

A Tutorial on Bayesian Networks

Weng-Keen Wong

School of Electrical Engineering and Computer Science

Oregon State University

Introduction



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

- The patient has a cough
- The patient has a fever
- The patient has difficulty breathing

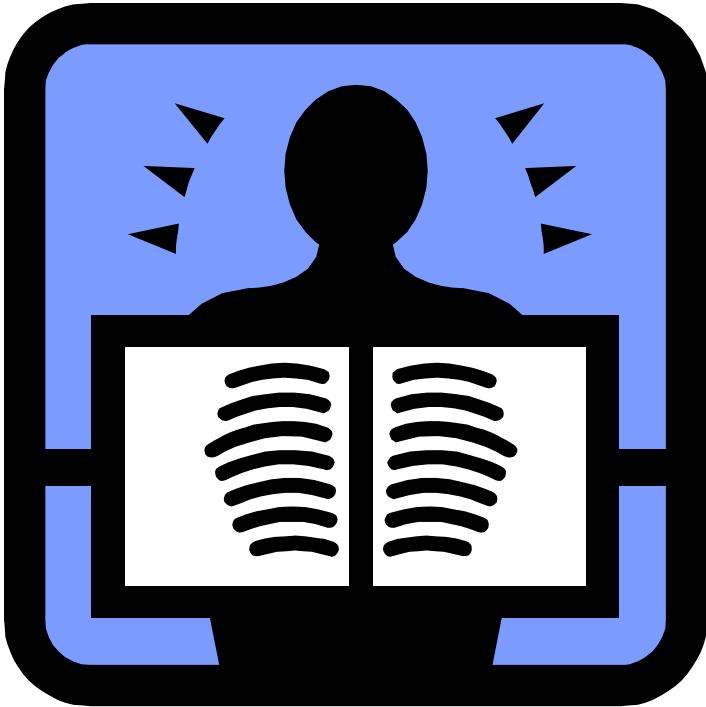
Introduction



You would like to determine how likely the patient has pneumonia given that the patient has a cough, a fever, and difficulty breathing

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

Introduction



Now suppose you order a chest x-ray and the results are positive.

Your belief that that the patient has pneumonia is now much higher.

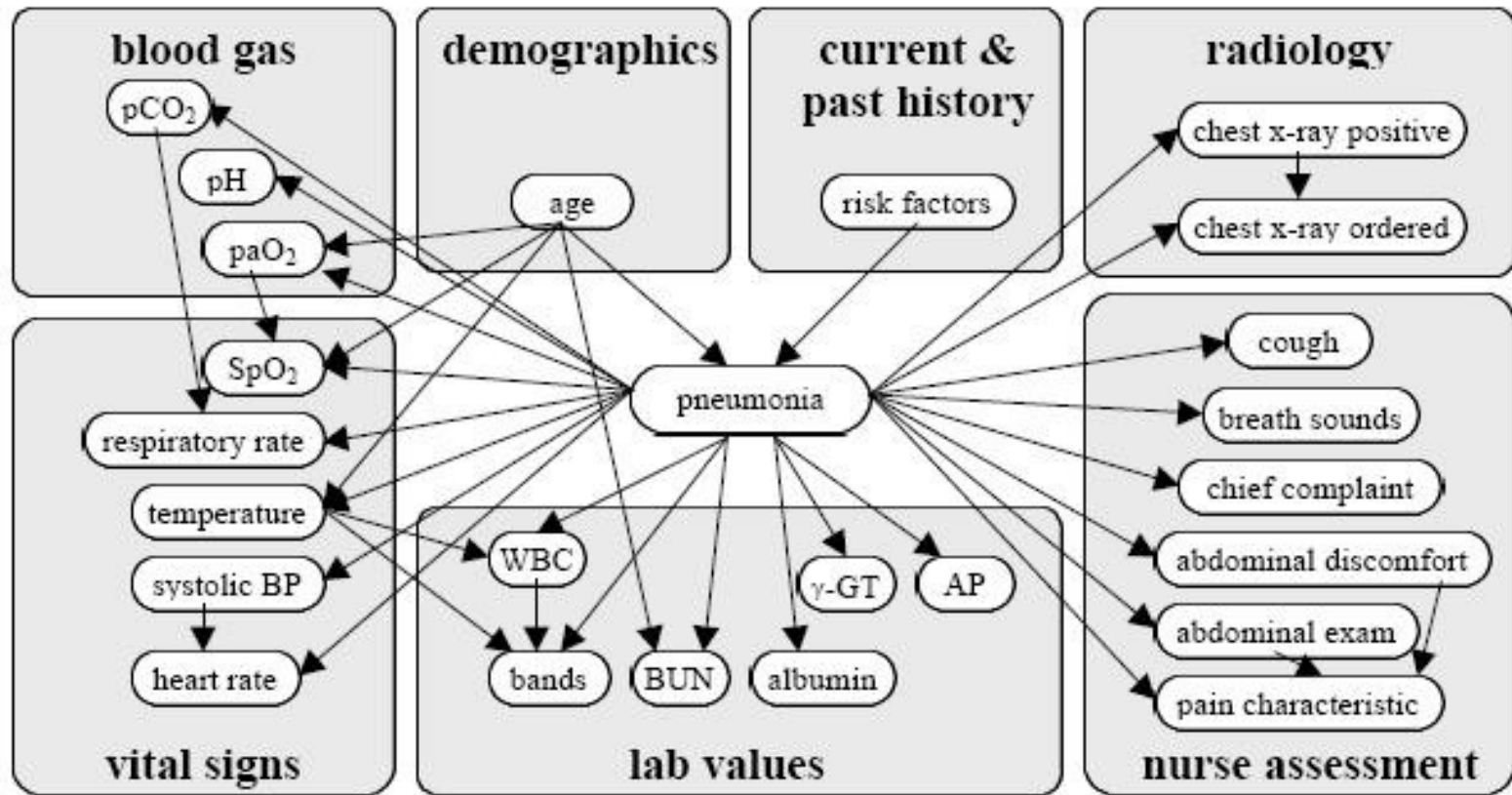
Introduction

- In the previous slides, what you observed affected your belief that the patient has pneumonia
- This is called reasoning with uncertainty
- Wouldn't it be nice if we had some methodology for reasoning with uncertainty? Why in fact, we do...

