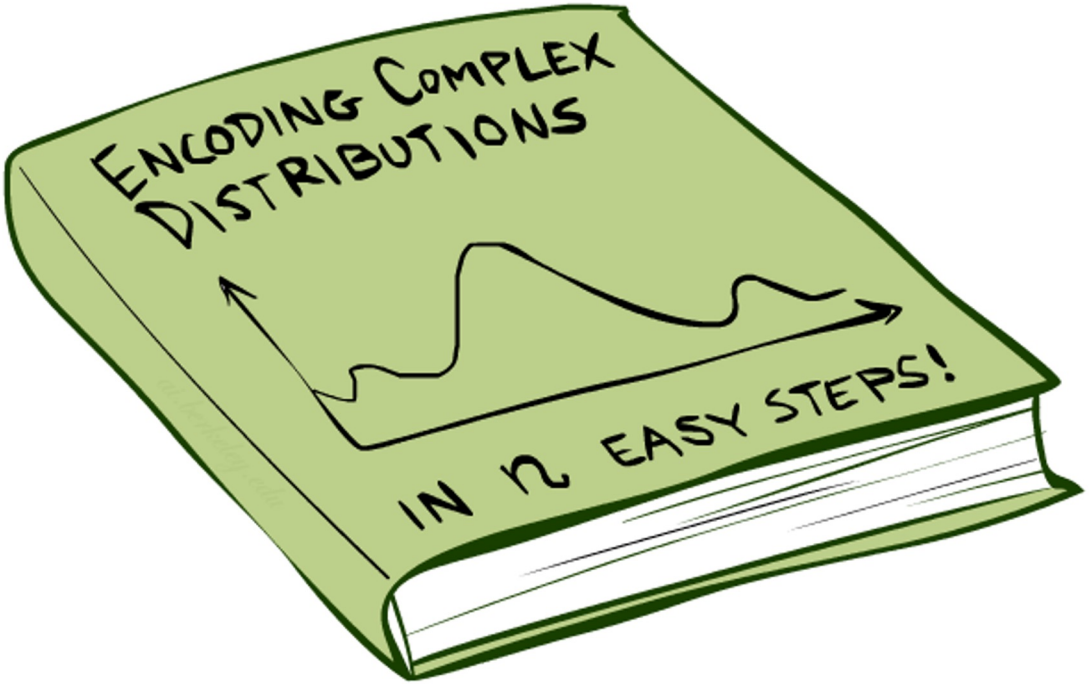
Bayesian Networks



Recap: Inference by Enumeration

General case:

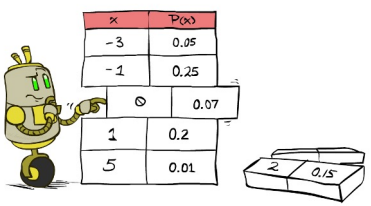
- Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
- Query variable: Q
- Hidden variables: $H_1 \dots H_r$

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

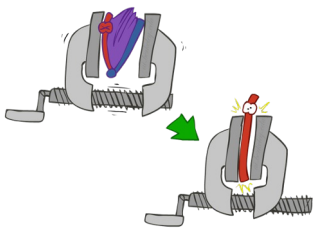
We want:

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

Step 1: Select the entries consistent with the evidence



Step 2: Sum out H to get joint of query and evidence



Step 3: Normalize

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

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

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

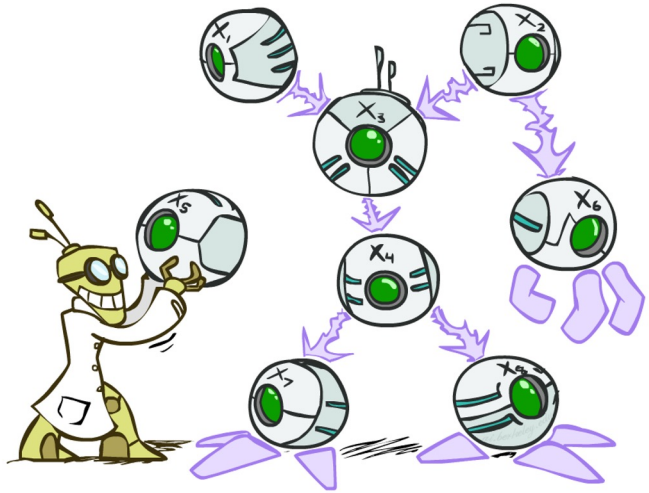
$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

Bayesian Networks: Motivation

- Two problems with using full joint distribution tables as our probabilistic models for inference:
  - Unless there are only a few variables, the joint distribution is WAY too big to represent explicitly;
  - Hard to learn (estimate) anything empirically about more than a few variables at a time.

Bayesian Networks: The Big Picture

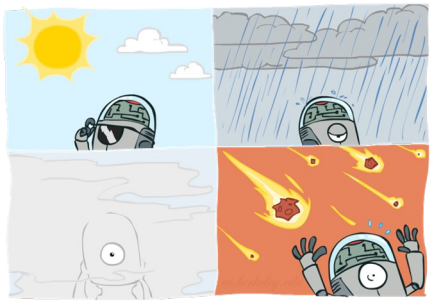
- Bayesian Networks: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probability tables, or CPTs).
  - Using a directed graph to represent full joint probability distribution in a compact way.
  - More properly called graphical models.

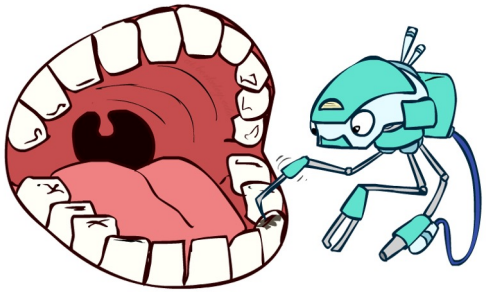


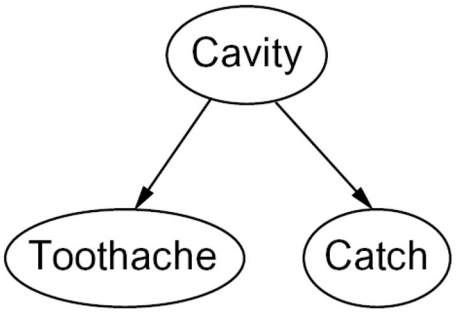
Graphical Model Notation

- Nodes: corresponding to random variables (with domains).
  - Can be discrete or continuous variables.
  - Can be observed or unobserved (e.g., “hidden” variable).
- Directed links or arrows: connecting pairs of nodes.
  - Can describe how variables locally interact, or formally conditional independence relationship.
  - If there is an arrow from node X to node Y, X is said to be a parent of Y.
  - The graph has no directed cycles and hence is a directed acyclic graph (DAG).
  - Arrows mean that there may (but not necessarily) be a causal relationship between the two variables.

Bayes Net Examples



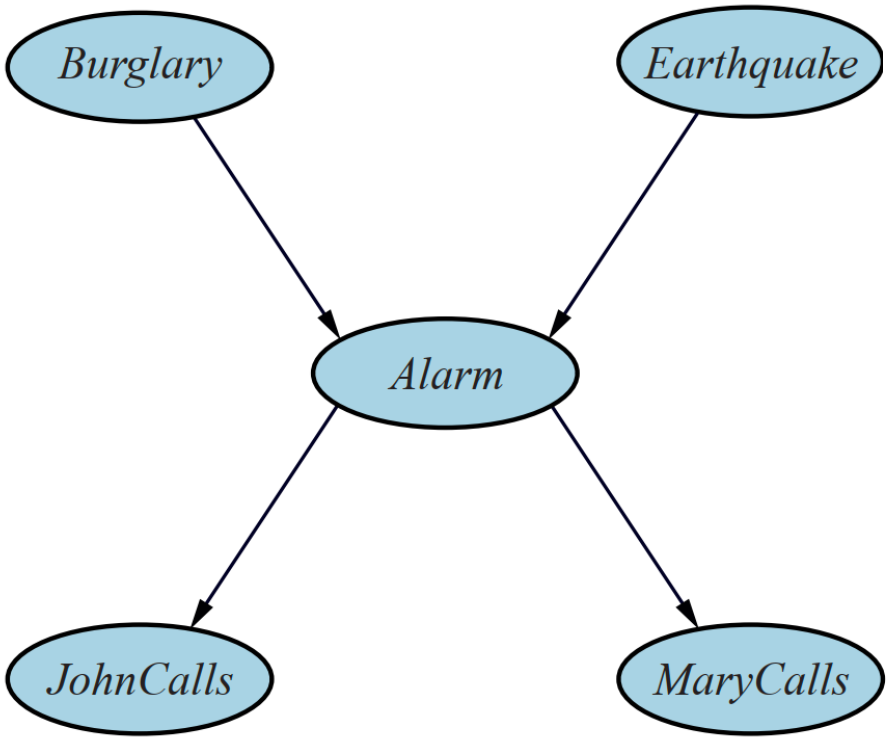




Naïve Bayes model can be represented by a simple Bayesian network.

