Bayesian inference

Computational Cognitive Science 2014
Lecture 2
Dan Navarro

Why start with Bayesian inference?

- The course needs it!
  - A lot of the psychological models rely directly or indirectly on Bayesian statistical methods
  - We'll start with the basics in these lectures, and move to the more complex stuff later

- It's just plain useful
  - A huge chunk of modern statistics relies on Bayes
  - Probabilistic AI is highly Bayesian
  - Etc.

Introduction to probability

Defining probability

- Intuitively: the "chances" of something happening
- Easy to convert to a formal system
  - e.g., Kolmogorov axioms, Cox axioms

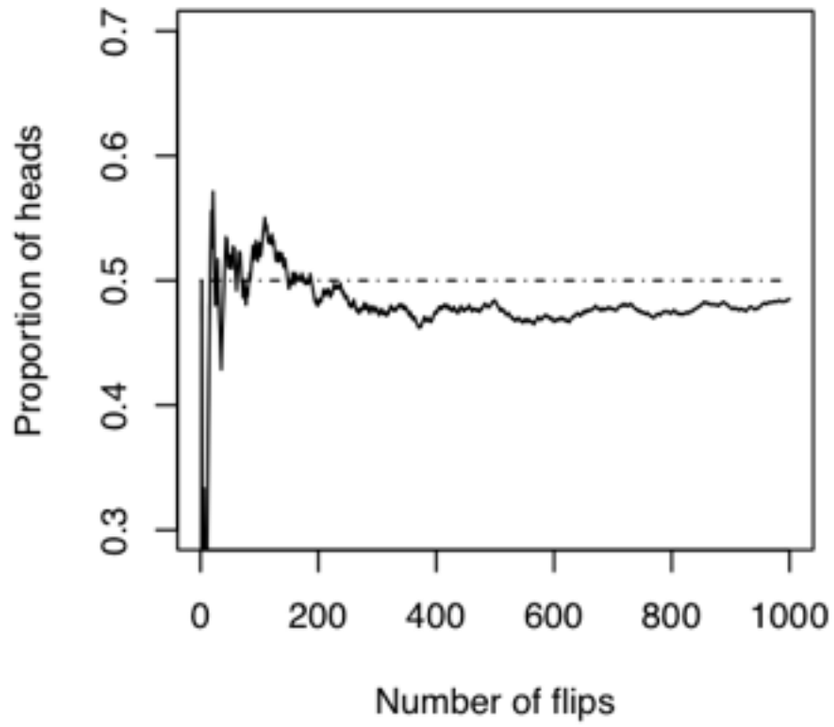
- P(X) is a number between 0 and 1
  - P(X) = 0 means X definitely will not happen
  - P(X) = 1 means X definitely will happen
  - + some other rules that we'll discuss

Understanding probability

- Not simple to interpret psychologically
- Two main schools of thought
  - Probability = long run frequency
  - Probability = degree of belief

Long run frequency

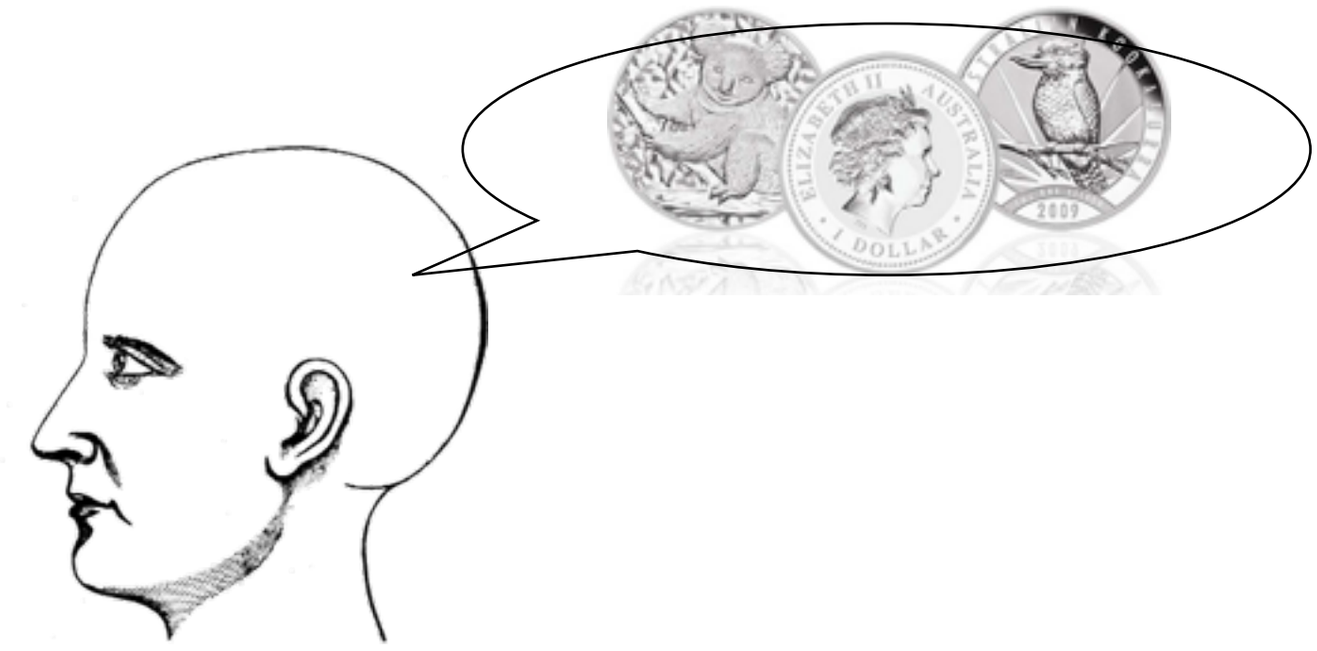
- The "frequentist" view...



