Hierarchical Bayesian Models

Computational Cognitive Science, Lecture 14

Stella Frank, stella.frank@ed.ac.uk

October 31 2019

Hierarchical Bayesian Models

- Modelling concept: "priors over priors"
  Leads to extremely powerful & flexible models

- Cognitive concept: "overhypotheses"
  We can have higher-level hypotheses about the probability of lower-level hypotheses

- Concrete implementation: Dirichlet priors over multinomial likelihoods

Computational Cognitive Science

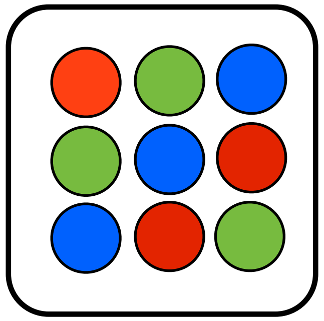
material
shape

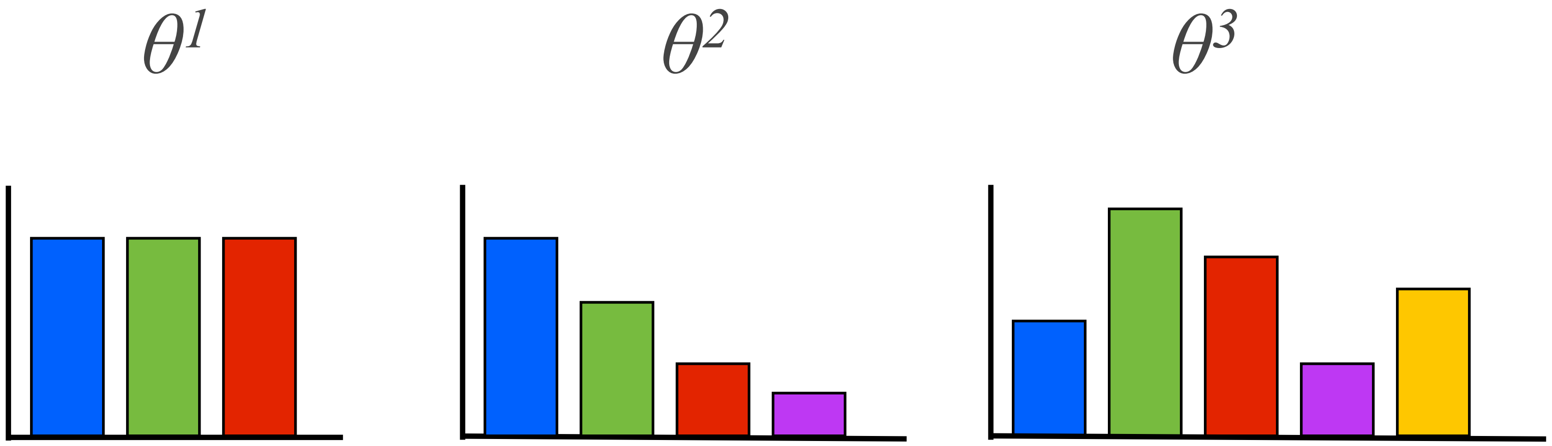
Slides today are from Amy Perfors and Danielle Navarro's excellent Computational Cognitive Science course, online at https://compcogscisydney.org/ccs/

Lecture 11: Higher order knowledge

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?

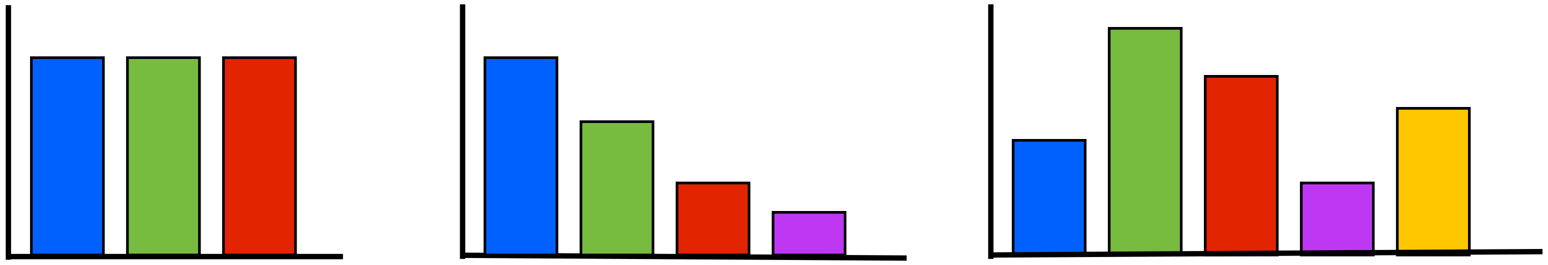
We can also form theories about the nature of the hypotheses themselves: these theories are called *overhypotheses*

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

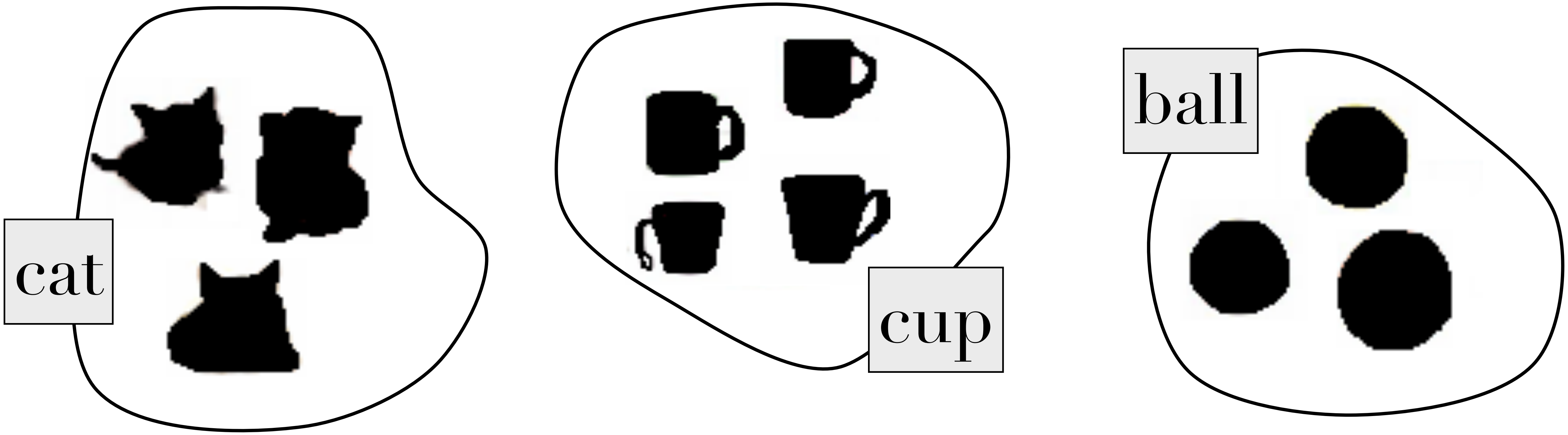
Simple categorisation:
figuring out which things “go together”

Feature variability: one kind of overhypothesis

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

Second-order knowledge (over-hypothesis):

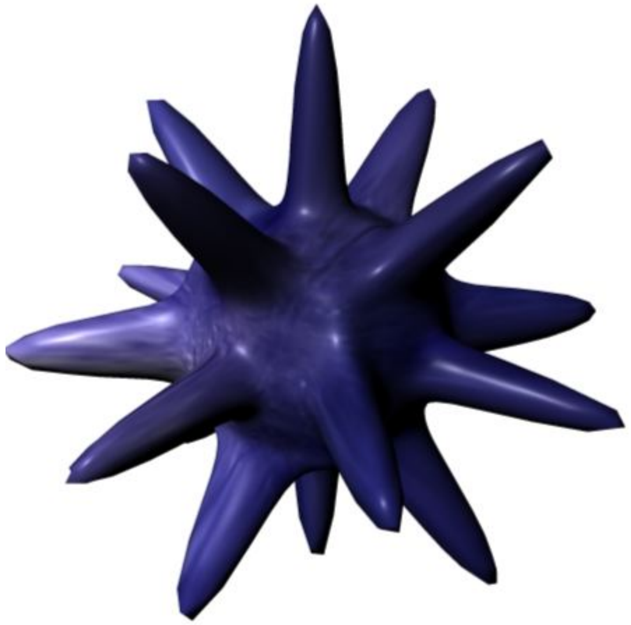
(solid) noun categories are organised by shape

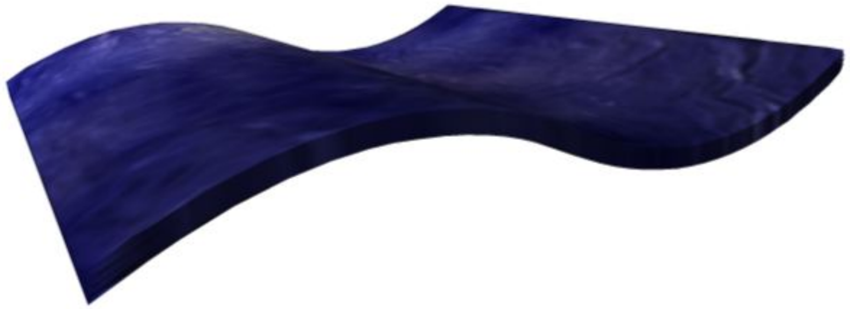


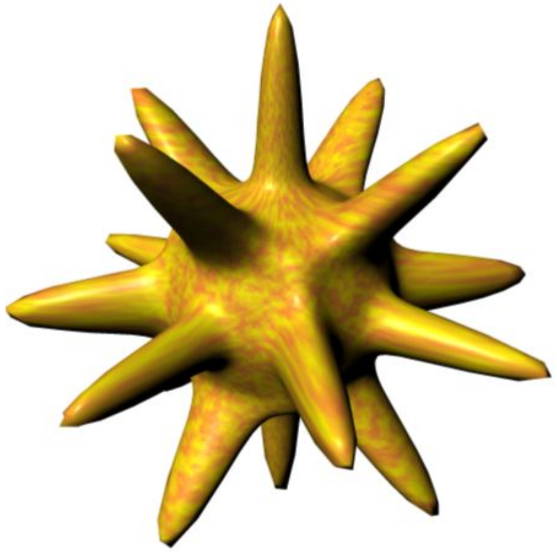
Allows generalisation based on one data point!

