Daniel Kahneman

Nobel Memorial Prize in Economic Sciences 2002

For having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty

Presidential Medal of Freedom 2013

Judea Pearl

ACM A.M. Turing Award 2011

For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

AI101

Lecture 10: Bayesian networks

Recap
Uncertainty

Agents must deal with uncertainties. Typically this is done using probability theory.

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

Today

- Bayesian Networks
- Inference in Bayesian Networks

Motivation: Uncertainty in AI

How can we deal with uncertainty on a computer?

Recall Joint distribution is enumerating everything

- Worst-case run time: $O(2^n)$
  - n = # of RVs
- Space is $O(2^n)$ too
  - Size of the table of the joint distribution

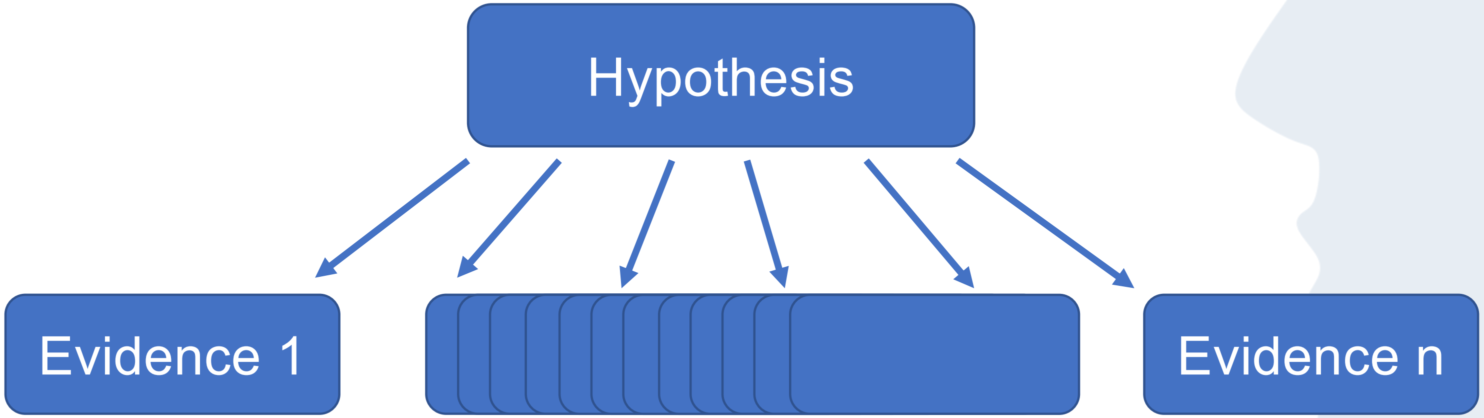
Mission over? No!! Our mission has just started

Main idea: make use of independencies to compress the representation

Naïve Bayes model
Bayes Rule and Independence

A naïve Bayes model assumes that all effects are independent given the cause

$$P(hypothesis, evidence_1, evidence_2, ..., evidence_n) = P(hypothesis) \prod_i P(evidence_1 | hypothesis)$$



The total number of parameters is linear in n

Graphical encoding of conditional distributions

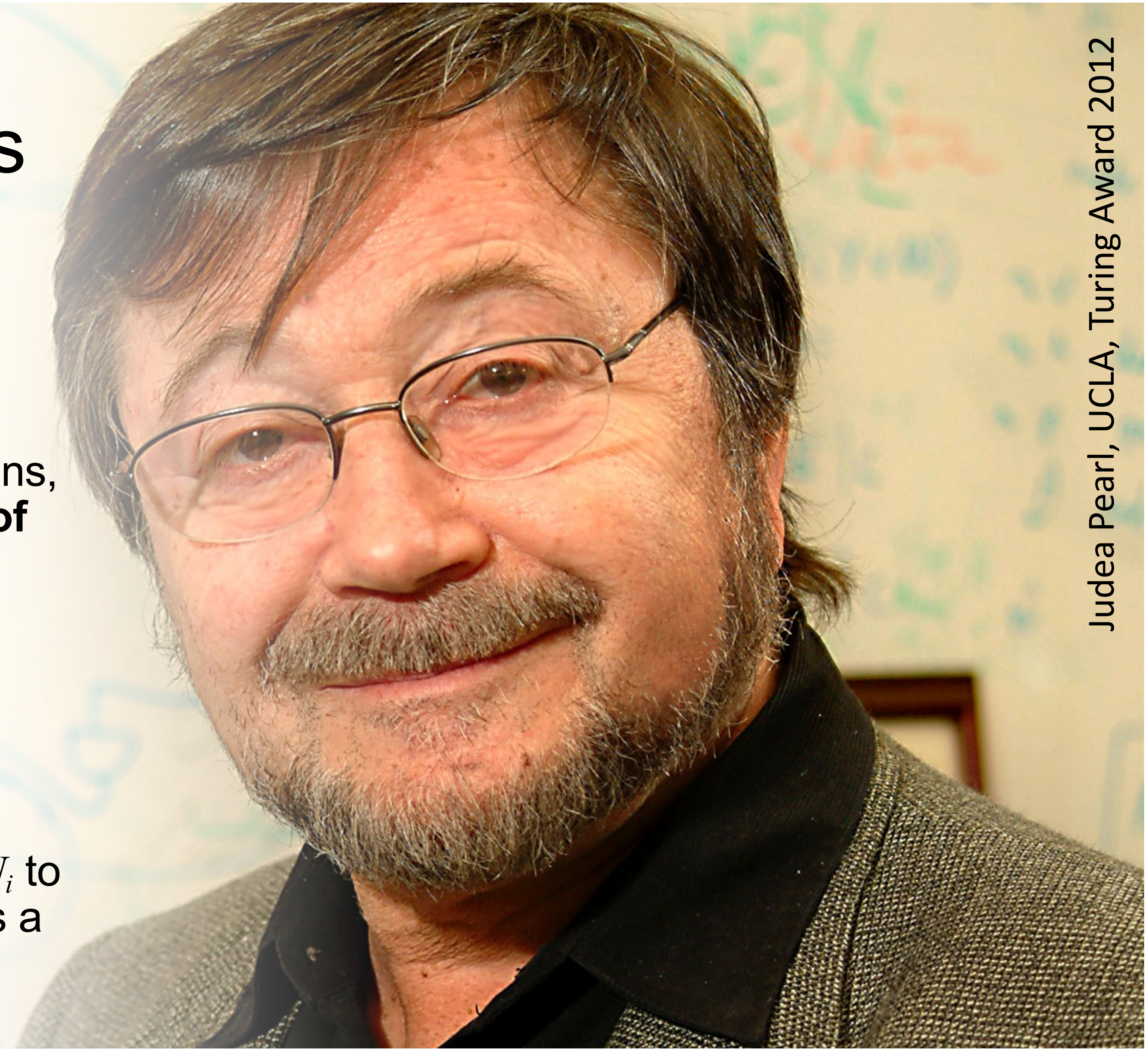
Bayesian Networks

Are a simple, graphical notation for **conditional independence** assertions, hence for **compact specifications of full joint distributions**

A BN is a directed acyclic graph with the following components:

Nodes: one node for each variable

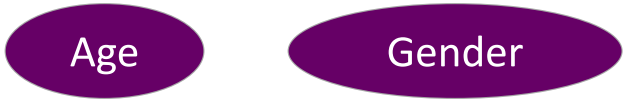
Edges: a directed edge from node N_i to node N_j indicates that variable X_i has a direct influence upon variable X_j



Independency

Let us develop this step by step

(Current) age and the gender of a person are independent



$$P(G, A) = P(G) \cdot P(A)$$

$$P(A \mid G) = P(A)$$

$$P(G \mid A) = P(G)$$

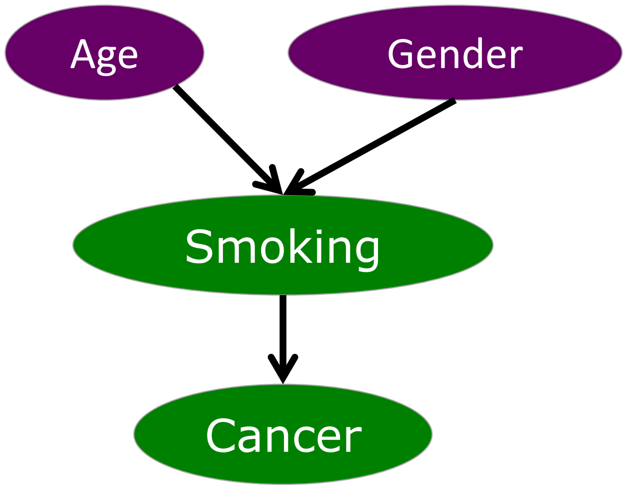
You would not give me money for information on the gender to know the age of a person!

Conditional Independence

Recall, absolute independencies are rare

Cance is independent of age and gender, if the person smokes.

If you have not observed anything,age and gender are independent.



Less entries and consequently lower complexity

$$P(C|S,G,A) = P(C|S)$$

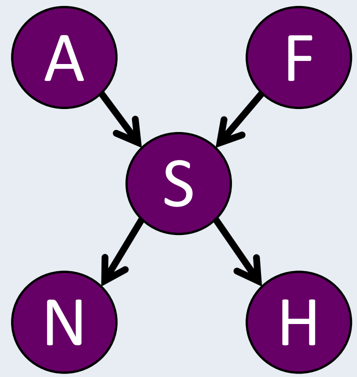
Bayesian Networks [Pearl 1989]

Set of random variables $\{X_1, \ldots, X_n\}$

Directed, acyclic graph (DAG)

To each RV X_1 we associate the

conditional probabilitiy distribution: $P\left(X_i \mid \mathrm{Pa}\left(X_i\right)\right)$



The **joint distribution** is $P(X_1, \ldots, X_n) = \boxed{\prod_{i=1}^{n} P(X_i | Pa(X_i))}$

BN semantics

Local Markov Assumption

Each RV X is independent of ist „non-descendant“ given its parents ($X_i \perp$ nonDescendants $|$ Pa_{X_i})

Example
A very simple one

$S \in \{no, few, many\}$ $C \in \{no, benigne, maligne\}$

P(S=n)	0.80
P(S=f)	0.15
P(S=m)	0.05

Smoking=	n	f	m
P(C=n)	0.96	0.88	0.60
P(C=b)	0.03	0.08	0.25
P(C=m)	0.01	0.04	0.15

But how do we do inference?

What is Inference in Bayesian Networks?

Query: $P(X \mid e)$

Definiton of conditional probability $P(X \mid e) = \dfrac{P(X, e)}{P(e)}$

Up to normalization $P(X \mid e) \propto P(X, e)$

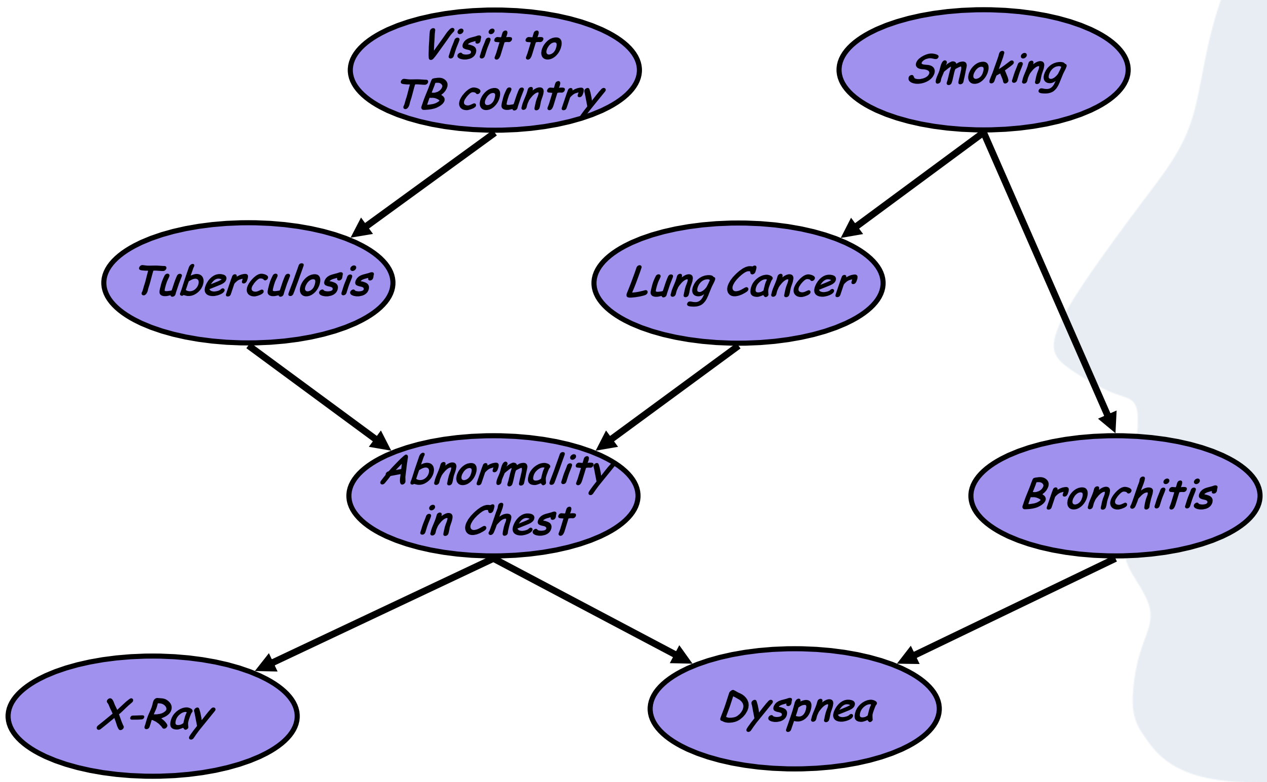
Hence, this rewrites to

$$P(\mathbf{Y}) = \underbrace{\sum_{X_i \notin \mathbf{Y}} \underbrace{\prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))}_{\text{BN semantics}}}_{\text{Marginalization}}$$

