香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

COMP7015 Artificial Intelligence

Lecture 10: Bayesian Learning

Instructor: Dr. Kejing Yin

November 17, 2022

Logistics

- Sample Solution to PA1 is released. Check out in Moodle.
- Programming Assignment 2 *Due: 23:59 pm on Nov. 27*
- Written Assignment 2 will be out on Nov. 18.
  Due: 23:59 pm on Nov. 30 (submission cut-off at 1 pm on Dec. 1)
- Remaining Office Hours Moved to Mondays:
  Nov. 21, Nov. 28, and Dec. 5

Logistics

- Group Project:
  - Final deadline is approaching (Dec. 3).
  - Submission box will open on Nov. 22.
  - No late submissions for group project. You are suggested to submit a preliminary version *well ahead of the deadline* to avoid last-minute accidents (internet issues, etc.)
  - Assessment Rubrics are out. Check them out in Moodle!

Agenda for Today

- Basics of Probability: A Brief Recap
- Bayesian Decision Theory
- Naïve Bayes Classifier
- Bayesian Networks

- Genetic Algorithms

Why use probability?

- Stochastic environment: outcome of an action might be truly random.
- Multi-agent environment:
  - If other players are rational and their goals are known, then you don't need probability; you just work out what their rational actions will be.
  - If other players have unknown goals, then model them as random.
- Unknown environment: outcome of an action is not truly random, but you don't know what the outcome will be.
  - In this case, "probability" measures your belief: P(Q|A)=the degree to which you believe that action A will produce outcome Q.
- Computational complexity:
  - Instead of searching 1b paths using A*, you could randomly choose 1k paths to try, and then choose the best of those.

Why NOT use probability?

- Multi-agent environment:
  - Maybe it's better to find out what the other players really want?
- Unknown environment:
  - Maybe it's better to learn the rules of the game?
- Computational complexity:
  - Maybe it's better to do a complete search, instead of just a partial search?

Notice: these are quantitative questions. "Better" requires some metric: how much better, and with what probability?

Basic of Probability: A Brief Recap

Credits: this section is partially adapted from Lecture 5 of CS440/CEC448 (UIUC) under CC-BY-4.0 license

What is probability?

- Latin *probabilis* = probable, commendable, believable, from *probare* = to test something
- If tested, it will (probably) turn out to be true

The axioms of probability

A = some future event, e.g., “it will rain tomorrow.”

$P(A)$ = the degree to which we believe that event A, if tested, will turn out to be true.

The axioms of probability

Axiom 1: every event has a non-negative probability.

$$P(A) \geq 0$$

Axiom 2: a certain event has probability 1.

$$P(\text{True}) = 1$$

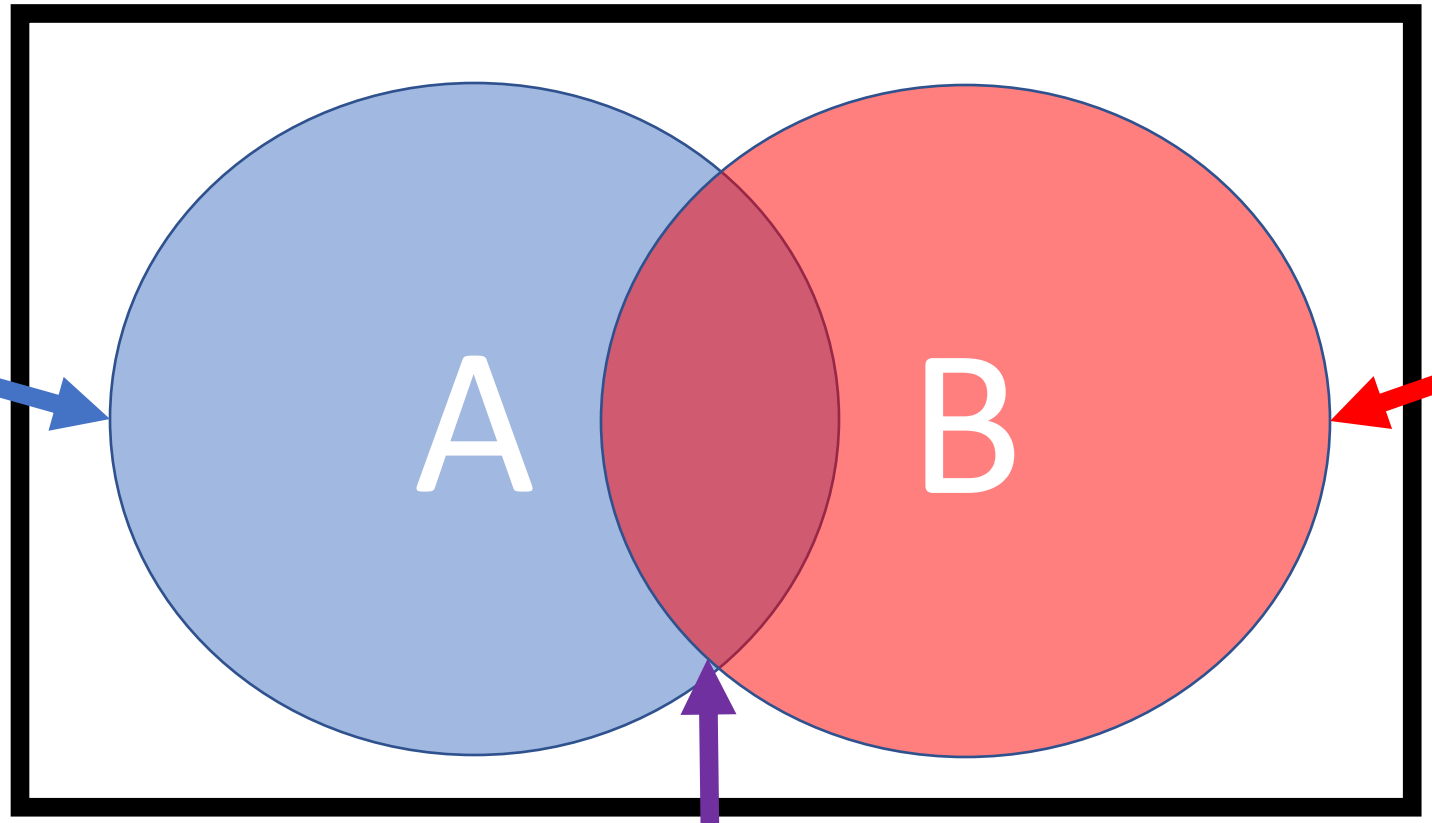
Axiom 3: probability measures behave like set measures.

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Axiom 3: probability measures behave like set measures.

Area of the whole rectangle is $P(\text{True}) = 1$.

Area of this circle is $P(A)$.



Area of this circle is $P(B)$.

Area of their intersection is $P(A \wedge B)$.

Area of their union is $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Example

- A = “it will rain tomorrow.” Suppose $P(A) = 0.4$.
- B = “it will snow tomorrow.” Suppose $P(B) = 0.2$.
- $A \wedge B$ = “it will both rain and snow tomorrow.” Suppose

$$P(A \wedge B) = 0.1$$

Then the probability that it will either rain or snow tomorrow is

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) = 0.4 + 0.2 - 0.1 = 0.5$$

Outline

- Motivation: Why use probability?
- The axioms of probability
- Random variables
- Conditional probability
- Mutually exclusive vs. Independent vs. Conditionally Independent

Random variables

- A random variable is a function that maps from the outcomes of an experiment to a set of values
- Example: throw four dice, all different colors. X = number of pips showing on the green die.
- Then run the experiment…

- In this particular outcome, X=3.
- In some other outcome, X would have taken a different value.

CC-BY-3.0 Image by Diacritica, 2010
https://commons.wikimedia.org/wiki/File:6sided_dice.jpg

Notation: $P(X = x)$

- Capital letters are random variables. Small letters are values that the random variable might take.
- "$X = x$" is an event. As such, it has a probability. For example, we can talk about the probability $P(X = x)$:

$$P(X = x) = \frac{1}{6} \qquad \forall\, x \in \{1,2,3,4,5,6\}$$

- $\forall$ means "for all." The equation above is shorthand for these six equations:

$$P(X = 1) = \frac{1}{6}, \qquad P(X = 2) = \frac{1}{6}, \qquad P(X = 3) = \frac{1}{6},$$

$$P(X = 4) = \frac{1}{6}, \qquad P(X = 5) = \frac{1}{6}, \qquad P(X = 6) = \frac{1}{6}$$

Notation: $P(X)$

X is not an event, and it's not a value; it's a function. So $P(X)$ is NOT a number. Instead, $P(X)$ is a table, showing all of the values X might take, and the probabilities of each. For example:

$P(X) =$

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Abuse of Notation: Events and Binary Random Variables

- There's one confusing thing. A capital letter might be either an event (A="it will rain tomorrow"), or a random variable (X="number of pips showing").
- P(A) is a number, but P(X) is a table.
- You have to pay attention to whether the capital letter is an event, or a random variable.

Outline

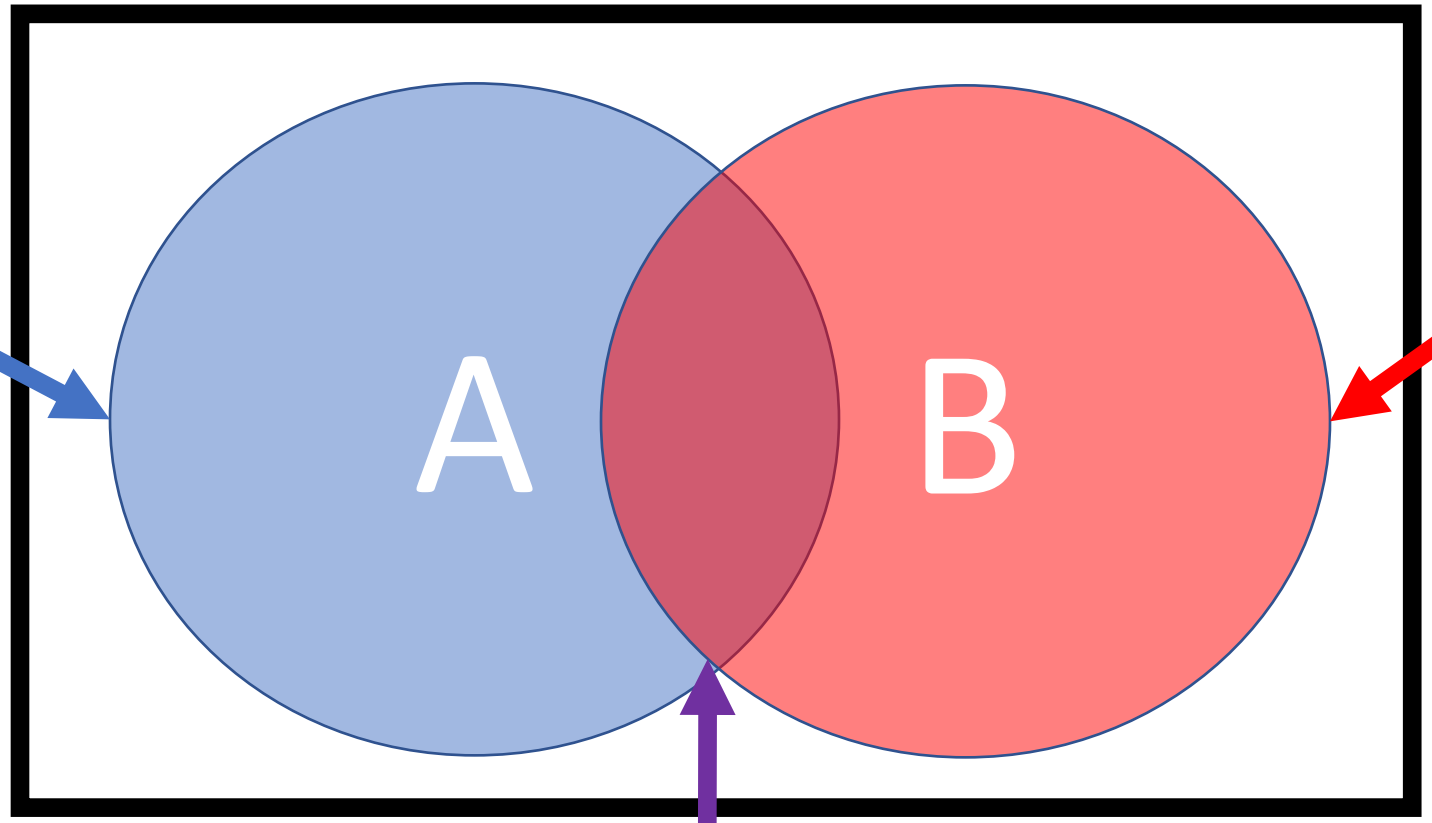
- Motivation: Why use probability?
- The axioms of probability
- Random variables
- Conditional probability
- Mutually exclusive vs. Independent vs. Conditionally Independent

Joint and Conditional probabilities: definitions

- $P(A \wedge B)$ is the probability that both event A and event B happen. This is called their **joint probability**.
- $P(B|A)$ is the probability that event B happens, given that event A happens. This is called the **conditional probability** of B given A.
- Example:
  - A = "it will rain tomorrow"
  - B = "it will snow tomorrow"
  - $P(A \wedge B)$ = probability that it will both snow and rain
  - $P(B|A)$ = probability that it will snow, given that it rains

Joint probabilities are usually given in the problem statement

Area of the whole rectangle is $P(\text{True}) = 1$.



Suppose $P(A) = 0.4$

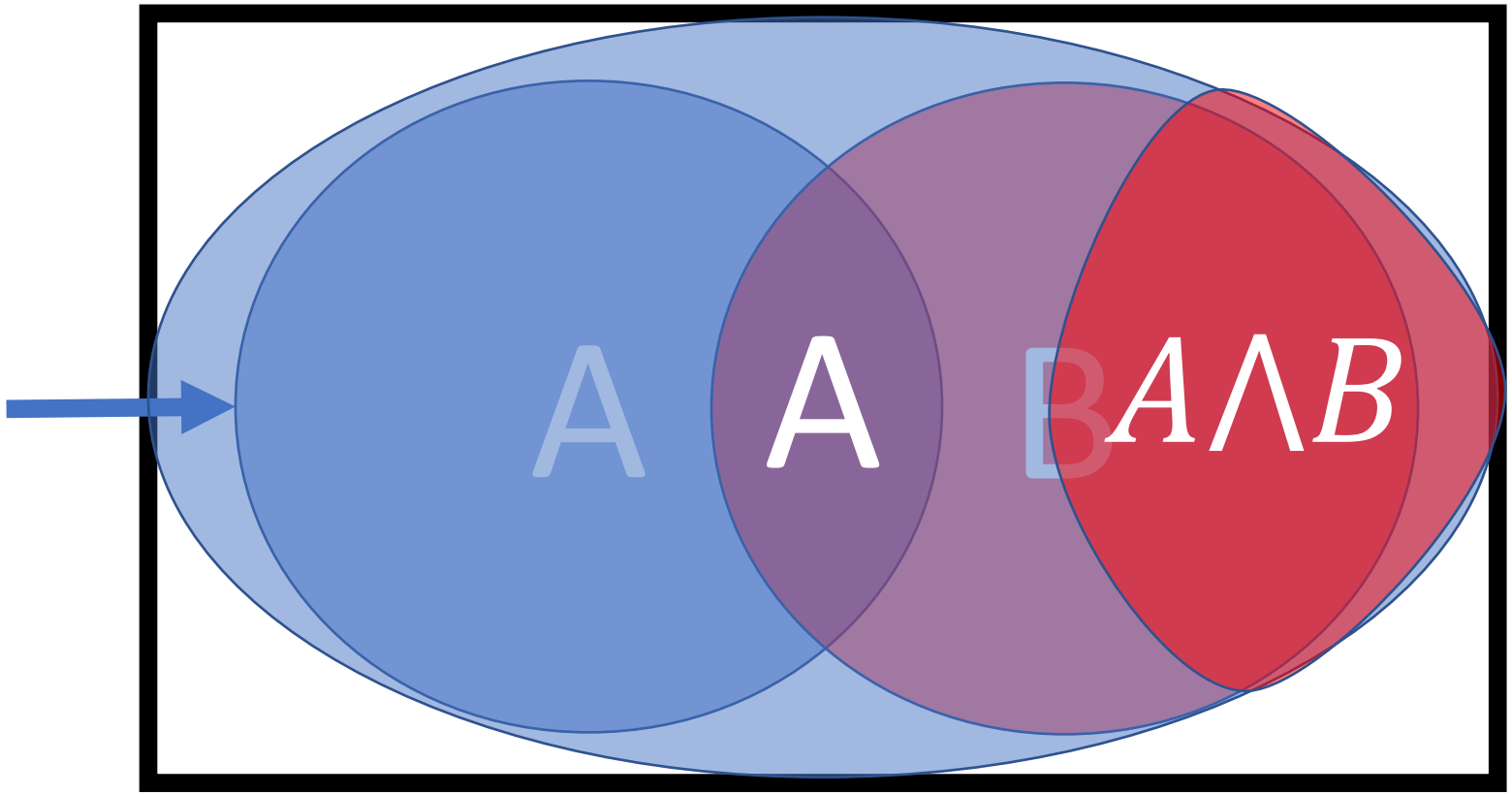
Suppose $P(B) = 0.2$

Suppose $P(A \wedge B) = 0.1$

Conditioning events change our knowledge!
For example, given that A is true…

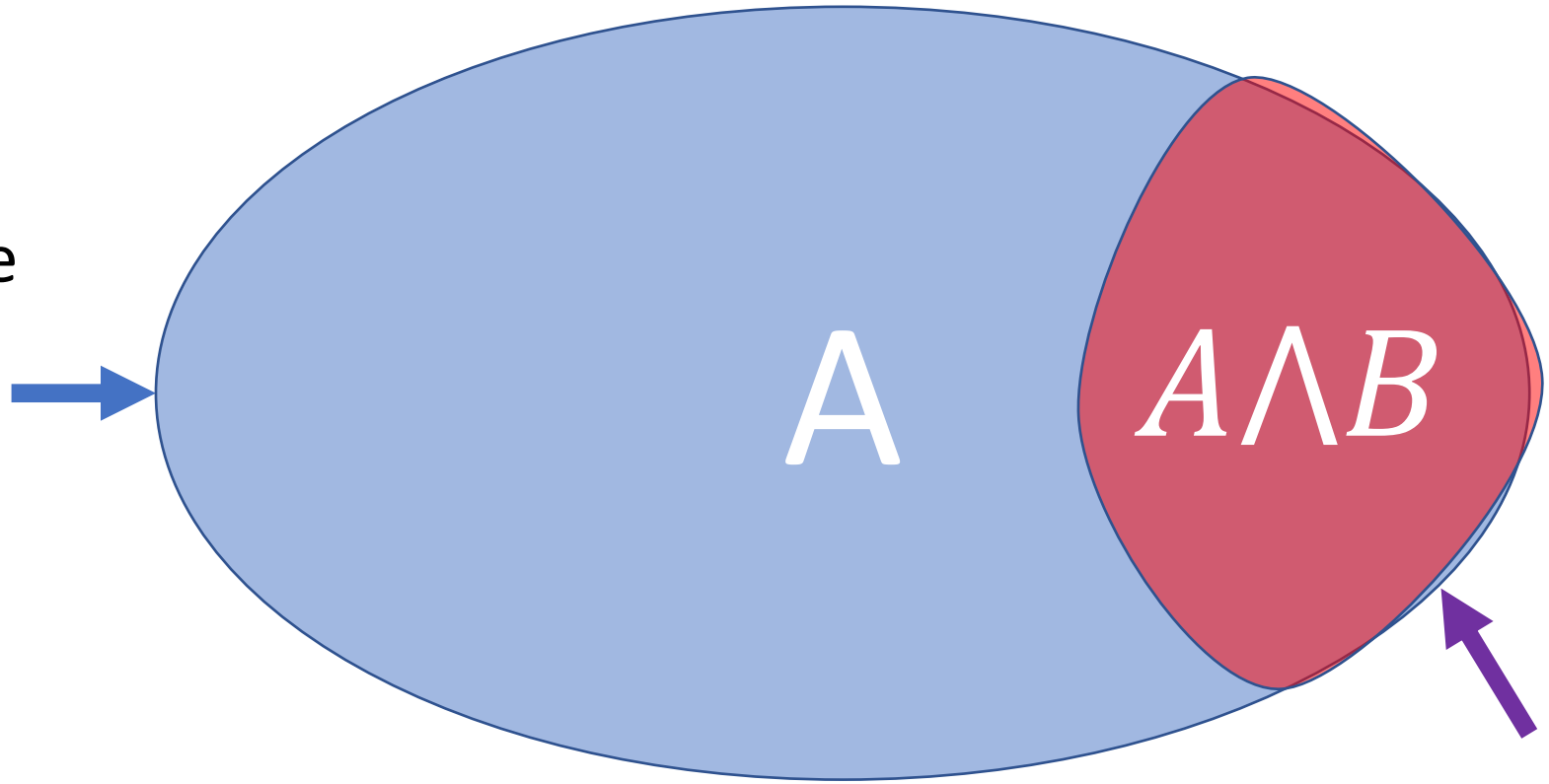
Most of the events in this rectangle are no longer possible!

Only the events inside this circle are now possible.



Conditioning events change our knowledge!
For example, given that A is true…

Given certain knowledge that A has occurred, we now have $P(A) = 1$.



Given certain knowledge that A has occurred, the probability that B also occurs is now

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

Joint and Conditional distributions of random variables

- $P(X, Y)$ is the **joint probability distribution** over all possible outcomes $P(X = x, Y = y)$.

- $P(X|Y)$ is the **conditional probability distribution** of outcomes $P(X = x|Y = y)$.

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

Joint and Conditional distributions of random variables

Example:

X = number of pips on the bone die.

Y = X modulo 2.

The joint probability $\boldsymbol{P(X = 5, Y = 1) = \frac{1}{6}}$.

Their joint distribution is:

Bone die found at Cantonment Clinch.
CC-BY-3.0, Colby Kirk, 2007.

$P(X, Y) =$

$P(X = x, Y = y)$		x					
		1	**2**	**3**	**4**	**5**	**6**
y	**0**	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$
	1	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0

Joint and Conditional distributions of random variables

Example:

X = number of pips on the bone die.

Y = X modulo 2.

The conditional probability $\boldsymbol{P(X = 5|Y = 1) = \frac{1}{3}}$.

Their joint distribution is:

Bone die found at Cantonment Clinch.
CC-BY-3.0, Colby Kirk, 2007.