Which of the two items on the right are daxes?

dax

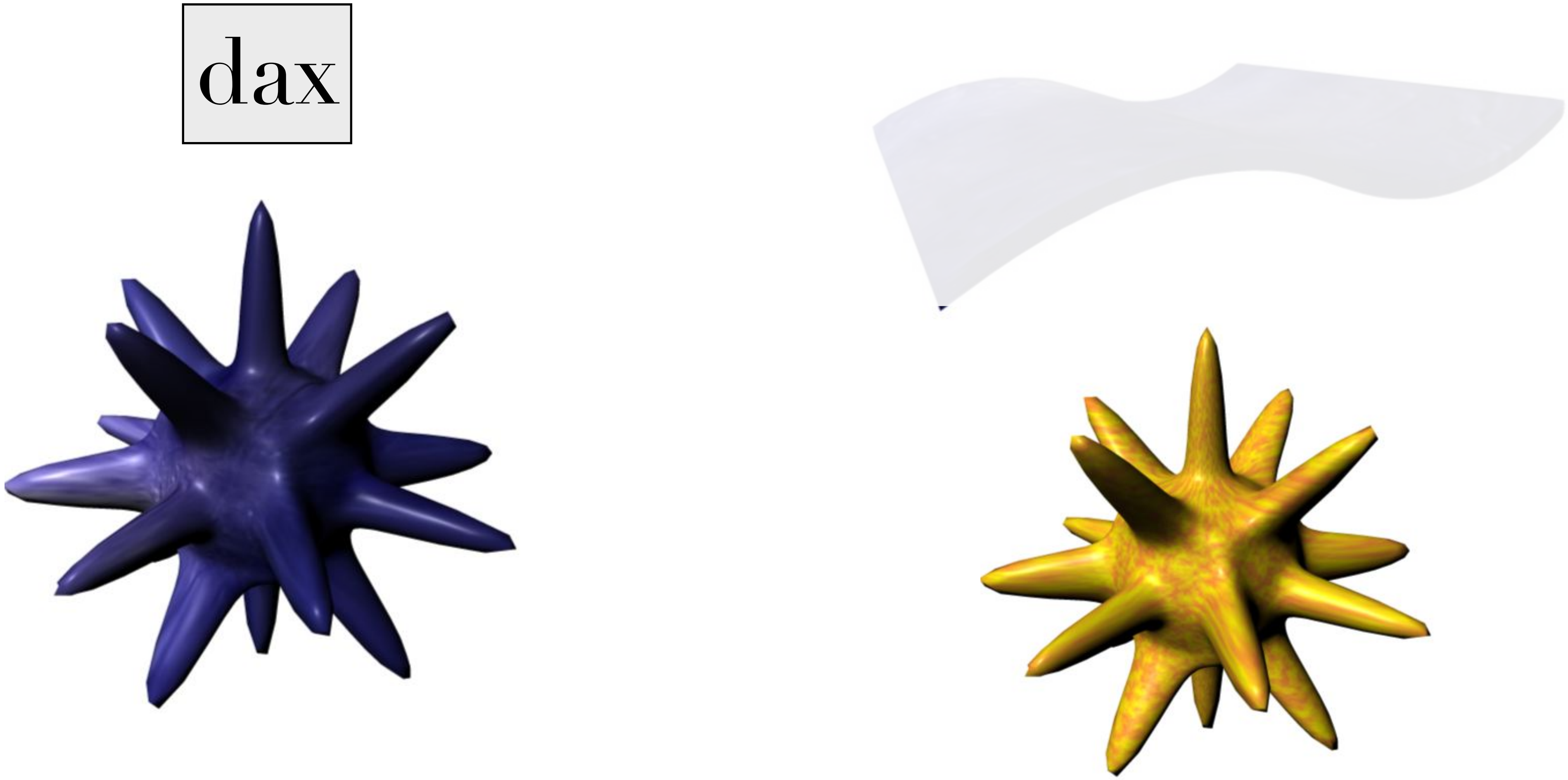






Allows generalisation based on one data point!

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

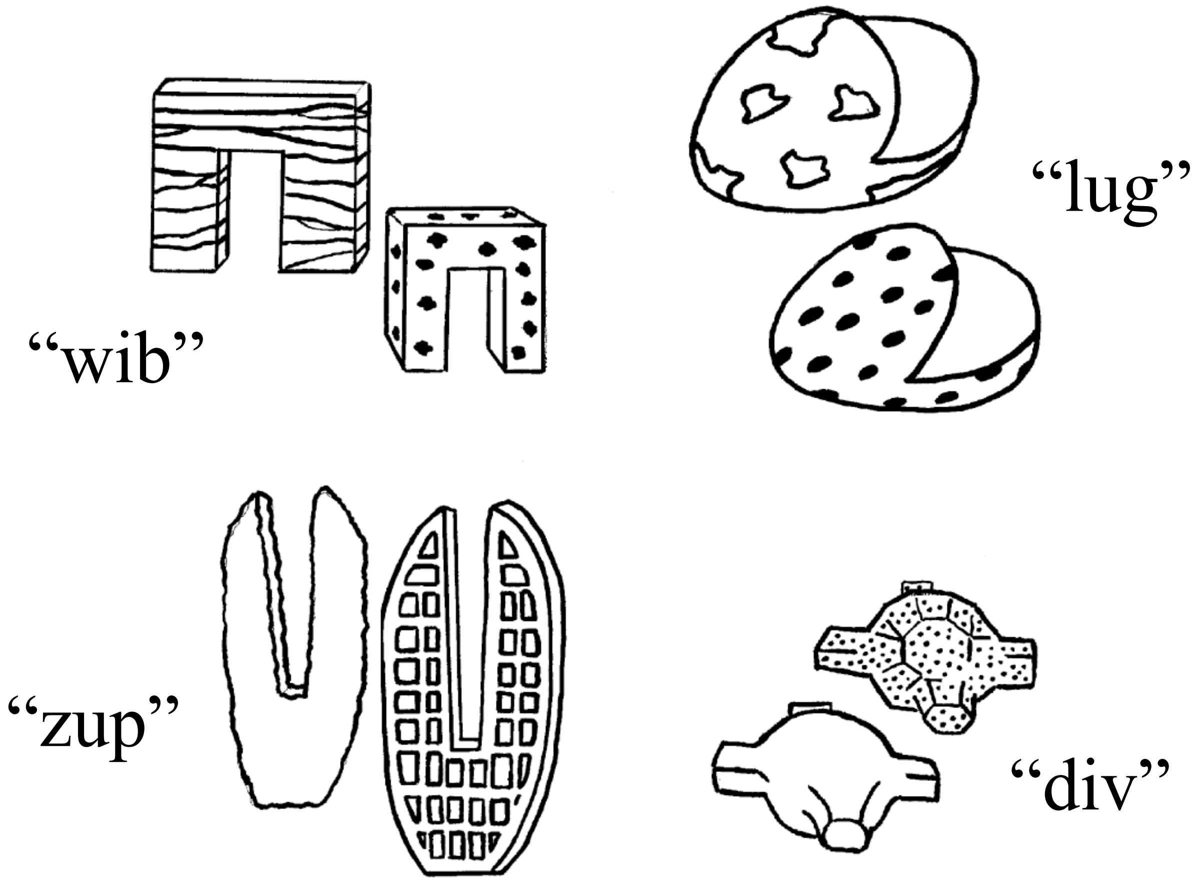


The shape bias emerges at around two years

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

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

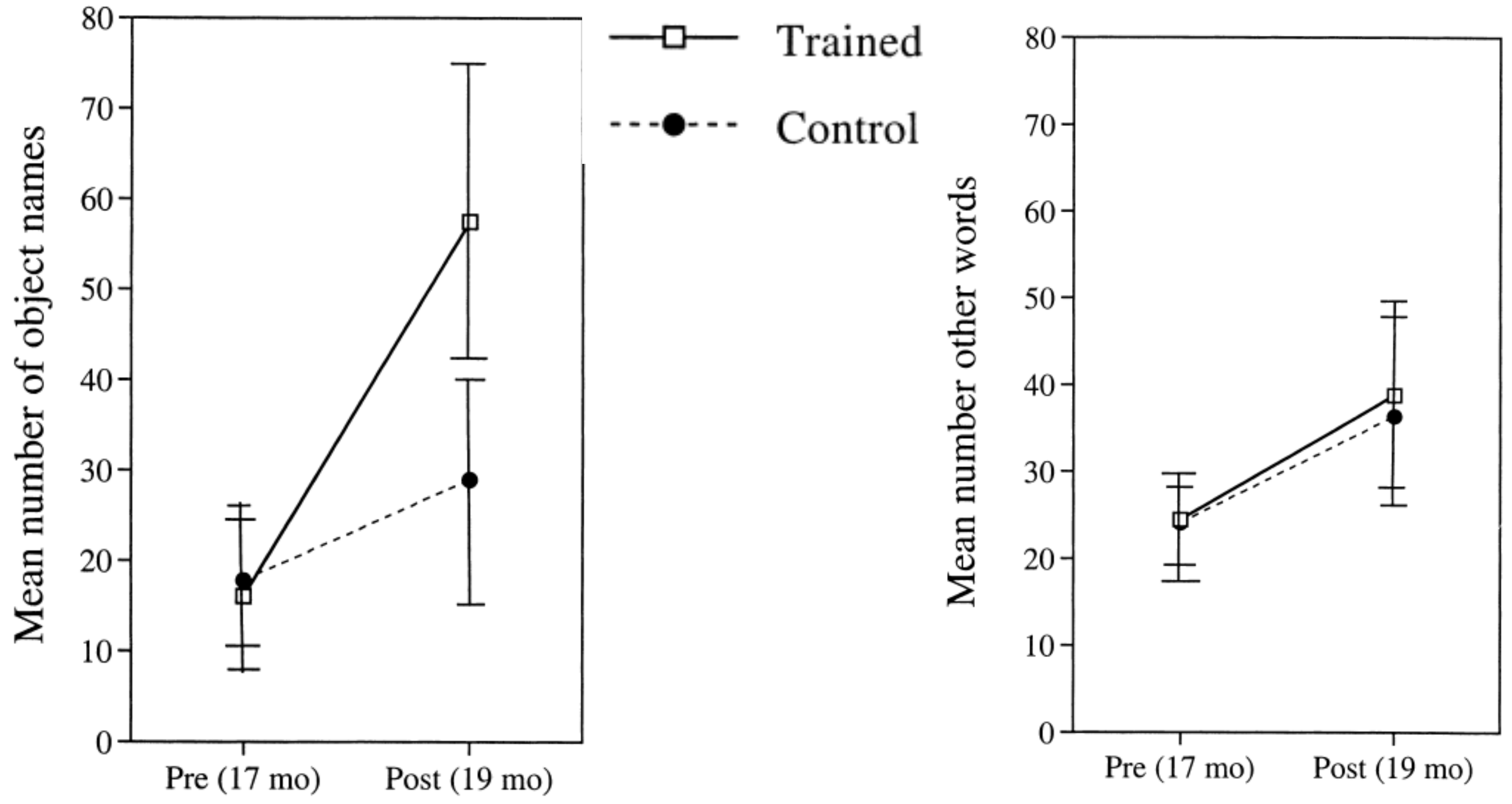
After 8 weeks of training,19-month-olds show the shape bias.



