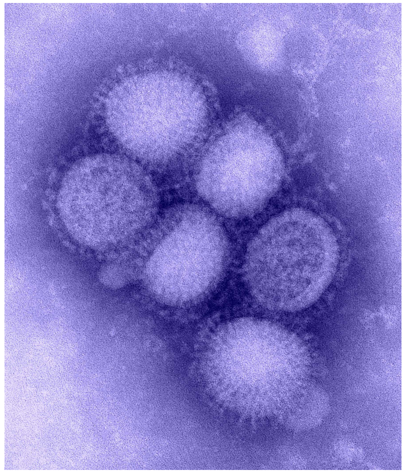
# Reasoning with Uncertainty

Topics:

- Why is uncertainty important?
- How do we represent and reason with uncertain knowledge?
- Progress in research:
  - 1980s: rule-based representation of uncertainty (MYCIN, Prospector)
  - 1990s to present: graphical models, probabilistic expert systems (Munin, Promedas)
  - latest developments: integration of probability theory and logic

# Why important: biomedical

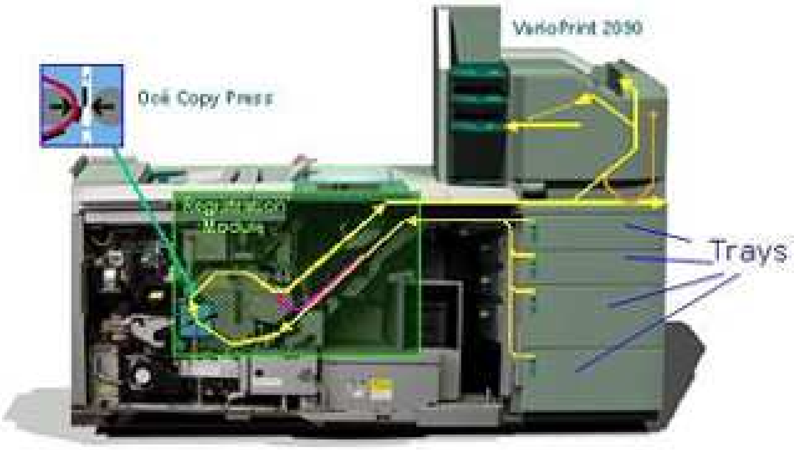
Have you got Mexican Flu?

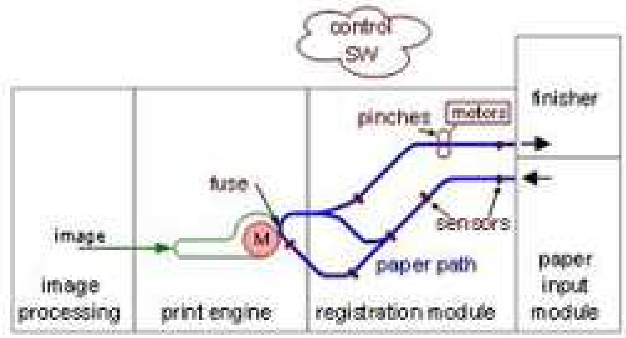



- $M$: mexican flu; $C$: chills; $S$: sore throat
- Probability of mexican flu given sore throat?

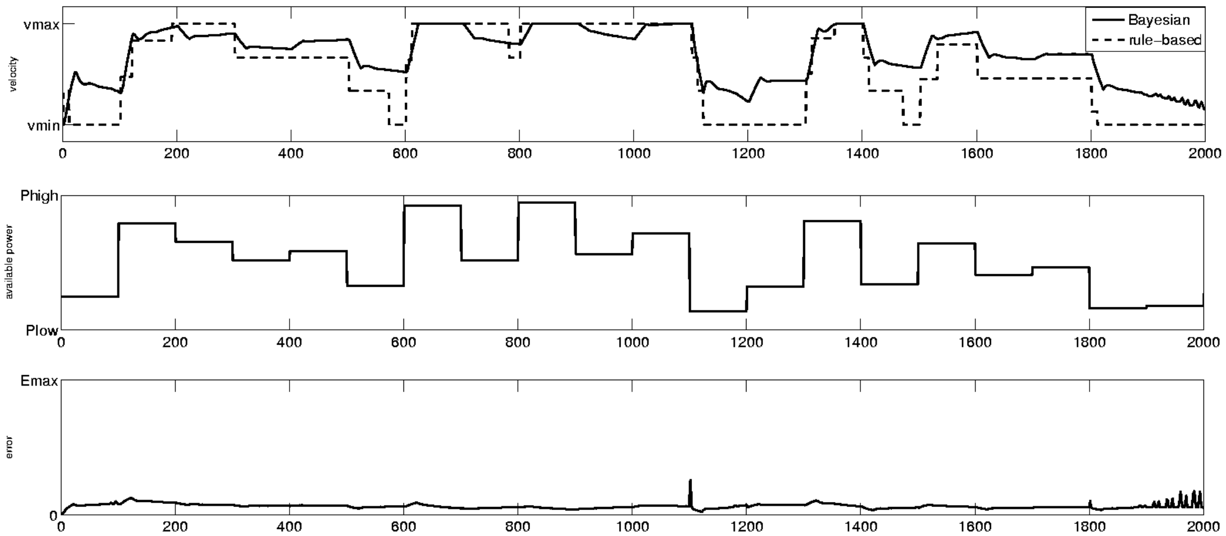
# Why important: embedded systems

Control of behaviour of large production printer



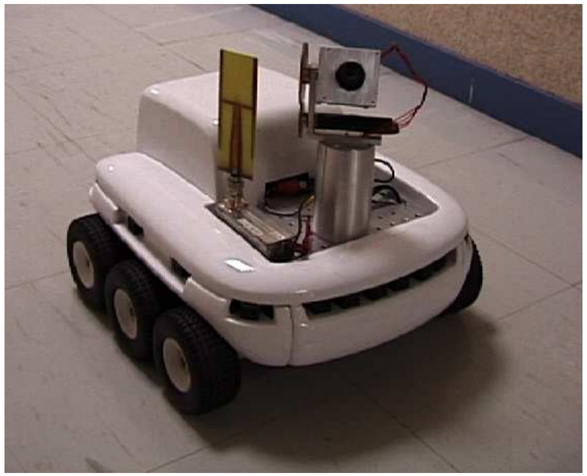



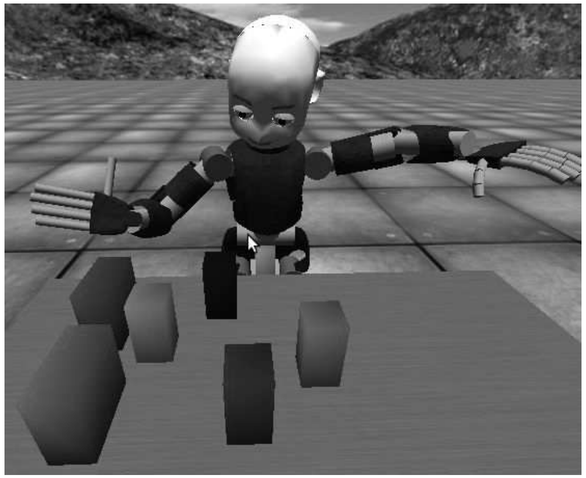

Speed $v$ given available power $P$ and required energy:



# Why important: agents

- Agents (robots) perceive an incomplete image of the world using sensors that are inherently unreliable
- Partially observable worlds
- Noisy computer vision (Lenna: famous image)
- Uncertain, noisy action outcomes





# Representation of uncertainty

- Representation of uncertainty is clearly important!
- How to do it? For example rule based:
  - $e$: evidence
  - $h$: hypothesis

$$e_1 \wedge \cdots \wedge e_n \to h_x$$

If $e_1, e_2, \ldots, e_n$ are true (observed), then conclusion $h$ is true with certainty $x$

  - How to proceed when $e_i, i = 1, \ldots, n$ are uncertain?

$\Rightarrow$ uncertainty propagation/inference/reasoning

# Theory

- We need a basic ‘theory’, e.g.
  - Certainty-factor model (Mycin)
  - Subjective Bayesian method (Prospector) – not discussed
  - Dempster-Shafer theory – not discussed
  - Probability theory
- This theory should tell us how to draw inferences with uncertainty statements
- Many systems (Fuzzy, Plausibility, Probability, Intervals, etc.)
- Much philosophical and technical debate on *semantics* and *truthfulness* of various representation theories.

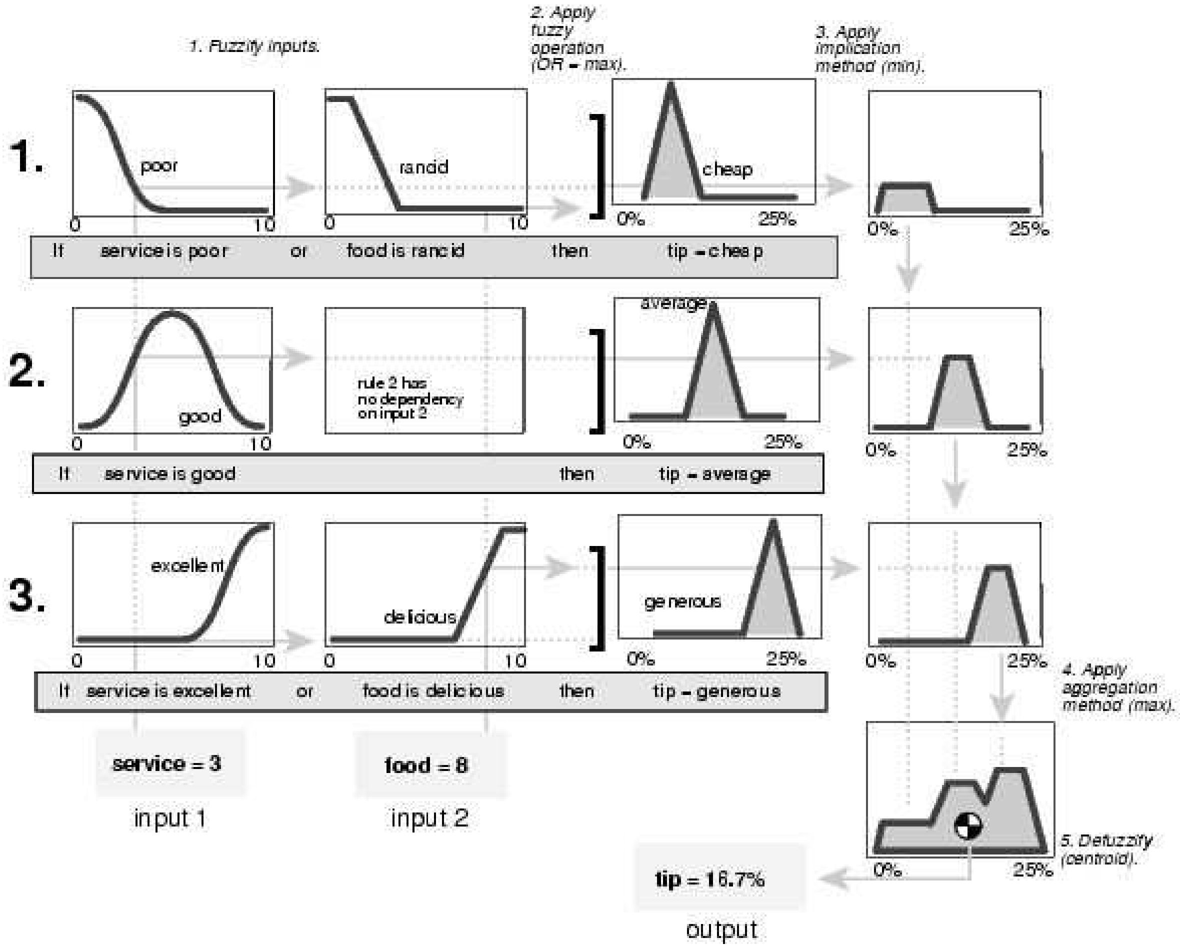
# Rule-based uncertain knowledge

Early, simple approach – certainty-factor calculus:

- *fever* $\wedge$ *myalgia* $\rightarrow$ *flu*$_{\mathrm{CF}=0.8}$
- Example how its works:
  - $\mathrm{CF}(\textit{fever}, e) = 0.6$;
    $\mathrm{CF}(\textit{myalgia}, e) = 1$
    ($e$ is evidence; background knowledge)
  - Combination functions:
    $$\mathrm{CF}(\textit{flu}, e)$$
    $$= 0.8 \cdot \max\{0, \min\{\mathrm{CF}(\textit{fever}, e), \mathrm{CF}(\textit{myalgia}, e)\}\}$$
    $$= 0.8 \cdot \max\{0, \min\{0.6, 1\}\} = 0.48$$

# Fuzzy Logic

- Well-known AI rule-based: Fuzzy Logic
- Fuzzy technology: in cars, washing machines, etc.



# Certainty factor calculus

- Developed by E.H. Shortliffe and B.G. Buchanan for rule-based expert systems
- Applied in MYCIN, the expert system for the diagnosis of infectious disease
- Probability theory was seen as unsatisfactory:
  - Not enough data to obtain sufficient statistics
  - Medical knowledge must be explicitly represented
  - Line of reasoning should be explained by the system

# Inference rules

- Define combination functions $f_\wedge$, $f_\vee$, $f_{\rm prop}$, $f_{\rm co}$, where:
  - $f_\wedge$: combines uncertainty w.r.t. conjunctions of uncertain evidence
  - $f_\vee$: combines uncertainty w.r.t. disjunctions of uncertain evidence
  - $f_{\rm co}$: combines uncertainty for two co-concluding rules:

$$\begin{array}{ll} e_1 \to h_x & \textit{contact\_chicken} \to \textit{flu}_{0.01} \\ e_2 \to h_y & \textit{train\_contact\_humans} \to \textit{flu}_{0.1} \end{array}$$

  - $f_{\rm prop}$: propagation of uncertain evidence $e$ to a hypothesis $h$

# Certainty factor calculus

- Weak relationship to probability theory
- Certainty factors (CFs): subjective estimates of uncertainty with $\mathrm{CF}(x,e) \in [-1,1]$ ($\mathrm{CF}(x,e) = -1$ false, $\mathrm{CF}(x,e) = 0$ unknown, and $\mathrm{CF}(x,e) = 1$ true)
- CF-calculus offers fill-in for combination functions: $f_\wedge$, $f_\vee$, $f_{\mathrm{co}}$, $f_{\mathrm{prop}}$

# Combination functions

- $f_\wedge$
  - rule: $e_1 \wedge e_2 \rightarrow h_{\mathrm{CF}(h,e)}$ with
  - uncertain evidence $\mathrm{CF}(e_1, e')$ and $\mathrm{CF}(e_2, e')$

  then:

$$\mathrm{CF}(e_1 \wedge e_2, e') = \min\{\mathrm{CF}(e_1, e'), \mathrm{CF}(e_2, e')\}$$

- $f_\vee$
  - rule: $e_1 \vee_2 \rightarrow h_{\mathrm{CF}(h,e)}$ with
  - uncertain evidence $\mathrm{CF}(e_1, e')$ and $\mathrm{CF}(e_2, e')$

  then:

$$\mathrm{CF}(e_1 \vee e_2, e') = \max\{\mathrm{CF}(e_1, e'), \mathrm{CF}(e_2, e')\}$$

# Combination functions

- $f_{\text{prop}}$
  - rule $e \rightarrow h_{\text{CF}(h,e)}$
  - uncertain evidence w.r.t. $e$, i.e. $\text{CF}(e, e')$ ($e'$ includes all evidence so far)

then:

$$\text{CF}(h, e') = \text{CF}(h, e) \cdot \max\{0, \text{CF}(e, e')\}$$

# Combination functions

- $f_{\text{co}}$:
  - two rules:

    $e_1 \to h_{\text{CF}(h,e_1)}$
    $e_2 \to h_{\text{CF}(h,e_2)}$
  - uncertain evidence $\text{CF}(e_1, e')$ and $\text{CF}(e_2, e')$
  - Let $\text{CF}(h, e_1') = x$ via rule 1 and $\text{CF}(h, e_2') = y$ via rule 2 (using $f_{\text{prop}}$)
  - Then:

$$\text{CF}(h, e') = \begin{cases} x + y(1 - x) & \text{if } x, y \geq 0 \\ x + y(1 + x) & \text{if } x, y < 0 \\ \frac{x+y}{1-\min\{|x|,|y|\}} & \text{otherwise} \end{cases}$$

# Example

$$\mathcal{R} = \{R_1 : \mathit{flu} \to \mathit{fever}_{\mathrm{CF}(\mathit{fever},\mathit{flu})=0.8},\\ R_2 : \mathit{common\text{-}cold} \to \mathit{fever}_{\mathrm{CF}(\mathit{fever},\mathit{common\text{-}cold})=0.3}\}$$

- Evidence: $\mathrm{CF}(\mathit{flu}, e') = 0.6$ and $\mathrm{CF}(\mathit{common\text{-}cold}, e') = 1$
- What is the certainty factor for *fever*?

# Solution

Application of $f_{\text{prop}}$

Evidence: $\text{CF}(\mathit{flu}, e') = 0.6$ and $\text{CF}(\mathit{common\text{-}cold}, e') = 1$

For rule $R_1$:

$$\begin{aligned} \text{CF}(\mathit{fever}, e'_1) &= \text{CF}(\mathit{fever}, \mathit{flu}) \cdot \max\{0, \text{CF}(\mathit{flu}, e')\} \\ &= 0.8 \cdot 0.6 = 0.48 \end{aligned}$$

for rule $R_2$ this yields $\text{CF}(\mathit{fever}, e'_2) = 0.3$

Application of $f_{\text{co}}$:

$$\begin{aligned} \text{CF}(\mathit{fever}, e') &= \text{CF}(\mathit{fever}, e'_1) + \text{CF}(\mathit{fever}, e'_2)(1 - \text{CF}(\mathit{fever}, e'_1)) \\ &= 0.48 + 0.3(1 - 0.48) = 0.636 \end{aligned}$$

# However . . .

$$\mathit{fever} \wedge \mathit{myalgia} \rightarrow \mathit{flu}_{\mathrm{CF}=0.8}$$

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

# Problems with the CF model

- CF model requires rules to be encoded in the direction in which they are used.
- CF reasoning becomes unsound if strong assumptions fail to hold (consequence of combination functions)
- Assumption of modularity: A rule *if* $e$ *then* $h$ conforms to the following:
  - Detachment: given $e$ we can conclude $h$ no matter how we established $e$
  - Locality: given $e$ we can conclude $h$ no matter what else we know to be true
- Holds for logic but not for probability theory!
- Illogical results are obtained such as the dependence of a diagnosis on the order in which findings are entered

# The inevitability of probability theory

*Probability theory is nothing but common sense reduced to calculation.*
Laplace, 1819

- Basic postulates for any measure of belief (Cox, 1946; Jaynes, 2003):
  1. Representation of degrees of plausibility by real numbers
  2. Qualitative correspondence with common sense
  3. Consistency
- Axioms of probability theory follow as a logical consequence from these postulates
- If you do not reason according to Probability Theory, you can be made to act irrationally (de Finetti)

# Probability space

- A probability space represents our uncertainty regarding an *experiment* (DB query) and consists of:
  - A *sample space* $\Omega$ consisting of a set of outcomes
  - A *probability measure* $P$ which is a real function of the subsets of $\Omega$
- A set of outcomes $A \subseteq \Omega$ is called an event
- $P(A)$ represents how likely it is that an experiment's outcome will be a member of $A$.

# Example

- Suppose our experiment is to examine whether someone has a cold and its related symptom fever.
- The outcomes are defined by

$$\begin{aligned}\Omega &= \{(\text{cold}, \text{fever}), (\text{no cold}, \text{fever}),\\ &\quad (\text{cold}, \text{no fever}), (\text{no cold}, \text{no fever})\}\end{aligned}$$

- and we may define probabilities

$$\begin{aligned}P(\{(\text{cold}, \text{fever}), (\text{cold}, \text{no fever})\}) &= 0.001\\ P(\{(\text{no cold}, \text{fever}), (\text{cold}, \text{fever})\}) &= 0.01\\ &\vdots\end{aligned}$$

- A probability measure $P$ can be completely described by assigning a probability to each event $\omega \in \Omega$

# Axioms of probability theory

- $P$ should obey three axioms:
  1. $P(A) \geq 0$ for all events $A$
  2. $P(\Omega) = 1$
  3. $P(A \cup B) = P(A) + P(B)$ for disjoint events $A$ and $B$
- Some consequences:
  - $P(A) = 1 - P(\Omega \setminus A)$
  - $P(\emptyset) = 0$
  - If $A \subseteq B$ then $P(A) \leq P(B)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$
- Given these axioms and a completely defined probability measure any quantity of interest can be computed!

