Bayes Nets: Independence

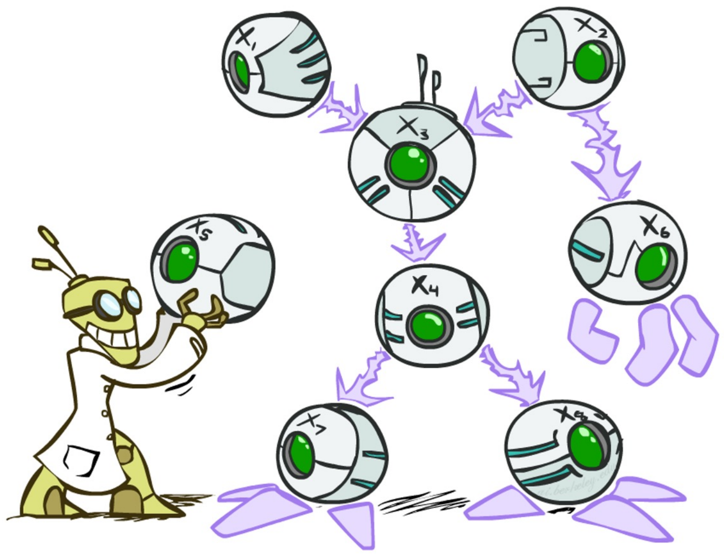
Instructor: Evgeny Pobachienko — UC Berkeley

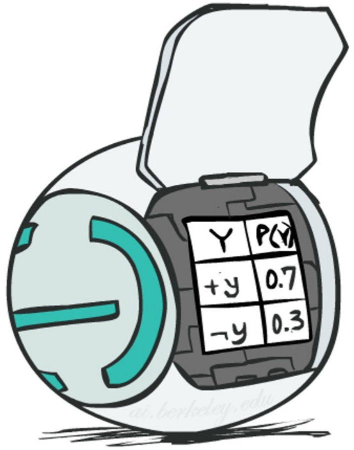
[Slides credit: Dan Klein, Pieter Abbeel, Anca Dragan, Stuart Russell, Satish Rao, and many others]

Bayesian Networks: Recall…

- A directed acyclic graph (DAG), one node per random variable
- A conditional probability table (CPT) for each node
  - Probability of X, given a combination of values for parents. $P(X|a_1 \ldots a_n)$
- Bayes nets implicitly encode joint distributions as a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

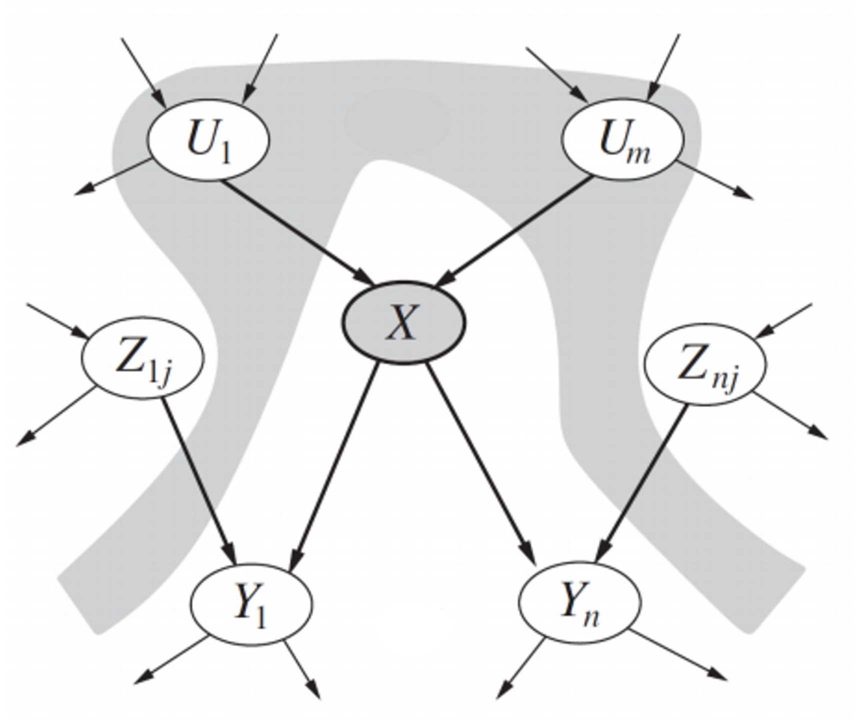
$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\mathit{parents}(X_i))$$



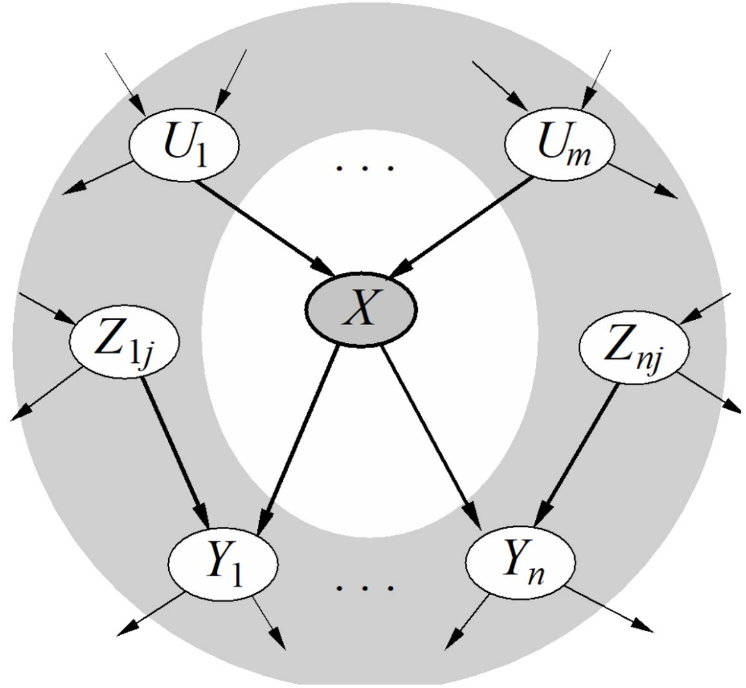


Independence Assumptions so far…

- Each node, given its parents, is conditionally independent of all its non-descendants in the graph



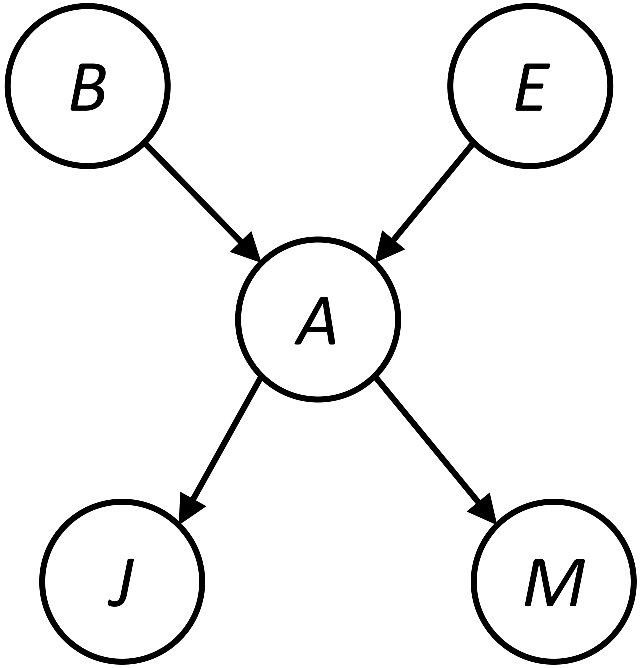
Each node, given its MarkovBlanket, is conditionally independent of all other nodes in the graph



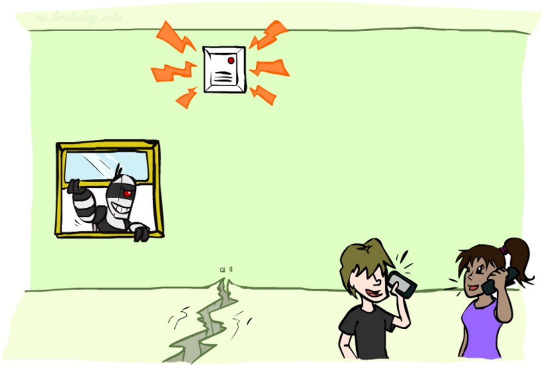
MarkovBlanket refers to the parents, children, and children's other parents.

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

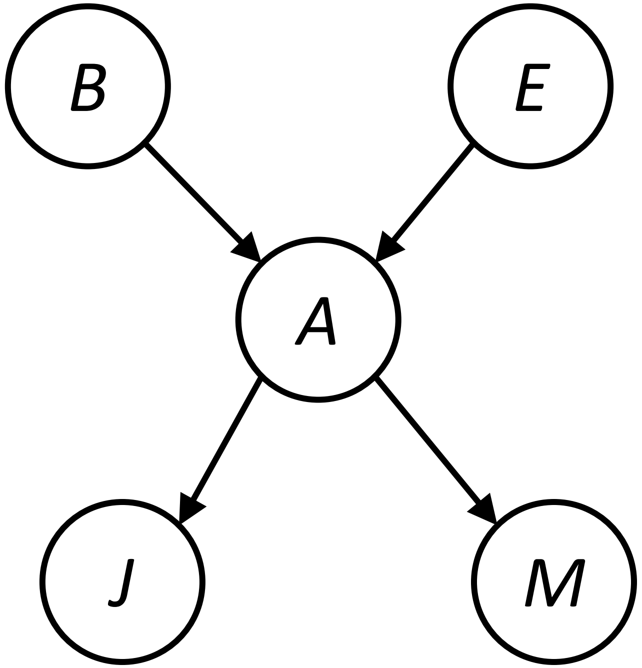
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

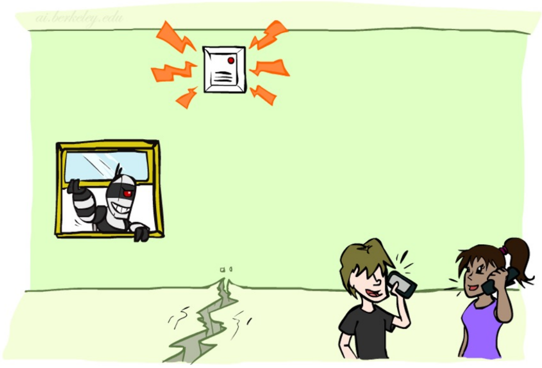
$$P(+b, -e, +a, -j, +m) =$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

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

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

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

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

$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Conditional Independence

- X and Y are independent iff

$$\forall x, y \;\; P(x,y) = P(x)P(y) \qquad\qquad X \perp\!\!\!\perp Y$$

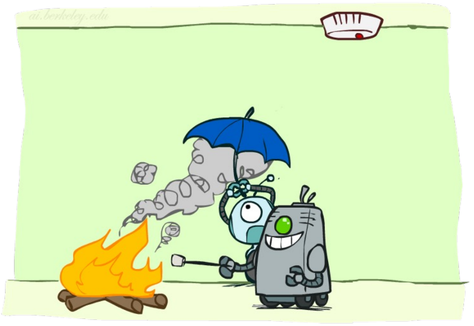
- Given Z, we say X and Y are conditionally independent iff

$$\forall x, y, z \;\; P(x,y|z) = P(x|z)P(y|z) \dashrightarrow X \perp\!\!\!\perp Y | Z$$

- (Conditional) independence is a property of a distribution

- Example:

$$Alarm \perp\!\!\!\perp Fire | Smoke$$

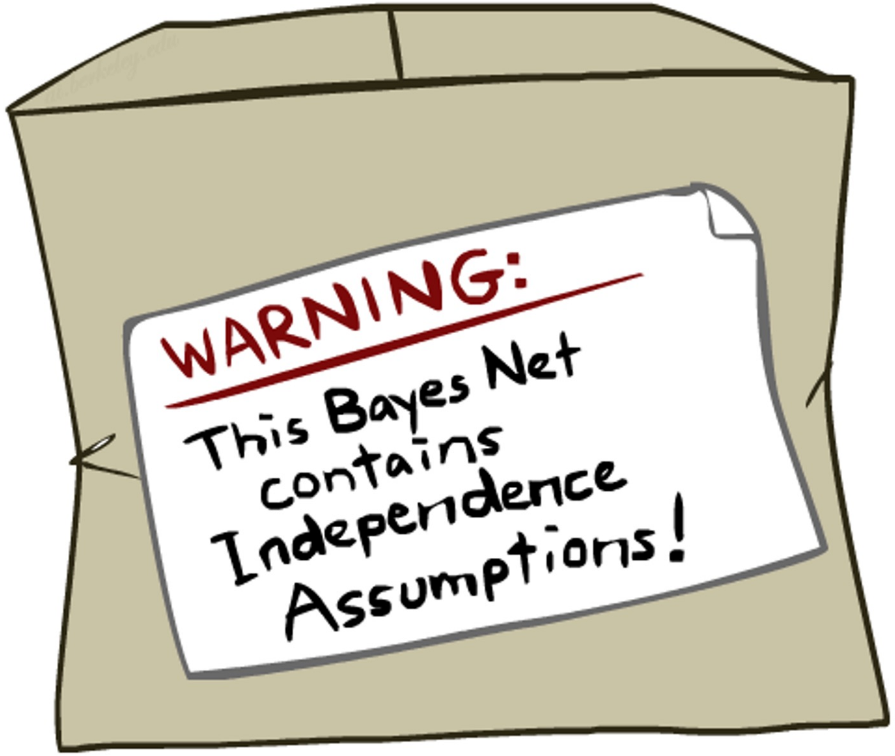


Bayes Nets: Assumptions

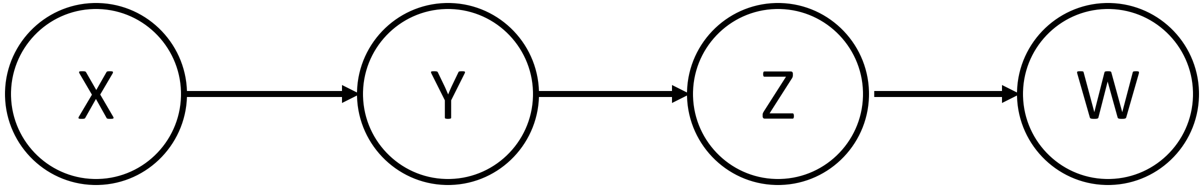
- Assumptions we are required to make to define the Bayes net when given the graph:

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

- Important for modeling: understand assumptions made when choosing a Bayes net graph



Example



- Conditional independence assumptions directly from simplifications in chain rule:

$$X \perp\!\!\!\perp Z | Y \qquad P(x,y,z,w) = P(x)P(y|x)P(z|x,y)P(w|x,y,z)$$

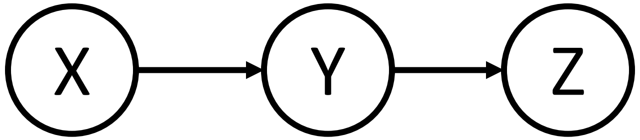
$$W \perp\!\!\!\perp \{X,Y\} | Z \qquad P(x,y,z,w) = P(x)P(y|x)P(z|y)P(w|z)$$

- Additional implied conditional independence assumptions?

