Probabilistic Reasoning

Outline

- Probability theory - basics
- Probabilistic Inference
  - Inference by enumeration
- Conditional Independence
- Bayesian Networks
- Inference in Bayesian Networks
  - Exact Inference
  - Approximate Inference

Non-monotonic logic

- Traditional logic is monotonic
  - The set of legal conclusions grows monotonically with the set of facts appearing in our initial database
- When humans reason, we use defeasible logic
  - Almost every conclusion we draw is subject to reversal
  - If we find contradicting information later, we'll want to retract earlier inferences
- Nonmonotonic logic, or defeasible reasoning, allows a statement to be retracted
- Solution: Truth Maintenance
  - Keep explicit information about which facts/inferences support other inferences
  - If the foundation disappears, so must the conclusion

Uncertainty

- On the other hand, the problem might not be in the fact that T/F values can change over time but rather that we are not certain of the T/F value
- Agents almost never have access to the whole truth about their environment
- Agents must act in the presence of uncertainty
  - Incompleteness and/or incorrectness of rules used by the agent
  - Limited and ambiguous sensors
  - Imperfection/noise in agent's actions
  - Dynamic nature of the environment

Pitfalls of pure logic

- Laziness
  - Too much work to list all conditions needed to ensure an exceptionless rule
- Theoretical ignorance
  - Science has no complete theory for the domain
- Practical ignorance
  - Even if we know all the rules, we may be uncertain about them
  - We only have a degree of belief in them

Probability Theory vs. Logic

- Probability - tool for handling degrees of belief
  - Summarizes uncertainty due to laziness and ignorance

	Ontological commitments	Epistemological commitments
Logic	world is composed of facts that do or do not hold	each sentence is true or false or unknown
Probability Theory	world is composed of facts that do or do not hold	numerical degree of belief between 0(sentence for certainly false) and 1(sentences are certainly true)

Rational Agent Approach

- Choose action A that maximizes expected utility
  - Maximizes Prob(A)*Utility(A)
- Prob(A) - probability that A will succeed
- Utility(A) - utility value to agent of A's outcomes

$$P(A_{25} \text{ gets me there on time} | \ldots) = 0.04$$
$$P(A_{90} \text{ gets me there on time} | \ldots) = 0.70$$
$$P(A_{120} \text{ gets me there on time} | \ldots) = 0.95$$
$$P(A_{1440} \text{ gets me there on time} | \ldots) = 0.9999$$

Which action to choose?

Depends on my preferences for missing flight vs. airport cuisine, etc.

Utility theory is used to represent and infer preferences

Decision theory = utility theory + probability theory

Probability Basics

- Begin with a set Ω – the sample space
  - Example: 6 possible rolls of a die
  - $\omega \in \Omega$ is a sample point/possible world/atomic event
- A probability space or model is a sample space with an assignment $P(\omega) \forall \omega \in \Omega$ such that
  - $0 < P(\omega) < 1$
  - $\sum_{\omega} P(\omega) = 1$
- An event A is any subset of Ω and
  - $P(A) = \sum_{(\omega \in A)} P(\omega)$

Random Variables

- A random variable is a function from sample points to some range (e.g., reals or Boolean)
  - Example $Odd(1) = true$
- P induces a probability distribution for random variable (r.v.) X:
  - $P(X = x_i) = \sum_{\omega:X(\omega)=x_i} P(\omega)$
  - Example: $P(Odd = true) = P(1) + P(3) + P(5)$

Propositions

- Think of a proposition as the event (set of sample points) where the proposition is true
- Given Boolean r.v. A and B:
  - Event a = set of sample points where $A(\omega) = true$
  - Event $\neg a$ = set of sample points where $A(\omega) = false$
  - Event $a \wedge b$ = sample points were $A(\omega) = true$ and $B(\omega) = true$
- Often in AI applications, the sample points are defined by the values of a set of random variables i.e., the sample space is the Cartesian product of the ranges of the variables
- With Boolean variables, sample point is propositional logic model
  - $A = true, B = false, a \vee \neg b$
- Proposition is disjunctions of atomic events in which it is true
  - $(a \vee b) = (a \wedge b) \vee (a \wedge \neg b) \vee (\neg a \wedge b)$
  - $\Rightarrow P(a \vee b) = P(a \wedge b) + P(a \wedge \neg b) + P(\neg a \wedge b)$

Syntax for Propositions

- Propositional or Boolean random variables
  - Example: Cavity (do I have a cavity?)
  - Cavity = true is a proposition, also written as $cavity$
- Discrete random variables (finite or infinite)
  - Example: Weather is one of {sunny, rain, cloudy, snow}
  - $Weather = rain$ is a proposition
  - Values must be exhaustive and mutually exclusive
- Continuous random variables
  - Example: Temperature = 37
- Arbitrary Boolean combinations of basic propositions

Prior Probability

- Prior or unconditional probabilities of propositions – correspond to belief prior to arrival of any (new) evidence
  - $P(cavity) = 0.1$ and $P(sunny) = 0.75$
- Probability distribution gives values for all possible assignments
  - $P(Weather) = (0.75, 0.1, 0.1, 0.05)$ (normalized i.e., sums to 1)
- Joint probability distribution for a set of r.v.s (i.e., every sample point)
  - $P(Weather, Cavity)$ = 4×2 matrix of values
- Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Conditional Probability

- Conditional or posterior probability: probability after witnessing some evidence
  - Example: $P(cavity|toothache) = 0.8$
  - Read it as Given that $toothache$ is all I know, and not If toothache then 80% change of cavity (does not represent causality)
- $P(Cavity|Toothache)$ = 2 element vector of 2-element vectors
- If we know more, e.g., $cavity$ is also given, then we have
  - $P(cavity|toothache, cavity) = 1$
    - The less specific belief is still valid after more evidence arrives, but is not always useful
- The new evidence may be irrelevant, allowing simplification
  - $P(cavity|toothache, KingsXIPunjabWin) = P(cavity|toothache) = 0.8$

Conditional Probability

- What is the probability of a cavity given a toothache?
- What is the probability of a cavity given the probe catches?
- Three Random Variables – $Cavity, Toothache$, and $Catch$
  - Each is either T or F

Inference by Enumeration (1)

- Start with the joint distribution

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

- For any proposition ϕ, sum the atomic events where it is true:
  - $P(\phi) = \sum_{\omega \mid \omega \models \phi} P(\omega)$

Inference by Enumeration (2)

- Start with the joint distribution

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

- For any proposition ϕ, sum the atomic events where it is true:
  - $P(\phi) = \sum_{\omega \mid \omega \models \phi} P(\omega)$
- $P(toothache) = 0.2$

Inference by Enumeration (3)

- Start with the joint distribution

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

- For any proposition ϕ, sum the atomic events where it is true:
  - $P(\phi) = \sum_{\omega \mid \omega \models \phi} P(\omega)$
- $P(toothache) = 0.2$
- $P(cavity \vee toothache) = 0.28$

Inference by Enumeration (4)

- Start with the joint distribution

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

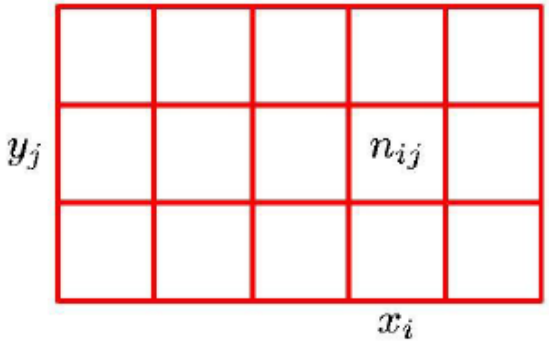
- Can also compute conditional probabilities
- $P(\neg cavity | toothache) = 0.4$

Problems with Enumeration

- Worst case time: $O(d^n)$
  - d maximum arity of the random variables
  - n - number of random variables
- Space complexity is also $O(d^n)$
  - Size of the joint distribution
- Problem: Hard/impossible to estimate all $O(d^n)$ entries of joint distribution for large problems

Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

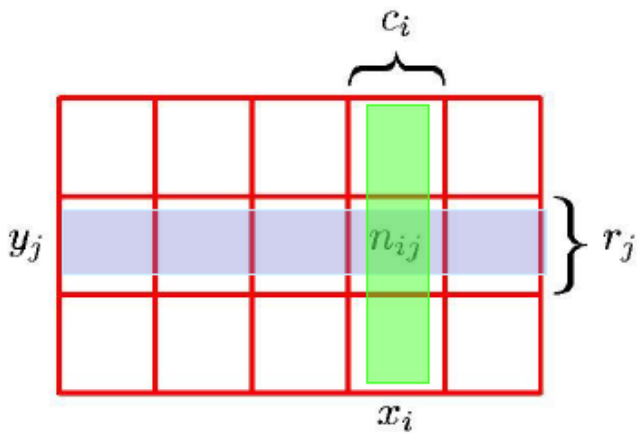


Total number of events = *N*

Marginal Probability

$$P(X = x_i) = \sum_j P(x_i, y_j) = \frac{c_i}{N}$$

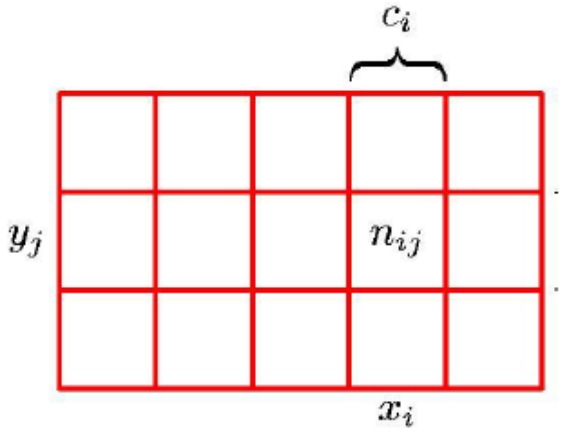
$$P(Y = y_j) = \sum_i P(x_i, y_i) = \frac{r_j}{N}$$



Summing out a variable is called *marginalization*

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$



Marginalization, Conditioning and Normalization

- Given full joint probabilities, marginalization or summing out is

$$P(Y) = \sum_{z \in Z} P(Y, z)$$

- Suppose we are given the conditional probabilities, then conditioning is

$$P(Y) = \sum_{z} P(Y|z)P(z)$$

- Normalization - α - normalization constant

$$P(X|y) = \alpha P(X, y)$$

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

$$P(\neg x|y) = \frac{P(\neg x, y)}{P(y)}$$

Quick Numerical

$P(X, Y)$

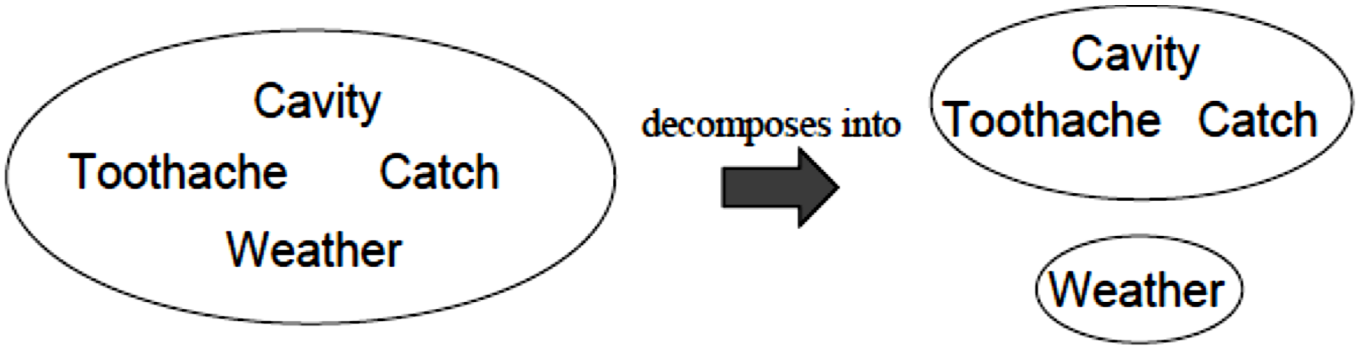
X	Y	P
x	y	0.4
x	$\neg y$	0.2
$\neg x$	y	0.3
$\neg x$	$\neg y$	0.1

- $P(x) =$
- $P(\neg x) =$
- $P(y) =$
- $P(\neg y) =$
- $P(x|y) =$
- $P(\neg x|y) =$
- $P(x|\neg y) =$
- $P(\neg x|\neg y) =$

Independence

- A and B are independent iff

$P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A, B) = P(A)P(B)$



- $P(Toothache, Catch, Cavity, Weather) = P(Toothache, Catch, Cavity)P(Weather)$
- Size of the joint probability distribution table?
- Absolute independences are rare.

Conditional Independence

- $P(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries.
- Suppose it is given that I have cavity - $Cavity = True$, then

$$P(catch|toothache, cavity) = P(catch|cavity)$$

- And similarly

$$P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$$

- Which results in $Catch$ being conditionally independent of $Toothache$ given $Cavity$.

$$P(Catch|Toothache, Cavity) = P(Catch|Cavity)$$

- Equivalent statements are

$$P(Toothache|Catch, Cavity) = P(Toothache|Cavity)$$

$$P(Toothache, Catch|Cavity)$$
$$= P(Toothache|Cavity)P(Catch|Cavity)$$

Conditional Independence

- Thus factoring the joint probability distribution

$$P(Toothache, Catch, Cavity) =$$

Conditional Independence

- Thus factoring the joint probability distribution

$$P(Toothache, Catch, Cavity)$$
$$= P(Toothache|Catch, Cavity)P(Catch, Cavity)$$
$$= P(Toothache|Cavity)P(Catch, Cavity)$$
$$= P(Toothache|Cavity)P(Catch|Cavity)P(Cavity)$$

- How many independent entries need to be stored?

Conditional Independence

- Thus factoring the joint probability distribution

$$P(Toothache, Catch, Cavity)$$
$$= P(Toothache|Catch, Cavity)P(Catch, Cavity)$$
$$= P(Toothache|Cavity)P(Catch, Cavity)$$
$$= P(Toothache|Cavity)P(Catch|Cavity)P(Cavity)$$

- How many independent entries need to be stored? -5
- Conditional Independence reduces the size of the representation in JPT that is exponential in N to linear in N.