$\Sigma_a (P_1 \times P_2) = (\Sigma_a P_1) \times P_2$ if A is not in P_2

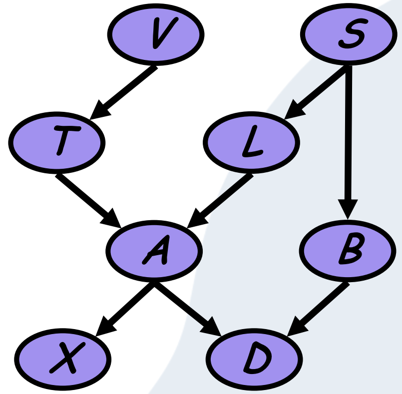
Let us have look at an example

“Tuberculosis” network:



Variable Elimination

- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*



Initial factors

$$\underline{P(v)}P(s)\underline{P(t\,|\,v)}P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

Eliminate: *v*

Compute: $f_v(t) = \sum_v P(v)P(t\,|\,v)$

$$\Rightarrow \underline{f_v(t)}P(s)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

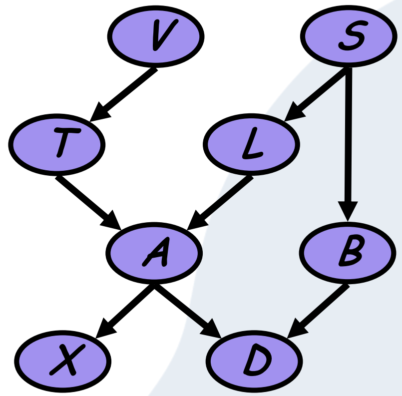
Note: $f_v(t) = P(t)$

In general, result of elimination is not necessarily a probability term

12

Variable Elimination

- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*



Initial factors

$$P(v)P(s)P(t \mid v)P(l \mid s)P(b \mid s)P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

$$\Rightarrow f_v(t)\underline{P(s)}\,\underline{P(l \mid s)}\,\underline{P(b \mid s)}P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

Eliminate: *s*

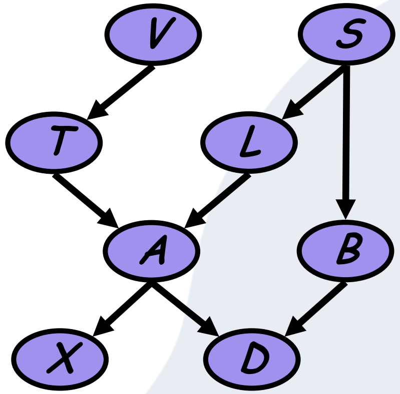
Compute: $f_s(b,l) = \sum_s P(s)P(b \mid s)P(l \mid s)$

$$\Rightarrow f_v(t)\underline{f_s(b,l)}P(a \mid t,l)P(x \mid a)P(d \mid a,b)$$

Summing on *s* results in a factor with two arguments $f_s(b,l)$
In general, result of elimination may be a function of several variables

Variable Elimination

- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*



Initial factors

$$P(v)P(s)P(t\,|\,v)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

$$\Rightarrow f_v(t)P(s)P(l\,|\,s)P(b\,|\,s)P(a\,|\,t,l)P(x\,|\,a)P(d\,|\,a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a\,|\,t,l)\underline{P(x\,|\,a)}P(d\,|\,a,b)$$

Eliminate: *x*

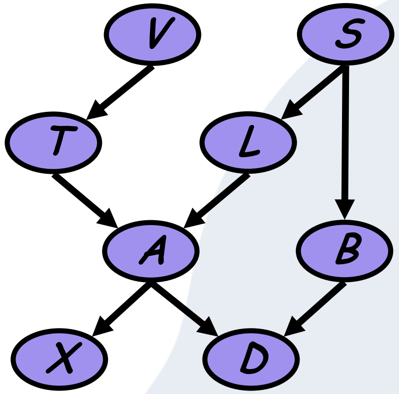
Compute: $f_x(a) = \sum_x P(x\,|\,a)$

$$\Rightarrow f_v(t)f_s(b,l)\underline{f_x(a)}P(a\,|\,t,l)P(d\,|\,a,b)$$

Note: $f_x(a) = 1$ for all values of *a* !!

Variable Elimination

- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*



Initial factors

$$P(v)P(s)P(t|v)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)P(s)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow \underline{f_v(t)}f_s(b,l)f_x(a)\underline{P(a|t,l)}P(d|a,b)$$

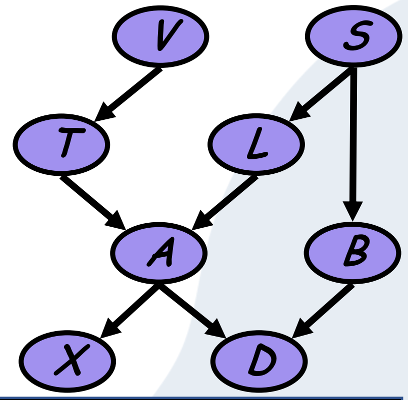
Eliminate: *t*

Compute: $f_t(a,l) = \sum_t f_v(t)P(a|t,l)$

$$\Rightarrow f_s(b,l)f_x(a)\underline{f_t(a,l)}P(d|a,b)$$

Variable Elimination

- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*



Initial factors

$$P(v)P(s)P(t|v)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)P(s)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a|t,l)P(d|a,b)$$

$$\Rightarrow \underline{f_s(b,l)}f_x(a)\underline{f_t(a,l)}P(d|a,b)$$

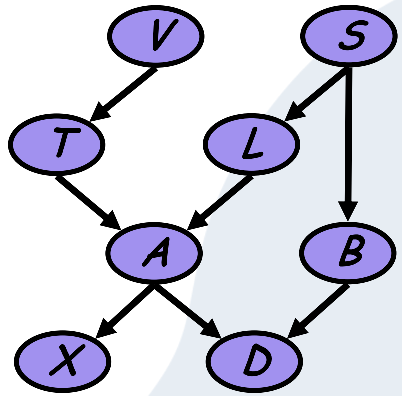
Eliminate: *l*

Compute: $f_l(a,b) = \sum_l f_s(b,l)f_t(a,l)$

$$\Rightarrow \underline{f_l(a,b)}f_x(a)P(d|a,b)$$

Variable Elimination

- We want to compute *P(d)*
- Need to eliminate: *v,s,x,t,l,a,b*



Initial factors

$$P(v)P(s)P(t|v)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)P(s)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)P(a|t,l)P(x|a)P(d|a,b)$$

$$\Rightarrow f_v(t)f_s(b,l)f_x(a)P(a|t,l)P(d|a,b)$$

$$\Rightarrow f_s(b,l)f_x(a)f_t(a,l)P(d|a,b)$$

$$\Rightarrow f_l(a,b)f_x(a)P(d|a,b) \Rightarrow f_a(b,d) \Rightarrow f_b(d)$$

Eliminate: *a,b*

Compute: $f_a(b,d) = \sum_a f_l(a,b)f_x(a)p(d|a,b) \quad f_b(d) = \sum_b f_a(b,d)$

As an algorithm, this is called: Variable elimination

Given a BN and a query P(X|e) / P(X,e)