- ... probability can only be assigned to events that can be replicated (e.g. coin flipping) not to one-off events (e.g., the mass of the Higgs boson)

Bayesian probability

- Degree of belief held by an intelligent agent...



- ... probability does not exist as a property of the objective world, it only expresses our beliefs about what will happen in the world

Big argument in statistics!

- Frequentist probability
  - Orthodox null hypothesis testing
  - Probability cannot be assigned to scientific theories

- Bayesian probability
  - Bayesian methods
  - Probability can be assigned to any hypothesis

- We'll sidestep the controversy
  - Focus on Bayesian methods because they're useful!

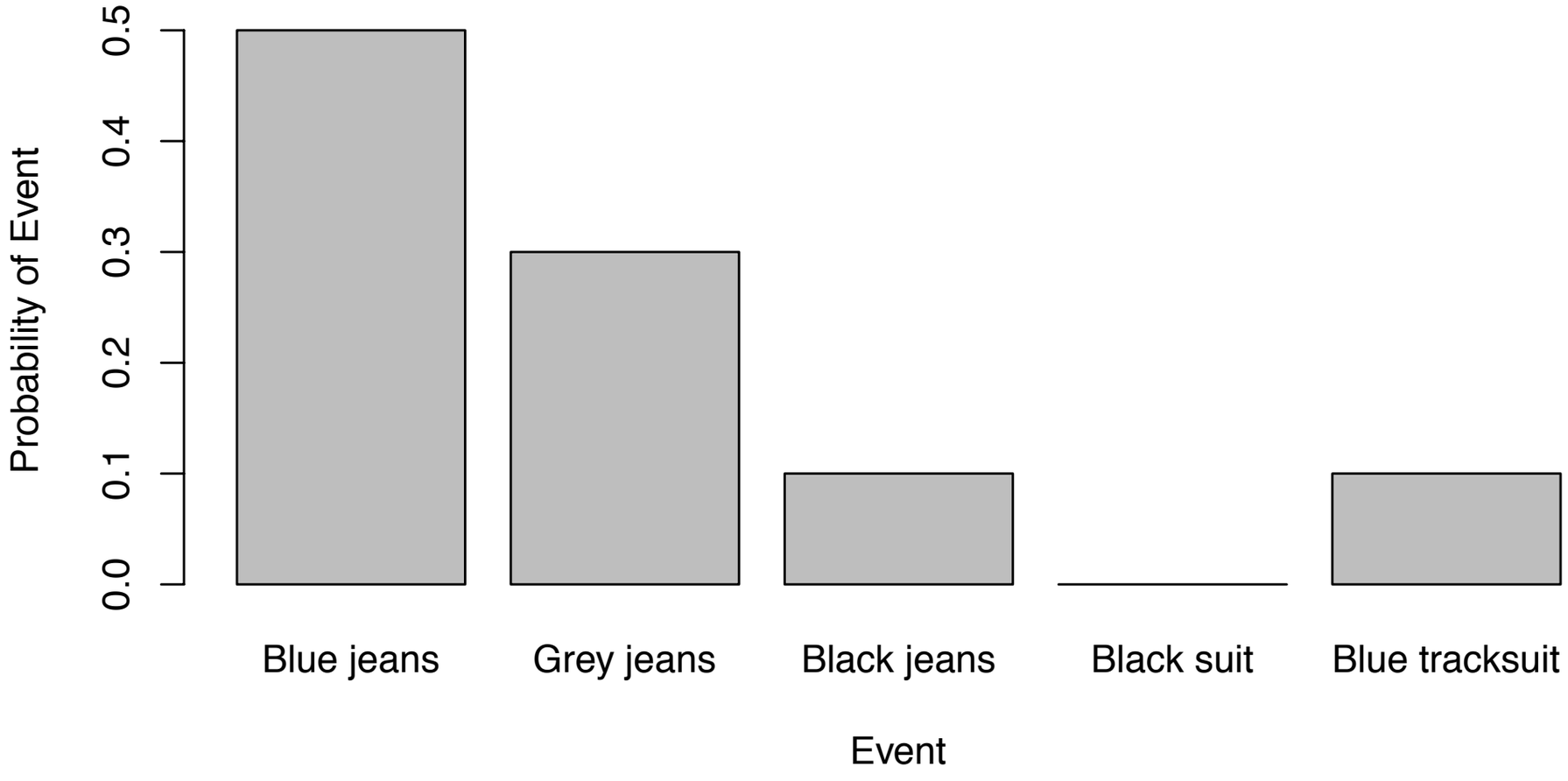
Other rules for probability

- Law of total probability:
  - Probabilities must sum to one, assuming the "sum" is across mutually exclusive and exhaustive possibilities (i.e. exactly one of them must occur)

$$\sum_{X \in \Omega} P(X) = 1$$

An example

- Suppose I own five sets of pants
  - I always wear pants (probably for the best)
  - I can't wear more than one pair of pants
  - So the "pants probabilities" sum to 1