Bayesian Networks

- Bayesian networks help us reason with uncertainty
- In the opinion of many AI researchers, Bayesian networks are the most significant contribution in AI in the last 10 years
- They are used in many applications eg.:
 - Spam filtering / Text mining
 - Speech recognition
 - Robotics
 - Diagnostic systems
 - Syndromic surveillance

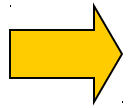
Bayesian Networks (An Example)



From: Aronsky, D. and Haug, P.J., Diagnosing community-acquired pneumonia with a Bayesian network, In: *Proceedings of the Fall Symposium of the American Medical Informatics Association*, (1998) 632-636.

Outline

1. Introduction



2. Probability Primer

3. Bayesian networks

4. Bayesian networks in syndromic surveillance

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 heads on a coin flip



Boolean Random Variables

- We deal 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 heads on a coin flip
 - A = It will rain today
 - A = There is a typo in these slides

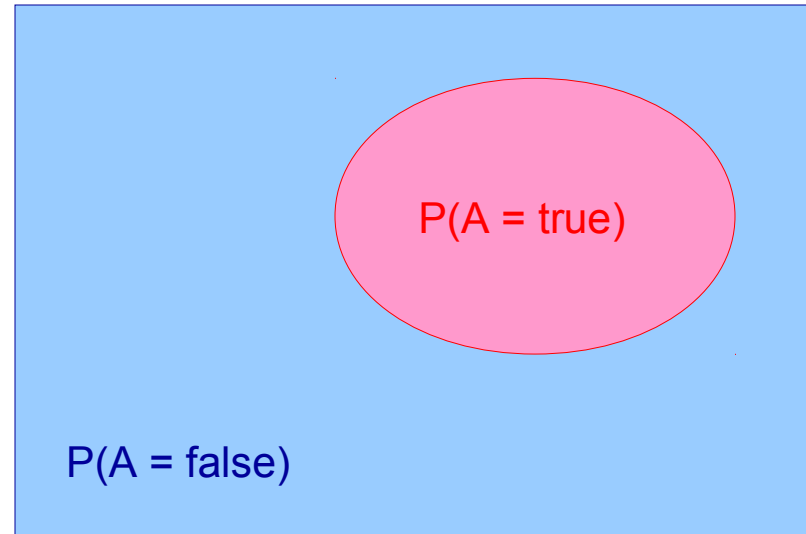
Probabilities

We will write $P(A = \text{true})$ to mean the probability that $A = \text{true}$.

What is probability? It is the relative frequency with which an outcome would be obtained if the process were repeated a large number of times under similar conditions*

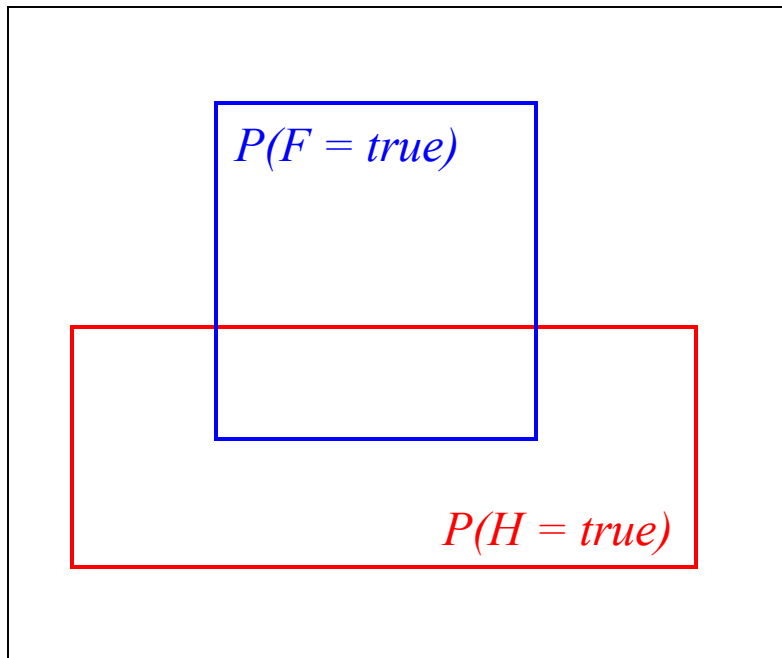
The sum of the red and blue areas is 1

*Ahem...there's also the Bayesian definition which says probability is your degree of belief in an outcome



Conditional Probability

- $P(A = \text{true} \mid B = \text{true})$ = Out of all the outcomes in which B is true, how many also have A equal to true
- Read this as: “Probability of A conditioned on B ” or “Probability of A given B ”



H = “Have a headache”

F = “Coming down with Flu”

