

# Artificial general intelligence

Uncertainty in artificial intelligence:  
from associations to causal reasoning

Antal Péter

ComBine Lab

Artificial Intelligence group

Department of Measurement and Information Systems



# Agenda

- Reminder:
  - Last lecture: axiomatizations in probability theory
  - Now: axiomatic approaches to causal inference ( $\sim$ causality)
- Examples for probabilistic inference
- From associations to direct dependencies
- The ultimate limit of observational learning
- Causal inference
- Learning causal relations and models
- Counterfactual inference

# Probabilistic inference

# Reminder: Naive Bayesian network

- Definition: conditional independence of „effects”  $X_i$  given „cause”  $Y$ .
- Properties:
  - Number of parameters ( $\sim$ model complexity): linear
  - Complexity of inference: linear

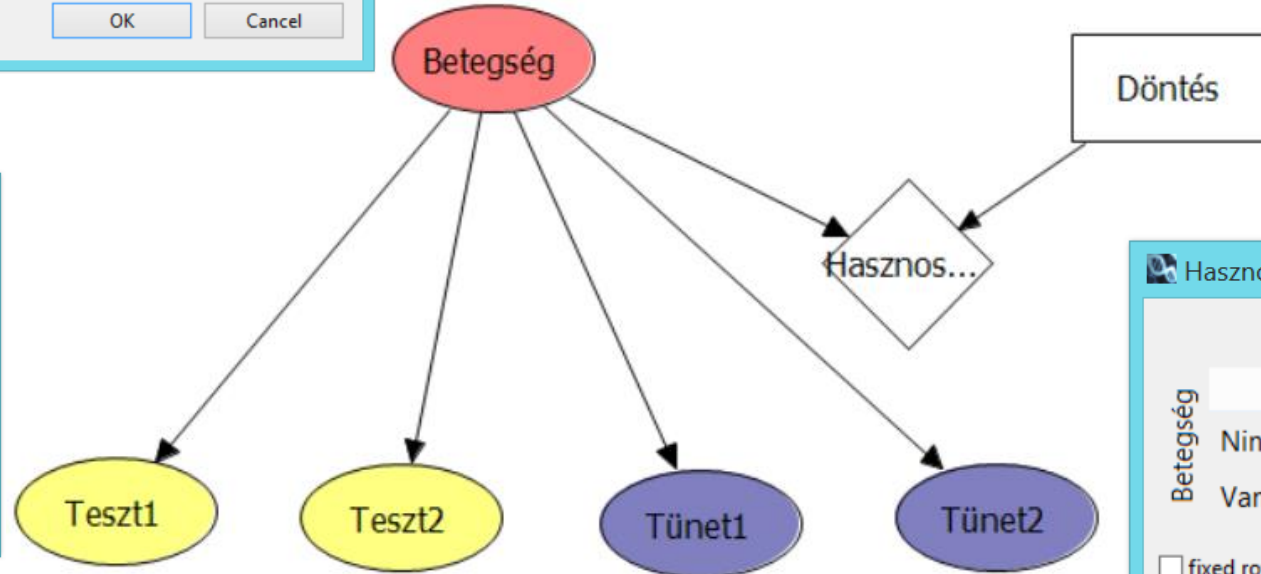
# Döntési (naív) hálózat

P(Betegség)

0	Nincs	Van
0	0.3	0.7

fixed row height  sample size/probability

OK Cancel



P(Teszt1|Betegség)

(Betegség)	Normal	Alacsony	Magas
(Nincs)	0.55	0.45	0.0
(Van)	0.05	0.45	0.5

fixed row height  sample size/probability

OK Cancel

Hasznosság(Betegség, Dö...)

		Döntés	
		Nincs	Van
Betegség	Nincs	0	-100
	Van	-5	10

fixed row height  matrix view

OK Cancel

# Universal (observational) inference

Any question about observable events in the domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the **query variables**  $\mathbf{Y}$  given specific values  $\mathbf{e}$  for the **evidence variables**  $\mathbf{E}$

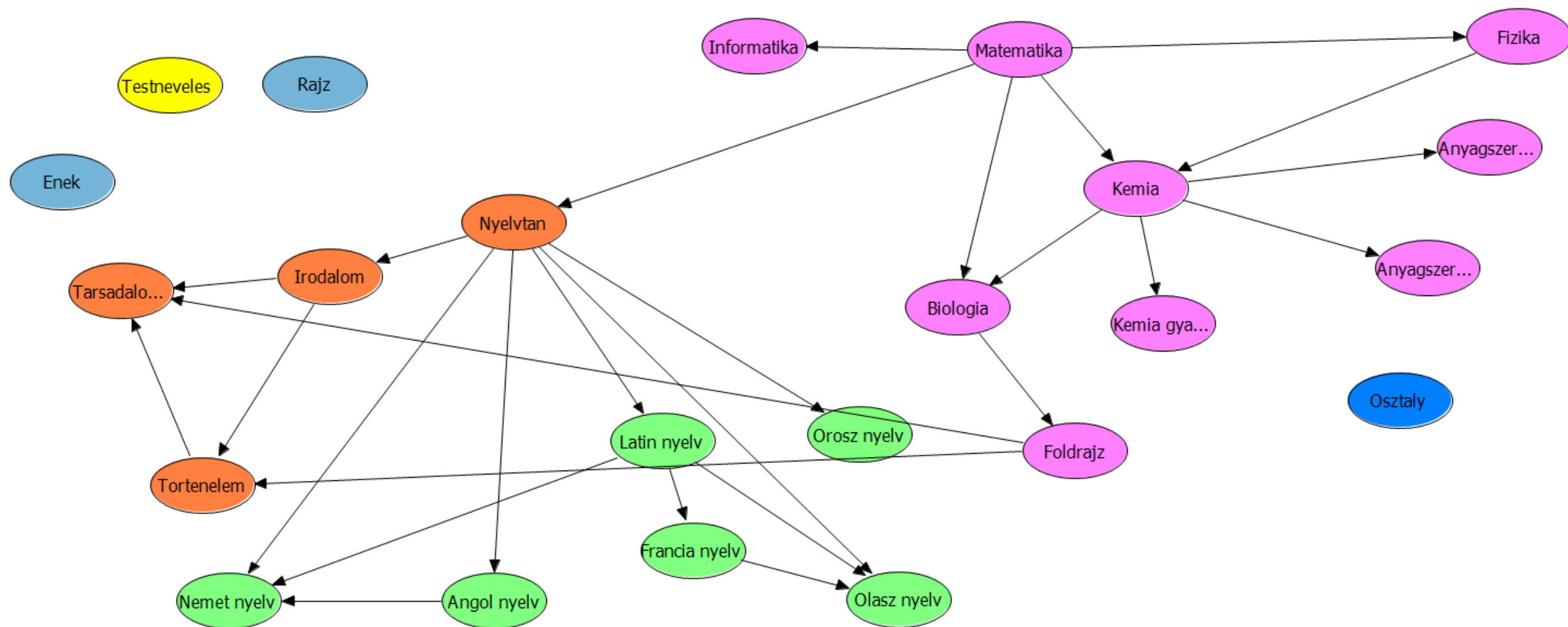
Let the **hidden variables** be  $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the hidden variables:

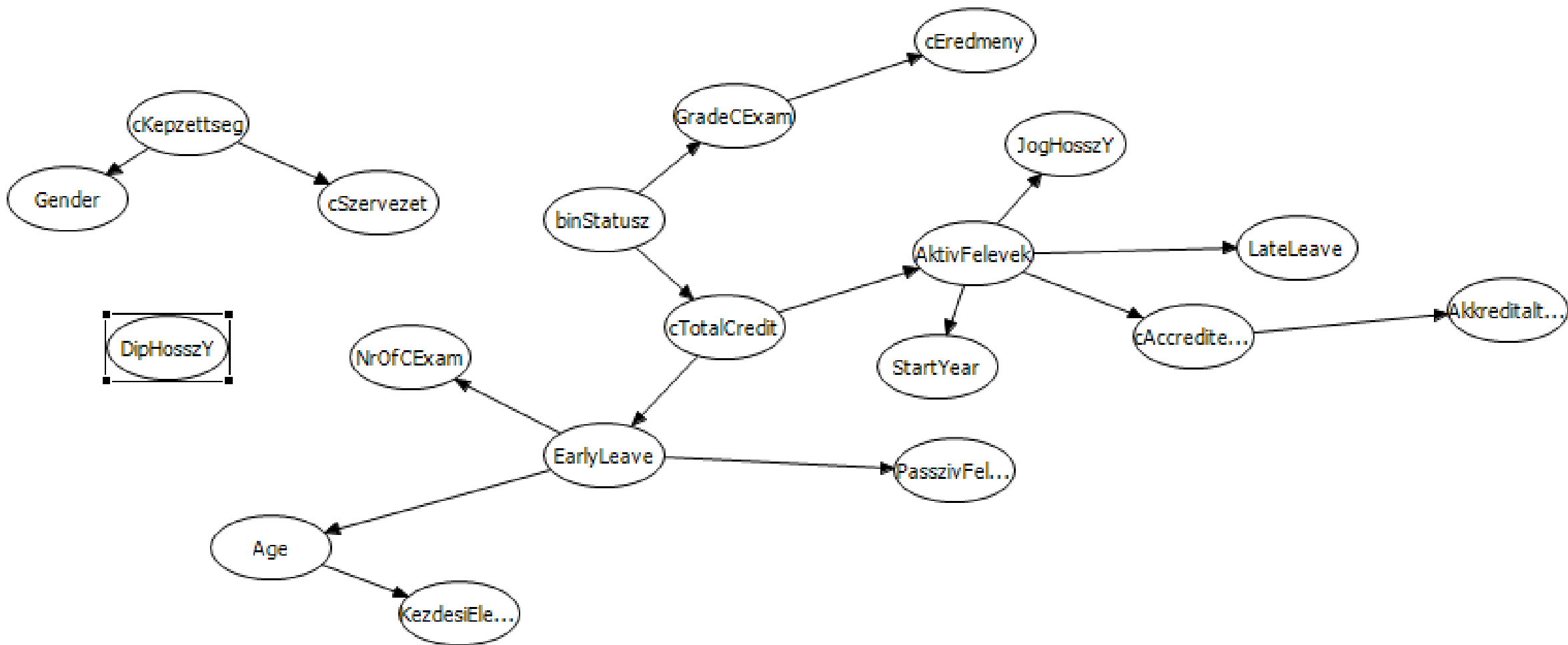
$$P(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha P(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \sum_{\mathbf{h}} P(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$$

- The terms in the summation are joint entries because  $\mathbf{Y}$ ,  $\mathbf{E}$  and  $\mathbf{H}$  together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity  $O(d^n)$  where  $d$  is the largest arity
  2. Space complexity  $O(d^n)$  to store the joint distribution
  3. How to find the numbers for  $O(d^n)$  entries?

