

MAI 5100: Fundamentals of Artificial Intelligence

Instructor: Dr. Christopher Clarke

Overview & Agenda (week 11)

- **General:**
 - HW3 will be released on 06/02, due on 06/16.
 - Project final presentations will be on 06/21.
 - Final report due on 06/23.

Overview & Agenda (week 11)

- This week:
 - Review of probability
 - Random Variables
 - Joint and Marginal Distributions
 - Conditional Distributions
 - Product Rule, Chain Rule, Bayes Rule
 - Inference
 - Independence

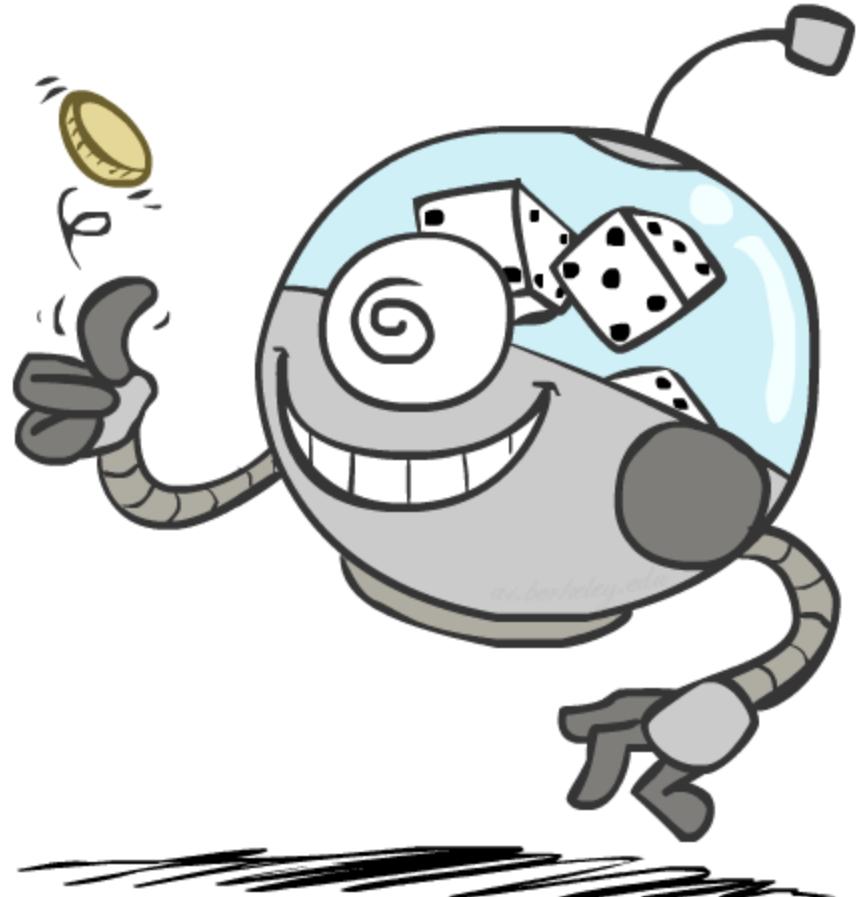
- This week:
 - Bayesian Networks
 - Definition
 - Structure and Parameters
 - Inference in Bayesian Networks
 - Applications of Bayesian Networks

Overview & Agenda (week 11)

- **This week:**
 - Markov Models
 - Markov Chains
 - Hidden Markov Models
 - Applications of HMMs

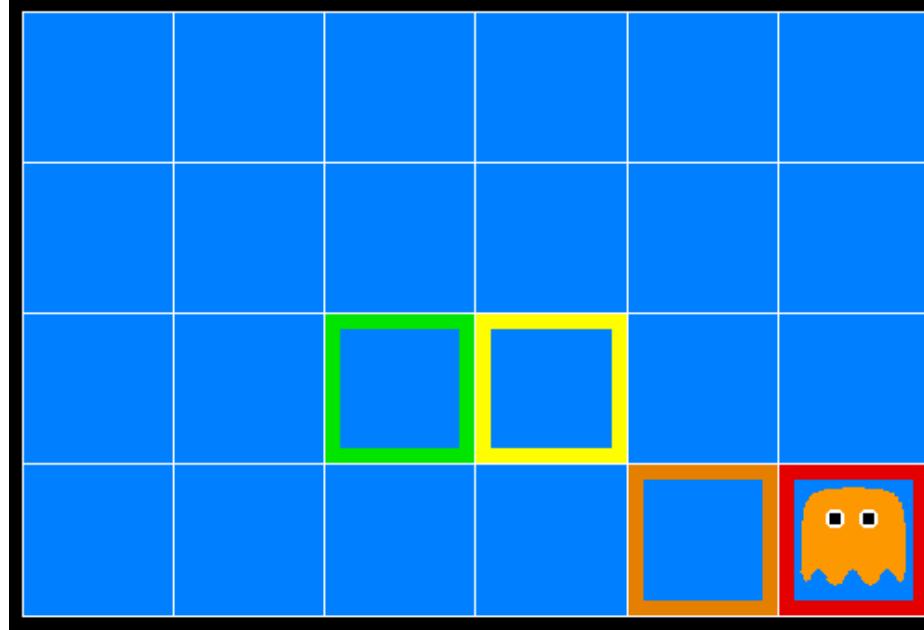
Today

- Probability
 - Random Variables
 - Joint and Marginal Distributions
 - Conditional Distribution
 - Product Rule, Chain Rule, Bayes' Rule
 - Inference
 - Independence
- You'll need all this stuff A LOT for the next few weeks, so make sure you go over it now!



Inference in Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
 - On the ghost: red
 - 1 or 2 away: orange
 - 3 or 4 away: yellow
 - 5+ away: green
- Sensors are noisy, but we know $P(\text{Color} \mid \text{Distance})$



$P(\text{red} \mid 3)$	$P(\text{orange} \mid 3)$	$P(\text{yellow} \mid 3)$	$P(\text{green} \mid 3)$
0.05	0.15	0.5	0.3

[Demo: Ghostbuster – no probability (L12D1)]

Uncertainty

- General situation:
 - **Observed variables (evidence):** Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
 - **Unobserved variables:** Agent needs to reason about other aspects (e.g. where an object is or what disease is present)
 - **Model:** Agent knows something about how the known variables relate to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

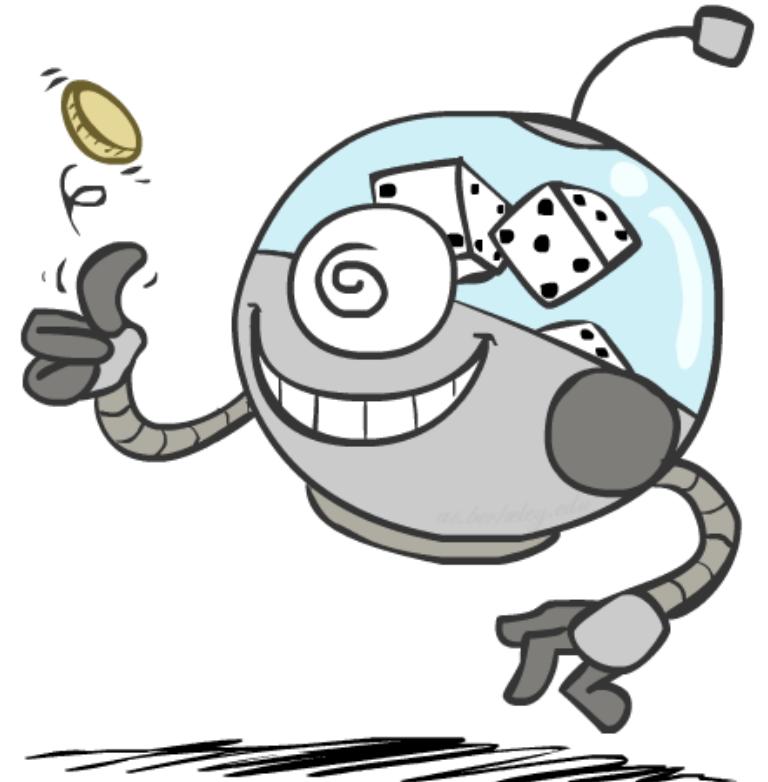
0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

0.17	0.10	0.10
0.09	0.17	0.10
<0.01	0.09	0.17

<0.01	<0.01	0.03
<0.01	0.05	0.05
<0.01	0.05	0.81

Random Variables

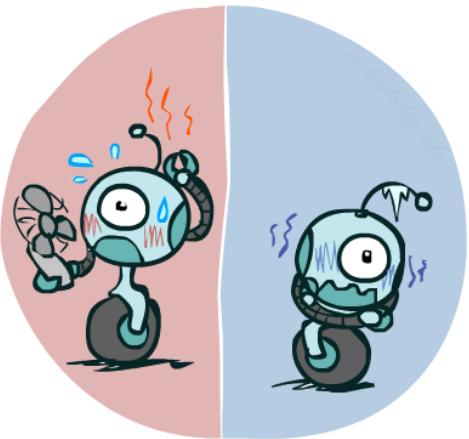
- A random variable is some aspect of the world about which we (may) have uncertainty
 - R = Is it raining?
 - T = Is it hot or cold?
 - D = How long will it take to drive to work?
 - L = Where is the ghost?
- We denote random variables with capital letters
- Like variables in a CSP, random variables have domains
 - R in {true, false} (often write as {+r, -r})
 - T in {hot, cold}
 - D in $[0, \infty)$
 - L in possible locations, maybe $\{(0,0), (0,1), \dots\}$



Probability Distributions

- Associate a probability with each value

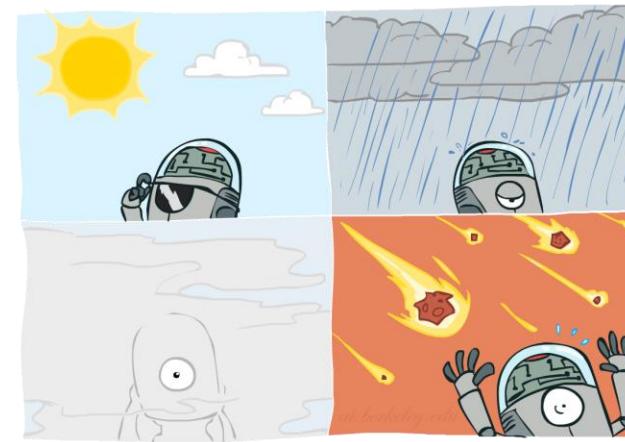
- Temperature:



$P(T)$

T	P
hot	0.5
cold	0.5

- Weather:



$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Probability Distributions

- Unobserved random variables have distributions

T	P
hot	0.5
cold	0.5

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

- A distribution is a TABLE of probabilities of values

- A probability (lower case value) is a single number

$$P(W = \text{rain}) = 0.1$$

- Must have: $\forall x \ P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Shorthand notation:

$$P(\text{hot}) = P(T = \text{hot}),$$

$$P(\text{cold}) = P(T = \text{cold}),$$

$$P(\text{rain}) = P(W = \text{rain}),$$

...

OK if all domain entries are unique

Joint Distributions

- A *joint distribution* over a set of random variables: X_1, X_2, \dots, X_n specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

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

- Must obey: $P(x_1, x_2, \dots, x_n) \geq 0$

$$\sum_{(x_1, x_2, \dots, x_n)} P(x_1, x_2, \dots, x_n) = 1$$

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- Size of distribution if n variables with domain sizes d ?
 - For all but the smallest distributions, impractical to write out!

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables
- Probabilistic models:
 - (Random) variables with domains
 - Assignments are called *outcomes*
 - Joint distributions: say whether assignments (outcomes) are likely
 - *Normalized*: sum to 1.0
 - Ideally: only certain variables directly interact
- Constraint satisfaction problems:
 - Variables with domains
 - Constraints: state whether assignments are possible
 - Ideally: only certain variables directly interact

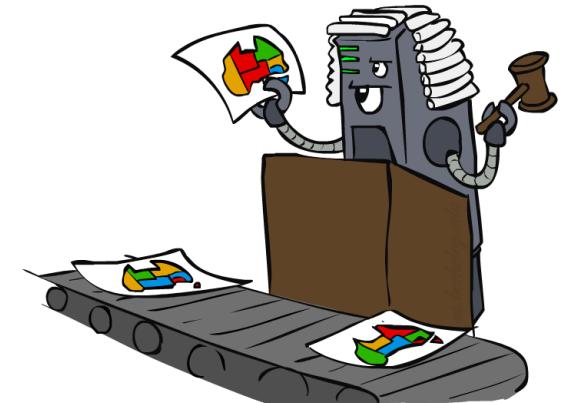
Distribution over T,W

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3



Constraint over T,W

T	W	P
hot	sun	T
hot	rain	F
cold	sun	F
cold	rain	T



Events

- An *event* is a set E of outcomes

$$P(E) = \sum_{(x_1 \dots x_n) \in E} P(x_1 \dots x_n)$$

- From a joint distribution, we can calculate the probability of any event

- Probability that it's hot AND sunny?
- Probability that it's hot?
- Probability that it's hot OR sunny?

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- Typically, the events we care about are *partial assignments*, like $P(T=\text{hot})$

Quiz: Events

- $P(+x, +y) ?$

$$P(X, Y)$$

- $P(+x) ?$

- $P(-y \text{ OR } +x) ?$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3



$$P(t) = \sum_s P(t, s)$$

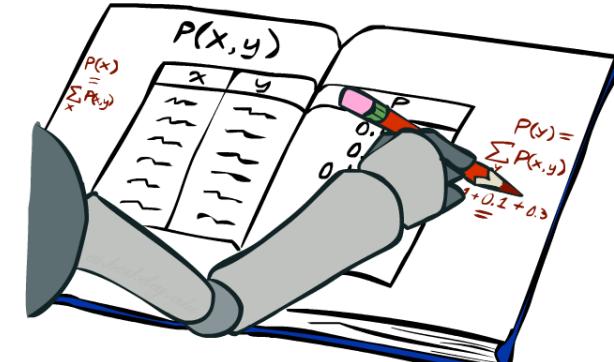


$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4



$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Quiz: Marginal Distributions

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

$$P(x) = \sum_y P(x, y)$$

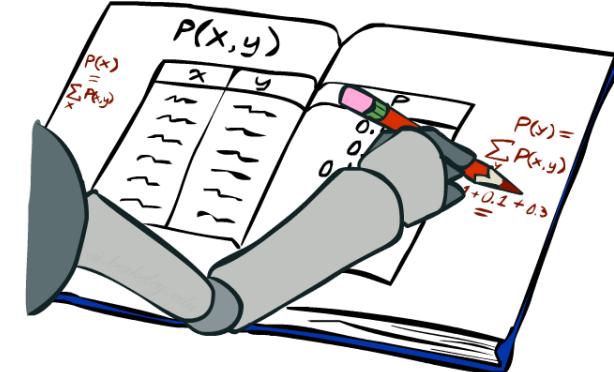
$$P(y) = \sum_x P(x, y)$$

$P(X)$

X	P
+x	
-x	

$P(Y)$

Y	P
+y	
-y	



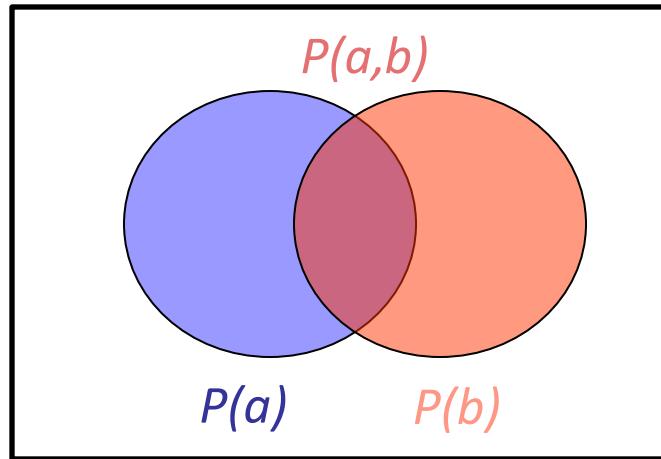
Conditional Probabilities

- A simple relation between joint and conditional probabilities
 - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3



$$P(W = s | T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$\begin{aligned} &= P(W = s, T = c) + P(W = r, T = c) \\ &= 0.2 + 0.3 = 0.5 \end{aligned}$$

Quiz: Conditional Probabilities

- $P(+x \mid +y) ?$

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

- $P(-x \mid +y) ?$

- $P(-y \mid +x) ?$

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$$P(W|T) \left[\begin{array}{l} P(W|T = hot) \\ \begin{array}{|c|c|} \hline W & P \\ \hline sun & 0.8 \\ \hline rain & 0.2 \\ \hline \end{array} \\ \\ P(W|T = cold) \\ \begin{array}{|c|c|} \hline W & P \\ \hline sun & 0.4 \\ \hline rain & 0.6 \\ \hline \end{array} \end{array} \right]$$

Joint Distribution

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Normalization Trick

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$\begin{aligned} P(W = s|T = c) &= \frac{P(W = s, T = c)}{P(T = c)} \\ &= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)} \\ &= \frac{0.2}{0.2 + 0.3} = 0.4 \end{aligned}$$



$$\begin{aligned} P(W = r|T = c) &= \frac{P(W = r, T = c)}{P(T = c)} \\ &= \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)} \\ &= \frac{0.3}{0.2 + 0.3} = 0.6 \end{aligned}$$

$P(W|T = c)$

W	P
sun	0.4
rain	0.6

Normalization Trick

$$\begin{aligned} P(W = s|T = c) &= \frac{P(W = s, T = c)}{P(T = c)} \\ &= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)} \\ &= \frac{0.2}{0.2 + 0.3} = 0.4 \end{aligned}$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

SELECT the joint probabilities matching the evidence
→

T	W	P
cold	sun	0.2
cold	rain	0.3

$P(c, W)$

NORMALIZE the selection
(make it sum to one)
→

W	P
sun	0.4
rain	0.6

$P(W|T = c)$

$$\begin{aligned} P(W = r|T = c) &= \frac{P(W = r, T = c)}{P(T = c)} \\ &= \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)} \\ &= \frac{0.3}{0.2 + 0.3} = 0.6 \end{aligned}$$

Normalization Trick

$P(T, W)$		
T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

SELECT the joint probabilities matching the evidence
→

$P(c, W)$		
T	W	P
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection
(make it sum to one)
→

$P(W T = c)$	
W	P
sun	0.4
rain	0.6

- Why does this work? Sum of selection is $P(\text{evidence})!$ ($P(T=c)$, here)

$$P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1, x_2)}{\sum_{x_1} P(x_1, x_2)}$$

Quiz: Normalization Trick

- $P(X | Y=-y) ?$

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

SELECT the joint
probabilities
matching the
evidence



NORMALIZE the
selection
(make it sum to one)



To Normalize

- (Dictionary) To bring or restore to a **normal condition**

All entries sum to ONE

- Procedure:

- Step 1: Compute $Z = \text{sum over all entries}$
- Step 2: Divide every entry by Z

- Example 1

W	P
sun	0.2
rain	0.3

Normalize  $Z = 0.5$

W	P
sun	0.4
rain	0.6

- Example 2

T	W	P
hot	sun	20
hot	rain	5
cold	sun	10
cold	rain	15

Normalize  $Z = 50$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)
- We generally compute conditional probabilities
 - $P(\text{on time} \mid \text{no reported accidents}) = 0.90$
 - These represent the agent's *beliefs* given the evidence
- Probabilities change with new evidence:
 - $P(\text{on time} \mid \text{no accidents, 5 a.m.}) = 0.95$
 - $P(\text{on time} \mid \text{no accidents, 5 a.m., raining}) = 0.80$
 - Observing new evidence causes *beliefs to be updated*



Inference by Enumeration

- General case:

- Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
- Query* variable: Q
- Hidden variables: $H_1 \dots H_r$

$$\left. \begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array} \right\} X_1, X_2, \dots, X_n$$

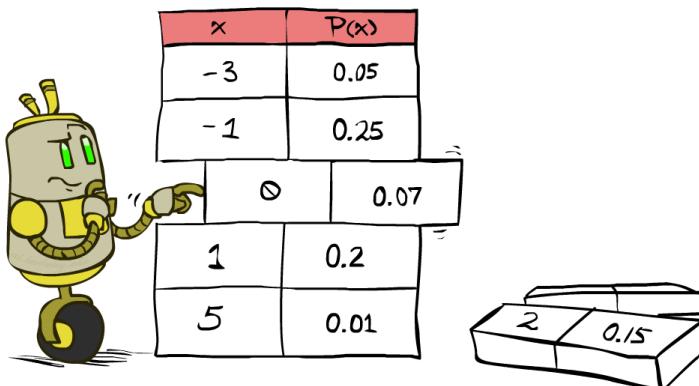
All variables

- We want:

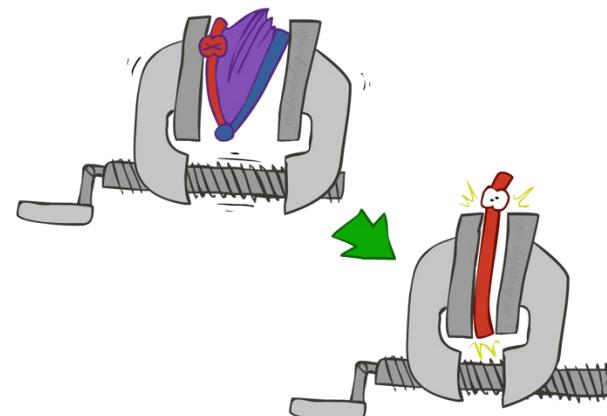
$$P(Q|e_1 \dots e_k)$$

* Works fine with multiple query variables, too

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \dots e_k)$$

$$P(Q|e_1 \dots e_k) = \frac{1}{Z} P(Q, e_1 \dots e_k)$$

$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Q, h_1 \dots h_r, e_1 \dots e_k)}_{X_1, X_2, \dots, X_n}$$

Inference by Enumeration

- $P(W)$?
- $P(W \mid \text{winter})$?
- $P(W \mid \text{winter, hot})$?

S	T	W	P
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

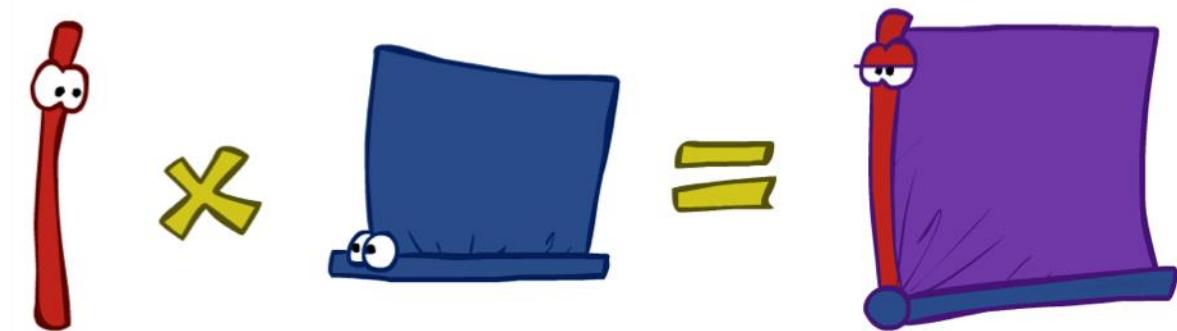
Inference by Enumeration

- Obvious problems:
 - Worst-case time complexity $O(d^n)$
 - Space complexity $O(d^n)$ to store the joint distribution

The Product Rule

- Sometimes have conditional distributions but want the joint

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



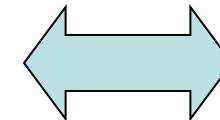
The Product Rule

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

- Example:

R	P
sun	0.8
rain	0.2

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3



$$P(D, W)$$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

The Chain Rule

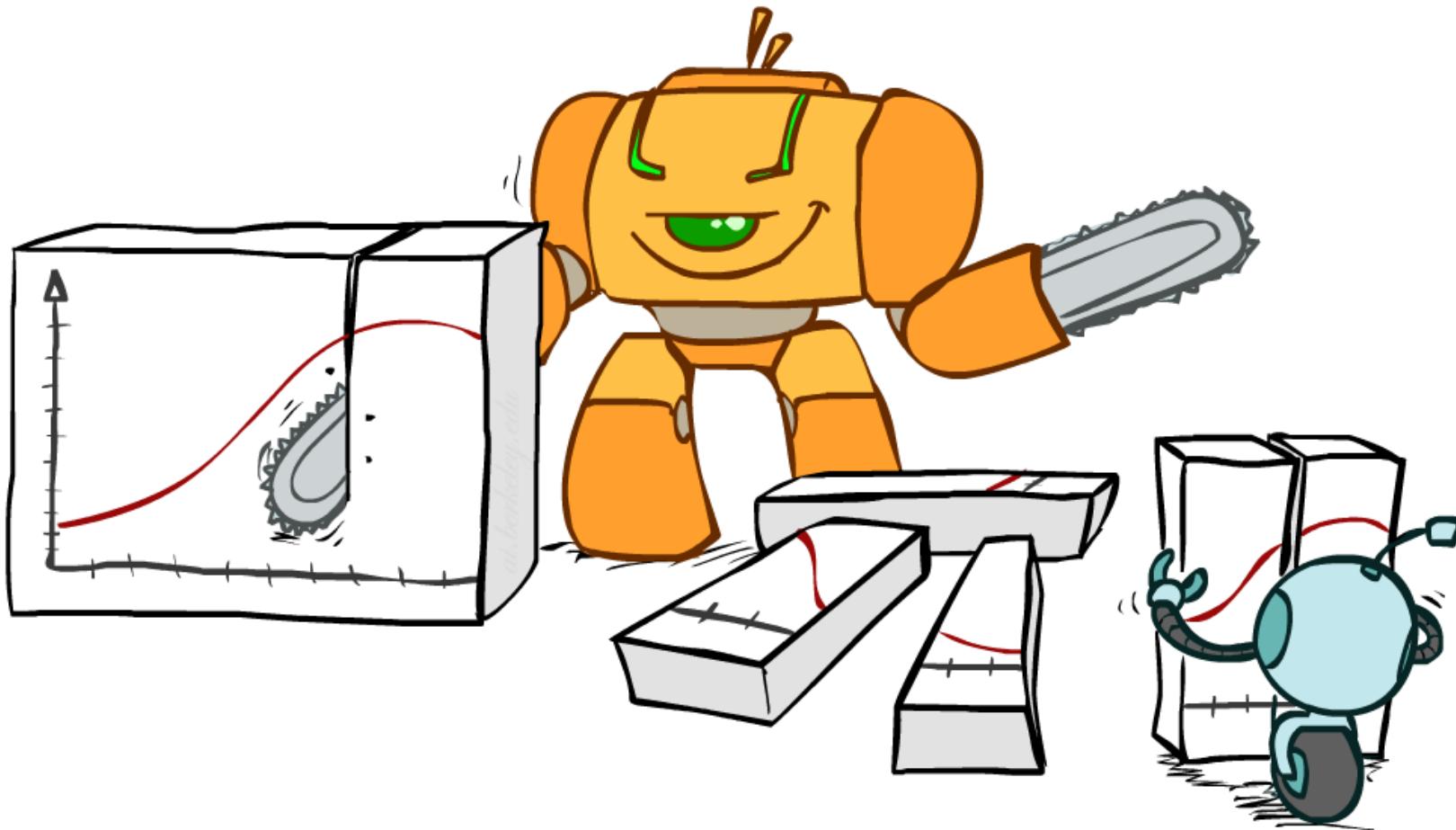
- More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i|x_1 \dots x_{i-1})$$

- Why is this always true?

Bayes Rule



Bayes' Rule

- Two ways to factor a joint distribution over two variables:

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

That's my rule!

- Dividing, we get:

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

- Why is this at all helpful?

- Lets us build one conditional from its reverse
- Often one conditional is tricky but the other one is simple
- Foundation of many systems we'll see later (e.g. ASR, MT)

- In the running for most important AI equation!



Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:

- M: meningitis, S: stiff neck

$$\left. \begin{array}{l} P(+m) = 0.0001 \\ P(+s|m) = 0.8 \\ P(+s|-m) = 0.01 \end{array} \right\} \text{Example givens}$$

$$P(+m|s) = \frac{P(+s|m)P(+m)}{P(+s)} = \frac{P(+s|m)P(+m)}{P(+s|m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

- Note: posterior probability of meningitis still very small
 - Note: you should still get stiff necks checked out! Why?

Quiz: Bayes' Rule

- Given:

$$P(W)$$

R	P
sun	0.8
rain	0.2

$$P(D|W)$$

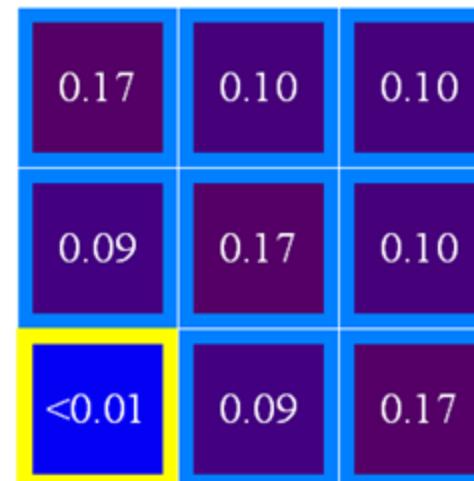
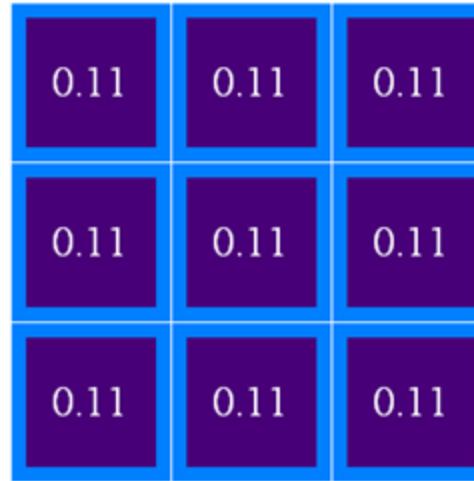
D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

- What is $P(W | \text{dry})$?

Ghostbusters, Revisited

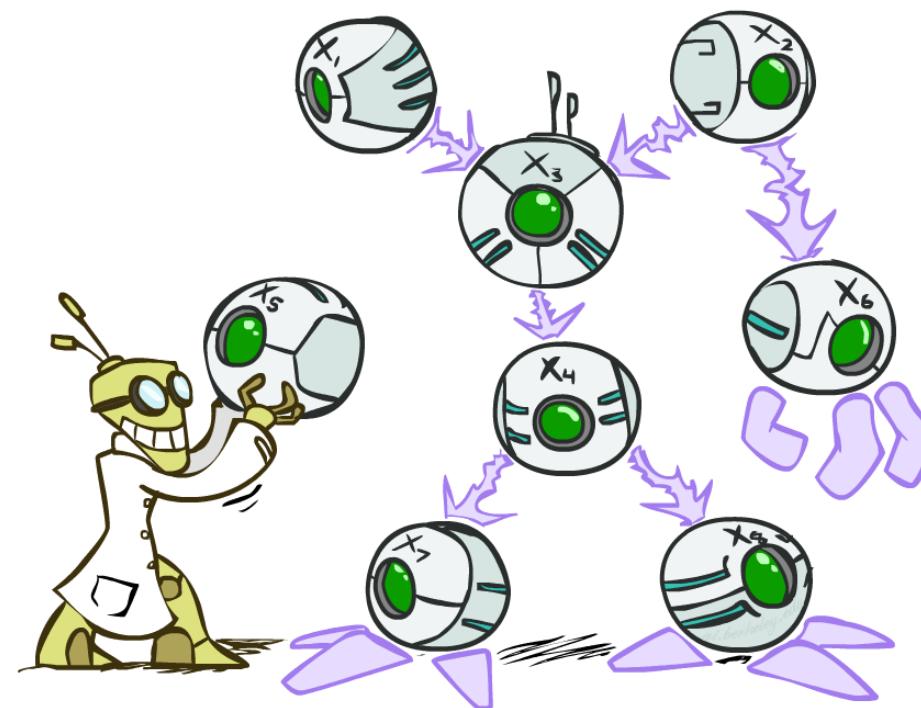
- Let's say we have two distributions:
 - Prior distribution over ghost location: $P(G)$
 - Let's say this is uniform
 - Sensor reading model: $P(R | G)$
 - Given: we know what our sensors do
 - R = reading color measured at (1,1)
 - E.g. $P(R = \text{yellow} | G=(1,1)) = 0.1$
- We can calculate the posterior distribution $P(G|r)$ over ghost locations given a reading using Bayes' rule:

$$P(g|r) \propto P(r|g)P(g)$$



[Demo: Ghostbuster – with probability (L12D2)]

Bayes' Nets

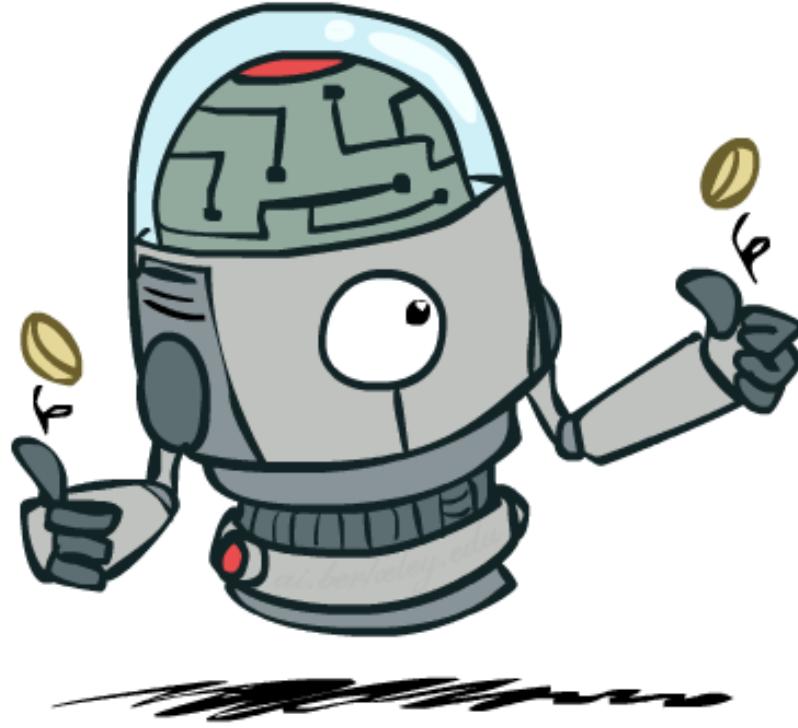


Probabilistic Models

- Models describe how (a portion of) the world works
- **Models are always simplifications**
 - May not account for every variable
 - May not account for all interactions between variables
 - “All models are wrong; but some are useful.”
 - George E. P. Box
- What do we do with probabilistic models?
 - We (or our agents) need to reason about unknown variables, given evidence
 - Example: explanation (diagnostic reasoning)
 - Example: prediction (causal reasoning)
 - Example: value of information



Independence



Independence

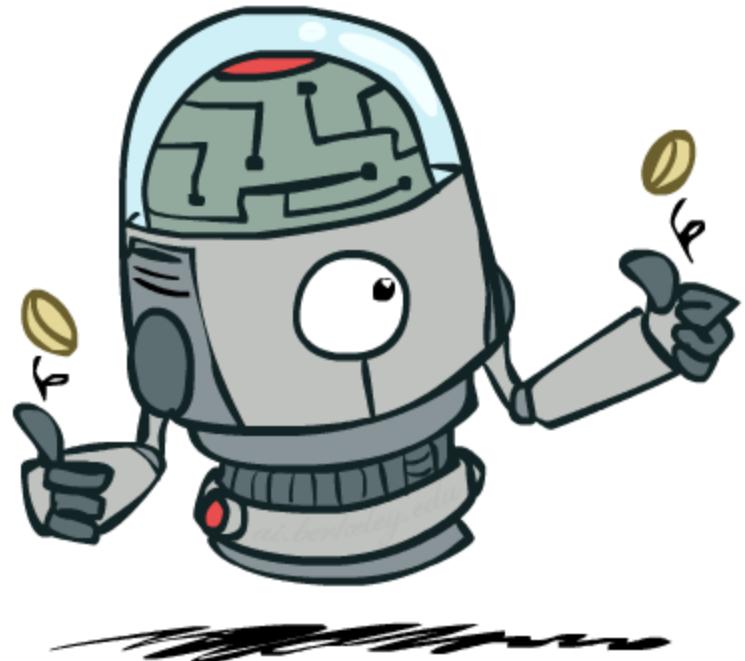
- Two variables are *independent* if:

$$\forall x, y : P(x, y) = P(x)P(y)$$

- This says that their joint distribution *factors* into a product two simpler distributions
- Another form:

$$\forall x, y : P(x|y) = P(x)$$

- We write: $X \perp\!\!\!\perp Y$
- Independence is a simplifying *modeling assumption*
 - *Empirical* joint distributions: at best “close” to independent
 - What could we assume for {Weather, Traffic, Cavity, Toothache}?



