

# Bayes Nets for representing and reasoning about uncertainty

**Andrew W. Moore**

**Associate Professor**

**School of Computer Science**

**Carnegie Mellon University**

[www.cs.cmu.edu/~awm](http://www.cs.cmu.edu/~awm)

[awm@cs.cmu.edu](mailto:awm@cs.cmu.edu)

412-268-7599

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: <http://www.cs.cmu.edu/~awm/tutorials> . Comments and corrections gratefully received.

# What we'll discuss

- Recall the numerous and dramatic benefits of Joint Distributions for describing uncertain worlds
- Reel with terror at the problem with using Joint Distributions
- Discover how Bayes Net methodology allows us to build Joint Distributions in manageable chunks
- Discover there's still a lurking problem...
- ...Start to solve that problem

# Why this matters

- In Andrew's opinion, the most important technology in the Machine Learning / AI field to have emerged in the last 10 years.
- A clean, clear, manageable language and methodology for expressing what you're certain and uncertain about
- Already, many practical applications in medicine, factories, helpdesks:
  - P(this problem | these symptoms)
  - anomalousness of this observation
  - choosing next diagnostic test | these observations

# Why this matters

- In Andrew's opinion, the most important technology in the Machine Learning / AI field emerged in the last 10 years.

Active Data  
Collection

Inference

Anomaly  
Detection

- Already, many practical applications in medicine, factories, helpdesks:

$P(\text{this problem} \mid \text{these symptoms})$

anomalousness of this observation

choosing next diagnostic test  $\mid$  these observations

# Ways to deal with Uncertainty

- Three-valued logic: True / False / Maybe
- Fuzzy logic (truth values between 0 and 1)
- Non-monotonic reasoning (especially focused on Penguin informatics)
- Dempster-Shafer theory (and an extension known as quasi-Bayesian theory)
- Possibilistic Logic
- Probability

# Discrete Random Variables

- A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
- Examples
  - A = The US president in 2023 will be male
  - A = You wake up tomorrow with a headache
  - A = You have Ebola

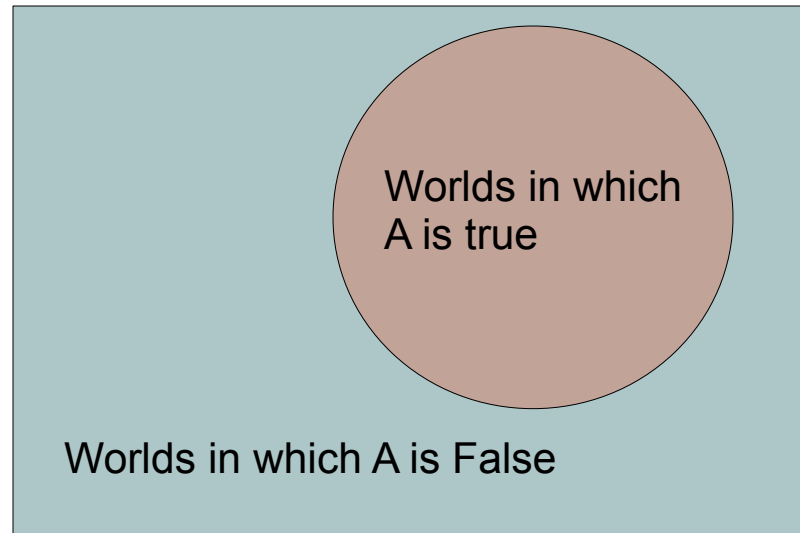
# Probabilities

- We write  $P(A)$  as “the fraction of possible worlds in which  $A$  is true”
- We could at this point spend 2 hours on the philosophy of this.
- But we won't.

# Visualizing A

Event space of  
all possible  
worlds

Its area is 1

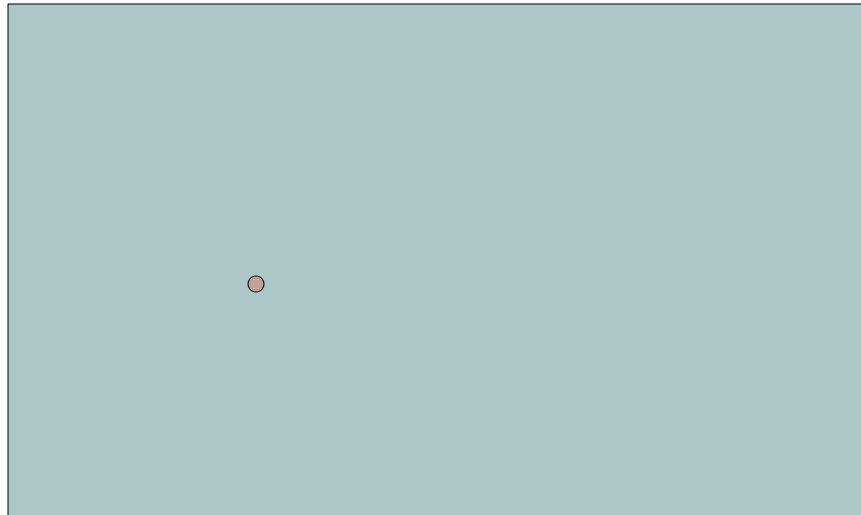


$P(A)$  = Area of  
reddish oval



# Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$



The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

# Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

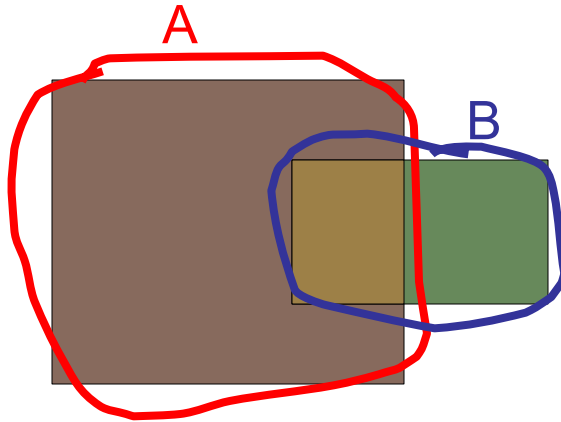


The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true

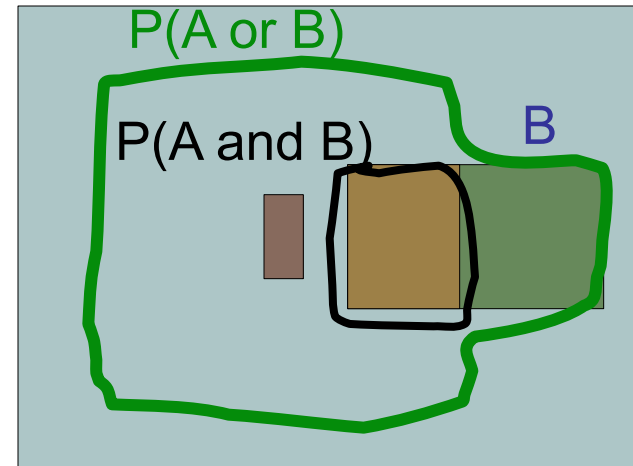
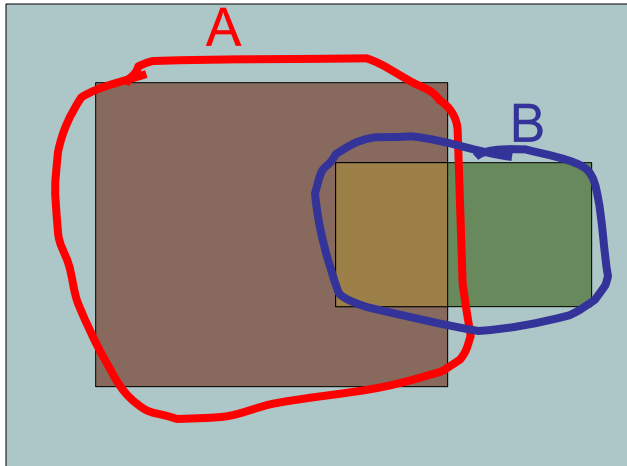
# Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$



# Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$



Simple addition and subtraction

# These Axioms are Not to be Trifled With

- There have been attempts to do different methodologies for uncertainty
  - Fuzzy Logic
  - Three-valued logic
  - Dempster-Shafer
  - Non-monotonic reasoning
- But the axioms of probability are the only system with this property:

If you gamble using them you can't be unfairly exploited by an opponent using some other system [di Finetti 1931]

# Theorems from the Axioms

- $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$
- $P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$

From these we can prove:

$$P(\text{not } A) = P(\sim A) = 1 - P(A)$$

- How?

# Side Note

- I am inflicting these proofs on you for two reasons:
  1. These kind of manipulations will need to be second nature to you if you use probabilistic analytics in depth
  2. Suffering is good for you

# Another important theorem

- $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$
- $P(\text{A or B}) = P(\text{A}) + P(\text{B}) - P(\text{A and B})$

From these we can prove:

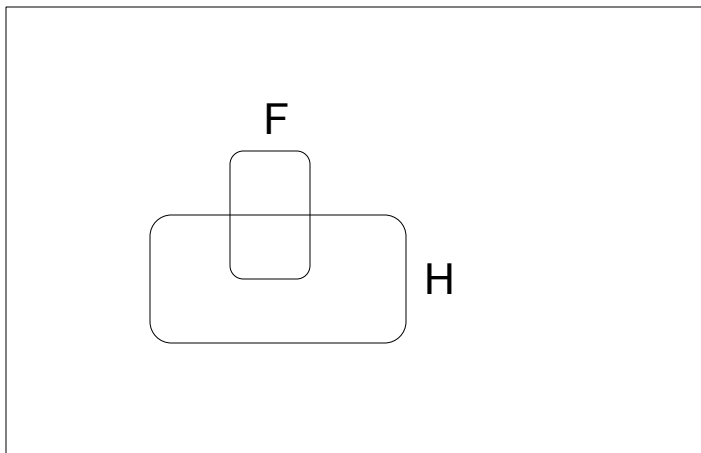
$$P(A) = P(A \wedge B) + P(A \wedge \sim B)$$

- How?



