CS 4100 // artificial intelligence

instructor: byron wallace

Probability

Attribution: many of these slides are modified versions of those distributed with the *UC Berkeley CS188* materials
Thanks to John DeNero and Dan Klein

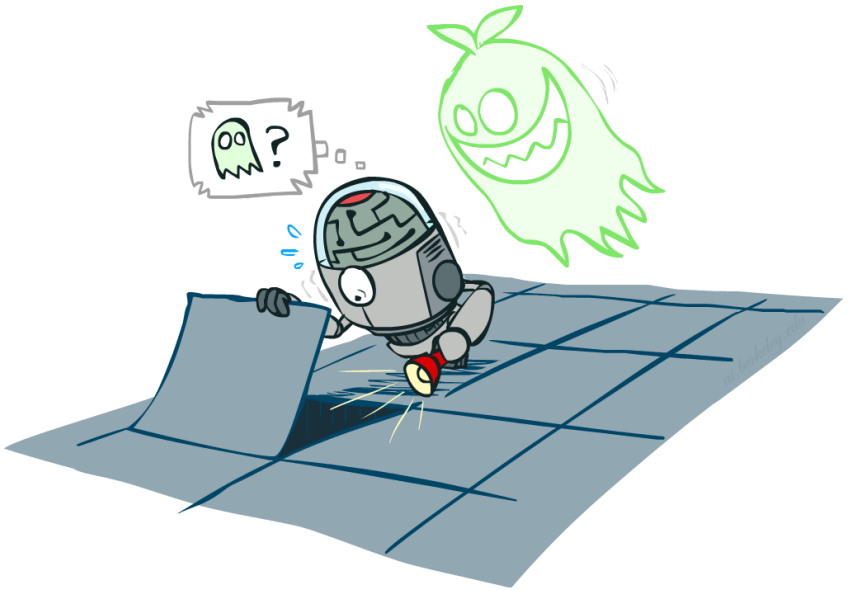
Where are we?

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

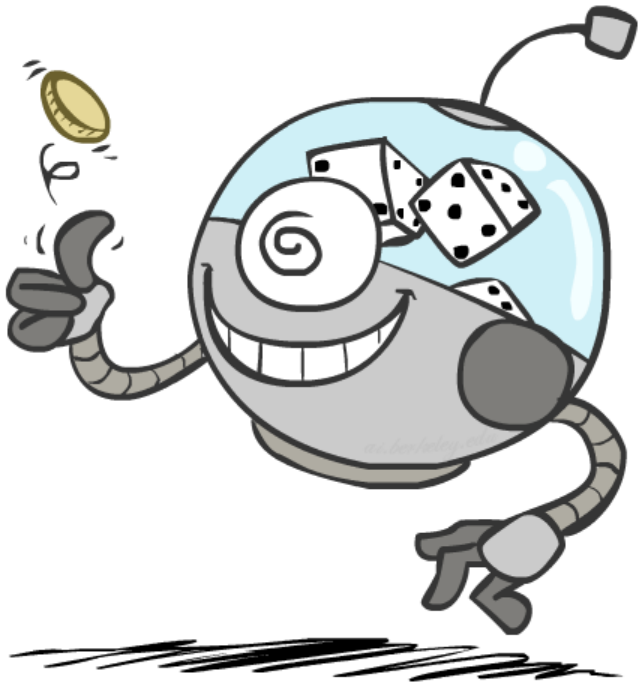


Today

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!



But wait, why do we need to do all this math?

Eugene Charniak, Brown University

Two eras of Natural Language Processing: Before Statistics (BS) and After Statistics (AS). The BS stuff doesn't work.

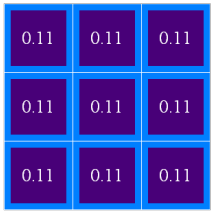
(I'm paraphrasing)

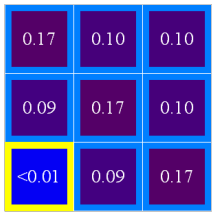
Uncertainty

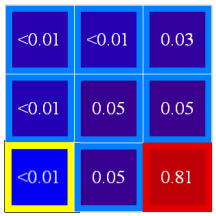
General situation:

- **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

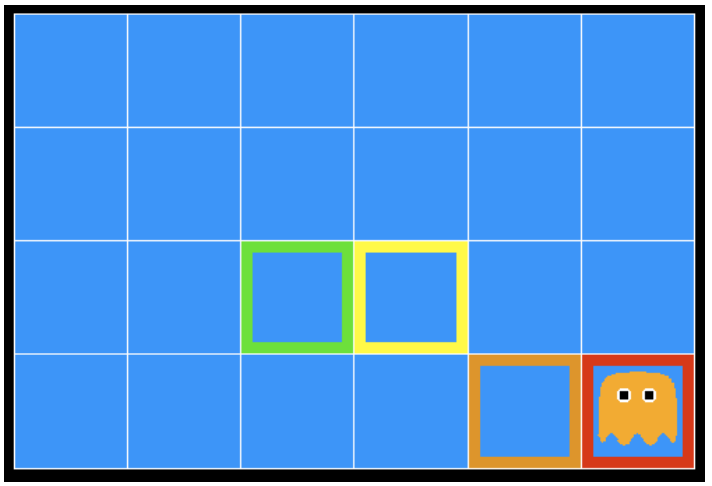






Example (&hw 4 preview!): “ghostbusters”

