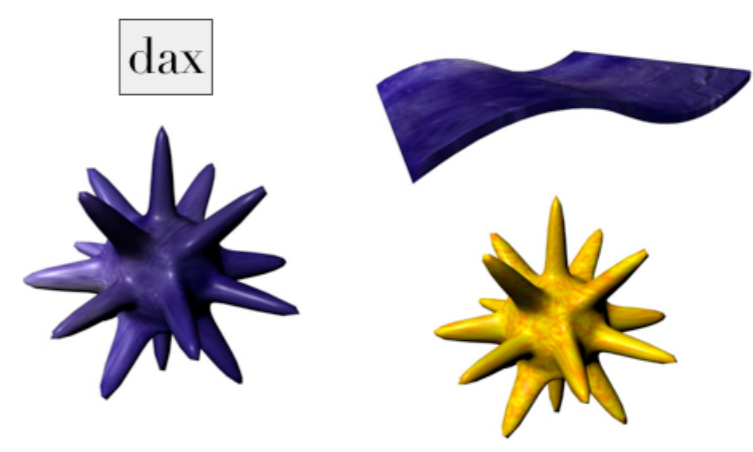
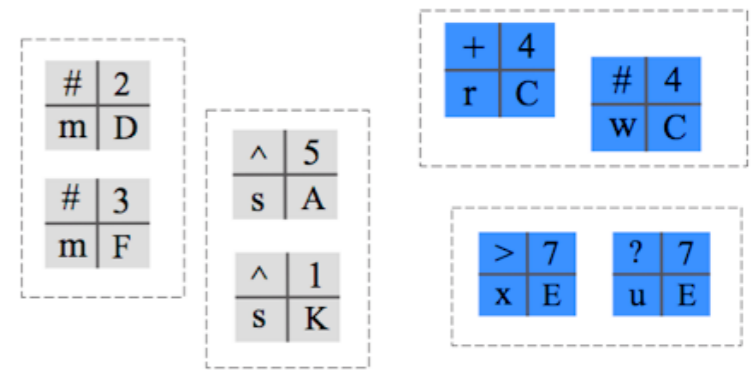
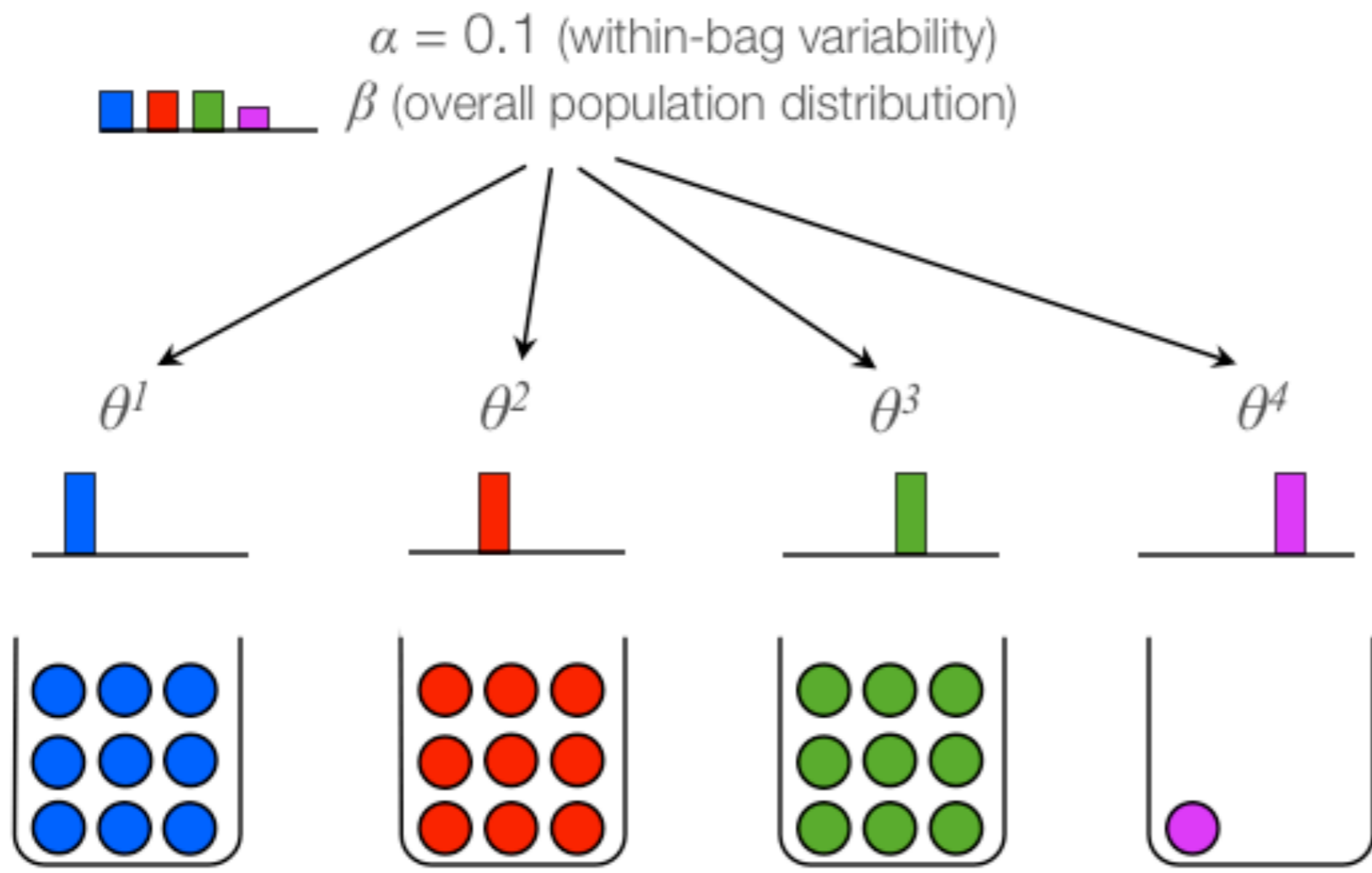
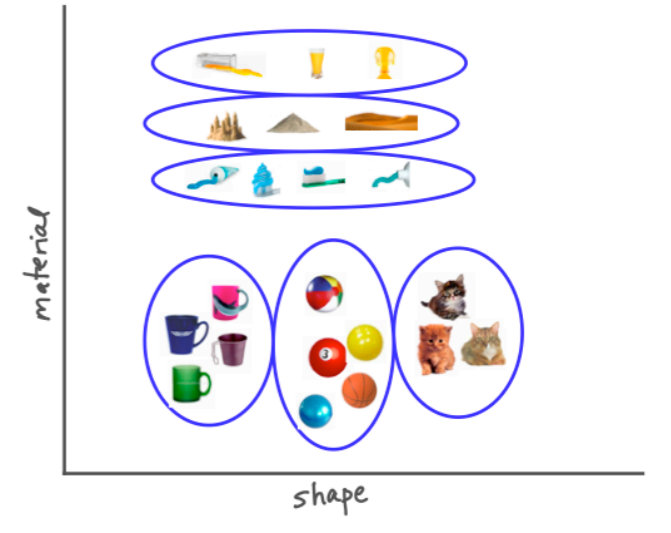


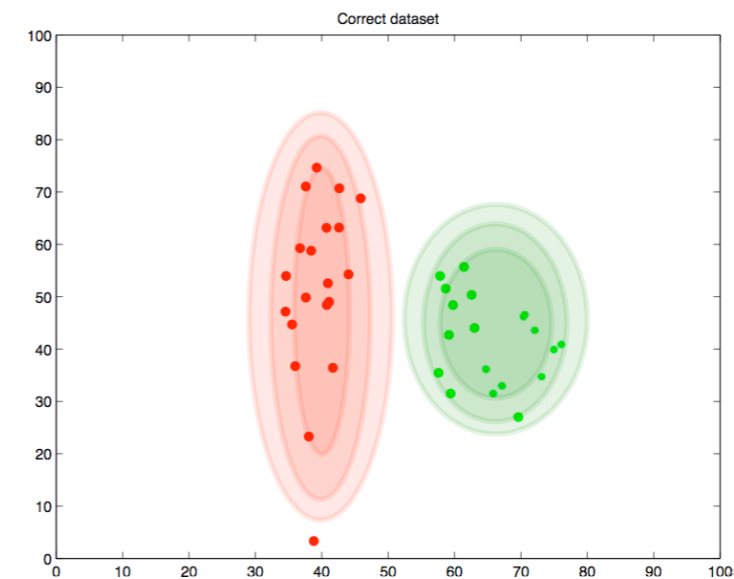
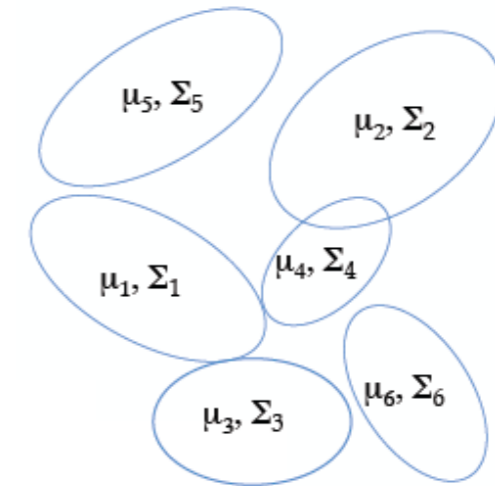
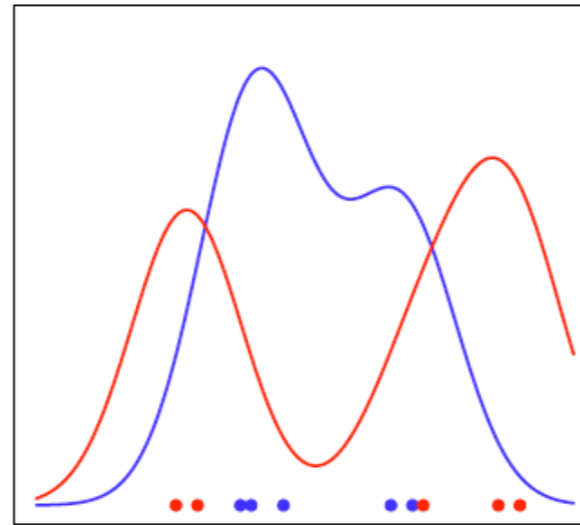
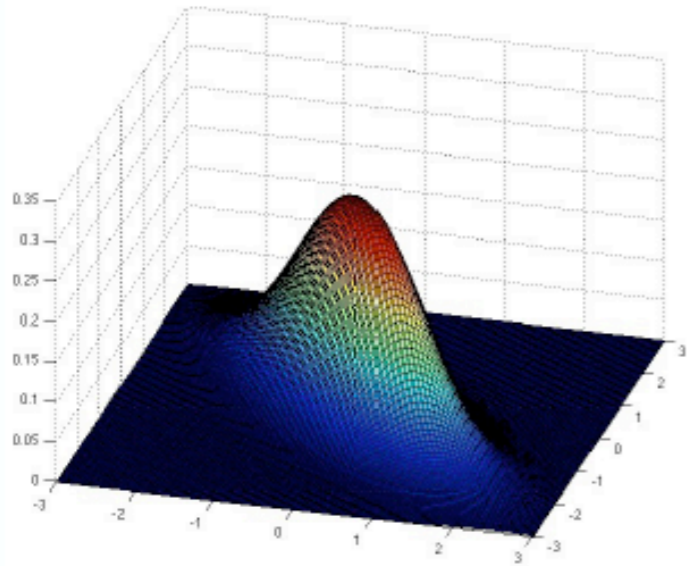
Computational Cognitive Science



Lecture 11: Higher order knowledge

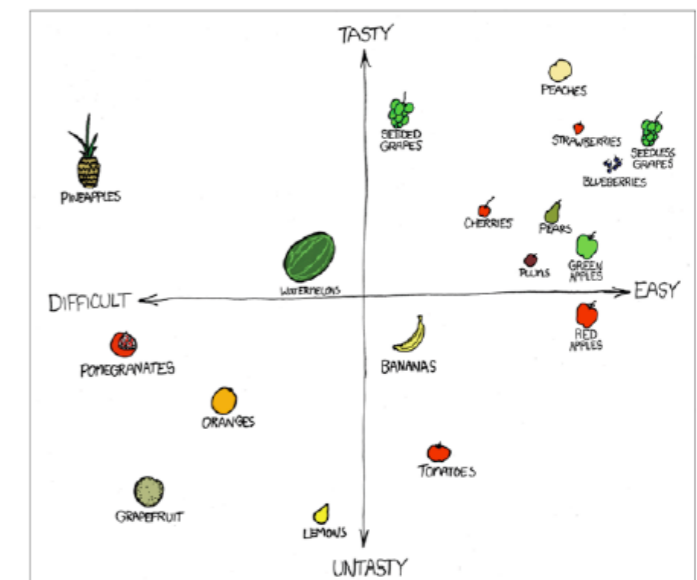
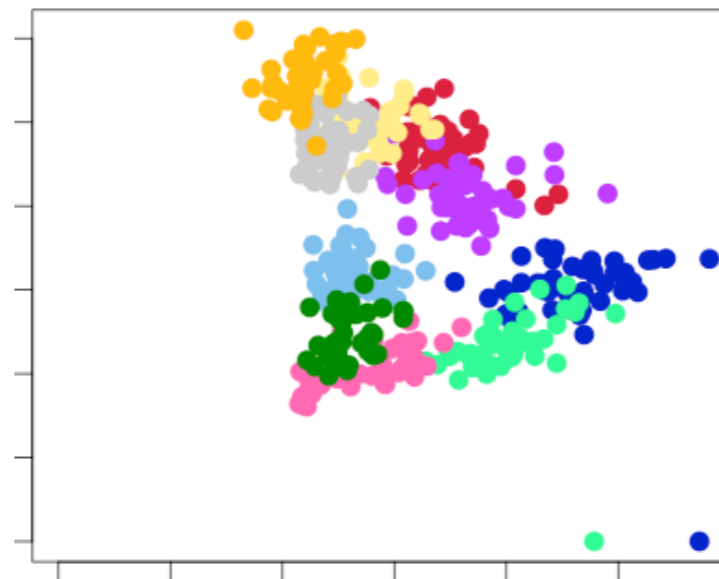
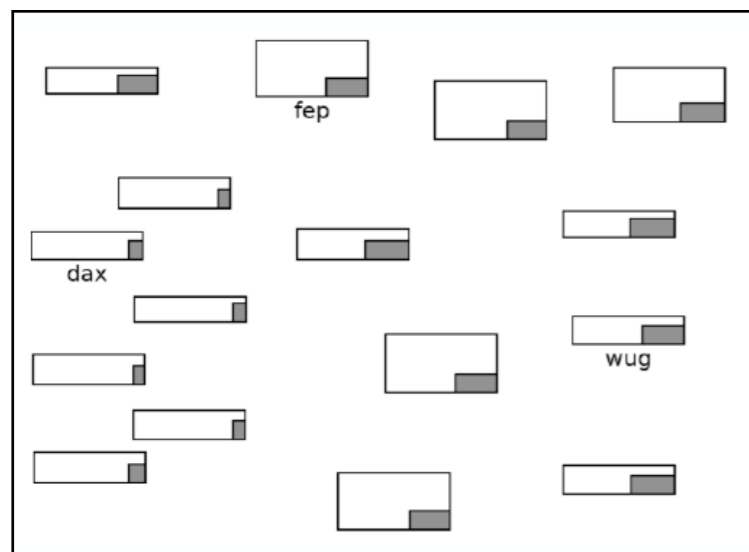
Classification so far

- ▶ We've seen a bunch of classification algorithms



Classification so far

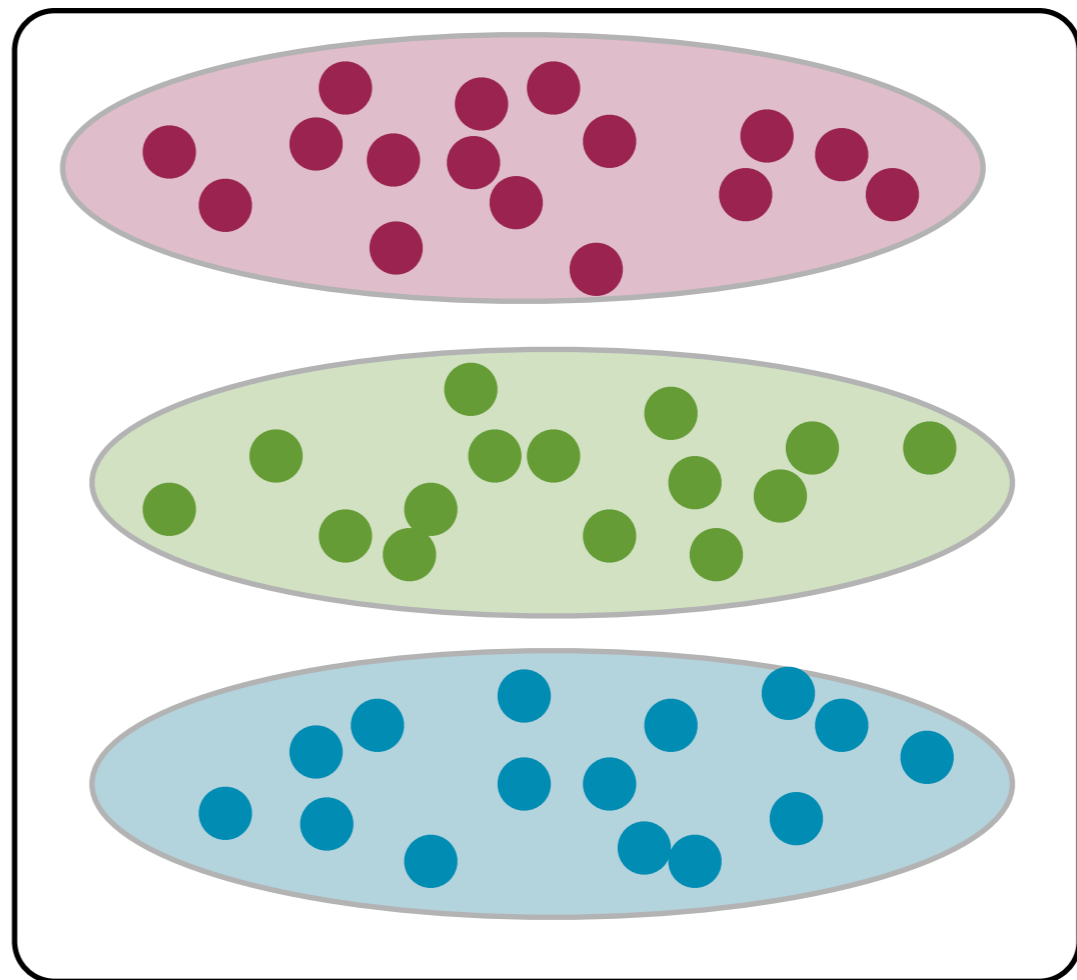
- ▶ We've seen a bunch of classification algorithms... and linked them people's tasks and performance



Classification so far

- ▶ But all of these algorithms don't incorporate a lot of additional knowledge that people do

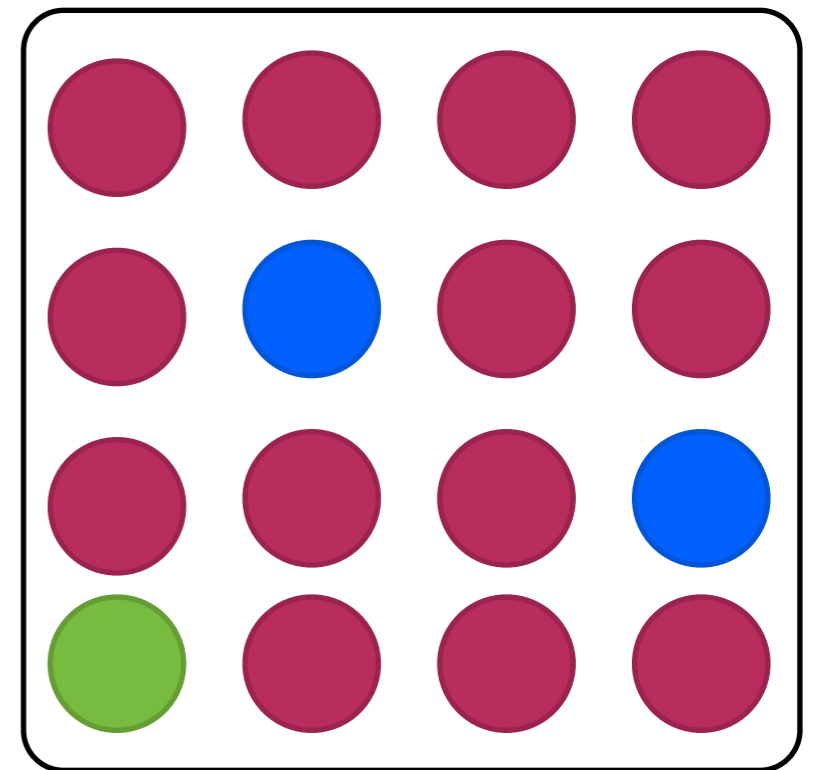
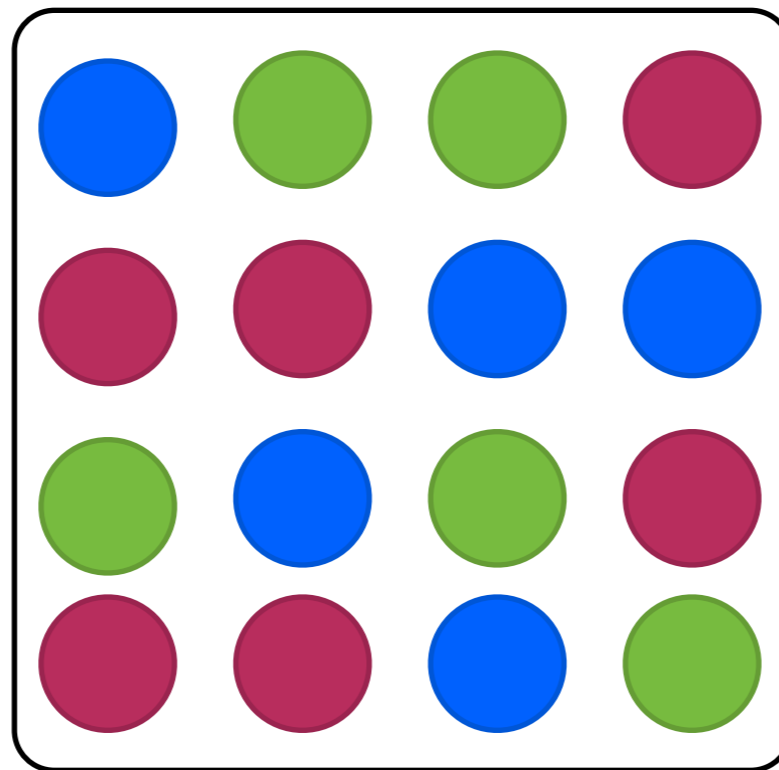
Higher-level
knowledge about
which dimensions
tend to matter



Classification so far

- ▶ But all of these algorithms don't incorporate a lot of additional knowledge that people do

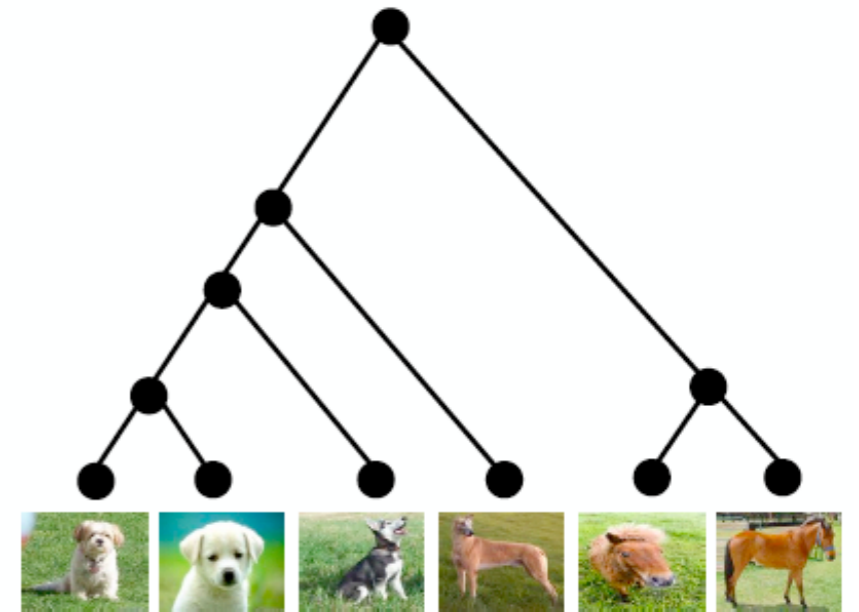
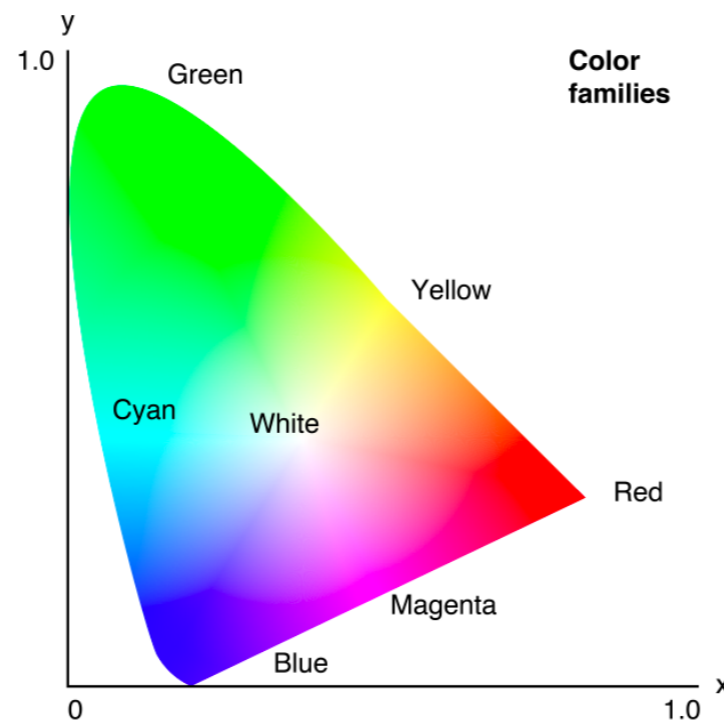
Higher-level
knowledge about
how things tend
to be distributed



Classification so far

- ▶ But all of these algorithms don't incorporate a lot of additional knowledge that people do

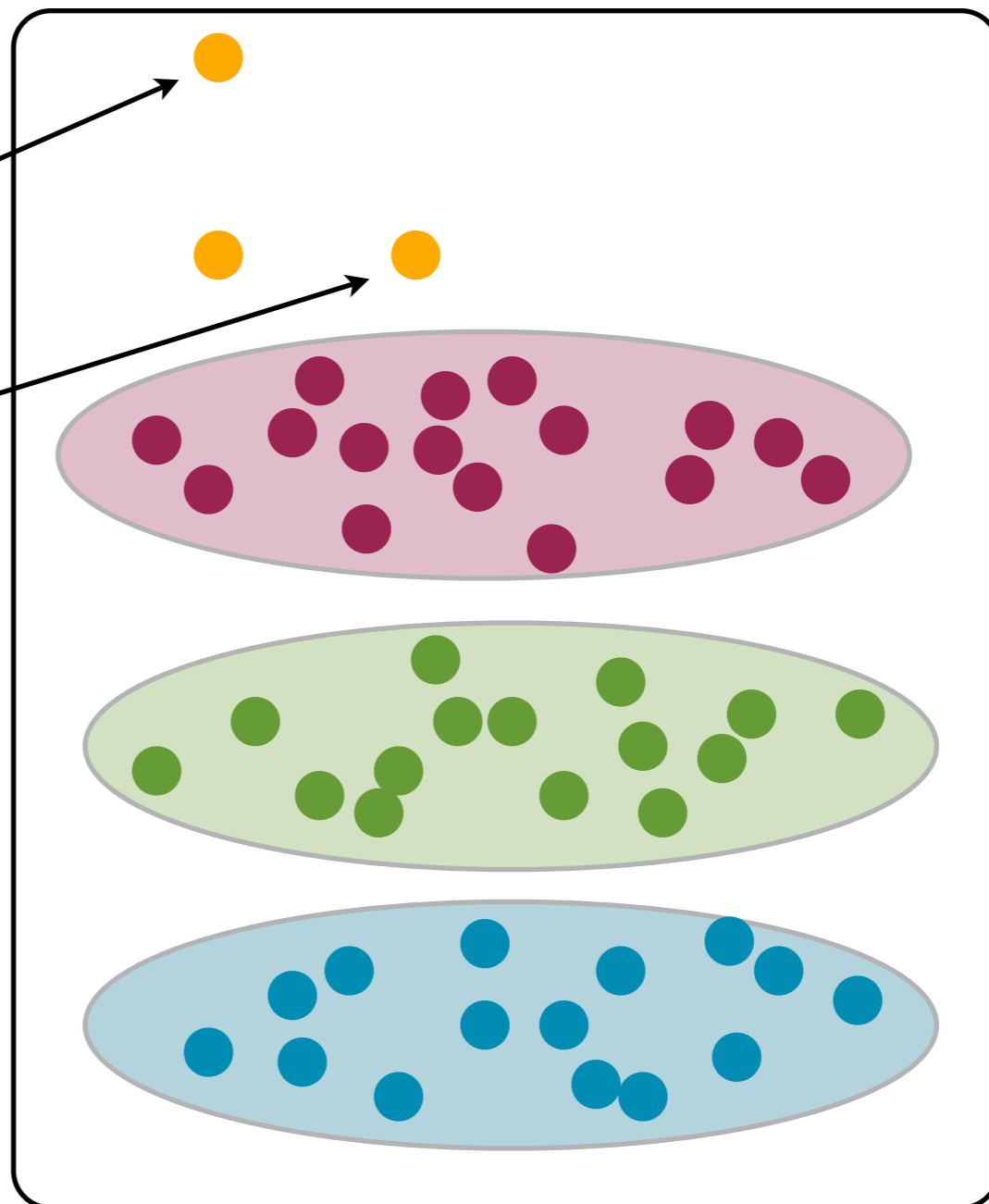
Higher-level knowledge about what the underlying structure is



Classification so far

- ▶ This knowledge licenses much more sensible inferences, which these algorithms cannot make

Which of these points is more likely to belong to a *new* category?