Smith et al., 2002. Object name learning provides on-the-job training for attention. Psych. Sci.

The shape bias emerges at around two years

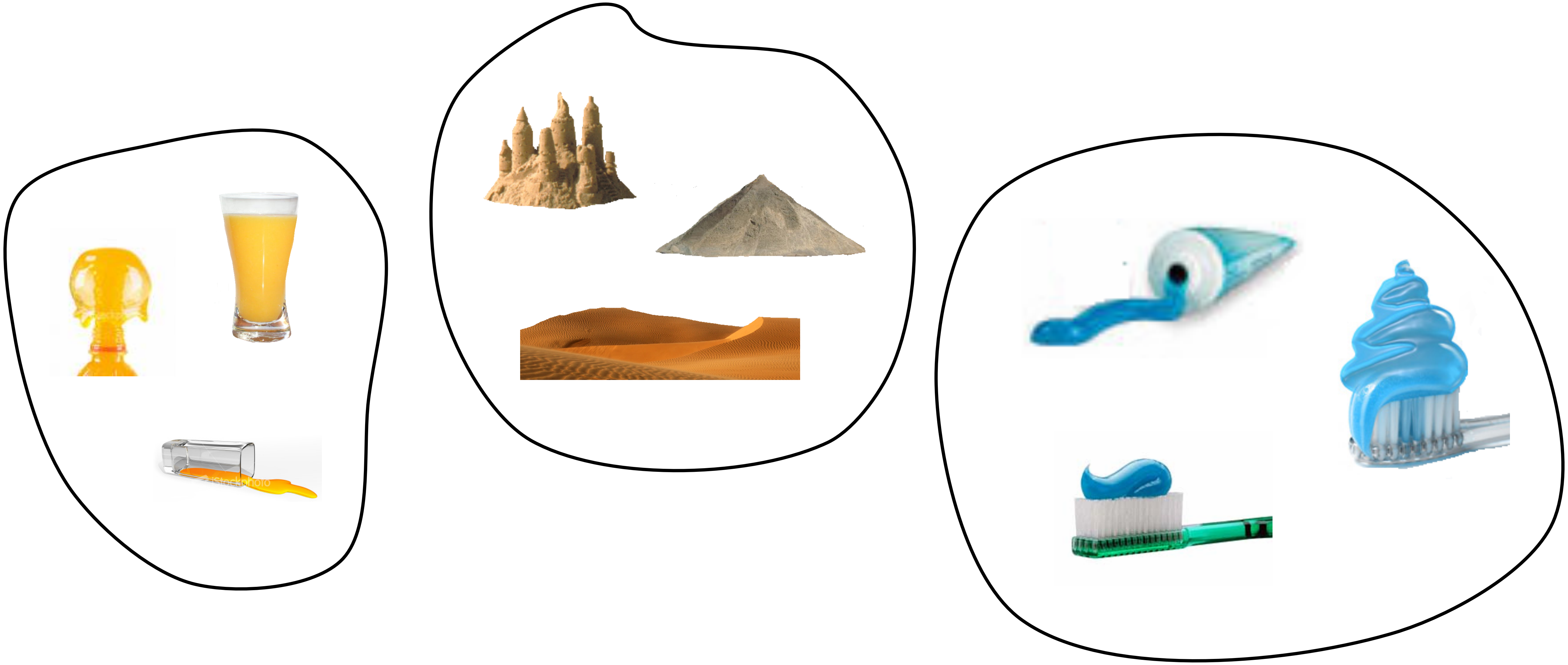
Plus, it helps them when learning other vocabulary!



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!

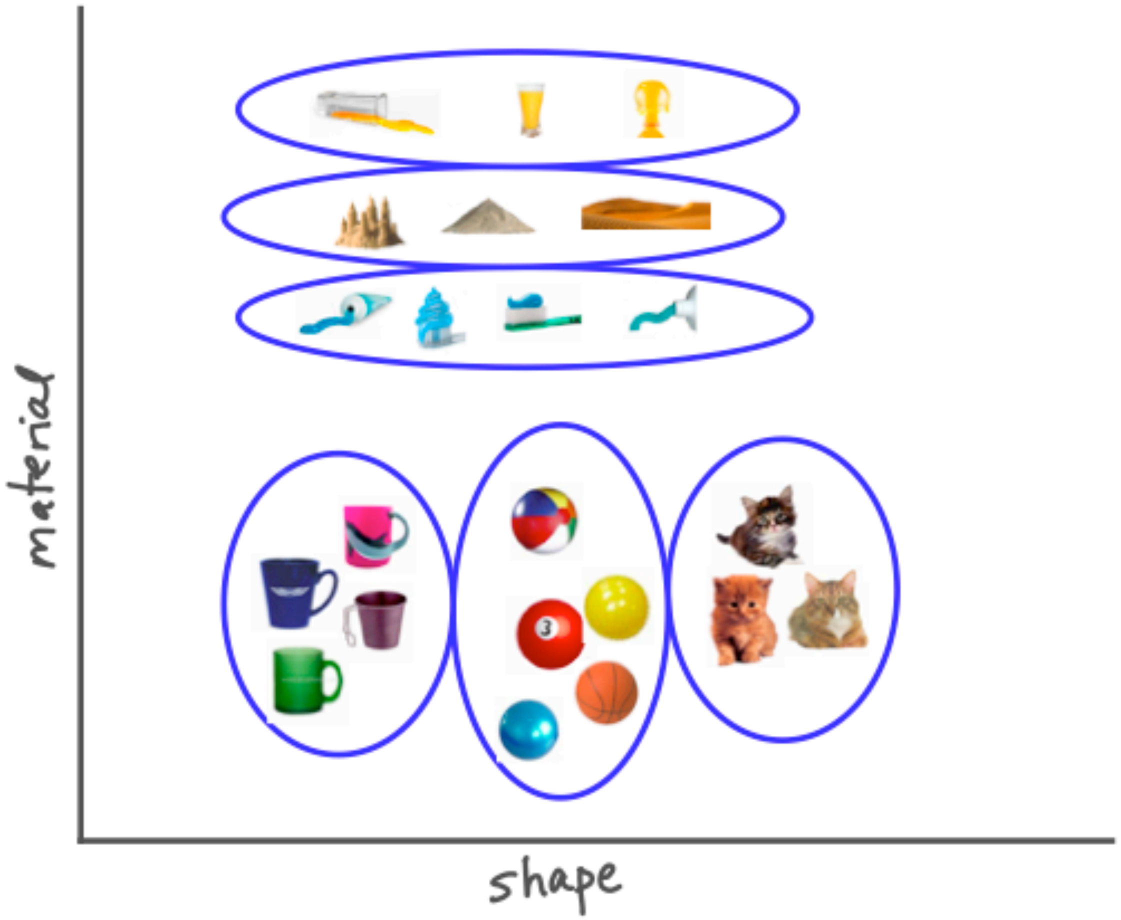


Non-solid substances tend to be organised by colour or texture, not shape

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?

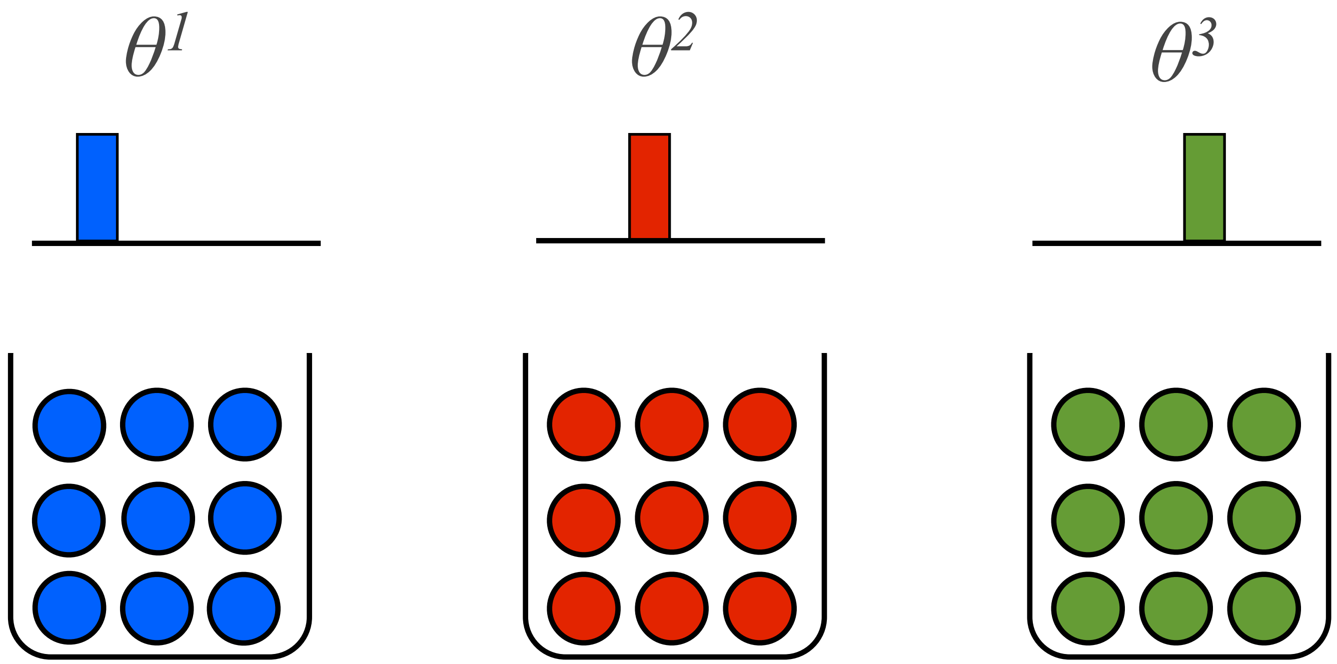
A Bayesian model for overhypothesis learning

