Bayes Nets

With slides from Dan Klein and Stuart Russell

Today

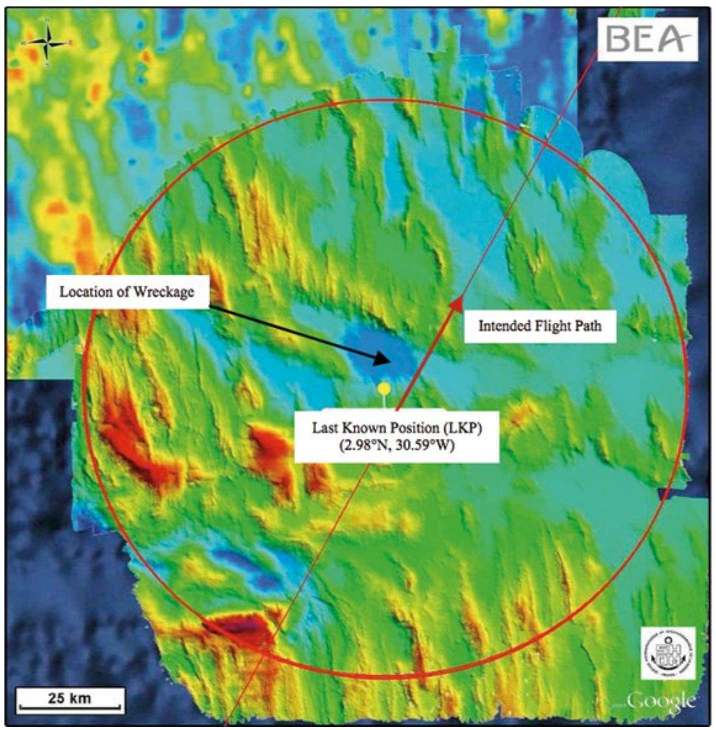
- Bayes Nets
  - BN Inference: Enumeration vs. VE

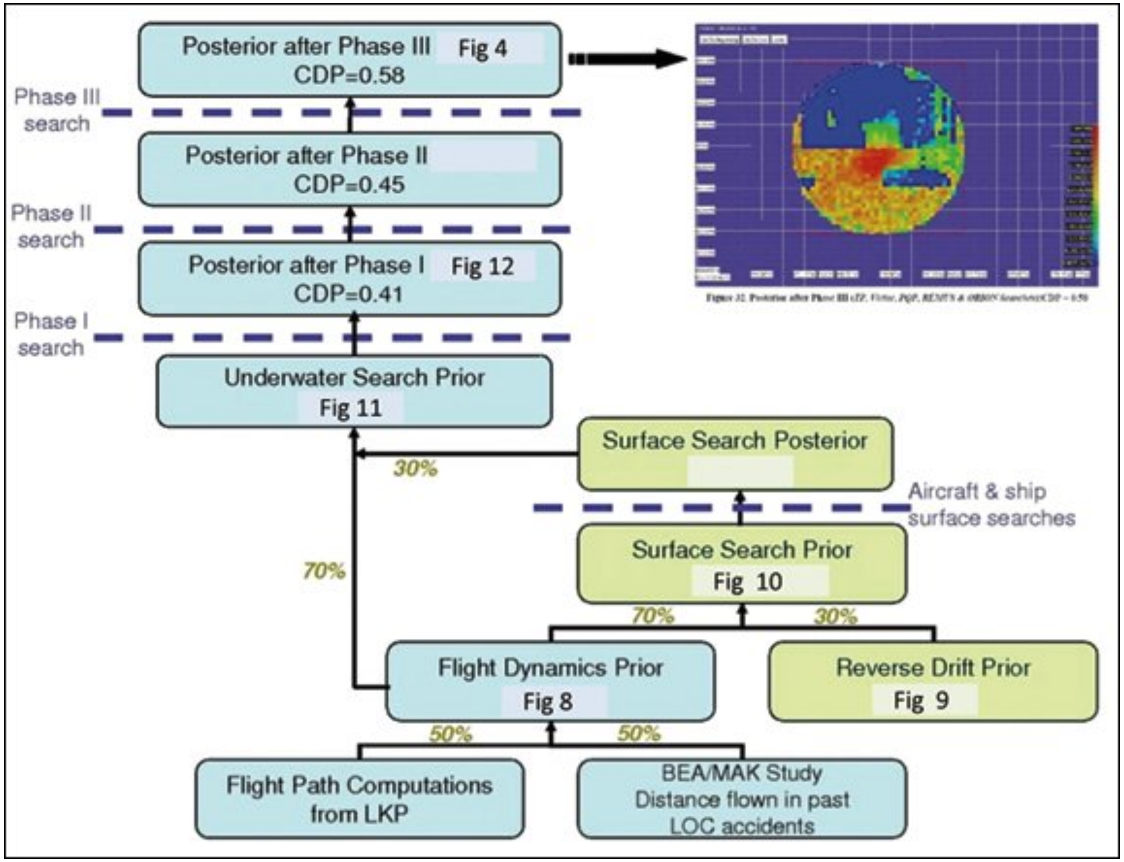
- BN Inference with time
  - Markov Chains
  - Hidden Markov Models

