Independence

Two variables are **independent** if: $\forall x, y \, P(x, y) = P(x)P(y)$

We denote this as $X \perp\!\!\!\perp Y$

Conditional Independence

X is **conditionally independent** of Y given Z

if and only if: $\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$

or, equivalently, if and only if $\forall x, y, z : P(x|z, y) = P(x|z)$

$$X \perp\!\!\!\perp Y | Z$$

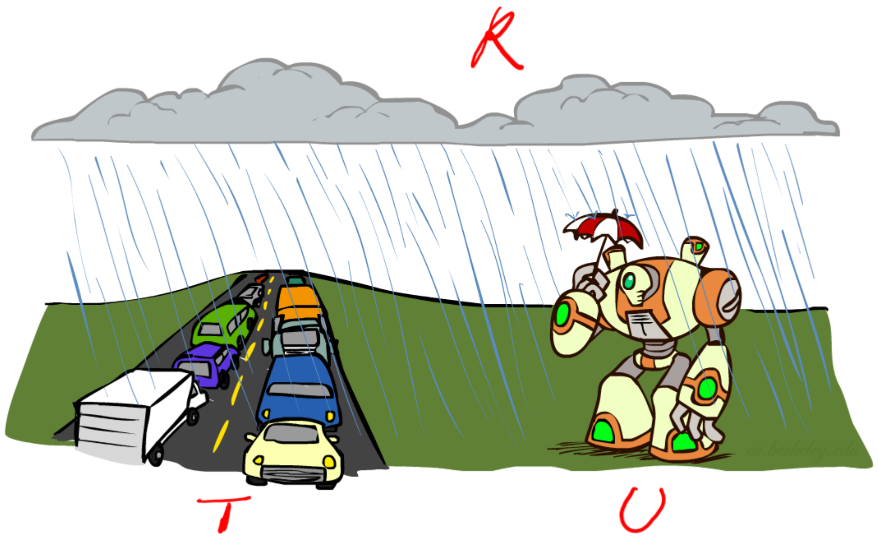
Conditional Independence

Traffic, Umbrella, Raining

$X \perp\!\!\!\perp Y \mid Z$

↑

Raining



$T \perp\!\!\!\perp U$?

$T \perp\!\!\!\perp U \mid R$

$\Updownarrow$

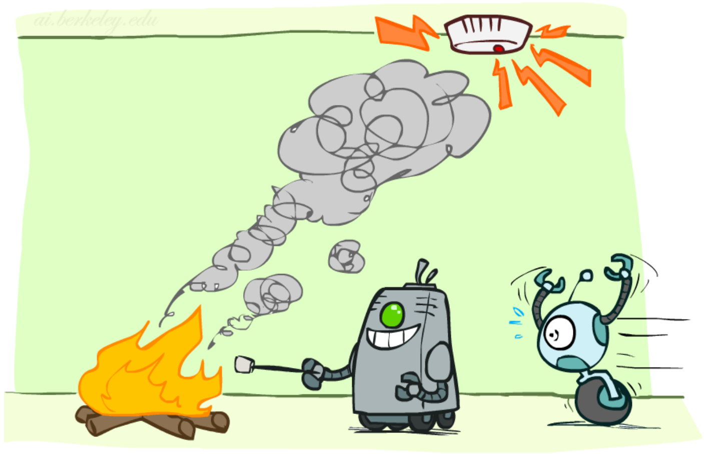
$P(T \mid R, U) = P(T \mid R)$

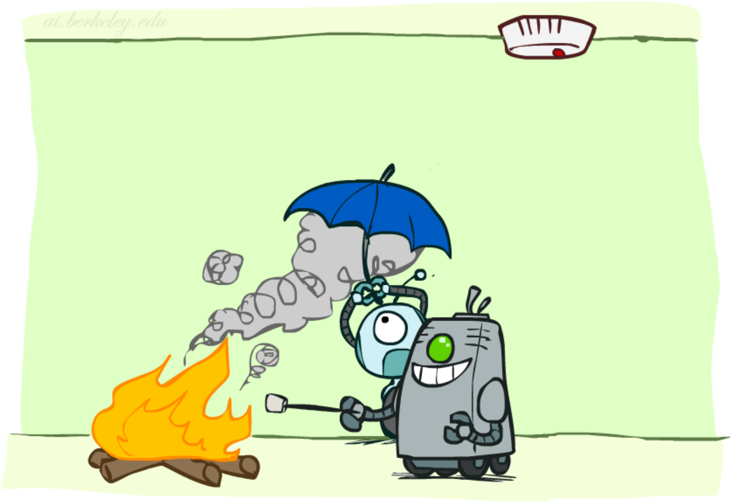
Conditional Independence

Fire, Smoke, Alarm (Smoke detector)

$X \perp\!\!\!\perp Y \mid Z$

$$P(\text{Alarm} \mid \text{smoke}) \overset{?}{=} P(\text{Alarm} \mid \text{smoke}, \text{Fire})$$

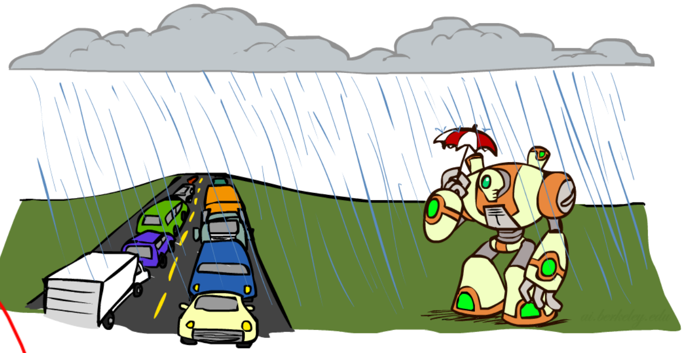




Independence vs. Conditional Independence

Rain
Traffic
Pedestrian holding umbrella
Flood in the house
Trip cancelled
…

$\Bigg\}$ **Dependent**



R

T U F C

P(Traffic | Rain, Umbrella) = P(Traffic | Rain) **Conditional Independent**

Conditional distribution / independence allows us to model the probability of a certain event only using relevant factors.

Bayesian Networks

Bayes Net

Bayesian Network Example

Traffic, Umbrella, Raining

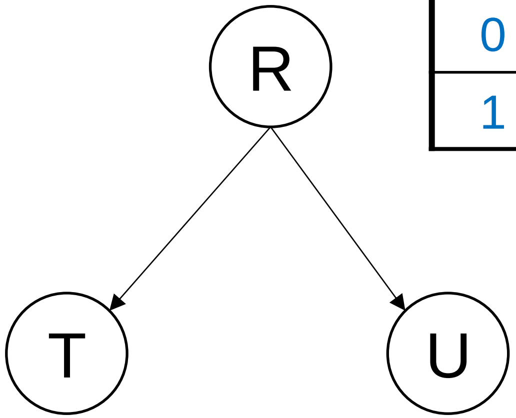
$P(t, u, r)$

$= P(r)\ P(t \mid r)\ P(u \mid r, t)$ (always hold by chain rule)

$= P(r)\ P(t \mid r)\ P(u \mid r)$

$T \perp\!\!\!\perp U \mid R$

R	P(R)
0	0.7
1	0.3



R	T	P(T|R)
0	0	0.5
0	1	0.5
1	0	0.2
1	1	0.8

R	U	P(U|R)
0	0	0.8
0	1	0.2
1	0	0.1
1	1	0.9

Bayesian Network (BN)

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - Suppose a node as m parents, and suppose each random variable can take d different values
  - What is the size of the table?
- The BN models the joint probability as

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

$Y_1 \in \{1, \ldots, d\}$

$Y_2 \in \{1, \ldots, d\}$

$X \in \{1, \ldots, d\}$

$\boxed{p(X \mid Y_1, Y_2, \ldots, Y_m)}$

X	Y_1	Y_2	...	Y_m
1	1	1	1	1
1	1	1	1	2
⋮				⋮
d	d	d	d	d

$\# \text{rows} = d^{m+1}$

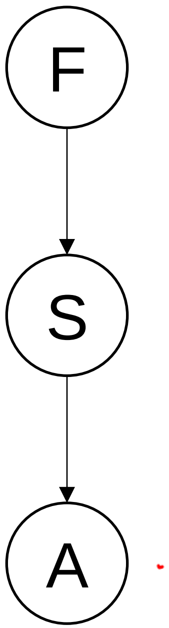
Bayesian Network Example

Fire, Smoke, Alarm

$P(f, s, a) = P(f)\,P(s \mid f)\,P(a \mid s)$ (by BN semantics)

Prove $F \perp\!\!\!\perp A \mid S$?

$p(f)\,p(s \mid f)\,p(a \mid s, f)$



10^{-6}

Bayesian Network Example

0.001 0.001

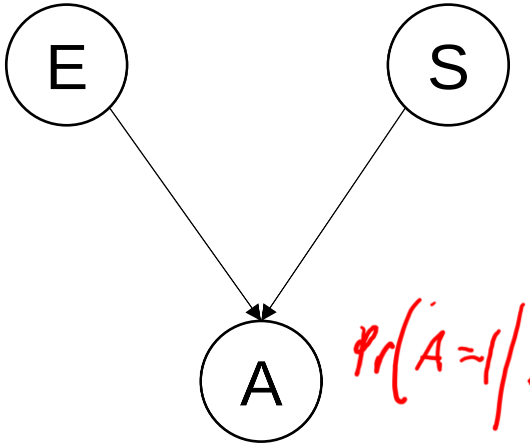
Earthquake, Smoke, Alarm

$P(e, s, a) = P(e)\,P(s)\,P(a \mid e, s)$

$E \perp\!\!\!\perp S$? Yes $E \perp\!\!\!\perp S \mid A$? No

E	p(E)
0	0.999
1	0.001

S	p(S)
0	0.999
1	0.001



$$\Pr(A=1 \mid E, S) = \begin{cases} 1, & \text{if } E=1 \text{ or } S=1 \\ 0, & \text{otherwise} \end{cases}$$

Pr(Earthquake | Alarm)	?	Pr(Earthquake | Alarm, Smoke)
$\approx 1/2$		≈ 0.001

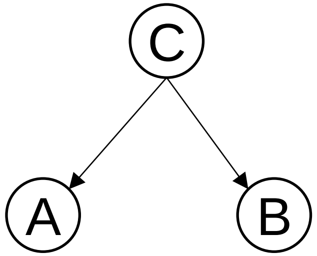
"Explain away"