Example: Independence?

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(T)$

T	P
hot	0.5
cold	0.5

$P_2(T, W)$

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

$P(W)$

W	P
sun	0.6
rain	0.4

Example: Independence

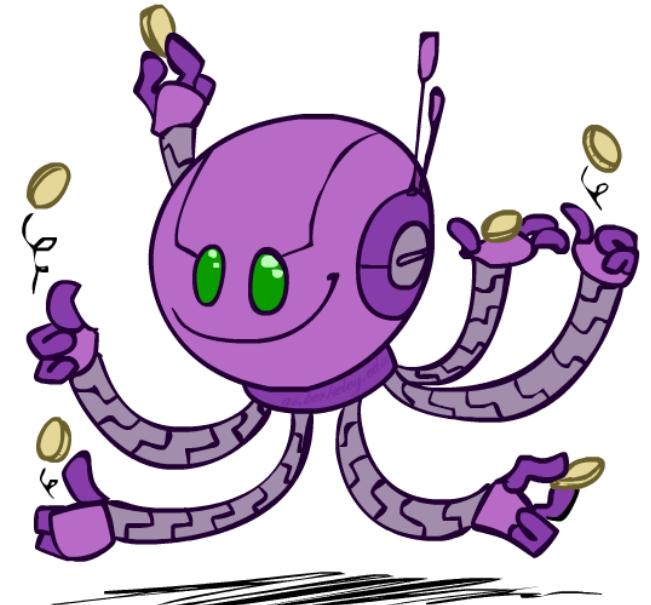
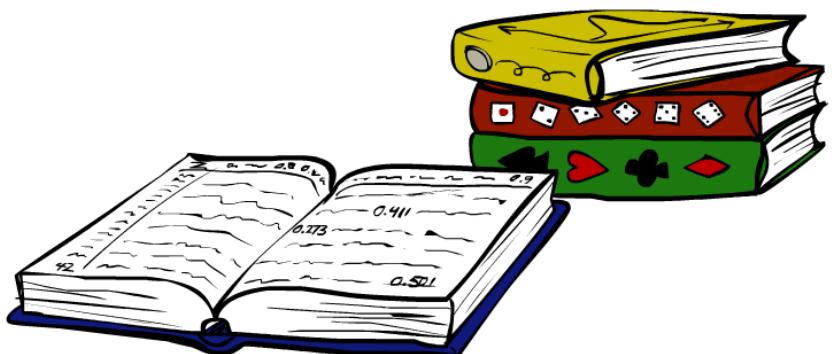
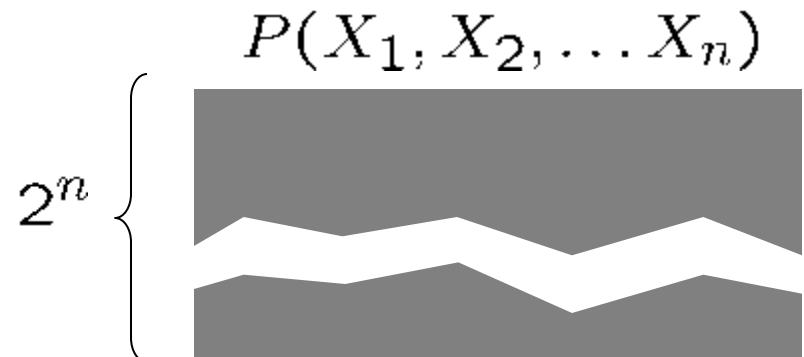
- N fair, independent coin flips:

$P(X_1)$	
H	0.5
T	0.5

$P(X_2)$	
H	0.5
T	0.5

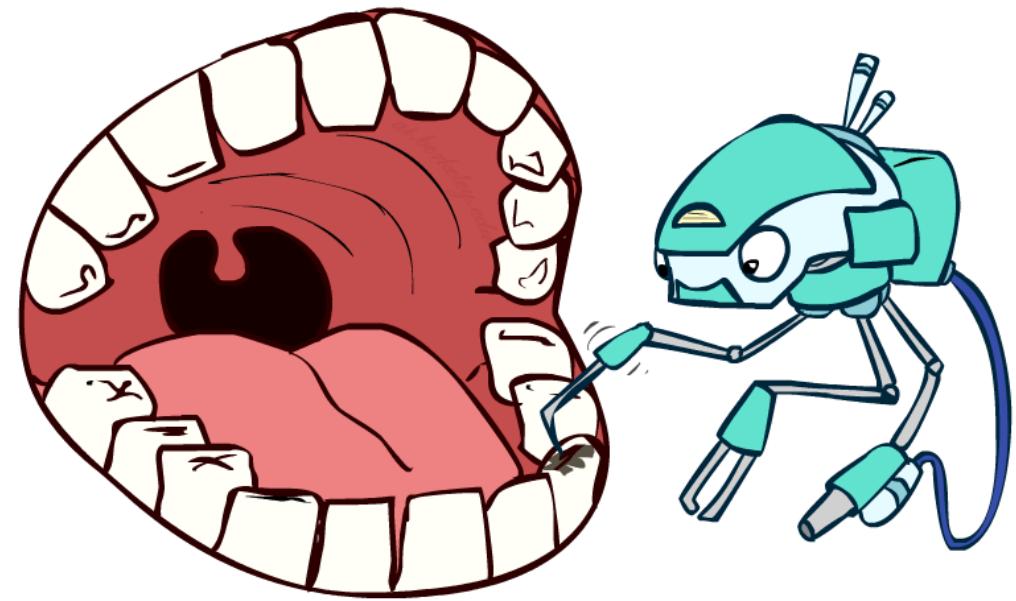
...

$P(X_n)$	
H	0.5
T	0.5



Conditional Independence

- $P(\text{Toothache}, \text{Cavity}, \text{Catch})$
- If I have a cavity, the probability that the probe catches it doesn't depend on whether I have a toothache:
 - $P(+\text{catch} | +\text{toothache}, +\text{cavity}) = P(+\text{catch} | +\text{cavity})$
- The same independence holds if I don't have a cavity:
 - $P(+\text{catch} | +\text{toothache}, -\text{cavity}) = P(+\text{catch} | -\text{cavity})$
- Catch is *conditionally independent* of Toothache given Cavity:
 - $P(\text{Catch} | \text{Toothache}, \text{Cavity}) = P(\text{Catch} | \text{Cavity})$
- Equivalent statements:
 - $P(\text{Toothache} | \text{Catch}, \text{Cavity}) = P(\text{Toothache} | \text{Cavity})$
 - $P(\text{Toothache}, \text{Catch} | \text{Cavity}) = P(\text{Toothache} | \text{Cavity}) P(\text{Catch} | \text{Cavity})$
 - One can be derived from the other easily



Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z $X \perp\!\!\!\perp Y | Z$

if and only if:

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

Conditional Independence

- What about this domain:

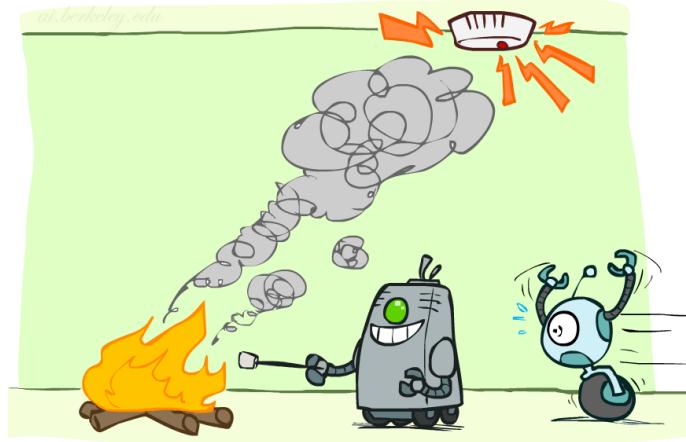
- Traffic
- Umbrella
- Raining



Conditional Independence

- What about this domain:

- Fire
- Smoke
- Alarm



Conditional Independence and the Chain Rule

- Chain rule:

$$P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \dots$$

- Trivial decomposition:

$$P(\text{Traffic, Rain, Umbrella}) =$$

$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain, Traffic})$$

- With assumption of conditional independence:

$$P(\text{Traffic, Rain, Umbrella}) =$$

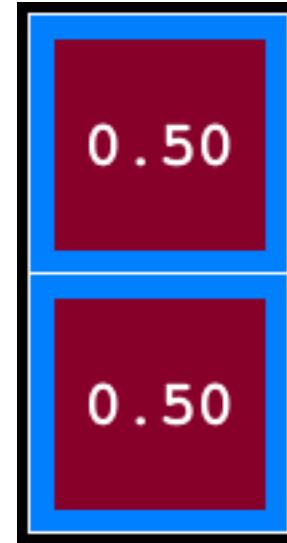
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$



- Bayes' nets / graphical models help us express conditional independence assumptions

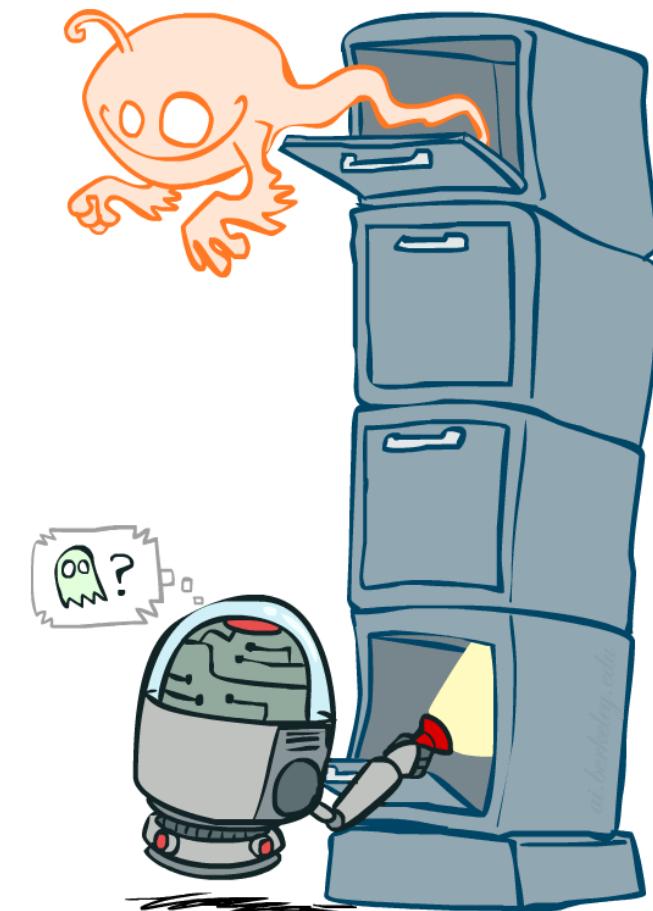
Ghostbusters Chain Rule

- Each sensor depends only on where the ghost is
- That means, the two sensors are conditionally independent, given the ghost position
- T: Top square is red
B: Bottom square is red
G: Ghost is in the top
- Givens:
 $P(+g) = 0.5$
 $P(-g) = 0.5$
 $P(+t | +g) = 0.8$
 $P(+t | -g) = 0.4$
 $P(+b | +g) = 0.4$
 $P(+b | -g) = 0.8$

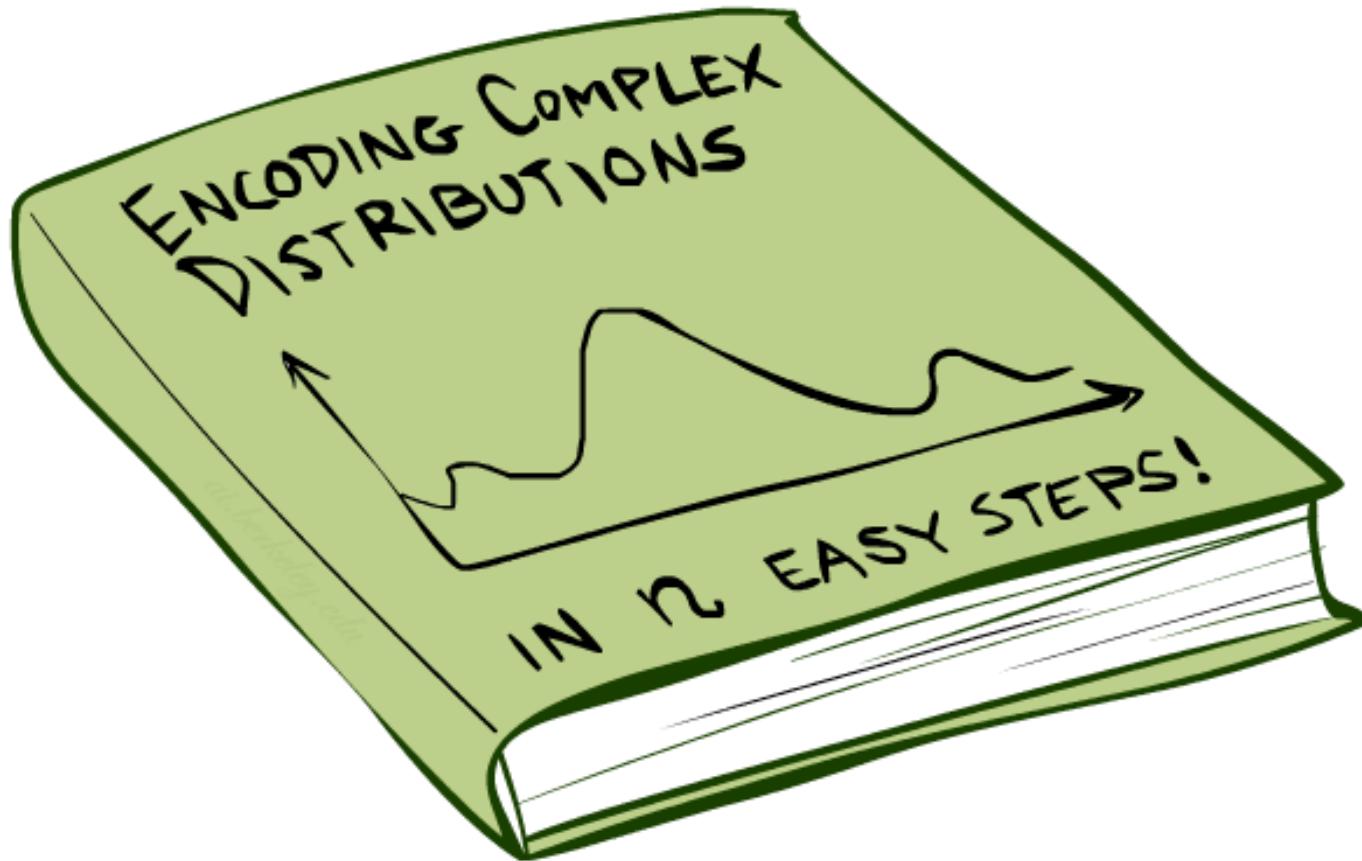


$$P(T, B, G) = P(G) P(T|G) P(B|G)$$

T	B	G	$P(T, B, G)$
+t	+b	+g	0.16
+t	+b	-g	0.16
+t	-b	+g	0.24
+t	-b	-g	0.04
-t	+b	+g	0.04
-t	+b	-g	0.24
-t	-b	+g	0.06
-t	-b	-g	0.06

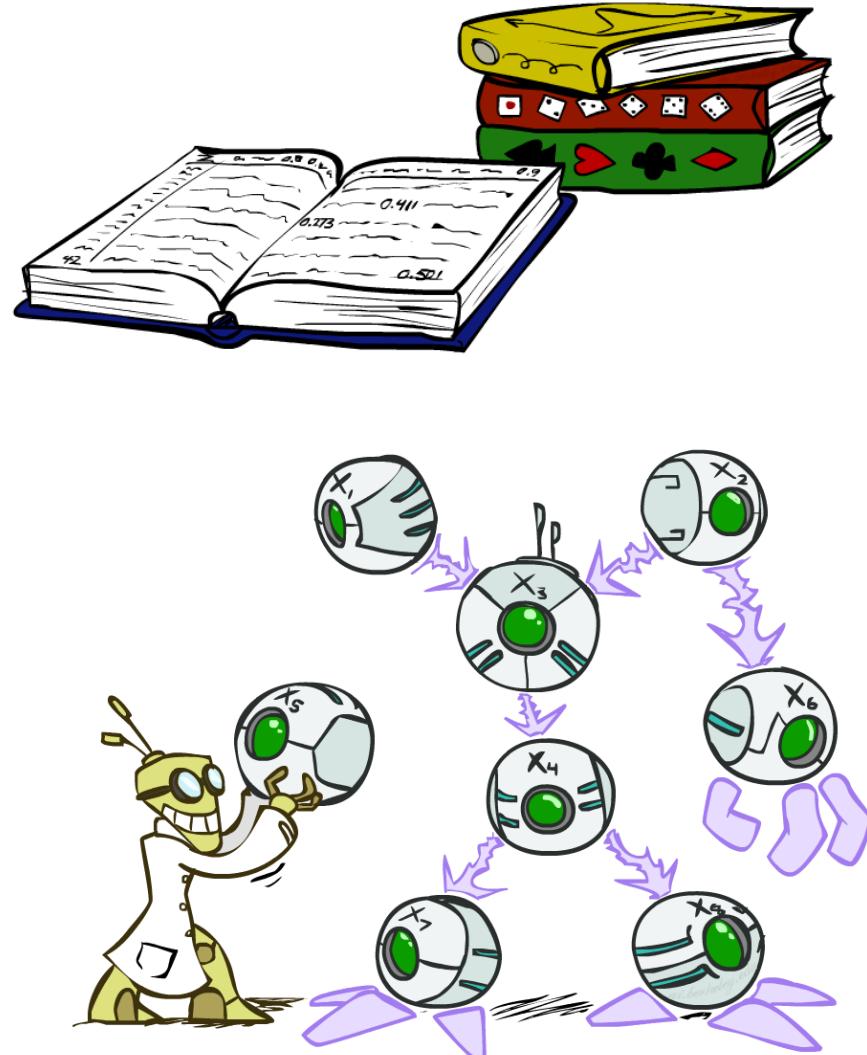


Bayes'Nets: Big Picture

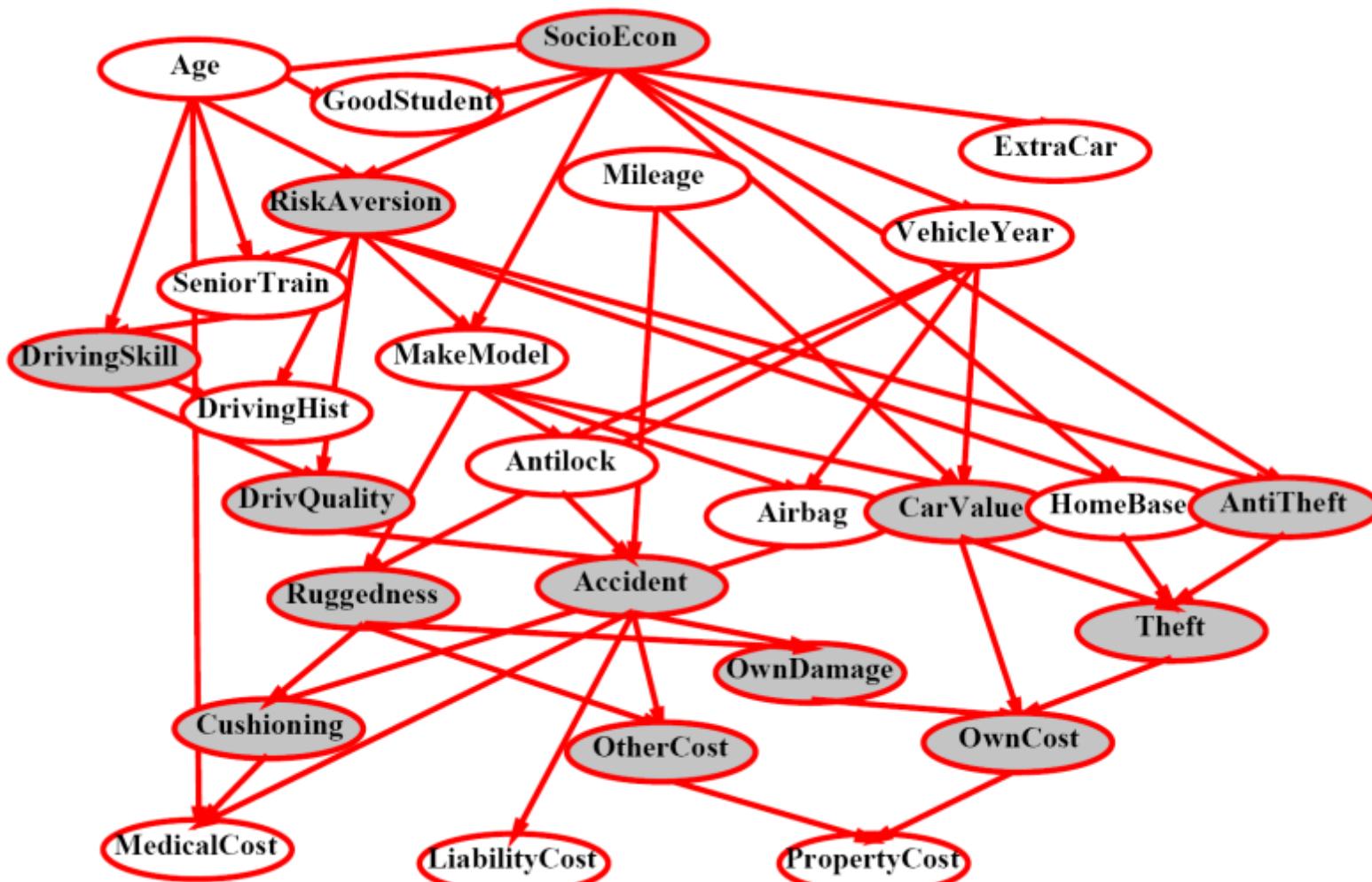


Bayes' Nets: Big Picture

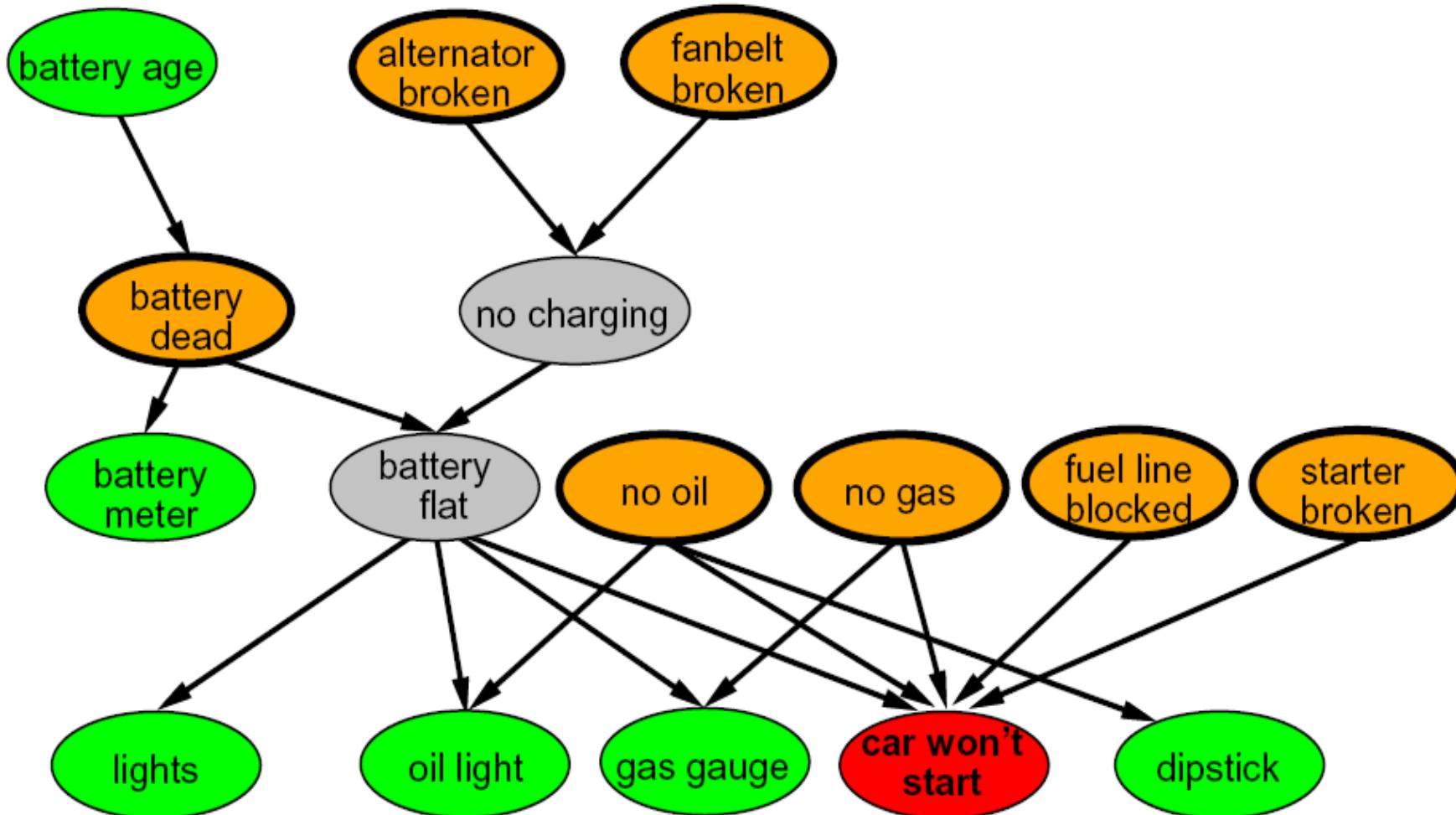
- Two problems with using full joint distribution tables as our probabilistic models:
 - Unless there are only a few variables, the joint is WAY too big to represent explicitly
 - Hard to learn (estimate) anything empirically about more than a few variables at a time
- **Bayes' nets:** a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
 - More properly called **graphical models**
 - We describe how variables locally interact
 - Local interactions chain together to give global, indirect interactions
 - For about 10 min, we'll be vague about how these interactions are specified



Example Bayes' Net: Insurance

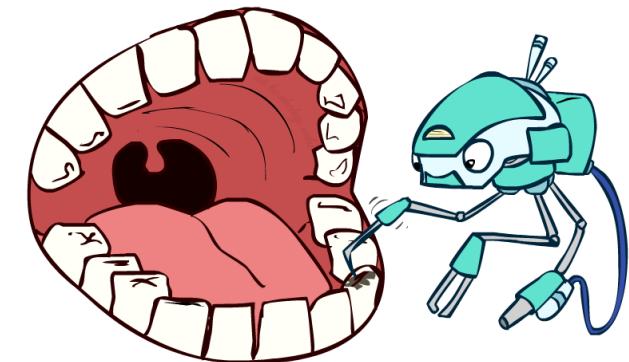
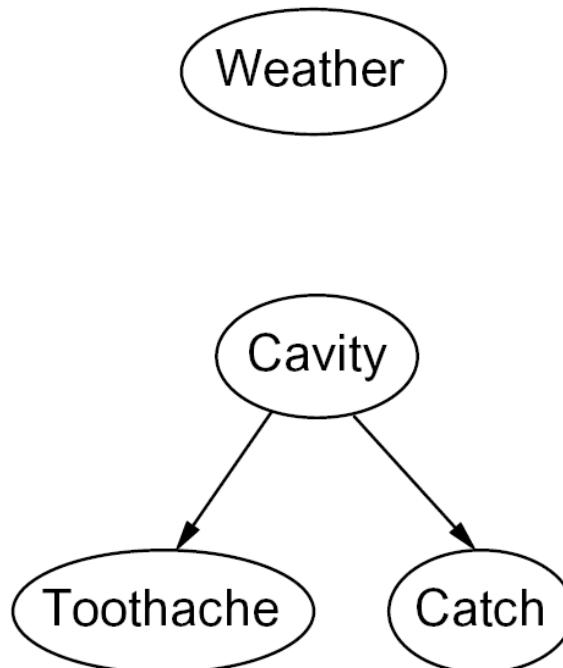


Example Bayes' Net: Car



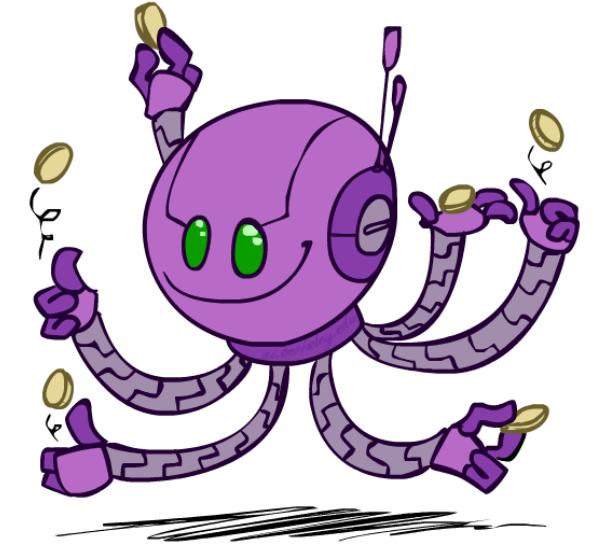
Graphical Model Notation

- Nodes: variables (with domains)
 - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
 - Similar to CSP constraints
 - Indicate “direct influence” between variables
 - Formally: encode conditional independence (more later)
- For now: imagine that arrows mean direct causation (in general, they don’t!)



Example: Coin Flips

- N independent coin flips



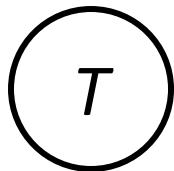
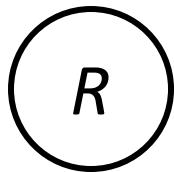
- No interactions between variables: **absolute independence**

Example: Traffic

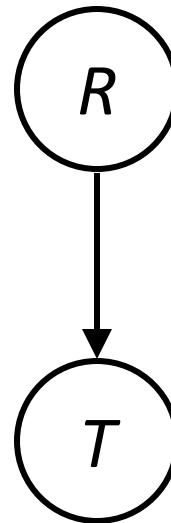
- Variables:
 - R : It rains
 - T : There is traffic



- Model 1: independence



- Model 2: rain causes traffic



- Why is an agent using model 2 better?

Example: Traffic II

- Let's build a causal graphical model!