- 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

Level 1: Bag proportions

Data



$\boldsymbol{\theta}^1 = [0.98\ 0.01\ 0.01]$ $\boldsymbol{\theta}^2 = [0.01\ 0.98\ 0.01]$ $\boldsymbol{\theta}^3 = [0.01\ 0.01\ 0.98]$

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

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

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

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

Likelihoods: p(y | θ)

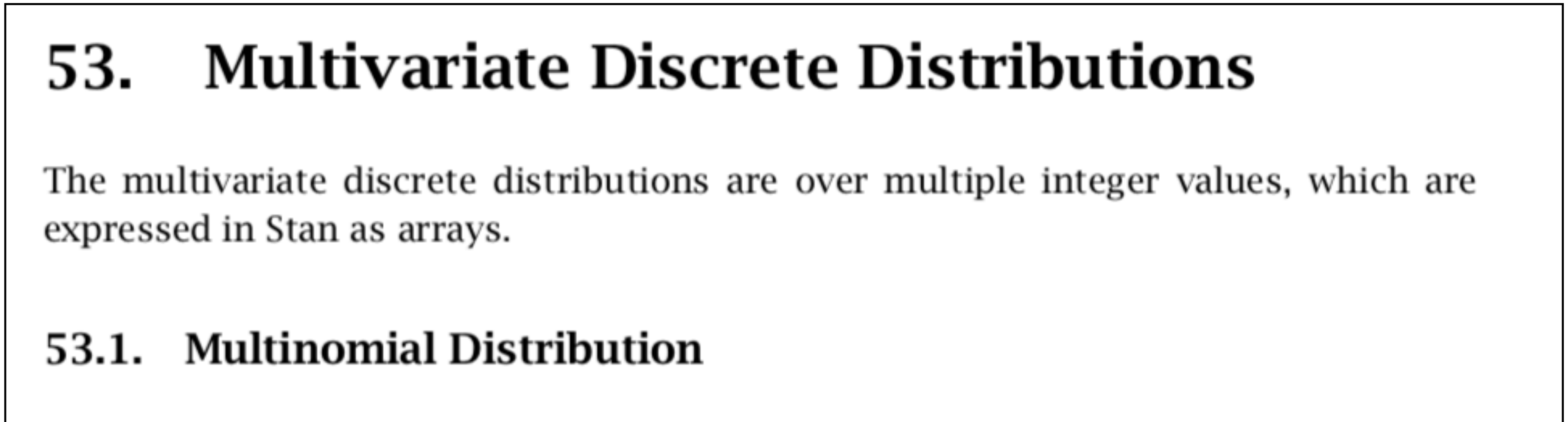
Form of the likelihood (i.e. which distribution to use) is determined (or at least constrained) by the form of the data

- Marbles world: 🔵 🔴
  - discrete outcomes (colours)
  - multiple outcomes (sets of draws)

Likelihoods: p(y | θ)

Likelihoods are usually mostly determined by the question (plus domain knowledge).

- Marbles world:
  - discrete outcomes (colours)
  - multiple outcomes (sets of draws)



53. Multivariate Discrete Distributions

The multivariate discrete distributions are over multiple integer values, which are expressed in Stan as arrays.

53.1. Multinomial Distribution

Other uses for the multinomial

- What's the probability of a bunch of words in a text? (ignoring syntax: "bag of words" model)
- Birthday paradox: what's the probability of two of you having the same birthday?
- Distribution of students to tutorial groups

What's the probability of a bunch of marbles? (i.e., likelihood, aka probability of the evidence)

- Assume we know the distribution of marbles in the bag

$$\theta = [\text{red} = 0.1, \text{green} = 0.4, \text{blue} = 0.5]$$

- and we see 🔵 - what's p(🔵 | θ)?

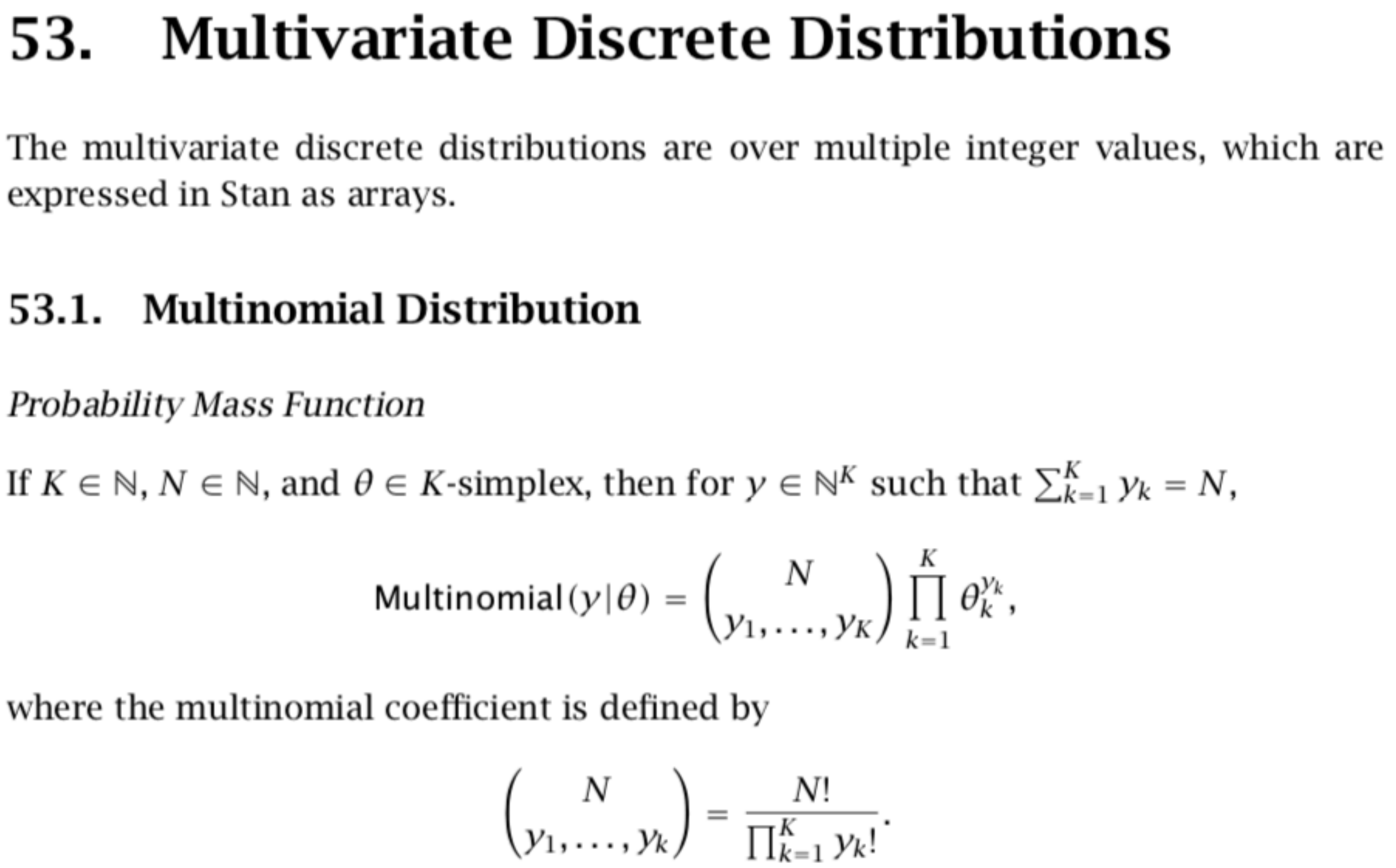
What's the probability of a bunch of marbles? (i.e., likelihood, aka probability of the evidence)

- Assume we know the distribution of marbles in the bag

$$\theta = [\text{red} = 0.1, \text{green} = 0.4, \text{blue} = 0.5]$$

- and we see 🔵 - what's p(🔵 | θ)?
- What about 🔵🔴 ?
- What about 🔵🔵 ?

What's the probability of a bunch of marbles? (i.e., likelihood, aka probability of the evidence)



53. Multivariate Discrete Distributions

The multivariate discrete distributions are over multiple integer values, which are expressed in Stan as arrays.

