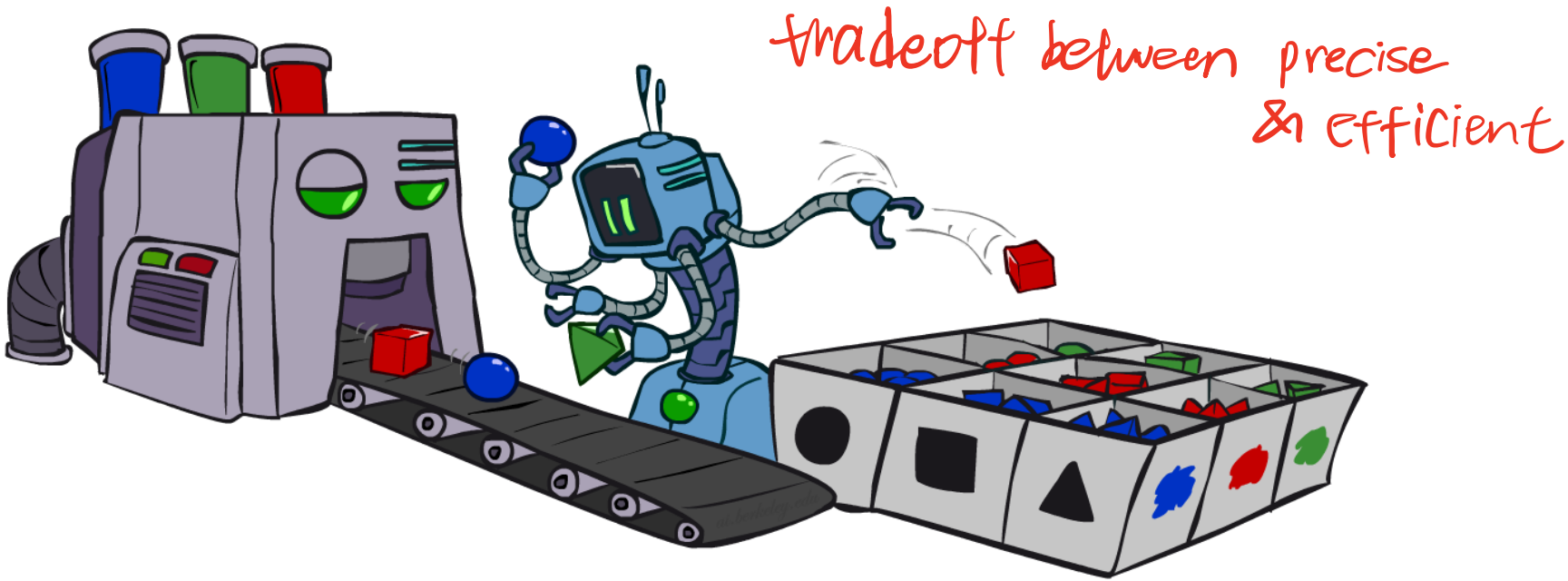
Bayes Nets: Approximate Inference



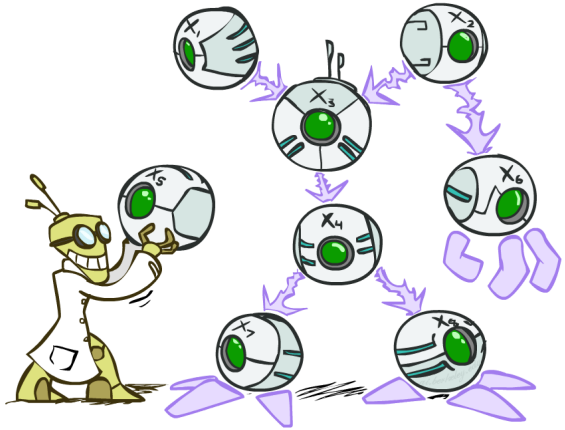
AIMA Chapter 14.5, PRML Chapter 11

[Adapted from slides by Dan Klein and Pieter Abbeel at UC Berkeley]

Bayes' Net Representation

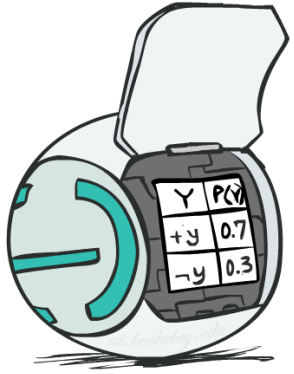
- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over X, one for each combination of parents' values

$$P(X|a_1 \ldots a_n)$$



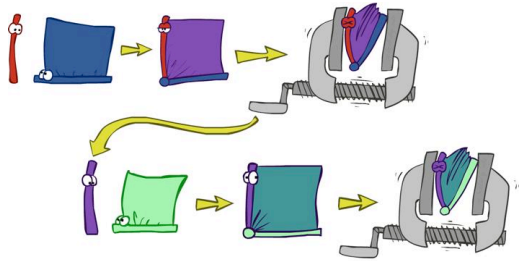
- Bayes' nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

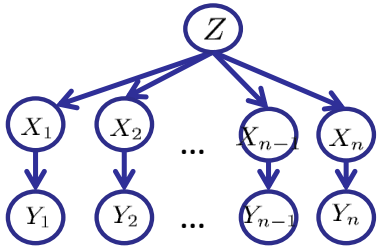
$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|parents(X_i))$$

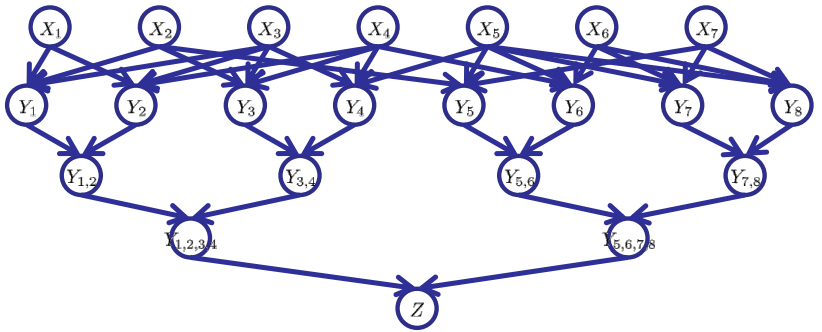


Variable Elimination

- Interleave joining and marginalizing
- d^k entries computed for a factor over k variables with domain sizes d
- Ordering of elimination of hidden variables can affect size of factors generated
- Worst case: running time exponential in the size of the Bayes' net