Bayes' Rule

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$\Rightarrow$ Bayes' rule $P(a|b) = \frac{P(b|a)P(a)}{P(b)}$

or in distribution form

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X|Y)\mathbf{P}(Y)$$

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Bayes' rule is used to Compute *Diagnostic* Probability from *Causal* Probability

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g. let M be meningitis, S be stiff neck

P(M) = 0.0001,

P(S) = 0.1,

P(S|M)= 0.8 (note: these can be estimated from patients)

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

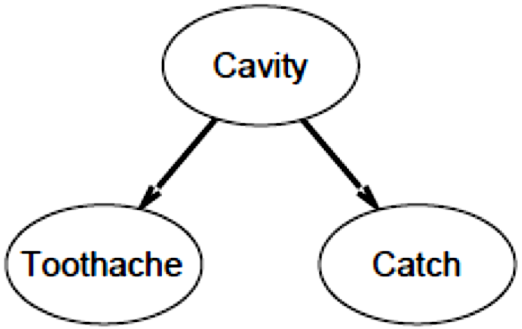
(But chance of M did increase from 0.0001 to 0.0008 given stiff neck)

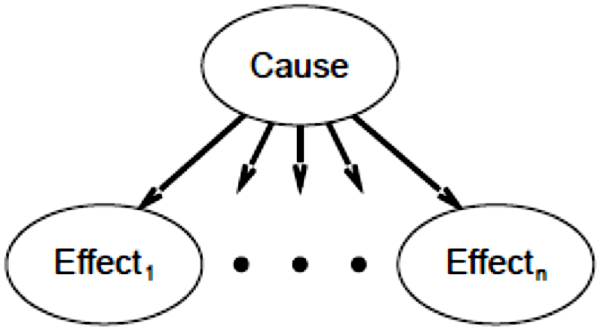
Bayes' Rule and conditional independence

$$\begin{aligned}
&\mathbf{P}(Cavity|toothache \wedge catch) \\
&= \alpha\,\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity) \\
&= \alpha\,\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

This is an example of a naive Bayes model:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\prod_i \mathbf{P}(Effect_i|Cause)$$





Total number of parameters is **linear** in n

Outline

- Probability theory - basics
- Probabilistic Inference
  - Inference by enumeration
- Conditional Independence
- Bayesian Networks
- Inference in Bayesian Networks
  - Exact Inference
  - Approximate Inference

Bayesian Networks

Material adapted from Dan Klein

Probabilistic Models

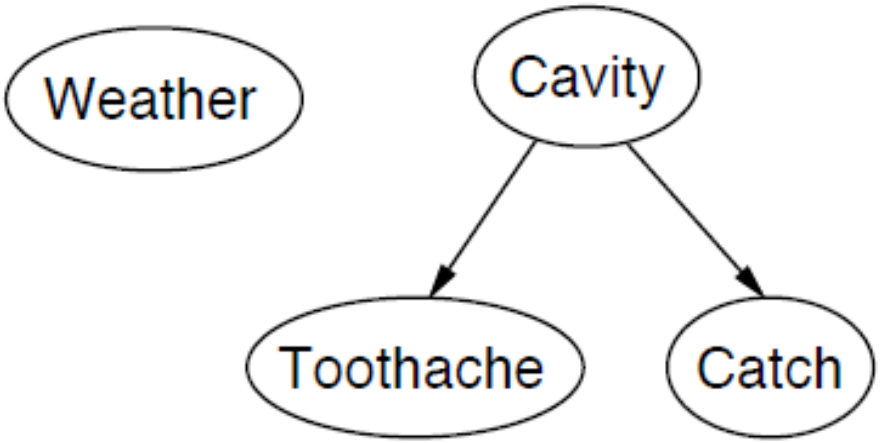
- Models that describe how a portion of the real world works
  - Simplifications of the real world
- What do we do with probabilistic models
  - Agents reason about unknown variables, given evidence
    - Reason about unknown variables
      - Explanation (diagnostic reasoning)
      - Prediction (causal reasoning)
    - Value of information

Bayesian Networks

- Graphical notation for conditional independence assumptions
  - Compact specification of full joint distributions
- Syntax
  - One node per random variable
  - A directed edge from one node to another (influences)
  - A conditional distribution at each node given its parents - $P\left(X_i | Parents(X_i)\right)$
    - In the simplest case, represented as a conditional probability distribution table (CPT)
  - Directed Acyclic Graph

Example (1)

- Topology of the network encodes the conditional independence assumptions

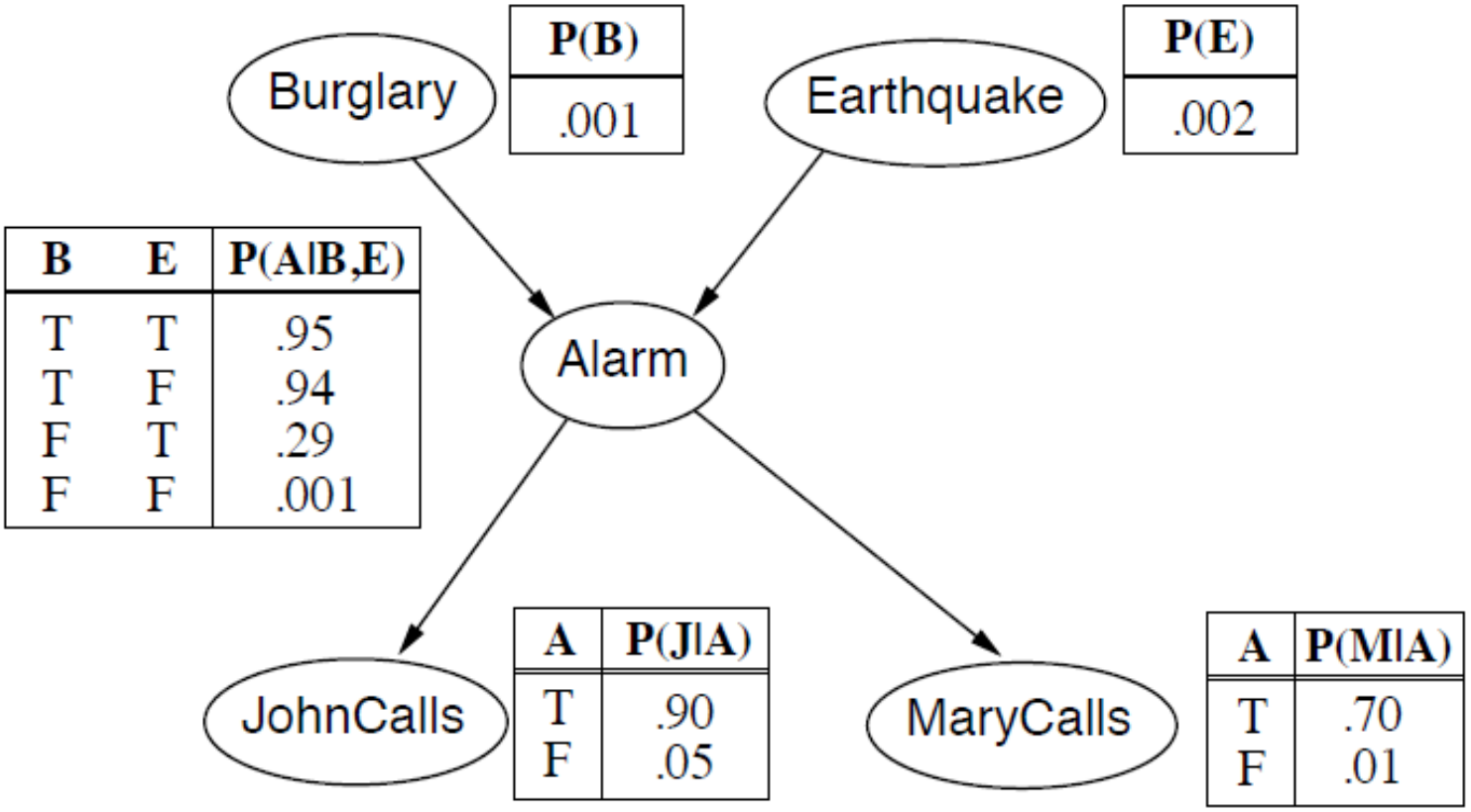


- Weather is independent of other variables
- $Toothache$ and $Catch$ are conditionally independent given $Cavity$

Example (2)

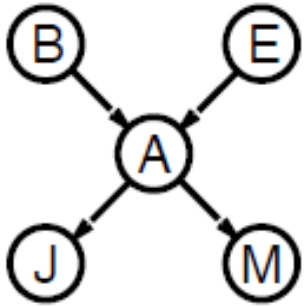
- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes the alarm is set off by earthquakes. Is there a burglar?
- Variables – $Burglar, Earthquake, Alarm, JohnCalls, MaryCalls$
- Network topology reflects the "casual" knowledge
  - Burglar can set off the alarm
  - Earthquake can set off the alarm
  - The alarm can cause John to call
  - The alarm can cause Mary to call

Example (2)


Burglary
P(B)
.001
Earthquake
P(E)
.002
B E P(A|B,E)
T T .95
T F .94
F T .29
F F .001
Alarm
JohnCalls
A P(J|A)
T .90
F .05
MaryCalls
A P(M|A)
T .70
F .01

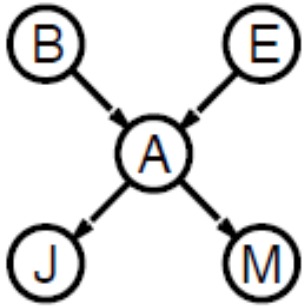
Compactness

- A CPT for X_i with k Boolean parents has 2^k rows for the combination of parent values.
- Each row requires one number p for $X_i = True$
  - $X_i = False$ can be computed from p
- If each variable has no more than k parents, the complete network requires $O(n2^k)$ entries
  - Thus the representation grows linearly in n, in contrast to full joint probability distribution table.
- For the burglary net



Global Semantics

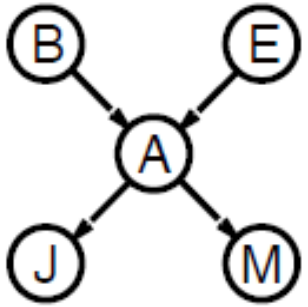
- Define the full joint distribution as a product of the local conditional distributions
- $P(X_1, \ldots X_N) = \prod_{i=1}^{N} P\left(X_i | Parents(X_i)\right)$
- For example
- $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) =$



Global Semantics

- Define the full joint distribution as a product of the local conditional distributions
- $P(X_1, \ldots X_N) = \prod_{i=1}^{N} P\left(X_i | Parents(X_i)\right)$
- For example
- $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$
$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \approx 0.0006$$



Bayes Nets – Causality?

- When BNs reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean
  - Topology may happen to encode the causal structure
  - *Topology really encodes conditional independence*.

Constructing Bayesian Networks

- A series of locally testable assertions of conditional independence guarantees global semantics (full joint)
- Choose an ordering of the variables - $X_1, \ldots, X_N$
- For $i = 1, \ldots, N$
  - Add variable X_i to the network
  - Select parents from $X_1, \ldots, X_{i-1}$ such that

$$P\big(X_i | Parents(X_i)\big) = P(X_i | X_1, \ldots, X_{i-1})$$

- This choice guarantees global semantics

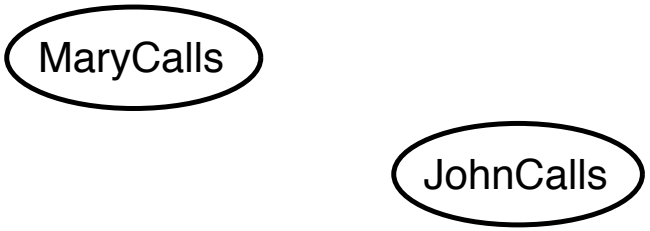
$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_1, \ldots, X_{i-1}) = \prod_{i=1}^{N} P(X_i | Parents(X_i))$$

Example

- Suppose we choose the ordering M, J, A, B, E

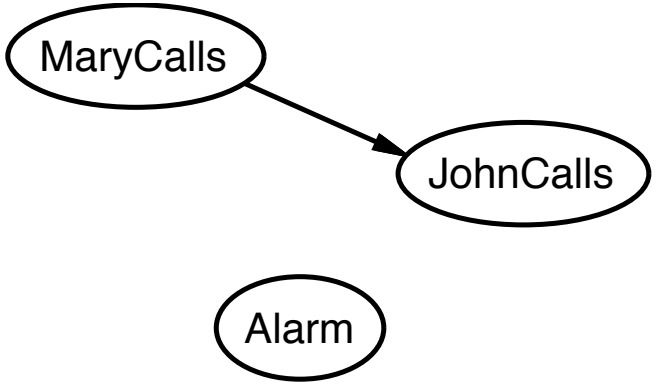
Example

- Suppose we choose the ordering M, J, A, B, E



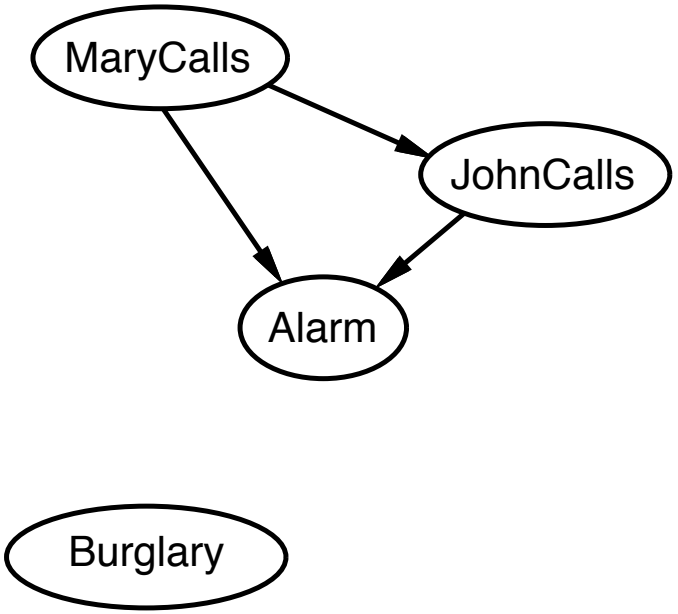
Example

- Suppose we choose the ordering M, J, A, B, E



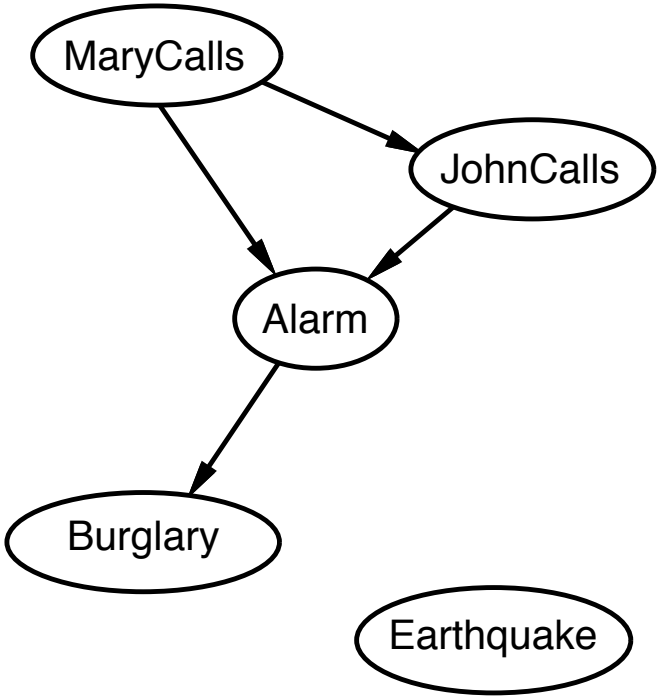
Example

- Suppose we choose the ordering M, J, A, B, E



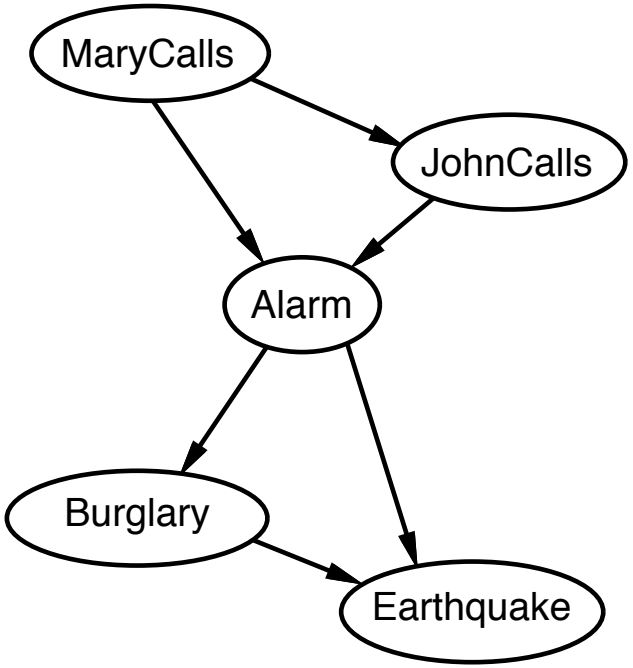
Example

- Suppose we choose the ordering M, J, A, B, E



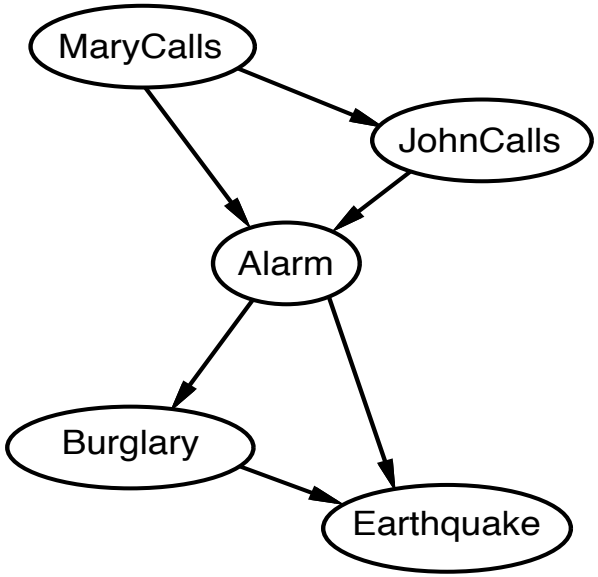
Example

- Suppose we choose the ordering M, J, A, B, E



Example

- Deciding conditional independence is hard in non-causal directions
  - Causal models and conditional independencies seems hardwired in humans
- Assessing conditional independencies in non-causal directions is hard
- Size of the network?