Other rules

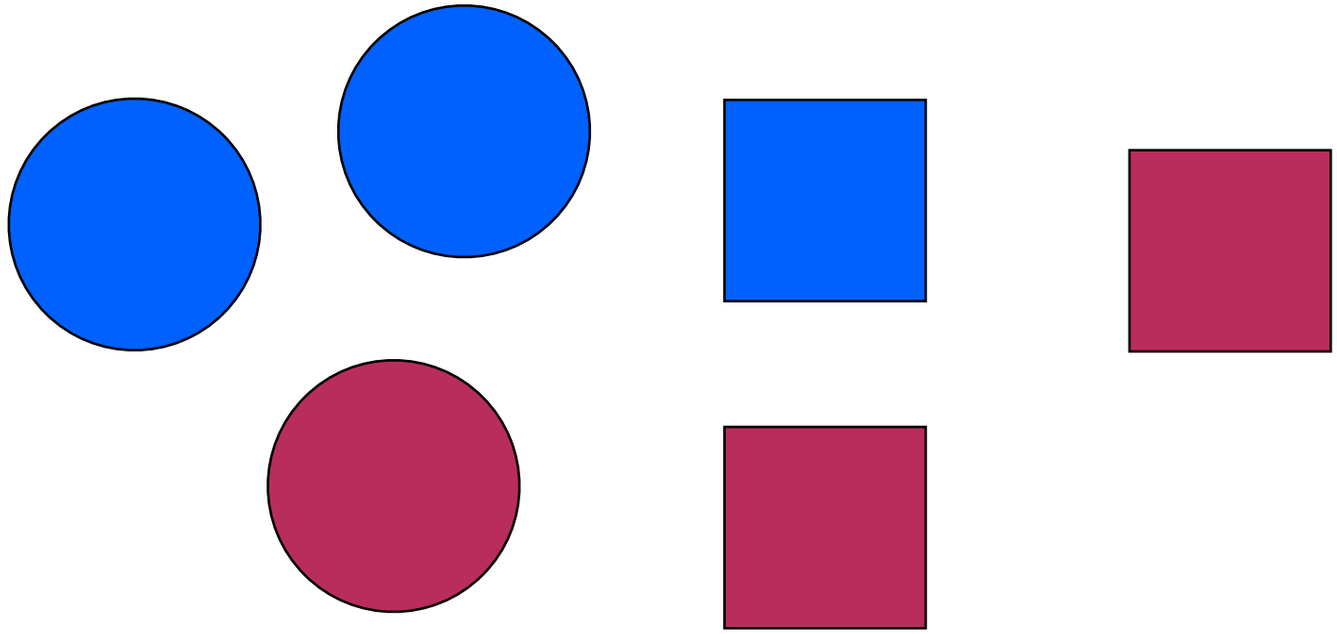
- Probability that "not A" occurs
  - P(not A) = 1 - P(A)

- Probability that "either A or B" occurs
  - P (A or B) = P(A) + P(B) - P(A and B)

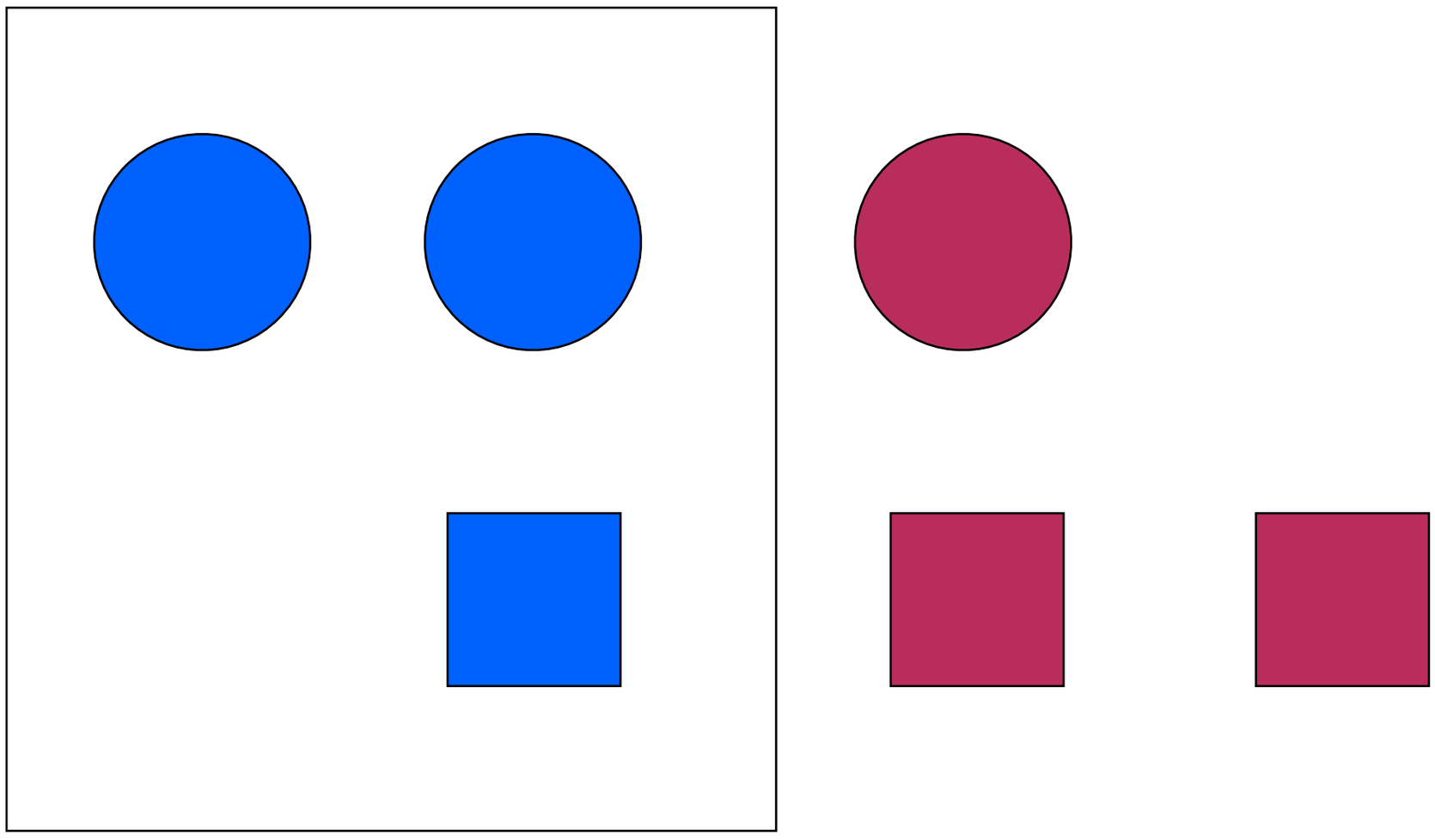
- Probability that "A and B" both occur
  - P(A and B) = P(A) P(B | A)
  - P (A and B) is usually written as P(A, B)

