Bayesian Networks

Probabilistic Graphical Models in AI

Introduction

Peter Lucas

peterl@cs.ru.nl

LIACS

Leiden University

Course Organisation

- Lecturer: Peter Lucas
- Where I am located: Room 123, Snellius
- Structure of course:
  - Lectures
  - Seminar: group research, individual scientific paper, and discussions
  - Practical assignment: develop your own Bayesian network; experiment with learning (structure and classifiers)
- Assessment:
  - Exam: 35%; seminar: 35% (or Exam only: 60%)
  - Practical assignment 1 and 2: 15% each (or 20% each when no seminar)
- Course information: www.cs.ru.nl/∼peterl/BN

Course Aims

- Develop complete understanding of basic probability theory (theory)
- Knowledge and understanding of differences and similarities between various probabilistic graphical models (theory)
- Know how to build Bayesian networks from expert knowledge (theory and practice)
- Being familiar with basic inference algorithms (theory and practice)
- Understand the basic issues of learning Bayesian networks from data (theory and practice)
- Be familiar with typical applications (practice)
- Critical appraisal of a specialised topic (theory, possibly practice)

Literature

- Compulsory:
  - K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence*, Chapman & Hall, Boca Raton, 2004 or 2010
- Background:
  - R.G. Cowell, A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Springer, New York, 1999
  - F.V. Jensen and T. Nielsen, *Bayesian Networks and Decision Graphs*, Springer, New York, 2007
  - D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, Cambridge, MA, 2009
- *Various research papers* on the mentioned topics

Uncertainty in Daily Life

- Empirical evidence:

  “If symptoms of fever, shortness of breath (dyspnoea), and coughing are present, and the patient has recently visited China, then the patient has *probably* SARS”

- Subjective belief:

  “The Rutte government will resign soon (and after the elections will be *likely* replaced by a VVD, D66, GL, and PvdA government)”

- Temporal dimension:

  “There is more than *60% chance* that the Dutch economy will fully recover in the next two years”

Uncertainty Representation

- Methods for dealing with uncertainty are not new:
  - 17th century: Fermat, Pascal, Huygens, Leibniz, Bernoulli
  - 18th century: Laplace, De Moivre, Bayes
  - 19th century: Gauss, Boole
- Most important research question in early AI (1970–1987):
  - How to incorporate uncertainty reasoning in logical deduction?
- Again an important research question in modern AI (e.g. Markov logic)

Early AI Methods of Uncertainty

- Rule-based uncertainty representation:

  $(\textit{fever} \;\wedge\; \textit{dyspnoea}) \;\Rightarrow \text{SARS}_{\mathrm{CF}=0.4}$

- Uncertainty calculus (certainty-factor (CF) model, subjective Bayesian method):
  - $\mathrm{CF}(\textit{fever}, B) = 0.6$; $\mathrm{CF}(\textit{dyspnoea}, B) = 1$ (B is background knowledge)
  - Combination functions:

    $$\begin{aligned}\mathrm{CF}(&\text{SARS}, \{\textit{fever}, \textit{dyspnoea}\} \cup B) \\ &= 0.4 \cdot \max\{0, \min\{\mathrm{CF}(\textit{fever}, B), \mathrm{CF}(\textit{dyspnoea}, B)\}\} \\ &= 0.4 \cdot \max\{0, \min\{0.6, 1\}\} = 0.24\end{aligned}$$

However · · ·

$$(\textit{fever} \wedge \textit{dyspnoea}) \Rightarrow \text{SARS}_{\mathrm{CF}=0.4}$$

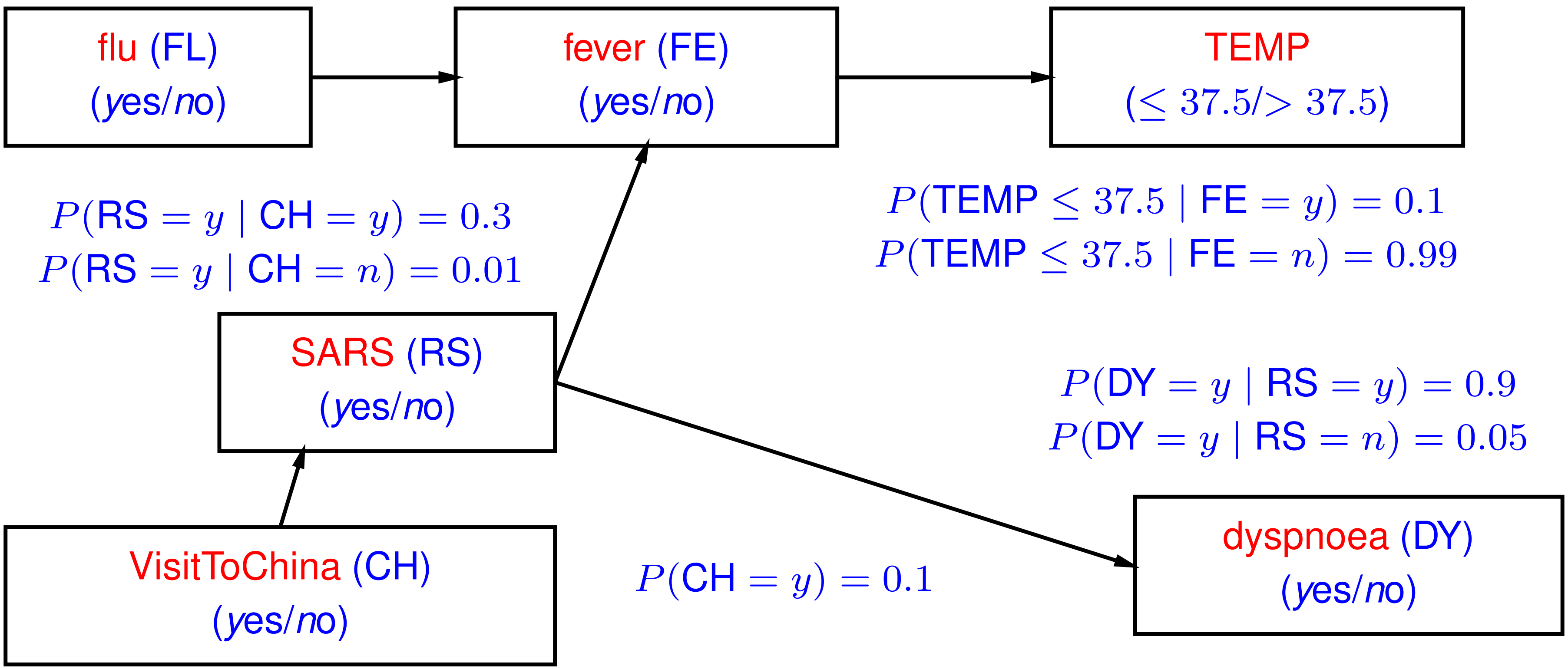
- How likely is the occurrence of *fever* or *dyspnoea* given that the patient has SARS?
- How likely is the occurrence of *fever* or *dyspnoea* in the absence of SARS?
- How likely is the presence of SARS when just *fever* is present?
- How likely is no SARS when just *fever* is present?