Recap

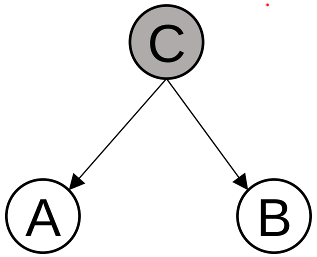
- Common cause

A and B are not independent *in general*

They could still be independent *in special cases*

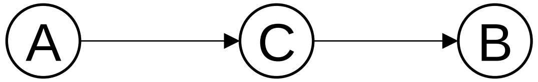


$A \cancel{\perp\!\!\!\perp} B$

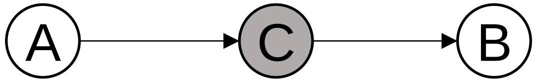


$A \perp\!\!\!\perp B \mid C$

- Causal chain

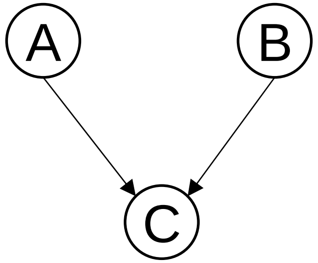


$A \cancel{\perp\!\!\!\perp} B$

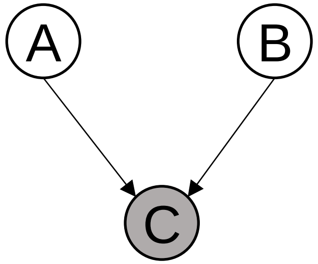


$A \perp\!\!\!\perp B \mid C$

- Common effect

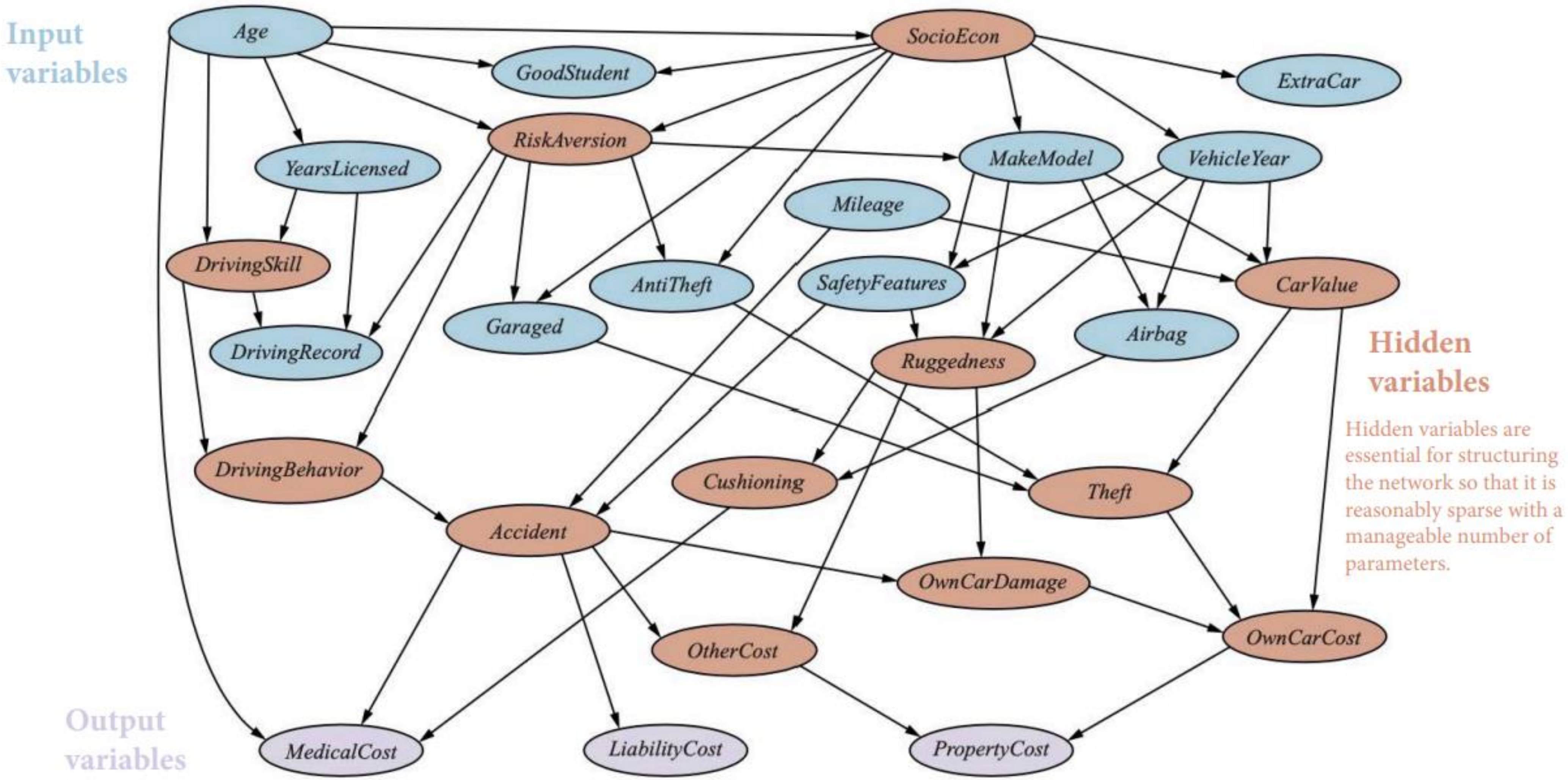


$A \perp\!\!\!\perp B$

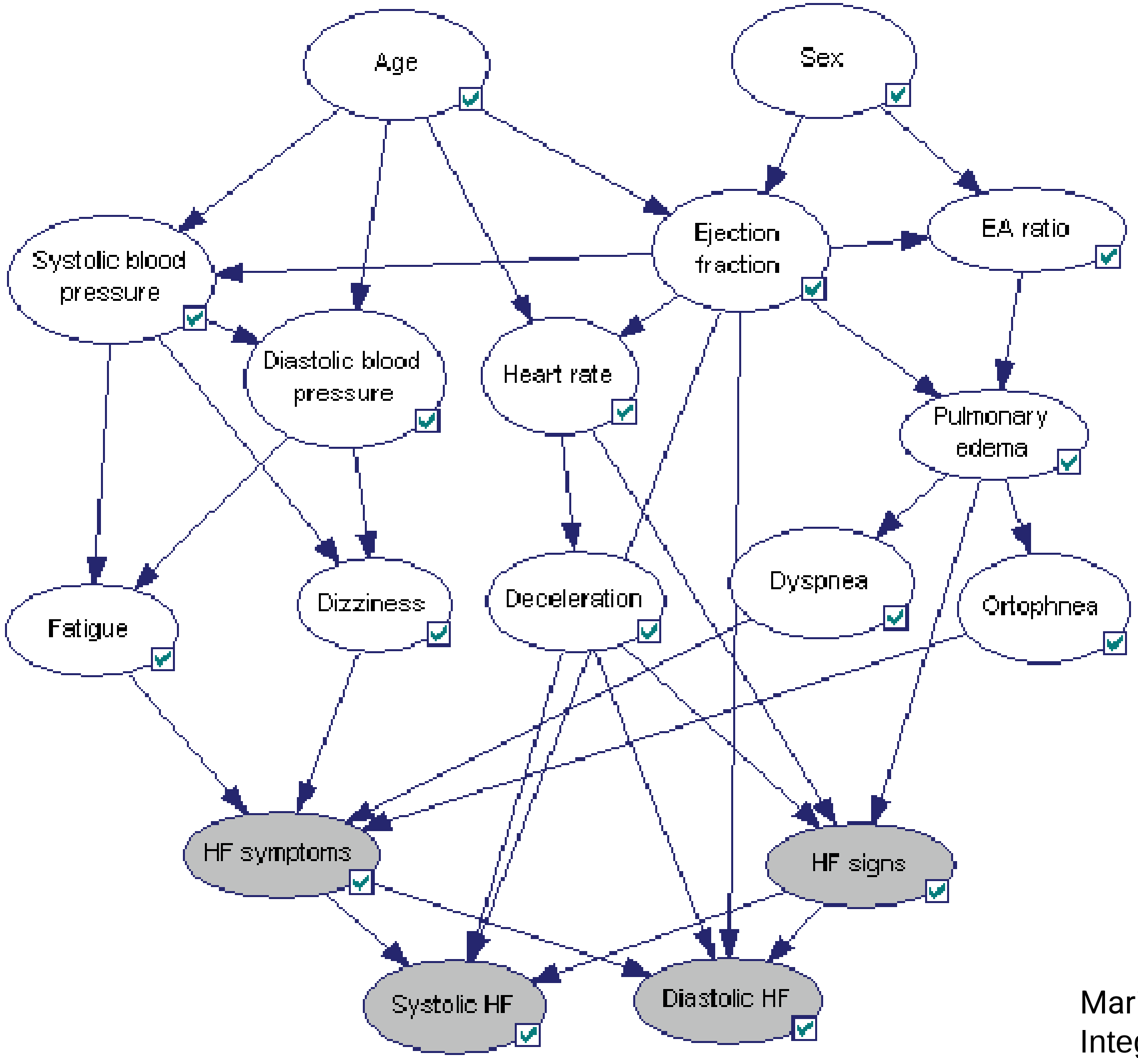


$A \cancel{\perp\!\!\!\perp} B \mid C$

Example: Car Insurance



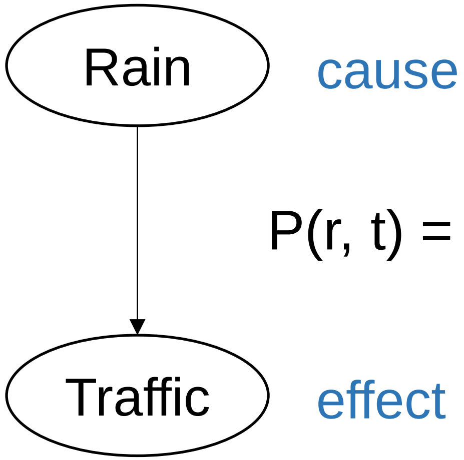
Example: Medical Diagnosis



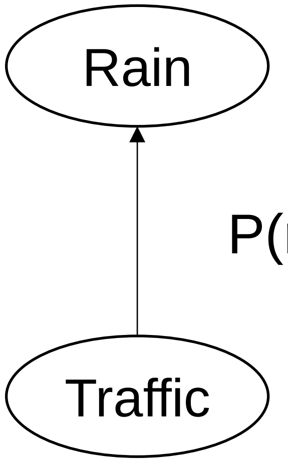
Marin Prcela et al. Information Gain of Structured Medical Diagnostic Tests - Integration of Bayesian Networks and Ontologies

Causality?

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents) and easier to think about
- BNs need not be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - Arrows that reflect correlation, but not necessary causality

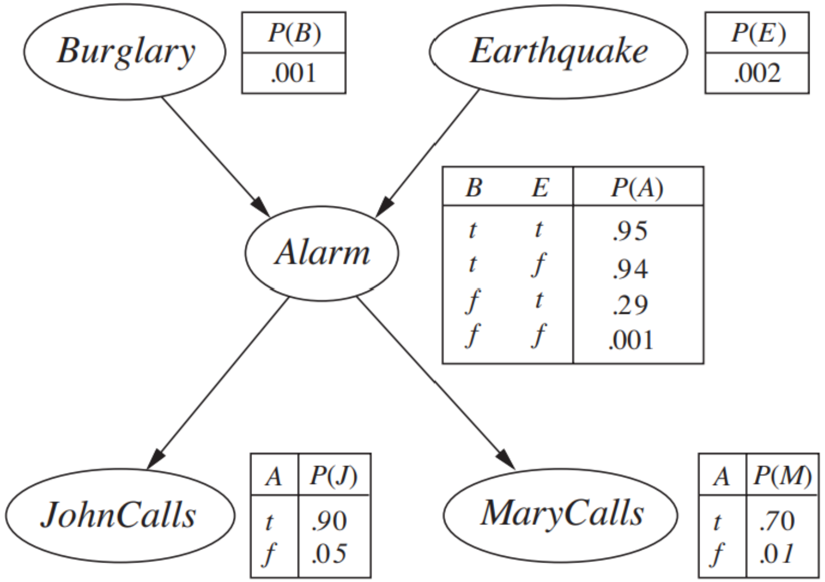


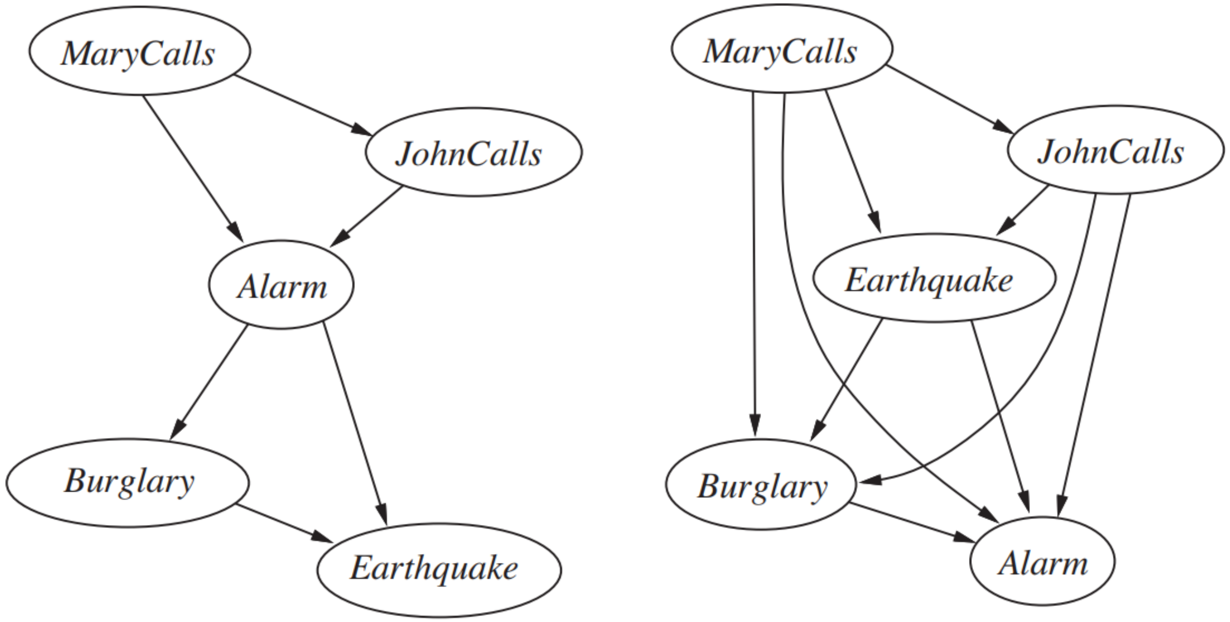
$P(r, t) = P(r)\, P(t \mid r)$



$P(r, t) = P(t)\, P(r \mid t)$

Causality?





(a) (b)

Independence Given Evidence

General question: Are two variables $X, \ Y$ independent of each other conditioned on $Z = \{Z_1, Z_2, \dots\}$?

Or: Are X and Y "D-separated" by Z?

Algorithm

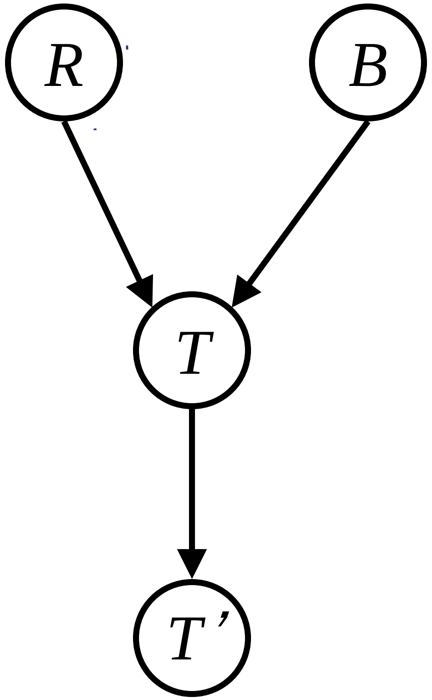
1. Consider just the **ancestral subgraph** consisting of X, Y, Z, and their ancestors.
2. Add links between any unlinked pair of nodes that share a common child; now we have the so-called **moral graph**.
3. Replace all directed links by undirected links.
4. If Z blocks all paths between X and Y in the resulting graph, then Z d-separates X and Y.

Example

$R \perp\!\!\!\perp B$ *Yes*

$R \perp\!\!\!\perp B | T$

$R \perp\!\!\!\perp B | T'$



Example

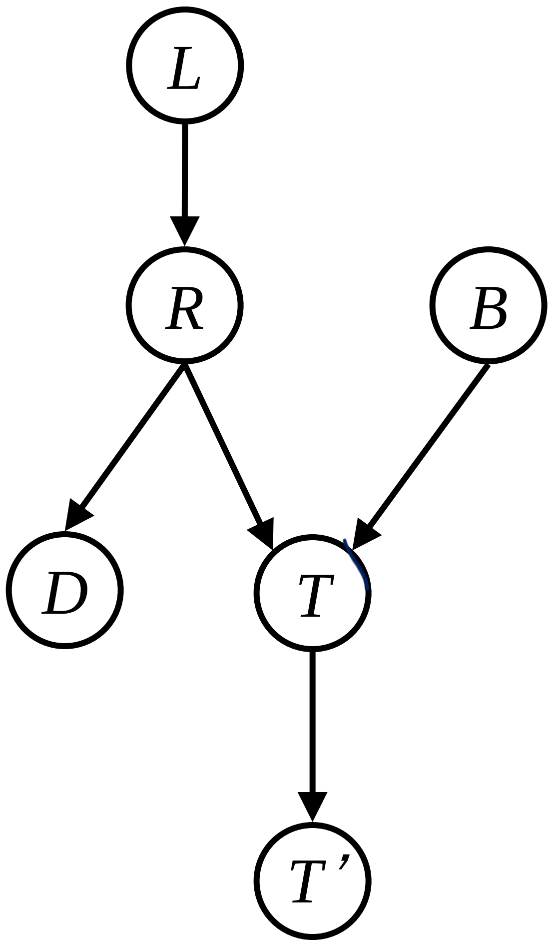
$L \perp\!\!\!\perp T' | T$ *Yes*

$L \perp\!\!\!\perp B$ *Yes*

$L \perp\!\!\!\perp B | T$

$L \perp\!\!\!\perp B | T'$

$L \perp\!\!\!\perp B | T, R$ *Yes*



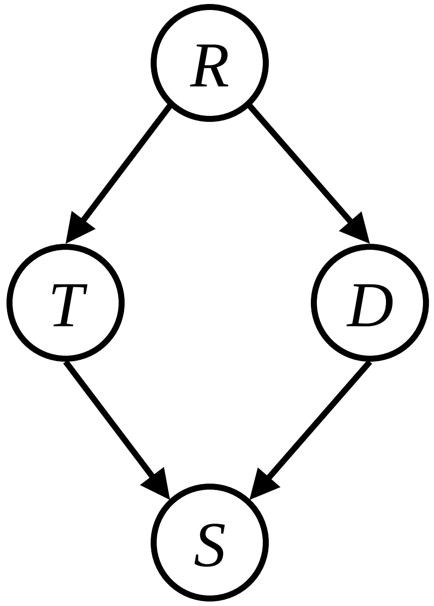
Example

- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I'm sad
- Questions:

$$T \perp\!\!\!\perp D$$

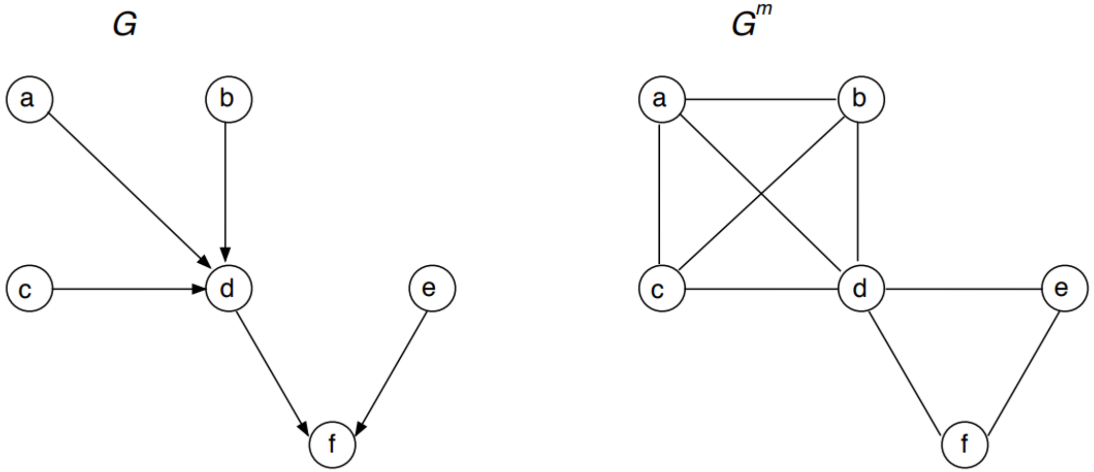
$$T \perp\!\!\!\perp D | R$$ Yes

$$T \perp\!\!\!\perp D | R, S$$



Proof Sketch

Statement: If X and Y and separated by Z in the moral graph, then X ⫫ Y | Z



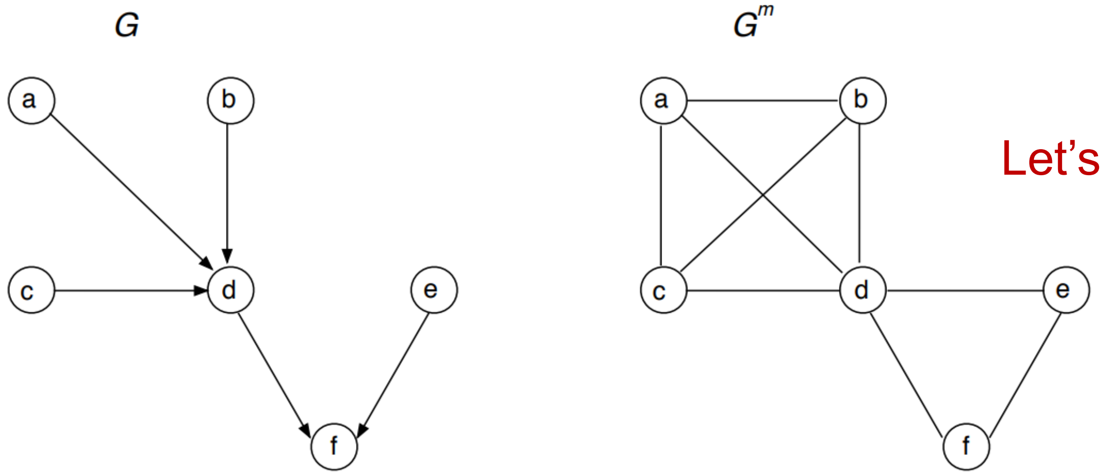
The moral graph gives a way to **"factorize"** the joint distribution of BN.