Classification so far

What kinds of higher-level knowledge do people make use of?

How can we understand such knowledge from a computational perspective? What kind of model could learn the same thing?

Lecture outline (next three lectures)

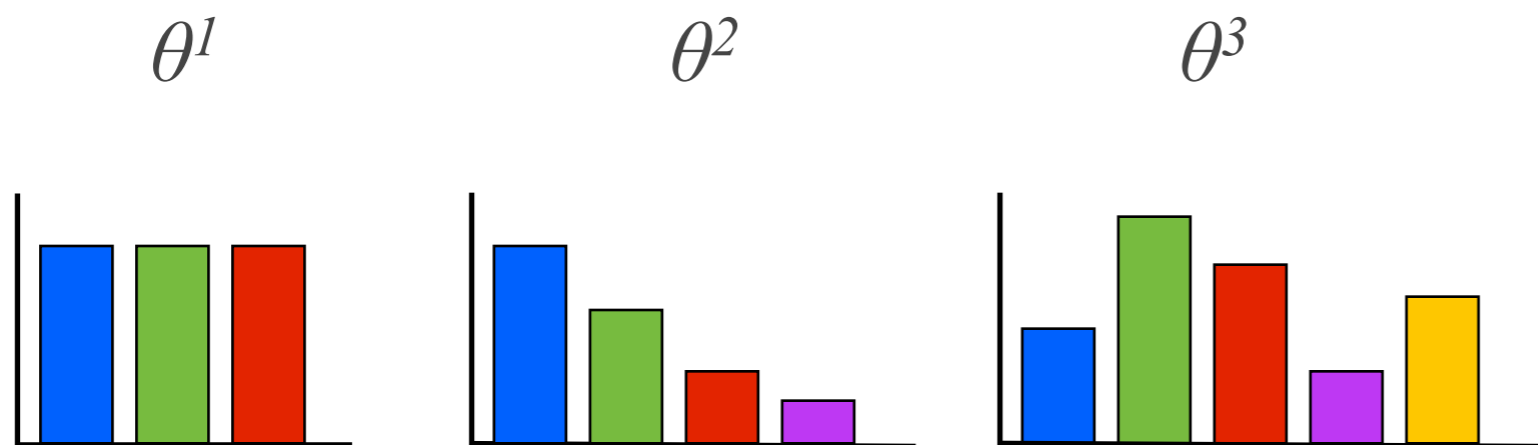
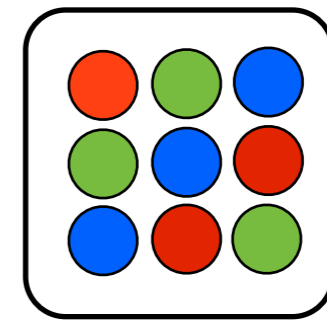
- ▶ Today: Learning about category variability
 - This kind of learning in children and adults
 - A model for this kind of learning
 - Limitations of this model
- ▶ Lecture 12: Learning about distributions of categories
 - This kind of learning in adults
 - Failure of current models
 - A model for this kind of learning
- ▶ Lecture 13: Learning about category structure
 - This kind of learning in people
 - A model for this kind of learning

Lecture outline (next three lectures)

- ➔ Today: Learning about category variability
 - ➔ This kind of learning in children and adults
 - A model for this kind of learning
 - Performance of this model
- ▶ Lecture 12: Learning about distributions of categories
 - This kind of learning in adults
 - Failure of current models
 - A model for this kind of learning
- ▶ Lecture 13: Learning about category structure
 - A model for this kind of learning
 - This kind of learning in people

What do we mean by higher-level knowledge?

Hypothesis space is the set of possible “true” distributions from which the colours in the box were drawn



Each hypothesis is one possible distribution θ

Put another way, each hypothesis is a theory about the nature of the true situation

What do we mean by higher-level knowledge?

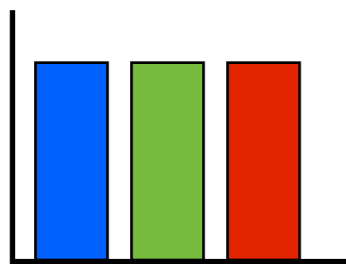
We can also form theories about the nature of the hypotheses themselves: these theories are called *overhypotheses*

bags tend to have multiple colours

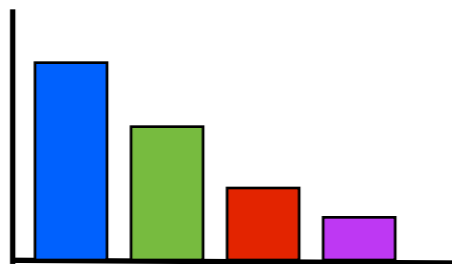
colours tend (not to) be uniformly distributed

overhypotheses

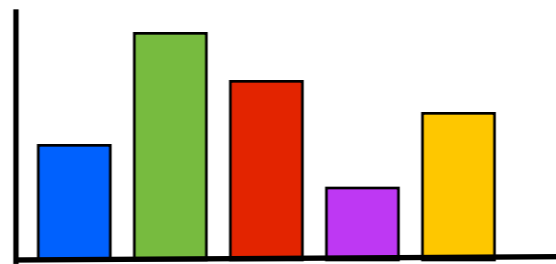
θ^1



θ^2



θ^3



hypotheses

Feature variability: one kind of overhypothesis

Simple categorisation:
figuring out which things “go together”



Feature variability: one kind of overhypothesis

Simple categorisation:
figuring out which things “go together”

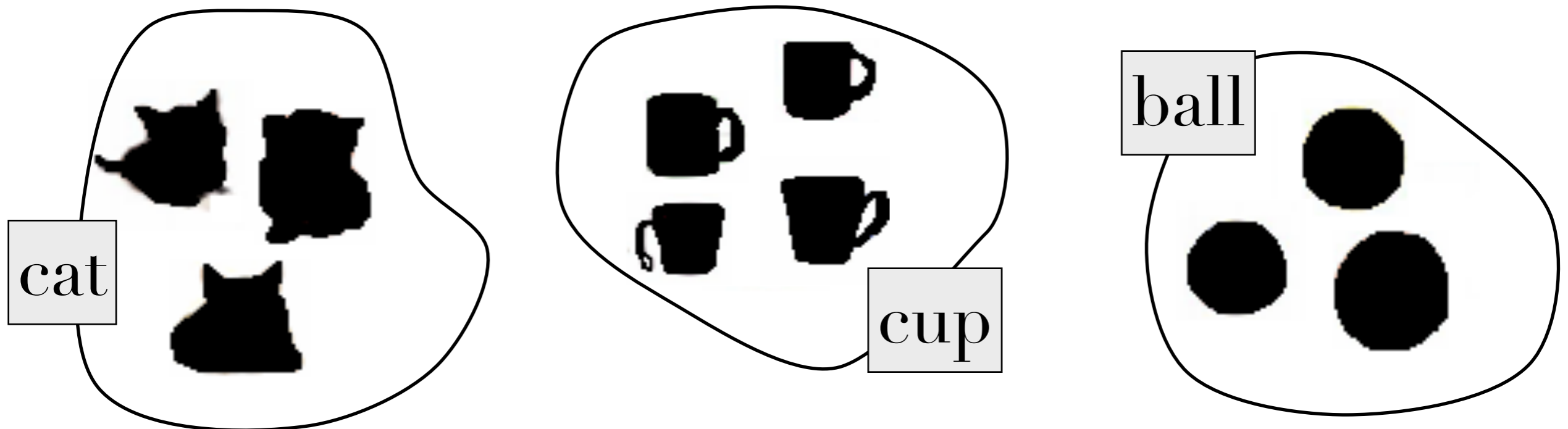
First-order inferences (hypothesis):
which categories individual items go in
cats are cat-shaped, cups are cup-shaped



Feature variability: one kind of overhypothesis

More complex categorisation:
learning what kind of rules or tendencies govern
how categories are organised

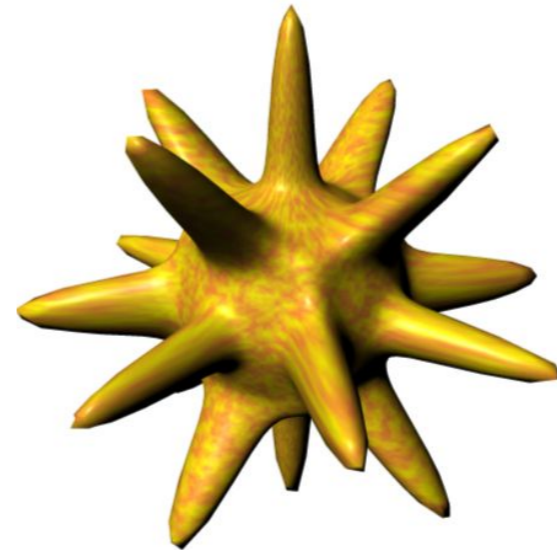
Second-order knowledge (over-hypothesis):
(solid) noun categories are organised by shape



Allows generalisation based on one data point!

Which of the two items on the right are daxes?

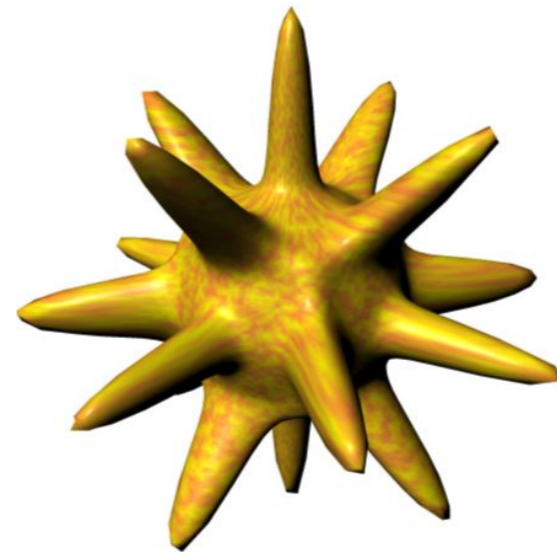
dax



Allows generalisation based on one data point!

The bias to categorise by shape is called the **shape bias**.

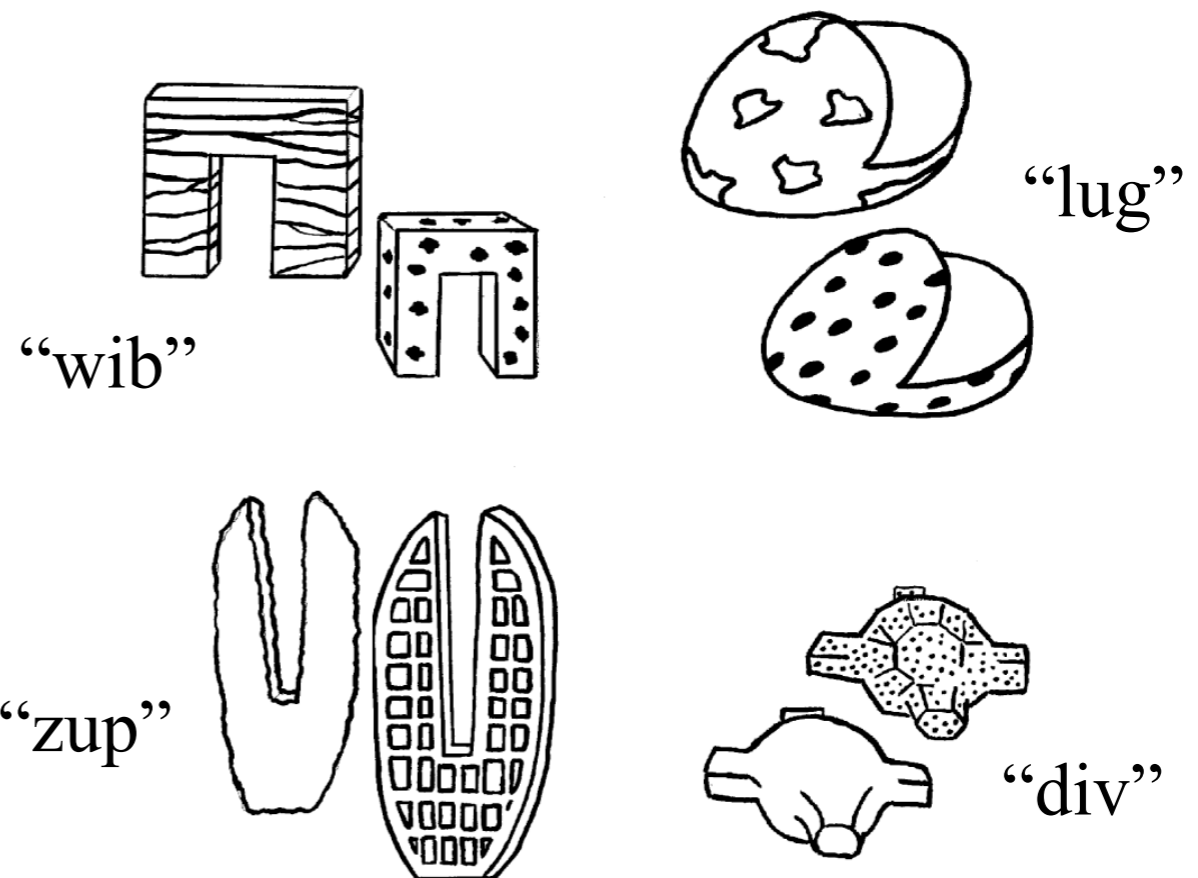
dax



The shape bias emerges at around two years

We know it is learned because it emerges more rapidly if children receive special training

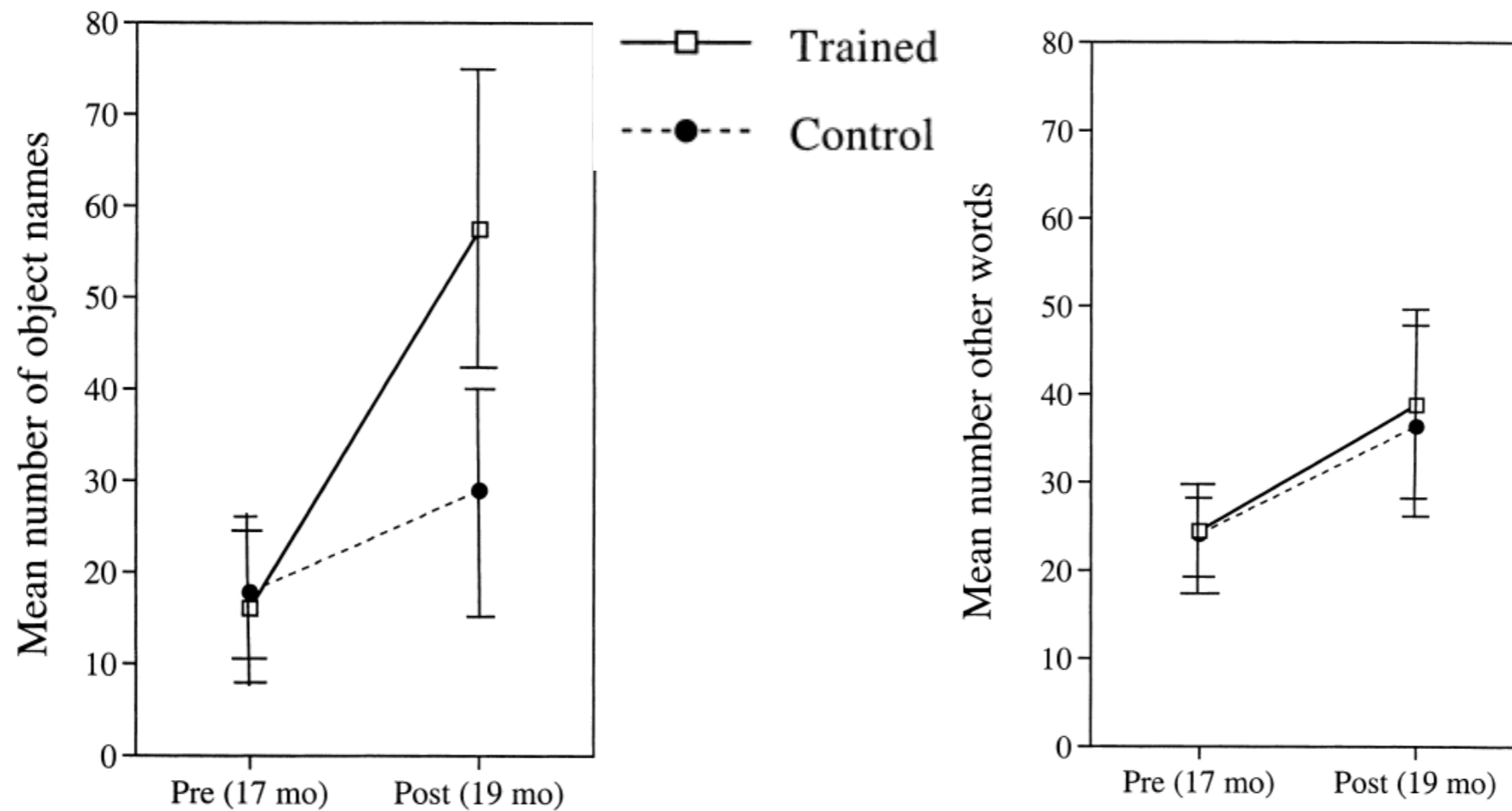
17-month-olds given labels for 4 artificial categories:



After 8 weeks of training, 19-month-olds show the shape bias.

The shape bias emerges at around two years

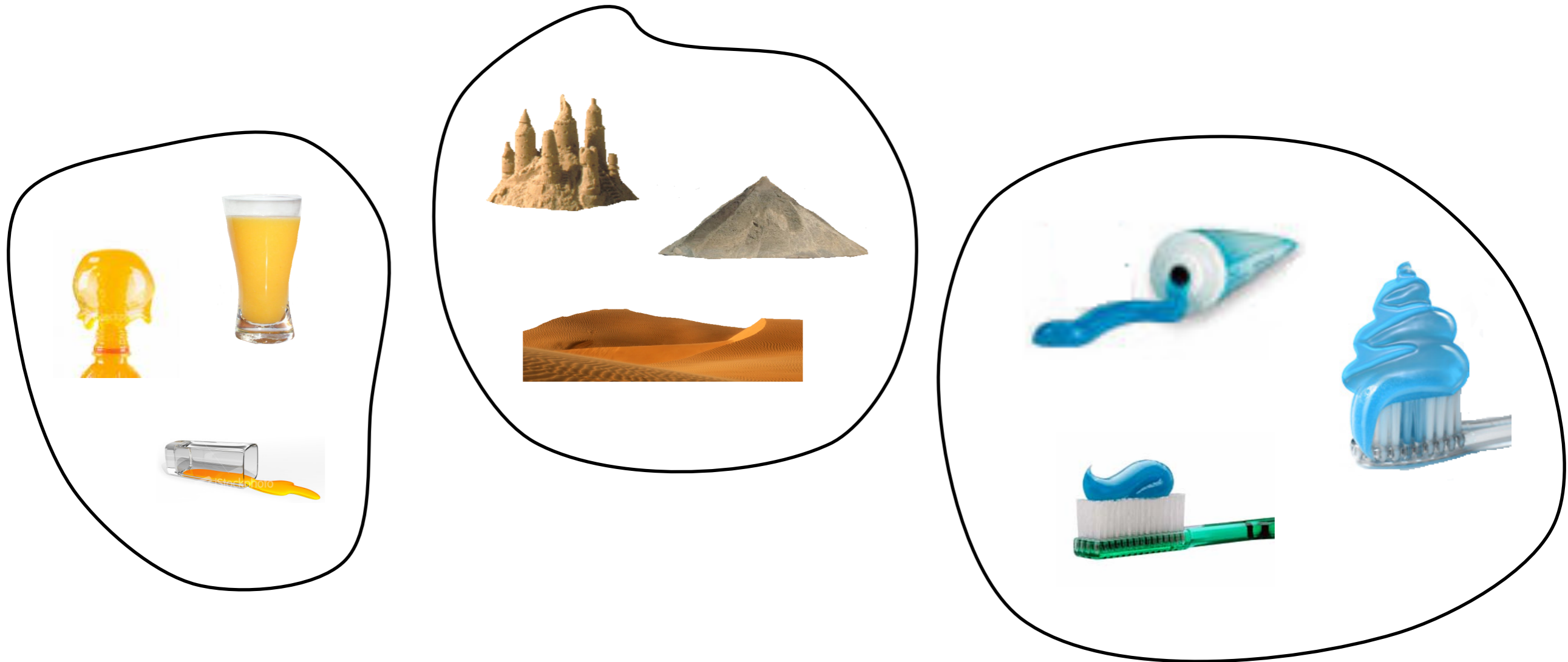
Plus, it helps them when learning other vocabulary!



Intuitively, children must be learning this overhypothesis about nouns based on the distribution of shape features in early words

It's not just the shape bias though..

Other categories are organised in different ways!

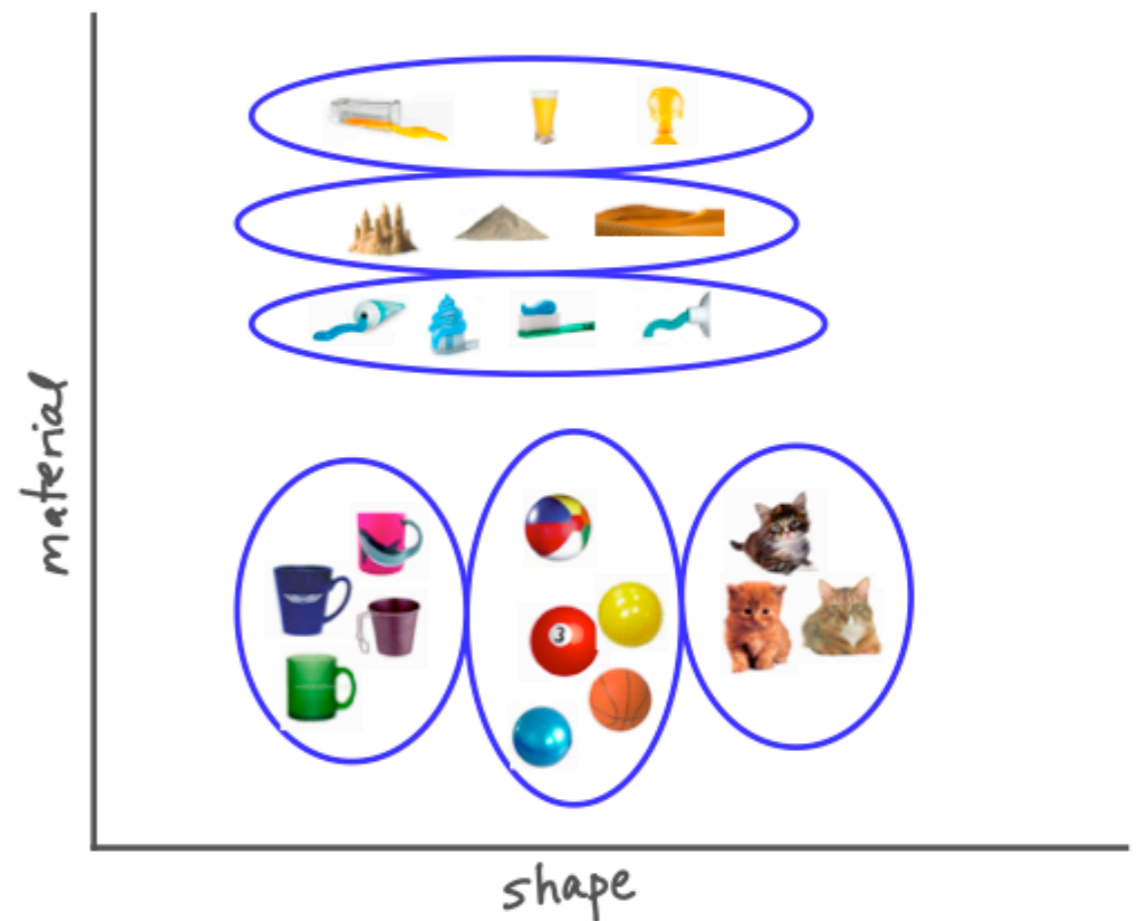


Non-solid substances tend to be organised by colour or texture,
not shape

It's not just the shape bias though..

Over developmental time, children learn multiple categories along with multiple ways of categorising them (“kinds”)

- ▶ 24 months: count nouns organised by shape
- ▶ 24 months: foods organised by colour
- ▶ 30 months: non-solids organised by texture
- ▶ 30 months: animates organised by shape and texture



How can we understand this learning?

What kind of model can -- like people -- learn on multiple levels of abstraction (both hypotheses and overhypotheses), with multiple kinds at once?

Lecture outline (next three lectures)

- ➔ Today: Learning about category variability
 - This kind of learning in children and adults
- ➔ A model for this kind of learning
 - Performance of this model
- ▶ Lecture 12: Learning about distributions of categories
 - This kind of learning in adults
 - Failure of current models
 - A model for this kind of learning
- ▶ Lecture 13: Learning about category structure
 - A model for this kind of learning
 - This kind of learning in people

There are many models

- ▶ ... at least for simpler types of abstract knowledge
- ▶ Non-Bayesian: models of **selective attention** (e.g., the GCM - equivalent to an exponential kernel classifier)
- ▶ However, these have difficulty learning multiple kinds at once

shape



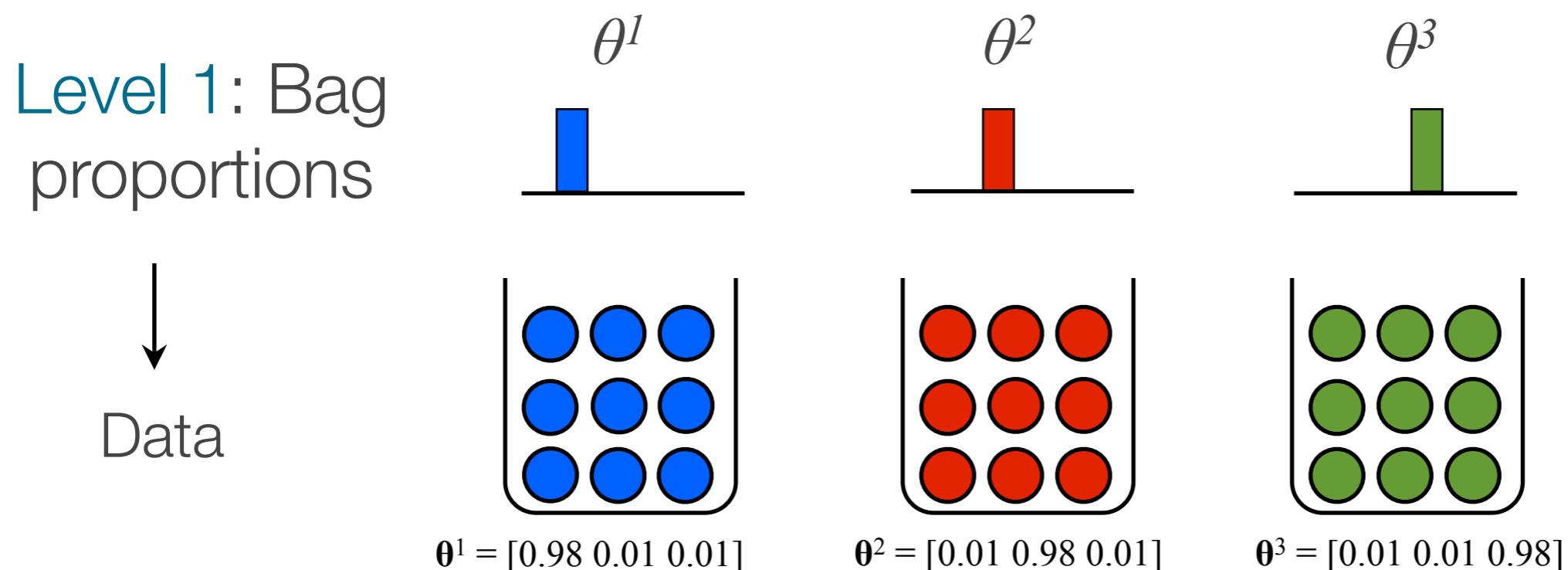
colour/texture



A Bayesian model for overhypothesis learning

- ▶ Visualise categories as bags of features; to keep things simple let's restrict ourselves to one kind and one feature
- ▶ First-order learning involves realising that category 1 is all blue, category 2 is all red, and so forth

We capture this raw data y with a multinomial distribution. In essence, each multinomial θ gives the probability distribution over each colour



A Bayesian model for overhypothesis learning

- ▶ Visualise categories as bags of features; to keep things simple let's restrict ourselves to one kind and one feature
- ▶ First-order learning involves realising that category 1 is all blue, category 2 is all red, and so forth

We capture this raw data y with a multinomial distribution. In essence, each multinomial θ gives the probability distribution over each colour

$$\mathbf{y} \sim \text{Multinomial}(\theta)$$

$$p(\mathbf{y}|\theta) = \begin{cases} \frac{n!}{y_1! \dots y_k!} \theta_1^{y_1} \dots \theta_k^{y_k} & \text{when } \sum_{i=1}^k y_i = n \\ 0 & \text{otherwise} \end{cases}$$

Here, n is the number of balls, k is the number of feature values there are in total, and y_i is the number of balls with that feature value

We need a prior!

$$p(\mathbf{y}|\theta) = \begin{cases} \frac{n!}{y_1! \dots y_k!} \theta_1^{y_1} \dots \theta_k^{y_k} & \text{when } \sum_{i=1}^k y_i = n \\ 0 & \text{otherwise} \end{cases}$$

However, in order to calculate $p(\theta|\mathbf{y})$, which is what we need to be able to go from the raw data \mathbf{y} to the inferred category features, we need a prior over those features.