$$P(H = \text{true}) = 1/10$$

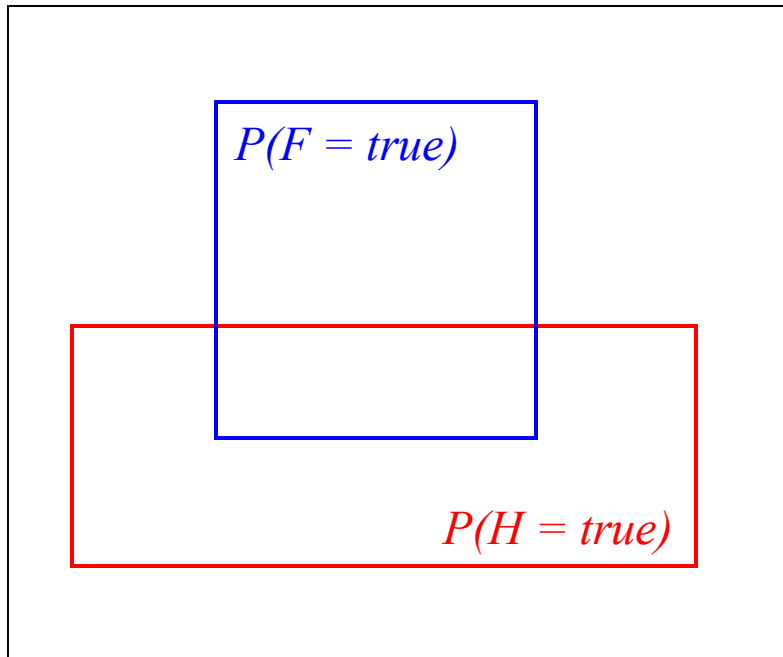
$$P(F = \text{true}) = 1/40$$

$$P(H = \text{true} \mid F = \text{true}) = 1/2$$

“Headaches are rare and flu is rarer, but if you’re coming down with flu there’s a 50-50 chance you’ll have a headache.”

The Joint Probability Distribution

- We will write $P(A = \text{true}, B = \text{true})$ to mean “the probability of $A = \text{true}$ **and** $B = \text{true}$ ”
- Notice that:



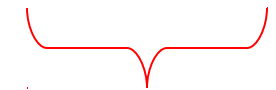
$$\begin{aligned} &P(H = \text{true} | F = \text{true}) \\ &= \frac{\text{Area of "H and F" region}}{\text{Area of "F" region}} \\ &= \frac{P(H = \text{true}, F = \text{true})}{P(F = \text{true})} \end{aligned}$$

In general, $P(X | Y) = P(X, Y) / P(Y)$

The Joint Probability Distribution

- Joint probabilities can be between any number of variables
eg. $P(A = \text{true}, B = \text{true}, C = \text{true})$
- 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

- 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)

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

Examples of things you can compute:

- $P(A=true) = \text{sum of } P(A,B,C) \text{ in rows with } A=true$
- $P(A=true, B = true \mid C=true) =$
 $P(A = true, B = true, C = true) / P(C = true)$

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 $P(C_1, \dots, C_n)$
- If the coin flips are not independent, you need 2^n values in the table
- If the coin flips are independent, then

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

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

Conditional Independence

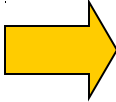
Variables A and B are conditionally independent given C if any of the following hold:

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



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)

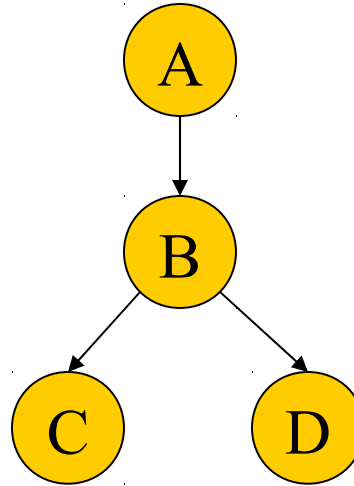
Outline

1. Introduction
2. Probability Primer
-  3. Bayesian networks
4. Bayesian networks in syndromic surveillance

A Bayesian Network

A Bayesian network is made up of:

1. A Directed Acyclic Graph



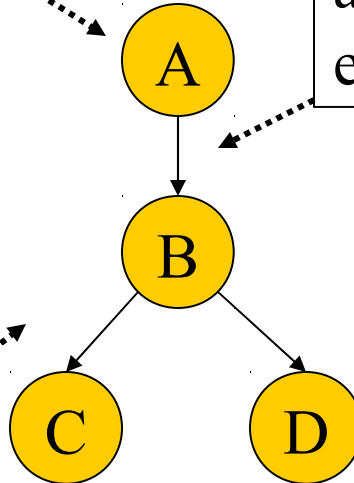
2. A set of tables for each node in the graph

A	P(A)	A	B	P(B A)	B	D	P(D B)	B	C	P(C B)
false	0.6	false	false	0.01	false	false	0.02	false	false	0.4
true	0.4	false	true	0.99	false	true	0.98	false	true	0.6
		true	false	0.7	true	false	0.05	true	false	0.9
		true	true	0.3	true	true	0.95	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



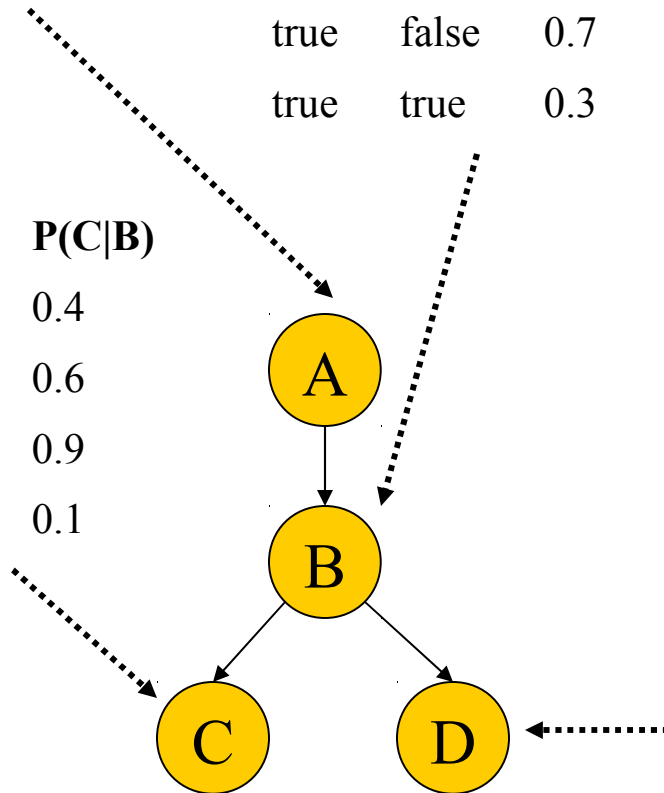
Informally, an arrow from node X to node Y means X has a direct influence on Y