# Conditional Probability

- $P(A|B)$  = Fraction of worlds in which B is true that also have A true



H = “Have a headache”

F = “Coming down with Flu”

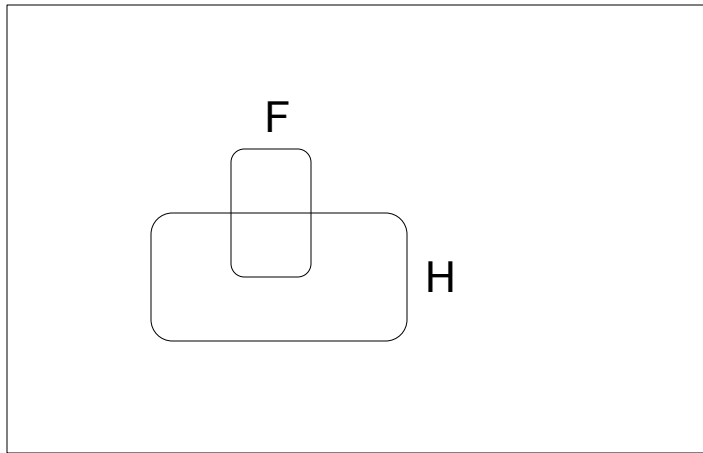
$$P(H) = 1/10$$

$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

“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.”

# Conditional Probability



H = "Have a headache"  
F = "Coming down with Flu"

$$P(H) = 1/10$$
$$P(F) = 1/40$$
$$P(H|F) = 1/2$$

$P(H|F)$  = Fraction of flu-inflicted worlds in which you have a headache

$$= \frac{\text{\#worlds with flu and headache}}{\text{\#worlds with flu}}$$

$$= \frac{\text{Area of "H and F" region}}{\text{Area of "F" region}}$$

$$= \frac{P(H \wedge F)}{P(F)}$$

# Definition of Conditional Probability

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

## Corollary: The Chain Rule

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

# Bayes Rule

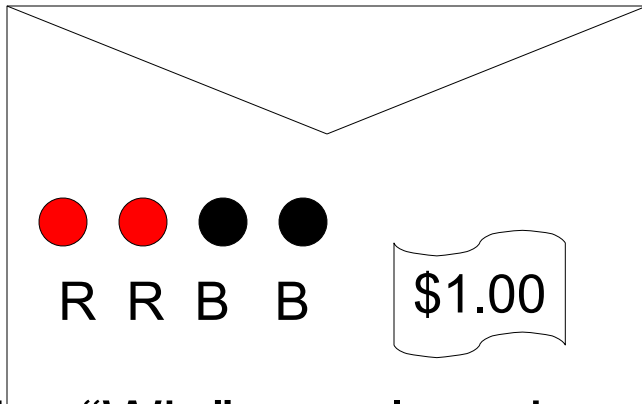
$$P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{P(A|B) P(B)}{P(A)}$$

This is Bayes Rule

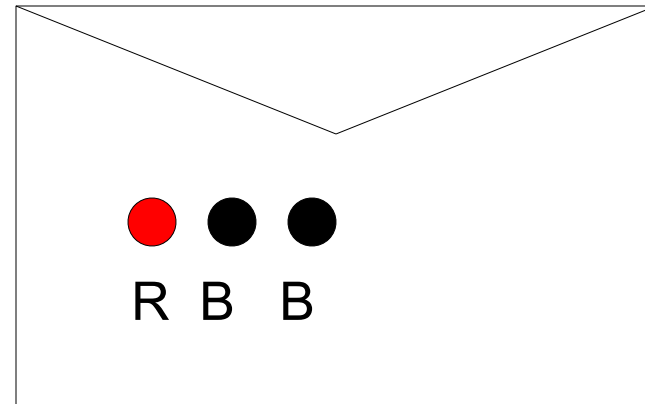
**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**



# Using Bayes Rule to Gamble



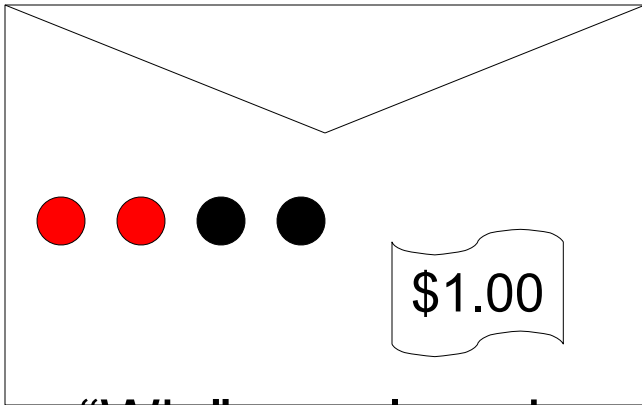
The “Win” envelope has  
a dollar and four  
beads in it



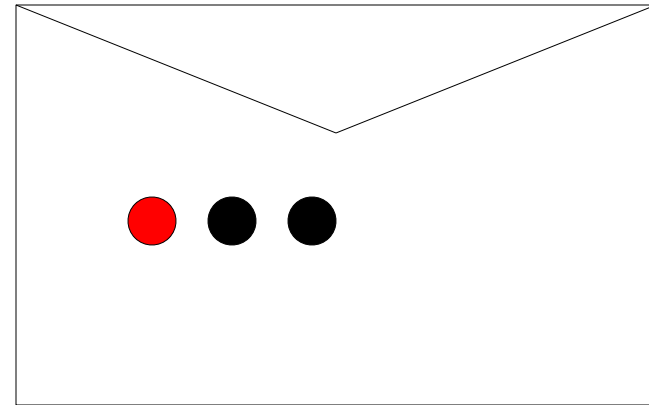
The “Lose” envelope  
has three beads and  
no money

Trivial question: someone draws an envelope at random and offers to sell it to you. How much should you pay?

# Using Bayes Rule to Gamble



The “Win” envelope has  
a dollar and four  
beads in it



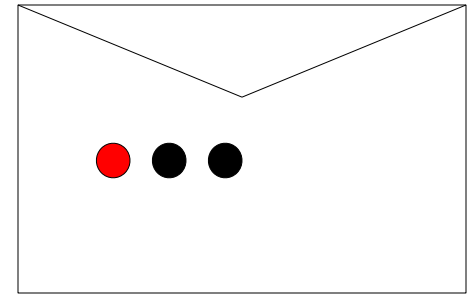
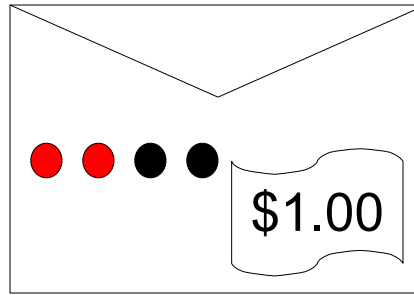
The “Lose” envelope  
has three beads and  
no money

Interesting question: before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?

Suppose it's red: How much should you pay?

# Calculation...



# Multivalued Random Variables

- Suppose  $A$  can take on more than 2 values
- $A$  is a *random variable with arity  $k$*  if it can take on exactly one value out of  $\{v_1, v_2, \dots, v_k\}$
- Thus...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$



# An easy fact about Multivalued Random Variables:

- Using the axioms of probability...  
 $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$   
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- And assuming that A obeys...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

- It's easy to prove that

$$P(A = v_1 \vee A = v_2 \vee A = v_i) = \sum_{j=1}^i P(A = v_j)$$

# An easy fact about Multivalued Random Variables:

- Using the axioms of probability...  
 $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$   
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- And assuming that A obeys...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

- It's easy to prove that

$$P(A = v_1 \vee A = v_2 \vee A = v_i) = \sum_{j=1}^i P(A = v_j)$$

- And thus we can prove

$$\sum_{j=1}^k P(A = v_j) = 1$$

## Another fact about Multivalued Random Variables:

- Using the axioms of probability...  
 $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$   
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- And assuming that A obeys...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee A = v_k) = 1$$

- It's easy to prove that

$$P(B \wedge [A = v_1 \vee A = v_2 \vee A = v_i]) = \sum_{j=1}^i P(B \wedge A = v_j)$$

## Another fact about Multivalued Random Variables:

- Using the axioms of probability...  
 $0 \leq P(A) \leq 1$ ,  $P(\text{True}) = 1$ ,  $P(\text{False}) = 0$   
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- And assuming that A obeys...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$
$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

- It's easy to prove that

$$P(B \wedge [A = v_1 \vee A = v_2 \vee \dots \vee A = v_i]) = \sum_{j=1}^i P(B \wedge A = v_j)$$

- And thus we can prove

$$P(B) = \sum_{j=1}^k P(B \wedge A = v_j)$$

# More General Forms of Bayes Rule

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

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

# More General Forms of Bayes Rule

$$P(A = v_i | B) = \frac{P(B | A = v_i)P(A = v_i)}{\sum_{k=1}^{n_A} P(B | A = v_k)P(A = v_k)}$$

# Useful Easy-to-prove facts

$$P(A \mid B) + P(\neg A \mid B) = 1$$

$$\sum_{k=1}^{n_A} P(A = v_k \mid B) = 1$$

# The Joint Distribution

*Example: Boolean  
variables  $A, B, C$*

Recipe for making a joint distribution  
of  $M$  variables:



# The Joint Distribution

*Example: Boolean variables A, B, C*

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have  $2^M$  rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

# The Joint Distribution

*Example: Boolean variables A, B, C*

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have  $2^M$  rows).
2. For each combination of values, say how probable it is.