Conditional probability, P(A|B)

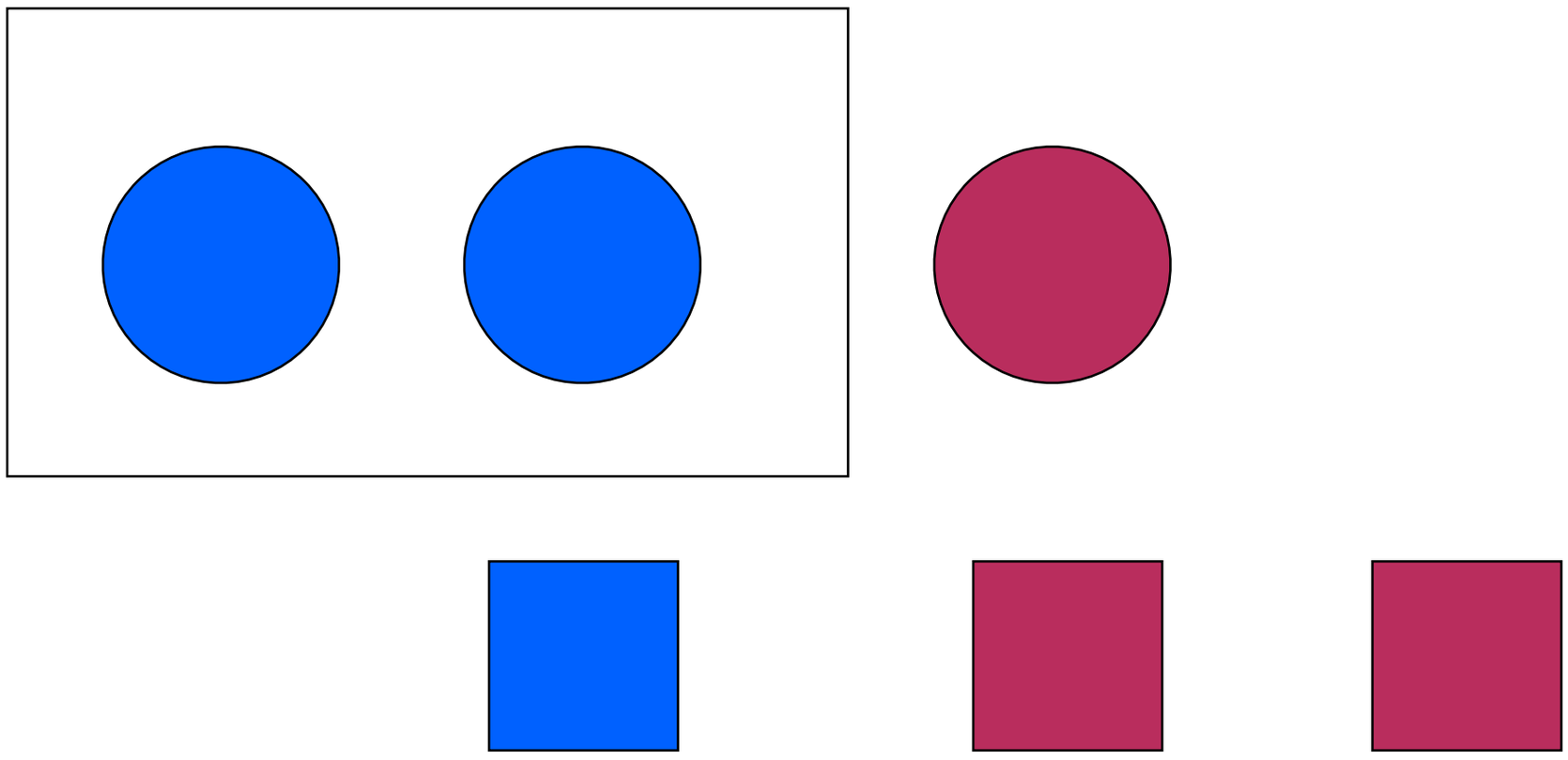
- Probability that A is true, given that B is true
  - Not the same thing as P(A,B) or P(A)
  - Often the cause of some confusion



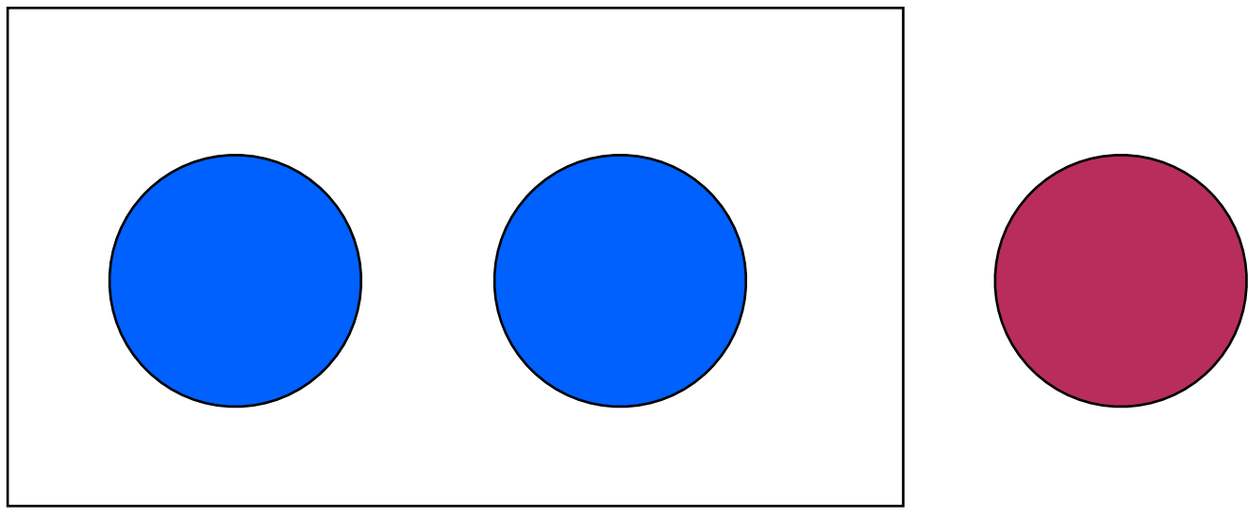
P(blue) = 0.5



P(blue, circle) = 0.333



P(blue | circle) = 0.667



The squares are ignored, because it is a given that the object is a circle

Learning via Bayesian inference

Bayes' rule

- Consider two events, A and B
- We can write the joint probability P(A , B) in two different ways:

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

- Therefore,

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

- And so,

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

Bayesian inference

- The trick:
  - Let B = "hypothesis h about the world is true"
  - Let A = "data set d is observed"

- What it implies:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

- Bayes' theorem gives a learning rule... what does data d tell us about the plausibility of hypothesis h

Bayesian inference

P(d|h) : the <u>likelihood</u> of observing d if h is true

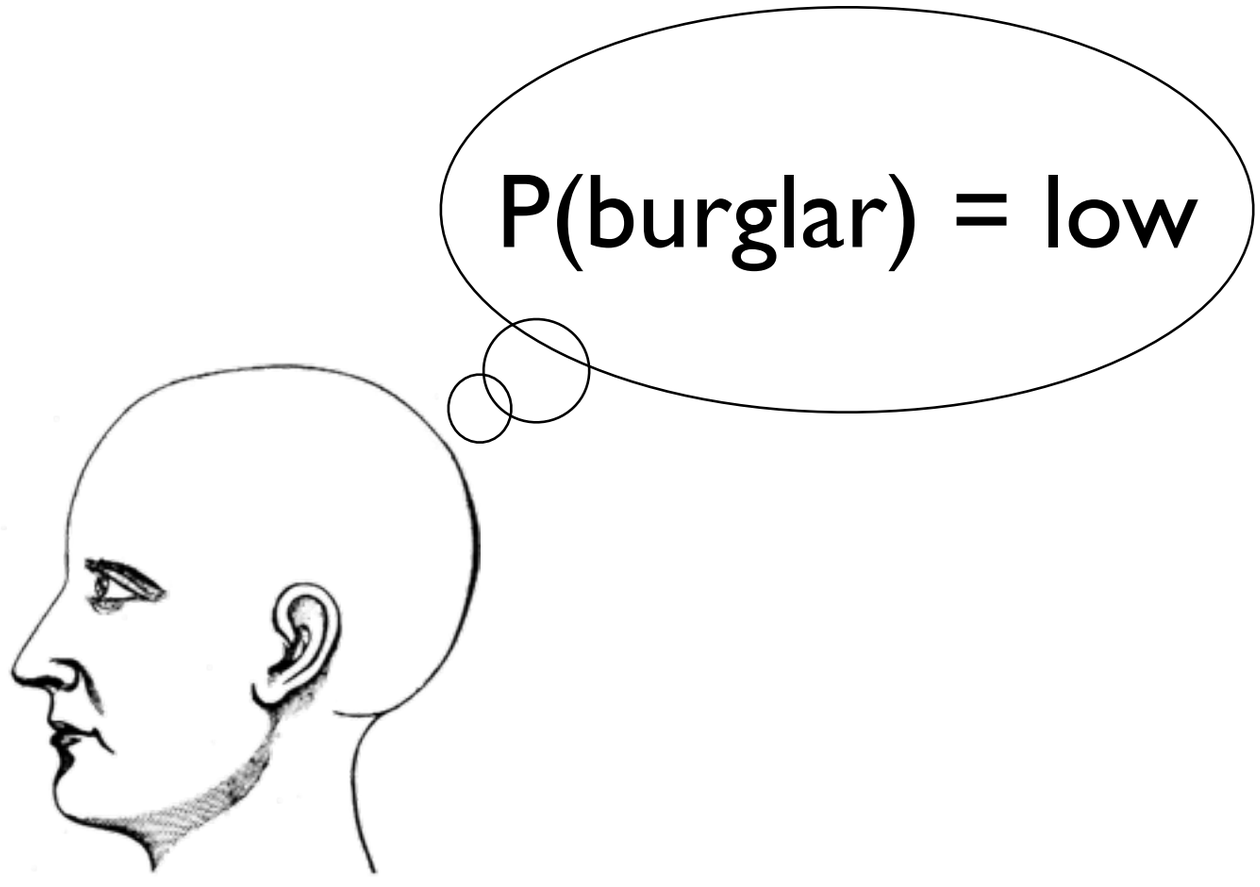
P(h) : the <u>prior probability</u> that h is true