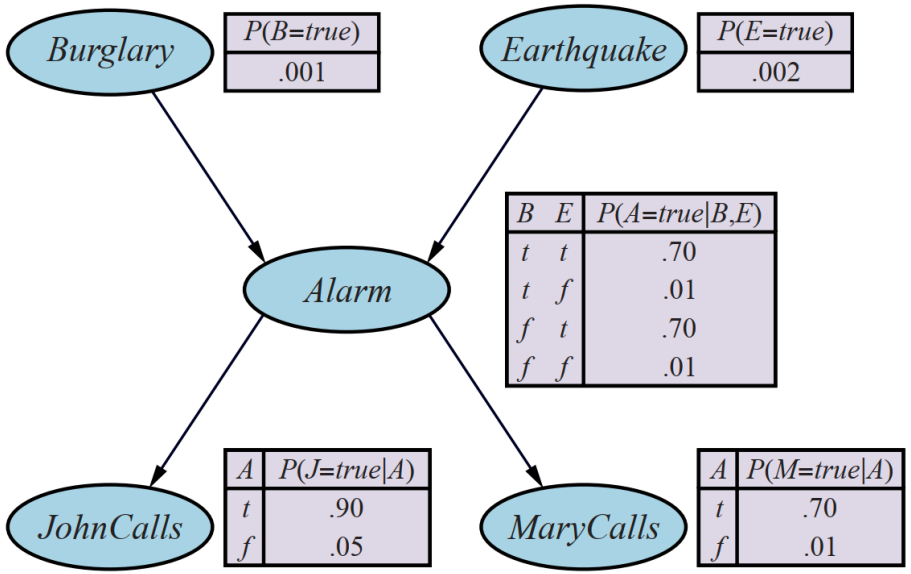
Example: Alarm Network

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!



Conditional Probability Table (CPT)

- The local probability information attached to each node can take the form of conditional probability table (CPT).
  - Each row in a CPT contains the conditional probability of each node value for a conditioning case.
  - CPT can only be used for discrete random variables as in this example.



Are there hidden variables in this Bayes net?

Bayes Net Semantics

- The syntax of a Bayes net consists of a directed acyclic graph with some local probability information attached to each node.
- The semantics defines how the syntax corresponds to a joint distribution over the variables of the network.

$$P(x_i|x_1 \cdots x_{i-1}) = P(x_i|parents(X_i))$$

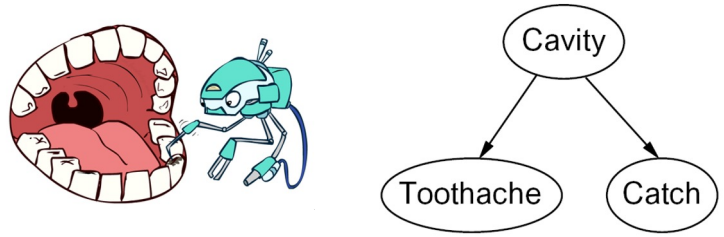
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \,|\, parents(X_i))$$

Bayes Net Semantics

- The syntax of a Bayes net consists of a directed acyclic graph with some local probability information attached to each node.
- The semantics defines how the syntax corresponds to a joint distribution over the variables of the network.

$$P(x_i|x_1 \cdots x_{i-1}) = P(x_i|parents(X_i))$$

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



$P(+cavity, +catch, -toothache)$

- Recall that we can infer by using full joint distributions of all random variables. If a Bayes net is a representation of the joint distribution, then it can also be used to answer any query – which is a great news!

Constructing Bayes Net

- Assume there is an ordering of all variables, the parents of node X_i should contain all those nodes in $\{X_1, \cdots, X_{i-1}\}$ that directly influence X_i.
- A Bayesian network accurately represents a domain only if each node is conditionally independent of its other preceding variables in the chosen ordering, given its parents.

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

vs.

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_n \mid x_{n-1}, \ldots, x_1) P(x_{n-1} \mid x_{n-2}, \ldots, x_1) \cdots P(x_2 \mid x_1) P(x_1) \\ &= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1). \end{aligned}$$

Constructing Bayes Net

1. Determine the set of variables that are required to model the domain.
2. Order all variables as $\{X_1, \cdots, X_n\}$
   - Any order of the variables will work, but the resulting network will be more compact if the variables are ordered such that causes precede effects.
3. For i = 1 to n do:
   - Choose a minimal set of parents for X_i from $\{X_1, \cdots, X_{i-1}\}$ to satisfy

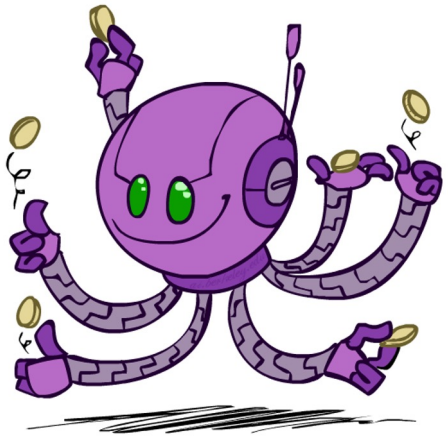
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))$$

   - For each parent insert a link from the parent to X_i
   - Write down the conditional probability table (CPT) for $P(x_i \mid parents(X_i))$

Example: Coin Flips

- *N* independent coin flips

X_1 X_2 . . . X_n



Example: Coin Flips

X_1 X_2 . . . X_n

$P(X_1)$

h	0.5
t	0.5

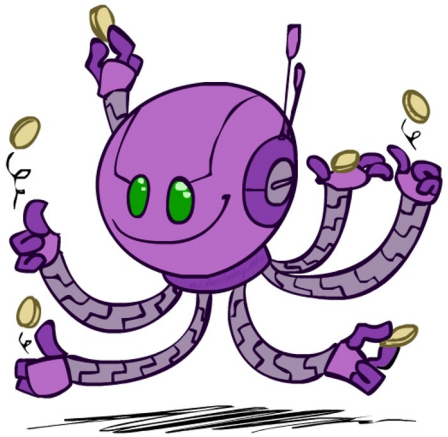
$P(X_2)$

h	0.5
t	0.5

. . .

$P(X_n)$

h	0.5
t	0.5

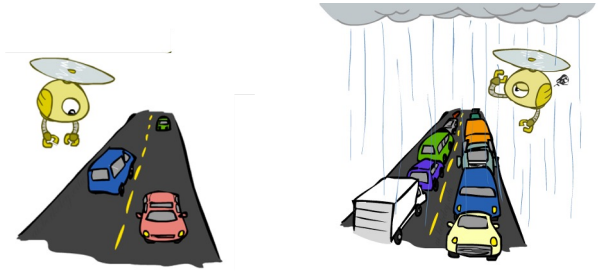


$$P(h, h, t, h) = P(h)P(h)P(t)P(h)$$

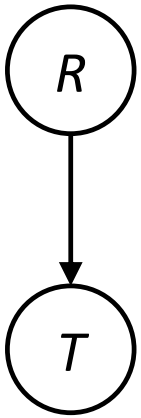
Only distributions whose variables are absolutely independent can be represented by a Bayes net with no edge links.

Example: Traffic

- Variables:
  - *R*: It rains
  - *T*: There is traffic

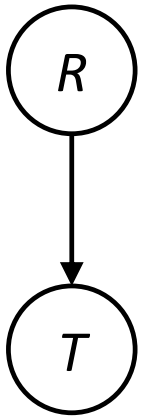


Model 1: independence



Model 2: rain may cause traffic

Example: Traffic



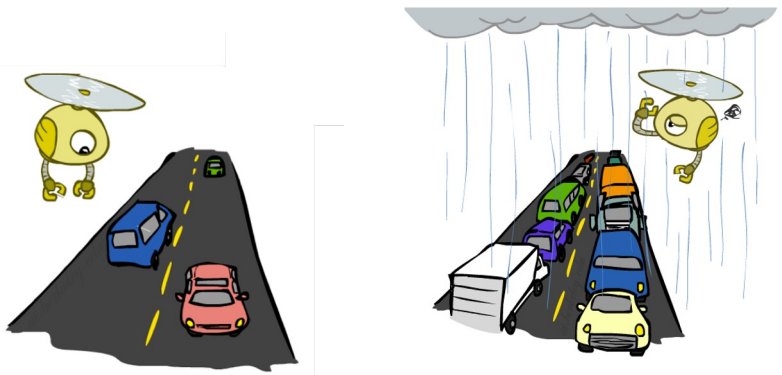
$P(R)$

$+r$	1/4
$-r$	3/4

$P(T|R)$

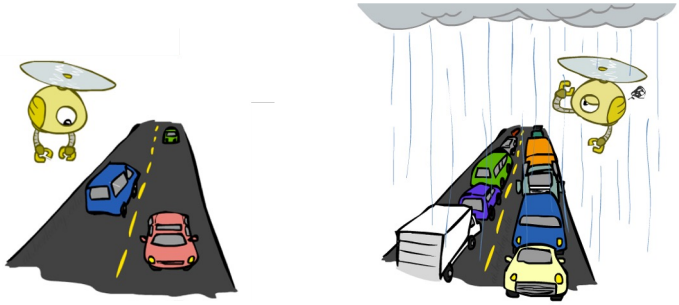
$+r$	$+t$	3/4
	$-t$	1/4
$-r$	$+t$	1/2
	$-t$	1/2