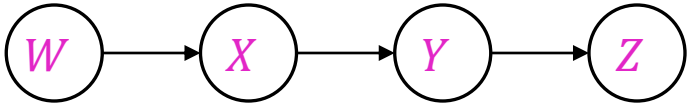


Bayes Nets: Conditional Independence

- Fundamental Assumption

$$P(x_i|x_1, \ldots, x_{i-1}) = P(x_i|Parents(X_i))$$

- Beyond the Bayes net Chain rule conditional independence assumptions,
  - Additional conditional independencies, that can be read off the graph.
  - Example

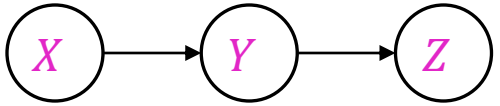


CI assumptions directly from simplifications in chain rule:

Other implied CI assumptions

Bayes Nets: Conditional Independence (2)

- Given a BN, a common yet important question would be
  - Are two nodes independent given certain evidence?
    - If yes, prove using probabilistic algebra
    - If no, give a counter example.
  - Consider the following example
    - Are X and Z independent?

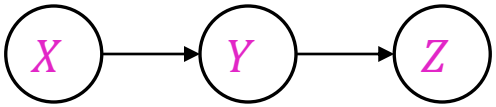


D-Separation

- Answer independence queries regarding variables in a BN.
  - Study independence properties for triples
    - Causal chain, common cause and common effect
  - Analyze complex cases in terms of member triples
  - D-Separation: a condition/algorithm for answering these queries.

D-Separation – Causal Chains

- X – Transfer money, Y - non-zero balance, Z – withdraw money



$$P(X,Y,Z) = P(X)P(Y|X)P(Z|Y)$$

- Is it guaranteed that X is independent of Z?
  - No! - Proof by counter example
    - Consider the case- Transferring money causes non-zero balance causes withdrawal

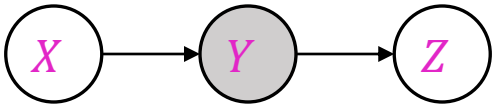
$$P(y|x) = 1, P(\neg y|\neg x) = 1$$
$$P(z|y) = 1, P(\neg z|\neg y) = 1$$

And the evidence

$$P(x) = 0.8$$
$$P(x,z) \neq P(x)P(z)$$

D-Separation – Causal Chains

- X – Transfer money, Y - non-zero balance, Z – withdraw money
- Is it guaranteed that X is independent of Z given Y?

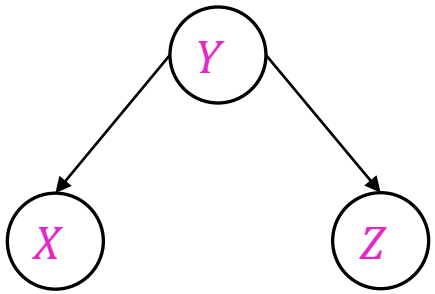


$$P(X, Y, Z) = P(X)P(Y|X)P(Z|Y)$$

- Evidence along the chain "blocks" the influence

D-Separation – Common Cause

- X – thin attendance, Y- project is due, Z – students are busy



$$P(X,Y,Z) = P(Y)P(X|Y)P(Z|Y)$$

- Is it guaranteed that X is independent of Z?
  - No! - Proof by counter example
    - Consider the scenario Project due causes both thin attendance and students busy

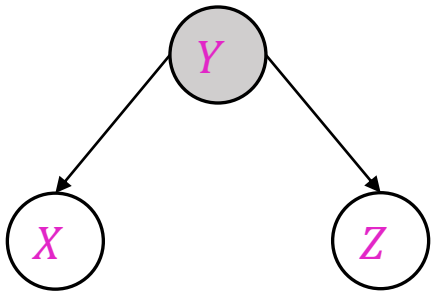
$$P(x|y) = 1, P(\neg x|\neg y) = 1$$
$$P(z|y) = 1, P(\neg z|\neg y) = 1$$

And the evidence

$$P(y) = 0.8$$
$$P(x,z) \neq P(x)P(z)$$

D-Separation – Common Cause

- X – thin attendance, Y- project is due, Z – students are busy
- Is it guaranteed that X is independent of Z given Y?

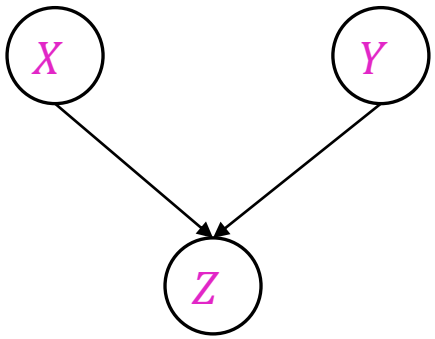


$$P(X, Y, Z) = P(Y)P(X|Y)P(Z|Y)$$

- Observing the cause blocks the influence between effects

D-Separation – Common Effect

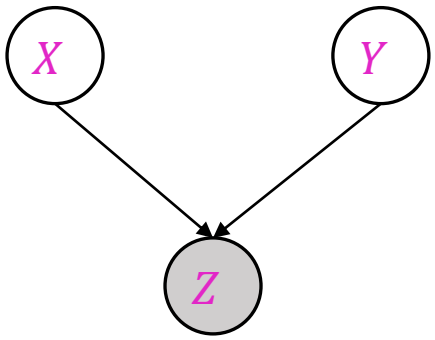
- X – Raining, Y- Cricket Match, Z - Traffic
- Is it guaranteed that X is independent of Y?



$$P(X, Y, Z) = P(X)P(Y)P(Z|X, Y)$$

D-Separation – Common Effect

- X – Raining, Y- Cricket Match, Z - Traffic

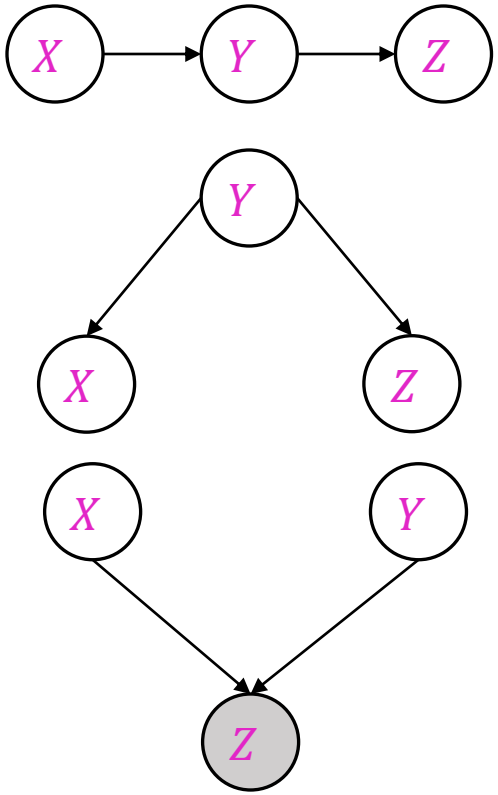


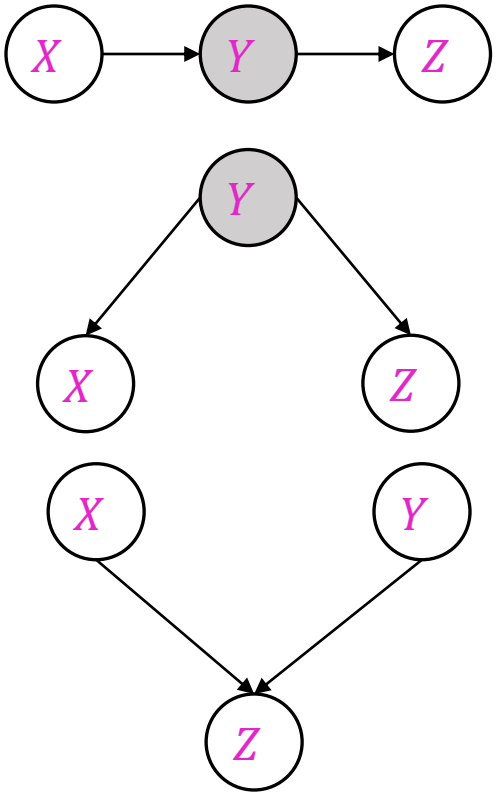
$$P(X, Y, Z) = P(X)P(Y)P(Z|X, Y)$$

- Is it guaranteed that X is independent of Y given Z?
  - No, seeing traffic puts the rain and the cricket match as explanation
- This is different from the other cases
  - Observing an effect activates influence between possible causes

D-Separation – Active/Inactive Paths

- A path is active if each triple in the path is active - dependencies between nodes.



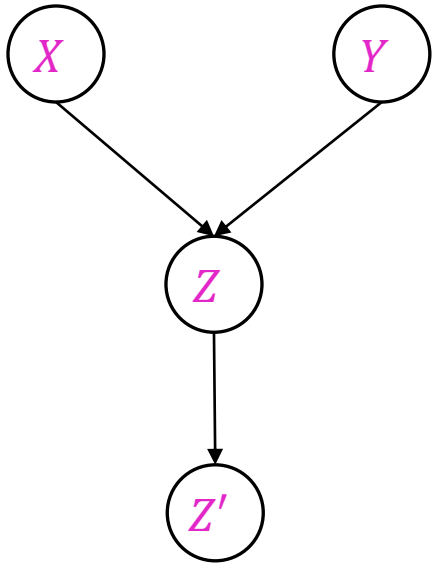


D-separation

- Given a BN, are two variables X and Y independent (given evidence Z)?
  - Yes if X and Y are ‘d-separated’ by Z
  - All paths are inactive implies independence
    - A single inactive segment is sufficient to block the path
  - If one or more paths are active, then independence is not guaranteed
- Analyze the graph
  - Check for the three canonical forms all through the path

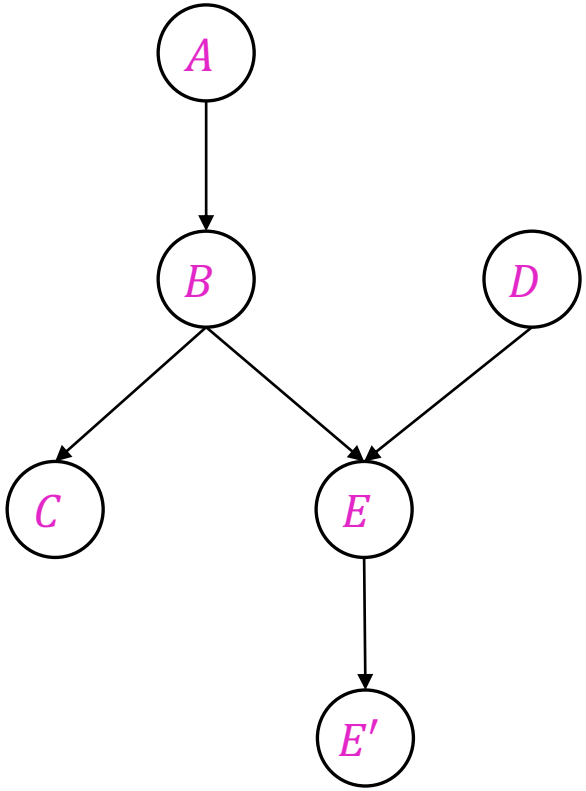
Example

- X independent of Y
- X independent of Y given Z
- X independent of Y given Z'

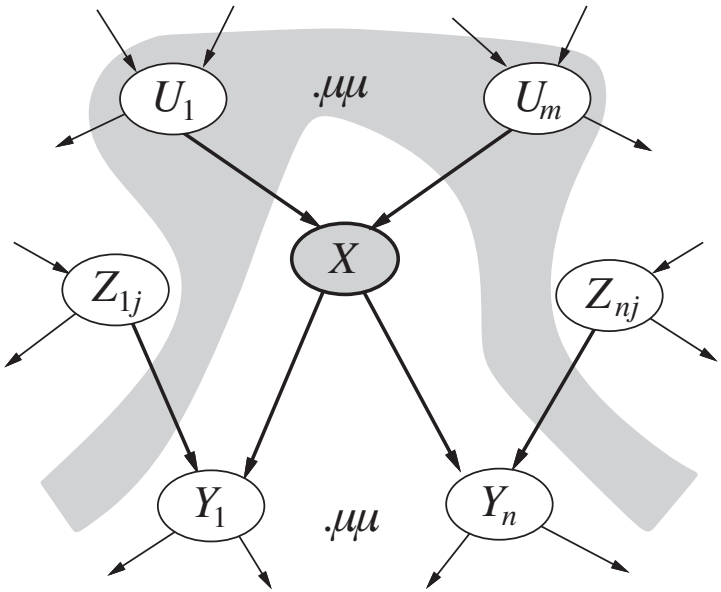


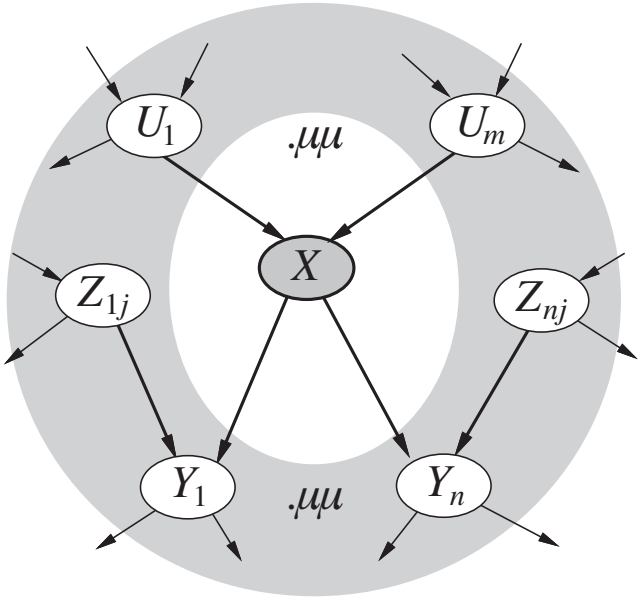
Example (2)

- C independent of D
- A independent of E' given E
- A independent of D
- A independent of D given E
- A independent of D given B, E



Conditional Independence in BNs – Markov Blanket





Outline

- Probability theory - basics
- Probabilistic Inference
  - Inference by enumeration
- Conditional Independence
- Bayesian Networks
- Inference in Bayesian Networks
  - Exact Inference
  - Approximate Inference

Probabilistic Inference in BNs

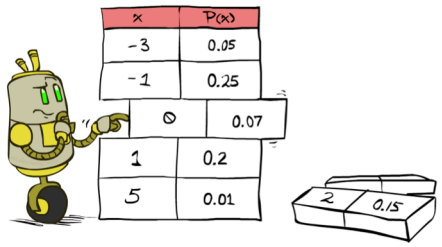
- Graphical independence representation yields efficient inference schemes
- We generally want to compute
  - $P(X|E)$, where E is the evidence from the sensory measurements (known values for variables)
  - Sometimes, we may want to compute just $P(X)$
- Three approaches
  - Enumeration
  - Variable Elimination
  - Approximate Inference
    - sampling

