UNCERTAINTY

Introduction

- So far we have considered mainly accessible/observable environments.
  - Or pretended that environments were accessible/observable.
- Clearly not true of the real world:
  - Is is raining in Manhattan?
- Partial observability can arise for many reasons.
  - World structure *vs.* sensor ability.
  - Sensor noise.
  - Computational complexity.

Types of problem

- Who is outside in the corridor?
  - Uncertainty
- The radio says it is raining in Manhattan, but when I phone my wife she says it isn't raining.
  - Ambiguity
  - Contradiction.
- Is it true that "Simon is tall"
  - Vagueness
- Who will be in next year's World Series?
  - Ignorance

Uncertainty

- Let action A_t = leave for airport t minutes before flight
  - Will A_t get me there on time?
- Problems:
  1. partial observability (road state, other drivers' plans, etc.)
  2. noisy sensors (1010 WINS traffic reports)
  3. uncertainty in action outcomes (flat tire, etc.)
  4. immense complexity of modelling and predicting traffic

- Hence a purely logical approach either:
    1. risks falsehood: "A_{90} will get me there on time"
    2. leads to conclusions that are too weak for decision making: "A_{90} will get me there on time if there's no accident on the Williamsburg bridge, and it doesn't rain and my tires remain intact etc etc."
- (A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport . . .)

- How could an agent cope with this?

Methods for handling uncertainty

- *Nonmonotonic* logic
  - Assume my car does not have a flat tire
  - Assume A_{25} works unless contradicted by evidence
- Issues: What assumptions are reasonable? How to handle contradiction?

- *Rules with fudge factors*:
  - $A_{25} \mapsto_{0.3}$ *AtAirportOnTime*
  - *Sprinkler* $\mapsto_{0.99}$ *WetGrass*
  - *WetGrass* $\mapsto_{0.7}$ *Rain*
- Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*??
- Semantics?

- *Probability*
  - Given the available evidence,
    A_{25} will get me there on time with probability 0.04
- Issues: Computational complexity, obtaining values, semantics.
  - We will consider the computational issues in some detail.

- Nate Silver, Five Thirty Eight

Probability

- Probabilistic assertions *summarize* effects of
  - laziness: failure to enumerate exceptions, qualifications, etc.
  - ignorance: lack of relevant facts, initial conditions, etc.
- *Subjective* or *Bayesian* probability:
  - Probabilities relate propositions to one's own state of knowledge

$$P(A_{25}|\text{no reported accidents}) = 0.06$$

- Probabilities of propositions change with new evidence:

$$P(A_{25}|\text{no reported accidents}, \ 5\text{ a.m.}) = 0.15$$

(Analogous to logical entailment status $KB \models \alpha$, not truth.)

Making decisions under uncertainty

- Suppose I believe the following:

$$\begin{aligned}
P(A_{25} \text{ gets me there on time}|\ldots) &= 0.04 \\
P(A_{90} \text{ gets me there on time}|\ldots) &= 0.70 \\
P(A_{120} \text{ gets me there on time}|\ldots) &= 0.95 \\
P(A_{1440} \text{ gets me there on time}|\ldots) &= 0.9999
\end{aligned}$$

Which action to choose?

- Depends on my *preferences* for missing flight vs. airport cuisine, sleeping on a bench, and so on.
- *Utility theory* is used to represent and infer preferences
- *Decision theory* = utility theory + probability theory
- We will come back to decision theory with a vengence next time.

Probability basics

- Begin with a set Ω—the *sample space*.
- This is all the possible things that could happen.
  - 6 possible rolls of a die.
- $\omega \in \Omega$ is a *sample point*, *possible world*, *atomic event*.

- A *probability space* or *probability model* is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ such that:

$$0 \leq P(\omega) \leq 1$$
$$\sum_{\omega} P(\omega) = 1$$

$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6.$

- An *event* A is any subset of Ω

$$P(A) = \sum_{\{\omega \in A\}} P(\omega)$$

$$P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$$

Random variables

- A *random variable* is a function from sample points to some range.
  - $raining(Brooklyn) = true$.
  - $temperature(234NE) = 73$
- P induces a *probability distribution* for any r.v. X:

$$P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

$$P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$$

Propositions

- Think of a proposition as the event (set of sample points) where the proposition is true
- Given Boolean random variables A and B:

  event a = set of sample points where $A(\omega) = true$

  event $\neg a$ = set of sample points where $A(\omega) = false$

  event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$

- Often in AI applications, the sample points are *defined* by the values of a set of random variables.
- A state can be defined by a set of Boolean variables.

$$a \wedge b \wedge \neg c \qquad A = true, B = true, C = false$$

  This is then just a sample point.

- Thus, with Boolean variables, sample point = propositional logic model

  $A = true, B = false \qquad a \wedge \neg b$

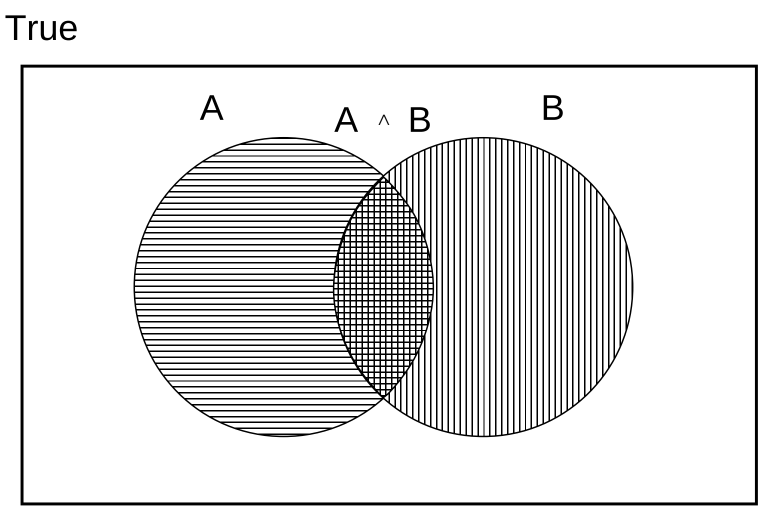
- Proposition = disjunction of atomic events in which it is true

  $$(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$$
  $$\Rightarrow \; P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$$

Why use probability?

- The definitions imply that certain logically related events must have related probabilities

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$



- de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.
  - Dutch book argument

Syntax for propositions

- *Propositional* or *Boolean* random variables
  - *Cavity* (do I have a cavity?)
  - $Cavity = true$ is a proposition, also written *cavity*
- *Discrete* random variables (finite or infinite)
  - *Weather* is one of $\langle sunny, rain, cloudy, snow\rangle$
  - $Weather = rain$ is a proposition

  Values must be exhaustive and mutually exclusive
- *Continuous* random variables (bounded or unbounded)
  - $Temp = 21.6$; also allow, e.g., $Temp < 22.0$.
- Arbitrary Boolean combinations of basic propositions

Prior probability

- *Prior* or *unconditional probabilities* of propositions

  $P(Cavity = true) = 0.1$ and $P(Weather = sunny) = 0.72$

  correspond to belief before (prior) to arrival of any (new) evidence.

- *Probability distribution* gives values for all possible assignments:

  $\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$

- Distribution is normalized, i.e., sums to 1

- *Joint probability distribution* for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)

  $\mathbf{P}(Weather, Cavity)$ = a 4×2 matrix of values

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

- *Every question about a domain can be answered by the joint distribution because every event is a sum of sample points*

Conditional probability

- *Conditional* or *posterior probabilities*

  $P(cavity|toothache) = 0.8$

  given that *toothache* *is all I know NOT* "if *toothache* then 80% chance of *cavity*"

- Notation for conditional distributions:

  $\mathbf{P}(Cavity|Toothache)$

  A 2-element vector of 2-element vectors.