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



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

Random variables

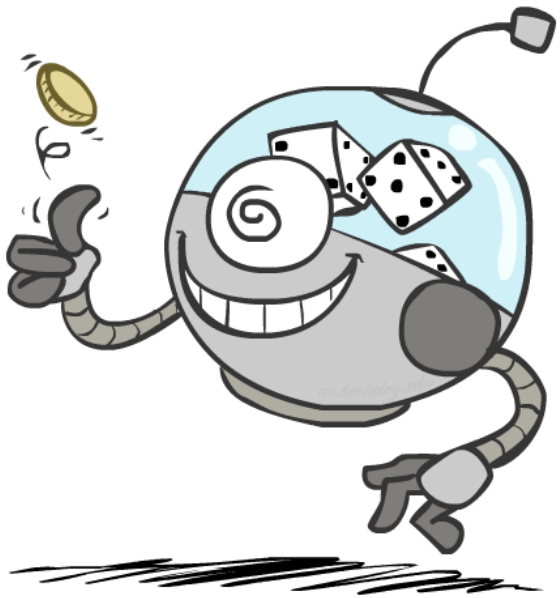
A random variable is some aspect of the world about which we (may) have uncertainty

- 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

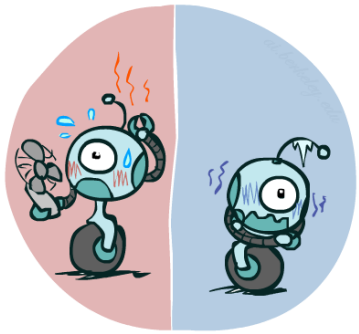
- R in {true, false} (often write as {+r, -r})
- T in {hot, cold}
- D in [0, ∞)
- L in possible locations, maybe {(0,0), (0,1), …}



Probability distributions

Associate a probability with each value

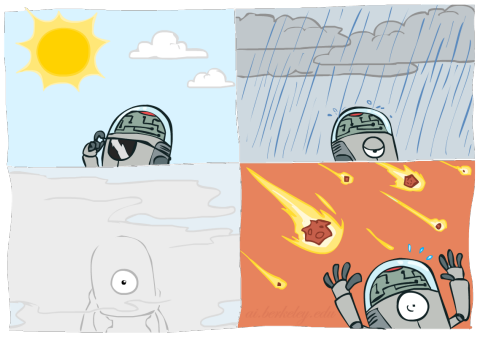
Temperature



$P(T)$

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

Unobserved 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:

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

A (discrete) distribution is a TABLE of probabilities of values

A probability (lower case value) is a single number

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

Note that: $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Joint distributions

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

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

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

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, this is impractical to write out!

$P(T, W)$

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

Probabilistic models v CSPs

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

Probabilistic models:

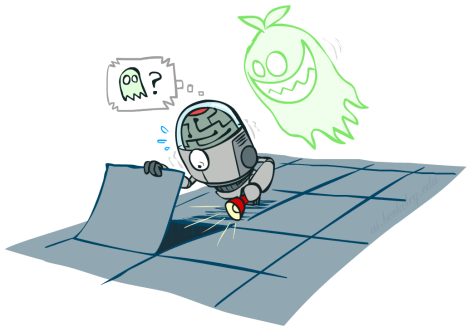
- (Random) variables with domains
- Assignments are called *outcomes*
- Joint distributions: say whether assignments (outcomes) are likely
- *Normalized:* sum to 1.0
- Ideally: only certain variables directly interact

Constraint satisfaction problems:

- Variables with domains
- Constraints: state whether assignments are possible
- Ideally: only certain variables directly interact

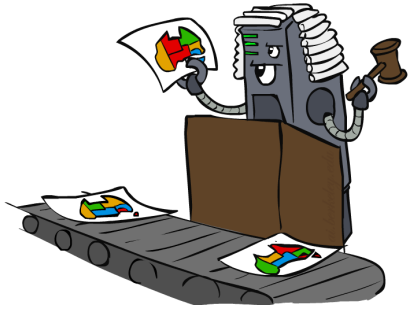
Distribution over T,W

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

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

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

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

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) ?
- P(+x) ?
- P(-y OR +x) ?