<b>A</b>	<b>B</b>	<b>C</b>	<b>Prob</b>
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

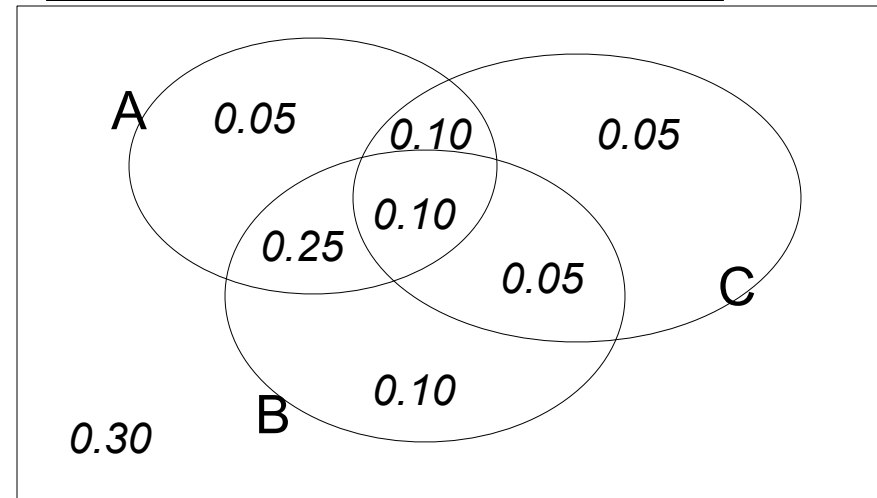
# The Joint Distribution

*Example: Boolean variables A, B, C*









Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have  $2^M$  rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10











# Using the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

Once you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$









# Using the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

$$P(\text{Poor Male}) = 0.4654$$

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

# Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122 
		rich	0.0245895 
	v1:40.5+	poor	0.0421768 
		rich	0.0116293 
Male	v0:40.5-	poor	0.331313 
		rich	0.0971295 
	v1:40.5+	poor	0.134106 
		rich	0.105933 

$$P(\text{Poor}) = 0.7604$$

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

# Inference with the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	<div></div>
		rich	0.0245895	<div></div>
	v1:40.5+	poor	0.0421768	<div></div>
		rich	0.0116293	<div></div>
Male	v0:40.5-	poor	0.331313	<div></div>
		rich	0.0971295	<div></div>
	v1:40.5+	poor	0.134106	<div></div>
		rich	0.105933	<div></div>

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

# Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

$$P(\text{Male} | \text{Poor}) = 0.4654 / 0.7604 = 0.612$$



# Joint distributions

- Good news

Once you have a joint distribution, you can ask important questions about stuff that involves a lot of uncertainty

- Bad news

Impossible to create for more than about ten attributes because there are so many numbers needed when you build the damn thing.

# Using fewer numbers

Suppose there are two events:

- M: Manuela teaches the class (otherwise it's Andrew)
- S: It is sunny

The joint p.d.f. for these events contain four entries.

If we want to build the joint p.d.f. we'll have to invent those four numbers. OR WILL WE??

- We don't have to specify with bottom level conjunctive events such as  $P(\sim M \wedge S)$  IF...
- ...instead it may sometimes be more convenient for us to specify things like:  $P(M)$ ,  $P(S)$ .

But just  $P(M)$  and  $P(S)$  don't derive the joint distribution. So you can't answer all questions.

# Using fewer numbers

Suppose there are two events:

- M: Manuela teaches the class (otherwise it's Andrew)
- S: It is sunny

The joint p.d.f. for these events contain four entries.

If we want to build the joint p.d.f. we'll have to invent those four numbers. OR WILL WE??

- We don't have to specify with bottom level conjunctive events such as  $P(\sim M \wedge S)$  IF...
- ...instead it may sometimes be more convenient for us to specify things like:  $P(M)$ ,  $P(S)$ .

But just specifying  $P(M)$  and  $P(S)$  don't derive the joint distribution. So you can't make a joint distribution.

What extra assumption can you make?

# Independence

“The sunshine levels do not depend on and do not influence who is teaching.”

This can be specified very simply:

$$P(S \mid M) = P(S)$$

This is a powerful statement!

It required extra domain knowledge. A different kind of knowledge than numerical probabilities. It needed an understanding of causation.

# Independence

From  $P(S \mid M) = P(S)$ , the rules of probability imply: (*can you prove these?*)

- $P(\sim S \mid M) = P(\sim S)$
- $P(M \mid S) = P(M)$
- $P(M \wedge S) = P(M) P(S)$
- $P(\sim M \wedge S) = P(\sim M) P(S)$ ,  $P(M \wedge \sim S) = P(M)P(\sim S)$ ,  
 $P(\sim M \wedge \sim S) = P(\sim M)P(\sim S)$

# Independence

From  $P(S \mid M) = P(S)$ , the rules of probability imply: (*can you prove these?*)

And in general:

- $P(\sim S)$

$$P(M=u \wedge S=v) = P(M=u) P(S=v)$$

- $P(M)$

for each of the four combinations of

$u = \text{True/False}$

- $P(M)$

$v = \text{True/False}$

- $P(\sim M \wedge S) = P(\sim M) P(S)$ ,  $P(M \wedge \sim S) = P(M) P(\sim S)$ ,  
 $P(\sim M \wedge \sim S) = P(\sim M) P(\sim S)$

# Independence

We've stated:

$$P(M) = 0.6$$

$$P(S) = 0.3$$

$$P(S \mid M) = P(S)$$

From these statements, we can derive the full joint pdf.

M	S	Prob
T	T	
T	F	
F	T	
F	F	

And since we now have the joint pdf, we can make any queries we like.

# A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.



# A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

Let's begin with writing down knowledge we're happy about:

$$P(S \mid M) = P(S), \quad P(S) = 0.3, \quad P(M) = 0.6$$

Lateness is not independent of the weather and is not independent of the lecturer.

# A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

Let's begin with writing down knowledge we're happy about:

$$P(S \mid M) = P(S), \quad P(S) = 0.3, \quad P(M) = 0.6$$

Lateness is not independent of the weather and is not independent of the lecturer.

We already know the Joint of S and M, so all we need now is

$$P(L \mid S=u, M=v)$$

in the 4 cases of  $u/v = \text{True/False}$ .

# A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

$P(S \mid M) = P(S)$	$P(L \mid M \wedge S) = 0.05$
$P(S) = 0.3$	$P(L \mid M \wedge \sim S) = 0.1$
$P(M) = 0.6$	$P(L \mid \sim M \wedge S) = 0.1$
	$P(L \mid \sim M \wedge \sim S) = 0.2$

Now we can derive a full joint p.d.f. with a “mere” six numbers instead of seven\*

*\*Savings are larger for larger numbers of variables.*

# A more interesting case

- M : Manuela teaches the class
- S : It is sunny
- L : The lecturer arrives slightly late.

Assume both lecturers are sometimes delayed by bad weather. Andrew is more likely to arrive late than Manuela.

$$P(S \mid M) = P(S)$$

$$P(S) = 0.3$$

$$P(M) = 0.6$$

$$P(L \mid M \wedge S) = 0.05$$

$$P(L \mid M \wedge \sim S) = 0.1$$

$$P(L \mid \sim M \wedge S) = 0.1$$

$$P(L \mid \sim M \wedge \sim S) = 0.2$$

Question: Express

$$P(L=x \wedge M=y \wedge S=z)$$

in terms that only need the above expressions, where  $x, y$  and  $z$  may each be True or False.

# A bit of notation

$$P(S \mid M) = P(S)$$

$$P(S) = 0.3$$

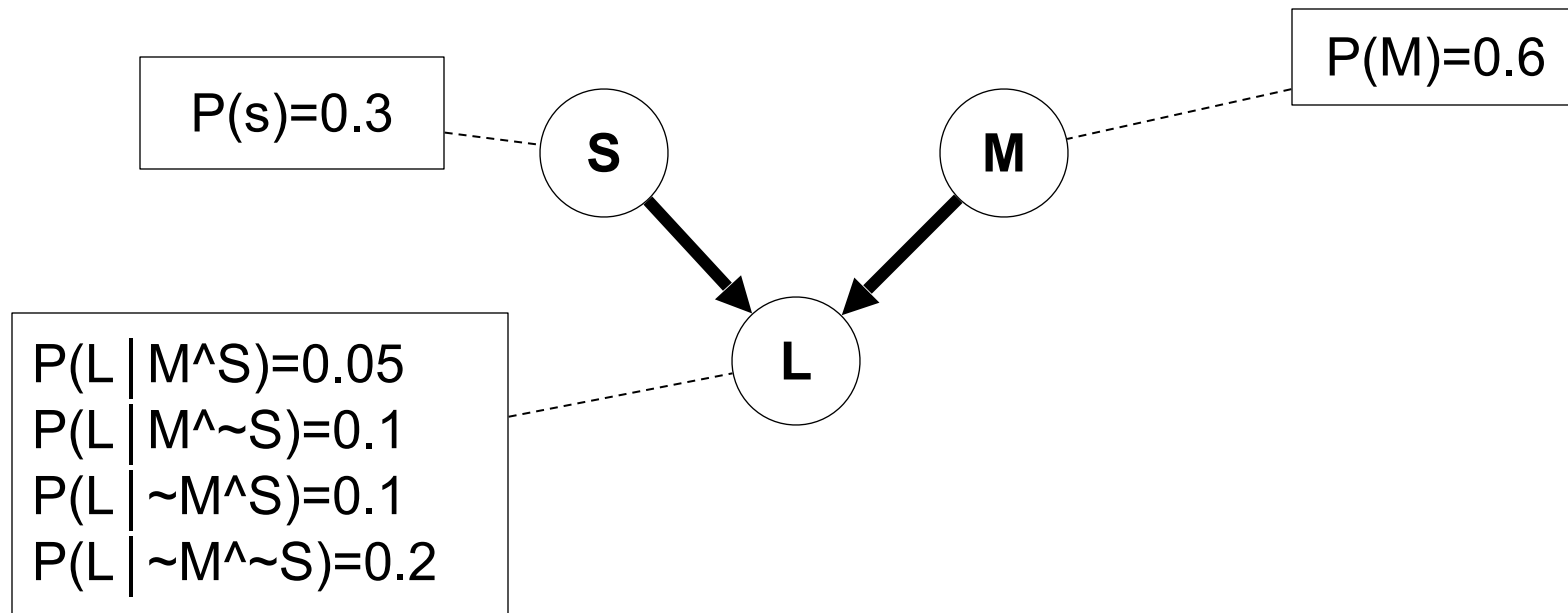
$$P(M) = 0.6$$

$$P(L \mid M \wedge S) = 0.05$$

$$P(L \mid M \wedge \sim S) = 0.1$$

$$P(L \mid \sim M \wedge S) = 0.1$$

$$P(L \mid \sim M \wedge \sim S) = 0.2$$



# A bit of notation

$$P(S \mid M) = P(S)$$

$$P(S) = 0.3$$

$$P(M) = 0.6$$

$$P(L \mid M \wedge S) = 0.05$$

$$P(L \mid M \wedge \sim S)$$

$$P(L \mid \sim M \wedge S)$$

$$P(L \mid \sim M \wedge \sim S)$$

Read the absence of an arrow between S and M to mean “it would not help me predict M if I knew the value of S”