Bayesian Networks

$P(\text{CH}, \text{FL}, \text{RS}, \text{DY}, \text{FE}, \text{TEMP})$

$P(\text{FE} = y \mid \text{FL} = y, \text{RS} = y) = 0.95$
$P(\text{FE} = y \mid \text{FL} = n, \text{RS} = y) = 0.80$
$P(\text{FE} = y \mid \text{FL} = y, \text{RS} = n) = 0.88$
$P(\text{FE} = y \mid \text{FL} = n, \text{RS} = n) = 0.001$

$P(\text{FL} = y) = 0.1$



$P(\text{TEMP} \leq 37.5 \mid \text{FE} = y) = 0.1$
$P(\text{TEMP} \leq 37.5 \mid \text{FE} = n) = 0.99$

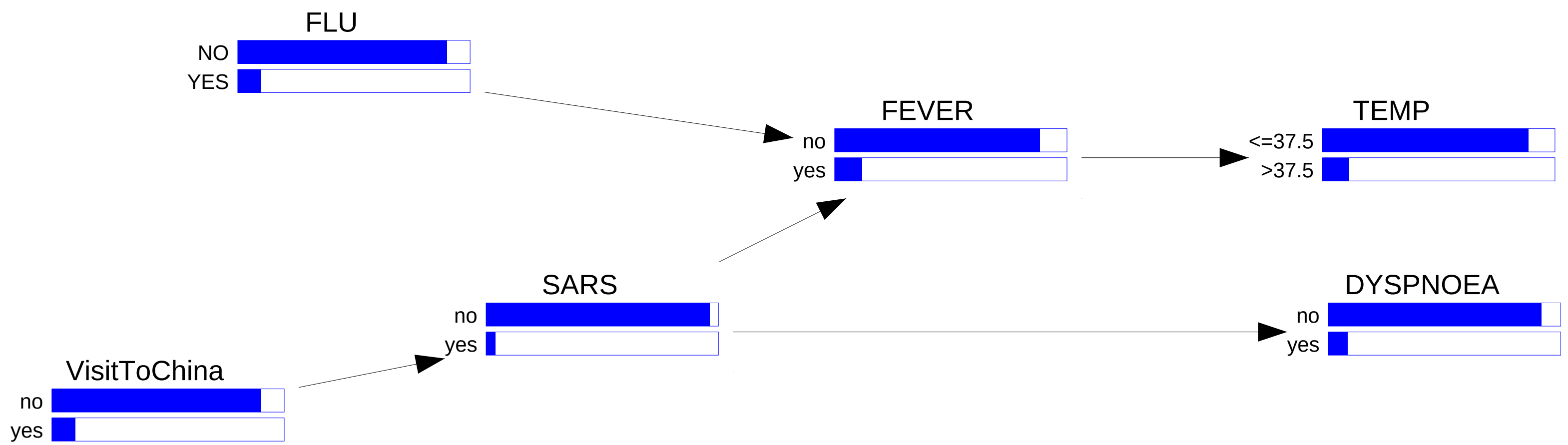
$P(\text{RS} = y \mid \text{CH} = y) = 0.3$
$P(\text{RS} = y \mid \text{CH} = n) = 0.01$

$P(\text{DY} = y \mid \text{RS} = y) = 0.9$
$P(\text{DY} = y \mid \text{RS} = n) = 0.05$

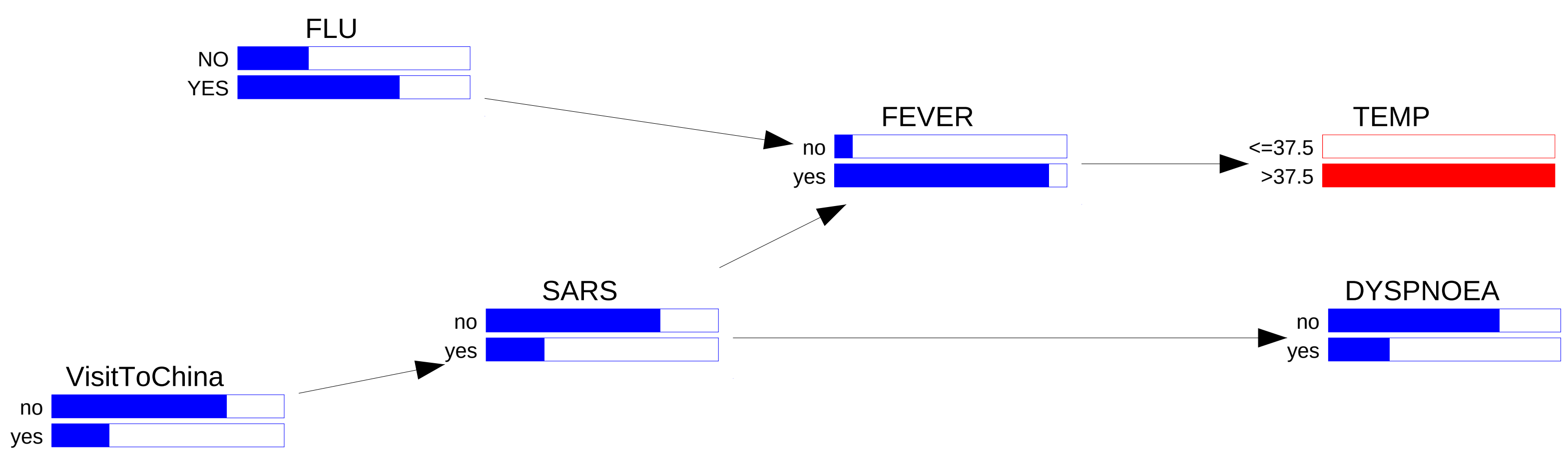
$P(\text{CH} = y) = 0.1$

Reasoning: Evidence Propagation

- Nothing known:

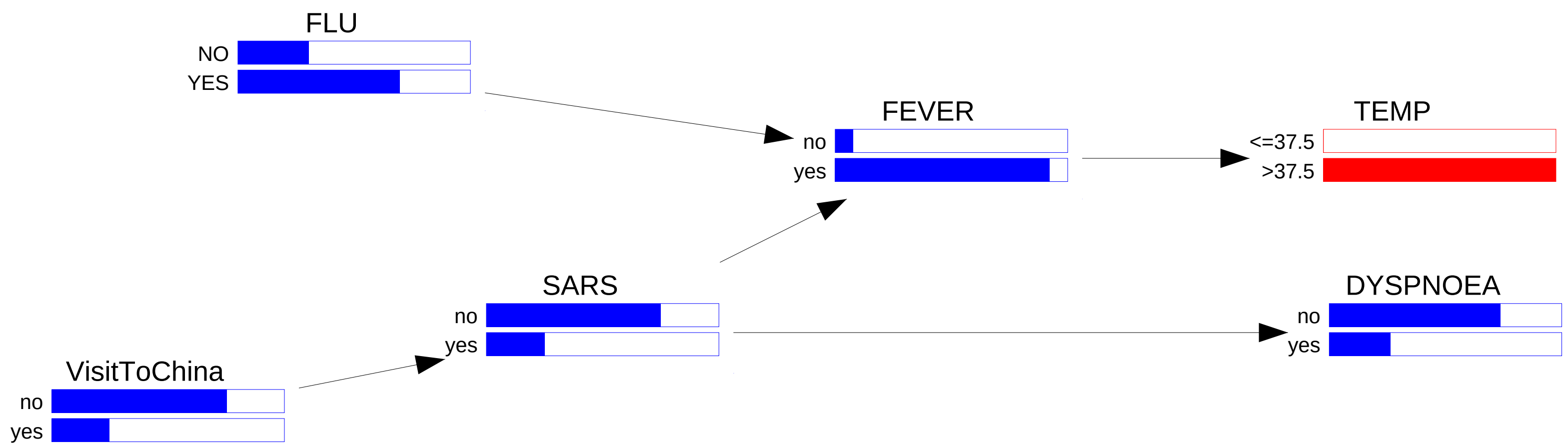


- Temperature >37.5 °C:

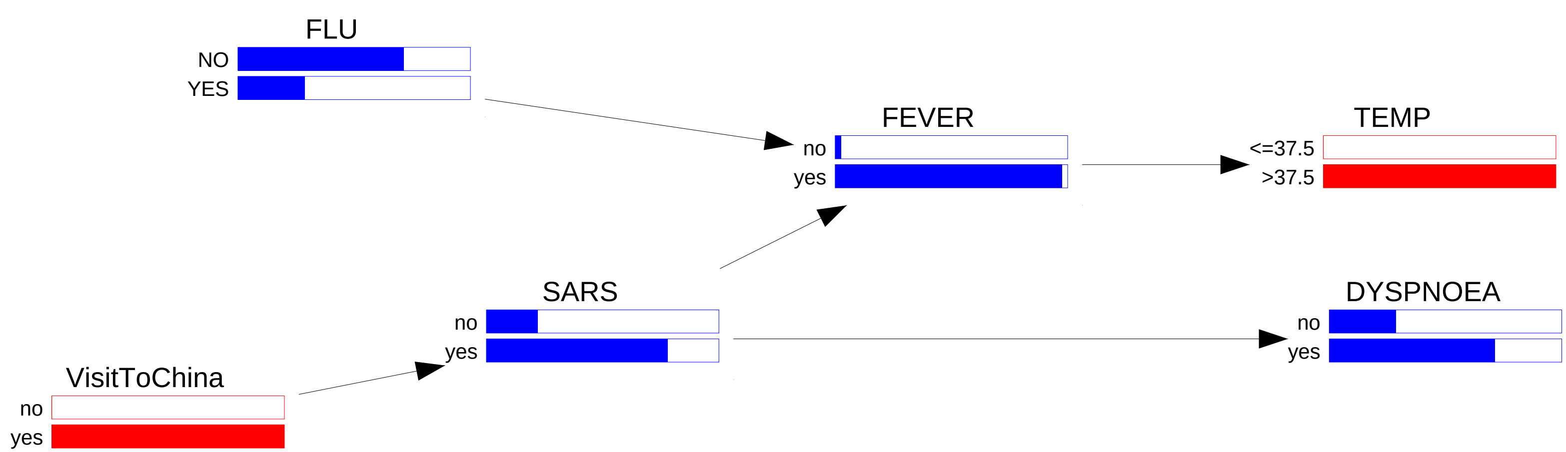


Reasoning: Evidence Propagation

- Temperature >37.5 °C:



- I just returned from China:



Independence Representation in Graphs

The set of variables X is conditionally independent of the set Z *given* the set Y, notation $X \perp\!\!\!\perp Z \mid Y$, iff

$$\boxed{P(X \mid Y, Z) = P(X \mid Y)}$$

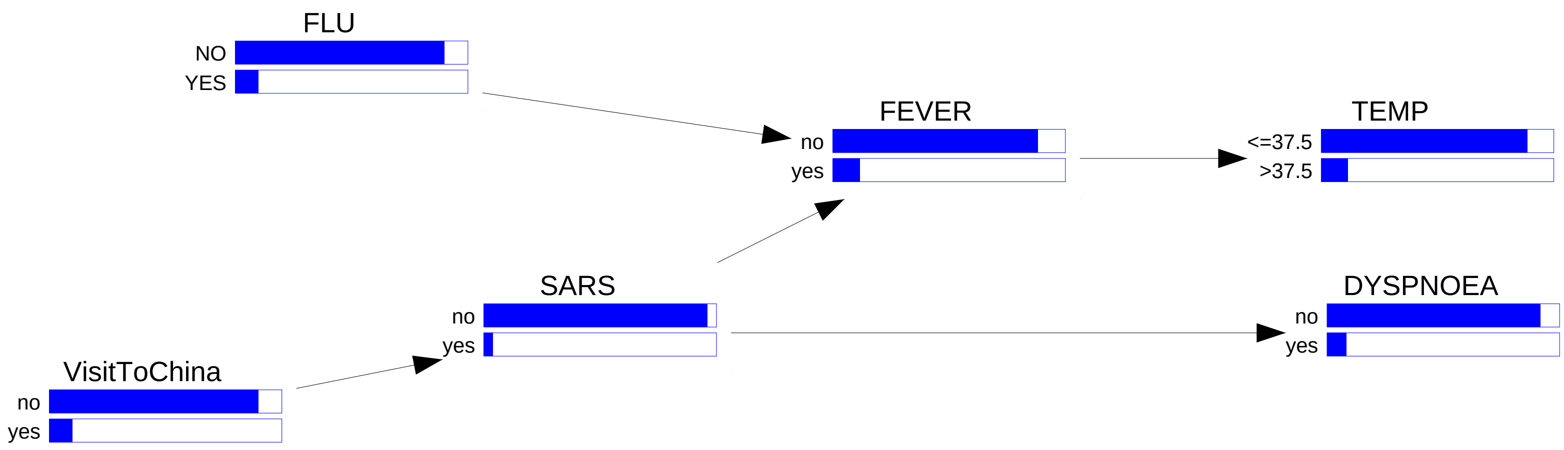
Meaning:

> "If we know Y then Z does not have any (extra) effect on our knowledge concerning X (and thus can be omitted)"

Example

If we know that John has fever, then also knowing that he has a high body temperature has no effect on our knowledge about flu

Find the Independences



Examples:

- FLU $\perp\!\!\!\perp$ VisitToChina $\mid \emptyset$
- FLU $\perp\!\!\!\perp$ SARS $\mid \emptyset$
- FLU $\not\perp\!\!\!\perp$ SARS $\mid$ FEVER, also FLU $\not\perp\!\!\!\perp$ SARS $\mid$ TEMP
- SARS $\perp\!\!\!\perp$ TEMP $\mid$ FEVER
- VisitToChina $\perp\!\!\!\perp$ DYSPNOEA $\mid$ SARS

Probabilistic Reasoning

- Interested in conditional probability distributions:

  $$P(X_W \mid \mathcal{E}) = P^{\mathcal{E}}(X_W)$$

  with W set of vertices, for (possibly empty) evidence $\mathcal{E}$ (instantiated variables)

  Examples

  $$P(\text{FLU} = \textit{yes} \mid \text{TEMP} < 37.5)$$

  $$P(\text{FLU} = \textit{yes}, \text{VisitToAsia} = \textit{yes} \mid \text{TEMP} < 37.5)$$

- Tendency to focus on conditional probability distributions of single variables

Probabilistic Reasoning (cont)

- Joint probability distribution $P(X)$:
  $P(X) = P(X_1, X_2, \ldots, X_n)$
  - marginalisation:

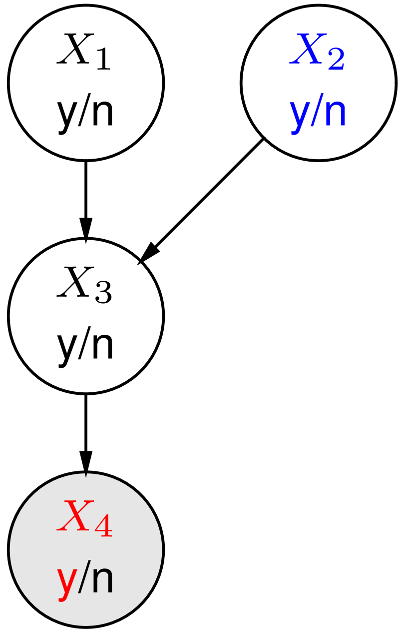
$$P(Y) = \sum_{X \setminus Y} P(X) = \sum_{X \setminus Y} \prod_{v \in V} P(X_v \mid X_{\pi(v)})$$

  - conditional probabilities and Bayes' rule:

$$P(Y, Z \mid X) = \frac{P(X \mid Y, Z) P(Y, Z)}{P(X)}$$

- Many efficient Bayesian reasoning algorithms exist

Naive Probabilistic Reasoning: Evidence

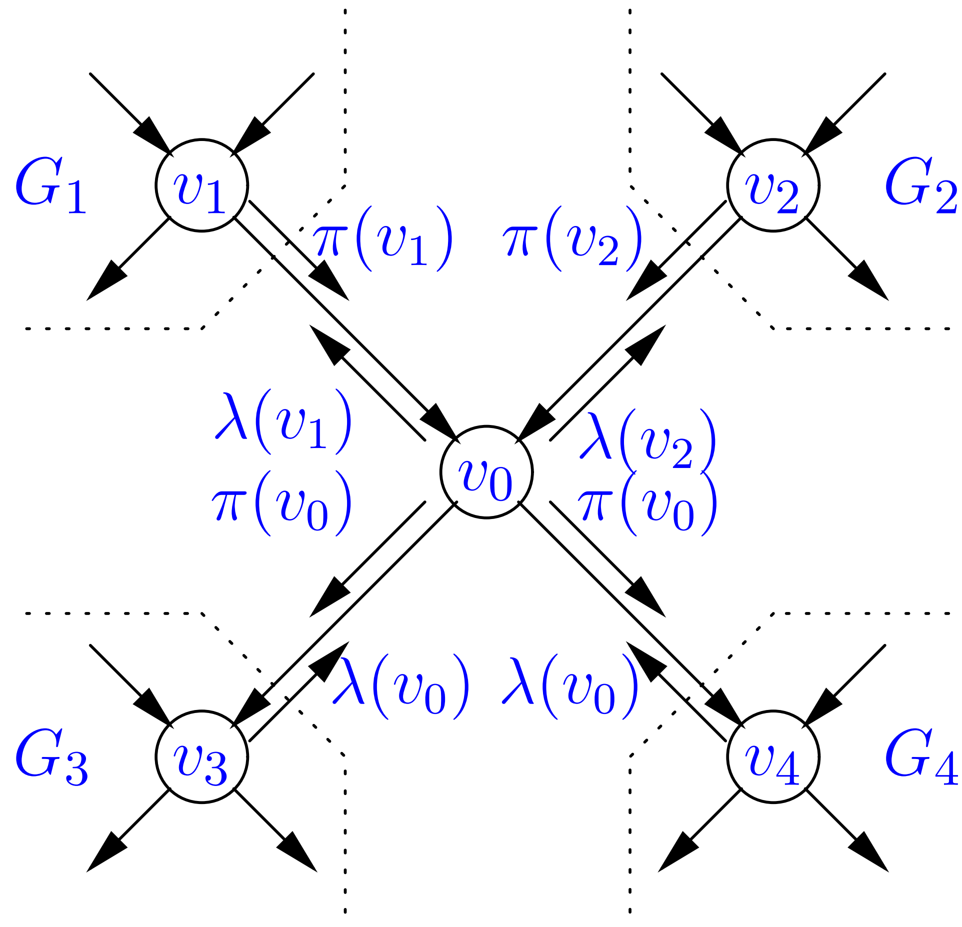


$P(x_4 \mid x_3) = 0.4$
$P(x_4 \mid \neg x_3) = 0.1$
$P(x_3 \mid x_1, x_2) = 0.3$
$P(x_3 \mid \neg x_1, x_2) = 0.5$
$P(x_3 \mid x_1, \neg x_2) = 0.7$
$P(x_3 \mid \neg x_1, \neg x_2) = 0.9$
$P(x_1) = 0.6$
$P(x_2) = 0.2$

$$P^{\mathcal{E}}(x_2) = P(x_2 \mid x_4) = \frac{P(x_4 \mid x_2)P(x_2)}{P(x_4)}$$ (Bayes' rule)