P(h|d) : the <u>posterior probability</u> that h is true

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

P(d) : discussed later

The priors P(h) describe the learner's initial beliefs



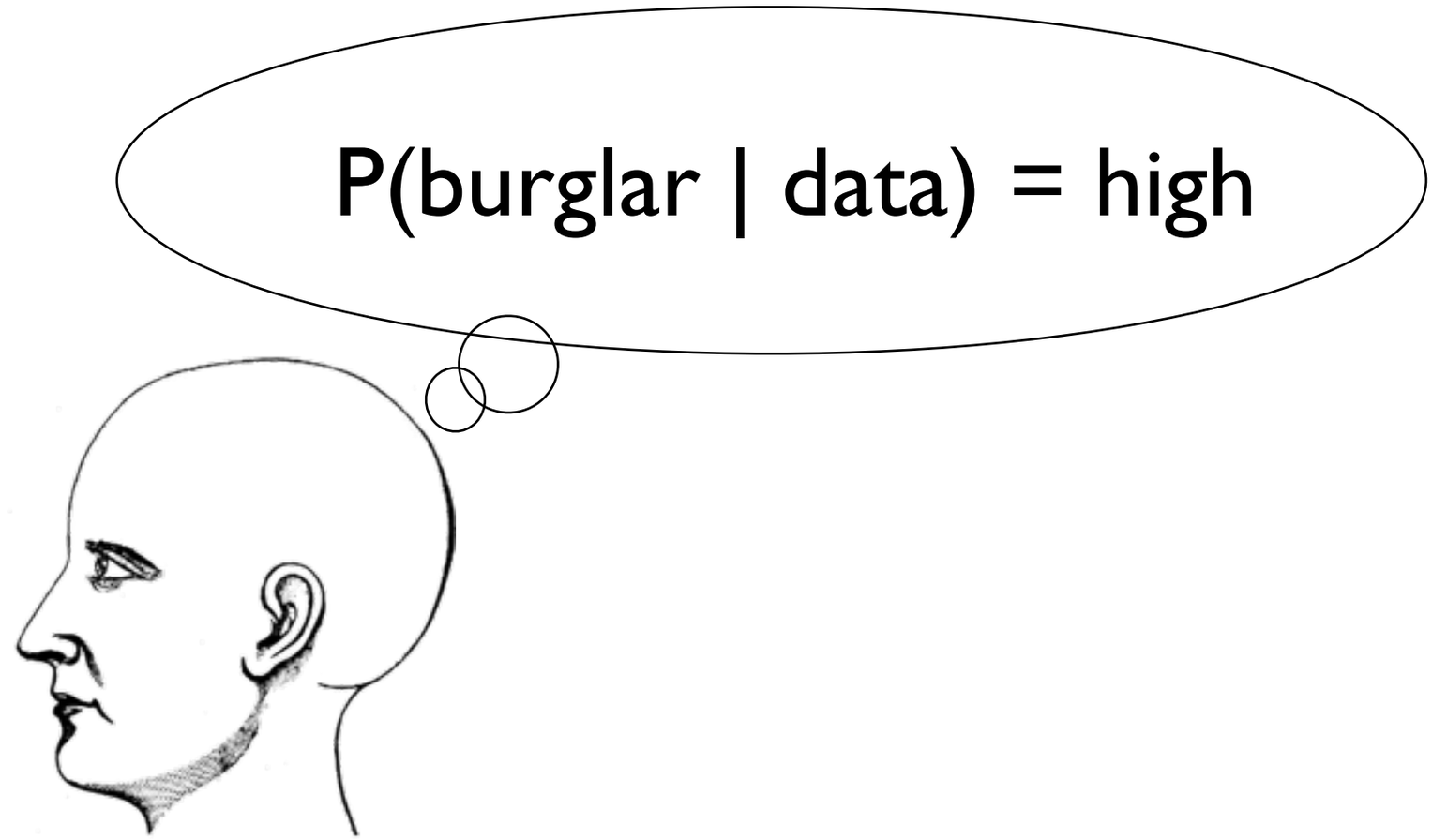
The likelihood P(d|h) is the plausibility of the data if the hypothesis is true

data = (
smashed my window,
holding my TV,
in my lounge room
)

P(data | burglar) = moderate

P(data | not burglar) = very low

The posteriors P(h|d) describe the



More formally

- The data d belongs to a sample space D of possible data sets that you might have observed
- The hypothesis h belongs to a hypothesis space H of theories that might be true
- The prior distribution P(h) is a probability distribution over possible hypotheses
- Every hypothesis h specifies its own likelihood function P(d|h), which is a probability distribution over possible data sets

Probability of the data, P(d)

- Simplest way to think about it is to note that the posterior distribution P(h|d) needs to be a proper probability distribution
- So the sum (over h) of P(h|d) must equal 1
- Gives:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}$$

Side note

- Notational issue
  - Sometimes in this class we'll use **d** for data
  - At other times we'll denote data with **x**

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}$$

$$P(h|x) = \frac{P(x|h)P(h)}{\sum_{h' \in \mathcal{H}} P(x|h')P(h')}$$

The taxi cab problem
(a.k.a. “hey, the prior matters!”)

Most (80%) of the taxis in Simpletown are green, with the rest (20%) being yellow. In a traffic accident involving a hit-and-run taxi, a witness claims the taxi was yellow. Careful testing shows that the witness can successfully identify the colour of a taxi only 75% of the time due to bad eyesight.

On the balance of probabilities, should we hold Yellow Taxi Company liable?

Specification of the problem

- Hypothesis space
  - h1: taxi is yellow
  - h2: taxi is green

- Priors
  - P(h1) = 0.2
  - P(h2) = 0.8

Specification of the problem

- Hypothesis space
  - h1: taxi is yellow
  - h2: taxi is green
- Priors
  - P(h1) = 0.2
  - P(h2) = 0.8
- Data
  - d: witness says yellow
- Likelihood
  - P(d | h1) = 0.75
  - P(d | h2) = 0.25

$$
\begin{aligned}
P(h_1|x) &= \frac{P(x|h_1)P(h_1)}{P(x|h_1)P(h_1) + P(x|h_2)P(h_2)} \\
&= \frac{.75 \times .2}{.75 \times .2 + .25 \times .8} \\
&\approx 0.43
\end{aligned}
$$

No. Only a 43 percent chance it was yellow!

$$
\begin{aligned}
P(h_1|x) &= \frac{P(x|h_1)P(h_1)}{P(x|h_1)P(h_1) + P(x|h_2)P(h_2)} \\
&= \frac{.75 \times .2}{.75 \times .2 + .25 \times .8} \\
&\approx 0.43
\end{aligned}
$$

No. Only a 43 percent chance it was yellow!

psychology: people ignore the base rate (prior) in this problem

The Monty Hall problem
(a.k.a. “the likelihood matters too!”)