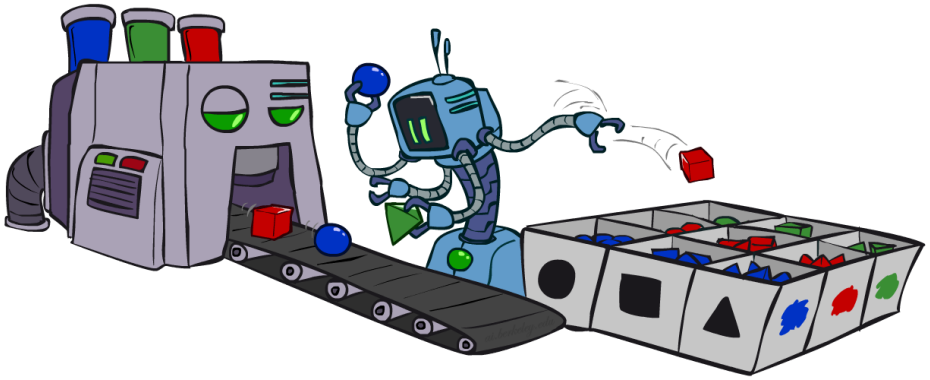
随机采样，蒙特卡洛方法

Sampling

频率以概率

- Goal: probability *P*
- Basic idea
  - Draw *N* samples from a sampling distribution *S*
  - Compute some quantity from the samples
  - Show this converges to the true probability *P*

- Why sample?
  - Often very fast to get a decent approximate answer
  - The algorithms are very simple and general (easy to apply to fancy models)
  - They require very little memory (O(n))



Sampling from a discrete distribution

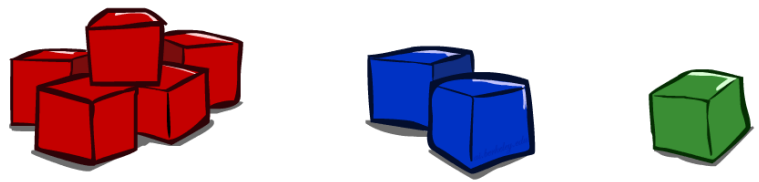
- Sampling from given distribution
  - Step 1: Get sample *u* from uniform distribution over [0, 1)
    - Random() in many programing languages
  - Step 2: Convert this sample *u* into an outcome for the given distribution by associating each outcome *x* with a *P*(*x*)-sized sub-interval of [0,1)

- Example

C	P(C)
red	0.6
green	0.1
blue	0.3

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

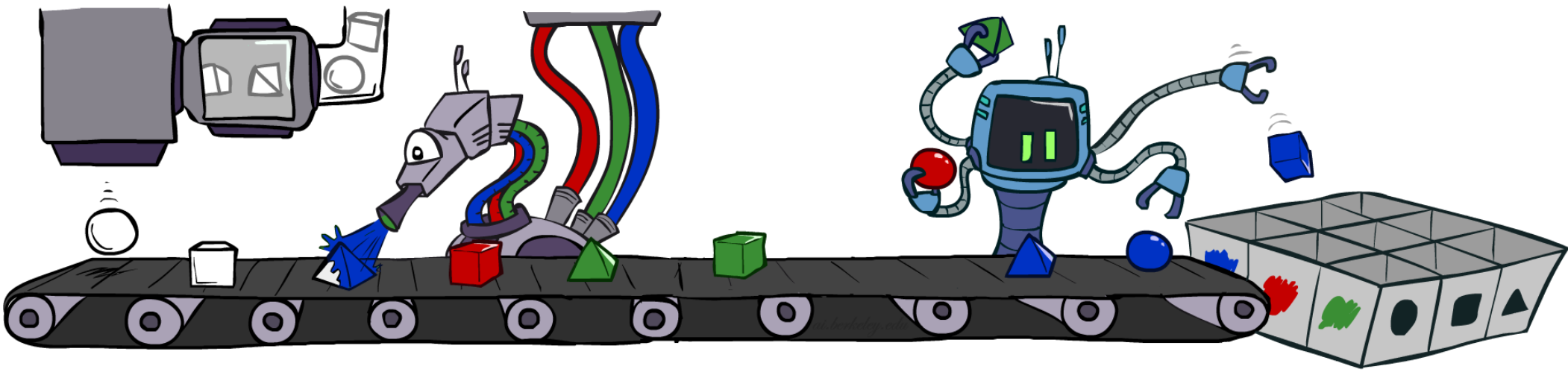
  - If random() returns *u* = 0.83, then our sample is *C* = blue
  - E.g, after sampling 8 times:



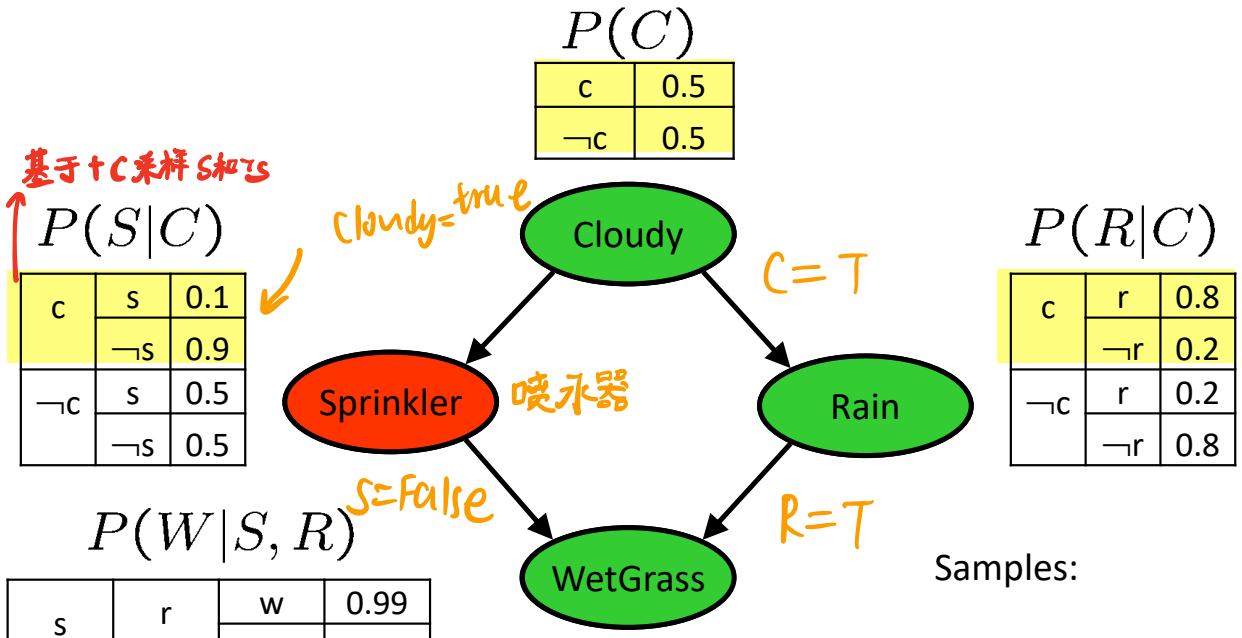
Sampling in Bayes Nets

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

Prior Sampling



Prior Sampling



Samples:

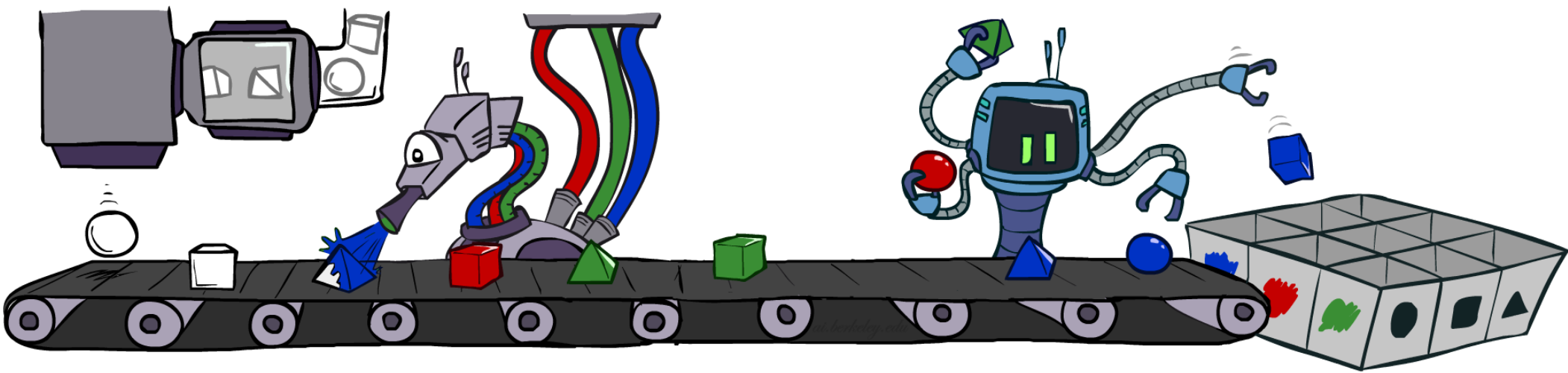
c, ¬s, r, w

¬c, s, ¬r, w

…

Prior Sampling

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



Using samples

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

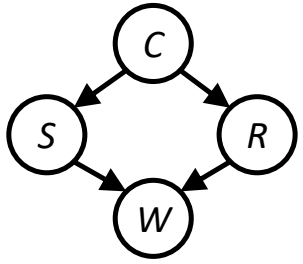
  c, ¬s, r, w

  ¬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
- If we want to know P(*C*| *r, w*)
  - Count (*c, r, w*) and (¬*c, r, w*)
  - Normalize to get P(*C*| *r, w*) = <c:0.67, ¬c:0.33>

Prior Sampling

- This process generates samples with probability:

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

采样概率 生成特定事件的概率

…i.e. the BN's joint probability

特定事件在样本中出现次数

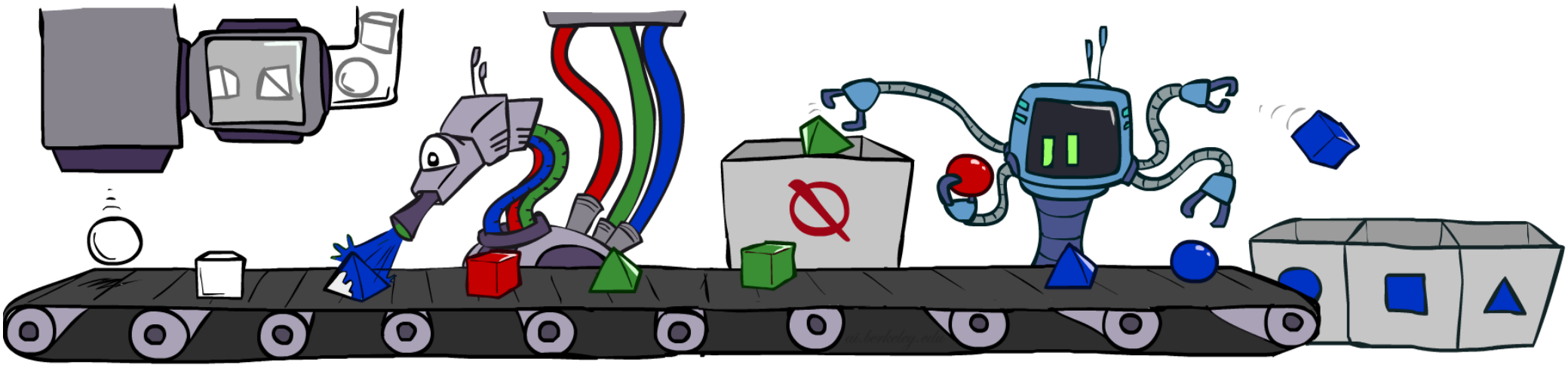
- Let the number of samples of an assignment be $N_{PS}(x_1 \ldots x_n)$
- So $\hat{P}(x_1, \ldots, x_n) = N_{PS}(x_1, \ldots, x_n)/N$
- Then

