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.

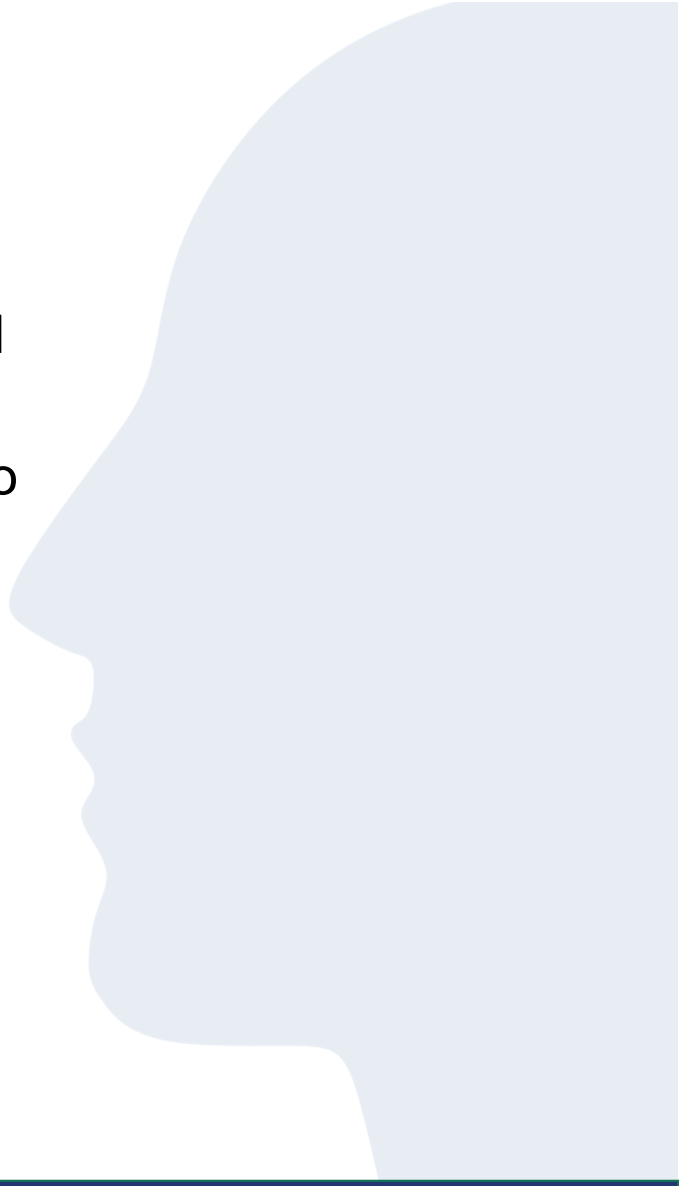
AI101

Lecture 8: Uncertainty

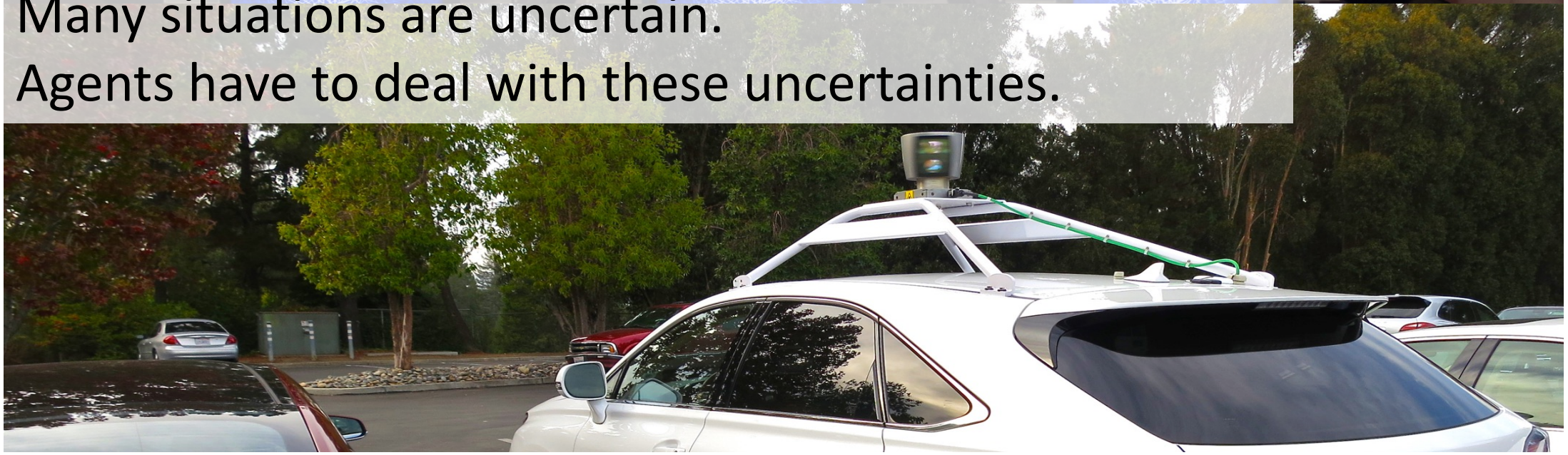
Recap
First Order Logic

Would like our AI to have knowledge about the world, and logically draw conclusions from it

- Notion of an object and use variables (placeholders) to specify abstract knowledge to specify true regularities
- Syntax & Semantics
- Unification
- Skolem
- Resolution
- Gödel's Incompleteness Theorem



Many situations are uncertain.
Agents have to deal with these uncertainties.



So far…

So far, agents believed that:

- (logical) statements are true or false (maybe unknown)
- actions will always do what we think they do

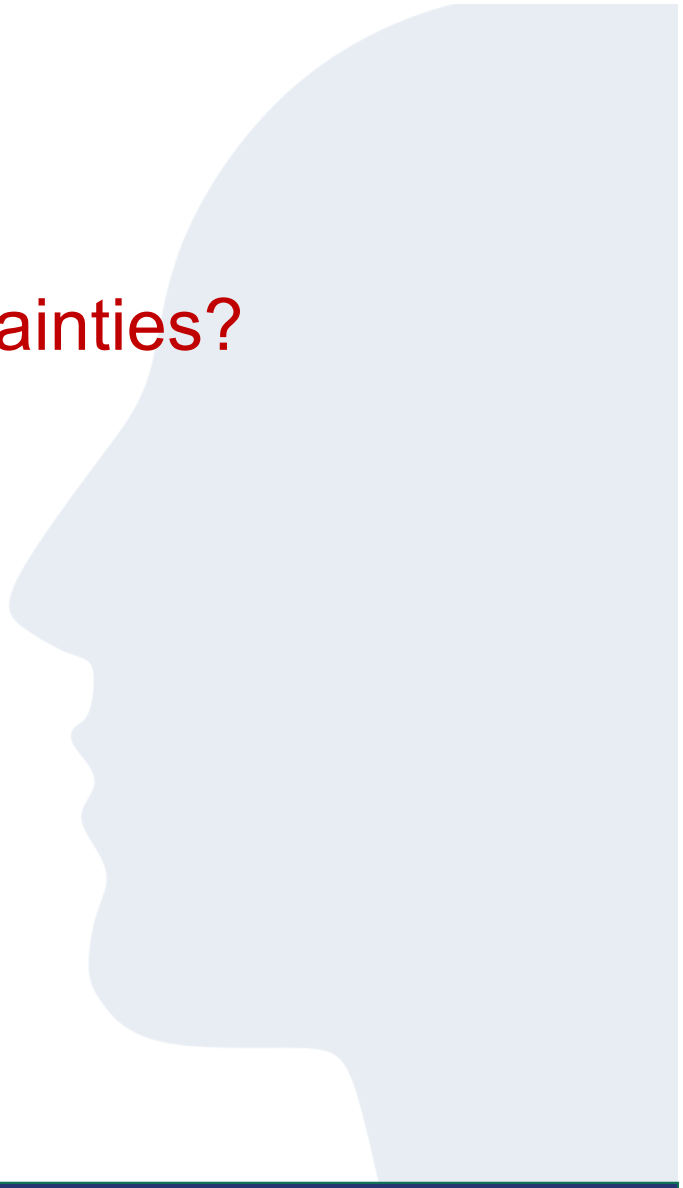
Unfortunately, the real world is not like that:

- Agents almos never have access to the whole truth, i.e. the complete/perfect information

Agents must deal with **uncertainty**

Uncertainty
Outline

How can agents deal with uncertainties?



Today

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule
- Bayesian Networks
- Inference in Bayesian Networks

So far…

Example: Getting to the Airport

“We want to get to the airport to take a flight. When to we leave?”

- We have different actions for getting to the airport

  → action A_t = leave for the airport t minutes before departure

- Typical problems
  - Will a given action A_t get me to the airport in time?
  - Which action is the best choice for getting me to the airport

So far…

Example: Getting to the Airport

Risks involved in the plan A_{90} will get me to the airport (leaving 90min before departure)

- partial observability (road state, other drivers' plans, etc.)
- noisy sensors (traffic reports may be wrong)
- uncertainty in action outcomes (flat tire, accident, etc.)
- immense complexity of modeling and predicting traffic

A logically correct plan: *A_{90} will get me to the airport as long as my car doesn't break down, I don't run out of gas, no accident, the bridge doesn't fall down,* **etc**

Unfortunately, it is impossible to model all things that can go wrong (**qualification problem**)

A more cautious plan: *A_{1440} will get me to the airport*

What would you do?

Probabilities

Probabilities are **one way** of handling uncertainty

- E.g. A_{90} will get me to the airport with probability 0.5

They **summarize the effects** that are due to

- **Laziness**
  - I don't want to list all things that must not go wrong
- **Theoretical Ignorance**
  - Some things just can't be known
  - e.g.: We cannot completely model the weather
- **Practical Ignorance**
  - Some things might not be known about the particular situation
  - e.g. Is there a traffic jam at A5?