- If we know more, e.g., *cavity* is also given, then we have

  $$P(cavity|toothache, cavity) = 1$$

  Note: the less specific belief *remains valid* after more evidence arrives, but is not always *useful*

- New evidence may be irrelevant, allowing simplification

  $$P(cavity|toothache, jetsWin) = P(cavity|toothache) = 0.8$$

- This kind of inference, sanctioned by domain knowledge, is crucial

- Definition of conditional probability:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

- *Product rule* gives an alternative formulation:

$$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

- A general version holds for whole distributions,

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$$

(View as a 4×2 set of equations, *not* matrix multiplication)

- *Chain rule* is derived by successive application of product rule:

$$\begin{aligned}
\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\
&= \mathbf{P}(X_1, \ldots, X_{n-2})\,\mathbf{P}(X_{n-1}|X_1, \ldots, X_{n-2}) \\
&\qquad \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\
&= \ldots \\
&= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})
\end{aligned}$$

Inference by enumeration

- Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

- For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$$

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

- For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$$

- $P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

- For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$$

- $P(cavity \vee toothache) =$
  $0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

- Can also compute conditional probabilities:

$$
\begin{aligned}
P(\neg cavity|toothache) &= \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \\
&= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\end{aligned}
$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

- Denominator can be viewed as a normalization constant α

$$\begin{aligned} \mathbf{P}(Cavity|toothache) &= \alpha\, \mathbf{P}(Cavity, toothache) \\ &= \alpha\, [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\ &= \alpha\, [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\ &= \alpha\, \langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle \end{aligned}$$

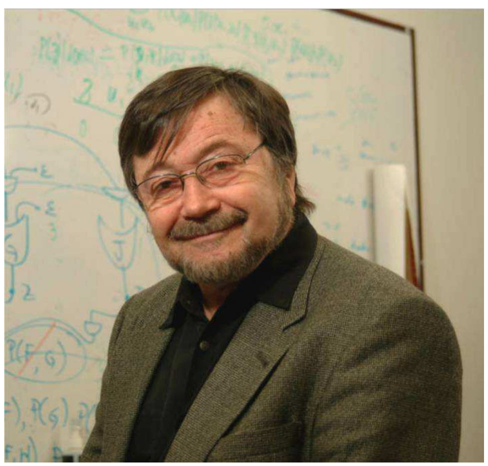
Inference by enumeration

- Let $\mathbf{X}$ be all the variables.
- Typically, we want 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:

$$\begin{aligned}\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) &= \alpha\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) \\ &= \alpha\sum_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})\end{aligned}$$

- The terms in the summation are joint entries because **Y**, **E**, and **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???
- This problem effectively stopped the use of probability in AI until the mid 80s

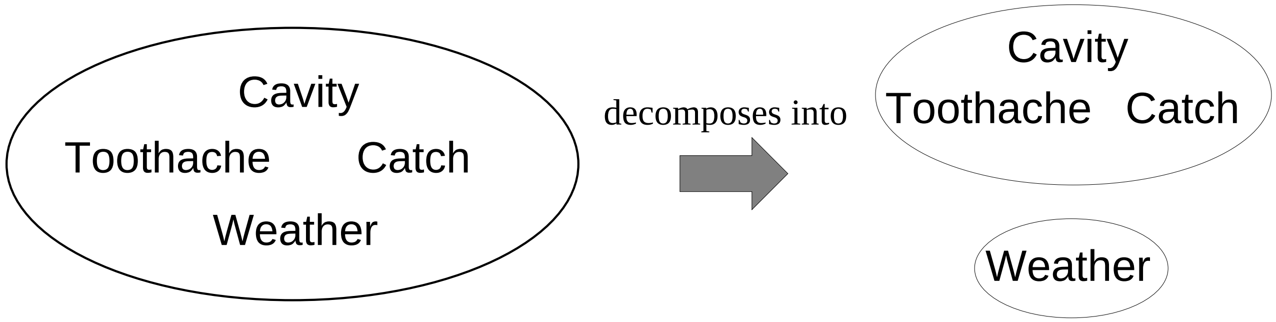


Independence

- A and B are *independent* iff

$$\begin{aligned} \mathbf{P}(A|B) &= \mathbf{P}(A), \text{ or} \\ \mathbf{P}(B|A) &= \mathbf{P}(B), \text{ or} \\ \mathbf{P}(A, B) &= \mathbf{P}(A)\mathbf{P}(B) \end{aligned}$$

- Why is this interesting?
  - Can help with the size of the problem.



- $\mathbf{P}(Toothache, Catch, Cavity, Weather)$
  $= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$
- 32 entries reduced to 12; for n independent biased coins, $2^n \rightarrow n$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

- $\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1$ = 7 independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

$$\mathbf{P}(catch|toothache, cavity) = \mathbf{P}(catch|cavity) \tag{1}$$

- The same independence holds if I haven't got a cavity:

$$\mathbf{P}(catch|toothache, \neg cavity) = \mathbf{P}(catch|\neg cavity) \tag{2}$$

- *Catch* is *conditionally independent* of *Toothache* given *Cavity*

$$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$$

- Equivalent statements:

$$\begin{aligned}
\mathbf{P}(Toothache|Catch, Cavity) &= \mathbf{P}(Toothache|Cavity) \\
\mathbf{P}(Toothache, Catch|Cavity) &= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)
\end{aligned}$$

- Write out full joint distribution using chain rule:

$$\begin{aligned}
&\mathbf{P}(Toothache, Catch, Cavity) \\
&= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch, Cavity) \\
&= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity) \\
&= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

  2 + 2 + 1 = 5 independent numbers (equations 1 and 2 remove 2)

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.
  - Can often make conditional independence statements when little else is known.

Bayes' Rule

- Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Rightarrow \textit{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X|Y)\mathbf{P}(Y)$$

Rev. T. Bayes

- Useful for assessing *diagnostic* probability from *causal* probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

- Let M be meningitis, S be stiff neck:

$$\begin{aligned} P(m|s) &= \frac{P(s|m)P(m)}{P(s)} \\ &= \frac{0.8 \times 0.0001}{0.1} \\ &= 0.0008 \end{aligned}$$

- Note: posterior probability of meningitis still very small!

Bayes' Rule and conditional independence

- So, in our running example

$$\begin{aligned}&\mathbf{P}(\mathit{Cavity}|\mathit{toothache} \wedge \mathit{catch})\\ &= \alpha\,\mathbf{P}(\mathit{toothache} \wedge \mathit{catch}|\mathit{Cavity})\mathbf{P}(\mathit{Cavity})\\ &= \alpha\,\mathbf{P}(\mathit{toothache}|\mathit{Cavity})\mathbf{P}(\mathit{catch}|\mathit{Cavity})\mathbf{P}(\mathit{Cavity})\end{aligned}$$

- This is an example of a *naive Bayes* model:

$$\mathbf{P}(\mathit{Cause}, \mathit{Effect}_1, \ldots, \mathit{Effect}_n) = \mathbf{P}(\mathit{Cause}) \prod_i \mathbf{P}(\mathit{Effect}_i|\mathit{Cause})$$

- Total number of parameters is *linear* in n

What we learned so far

- Probability is a rigorous formalism for uncertain knowledge
- *Joint probability distribution* specifies probability of every *atomic event*
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size
- *Independence* and *conditional independence* provide the tools
- Next we'll look at how this is used.