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

- I.e., the sampling procedure is consistent

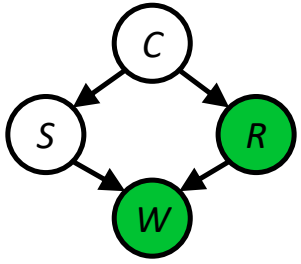
Rejection Sampling 拒绝采样

给定一个易于采样的分布，为一个难于采样的分布生成采样样本的通法



Rejection Sampling

- A simple modification of prior sampling for conditional probabilities
- Let's say we want P(*C*| *r, w*)
- When generating a sample, reject it immediately if not *R*=true, *W*=true
- It is consistent for conditional probabilities (i.e., correct in the limit)



c, ¬s, r, w

~~c, s, ¬r~~

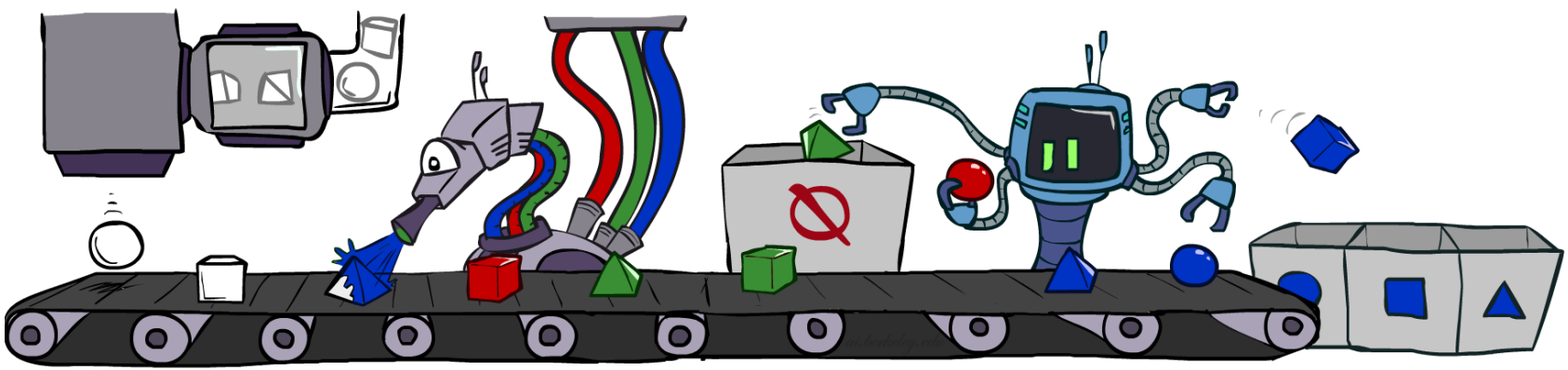
~~¬c, s, r, ¬w~~

~~c, ¬s, ¬r~~

¬c, ¬s, r, w

Rejection Sampling

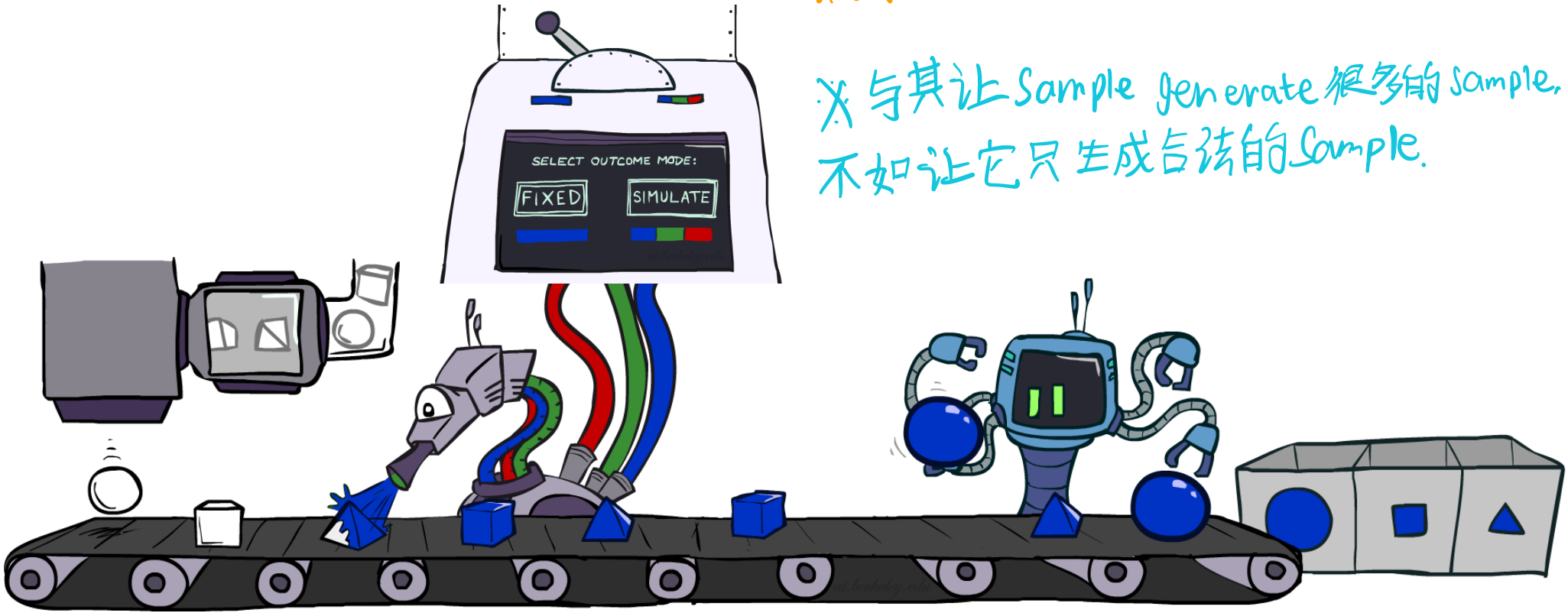
- Input: evidence $e_1,..,e_k$
- For i=1, 2, …, n
  - Sample X_i from $P(X_i \mid parents(X_i))$
  - If x_i not consistent with evidence
    - Reject: Return, and no sample is generated in this cycle
- Return $(x_1, x_2, …, x_n)$



Likelihood Weighting 似然加权

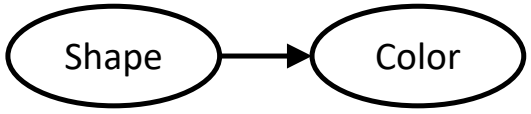
属于重要性采样的一部分

※ 与其让 Sample generate 很多的 sample，不如让它只生成合法的 Sample.

