Categories and Concepts - Spring 2019

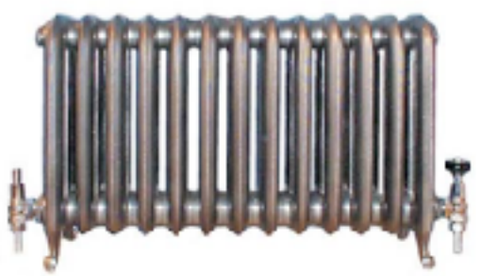
Computational models (part 2)

Brenden Lake

PSYCH-GA 2207

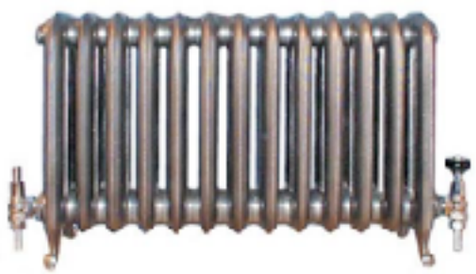
Concept learning from just a few positive examples

where are the others?



Concept learning from just a few positive examples

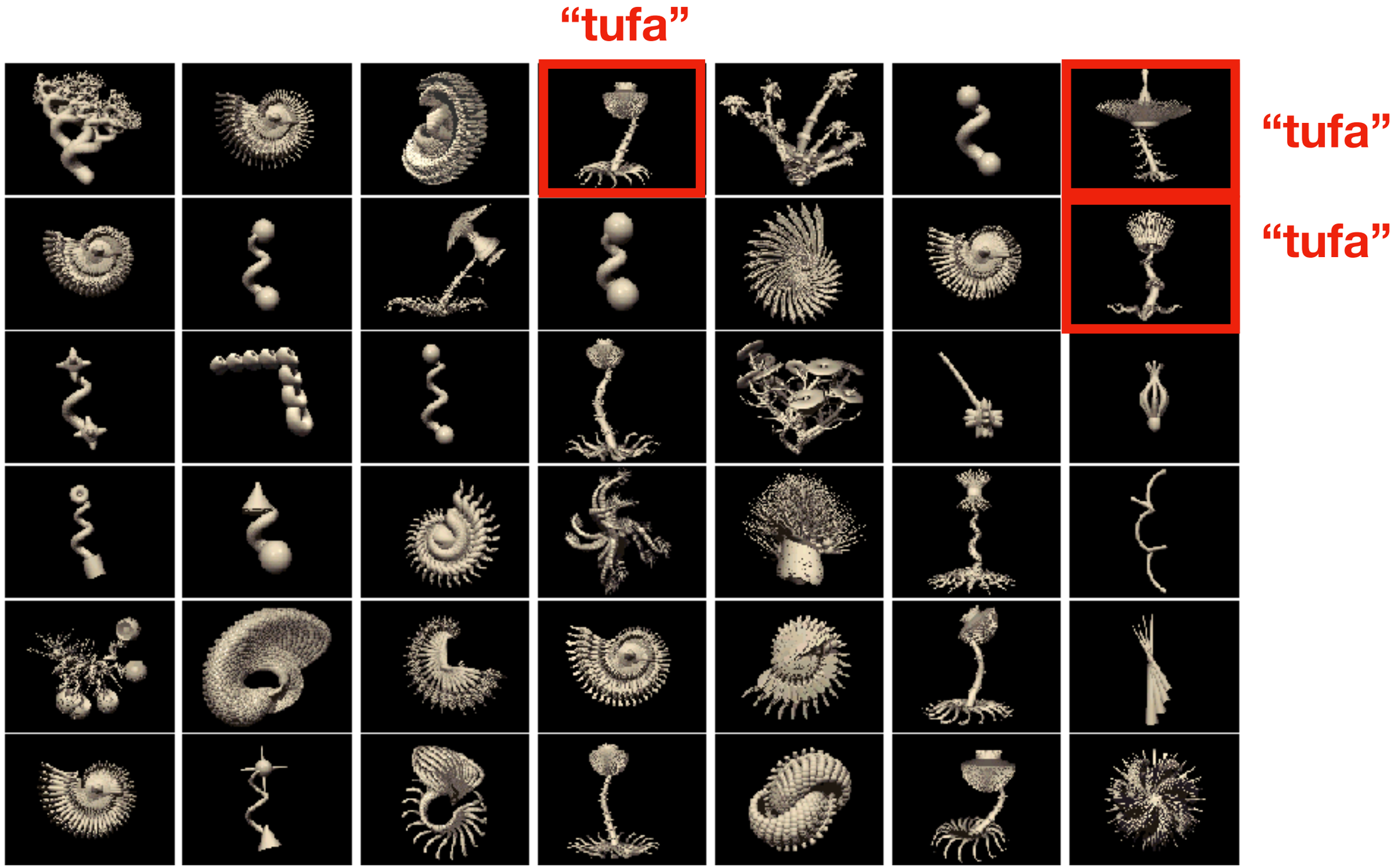
where are the others?



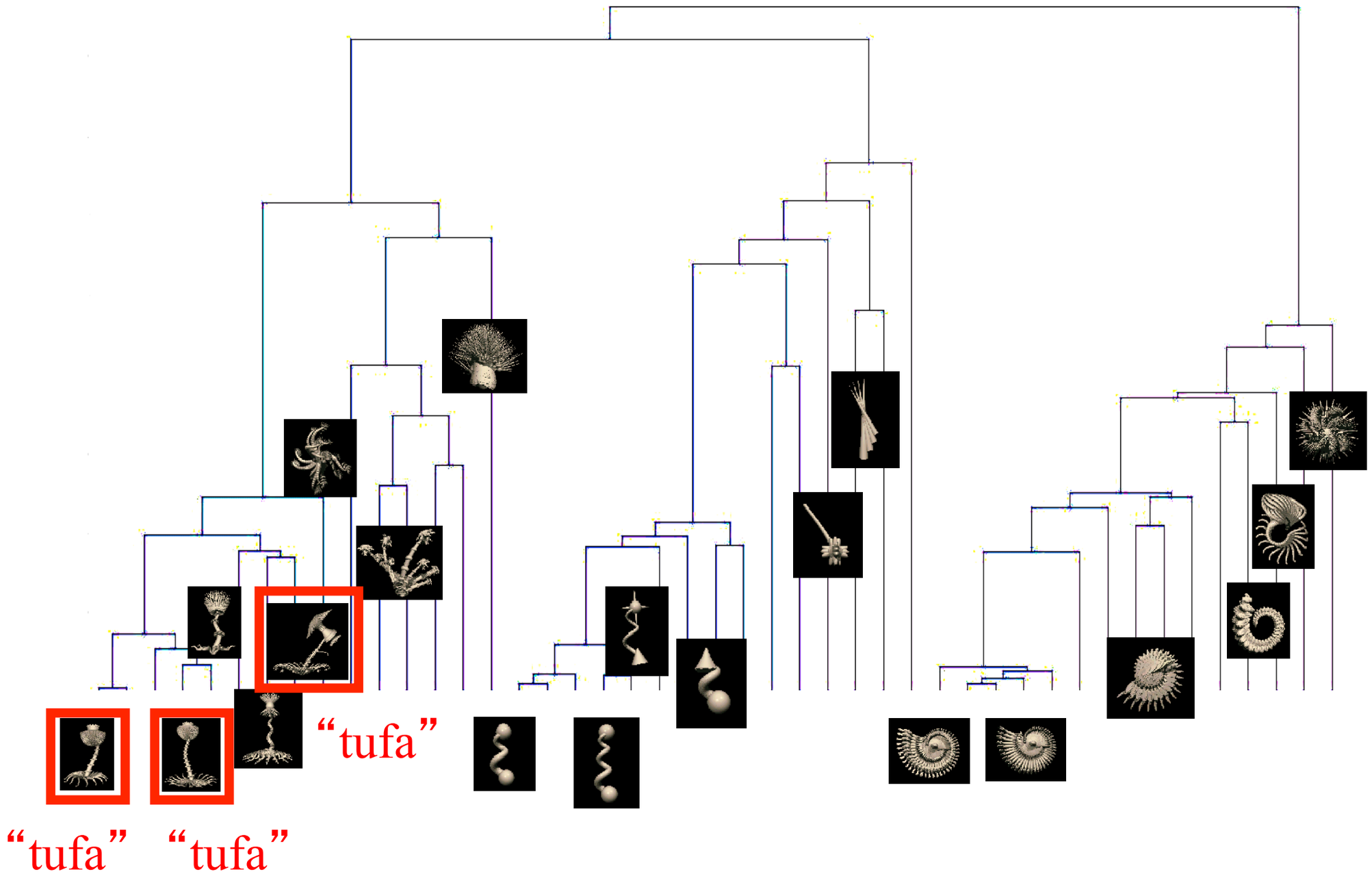
Concept learning from just a few positive examples

(Example from Josh Tenenbaum)

Concept learning from just a few positive examples



Concept learning from just a few positive examples



Psychological Review
2007, Vol. 114, No. 2, 245–272

0033-295X/07/$12.00 DOI: 10.1037/0033-295X.114.2.245

Word Learning as Bayesian Inference

Fei Xu
University of British Columbia

Joshua B. Tenenbaum
Massachusetts Institute of Technology

The authors present a Bayesian framework for understanding how adults and children learn the meanings of words. The theory explains how learners can generalize meaningfully from just one or a few positive examples of a novel word's referents, by making rational inductive inferences that integrate prior knowledge about plausible word meanings with the statistical structure of the observed examples. The theory addresses shortcomings of the two best known approaches to modeling word learning, based on deductive hypothesis elimination and associative learning. Three experiments with adults and children test the Bayesian account's predictions in the context of learning words for object categories at multiple levels of a taxonomic hierarchy. Results provide strong support for the Bayesian account over competing accounts, in terms of both quantitative model fits and the ability to explain important qualitative phenomena. Several extensions of the basic theory are discussed, illustrating the broader potential for Bayesian models of word learning.

Keywords: word learning, Bayesian inference, concepts, computational modeling

Learning even the simplest names for object categories presents a difficult induction problem (Quine, 1960). Consider a typical dilemma faced by a child learning English. Upon observing a competent adult speaker use the word *dog* in reference to Max, a particular Dalmatian running by, what can the child infer about the meaning of the word *dog*? The potential hypotheses appear endless. The word could refer to all (and only) dogs, all mammals, all animals, all Dalmatians, this individual Max, all dogs plus the Lone Ranger's horse, all dogs except Labradors, all spotted things, all running things, the front half of a dog, undetached dog parts, things that are dogs if first observed before next Monday but cats if first observed thereafter, and on and on. Yet despite this severe underdetermination, even 2- or 3-year-olds seem to be remarkably successful at learning the meanings of words from examples. In particular, children or adults can often infer the approximate extensions of words such as *dog* given only a few relevant examples of how the word can be used and no systematic evidence of how words are not to be used (Bloom, 2000; Carey, 1978; Markman, 1989; Regier, 1996). How do they do it?

Two broad classes of proposals for how word learning works have been dominant in the literature: *hypothesis elimination* and *associative learning*. Under the hypothesis elimination approach, the learner effectively considers a hypothesis space of possible concepts onto which words will map and (leaving aside for now the problem of homonyms and polysemy) assumes that each word maps onto exactly one of these concepts. The act of learning consists of eliminating incorrect hypotheses about word meaning

Review: Bayes' rule for updating beliefs in light of data

Data (D): John is coughing

Hypotheses:

$h_1 =$ John has a cold

$h_2 =$ John has emphysema

$h_3 =$ John has a stomach flu

"Bayes' rule"

posterior: $P(h_i|D)$ — likelihood: $P(D|h_i)$ — prior: $P(h_i)$

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

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

Bayesian concept learning

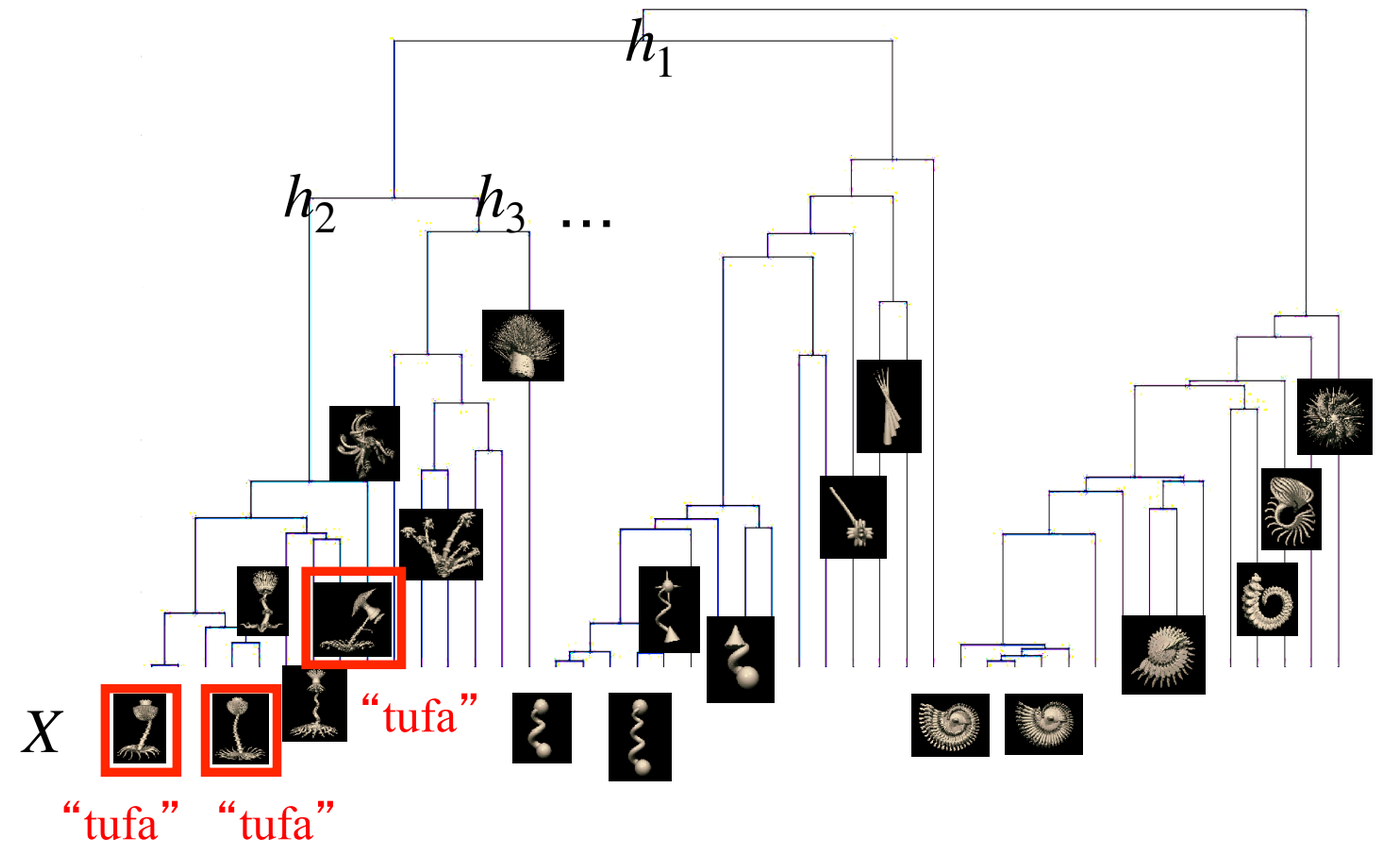
$h \in H$: **hypothesis about meaning of word (e.g., node in tree structure)**