# Joint density

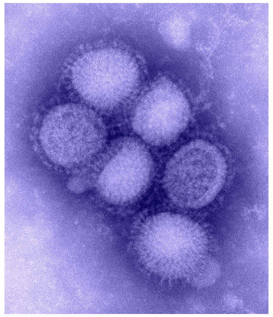
- The joint density for two random variables $X$ and $Y$ is given by

$$p_{XY}(x, y) = P(\{\omega\colon X(\omega) = x, Y(\omega) = y\})$$

- Often written as $P(X = x, Y = y)$, $P(x, y)$, $p(x, y)$, $\ldots$
- Generalizes to multiple random variables
- From now on we work with random variables and (joint) densities instead of events

# Example

## Have you got Mexican Flu?

$$
\begin{aligned}
P(m, c, s) &= 0.009215 \\
P(m, \bar{c}, s) &= 0.000485 \\
P(m, c, \bar{s}) &= 0.000285 \\
P(m, \bar{c}, \bar{s}) &= 1.5 \cdot 10^{-5} \\
P(\bar{m}, c, s) &= 9.9 \cdot 10^{-6} \\
P(\bar{m}, \bar{c}, s) &= 0.0098901 \\
P(\bar{m}, c, \bar{s}) &= 0.0009801 \\
P(\bar{m}, \bar{c}, \bar{s}) &= 0.97912
\end{aligned}
$$

- $M$: mexican flu; $C$: chills; $S$: sore throat

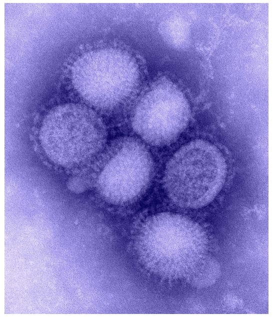
# Marginalization

- Joint probability distribution $P(X) = P(X_1, X_2, \ldots, X_n)$
- $U$ and $V$ are mutually exclusive and collectively exhaustive subsets of $X$.
  - Marginalization:

$$P(u) = \sum_{v \in \mathrm{dom}(v)} P(u, v)$$

# Example

## Have you got Mexican Flu?

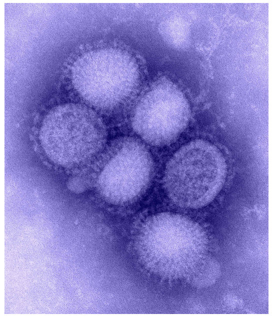



$$
\begin{aligned}
P(m, c, s) &= 0.009215 \\
P(m, \bar{c}, s) &= 0.000485 \\
P(m, c, \bar{s}) &= 0.000285 \\
P(m, \bar{c}, \bar{s}) &= 1.5 \cdot 10^{-5} \\
P(\bar{m}, c, s) &= 9.9 \cdot 10^{-6} \\
P(\bar{m}, \bar{c}, s) &= 0.0098901 \\
P(\bar{m}, c, \bar{s}) &= 0.0009801 \\
P(\bar{m}, \bar{c}, \bar{s}) &= 0.97912
\end{aligned}
$$

- $M$: mexican flu; $C$: chills; $S$: sore throat
- Probability of mexican flu and sore throat?

# Example

## Have you got Mexican Flu?

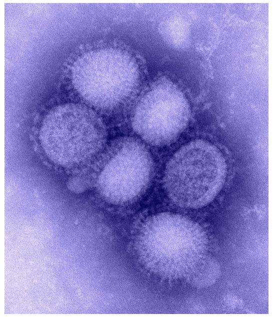
- $M$: mexican flu; $C$: chills; $S$: sore throat
- Probability of mexican flu and sore throat?

$$
\begin{aligned}
P(m,c,s) &= 0.009215 \\
P(m,\bar{c},s) &= 0.000485 \\
P(m,c,\bar{s}) &= 0.000285 \\
P(m,\bar{c},\bar{s}) &= 1.5 \cdot 10^{-5} \\
P(\bar{m},c,s) &= 9.9 \cdot 10^{-6} \\
P(\bar{m},\bar{c},s) &= 0.0098901 \\
P(\bar{m},c,\bar{s}) &= 0.0009801 \\
P(\bar{m},\bar{c},\bar{s}) &= 0.97912
\end{aligned}
$$

$$
\begin{aligned}
P(m,s) &= P(m,c,s) + P(m,\bar{c},s) \\
&= 0.009215 + 0.000485 \\
&= 0.0097
\end{aligned}
$$

# Example

Have you got Mexican Flu?

$$
\begin{aligned}
P(m, c, s) &= 0.009215 \\
P(m, \bar{c}, s) &= 0.000485 \\
P(m, c, \bar{s}) &= 0.000285 \\
P(m, \bar{c}, \bar{s}) &= 1.5 \cdot 10^{-5} \\
P(\bar{m}, c, s) &= 9.9 \cdot 10^{-6} \\
P(\bar{m}, \bar{c}, s) &= 0.0098901 \\
P(\bar{m}, c, \bar{s}) &= 0.0009801 \\
P(\bar{m}, \bar{c}, \bar{s}) &= 0.97912
\end{aligned}
$$

- $M$: mexican flu; $C$: chills; $S$: sore throat
- Probability of mexican flu given sore throat?

# Conditioning

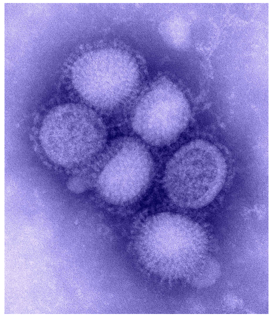
- Conditioning specifies how to revise beliefs based on new information.
- The conditional probability of a $A$ given $B$ is

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

  where $P(B) = \sum_a P(a, B)$.
- Information $B$ rules out possible worlds incompatible with $B$ and induces a new measure over possible worlds in which $B$ holds
- Often, $B$ is available evidence and $A$ is a hypothesis of interest (e.g., disease given symptoms)

# Example

Have you got Mexican Flu?

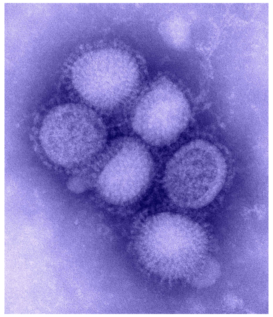



$$
\begin{aligned}
P(m,c,s) &= 0.009215 \\
P(m,\bar{c},s) &= 0.000485 \\
P(m,c,\bar{s}) &= 0.000285 \\
P(m,\bar{c},\bar{s}) &= 1.5 \cdot 10^{-5} \\
P(\bar{m},c,s) &= 9.9 \cdot 10^{-6} \\
P(\bar{m},\bar{c},s) &= 0.0098901 \\
P(\bar{m},c,\bar{s}) &= 0.0009801 \\
P(\bar{m},\bar{c},\bar{s}) &= 0.97912
\end{aligned}
$$

- $M$: mexican flu; $C$: chills; $S$: sore throat
- Probability of mexican flu given sore throat?

# Example

Have you got Mexican Flu?

- $M$: mexican flu; $C$: chills; $S$: sore throat
- Probability of mexican flu given sore throat?

$$
\begin{aligned}
P(m, c, s) &= 0.009215 \\
P(m, \bar{c}, s) &= 0.000485 \\
P(m, c, \bar{s}) &= 0.000285 \\
P(m, \bar{c}, \bar{s}) &= 1.5 \cdot 10^{-5} \\
P(\bar{m}, c, s) &= 9.9 \cdot 10^{-6} \\
P(\bar{m}, \bar{c}, s) &= 0.0098901 \\
P(\bar{m}, c, \bar{s}) &= 0.0009801 \\
P(\bar{m}, \bar{c}, \bar{s}) &= 0.97912
\end{aligned}
$$

$$
\begin{aligned}
P(m \mid s) &= P(m, s)/P(s) \\
&= 0.0097/0.0196 \\
&= 0.495
\end{aligned}
$$

# Product rule

- Conditional probability: $P(A|B) = \frac{P(A,B)}{P(B)}$
- Therefore: $P(A,B) = P(A|B)P(B)$

# Chain rule

- Extension of the product rule:

$$
\begin{aligned}
&P(X_1, X_2, \ldots, X_n) \\
&= P(X_n \mid X_1, X_2, \ldots, X_{n-1}) \times P(X_1, X_2, \ldots, X_{n-1}) \\
&= P(X_n \mid X_1, X_2, \ldots, X_{n-1}) \times \\
&\qquad P(X_{n-1} \mid X_1, X_2, \ldots, X_{n-2}) \times P(X_1, X_2, \ldots, X_{n-2}) \\
&= P(X_n \mid X_1, X_2, \ldots, X_{n-1}) \times P(X_{n-1} \mid X_1, X_2, \ldots, X_{n-2}) \times \\
&\qquad \cdots \times P(X_3 \mid X_1, X_2) \times P(X_2 \mid X_1) \times P(X_1) \\
&= \prod_{i=1}^{n} P(X_i \mid X_1, \cdots, X_{i-1})
\end{aligned}
$$

# Bayes' rule

- The chain rule and commutativity of conjunction ($P(A, B)$ is equivalent to $P(B, A)$) gives us:

$$P(A, B) = P(A \mid B) \times P(B) = P(B \mid A) \times P(A).$$

- If $P(B) \neq 0$, you can divide the right hand sides by $P(B)$:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- This is Bayes' rule.

# Bayes' rule

- Why is Bayes' rule interesting?
- Often you have causal knowledge:

$$P(\text{symptom} \mid \text{disease}), P(\text{disease})$$
$$P(\text{alarm} \mid \text{fire}), P(\text{fire})$$
$$P(\text{image} \mid \text{a tree is in front of a car}), P(\text{a tree is in front of a car})$$

and want to do evidential reasoning:

$$P(\text{disease} \mid \text{symptom})$$
$$P(\text{fire} \mid \text{alarm})$$
$$P(\text{a tree is in front of a car} \mid \text{image})$$

- Reasoning 'against the direction of the arrows' is not possible using e.g. certainty factors.

# Bayes' rule in practice

- A drug test is 99% sensitive (the test returns a positive result for a user 99% of the time)
- A drug test is 99% specific (the test returns a negative result for a non-user 99% of the time)
- Suppose that 0.5% of people are users of the drug
- If an individual tests positive, what is the probability they are a user?

# Bayes' rule in practice

- $d$ = drug user, $p$ = positive test, $P(p \mid d) = 0.99$
- $P(\neg p \mid \neg d) = 0.99$, $p(d) = 0.005$.

$$
\begin{aligned}
P(d \mid p) &= \frac{P(p \mid d)P(d)}{P(p)} \\
&= \frac{P(p \mid d)P(d)}{P(p \mid d)P(d) + P(p \mid \neg d)p(\neg d)} \\
&= \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.01 \cdot 0.995} \\
&= 33.2\%
\end{aligned}
$$

# Independence

- Random variable $X$ is independent of random variable $Y$ if for all $x$ and $y$

$$P(x \mid y) = P(x)$$

- This is written as $X \perp\!\!\!\perp Y$
- Examples:
  - $\text{Flu} \perp\!\!\!\perp \text{Haircolor}$ since $P(\text{Flu} \mid \text{Haircolor}) = P(\text{Flu})$.
  - $\text{Myalgia} \not\!\perp\!\!\!\perp \text{Fever}$ since $P(\text{Myalgia} \mid \text{Fever}) \neq P(\text{Myalgia})$.

# Independence

- Independence is very powerful because it allows us to reason about aspects of a system in isolation.
- However, it does not often occur in complex systems. For example, try and think of two medical symptoms that are independent.
- A generalization of independence is conditional independence, where two aspects of a system become independent once we observe a third aspect.
- Conditional independence does often arise and can lead to significant representational and computational savings.