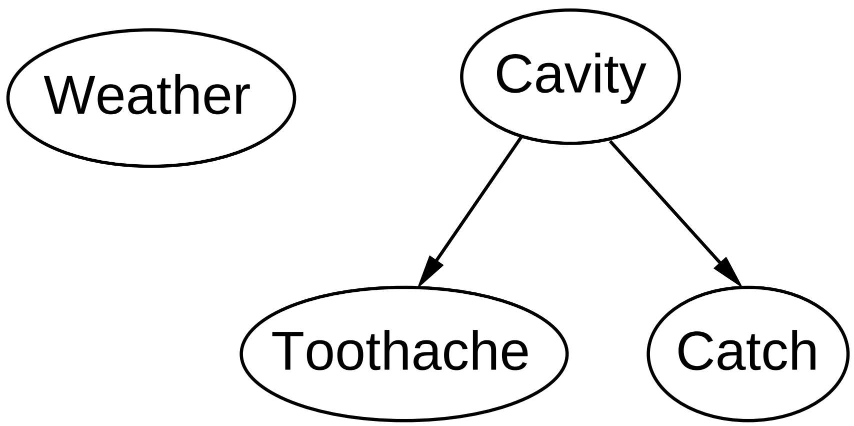
Bayesian networks

- 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 ≈ "directly influences") a conditional distribution for each node given its parents

$$\mathbf{P}(X_i | Parents(X_i))$$

- 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

- Topology of network encodes conditional independence assertions:

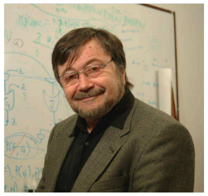


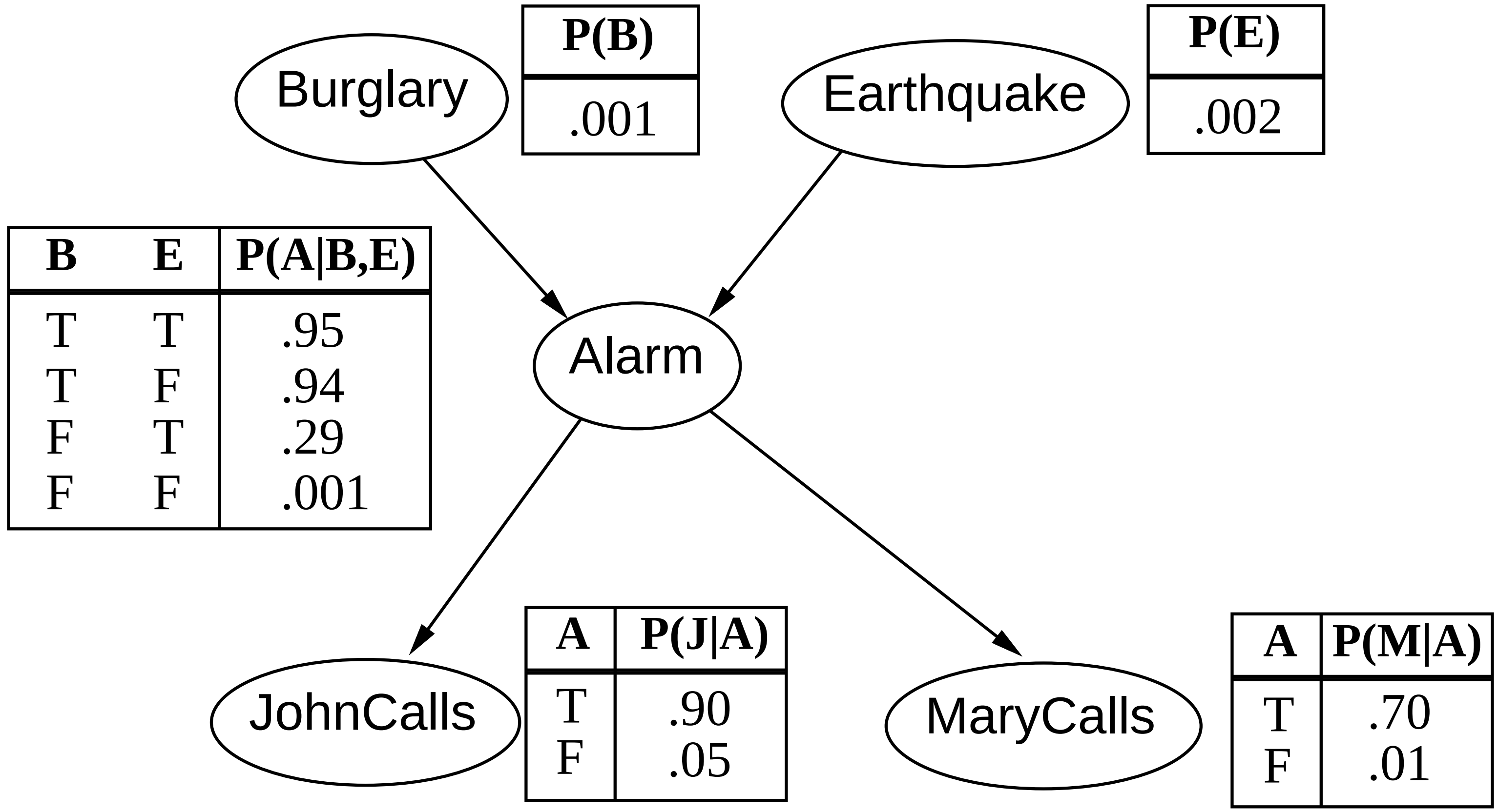
- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

- An example (from California):

  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: *Burglar*, *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





P(B)
.001

P(E)
.002

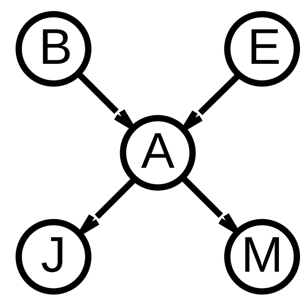
B	**E**	**P(A|B,E)**
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	**P(J|A)**
T	.90
F	.05

A	**P(M|A)**
T	.70
F	.01

Compactness

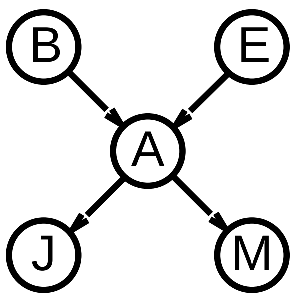
- A CPT for Boolean X_i with k Boolean parents has 2^krows for the combinations of parent values
- Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1 - p$)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
  - grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1+1+4+2+2 = 10$ numbers (vs. $2^5 - 1 = 31$)



Global semantics

- *Global* semantics defines the full joint distribution as the product of the local conditional distributions:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$



- $P(j \land m \land a \land \neg b \land \neg e)$

$$=$$

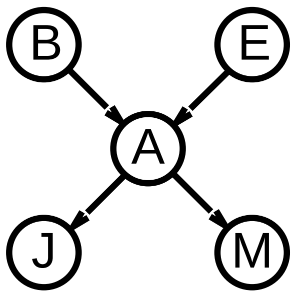
Global semantics

- *Global* semantics defines the full joint distribution as the product of the local conditional distributions:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