$P(X|Y) =$

$P(X = x \mid Y = y)$		x					
		1	**2**	**3**	**4**	**5**	**6**
y	**0**	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$
	1	$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0

Normalization trick

If you're given the joint probability distribution and want to find the conditional distribution, just renormalize so that each row sums to 1.

$P(X = x, Y = y)$		x					
		1	**2**	**3**	**4**	**5**	**6**
y	**0**	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$
	1	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0

$P(X = x \mid Y = y)$		x					
		1	**2**	**3**	**4**	**5**	**6**
y	**0**	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$
	1	$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0

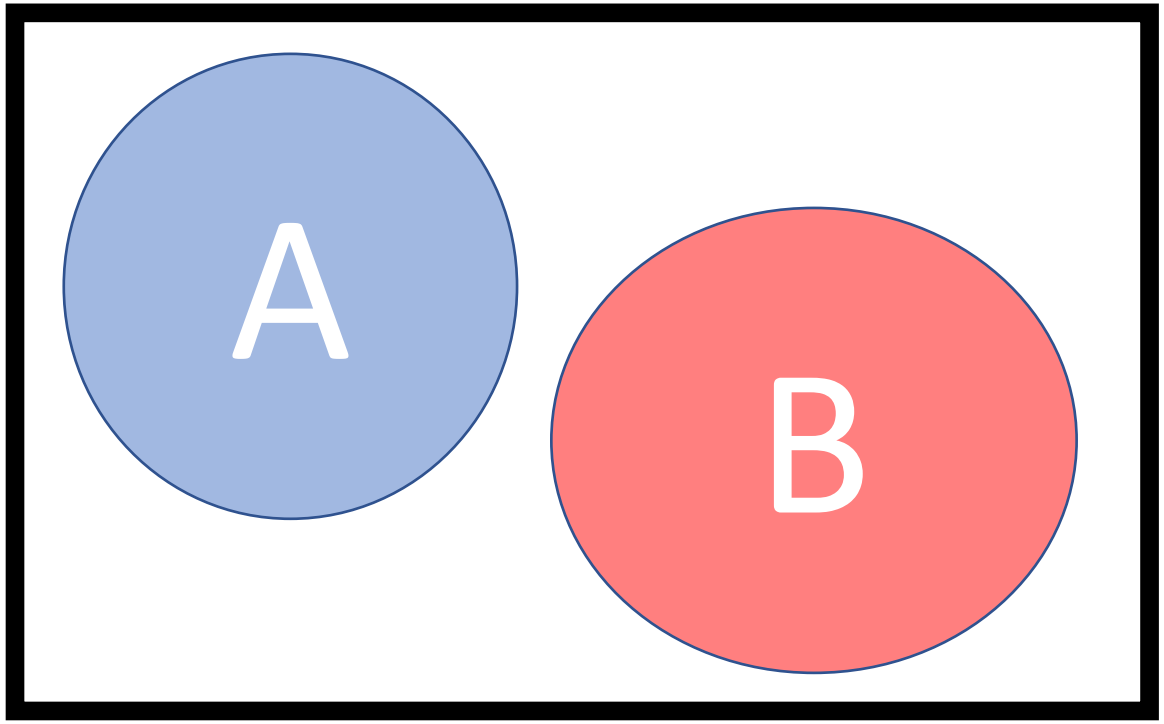
Outline

- Motivation: Why use probability?
- The axioms of probability
- Random variables
- Conditional probability
- Mutually exclusive vs. Independent vs. Conditionally Independent

Mutually exclusive events

Mutually exclusive events never occur simultaneously:

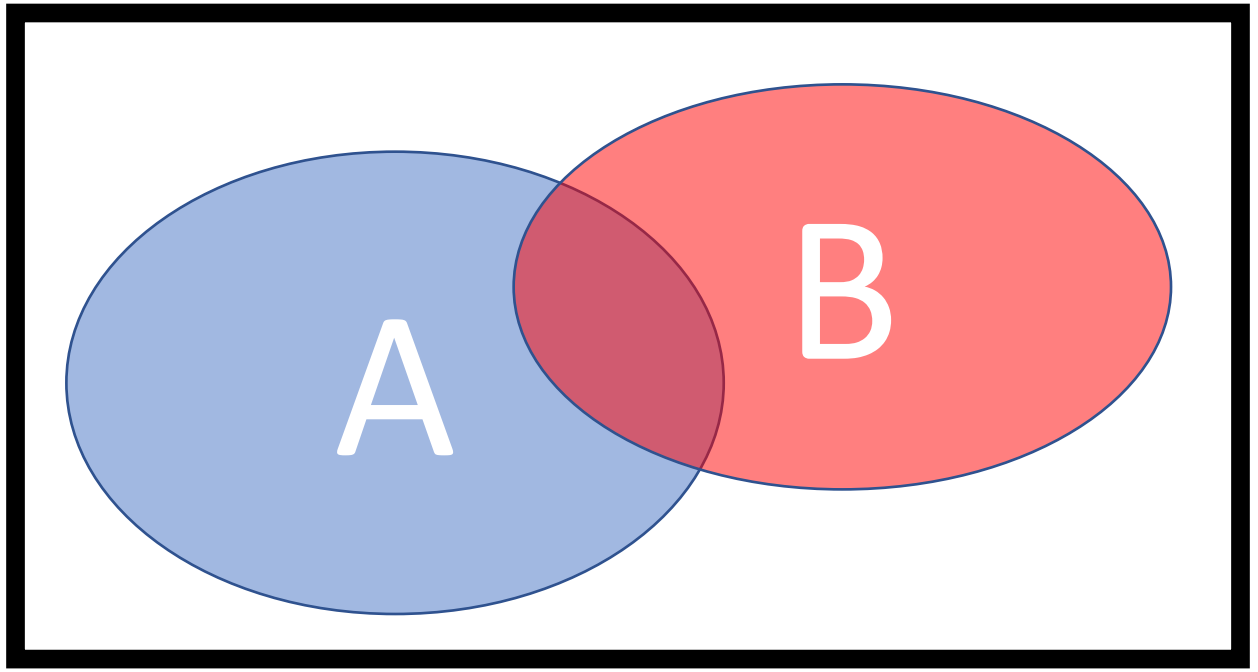
$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) = P(A) + P(B)$$



Independent events

Independent events occur with equal probability, regardless of whether or not the other event has occurred:

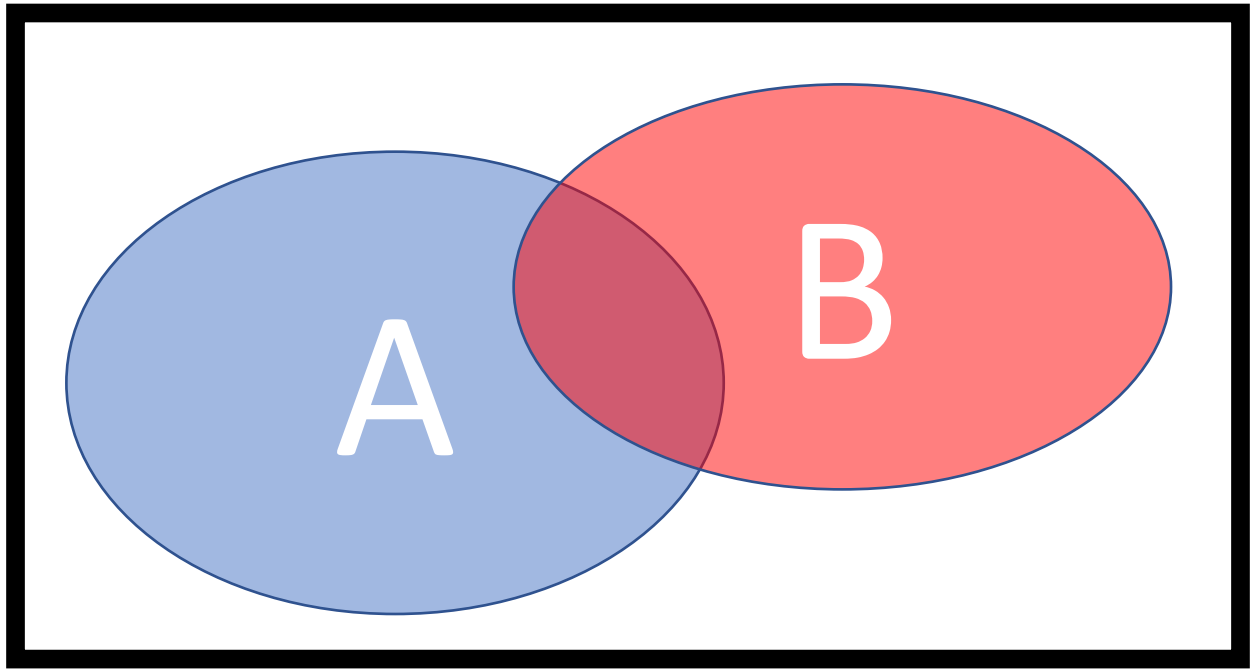
$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = P(A), \qquad P(B|A) = \frac{P(A \wedge B)}{P(A)} = P(B)$$



Independent events: A more useful definition

Re-arranging terms in the previous slide gives us this more useful definition of independent events:

$$P(A \wedge B) = P(A)P(B)$$



Independent vs. Mutually Exclusive

- Independent events:

$$P(A \wedge B) = P(A)P(B)$$

- Mutually exclusive events:

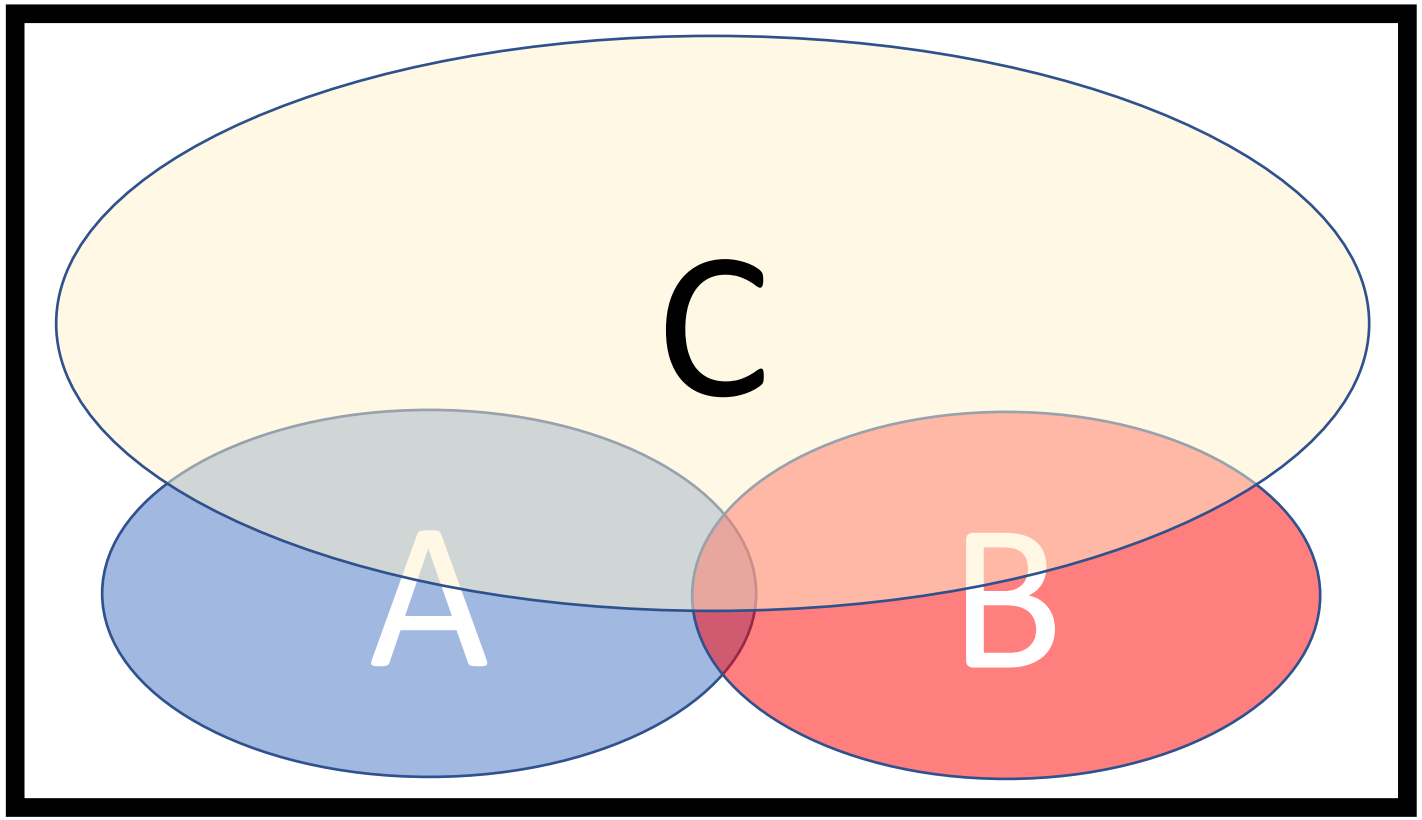
$$P(A \vee B) = P(A) + P(B)$$

Don't confuse them! Mutually exclusive events are not independent. Quite the contrary.

Conditionally independent events

Events A and B are conditionally independent, given C, if

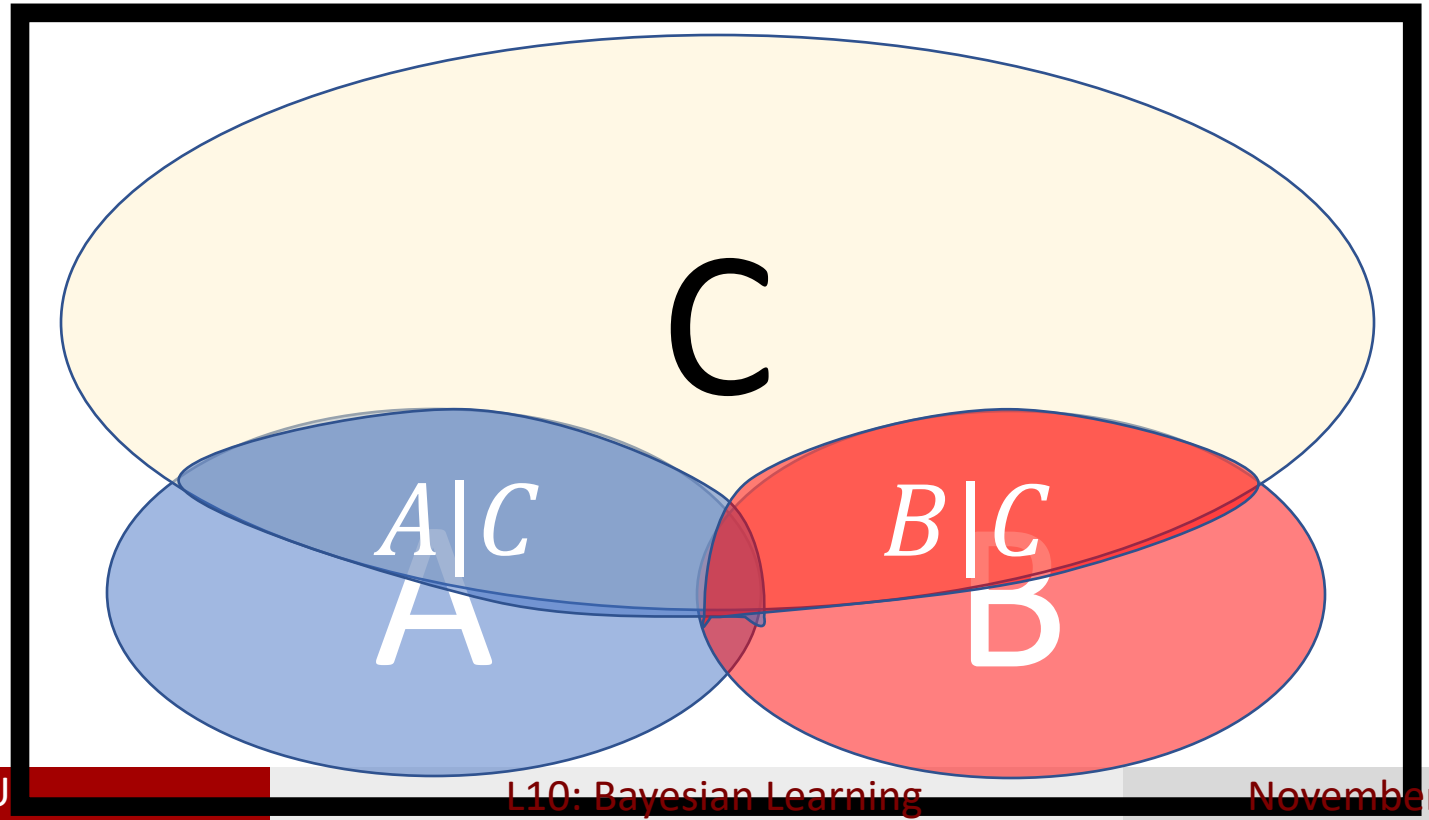
$$P(A|B,C) = P(A|C)$$



Conditionally independent events

Events A and B are conditionally independent, given C, if

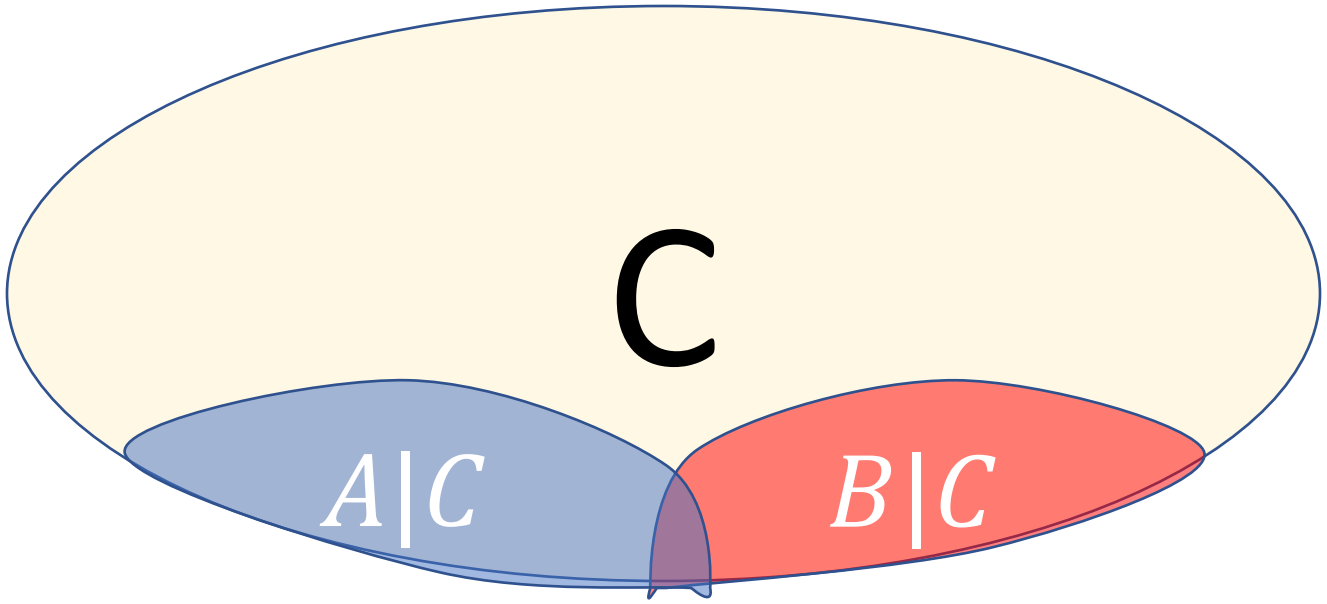
$$P(A|B,C) = \frac{P(A \wedge B|C)}{P(B|C)} = P(A|C)$$



Conditionally independent events

Events A and B are conditionally independent, given C, if

$$P(A, B|C) = P(A|C)P(B|C)$$

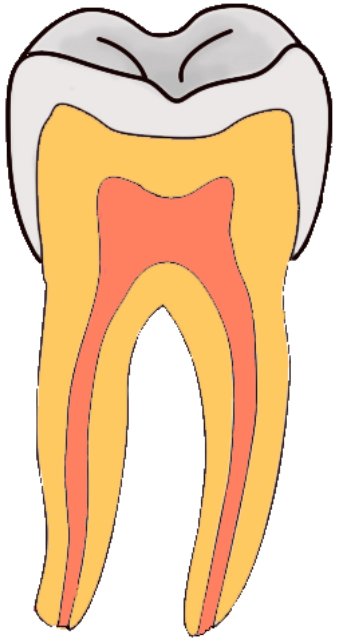


Independence ≠ Conditional Independence

Toothache= patient has a toothache

By William Brassey Hole(Died:1917)

Cavity= the patient has a cavity

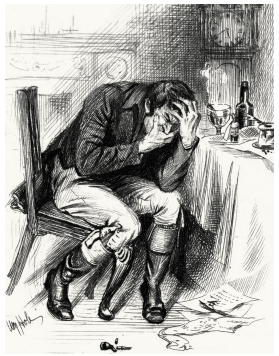


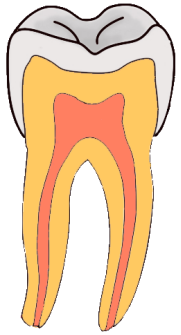
By Aduran, CC-SA 3.0

Catch= dentist's probe catches on something in the mouth

By Dozenist, CC-SA 3.0

These Events are not Independent





- If the patient has a toothache, then it's likely he has a cavity. Having a cavity makes it more likely that the probe will catch on something.