Instantiate evidence **e**

Choose an elimination order over the variables, e.g., $X_1, \ldots, X_n$

Initial *factors* $\{f_1,\ldots,f_n\}$: $f_i = P(X_i|\mathbf{Pa}_{X_i})$ (CPT for X_i)

For i = 1 to n, if $X_i \notin \{X, \mathbf{E}\}$

- Collect factors $f_1,\ldots,f_k$ that include X_i
- Generate a new factor by eliminating X_i from these factors

$$g = \sum_{X_i} \prod_{j=1}^{k} f_j$$

- Variable X_i has been eliminated! Add g to the set of factors

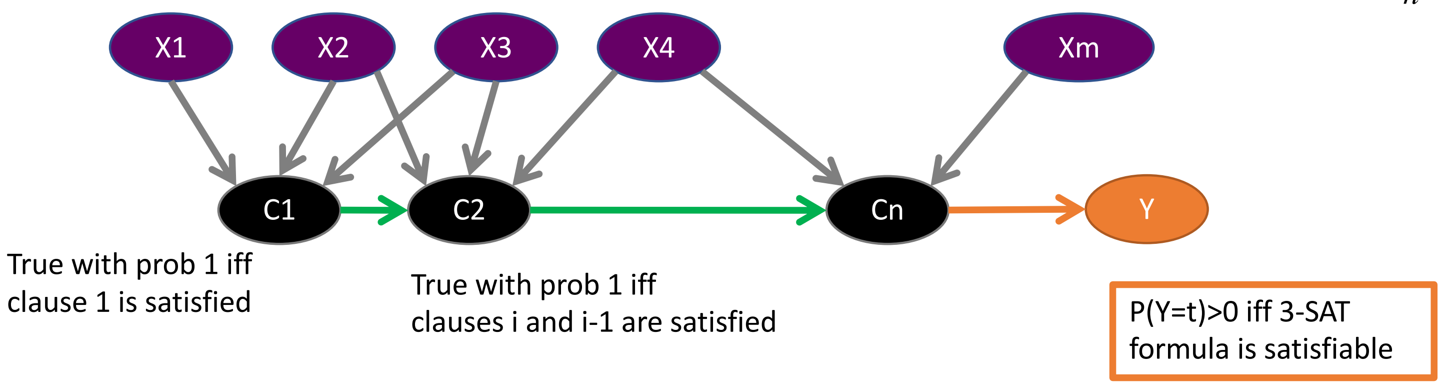
Normalize (everthing sums to 1) P(X,**e**) to obtain P(X|**e**)

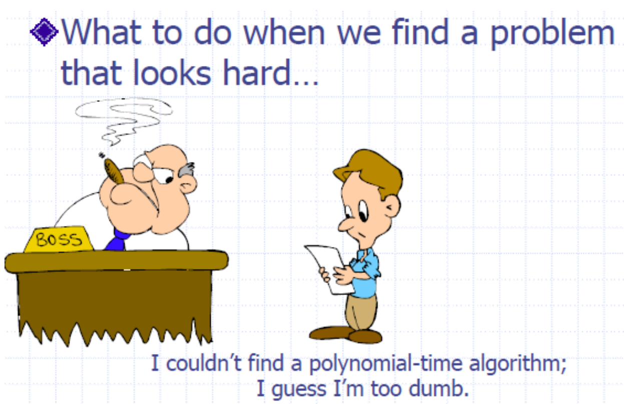
Mission Completed? No …

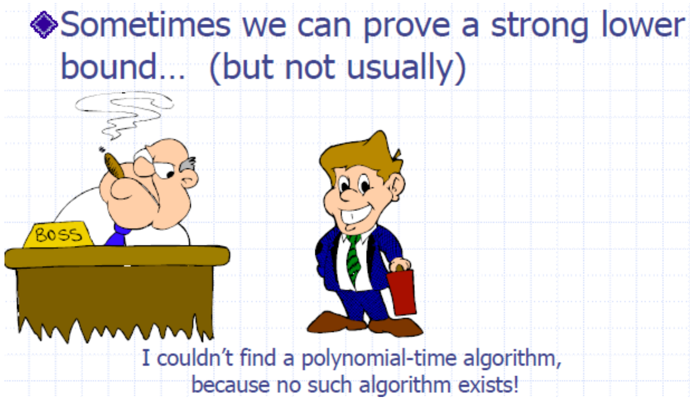
Theorem: Inference (even approximate) in Bayesian networks is NP-hard (#P; via reduction to 3-SAT)

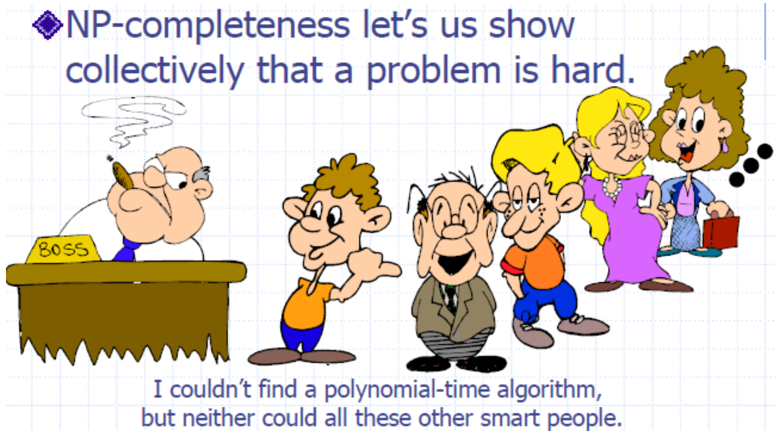
$$\underbrace{(\overline{X}_1 \vee X_2 \vee X_3)}_{C_1} \wedge \underbrace{(\overline{X}_2 \vee X_3 \vee X_4)}_{C_2} \wedge \underbrace{\ldots}_{C_n}$$