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

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



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

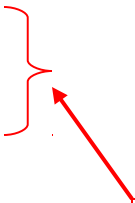
The parameters are the probabilities in these conditional probability tables (CPTs)

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 $P(C=\text{true} \mid B)$ and $P(C=\text{false} \mid B)$ must add up to 1
eg. $P(C=\text{true} \mid B=\text{false}) + P(C=\text{false} \mid B=\text{false}) = 1$

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)

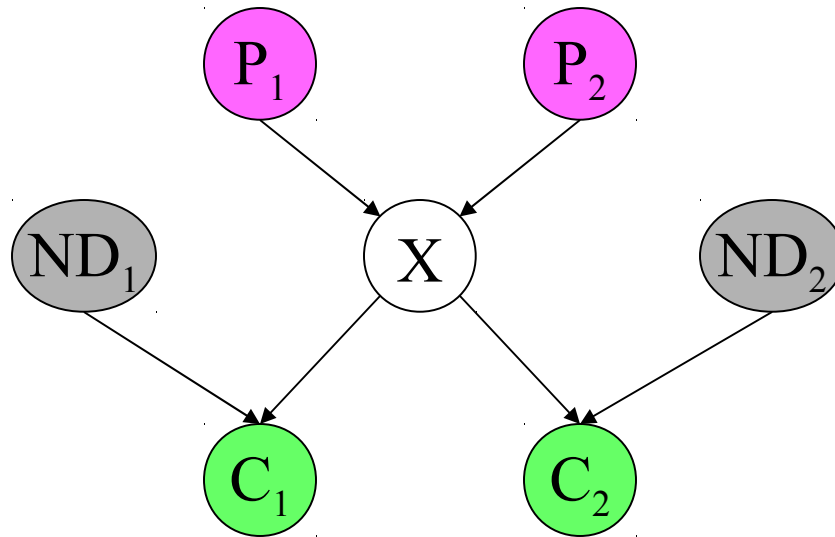
Bayesian Networks

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, \dots, X_n in the Bayesian net using the formula:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i \mid \text{Parents}(X_i))$$

Where $\text{Parents}(X_i)$ means the values of the Parents of the node X_i with respect to the graph

Using a Bayesian Network Example

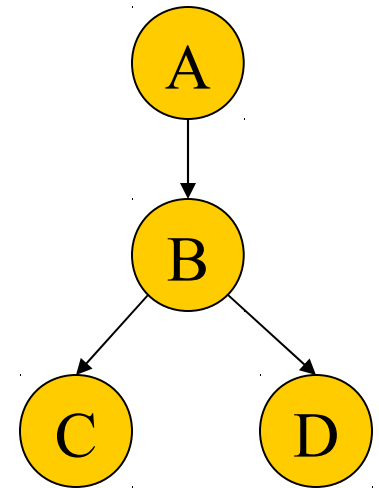
Using the network in the example, suppose you want to calculate:

$$P(A = \text{true}, B = \text{true}, C = \text{true}, D = \text{true})$$

$$= P(A = \text{true}) * P(B = \text{true} \mid A = \text{true}) *$$

$$P(C = \text{true} \mid B = \text{true}) P(D = \text{true} \mid B = \text{true})$$

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



Using a Bayesian Network Example

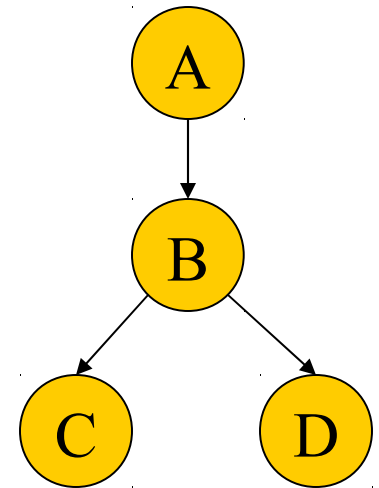
Using the network in the example, suppose you want to calculate:

$$\begin{aligned} &P(A = \text{true}, B = \text{true}, C = \text{true}, D = \text{true}) \\ &= P(A = \text{true}) * P(B = \text{true} \mid A = \text{true}) * \\ &\quad P(C = \text{true} \mid B = \text{true}) P(D = \text{true} \mid B = \text{true}) \\ &= (0.4)*(0.3)*(0.1)*(0.95) \end{aligned}$$

This is from the
graph structure



These numbers are from the
conditional probability tables



Inference

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

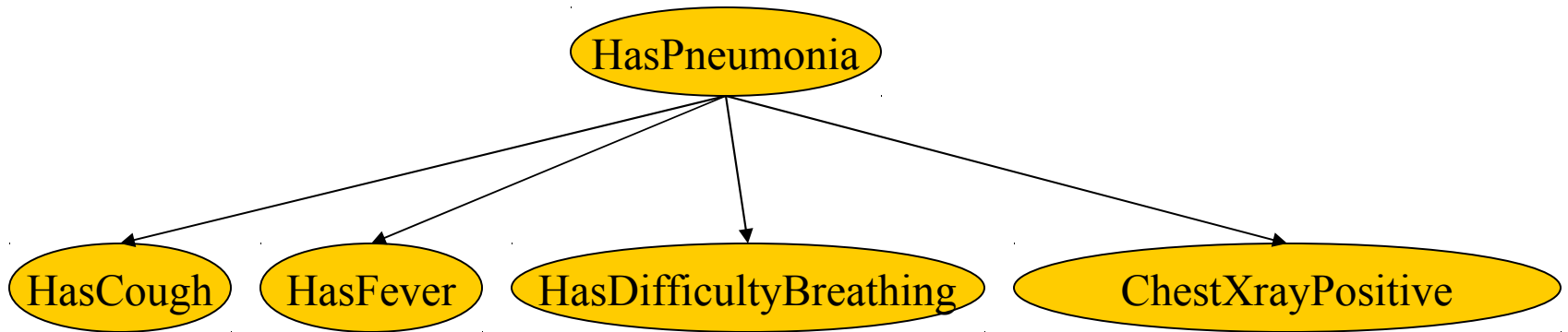
$$P(X \mid E)$$



E = The evidence variable(s)

X = The query variable(s)

Inference



- An example of a query would be:
 $P(\text{HasPneumonia} = \text{true} \mid \text{HasFever} = \text{true}, \text{HasCough} = \text{true})$
- Note: Even though *HasDifficultyBreathing* and *ChestXrayPositive* 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

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

How is the Bayesian network created?

