

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

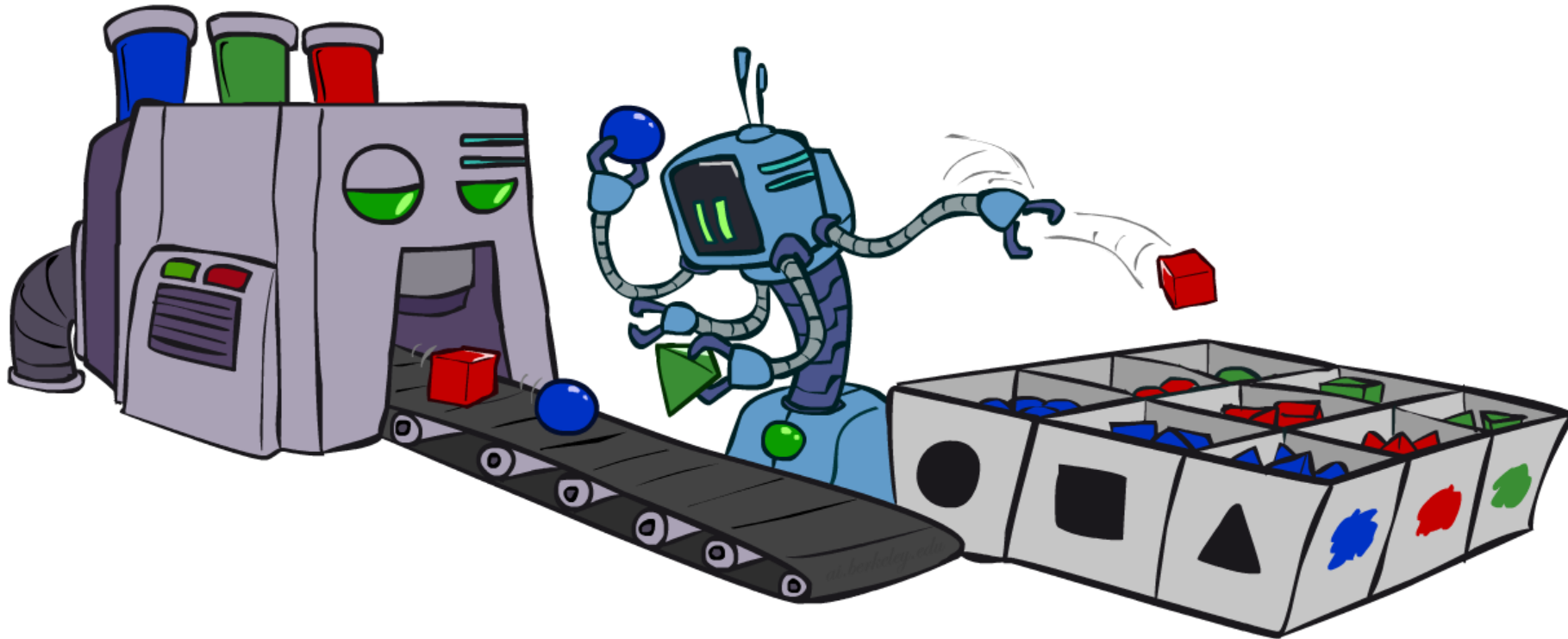
$-a, -b, +c$
 $+a, -b, +c$
 $-a, -b, +c$
 $-a, +b, +c$
 $+a, -b, +c$
 $-a, +b, -c$
 $-a, +b, +c$
 $-a, +b, +c$
 $+a, -b, +c$
 $-a, +b, +c$

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

AI: Representation and Problem Solving

Bayes Nets Sampling



Instructor: Pat Virtue

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

Review: Bayes Nets

Joint distributions $\rightarrow$ answer any query

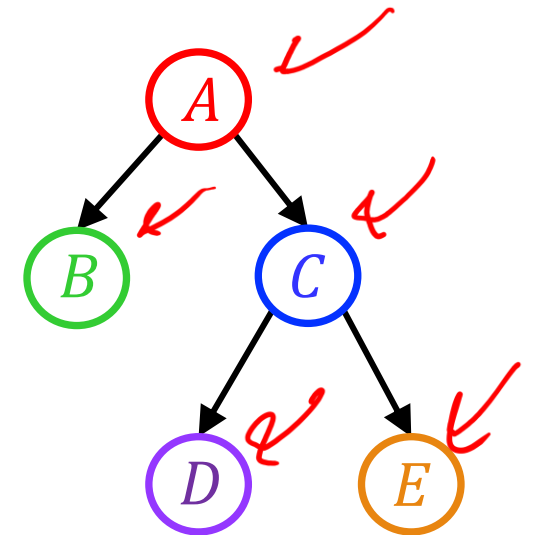
$$P(a \mid e) = \frac{1}{Z} P(a, e) = \frac{1}{Z} \sum_b \sum_c \sum_d P(a, b, c, d, e)$$

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

$$\underline{P(A, B, C, D, E)} = P(A) P(B|A) P(C|A) P(D|C) P(E|C)$$



Variable Elimination: Order matters

- 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)

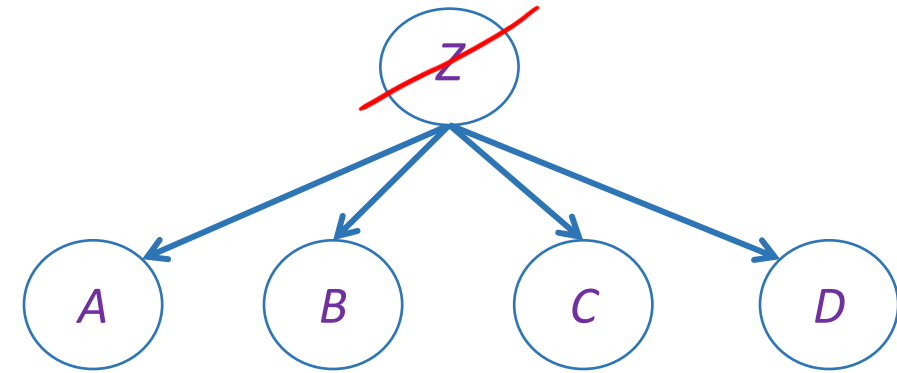
- Elimination Order: Z, C, B, A

- $P(D) = \alpha \sum_{a,b,c,z} P(a|z) P(b|z) P(c|z) P(D|z) P(z)$

- $= \alpha \sum_a \sum_b \sum_c \sum_z P(a|z) P(b|z) P(c|z) P(D|z) P(z)$

- Largest factor has 4 variables (A,B,C,D) (or 5 if you count pre-summation over Z)

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



$f(a, b, c, d, z)$

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

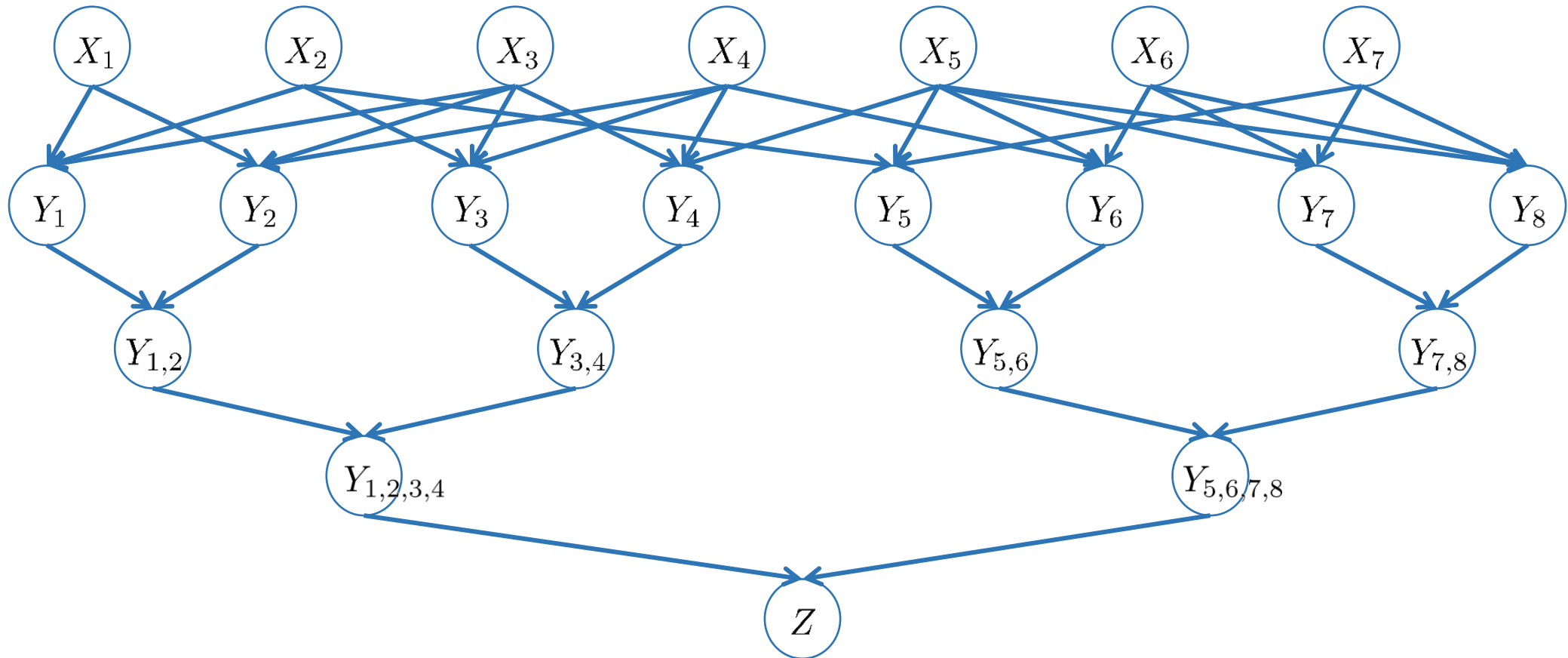
Does there always exist an ordering that only results in small factors?

- 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 and Independence

✓ 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 $P(-c \mid -a, +b)$?

- 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

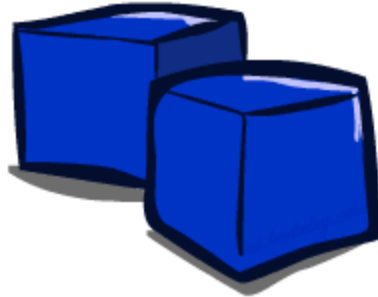


+a	+b	+c	0
+a	+b	-c	0
+a	-b	+c	3
+a	-b	-c	0
-a	+b	+c	4
-a	+b	-c	1
-a	-b	+c	2
-a	-b	-c	0

