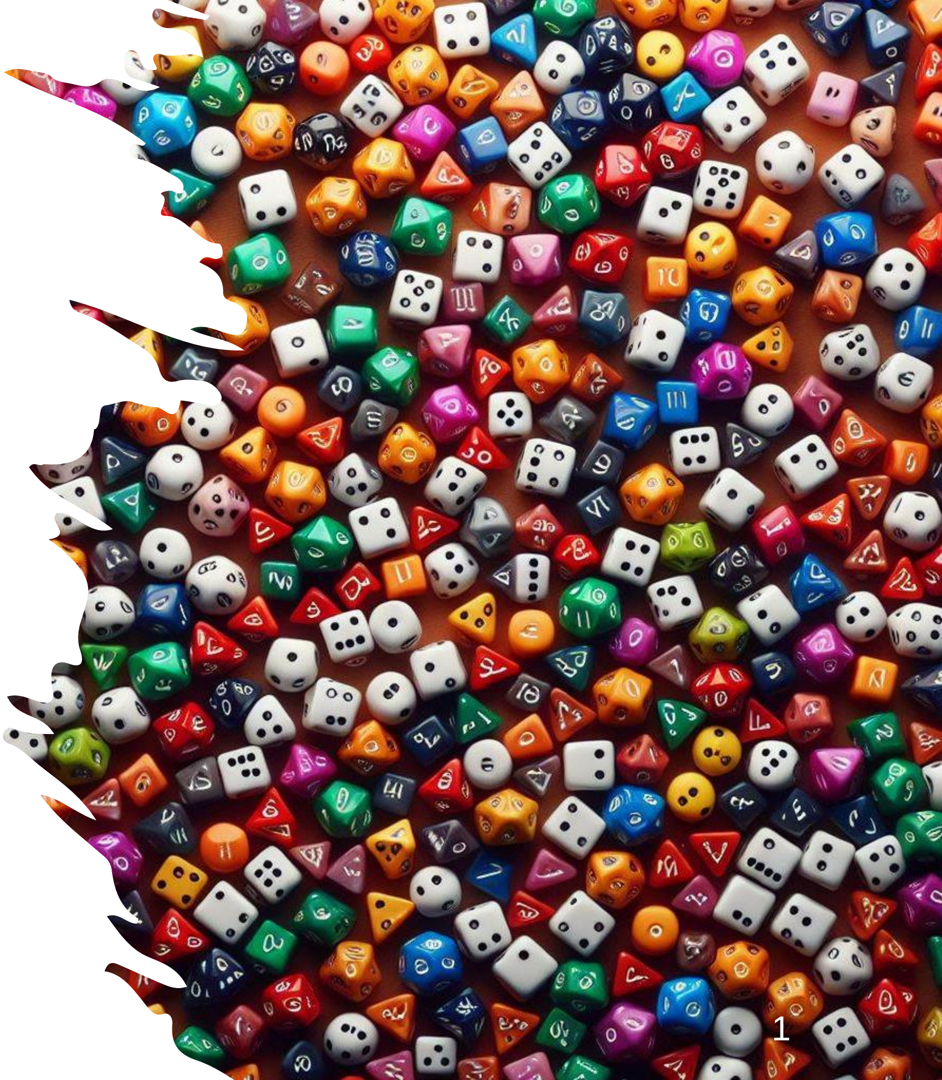
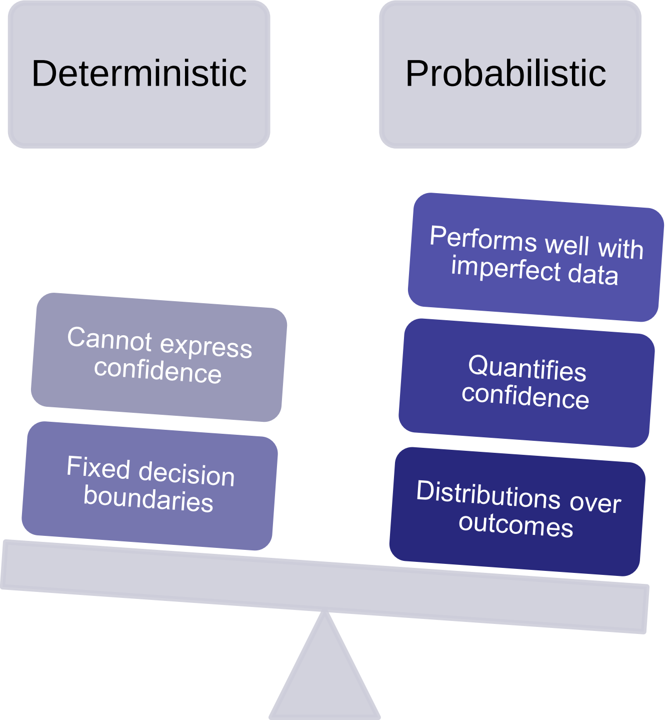
Handling Uncertainty

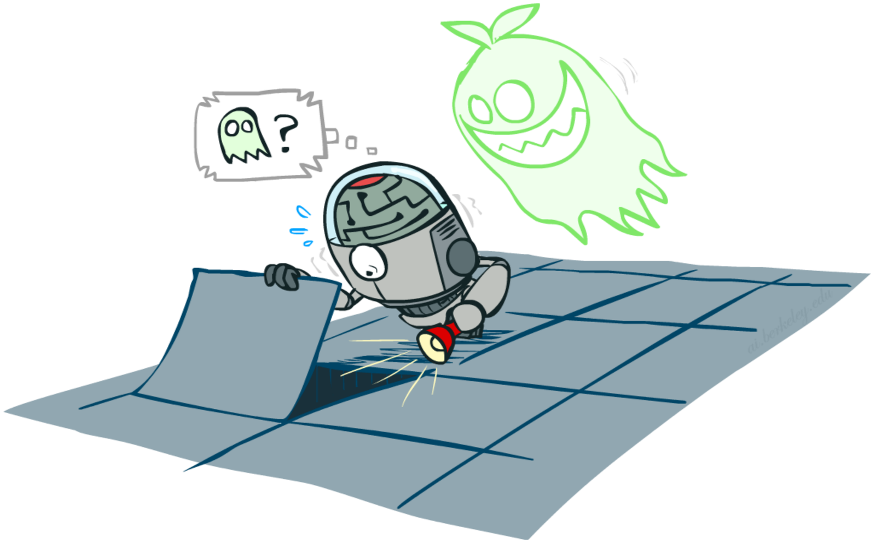
- Probability
- Probabilistic Inference
- Bayes Rule
- Conditional Independence
- Bayesian Networks
- Other approaches



These slides were adapted from slides created by Dan Klein and Pieter Abbeel forCS188 Intro to AI at UC Berkeley.

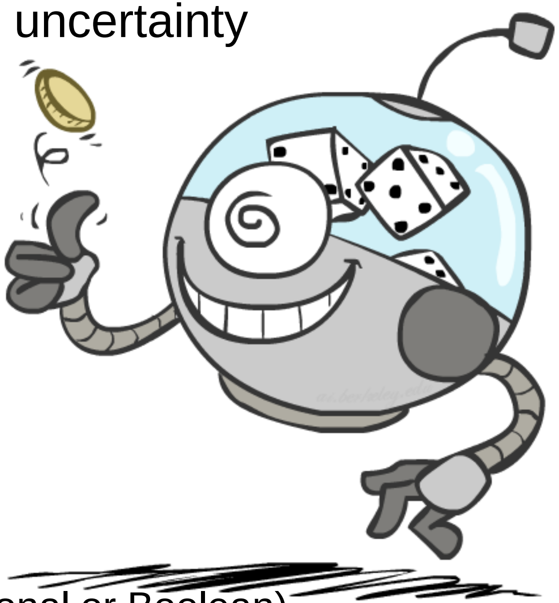
Deterministic vs Probabilistic





Random Variables

- Some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
  - L = Where is the ghost?