# Example: Network of grades („érdemjegyháló”)



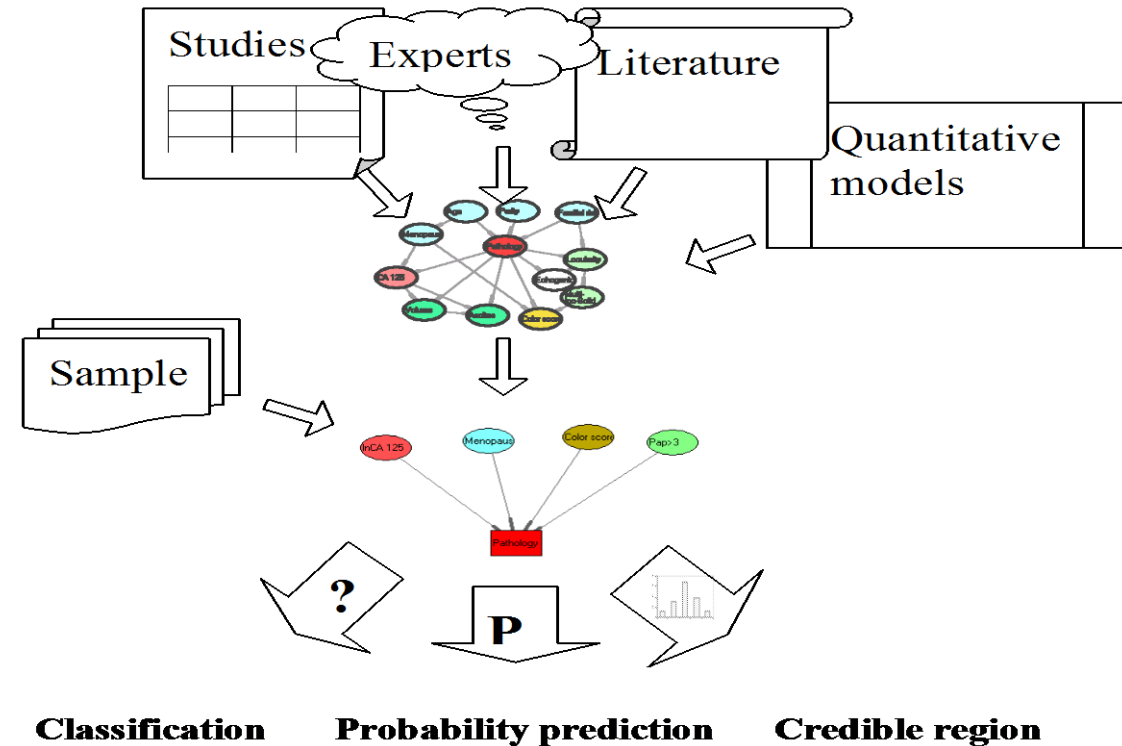
# Example: student attrition (lemorzsolódás)





# Petefészekrák nem-invazív diagnosztikája

- International Ovarian Tumor Analysis (IOTA)
  - 1996-1999: ~300 minta (K.U.Leuven)
  - 1998-2004: 68 paraméter, 1066 minta
- Kérdések:
  - Predikciós teljesítmény
  - Biomarkerek
    - Szükségesség, elégségesség, helyettesíthetőség
    - költséghatékonyság

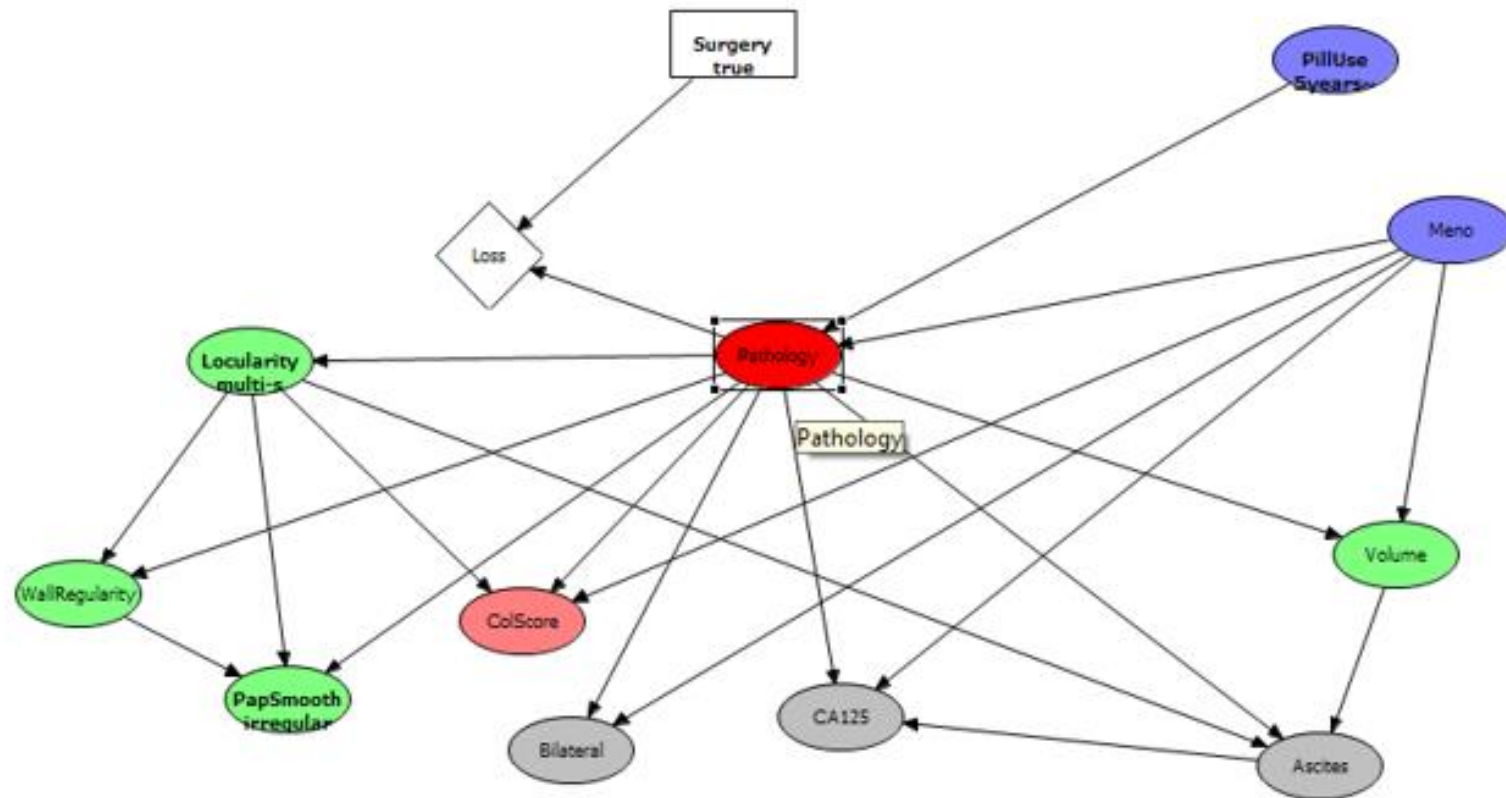


Antal, P. et al. (2000). **How might we combine the information we know** about a mass better? In *Proc. of the 1st Montecarlo Conference on updates in Gynaecology (MCG)* (pp. 1-3).

Timmerman et al. **Terms, definitions and measurements** to describe the sonographic features of adnexal tumors: a consensus opinion from the IOTA group. *Utrasound Obstet Gynecol*, 16:500-505, 2000

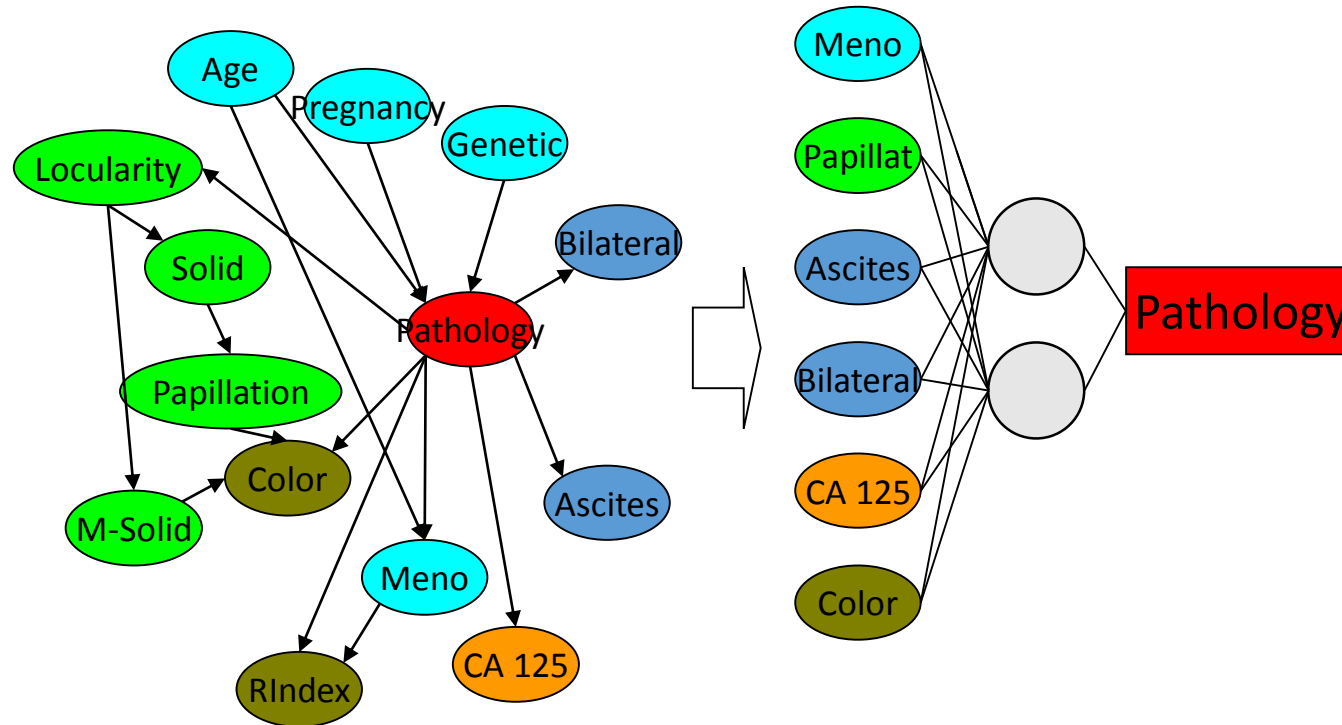
Timmerman, D. (2004). The use of mathematical models to evaluate pelvic masses; **can they beat an expert operator?**. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 18(1), 91-104.

