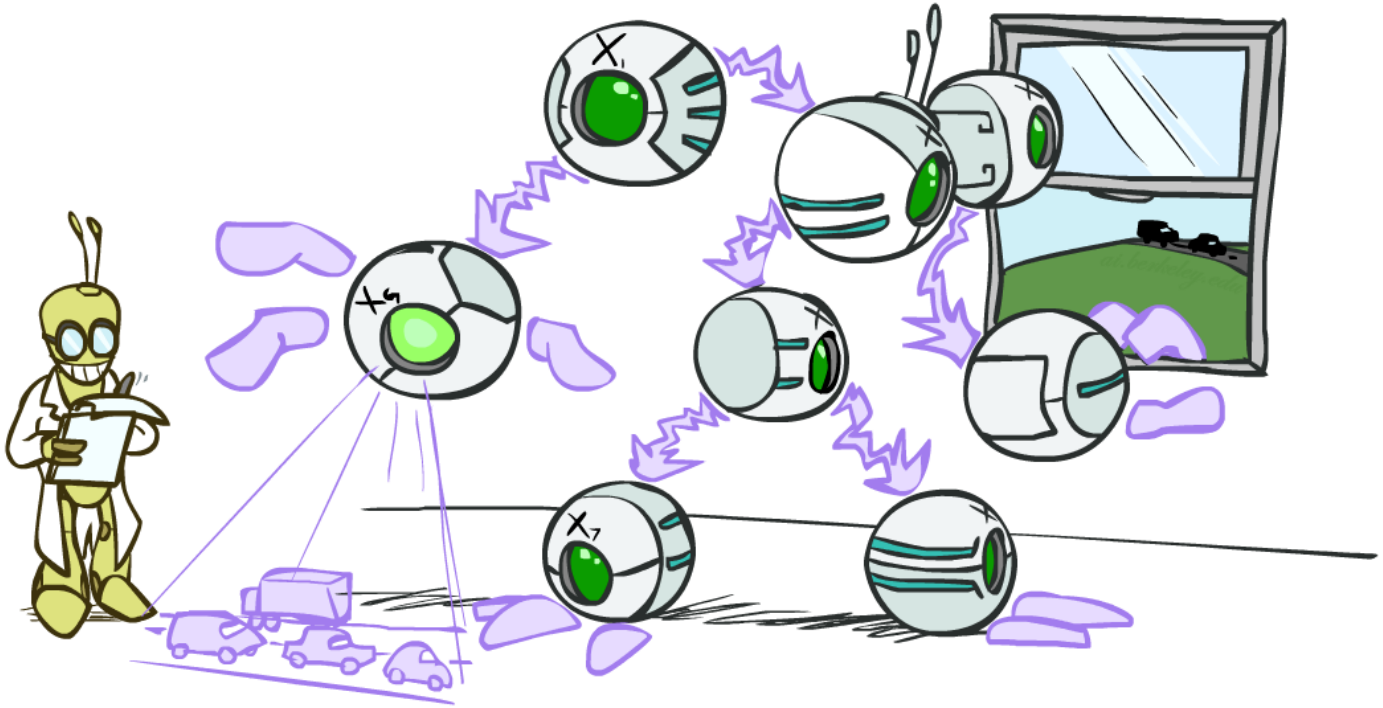
07-380 AI & ML II

Bayes Nets



Instructors: Pat Virtue & Mohammad Salameh

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

A) Log Reg (Neural nets)

SGD

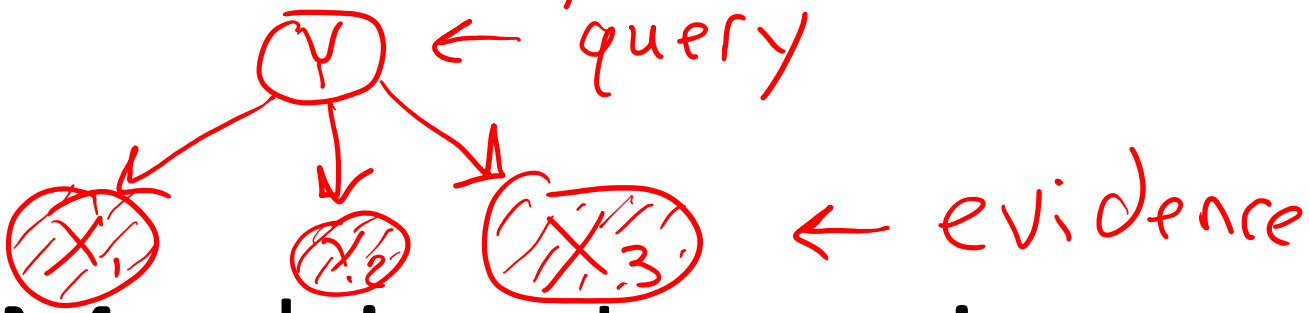
Given $\vec{x}$

$\rightarrow f(\vec{x})$

B) Naive Bayes

MLE

Given $\vec{x}$



$f(\vec{x}) : p(y \mid \vec{x})$

$: p(y)\,p(\vec{x} \mid y)$

Machine Learning

Select Model → Train Model → Use Model (Inference)

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

$P(\text{query} \mid \text{evidence})$

Bayes Nets

Representation and Independence

(Training?)

Exact Inference

- Enumeration (always exponential complexity)
- Some shortcuts (worst-case exponential complexity, often better)
- Inference is NP-hard in general

Approximate Inference

- Prior sampling
- Rejection sampling
- Likelihood-weighted sampling
- Gibbs sampling

Graphical Model Notation

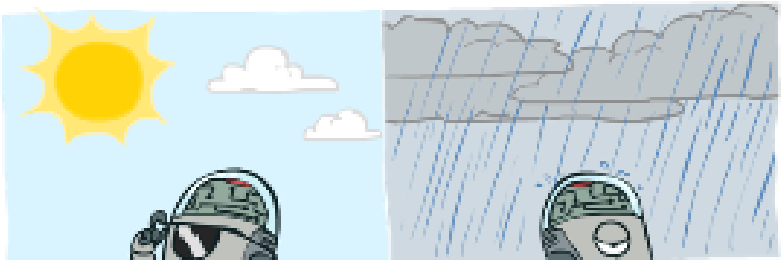
Nodes: variables (with domains)

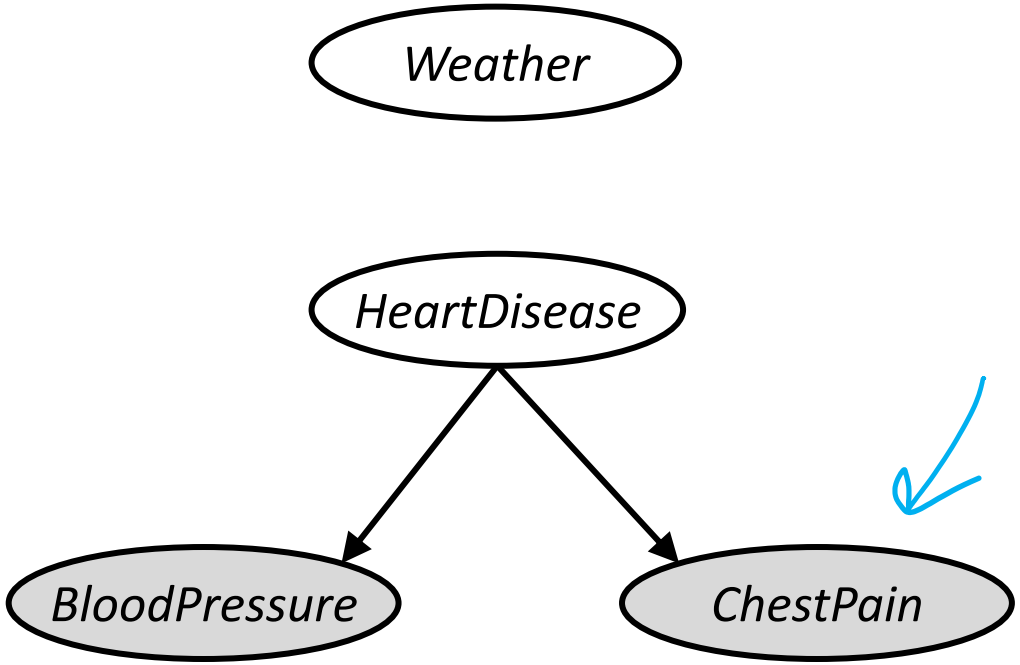
- Can be assigned (observed) or unassigned (unobserved)
- We'll shade nodes to indicate observed variables
- Observed **does not** mean $Variable = true$
  Observed just means that at inference time, we will have the value for that variable

Edges

- Indicate "direct influence" between variables
- Absence of edges: encode conditional independence

For now: imagine that arrows mean direct causation (in general, they don't!)





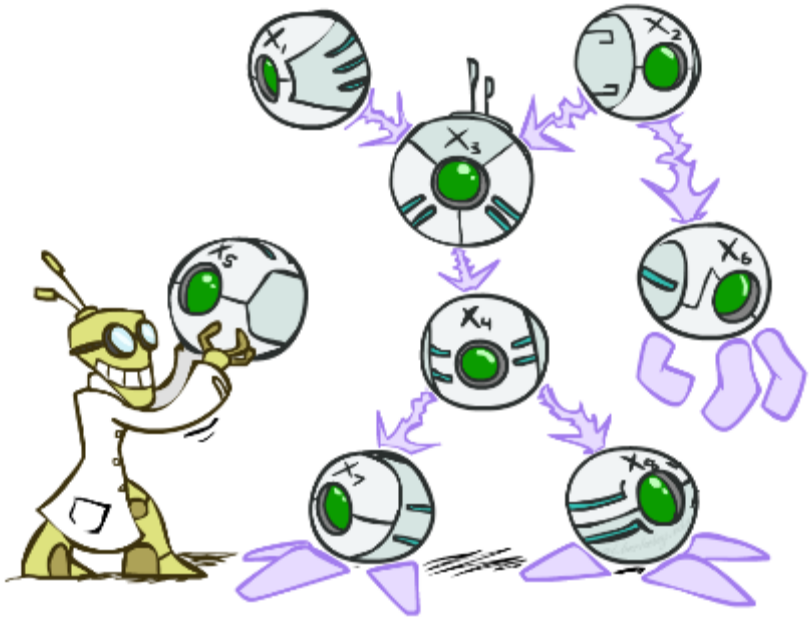
Bayes' Nets: Big Picture

Two problems with using full joint distribution tables as our probabilistic models:

- Usually, the joint is WAY too big to represent explicitly
- Hard to learn (estimate) anything empirically about more than a few variables at a time

Bayes nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)

- A type of probabilistic graphical models
- We describe how variables locally interact
- Local interactions chain together to give global, indirect interactions

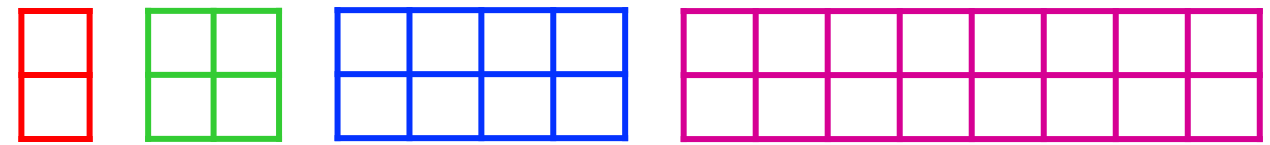


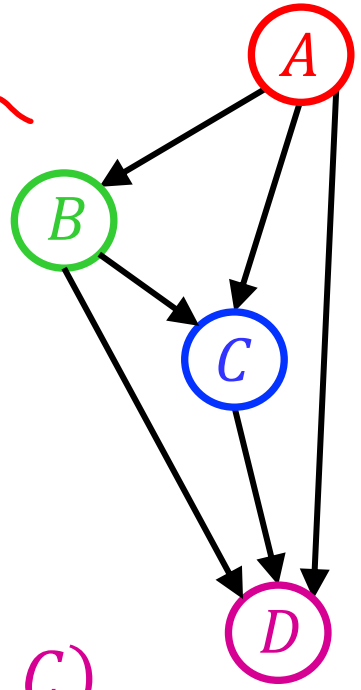
Bayesian Networks

Bayes net

Example of full chain rule Bayes net

are parameters of Categorical distribution





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

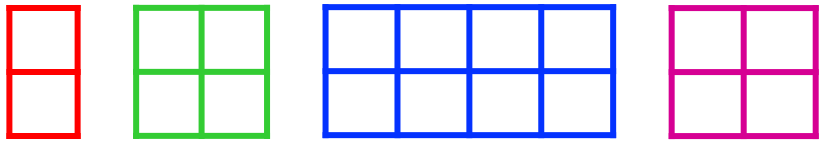
Encode joint distributions as product of conditional distributions on each variable

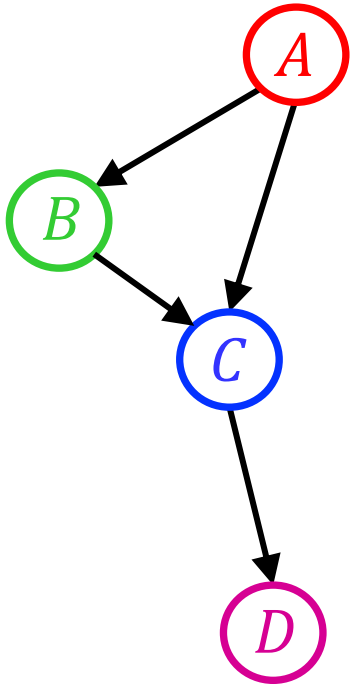
$$P(X_1,\ldots,X_N) = \prod_i P(X_i|\,Parents(X_i))$$

Bayesian Networks

Bayes net

Missing edges (compared to full chain rule) encode our independence assumptions



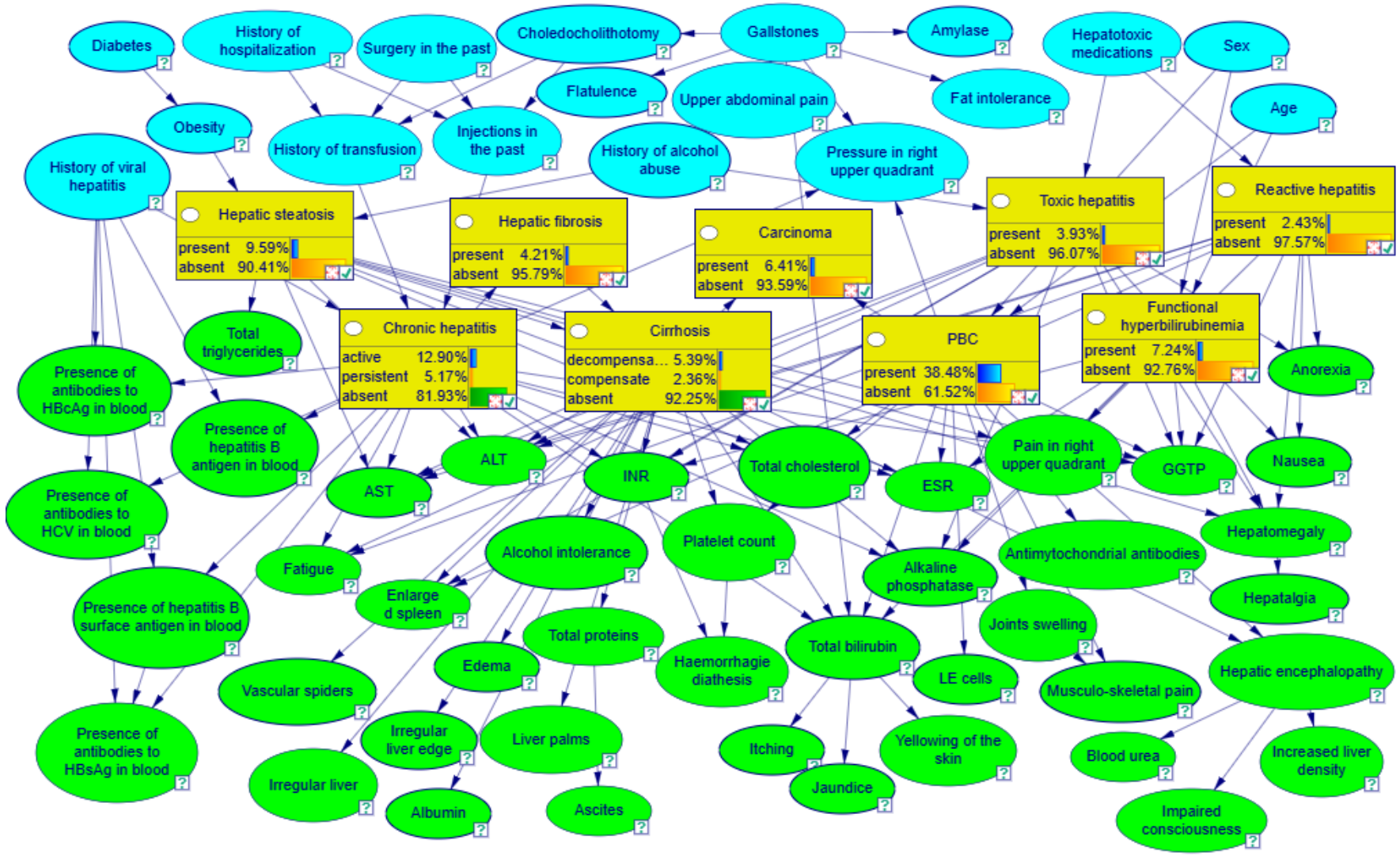


$$P(A,B,C,D) = P(A)\,P(B|A)\,P(C|A,B)\,P(D|C)$$

Encode joint distributions as product of conditional distributions on each variable

