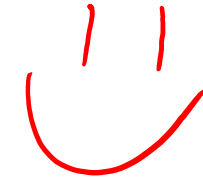


# Warm-up as you walk in



Given these  $N=10$  observations of the world:

What is the approximate value for  $P(-c \mid -a, +b)$ ?

- A.  $1/10$
- B.  $5/10$
- C.  $1/4$
- D.  $1/5$
- E. I'm not sure

$-a, -b, +c$   
 $+a, -b, +c$   
 $-a, -b, +c$   
 $-a, +b, +c$   
 $+a, -b, +c$   
 $-a, +b, -c$   
 $-a, +b, +c$   
 $-a, +b, +c$   
 $+a, -b, +c$   
 $-a, +b, +c$

Counts

+a	+b	+c	0
+a	+b	-c	0
+a	-b	+c	3
+a	-b	-c	0
-a	+b	+c	4
-a	+b	-c	1
-a	-b	+c	2
-a	-b	-c	0

# Announcements

## Assignments:

- HW10 and P5 out after midterm

## Midterm:

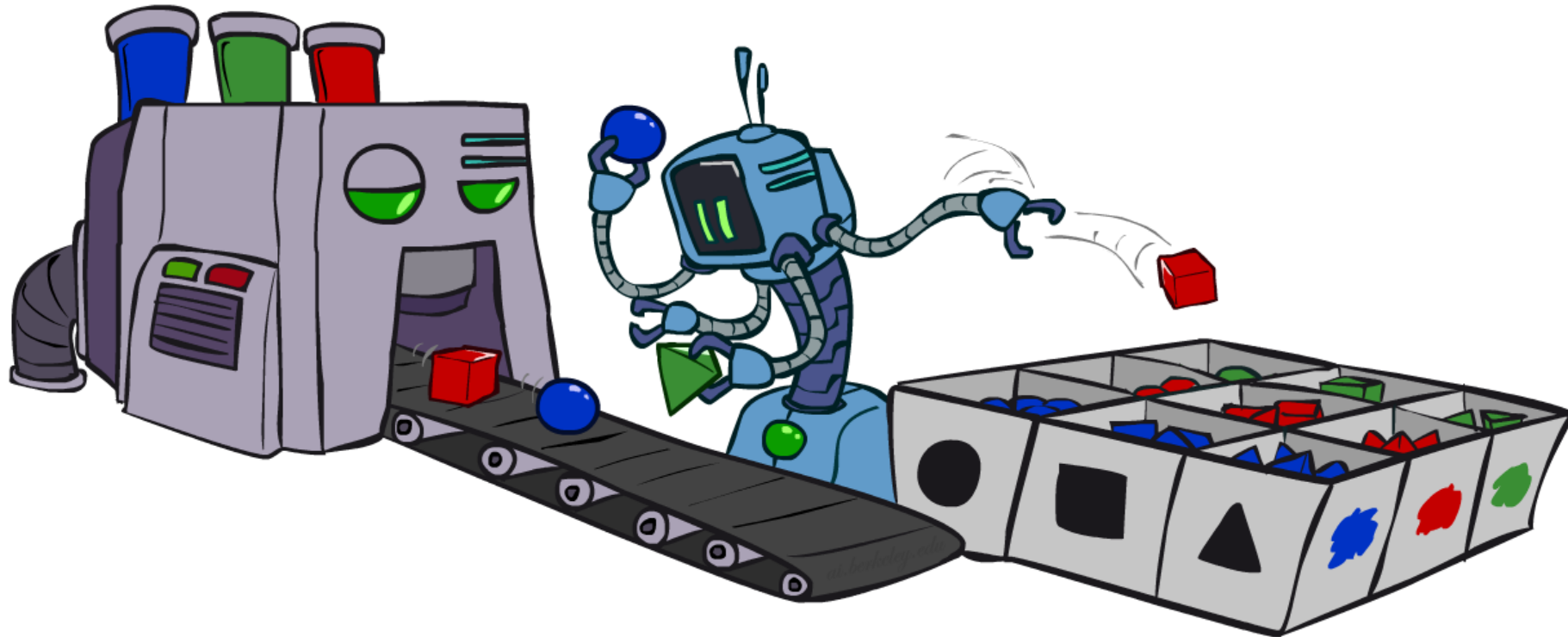
- Tue 11/12, in-class
  - See Piazza for details

## Piazza:

- Lecture thread on piazza

# AI: Representation and Problem Solving

## Bayes Nets Sampling



Instructors: Pat Virtue & Fei Fang

Slide credits: CMU AI and <http://ai.berkeley.edu>

# Review: Bayes Nets

Joint distributions  $\rightarrow$  answer any query

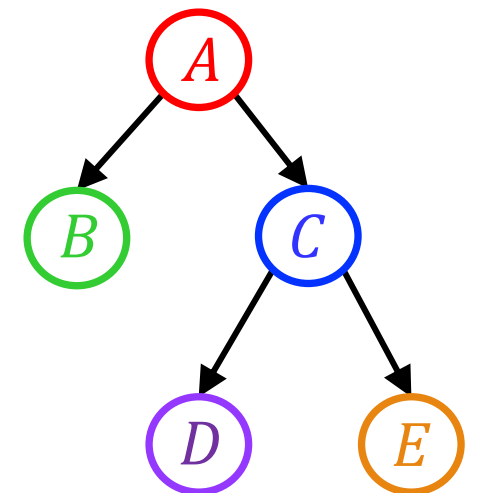
$$P(a | e) = \frac{1}{Z} P(a, e) = \frac{1}{Z} \sum_b \sum_c \sum_d P(a, b, c, d, e)$$

Break down joint using chain rule

$$P(A, B, C, D, E) = P(A) P(B|A) P(C|A, B) P(D|A, B, C) P(E|A, B, C, D)$$

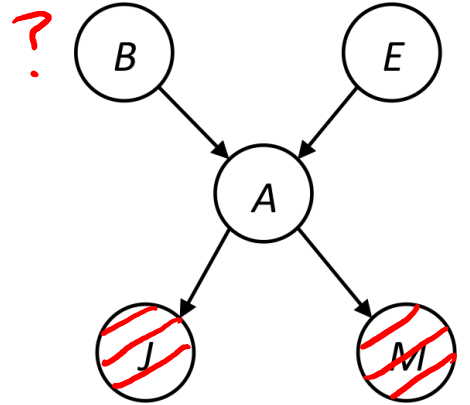
With Bayes nets

$$P(A, B, C, D, E) = P(A) P(B|A) P(C|A) P(D|C) P(E|C)$$



# Example

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



$P(B)$	<u><math>P(E)</math></u>	<u><math>P(A \mid B, E)</math></u>	$P(j \mid A)$	$P(m \mid A)$
--------	--------------------------	------------------------------------	---------------	---------------

Choose A

$$\begin{array}{l} P(A \mid B, E) \\ P(j \mid A) \\ P(m \mid A) \end{array} \xrightarrow{\times} \underline{P(A, j, m \mid B, E)} \xrightarrow{\Sigma} \underline{P(j, m \mid B, E)}$$

$P(B)$	$P(E)$	$P(j, m \mid B, E)$
--------	--------	---------------------

# Example

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

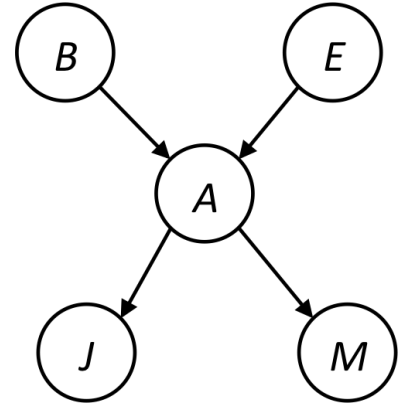
Choose E

$$P(E) \quad P(j,m|B,E) \xrightarrow{\times} P(E,j,m|B) \xrightarrow{\Sigma} P(j,m|B)$$

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

Finish with B

$$P(B) \quad P(j,m|B) \xrightarrow{\times} P(j,m,B) \xrightarrow{\text{Normalize}} P(B|j,m)$$



# Variable Elimination order matters

$P(D)$

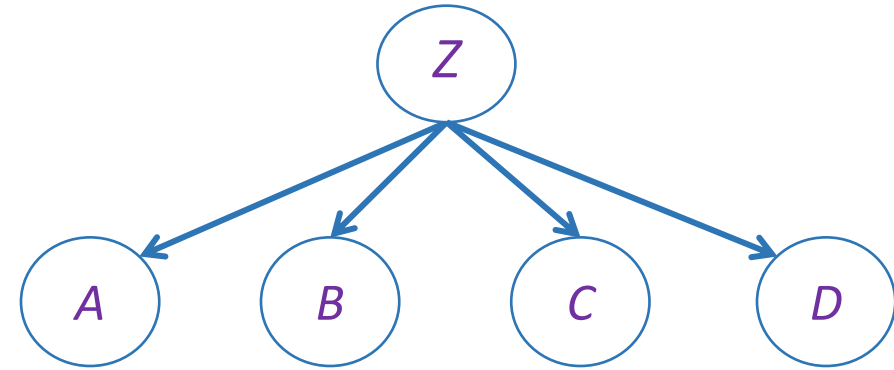
- Elimination order: C, B, A, Z

- $P(D) = \alpha \sum_{z,a,b,c} P(D|z) P(z) P(a|z) P(b|z) P(c|z)$
- $= \alpha \sum_z P(D|z) P(z) \sum_a P(a|z) \sum_b P(b|z) \sum_c P(c|z)$
- Largest factor has 2 variables (D,Z)

- Elimination order: Z, C, B, A

- $P(D) = \alpha \sum_{a,b,c,z} P(a|z) P(b|z) P(c|z) P(D|z) P(z)$
- $= \alpha \sum_a \sum_b \sum_c \sum_z P(a|z) P(b|z) P(c|z) P(D|z) P(z)$
- Largest factor has 4 variables (A,B,C,D)

- In general, with  $n$  leaves, factor of size  $2^n$



# VE: Computational and Space Complexity

The computational and space complexity of variable elimination is determined by the largest factor (and it's space that kills you)

The elimination ordering can greatly affect the size of the largest factor.