You're a contestant on a game show, and there is a prize behind one of three doors, labelled A, B and C. The host asks you to pick a door, and you choose A. He then opens up door B, and shows you that there is no prize there. Finally, he asks if you would like to switch to door C, or if you want to stay with door A.

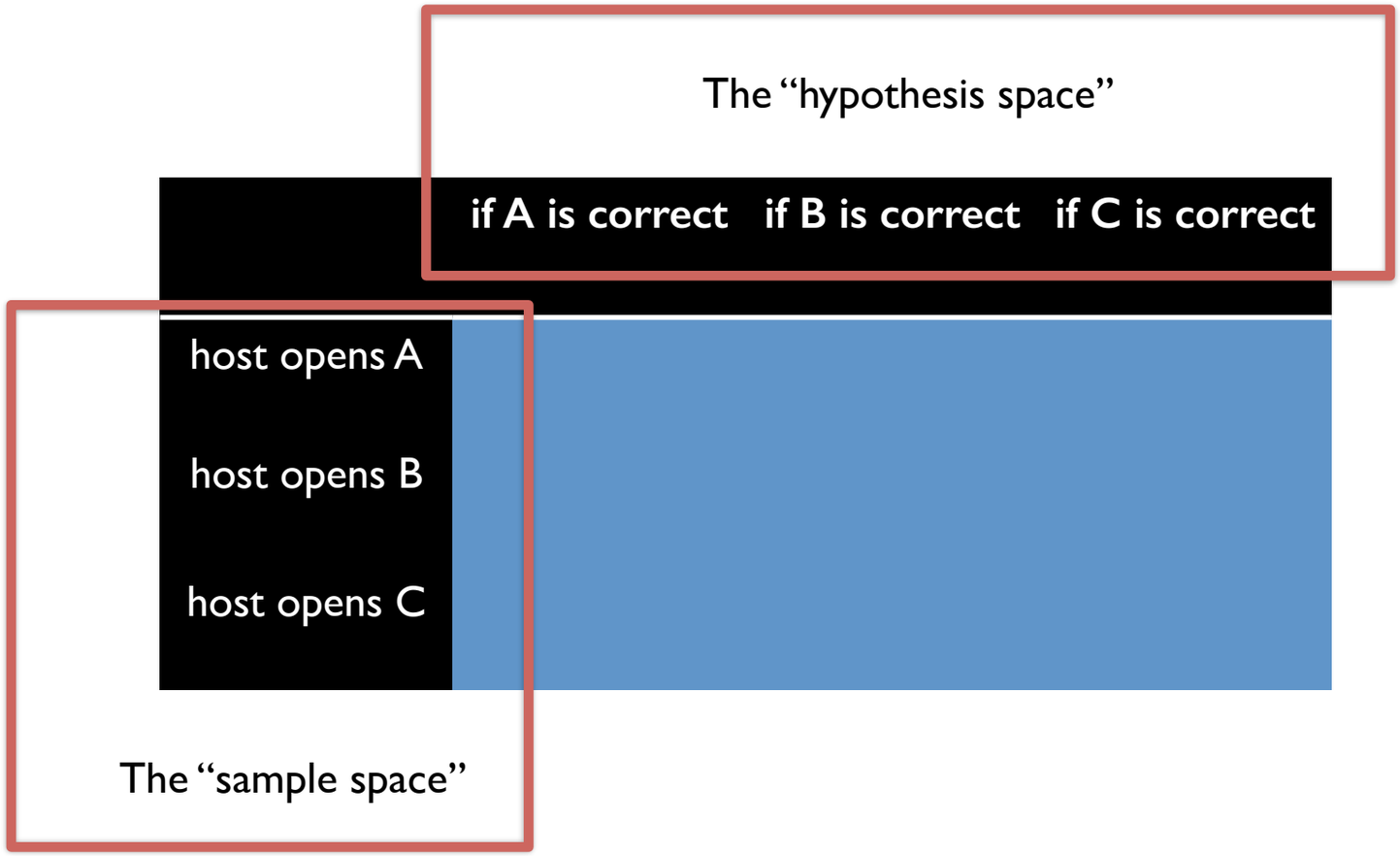
What should you do? Does it matter?

$$
\begin{aligned}
&P(\text{C wins}|\text{B revealed empty, you chose A}) \\
&= \frac{P(\text{B revealed empty}|\text{C wins, you chose A})P(\text{C wins}|\text{you chose A})}{\sum_{X\in\{A,B,C\}} P(\text{B revealed empty}|\text{X wins, you chose A})P(\text{X wins}|\text{you chose A})} \\
&= \frac{P(\text{B revealed empty}|\text{C wins, you chose A})P(\text{C wins})}{\sum_{X\in\{A,B,C\}} P(\text{B revealed empty}|\text{X wins, you chose A})P(\text{X wins})} \\
&= \frac{P(\text{B revealed empty}|\text{C wins, you chose A}) \times 0.33}{\sum_{X\in\{A,B,C\}} P(\text{B revealed empty}|\text{X wins, you chose A}) \times 0.33} \\
&= \frac{P(\text{B revealed empty}|\text{C wins, you chose A})}{\sum_{X\in\{A,B,C\}} P(\text{B revealed empty}|\text{X wins, you chose A})} \\
&= \frac{P(\text{B r.e.}|\text{C w., y.c. A})}{\sum_{X\in\{A,B,C\}} P(\text{B r.e.}|\text{X w., y.c. A})} \\
&= \frac{P(\text{B r.e.}|\text{C w., y.c. A})}{P(\text{B r.e.}|\text{A w., y.c. A}) + P(\text{B r.e.}|\text{B w., y.c. A}) + P(\text{B r.e.}|\text{C w., y.c. A})} \\
&= \frac{1}{.5+0+1} \\
&= 0.66
\end{aligned}
$$

You should switch to door C. It has a 2/3 chance of winning

This problem is all about the likelihood

This problem is all about the likelihood



This problem is all about the likelihood

	if A is correct	if B is correct	if C is correct
host opens A	0%	0%	0%
host opens B			
host opens C			