This kind of stuff will be thoroughly formalized later

$$P(s)=0.3$$

S

M

$$P(M)=0.6$$

L

$$P(L \mid M \wedge S) = 0.05$$

$$P(L \mid M \wedge \sim S) = 0.1$$

$$P(L \mid \sim M \wedge S) = 0.1$$

$$P(L \mid \sim M \wedge \sim S) = 0.2$$

Read the two arrows into L to mean that if I want to know the value of L it may help me to know M and to know S.

# An even cuter trick

Suppose we have these three events:

- M : Lecture taught by Manuela
- L : Lecturer arrives late
- R : Lecture concerns robots

Suppose:

- Andrew has a higher chance of being late than Manuela.
- Andrew has a higher chance of giving robotics lectures.

What kind of independence can we find?

How about:

- $P(L \mid M) = P(L) ?$
- $P(R \mid M) = P(R) ?$
- $P(L \mid R) = P(L) ?$

# Conditional independence

Once you know who the lecturer is, then whether they arrive late doesn't affect whether the lecture concerns robots.

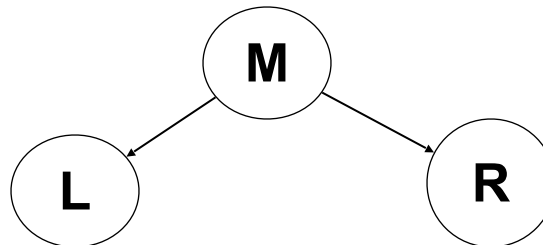
$$P(R \mid M, L) = P(R \mid M) \text{ and}$$

$$P(R \mid \sim M, L) = P(R \mid \sim M)$$

We express this in the following way:

“R and L are conditionally independent given M”

..which is also  
notated by the  
following diagram.



Given knowledge of M,  
knowing anything else in  
the diagram won't help  
us with L, etc.



# Conditional Independence formalized

R and L are conditionally independent given M if  
for all  $x, y, z$  in  $\{T, F\}$ :

$$P(R=x \mid M=y \wedge L=z) = P(R=x \mid M=y)$$

More generally:

Let  $S_1$  and  $S_2$  and  $S_3$  be sets of variables.

Set-of-variables  $S_1$  and set-of-variables  $S_2$  are  
conditionally independent given  $S_3$  if for all  
assignments of values to the variables in the sets,

$$P(S_1\text{'s assignments} \mid S_2\text{'s assignments} \ \& \ S_3\text{'s assignments}) = \\ P(S_1\text{'s assignments} \mid S_3\text{'s assignments})$$

## Example:

R and L are  
for all  $x, y, z$

$P(R=$

More gener

“Shoe-size is conditionally independent of Glove-size given  
height weight and age”

means

forall  $s, g, h, w, a$

$$P(\text{ShoeSize}=s | \text{Height}=h, \text{Weight}=w, \text{Age}=a) \\ =$$

$$P(\text{ShoeSize}=s | \text{Height}=h, \text{Weight}=w, \text{Age}=a, \text{GloveSize}=g)$$

Let  $S_1$  and  $S_2$  and  $S_3$  be sets of va

Set-of-variables  $S_1$  and set-of-variables  $S_2$  are  
conditionally independent given  $S_3$  if for all  
assignments of values to the variables in the sets,

$$P(S_1\text{'s assignments} \mid S_2\text{'s assignments} \ \& \ S_3\text{'s assignments}) = \\ P(S_1\text{'s assignments} \mid S_3\text{'s assignments})$$

## Example:

R and L are  
for all  $x, y, z$

$P(R=$

More gener

Let  $S_1$  and  $S_2$  and  $S_3$  be sets of va

“Shoe-size is conditionally independent of Glove-size given  
height weight and age”

does not mean

forall  $s, g, h$

$P(\text{ShoeSize}=s | \text{Height}=h)$

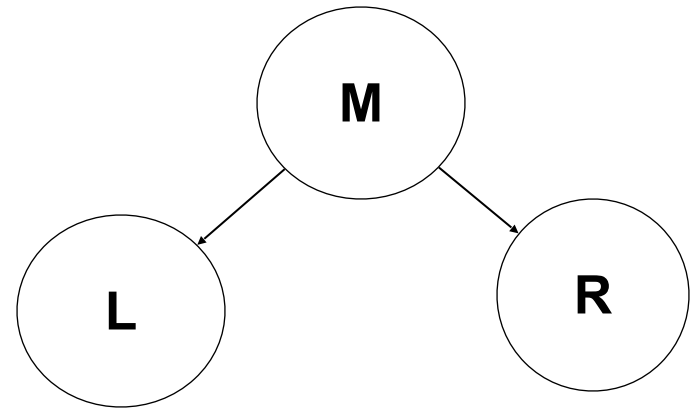
=

$P(\text{ShoeSize}=s | \text{Height}=h, \text{GloveSize}=g)$

Set-of-variables  $S_1$  and set-of-variables  $S_2$  are  
conditionally independent given  $S_3$  if for all  
assignments of values to the variables in the sets,

$P(S_1\text{'s assignments} \mid S_2\text{'s assignments} \ \& \ S_3\text{'s assignments}) =$   
 $P(S_1\text{'s assignments} \mid S_3\text{'s assignments})$

# Conditional independence



We can write down  $P(M)$ . And then, since we know L is only directly influenced by M, we can write down the values of  $P(L | M)$  and  $P(L | \sim M)$  and know we've fully specified L's behavior. Ditto for R.

$$P(M) = 0.6$$

$$P(L | M) = 0.085$$

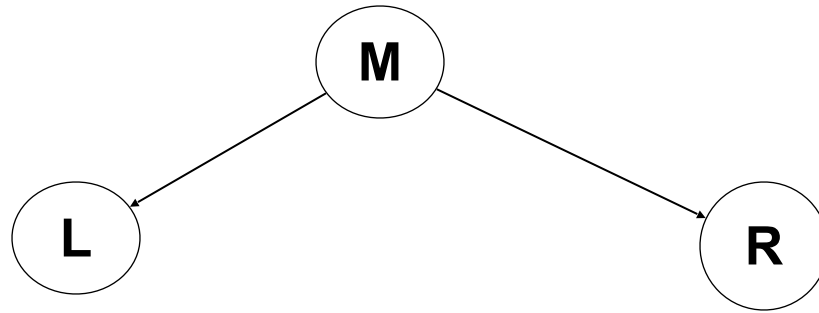
$$P(L | \sim M) = 0.17$$

$$P(R | M) = 0.3$$

$$P(R | \sim M) = 0.6$$

'R and L conditionally independent given M'

# Conditional independence



$$P(M) = 0.6$$

$$P(L \mid M) = 0.085$$

$$P(L \mid \sim M) = 0.17$$

$$P(R \mid M) = 0.3$$

$$P(R \mid \sim M) = 0.6$$

Conditional Independence:

$$P(R \mid M, L) = P(R \mid M),$$

$$P(R \mid \sim M, L) = P(R \mid \sim M)$$

Again, we can obtain any member of the Joint prob dist that we desire:

$$P(L=x \wedge R=y \wedge M=z) =$$

# Assume five variables

T: The lecture started by 10:35

L: The lecturer arrives late

R: The lecture concerns robots

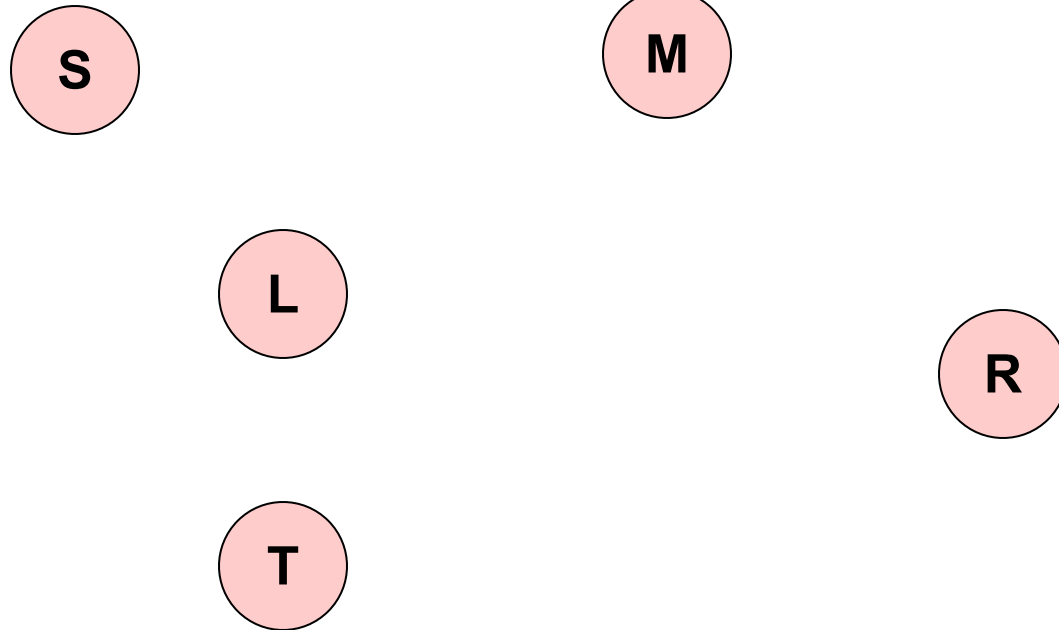
M: The lecturer is Manuela

S: It is sunny

- T only directly influenced by L (i.e. T is conditionally independent of R,M,S given L)
- L only directly influenced by M and S (i.e. L is conditionally independent of R given M & S)
- R only directly influenced by M (i.e. R is conditionally independent of L,S, given M)
- M and S are independent