- Project 3: Ghostbusters

BN for Finding Air France Flight 447

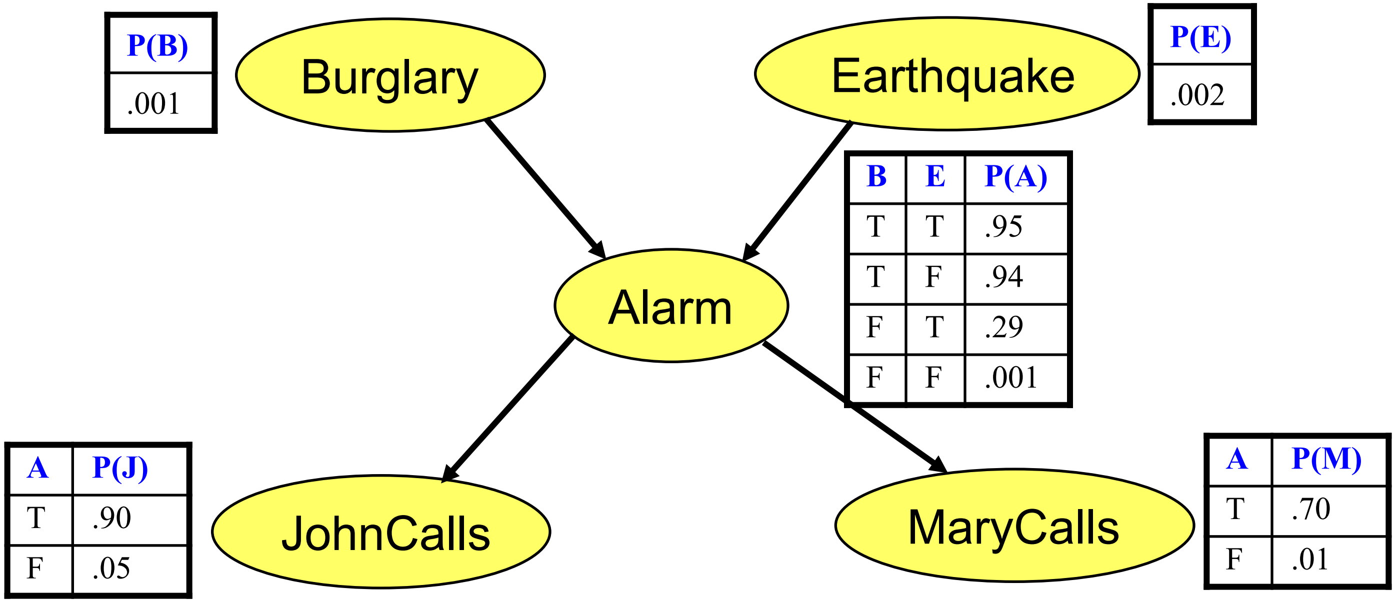
- http://fivethirtyeight.com/features/how-statisticians-could-help-find-flight-370/
- https://www.informs.org/ORMS-Today/Public-Articles/August-Volume-38-Number-4/In-Search-of-Air-France-Flight-447





Recap: BN Alarm Network

- Each node has a **conditional probability table** (**CPT**) that gives the probability of each of its values given every possible combination of values for its parents (conditioning case).
  - Roots (sources) of the DAG that have no parents are given prior probabilities.



P(B)
.001

P(E)
.002

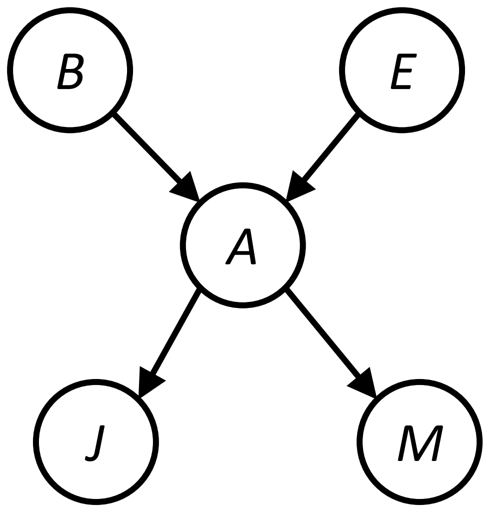
B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

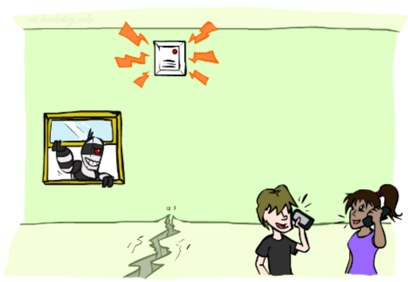
A	P(M)
T	.70
F	.01

Inference Example

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

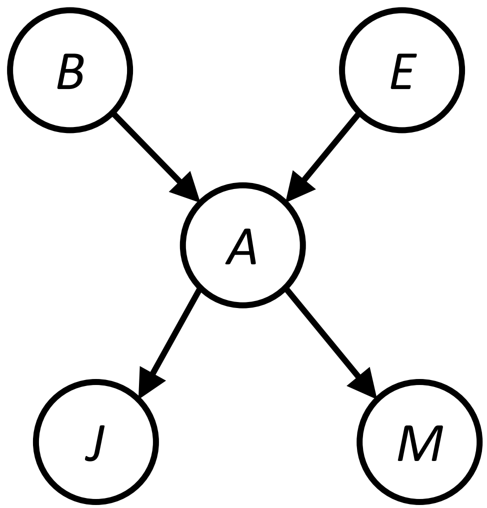
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

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

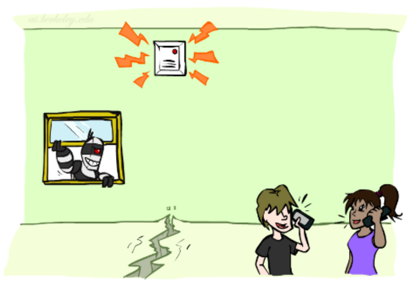
$$P(+b, -e, +a, -j, +m) =$$

Inference Example (2)

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

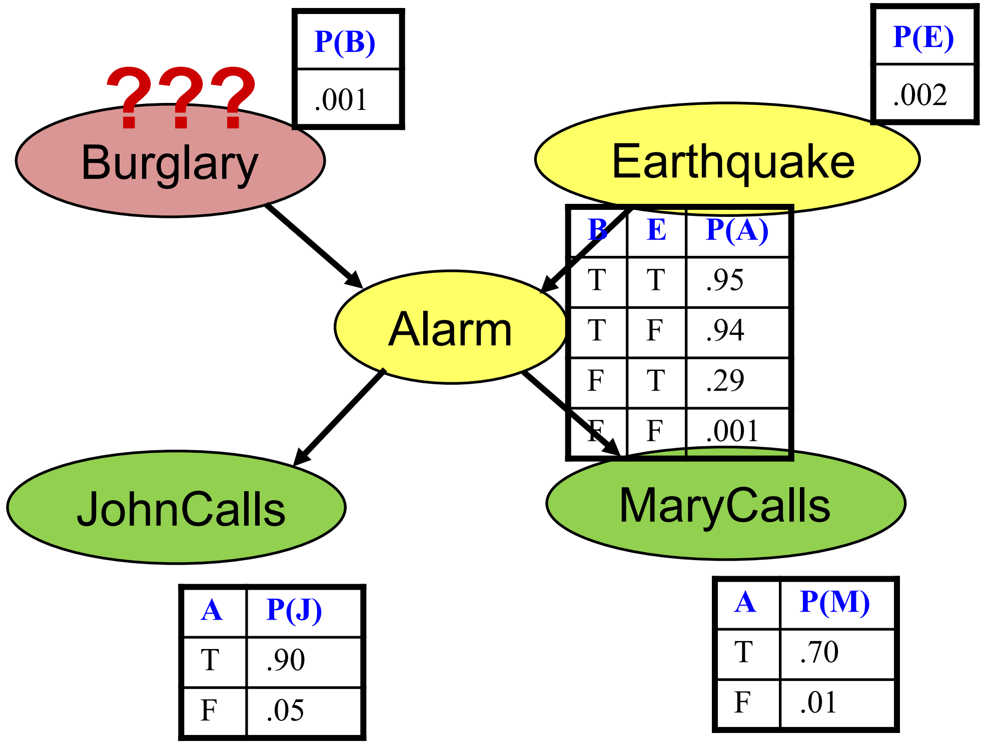
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