Probabilities
How to Understand Probabilities

Probabilities are related to one's (subjective) beliefs

- A probability p means that I believe that the statement will be true in $p \cdot 100\%$ of the cases.
- E.g. There is a traffic jam in *10%* of the cases
- It does not mean that the street is jammed by a degree of *10%*

Probability Theory is about the **degree of belief** not the **degree of truth**

Probabilities of propositions change with new evidence:

- $P(A_{45}$ gets me there in time | no reported accidents$) = 0.06$
- $P(A_{45}$ gets me there in time | no reported accidents, 5 a.m.$) = 0.15$

Probabilities

The **degree of belief** view resolves tricky issues

Consider the probability that the sun will still exist tomorrow.

- Difficult to observe by an experiment

What is the chance that a patient has a particular disease?

- A medical doctor wants to consider other patients who are similar. But if you gather too much information to compare patients, there are no similar patients left

Probabilities
Basics

The state or sample space can be seen as a set of all samples:

- E.g. the sample space of a die roll is 1,2,3,4,5,6 (sides)

$$\Omega$$

$$\omega \in \Omega$$

A probability space or probability model is a sample space with an assignment of probabilities per possible sample

E.g. $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$

$$P(\omega) \text{ for every } \omega \in \Omega$$

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

An event A is any subset of the sample space:

E.g. $P(\text{roll greater } 4) = P(5) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

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

Probabilities

Kolmogorov's Axioms of Probability

1. **All probabilities are between 0 and 1**

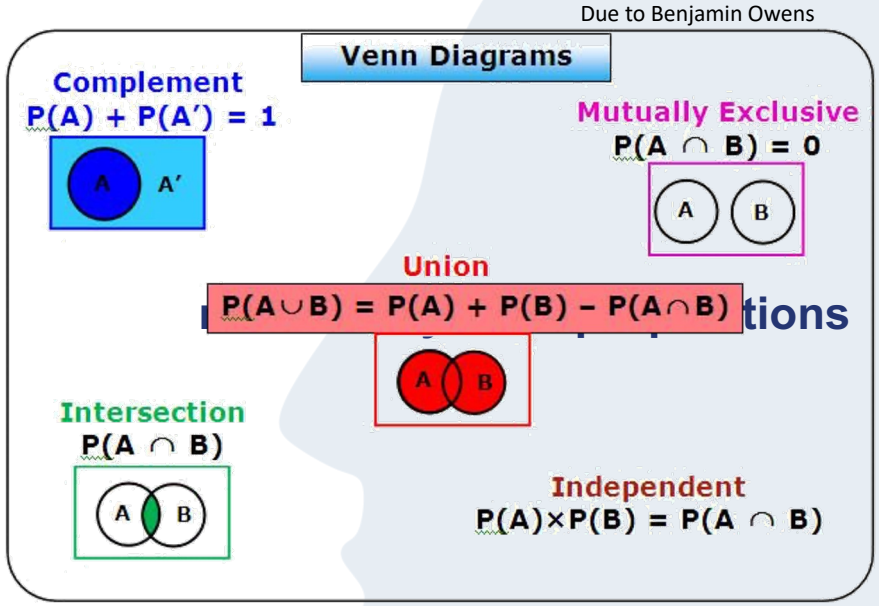
$$0 \leq P(a) \leq 1$$

2. **Necessarily true propositions have probability 1, have probability 0**

$$P(false) = 0, P(tue) = 1$$

3. **The probability of a disjunction is**

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



4. **These axioms restrict the set of probabilistic beliefs that an agent can (reasonably) hold.**
   - similar to logical constraints like A and $\neg A$ can't both be true

Probabilities

You better do not violate the axioms of probability

Dutch Book* Theorem, Bruno de Finetti (1931)

- an agent (in the example it is Agent 1) who bets according to probabilities that violate the axioms of probability can be forced to bet so as to lose money *regardless of outcome!*

Example:

- suppose Agent 1 believes the following: $P(a) = 0.4, P(b) = 0.3, P(a \vee b) = 0.8$
- Agent 2 can now select a set of events and bet on them according to these probabilities so that she cannot loose

Agent 1		Agent 2		Outcome for Agent 1			
proposition	belief	bet	stakes	$a \wedge b$	$a \wedge \neg b$	$\neg a \wedge b$	$\neg a \wedge \neg b$
a	0.4	a	4:6	-6	-6	4	4
b	0.3	b	3:7	-7	3	-7	3
$a \vee b$	0.8	$\neg(a \vee b)$	2:8	2	2	2	-8
				-11	**-1**	**-1**	**-1**

Bruno de Finetti

Axioms of Probability are violated because $P(a \vee b) > P(a) + P(b)$

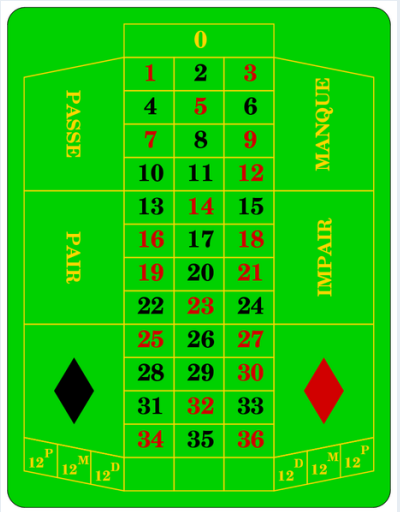
*[Wikipedia] In gambling, a Dutch book is a set of odds and bets, established by the bookmaker, that ensures that the bookmaker will profit, at the expense of the gamblers, regardless of the outcome of the event (a horse race, for example) on which the gamblers bet.

Probabilities

Let's make our life easier: Instead of events, let us use random variables

A random variable is a function from atomic events to some range of values

Example: Roulette



- Atomic events: numbers 0-36
- Random variables with outcomes true or false
  - Rouge | Noir, Pair | Impair, Passe | Manque
  - Transversale, Carre, Cheval
  - Douzaines premier | Milieu | Dernier
  - …

  E.g. rouge(36) = true

The probability function P over atomic events induces a probability distribution over all random variables X

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

Probabilities

Propositions, or towards uncertain knowledge

Think of a proposition as the event where the proposition is true:

Often in AI applications, the sample points are defined by the values of a set of random variables

- i.e. the sample space is the Cartesian product of the ranges of the variables

With Boolean variables, sample points = propositional logic model

- E.g. A=*true*, B=*false*, or $a \wedge \neg b$

Proposition = disjunction of atomic events in which it is true

- E.g. $(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b) \quad \rightarrow P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Probabilities
Syntax of Propositions

Propositional or Boolean random can be true or false

- E.g. hasUmbrella,
- *hasUmbrella=true* is a proposition, and can be simply written as *hasUmbrella*

Discrete random variables (finite or infinite)

- E.g. *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)

- E.g *Temp* is an unbound variable
- *Temp=25.5, Temp>23* are propositions

Joint Distributions
Uncertain/Quantified Truth Table

A joint distribution gives the probability of combined events

- E.g. The probability that X=x and Y=y is true $P(x, y) \equiv P(X = x \wedge Y = y)$

Cancer

Smoking	no	benigne	maligne
no	0.768	0.024	0.008
few	0.132	0.012	0.006
many	0.035	0.010	0.005

The joint distribution allows us to answer any question! But how?

Marginalization (or Summing Out)

We do not want to talk always about all variable!

For any set of variables X and Y we can compute the probability

$$P(Y) = \sum_{i=1}^{n} P(x_i, Y)$$

The resulting distribution is called marginal distribution and its probabilities are the marginal probabilities

Cancer

Smoking	no	benigne	maligne
no	0.768	0.024	0.008
few	0.132	0.012	0.006
many	0.035	0.010	0.005

$$P(Y = few) = P(no, few) + P(benigne, few) + P(maligne, few) = 0.15$$

Conditional Probabilities

are kind of “probability distribution for a sub-population”

A **conditional probability** can be described as the probability of $X{=}x$ under the assumption that $Y{=}y$ is true:

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

The **Product rule** gives us an alternative formulation

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

The **Chain rule** can be derived by successive application of the Product rule:

$$\begin{aligned} &P(X_1, X_2, ..., X_n) = P(X_1, ..., X_{n-1})P(X_n|X_1, ..., X_{n-1}) \\ &= P(X_1, ..., X_{n-2})P(X_{n-1}|X_1, ..., X_{n-2})P(X_n|X_1, ..., X_{n-1}) \\ &= \prod_{i=1}^{n} P(X_i|X_1, ..., X_{i-1}) \end{aligned}$$

Conditional Probabilities

Example

Smoking \ Cancer	no	benigne	maligne
no	0.768	0.024	0.008
few	0.132	0.012	0.006
many	0.035	0.010	0.005

Lets assume $P(Y = few)$

We already now that $P(Y = few) = 0.15$

Now we want to calculate $P(X = maligne|Y = few) = \frac{P(maligne, few)}{P(Y = few)} = \frac{0.006}{0.15} = 0.04$

Lets assume $P(X = maligne)$

We calculate $P(maligne) = P(maligne, no) + P(maligne, few) + P(maligne, many) = 0.019$

Now we want to calculate $P(Y = few|X = maligne) = \frac{P(maligne, few)}{P(maligne)} = \frac{0.006}{0.019} = 0.316$

Joint Distributions

Can we reduce the complexity of the joint distribution?

