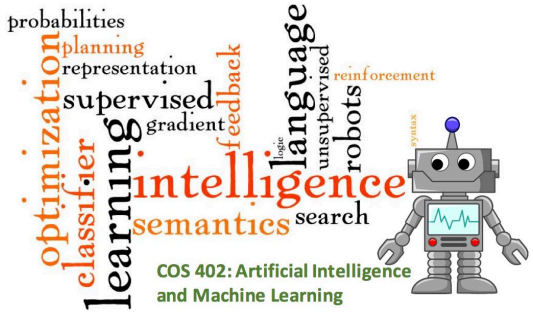


COS 402 – Machine Learning and Artificial Intelligence
Fall 2016

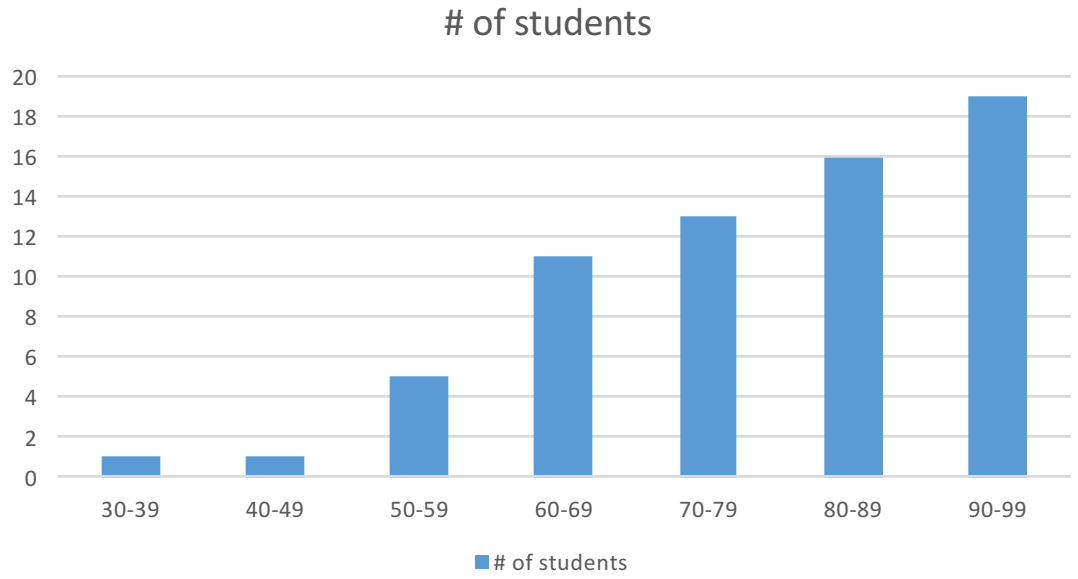
Lecture 13: Knowledge Representation and Reasoning Part 2

Sanjeev Arora Elad Hazan

PRINCETON UNIVERSITY

(Borrows from slides of Percy Liang, Stanford U.)

Midterm Histogram

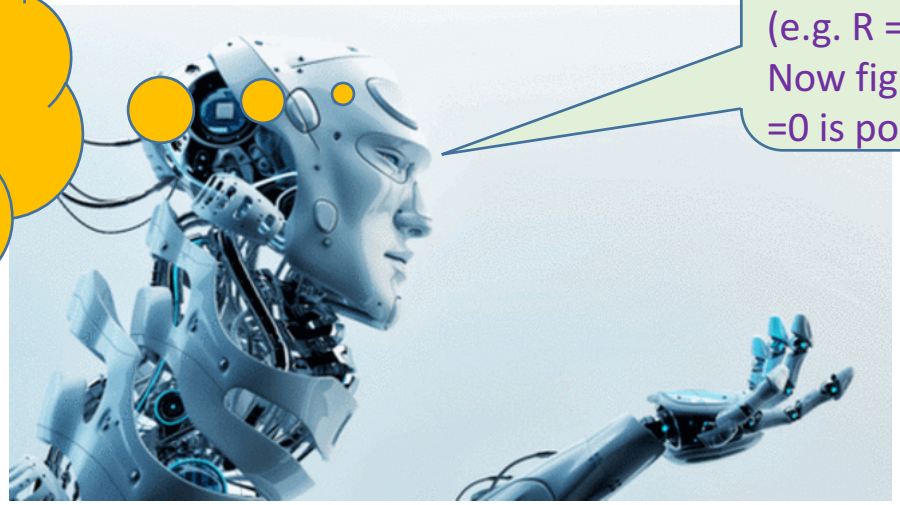


Mean = 78.42

Questions? Concerns? Pls come talk to us.
Grading mistake? Pls talk to Dr. Li and Mr. Singh first.