# Döntési hálózatok

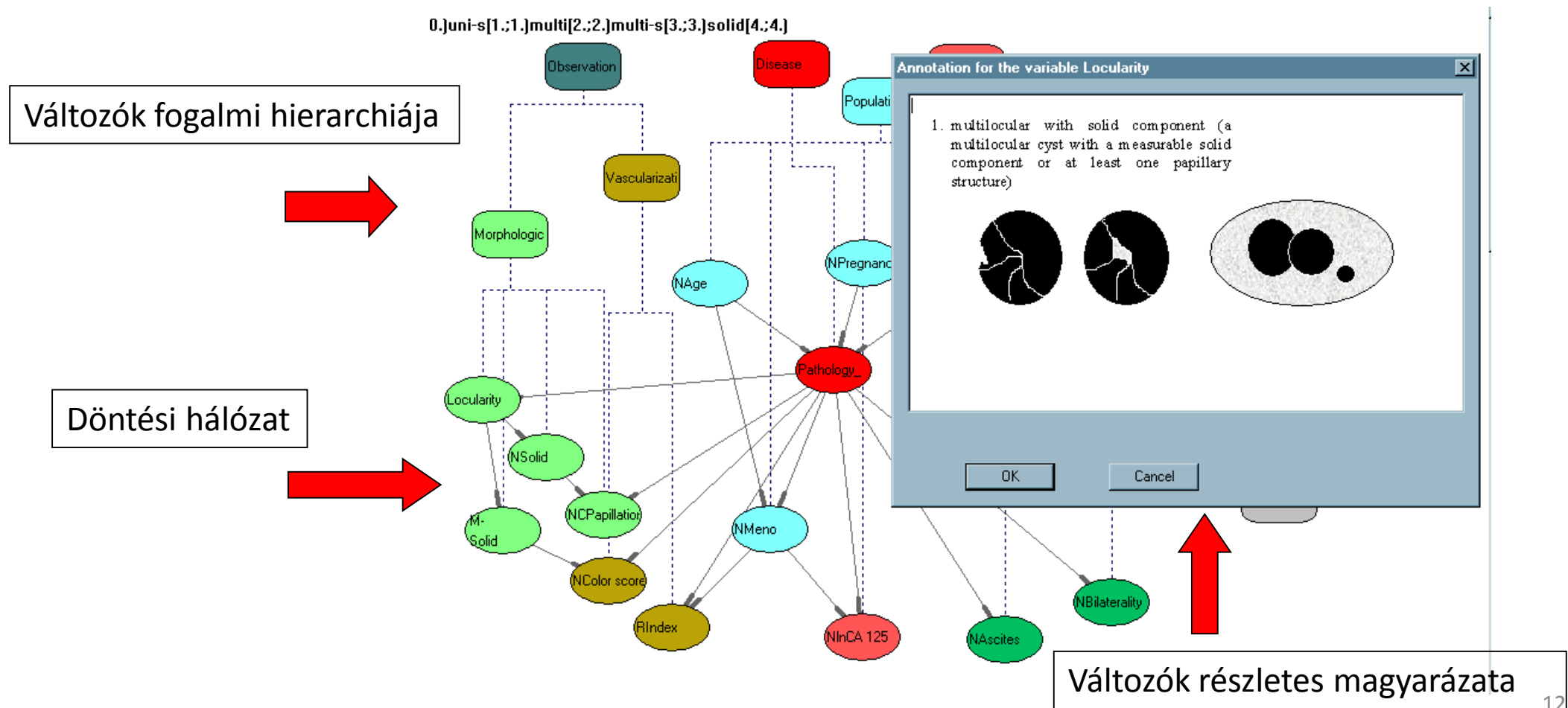


Antal, P., Fannes, G., Timmerman, D., Moreau, Y. and De Moor, B., 2004. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in medicine*, 30(3), pp.257-281.

# „Tudásgazdag” neurális hálózatok



# Bizonyíték-alapú, értelmezhető döntési hálózatok



Antal, P., Mészáros, T., De Moor, B., & Dobrowiecki, T. (2001). Annotated Bayesian Networks: a tool to integrate textual and probabilistic medical knowledge. In *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001* (pp. 177-182). IEEE.

# Érthető, megmagyarázott döntések

Kérdés

Bizonyíték-alapú döntési hálózat

The screenshot displays the BNet software interface. At the top, a menu bar includes 'File', 'Edit', 'Prior acquisition', 'Data preparation', 'Knowledge modelling', 'Inference', 'Model refinement', 'Model transformation', 'View', 'Options', 'Window', and 'Help'. Below the menu, a window titled 'Model-based IR for C:\research\ABN\model\OC\_large\_v1.bn' shows a query: '0-50[40.;50.]50<=[50.;100]'. A 'Result list' window is open, displaying a list of search results with titles like 'CA 125 in gynecological pathology - a review' and 'Serum CA 125 levels and surgical findings in patients undergoing secondary operations for epithelial ovarian cancer'. A 'Result: ml\_oc\_v1:AN92267045' window is also open, showing a detailed view of a specific article with its title, author, journal, and abstract. The abstract text is highlighted in red. To the right, a decision network diagram is visible, with nodes labeled 'Indirect Fa', 'Hormonal', 'PillUse', 'HormTher', 'PostMeno', 'Meno', and 'Infant'. Red arrows point from the labels 'Kérdés', 'Bizonyíték-alapú döntési hálózat', 'Magyarozat', and 'Szakcikkek' to their respective elements in the screenshot.

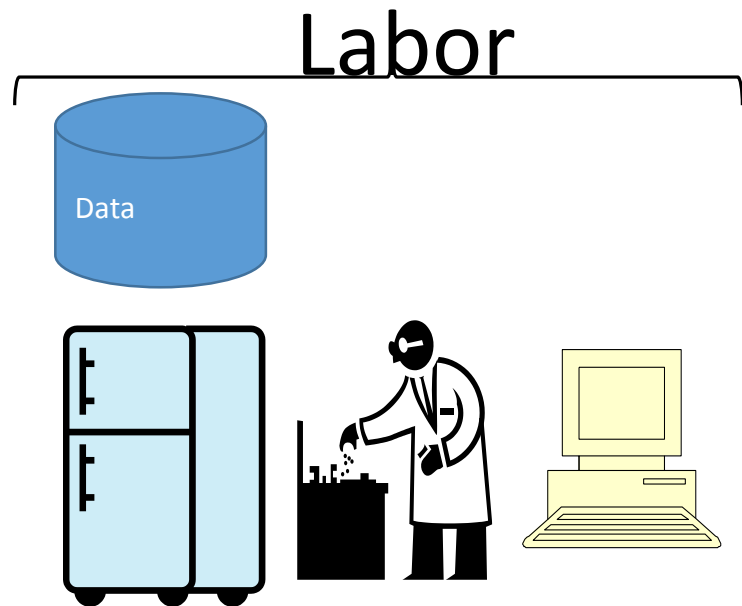
Magyarozat

Szakcikkek

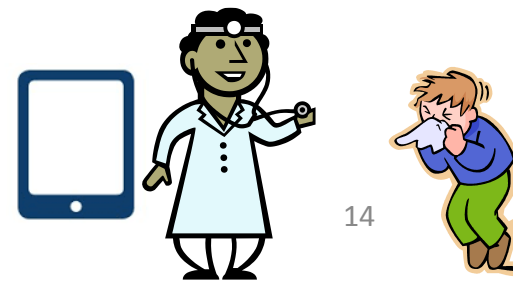
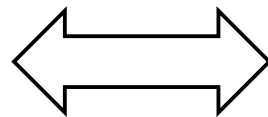
Antal, P., De Moor, B., Timmerman, D., Mészáros, T., & Dobrowiecki, T. (2002). Domain knowledge based information retrieval language: an application of annotated Bayesian networks in ovarian cancer domain. In *Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)* (pp. 213-218). IEEE.