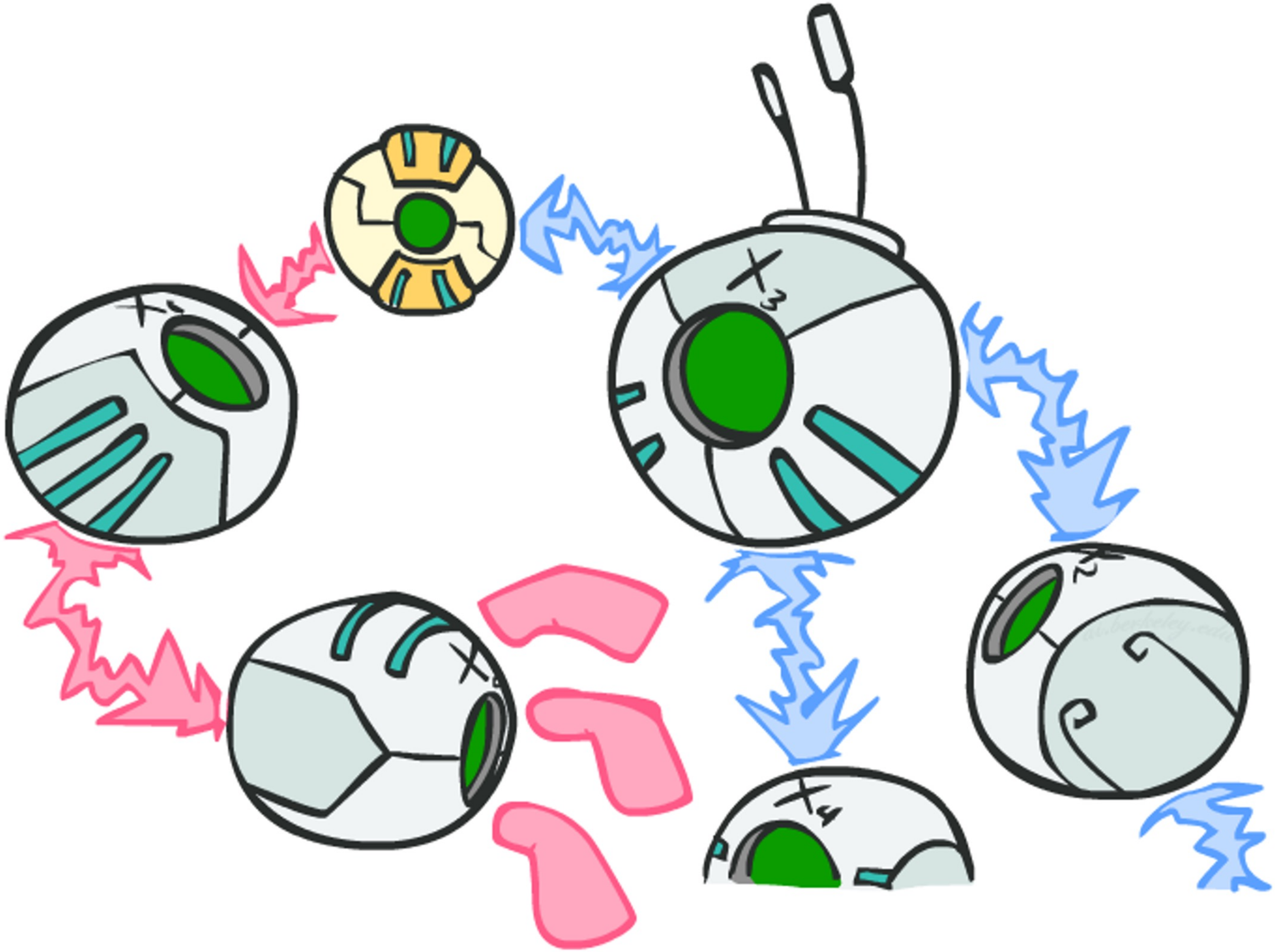
$$W \perp\!\!\!\perp X | Y$$

Independence in a BN

- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - Question: are X and Z guaranteed to be independent?
    - Answer: no. Example: low pressure causes rain, which causes traffic.
    - X can influence Z, Z can influence X (via Y)
    - Addendum: they *could* be independent: how?



D-separation: Outline



D-separation: Outline

- Study independence properties for triples
  - Why triples?

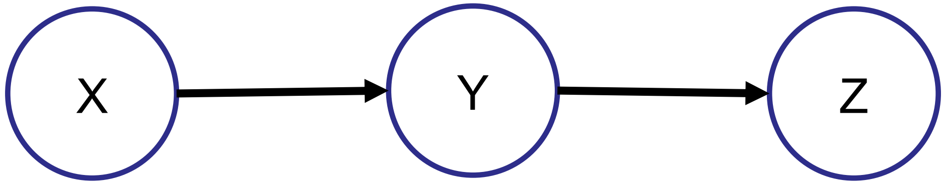
- Analyze complex cases in terms of member triples

- D-separation: a condition / algorithm for answering such queries

Causal Chains

- This configuration is a "causal chain"

X: Low pressure Y: Rain Z: Traffic



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

Is X guaranteed to be independent of Z?

No!

One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.

Example:

Low pressure causes rain causes traffic, high pressure causes no rain causes no traffic

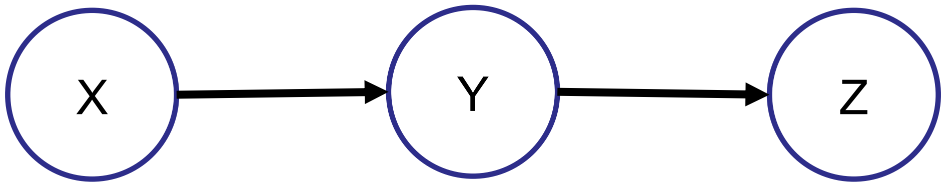
In numbers:

P(+y | +x) = 1, P(-y | - x) = 1,
P(+z | +y) = 1, P(-z | -y) = 1

Causal Chains

- This configuration is a "causal chain"

X: Low pressure Y: Rain Z: Traffic



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

Given Y, is X guaranteed to be independent of Z?

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

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

$$= P(z|y)$$

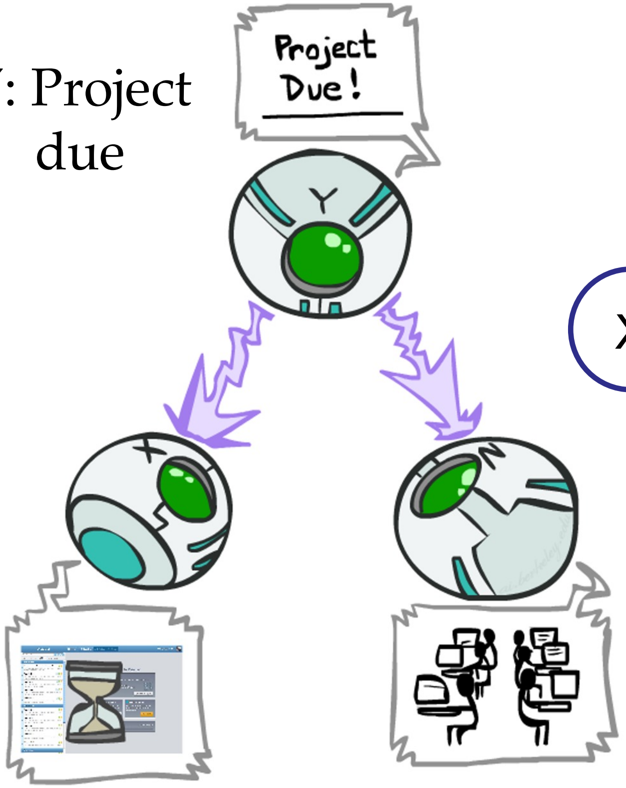
Yes!

Evidence along the chain "blocks" the influence

Common Causes

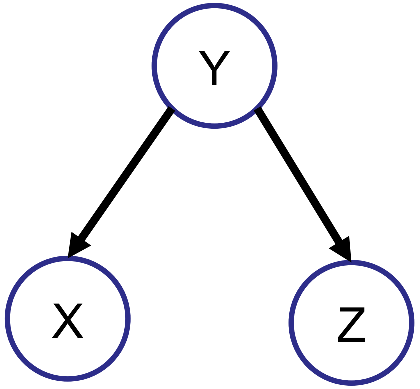
- This configuration is a "common cause"

Y: Project due



X: Forums busy

Z: Lab full



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

Guaranteed X independent of Z ?

No!

One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.

Example:

Project due causes both forums busy and lab full

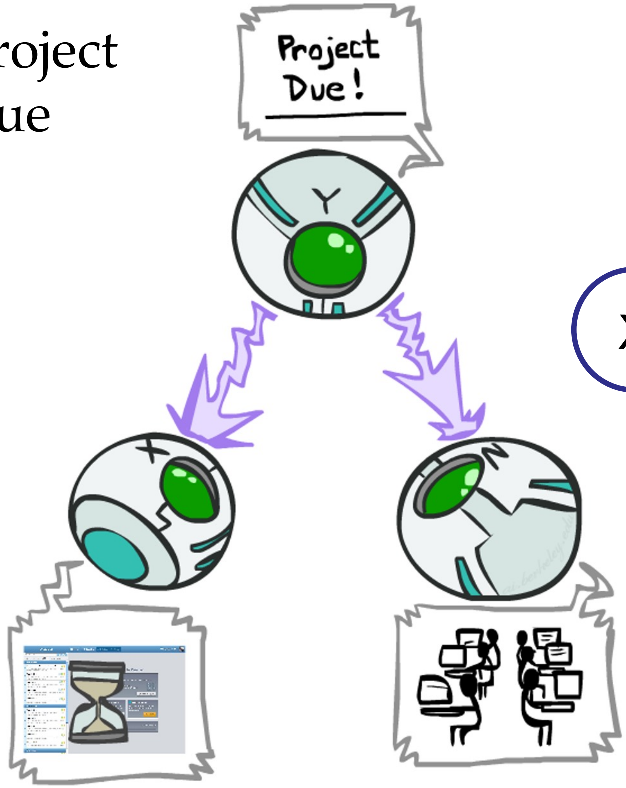
In numbers:

P(+x | +y) = 1, P(-x | -y) = 1,
P(+z | +y) = 1, P(-z | -y) = 1

Common Cause

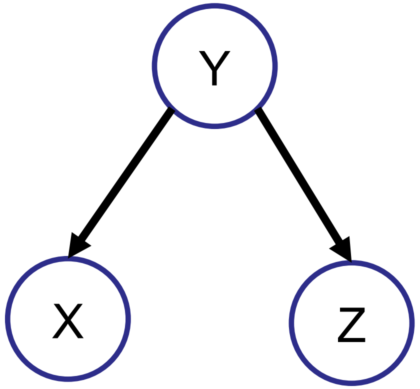
- This configuration is a “common cause”

Y: Project due



X: Forums busy

Z: Lab full



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

Guaranteed X and Z independent given Y?

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

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

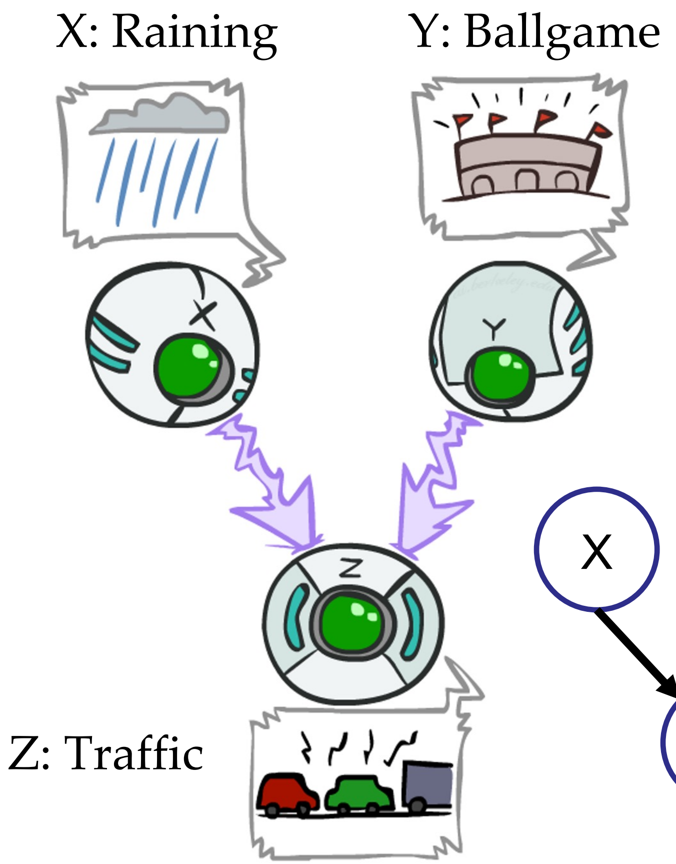
$$= P(z|y)$$

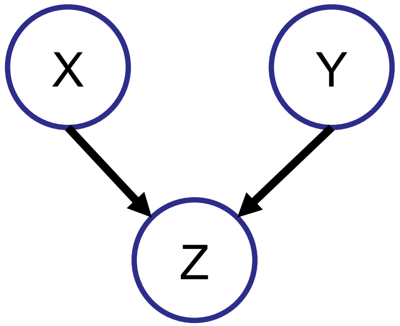
Yes!

Observing the cause blocks influence between effects.

Common Effect

- Last configuration: two causes of one effect (v-structures)





Are X and Y independent?

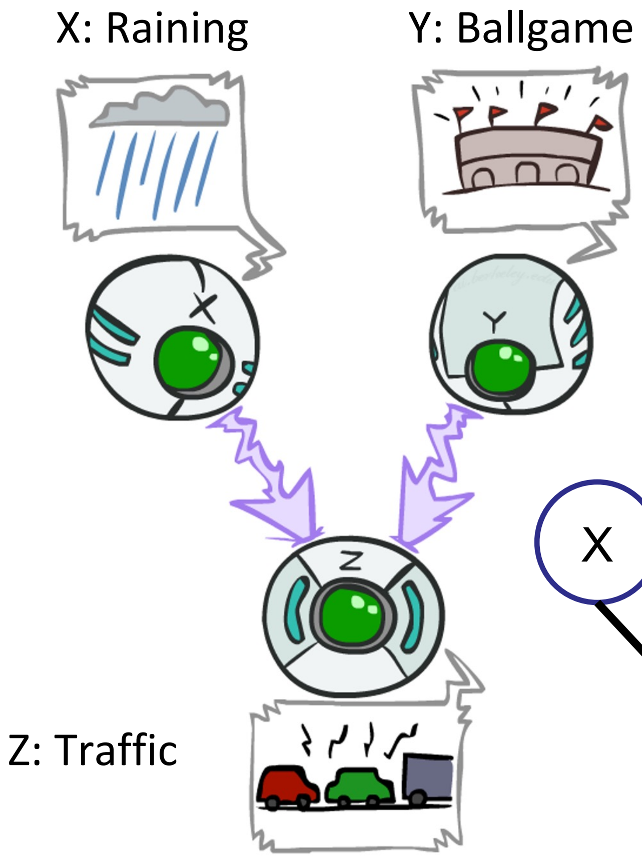
Yes: the ballgame and the rain cause traffic, but they are not correlated

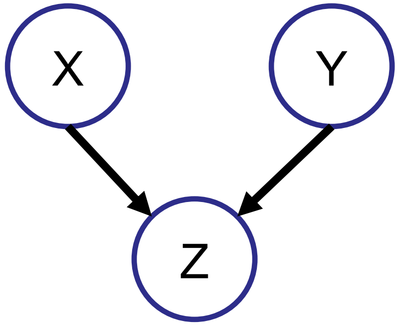
Proof:

$$P(x, y) = \sum P(x, y, z)$$

Common Effect

- Last configuration: two causes of one effect (v-structures)





Are X and Y independent?

Yes: the ballgame and the rain cause traffic, but they are not correlated

(Proved previously)

Are X and Y independent given Z?

No: seeing traffic puts the rain and the ballgame in competition as explanation.

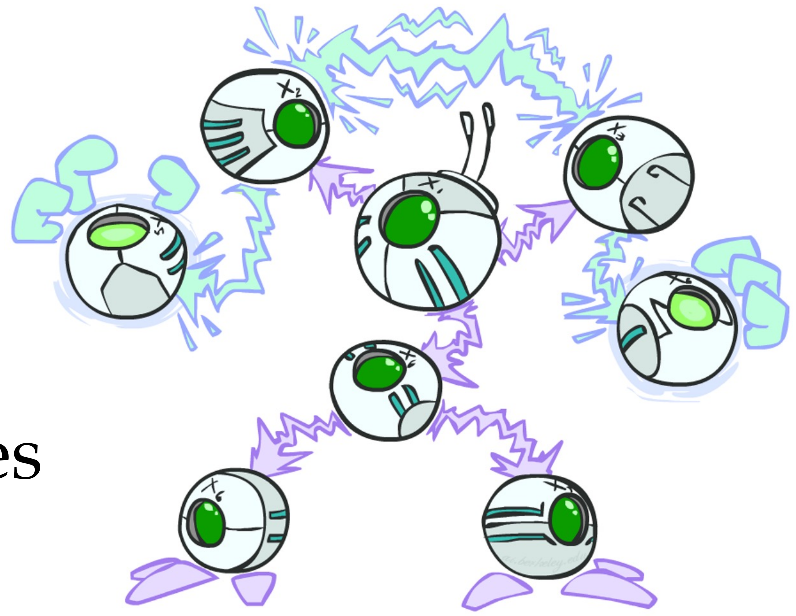
This is backwards from the other cases

Observing an effect activates influence between possible causes.

