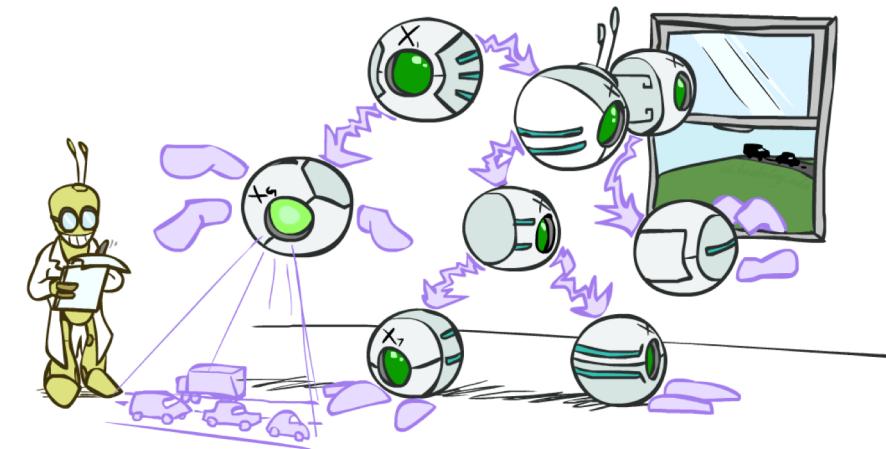


# CSE 573: Artificial Intelligence

Hanna Hajishirzi  
Bayes Net Inference

slides adapted from  
Dan Klein, Pieter Abbeel ai.berkeley.edu  
And Dan Weld, Luke Zettlemoyer



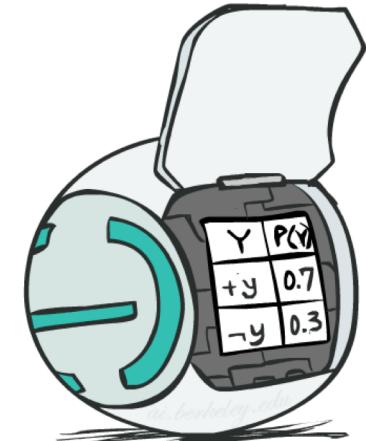
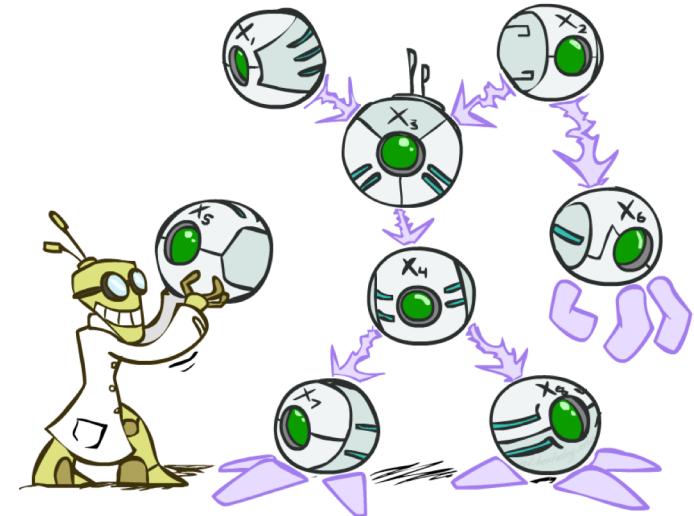
# Bayes' Net Representation

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over X, one for each combination of parents' values

$$P(X|a_1 \dots a_n)$$

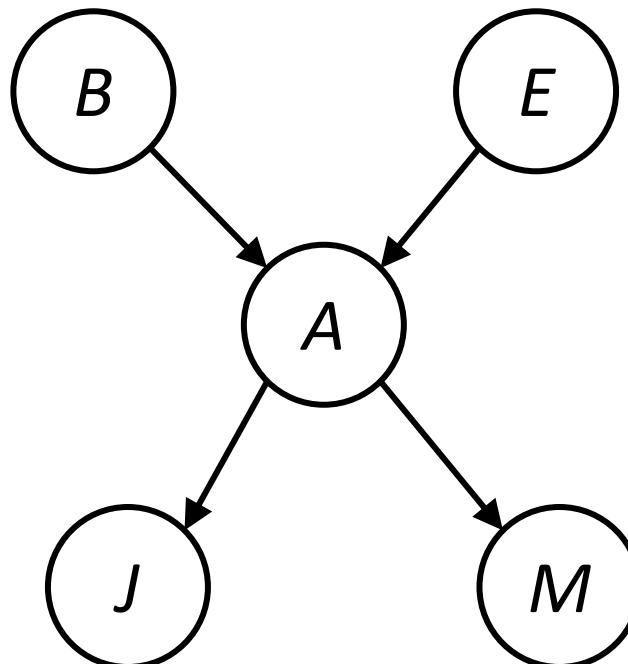
- Bayes' nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parents}(X_i))$$



# Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999

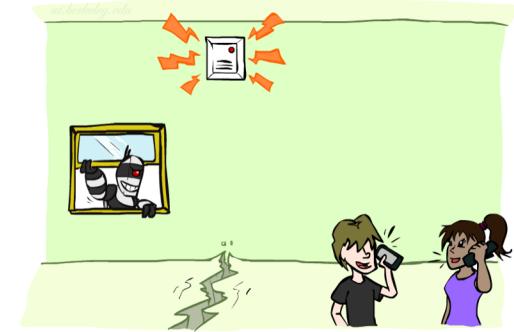


E	P(E)
+e	0.002
-e	0.998

A	J	P(J A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

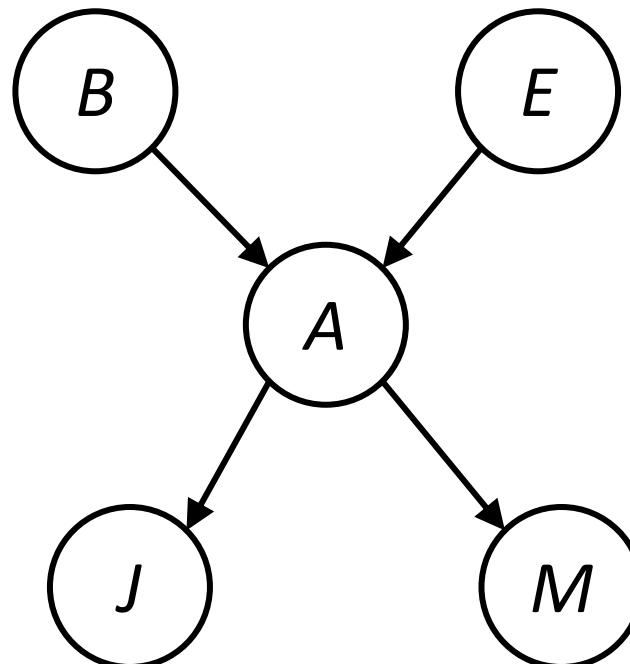
$$P(+b, -e, +a, -j, +m) =$$



B	E	A	P(A B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

# Example: Alarm Network

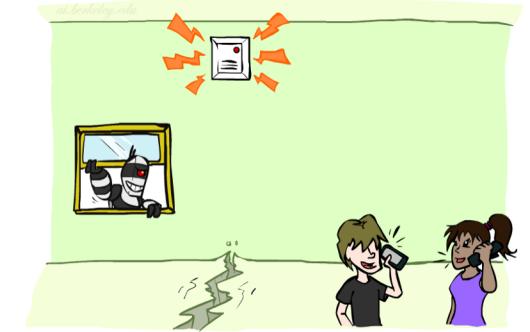
B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99



$$P(+b, -e, +a, -j, +m) =$$

$$P(+b)P(-e)P(+a|+b, -e)P(-j|+a)P(+m|+a) =$$

$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

B	E	A	P(A B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

# Inference

- Inference: calculating some useful quantity from a joint probability distribution

