Bayesian Networks

Courtesy: Weng-Keen Wong, Oregon State University

Introduction

Suppose you are trying to determine if a patient has inhalational anthrax. You observe the following symptoms:

- The patient has a cough

- The patient has a fever

- The patient has difficulty breathing

Introduction

You would like to determine how likely the patient is infected with inhalational anthrax given that the patient has a cough, a fever, and difficulty breathing

We are not 100% certain that the patient has anthrax because of these symptoms. We are dealing with uncertainty!
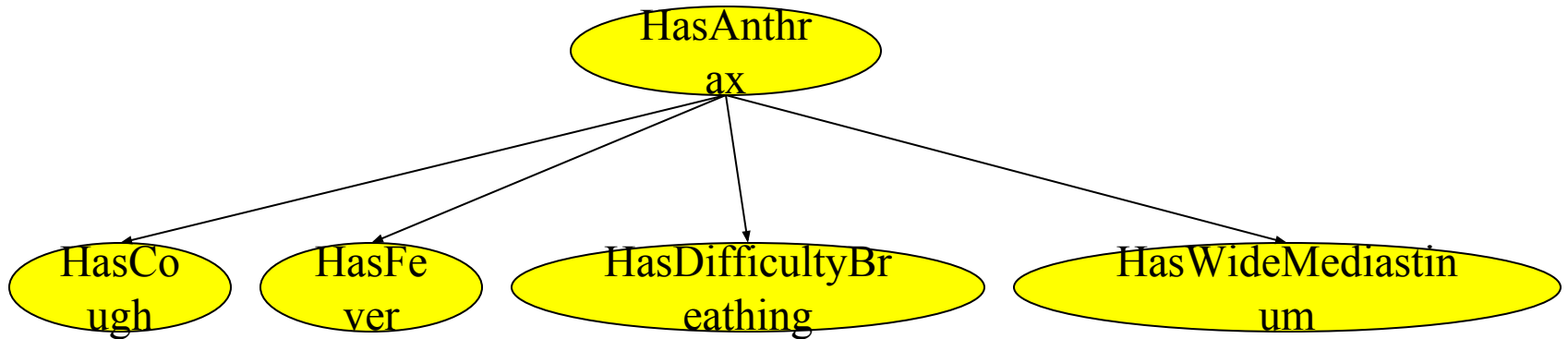
Introduction

Now suppose you order an x-ray and observe that the patient has a wide mediastinum.

Your belief that that the patient is infected with inhalational anthrax is now much higher.
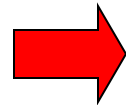
Introduction

- In the previous slides, what you observed affected your belief that the patient is infected with anthrax

- This is called reasoning with uncertainty

- Wouldn't it be nice if we had some methodology for reasoning with uncertainty? Well in fact, we do…

Bayesian Networks

- Bayesian networks are used in many applications eg. spam filtering, speech recognition, robotics, diagnostic systems and even syndromic surveillance

Outline

1. Introduction
2. Probability Primer
3. Bayesian networks

Probability Primer: Random Variables

- A random variable is the basic element of probability

- Refers to an event and there is some degree of uncertainty as to the outcome of the event

- For example, the random variable A could be the event of getting a head on a coin flip

Boolean Random Variables

- We will start with the simplest type of random variables – Boolean ones
- Take the values *true* or *false*
- Think of the event as occurring or not occurring
- Examples (Let A be a Boolean random variable):

  A = Getting a head on a coin flip

  A = It will rain today

The Joint Probability Distribution

- Joint probabilities can be between any number of variables

  eg. *P(A = true, B = true, C = true)*

- For each combination of variables, we need to say how probable that combination is

- The probabilities of these combinations need to sum to 1

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Sums to 1

The Joint Probability Distribution

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

- Once you have the joint probability distribution, you can calculate any probability involving *A*, *B*, and *C*
- Note: May need to use marginalization and Bayes rule, (both of which are not discussed in these slides)