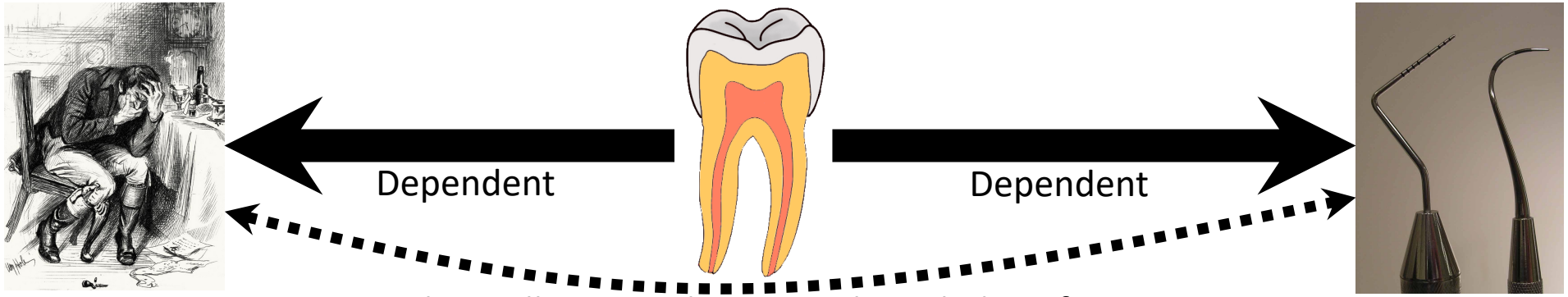
$$P(\text{Catch}|\text{Toothache}) > P(\text{Catch})$$

- If the probe catches on something, then it's likely that the patient has a cavity. If he has a cavity, then he might also have a toothache.

$$P(\text{Toothache}|\text{Catch}) > P(\text{Toothache})$$

- So Catch and Toothache are not independent

…but they are Conditionally Independent

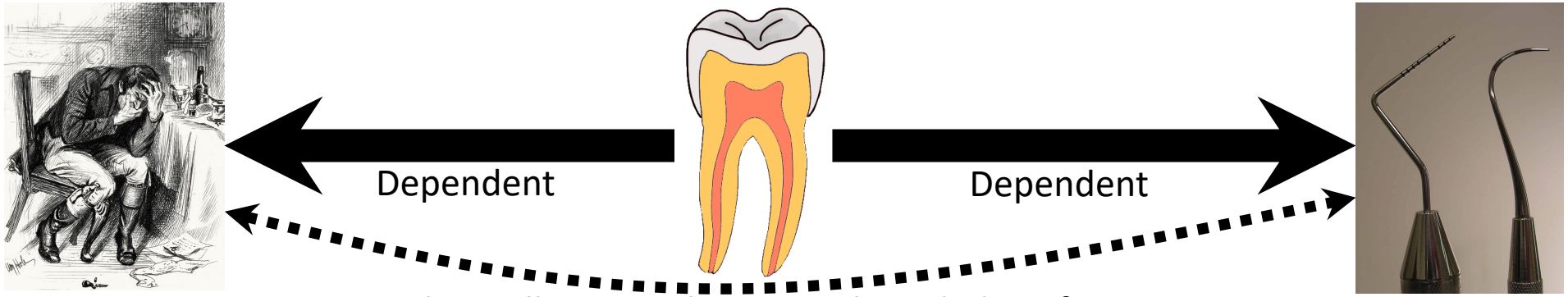


- Here are some reasons the probe might not catch, despite having a cavity:
  - The dentist might be really careless
  - The cavity might be really small
- Those reasons have nothing to do with the toothache!

$$P(\text{Catch}|\text{Cavity}, \text{Toothache}) = P(\text{Catch}|\text{Cavity})$$

- **Catch** and **Toothache** are **conditionally independent** given knowledge of **Cavity**

…but they are Conditionally Independent



These statements are all equivalent:

$$P(\text{Catch}|\text{Cavity}, \text{Toothache}) = P(\text{Catch}|\text{Cavity})$$

$$P(\text{Toothache}|\text{Cavity}, \text{Catch}) = P(\text{Toothache}|\text{Cavity})$$

$$P(\text{Toothache}, \text{Catch}|\text{Cavity}) = P(\text{Toothache}|\text{Cavity})\, P(\text{Catch}|\text{Cavity})$$

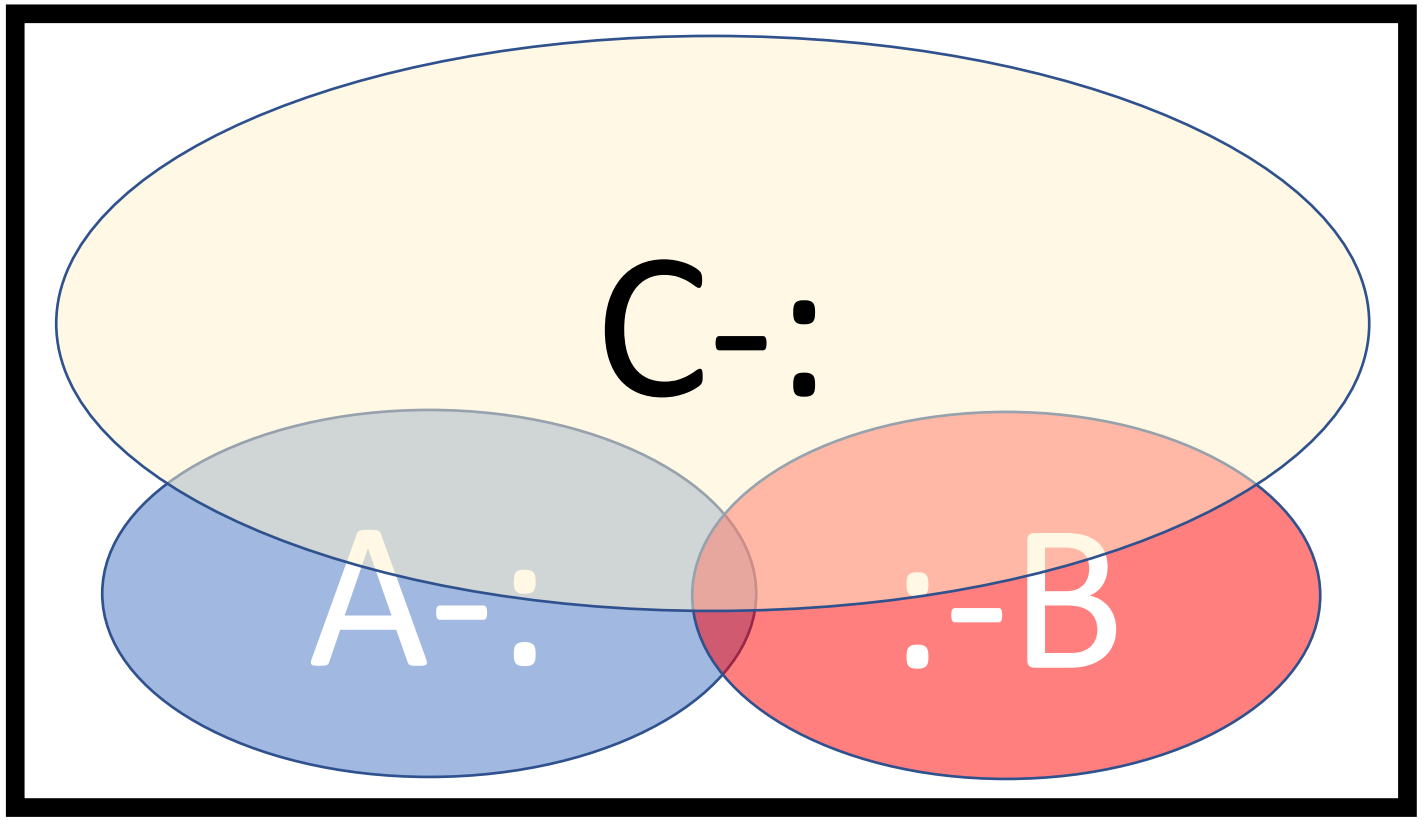
…and they all mean that **Catch** and **Toothache** are **conditionally independent** given knowledge of **Cavity**

Summary

Here's today's most important equation:

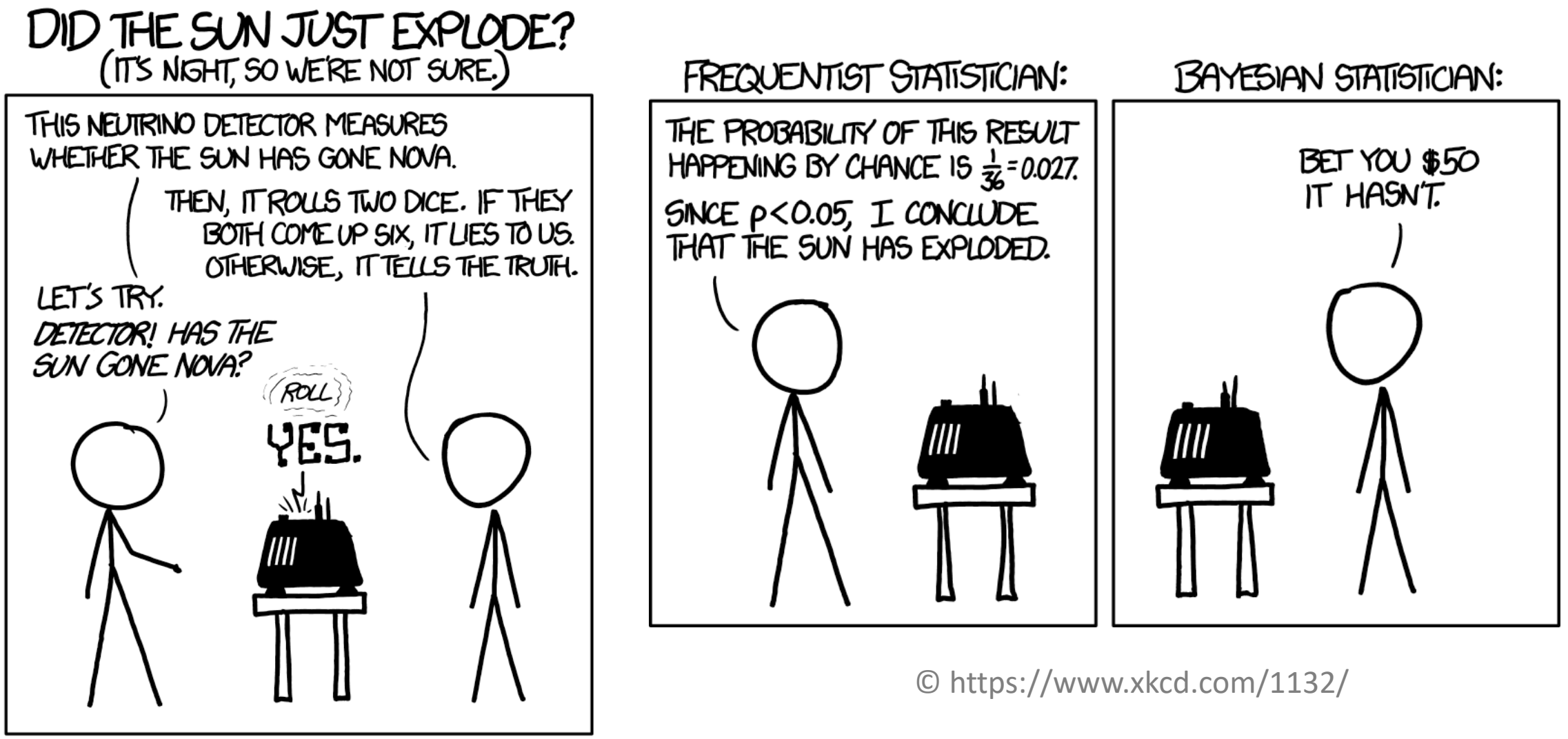
$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

If you haven't seen this stuff since high school, read appendix A.3.



Bayesian Decision Theory and Naïve Bayes Classifier

Credits: this section is partially adapted from Lecture 6 of CS440/CEC448 (UIUC) under CC-BY-4.0 license



© https://www.xkcd.com/1132/

Bayes' Rule

Rev. Thomas Bayes (1702-1761)

By Unknown - [2][3], Public Domain, https://commons.wikimedia.org/w/index.php?curid=14532025

- The product rule gives us two ways to factor a joint probability:

$$P(A,B) = P(B|A)P(A) = P(A|B)P(B)$$

- Therefore,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Why is this useful?
  - "A" is something we care about, but P(A|B) is really really hard to measure (example: the sun exploded)
  - "B" is something less interesting, but P(B|A) is easy to measure (example: the amount of light falling on a solar cell)
  - Bayes' rule tells us how to compute the probability we want (P(A|B)) from probabilities that are much, much easier to measure (P(B|A)).

Bayes Rule example

Eliot & Karson are getting married tomorrow, at an outdoor ceremony in the desert. Unfortunately, the weatherman has predicted rain for tomorrow.

- In recent years, it has rained (event R) only 5 days each year (5/365 = 0.014).

$$P(R) = 0.014$$

- When it actually rains, the weatherman forecasts rain (event F) 90% of the time.

$$P(F|R) = 0.9$$

- When it doesn't rain, he forecasts rain (event F) only 10% of the time.

$$P(F|\neg R) = 0.1$$

- What is the probability that it will rain on Eliot's wedding?

$$P(R|F) = \frac{P(F|R)P(R)}{P(F)} = \frac{P(F|R)P(R)}{P(F,R) + P(F,\neg R)} = \frac{P(F|R)P(R)}{P(F|R)P(R) + P(F|\neg R)P(\neg R)}$$

$$= \frac{(0.9)(0.014)}{(0.9)(0.014) + (0.1)(0.956)} = 0.116$$

The More Useful Version of Bayes' Rule

Rev. Thomas Bayes
(1702-1761)

By Unknown - [2][3], Public Domain, https://commons.wikimedia.org/w/index.php?curid=14532025

This version is what you memorize.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Remember, P(B|A) is easy to measure (the probability that light hits our solar cell, if the sun still exists and it's daytime). Let's assume we also know P(A) (the probability the sun still exists).
- But suppose we don't really know P(B) (what is the probability light hits our solar cell, if we don't really know whether the sun still exists or not?)

This version is what you actually use.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

The Misdiagnosis Problem

- 1% of women at age 40 who participate in routine screening have breast cancer.
- 80% of women with breast cancer will get positive mammographies.
- 9.6% of women without breast cancer will also get positive mammographies.
- Question: A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

$$P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})}$$

$$= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive} \mid \text{cancer})P(\text{cancer}) + P(\text{positive} \mid \neg\text{cancer})P(\neg\text{Cancer})}$$

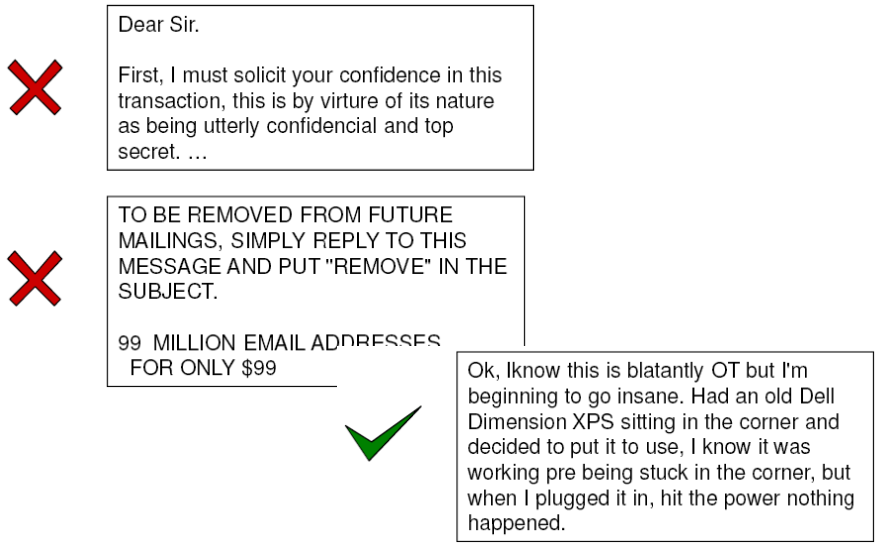
$$= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.096 \times 0.99} = \frac{0.008}{0.008 + 0.095} = 0.0776$$

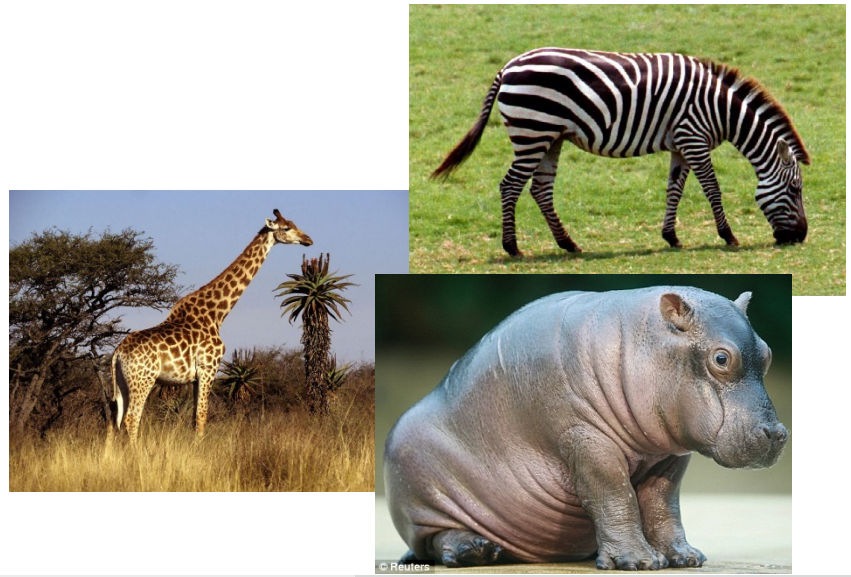
Second Opinions

If your doctor tells you that you have a health problem or suggests a treatment for an illness or injury, you might want a second opinion. This is especially true when you're considering surgery or major procedures.

Bayesian Decision Theory

- The agent is given some evidence, E.
- The agent has to make a decision about the value of an unobserved variable Y. Y is called the "class variable" or the "category" ("label").
  - Partially observable, stochastic, episodic environment
  - Example: $Y \in$ {spam, not spam}, E = email message.
  - Example: $Y \in$ {zebra, giraffe, hippo}, E = image features





Classification using probabilities

- Suppose you know that you have a toothache.
- Should you conclude that you have a cavity?
- Goal: make a decision that **minimizes your probability of error**.
- Equivalent: make a decision that **maximizes the probability of being correct**. This is called a MAP (maximum a posteriori) decision.

- You decide that you have a cavity if and only if

$$P(\text{Cavity}|\text{Toothache}) > P(\neg\text{Cavity}|\text{Toothache})$$

Label Y *Evidence E*

Bayesian Decisions

- What if we don't know $P(\text{Cavity}|\text{Toothache})$? Instead, we only know $P(\text{Toothache}|\text{Cavity})$, $P(\text{Cavity})$, and $P(\text{Toothache})$?
- Then we choose to believe we have a Cavity if and only if

$$P(\text{Cavity}|\text{Toothache}) > P(\neg\text{Cavity}|\text{Toothache})$$

Which can be re-written as

$$\frac{P(\text{Toothache}|\text{Cavity})P(\text{Cavity})}{P(\text{Toothache})} > \frac{P(\text{Toothache}|\neg\text{Cavity})P(\neg\text{Cavity})}{P(\text{Toothache})}$$

MAP decision

The action, “a”, should be the value of C that has the highest posterior probability given the observation $E = e$:

$$a = \text{argmax}\, P(Y = a|E = e) = \text{argmax}\frac{P(E = e|Y = a)P(Y = a)}{P(E = e)}$$
$$= \text{argmax}\, P(E = e|Y = a)P(Y = a)$$

$$P(Y = a|E = e) \propto P(E = e|Y = a)P(Y = a)$$

posterior likelihood prior

The Bayesian Terms

- $P(Y)$ is called the "**prior**" (*a priori*, in Latin) because it represents your belief about the query variable before you see any observation.
- $P(Y|E)$ is called the "**posterior**" (*a posteriori*, in Latin), because it represents your belief about the query variable after you see the observation.
- $P(E|Y)$ is called the "**likelihood**" because it tells you how much the observation, E=e, is like the observations you expect if Y=y.
- $P(E)$ is called the "**evidence distribution**" because E is the evidence variable, and $P(E)$ is its marginal distribution.

$$P(Y = y|E = e) = \frac{P(E = e|Y = y)P(Y = y)}{P(E = e)}$$

Naïve Bayes Classifier

- Suppose we have many different types of observations (symptoms, features) $E_1, \ldots, E_n$ that we want to use to obtain evidence about an underlying hypothesis Y.
- MAP decision:

$$P(Y|E_1, \ldots, E_n) \propto P(Y)P(E_1, \ldots, E_n|Y)$$

- THE BIG PROBLEM: If each feature E_i can take on K values, how many entries are in the probability table $P(E_1, \ldots, E_n|Y)$? How can we calculate this from finite training samples?

Naïve Bayes Classifier

- Suppose we have many different types of observations (symptoms, features) $E_1, \ldots, E_n$ that we want to use to obtain evidence about an underlying hypothesis Y.

- The naïve Bayes classifier makes the "attribute conditional independence assumption": **given any known class, assume all attributes are independent of each other**. *(assume that each attribute influences the prediction result independently.)*

$$
\begin{aligned}
a &= \operatorname{argmax} p(Y = a | E_1 = e_1, \ldots, E_n = e_n) \\
&= \operatorname{argmax} p(Y = a) p(E_1 = e_1, \ldots, E_n = e_n | Y = a) \\
&\approx \operatorname{argmax} p(Y = a) p(E_1 = e_1 | Y = a) \ldots p(E_n = e_n | Y = a)
\end{aligned}
$$

Naïve Bayes Classifier Example

- You are a robot in an animal shelter and must learn to discriminate Dogs from Cats. You are given the following examples. Classify a new example (Sound=Meow, Fur=Fine, and Color=Black) using naïve Bayes classifier.

Example	Sound	Fur	Color	Class
Example #1	Meow	Coarse	Brown	Dog
Example #2	Bark	Fine	Brown	Dog
Example #3	Bark	Coarse	Black	Dog
Example #4	Bark	Coarse	Black	Dog
Example #5	Meow	Fine	Brown	Cat
Example #6	Meow	Coarse	Black	Cat
Example #7	Bark	Fine	Black	Cat
Example #8	Meow	Fine	Brown	Cat

Naïve Bayes Classifier Example

- You are a robot in an animal shelter and must learn to discriminate Dogs from Cats. You are given the following examples. Classify a new example (Sound=Meow, Fur=Fine, and Color=Black) using naïve Bayes classifier.