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

$$P(+b, -e, +a, -j, +m) =$$
$$P(+b)P(-e)P(+a| + b, -e)P(-j| + a)P(+m| + a) =$$
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Bayes Net Inference: Example

- Example: Given that John and Mary call (+j, +m), what is the probability that there is a Burglary (+b)?



P(B)
.001

P(E)
.002

B	E	P(A)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

$$P(+b \mid +j, +m) = ?$$

Queries: Inference by Enumeration

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

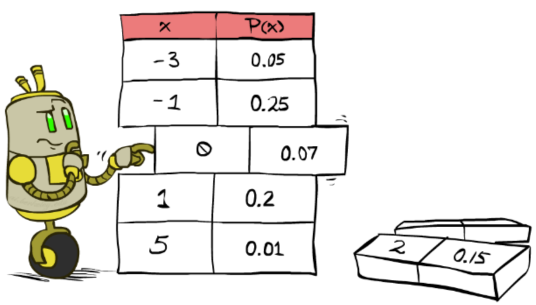
$\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\}$ $X_1, X_2, \ldots X_n$ *All variables*

- We want:

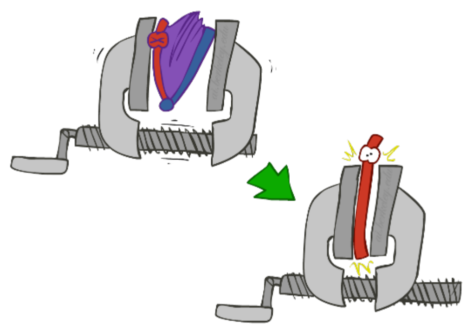
** Works fine with multiple query variables, too*

$$\boxed{P(Q|e_1 \ldots e_k)}$$

- Step 1: Select the entries consistent with the evidence



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



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

- Step 3: Normalize

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

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

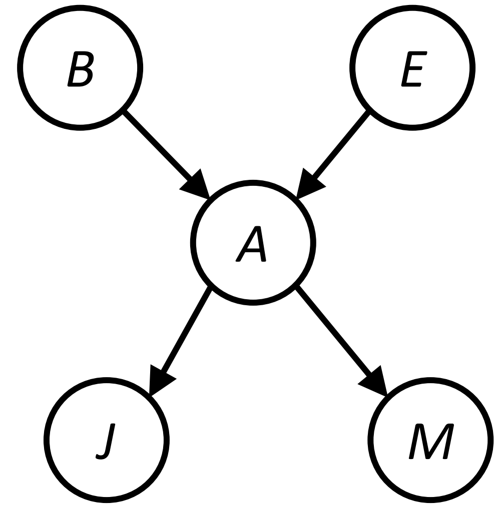
$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

Why? -- Bayes rule:

P(cause|effect) = P(effect|cause)p(cause)/p(effect) = p(cause,effect)/p(effect):

Inference by Enumeration: Example

- Given unlimited time, inference in BNs is easy
- Use inference by enumeration:



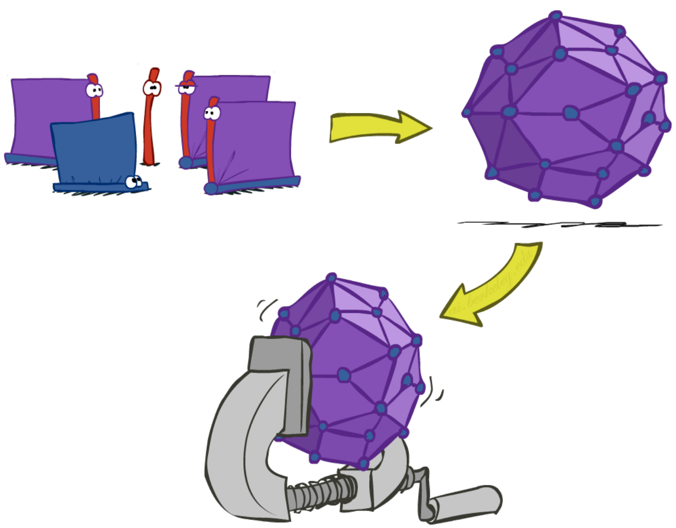
$$\boxed{P(B \mid +j, +m)} \propto_B P(B, +j, +m)$$
$$= \sum_{e,a} P(B, e, a, +j, +m)$$
$$= \sum_{e,a} P(B)P(e)P(a|B,e)P(+j|a)P(+m|a)$$

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

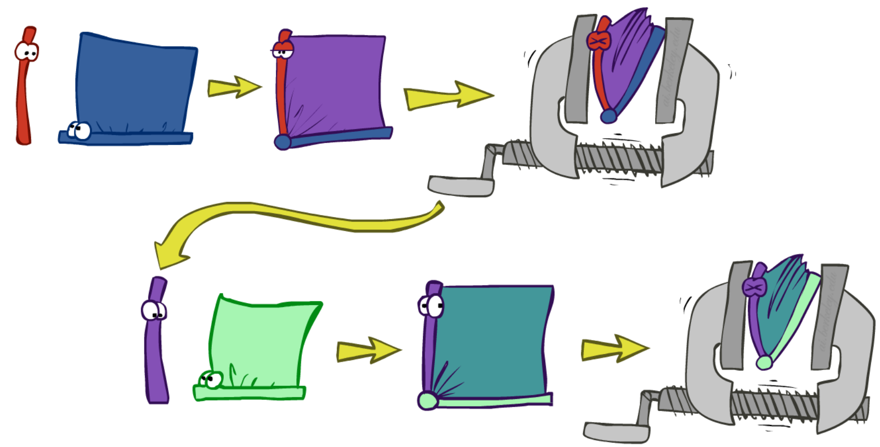
- Then normalize (divide) by prob(evidence), p(+j,+m)