Examples of things you can compute:

- *P(A=true)* = sum of *P(A,B,C)* in rows with *A=true*

- *P(A=true, B = true | C=true) =*

P(A = true, B = true, C = true) / P(C = true)

11

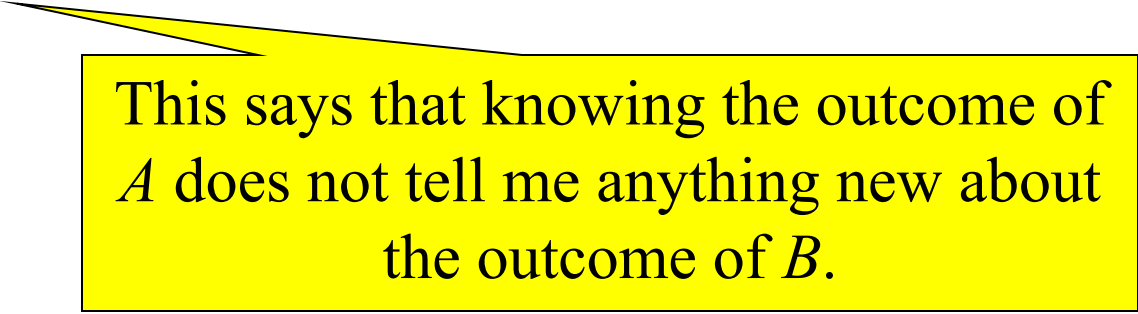# The Problem with the Joint Distribution

- Lots of entries in the table to fill up!
- For *k* Boolean random variables, you need a table of size 2^k
- How do we use fewer numbers? Need the concept of independence

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Independence

Variables *A* and *B* are independent if any of the following hold:

- $P(A,B) = P(A)\,P(B)$

- $P(A \mid B) = P(A)$

- $P(B \mid A) = P(B)$

This says that knowing the outcome of *A* does not tell me anything new about the outcome of *B*.

Independence

How is independence useful?

- Suppose you have n coin flips and you want to calculate the joint distribution $P(C_1, \ldots, C_n)$

- If the coin flips are not independent, you need 2^n values in the table

- If the coin flips are independent, then

$$P(C_1, \ldots, C_n) = \prod_{i=1}^{n} P(C_i)$$

Each $P(C_i)$ table has 2 entries and there are n of them for a total of $2n$ values
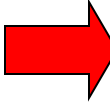
Conditional Independence

Variables *A* and *B* are conditionally independent given *C* if any of the following hold:

- $P(A, B \mid C) = P(A \mid C) \, P(B \mid C)$
- $P(A \mid B, C) = P(A \mid C)$
- $P(B \mid A, C) = P(B \mid C)$

Knowing *C* tells me everything about *B*. I don't gain anything by knowing *A* (either because *A* doesn't influence *B* or because knowing *C* provides all the information knowing *A* would give)
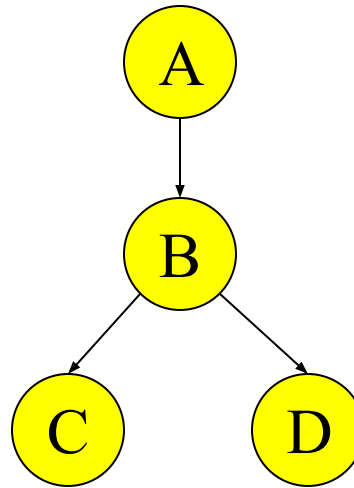
Outline

1. Introduction
2. Probability Primer
3. Bayesian networks

A Bayesian Network

A Bayesian network is made up of:

1. A Directed Acyclic Graph

2. A set of tables for each node in the graph

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

A Directed Acyclic Graph

Each node in the graph is a random variable

A node *X* is a parent of another node *Y* if there is an arrow from node *X* to node *Y* eg. *A* is a parent of *B*
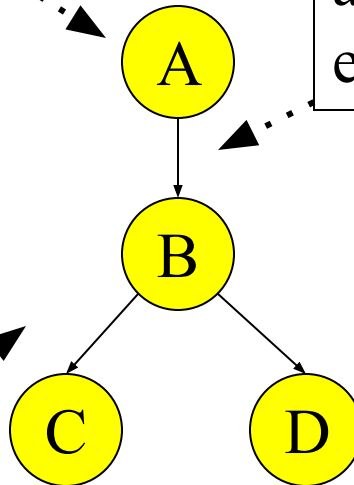
Informally, an arrow from node *X* to node *Y* means *X* has a direct influence on *Y*
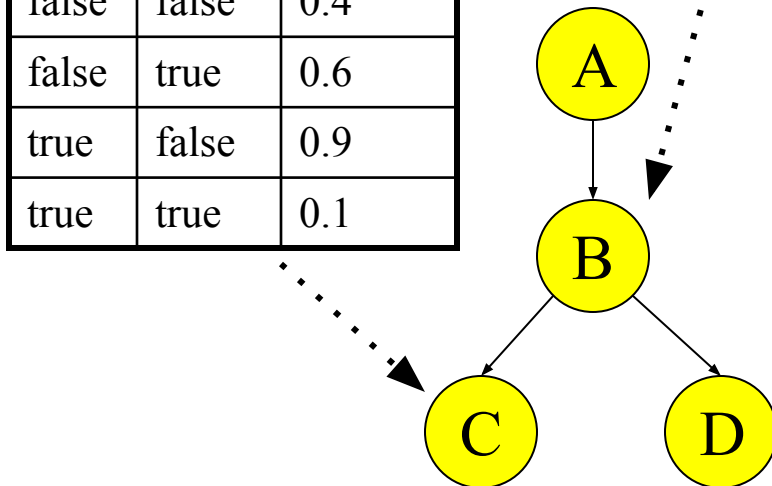
A Set of Tables for Each Node

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

Each node X_i has a conditional probability distribution $P(X_i \mid \text{Parents}(X_i))$ that quantifies the effect of the parents on the node

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

The parameters are the probabilities in these conditional probability tables (CPTs)

A

B

C   D
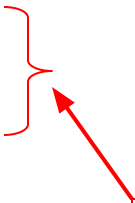
B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

A Set of Tables for Each Node

Conditional Probability
Distribution for C given B

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

For a given combination of values of the parents (B in this example), the entries for P(C=true | B) and P(C=false | B) must add up to 1
eg. P(C=true | B=false) + P(C=false |B=false)=1

If you have a Boolean variable with k Boolean parents, this table has 2^{k+1} probabilities (but only 2^k need to be stored)
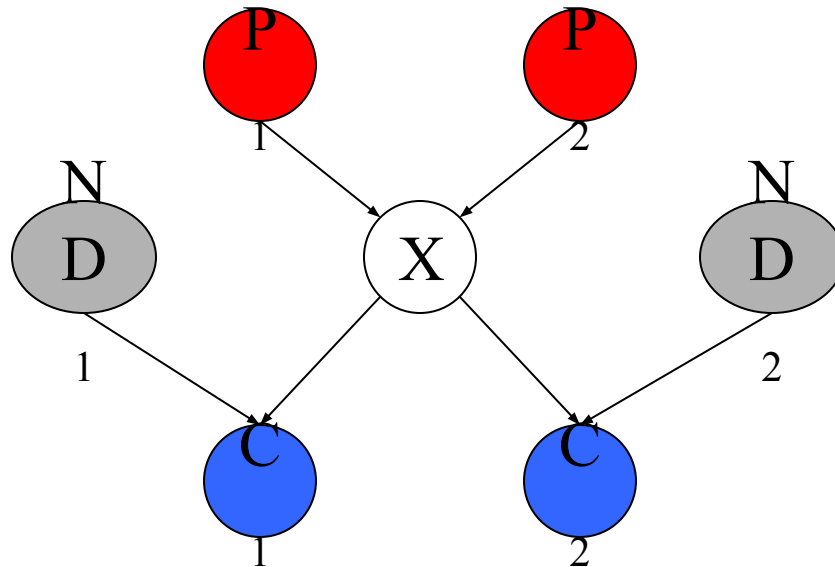
Bayesian Networks

