. $(X \perp\!\!\!\perp Y|Z)_{G_1}) \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$).*

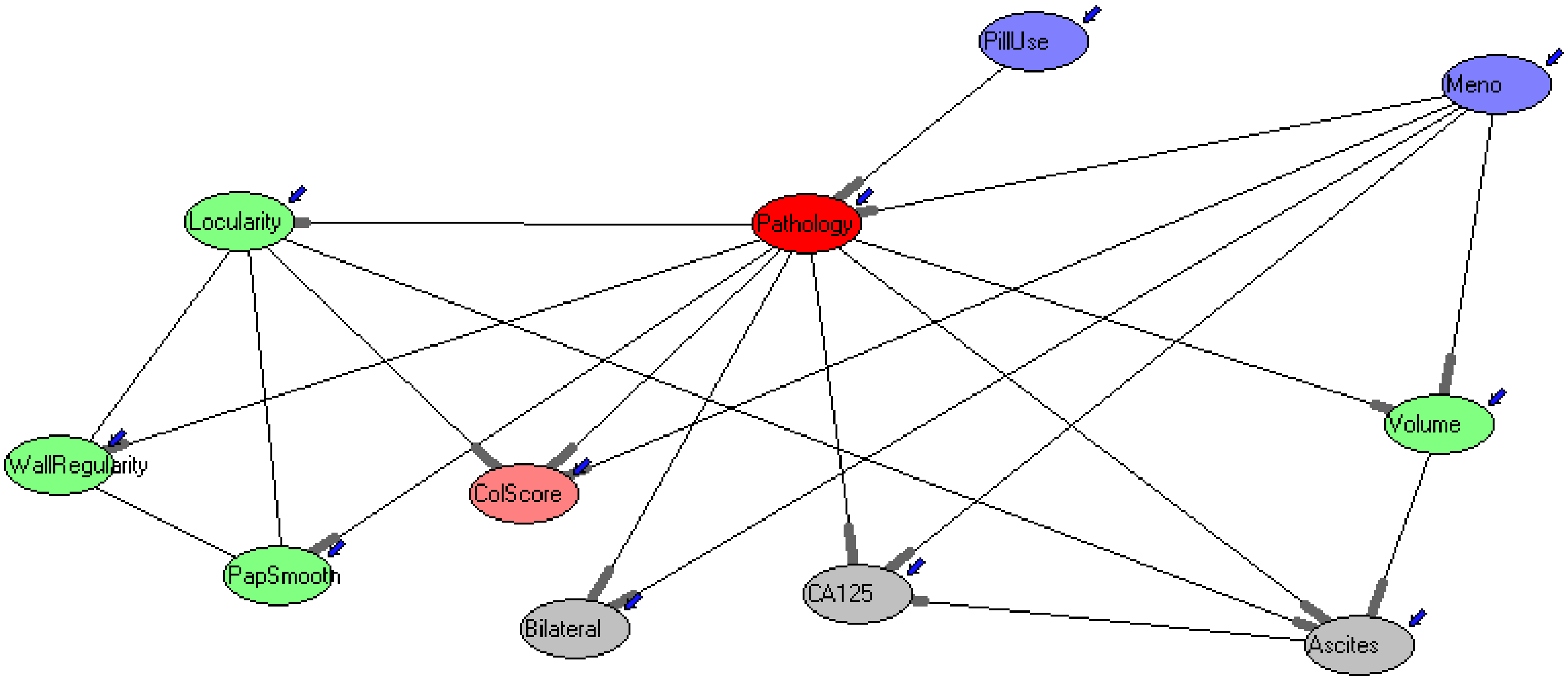
The implied equivalence classes may contain $n!$ number of DAGs (e.g. all the full networks representing no independencies) or just 1.