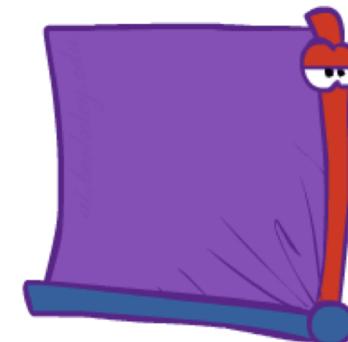
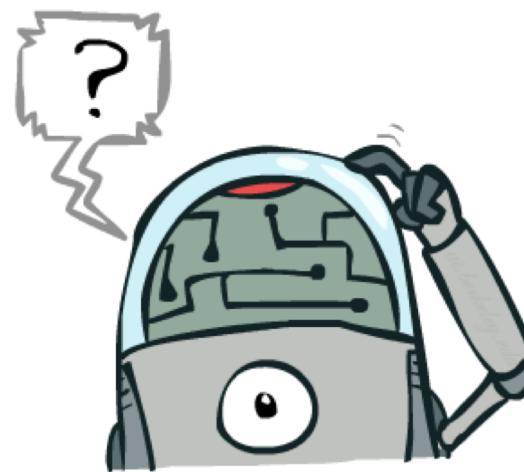
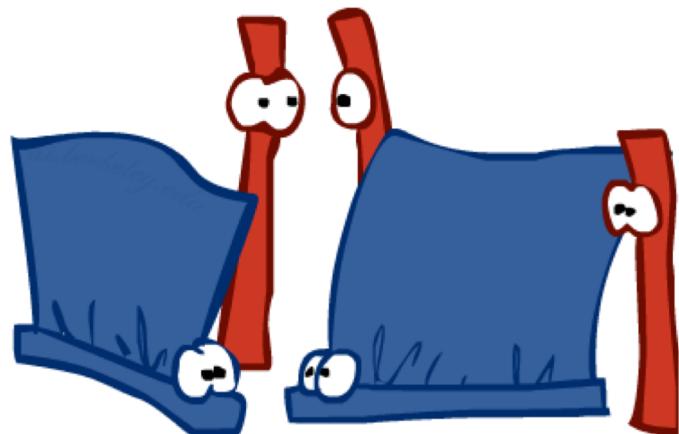
- Examples:

- Posterior probability

$$P(Q|E_1 = e_1, \dots, E_k = e_k)$$

- Most likely explanation:

$$\operatorname{argmax}_q P(Q = q|E_1 = e_1 \dots)$$



# Inference by Enumeration in Bayes' Net

- Given unlimited time, inference in BNs is easy

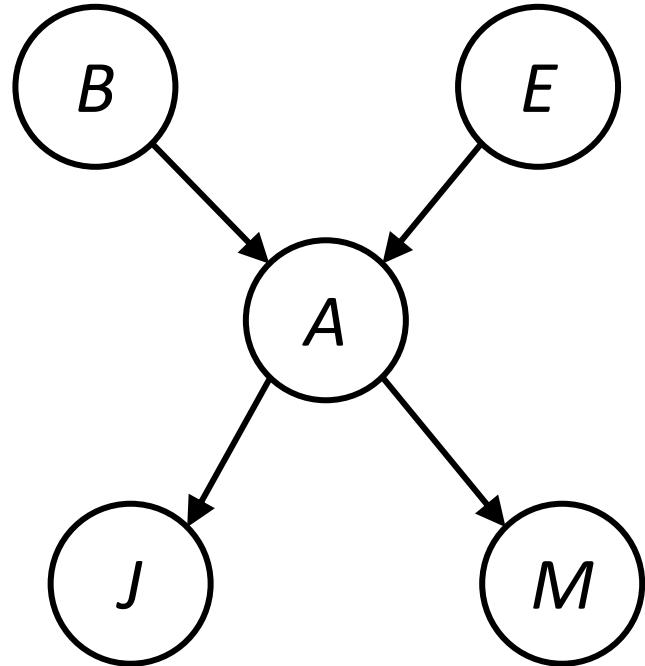
$$P(B \mid +j, +m) \propto_B P(B, +j, +m)$$

$$= \sum_{e,a} P(B, e, a, +j, +m)$$

$$= \sum_{e,a} P(B)P(e)P(a|B, e)P(+j|a)P(+m|a)$$

$$= P(B)P(+e)P(+a|B, +e)P(+j|+a)P(+m|+a) + P(B)P(+e)P(-a|B, +e)P(+j|-a)P(+m|-a)$$

$$P(B)P(-e)P(+a|B, -e)P(+j|+a)P(+m|+a) + P(B)P(-e)P(-a|B, -e)P(+j|-a)P(+m|-a)$$



# Example: Traffic Domain

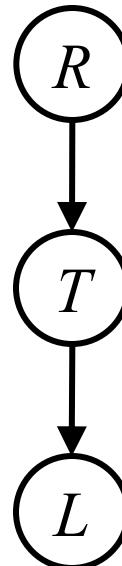
- Random Variables

- R: Raining
- T: Traffic
- L: Late for class!

$$P(L) = ?$$

$$= \sum_{r,t} P(r,t,L)$$

$$= \sum_{r,t} P(r)P(t|r)P(L|t)$$



$$P(R)$$

+r	0.1
-r	0.9

$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

# Inference by Enumeration: Procedural Outline

- Track objects called **factors**
- Initial factors are local CPTs (one per node)

$$P(R)$$

+r	0.1
-r	0.9

$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- Any known values are selected
  - E.g. if we know  $L = +\ell$ , the initial factors are

$$P(R)$$

+r	0.1
-r	0.9

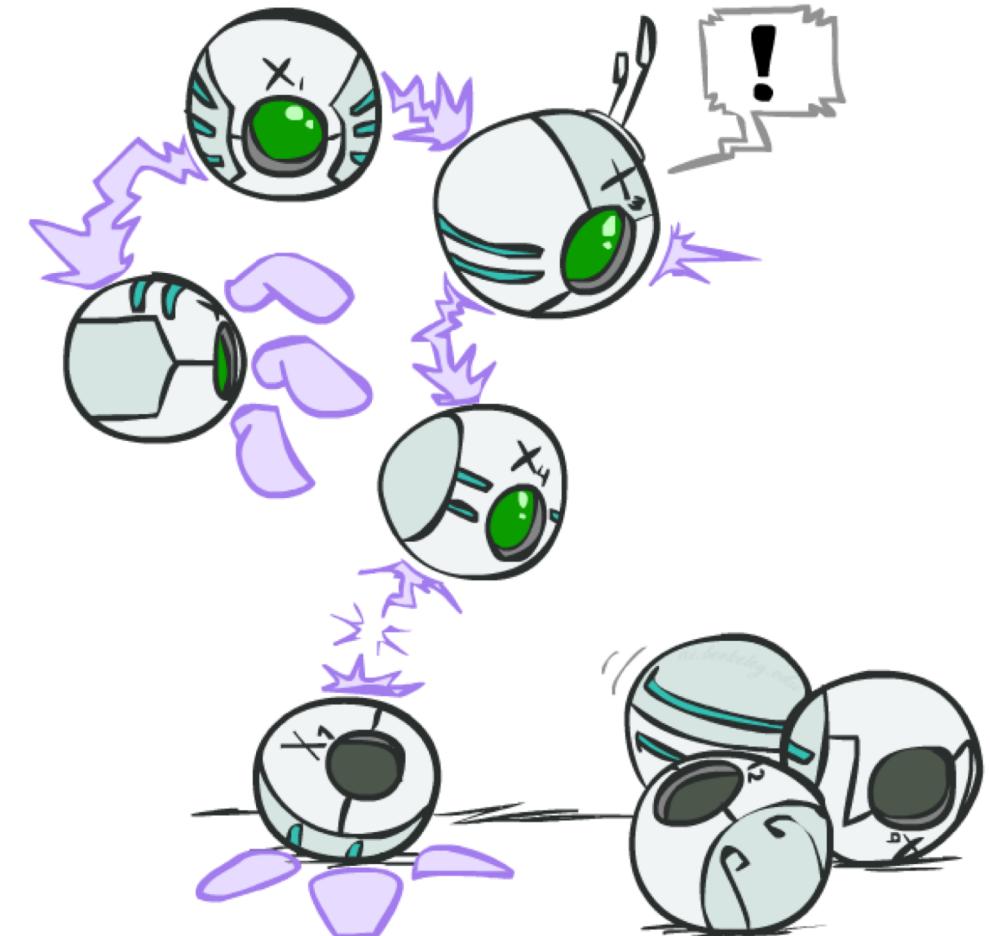
$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(+\ell|T)$$