1. Get an expert to design it

- Expert must determine the structure of the Bayesian network
 - This is best done by modeling direct causes of a variable as its parents
- Expert must determine the values of the CPT entries
 - These values could come from the expert's informed opinion
 - Or an external source eg. census information
 - Or they are estimated from data
 - Or a combination of the above

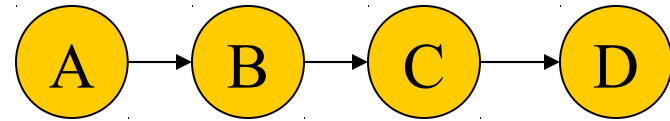
1. Learn it from data

- This is a much better option but it usually requires a large amount of data
- This is where Bayesian statistics comes in!

Learning Bayesian Networks from Data

Given a data set, can you learn what a Bayesian network with variables A, B, C and D would look like?

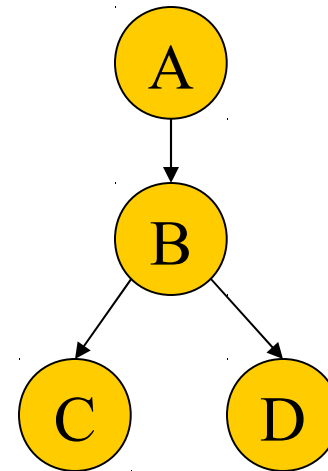
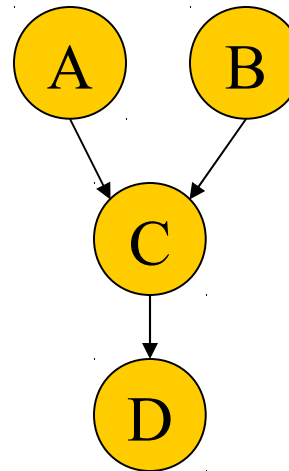
A	B	C	D
true	false	false	true
true	false	true	false
true	false	false	true
false	true	false	false
false	true	false	true
false	true	false	false
false	true	false	false
:	:	:	:



or

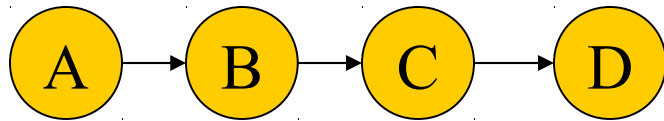
or

or



?

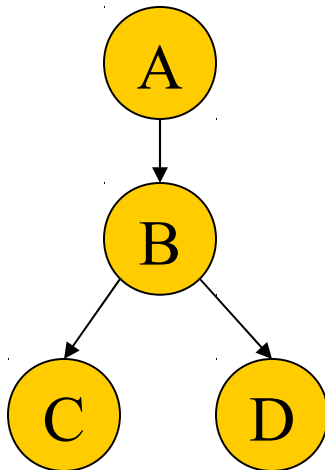
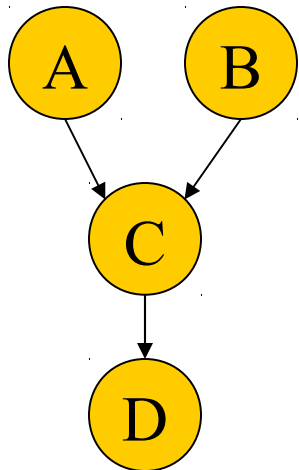
Learning Bayesian Networks from Data



or

or

or

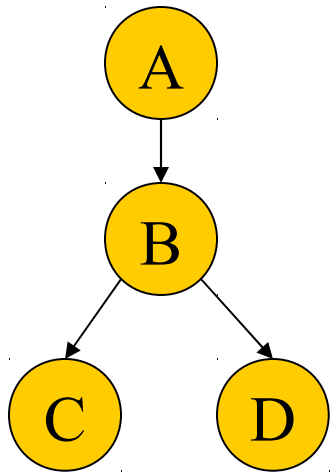


?

- Each possible structure contains information about the conditional independence relationships between A, B, C and D
- We would like a structure that contains conditional independence relationships that are supported by the data
- Note that we also need to learn the values in the CPTs from data

Learning Bayesian Networks from Data

How does Bayesian statistics help?



1. I might have a prior belief about what the structure should look like.

2. I might have a prior belief about what the values in the CPTs should be.

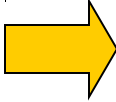
These beliefs get updated as I see more data