- E.g., previous slide's example  $2^n$  vs. 2

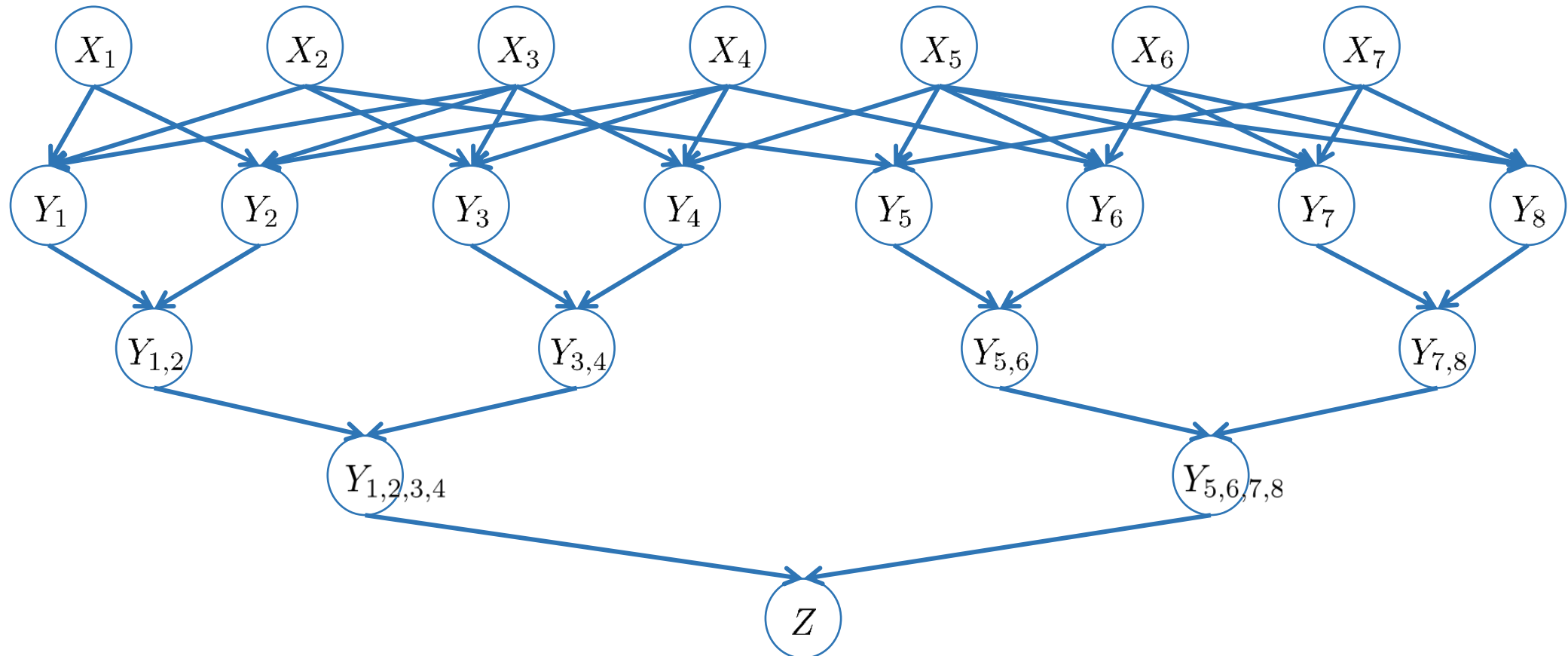
Does there always exist an ordering that only results in small factors?

- **No!**

# VE: Computational and Space Complexity

Inference in Bayes' nets is NP-hard.

No known efficient probabilistic inference in general.



# Bayes Nets

✓ Part I: Representation

✓ Part II: Exact inference

✓ ■ Enumeration (always exponential complexity)

✓ ■ Variable elimination (worst-case exponential complexity, often better)

✓ ■ Inference is NP-hard in general

Part III: Approximate Inference

# Warm-up as you walk in

Given these  $N=10$  observations of the world:

What is the approximate value for  $P(-c | \underline{-a, +b})$ ?

- A.  $1/10$
- B.  $5/10$
- C.  $1/4$
- D.  $1/5$
- E. I'm not sure

$-a, -b, +c$   
 $+a, -b, +c$   
 $-a, -b, +c$   
 $-a, +b, +c$   
 $+a, -b, +c$   
 $-a, +b, -c$   
 $-a, +b, +c$   
 $-a, +b, +c$   
 $+a, -b, +c$   
 $-a, +b, +c$

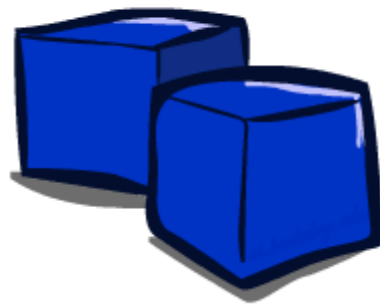
Counts

+a	+b	+c	0
+a	+b	-c	0
+a	-b	+c	3
+a	-b	-c	0
-a	+b	+c	4
-a	+b	-c	1
-a	-b	+c	2
-a	-b	-c	0