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$$

A natural prior to use is called the [Dirichlet](#) prior.

* The reason it is natural is that when combined with the multinomial, the result is still a multinomial, so the math is a lot easier. (This property is called *conjugacy*). Also, it's straightforwardly interpretable

We need a prior!

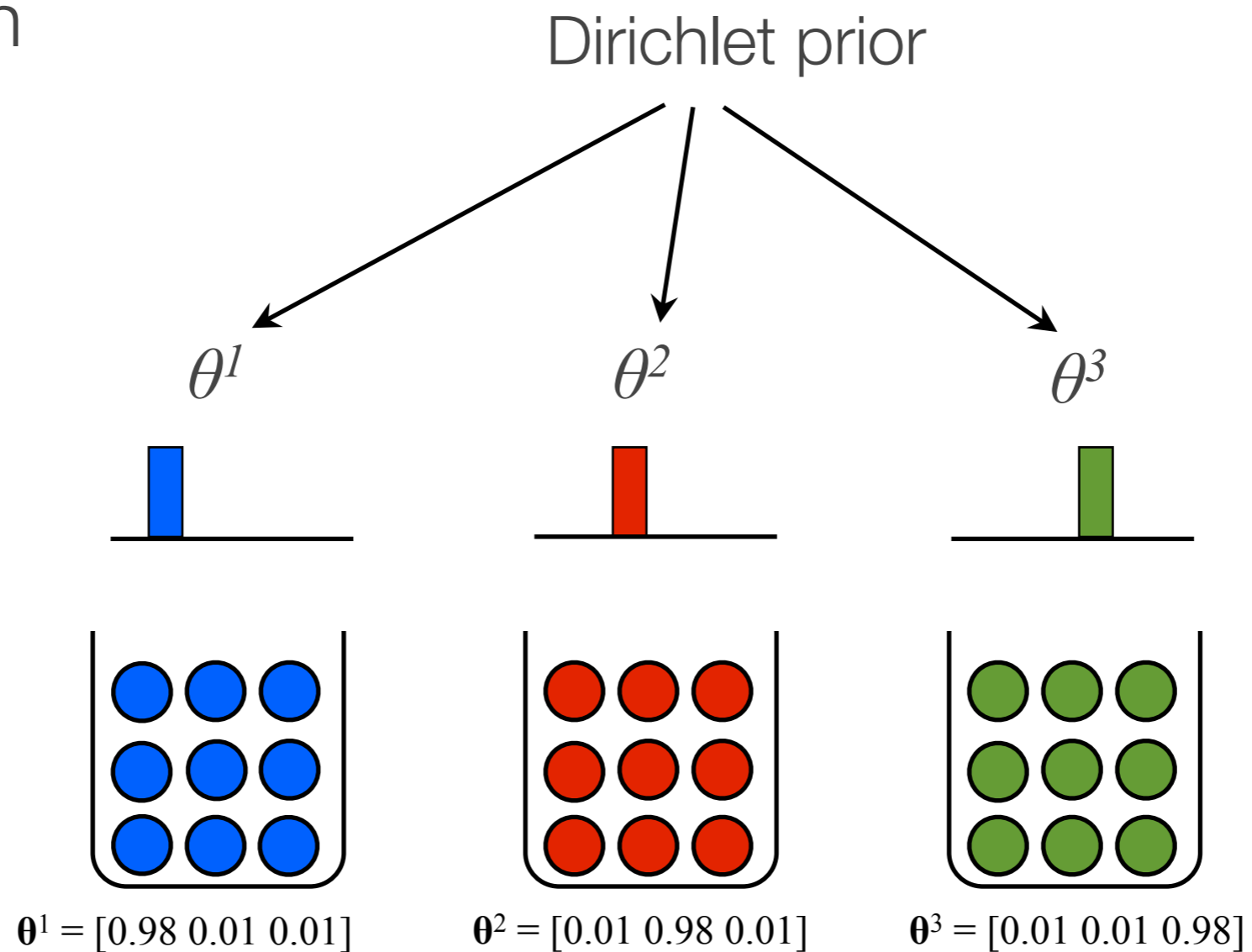
Level 2: Bags in general



Level 1: Bag proportions



Data



We need a prior!

A Dirichlet distribution consists of two elements:

α = concentration parameter

β = base distribution

$$\theta^j \sim \text{Dirichlet}(\alpha, \beta)$$

distribution of features
amongst the entire
dataset

tendency for features
to be uniform in any
one category

We need a prior!

If you make make prior choices about what α and β should be, you end up with a standard category-learning model (similar to the ones we have already discussed, with multinomials instead of Gaussians)

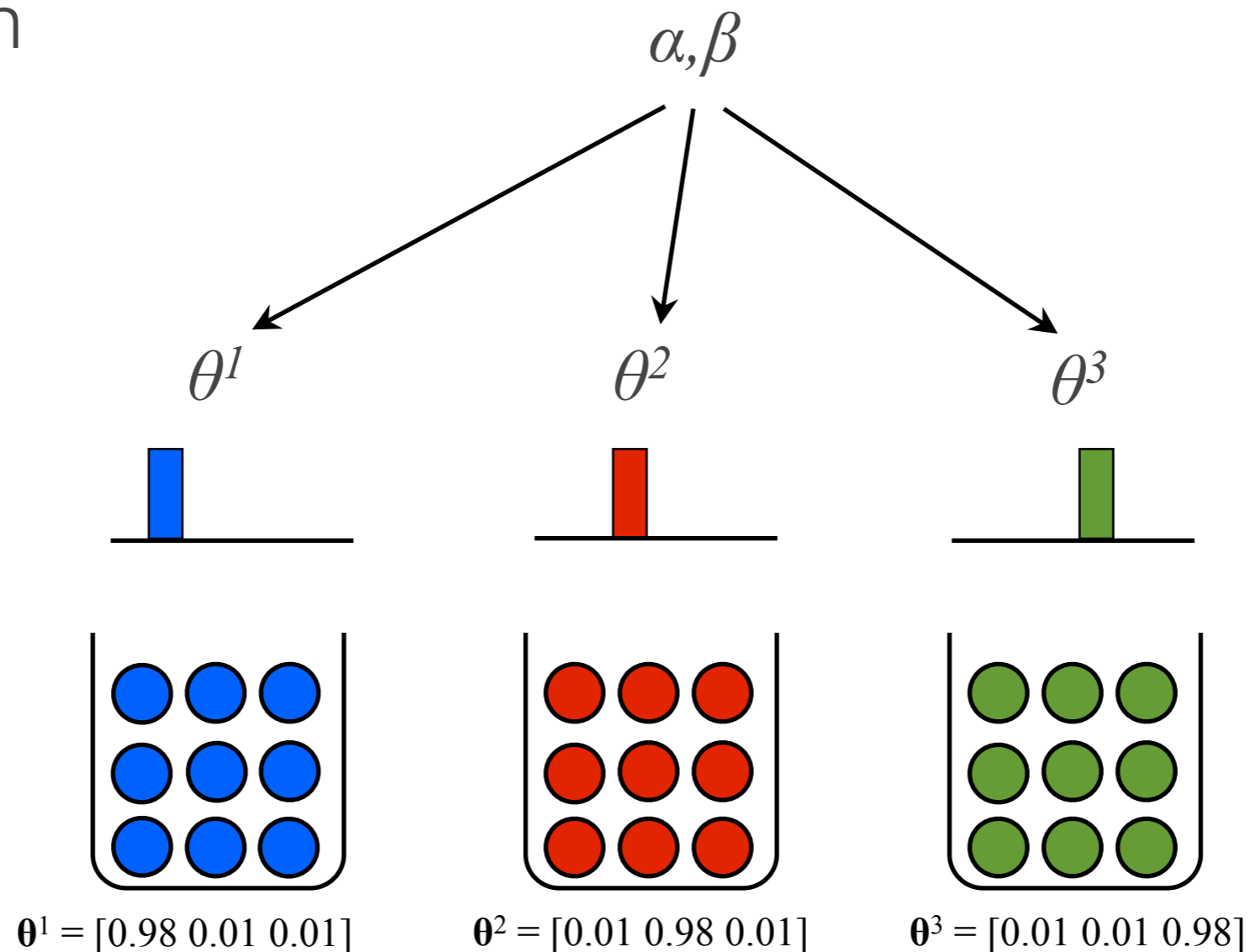
Level 2: Bags in general



Level 1: Bag proportions



Data



We need a prior!

However, such a model cannot learn based on this data that categories tend to be uniform (or not). As a result, it cannot generalise correctly given new data (unless that is built into the prior).

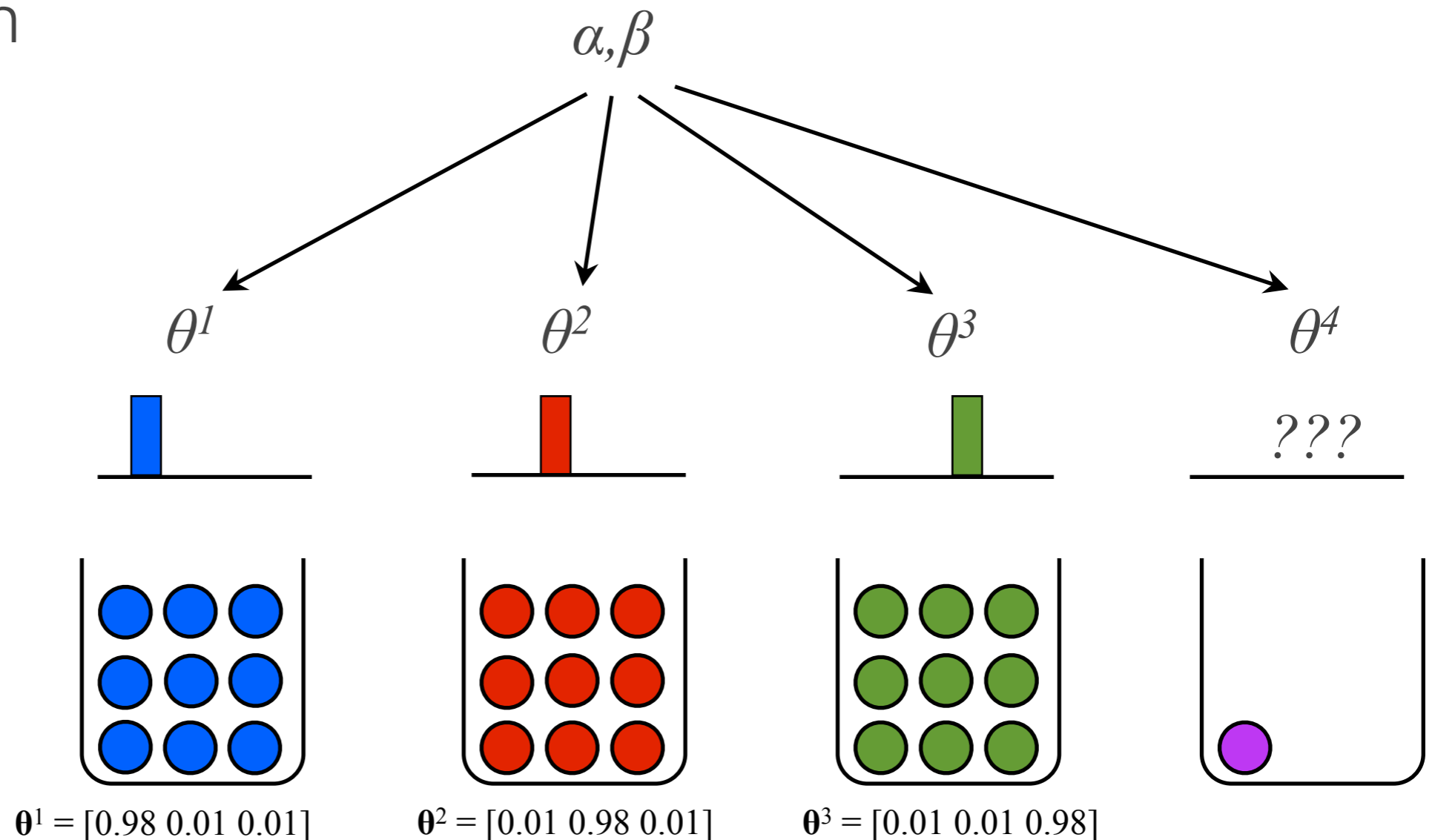
Level 2: Bags in general



Level 1: Bag proportions



Data



We need a prior!

However, such a model cannot learn based on this data that categories tend to be uniform (or not). As a result, it cannot generalise correctly given new data (unless that is built into the prior).

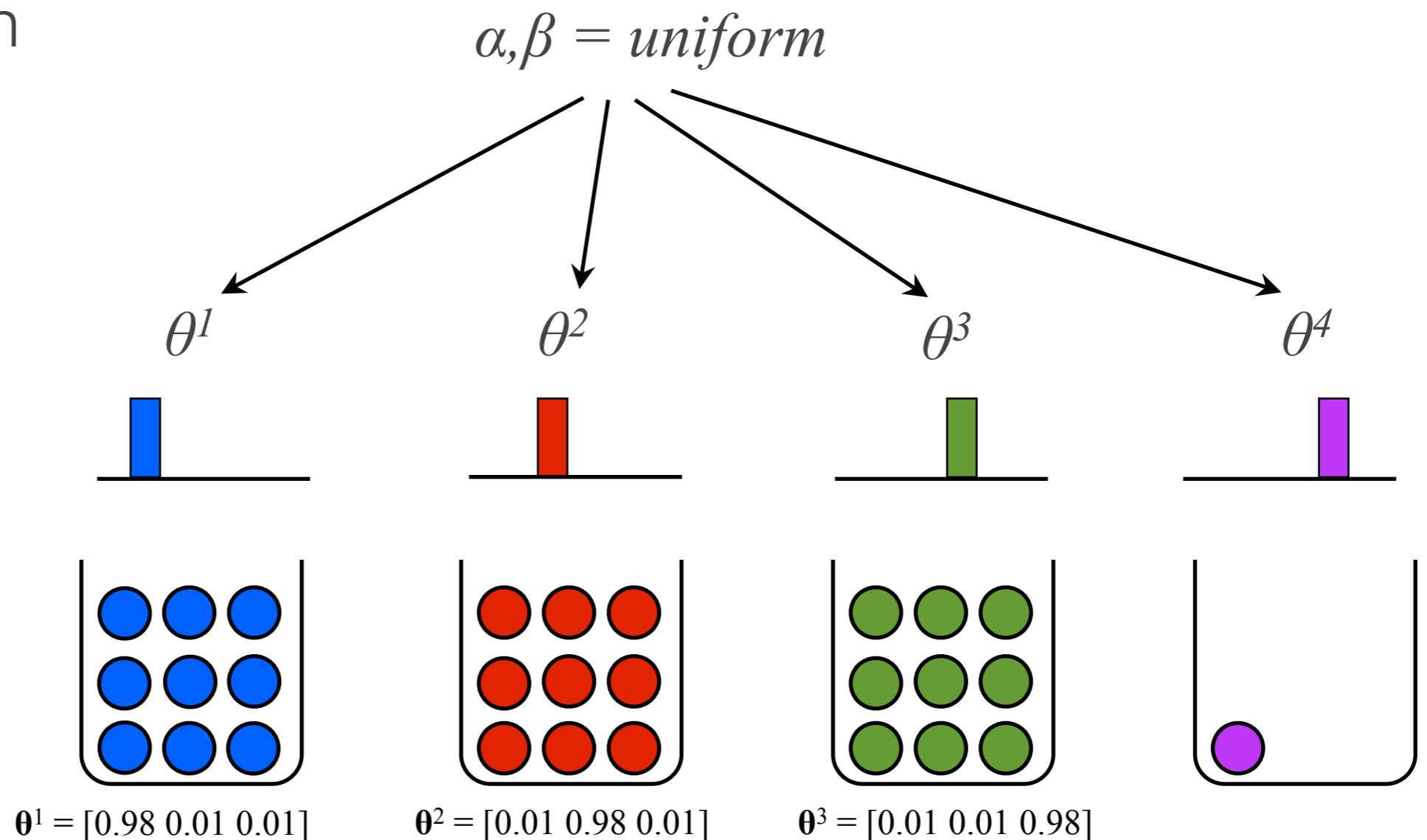
Level 2: Bags in general



Level 1: Bag proportions



Data



We need a prior!

However, such a model cannot learn based on this data that categories tend to be uniform (or not). As a result, it cannot generalise correctly given new data (unless that is built into the prior).

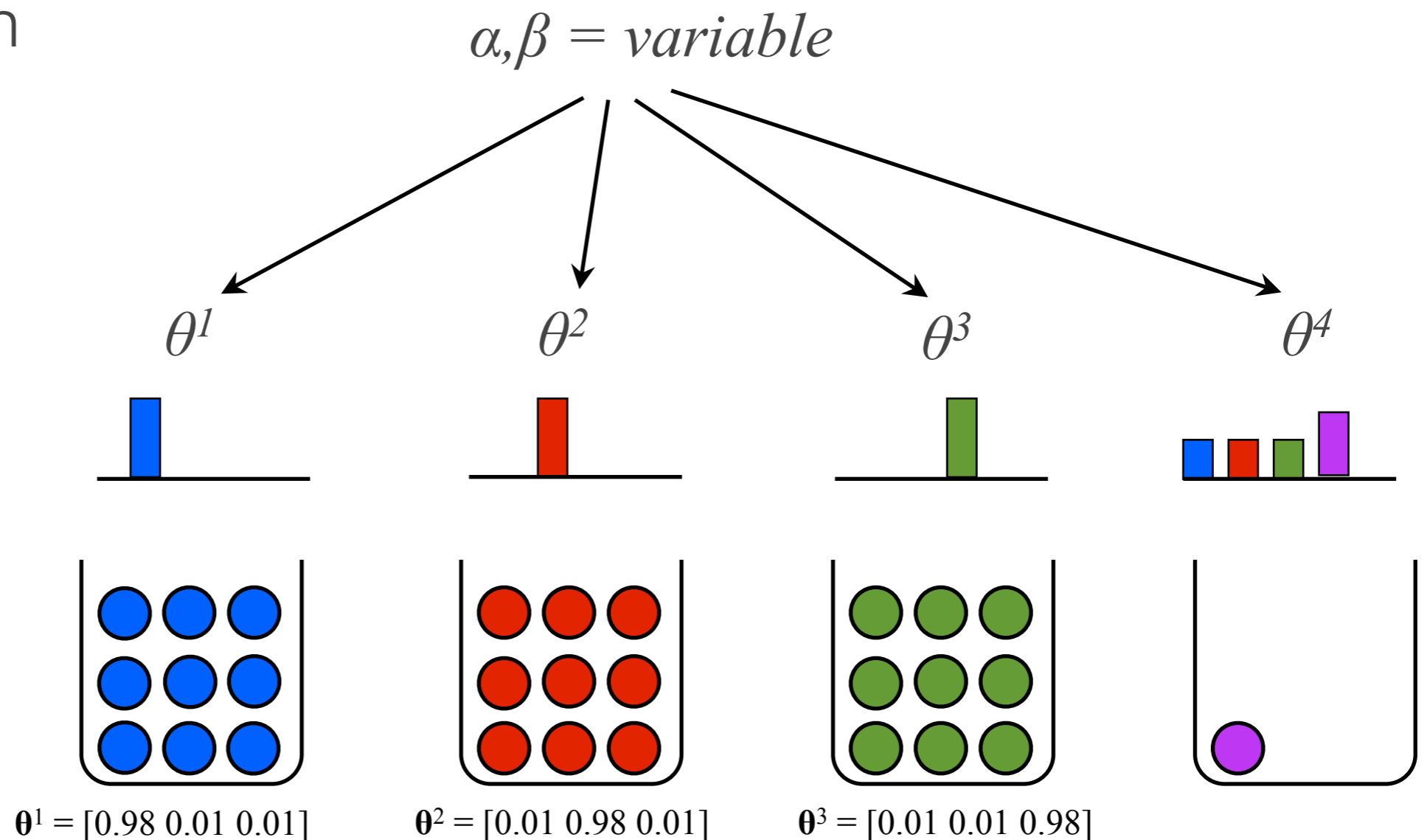
Level 2: Bags in general



Level 1: Bag proportions



Data



We need a prior!

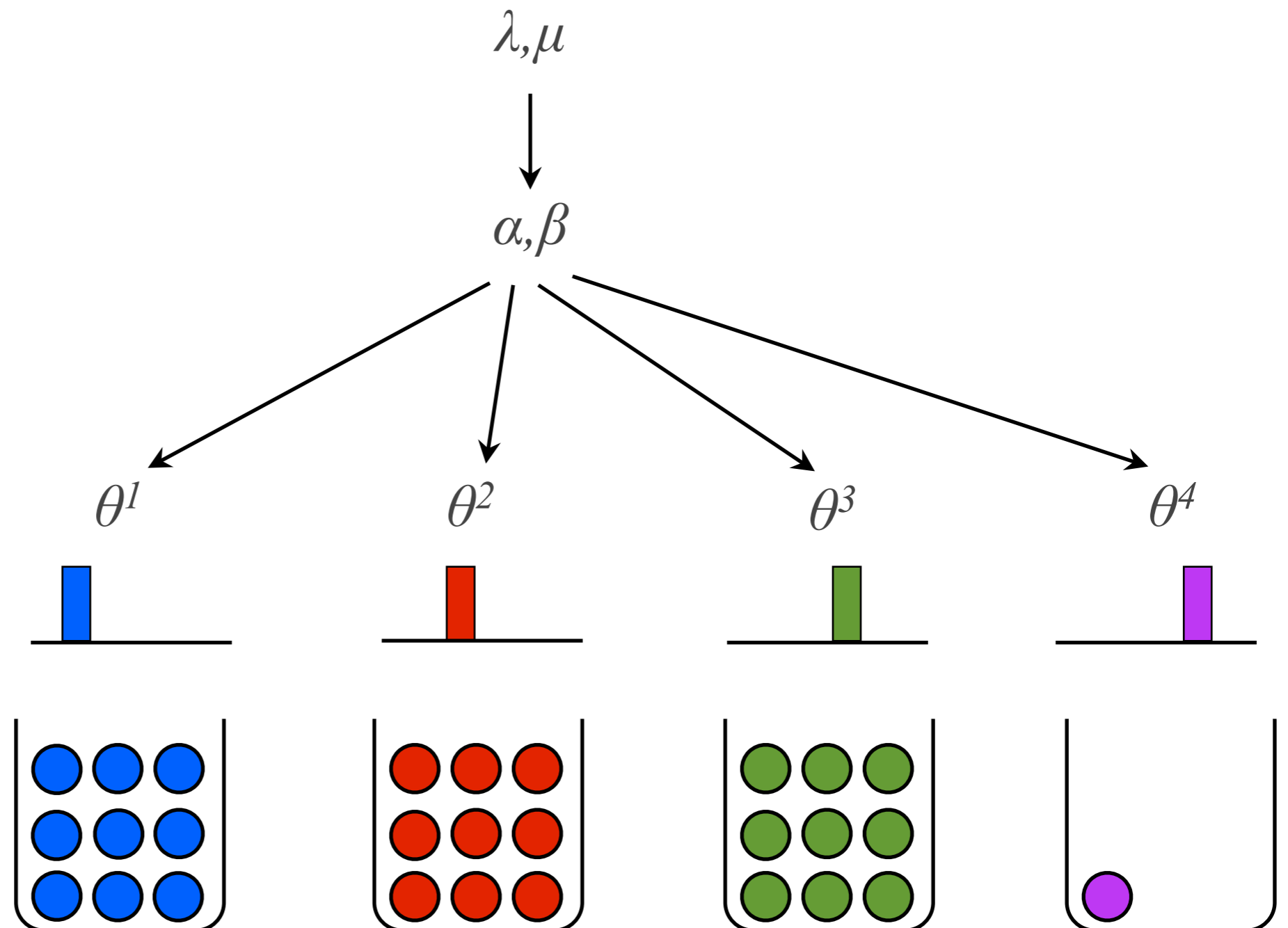
What we want is to *learn* this knowledge by putting a prior on our prior

Level 3: Prior about bags in general

Level 2: Bags in general

Level 1: Bag proportions

Data

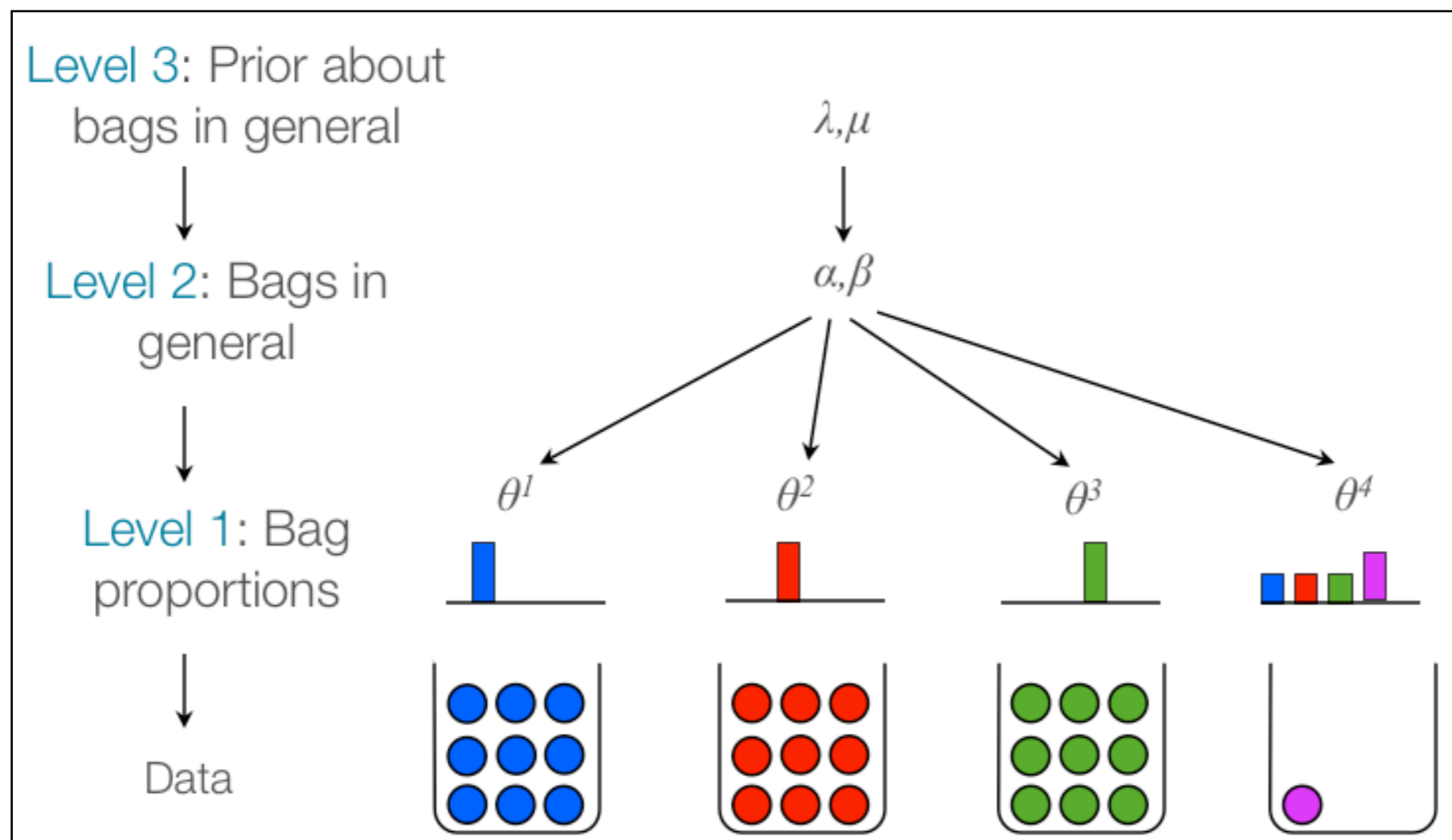


The full model

α is a scalar $\longrightarrow \alpha \sim \text{Exponential}(\lambda)$

The Dirichlet is conjugate to the Dirichlet $\longrightarrow \beta \sim \text{Dirichlet}(\mu)$

This is called a **hierarchical Bayesian model**, and in principle you can keep adding additional levels however much you want



