Computational cognitive modeling

Bayesian modeling

The problem of induction

Original thought experiment due to W. V. Quine (1960).

The problem of induction

A mug?

A mug *on* a table?

Coffee?

A white mug on a white marble table?

3 pm?

Marble?

Handle?

Smell?

White objects?

A beverage?

A mug filled with coffee?

“gavagai”

Location?

Ceramics?

Original thought experiment due to W. V. Quine (1960).

The problem of induction

now you get more data…

The problem of induction

now you get more data…

A mug?

A mug *on* a table?

Coffee?

A white mug on a white marble table?

3 pm?

Marble?

Handle?

Smell?

White objects?

A beverage?

A mug filled with coffee?

“gavagai”

Location?

Ceramics?

Review: an argument for the need for biases

If totally unbiased generalization systems are incapable of making the inductive leap to characterize the new instances, then **the power of a generalization system follows directly from its biases – from decisions based on criteria other than consistency with the training instances.** Therefore, progress toward understanding learning mechanisms depends upon understanding the sources of, and justification for, various biases.

Mitchell (1980)

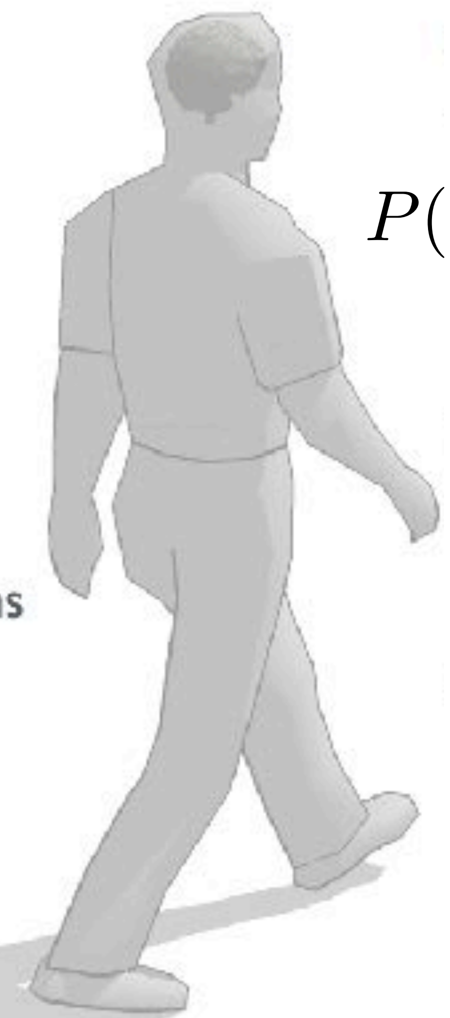
possible biases

- Domain knowledge
- Intended use/goal of generalization (e.g., cost of being incorrect... i.e., risk sensitive)
- Knowledge about the source of training data
- Biases towards simplicity/generality
- Analogy with previous generalizations

Bayesian modeling is an approach for understanding inductive problems, and it typically takes a strong "top-down" strategy

Three levels of description *(David Marr, 1982)*

Computational
Why do things work the way they do?
What is the goal of the computation?
What are the unifying principles?



$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

Algorthmic
What representations can implement such computations?
How does the choice of representations determine the algorithm?

Implementational
How can such a system be built in hardware?
How can neurons carry out the computations?

Key principles of Bayesian models of cognition

- Start with analyzing the computational problem that has to be solved, and describe it as a problem of Bayesian inference

- Use behavioral data to refine your computational level model, and also to build an algorithmic and implementational level account

- Bayesian inference provides a flexible framework for testing different hypotheses about representation, without having to worry about how to define inference and learning

Bayesian inference for evaluating hypotheses in light of data

Data (D): John is coughing

"Bayes' rule"

posterior likelihood prior

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

Hypotheses:

$h_1 =$ John has a cold

$h_2 =$ John has emphysema

$h_3 =$ John has a stomach flu

Which hypotheses should we believe, and with what certainty?

We want to calculate the posterior probabilities: $P(h_1|D)$, $P(h_2|D)$, and $P(h_3|D)$

Example from Josh Tenenbaum

Bayesian inference

posterior, likelihood, prior

$$P(h_i|D) = \frac{P(D|h_i)P(h_i)}{\sum_j P(D|h_j)P(h_j)}$$

Data (D): John is coughing

Hypotheses:

$h_1 =$ John has a cold

$h_2 =$ John has emphysema

$h_3 =$ John has a stomach flu

$P(h_1) = .75 \quad P(D|h_1) = 1$

$P(h_2) = .05 \quad P(D|h_2) = 1$

$P(h_3) = .2 \quad P(D|h_3) = .2$

Prior favors h_1 and h_3, over h_2

Likelihood favors h_1 and h_2, over h_3

Posterior favors h_1, over h_2 and h_3

$$P(h_1|D) = .89 = \frac{.75(1)}{.75(1) + .05(1) + .2(.2)}$$

$$P(h_2|D) = .06$$

$$P(h_3|D) = .05$$

Where does Bayes' rule come from?

Definition of conditional probability:

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

"product rule"

$$P(a,b) = P(a|b)P(b)$$

Derivation

$$P(h,D) = P(h|D)P(D)$$

$$P(h,D) = P(D|h)P(h)$$

product rule applied twice

$$\Downarrow$$

$$P(h|D)P(D) = P(D|h)P(h)$$

Equating the two right hand sides

$$\Downarrow$$

$$\boxed{P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{P(D|h)P(h)}{\sum_{h'} P(D|h')P(h')}}$$

Divide by $P(D)$

Why is this a reasonable way to represent beliefs?

- If your beliefs are inconsistent with axioms of probability, someone can take advantage of you in gambling (see example from Russel and Norvig reading)

Agent 1		Agent 2		Outcome for Agent 1			
Proposition	Belief	Bet	Stakes	$a \wedge b$	$a \wedge \neg b$	$\neg a \wedge b$	$\neg a \wedge \neg b$
a	0.4	a	4 to 6	–6	–6	4	4
b	0.3	b	3 to 7	–7	3	–7	3
$a \vee b$	0.8	$\neg(a \vee b)$	2 to 8	2	2	2	–8
				–11	–1	–1	–1

Figure 13.2 Because Agent 1 has inconsistent beliefs, Agent 2 is able to devise a set of bets that guarantees a loss for Agent 1, no matter what the outcome of a and b.

- Bayes' rule provides a very general account of learning, where prior knowledge can be combined with data to update beliefs

Bayesian concept learning with the number game

Rules and Similarity in Concept Learning

Joshua B. Tenenbaum
Department of Psychology
Stanford University, Stanford, CA 94305
jbt@psych.stanford.edu

In *Advances in neural information processing systems (1999)*

Abstract

This paper argues that two apparently distinct modes of generalizing concepts – abstracting rules and computing similarity to exemplars – should both be seen as special cases of a more general Bayesian learning framework. Bayes explains the specific workings of these two modes – which rules are abstracted, how similarity is measured – as well as why generalization should appear rule- or similarity-based in different situations. This analysis also suggests why the rules/similarity distinction, even if not computationally fundamental, may still be useful at the algorithmic level as part of a principled approximation to fully Bayesian learning.

(You will work with this model in homework 4)

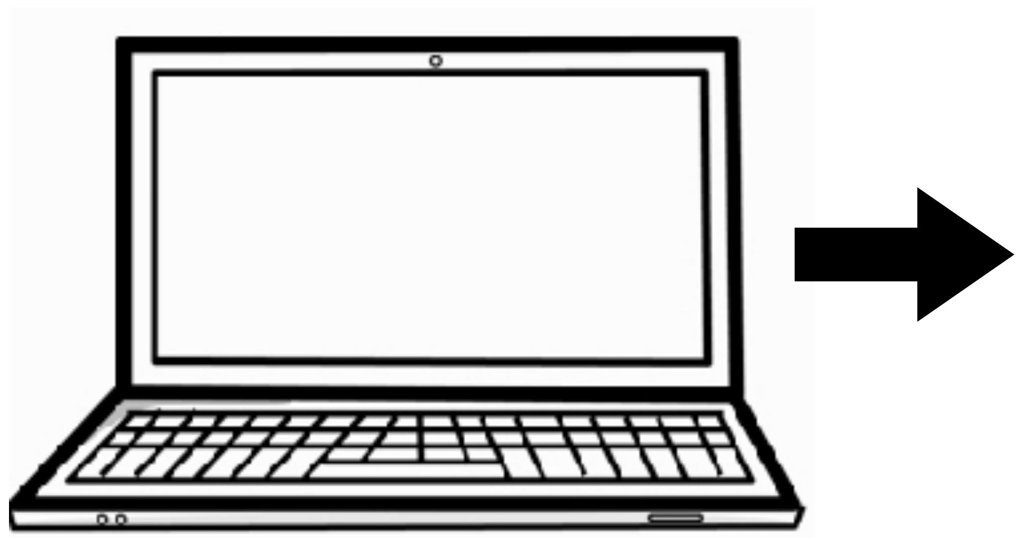
1 Introduction

In domains ranging from reasoning to language acquisition, a broad view is emerging of cognition as a hybrid of two distinct modes of computation, one based on applying abstract rules and the other based on assessing similarity to stored exemplars [7]. Much support for this view comes from the study of concepts and categorization. In generalizing concepts, people's judgments often seem to reflect both rule-based and similarity-based computations [9], and different brain systems are thought to be involved in each case [8]. Recent psychological models of classification typically incorporate some combination of rule-based and similarity-based modules [1,4]. In contrast to this currently popular modularity position, I will argue here that rules and similarity are best seen as two ends of a continuum of possible concept representations. In [11,12], I introduced a general theoretical framework to account

The number game

There is an unknown computer program that generates numbers in the range 1 to 100. You are provided with a small set of random examples from this program.

1 random "yes" example



60

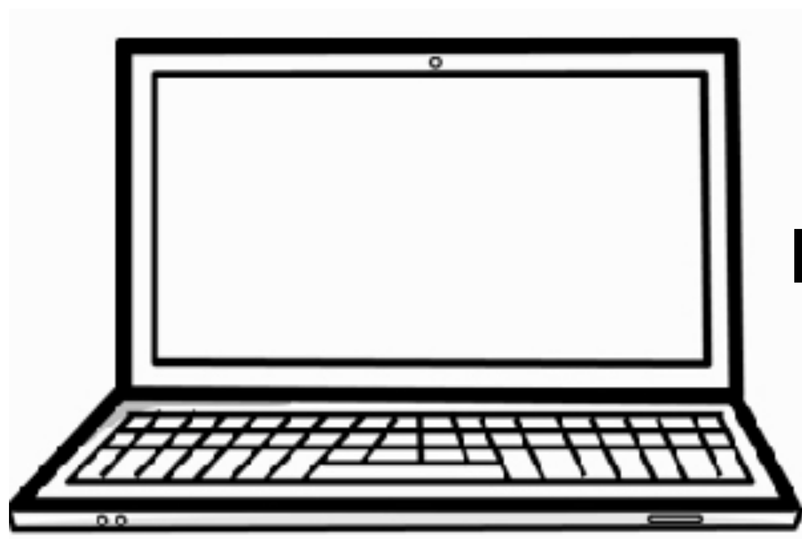
Which numbers will be accepted by the same computer program?

51? 58? 20?

The number game

There is an unknown computer program that generates numbers in the range 1 to 100. You are provided with a small set of random examples from this program.

4 random "yes" examples



60 80 10 30

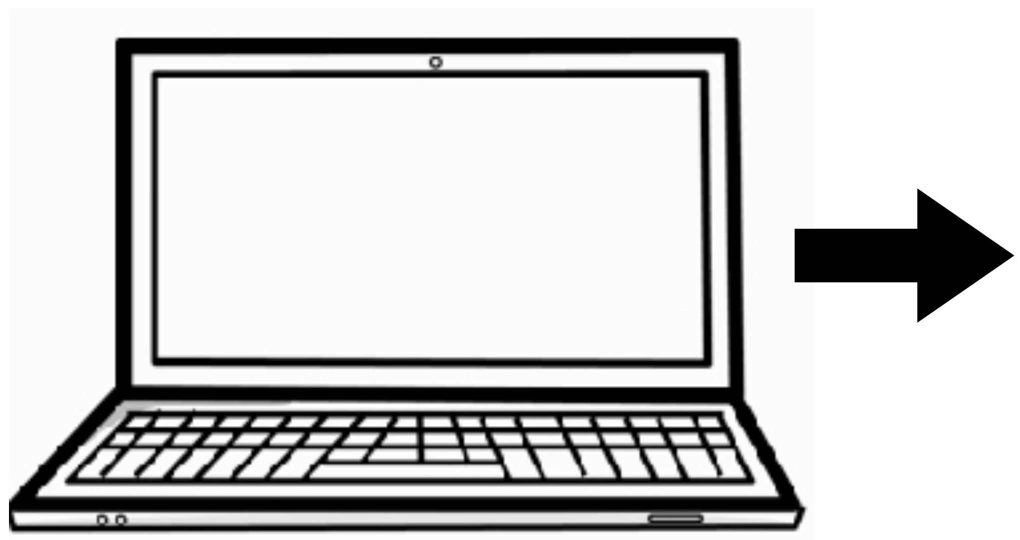
Which numbers will be accepted by the same computer program?

51? 58? 20?

The number game

There is an unknown computer program that generates numbers in the range 1 to 100. You are provided with a small set of random examples from this program.

4 random “yes” examples



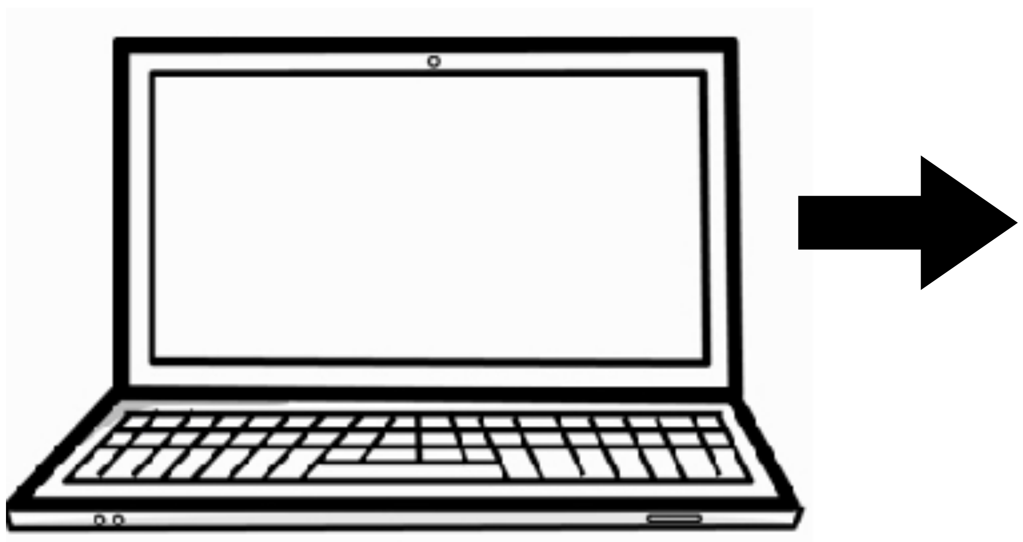
60 52 57 55

Which numbers will be accepted by the same computer program?

51? 58? 20?

A Bayesian model of the number game

random "yes" examples of an **unknown concep**t ***C***



Observations

$$X = \{x^{(1)}, ..., x^{(n)}\}$$

60 52 57 55

Predictions:

Which numbers ***y*** will be accepted by the same computer program ***C***?

$$P(y \in C \mid X)$$

51? 58? 20?

A Bayesian model of the number game

We have observations:

$$X = \{x^{(1)}, ..., x^{(n)}\}$$

We have a space of hypotheses, which are sets of numbers $h \in H$

and prior $P(h)$ (more details next slide)

- *mathematical hypotheses:* odd numbers (h = [1, 3, 5, …, 99]), even numbers (h = [2, 4, 6, … ,100]), square numbers (h = [1, 4, 9, 16]), cube numbers, primes, multiples of n, etc.

