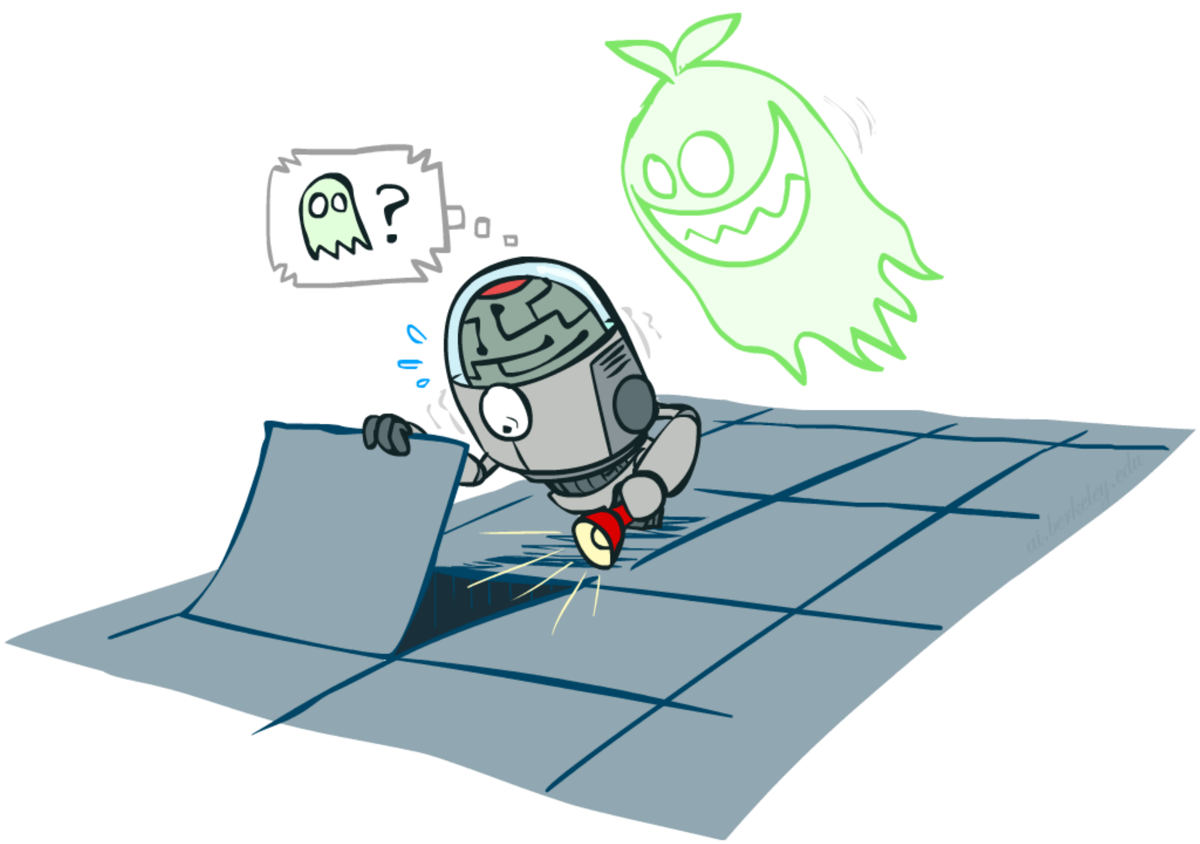
Where are we now in Ve492?

- We're done with Part I Search and Planning!
- Part II: Probabilistic Reasoning
  - Diagnosis
  - Speech recognition
  - Tracking objects
  - Robot mapping
  - Genetics
  - Error correcting codes
  - … lots more!
- Part III: Machine Learning
- Part IV: Logic



Ve492: Introduction to Artificial Intelligence
Probability Review

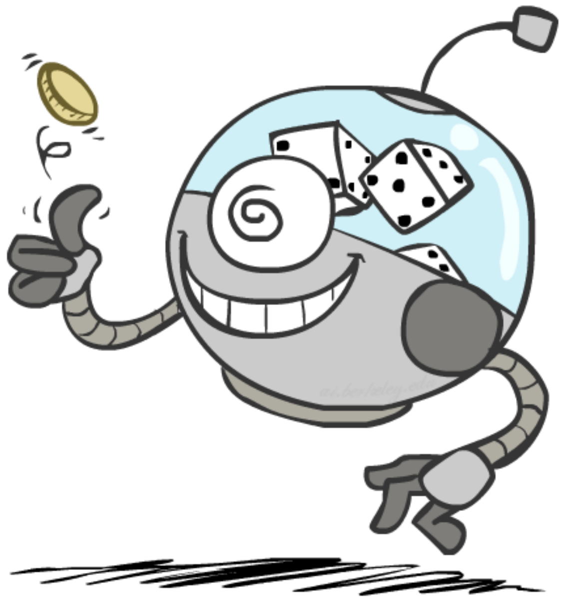
Paul Weng

UM-SJTU Joint Institute

Slides adapted from http://ai.berkeley.edu, CMU, AIMA, UM

Today

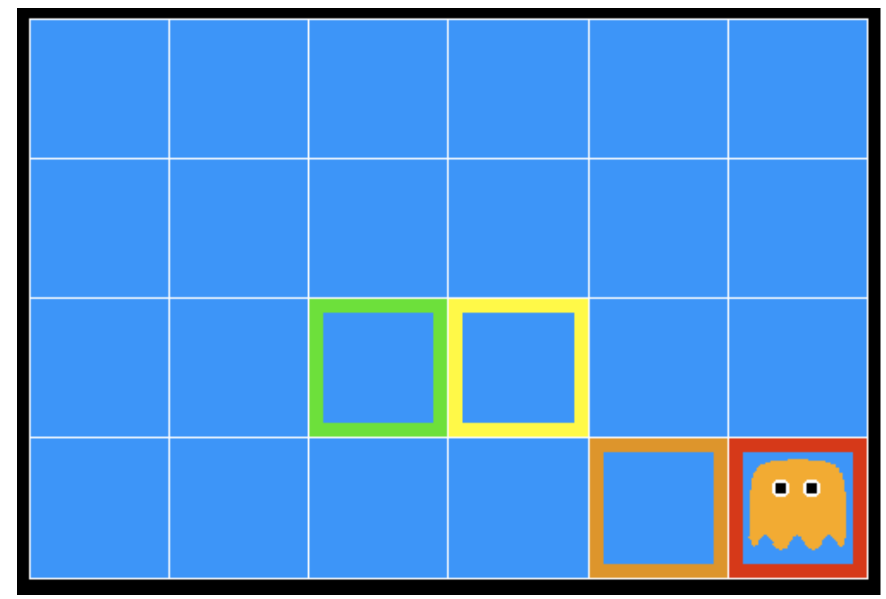
- Probability
  - Random Variables
  - Joint and Marginal Distributions
  - Conditional Distribution
  - Product Rule, Chain Rule, Bayes' Rule
  - Inference
  - (Conditional) Independence



- You'll need all this stuff A LOT for the next few weeks, so make sure you go over it now!