Each **clique** in the moral graph is a **factor**.

$$\underbrace{P(a)\ P(b)\ P(c)\ P(d \mid a, b, c)}_{\phi(a,\, b,\, c,\, d)}\ \underbrace{P(e)\ P(f \mid d, e)}_{\phi(d,\, e,\, f)} = \phi(a, b, c, d)\ \phi(d, e, f)$$

Proof Sketch

Statement: If X and Y and separated by Z in the moral graph, then X ⫫ Y | Z



Let's try to prove a ⫫ f | d

$$P(a|d) = \frac{P(a,d)}{P(d)} = \frac{\sum_f \phi(a,d)\phi(d,f)}{\sum_{a,f} \phi(a,d)\phi(d,f)} = \frac{\phi(a,d)\sum_f \phi(d,f)}{\sum_a \phi(a,d)\sum_f \phi(d,f)} = \frac{\phi(a,d)}{\sum_a \phi(a,d)}$$

$$P(a|d,f) = \frac{P(a,d,f)}{P(d,f)} = \frac{\phi(a,d)\phi(d,f)}{\sum_a \phi(a,d)\phi(d,f)} = \frac{\phi(a,d)}{\sum_a \phi(a,d)}$$

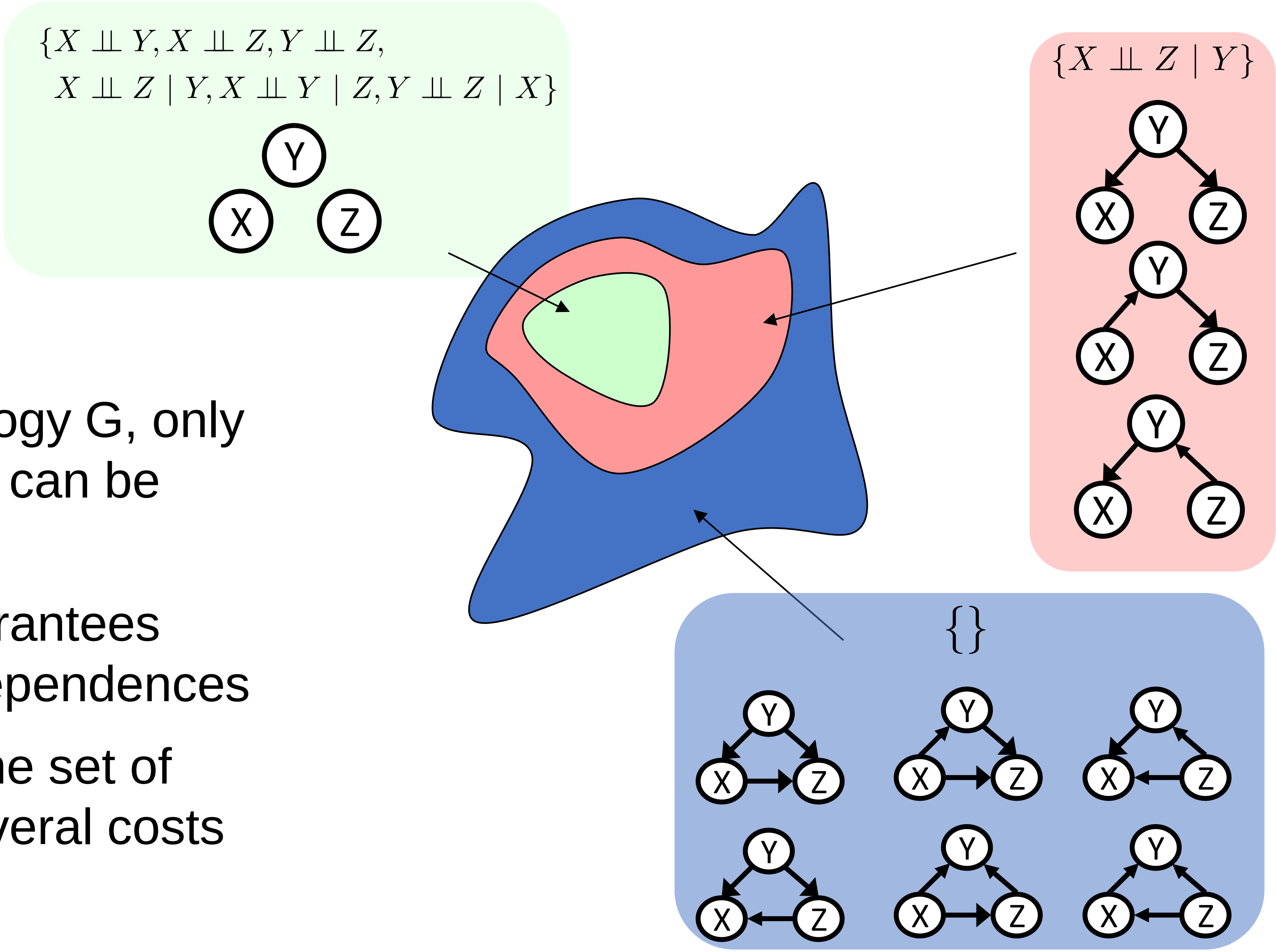
Structure Implications

- Given a Bayes net structure, can run d-separation algorithm to build a complete list of conditional independences that are necessarily true of the form

$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

- This list determines the set of probability distributions that can be represented

Topology Limits Distributions



- Given some graph topology G, only certain joint distributions can be encoded
- The graph structure guarantees certain (conditional) independences
- Adding arcs increases the set of distributions, but has several costs

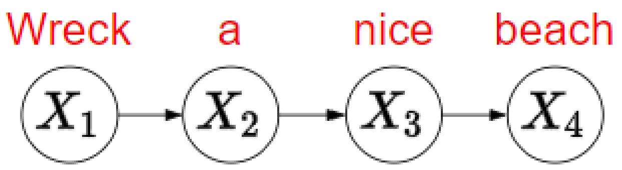
Application: Language Modeling

- Markov Model

Probabilistic program: Markov model

For each position $i = 1, 2, \ldots, n$:

    Generate word $X_i \sim p(X_i \mid X_{i-1})$



Application: Object Tracking

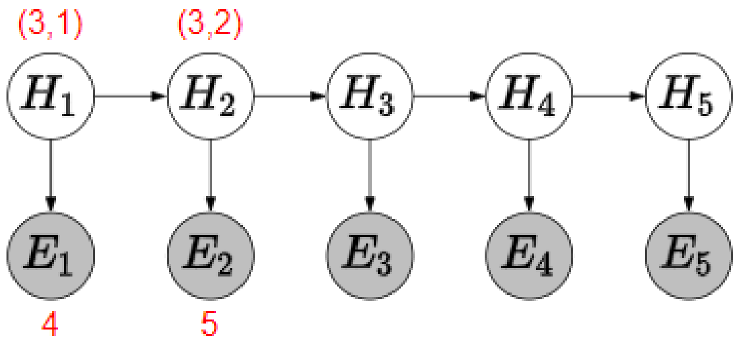
- Hidden Markov Model

Probabilistic program: hidden Markov model (HMM)

For each time step $t = 1, \ldots, T$:

Generate object location $H_t \sim p(H_t \mid H_{t-1})$

Generate sensor reading $E_t \sim p(E_t \mid H_t)$



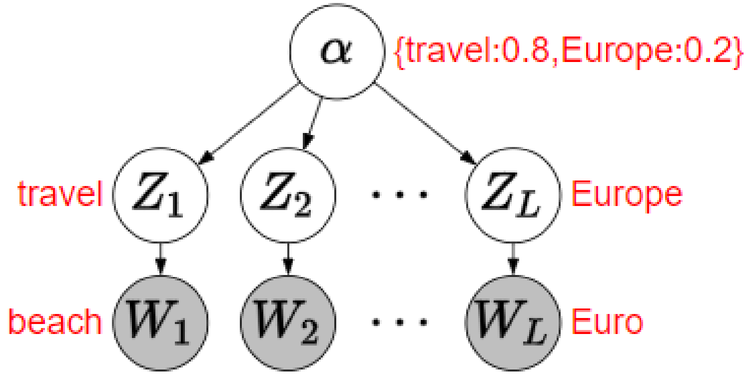
Inference: given sensor readings, where is the object?

Application: Topic Modeling

- Latent Dirichlet Allocation

Probabilistic program: latent Dirichlet allocation

Generate a distribution over topics $\alpha \in \mathbb{R}^K$
For each position $i = 1, \ldots, L$:
  Generate a topic $Z_i \sim p(Z_i \mid \alpha)$
  Generate a word $W_i \sim p(W_i \mid Z_i)$

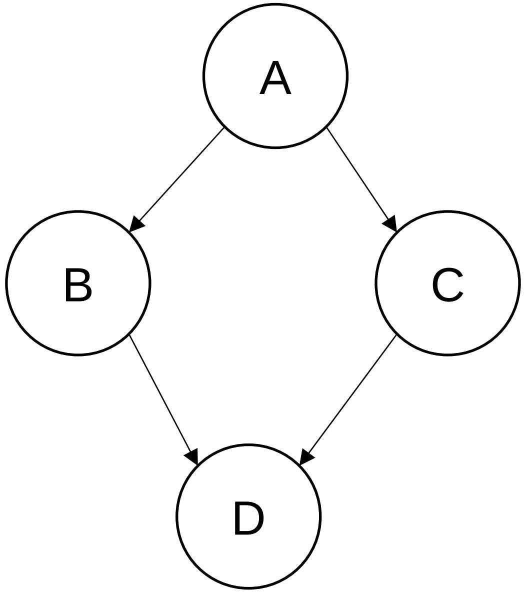


Document classification, information retrieval, customer segmentation, …

Inference: given a text document, what topics is it about?

Exact Inference in Bayesian Networks

The “Join” Operation in Bayesian Network

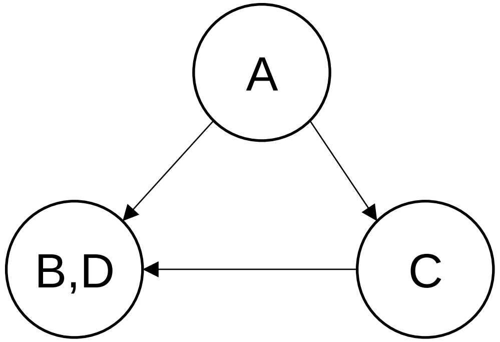


The BN defines four factors P(A), P(B|A), P(C|A), P(D|B,C)

Join on B: Combine all factors that involve B

P(A), P(B|A), P(C|A), P(D|B,C)

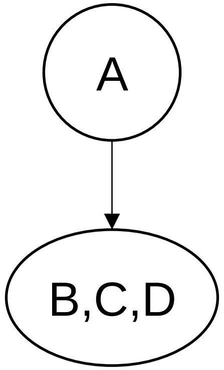
P(A), P(B,D | A,C), P(C|A)



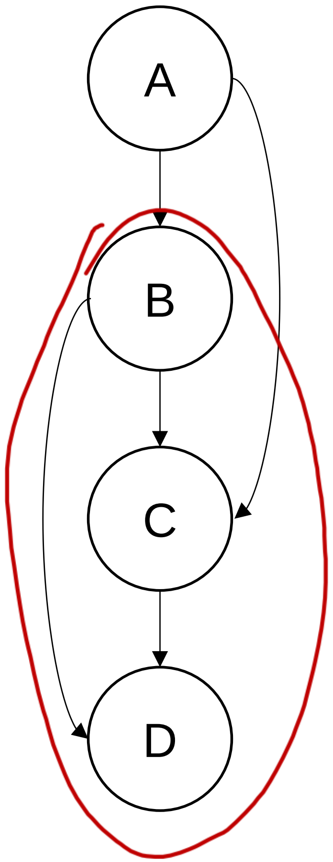
Further join on C: Combine all factors that involve C

P(A), P(B,D | A,C), P(C|A)

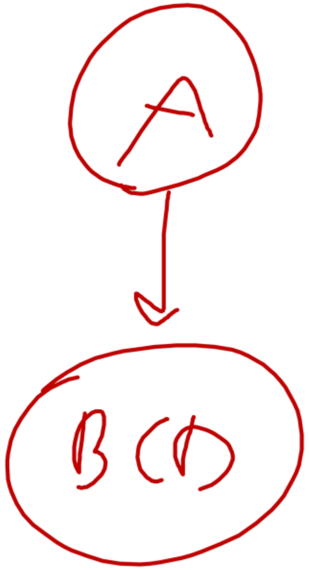
P(A), P(B,C,D | A)



Exercise



What are the factors after joining on B?



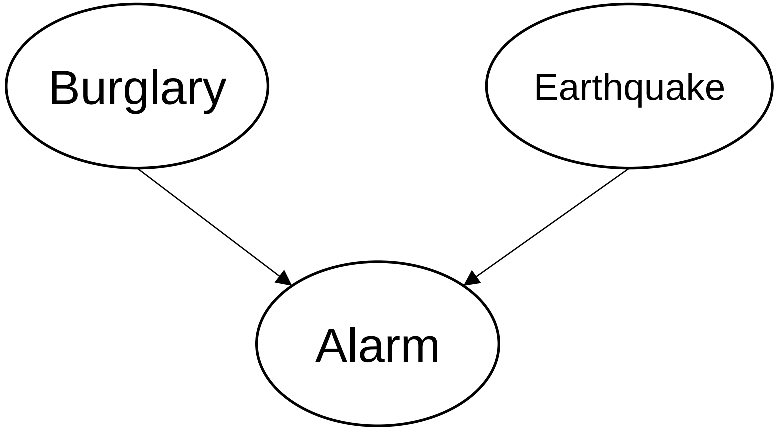
$$P(A)\ \underbrace{P(B|A)\,P(C|A,B)\,P(D|B,C)}$$

$$P(A)\qquad P(B,C,D|A)$$

Exercise

$P(b,a|e) = p(b)p(a|b,e)$

B	P(B)
T	0.001
F	0.999



E	P(E)
T	0.002
F	0.998

B	A	E	P(B,A|E)
T	T	T	
T	T	F	
T	F	T	
T	F	F	
F	T	T	
F	T	F	
F	F	T	
F	F	F	

B	E	A	P(A|B,E)
T	T	T	0.95
T	T	F	0.05
T	F	T	0.94
T	F	F	0.06
F	T	T	0.29
F	T	F	0.71
F	F	T	0.001
F	F	F	0.999

Can you calculate P(B, A|E)?