Theorem 2 *Two DAGs G_1, G_2 are* observationally equivalent *, iff they have the same skeleton (i.e. the same edges without directions) and the same set of* v-structures *(i.e. two converging arrows without an arrow between their tails).*

Definition 12 *The* essential graph *representing observationally equivalent DAGs is a partially oriented DAG (PDAG), that represents the identically oriented edges called* compelled edges *of the observationally equivalent DAGs (i.e. in the equivalence class), such a way that in the common skeleton only the compelled edges are directed (the others are undirected representing inconclusiveness).*

A limits of learnability: compelled edges

(“can we interpret edges as causal relations?”➔compelled edges)



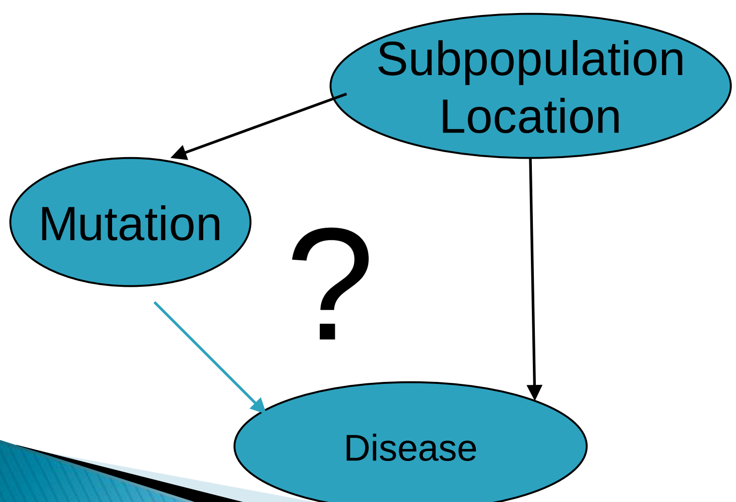
Interventional inference in causal Bayesian networks

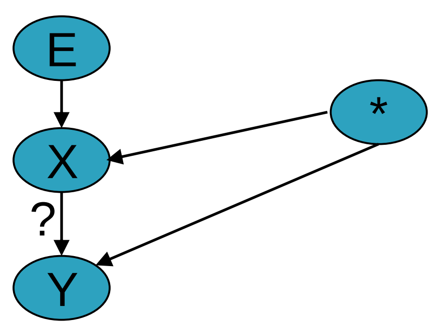
- (Passive, observational) inference
  - P(Query|Observations)
- **Interventionist inference**
  - **P(Query|Observations, Interventions)**
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

Interventions and graph surgery

If G is a causal model, then compute p(Y|do(X=x)) by

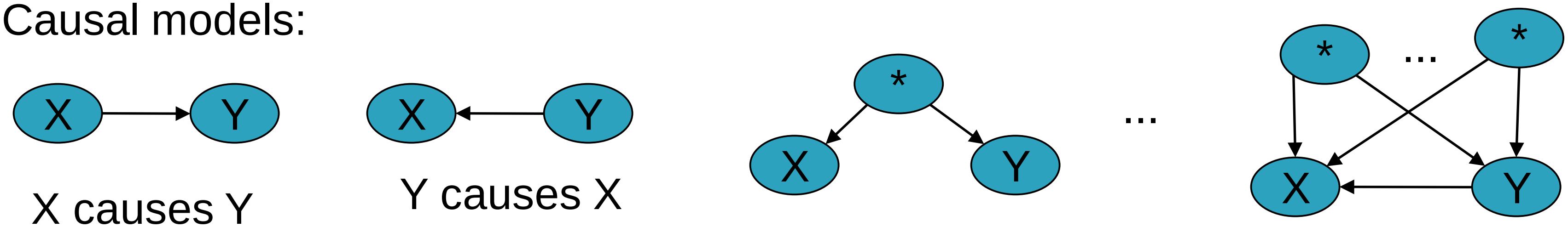
1. deleting the incoming edges to X
2. setting X=x
3. performing standard Bayesian network inference.