Last time: rule-based system based upon logic

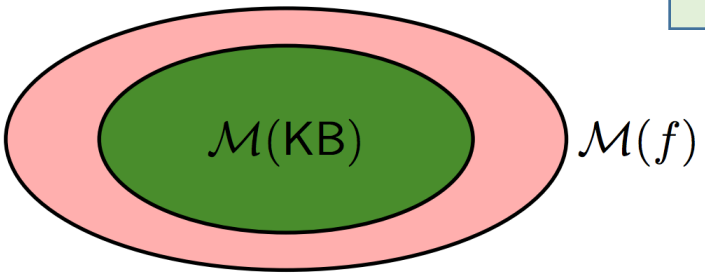
KB: $R \vee S \vee C$;
$R \rightarrow C \wedge \neg S$;
$C \leftrightarrow \neg S$
$R \rightarrow U$
$S \rightarrow \neg U$



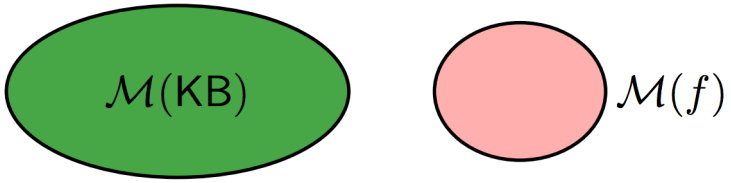
Sensors give some values to variables (e.g. R =1)
Now figure out if U =0 is possible

(contd)Venn diagram view

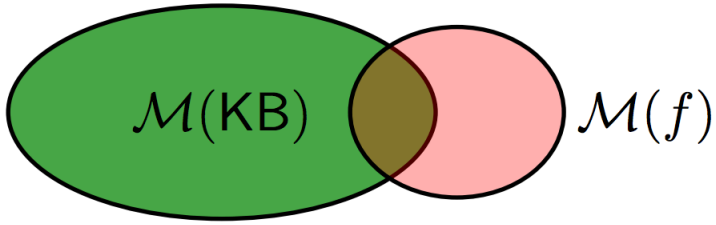
Knowledge base : Set of formulae $\{f_1, f_2, \ldots, f_n\}$
M(KB) = All possible models for $f_1 \wedge f_2 \wedge \ldots \wedge f_n$



Entailment KB $\models$ f



KB contradicts f



KB is consistent with f

(sometimes also phrased as "contingency")

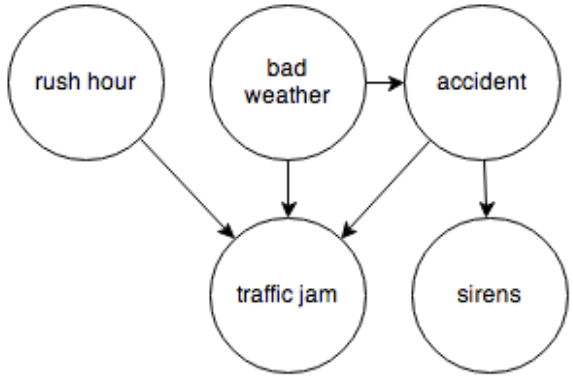
Last time: Resolution procedure

- Decides if a given KB is satisfiable (has a model)
- Can be used to decide KB $\models$ f by adding $\neg$f to KB and checking if the new KB is satisfiable. (General purpose procedure!)
- Simple algorithm; often works in practice; known to take exponential time in worst-case. (If we had an efficient algorithm that always works in practice, P = NP, which is believed to be false.)

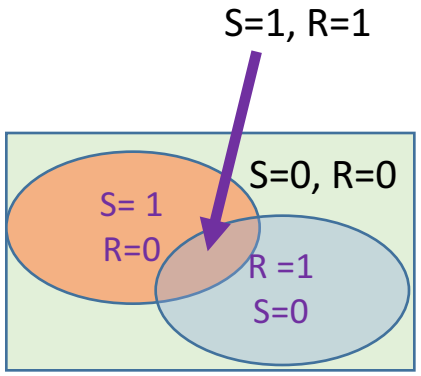
Today

Bayesian nets (aka “Belief nets” and “Graphical models”)

Can be seen as adding probabilities to logic.
(Reasoning involves numerical calculations instead of resolution)



Review: Probabilities (example)



Random variables: Sunshine $S \in \{0, 1\}$; Rainy $R \in \{0, 1\}$.

Joint Distribution

$\mathbb{P}(S, R) =$

s	r	$\mathbb{P}(S = s, R = r)$
0	0	0.20
0	1	0.08
1	0	0.70
1	1	0.02

Marginal Distribution

$\mathbb{P}(S) =$

s	$\mathbb{P}(S = s)$
0	0.28
1	0.72

Conditional Distribution

$\mathbb{P}(S \mid R = 1) =$

s	$\mathbb{P}(S = s \mid R = 1)$
0	0.8
1	0.2

Review (contd)

Random variables:

$X = (X_1, \ldots, X_n)$ partitioned into (A, B)

Joint distribution:

$\mathbb{P}(X) = \mathbb{P}(X_1, \ldots, X_n)$

Marginal distribution:

$\mathbb{P}(A) = \sum_b \mathbb{P}(A, B = b)$

Conditional distribution:

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

Example: Medical diagnosis

Patient with dyspnoea (shortness of breath), concerned about lung cancer

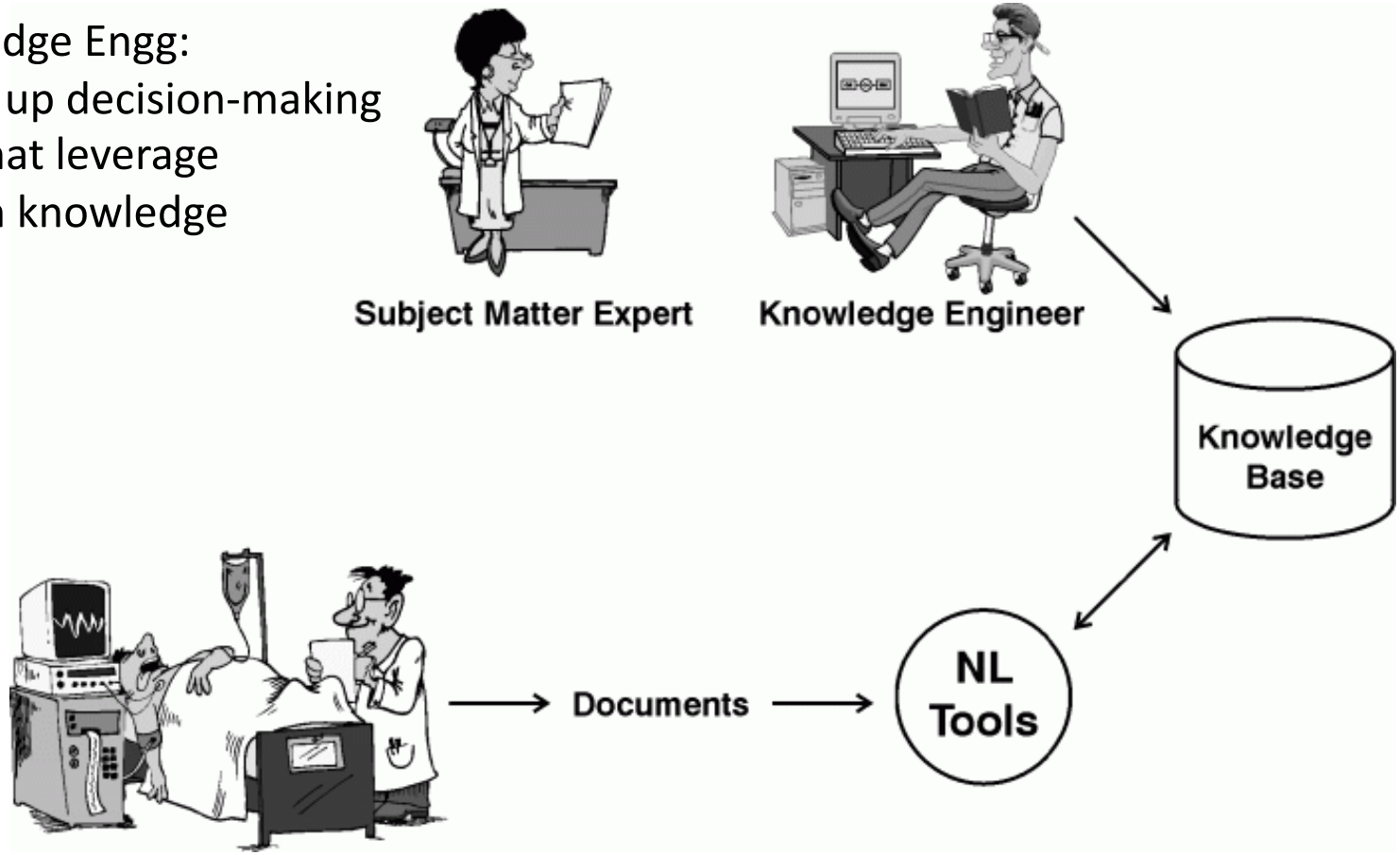
- Other explanations: bronchitis, tuberculosis
- Risk factors: smoking, exposure to air pollution
- X-ray can show formations in lungs (but similar imaging in tuberculosis)

What is a ”calculus” for estimating the chance that he has lung cancer?

Or, assuming cancer has been diagnosed, the chance that it was caused by exposure to air pollution?

[Example taken from *Bayesian Artificial Intelligence*, by Korb and Nicholson.]

Knowledge Engg:
Setting up decision-making tools that leverage domain knowledge



Last time: Knowledge engineering via propositional logic

Today: Via Bayesian nets

Main goals: (a) Compact representation.
(b) Explicit modeling of probability and "causality."

Step 1: Identify variables associated with the problem

Need quantitative versions of facts like
"Smokers are likelier to get cancer, but not all smokers get cancer."