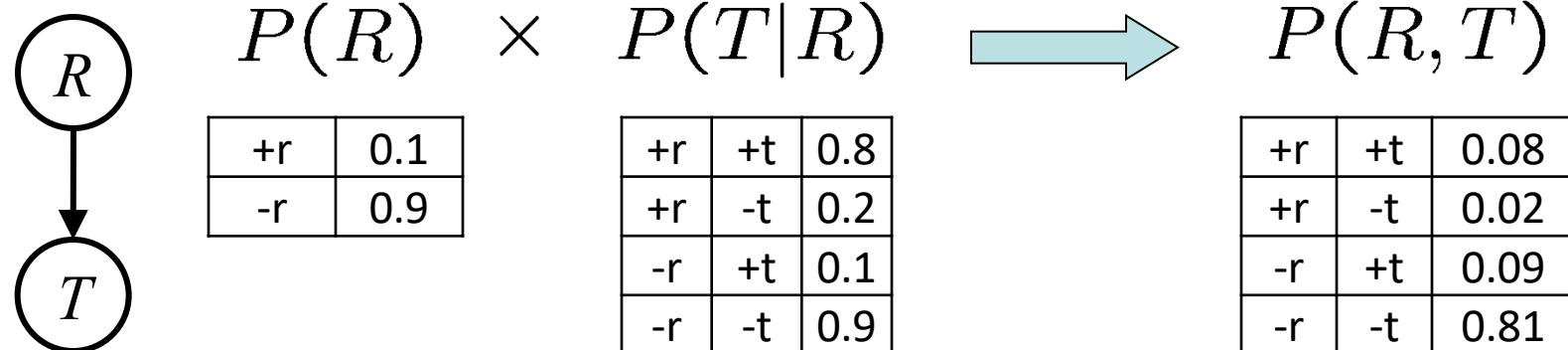
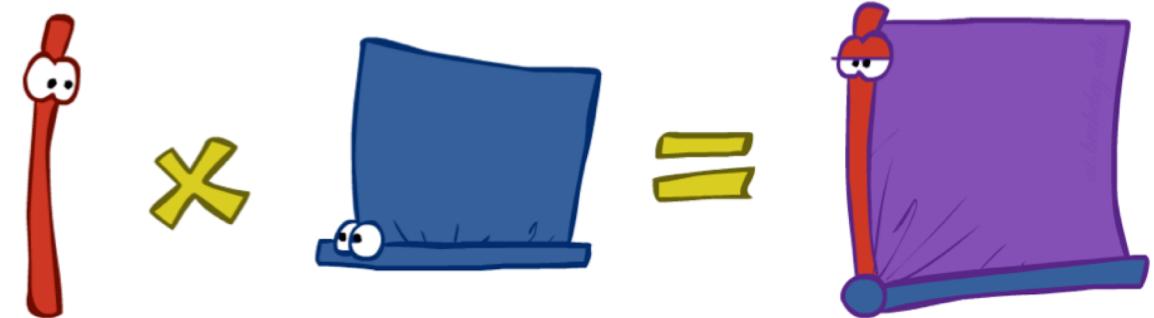
+t	+l	0.3
-t	+l	0.1

- Procedure: Join all factors, then sum out all hidden variables



# Operation 1: Join Factors

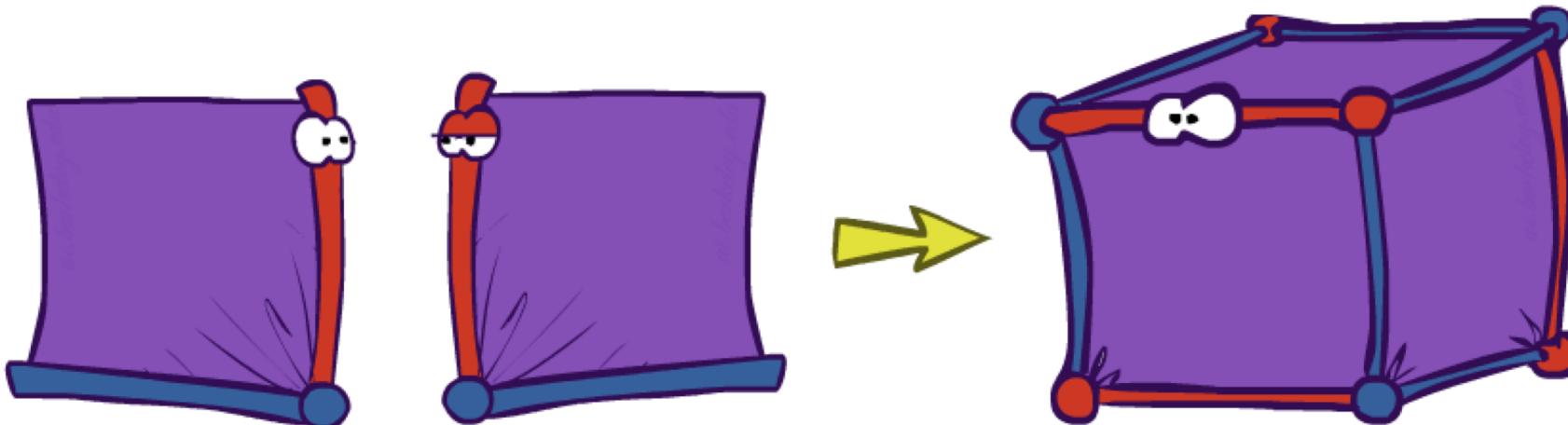
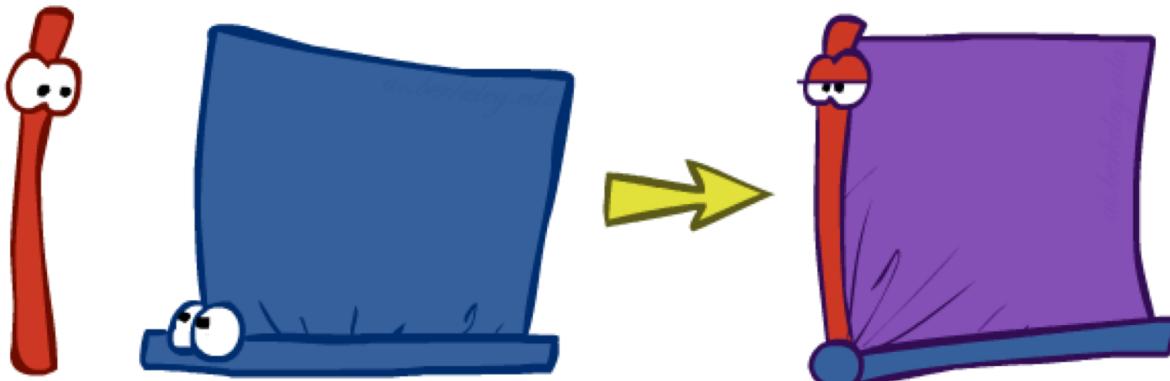
- First basic operation: **joining factors**
- Combining factors:
  - Just like a database join
  - Get all factors over the joining variable
  - Build a new factor over the union of the variables involved
- Example: Join on R