$$P(X, Y)$$

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

Quiz: events

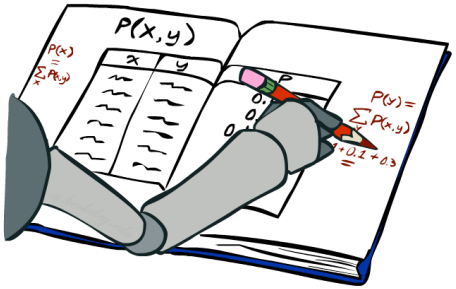
- P(+x, +y) ?
  .2
- P(+x) ?
  .5
- P(-y OR +x) ?
  .6

$$P(X, Y)$$

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

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

$P(T)$

T	P
hot	0.5
cold	0.5

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

$P(W)$

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

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

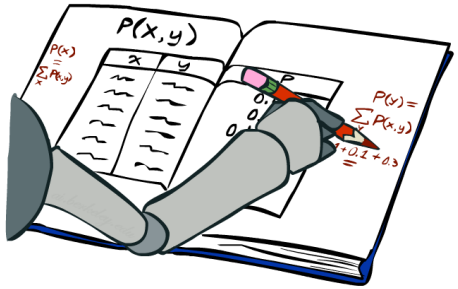
$$P(y) = \sum_x P(x,y)$$

$P(X)$

X	P
+x	
-x	

$P(Y)$

Y	P
+y	
-y	

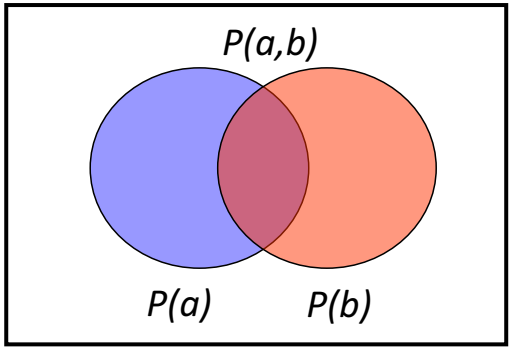


Conditional probabilities

A simple relation between joint and conditional probabilities

- In fact, this is taken as the *definition* of a conditional probability

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

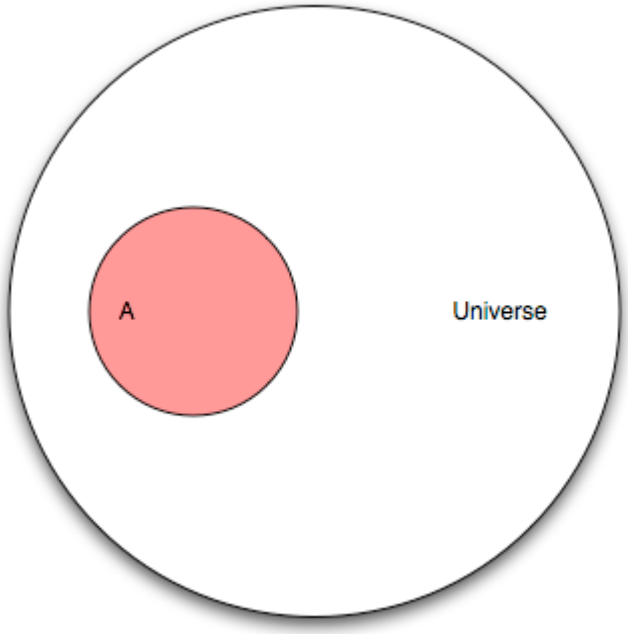


$P(T, W)$

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

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

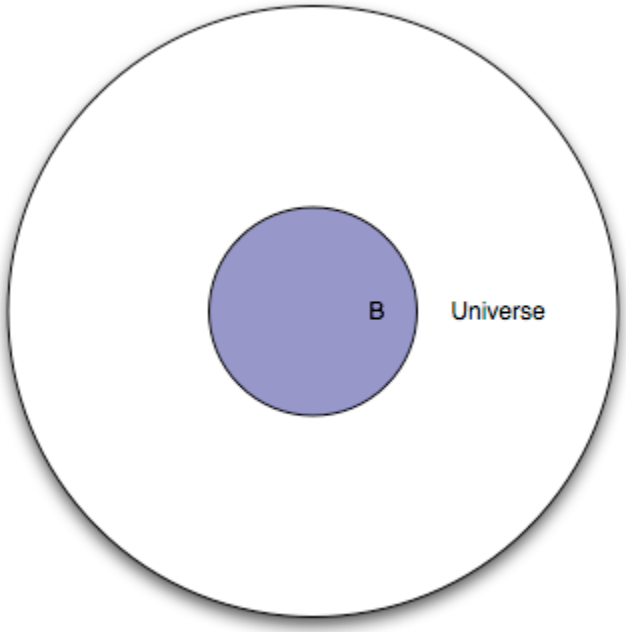
$$= P(W = s, T = c) + P(W = r, T = c)$$
$$= 0.2 + 0.3 = 0.5$$