Approximate Inference: Sampling



$$\frac{5}{8}$$

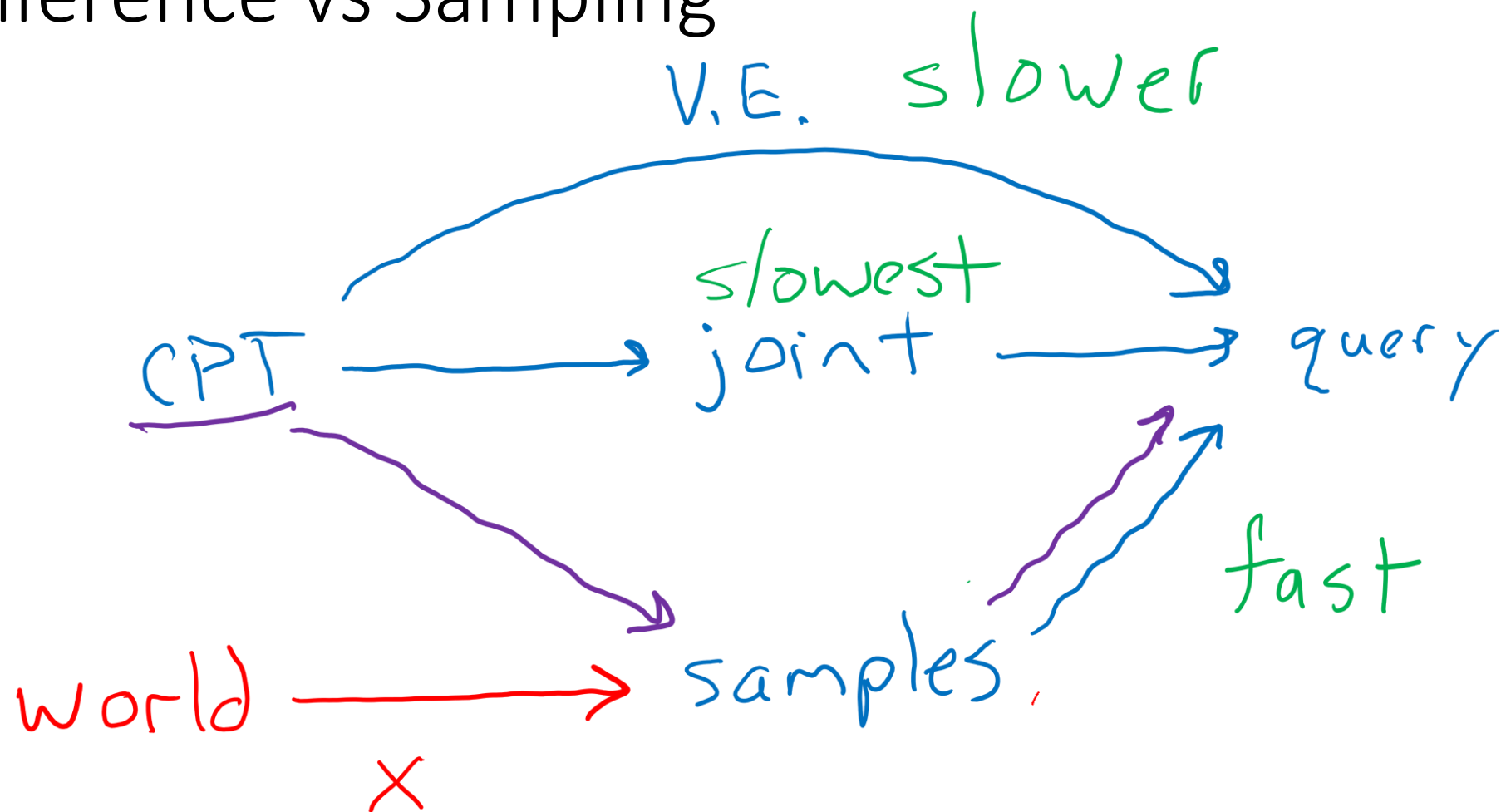


$$\frac{2}{8}$$

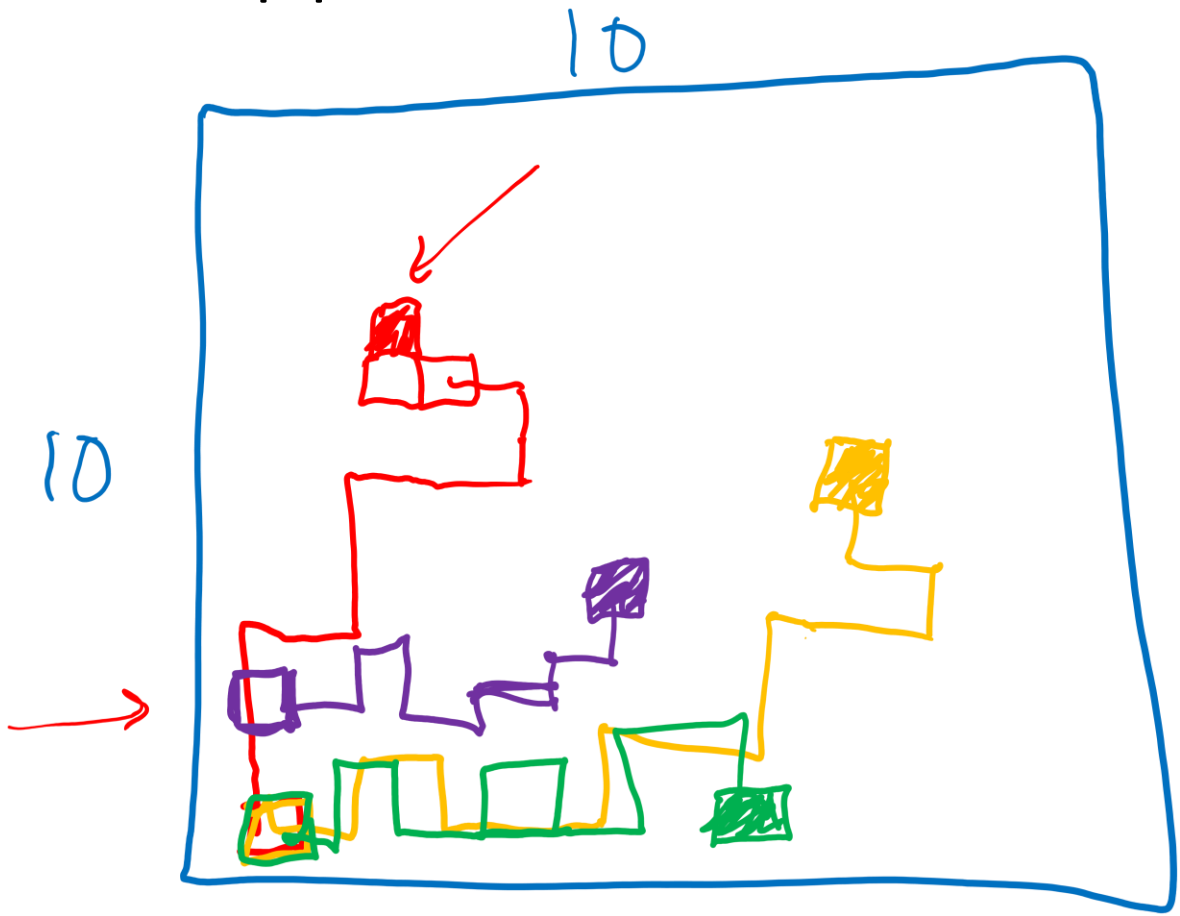


$$\frac{1}{8}$$

Inference vs Sampling

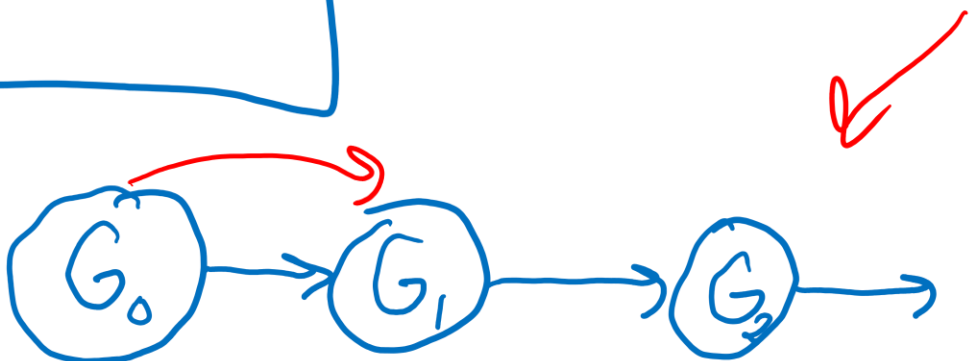


Motivation for Approximate Inference



$$\frac{P(G_{t+1} | G_t)}{P(G_0)}$$

$$P(G_{20})$$



Sampling

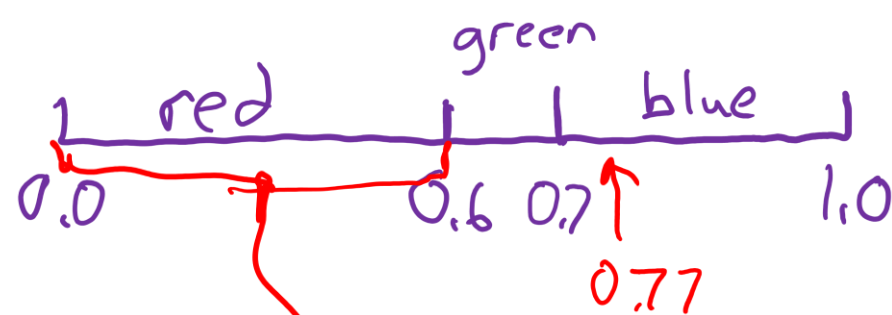
Sampling from given distribution

- 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

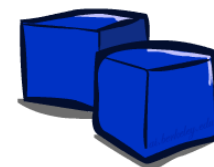


$0 \leq u < 0.6, \rightarrow C = \text{red}$

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

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

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

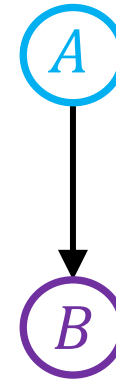
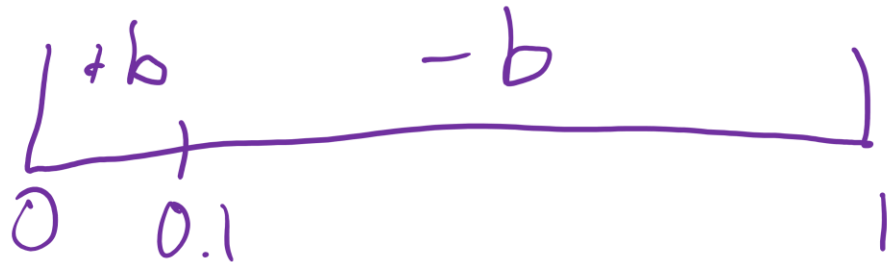


Sampling

How would you sample from a conditional distribution?



+a



$P(A)$

+a	1/2
-a	1/2

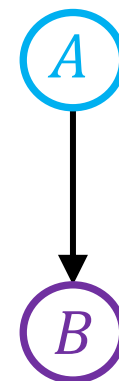
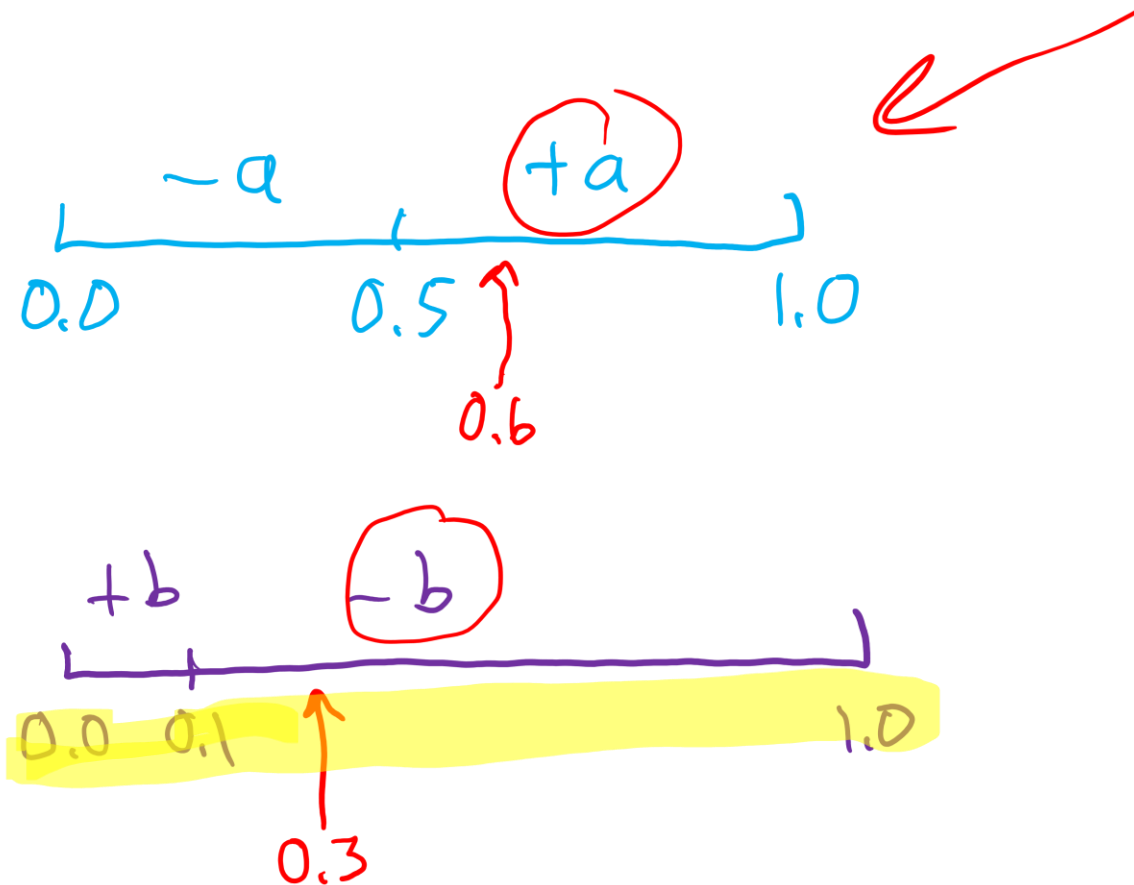
$P(B|A)$