- We denote random variables with capital letters

- Random variables have domains
  - R in {true, false} (often write as {+r, -r}) or {r, $\neg$r} (Propositional or Boolean)
  - T in {hot, cold} (Discrete)
  - D in $[0, \infty)$ (Continuous)
  - L in possible locations, maybe {(0,0), (0,1), …}

Probability Distributions

- Probability distribution gives values for all possible assignments:

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
cloud	0.3
snow	0.0

  – Distribution is a table P(Weather) = ⟨0.6,0.1,0.3,0.0⟩

    Probability is a single value $P(W = rain) = 0.1$

Joint Distributions

- Joint probability distribution for a set of random variables $X_1, X_2, \ldots X_n$

- gives the probability of every atomic event on those values

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Events

- An *event* is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?

- Typically, the events we care about are *partial assignments*, like P(T=hot)

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Your turn: Events

- P(+x, +y) ?

- P(+x) ?

- P(-y OR +x) ?

- P(-y IF +x) ?

$$P(X, Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_{w} P(t,w)$$

$$P(w) = \sum_{t} P(t,w)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probability

- Definition $$P(a|b) = \frac{P(a,b)}{P(b)}$$
- Product Rule $$P(a,b) = P(a|b)P(b) \qquad P(b,a) = P(b|a)P(a)$$
- Bayes Rule $$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

P(hypothesis|data) = P(data|hypothesis) × P(hypothesis) / P(data)

Posterior	Likelihood × Prior	Evidence
Updated belief	Model × Initial belief	Normalization factor

Conditional Probabilities

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = s | T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$\boxed{\begin{aligned} &= P(W = s, T = c) + P(W = r, T = c) \\ &= 0.2 + 0.3 \quad = 0.5 \end{aligned}}$$

$$P(a|b) = \frac{P(a, b)}{P(b)}$$

Your turn: Conditional Probabilities

$$P(X, Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

- P(+x | +y) ?
- P(-x | +y) ?
- P(-y | +x) ?

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$$P(W|T)$$

$$P(W|T = hot)$$

W	P
sun	0.8
rain	0.2

$$P(W|T = cold)$$

W	P
sun	0.4
rain	0.6

Joint Distribution

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Normalization Trick

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$\begin{aligned} P(W=s|T=c) &= \frac{P(W=s,T=c)}{P(T=c)} \\ &= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\ &= \frac{0.2}{0.2+0.3} = 0.4 \end{aligned}$$

$$\begin{aligned} P(W=r|T=c) &= \frac{P(W=r,T=c)}{P(T=c)} \\ &= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\ &= \frac{0.3}{0.2+0.3} = 0.6 \end{aligned}$$

$$P(W|T=c)$$

W	P
sun	0.4
rain	0.6

Normalization Trick

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)}$$
$$= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)}$$
$$= \frac{0.2}{0.2+0.3} = 0.4$$

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

SELECT the joint probabilities matching the evidence

→

$P(c,W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection (make it sum to one)

→

$P(W|T=c)$

W	P
sun	0.4
rain	0.6

$$P(W=r|T=c) = \frac{P(W=r,T=c)}{P(T=c)}$$
$$= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)}$$
$$= \frac{0.3}{0.2+0.3} = 0.6$$

Your turn: Normalization Trick

- P(X | Y=-y) ?

$$P(X, Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

SELECT the joint probabilities matching the evidence

NORMALIZE the selection (make it sum to one)

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

- We generally compute conditional probabilities
  - P(on time | no reported accidents) = 0.90
  - These represent the agent's *beliefs* given the evidence

- Probabilities change with new evidence:
  - P(on time | no accidents, 5 a.m.) = 0.95
  - P(on time | no accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes *beliefs to be updated*

Inference by Enumeration

- General case:
  - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
  - Query* variable: Q
  - Hidden variables: $H_1 \dots H_r$

$\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} X_1, X_2, \dots X_n$ *All variables*

- We want:

** Works fine with multiple query variables, too*

$$P(Q|e_1 \dots e_k)$$

- Step 1: Select the entries consistent with the evidence
- Step 2: Sum out H to get joint of Query and evidence
- Step 3: Normalize

$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} P(\underbrace{Q, h_1 \dots h_r, e_1 \dots e_k}_{X_1, X_2, \dots X_n})$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

Inference by Enumeration

- P(W)?

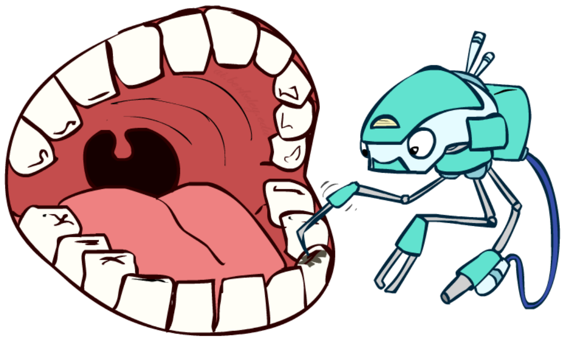
- P(W | winter)?

- P(W | winter, hot)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Probabilistic Inference

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576



$$P(cavity|toothache) = \frac{P(cavity, toothache)}{P(toothache)}$$

P(Cavity|toothache)=αP(Cavity,toothache)

=α[P(Cavity,toothache,catch)+P(Cavity,toothache,¬catch)]

α is the normalization constant

Catch is the hidden variable

The Chain Rule

- More generally, we can always write any joint distribution as an incremental product of conditional distributions (successive applications of the product rule)

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$

P(Toothache,Catch,Cavity)

=P(Toothache|Catch,Cavity) P(Catch,Cavity)

=P(Toothache|Catch,Cavity) P(Catch|Cavity)P(Cavity)

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later (e.g. ASR, MT)

Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\} \text{Example givens}$$

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

  - Note: Probability of meningitis still very small
  - Note: you should still get stiff necks checked out!

Your turn: Bayes' Rule

- Given:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

- What is P(W | dry) ?

Independence

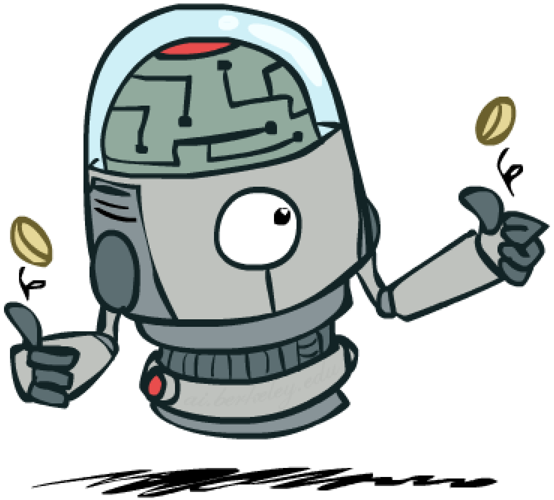
- Two variables are *independent* if:

$$\forall x, y : P(x, y) = P(x)P(y)$$

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form:

$$\forall x, y : P(x|y) = P(x)$$

  - We write: $X \perp\!\!\!\perp Y$



- Independence is a simplifying *modeling assumption*
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?