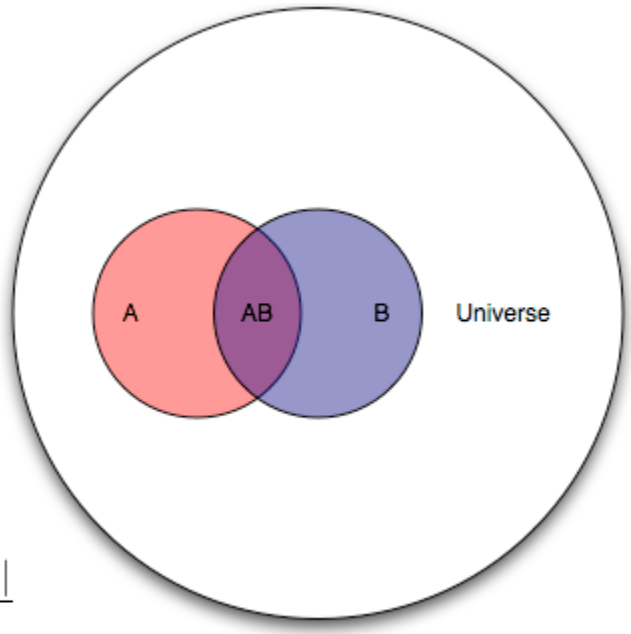


$$P(A) = \frac{|A|}{|U|}$$

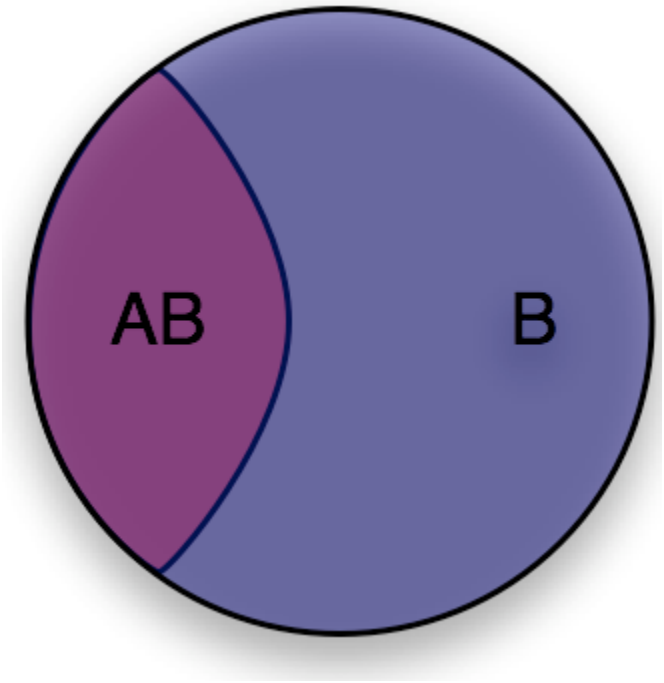
Credit: http://oscarbonilla.com/2009/05/visualizing-bayes-theorem/

$$P(B) = \frac{|B|}{|U|}$$





$$P(AB) = \frac{|AB|}{|U|}$$



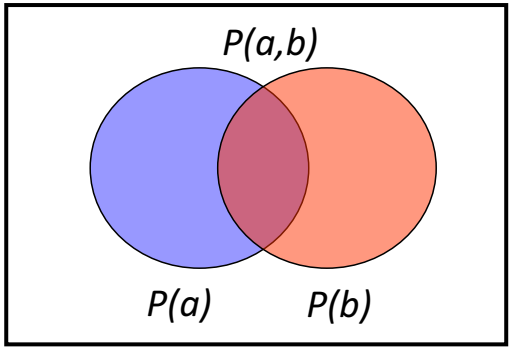
$$P(A|B) = \frac{P(AB)}{P(B)}$$

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

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

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

Quiz: conditional probabilities

$P(X, Y)$

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

- P(+x | +y) ?
- P(+y | +x) ?
- P(-y | +x) ?

Conditional *distributions*

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

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

Normalization trick

$P(T, W)$

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

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

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

$P(W|T = c)$

W	P
sun	0.4
rain	0.6

Normalization trick

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

$P(T, W)$

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

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

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 | 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 **1**

Procedure:

- Step 1: Compute Z = 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(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

** Works fine with multiple query variables, too*

General case:

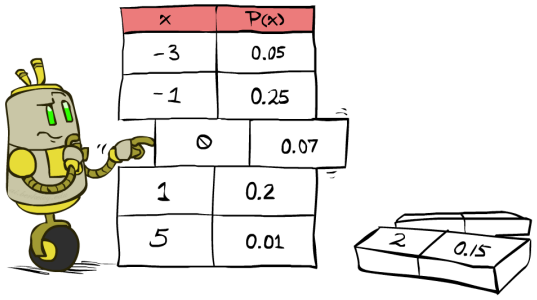
- Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
- Query* variable: Q
- Hidden variables: $H_1 \dots H_r$