Yes, if we can make use of independencies

X and Y are independent from another if one of the following is true;

$$P(X|Y) = P(X)$$
$$P(Y|X) = P(Y)$$
$$P(X,Y) = P(X)P(Y)$$

Independent variables are not effected by the other variable

- This reduces the amount of possible values

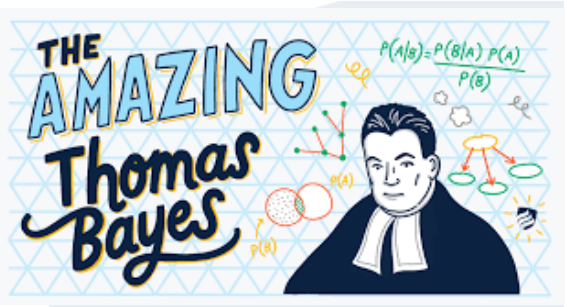
But…

- **Absolut independent variables are rare**
- **i.e. in cancer research there are a lot of variables, none of which are independent**

Bayes Rule

In contrast to logic, no input & output (if P>0)!

Product rule: $P(x,y) = P(x|y)P(y) = P(y|x)P(x)$



Rev. Thomas Bayes (c. 1702 – 1761)
English theologian and mathematician

Bayer' rule:

Probability of the evidence, given the belief is true. This is called **Likelihood**

Probability of the hypothesis before considering the evidence. This is called **Prior**

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

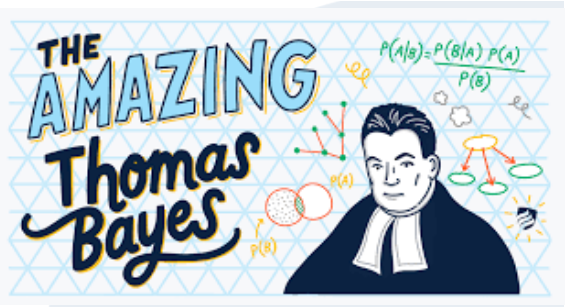
Probability of the hypothesis X after the evidence Y. This is called **Posterior**

Probability of the evidence Y under any circumstance. This is called **Marginalization**

Bayes Rule

Example: AIDS-Test

Bayer' rule: $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$



Rev. Thomas Bayes (c. 1702 – 1761)
English theologian and mathematician

Events

- Aids = a person is infected or not
- Positive = a person has a positive test result

Probabilities

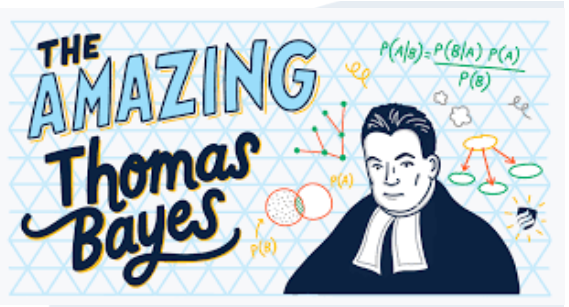
$P(positive, aids) = 0.99,$
$P(negative, aids) = 0.01,$
$P(positive, \neg aids) = 0.005,$
$P(negative, \neg aids) = 0.995$

Is this test reliable?

Bayes Rule

Example: AIDS-Test

Bayer' rule: $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$



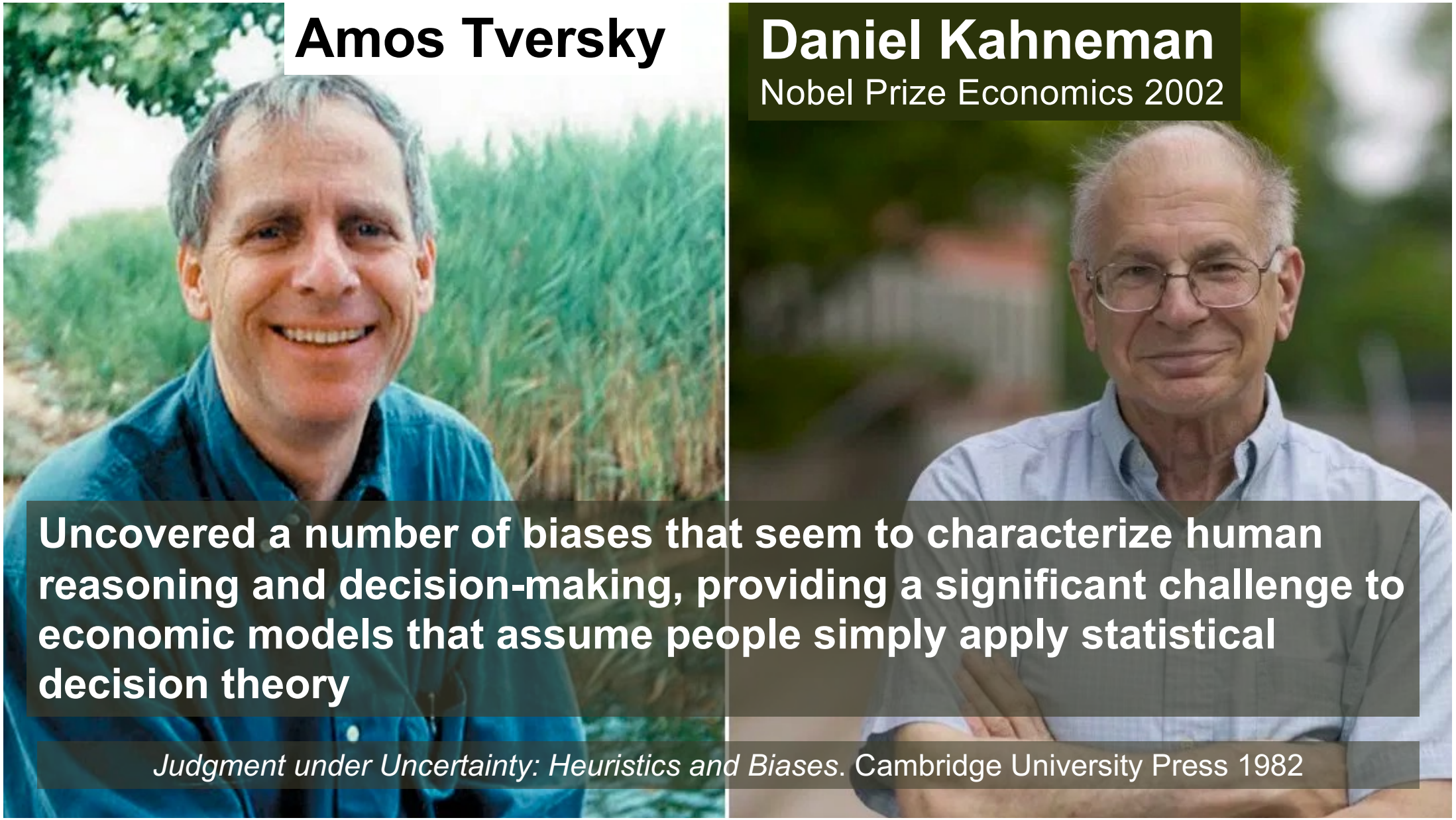
Rev. Thomas Bayes (c. 1702 – 1761)
English theologian and mathematician

Probabilities

$$P(positive, aids) = 0.99,$$
$$P(negative, aids) = 0.01,$$
$$P(positive, \neg aids) = 0.005,$$
$$P(negative, \neg aids) = 0.995$$

Lets assume the risk for you to having aids is $P(aids) = 0.0001$ *(Prior)* and now you have a positive test result. Should you panic?

$$P(a|p) = \frac{P(p|a)P(a)}{P(p)} = \frac{P(p|a)P(a)}{P(p|a)P(a) + P(p|\neg a)P(\neg a)}$$
$$= \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.005 \cdot 0.9999} = \mathbf{0.0194}$$


Amos Tversky
Daniel Kahneman
Nobel Prize Economics 2002
Uncovered a number of biases that seem to characterize human reasoning and decision-making, providing a significant challenge to economic models that assume people simply apply statistical decision theory
Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press 1982

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

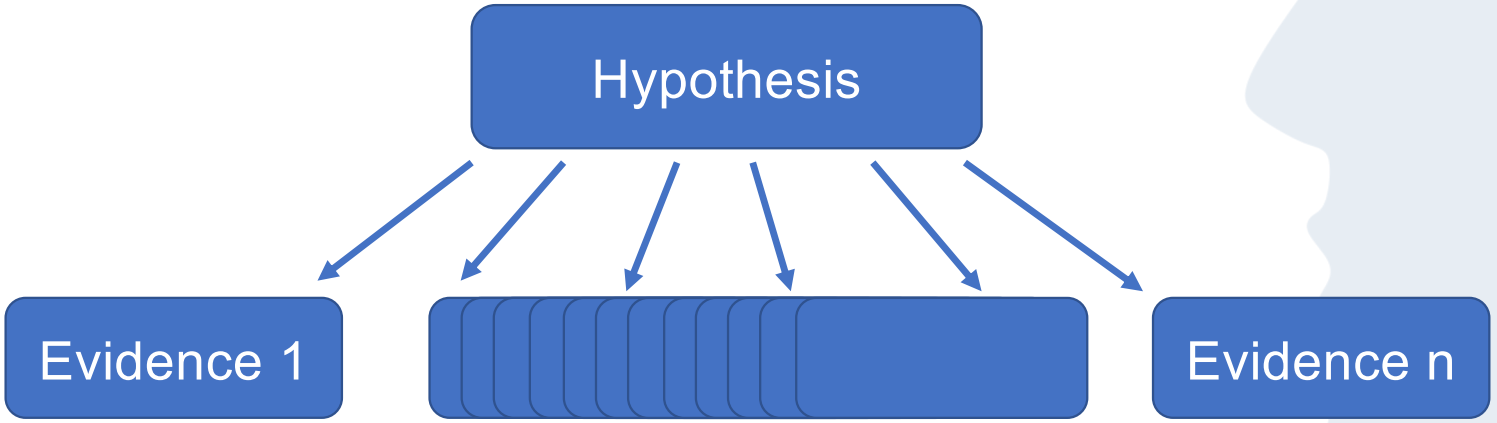
Main idea: make use of independencies to compress the representation

