

# 6.034 Artificial Intelligence: Lecture 10

## Probabilistic reasoning I: the crash course

Robert C. Berwick

September 25, 2020



### Menu for today

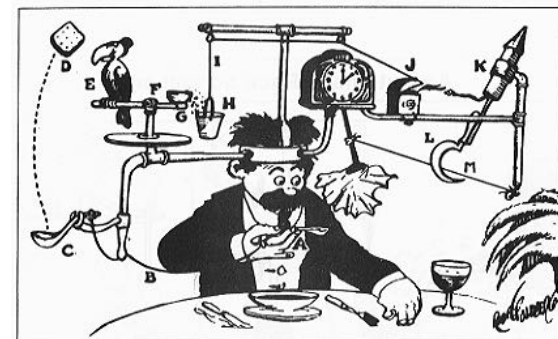
- ❑ Motivation: Why probabilistic intelligence?
- ❑ Probability basics
  - Probability Axioms
  - Conditional Probability
  - Chain Rule
  - Independence
  - Conditional Independence
- ❑ Result: probabilistic reasoning central to intelligent reasoning
  - Bayes rule: the simplest possible machine learning engine
  - Bayesian inference: selecting the best theory
  - Application challenge: kids learning words
- ❑ Discussion: Is Bayes always right?

## Motivation: Why Probability in (A)I ?

- Intelligent agents ought to handle uncertainty because the world is uncertain
- Probabilities – a (the) logically coherent algebra for uncertainty
- How to combine/weigh evidence from disparate sources, signals, and representations; focus on particular constraints
- Change in face of evidence/input → learning
- Goal: efficient models of cause and effect – “Belief nets”/Bayes nets– so agents can reason about...

## Rube Goldberg machine

Self-Operating Napkin



*Professor Butts and the Self-Operating Napkin* (1931). [Soup spoon](#) (A) is raised to mouth, pulling [string](#) (B) and thereby jerking [ladle](#) (C), which throws [cracker](#) (D) past [toucan](#) (E). Toucan jumps after cracker and [perch](#) (F) tilts, upsetting [seeds](#) (G) into [pail](#) (H). Extra weight in pail pulls [cord](#) (I), which opens and ignites [lighter](#) (J), setting off [skyrocket](#) (K), which causes [sickle](#) (L) to cut [string](#) (M), allowing [pendulum](#) with attached [napkin](#) to swing back and forth, thereby wiping chin.

From Wikipedia

### How children learn words/concepts – one classical view

“For if we will observe how children learn languages, we shall find that, to make them understand what the names of simple ideas or substances stand for, people ordinarily show them the thing whereof they would have them have the idea; and then repeat to them the name that stands for it; as white, sweet, milk, sugar, cat, dog.” (John Locke, 1690, Book 3, IX.9)

“shoe”

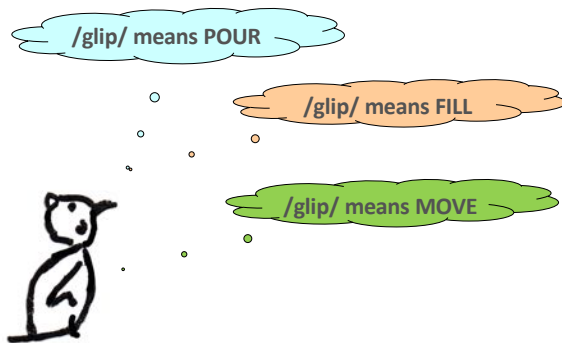
### Really? The real shoe world



Suppose kid sees this and hears some sound in context...