$X_1, X_2, \dots X_n$ *All variables*

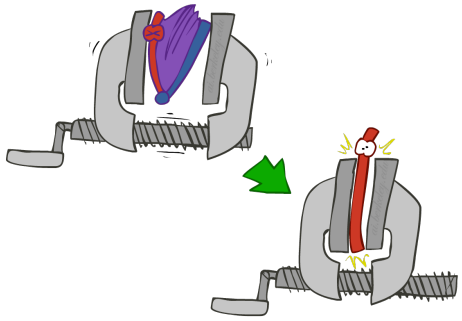
We want:

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

Step 1: Select the entries consistent with the evidence



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



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

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

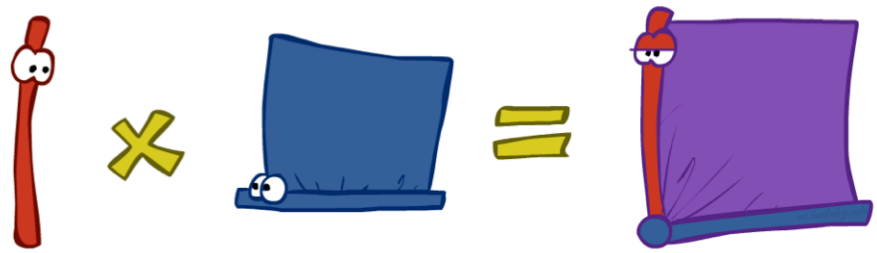
- P(W)?
- P(W | winter)?
- P(W | winter, hot)?

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

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

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

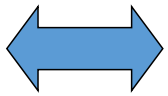
Example:

$P(W)$

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



$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

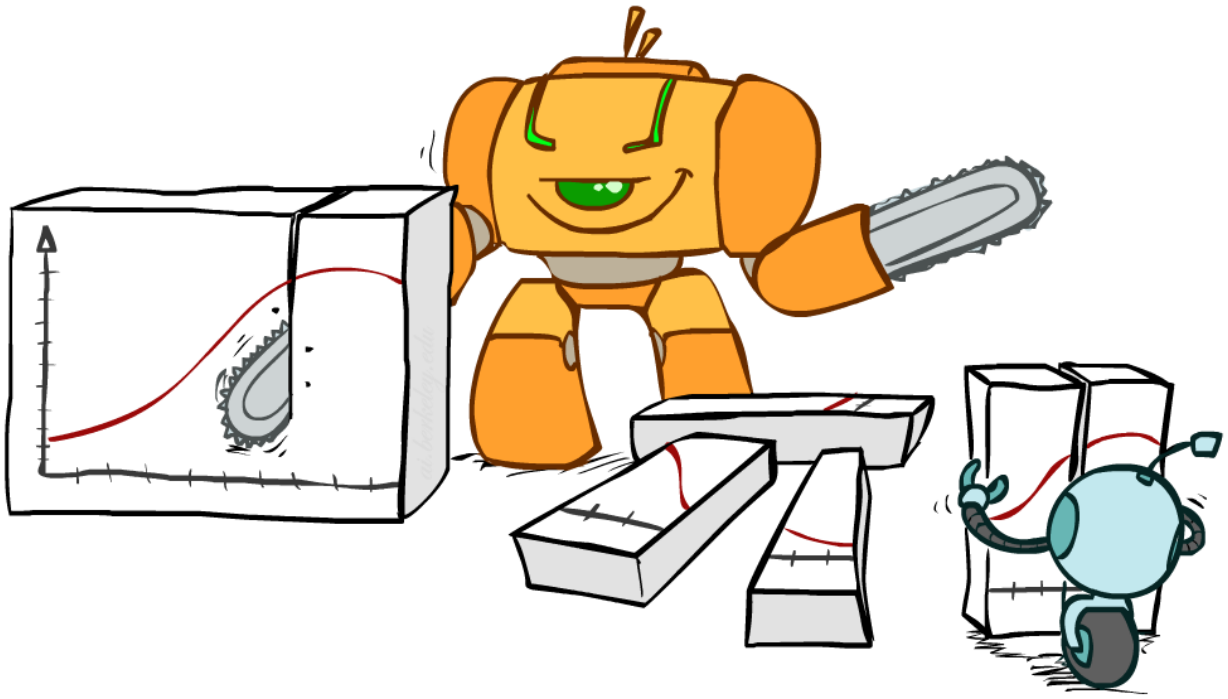
The chain rule

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

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

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

Bayes rule



Bayes' rule

Two ways to factor a joint distribution over two variables:

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

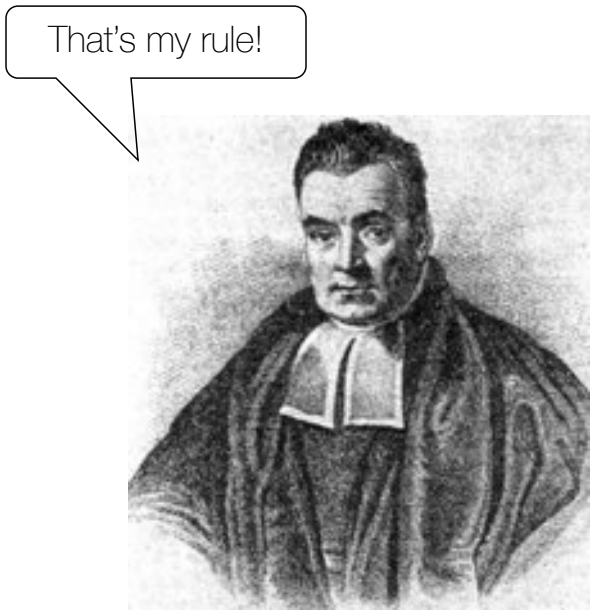
Dividing, we get:

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

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!




BAYES
RULES

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

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

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

- Note: posterior probability of meningitis still very small

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 | dry) ?

Drug testing example

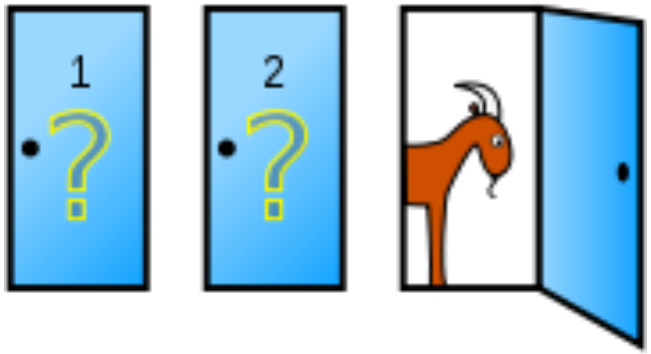
- Assume 0.4% of the Mass population uses marijuana
- Drug test: 99% true positive results for drug users; 99% true negative results for non-users
- If a randomly selected individual is tested positive, what is the probability he or she is actually a user?

Drug testing example

$$
\begin{aligned}
P(User|+) &= \frac{P(+|User)P(User)}{P(+)} \\
&= \frac{P(+|User)P(User)}{P(+|User)P(User)+P(+|!User)P(!User)} \\
&= \frac{0.99\times 0.004}{0.99\times 0.004+0.01\times 0.996} \\
&= 28.4\%
\end{aligned}
$$

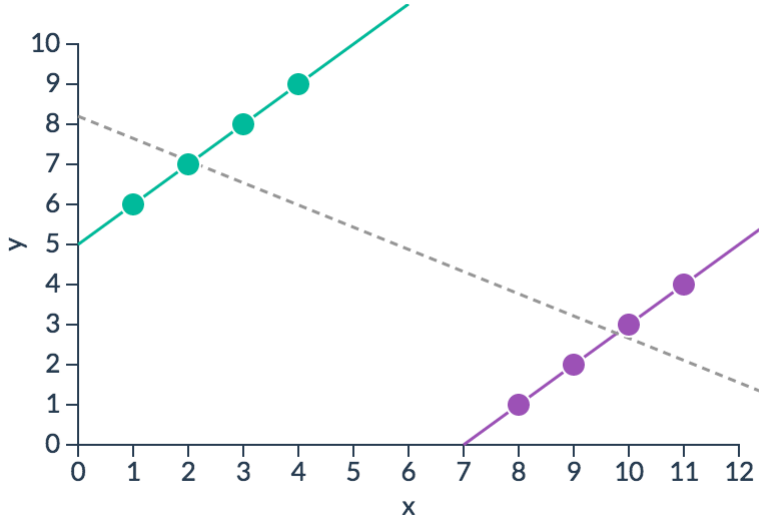
Brief aside; some “gotchas”

Monty hall

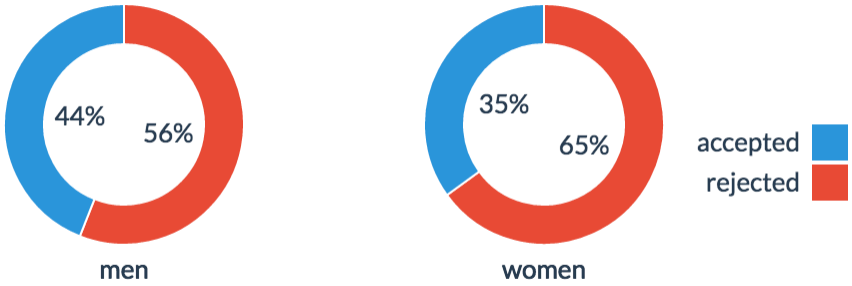


Simpson's paradox

- When a relationship is reversed at a higher level of data aggregation compared with the lower level



An example



In 1973, the University of California-Berkeley was sued for sex discrimination: they had accepted 44% of male applicants and only 35% of female applicants.

credit: http://vudlab.com/simpsons/

But…

It turns out women were applying to more competitive programs!