# Adaptív klinikai laboratóriumi diagnosztika

1. Döntéstámogató rendszer
2. Költséghatékonysági kontroll
3. Orvoskontroll biztosítása
4. Labormérések indításának segítése adott minta esetén
5. Labor kapacitásának automatikus figyelembevétele



Tájékoztatás a keletkező eredményekről és javaslatokról



# Independence models

# 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:  $D_p(X;Y|Z) = \text{def} = \neg I_p(X;Y|Z)$

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:

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

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 semi-graphoid axioms

1. Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\mathbf{X}; \mathbf{Y} | \mathbf{Z}) \text{ iff } I_p(\mathbf{Y}; \mathbf{X} | \mathbf{Z})$$

2. Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\mathbf{X}; \mathbf{Y} \cup \mathbf{W} | \mathbf{Z}) \Rightarrow I_p(\mathbf{X}; \mathbf{Y} | \mathbf{Z}) \text{ and } I_p(\mathbf{X}; \mathbf{W} | \mathbf{Z})$$

3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\mathbf{X}; \mathbf{Y} \cup \mathbf{W} | \mathbf{Z}) \Rightarrow I_p(\mathbf{X}; \mathbf{Y} | \mathbf{Z} \cup \mathbf{W})$$

4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\mathbf{X}; \mathbf{Y} | \mathbf{Z}) \text{ and } I_p(\mathbf{X}; \mathbf{W} | \mathbf{Z} \cup \mathbf{Y}) \Rightarrow I_p(\mathbf{X}; \mathbf{Y} \cup \mathbf{W} | \mathbf{Z})$$

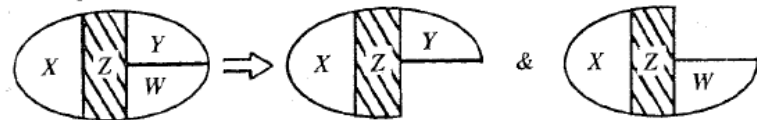
# Graphoids

Graphoids: Semi-graphoids+Intersection (holds only in strictly positive distribution)

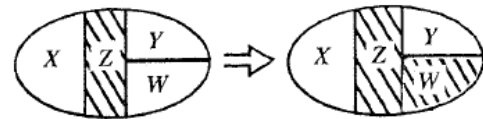
Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies.

$$I_p(X; Y|Z \cup W) \text{ and } I_p(X; W|Z \cup Y) \Rightarrow I_p(X; Y \cup W|Z)$$

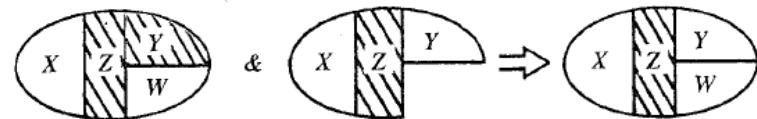
Decomposition



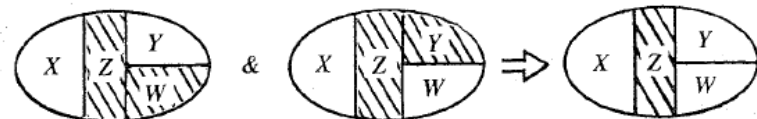
Weak Union



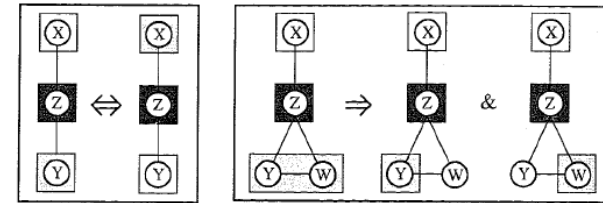
Contraction



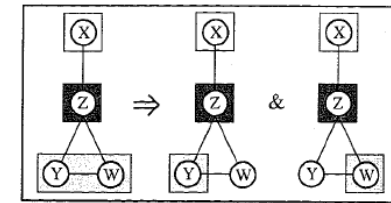
Intersection



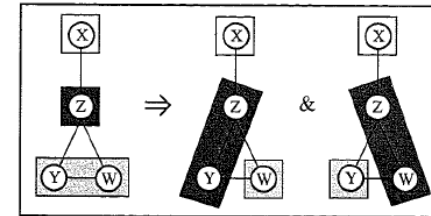
J.Pearl: Probabilistic Reasoning in intelligent systems, 1998



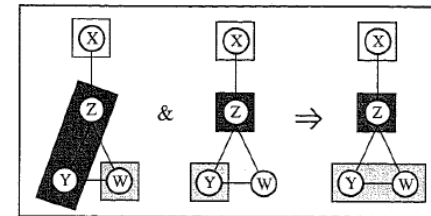
(a) Symmetry



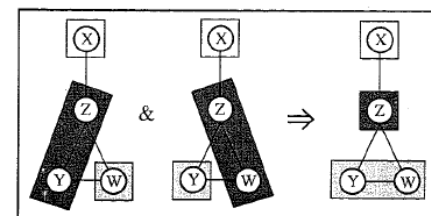
(b) Decomposition



(c) Weak Union



(d) Contraction



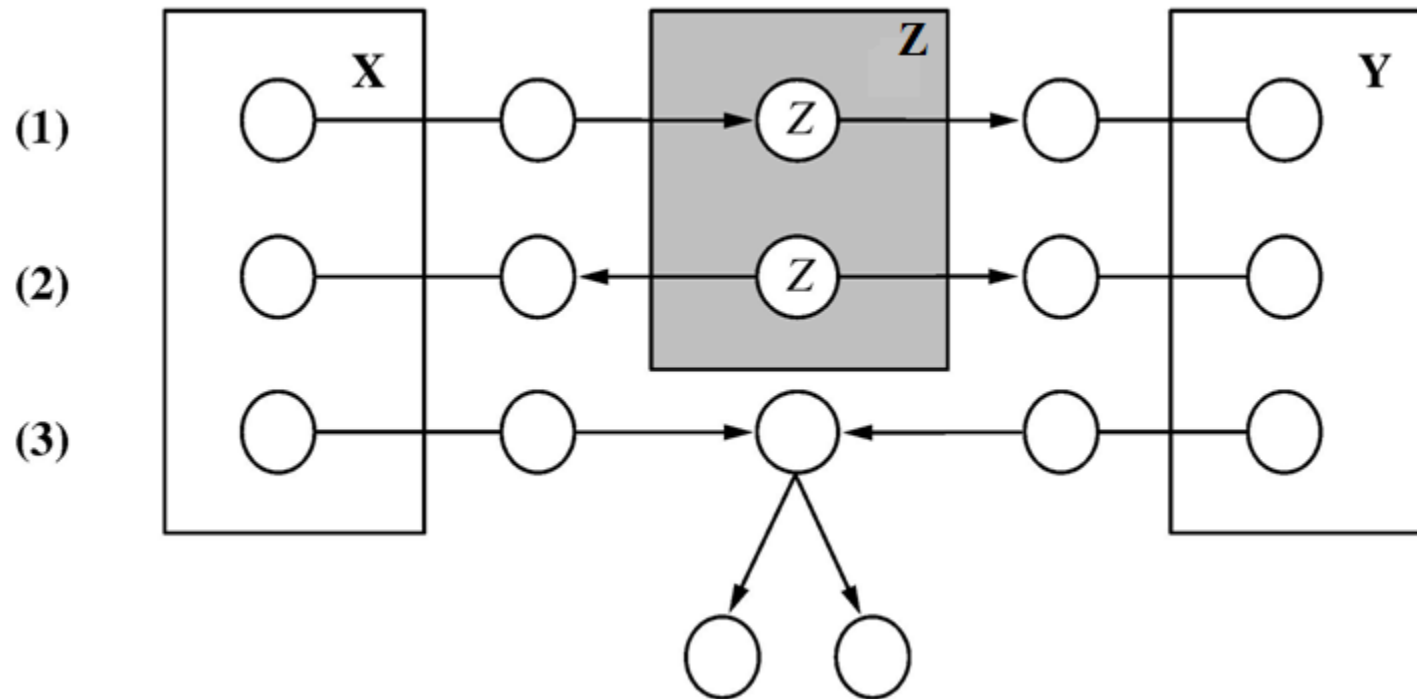
(e) Intersection

# Separation in undirected graphs

$I_G(X;Y|Z)$  denotes that  $X$  is separated from  $Y$  by  $Z$  in undirected graph  $G$ , i.e. every path between  $X$  and  $Y$  is blocked by  $Z$  (it contains a node from  $Z$ ).

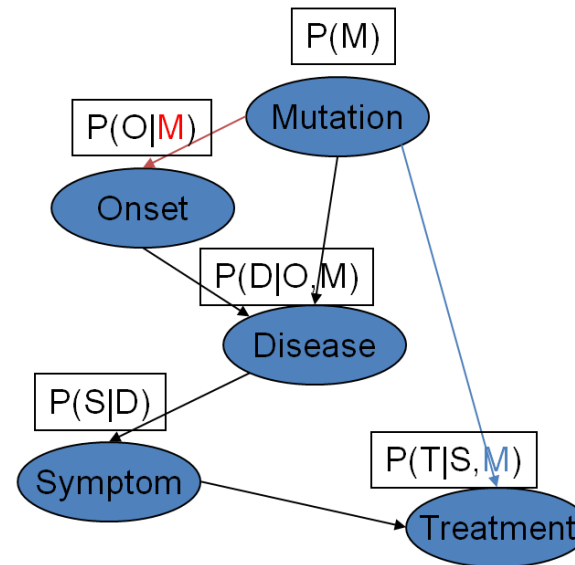
# Directed separation in directed graphs