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

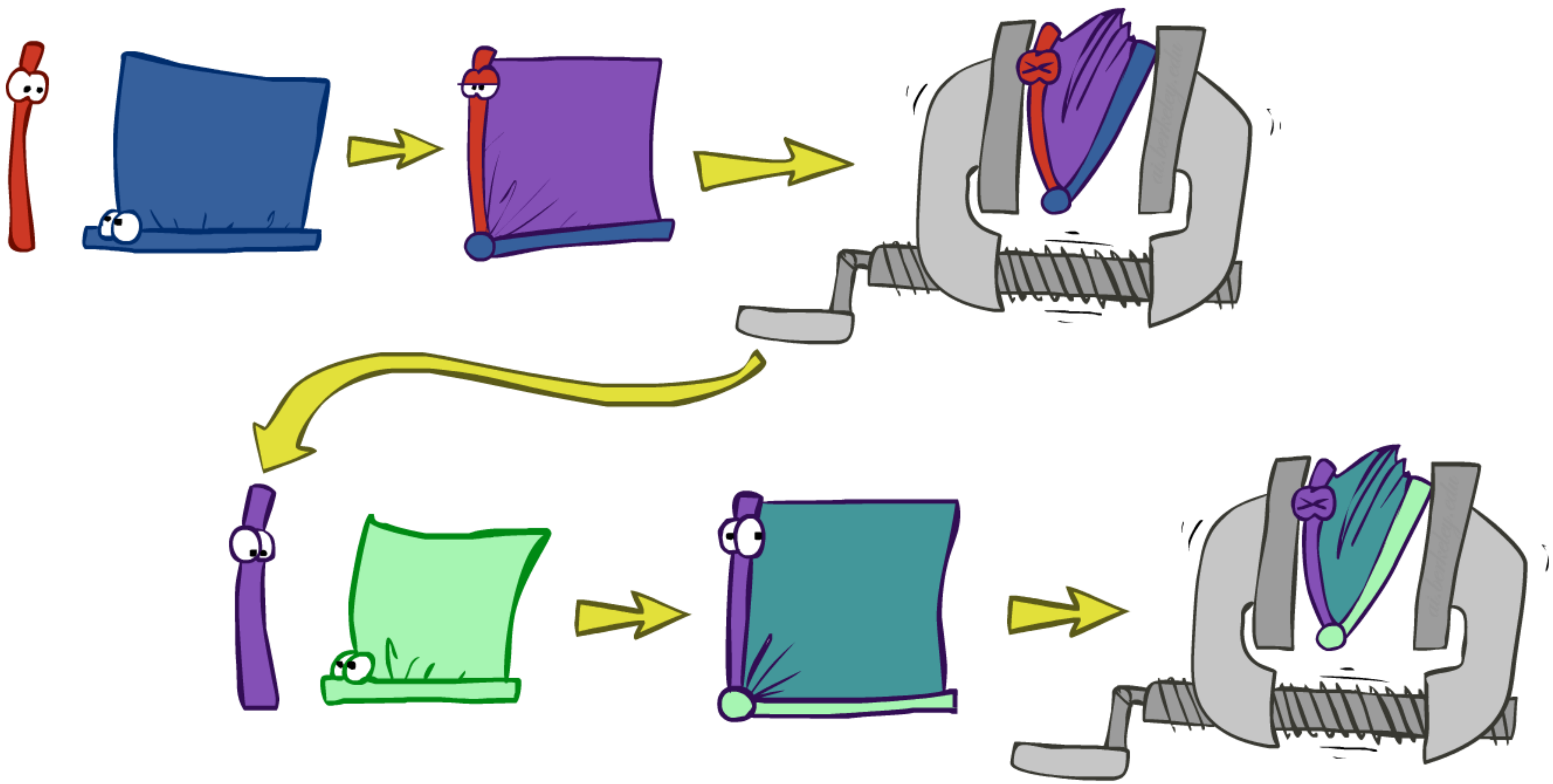


- Idea: interleave joining and marginalizing!
  - Called "Variable Elimination"
  - Still NP-hard, but usually much faster than inference by enumeration

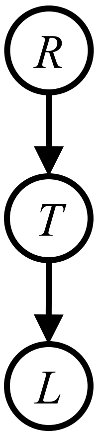


  - First we'll need some new notation: factors

Marginalizing Early (= Variable Elimination)



Marginalizing Early! (aka 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

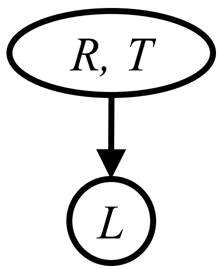
$P(L|T)$

+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



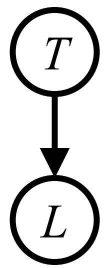
$P(L|T)$

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

Sum out R

$P(T)$

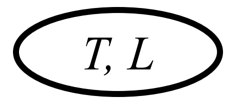
+t	0.17
-t	0.83



$P(L|T)$

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

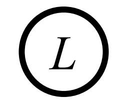
Join T



$P(T, L)$

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

Sum out T



$P(L)$

+l	0.134
-l	0.866

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:

$P(+r)$

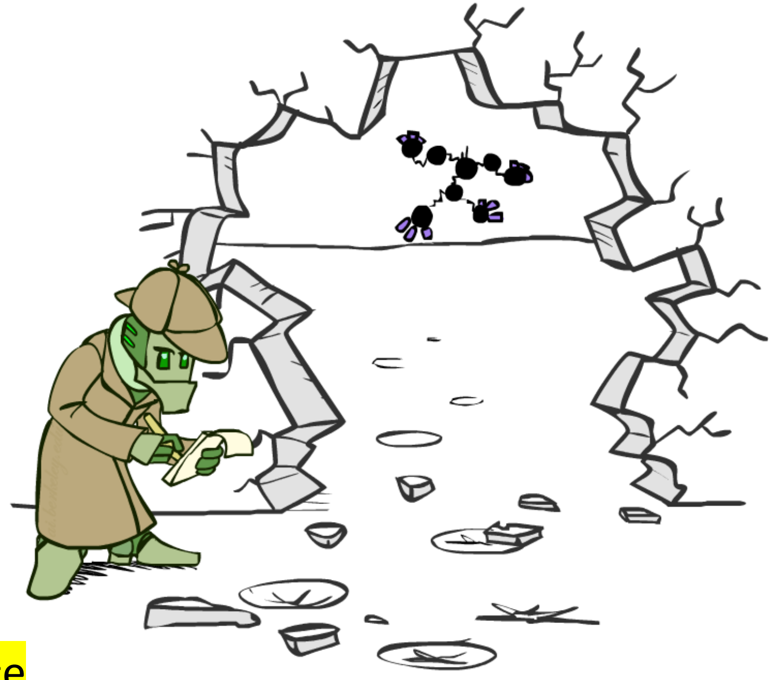
+r	0.1

$P(T|+r)$

+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



- Eliminate all variables 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:

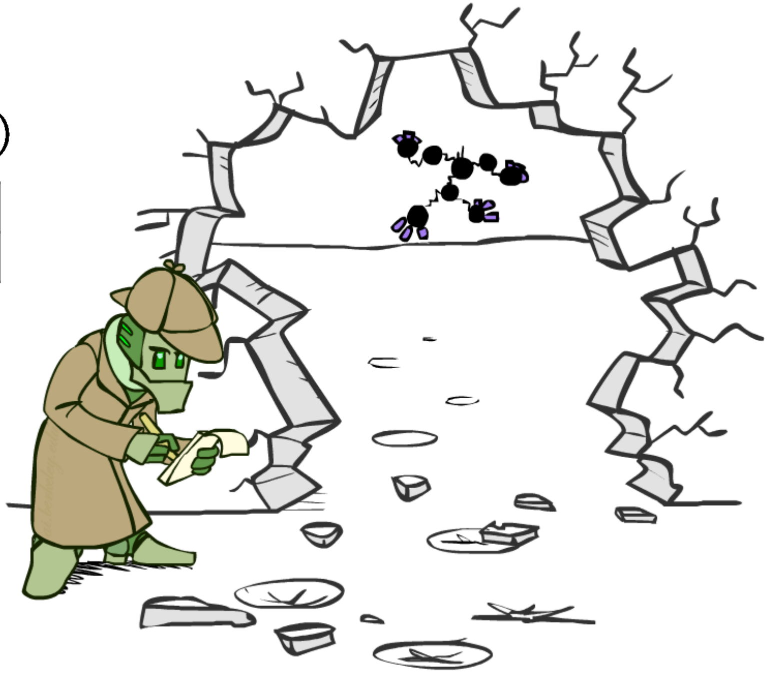
$P(+r, L)$

+r	+l	0.026
+r	-l	0.074

Normalize

$P(L|+r)$

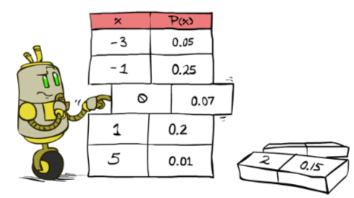
+l	0.26
-l	0.74



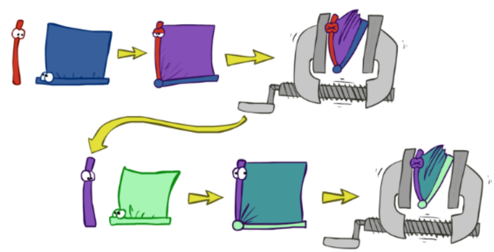
- To get our answer, just normalize this!
- That 's it!

General Variable Elimination

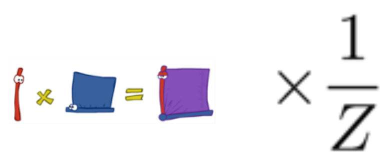
- Query: $P(Q|E_1 = e_1, \ldots E_k = e_k)$



- Start with initial factors:
  - Local CPTs (but instantiated by evidence)



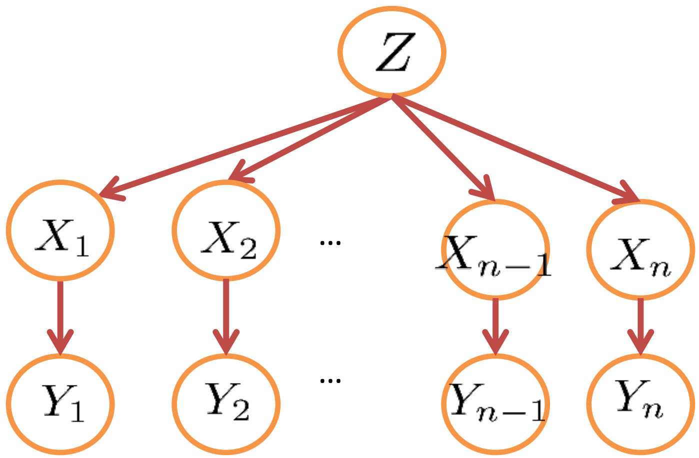
- While there are still hidden variables (not Q or evidence):
  - Pick a hidden variable H
  - Join all factors mentioning H
  - Eliminate (sum out) H



- Join all remaining factors and normalize

Variable Elimination Ordering

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



- Answer: 2^{n+1} versus 2^2 (assuming binary)
- In general: the ordering can greatly affect efficiency
- Remind you of anything from Search?

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.
  - Try to marginalize (eliminate) small factors first
- Does there always exist an ordering that only results in small factors?
  - No!