X : **data (often just labels of positive examples)**

n : **number of examples**

Posterior over word meanings

$$p(h \mid X) = \frac{P(X \mid h)P(h)}{P(X)}$$



Likelihood

$$P(X \mid h) = [\frac{1}{\textbf{size}(h)}]^n$$

("the size principle", uniform random sample from appropriate set)

Prior

$P(h)$

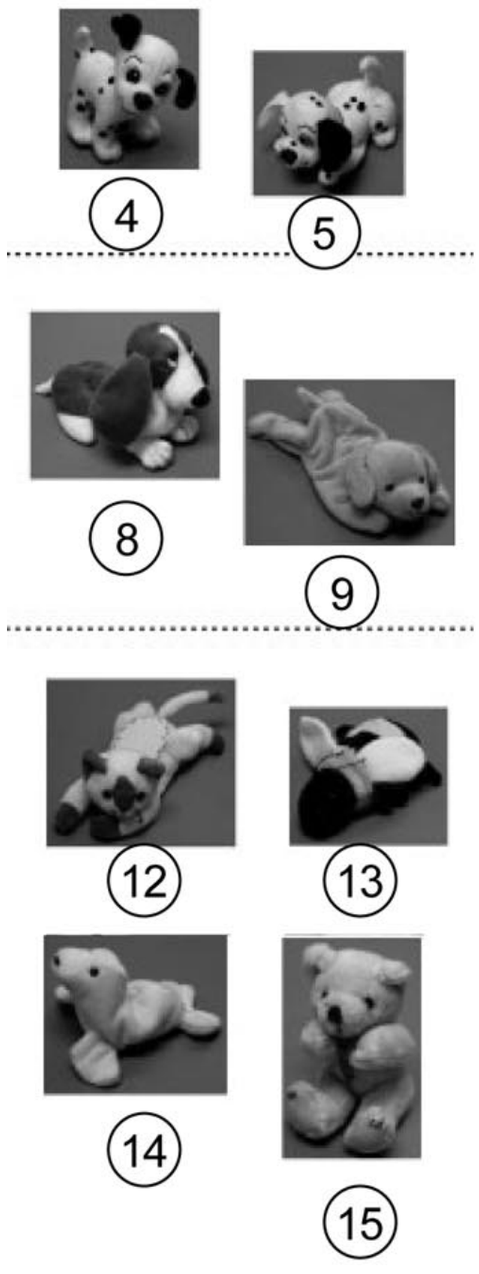
could be uniform over nodes tree,
or favor more distinctive nodes,
or favor nodes at a certain level….

Xu and Tenenbaum Exp 1: Adults

Helping Mr. Frog who speaks a different language pick out the objects he wants.

"Here is a fep"

Which others are feps?

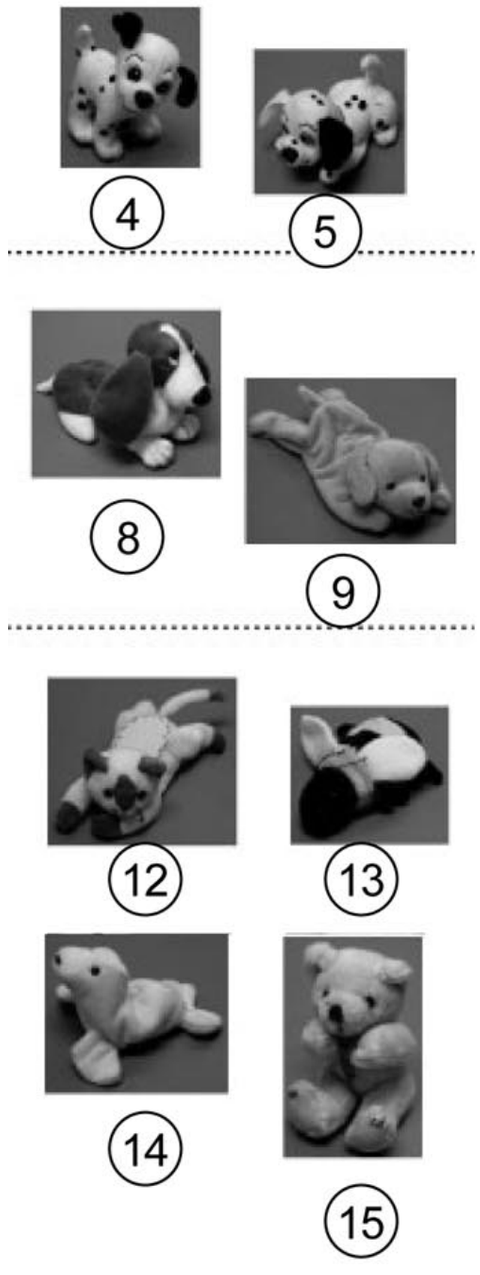


Xu and Tenenbaum Exp 1

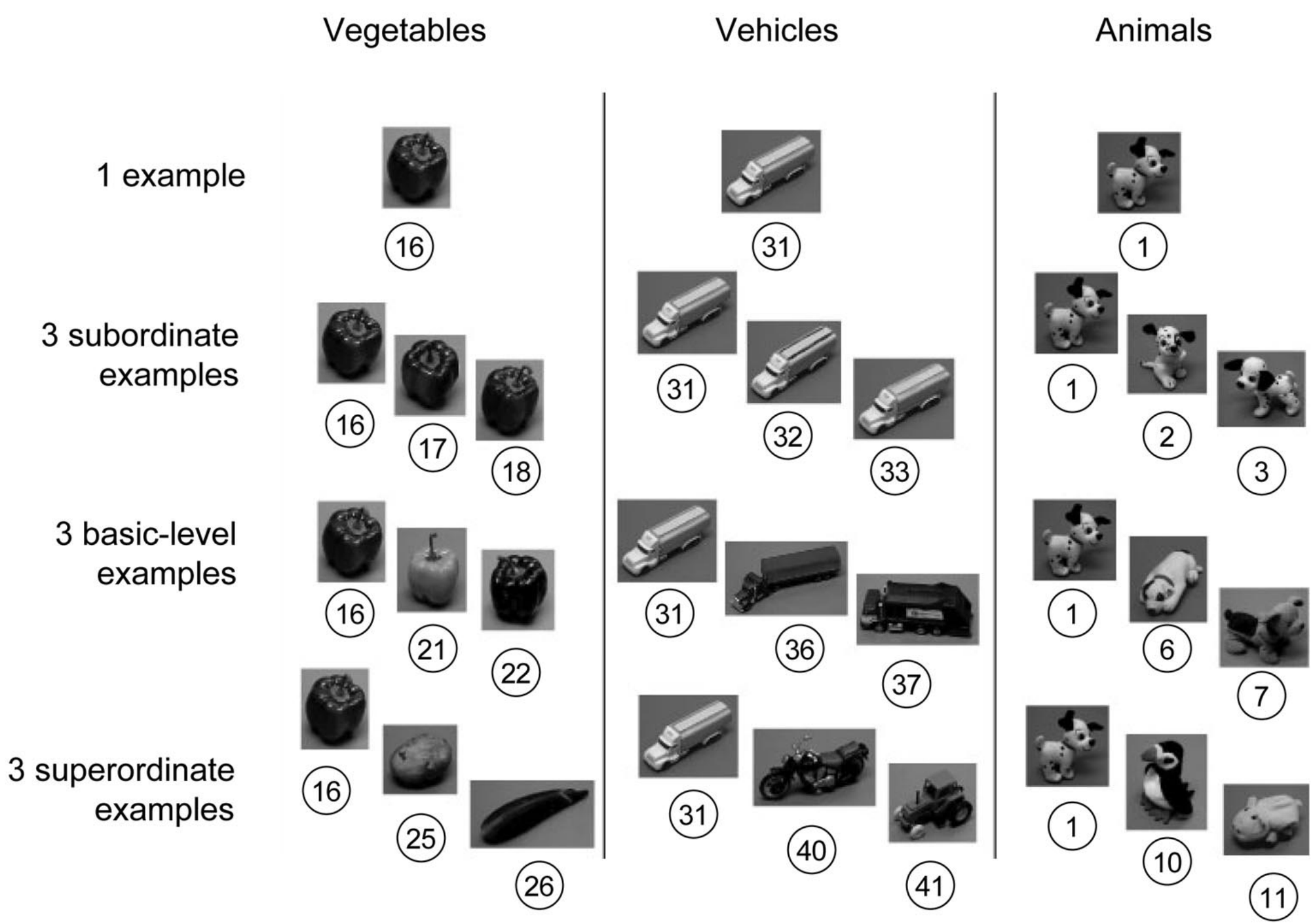
Helping Mr. Frog who speaks a different language pick out the objects he wants.

"Here are three feps"

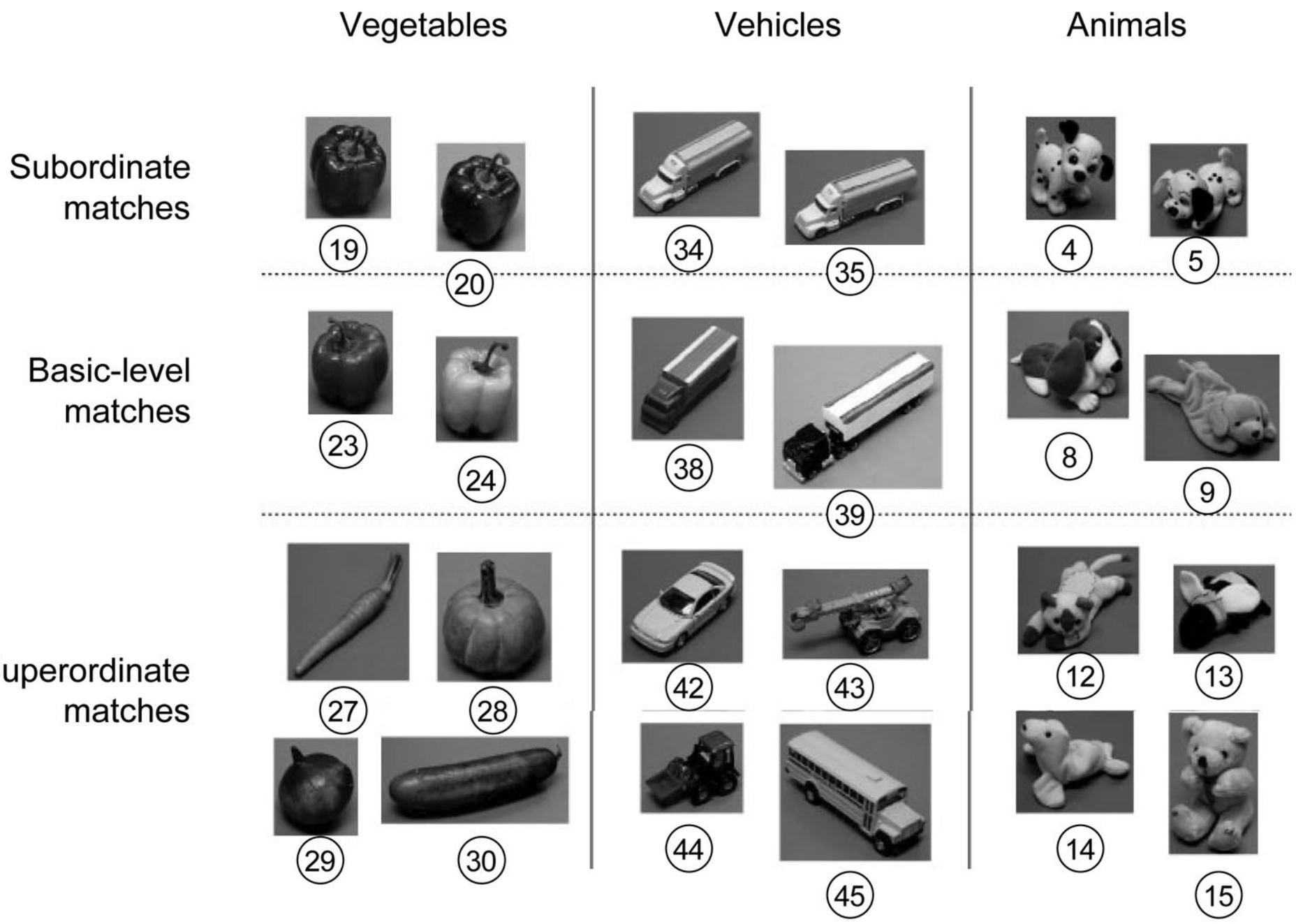
Which others are feps?