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



# Bayesian networks: three facets

3. Concise representation of joint distributions

$$P(M, O, D, S, T) = P(M)P(O | M)P(D | O, M)P(S | D)P(T | S, M)$$


1. Causal model

$M_p = \{I_{p,1}(X_1; Y_1 | Z_1), \dots\}$

2. Graphical representation of (in)dependencies

# Bayesian networks

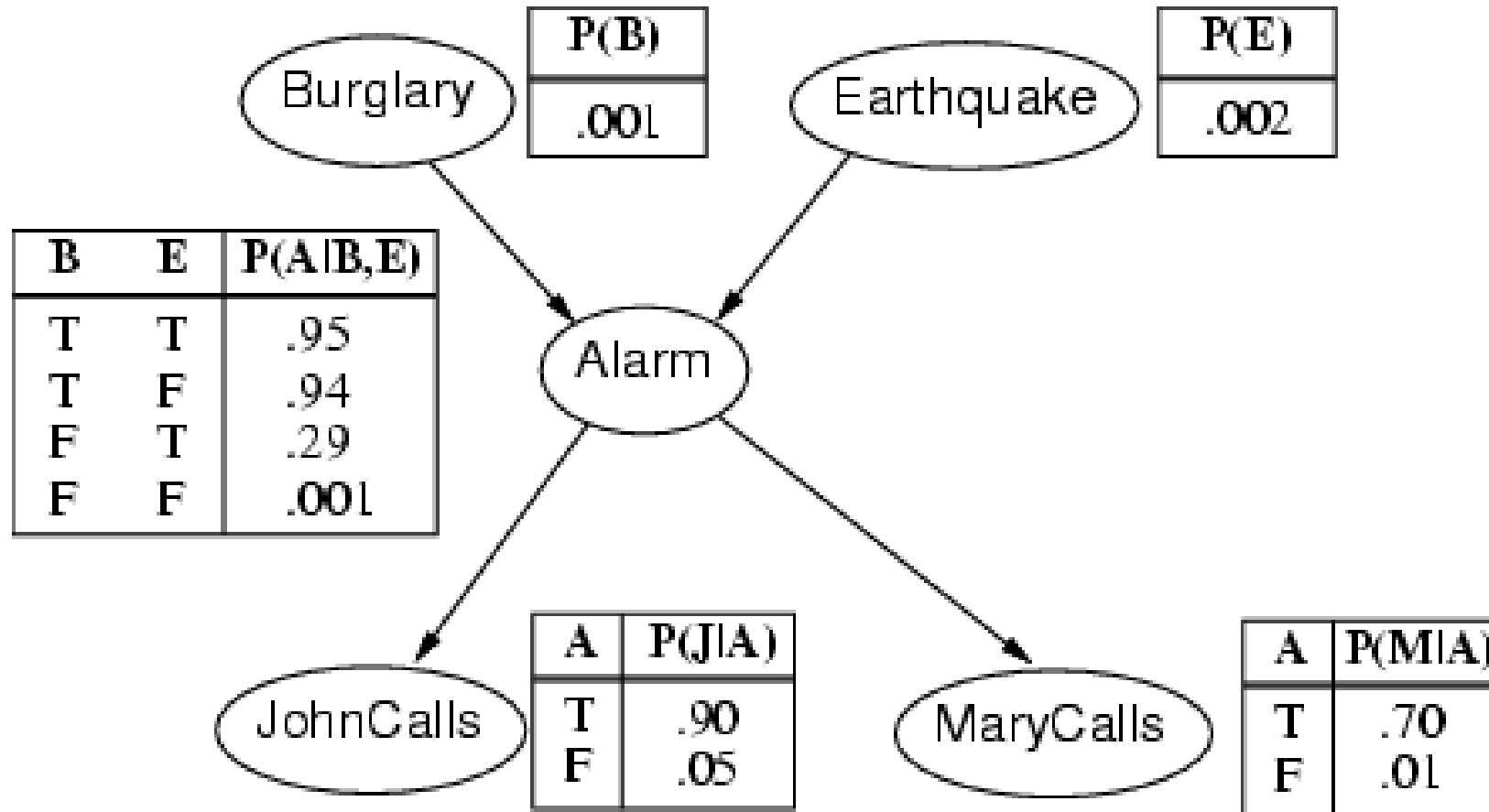
- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- Syntax:
  - a set of nodes, one per variable
  - 
  - a directed, acyclic graph (link  $\approx$  "directly influences")
  - a conditional distribution for each node given its parents:  
$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$
- In the simplest case, conditional distribution represented as a **conditional probability table** (CPT) giving the distribution over  $X_i$  for each combination of parent values

# Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary, Earthquake, Alarm, JohnCalls, MaryCalls*
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call



# Example contd.



# Constructing Bayesian networks

- 1. Choose an ordering of variables  $X_1, \dots, X_n$
- 2. For  $i = 1$  to  $n$ 
  - add  $X_i$  to the network
  - select parents from  $X_1, \dots, X_{i-1}$  such that
$$\mathbf{P}(X_i \mid \text{Parents}(X_i)) = \mathbf{P}(X_i \mid X_1, \dots, X_{i-1})$$

This choice of parents guarantees:

$$\begin{aligned} \mathbf{P}(X_1, \dots, X_n) &= \prod_{i=1}^n \mathbf{P}(X_i \mid X_1, \dots, X_{i-1}) && \text{//(chain rule)} \\ &= \prod_{i=1}^n \mathbf{P}(X_i \mid \text{Parents}(X_i)) && \text{//(by construction)} \end{aligned}$$

# Representation of independencies

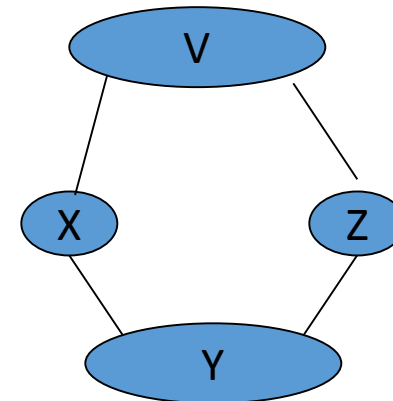
D-separation provides a sound and complete, computationally efficient algorithm to read off an (in)dependency model consisting the independencies that are valid in all distributions Markov relative to  $G$ , that is  $\forall X, Y, Z \subseteq V$

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

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

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

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



An almost always complete calculus for independencies

# Markov conditions

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

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

where  $\text{Pa}(X_i)$  denotes the parents of  $X_i$  in  $G$ .

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

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

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

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

$$\forall i = 1, \dots, n : (X_i \perp\!\!\!\perp \text{Nondescendants}(X_i) | \text{Pa}(X_i))_P, \quad (8)$$

where  $\text{Nondescendants}(X_i)$  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  $G$ ,*

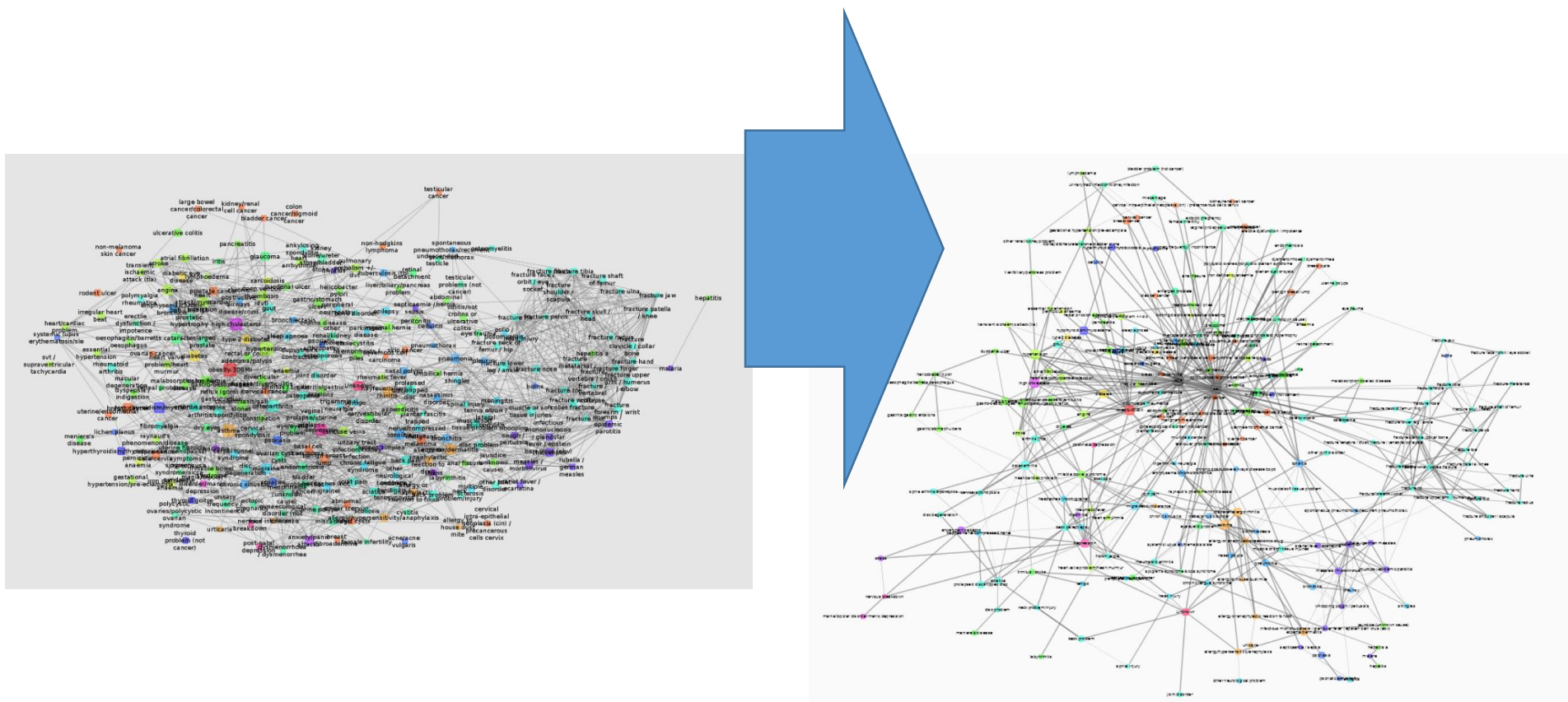
*O  $P$  obeys the ordered Markov condition w.r.t.  $G$ ,*

*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  $F, O, L, G$ ).*

# Multimorbidity network



Marx, P., Antal, P., Bolgar, B., Bagdy, G., Deakin, B. and Juhász, G., 2017. Comorbidities in the diseaseome are more apparent than real: What Bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, 13(6), p.e1005487.

# Relativity of the interpretations

1. The presence of unobserved (hidden) variables as potential confounders.
2. Selection bias can occur if the observation depends on the joint combination of otherwise independent events, inducing non-causal dependencies between them.
3. The mixture of causal models, if conditionally both  $X$  causes  $Y$  and vice versa. A similar problem is the presence of feedback (and indirectly temporality).
4. Global physical and semantic constraints between the variables.
5. Stability can be also questioned, because of deterministic dependencies, resulting in the lack of guarantee for the uniqueness and exactness of the representation.
6. The (in)dependencies are relative to the set of variables and specifically, also to the values of the variables



Towards causal inference

# Principles of causality

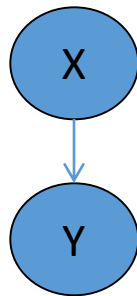
- strong association,
  - X precedes temporally Y,
  - plausible explanation without alternative explanations based on confounding,
  - necessity (generally: if cause is removed, effect is decreased or actually: y would not have been occurred with that much probability if x had not been present),
  -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

# Questions

- Can we represent exactly (in)dependencies by a BN?
  - From a causal model? Suff.&nec.?
- Can we interpret
  - 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?

## Motivation: from observational inference...

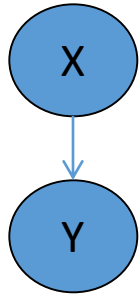
- In a Bayesian network, any query can be answered corresponding to passive observations:  $p(Q=q | E=e)$ .
  - What is the (conditional) probability of  $Q=q$  given that  $E=e$ .
  - Note that  $Q$  can precede temporally  $E$ .