- Variables

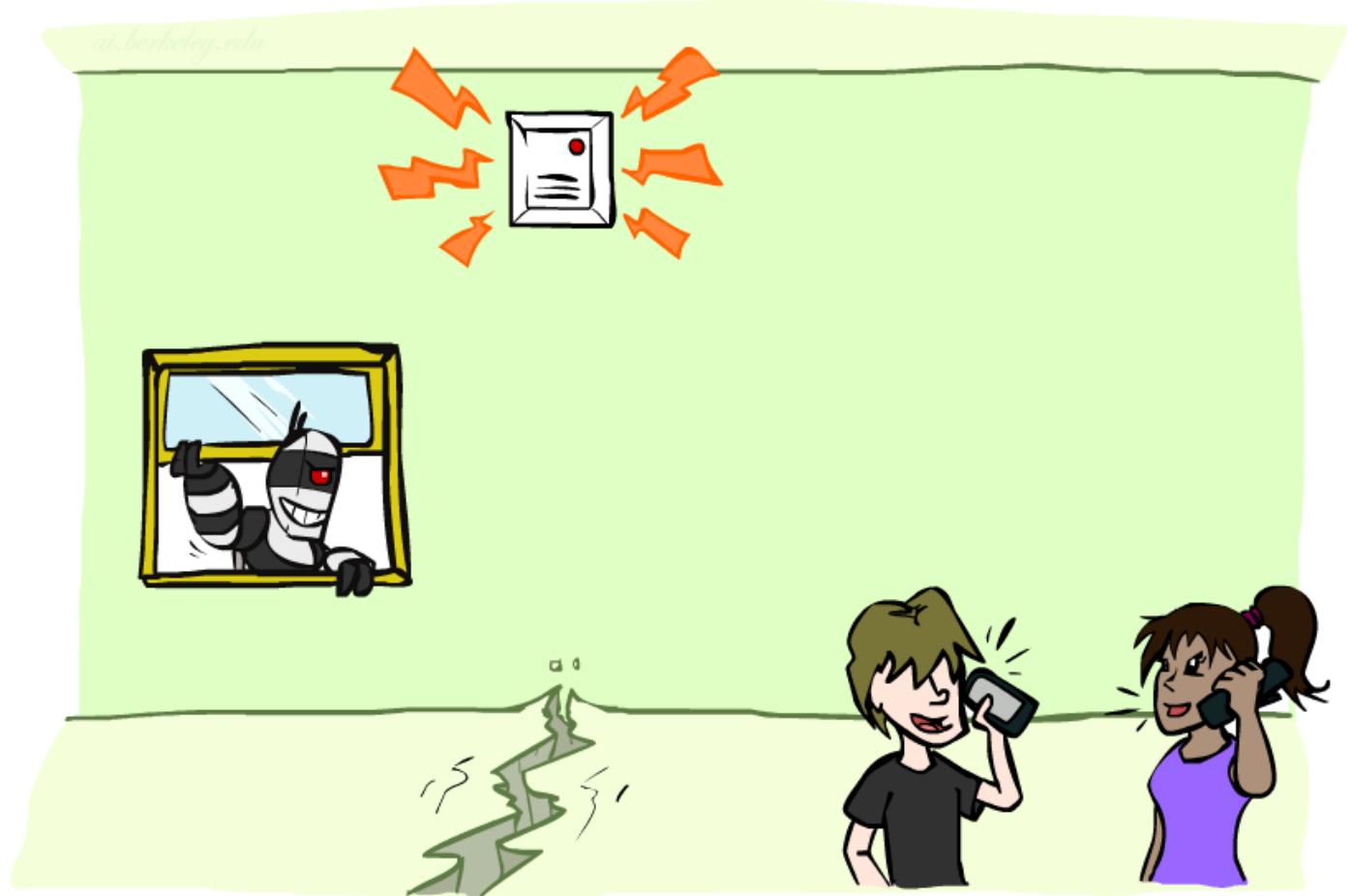
- T: Traffic
- R: It rains
- L: Low pressure
- D: Roof drips
- B: Ballgame
- C: Cavity



Example: Alarm Network

■ Variables

- B: Burglary
- A: Alarm goes off
- M: Mary calls
- J: John calls
- E: Earthquake!



Bayes' Net Semantics



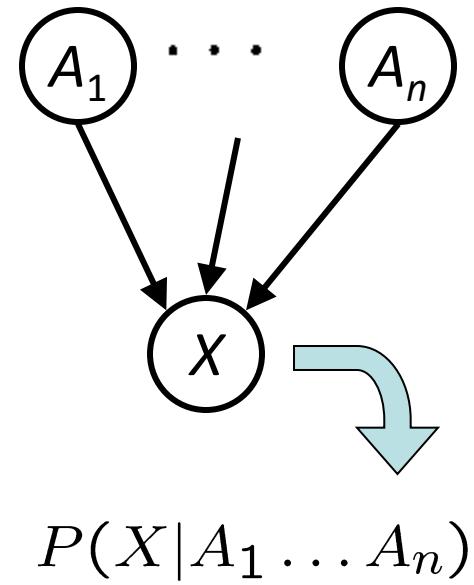


Bayes' Net Semantics

- A set of nodes, one per variable X
- A directed, acyclic graph
- A conditional distribution for each node
 - A collection of distributions over X , one for each combination of parents' values

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

- CPT: conditional probability table
- Description of a noisy “causal” process



$$P(X|A_1 \dots A_n)$$

A Bayes net = Topology (graph) + Local Conditional Probabilities

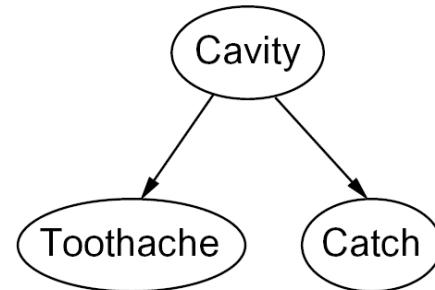
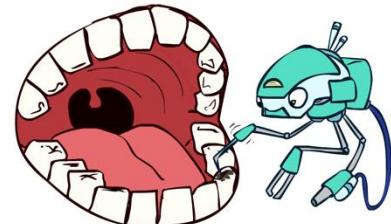
Probabilities in BNs



- 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:

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

- Example:



$P(+\text{cavity}, +\text{catch}, -\text{toothache})$

Probabilities in BNs



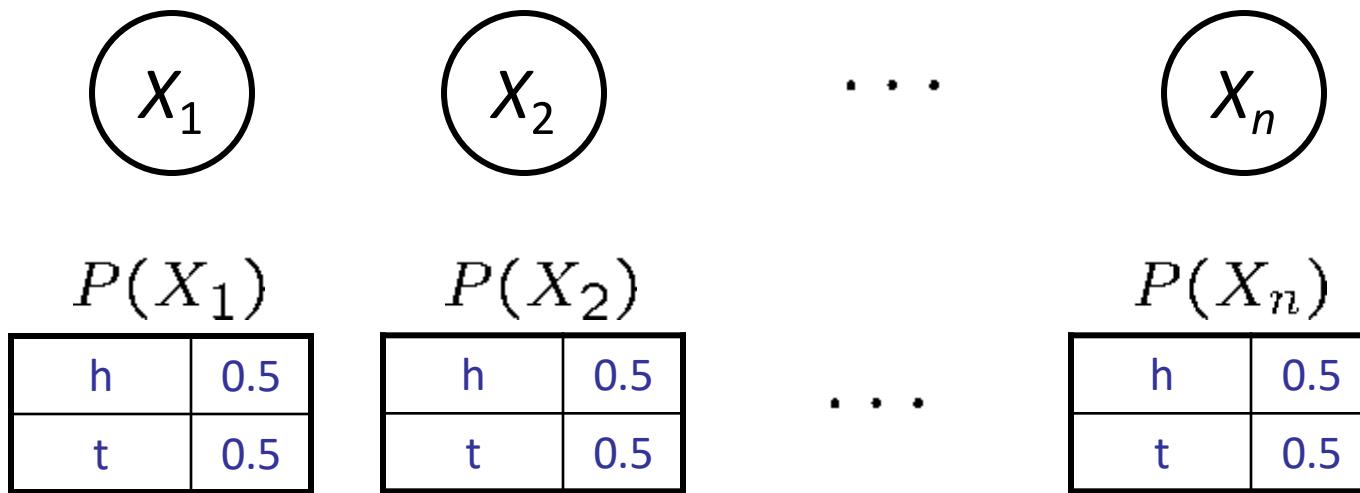
- Why are we guaranteed that setting

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

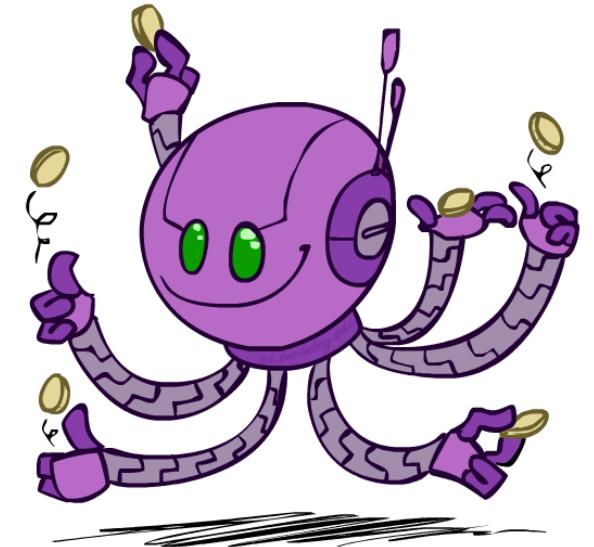
results in a proper joint distribution?

- Chain rule (valid for all distributions): $P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | x_1 \dots x_{i-1})$
- Assume conditional independences: $P(x_i | x_1, \dots, x_{i-1}) = P(x_i | \text{parents}(X_i))$
→ Consequence: $P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parents}(X_i))$
- Not every BN can represent every joint distribution
 - The topology enforces certain conditional independencies

Example: Coin Flips

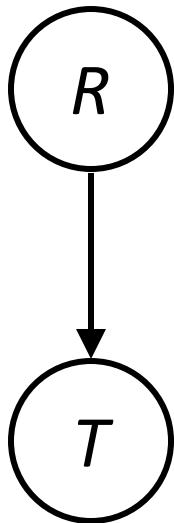


$$P(h, h, t, h) =$$



Only distributions whose variables are absolutely independent can be represented by a Bayes' net with no arcs.

Example: Traffic



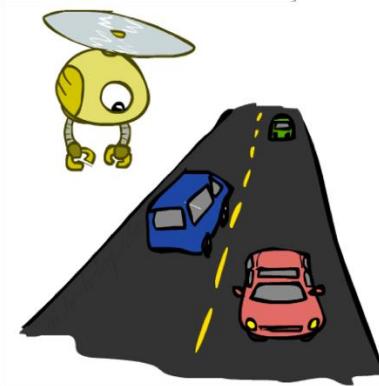
$P(R)$

+r	1/4
-r	3/4

$P(+r, -t) =$

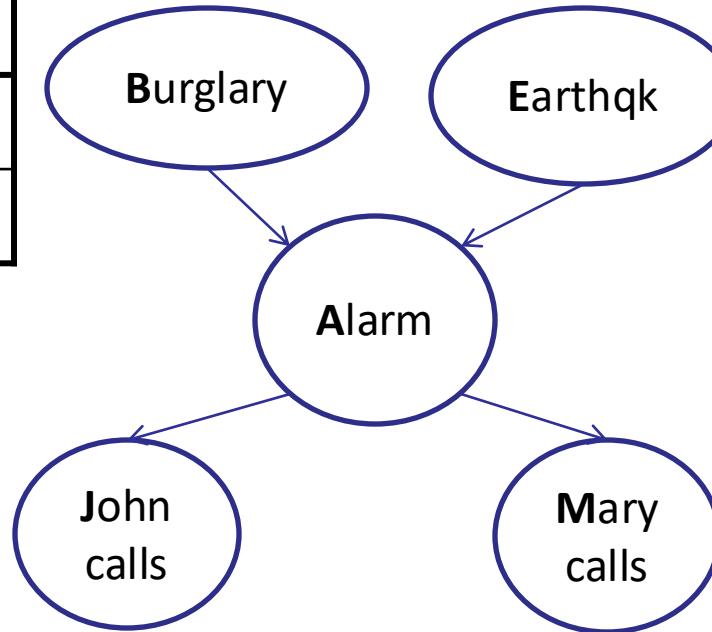
$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2



Example: Alarm Network

B	$P(B)$
+b	0.001
-b	0.999



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

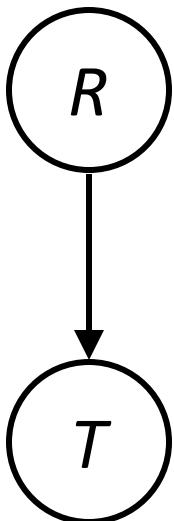
E	$P(E)$
+e	0.002
-e	0.998



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: Traffic

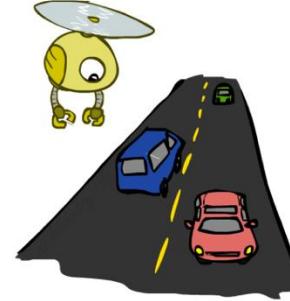
- Causal direction

 $P(R)$

+r	1/4
-r	3/4

 $P(T|R)$

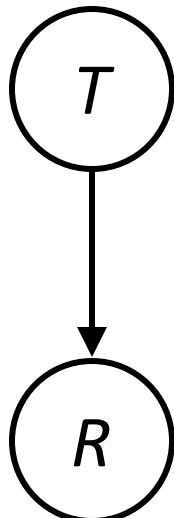
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

 $P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Example: Reverse Traffic

- Reverse causality?



$P(T)$

+t	9/16
-t	7/16

$P(R|T)$

+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7



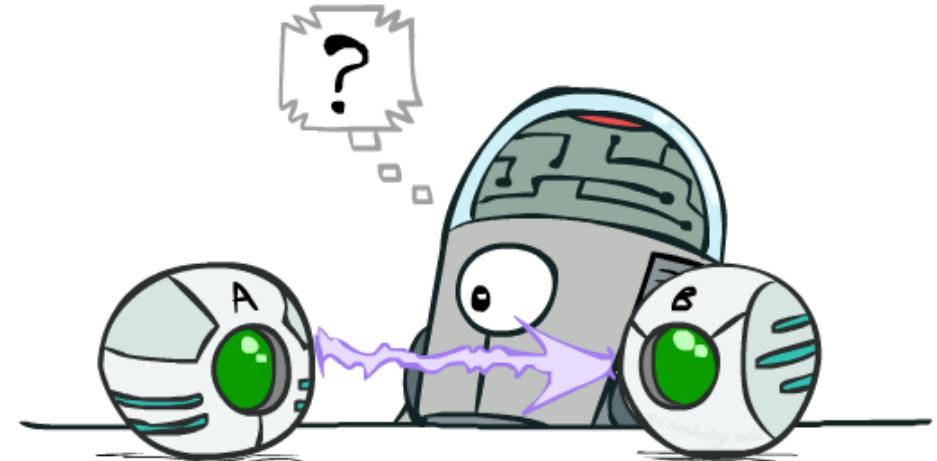
$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Causality?

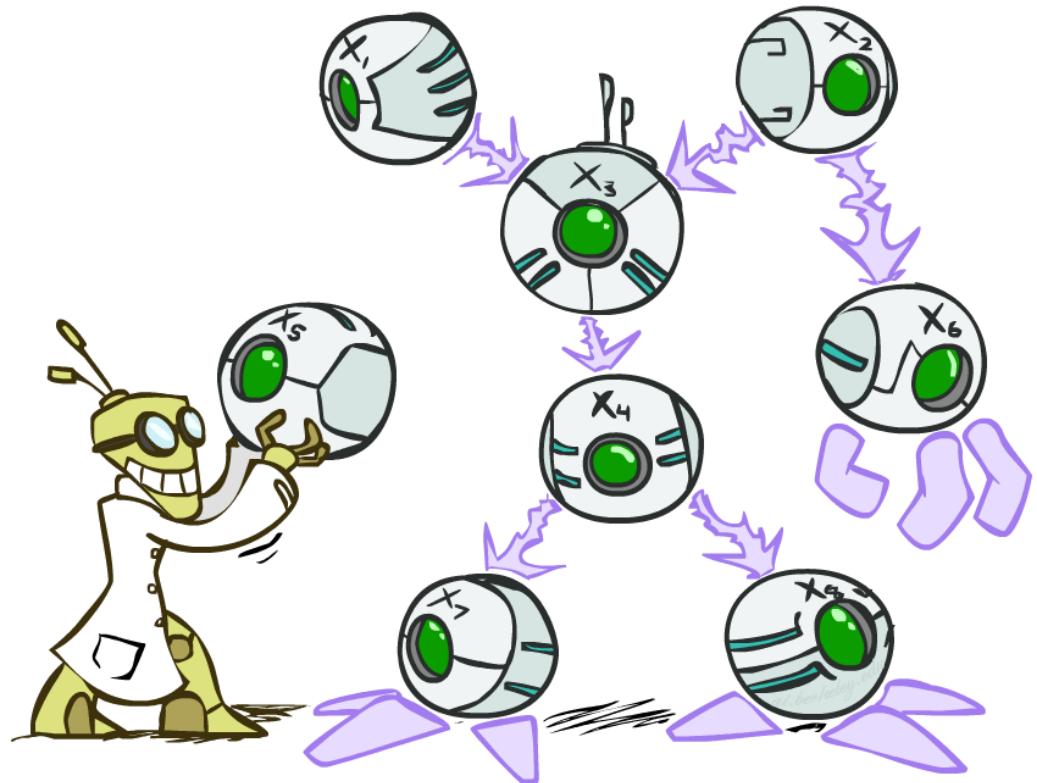
- When Bayes' nets reflect the true causal patterns:
 - Often simpler (nodes have fewer parents)
 - Often easier to think about
 - Often easier to elicit from experts
- BNs need not actually be causal
 - Sometimes no causal net exists over the domain (especially if variables are missing)
 - E.g. consider the variables *Traffic* and *Drips*
 - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
 - Topology may happen to encode causal structure
 - **Topology really encodes conditional independence**

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

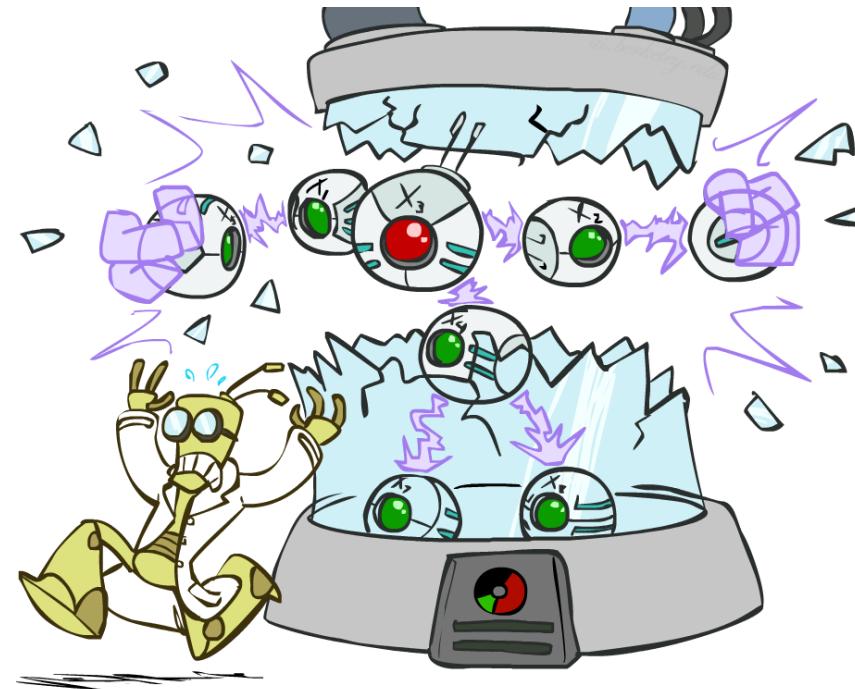


Bayes' Nets

- So far: how a Bayes' net encodes a joint distribution
- Next: how to answer queries about that distribution
 - Today:
 - First assembled BNs using an intuitive notion of conditional independence as causality
 - Then saw that key property is conditional independence
 - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)



Bayes' Nets: Independence

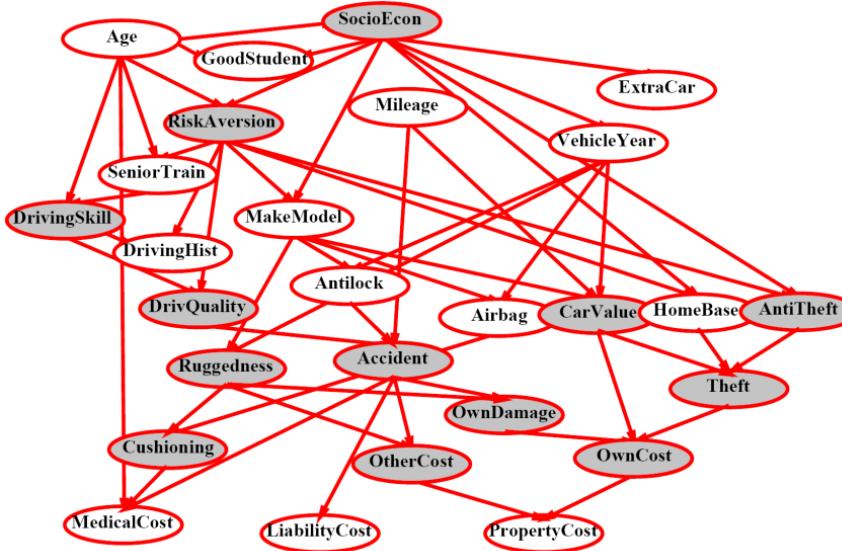


Bayes' Nets

- A Bayes' net is an efficient encoding of a probabilistic model of a domain

- Questions we can ask:

- Inference: given a fixed BN, what is $P(X | e)$?
- Representation: given a BN graph, what kinds of distributions can it encode?
- Modeling: what BN is most appropriate for a given domain?



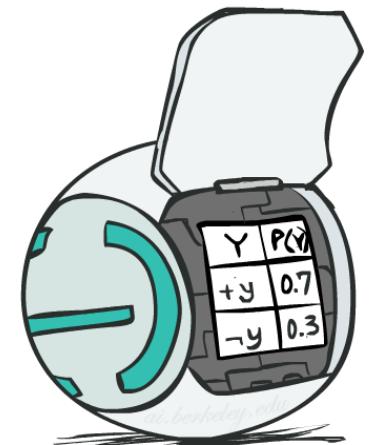
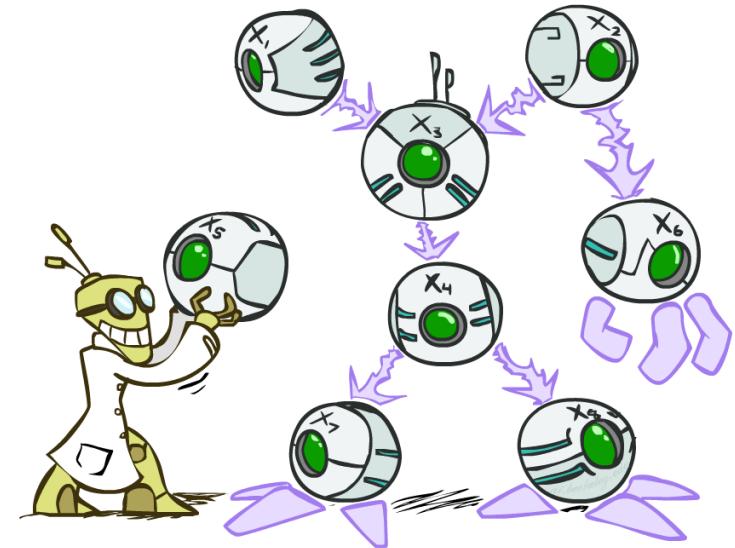
Bayes' Net Semantics

- 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

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

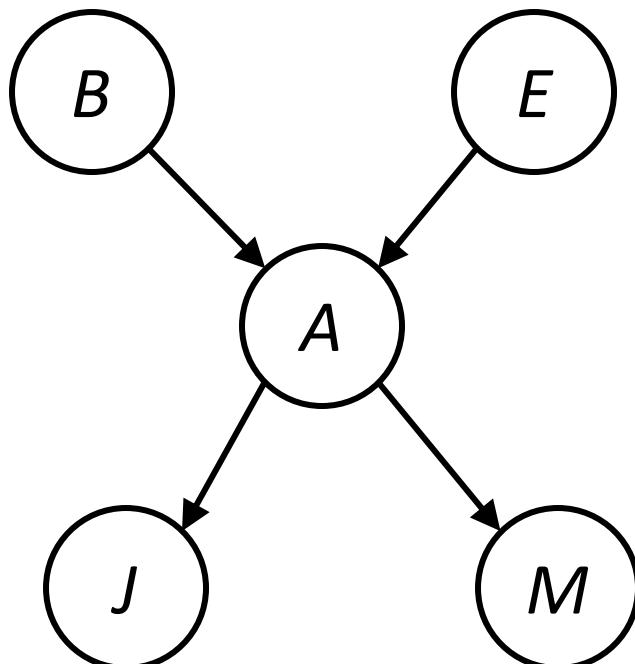
- Bayes' nets implicitly encode joint distributions
 - As a product of local conditional distributions
 - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

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



Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999

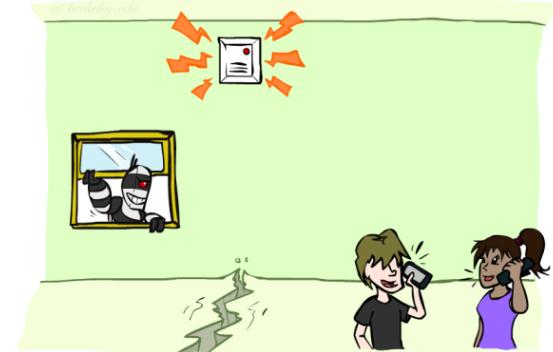


E	P(E)
+e	0.002
-e	0.998

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

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

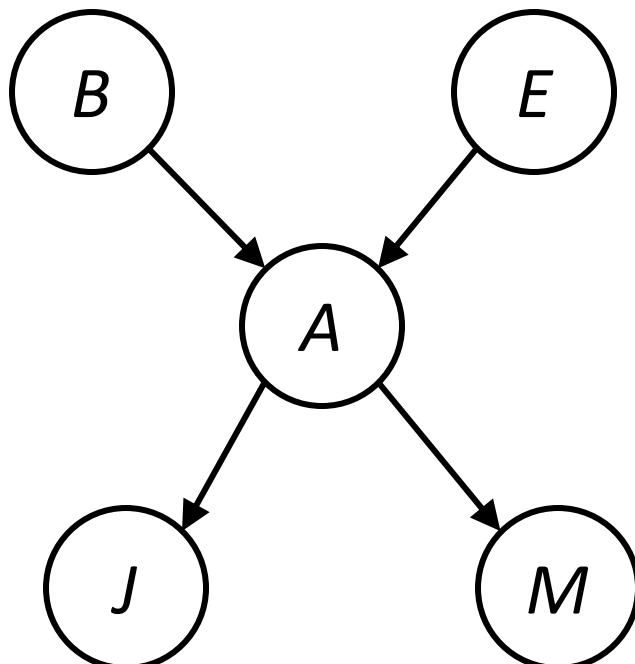
$$P(+b, -e, +a, -j, +m) =$$



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

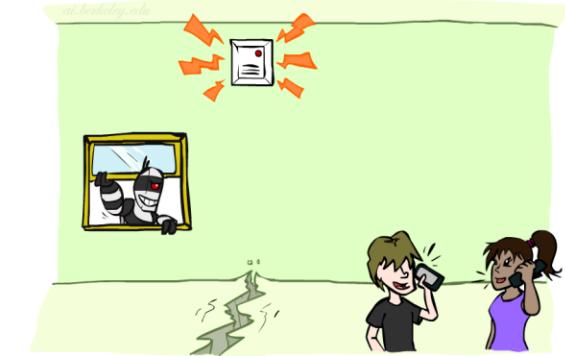
B	P(B)
+b	0.001
-b	0.999



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



$$\begin{aligned}
 P(+b, -e, +a, -j, +m) &= \\
 P(+b)P(-e)P(+a|+b, -e)P(-j|+a)P(+m|+a) &= \\
 0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7
 \end{aligned}$$

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

Size of a Bayes' Net

- How big is a joint distribution over N Boolean variables?

$$2^N$$

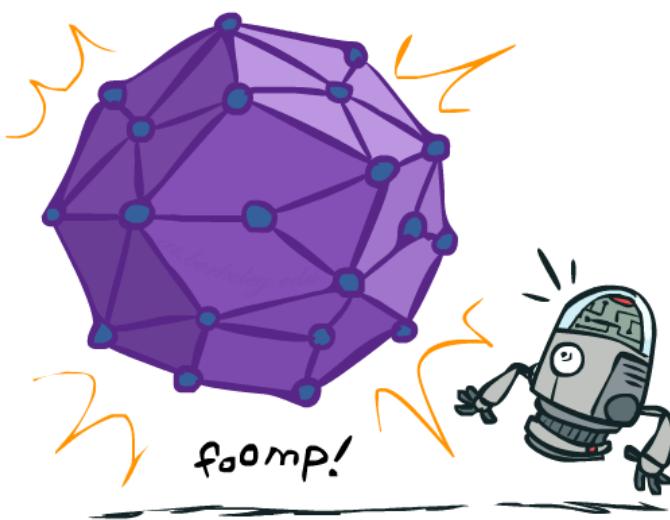
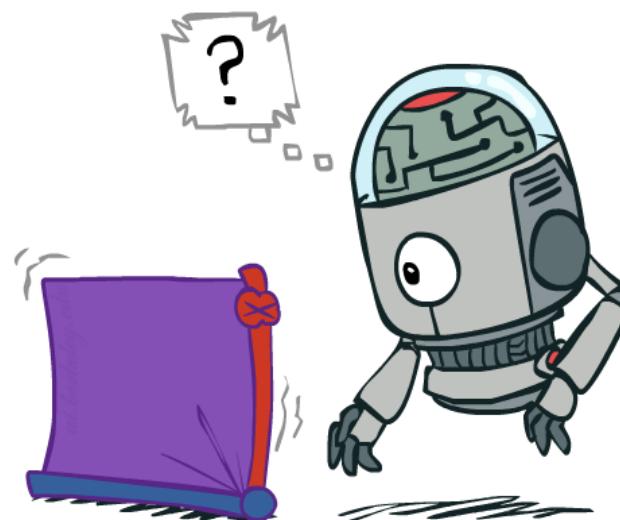
- How big is an N-node net if nodes have up to k parents?

$$O(N * 2^{k+1})$$

- Both give you the power to calculate

$$P(X_1, X_2, \dots, X_n)$$

- BNs: Huge space savings!
- Also easier to elicit local CPTs
- Also faster to answer queries (coming)



Bayes' Nets

✓ Representation

- Conditional Independences
- Probabilistic Inference
- Learning Bayes' Nets from Data

Conditional Independence

- X and Y are **independent** if

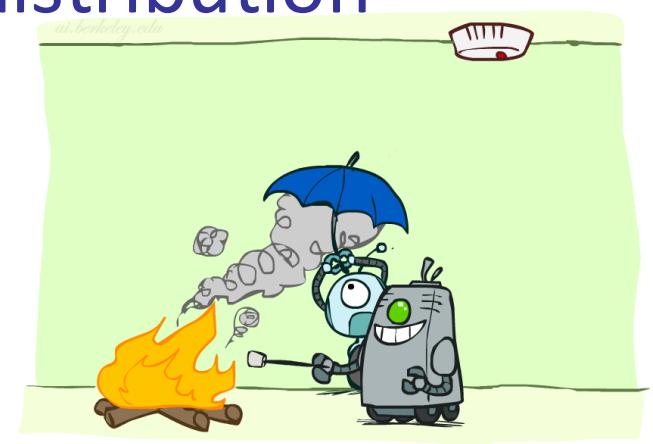
$$\forall x, y \ P(x, y) = P(x)P(y) \dashrightarrow X \perp\!\!\!\perp Y$$

- X and Y are **conditionally independent given Z**

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

- (Conditional) independence is a property of a distribution

■ Example: *Alarm* $\perp\!\!\!\perp$ *Fire* | *Smoke*

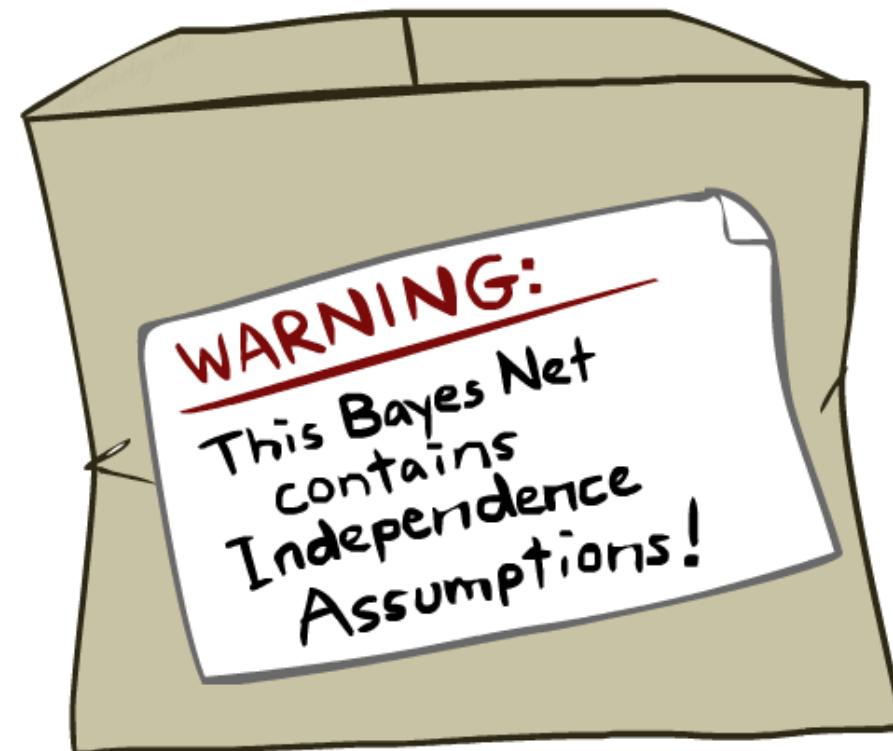


Bayes Nets: Assumptions

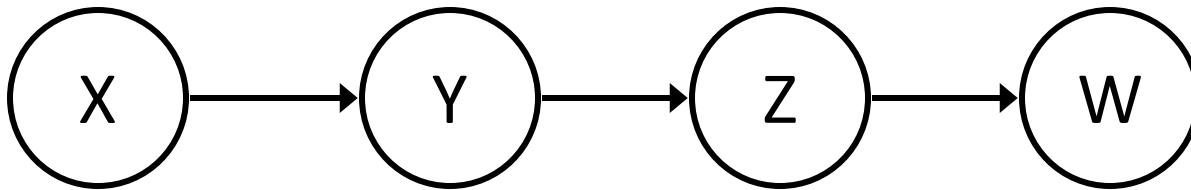
- Assumptions we are required to make to define the Bayes net when given the graph:

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

- Beyond above “chain rule → Bayes net” conditional independence assumptions
 - Often additional conditional independences
 - They can be read off the graph
- Important for modeling: understand assumptions made when choosing a Bayes net graph



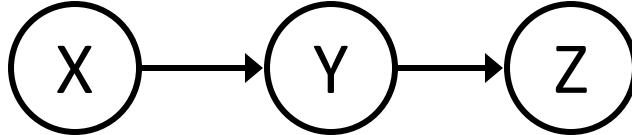
Example



- Conditional independence assumptions directly from simplifications in chain rule:
- Additional implied conditional independence assumptions?