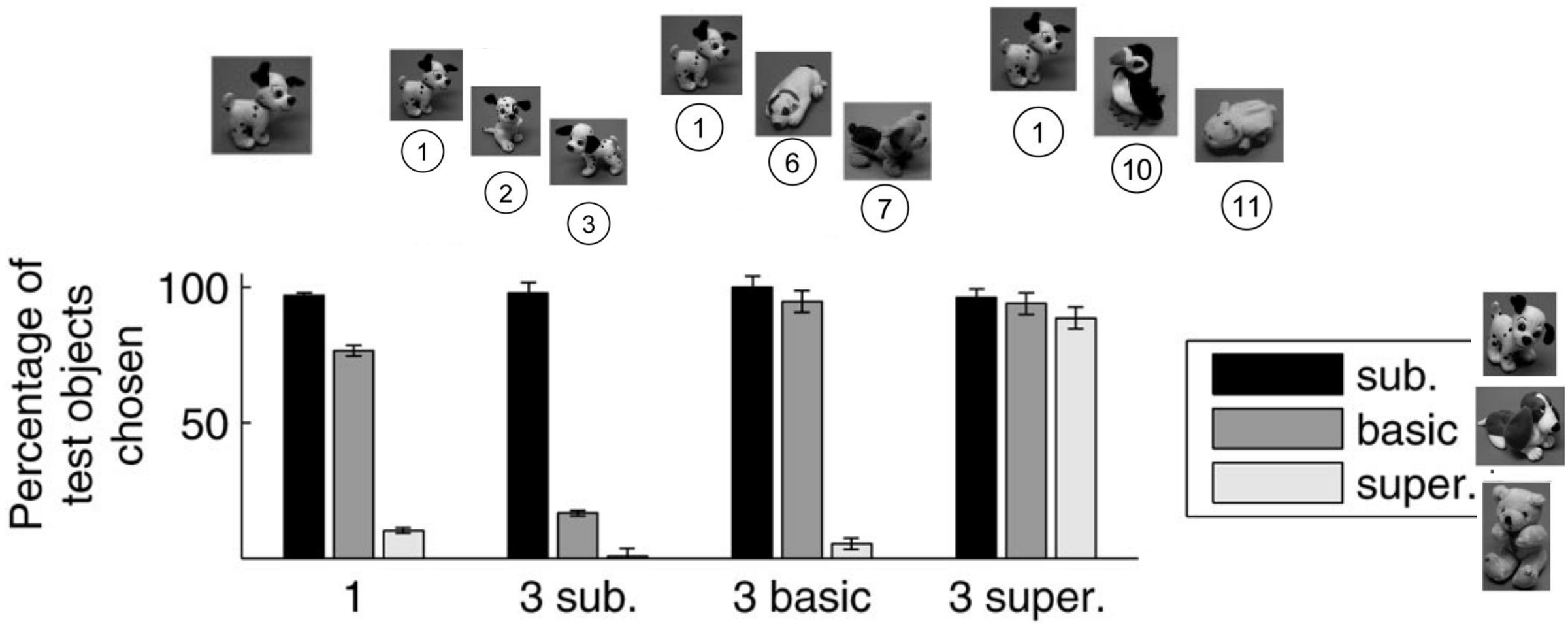
Xu and Tenenbaum Exp 1: Training set



Xu and Tenenbaum Exp 1: Test set



Adult results (Exp 1)



- Graded generalization with one subordinate example, and rule-like generalization with 3 examples
- Rapid learning from only sparse, positive examples

Xu and Tenenbaum Exp 3: Children

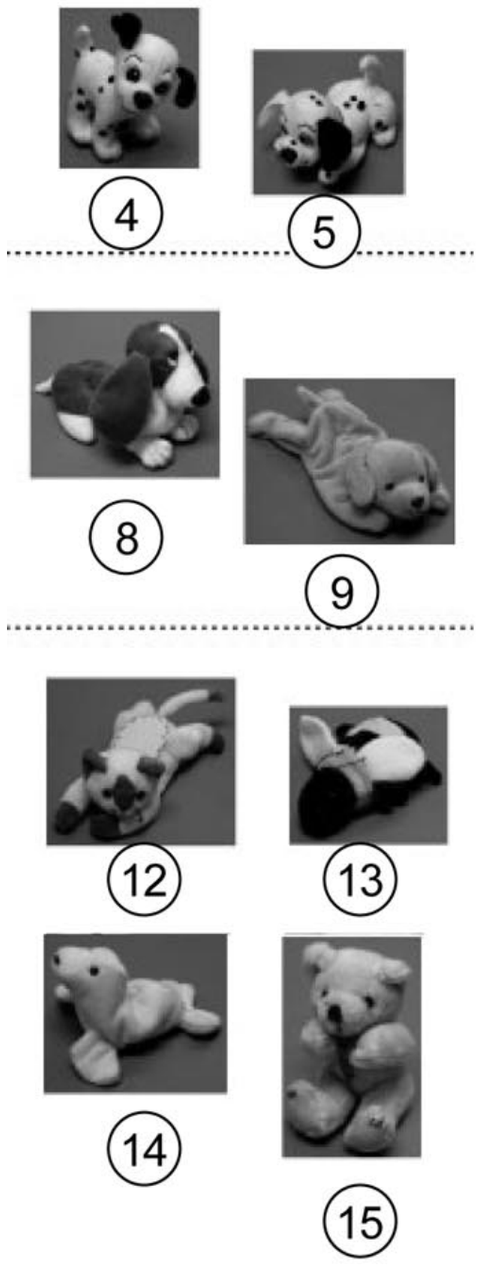
Helping Mr. Frog who speaks a different language pick out the objects he wants.

"Here is a fep"

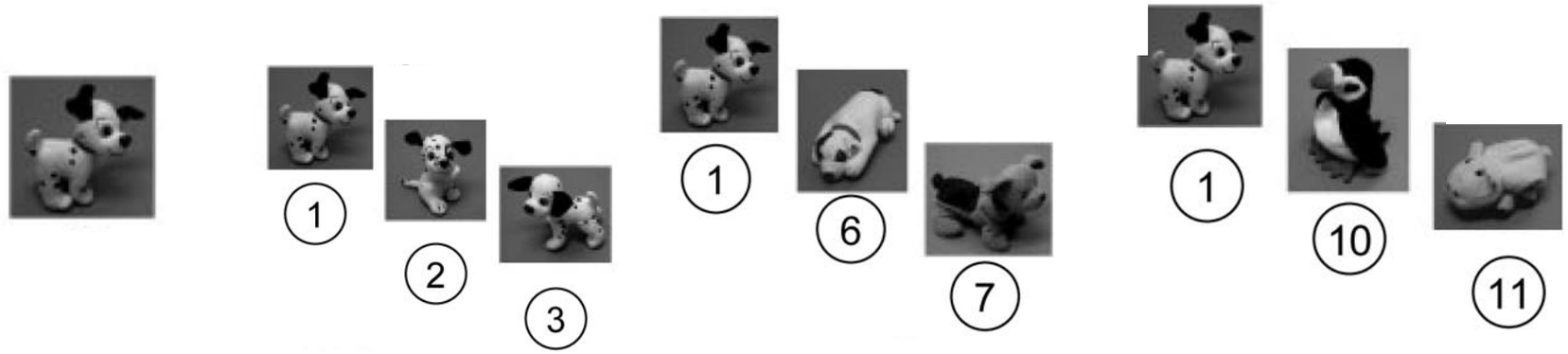
- Can 3-4 year olds learn from sparse, positive examples?

- If only one example, it is labeled 3 times to control number of labeling events

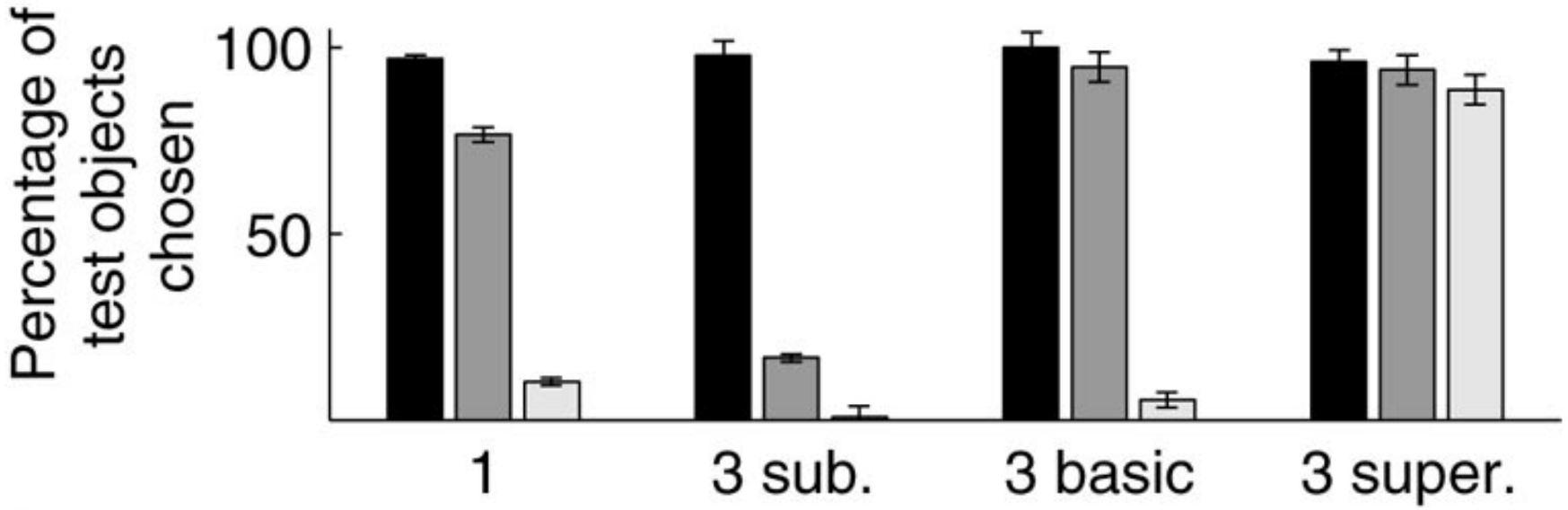
Which others are feps?
(YES/NO for each)



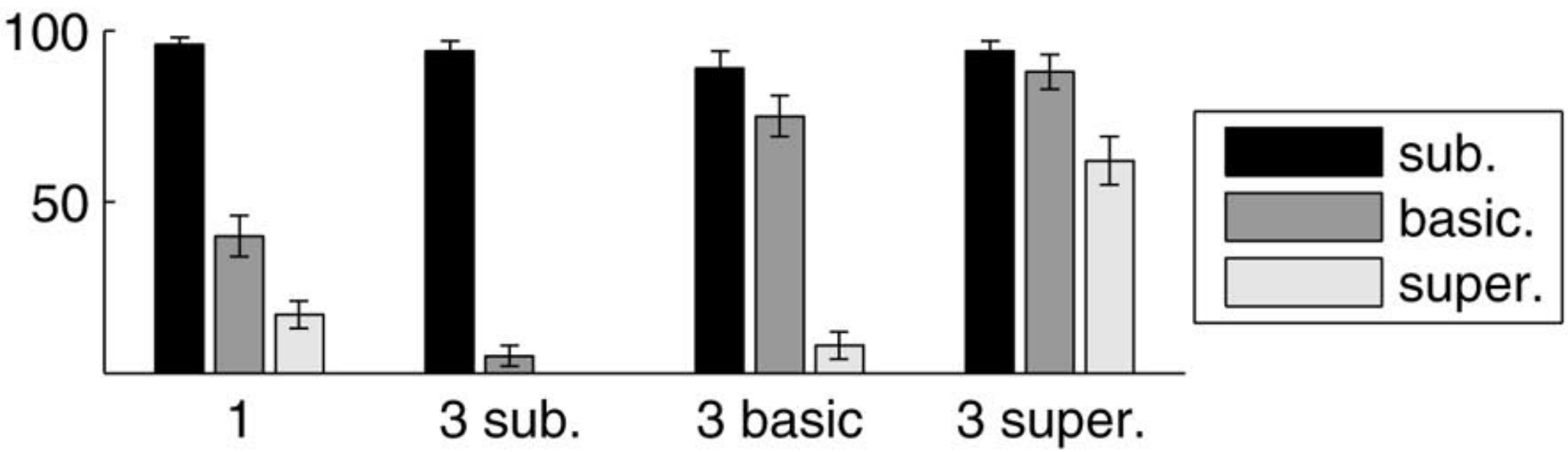
Xu and Tenenbaum Exp 3



Adults



Children



Bayesian concept learning

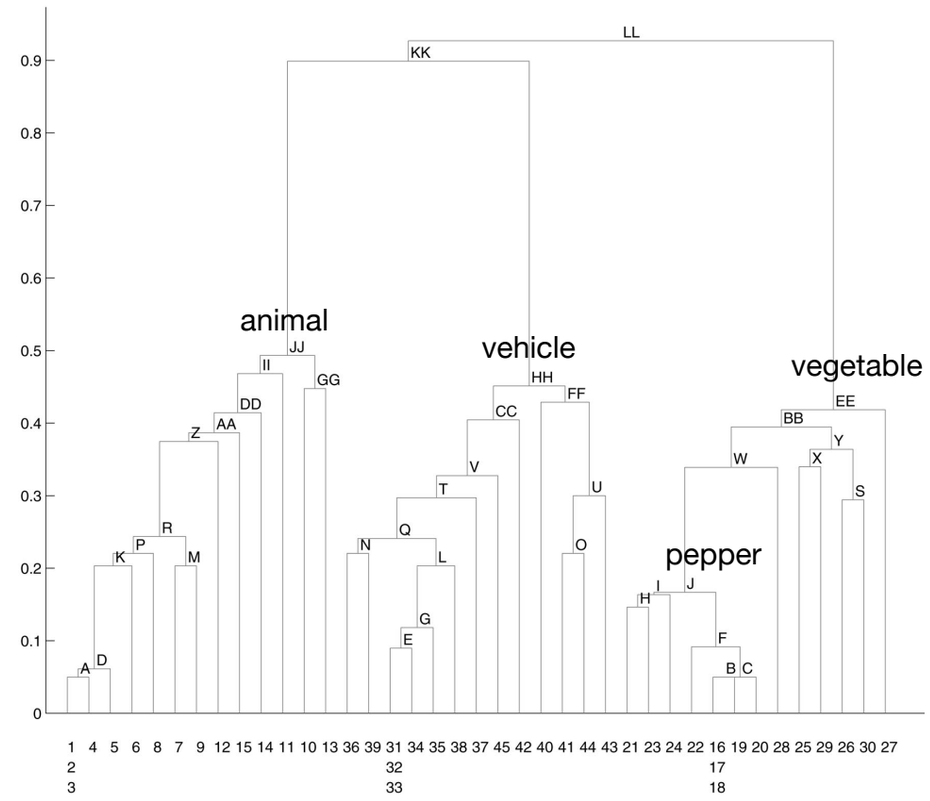
$h \in H$: **hypothesis about the meaning of word (node in tree structure)**

X : **1 or 3 positive examples**

n : **number of examples**

Posterior over word meanings

$$p(h \mid X) = \frac{P(X \mid h)P(h)}{P(X)}$$



Likelihood

$$P(X \mid h) = [\frac{1}{\textbf{size}(h)}]^n \approx [\frac{1}{\textbf{height}(h) + \epsilon}]^n$$

(height is the average within-node distance between examples)

Prior

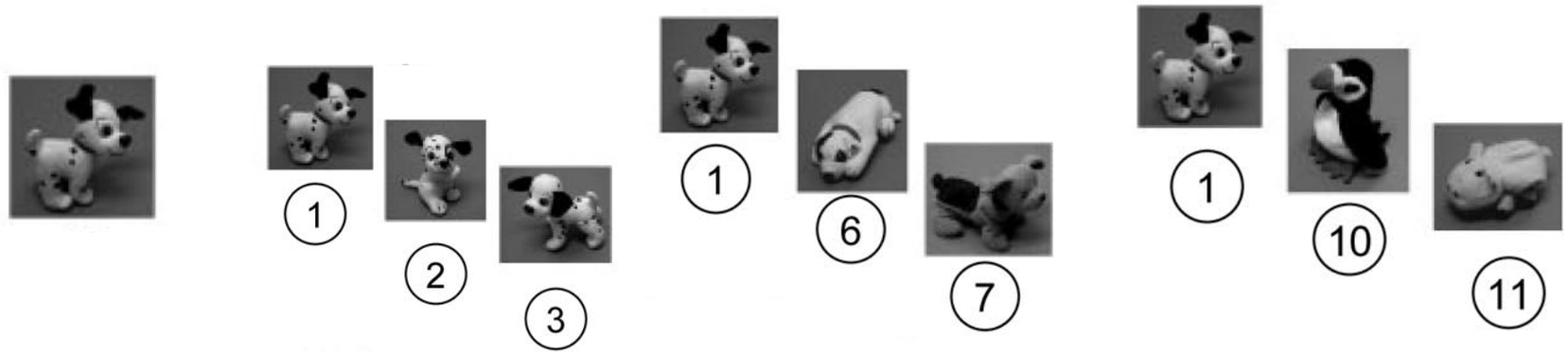
$$P(h) \propto \textbf{height}(\textbf{parent}[h])) - \textbf{height}(h)$$

favors more distinctive nodes,
or favor nodes at a certain level….