$p(b,a|e) = p(b|e)p(a|b,e)$

Review: Inference by Enumeration

General case:

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

$$\left.\begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array}\right\} \text{All variables}$$

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

$P(E_1, \dots, E_k, Q, H_1, \dots, H_r)$

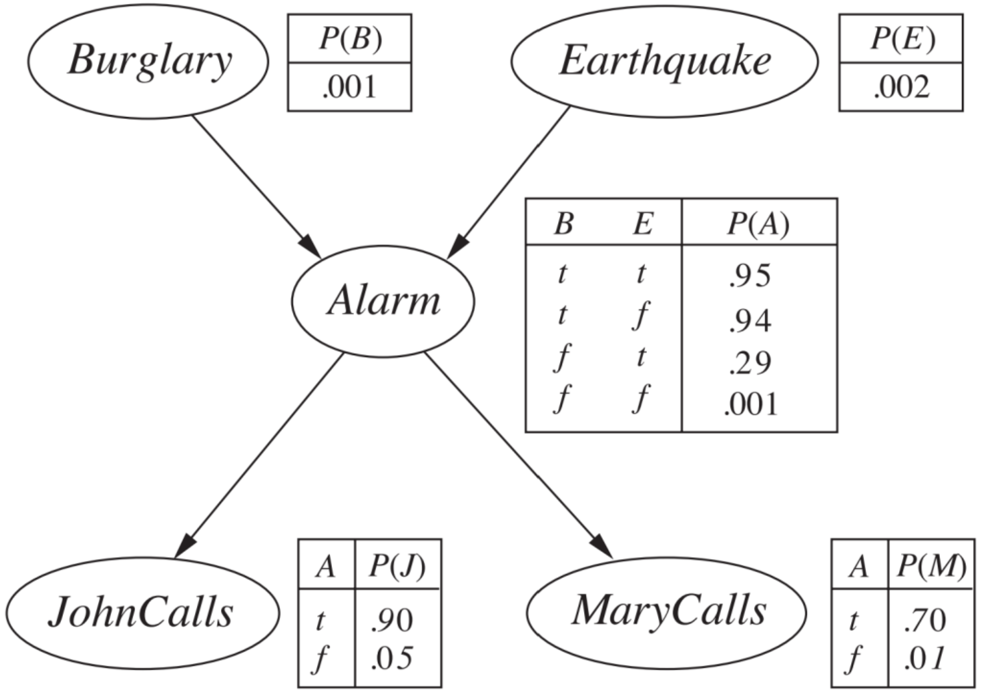
Inference by Enumeration

Step 1. Select the entries consistent with the evidence

Step 2. Sum out H to get joint probability of Query and evidence

Step 3. Normalize

Inference by Enumeration



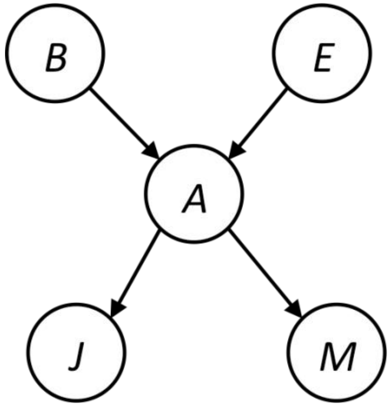
Step 0. Create a joint probability table

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

B	E	A	J	M	P(B,E,A,J,M)
T	T	T	T	T	0.001 * 0.002 * 0.95 * 0.90 * 0.70
T	T	T	T	F	0.001 * 0.002 * 0.95 * 0.90 * 0.30
T	T	T	F	T	0.001 * 0.002 * 0.95 * 0.10 * 0.70
…	…	…	…	…	
F	F	F	F	F	0.999 * 0.998 * 0.999 * 0.95 * 0.99

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

Step 0: Create a Joint Probability Table



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

B	P(B)
T	
F	

B	E	A	P(A|B,E)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

Join on B

B	A	E	P(B,A|E)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

E	P(E)
T	
F	

Join on E

B	E	A	P(B,E,A)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

A	J	P(J|A)
T	T	
…	…	

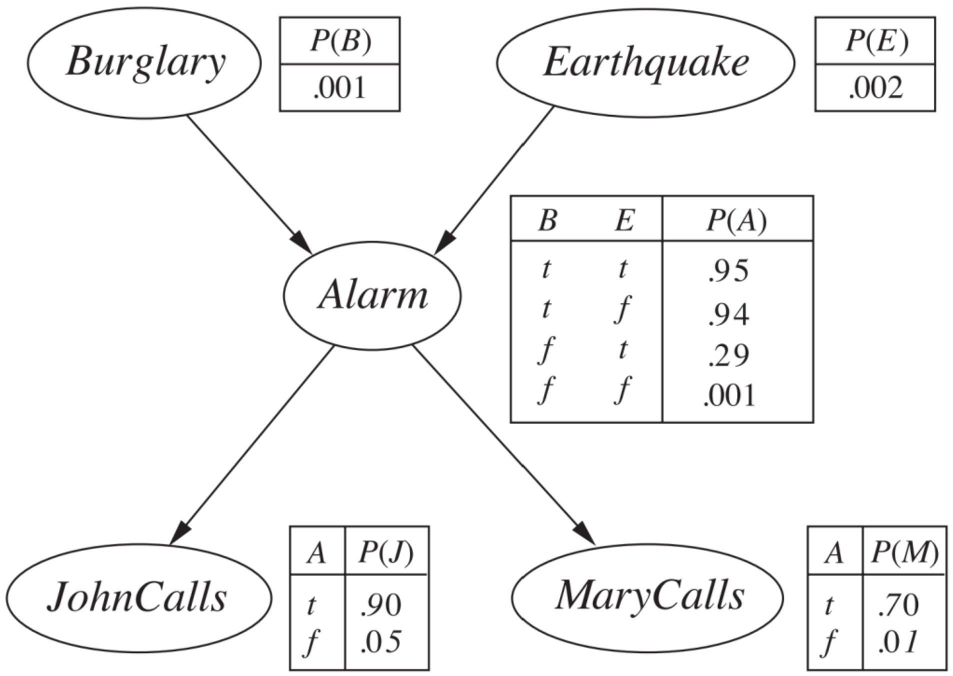
A	M	P(M|A)
T	T	
…	…	

Join on A

P(B,E,A,J,M)

Inference by Enumeration

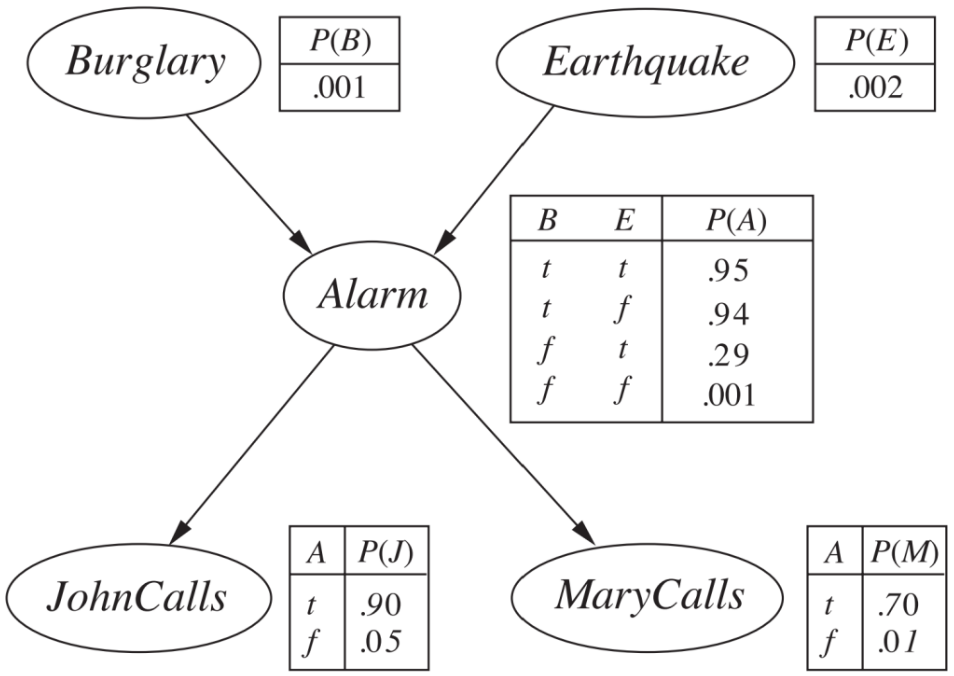
Step 1. Select the entries consistent with the evidence



B	E	A	J	M	P(B,E,A,J,M)
T	T	T	T	T	0.001 * 0.002 * 0.95 * 0.90 * 0.70
T	T	F	T	T	0.001 * 0.002 * 0.05 * 0.05 * 0.01
T	F	T	T	T	0.001 * 0.998 * 0.94 * 0.90 * 0.70
T	F	F	T	T	0.001 * 0.998 * 0.06 * 0.05 * 0.01
F	T	T	T	T	0.999 * 0.002 * 0.29 * 0.90 * 0.70
F	T	F	T	T	0.999 * 0.002 * 0.71 * 0.05 * 0.01
F	F	T	T	T	0.999 * 0.998 * 0.001 * 0.90 * 0.70
F	F	F	T	T	0.999 * 0.998 * 0.999 * 0.05 * 0.01

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

Inference by Enumeration



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

Step 2. Sum out hidden variable to get joint probability of query and evidence **(Marginalize)**

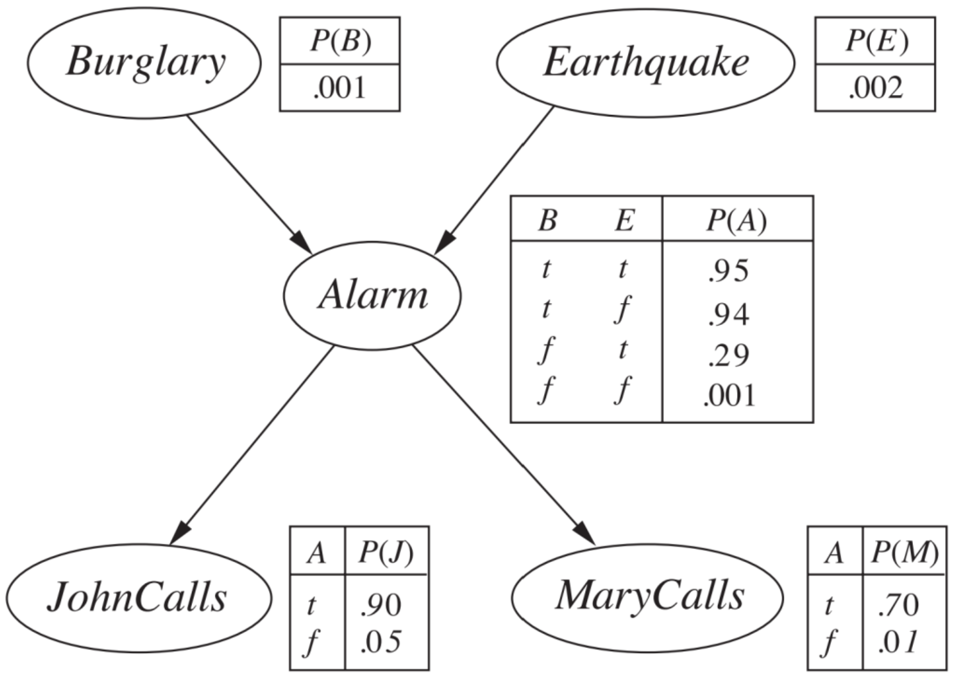
B	E	A	J	M	P(B,E,A,J,M)
T	T	T	T	T	0.001 * 0.002 * 0.95 * 0.90 * 0.70
T	T	F	T	T	0.001 * 0.002 * 0.05 * 0.05 * 0.01
T	F	T	T	T	0.001 * 0.998 * 0.94 * 0.90 * 0.70
T	F	F	T	T	0.001 * 0.998 * 0.06 * 0.05 * 0.01
F	T	T	T	T	0.999 * 0.002 * 0.29 * 0.90 * 0.70
F	T	F	T	T	0.999 * 0.002 * 0.71 * 0.05 * 0.01
F	F	T	T	T	0.999 * 0.998 * 0.001 * 0.90 * 0.70
F	F	F	T	T	0.999 * 0.998 * 0.999 * 0.05 * 0.01

+

+

B	J	M	P(B,J,M)
T	T	T	0.0006
F	T	T	0.0015

Inference by Enumeration



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

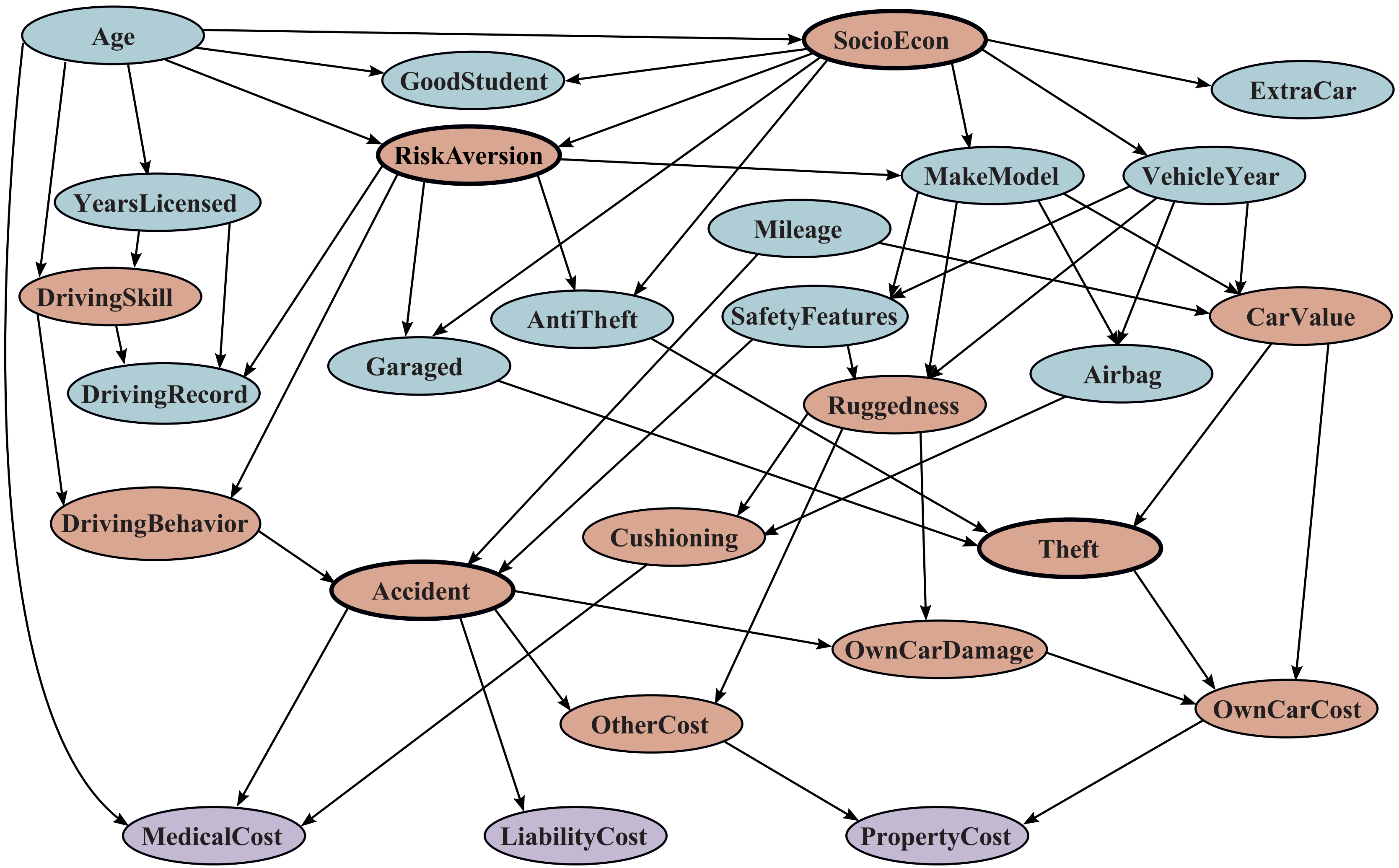
Step 3. Normalize

B	E	A	J	M	P(B,E,A,J,M)
T	T	T	T	T	0.001 * 0.002 * 0.95 * 0.90 * 0.70
T	T	F	T	T	0.001 * 0.002 * 0.05 * 0.05 * 0.01
T	F	T	T	T	0.001 * 0.998 * 0.94 * 0.90 * 0.70
T	F	F	T	T	0.001 * 0.998 * 0.06 * 0.05 * 0.01
F	T	T	T	T	0.999 * 0.002 * 0.29 * 0.90 * 0.70
F	T	F	T	T	0.999 * 0.002 * 0.71 * 0.05 * 0.01
F	F	T	T	T	0.999 * 0.998 * 0.001 * 0.90 * 0.70
F	F	F	T	T	0.999 * 0.998 * 0.999 * 0.05 * 0.01