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

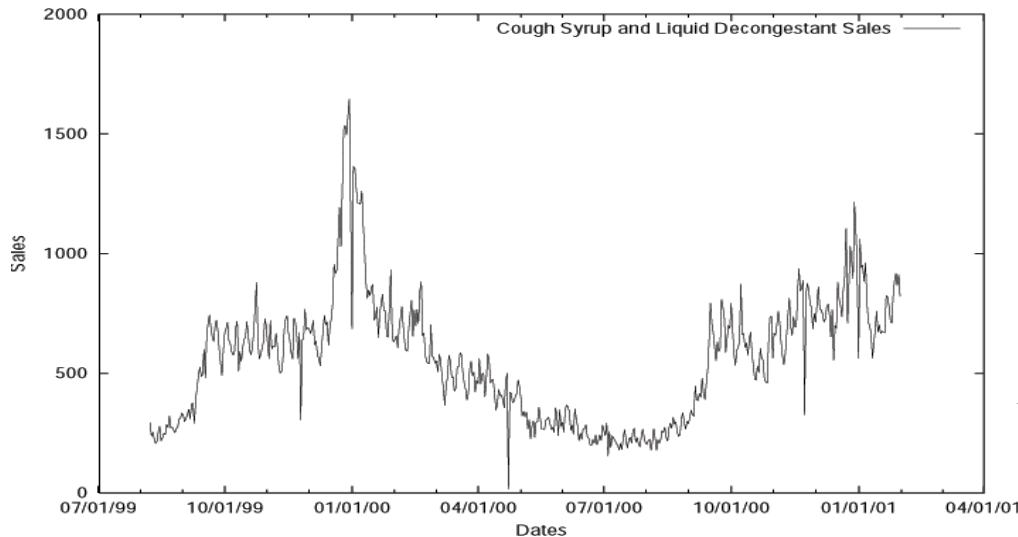
Learning Bayesian Networks from Data

- We won't have enough time to describe how we actually learn Bayesian networks from data
- If you are interested, here are some references:
 - Gregory F. Cooper and Edward Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9:309-347, 1992.
 - David Heckerman. A Tutorial on Learning Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research. 1995. (Available online)

Outline

1. Introduction
2. Probability Primer
3. Bayesian networks
-  4. Bayesian networks in syndromic surveillance

Bayesian Networks in Syndromic Surveillance



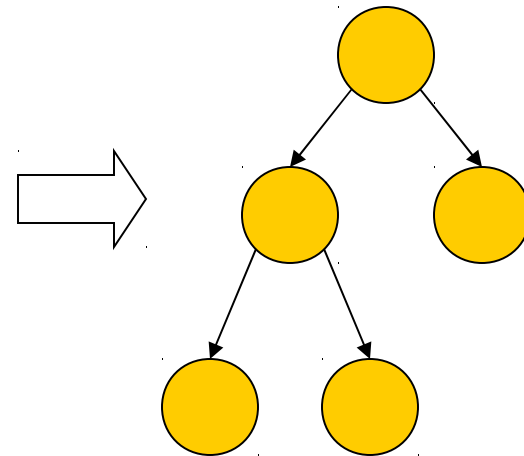
From: Goldenberg, A., Shmueli, G., Caruana, R. A., and Fienberg, S. E. (2002). Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Sciences* (pp. 5237-5249)

- Syndromic surveillance systems traditionally monitor univariate time series
- With Bayesian networks, it allows us to model multivariate data and monitor it

What's Strange About Recent Events (WSARE) Algorithm

Bayesian networks used to model the multivariate baseline distribution for ED data

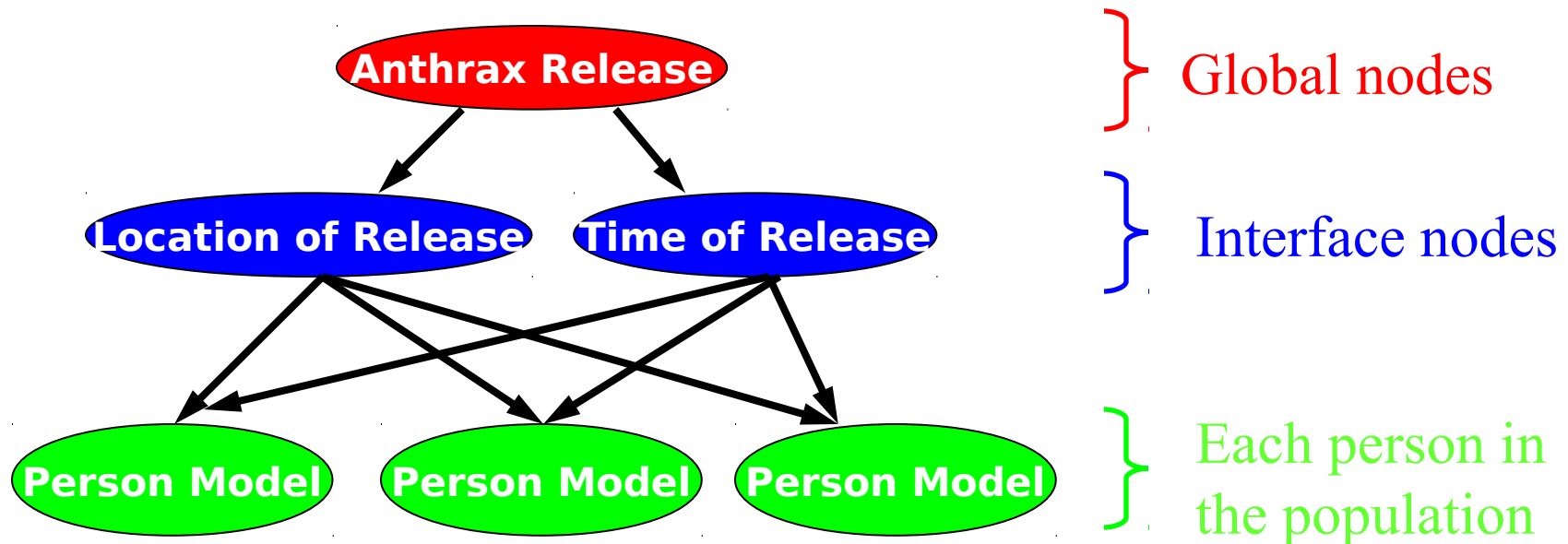
Date	Time	Gender	Age	Home Location	Many more...
6/1/03	9:12	M	20s	NE	...
6/1/03	10:45	F	40s	NE	...
6/1/03	11:03	F	60s	NE	...
6/1/03	11:07	M	60s	E	...
6/1/03	12:15	M	60s	E	...
:	:	:	:	:	:



Population-wide ANomaly Detection and Assessment (PANDA)