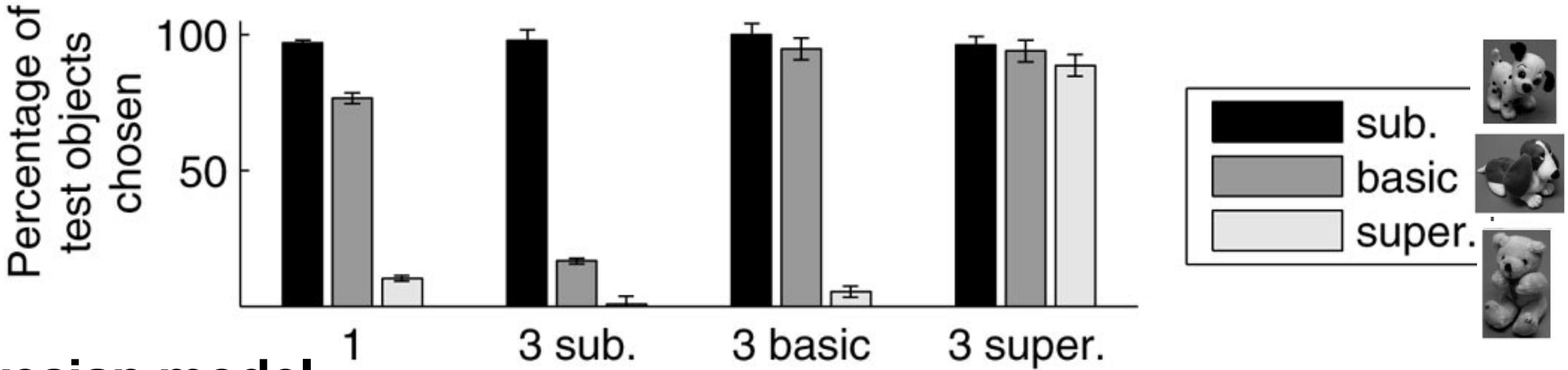
Generalizing to a new example *y*

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

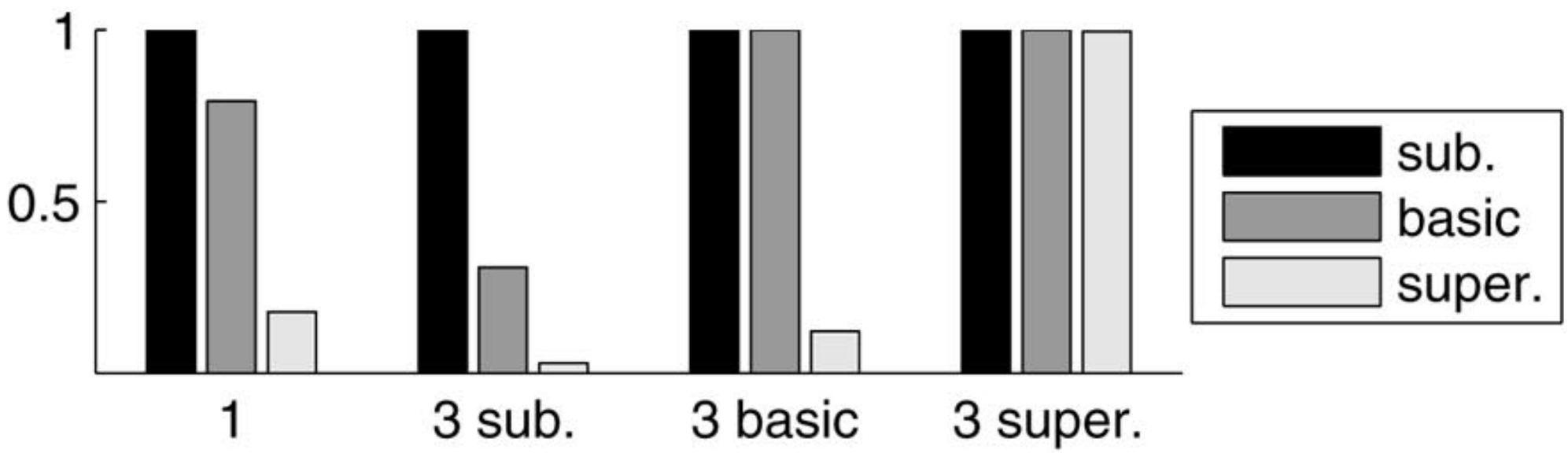
Model results



Adults

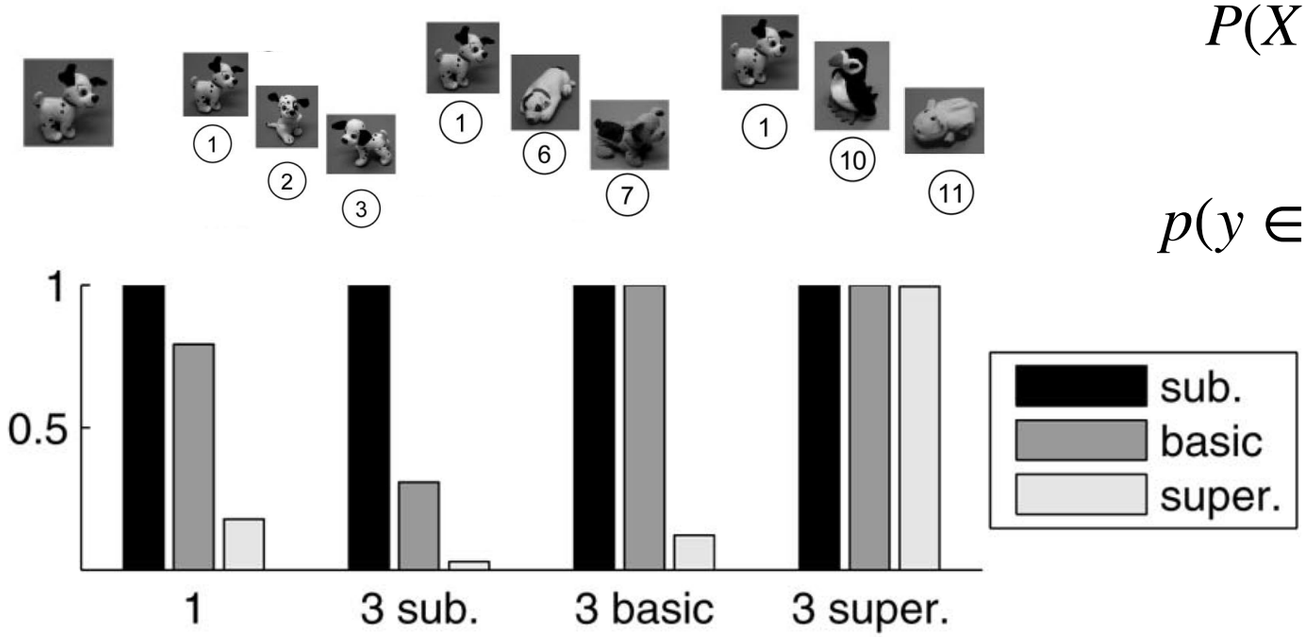


Bayesian model



Alternative Bayesian models

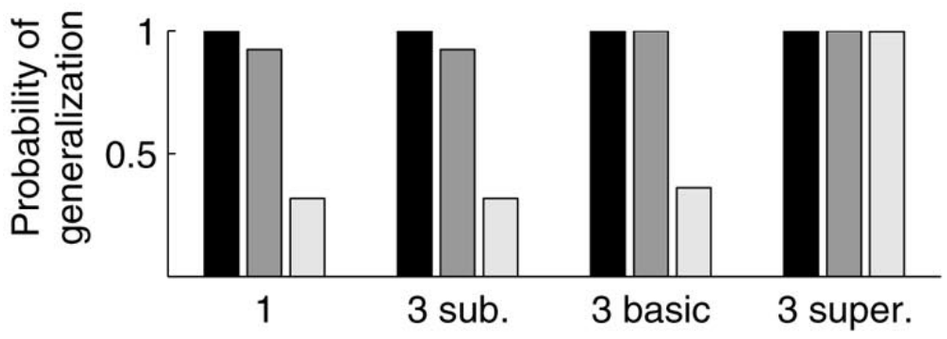
Full Bayesian model



$$P(X|h) = [\frac{1}{\textbf{height}(h) + \epsilon}]^n$$

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

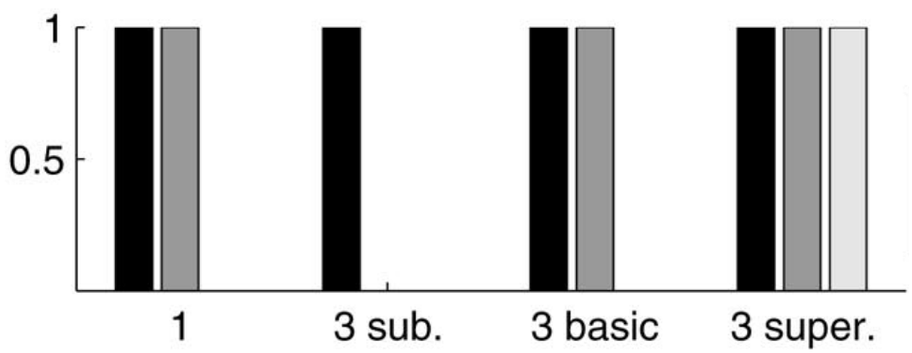
Weak Bayesian model



$$P(X|h) = 1$$

generalization isn't sharp enough

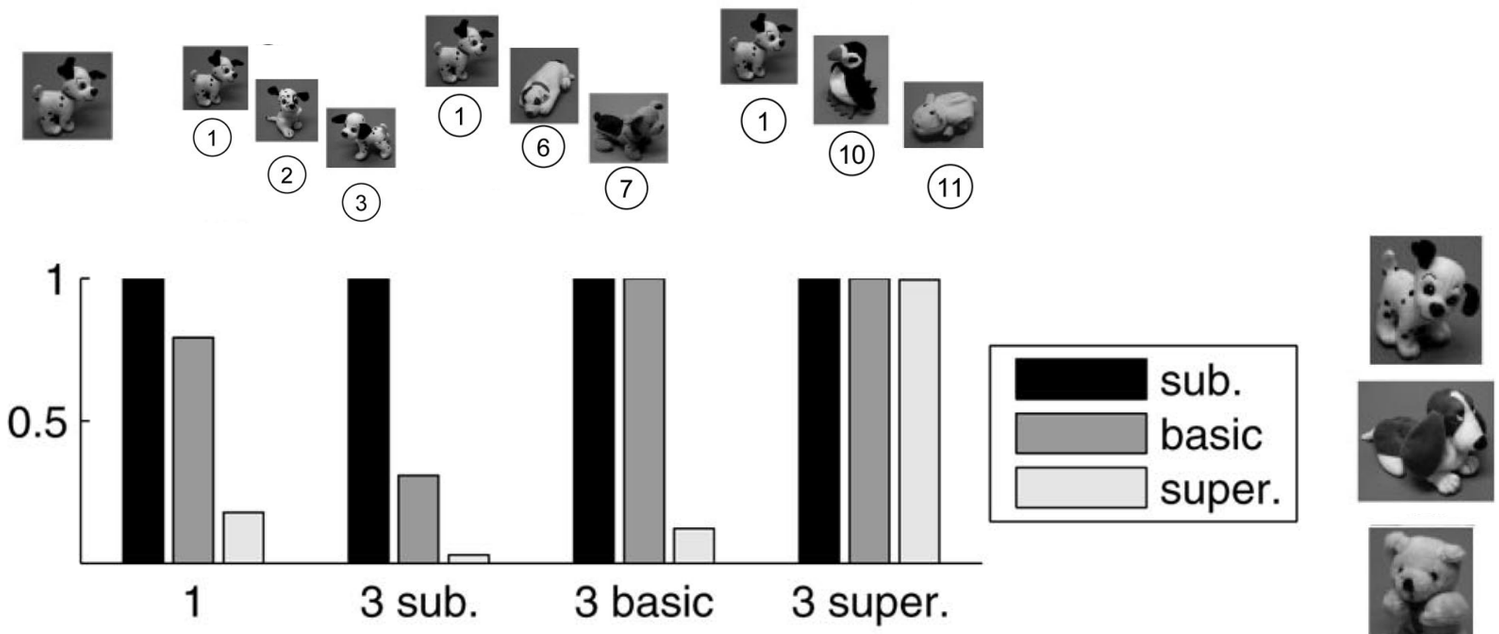
Most likely hypothesis only



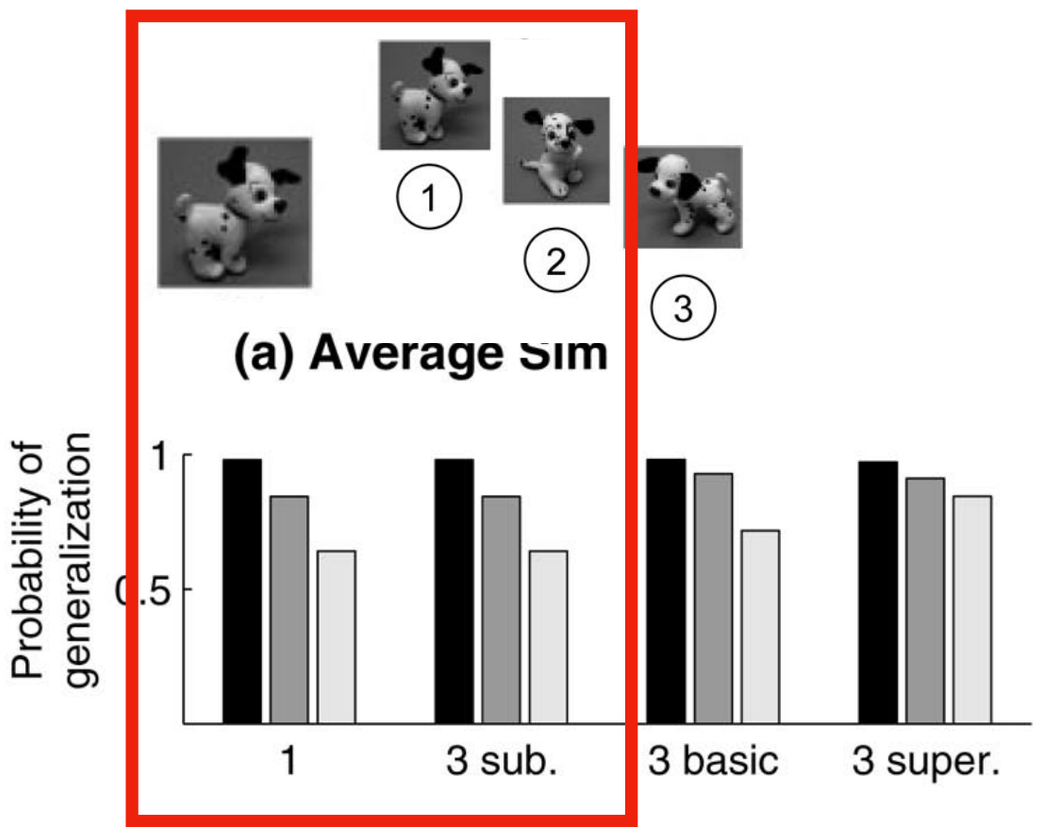
generalization is too sharp

What about an exemplar model?