$$P(X_1,\ldots,X_N) = \prod_i P(X_i|\,Parents(X_i))$$

Example: Liver Disorders



https://demo.bayesfusion.com/bayesbox.html

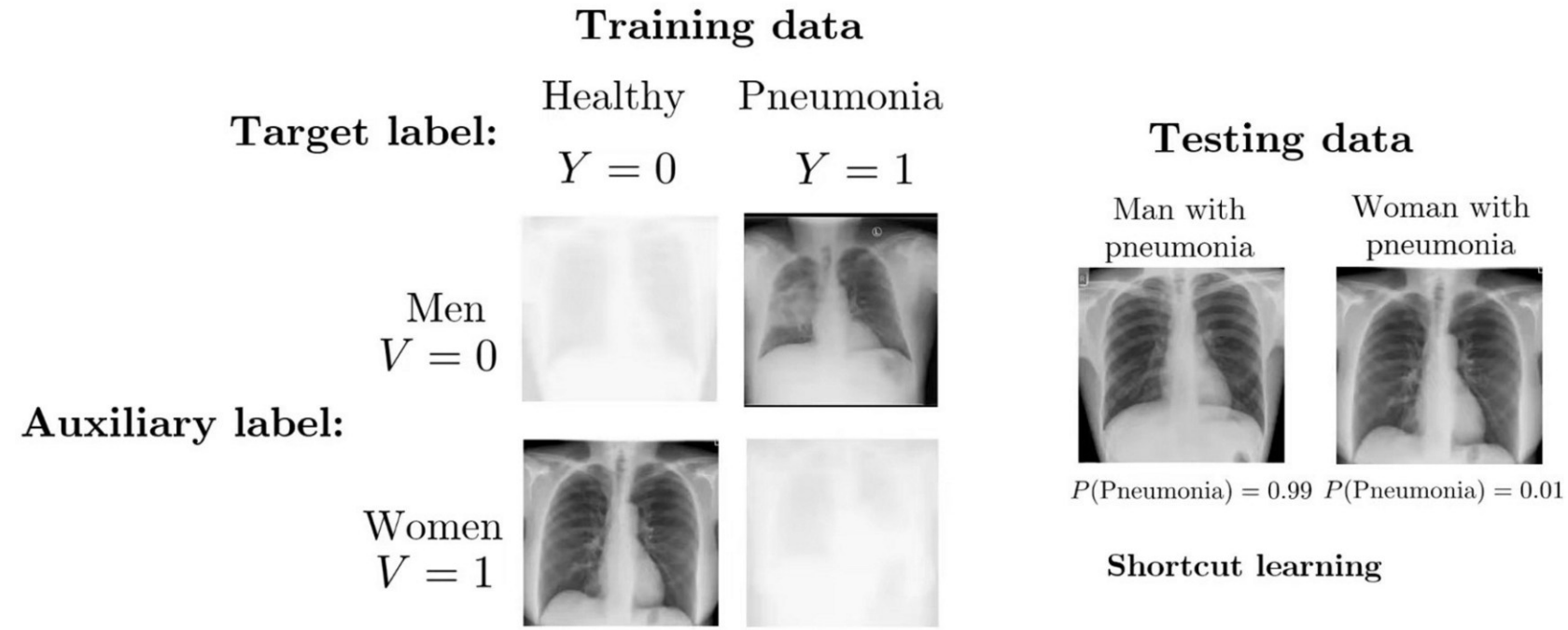
Maggie Makar, University of Michigan

Assistant Professor
Computer Science and Engineering
University of Michigan

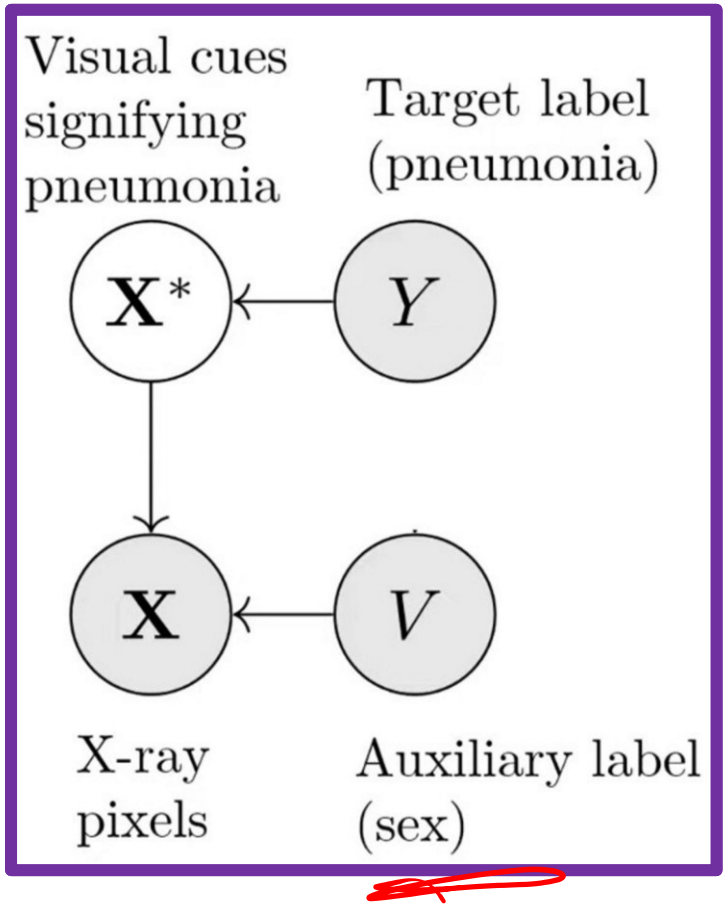
https://mymakar.github.io/

Causally-motivated shortcut removal using auxiliary labels
M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, A. D'Amour
AIStats 2022 [paper]

Pneumonia detection under biased sampling



Cases courtesy of Dr. Andrew Dixon, Dr Henry Knip, Dr. Usman Bashir, and Dr. Ian Bickle , Radiopaedia.org, rID: 48366, 31388, 18394 and rID: 50318



Causally-motivated shortcut removal using auxiliary labels

M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, A. D'Amour

AIStats 2022 [paper]

Danielle Belgrave, Microsoft Research

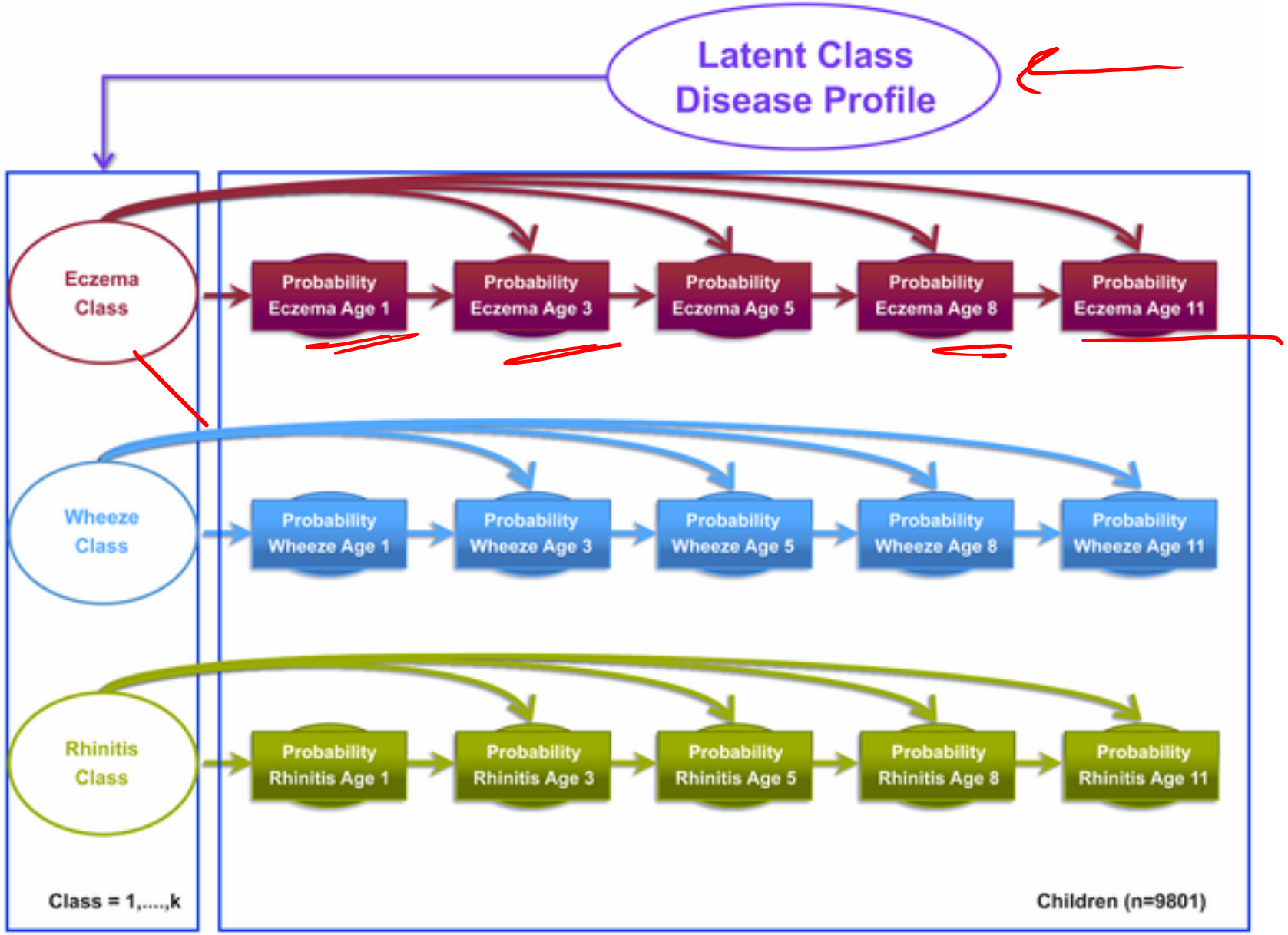
Deepmind

GSK

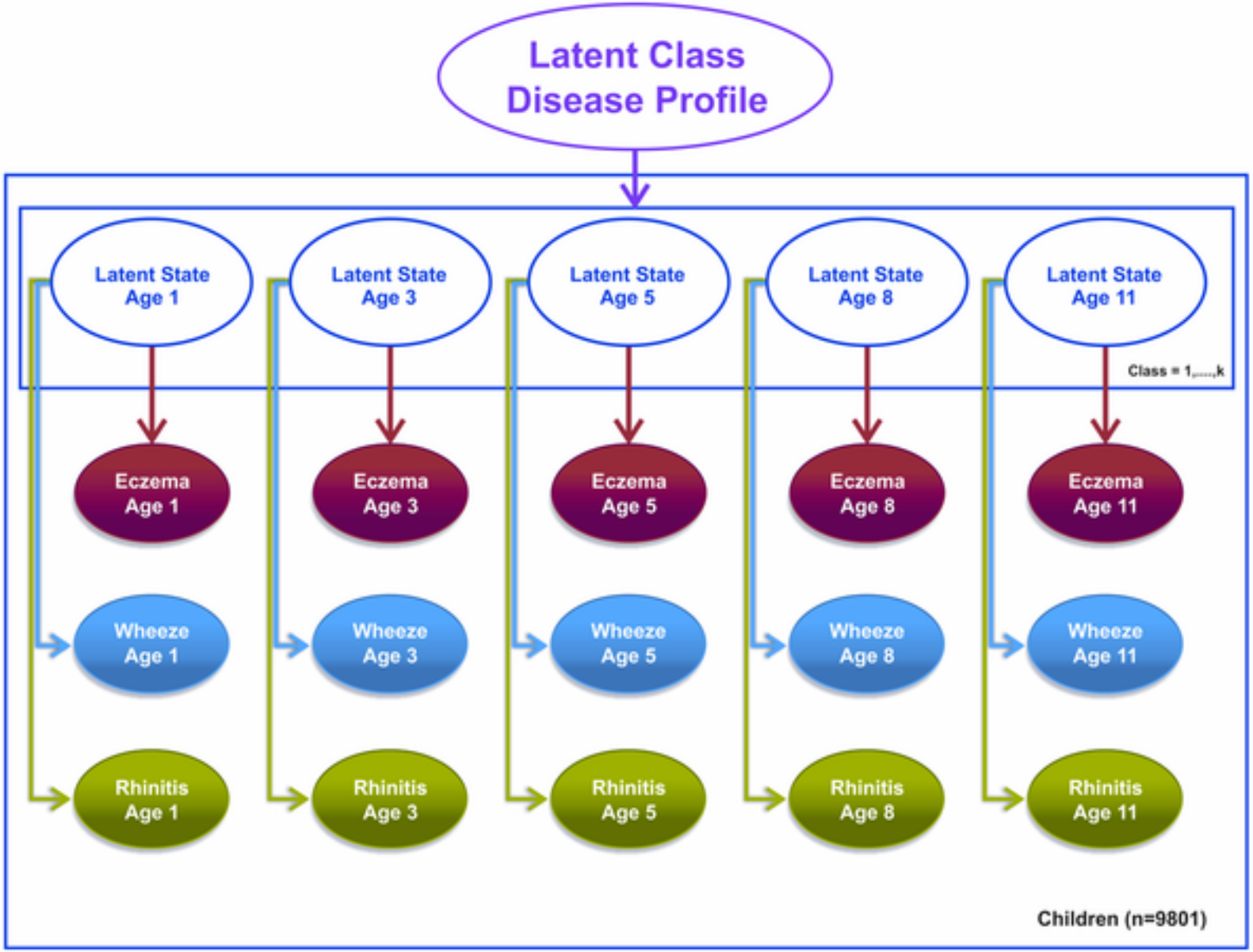
Vice President
AI/ML
GSK

https://www.daniellebelgrave.com

Developmental Profiles of Eczema, Wheeze, and Rhinitis:
Two Population-Based Birth Cohort Studies
Danielle Belgrave, et al. *PLOS Medicine*, 2014
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001748



Danielle Belgrave, et al. *PLOS Medicine*, 2014
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001748



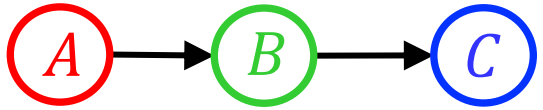
Danielle Belgrave, et al. *PLOS Medicine*, 2014
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001748

Bayes Nets Independence

Conditional Independence Semantics

For the following Bayes nets, write the joint $P(A, B, C)$

1. Using the chain rule (with top-down order A,B,C)
2. Using Bayes net semantics (product of CPTs)



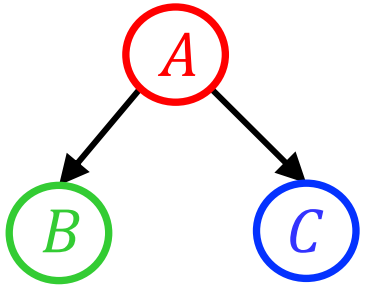
$P(A)\ P(B|A)\ P(C|A,B)$

$P(A)\ P(B|A)\ P(C|B)$

Assumption:

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

C is independent from A given B



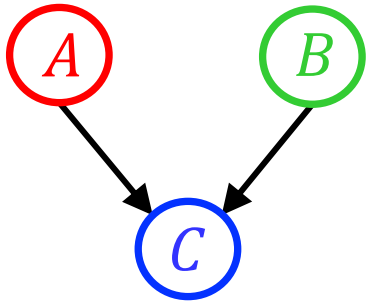
$P(A)\ P(B|A)\ P(C|A,B)$

$P(A)\ P(B|A)\ P(C|A)$

Assumption:

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

C is independent from B given A



$P(A)\ P(B|A)\ P(C|A,B)$

$P(A)\ P(B)\ P(C|A,B)$

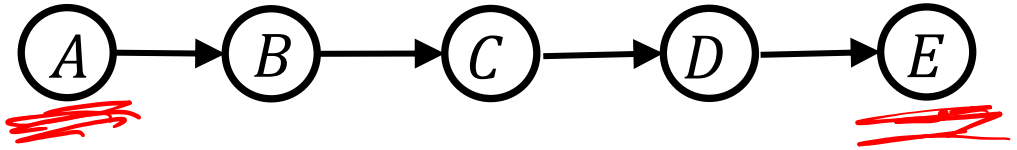
Assumption:

$P(B|A) = P(B)$

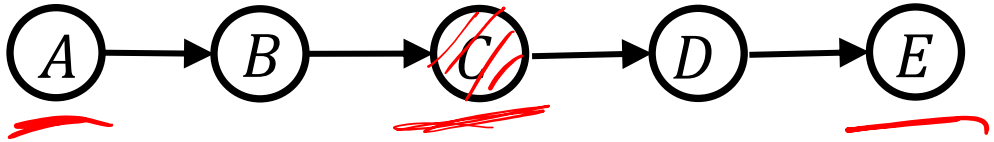
A is independent from B given { }

Answering Independence Questions

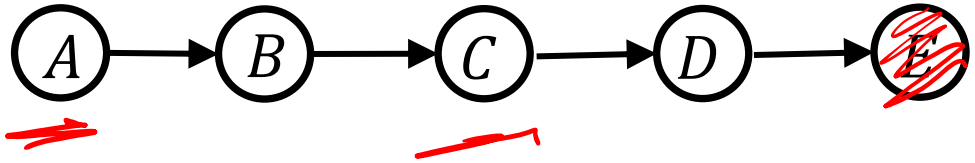
- Is A independent from E? No $A \perp\!\!\!\perp E$



- Is A independent from E given C? Yes



- Is A independent from C given E?