Two important properties:

1. Encodes the conditional independence relationships between the variables in the graph structure

2. Is a compact representation of the joint probability distribution over the variables

Conditional Independence

The Markov condition: given its parents (P_1, P_2), a node (X) is conditionally independent of its non-descendants (ND_1, ND_2)

The Joint Probability Distribution

Due to the Markov condition, we can compute the joint probability distribution over all the variables X_1, ..., X_n in the Bayesian net using the formula:
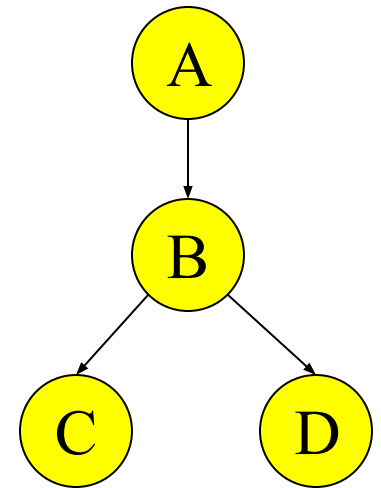
$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid Parents(X_i))$$

Where Parents(X_i) means the values of the Parents of the node X_i with respect to the graph

Using a Bayesian Network Example

Using the network in the example, suppose you want to calculate:

P(A = true, B = true, C = true, D = true)

= P(A = true) * P(B = true | A = true) *

  P(C = true | B = true) P(D = true | B = true)

= (0.4)*(0.3)*(0.1)*(0.95)

Using a Bayesian Network Example

Using the network in the example, suppose you want to calculate:

P(A = true, B = true, C = true, D = true)

= P(A = true) * P(B = true | A = true) *

  P(C = true | B = true) P(D = true | B = true)

= (0.4)*(0.3)*(0.1)*(0.95)

This is from the graph structure

These numbers are from the conditional probability tables
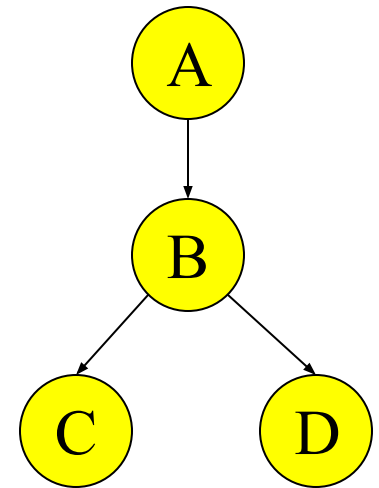
Joint Probability Factorization

For any joint distribution of random variables the following factorization is always true:

$$P(A,B,C,D) = P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)$$

We derive it by repeatedly applying the Bayes' Rule P(X,Y)=P(X|Y)P(Y):

$$P(A,B,C,D) = P(B,C,D \mid A)P(A)$$
$$= P(C,D \mid B,A)P(B \mid A)P(A)$$
$$= P(D \mid C,B,A)P(C \mid B,A)P(B \mid A)P(A)$$
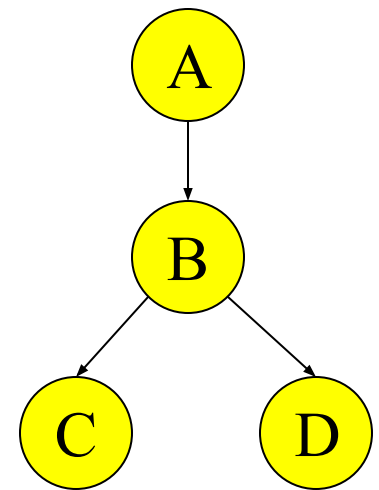$$P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)$$