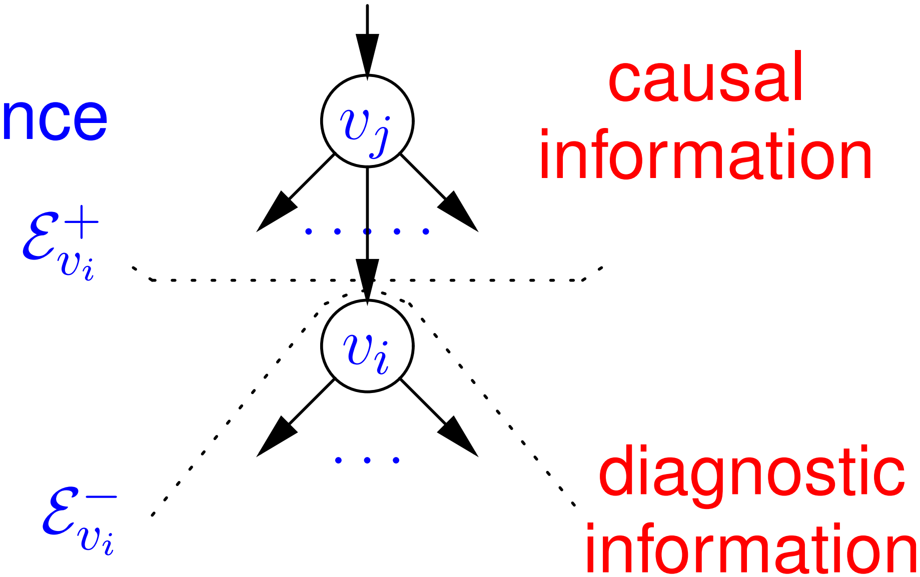
$$= \frac{\sum_{X_3} P(x_4|X_3) \sum_{X_1} P(X_3|X_1, x_2)P(X_1)P(x_2)}{\sum_{X_3} P(x_4 \mid X_3) \sum_{X_1,X_2} P(X_3 \mid X_1, X_2)P(X_1)P(X_2)} \approx 0.14$$

Judea Pearl's Algorithm



- Object-oriented approach: vertices are objects, which have local information and carry out local computations
- Updating of probability distribution by message passing: arcs are communication channels

Data Fusion Lemma



Data fusion:

$P^{\mathcal{E}}(X_{v_i}) = P(X_{v_i} \mid \mathcal{E})$

$= \alpha \cdot$ causal info for $X_{v_i} \cdot$ diagnostic info for X_{v_i}

$= \alpha \cdot \pi(v_i) \cdot \lambda(v_i)$

where:

- $\mathcal{E} = \mathcal{E}^{+}_{v_i} \cup \mathcal{E}^{-}_{v_i}$: evidence
- α: normalisation constant

Problem Solving

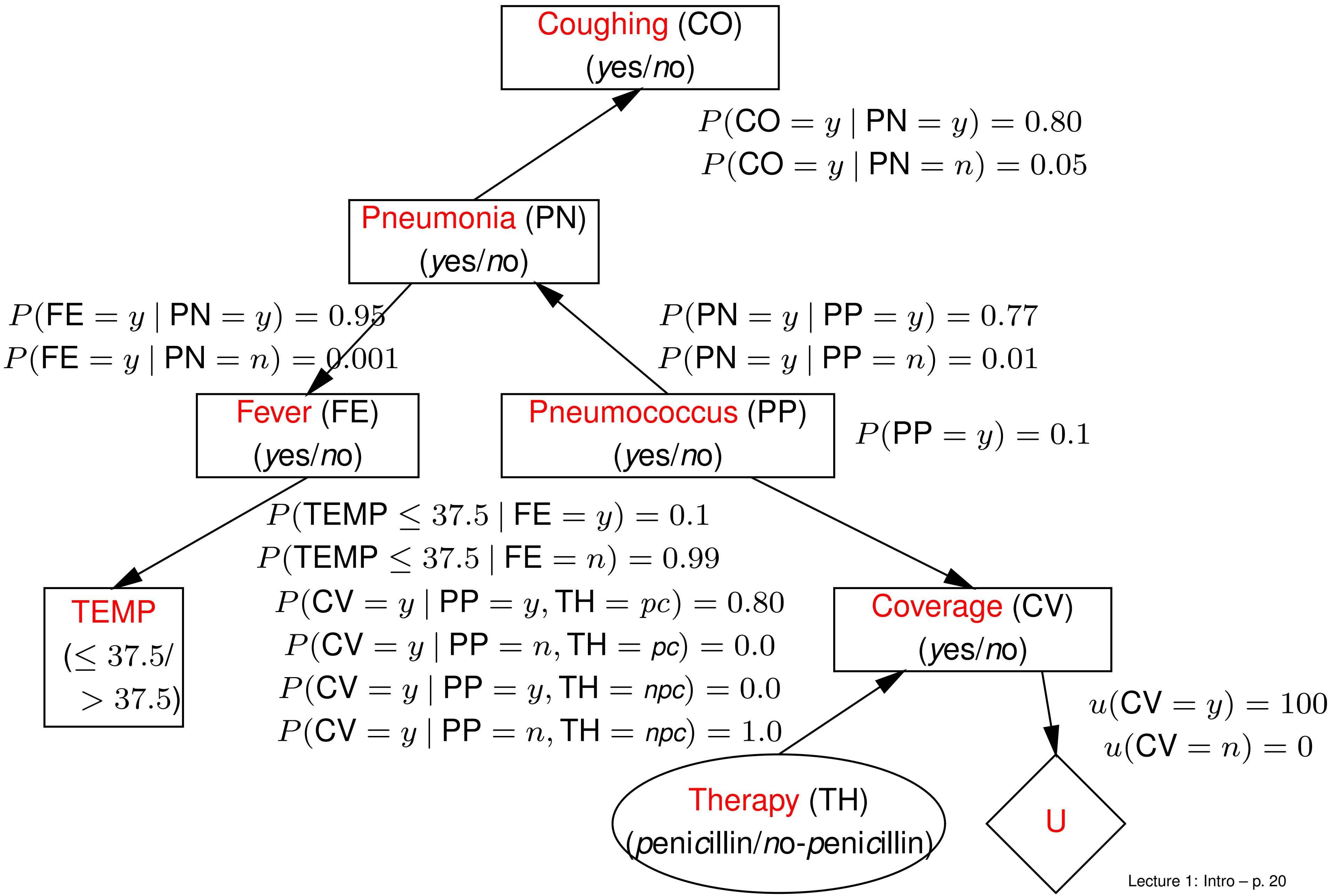
Bayesian networks are declarative, i.e.:

- mathematical basis
- problem to be solved determined by (1) entered evidence $\mathcal{E}$ (may include decisions); (2) given hypothesis H: $P(H \mid \mathcal{E})$ (cf. KB $\wedge H \vDash \mathcal{E}$)

Examples:

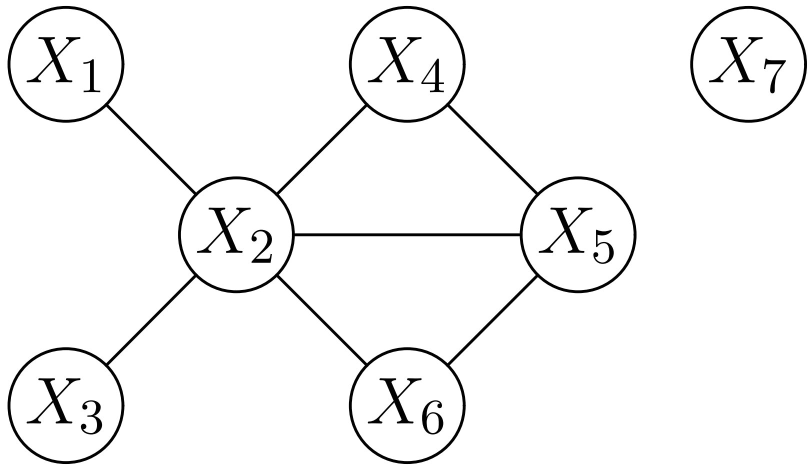
- Description of populations
- Maximum a Posteriori (MAP) Assignment for classification and diagnosis: $D = \arg\max_H P(H \mid \mathcal{E})$
- Temporal reasoning, prediction, *what-if* scenarios
- Decision-making based on decision theory MEU$(D \mid \mathcal{E}) = \max_{d \in D} \sum_x u(x) P(x \mid d, \mathcal{E})$

Decision Networks



Markov Networks

- Structure of a joint probability distribution P can also be described by undirected graphs (instead of directed graphs as in Bayesian networks)

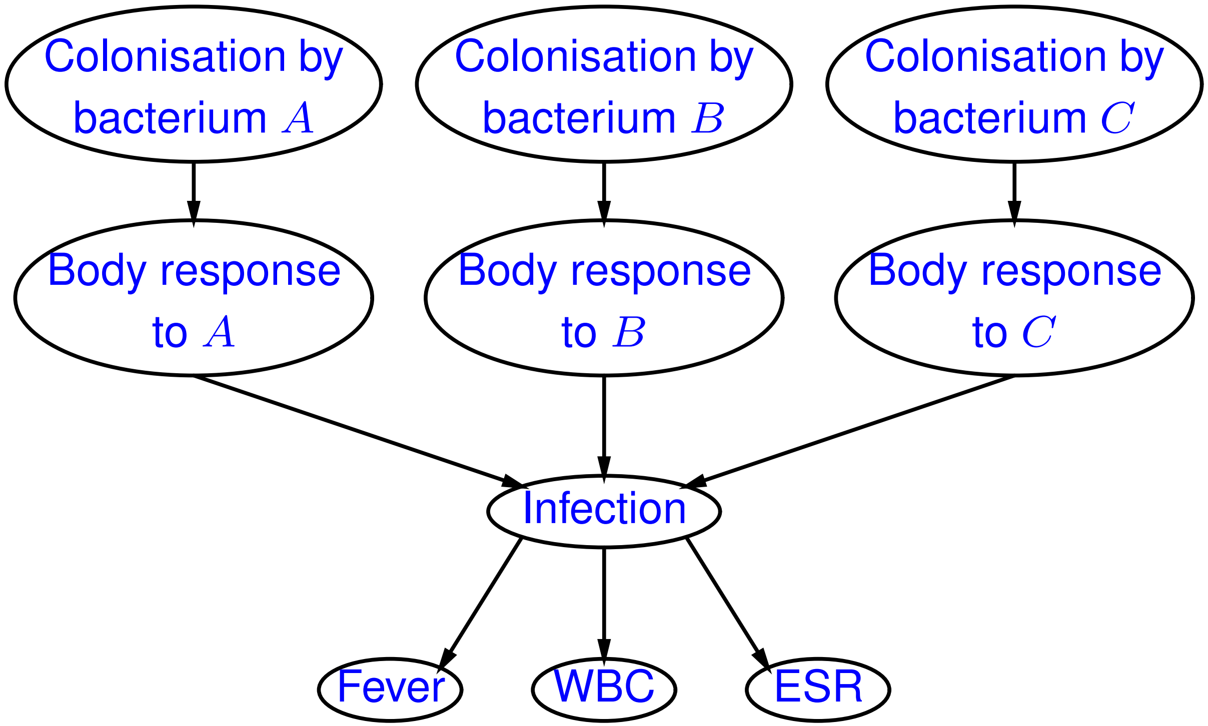


- Together with $P(V) = P(X_1, X_2, X_3, X_4, X_5, X_6, X_7)$: Markov network
- Marginalisation (example):

$$P(\neg x_2) = \sum_{X_1, X_3, X_4, X_5, X_6, X_7} P(X_1, \neg x_2, X_3, X_4, X_5, X_6, X_7)$$

Manual Construction

Qualitative modelling:

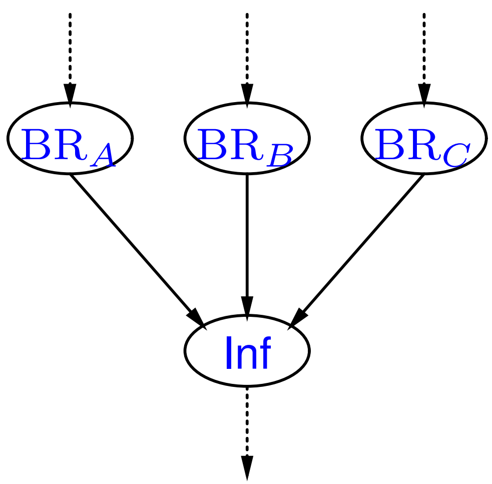


People become colonised by bacteria when entering a hospital, which may give rise to infection