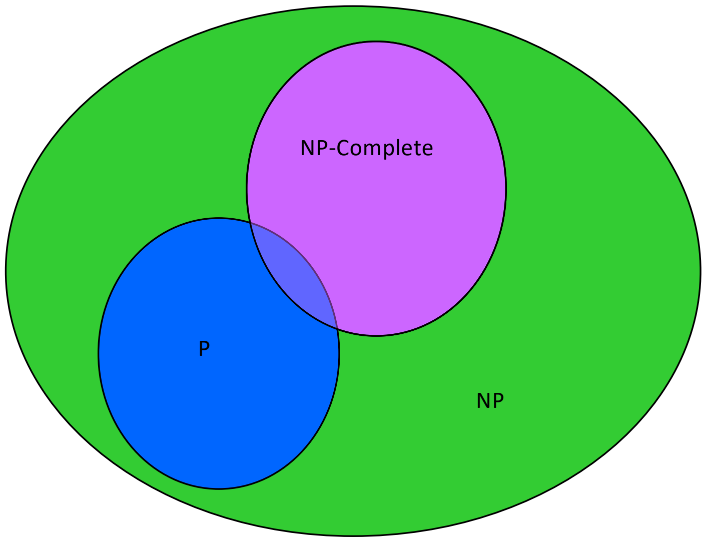
0.5/0.5 prior

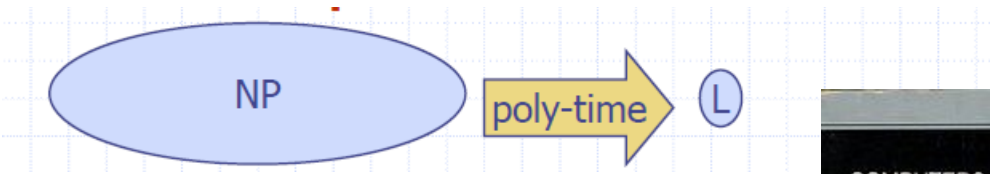


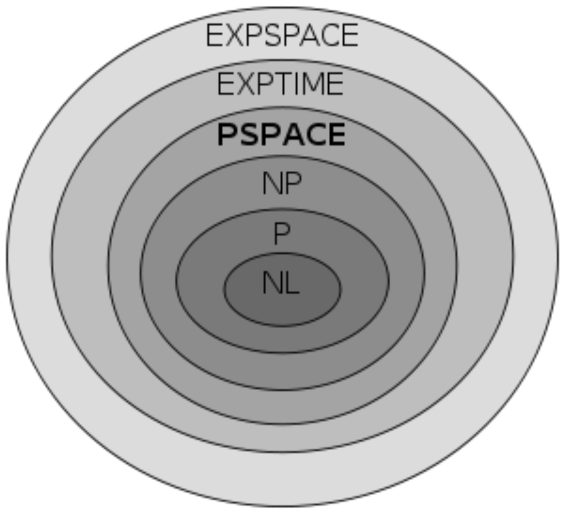

What to do when we find a problem that looks hard…
BOSS
I couldn't find a polynomial-time algorithm;
I guess I'm too dumb.

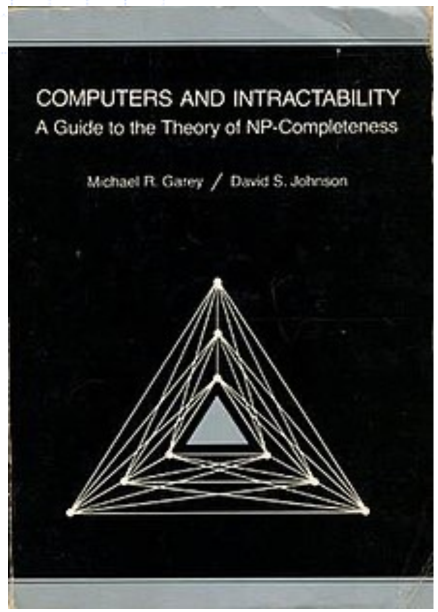

Sometimes we can prove a strong lower bound… (but not usually)
BOSS
I couldn't find a polynomial-time algorithm,
because no such algorithm exists!


NP-completeness let's us show collectively that a problem is hard.
BOSS
I couldn't find a polynomial-time algorithm,
but neither could all these other smart people.


NP-Complete
P
NP


NP
poly-time
L


EXPSPACE
EXPTIME
PSPACE
NP
P
NL


COMPUTERS AND INTRACTABILITY
A Guide to the Theory of NP-Completeness
Michael R. Garey / David S. Johnson

Complexity of Inference

Theorem:

Inference in Bayesian networks
(even approximate, without proof) is NP-hard

Approximate Inference
Inference by Stochastic Sampling (Sampling from a BN)

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

Outline:

- Sampling from an empty network
- Rejection sampling: reject samples disagreeing with evidence
- Likelihood weighting: use evidence to weight samples
- Markov Chain Monte Carlo (MCMC): Sample from a stochastic process whose stationary distribution is the true posterior

How to draw a sample?

Given:

- Random variable X, $D(X)=\{0, 1\}$
- $P(X) = \{0.3, 0.7\}$

Sample $X \leftarrow P(X)$

- Draw a random number $r \in [0, 1]$
- If ($r < 0.3$) then set $X=0$
- Else set $X=1$

Can generalize of any domain size

How to draw a sample?
Sampling from an "Empty Network"

Generating samples from a network that has no evidence associated with it (*empty* network)

Basic idea:

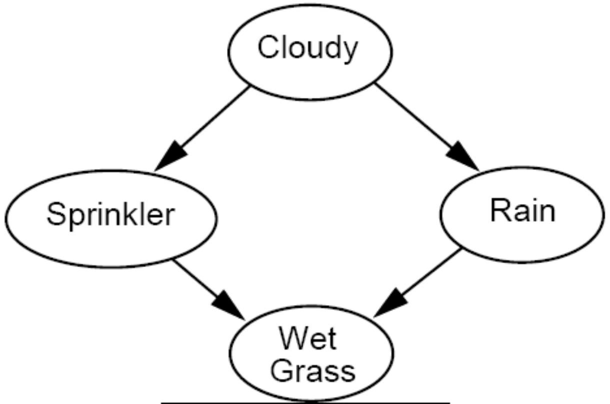
- sample a value for each variable in topological order
- using the specified conditional probabilities