Independence in a BN

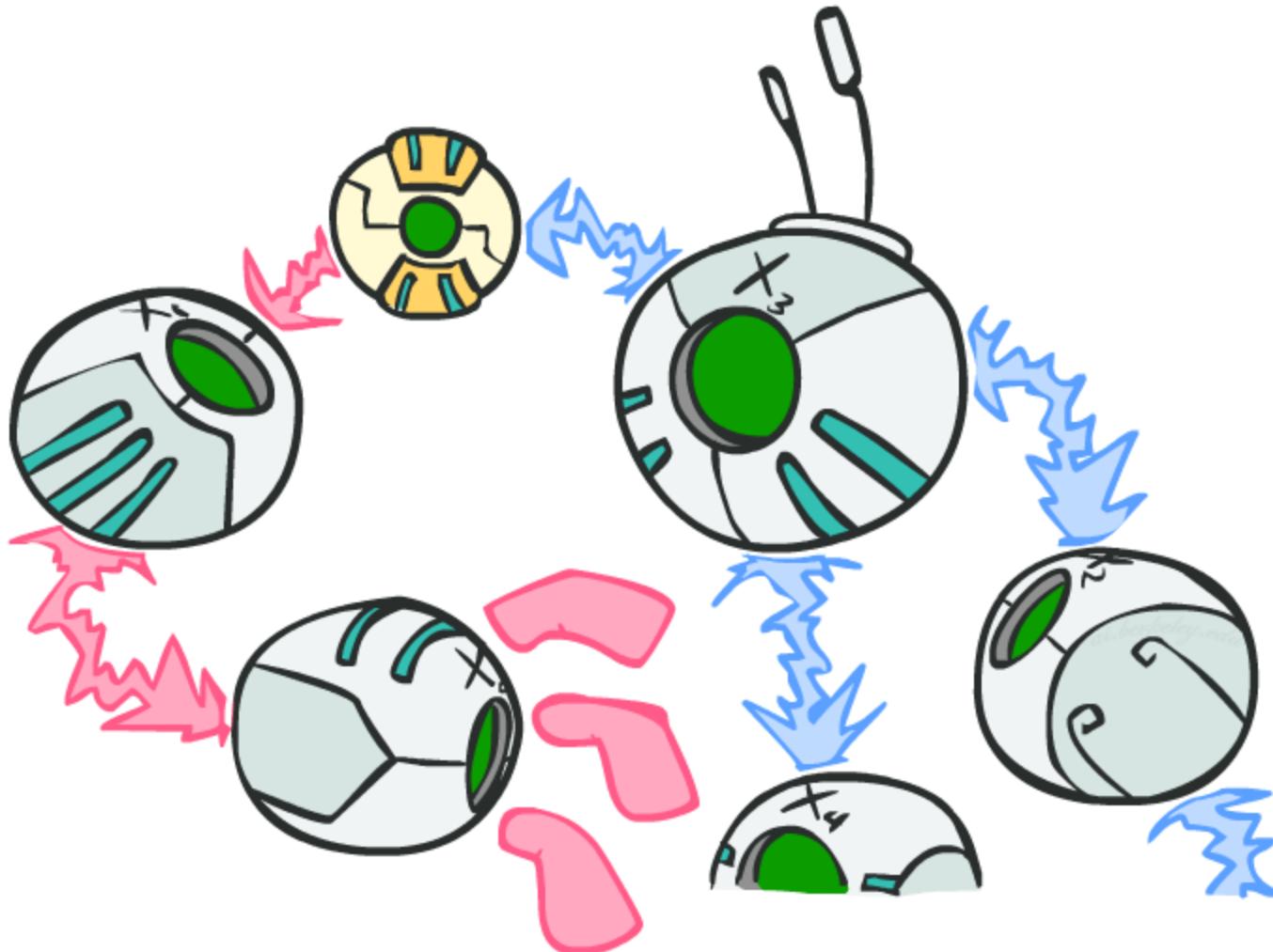
- Important question about a BN:
 - Are two nodes independent given certain evidence?
 - If yes, can prove using algebra (tedious in general)
 - If no, can prove with a counter example
 - Example:



```
graph LR; X((X)) --> Y((Y)); Y --> Z((Z))
```

- Question: are X and Z necessarily independent?
 - Answer: no. Example: low pressure causes rain, which causes traffic.
 - X can influence Z, Z can influence X (via Y)
 - Addendum: they *could* be independent: how?

D-separation: Outline



D-separation: Outline

- Study independence properties for triples
- Analyze complex cases in terms of member triples
- D-separation: a condition / algorithm for answering such queries

Causal Chains

- This configuration is a “causal chain”



X: Low pressure

Y: Rain

Z: Traffic

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

- Guaranteed X independent of Z ? **No!**

- One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.
- Example:
 - Low pressure causes rain causes traffic, high pressure causes no rain causes no traffic
- In numbers:

$$\begin{aligned}P(+y | +x) &= 1, P(-y | -x) = 1, \\P(+z | +y) &= 1, P(-z | -y) = 1\end{aligned}$$

Causal Chains

- This configuration is a “causal chain”
- Guaranteed X independent of Z given Y?



X: Low pressure Y: Rain Z: Traffic

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

$$\begin{aligned} P(z|x, y) &= \frac{P(x, y, z)}{P(x, y)} \\ &= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} \end{aligned}$$

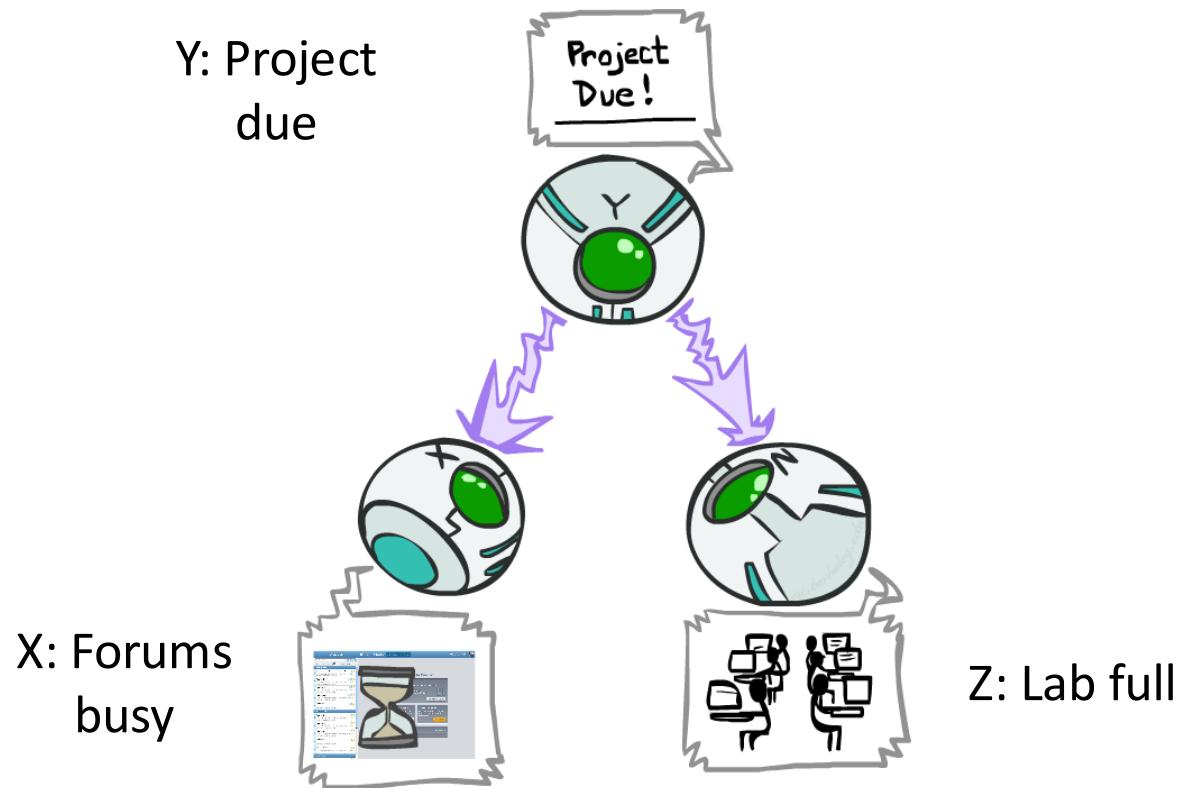
$$= P(z|y)$$

Yes!

- Evidence along the chain “blocks” the influence

Common Cause

- This configuration is a “common cause”



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

- Guaranteed X independent of Z ? *No!*

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

- Example:

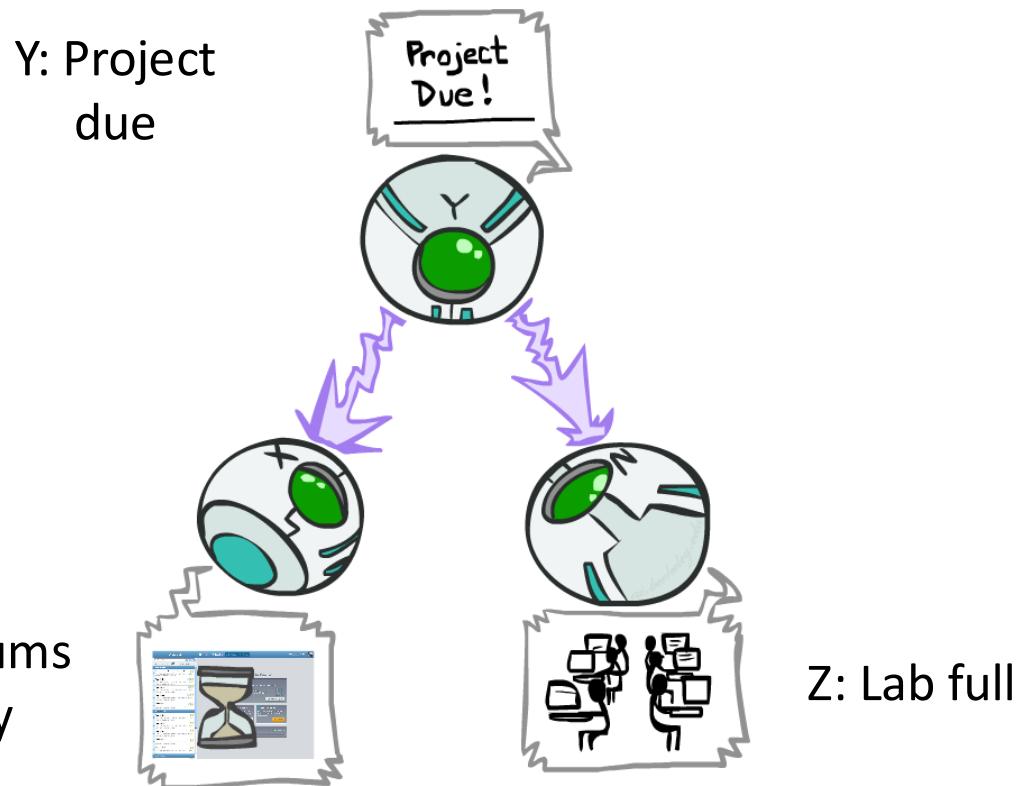
- Project due causes both forums busy and lab full

- In numbers:

$$\begin{aligned}P(+x | +y) &= 1, P(-x | -y) = 1, \\P(+z | +y) &= 1, P(-z | -y) = 1\end{aligned}$$

Common Cause

- This configuration is a “common cause”
- Guaranteed X and Z independent given Y?



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

$$\begin{aligned} P(z|x, y) &= \frac{P(x, y, z)}{P(x, y)} \\ &= \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)} \end{aligned}$$

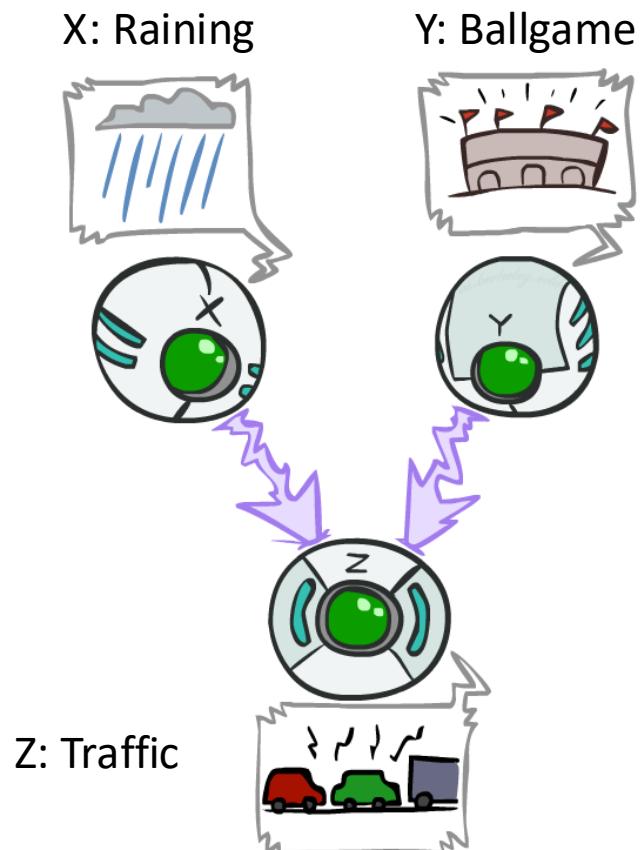
$$= P(z|y)$$

Yes!

- Observing the cause blocks influence between effects.

Common Effect

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



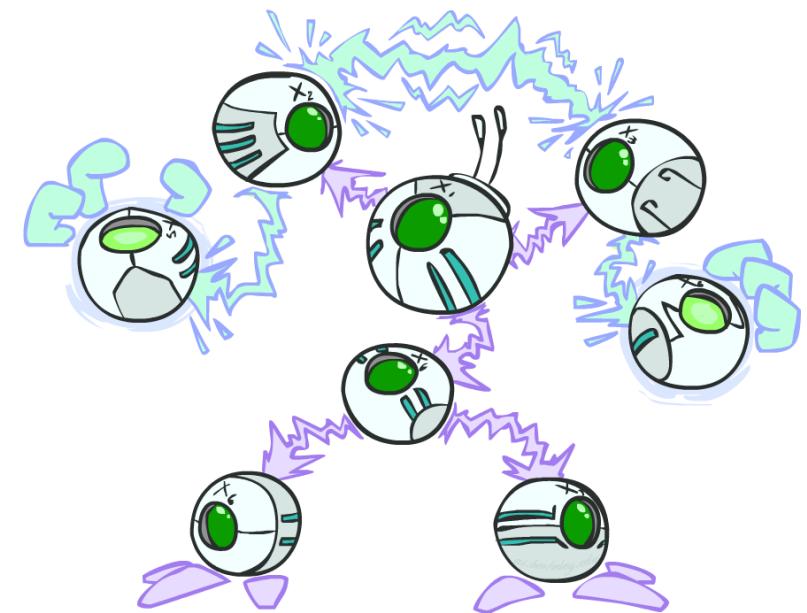
- Are X and Y independent?
 - *Yes*: the ballgame and the rain cause traffic, but they are not correlated
 - Still need to prove they must be (try it!)
- Are X and Y independent given Z?
 - *No*: seeing traffic puts the rain and the ballgame in competition as explanation.
- This is backwards from the other cases
 - Observing an effect **activates** influence between possible causes.

The General Case



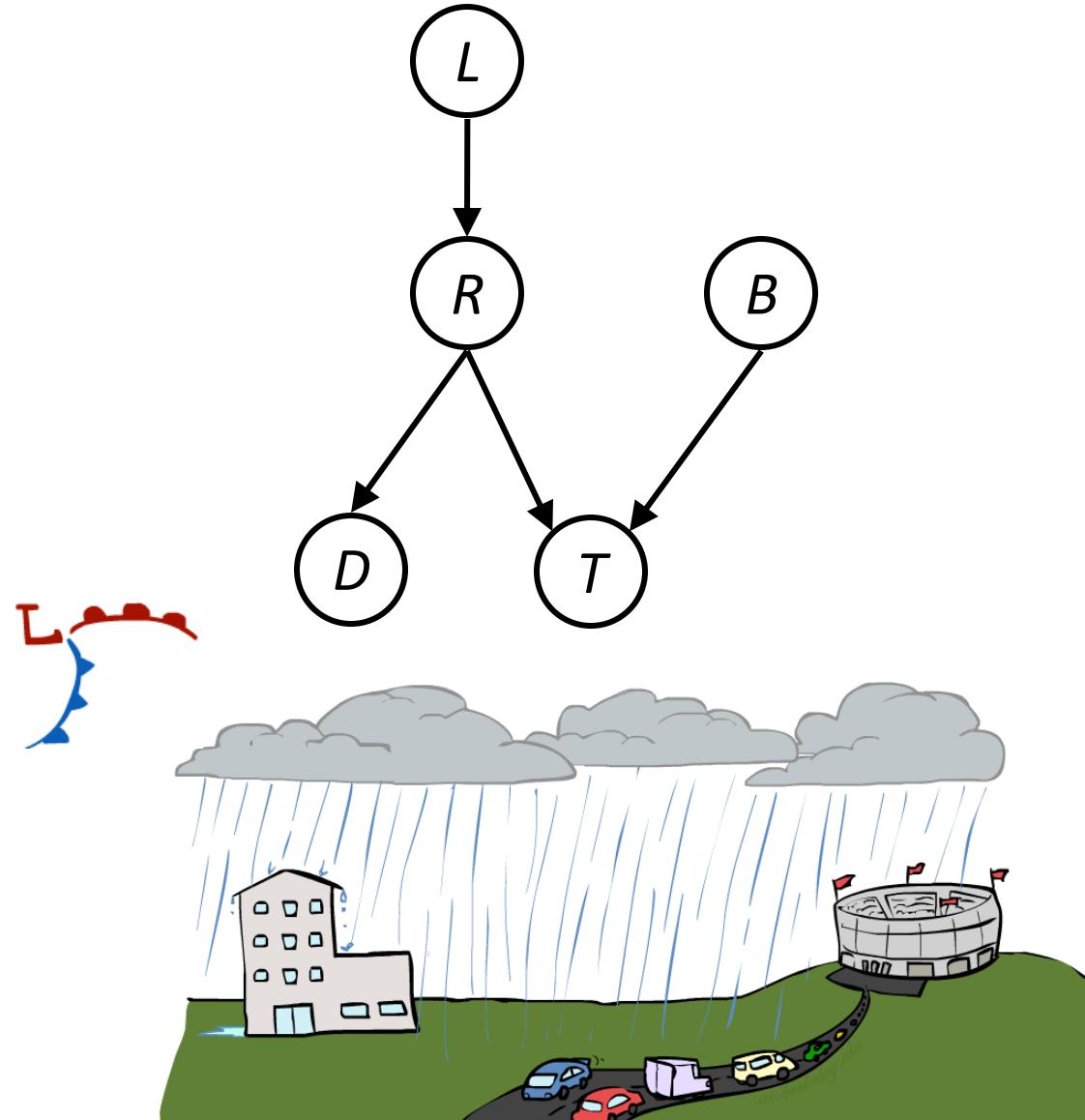
The General Case

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



Reachability

- Recipe: shade evidence nodes, look for paths in the resulting graph
- Attempt 1: if two nodes are connected by an undirected path not blocked by a shaded node, they are conditionally independent
- Almost works, but not quite
 - Where does it break?
 - Answer: the v-structure at T doesn't count as a link in a path unless "active"



Active / Inactive Paths

- Question: Are X and Y conditionally independent given evidence variables $\{Z\}$?

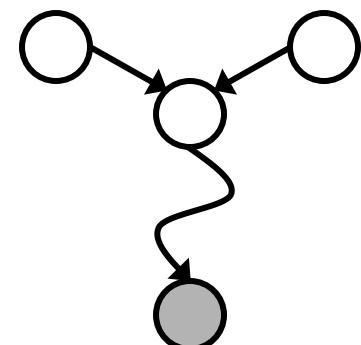
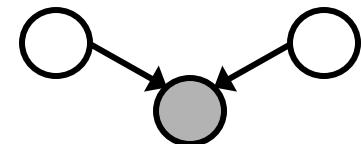
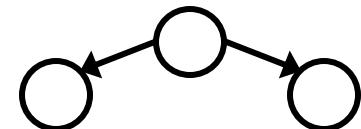
- Yes, if X and Y “d-separated” by Z
- Consider all (undirected) paths from X to Y
- No active paths = independence!

- A path is active if each triple is active:

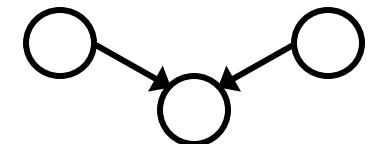
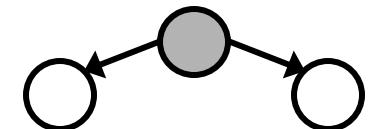
- Causal chain $A \rightarrow B \rightarrow C$ where B is unobserved (either direction)
- Common cause $A \leftarrow B \rightarrow C$ where B is unobserved
- Common effect (aka v-structure)
 $A \rightarrow B \leftarrow C$ where B or one of its descendants is observed

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

Active Triples



Inactive Triples



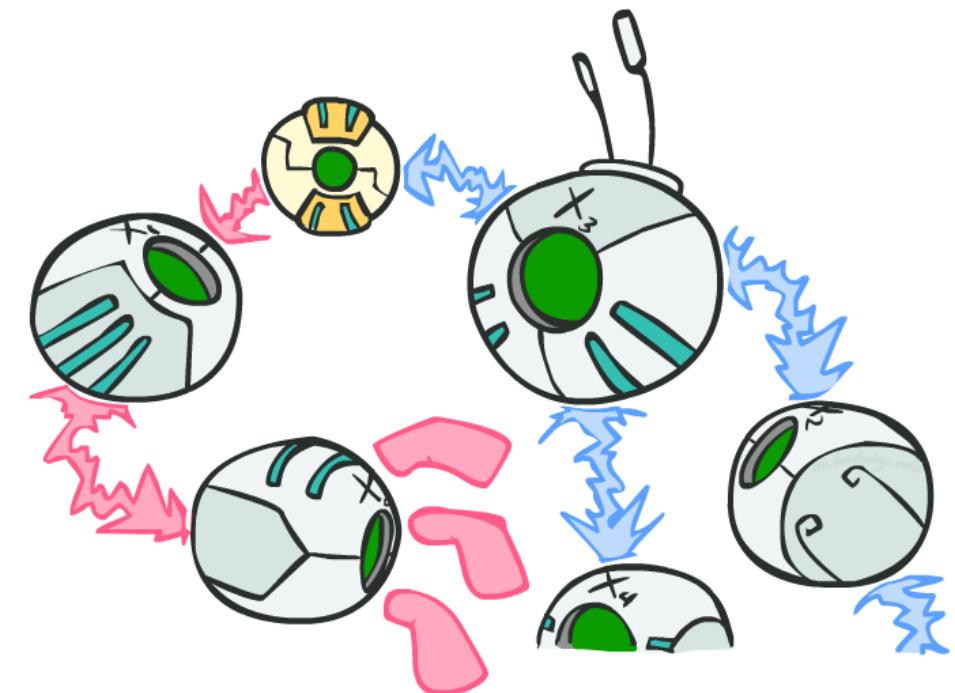
D-Separation

- Query: $X_i \perp\!\!\!\perp X_j | \{X_{k_1}, \dots, X_{k_n}\}$?
- Check all (undirected!) paths between X_i and X_j
 - If one or more active, then independence not guaranteed

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

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

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



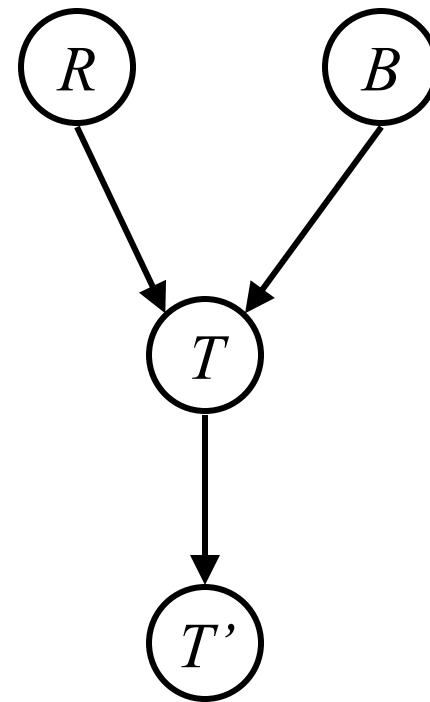
Example

$R \perp\!\!\!\perp B$

Yes

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

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



Example

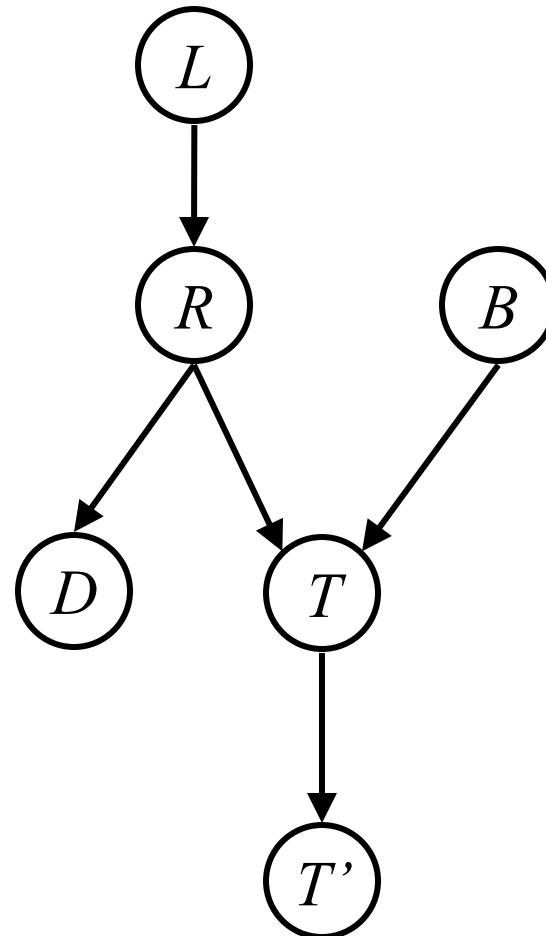
$L \perp\!\!\!\perp T' | T$ Yes

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

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

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

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



Example

- Variables:

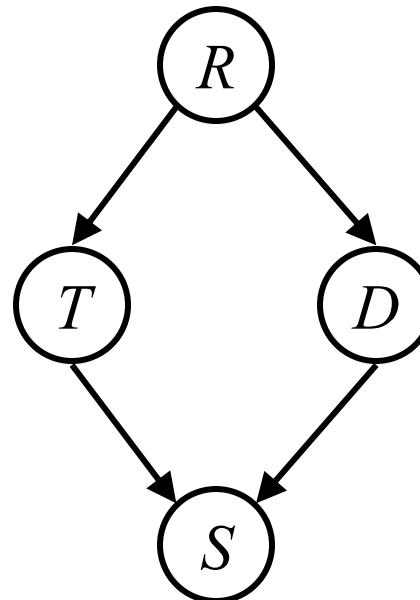
- R: Raining
- T: Traffic
- D: Roof drips
- S: I'm sad

- Questions:

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

$$T \perp\!\!\!\perp D | R \quad \text{Yes}$$

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

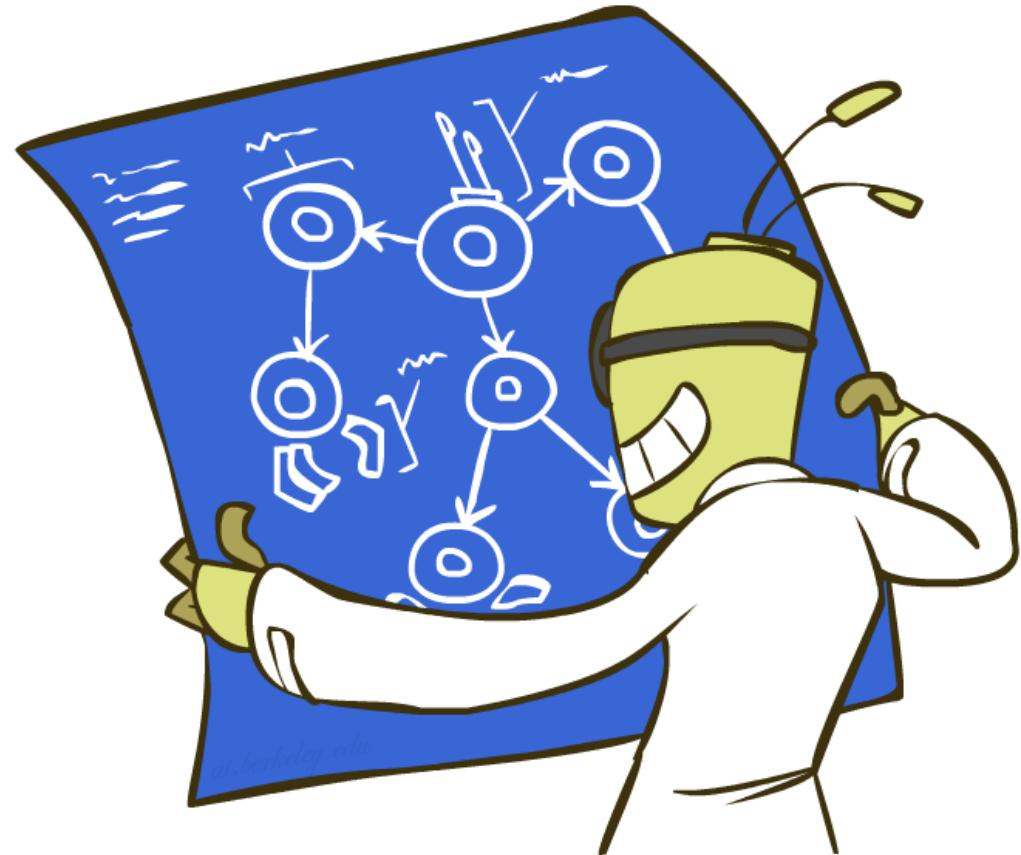


Structure Implications

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

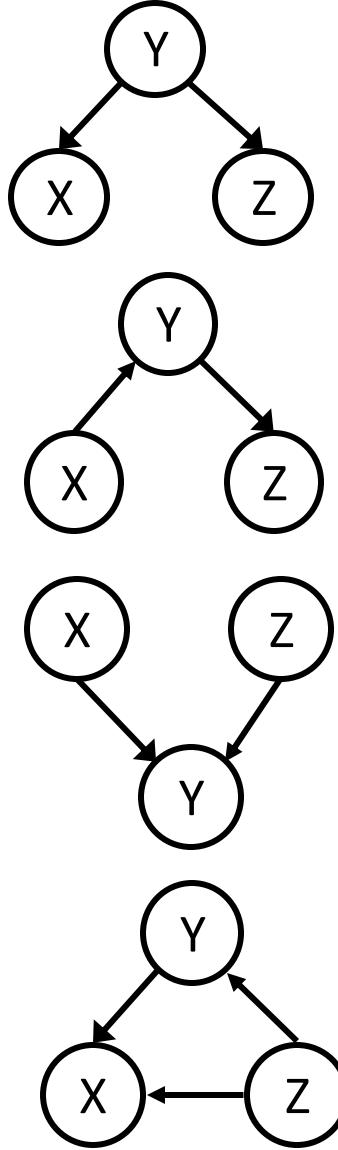
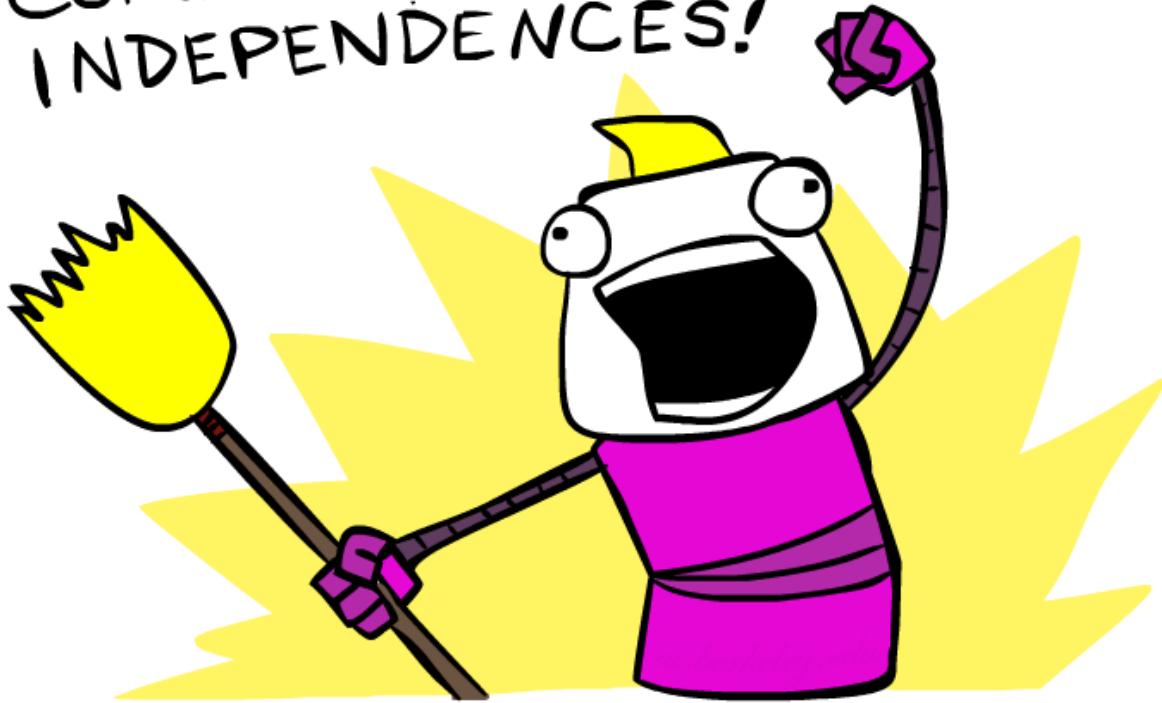
$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, \dots, X_{k_n}\}$$

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



Computing All Independences

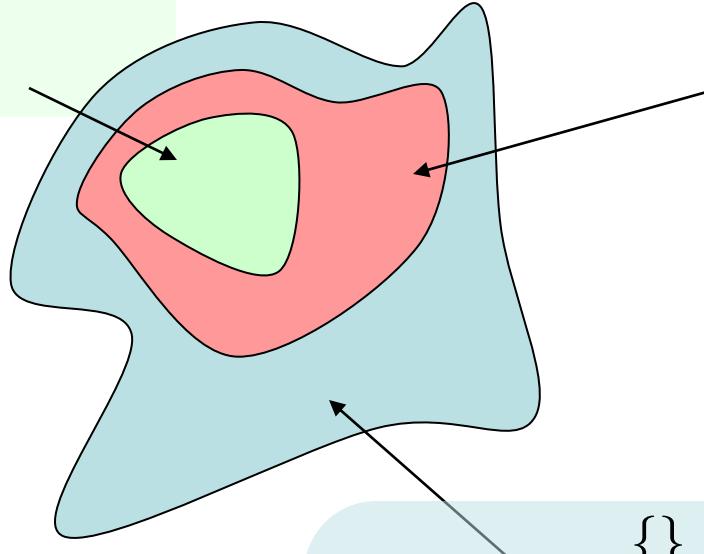
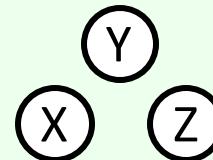
COMPUTE ALL THE INDEPENDENCES!