Inference by Enumeration

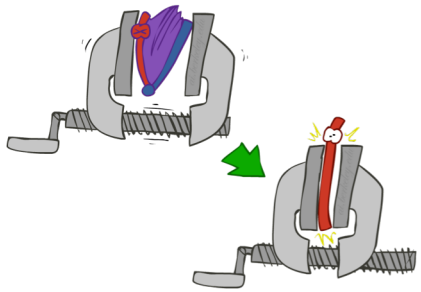
- General case - the set of random variables $X_1, \ldots, X_n$ is broken into three categories
  - Evidence variables - $E_1, \ldots, E_k = e_1, \ldots, e_k$
  - Query variables - $Q_1, \ldots, Q_L = Q$
  - Hidden variables - $H_1, \ldots, H_r$
- We want to estimate - $P(Q|e_1, \ldots, e_k)$

Inference by Enumeration

- We want to estimate - $P(Q|e_1, \dots, e_k)$
- Step 1 – select the consistent entries with the evidence



- Step 2 - sum of the hidden variables to get the joint of Query and evidence
  - $P(Q, e_1, \dots, e_k) = \sum_{H_1, \dots H_r} P(Q, h_1, \dots, h_r, e_1, \dots, e_k)$



- Step 3 – Normalize
  - $Z = \sum_q P(Q, e_1, \dots, e_k)$
  - $P(Q|e_1, \dots, e_k) = \frac{1}{Z} P(Q, e_1, \dots, e_k)$

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

Marginalization, Conditioning and Normalization

- Given full joint probabilities, marginalization or summing out is

$$P(Y) = \sum_{z \in Z} P(Y, z)$$

- Suppose we are given the conditional probabilities, then conditioning is

$$P(Y) = \sum_{z} P(Y|z)P(z)$$

- Normalization - α - normalization constant

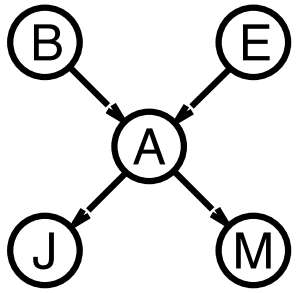
$$P(X|y) = \alpha P(X, y)$$

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

$$P(\neg x|y) = \frac{P(\neg x, y)}{P(y)}$$

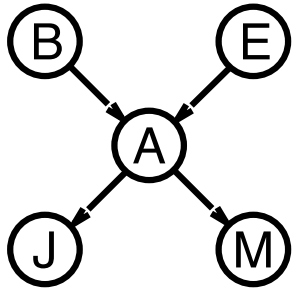
Inference by Enumeration

- Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation.
- Consider the simple query on the burglary network
  - $P(B|j,m)$



Inference by Enumeration (2)

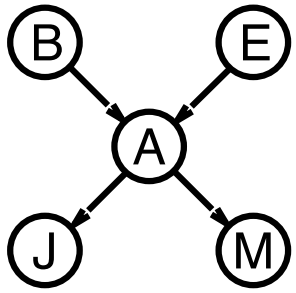
- Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation.
- Consider the simple query on the burglary network
  - $P(B|j,m) = P(B,j,m)/P(j,m)$
  - $= \frac{1}{Z} P(B,j,m)$
  - $= \frac{1}{Z} \sum_e \sum_a P(B,e,a,j,m)$



- Rewrite full joint entries using product of the CPT entries

Inference by Enumeration (3)

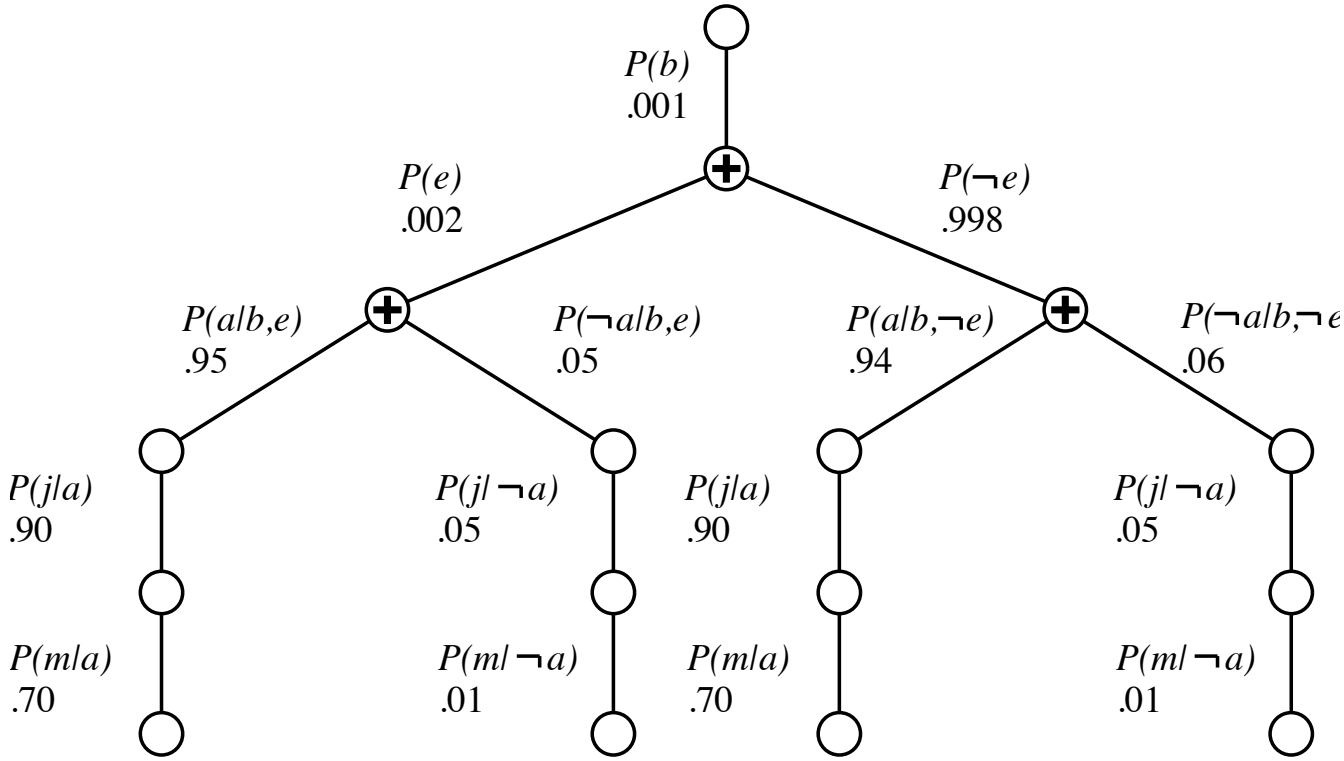
- Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation.
- Consider the simple query on the burglary network
  - $P(B|j,m) = P(B,j,m)/P(j,m)$
  - $= \frac{1}{Z} P(B,j,m)$
  - $= \frac{1}{Z} \sum_e \sum_a P(B,e,a,j,m)$



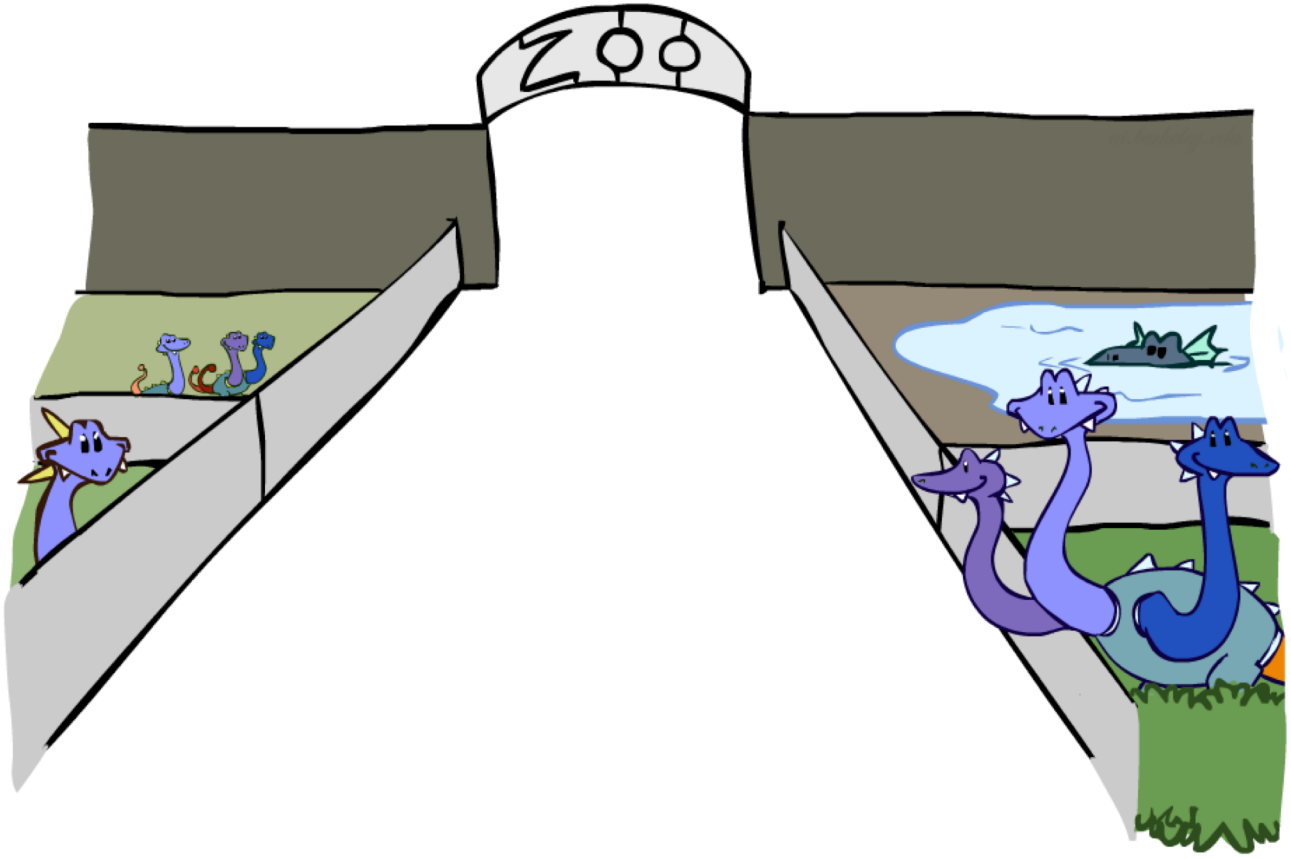
- Rewrite full joint entries using product of the CPT entries
  - $= \frac{1}{Z} P(B) \sum_e P(e) \sum_a P(a|B,e)\, P(j|a) P(m|a)$

Evaluation Tree

- Recursive depth first enumeration
  - $= \frac{1}{Z} P(B) \sum_e P(e) \sum_a P(a|B,e)\, P(j|a) P(m|a)$



Factor Zoo



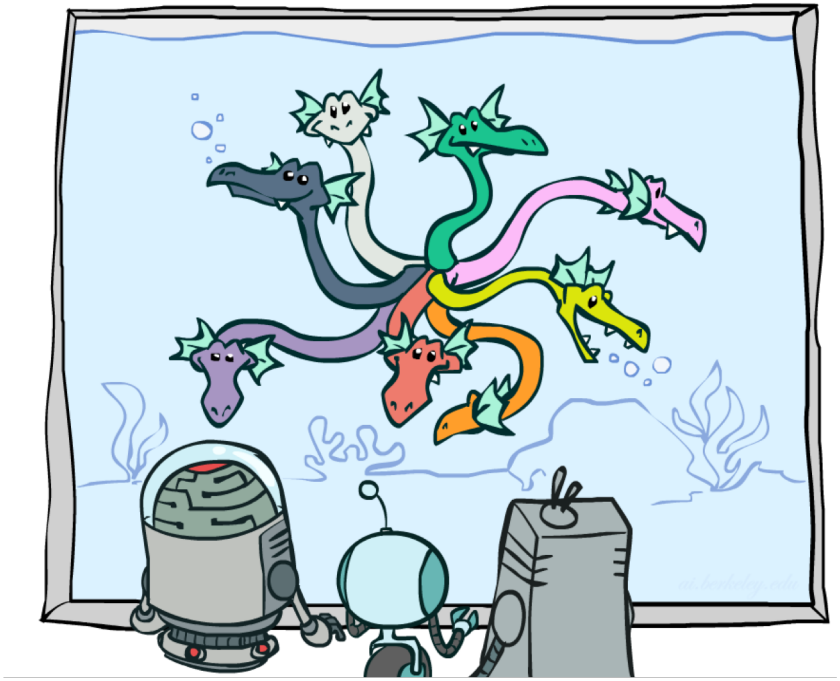
Factor Zoo I

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

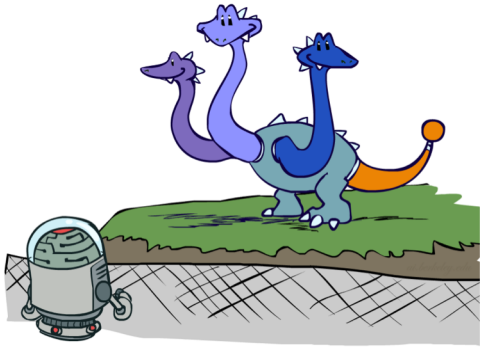
$P(cold, W)$

T	W	P
cold	sun	0.2
cold	rain	0.3



- Joint distribution: $P(X, Y)$
  - Entries $P(x, y)$ for all x, y
  - Sums to 1
- Selected joint: $P(x, Y)$
  - A slice of the joint distribution
  - Entries $P(x, y)$ for fixed x, all y
  - Sums to $P(x)$
- Number of capitals determines dimensionality of the table

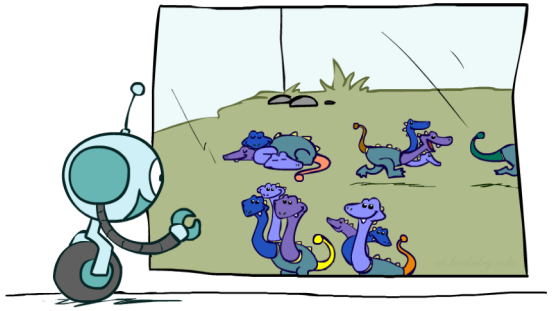
Factor Zoo II



$P(W|cold)$

T	W	P
cold	sun	0.4
cold	rain	0.6

- **Single conditional: $P(Y \mid x)$**
  - Entries $P(y \mid x)$ for fixed x, all y
  - Sums to 1

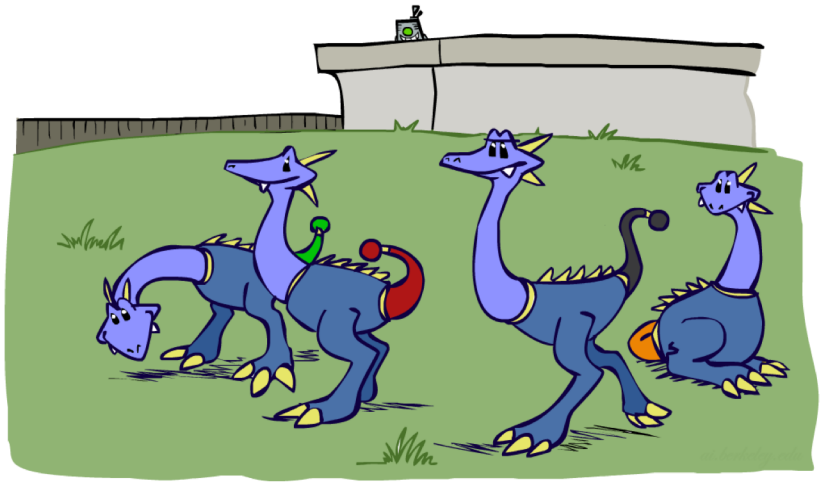


$P(W|T)$

T	W	P
hot	sun	0.8
hot	rain	0.2
cold	sun	0.4
cold	rain	0.6

- **Family of conditionals: $P(Y \mid X)$**
  - Multiple conditionals
  - Entries $P(y \mid x)$ for all x, y
  - Sums to $|X|$

Factor Zoo III



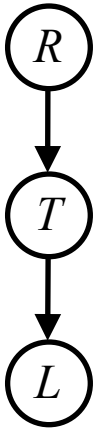
$P(rain|T)$

T	W	P
hot	rain	0.2
cold	rain	0.6

- Specified family: $P(y \mid X)$
  - Entries $P(y \mid x)$ for fixed y, but for all x
  - Sums to … who knows!