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

Sampling

How would you sample from a conditional distribution?

$$P(A, B)$$



$$P(A)$$

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

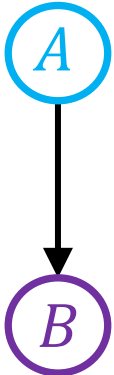
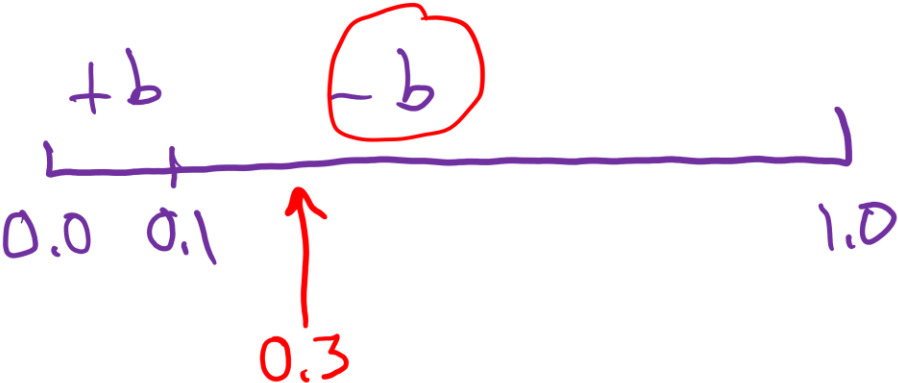
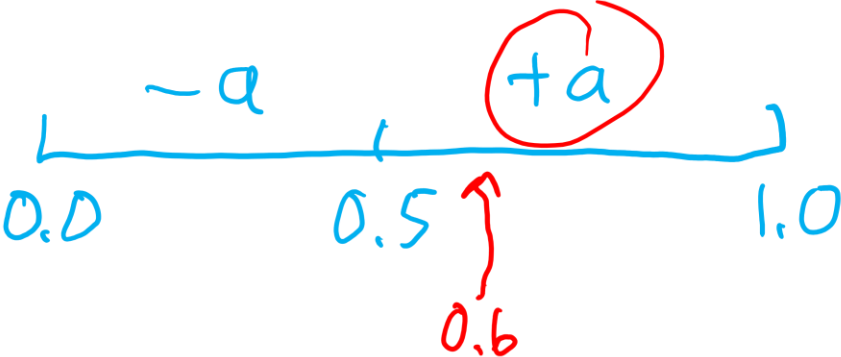
$$P(B|A)$$

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

Sampling

$$P(A, B)$$

How would you sample from a conditional distribution?

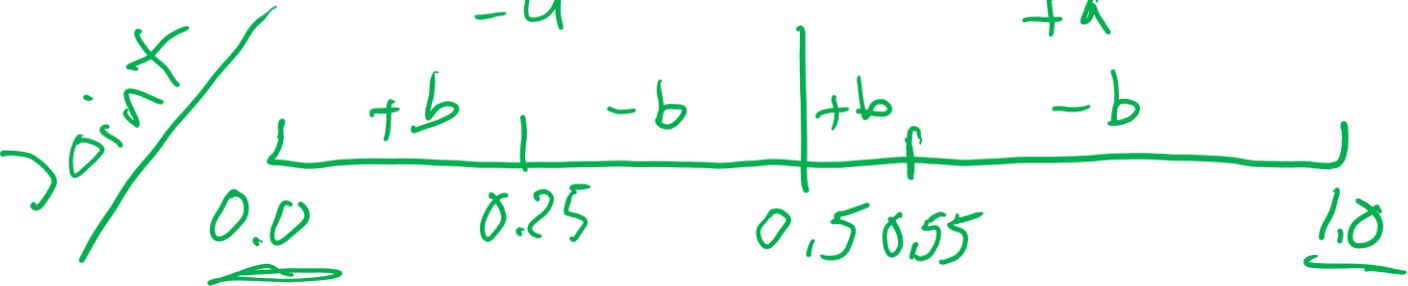


$P(A)$

+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

Handout

Create 1 sample

No evidence:
Prior Sampling

Input: no evidence

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

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

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

nodes
in Bayes net

Some evidence:

Likelihood Weighted Sampling

Input: evidence instantiation

$w = 1.0$

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

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 $(x_1, x_2, \dots, x_m), w$

All evidence:

Likelihood Weighted

Input: evidence instantiation

$w = 1.0$

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

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

return w

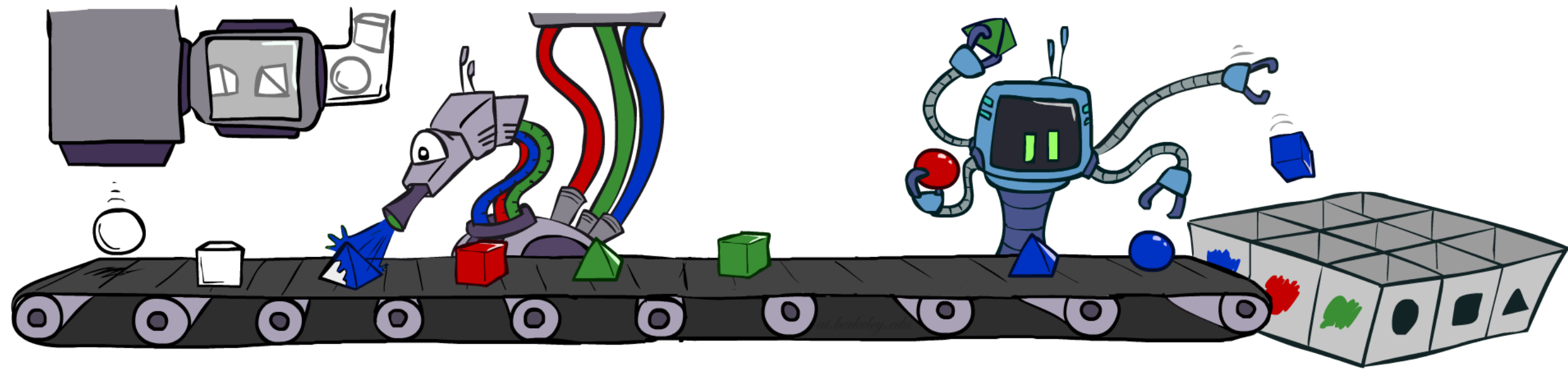
N samples

Hist of counts
counts/ N

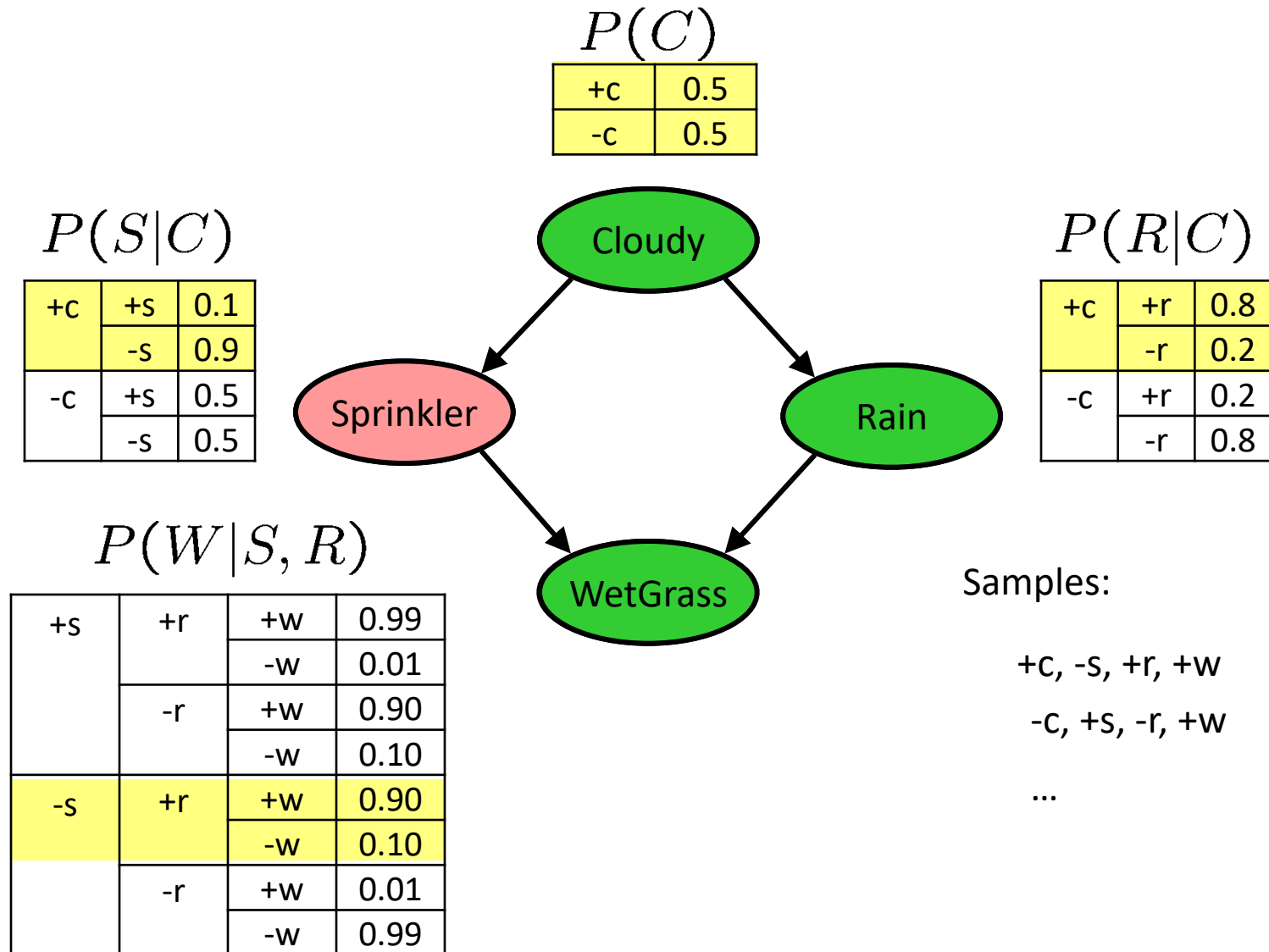
$\hat{P}(x_1, x_2, x_3, x_4)$

$P(x_1, x_2, x_3, x_4)$

Prior Sampling



Prior Sampling



Prior Sampling

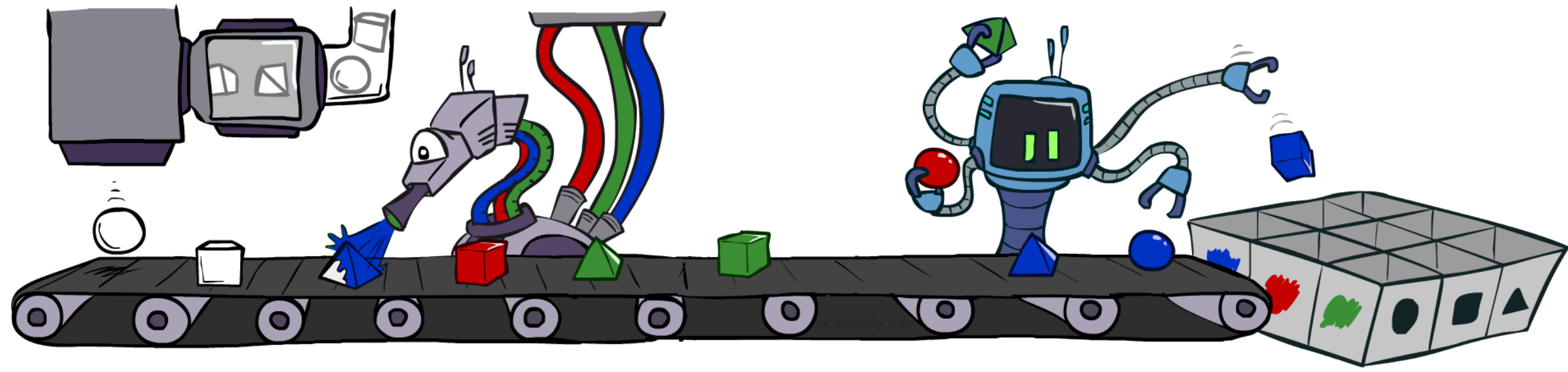
for $n = 1 \dots N$

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

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

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

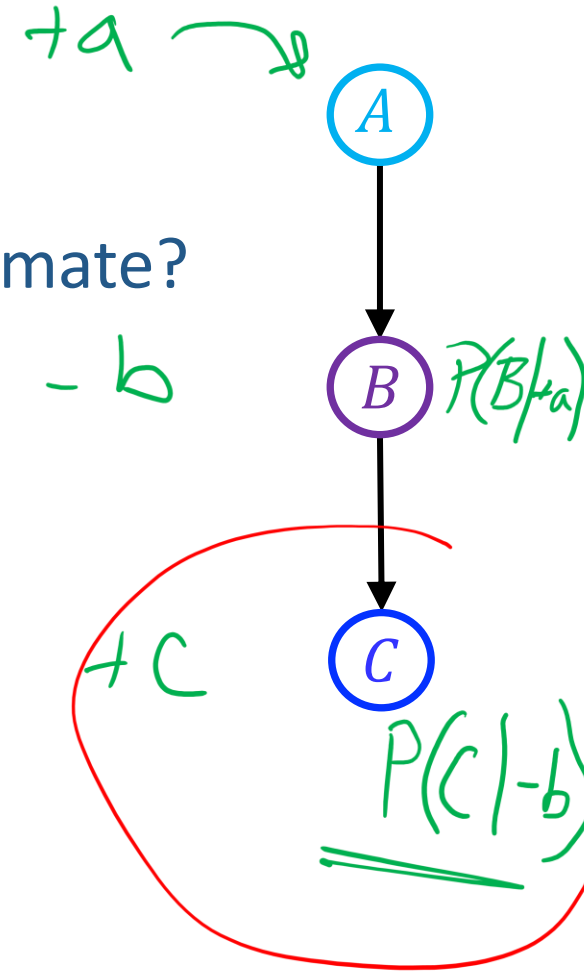
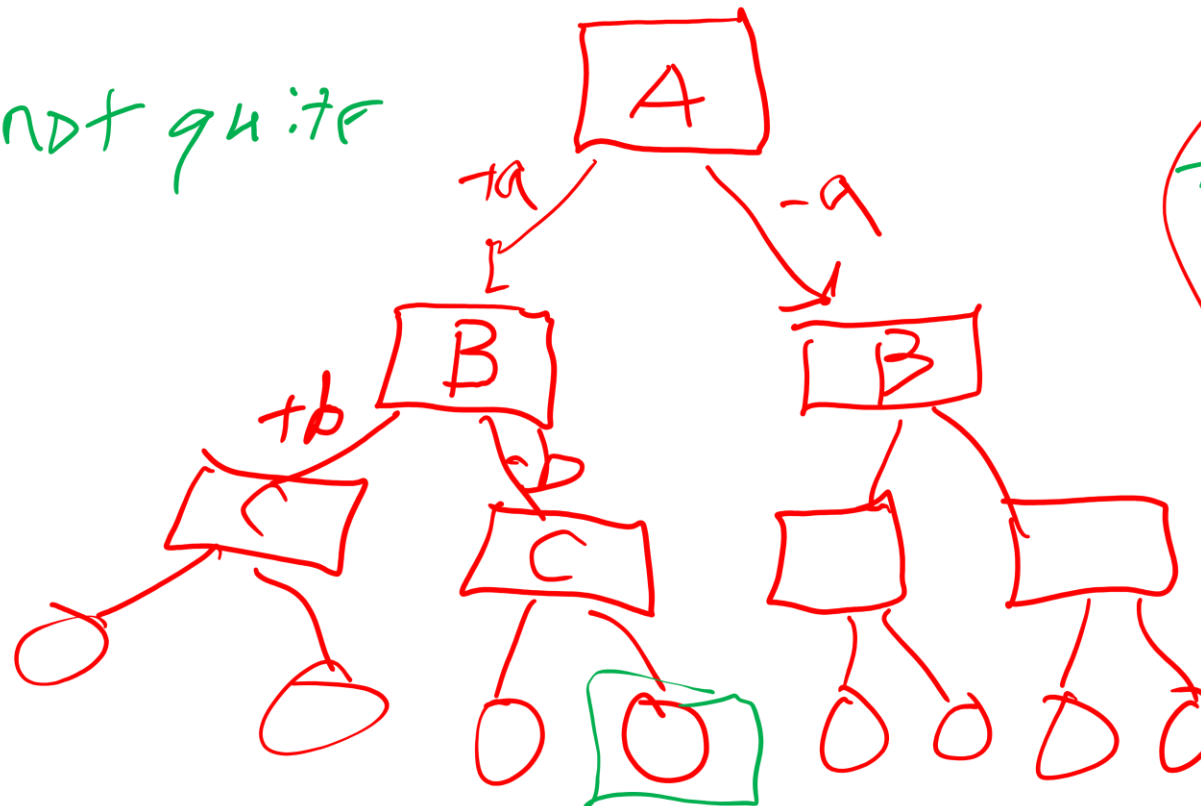
↓
Use samples
to get
approx
joint dist



Poll 1

Prior Sampling: What does the value $\frac{N(+a, -b, +c)}{N}$ approximate?

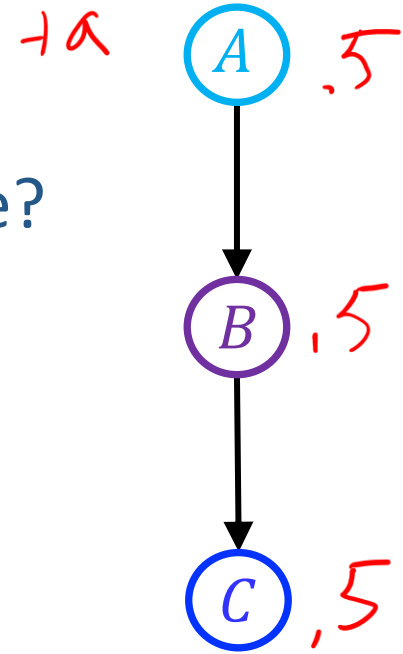
- A. $P(+a, -b, +c)$
- B. $P(+c \mid +a, -b)$
- C. $P(+c \mid -b)$ ← not quite
- D. $P(+c)$
- E. I don't know



Poll 1

Prior Sampling: What does the value $\frac{N(+a, -b, +c)}{N}$ 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



Let's assume for simplicity
that all probabilities are 0.5:

$$P(a) = 0.5 \forall a$$

$$P(b \mid a) = 0.5 \forall a, b$$

$$P(c \mid b) = 0.5 \forall b, c$$

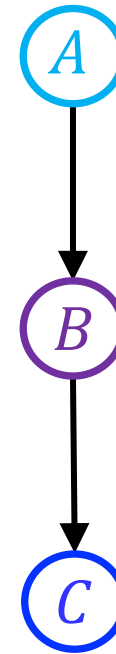
Let's take 1000 samples:

$+a$	$-b$	$+c$
498	251	125
1000	498	251

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

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



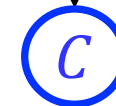
$P(A)$

+a	1/2
-a	1/2



$P(B|A)$

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



$P(C|B)$

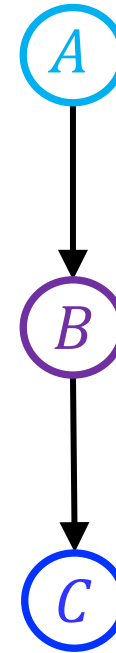
+b	+c	4/5
	-c	1/5
-b	+c	1
	-c	0

Probability of a sample

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

Algorithm: Multiply probability of
each node given parents:

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

Example

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

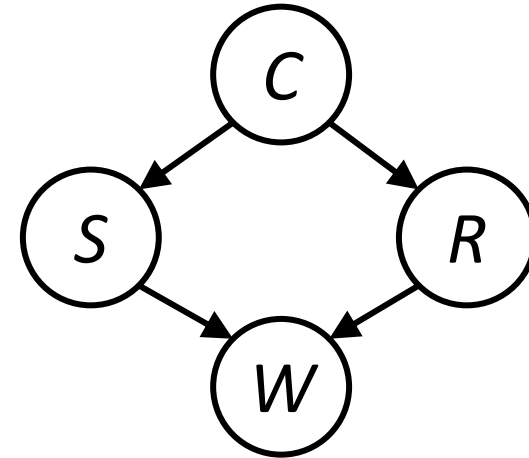
+C, -S, +r, +W

+C, +S, +r, +W

-C, +S, +r, -W

+C, -S, +r, +W

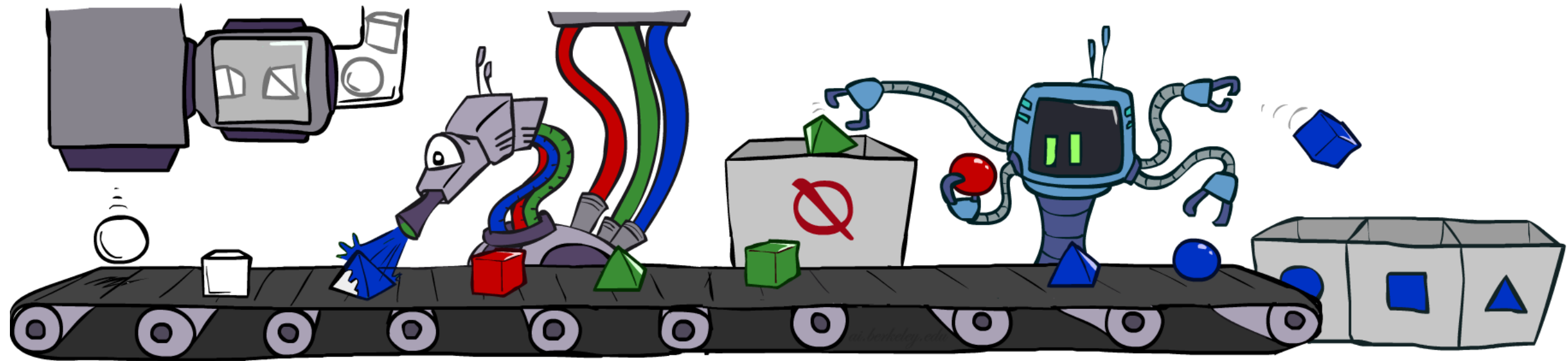
-C, -S, -r, +W



If we want to know $P(W)$

- We have counts <+w:4, -w:1>
- Normalize to get $P(W) = \langle +w:0.8, -w:0.2 \rangle$
- 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?)