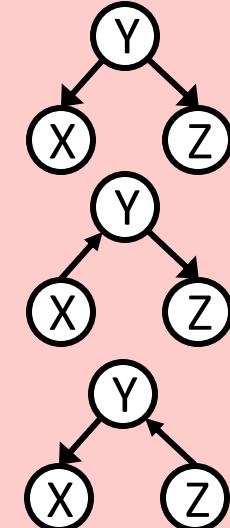
Topology Limits Distributions

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

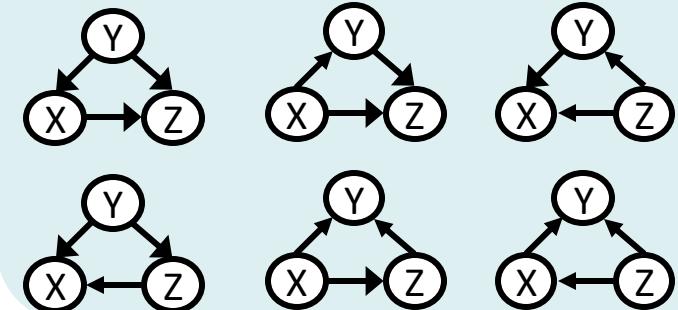
$$\{X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z, \\ X \perp\!\!\!\perp Z \mid Y, X \perp\!\!\!\perp Y \mid Z, Y \perp\!\!\!\perp Z \mid X\}$$



$$\{X \perp\!\!\!\perp Z \mid Y\}$$



$$\{\}$$



Bayes Nets Representation Summary

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

Bayes' Nets



Representation



Conditional Independences

- Probabilistic Inference
 - Enumeration (exact, exponential complexity)
 - Variable elimination (exact, worst-case exponential complexity, often better)
 - Probabilistic inference is NP-complete
 - Sampling (approximate)
- Learning Bayes' Nets from Data

Inference

- Inference: calculating some useful quantity from a joint probability distribution

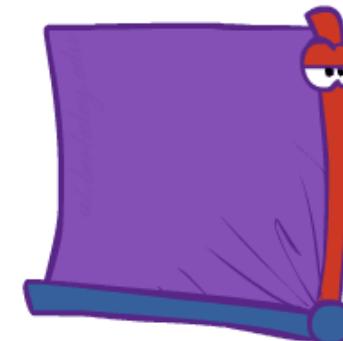
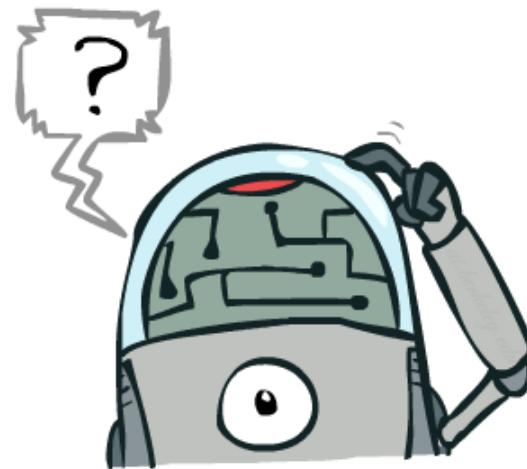
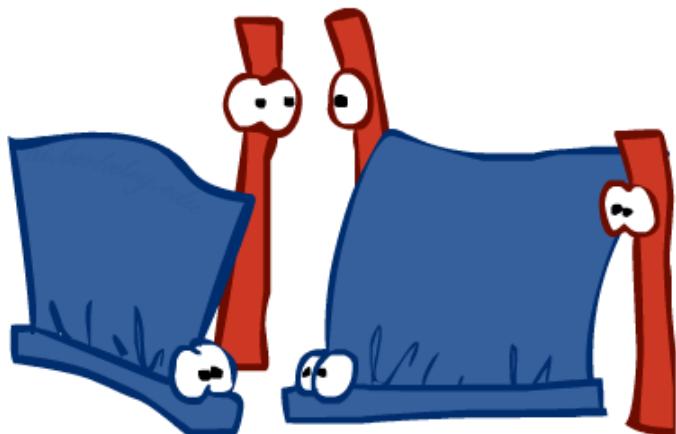
- Examples:

- Posterior probability

$$P(Q|E_1 = e_1, \dots, E_k = e_k)$$

- Most likely explanation:

$$\operatorname{argmax}_q P(Q = q|E_1 = e_1 \dots)$$



Inference by Enumeration

- General case:

- Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
- Query* variable: Q
- Hidden variables: $H_1 \dots H_r$

$$\left. \begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array} \right\} X_1, X_2, \dots, X_n$$

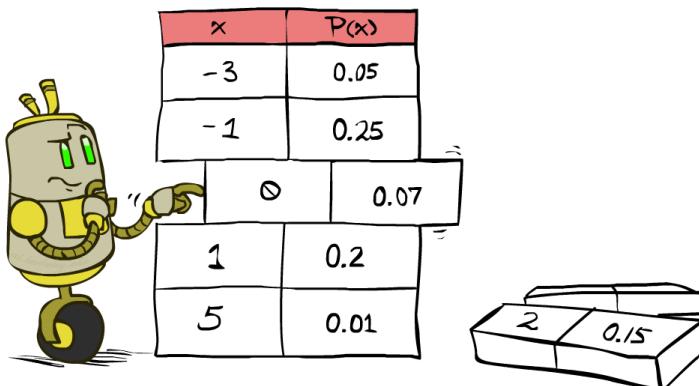
All variables

- We want:

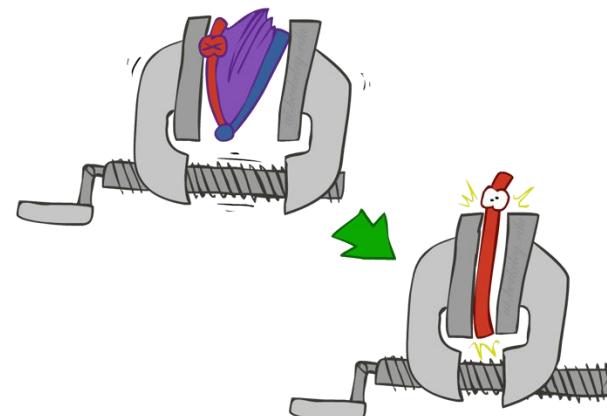
$$P(Q|e_1 \dots e_k)$$

* Works fine with multiple query variables, too

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \dots e_k)$$

$$P(Q|e_1 \dots e_k) = \frac{1}{Z} P(Q, e_1 \dots e_k)$$

$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Q, h_1 \dots h_r, e_1 \dots e_k)}_{X_1, X_2, \dots, X_n}$$

Inference by Enumeration in Bayes' Net

- Given unlimited time, inference in BNs is easy
- Reminder of inference by enumeration by example:

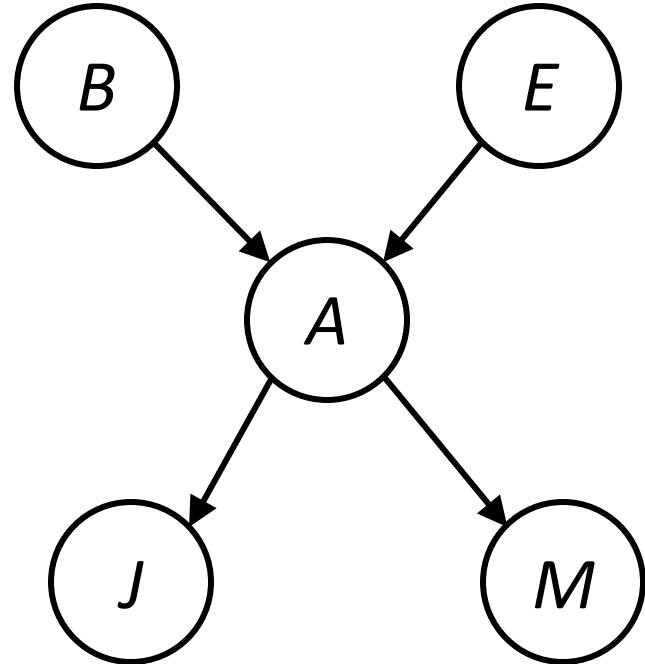
$$P(B \mid +j, +m) \propto_B P(B, +j, +m)$$

$$= \sum_{e,a} P(B, e, a, +j, +m)$$

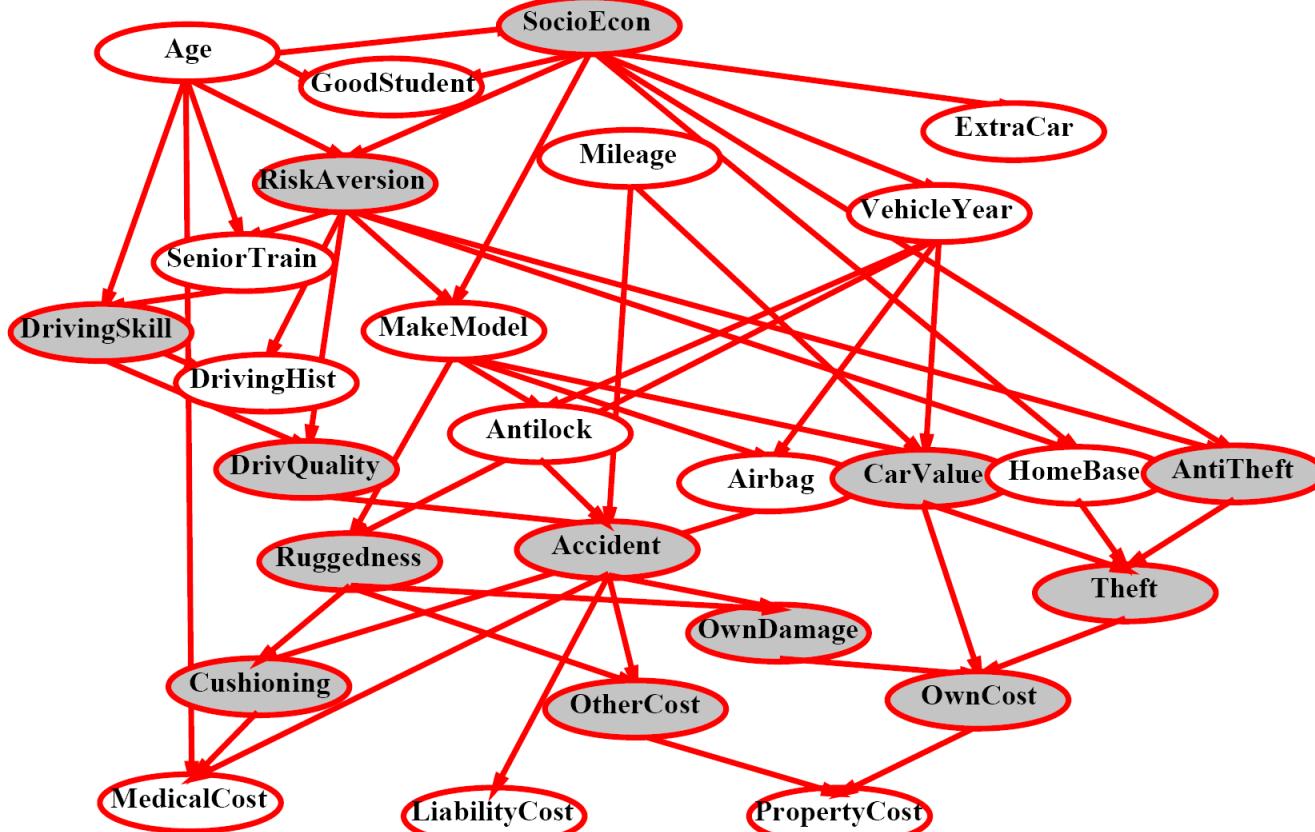
$$= \sum_{e,a} P(B)P(e)P(a|B, e)P(+j|a)P(+m|a)$$

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

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



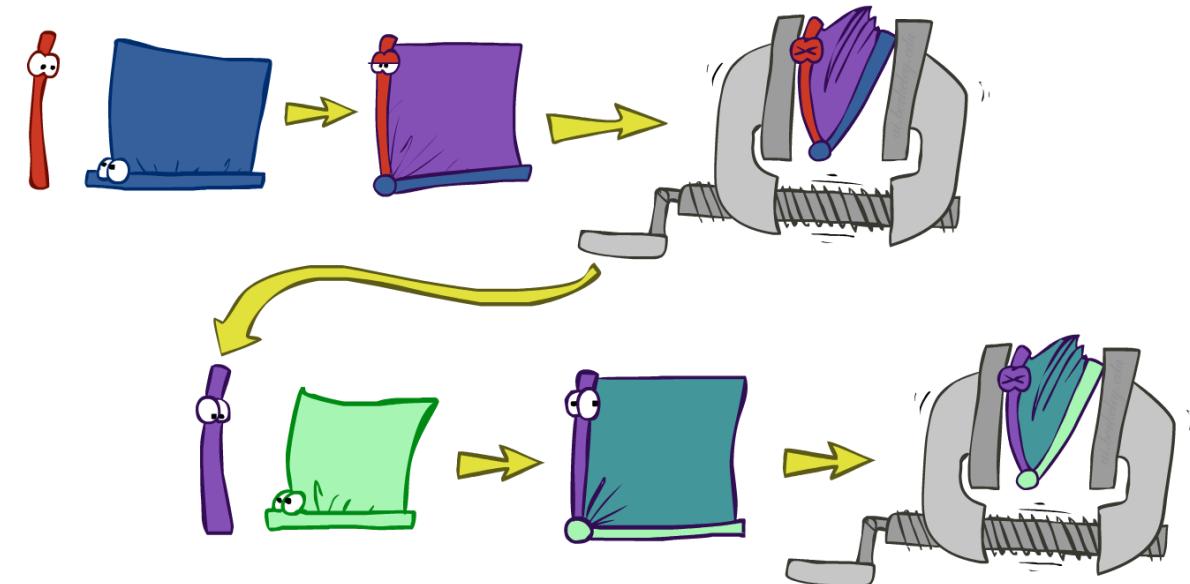
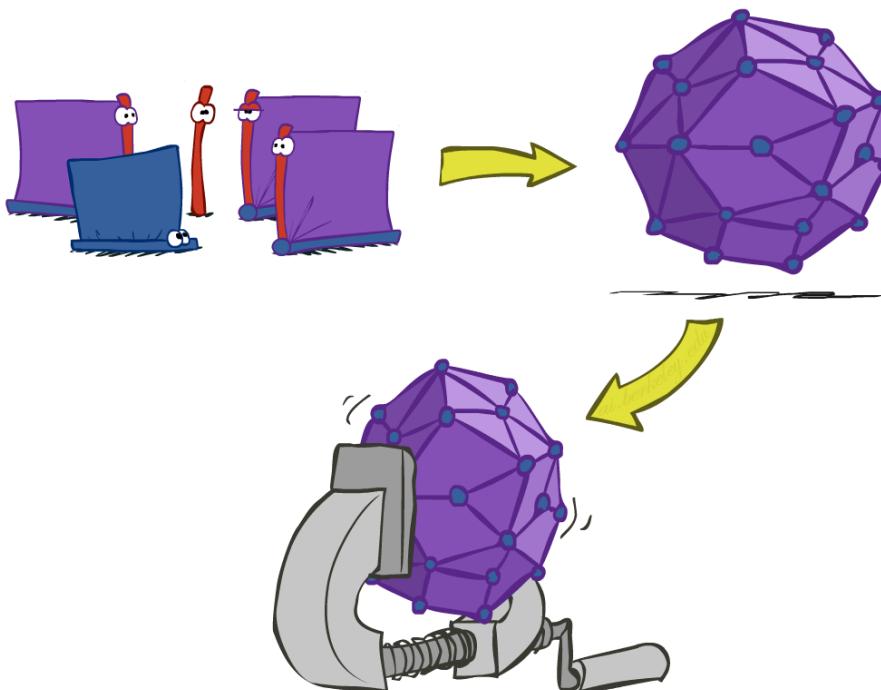
Inference by Enumeration?



$$P(\text{Antilock} | \text{observed variables}) = ?$$

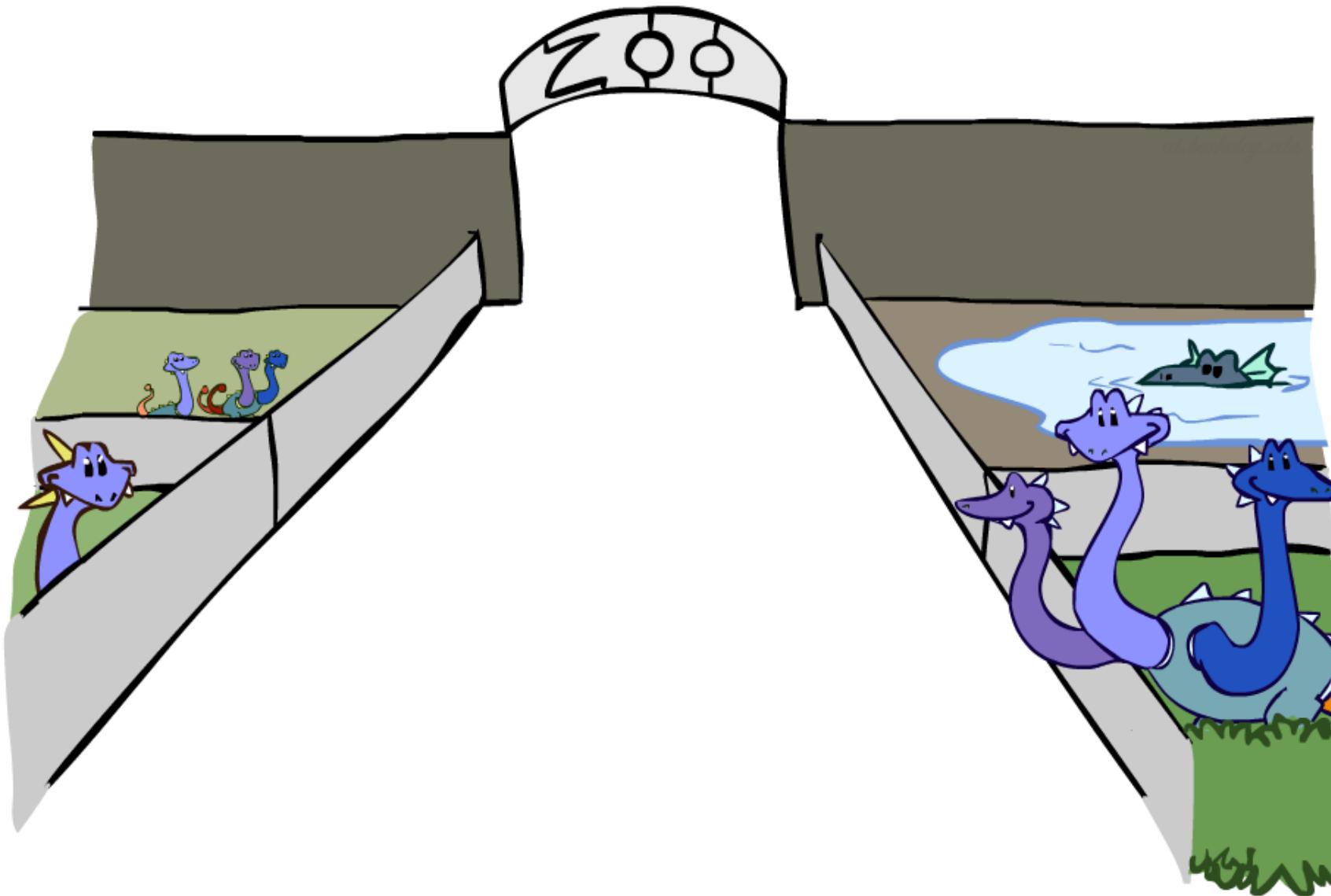
Inference by Enumeration vs. Variable Elimination

- Why is inference by enumeration so slow?
 - You join up the whole joint distribution before you sum out the hidden variables
- Idea: interleave joining and marginalizing!
 - Called “Variable Elimination”
 - Still NP-hard, but usually much faster than inference by enumeration



- First we'll need some new notation: factors

Factor Zoo



Factor Zoo I

- Joint distribution: $P(X, Y)$

- Entries $P(x, y)$ for all x, y
- Sums to 1

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

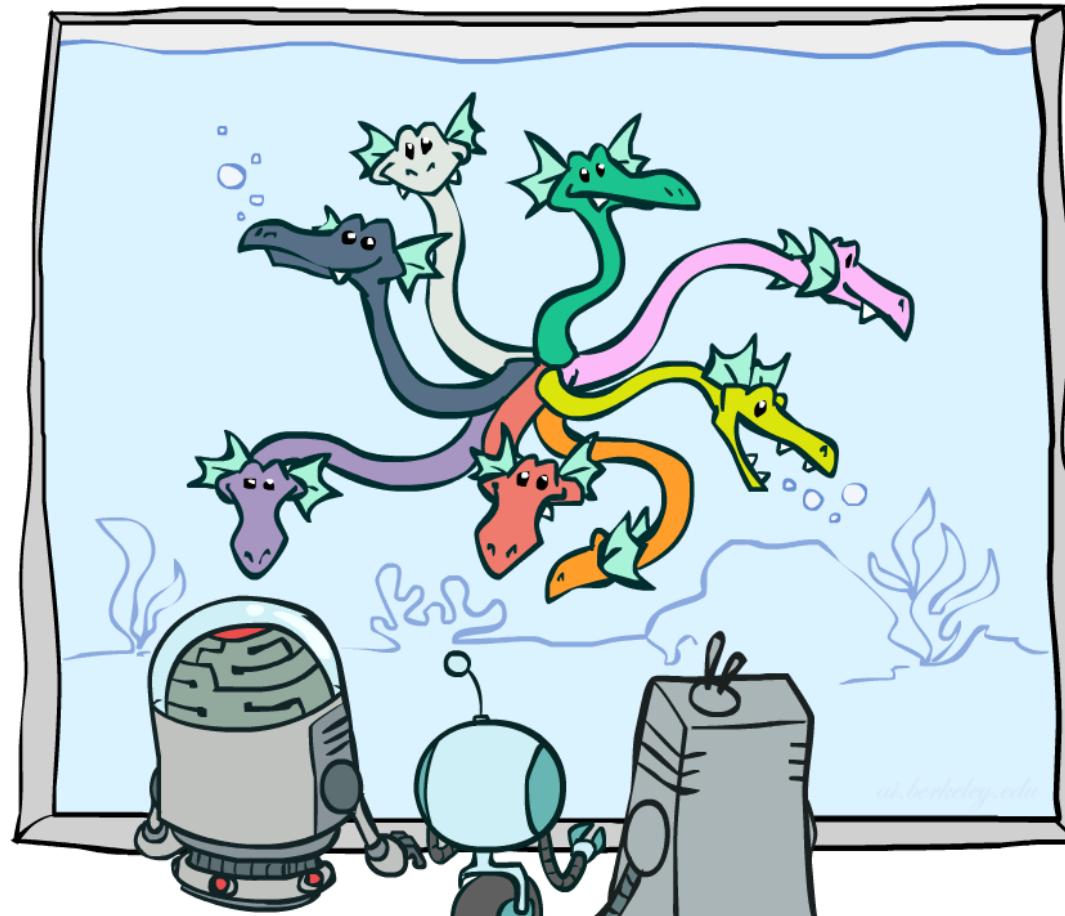
- Selected joint: $P(x, Y)$

- A slice of the joint distribution
- Entries $P(x, y)$ for fixed x , all y
- Sums to $P(x)$

$$P(\text{cold}, W)$$

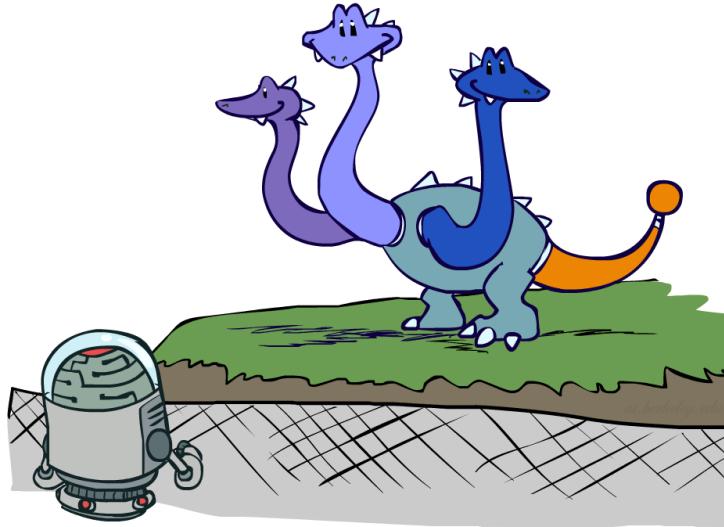
T	W	P
cold	sun	0.2
cold	rain	0.3

- Number of capitals = dimensionality of the table

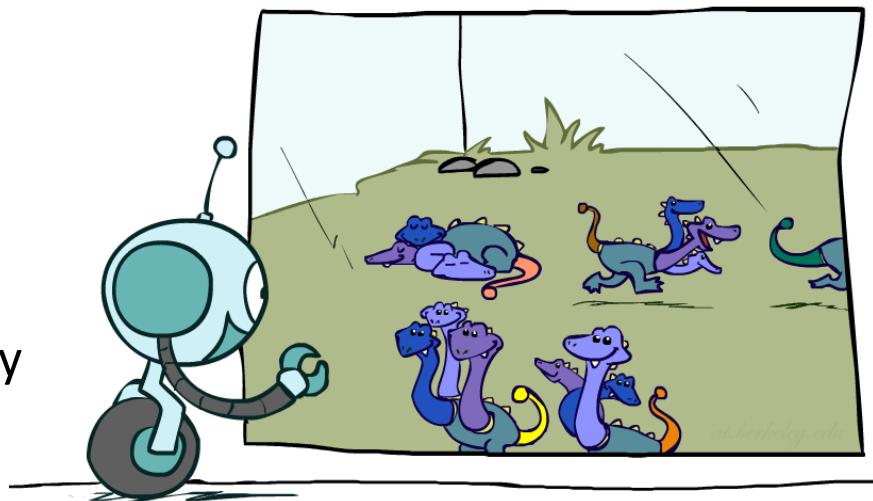


Factor Zoo II

- Single conditional: $P(Y | x)$
 - Entries $P(y | x)$ for fixed x , all
 - Sums to 1



- Family of conditionals:
 $P(X | Y)$
 - Multiple conditionals
 - Entries $P(x | y)$ for all x, y
 - Sums to $|Y|$



$P(W|cold)$

T	W	P
cold	sun	0.4
cold	rain	0.6

$P(W|T)$

T	W	P
hot	sun	0.8
hot	rain	0.2
cold	sun	0.4
cold	rain	0.6

$P(W|hot)$

$P(W|cold)$

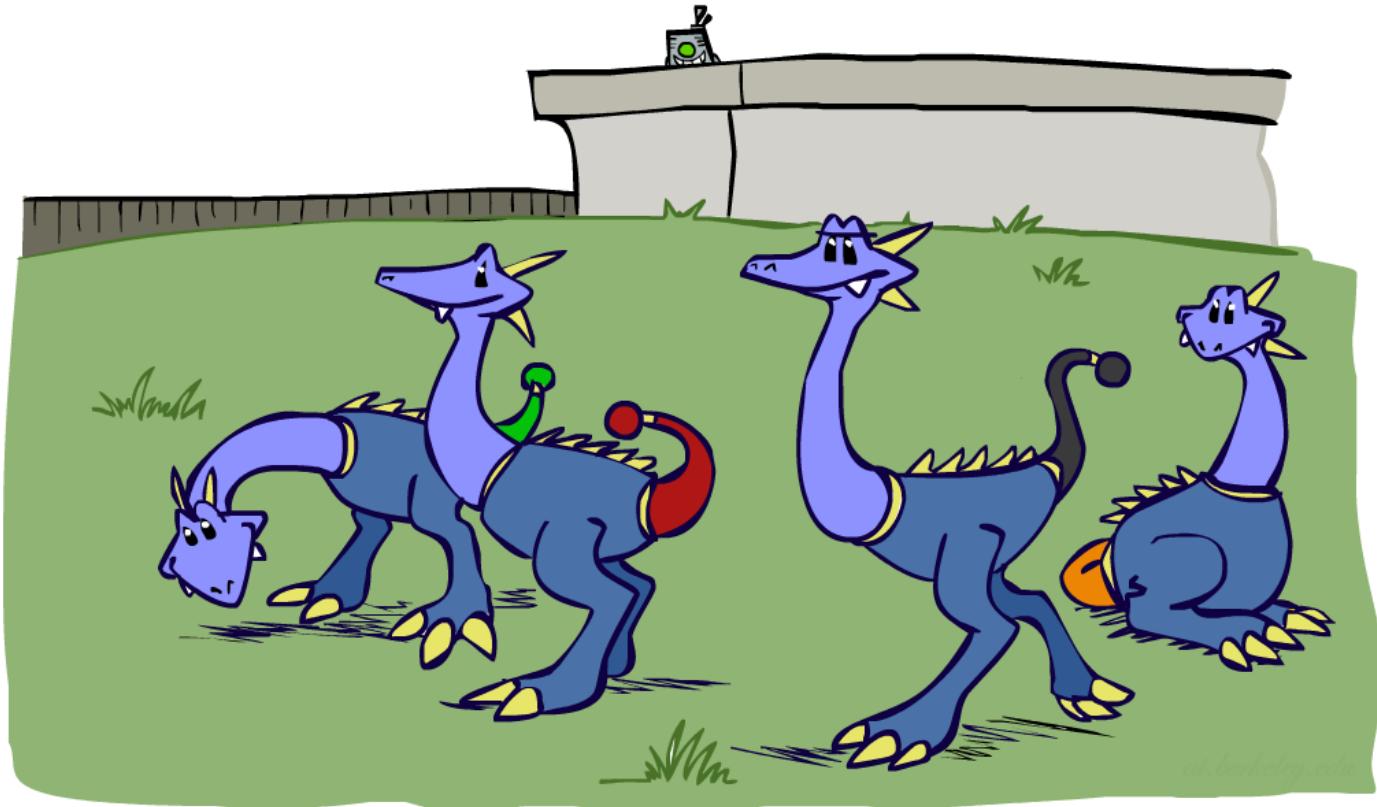
Factor Zoo III

- Specified family: $P(y | X)$
 - Entries $P(y | x)$ for fixed y ,
but for all x
 - Sums to ... who knows!

$$P(\text{rain} | T)$$

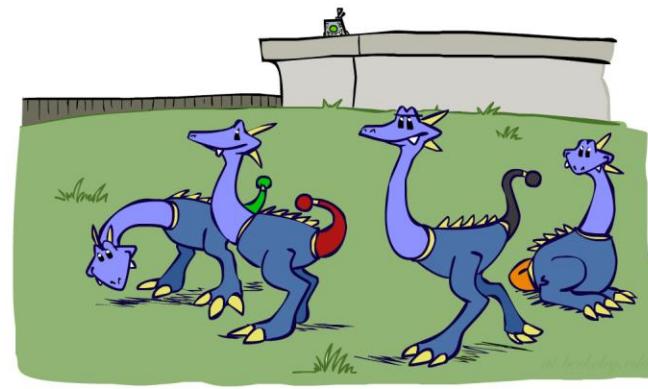
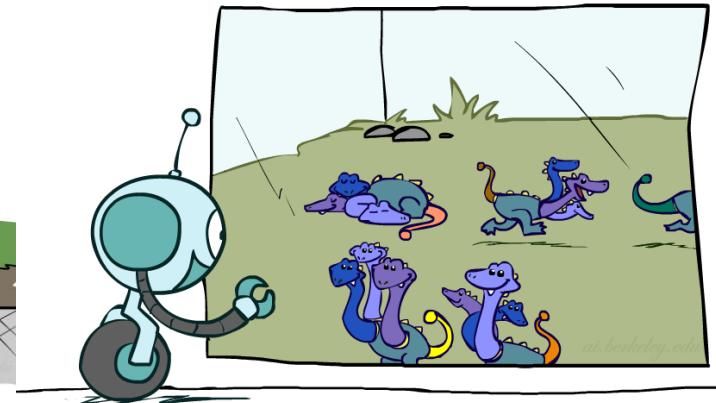
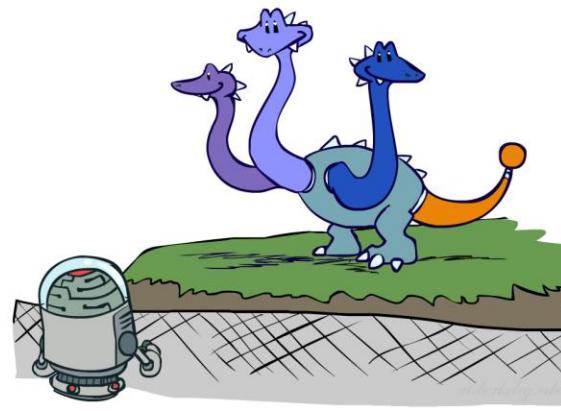
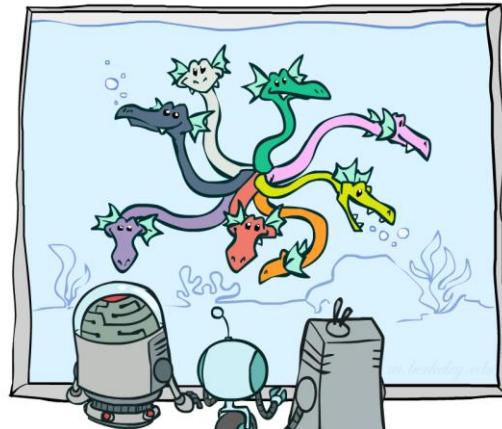
T	W	P
hot	rain	0.2
cold	rain	0.6

$$\left. \begin{array}{l} P(\text{rain} | \text{hot}) \\ P(\text{rain} | \text{cold}) \end{array} \right\}$$