# Conditional independence

- Random variable $X$ is conditionally independent of random variable $Y$ given random variable $Z$ if

$$P(x \mid y, z) = P(x \mid z)$$

whenever $P(y, z) > 0$. That is, knowledge of $Y$ doesn't affect your belief in the value of $X$, given a value of $Z$.

- This is written as $X \perp\!\!\!\perp Y \mid Z$
- Example:
  - Symptoms are conditionally independent given the disease:

$$\text{Myalgia} \perp\!\!\!\perp \text{Fever} \mid \text{Flu}$$

since $P(\text{Myalgia} \mid \text{Fever}, \text{Flu}) = P(\text{Myalgia} \mid \text{Flu})$

# Conditional independence

- An intuitive test of conditional independence (Paskin):

  *Imagine that you know the value of $Z$ and you are trying to guess the value of $X$. In your pocket is an envelope containing the value of $Y$. Would opening the envelope help you guess $X$? If not, then $X \perp\!\!\!\perp Y \mid Z$.*

# Example

- Assume we have a joint density over the following five variables:
  - Temperature: $\mathrm{temp} \in \{\mathrm{high}, \mathrm{low}\}$
  - Fever: $\mathrm{fe} \in \{\mathrm{y}, \mathrm{n}\}$
  - Myalgia: $\mathrm{my} \in \{\mathrm{y}, \mathrm{n}\}$
  - Flu: $\mathrm{fl} \in \{\mathrm{y}, \mathrm{n}\}$
  - Pneumonia: $\mathrm{pn} \in \{\mathrm{y}, \mathrm{n}\}$
- Probabilistic inference amounts to computing one or more (conditional) densities given (possibly empty) observations.

# Inference problem

$$P(\text{pn} \mid \text{temp}{=}\text{high}) \;=\; \frac{1}{Z}\sum_{\text{fe}}\sum_{\text{my}}\sum_{\text{fl}} P(\text{temp}{=}\text{high}, \text{fe}, \text{my}, \text{fl}, \text{pn})$$

- We don't need to compute $Z$. We just compute

  $$P(\text{pn} \mid \text{temp}{=}\text{high}) \times P(\text{temp}{=}\text{high})$$

  and renormalize.
- We do need to compute the sums, which becomes expensive very fast (nested **for** loops)!

# Representation problem

- In order to specify the joint density $P(\text{temp}, \text{fe}, \text{my}, \text{fl}, \text{pn})$ we need to estimate $31(2^n - 1)$ probabilities
- Probabilities can be estimated by means of knowledge engineering or by parameter learning
- This doesn't solve the problem
  - How does an expert estimate $P(\text{temp}{=}\text{low}, \text{fe}{=}\text{y}, \text{my}{=}\text{n}, \text{fl}{=}\text{y}, \text{pn}{=}\text{y})$?
  - Parameter learning requires huge databases containing multiple instances of each configuration
- Solution: conditional independence!

# Chain rule revisited

The chain rule allows us to write:

$$P(\text{temp}, \text{fe}, \text{my}, \text{fl}, \text{pn})$$

$$= P(\text{temp} \mid \text{fe}, \text{my}, \text{fl}, \text{pn}) P(\text{fe} \mid \text{my}, \text{fl}, \text{pn}) P(\text{my} \mid \text{fl}, \text{pn}) P(\text{fl} \mid \text{pn}) P(\text{pn})$$

This requires $16 + 8 + 4 + 2 + 1 = 31$ probabilities

We now make the following (conditional) independence assumptions:

- $\text{fl} \perp\!\!\!\perp \text{pn}$
- $\text{my} \perp\!\!\!\perp \{\text{temp}, \text{fe}, \text{pn}\} \mid \text{fl}$
- $\text{temp} \perp\!\!\!\perp \{\text{my}, \text{fl}, \text{pn}\} \mid \text{fe}$
- $\text{fe} \perp\!\!\!\perp \{\text{my}\} \mid \{\text{fl}, \text{pn}\}$

# Chain rule revisited

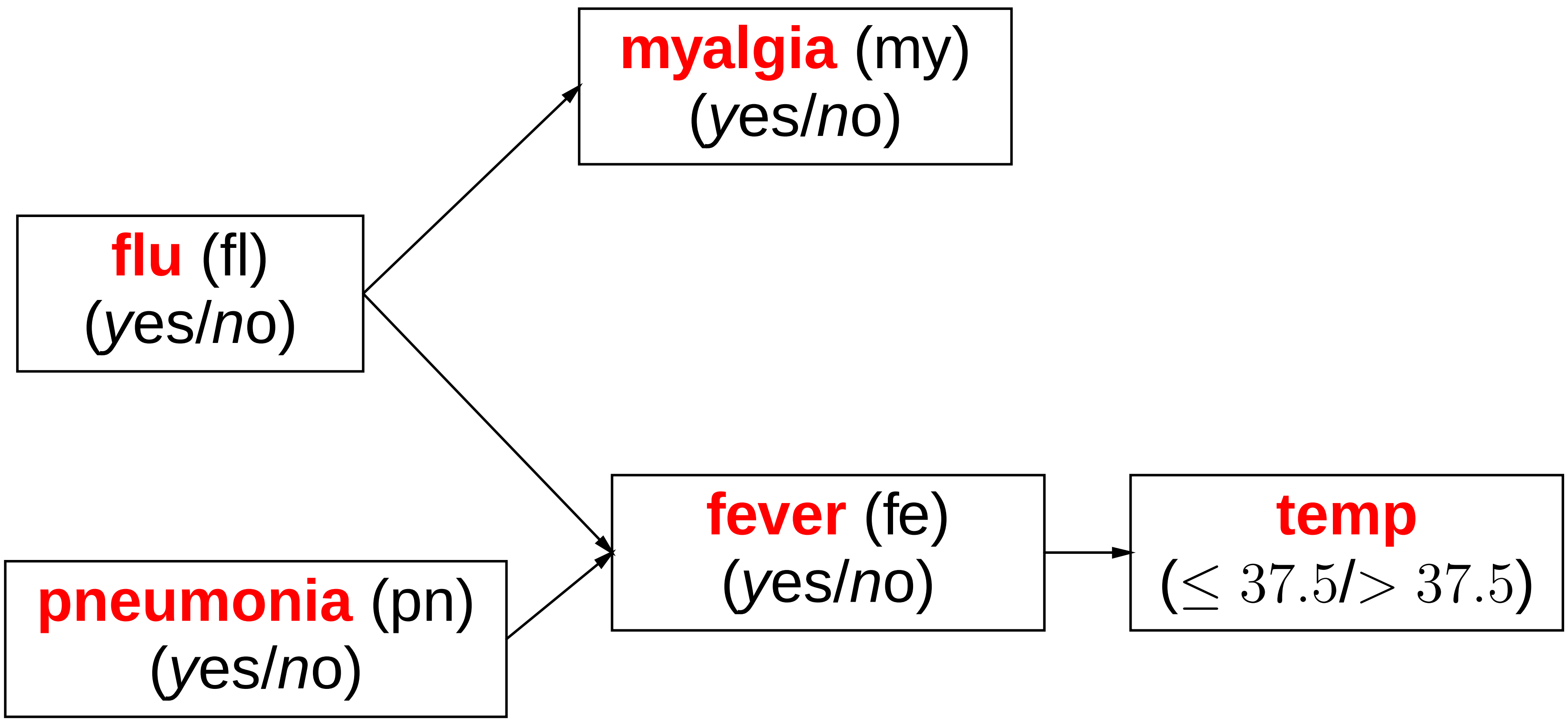
- By definition of conditional independence:

$$
\begin{aligned}
& P(\text{temp}, \text{fe}, \text{my}, \text{fl}, \text{pn}) \\
& = P(\text{temp} \mid \text{fe}) P(\text{fe} \mid \text{fl}, \text{pn}) P(\text{my} \mid \text{fl}) P(\text{fl}) P(\text{pn})
\end{aligned}
$$

- This requires just $2 + 4 + 2 + 1 + 1 = 10$ instead of $31$ probabilities
- Conditional independence assumptions reduce the number of required probabilities and makes the specification of the remaining probabilities easier:
  - $P(\text{my} \mid \text{fl})$: the probability of myalgia given that someone has flu
  - $P(\text{pn})$: the prior probability that a random person suffers from pneumonia

# Bayesian networks

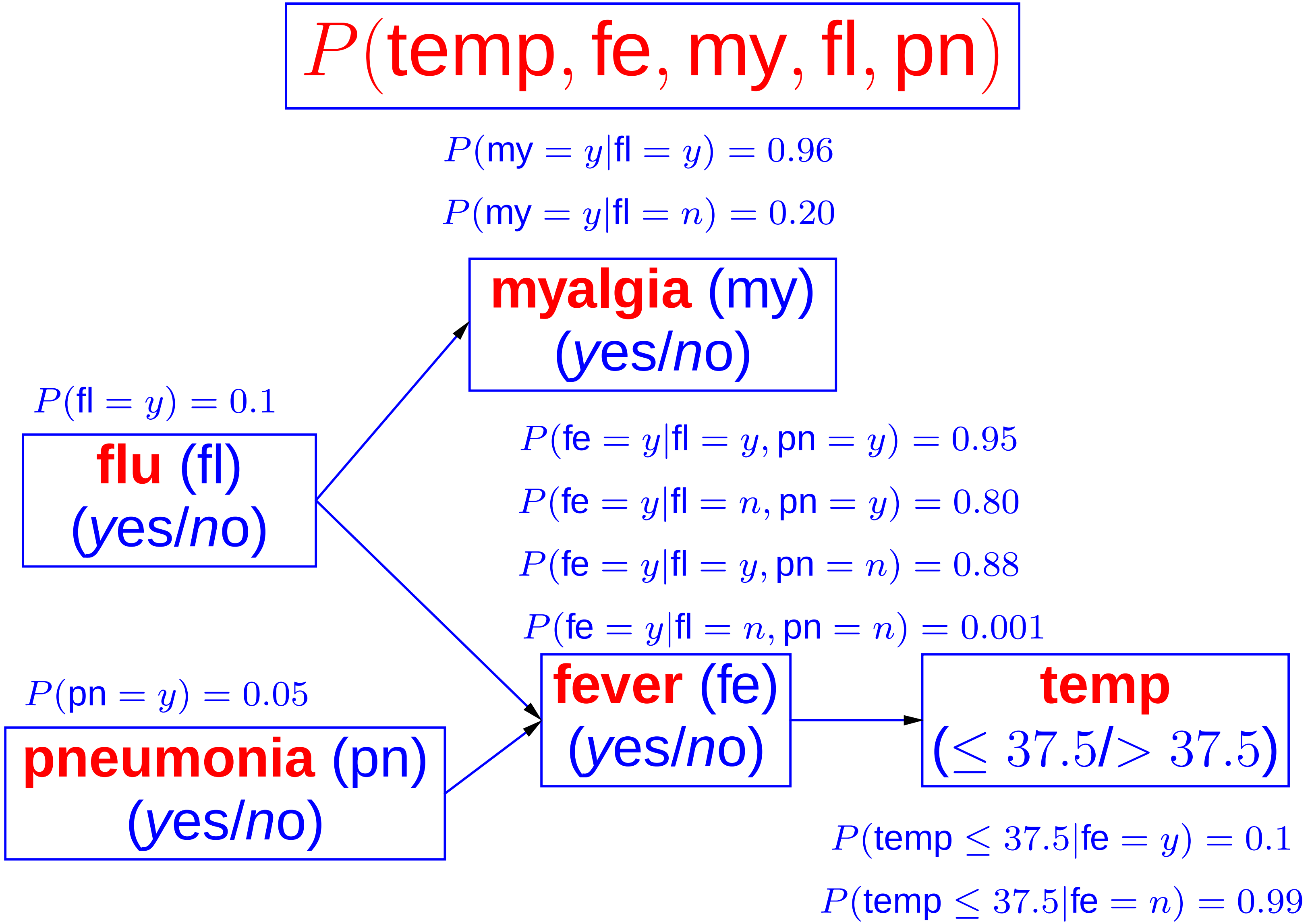
A Bayesian (belief) network is a convenient graphical representation of the independence structure of a joint density