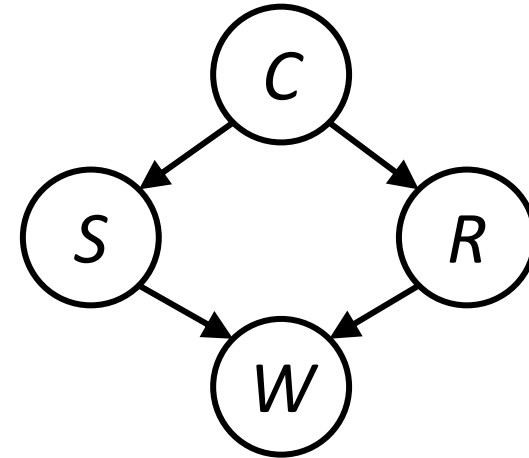
Rejection Sampling



Rejection Sampling

Let's say we want $P(C \mid +s)$

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



$+C \quad \frac{1}{2}$
 $-C \quad \frac{1}{2}$

~~$+C, -S, +r, +W$~~
 $+C, +S, +r, +W$
 $-C, +S, +r, -W$
 ~~$+C, -S, +r, +W$~~
 ~~$-C, -S, -r, +W$~~

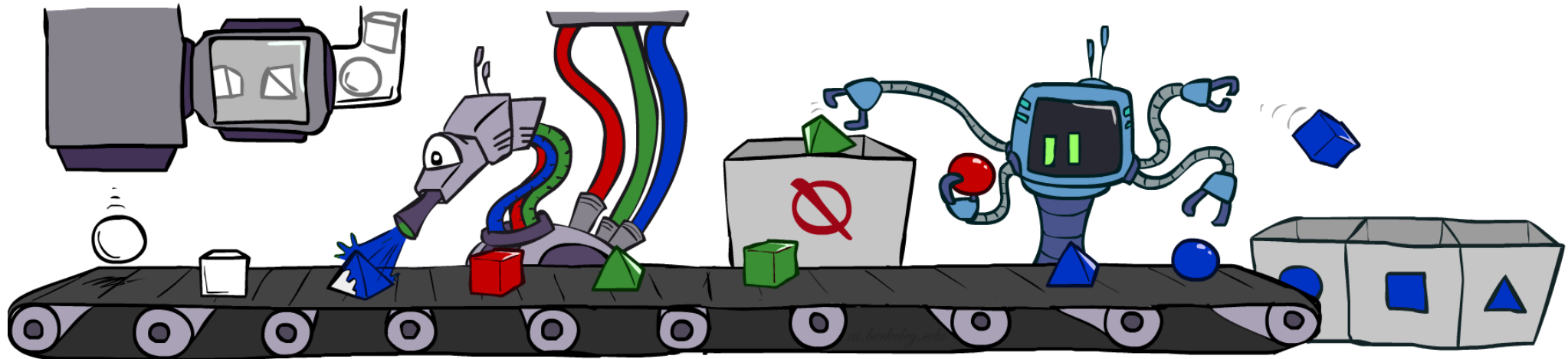
Rejection Sampling

IN: evidence instantiation

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

- 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 $(x_1, x_2, \dots, x_m)$



Poll 3

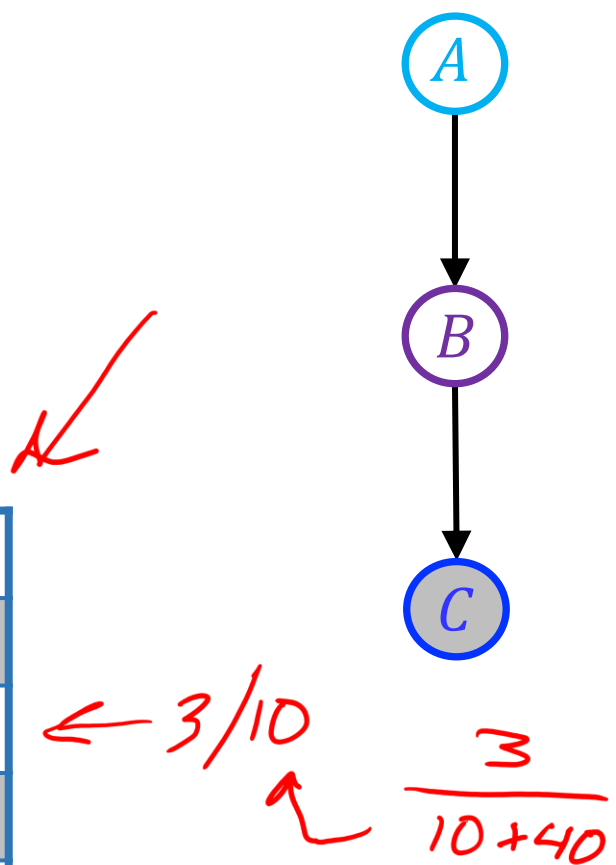
? [40]

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

- ☒ A. $P(+a, -b, +c)$
- ☒ B. $P(+a, -b \mid +c)$ $\approx 3/10$
- ☒ C. $P(+a, -b \mid -c)$
- ☐ D. None of the above

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	?



Poll 3

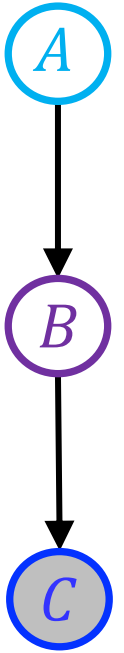
What queries can we (approximately) answer with rejection sampling samples (evidence: $+c$)? Select all that apply

- A. $P(+a, -b, +c)$
- B. $P(+a, -b \mid +c)$
- C. $P(+a, -b \mid -c)$
- D. None of the above

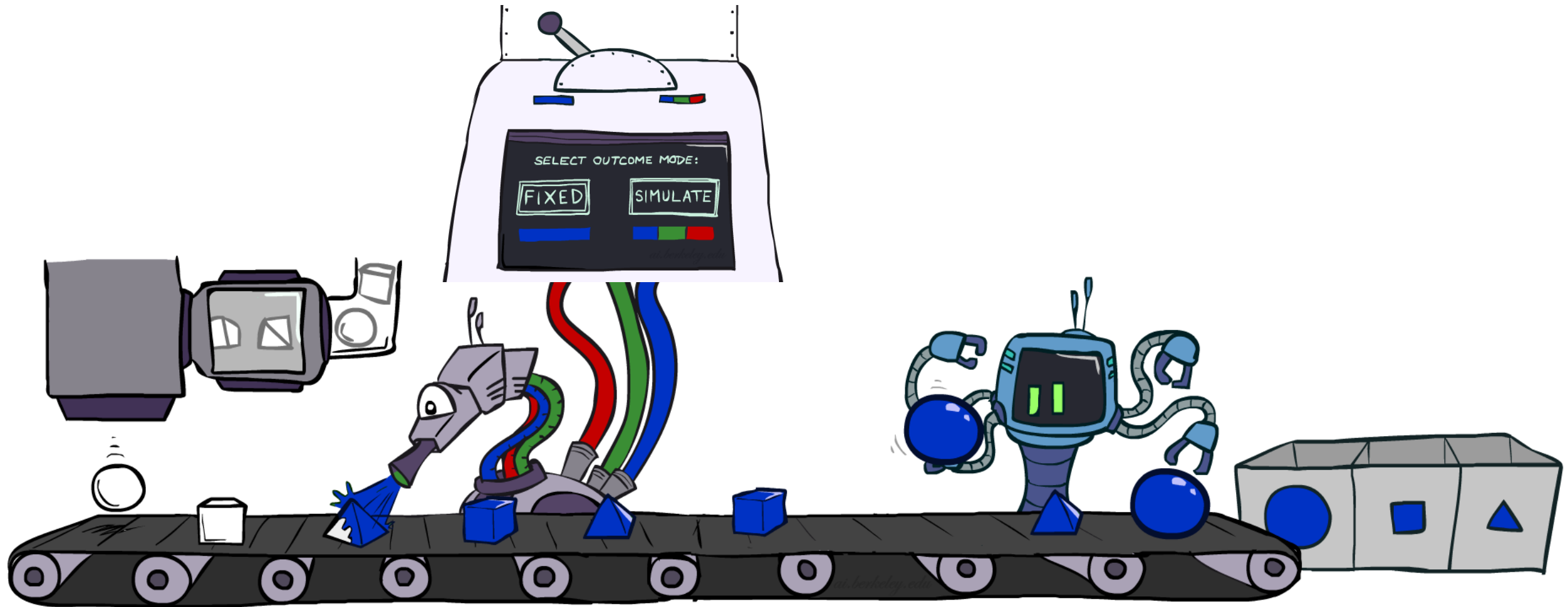
A: only if we also have total number of attempts

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	



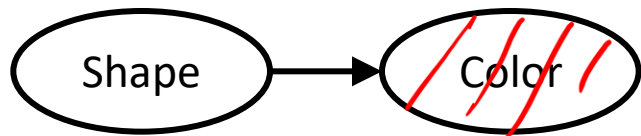
Likelihood Weighting



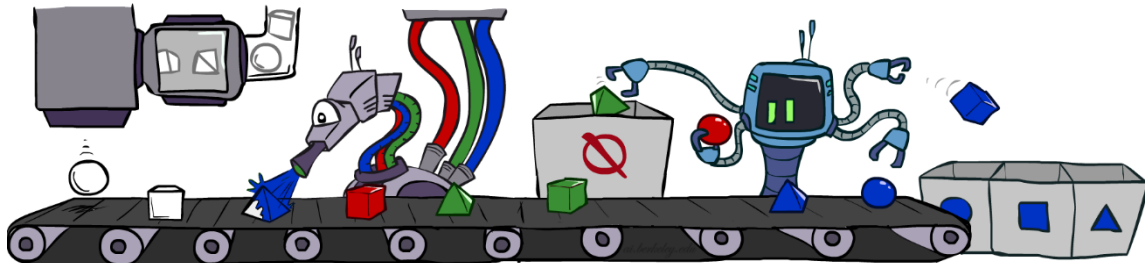
Likelihood Weighting

Problem with rejection sampling:

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

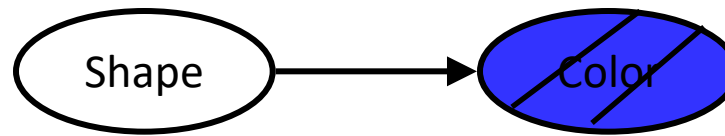


~~pyramid, green~~
~~pyramid, red~~
sphere, blue
cube, red
~~sphere, green~~

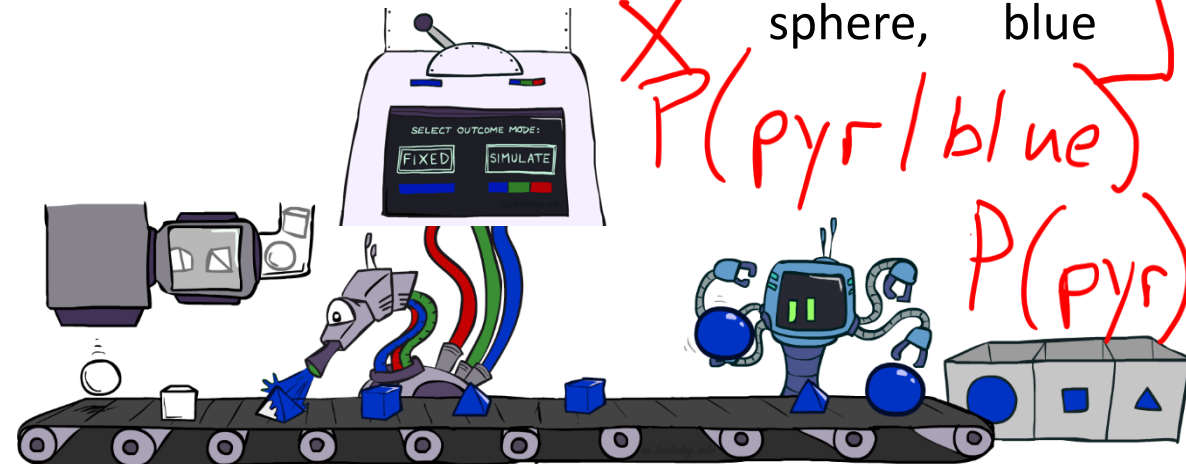


▪ Idea: fix evidence variables and sample the rest

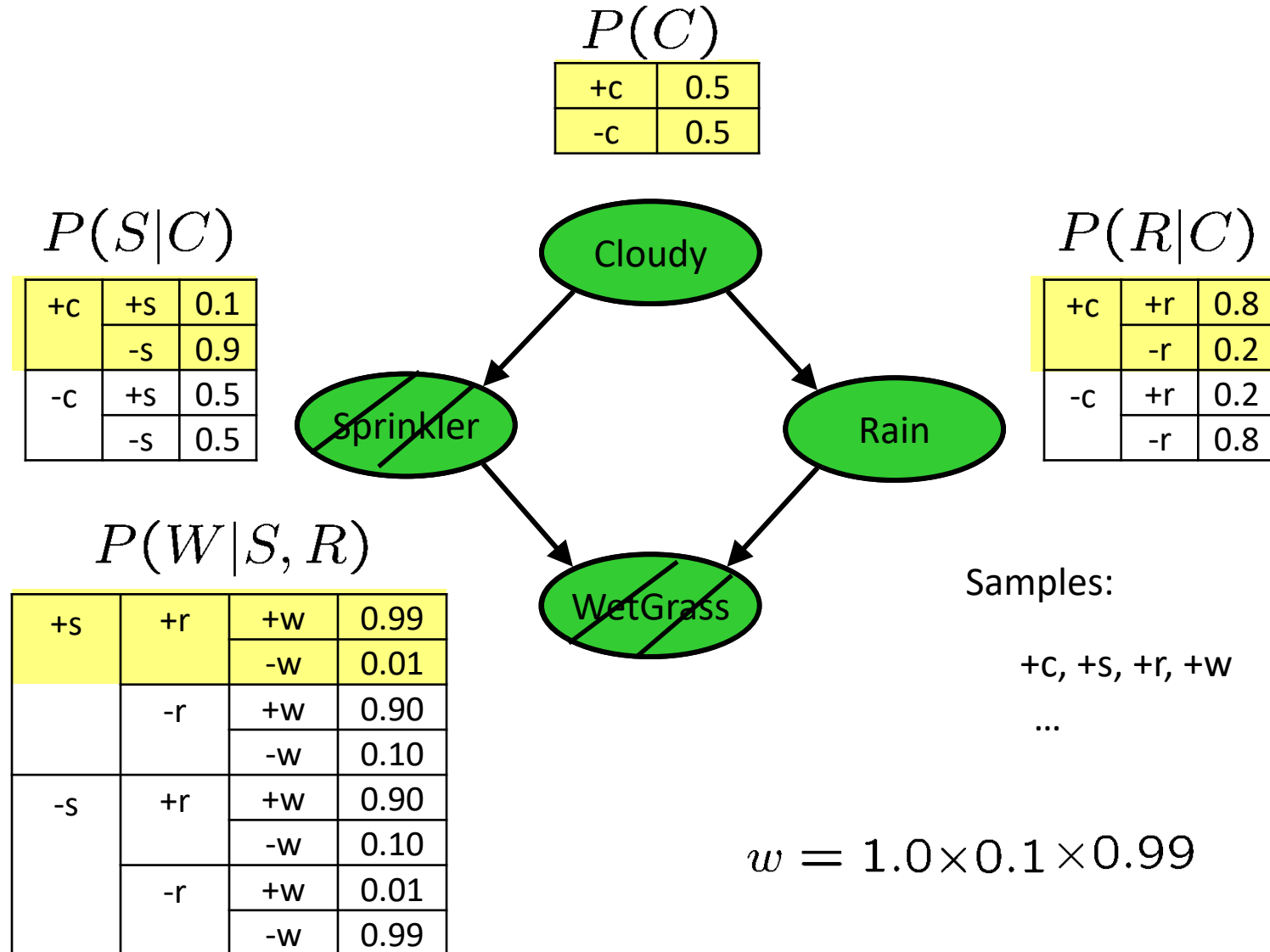
- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents



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



Likelihood Weighting



Likelihood Weighting

IN: evidence instantiation

$w = 1.0$

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

- 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 $(x_1, x_2, \dots, x_m), w$

$N(s, c)$

cube
pyr
sph

b
b
b

100
300
600

$w = .1$
 $w = .2$
 $w = .3$

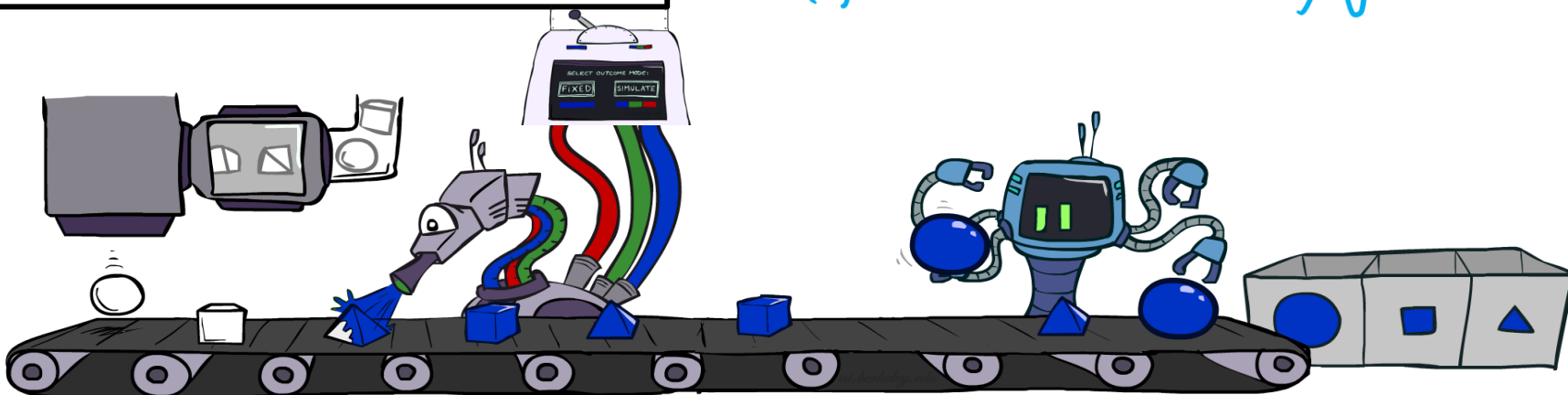
$\leftarrow P(b|cu)$
 $\leftarrow P(b|py)$
 $\leftarrow P(b|sp)$

$\hat{P}(s)$

$P(b|s)$

$$\hat{P}(\text{pyr}, \text{blue}) = \frac{N(\text{pyr}, \text{blue})}{N}$$

$w_{\text{pyr}, \text{blue}}$



Likelihood Weighting

No evidence:

Prior Sampling

Input: no evidence

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

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

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

$$P(x_1, x_2, \dots, x_m) \approx \frac{N(x_1, x_2, \dots, x_m)}{N}$$

Some evidence:

Likelihood Weighted Sampling

Input: evidence instantiation

$w = 1.0$

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

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 $(x_1, x_2, \dots, x_m), w$

$$P(x_1, x_2, \dots, x_m) \approx w \cdot \frac{N(x_1, x_2, \dots, x_m)}{N}$$

All evidence:

Likelihood Weighted

Input: evidence instantiation

$w = 1.0$

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

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

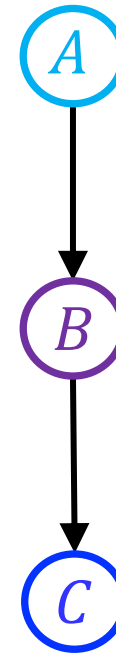
return w

$$P(x_1, x_2, \dots, x_m) = w$$

Remember 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

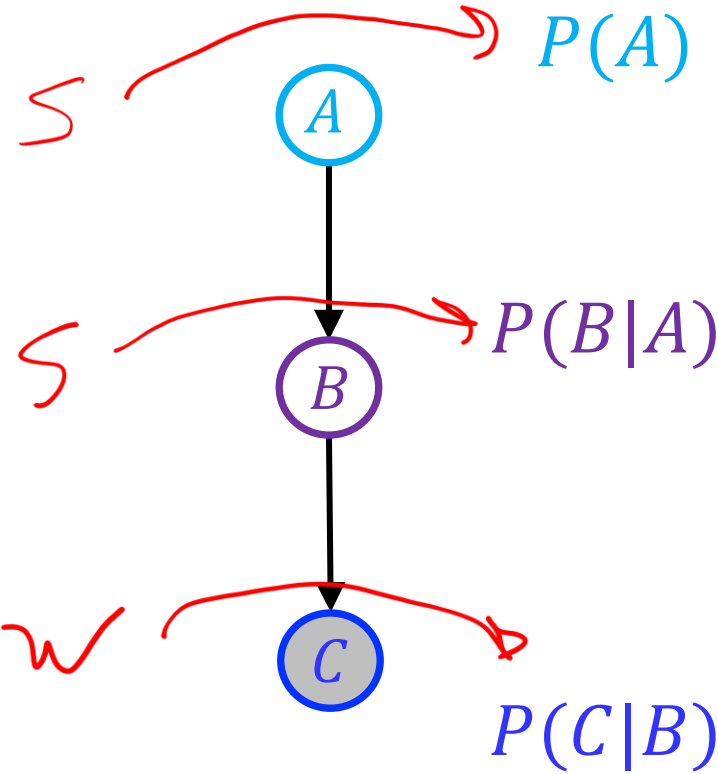
$P(B|A)$

+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



+a	1/2
-a	1/2

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

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

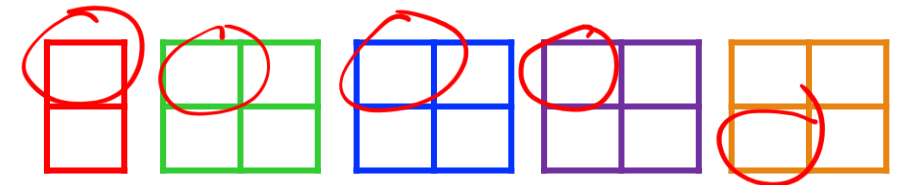
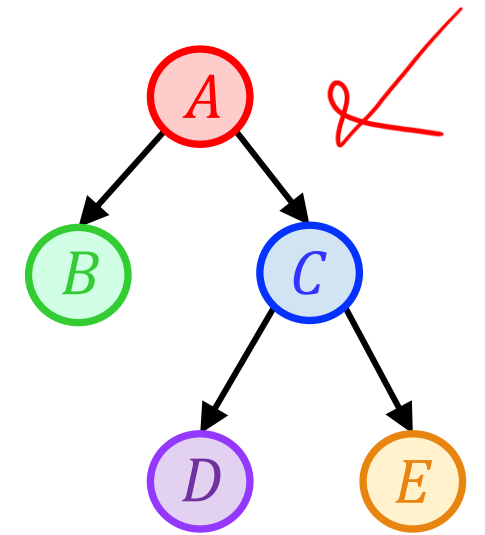
Likelihood Weighting Distribution

Consistency of likelihood weighted sampling distribution

Evidence: $+a$, $+b$, $-c$, $-d$, $+e$

Joint from Bayes nets

$$P(A, B, C, D, E) = P(+a) P(+b|+a) P(-c|+a) P(-d|-c) P(+e|-c)$$



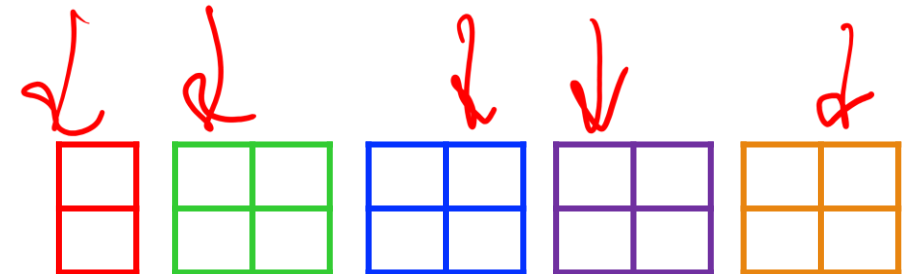
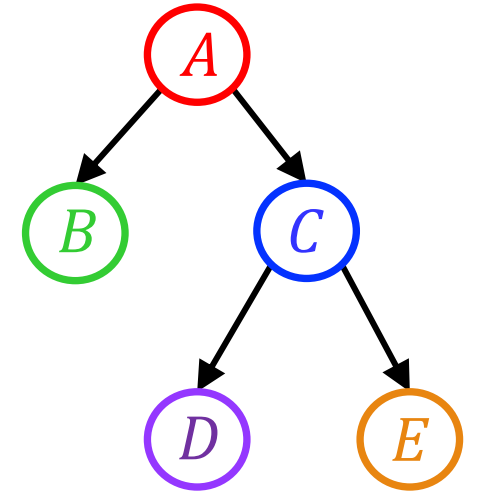
Likelihood Weighting Distribution