Full Bayesian model



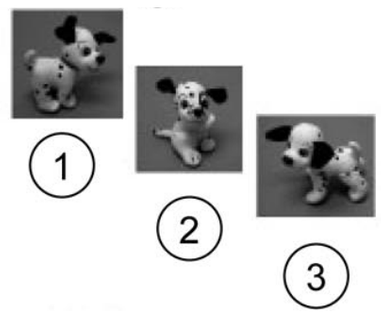
Exemplar model



Key point:

Exemplar model doesn't care between 1 example and 3 examples

$$\textbf{sim}(y, C) = \sum_{x \in C} \textbf{sim}(y, x)$$



Conclusions

- Children can learn new concepts from just one or a few positive examples

- Children are sensitive to sampling assumptions, whether it's three labels of the same exemplar or three different exemplars

- Bayesian models of concept learning provide an explicit model of sampling assumptions, and can learn concepts from just one or a few examples

- "Only a combination of sophisticated mental representations and sophisticated statistical inference machinery will be able to explain how adults and children can learn so many words so fast so accurately" (Xu & Tenenbaum)

    - Implication: neural network models and exemplar models (e.g., ALCOVE) are not up to the challenge

Developmental Science 10:3 (2007), pp 288–297 DOI: 10.1111/j.1467-7687.2007.00590.x

BAYESIAN SPECIAL SECTION

Sensitivity to sampling in Bayesian word learning

Fei Xu[1] and Joshua B. Tenenbaum[2]

1. *Department of Psychology, University of British Columbia, Canada*
2. *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA*

Abstract

We report a new study testing our proposal that word learning may be best explained as an approximate form of Bayesian inference (Xu & Tenenbaum, in press). Children are capable of learning word meanings across a wide range of communicative contexts. In different contexts, learners may encounter different sampling processes generating the examples of word–object pairings they observe. An ideal Bayesian word learner could take into account these differences in the sampling process and adjust his/her inferences about word meaning accordingly. We tested how children and adults learned words for novel object kinds in two sampling contexts, in which the objects to be labeled were sampled either by a knowledgeable teacher or by the learners themselves. Both adults and children generalized more conservatively in the former context; that is, they restricted the label to just those objects most similar to the labeled examples when the exemplars were chosen by a knowledgeable teacher, but not when chosen by the learners themselves. We discuss how this result follows naturally from a Bayesian analysis, but not from other statistical approaches such as associative word-learning models.

Introduction

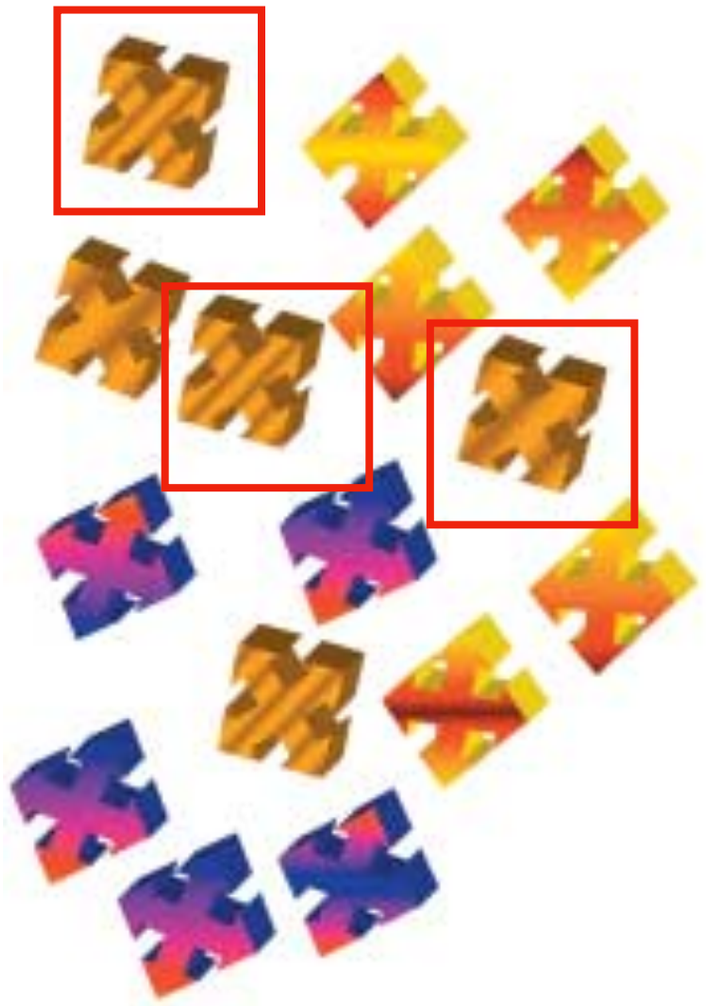
Models for how children learn the meanings of words traditionally fall into two classes. One class of models treats the process as inferential in nature, akin to reasoning. Although the child presumably is not consciously working out each step of the reasoning process and the computations may be done implicitly, the child learner is assumed to draw on a set of hypotheses about candidate word meanings and to evaluate these hypotheses based on observed input using one or more principles of rational inference (e.g. Bloom, 2000; Carey, 1978; Markman, 1989; Siskind, 1996). In contrast, associative models assume that the learner represents a matrix of graded word–object mappings, and the strengths of these mappings are incrementally increased or decreased over time given repeated exposures (e.g. Colunga & Smith, 2005; Gasser & Smith, 1998; Regier, 2003, 2005).

Confronted with a novel word, the learner constructs a hypothesis space of candidate word meanings (i.e. lexicalizable concepts) and a prior probability distribution over that hypothesis space. Given one or more examples of objects labeled by the new word, the learner updates the prior to a posterior distribution of beliefs based on the likelihood of observing these examples under each candidate hypothesis. The prior represents any knowledge (due to previous learning or innate endowment) about which meanings are more or less likely to be the target of the new word, independent of the observed examples. The likelihood is based on the sampling process presumed to have generated the observed object–label pairs.

Recent studies of word learning with adults and children provide some initial evidence for this account. These studies test generalization: participants are shown one or more examples of a novel word (e.g. ‘blicket’) and are asked to judge which objects from a test set the word

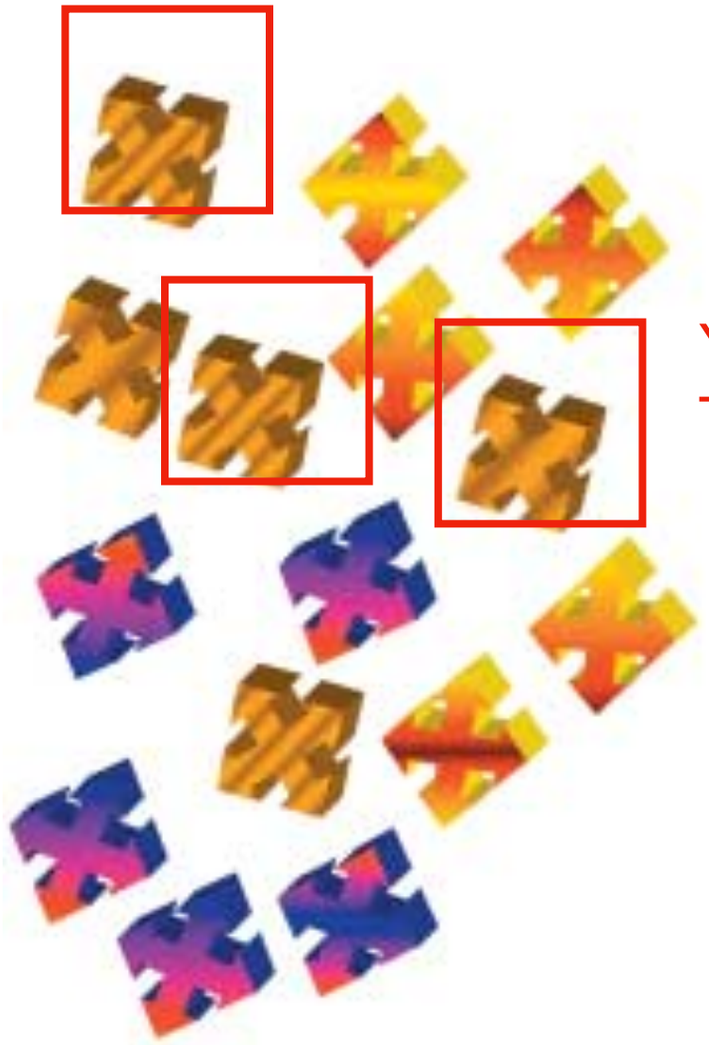
Teacher-driven condition

"Here are three feps"

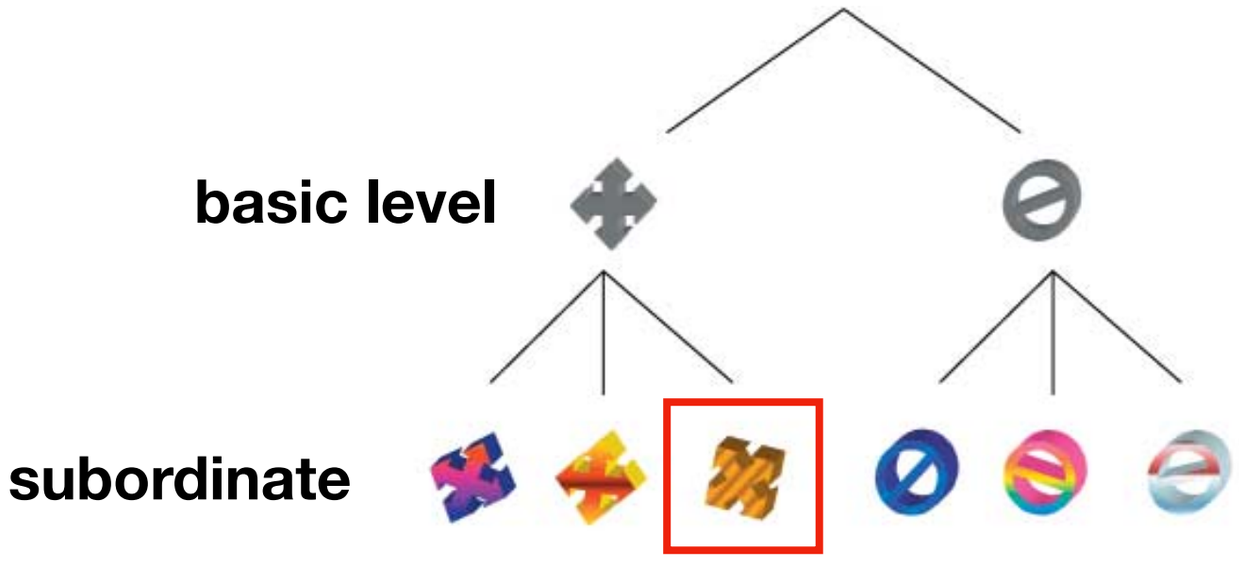


Learner-driven condition

“Here is a fep. Can you point to two more?”



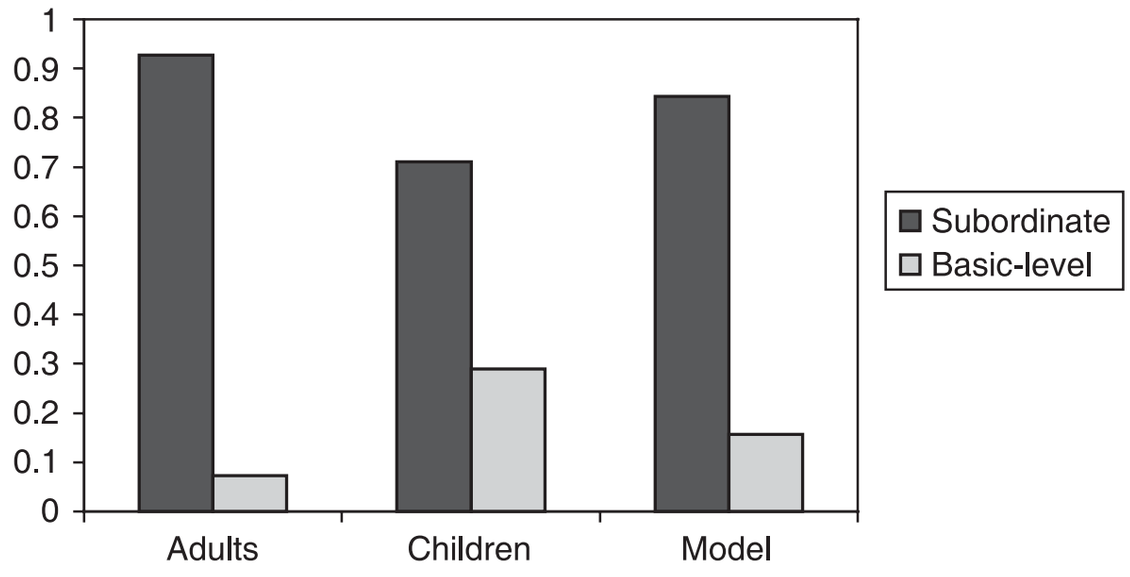
Yes, that’s right.
That’s right too



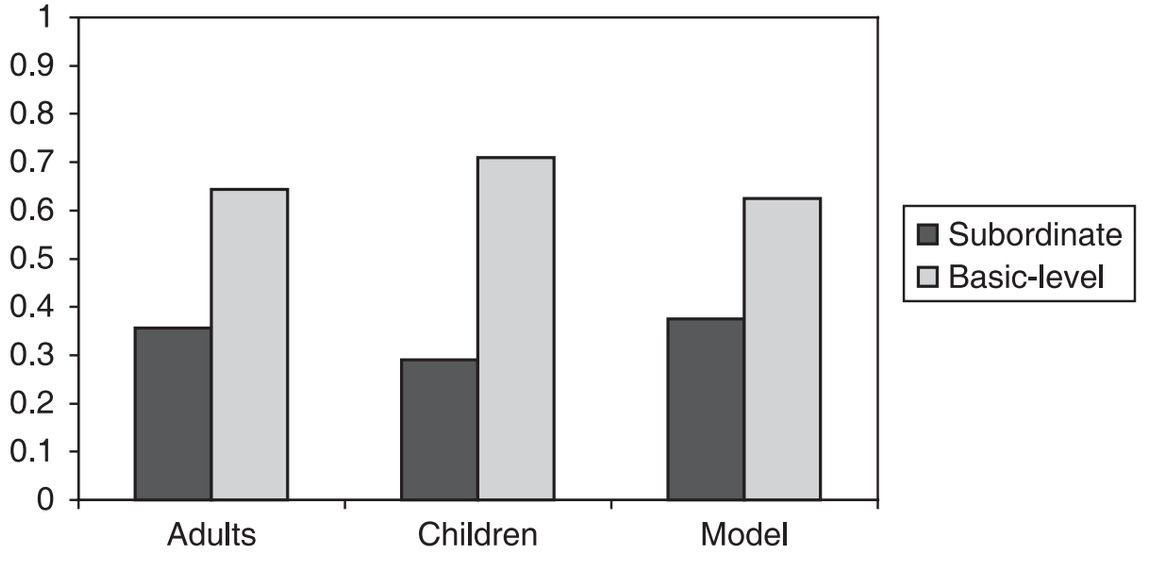
"Here are three feps"

"Here is a fep. Can you point to two more?"