The General Case

The General Case

- General question: in a given BN, are two variables independent (given evidence)?
- Solution: analyze the graph
- Any complex example can be broken into repetitions of the three canonical cases

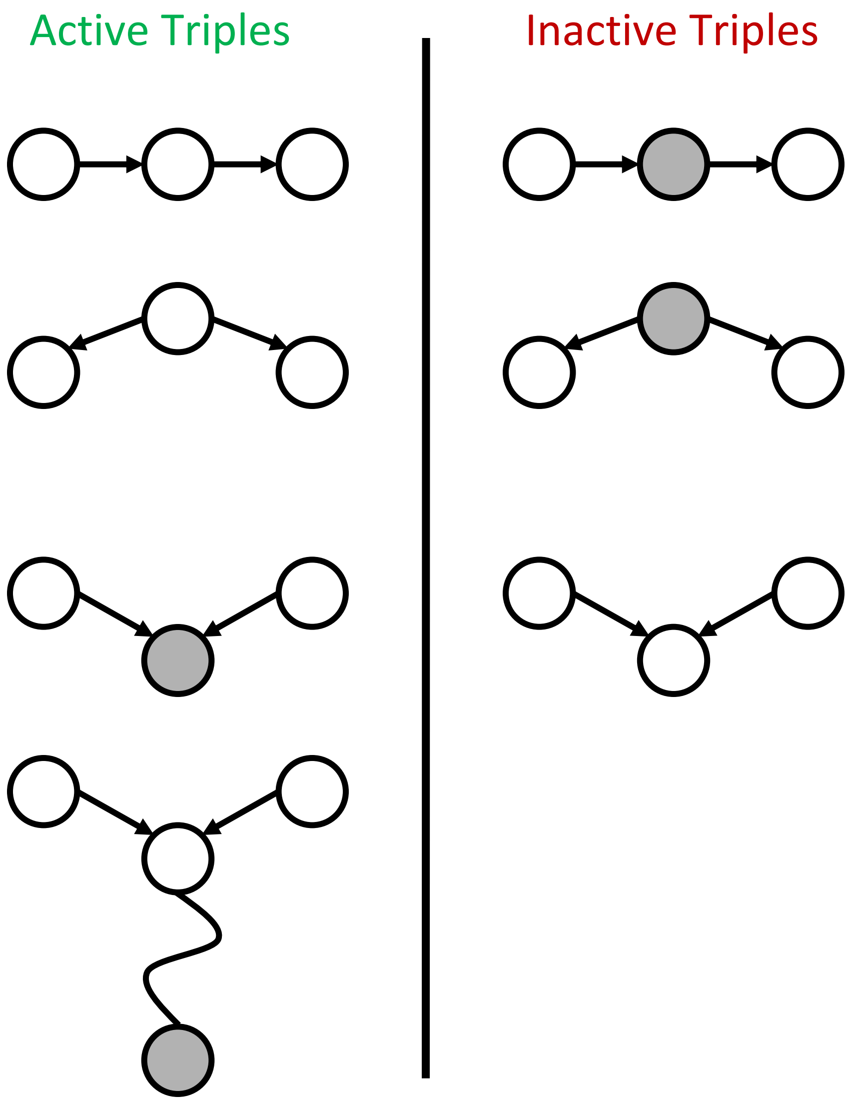


Active / Inactive Paths

- Question: Are X and Y conditionally independent given evidence variables {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 A -> B -> C where B is unobserved (either direction)
  - Common cause A <- B -> C where B is unobserved
  - Common effect (aka v-structure)
    A -> B <- C where B *or one of its descendants* is observed

- All it takes to block a path is a single inactive segment



D-Separation

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

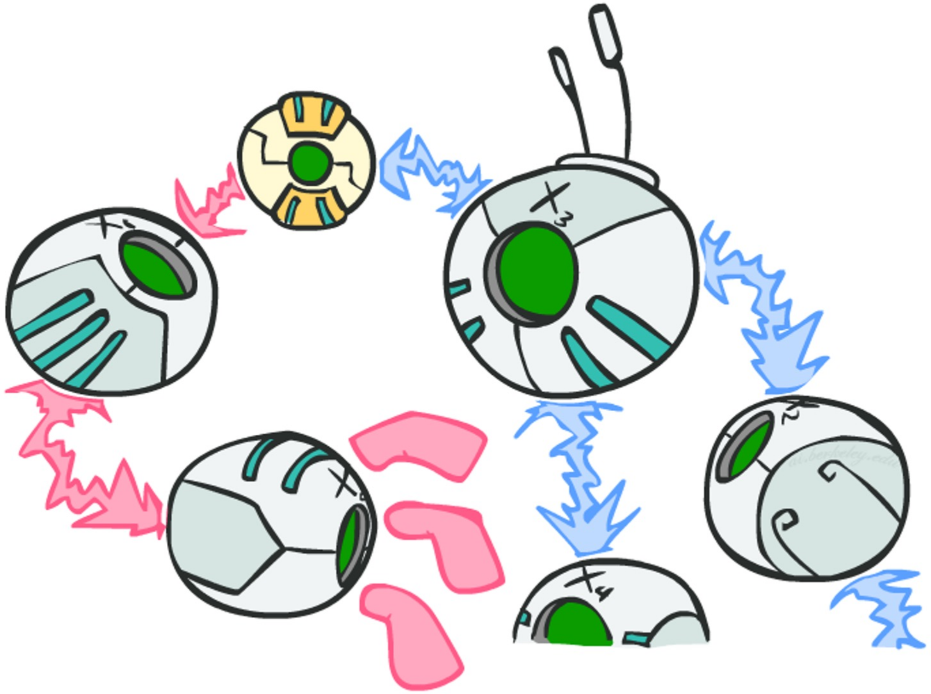
Check all (undirected!) paths between X_i and X_j

If one or more active paths, then independence not guaranteed

$$X_i \not\perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

Otherwise (i.e. if all paths are inactive), then independence is guaranteed

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

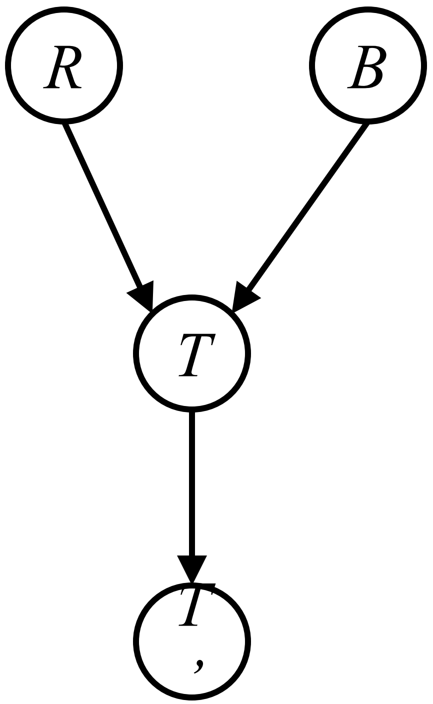


Example

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

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

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



Example

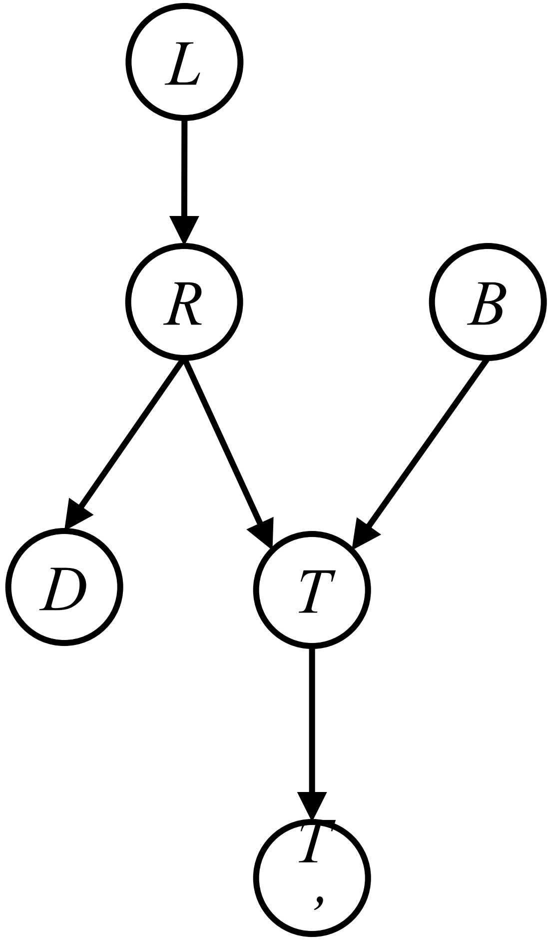
$L \perp\!\!\!\perp T' | T$ Yes

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

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

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

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



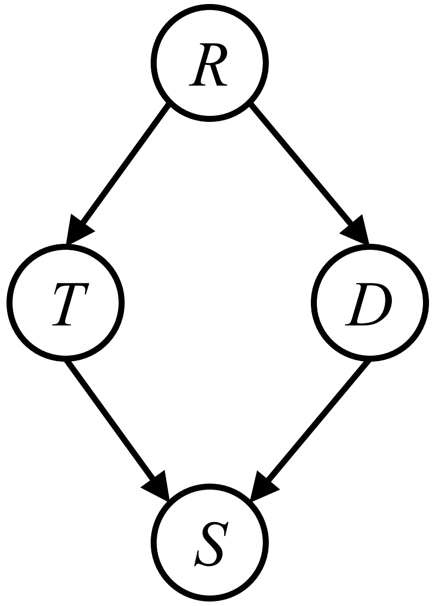
Example

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

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

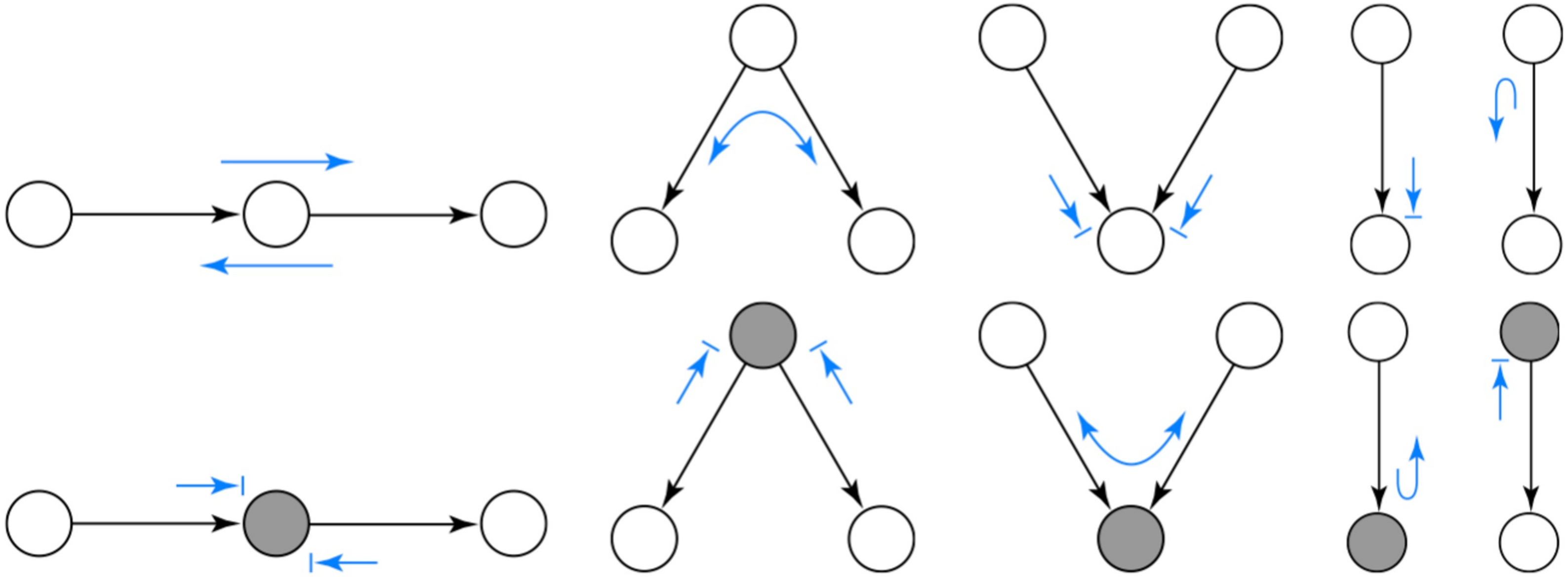
$T \perp\!\!\!\perp D | R$ *Yes*

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



Another Perspective: Bayes Ball

An undirected path is active if a Bayes ball travelling along it never encounters the "stop" symbol: $\longrightarrow\!\!|$



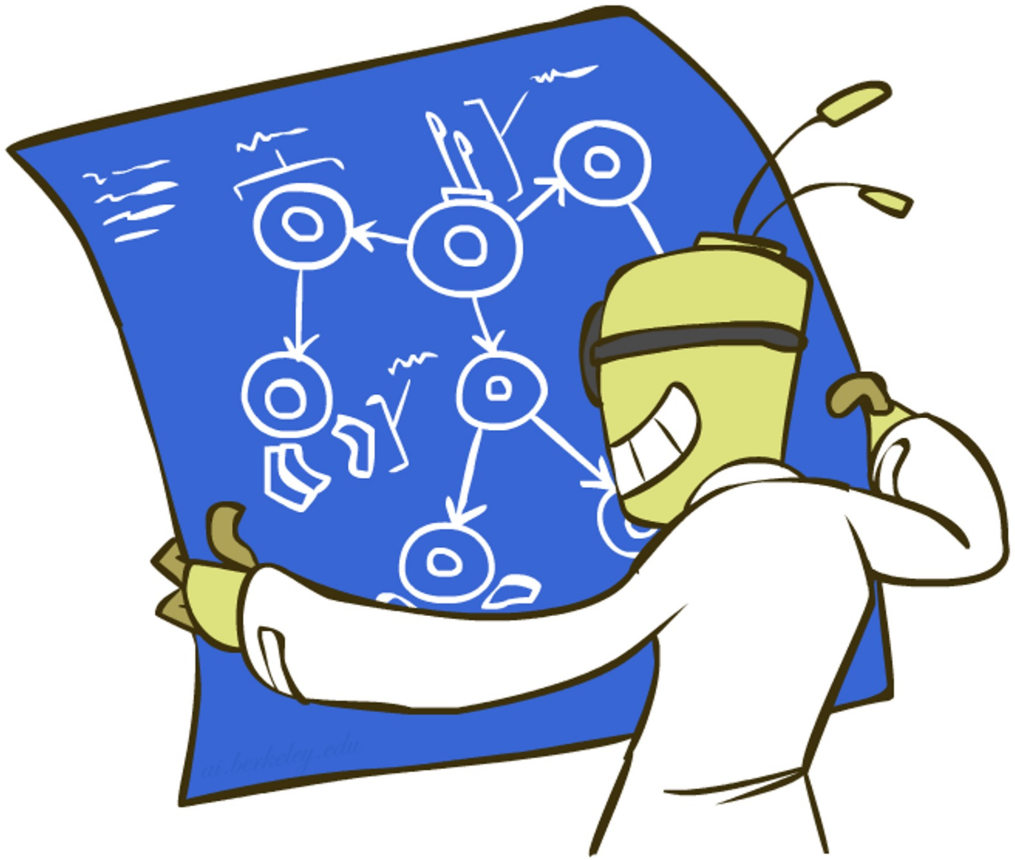
If there are no active paths from X to Y when $\{Z_1, \ldots, Z_k\}$ are shaded, then $X \perp\!\!\!\perp Y \mid \{Z_1, \ldots, Z_k\}$.