Conditional Independence

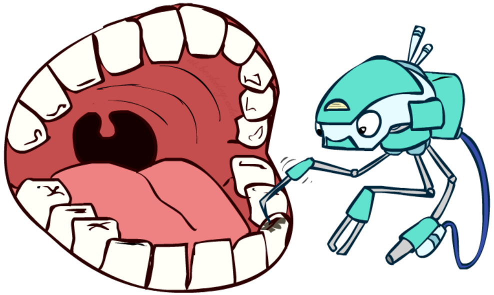
- P(Toothache, Cavity, Catch)

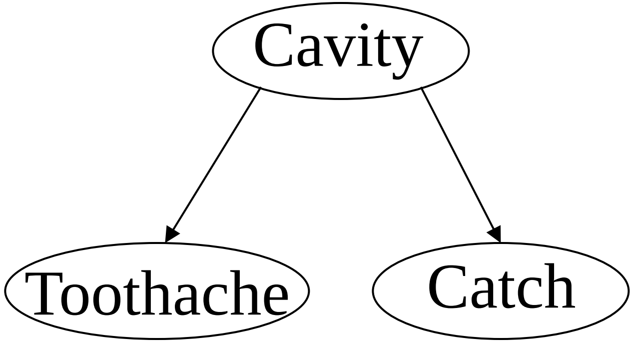
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(catch | toothache, cavity) = P(catch | cavity)

- The same independence holds if I don't have a cavity:
  - P(catch | toothache, $\neg$cavity) = P(catch| $\neg$cavity)

- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)

- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch |Cavity)
  - One can be derived from the other easily





Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z

$$X \perp\!\!\!\perp Y | Z$$

if and only if:

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

$$P(x|z,y) = \frac{P(x,z,y)}{P(z,y)}$$

$$= \frac{P(x,y|z)P(z)}{P(y|z)P(z)}$$

$$= \frac{P(x|z)P(y|z)P(z)}{P(y|z)P(z)}$$

Conditional Independence and the Chain Rule

- Chain rule: $P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$

- Trivial decomposition:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{Traffic})$$

- With assumption of conditional independence:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$

- Bayes' nets / graphical models help us express conditional independence assumptions

Bayes' Rule and Conditional Independence

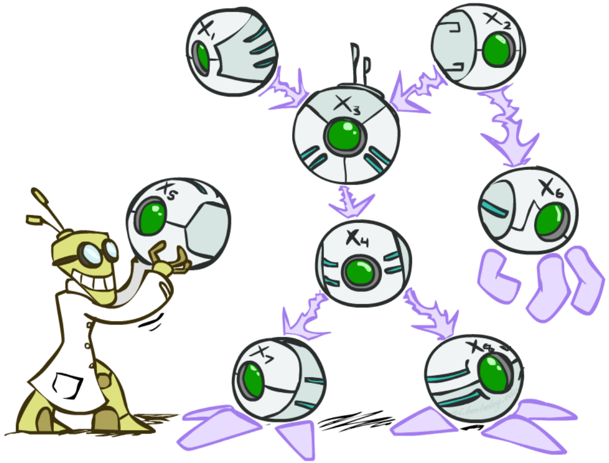
- $P(Cavity|toothache \wedge catch)$
  $= \alpha P(toothache \wedge catch|Cavity)P(Cavity)$
  $= \alpha P(toothache|Cavity)P(catch|Cavity)P(Cavity)$

- This is an example of a naïve Bayes model
- $P(Cause, Effect_1, Effect_2, \ldots Effect_n) = P(Cause) \Pi_i P(Effect_i|Cause)$

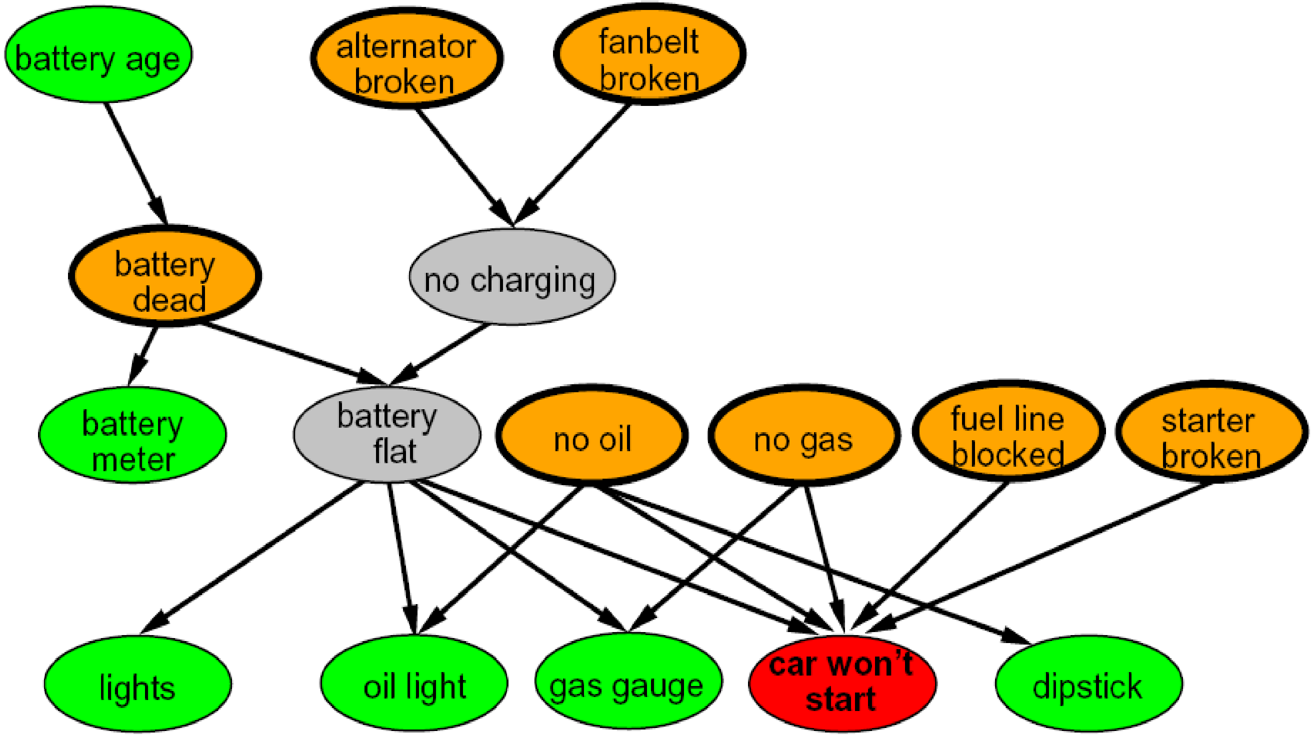
Bayes' Nets: Big Picture

- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

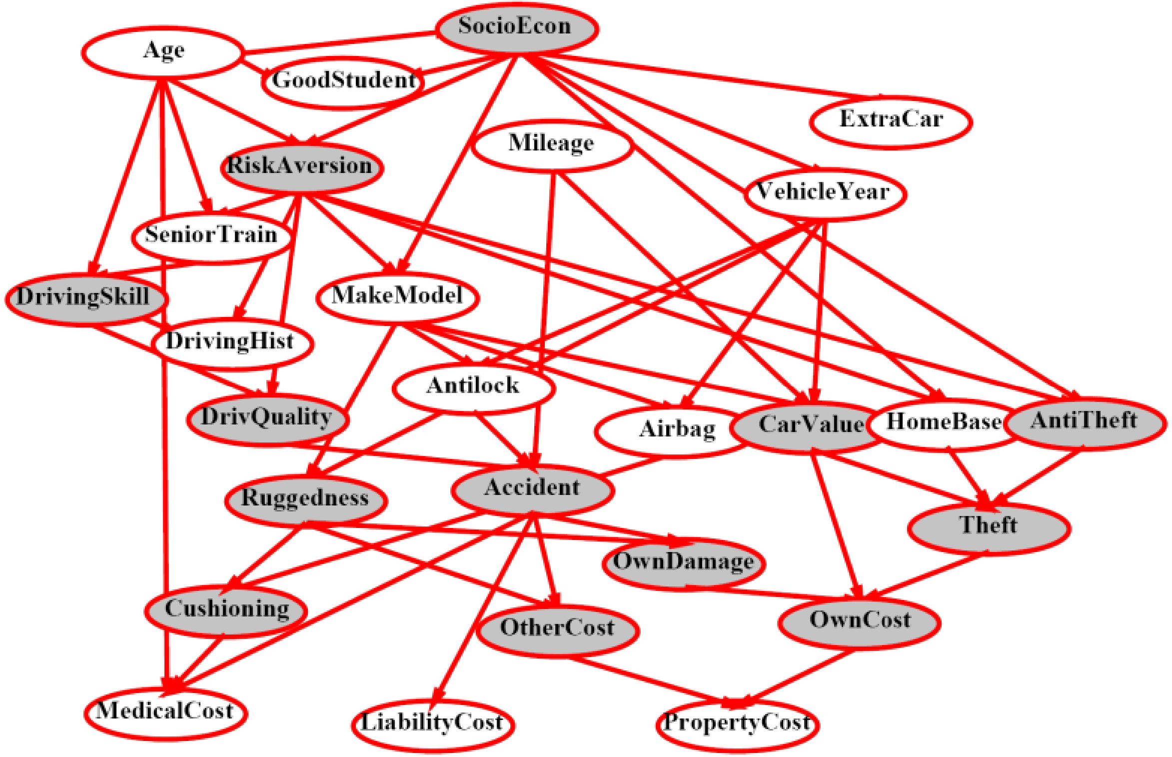
- Bayes' nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models