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 $\mathbf{x}$

How to draw a sample?
Example

P(C)
.50

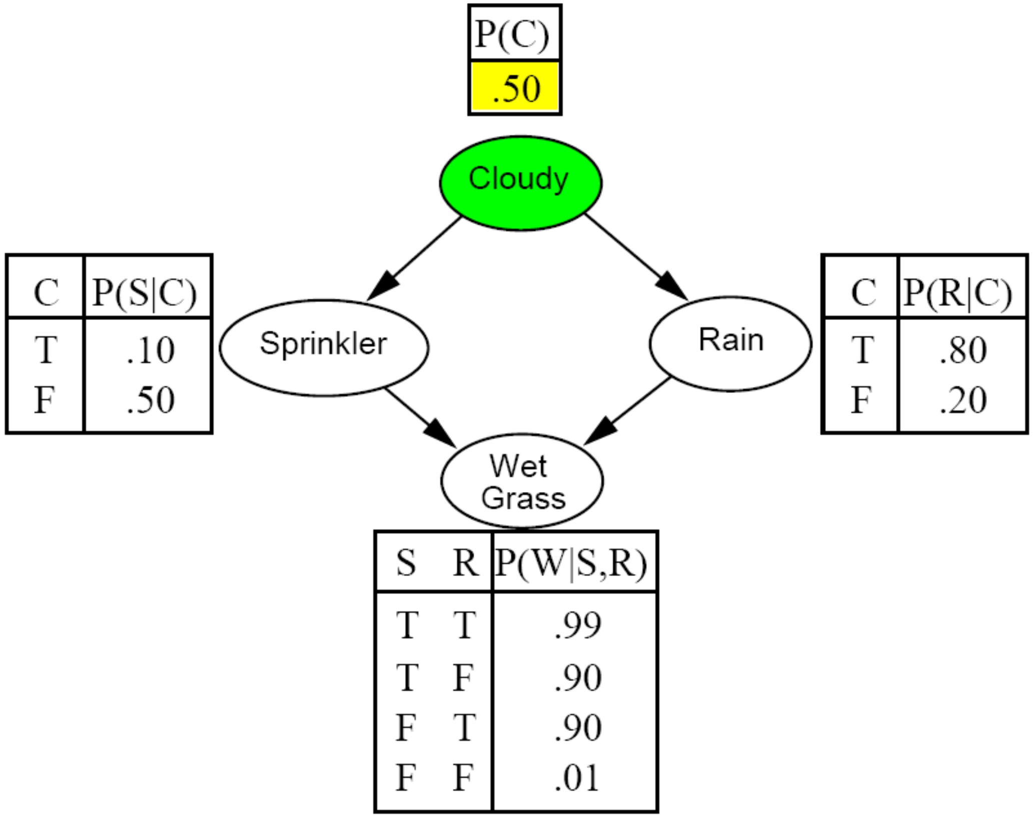


C	P(S|C)
T	.10
F	.50

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

How to draw a sample?
Example



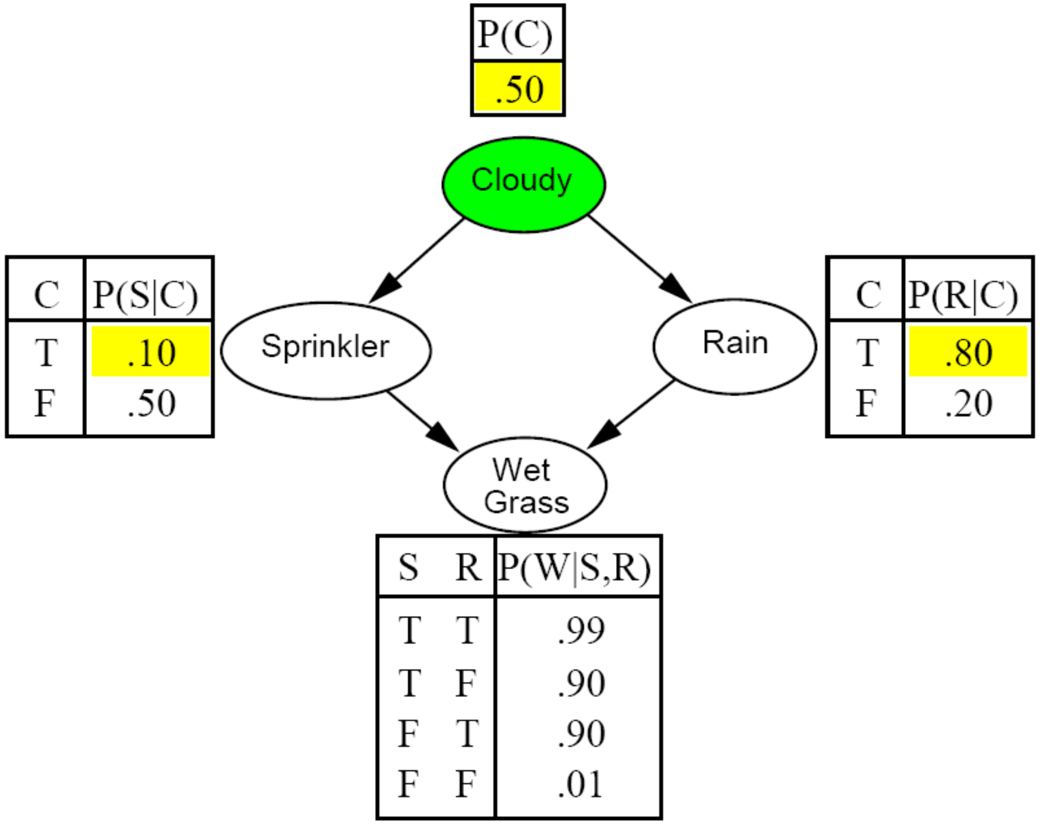
P(C)
.50

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

How to draw a sample?
Example



P(C)
.50

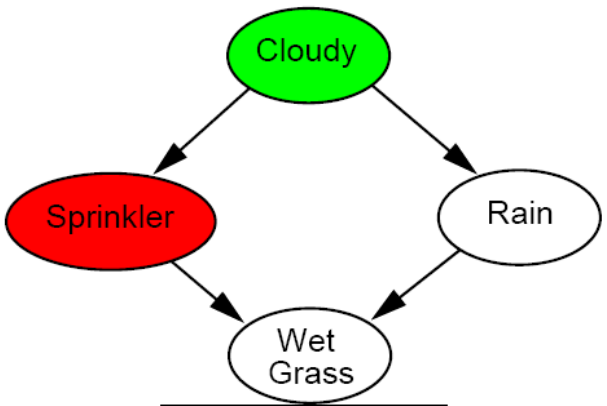
C	P(S|C)
T	.10
F	.50

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

How to draw a sample?
Example

P(C)
.50

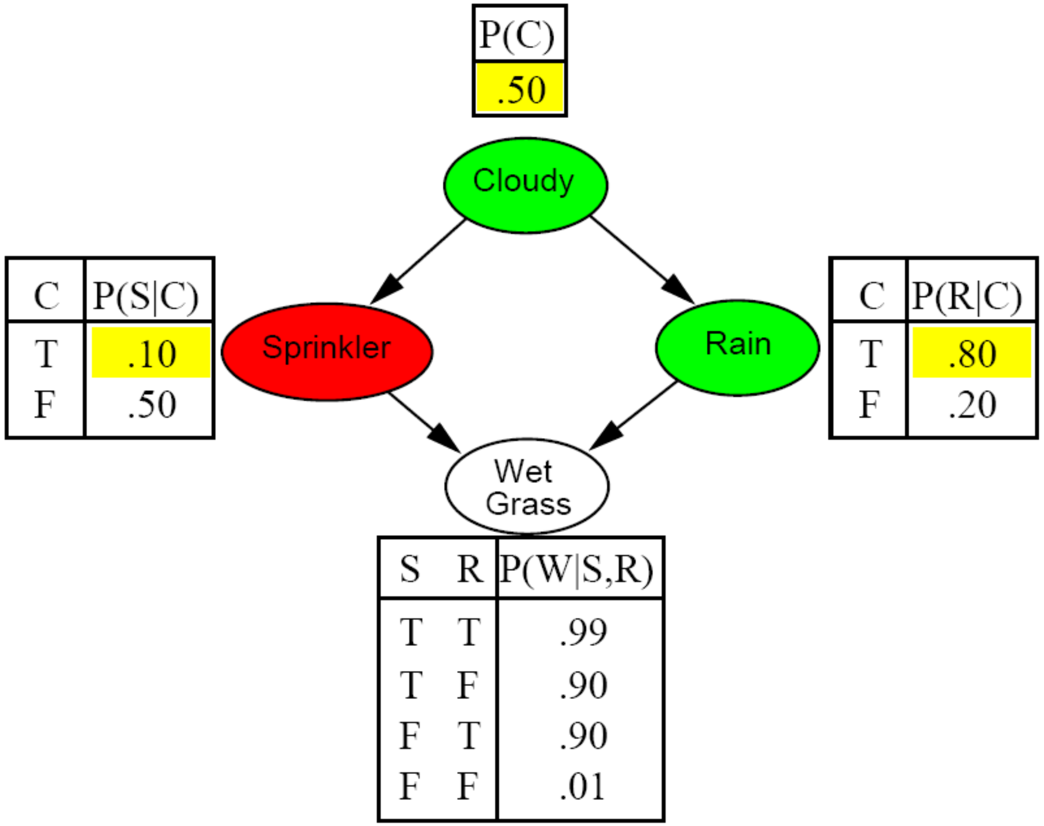


C	P(S|C)
T	.10
F	.50

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

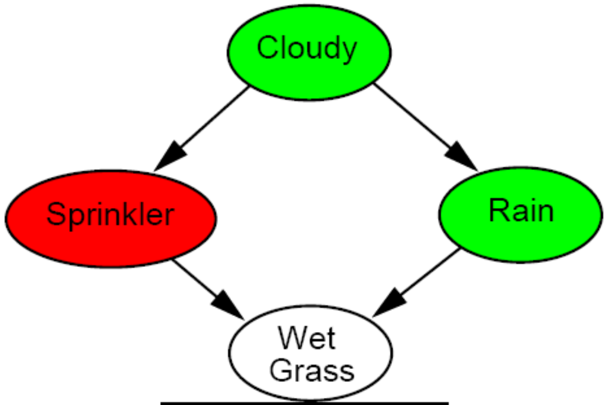
S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

How to draw a sample?
Example



How to draw a sample?
Example

P(C)
.50

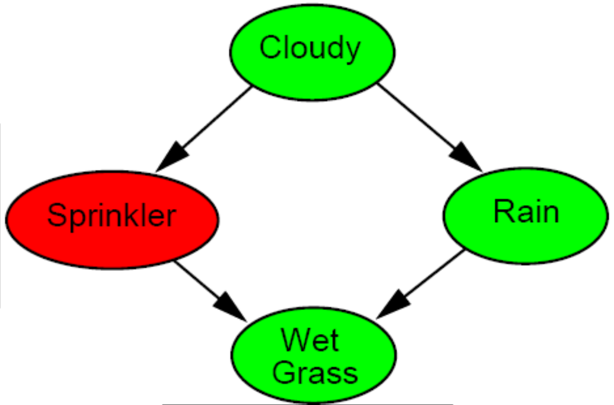


C	P(S|C)
T	.10
F	.50

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

How to draw a sample?
Example



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

Probability Estimation using Sampling

How do we calculate a probability estimation?

- Sample many points using the above algorithm
- count how often each possible combination $x_1, x_2, \ldots, x_n$ appears
- estimate the probability by the observed percentages

  $\hat{P}_{PS}(x_1 \ldots x_n) = N_{PS}(x_1 \ldots x_n) / N$

This converges towards the joint probability function!

Markov Chain Monte Carlo (MCMC) Sampling

"State" of network = current assignment to all variables.

Generate next state by sampling one variable given Markov blanket