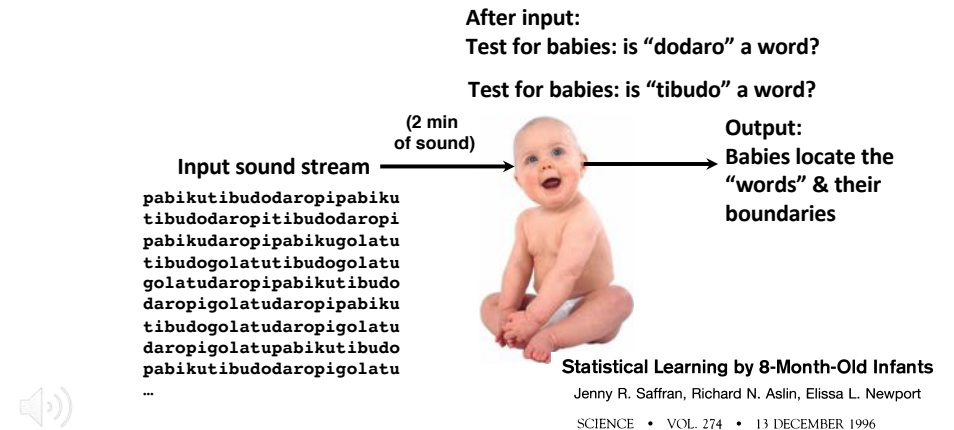
...glip...



What “concept” should the sound “glip” correspond to?  
(How to fill in “frames” as in Representation lecture?)

Slide courtesy of Sourabh Niyogi

The real world is a noisy world that requires probability for intelligence



The 8 month-old babies seem to succeed at this using probabilities  
If 8 month old babies can use probabilities, then so can you...

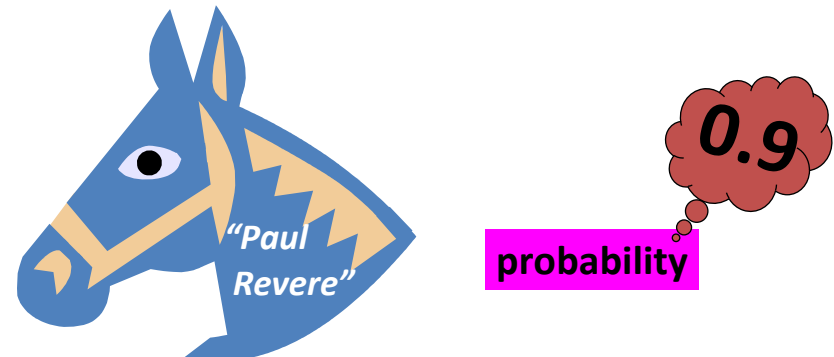
But people are lousy at (conscious?) reasoning with probabilities...

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more likely:

1. Linda is a bank teller
2. Linda is a bank teller and works with a local foodbank

**Probability (X & Y)  $\leq$  Probability(X)**

**Probability: the crash (or repeat) course**



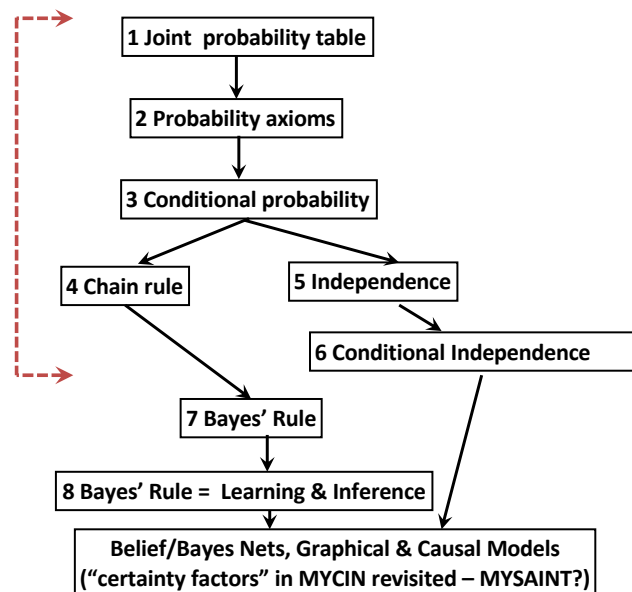
**P(Paul-Revere-wins) = 0.9**

**What does that mean?**

**How do we compute with this stuff?**

**What can we do with it?**

## Winston's probability roadmap



## Discrete Random Variables

- E is a (Boolean) random variable if it denotes an uncertain event
  - A statue appears at MIT
  - Global warming will cause Florida to be under water by 2100
- We can extend this to discrete variables with more than two possible values
- Random variables can also be continuous
  - Your first child will be 6'3" in adulthood