Consistency of likelihood weighted sampling distribution

Evidence: None

Joint from Bayes nets

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



Likelihood Weighting Distribution

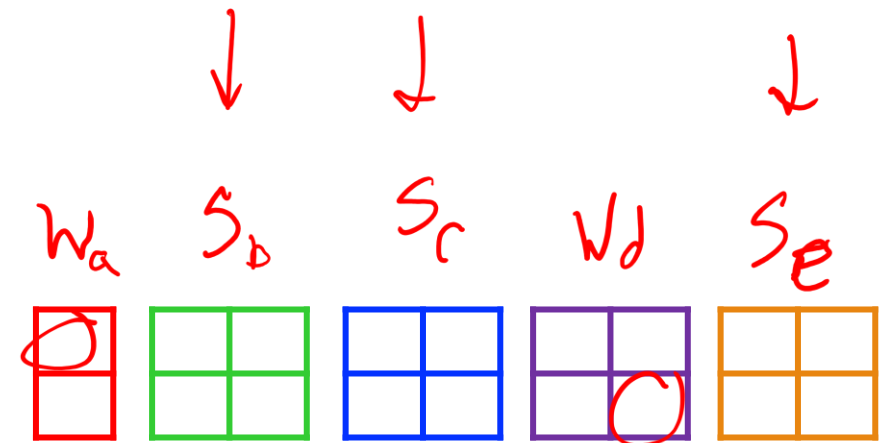
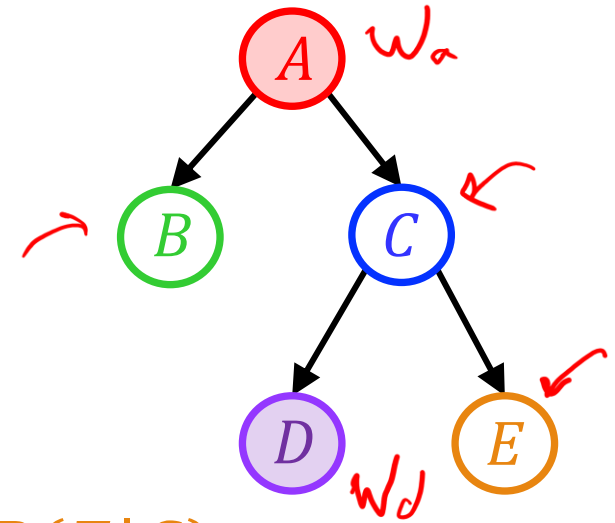
Consistency of likelihood weighted sampling distribution

Evidence: $+a$, $-d$

Joint from Bayes nets

$$P(A, B, C, D, E) = P(+a) P(B|+a) P(C|+a) P(-d|C) P(E|C)$$

$$W = w_a \times w_d$$



Poll 4

Given a fixed query, 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

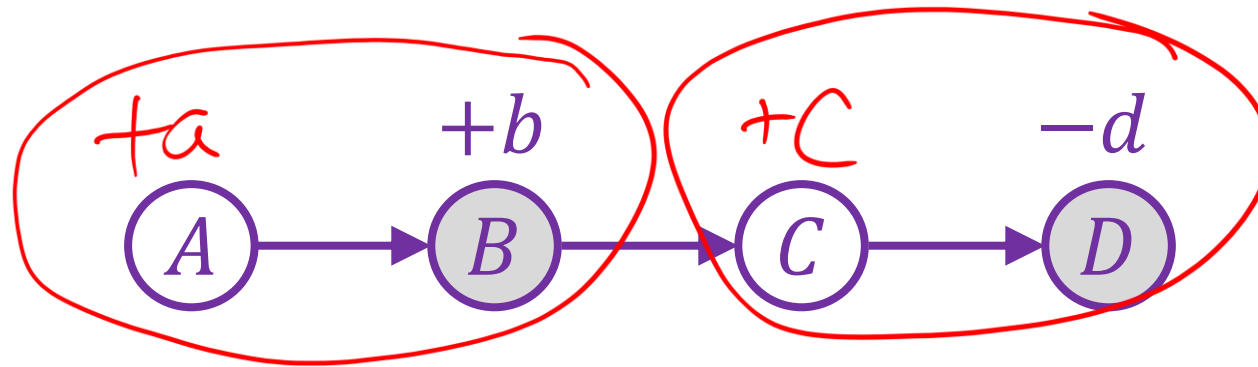
Poll 4

Given a fixed query, 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

Example:

$$P(A, C \mid +b, -d)$$



Poll 5

Assuming an observed value +c, 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

Poll 5

Assuming an observed value +c, 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

$$\text{weight}_{(x_1, x_2, x_3, e_1, e_2, e_3)} = \prod_i P(e_i \mid \text{Parents}(e_i))$$
$$\frac{N(x_1, x_2, x_3, e_1, e_2, e_3)}{N} \approx \prod_j P(x_j \mid \text{Parents}(x_j))$$

$$\text{weight}_{(x_1, x_2, x_3, e_1, e_2, e_3)} \cdot \frac{N(x_1, x_2, x_3, e_1, e_2, e_3)}{N} \approx P(x_1, x_2, x_3, e_1, e_2, e_3)$$

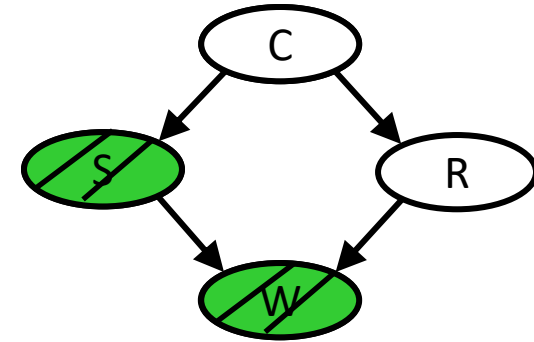
Likelihood Weighting

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

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



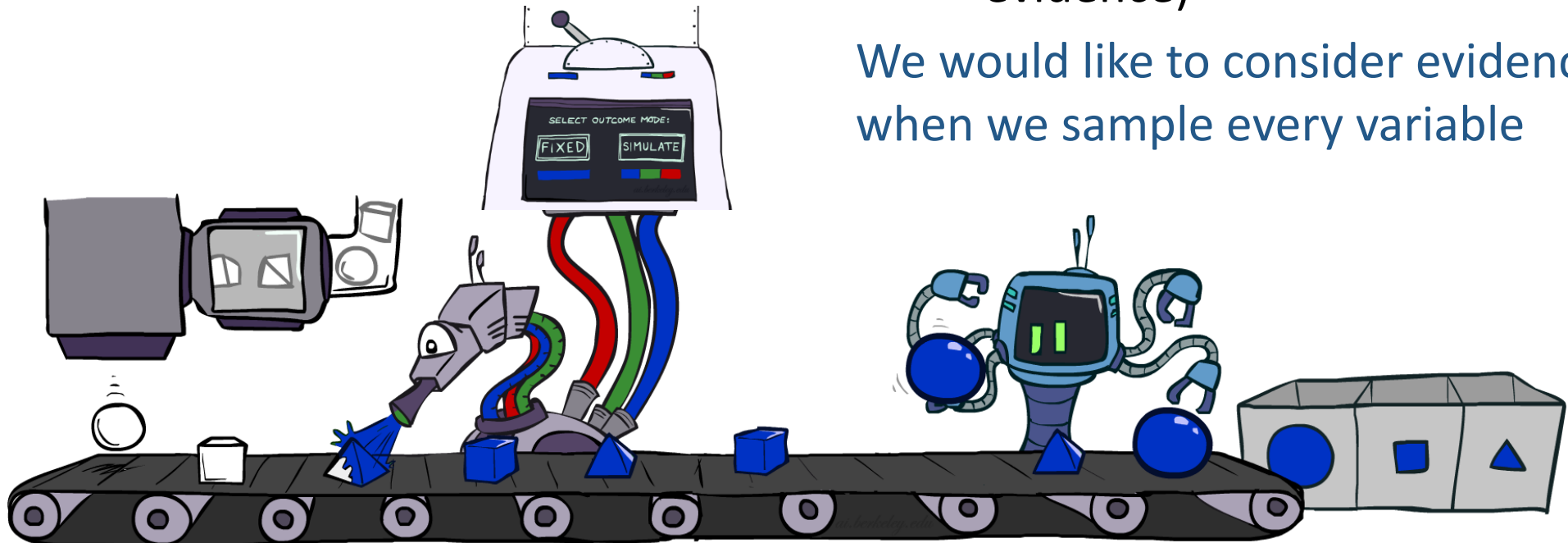
Together, weighted sampling distribution is consistent