# Bayesian networks

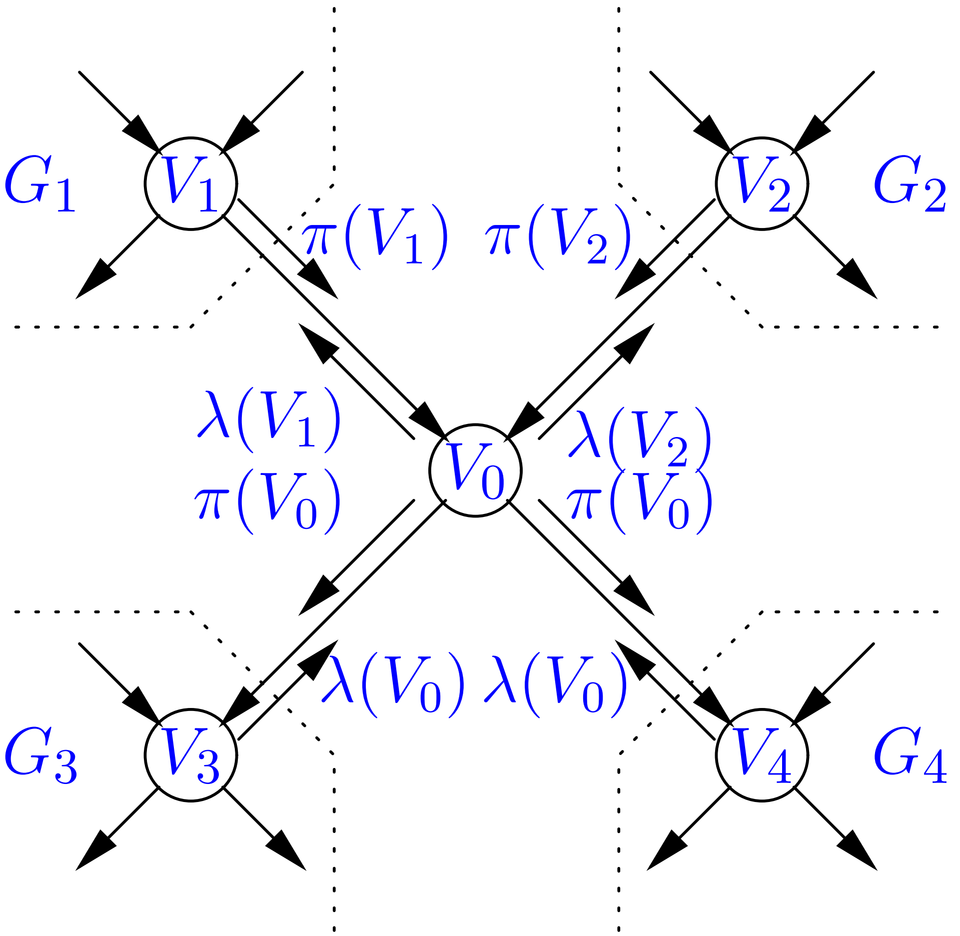
- A Bayesian network consists of:
  - a directed acyclic graph with nodes labeled with random variables
  - a domain for each random variable
  - a set of (conditional) densities for each variable given its parents
- Bayesian networks may consist of discrete or continuous random variables, or both
- We focus on the discrete case
- A Bayesian network is a particular kind of probabilistic graphical model
- Many statistical methods can be represented as graphical models

# Specification of probabilities



# Algorithm: Belief propagation

- Breakthrough algorithm due to Pearl (1988)



# Algorithm: Variable Elimination

- Based on the notion that a belief network specifies a factorization of the joint probability distribution
- See also Daphne Koller's online lectures at Youtube (http://www.youtube.com/watch?v=jz02X3hByac)
- Poole and Mackworth - AI book - Section 6.4.
- Computes factors: functions of variables
- For small networks matches informal procedure for calculating probabilities and utilities

# Adding Utility

- Preferences, utility, decisions (see: AIPSML + end slides)
- Bayesian networks represent joint distributions
- decision networks add
  - decision nodes
  - utility nodes
- Inference: variable elimination, etc.
- Models: single-decision, MDP, POMDP, etc.

# Probabilistic interpretation CF calculus?

- Rule-based uncertainty: $e \to h_x$
  - propagation from antecedent $e$ to conclusion $h$ ($f_{\text{prop}}$)
  - combination of $\wedge$ and $\vee$ evidence in $e$ ($f_\wedge$ and $f_\vee$)
  - co-concluding rules ($f_{\text{co}}$):

$$e_1 \to h_x$$
$$e_2 \to h_y$$

- Bayesian networks: joint probability distribution $P(X_1, \ldots, X_n)$ with marginalisation $\sum_Y P(Y, Z)$ and conditioning $P(Y \mid Z)$
- (Based on Lucas, KB Systems 14 (2001) pp 327–335)

# Propagation

- $f_{\text{prop}}$ (propagation):

$$e' \xrightarrow{\text{CF}(e,e')} e \xrightarrow{\text{CF}(h,e)} h$$

$$\text{CF}(h,e') = \text{CF}(h,e) \cdot \max\{0, \text{CF}(e,e')\}$$

- corresponding Bayesian network (with $P(e')$ extra):

$$E' \xrightarrow{P(E \mid E')} E \xrightarrow{P(H \mid E)} H$$

$$P(h \mid e') = P(h \mid e)P(e \mid e') + P(h \mid \neg e)P(\neg e \mid e')$$

$\Rightarrow P(h \mid \neg e) = 0$ (assumption of CF-model)

# Co-concluding

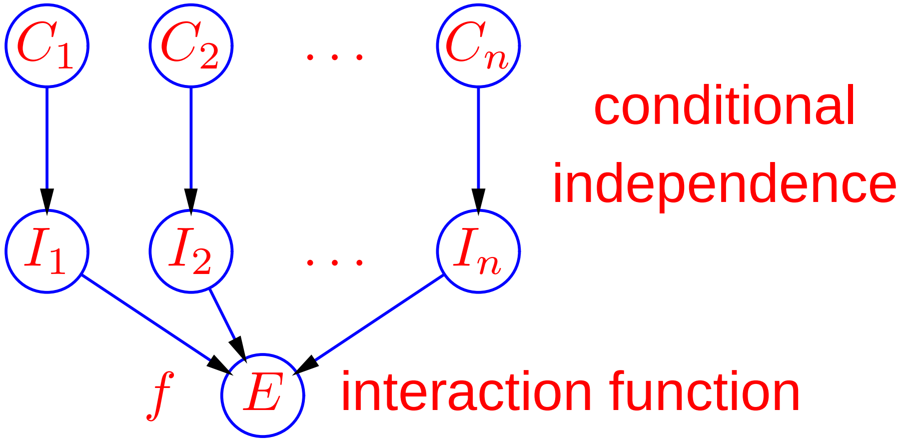
- $f_{\text{co}}$ (co-concluding):

$$\begin{tikzcd}
e'_1 \arrow[dr, "\mathbf{CF}(h, e'_1)"] & \\
& h \\
e'_2 \arrow[ur, "\mathbf{CF}(h, e'_2)"'] &
\end{tikzcd}$$

- idea: see this as uncertain deterministic interaction $\Rightarrow$ causal independence model

$$\begin{tikzcd}
E'_1 \arrow[r] & I_1 \arrow[dr] & & \\
& & H & \text{function } f \\
E'_2 \arrow[r] & I_2 \arrow[ur] & &
\end{tikzcd}$$

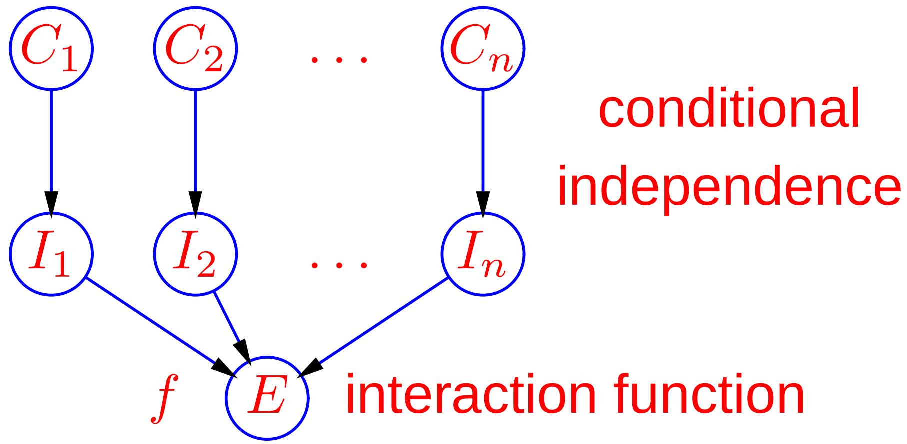
# Causal Independence




$$
\begin{aligned}
P(e \mid C_1, \ldots, C_n) &= \sum_{I_1,\ldots,I_n} P(e \mid I_1, \ldots, I_n) \prod_{k=1}^{n} P(I_k \mid C_k) \\
&= \sum_{f(I_1,\ldots,I_n)=e} \prod_{k=1}^{n} P(I_k \mid C_k)
\end{aligned}
$$

Boolean functions: $P(E \mid I_1, \ldots, I_n) \in \{0, 1\}$ with
$f(I_1, \ldots, I_n) = 1$ if $P(e \mid I_1, \ldots, I_n) = 1$