- ▶ 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:  $\text{do}(X=x)$  as set  $X$  to  $x$ .
- What is the relation of  $p(Q=q | E=e)$  and  $p(Q=q | \text{do}(E=e))$ ?



- ▶ Specification:  $p(X), p(Y|X)$
- ▶ Joint distribution:  $p(X,Y)$
- ▶ Inferences:
  - ▶  $p(Y|X=x)=p(Y|\text{do}(X=x))$
  - ▶  $p(X|Y=y)\neq p(X|\text{do}(Y=y))$

- 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 had 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

$$P(X_1, \dots, X_n \mid \text{Structure, parameters})$$

## 1. Representation of parameters

„small number of parameters“

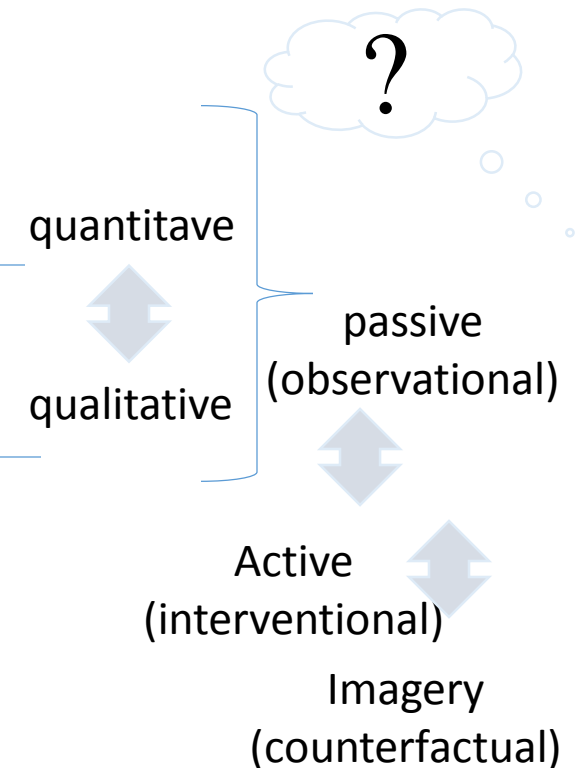
## 2. Representation of independencies

„what is relevant for diagnosis“

## 3. Representation of causal relations

„what is the effect of a treatment“

## 4. Representation of possible worlds



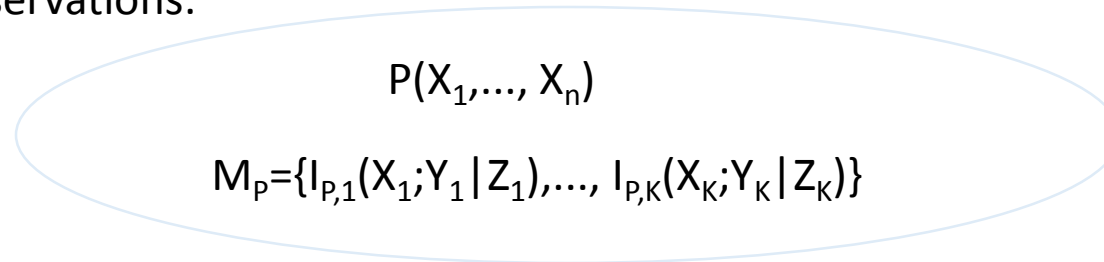
# Observational equivalence of causal models

Causal models:



J.Pearl:  
~ „3D objects”

From passive observations:



„2D projection”

Different causal models can have the same independence map!

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

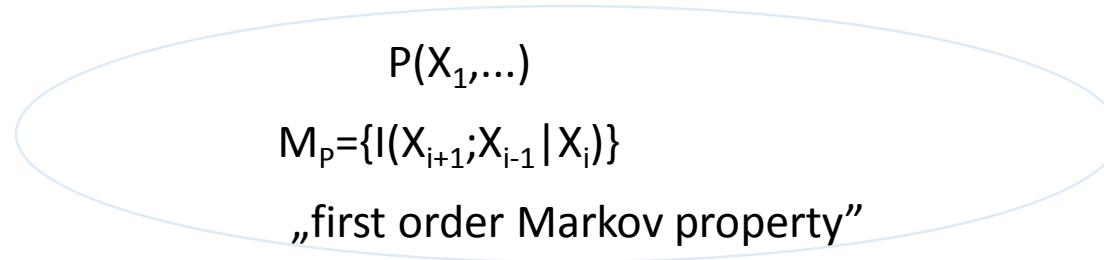


# Association vs. Causation: Markov chain

Causal models:

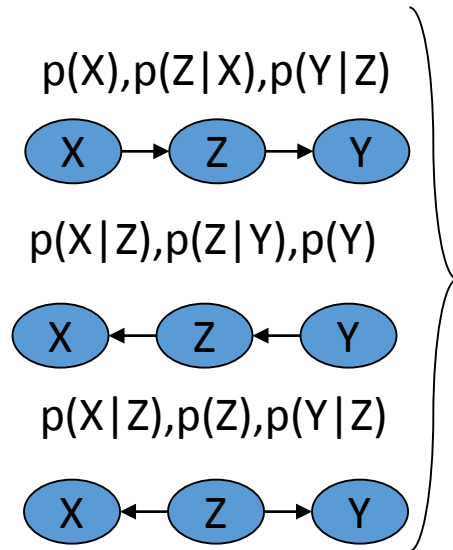


Markov chain

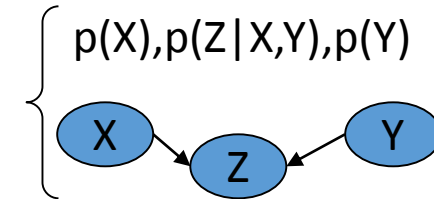


Flow of time?

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



“transitive”  $M \neq$  „intransitive”  $M$



„v-structure”

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

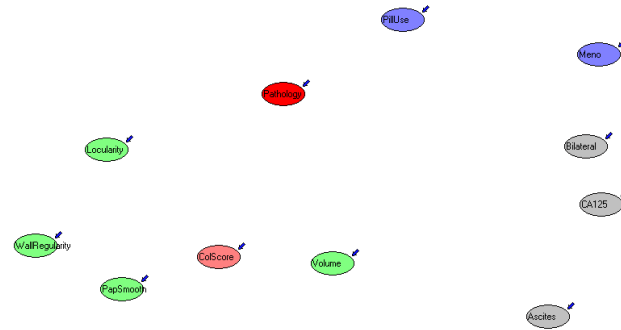
$$M_p = \{D(X;Z), D(Y;Z), I(X;Y), D(X;Y|Z)\}$$

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

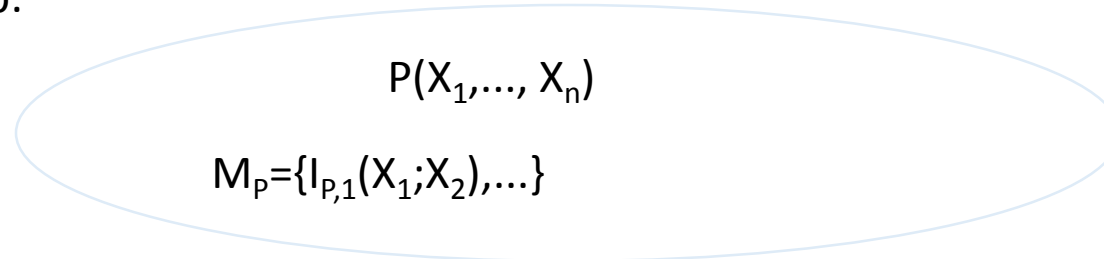
Ever(?): present knowledge renders past states conditionally independent.  
(backward/atemporal confounding)

# Observational equivalence: total independence

„Causal” model:

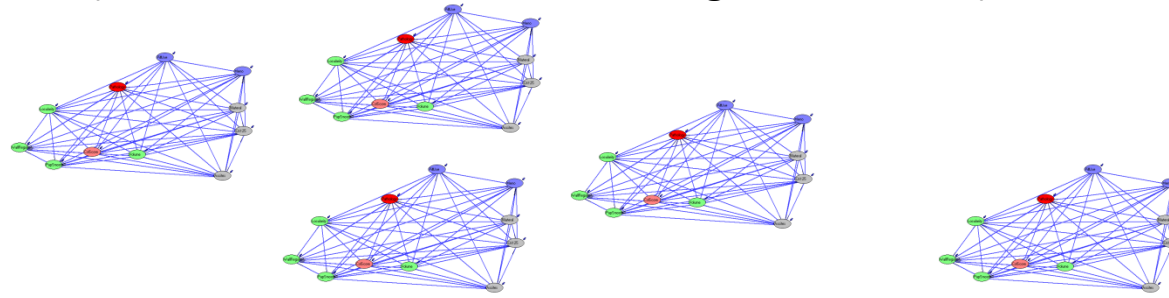


Dependency map:



# Observational equivalence: full dependence

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



Dependency map:



One-to-many relation

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

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

# 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.  $(X \perp\!\!\!\perp Y|Z)_{G_1} \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$ ).*