- *interval hypotheses:* continuous intervals of the number line

Likelihood $P(X|h)$

$$P(X|h) = \prod_{i=1}^{n} P(x^{(i)}|h)$$

(assumption that examples are independent)

$$P(x^{(i)}|h) = \frac{1}{|h|} \text{ if } x^{(i)} \in h$$

|*h*| is the "size" of *h*

$$= 0 \text{ otherwise}$$

Bayes' rule for computing posterior beliefs:

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

A Bayesian model of the number game

The hypothesis space and prior

Mathematical hypotheses

- odd numbers
- even numbers
- square numbers
- cube numbers
- primes
- multiples of n, such that 3 ≤ n ≤ 12
- powers of n, such that 2 ≤ n ≤ 10
- numbers ending in n, such that 0 ≤ n ≤ 9

(Mathematical hypotheses are equally likely in the prior)

$$P(h)$$

Interval hypotheses

- Intervals between n and m, such that 1 ≤ n ≤ 100; n ≤ m ≤ 100

(Interval hypotheses reweighted to favor intermediate sizes) $P(h)$

λ is free parameter that trades off “math” vs. “interval” hypotheses

A Bayesian model of the number game

We have observations:

$$X = \{x^{(1)}, ..., x^{(n)}\}$$

We want to make predictions for new numbers y:

$$P(y \in C \mid X)$$

Bayes' rule for computing posterior beliefs:

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

Posterior predictions about new example ***y:***

$$P(y \in C \mid X) = \sum_{h \in H} P(y \in C \mid h) P(h|X)$$

first term is 1 or 0 based on membership

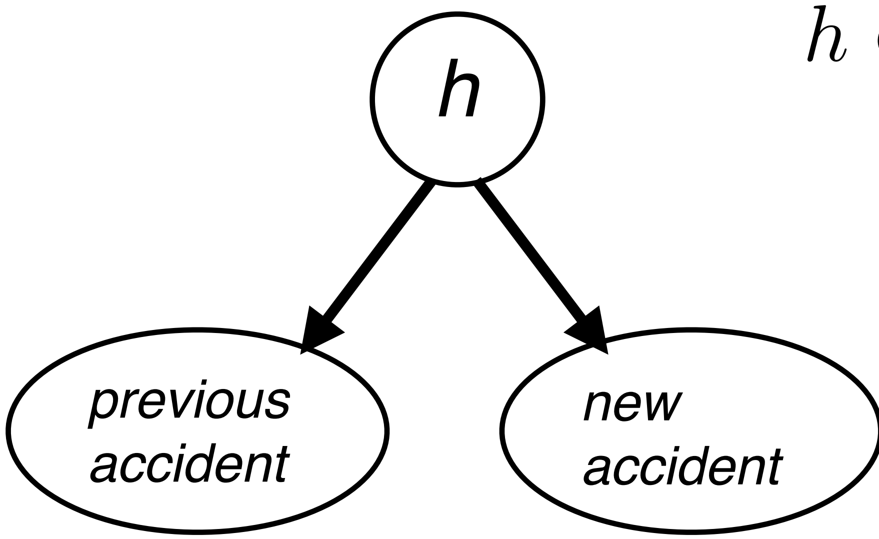
Bayesian hypothesis averaging : when making Bayesian predictions, one must average over all possible hypotheses, weighted by their posterior belief

Examples: Bayesian hypothesis averaging

Say you are an insurance company, and you want to predict which customers are more likely to get into a car accident.

$$P(\text{new accident}|\text{previous accident}) = \sum_h P(\text{new accident}|h)P(h|\text{previous accident})$$

$$h \in \{\text{good driver}, \text{bad driver}\}$$



also known as "marginalization" of a variable (*h*)

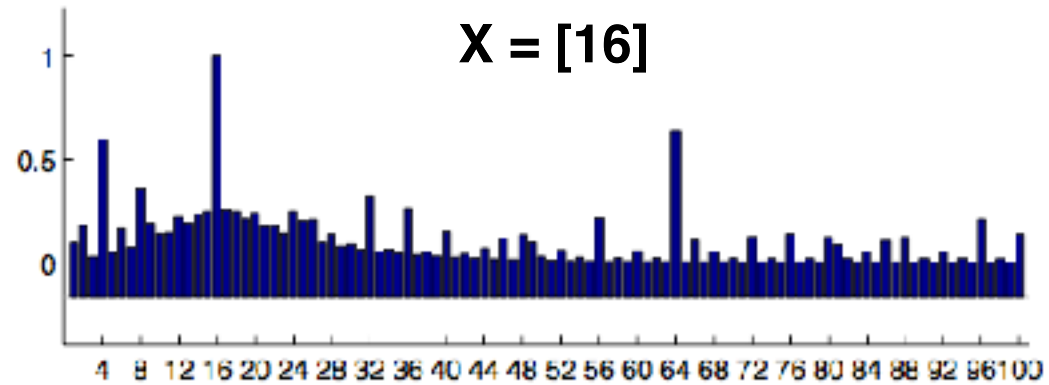
Another example (the previous evidence does not necessarily need to be relevant)

$$P(\text{new accident}|\text{born in Feb.}) = \sum_h P(\text{new accident}|h)P(h|\text{born in Feb.})$$

$$h \in \{\text{good driver}, \text{bad driver}\}$$

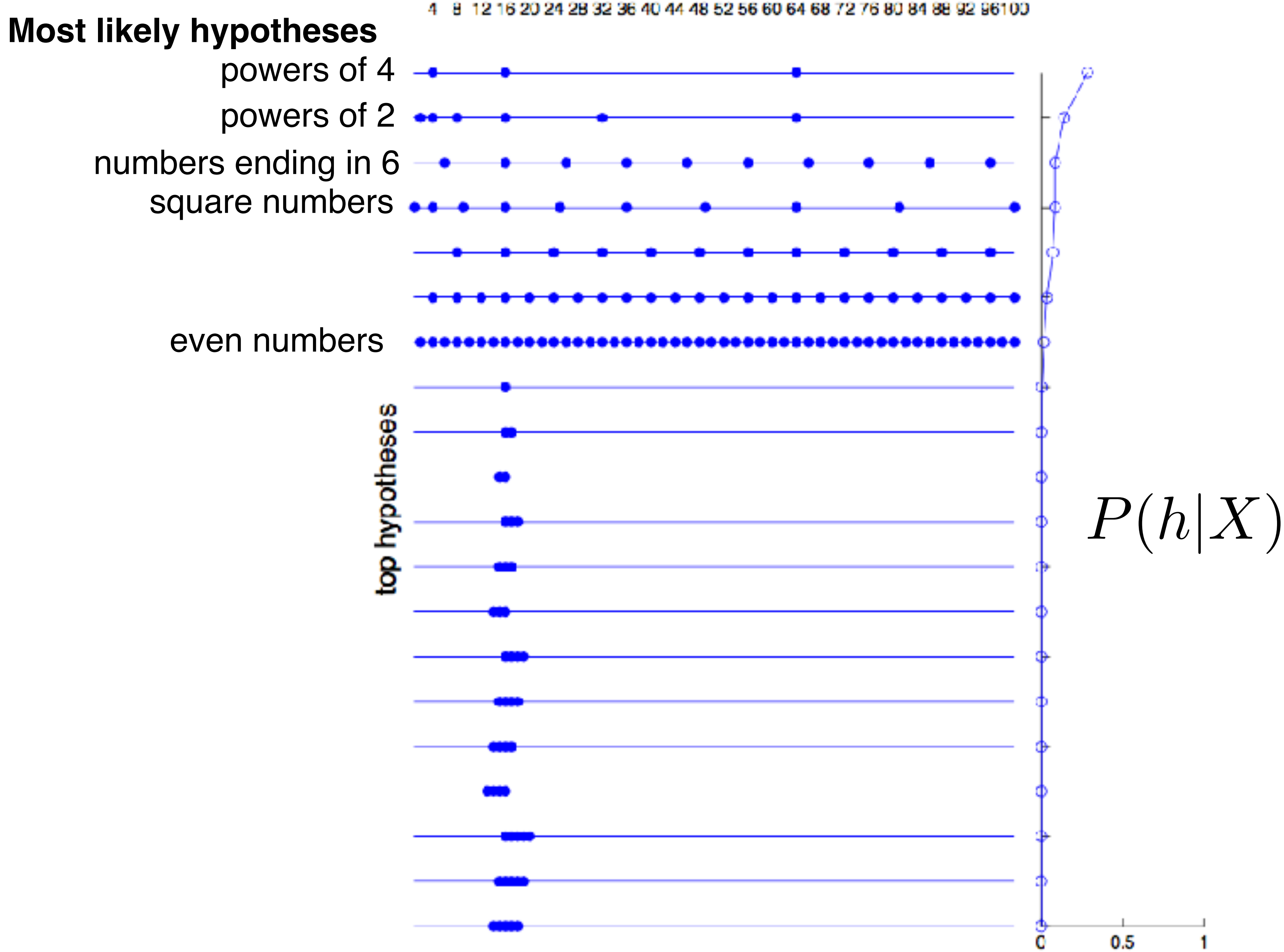
The **size principle:** hypotheses with smaller extensions are more likely than hypotheses with larger extensions

$P(y \in C|X)$



Likelihood

$$P(x^{(i)}|h) = \frac{1}{|h|} \text{ if } x^{(i)} \in h$$
$$= 0 \text{ otherwise}$$



The **size principle:** hypotheses with smaller extensions are more likely than hypotheses with larger extensions

Likelihood **compounding evidence**

$$\boxed{P(X|h) = \prod_{i=1}^{n} P(x^{(i)}|h)}$$

$$P(x^{(i)}|h) = \frac{1}{|h|} \text{ if } x^{(i)} \in h$$
$$= 0 \text{ otherwise}$$

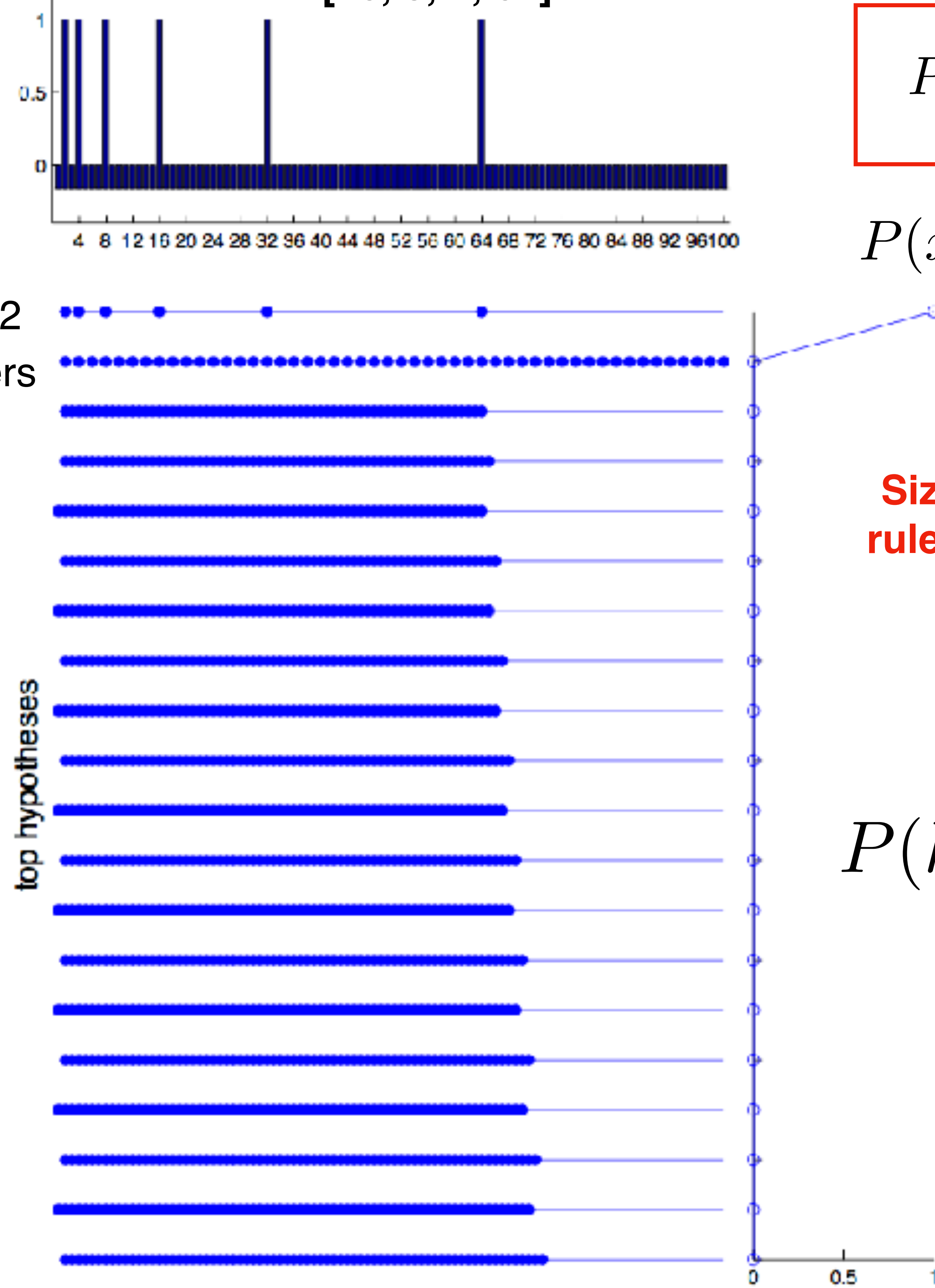
$P(y \in C|X)$

X = [16, 8, 2, 64]

Most likely hypotheses

powers of 2

even numbers



$P(h|X)$

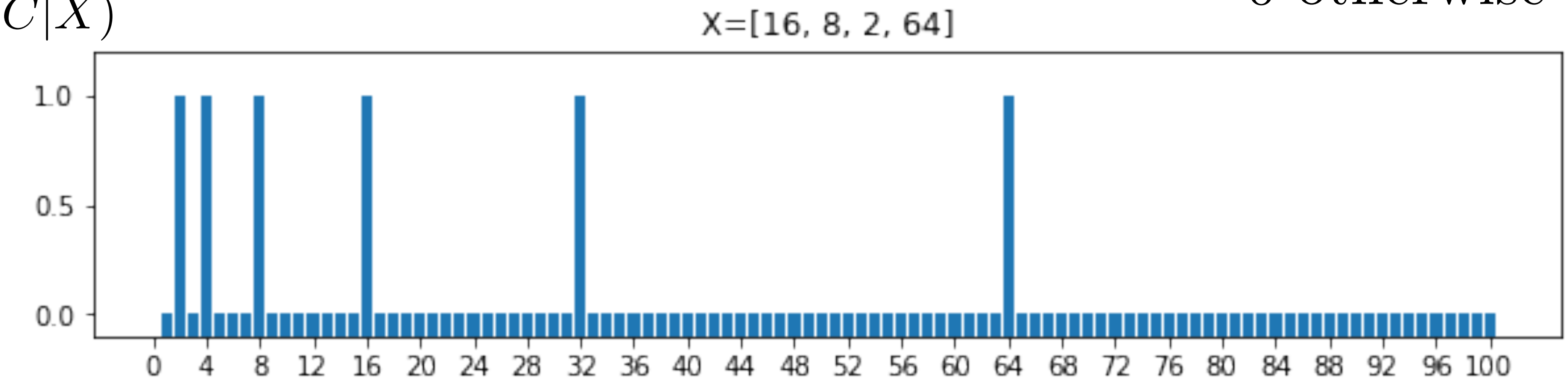
Size principle leads to very sharp, rule-like generalizations from just a few examples

How the size principle influences generalization

With size principle (strong sampling):

$$P(x^{(i)}|h) = \frac{1}{|h|} \text{ if } x^{(i)} \in h$$
$$= 0 \text{ otherwise}$$