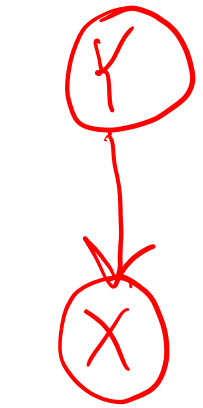
Leveraging Independence from a Bayes Net

Is X_2 independent from X_3 given X_1 and X_6?

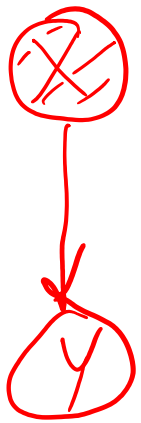
$P(X_2 \mid X_1, X_6)$

$= P(X_2 \mid X_1, X_6, X_3)$

Gen



Disc

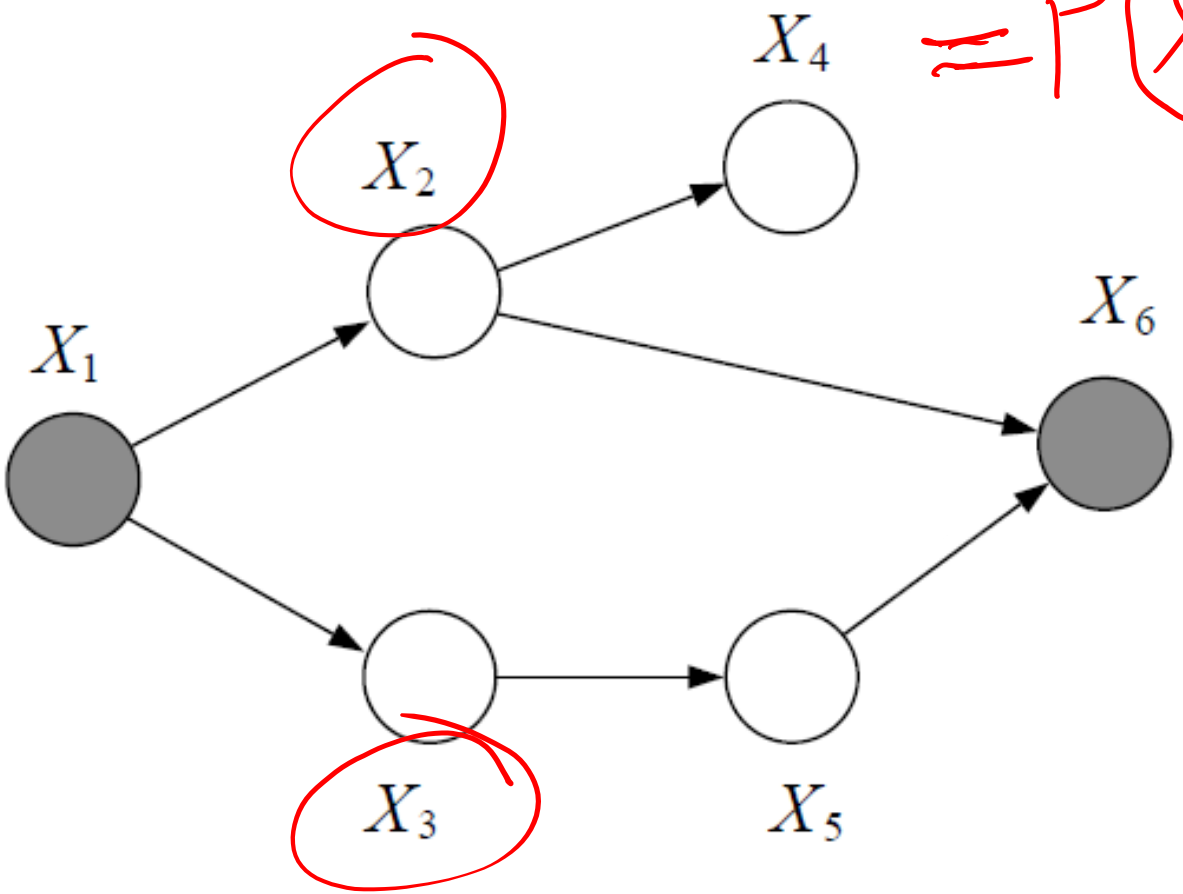


$P(y,x)$

$P(x \mid y)$

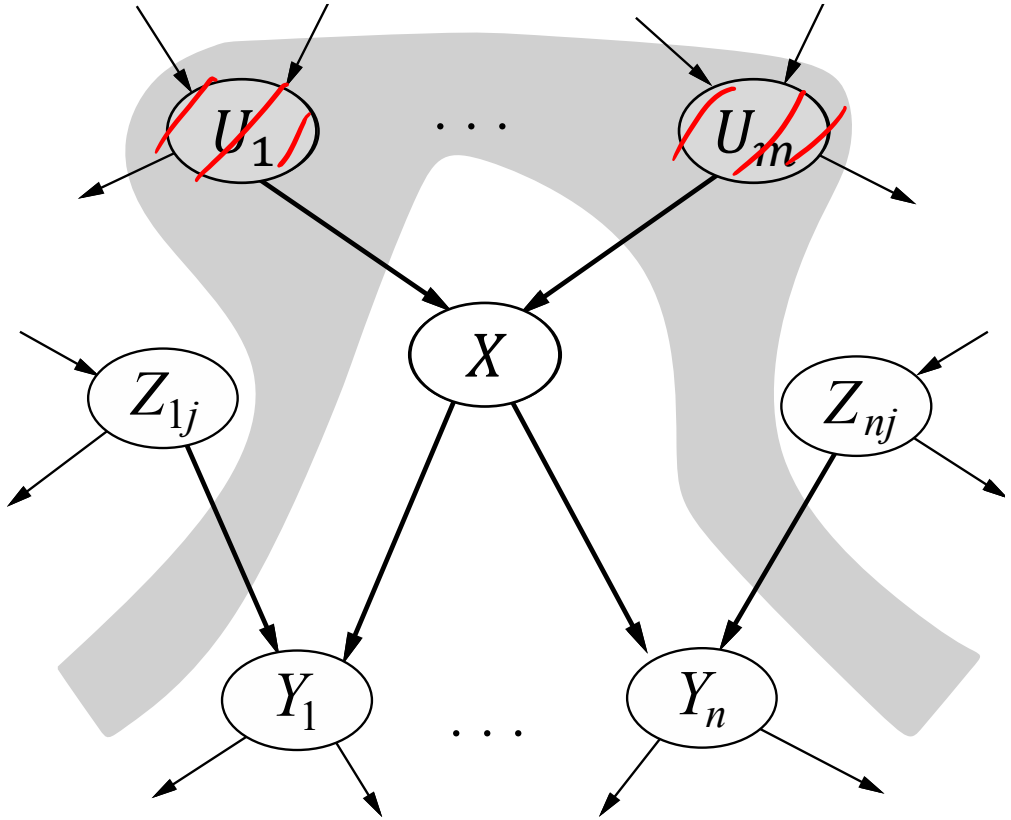
$P(y \mid x)$

$P(y \mid x)$



Conditional Independence Semantics

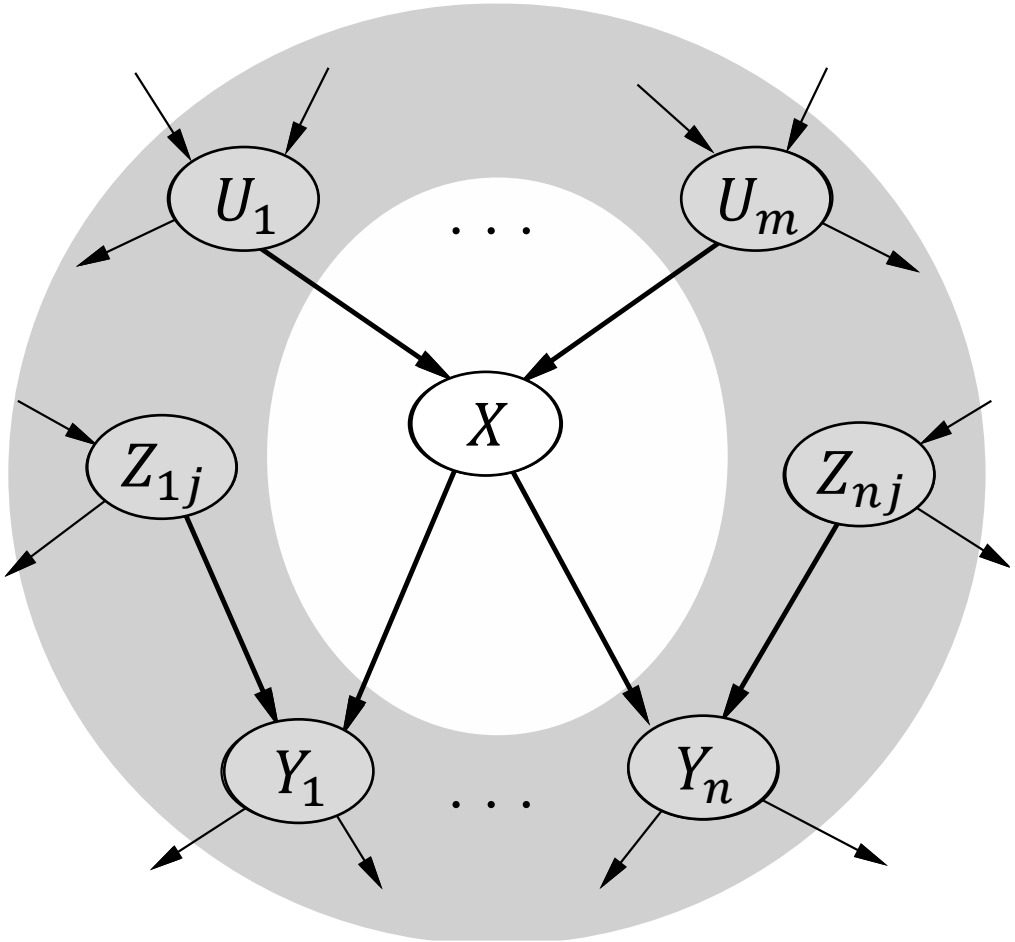
Every variable is conditionally independent of its non-descendants given its parents



Markov blanket

A variable's Markov blanket consists of parents, children, children's other parents

Every variable is conditionally independent of all other variables given its Markov blanket



Machine Learning

Select Model → Train Model → Use Model (Inference)

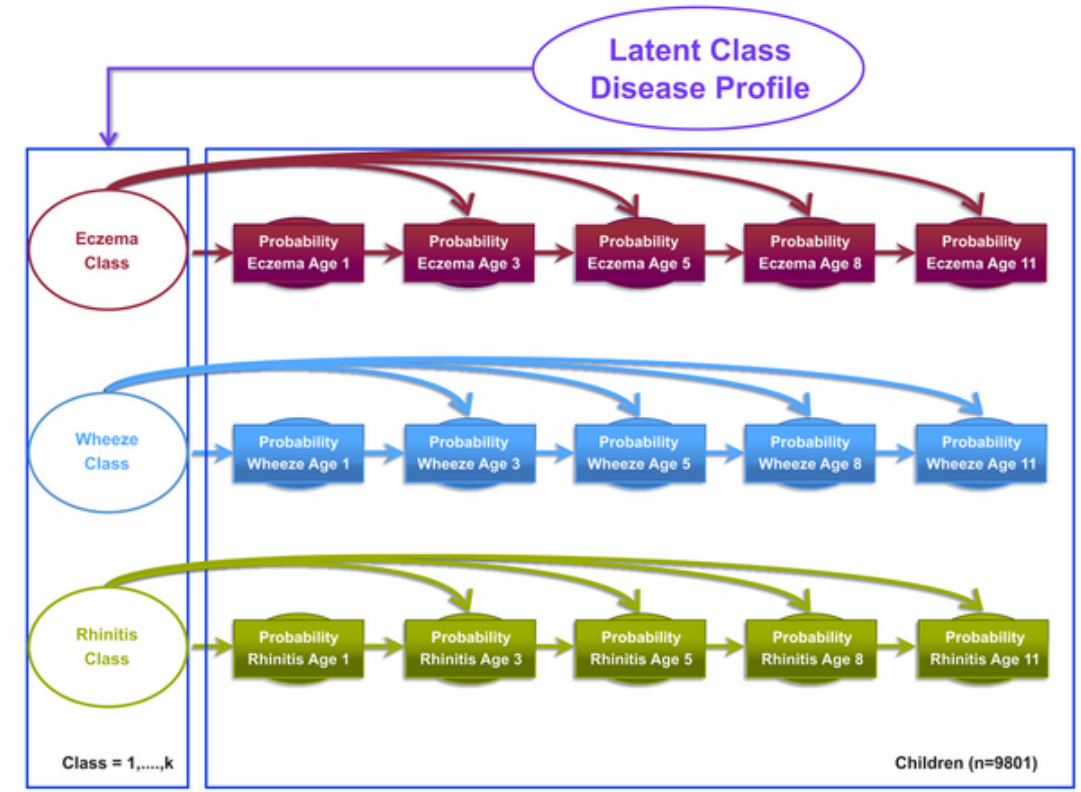
✓ ?

Bayes Nets Training

Bayes Nets Training

Where do all these probabilities come from?

Data $\xrightarrow{\text{MLE}}$ Parameter



Danielle Belgrave, et al. *PLOS Medicine*, 2014
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001748

Characteristic	MAAS Cohort		ALSPAC Cohort		Joint MAAS and ALSPAC Cohort	
	n/Total	Percent	*n*/Total	Percent	*n*/Total	Percent
Gender (Female)	617/1,136	54.3	4,212/8,665	48.61	4,829/9,801	49.3
Eczema						
Age 1 y	383/1,077					
Age 3 y	355/1,061					
Age 5 y	340/1,050					
Age 8 y	285/1,027					
Age 11 y	216/924					
Wheeze						
Age 1 y	300/1,087					
Age 3 y	257/1,095					
Age 5 y	238/1,056					
Age 8 y	185/1,024					
Age 11 y	173/916					
Rhinitis						
Age 1 y	8/943					
Age 3 y	49/1,075					
Age 5 y	292/1,039					
Age 8 y	297/1,027					
Age 11 y	321/927					

	n/Total	Percent
Gender (Female)	617/1,136	54.3
Eczema		
Age 1 y	383/1,077	35.6
Age 3 y	355/1,061	33.5
Age 5 y	340/1,050	32.4
Age 8 y	285/1,027	27.8
Age 11 y	216/924	23.4

Danielle Belgrave, et al. *PLOS Medicine*, 2014
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001748

Bayes Nets

Representation and Independence

(Training?)

Exact Inference

- Enumeration (always exponential complexity)
- Some shortcuts (worst-case exponential complexity, often better)
- Inference is NP-hard in general

Approximate Inference

- Prior sampling
- Rejection sampling
- Likelihood-weighted sampling
- Gibbs sampling

Bayes Nets Inference

Exact Inference

Answer Any Query from Joint Distribution

P(S, T, W)

P(Weather)?

P(Weather | winter)?

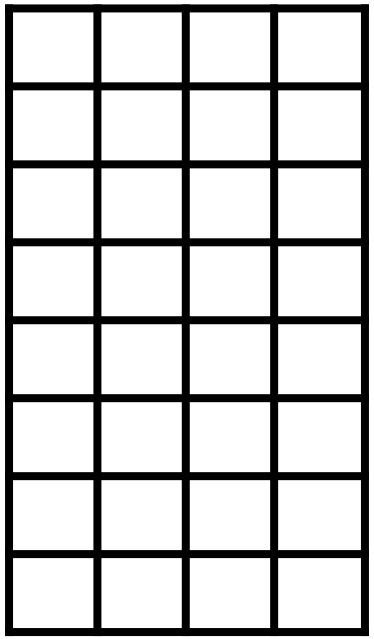
P(Weather | winter, hot)?

Season	Temp	Weather	P(S, T, W)
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Answer Any Query from Joint Distribution

Joint distributions are the best!

Joint



Query

$$P(q_1, q_2 \mid e_1, e_2, e_3)$$

Answer Any Query from Joint Distribution

Two tools to go from joint to query

1. Definition of conditional probability

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

Joint

$\xleftarrow{\text{sum } c}$ $P(A, B, C)$

$\xleftarrow{\text{sum } a\, c}$ $P(A, B, C)$

2. Law of total probability (marginalization, summing out)

$$P(A) = \sum_b P(A,b)$$

$$P(Y \mid U,V) = \sum_x \sum_z P(x,Y,z \mid U,V)$$

Answer Any Query from Joint Distribution

Two tools to go from joint to query

Joint: $P(H_1, H_2, Q, E)$

Query: $P(Q \mid e)$

1. Definition of conditional probability

$$P(Q \mid e) = \frac{P(Q, e)}{P(e)}$$

2. Law of total probability (marginalization, summing out)

$$P(Q, e) = \sum_{h_1} \sum_{h_2} P(h_1, h_2, Q, e)$$

$$P(e) = \sum_{q} \sum_{h_1} \sum_{h_2} P(h_1, h_2, q, e)$$

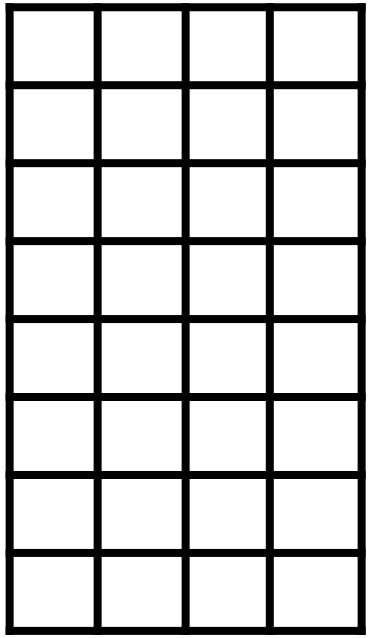
Answer Any Query from Joint Distribution

Joint distributions are the best!