- See http://vudlab.com/simpsons/

Holy wars: Bayesians v Frequentists

- *Frequentists* believe underlying parameters (e.g., μ) are **fixed.** This is the world of *p*-values.

- *Bayesians* think of parameters as random variables – so μ is an RV following some distribution.

The Bayesian approach

- Taking the Bayesian approach, we incorporate *prior* knowledge, wherever it may come from, into inference.

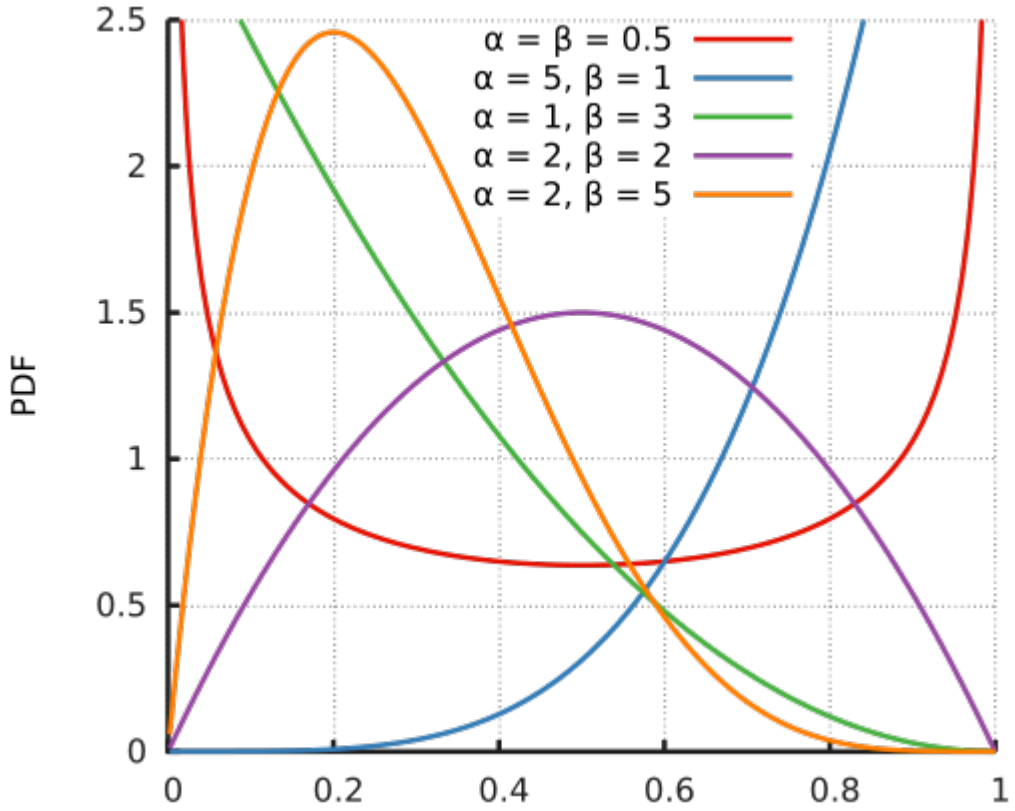
- Let's go back to coin flipping

Coin flipping re-visited

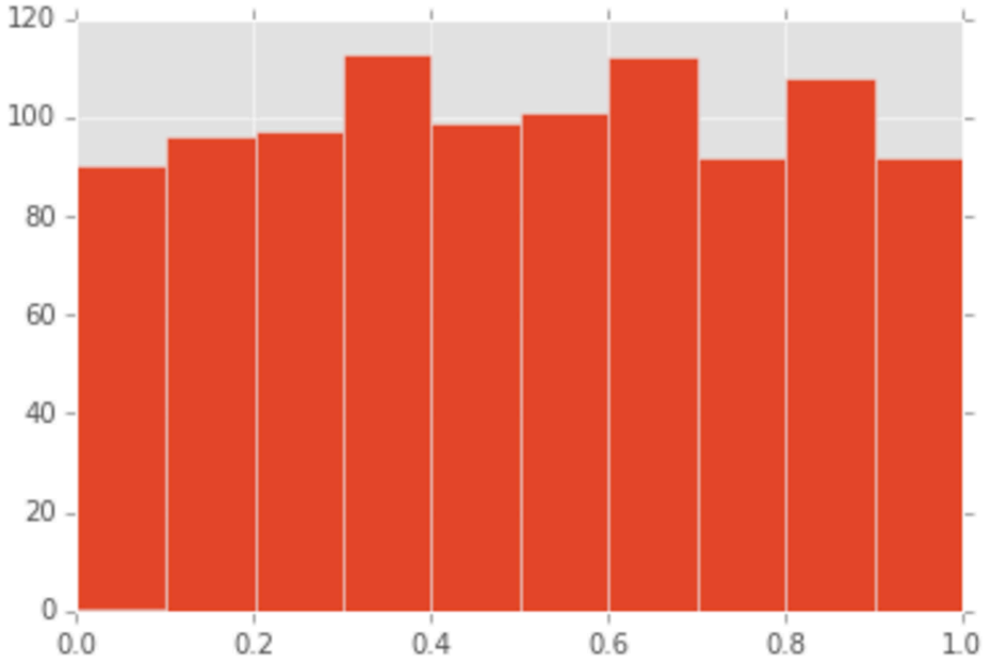
- Suppose we have two observed coin flips, both heads. The standard frequentist (ML) estimate would say p = 0.0. This seems extreme, no?