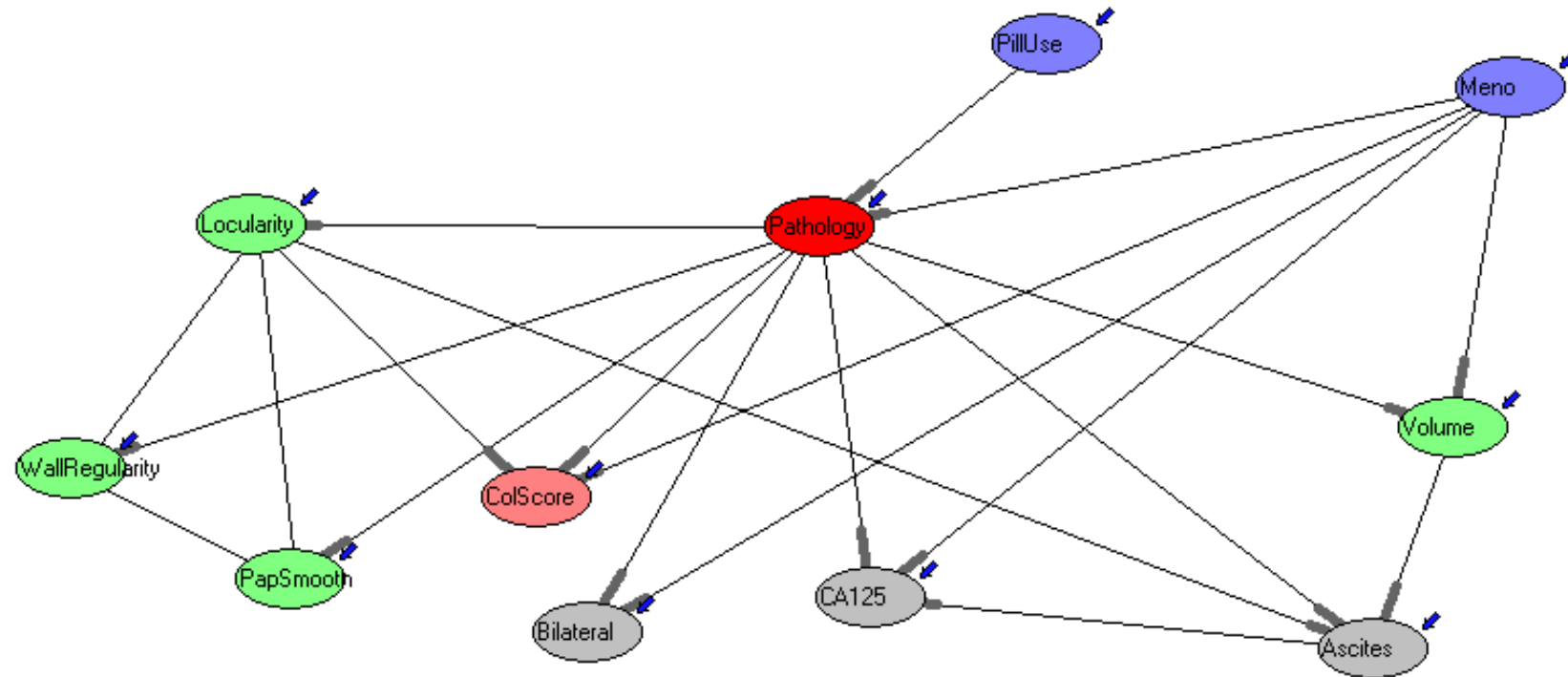
The implied equivalence classes may contain  $n!$  number of DAGs (e.g. all the full networks representing no independencies) or just 1.

**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).*

# Compelled edges and PDAG

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



# The Causal Markov Condition

- A DAG is called a *causal structure* over a set of variables, if each node represents a variable and edges direct influences. A *causal model* is a causal structure extended with local probabilistic models.
- A causal structure  $G$  and distribution  $P$  satisfies the Causal Markov Condition, if  $P$  obeys the local Markov condition w.r.t.  $G$ .
- The distribution  $P$  is said to stable (or faithful), if there exists a DAG called *perfect map* exactly representing its (in)dependencies (i.e.  $I_G(X;Y|Z) \Leftrightarrow I_P(X;Y|Z) \forall X,Y,Z \subseteq V$ ).
- CMC: **sufficiency** of  $G$  (there are no extra, acausal dependencies)
- Faithfulness/stability: **necessity** of  $G$  (there are no extra, parametric independencies)

# Interventional inference in causal Bayesian networks

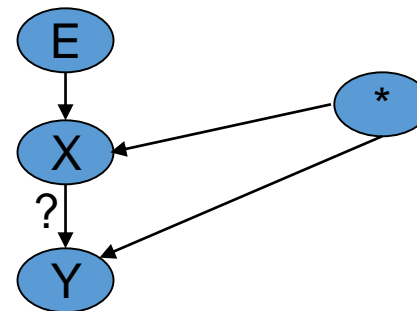
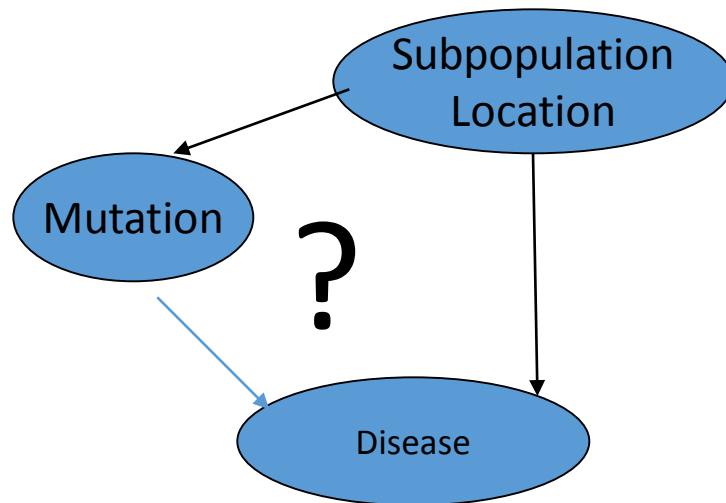
- (Passive, observational) inference
  - $P(\text{Query} | \text{Observations})$
- **Interventionist inference**
  - $P(\text{Query} | \text{Observations}, \text{Interventions})$
- Counterfactual inference
  - $P(\text{Query} | \text{Observations}, \text{Counterfactual conditionals})$



# Interventions and graph surgery

If  $G$  is a causal model, then compute  $p(Y | \text{do}(X=x))$  by

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



# Learning causal relations and models

# Inductive Causation (asymptotic, no hidden)

1. *Skeleton*: Construct an undirected graph (skeleton), such that variables  $X, Y \in V$  are connected with an edge iff  $\forall S(X \perp\!\!\!\perp Y|S)_p$ , where  $S \subseteq V \setminus \{X, Y\}$ .
2. *v-structures*: Orient  $X \rightarrow Z \leftarrow Y$  iff  $X, Y$  are nonadjacent,  $Z$  is a common neighbour and  $\neg \exists S$  that  $(X \perp\!\!\!\perp Y|S)_p$ , where  $S \subseteq V \setminus \{X, Y\}$  and  $Z \in S$ .
3. *propagation*: Orient undirected edges without creating new v-structures and directed cycle.

## Theorem

*The following four rules are necessary and sufficient.*

$R_1$  if  $(a \neq c) \wedge (a \rightarrow b) \wedge (b - c)$ , then  $b \rightarrow c$

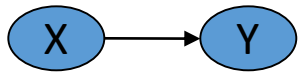
$R_2$  if  $(a \rightarrow c \rightarrow b) \wedge (a - b)$ , then  $a \rightarrow b$

$R_3$  if  $(a - b) \wedge (a - c \rightarrow b) \wedge (a - d \rightarrow b) \wedge (c \neq d)$ , then  $a \rightarrow b$

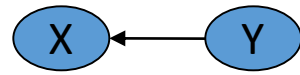
$R_4$  if  $(a - b) \wedge (a - c \rightarrow d) \wedge (c \rightarrow d \rightarrow b) \wedge (c \neq b) \wedge (a - d)$ , then  $a \rightarrow b$

# Association vs. Causation

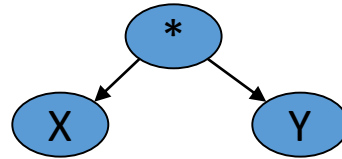
Causal models:



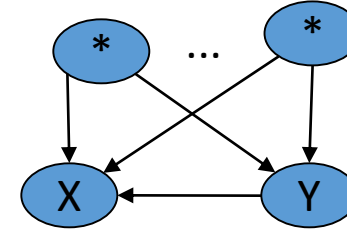
X causes Y



Y causes X

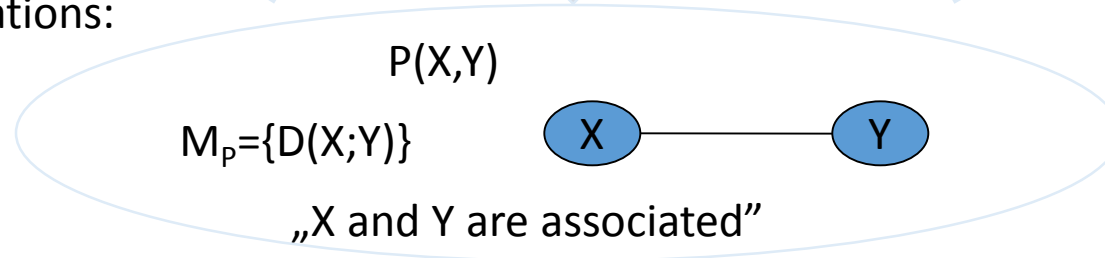


There is a common cause  
(pure confounding)



Causal effect of Y on X  
is confounded by many  
factors

From passive observations:



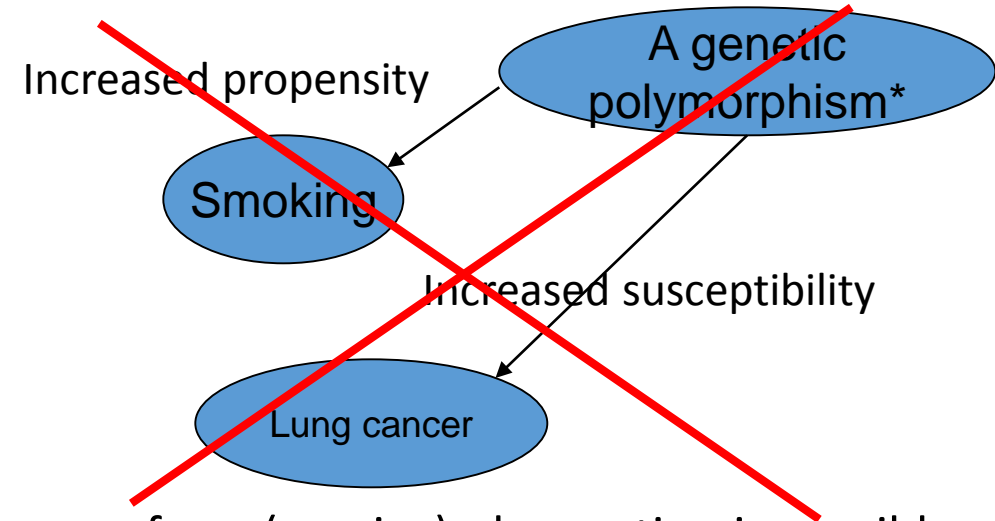
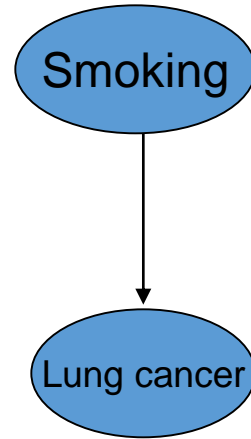
## Reichenbach's Common Cause Principle:

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.