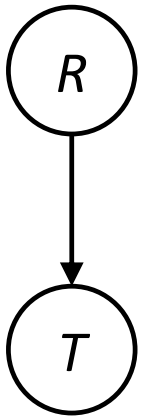
$$P(+r,-t) = P(+r)P(-t|+r) \;=\; ¼ * ¼$$



Example: Traffic

- Causal direction





$P(R)$

$+r$	1/4
$-r$	3/4

$P(T|R)$

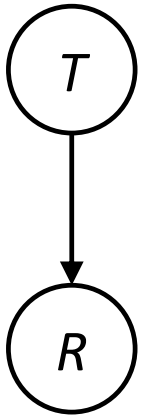
$+r$	$+t$	3/4
	$-t$	1/4
$-r$	$+t$	1/2
	$-t$	1/2

$P(T, R)$

$+r$	$+t$	3/16
$+r$	$-t$	1/16
$-r$	$+t$	6/16
$-r$	$-t$	6/16

Example: Reverse Traffic

- Reverse causality?



$P(T)$

$+t$	9/16
$-t$	7/16

$P(R|T)$

$+t$	$+r$	1/3
	$-r$	2/3
$-t$	$+r$	1/7
	$-r$	6/7

$P(T, R)$

$+r$	$+t$	3/16
$+r$	$-t$	1/16
$-r$	$+t$	6/16
$-r$	$-t$	6/16

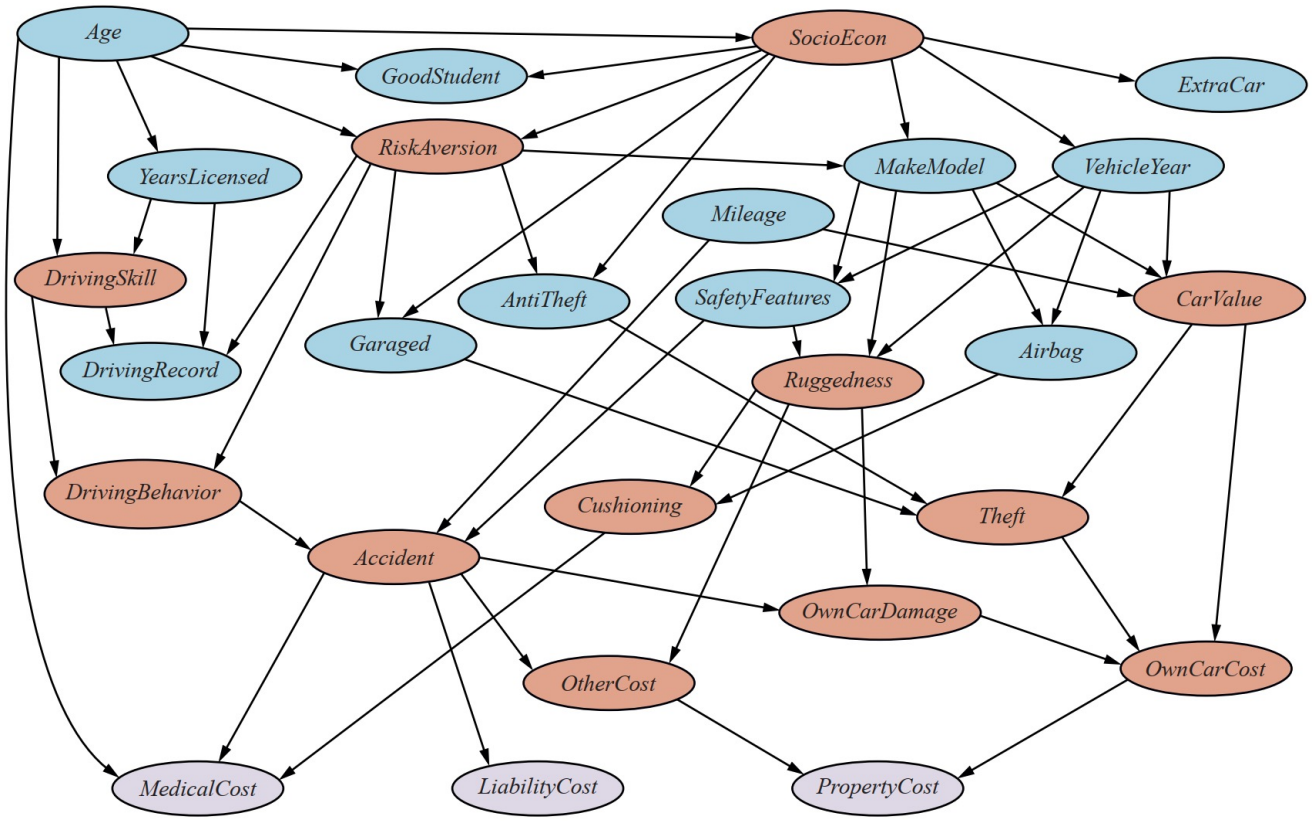
Causality in Bayes Net

- Bayesian networks do not necessarily encode causality.
  - The arrows may reflect statistical correlation, not causal influence.
  - In some domains, especially when variables are missing, a true causal network may not exist or be hard to define.
- What do the arrows really mean in Bayes nets?
  - The topology of a Bayes net primarily encodes conditional independence.
  - The structure may happen to align with causal relationships, but this is not guaranteed.

Causality in Bayes Net

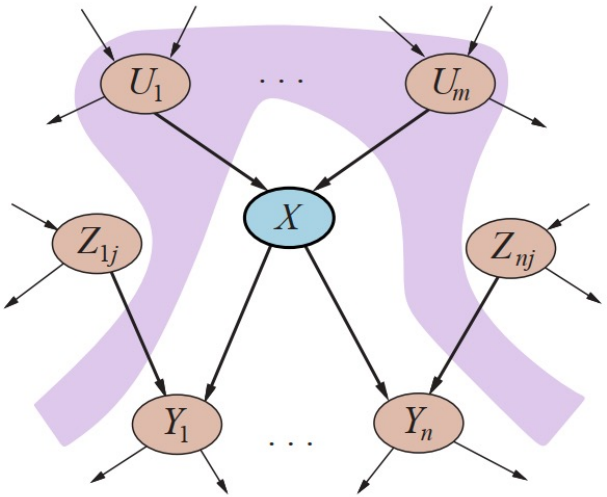
- When Bayes Nets Reflect True Causality,
  - The network is often simpler (fewer parents per node);
  - Easier to understand and interpret;
  - Easier to elicit from domain experts, especially in fields like medicine or engineering.

Example: Car Insurance



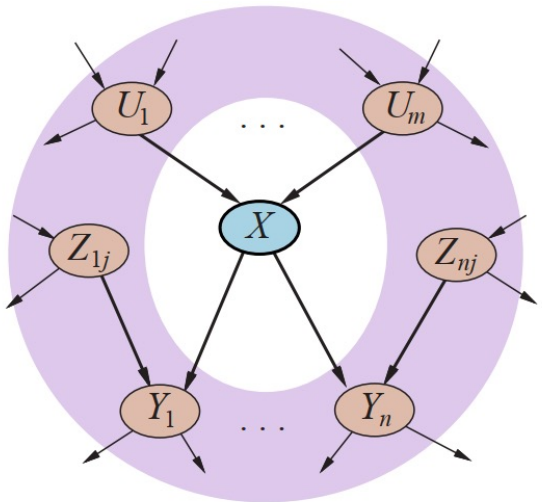
Conditional Inpendence in Bayes Net

- Each variable is conditionally independent of its other predecessors in the ordered variables, given its parents.
- Each variable is conditionally independent of its non-descendants, given its parents.



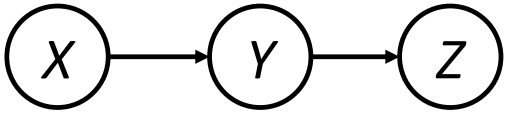
Conditional Inpendence in Bayes Net

- Each variable is conditionally independent of all other nodes in the network, given its Markov blanket.
  - Markov blanket refers to the parents, children, and children's parents.



Example: Causal Chain

- Question: Are *X* and *Z* guaranteed to be independent?
  - Answer: No (Example: low pressure causes rain, which further causes traffic)
  - *X* can influence *Z*, *Z* can influence *X* (via *Y*).