Problems with joints

- We aren't given the joint table
  - Usually some set of conditional probability tables
- Huge
  - N variables with D values
  - D^N entries

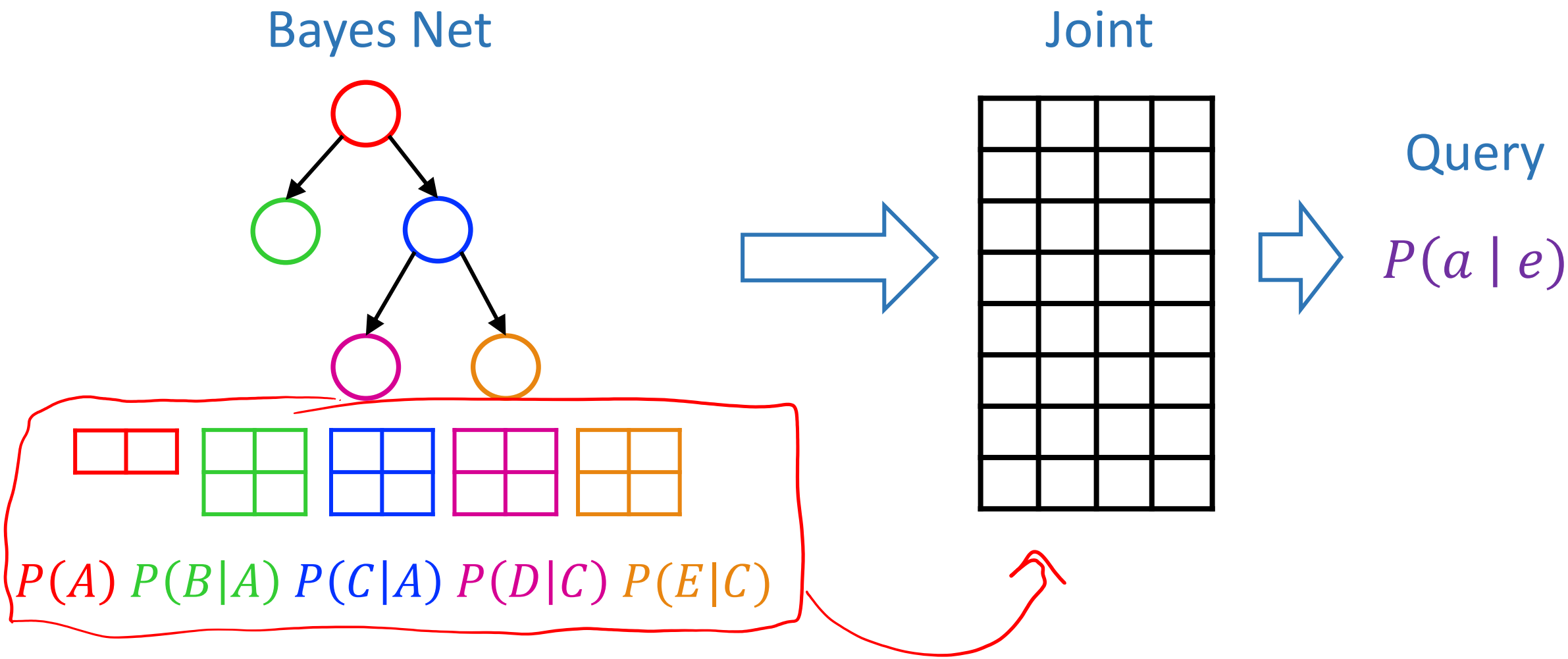
Joint



Query

$P(a \mid e)$

Answer Any Query from Bayes Net



CAUTION: Multipling Probability Tables!

We're going to write things like this quite a bit:

$$P(A,B,C) = P(A)\,P(\,B \mid A\,)\,P(\,C \mid A,B\,)$$

Is this regular multiplication?

We have tables... is this matrix multiplication?

$P(A,B,C)$

	$+a$	$-a$
$+b,+c$	.015	
$-b,+c$		
$+b,-c$		
$-b,-c$		

$P(A)$

$+a$	$-a$
0.1	0.9

$P(\,B \mid A\,)$

	$+b$	$-b$
$+a$	0.5	0.5
$-a$	0.2	0.8

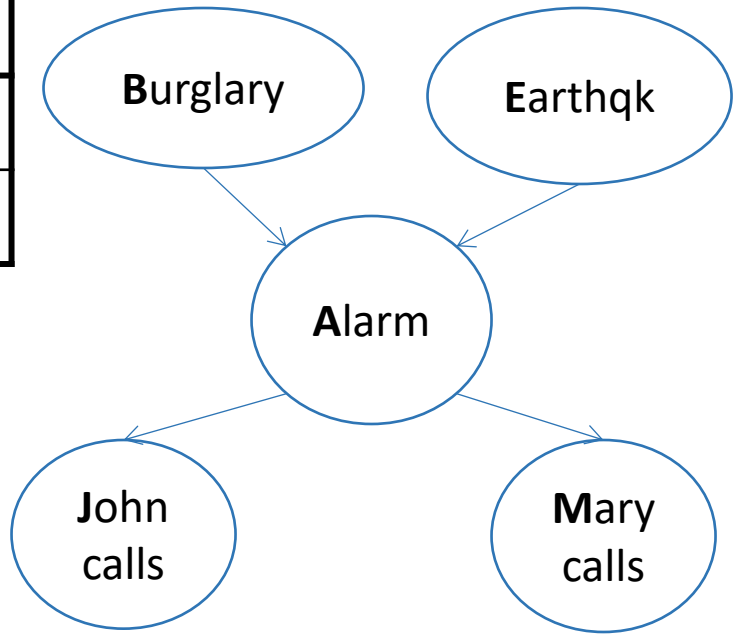
$P(\,C \mid A,B\,)$

	$+c$	$-c$
$+a,+b$	0.3	0.7
$-a,+b$	0.0	1.0
$+a,-b$	0.7	0.3
$-a,-b$	1.0	0.0

Example: Alarm Network

$P(+b, -e, -a, -j, -m) =$ ✓

B	P(B)
+b	0.001
-b	0.999

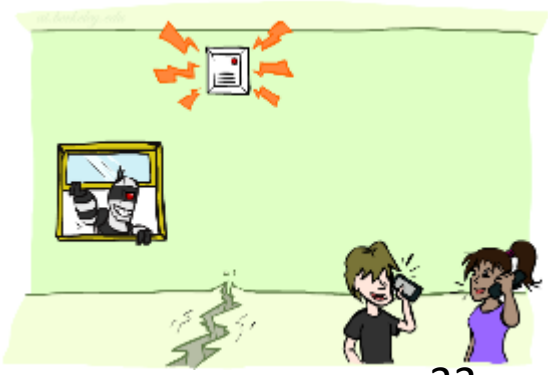


E	P(E)
+e	0.002
-e	0.998

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



Example: Alarm Network

$$P(+b,-e,-a,-j,-m) = P(+b) * P(-e) * P(-a|+b,-e) * P(-j|-a) * P(-m|-a)$$
$$= 0.001 * 0.998 * 0.06 * 0.95 * 0.99$$

$$P(+b,-e,-a,-j,-m) = P(-e) * P(-a|+b,-e) * P(+b) * P(-j|-a) * P(-m|-a)$$
$$= 0.998 * 0.06 * 0.001 * 0.95 * 0.99$$

$$P(+b,-e,-a,-j,-m) = P(-e) * P(-a|+b,-e) * P(+b)\, P(-j|-a) * P(-m|-a)$$
$$= 0.998 * 0.06 * 0.0095 * 0.99$$

- Multiplication order can change (commutativity)
- Multiplication pairs don't have to make sense (associativity)

Exact

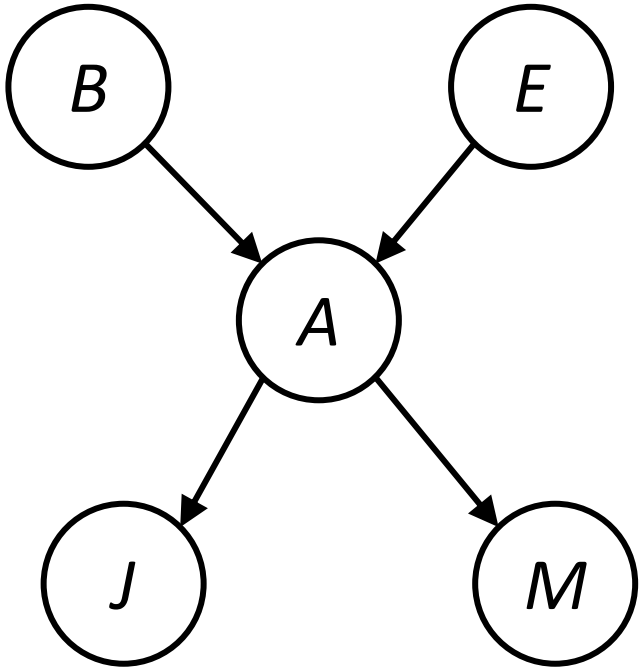
Inference by Enumeration in Bayes Net

Brute force

Reminder of inference by enumeration:

- Any probability of interest can be computed by summing entries from the joint distribution
- Entries from the joint distribution can be obtained from a BN by multiplying the corresponding conditional probabilities

$$
\begin{aligned}
P(B \mid j, m) &= \alpha\, P(B, j, m) \\
&= \alpha \sum_{e,a} P(B, e, a, j, m) \\
&= \alpha \sum_{e,a} P(B)\, P(e)\, P(a \mid B, e)\, P(j \mid a)\, P(m \mid a)
\end{aligned}
$$



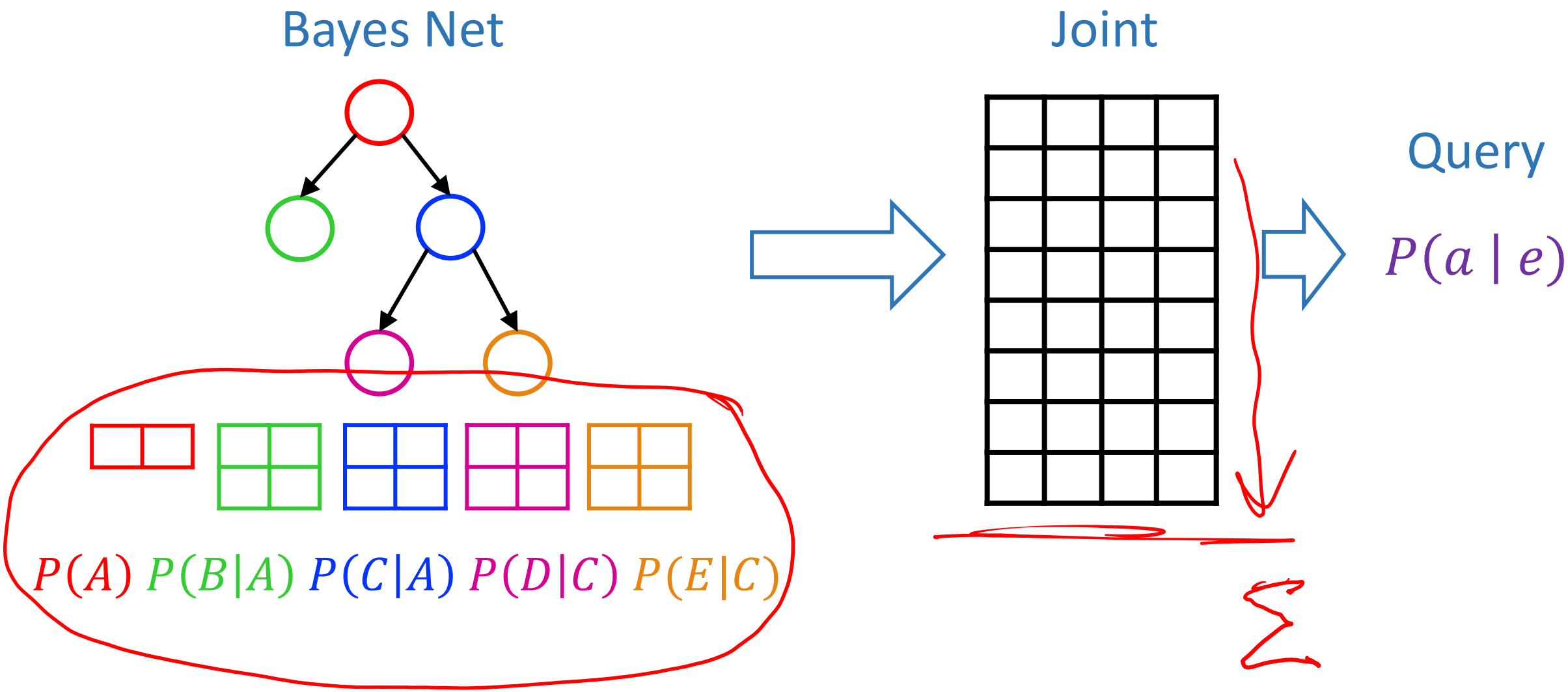
So inference in Bayes nets means computing sums of products of numbers: sounds easy!!

Problem: sums of ***exponentially many*** products!

Bayes Net Inference

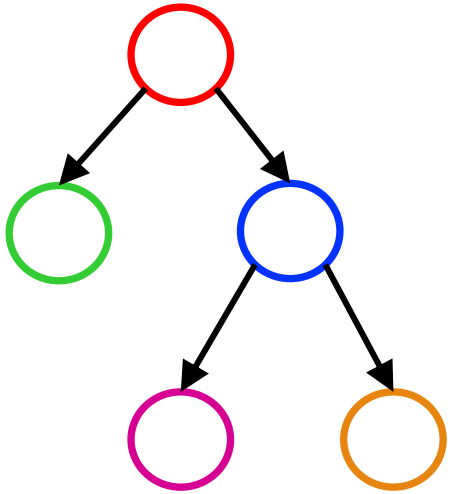
Exact inference shortcuts: Variable elimination

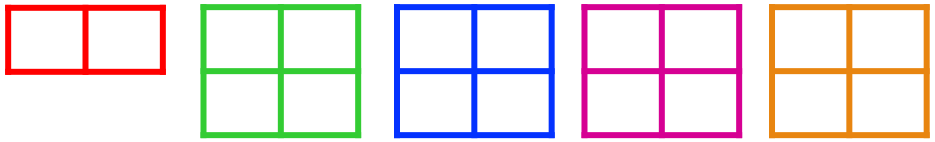
Answer Any Query from Bayes Net



Next: Answer Any Query from Bayes Net

Bayes Net





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

Query

$P(a \mid e)$

Can we do better?

$x_1 \quad x_2$
$y_1 \quad y_2$
$z_1 \quad z_2$

Consider

- $x_1y_1z_1 + x_1y_1z_2 + x_1y_2z_1 + x_1y_2z_2 + x_2y_1z_1 + x_2y_1z_2 + x_2y_2z_1 + x_2y_2z_2$
- 16 multiplies, 7 adds
- Lots of repeated subexpressions!

Rewrite as

- $(x_1 + x_2)(y_1 + y_2)(z_1 + z_2)$
- 2 multiplies, 3 adds

$$\sum_e \sum_a P(B)\, P(e)\, P(a \mid B, e)\, P(j \mid a)\, P(m \mid a)$$

$$\begin{aligned}
= \; & P(B)\, P(+e)\, P(+a \mid B, +e)\, P(j \mid +a)\, P(m \mid +a) \\
+ \; & P(B)\, P(-e)\, P(+a \mid B, -e)\, P(j \mid +a)\, P(m \mid +a) \\
+ \; & P(B)\, P(+e)\, P(-a \mid B, +e)\, P(j \mid -a)\, P(m \mid -a) \\
+ \; & P(B)\, P(-e)\, P(-a \mid B, -e)\, P(j \mid -a)\, P(m \mid -a)
\end{aligned}$$

- Lots of repeated subexpressions!