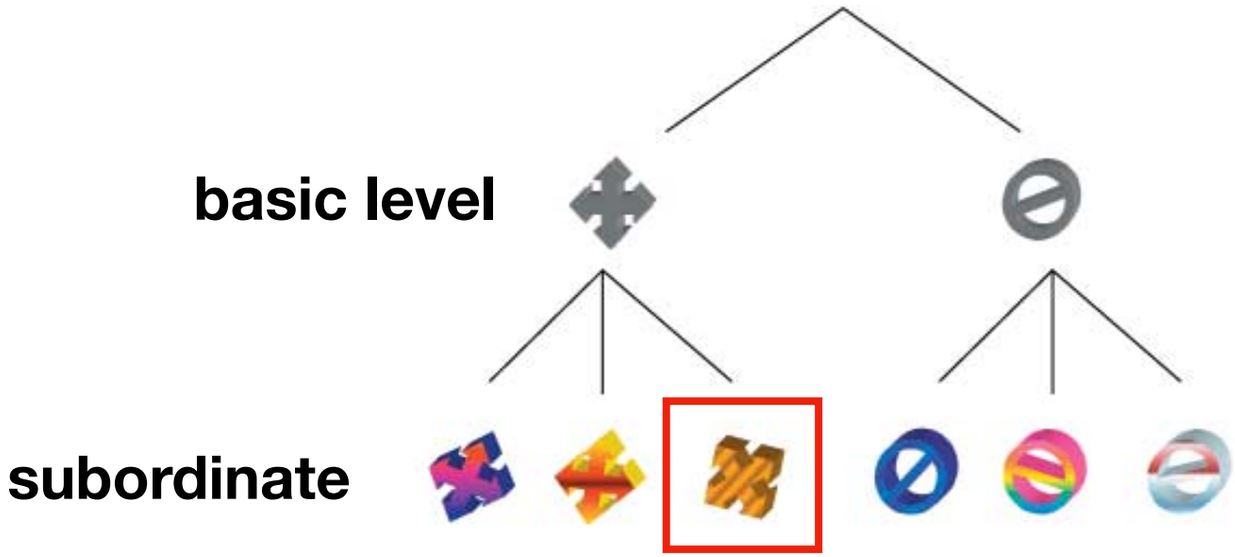
(a) Teacher-driven Condition



(b) Learner-driven Condition

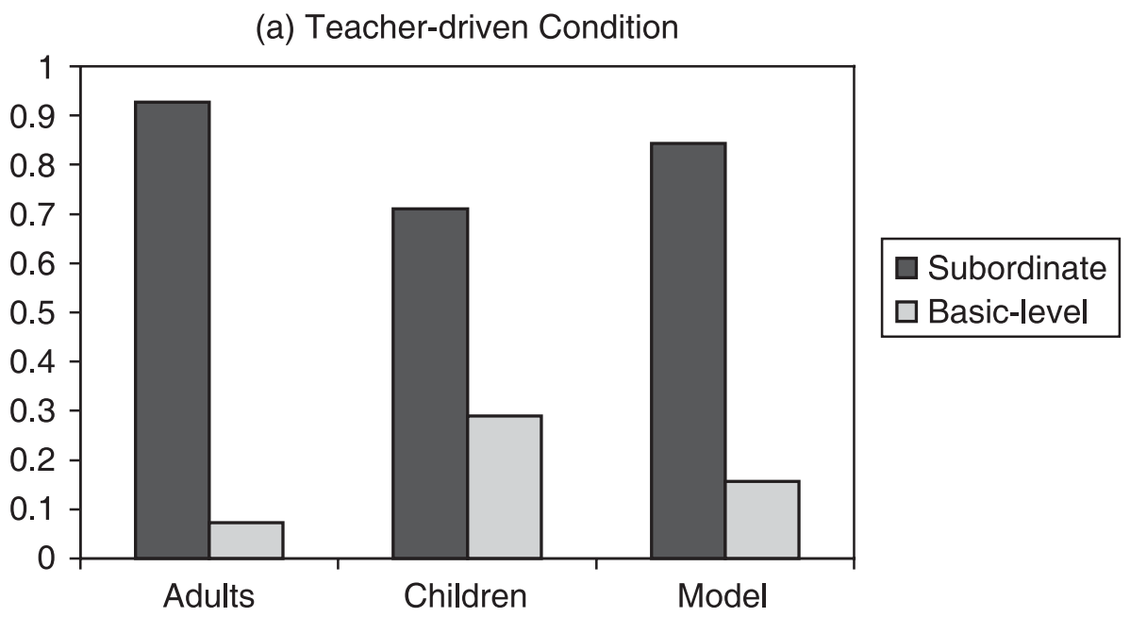


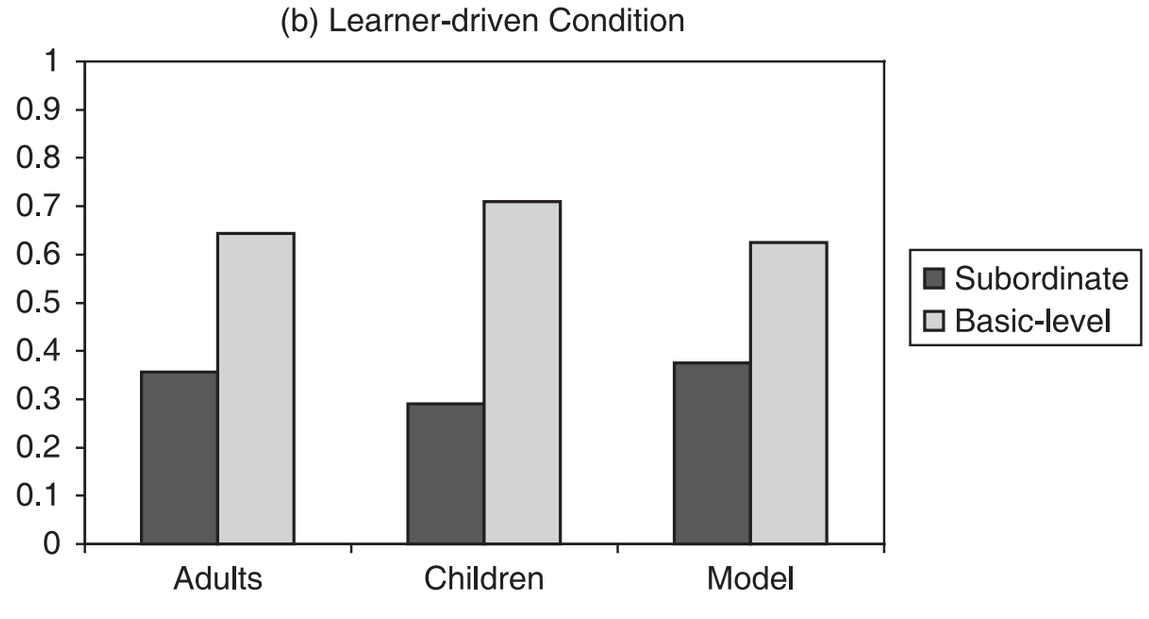
Key point: exemplar models, ALCOVE, and neural networks don't differentiate based on sampling process



"Here are three feps"

"Here is a fep. Can you point to two more?"





Key point: ALCOVE and other gradient-based algorithms don't differentiate between two sampling conditions

response $P(y \in A)$ **0.503**

$$\textbf{sim}(y, C) = \sum_{x \in X} w_{cx}\ \textbf{sim}(y, x)$$

association weights w_{Ax} 0.015 0.002 0.002 0.0

$\textbf{sim}(y, x)$ 1.0 0.11 0.11 0.01

$$\textbf{sim}(y, x) = e^{\sum_{D_i} \alpha_i |x_i - y_i|}$$

w_{Ax} 0.002 0.0 0.0 0.0

$\textbf{sim}(y, x)$ 0.11 0.01 0.01 0.0

} Exemplars X

attention weights

α_{color} 0.33 $\alpha_{texture}$ 0.33 α_{slash} 0.33

current stimulus y

Cognitive Science 32 (2008) 108–154

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210701802071

A Rational Analysis of Rule-Based Concept Learning

Noah D. Goodman[a], Joshua B. Tenenbaum[a], Jacob Feldman[b], Thomas L. Griffiths[c]

[a]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*
[b]*Department of Psychology, Rutgers University*
[c]*Department of Psychology, University of California, Berkeley*

Abstract

This article proposes a new model of human concept learning that provides a rational analysis of learning feature-based concepts. This model is built upon Bayesian inference for a grammatically structured hypothesis space—a concept language of logical rules. This article compares the model predictions to human generalization judgments in several well-known category learning experiments, and finds good agreement for both average and individual participant generalizations. This article further investigates judgments for a broad set of 7-feature concepts—a more natural setting in several ways—and again finds that the model explains human performance.

Keywords: Concept learning; Categorization; Bayesian induction; Probabilistic grammar; Rules; Language of thought

Motivating rational rules

- The author's lay out three themes in concept learning
  - Concepts are used to discriminate between objects, events, relationships, etc.
  - Concepts are learned inductively from sparse and noisy data
  - Concepts are compositional and are formed by combining simpler concepts

Rational rules is an attempt to combine these themes into a single model by combining the "classical view"/rule learning with probabilistic inference

Key question: The classical view (e.g., rules and definitions) does not account for graded effects in categorization and learning. Can we account for these effects by performing Bayesian inference over rules and definitions?

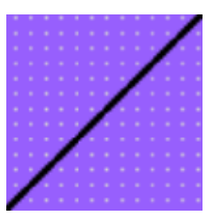
Model's aim is to learn defining rules for concepts

Type I

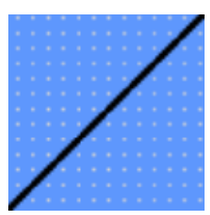
$(f_1(x) = 0)$

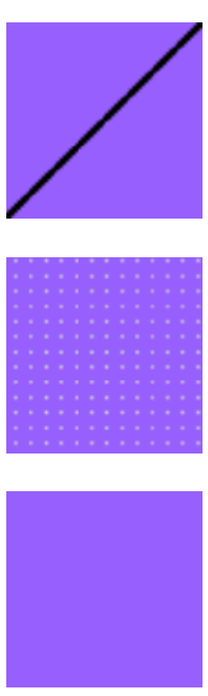
(color is purple)

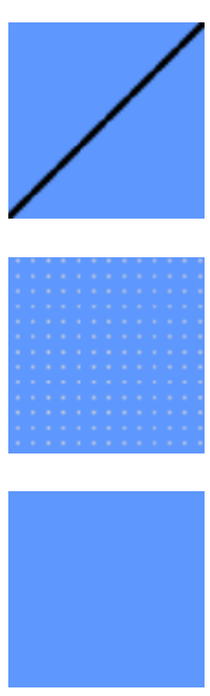
A



B





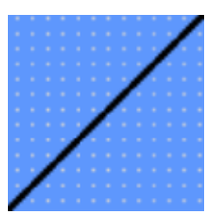


Type II

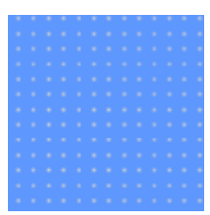
$((f_1(x) = 1) \wedge (f_2(x) = 1)) \vee ((f_1(x) = 0) \wedge (f_2(x) = 0))$

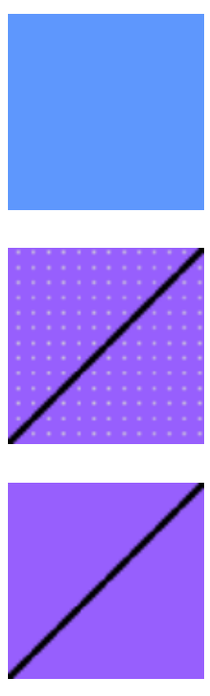
(color is blue AND has slash) OR (color is purple AND has no slash)

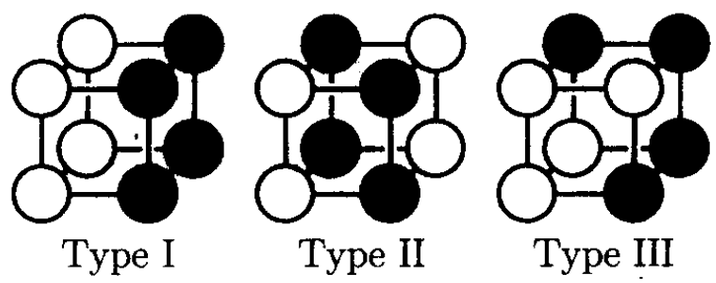
A

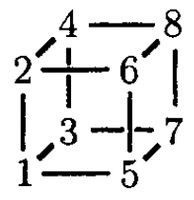


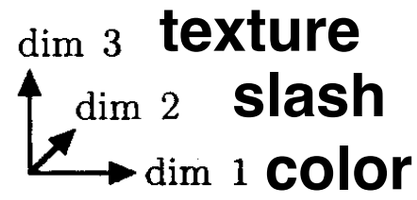
B











$f_1(x)$: **color**

$f_2(x)$: **slash**

$f_3(x)$: **texture**

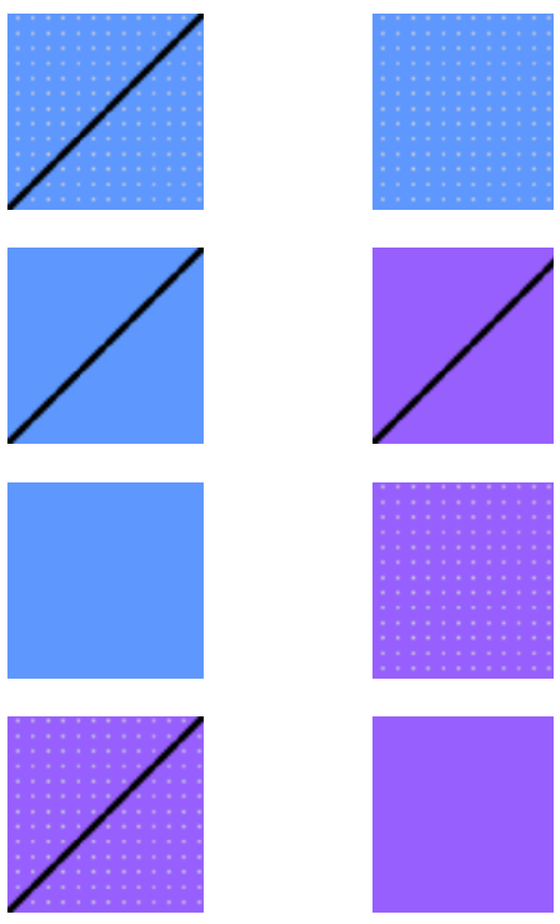
Model's aim is to learn defining rules for concepts