Discrete  
Random  
Variables

STATUE

HACK

DONOR

Joint probability table = a complete probabilistic description of the state of all events, here 3: S, D, H

 Sculpture calculations

11:50:18 EDT 11-Aug-2025

Sculpture	Donor	Hack	Tally	P	Selected
false	false	false	405	0.405	<input type="checkbox"/>
false	false	true	225	0.225	<input type="checkbox"/>
false	true	false	10	0.010	<input type="checkbox"/>
false	true	true	5	0.005	<input type="checkbox"/>
true	false	false	0	0.000	<input checked="" type="checkbox"/>
true	false	true	225	0.225	<input checked="" type="checkbox"/>
true	true	false	80	0.080	<input checked="" type="checkbox"/>
true	true	true	50	0.050	<input checked="" type="checkbox"/>
<input type="radio"/> T <input type="radio"/> F <input type="radio"/> ?	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> ?	<input type="radio"/> T <input type="radio"/> F <input type="radio"/> ?	1000	1.000	0.355

Marginal probability: likelihood of some incompletely specified state of events, e.g., Probability(Sculpture), summed over the Donor and Hack events

## Conditional probability: Prob Hack given Sculpture (one event contingent on another – a cause?)

# (one event contingent on another – a cause?)

## Sculpture calculations

12:54:23 EDT 11-Aug-2020

- Goal trees
- Search
- Constraint satisfaction
- Biological mimetics
- Learning
- Neural nets
- Bayes nets
- Sculpture calculations**
- Dog barking calculatio...
- Probability acquisition
- Coin modeling
- Parent's party
- Model comparison
- Structure discovery
- Miscellaneous

Sculpture	Donor	Hack	Tally	P	Selected
false	false	false	0	0.000	<input type="checkbox"/>
false	false	true	0	0.000	<input type="checkbox"/>
false	true	false	0	0.000	<input type="checkbox"/>
false	true	true	0	0.000	<input type="checkbox"/>
true	false	false	0	0.000	<input type="checkbox"/>
true	false	true	225	0.634	<input checked="" type="checkbox"/>
true	true	false	80	0.225	<input type="checkbox"/>
true	true	true	50	0.141	<input checked="" type="checkbox"/>
<input checked="" type="radio"/> T	<input type="radio"/> F	<input type="radio"/> ?	<input type="radio"/> T	<input type="radio"/> F	<input type="radio"/> ?
			355	1.000	0.775

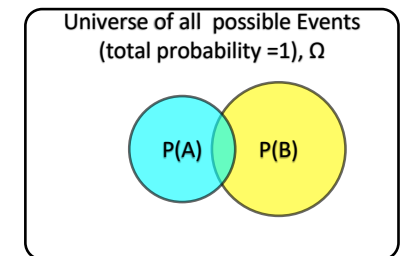
441

## Lightning review of probability

### (1) Probability Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{TRUE})=1$ ;  $P(\text{FALSE})=0$
- $P(A)+P(B)-P(A,B)=P(A \cup B)$

Note: if  $P(A,B) \equiv P(A \cap B) = \emptyset$ , then  
 $P(A)+P(B) = P(A \cup B)$