Structure Implications

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

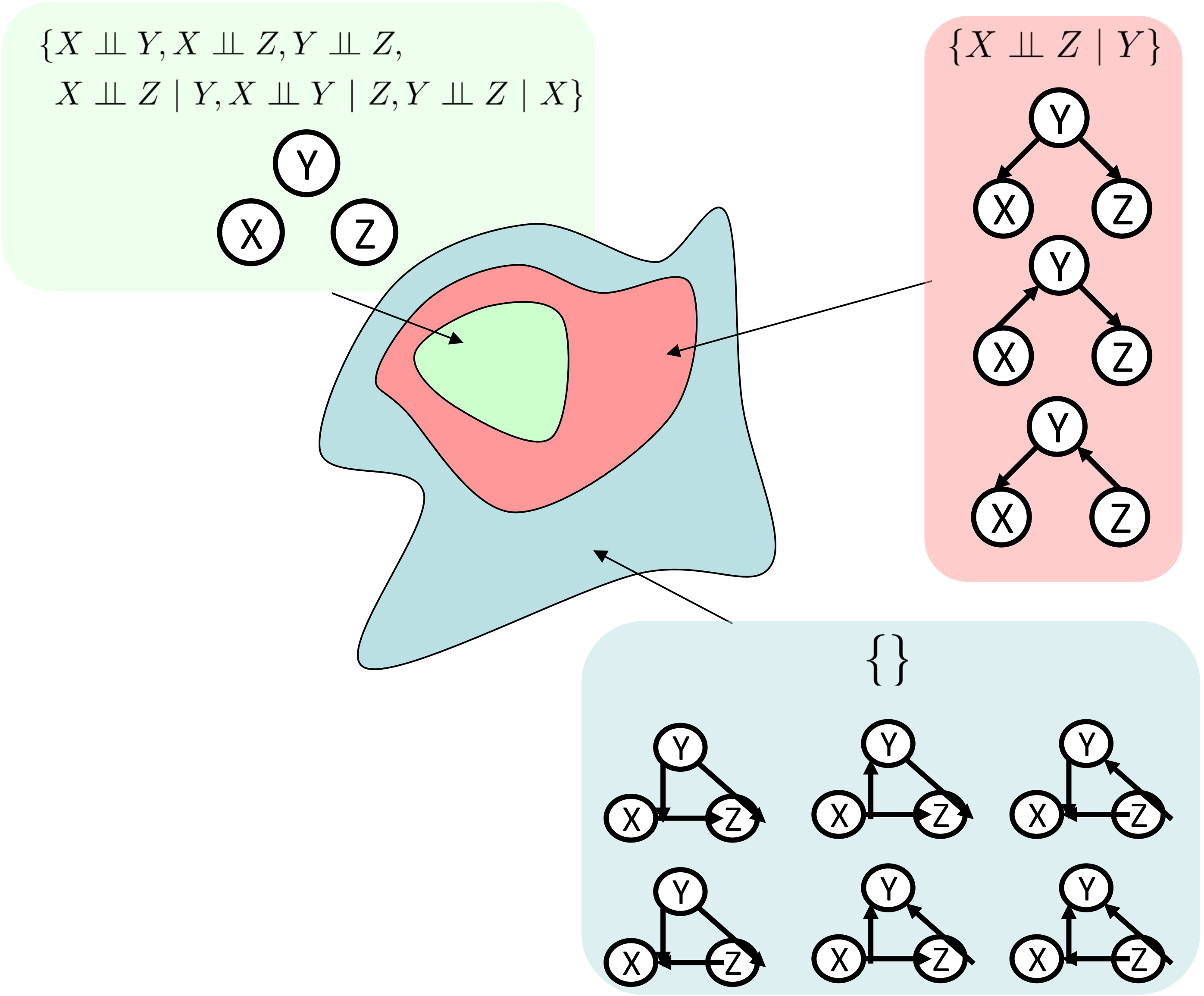
$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, \ldots, X_{k_n}\}$$

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



Topology Limits Distributions

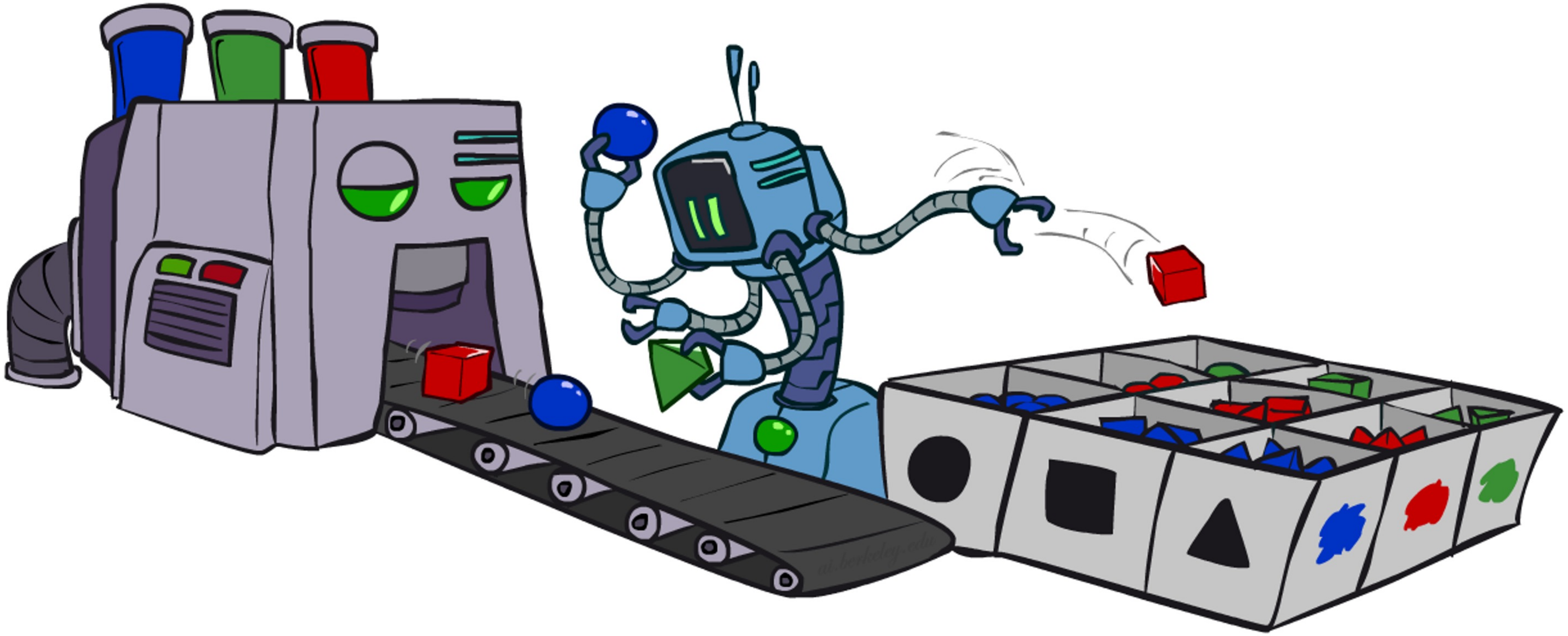
- Given some graph topology G, only certain joint distributions can be encoded
- The graph structure guarantees certain (conditional) independences
- (There might be more independence)
- Adding arcs increases the set of distributions, but has several costs
- Full conditioning can encode any distribution



Bayes Nets Representation Summary

- Bayes nets compactly encode joint distributions (by making use of conditional independences!)
- Guaranteed independencies of distributions can be deduced from BN graph structure
- D-separation gives precise conditional independence guarantees from graph alone
- A Bayes net's joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution

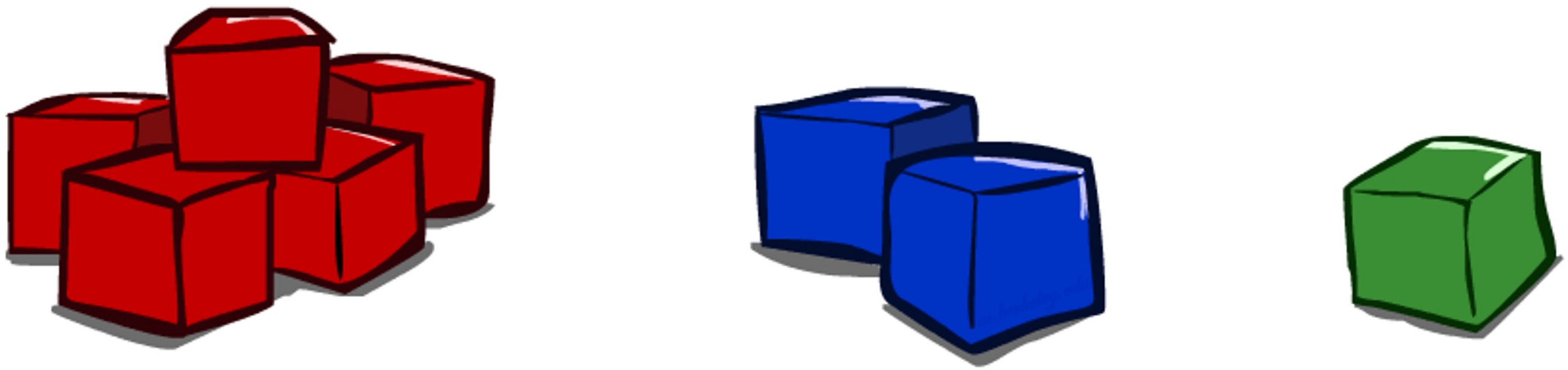
Bayesian Networks: Sampling



Instructor: Evgeny Pobachienko – UC Berkeley

[Slides credit: Dan Klein, Pieter Abbeel, Anca Dragan, Stuart Russell, Ketrina Yim, and many others]

Approximate Inference: Sampling



Sampling

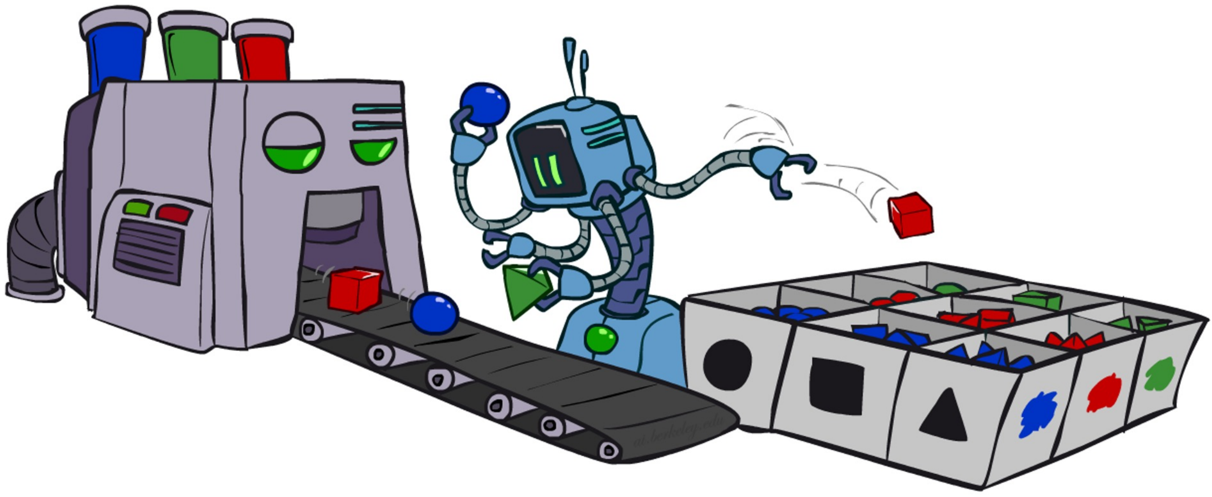
- Sampling is a lot like repeated simulation
  - Predicting the weather, basketball games, …

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

Why sample?

Learning: get samples from a distribution you don't know

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



Sampling

- Sampling from given distribution
  - Step 1: Get sample u from uniform distribution over [0, 1)
    - E.g. random() in python
  - Step 2: Convert this sample u into an outcome for the given distribution by having each target outcome associated with a sub-interval of [0,1) with sub-interval size equal to probability of the outcome

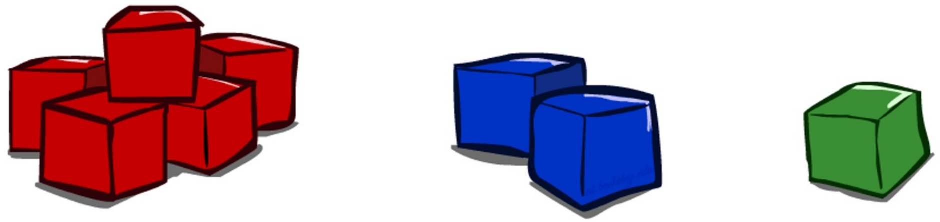
Example

C	P(C)
red	0.6
green	0.1
blue	0.3

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

If random() returns $u = 0.83$, then our sample is C = blue

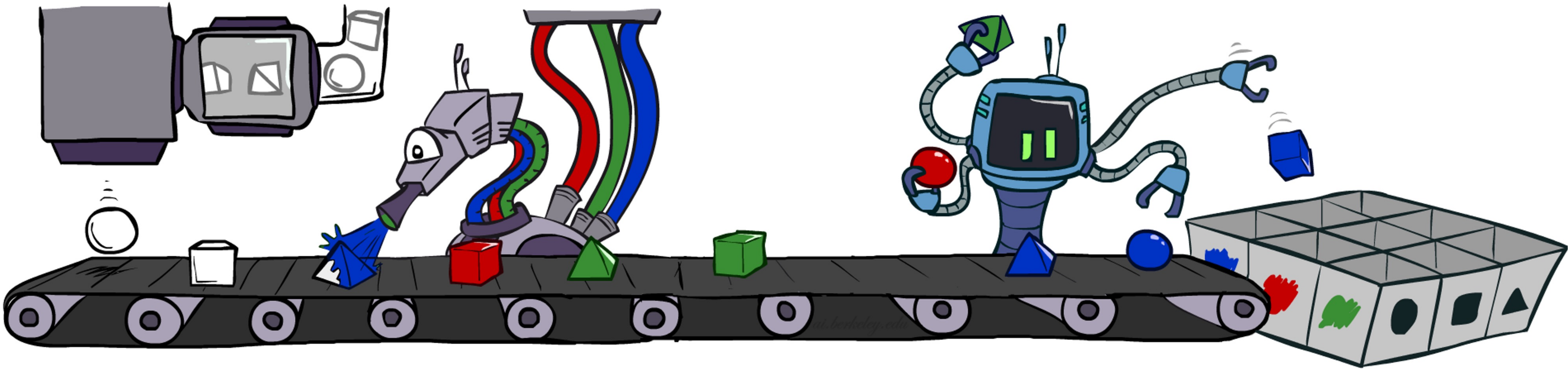
E.g, after sampling 8 times:



Sampling in Bayes' Nets

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

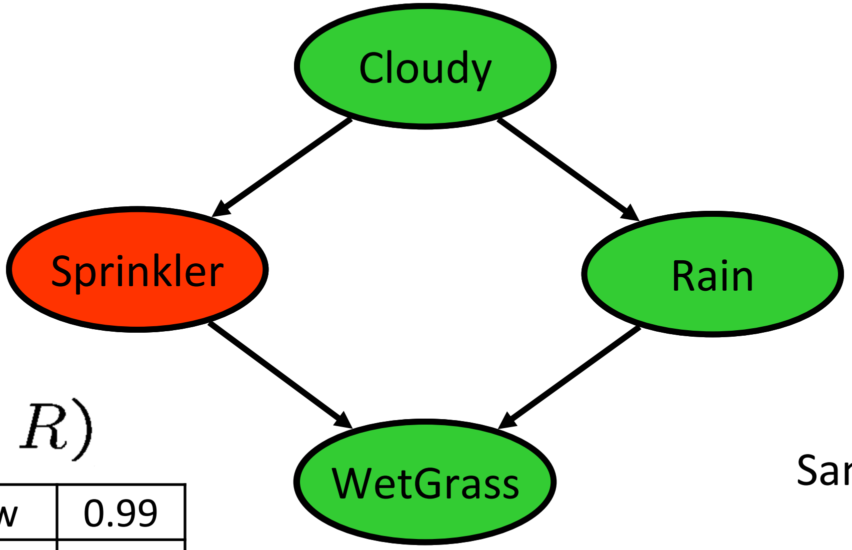
Prior Sampling



Prior Sampling

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

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

$P(R|C)$

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

$P(W|S, R)$

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

Samples:

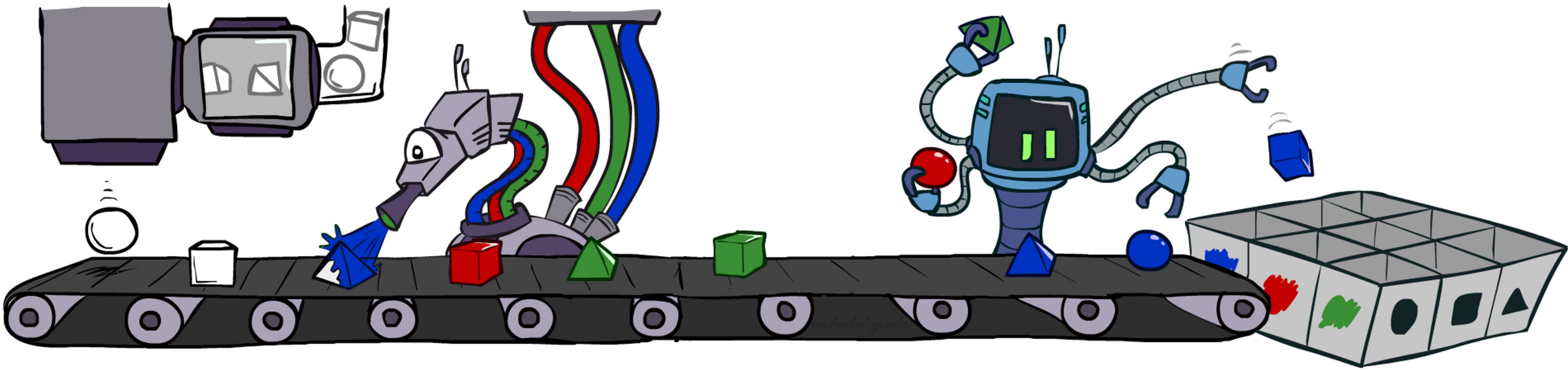
+c, -s, +r, +w

-c, +s, -r, +w

…

Prior Sampling

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



Prior Sampling

- This process generates samples with probability:

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

…i.e. the BN's joint probability

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

- Then

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

- I.e., the sampling procedure is consistent

Example

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

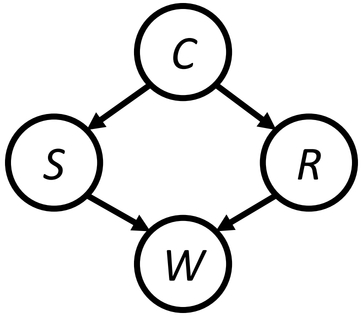
  +c, -s, +r, +w

  +c, +s, +r, +w

  -c, +s, +r, -w

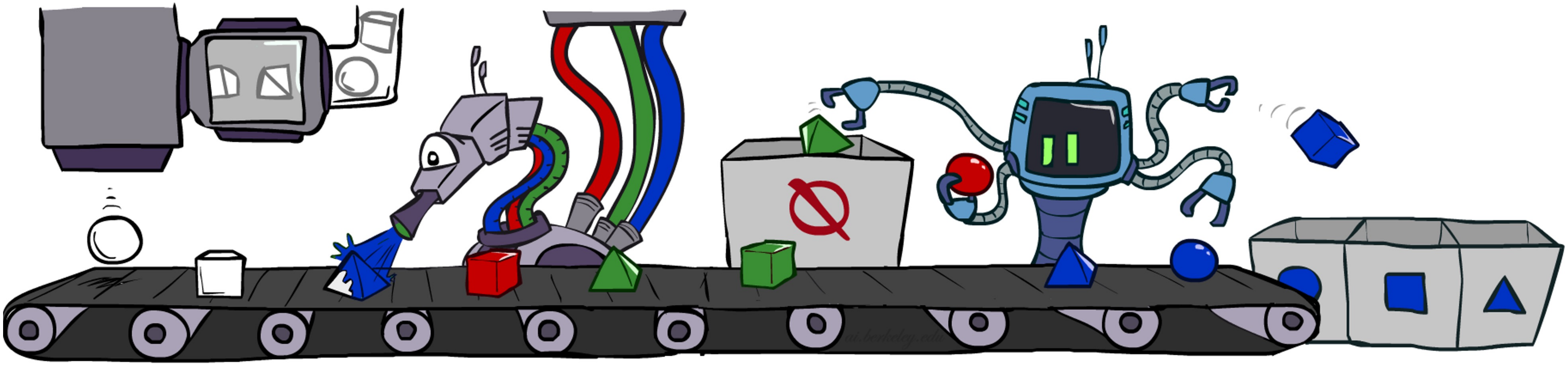
  +c, -s, +r, +w

  -c, -s, -r, +w



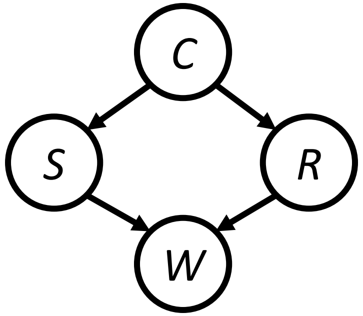
- If we want to know P(W)
  - We have counts <+w:4, -w:1>
  - Normalize to get P(W) = <+w:0.8, -w:0.2>
  - This will get closer to the true distribution with more samples
- What about P(C | +r, +w)?

Rejection Sampling



Rejection Sampling

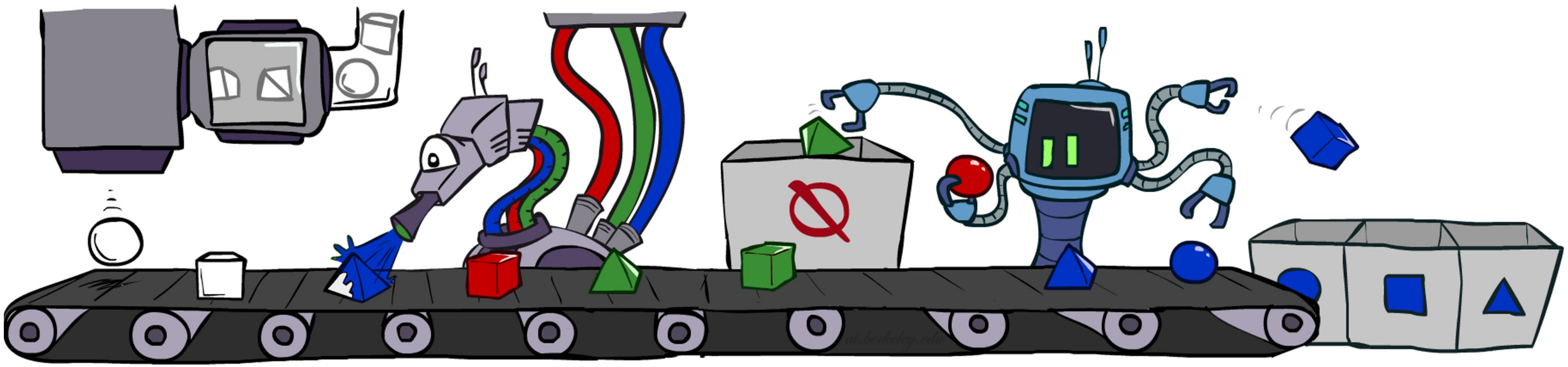
- Let's say we want P(C)
  - Just tally counts of C as we go



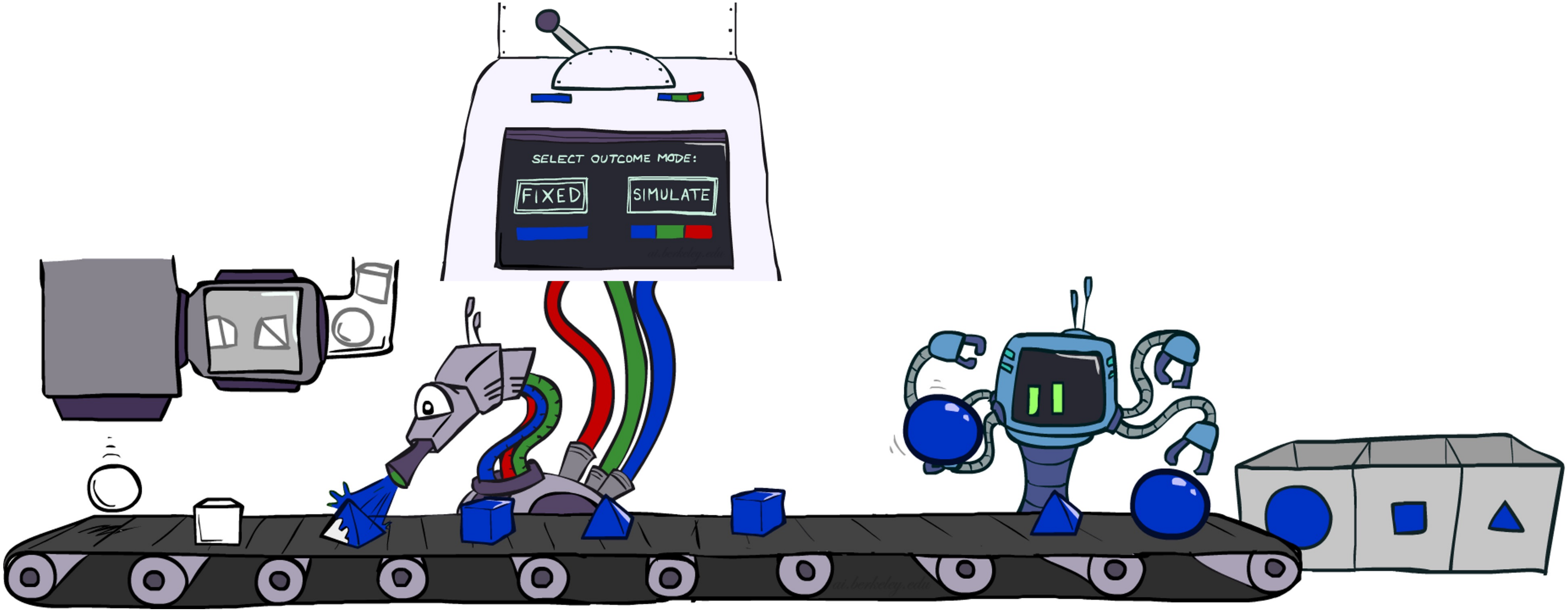
- Let's say we want P(C | +s)
  - Same thing: tally C outcomes, but ignore (reject) samples which don't have S=+s
  - We can toss out samples early!
  - It is also consistent for conditional probabilities (i.e., correct in the limit)

Rejection Sampling

- Input: evidence instantiation
- For i = 1, 2, …, n in topological order
  - Sample x_i from $P(X_i \mid \text{Parents}(X_i))$
  - If x_i not consistent with evidence
    - Reject: return – no sample is generated in this cycle
- Return $(x_1, x_2, \ldots, x_n)$

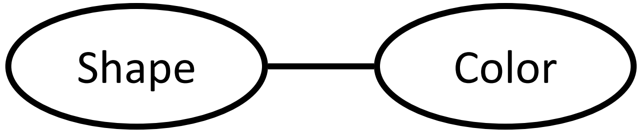


Likelihood Weighting

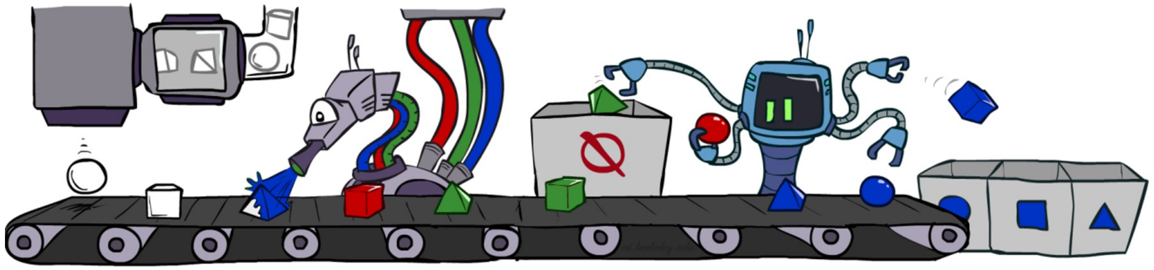


Likelihood Weighting

- Problem with rejection sampling:
  - If evidence is unlikely, rejects lots of samples
  - Consider P(Shape | blue)



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

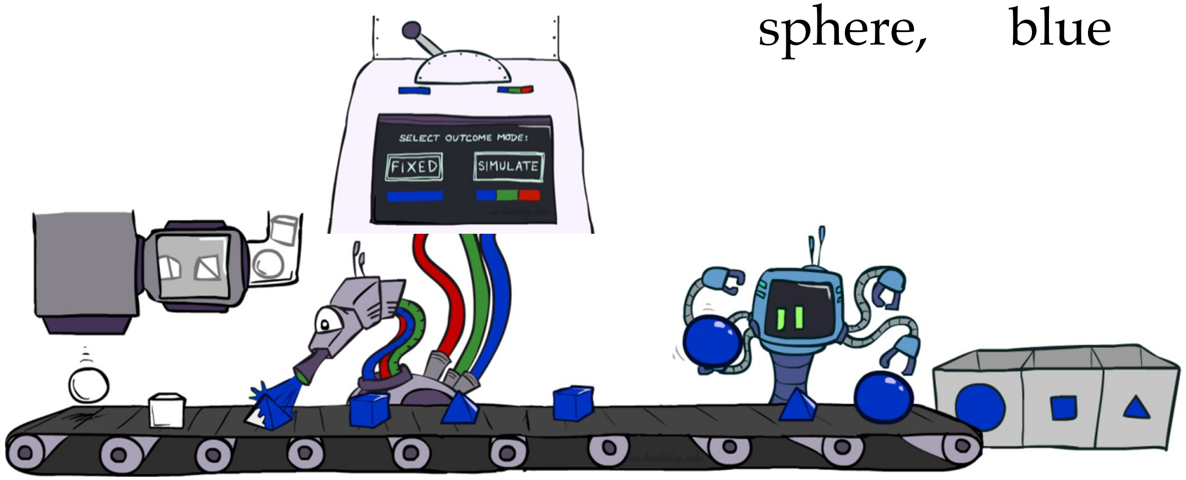


Idea: fix evidence variables and sample the rest

Problem: sample distribution not consistent!

Solution: weight by probability of evidence given parents

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



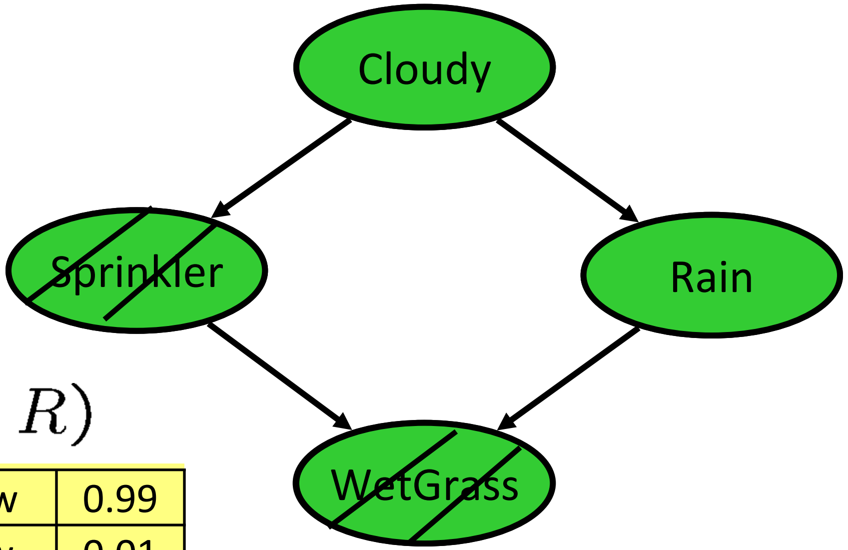
Likelihood Weighting

$P(C)$

+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:

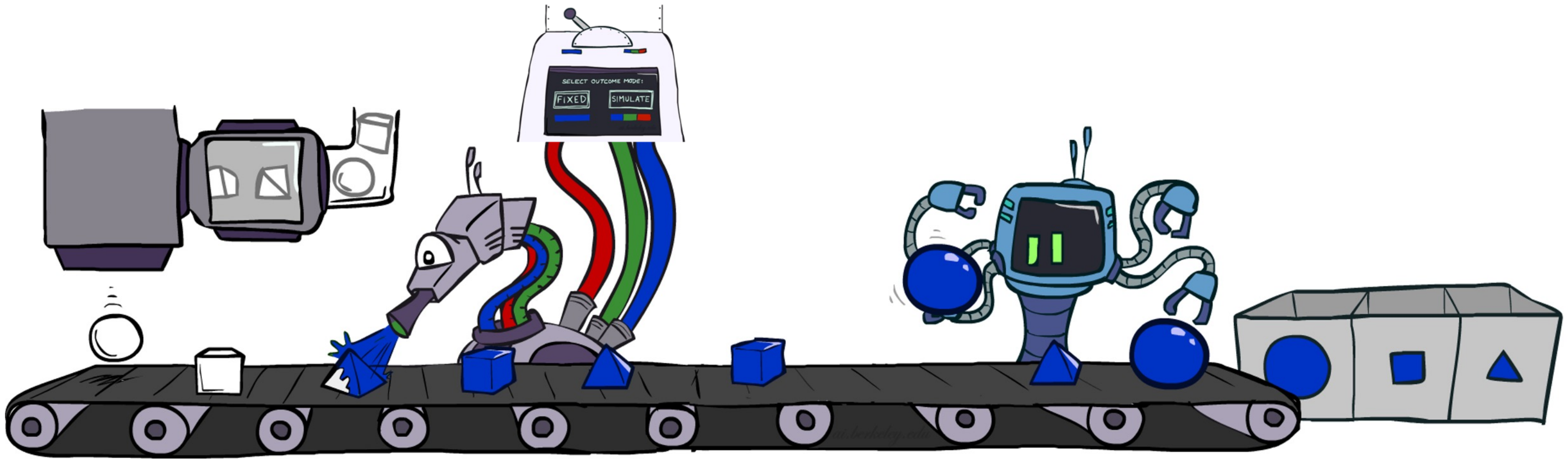
+c, +s, +r, +w   w = 1.0

-c, +s, -r, +w   w = 1.0 x 0.5 x 0.90

…

Likelihood Weighting

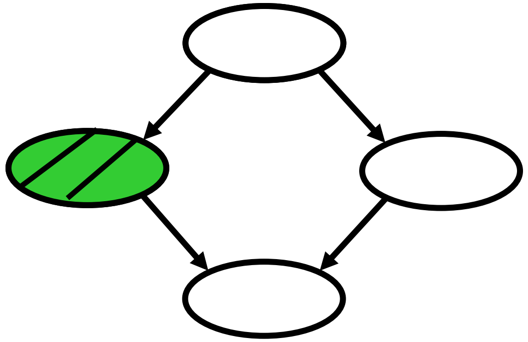
- Input: evidence instantiation
- w = 1.0
- for i = 1, 2, …, n in topological order
  - if X_i is an evidence variable
    - X_i = observation x_i for X_i
    - Set w = w * $P(x_i \mid \text{Parents}(X_i))$
  - else
    - Sample x_i from $P(X_i \mid \text{Parents}(X_i))$
- return $(x_1, x_2, \ldots, x_n)$, w