Sample each variable in turn, keeping evidence fixed

function MCMC-ASK($X, \mathbf{e}, bn, N$) **returns** an estimate of $P(X|\mathbf{e})$
  local variables: $\mathbf{N}[X]$, a vector of counts over X, initially zero
    $\mathbf{Z}$, the nonevidence variables in bn
    $\mathbf{x}$, the current state of the network, initially copied from $\mathbf{e}$

  initialize $\mathbf{x}$ with random values for the variables in $\mathbf{Y}$
  for $j = 1$ to N **do**
    for each Z_i in $\mathbf{Z}$ **do**
      sample the value of Z_i in $\mathbf{x}$ from $\mathbf{P}(Z_i|mb(Z_i))$
        given the values of $MB(Z_i)$ in $\mathbf{x}$
      $\mathbf{N}[x] \leftarrow \mathbf{N}[x] + 1$ where x is the value of X in $\mathbf{x}$
  return NORMALIZE($\mathbf{N}[X]$)

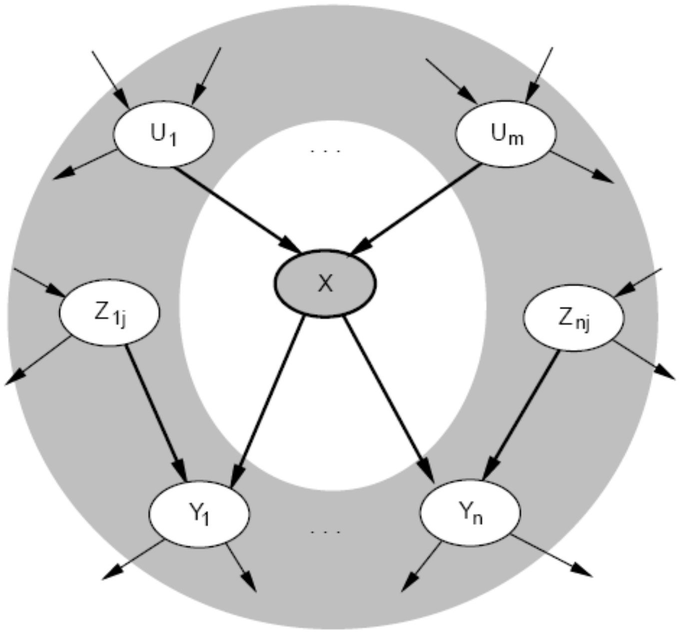
Gibbs Sampling

Can also choose a variable to sample at random each time

Markov Blanket

Markov Blanket = parents + children + children's parents

Each node is conditionally independent of all other nodes given its Markov blanket



$$\mathbf{P}\big(X \mid U_{1,}\ldots, U_m, Y_{1,}\ldots, Y_n, Z_{1\mathrm{j}}, \ldots, Z_{nj}\big) = \boldsymbol{P}(X \mid allvariables)$$

Ordered Gibbs Sampler

Generate sample x^{t+1} from x^t : Process all variables in some order

$$X_1 = x_1^{t+1} \leftarrow P(x_1 \mid x_2^t, x_3^t, \ldots, x_N^t, e)$$

$$X_2 = x_2^{t+1} \leftarrow P(x_2 \mid x_1^{t+1}, x_3^t, \ldots, x_N^t, e)$$

...

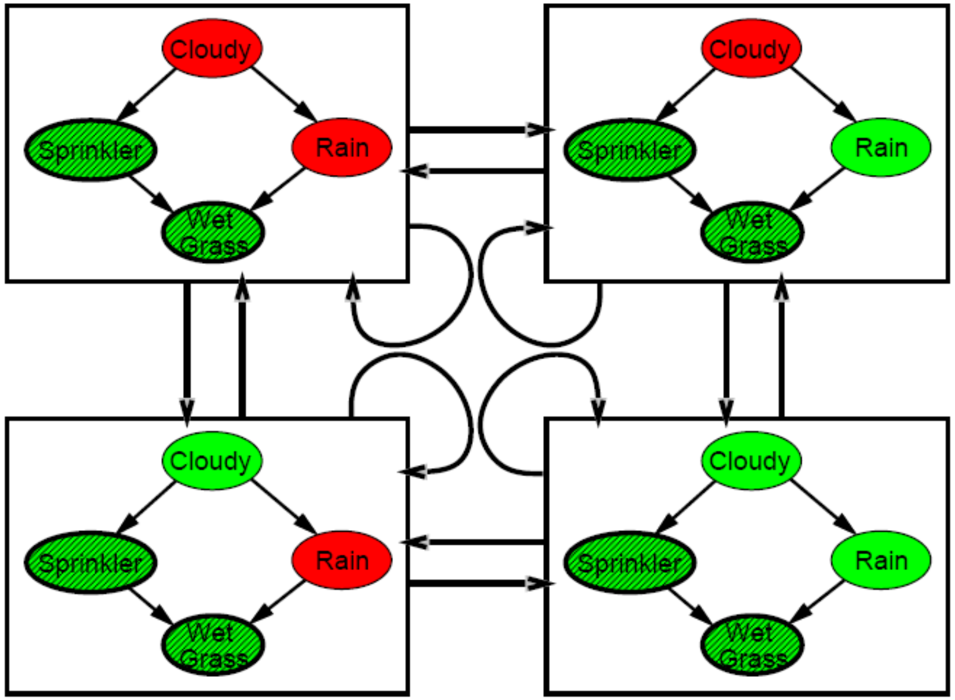
$$X_N = x_N^{t+1} \leftarrow P(x_N \mid x_1^{t+1}, x_2^{t+1}, \ldots, x_{N-1}^{t+1}, e)$$