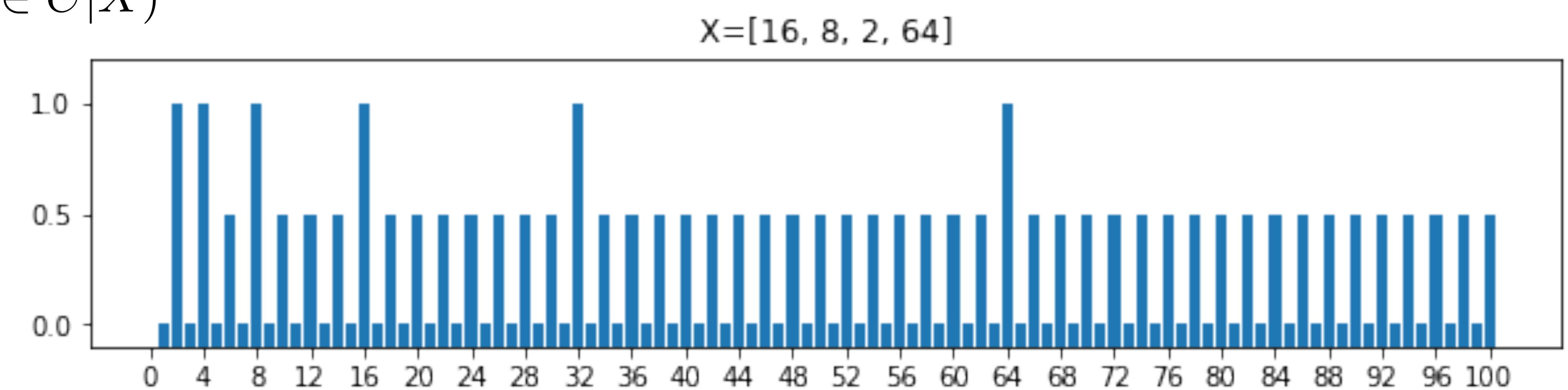
$P(y \in C|X)$



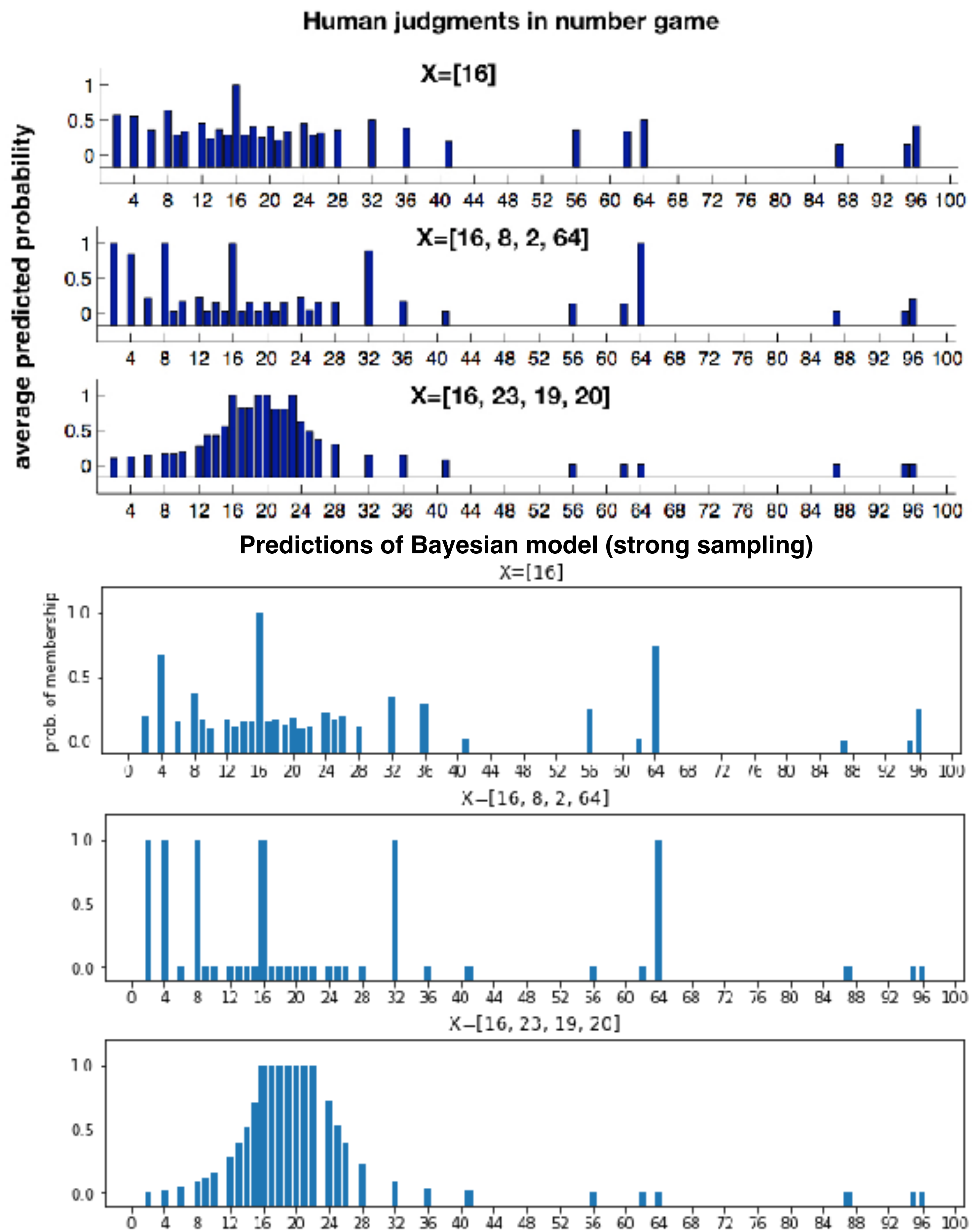
Without size principle (weak sampling):

$$P(X|h) = 1 \text{ if } x^{(i)} \in h \text{ for all } i$$
$$= 0 \text{ otherwise}$$

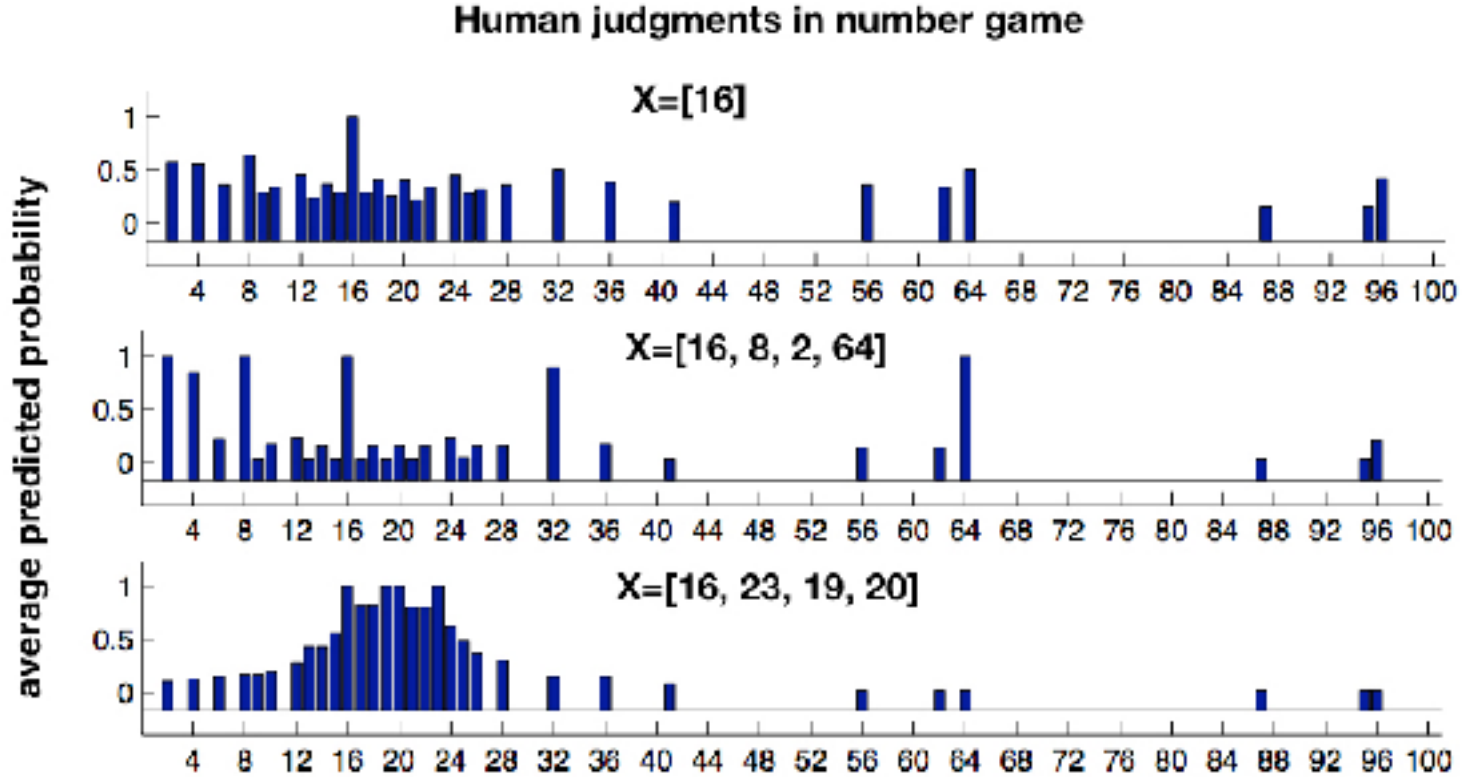
$P(y \in C|X)$

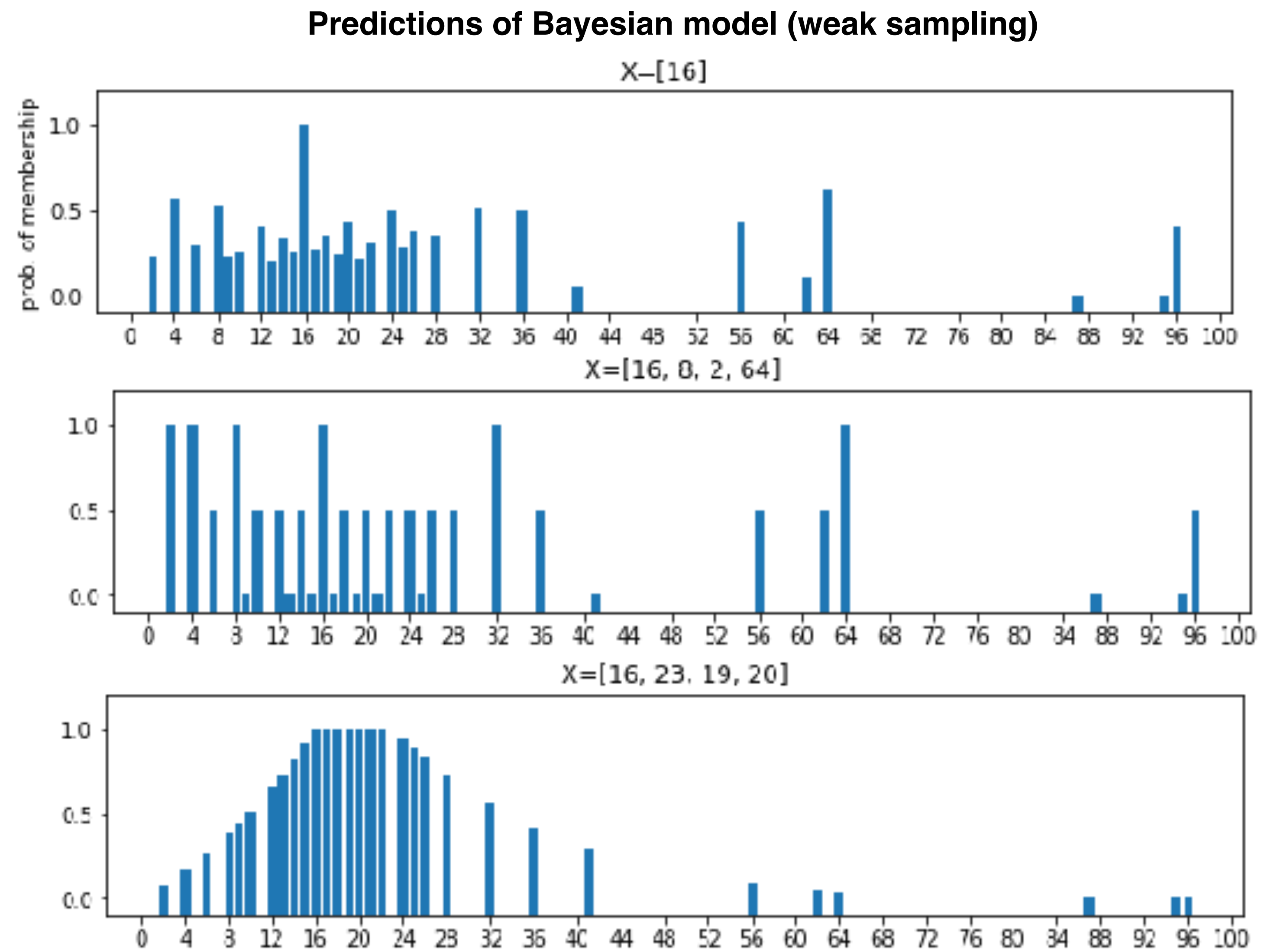


Human vs. model predictions in the number game



Human vs. model predictions in the number game (weak sampling)





(weak sampling does not capture the sharpness of people's generalization curves)

Conclusions from Bayesian concept learning and the number game

• People can make meaningful predictions from very sparse evidence, aided by strong assumptions for how the data is generated (strong sampling)

• People display a mixture of both “rule-like” and “similarity-like” generalizations, depending on what the data entails

• A Bayesian account of concept learning displays both of these characteristics, and can make quantitative predictions regarding how people generalize to new examples.

• Discussion point: Where does the hypothesis space come from?