Likelihood Weighting

- Sampling distribution if z sampled and e fixed evidence

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



- Now, samples have weights

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

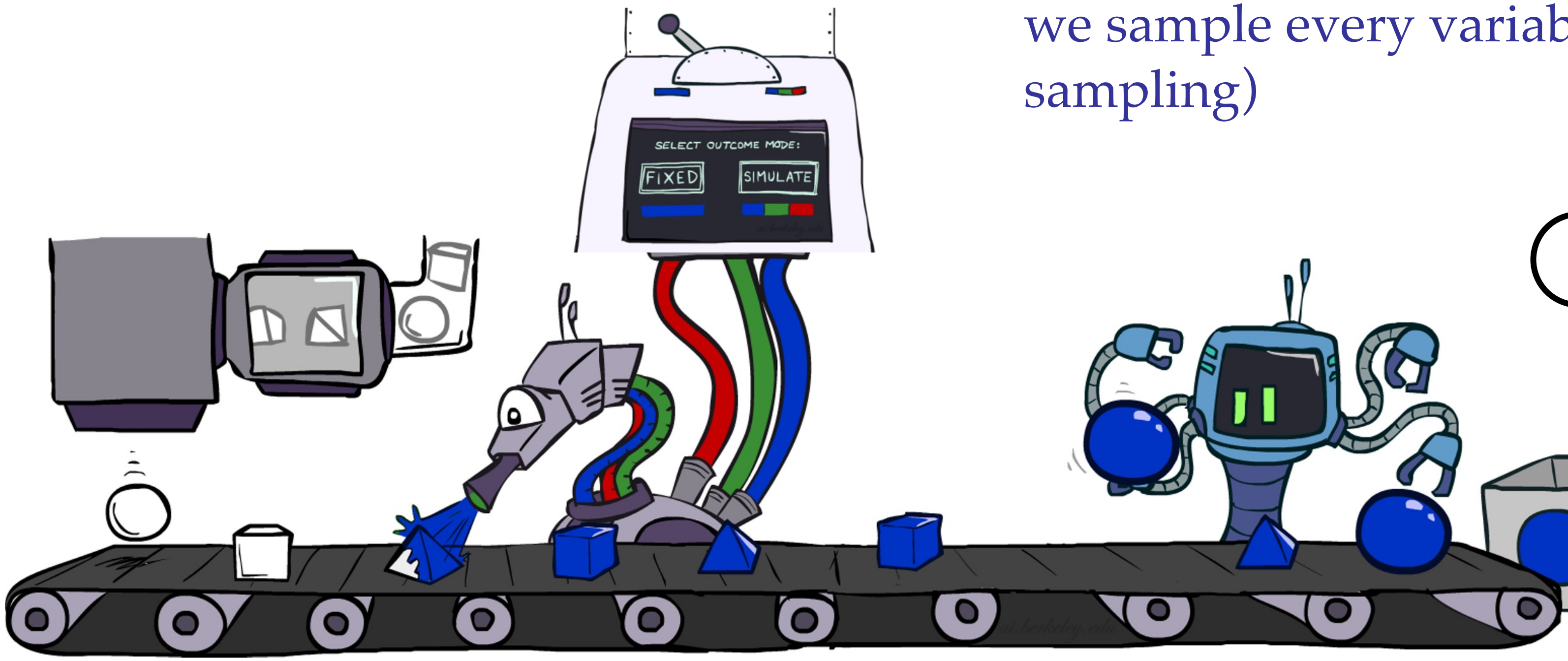
- Together, weighted sampling distribution is consistent

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

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

Likelihood Weighting

- Likelihood weighting is good
  - All samples are used
  - More of our samples will reflect the state of the world suggested by the evidence
  - Values of downstream variables are influenced by upstream evidence



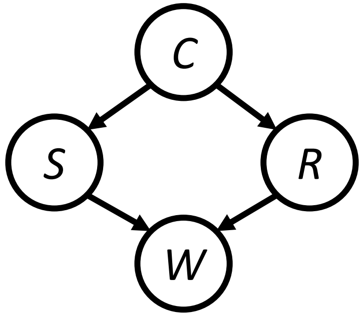
Likelihood weighting doesn't solve all our problems

The values of upstream variables are unaffected by downstream evidence

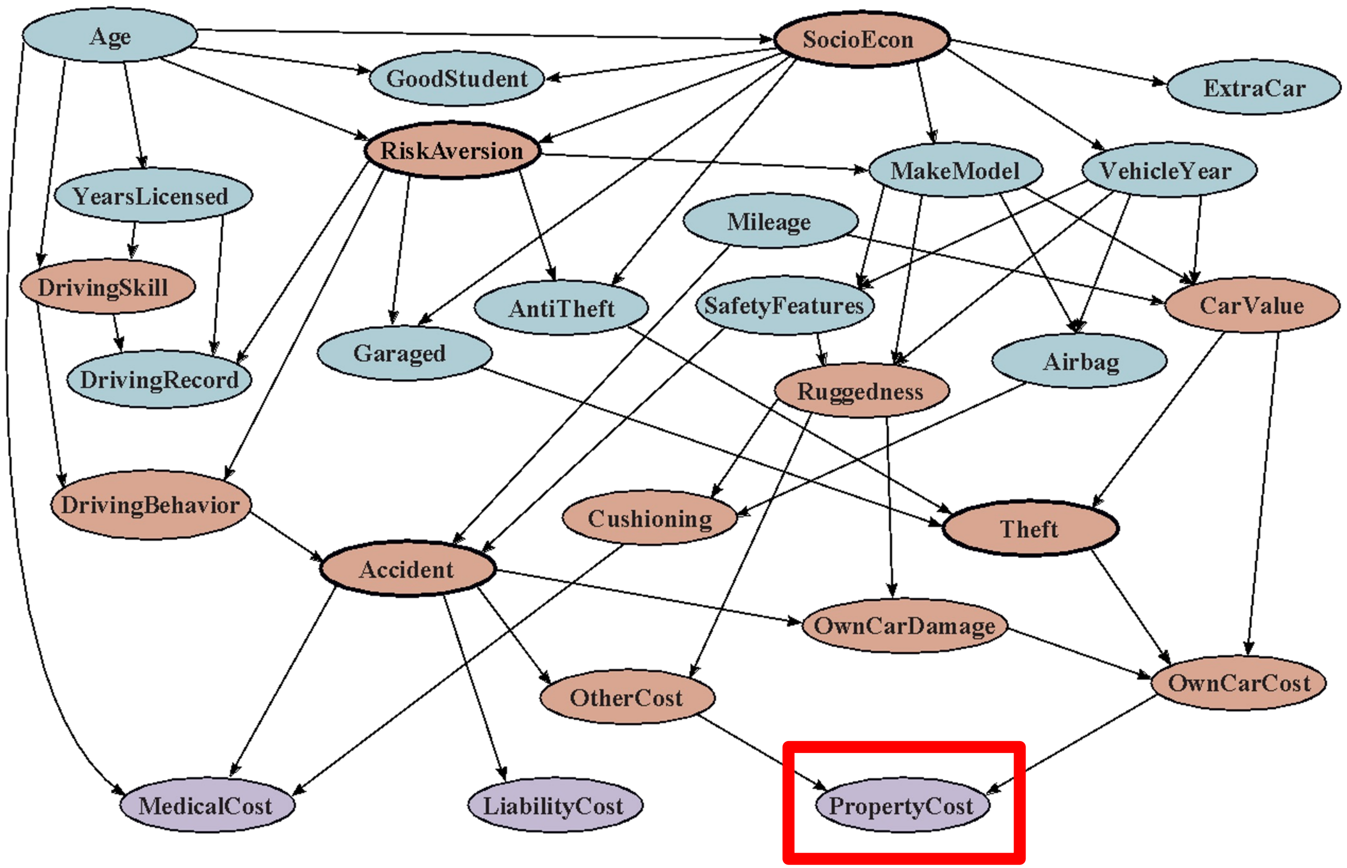
With evidence in k leaf nodes, weights will be $O(2^{-k})$

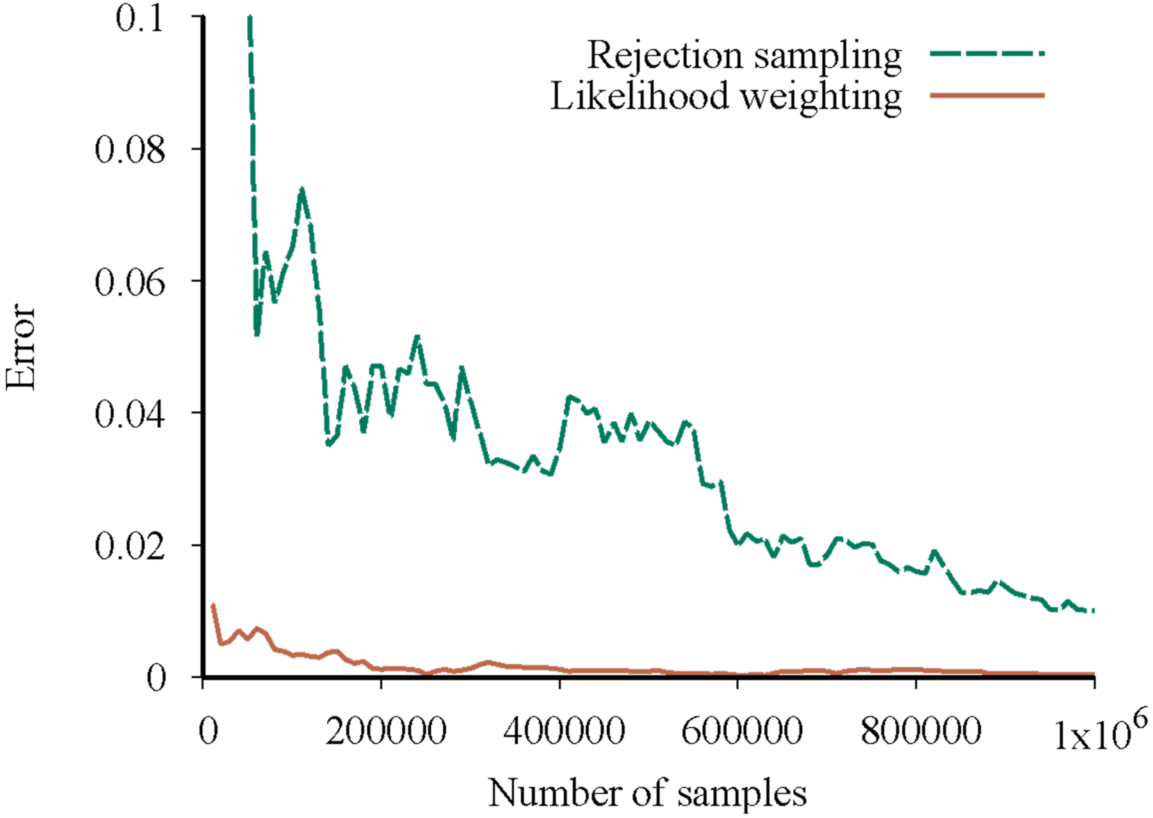
With high probability, one lucky sample will have much larger weight than the others, dominating the result

We would like to consider evidence when we sample every variable (leads to Gibbs sampling)

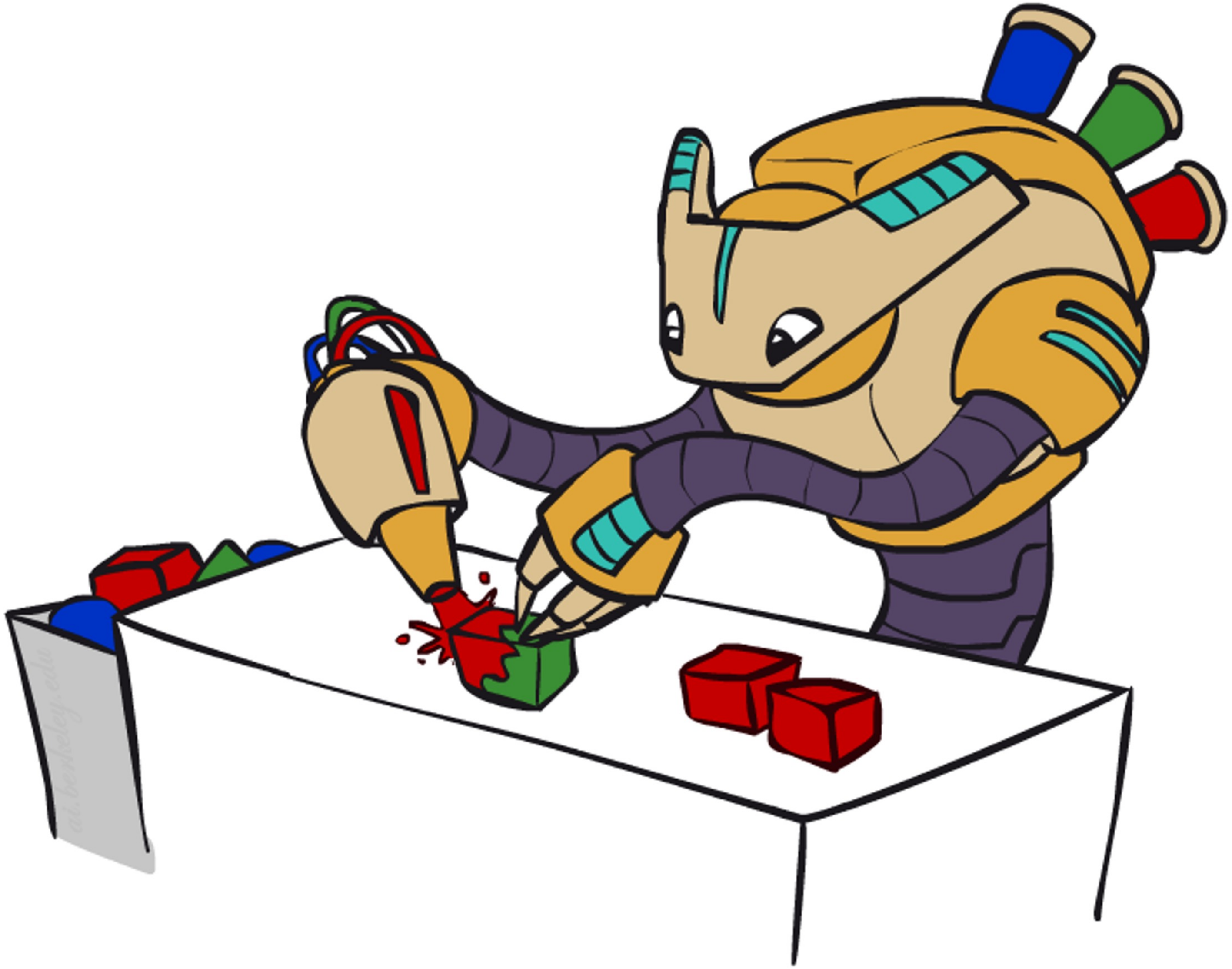


Example: Car Insurance: $P(PropertyCost \mid \boldsymbol{e})$





Gibbs Sampling



Markov Chain Monte Carlo

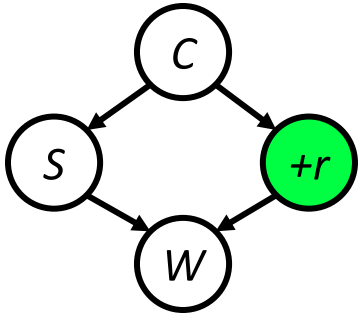
- Gibbs sampling is a MCMC technique (Metropolis-Hastings)
- MCMC (Markov chain Monte Carlo) is a family of randomized algorithms for approximating some quantity of interest over a very large state space
  - Markov chain = a sequence of randomly chosen states ("random walk"), where each state is chosen conditioned on the previous state
  - Monte Carlo = a very expensive city in Monaco with a famous casino
  - Monte Carlo = an algorithm (usually based on sampling) that has some probability of producing an incorrect answer
- MCMC = wander around for a bit, average what you see