Inference in Ghostbusters

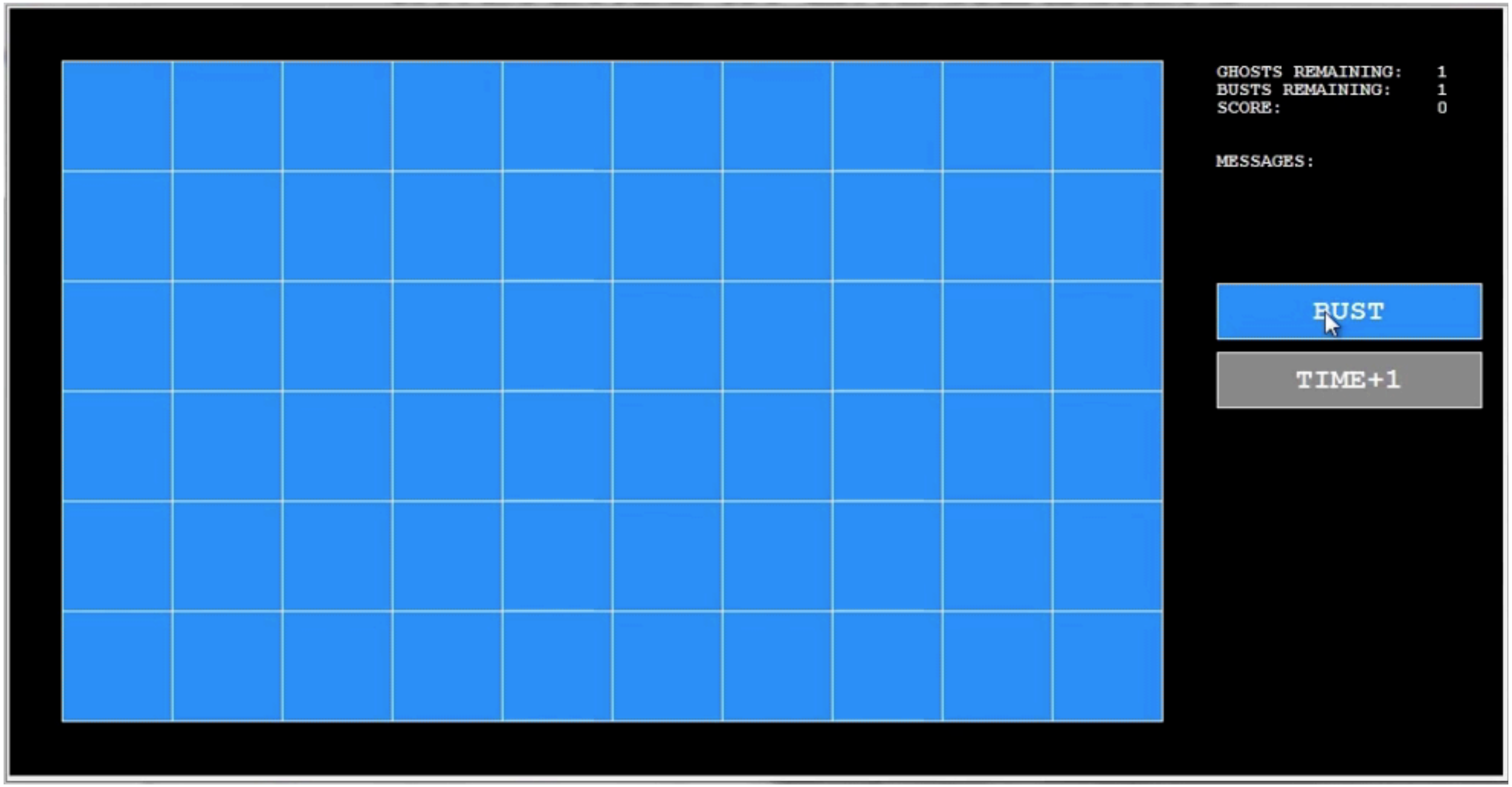
- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - On the ghost: red
  - 1 or 2 away: orange
  - 3 or 4 away: yellow
  - 5+ away: green



- Sensors are noisy, but we know P(Color | Distance)

P(red | 3)	P(orange | 3)	P(yellow | 3)	P(green | 3)
0.05	0.15	0.5	0.3

Video of Demo Ghostbuster – No probability



Uncertainty

- General situation:

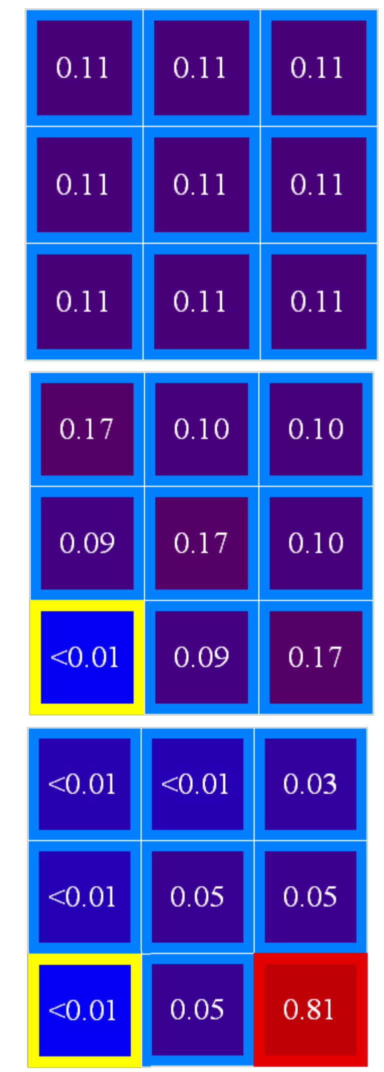
  - **Observed variables (evidence)**: Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)

  Hidden / Latent

  - **Unobserved variables**: Agent needs to reason about other aspects (e.g. where an object is or what disease is present)

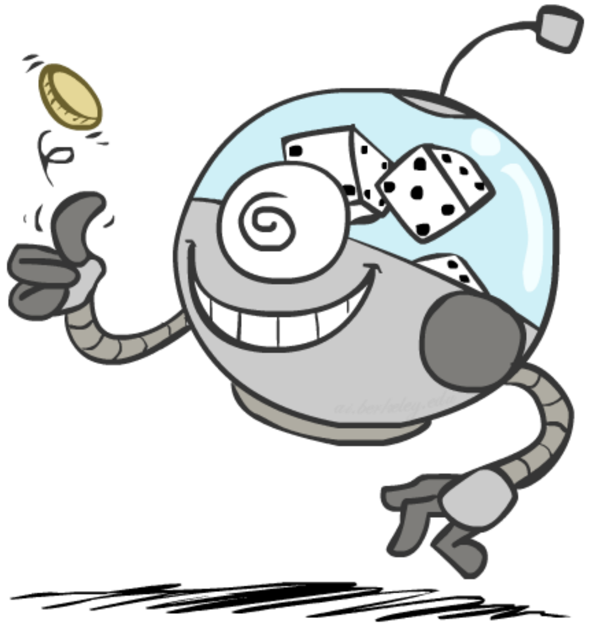
  - **Model**: Agent knows something about how the known variables relate to the unknown variables

- Probabilistic reasoning gives us a framework for managing uncertain beliefs and knowledge



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?

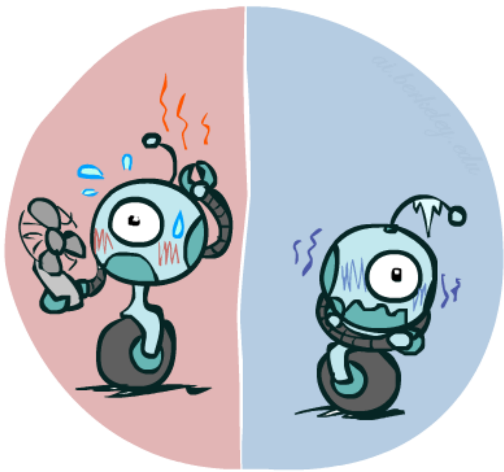


- We denote random variables with capital letters
- Like variables in a CSP, random variables have domains
  - R in {true, false}   (often write as {+r, -r})
  - T in {hot, cold}
  - D in $[0, \infty)$
  - L in possible locations, maybe {(0,0), (0,1), …}

Probability Distributions

- Associate a probability with each value

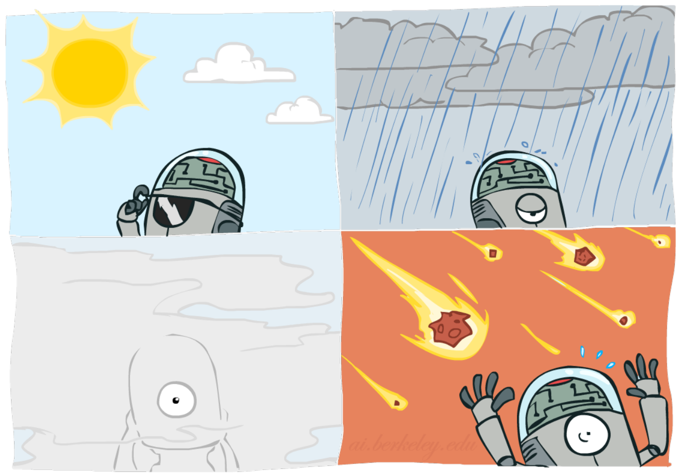
  - Temperature:



$P(T)$

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

- Random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\ldots$$

OK *if* all domain entries are unique

- A distribution is a TABLE of probabilities of values

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

$$P(W = rain) = 0.1$$

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

Joint Distributions

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

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

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

$P(T, W)$

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

  - Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

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

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

$O(d^n)$

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables

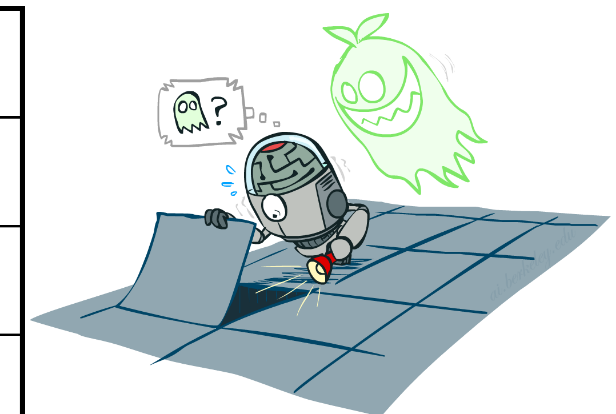
- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called *outcomes*
  - Joint distributions: say whether assignments (outcomes) are likely
  - *Normalized:* sum to 1.0
  - Ideally: only certain variables directly interact

- 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

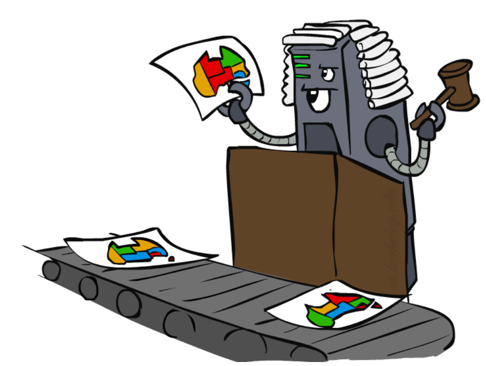
[0, 1]



Constraint over T,W

T	W	P	
hot	sun	T	1
hot	rain	F	0
cold	sun	F	
cold	rain	T	

{0, 1}



Events

- An *event* is a set E of outcomes

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

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny? 0.4
  - Probability that it's hot? 0.5
  - Probability that it's hot OR sunny? 0.7
- Typically, the events we care about are *partial assignments*, like P(T=hot)

$P(T, W)$

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

Quiz: Events

- P(+x, -y) ?   0.3

- P(+y) ?   0.6

- P(-y OR +x) ?   0.6

$P(X, Y)$

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

$$P(t) = \sum_w P(t, w)$$

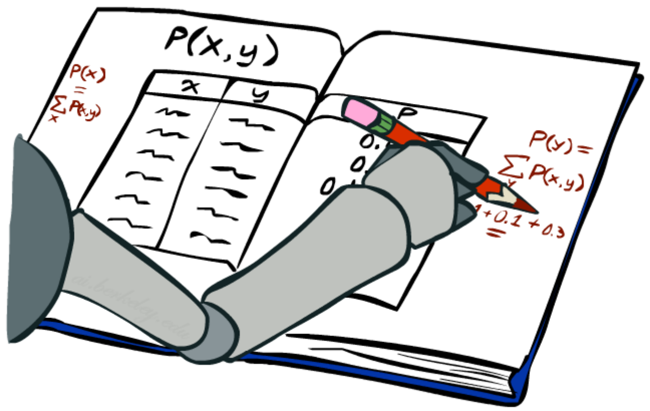
$$P(w) = \sum_t P(t, w)$$

$P(T)$

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

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

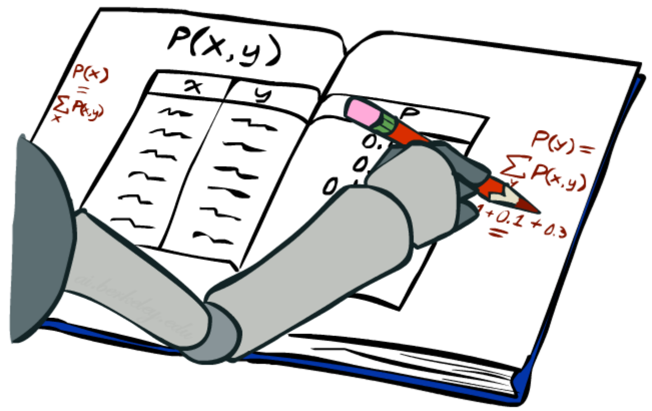
$$P(y) = \sum_x P(x, y)$$

$P(X)$

X	P
+x	0.5
-x	0.5

$P(Y)$

Y	P
+y	0.6
-y	0.4



Conditional Distributions

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

$P(\underline{X}, \underline{Y})$

$\underline{P}(\underline{X} \mid \underline{Y})$

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

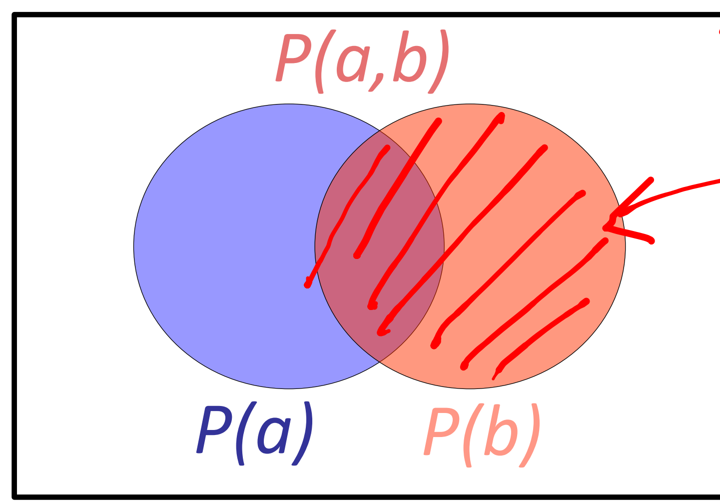
T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Conditional Probabilities

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

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

Ω



$P(T, W)$

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

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

$$= P(W=s, T=c) + P(W=r, T=c)$$
$$= 0.2 + 0.3 \quad = 0.5$$

Quiz: Conditional Probabilities

- Provide answers as fractions
- $P(+x \mid +y)$? $= \dfrac{0.2}{0.6} = \dfrac{1}{3}$
- $P(-x \mid +y)$? $= \dfrac{2}{3}$
- $P(-y \mid +x)$? $= \dfrac{3}{5}$

$$P(X,Y)$$

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

How to compute a whole conditional distribution at once?

$$P(T, W)$$

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

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

$$\longrightarrow$$

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

$$P(W|T=c)$$

W	P
sun	0.4
rain	0.6

Normalization Trick

$P(T, W)$

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=c

$P(c, W)$

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

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

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

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(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

- $P(X=+x \mid Y=-y)$? Provide the answer as a fraction.