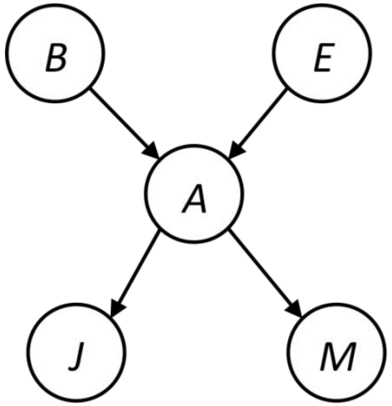
B	J	M	P(B,J,M)
T	T	T	0.0006
F	T	T	0.0015

B	P(B | +j, +m)
T	0.285
F	0.715

Inference by Enumeration?



How did we do Inference by Enumeration?



$$P(B,E,A,J,M) = P(B)\ P(E)\ P(A \mid B,E)\ P(J \mid A)\ P(M \mid A)$$

B	P(B)
T	
F	

B	E	A	P(A|B,E)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

Join on B

B	A	E	P(B,A|E)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

E	P(E)
T	
F	

Join on E

B	E	A	P(B,E,A)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

A	J	P(J|A)
T	T	
…	…	

A	M	P(M|A)
T	T	
…	…	

Join on A

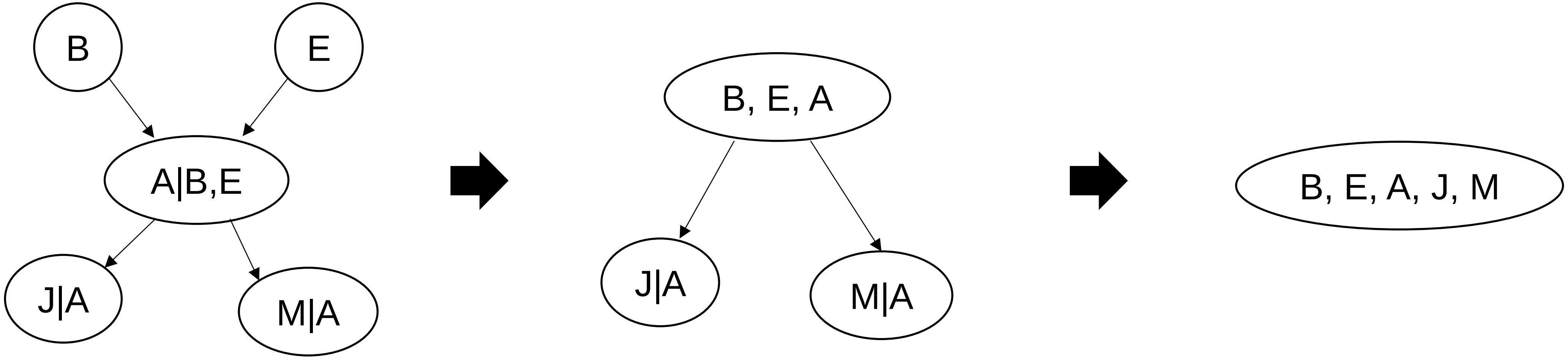
P(B,E,A,J,M)

We first create a big table by **joining all variables**, and then

1) Removing entries inconsistent with the evidence
2) Perform marginalization to eliminate hidden variables

How did we do Inference by Enumeration?

Each node here represents a “table”



Joining all variables **(Step 0)**

1) only keep rows consistent
with the evidence **(Step 1)**

2) Marginalize hidden variables
(Step 2)

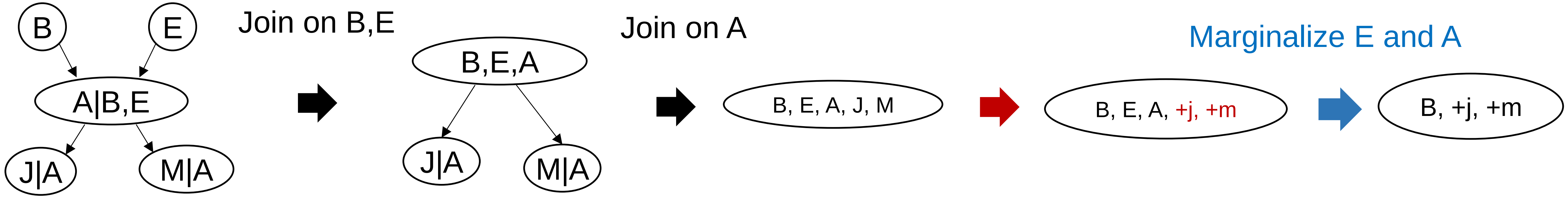
Improving the Algorithm

- First improvement: Instead of eliminating rows inconsistent with the evidence at the end, we will only keep rows consistent with evidence **from the beginning**.
- Second improvement: Instead of marginalize all hidden variables at the end after joining all variables, we will **interleave joining and marginalization**.

Improving the Algorithm

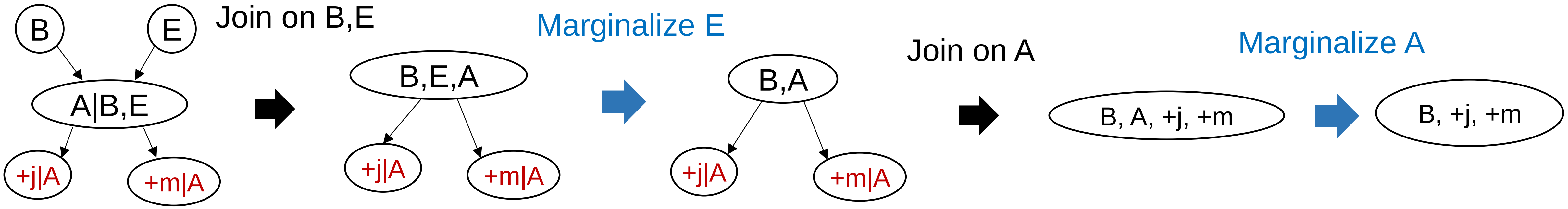
$P(B \mid +j, +m)$

Inference by Enumeration



A variable can only be marginalized when it's only involved in one factor. Otherwise, it has to be joined first.

Variable Elimination

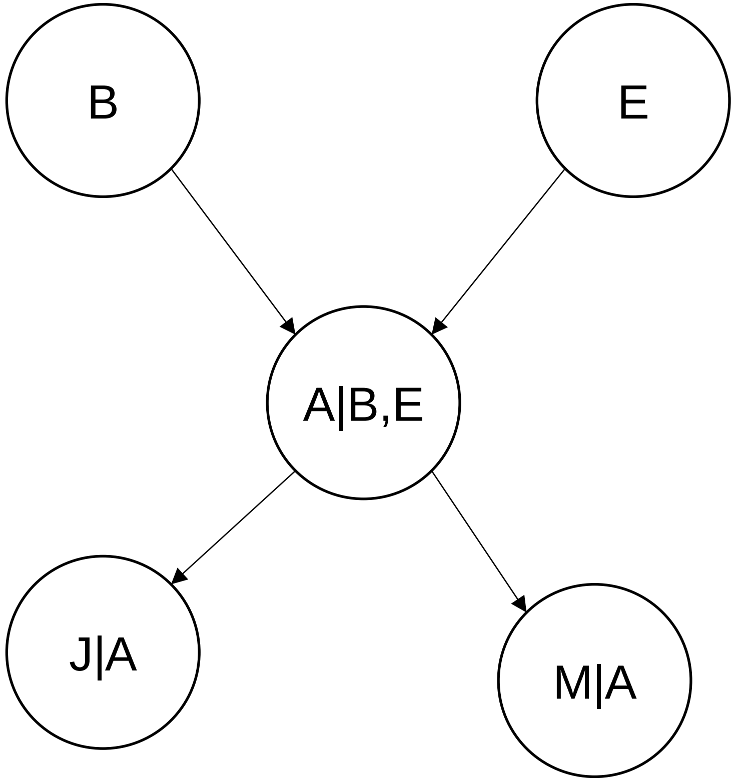


Variable Elimination

Query: P(B | +j, +m) = ?

B	P(B)
T	
F	

E	P(E)
T	
F	



B	E	A	P(A|B,E)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

A	J	P(J|A)
T	T	
T	F	
F	T	
F	F	

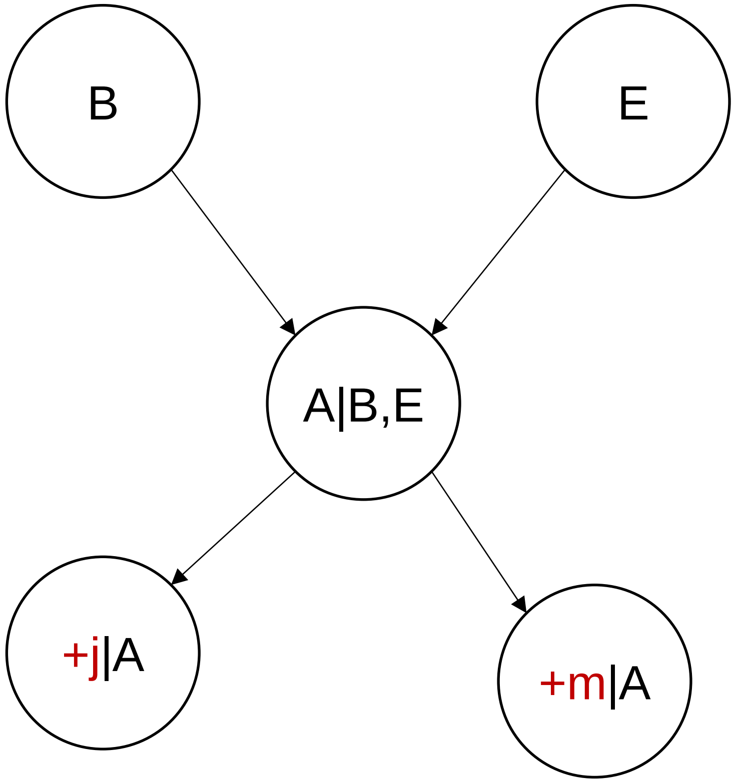
A	M	P(M|A)
T	T	
T	F	
F	T	
F	F	

Variable Elimination

Query: P(B | +j, +m) = ?

B	P(B)
T	
F	

E	P(E)
T	
F	



B	E	A	P(A|B,E)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

A	J	P(J|A)
T	T	
F	T	

A	M	P(M|A)
T	T	
F	T	

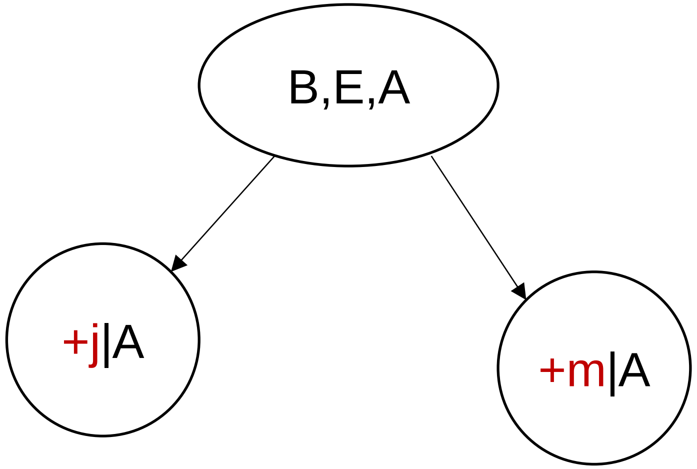
1) Only keep rows consistent with the evidence

Variable Elimination

Query: $P(B \mid +j, +m) = ?$

B	E	A	P(B,E,A)
T	T	T	
T	T	F	
…	…	…	
F	F	F	

Join on B and E



A	J	P(J|A)
T	T	
F	T	

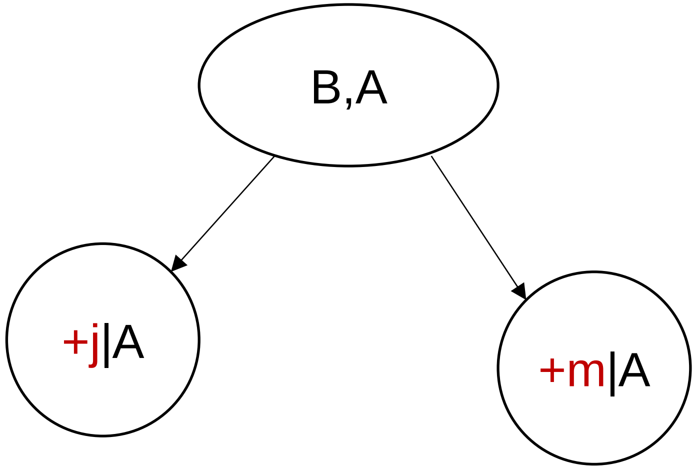
A	M	P(M|A)
T	T	
F	T	

Variable Elimination

Query: P(B | +j, +m) = ?

2) Marginalize E (earlier than inference by enumeration)

B	A	P(B,A)
T	T	
T	F	
F	T	
F	F	



A	J	P(J|A)
T	T	
F	T	

A	M	P(M|A)
T	T	
F	T	

Variable Elimination

Query: P(B | +j, +m) = ?

