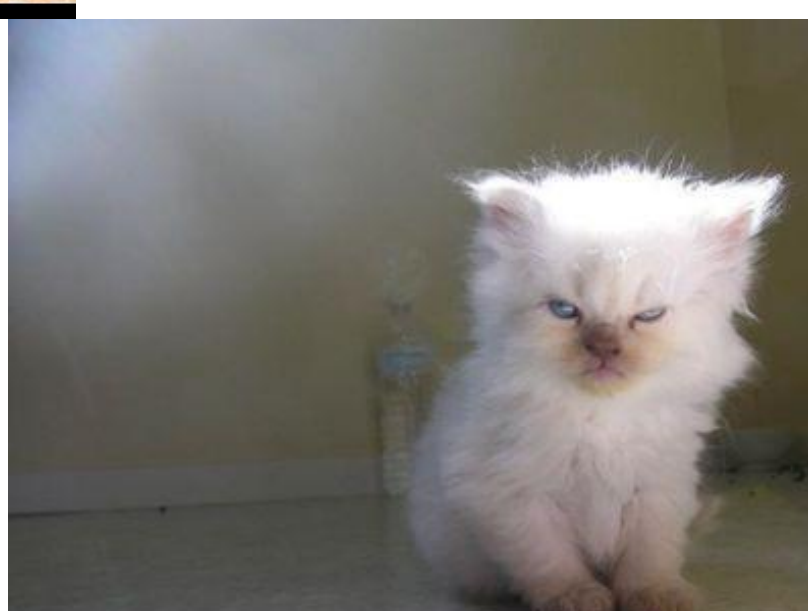
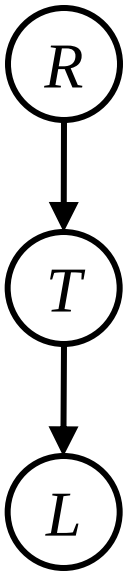
# Last Time: Factor Zoo





# Last Time: VE




$P(R)$

| | |
|---|---|
| +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 |

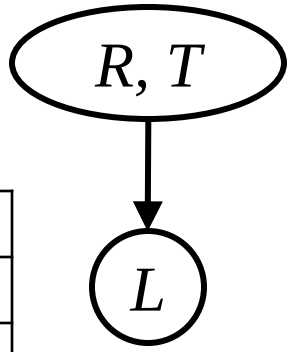
Join R

$P(R,T)$

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

$P(L|T)$

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



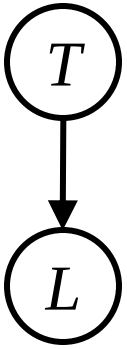

Sum out R

$P(T)$

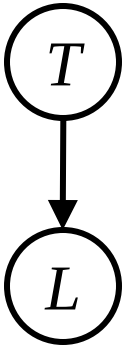
| | |
|---|---|
| +t | 0.17 |
| -t | 0.83 |

$P(L|T)$

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



# Last Time: VE




Join T


Sum out T


$P(T)$

| | |
|---|---|
| +t | 0.17 |
| -t | 0.83 |

$P(L|T)$

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

$P(T, L)$

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

$P(L)$

| | |
|---|---|
| +l | 0.134 |
| -l | 0.886 |

# 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'd 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!

# General Variable Elimination

- Query: $P(Q|E_1 = e_1, \ldots E_k = e_k)$

- 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

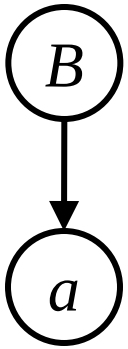
- Join all remaining factors and normalize

# Variable Elimination Bayes Rule

## Start / Select

$P(B)$

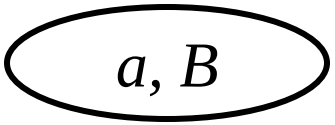
| B | P |
|---|---|
| +b | 0.1 |
| ¬b | 0.9 |




$P(A|B) \rightarrow P(a|B)$

| B | A | P |
|---|---|---|
| +b | +a | 0.8 |
| ~~+b~~ | ~~¬a~~ | ~~0.2~~ |
| ¬b | +a | 0.1 |
| ~~¬b~~ | ~~¬a~~ | ~~0.9~~ |

## Join on B




$P(a, B)$

| A | B | P |
|---|---|---|
| +a | +b | 0.08 |
| +a | ¬b | 0.09 |

## Normalize

$P(B|a)$

| A | B | P |
|---|---|---|
| +a | +b | 8/17 |
| +a | ¬b | 9/17 |

14.15) P(B | +j, +m) = ?