Example	Sound	Fur	Color	Class
Example #1	Meow	Coarse	Brown	Dog
Example #2	Bark	Fine	Brown	Dog
Example #3	Bark	Coarse	Black	Dog
Example #4	Bark	Coarse	Black	Dog
Example #5	Meow	Fine	Brown	Cat
Example #6	Meow	Coarse	Black	Cat
Example #7	Bark	Fine	Black	Cat
Example #8	Meow	Fine	Brown	Cat

Priors: P(class=Cat) = 4/8 = 0.5; P(class=Dog) = 4/8 = 0.5

Conditional probability of the features:

P(Sound=Meow | class=Cat) = 3/4 = 0.75
P(Sound=Meow | class=Dog) = 1/4 = 0.25

P(Fur=Fine | class=Cat) = 3/4 = 0.75
P(Fur=Fine | class=Dog) = 1/4 = 0.25

P(Color=Black | class=Cat) = 2/4 = 0.5
P(Color=Black | class=Dog) = 2/4 = 0.5

P(c=Dog)P(Sound=Meow|c=Dog) P(Fur=Fine|c=Dog) P(Color=Black|c=Dog) =0.5*0.25*0.25*0.5=0.016

P(c=Cat)P(Sound=Meow|c= Cat) P(Fur=Fine|c= Cat) P(Color=Black|c= Cat) =0.5*0.75*0.75*0.5= 0.14

Since 0.14 > 0.016, the naïve Bayes classifier will classify this new example as a Cat.

Naïve Bayes Classifier Exericise

- Given the following dataset, what would be the prediction generated by a naïve Bayes classifier for the new sample (blue, round, small)?

	color	shape	size	class
1	red	square	big	+
2	blue	square	big	+
3	red	round	small	-
4	green	square	small	-
5	red	round	big	+
6	green	round	big	-

Naïve Bayes Classifier: Pros and Cons

- Pros:
  - Naïve Bayes classifier is fast and simple. We just calculate the probabilities.
  - Easy to implement.
  - Incremental learning: just update the probabilities when new data arrive.
  - Suitable for large datasets: since it is very efficient.
- Cons:
  - The attribute conditional independence assumption is a strong assumption and may not be satisfied by real data.
  - Cannot do regression (can only make predictions for categorical labels).
  - Zero frequency problem: cannot handle new values of a feature in the test set as naïve Bayes classifier will assign zero probability (Color=white in the previous example).

Case study: Text document classification

- **MAP decision:** assign a document to the class with the highest posterior
  P(class | document)

- Example: spam classification
  - Classify a message as spam if P(spam | message) > P(¬spam | message)

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

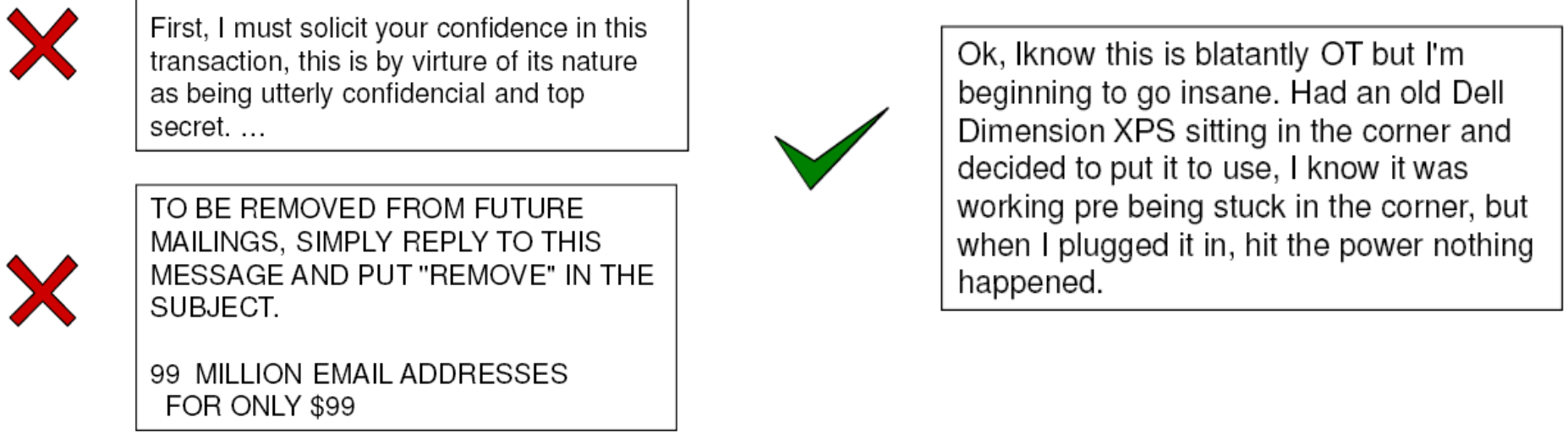
Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Case study: Text document classification

- **MAP decision:** assign a document to the class with the highest posterior
  $P(\text{class} \mid \text{document})$
- We have $P(\text{class} \mid \text{document}) \propto P(\text{document} \mid \text{class})P(\text{class})$
- To enable classification, we need to be able to estimate the likelihoods $P(\text{document} \mid \text{class})$ for all classes and priors $P(\text{class})$

Naïve Bayes Representation

- Goal: estimate likelihoods P(document | class) and priors P(class)
- Likelihood: ***bag of words*** representation
  - The document is a sequence of words $(w_1, \ldots, w_n)$
  - The order of the words in the document is not important
  - Each word is conditionally independent of the others given document class



Naïve Bayes Representation

- Goal: estimate likelihoods P(document | class) and priors P(class)
- Likelihood: ***bag of words*** representation
  - The document is a sequence of words ($E_1 = w_1, \ldots, E_n = w_n$)
  - The order of the words in the document is not important
  - Each word is conditionally independent of the others given document class

$$P(\text{document}|\text{class}) = P(w_1, \ldots, w_n|\text{class}) \approx \prod_{i=1}^{n} P(w_i|\text{class})$$

  - Thus, the problem is reduced to estimating marginal likelihoods of individual words $p(w_i \mid \text{class})$

Parameter estimation

- Model parameters: feature likelihoods p(word | class) and priors p(class)
- How do we obtain the values of these parameters?

prior

spam:	0.33
¬spam:	0.67

P(word | spam)

the :	0.0156
to :	0.0153
and :	0.0115
of :	0.0095
you :	0.0093
a :	0.0086
with:	0.0080
from:	0.0075
...	

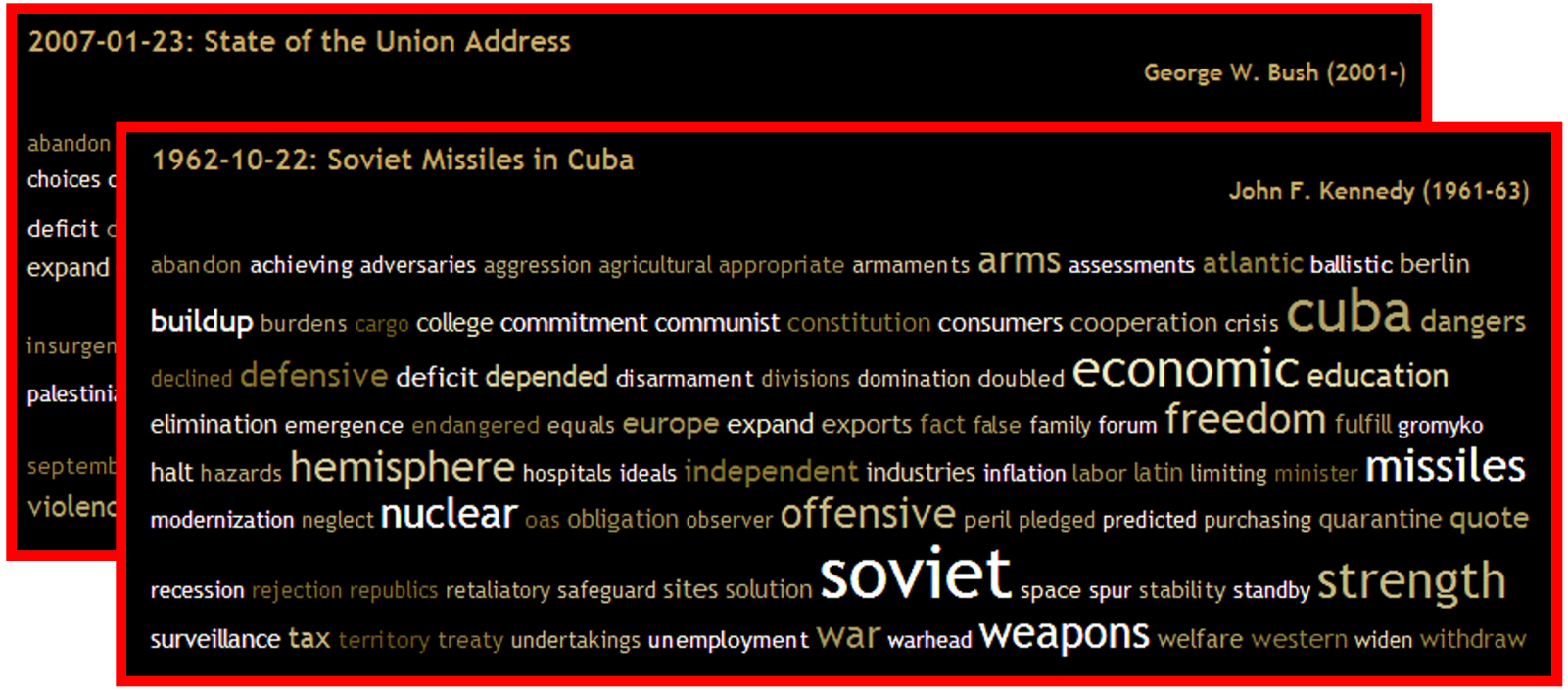
P(word | ¬spam)

the :	0.0210
to :	0.0133
of :	0.0119
2002:	0.0110
with:	0.0108
from:	0.0107
and :	0.0105
a :	0.0100
...	

Bag of words illustration

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

Bag of words illustration



US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

Bag of words illustration

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian "Decision"
  - The "Naïve Bayesian" Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Laplace Smoothing

Bayesian Learning

- Model parameters: feature likelihoods P(word | class) and priors P(class)
  - How do we obtain the values of these parameters?
  - Need *training set* of labeled samples from both classes

$$P(\text{word} \mid \text{class}) = \frac{\#\text{ of occurrences of this word in docs from this class}}{\text{total \# of words in docs from this class}}$$

  - This is the *maximum likelihood* (ML) estimate. It is the estimate that maximizes the probability of the training data, which is defined as:

$$\prod_{d=1}^{D} \prod_{i=1}^{n_d} P(w_{d,i} \mid class_{d,i})$$

d: index of training document, *i*: index of a word

Bayesian Learning

The data likelihood

$$\mathrm{P}(\text{training data}) = \prod_{d=1}^{D} \prod_{i=1}^{\#\, words\ in\ d} P\left(E = w_{d,i} | Y = c_d\right)$$

is maximized (subject to the constraint that $\sum_w P(w|c) = 1$) if we choose:

$$P(E = w|Y = c) = \frac{\#\text{ occurrences of word } w \text{ in documents of type } c}{\text{total number of words in all documents of type } c}$$

$$P(Y = c) = \frac{\#\text{ documents of type } c}{\text{total number of documents}}$$

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian "Decision"
  - The "Naïve Bayesian" Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Laplace Smoothing

What is the probability that the sun will fail to rise tomorrow?

- # times we have observed the sun to rise = 100,000,000
- # times we have observed the sun not to rise = 0
- Estimated probability the sun will not rise = $\frac{0}{0+100,000,000} = 0$

Oops….

Laplace Smoothing

- The basic idea: add 1 “unobserved observation” to every possible event
- # times the sun has risen or might have ever risen = 100,000,000+1 = 100,000,001
- # times the sun has failed to rise or might have ever failed to rise = 0+1 = 1
- Estimated probability the sun will not rise = $\frac{1}{1+100,000,001} = 0.000000009999998$

Parameter estimation

- ML (Maximum Likelihood) parameter estimate:

$$P(\text{word} \mid \text{class}) = \frac{\#\text{ of occurrences of this word in docs from this class}}{\text{total \# of words in docs from this class}}$$

- Laplacian Smoothing estimate
  - How can you estimate the probability of a word you never saw in the training set? (Hint: what happens if you give it probability 0, then it actually occurs in a test document?)
  - **Laplace smoothing:** pretend you have seen every vocabulary word one more time than you actually did

$$P(\text{word} \mid \text{class}) = \frac{\#\text{ of occurrences of this word in docs from this class} + 1}{\text{total \# of words in docs from this class} + V}$$

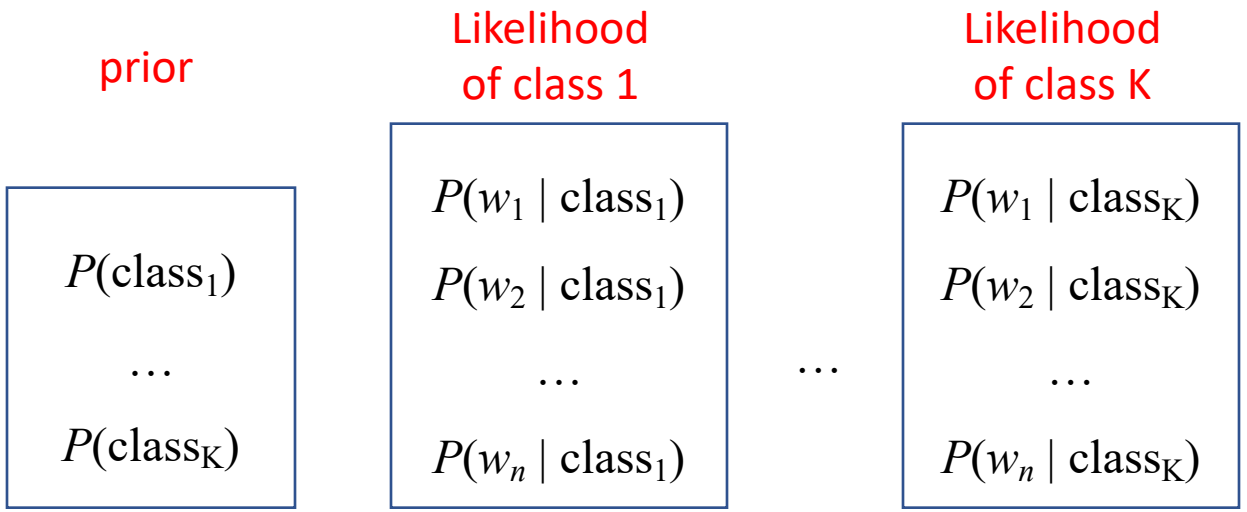
(V: total number of unique words)

Summary: Naïve Bayes for Document Classification

- Naïve Bayes model: assign the document to the class with the highest posterior

$$P(\text{class}|\text{document}) \propto P(\text{class}) \prod_{i=1}^{n} P(w_i|\text{class})$$

- Model parameters:



Bayesian Networks

Credits: this section is partially adapted from Lecture 13&14 of CS440/CEC448 (UIUC) under CC-BY-4.0 license

Bayesian Classifier

- Class label $Y = y$, drawn from some set of labels
- Observation $X = x$, drawn from some set of features
- Bayesian classifier: choose the class label, y, that minimizes your probability of making a mistake:

$$\hat{y} = \underset{y}{\operatorname{argmin}} P(Y \neq y | X = x)$$

Minimum Probability of Error = Maximum A Posteriori

- The minimum probability of error (MPE) classifier is the one that minimizes your probability of making a mistake:

$$\hat{y} = \underset{y}{\operatorname{argmin}}\, P(Y \neq y | X = x)$$

- The maximum a posteriori (MAP) classifier is the one that maximizes your probability of being correct:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(Y = y | X = x)$$

- Notice: they're the same! This is called the MPE=MAP rule.

Beyond naïve assumptions: What if P(X, Y) is complicated?

Very, very common problem: P(X,Y) is complicated because both X and Y depend on some hidden variable H

$$P(Y = y|X = x) = \frac{\sum_h P(X = x, H = h, Y = y)}{\sum_{h,y'} P(X = x, H = h, Y = y')}$$

Why is this a problem?

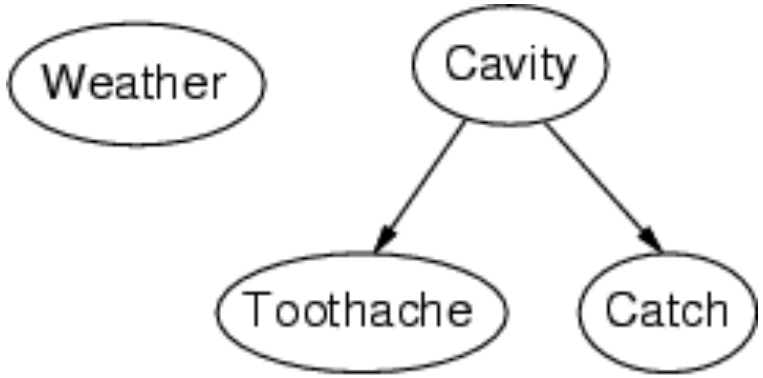
1. **SPACE COMPLEXITY**: $P(X = x, H = h, Y = y)$ requires $|X| \cdot |H| \cdot |Y|$ entries
   - Example: X has cardinality 1000, H has cardinality 1000, Y has cardinality 1000, then $P(X = x, H = h, Y = y)$ is a probability table with 1 billion entries.

2. **TIME COMPLEXITY**: The summation requires a lot of time.

Bayesian networks: Structure

- **Nodes:** random variables

- **Arcs:** interactions
  - An arrow from one variable to another indicates direct ***causal*** influence of variable #1 on variable #2
  - Must form a directed, acyclic graph



Conditional independence and the joint distribution

- Key property: each node is conditionally independent of its *non-descendants* given its *parents*
- Suppose the nodes $X_1, \ldots, X_n$ are sorted in topological order
- To get the joint distribution $P(X_1, \ldots, X_n)$, use chain rule:

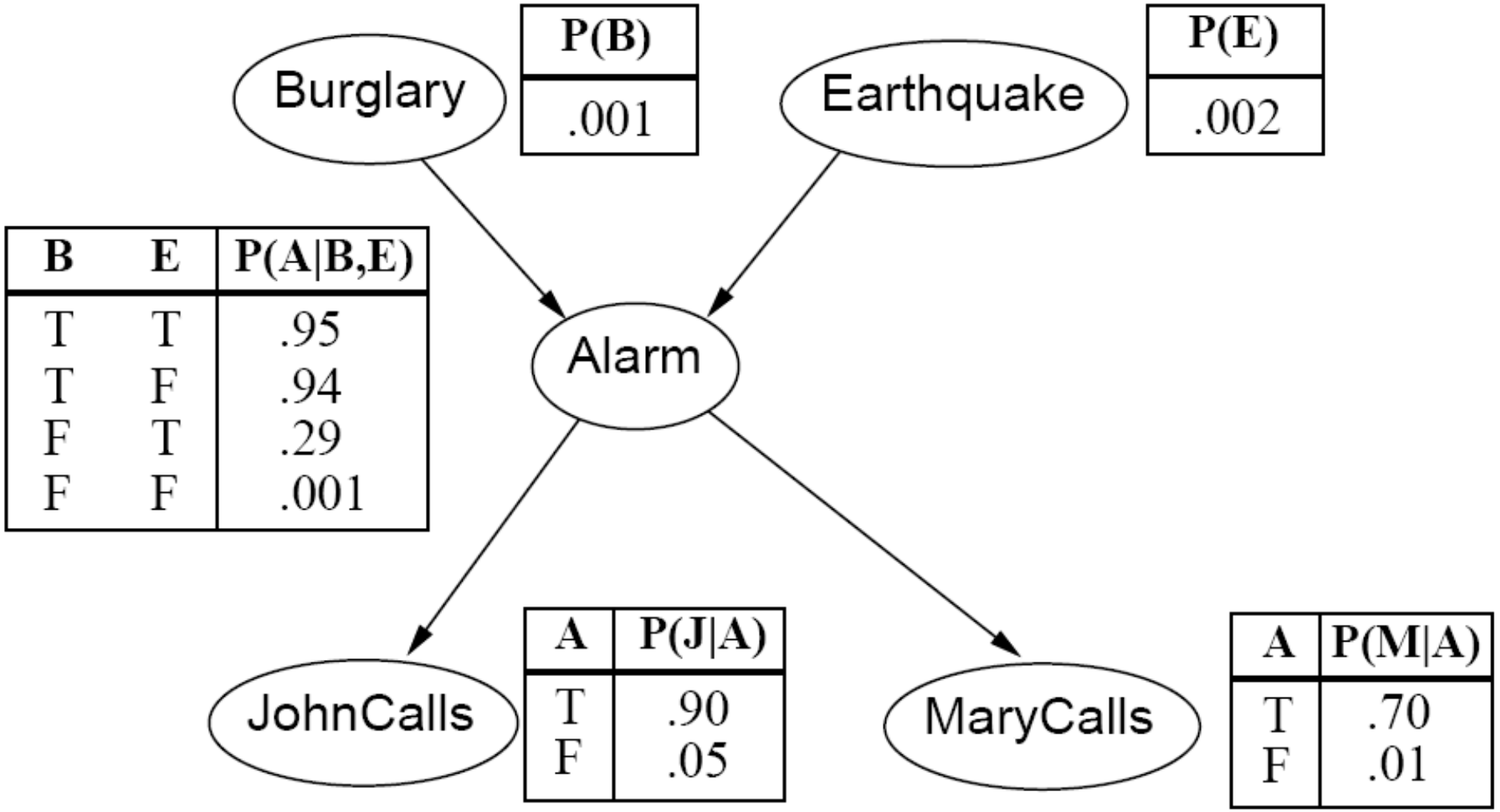
$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid X_1,\ldots,X_{i-1}\right)$$

$$= \prod_{i=1}^{n} P\left(X_i \mid Parents(X_i)\right)$$

Example: Los Angeles Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  - Example inference task: suppose Mary calls and John doesn't call. What is the probability of a burglary?
- What are the random variables?
  - *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- What are the direct influence relationships?
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example: Burglar Alarm



Space complexity: LA Burglar Alarm

- How much space do we need to store the model without dependencies?
  - 5 variables
  - Each is binary
  - $P(B, E, A, J, M)$ is a table with $2^5 = 32$ entries
  - Since they add up to 1, we could store just $2^5 - 1 = 31$ entries
- How much space do we need to store the Bayes net parameters?
  - $P(B), P(E)$: two numbers
  - $P(A|B = b, E = e)$: one entry for each setting of $b \in \{F, T\}$, $e \in \{F, T\}$
  - $P(J|A = a), P(M|A = a)$: two numbers for each setting of $a \in \{F, T\}$
  - Total: $1 + 1 + 4 + 2 + 2 = 10$ entries