BN Quiz

- 5 minutes, can work in pairs

Reasoning over Time or Space

- Often, we want to reason about a sequence of observations
  - Speech recognition
  - Robot localization
  - User attention
  - Medical monitoring

- Need to introduce time (or space) into our models

Roadmap

- Markov Models

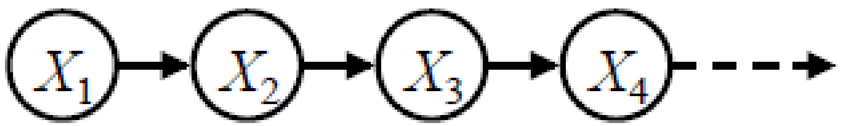
  (= a particular Bayes net)
- Hidden Markov Models (HMMs)
  - Representation

    (= another particular Bayes net)
  - Inference
    - Forward algorithm (= variable elimination)
    - Particle filtering (= likelihood weighting with some tweaks)
    - Viterbi (= variable elimination, but replace sum by max

      = graph search)
- Dynamic Bayes' Nets
  - Representation
    - (= yet another particular Bayes' net)
  - Inference: forward algorithm and particle filtering

Markov Models

- A Markov model is a chain-structured BN
  - Each node is identically distributed (stationarity)
  - Value of X at a given time is called the state
  - As a BN:

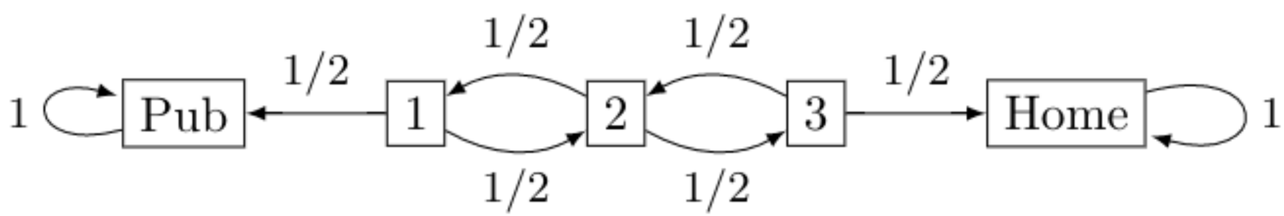


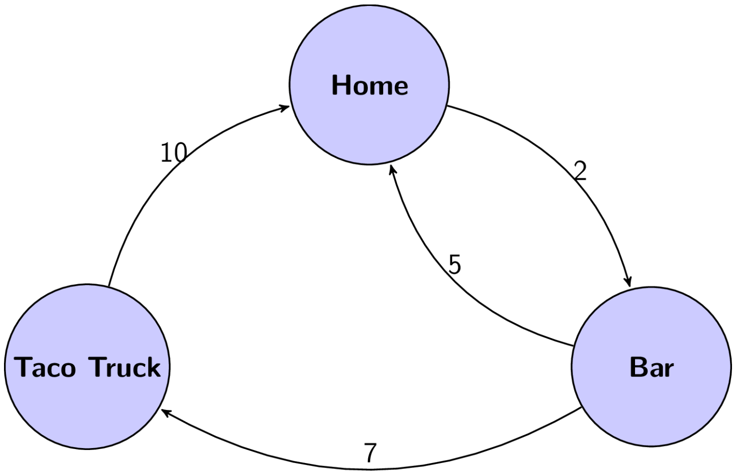
$$P(X_1) \qquad P(X_t|X_{t-1})$$

  - Parameters: called transition probabilities or dynamics, specify how the state evolves over time (also, initial state probabilities)
  - Same as MDP transition model, but no choice of action

What's "Markov" about it?

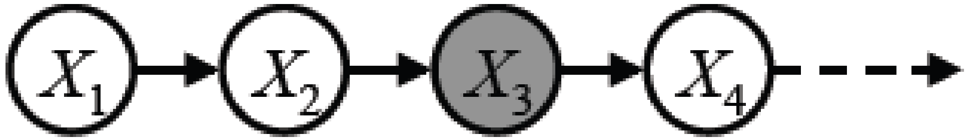
- Famous Markov model: Drunkard's walk





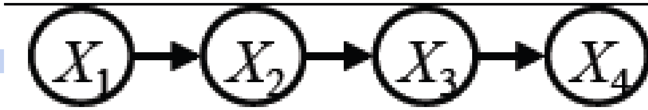
Andrey Markov

Conditional Independence in MMs



- Basic conditional independence:
  - Past and future independent of the present
  - Each time step only depends on the previous
  - This is called the (first order) Markov property
- Note that the chain is just a (growing) BN
  - We can always use generic BN reasoning on it if we truncate the chain at a fixed length

Example: P(X4)?



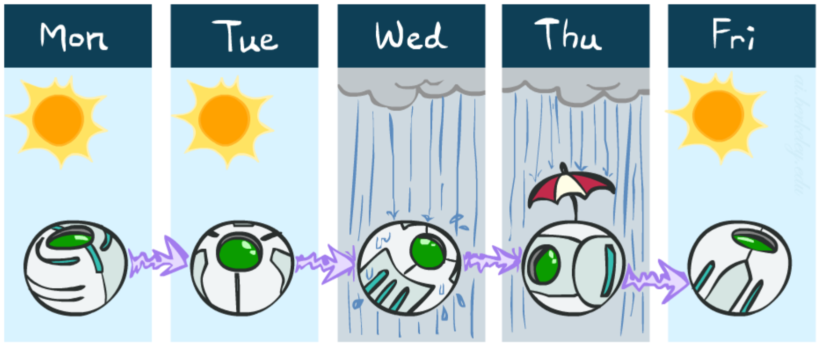
- Slow answer: inference by enumeration
  - Enumerate all sequences of length t which end in s
  - Add up their probabilities

$$P(X_4) = \sum_{x_1, x_2, x_3} P(x_1, x_2, x_3, X_4)$$

$$\begin{aligned}
&= P(X_1 = +x_1)P(X_2 = +x_2|X_1 = +x_1)P(X_3 = +x_3|X_2 = +x_2)P(X_4|X_3 = +x_3) \\
&+ P(X_1 = +x_1)P(X_2 = +x_2|X_1 = +x_1)P(X_3 = -x_3|X_2 = +x_2)P(X_4|X_3 = -x_3) \\
&+ P(X_1 = +x_1)P(X_2 = -x_2|X_1 = +x_1)P(X_3 = +x_3|X_2 = -x_2)P(X_4|X_3 = +x_3) \\
&+ P(X_1 = +x_1)P(X_2 = -x_2|X_1 = +x_1)P(X_3 = -x_3|X_2 = -x_2)P(X_4|X_3 = -x_3) \\
&+ P(X_1 = -x_1)P(X_2 = +x_2|X_1 = -x_1)P(X_3 = +x_3|X_2 = +x_2)P(X_4|X_3 = +x_3) \\
&+ P(X_1 = -x_1)P(X_2 = +x_2|X_1 = -x_1)P(X_3 = -x_3|X_2 = +x_2)P(X_4|X_3 = -x_3) \\
&+ P(X_1 = -x_1)P(X_2 = -x_2|X_1 = -x_1)P(X_3 = -x_3|X_2 = -x_2)P(X_4|X_3 = -x_3) \\
&+ P(X_1 = -x_1)P(X_2 = -x_2|X_1 = -x_1)P(X_3 = -x_3|X_2 = -x_2)P(X_4|X_3 = -x_3)
\end{aligned}$$