Association vs. Causation

Causal models:



X causes Y

Y causes X

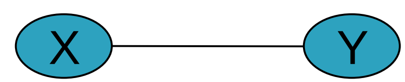
There is a common cause (pure confounding)

Causal effect of Y on X is confounded by many factors

From passive observations:

P(X,Y)

M_P={D(X;Y)}



„X and Y are associated”

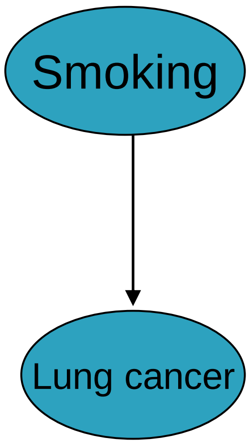
Reichenbach's Common Cause Principle:

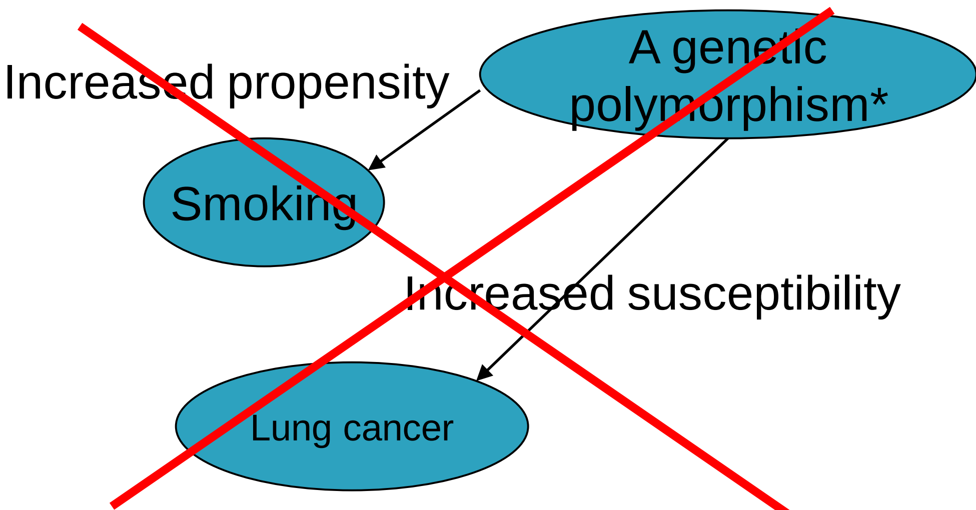
a correlation between events *X* and *Y* indicates either that *X* causes *Y*, or that *Y* causes *X*, or that *X* and *Y* have a common cause.

Local Causal Discovery

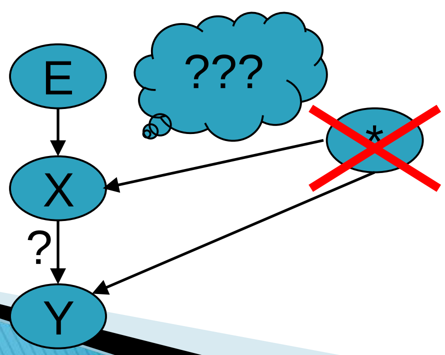
"can we interpret edges as causal relations in the presence of hidden variables?"

- Can we learn causal relations from observational data in presence of confounders???





- Automated, tabula rasa causal inference from (passive) observation is possible, i.e. hidden, confounding variables can be excluded



Summary

- Can we represent exactly (in)dependencies by a BN?
  - ***almost always***
- Can we interpret
  - edges as causal relations
    - with no hidden variables?
      - ***compelled edges as a filter***
    - in the presence of hidden variables?
      - ***Sometimes, e.g. confounding can be excluded in certain cases***
    - in local models as autonomous mechanisms?
      - ***a priori knowledge, e.g. Causal Markov Assumption***
- Can we infer the effect of interventions in a causal model?
  - ***Graph surgery with standard inference in BNs***
- Suggested reading
  - J. Pearl: Causal inference in statistics, 2009