# Approximate Inference: Sampling



5

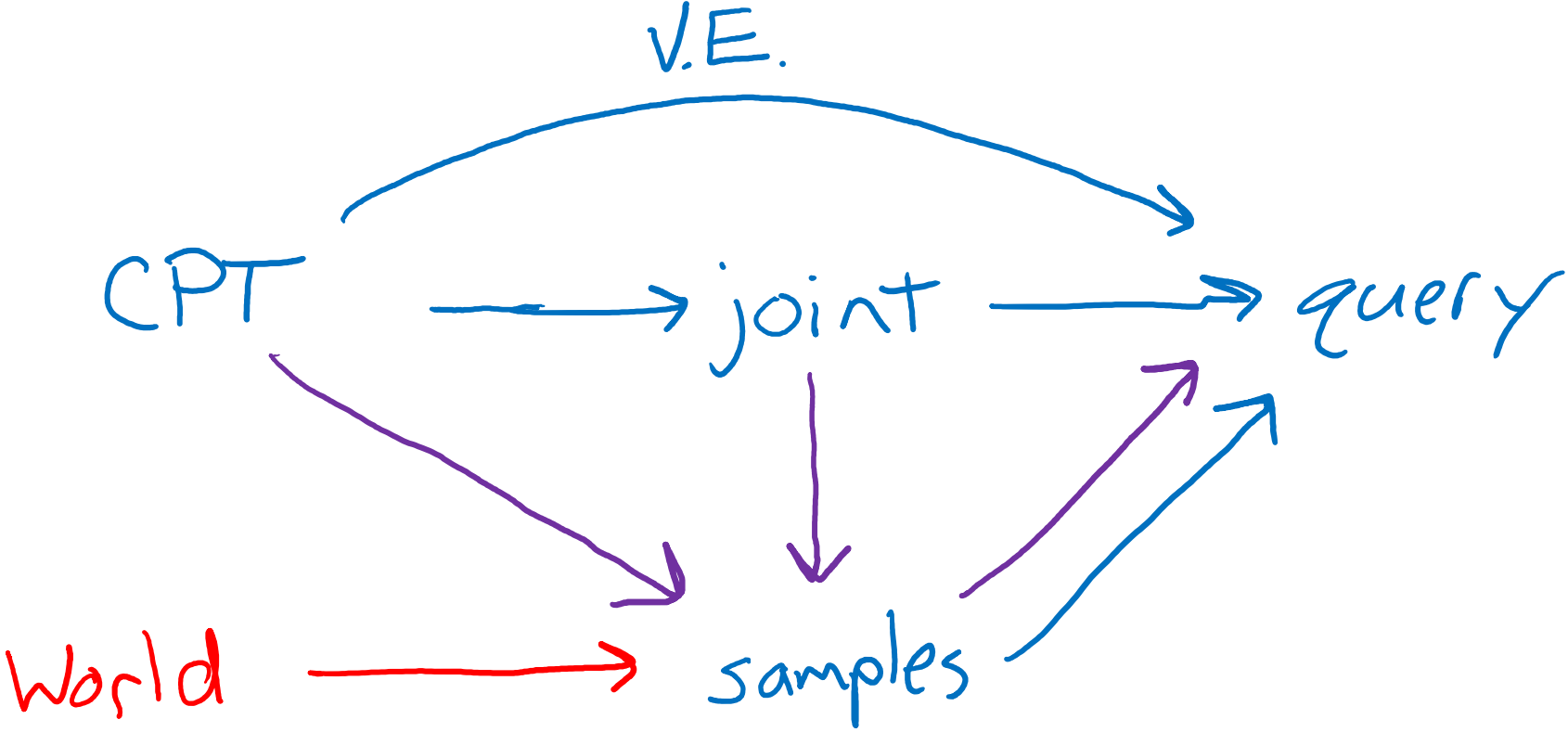


2

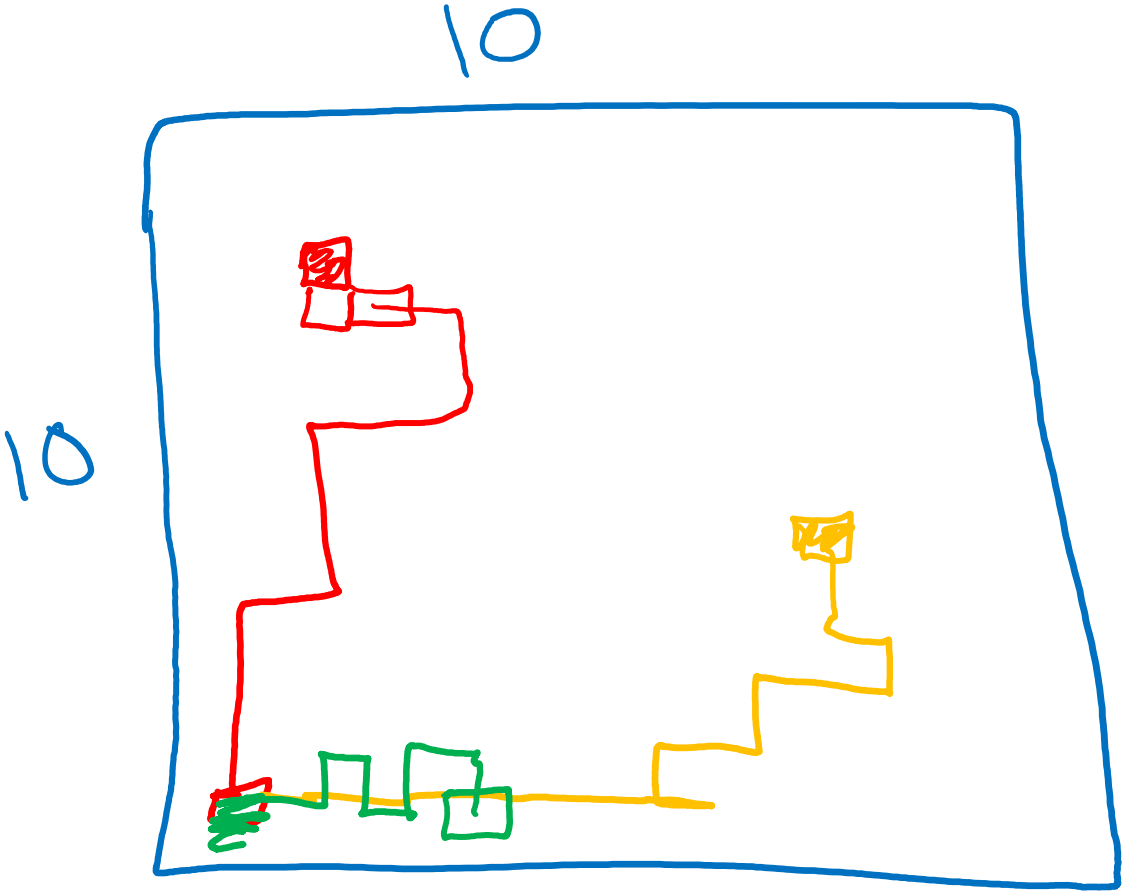


1

# Inference vs Sampling



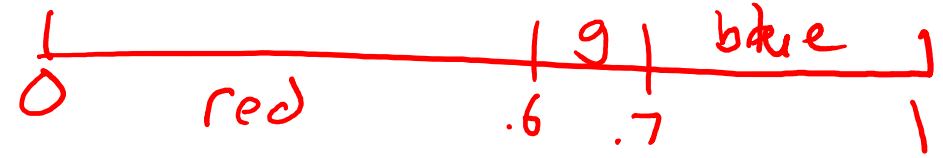
# Motivation for Approximate Inference



$$P(G_{t=20})$$



# 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 outcome associated with a sub-interval of  $[0,1)$  with sub-interval size equal to probability of the outcome



## Example

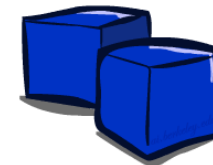
C	P(C)
red	0.6
green	0.1
blue	0.3

$$0 \leq u < 0.6, \rightarrow C = \text{red}$$

$$0.6 \leq u < 0.7, \rightarrow C = \text{green}$$

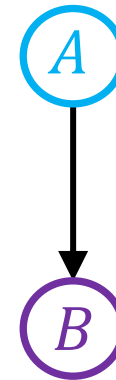
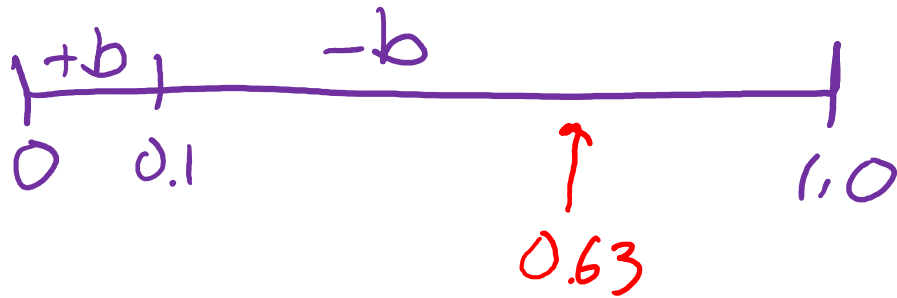
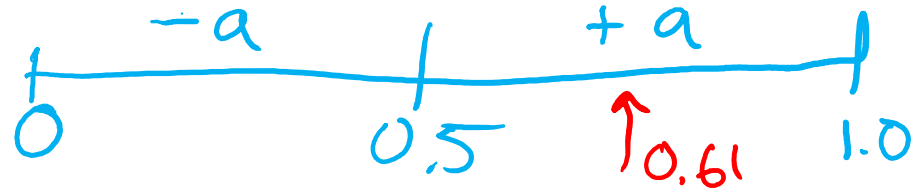
$$0.7 \leq u < 1, \rightarrow C = \text{blue}$$

- If `random()` returns  $u = 0.83$ , then our sample is  $C = \text{blue}$
- E.g, after sampling 8 times:



# Sampling

How would you sample from a conditional distribution?