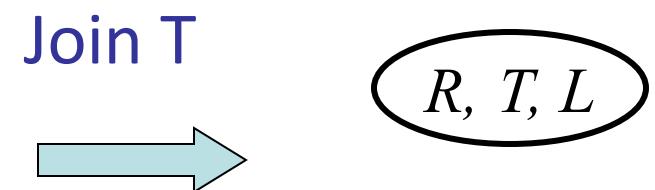
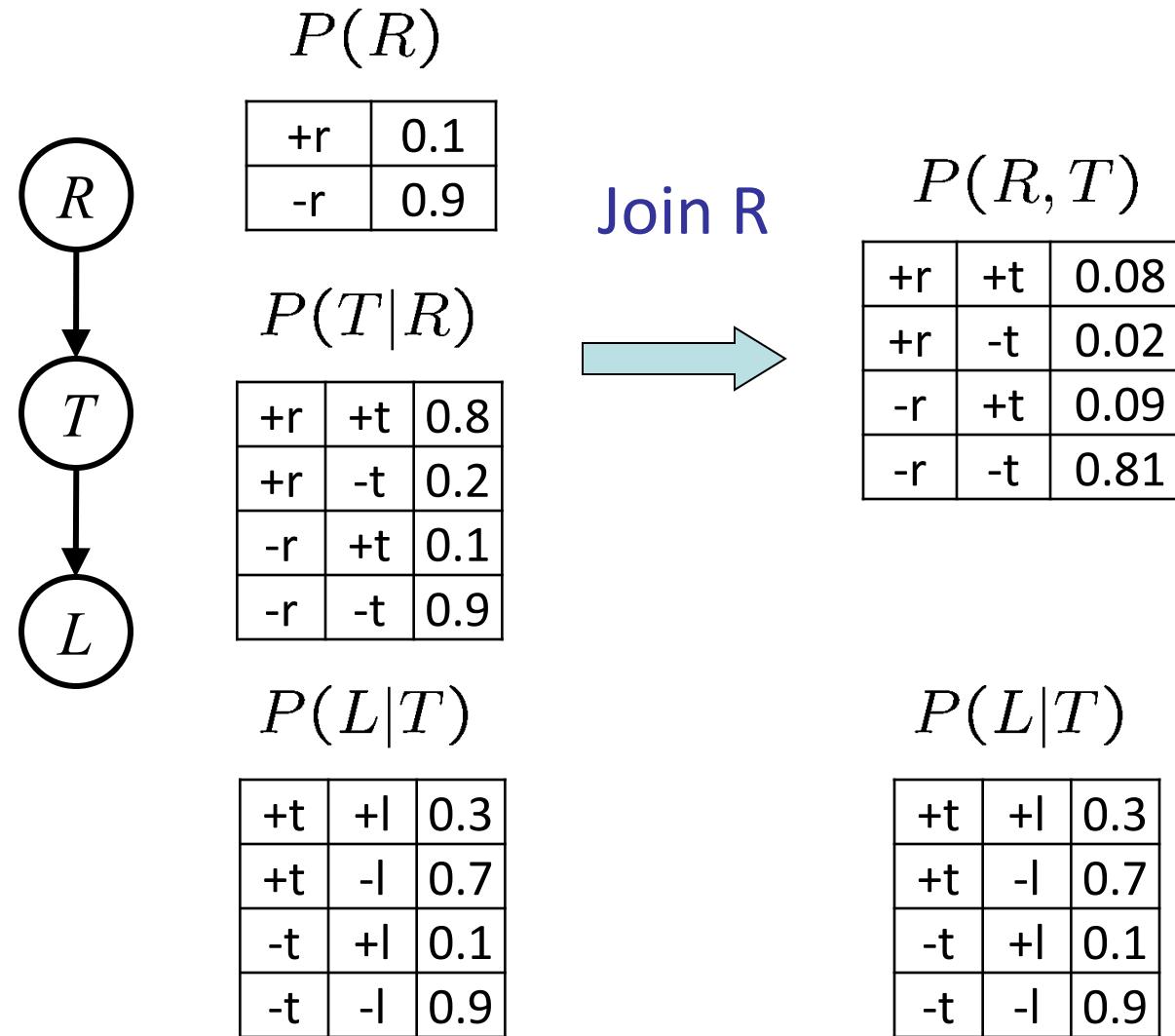
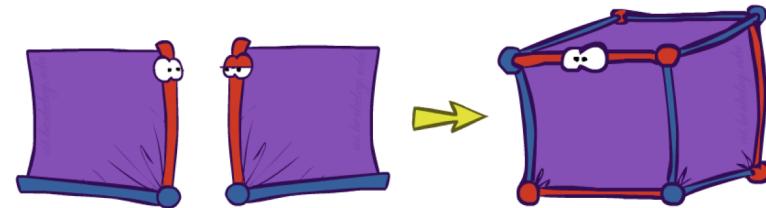
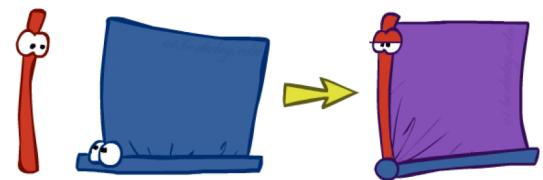
- Computation for each entry: pointwise products  $\forall r, t : P(r, t) = P(r) \cdot P(t|r)$

# Example: Multiple Joins

---



# Example: Multiple Joins



$P(R, T, L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

# Operation 2: Eliminate

- Second basic operation: **marginalization**
- Take a factor and sum out a variable
  - Shrinks a factor to a smaller one
  - A **projection** operation
- Example:

$$P(R, T)$$

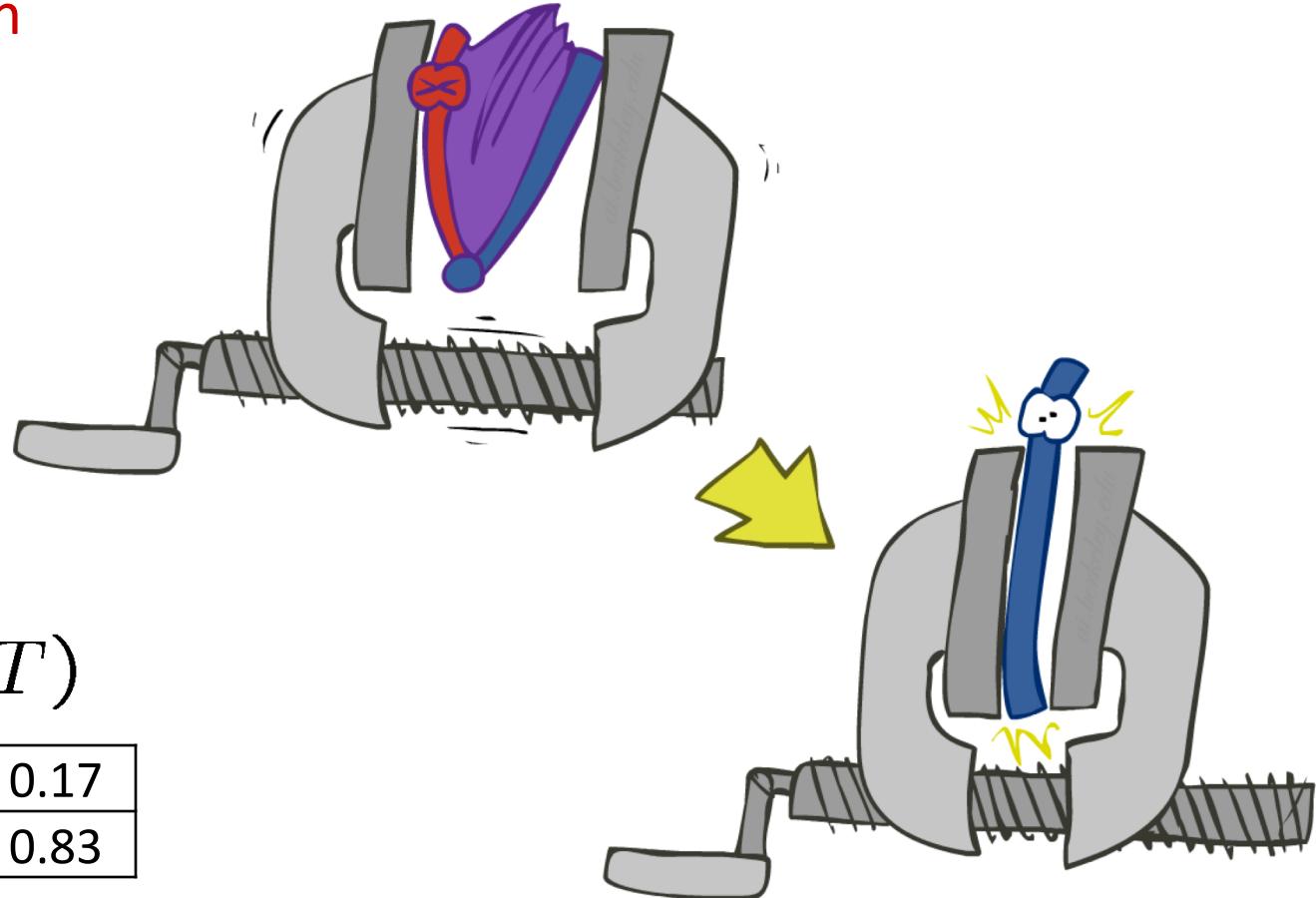
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

sum  $R$



$$P(T)$$

+t	0.17
-t	0.83



# Multiple Elimination

$P(R, T, L)$

$+r$	$+t$	$+l$	$0.024$
$+r$	$+t$	$-l$	$0.056$
$+r$	$-t$	$+l$	$0.002$
$+r$	$-t$	$-l$	$0.018$
$-r$	$+t$	$+l$	$0.027$
$-r$	$+t$	$-l$	$0.063$
$-r$	$-t$	$+l$	$0.081$
$-r$	$-t$	$-l$	$0.729$