**Enumeration?**

*Sum over what I don't know*

**VE?**

*Join A*

*Eliminate A*

*Joint E*

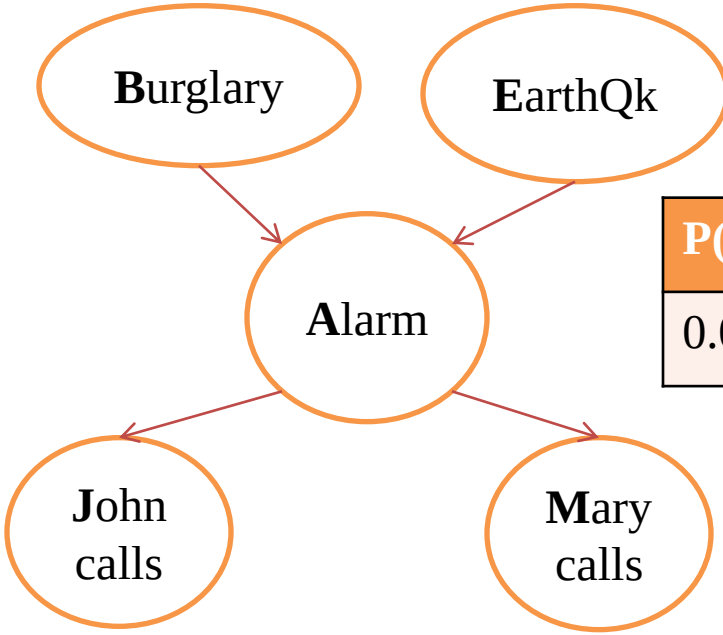
*Eliminate E*

*Join B*

*Normalize*

| P(B) |
|---|
| 0.001 |

| B | E | P(A\|B,E) |
|---|---|---|
| +b | +e | 0.95 |
| +b | ¬e | 0.94 |
| ¬b | +e | 0.29 |
| ¬b | ¬e | 0.001 |




| P(E) |
|---|
| 0.002 |

| A | P(J\|A) |
|---|---|
| +a | 0.9 |
| ¬a | 0.05 |

| A | P(M\|A) |
|---|---|
| +a | 0.7 |
| ¬a | 0.01 |

# 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

$$\begin{matrix} P(A|B,E) \\ P(j|A) \\ P(m|A) \end{matrix} \xrightarrow{\times} P(j,m,A|B,E) \xrightarrow{\Sigma} P(j,m|B,E)$$

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

# Example

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

Choose E

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

$$\xrightarrow{\times} \quad P(j,m,E|B) \quad \xrightarrow{\Sigma} \quad P(j,m|B)$$

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

Finish with B

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

$$\xrightarrow{\times} \quad P(j,m,B) \quad \xrightarrow{\text{Normalize}} \quad P(B|j,m)$$

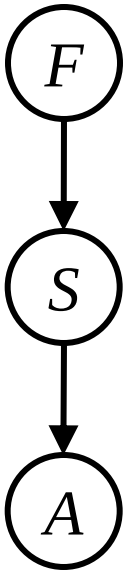
# Variable Elimination

- What you need to know:
  - Should be able to run it on small examples, understand the factor creation / reduction flow
  - Better than enumeration: saves time by marginalizing variables as soon as possible rather than at the end

- We will see special cases of VE later
  - On tree-structured graphs, variable elimination is poly. time
  - Implement a tree-structured special case to track invisible ghosts

# Approximate Inference

- Simulation has a name: sampling
- Sampling is a hot topic in machine learning, and it's really simple
- Basic idea:
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability
  - Show this converges to the true probability P
- Why sample?
  - Learning: get samples from a distribution you don't know
  - Inference: getting a sample is faster than computing the right answer (e.g., with variable elimination)