Variable elimination: The basic ideas

Move summations inwards as far as possible

$$P(B \mid j,m) = \alpha \sum_e \sum_a P(B,e,a,j,m)$$

$$= \alpha \sum_e \sum_a P(j|a)\, P(e)\, P(m|a)\, P(a|B,e)\, P(B)$$

$$= \alpha \sum_e P(e)\;\; P(B) \sum_a P(j|a)\, P(m|a)\, P(a|B,e)$$

$$= \alpha\; P(B) \sum_e P(e) \sum_a P(j|a)\, P(m|a)\, P(a|B,e)$$

Inference Overview

Given random variables Q, H, E (query, hidden, evidence)

Enumeration

We know how to do inference on a joint distribution

$$P(q|e) = \alpha\, P(q,e)$$
$$= \alpha \sum_{h \in \{h_1,h_2\}} P(q,h,e)$$

We know Bayes nets can break down joint in to CPT factors

$$H \rightarrow Q \rightarrow E$$

$$P(q|e) = \alpha \sum_{h \in \{h_1,h_2\}} P(h)\, P(q|h)\, P(e|q)$$
$$= \alpha\, [P(h_1)\, P(q|h_1)\, P(e|q) + P(h_2)\, P(q|h_2)\, P(e|q)]$$

Variable Elimination

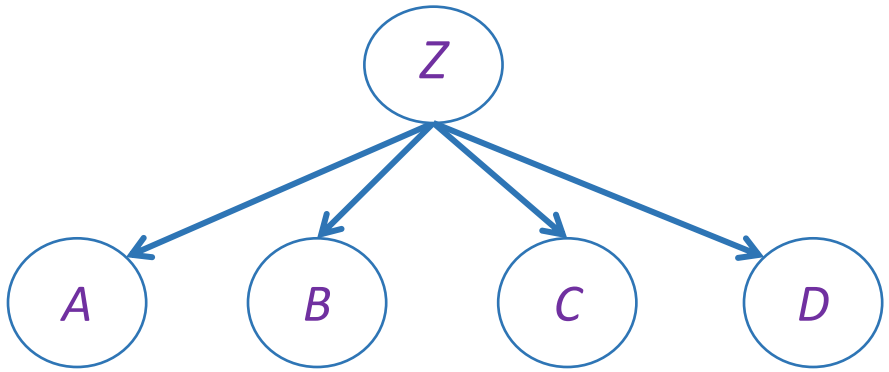
But we can be more efficient

$$P(q|e) = \alpha\, P(e|q) \sum_{h \in \{h_1,h_2\}} P(h) P(q|h)$$
$$= \alpha\, P(e|q)\, [P(h_1) P(q|h_1) + P(h_2) P(q|h_2)]$$
$$= \alpha\, P(e|q)\, P(q)$$

Now just extend to larger Bayes nets and a variety of queries

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)

Variable Elimination: 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

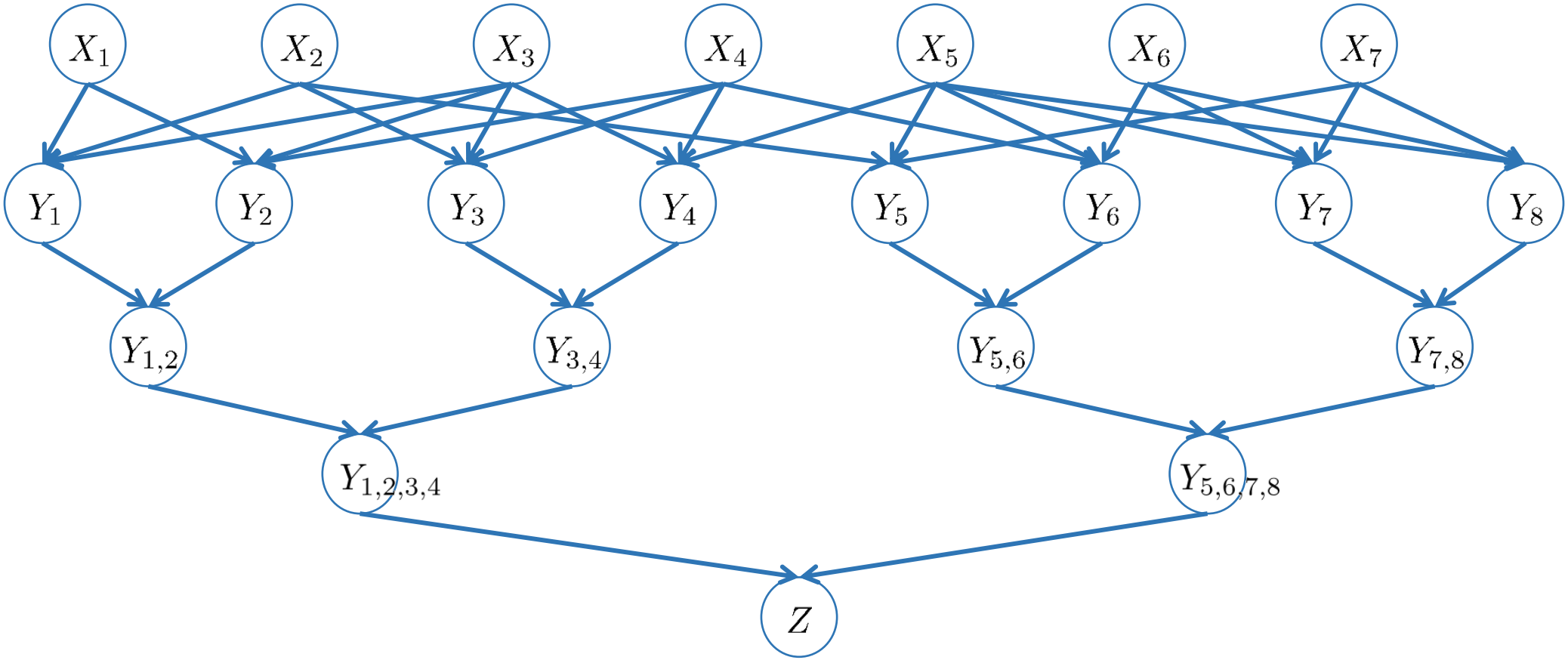
Does there always exist an ordering that only results in small factors?

- No!

VE: Computational and Space Complexity

Inference in Bayes' nets is NP-hard.

No known efficient probabilistic inference in general.



Bayes Nets

Representation and Independence

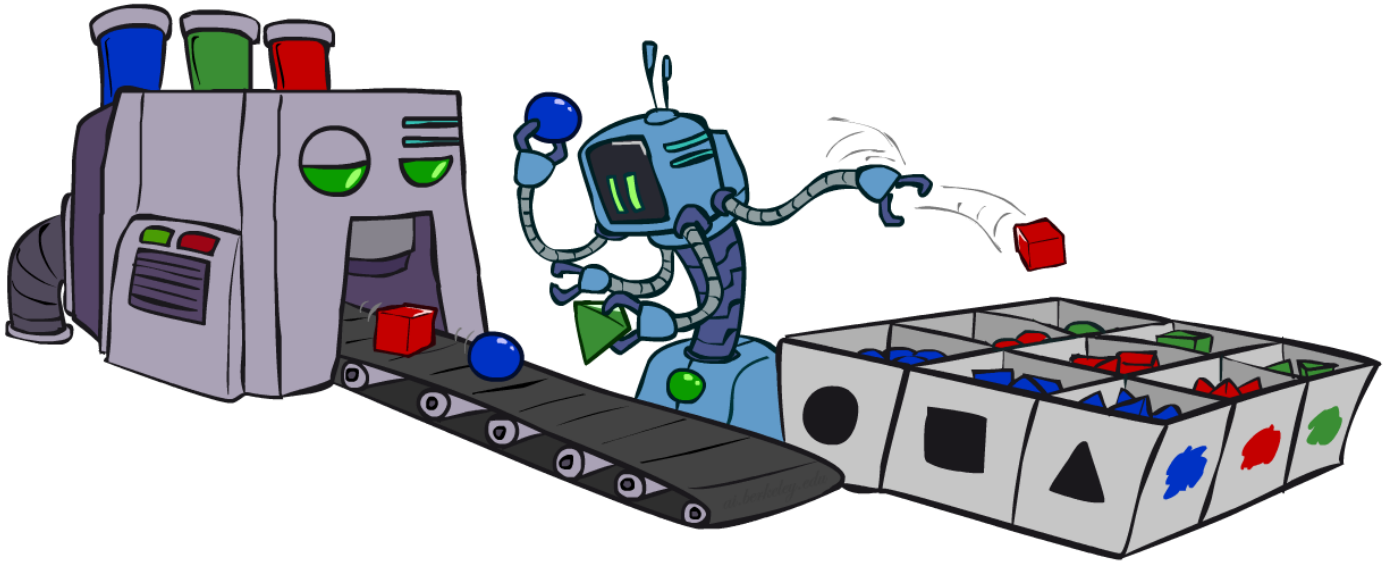
(Training?)

Exact Inference

- Enumeration (always exponential complexity)
- Some shortcuts (worst-case exponential complexity, often better)
- Inference is NP-hard in general

Approximate Inference

- Prior sampling
- Rejection sampling
- Likelihood-weighted sampling
- Gibbs sampling



Bayes Nets Inference

Approximate Inference (Sampling)