Joint Probability Factorization

Our example graph carries additional independence information, which simplifies the joint distribution:

$$P(A,B,C,D) = P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)$$
$$= P(A)P(B \mid A)P(C \mid B)P(D \mid B)$$

This is why, we only need the tables for

P(A), P(B|A), P(C|B), and P(D|B)

and why we computed

P(A = true, B = true, C = true, D = true)

= P(A = true) * P(B = true | A = true) *

P(C = true | B = true) P(D = true | B = true)

= (0.4)*(0.3)*(0.1)*(0.95)

Inference

- Using a Bayesian network to compute probabilities is called inference
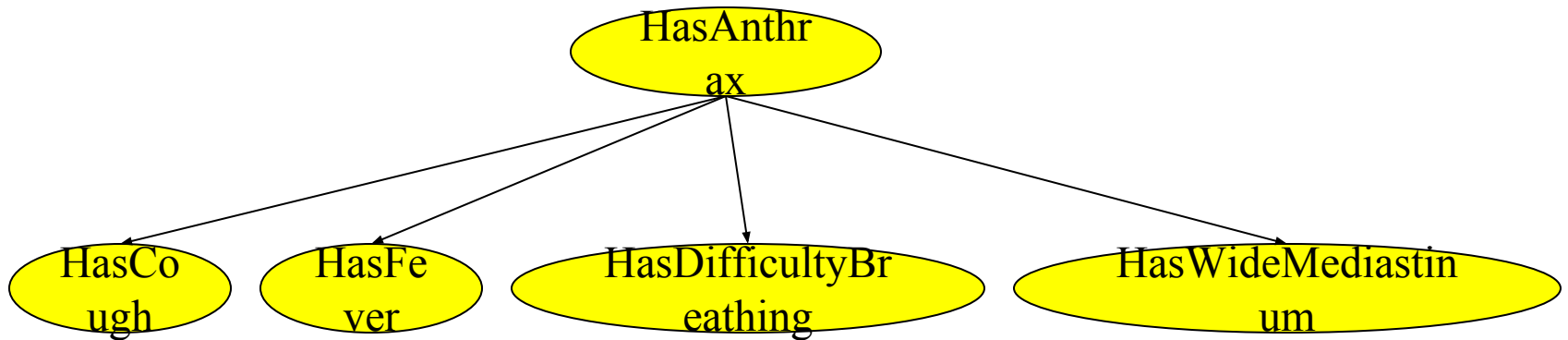- In general, inference involves queries of the form:

$P(X \mid E)$

E = The evidence variable(s)

X = The query variable(s)

Inference

- An example of a query would be:

  P(*HasAnthrax = true | HasFever = true, HasCough = true*)

- Note: Even though *HasDifficultyBreathing* and *HasWideMediastinum* are in the Bayesian network, they are not given values in the query (ie. they do not appear either as query variables or evidence variables)

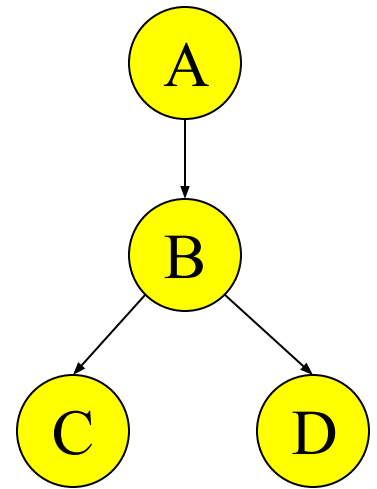- They are treated as unobserved variables and summed out.