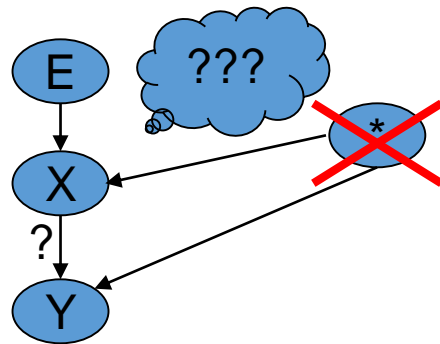
# Local Causal Discovery

“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



# A deterministic concept of causation

- H.Simon
  - $X_i = f_i(X_1, \dots, X_{i-1})$  for  $i=1..n$
  - In the linear case the system of equations indicates a natural causal ordering (flow of time?)

					X
				X	X
			X	X	X
		X	X	X	X
	....				



The probabilistic conceptualization is its generalization:

$$P(X_i | X_1, \dots, X_{i-1}) \sim X_i = f_i(X_1, \dots, X_{i-1})$$

*A posteriori* probability of a „causal” ordering...

Towards counterfactual inference

# Functional (causal) Bayesian network

The axiomatic foundation for the graph surgery semantics of the  $P(.|do(.), .)$  notation.

## Definition

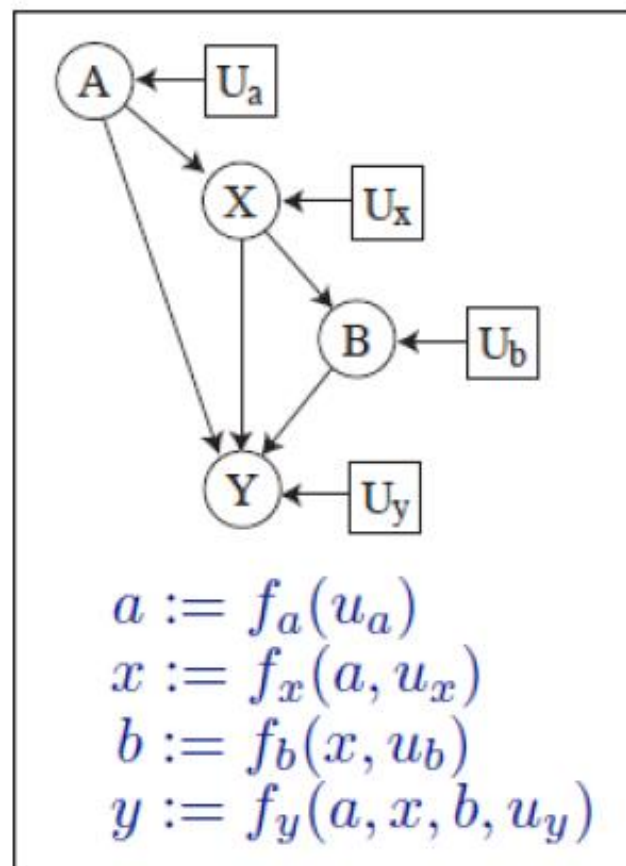
Let  $p(V|do(\mathbf{x}))$  denote an interventional distribution corresponding to setting variable(s)  $\mathbf{X} \subseteq V$  to value  $\mathbf{x}$  and  $P_*$  the set of all interventional distributions (including  $p(V|do(0))$  the observational target distribution without intervention). A DAG  $G$  is said to be a causal Bayesian network compatible with  $P_*$  iff for each  $p(V|do(\mathbf{x})) \in P_*$  the following three conditions hold

1.  $p(V|do(\mathbf{x}))$  is Markov relative to  $G$ ,
2.  $\forall X_i \in \mathbf{X} p(x_i|do(\mathbf{x})) = 1$  if value  $x_i$  and  $\mathbf{x}$  is compatible,
3.  $\forall X_i \notin \mathbf{X} p(x_i|pa_i, do(\mathbf{x})) = p(x_i|pa_i)$  if value(s)  $pa_i$  and  $\mathbf{x}$  is compatible.



# Counterfactuals I.

- Observe  $X = x$  and  $Y = y$
- What is the probability, that  $Y$  would have attained the value  $y'$  if  $X$  had been  $x'$ ? (here  $y$  and  $y'$  can be equal but  $x' \neq x$ )
- Variables  $A$  and  $B$  can be either observed or hidden, but **the full model (graph, functions, and  $P(\mathbf{U})$ ) is assumed to be known**
- Interpreting the question: We assume a **minimal** change of mechanism, i.e. we set  $X$  into state  $x'$  without changing anything else, i.e:  $\text{do}(X = x')$
- Interpreting the question: We assume that the disturbance variables  $\mathbf{U} \setminus U_x = \{U_a, U_b, U_y\}$  are **persistent**, i.e. do not change

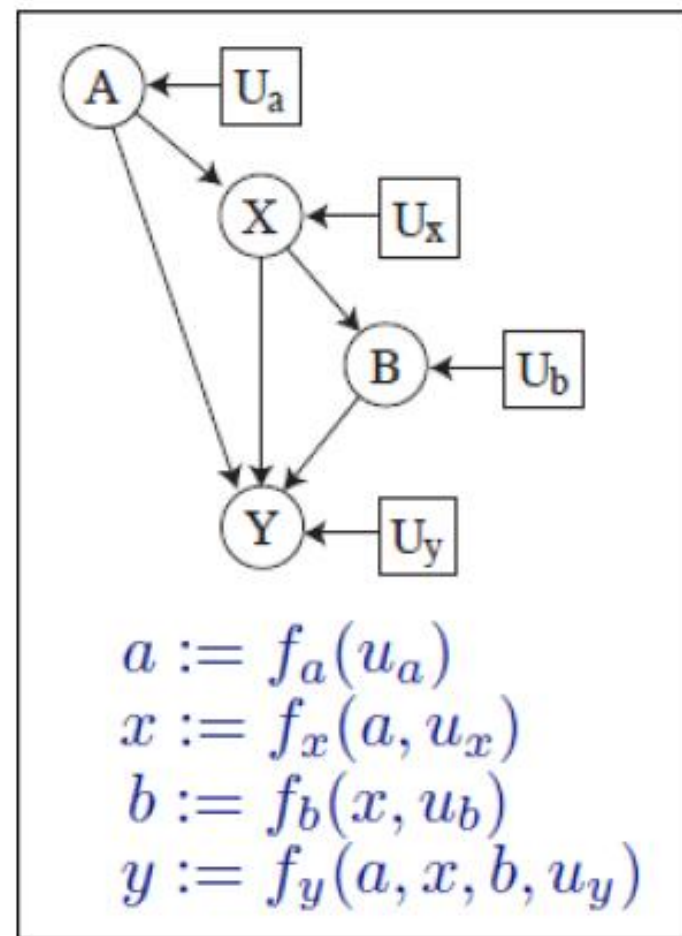


## Counterfactuals II.

- With these specifications, we have a well defined probability:

$$P(Y_{x'} = y' \mid X = x, Y = y)$$

$Y_x(u)$  = 'the value of  $Y$ , when the disturbance variables attain the values  $u$  and  $X$  is **set** to equal  $x$ '.



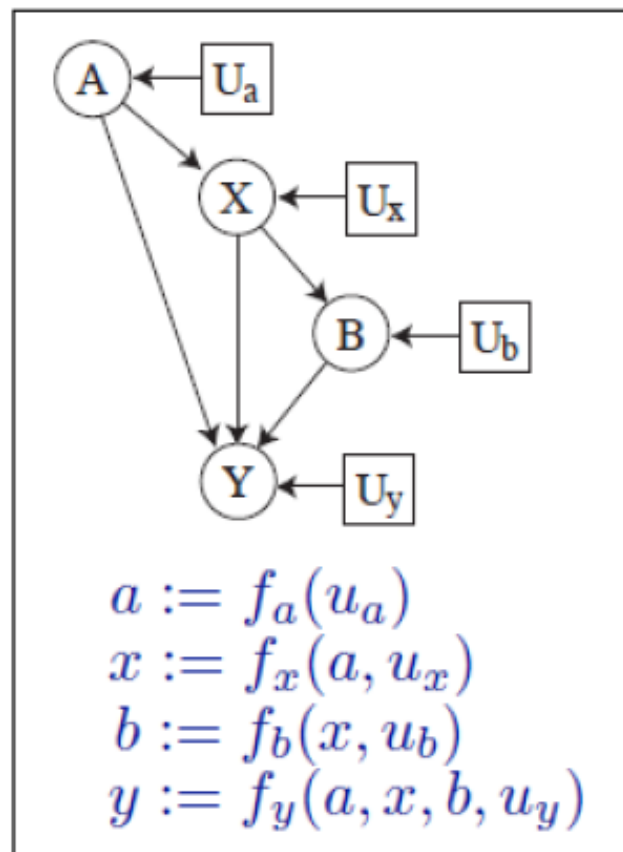
## Counterfactuals III.

- But, how to calculate  $P(Y_{x'} = y' \mid X = x, Y = y)$ ?

(Pearl theorem 7.1.7)

Three steps:

- (‘abduction’): Calculate the probability distribution over all disturbances, given the evidence  $e$ , i.e.  $P(\mathbf{U} \mid e)$
- (‘action’): Change the model by the intervention  $\text{do}(X = x')$ , i.e. remove all arrows into  $X$  and set its value to  $x'$
- (‘prediction’) Using the updated model, and the probability distribution  $P(\mathbf{U} \mid e)$ , calculate  $P(Y = y')$



# Summary

- Independence models
- Probabilistic graphical models
  - Bayesian networks
  - Causal interpretation
- Causal inference
- Counterfactual inference