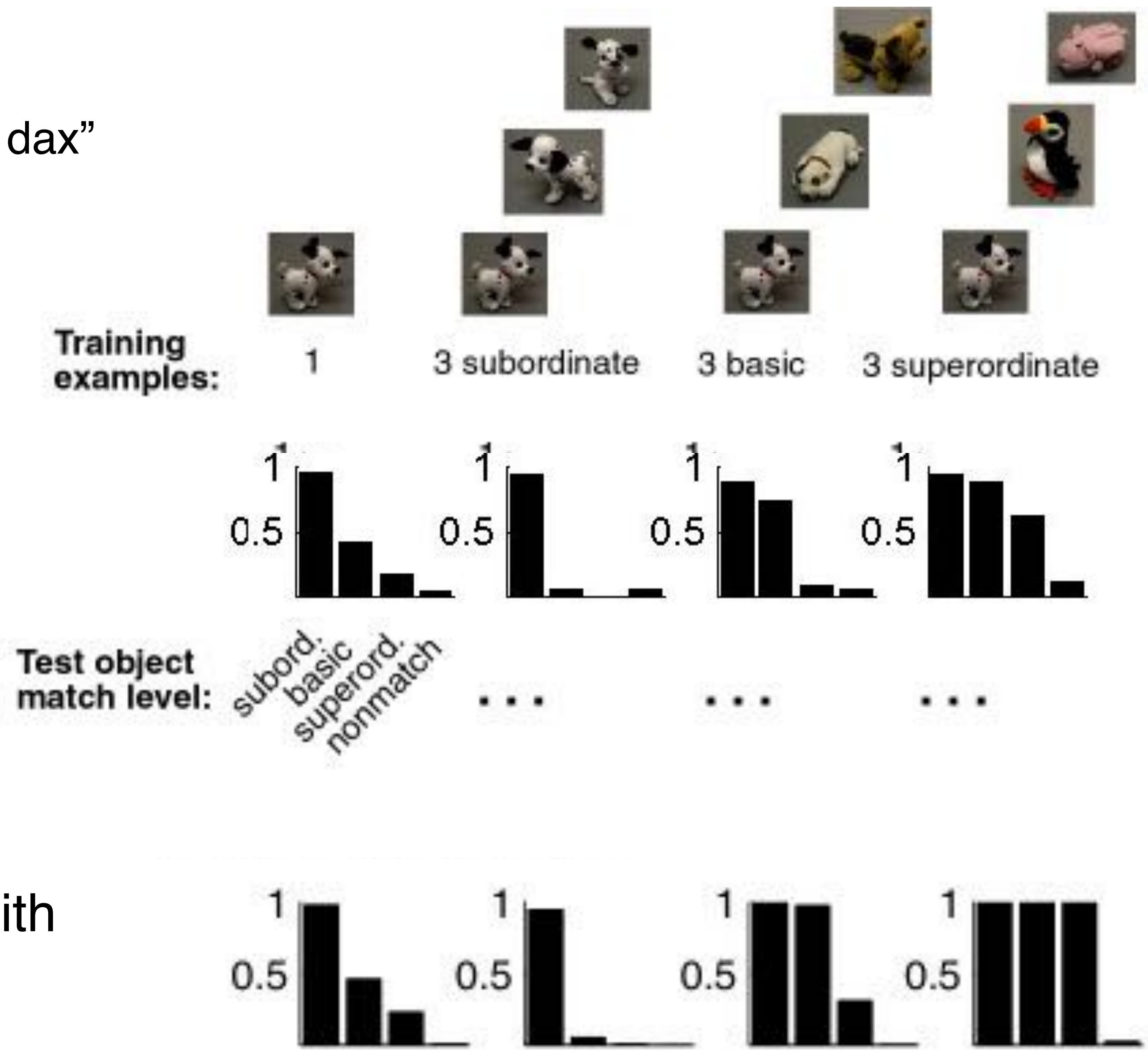
Word learning as Bayesian inference

(Xu and Tenenbaum, 2007, *Psychological Review*)

Prompt: "This is a dax"

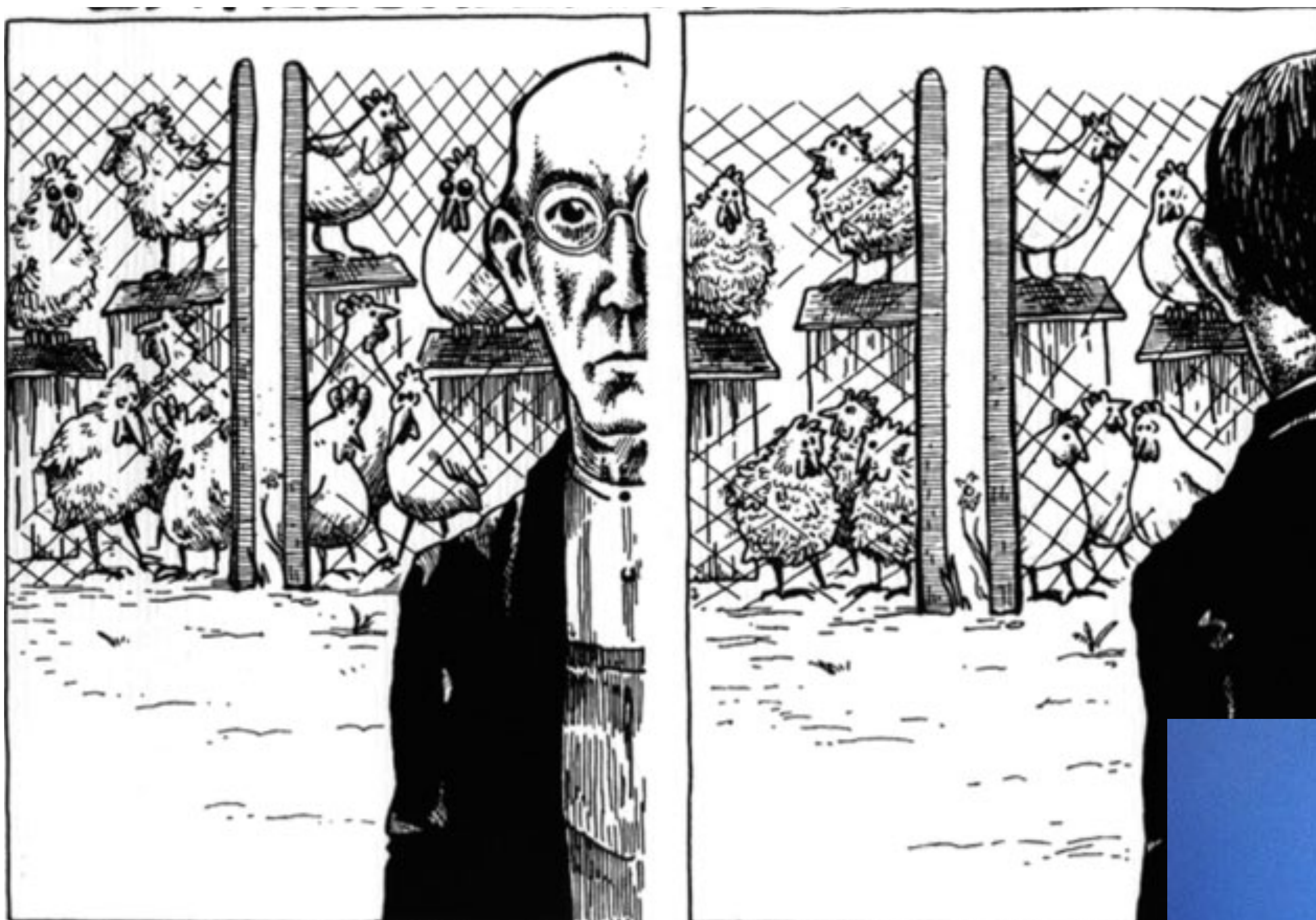
Children's
generalizations

Bayesian
concept learning with
tree-structured
hypothesis space

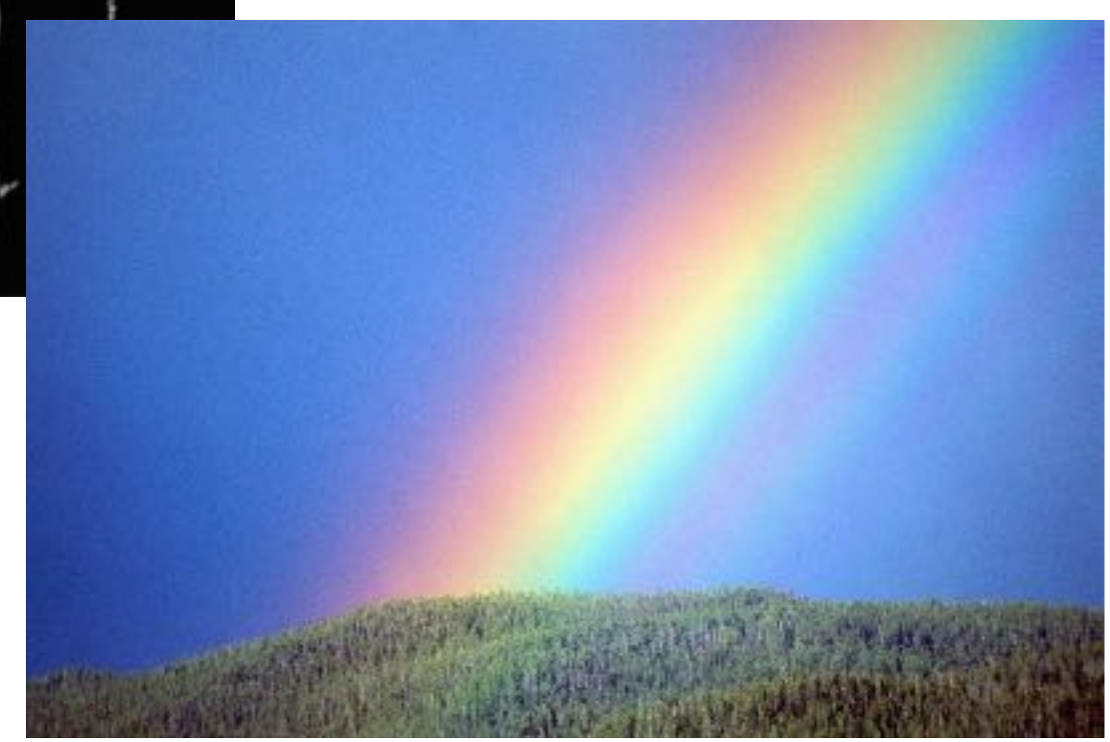


Slide credit: Josh Tenenbaum

Categorical perception: A link between categorization and perception/discrimination



From Goldstone and Hendrickson (2009)

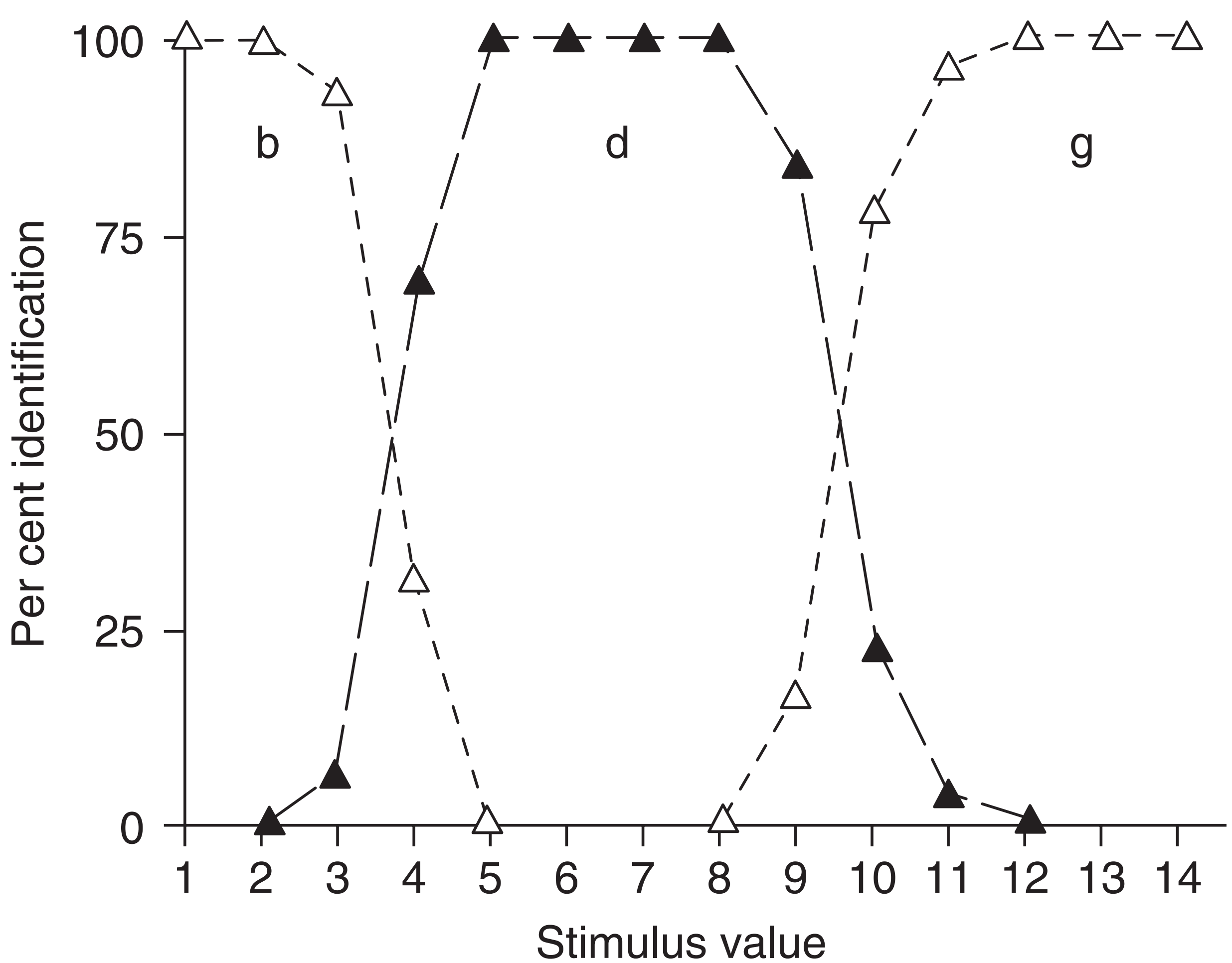


Categorical perception in speech

A link between categorization and discrimination

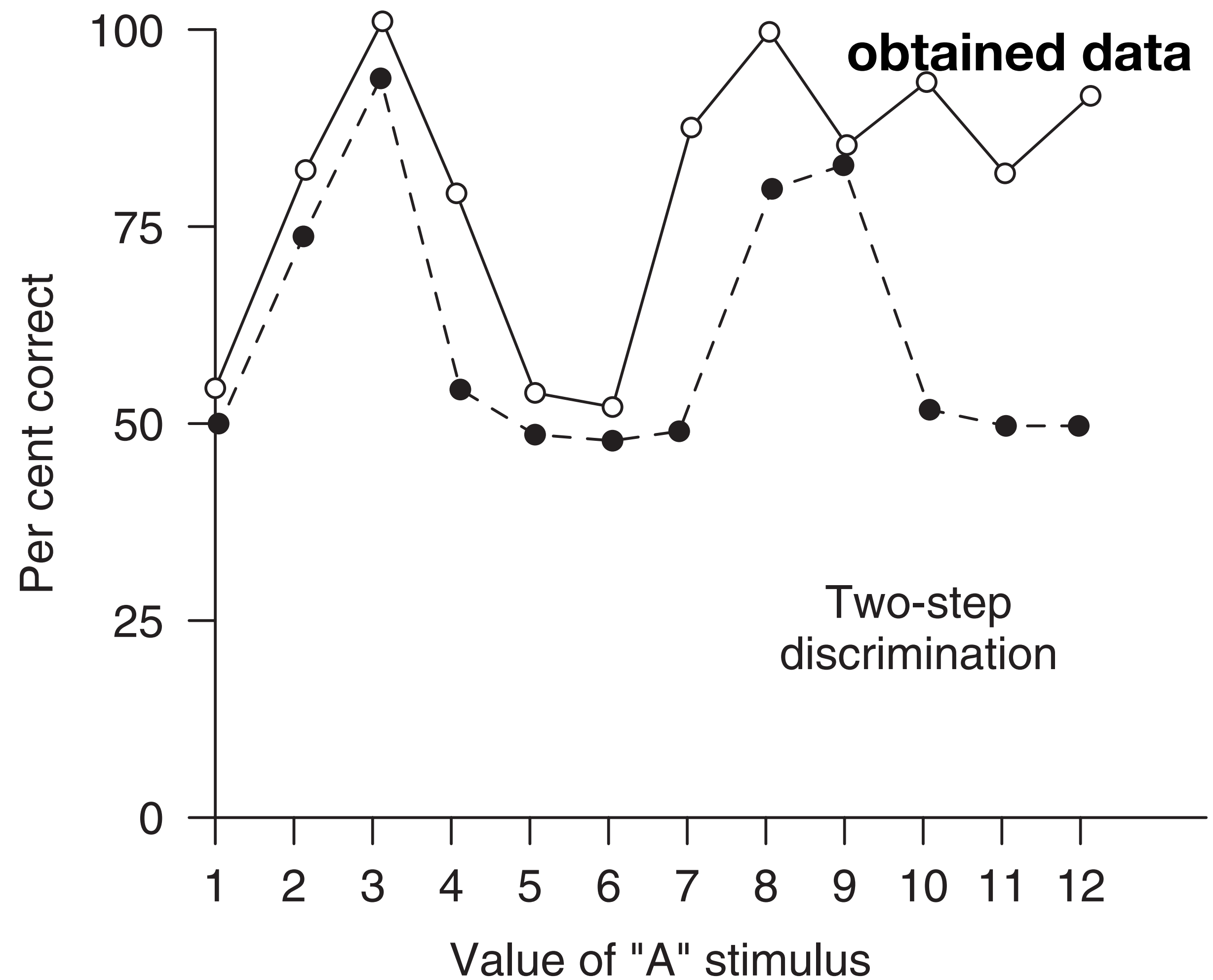
Identification (labeling) task

(“ba” vs. “da” vs. “ga”)