$P(A)$

$+a$	$1/2$
$-a$	$1/2$

$P(B|A)$

$+a$	$+b$	$1/10$
	$-b$	$9/10$
$-a$	$+b$	$1/2$
	$-b$	$1/2$

# Sampling in Bayes' Nets

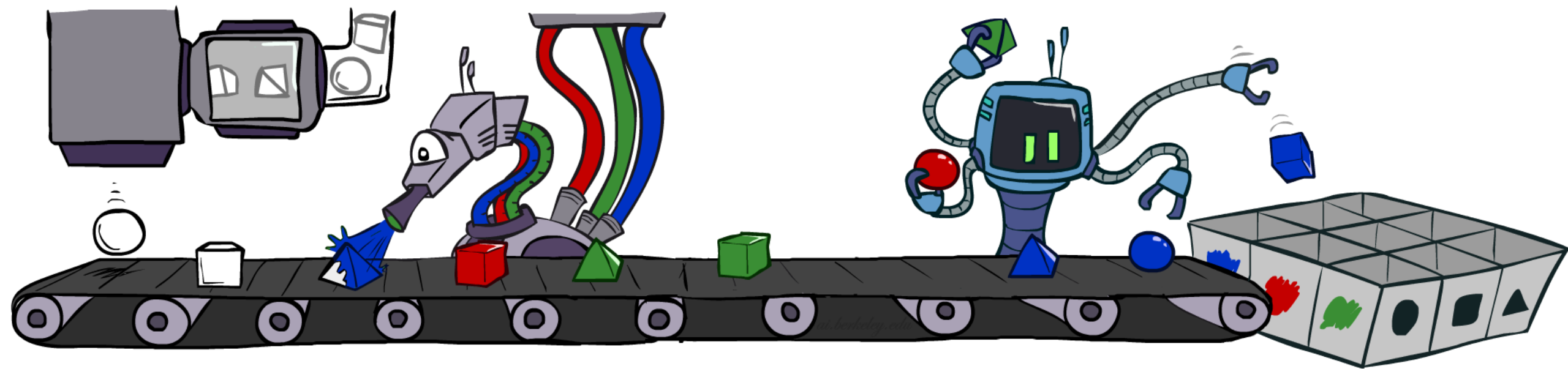
Prior Sampling

Rejection Sampling

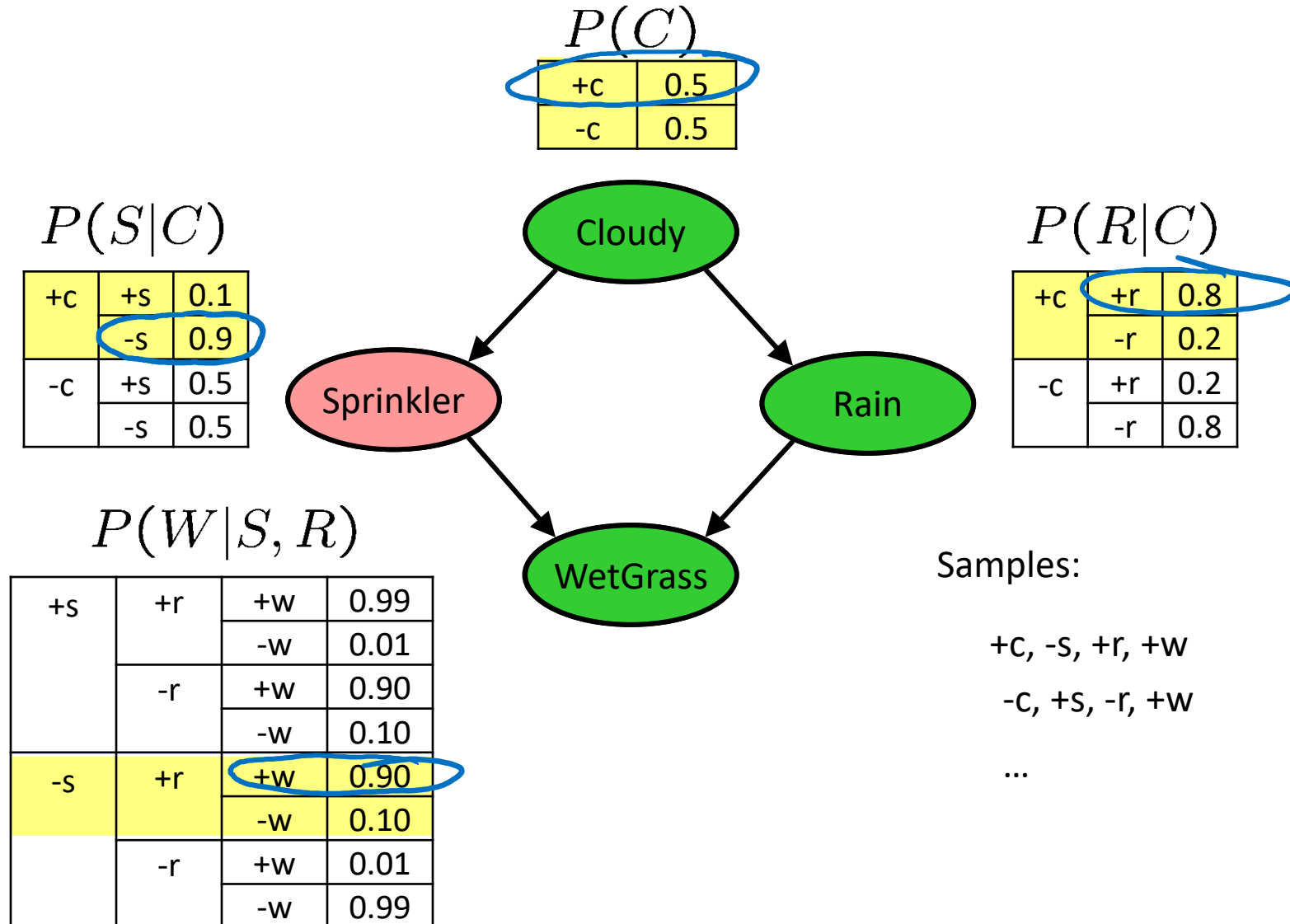
Likelihood Weighting

Gibbs Sampling

# Prior Sampling



# Prior Sampling

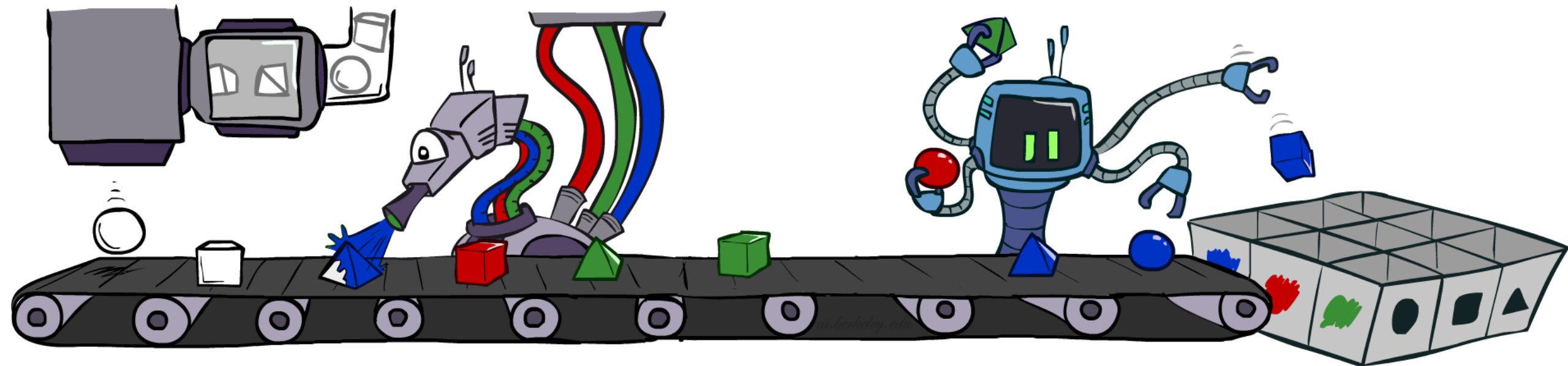


# Prior Sampling

For  $i=1, 2, \dots, n$

- Sample  $x_i$  from  $P(X_i \mid \text{Parents}(X_i))$

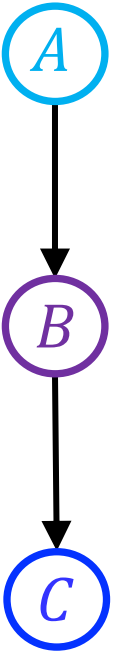
Return  $(x_1, x_2, \dots, x_n)$



# Piazza Poll 1

$$\frac{N(+a, -b, +c)}{N}$$

Prior Sampling: What does the value approximate?



- A.  $P(+a, -b, +c)$
- B.  $P(+c \mid +a, -b)$
- C.  $P(+c \mid -b,)$
- D.  $P(+c)$
- E. I don't know



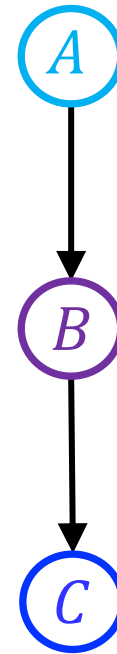
$$P(A)P(B|A)P(C|B)$$
$$+a \quad -b|+a \quad +c|-b$$
$$\frac{500}{1000} \times \frac{252}{500} \times \frac{248}{504} \approx \frac{124}{1000}$$

# Piazza Poll 2

How many  $\{-a, +b, -c\}$  samples out of  $N=1000$  should we expect?

- A. 1
- B. 50
- C. 125
- D. 200
- E. I have no idea

$$1000 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{5}$$



$P(A)$

+a	1/2
-a	1/2

$P(B|A)$

+a	+b	1/10
	-b	9/10
-a	+b	1/2
	-b	1/2

$P(C|B)$

+b	+c	4/5
	-c	1/5
-b	+c	1
	-c	0

# Probability of a sample

Given this Bayes Net & CPT,  
what is  $P(+a, +b, +c)$ ?

Algorithm: Multiply likelihood of  
each node given parents:

- $w = 1.0$
- for  $i=1, 2, \dots, n$ 
  - Set  $w = w * P(x_i | \text{Parents}(X_i))$
- return  $w$



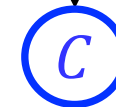
$P(A)$

+a	1/2
-a	1/2



$P(B|A)$

+a	+b	1/10
	-b	9/10
-a	+b	1/2
	-b	1/2



$P(C|B)$

+b	+c	4/5
	-c	1/5
-b	+c	1
	-c	0

# Prior Sampling

This process generates samples with probability:

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

...i.e. the BN's joint probability

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

$$\begin{aligned} \text{Then } \lim_{N \rightarrow \infty} \hat{P}(x_1, \dots, x_n) &= \lim_{N \rightarrow \infty} N_{PS}(x_1, \dots, x_n) / N \\ &= S_{PS}(x_1, \dots, x_n) \\ &= P(x_1 \dots x_n) \end{aligned}$$