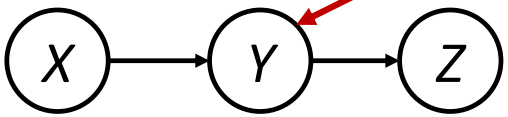
$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

X: Low pressure *Y*: Rain *Z*: Traffic

Example: Causal Chain

- Question: Is *X* guaranteed to be independent of Z, given Y?
  - Answer: Yes.

Evidence along the chain "blocks" the influence.



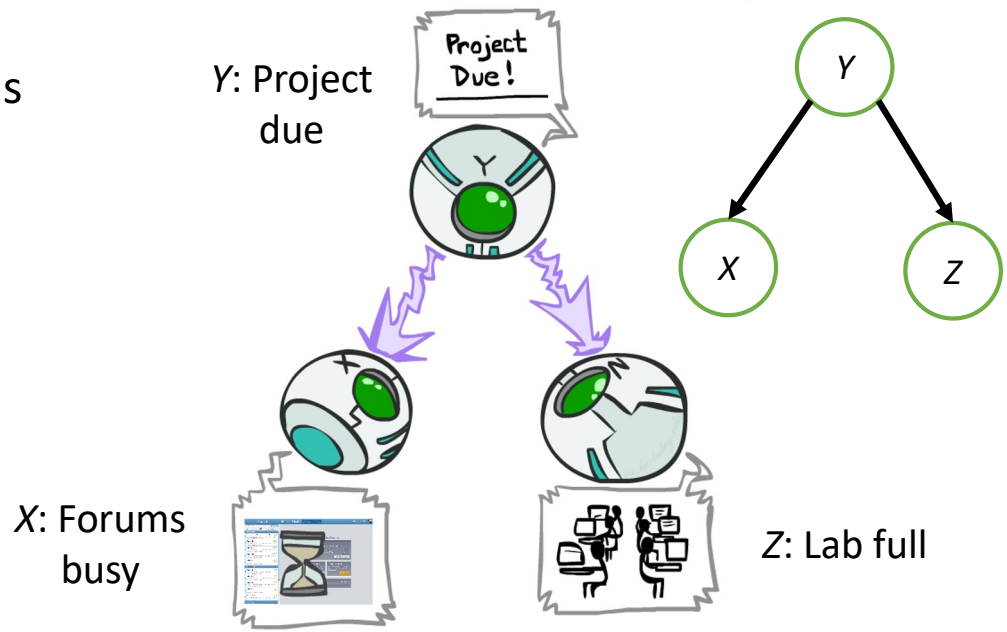
$$P(z|x,y) = \frac{P(x,y,z)}{P(x,y)}$$

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

X: Low pressure *Y*: Rain *Z*: Traffic

Example: Common Cause

- Question: Are X and Z guaranteed to be independent?
  - Answer: No.
  - E.g.: Project due causes both forums busy and lab full.



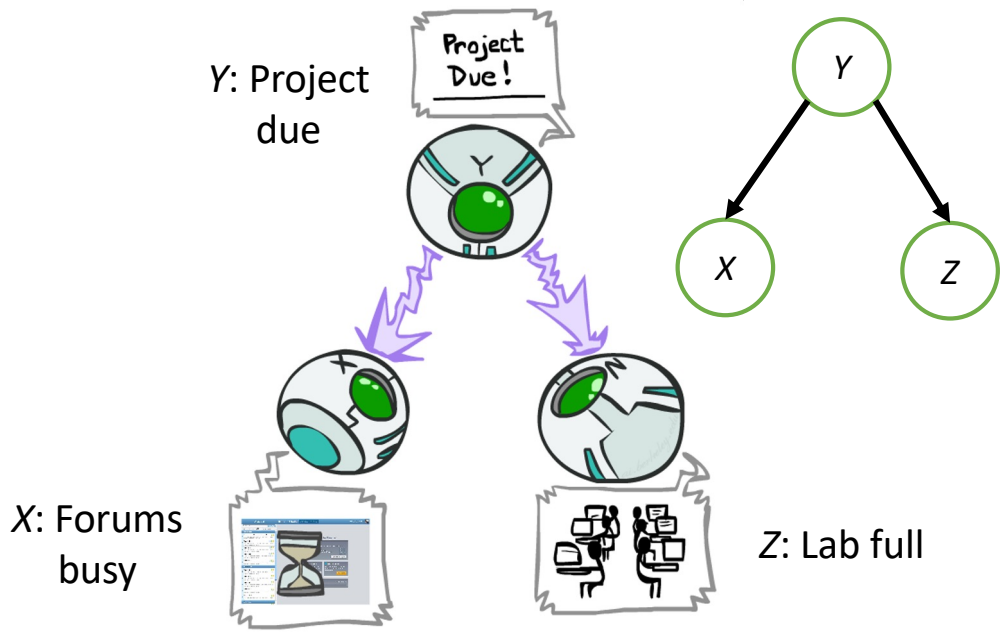
Example: Common Cause

- Question: Are *X* and *Z* guaranteed to be independent, given *Y*?
  - Answer: Yes.

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

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

$$= P(z|y)$$



Observing the cause blocks influence between effects.

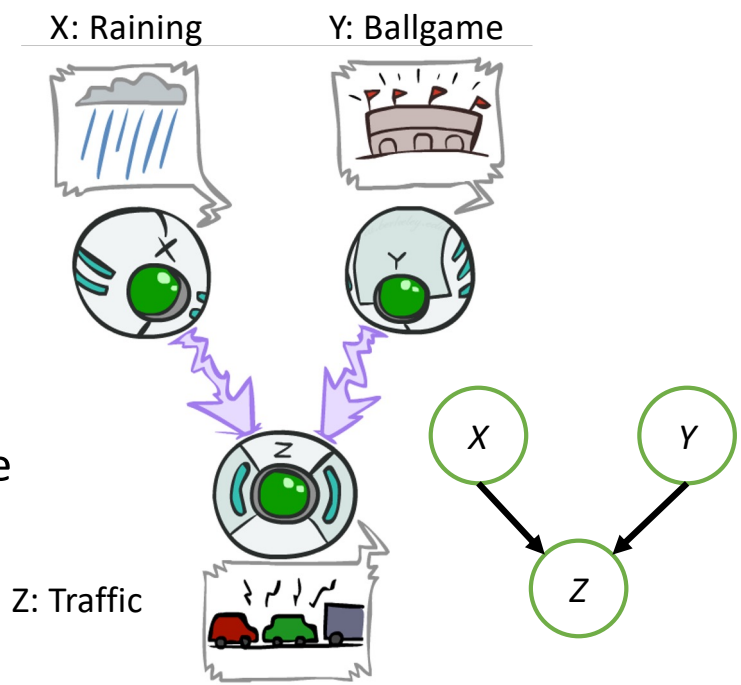
Example: Common Effect

- Question: Are *X* and *Y* guaranteed to be independent?
  - Answer: Yes.
  - E.g.: the ballgame and the rain cause traffic, but they are not correlated.

$$P(x, y) = \sum_{z} P(x, y, z)$$

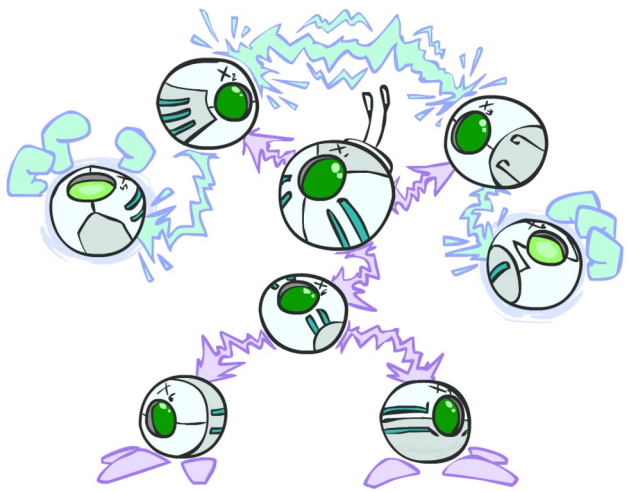
- Question: Are *X* and *Y* independent given *Z*?
  - Answer: No.
  - E.g.: seeing traffic puts the rain and the ballgame in competition as explanation.

Observing an effect activates influence between possible causes.