Discrimination task

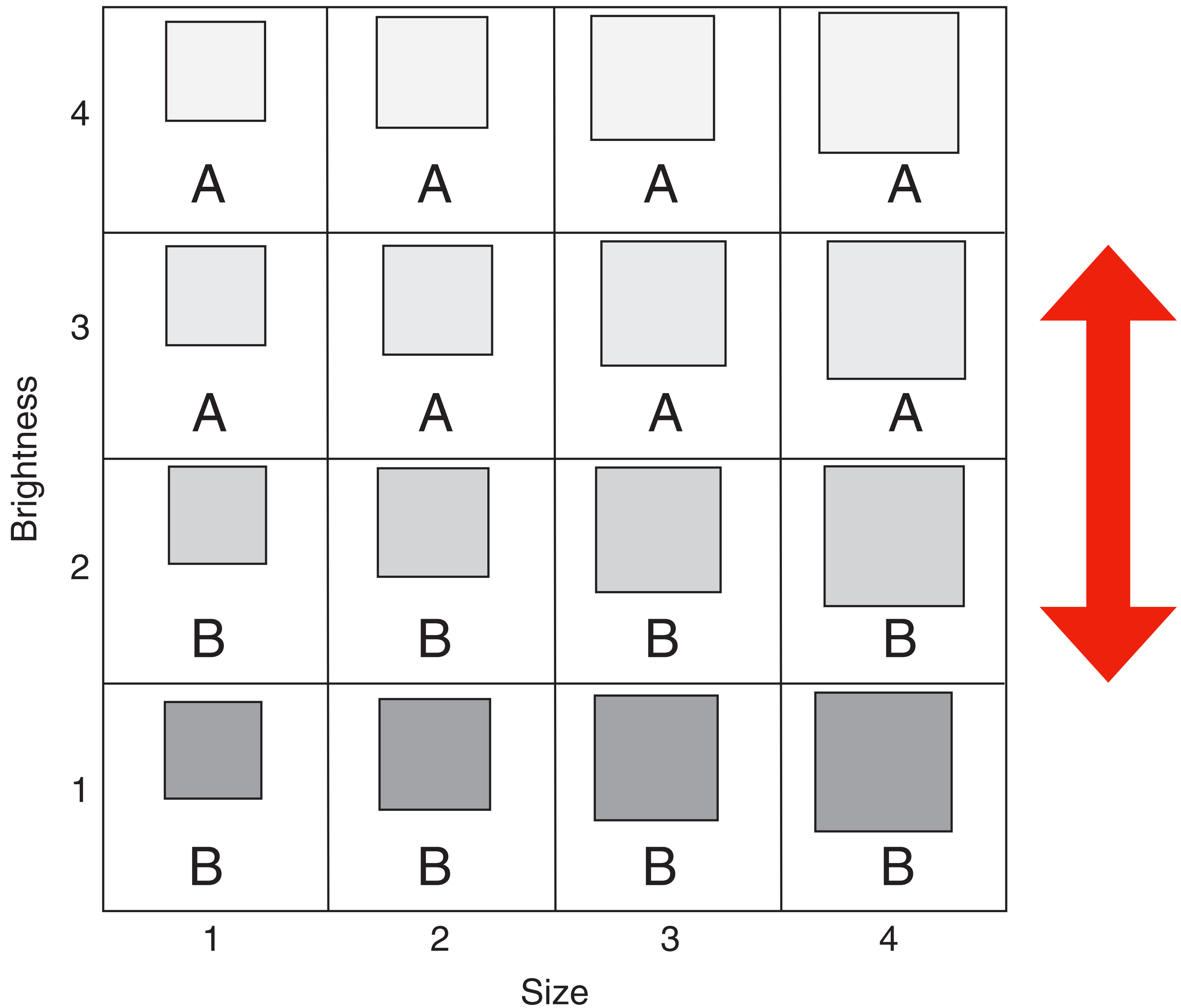
(ABX; which is X identical to, A or B?)



From Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957)

Categorical perception for artificial visual categories

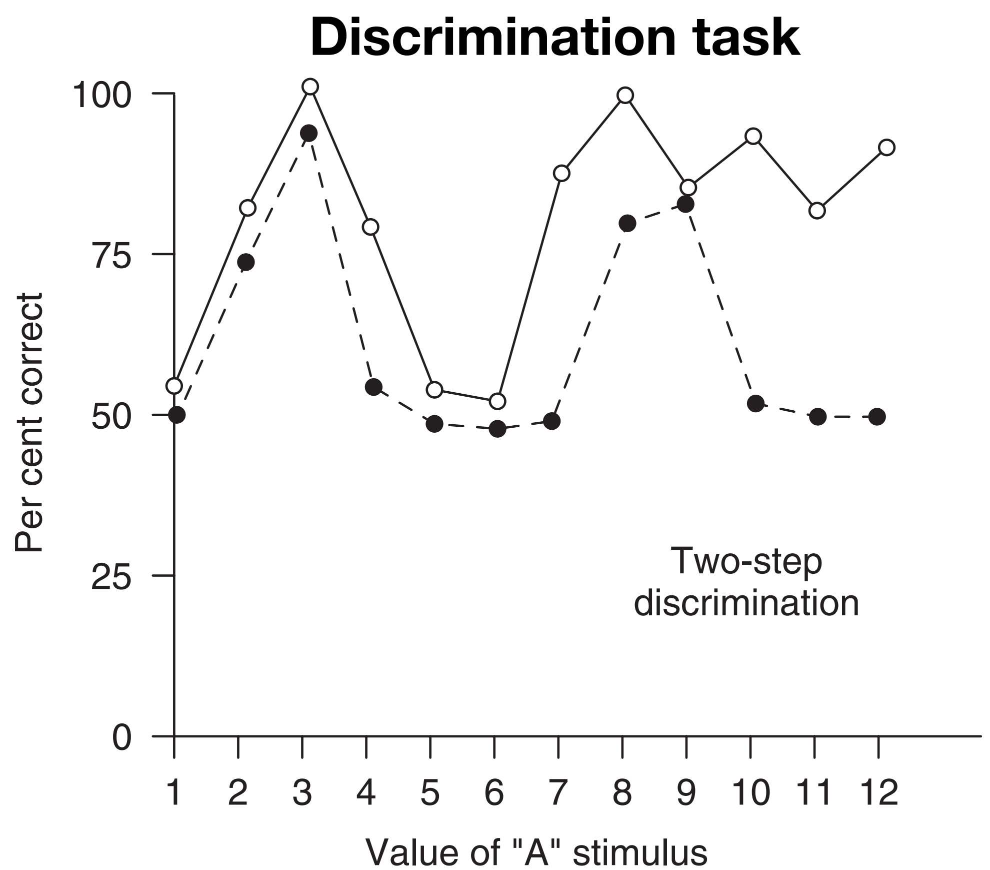
A link between categorization and discrimination



enhanced discrimination along this dimension, especially at the “A” vs. “B” boundary

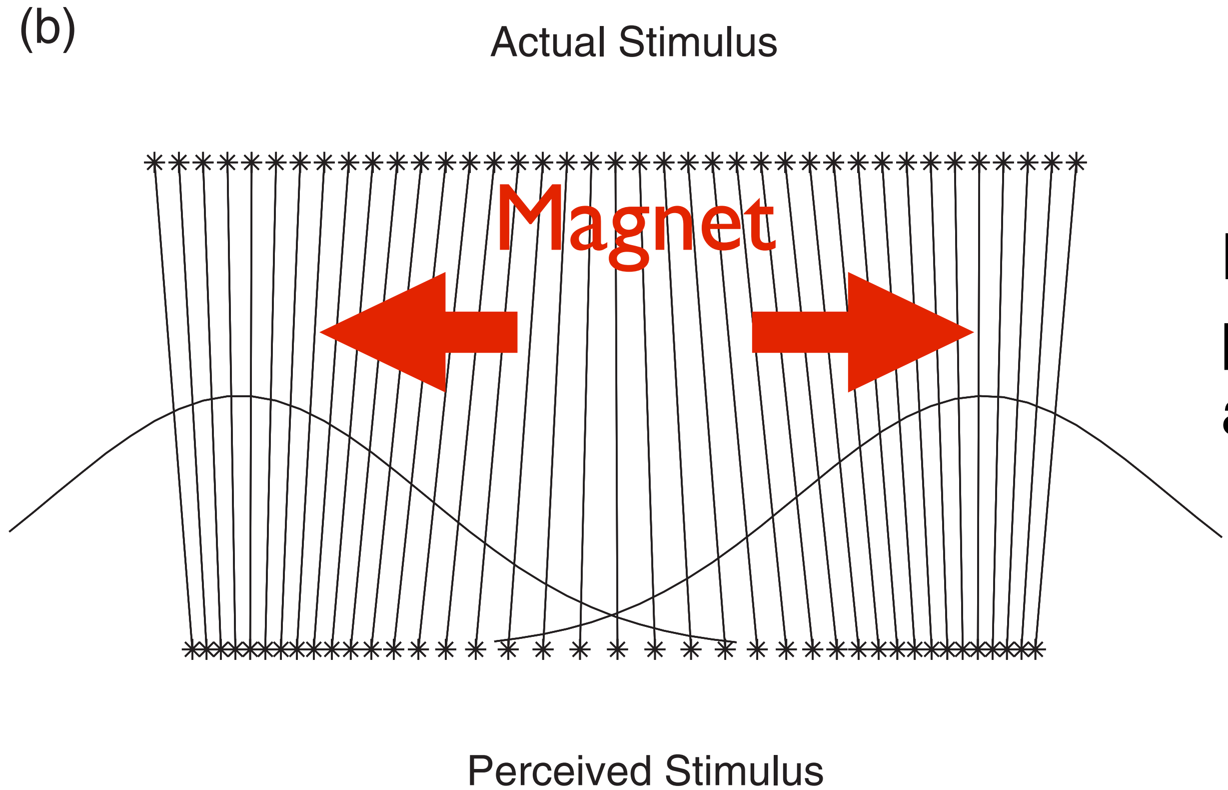
From Rob Goldstone (1994)

Categorical perception: A link between categorization and discrimination



Categorical perception is very closely related to the “perceptual magnet effect” in speech studied by Pat Kuhl et al.

(b)



Let’s try to understand the perceptual magnet effect through a Bayesian model

Psychological Review
2009, Vol. 116, No. 4, 752–782

0033-295X/09/$12.00 DOI: 10.1037/a0017196

The Influence of Categories on Perception: Explaining the Perceptual Magnet Effect as Optimal Statistical Inference

Naomi H. Feldman
Brown University

Thomas L. Griffiths
University of California, Berkeley

James L. Morgan
Brown University

A variety of studies have demonstrated that organizing stimuli into categories can affect the way the stimuli are perceived. We explore the influence of categories on perception through one such phenomenon, the perceptual magnet effect, in which discriminability between vowels is reduced near prototypical vowel sounds. We present a Bayesian model to explain why this reduced discriminability might occur: It arises as a consequence of optimally solving the statistical problem of perception in noise. In the optimal solution to this problem, listeners' perception is biased toward phonetic category means because they use knowledge of these categories to guide their inferences about speakers' target productions. Simulations show that model predictions closely correspond to previously published human data, and novel experimental results provide evidence for the predicted link between perceptual warping and noise. The model unifies several previous accounts of the perceptual magnet effect and provides a framework for exploring categorical effects in other domains.

Keywords: perceptual magnet effect, categorical perception, speech perception, Bayesian inference, rational analysis

The influence of categories on perception is well known in domains ranging from speech sounds to artificial categories of objects. Liberman, Harris, Hoffman, and Griffith (1957) first described categorical perception of speech sounds, noting that listeners' perception conforms to relatively sharp identification boundaries between categories of stop consonants and that whereas between-category discrimination of these sounds is nearly perfect, within-category discrimination is little better than chance. Similar patterns have been observed in the perception of colors (Davidoff, Davies, & Roberson, 1999), facial expressions (Etcoff & Magee, 1992), and familiar faces (Beale & Keil, 1995), as well as the representation of objects belonging to artificial categories that are learned over the course of an experiment (Goldstone, 1994; Goldstone, Lippa, & Shiffrin, 2001). All of these categorical effects are characterized by better discrimination of between-category contrasts than within-category contrasts, although the magnitude of the effect varies between domains.

In this article, we develop a computational model of the influence of categories on perception through a detailed investigation of one such phenomenon, the *perceptual magnet effect* (Kuhl, 1991), which has been described primarily in vowels. The perceptual magnet effect involves reduced discriminability of speech sounds near phonetic category prototypes. For several reasons, speech sounds, particularly vowels, provide an excellent starting point for assessing a model of the influence of categories on perception. Vowels are naturally occurring, highly familiar stimuli that all listeners have categorized. As discussed later, a precise two-

Naomi H. Feldman and James L. Morgan, Department of Cognitive and Linguistic Sciences, Brown University; Thomas L. Griffiths, Department of Psychology, University of California, Berkeley.

Bayesian model of speech perception

speaker

noise

listener

T

S

P(TlS)

perception

noise

speech sound from speaker

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)}$$

The speaker makes an intended sound production T.
Noise in the air perturbs T into S.
The listener calculates the posterior P(TlS)

Bayesian model of speech perception

The speaker makes a sound production T.
Noise in the air perturbs T into S.

Prior

$$T|c \sim N(\mu_c, \sigma_c^2)$$

Likelihood (normal distribution)

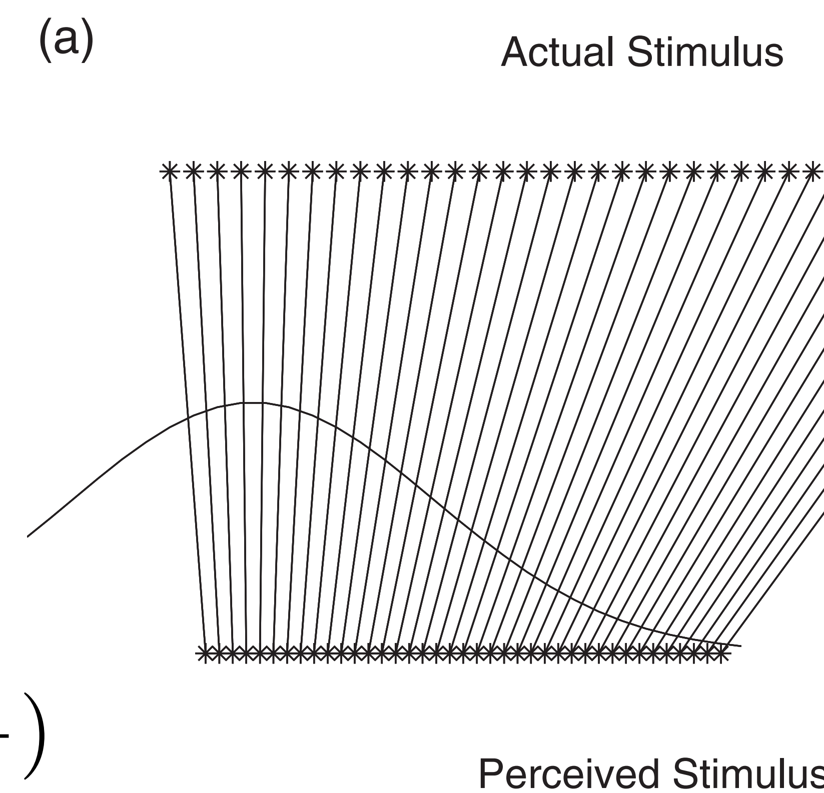
$$S|T \sim N(T, \sigma_S^2)$$

Posterior

$$P(T|S,c) = \frac{P(S|T)P(T|c)}{P(S|c)}$$

$$P(T|S,c) = N(\frac{\sigma_c^2 S + \sigma_S^2 \mu_c}{\sigma_c^2 + \sigma_S^2}, \frac{\sigma_c^2 \sigma_S^2}{\sigma_c^2 + \sigma_S^2})$$

If the stimulus is noisy, pull your perception towards the category you think it comes from.