The parameters on the higher levels are called **hyperparameters**

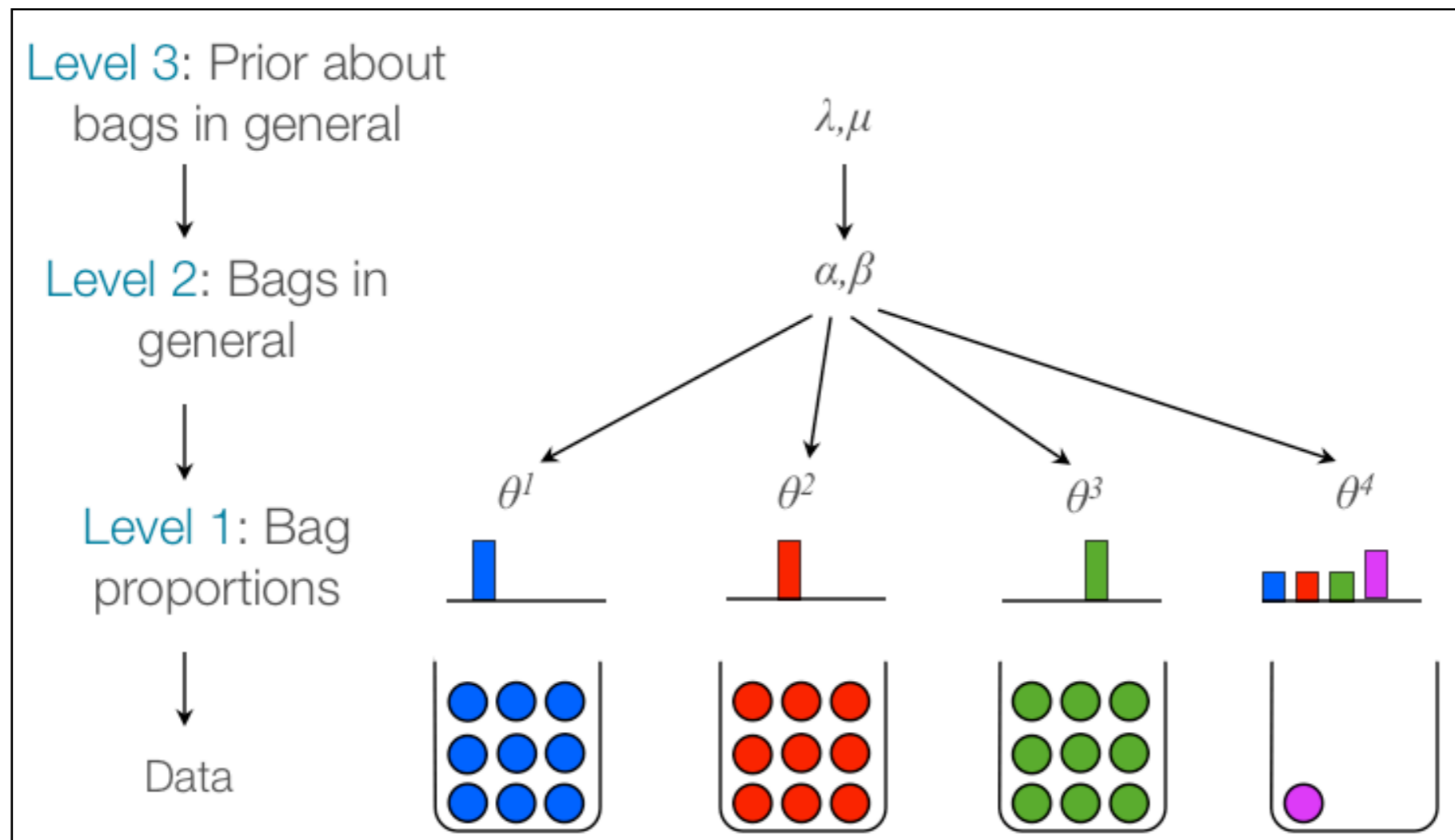
The full model

Make inferences about the category-specific distributions by:

$$p(\theta^j | \mathbf{y}) = \int_{\alpha, \beta, \lambda, \mu} p(\theta^j | \alpha, \beta, \lambda, \mu) p(\alpha, \beta, \lambda, \mu | \mathbf{y}) d\alpha d\beta d\lambda d\mu$$

Simultaneously inferring:

$$p(\alpha, \beta, \lambda, \mu | \mathbf{y}) \propto p(\mathbf{y} | \alpha, \beta) p(\alpha | \lambda) p(\beta | \mu) p(\lambda) p(\mu)$$



This model can learn which features “matter”

Level 3: Prior about bags in general



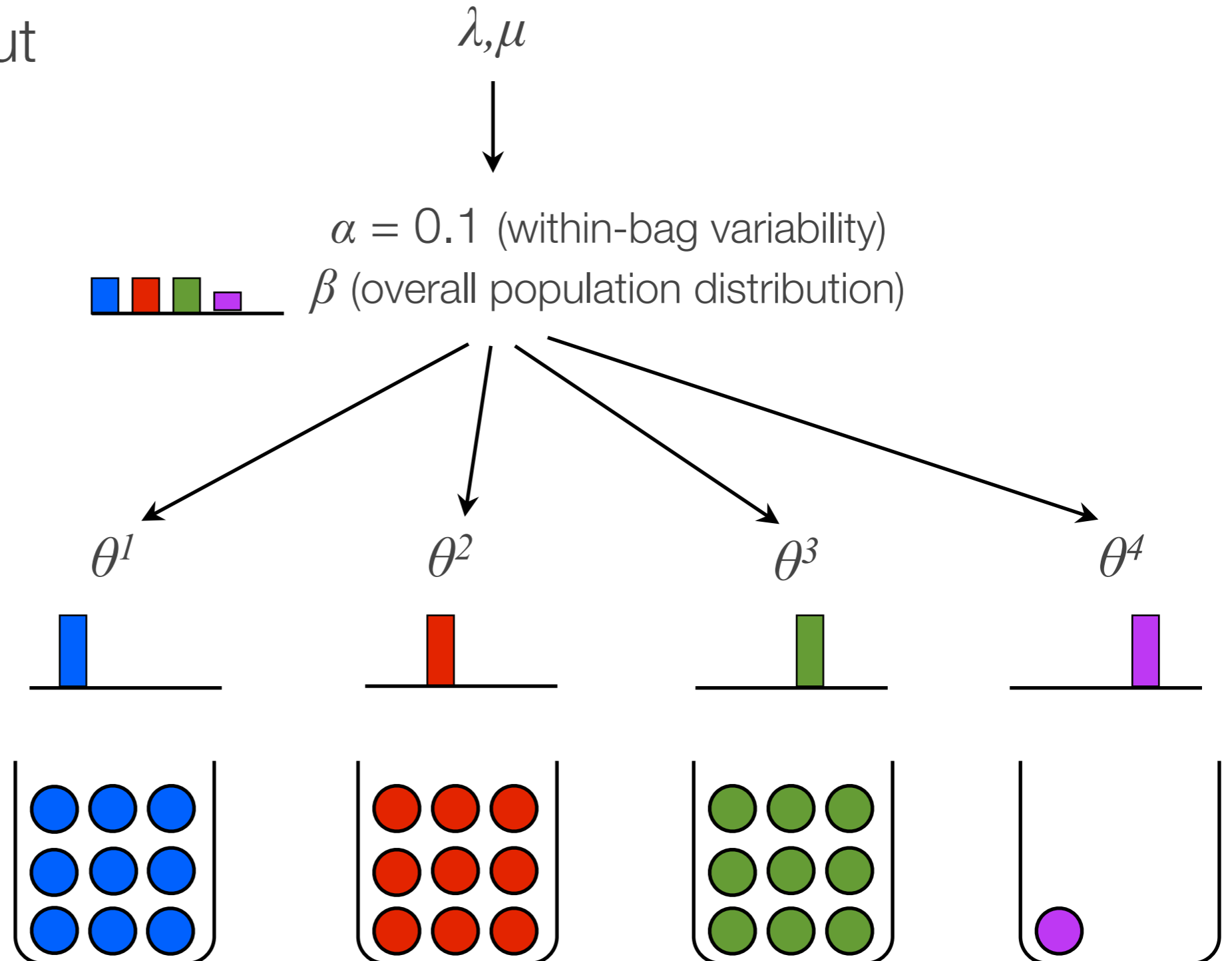
Level 2: Bags in general



Level 1: Bag proportions



Data



This model can learn which features “matter”

Level 3: Prior about bags in general



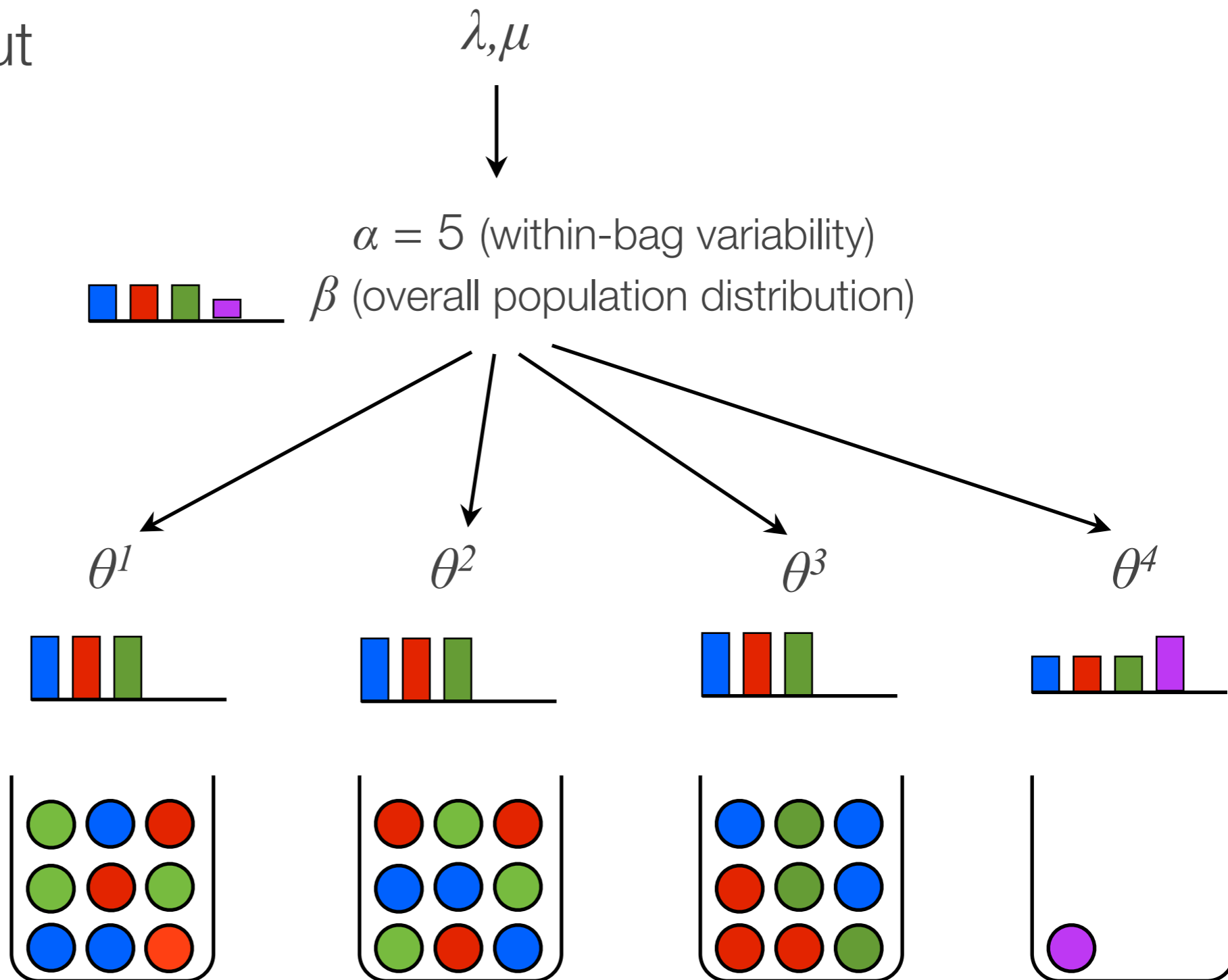
Level 2: Bags in general



Level 1: Bag proportions



Data



This model can learn which features “matter”

...but it still can't learn multiple different overhypotheses for multiple different kinds

shape



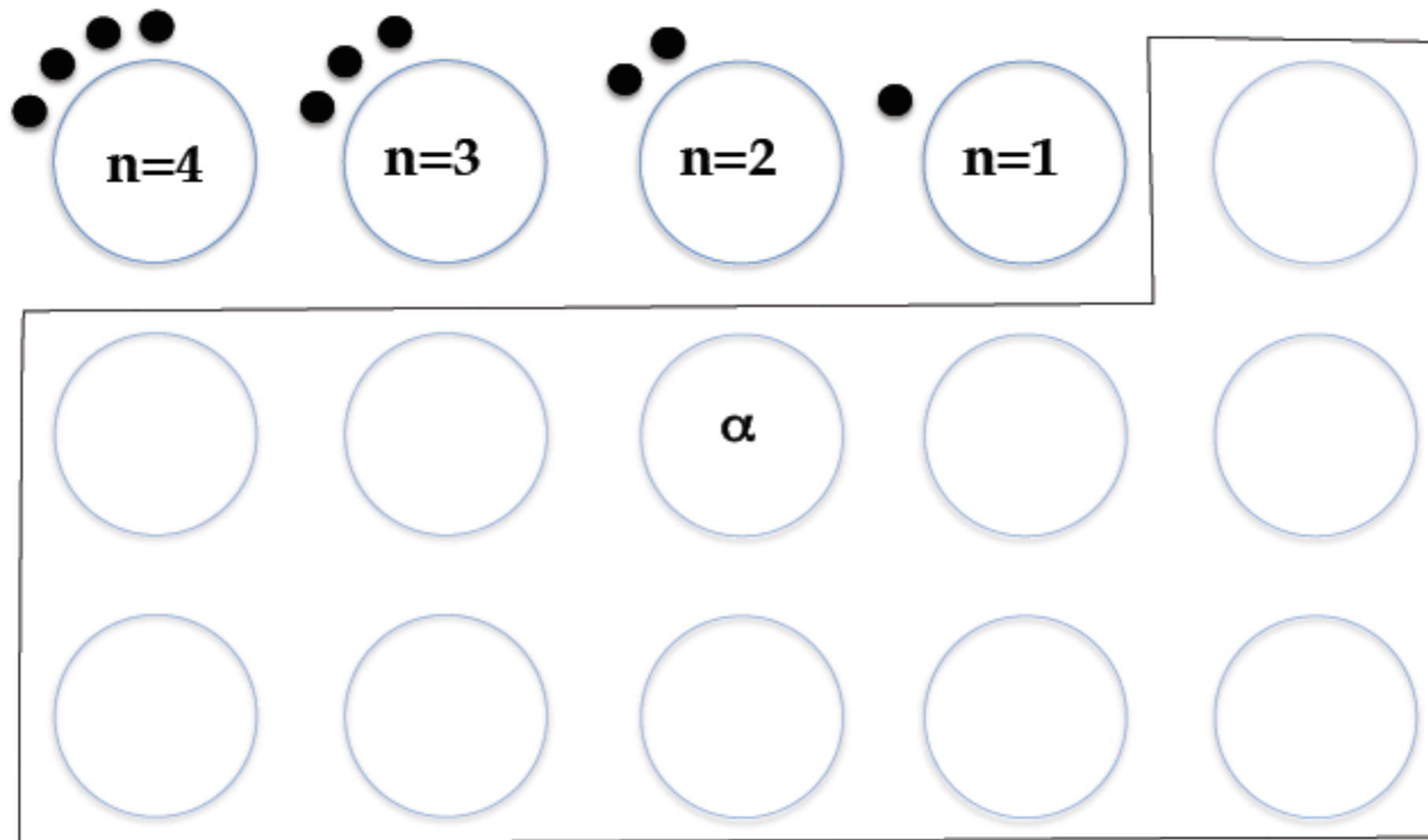
colour/texture



This model cannot do so; it can only learn one overhypothesis at a time. What we want is to be able to cluster items in different kinds, but have a prior that favours fewer kinds.

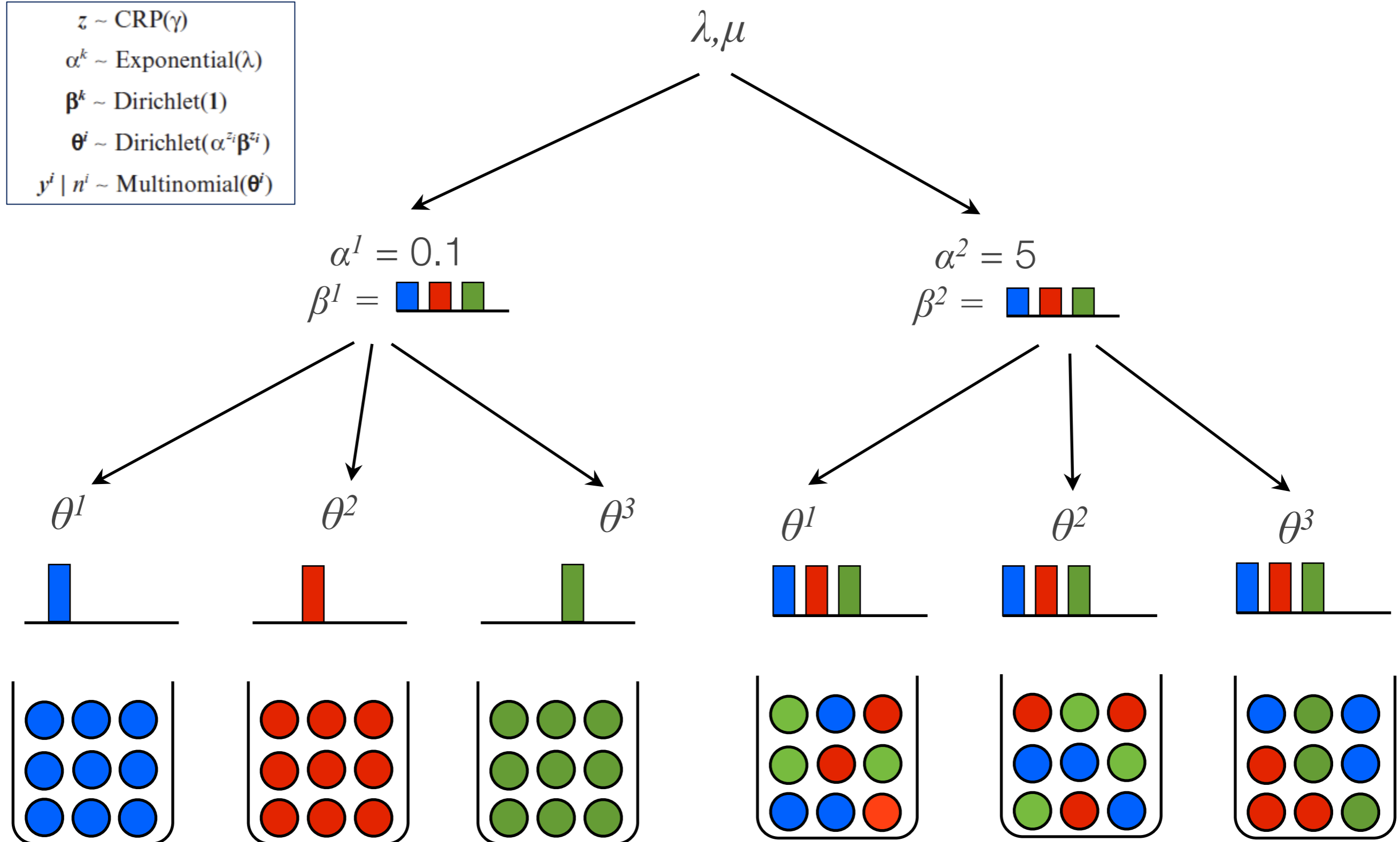
What kind of prior might that be?

Well, really, what else?

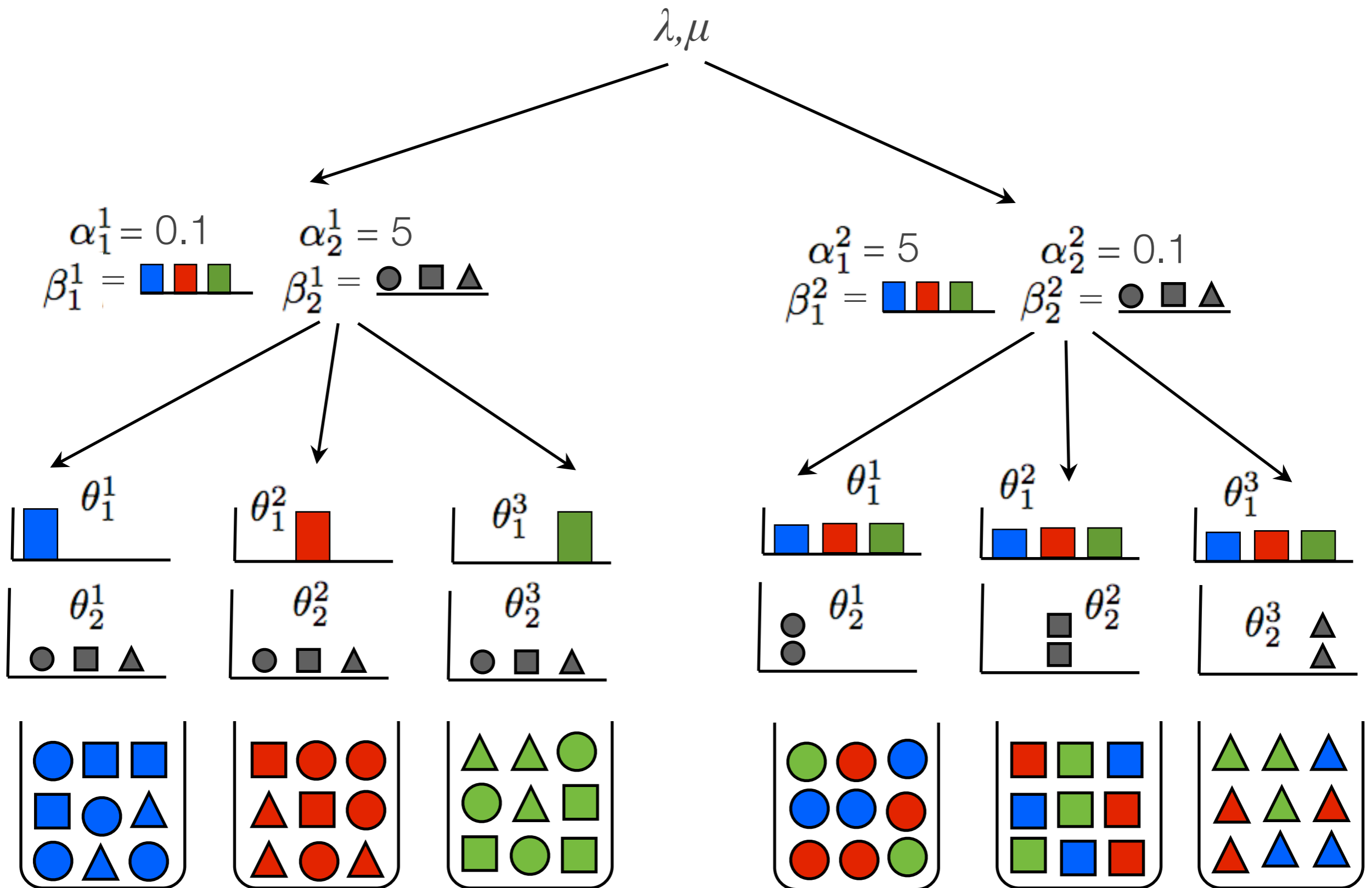


Learning multiple kinds

$z \sim \text{CRP}(\gamma)$
 $\alpha^k \sim \text{Exponential}(\lambda)$
 $\beta^k \sim \text{Dirichlet}(\mathbf{1})$
 $\theta^i \sim \text{Dirichlet}(\alpha^{z_i} \beta^{z_i})$
 $y^i | n^i \sim \text{Multinomial}(\theta^i)$



Extendible to having multiple features



Another problem

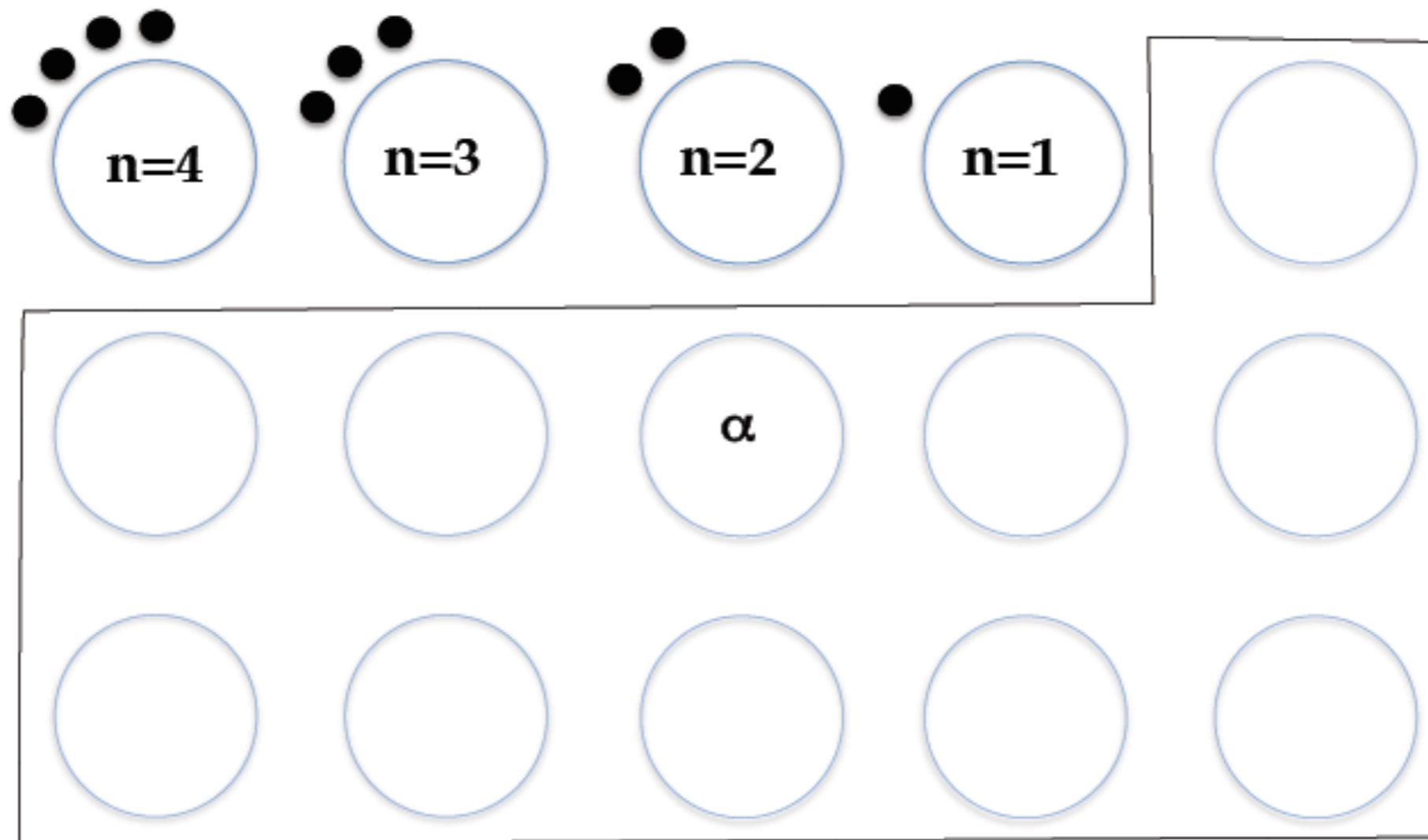
This assumes that it is given, as data, what each of the categories are (i.e., that it's all supervised)

This is clearly not accurate; we want to be able to have it search over the space of assignments of items to categories, and simultaneously figure out the best category assignments as well as the appropriate numbers of kinds

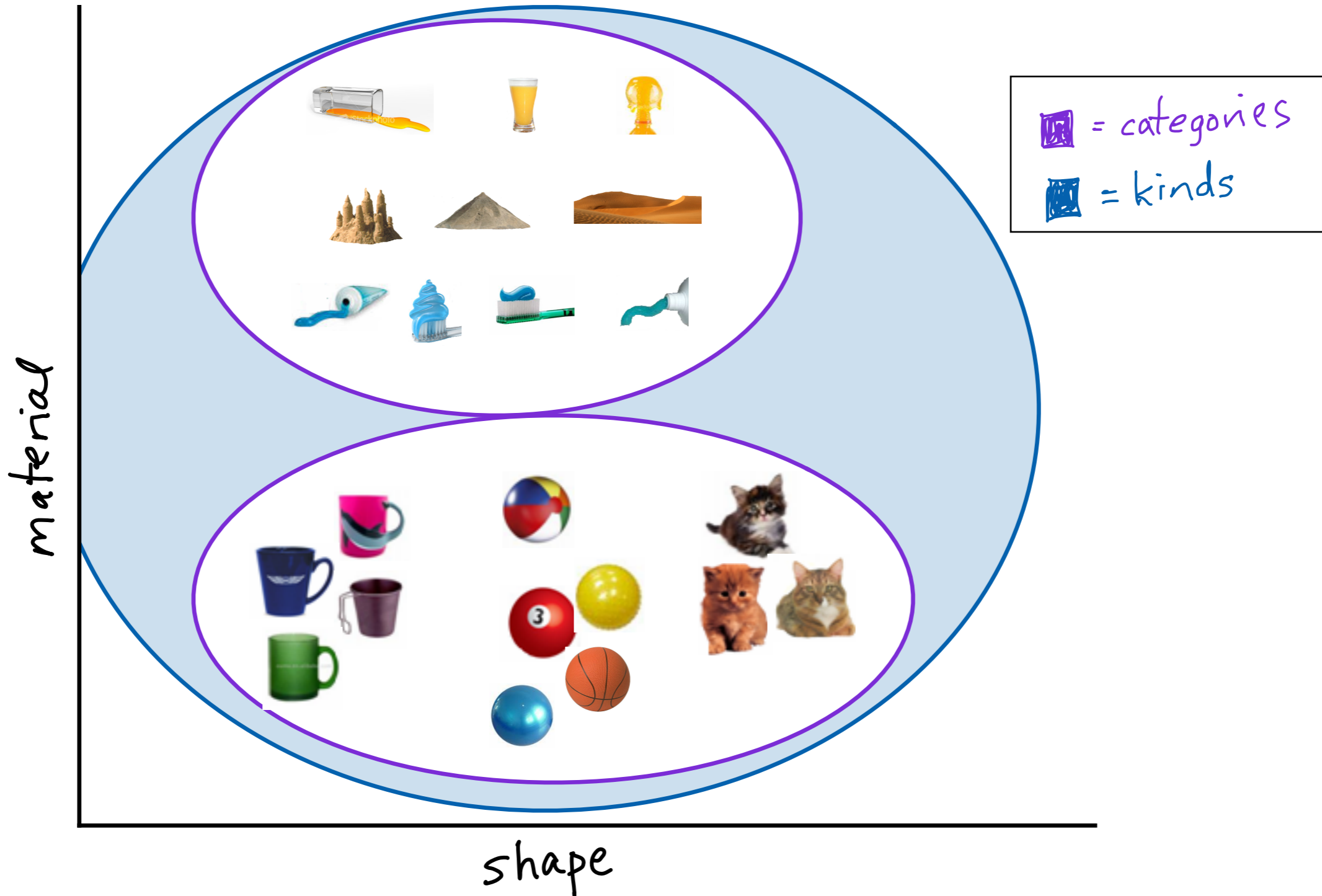
Hmmm... maybe some sort of prior that favours fewer categories, but can put items into arbitrarily many....