# Prior Sampling



$P(A)$

+a	1/2
-a	1/2

$P(B|A)$

+a	+b	1/10
	-b	9/10
-a	+b	1/2
	-b	1/2

$P(C|B)$

+b	+c	4/5
	-c	1/5
-b	+c	1
	-c	0

# Example

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

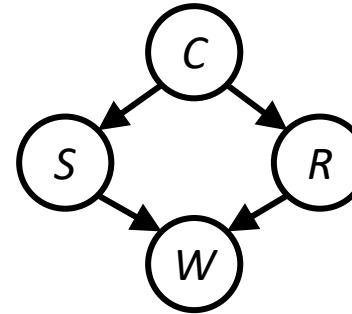
+c, -s, +r, +w

+c, +s, +r, +w

-c, +s, +r, -w

+c, -s, +r, +w

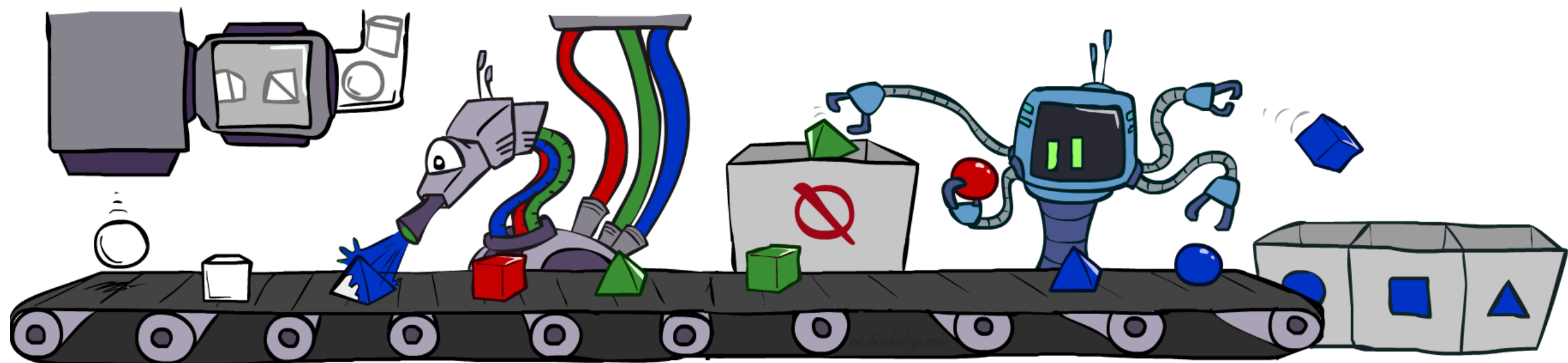
-c, -s, -r, +w



If we want to know  $P(W)$

- We have counts  $\langle +w:4, -w:1 \rangle$
- Normalize to get  $P(W) = \langle +w:0.8, -w:0.2 \rangle$
- This will get closer to the true distribution with more samples
- Can estimate anything else, too
- What about  $P(C \mid +w)$ ?  $P(C \mid +r, +w)$ ?  $P(C \mid -r, -w)$ ?
- Fast: can use fewer samples if less time (what's the drawback?)

# Rejection Sampling



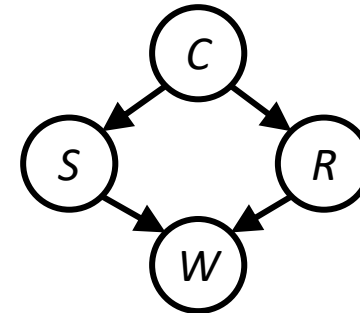
# Rejection Sampling

Let's say we want  $P(C)$

- No point keeping all samples around
- 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  $S=+s$
- 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~~

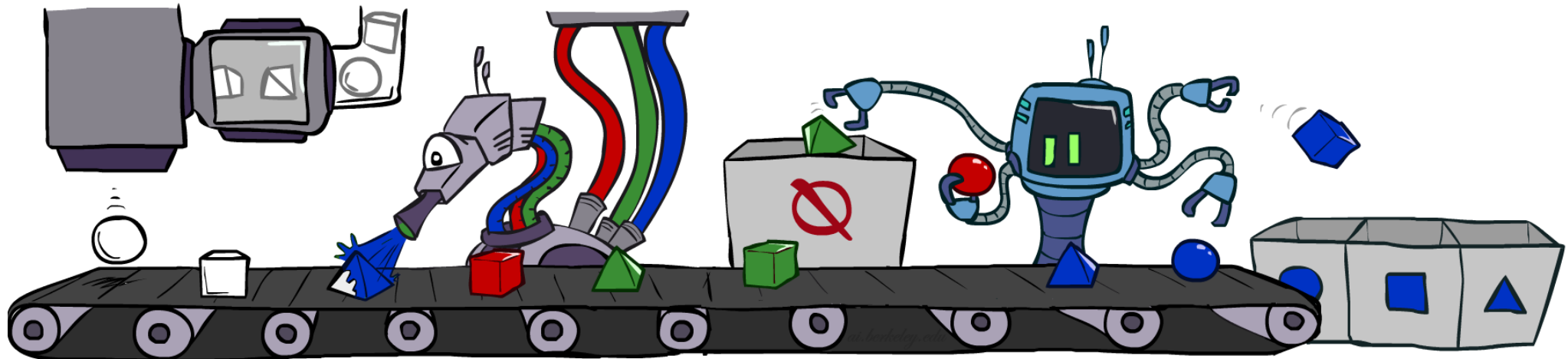
# Rejection Sampling

IN: evidence instantiation

For  $i=1, 2, \dots, n$

- Sample  $x_i$  from  $P(X_i \mid \text{Parents}(X_i))$
- If  $x_i$  not consistent with evidence
  - Reject: Return, and no sample is generated in this cycle

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



# Piazza Poll 3

What queries can we (approximately) answer with rejection sampling samples (evidence:  $+c$ )?

A.  $P(+a, -b, +c)$

B.  $P(+a, -b \mid +c)$   $P(+c)$

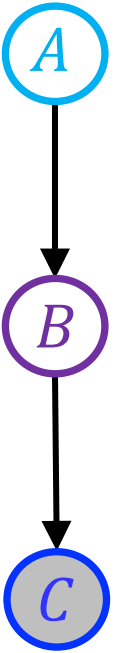
C. Both

D. Neither

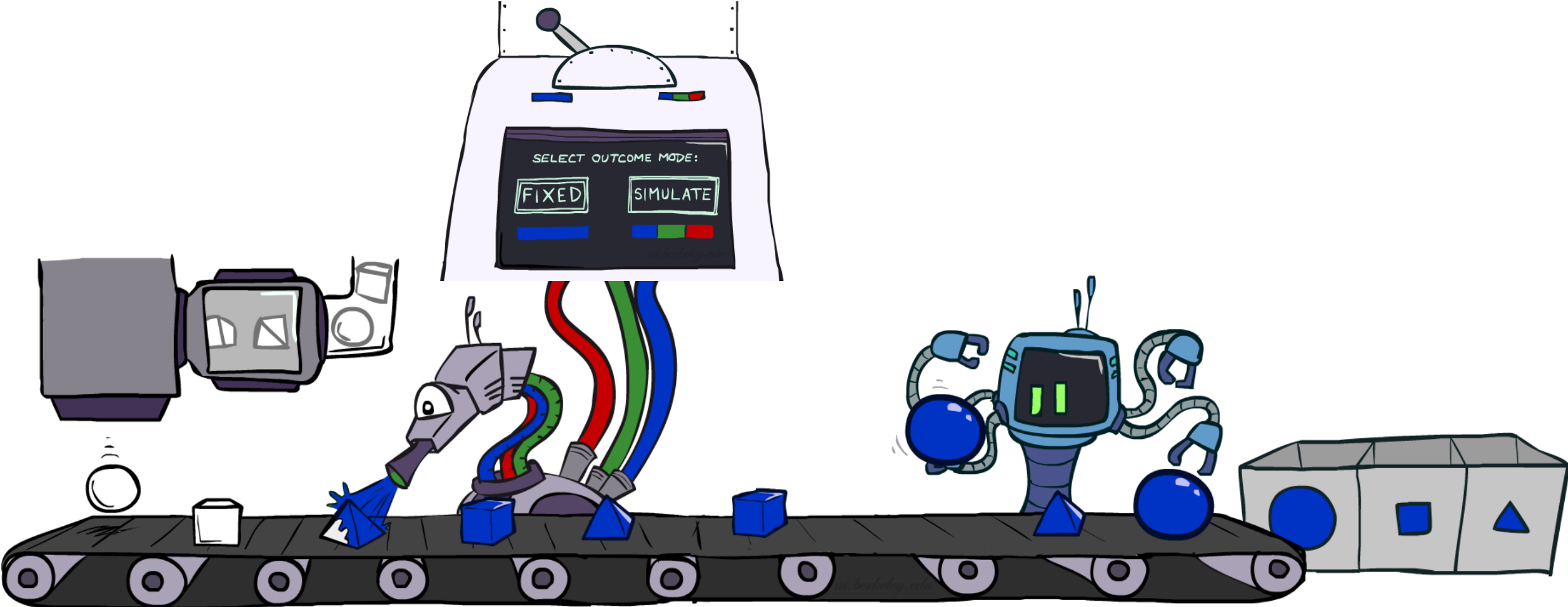
E. I have no idea

Counts  $N(A, B, C)$

+a	+b	+c	4
+a	+b	-c	
+a	-b	+c	3
+a	-b	-c	
-a	+b	+c	2
-a	+b	-c	
-a	-b	+c	1
-a	-b	-c	



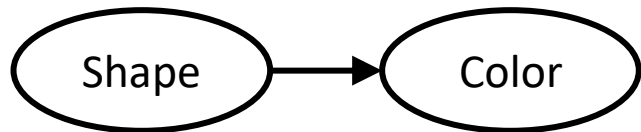
# Likelihood Weighting



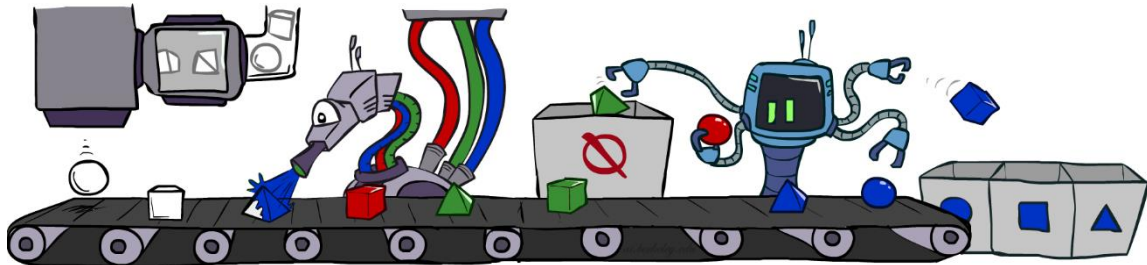
# Likelihood Weighting

## Problem with rejection sampling:

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



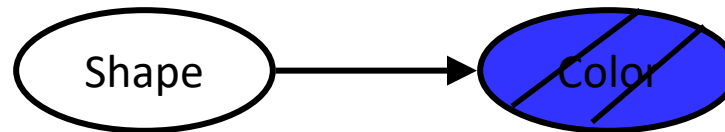
~~pyramid, green~~  
~~pyramid, red~~  
sphere, blue  
cube, red  
~~sphere, green~~