Posterior is Gaussian, where the mean is a **weighted average** between the **actual stimulus *S* and the prior mean μ_c.**

- If the noise is high (high σ_S), rely more on the prior category mean
- If the category is highly variable (high σ_c), rely more on the actual stimulus S

Key technical concept: Conjugate priors

When prior and posterior are in the same family, then we have a **conjugate prior** for the likelihood function.

prior	likelihood	posterior	common use case
Normal	Normal (unknown mean, known variance)	Normal	estimating mean of a continuous sample
Beta	Binomial	Beta	estimating fairness of a coin based on counts
Dirichelt	Multinomial	Dirichelt	estimating weights on k-sided dice based on counts
…			

This makes it very easy to compute posterior distributions, as it can be done in closed form with standard formulas.

https://en.wikipedia.org/wiki/Conjugate_prior

Bayesian model of speech perception: Multiple categories

The speaker makes a sound production T.
Noise in the air perturbs T into S.

Step 1) Bayesian classification of the speech sound in category *c*

$$p(c|S) = \frac{p(S|c)p(c)}{\sum_c p(S|c)p(c)}$$

$$p(S|c) = \int p(S|T)p(T|c)\ dT$$

(this term is another Gaussian distribution)

Step 2) Compute reconstruction of *T as weighted mixture of posteriors*

$$p(T|S) = \Sigma_c p(T|S,c)p(c|S)$$

(mixture of Gaussians)

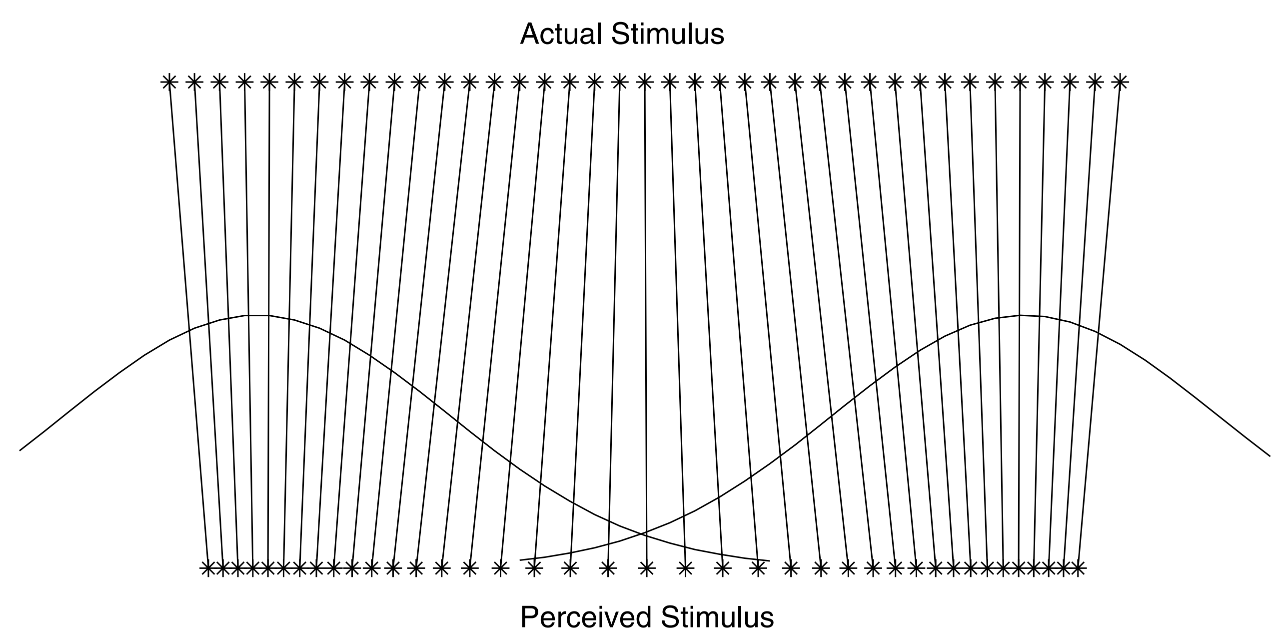
$$P(T|S,c) = N\left(\frac{\sigma_c^2 S + \sigma_S^2 \mu_c}{\sigma_c^2 + \sigma_S^2}, \frac{\sigma_c^2 \sigma_S^2}{\sigma_c^2 + \sigma_S^2}\right)$$

(term is posterior from previous slide with known category)

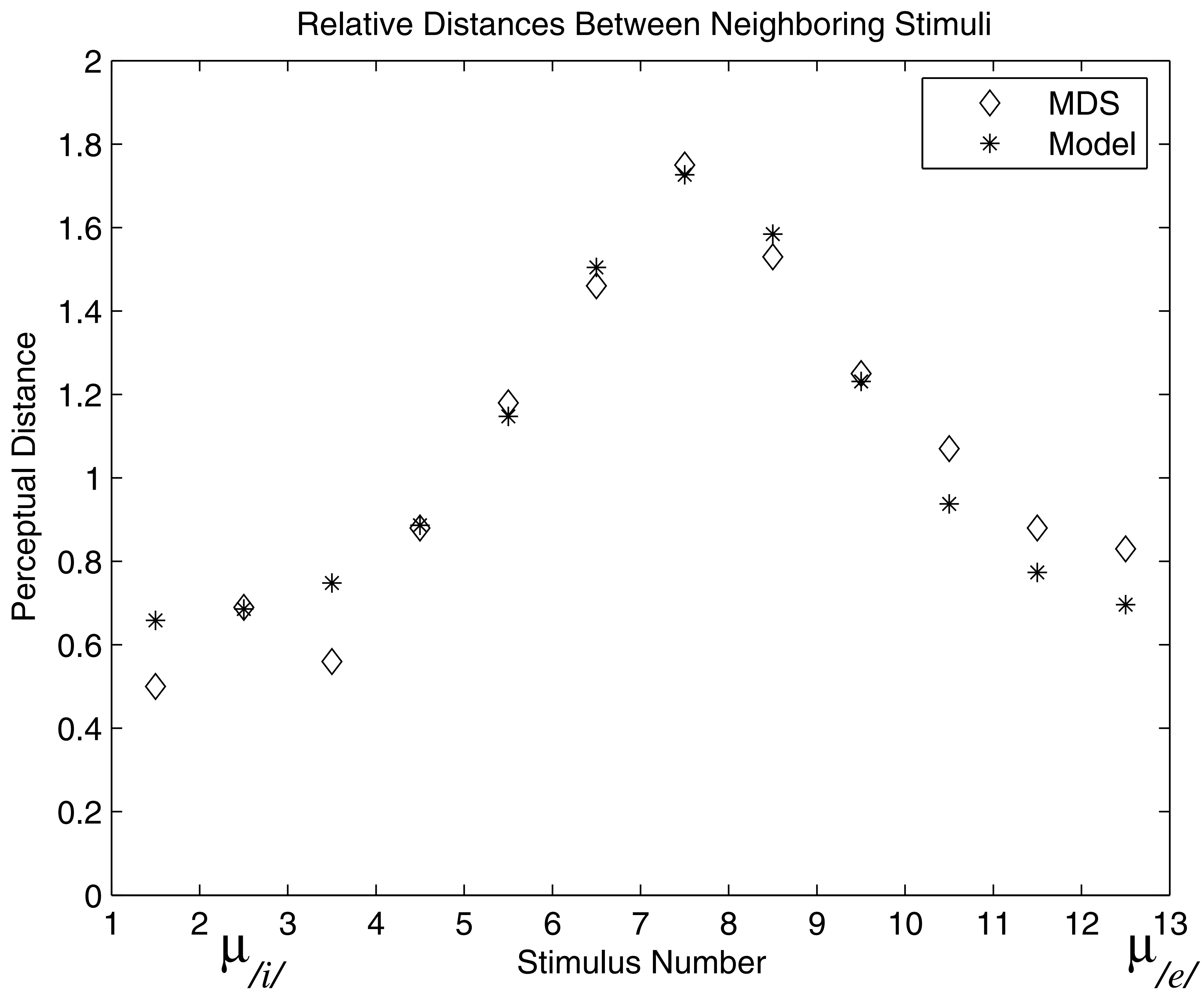
Posterior mean (expected value *E* [ ])

$$E[T|S] = \frac{\sigma_c^2}{\sigma_c^2+\sigma_S^2} S + \frac{\sigma_S^2}{\sigma_c^2+\sigma_S^2} \Sigma_c p(c|S)\mu_c$$

- If the noise is high (high σ_S), rely more on the category means
- If the category is highly variable (high σ_c), rely more on the actual stimulus *S*



Comparing the Bayesian model to perceptual data



participant data
Bayesian model

Computing perceptual distance

- Participant discrimination judgments converted to perceptual distance using multi-dimensional scaling (MDS)
- Bayesian model predictions computed as $E[T|S]$

Conclusions from the Bayesian models of the perceptual magnet effect

- Categories influence perception in a range of domains: speech, color, faces, etc...

Although it’s clear that categories influence perception, it’s not clear WHY they should

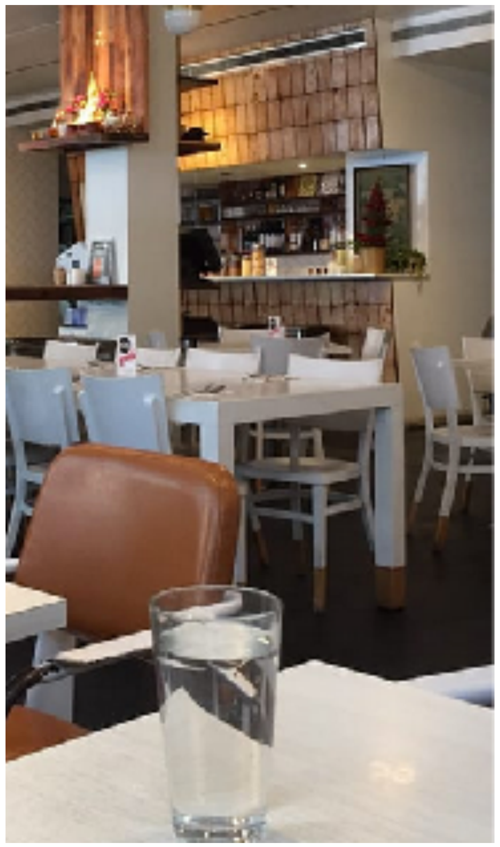
- There are many other models of categorical perception and perceptual magnet effect, but they don’t really answer the “why” question.

The Bayesian model suggests why perception should have this characteristic: It’s a rational adaption for perceiving/reconstructing stimuli under noise.

Implications for understanding behavioral data: User ratings

Actual object

T

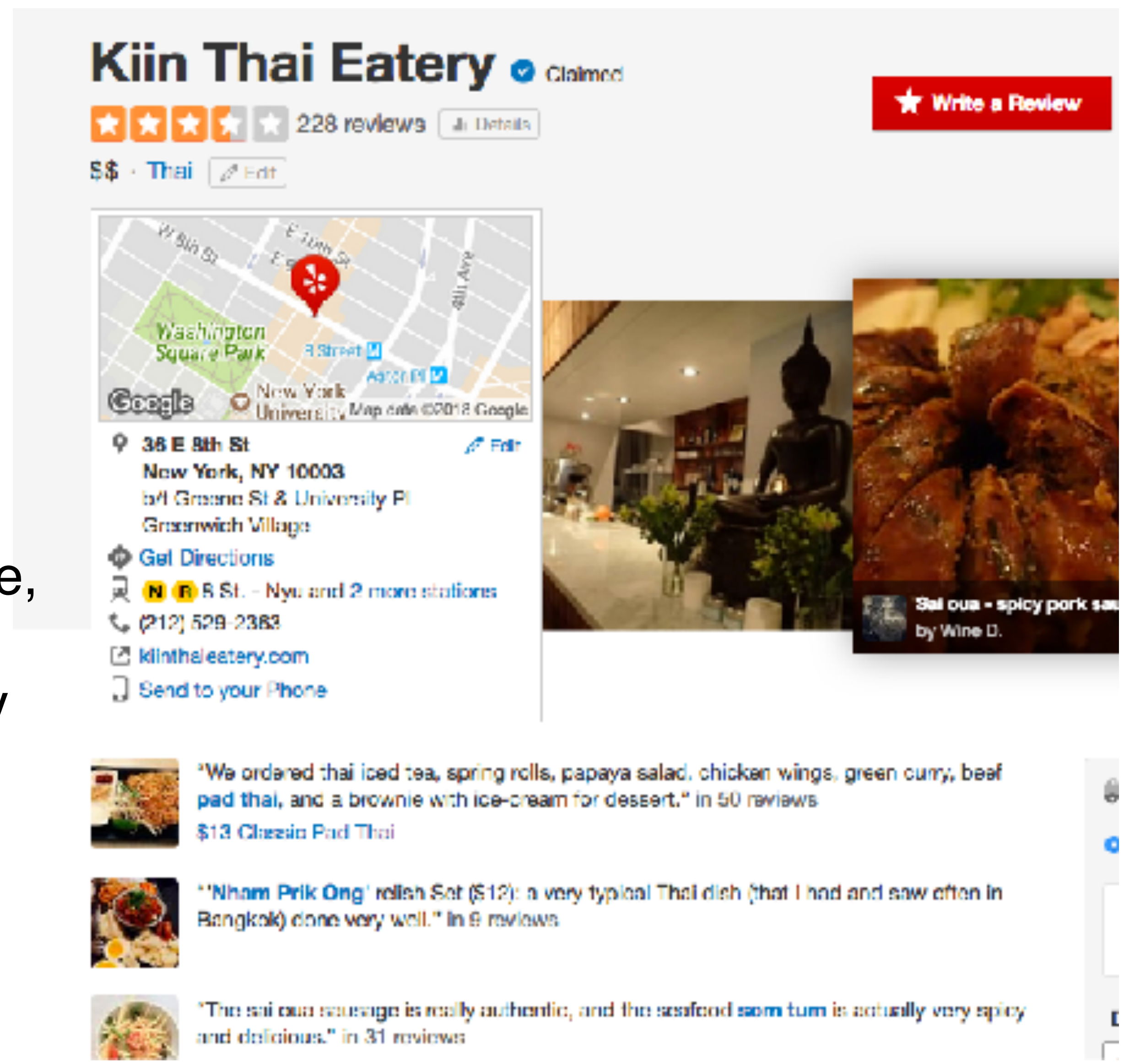


noise
(in perception,
memory, etc.)

S

P(T|S,c)

Perceived experience may **warped by the category c**



Price, quality, and many other attribute ratings (attire, good for groups, etc.) maybe warped by category knowledge (Thai restaurant), especially if ratings are entered with a delay.

More business info

Takes Reservations **Yes**
Delivery **Yes**
Take-out **Yes**
Accepts Credit Cards **Yes**
Accepts Apple Pay **No**
Good For **Lunch, Dinner**
Parking **Street**
Bike Parking **Yes**
Good for Kids **Yes**
Good for Groups **Yes**
Attire **Casual**
Ambience **Casual**
Noise Level **Average**
Alcohol **Full Bar**
Outdoor Seating **No**
Wi-Fi **Free**
Has TV **No**
Waiter Service **Yes**
Caters **No**
Gender Neutral Restrooms **Yes**

Predicting the future with Bayesian inference

PSYCHOLOGICAL SCIENCE

Research Article

Optimal Predictions in Everyday Cognition

Thomas L. Griffiths[1] and Joshua B. Tenenbaum[2]

[1]Department of Cognitive and Linguistic Sciences, Brown University, and [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

ABSTRACT—*Human perception and memory are often explained as optimal statistical inferences that are informed by accurate prior probabilities. In contrast, cognitive judgments are usually viewed as following error-prone heuristics that are insensitive to priors. We examined the optimality of human cognition in a more realistic context than typical laboratory studies, asking people to make predictions about the duration or extent of everyday phenomena such as human life spans and the box-office take of movies. Our results suggest that everyday cognitive judgments follow the same optimal statistical principles as perception and memory, and reveal a close correspondence between people's implicit probabilistic models and the statistics of the world.*

If you were assessing the prospects of a 60-year-old man, how

Perry, Super, & Gallogly, 2001; Huber, Shiffrin, Lyle, & Ruys, 2001; Knill & Richards, 1996; Körding & Wolpert, 2004; Shiffrin & Steyvers, 1997; Simoncelli & Olshausen, 2001; Weiss, Simoncelli, & Adelson, 2002). In contrast—perhaps as a result of the great attention garnered by the work of Kahneman, Tversky, and their colleagues (e.g., Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1974)—cognitive judgments under uncertainty are often characterized as the result of error-prone heuristics that are insensitive to prior probabilities. This view of cognition, based on laboratory studies, appears starkly at odds with the near-optimality of other human capacities, and with people's ability to make smart predictions from sparse data in the real world.