53.1. Multinomial Distribution

Probability Mass Function

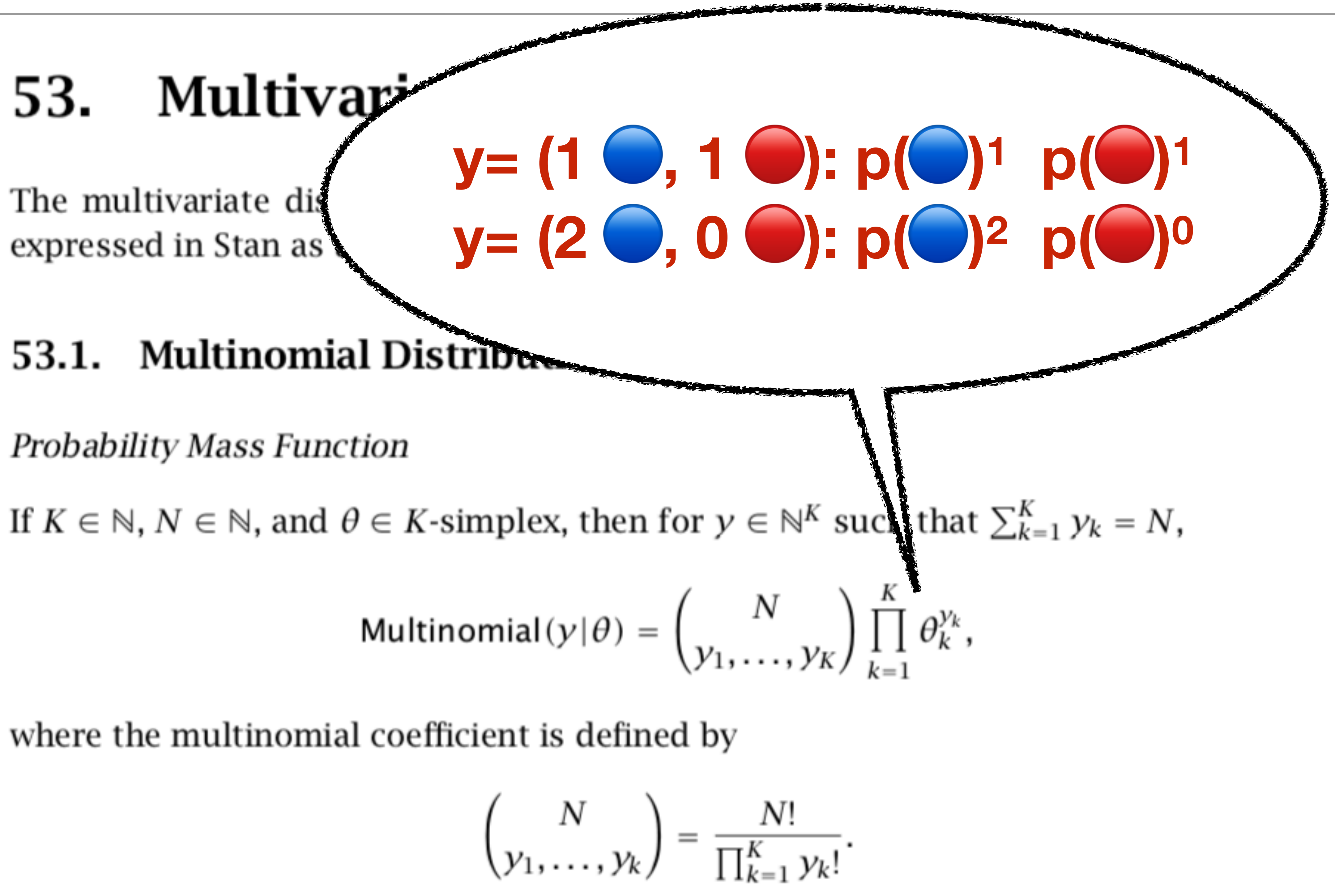
If $K \in \mathbb{N}$, $N \in \mathbb{N}$, and $\theta \in K$-simplex, then for $y \in \mathbb{N}^K$ such that $\sum_{k=1}^{K} y_k = N$,

$$\text{Multinomial}(y|\theta) = \binom{N}{y_1, \ldots, y_K} \prod_{k=1}^{K} \theta_k^{y_k},$$

where the multinomial coefficient is defined by

$$\binom{N}{y_1, \ldots, y_k} = \frac{N!}{\prod_{k=1}^{K} y_k!}.$$

What's the probability of a bunch of marbles?



y= (1 🔵, 1 🔴): p(🔵)^{1} p(🔴)^{1}

y= (2 🔵, 0 🔴): p(🔵)^{2} p(🔴)^{0}

53. Multivari

The multivariate di
expressed in Stan as

53.1. Multinomial Distribu

Probability Mass Function

If $K \in \mathbb{N}$, $N \in \mathbb{N}$, and $\theta \in K$-simplex, then for $y \in \mathbb{N}^K$ such that $\sum_{k=1}^{K} y_k = N$,

$$\text{Multinomial}(y|\theta) = \binom{N}{y_1,\ldots,y_K} \prod_{k=1}^{K} \theta_k^{y_k},$$

where the multinomial coefficient is defined by

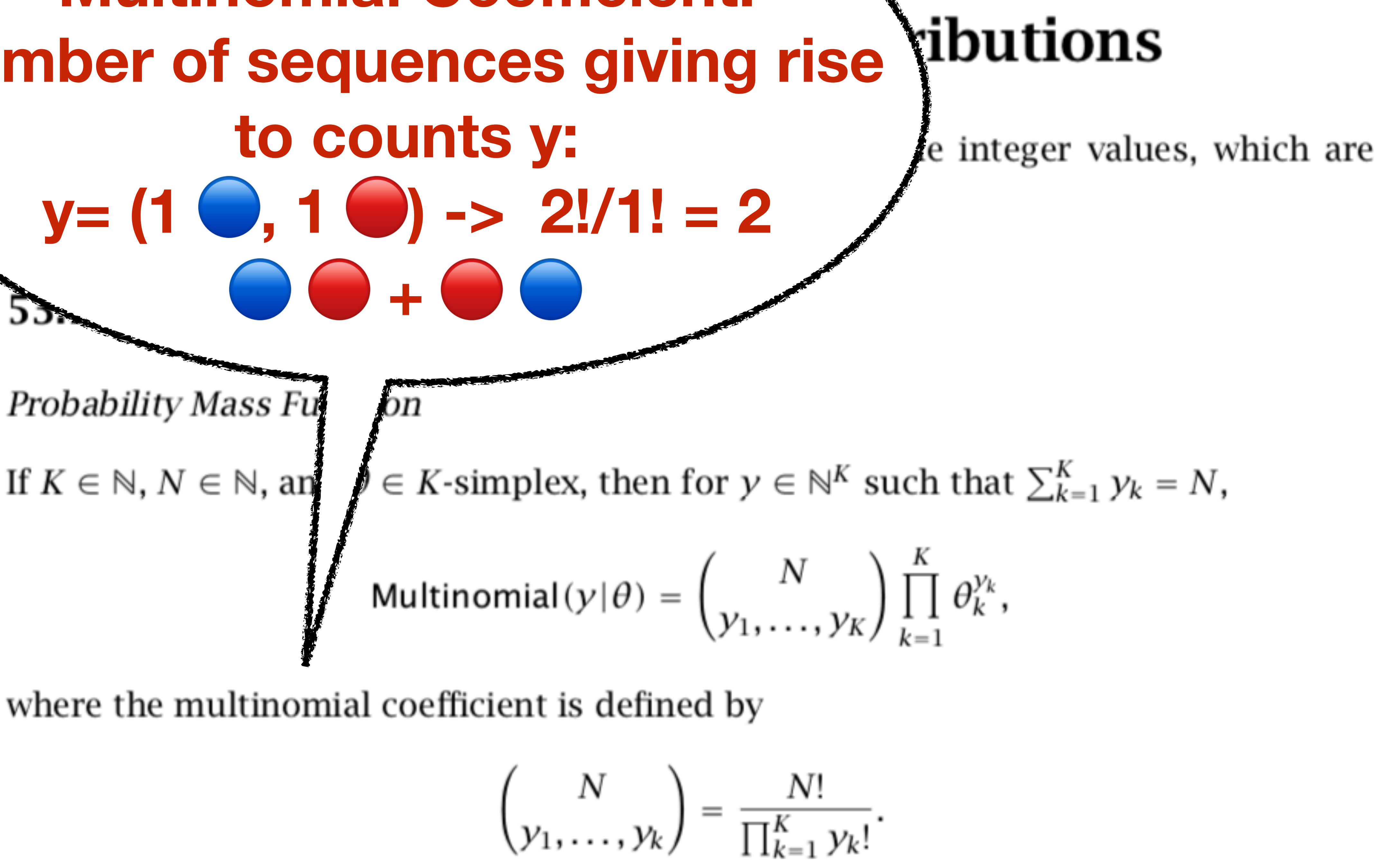
$$\binom{N}{y_1,\ldots,y_k} = \frac{N!}{\prod_{k=1}^{K} y_k!}.$$

What's ... bunch of marbles?

Multinomial Coefficient: number of sequences giving rise to counts y:

y= (1 🔵, 1 🔴) -> 2!/1! = 2

🔵🔴 + 🔴🔵



ributions

e integer values, which are

53.

Probability Mass Fu on

If $K \in \mathbb{N}$, $N \in \mathbb{N}$, an $\in K$-simplex, then for $y \in \mathbb{N}^K$ such that $\sum_{k=1}^{K} y_k = N$,

$$\text{Multinomial}(y|\theta) = \binom{N}{y_1, \ldots, y_K} \prod_{k=1}^{K} \theta_k^{y_k},$$

where the multinomial coefficient is defined by

$$\binom{N}{y_1, \ldots, y_k} = \frac{N!}{\prod_{k=1}^{K} y_k!}.$$

What's the probability of a *sequence* of marbles?

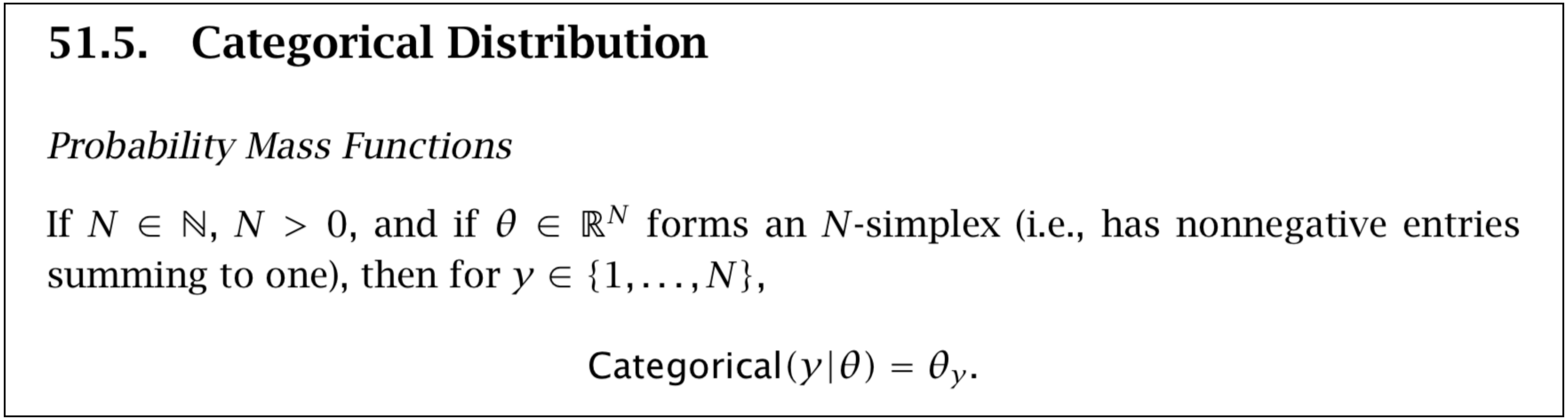
- Assume we know the distribution of marbles in the bag

$$\theta = [\text{red} = 0.1,\ \text{green} = 0.4,\ \text{blue} = 0.5]$$

- and we see 🔵 and then 🔴:

  p(🔵 , 🔴 | θ) = $\theta_{🔵}$ x $\theta_{🔴}$ = 0.05

  Note: no multinomial coefficient here!