What kind of prior might that be?

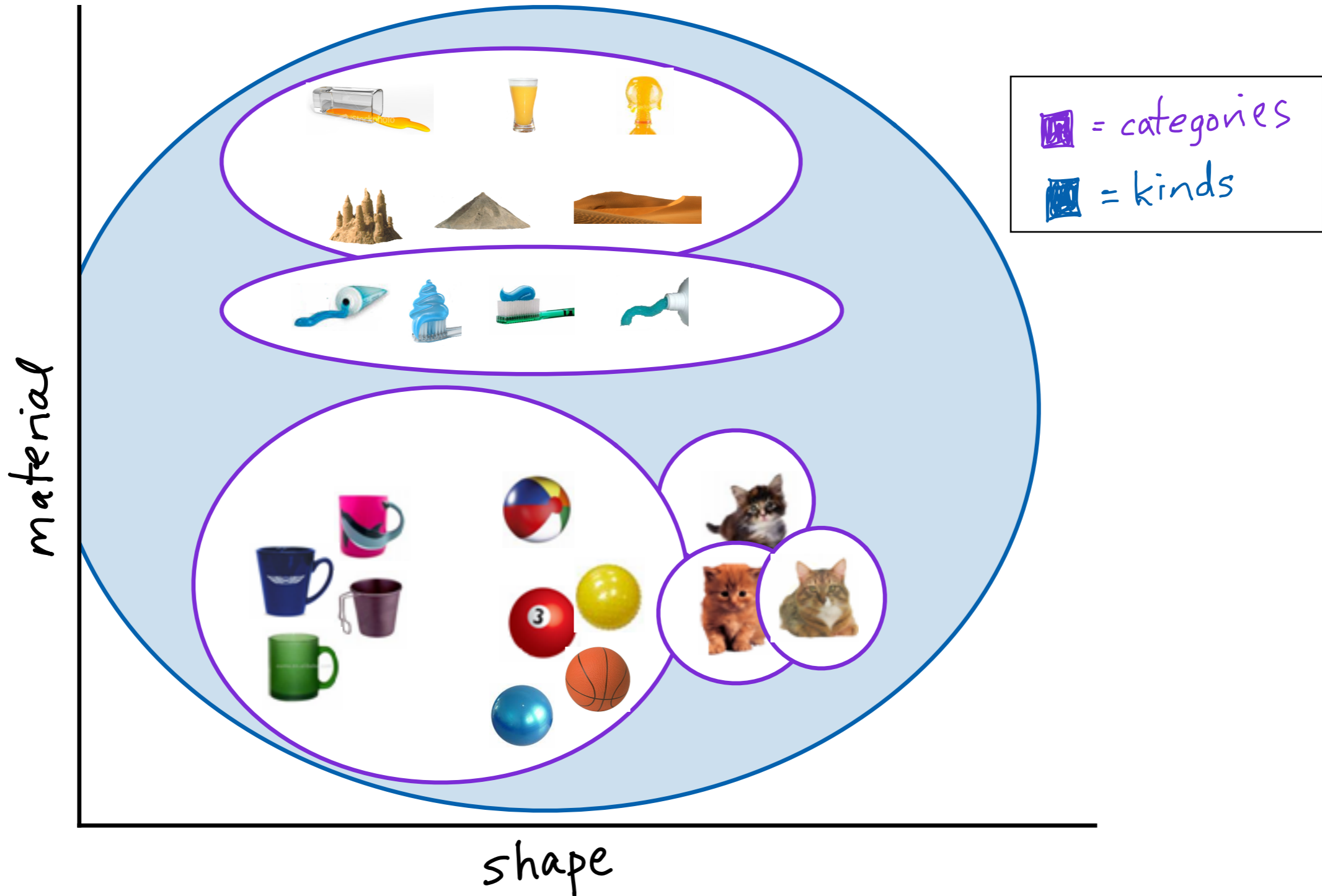
Well, really, what else?



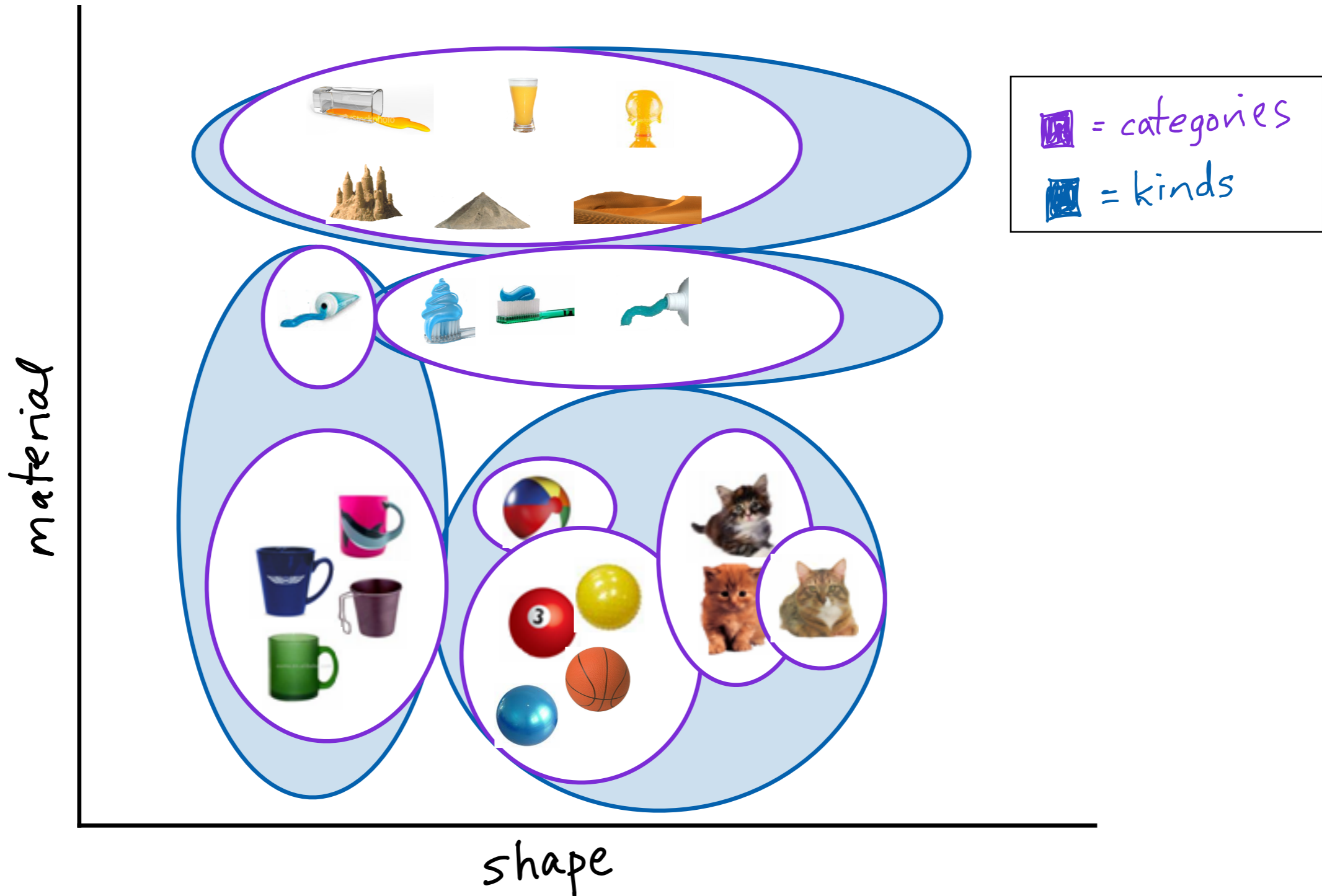
A sketch of the model



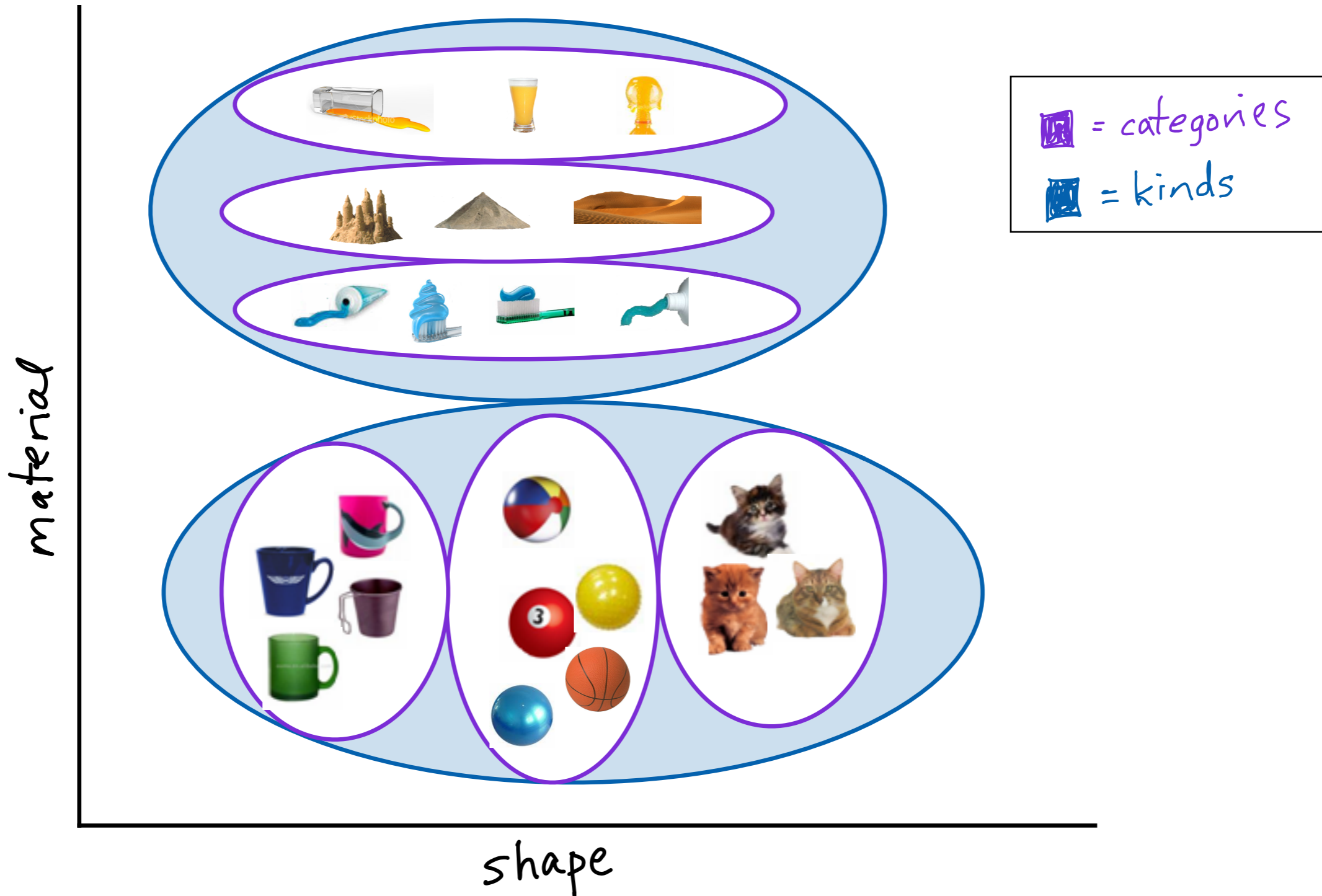
A sketch of the model



A sketch of the model



A sketch of the model



A sketch of the model

That's all well and good...

how does it do?

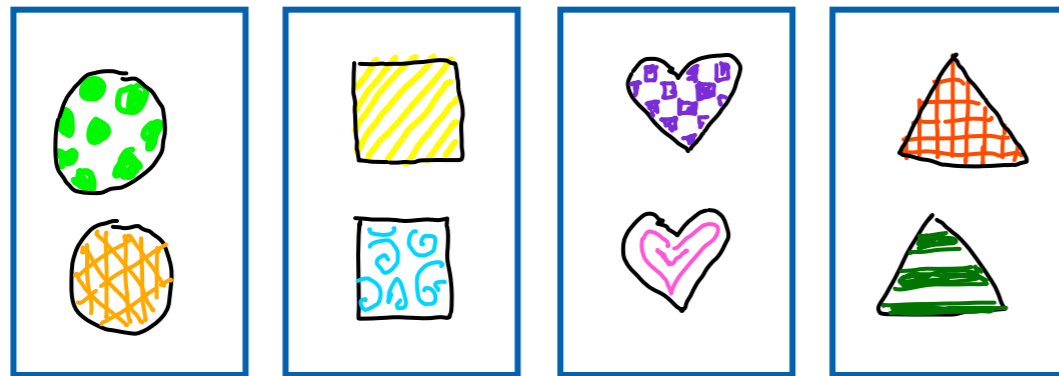
Lecture outline (next three lectures)


- ➔ Today: Learning about category variability
 - This kind of learning in children and adults
 - A model for this kind of learning
- ➔ Performance of this model
- ▶ Lecture 12: Learning about distributions of categories
 - This kind of learning in adults
 - Failure of current models
 - A model for this kind of learning
- ▶ Lecture 13: Learning about category structure
 - A model for this kind of learning
 - This kind of learning in people

Captures the acquisition of the shape bias

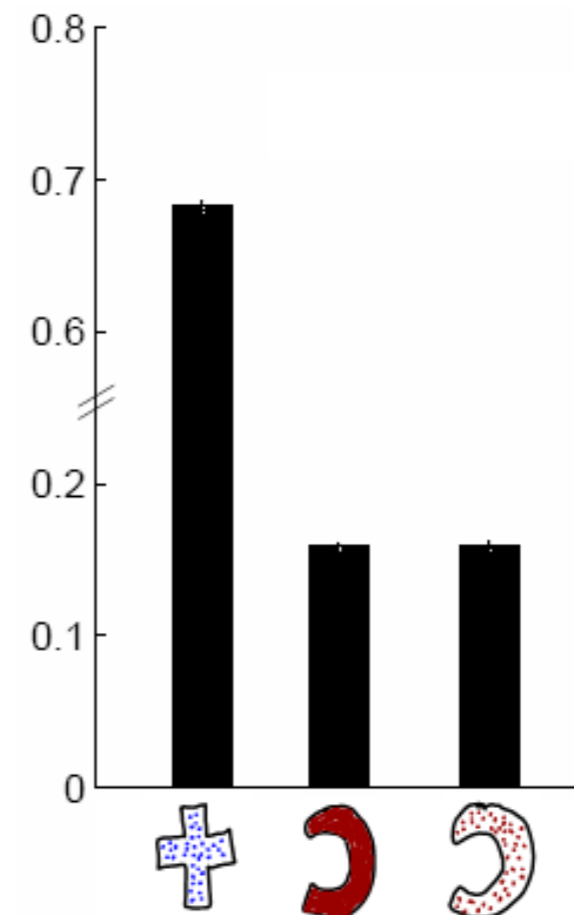
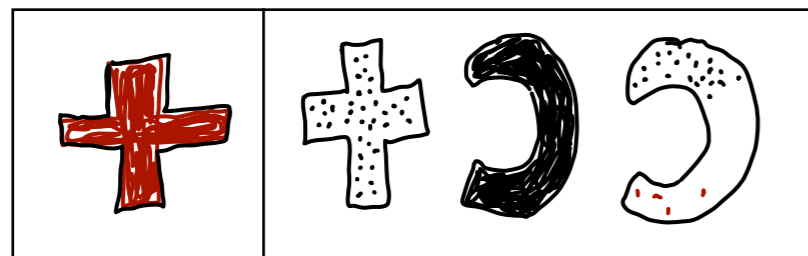
Classifies by shape based on four training categories

Training



Probability that object belongs to the same category as 

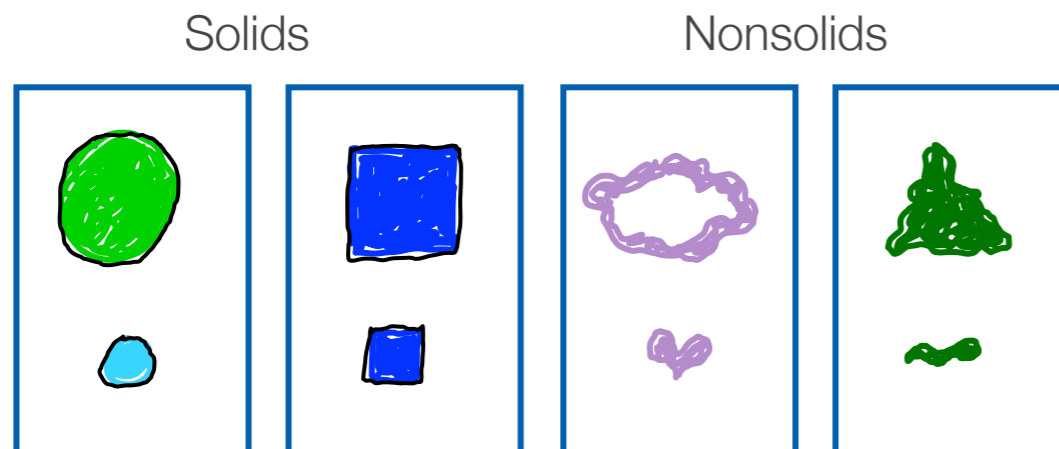
Testing



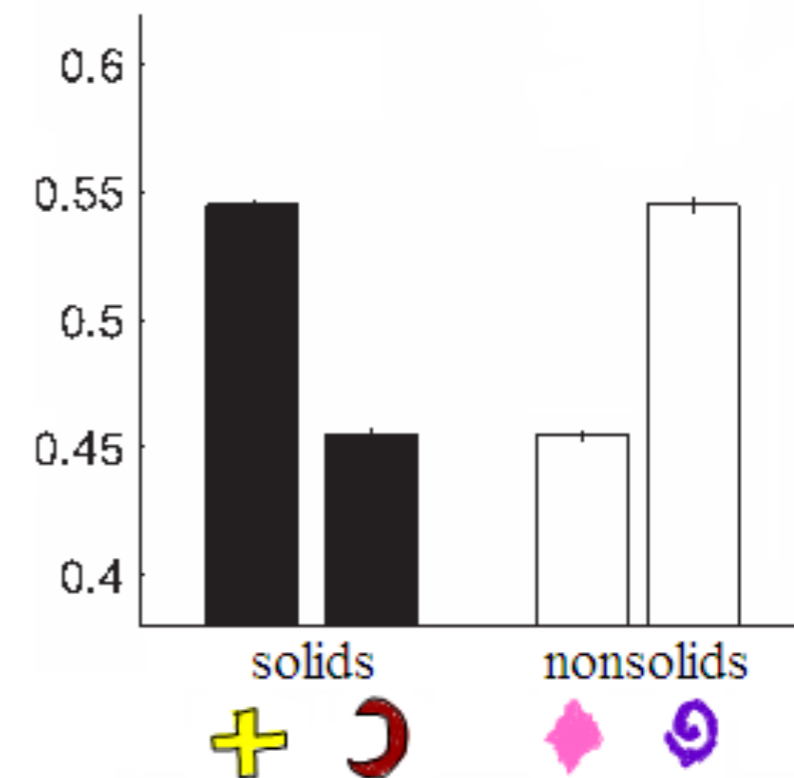
Captures the acquisition of the shape bias

Can also learn multiple different kinds

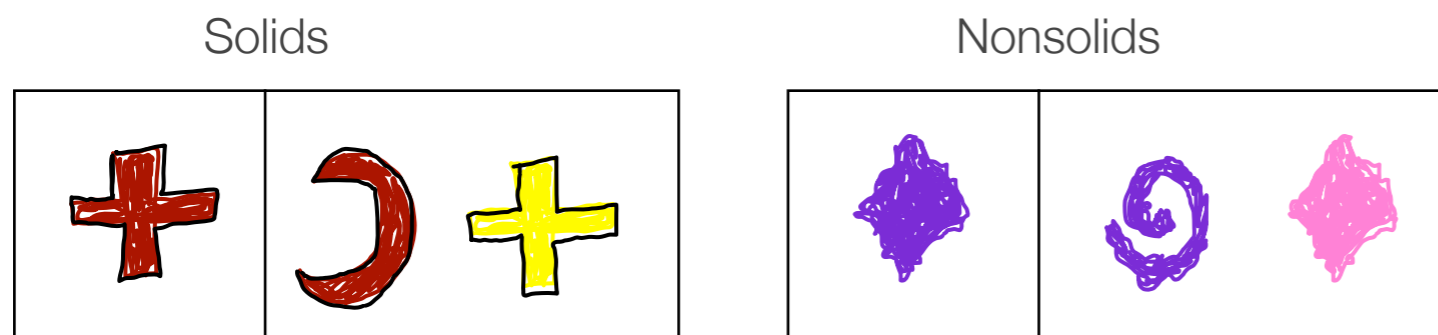
Training



Probability that object belongs to the same category as the test



Testing



One empirical question

Children obviously learn this somehow based on the data. Two possibilities present themselves:

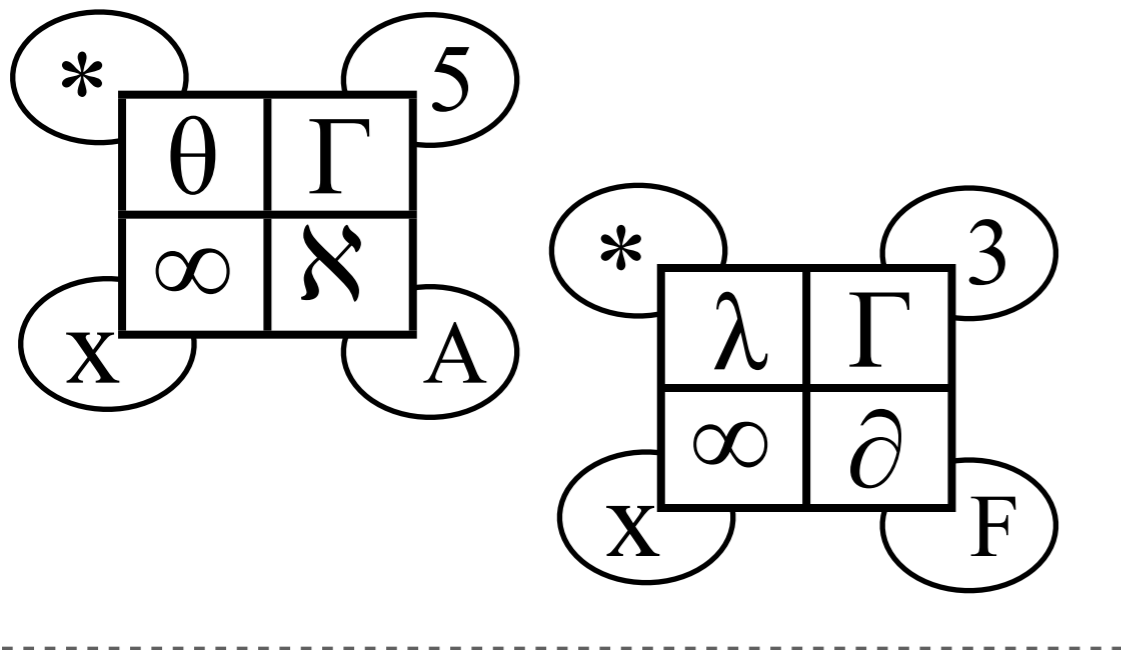
- ▶ Children can learn multiple overhypothesis because they have a bias to consider only certain kinds of features

OR

- ▶ People in general are able to learn many different kinds of arbitrary overhypotheses based on relatively little data

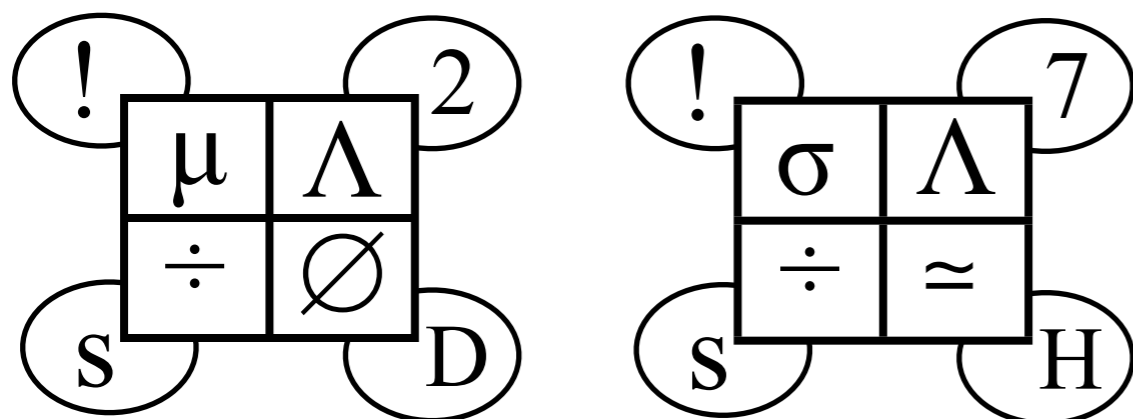
Test: Can people learn arbitrary overhypotheses?

Category learning task with arbitrary, weird stimuli (with adults)



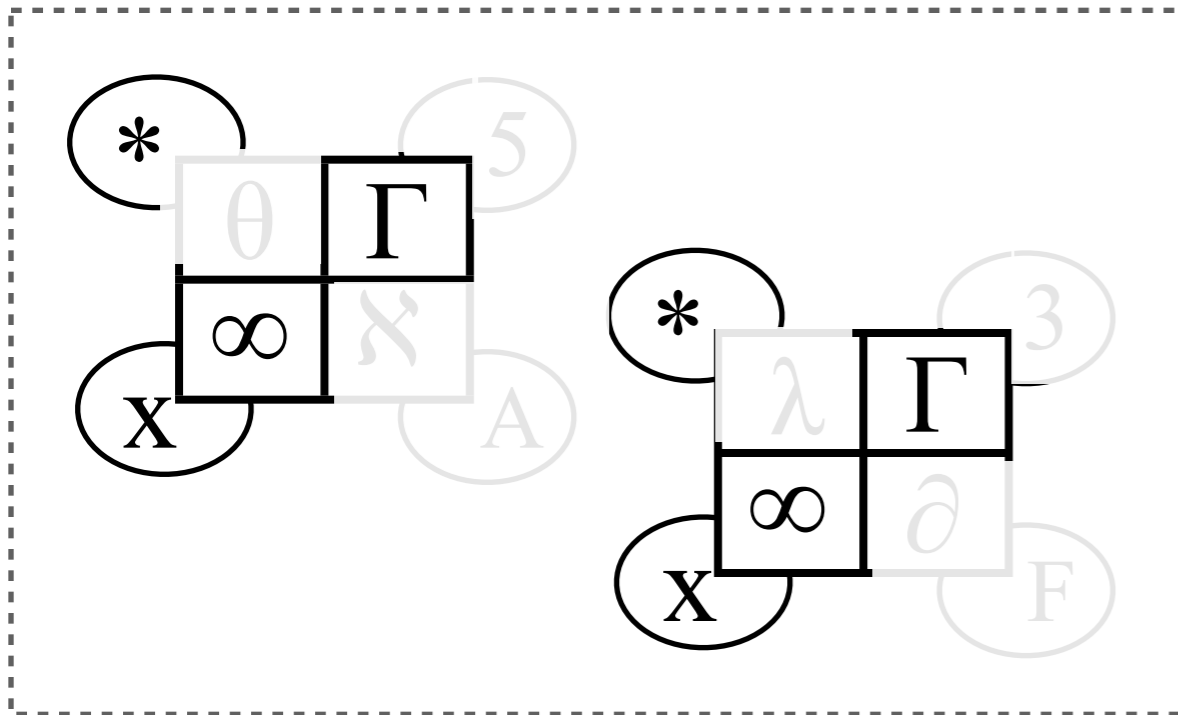
Eight features with 10 possible values

Four features (at random) are used to classify the objects

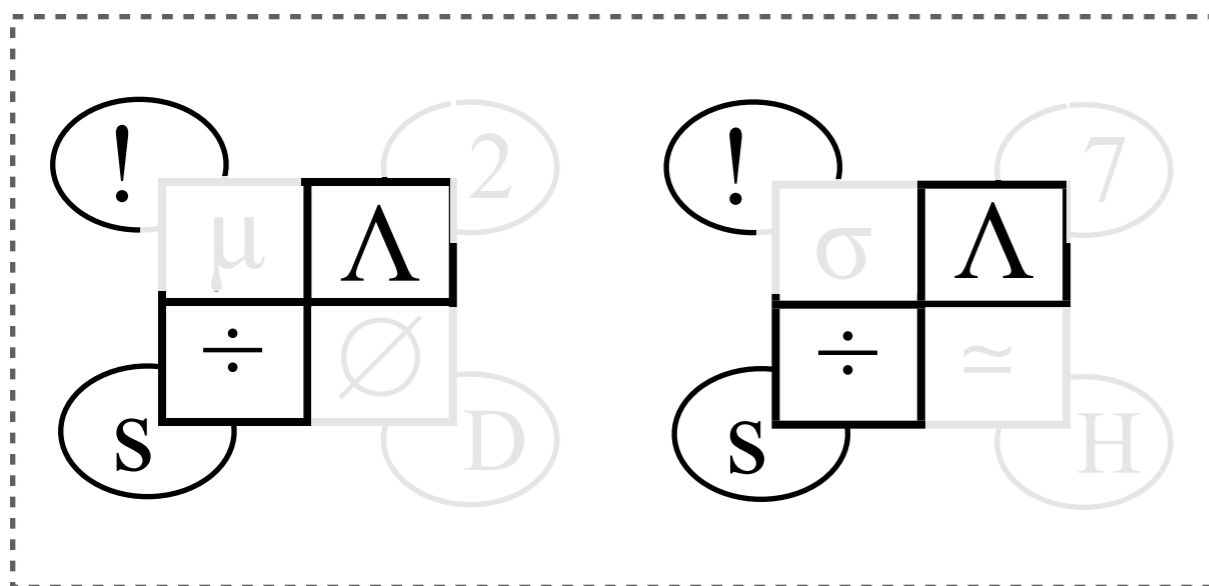


Test: Can people learn arbitrary overhypotheses?