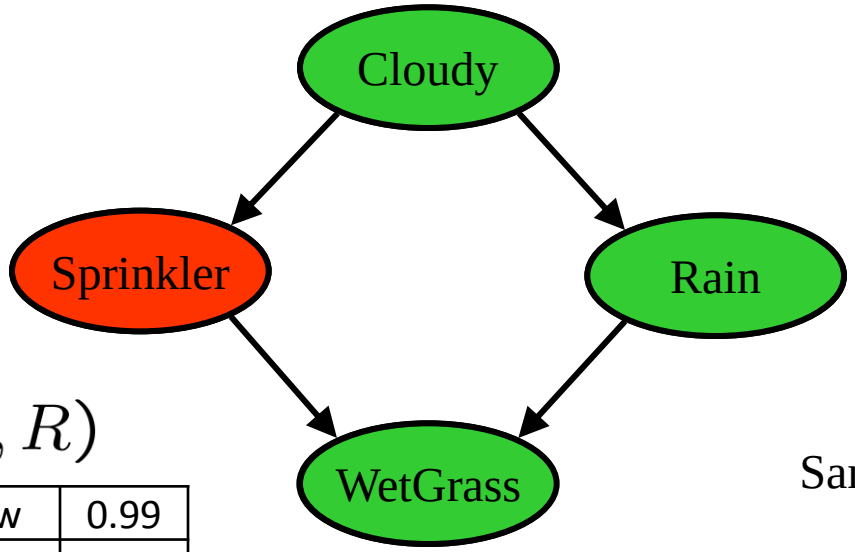
# Prior Sampling

$P(C)$

| | |
|---|---|
| +c | 0.5 |
| -c | 0.5 |

$P(S|C)$

| | | |
|---|---|---|
| +c | +s | 0.1 |
| | -s | 0.9 |
| -c | +s | 0.5 |
| | -s | 0.5 |




$P(R|C)$

| | | |
|---|---|---|
| +c | +r | 0.8 |
| | -r | 0.2 |
| -c | +r | 0.2 |
| | -r | 0.8 |

$P(W|S,R)$

| | | | |
|---|---|---|---|
| +s | +r | +w | 0.99 |
| | | -w | 0.01 |
| | -r | +w | 0.90 |
| | | -w | 0.10 |
| -s | +r | +w | 0.90 |
| | | -w | 0.10 |
| | -r | +w | 0.01 |
| | | -w | 0.99 |

Samples:

+c, -s, +r, +w

-c, +s, -r, +w

…

# Prior Sampling

- This process generates samples with probability:

$$S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i | \mathsf{Parents}(X_i)) = P(x_1 \ldots x_n)$$

…i.e. the BN's joint probability

- Let the number of samples of an event be $N_{PS}(x_1 \ldots x_n)$

- Then

$$\begin{aligned} \lim_{N \to \infty} \hat{P}(x_1, \ldots, x_n) &= \lim_{N \to \infty} N_{PS}(x_1, \ldots, x_n)/N \\ &= S_{PS}(x_1, \ldots, x_n) \\ &= P(x_1 \ldots x_n) \end{aligned}$$

- i.e., the sampling procedure is consistent

# Example

- We'll get a bunch of samples from the BN:

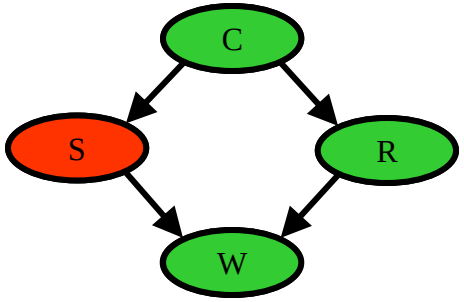
  +c, -s, +r, +w

  +c, +s, +r, +w

  -c, +s, +r, -w

  +c, -s, +r, +w

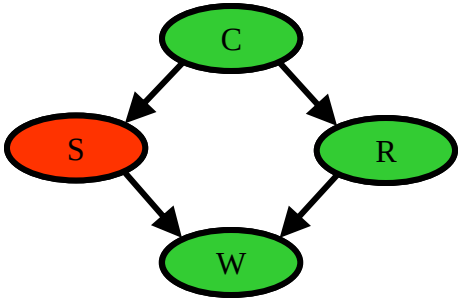
  -c, -s, -r, +w




- If we want to know P(W)
  - We have counts <+w:4, -w:1>
  - Normalize to get P(W) = <+w:0.8, -w:0.2>
  - This will get closer to the true distribution with more samples
  - Can estimate anything else, too
  - What about P(C| +w)? P(C| +r, +w)? P(C| -r, -w)?
  - Fast: can use fewer samples if less time (what's the drawback?)

# Rejection Sampling

- Let's say we want P(C)
  - No point keeping all samples around
  - Just tally counts of C as we go




- Let's say we want P(C| +s)
  - Same thing: tally C outcomes, but ignore (reject) samples which don't have S=+s
  - This is called rejection sampling
  - It is also consistent for conditional probabilities (i.e., correct in the limit)

+c, -s, +r, +w
+c, +s, +r, +w
-c, +s, +r, -w
+c, -s, +r, +w
-c, -s, -r, +w