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

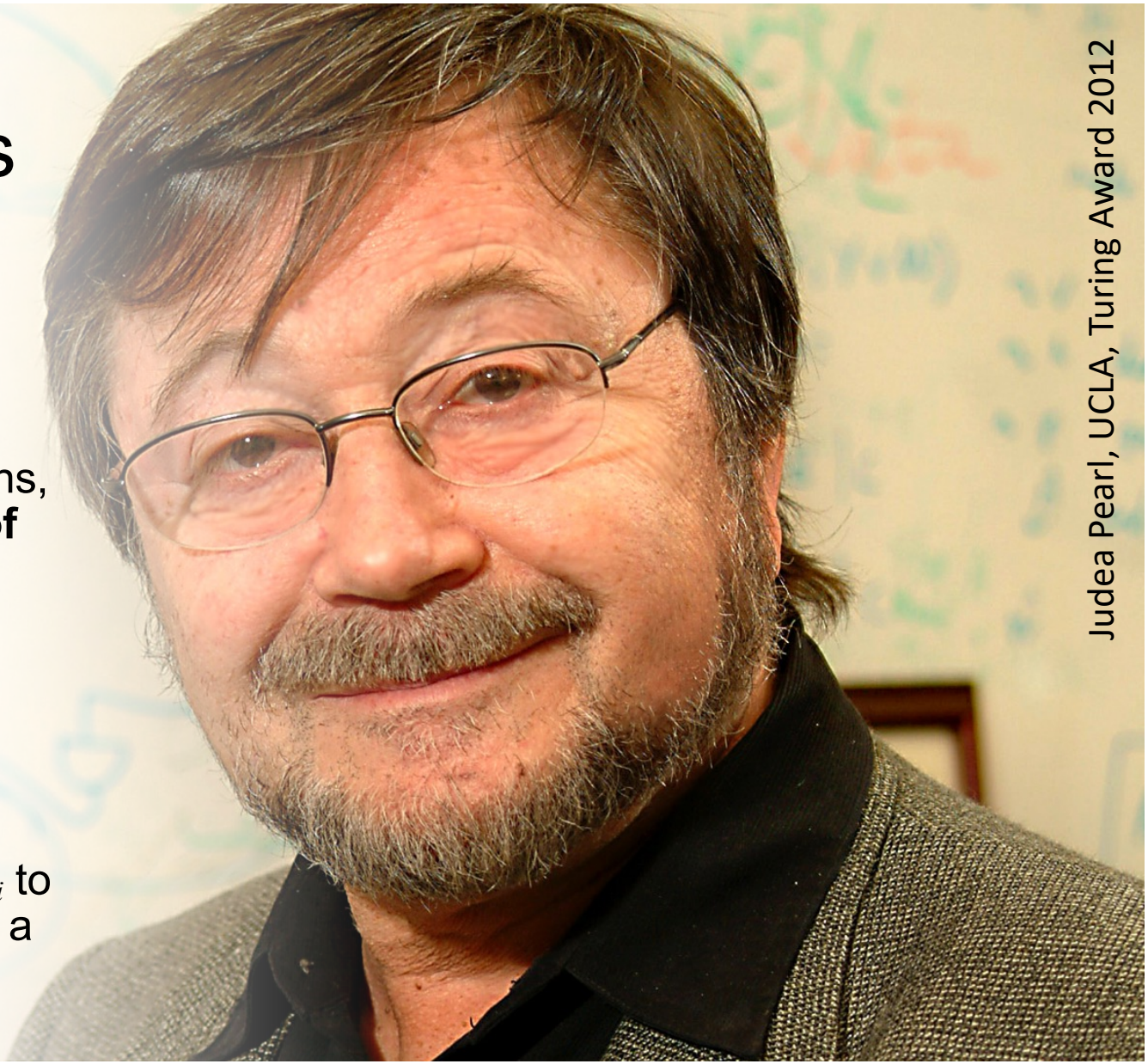
Bayesian Networks

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



Judea Pearl, UCLA, Turing Award 2012

Independency

Let us develop this step by step

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

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

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

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

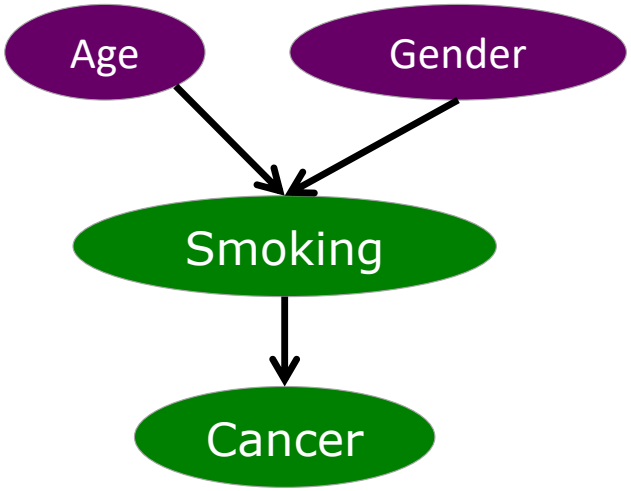
You would not give me money for information on the gender to know the age of a person!

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

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

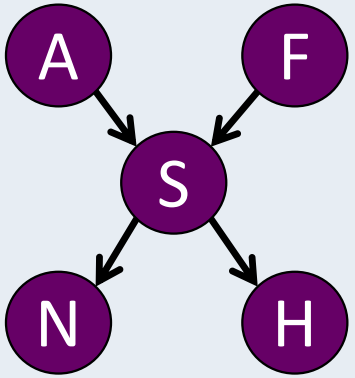
Bayesian Networks [Pearl 1989]

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

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 $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i))$

BN semantics

Local Markov Assumption

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

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

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

Hence, this rewrites to

$$P(\mathbf{Y}) = \sum_{X_i \notin \mathbf{Y}} \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))$$

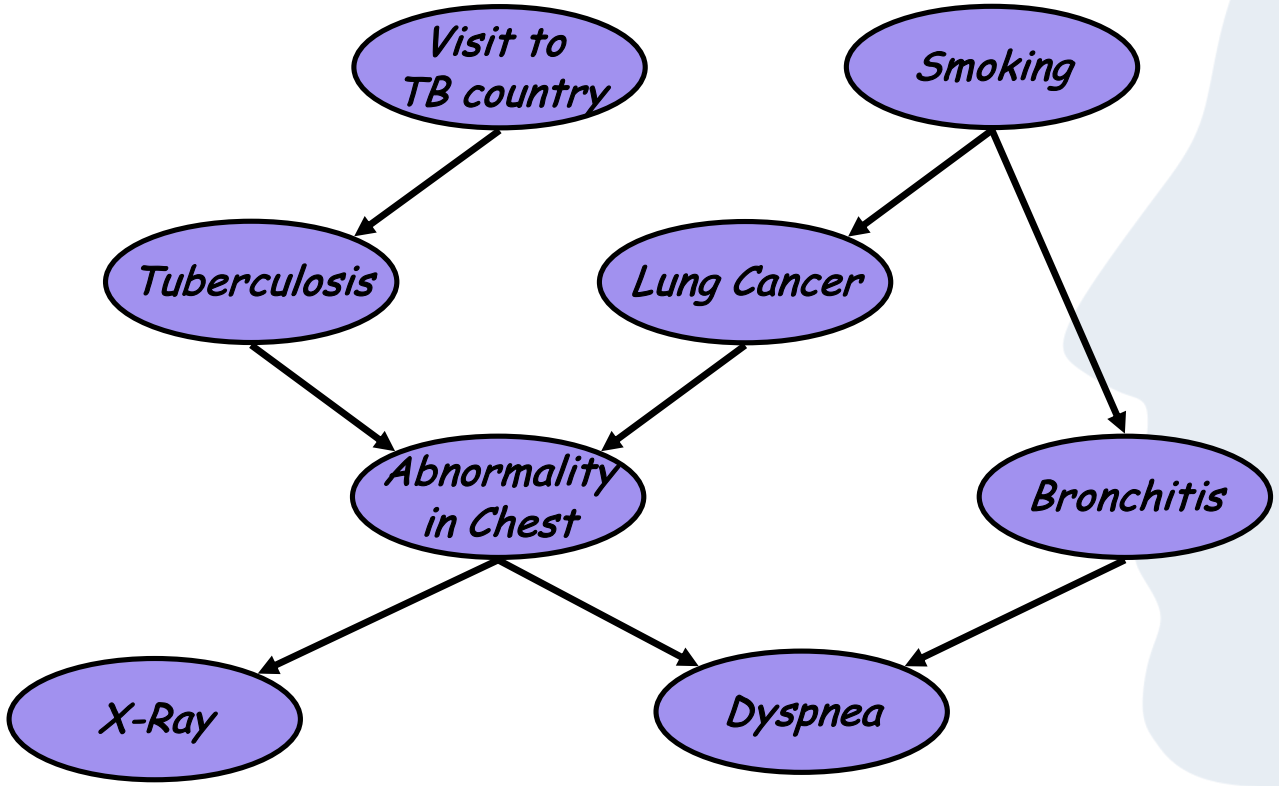
BN semantics

Marginalization

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

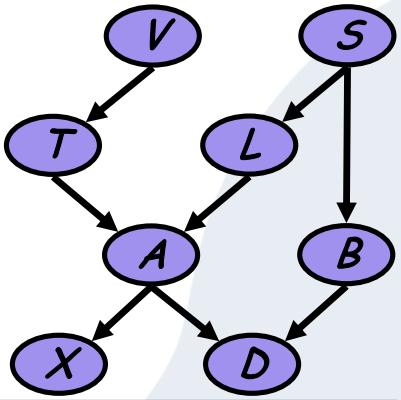
Let us have look at an example

"Tuberculosis" network:



Variable Elimination

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



Initial factors

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

Eliminate: *v*

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

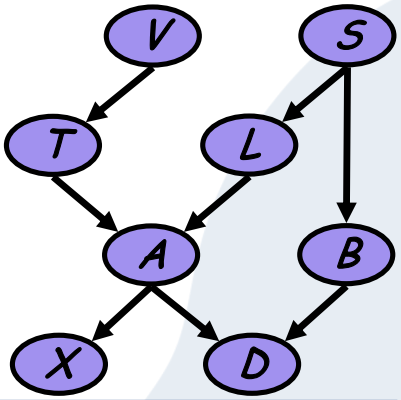
$$\Rightarrow f_v(t)P(s)P(l|s)P(b|s)P(a|t,l)P(x|a)P(d|a,b)$$

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

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

Variable Elimination

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



Initial factors

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

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

Eliminate: *s*

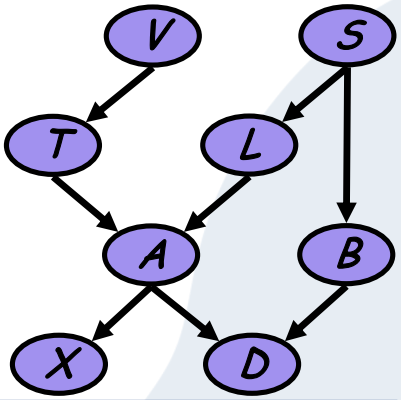
Compute: $f_s(b,l) = \sum_s P(s)P(b|s)P(l|s)$

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

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

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

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

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

Eliminate: *x*

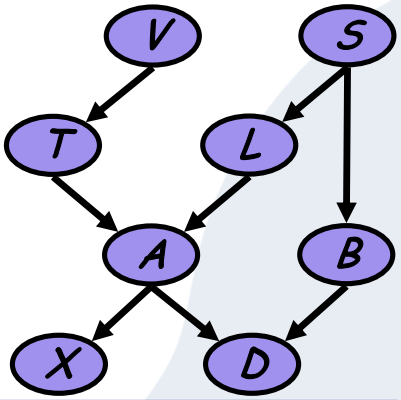
Compute: $f_x(a) = \sum_x P(x|a)$

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

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

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

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

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

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

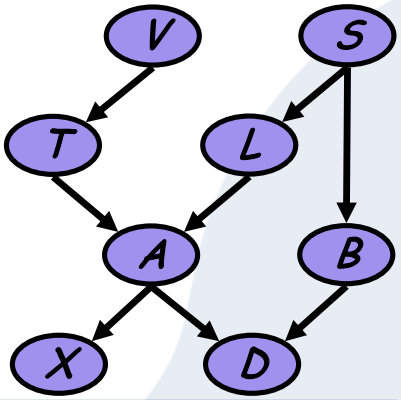
Eliminate: *t*

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

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

Variable Elimination

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



Initial factors

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

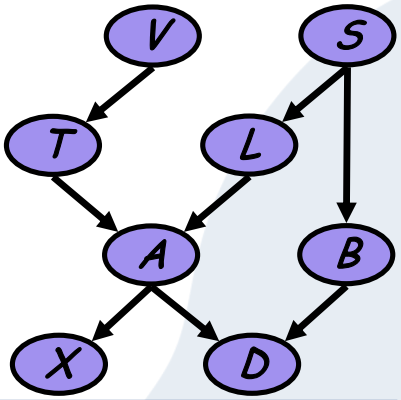
Eliminate: *l*

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

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

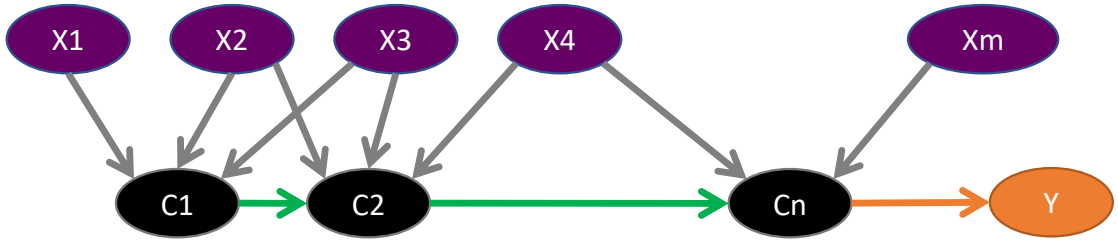
Normalize (everthing sums to 1) P(X,**e**) to obtain P(X|**e**)

Mission Completed? No …

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

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

0.5/0.5 prior

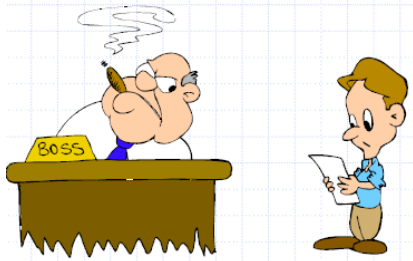


True with prob 1 iff clause 1 is satisfied

True with prob 1 iff clauses i and i-1 are satisfied

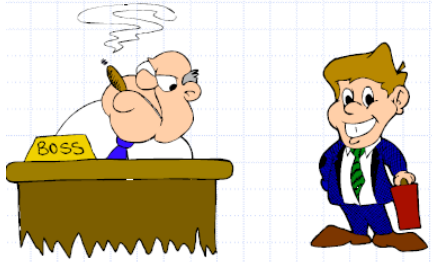
P(Y=t)>0 iff 3-SAT formula is satisfiable

- What to do when we find a problem that looks hard...



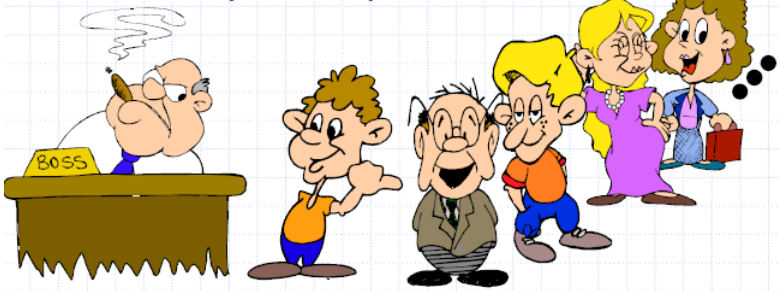
I couldn't find a polynomial-time algorithm;
I guess I'm too dumb.

- Sometimes we can prove a strong lower bound... (but not usually)

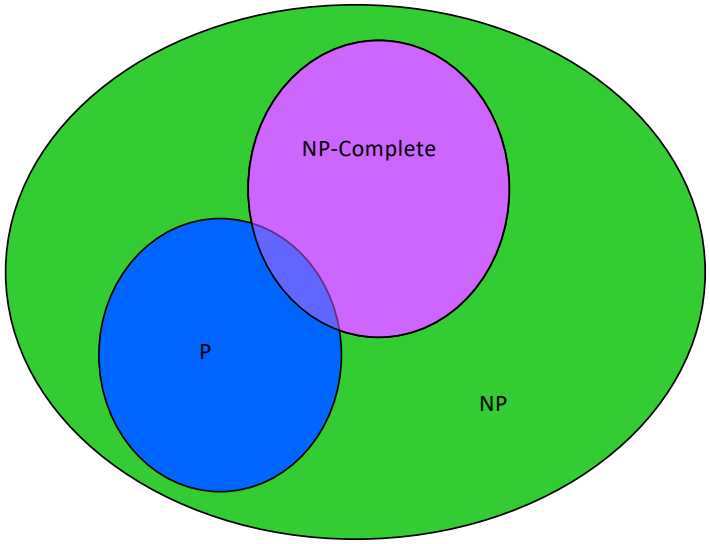


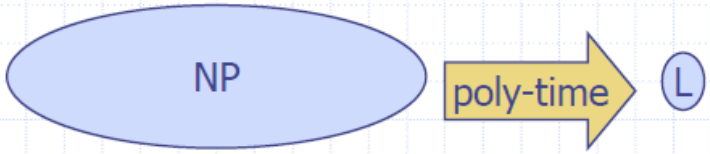
I couldn't find a polynomial-time algorithm,
because no such algorithm exists!

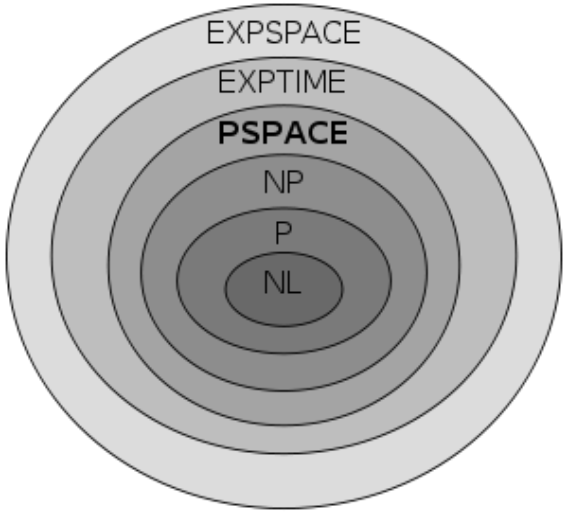
- NP-completeness let's us show collectively that a problem is hard.