$P(X, Y)$

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

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

NORMALIZE the selection (make it sum to one)

X	P
+x	3/4
-x	1/4

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

- We generally compute conditional probabilities
  - P(on time | no reported accidents) = 0.90
  - These represent the agent's *beliefs* given the evidence

- Probabilities change with new evidence:
  - P(on time | no accidents, 5 a.m.) = 0.95
  - P(on time | 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: $H_1 \dots H_r$

$X_1, X_2, \dots X_n$

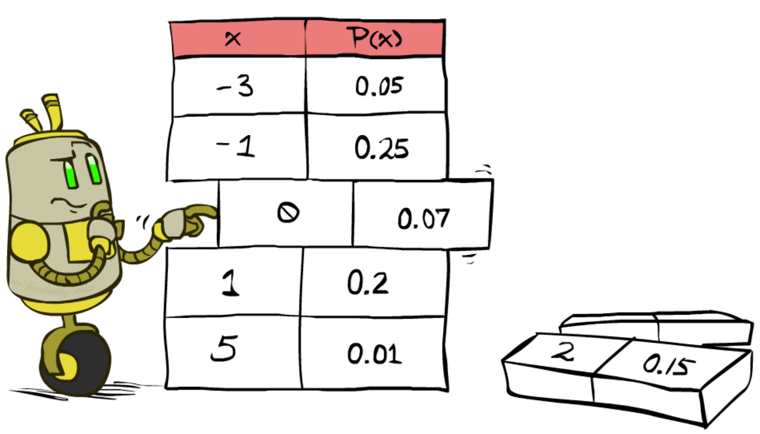
All variables

- We have the joint and we want:

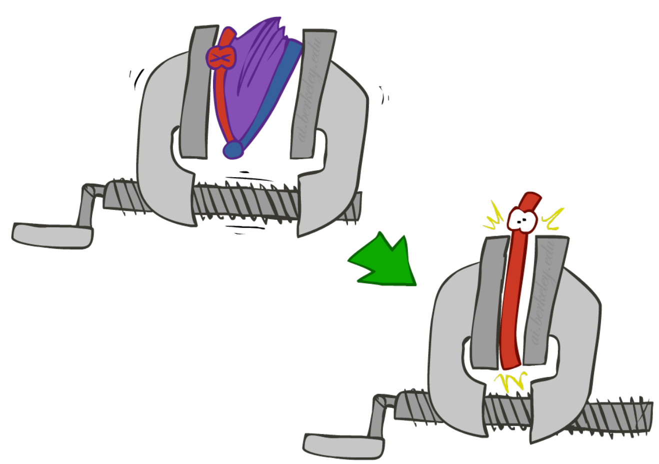
$$P(Q|e_1 \dots e_k)$$

* *Works fine with multiple query variables, too*

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out $H_1, \dots, H_r$ to get the joint of Q and evidence



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

- Step 3: Normalize

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

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

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

Inference by Enumeration

- P(W)?

  P(W = sun) = 0.65

  P(W = rain) = 0.35

- P(W | winter)?

  P(W = sun | winter) = 0.5

  P(W = rain | winter) = 0.5

- P(W | winter, hot)?

  sun = $\frac{2}{3}$

  rain = $\frac{1}{3}$

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

0.25

0.25

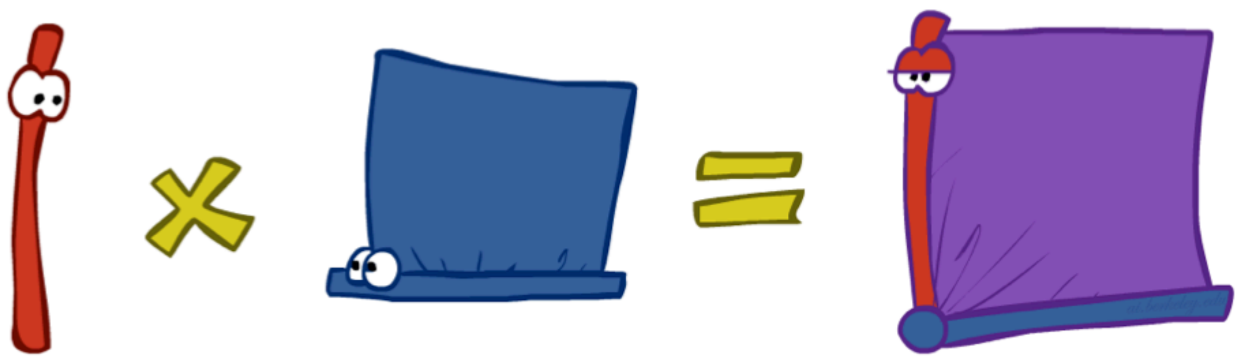
Inference by Enumeration

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

The Product Rule

- Sometimes we have conditional and marginal 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:

$P(W)$

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

$\Leftrightarrow$

$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

The Chain Rule

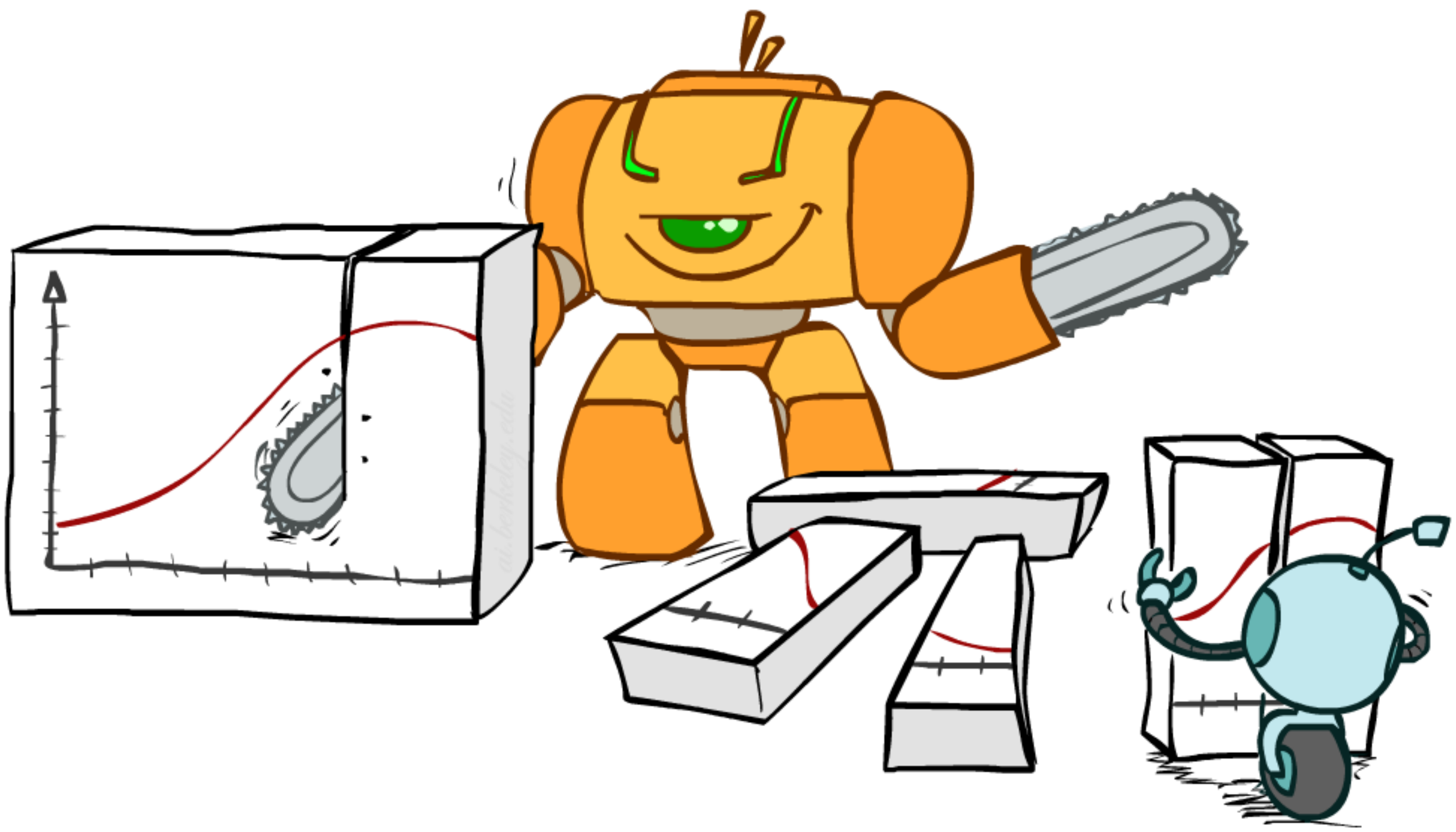
- More generally, we can always write any joint distribution as an incremental product of conditional distributions

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

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

- Why is this always true?