Example Bayes' Net: Car

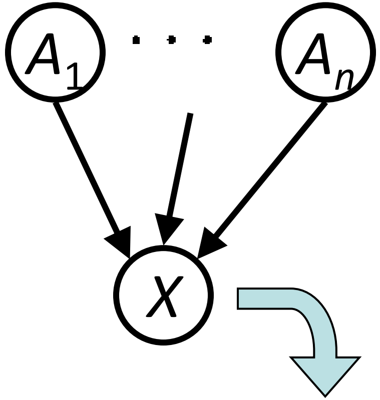


Example Bayes' Net: Insurance



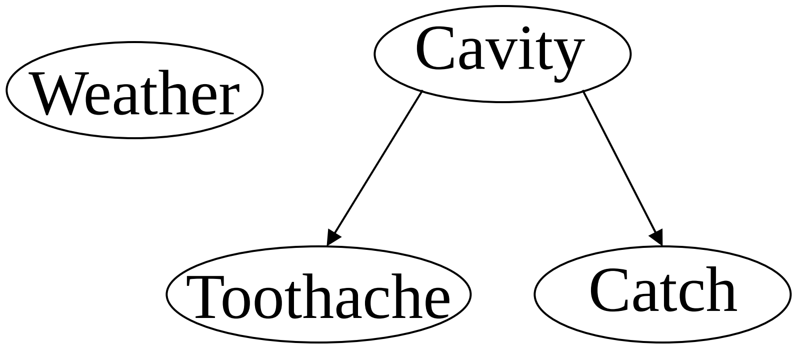
Bayes' Nets Semantics

- 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
  - A conditional distribution for each node given its parents:
    - $P(X_i \mid Parents(X_i))$
    - Conditional distribution could be represented as Conditional Probability Table



$$P(X|A_1 \ldots A_n)$$

Example

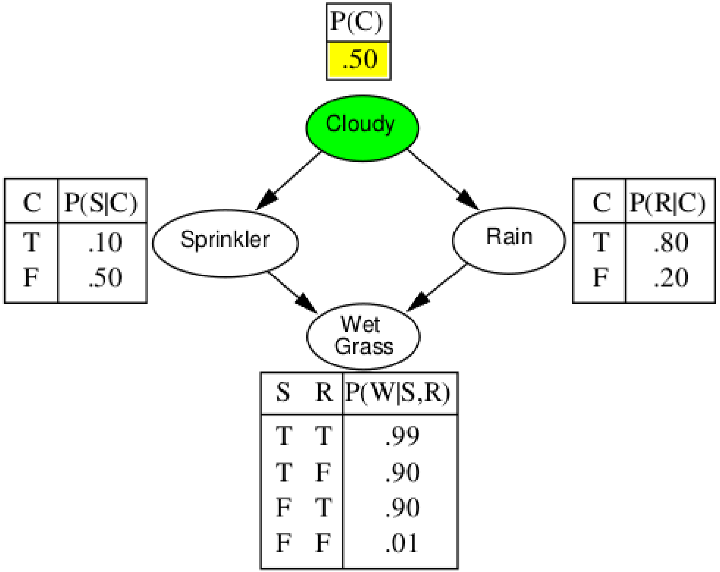


- Topology of the network encodes conditional independence assertions
- Weather is independent of the other variables
- Toothache and Catch are conditionally independent given Cavity

$P(+cavity, +catch, -toothache)$

=P(-toothache|+cavity)P(+catch|+cavity)P(+cavity)

Example



P(C)
.50

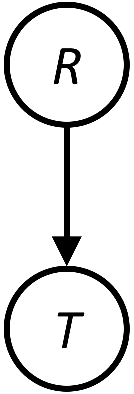
C	P(S|C)
T	.10
F	.50

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

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

Example: Traffic



$P(R)$

+r	1/4
-r	3/4

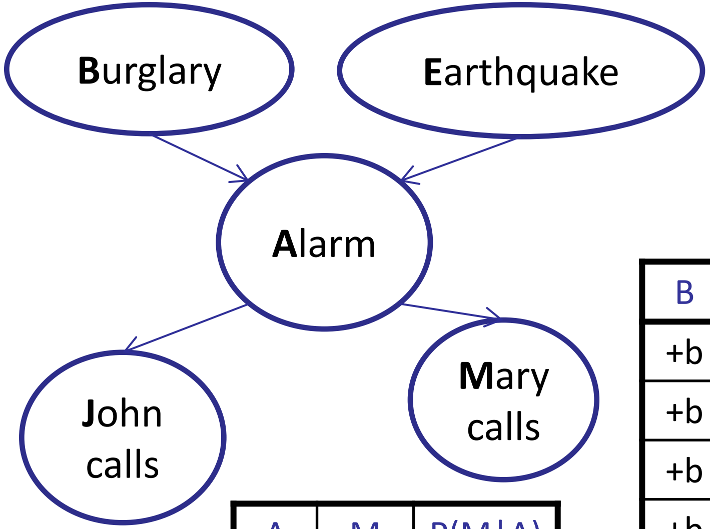
$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(+r, -t) =$ P(+r)P(-t|+r) = 1/4*1/4

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

$P(M|A)P(J|A)P(A|B,E)$

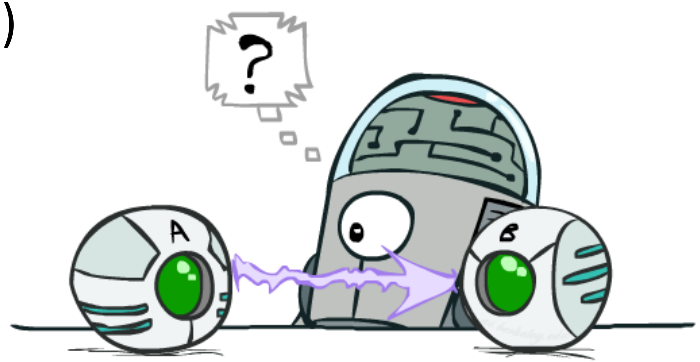
A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

Causality?