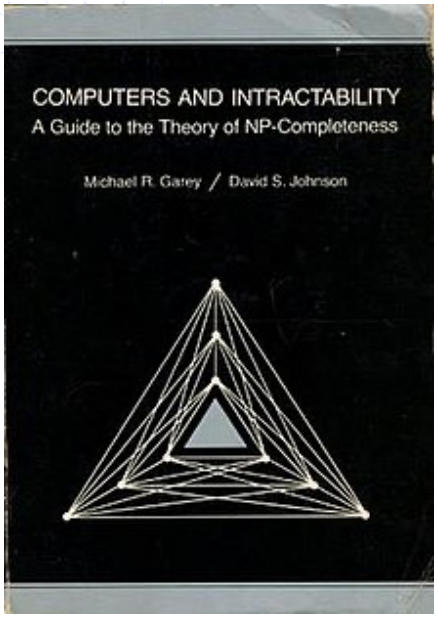


I couldn't find a polynomial-time algorithm,
but neither could all these other smart people.









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)

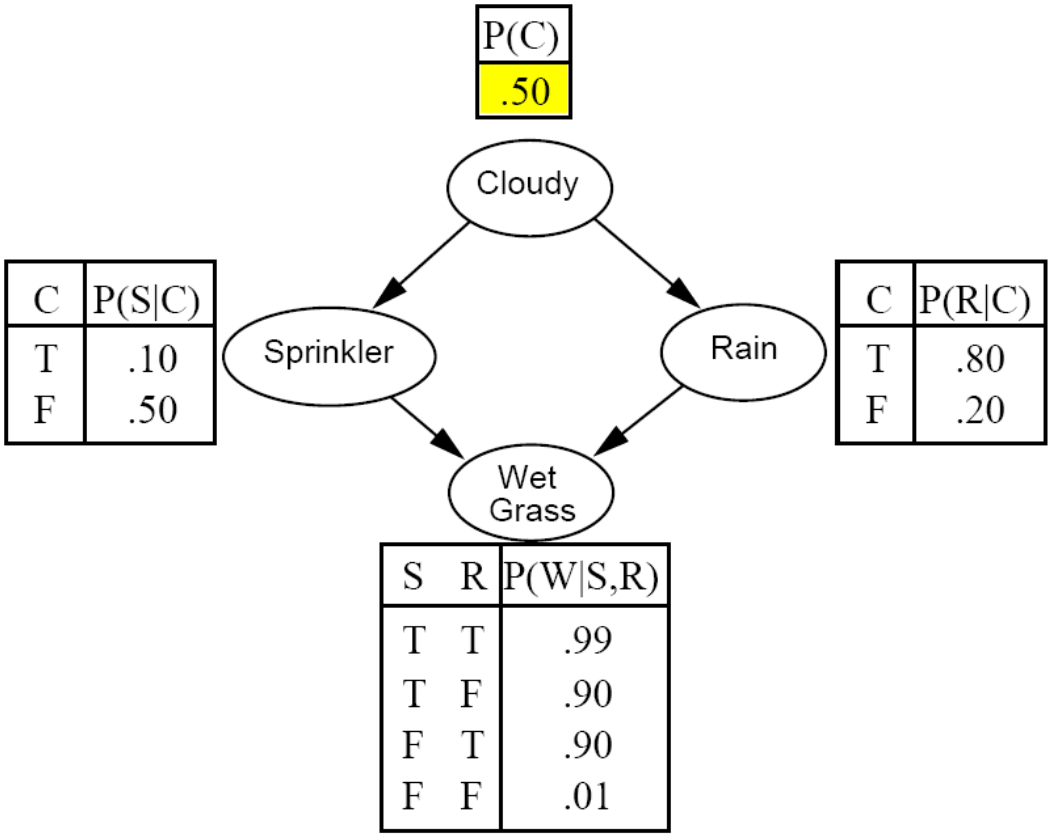
Basic idea:

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

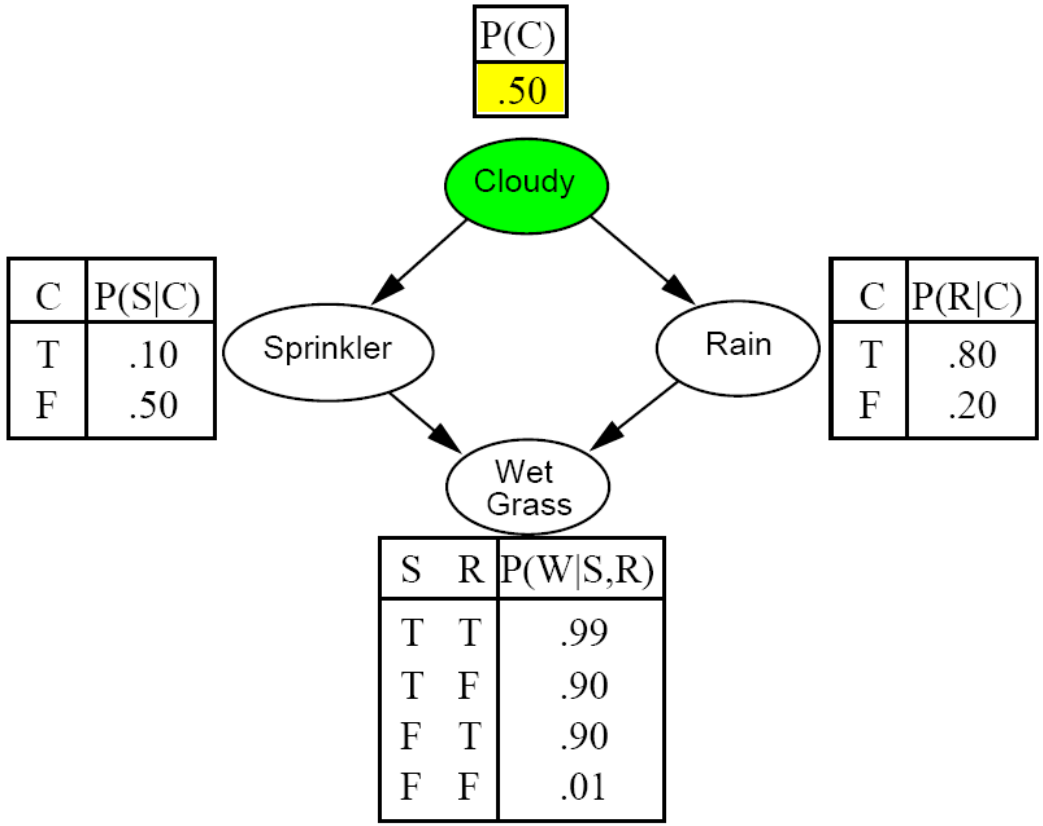
function PRIOR-SAMPLE(bn) **returns** an event sampled from bn
  inputs: bn, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

  $\mathbf{x} \leftarrow$ 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 $\mathbf{x}$