Bayes Rule



Bayes' Rule

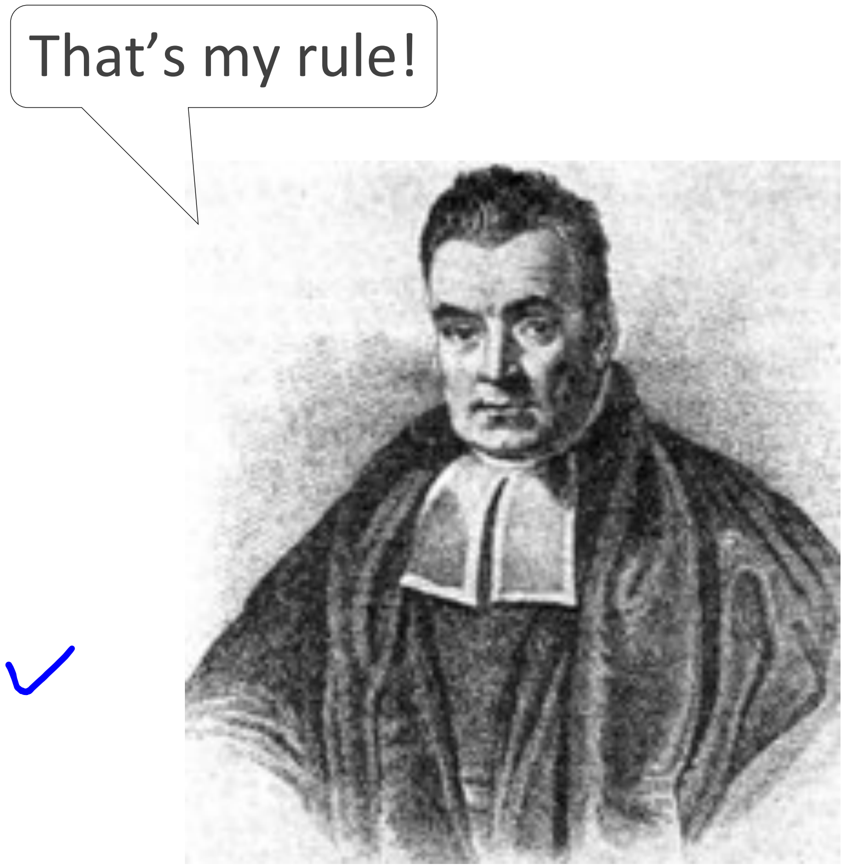
- Two ways to factor a joint distribution over two variables:

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

- Dividing, we get:

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

Bayesian



- Why is this at all helpful?
  - Let 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:

$$\underbrace{P(\text{cause}|\text{effect})}_{\text{posterior}} = \frac{P(\text{effect}|\text{cause})\overbrace{P(\text{cause})}^{\text{prior}}}{P(\text{effect})}$$

- Example:
  - M: meningitis, S: stiff neck

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

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999} \approx 0.008$$

  - 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

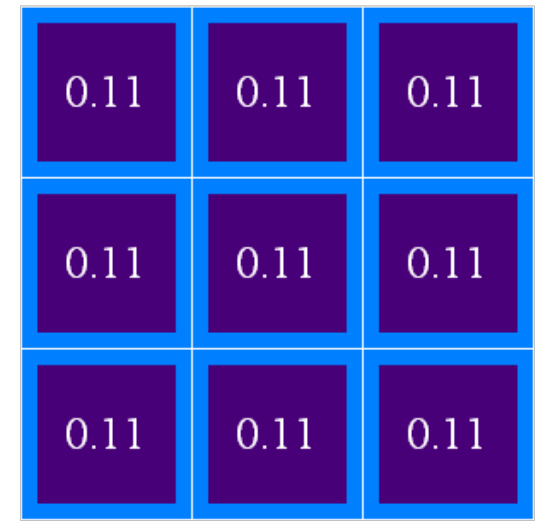
- What is P(sun | dry)? Give only the first two decimals

$$= \frac{P(sun, dry)}{P(dry)} = \frac{P(dry \mid sun) \times P(sun)}{P(dry|sun)P(sun) + P(dry \mid rain) P(rain)}$$

$$= 0.92$$

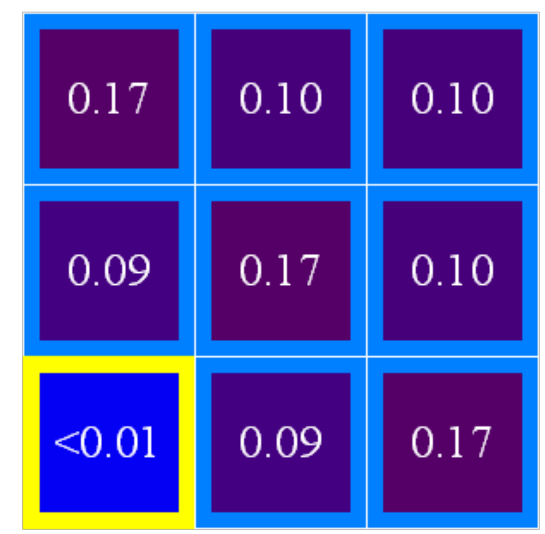
Ghostbusters, Revisited

- 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 = yellow | G=(1,1)) = 0.1



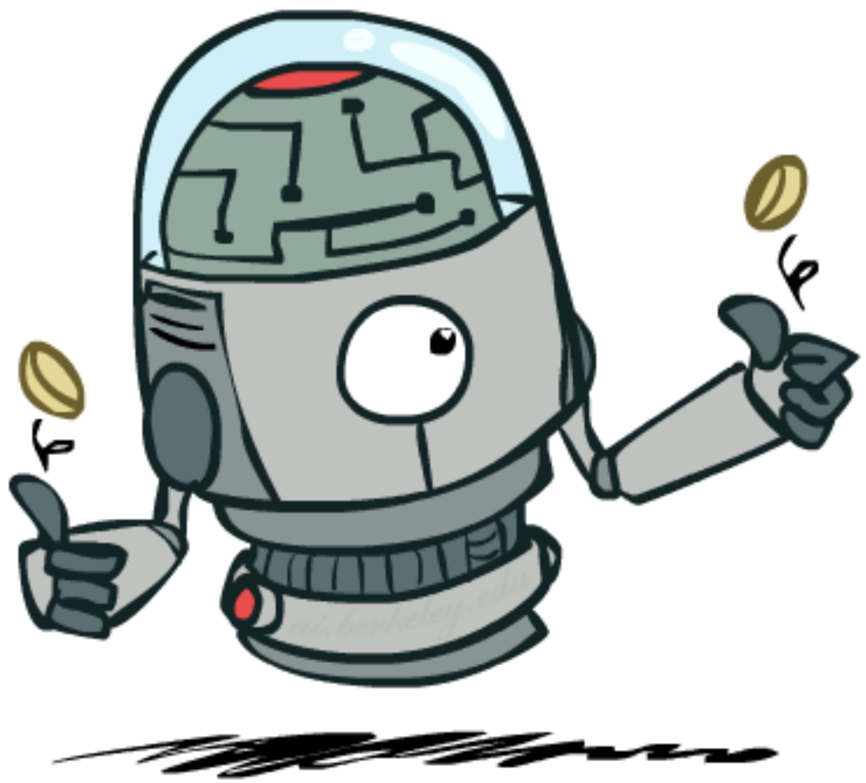
- We can calculate the posterior distribution P(G | r) over ghost locations given a reading using Bayes' rule:

$$P(g|r) \propto \underline{P(r|g)P(g)}$$



- What about two readings?
  What is $P(r_1, r_2|g)$?