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts

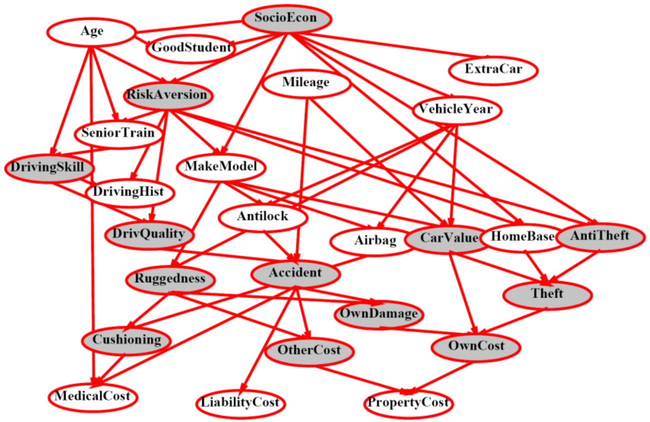


- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - End up with arrows that reflect correlation, not causation

- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence

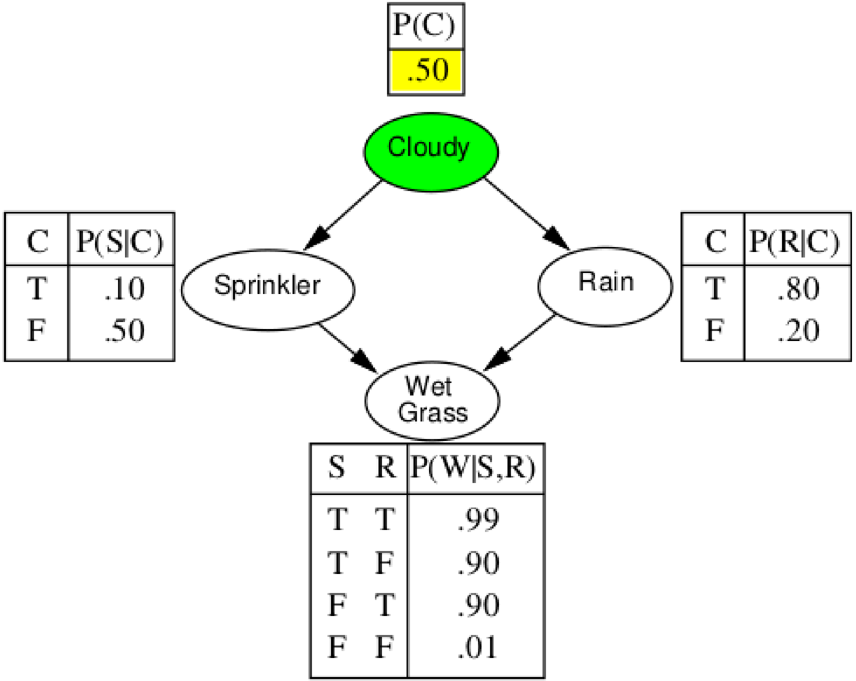
Why Bayes' Nets?

- A Bayes net is an efficient encoding of a probabilistic model of a domain



- Questions we can ask:
  - Inference: given a fixed BN, what is P(X | e)?
  - Representation: given a BN graph, what kinds of distributions can it encode?
  - Modeling: what BN is most appropriate for a given domain?

Lawn Example



Exact Inference

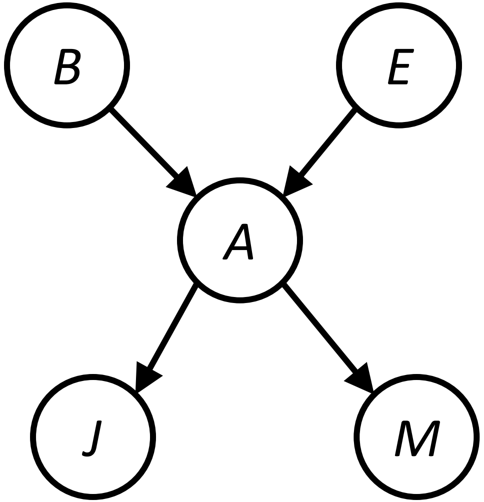
- We observe the grass is wet and wish to know the cause
- There are two possible causes
  - Raining or Sprinkler is on

$$\mathrm{P(S = t|W = t)} = \frac{\sum_{\mathrm{c,r}} \mathrm{P(C = c, S = t, R = r, W = t)}}{P(W = t)} = 0.2781/0.64$$

$$\mathrm{P(R = t|W = t)} = \frac{\sum_{\mathrm{c,s}} \mathrm{P(C = c, S = s, R = t, W = t)}}{P(W = t)} = 0.4581/0.64$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

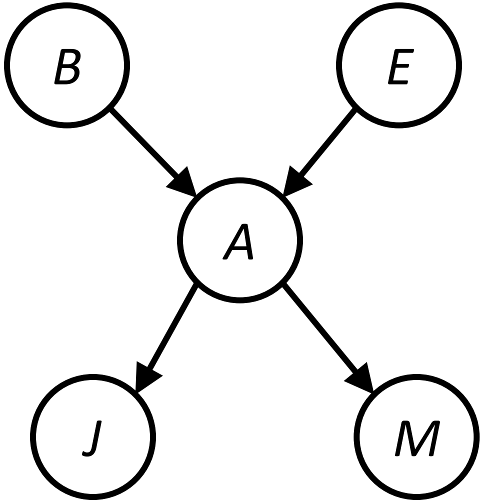
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(+b, -e, +a, -j, +m) =$$
$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

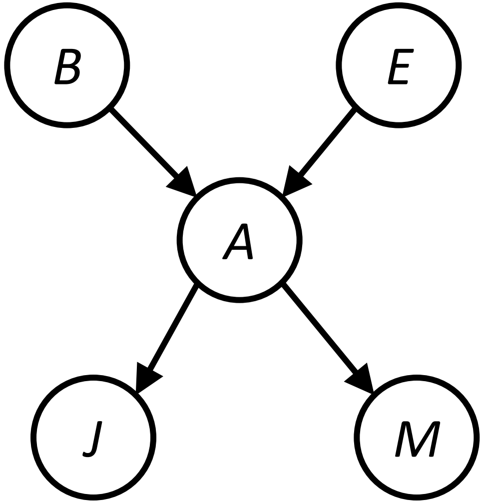
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(b \mid j,m) = P(b,j,m) / P(j,m) = \sum_{ea} P(b,e,a,j,m) / P(j,m)$$

Inference by Enumeration in Bayes' Net

- Given unlimited time, inference in BNs is easy
- Reminder of inference by enumeration by example:



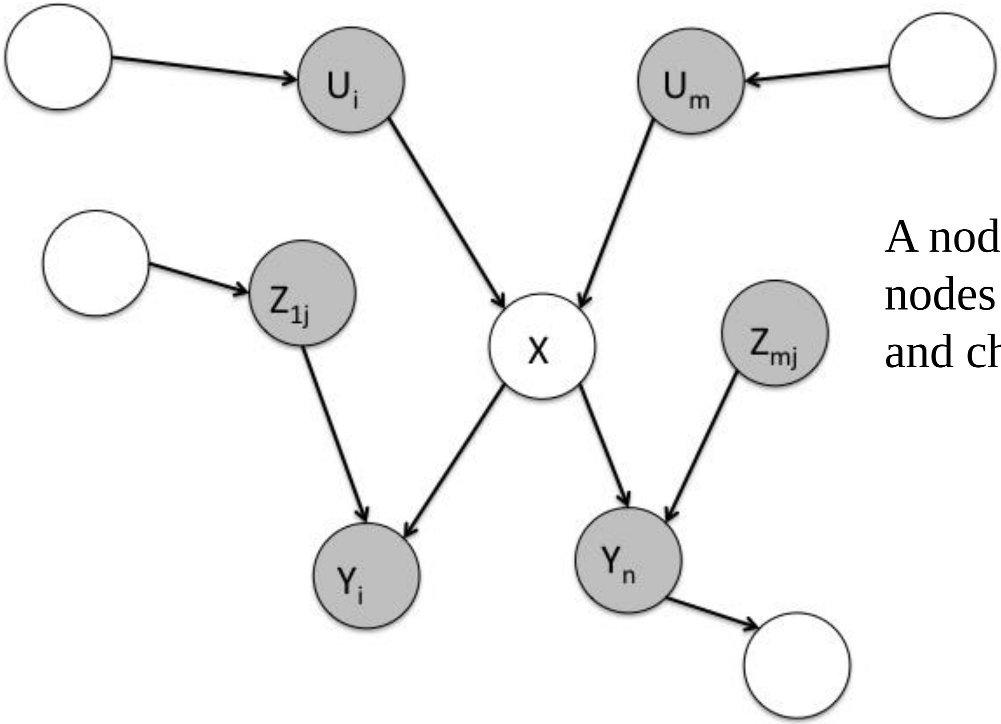
$$P(B \mid +j, +m) \;\; \propto_B P(B, +j, +m)$$

$$= \sum_{e,a} P(B, e, a, +j, +m)$$

$$= \sum_{e,a} P(B)P(e)P(a|B,e)P(+j|a)P(+m|a)$$

$$= P(B)P(+e)P(+a|B,+e)P(+j|+a)P(+m|+a) + P(B)P(+e)P(-a|B,+e)P(+j|-a)P(+m|-a)$$
$$P(B)P(-e)P(+a|B,-e)P(+j|+a)P(+m|+a) + P(B)P(-e)P(-a|B,-e)P(+j|-a)P(+m|-a)$$