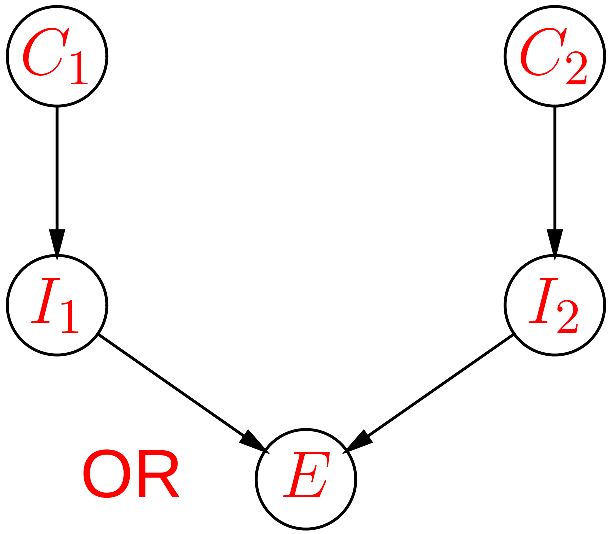
# Causal Independence




$$P(e \mid C_1, \ldots, C_n) \;=\; \sum_{f(I_1,\ldots,I_n)=e} \prod_{k=1}^{n} P(I_k \mid C_k)$$

- Requires specification of one Boolean function and just $n$ probabilities (assuming $P(i_k \mid \neg c_k) = 0$)
- Compare with $2^n$ probabilities for arbitrary $P(e \mid C_1, \ldots, C_n)$
- Simplifies BN construction/facilitates inference

# Example: noisy OR




- Interactions between ‘causes’: logical OR
- Meaning: presence of the intermediate causes $I_k$ produces effect $e$ (i.e. $E = \mathit{true}$)

$$
\begin{aligned}
P(e|C_1, C_2) &= \sum_{I_1 \vee I_2 = e} P(e|I_1, I_2) \prod_{k=1,2} P(I_k \mid C_k) \\
&= P(i_1|C_1)P(i_2|C_2) + P(\neg i_1|C_1)P(i_2|C_2) \\
&\quad + P(i_1|C_1)P(\neg i_2|C_2)
\end{aligned}
$$

# Noisy OR and $f_{\mathbf{co}}$

- $f_{\text{co}}$:

$$\text{CF}(h, e'_1 \textbf{ co } e') = \text{CF}(h, e'_1) + \text{CF}(h, e'_2)(1 - \text{CF}(h, e'_1))$$

for $\text{CF}(h, e'_1) \in [0, 1]$ and $\text{CF}(h, e'_2) \in [0, 1]$

- causal independence with logical OR (noisy OR):

$$\begin{aligned} P(e|C_1, C_2) &= \sum_{I_1 \vee I_2 = e} P(e|I_1, I_2) \prod_{k=1,2} P(I_k \mid C_k) \\ &= P(i_1|C_1)P(i_2|C_2) + P(\neg i_1|C_1)P(i_2|C_2) \\ &\quad + P(i_1|C_1)P(\neg i_2|C_2) \\ &= P(i_1|C_1) + P(i_2|C_2)(1 - P(i_1|C_1)) \end{aligned}$$

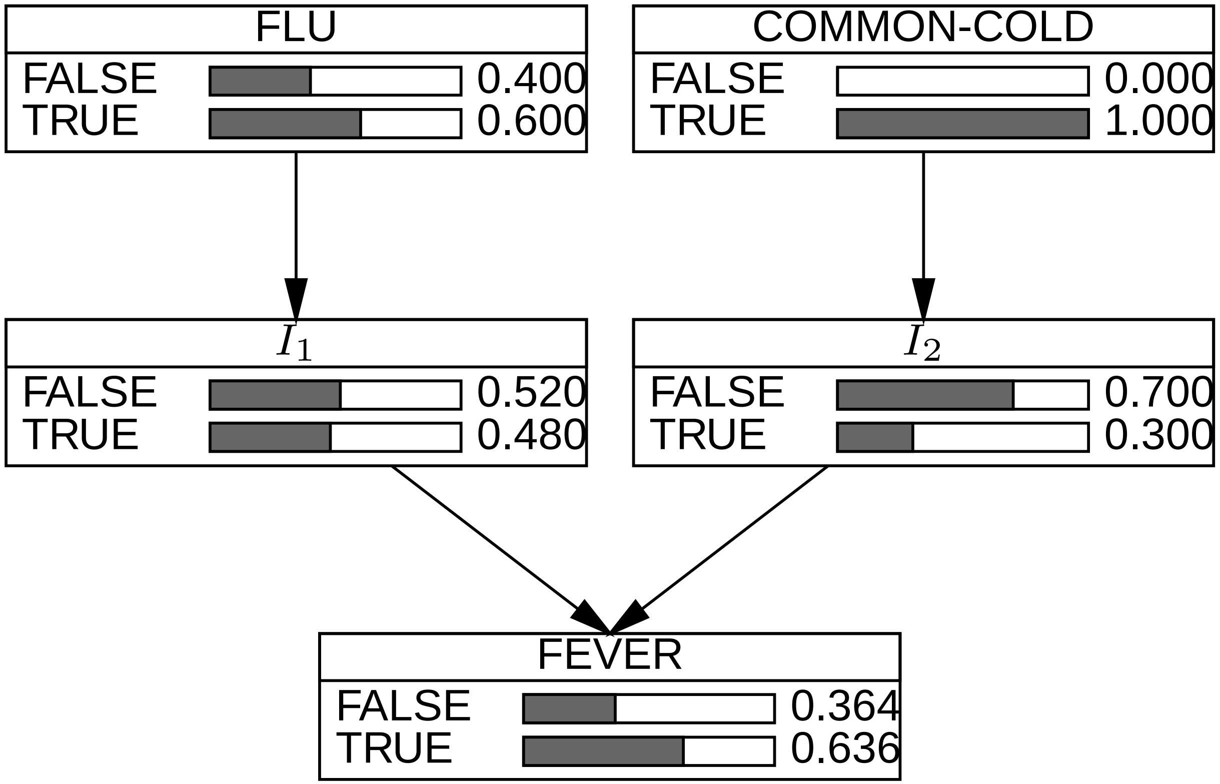
# Example

- The consequences of ‘flu’ and ‘common cold’ on ‘fever’ are modelled by the variables $I_1$ and $I_2$:
  - $P(i_1 \mid \textit{flu}) = 0.8$, and
  - $P(i_2 \mid \textit{common-cold}) = 0.3$
- Furthermore, $P(i_k \mid w) = 0$, $k = 1, 2$, if $w \in \{\neg\textit{flu}, \neg\textit{common-cold}\}$
- Interaction between FLU and COMMON-COLD as noisy-OR:

$$P(\textit{fever} \mid I_1, I_2) = \begin{cases} 0 & \text{if } I_1 = \textit{false} \text{ and } I_2 = \textit{false} \\ 1 & \text{otherwise} \end{cases}$$

# Result

- Bayesian network:




- Fragment CF model:

$$\mathrm{CF}(\mathit{fever}, e_1' \,\mathbf{co}\, e_2') = \mathrm{CF}(\mathit{fever}, e_1') + \mathrm{CF}(\mathit{fever}, e_2')(1 - \mathrm{CF}(\mathit{fever}, e_1'))$$
$$= 0.48 + 0.3(1 - 0.48) = 0.636$$

# Conclusions

- Early rule-based (logical) approach to reasoning with uncertainty was attractive
- However, naive rules + probability can lead to problems
- Bayesian networks and other probabilistic graphical models (Markov networks, chain graphs) are the state of the art for reasoning with uncertainty
- Therefore exploitation of probability theory (also for decisions)
- Although still various rule-based systems are useful for various purposes
- Many rule-based uncertainty reasoning can be mapped (partially) to specific Bayesian network structures
- Next week: probability + logic

# Blanc Slide

empty...

# Decision making

- (The next 20-ish slides are mainly a recap from AIPSML)
- We know how to reason about the state of the world
- Is that enough to implement an intelligent agent?
- No:
  - reasoning without action is void
  - reasoning may require action to gain information
  - action selection requires preferences

# Preferences

- Actions result in outcomes
- Agents have preferences over outcomes
- A rational agent will take the action that has the best outcome for them
- Sometimes agents don't know the outcomes of the actions, but they still need to compare actions
- Agents have to act (doing nothing is often an action).

If $o_1$ and $o_2$ are outcomes

- $o_1 \succeq o_2$ means $o_1$ is at least as desirable as $o_2$
- $o_1 \sim o2$ means $o_1 \succeq o_2$ and $o_2 \succeq o_1$
- $o_1 \succ o_2$ means $o_1 \succeq o_2$ but not $o_2 \succeq o_1$

# Lotteries

- An agent may not know the outcomes of their actions but only have a probability distribution of the outcomes.
- A lottery is a probability distribution over outcomes. It is written

$$p_1 : o_1; p_2 : o_2; \ldots ; p_k : o_k$$

where the $o_i$ are outcomes and $p_i > 0$ such that

$$\sum_i p_i = 1$$

The lottery specifies that outcome $o_i$ occurs with probability $p_i$.
- E.g. $0.1 : cured; 0.9 : uncured$ when receiving treatment

(+Neumann-Morgenstern axioms for utility)

# Rational agents

- If an agent respects the von Neumann-Morgenstern axioms then it is said to be rational
- If an agent is rational, then the preference of an outcome can be quantified using a utility function:

$$U: \text{outcomes} \rightarrow [0,1]$$

such that:

- $o_1 \succeq o_2$ if and only if $U(o_1) \geq U(o_2)$.
- $U([p_1 : o_1, p_2 : o_2, \ldots, p_k : o_k]) = \sum_{i=1}^{k} p_i \cdot U(o_i)$

# Utilities

$$U : \text{outcomes} \rightarrow [0, 1]$$

- Utility is a measure of desirability of outcomes to an agent.
- Let $u(o)$ be the utility of outcome $o$ to the agent.
- Simple goals can be specified by: outcomes that satisfy the goal have utility 1; other outcomes have utility 0.
- Often utilities are more complicated: for example some function of the amount of damage to a robot, how much energy is left, what goals are achieved, and how much time it has taken.

# Decision-making under uncertainty

What an agent should do depends on:

- The agent's beliefs: the ways the world could be, given the agent's knowledge.
- The agent's preferences: what the agent wants and tradeoffs when there are risks.
- The agent's ability: what actions are available to it.

Decision theory specifies how to trade off the desirability and probabilities of the possible outcomes for competing actions.

# Single decisions

- Decision variables are like random variables that an agent gets to choose a value for.
- For a single decision variable, the agent can choose $D = d$ for any $d \in \mathrm{dom}(D)$.
- The expected utility of decision $D = d$ is

$$E(U \mid d) = \sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid d) U(x_1,\ldots,x_n,d)$$

- An optimal single decision is the decision $D = d_{\max}$ whose expected utility is maximal:

$$d_{\max} = \arg \max_{d \in \mathrm{dom}(D)} E(U \mid d)$$

# Example

Suppose:

- $P =$ throw party
- $R =$ rain
- $U(p, \neg r) = 500$, $U(p, r) = -100$, $U(\neg p, r) = 0$, $U(\neg p, \neg r) = 50$
- $P(r \mid P) = P(r) = 0.6$

Then:

$$\begin{aligned} E(U \mid p) &= 0.6 \cdot -100 + 0.4 \cdot 500 = 140 \\ E(U \mid \neg p) &= 0.6 \cdot 0 + 0.4 \cdot 50 = 20 \end{aligned}$$

Conclusion: Party!