Independence



Independence

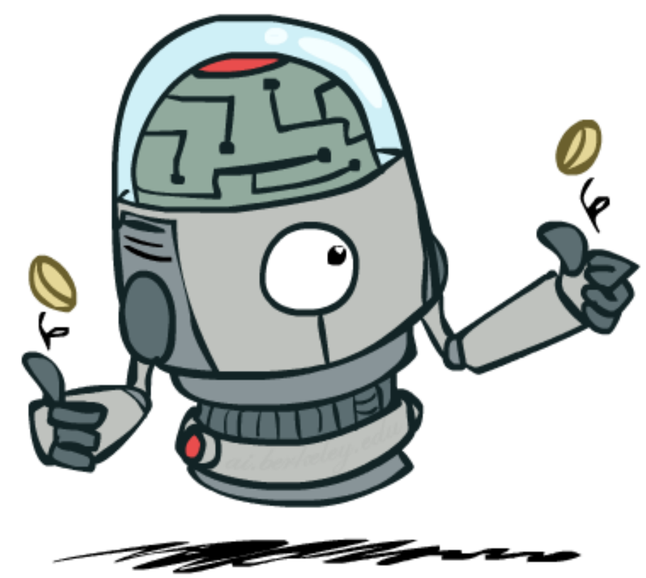
- Two variables are *independent* if:

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

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form: $\forall x, y : P(x|y) = P(x)$

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

  - We write: $X \perp\!\!\!\perp Y$



- Independence can be used as 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

$P(W)$

W	P
sun	0.6
rain	0.4

$P_2(T,W)$

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

Example: Independence

- N fair, independent coin flips:

iid

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

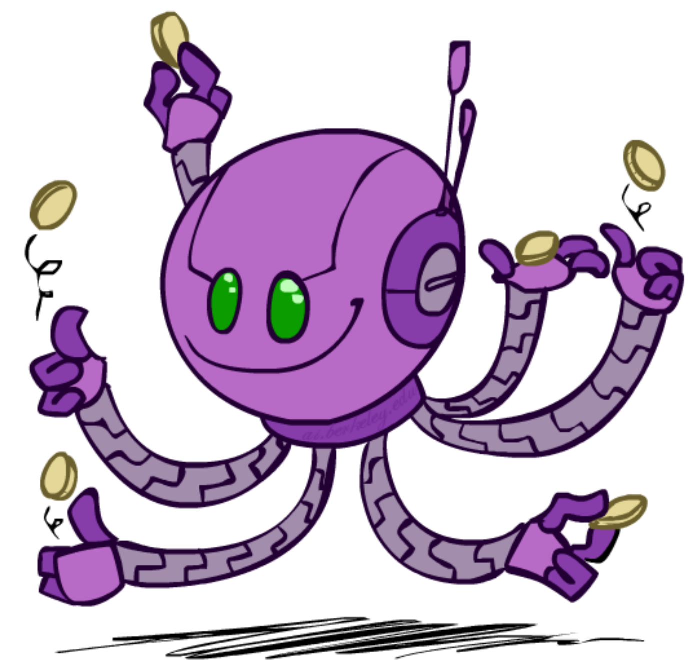
H	0.5
T	0.5

$\dots$

$P(X_n)$

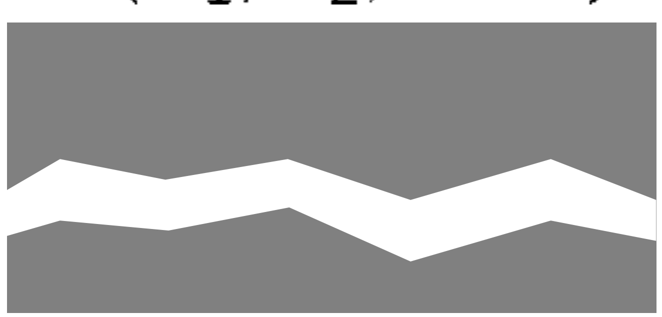
H	0.5
T	0.5

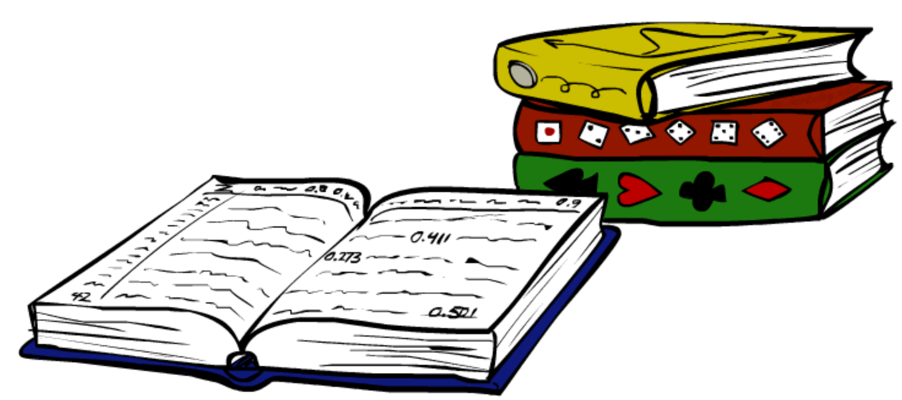
1



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

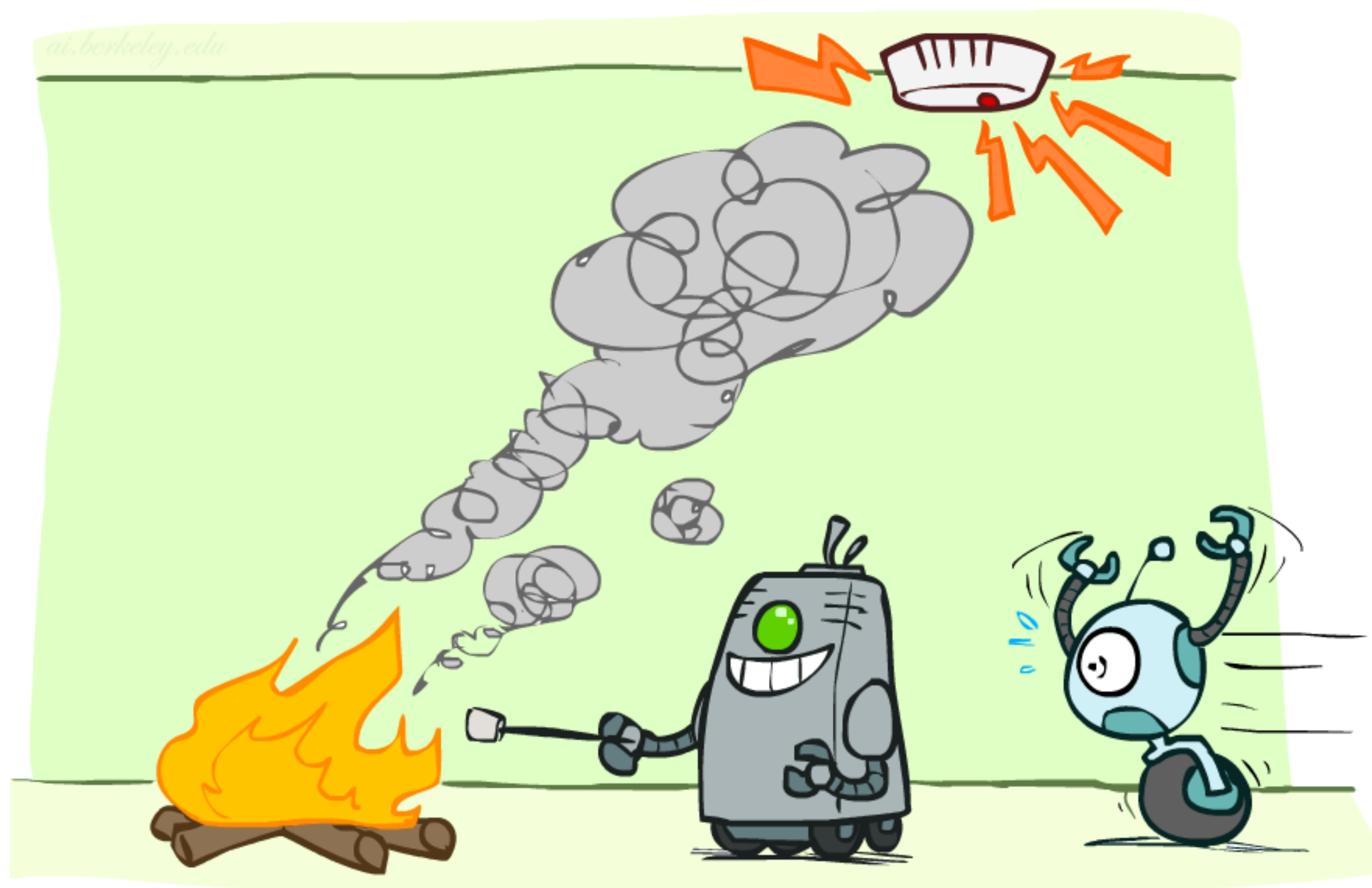
2^n





$2^n - 1$

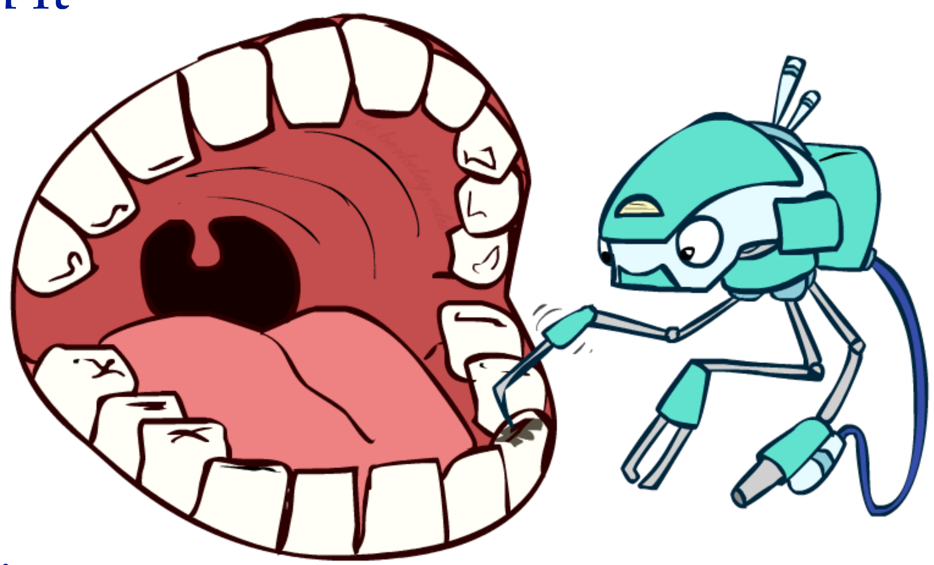
Conditional Independence



Conditional Independence

- P(Toothache, Cavity, Catch)

- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)



- The same independence holds if I don't have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch | -cavity)

- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)

- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily

Conditional Independence

- Unconditional (absolute) independence very rare
- *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:

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

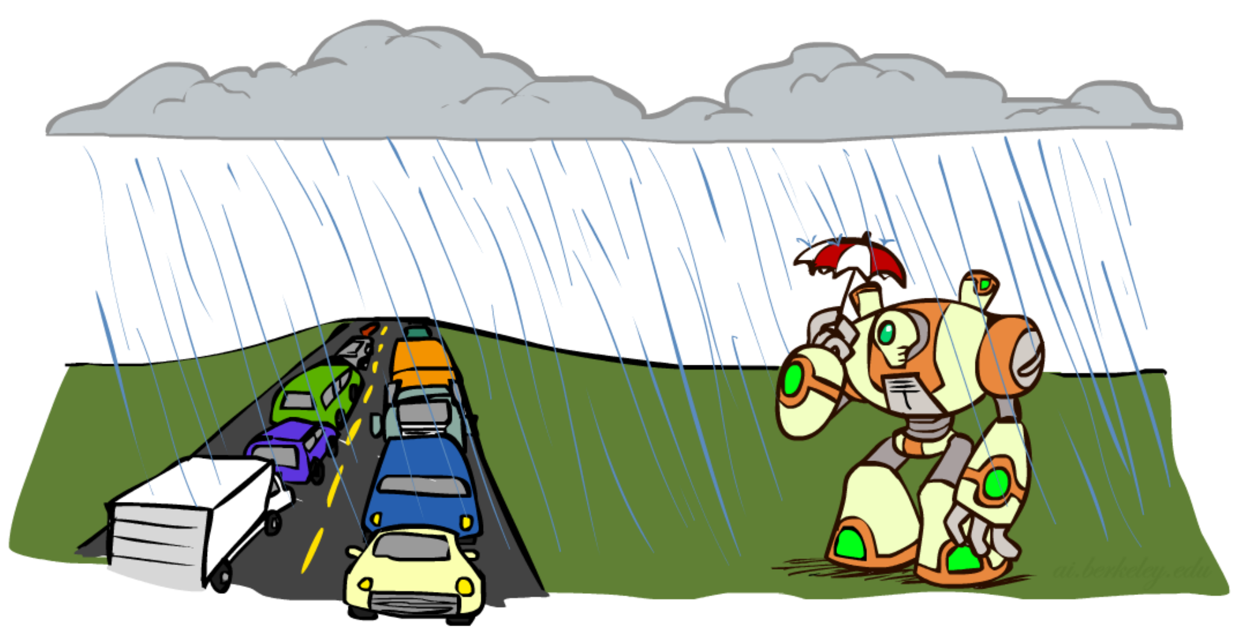
or, equivalently, if and only if

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

Conditional Independence

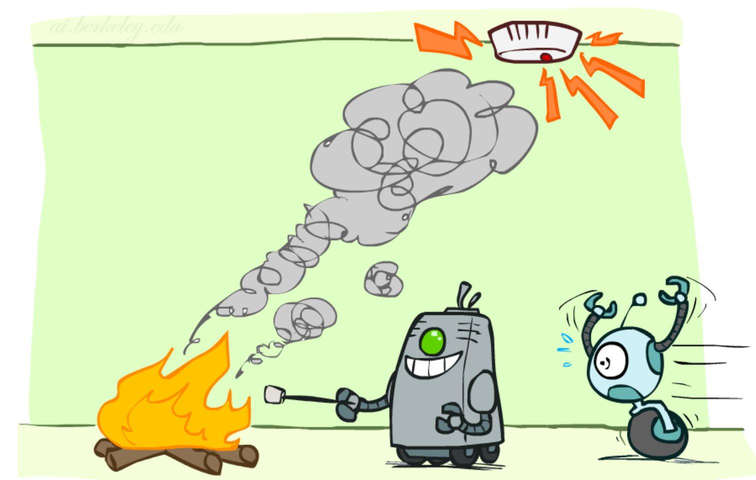
- What about this domain:
  - Traffic
  - Umbrella
  - Raining