- In the Bayesian world, we'll instead factor in our *prior* knowledge, specifically through a *prior* distribution on p

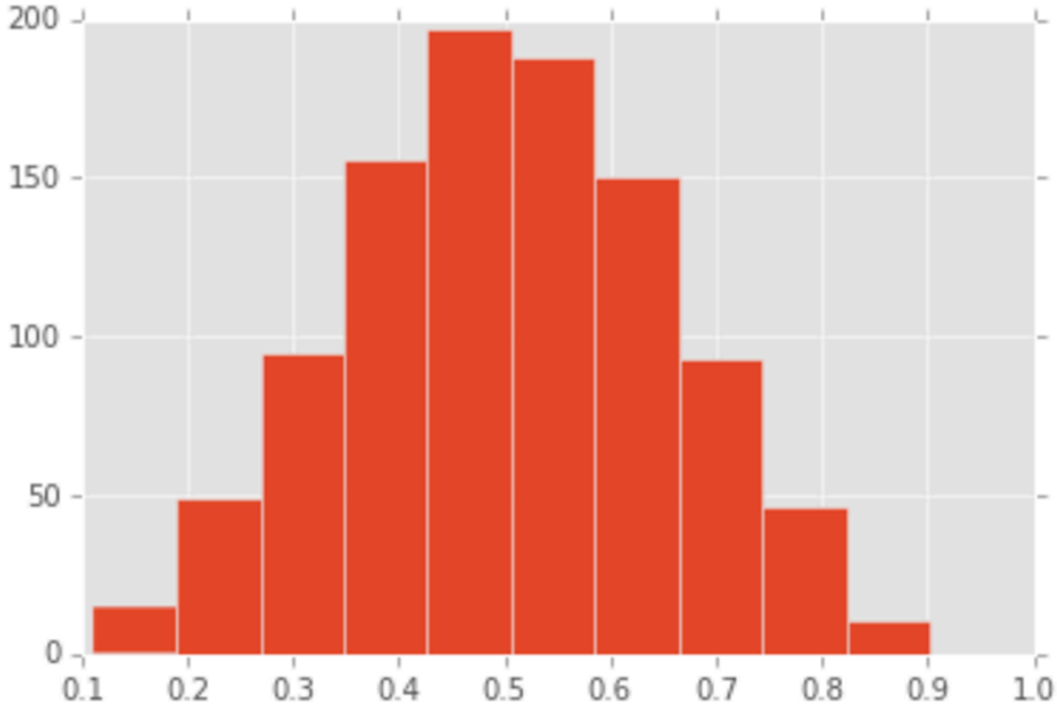
The Beta-Bernoulli



Uninformative prior – Beta(1,1)

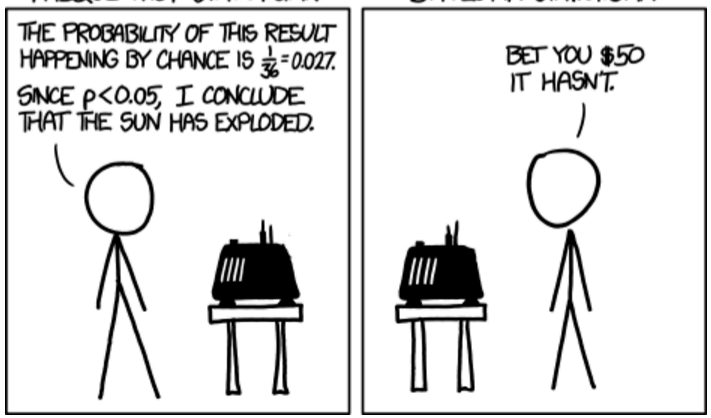


Informative prior -- Beta(5,5)



DID THE SUN JUST EXPLODE?

(IT'S NIGHT, SO WE'RE NOT SURE.)



That's it for today!

- Next time: Markov models!

- HWs due *next* ***Sunday***