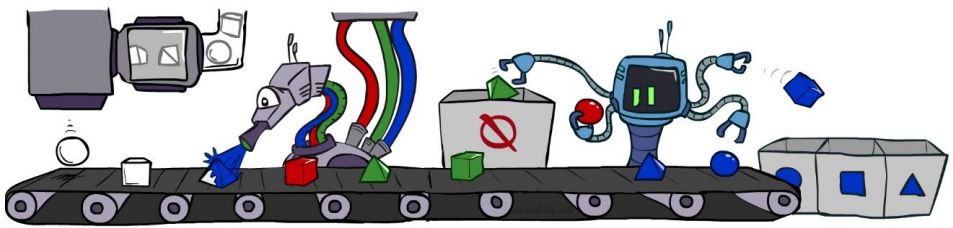


Likelihood Weighting

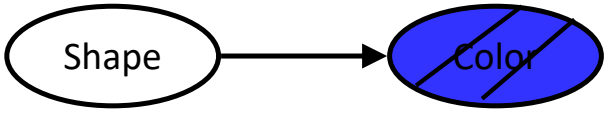
- Problem with rejection sampling:
  - If evidence is unlikely, rejects lots of samples
  - Evidence not exploited as you sample
  - Consider P(*Shape*|*Color=blue*)



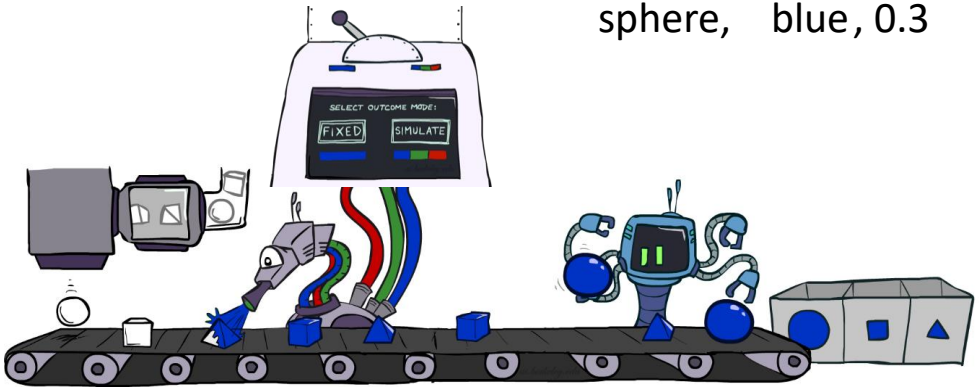
~~pyramid, green~~
~~pyramid, red~~
sphere, blue
~~cube, red~~
~~sphere, green~~



- Idea: fix evidence variables, sample the rest
  - Problem: sample distribution not consistent!
  - Solution: ***weight*** each sample by probability of evidence variables given parents



pyramid, blue , 0.4
pyramid, blue , 0.4
sphere, blue , 0.3
cube, blue , 0.8
sphere, blue , 0.3



Likelihood Weighting

$P(C|+s,+w)$

$P(C)$

c	0.5
¬c	0.5

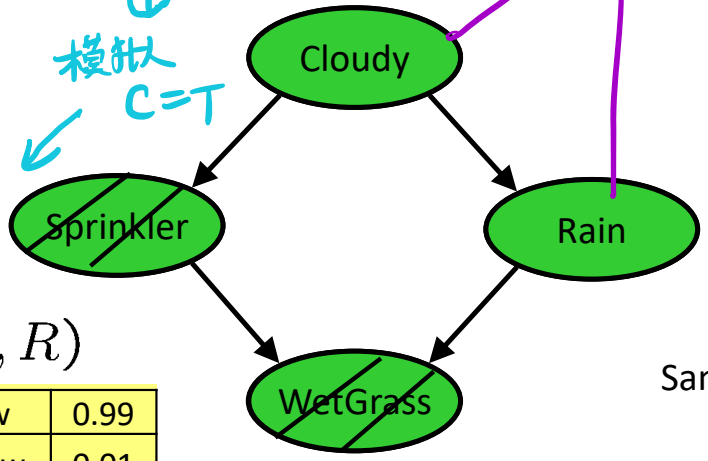
只对非观察的变量进行采样，对观察的常量进行加权.

①模拟 C=T

$P(S|C)$

c	s	0.1
	¬s	0.9
¬c	s	0.5
	¬s	0.5

这里强制让S变成正的。但由于在现实生活中，我们大概率会取到T，所以强制让结果成为S，需要使用权重来规避掉这个问题.



$P(R|C)$

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

...

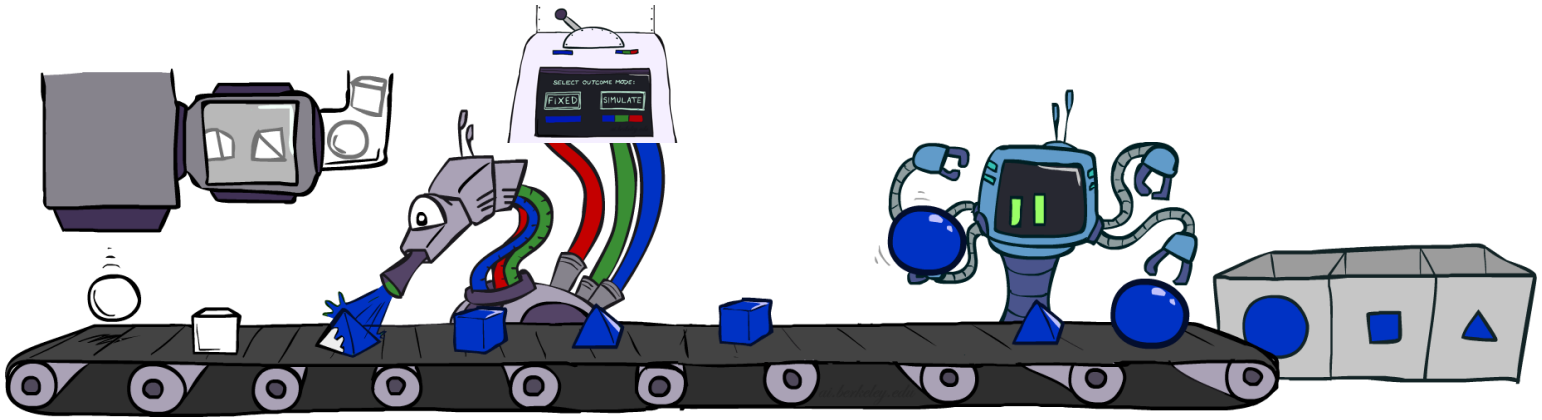
①采样了 C:(T) R:(T) ②加权

$w = 1.0 \times 0.1 \times 0.99$

S:(T)	C:(T)	W(T)	R(T) S(T)
	0.1		0.99

Likelihood Weighting

- Input: evidence $e_1,..,e_k$
- $w = 1.0$
- for i=1, 2, …, n
  - if X_i is an evidence variable 对evidence加权
    - x_i = observed value$_i$ for X_i
    - Set $w = w * P(x_i \mid Parents(X_i))$
  - else 对非evidence采样
    - Sample x_i from $P(X_i \mid Parents(X_i))$
- return $(x_1, x_2, …, x_n)$, w