# Making a Bayes net

T: The lecture started by 10:35  
L: The lecturer arrives late  
R: The lecture concerns robots  
M: The lecturer is Manuela  
S: It is sunny

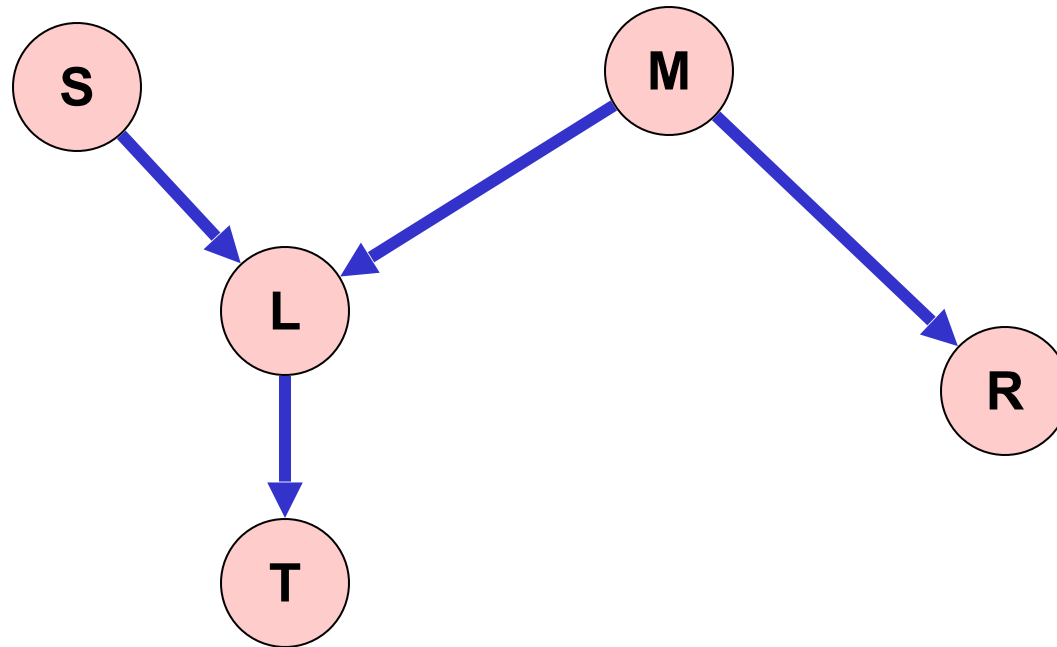


Step One: add variables.

- Just choose the variables you'd like to be included in the net.

# Making a Bayes net

T: The lecture started by 10:35  
L: The lecturer arrives late  
R: The lecture concerns robots  
M: The lecturer is Manuela  
S: It is sunny



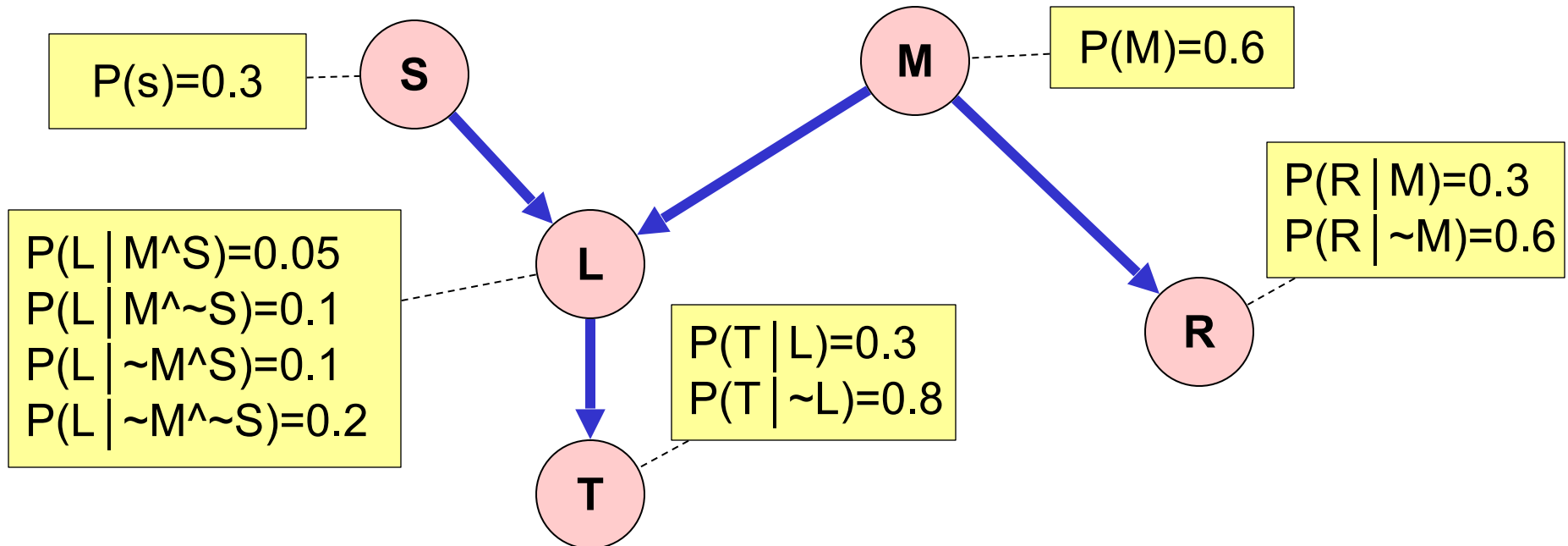
Step Two: add links.

- The link structure must be acyclic.
- If node  $X$  is given parents  $Q_1, Q_2, \dots, Q_n$  you are promising that any variable that's a non-descendent of  $X$  is conditionally independent of  $X$  given  $\{Q_1, Q_2, \dots, Q_n\}$



# Making a Bayes net

T: The lecture started by 10:35  
L: The lecturer arrives late  
R: The lecture concerns robots  
M: The lecturer is Manuela  
S: It is sunny

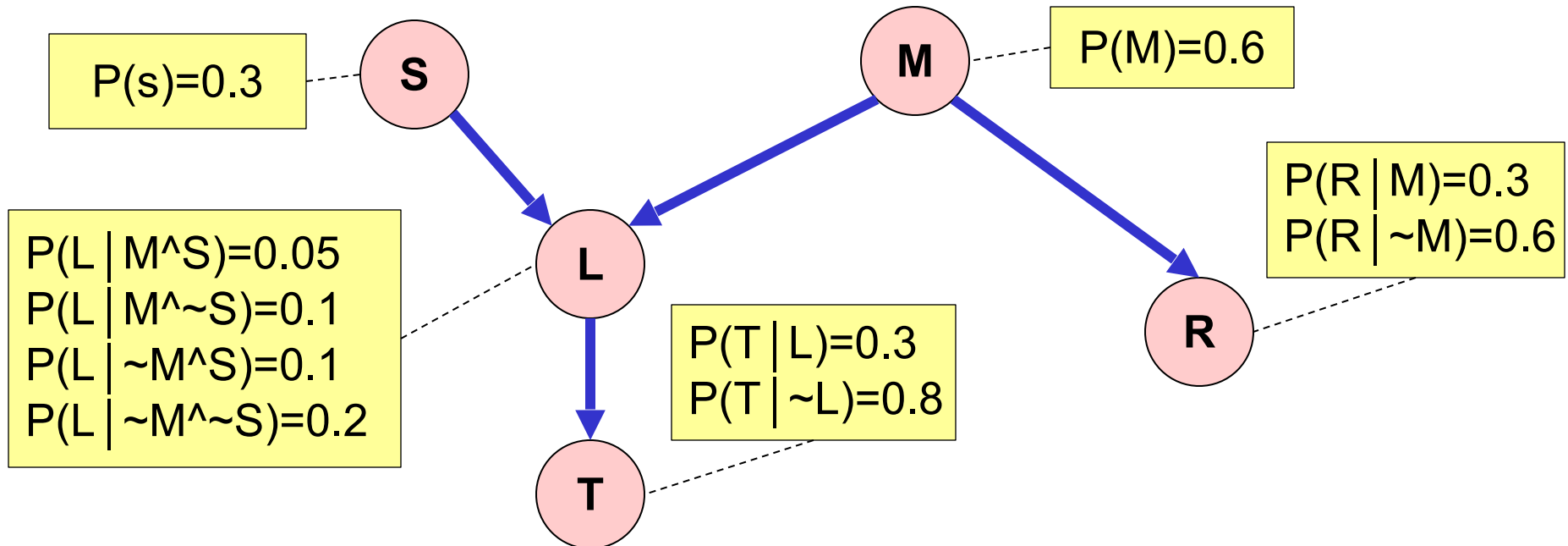


Step Three: add a probability table for each node.