$R, T, L$

Sum  
out R

$T, L$

Sum  
out T

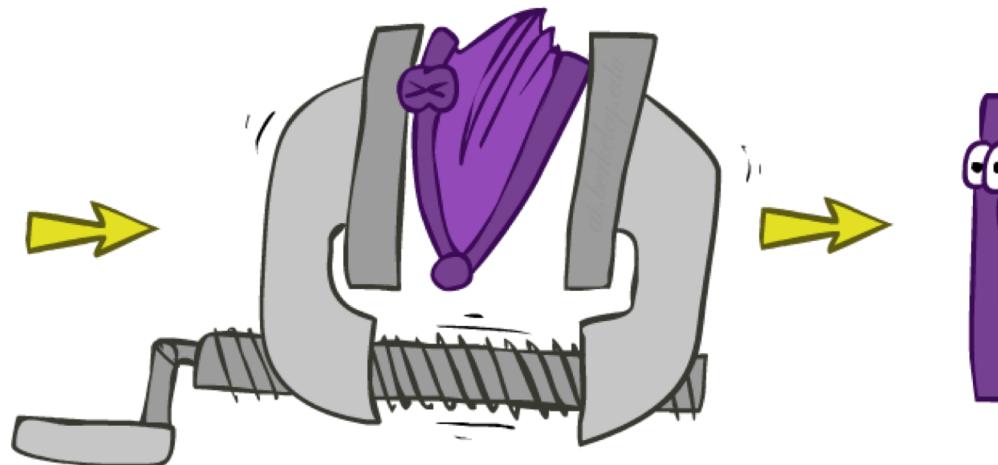
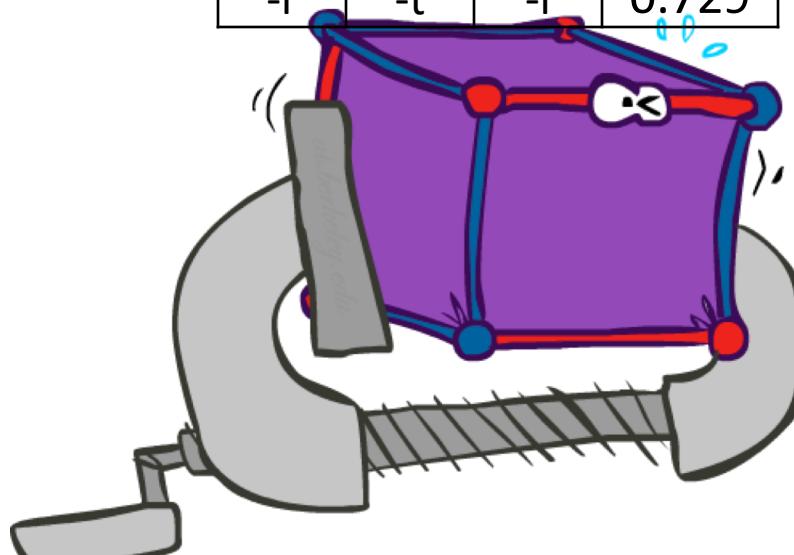
$L$

$P(T, L)$

$+t$	$+l$	$0.051$
$+t$	$-l$	$0.119$
$-t$	$+l$	$0.083$
$-t$	$-l$	$0.747$

$P(L)$

$+l$	$0.134$
$-l$	$0.866$



# Thus Far: Multiple Join, Multiple Eliminate (= Inference by Enumeration)

$P(R)$

$P(T|R)$

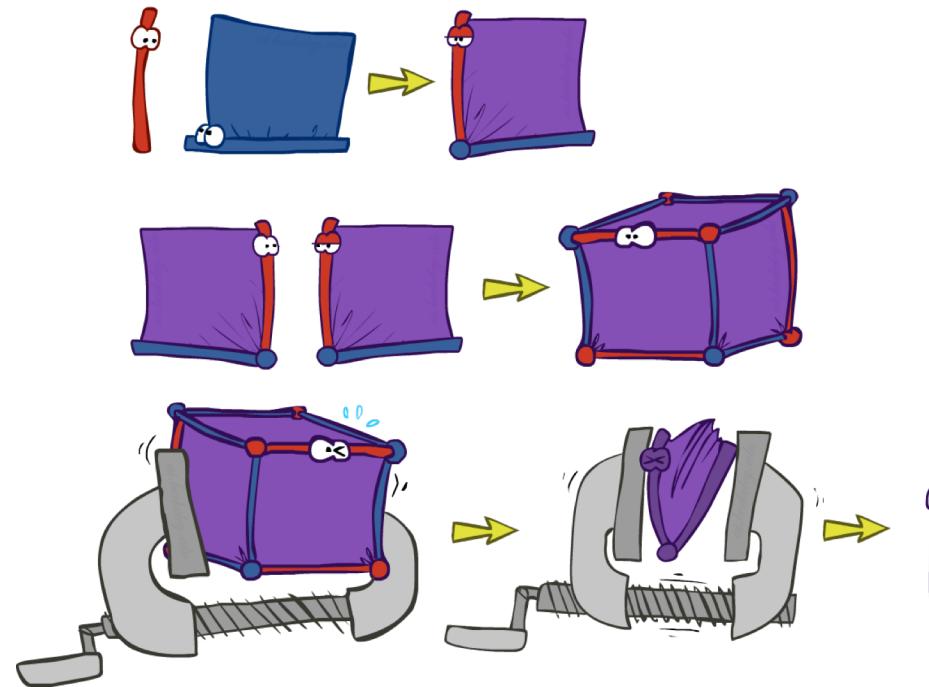


$P(R, T, L)$



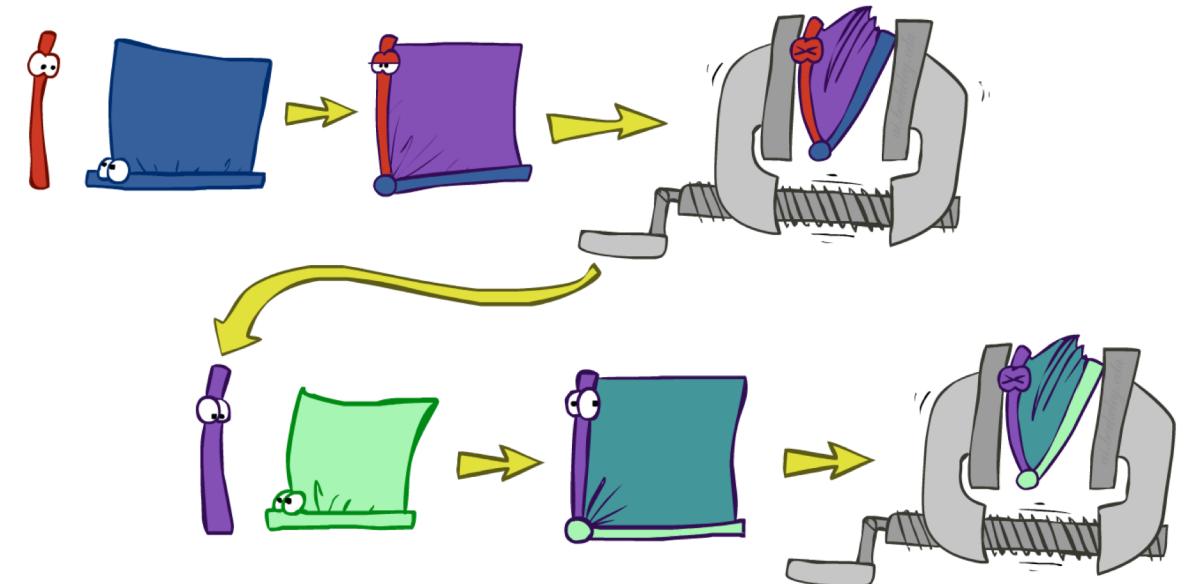
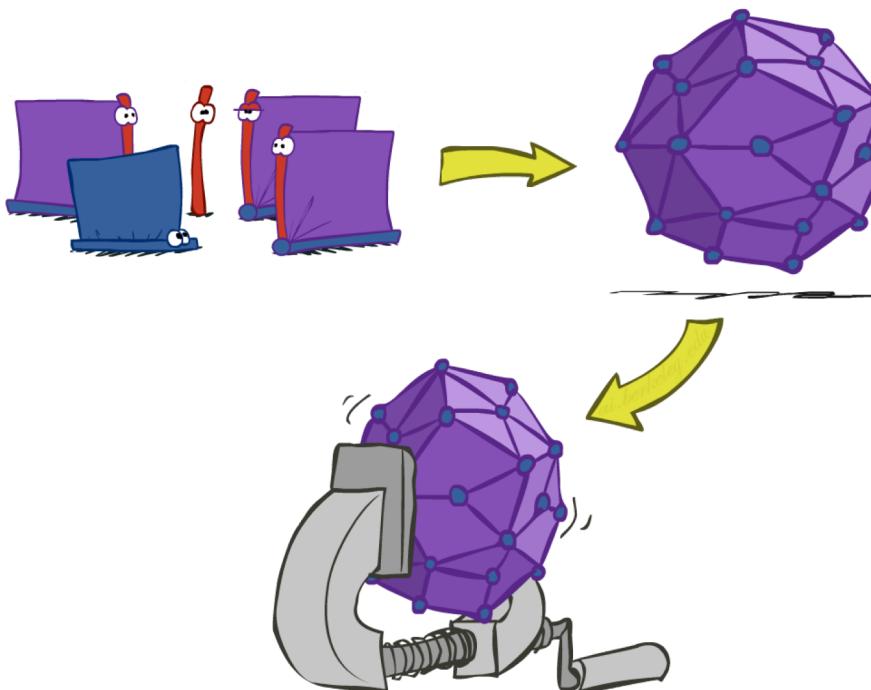
$P(L)$