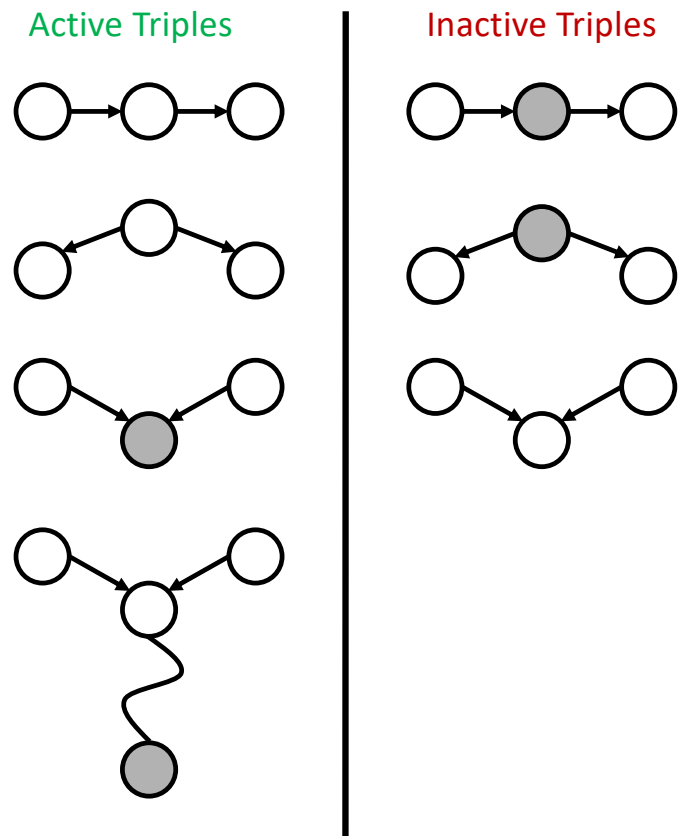
The General Case

- General question:
  In a given Bayes net, are two variables independent (given evidence)?
- Solution: analyze the graph.
- Any complex example can be broken into repetitions of the three canonical cases.
  - Causal chain;
  - Common cause;
  - Common effect.

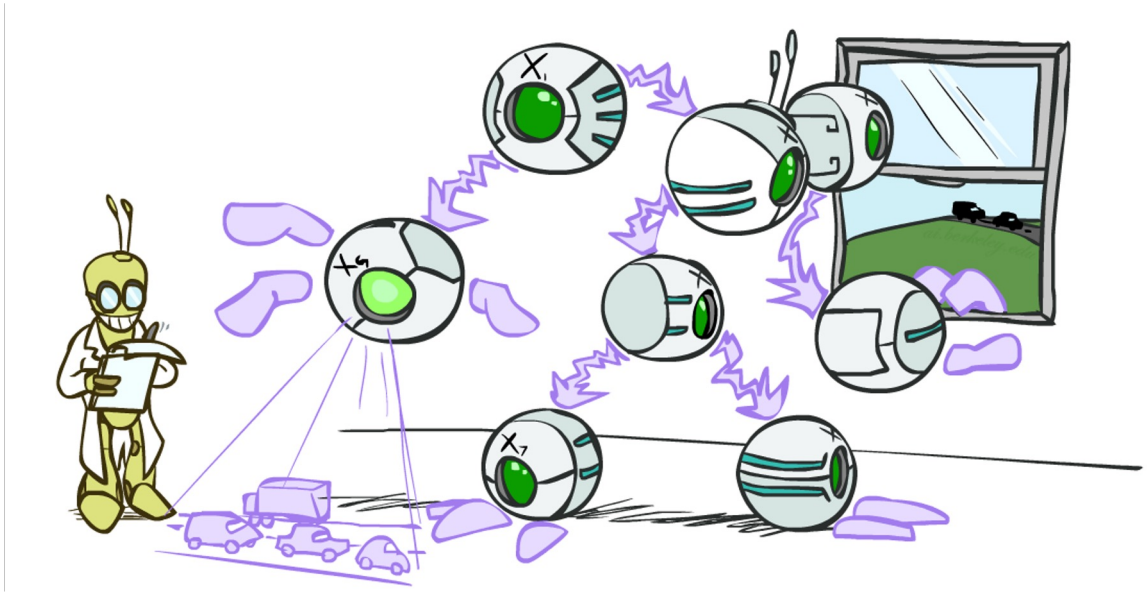


Active / Inactive Paths

- Question: Are *X* and *Y* conditionally independent given evidence {*Z*}?
  - Yes, if *X* and *Y* "d-separated" by *Z.*
    - Consider all (undirected) paths from *X* to *Y.*
    - No active paths = independence!
  - A path is active if each triple is active:
    - Causal chain *X* -> *Y* -> *Z* where *Y* is unobserved.
    - Common cause *X* <- *Y* -> *Z* where *Y* is unobserved.
    - Common effect *X* -> *Y* <- *Z* where *Y* (*or one of its descendants*) is observed.



Inference with Bayes Net

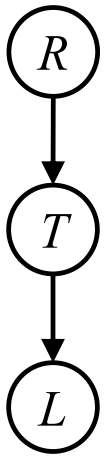


Inference

- Inference is the process of computing the probability distribution of one or more variables, given observed evidence about other variables in the network.
  - It answers queries like "What is the probability of *X* given *Y*, or $P(X|Y)$?"
- Why inference is needed?
  - Reasoning under uncertainty: Diagnose diseases from symptoms; Detect faults in engineering systems; Assess risks in financial decision-making.
  - Predicting future events: Forecast traffic conditions based on weather; Estimate recovery outcomes based on treatments.
  - Identifying probable causes and initials: Trace the root cause of machine failure using sensor data; Understand underlying factors behind observed effects.

Example: Traffic Domain

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



$$P(L) = ?$$

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

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

$P(R)$

$+r$	0.1
$-r$	0.9

$P(T|R)$

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

$P(L|T)$

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

Inference by Enumeration: Procedural Outline

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

$P(R)$

$+r$	0.1
$-r$	0.9

$P(T|R)$

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

$P(L|T)$

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

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

$P(R)$

$+r$	0.1
$-r$	0.9

$P(T|R)$

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

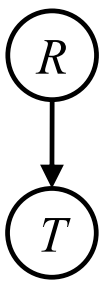
$P(+\ell|T)$

$+t$	$+l$	0.3
$-t$	$+l$	0.1

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

Operation 1: Join Factors

- First basic operation: joining factors
- Combining factors:
  - 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

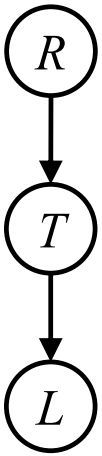
$+r$	$+t$	0.8
$+r$	$-t$	0.2
$-r$	$+t$	0.1
$-r$	$-t$	0.9

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

$$\forall r, t: \quad P(r,t) = P(r) \cdot P(t|r)$$

- Computation for each entry: pointwise products

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

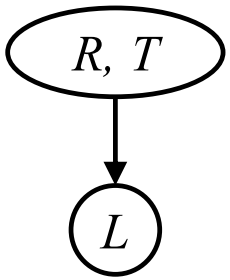
Join R

$P(R,T)$

$+r$	$+t$	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:
  - Shrink 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 out 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