- The table for node X must list  $P(X | \text{Parent Values})$  for each possible combination of parent values

# Making a Bayes net

T: The lecture started by 10:35  
L: The lecturer arrives late  
R: The lecture concerns robots  
M: The lecturer is Manuela  
S: It is sunny



- Two unconnected variables may still be correlated
- Each node is conditionally independent of all non-descendants in the tree, given its parents.
- You can deduce many other conditional independence relations from a Bayes net. See the next lecture.

# Bayes Nets Formalized

A Bayes net (also called a belief network) is an augmented directed acyclic graph, represented by the pair  $V, E$  where:

- $V$  is a set of vertices.
- $E$  is a set of directed edges joining vertices. No loops of any length are allowed.

Each vertex in  $V$  contains the following information:

- The name of a random variable
- A probability distribution table indicating how the probability of this variable's values depends on all possible combinations of parental values.

# Building a Bayes Net

1. Choose a set of relevant variables.
2. Choose an ordering for them
3. Assume they're called  $X_1 \dots X_m$  (where  $X_1$  is the first in the ordering,  $X_2$  is the second, etc)
4. For  $i = 1$  to  $m$ :
  1. Add the  $X_i$  node to the network
  2. Set  $Parents(X_i)$  to be a minimal subset of  $\{X_1 \dots X_{i-1}\}$  such that we have conditional independence of  $X_i$  and all other members of  $\{X_1 \dots X_{i-1}\}$  given  $Parents(X_i)$
  3. Define the probability table of  $P(X_i = k \mid \text{Assignments of } Parents(X_i))$ .

# Example Bayes Net Building

Suppose we're building a nuclear power station.

There are the following random variables:

GRL : Gauge Reads Low.

CTL : Core temperature is low.

FG : Gauge is faulty.

FA : Alarm is faulty

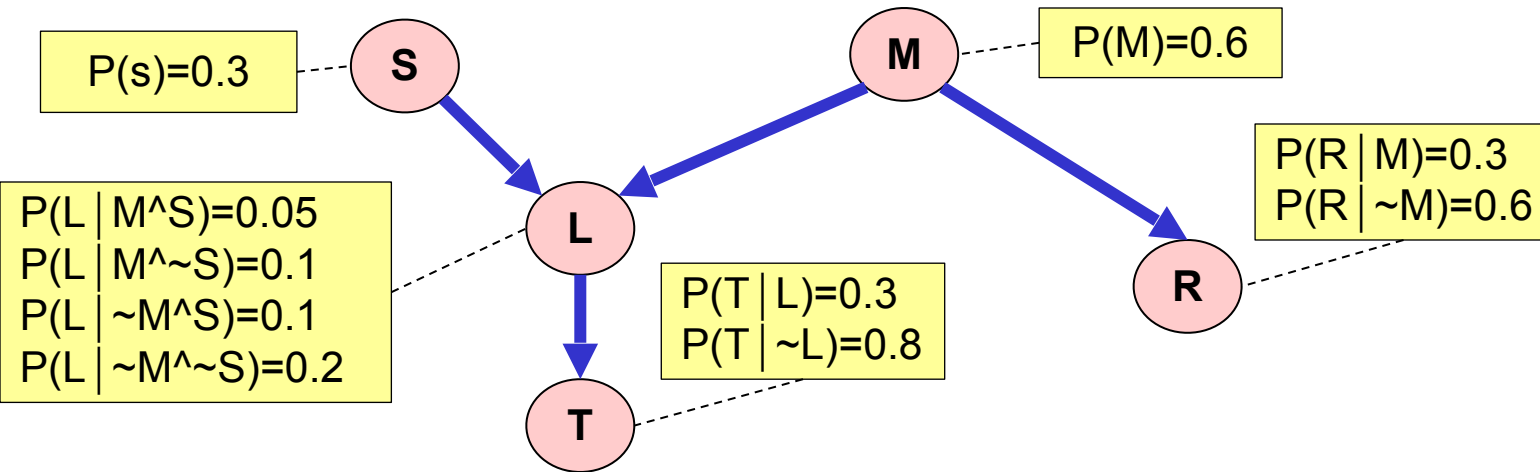
AS : Alarm sounds

- If alarm working properly, the alarm is meant to sound if the gauge stops reading a low temp.
- If gauge working properly, the gauge is meant to read the temp of the core.

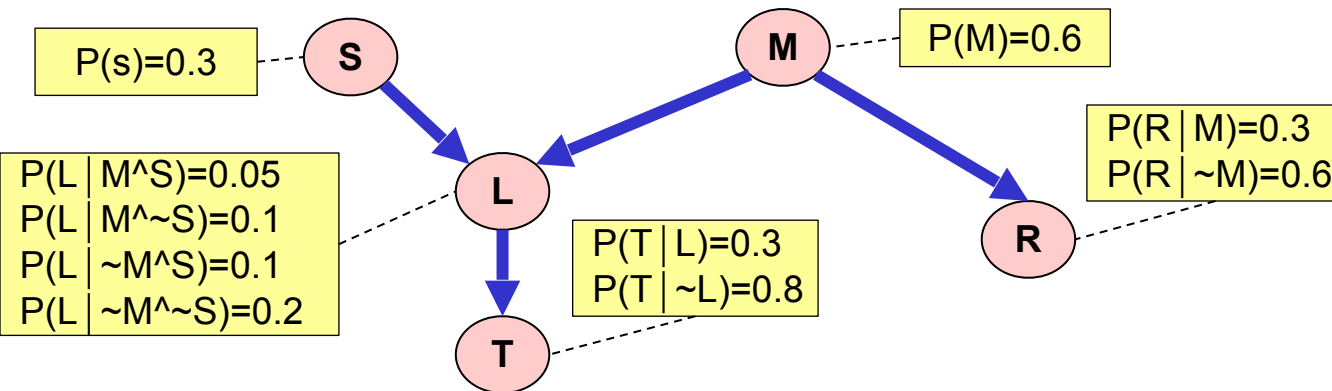
# Computing a Joint Entry

How to compute an entry in a joint distribution?

E.G: What is  $P(S \wedge \sim M \wedge L \wedge \sim R \wedge T)$ ?



# Computing with Bayes Net



$$\begin{aligned}
 &P(T \wedge \sim R \wedge L \wedge \sim M \wedge S) = \\
 &P(T \mid \sim R \wedge L \wedge \sim M \wedge S) * P(\sim R \wedge L \wedge \sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \wedge L \wedge \sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \mid L \wedge \sim M \wedge S) * P(L \wedge \sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \mid \sim M) * P(L \wedge \sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \mid \sim M) * P(L \mid \sim M \wedge S) * P(\sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \mid \sim M) * P(L \mid \sim M \wedge S) * P(\sim M \mid S) * P(S) = \\
 &P(T \mid L) * P(\sim R \mid \sim M) * P(L \mid \sim M \wedge S) * P(\sim M) * P(S).
 \end{aligned}$$

# The general case

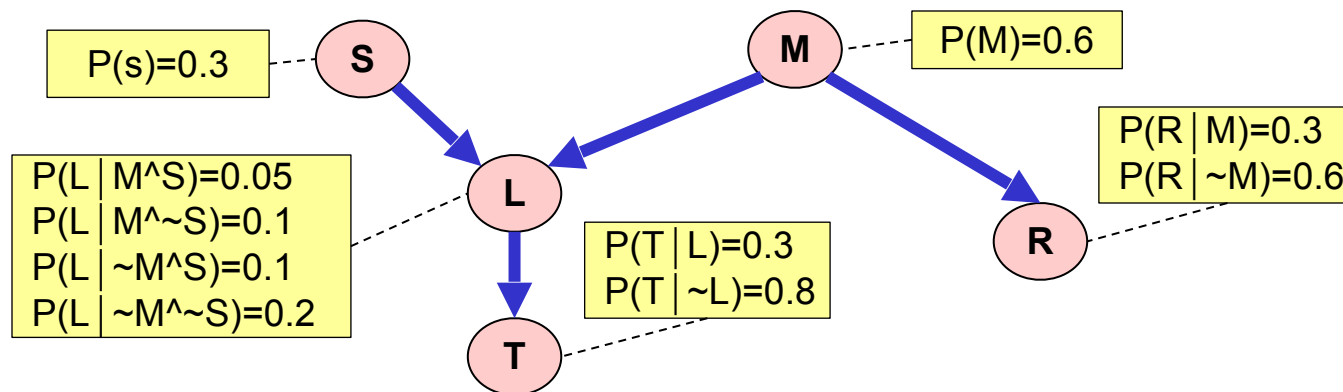
$$\begin{aligned}
 &P(X_1=x_1 \wedge X_2=x_2 \wedge \dots X_{n-1}=x_{n-1} \wedge X_n=x_n) = \\
 &P(X_n=x_n \wedge X_{n-1}=x_{n-1} \wedge \dots X_2=x_2 \wedge X_1=x_1) = \\
 &P(X_n=x_n \mid X_{n-1}=x_{n-1} \wedge \dots X_2=x_2 \wedge X_1=x_1) * P(X_{n-1}=x_{n-1} \wedge \dots X_2=x_2 \wedge X_1=x_1) = \\
 &P(X_n=x_n \mid X_{n-1}=x_{n-1} \wedge \dots X_2=x_2 \wedge X_1=x_1) * P(X_{n-1}=x_{n-1} \mid \dots X_2=x_2 \wedge X_1=x_1) * \\
 &P(X_{n-2}=x_{n-2} \wedge \dots X_2=x_2 \wedge X_1=x_1) = \\
 &\vdots \\
 &\vdots \\
 &= \prod_{i=1}^n P((X_i = x_i) \mid ((X_{i-1} = x_{i-1}) \wedge \dots (X_1 = x_1))) \\
 &= \\
 &\prod_{i=1}^n P((X_i = x_i) \mid \text{Assignments of Parents}(X_i))
 \end{aligned}$$

So any entry in joint pdf table can be computed. And so **any conditional probability** can be computed.