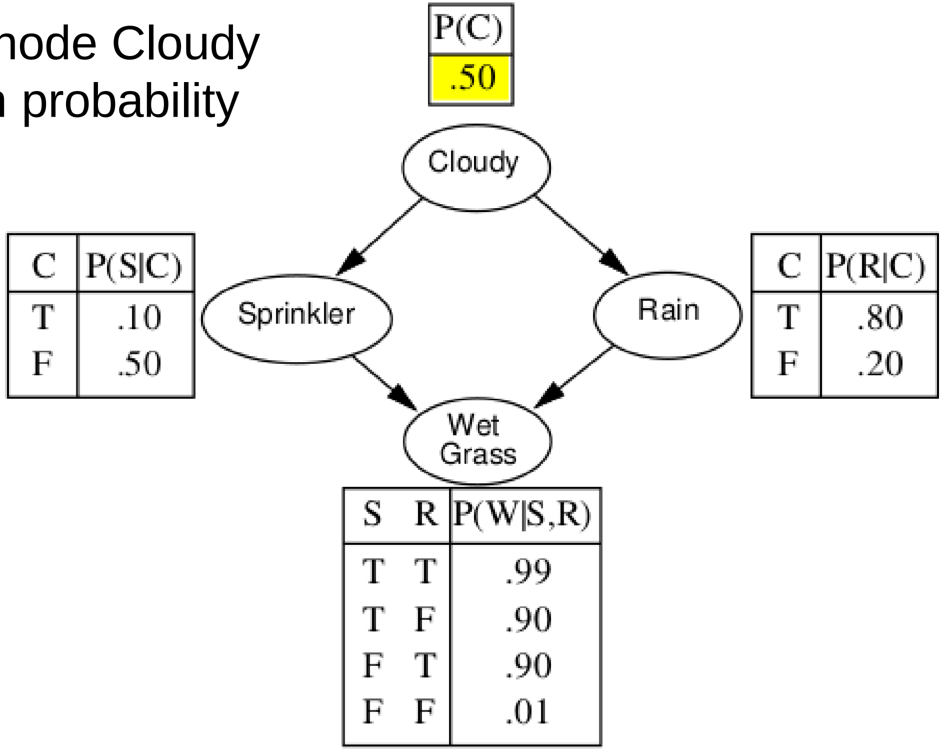
Markov Blanket



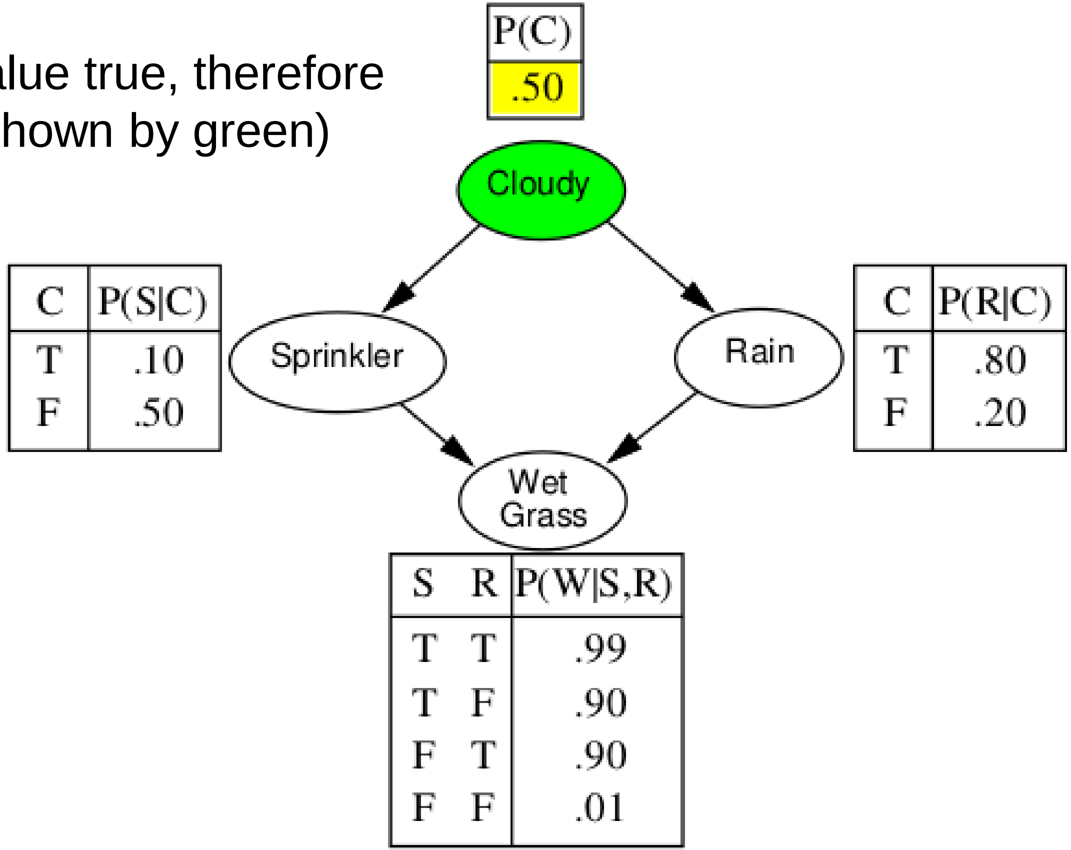
A node is conditionally independent of all other nodes in a network, given its parents, children and children's parents

Approximate Inference (Direct Sampling)

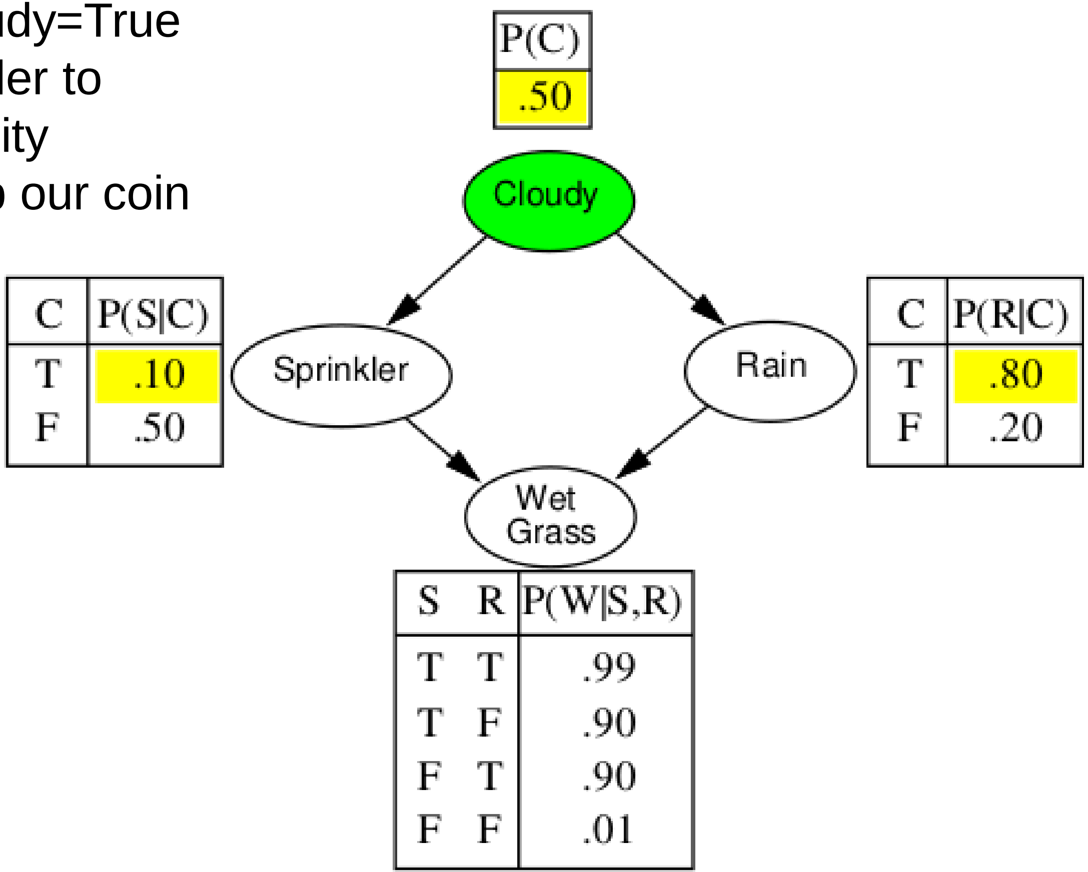
We start at the root node Cloudy and flip our coin with probability P(True)= 0.5