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

$P(T,L)$

Sum out T

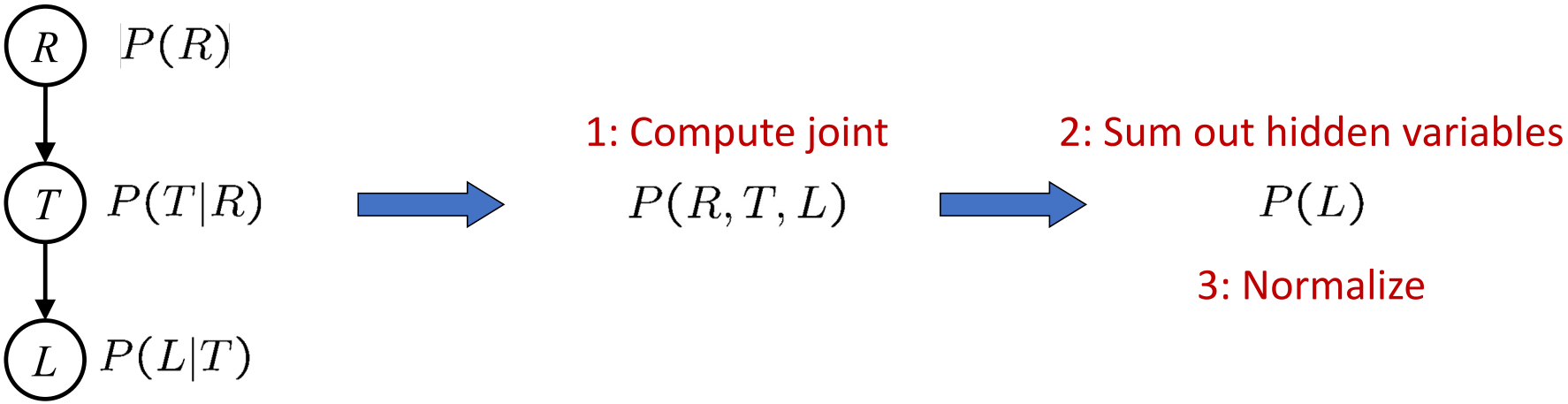
L

$+l$	0.134
$-l$	0.866

$P(L)$

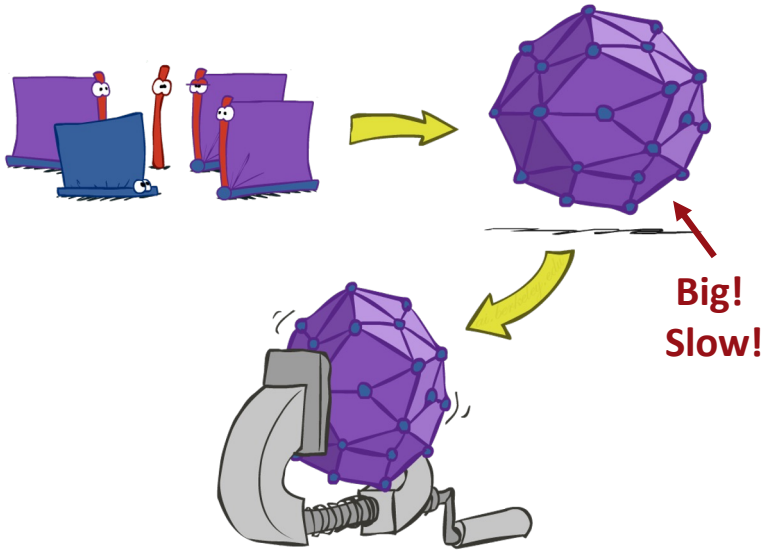
Multiple Join, Multiple Eliminate

- The procedure of multiple-join-multiple-eliminate is equivalent to inference by enumeration.

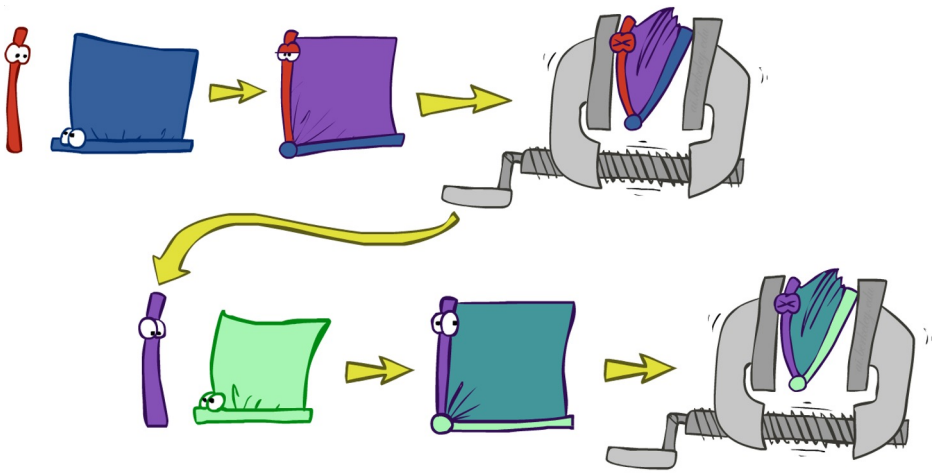


Variable Elimination

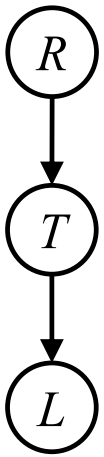
- Why is inference by enumeration slow?
  - Join up the whole joint distribution before you summing out the hidden variables.



- Idea: interleave joining and marginalizing!
  - Called "Variable Elimination".



Example: Traffic Domain



$$P(L) = ?$$

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

Join on r

Join on t

Eliminate r

Eliminate t

Inference by Enumeration

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

Join on r

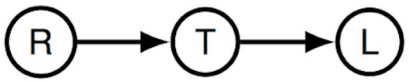
Eliminate r

Join on t

Eliminate t

Variable Elimination

The computational and space complexity of variable elimination is determined by the largest factor (e.g., check *P(R,T,L)*).



$P(R)$

T	L
+r	0.1
-r	0.9

$P(T|R)$

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

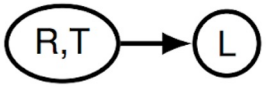
$P(L|T)$

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

Join R

$P(R,T)$

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

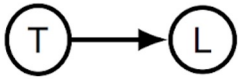


$P(L|T)$

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

Sum out R

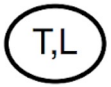
+t	0.17
-t	0.83



$P(L|T)$

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

Join T



$P(T,L)$

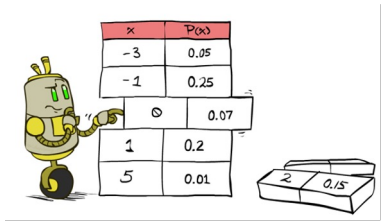
+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

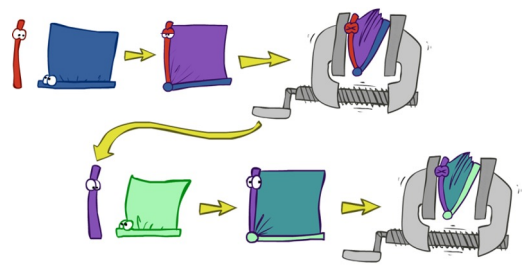
Sum out T

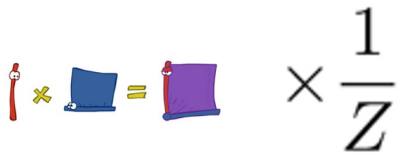
+l	0.134
-l	0.866

General Variable Elimination

- Query: $P(Q|E_1 = e_1, \ldots E_k = e_k)$
- Start with initial factors:
  - Local CPTs (instantiated by *Evidence* if needed)
- While there are still hidden variables (not *Q* or *Evidence*):
  - Pick a hidden variable *H*;
  - Join all factors mentioning *H*;
  - Eliminate (i.e., sum out) *H*.
- Join all remaining factors and normalize.

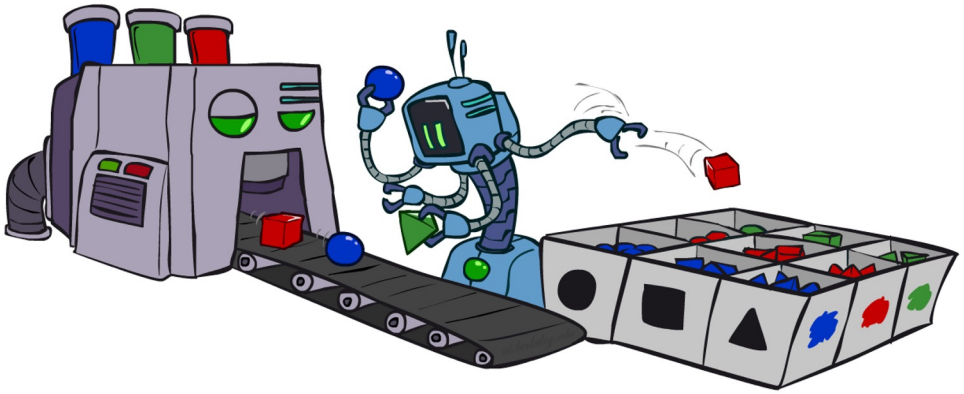






Approximate Inference: Sampling

- Variable elimination is for exact inference, which is often impossible.
  - So we need approximate inference, i.e., by sampling.
- Basic idea of sampling for approximate inference:
  - Draw *N* samples from a sampling distribution *S*;
  - Compute an approximate posterior probability;
  - Show this converges to the true probability *P*.

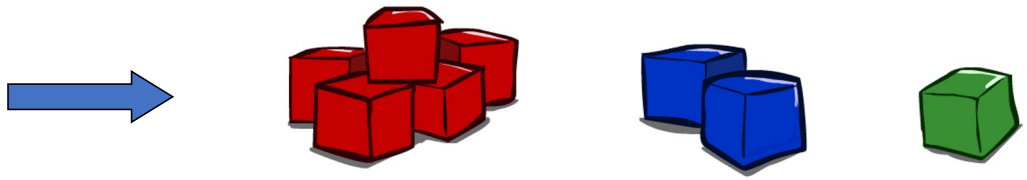


Sampling from Given Distribution

- Step 1: Get sample u from uniform distribution over [0, 1).
- Step 2: Convert this sample u into an outcome for the given distribution.
  - Associate each target outcome with a sub-interval of [0,1).
  - The sub-interval is size equal to probability of the outcome.