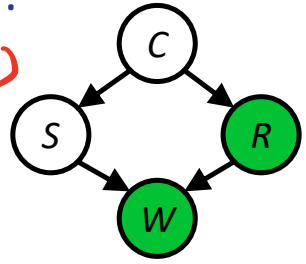
Using samples

0.8 X

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

c, ¬s, r, w	0.1
c, s, r, w	0.2
¬c, s, r, w	0.3
c, ¬s, r, w	0.1
¬c, ¬s, r, w	0.5

w: $P(r|c) \cdot P(w|s,r)$

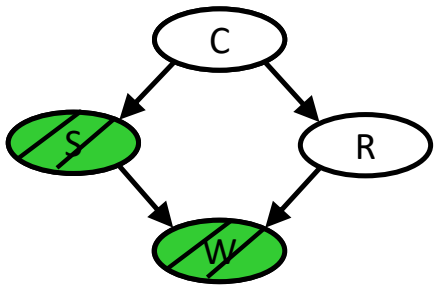


- If we want to know $P(C| r, w)$
  - We have weight sums <(c, r, w): 0.4, $(\neg c, r, w)$: 0.8>
  - Normalize to get $P(C| r, w)$ = <c: 0.33, $\neg c$: 0.67>
  - This will get closer to the true distribution with more samples

Likelihood Weighting

- Sampling distribution (z is sampled and e is fixed evidence)

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



- Now, samples have weights

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

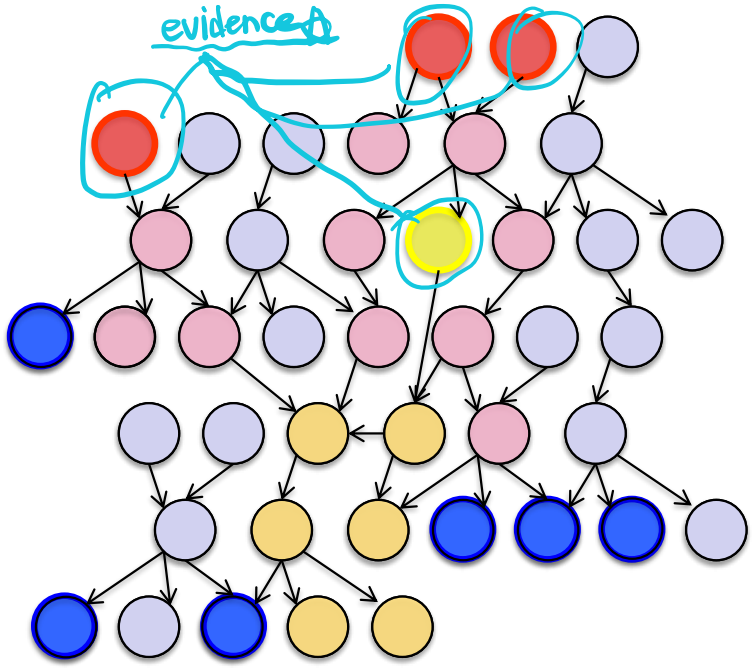
- Together, weighted sampling distribution is consistent

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

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

Likelihood Weighting

- Likelihood weighting is good
  - All samples are used
  - The values of ***downstream*** variables are influenced by ***upstream*** evidence



- Likelihood weighting still has weaknesses
  - The values of ***upstream*** variables are unaffected by ***downstream*** evidence 下游证据不影响上游变量
  - With many downstream evidence, we may
    - mostly get samples that are inconsistent with the evidence and thus have very small weights
    - get a few lucky samples with very large weights, which dominate the result

- We would like each variable to "see" ***all*** the evidence!

问题：①大多数上游变量采出来之后，满足下游所有证据的样本太少。

②即使采出符合证据的样本，这些样本一家独大，影响结果

Example: Fire Alarm

- In likelihood weighting, evidence influences the choice of downstream variables, but not upstream ones
- Example: P(F|+a)