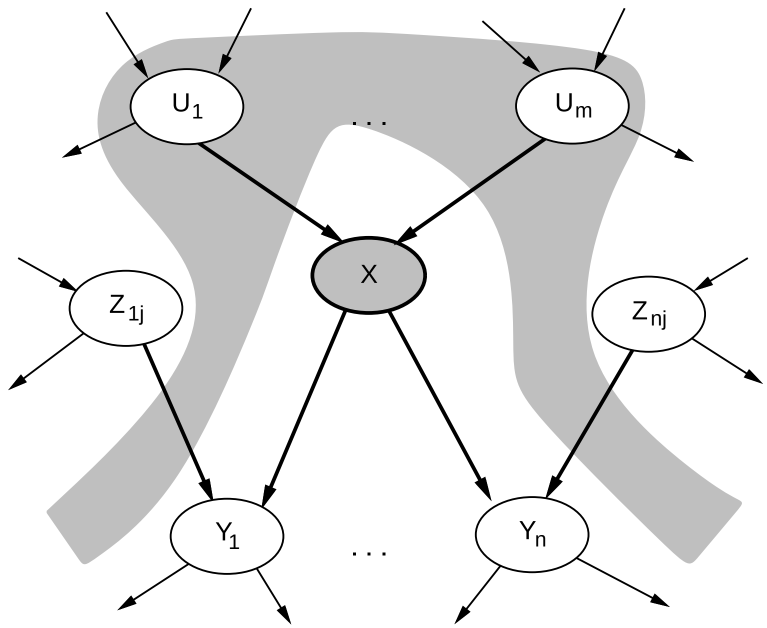
- $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$\begin{aligned} &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$



Local semantics

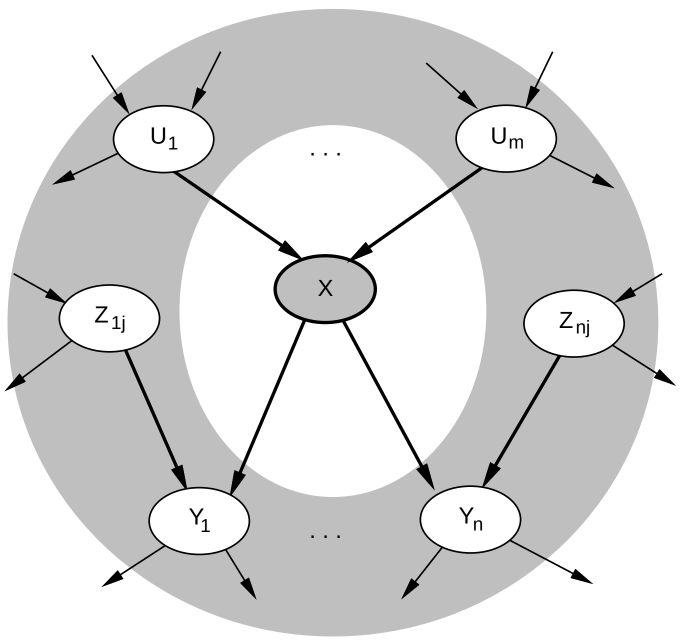
- *Local* semantics: each node is conditionally independent of its nondescendants given its parents

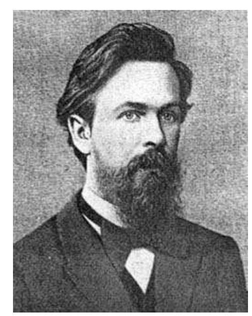


- Theorem: Local semantics $\Leftrightarrow$ global semantics

Markov blanket

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





Compact conditional distributions

- CPT grows exponentially with number of parents
  - Use *canonical* distributions that are defined compactly
- *Deterministic* nodes are the simplest case.
- $X = f(Parents(X))$ for some function f
  - Boolean functions:

$$NorthAmerican \;\Leftrightarrow\; Canadian \vee US \vee Mexican$$

  - Numerical relationships among continuous variables

$$\frac{\partial Level}{\partial t} = \text{inflow + precipitation - outflow - evaporation}$$

- *Noisy-OR* distributions model multiple noninteracting causes
  1. Parents $U_1 \ldots U_k$ include all causes (can add *leak node*)
  2. Independent failure probability q_i for each cause alone
     $\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - prod_{i=1}^{j} q_i$

Cold	*Flu*	*Malaria*	$P(Fever)$	$P(\neg Fever)$
F	F	F	**0.0**	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

- Number of parameters *linear* in number of parents

Inference tasks

- *Simple queries*: compute posterior marginal $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$

  $P(NoGas|Gauge = empty, Lights = on, Starts = false)$

- *Conjunctive queries*

  $\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$

- *Optimal decisions*: decision networks include utility information; probabilistic inference required for $P(outcome|action, evidence)$

- *Value of information*: which evidence to seek next?

- *Sensitivity analysis*: which probability values are most critical?

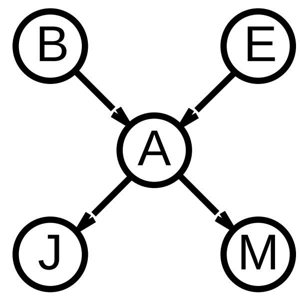
- *Explanation*: why do I need a new starter motor?

Inference by enumeration

- Simplest approach to evaluating the network is to do just as we did for the dentist example.
- Difference is that we use the structure of the network to tell us which sets of joint probabilities to use.
  - Thanks Professor Markov
- Gives us a slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

- Simple query on the burglary network.

$$\mathbf{P}(B|j,m)$$
$$= \mathbf{P}(B,j,m)/P(j,m)$$
$$= \alpha\mathbf{P}(B,j,m)$$
$$= \alpha\ \Sigma_e\ \Sigma_a\ \mathbf{P}(B,e,a,j,m)$$



- Rewrite full joint entries using product of CPT entries:
  - $\mathbf{P}(B|j,m)$
  - $= \alpha\ \Sigma_e\ \Sigma_a\ \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)$
  - $= \alpha\mathbf{P}(B)\ \Sigma_e\ P(e)\ \Sigma_a\ \mathbf{P}(a|B,e)P(j|a)P(m|a)$

Enumeration algorithm

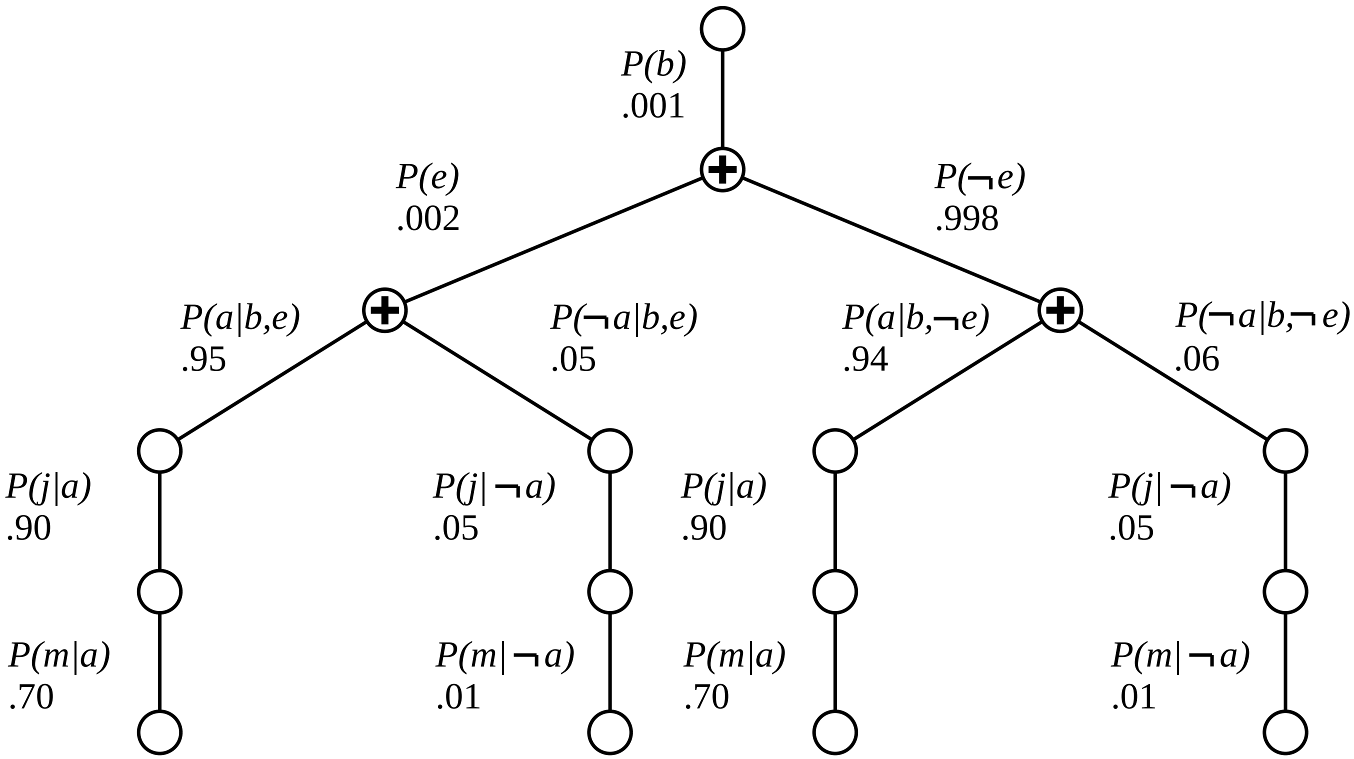
function ENUMERATION-ASK(X, $\mathbf{e}$, bn) **returns** a distribution over X