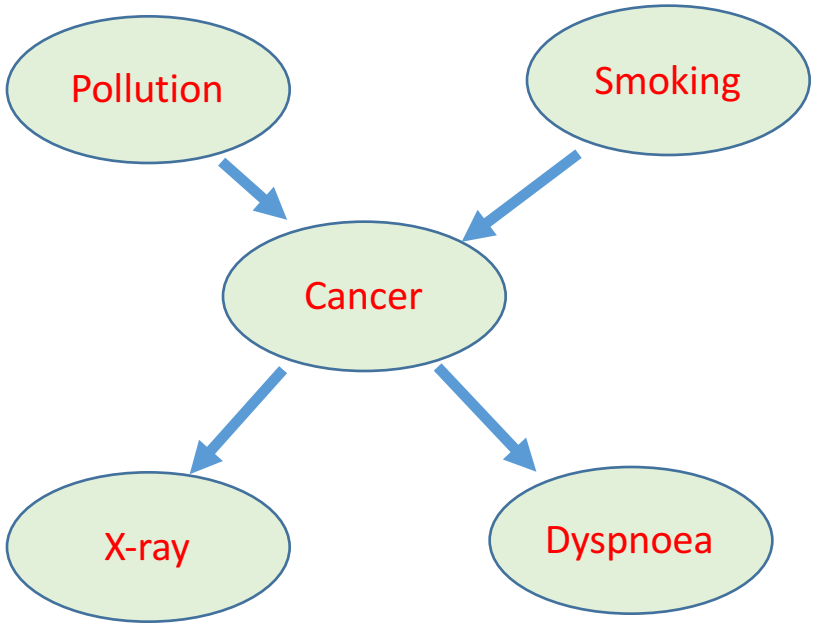
"Lung cancer can lead to the symptom Dyspnoea, but not all patients may exhibit it, and some patients who exhibit it may not have Lung Cancer.."

Node name	Type	Values
Pollution	Binary	{*low, high*}
Smoker	Boolean	{*T, F*}
Cancer	Boolean	{*T, F*}
Dyspnoea	Boolean	{*T, F*}
X-ray	Binary	{*pos, neg*}

How many numbers do you need to specify a probability distribution on 5 binary variables?

$2^5 - 1 = 31$

Step 2: Identify causality structure.

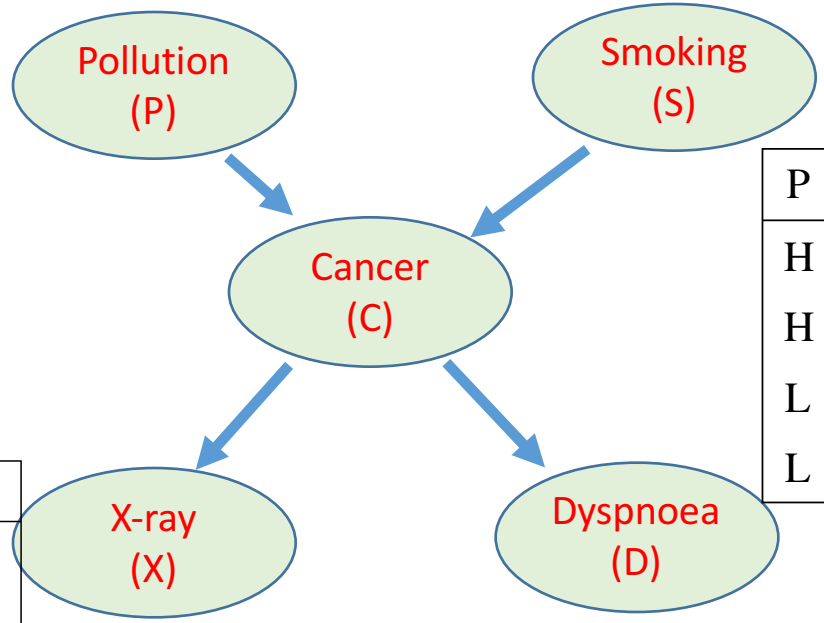


Node name	Type	Values
Pollution	Binary	{*low, high*}
Smoker	Boolean	{*T, F*}
Cancer	Boolean	{*T, F*}
Dyspnoea	Boolean	{*T, F*}
X-ray	Binary	{*pos, neg*}

Step 3: Put in conditional probabilities.

(estimated from patient studies)

Node name	Type	Values
Pollution	Binary	{*low, high*}
Smoker	Boolean	{*T, F*}
Cancer	Boolean	{*T, F*}
Dyspnoea	Boolean	{*T, F*}
X-ray	Binary	{*pos, neg*}