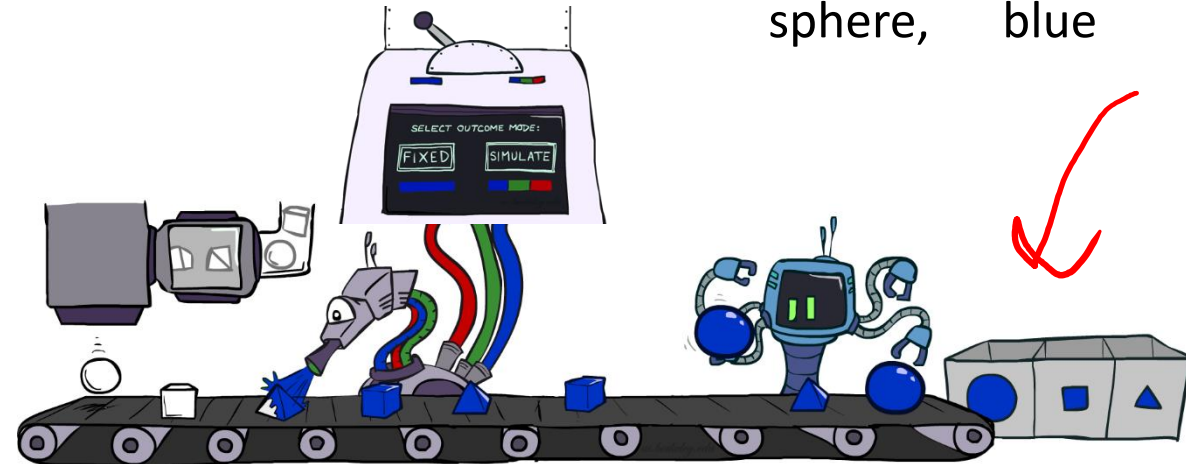
$$P(\text{cube, blue})$$
$$P(\text{cube} | \text{blue})$$

- Idea: fix evidence variables and sample the rest

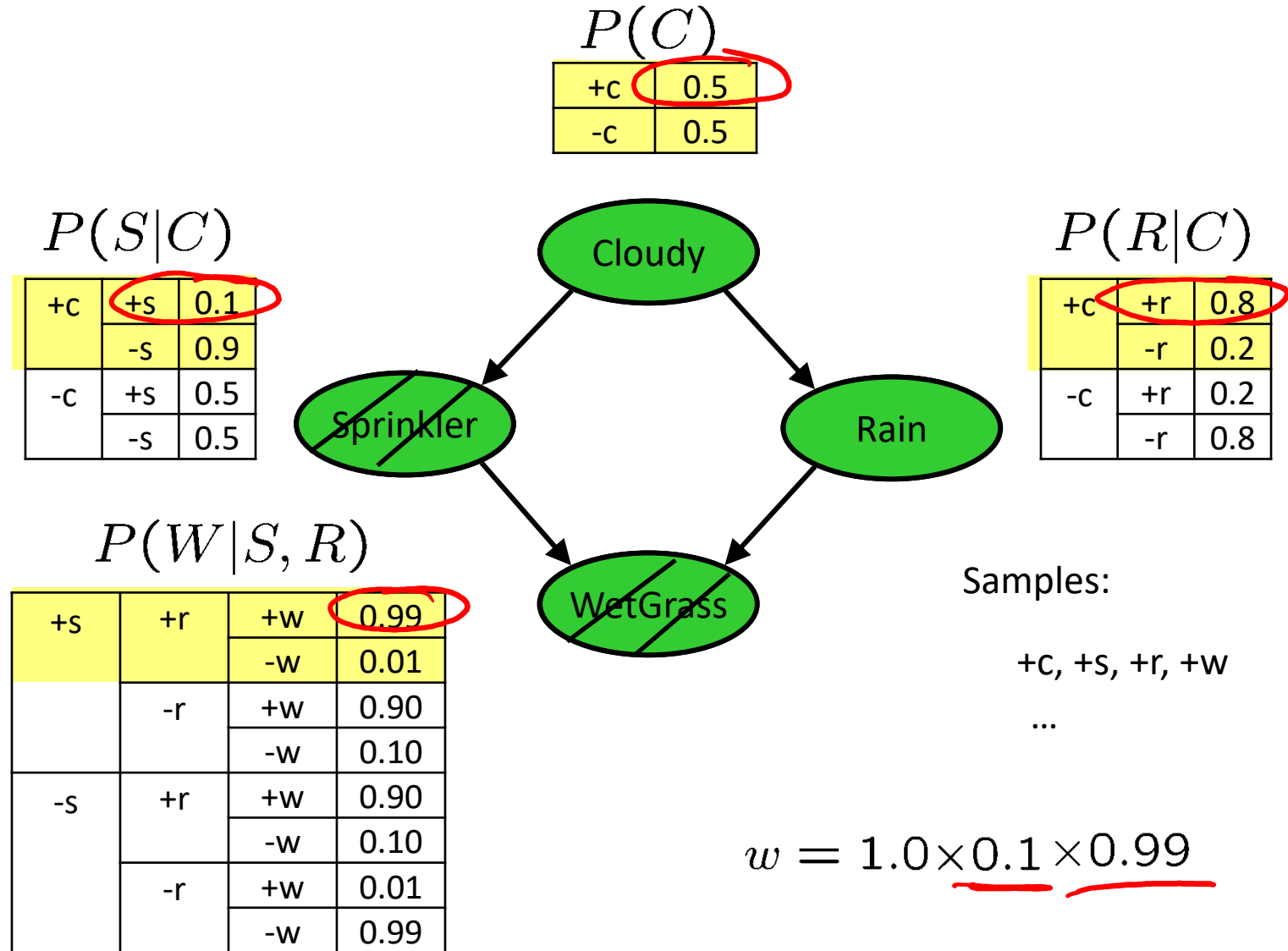
- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents



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



# Likelihood Weighting



# Likelihood Weighting

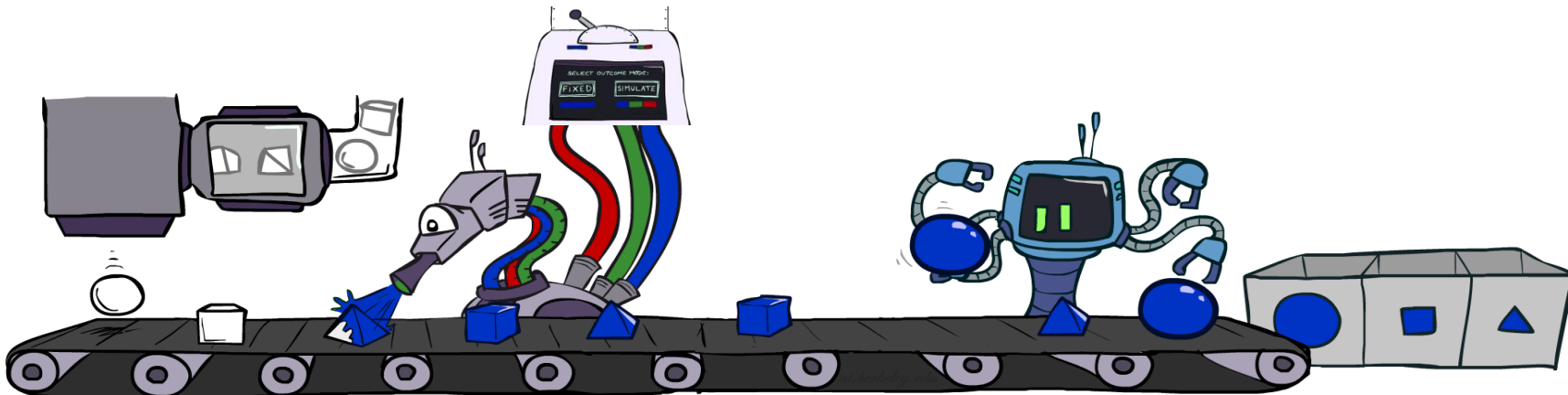
IN: evidence instantiation

$w = 1.0$

for  $i=1, 2, \dots, n$

- if  $X_i$  is an evidence variable
  - $X_i =$  observation  $x_i$  for  $X_i$
  - Set  $w = w * P(x_i | \text{Parents}(X_i))$
- else
  - Sample  $x_i$  from  $P(X_i | \text{Parents}(X_i))$

return  $(x_1, x_2, \dots, x_n), w$



# Likelihood Weighting

No evidence:  
Prior Sampling

Input: no evidence

for  $i=1, 2, \dots, n$

- Sample  $x_i$  from  $P(X_i \mid \text{Parents}(X_i))$

return  $(x_1, x_2, \dots, x_n)$

Some evidence:  
Likelihood Weighted Sampling

Input: evidence instantiation

$w = 1.0$

for  $i=1, 2, \dots, n$

if  $X_i$  is an evidence variable

- $X_i = \text{observation } x_i \text{ for } X_i$
- 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  $(x_1, x_2, \dots, x_n), w$

All evidence:  
Likelihood Weighted

Input: evidence instantiation

$w = 1.0$

for  $i=1, 2, \dots, n$

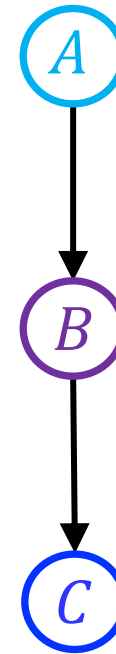
- Set  $w = w * P(x_i \mid \text{Parents}(X_i))$

return  $w$

# Remember Piazza Poll 2

How many  $\{-a, +b, -c\}$  samples out of  $N=1000$  should we expect?