Join on A

B,A,+j,+m

B	A	J	M	P(B,A,J,M)
T	T	T	T	
T	F	T	T	
F	T	T	T	
F	F	T	T	

Variable Elimination

Query: $P(B \mid +j, +m) = ?$

Marginalize A

B,+j,+m

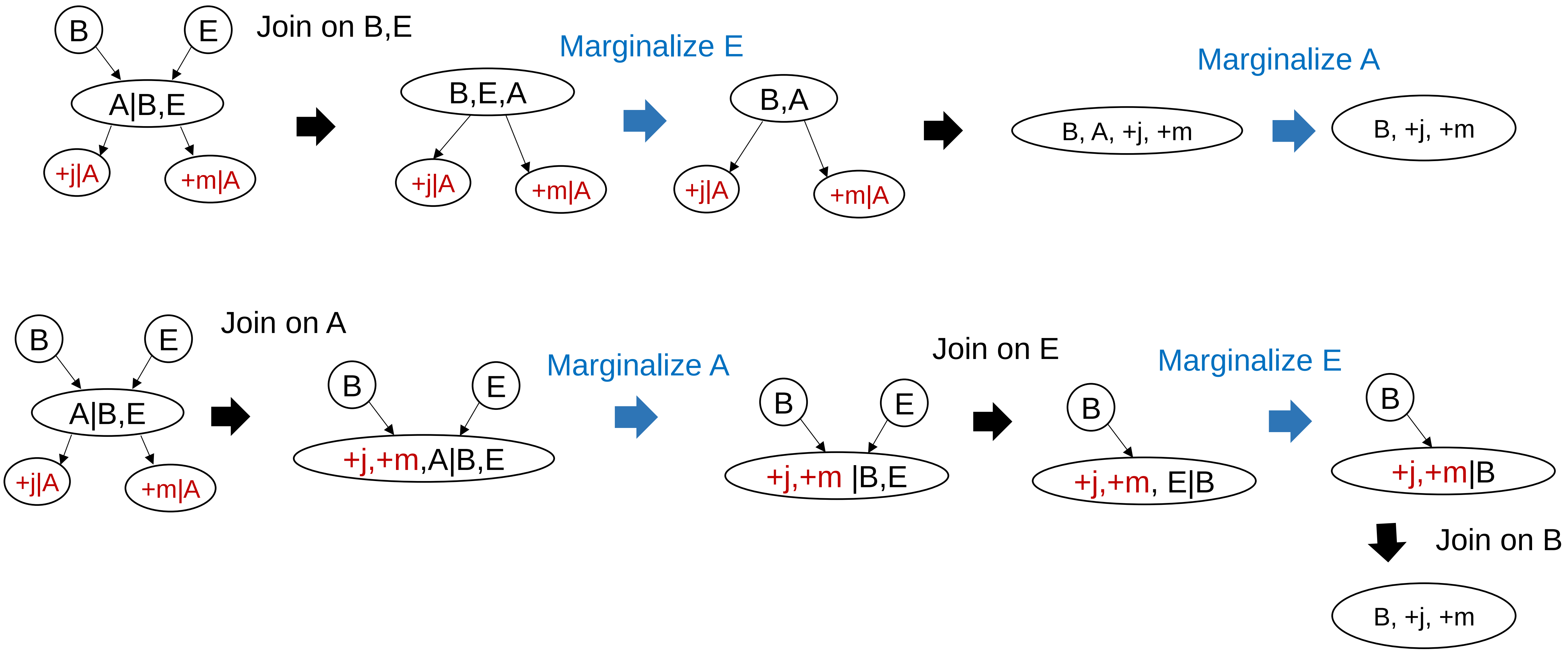
B	J	M	P(B,J,M)
T	T	T	
F	T	T	

We can then get $P(B \mid +j, +m)$ by normalizing this table

Variable Elimination

Query: P(B | +j, +m) = ?

Can be done in different orders



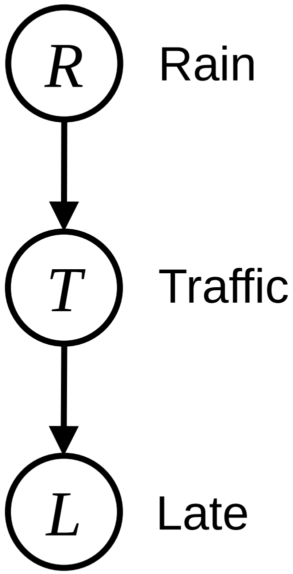
Variable Elimination

- Start with initial factors but instantiated by evidence
- While there are still hidden variables:
  - Pick a hidden variable X
  - Join all factors mentioning X
  - Eliminate (sum out) X   (i.e., marginalize X)
- Join all the remaining factors
- Normalize

Ordering of the Join and Marginalize?

- The time and space of variable elimination are dominate by the **size of the largest factor** constructed during the algorithm.
- It’s hard to determine the optimal ordering
  - Heuristics: Choose the variable that minimize the size of the next factor to be constructed.

Exercise



$P(R)$

+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

Calculate P(L)
(Use the heuristic: minimize the size of the next constructed factor)

P(L)

+l	
-l	

P(R, T)

+r	+t
+r	-t
-r	+t
-r	-t

P(T)

+t
-t

P(L, T)

P(L)

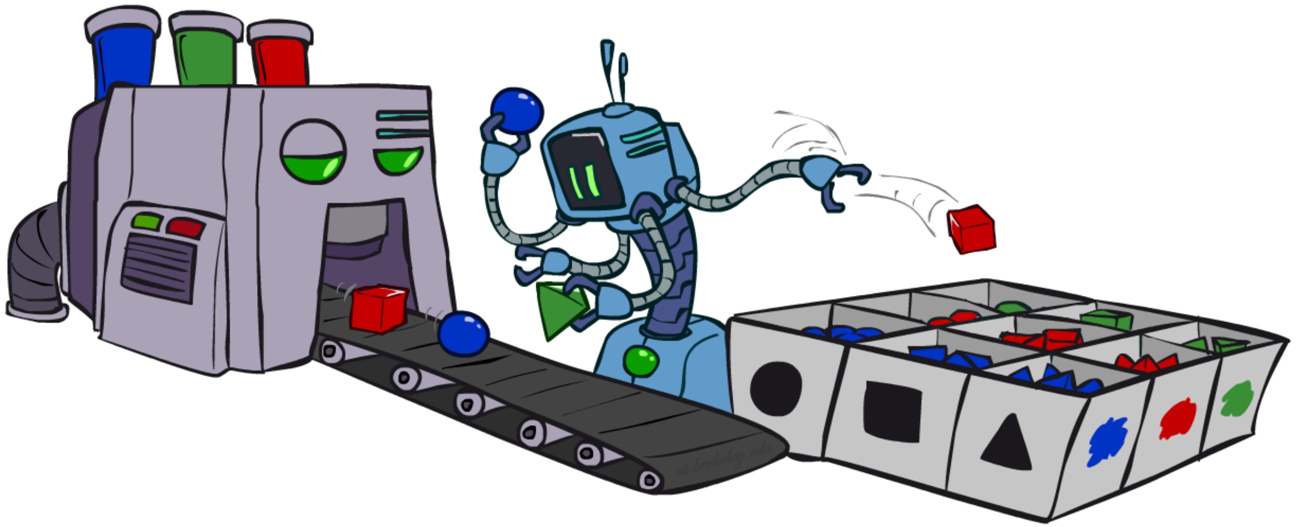
Approximate Inference in Bayesian Networks

- Still, the inference procedure may still be time consuming if the Bayesian network is dense.

Sampling

- Basic idea
  - Draw N samples from a ***sampling distribution*** S
  - Compute an approximate posterior probability
  - Show this converges to the true probability P

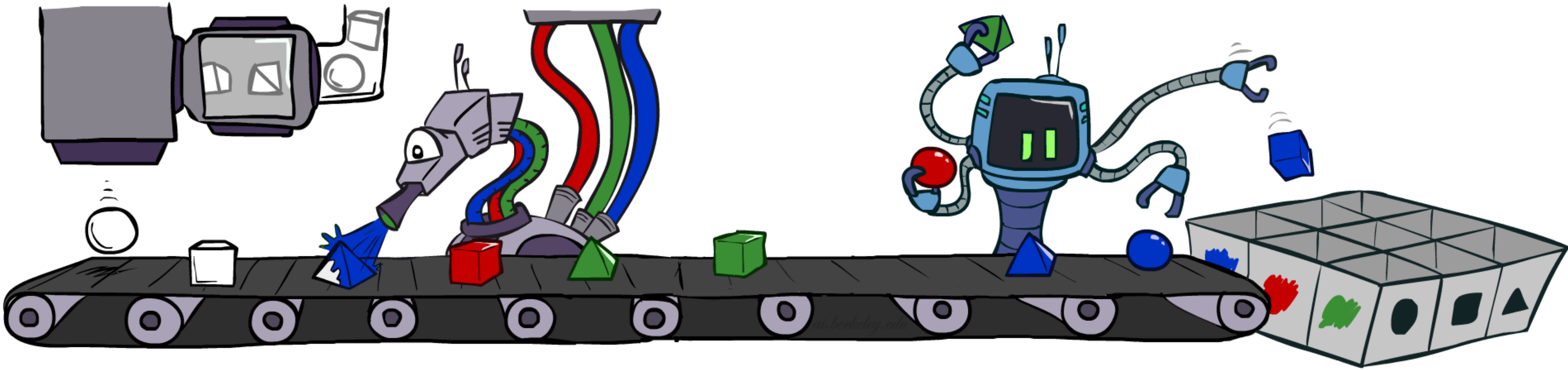
- Why sample?
  - Often very fast to get a decent approximate answer
  - The algorithms are very simple and general (easy to apply to fancy models)
  - They require very little memory ($O(n)$)
  - They can be applied to large models, whereas exact algorithms blow up



Sampling in Bayes nets

- Prior sampling
- Rejection sampling
- Likelihood weighting
- Gibbs sampling

Prior Sampling



Prior Sampling

$P(W)$

W

¬W

15/20

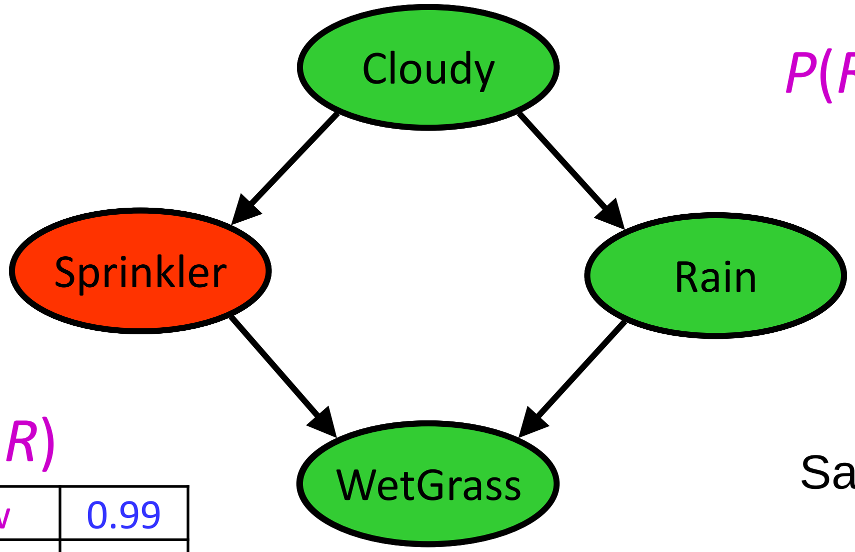
5/20

$P(C)$

c	0.5
¬c	0.5

$P(S \mid C)$

c	s	0.1
	¬s	0.9
¬c	s	0.5
	¬s	0.5



$P(R \mid C)$

c	r	0.8
	¬r	0.2
¬c	r	0.2
	¬r	0.8

$P(W \mid S,R)$

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

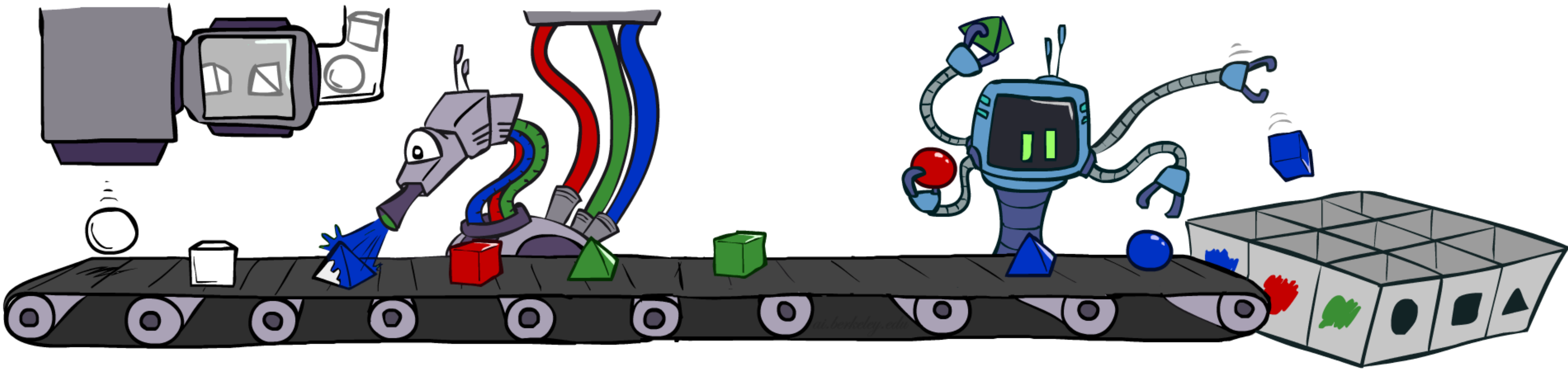
Samples:

$$\left.\begin{array}{l} c, \neg s, r, w \\ \neg c, s, \neg r, w \\ \ldots \end{array}\right\} \begin{array}{l} 15\ w \\ 5\ \neg w \end{array}$$

$S_{PS}(c, \neg s, r, w) =$

Prior Sampling

- For i=1, 2, …, n (in topological order)
  - Sample X_i from $P(X_i \mid parents(X_i))$
- Return $(x_1, x_2, \ldots, x_n)$



Prior Sampling

- This process generates samples with probability:

  $$S_{PS}(x_1,\ldots,x_n) = \prod_i P(x_i \mid parents(X_i)) = P(x_1,\ldots,x_n)$$

  …i.e. the BN's joint probability

- Let the number of samples of an event be $N_{PS}(x_1,\ldots,x_n)$
- Estimate from N samples is $Q_N(x_1,\ldots,x_n) = N_{PS}(x_1,\ldots,x_n)/N$
- Then $\lim_{N\to\infty} Q_N(x_1,\ldots,x_n) = \lim_{N\to\infty} N_{PS}(x_1,\ldots,x_n)/N$

  $$= S_{PS}(x_1,\ldots,x_n)$$

  $$= P(x_1,\ldots,x_n)$$

- I.e., the sampling procedure is ***consistent***

Example

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

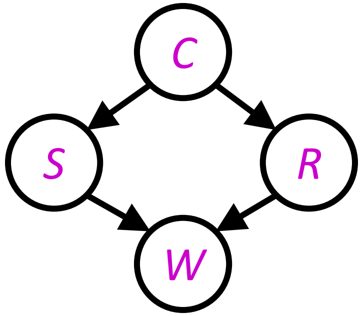
  c, ¬s, r, w

  c, s, r, w

  ¬c, s, r, ¬w

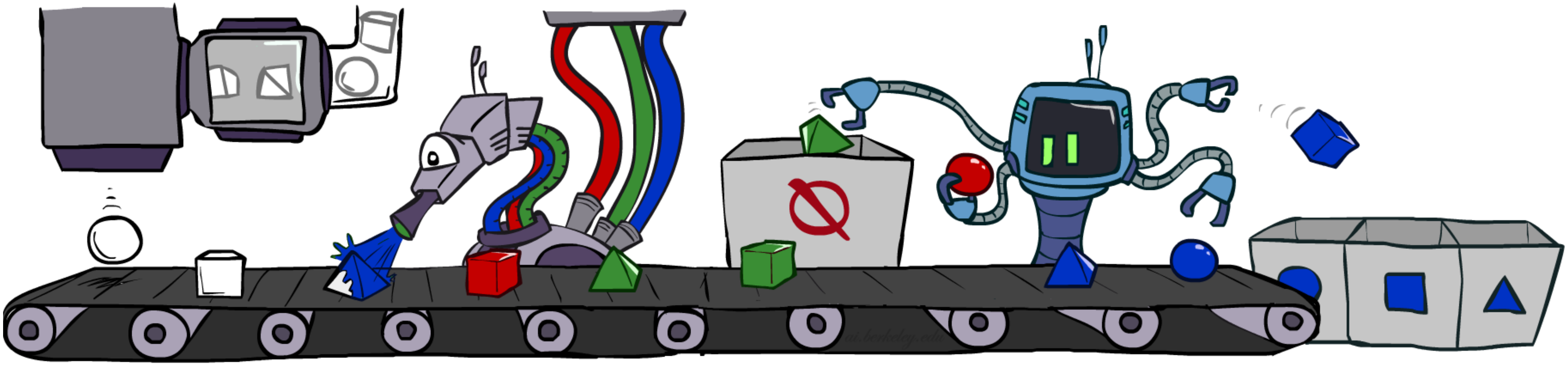
  c, ¬s, r, w

  ¬c, ¬s, ¬r, w



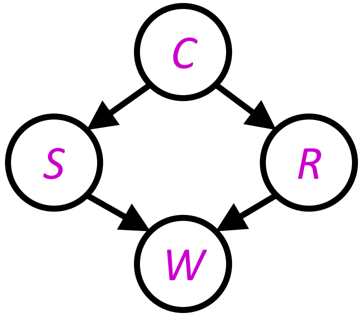
- If we want to know $P(W)$
  - We have counts <w:4, ¬w:1>
  - Normalize to get $P(W)$ = <w:0.8, ¬w:0.2>
  - This will get closer to the true distribution with more samples