Category learning task with arbitrary, weird stimuli (with adults)



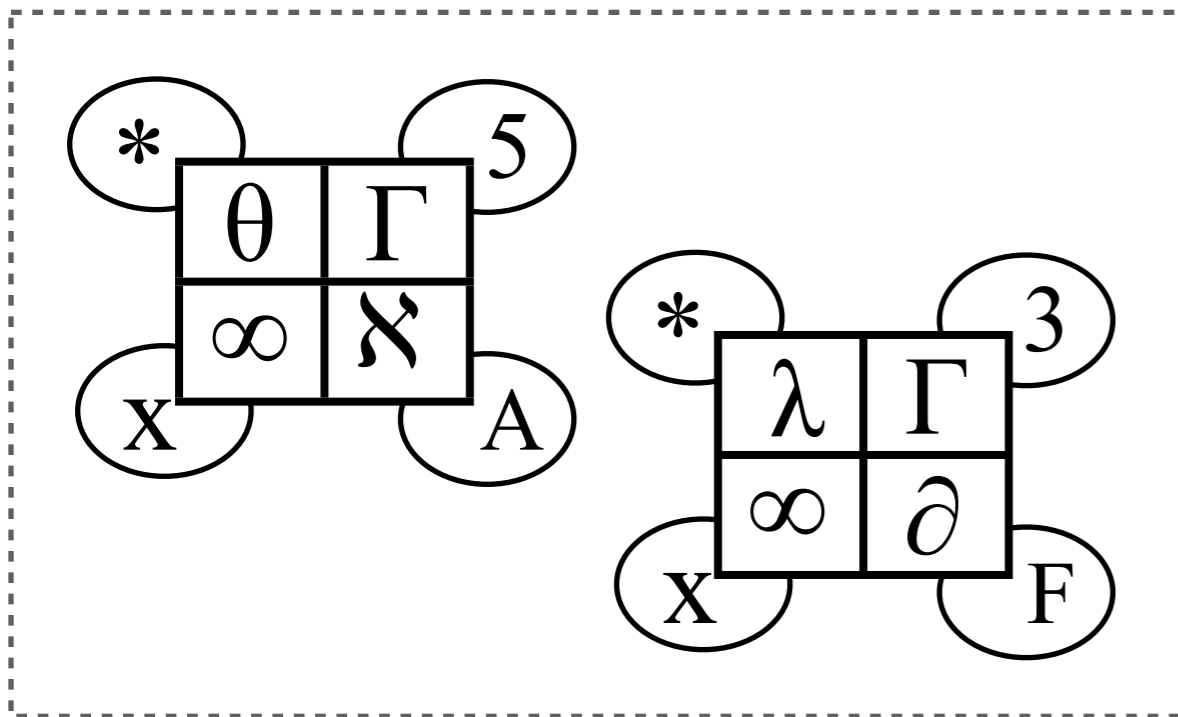
Eight features with 10 possible values



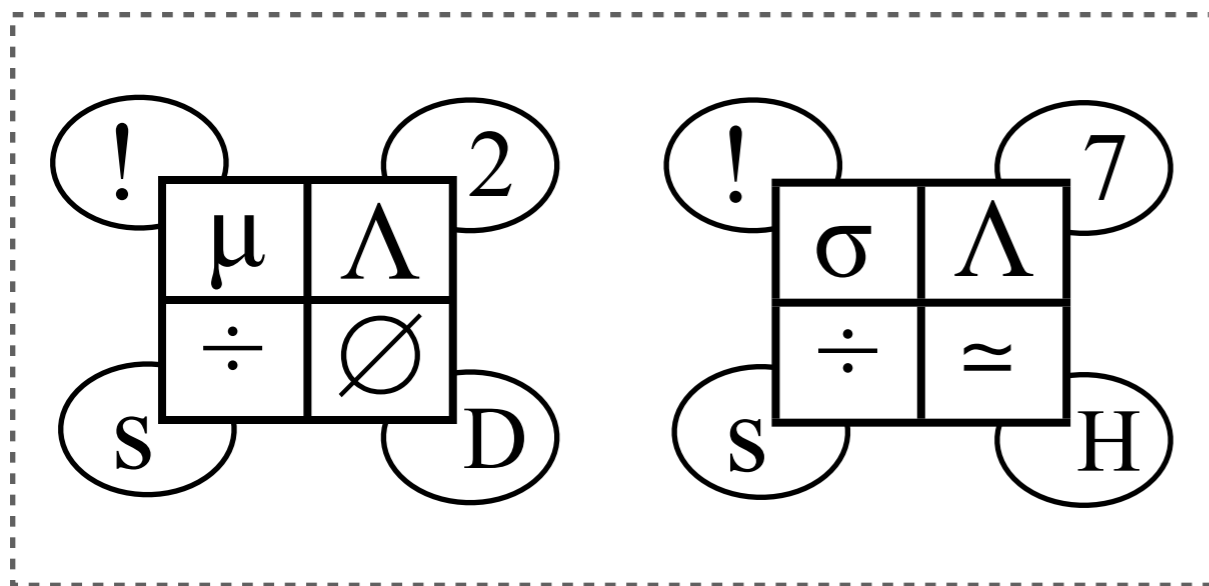
Four features (at random) are used to classify the objects

Test: Can people learn arbitrary overhypotheses?

Category learning task with arbitrary, weird stimuli (with adults)

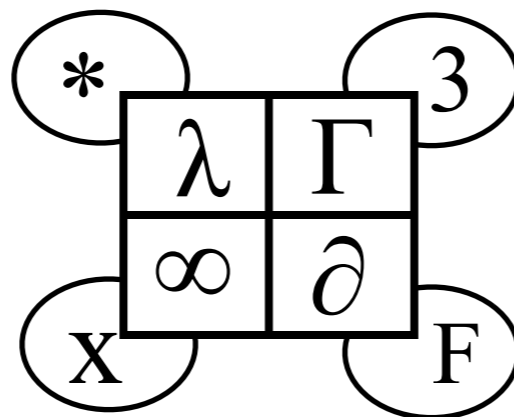
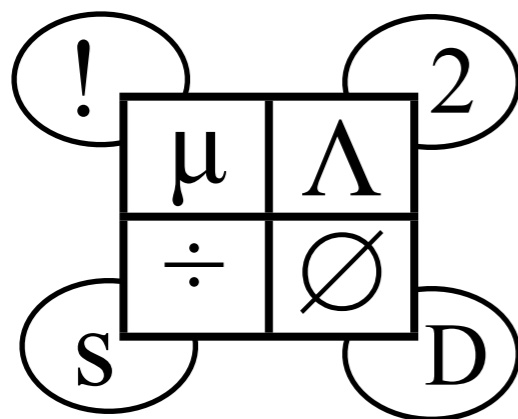
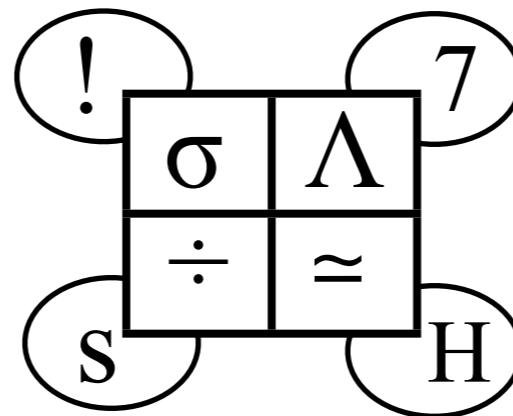
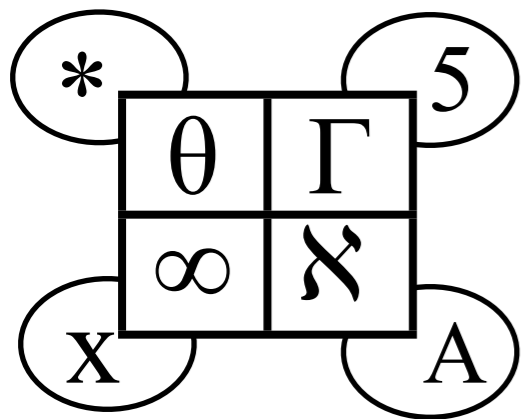


Some trials were supervised (drawn with boxes to indicate the categories)



Test: Can people learn arbitrary overhypotheses?

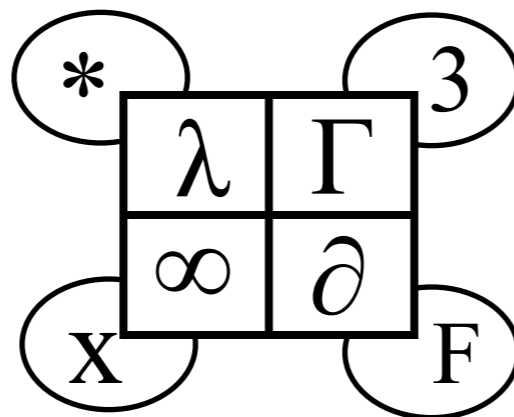
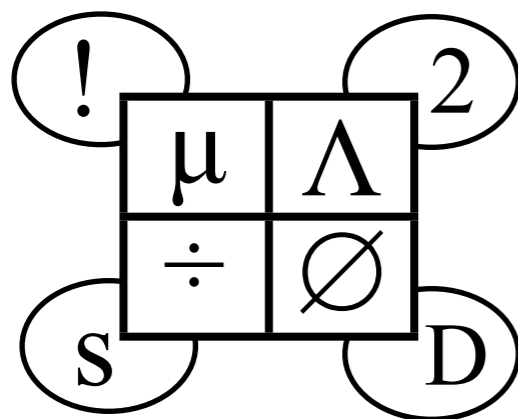
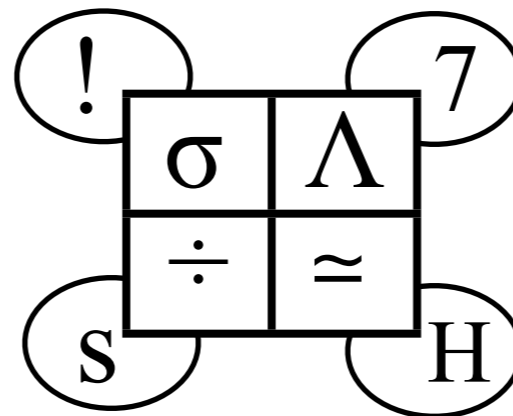
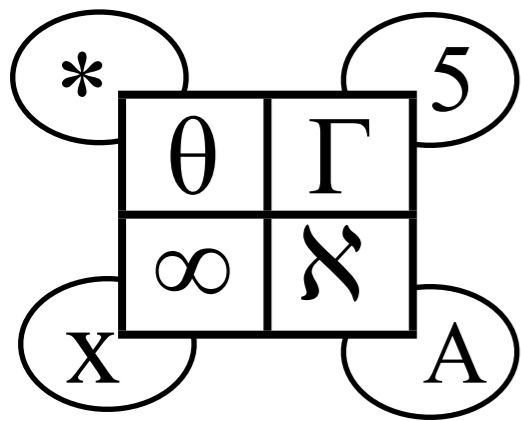
Category learning task with arbitrary, weird stimuli (with adults)



Others were totally unsupervised (people had to sort items themselves)

Test: Can people learn arbitrary overhypotheses?

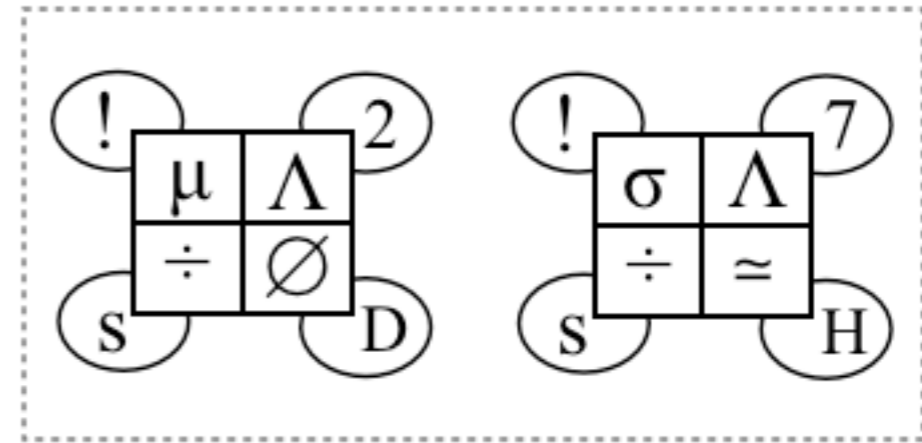
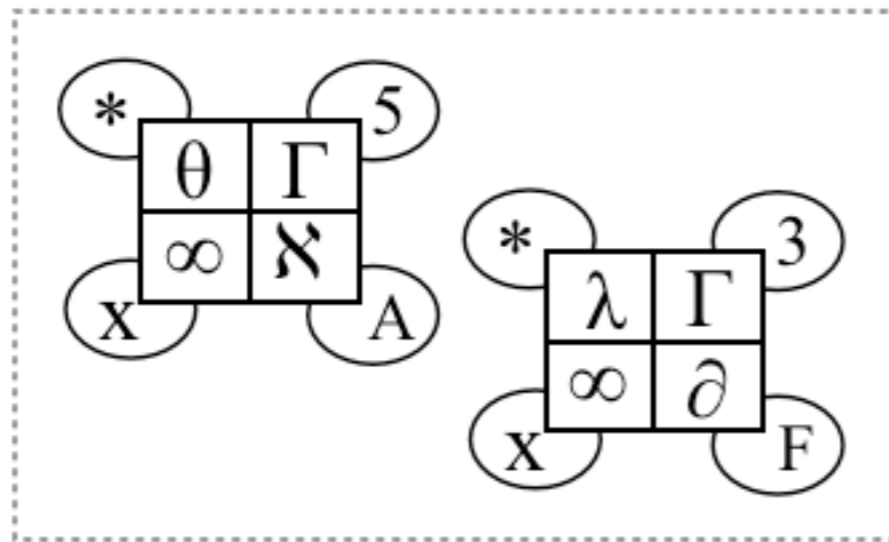
Category learning task with arbitrary, weird stimuli (with adults)



Varied # of items (4, 8, 16)
and number of “true”
categories (2, 4, 8)

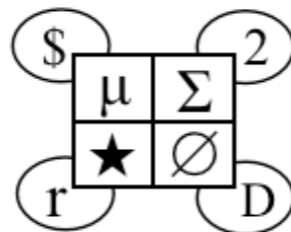
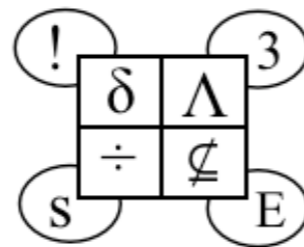
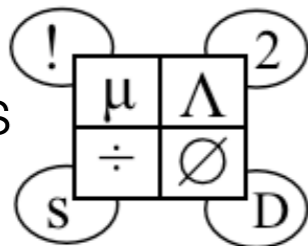
Two kinds of test questions

Given



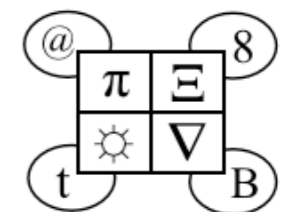
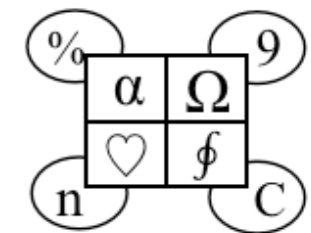
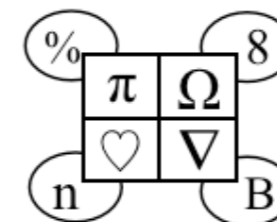
First

Which of the following two items is in the same category as



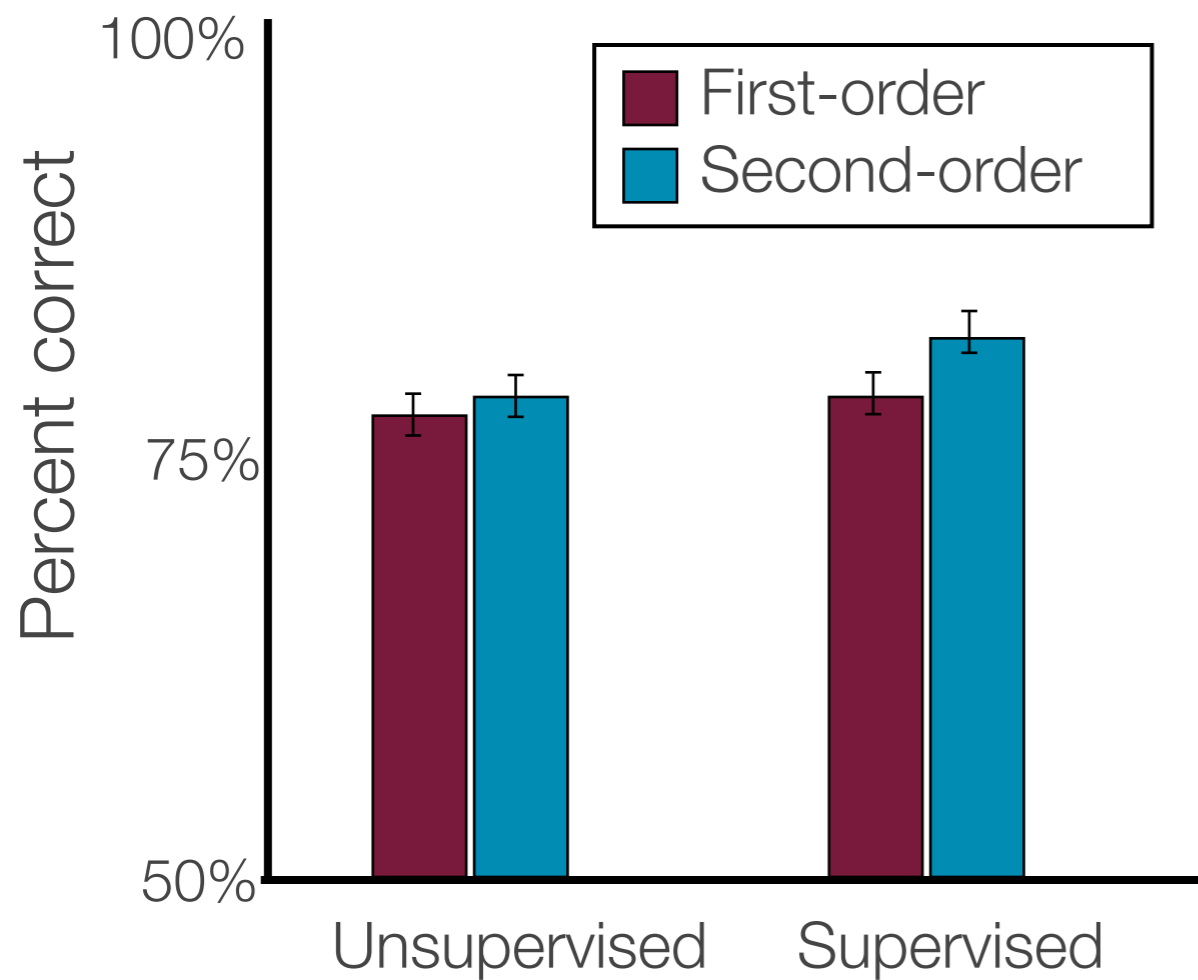
Second

Which of the following two items is in the same category as

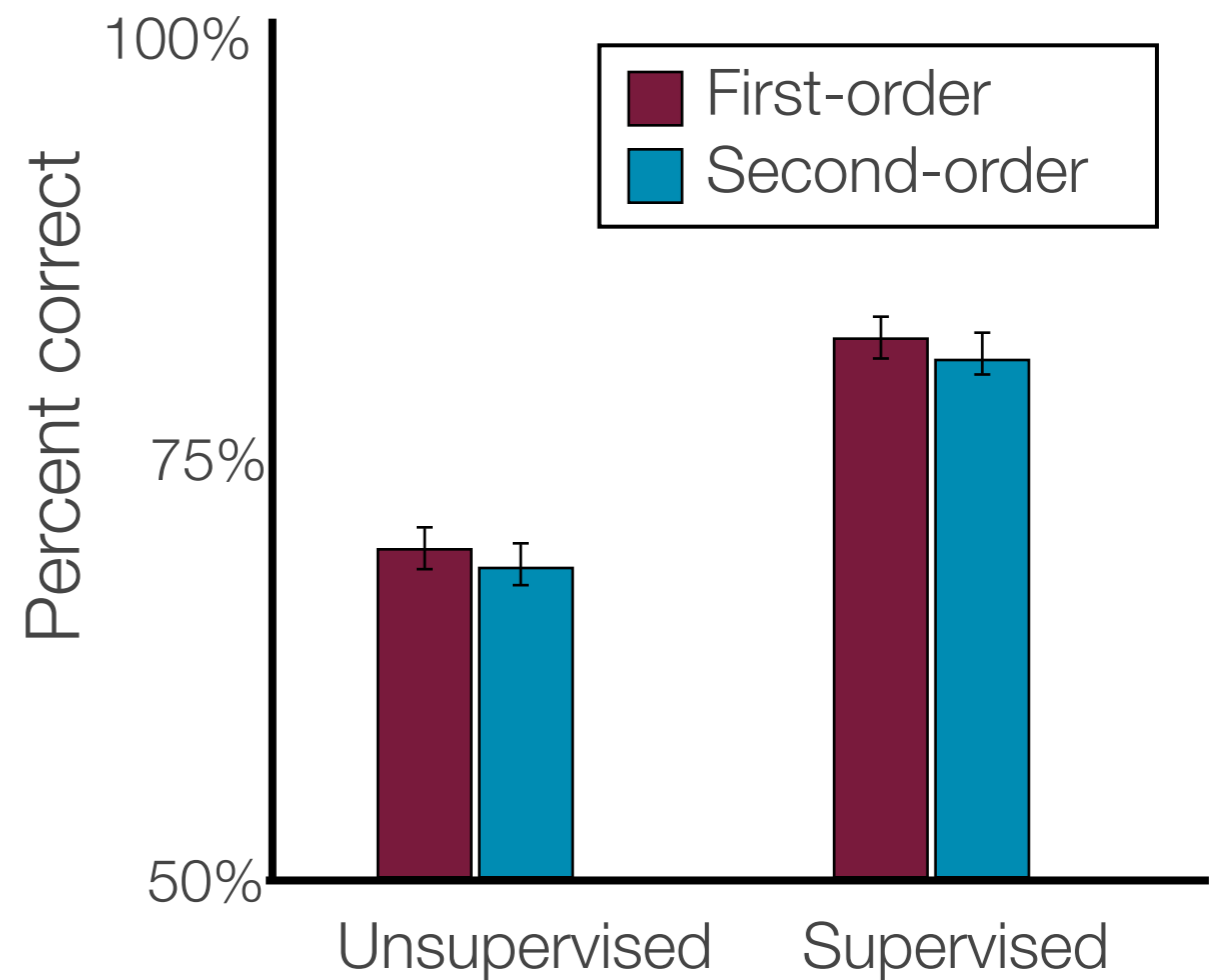


Test: Can people learn arbitrary overhypotheses?

Model



Human

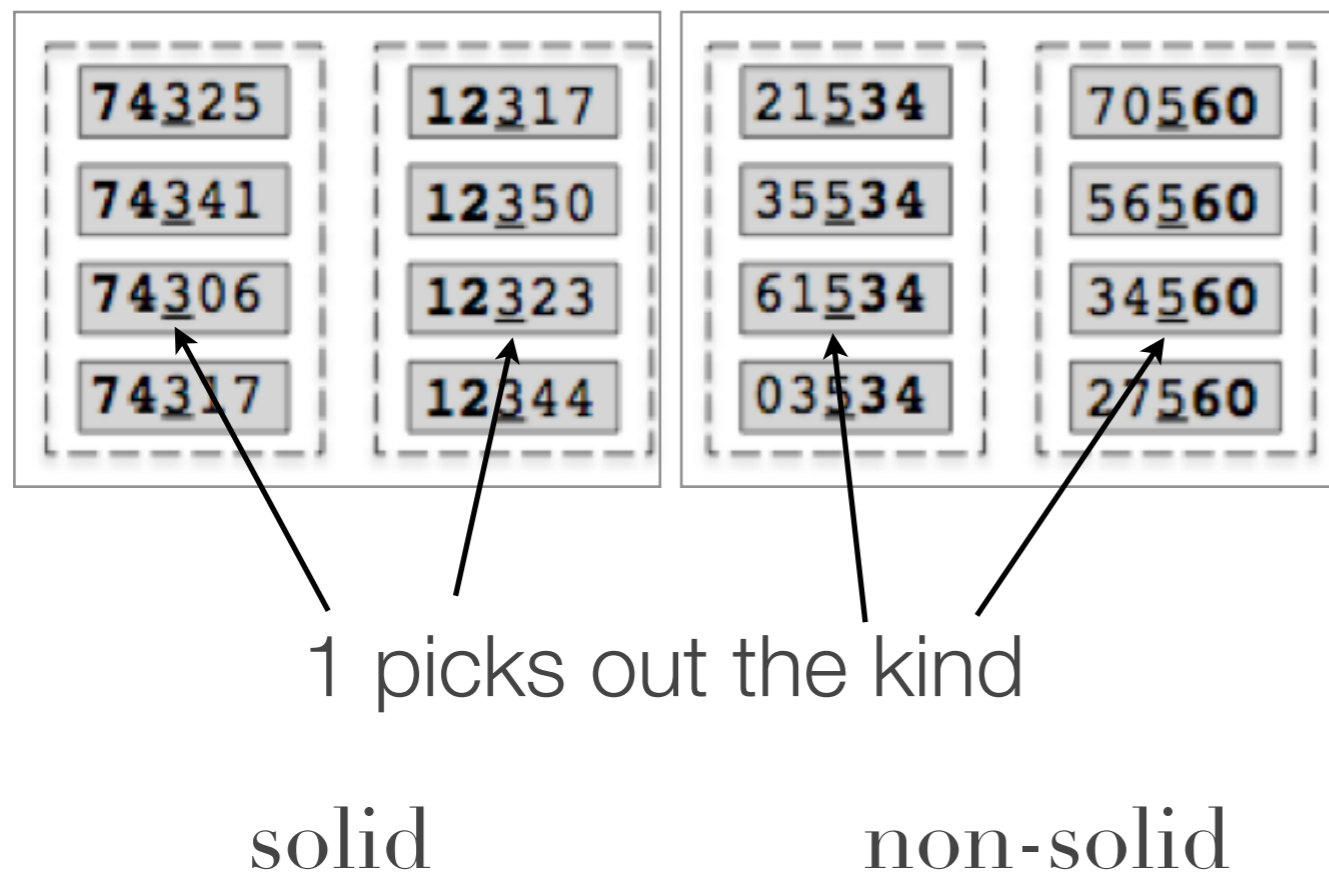


Test: Can people learn arbitrary overhypotheses?

But here they just saw one kind at once
(which lots of models can handle)...

Can people learn multiple kinds, each
with its own overhypothesis?

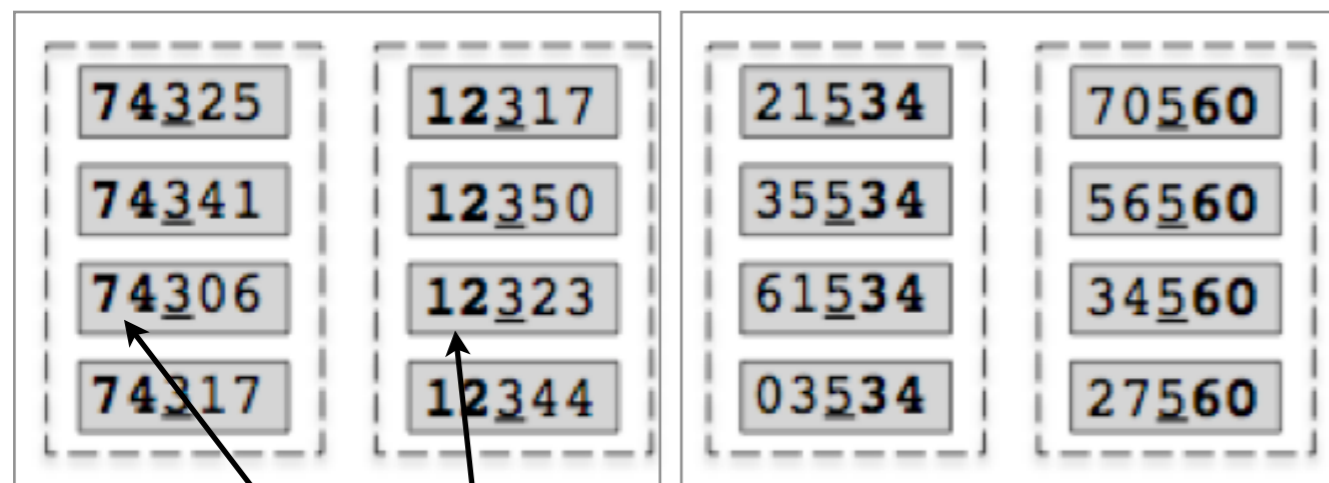
Test: Can people learn arbitrary overhypotheses?