Example: Traffic Domain

- Random Variables
  - R: Raining
  - T: Traffic
  - L: Late for flight!
- $P(L) =$
  - $= \sum_{r,t} P(r, t, L)$
  - $= \sum_{r,t} P(r)P(t|r)P(L|t)$



$P(R)$

+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

- Procedure tracks 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
  - For example if we know $L = true$,

$P(R)$

+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

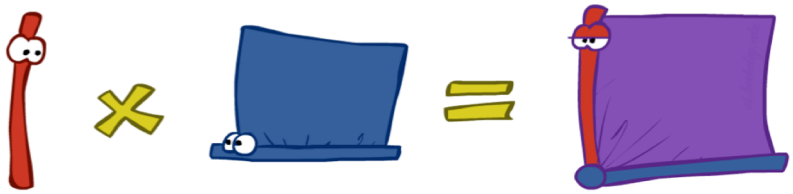
$P(+\ell|T)$

+t	+l	0.3
-t	+l	0.1

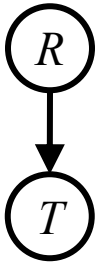
- Procedure – Join all factors and eliminate hidden variables

Operation 1 – Join Factors

- Similar to database join
- Get all factors over the joining variable
  - Build a new factor over the union of the variables involved



- Example – Join on R



$$P(R) \times P(T|R) \rightarrow P(R,T)$$

+r	0.1
-r	0.9

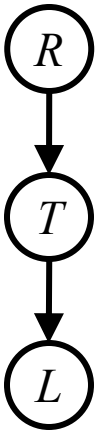
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

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

R,T

  - Computation of each entry is a pointwise product
    - $\forall r,t\ P(r,t) = P(r)\,.\,P(t|r)$

Example: Multiple Joins



$P(R)$

+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

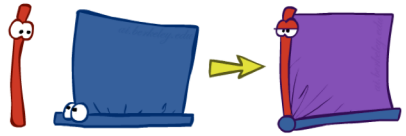
Join R

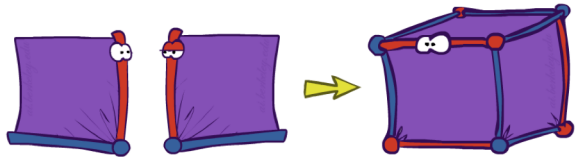
$P(R,T)$

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

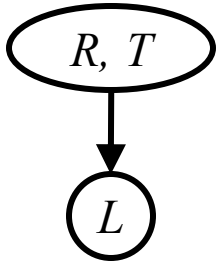
$P(L|T)$

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





Join T



$P(R,T,L)$

+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:

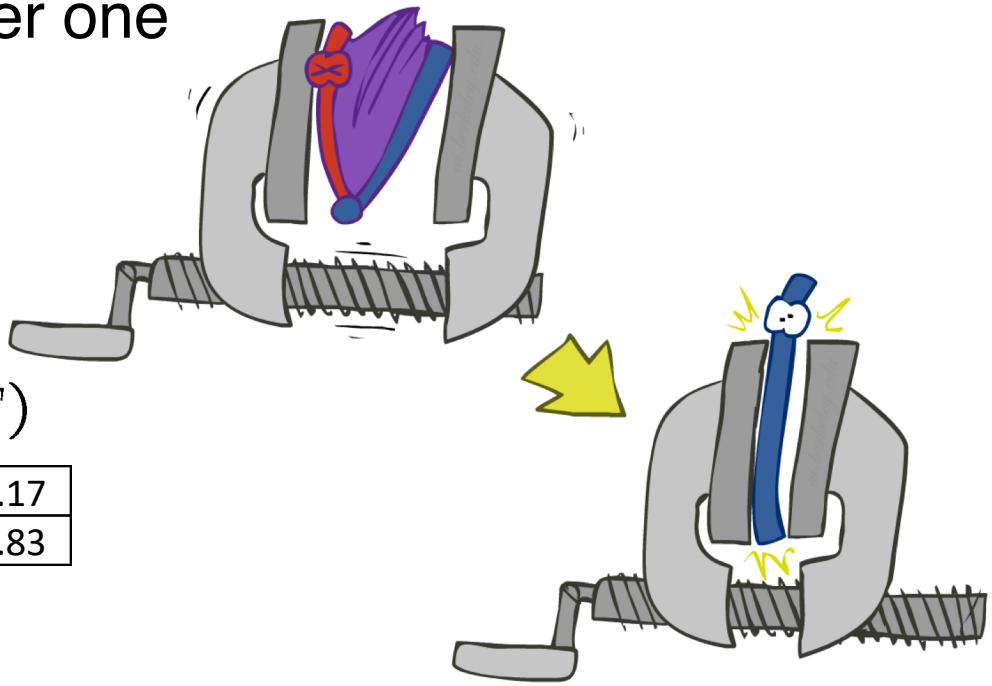
$P(R,T)$

+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

+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

Sum out R

T, L

$P(T, L)$

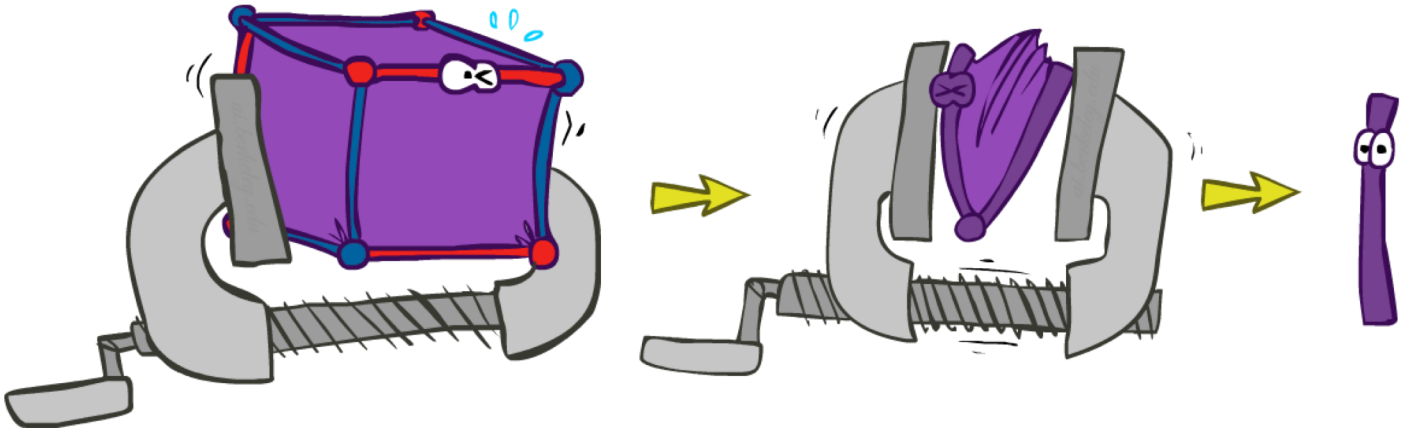
+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T