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

Likelihood Weighting

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

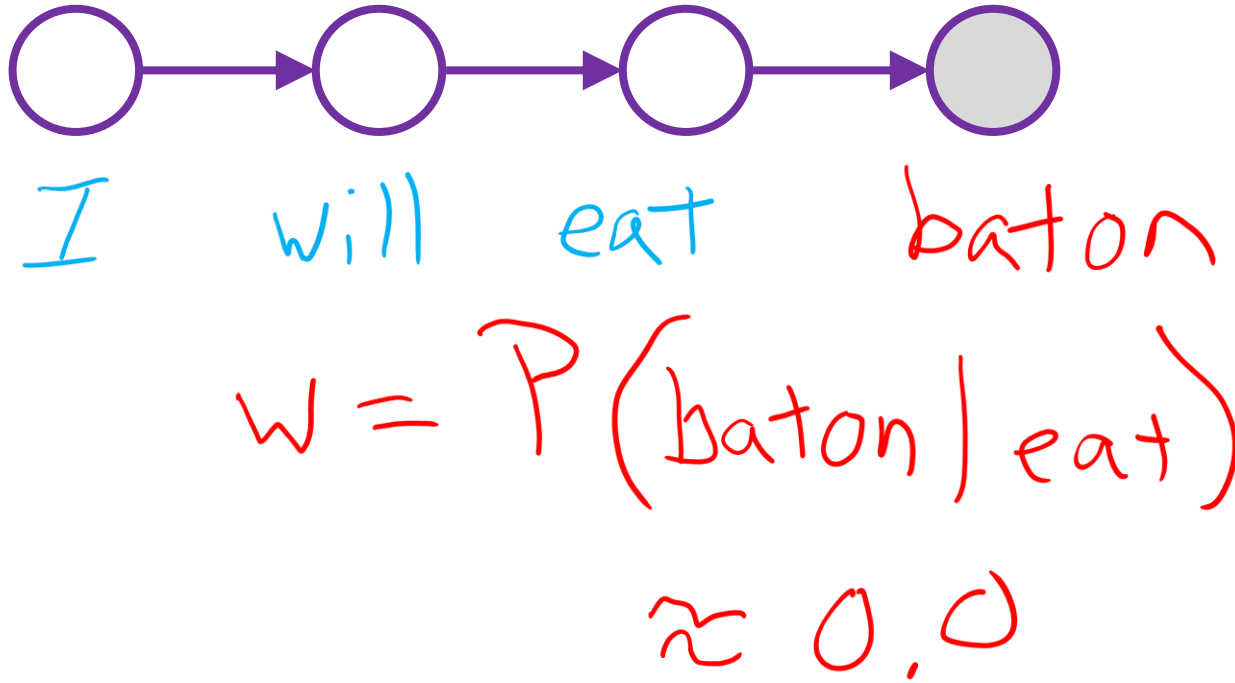


Likelihood weighting doesn't solve all our problems

- 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



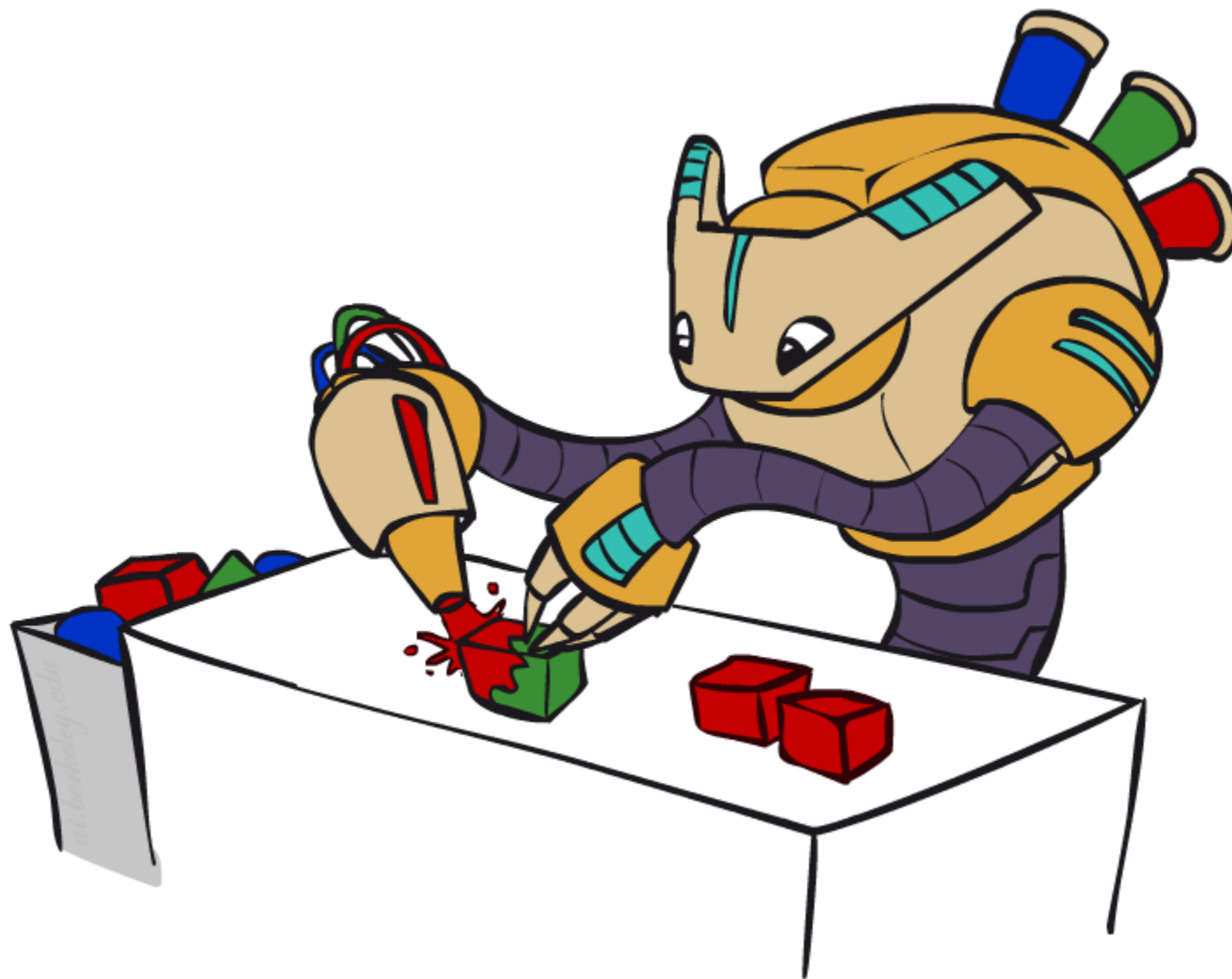
Likelihood weighting doesn't solve all our problems

- 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

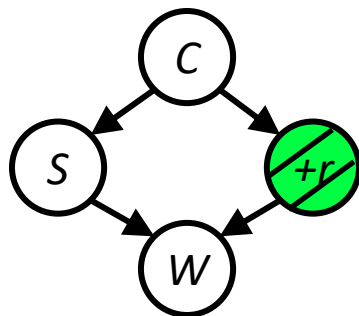
Procedure: keep track of a full instantiation $x_1, x_2, \dots, x_m$.

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.

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

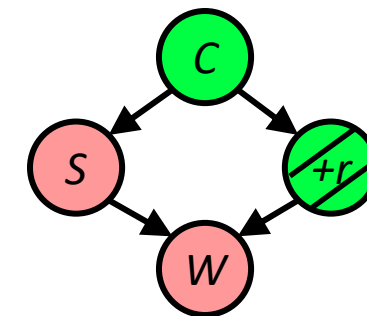
Step 1: Fix evidence

- $R = +r$



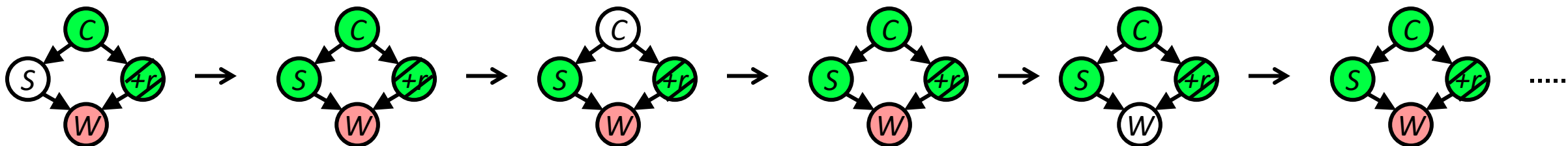
Step 2: Initialize other variables

- Randomly



Steps 3: Repeat

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



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

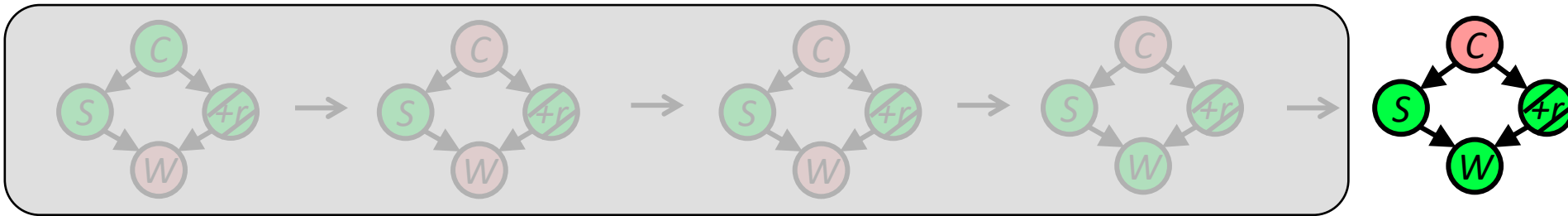
Sample from $P(C \mid +s, -w, +r)$

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

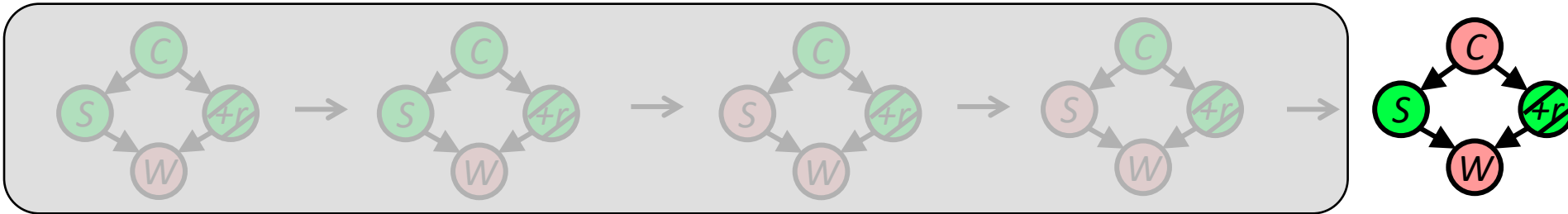
Gibbs Sampling Example: $P(S \mid +r)$

Keep only the last sample from each iteration:

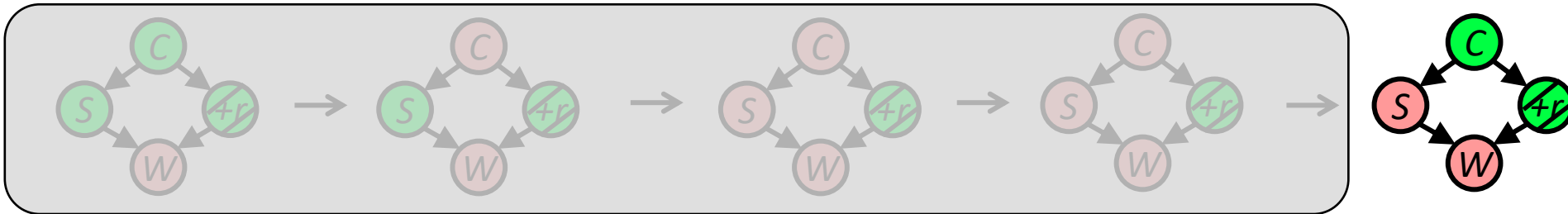
1.



2.



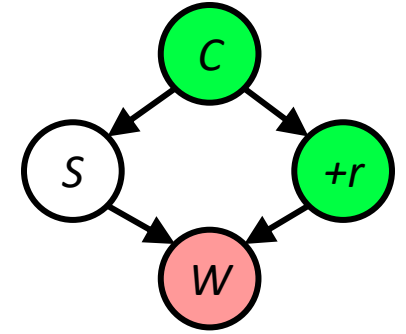
3.



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

Gibbs Sampling

Procedure: keep track of a full instantiation $x_1, x_2, \dots, x_m$.

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.

Further Reading on Gibbs Sampling

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

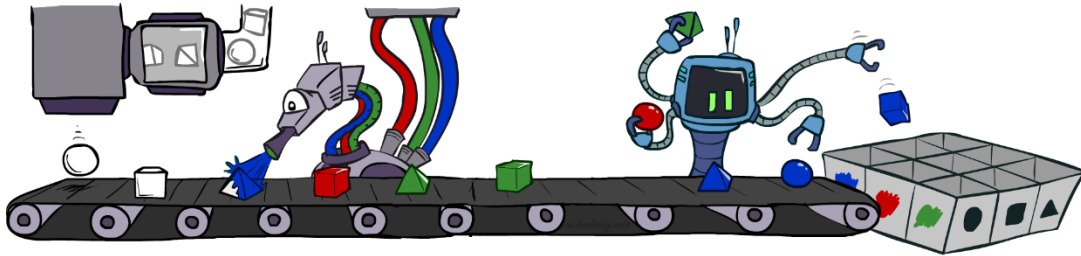
Gibbs sampling is a special case of more general methods called Markov chain Monte Carlo (MCMC) methods

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

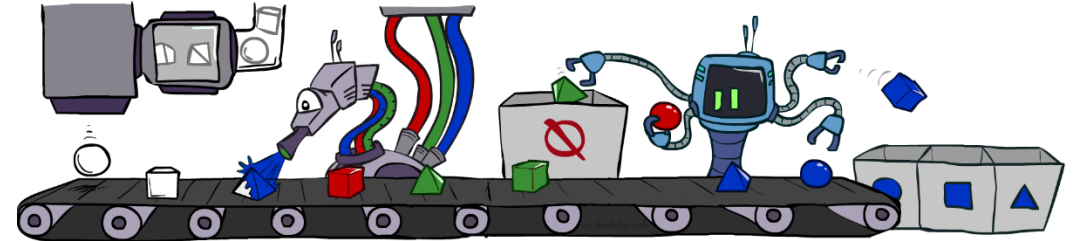
You may read about Monte Carlo methods – they're just sampling

Bayes Net Sampling Summary

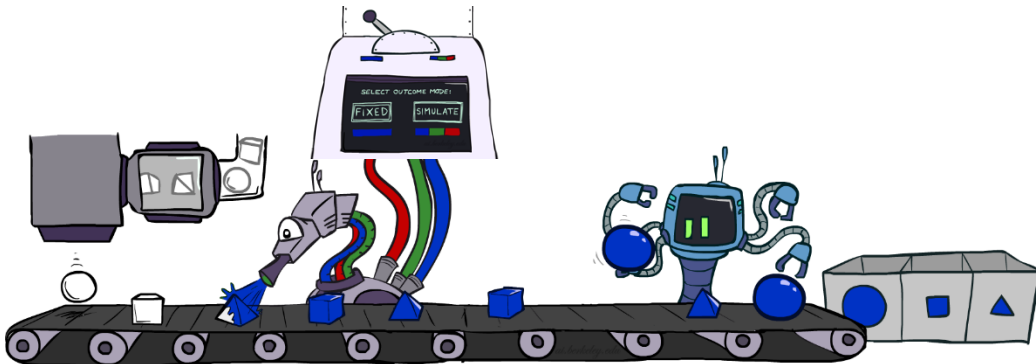
Prior Sampling $P(Q, E)$



Rejection Sampling $P(Q | e)$



Likelihood Weighting $P(Q, e)$



Gibbs Sampling $P(Q | e)$