In short, for i=1 to N:

$$X_i = x_i^{t+1} \leftarrow \textbf{sampled from } P(x_i \mid x^t \setminus x_i, e)$$

The Markov Chain

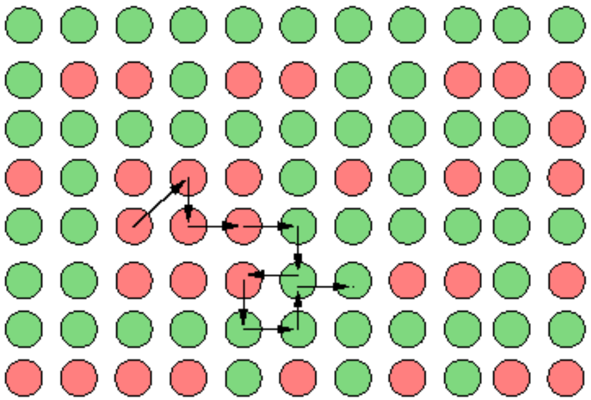
With $Sprinkler = true, WetGrass = true$, there are four states:



Wander about for a while, average what you see

Gibbs Sampling: Illustration

The process of Gibbs sampling can be understood as a *random walk* in the space of all instantiations with $\boldsymbol{Y} = \boldsymbol{u}$:



Reachable in one step: instantiations that differ from current one by value assignment to at most one variable (assume randomized choice of variable X_k).

Guaranteed to converge iff chain is :

irreducible (every state reachable from every other state)
aperiodic (returns to state i can occur at irregular times)
ergodic (returns to every state with probability 1)

How to get a Probability Distribution from Sampling
Example

Task: Estimate $P(Rain|Sprinkler = true, WetGrass = true)$

1. Sample $Cloudy$ or $Rain$ given its Markov Blanket, repeat n times.
2. Count number of times $Rain$ is true and false in the samples.

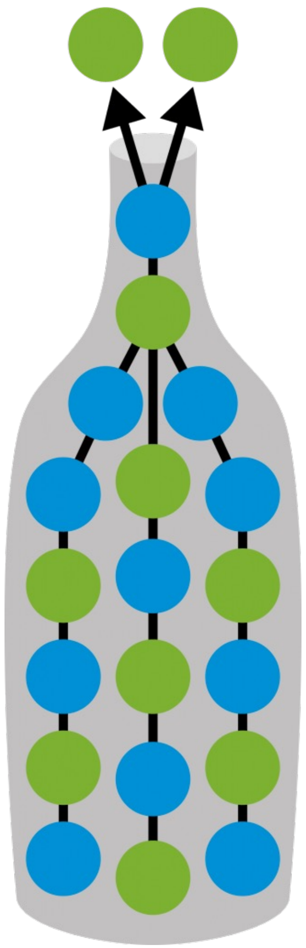
E.g. sample 100 states and count 31 times $Rain$ and 69 without $Rain$

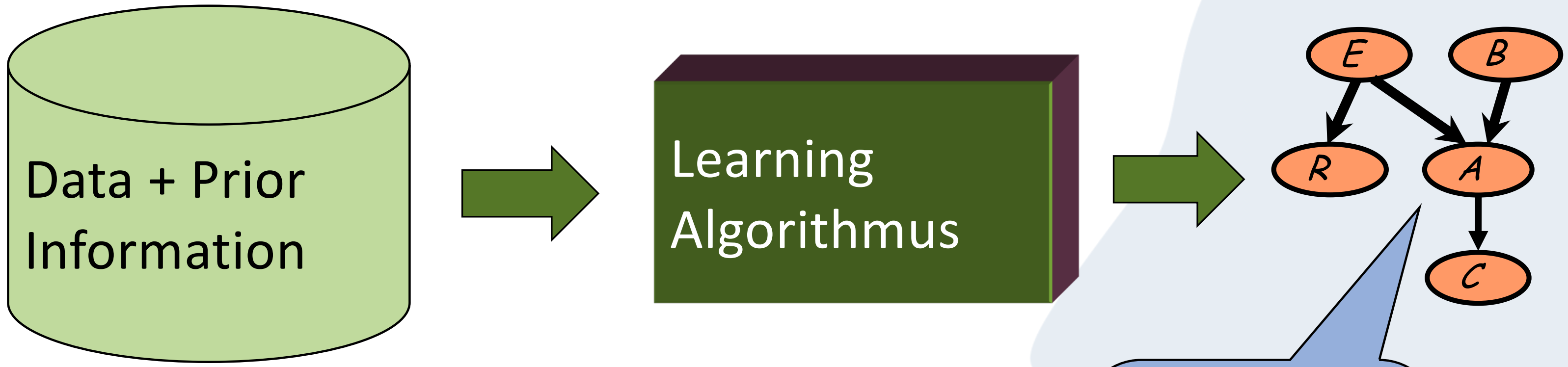
$$\hat{P}(Rain|Sprinkler = true, WetGrass = true)$$
$$= NORMALIZE(\langle 31, 69\rangle) = \langle 0.31, 0.69\rangle$$

Theorem: Chain approaches stationary distribution:
long-run fraction of time spent in each state is exactly proportional to its posterior probability

How to get a Probability Distribution from Sampling

Where do the numbers come from?





E	B	$P(A \mid E,B)$	
e	b	.9	.1
e	$\bar{b}$	.7	.3
$\bar{e}$	b	.8	.2
$\bar{e}$	$\bar{b}$	.99	.01

Knowledge acquisition bottleneck

- Experts are expensive
- It is difficult to get hands on experts

Summary

- Uncertainty is omnipresent
- Uncertainty can be captured using probability distributions
- Graphical models are compact encodings of probability distributions
- They lead to effective algorithms for inference such as Variablen-Elimination
- Inference in Bayesian Netorks is NP-hard

Next Week: special lecture together with Johannes “Juffi” Fürnkranz, JKU Linz, previously with TU Darmstadt

You should be able to:

- Argue why not following the axoims of probabilties is bad
- Compute marginals from joint distriubtions
- Specify a Bayesian network
- Run Variable Elimination