Five total features,
each with 10 possible
values

(in this schematic
diagram, each feature
is indicated by the
location in the vector;
the value is indicated
by the number)

Test: Can people learn arbitrary overhypotheses?



2 pick out the category in kind A

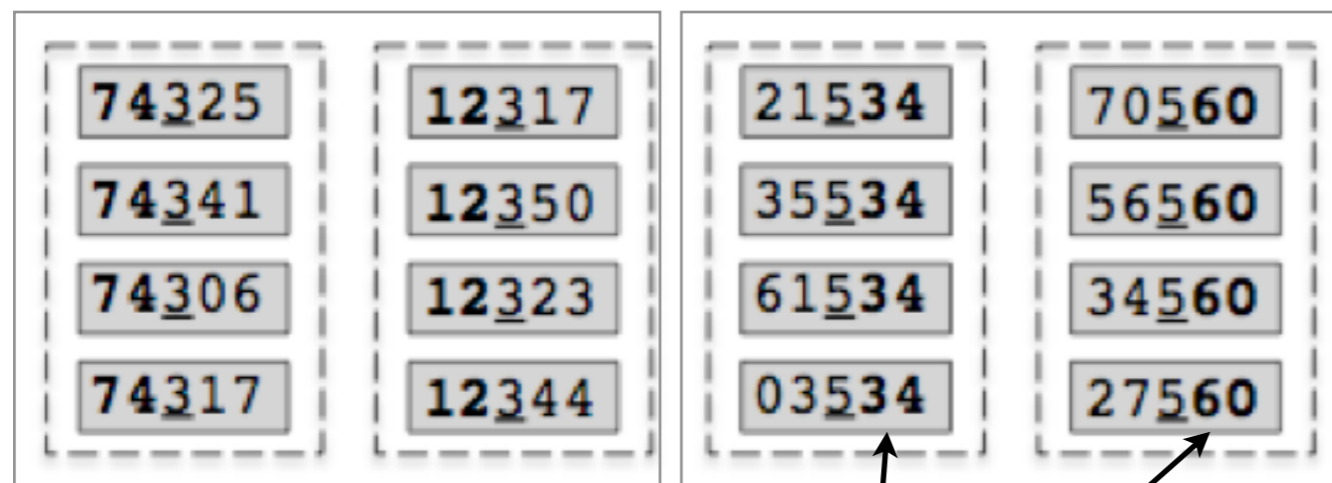
solid
shape

non-solid

Five total features,
each with 10 possible
values

(in this schematic
diagram, each feature
is indicated by the
location in the vector;
the value is indicated
by the number)

Test: Can people learn arbitrary overhypotheses?



2 pick out the category in kind B

solid
shape

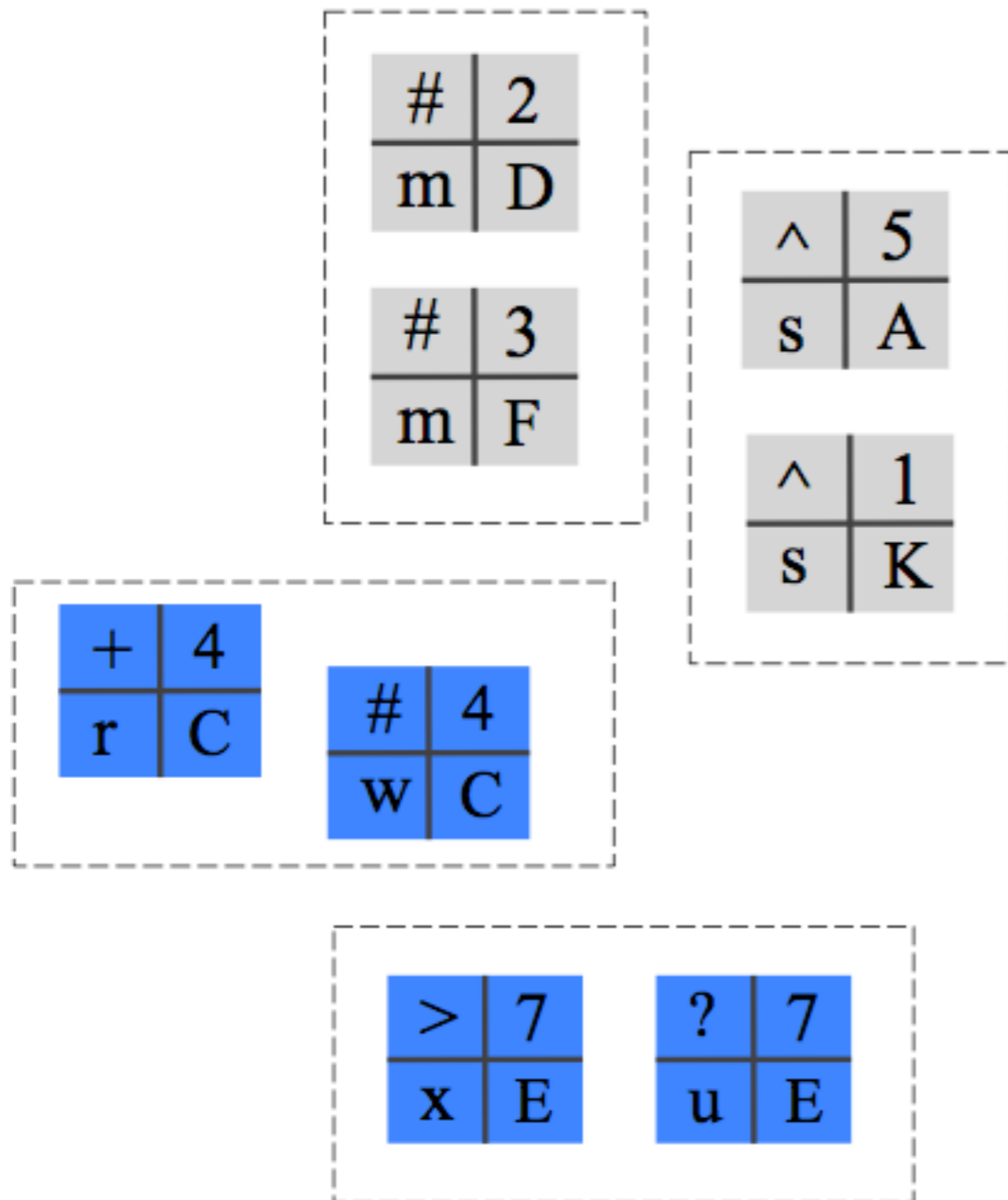
non-solid
texture/colour

Five total features,
each with 10 possible
values

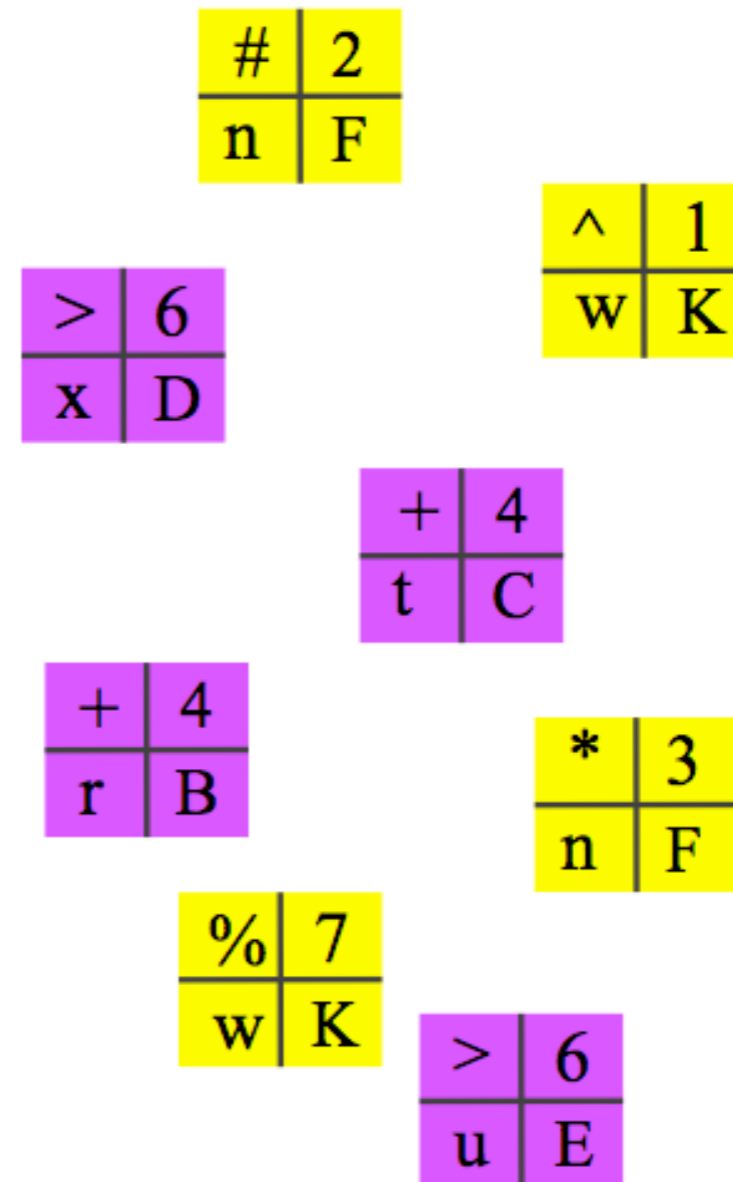
(in this schematic
diagram, each feature
is indicated by the
location in the vector;
the value is indicated
by the number)

Test: Can people learn arbitrary overhypotheses?

Supervised



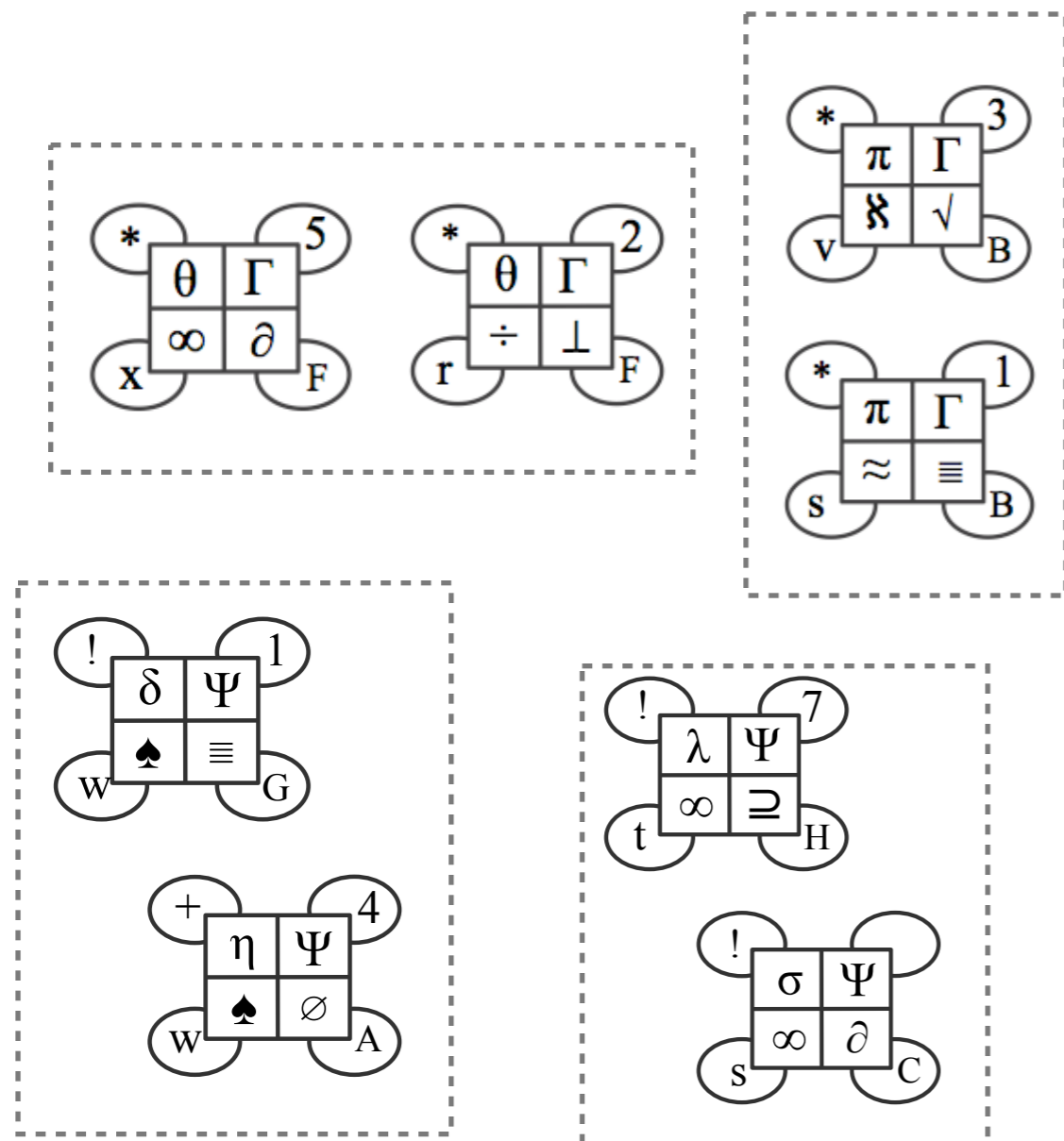
Unsupervised



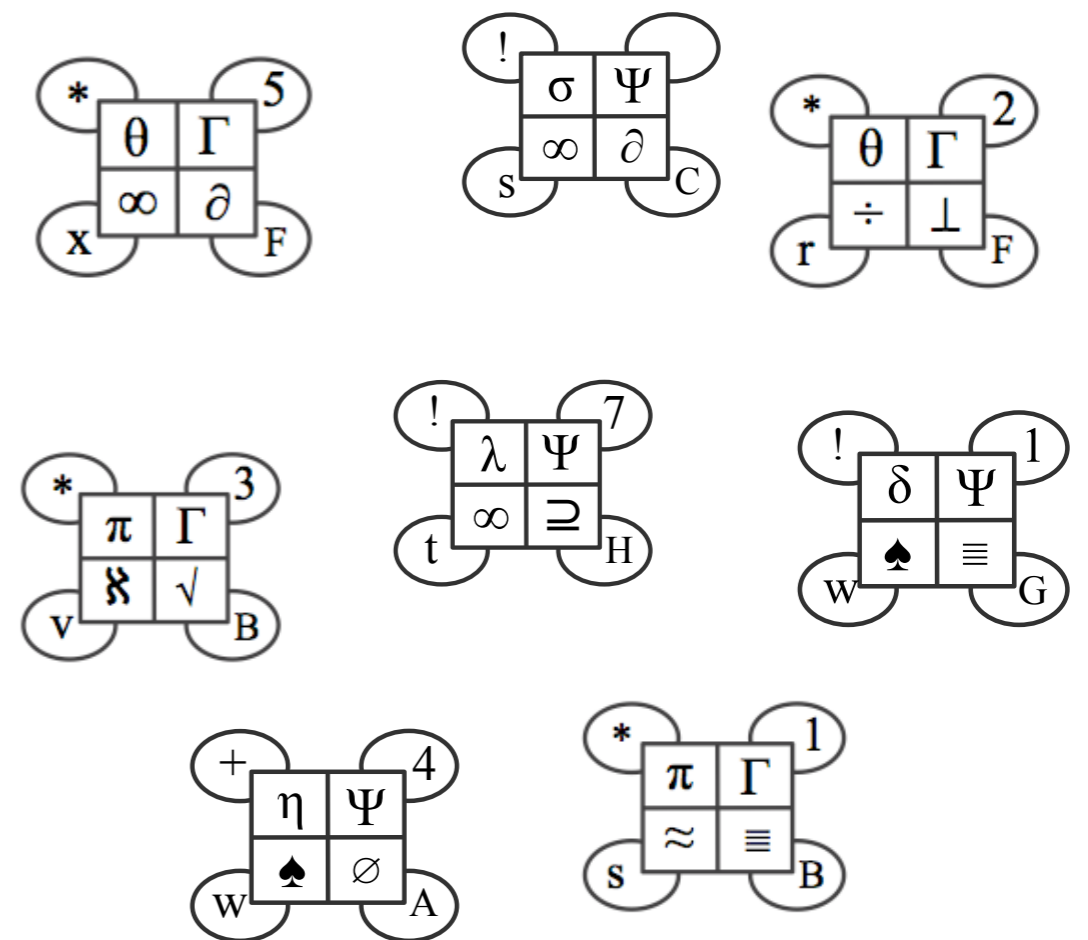
Test: Can people learn arbitrary overhypotheses?

(also did a harder condition with much more challenging stimuli)

Supervised

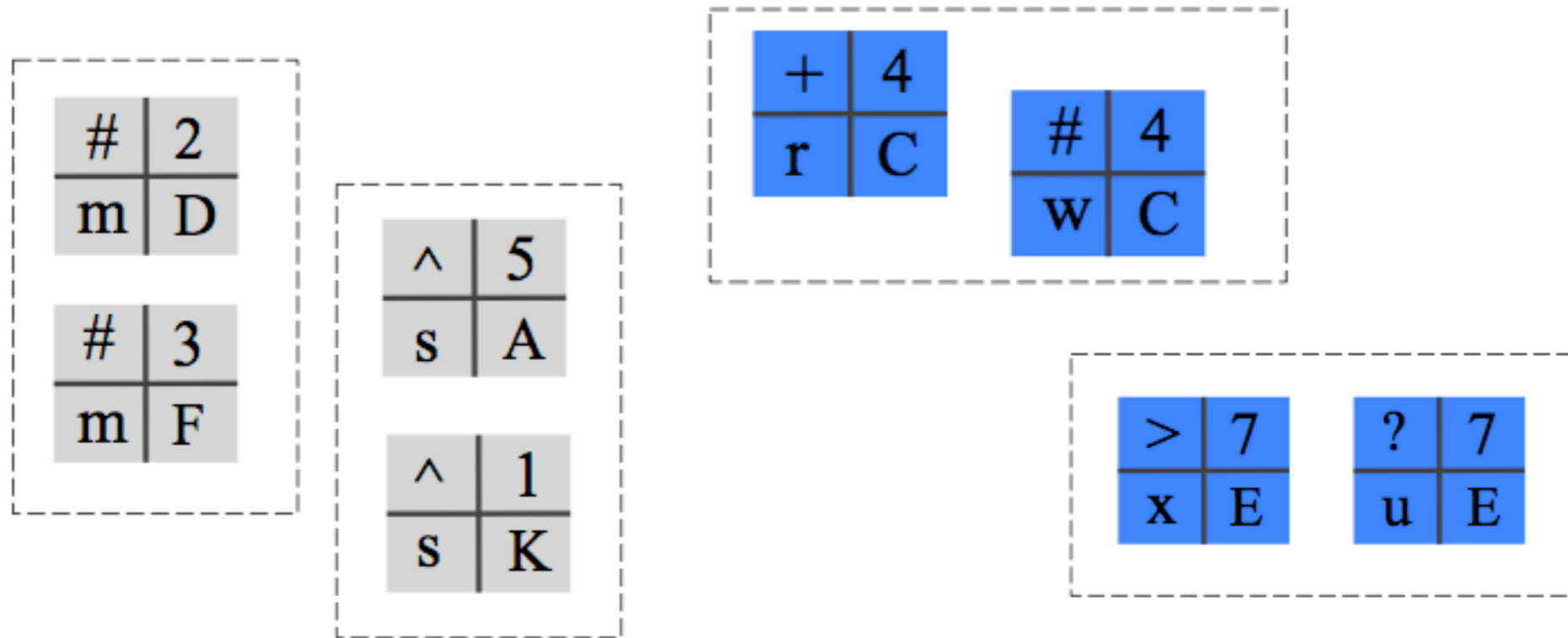


Unsupervised



Test: Can people learn arbitrary overhypotheses?

Given



First

Which of the following two items is in the same category as

#	2
m	D

#	4
m	E

+	2
u	D

Second

Which of the following two items is in the same category as

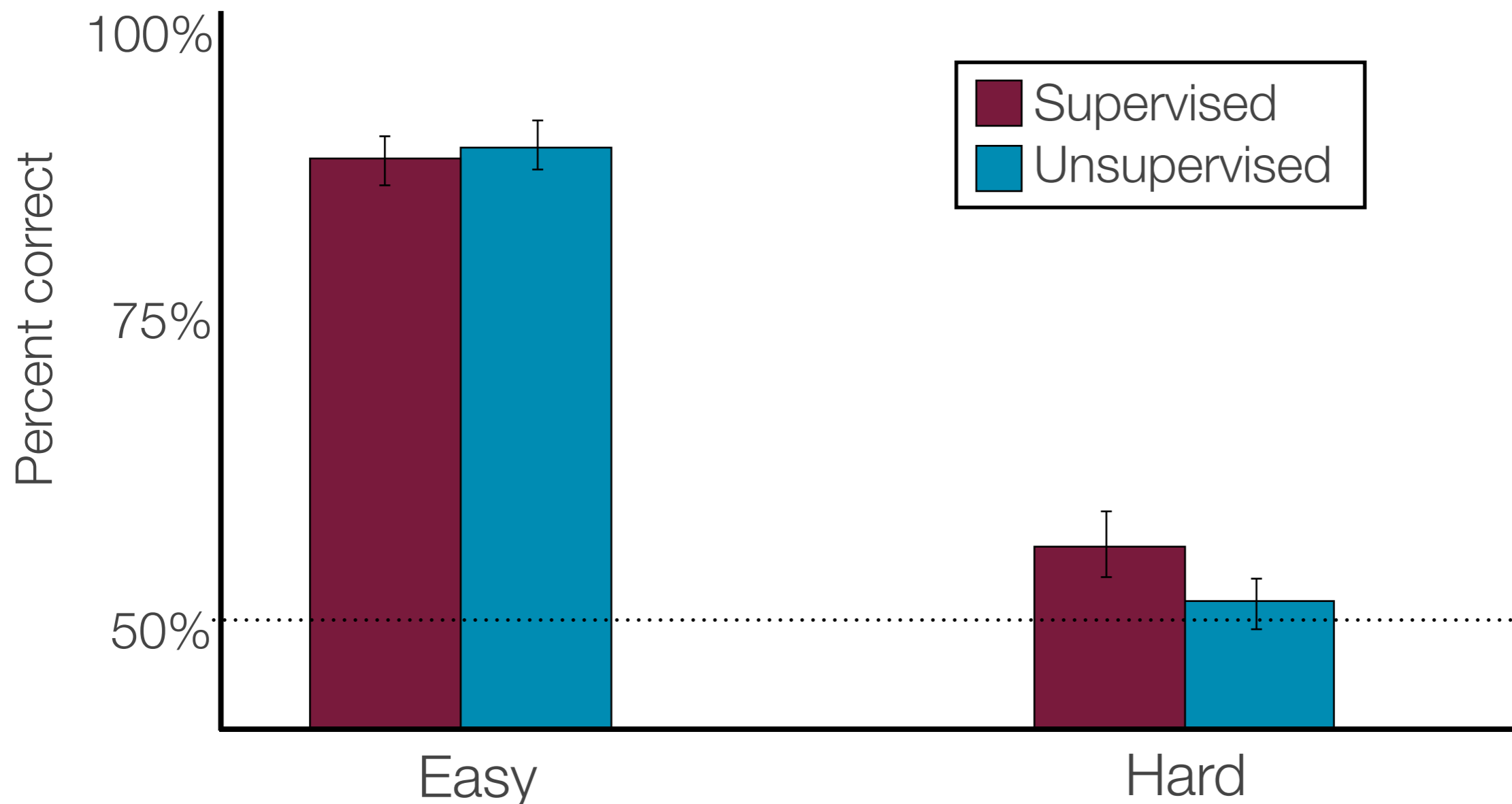
!	8
n	B

!	9
n	G

%	8
t	B

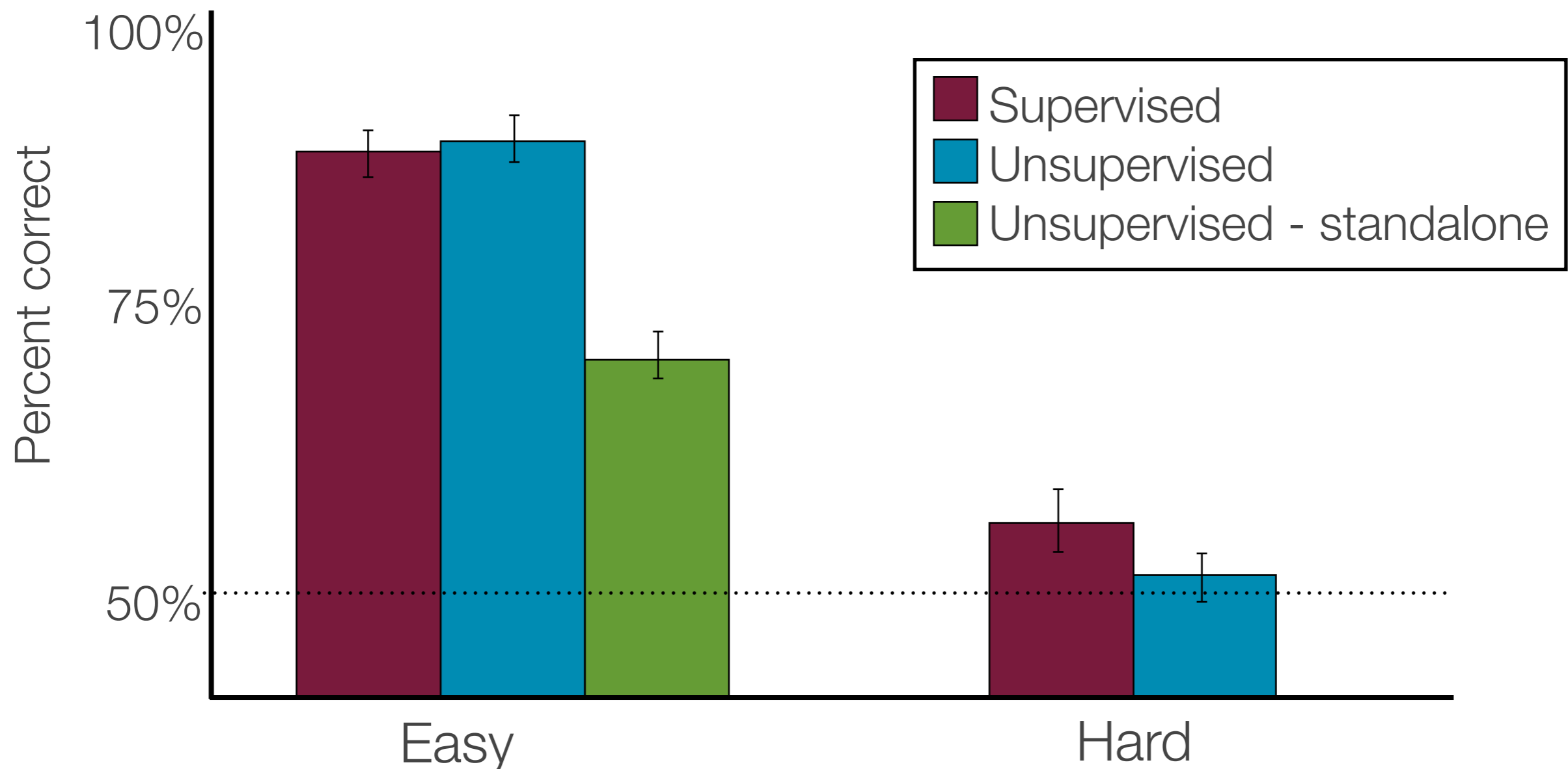
Test: Can people learn arbitrary overhypotheses?

- ▶ People can learn multiple arbitrary kinds and categories, at least if there are few features or they are very salient



Test: Can people learn arbitrary overhypotheses?