inputs: X, the query variable
$\mathbf{e}$, observed values for variables $\exists$
bn, a Bayesian network with variables $\{X\} \cup \exists \cup \mathbf{Y}$

$\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
for each value x_i of X **do**
  extend $\mathbf{e}$ with value x_i for X
  $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(VARS[bn], $\mathbf{e}$)
return NORMALIZE($\mathbf{Q}(X)$)

function ENUMERATE-ALL(*vars*, **e**) **returns** a real number
if EMPTY?(*vars*) **then return** 1.0
$Y \leftarrow$ FIRST(*vars*)
if Y has value y in **e**
then return $P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST(*vars*), **e**)
else return $\Sigma_y\ P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST(*vars*), $\mathbf{e}_y$)
where $\mathbf{e}_y$ is **e** extended with $Y = y$

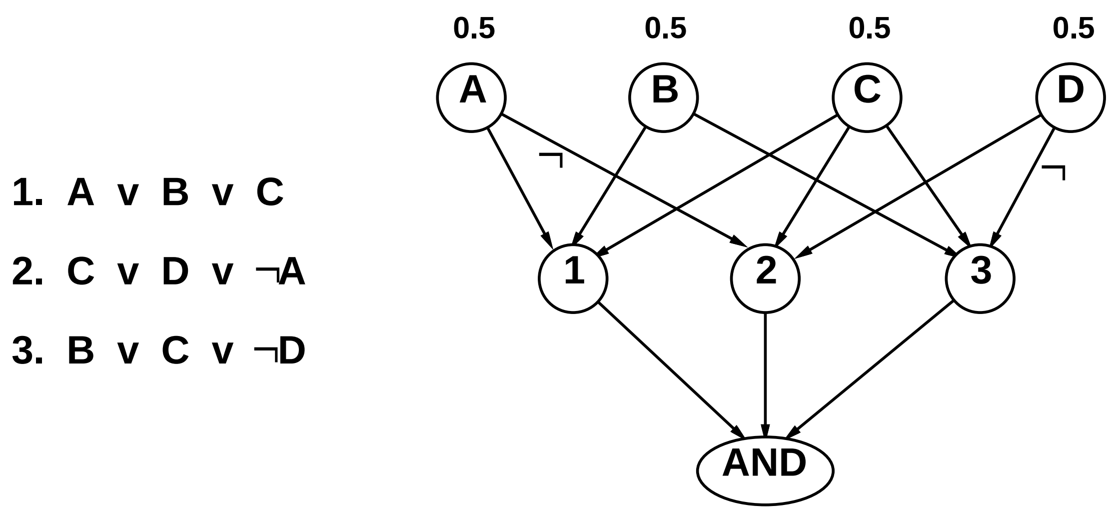
Evaluation tree



- Enumeration is inefficient: repeated computation
  - Computes $P(j|a)P(m|a)$ for each value of e

Complexity of exact inference

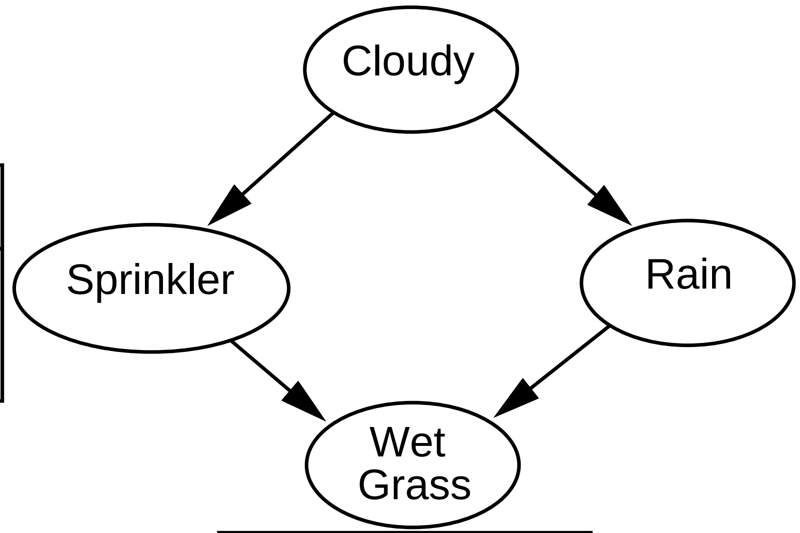
- *Singly connected* networks (or *polytrees*)
  - any two nodes are connected by at most one (undirected) path
  - time and space cost of variable elimination are $O(d^k n)$
- *Multiply connected* networks:
  - can reduce 3SAT to exact inference $\Rightarrow$ NP-hard
  - equivalent to *counting* 3SAT models $\Rightarrow$ #P-complete



Inference by stochastic simulation

- Basic idea:
  1. Draw N samples from a sampling distribution S
  2. Compute an approximate posterior probability $\hat{P}$
  3. Show this converges to the true probability P