P(P=L)
0.90

P(S=T)
0.30

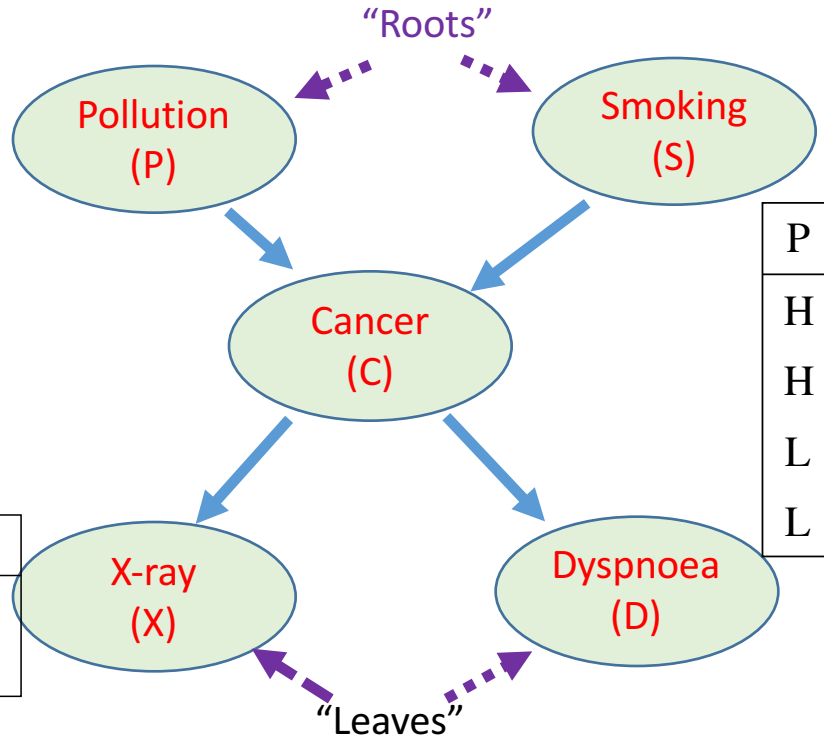
P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

NB: Bayes net has no directed cycle (why?)

Node name	Type	Values
Pollution	Binary	{*low, high*}
Smoker	Boolean	{*T, F*}
Cancer	Boolean	{*T, F*}
Dyspnoea	Boolean	{*T, F*}
X-ray	Binary	{*pos, neg*}



P(P=L)
0.90

P(S=T)
0.30

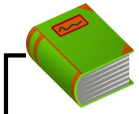
P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

Roots are all independent of each other.

Bayesian Net: Formal Definition



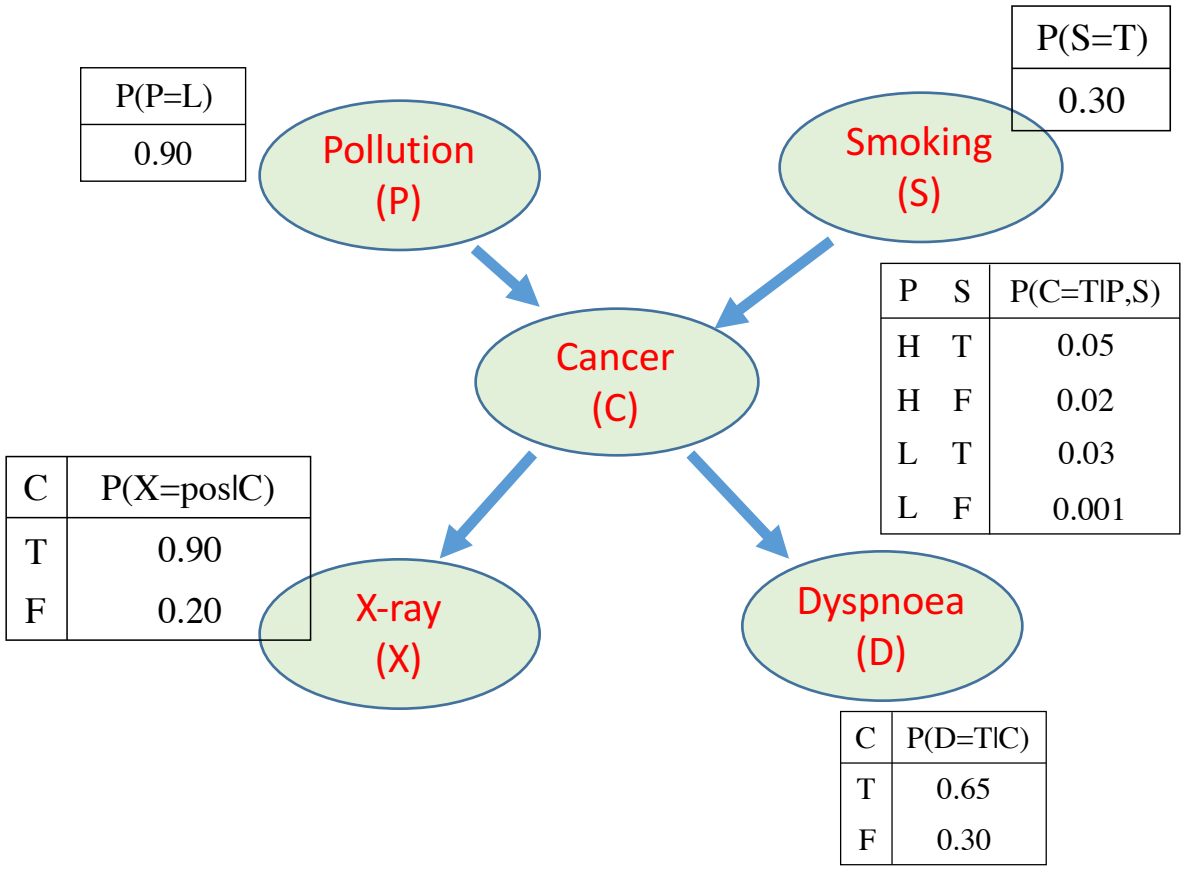
Definition: Bayesian network

Let $X = (X_1, \ldots, X_n)$ be random variables.