Inference Example

Supposed we know that A=true.
What is more probable C=true or D=true?
For this we need to compute
P(C=t | A =t) and P(D=t | A =t).
Let us compute the first one.

$$P(C = t \mid A = t) = \frac{P(A = t, C = t)}{P(A = t)} = \frac{\displaystyle\sum_{b,d} P(A = t, B = b, C = t, D = d)}{P(A = t)}$$

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

30

What is P(A=true)?
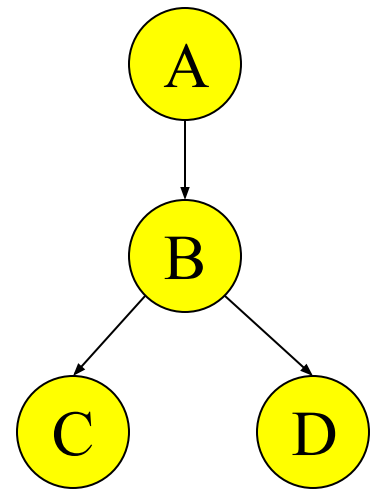
$$P(A=t) = \sum_{b,c,d} P(A=t, B=b, C=c, D=d)$$

$$= \sum_{b,c,d} P(A=t)P(B=b \mid A=t)P(C=c \mid B=b)P(D=d \mid B=b)$$

$$= P(A=t)\sum_{b,c,d} P(B=b \mid A=t)P(C=c \mid B=b)P(D=d \mid B=b)$$

$$= P(A=t)\sum_{b} P(B=b \mid A=t)\sum_{c,d} P(C=c \mid B=b)P(D=d \mid B=b)$$

$$= P(A=t)\sum_{b} P(B=b \mid A=t)\sum_{c} P(C=c \mid B=b)\sum_{d} P(D=d \mid B=b)$$

$$= P(A=t)\sum_{b} P(B=b \mid A=t)\sum_{c} P(C=c \mid B=b)*1$$

$$= 0.4(P(B=t \mid A=t)\sum_{c} P(C=c \mid B=t) + P(B=f \mid A=t)\sum_{c} P(C=c \mid B=f)) = ...$$

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

31

What is P(C=true, A=true)?
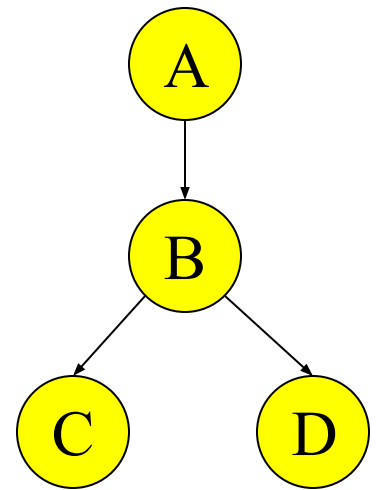
$$P(A=t, C=t) = \sum_{b,d} P(A=t, B=b, C=t, D=d)$$

$$= \sum_{b,d} P(A=t)P(B=b \mid A=t)P(C=t \mid B=b)P(D=d \mid B=b)$$

$$= P(A=t)\sum_{b} P(B=b \mid A=t)P(C=t \mid B=b)\sum_{d} P(D=d \mid B=b)$$

$$= 0.4\bigl(P(B=t \mid A=t)P(C=t \mid B=t)\sum_{d} P(D=d \mid B=t)$$

$$+ P(B=f \mid A=t)P(C=t \mid B=f)\sum_{d} P(D=d \mid B=f)\bigr)$$

$$= 0.4(0.3*0.1*1 + 0.7*0.6*1) = 0.4(0.03+0.42) = 0.4*0.45 = 0.18$$

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

32

The Bad News

- Exact inference is feasible in small to medium-sized networks

- Exact inference in large networks takes a very long time

- We resort to approximate inference techniques which are much faster and give pretty good results

One last unresolved issue…

We still haven't said where we get the Bayesian network from. There are two options:

- Get an expert to design it

- Learn it from data, e.g., the same way as in the lecture on Bayes Classifier in Ch. 8.