Poll 1

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

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

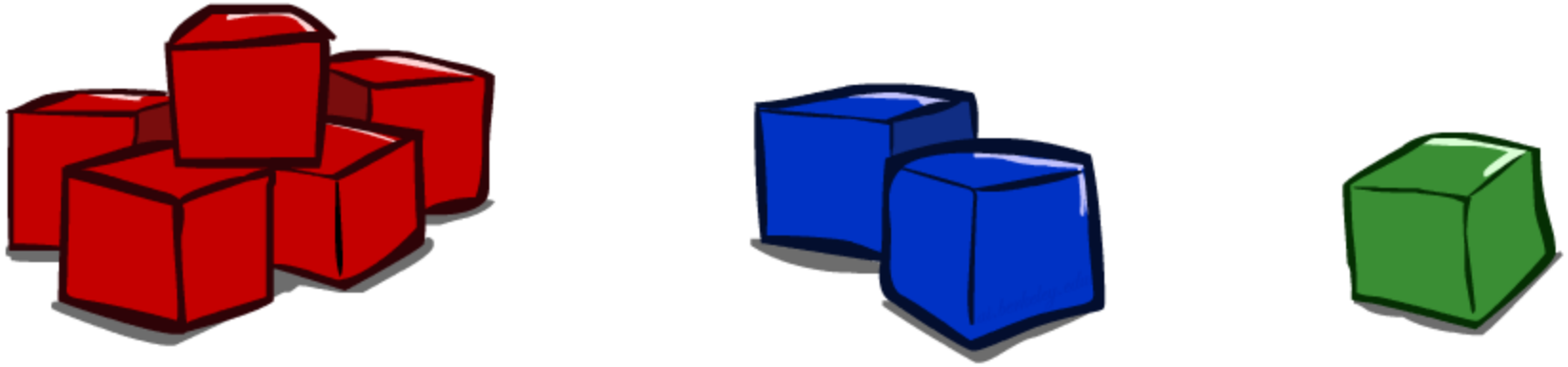
Joint

N=10

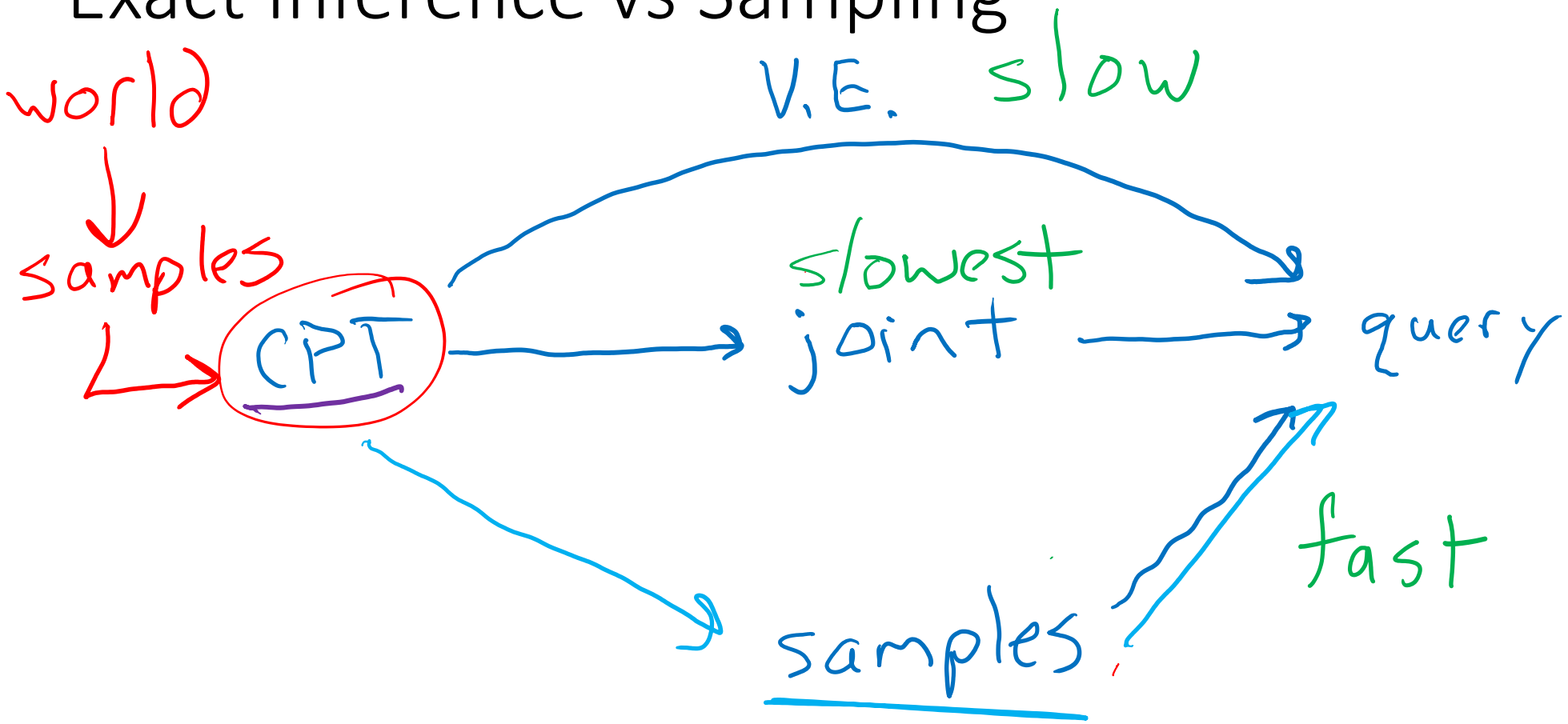
Counts

+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



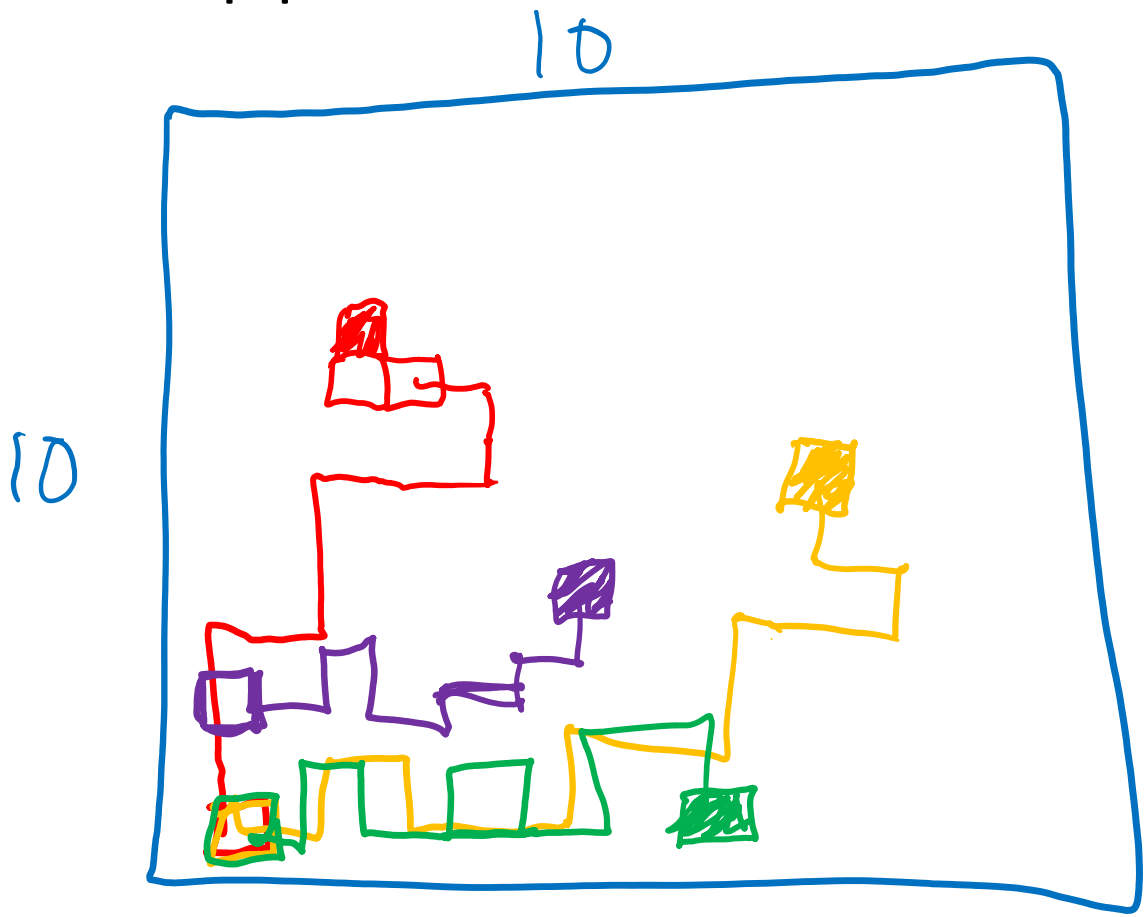
Exact Inference vs Sampling



Motivation for Approximate Inference

$P(G_{t+1} \mid G_t)$

$P(G_0)$

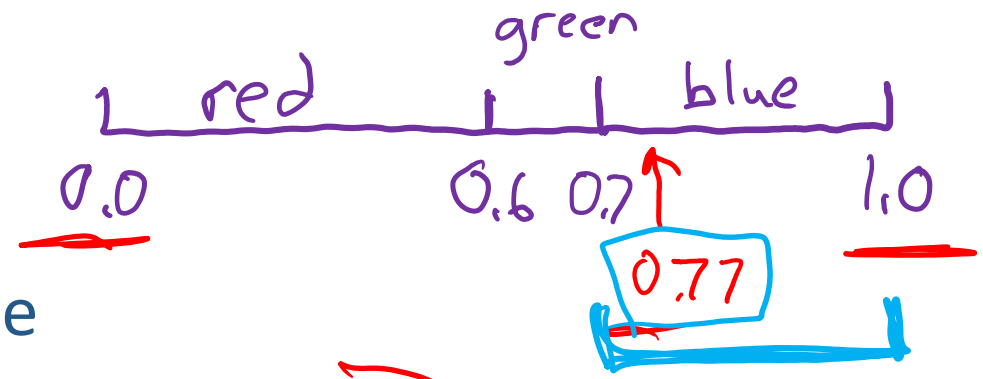


$P(G_{20})$

$$G_0 \rightarrow G_1 \rightarrow G_2 \rightarrow$$

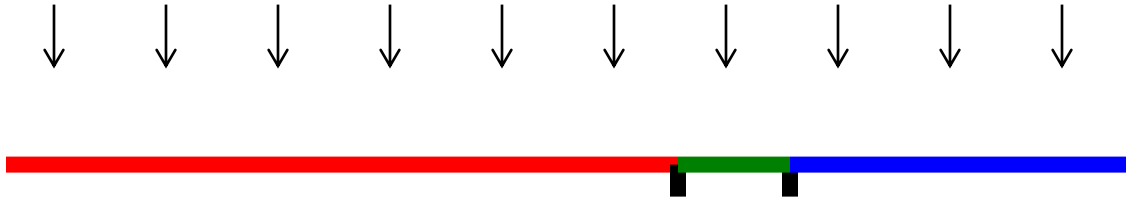
Sampling

rand



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

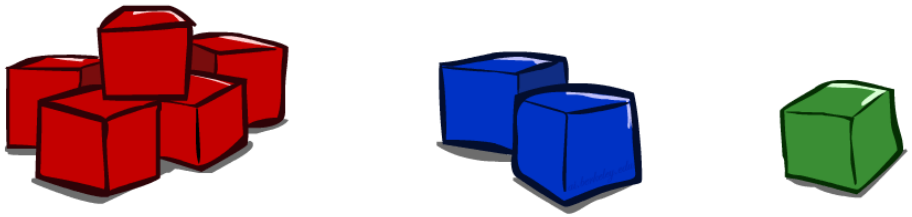


Example

C	P(C)
red	0.6
green	0.1
blue	0.3

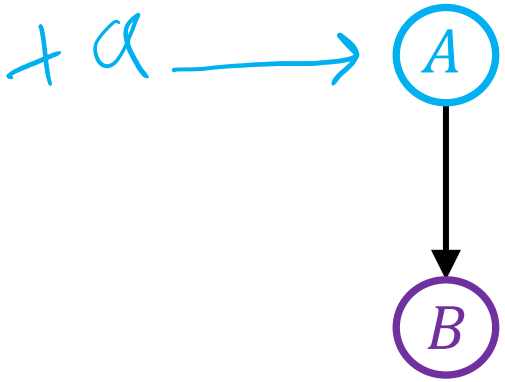
$$0 \leq u < 0.6, \rightarrow C = red$$
$$0.6 \leq u < 0.7, \rightarrow C = green$$
$$0.7 \leq u < 1, \rightarrow C = blue$$

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



Sampling

How would you sample from a conditional distribution?



$P(A)$

+a	1/2
-a	1/2

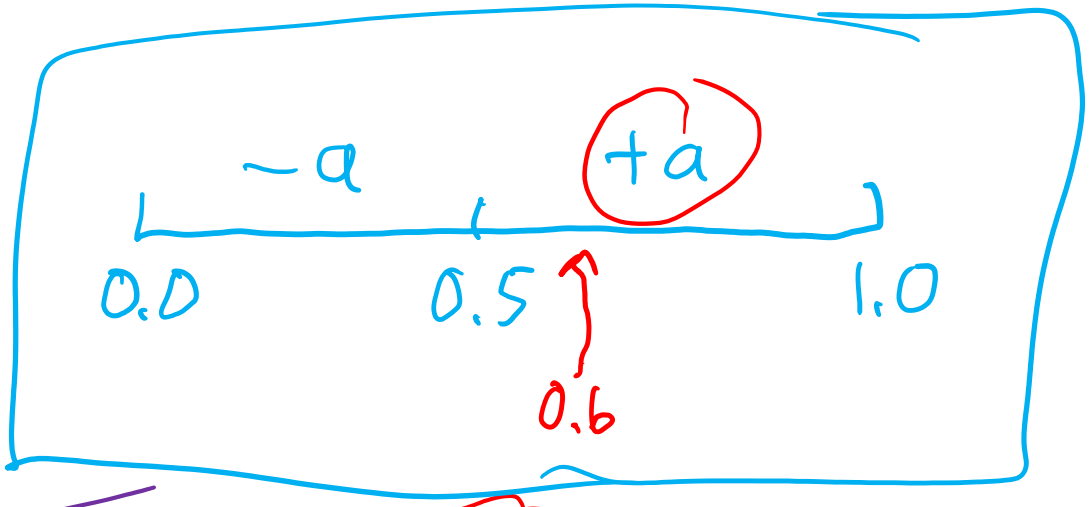
$P(B|A)$

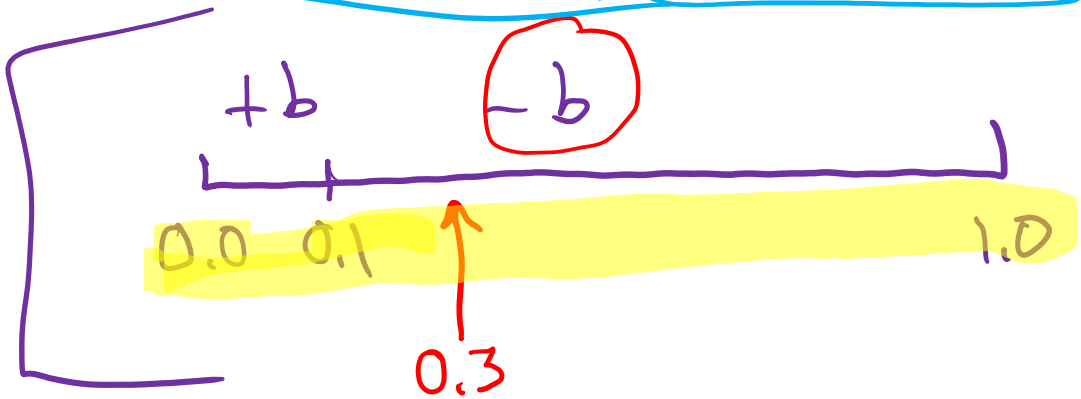
+a	+b	1/10
	-b	9/10
-a	+b	1/2
	-b	1/2

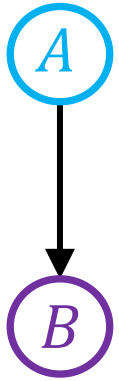
Sampling

$P(A, B)$

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

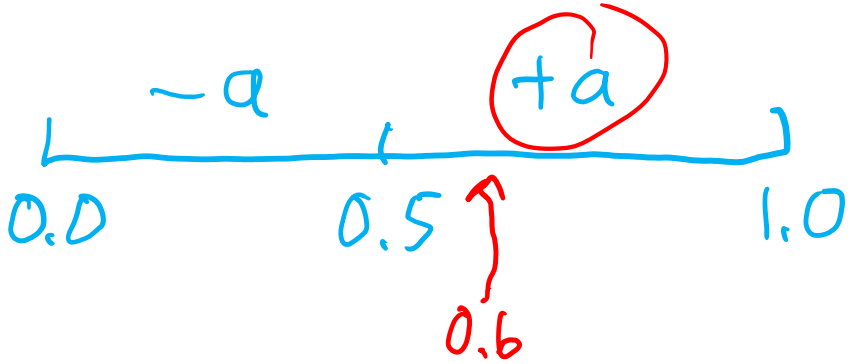
$P(B|+a)$

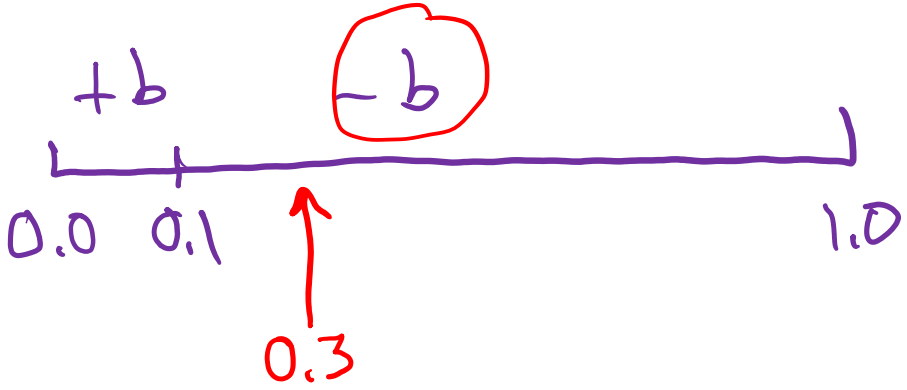
$P(B|-a)$

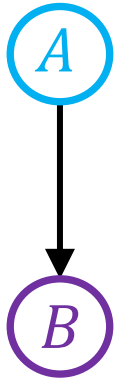
Sampling

$P(A, B)$

How would you sample from a conditional distribution?





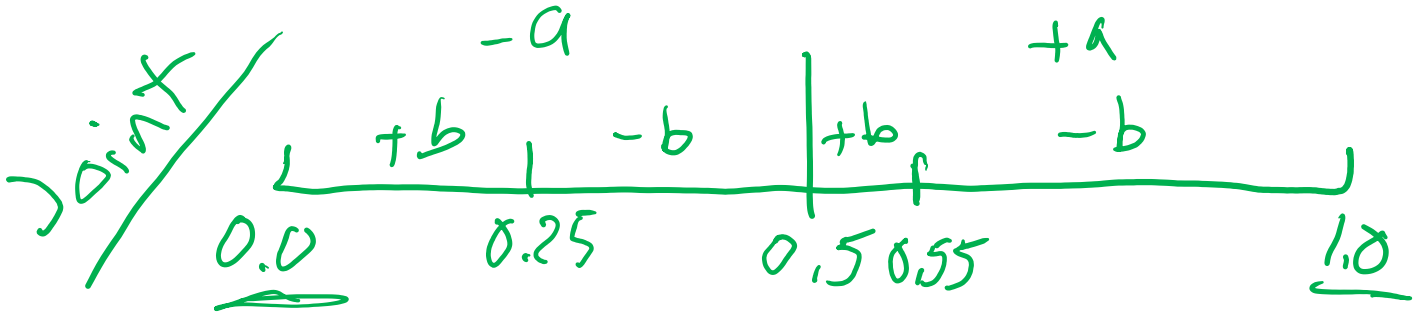


$P(A)$

+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 ~ mch

Likelihood Weighting

Out of of scope

Gibbs Sampling

Handout

No evidence: Prior Sampling

Input: no evidence

node

for i=1, 2, …, m

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

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

Some evidence: Likelihood Weighted Sampling

Input: evidence instantiation

w = 1.0

for i=1, 2, …, m

if X_i is an evidence variable

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

else

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

return $(x_1, x_2, \ldots, x_m)$, w

All evidence: Likelihood Weighted

Input: evidence instantiation

w = 1.0

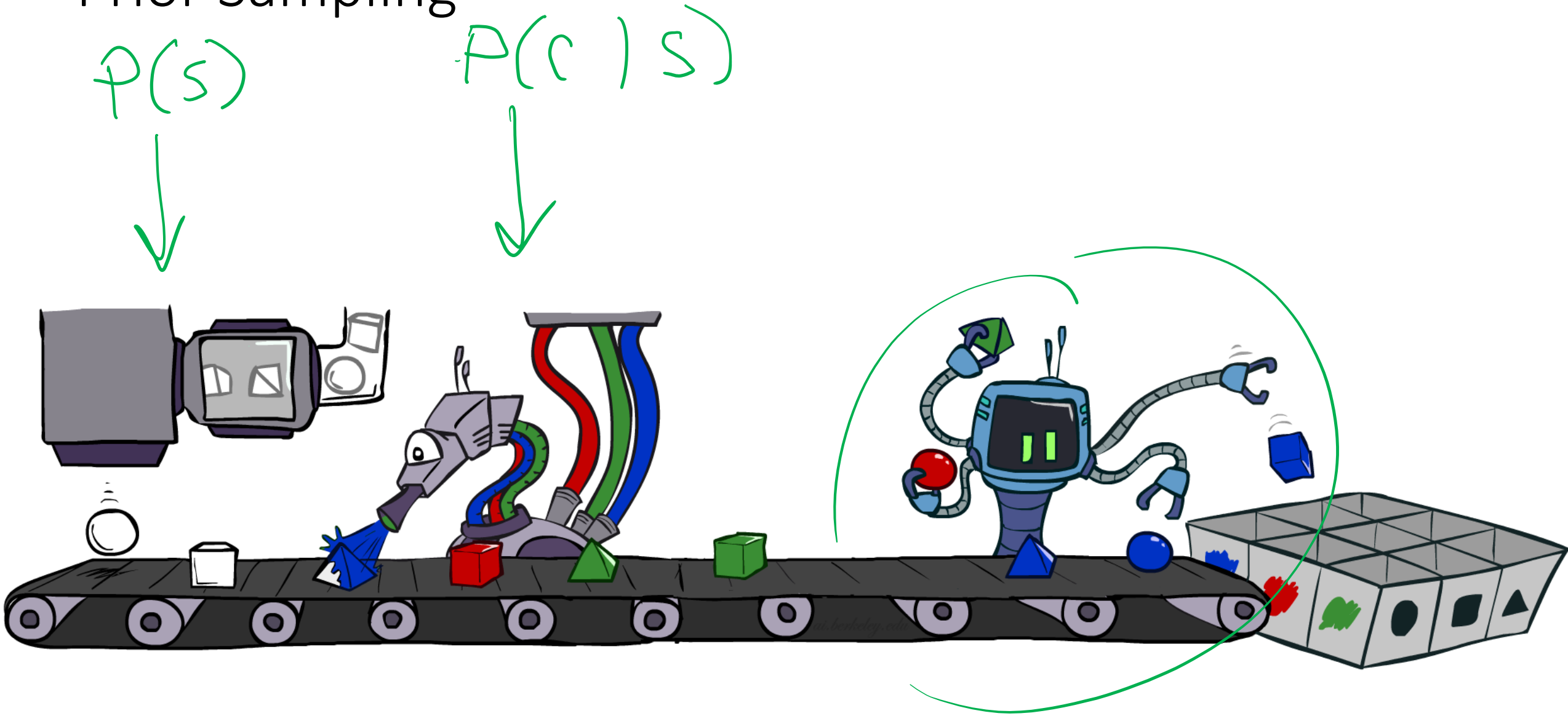
for i=1, 2, …, m

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

return w

1 sample

Prior Sampling



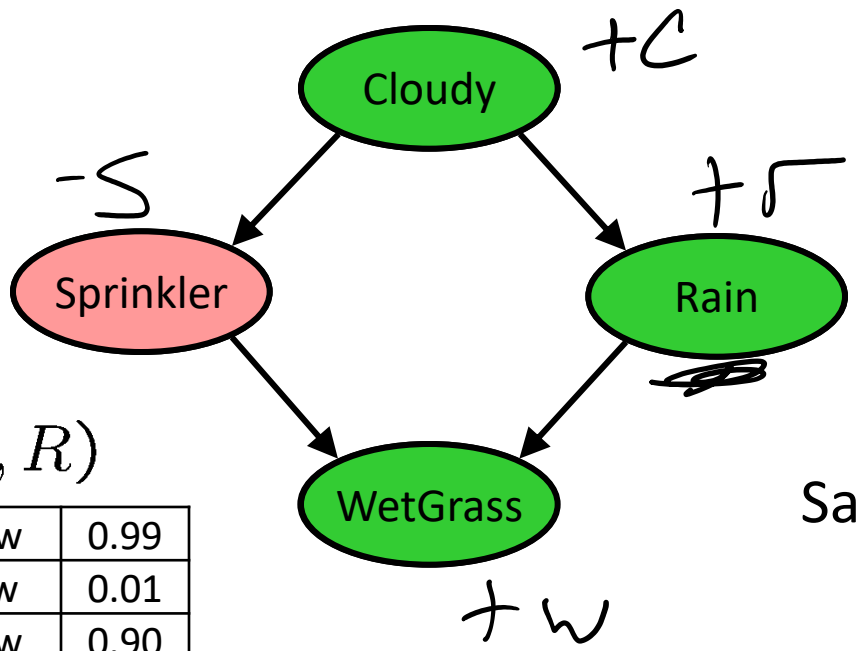
Prior Sampling Demo

$P(C)$

+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

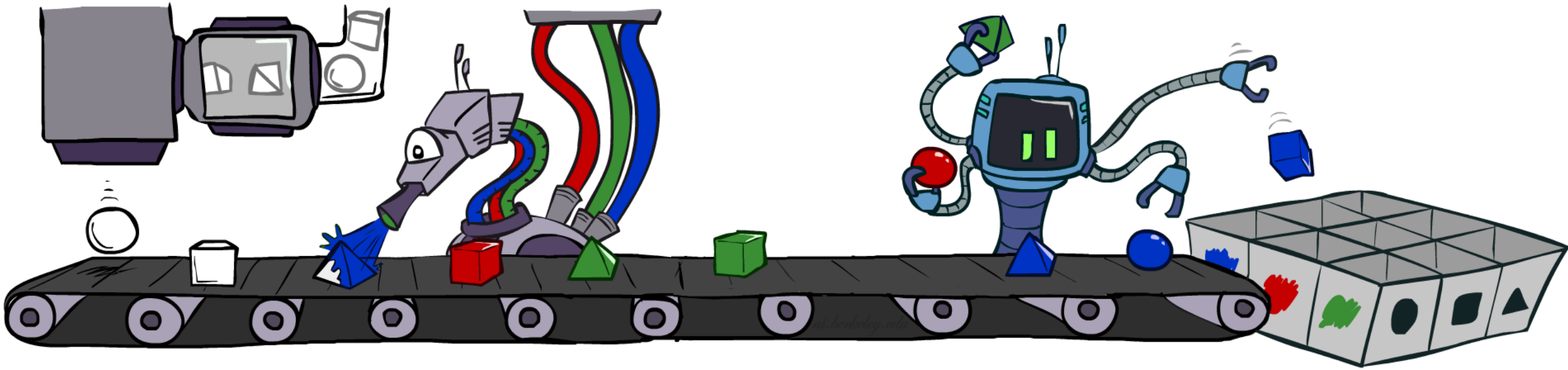
...