P(C)
.50

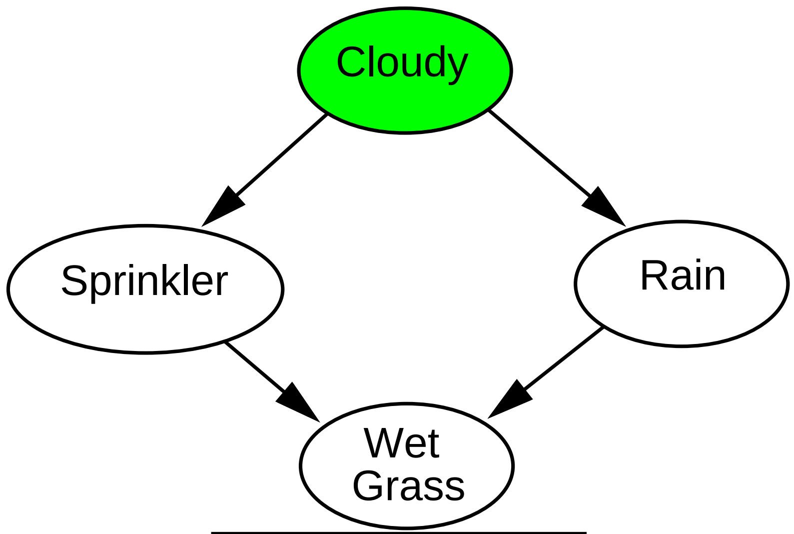


C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

P(C)
.50

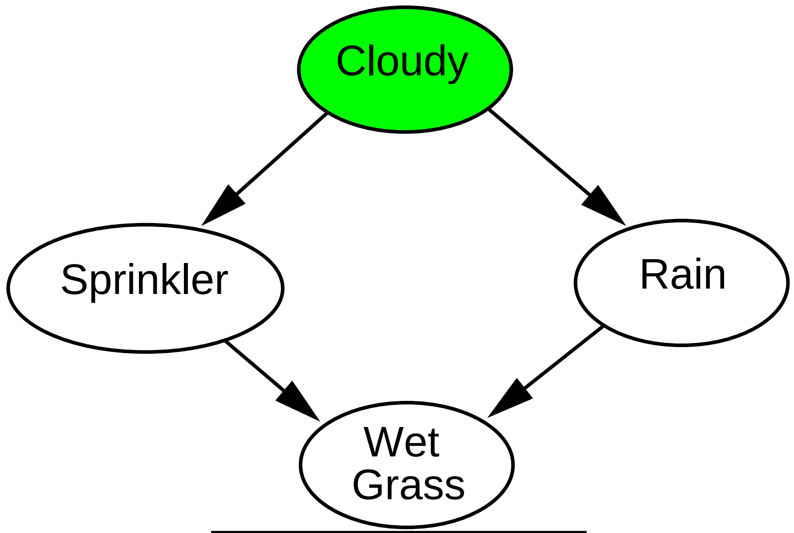


C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

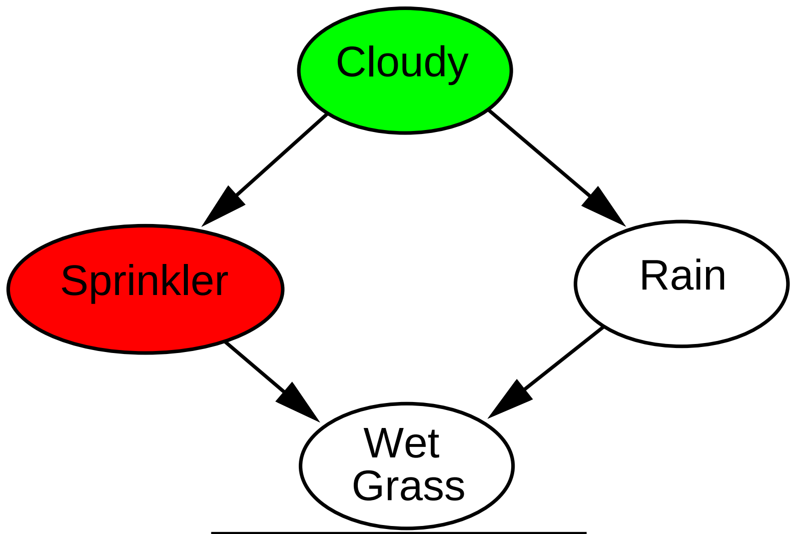
P(C)
.50



C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01



P(C)
.50

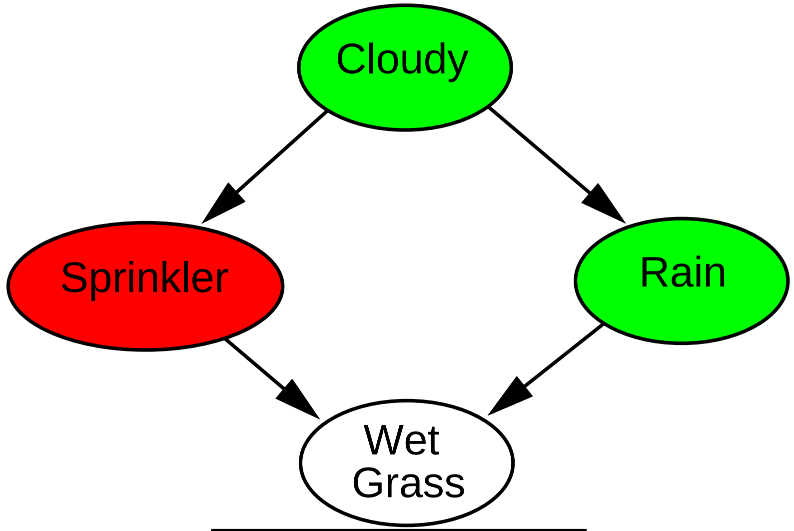
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

P(C)
.50

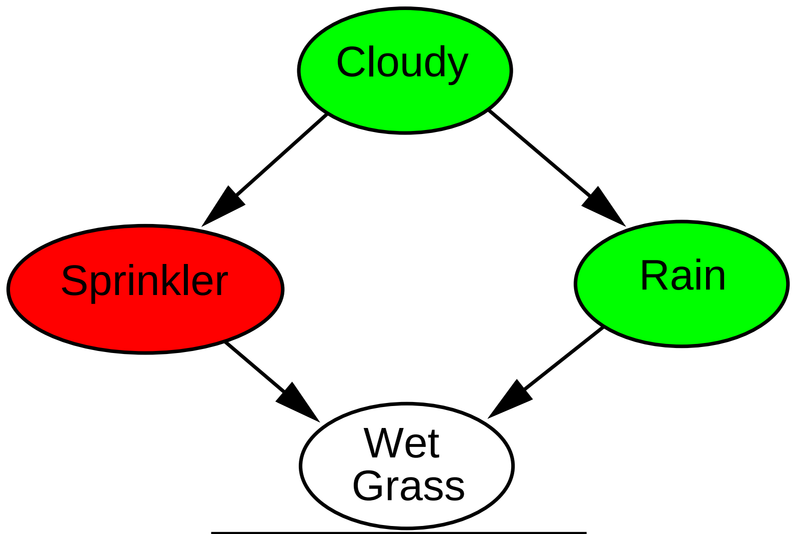
C	P(S|C)
T	.10
F	.50



C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

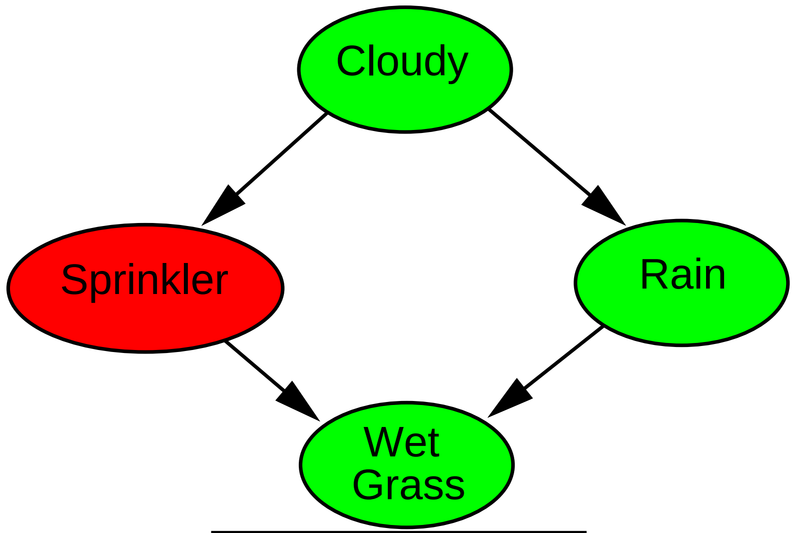
P(C)
.50



C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01



P(C)
.50

C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

- So, this time we get the event

$$[true, false, true, true]$$

- If we repeat the process many times, we can count the number of times $[true, false, true, true]$ is the result.
- The proportion of this to the total number of runs is:

$$\mathbf{P}(c, \neg s, r, w)$$

- The more runs, the more accurate the probability.