L

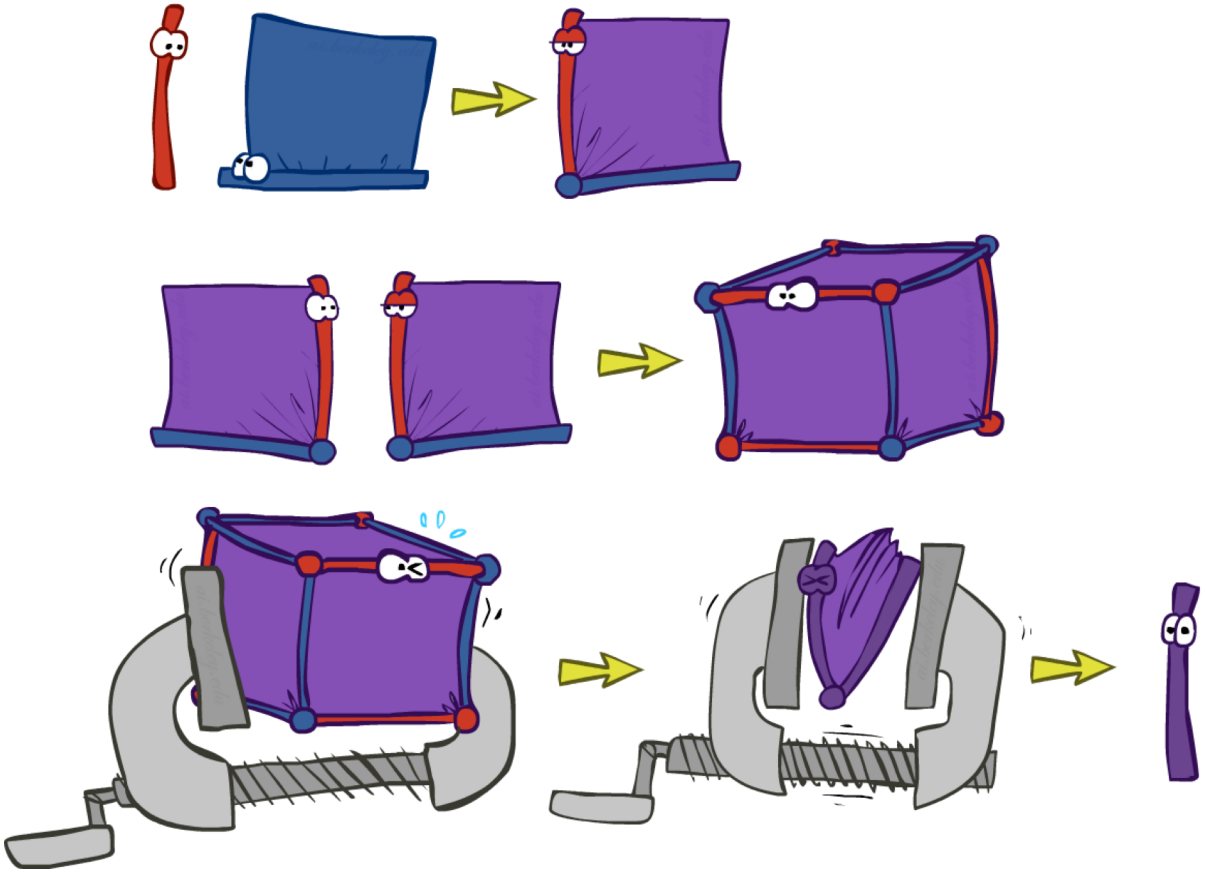
$P(L)$

+l	0.134
-l	0.866

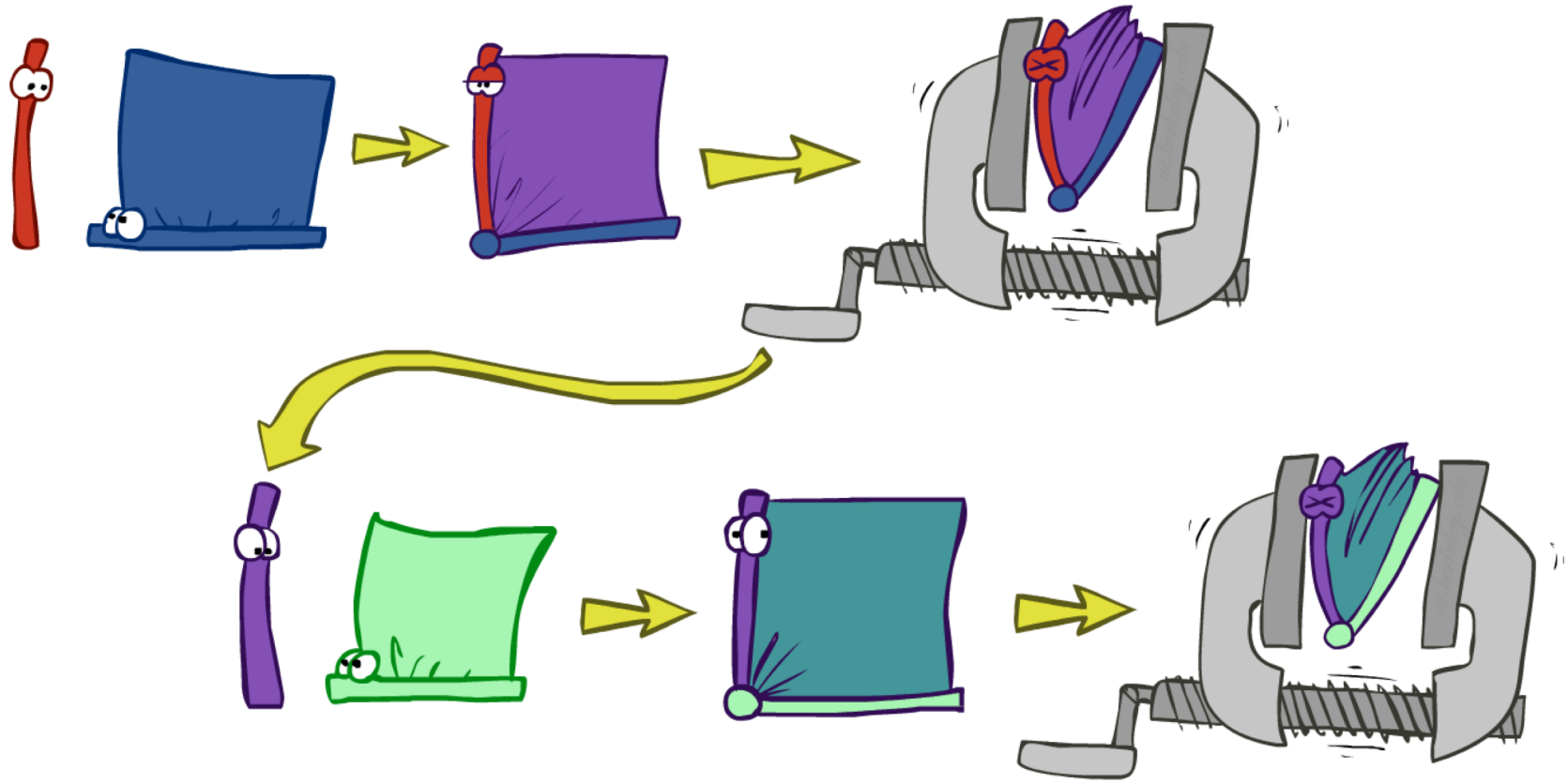


Inference by Enumeration

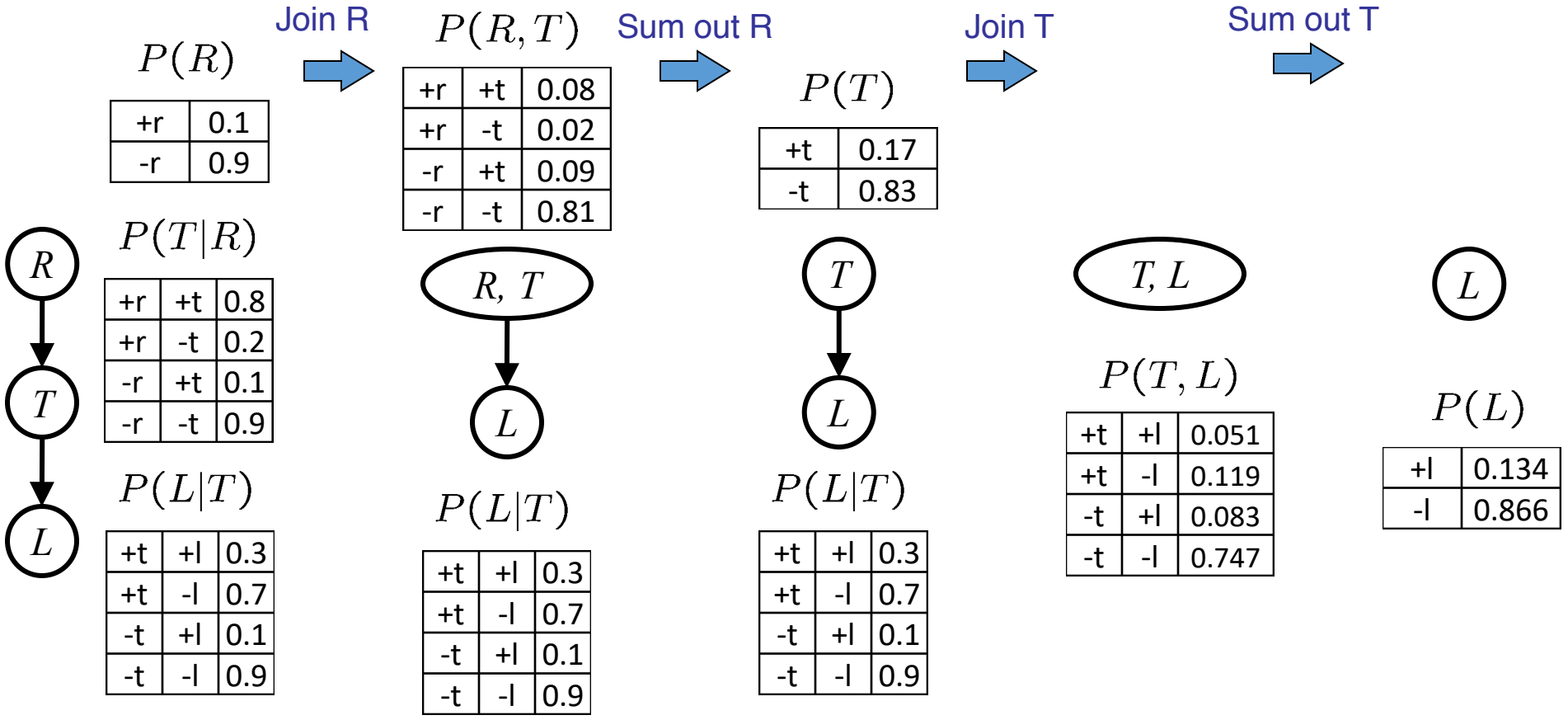
- Multiple joins and multiple eliminations



Marginalizing Early (= Variable Elimination)

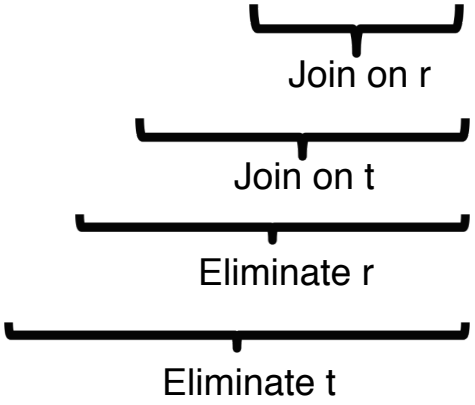


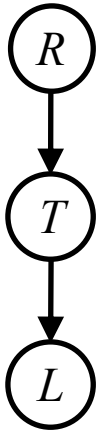
Marginalizing Early! (aka VE)



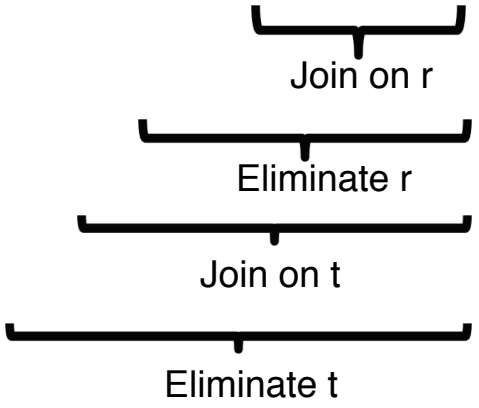
Traffic Domain

- Inference by enumeration
- $P(L) =$
  - $= \sum_{r,t} P(r,t,L)$
  - $= \sum_{r,t} P(L|t)P(r)P(t|r)$





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



Evidence I

- 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 variables other than query + evidence

Evidence II

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

$P(+r, L)$

+r	+l	0.026
+r	-l	0.074

Normalize

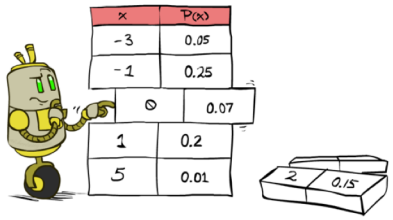
$P(L \mid +r)$

+l	0.26
-l	0.74

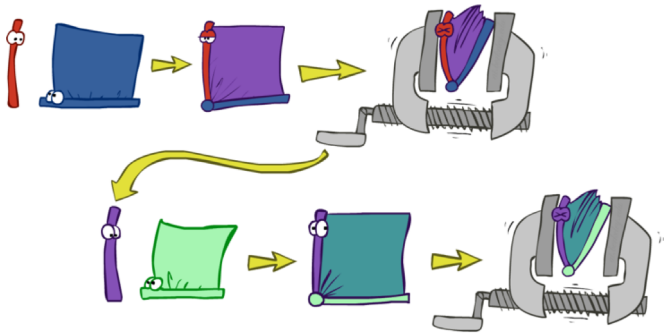
- To get our answer, just normalize this!
- That 's it!

General Variable Elimination

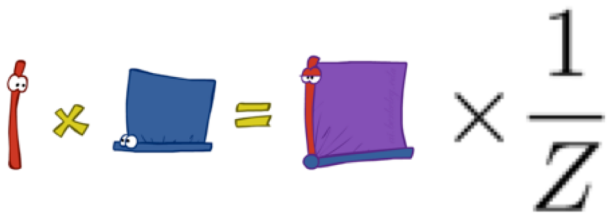
- Query: $P(Q|e_1, \ldots, e_k)$
- Start with initial factors:
  - Local CPTs (but instantiated by evidence)



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

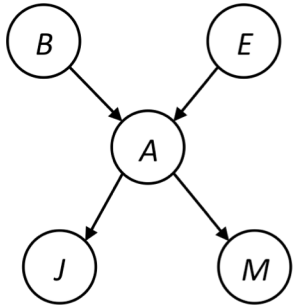


- Join all remaining factors and normalize



Example

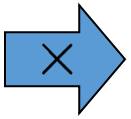
$$P(B|j,m) \propto P(B,j,m)$$



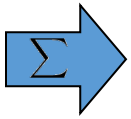
$P(B)$	$P(E)$	$P(A\|B,E)$	$P(j\|A)$	$P(m\|A)$

Choose A

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



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

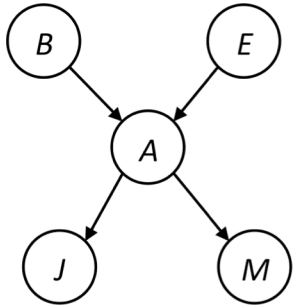


$P(j,m|B,E)$

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

Example

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



Choose E

$P(E)$

$P(j,m|B,E)$ → × → $P(j,m,E|B)$ → $\sum$ → $P(j,m|B)$

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

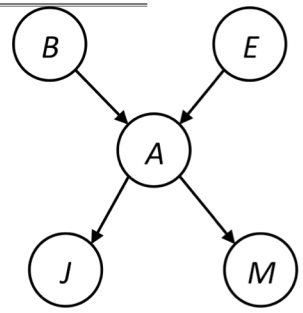
Finish with B

$P(B)$

$P(j,m|B)$ → × → $P(j,m,B)$ → Normalize → $P(B|j,m)$

Same Example in Equations

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



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

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

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

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

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

$$= \sum_{e} P(B)P(e)f_1(j,m|B,e)$$

$$= P(B)\sum_{e} P(e)f_1(j,m|B,e)$$

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

marginal can be obtained from joint by summing out

use Bayes' net joint distribution expression

use x*(y+z) = xy + xz

joining on a, and then summing out gives f_1

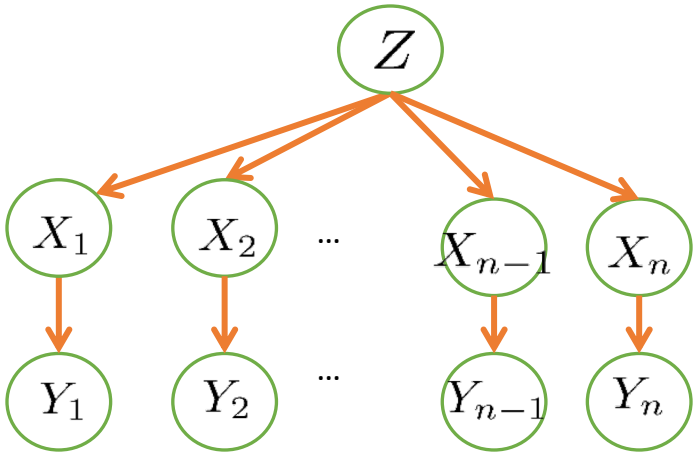
use x*(y+z) = xy + xz

joining on e, and then summing out gives f_2

All we are doing is exploiting uwy + uwz + uxy + uxz + vwy + vwz + vxy +vxz = (u+v)(w+x)(y+z) to improve computational efficiency!

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?



Another Variable Elimination Example

Query: $P(X_3|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

Start by inserting evidence, which gives the following initial factors:

$$P(Z), P(X_1|Z), P(X_2|Z), P(X_3|Z), P(y_1|X_1), P(y_2|X_2), P(y_3|X_3)$$

Eliminate X_1, this introduces the factor $f_1(y_1|Z) = \sum_{x_1} P(x_1|Z)P(y_1|x_1)$, and we are left with:

$$P(Z), P(X_2|Z), P(X_3|Z), P(y_2|X_2), P(y_3|X_3), f_1(y_1|Z)$$

Eliminate X_2, this introduces the factor $f_2(y_2|Z) = \sum_{x_2} P(x_2|Z)P(y_2|x_2)$, and we are left with:

$$P(Z), P(X_3|Z), P(y_3|X_3), f_1(y_1|Z), f_2(y_2|Z)$$

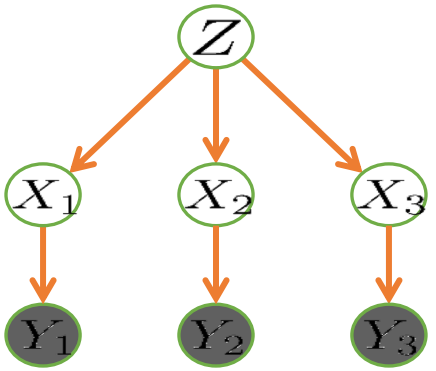
Eliminate Z, this introduces the factor $f_3(y_1, y_2, X_3) = \sum_z P(z)P(X_3|z)f_1(y_1|Z)f_2(y_2|Z)$, and we are left with:

$$P(y_3|X_3), f_3(y_1, y_2, X_3)$$

No hidden variables left. Join the remaining factors to get:

$$f_4(y_1, y_2, y_3, X_3) = P(y_3|X_3), f_3(y_1, y_2, X_3)$$

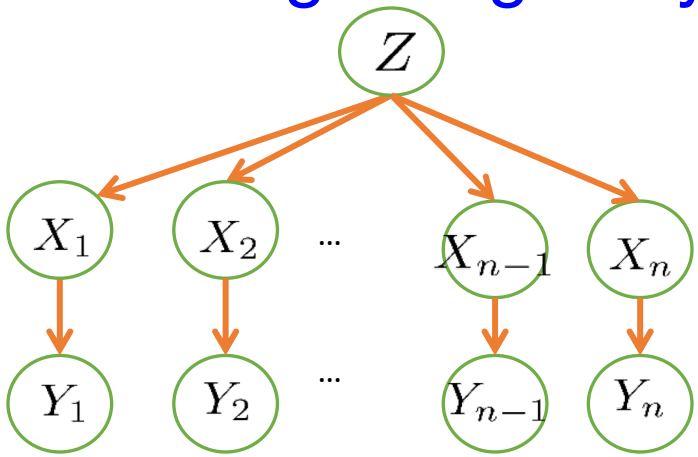
Normalizing over X_3 gives $P(X_3|y_1, y_2, y_3) = f_4(y_1, y_2, y_3, X_3)/\sum_{x_3} f_4(y_1, y_2, y_3, x_3)$



Computational complexity critically depends on the largest factor being generated in this process. Size of factor = number of entries in table. In example above (assuming binary) all factors generated are of size 2 --- as they all only have one variable (Z, Z, and X_3 respectively).

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



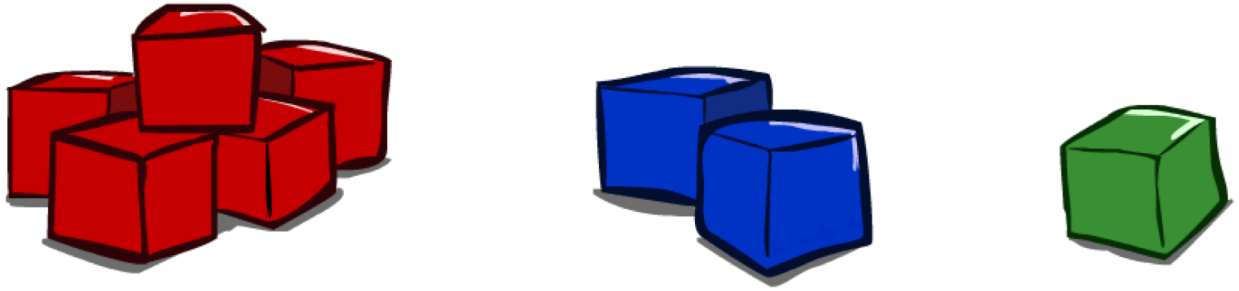
VE: Computational and Space Complexity

- 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 2n vs. 2
- Does there always exist an ordering that only results in small factors?
  - No!
- Exact inference in BNs is NP-Hard

Outline

- Probability theory - basics
- Probabilistic Inference
  - Inference by enumeration
- Conditional Independence
- Bayesian Networks
- Inference in Bayesian Networks
  - Exact Inference
  - Approximate Inference

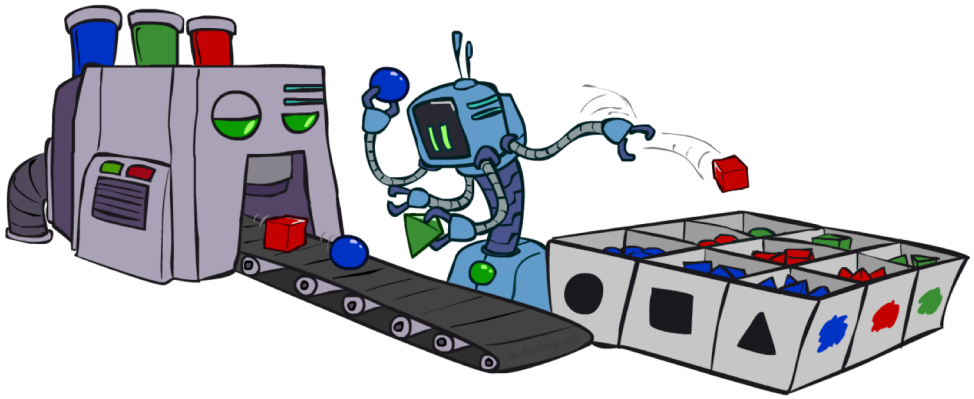
Approximate Inference: Sampling



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?
  - Inference: getting a sample is faster than computing the right answer (e.g. with variable elimination)