# Where are we now?

- We have a methodology for building Bayes nets.
- We don't require exponential storage to hold our probability table. Only exponential in the maximum number of parents of any node.
- We can compute probabilities of any given assignment of truth values to the variables. And we can do it in time linear with the number of nodes.
- So we can also compute answers to any questions.



E.G. What could we do to compute  $P(R \mid T, \sim S)$ ?

# Where are we now?

Step 1: Compute  $P(R \wedge T \wedge \sim S)$

Step 2: Compute  $P(\sim R \wedge T \wedge \sim S)$

Step 3: Return

$$P(R \wedge T \wedge \sim S)$$

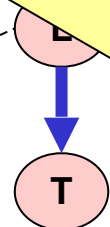
$$P(R \wedge T \wedge \sim S) + P(\sim R \wedge T \wedge \sim S)$$

$P(L \mid M \wedge S)$	$=0.05$
$P(L \mid M \wedge \sim S)$	$=0.1$
$P(L \mid \sim M \wedge S)$	$=0.1$
$P(L \mid \sim M \wedge \sim S)$	$=0.2$

$P(T \mid L)$	$=0.3$
$P(T \mid \sim L)$	$=0.8$

$$P(M)=0.6$$

$P(R \mid M)$	$=0.3$
$P(R \mid \sim M)$	$=0.6$



E.G. What could we do to compute  $P(R \mid T, \sim S)$ ?

ing Bayes nets.

ge to hold our probability table. Only  
er of parents of any node.

ny given assignment of truth values to  
time linear with the number of nodes.

to any questions.

# Where are we now?

Step 1: Compute  $P(R \wedge T \wedge \sim S)$

Step 2: Compute  $P(\sim R \wedge T \wedge \sim S)$

Step 3: Return

$$P(R \wedge T \wedge \sim S)$$

$$P(R \wedge T \wedge \sim S) + P(\sim R \wedge T \wedge \sim S)$$

Sum of all the rows in the Joint that match  $R \wedge T \wedge \sim S$

Sum of all the rows in the Joint that match  $\sim R \wedge T \wedge \sim S$

$P(L \mid M \wedge S)$	$=0.05$
$P(L \mid M \wedge \sim S)$	$=0.1$
$P(L \mid \sim M \wedge S)$	$=0.1$
$P(L \mid \sim M \wedge \sim S)$	$=0.2$

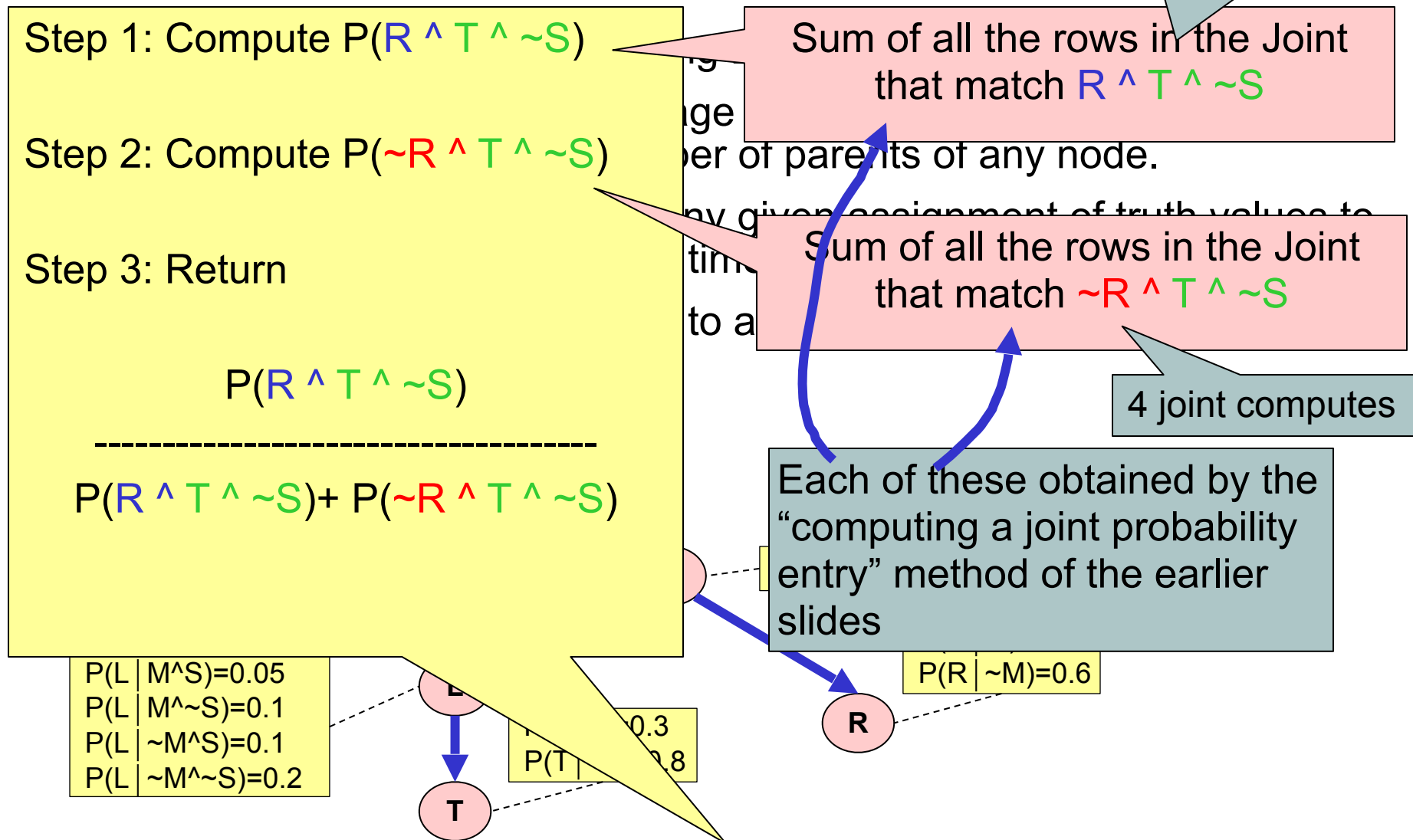
$P(T \mid L)$	$=0.3$
$P(T \mid \sim L)$	$=0.8$

$$P(M)=0.6$$

$P(R \mid M)$	$=0.3$
$P(R \mid \sim M)$	$=0.6$

E.G. What could we do to compute  $P(R \mid T, \sim S)$ ?

# Where are we now?



E.G. What could we do to compute  $P(R \mid T, \sim S)$ ?

# The good news

We can do inference. We can compute any conditional probability:

$P(\text{Some variable} \mid \text{Some other variable values})$

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{joint entries matching } E_1 \text{ and } E_2} P(\text{joint entry})}{\sum_{\text{joint entries matching } E_2} P(\text{joint entry})}$$

# The good news

We can do inference. We can compute any conditional probability:

$P(\text{Some variable} \mid \text{Some other variable values})$

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{joint entries matching } E_1 \text{ and } E_2} P(\text{joint entry})}{\sum_{\text{joint entries matching } E_2} P(\text{joint entry})}$$

Suppose you have  $m$  binary-valued variables in your Bayes Net and expression  $E_2$  mentions  $k$  variables.

How much work is the above computation?

# The sad, bad news

Conditional probabilities by enumerating all matching entries in the joint are expensive:

**Exponential in the number of variables.**

# The sad, bad news

Conditional probabilities by enumerating all matching entries in the joint are expensive:

**Exponential in the number of variables.**

But perhaps there are faster ways of querying Bayes nets?

- In fact, if I ever ask you to manually do a Bayes Net inference, you'll find there are often many tricks to save you time.
- So we've just got to program our computer to do those tricks too, right?



# The sad, bad news

Conditional probabilities by enumerating all matching entries in the joint are expensive:

**Exponential in the number of variables.**

But perhaps there are faster ways of querying Bayes nets?

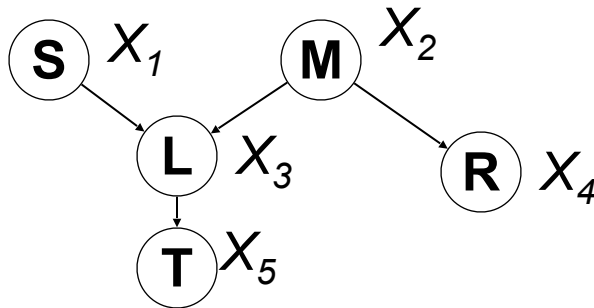
- In fact, if I ever ask you to manually do a Bayes Net inference, you'll find there are often many tricks to save you time.
- So we've just got to program our computer to do those tricks too, right?

## Sadder and worse news:

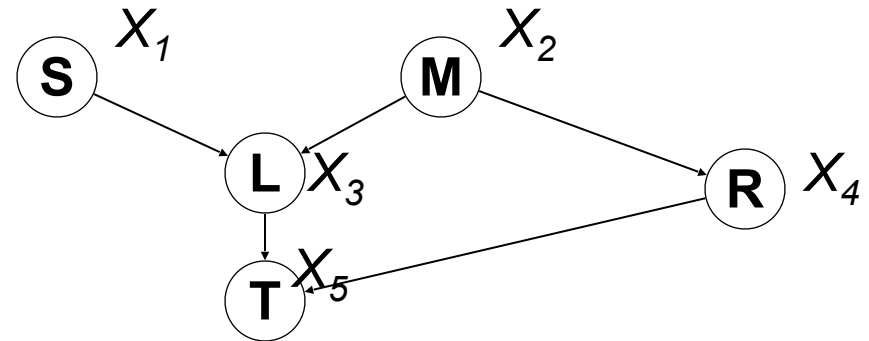
**General querying of Bayes nets is NP-complete.**

# Bayes nets inference algorithms

A poly-tree is a directed acyclic graph in which no two nodes have more than one path between them.