C	$P(C)$
red	0.6
green	0.1
blue	0.3

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



Example: after sampling 8 times

Sampling in Bayesian Networks

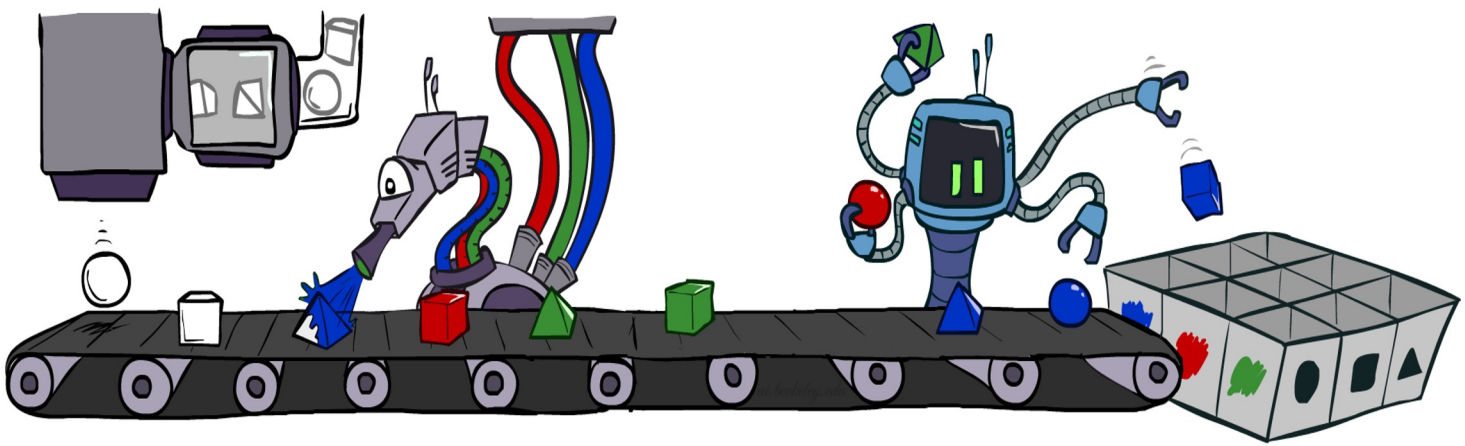
- Prior Sampling
- Rejection Sampling
- Likelihood Weighting
- Gibbs Sampling

Prior Sampling

For $i = 1, 2, \ldots, n$ in topological order

Sample x_i from $P(X_i \mid Parents(X_i))$

Return $(x_1, x_2, \ldots, x_n)$



Example: Prior Sampling

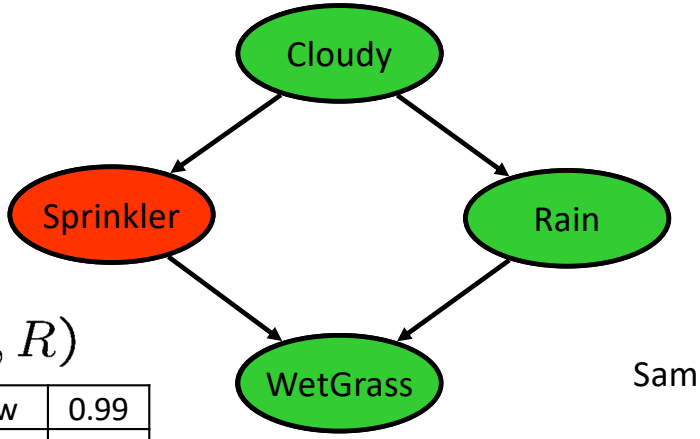
- What is $P(W)$?

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

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



$P(R|C)$

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

$P(W|S,R)$

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

Samples:

+c, -s, +r, +w

-c, +s, -r, +w

…

Example: Prior Sampling

- After several sampling, we'll get a bunch of samples from the BN:

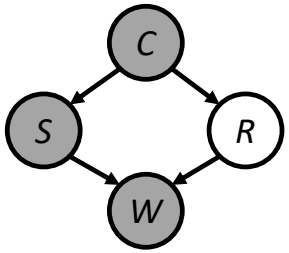
  +c, -s, +r, +w

  +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.
- What about $P(C \mid +r, +w)$?
  - Some evidences are instantiated here.

Rejection Sampling

Input: evidence instantiation

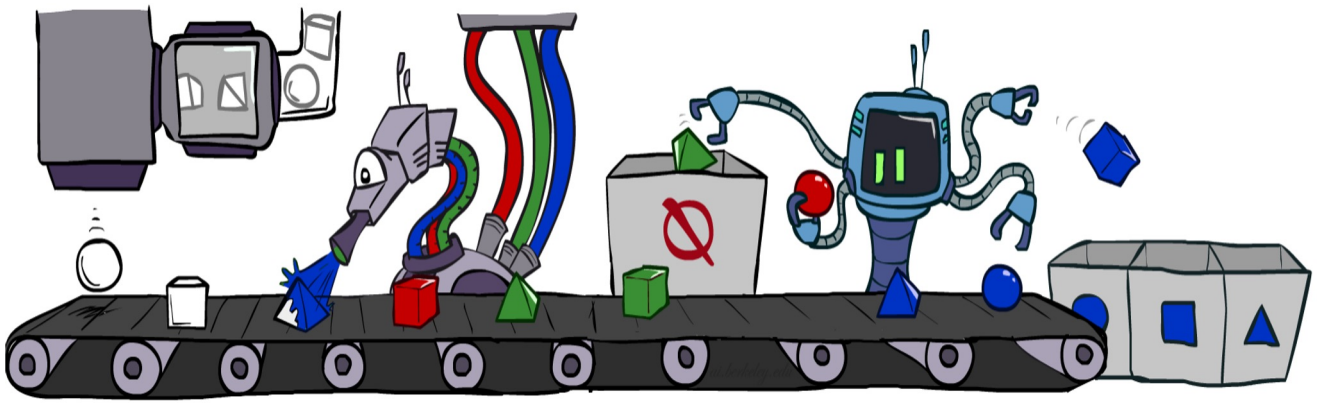
For $i = 1, 2, \ldots, n$ in topological order

Sample x_i from $P(X_i \mid Parents(X_i))$

If x_i not consistent with evidence

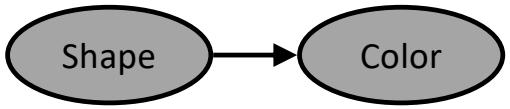
Reject (that means no sample is generated in this iteration)

Return $(x_1, x_2, \ldots, x_n)$ (exclude these rejected)



Rejection Sampling

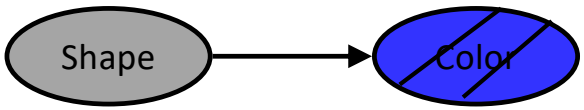
- Problem with rejection sampling:
  - If evidence is unlikely to happen, we will have to reject lots of samples!
  - Example: consider $P(\, Shape \mid blue \,)$



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

Likelihood Weighting

- Idea: fix evidence variables first, and then sample the rest.
  - Problem: sample distribution is not consistent with real distribution, because of fixed evidence.
  - Solution: weight samples by the probability of the evidence given parents.



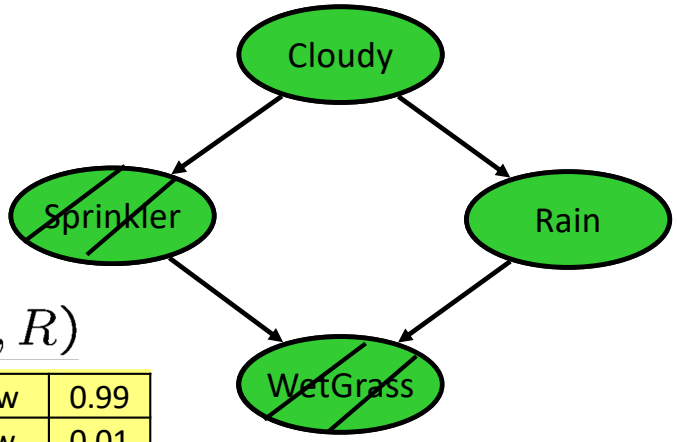
pyramid,	blue
pyramid,	blue
sphere,	blue
cube,	blue
sphere,	blue

$P(C)$

+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 (+s, +w are fixed evidences):