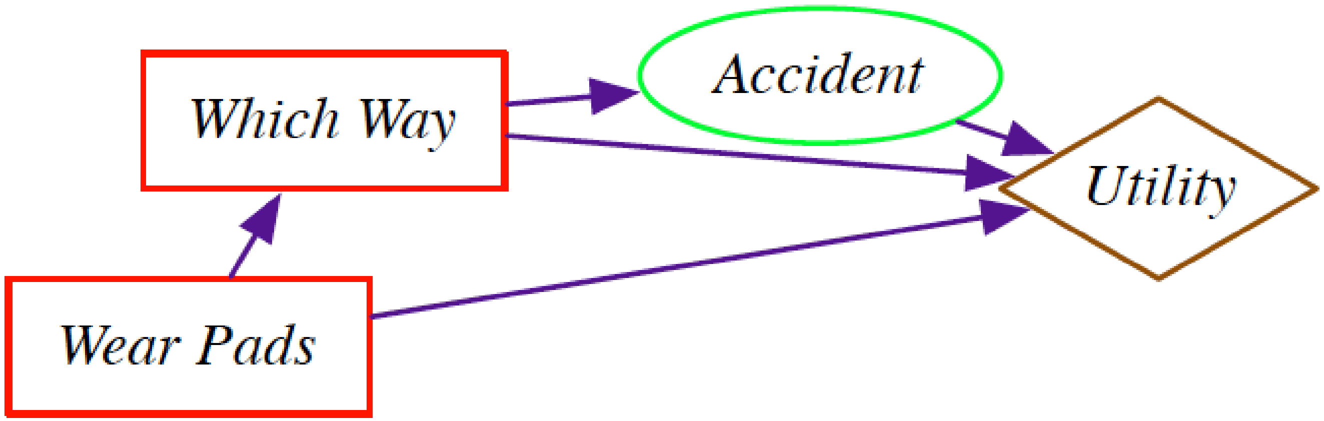
# Sequential decisions

- Multiple decisions made in parallel can be regarded as one big single decision.
- An intelligent agent doesn't carry out just one action or ignore intermediate information
- A more typical scenario is where the agent: observes, acts, observes, acts, . . .
- Subsequent actions can depend on what is observed.
- What is observed depends on previous actions.
- Some actions purely intended to gather information (e.g. diagnostic tests, sensing)
- Sequential decision making (AIPSML): value iteration, reinforcement learning, etc.

# Influence diagram

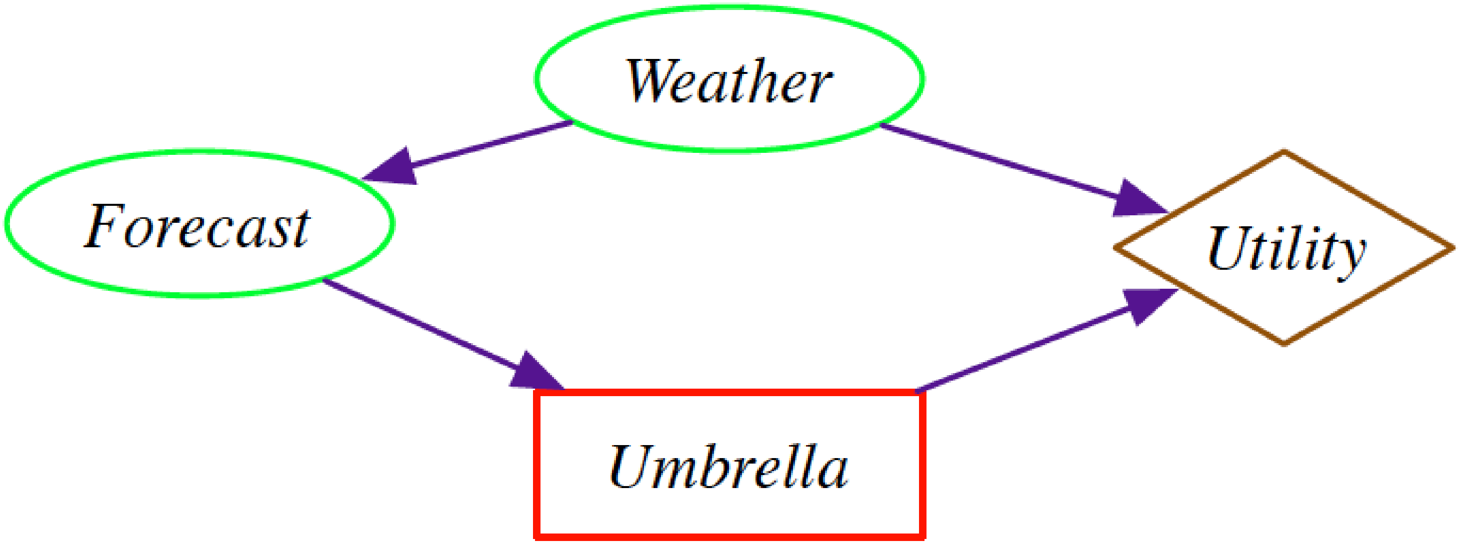
Extend belief networks with:

- Decision nodes, that the agent chooses the value for. Domain is the set of possible actions.Drawn as rectangle.
- Utility node, the parents are the variables on which the utility depends. Drawn as a diamond.




- Shows explicitly which nodes affect whether there is an accident.

# Umbrella network




- You don’t get to observe the weather when you have to decide whether to take your umbrella. You do get to observe the forecast.

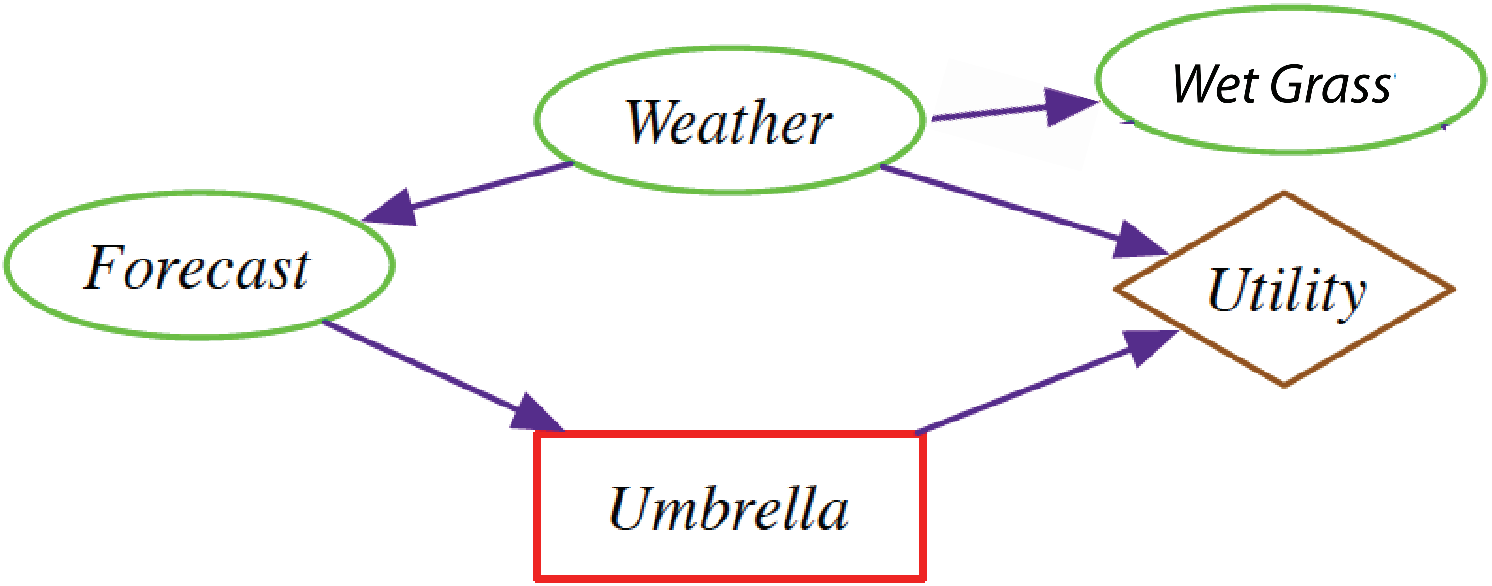
# Finding the optimal policy

- Partial order: $\mathcal{X}_1 \prec D_2, \ldots, \mathcal{X}_{n-1} \prec D_n \prec \mathcal{X}_n$
- Recall:
  $E(U \mid d) = \sum_{x_1,\ldots,x_n} P(x_1, \ldots, x_n \mid d) U(x_1, \ldots, x_n, d)$
- The maximal expected utility $U^*$ is given by

$$U^* = \sum_{\mathcal{X}_1} \max_{D_2} \cdots \sum_{\mathcal{X}_{n-1}} \max_{D_n} \sum_{\mathcal{X}_n} \prod_{i \in \mathcal{I}} P(x_i \mid \pi(x_i)) \sum_{j \in \mathcal{J}} U_j(\pi(u_j))$$

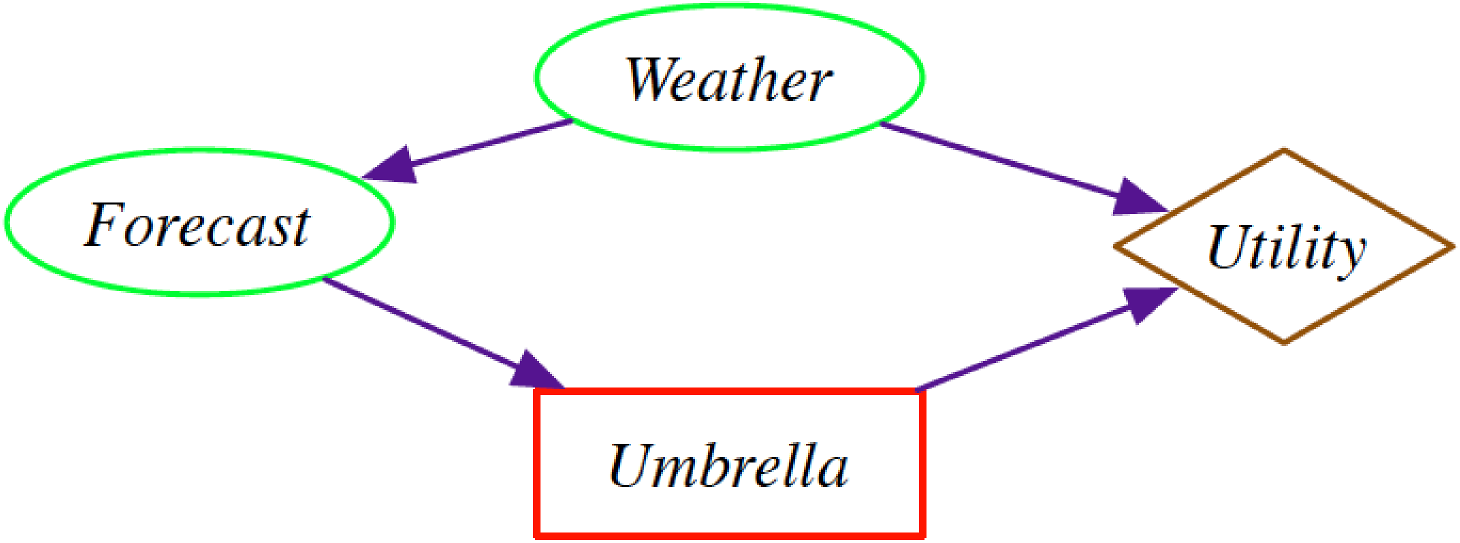
- The optimal policy can be found by variable elimination while maximizing over decisions:
  - first consider the last decision
  - find an optimal decision for each value of its parents and produce a factor of these maximum values.
  - recursively solve for the remaining decisions

# Umbrella network




- You don’t get to observe the weather when you have to decide whether to take your umbrella. You do get to observe the forecast. Rain will cause wet grass.