Factor Zoo Summary

- In general, when we write $P(Y_1 \dots Y_N | X_1 \dots X_M)$
 - It is a “factor,” a multi-dimensional array
 - Its values are $P(y_1 \dots y_N | x_1 \dots x_M)$
 - Any assigned (=lower-case) X or Y is a dimension missing (selected) from the array



Example: Traffic Domain

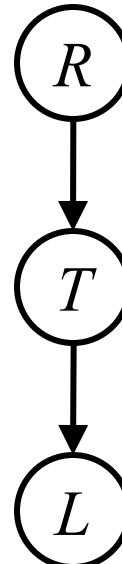
- Random Variables

- R: Raining
- T: Traffic
- L: Late for class!

$$P(L) = ?$$

$$= \sum_{r,t} P(r,t,L)$$

$$= \sum_{r,t} P(r)P(t|r)P(L|t)$$



$$P(R)$$

+r	0.1
-r	0.9

$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Inference by Enumeration: Procedural Outline

- Track objects called **factors**
- Initial factors are local CPTs (one per node)

$$P(R)$$

+r	0.1
-r	0.9

$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- Any known values are selected
 - E.g. if we know $L = +\ell$, the initial factors are

$$P(R)$$

+r	0.1
-r	0.9

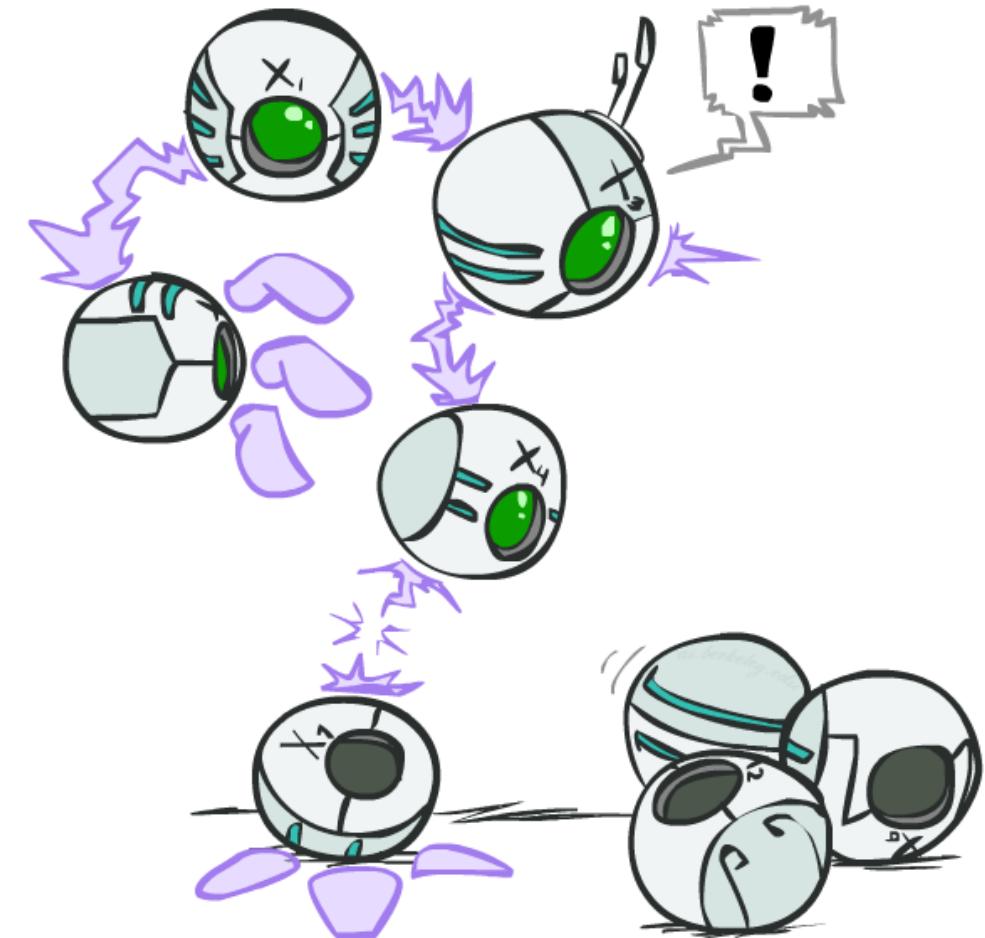
$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(+\ell|T)$$

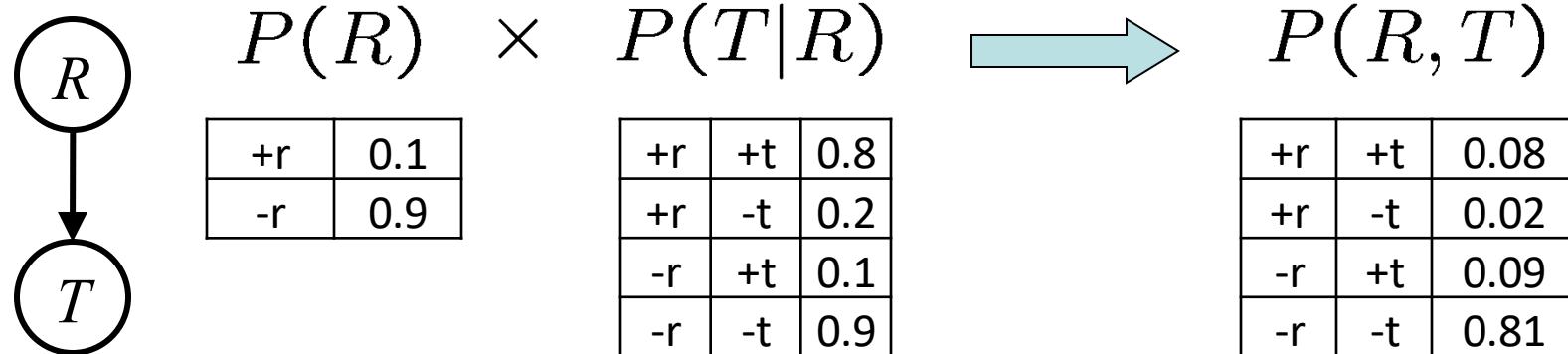
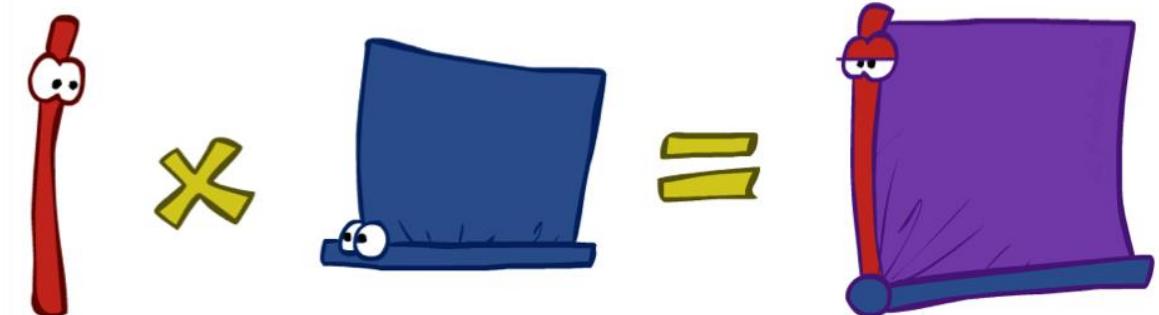
+t	+l	0.3
-t	+l	0.1

- Procedure: Join all factors, then eliminate all hidden variables



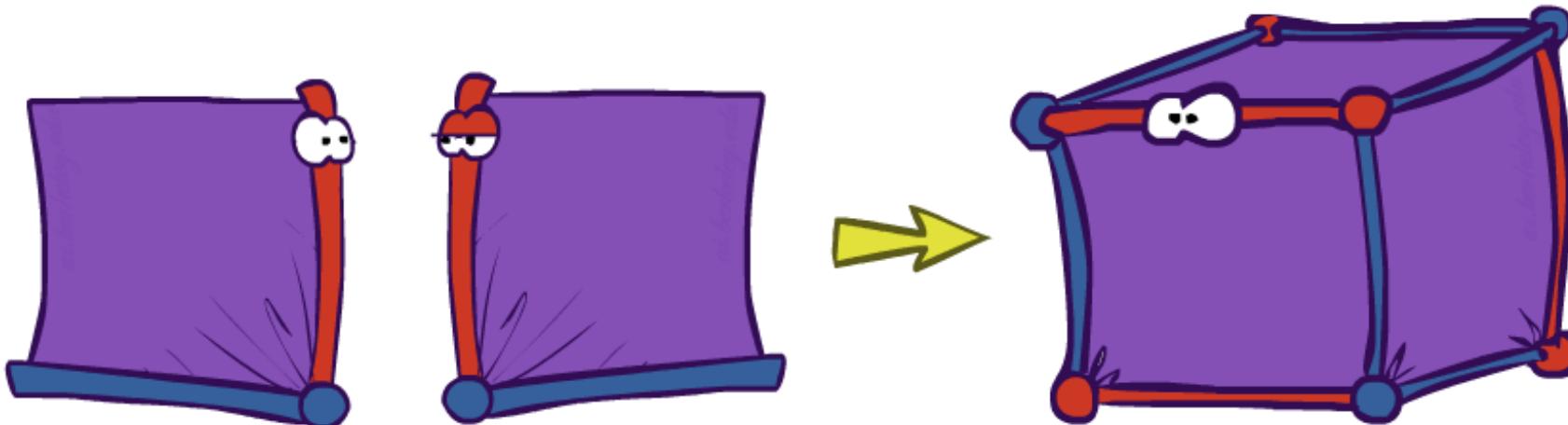
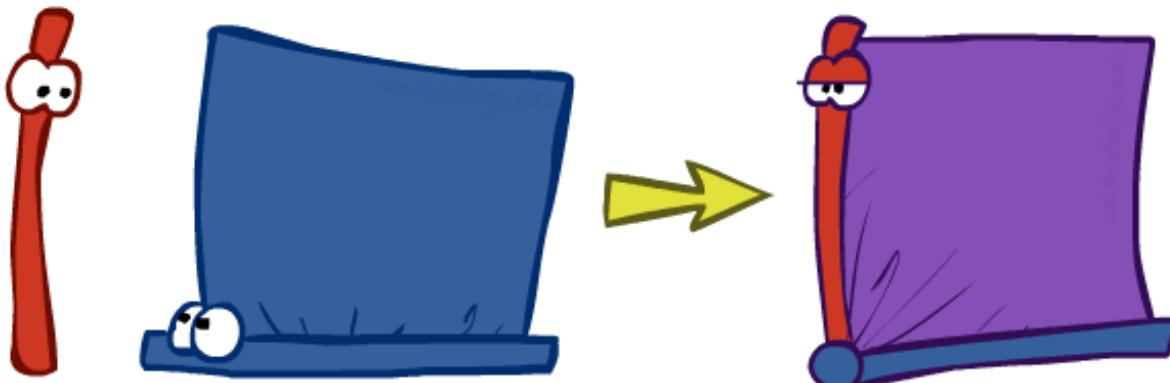
Operation 1: Join Factors

- First basic operation: **joining factors**
- Combining factors:
 - Just like a database join
 - Get all factors over the joining variable
 - Build a new factor over the union of the variables involved
- Example: Join on R

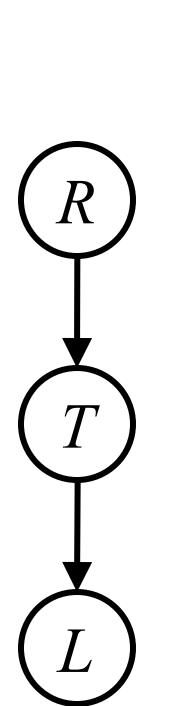
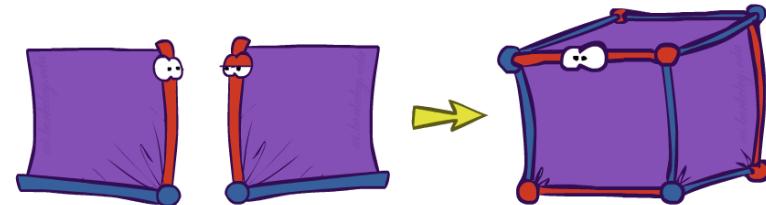
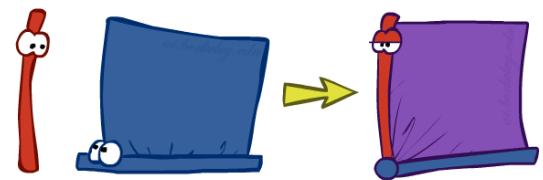


- Computation for each entry: pointwise products $\forall r, t : P(r, t) = P(r) \cdot P(t|r)$

Example: Multiple Joins



Example: Multiple Joins



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join R

$P(R, T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

Join T

R, T

L

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(R, T, L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

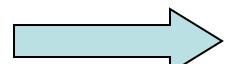
Operation 2: Eliminate

- Second basic operation: **marginalization**
- Take a factor and sum out a variable
 - Shrinks a factor to a smaller one
 - A **projection** operation
- Example:

$$P(R, T)$$

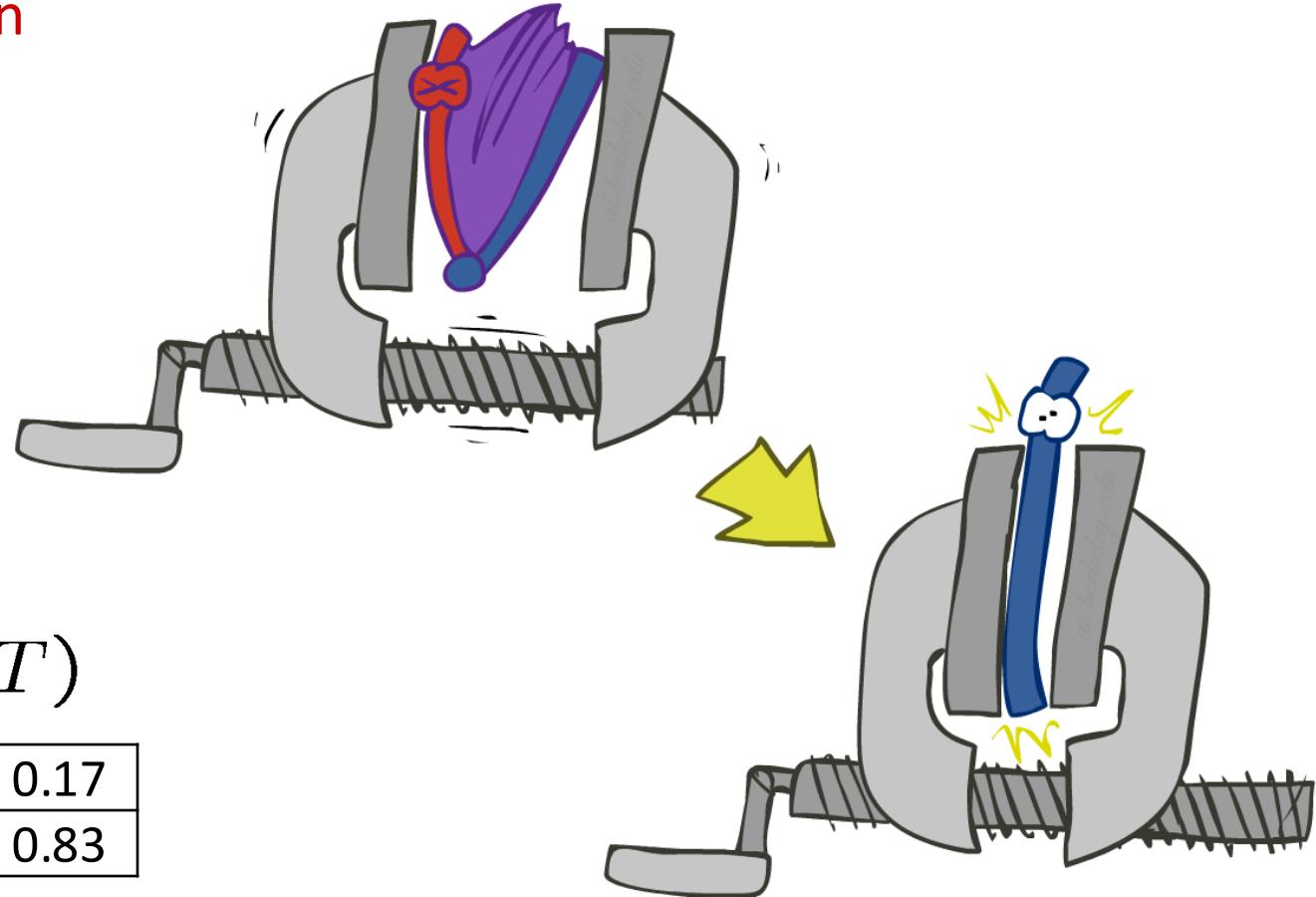
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

sum R



$$P(T)$$

+t	0.17
-t	0.83



Multiple Elimination

$P(R, T, L)$

$+r$	$+t$	$+l$	0.024
$+r$	$+t$	$-l$	0.056
$+r$	$-t$	$+l$	0.002
$+r$	$-t$	$-l$	0.018
$-r$	$+t$	$+l$	0.027
$-r$	$+t$	$-l$	0.063
$-r$	$-t$	$+l$	0.081
$-r$	$-t$	$-l$	0.729

R, T, L

Sum
out R

T, L

Sum
out T

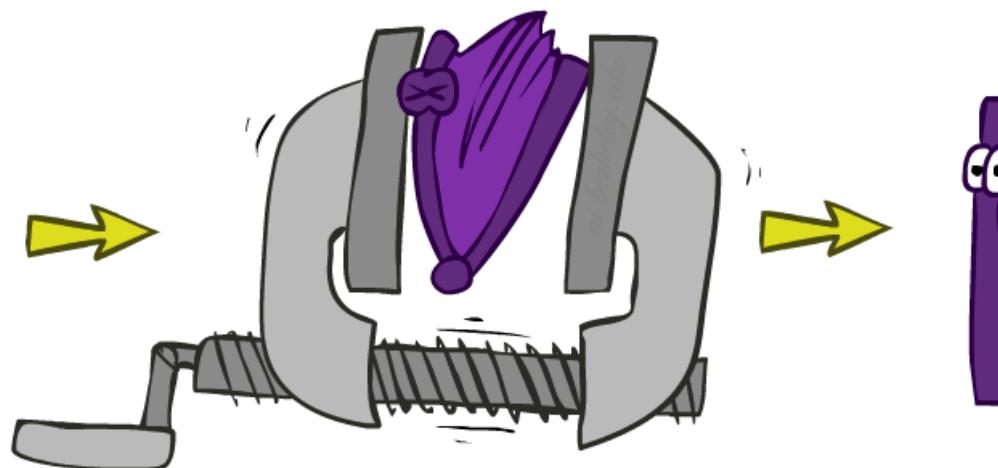
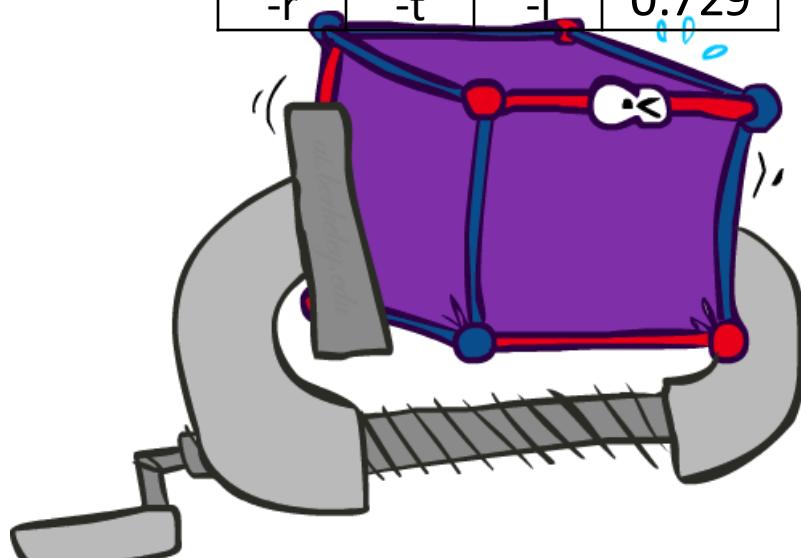
L

$P(T, L)$

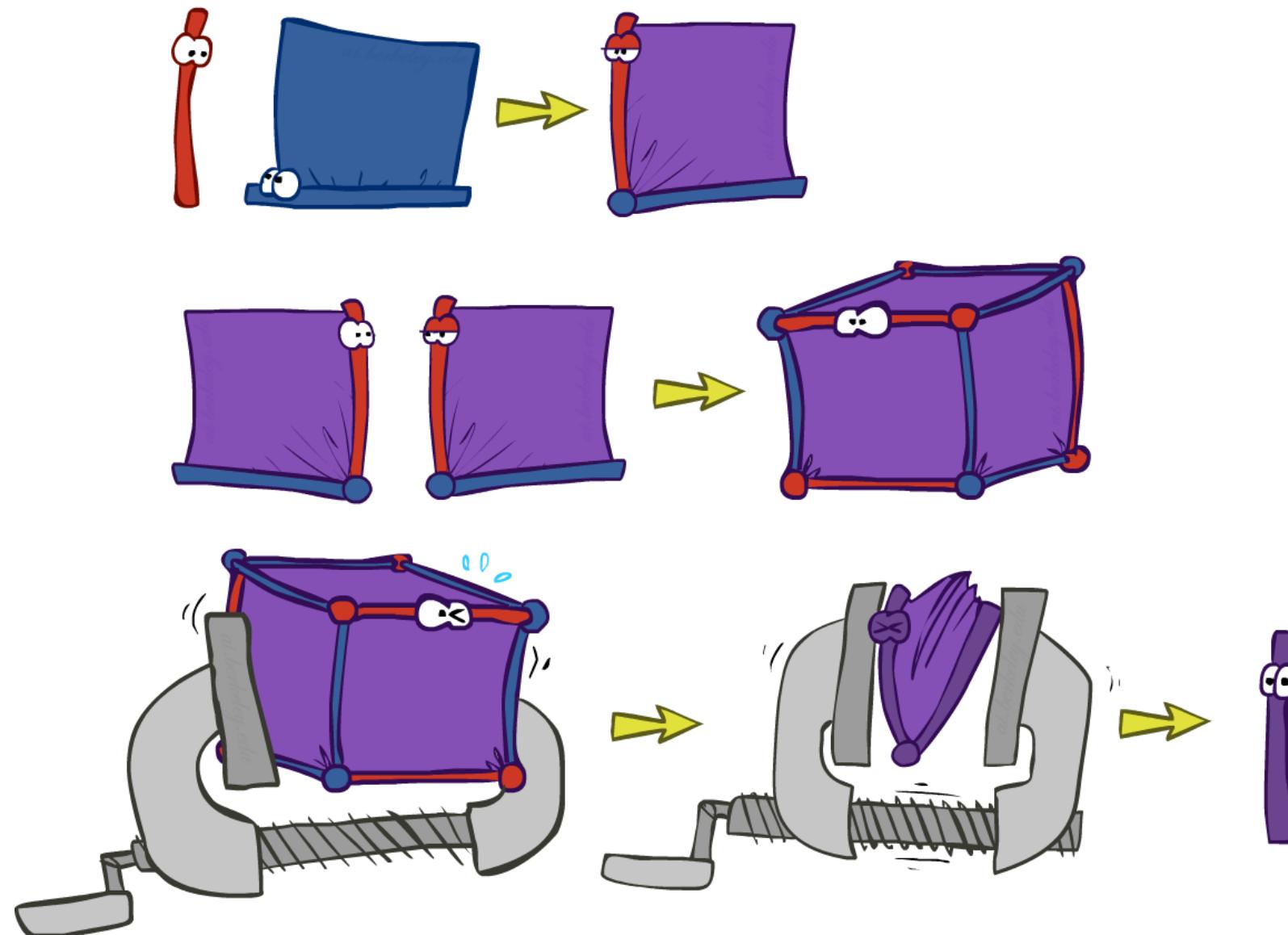
$+t$	$+l$	0.051
$+t$	$-l$	0.119
$-t$	$+l$	0.083
$-t$	$-l$	0.747

$P(L)$

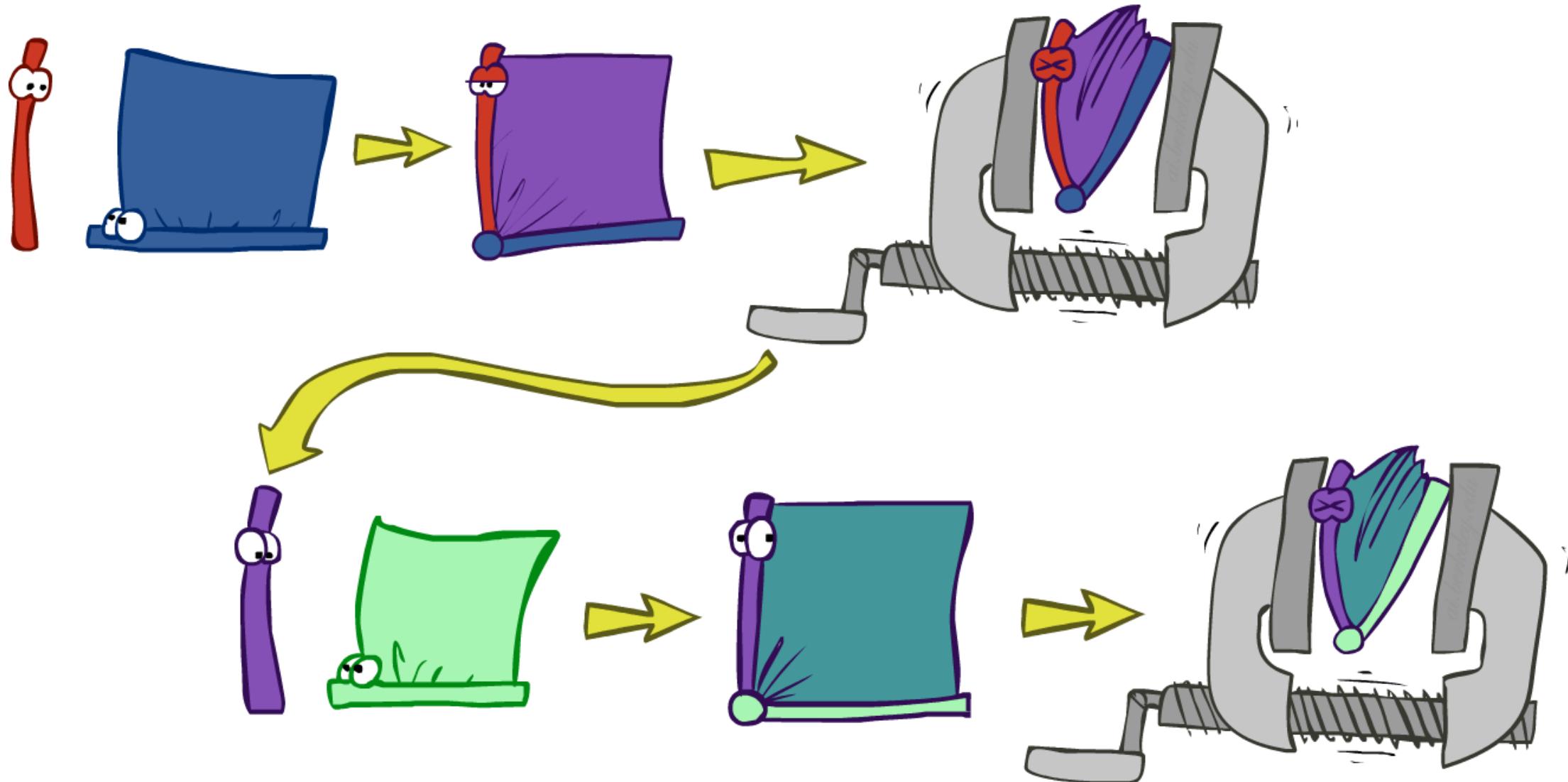
$+l$	0.134
$-l$	0.886



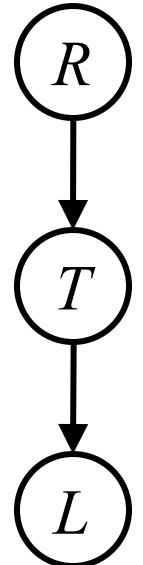
Thus Far: Multiple Join, Multiple Eliminate (= Inference by Enumeration)



Marginalizing Early (= Variable Elimination)



Traffic Domain



$$P(L) = ?$$

- Inference by Enumeration

$$= \sum_t \sum_r P(L|t) P(r) P(t|r)$$

Join on r

Join on t

Eliminate r

Eliminate t

- Variable Elimination

$$= \sum_t P(L|t) \sum_r P(r) P(t|r)$$

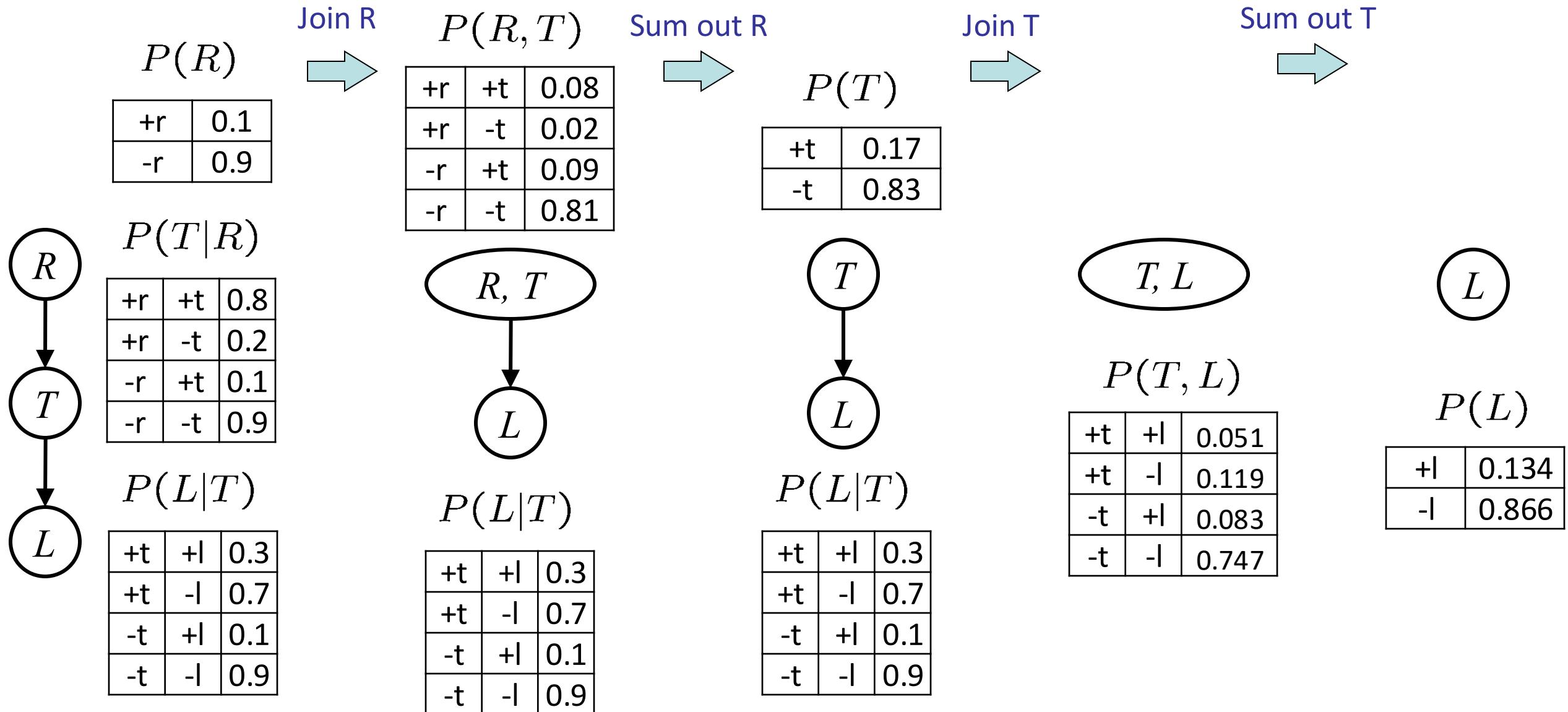
Join on r

Eliminate r

Join on t

Eliminate t

Marginalizing Early! (aka VE)