Example Markov Chain: Weather

- States: X = {rain, sun}

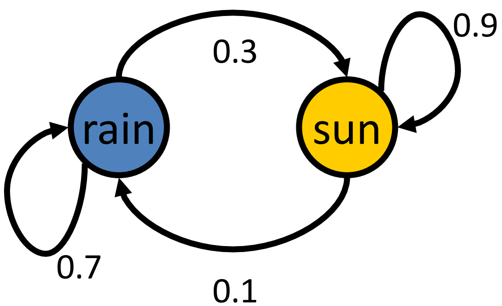


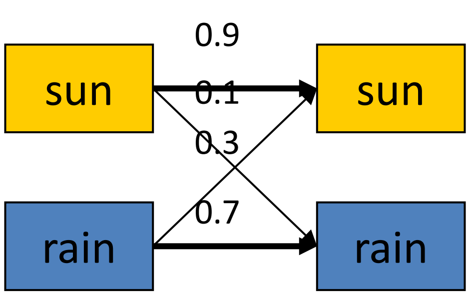
- Initial distribution: 1.0 sun

- CPT $P(X_t \mid X_{t-1})$:

X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	rain	0.7

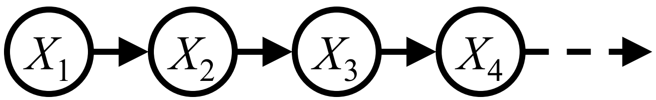
Two new ways of representing the same BN

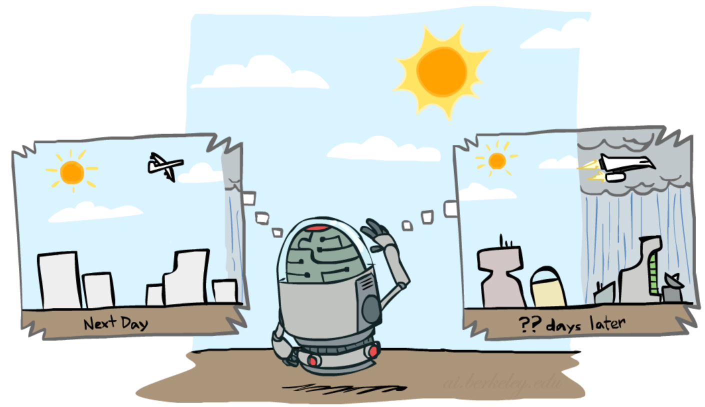




Mini-Forward Algorithm

- Question: What's P(X) on some day t?





$$P(x_1) = \text{known}$$

$$P(x_t) = \sum_{x_{t-1}} P(x_{t-1}, x_t)$$
$$= \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1})$$

Forward simulation

Example Run of Mini-Forward Algorithm

- From initial observation of sun

$$\left\langle \begin{matrix} 1.0 \\ 0.0 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.9 \\ 0.1 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.84 \\ 0.16 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.804 \\ 0.196 \end{matrix} \right\rangle \Rightarrow \left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle$$

$$\mathrm{P}(X_1) \quad \mathrm{P}(X_2) \quad \mathrm{P}(X_3) \quad \mathrm{P}(X_4) \quad \mathrm{P}(X_\infty)$$

- From initial observation of rain

$$\left\langle \begin{matrix} 0.0 \\ 1.0 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.3 \\ 0.7 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.48 \\ 0.52 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.588 \\ 0.412 \end{matrix} \right\rangle \Rightarrow \left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle$$

$$\mathrm{P}(X_1) \quad \mathrm{P}(X_2) \quad \mathrm{P}(X_3) \quad \mathrm{P}(X_4) \quad \mathrm{P}(X_\infty)$$

- From yet another initial distribution $\mathrm{P}(X_1)$:

$$\left\langle \begin{matrix} p \\ 1-p \end{matrix} \right\rangle \quad \ldots \quad \Rightarrow \left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle$$

$$\mathrm{P}(X_1) \qquad\qquad \mathrm{P}(X_\infty)$$

Stationary Distributions

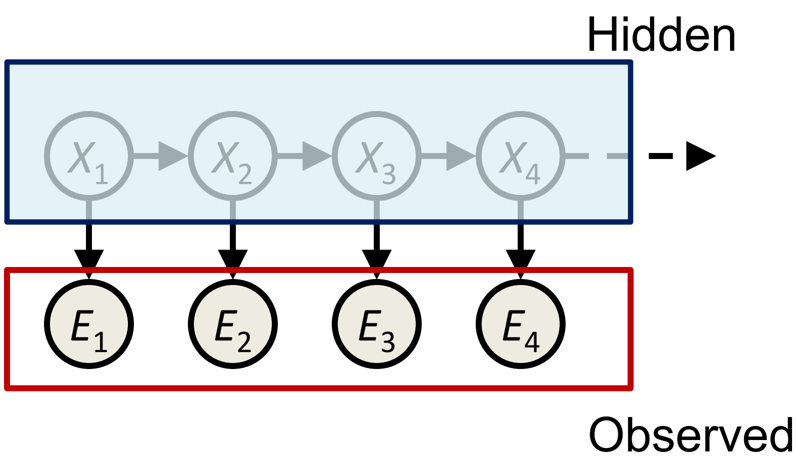
- For most chains:
  - influence of initial distribution gets less and less over time.
  - the distribution we end up in is independent of the initial distribution
- Stationary distribution:
  - Distribution we end up with is called the stationary distribution P_{∞} of the chain
  - It satisfies