Prior Sampling

For i = 1, 2, …, m

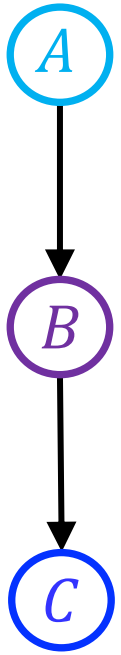
Sample x_i from $P(X_i \mid \text{Parents}(X_i))$

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



Poll 2

Prior Sampling: What does the value $\frac{N(+a,-b,+c)}{N=1000}$ 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

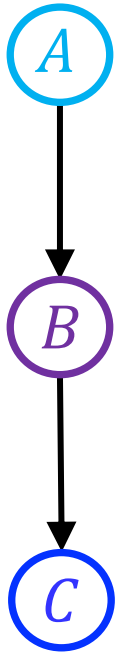
$P(A)$ $P(B \mid +a)$ $P(C \mid +a, -b)$

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

0.5 0.5 0.5 $= 0.125$

Poll 2

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:

$$\mathrm{P}(a) = 0.5\ \forall a$$
$$\mathrm{P}(b \mid a) = 0.5\ \forall a, b$$
$$\mathrm{P}(c \mid b) = 0.5\ \forall b, c$$

Let's take 1000 samples:

Poll 3

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, …, m
    Set w = w * P(x_i | 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

$P(C|B)$

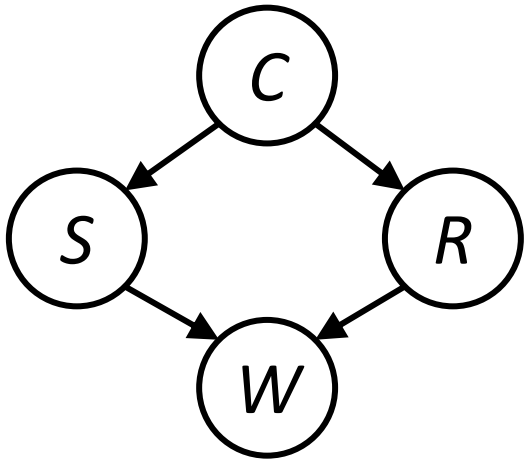
+b	+c	4/5
	-c	1/5
-b	+c	1
	-c	0

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

$\frac{3}{4}$	P(+c | +w)
1/4	P(-c | +w)



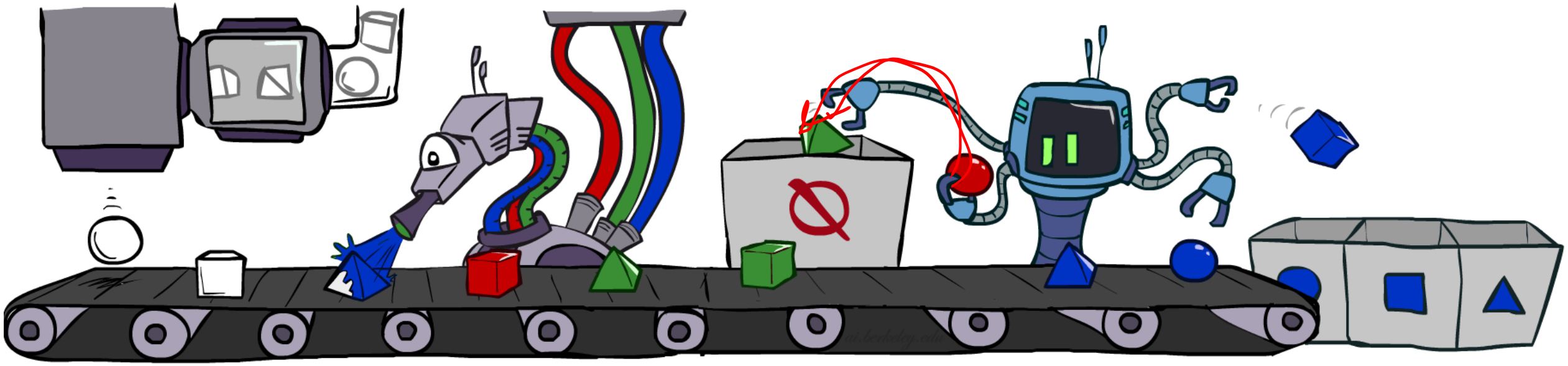
If we want to know P(W)

- 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
- Can estimate anything else, too
- What about P(C| +w)? P(C| +r, +w)? P(C| -r, -w)?
- Fast: can use fewer samples if less time (what's the drawback?)

Rejection Sampling

$\boxed{P(S \mid \text{blue})}$

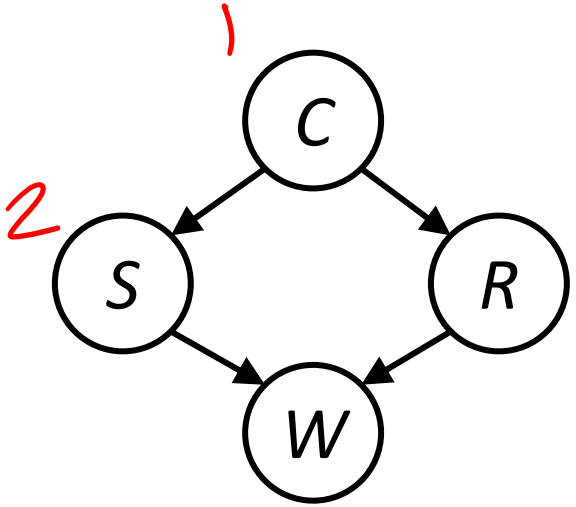
keep only blue



Rejection Sampling

Let's say we want P(C | +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, -s, +r, -w~~

+c, +s, +r, +w

-c, +s, +r, -w

~~+c, -s, +r, +w~~

~~-c, -s, -r, +w~~

Rejection Sampling

IN: evidence instantiation

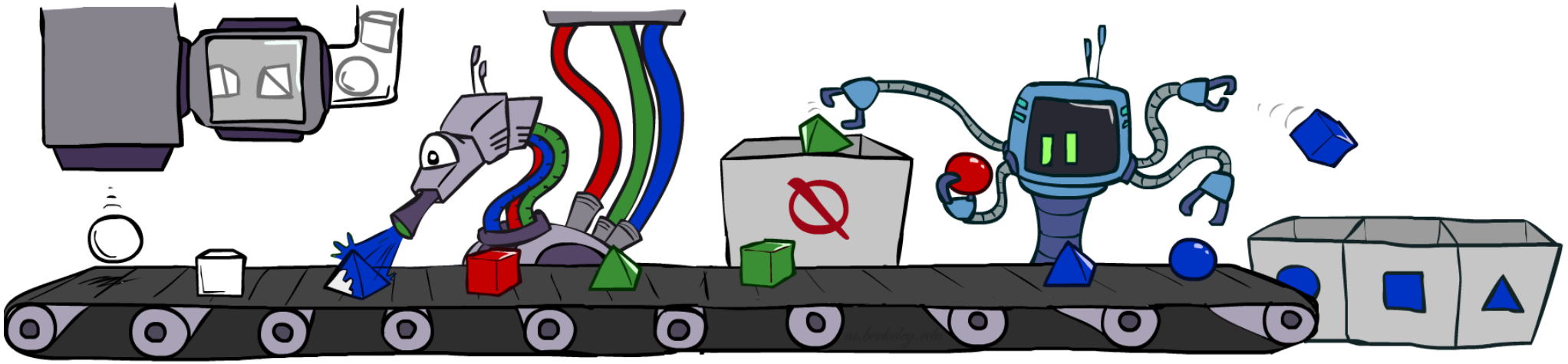
For i=1, 2, …, m

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 $(x_1, x_2, \ldots, x_m)$



Poll 3

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

Select all that apply

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

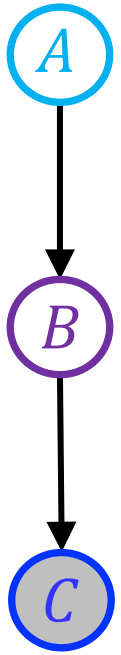
B. $P(+a, -b \mid +c)$ $\frac{3}{10}$

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

D. None of the above

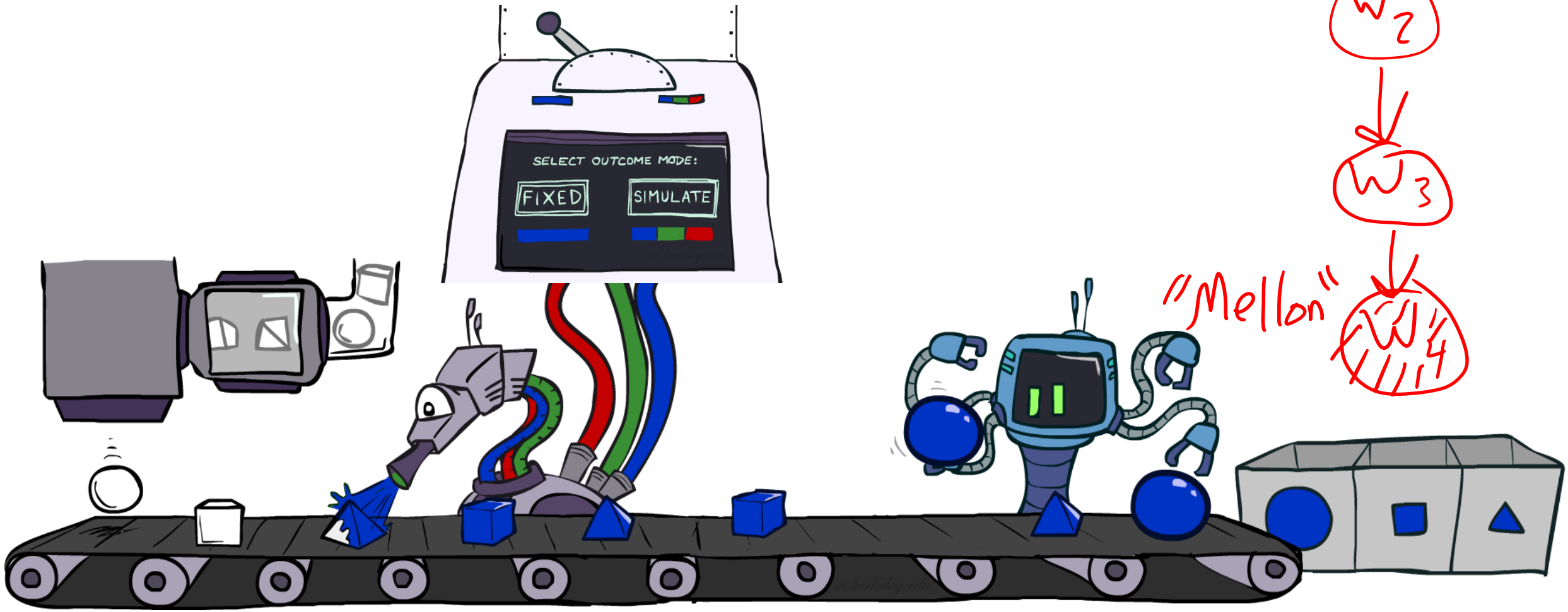
Counts $N(A, B, C)$

+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	



Likelihood Weighting

$P(w_1\ w_2\ w_3 \mid W_4 = \text{"Mellon"})$



Likelihood Weighting

$P(S)\,\underline{\underline{P(b|S)}} = \hat{P}(S, \text{blue})$

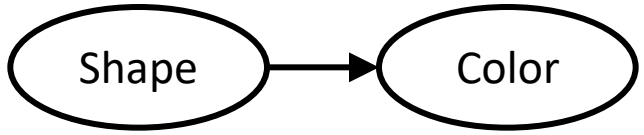
sample

S	color
pyr	b
sph	b
cub	b

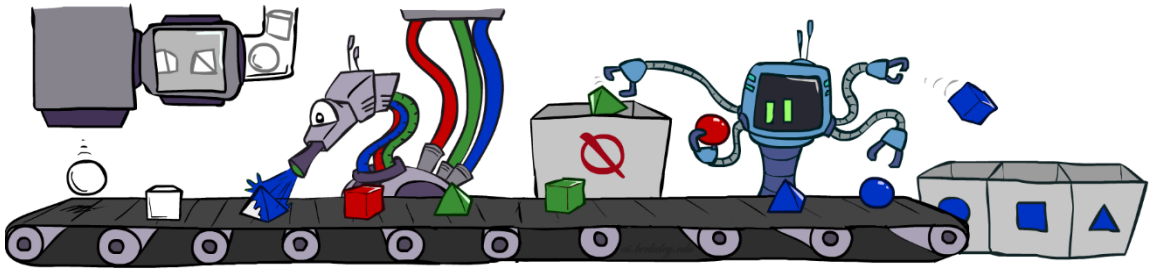
$\boxed{\begin{matrix} 2/5 \\ 2/5 \\ 1/5 \end{matrix}}$ $\hat{P}(S)$

Problem with rejection sampling:

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



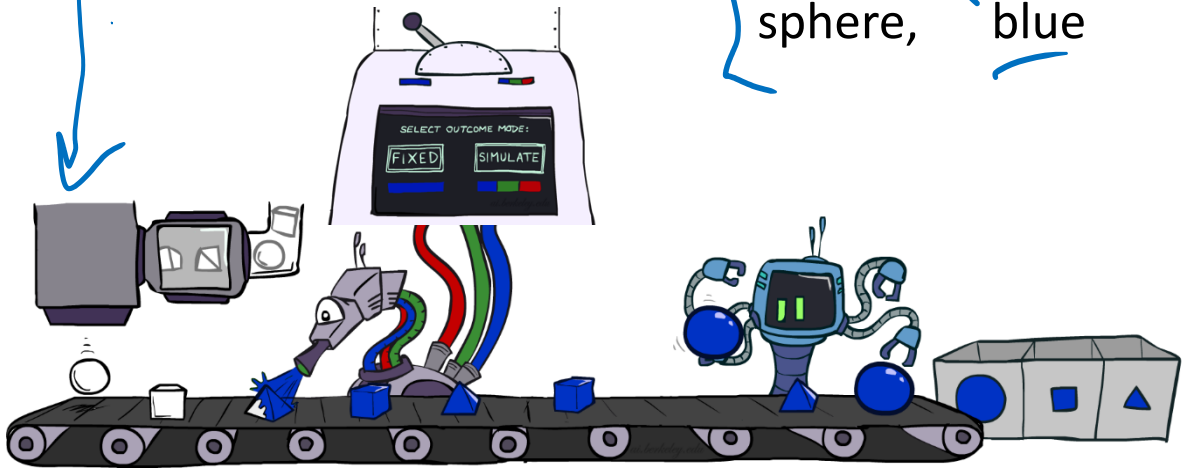
~~pyramid, green~~
~~pyramid, red~~
sphere, blue
~~cube, red~~
~~sphere, green~~