Bayesian-network Modelling

Qualitative
causal modelling

Cause $\rightarrow$ Effect

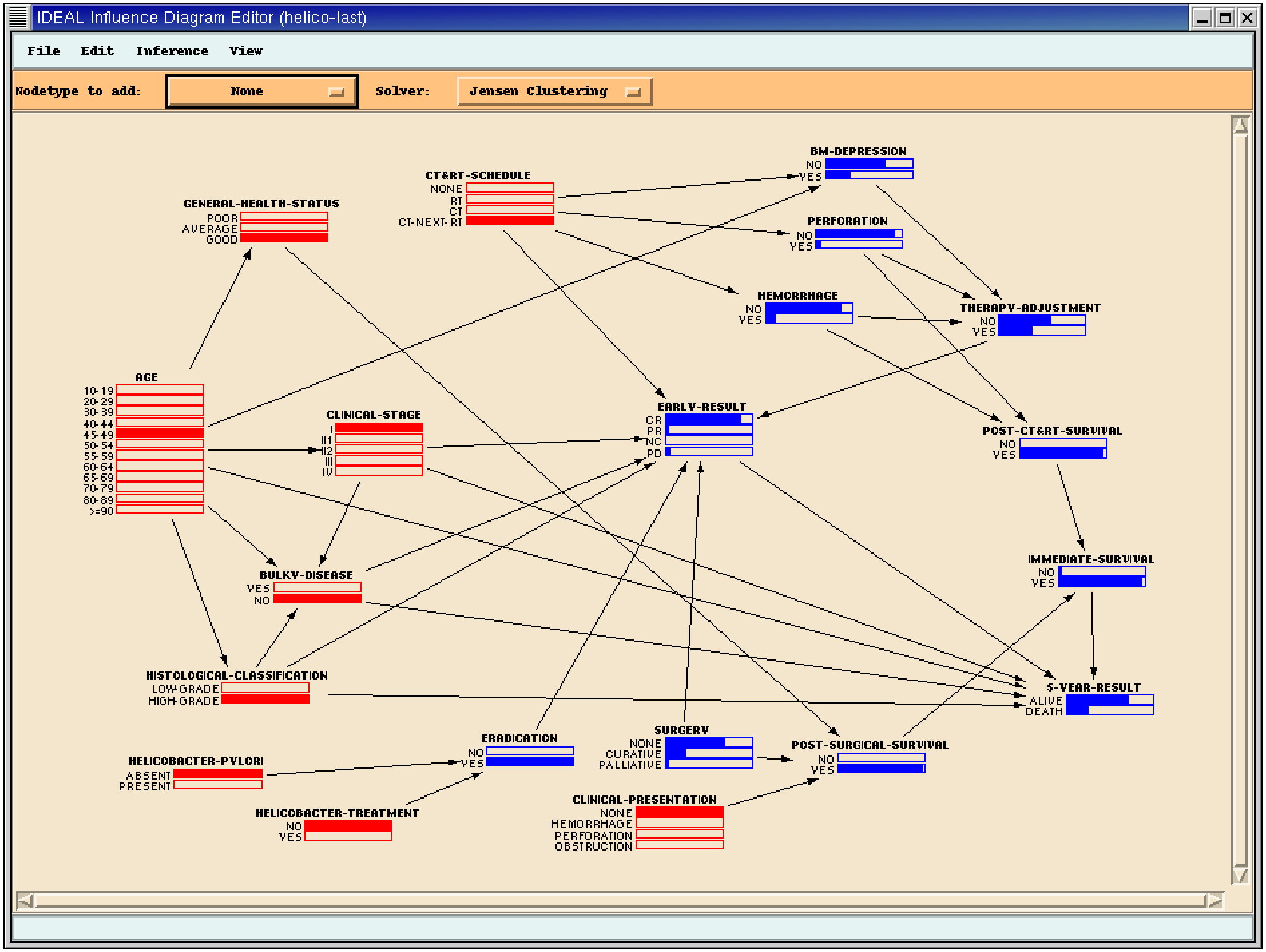


Quantitative
interaction modelling

$$P(\mathrm{Inf} \mid \mathrm{BR}_A, \mathrm{BR}_B, \mathrm{BR}_C)$$

	BR_A							
	t				f			
	BR_B				BR_B			
	t		f		t		f	
	BR_C		BR_C		BR_C		BR_C	
Inf	t	f	t	f	t	f	t	f
t	0.8	0.6	0.5	0.3	0.4	0.2	0.3	0.1
f	0.2	0.4	0.5	0.7	0.6	0.8	0.7	0.9