To evaluate how cognitive judgments compare with optimal statistical inferences in real-world settings, we asked people to predict the duration or extent of everyday phenomena such as human life spans and the gross of movies. We varied the phe-

Let’s make some predictions

You stopped by a friend’s apartment, and she has been watching a movie for 15 minutes. What would you predict for the length of the movie?

You stopped by a friend’s apartment, and she has been watching a movie for 75 minutes. What would you predict for the length of the movie?

Let’s make some predictions

You stopped by a friend’s apartment, and she has been watching a movie for 15 minutes. What would you predict for the length of the movie?

You stopped by a friend’s apartment, and she has been watching a movie for 75 minutes. What would you predict for the length of the movie?

A movie has grossed 15 million dollars at the box office, but you don’t know how long it’s been running. How much will it gross total?

A movie has grossed 75 million dollars at the box office, but you don’t know how long it’s been running. How much will it gross total?

Let’s make some predictions

You stopped by a friend’s apartment, and she has been watching a movie for 15 minutes. What would you predict for the length of the movie?

You stopped by a friend’s apartment, and she has been watching a movie for 75 minutes. What would you predict for the length of the movie?

A movie has grossed 15 million dollars at the box office, but you don’t know how long it’s been running. How much will it gross total?

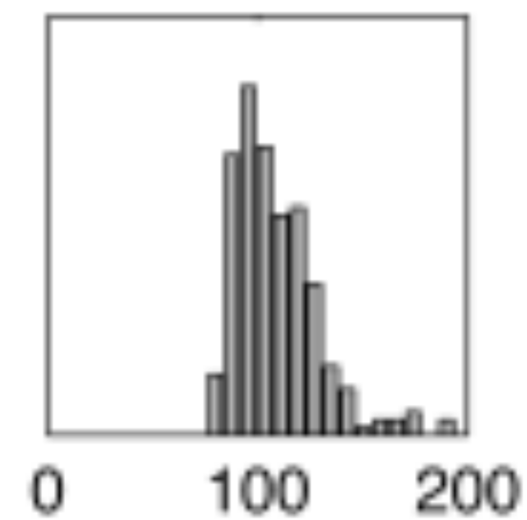
A movie has grossed 75 million dollars at the box office, but you don’t know how long it’s been running. How much will it gross total?

Let’s make some predictions

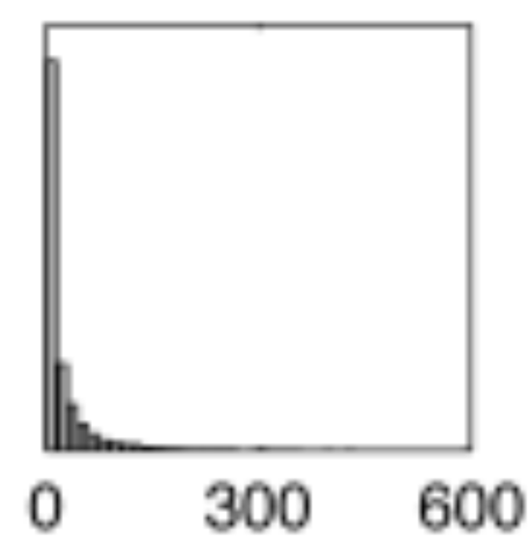
You stopped by a friend’s apartment, and she has been watching a movie for 15 minutes. What would you predict for the length of the movie?

You stopped by a friend’s apartment, and she has been watching a movie for 75 minutes. What would you predict for the length of the movie?

Movie runtimes



Movie grosses



A movie has grossed 15 million dollars at the box office, but you don’t know how long it’s been running. How much will it gross total?

A movie has grossed 75 million dollars at the box office, but you don’t know how long it’s been running. How much will it gross total?

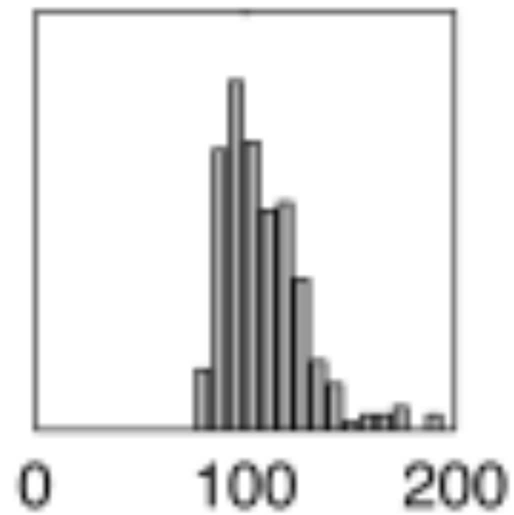
Simple Bayesian model of predicting the future

A movie has grossed 15 million dollars at the box office, but you don't know how long it's been running. How much will it gross total?

t_{total} : the total length of time you are estimating

t : the current time you are given (current runtime of movie, current gross)

Movie runtimes

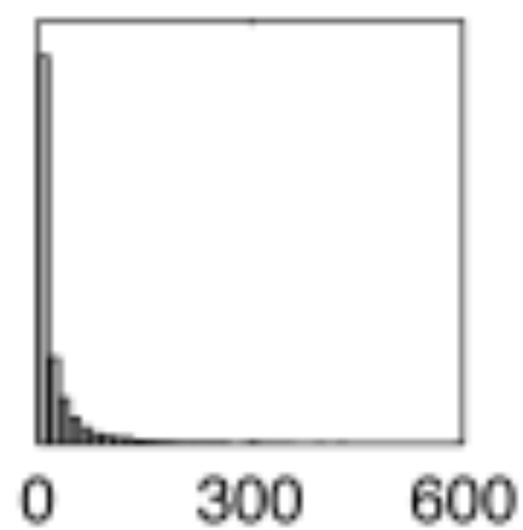


Bayesian estimation problem

posterior, likelihood, prior

$$P(t_{total}|t) = \frac{P(t|t_{total})P(t_{total})}{P(t)}$$

Movie grosses



Likelihood

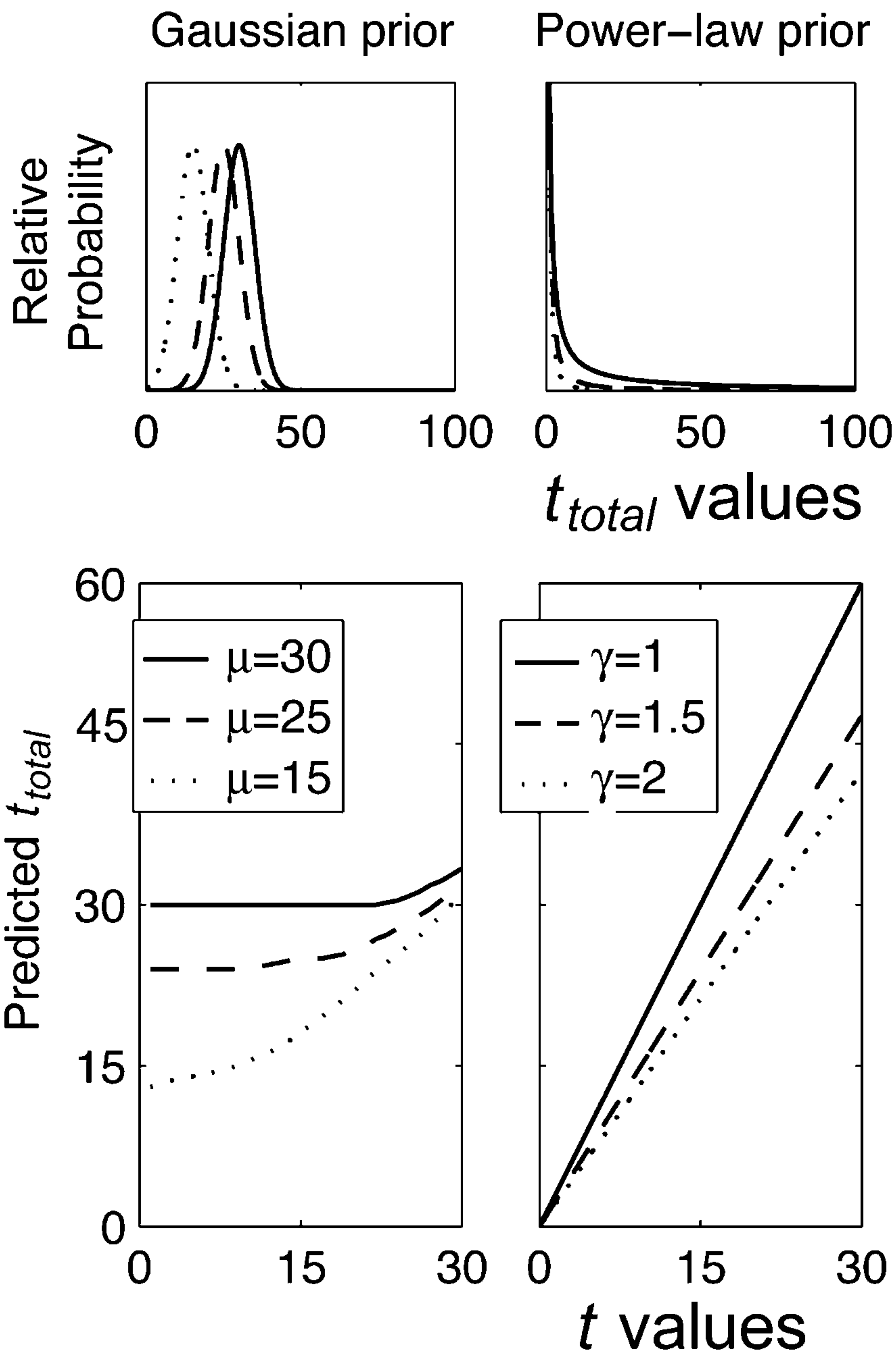
$$P(t|t_{total}) = 1/t_{total}$$

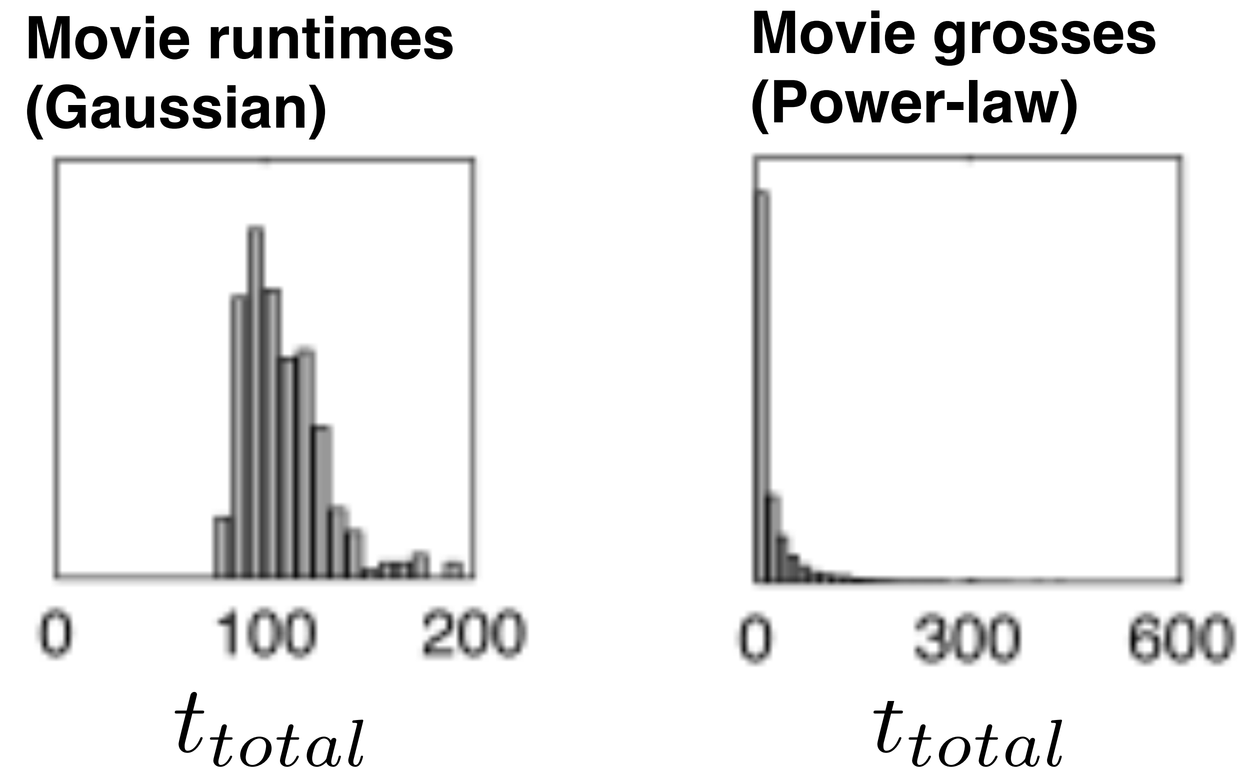
Assumption: you are equally likely to encounter a movie/person/etc. at any time