We obtain the value true, therefore
Cloudy = True (shown by green)

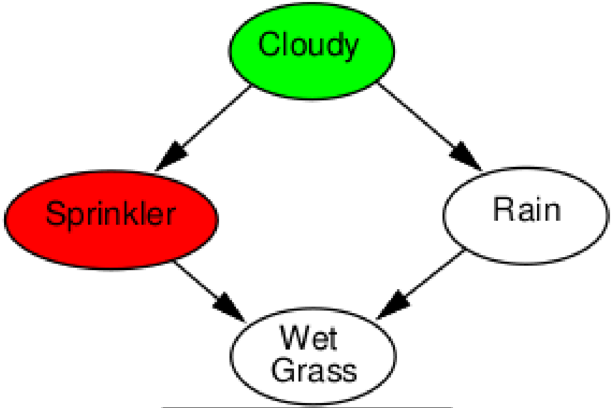


We use the value Cloudy=True and the CPT of Sprinkler to determine the probability Sprinkler=True and flip our coin with this probability



We obtain the value False, therefore
Sprinkler = False (shown by red)

P(C)
.50

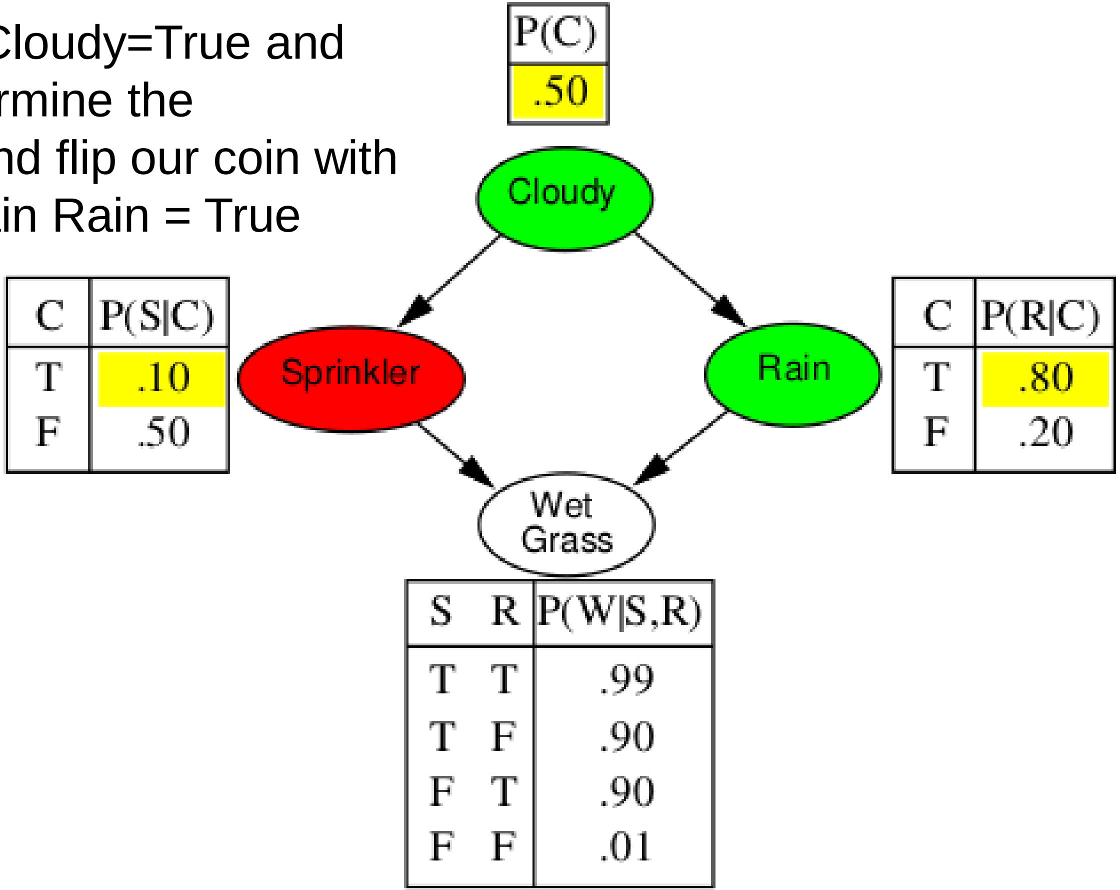


C	P(S|C)
T	.10
F	.50

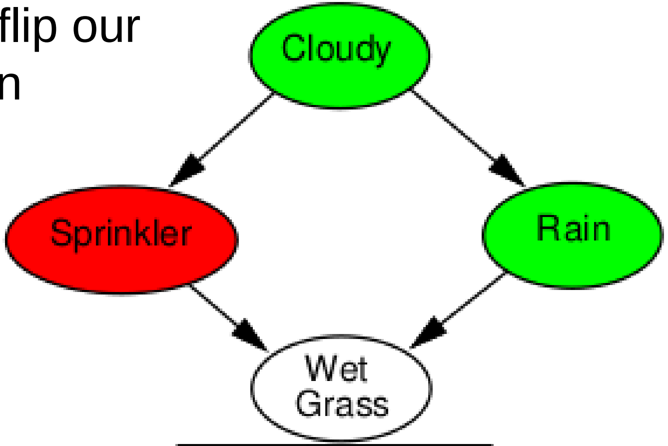
C	P(R|C)
T	.80
F	.20

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

We also use the value Cloudy=True and the CPT of Rain to determine the probability Rain=True and flip our coin with this probability. We obtain Rain = True (Green)



We can now use Sprinkler = False and Rain = True to look up the probability that Wet Grass is true from its CPT and flip our coin using this probability distribution