- A. 1
- B. 50
- C. 125
- D. 200
- E. I have no idea



$P(A)$

+a	1/2
-a	1/2

$P(B|A)$

+a	+b	1/10
	-b	9/10
-a	+b	1/2
	-b	1/2

$P(C|B)$

+b	+c	4/5
	-c	1/5
-b	+c	1
	-c	0

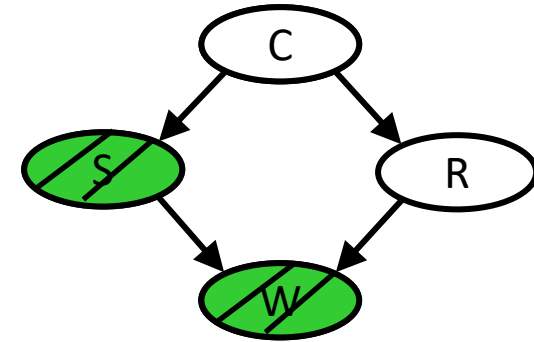
# Likelihood Weighting

Sampling distribution if  $z$  sampled and  $e$  fixed evidence

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

Now, samples have weights

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



Together, weighted sampling distribution is consistent

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

# Likelihood Weighting



$P(A)$

+a	1/2
-a	1/2

$P(B|A)$

+a	+b	1/10
	-b	9/10
-a	+b	1/2
	-b	1/2

$P(C|B)$

+b	+c	4/5
	-c	1/5
-b	+c	1
	-c	0

# Piazza Poll 4

Two identical samples from likelihood weighted sampling will have the same exact weights.

- A. True
- B. False
- C. It depends
- D. I don't know

## Piazza Poll 5

What does the following likelihood weighted value approximate?

$$\text{weight}_{(+a, -b, +c)} \cdot \frac{N(+a, -b, +c)}{N}$$

- A.  $P(+a, -b, +c)$
- B.  $P(+a, -b \mid +c)$
- C. I'm not sure

# Likelihood Weighting

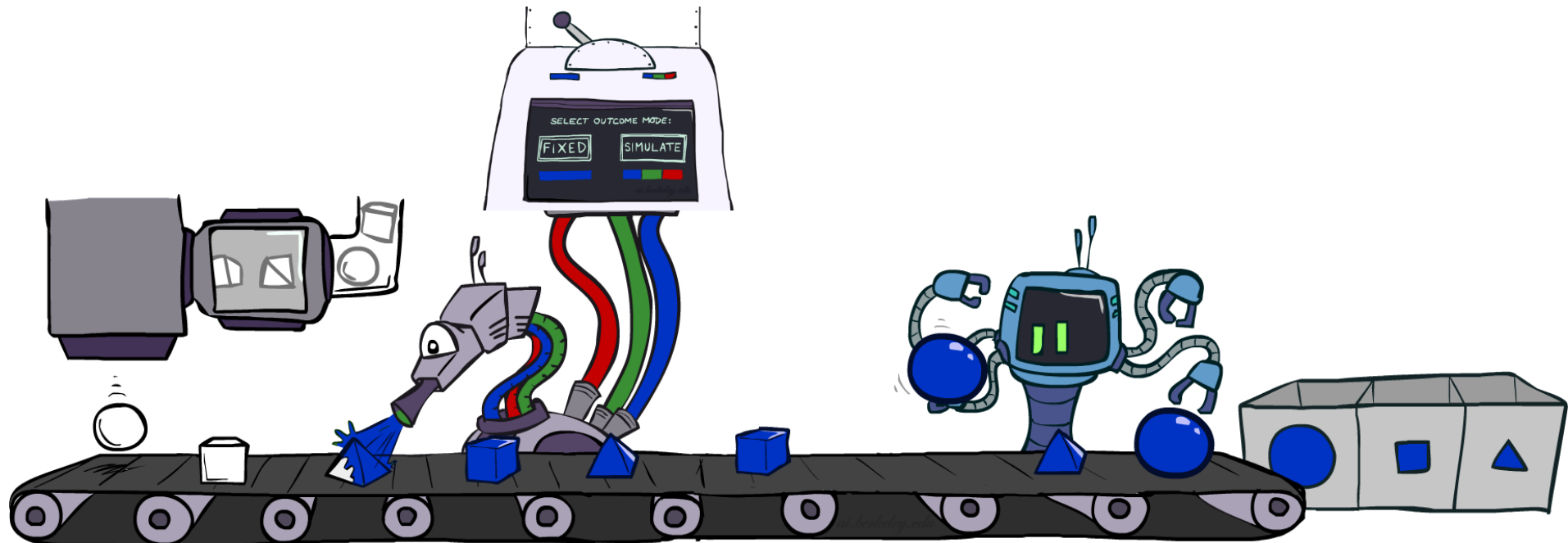
## Likelihood weighting is good

- 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



# Likelihood Weighting

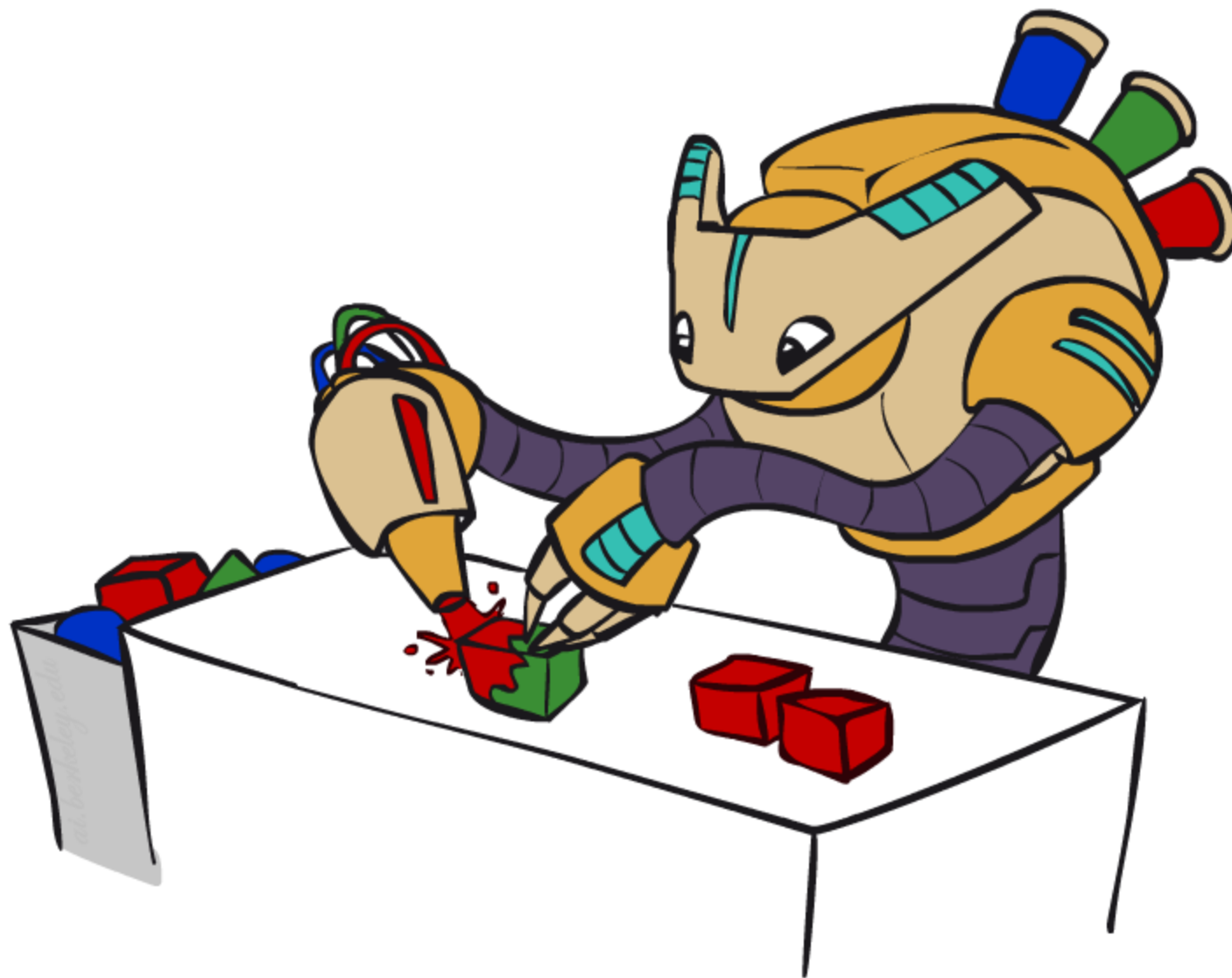
Likelihood weighting doesn't solve all our problems

- 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  $x_1, x_2, \dots, x_n$ .