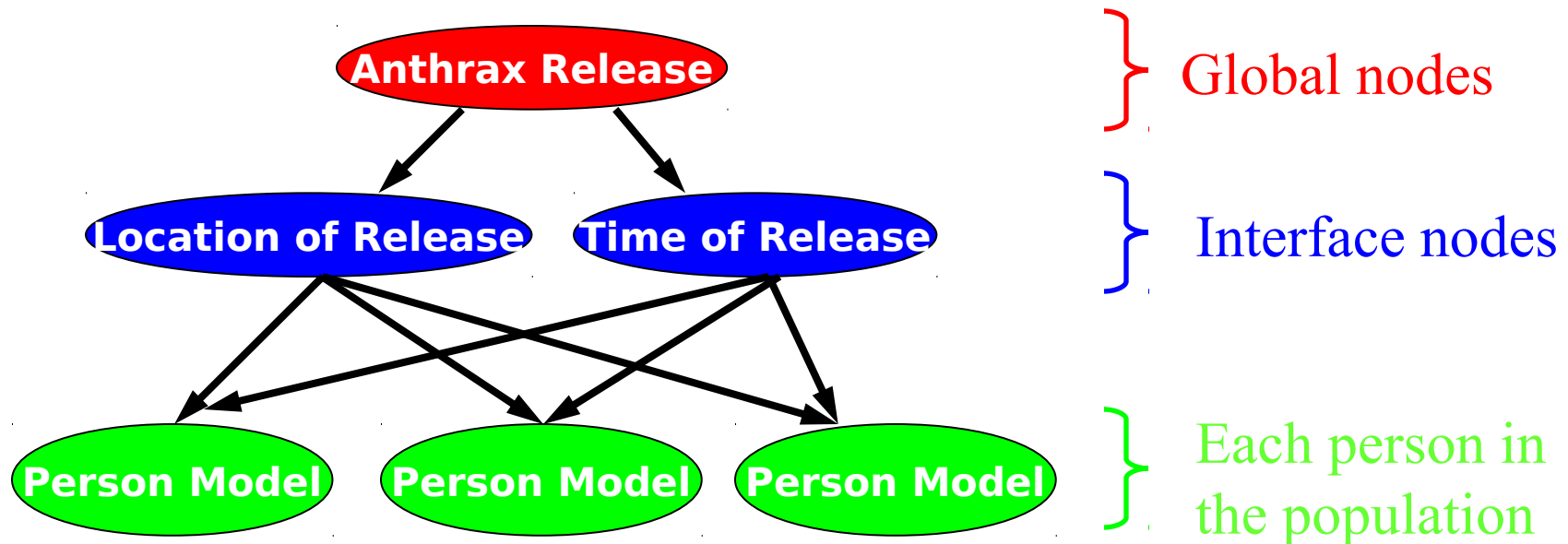
- A detector specifically for a large-scale outdoor release of inhalational anthrax
- Uses a massive causal Bayesian network
- **Population-wide approach**: each person in the population is represented as a subnetwork in the overall model

Population-Wide Approach



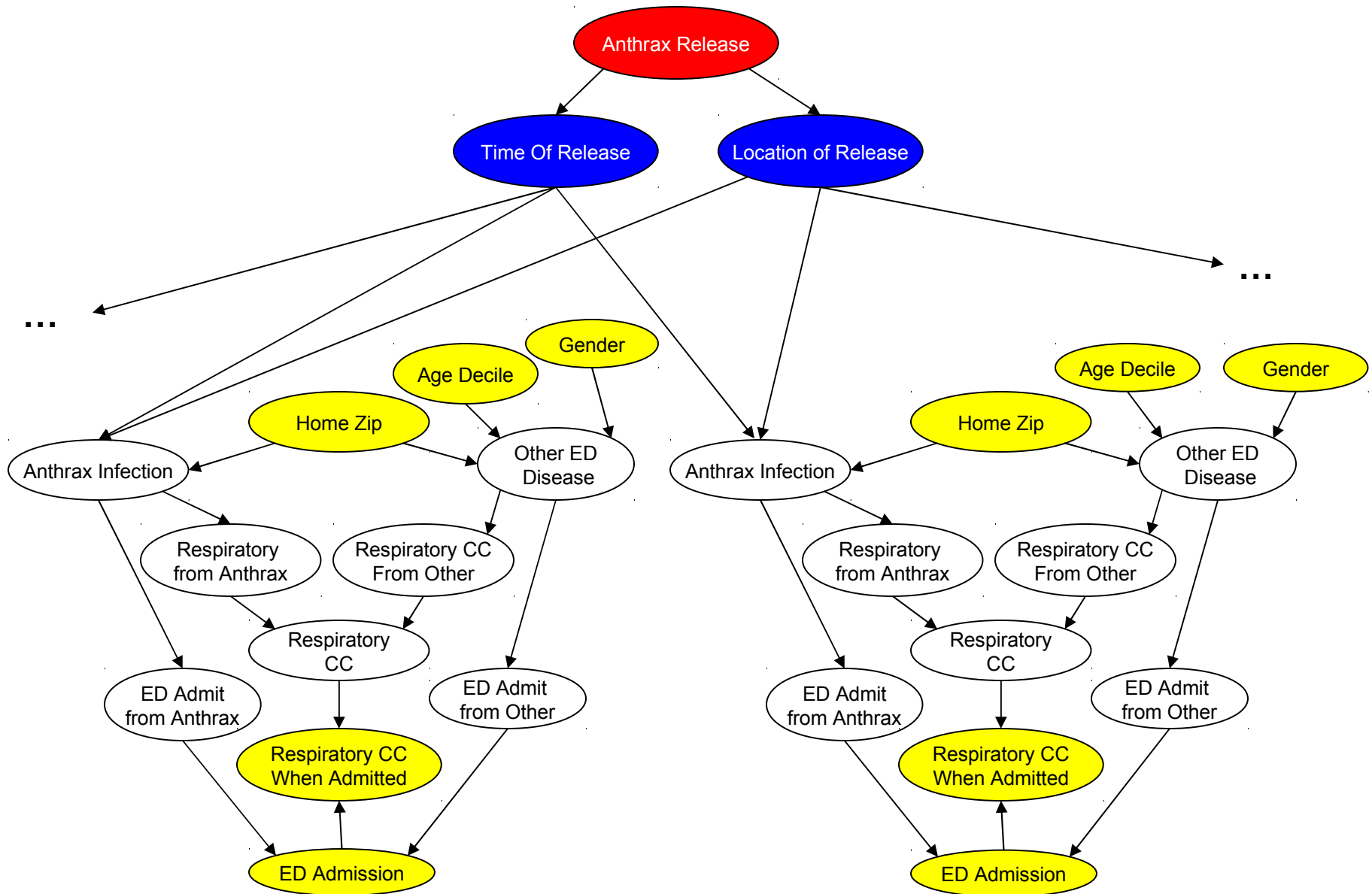
- Note the conditional independence assumptions
- Anthrax is infectious but non-contagious