$P(-f|+a) \propto 0.20/1.1$

$P(+f|+a) \propto 0.9/1.1$

P(+f)	0.001
P(-f)	0.999

P(+a|+f)	0.9
P(-a|+f)	0.1

P(+a|-f)	0.01
P(-a|-f)	0.99

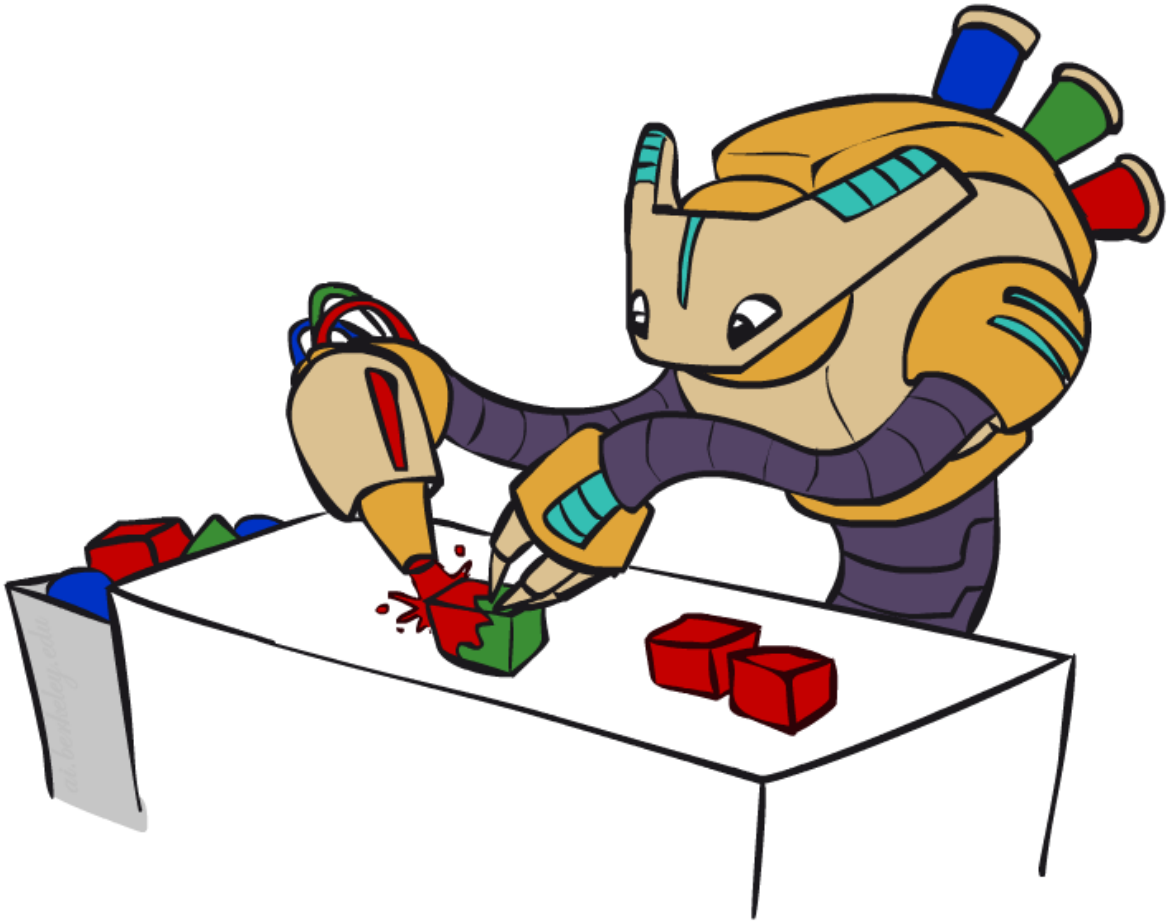
-f +a 0.01
-f +a 0.01
-f +a 0.01
} 20 times

+f +a 0.9 } 1 time

you'll get the likelihood weighting
你会得到似然加权。

现在有一个问题，如果该样本出现的机会特别小，
那么需要等很长很长的时间. 尤其是如果我的evidence在拓扑序靠后.

Gibbs Sampling



之前的采样：都是在父结点的基础上进行采样，如果父结点很深，downstream

Gibbs Sampling

① 生成一个完整样本 (down stream & up stream)

- *Procedure:* keep track of a full instantiation $x_1, x_2, \ldots, x_n$. Start with an arbitrary instantiation consistent with the evidence. Sample one variable at a time, conditioned on all the rest, but keep evidence fixed. Keep repeating this for a long time.
- *Property:* in the limit of repeating this infinitely many times the resulting samples come from the correct distribution (i.e. conditioned on evidence).
- *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 we want high weight.

you pick up a single value, and you resample that value

您选取一个数值，然后重新对该数值进行抽样

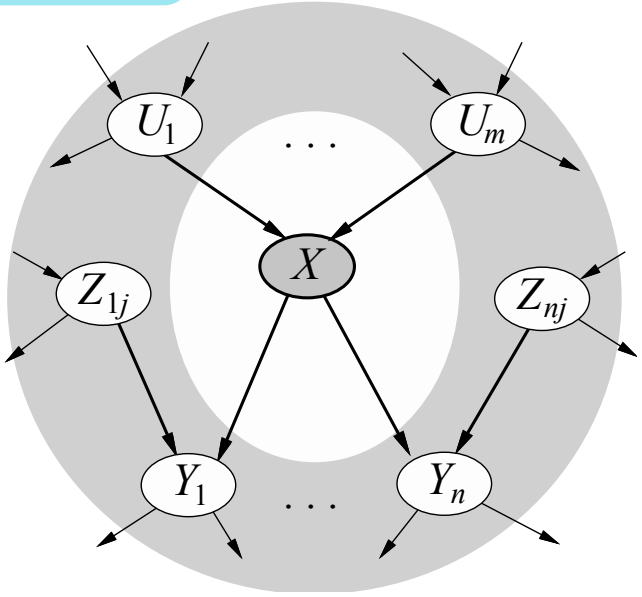
Gibbs Sampling

- Generate each sample by making a random change to the preceding sample
  - Evidence variables remain fixed. For each of the non-evidence variable, sample its value conditioned on all the other variables
  - $X_i' \sim P(X_i \mid x_1,..,x_{i-1},x_{i+1},..,x_n)$
  - In a Bayes net

    $P(X_i \mid x_1,..,x_{i-1},x_{i+1},..,x_n)$

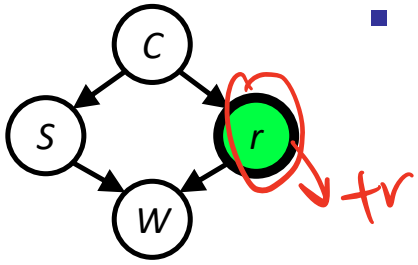
    $= P(X_i \mid markov\_blanket(X_i))$

    $= \alpha\ P(X_i \mid u_1,..,u_m)\ \prod_j P(y_j \mid parents(Y_j))$

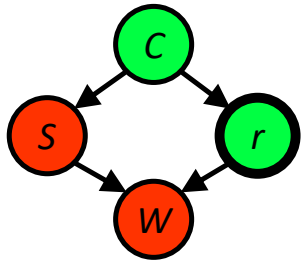


Gibbs Sampling Example: $P(S \mid r)$

- Step 1: Fix evidence
  - R = true

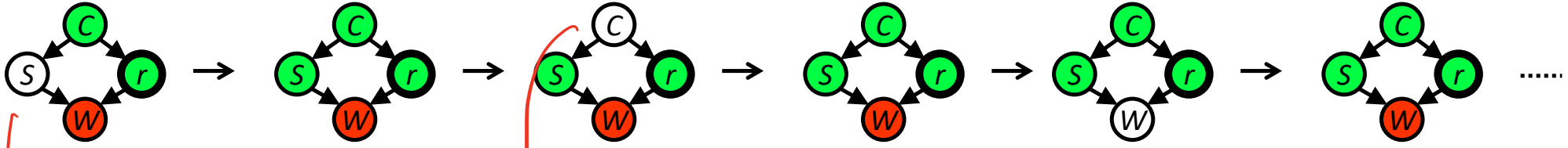


- Step 2: Initialize other variables
  - Randomly



- Step 3: Repeat
  - Choose an arbitrary non-evidence variable X
  - Resample X from $P(X \mid markov\_blanket(X))$

所有的样本都互相关联了.