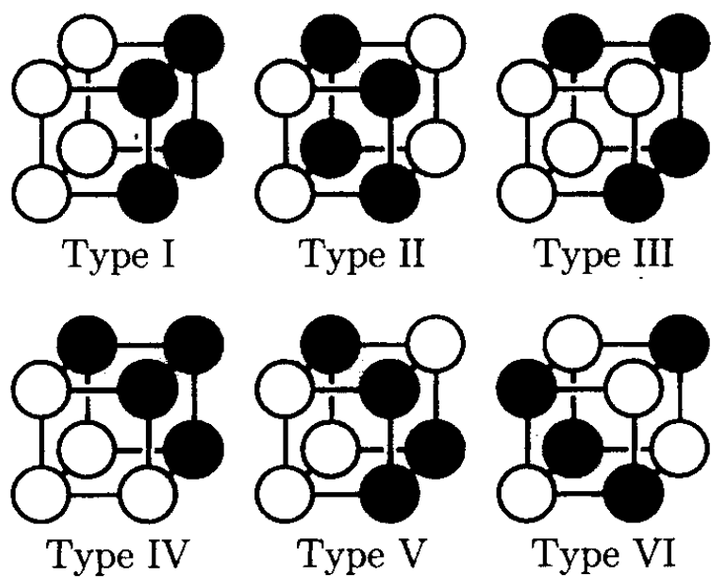
Type III

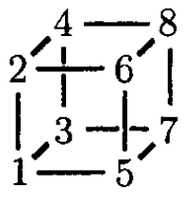
$$((f_1(x) = 0) \wedge (f_3(x) = 0)) \vee ((f_2(x) = 0) \wedge (f_3(x) = 1)),$$

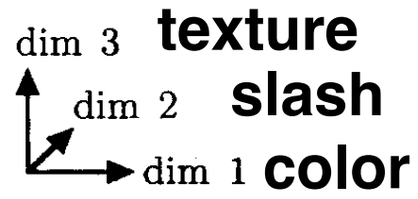
(color is blue AND has no texture) OR (has slash and has texture)

A B









$f_1(x)$: **color**

$f_2(x)$: **slash**

$f_3(x)$: **texture**

Rational rules model

F : **formula that defines a concept (Rule)**

X : **set of observed examples**

$l(X)$: **labels provided to examples**

Posterior over word meanings

$$p(F \mid l(X); X) \propto p(l(X) \mid F; X)p(F)$$

Likelihood (noisy labeling according to formula)

$$p(l(X) \mid F; X) \propto e^{-bQ_l(F)}$$

$Q_l(F)$ number of labels that have been misaligned according to formula

Prior

$p(F)$ Based on derivation of formula under a probability context free grammar (favors short formulas and re-use of derivational steps)

Generalizing to a new example *y*

$$p(y \in l \mid l(X); X) = \sum_F P(y \in C \mid F)p(F \mid l(X); X)$$

Deriving a formula

A grammar of rule-based concepts

$$
\begin{aligned}
&(\mathrm{S1}) \quad S \rightarrow \forall x\ \ell(x) \Leftrightarrow (D)\\
&(\mathrm{D1}) \quad D \rightarrow (C) \vee D\\
&(\mathrm{D2}) \quad D \rightarrow False\\
&(\mathrm{C1}) \quad C \rightarrow P \wedge C\\
&(\mathrm{C2}) \quad C \rightarrow True\\
&(\mathrm{P1}) \quad P \rightarrow F_1\\
&\qquad \vdots\\
&(\mathrm{PN}) \quad P \rightarrow F_N\\
&(\mathrm{F}_1\mathrm{1}) \quad F_1 \rightarrow f_1(x) = 1\\
&(\mathrm{F}_1\mathrm{2}) \quad F_1 \rightarrow f_1(x) = 0\\
&\qquad \vdots\\
&(\mathrm{F}_N\mathrm{1}) \quad F_N \rightarrow f_N(x) = 1\\
&(\mathrm{F}_N\mathrm{2}) \quad F_N \rightarrow f_N(x) = 0
\end{aligned}
$$

Derivation of rules:

$$
\begin{aligned}
& S\\
S1: \quad & \overbrace{\forall x\ \ell(x) \Leftrightarrow (D)}\\
D1: \quad & \forall x\ \ell(x) \Leftrightarrow (\overbrace{(C) \vee D}^{})\\
D2: \quad & \forall x\ \ell(x) \Leftrightarrow ((C) \vee \overbrace{\mathrm{False}})\\
C1: \quad & \forall x\ \ell(x) \Leftrightarrow (\overbrace{(P \wedge C)} \vee \mathrm{False})\\
C2: \quad & \forall x\ \ell(x) \Leftrightarrow ((P \wedge \overbrace{\mathrm{True}}) \vee \mathrm{False})\\
P1: \quad & \forall x\ \ell(x) \Leftrightarrow ((\overbrace{F_1} \wedge \mathrm{True}) \vee \mathrm{False})\\
F_1 2: \quad & \forall x\ \ell(x) \Leftrightarrow (((\overbrace{f_1(x)=0}) \wedge \mathrm{True}) \vee \mathrm{False})
\end{aligned}
$$

Accounts for SHJ due to boolean complexity

Type I

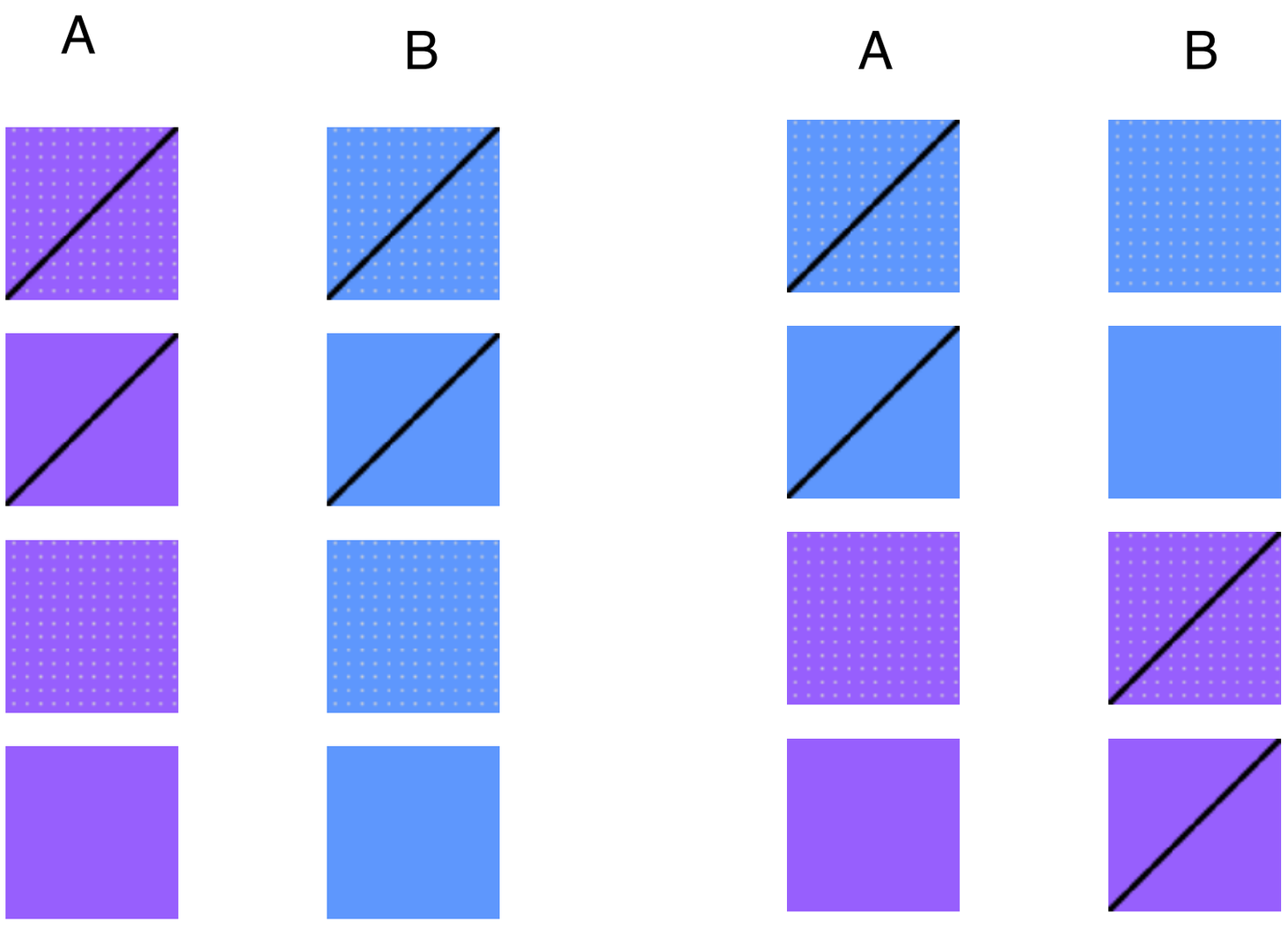
$(f_1(x) = 0)$

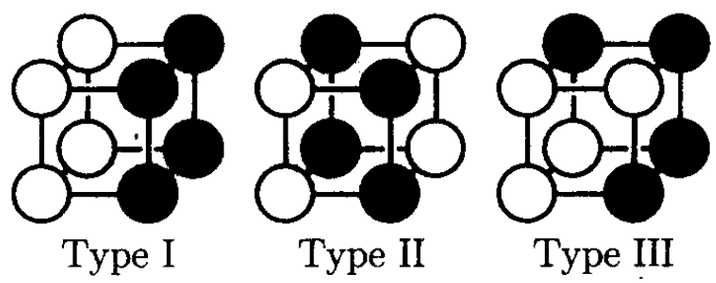
(color is purple)

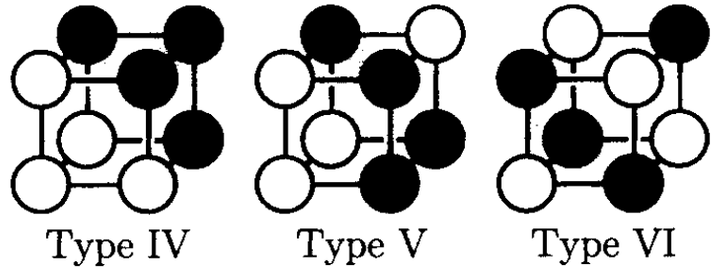
Type II

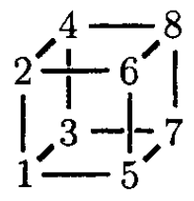
$((f_1(x) = 1) \wedge (f_2(x) = 1)) \vee ((f_1(x) = 0) \wedge (f_2(x) = 0))$

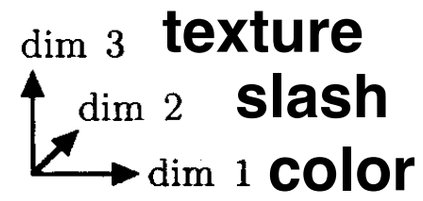
(color is blue AND has slash) OR (color is purple AND has no slash)











$f_1(x)$: **color**

$f_2(x)$: **slash**

$f_3(x)$: **texture**

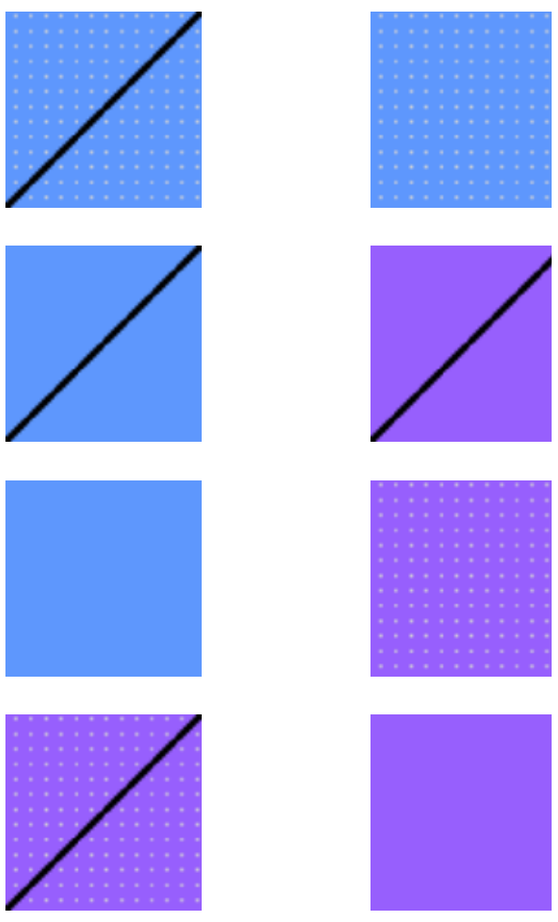
Accounts for SHJ due to boolean complexity

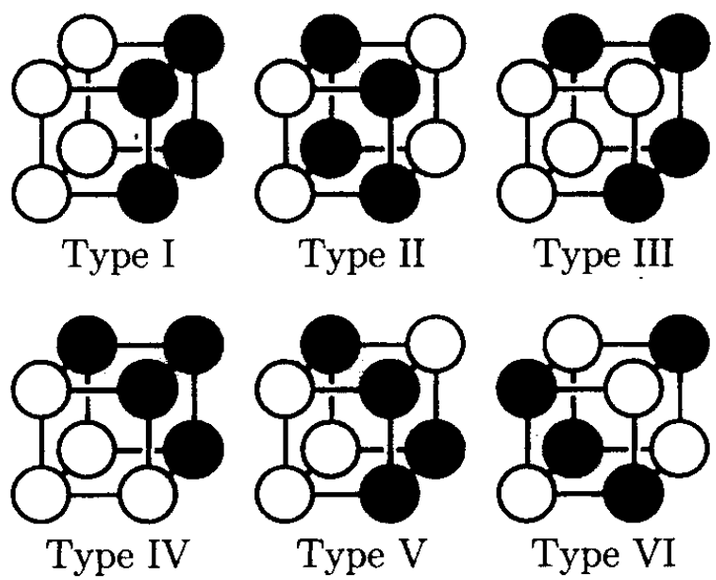
Type III

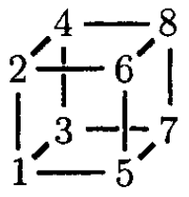
$$((f_1(x) = 0) \wedge (f_3(x) = 0)) \vee ((f_2(x) = 0) \wedge (f_3(x) = 1)),$$