Population-Wide Approach

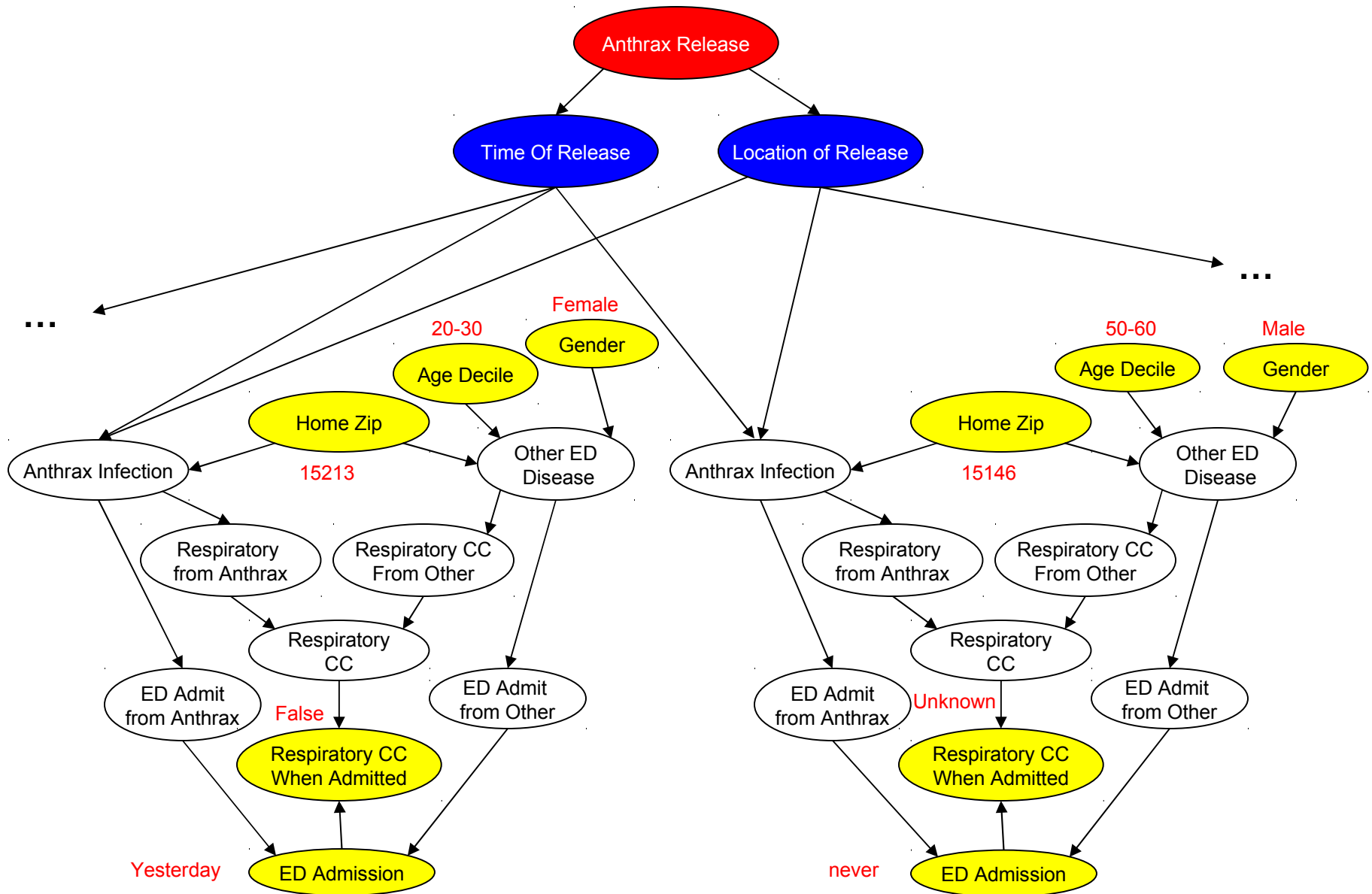


- Structure designed by expert judgment
- Parameters obtained from census data, training data, and expert assessments informed by literature and experience

Person Model (Initial Prototype)



Person Model (Initial Prototype)



What else does this give you?

1. Can model information such as the spatial dispersion pattern, the progression of symptoms and the incubation period
2. Can combine evidence from ED and OTC data
3. Can infer a person's work zip code from their home zip code
4. Can explain the model's belief in an anthrax attack

Acknowledgements

- These slides were partly based on a tutorial by Andrew Moore
- Greg Cooper, John Levander, John Dowling, Denver Dash, Bill Hogan, Mike Wagner, and the rest of the RODS lab

References

Bayesian networks:

- “Bayesian networks without tears” by Eugene Charniak
- “Artificial Intelligence: A Modern Approach” by Stuart Russell and Peter Norvig
- “Learning Bayesian Networks” by Richard Neapolitan
- “Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference” by Judea Pearl

Other references:

- My webpage
<http://www.eecs.oregonstate.edu/~wong>