1. Start with an arbitrary instantiation consistent with the evidence.
2. Sample one variable at a time, conditioned on all the rest, but keep evidence fixed.
3. 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

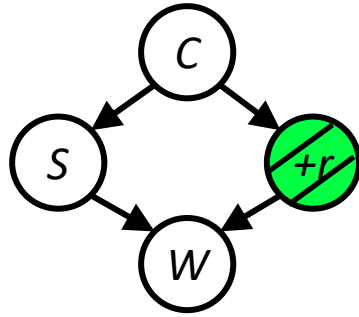
*Rationale:* both upstream and downstream variables condition on evidence.

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: $P(S | +r)$

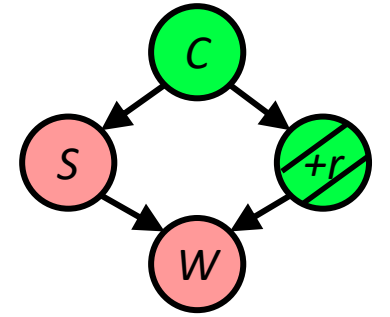
## Step 1: Fix evidence

- $R = +r$



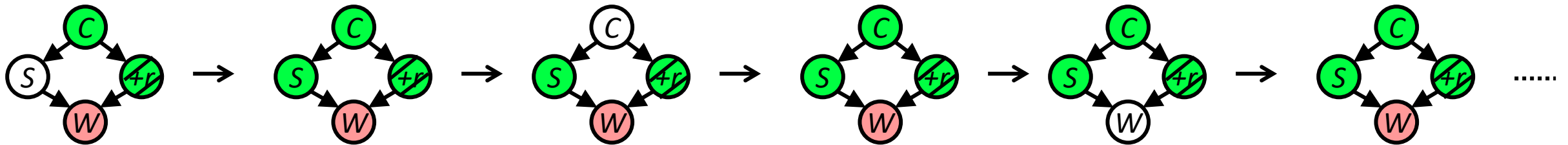
## Step 2: Initialize other variables

- Randomly



## Steps 3: Repeat

- Choose a non-evidence variable  $X$
- Resample  $X$  from  $P(X | \text{all other variables})$



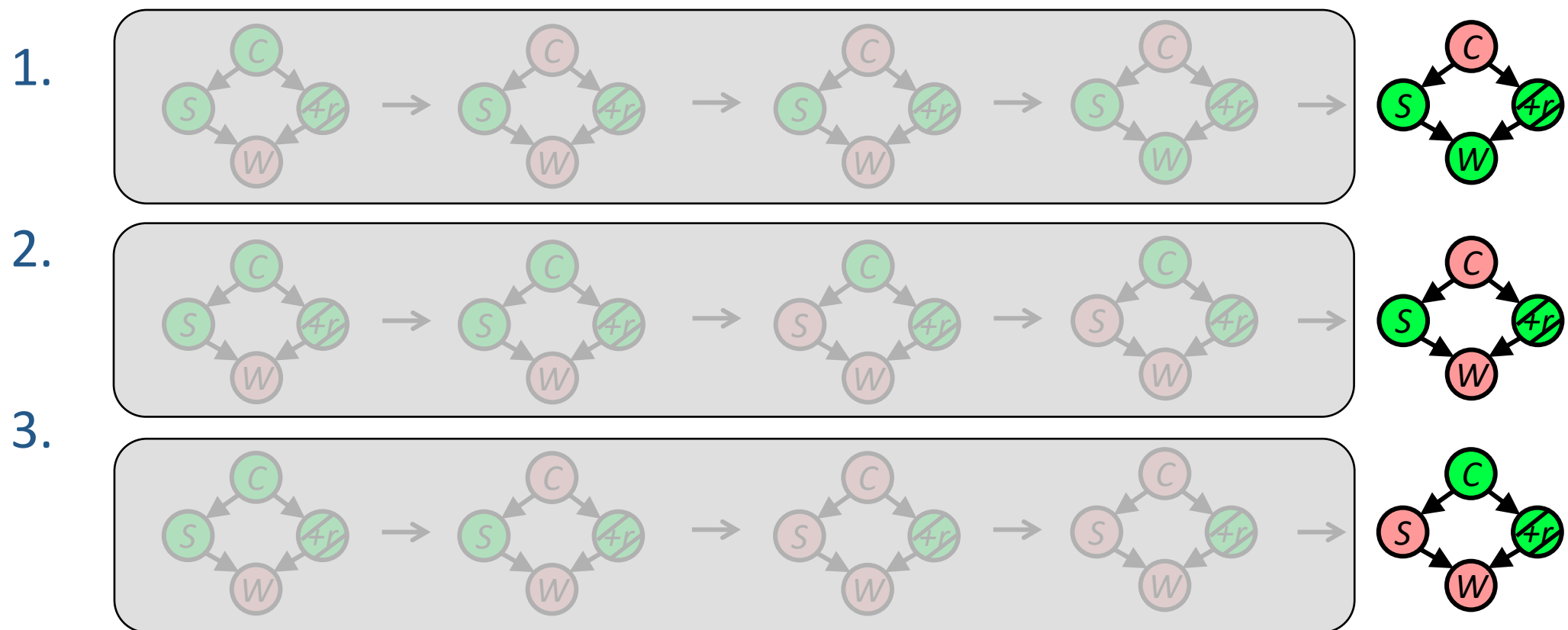
Sample from  $P(S | +c, -w, +r)$

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

Sample from  $P(W | +s, +c, +r)$

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

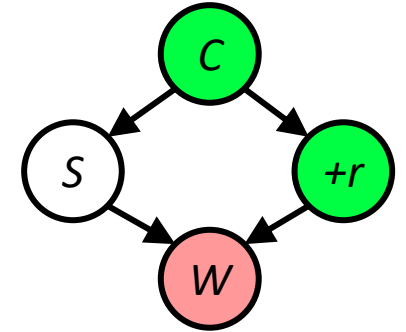
Keep only the last sample from each iteration:



# Efficient Resampling of One Variable

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

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



Many things cancel out – only CPTs with S remain!

More generally: only CPTs that have resampled variable need to be considered, and joined together

# Further Reading on Gibbs Sampling

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

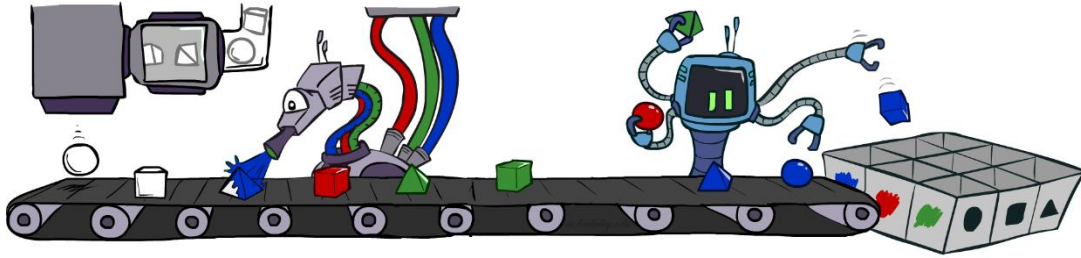
Gibbs sampling is a special case of more general methods called Markov chain Monte Carlo (MCMC) methods

- Metropolis-Hastings is one of the more famous MCMC methods (in fact, Gibbs sampling is a special case of Metropolis-Hastings)

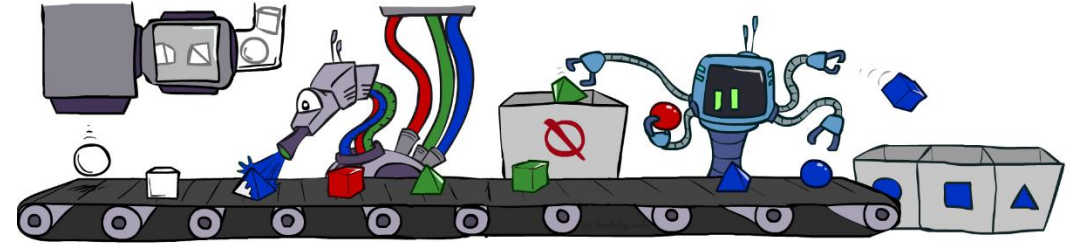
You may read about Monte Carlo methods – they're just sampling

# Bayes' Net Sampling Summary

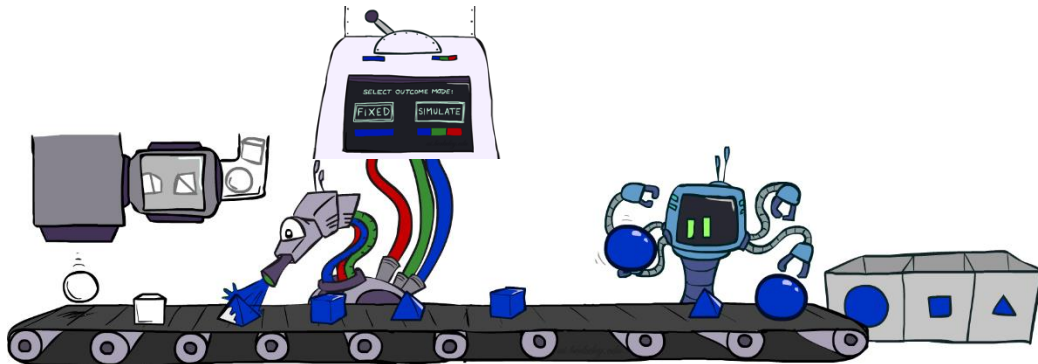
Prior Sampling  $P(Q, E)$



Rejection Sampling  $P(Q | e)$



Likelihood Weighting  $P(Q, e)$



Gibbs Sampling  $P(Q | e)$