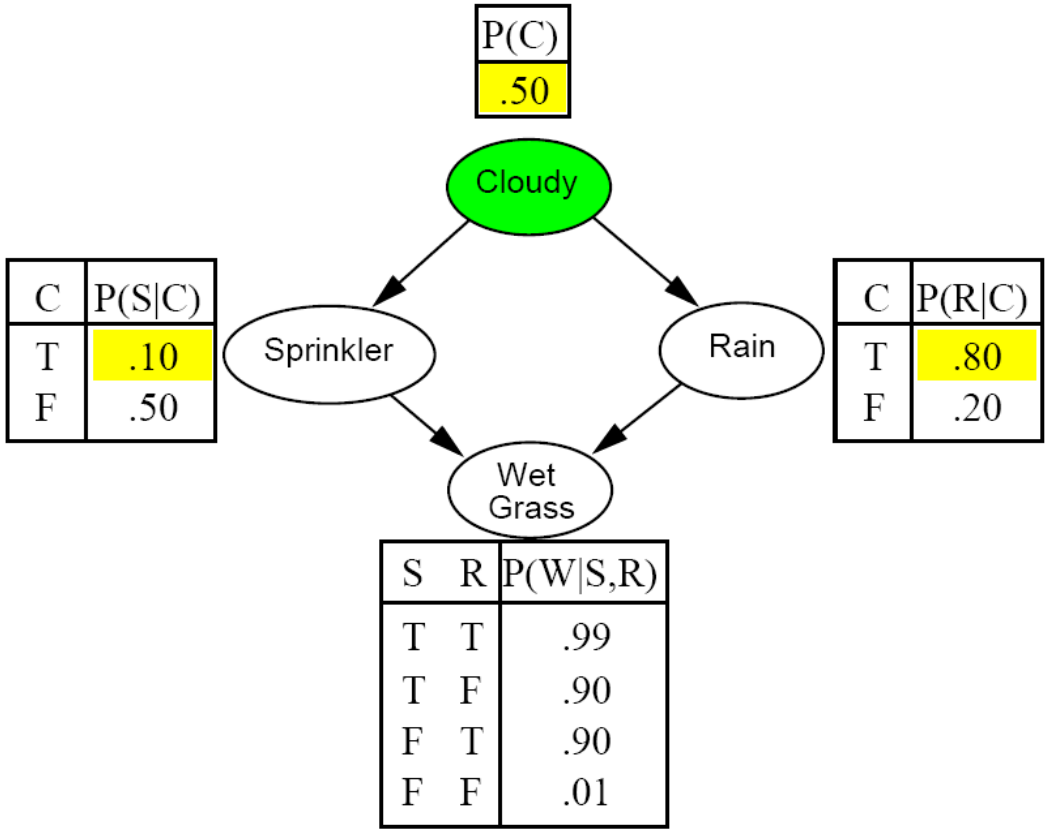
How to draw a sample?
Example



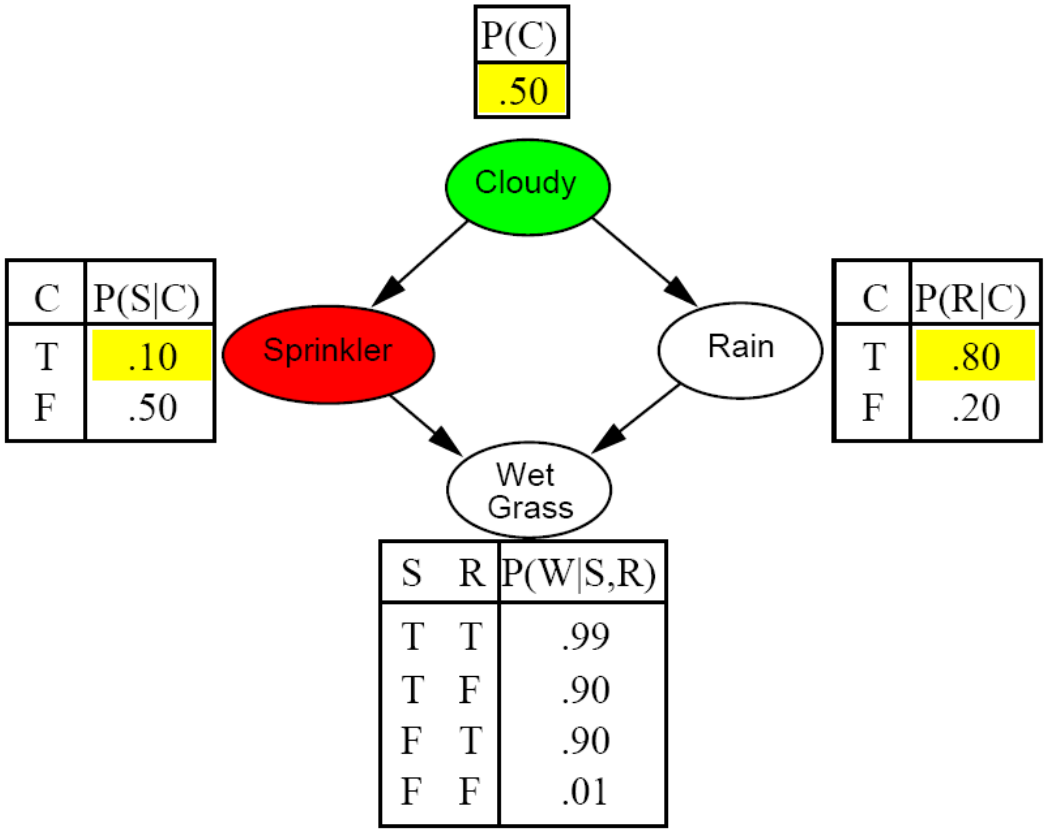
How to draw a sample?
Example



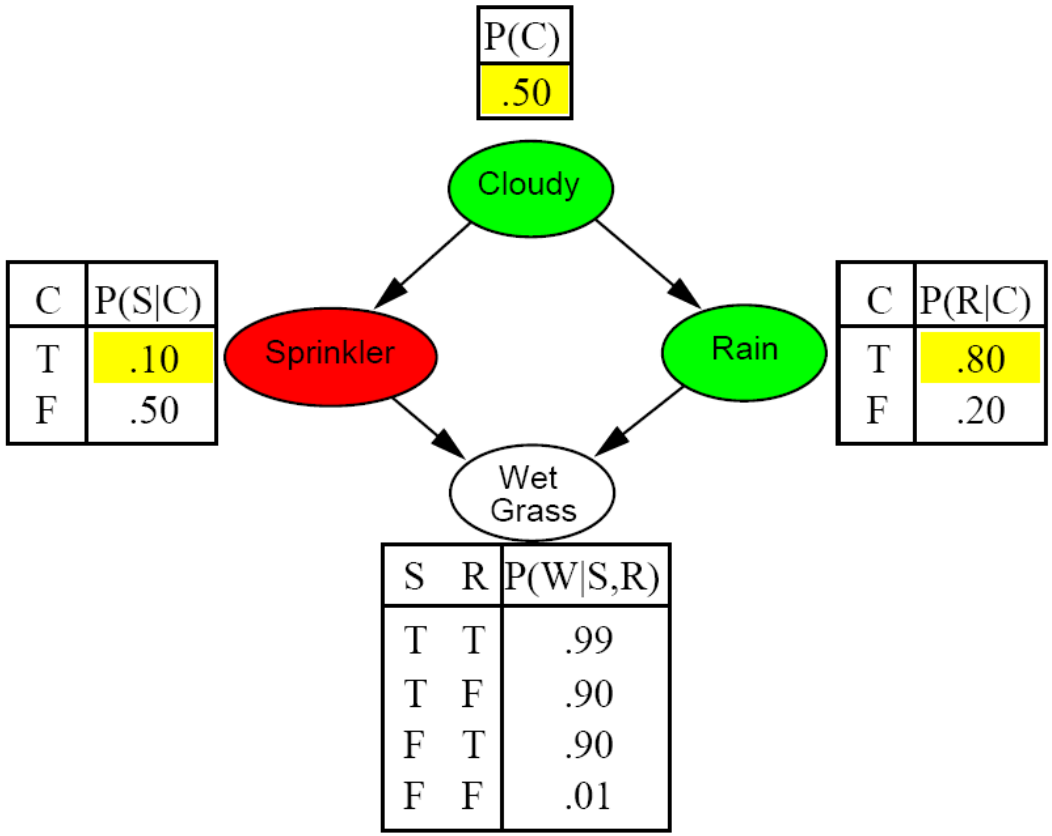
How to draw a sample?
Example



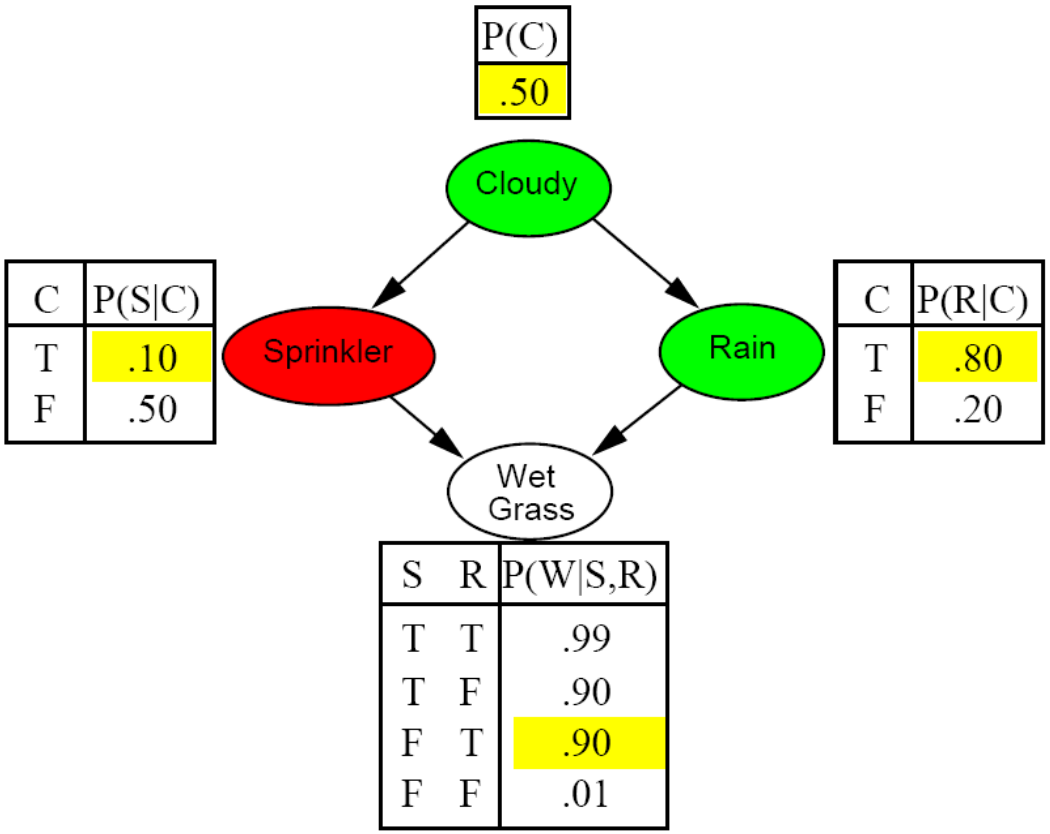
How to draw a sample?
Example



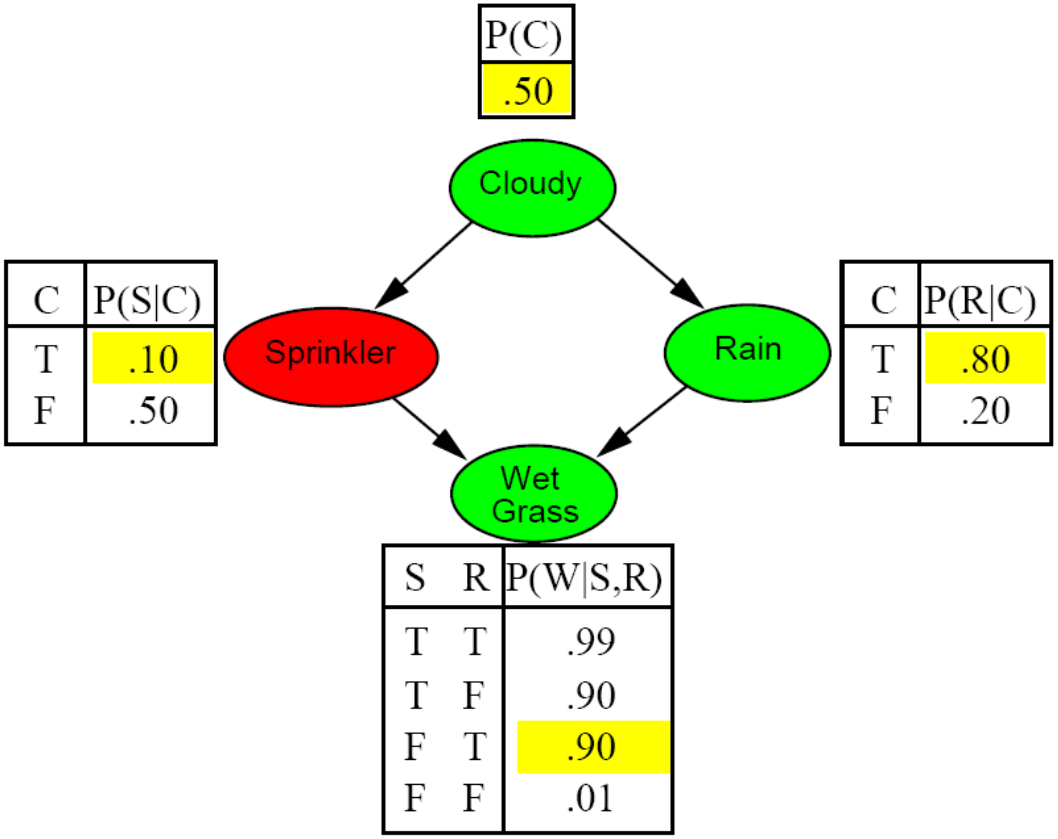
How to draw a sample?
Example



How to draw a sample?
Example



How to draw a sample?
Example



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($\mathbf{N}[X]$)

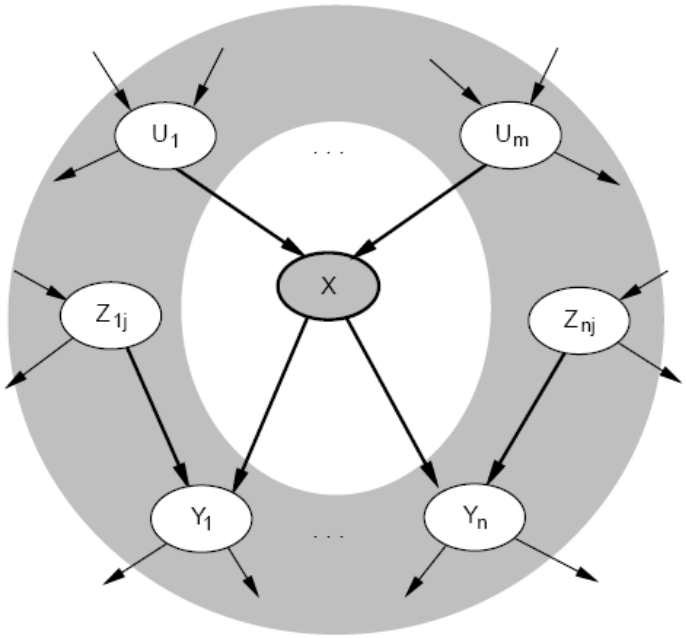
Gibbs Sampling

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



$$\mathbf{P}(X \mid U_1, \ldots, U_m, Y_1, \ldots, Y_n, Z_{1j}, \ldots, Z_{nj}) = \boldsymbol{P}(X \mid allvariables)$$

Ordered Gibbs Sampler

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

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

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

...

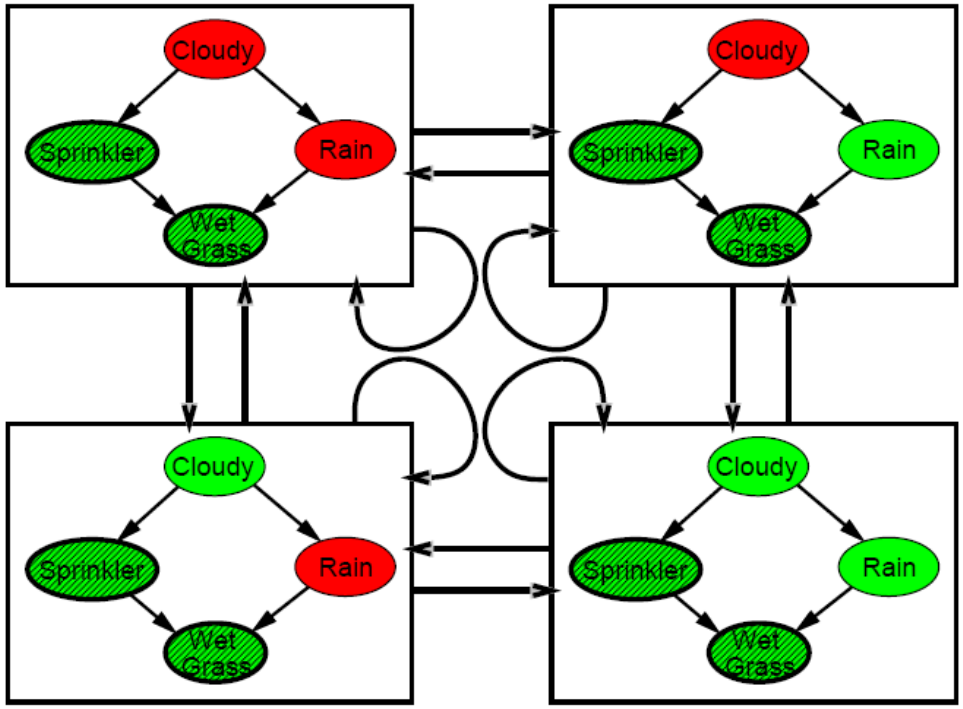
$$X_N = x_N^{t+1} \leftarrow P(x_N \mid x_1^{t+1}, x_2^{t+1}, \ldots, x_{N-1}^{t+1}, e)$$