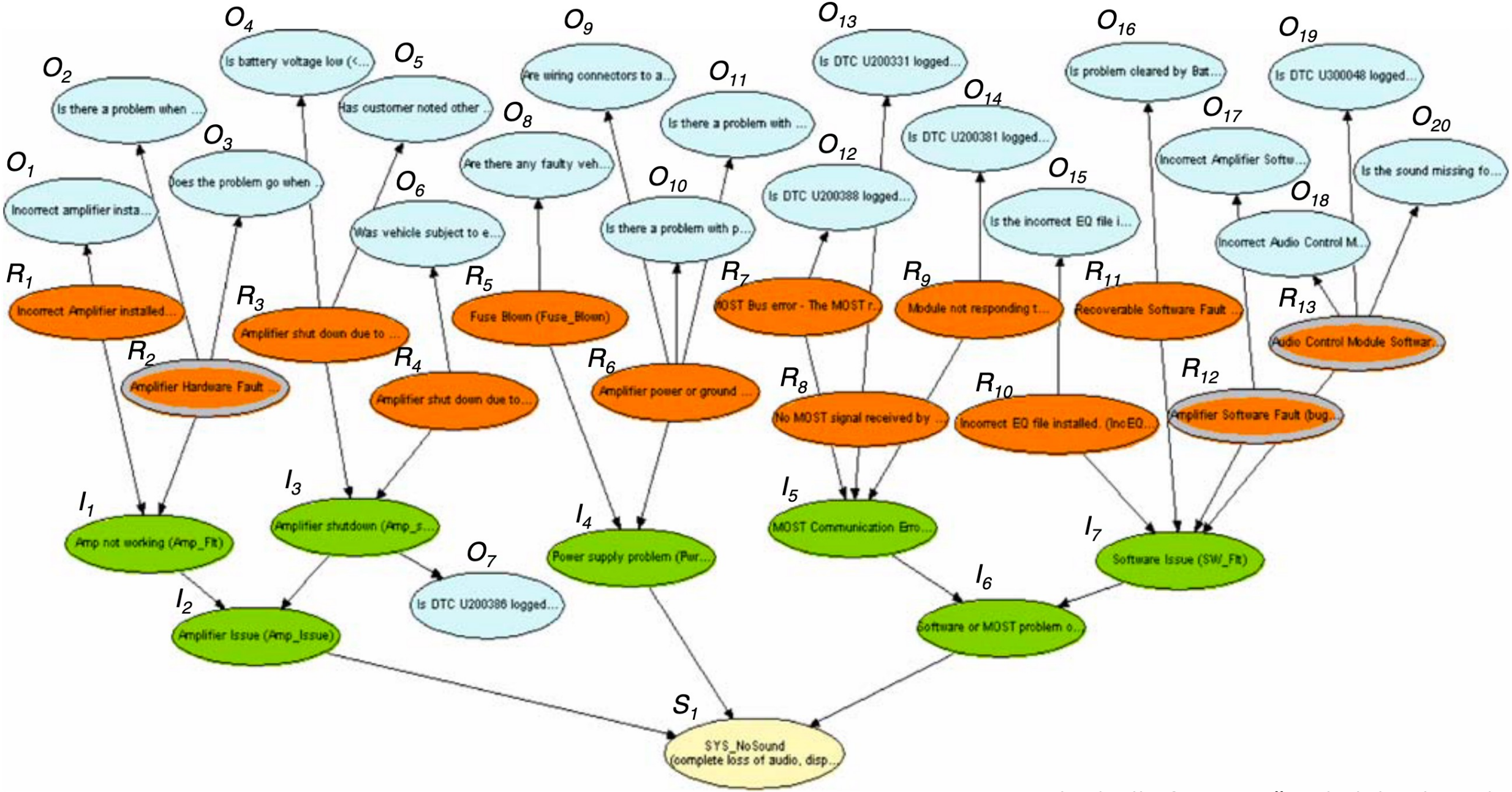


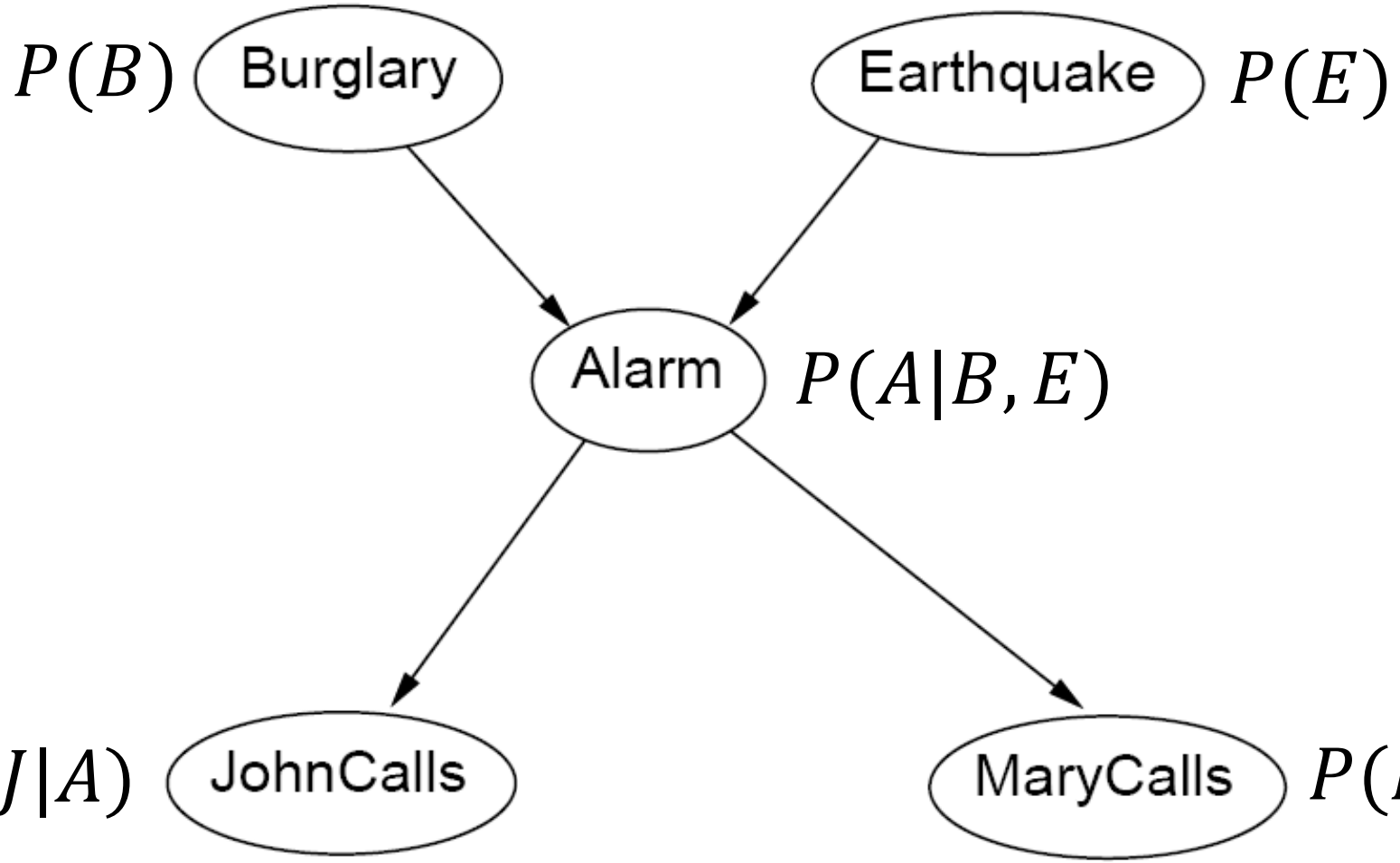
Fig. 6 Bayesian diagnostic model for the symptom "no sound"

Huang, McMurran, Dhadyalla & Jones, "Probability-based vehicle fault diagnosis: Bayesian network method," 2008

Space complexity, Huang et al. “no sound” diagnosis model

- How much space do we need to store the model without dependencies?
  - 41 binary variables: table would require $2^{41} - 1 = 2{,}199{,}023{,}255{,}551$ entries
- How much space do we need to store the Bayes net parameters?
  - One binary variable with four binary parents, requires one entry for each of the $2^4 = 16$ values of its parent variables
  - Two binary variable with three binary parents, each require 8 entries
  - Five binary variables with two binary parents, each require 4 entries
  - Twenty binary variables with one binary parent, each require 2 entries
  - Thirteen binary variables with no parents, each require 1 entry
  - Total: $16 + 2{\times}8 + 5{\times}4 + 20{\times}2 + 13 = 105$ entries

Example: Burglar Alarm



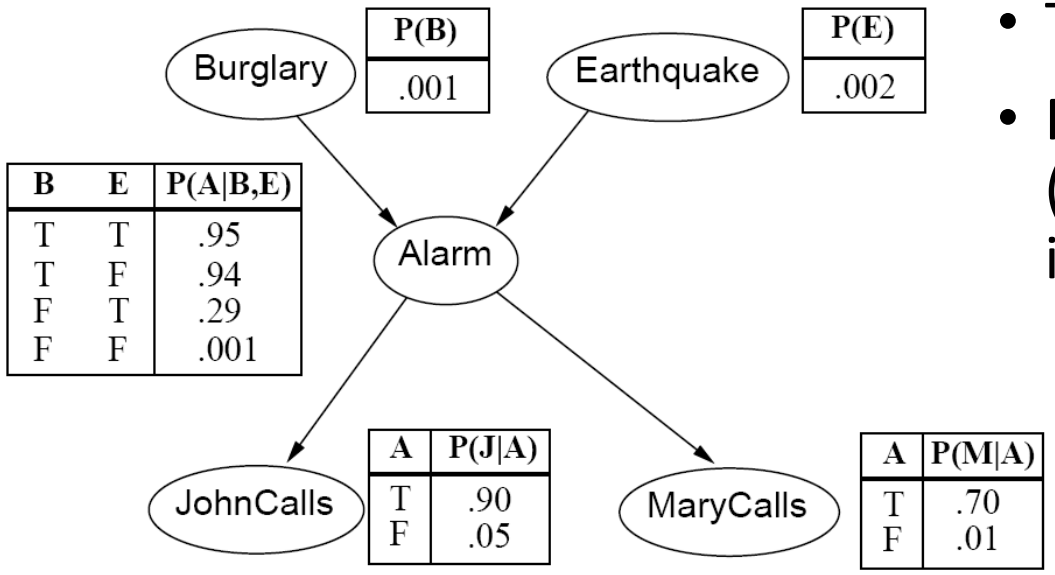
- A "model" is a complete specification of the dependencies.
- The conditional probability tables are the *model parameters.*

Classification using probabilities

- Suppose Mary has called to tell you that you had a burglar alarm. Should you call the police?
  - Make a decision that **maximizes the probability of being correct**. This is called a MAP (maximum a posteriori) decision. You decide that you have a burglar in your house if and only if

$$P(Burglary|Mary) > P(\neg Burglary|Mary)$$

Using a Bayes network to estimate a posteriori probabilities

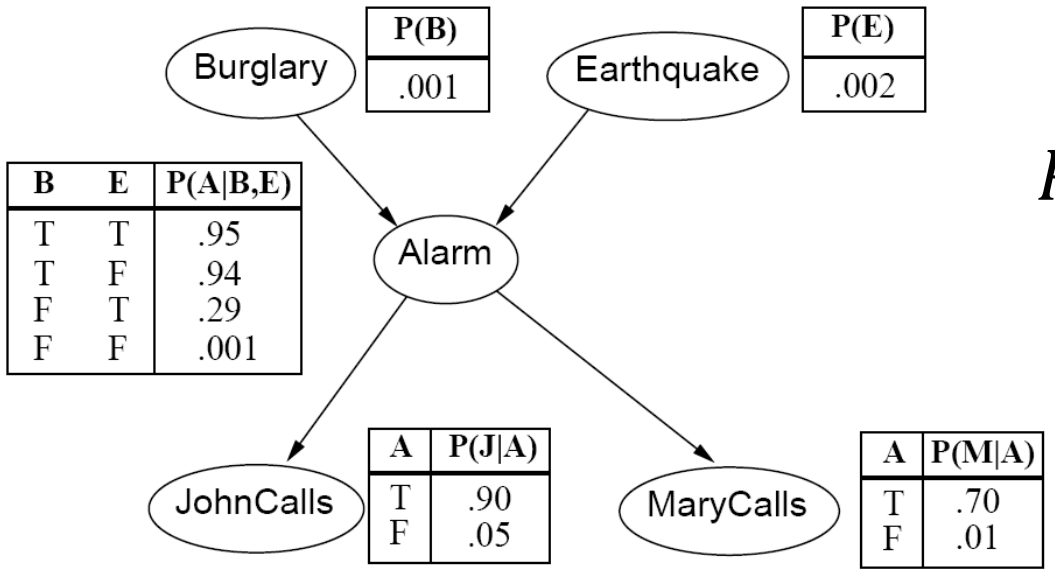


- Notice: we don't know $P(B|M)$! We have to figure out what it is.
- This is called "inference".
- First step: find the joint probability of B (and $\neg B$), M (and $\neg M$), and any other variables that are necessary in order to link these two together.

$$P(B,E,A,M) = P(B)P(E)P(A|B,E)P(M|A)$$

$P(BEAM)$	$\neg M, \neg A$	$\neg M, A$	$M, \neg A$	M, A
$\neg B, \neg E$	0.986045	2.99×10^{-4}	9.96×10^{-3}	6.98×10^{-4}
$\neg B, E$	1.4×10^{-3}	1.7×10^{-4}	1.4×10^{-5}	4.06×10^{-4}
$B, \neg E$	5.93×10^{-5}	2.81×10^{-4}	5.99×10^{-7}	6.57×10^{-4}
B, E	9.9×10^{-8}	5.7×10^{-7}	10^{-9}	1.33×10^{-6}

Using a Bayes network to estimate a posteriori probabilities



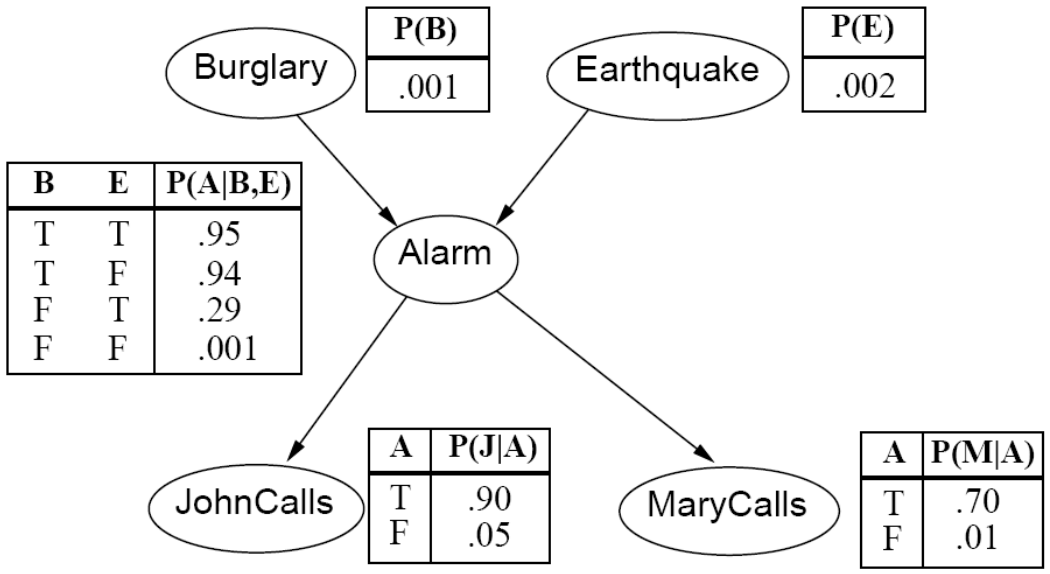
Second step: marginalize (add) to get rid of the variables you don't care about.

$$P(B, M) = \sum_{e \in \{F,T\}} \sum_{a \in \{F,T\}} P(B, E = e, A = a, M)$$

$P(B, M)$	$\neg M$	M
$\neg B$	0.987922	0.011078
B	0.000341	0.000659

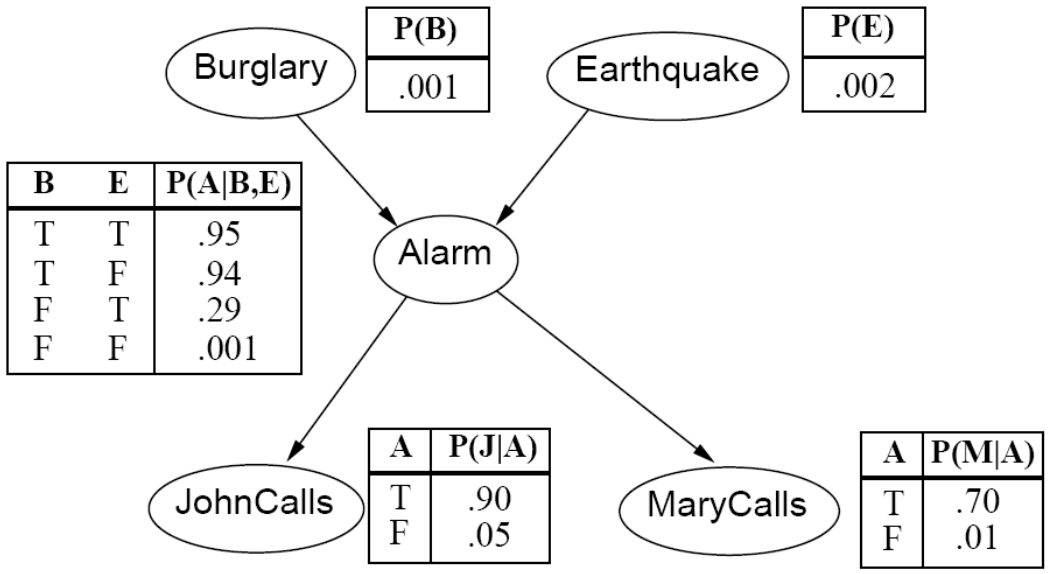
Using a Bayes network to estimate a posteriori probabilities

Third step: ignore (delete) the column that didn't happen.



$P(B, M)$	$\boldsymbol{M}$
$\neg B$	0.011078
B	0.000659

Using a Bayes network to estimate a posteriori probabilities



Fourth step: use the definition of conditional probability.

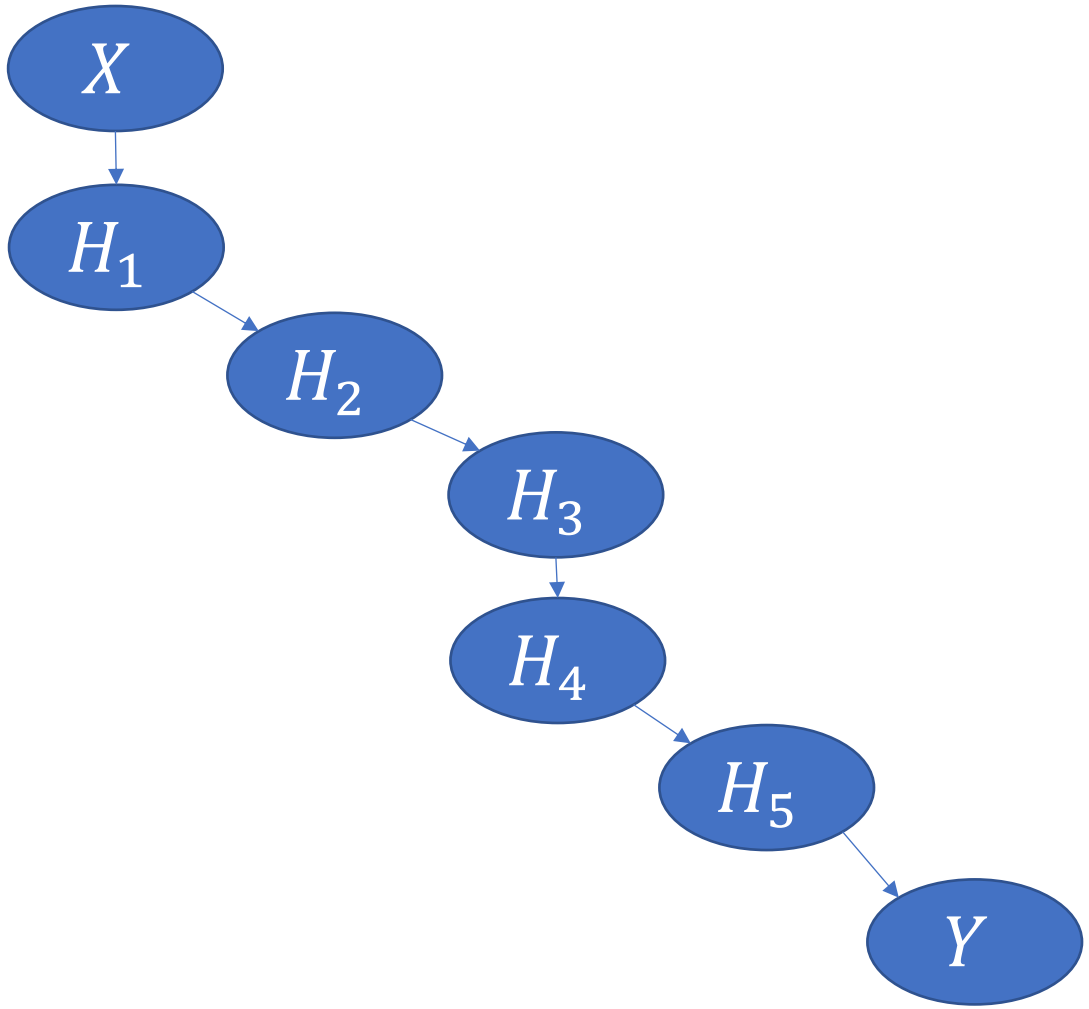
$$P(B|M) = \frac{P(B,M)}{P(B,M) + P(\neg B,M)}$$

$P(B\|M)$	M
$\neg B$	0.943883
B	0.056117

Some unexpected conclusions

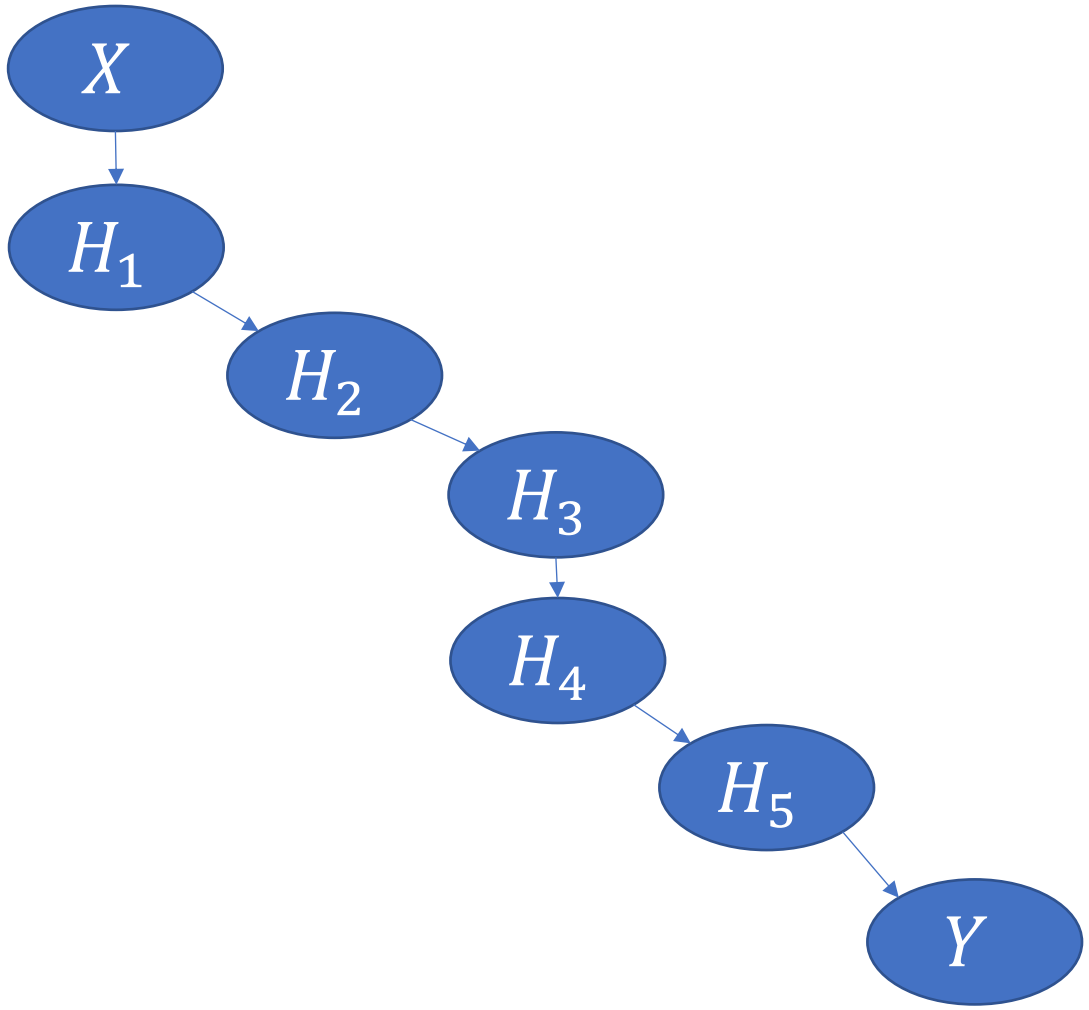
- Burglary is so unlikely that, if only Mary calls or only John calls, the probability of a burglary is still only about 5%.
- If both Mary and John call, the probability is ~50%.