Gibbs sampling

- A particular kind of MCMC
  - States are complete assignments to all variables
    - (local search: closely related to simulated annealing!)
  - Evidence variables remain fixed, other variables change
  - To generate the next state, pick a variable and sample a value for it conditioned on all the other variables: $X_i' \sim P(X_i \mid x_1,..,x_{i-1},x_{i+1},..,x_n)$
    - Will tend to move towards states of higher probability, but can go down too
    - In a Bayes net, $P(X_i \mid x_1,..,x_{i-1},x_{i+1},..,x_n) = P(X_i \mid markovblanket(X_i))$
- Theorem: Gibbs sampling is consistent*
- Provided all Gibbs distributions are bounded away from 0 and 1 and variable selection is fair

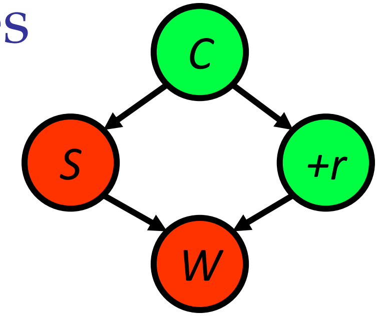
Gibbs Sampling Example: P(S | +r)

- Step 1: Fix evidence
  - R = +r



Step 2: Initialize other variables

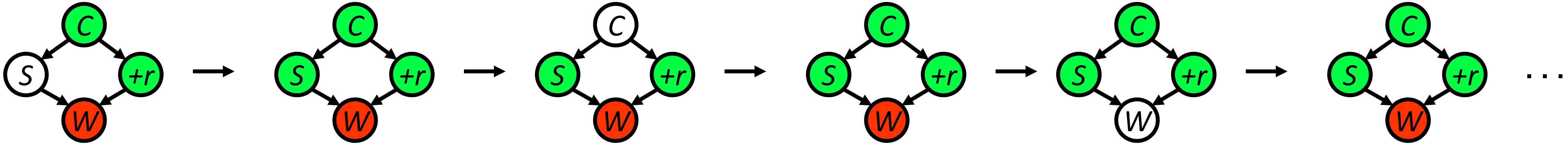
Randomly



Steps 3: Repeat:

Choose a non-evidence variable X

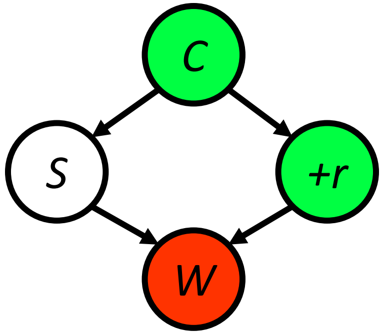
Resample X from P(X | MarkovBlanket(X))



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

Resampling of One Variable

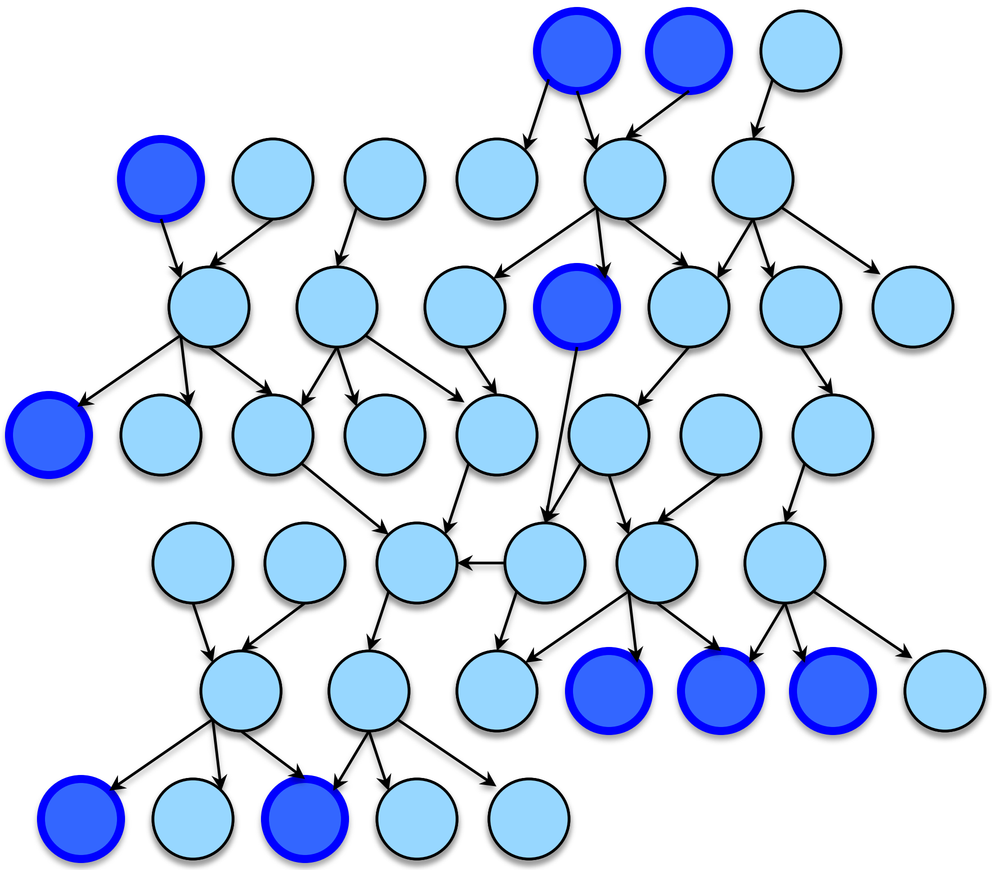
- Sample from P(S | +c, +r, -w)



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

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

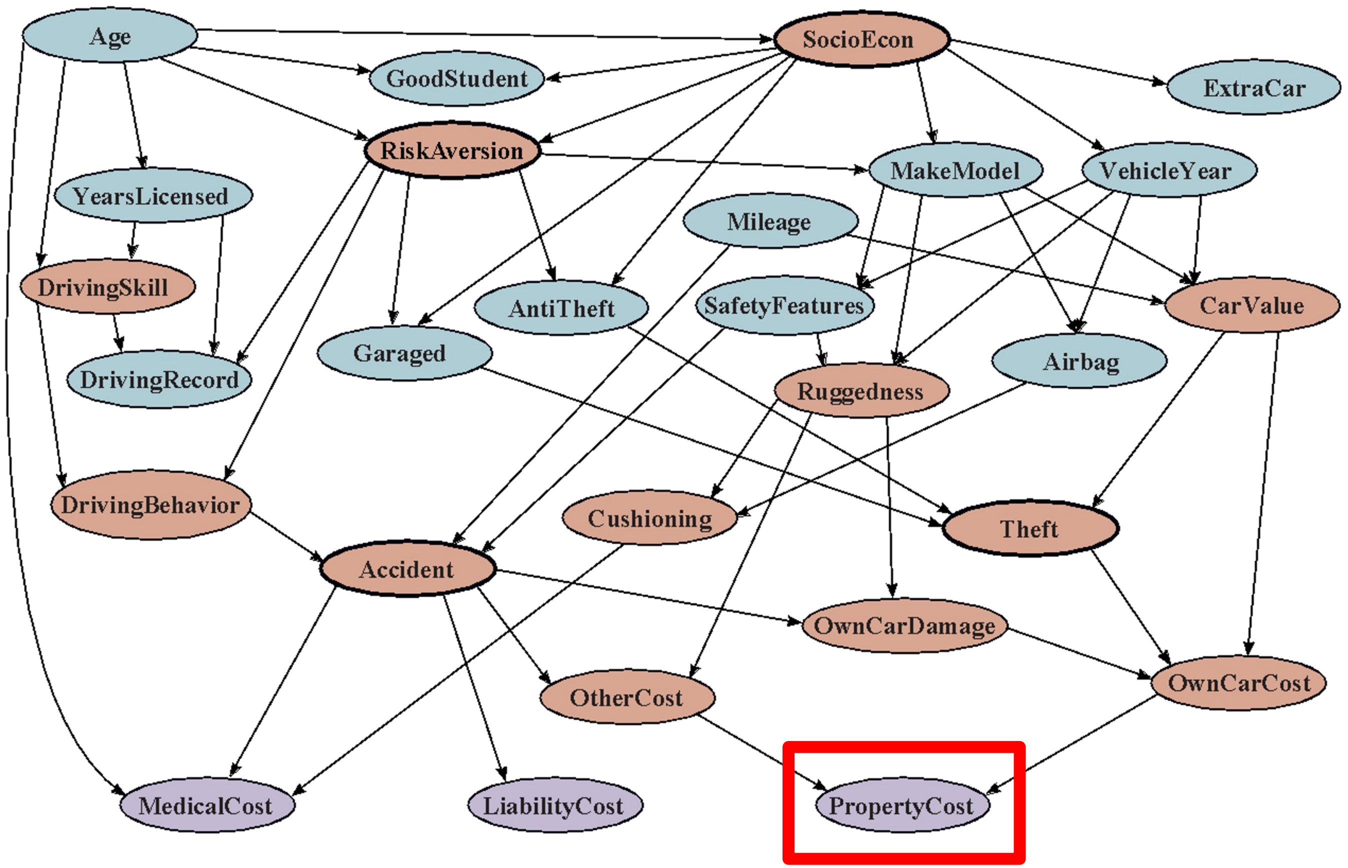
Why would anyone do this?

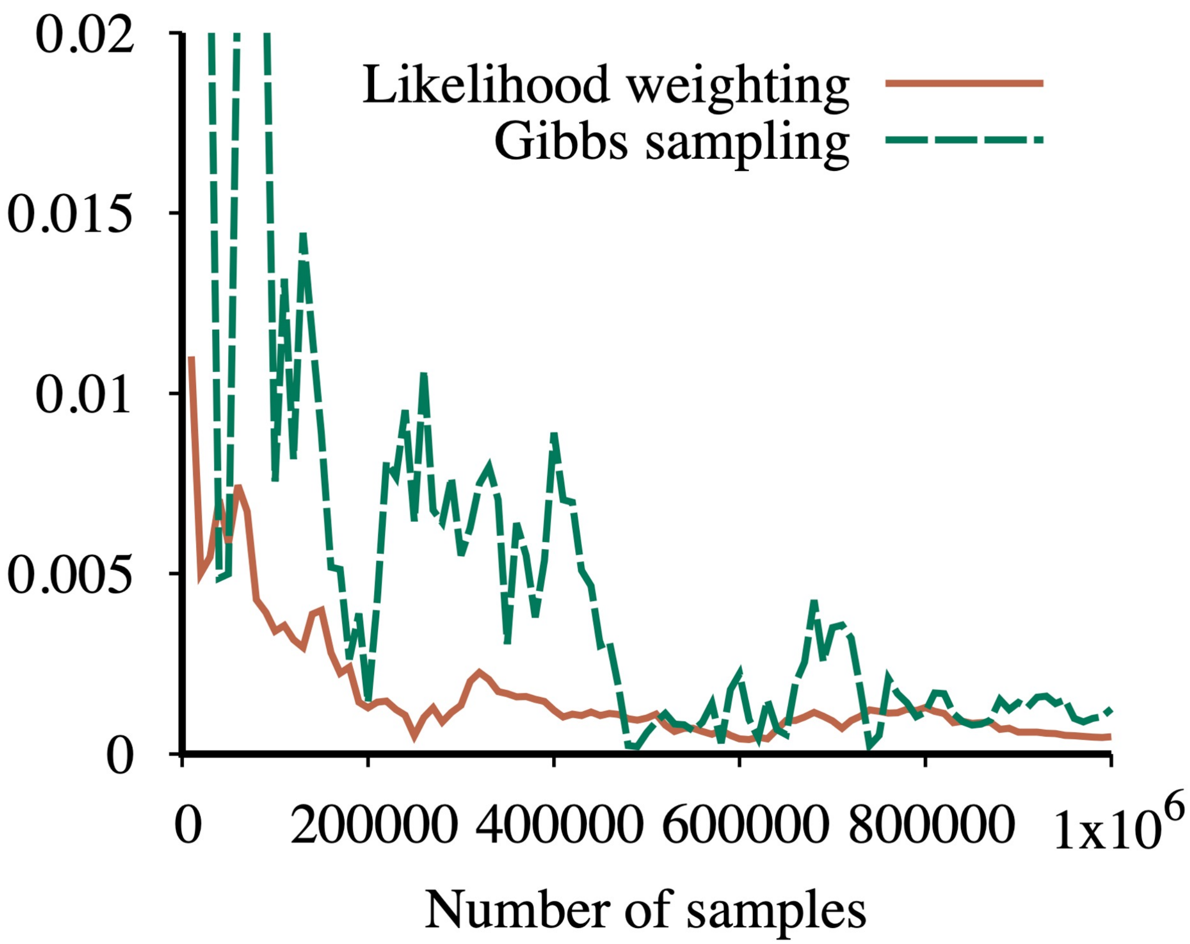


Samples soon begin to reflect all the evidence in the network

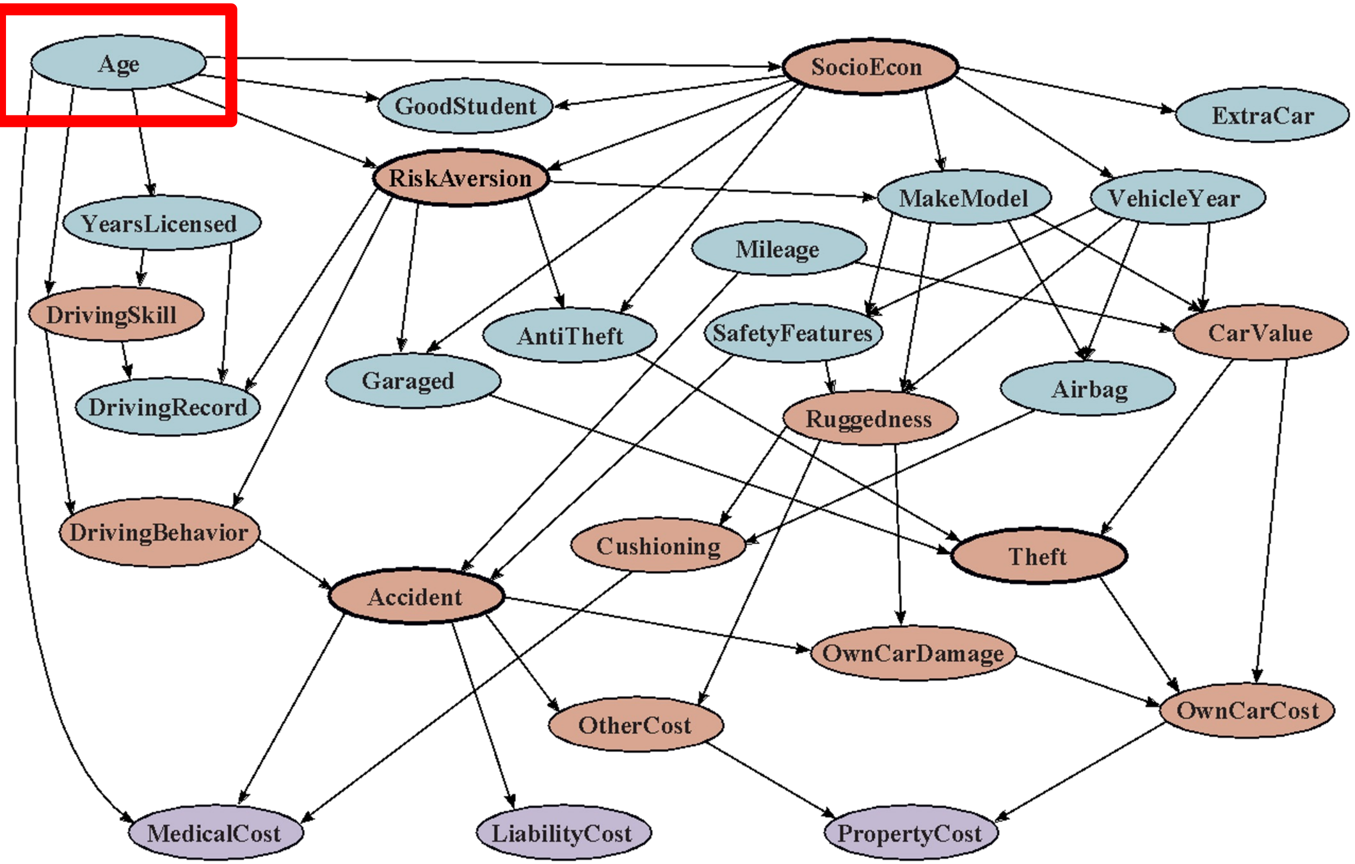
Eventually they are being drawn from the true posterior!