Venn  
Diagram

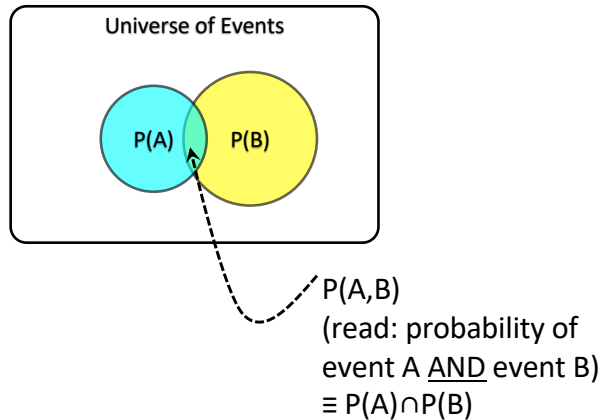
### • Useful Theorems

$$P(\neg A) = 1 - P(A)$$

$$P(A) = P(A, B) + P(A, \neg B)$$



## Joint probability as conjunction of events



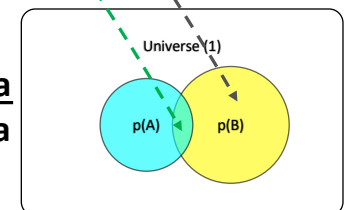
E.g., Paul Revere wins the race & the weather is clear

## (2) Conditional probability: a way to focus on what “matters”

- $P(A|B)$  = fraction of possible worlds in which B is true in which A is also true
- $P(A|B) = P(A,B)/P(B)$
- so,  $P(A, B) = P(B)P(A|B)$
- by symmetry,  
 $P(A,B) = P(A)P(B|A)$

Ratio of Green Area  
Yellow Area

E.g., (Paul Revere wins | weather clear)



**But typically, more than one conditioning variable that is a factor ...**

**P(Paul Revere wins | weather's clear, ~~ground is dry, jockey's brother my friend, Epitaph also in race, Epitaph was recently bought by Gonzalez, race is on May 17, ...~~ )**

BACKING OFF:

We strike out some of the conditioning factors

Perhaps not exactly what we want but at least we can get a reasonable estimate of it!

(i.e., more bias but less variance)

try to *keep* the conditions that we suspect will have the most influence on whether Paul Revere wins

**What about on the left-hand side of the conditioning bar?**

**P(Paul Revere wins, ~~Valentine places, Epitaph shows~~ | weather's clear)**

NOT ALLOWED!

but we can do something similar to help ...

### (3) Chain rule for conditional probabilities

- We already showed from  $P(A|B)$ , that  $P(A,B)=P(A|B) P(A)$
- This is the base case of an induction, for 1 event;  $n=1$ ; for the inductive step...
- What about  $P(A,B,C)$ ? Reduce B,C to one event this way:
- Let  $Y=(B,C)$  then:  
 $P(A,B,C)=P(A,Y) = P(A|Y)P(Y)$  (as before); now substitute B,C for Y  
 $= P(A|B,C)P(B,C)$  ;now rewrite  $P(B,C)$  as  $P(B|C)P(C)$   
 $= P(A|B,C)P(B|C)P(C)$  ... rinse & repeat for more events...  
 $P(A|B,C,D)=P(A|B,C,D)P(B|C,D)P(C|D)P(D)$
- **Chain rule: no event depends on any event to its left**
- $P(x_1,x_2,...,x_n)= \prod_{i=1}^n P(x_i | x_{i-1},...,x_1)$
- Moves variables to the right of the conditioning bar

### Example from our horse race

$$\begin{aligned} &P(\text{Revere wins, Valentine places, Epitaph shows, weather's clear}) \\ &= P(\text{Revere wins} | \text{Valentine, Epitaph, weather's clear}) \\ &\quad \times P(\text{Valentine places} | \text{Epitaph, weather's clear}) \\ &\quad \times P(\text{Epitaph shows} | \text{weather's clear}) \\ &\quad \times P(\text{weather's clear}) \end{aligned}$$

#### (4) (Probabilistic) Independence:

Holds when an event B does not affect the probability of another event A, and vice-versa

$$P(A \cap B) = P(A, B) \text{ iff } P(A) \times P(B)$$

Rewriting, if A, B are independent then:

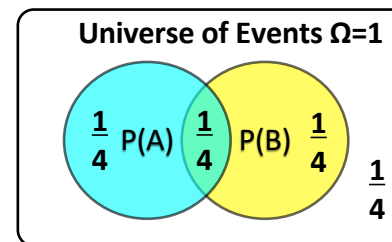
$$P(A) = \frac{P(A, B)}{P(B)} = P(A | B)$$

And, by symmetry:

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

Knowing about event B probability does not tell us more about event A probability, and vice-versa

#### Venn diagram of independence



Blue = Green  
White = Yellow

$$P(A) = \frac{1}{2} \quad P(B) = \frac{1}{2} \quad P(A | B) = P(A, B) / P(B) = \frac{1/4}{1/2} = 2/4 = 1/2 = P(A)$$
$$P(B | A) = P(A, B) / P(A) = \frac{1/4}{1/2} = 2/4 = 1/2 = P(B)$$

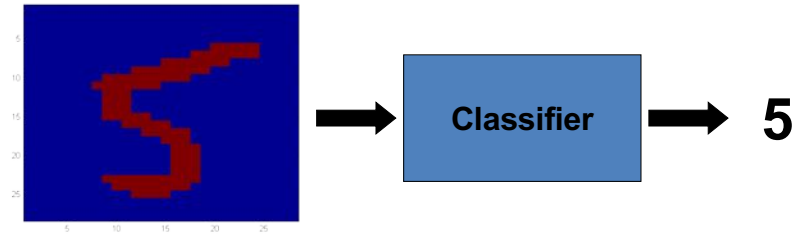
Note: If 2 events independent,  $\log P(A \cap B) = \log P(A) + \log P(B)$   
If negative log probability is interpreted as information content in bits,  
two events are independent if and only if the information content of the combined event  
equals the sum of information content of the individual events:  
 $\text{Information}(A, B) = \text{Information}(A) + \text{Information}(B)$

## (5) Conditional independence

- Definition of conditional independence:  
 $P(A,B|C) = P(A|C)P(B|C)$  ; it's as if event A can ignore B  
Compare: if A,B are independent, then  $P(A,B)=P(A)P(B)$   
A kind of modularity – can suppress probabilistically irrelevant info
- Useful for chain rule simplification  
If all the variables  $X_i$  below are all conditionally independent given Y, then:

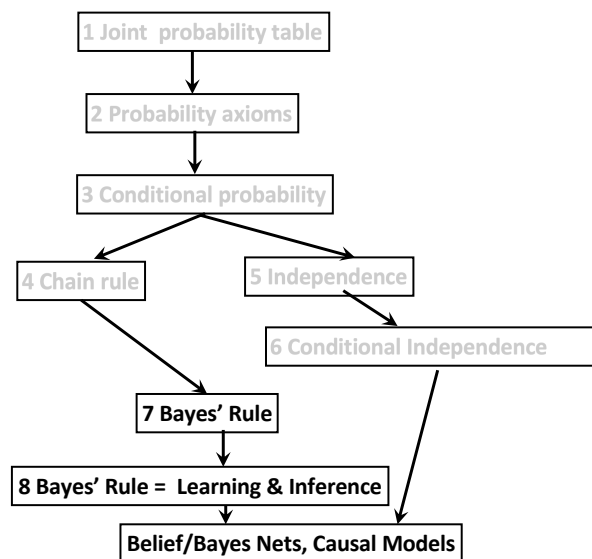
$$P(X_1, X_2, \dots, X_n | Y) = P(X_1 | Y) P(X_2 | Y) \dots P(X_n | Y)$$

- We will use this factoring next lecture for a machine learning task where the X's are features, and Y is some label we want to predict from the values of the  $X_i$ 's



Q: how many features  $X_i$  could there be? (in, say, a 50 x 50 pixel array)

## Winston's probability roadmap



## (6) Bayes' rule: A machine learning method

$$P(A|B) = \frac{P(A, B)}{P(B)} \implies P(A, B) = P(A|B) \cdot P(B)$$

||

$$P(B|A) = \frac{P(A, B)}{P(A)} \implies P(A, B) = P(B|A) \cdot P(A)$$

$\therefore P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$ ; now divide by  $P(B)$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad \textbf{Bayes' Rule}$$

Let's rewrite this in terms of updating some hypothesis  
given (new) evidence or facts...