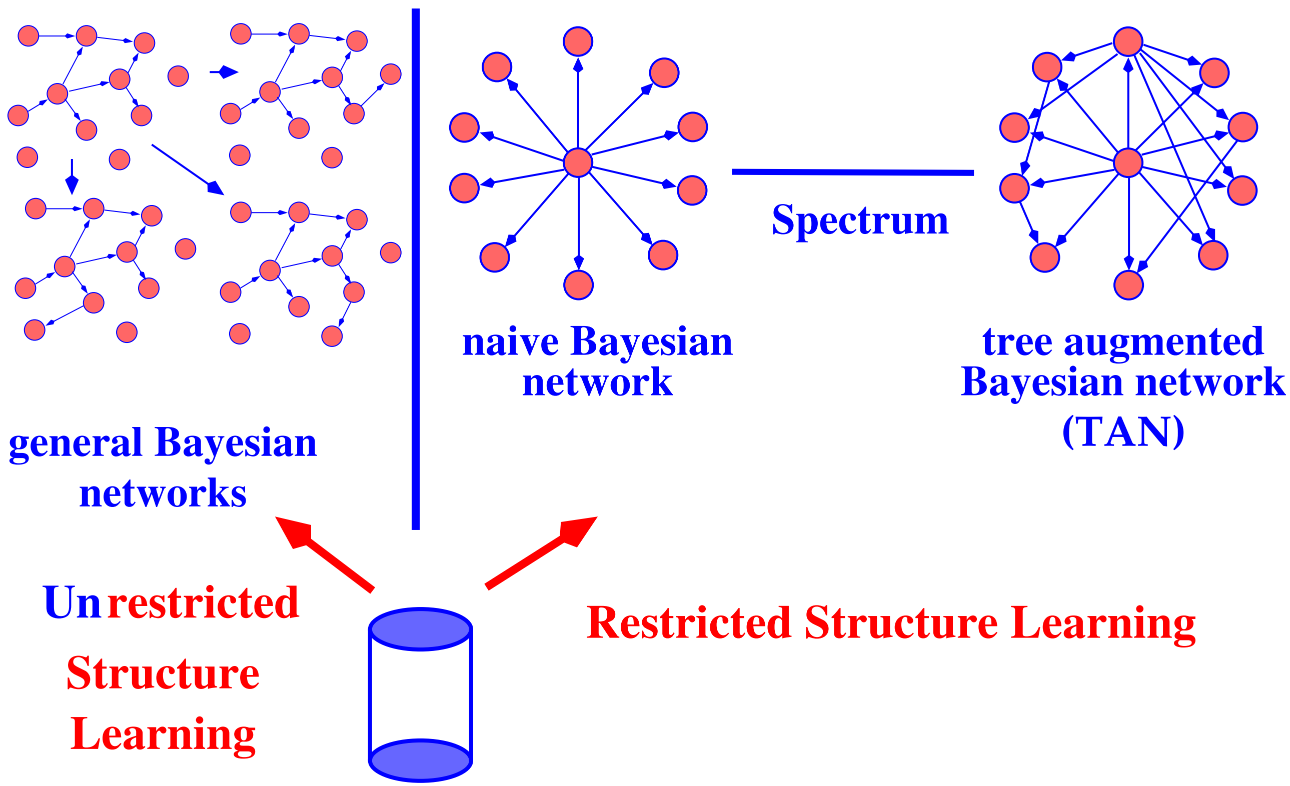
Example BN: non-Hodgkin Lymphoma



Bayesian Network Learning

Bayesian network $\mathcal{B} = (G, P)$, with

- digraph $G = (V(G), A(G))$, and
- probability distribution P



Learning Bayesian Networks

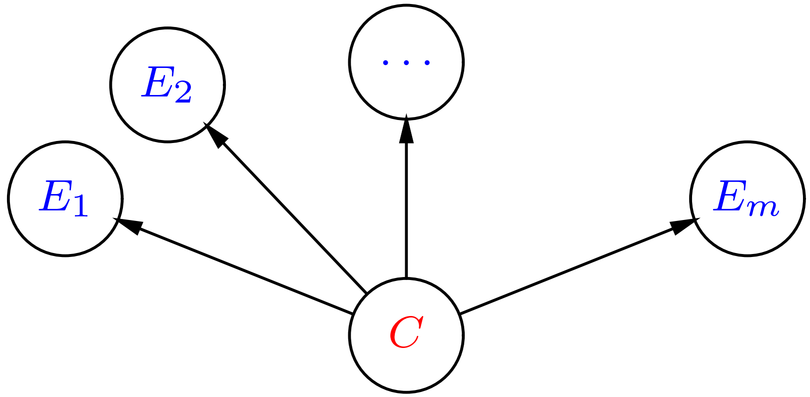
Problems:

- for many BNs too many probabilities have to be assessed
- complex BNs do not necessarily yield better classifiers
- complex BNs may yield better estimates of a probability distribution

Solution:

- use simple probabilistic models for classification:
  - *naive* (independent) form BN
  - *T*ree-*A*ugmented Bayesian *N*etwork (TAN)
  - *F*orest-*A*ugmented Bayesian *N*etwork (FAN)
- use background knowledge and clever heuristics

Naive (independent) form BN



- C is a class variable
- The evidence variables E_i in the evidence $\mathcal{E} \subseteq \{E_1, \dots, E_m\}$ are conditionally independent given the class variable C

This yields: $P(C \mid \mathcal{E}) = \frac{P(\mathcal{E}|C)P(C)}{P(\mathcal{E})} = \frac{\prod_{E \in \mathcal{E}} P(E|C)P(C)}{\sum_C \prod_{E \in \mathcal{E}} P(E|C)P(C)}$

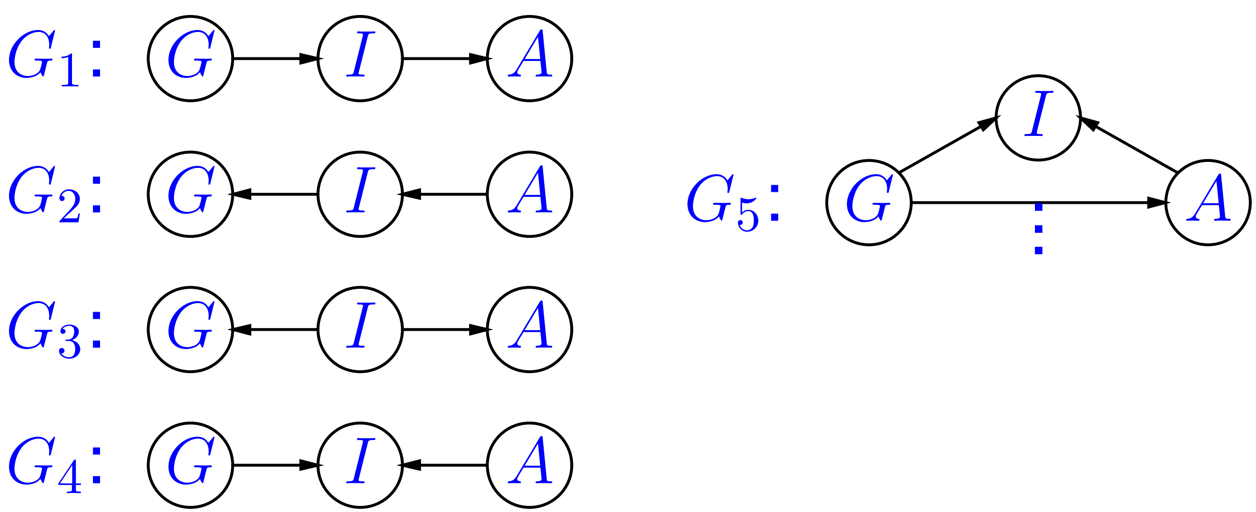
Classifier: $c_{\max} = \text{arg}\max_C P(C \mid \mathcal{E})$

Learning Structure from Data

Given the following dataset D:

Student	Gender	IQ	High Mark for Maths
1	male	low	no
2	female	average	yes
3	male	high	yes
4	female	high	yes

and the following Bayesian networks:

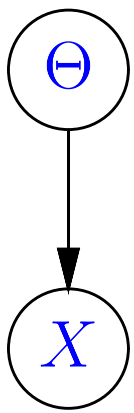


Which one is the best?

Being Bayesian about Bayesian Networks

Bayesian statistics: inherent uncertainty in parameters and exploitation of data to update knowledge:

- Uncertain parameters:

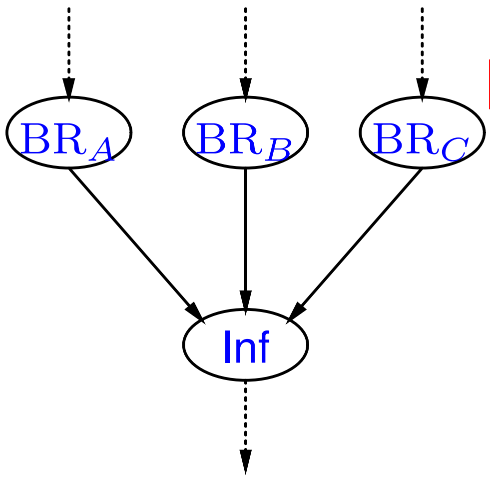


Probability distribution $P(X \mid \Theta)$, with Θ uncertain parameters with probability density $p(\Theta)$

- Assume the Bayesian network structure G comes from a probability distribution, based on data D:

$$P(G \mid D)$$

Research Issues



Modelling:

- To determine the structure of a network
- Generalisation of networks using logics (e.g. Markov logic networks)

Learning:

- Structure learning: determine the 'best' graph topology
- Parameter learning: determine the 'best' probability distribution (discrete or continuous)

Inference: increase speed, reduce memory requirements

$\Rightarrow$ you can contribute too $\cdots$