51.5. Categorical Distribution

Probability Mass Functions

If $N \in \mathbb{N}$, $N > 0$, and if $\theta \in \mathbb{R}^N$ forms an N-simplex (i.e., has nonnegative entries summing to one), then for $y \in \{1, \ldots, N\}$,

$$\text{Categorical}(y|\theta) = \theta_y.$$

Assumptions of multinomial/categorical distributions

What assumptions do these distributions make about how the data is generated?

Would you use a multinomial or a categorical to model:

- Someone's outfit (as draws from a closet)?
- Language (as draws from a vocabulary)?

Assumptions of multinomial/categorical distributions

Assumptions these distributions make about how the data is generated:

- Independence between draws (draws are iid: independently and identically distributed)

- Variables are categorical; no structure or relations between variables (e.g. no ordering, similarity).

We need a prior!

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

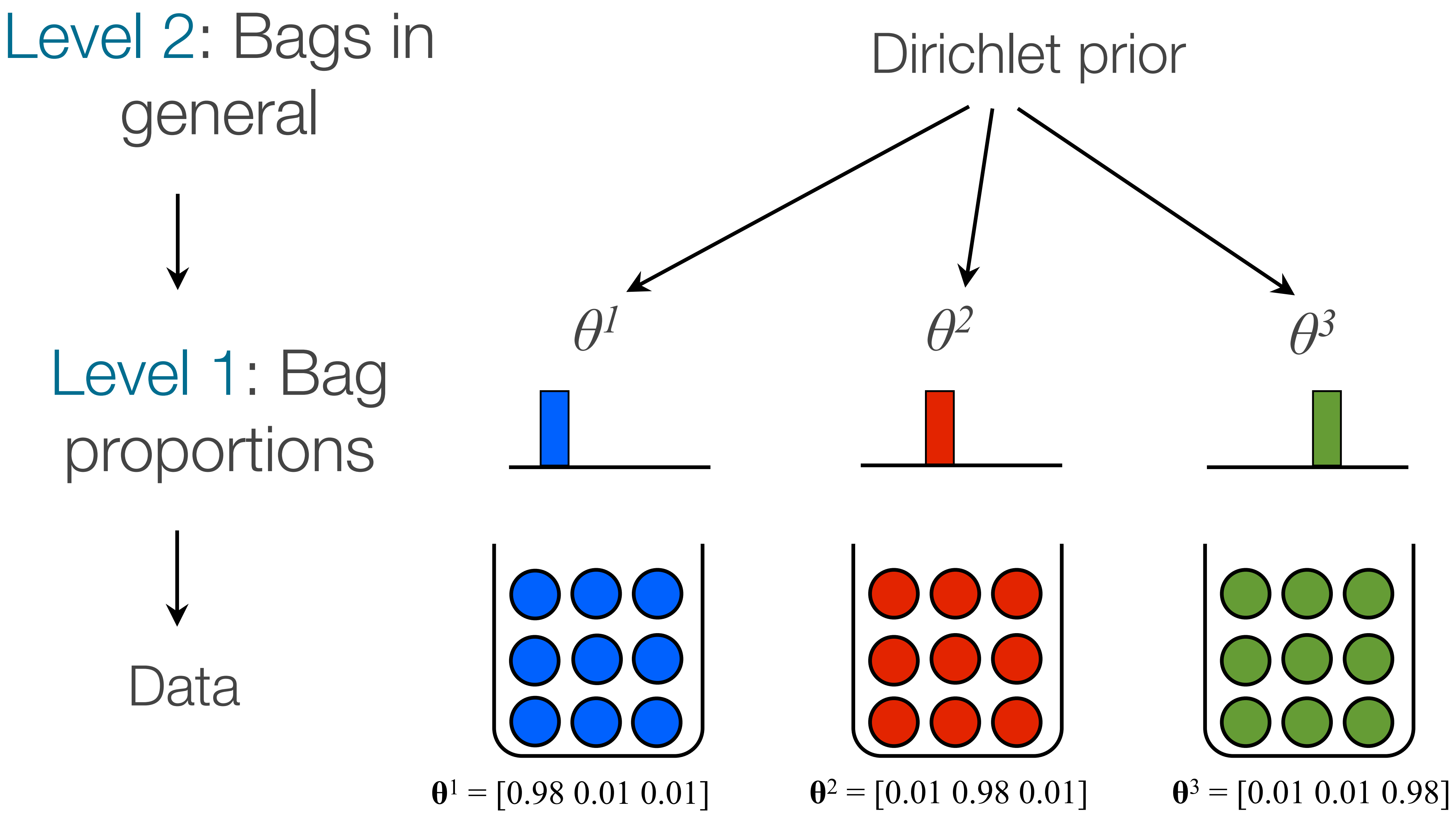
However, in order to calculate *p(θ|y)*, which is what we need to be able to go from the raw data y to the inferred category features, we need a prior over those features.

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

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

Dirichlet prior

θ^1 θ^2 θ^3

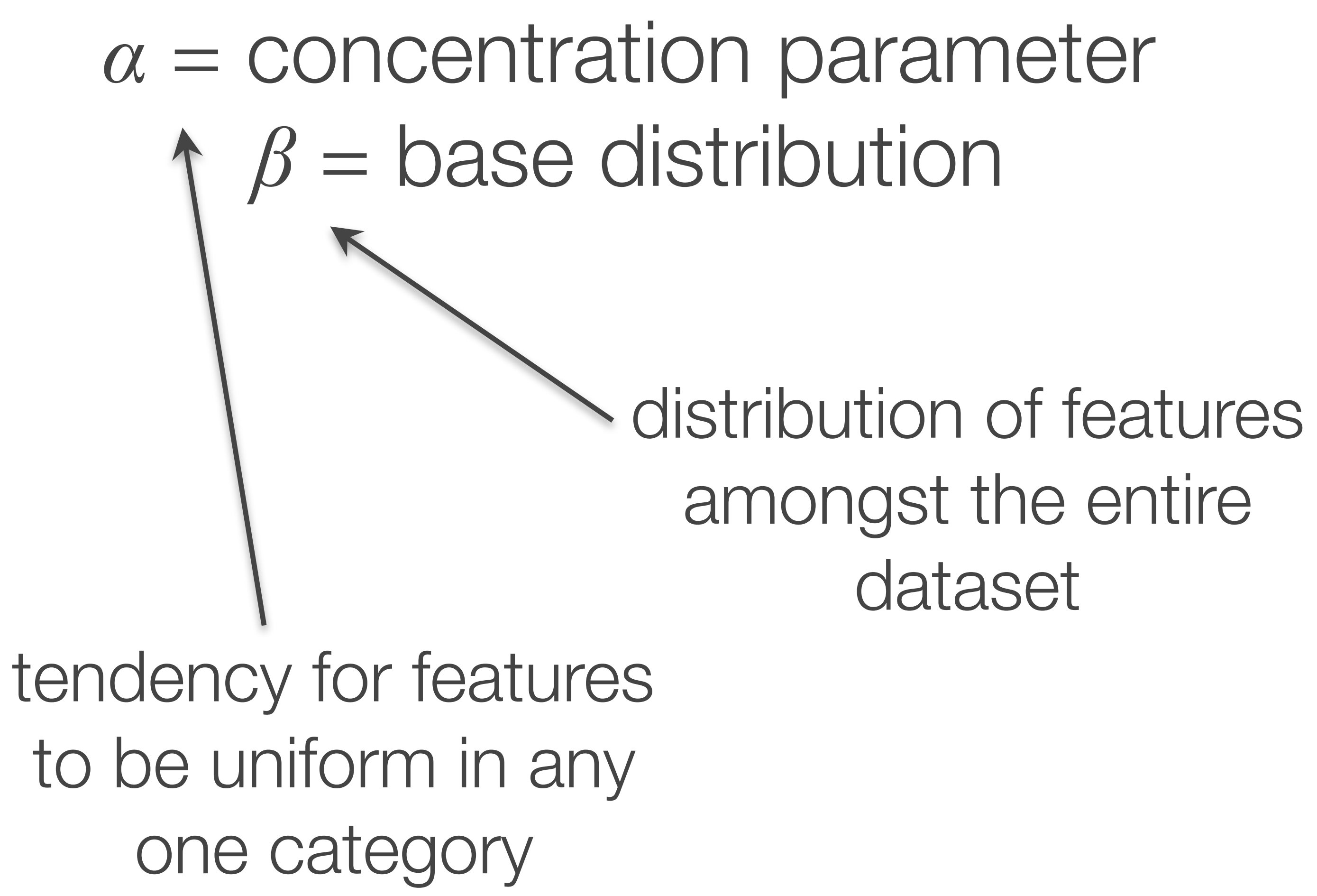
$\boldsymbol{\theta}^1 = [0.98\ 0.01\ 0.01]$

$\boldsymbol{\theta}^2 = [0.01\ 0.98\ 0.01]$

$\boldsymbol{\theta}^3 = [0.01\ 0.01\ 0.98]$

We need a prior!