Belief propagation: The general algorithm



Given an arbitrary Bayes net, you want to find the joint probability of two variables, X and Y, that are connected by a chain of intermediate variables, H_1 through H_N.

Belief propagation: The general algorithm



Initialize:

Start with P(X)

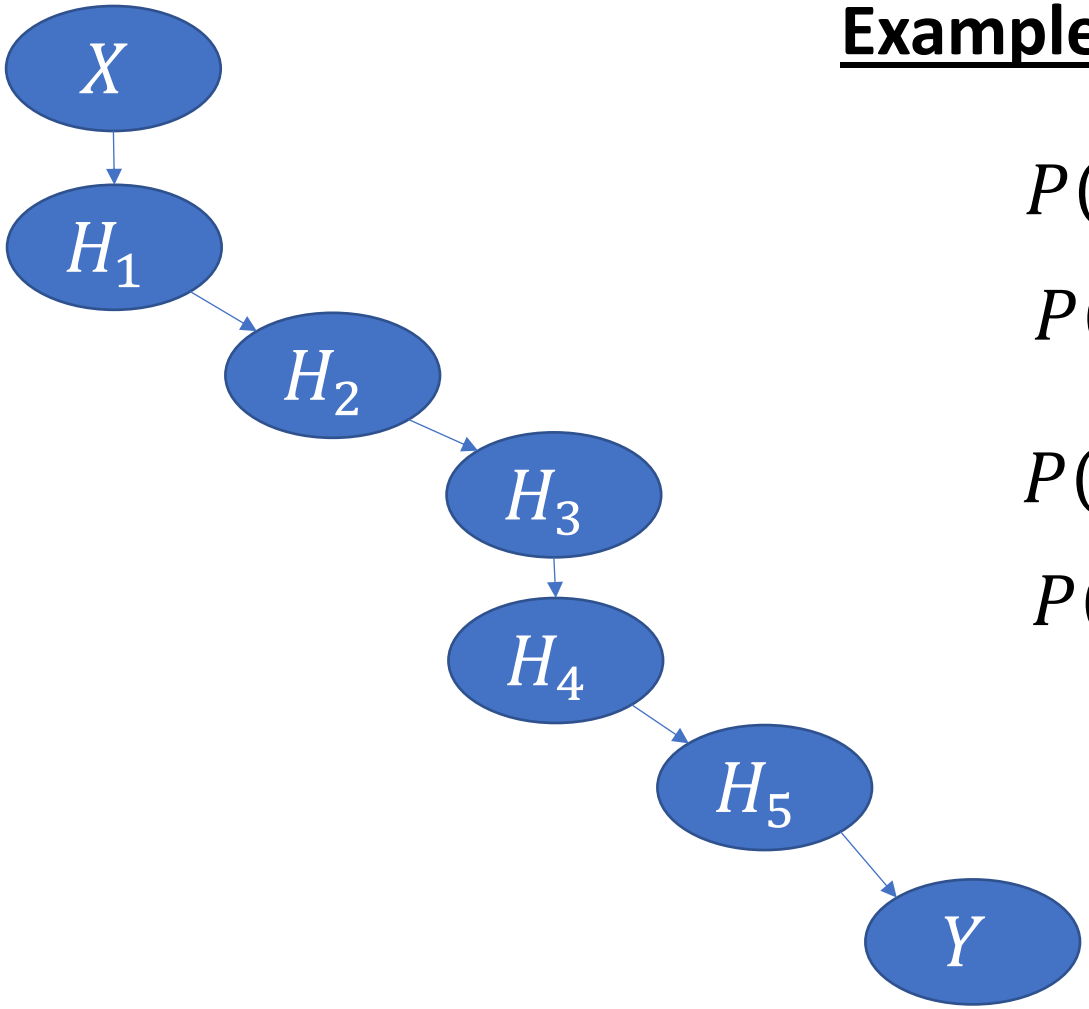
Iterate:

1. PRODUCT: Multiply in the next variable
2. SUM: Marginalize out any variables you no longer need

Terminate:

When you have P(X,Y)

Belief propagation: The general algorithm



Example:

$$P(X, H_1) = P(X)P(H_1|X)$$

$$P(X, H_1, H_2) = P(X, H_1)P(H_2|H_1)$$

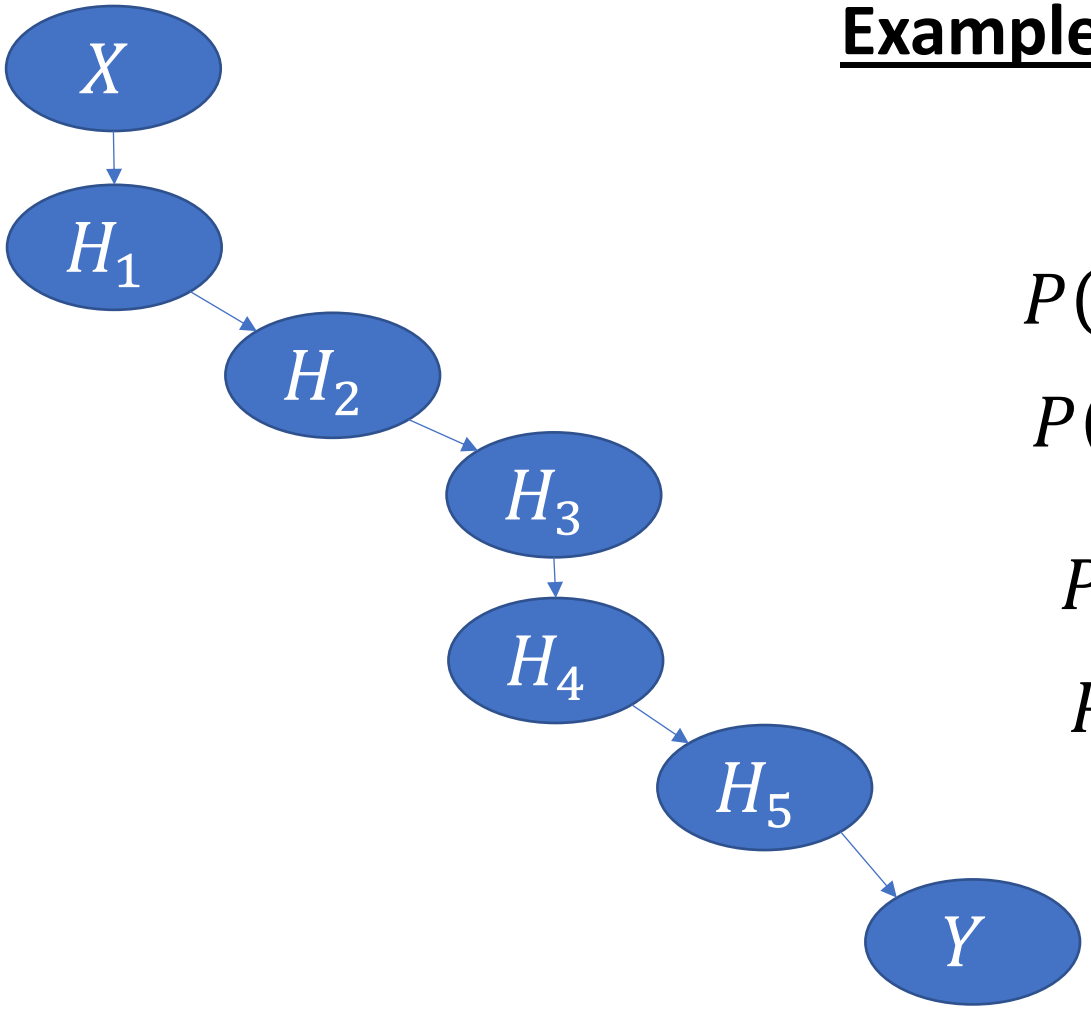
$$P(X, H_2) = \sum_{h_1} P(X, H_1 = h_1, H_2)$$

$$P(X, H_2, H_3) = P(X, H_2)P(H_3|H_2)$$

$$P(X, H_3) = \sum_{h_2} P(X, H_2 = h_2, H_3)$$

$$\vdots$$

Belief propagation: The general algorithm



Example:

$$\vdots$$

$$P(X, H_4, H_5) = P(X, H_4)P(H_5|H_4)$$

$$P(X, H_5) = \sum_{h_4} P(X, H_4 = h_4, H_5)$$

$$P(X, H_5, Y) = P(X, H_5)P(Y|H_5)$$

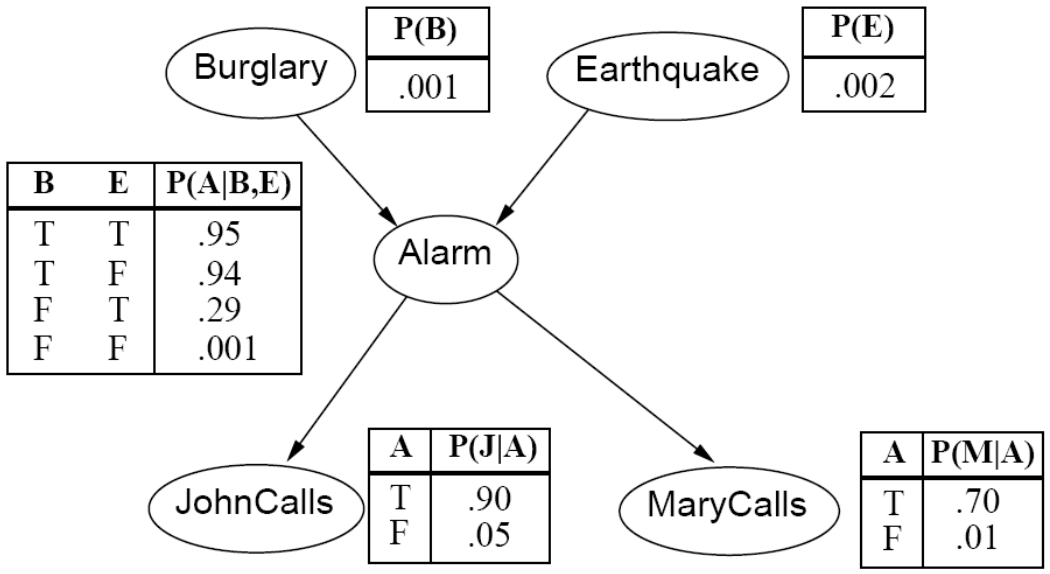
$$P(X, Y) = \sum_{h_5} P(X, H_5 = h_5, Y)$$

Belief propagation: Space and time complexity

- If there is just one path from X to Y (as shown in the example), then space and time complexity of belief propagation are each K^3, where K is the maximum cardinality of any of the random variables.
  - Each product operation results in a table of 3 variables, with $K^3 - 1$ entries
  - Each summation is over K entries, for each of K^2 combinations

- If there are multiple paths from X to Y, or if there are multiple X variables (many different relevant observations), then belief propagation becomes NP-complete
  - It's necessary to create a probability table containing all the variables in all the paths between X and Y
  - That table has $K^{2N+1} - 1$ entries, where N is the number of different paths that connect X to Y

The Los Angeles Burglar Alarm



Fourth step: use the definition of conditional probability.

$$P(B|M) = \frac{P(B,M)}{P(B,M) + P(B,\neg M)}$$

$P(B\|M)$	M
$\neg B$	0.943883
B	0.056117

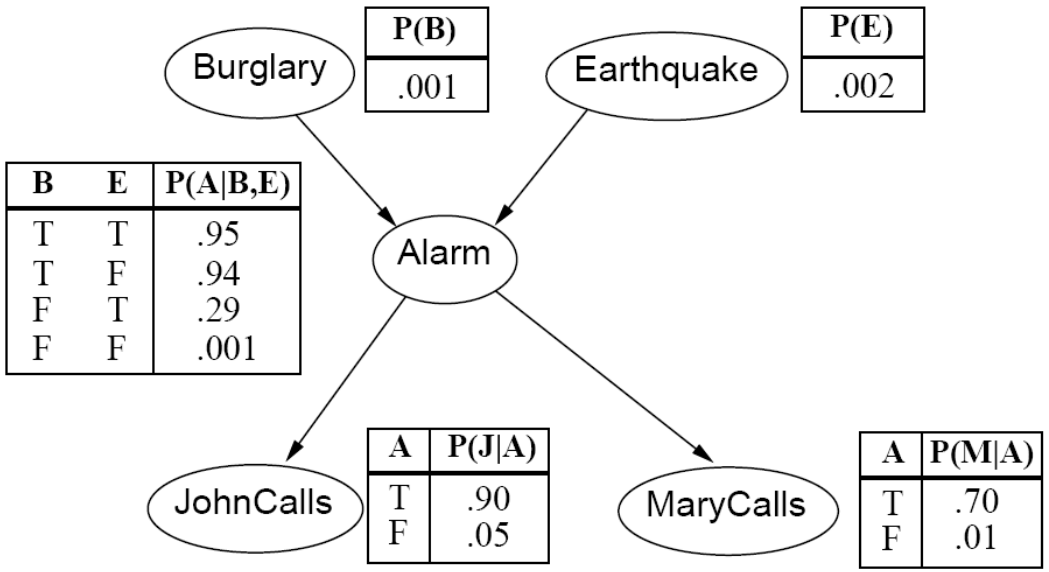
Some unexpected conclusions

- If only Mary calls or only John calls, the probability of a burglary is about 5% or 6%.

unless …

- If you know that there was an earthquake, then it's very likely that the alarm was caused by the earthquake. In that case, the probability you had a burglary is vanishingly small, even if twenty of your neighbors call you.
- This is called the "explaining away" effect. The earthquake "explains away" the burglar alarm.

The "Explaining Away" Effect



Probability of a Burglary, given that Mary called, and given a known earthquake:

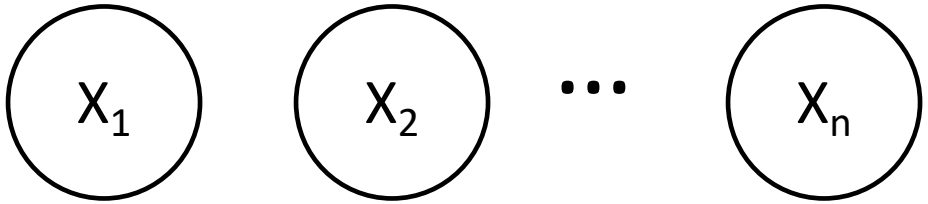
$$P(B|M,E) = \frac{\sum_{a\in\{F,T\}} P(M, A=a, E, B)}{\sum_{a\in\{F,T\},b\in\{F,T\}} P(M, A=a, E, B=b)}$$

$$= \frac{(0.001)(0.002)(0.95)(0.7) + (0.001)(0.002)(0.05)(0.01)}{\begin{pmatrix}(0.001)(0.002)(0.95)(0.7) + (0.001)(0.002)(0.05)(0.01)\\ +(0.999)(0.002)(0.29)(0.7) + (0.999)(0.002)(0.71)(0.01)\end{pmatrix}}$$

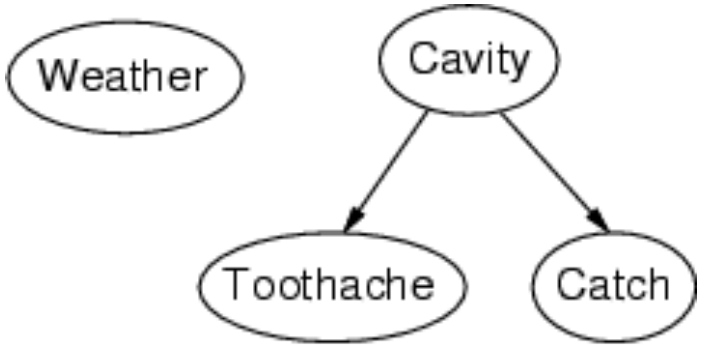
$$= 0.003$$

Independence

- By saying that X_i and X_j are independent, we mean that
$$\mathrm{P}(X_j, X_i) = \mathrm{P}(X_i)\mathrm{P}(X_j)$$
- X_i and X_j are independent if and only if they have no common ancestors
- Example: *independent coin flips*



- Another example: Weather is independent of all other variables in this model.

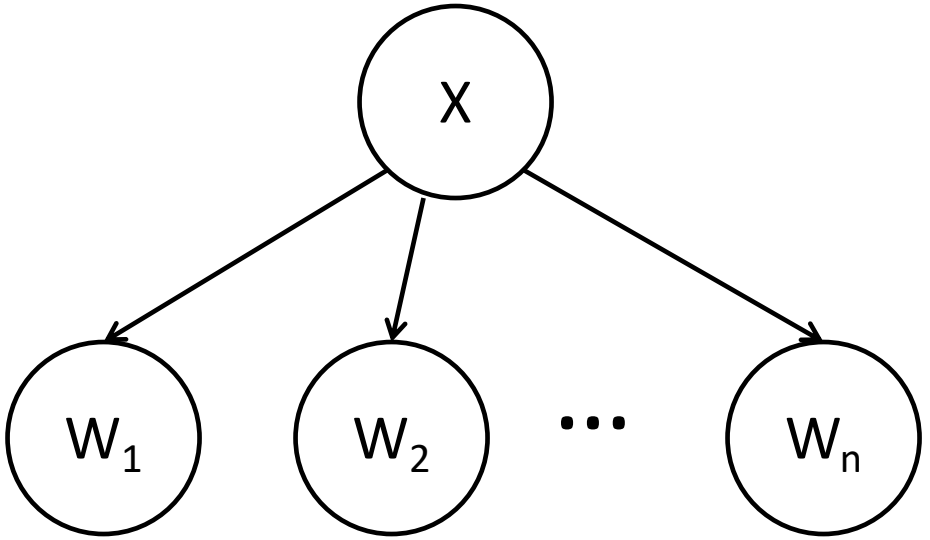


Conditional independence

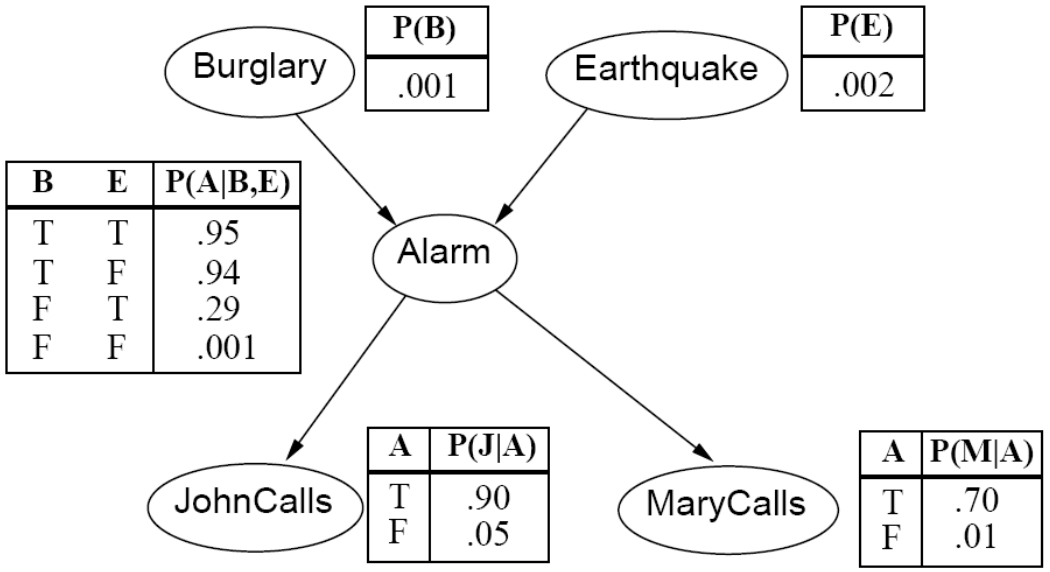
- By saying that W_i and W_j are conditionally independent given X, we mean that

$$\mathrm{P}(W_i, W_j | X) = \mathrm{P}(W_i | X)\mathrm{P}(W_j | X)$$

- W_i and W_j are conditionally independent given X if and only if they have no common ancestors other than the ancestors of X.
- Example: *naïve Bayes model:*