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

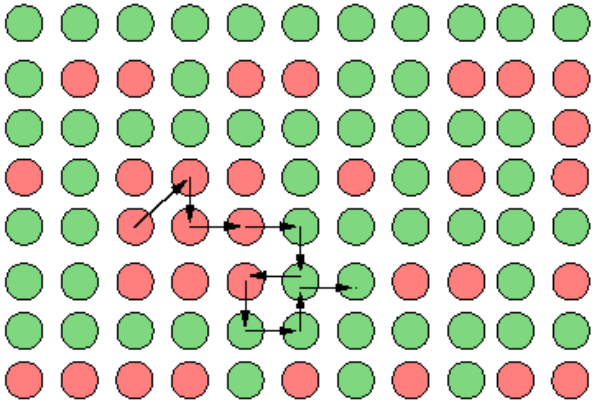
With $Sprinkler = true, WetGrass = true$, there are four states:



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.

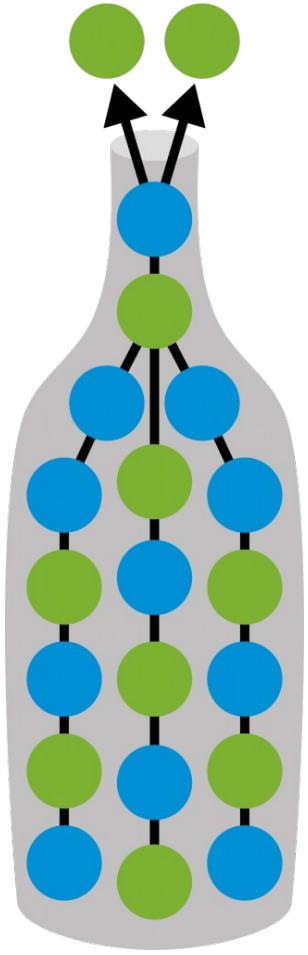
E.g. sample 100 states and count 31 times $Rain$ and 69 without $Rain$

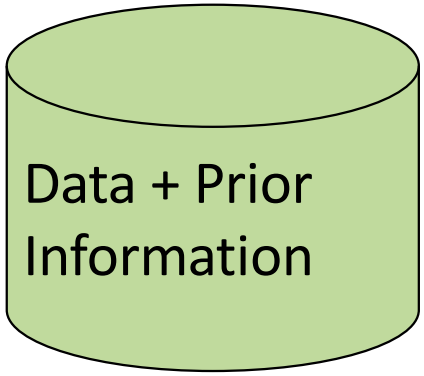
$$\hat{P}(Rain|Sprinkler = true, WetGrass = true)$$
$$= NORMALIZE(\langle 31, 69 \rangle) = \langle 0.31, 0.69 \rangle$$

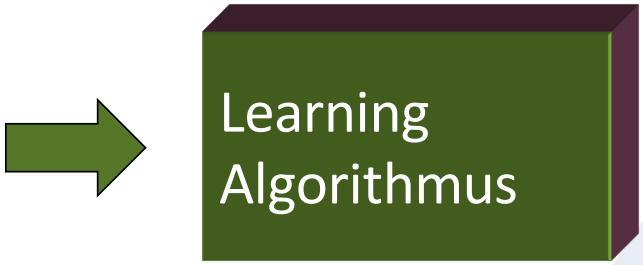
Theorem: Chain approaches stationary distribution:
long-run fraction of time spent in each state is exactly proportional to its posterior probability

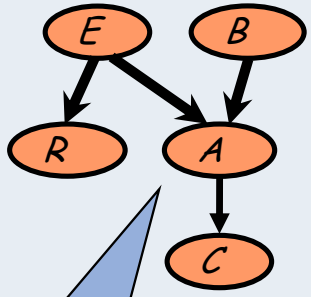
How to get a Probability Distribution from Sampling

Where do the numbers come from?









E	B	P(A | 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

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

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