A Dirichlet distribution consists of two elements:



α = concentration parameter

β = base distribution

distribution of features amongst the entire dataset

tendency for features to be uniform in any one category

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

Dirichlet parameterisations

- $\theta \sim$ Dirichlet($\overrightarrow{\alpha}$) : $\overrightarrow{\alpha}$ is a vector of non-negative values (and no constraint on $\sum_{i=0}^{K} \alpha_i$)

- $\theta \sim$ Dirichlet($\alpha \overrightarrow{\beta}$) : $\overrightarrow{\beta}$ is a probability distribution ($\sum_{i=0}^{K} \beta_i = 1$ and all $\beta_i \geq 0$); α is a scalar

- $\theta \sim$ Dirichlet(α) : scalar parameter with implied uniform $\overrightarrow{\beta}$ distribution

What's this Dirichlet doing in my prior?

- What's θ?
- What's P(θ)?

$$P(\theta | y) = \frac{P(y | \theta)P(\theta)}{\int_{\theta'} P(y | \theta')P(\theta')}$$

What's this Dirichlet doing in my prior?

- What's θ?
  - Parameters of the multinomial distribution over marbles
  - Form: a vector of on-negative numbers that sum to 1: a probability distribution
- What's P(θ)?
  - Prior probability distribution over θ
  - Form: a distribution over distributions
  - "Which θ are a-priori more probable?"

Enter Dirichlet

62. Simplex Distributions

The simplex probabilities have support on the unit K-simplex for a specified K. A K-dimensional vector θ is a unit K-simplex if $\theta_k \geq 0$ for $k \in \{1, \ldots, K\}$ and $\sum_{k=1}^{K} \theta_k = 1$.

62.1. Dirichlet Distribution

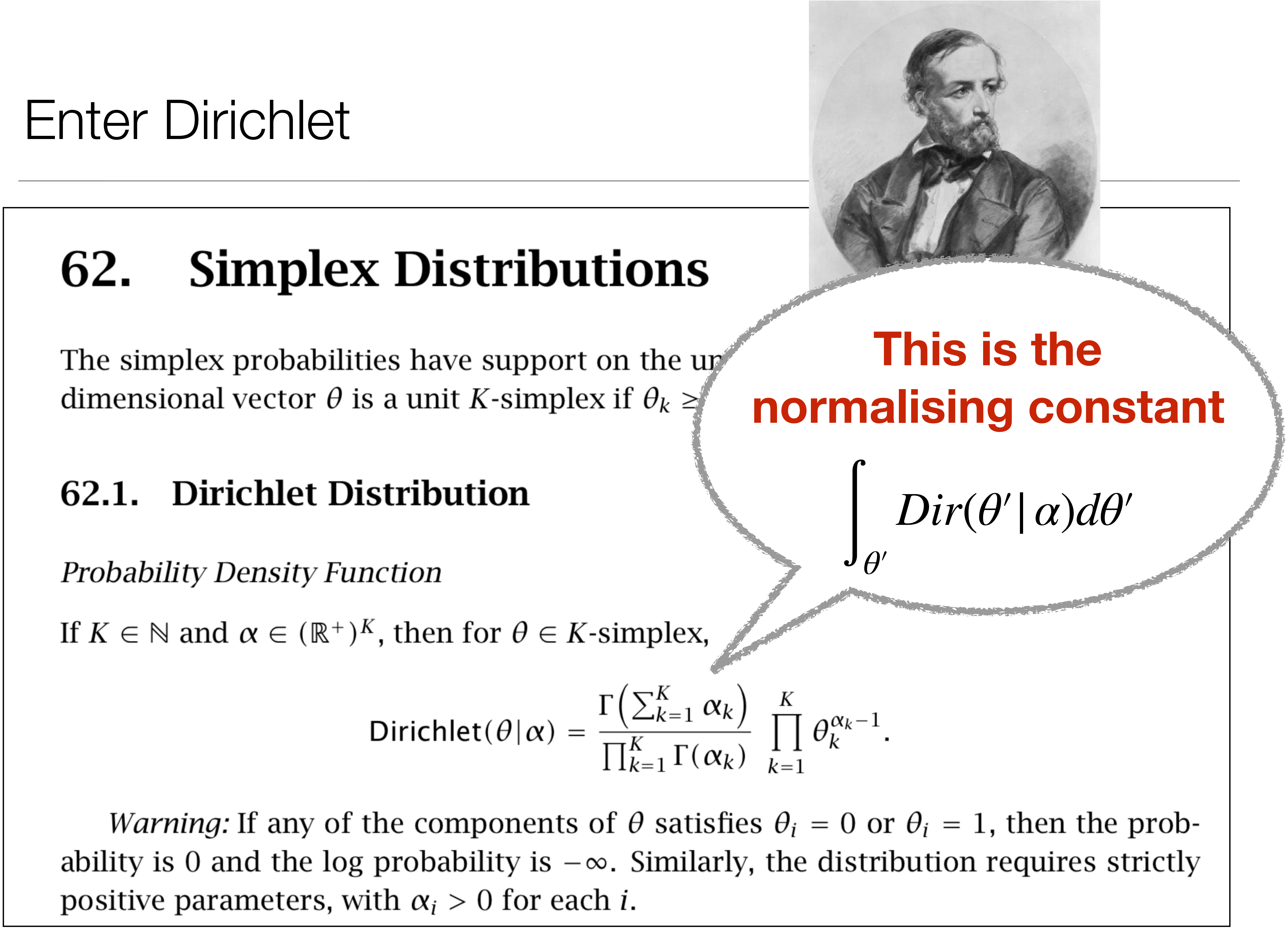
Probability Density Function

If $K \in \mathbb{N}$ and $\alpha \in (\mathbb{R}^+)^K$, then for $\theta \in K$-simplex,

$$\text{Dirichlet}(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}.$$

Warning: If any of the components of θ satisfies $\theta_i = 0$ or $\theta_i = 1$, then the probability is 0 and the log probability is $-\infty$. Similarly, the distribution requires strictly positive parameters, with $\alpha_i > 0$ for each i.

Enter Dirichlet



62. Simplex Distributions

The simplex probabilities have support on the un
dimensional vector θ is a unit K-simplex if $\theta_k \geq$

62.1. Dirichlet Distribution

Probability Density Function

If $K \in \mathbb{N}$ and $\alpha \in (\mathbb{R}^+)^K$, then for $\theta \in K$-simplex,

$$\text{Dirichlet}(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}.$$

Warning: If any of the components of θ satisfies $\theta_i = 0$ or $\theta_i = 1$, then the probability is 0 and the log probability is $-\infty$. Similarly, the distribution requires strictly positive parameters, with $\alpha_i > 0$ for each i.

This is the normalising constant

$$\int_{\theta'} Dir(\theta'|\alpha)d\theta'$$

Posterior distribution of Dirichlet-Categorical

$$\begin{aligned}
P(\theta|y, \boldsymbol{\alpha}) &\propto P(y|\theta)P(\theta|\boldsymbol{\alpha}) \\
&= \mathrm{Cat}(y|\theta)\mathrm{Dir}(\theta|\boldsymbol{\alpha}) \\
&= \prod_{k=1}^{K} \theta_k^{N_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{i=k}^{K} \Gamma(\alpha_k)} \\
&\propto \prod_{k=1}^{K} \theta_k^{N_k} \theta_k^{\alpha_k - 1} \\
&= \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}
\end{aligned}$$

where N_k is the number of items of category k in y.

Posterior distribution of Dirichlet-Categorical

$$\begin{aligned}
P(\theta|y,\boldsymbol{\alpha}) &\propto P(y|\theta)P(\theta|\boldsymbol{\alpha}) \\
&= \mathsf{Cat}(y|\theta)\mathsf{Dir}(\theta|\boldsymbol{\alpha}) \\
&= \prod_{k=1}^{K} \theta_k^{N_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{i=k}^{K} \Gamma(\alpha_k)} \\
&\propto \prod_{k=1}^{K} \theta_k^{N_k} \theta_k^{\alpha_k - 1} \\
&= \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}
\end{aligned}$$

where N_k is the number of items of category k in y.

Posterior is also a Dirichlet: $P(\theta|y,\alpha) = Dir(\alpha')$, where $\alpha' = \alpha + y$; i.e. $\alpha'_k = \alpha_k + N_k$

Posterior distribution of Dirichlet-Categorical

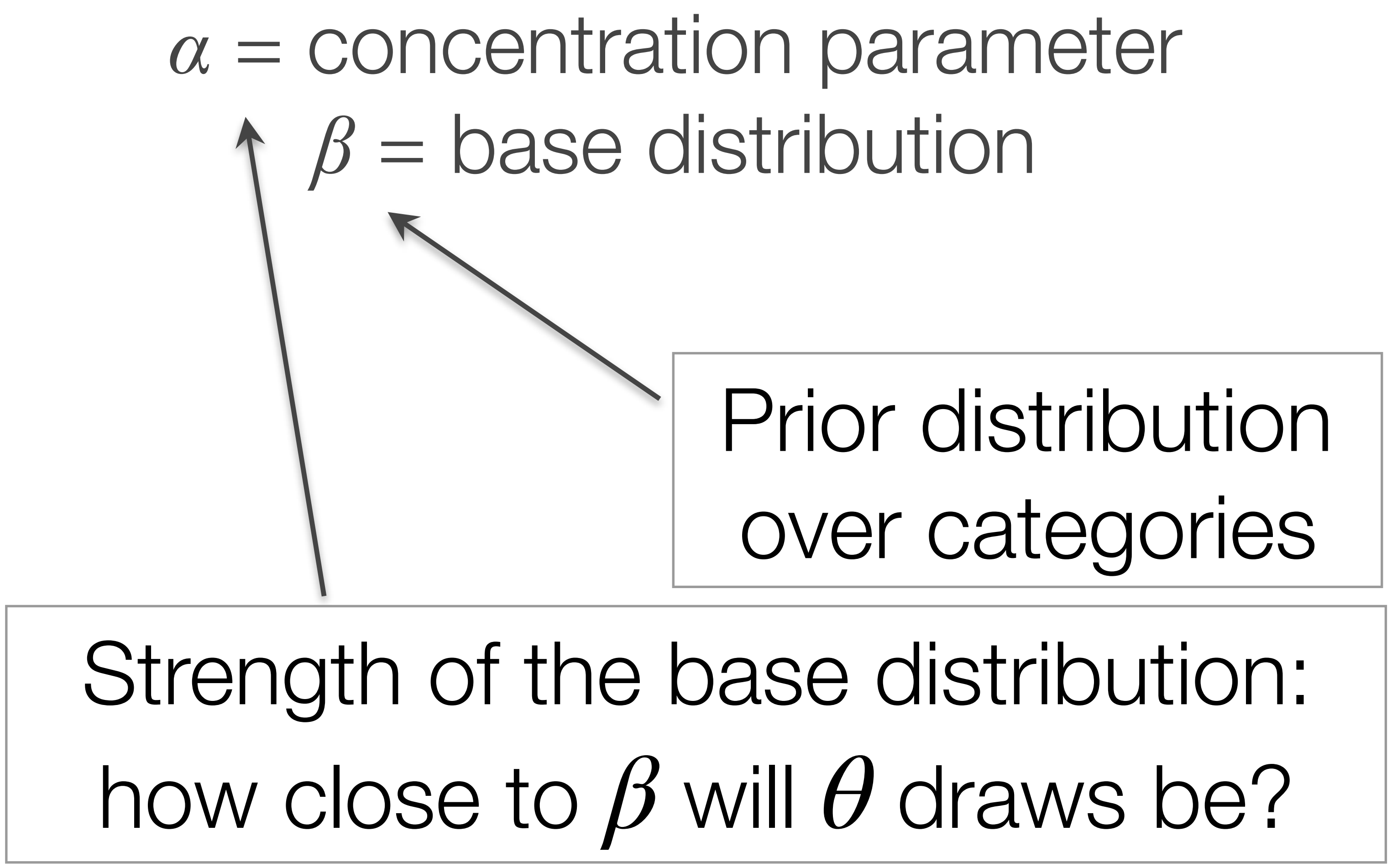
Posterior is also a Dirichlet: $P(\theta|y, \alpha) = Dir(\alpha')$,
where $\alpha' = \alpha + y$; i.e. $\alpha'_k = \alpha_k + N_k$