Prior

$P(t_{total})$ is estimated from real world statistics

Different priors have qualitatively different predictions





Gaussian prior

$$P(t_{total}) \propto \exp(-\frac{1}{2\sigma^2}(t_{total} - \mu)^2)$$

Power-law prior

$$P(t_{total}) \propto t_{total}^{-\gamma}$$

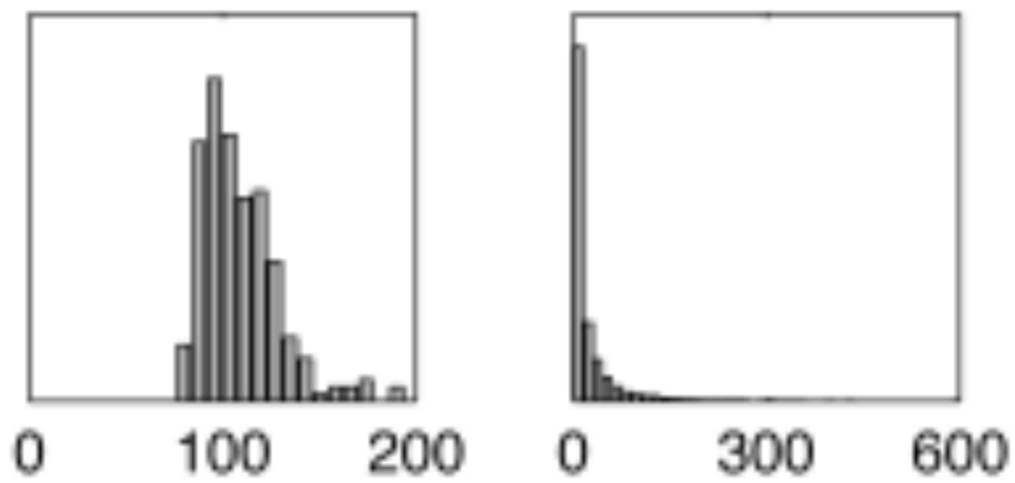
Posterior

$$P(t_{total}|t) = \frac{P(t|t_{total})P(t_{total})}{P(t)}$$

Different priors have qualitatively different predictions

Movie runtimes (Gaussian) **Movie grosses (Power-law)**

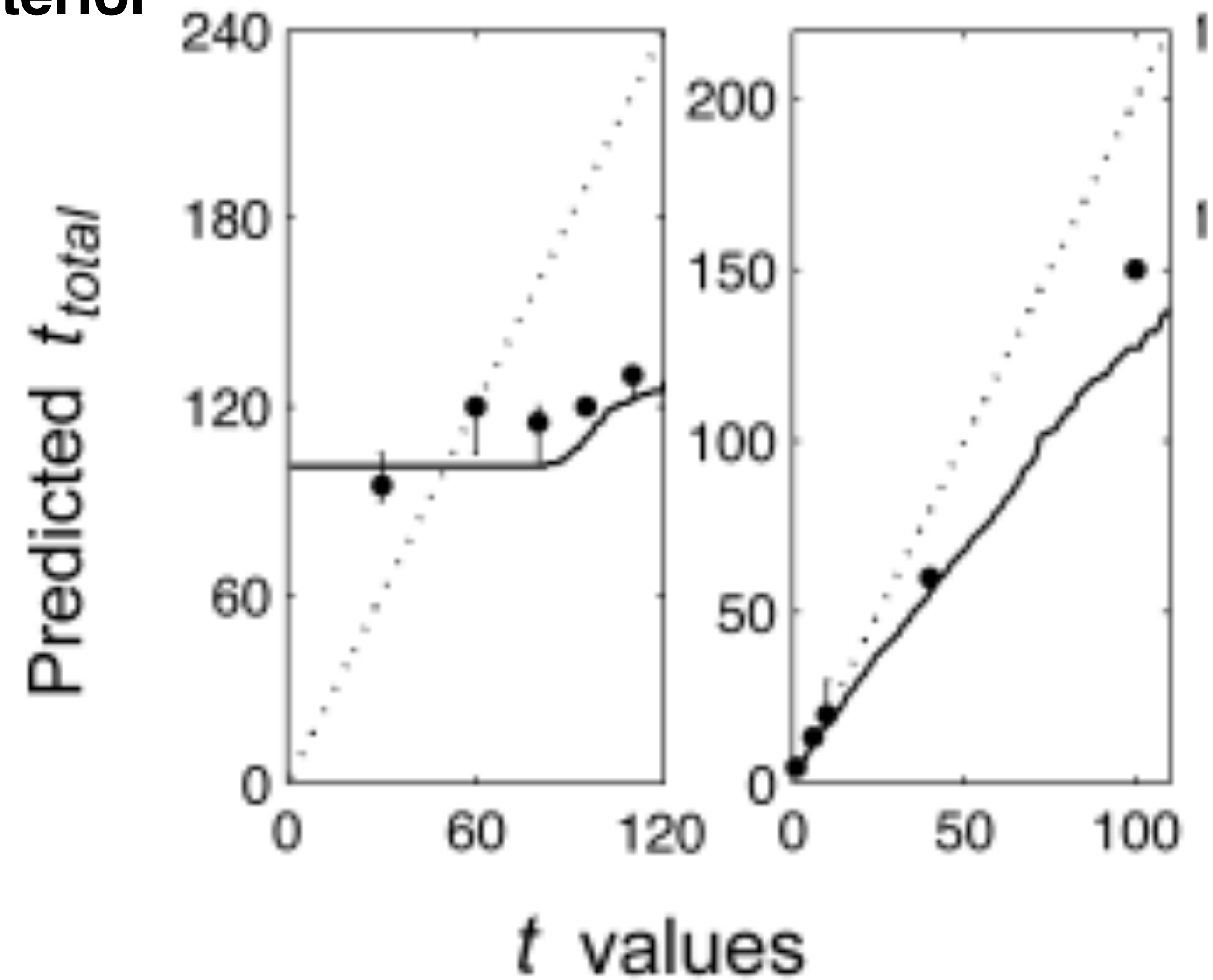
Prior



Black dots are median prediction of participants

Solid lines are Bayesian predictions

Posterior



Posterior

$$P(t_{total}|t) = \frac{P(t|t_{total})P(t_{total})}{P(t)}$$

Optimal predictions

For movie runtimes, predict the mean unless the runtime has already exceeded it.

For movie grosses, multiply the current gross by roughly 1.5

Patterns of prediction across a range of domains

Poem lengths: If your friend read you her favorite line of poetry, and told you it was line 5 of a poem, what would you predict for the total length of the poem?

Life spans: Insurance agencies employ actuaries to make predictions about people's life spans—the age at which they will die—based upon demographic information. If you were assessing an insurance case for an 18-year-old man, what would you predict for his life span?

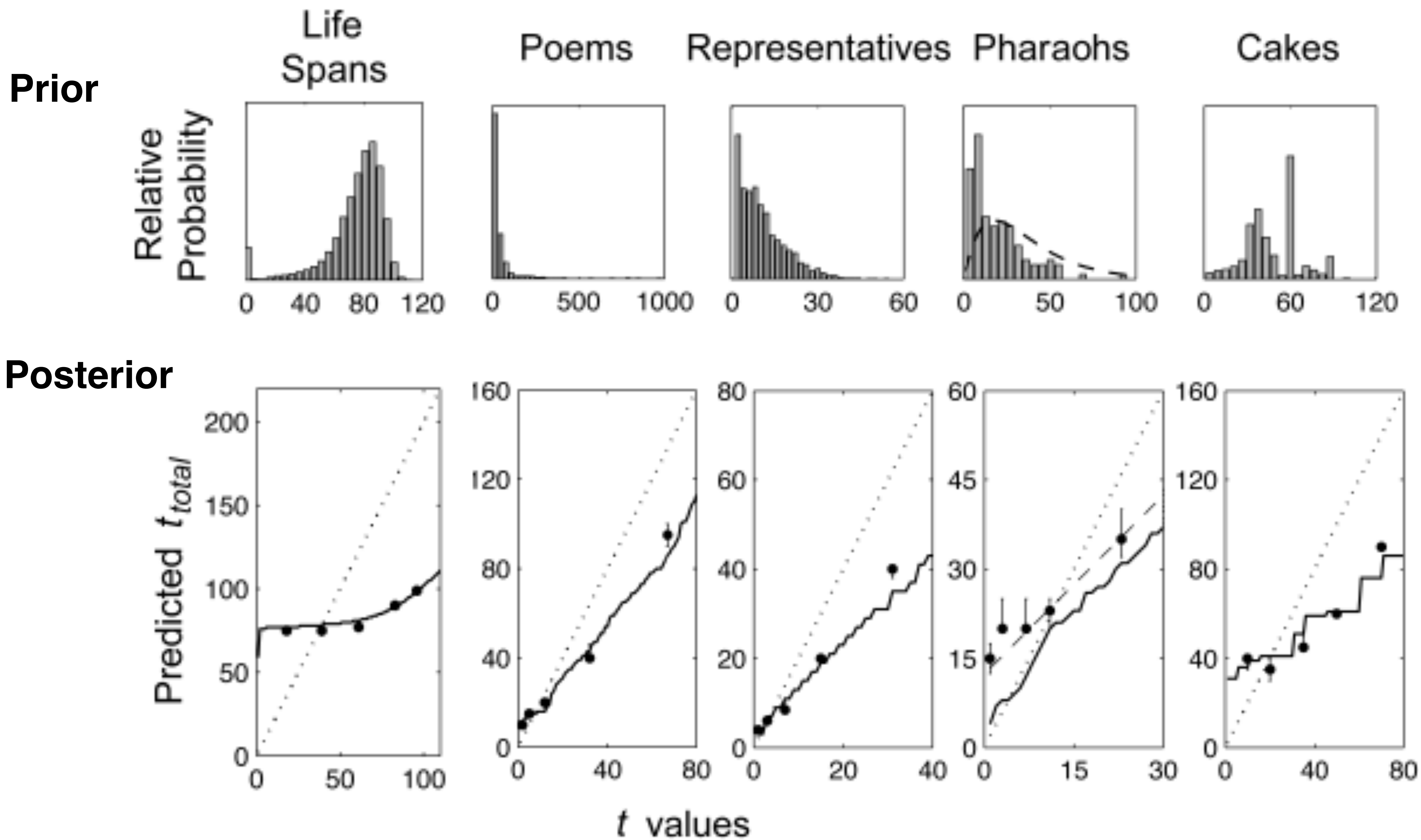
Baking times for cakes: Imagine you are in somebody's kitchen and notice that a cake is in the oven. The timer shows that it has been baking for 35 minutes. What would you predict for the total amount of time the cake needs to bake?

Waiting times: If you were calling a telephone box office to book tickets and had been on hold for 3 minutes, what would you predict for the total time you would be on hold?

Reigns of pharaohs: If you opened a book about the history of ancient Egypt to a page listing the reigns of the pharaohs, and noticed that at 4000 BC a particular pharaoh had been ruling for 11 years, what would you predict for the total duration of his reign?

Terms of U.S. representatives: If you heard a member of the House of Representatives had served for 15 years, what would you predict his total term in the House would be?

Patterns of prediction across a range of domains



Black dots are median prediction of participants

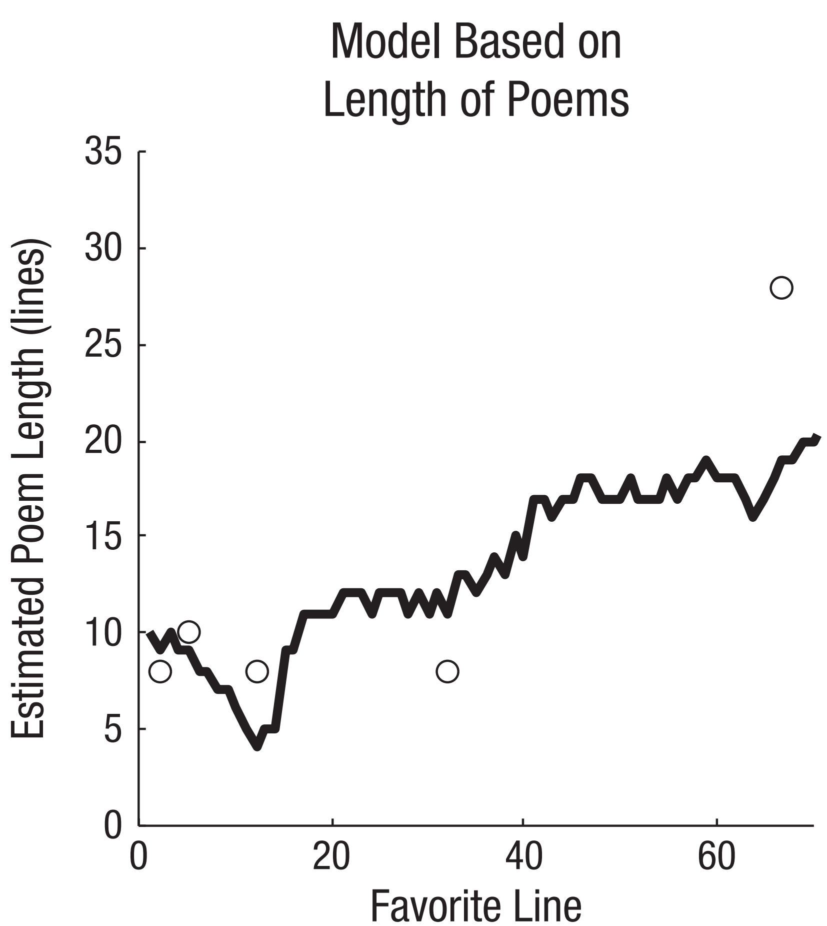
Solid lines are Bayesian predictions

Critique of “Optimal predictions in everyday cognition”

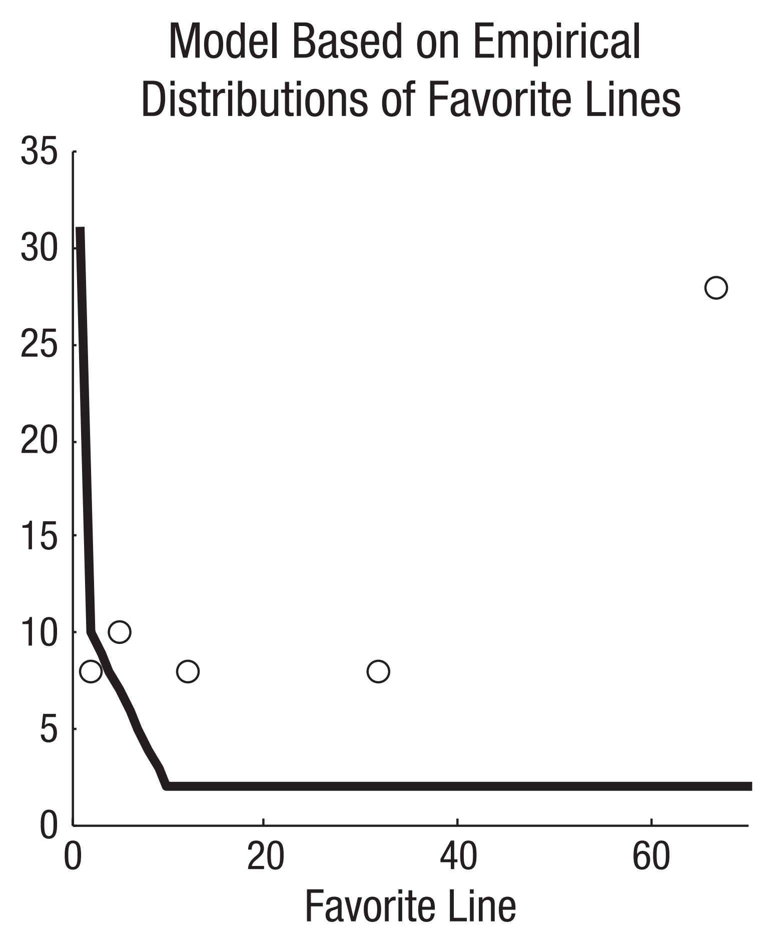
Marcus, G. and Davis, E. (2013) “How robust are probabilistic models of higher-level cognition?”

White dots are mean prediction of participants

Solid lines are Bayesian predictions

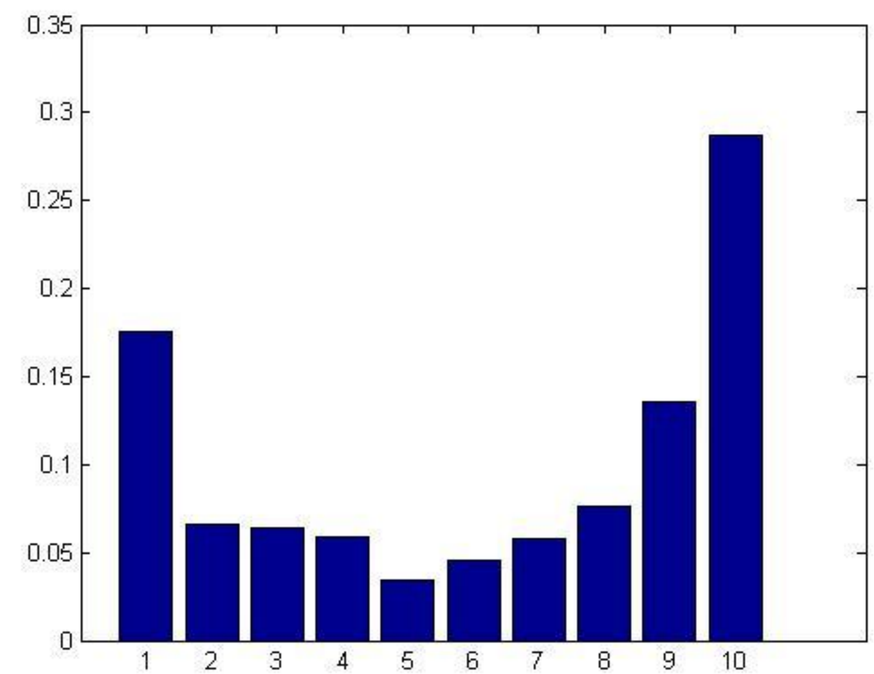


If Griffiths and Tenenbaum data is replotted to only show “additional length” on y-axis, the predictions are less impressive.



If distribution of favorite lines is taken into account, the predictions are way off.

Favorite lines are not uniformly distributed across poem.



Distribution of favorite lines per decile of poem.

Conclusions from optimal predictions in everyday cognition

- Critique for Marcus and Davis notwithstanding, there are surprisingly close fit between people's predictions and optimal Bayesian predictions.

- Implications

  1. People seem to accurately absorb the statistics of their environment for everyday quantities.

  2. In addition, people use these learned statics in accordance with Bayesian inference.

  3. The simplifying assumptions of "equal likelihood of encounter across timespan" could also be important, given Marcus and David critique.