## Bayes' rule dissected: 4 key components

Posterior Probability

$$P(A|B) = \frac{\text{Likelihood } P(B|A) \cdot \text{Prior Probability (= initial state bias) } P(A)}{\text{Normalization } P(B)}$$

Let's rewrite this in terms of updating some hypothesis given (new) evidence or facts...a learning model

## In terms of Evidence and Hypotheses, a learning model via updating initial probability

**Posterior Probability of Hypothesis<sub>1</sub> (after we observe evidence)**

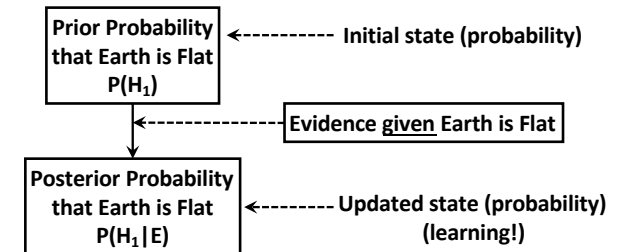
x

**Likelihood Probability of generating the evidence given Hypothesis<sub>1</sub>**

x

**Prior Probability of a particular Hypothesis<sub>1</sub> (before we observe any evidence)**

$$P(\text{Hypothesis}_1 | \text{Evidence}) = \frac{P(\text{Evidence} | \text{Hypothesis}_1) \cdot P(\text{Hypothesis}_1)}{P(\text{Evidence})}$$



### "Bayesian Inference"

NB: Crucially, note:  $P(\text{Evidence}) = \sum \text{all Evidence Probabilities}$

## Comparing alternative hypotheses – factor out denominator for likelihood comparison

Rank hypotheses to select the “most likely” one...

$$P(\text{Earth-is-flat}|\text{Evidence}) = \frac{P(\text{Evidence}|\text{Earth-is-flat}) \cdot P(\text{Earth-is-flat})}{P(\text{Evidence})}$$

$$P(\text{Earth-is-round}|\text{Evidence}) = \frac{P(\text{Evidence}|\text{Earth-is-round}) \cdot P(\text{Earth-is-round})}{P(\text{Evidence})}$$



**Bayes Factor: Ratio of Likelihoods of 2 models**

$$\frac{P(\text{Earth-is-round}|\text{Evidence})}{P(\text{Earth-is-flat}|\text{Evidence})} = \frac{P(\text{Evidence}|\text{Earth-is-round}) \cdot P(\text{Earth-is-round})}{P(\text{Evidence}|\text{Earth-is-flat}) \cdot P(\text{Earth-is-flat})}$$

**Can rewrite in gambler’s “odds notation”**

Odds = ratio of two probabilities  $P(A_1)/P(A_2)$ , written  $A_1 : A_2$

## Picking the “best” theories (“Bayesian model selection”)

Posterior Odds	Likelihood Ratio or Bayes factor	Prior Odds
$\frac{P(\text{Earth-is-round} \text{Evidence})}{P(\text{Earth-is-flat} \text{Evidence})}$	$\frac{P(\text{Evidence} \text{Earth-is-round}) \cdot P(\text{Earth-is-round})}{P(\text{Evidence} \text{Earth-is-flat}) \cdot P(\text{Earth-is-flat})}$	

- Odds = ratio of two probabilities  $P(A_1/A_2)$  (written  $A_1:A_2$ )
- In gambling, “3-to-1” odds means 75% chance of success; 1-to-1 odds is 50% chance (equal odds, 1:1)
- The posterior odds = the Bayes Factor times the prior odds  
In other words, the posterior is proportional to the likelihood (the ratio of each theory generating the observed evidence) times the prior probabilities of the theories
- Q: where do the priors come from?
- Q: Is this always the right way to do inference?



Let's apply our machine learning method!  
Learning words, but this time using Bayesian  
inference



... glip ...



/glip/ means POUR

/glip/ means FILL

/glip/ means MOVE

3 Hypotheses: H1, H2, H3

What "concept" should "glip" correspond to?  
(Remember frames and trajectories from Representation lec)

Can imagine different sorts of evidence that could be  
combined to infer what the sound /glip/ corresponds to



Q: How to combine these different kinds of evidence in the right way?