Rejection sampling



Rejection sampling

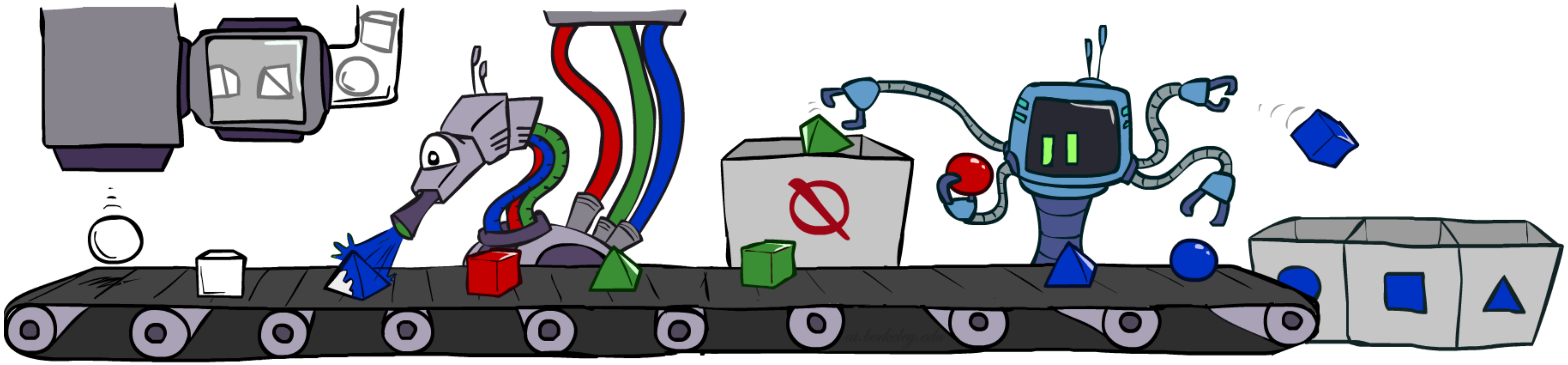
- A simple application of prior sampling for estimating conditional probabilities
  - Let's say we want $P(C|\ r, w) = \alpha\ P(C, r, w)$
  - For these counts, samples with $\neg r$ or $\neg w$ ***are not relevant***
  - So count the C outcomes for samples with r, w and reject all other samples
- This is called ***rejection sampling***
  - It is also consistent for conditional probabilities (i.e., correct in the limit)



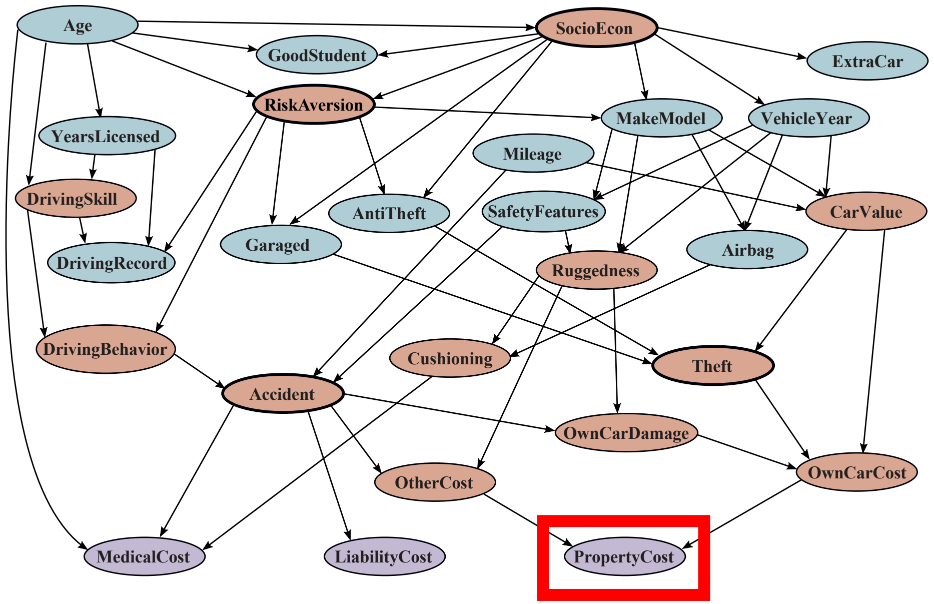
c, ¬s, r, w
~~c, s, ¬r~~
~~¬c, s, r, ¬w~~
~~c, ¬s, ¬r~~
¬c, ¬s, r, w

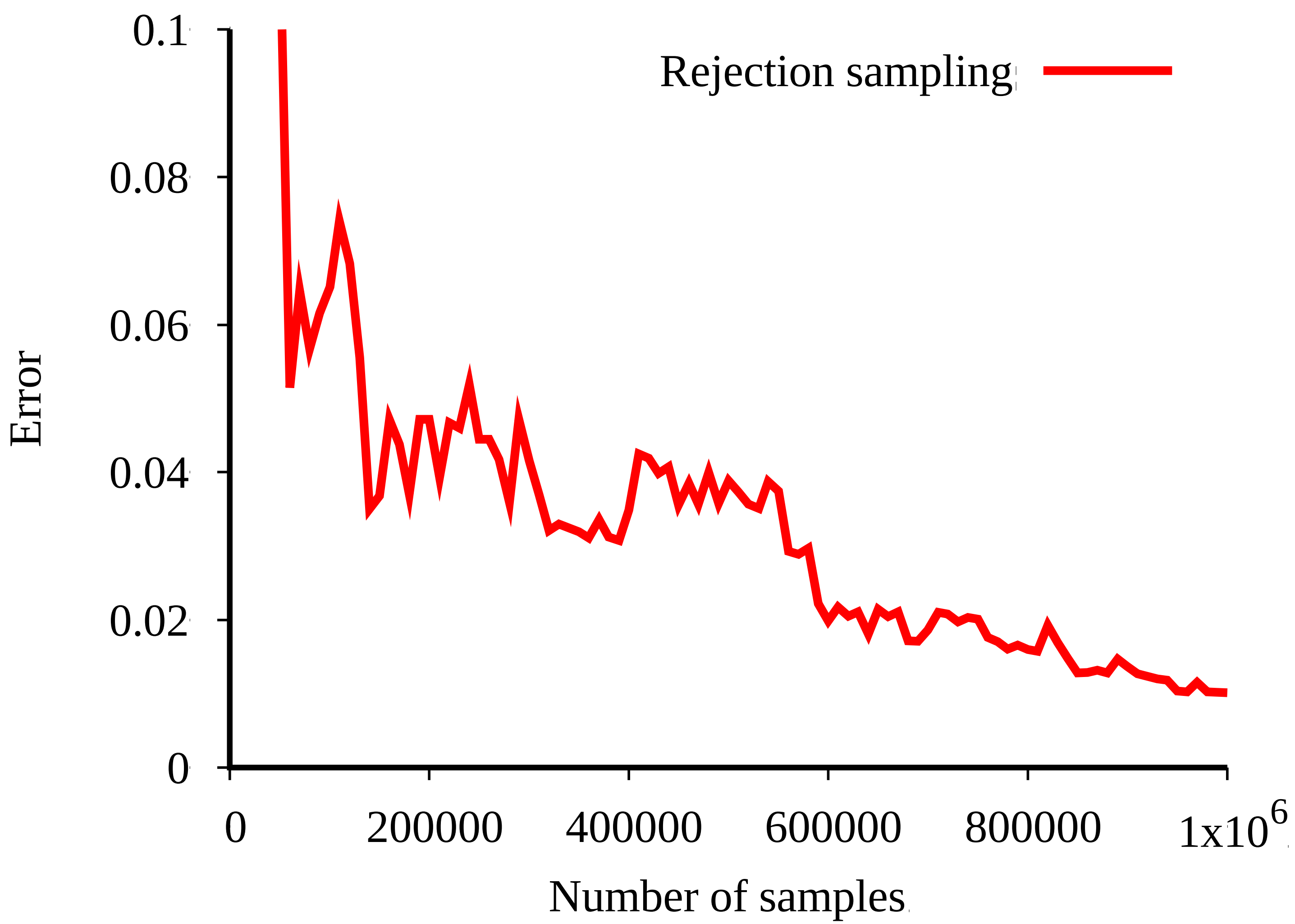
Rejection sampling

- Input: evidence $e_1,..,e_k$
- For i=1, 2, …, n
  - Sample X_i from $P(X_i \mid parents(X_i))$
  - If x_i not consistent with evidence
    - Reject: Return, and no sample is generated in this cycle
- Return $(x_1, x_2, \ldots, x_n)$

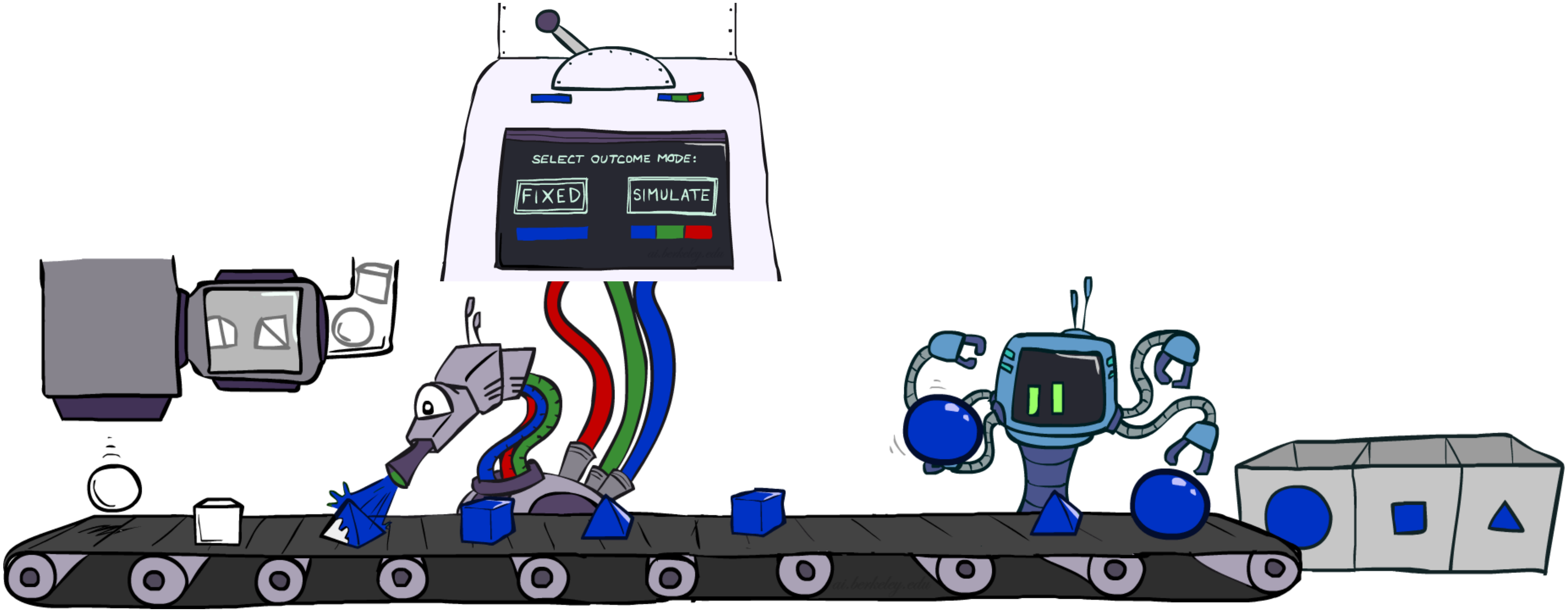


Car Insurance: *P(PropertyCost | e)*



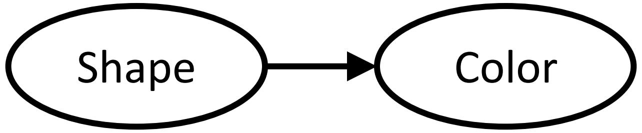


Likelihood weighting

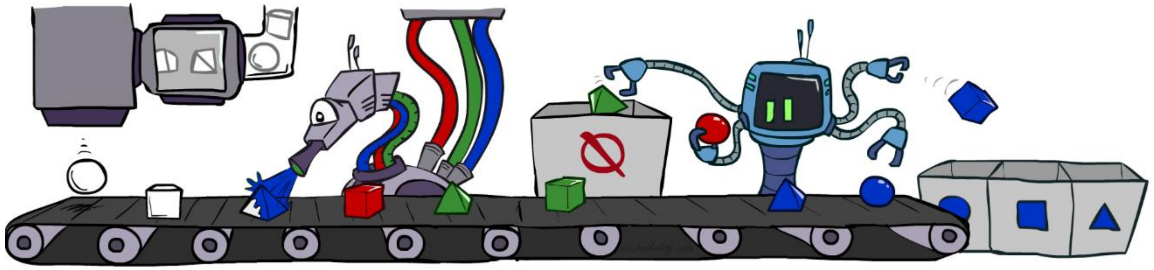


Likelihood weighting

- Problem with rejection sampling:
  - If evidence is unlikely, rejects lots of samples
  - Evidence not exploited as you sample
  - Consider P(*Shape*|*Color*=*blue*)