A **Bayesian network** is a directed acyclic graph (DAG) that specifies a joint distribution over X as a product of local conditional distributions, one for each node:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i \mid x_{\text{Parents}(i)})$$

Example revisited



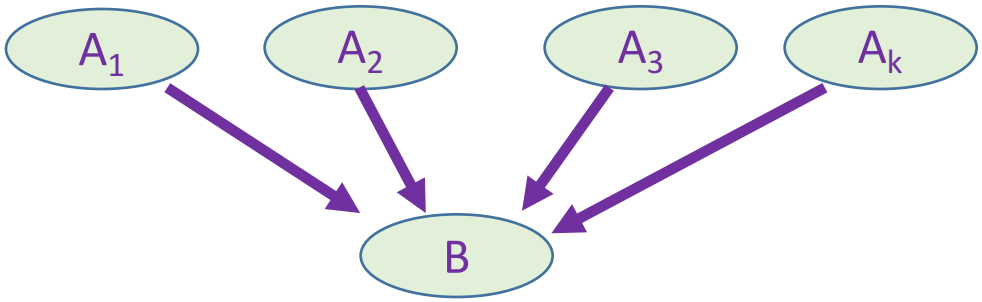
$\Pr[X, D, C, P, S]$

$= \Pr[X|C]\ \Pr[D|C]\ \ \Pr[C|P, S]\ \Pr[S]\ \Pr[C]$

Note: Distribution on
5 boolean variables; specified
using only 10 numbers

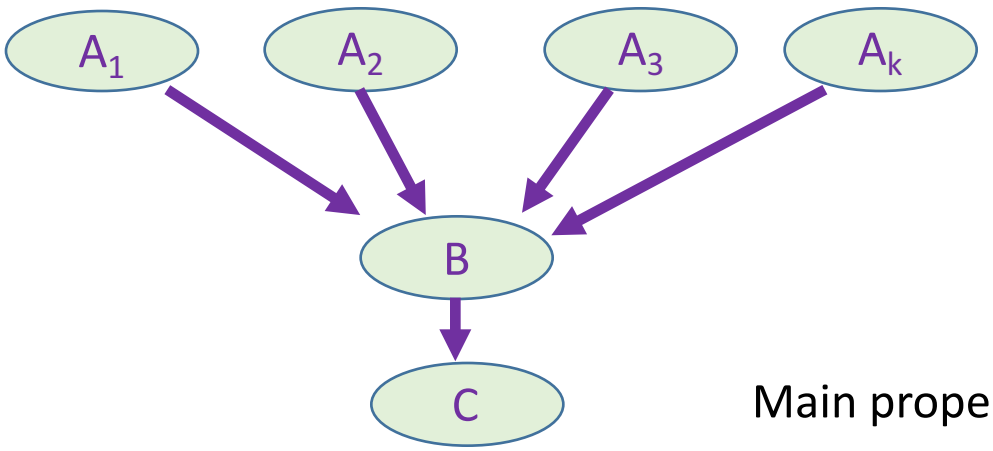
(instead of the trivial
$2^5 - 1 = 31$ numbers.)

Conditional probability table



- B depends upon all other variables only through A_1, .. A_k
- CPD at B gives Prob. B=1 or B=0 conditioned on all 2^k combinations of A_1, .. A_k
- Note: Sum of probabilities of all events in CPD 1.

Conditional probability table



- B depends upon all other variables only through $A_1, .. A_k$
- CPD gives Prob of B conditioned on all 2^k combinations of $A_1, .. A_k$

Main property: $Pr[C| B, A_1, .. A_k] = Pr[C|B]$

No way for A_1 to "cause" C except via B
(causation/influence travels via directed paths)

Key idea: locally normalized

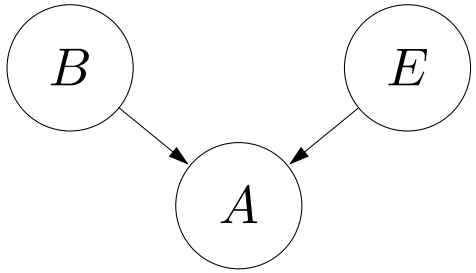
All factors (local conditional distributions) satisfy:

$$\sum_{x_i} p(x_i \mid x_{\text{Parents}(i)}) = 1 \text{ for each } x_{\text{Parents}(i)}$$

Implications:

- Consistency of sub-Bayesian networks
- Consistency of conditional distributions

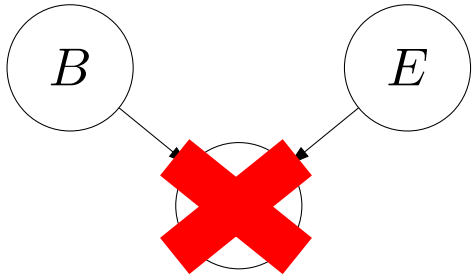
Consistency of sub-Bayesian networks



A short calculation:

$$\begin{aligned}
\mathbb{P}(B = b, E = e) &= \textstyle\sum_a \mathbb{P}(B = b, E = e, A = a) \\
&= \textstyle\sum_a p(b)p(e)p(a \mid b, e) \\
&= p(b)p(e) \textstyle\sum_a p(a \mid b, e) \\
&= p(b)p(e)
\end{aligned}$$

Consistency of sub-Bayesian networks



Suppose we remove a leaf node A.