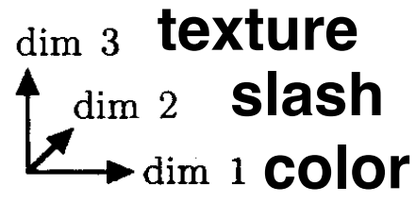
(color is blue AND has slash) OR (has slash and has texture)

A B









$f_1(x)$: **color**

$f_2(x)$: **slash**

$f_3(x)$: **texture**

Medin & Schaffer Experiment 2

famous “5-4” category structure

TRAINING STIMULI

"A" STIMULI

STIMULUS NUMBER	DIMENSION VALUES C	F	S	N	FE	RATING
4	1	1	1	0	4.9	4.8
7	1	0	1	0	3.3	5.4
15	1	0	1	1	3.2	5.1
13	1	1	0	1	4.8	5.2
5	0	1	1	1	4.5	5.2
Prototype:	**1**	**1**	**1**	**1**		

"B" STIMULI

STIMULUS NUMBER	DIMENSION VALUES C	F	S	N	FE	RATING
12	1	1	0	0	5.5	5.0
2	0	1	1	0	5.2	5.1
14	0	0	0	1	3.9	5.2
10	0	0	0	0	3.1	5.5
Prototype:	**0**	**0**	**0**	**0**		

Key comparison is stimulus 4 vs stimulus 7

- Prototype model would predict stimulus 4 is easier to learn
  - It’s more similar to the prototype
- Exemplar model would predict stimulus 7 is easier to learn
  - Has two near neighbors, 15 and 4
- **Behavior results favor exemplar model**: stimulus 7 had fewer error (FE) and higher confidence rating

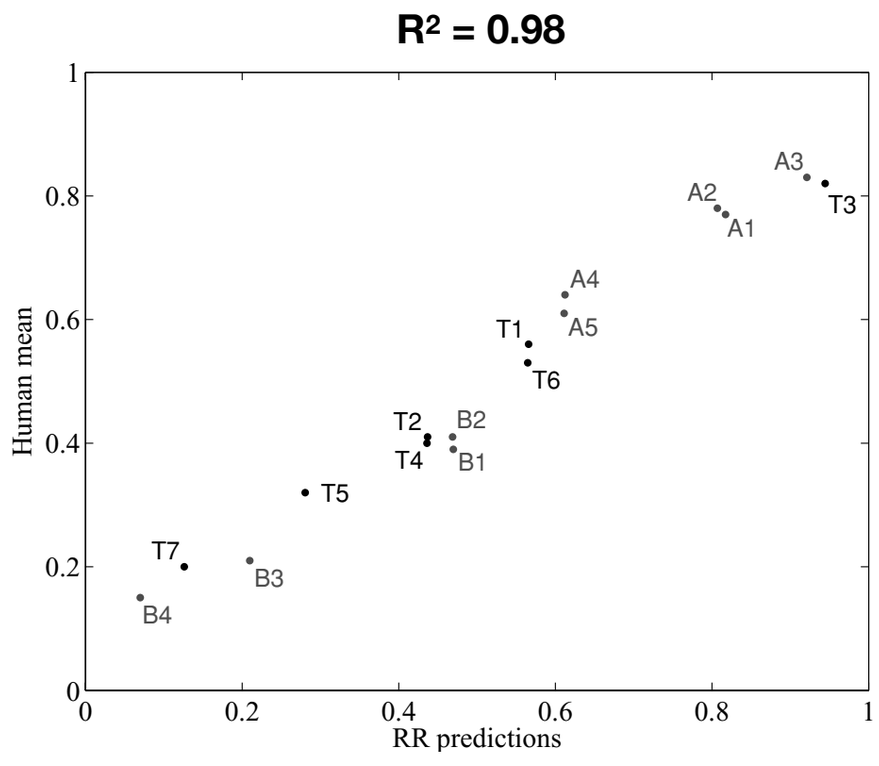
Medin & Schaffer Experiment 2

Rational rules can account for enhancement by neighboring exemplars

Table 3
The category structure of Medin & Schaffer (1978), with the human data of Nosofsky et al. (1994a), and the predictions of the Rational Rules model ($b = 1$).

Object		Feature Values	probability of classifying as category "A": Human	RR_{DNF}
A1	more like prototype (all 0's)	0001	0.77	0.82
A2	more like other exemplars	0101	0.78	0.81
A3		0100	0.83	0.92
A4		0010	0.64	0.61
A5		1000	0.61	0.61

Rational rules accounts for this by reliance on rules that often use Features 1 and 3, and not often Feature 2



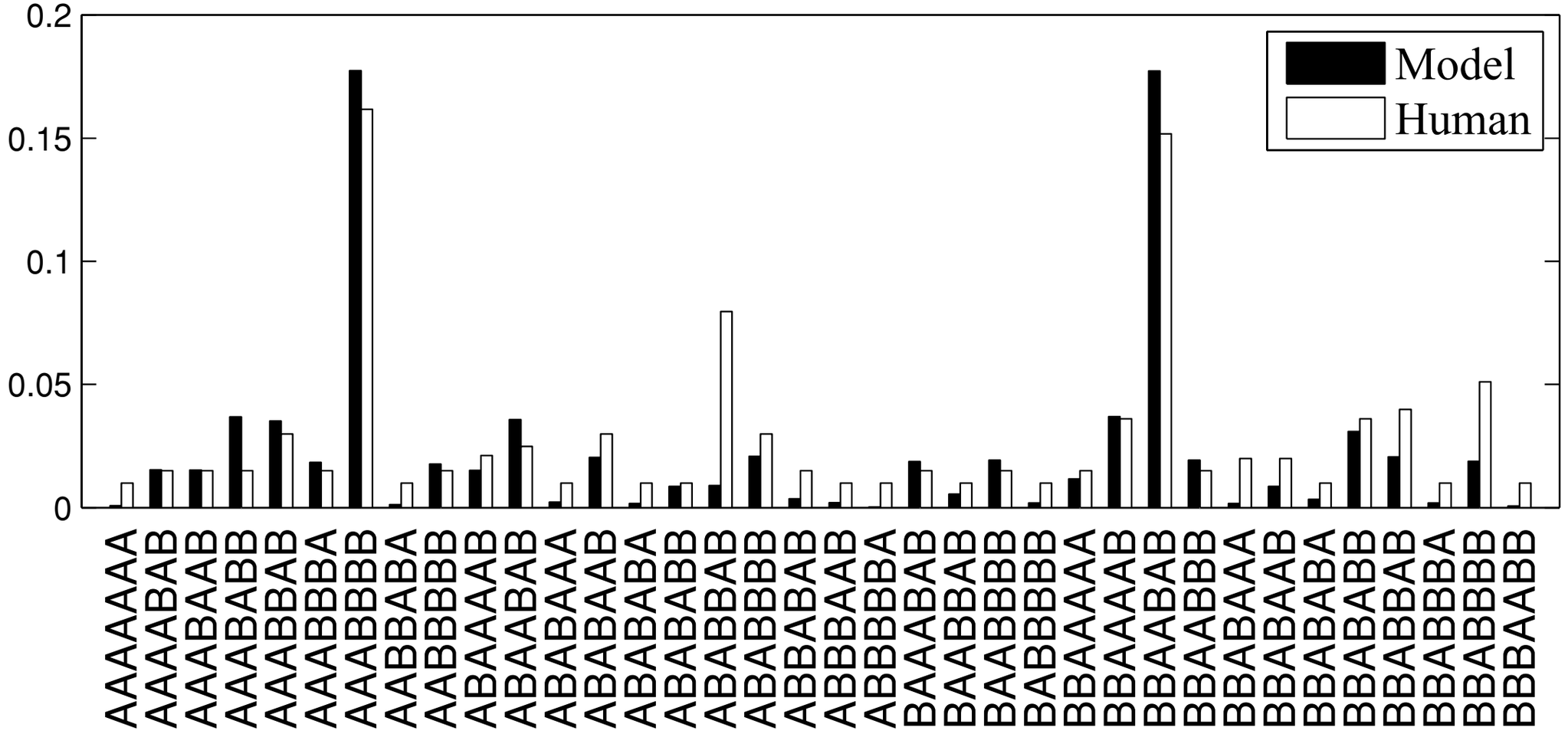
Individual generalization patterns

TRAINING STIMULI

"A" STIMULI							"B" STIMULI						
STIMULUS NUMBER	DIMENSION VALUES				FE	RAT-ING	STIMULUS NUMBER	DIMENSION VALUES				FE	RAT-ING
	C	F	S	N				C	F	S	N		
4	1	1	1	0	4.9	4.8	12	1	1	0	0	5.5	5.0
7	1	0	1	0	3.3	5.4	2	0	1	1	0	5.2	5.1
15	1	0	1	1	3.2	5.1	14	0	0	0	1	3.9	5.2
13	1	1	0	1	4.8	5.2	10	0	0	0	0	3.1	5.5
5	0	1	1	1	4.5	5.2							

famous “5-4” category structure

Probability of responding with the indicated categorizations of 7 transfer stimuli



People don’t necessarily favor linearly separable categories (as accounted for by ALCOVE)

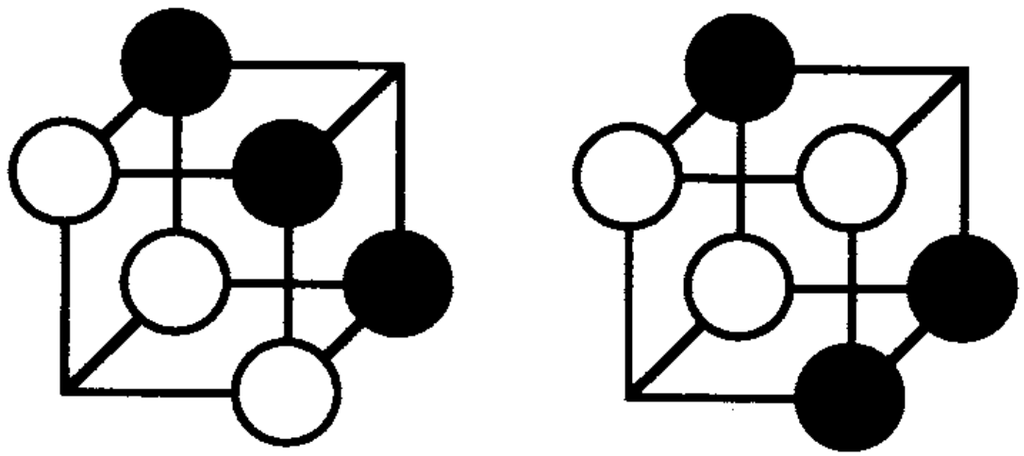


Figure 10. Category structures used by Medin and Schwanenflugel (1981, Experiment 4). (The linearly separable structure is a subset of Type IV in the Shepard, Hovland, and Jenkins, 1961, studies [cf. Figure 4], whereas the nonlinearly separable structure is the corresponding subset from Type III.)

linearly separable condition: people had 39.5 errors on average

non-linearly separable condition: people had 38.0 errors on average

(criterion was two error-free passes through stimuli)

Rational rules can learn NLS concepts more easily than LS

A complexity 4 rule perfectly discriminates the NLS case, but you need a more complex rule for the LS case

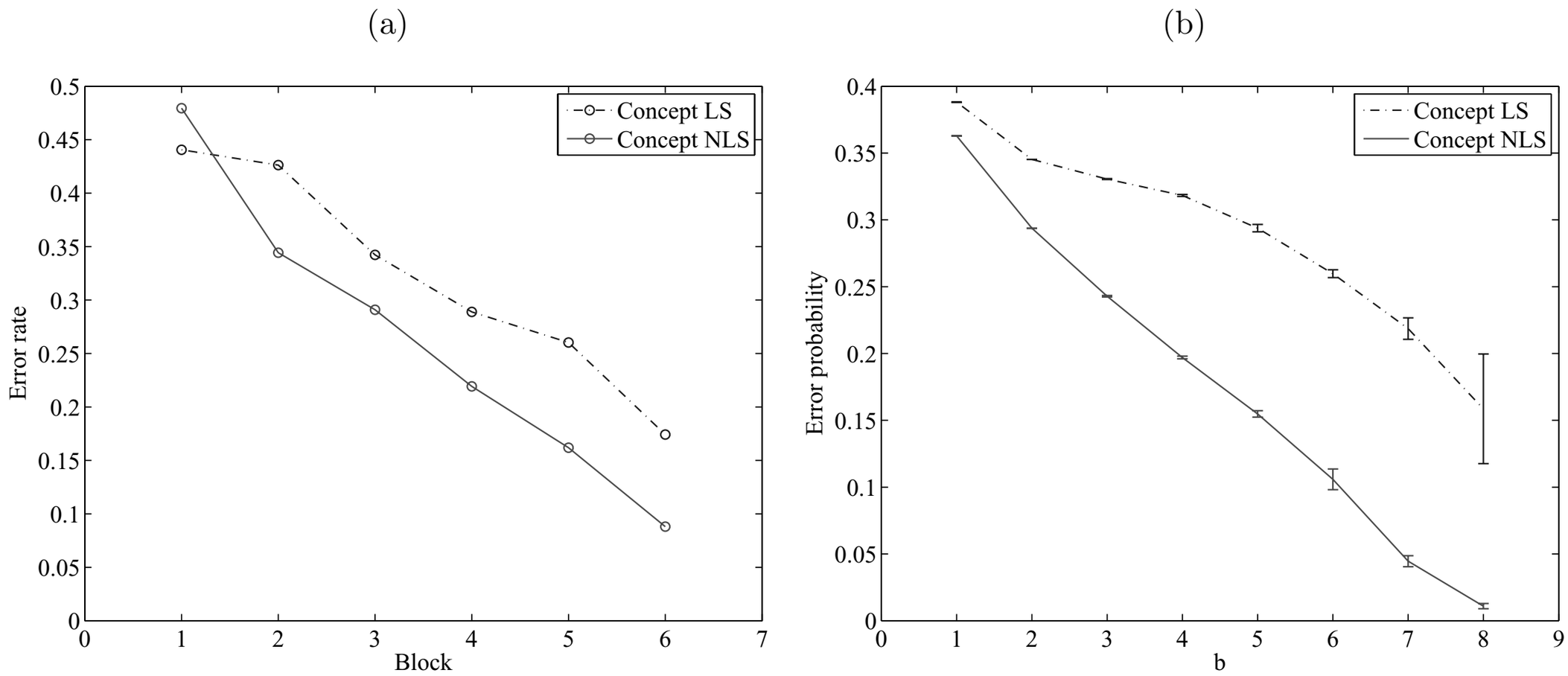


Fig. 5. (a) The human data from Medin and Schwanenflugel (1981) for the category structures in Table 4, showing that linearly separable Concept LS was more difficult to learn than Concept NLS, which is not linearly separable. (b) Predictions of the RR_{DNF} model: the probability of an incorrect response versus the outlier parameter *b*.