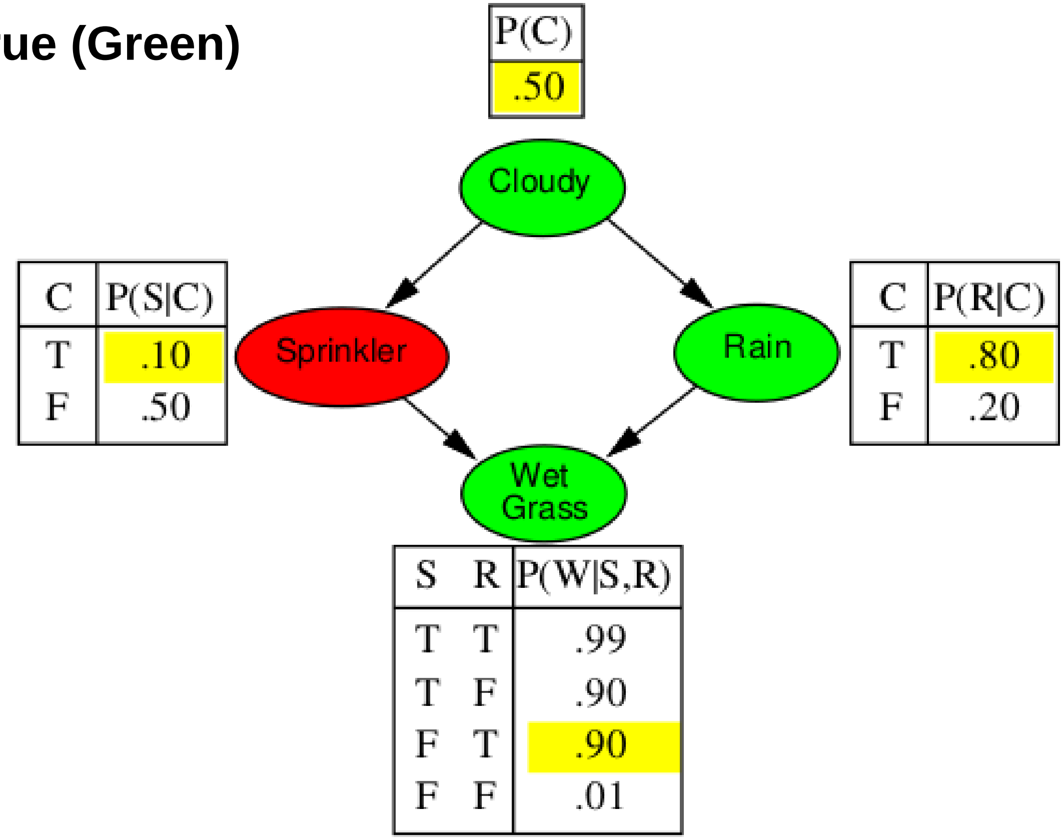
P(C)
.50

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

We obtain Wet Grass = True (Green)



Direct Sampling

- We can then repeat this process over and over again to get as many samples as we would like.
- Using the samples you can then answer queries by counting the number of samples you have which meet your query.

- If I wanted to know P(W=t,R=t,C=t) I can count the samples where both R and C and W are true and normalise this by the number of samples.
- The more samples you have the more confidence you will have in your estimate example.

Rejection Sampling

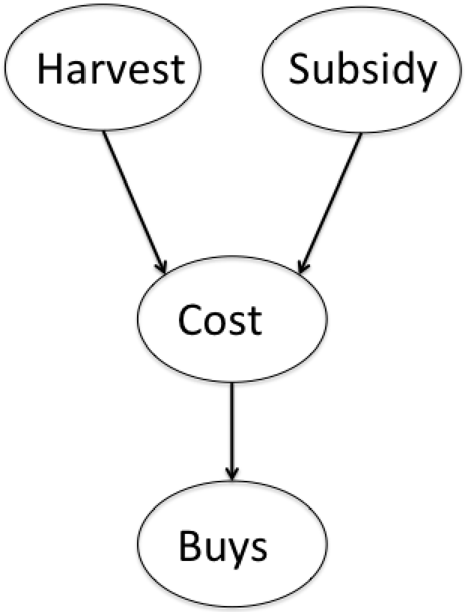
To estimate posterior probabilities P(R =t | S =t) we generate 100 events using direct sampling and observe that only 27 events have S=True.

This means that we need to reject the other 73 events (for which we have S=f). Of the 27 samples with S=t, 8 have R and 19 have R=f, therefore:

$$\hat{P}(R / S) = 8 / 27$$

The true answer is 0.3 and as more samples are collected the estimate will converge to the true answer.

Continuous Variables and Hybrid Networks



Customers choses whether or not to buy some fruit (Buys) depending on its cost (*Cost*). The cost depends on the yield of the harvest (*Harvest*) and whether or not a government subsidy (*Subsidy*) has been provided.

The variables *Subsidy* and *Buys* are discrete however *Cost* and *Harvest* are continuous. How do we compute a CPT for a continuous variable?

Russell & Norvig AI: A Modern Approach p520

Handling the discrete Parent

- For the discrete parent *subsidy* we can specify both:
  - P(Cost|Harvest,+subsidy) and
  - P(Cost|Harvest,-subsidy)
- defining a **Conditional Gaussian**.

Handling the continuous parent

- For the continuous parent *Harvest* we specify how the distribution of *Cost* depends on the value of the *Harvest*.
- The parameters of the *Cost* probability density function are therefore a function of the value of the parent

$$P(c \mid h, subsidy) = \mathcal{N}(c; a_t h + b_t, \sigma_t^2) = \frac{1}{\sigma_t \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2}$$

$$P(c \mid h, \neg subsidy) = \mathcal{N}(c; a_f h + b_f, \sigma_f^2) = \frac{1}{\sigma_f \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{c-(a_f h + b_f)}{\sigma_f}\right)^2}$$

Conditional distributions

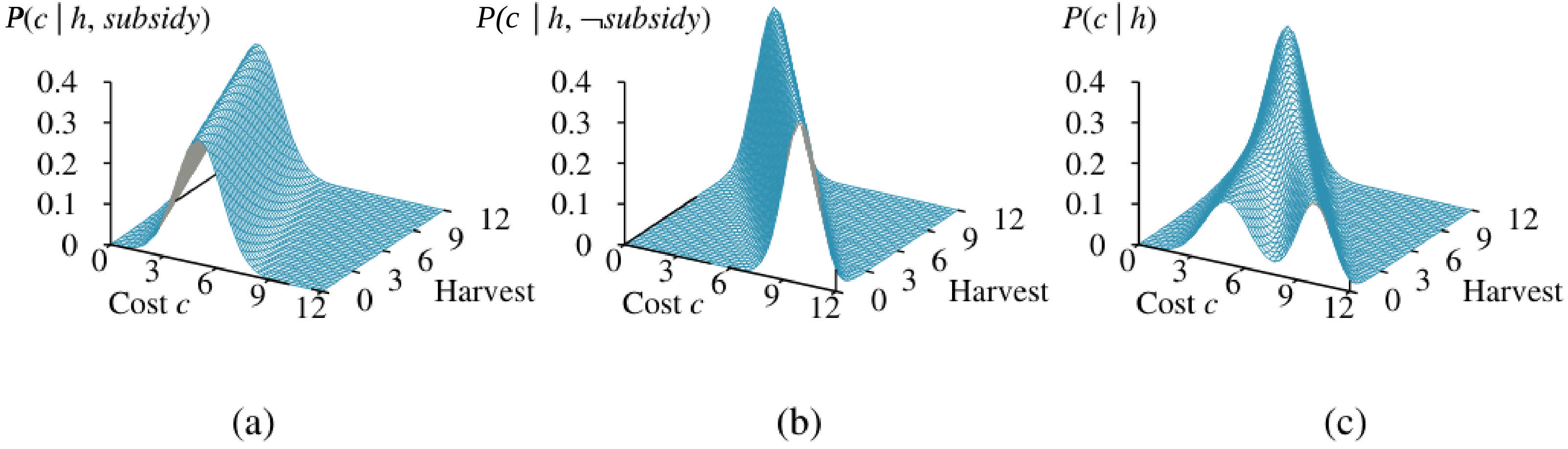


Figure 13.7 The graphs in (a) and (b) show the probability distribution over *Cost* as a function of *Harvest* size, with *Subsidy* true and false, respectively. Graph (c) shows the distribution $P(Cost \mid Harvest)$, obtained by summing over the two subsidy cases.

Handling discrete node with continuous parent.

We require a function that will be set:

Buys=true if Cost is low and
Buys=false if Cost is high

A common approach to this is to use a **Sigmoid** distribution

$$P(Buys = false|Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$