$P(L|T)$

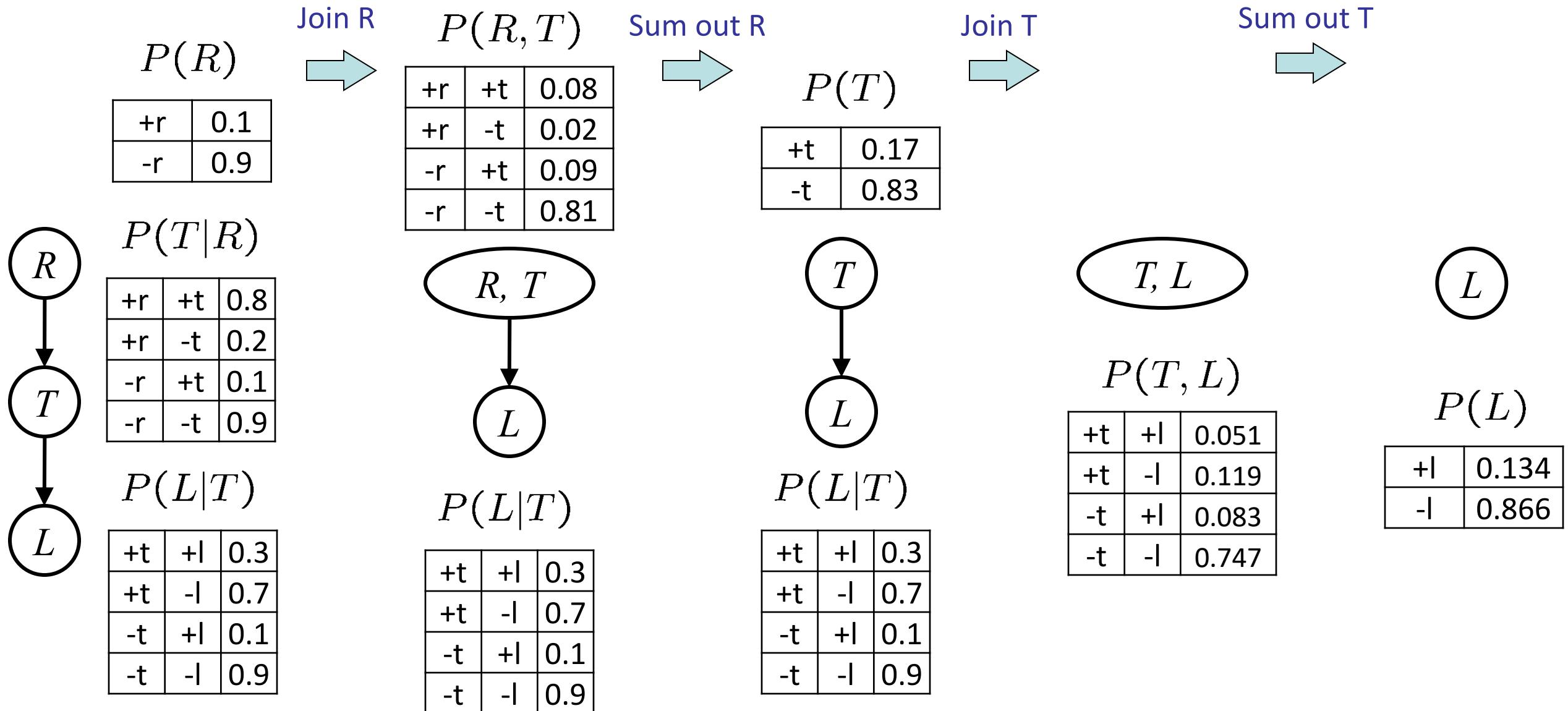


# Inference by Enumeration vs. Variable Elimination

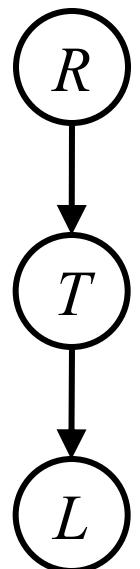
- Why is inference by enumeration so slow?
  - You join up the whole joint distribution before you sum out the hidden variables
- Idea: interleave joining and marginalizing!
  - Called “Variable Elimination”
  - Still NP-hard, but usually much faster than inference by enumeration



# Marginalizing Early! (aka VE)



# Traffic Domain



$$P(L) = ?$$

- Inference by Enumeration

$$= \sum_t \sum_r P(L|t) P(r) P(t|r)$$

Join on r

Join on t

Eliminate r

Eliminate t

- Variable Elimination

$$= \sum_t P(L|t) \sum_r P(r) P(t|r)$$

Join on r

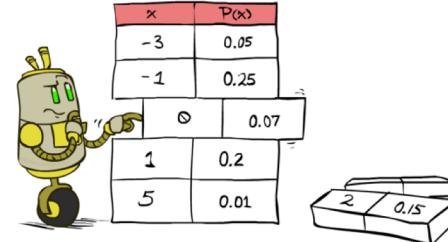
Eliminate r

Join on t

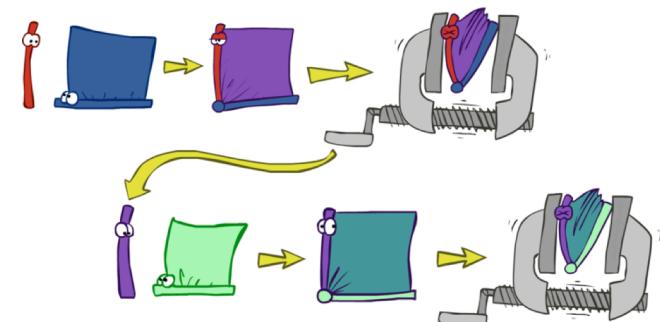
Eliminate t

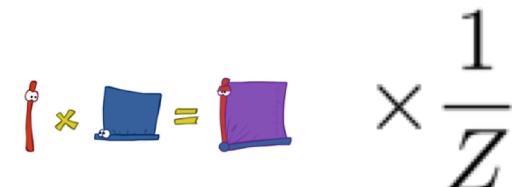
# General Variable Elimination

- Query:  $P(Q|E_1 = e_1, \dots, E_k = e_k)$
- Start with initial factors:
  - Local CPTs (but instantiated by evidence)
- While there are still hidden variables (not Q or evidence):
  - Pick a hidden variable H
  - Join all factors mentioning H
  - Eliminate (sum out) H
- Join all remaining factors and normalize



x	P(x)
-3	0.05
-1	0.25
0	0.07
1	0.2
5	0.01
2	0.15




$$\left( \times \text{ [blue]} = \text{ [purple]} \right) \times \frac{1}{Z}$$

# Evidence

- If evidence, start with factors that select that evidence

- No evidence uses these initial factors:

$$P(R)$$

+r	0.1
-r	0.9

$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- Computing  $P(L|+r)$  the initial factors become:

$$P(+r)$$

+r	0.1
----	-----

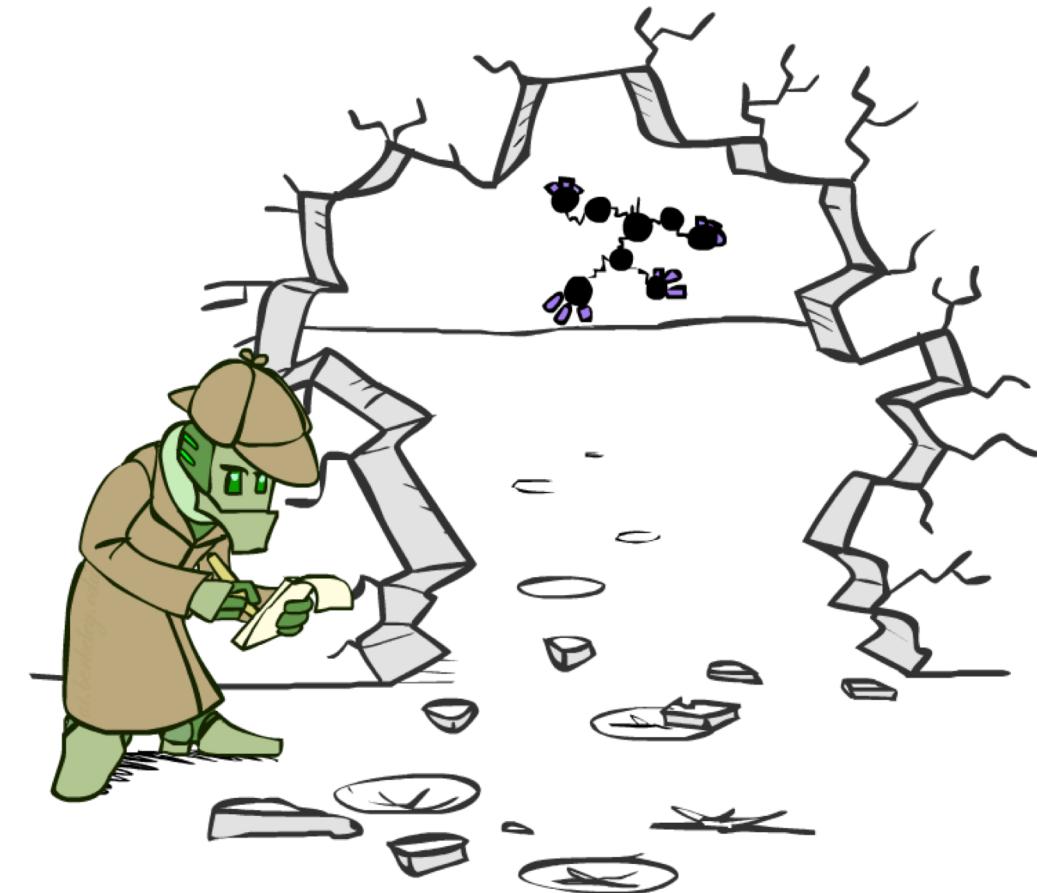
$$P(T|+r)$$

+r	+t	0.8
+r	-t	0.2

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- We eliminate all vars other than query + evidence



# Evidence II

- Result will be a selected joint of query and evidence
  - E.g. for  $P(L | +r)$ , we would end up with:

$P(+r, L)$

+r	+l	0.026
+r	-l	0.074

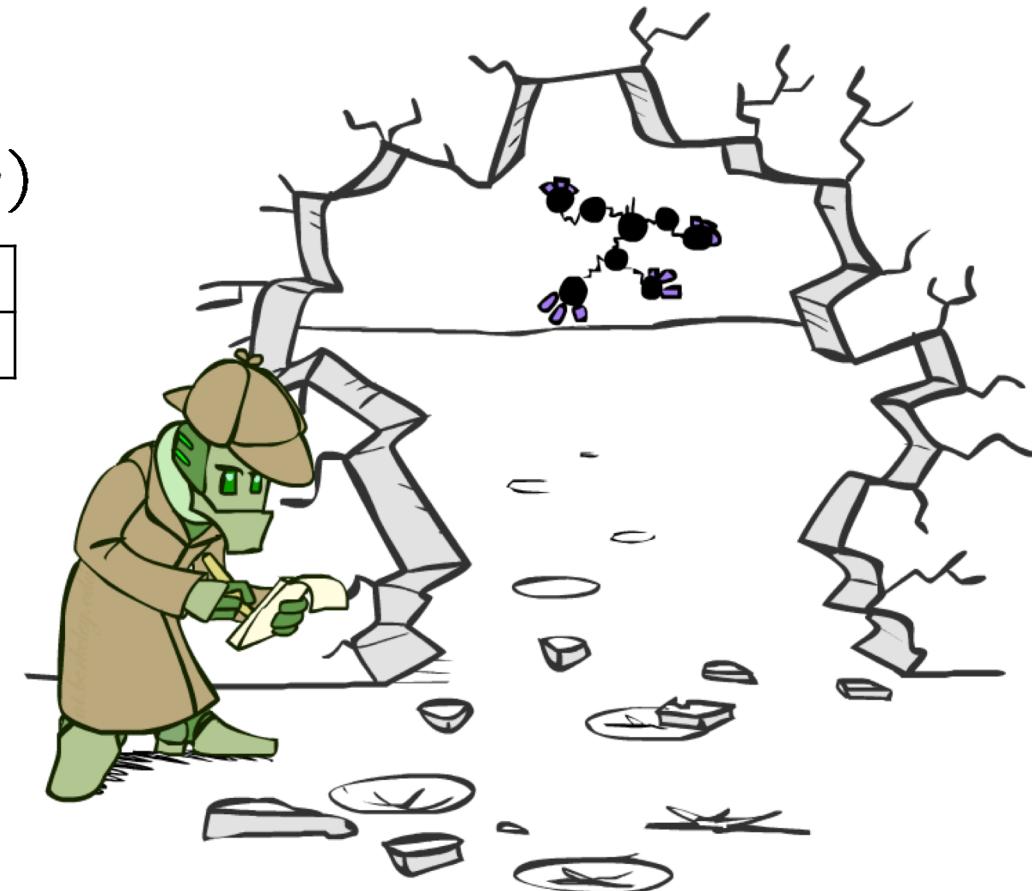
Normalize

$P(L | +r)$

+l	0.26
-l	0.74



- To get our answer, just normalize this!
- That's it!



# Inference by Enumeration

- General case:

- Evidence variables:  $E_1 \dots E_k = e_1 \dots e_k$
- Query\* variable:  $Q$
- Hidden variables:  $H_1 \dots H_r$

$$\left. \begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array} \right\} X_1, X_2, \dots, X_n$$

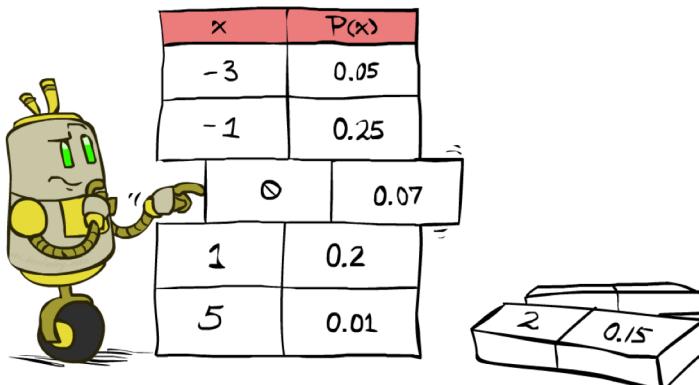
*All variables*

- We want:

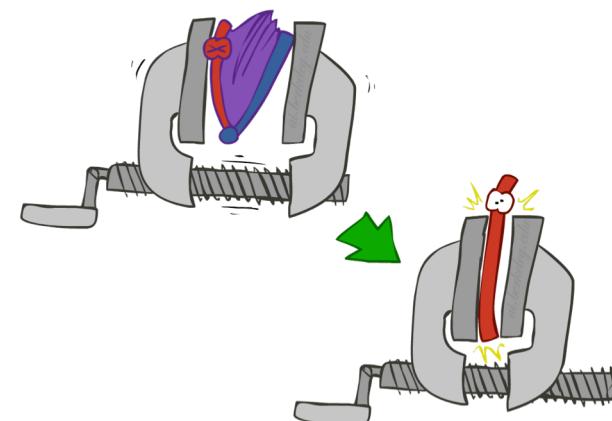
$$P(Q|e_1 \dots e_k)$$

\* Works fine with multiple query variables, too

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out  $H$  to get joint of Query and evidence



- Compute joint

$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Q, h_1 \dots h_r, e_1 \dots e_k)}_{\text{Sum out hidden variables } X_1, X_2, \dots, X_n}$$

- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \dots e_k)$$

$$P(Q|e_1 \dots e_k) = \frac{1}{Z} P(Q, e_1 \dots e_k)$$

# Variable Elimination

- General case:

- Evidence variables:  $E_1 \dots E_k = e_1 \dots e_k$
- Query\* variable:  $Q$
- Hidden variables:  $H_1 \dots H_r$

$$\left. \begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array} \right\} X_1, X_2, \dots, X_n$$