- This algorithm:

function PRIOR-SAMPLE(bn) **returns** an event sampled from bn
inputs: bn, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

$\mathbf{x} \leftarrow$ an event with n elements
for $i = 1$ **to** n **do**
$x_i \leftarrow$ a random sample from $\mathbf{P}(X_i \mid parents(X_i))$
given the values of $Parents(X_i)$ in $\mathbf{x}$
return $\mathbf{x}$

captures the *no evidence* case, which is what we just looked at.

- To get values with evidence, we need conditional probabilities

$$\mathbf{P}(X|\mathbf{e})$$

- Could just compute the joint probability and sum out the conditionals but that is inefficient.
- Better is to use *rejection sampling*
  - Sample from the network but reject samples that don't match the evidence.
  - If we want $\mathbf{P}(w|c)$ and our sample picks $\neg c$, we stop that run immediately.
  - For unlikely events, may have to wait a long time to get enough matching samples.
- Still inefficient.

- Likelihood weighting:
  - Version of *importance sampling*.
  - Fix evidence variable to *true*, so just sample relevant events.
  - Have to weight them with the likelihood that they fit the evidence.
  - Use the probabilities we know to weight the samples.

From probability to decision making

- What we have covered allows us to compute probabilities of interesting events.
- But *beliefs* alone are not so interesting to us.
- In the WW don't care so much if there is a pit in (2, 2), so much as we care whether we should go left or right.
- This is complicated because the world is uncertain.
  - Don't know the outcome of actions.
  - Non-deterministic as well as partially observable

DA MAYOR: Mookie.

MOOKIE: Gotta go.

DA MAYOR: C'mere, Doctor.

DA MAYOR: Doctor, this is Da Mayor talkin'.

MOOKIE: OK. OK.

DA MAYOR: Doctor, always try to do the right thing.

MOOKIE: That's it?

DA MAYOR: That's it.

MOOKIE: I got it.

(Spike Lee, Do the Right Thing)

- I offer you the chance to take part in this gamble:
  - $0 one time in one hundred;
  - $1 89 times in one hundred;
  - $5 10 times in one hundred.
- Would you prefer this to $1.00?

- I offer you the chance to take part in this gamble:
  - $0 one time in one hundred;
  - $1 89 times in one hundred;
  - $5 10 times in one hundred.
- Would you prefer this to $1.50?

- I offer you the chance to take part in this gamble:
  - \$0 one time in one hundred;
  - \$1 89 times in one hundred;
  - \$5 10 times in one hundred.
- Would you prefer this to \$1.20?

- I offer you the chance to take part in this gamble:
  - $0 one time in one hundred;
  - $1 89 times in one hundred;
  - $5 10 times in one hundred.
- Would you prefer this to $1.40?

- We can’t make this choice without thinking about how likely outcomes are.
- Although the first option is attractive, it isn’t necessarily the best course of action (especially if the choice is iterated).
- Decision theory gives us a way of analysing this kind of situation.

- Consider being offered a bet in which you pay \$2 if an odd number is rolled on a die, and win \$3 if an even number appears.
- To analyse this prospect we need a *random variable* X, as the function:

$$X : \Omega \mapsto \Re$$

from the sample space to the values of the outcomes. Thus for $\omega \in \Omega$:

$$X(\omega) = \begin{cases} 3, & \text{if } \omega = 2, 4, 6 \\ -2, & \text{if } \omega = 1, 3, 5 \end{cases}$$

- The probability that X takes the value 3 is:

$$\begin{aligned} \Pr(\{2, 4, 6\}) &= \Pr(\{2\}) + \Pr(\{4\}) + \Pr(\{6\}) \\ &= 0.5 \end{aligned}$$

- How do we analyse how much this bet is worth to us?

- To do this, we need to calculate the *expected value* of X.
- This is defined by:

$$E(X) = \sum_k k \Pr(X = k)$$

  where the summation is over all values of k for which $\Pr(X = k) \neq 0$.
- Here the expected value is:

$$E(X) = 0.5 \times 3 + 0.5 \times -2$$

- Thus the expected value of X is \$0.5, and we take this to be the value of the bet.
  - Not the value you will get.

- What is the expected value of this event:
  - \$0 one time in one hundred;
  - \$1 89 times in one hundred;
  - \$5 10 times in one hundred.
- Would you prefer this to \$1?

- And now we can make a first stab at defining what rational action is.
- Rational action is the choice of actions with the greatest expected value for the agent in question.
- The problem is then to decide what "value" is.

Decision theory

- One obvious way to define "value" is in terms of money.
- This has obvious applications in writing programs to trade stocks, or programs to play poker.
- The problem is that the value of a given amount of money to an individual is highly subjective.
- In addition, using monetary values does not take into account an individual's attitude to risk.

- As an example, consider a transaction which offered the following payoffs:
  - $0 one time in one hundred;
  - $1 million 89 times in one hundred;
  - $5 million 10 times in one hundred.
- Would you prefer this to a guaranteed $1 million?

- Utilities are a means of solving the problems with monetary values.
- Utilities are built up from preferences, and preferences are captured by a preference relation $\preceq$ which satisfies:

$$
\begin{array}{c}
a \preceq b \quad \text{or} \quad b \preceq a \\
a \preceq b \;\text{ and }\; b \preceq c \Rightarrow a \preceq c
\end{array}
$$

- You have to be able to state a preference.
- Preferences are transitive.

- A function:

$$u : \Omega \mapsto \Re$$

is a utility function representing a preference relation $\preceq$ if and only if:

$$u(a) \leq u(b) \ \leftrightarrow \ a \preceq b$$

- With additional assumptions on the preference relation (to do with preferences between lotteries) Von Neumann and Morgenstern identified a sub-class of utility functions.

- These “Von Neumann and Morgenstern utility functions”are such that calculating expected utility, and choosing the action with the maximum expected utility is the “best” choice according to the preference relation.
- This is “best” in the sense that any other choice would disagree with the preference order.
- This is why the *maximum expected utility* decision criterion is said to be rational.

- To relate this back to the problem of an agent making a rational choice, consider an agent with a set of possible actions A available to it.
- Each $a \in A$ has a sample space Ω_a associated with it, and a set of possible outcomes s_a where $s_a \subseteq S_a$ and $S_a = 2^{\Omega_a}$.
- (This is a simplification since each s_a will usually be conditional on the state of the environment the agent is in.)

- The action a^* which a rational agent should choose is that which maximises the agent's utility.
- In other words the agent should pick:

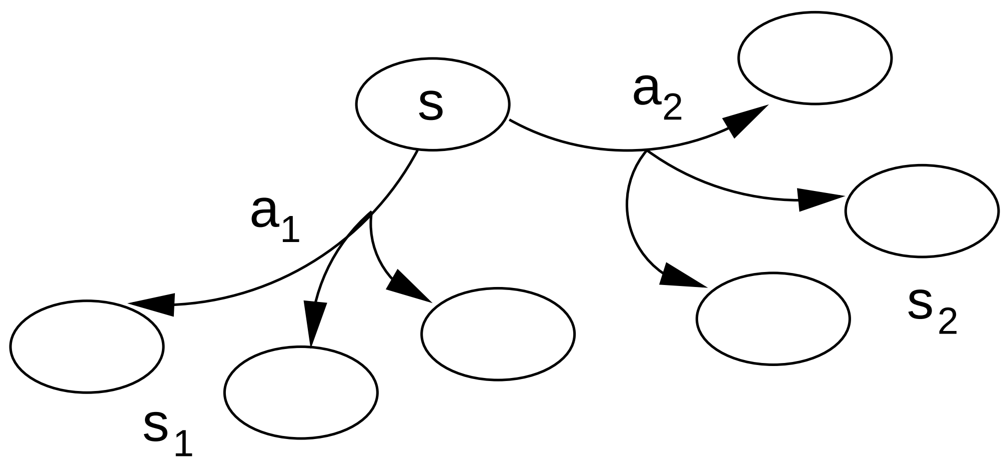
$$a^* = arg \max_{a \in A} u(s_a)$$

- The problem is that in any realistic situation, we don't know which s_a will result from a given a, so we don't know the utility of a given action.
- Instead we have to calculate the expected utility of each action and make the choice on the basis of that.