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

$P(S, \text{blue})$

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



Likelihood Weighting

Sample

$P(C)$

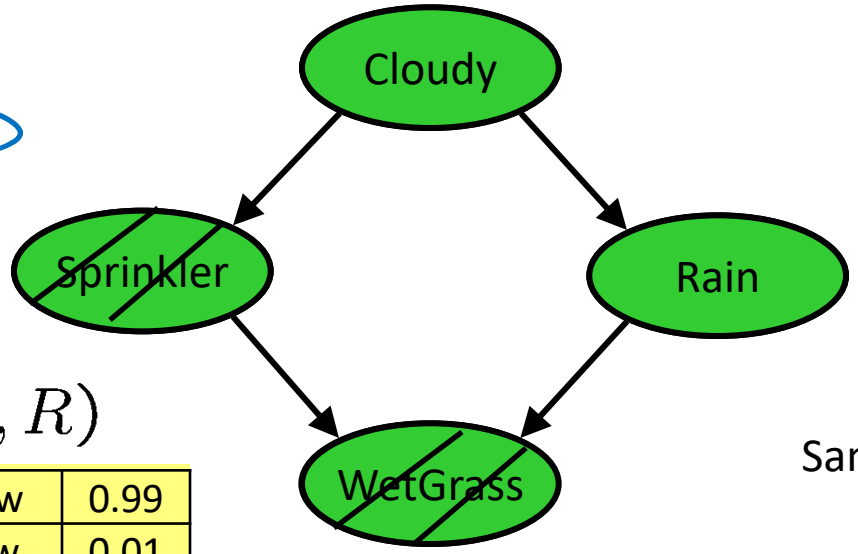
+c	0.5
-c	0.5

Don't sample

→ Grab weigh from table

$P(S|C)$

+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

…

$P(+s|+c) \times P(+w|+s,+r)$

$$w = 1.0 \times 0.1 \times 0.99$$

Likelihood Weighting

for N samples

IN: evidence instantiation

w = 1.0

for i=1, 2, …, m

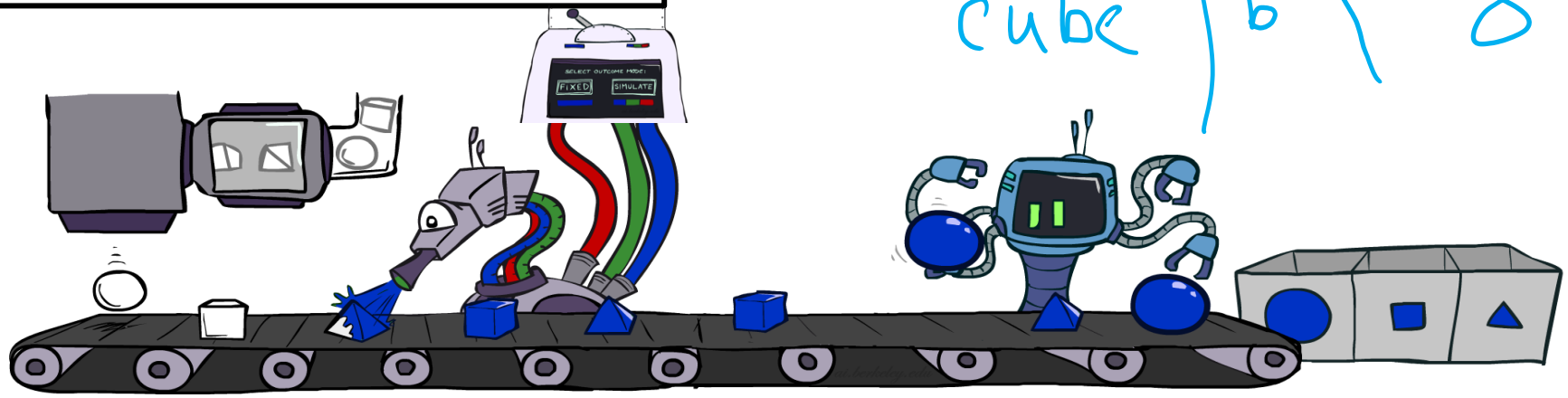
- if X_i is an evidence variable
  - X_i = observation x_i for X_i
  - Set w = w * $P(x_i \mid Parents(X_i))$
- else
  - Sample x_i from $P(X_i \mid Parents(X_i))$

return $(x_1, x_2, \ldots, x_m)$, w

Sampling Table

		Counts	W
sph	blue	2	0.1
pyr	blue	1	0.3

S	C	P(S, blue)
sph	b	2/3 * 0.1 = 0.066
pyr	b	1/3 * 0.3 0.099
cube	b	0



Likelihood Weighting

No evidence:

Prior Sampling

Input: no evidence

for i=1, 2, …, m

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

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

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

Some evidence:

Likelihood Weighted Sampling

Input: evidence instantiation

w = 1.0

for i=1, 2, …, m

if X_i is an evidence variable

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

else

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

return $(x_1, x_2, \ldots, x_m)$, w

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

All evidence:

Likelihood Weighted

Input: evidence instantiation

w = 1.0

for i=1, 2, …, m

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

return w

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

Remember Poll 3

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

$P(C|B)$

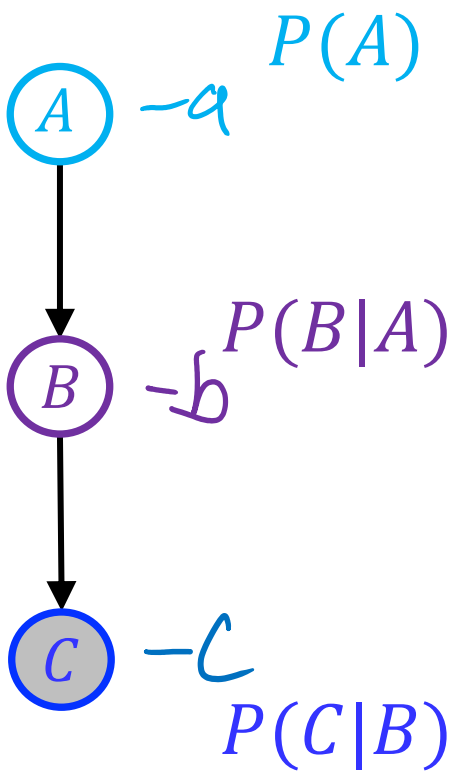
+b	+c	4/5
	-c	1/5
-b	+c	1
	-c	0

Likelihood Weighting

I

eat

"Mellon"

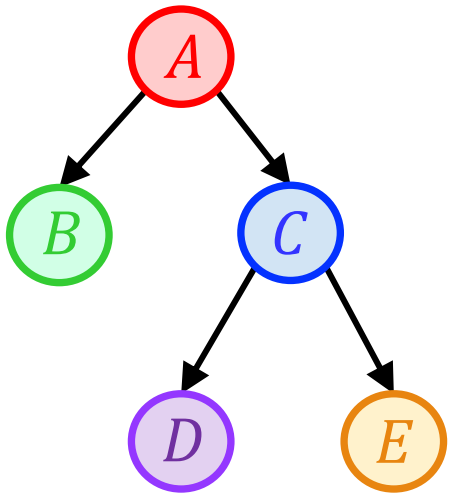


+a	1/2
-a	1/2

+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

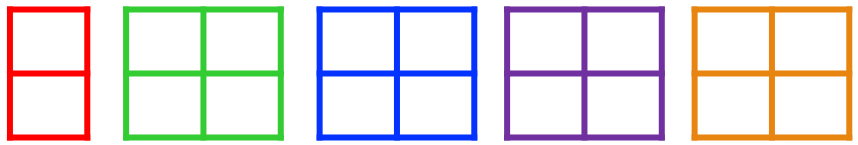


Consistency of likelihood weighted sampling distribution

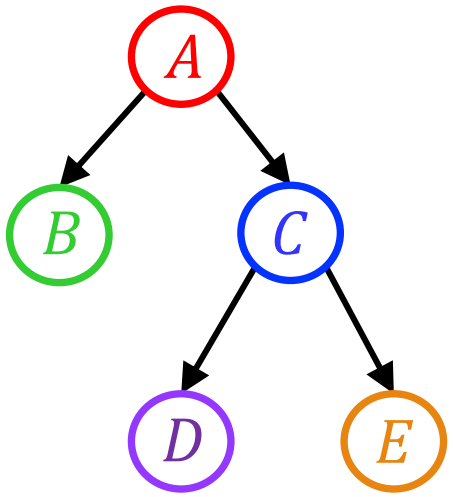
Evidence: $+a$, $+b$, $-c$, $-d$, $+e$

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

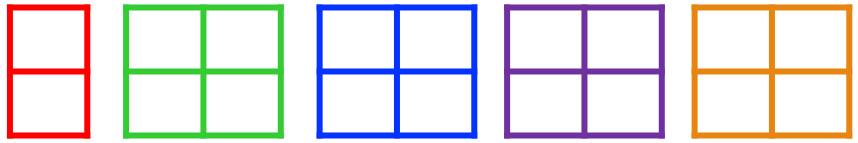


Consistency of likelihood weighted sampling distribution

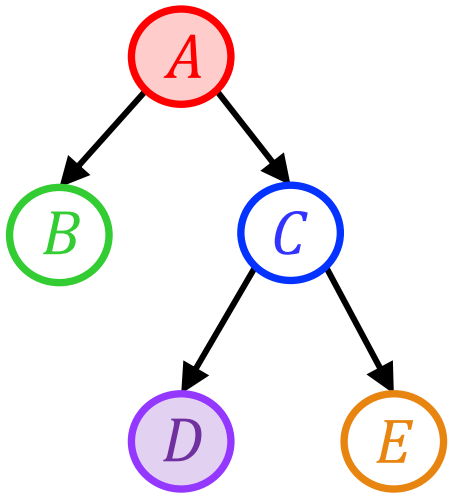
Evidence: None

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

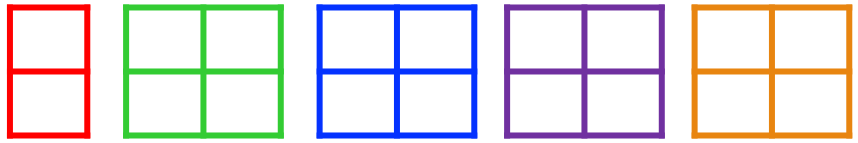


Consistency of likelihood weighted sampling distribution

Evidence: $+a$, $-d$

Joint from Bayes nets

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



Likelihood Weighting

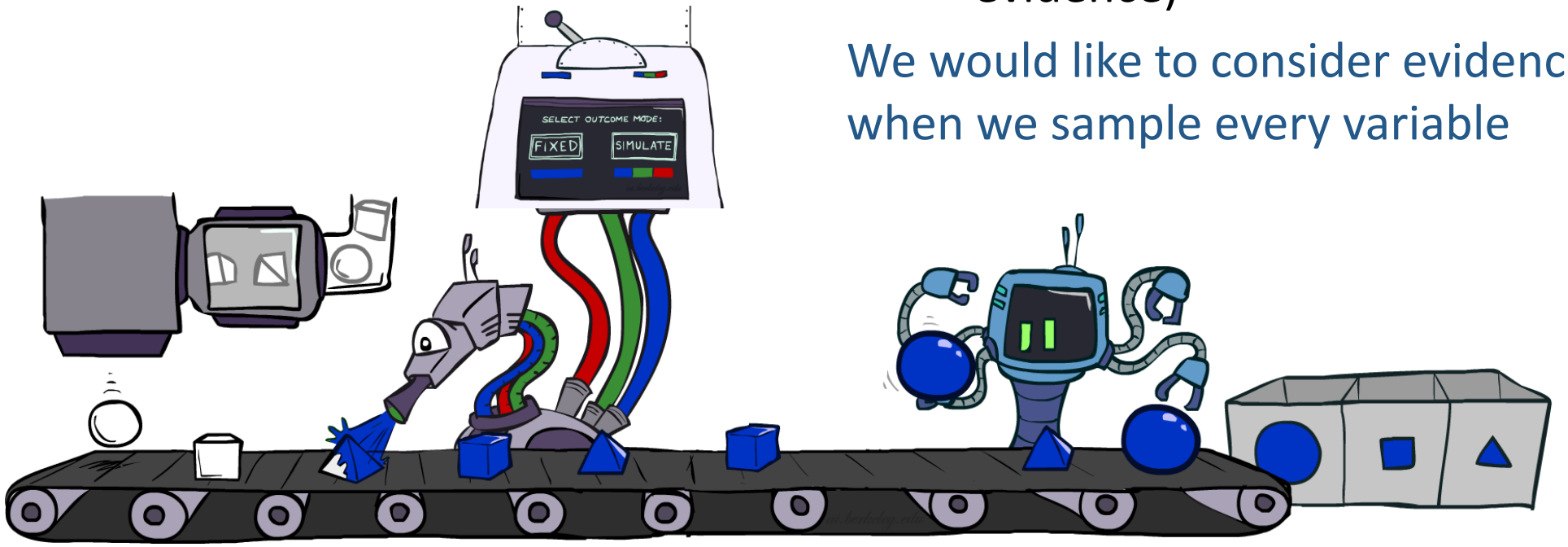
Likelihood weighting is good

- 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

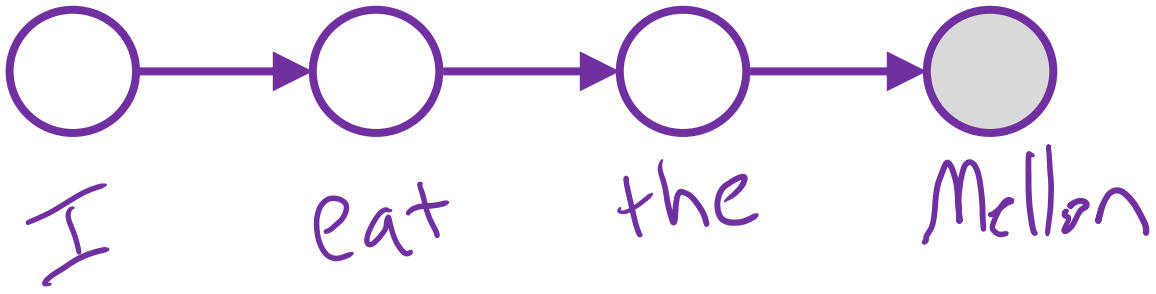
Likelihood weighting doesn't solve all our problems

- 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



$P(W_1 \mid W_4 = \text{"Mellon"})$

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

Out of scope: Further Reading on Gibbs Sampling

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

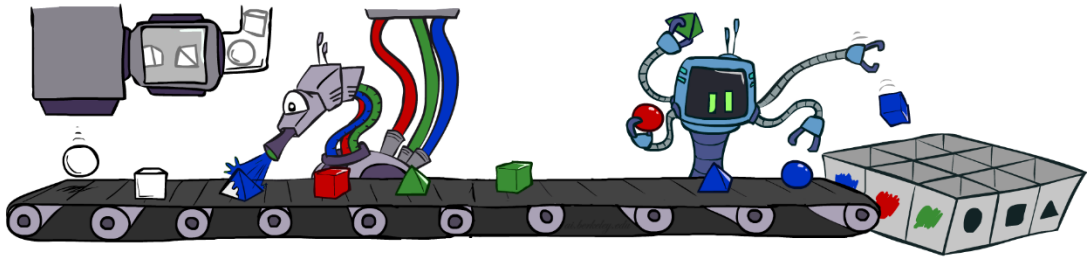
Gibbs sampling is a special case of more general methods called Markov chain Monte Carlo (MCMC) methods

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

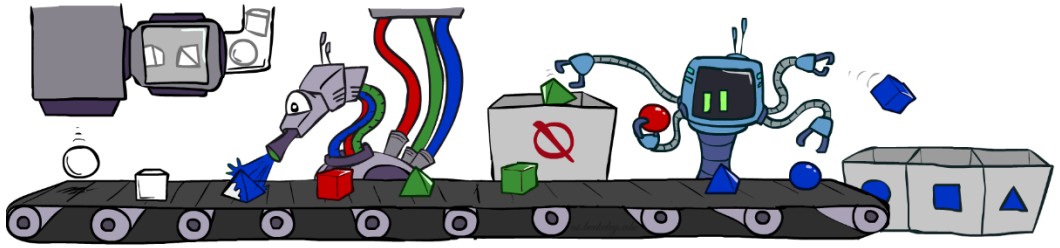
You may read about Monte Carlo methods – they're just sampling

Bayes Net Sampling Summary

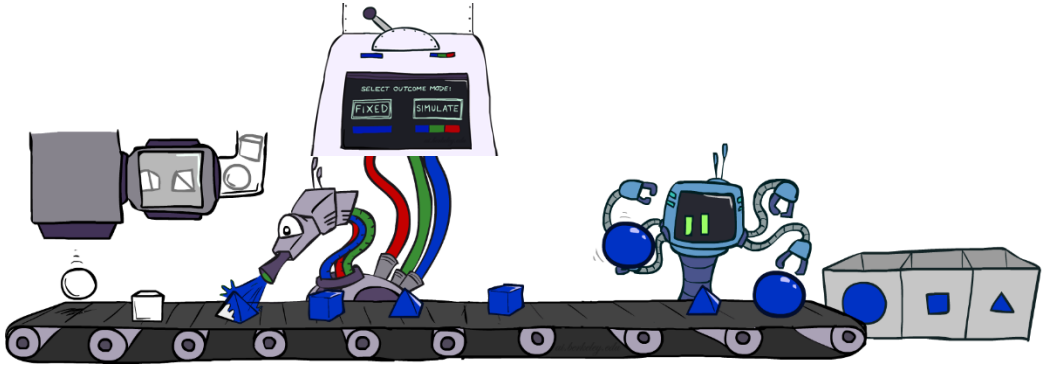
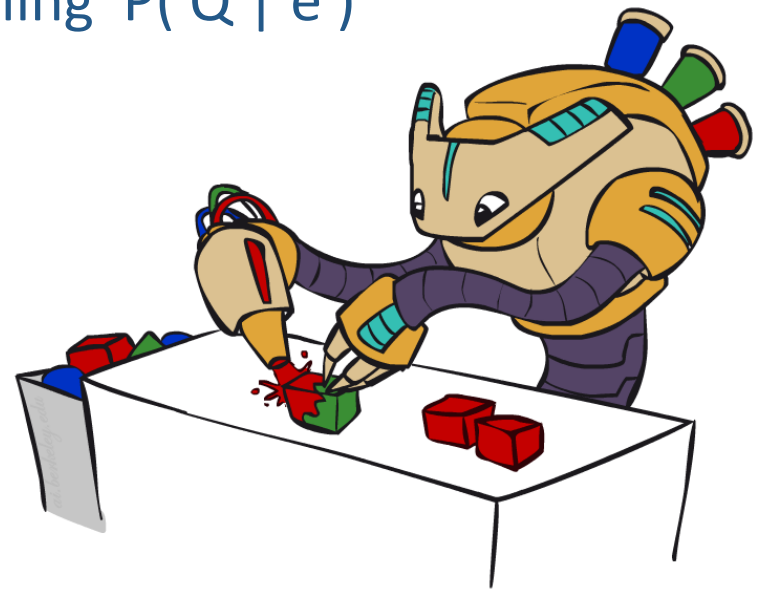
Prior Sampling P(Q, E)



Rejection Sampling P(Q | e)



Likelihood Weighting P(Q , e)



Gibbs Sampling P(Q | e)