A short calculation:

$$\mathbb{P}(B = b, E = e) = \sum_a \mathbb{P}(B = b, E = e, A = a)$$

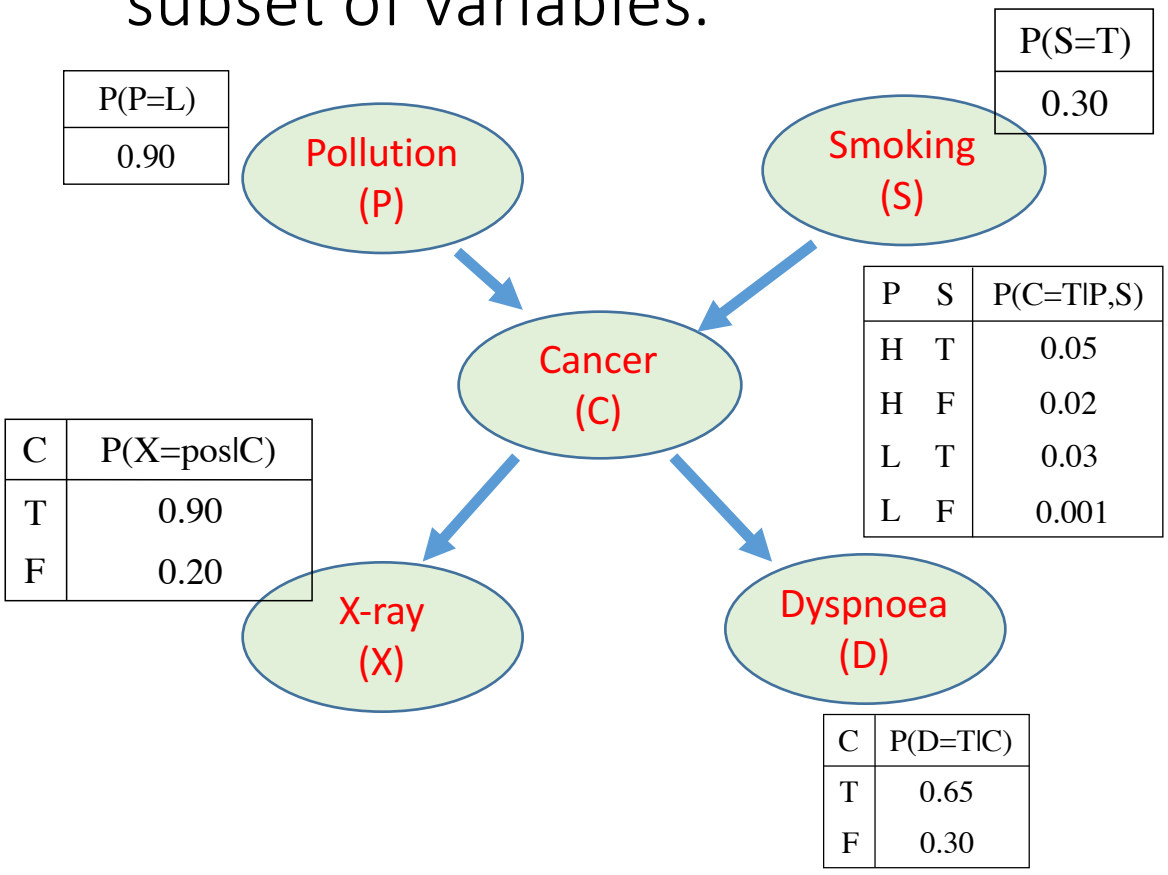
$$= \sum_a p(b)p(e)p(a \mid b, e)$$

$$= p(b)p(e) \sum_a p(a \mid b, e)$$

$$= p(b)p(e)$$

Marginalization gives new bayes net where B, E are independent with prob. p(b), p (e)

Marginalization: Compute subdistributions on some subset of variables.



$Pr[X, D, C, P, S]$

$= Pr[X|C]\ Pr[D|C]\ \ Pr[C|P, S]\ Pr[S]\ Pr[C]$

$Pr[C=0] = Pr[C=0, P=0, S=0] + Pr[C=0, P=0, S=1] + Pr[C=0, P=1, S=0] + Pr[C=0, P=1, S=1]$

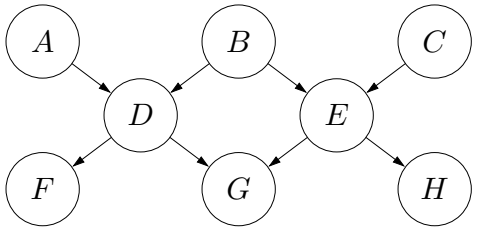
Disjoint Events

$= \sum_{b=0,1\ d=0,1} Pr[C=0, P=b, S=d]$

I've hinted already but will not prove formally…

Bayes nets define proper distributions, in the sense that
all marginal distributions are well-defined (meaning probabilities sum to 1).