# Finding the optimal policy




- Remove all variables not ancestors of the utility node

# Finding the optimal policy

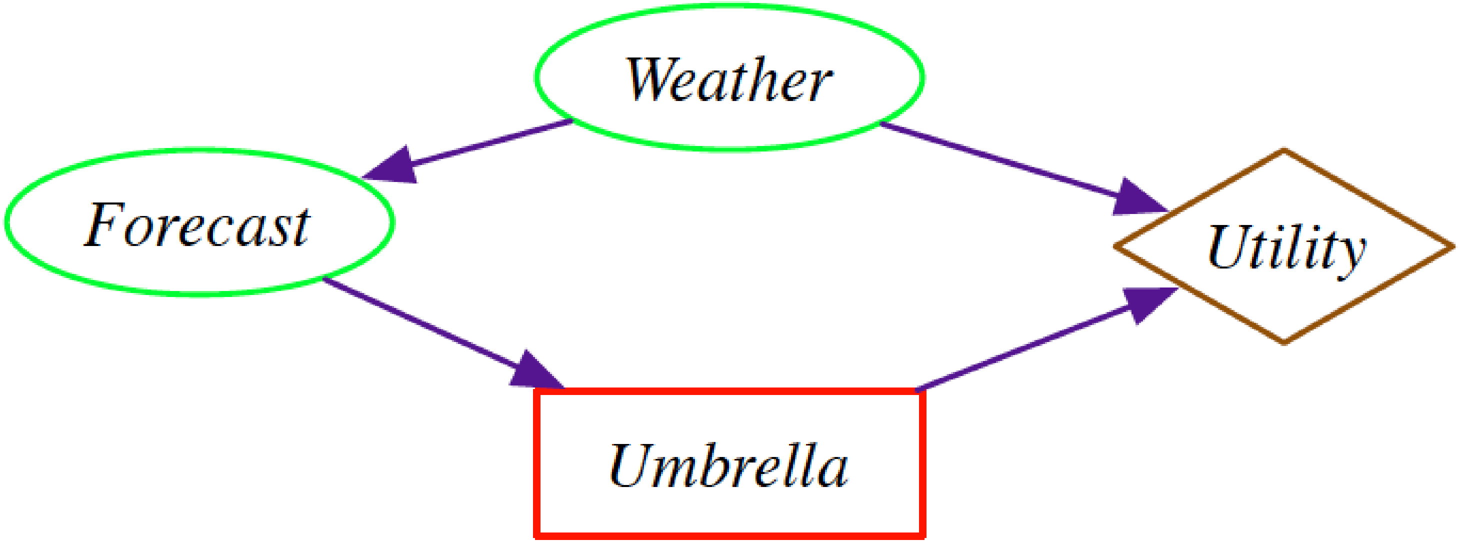
| Weather | Value |
|---|---|
| norain | 0.7 |
| rain | 0.3 |

| Weather | Fcast | Value |
|---|---|---|
| norain | sunny | 0.7 |
| norain | cloudy | 0.2 |
| norain | rainy | 0.1 |
| rain | sunny | 0.15 |
| rain | cloudy | 0.25 |
| rain | rainy | 0.6 |

| Weather | Umb | Value |
|---|---|---|
| norain | take | 20 |
| norain | leave | 100 |
| rain | take | 70 |
| rain | leave | 0 |

- Create a factor for each conditional probability table and a factor for the utility.

# Finding the optimal policy




$$U^* = \sum_{F,W} \max_{U} f_1(W) f_2(W,F) f_3(W,U)$$

# Finding the optimal policy

- Sum out variables not (parents of) a decision node $D$

$$
\begin{aligned}
U^* &= \sum_F \max_U \sum_W f_1(W) f_2(W,F) f_3(W,U) \\
&= \sum_F \max_U f_4(F,U)
\end{aligned}
$$

| *Forecast* | *Umbrella* | *Value* |
|---|---|---|
| *sunny* | *takeIt* | 12.95 |
| *sunny* | *leaveIt* | 49.0 |
| *cloudy* | *takeIt* | 8.05 |
| *cloudy* | *leaveIt* | 14.0 |
| *rainy* | *takeIt* | 14.0 |
| *rainy* | *leaveIt* | 7.0 |

# Finding the optimal policy

$$U^* = \sum_F \max_U f_4(F, U)$$

- Select $D$ that is in a factor $f$ with (some of) its parents
- Eliminate $D$ by maximizing. This returns:
  - the optimal decision function for $D$, $\arg\max_D f$
  - a new factor to use in VE, $\max_D f$

| Forecast | Umbrella |
|---|---|
| sunny | leaveIt |
| cloudy | leaveIt |
| rainy | takeIt |

| Forecast | Value |
|---|---|
| sunny | 49.0 |
| cloudy | 14.0 |
| rainy | 14.0 |

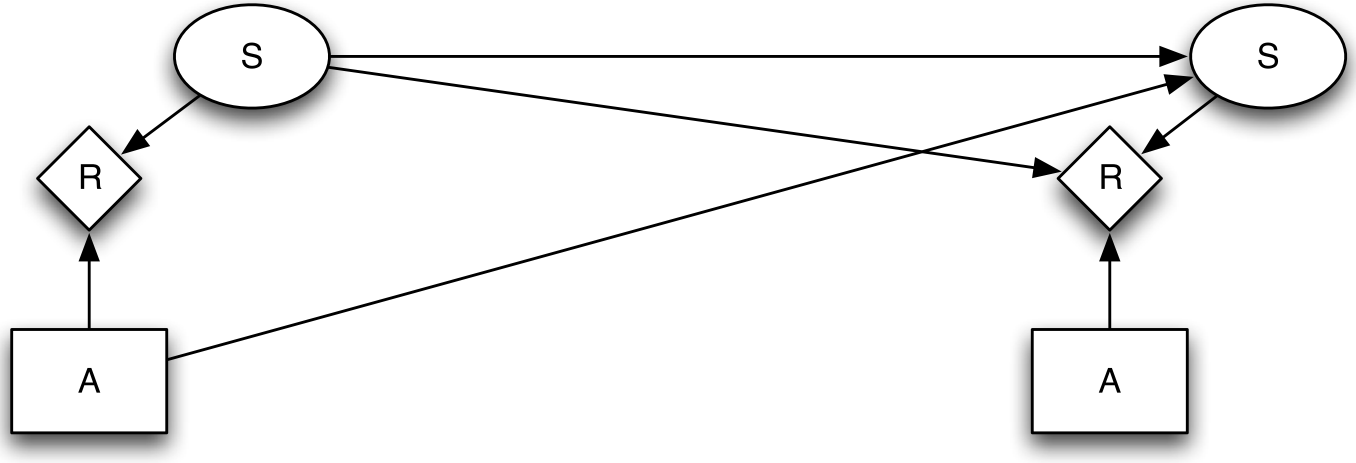
- the final sum returns the maximized expected utility:

$$U^* = \sum_F f_5(F) = 77$$

# Other properties

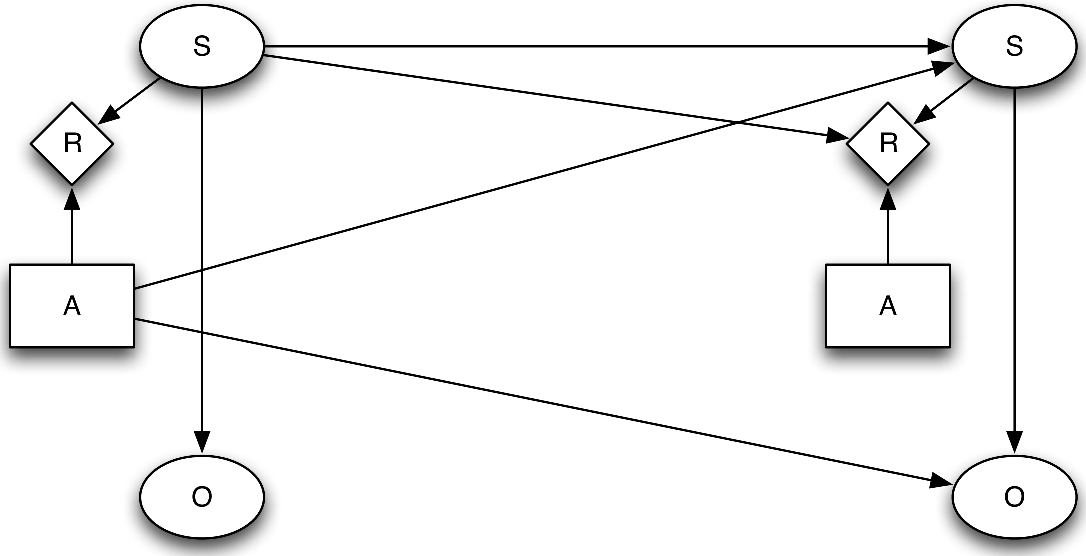
- Value of information:
  - The amount someone would be willing to pay for information on $X$ prior to making a decision $D$
  - The value of information on $X$ for decision $D$ is the expected utility of the network with an arc from $X$ to $D$ minus exp. util. of the network without the arc.
- Value of control:
  - The amount someone would be willing to pay in order to be able to control a random variable $X$
  - The value of control of a variable $X$ is the expected utility of the network when you make $X$ a decision variable minus the expected utility of the network when $X$ is a random variable.

# MDP




- $S$, a set of states of the world.
- $A$, a set of actions.
- $P : S \times S \times A \to [0, 1]$, written as $P(s'|s, a)$
- $R : S \times A \times S \to R$, written as $R(s, a, s')$

# POMDP




As an MDP but additionally:

- $O$, a set of possible observations;
- $P(s_0)$, which gives the probability distribution of the starting state
- $P(o \mid s, a)$, which gives the probability of observing $o$ given the state is $s$ and the previous action $a$.