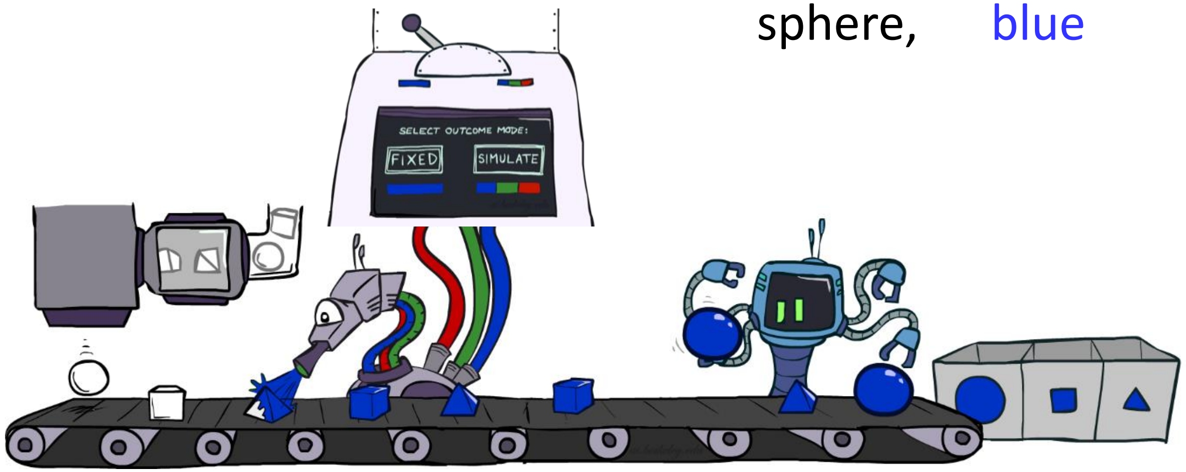


~~pyramid, green~~
~~pyramid, red~~
sphere, blue
~~cube, red~~
~~sphere, green~~



- Idea: fix evidence variables, sample the rest
  - Problem: sample distribution not consistent!
  - Solution: ***weight*** each sample by probability of evidence variables given parents

pyramid, blue
pyramid, blue
sphere, blue
cube, blue
sphere, blue



Likelihood Weighting

$P(C)$

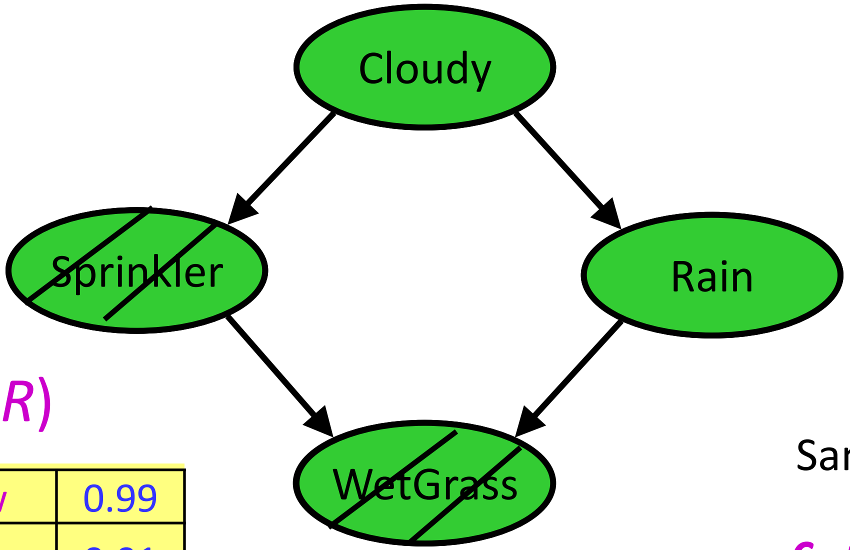
c	0.5
¬c	0.5

$P(S \mid C)$

c	s	0.1
	¬s	0.9
¬c	s	0.5
	¬s	0.5

$P(R \mid C)$

c	r	0.8
	¬r	0.2
¬c	r	0.2
	¬r	0.8



$P(W \mid S,R)$

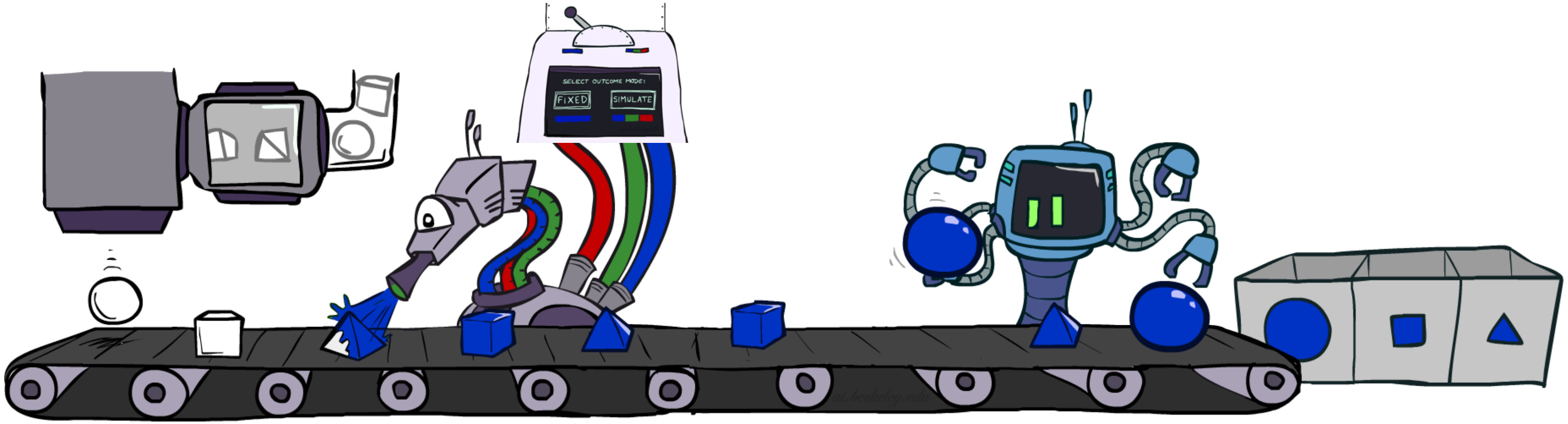
s	r	w	0.99
		¬w	0.01
	¬r	w	0.90
		¬w	0.10
¬s	r	w	0.90
		¬w	0.10
	¬r	w	0.01
		¬w	0.99

Samples:

c*, *s*, *r*, *w   ***w*** **= 1.0**   **x 0.1**   **x 0.99**

Likelihood weighting

- Input: evidence $e_1,..,e_k$
- $w = 1.0$
- for i=1, 2, …, n
  - if X_i is an evidence variable
    - x_i = observed value$_i$ for X_i
    - Set $w = w * P(x_i \mid parents(X_i))$
  - else
    - Sample x_i from $P(X_i \mid parents(X_i))$
- return $(x_1, x_2, \ldots, x_n)$, w



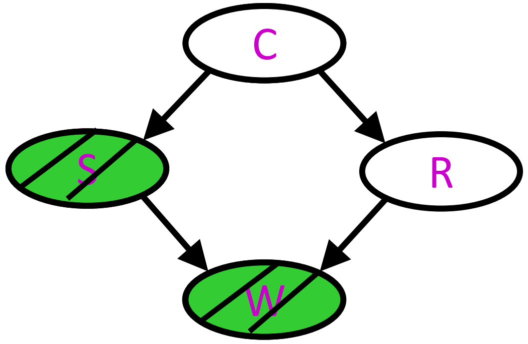
Likelihood weighting is consistent

- Sampling distribution if **Z** sampled and **e** fixed evidence

  $S_{WS}(\mathbf{z},\mathbf{e}) = \prod_j P(z_j \mid parents(Z_j))$

- Now, samples have weights

  $w(\mathbf{z},\mathbf{e}) = \prod_k P(e_k \mid parents(E_k))$



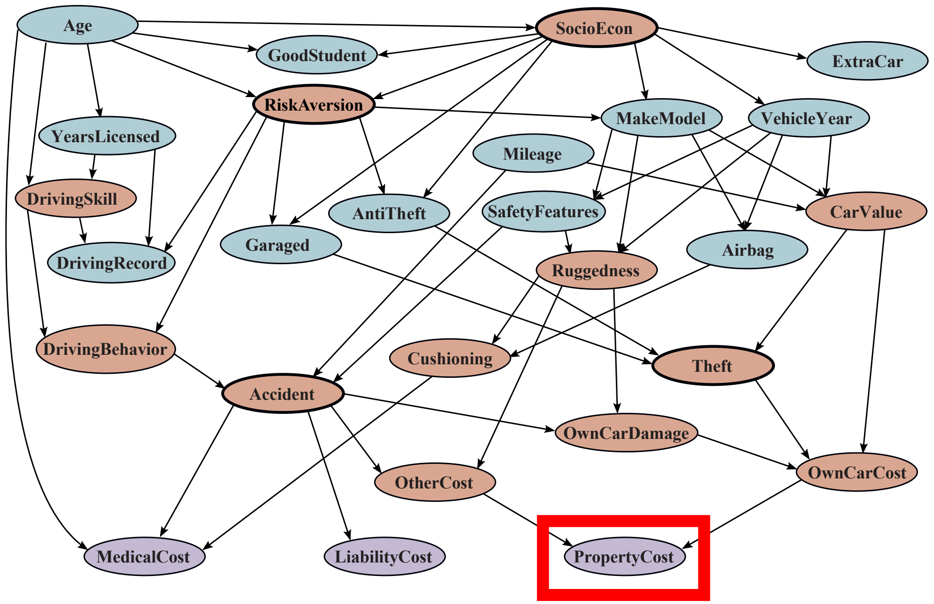
- Together, weighted sampling distribution is consistent

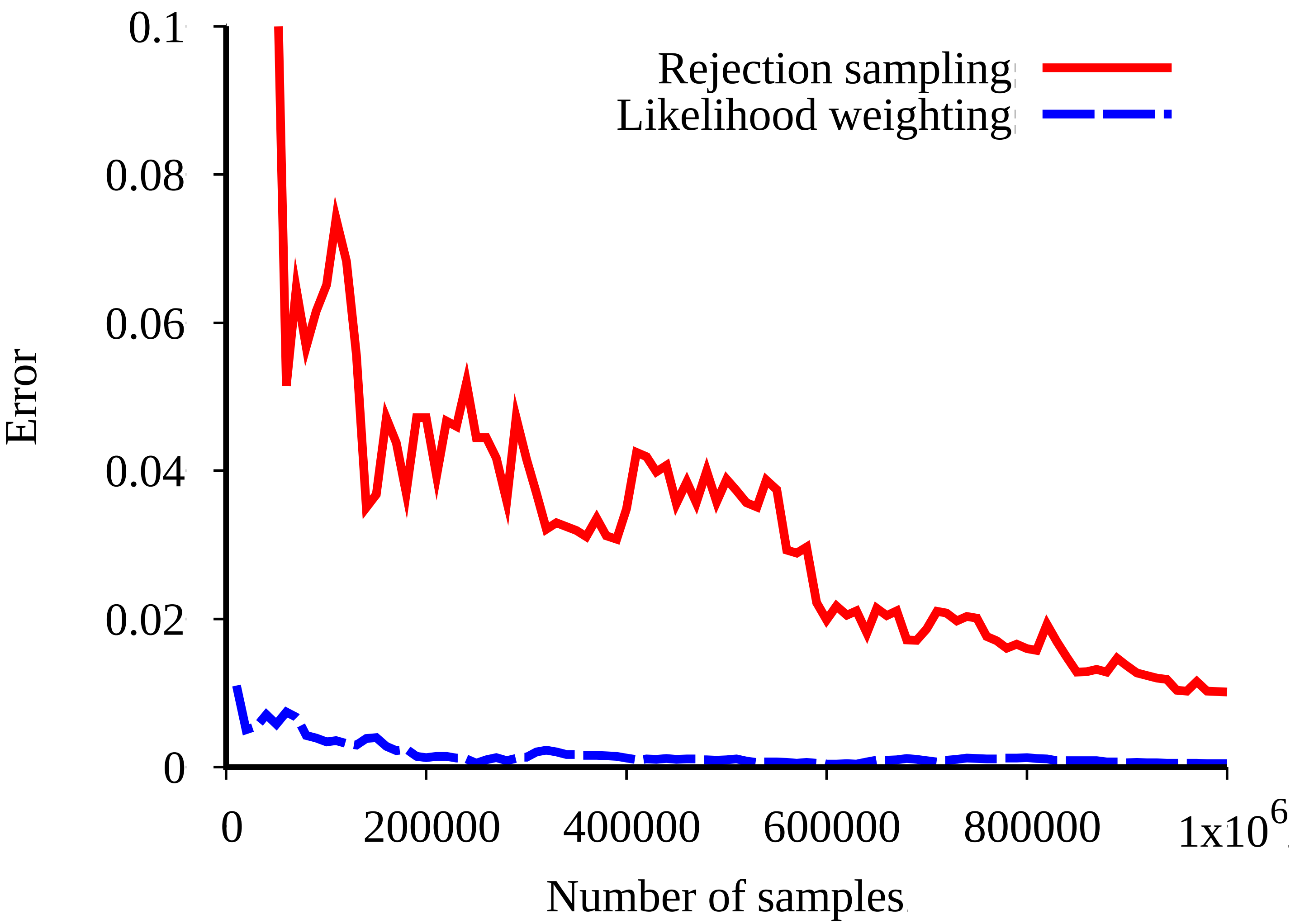
  $S_{WS}(\mathbf{z},\mathbf{e}) \cdot w(\mathbf{z},\mathbf{e}) = \prod_j P(z_j \mid parents(Z_j)) \prod_k P(e_k \mid parents(E_k))$

  $= P(\mathbf{z},\mathbf{e})$

- Likelihood weighting is an example of ***importance sampling***
  - Would like to estimate some quantity based on samples from P
  - P is hard to sample from, so use Q instead
  - Weight each sample x by $P(x)/Q(x)$

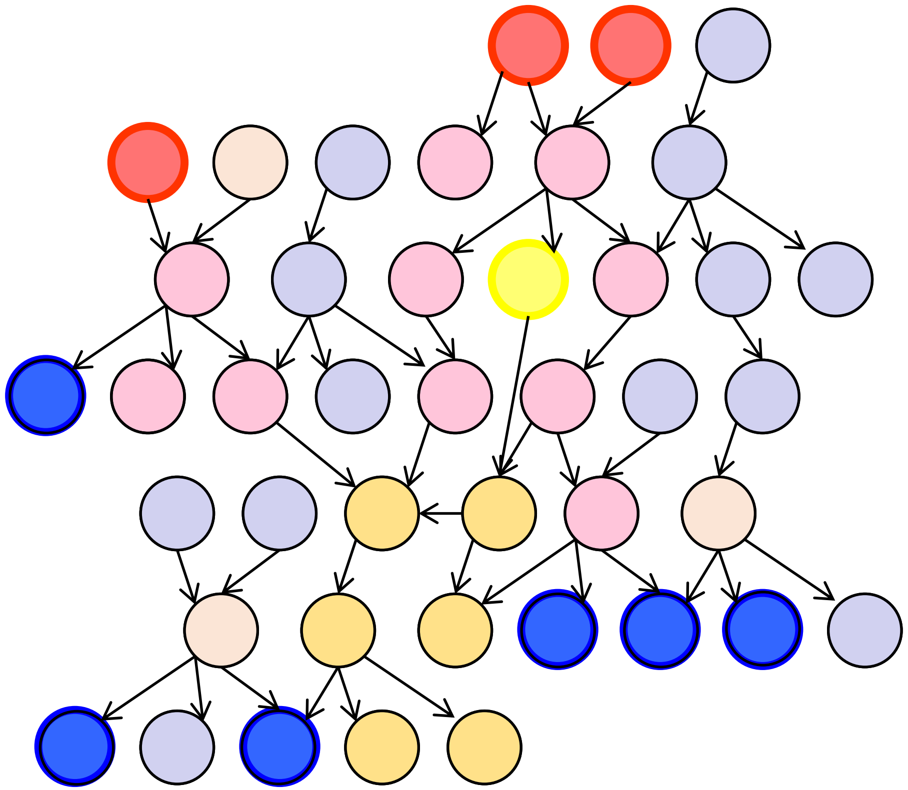
Car Insurance: *P*(*PropertyCost* | *e*)





Likelihood weighting

- Likelihood weighting is good
  - All samples are used
  - The values of ***downstream*** variables are influenced by ***upstream*** evidence



- Likelihood weighting still has weaknesses
  - The values of ***upstream*** variables are unaffected by ***downstream*** evidence
    - E.g., suppose evidence is a video of a traffic accident
  - With evidence in k leaf nodes, weights will be $O(2^{-k})$
  - With high probability, one lucky sample will have much larger weight than the others, dominating the result

- We would like each variable to "see" ***all*** the evidence!