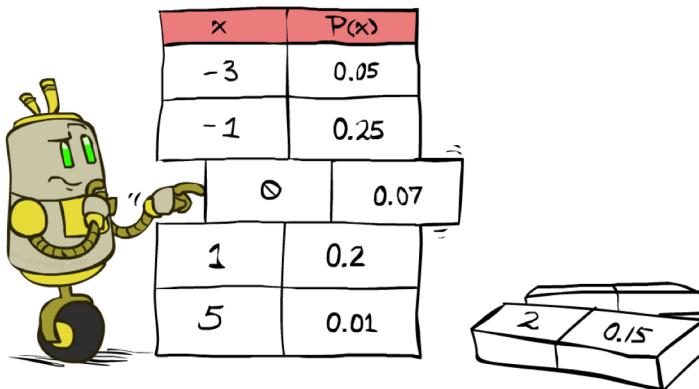
*All variables*

- We want:

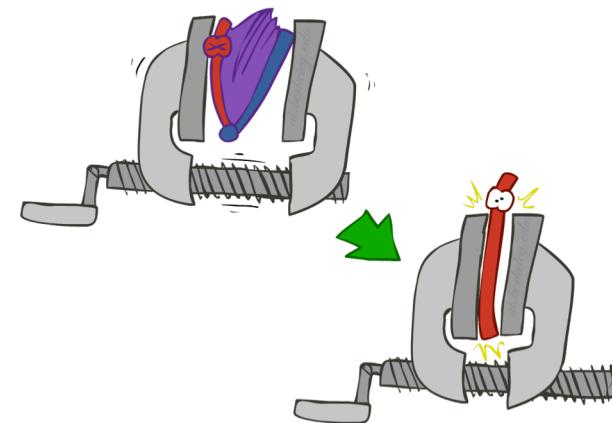
\* Works fine with multiple query variables, too

$$P(Q|e_1 \dots e_k)$$

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out  $H$  to get joint of Query and evidence



- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \dots e_k)$$

$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Q, h_1 \dots h_r, e_1 \dots e_k)}_{X_1, X_2, \dots, X_n}$$

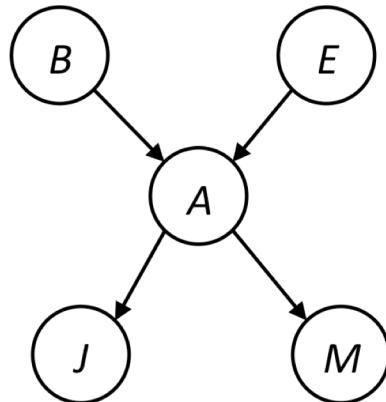
- Interleave joining and summing out  $X_1, X_2, \dots, X_n$

$$P(Q|e_1 \dots e_k) = \frac{1}{Z} P(Q, e_1 \dots e_k)$$

# Example

$$P(B|j, m) \propto P(B, j, m)$$

$P(B)$	$P(E)$	$P(A B, E)$	$P(j A)$	$P(m A)$
--------	--------	-------------	----------	----------

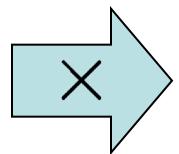


Choose A

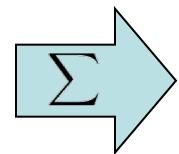
$$P(A|B, E)$$

$$P(j|A)$$

$$P(m|A)$$



$$P(j, m, A|B, E)$$



$$P(j, m|B, E)$$

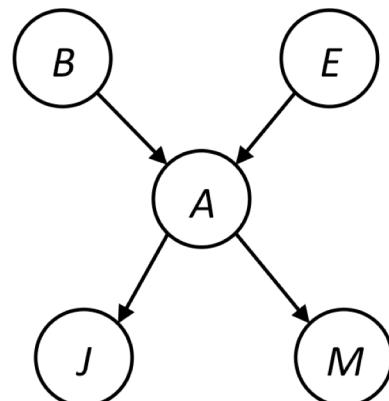
$P(B)$	$P(E)$	$P(j, m B, E)$
--------	--------	----------------

# Example

$P(B)$	$P(E)$	$P(j, m B, E)$
--------	--------	----------------

Choose E

$$\begin{array}{ccccc} P(E) & \xrightarrow{\times} & P(j, m, E|B) & \xrightarrow{\sum} & P(j, m|B) \\ P(j, m|B, E) & & & & \end{array}$$



$P(B)$	$P(j, m B)$
--------	-------------

Finish with B

$$\begin{array}{ccccc} P(B) & \xrightarrow{\times} & P(j, m, B) & \xrightarrow{\text{Normalize}} & P(B|j, m) \\ P(j, m|B) & & & & \end{array}$$

# Example

$$P(B|j, m) \propto P(B, j, m)$$

$P(B)$	$P(E)$	$P(A B, E)$	$P(j A)$	$P(m A)$
--------	--------	-------------	----------	----------

$$P(B|j, m) \propto P(B, j, m)$$

$$= \sum_{e,a} P(B, j, m, e, a)$$

$$= \sum_{e,a} P(B)P(e)P(a|B, e)P(j|a)P(m|a)$$

$$= \sum_e P(B)P(e) \sum_a P(a|B, e)P(j|a)P(m|a)$$

$$= \sum_e P(B)P(e)f_1(j, m|B, e)$$

$$= P(B) \sum_e P(e)f_1(j, m|B, e)$$

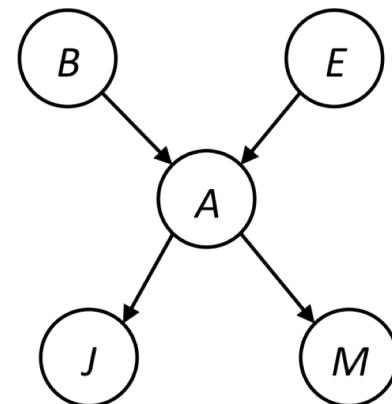
$$= P(B)f_2^e(j, m|B)$$

marginal can be obtained from joint by summing out

use Bayes' net joint distribution expression

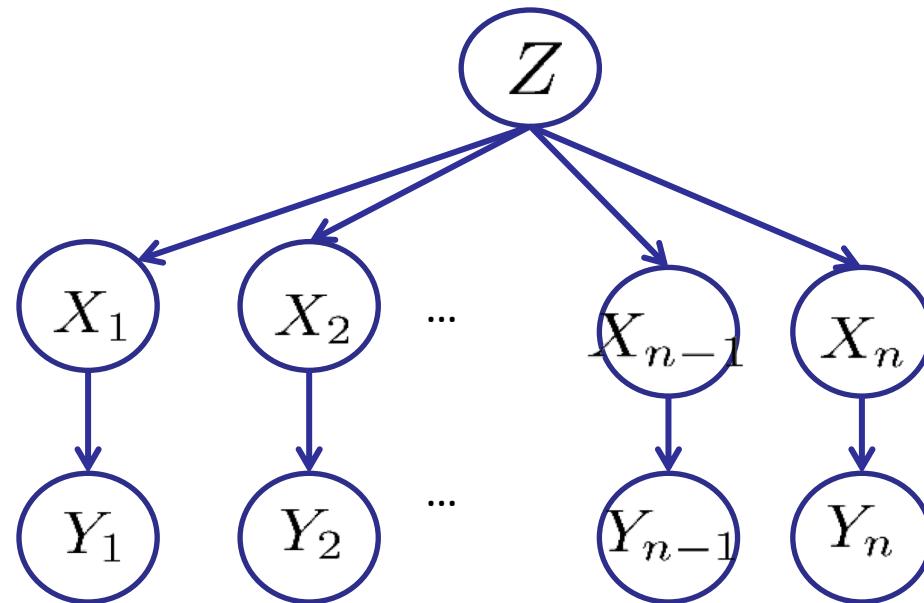
joining on a, and then summing out gives  $f_1$

joining on e, and then summing out gives  $f_2$



# Variable Elimination Ordering

- For the query  $P(X_n | y_1, \dots, y_n)$  work through the following two different orderings as done in previous slide:  $Z, X_1, \dots, X_{n-1}$  and  $X_1, \dots, X_{n-1}, Z$ . What is the size of the maximum factor generated for each of the orderings?



- Answer:  $2^n$  versus 2 (assuming binary)
- In general: the ordering can greatly affect efficiency.

# VE: Computational and Space Complexity

---

- The computational and space complexity of variable elimination is determined by the largest factor
- The elimination ordering can greatly affect the size of the largest factor.
  - E.g., previous slide's example  $2^n$  vs. 2
- Does there always exist an ordering that only results in small factors?
  - No!

# Variable Elimination

- Interleave joining and marginalizing
- $d^k$  entries computed for a factor over  $k$  variables with domain sizes  $d$
- Ordering of elimination of hidden variables can affect size of factors generated
- Worst case: running time exponential in the size of the Bayes' net