Sampling Basics

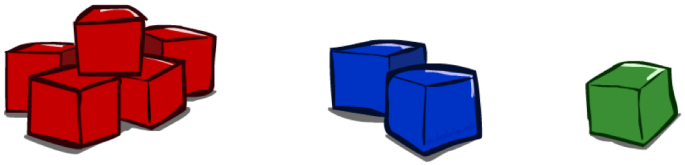
- Sampling from given distribution
  - Step 1: Get sample u from uniform distribution over [0, 1)
  - E.g. random() in python

C	P(C)
red	0.6
green	0.1
blue	0.3

$$0 \leq u < 0.6, \rightarrow C = red$$
$$0.6 \leq u < 0.7, \rightarrow C = green$$
$$0.7 \leq u < 1, \rightarrow C = blue$$

- Step 2: Convert this sample u into an outcome for the given distribution by having each outcome associated with a sub-interval of [0,1) with sub-interval size equal to probability of the outcome

- If random() returns *u* = 0.83, then our sample is *C* = blue
- E.g, after sampling 8 times:



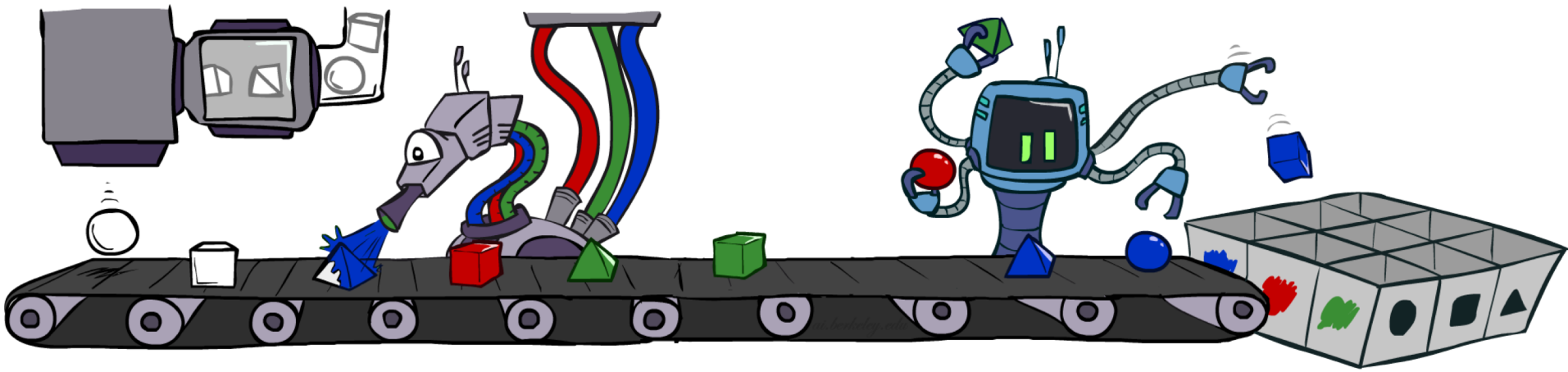
Sampling in Bayes' Nets

- Prior Sampling (Direct Sampling)
- Rejection Sampling
- Likelihood Weighting
- Gibbs Sampling

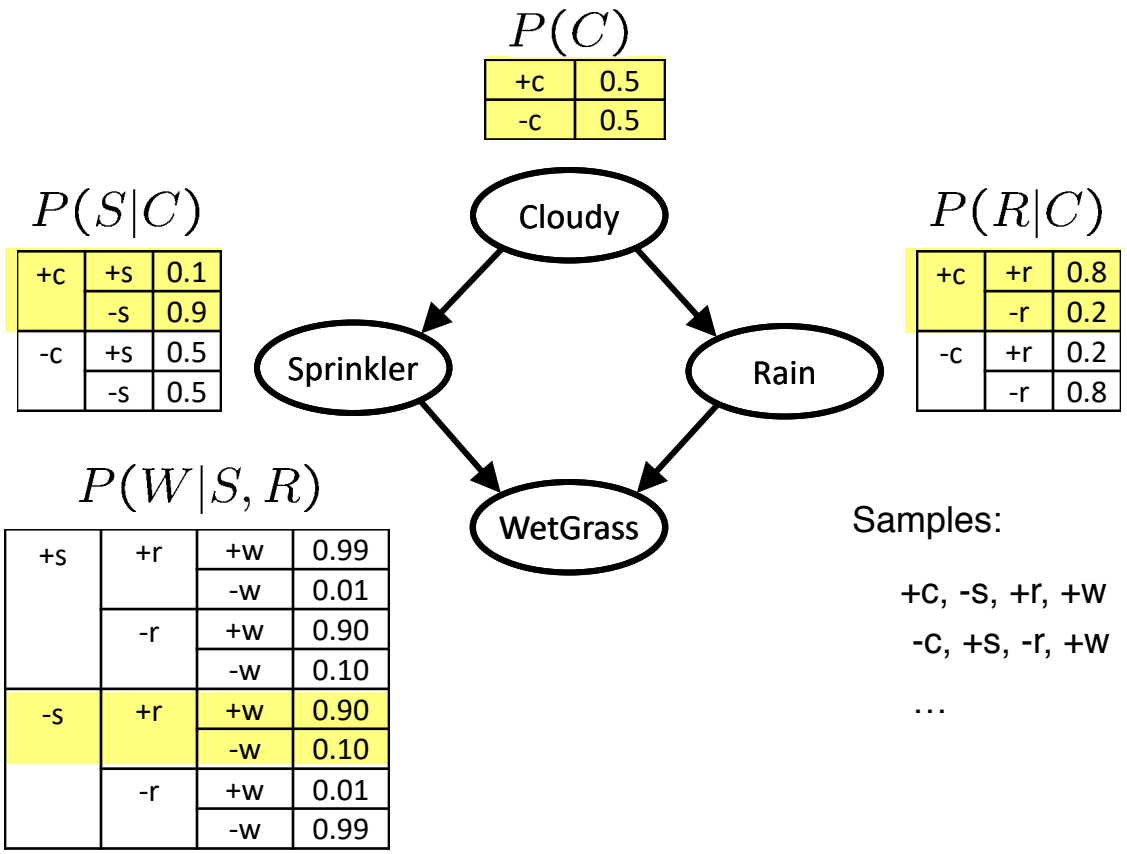
Prior (Direct) Sampling

- Ignore evidence. Sample from the joint probability.
- Do inference by counting the right samples.

- For $i = 1, 2, \ldots, n$
  - Sample x_i from $P(Xi \mid Parents(Xi))$
- Return $(x_1, x_2, \ldots, x_n)$



Prior Sampling



Example

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

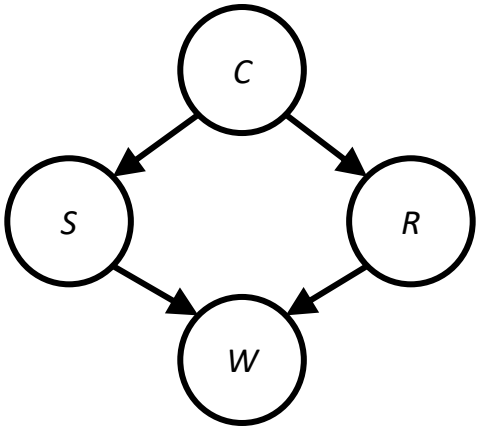
  +c, -s, +r, +w

  +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 $P(C|\ +w)$? $P(C|\ +r,+w)$? $P(C|\ -r,-w)$?

Prior Sampling Analysis

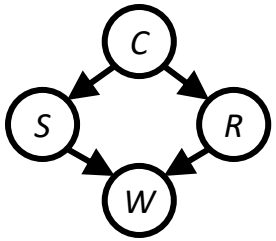
- This process generates samples with probability:

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

- …i.e. the BN's joint probability
- Let the number of samples of an event be $N_{PS}(x_1, \ldots, x_n)$
- Then $\lim_{N\to\infty} \hat{P}(x_1, \ldots, x_n) = \lim_{N\to\infty} N_{PS}(x_1, \ldots, x_n)/N$
$= S_{PS}(x_1, \ldots, x_n) = \mathrm{P}(x_1, \ldots, x_n)$
- I.e., the sampling procedure is consistent

Rejection Sampling

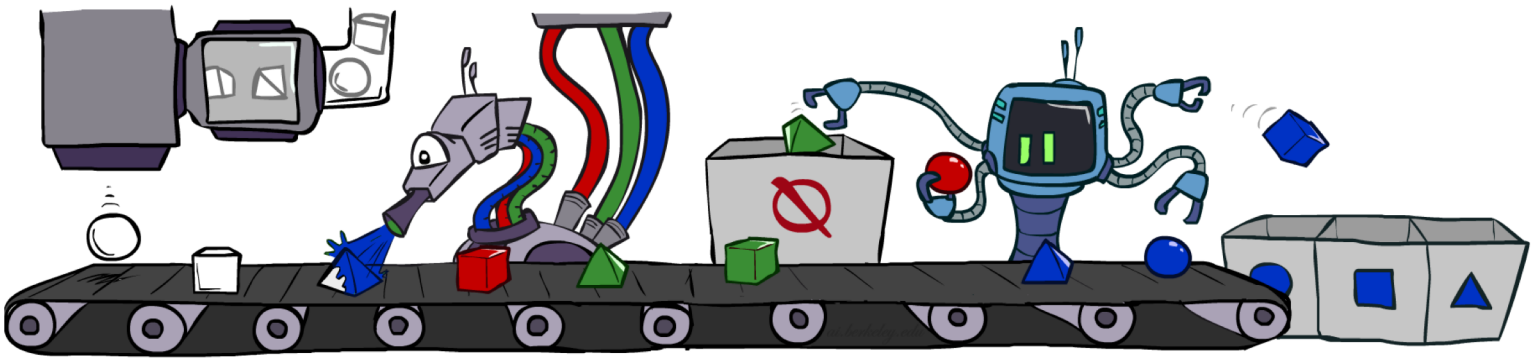
- Let's say we want $P(C| +s)$
  - Tally C outcomes, but ignore (reject) samples which don't have $S = +s$
  - This is called rejection sampling
  - It is also consistent for conditional probabilities (i.e., correct in the limit)



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

Rejection Sampling

- Evidence instantiation
- For $i = 1, 2, \ldots, n$
  - Sample x_i from $P(Xi \mid Parents(Xi))$
  - If x_i not consistent with evidence
  - Reject: Return, and no sample is generated in this cycle
- Return $(x_1, \ldots, x_n)$

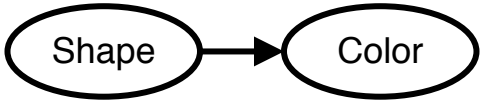


Rejection Sampling Analysis

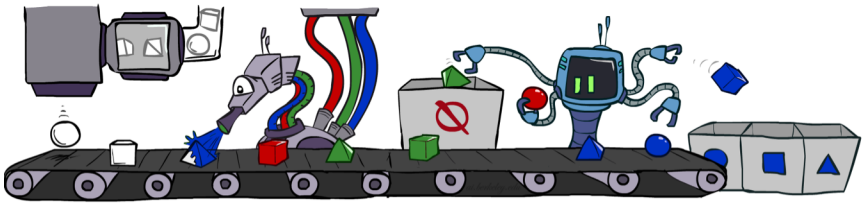
- Let $\hat{P}(X|e)$ be the estimated distribution that the algorithm returns.
- By definition: $\hat{P}(X|e) = \alpha N_{PS}(X, e) = \frac{N_{PS}(X,e)}{N_{PS}(e)}$
- We already know that
  - $N_{PS}(x_1, \ldots, x_n)/N \approx \mathrm{P}(x_1, \ldots, x_n)$ and
  - $N_{PS}(x_1, \ldots, x_m)/N \approx \mathrm{P}(x_1, \ldots, x_m)$ for $m < n$
- Thus $\hat{P}(X|e) \approx P(X|e)/P(e) = P(X|e)$
- Sampling procedure is consistent

Rejection Sampling

- Drawback?
  - If evidence is unlikely, rejects lots of samples
  - Evidence not exploited as you sample
  - Consider $P(Shape|blue)$

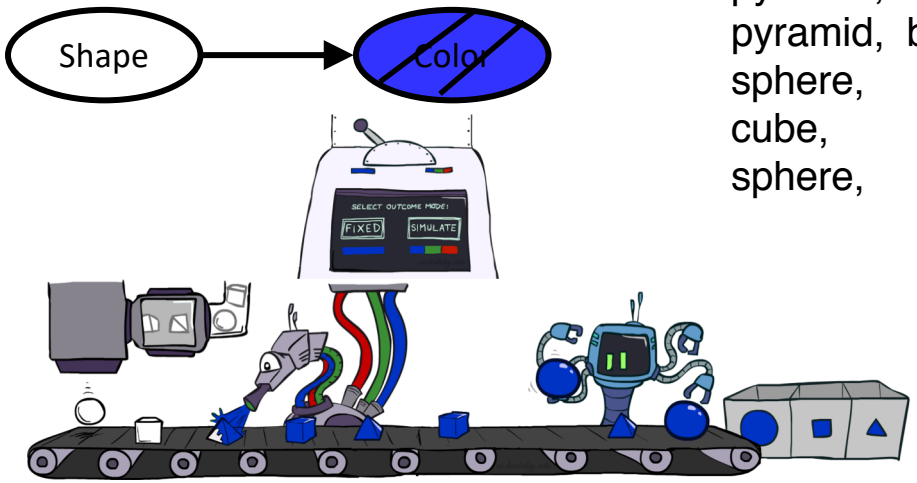


~~pyramid, green~~
~~pyramid, red~~
sphere, blue
~~cube, red~~
~~sphere, green~~



Likelihood Weighting or Importance Sampling

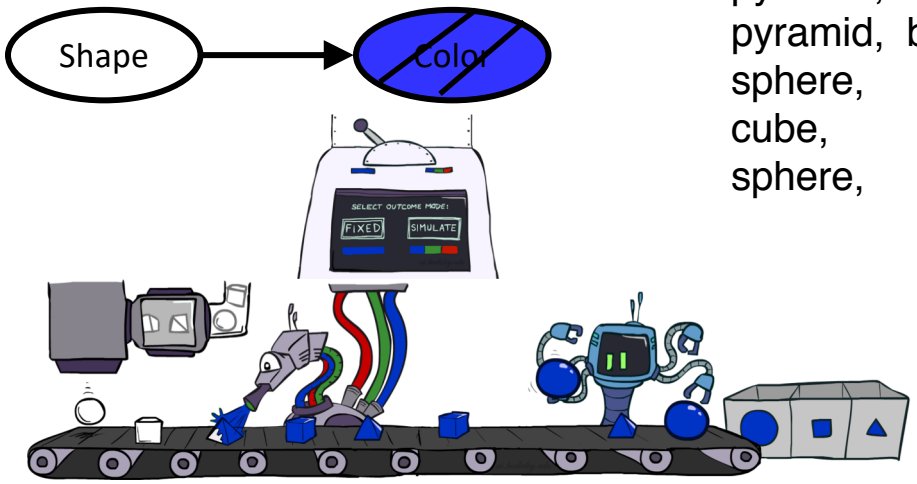
- Idea: fix evidence variables and sample the rest