A: Use our Bayesian inference engine...  
a universal data & theory grinder



Slide courtesy of Sourabh Niyogi

## Let Bayes decide the “world-to-word” mapping



### Evidence:

- (i) particular sounds corresponding to concept using glip  
“Mary is glipping water into the glass” (e.g., preceding slide)
- (ii) different scenes corresponding to the concept associated with /glip/ (e.g., preceding slide)

### Priors

$P(H_1)=1/3$  (pour)

$P(H_2)=1/3$  (fill)

$P(H_3)=1/3$  (move)

### Likelihoods

$P(\text{evidence} | H_1)$

$P(\text{evidence} | H_2)$

$P(\text{evidence} | H_3)$

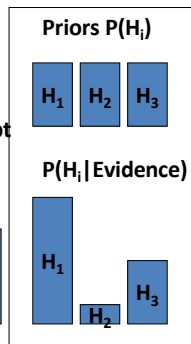
Bayesian Meat Grinder  
Posterior  $P(H_i) = \text{Likelihood} \times \text{Prior } P(H_i)$



Pick largest Posterior P

/glip/ means POUR

Learns with just a few examples – just like kids do?



Slide courtesy of Sourabh Niyogi

## DISCUSSION

If we act according to Bayes’ rule, will our inferences always be “rational”?

- Well, what does rational mean? If you’re a Bayesian...
- Consider this example...and think about priors & learning
- Rosencrantz is flipping a coin. Guildenstern is watching and is calling out "heads" or “tails”
- It is a **fair coin**—half the time it comes up heads, and half the time it comes up tails

## The certainty of the Bayesian fortune-teller

- Suppose Guildenstern is not a human being but rather is a Bayesian AI, and Guildenstern is *certain* that the coin is biased: it thinks that there is a 50% chance it is dealing with a coin that lands heads 3/4 of the time
- Conversely, G thinks there is 50% chance it is dealing with a coin that lands tails 3/4 of the time (those are its two priors)
- G's initial prediction will be exactly right: that the next flip is equally likely to be heads or tails (this depends on this initial 50-50 split in the prior hypotheses G has)

(Heads average:  $1/2 \times 3/4 + 1/2 \times 1/4 = 1/8 + 3/8 = 1/2$ )

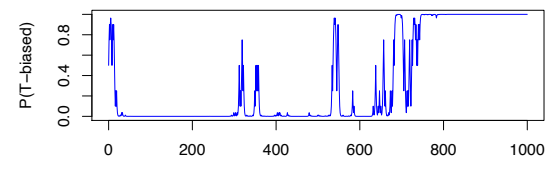
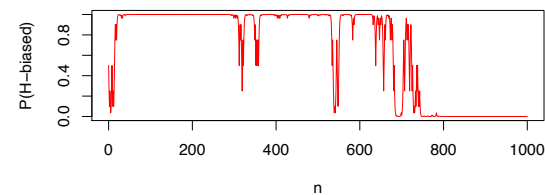
## But what would you do?

- Suppose you flipped the coin a billion times, and you see the long-run average of heads to tails is roughly 50-50...what would you now think? (Sure, the heads-tails ratio could be zig-zagging slightly above/below)
- What would be your prediction re the probability of the next coin flip?
- This is not what a Bayesian Guildenstern does!

## What does the Bayesian Guildenstern do?

- If the number of heads and tails is exactly even, then Guildenstern (correctly) forecasts that the odds on the next flip are 50-50
- But Guildenstern has seen 2 more heads than tails—*no matter how big  $n$  is*—then Guildenstern is 90% certain that it is dealing with a Head-biased coin and thinks that the chance the next flip will be heads is 70%
- If G sees 10 more heads than tails—again, *no matter how many flips  $n$  there have been*—Guildenstern is 99.9983% sure that it is dealing with an Head-biased coin and will forecast the odds of a head on the next flip at 74.9999%.

**As # of coins flips gets bigger, G never converges on the right answer that coin is unbiased**



**What has gone wrong? How to make it right? Or not?**