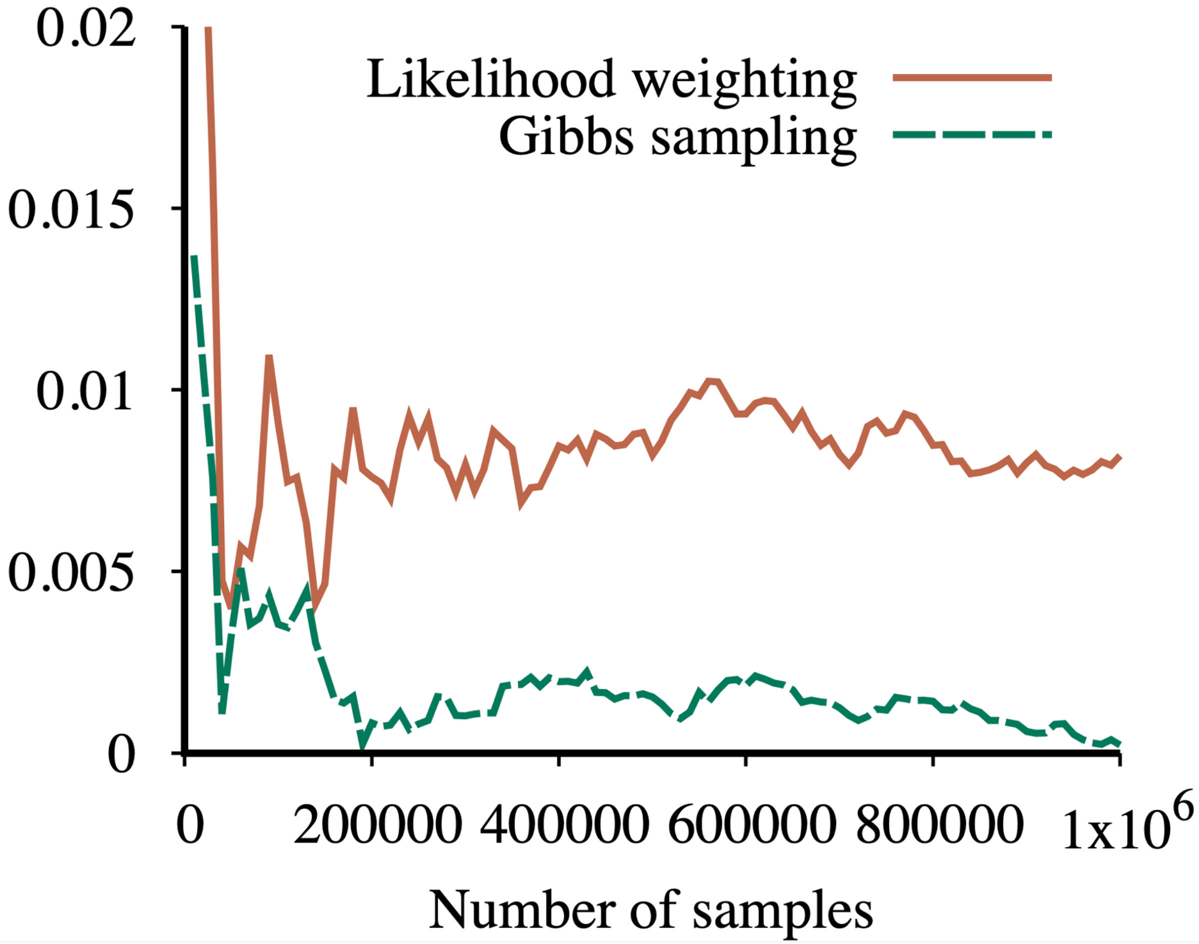
Car Insurance: $P(PropertyCost \mid \boldsymbol{e})$





Car Insurance: $P(Age \mid costs)$



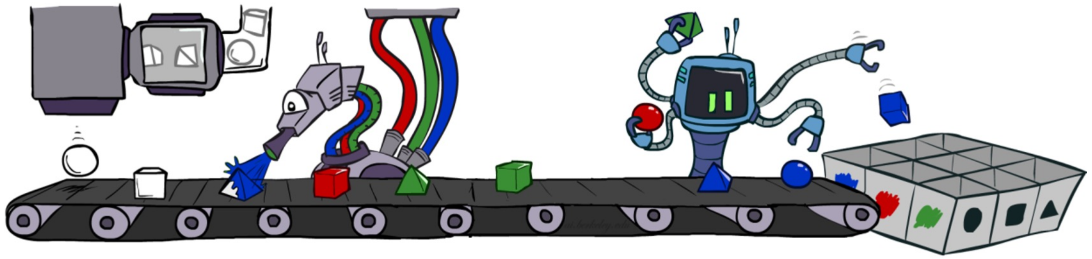


Why does it work? (see AIMA 13.4.2 for details)

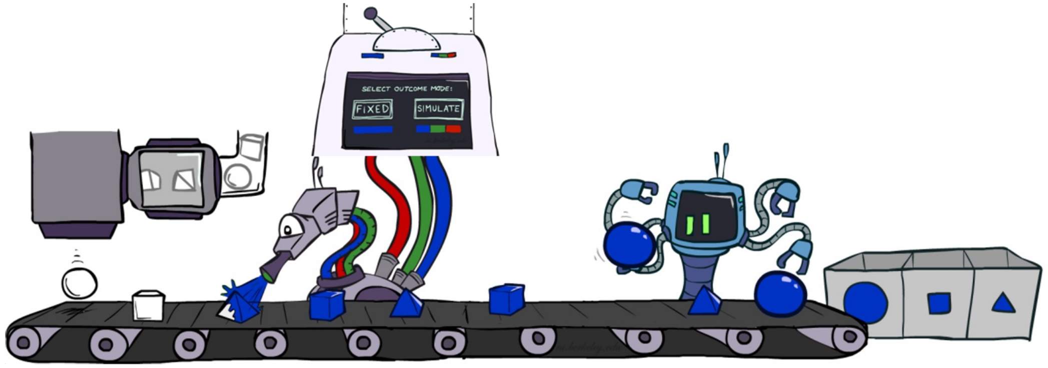
- Suppose we run it for a long time and predict the probability of reaching any given state at time t: $\pi_t(x_1,\ldots,x_n)$ or $\pi_t(\underline{\mathbf{x}})$
- Each Gibbs sampling step (pick a variable, resample its value) applied to a state $\underline{\mathbf{x}}$ has a probability $k(\underline{\mathbf{x}'} \mid \underline{\mathbf{x}})$ of reaching a next state $\underline{\mathbf{x}'}$
- So $\pi_{t+1}(\underline{\mathbf{x}'}) = \sum_{\underline{\mathbf{x}}} k(\underline{\mathbf{x}'} \mid \underline{\mathbf{x}})\, \pi_t(\underline{\mathbf{x}})$ or, in matrix/vector form $\pi_{t+1} = \mathbf{K}\pi_t$
- When the process is in equilibrium $\pi_{t+1} = \pi_t = \pi$ so $\mathbf{K}\pi = \pi$
- This has a unique* solution $\pi = \mathrm{P}(x_1, \ldots, x_n \mid e_1, \ldots, e_k)$
- So for large enough t the next sample will be drawn from the true posterior
  - "Large enough" depends on CPTs in the Bayes net; takes *longer* if nearly deterministic

Bayes' Net Sampling Summary

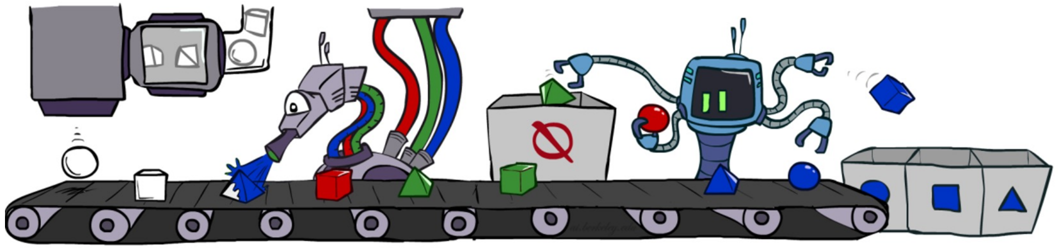
- Prior Sampling P(Q)



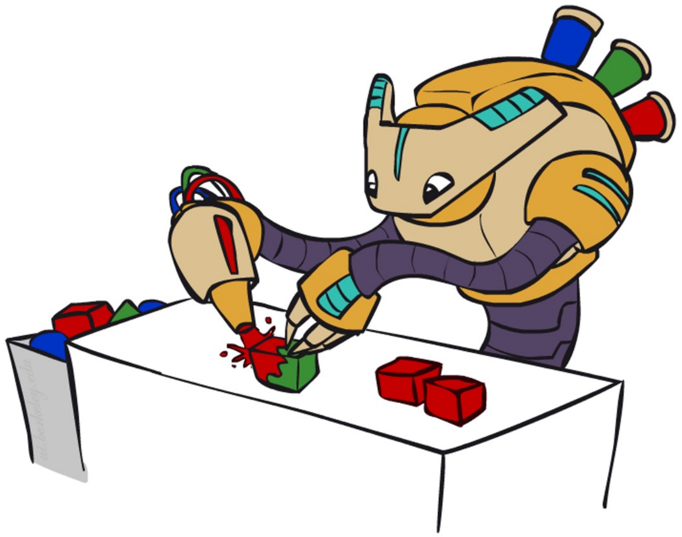
- Likelihood Weighting P(Q | e)



Rejection Sampling P(Q | e)



Gibbs Sampling P(Q | e)



CS 188: Artificial Intelligence

Hidden Markov Models

Instructor: Evgeny Pobachienko — UC Berkeley

[Slides Credit: Dan Klein, Pieter Abbeel, Anca Dragan, Stuart Russell, and many others]

Reasoning over Time or Space

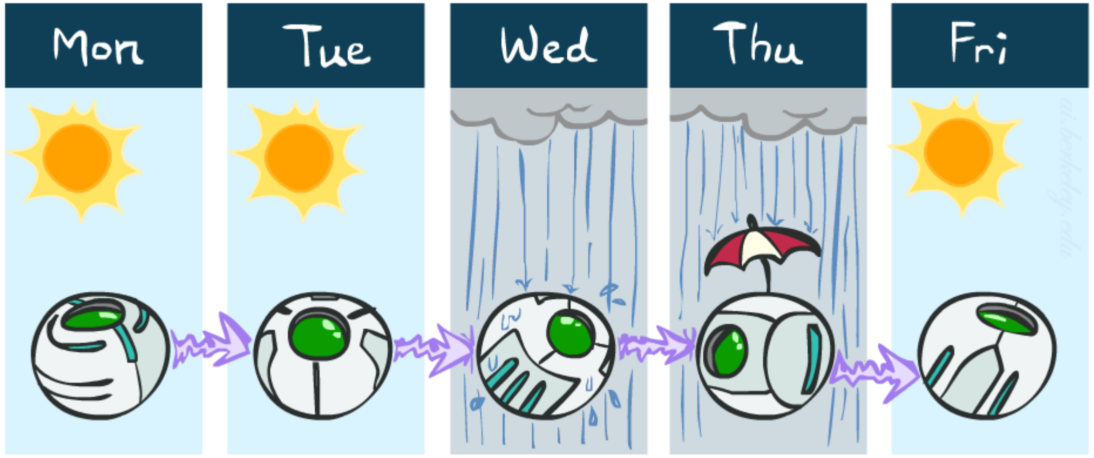
- Often, we want to reason about a sequence of observations
  - Speech recognition
  - Robot localization
  - User attention
  - Medical monitoring
- Need to introduce time (or space) into our models

Example Markov Chain: Weather

- States: X = {rain, sun}

- Initial distribution:

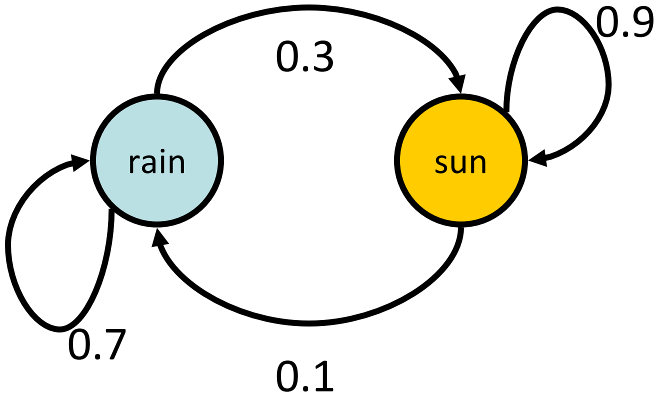
$P(X_0)$	
sun	rain
1	0.0

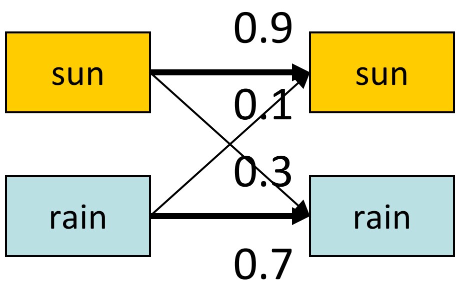


- CPT $P(X_t \mid X_{t-1})$:

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

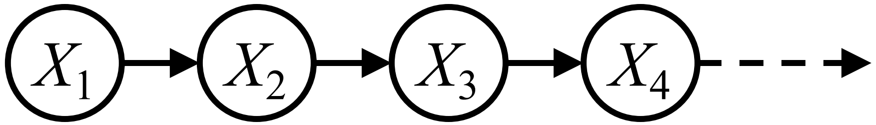
Two new ways of representing the same CPT





Markov Chains

- Value of X at a given time is called the state

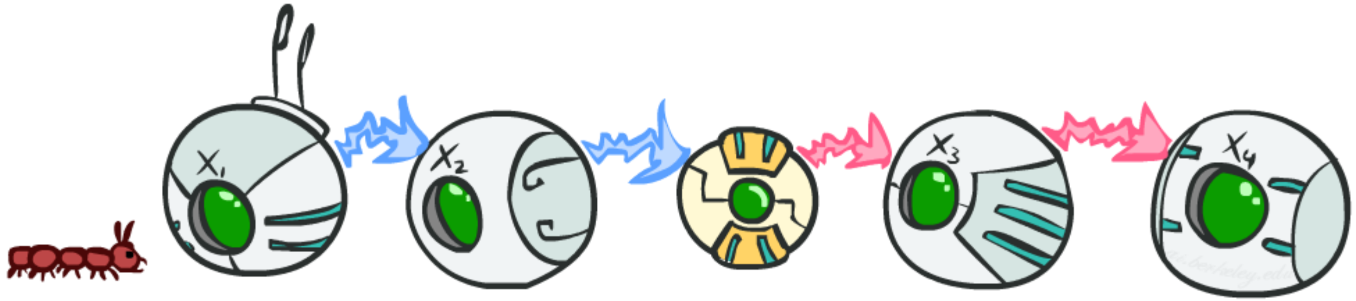


$$P(X_t) = ?$$

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

- Transition probabilities (dynamics): $P(X_t \mid X_{t-1})$ specify how the state evolves over time

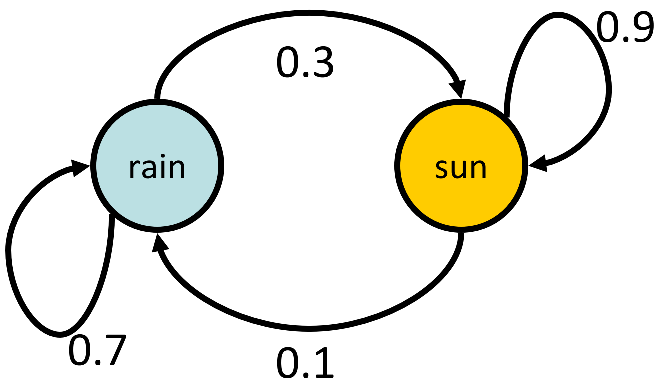
Markovian Assumption



- Basic conditional independence:
  - Given the present, the future is independent of the past!
  - Each time step only depends on the previous
  - This is called the (first order) Markov property

Example Markov Chain: Weather

- Initial distribution: 1.0 sun



- What is the probability distribution after one step?

$$P(X_2 = sun) = \sum_{x_1} P(x_1, X_2 = sun) = \sum_{x_1} P(X_2 = sun|x_1)P(x_1)$$

$$P(X_2 = \text{sun}) = \quad P(X_2 = \text{sun}|X_1 = \text{sun})P(X_1 = \text{sun}) + P(X_2 = \text{sun}|X_1 = \text{rain})P(X_1 = \text{rain})$$

$$0.9 \cdot 1.0 + 0.3 \cdot 0.0 = 0.9$$