You chose A... so an intelligent host won't open A

Host behaviour is not arbitrary:

	if A is correct	if B is correct	if C is correct
host opens A	0%	0%	0%
host opens B		0%	
host opens C			0%

The host isn't going to open the correct door... defeats the point of offering a choice

Host behaviour is not arbitrary:

	if A is correct	if B is correct	if C is correct
host opens A	0%	0%	0%
host opens B		0%	
host opens C			0%

If you were originally correct (in selecting A), then the host has two options... but if B or C is true, the host has only one option

Host behaviour is not arbitrary:

	if A is correct	if B is correct	if C is correct
host opens A	0%	0%	0%
host opens B	50%	0%	100%
host opens C	50%	100%	0%

If the host has no other biases at all, here's what we end up with as the likelihoods

Host behaviour is not arbitrary:

	if A is correct	if B is correct	if C is correct
host opens A	0%	0%	0%
host opens B	50%	0%	100%
host opens C	50%	100%	0%

If your choice of A was wrong, then the host really was more likely to have opted to open door B... so the host's behaviour is informative

Host behaviour is not arbitrary:

	if A is correct	if B is correct	if C is correct
host opens A	0%	0%	0%
host opens B	50%	0%	100%
host opens C	50%	100%	0%

psychology: people ignore the likelihood in this problem!

Suspicious coin flips
(a.k.a. “is there any psychology in this???”)

Notation

- Suppose we're flipping coins, and are trying to work out whether or not a coin is biased towards heads (H) or tails (T)
- Denote $P(H) = \theta$

- Now let's flip the coin five times...

Two possible data sets...

HHTHT

HHHHH

Comparing two hypotheses

- Suppose we have only two hypotheses:
  - h1: $\theta = .5$ "this is a fair coin"
  - h2: $\theta = 1$ "the coin always comes up heads"

- Bayesian inference:

$$P(h|x) = \frac{P(x|h)P(h)}{\sum_{h' \in \mathcal{H}} P(x|h')P(h')}$$

- This is easy to compute (only two hypotheses!)
- But I want to show a slightly different perspective

Comparing two hypotheses

- Posterior odds ratio:
  - Relative plausibility of two hypotheses
  - Divide the posterior probability of hypothesis 1 by the posterior probability of hypothesis 2
  - Can be used when there are lots of hypotheses
  - The "nasty" denominator term vanishes...

$$\frac{P(h_1|x)}{P(h_2|x)} = \frac{P(x|h_1)}{P(x|h_2)} \times \frac{P(h_1)}{P(h_2)}$$

posterior odds ratio | likelihood ratio | prior odds ratio

Comparing two hypotheses

- Suppose the prior odds are 999 to 1 in favour of the fair coin... Not completely unreasonable: most coins are pretty fair.

=999

$$\frac{P(h_1|x)}{P(h_2|x)} = \frac{P(x|h_1)}{P(x|h_2)} \times \boxed{\frac{P(h_1)}{P(h_2)}}$$

Data: HHTHT

- Hypothesis 1 (fair coin) says this is just as likely as any other sequence, and has a 1/32 chance
- Hypothesis 2 (always heads) says it's impossible, and so it's probability 0

$$\frac{P(h_1|x)}{P(h_2|x)} = \boxed{\begin{matrix} =1/32 \\ \frac{P(x|h_1)}{P(x|h_2)} \\ =0 \end{matrix}} \times \frac{P(h_1)}{P(h_2)}$$

- Posterior odds infinite... h1 is definitely superior

Data: HHHHH

- Hypothesis 1 (fair coin) says this is just as likely as any other sequence, and has a 1/32 chance
- Hypothesis 2 (always heads) says it's the only possibility, and so it's probability 1

=999

$$\frac{P(h_1|x)}{P(h_2|x)} = \underbrace{\frac{P(x|h_1)}{P(x|h_2)}}_{\substack{=1/32 \\ =1}} \times \frac{P(h_1)}{P(h_2)}$$

- Posterior odds still favour the fair coin, by a factor of about 30:1. The data are not informative enough to overwhelm the prior... yet

Data: HHHHHHHHHH

- Hypothesis 1 (fair coin) says this is just as likely as any other sequence, and has a 1/1024 chance
- Hypothesis 2 (always heads) says it's the only possibility, and so it's probability 1

$$\frac{P(h_1|x)}{P(h_2|x)} = \underbrace{\frac{P(x|h_1)}{P(x|h_2)}}_{\substack{=1/1024 \\ =1}} \times \frac{P(h_1)}{P(h_2)}$$

=999

- Posterior odds now favour the "always heads" hypothesis, though only barely

We can use this to do psychology!

- HHHHH looks like a "mere coincidence" while HHHHHHHHHH makes us suspicious.
- From a cognitive science perspective this gives us a measure of the strength of the prior belief…
  - If τ is the threshold (odds ratio) for suspicion, and x is the shortest suspicious sequence, the prior odds for a fair coin is roughly

$$\frac{\tau}{P(x|\text{“fair coin”})} = \tau \times 2^{|x|}$$

  - If τ = 1 and x is somewhere between 10 and 20 heads, prior odds are roughly between 1:1,000 and 1:1,000,000.

Structure in human beliefs

- Why is HHTHT "fair" but HHHHH isn't?
  - **Structured** prior beliefs...
  - We only get suspicious when we can construct a causal theory for the supposed "trick coin" hypothesis.
  - It is easy to imagine how a trick "all-heads" coin could work: low (but not negligible) prior probability.
  - It is hard to imagine how a trick "HHTHT" coin could work: extremely low (genuinely negligible) prior probability.

The main point!

- We want to use Bayesian inference as a tool to learn about how people think
  - We don't really believe that humans actually calculate these quantities... psychological models aren't that absurdly literal
  - But by coding up these models, we learn something interesting about how the mind operates