+c, +s, +r, +w $\quad w = 1.0$

-c, +s, -r, +w $\quad w = 1.0 \text{ x } 0.5 \text{ x } 0.90$

…

Likelihood Weighting

Input: evidence instantiation

Initial w = 1.0

for $i = 1, 2, \ldots, n$ in topological order

    if X_i is an evidence variable

        X_i = observation x_i for X_i

        Set $w = w * P(x_i \mid Parents(X_i))$

    else

        Sample x_i from $P(X_i \mid Parents(X_i))$

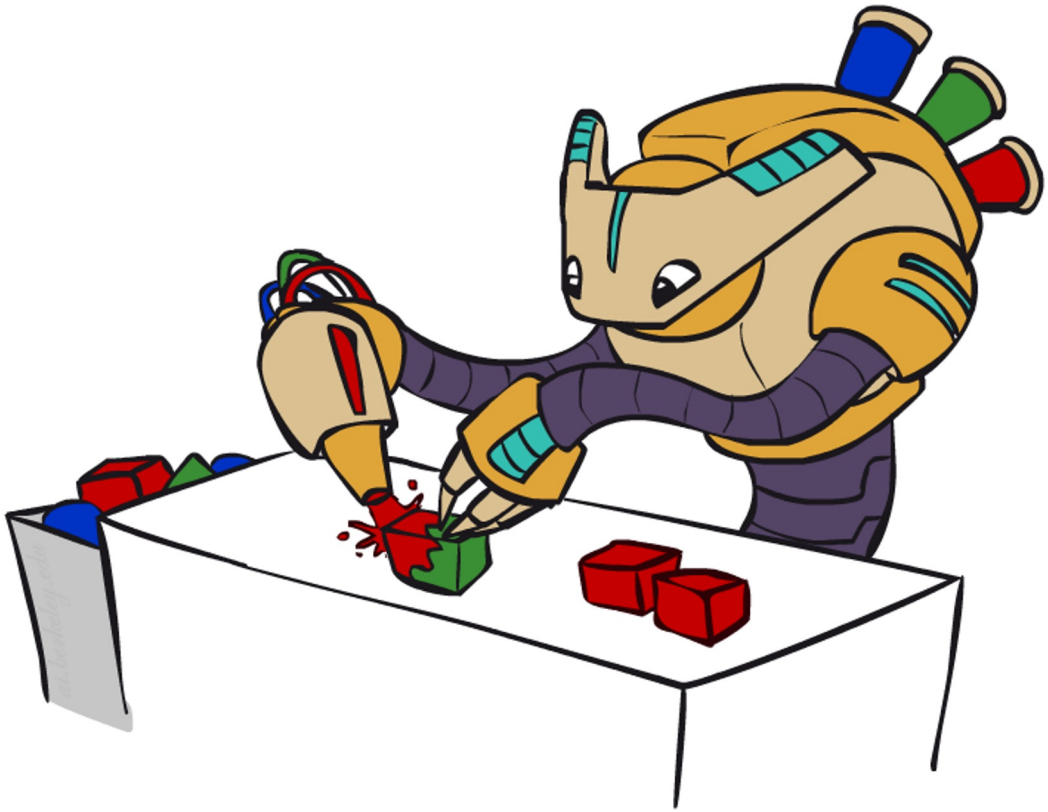
return $(x_1, x_2, \ldots, x_n)$, w's

In total every variable complies with its CPT

Likelihood Weighting

- Pros
  - All samples are used.
  - More samples will reflect the state of the world suggested by the evidence.
  - Values of downstream variables are influenced by upstream evidence.
- Cons
  - The values of upstream variables are unaffected by downstream evidence.
  - With high probability, one lucky sample will have much larger weight than the others, dominating the result.

Gibbs Sampling



Gibbs Sampling

- We start with a complete assignment of values to all variables in the network.
  - Each variable starts with an initial guess;
  - Evidence variables remain fixed throughout the process;
  - Non-evidence variables are updated during sampling.

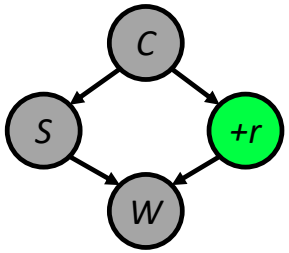
Gibbs Sampling

- Procedure of Gibbs sampling:
  - Initialize all variables with some values;
  - Repeat the following steps:
    - Pick a non-evidence variable X_i;
    - Sample a new value for X_i from its conditional distribution, given all other current variable values: $X_i' \sim P(X_i \mid x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$.
    - In a Bayes net, $P(X_i \mid x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) = P(X_i \mid \mathrm{MarkovBlanket}(X_i))$.
- After enough iterations, the sequence of sampled states will approximate the true joint distribution $P(X)$.
  - These samples can then be used to estimate probabilities, expectations, or make predictions.

Example: Gibbs Sampling

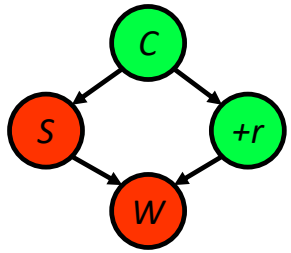
1: Fix evidence

- *R* = +r



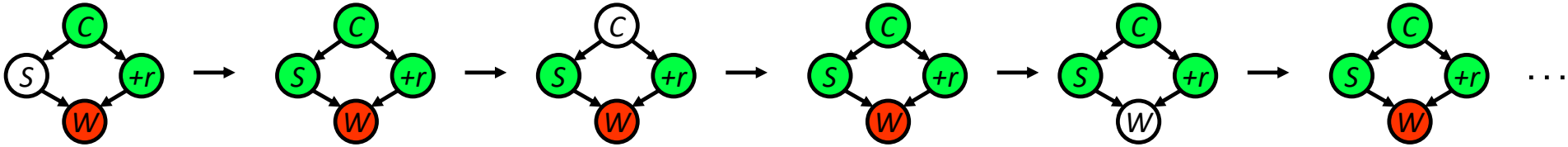
2: Initialize other variables

- Randomize



3: Repeat:

- Choose a non-evidence variable *X*
- Resample *X* from *P*(*X* | MarkovBlanket(*X*))



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

Markov Chain Monte Carlo

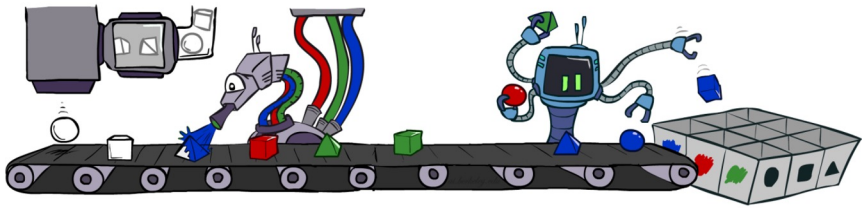
- MCMC methods are used to approximate complex probability distributions by generating samples, especially when exact computation is infeasible due to a large state space.
- Markov Chain
  - A sequence of states where each state is chosen based only on the previous one.
  - E.g., A full assignment to all variables in Gibbs sampling can be perceived as a state.
- Monte Carlo
  - A sampling-based algorithm that uses randomness to estimate answers.
  - May produce slightly inaccurate results, but becomes reliable with enough samples.

Gibbs Sampling: MCMC Example

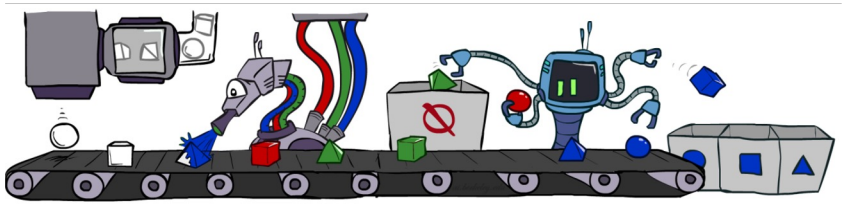
- Gibbs Sampling is a specific type of MCMC method.
  - Why Monte Carlo?
    - It uses randomness to generate samples;
    - It approximates expectations or probabilities using those samples.
  - It improves sampling by using a Markov chain to generate samples that are dependent but still converge to the target distribution.
    - This avoids the inefficiency of naive random sampling in high-dimensional spaces.

Summary: Bayes Net Sampling

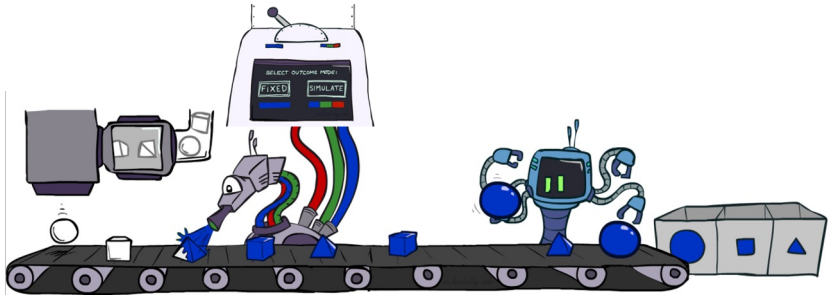
Prior Sampling P(Q)



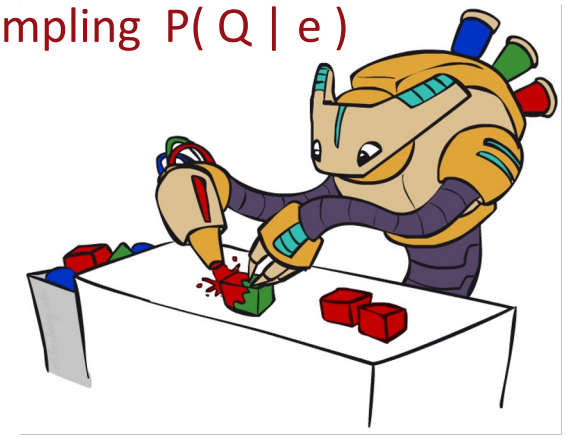
Rejection Sampling P(Q | e)



Likelihood Weighting P(Q | e)



Gibbs Sampling P(Q | e)



Thank You!