Evidence

- If evidence, start with factors that select that evidence

- No evidence uses these initial factors:

$$P(R)$$

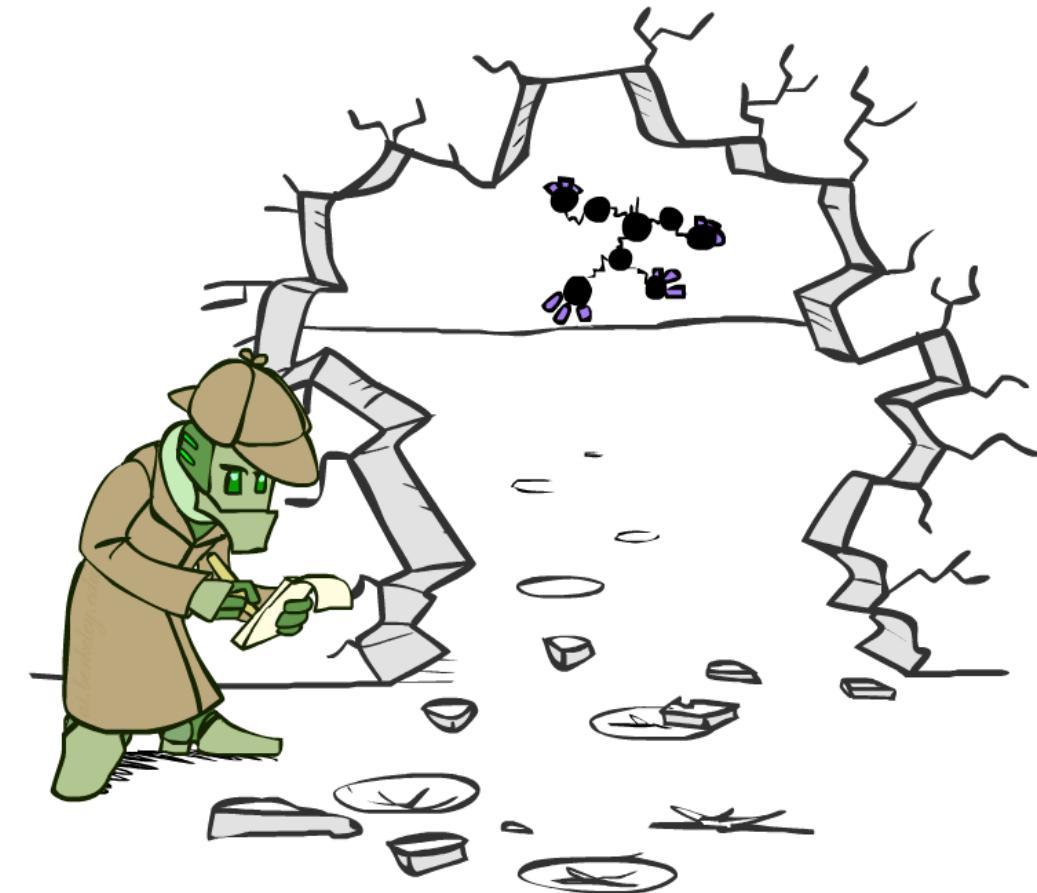
+r	0.1
-r	0.9

$$P(T|R)$$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9



- Computing $P(L|+r)$ the initial factors become:

$$P(+r)$$

+r	0.1
----	-----

$$P(T|+r)$$

+r	+t	0.8
+r	-t	0.2

$$P(L|T)$$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- We eliminate all vars other than query + evidence

Evidence II

- Result will be a selected joint of query and evidence
 - E.g. for $P(L | +r)$, we would end up with:

$$P(+r, L)$$

+r	+l	0.026
+r	-l	0.074

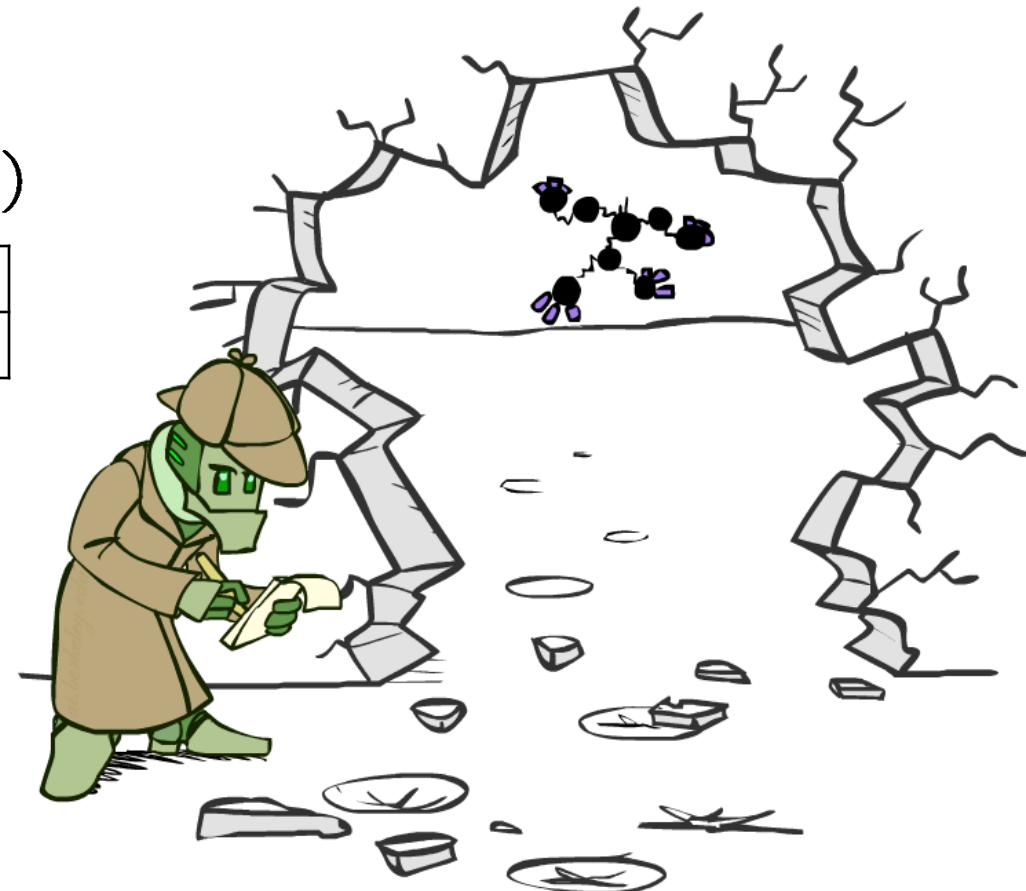
Normalize

$$P(L | +r)$$

+l	0.26
-l	0.74



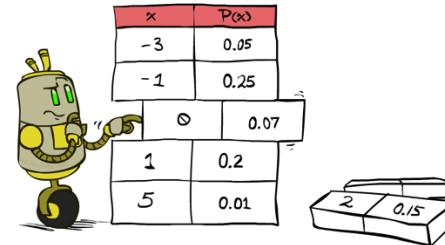
- To get our answer, just normalize this!
- That's it!



General Variable Elimination

- Query: $P(Q|E_1 = e_1, \dots, E_k = e_k)$

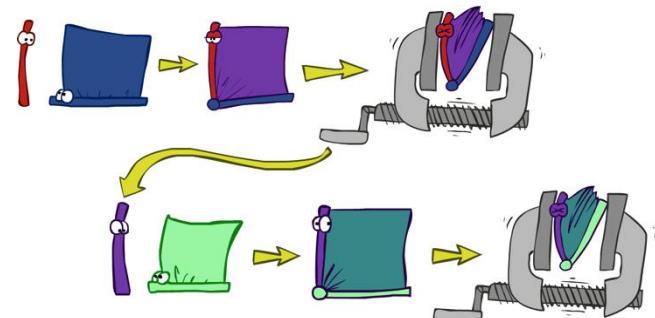
- Start with initial factors:
 - Local CPTs (but instantiated by evidence)



x	P(x)
-3	0.05
-1	0.25
0	0.07
1	0.2
5	0.01

- While there are still hidden variables (not Q or evidence):

- Pick a hidden variable H
 - Join all factors mentioning H
 - Eliminate (sum out) H



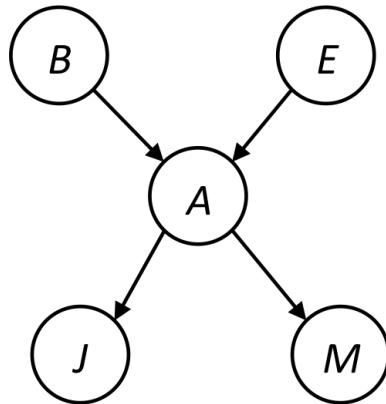
- Join all remaining factors and normalize

$$\text{f} \times \text{f} = \text{f} \times \frac{1}{Z}$$

Example

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

$P(B)$	$P(E)$	$P(A B, E)$	$P(j A)$	$P(m A)$
--------	--------	-------------	----------	----------

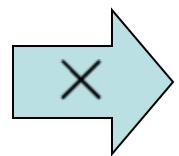


Choose A

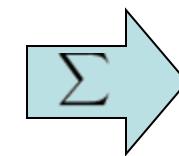
$$P(A|B, E)$$

$$P(j|A)$$

$$P(m|A)$$



$$P(j, m, A|B, E)$$



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

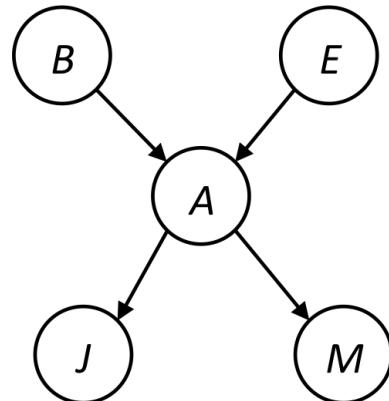
$P(B)$	$P(E)$	$P(j, m B, E)$
--------	--------	----------------

Example

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

Choose E

$$\begin{array}{ccccc} P(E) & \xrightarrow{\times} & P(j, m, E|B) & \xrightarrow{\sum} & P(j, m|B) \\ P(j, m|B, E) & & & & \end{array}$$



$P(B)$	$P(j, m B)$
--------	-------------

Finish with B

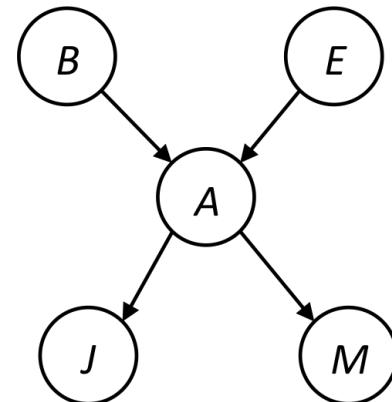
$$\begin{array}{ccccc} P(B) & \xrightarrow{\times} & P(j, m, B) & \xrightarrow{\text{Normalize}} & P(B|j, m) \\ P(j, m|B) & & & & \end{array}$$

Same Example in Equations

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

$P(B)$	$P(E)$	$P(A B, E)$	$P(j A)$	$P(m A)$
--------	--------	-------------	----------	----------

$$\begin{aligned}
 P(B|j, m) &\propto P(B, j, m) \\
 &= \sum_{e,a} P(B, j, m, e, a) \\
 &= \sum_{e,a} P(B)P(e)P(a|B, e)P(j|a)P(m|a) \\
 &= \sum_e P(B)P(e) \sum_a P(a|B, e)P(j|a)P(m|a) \\
 &= \sum_e P(B)P(e)f_1(B, e, j, m) \\
 &= P(B) \sum_e P(e)f_1(B, e, j, m) \\
 &= P(B)f_2(B, j, m)
 \end{aligned}$$



marginal can be obtained from joint by summing out

use Bayes' net joint distribution expression

use $x^*(y+z) = xy + xz$

joining on a, and then summing out gives f_1

use $x^*(y+z) = xy + xz$

joining on e, and then summing out gives f_2

Another Variable Elimination Example

Query: $P(X_3|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

Start by inserting evidence, which gives the following initial factors:

$$p(Z)p(X_1|Z)p(X_2|Z)p(X_3|Z)p(y_1|X_1)p(y_2|X_2)p(y_3|X_3)$$

Eliminate X_1 , this introduces the factor $f_1(Z, y_1) = \sum_{x_1} p(x_1|Z)p(y_1|x_1)$, and we are left with:

$$p(Z)f_1(Z, y_1)p(X_2|Z)p(X_3|Z)p(y_2|X_2)p(y_3|X_3)$$

Eliminate X_2 , this introduces the factor $f_2(Z, y_2) = \sum_{x_2} p(x_2|Z)p(y_2|x_2)$, and we are left with:

$$p(Z)f_1(Z, y_1)f_2(Z, y_2)p(X_3|Z)p(y_3|X_3)$$

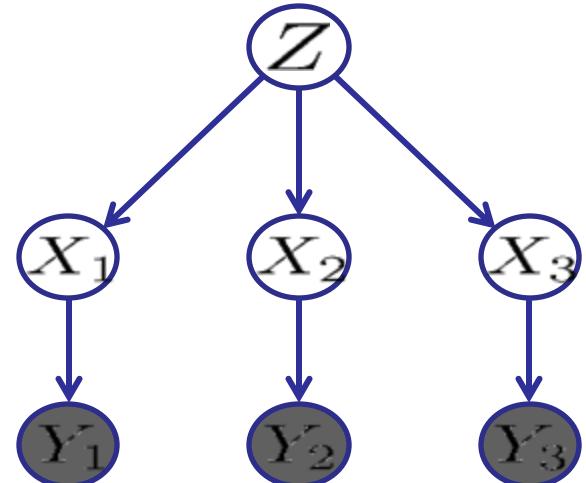
Eliminate Z , this introduces the factor $f_3(y_1, y_2, X_3) = \sum_z p(z)f_1(z, y_1)f_2(z, y_2)p(X_3|z)$, and we are left:

$$p(y_3|X_3), f_3(y_1, y_2, X_3)$$

No hidden variables left. Join the remaining factors to get:

$$f_4(y_1, y_2, y_3, X_3) = P(y_3|X_3)f_3(y_1, y_2, X_3).$$

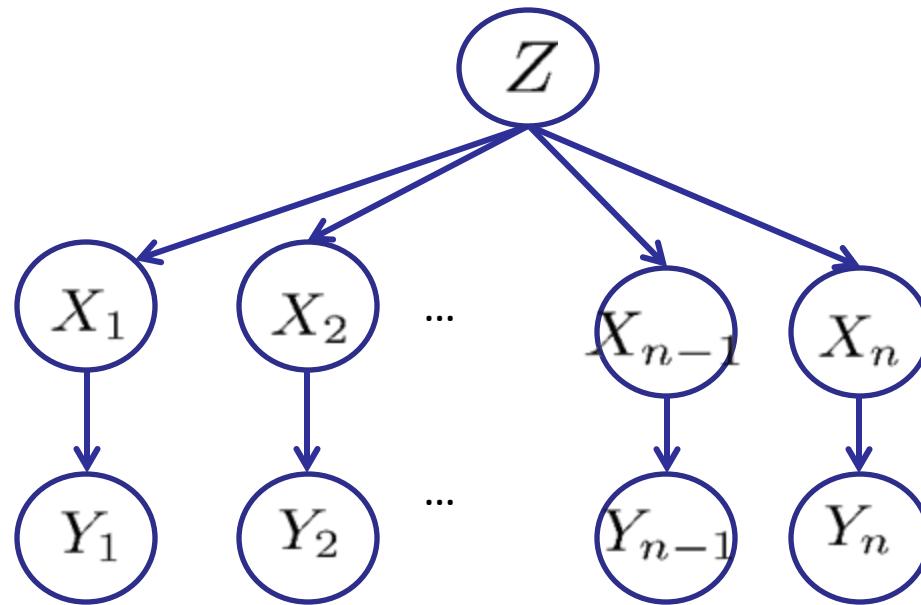
Normalizing over X_3 gives $P(X_3|y_1, y_2, y_3)$.



Computational complexity critically depends on the largest factor being generated in this process. Size of factor = number of entries in table. In example above (assuming binary) all factors generated are of size 2 --- as they all only have one variable (Z , Z , and X_3 respectively).

Variable Elimination Ordering

- For the query $P(X_n | y_1, \dots, y_n)$ work through the following two different orderings as done in previous slide: Z, X_1, \dots, X_{n-1} and X_1, \dots, X_{n-1}, Z . What is the size of the maximum factor generated for each of the orderings?



- Answer: 2^{n+1} versus 2^2 (assuming binary)
- In general: the ordering can greatly affect efficiency.

VE: Computational and Space Complexity

- The computational and space complexity of variable elimination is determined by the largest factor
- 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!

Worst Case Complexity?

- CSP:

$$(x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1 \vee x_3 \vee \neg x_4) \wedge (x_2 \vee \neg x_2 \vee x_4) \wedge (\neg x_3 \vee \neg x_4 \vee \neg x_5) \wedge (x_2 \vee x_5 \vee x_7) \wedge (x_4 \vee x_5 \vee x_6) \wedge (\neg x_5 \vee x_6 \vee \neg x_7) \wedge (\neg x_5 \vee \neg x_6 \vee x_7)$$

$$P(X_i = 0) = P(X_i = 1) = 0.5$$

$$Y_1 = X_1 \vee X_2 \vee \neg X_3$$

$$\dots \\ Y_8 = \neg X_5 \vee X_6 \vee X_7$$

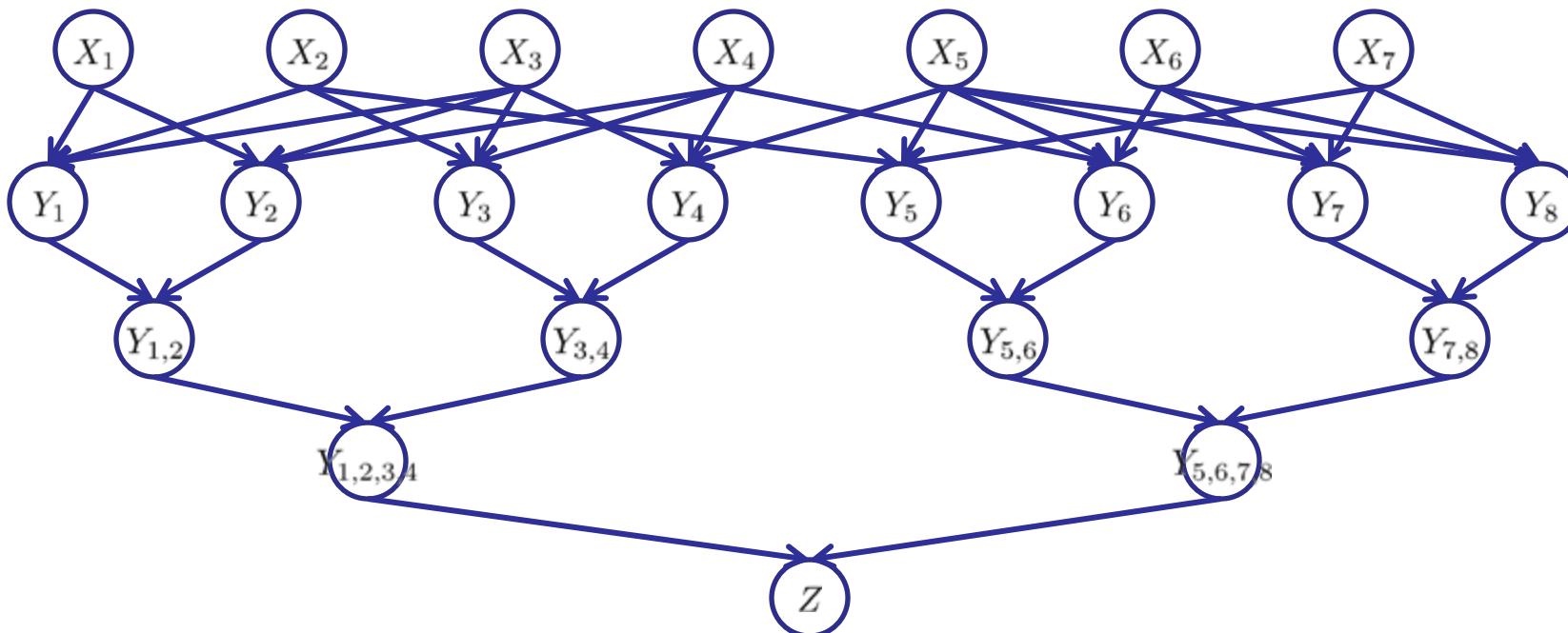
$$Y_{1,2} = Y_1 \wedge Y_2$$

$$\dots \\ Y_{7,8} = Y_7 \wedge Y_8$$

$$Y_{1,2,3,4} = Y_{1,2} \wedge Y_{3,4}$$

$$Y_{5,6,7,8} = Y_{5,6} \wedge Y_{7,8}$$

$$Z = Y_{1,2,3,4} \wedge Y_{5,6,7,8}$$



- If we can answer $P(z)$ equal to zero or not, we answered whether the 3-SAT problem has a solution.
- Hence inference in Bayes' nets is NP-hard. No known efficient probabilistic inference in general.

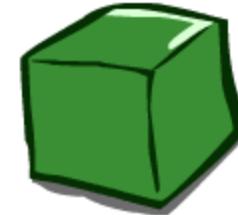
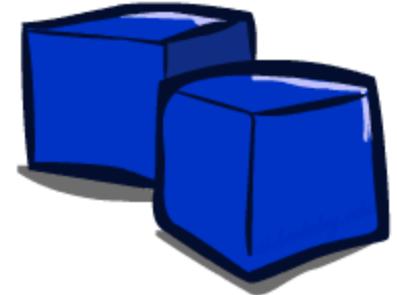
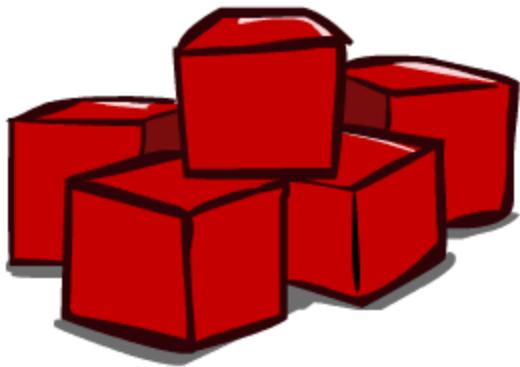
Polytrees

- A polytree is a directed graph with no undirected cycles
- For poly-trees you can always find an ordering that is efficient
 - Try it!!
- Cut-set conditioning for Bayes' net inference
 - Choose set of variables such that if removed only a polytree remains
 - Exercise: Think about how the specifics would work out!

Bayes' Nets

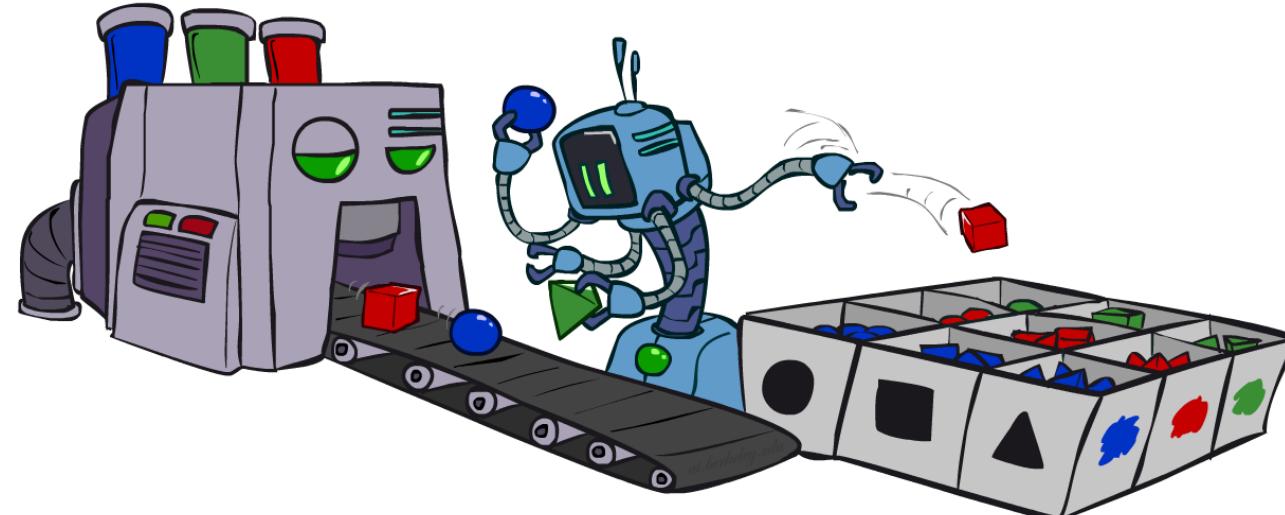
- ✓ Representation
- ✓ Conditional Independences
- Probabilistic Inference
 - ✓ Enumeration (exact, exponential complexity)
 - ✓ Variable elimination (exact, worst-case exponential complexity, often better)
 - ✓ Inference is NP-complete
 - Sampling (approximate)
- Learning Bayes' Nets from Data

Approximate Inference: Sampling



Sampling

- Sampling is a lot like repeated simulation
 - Predicting the weather, basketball games, ...
- Basic idea
 - Draw N samples from a sampling distribution S
 - Compute an approximate posterior probability
 - Show this converges to the true probability P
- Why sample?
 - Learning: get samples from a distribution you don't know
 - Inference: getting a sample is faster than computing the right answer (e.g. with variable elimination)



Sampling

- Sampling from given distribution

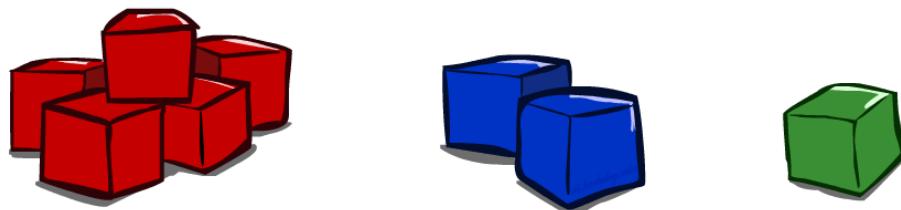
- Step 1: Get sample u from uniform distribution over $[0, 1]$
 - E.g. `random()` in python
- Step 2: Convert this sample u into an outcome for the given distribution by having each outcome associated with a sub-interval of $[0,1)$ with sub-interval size equal to probability of the outcome

- Example

C	P(C)
red	0.6
green	0.1
blue	0.3

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

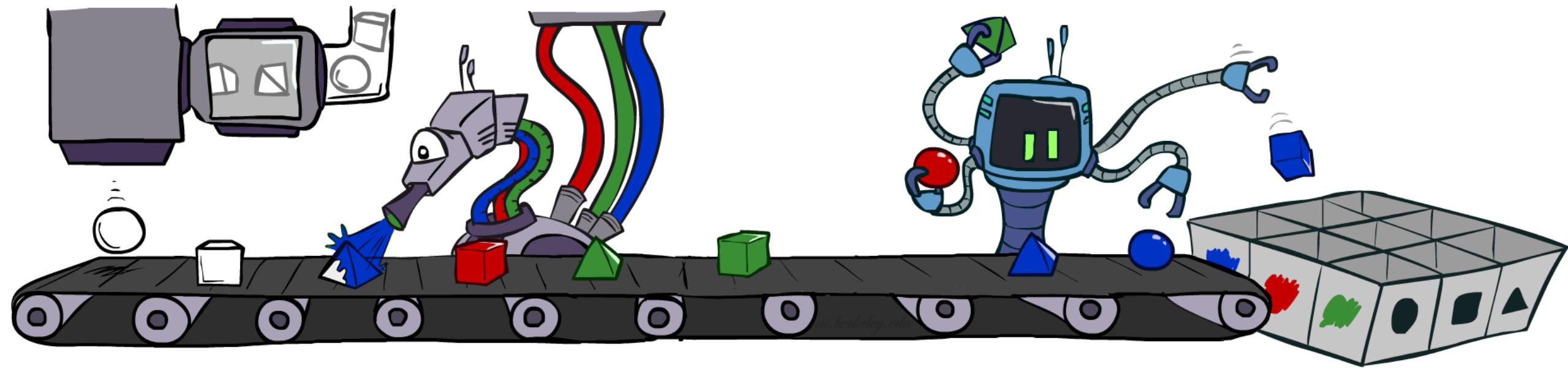
- If `random()` returns $u = 0.83$, then our sample is $C = \text{blue}$
- E.g, after sampling 8 times:



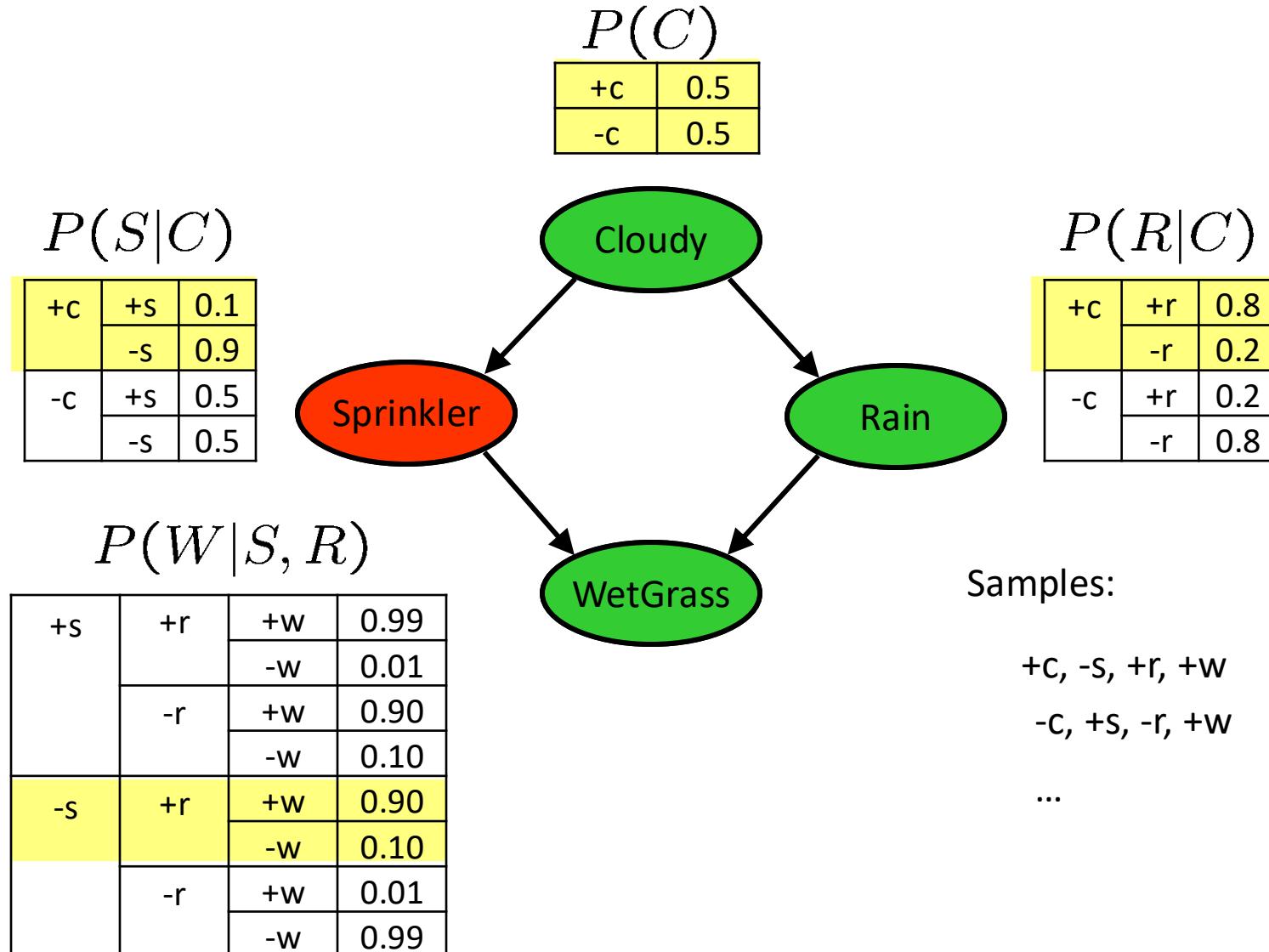
Sampling in Bayes' Nets

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

Prior Sampling

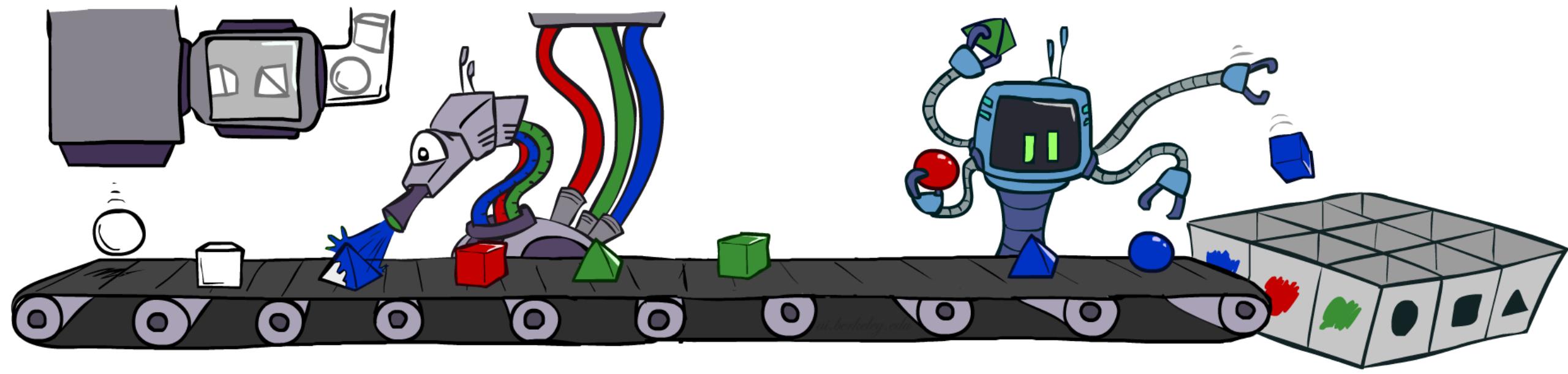


Prior Sampling



Prior Sampling

- For $i=1, 2, \dots, n$
 - Sample x_i from $P(X_i | \text{Parents}(X_i))$
 - Return (x_1, x_2, \dots, x_n)



Prior Sampling

- This process generates samples with probability:

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

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

- Let the number of samples of an event be $N_{PS}(x_1 \dots x_n)$
- Then $\lim_{N \rightarrow \infty} \hat{P}(x_1, \dots, x_n) = \lim_{N \rightarrow \infty} N_{PS}(x_1, \dots, x_n)/N$
 $= S_{PS}(x_1, \dots, x_n)$
 $= P(x_1 \dots x_n)$
- I.e., the sampling procedure is **consistent**

Example

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

+c, -s, +r, +w

+c, +s, +r, +w

-c, +s, +r, -w

+c, -s, +r, +w

-c, -s, -r, +w

- If we want to know $P(W)$

- We have counts $\langle +w:4, -w:1 \rangle$

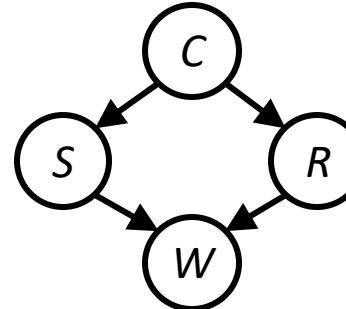
- Normalize to get $P(W) = \langle +w:0.8, -w:0.2 \rangle$

- This will get closer to the true distribution with more samples

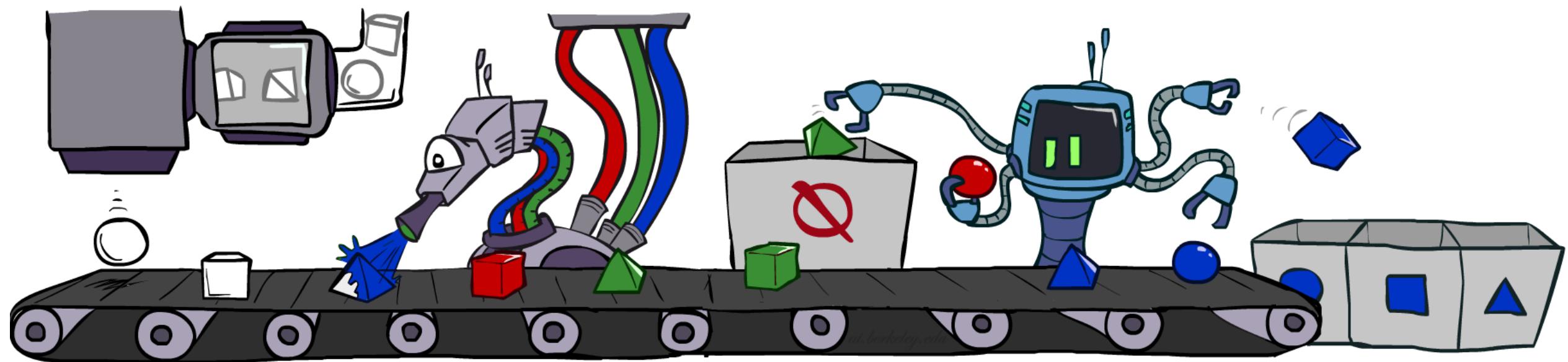
- Can estimate anything else, too

- What about $P(C| +w)$? $P(C| +r, +w)$? $P(C| -r, -w)$?