Conditional Independence ≠ Independence



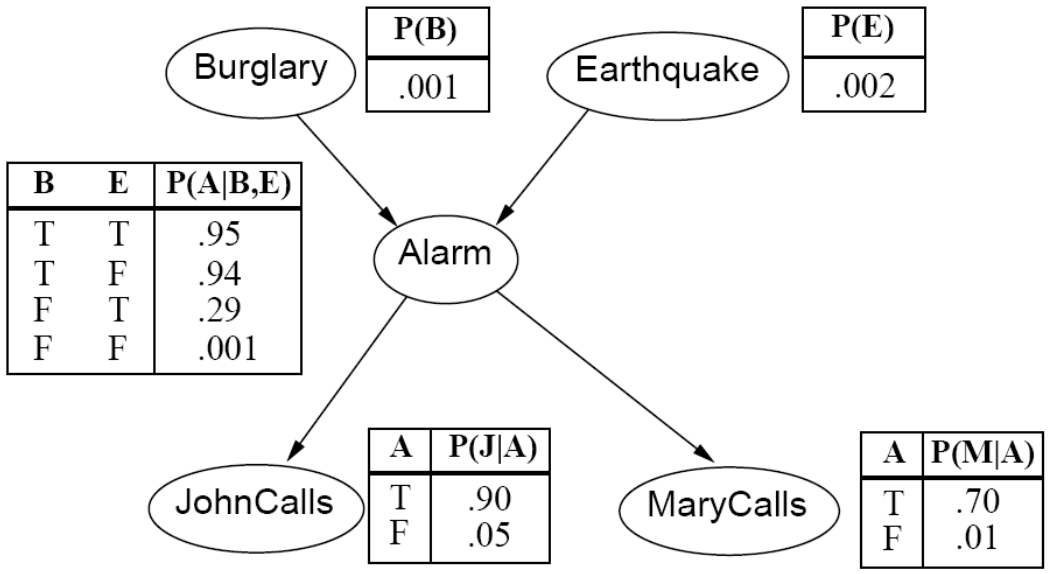
B and E are **independent**:

$$P(B|\neg E) = P(B) = 0.001$$

B and E are **not conditionally independent given A**:

$$P(B|\neg E, A) = 0.48 \neq P(B|\neg E) = 0.001$$

Conditional Independence ≠ Independence



J and M are **conditionally independent given A:**

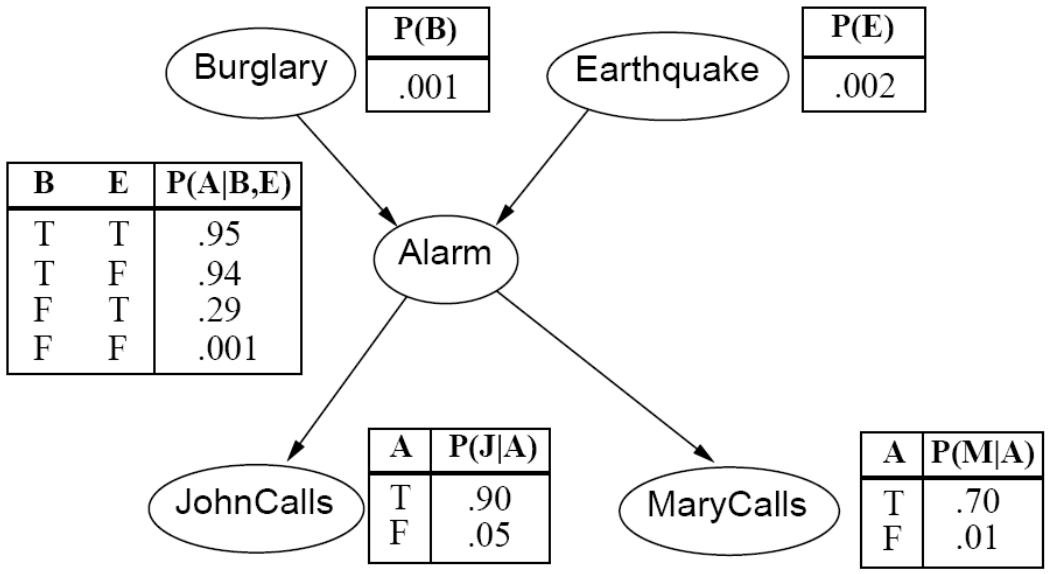
$$P(J|A,M) = P(J|A) = 0.9$$

$$P(M|A,J) = P(M|A) = 0.7$$

J and M are **not independent**!

$$P(J|M) = 0.18 \neq P(J) = 0.05$$

Conditional Independence ≠ Independence



B and M are **conditionally independent given A:**

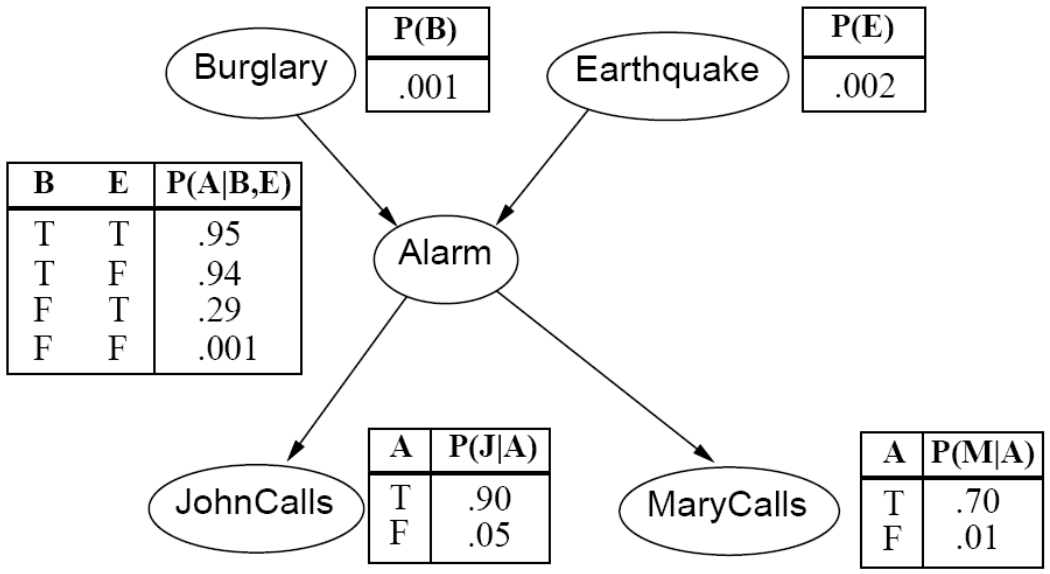
$$P(B|A,M) = P(B|A) = 0.37$$

$$P(M|A,B) = P(M|A) = 0.7$$

B and M are **not independent**!

$$P(B|M) = 0.056 \neq P(B) = 0.001$$

Conditional Independence ≠ Independence



- B and E (no common ancestor):
  - Independent
  - Not conditionally independent given A
- J and M (common ancestor):
  - Conditionally independent given A
  - Not independent
- B and M (one is ancestor of the other):
  - Conditionally independent given A
  - Not independent

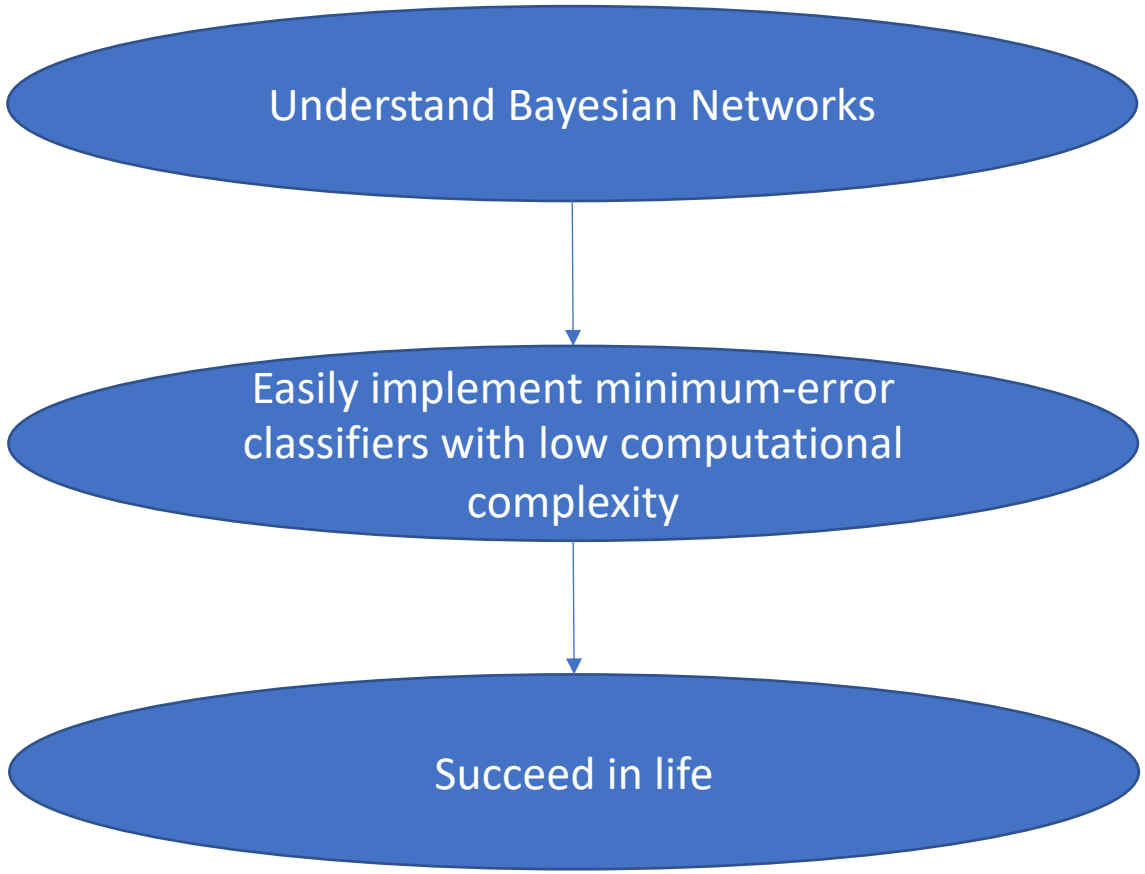
Conditional Independence ≠ Independence

- Variables in a Bayes net are **independent** if they have no common ancestors
  - If they have a common ancestor (e.g., J and M), they are not independent
  - If one is the ancestor of the other (e.g., B and M), they are not independent

- Variables in a Bayes net are **conditionally independent** given knowledge of:
  - Their common ancestors, and
  - A variable that is a descendant of one, and an ancestor of the other

Summary so far

- Why Bayes nets? The complexity of a true Bayes classifier
- Space complexity
- Time complexity
- Independence and Conditional independence



Parameter and Structure Learning for Bayesian Networks

- Parameter Learning
  - from Fully Observed data: Maximum Likelihood
  - from Partially Observed data: Expectation Maximization
  - from Partially Observed data: Hard EM

- Structure Learning
  - The usual method: knowledge engineering
  - An interesting recent method: causal analysis

Flying cows

The scenario:

Central Illinois has recently had a problem with flying cows.

Farmers have called to complain that their cows flew away.

Flying cows

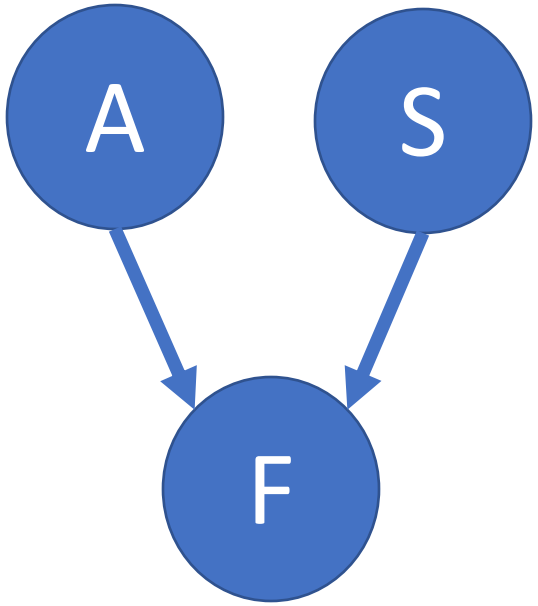
A team of experts determined that almost all flying cows were explained by one or both of the following causes:

- **Smart cows**. The cows learned how to fly, on their own, without help.
- **Alien intervention**. UFOs taught the cows how to fly.

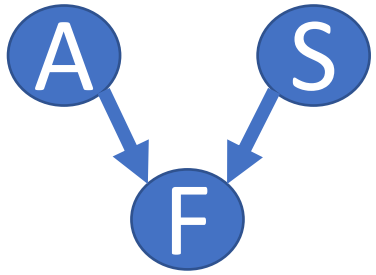
Flying cows

The experts created a Bayes net, to help them predict any future instances of cow flying:

- P(A) = Probability that aliens teach the cow.
- P(S) = Probability that a cow is smart enough to figure out how to fly on its own.
- P(F|S,A) = Probability that a cow learns to fly.



Flying cows

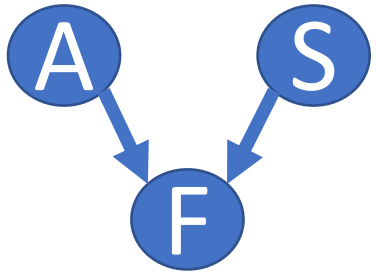


They went out to watch a nearby pasture for ten days.

- They reported the number of days on which A, S, and/or F occurred.
- Their results are shown in the table at left (True is marked as "T"; False is shown with a blank).

Day	A	S	F
1			
2		T	T
3			
4	T	T	T
5	T		
6			
7	T		T
8			
9			T
10			

Flying cows



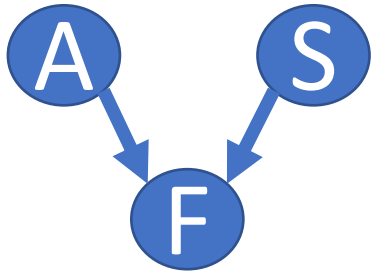
The experts now wish to estimate the parameters of their Bayes net

- P(A)
- P(S)
- P(F|S,A)

…so that they will be better able to testify before Congress about the relative dangers of aliens versus smart cows.

Day	A	S	F
1			
2		T	T
3			
4	T	T	T
5	T		
6			
7	T		T
8			
9			T
10			

Maximum Likelihood Estimation

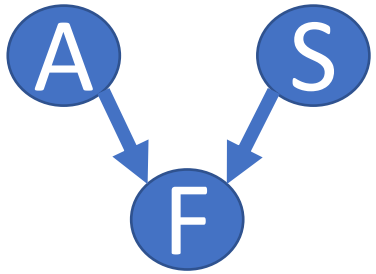


Suppose we have n training examples, $1 \leq i \leq n$, with known values for each of the random variables:

- A_i or $\neg A_i$
- S_i or $\neg S_i$
- F_i or $\neg F_i$

Day	A	S	F
1	$\neg A_1$	$\neg S_1$	$\neg F_1$
2	$\neg A_2$	S_2	F_2
3	$\neg A_3$	$\neg S_3$	$\neg F_3$
4	A_4	S_4	F_4
5	A_5	$\neg S_5$	$\neg F_5$
6	$\neg A_6$	$\neg S_6$	$\neg F_6$
7	A_7	$\neg S_7$	F_7
8	$\neg A_8$	$\neg S_8$	$\neg F_8$
9	$\neg A_9$	$\neg S_9$	F_9
10	$\neg A_{10}$	$\neg S_{10}$	$\neg F_{10}$

Maximum Likelihood Estimation

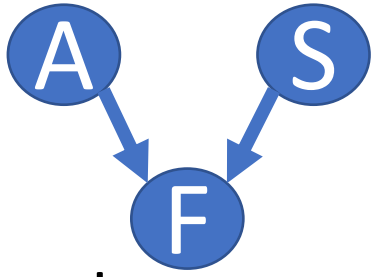


Day	A	S	F
1	$\neg A_1$	$\neg S_1$	$\neg F_1$
2	$\neg A_2$	S_2	F_2
3	$\neg A_3$	$\neg S_3$	$\neg F_3$
4	A_4	S_4	F_4
5	A_5	$\neg S_5$	$\neg F_5$
6	$\neg A_6$	$\neg S_6$	$\neg F_6$
7	A_7	$\neg S_7$	F_7
8	$\neg A_8$	$\neg S_8$	$\neg F_8$
9	$\neg A_9$	$\neg S_9$	F_9
10	$\neg A_{10}$	$\neg S_{10}$	$\neg F_{10}$

We can estimate model parameters to be the values that maximize the likelihood of the observations, subject to the constraints that

$$P(A) + P(\neg A) = 1$$
$$P(S) + P(\neg S) = 1$$
$$P(F|S,A) + P(\neg F|S,A) = 1$$

Maximum Likelihood Estimation



The maximum likelihood parameters are

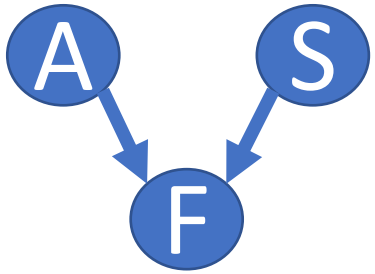
$$P(A) = \frac{\#\text{ days on which } A_i}{\#\text{ days total}}$$

$$P(S) = \frac{\#\text{ days on which } S_i}{\#\text{ days total}}$$

$$P(F|s,a) = \frac{\#\text{ days (A=}a\text{,S=}s\text{, F)}}{\#\text{ days (A=}a\text{,S=}s\text{)}}$$

Day	A	S	F
1	$\neg A_1$	$\neg S_1$	$\neg F_1$
2	$\neg A_2$	S_2	F_2
3	$\neg A_3$	$\neg S_3$	$\neg F_3$
4	A_4	S_4	F_4
5	A_5	$\neg S_5$	$\neg F_5$
6	$\neg A_6$	$\neg S_6$	$\neg F_6$
7	A_7	$\neg S_7$	F_7
8	$\neg A_8$	$\neg S_8$	$\neg F_8$
9	$\neg A_9$	$\neg S_9$	F_9
10	$\neg A_{10}$	$\neg S_{10}$	$\neg F_{10}$

Maximum Likelihood Estimation



The maximum likelihood parameters are

$$P(A) = \frac{3}{10}, \qquad P(S) = \frac{2}{10}$$

a	s	$P(F\|s,a)$
F	F	1/6
F	T	1
T	F	1/2
T	T	1

Day	A	S	F
1	$\neg A_1$	$\neg S_1$	$\neg F_1$
2	$\neg A_2$	S_2	F_2
3	$\neg A_3$	$\neg S_3$	$\neg F_3$
4	A_4	S_4	F_4
5	A_5	$\neg S_5$	$\neg F_5$
6	$\neg A_6$	$\neg S_6$	$\neg F_6$
7	A_7	$\neg S_7$	F_7
8	$\neg A_8$	$\neg S_8$	$\neg F_8$
9	$\neg A_9$	$\neg S_9$	F_9
10	$\neg A_{10}$	$\neg S_{10}$	$\neg F_{10}$

Conclusions: maximum likelihood estimation

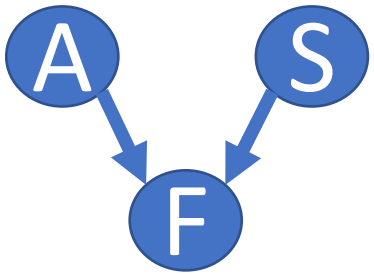
- Smart cows are far more dangerous than aliens.
- Maximum likelihood estimation is very easy to use, IF you have training data in which the values of ALL variables are observed.

- …but what if some of the variables can't be observed?
- For example: after the 6th day, the cows decide to stop responding to written surveys. Therefore, it's impossible to **observe**, on any given day, how smart the cows are. We don't know if $s_i = T$ or $s_i = F$…

Outline

- Parameter Learning
  - from Fully Observed data: Maximum Likelihood
  - from Partially Observed data: Expectation Maximization
  - from Partially Observed data: Hard EM
- Structure Learning
  - The usual method: knowledge engineering
  - An interesting recent method: causal analysis

Partially observed data



Day	A	S	F
1			
2		T	T
3			
4	T	T	T
5	T		
6			
7	T	?	T
8		?	
9		?	T
10		?	

Suppose that we have the following observations:

- We know whether A=True or False.
- We know whether F=True or False.
- After the 6^{th} day, we don't know whether S is True or False (shown as "?").

Expectation Maximization (EM): Main idea

Remember that maximum likelihood estimation counts examples:

$$P(F|S=s, A=a) = \frac{\#\text{ days } S=s, A=a, F}{\#\text{ days } S=s, A=a}$$

Expectation maximization is similar, but using "expected counts" instead of actual counts:

$$P(F|S=s, A=a) = \frac{E[\#\text{ days } S=s, A=a, F]}{E[\#\text{ days } S=s, A=a]}$$

Where E[X] means "expected value of X".

Expectation Maximization (EM): review

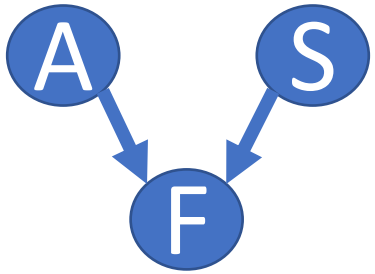
INITIALIZE: **guess** the model parameters.

ITERATE until convergence:

1. **E-Step**: $E[\#\text{ days } S = s, A = a, F = f] = \sum_{i:a_i=a, f_i=f} P(S = s|a, f)$
2. **M-Step:** $P(F = f|S = s, A = a) = \frac{E[\#\text{ days } S=s, A=a, F=f]}{E[\#\text{ days } S=s, A=a]}$

Continue the iteration, shown above, until the model parameters stop changing.