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

Likelihood Weighting or Importance Sampling

- Idea: fix evidence variables and sample the rest

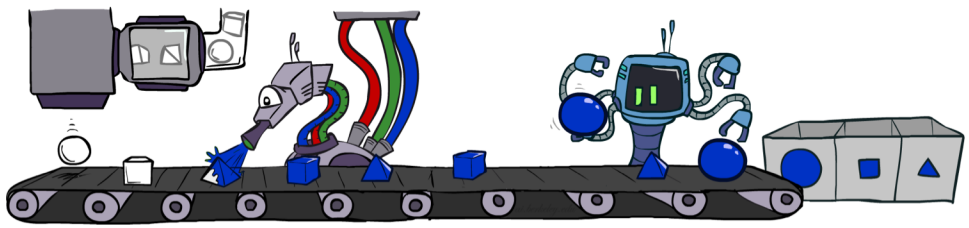


pyramid, blue
pyramid, blue
sphere, blue
cube, blue
sphere, blue

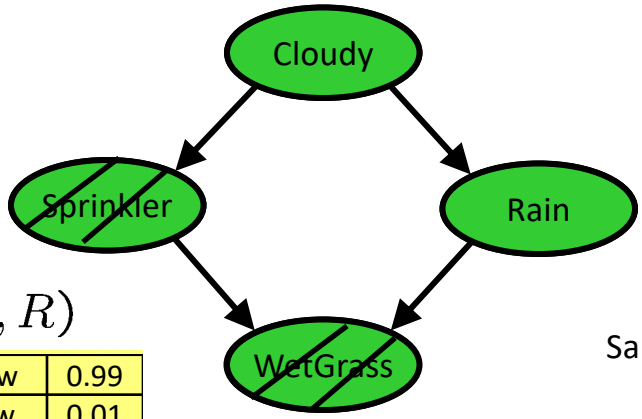
- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents

Likelihood Weighting

- Evidence instantiation
- $w = 1.0$
- for $i = 1, 2, \ldots, n$
  - if X_i is an evidence variable
    - X_i = observation x_i for X_i
    - Set $w = w \times P(x_i \mid Parents(X_i))$
  - else
    - Sample x_i from $P(X_i \mid Parents(X_i))$
- return $(x_1, \ldots, x_n)$, w



Likelihood Weighting



$P(C)$

+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

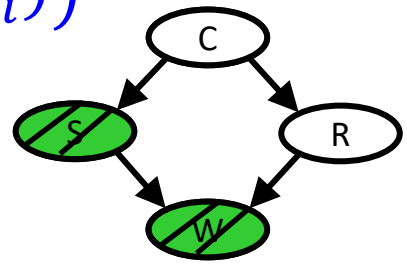
…

$$w = 1.0 \times 0.1 \times 0.99$$

Likelihood Weighting

- Sampling distribution if z sampled and e fixed evidence

$$S_{WS}(z,e) = \prod_{i=1}^{l} P\left(z_i | Parents(Z_i)\right)$$



- Now, samples have weights

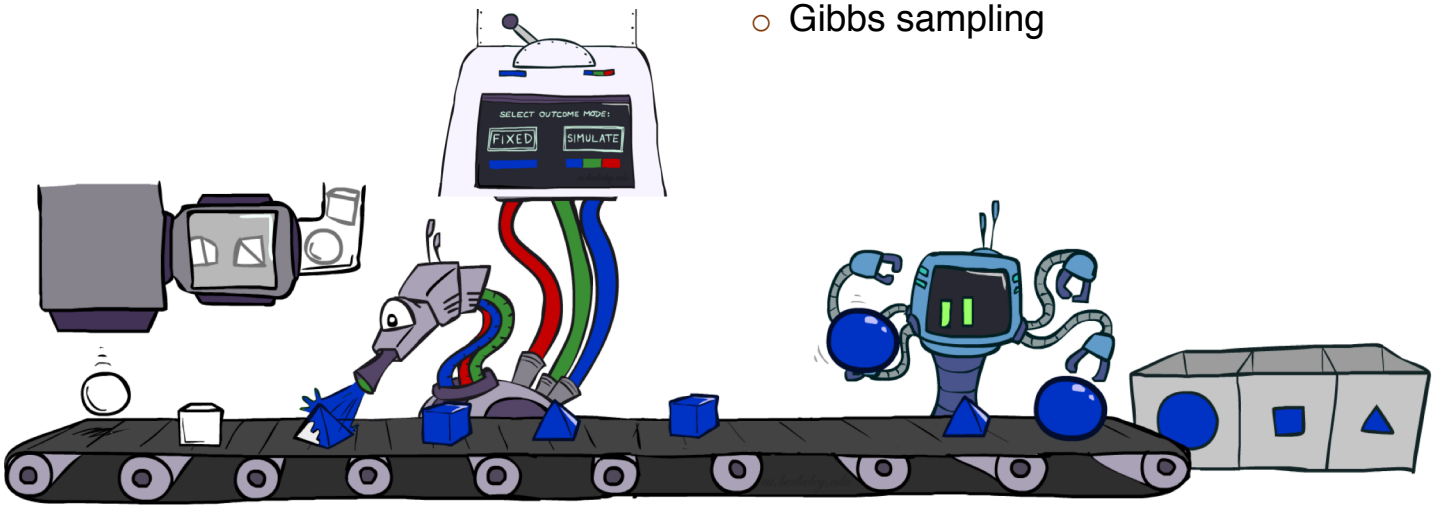
$$w(z,e) = \prod_{i=1}^{m} P\left(e_i | Parents(E_i)\right)$$

- Together, weighted sampling distribution is consistent

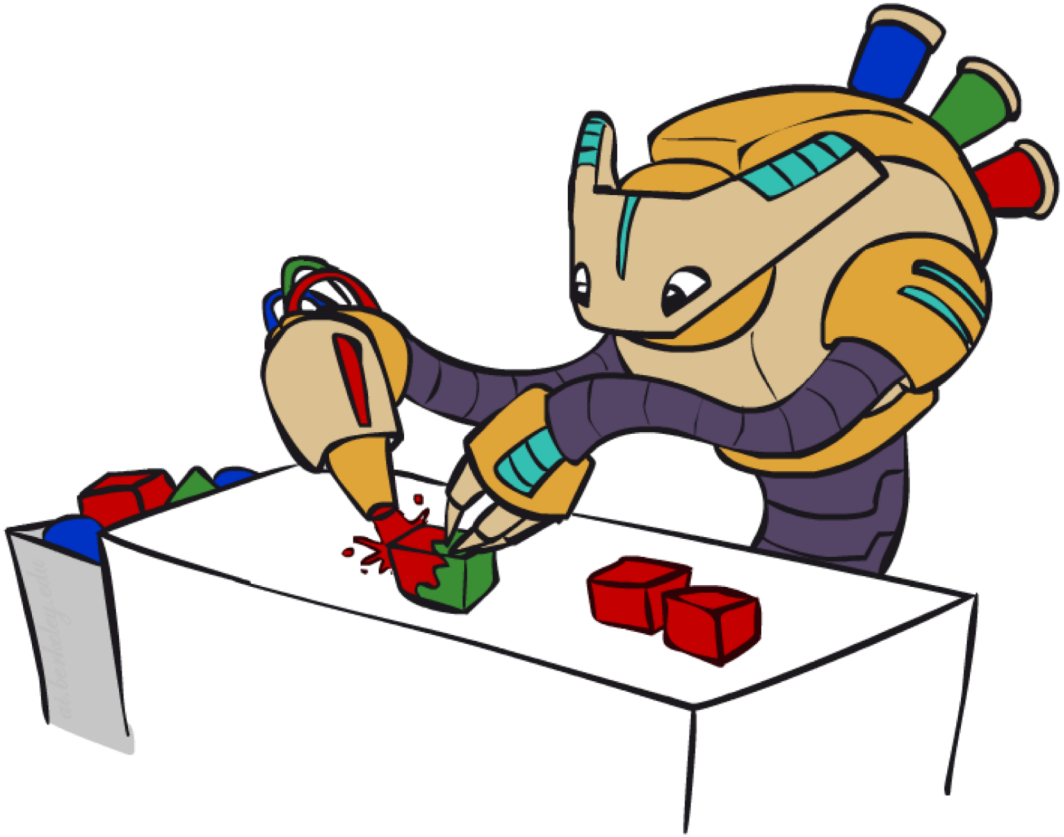
$$S_{WS}(z,e)\ w(z,e) = \prod_{i=1}^{l} P\left(z_i | Parents(Z_i)\right) \prod_{i=1}^{m} P\left(e_i | Parents(E_i)\right)$$
$$= P(z,e)$$

Likelihood Weighting

- Likelihood weighting is good
  - We have taken evidence into account as we generate the sample
  - E.g. here, W's value will get picked based on the evidence values of S, R
  - More of our samples will reflect the state of the world suggested by the evidence
- Likelihood weighting doesn't solve all our problems
  - Evidence influences the choice of downstream variables, but not upstream ones (C isn't more likely to get a value matching the evidence)
- We would like to consider evidence when we sample every variable
  - Gibbs sampling



Gibbs Sampling



Gibbs Sampling

- *Procedure:* keep track of a full instantiation $x_1, x_2, \ldots, x_n$. Start with an arbitrary instantiation consistent with the evidence. Sample one variable at a time, conditioned on all the rest, but keep evidence fixed. Keep repeating this for a long time.

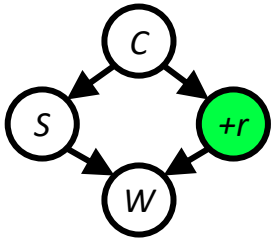
- *Property:* in the limit of repeating this infinitely many times the resulting sample is coming from the correct distribution

- *Rationale*: both upstream and downstream variables condition on evidence.

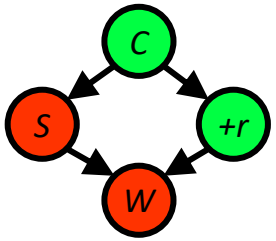
- In contrast: likelihood weighting only conditions on upstream evidence, and hence weights obtained in likelihood weighting can sometimes be very small. Sum of weights over all samples is indicative of how many "effective" samples were obtained, so want high weight.

Gibbs Sampling Example: $P(S \mid +r)$

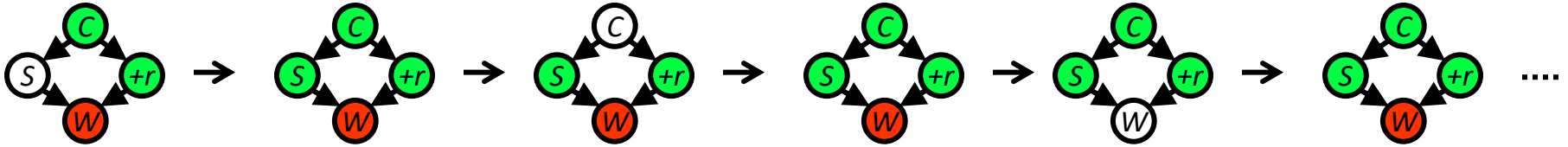
- Step 1: Fix evidence
  - R = +r



- Step 2: Initialize the other variables
  - Randomly



- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from $P(X \mid all\ other\ variables)$

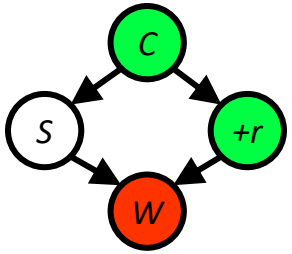


Sample from $P(S| +c, -w, +r)$ Sample from $P(C| +s, -w, +r)$ Sample from $P(W| +s, +c, +r)$

Efficient Resampling of One Variable

- Sample from $P(S \mid +c, +r, -w)$

$$
\begin{aligned}
P(S|+c,+r,-w) &= \frac{P(S,+c,+r,-w)}{P(+c,+r,-w)} \\
&= \frac{P(S,+c,+r,-w)}{\sum_s P(s,+c,+r,-w)} \\
&= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{\sum_s P(+c)P(s|+c)P(+r|+c)P(-w|s,+r)} \\
&= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{P(+c)P(+r|+c)\sum_s P(s|+c)P(-w|s,+r)} \\
&= \frac{P(S|+c)P(-w|S,+r)}{\sum_s P(s|+c)P(-w|s,+r)}
\end{aligned}
$$



- Many things cancel out – only CPTs with S remain!
- More generally: only CPTs that have the resampled variable need to be considered, and joined together

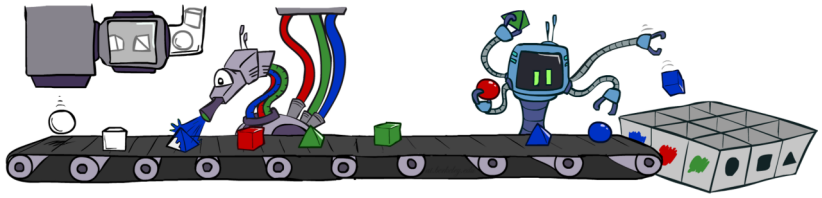
Further Reading on Gibbs Sampling

- Gibbs sampling produces sample from the query distribution $P(Q \mid e)$ in limit of re-sampling infinitely often

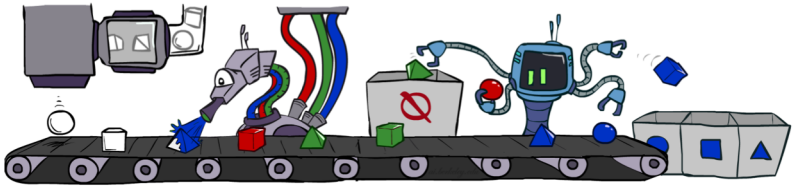
- Gibbs sampling is a special case of more general methods called Markov chain Monte Carlo (MCMC) methods
  - Metropolis-Hastings is one of the more famous MCMC methods (in fact, Gibbs sampling is a special case of Metropolis-Hastings)

Bayes' Net Sampling Summary

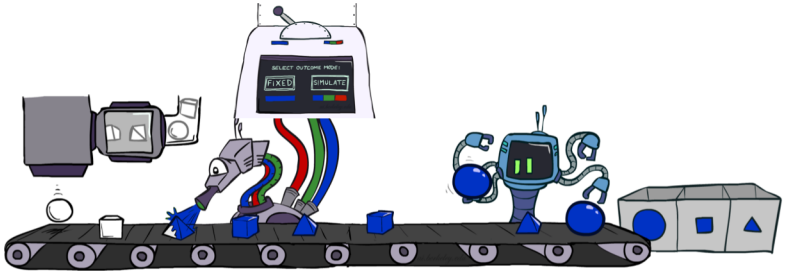
- Prior Sampling P



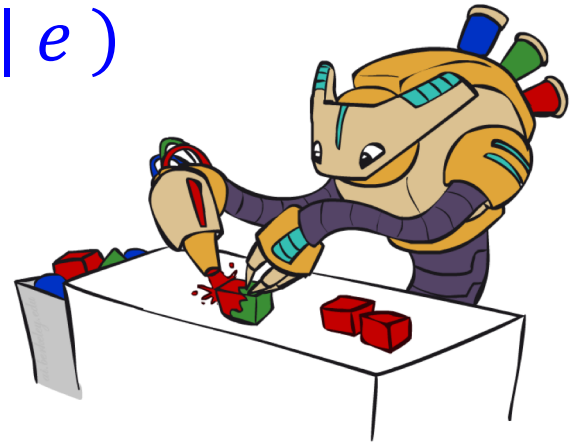
- Rejection Sampling $P(Q \mid e)$



- Likelihood Weighting $P(Q \mid e)$



- Gibbs Sampling $P(Q \mid e)$



Outline

- Probability theory - basics
- Probabilistic Inference
  - Inference by enumeration
- Conditional Independence
- Bayesian Networks
- Inference in Bayesian Networks
  - Exact Inference
  - Approximate Inference