Sample $S \sim P(S \mid c, r, \neg w)$

Choose S

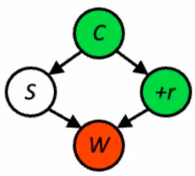
Sample $C \sim P(C \mid s, r)$

Choose C

Sample $W \sim P(W \mid s, r)$

Efficient Resampling of One Variable

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



$$P(S|+c,+r,-w) = \frac{P(S,+c,+r,-w)}{P(+c,+r,-w)}$$ def. of conditional probability

$$= \frac{P(S,+c,+r,-w)}{\sum_s P(s,+c,+r,-w)}$$ introduce summation

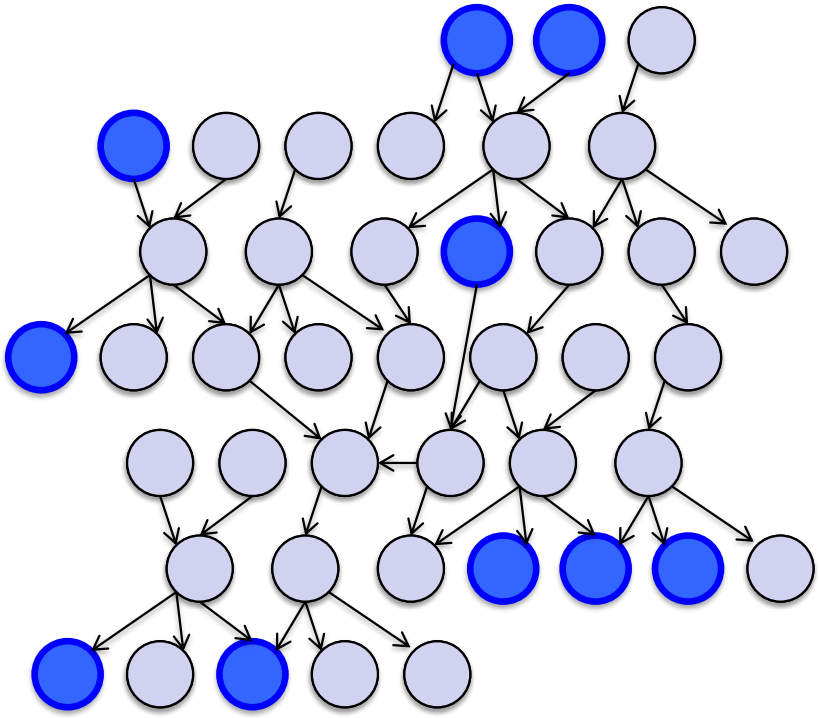
$$= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{\sum_s P(+c)P(s|+c)P(+r|+c)P(-w|s,+r)}$$ def. of Bayes' Nets

$$= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{P(+c)P(+r|+c)\sum_s P(s|+c)P(-w|s,+r)}$$ move summation term out

$$= \frac{P(S|+c)P(-w|S,+r)}{\sum_s P(s|+c)P(-w|s,+r)}$$ cancel out terms

So I'll walk you through it.
那我来带你一步步了解。

Why doing this?



- Samples soon begin to reflect all the evidence in the network
- Eventually they are being drawn from the true posterior!

- Theorem: Gibbs sampling is consistent

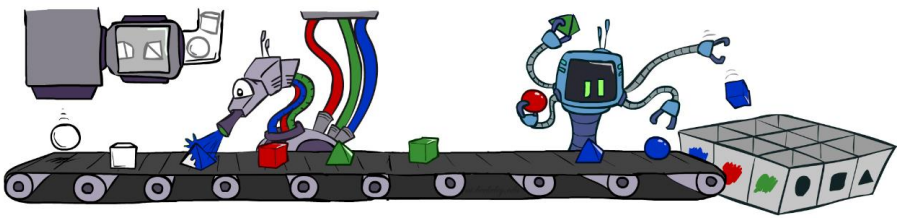
- MCM … ing som
  - M … walk”), w
  - M … ino

Markov Chain Monte Carlo (MCMC)

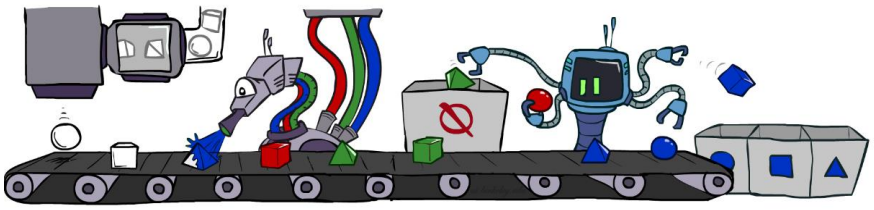
- MCMC is a family of randomized algorithms for approximating some quantity of interest over a very large state space
  - Markov chain = a sequence of randomly chosen states ("random walk"), where each state is chosen conditioned on the previous state
  - ~~Monte Carlo = a very expensive city in Monaco with a famous casino~~
  - Monte Carlo = an algorithm (usually based on sampling) that is likely to find a correct answer
- MCMC = sampling by constructing a Markov chain
- Gibbs, Metropolis-Hastings, Hamiltonian, Slice, etc.

Summary

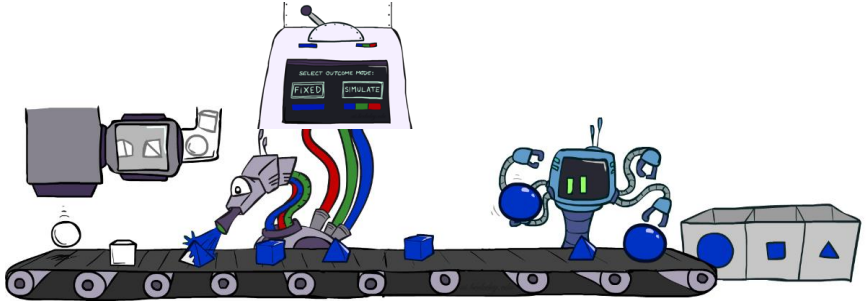
- Prior Sampling P



- Rejection Sampling P(Q | e)



- Likelihood Weighting P(Q | e)



- Gibbs Sampling P(Q | e)