Example: Initialize

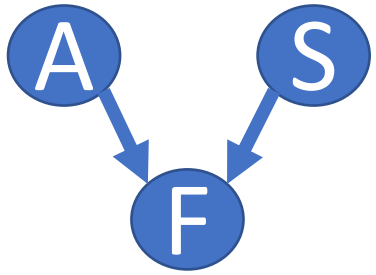


Marilyn Modigliani is a professional vaccavolatologist. She gives us these initial guesses about the possible model parameters (her guesses are probably not quite right, but they are as good a guess as anybody else's):

$$P(A) = \frac{1}{4}, \qquad P(S) = \frac{1}{4}$$

a	s	$P(F \mid s, a)$
F	F	0
F	T	1/2
T	F	1/2
T	T	1

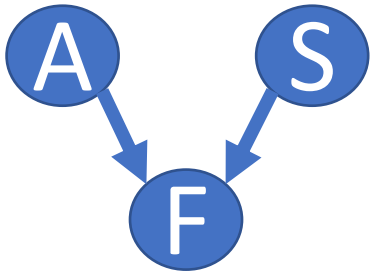
E-Step



Based on Marilyn's model, we calculate $P(S|a_i, f_i)$ for each of the missing days, as shown in the table at right.

Day	A	S	F
1			
2		T	T
3			
4	T	T	T
5	T		
6			
7	T	2/5	T
8		1/7	
9		1	T
10		1/7	

E-Step



The expected counts are

$$E[\#\text{ days } S = s, A = a, F = f] = \sum_{i:a_i=a,f_i=f} P(S = s|a, f)$$

a	f	$E[\#\ \textbf{days}\ S\|a, f]$	$E[\#\ \textbf{days}\ \neg S\|a, f]$
F	F	$0+0+0+\frac{1}{7}+\frac{1}{7}=\frac{2}{7}$	$1+1+1+\frac{6}{7}+\frac{6}{7}=\frac{33}{7}$
F	T	$1+1=2$	0+0=0
T	F	0	1
T	T	$1+\frac{2}{5}=\frac{7}{5}$	$0+\frac{3}{5}=\frac{3}{5}$

M-Step

a	f	$\boldsymbol{E}[\#\ \boldsymbol{days}\ \boldsymbol{S}\|a,f]$	$\boldsymbol{E}[\#\ \boldsymbol{days}\ \neg\boldsymbol{S}\|a,f]$
F	F	$0+0+0+\frac{1}{7}+\frac{1}{7}=\frac{2}{7}$	$1+1+1+\frac{6}{7}+\frac{6}{7}=\frac{33}{7}$
F	T	$1+1=2$	0+0=0
T	F	0	1
T	T	$1+\frac{2}{5}=\frac{7}{5}$	$0+\frac{3}{5}=\frac{3}{5}$

a	f	$E[\#\ days\ S|a,f]$	$E[\#\ days\ \neg S|a,f]$
F	F	$0+0+0+\frac{1}{7}+\frac{1}{7}=\frac{2}{7}$	$1+1+1+\frac{6}{7}+\frac{6}{7}=\frac{33}{7}$
F	T	$1+1=2$	0+0=0
T	F	0	1
T	T	$1+\frac{2}{5}=\frac{7}{5}$	$0+\frac{3}{5}=\frac{3}{5}$

M-Step

Now let's re-estimate the model parameters. For example,

$$P(F=1|S=0,A=0)=\frac{E[\#\text{ days } S=0,A=0,F=1]}{E[\#\text{ days } S=0,\ \ A=0]}$$

$$=\frac{0}{\frac{33}{7}+0}=0$$

a	f	$E[\#\ days\ S|a,f]$	$E[\#\ days\ \neg S|a,f]$
F	F	$0+0+0+\frac{1}{7}+\frac{1}{7}=\frac{2}{7}$	$1+1+1+\frac{6}{7}+\frac{6}{7}=\frac{33}{7}$
F	T	$1+1=2$	0+0=0
T	F	0	1
T	T	$1+\frac{2}{5}=\frac{7}{5}$	$0+\frac{3}{5}=\frac{3}{5}$

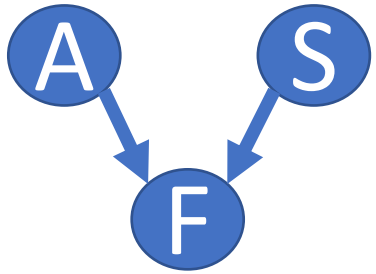
M-Step

Now let's re-estimate the model parameters. For example,

$$P(F=1|S=1,A=0)=\frac{E[\#\text{ days } S=1,A=0,F=1]}{E[\#\text{ days } S=1,\ \ A=0]}$$

$$=\frac{2}{\frac{2}{7}+2}=\frac{7}{8}$$

M-Step



The re-estimated probabilities are

$$P(A) = \frac{\#\text{ days } A}{\#\text{ days total}} = \frac{3}{10}$$

$$P(S) = \frac{E[\#\text{ days } S]}{\#\text{ days total}} = \frac{\frac{2}{7} + 2 + 0 + \frac{7}{5}}{10} = \frac{94}{350}$$

a	s	$P(F \mid S = s, A = a)$
F	F	$\frac{0}{\frac{33}{7} + 0} = 0$
F	T	$\frac{2}{\frac{2}{7} + 2} = \frac{7}{8}$
T	F	$\frac{3/5}{1 + \frac{3}{5}} = \frac{3}{8}$
T	T	$\frac{7/5}{0 + 7/5} = 1$

Expectation Maximization (EM): review

INITIALIZE: **guess** the model parameters.

ITERATE until convergence:

1. **E-Step**: $E[\#\text{ days } S = s, A = a, F = f] = \sum_{i:a_i=a, f_i=f} P(S = s|a, f)$
2. **M-Step:** $P(F = f|S = s, A = a) = \frac{E[\#\text{ days } S=s, A=a, F=f]}{E[\#\text{ days } S=s, A=a]}$

Continue the iteration, shown above, until the model parameters stop changing.

Properties of the EM algorithm

- It always converges.
- The parameters it converges to (P(A), P(S), and P(F|A,S)):
  - are guaranteed to be **at least as good as** your initial guess, but
  - They depend on your initial guess. Different initial guesses may result in different results, after the algorithm converges.
  - For example, Marilyn's initial guess was $P(F|\neg S, \neg A) = \mathbf{0}$. Notice that we ended up with the same value! According to the fully observed data we saw earlier, that might not be the best possible parameter for these data.

Structure Learning: Knowledge engineering

1. Find somebody who knows a lot about the problem you're trying to model (flying cows, or burglars in Los Angeles, or whatever).
2. Get them to tell you which variables depend on which others.
3. Draw corresponding circles and arrows.
4. Done! Proceed to parameter estimation.

Example: Bayesian diagnostic model for the symptom "no sound."

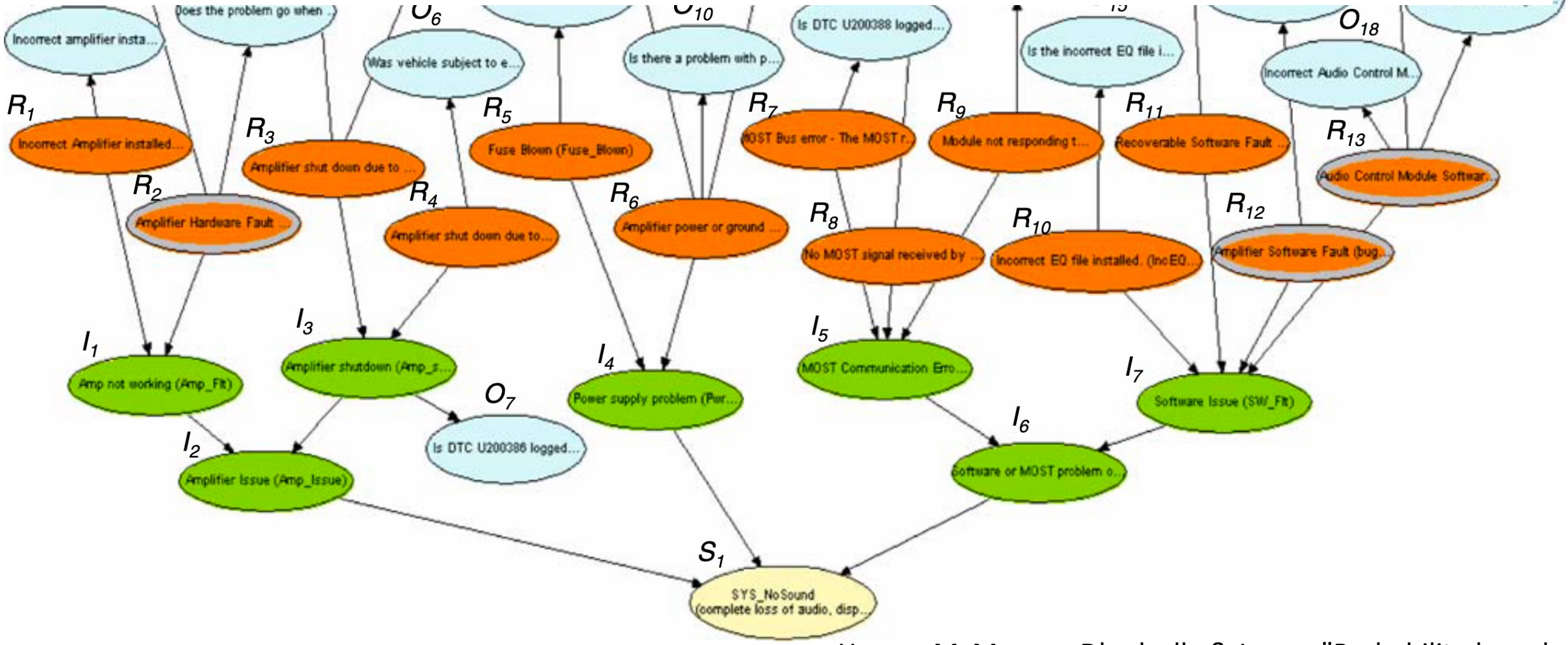
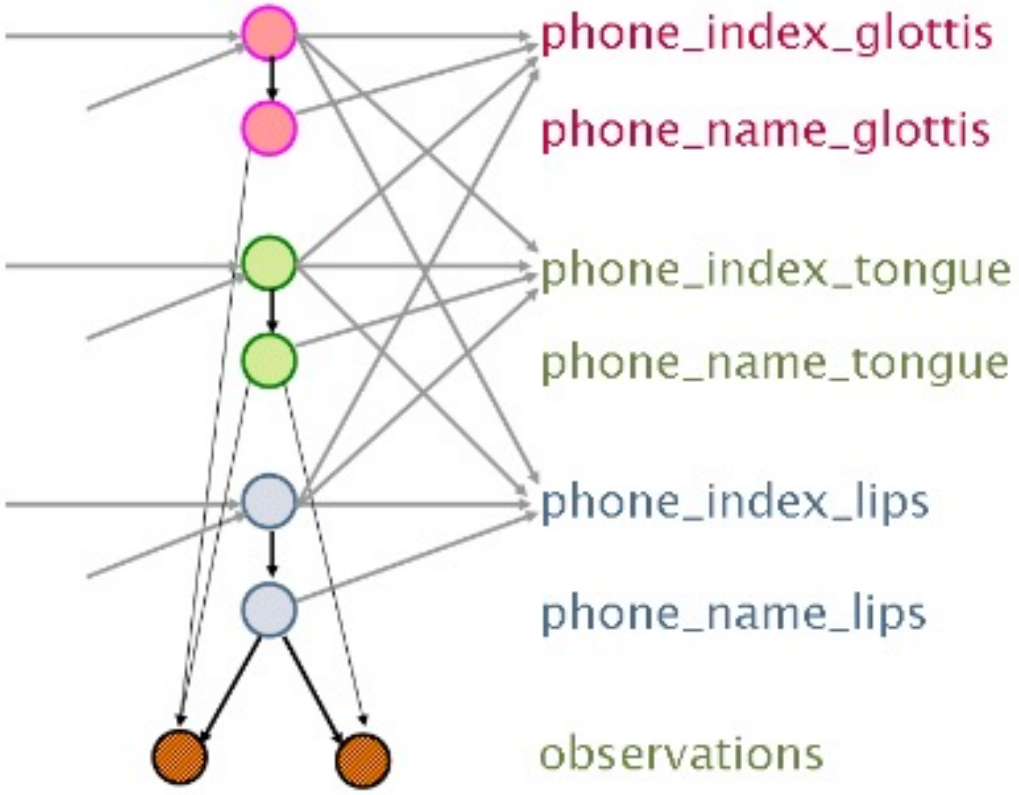


Fig. 6 Bayesian diagnostic model for the symptom "no sound"

Huang, McMurran, Dhadyalla & Jones, "Probability-based vehicle fault diagnosis: Bayesian network method," 2008.

Example Bayes Network: speech acoustics and speech appearance depend on glottis, tongue, and lip positions



Audiovisual Speech Recognition with Articulator Positions as Hidden Variables

Mark Hasegawa-Johnson, Karen Livescu, Partha Lal and Kate Saenko

International Congress on Phonetic Sciences 1719:299-302, 2007

Causal analysis

Suppose you know that you have V variables X_1, … ,X_V, but you don't know which variables depend on which others. You can learn this from the data:

For every possible ordering of the variables (there are V! possible orderings):

1. Create a blank initial network
2. For each variable in this ordering, i = 1 to V:
   a. add variable X_i to the network
   b. Check your training data. If there is any variable X_1, … ,X_{i-1} that CHANGES the probability of X_i=1, then add that variable to the set Parents(X_i) such that $P(X_i \mid Parents(X_i)) = P(X_i \mid X_1, \ldots X_{i-1})$
3. Count the number of edges in the graph with this ordering.

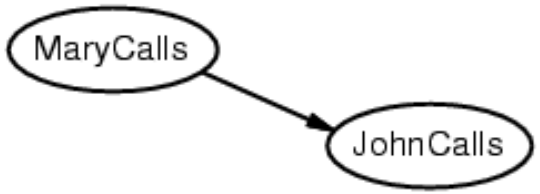
Choose the graph with the smallest number of edges.

Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E

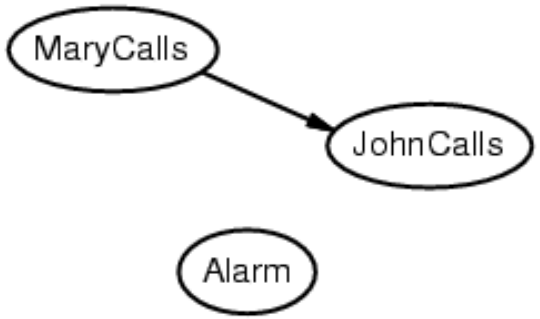
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



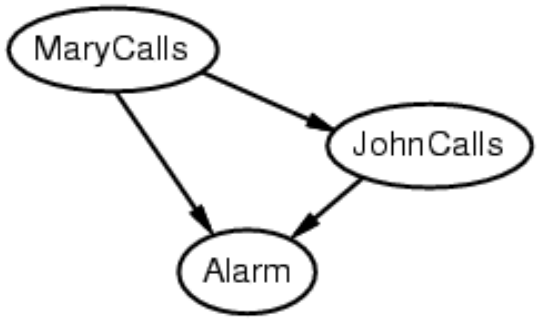
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



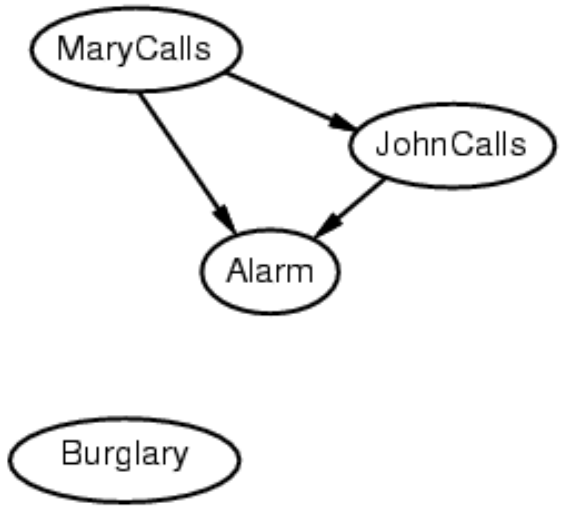
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



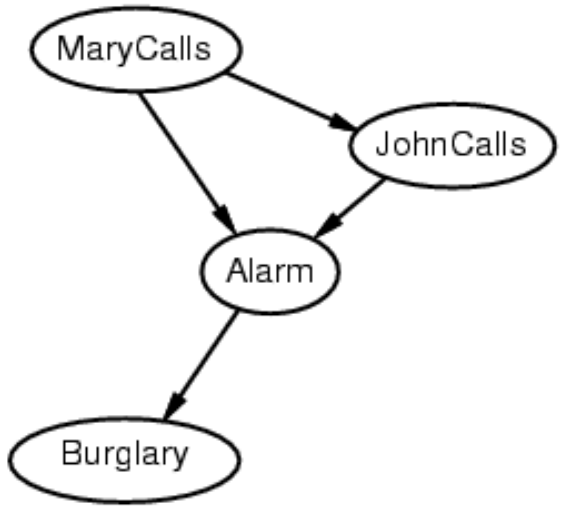
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



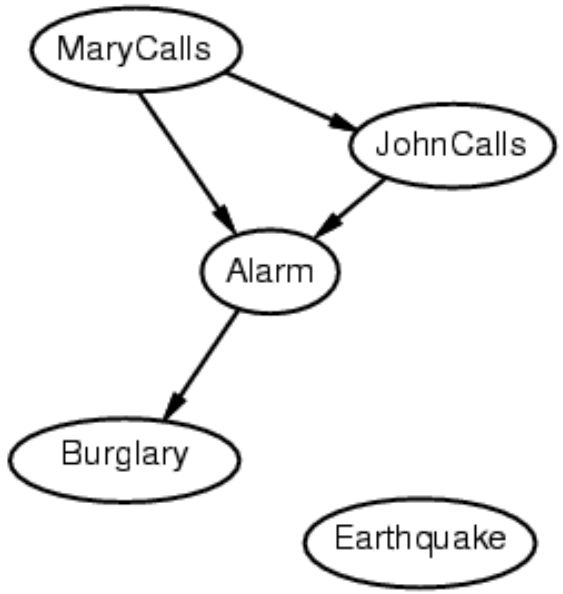
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E



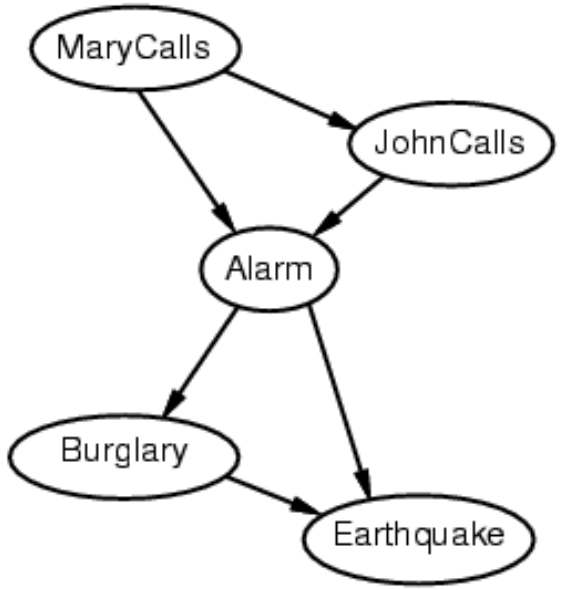
Example: The Los Angeles burglar alarm

- Suppose we choose the ordering M, J, A, B, E

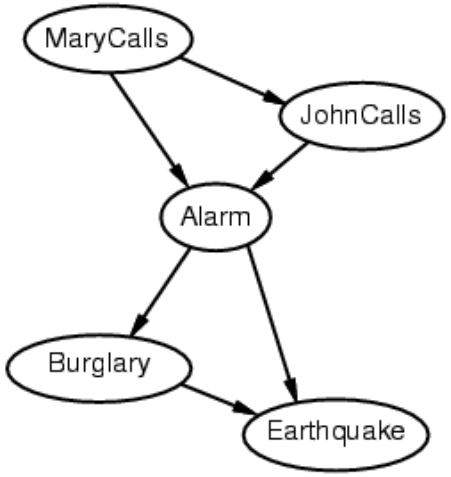


Example: The Los Angeles burglar alarm

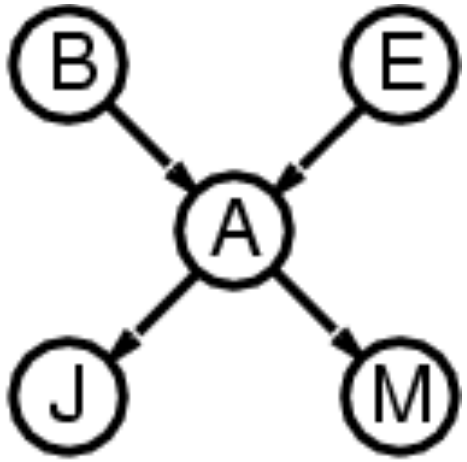
- Suppose we choose the ordering M, J, A, B, E



Example: The Los Angeles burglar alarm



versus

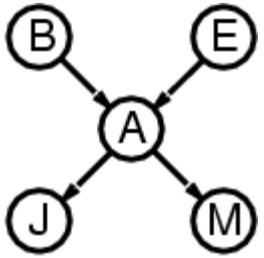


- Deciding conditional independence is hard in noncausal directions
  - The causal direction seems much more natural
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed (vs. 1+1+4+2+2=10 for the causal ordering)

Why store it in causal order? A: Saves memory

- Suppose we have a Boolean variable X_i with k Boolean parents. How many rows does its conditional probability table have?
  - 2^k rows for all the combinations of parent values
  - Each row requires one number for $P(X_i = \text{true} \mid \text{parent values})$
- If each variable has no more than k parents, how many numbers does the complete network require?
  - $O(n \cdot 2^k)$ numbers – vs. $O(2^n)$ for the full joint distribution
- How many nodes for the burglary network?

  1 + 1 + 4 + 2 + 2 = 10 numbers (vs. 2^5-1 = 31)



Parameter and Structure Learning for Bayesian Networks

- Maximum Likelihood (ML):

$$P(F|S=s, A=a) = \frac{\#\text{ days (A=}a\text{, S=}s\text{, F)}}{\#\text{ days (A=}a\text{, S=}s\text{)}}$$

- Expectation Maximization (EM):

$$P(F|S=s, A=a) = \frac{E[\#\text{ days } A=a, S=s, \ F]}{E[\#\text{ days } A=a, \ S=s]}$$

- Knowledge Engineering: ask an expert.
- Causal Analysis: construct all possible graphs, keep the one with the fewest edges.