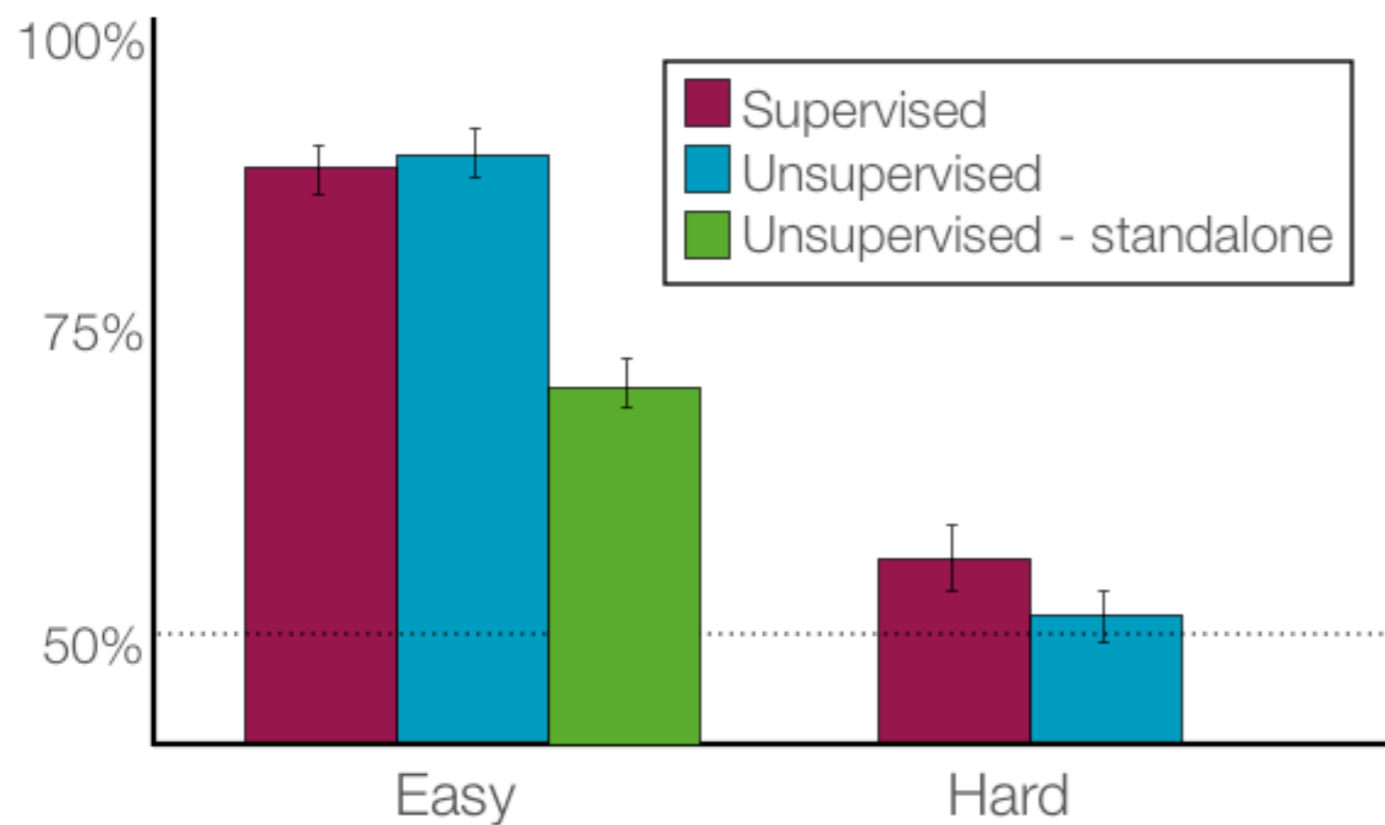
- ▶ People can learn multiple arbitrary kinds and categories, at least if there are few features or they are very salient



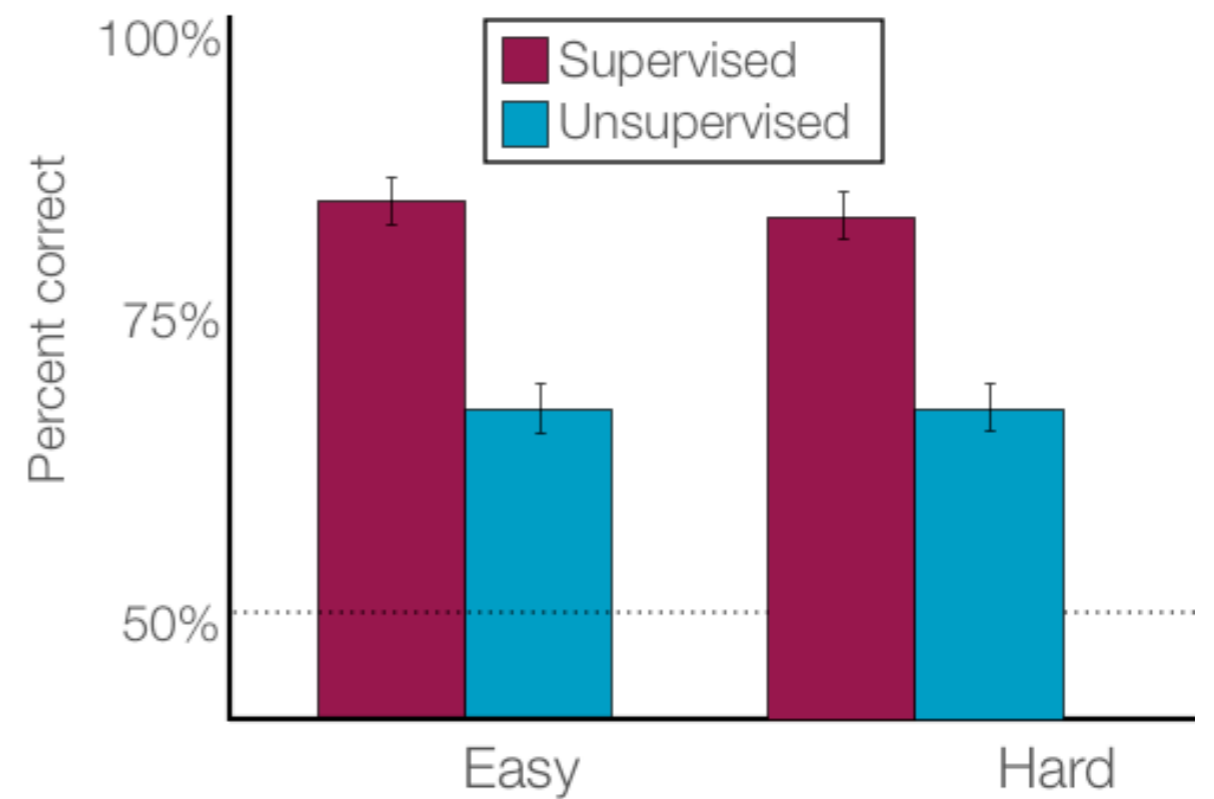
Model performance

- ▶ Model captures the difference between supervised and unsupervised, but not human failure in the “hard” condition

Humans



Model

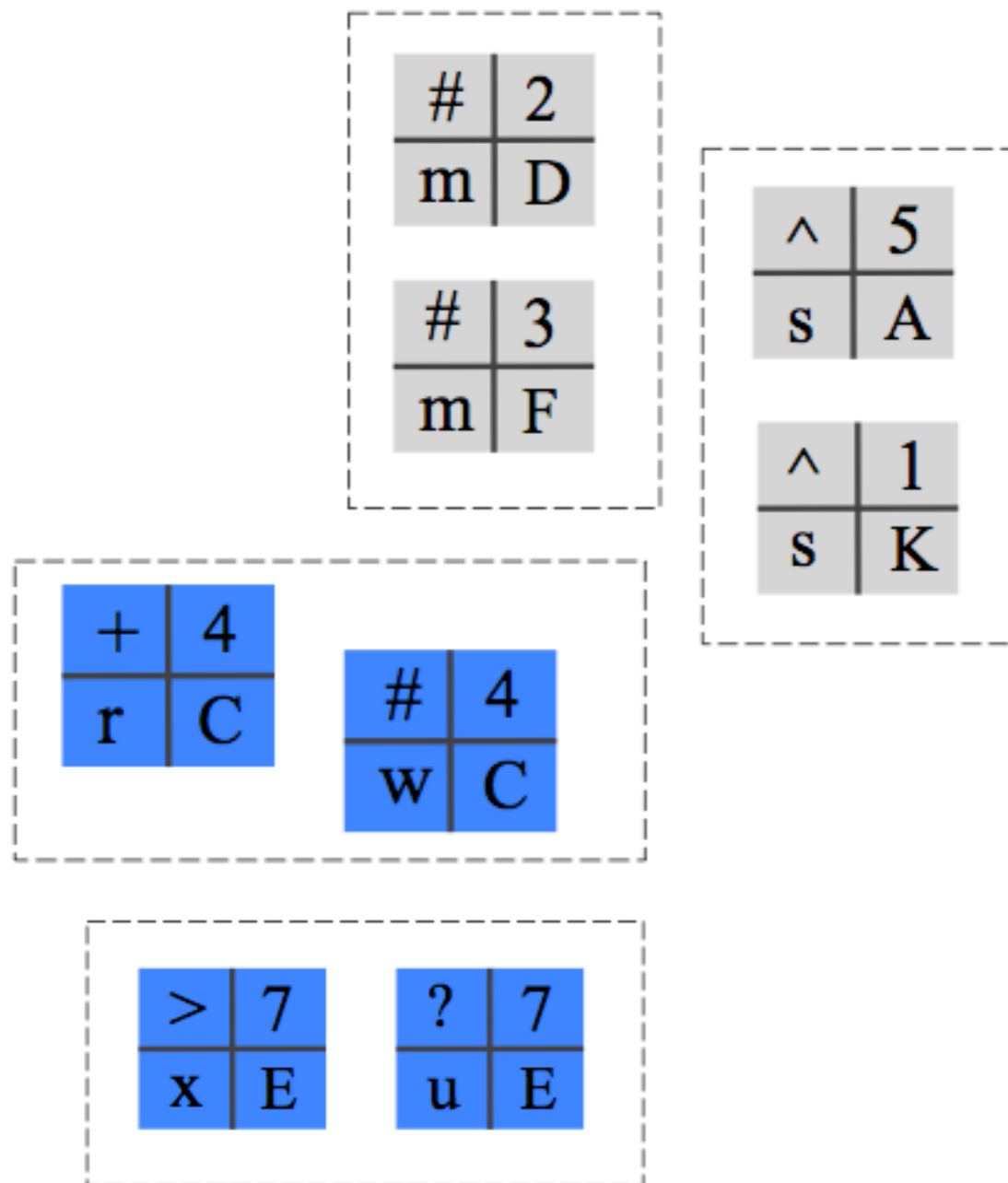


Model performance

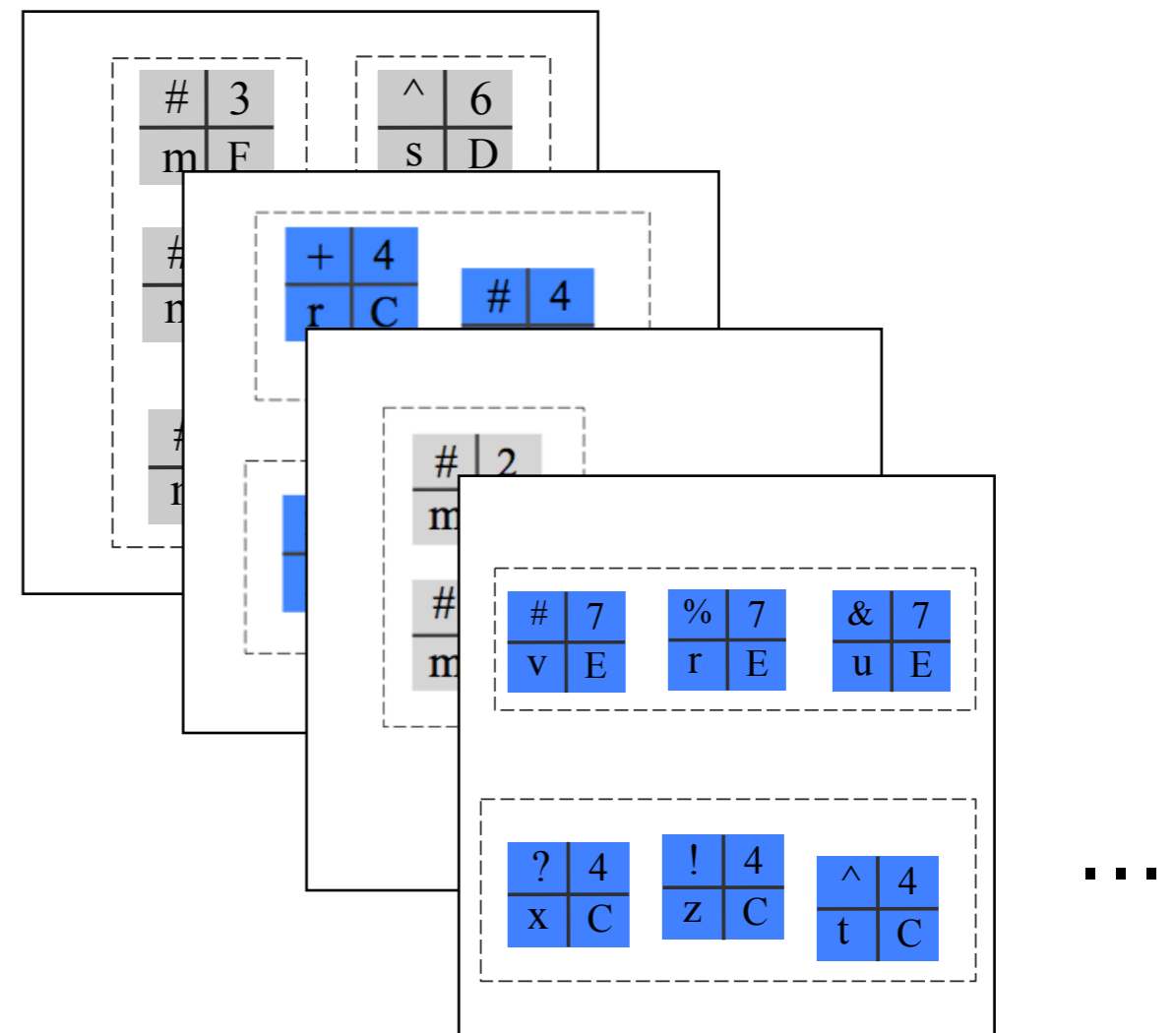
- ▶ Model captures the difference between supervised and unsupervised, but not human failure in the “hard” condition
- ▶ This is a somewhat common issue with many models, particularly Bayesian ones: they capture what performance would be in the ideal case, but don't capture limitations
- ▶ We can see this even more strongly with an extension to the existing experiment...

Experiment extension

Previous version: people saw all exemplars at once



New version: like real life, people need to rely on memory



Experiment extension

To anticipate: It didn't work.

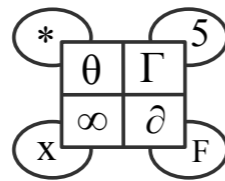
Nothing worked.

People couldn't learn this.

Experiment extension

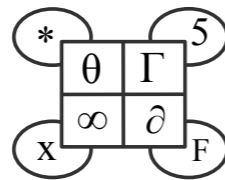
▶ 1 hard (one-kind control)

Eight total features
90% coherent
2 pick out kind, 2
categorise



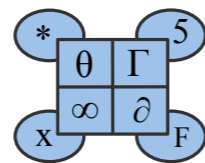
▶ 2 hard (two kinds)

Eight total features
90% coherent
2 pick out kind, 2
different categorise for
each kind



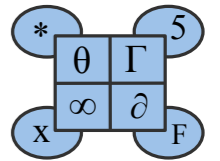
▶ 2 salient

Nine total features (one
identical for all)
One kind feature very
salient
90% coherent
2 pick out kind, 2 different
categorise for each kind



▶ 2 verbal

Just like 2 salient, with instructions
specifically telling people to look
for different ways of categorising.
Analogy of cutlery (shape) vs
shampoo (smell/colour)



▶ 2 few

Five total features
100% coherent
1 picks out kind, 2 different
categorise for each kind

+	4
r	C

▶ 2 grouped

Just like 2 few, but the
features for each category
are grouped in a line

+	4
r	C

▶ 2 mixed

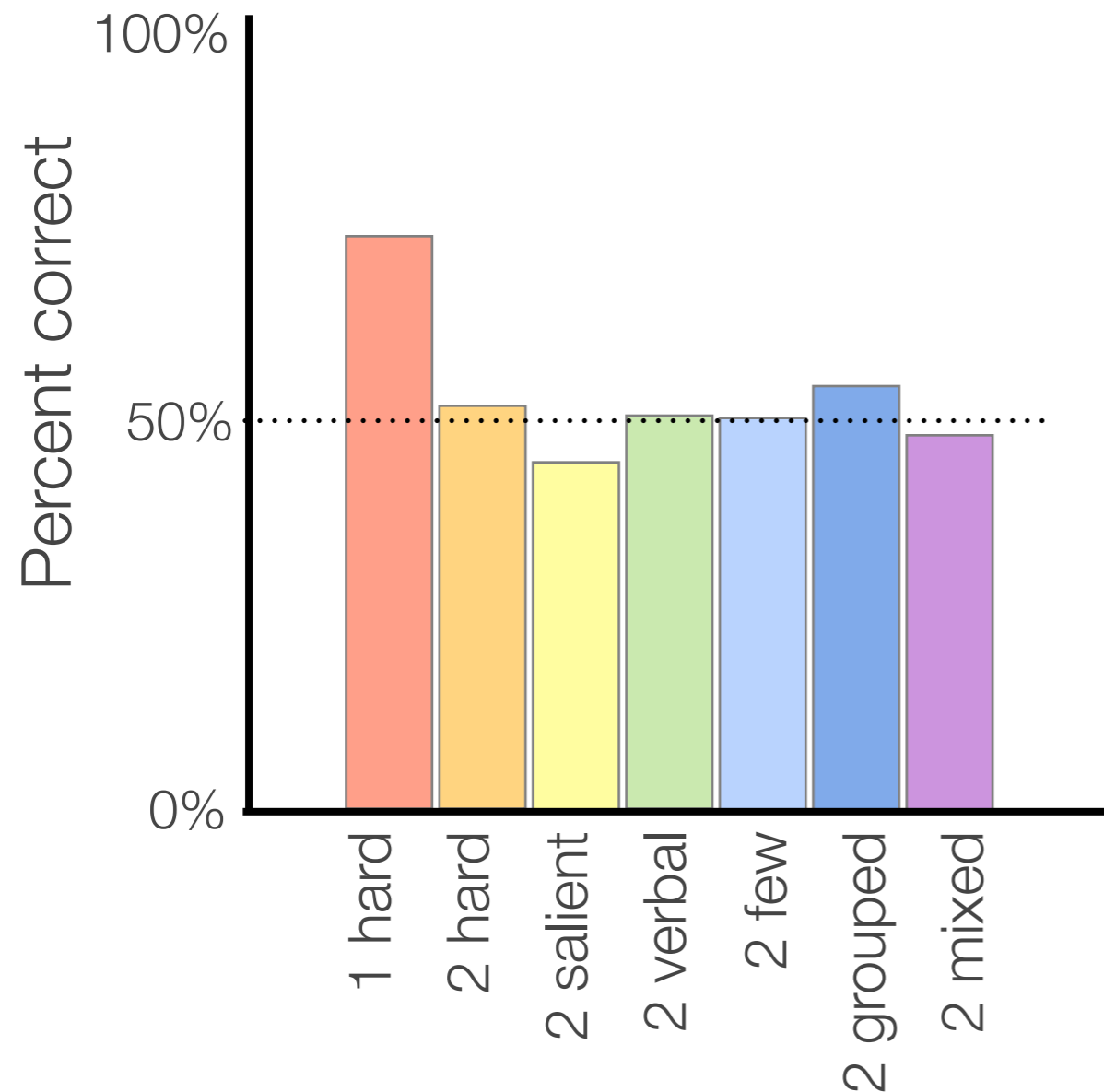
Just like 2 grouped, but
multiple kinds appeared on
screen at once

+	4
r	C

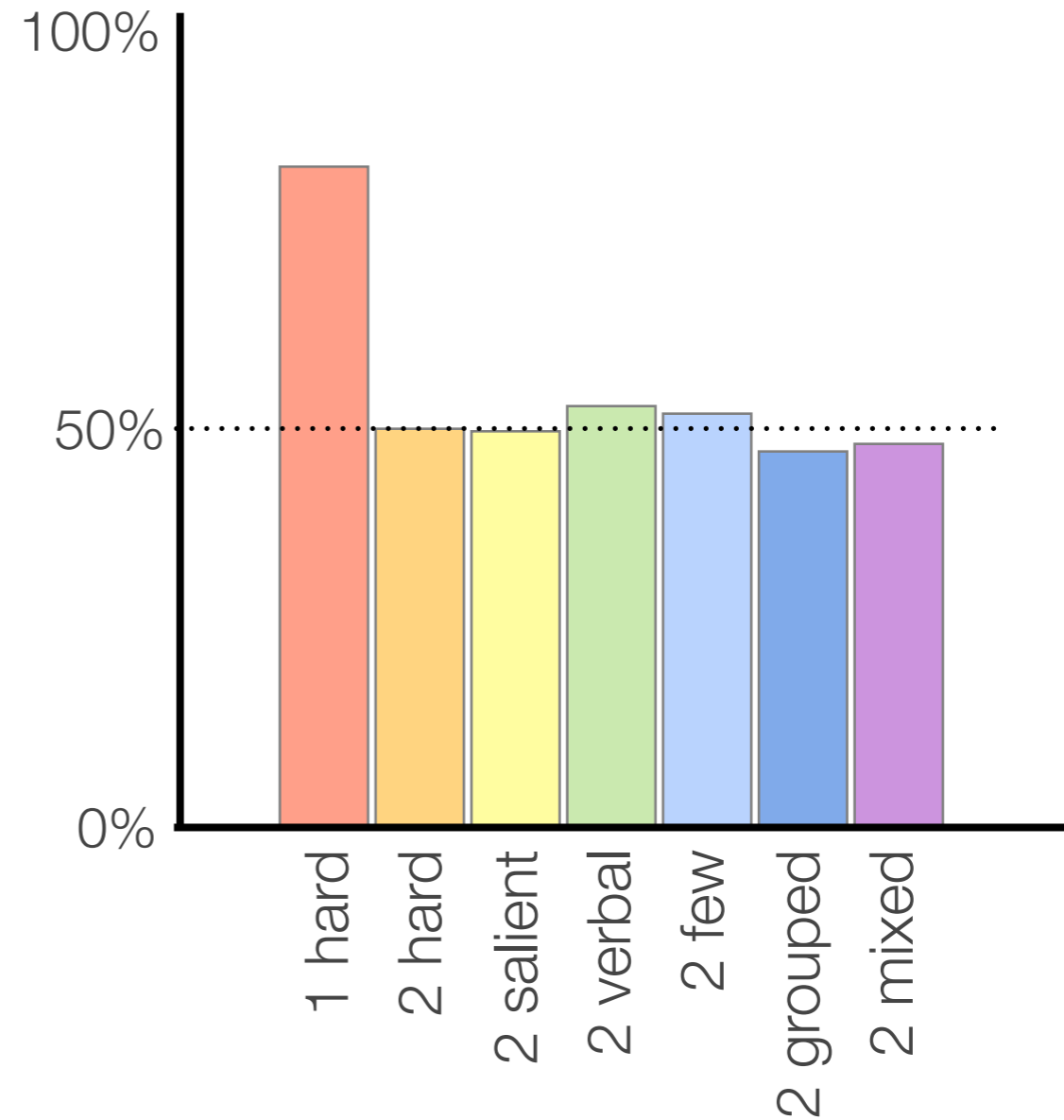
Results (many details elided)

People were unable to learn two kinds simultaneously when they had to rely on their memory!

Unsupervised



Supervised



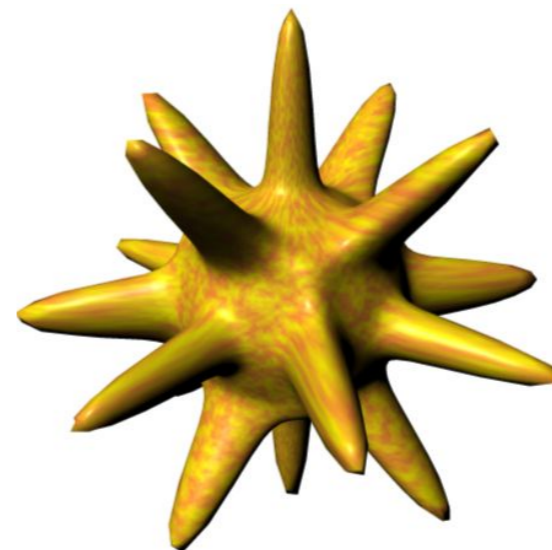
What does this mean?

- ▶ From the psychological perspective
 - There must be some difference between the processes that occur during explicit, fast learning in the lab vs. long-term developmental learning
- ▶ From the computational perspective
 - It would be nice if, in addition to knowing what models can capture performance in the ideal situation, we can account for the limitations people show
 - There are ways to do this, which we'll discuss later. Deliberate “hobbling” of the inference algorithms / muddying data / models of forgetting / etc, added onto these models

Summary

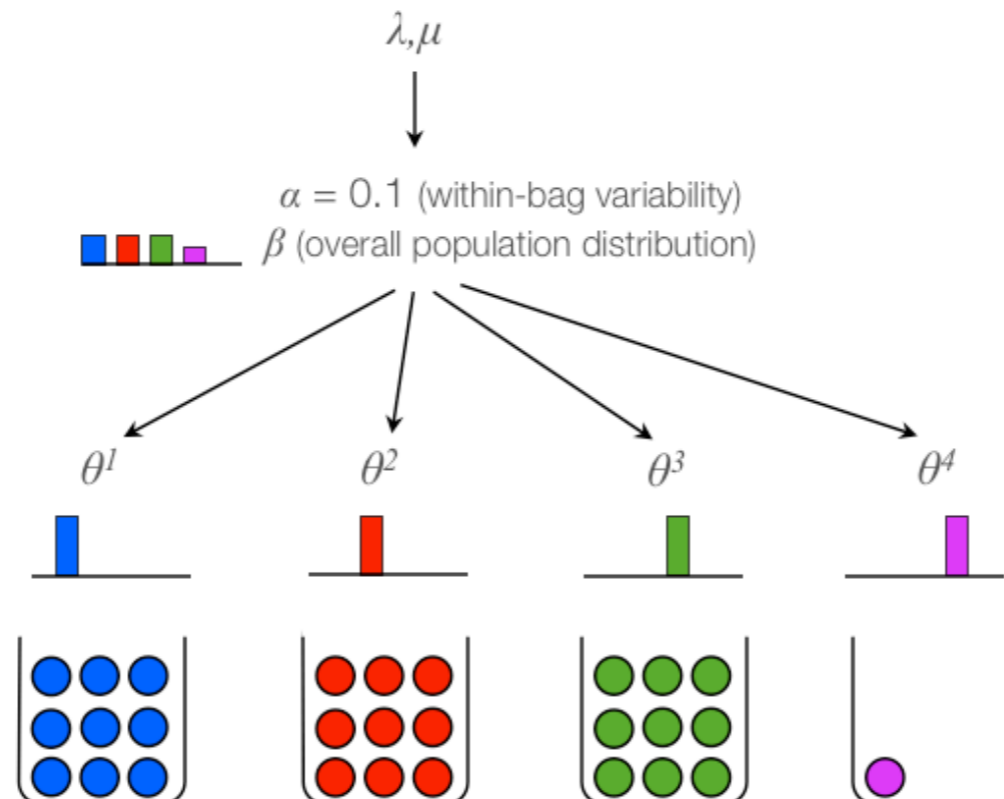
- ▶ People are capable of more complicated inferences than the models we have seen -- learning abstract knowledge about hypotheses (overhypotheses)

dax



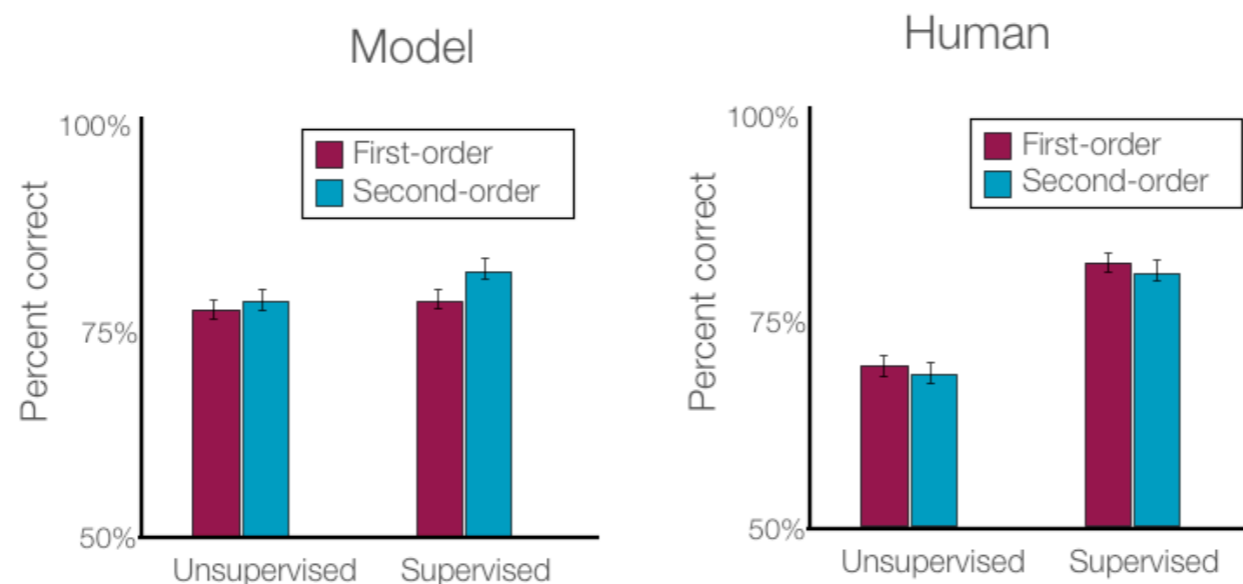
Summary

- ▶ People are capable of more complicated inferences than the models we have seen -- learning abstract knowledge about hypotheses (overhypotheses)
- ▶ We can capture the acquisition of overhypotheses with hierarchical Bayesian models



Summary

- ▶ People are capable of more complicated inferences than the models we have seen -- learning abstract knowledge about hypotheses (overhypotheses)
- ▶ We can capture the acquisition of overhypotheses with hierarchical Bayesian models
- ▶ These models do well at capturing many aspects of human behaviour...



Summary

- ▶ People are capable of more complicated inferences than the models we have seen -- learning abstract knowledge about hypotheses (overhypotheses)
- ▶ We can capture the acquisition of overhypotheses with hierarchical Bayesian models
- ▶ These models do well at capturing many aspects of human behaviour...
- ▶ ... except when humans have to struggle against their capacity limitations (e.g., memory) and the model doesn't. Modifying models to take these issues into account is something we will be returning to.

Additional references (not required)

Shape bias

- ▶ Smith, L., Jones, S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002) Object name learning provides on-the-job training for attention. *Psychological Science* 13: 13-19.

Overhypothesis model

- ▶ Kemp, C., Perfors, A., & Tenenbaum, J. (2007) Learning overhypotheses with hierarchical Bayesian models. *Developmental Science* 10: 307-321.
- ▶ Perfors, A. & Tenenbaum, J. (2009) Learning to learn categories. In Taatgen, N., van Rijn, H., Schomaker, L., & Nerbonne, J. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*: 136-141.