- Fast: can use fewer samples if less time (what's the drawback?)



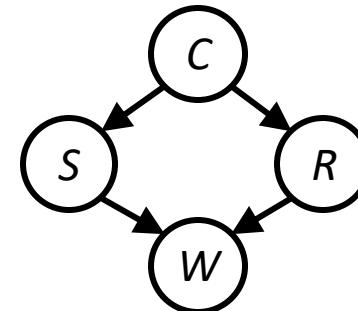
Rejection Sampling



Rejection Sampling

- Let's say we want $P(C)$
 - No point keeping all samples around
 - Just tally counts of C as we go

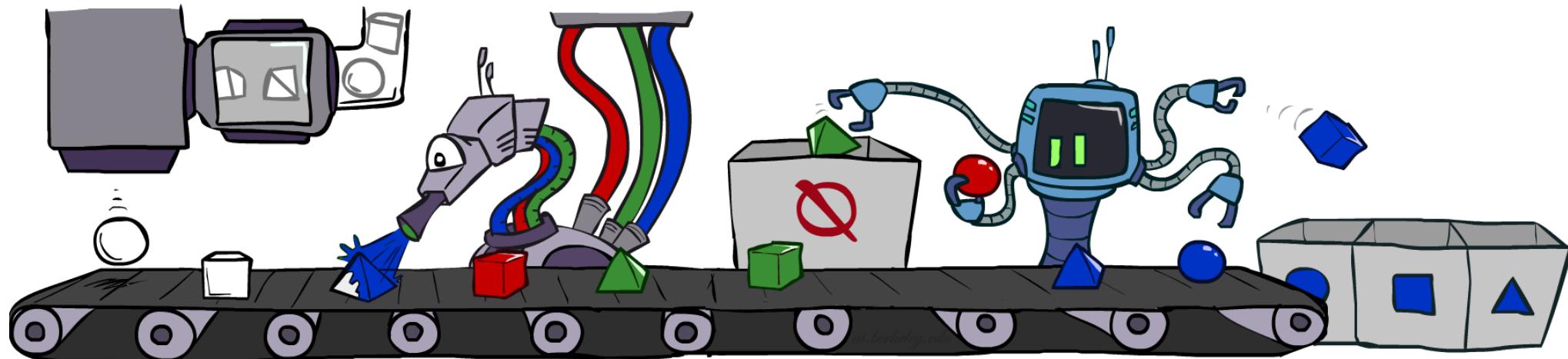
- Let's say we want $P(C| +s)$
 - Same thing: tally C outcomes, but ignore (reject) samples which don't have $S=+s$
 - This is called rejection sampling
 - It is also consistent for conditional probabilities (i.e., correct in the limit)



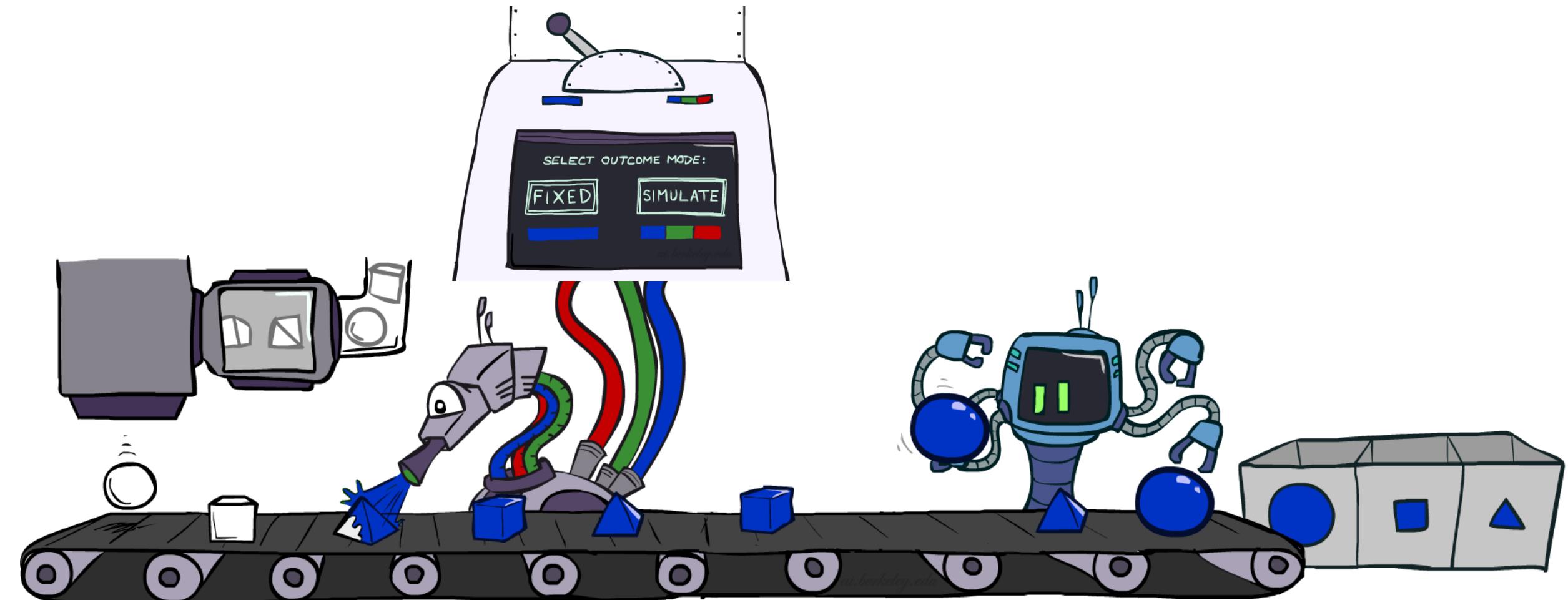
+c, -s, +r, +w
+c, +s, +r, +w
-c, +s, +r, -w
+c, -s, +r, +w
-c, -s, -r, +w

Rejection Sampling

- IN: evidence instantiation
- For $i=1, 2, \dots, n$
 - Sample x_i from $P(X_i | \text{Parents}(X_i))$
 - If x_i not consistent with evidence
 - Reject: Return, and no sample is generated in this cycle
 - Return (x_1, x_2, \dots, x_n)

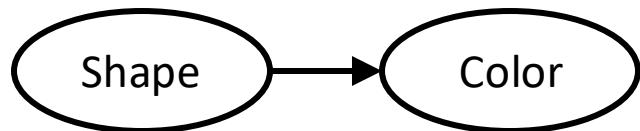


Likelihood Weighting



Likelihood Weighting

- Problem with rejection sampling:
 - If evidence is unlikely, rejects lots of samples
 - Evidence not exploited as you sample
 - Consider $P(\text{Shape} \mid \text{blue})$
- Idea: fix evidence variables and sample the rest
 - Problem: sample distribution not consistent!
 - Solution: weight by probability of evidence given parents

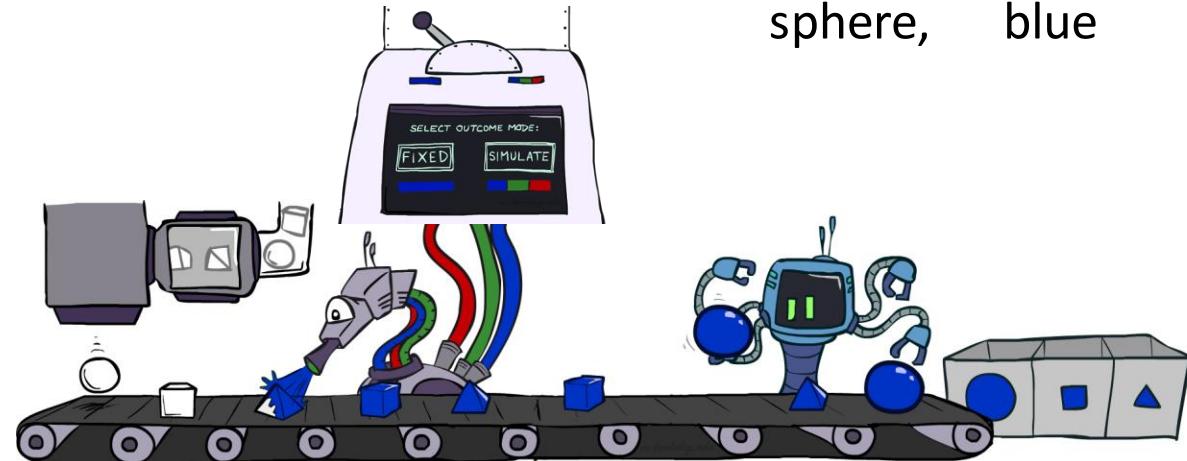
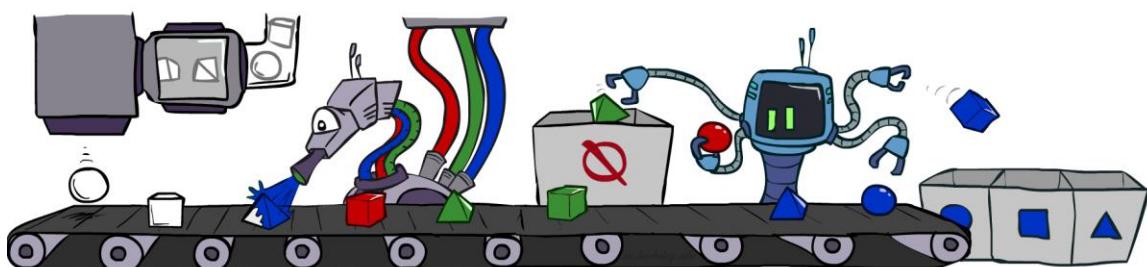


pyramid, green
pyramid, red
sphere, blue
cube, red
sphere, green



Shape → ~~Color~~

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



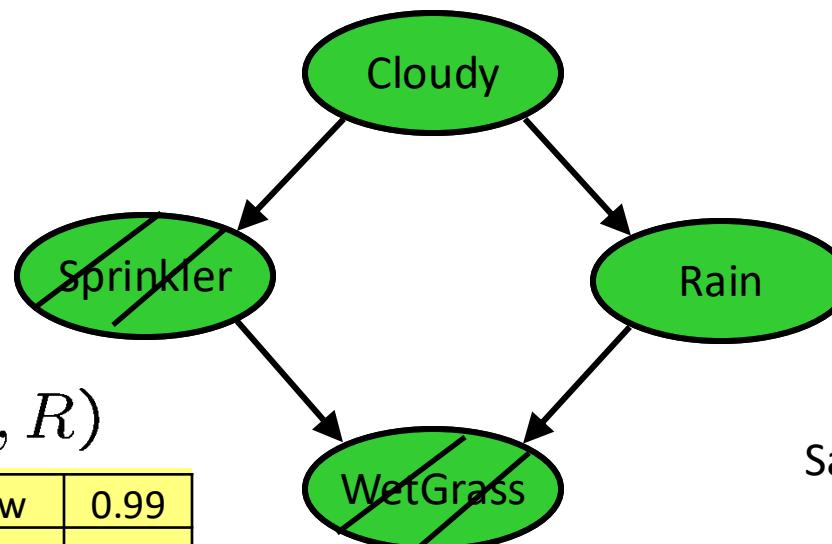
Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

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



$P(R|C)$

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

$P(W|S, R)$

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

Samples:

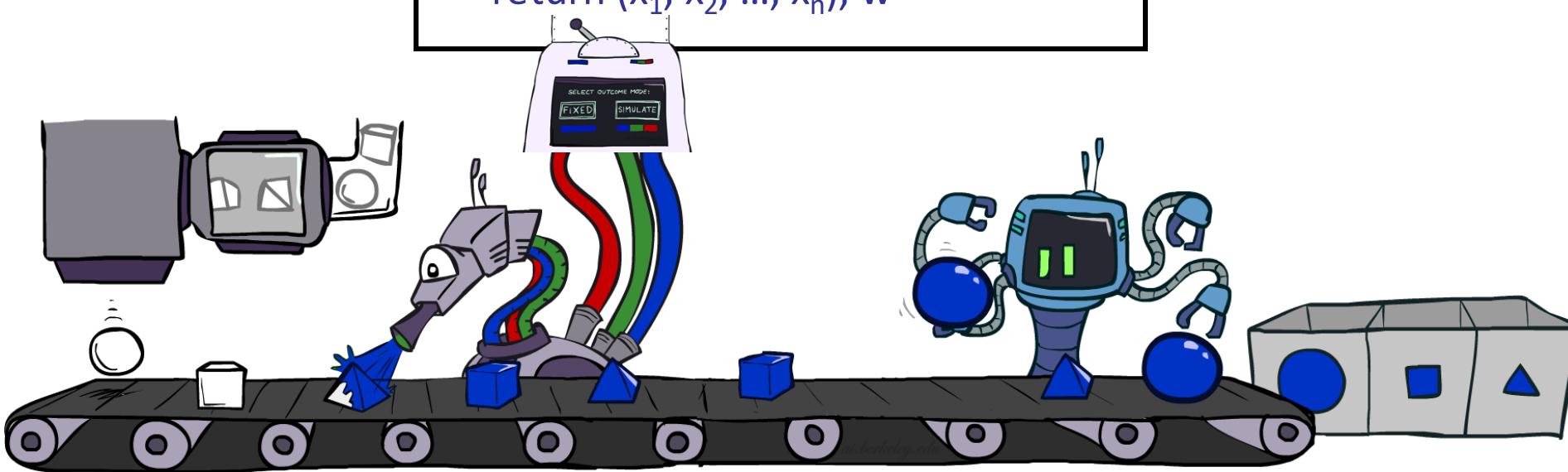
+c, +s, +r, +w

...

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

Likelihood Weighting

- IN: evidence instantiation
- $w = 1.0$
- for $i=1, 2, \dots, n$
 - if X_i is an evidence variable
 - $X_i = \text{observation } x_i \text{ for } X_i$
 - Set $w = w * P(x_i | \text{Parents}(X_i))$
 - else
 - Sample x_i from $P(X_i | \text{Parents}(X_i))$
- return $(x_1, x_2, \dots, x_n), w$



Likelihood Weighting

- Sampling distribution if \mathbf{z} sampled and \mathbf{e} fixed evidence

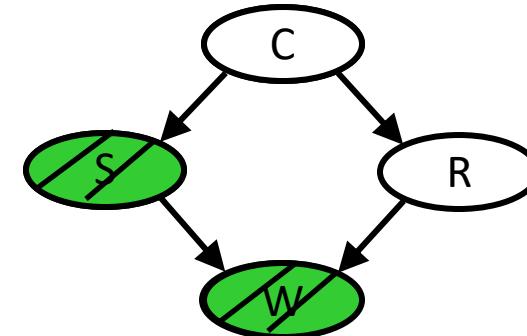
$$S_{WS}(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^l P(z_i | \text{Parents}(Z_i))$$

- Now, samples have weights

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

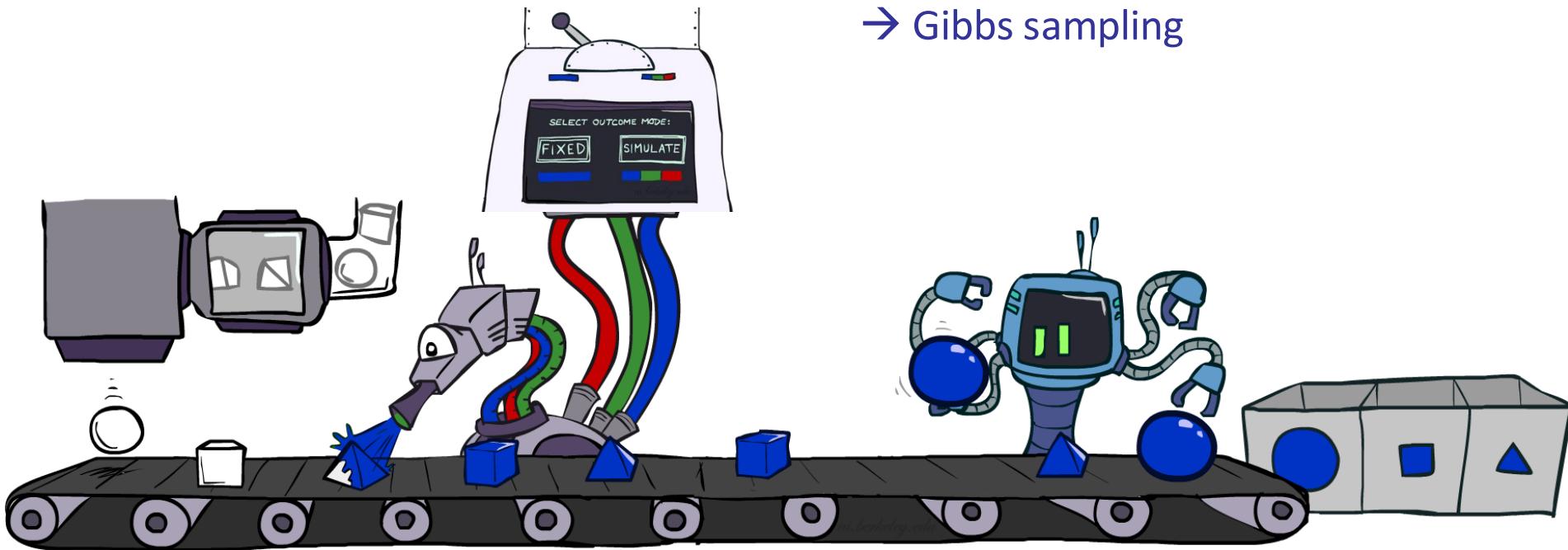
- Together, weighted sampling distribution is consistent

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

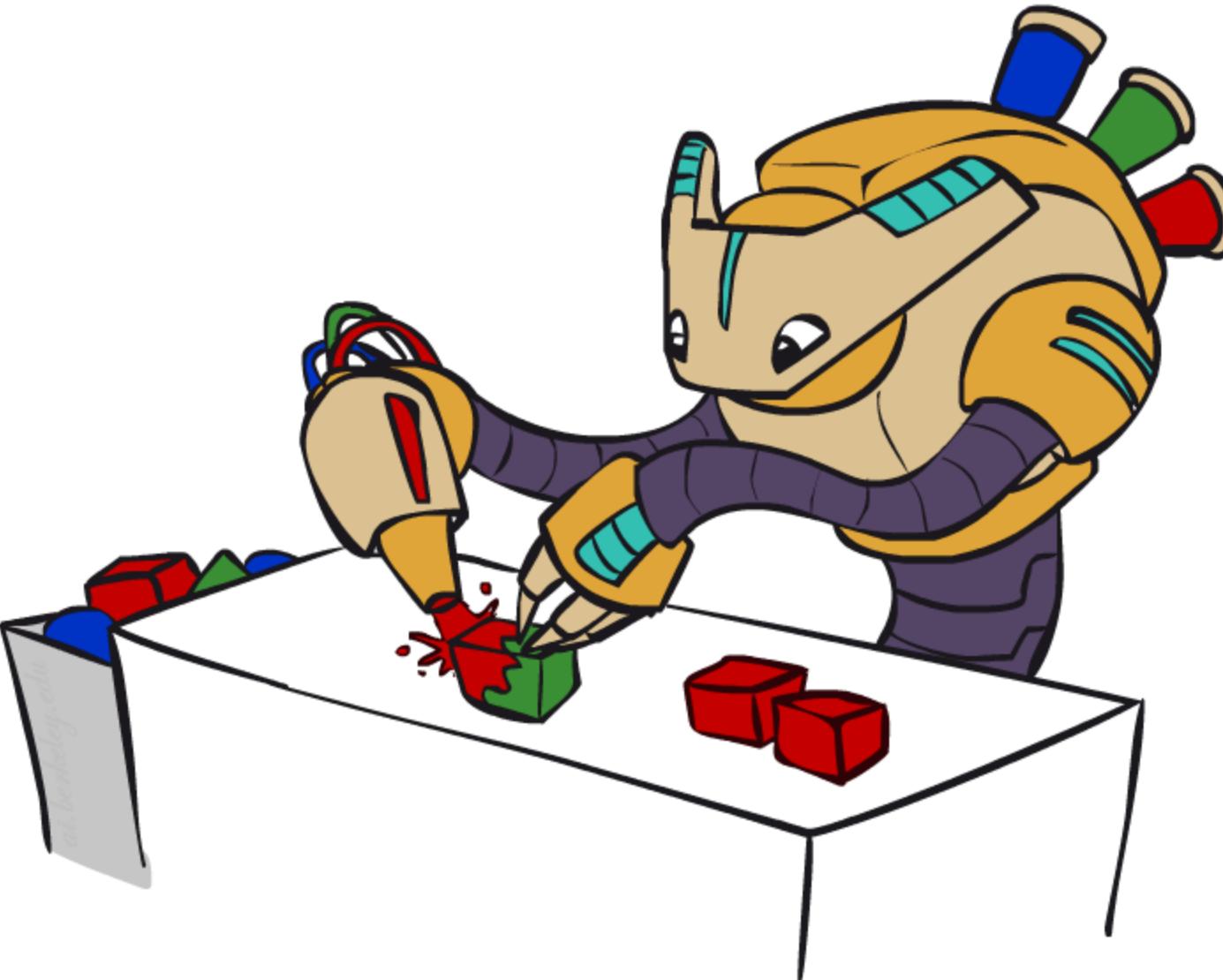


Likelihood Weighting

- 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
→ Gibbs sampling



Gibbs Sampling



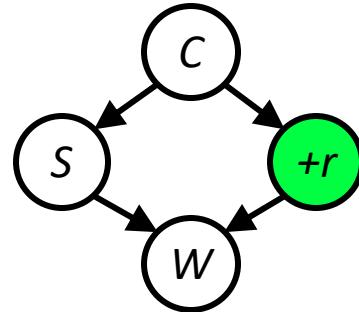
Gibbs Sampling

- *Procedure:* keep track of a full instantiation x_1, x_2, \dots, 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 sample is coming from the correct distribution
- *Rationale:* both upstream and downstream variables condition on evidence.
- In contrast: likelihood weighting only conditions on upstream evidence, and hence weights obtained in likelihood weighting can sometimes be very small. Sum of weights over all samples is indicative of how many “effective” samples were obtained, so want high weight.

Gibbs Sampling Example: $P(S | +r)$

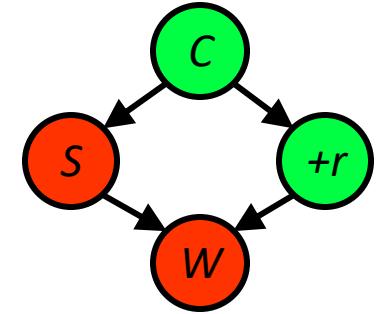
- Step 1: Fix evidence

- $R = +r$



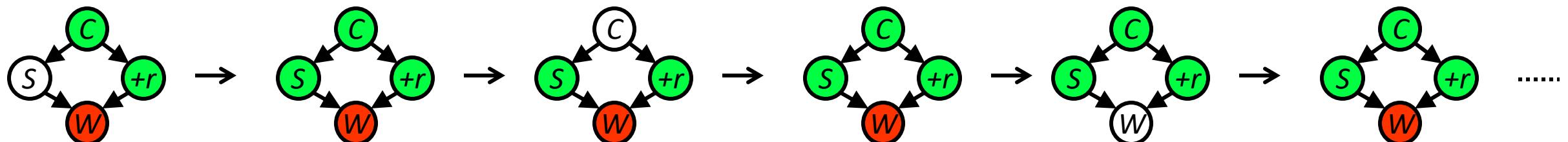
- Step 2: Initialize other variables

- Randomly



- Steps 3: Repeat

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



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

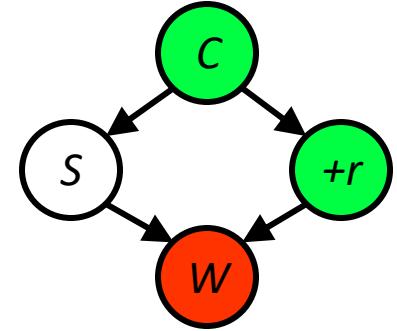
Sample from $P(C | +s, -w, +r)$

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

Efficient Resampling of One Variable

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

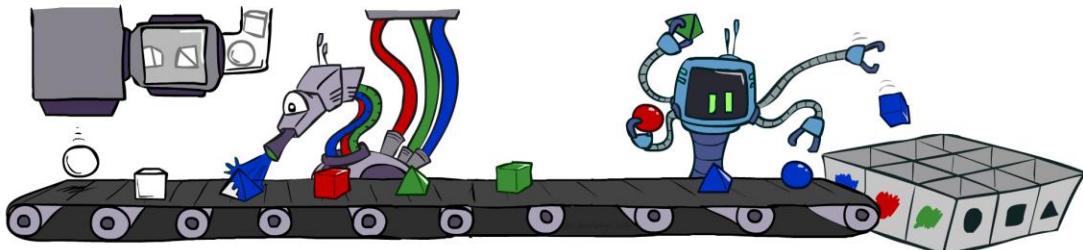
$$\begin{aligned} P(S | +c, +r, -w) &= \frac{P(S, +c, +r, -w)}{P(+c, +r, -w)} \\ &= \frac{P(S, +c, +r, -w)}{\sum_s P(s, +c, +r, -w)} \\ &= \frac{P(+c)P(S | +c)P(+r | +c)P(-w | S, +r)}{\sum_s P(+c)P(s | +c)P(+r | +c)P(-w | s, +r)} \\ &= \frac{P(+c)P(S | +c)P(+r | +c)P(-w | S, +r)}{P(+c)P(+r | +c) \sum_s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



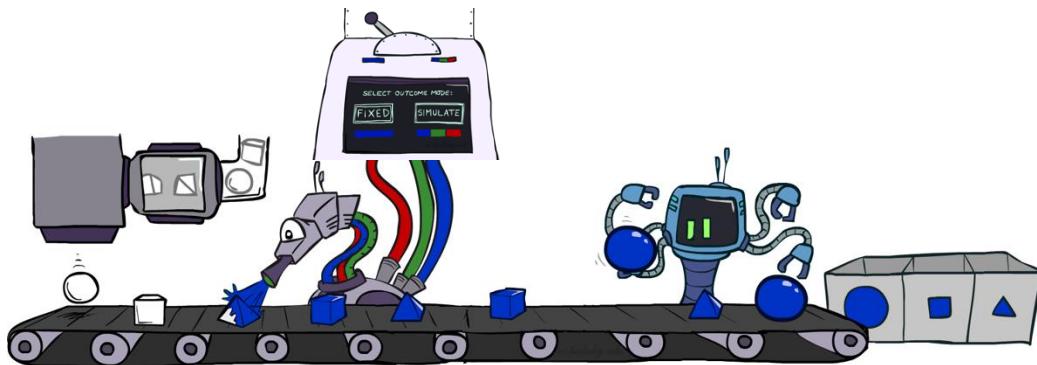
- Many things cancel out – only CPTs with S remain!
- More generally: only CPTs that have resampled variable need to be considered, and joined together

Bayes' Net Sampling Summary

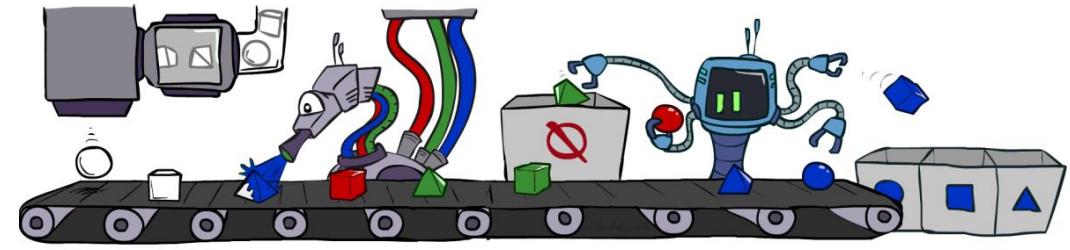
- Prior Sampling P



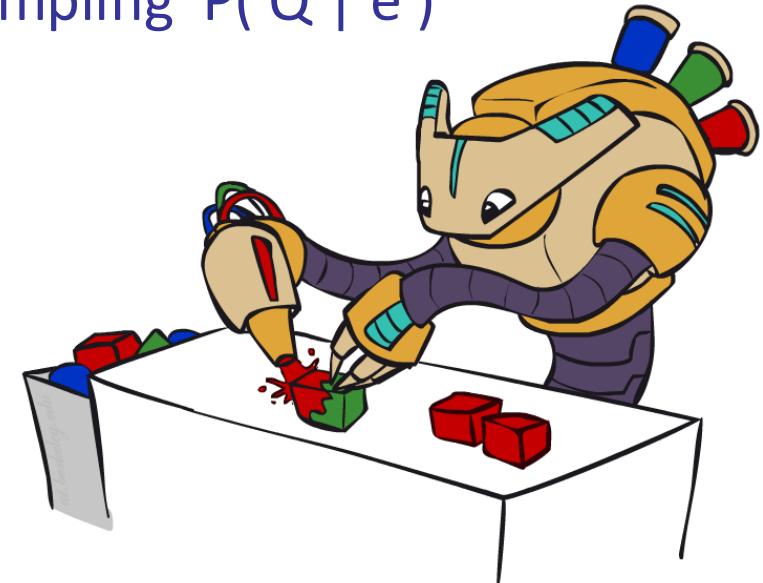
- Likelihood Weighting $P(Q | e)$



- Rejection Sampling $P(Q | e)$



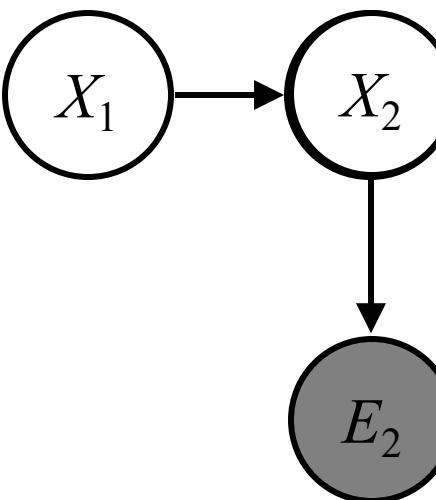
- Gibbs Sampling $P(Q | e)$



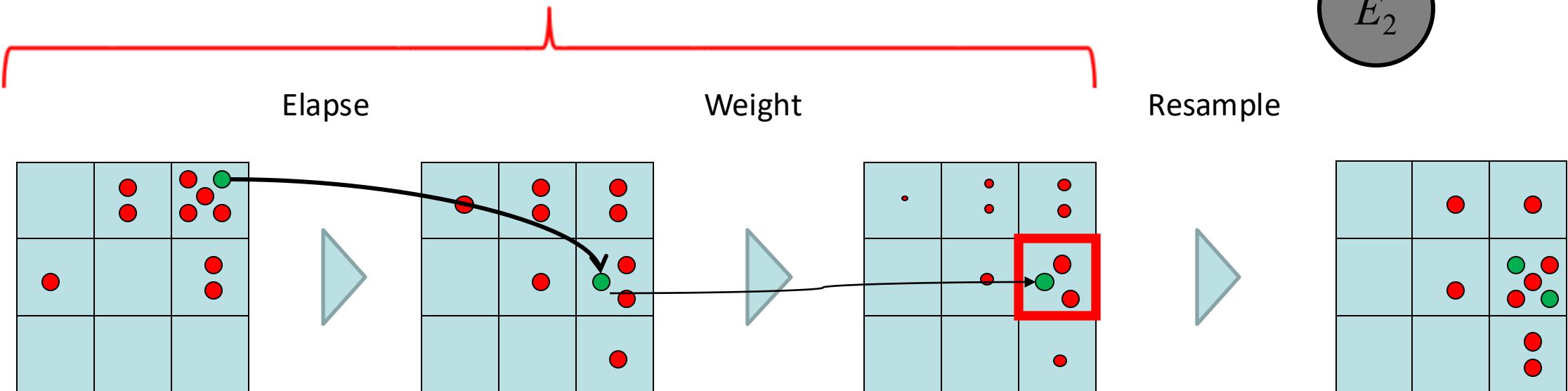
Further Reading on Gibbs Sampling*

- 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

How About Particle Filtering?



= likelihood weighting



Particles:

(3,3)

(2,3)

(3,3)

(3,2)

(3,3)

(3,2)

(1,2)

(3,3)

(3,3)

(2,3)

Particles:

(3,2)

(2,3)

(3,2)

(3,1)

(3,3)

(3,2)

(1,3)

(2,3)

(3,2)

(2,2)

Particles:

(3,2) w=.9

(2,3) w=.2

(3,2) w=.9

(3,1) w=.4

(3,3) w=.4

(3,2) w=.9

(1,3) w=.1

(2,3) w=.2

(3,2) w=.9

(2,2) w=.4

(New) Particles:

(3,2)

(2,2)

(3,2)

(2,3)

(3,3)

(3,2)

(1,3)

(2,3)

(3,2)

(3,2)

Particle Filtering

- Particle filtering operates on ensemble of samples
 - Performs likelihood weighting for each individual sample to elapse time and incorporate evidence
 - Resamples from the weighted ensemble of samples to focus computation for the next time step where most of the probability mass is estimated to be