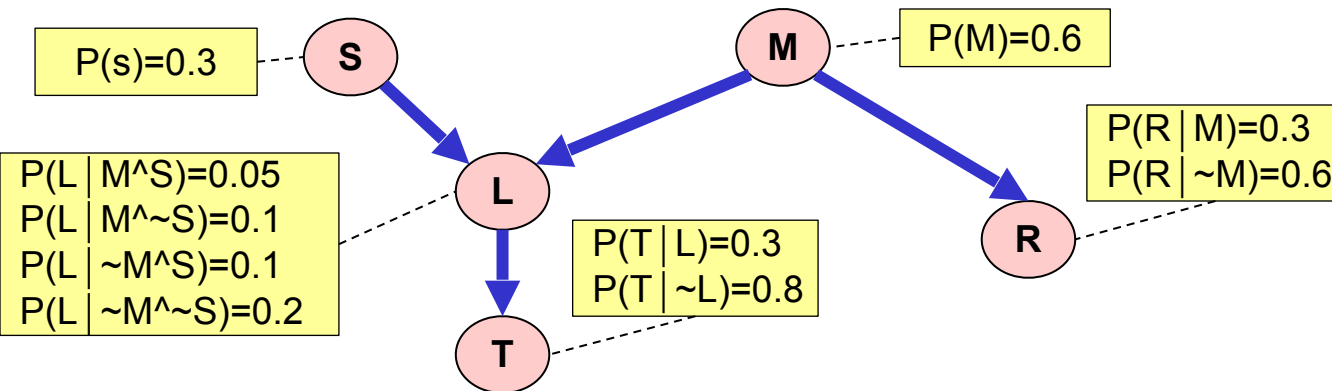
A poly tree



Not a poly tree  
(but still a legal Bayes net)

- If net is a poly-tree, there is a linear-time algorithm (see a later Andrew lecture).
- The best general-case algorithms convert a general net to a poly-tree (often at huge expense) and calls the poly-tree algorithm.
- Another popular, practical approach (doesn't assume poly-tree): Stochastic Simulation.

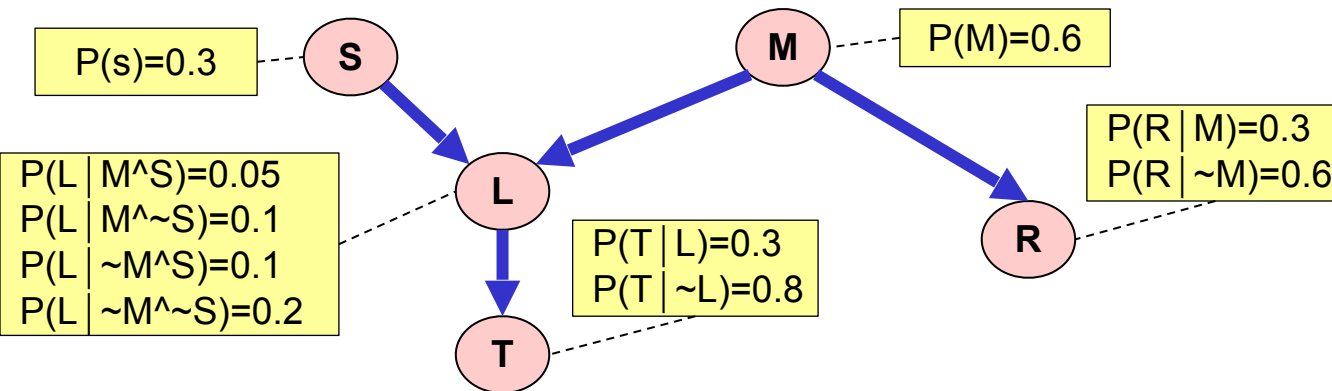
# Sampling from the Joint Distribution



It's pretty easy to generate a set of variable-assignments at random with the same probability as the underlying joint distribution.

How?

# Sampling from the Joint Distribution



1. Randomly choose S. S = True with prob 0.3
2. Randomly choose M. M = True with prob 0.6
3. Randomly choose L. The probability that L is true depends on the assignments of S and M. E.G. if steps 1 and 2 had produced S=True, M=False, then probability that L is true is 0.1
4. Randomly choose R. Probability depends on M.
5. Randomly choose T. Probability depends on L

# A general sampling algorithm

Let's generalize the example on the previous slide to a general Bayes Net.

As in Slides 16-17, call the variables  $X_1 \dots X_n$ , where  $Parents(X_i)$  must be a subset of  $\{X_1 \dots X_{i-1}\}$ .

For  $i=1$  to  $n$ :

1. Find parents, if any, of  $X_i$ . Assume  $n(i)$  parents. Call them  $X_{p(i,1)}, X_{p(i,2)}, \dots, X_{p(i,n(i))}$ .
2. Recall the values that those parents were randomly given:  $x_{p(i,1)}, x_{p(i,2)}, \dots, x_{p(i,n(i))}$ .
3. Look up in the lookup-table for:  
|  $X_{p(i,1)}=x_{p(i,1)}, X_{p(i,2)}=x_{p(i,2)} \dots X_{p(i,n(i))}=x_{p(i,n(i))}$   $P(X_i=True)$
4. Randomly set  $x_i=True$  according to this probability

$x_1, x_2, \dots, x_n$  are now a sample from the joint distribution of  $X_1, X_2, \dots, X_n$ .

# Stochastic Simulation Example

Someone wants to know  $P(R = \text{True} \mid T = \text{True} \wedge S = \text{False})$

We'll do lots of random samplings and count the number of occurrences of the following:

- $N_c$  : Num. samples in which  $T=\text{True}$  and  $S=\text{False}$ .
- $N_s$  : Num. samples in which  $R=\text{True}$ ,  $T=\text{True}$  and  $S=\text{False}$ .
- $N$  : Number of random samplings

Now if  $N$  is big enough:

$N_c / N$  is a good estimate of  $P(T=\text{True} \text{ and } S=\text{False})$ .

$N_s / N$  is a good estimate of  $P(R=\text{True}, T=\text{True}, S=\text{False})$ .

$P(R \mid T \wedge \sim S) = P(R \wedge T \wedge \sim S) / P(T \wedge \sim S)$ , so  $N_s / N_c$  can be a good estimate of  $P(R \mid T \wedge \sim S)$ .

# General Stochastic Simulation

Someone wants to know  $P(E_1 \mid E_2)$

We'll do lots of random samplings and count the number of occurrences of the following:

- $N_c$  : Num. samples in which  $E_2$
- $N_s$  : Num. samples in which  $E_1$  and  $E_2$
- $N$  : Number of random samplings

Now if  $N$  is big enough:

$N_c / N$  is a good estimate of  $P(E_2)$ .

$N_s / N$  is a good estimate of  $P(E_1, E_2)$ .

$P(E_1 \mid E_2) = P(E_1 \wedge E_2) / P(E_2)$ , so  $N_s / N_c$  can be a good estimate of  $P(E_1 \mid E_2)$ .

# Likelihood weighting

## Problem with Stochastic Sampling:

With lots of constraints in  $E$ , or unlikely events in  $E$ , then most of the simulations will be thrown away, (they'll have no effect on  $N_c$ , or  $N_s$ ).

Imagine we're part way through our simulation.

In  $E_2$  we have the constraint  $X_i = v$

We're just about to generate a value for  $X_i$  at random. Given the values assigned to the parents, we see that  $P(X_i = v \mid \text{parents}) = p$ .

Now we know that with stochastic sampling:

- we'll generate " $X_i = v$ " proportion  $p$  of the time, and proceed.
- And we'll generate a different value proportion  $1-p$  of the time, and the simulation will be wasted.

Instead, always generate  $X_i = v$ , but weight the answer by weight " $p$ " to compensate.



# Likelihood weighting

Set  $N_c := 0$ ,  $N_s := 0$

1. Generate a random assignment of all variables that matches  $E_2$ . This process returns a weight  $w$ .
2. Define  $w$  to be the probability that this assignment would have been generated instead of an unmatching assignment during its generation in the original algorithm. Fact:  $w$  is a product of all likelihood factors involved in the generation.
3.  $N_c := N_c + w$
4. If our sample matches  $E_1$  then  $N_s := N_s + w$
5. Go to 1

Again,  $N_s / N_c$  estimates  $P(E_1 \mid E_2)$

# Case Study I

Pathfinder system. (Heckerman 1991, Probabilistic Similarity Networks, MIT Press, Cambridge MA).

- Diagnostic system for lymph-node diseases.
- 60 diseases and 100 symptoms and test-results.
- 14,000 probabilities
- Expert consulted to make net.
  - 8 hours to determine variables.
  - 35 hours for net topology.
  - 40 hours for probability table values.
- Apparently, the experts found it quite easy to invent the causal links and probabilities.
- Pathfinder is now outperforming the world experts in diagnosis. Being extended to several dozen other medical domains.

# Questions

- What are the strengths of probabilistic networks compared with propositional logic?
- What are the weaknesses of probabilistic networks compared with propositional logic?
- What are the strengths of probabilistic networks compared with predicate logic?
- What are the weaknesses of probabilistic networks compared with predicate logic?
- (How) could predicate logic and probabilistic networks be combined?

# What you should know

- The meanings and importance of independence and conditional independence.
- The definition of a Bayes net.
- Computing probabilities of assignments of variables (i.e. members of the joint p.d.f.) with a Bayes net.
- The slow (exponential) method for computing arbitrary, conditional probabilities.
- The stochastic simulation method and likelihood weighting.