$$P_{\infty}(X) = P_{\infty+1}(X) = \sum_{x} P_{t+1|t}(X|x) P_{\infty}(x)$$

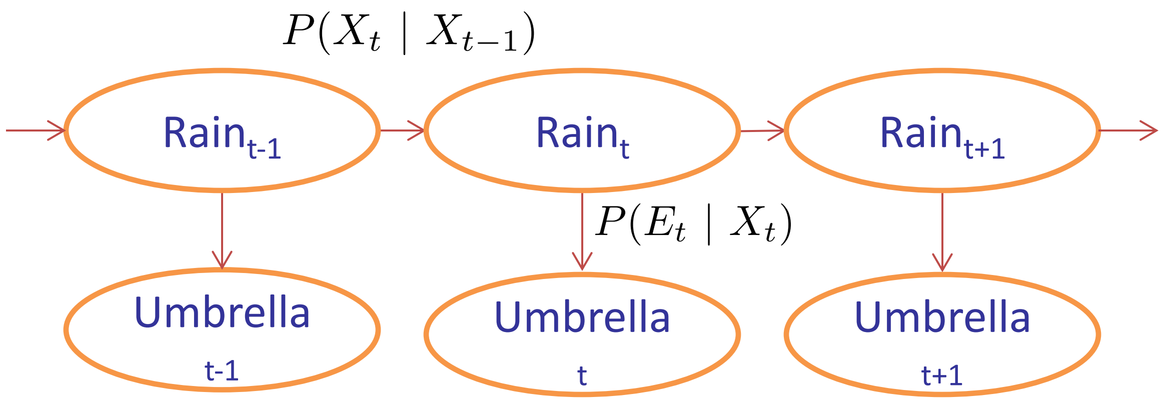
Hidden Markov Models

- Markov chains not so useful for most agents
  - Need observations to update your beliefs

- Hidden Markov models (HMMs)
  - Underlying Markov chain over states X
  - **You observe outputs (effects) at each time step**



Example: Weather HMM

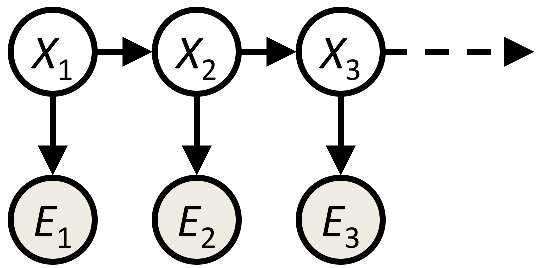


- An HMM is defined by:
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X_t \mid X_{t-1})$
  - **Emissions/observations:** $\boxed{P(E_t \mid X_t)}$

R_t	R_{t+1}	$P(R_{t+1} \mid R_t)$
+r	+r	0.7
+r	-r	0.3
-r	+r	0.3
-r	-r	0.7

R_t	U_t	$P(U_t \mid R_t)$
+r	+u	0.9
+r	-u	0.1
-r	+u	0.2
-r	-u	0.8

Joint Distribution of an HMM



– Joint distribution:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)P(X_3|X_2)P(E_3|X_3)$$

– More generally:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(E_t|X_t)$$

– Questions to be resolved:
- Does this indeed define a joint distribution?
- Can every joint distribution be factored this way, or are we making some assumptions about the joint distribution by using this factorization?

Filtering / Monitoring

- Filtering, or monitoring, is the task of tracking the distribution $B_t(X) = P_t(X_t \mid e_1, \ldots, e_t)$ (the belief state) over time
- We start with $B_1(X)$ in an initial setting, usually uniform
- As time passes, or we get observations, we update $B(X)$

Filtering: Weather HMM

Marginalize

B(+r)=B(+r)*P(+r|+r) + B(-r)*P(+r|-r)

B'(+r) = 0.5
B'(-r) = 0.5

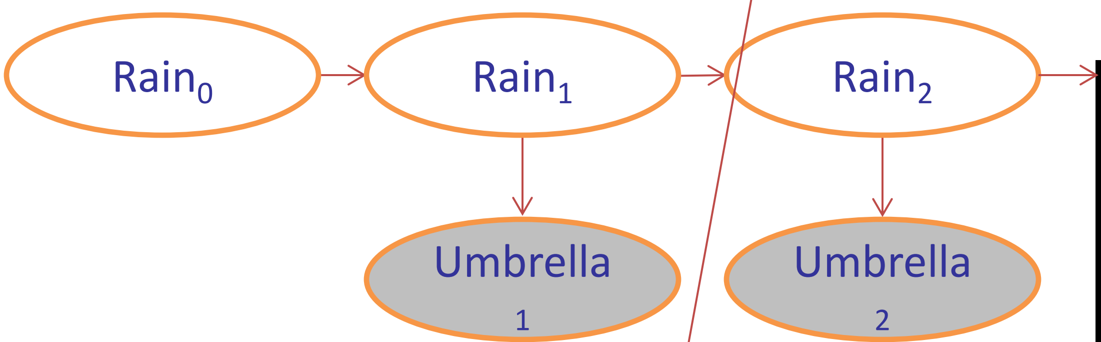
B'(+r) = 0.627
B'(-r) = 0.373

Observe

B(+r) = 0.5
B(-r) = 0.5

B(+r)=0.5*0.9
B(-r)=0.5*0.2

B(+r) = 0.883
B(-r) = 0.117



R_t	R_{t+1}	$P(R_{t+1}\|R_t)$
+r	+r	0.7
+r	-r	0.3
-r	+r	0.3
-r	-r	0.7

R_t	U_t	$P(U_t\|R_t)$
+r	+u	0.9
+r	-u	0.1
-r	+u	0.2
-r	-u	0.8

Normalize
B(+r)=0.818
B(-r)=0.182

Ghostbusters Filtering (project 3)

- Let’s say we have two distributions:
  - Prior distribution over ghost location: P(G)
    - Let’s say this is uniform
  - Sensor reading model: P(R | G)
    - Given: we know what our sensors do
    - R = reading color measured at (1,1)
    - E.g. P(R = yellow | G=(1,1)) = 0.1

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

0.17	0.10	0.10
0.09	0.17	0.10
<0.01	0.09	0.17

- Can calculate posterior distribution P(G|r) over ghost locations given a sensor reading, with Bayes’ rule:

$$P(g|r) \propto P(r|g)P(g)$$

Project 3: Ghostbusters!

- Due **Wed March 21st**

  http://www.mathcs.emory.edu/~eugene/cs425/p3/

- **Next week:**
  - HMMs: filtering, decoding (last topic on mid-term)
  - Approximate inference
  - [optional] Midterm review: Tuesday 5:30-6:30, room tbd
  - Midterm Exam (March 8th)