- In other words, for the set of outcomes s_a of each action each a, the agent should calculate:

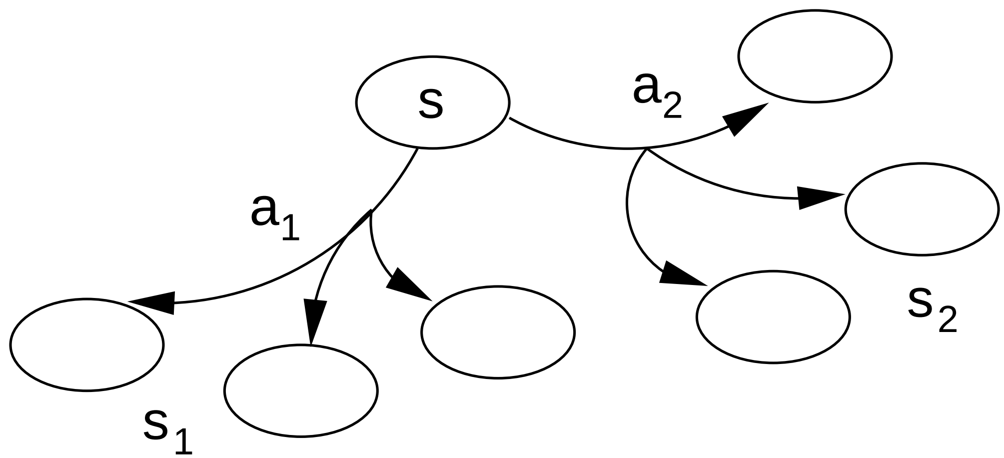
$$E(u(s_a)) = \sum_{s' \in s_a} u(s'). \Pr(s_a = s')$$

and pick the best.



- Thus to be rational, an agent needs to choose a^* such that:

$$a^* = arg \max_{a \in A} \sum_{s' \in s_a} u(s'). \Pr(s_a = s')$$



- As an example, consider an agent which has to choose between tossing a coin, rolling a die, or receiving a payoff of \$ 1.
- If the coin is chosen, then the agent gets \$1.50 a head and \$0.5 for a tail.
- If the die is chosen, the agent gets \$5 if a six is rolled, \$1 if a two or three is rolled, and nothing otherwise.
- What is the rational choice, assuming that the agent's preferences are (for once) modelled by monetary value?

- Well, we need to calculate the expected outcome of each choice.
- For doing nothing, we have a_1 = "receive payoff", s_{a_1} = {"get \$1"}, $u(\text{“get \$1})=1$ and $\Pr(s_{a_1} = \text{“get \$1}) = 1$.
- Thus:

$$E(u(s_{a_1})) = 1$$

- If the coin is chosen, we have a_2 = "coin", $s_{a_2} = \{\text{head}, \text{tail}\}$,

$$\begin{aligned} u(\text{head}) &= \$1.50 \\ u(\text{tail}) &= \$0.5 \end{aligned}$$

and

$$\begin{aligned} \Pr(s_{a_2} &= \text{head}) = 0.5 \\ \Pr(s_{a_2} &= \text{tail}) = 0.5 \end{aligned}$$

- Thus the expected utility is:

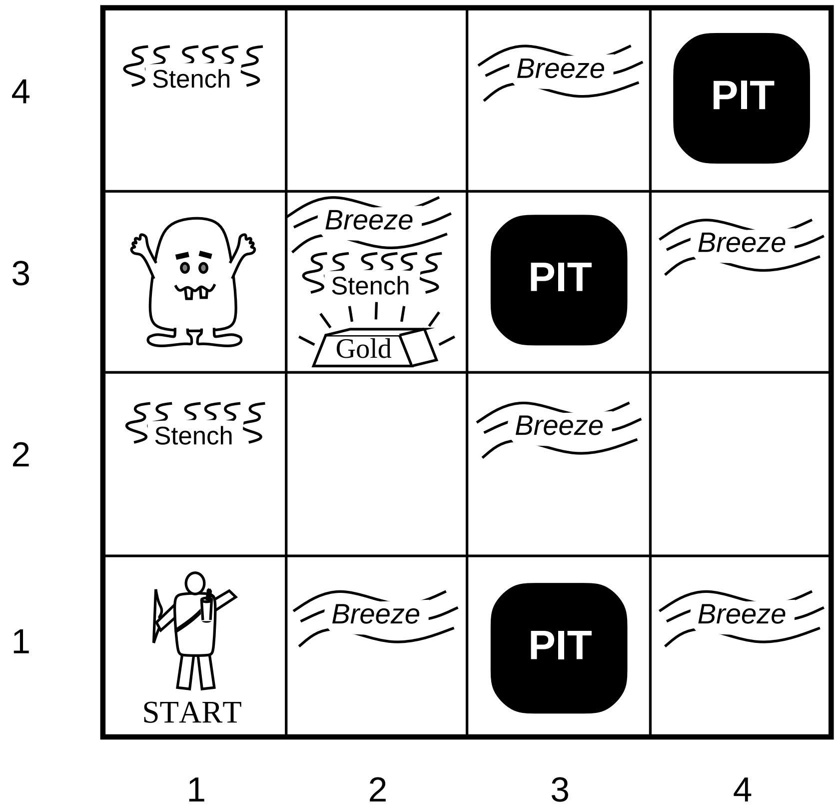
$$\begin{aligned} E(u(s_{a_2})) &= 0.5 \times 1.5 + 0.5 \times 0.5 \\ &= 1 \end{aligned}$$

- Action a_3, rolling the die, can be analysed in a similar way, giving:

$$E(u(s_{a_3})) = 1.17$$

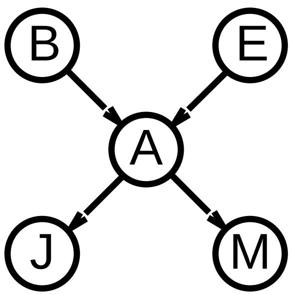
- Choosing to roll the die is the rational choice.

Decisions in the WW



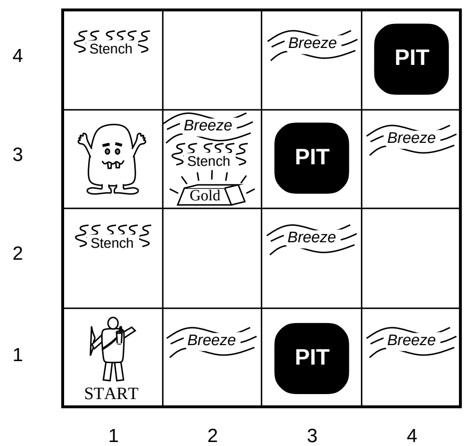
- Actions have a range of outcomes.
- Forward has some probability of moving sideways
  - Not so silly with a robot
- Probabilities across action outcomes.
  - Given an action, probability of getting to some states
- Utilities for states.

- Given what we know about Bayesian networks, we can clearly deal with complex situations as far as probability is concerned.



- Should I go home given that John calls and Mary doesn't?

- But what about more complex decisions?



- What is the best sequence of actions to carry out to get the gold?

Summary

- This lecture looked at dealing with uncertainty.
  - Non-deterministic environments
- It looked at handling this uncertainty using probability and then went on to look at how decision theory allows us to decide what to do.
- We now know how to make a decision about the best action to carry out.
  - But we can only choose one action at a time.
- Next time we'll look at sequential decision problems.