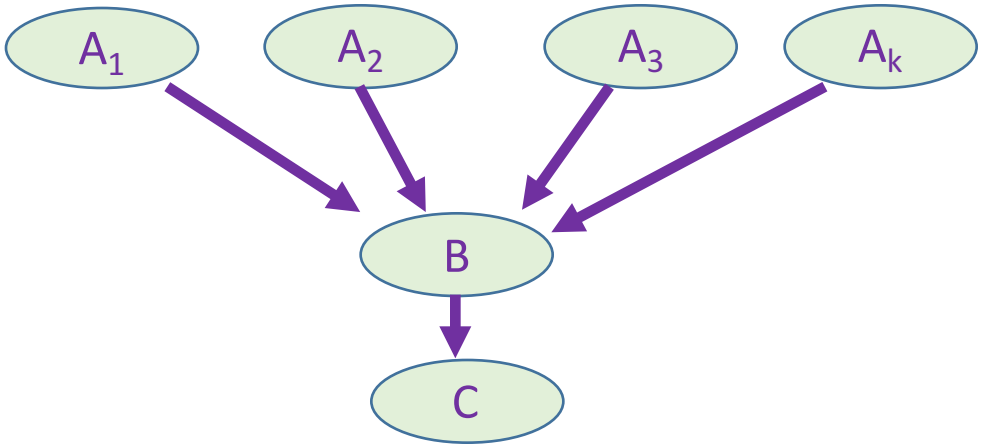
e.g,



$$\underbrace{\mathbb{P}(D = d \mid A = a, B = b)}_{\text{From marginalization calculation}} = \underbrace{p(d \mid a, b)}_{\text{local factor}}$$

From CPD table

Bayes nets as models of probabilistic processes



Step 1: Coins tossed at each A_i node to decide if A_i happened.
($Pr[Heads] = Pr[A_i = 1]$)

Step 2: Coins tossed at B node to decide if B =1. (Pr[Heads] looked up from CPD table)

Step 3: Coins tossed at C node to decide if C =1.

Uses: models of language/text, social processes, disease propagation etc.

Example: Bayes net for a 5-word sentence according to bigram model.

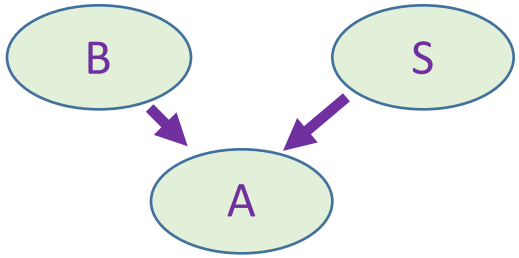
Instead of binary variables, use N-ary variables (N = # Words in Dictionary)

$$Pr[W_i \mid W_1 W_2 \ldots W_{i-1}] = Pr[W_i \mid W_{i-1}]$$

"Explaining away" phenomenon

(Aka Berkson's paradox, and "selection bias.")

Whiz-bang U admits students who are either Brainy (B) or Sporty (S) (or both)



$$\Pr[S=1 \mid B=1, A=1] \leq \Pr[S=1 \mid A=1]$$

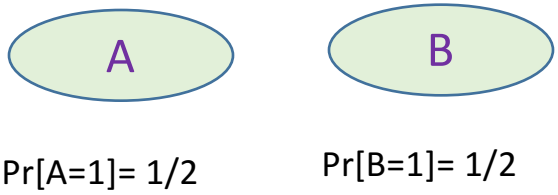
Conditional on having been admitted:
Being brainy implies you are less likely to be sporty (and vice versa)

Key idea: explaining away

Suppose two causes positively influence an effect. Conditioned on the effect, conditioning on one cause reduces the probability of the other cause.

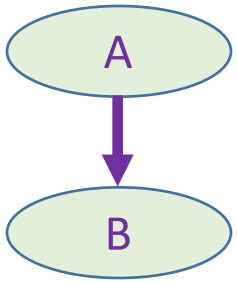
Nonuniqueness of bayes net

Suppose A, B are independent coin tosses.
Following bayes nets are all correct



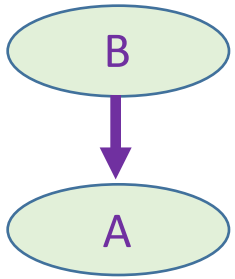
Pr[A=1]= 1/2

Pr[B=1]= 1/2



Pr[A=1]= 1/2

Pr[B =1| A]= 1/2



Pr[B=1]= 1/2

Pr[A =1| B]= 1/2

General lesson: “Causality” is not easy to pin down from correlation.

Next time: Doing calculations/predictions with bayesian nets.