Traffic $\perp\!\!\!\perp$ Umbrella | Raining

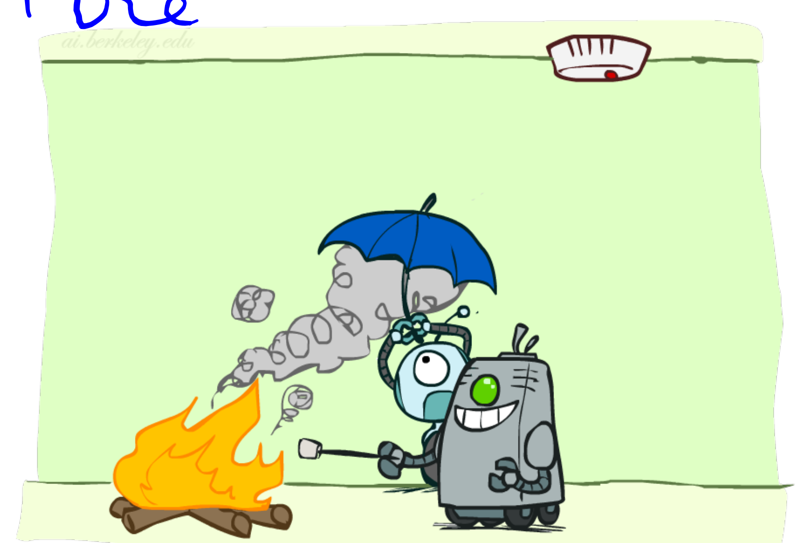


Conditional Independence

- What about this domain:
  - Fire
  - Smoke
  - Alarm



$\text{Alarm} \perp\!\!\!\perp \text{Smoke} \mid \text{Fire}$



Ghostbusters, Revisited

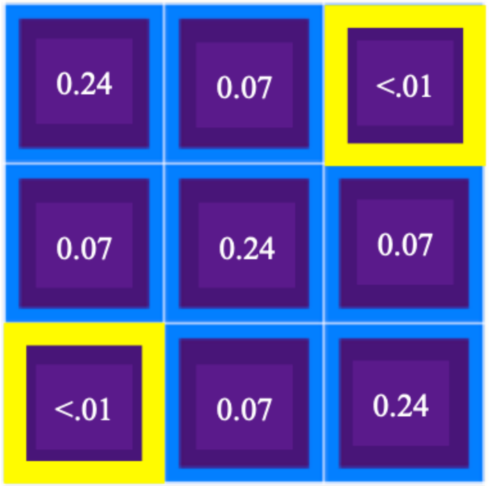
- What about two readings?
  What is $P(r_1, r_2 | g)$?
- Readings are conditionally independent given the ghost location!

$$P(r_1, r_2 | g) = P(r_1 | g) P(r_2 | g)$$

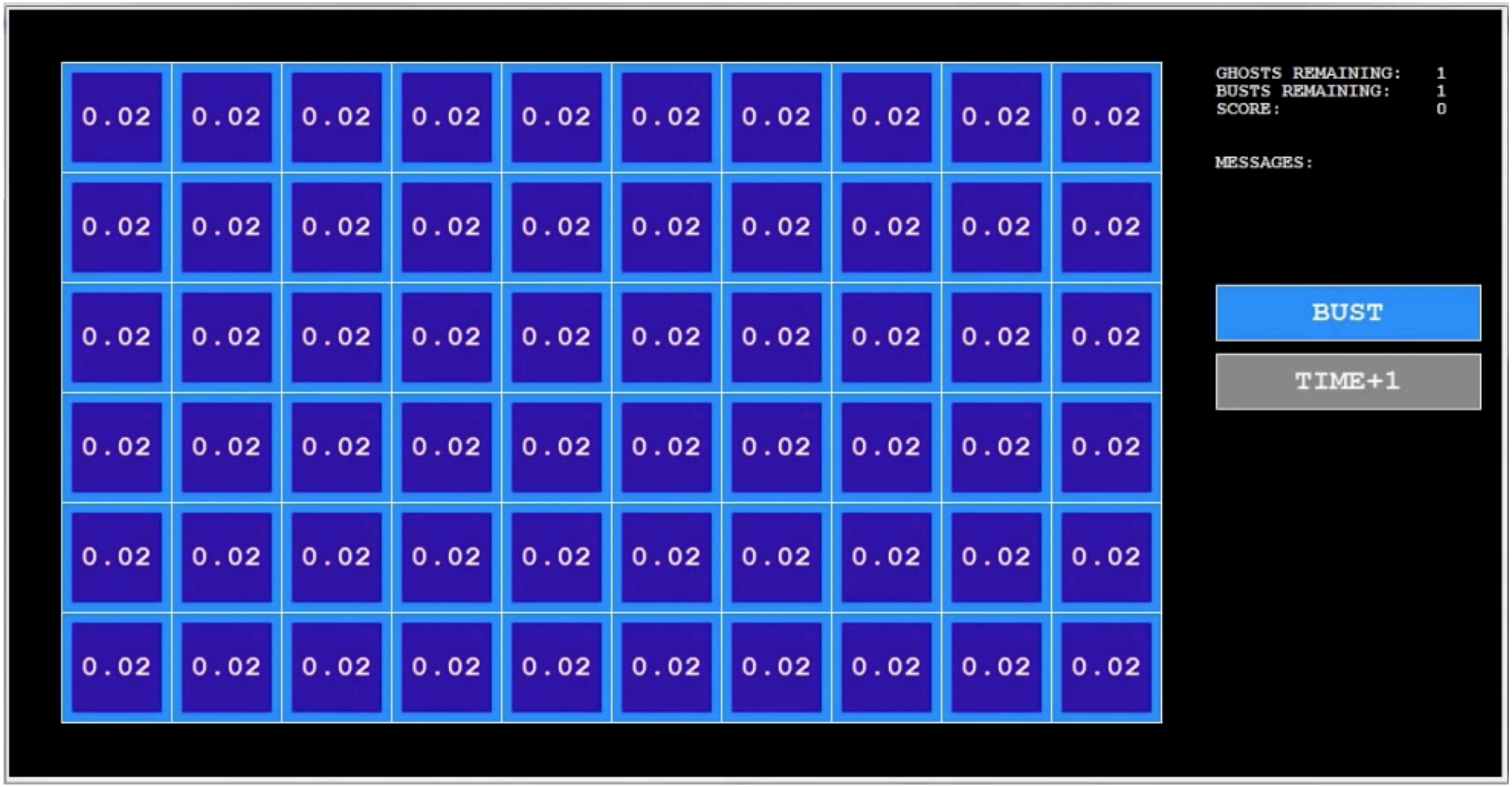
- Applying Bayes' rule in full:

$$P(g | r1, r2) \propto P(r1, r2 | g) P(g)$$
$$= P(g) P(r1 | g) P(r2 | g)$$

- Bayesian updating



Video of Demo Ghostbusters with Probability

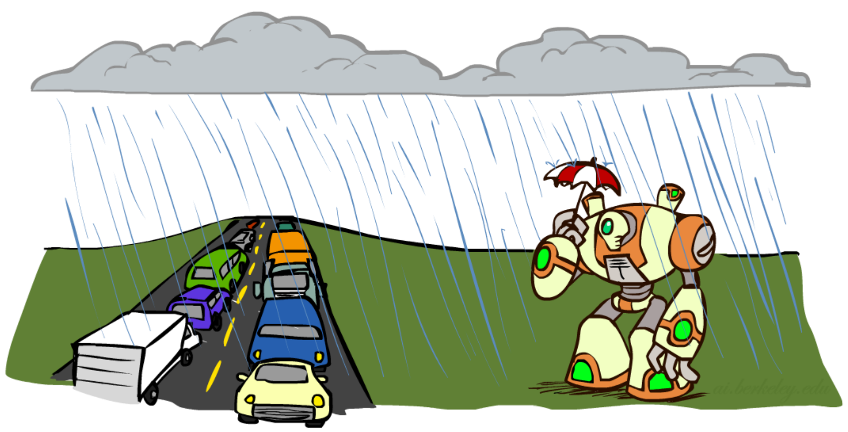


Conditional Independence and the Chain Rule

- Chain rule:

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

- Trivial decomposition:



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

- With assumption of conditional independence:

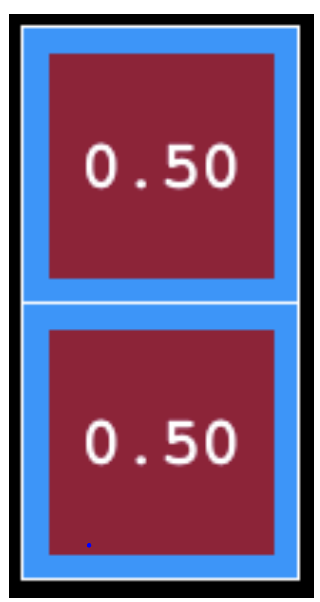
$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) = P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$

- Bayesian 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
- Given:
  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

$$P(T,B,G) = P(G)\,\underline{P(T|G)\,P(B|G)}$$



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