This is called *conjugacy*: Prior and posterior are same type of distribution, given a certain type of likelihood.
(Dirichlet is the conjugate prior for the categorical and the multinomial.)

Posterior Dirichlet is parameterised by counts in data y plus “pseudocounts” $\boldsymbol{\alpha}$.
Large values of $\boldsymbol{\alpha}$ (or $\alpha\boldsymbol{\beta}$) can thus overwhelm the data; these are “stronger” priors.

We need a prior!

A Dirichlet distribution consists of two elements:



α = concentration parameter

β = base distribution

Prior distribution over categories

Strength of the base distribution: how close to β will θ draws be?

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

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)

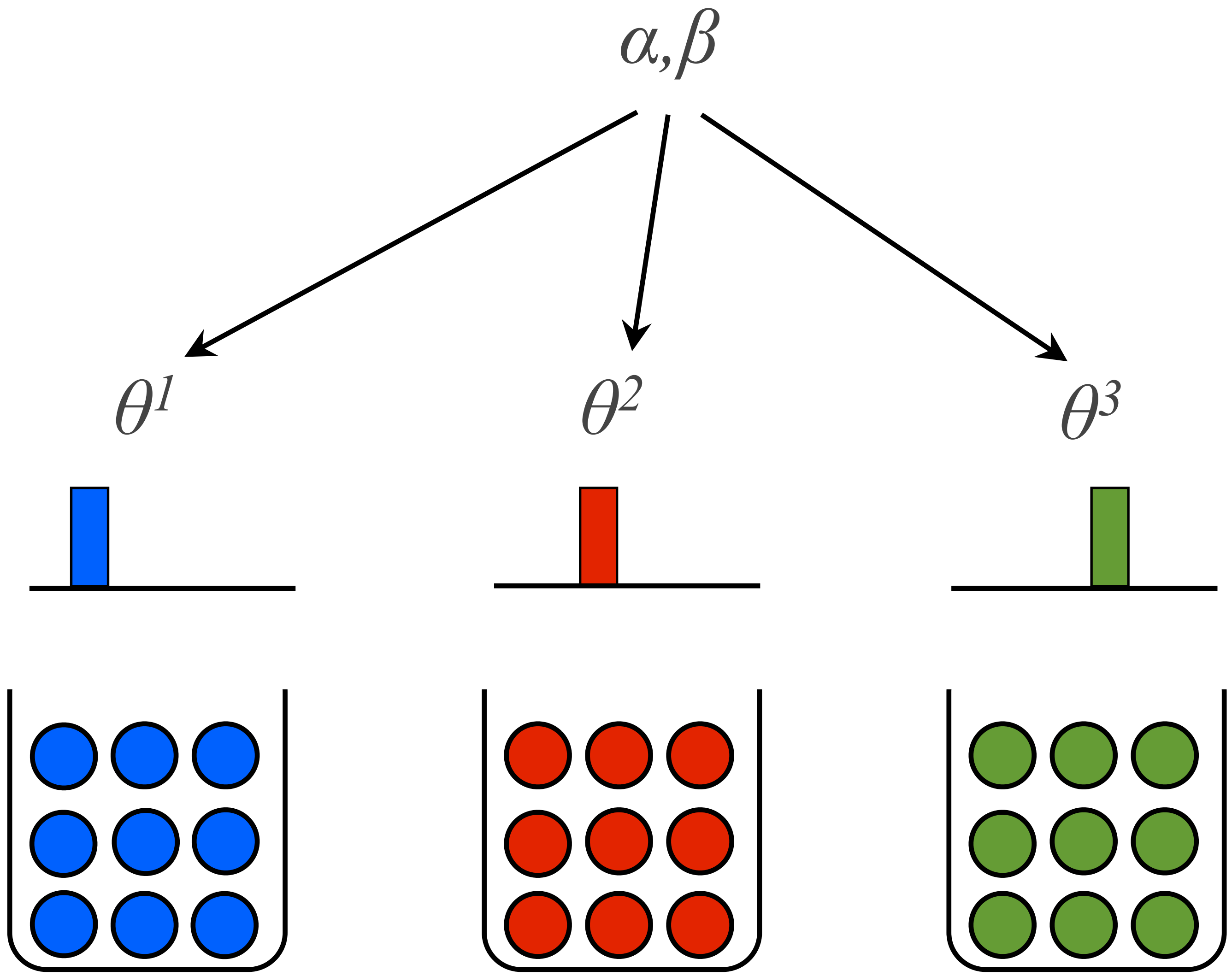
Level 2: Bags in general

↓

Level 1: Bag proportions

↓

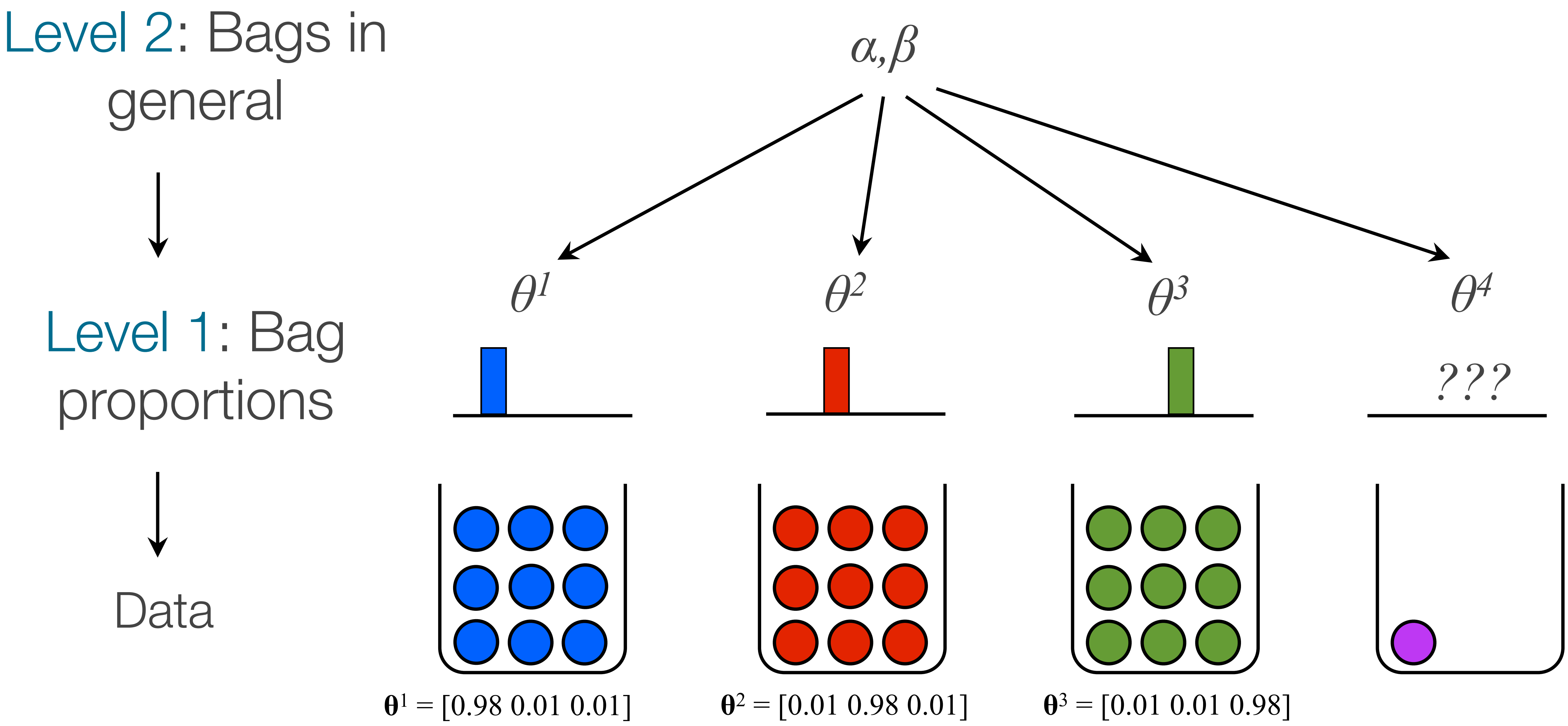
Data



$\boldsymbol{\theta}^1$ = [0.98 0.01 0.01] $\boldsymbol{\theta}^2$ = [0.01 0.98 0.01] $\boldsymbol{\theta}^3$ = [0.01 0.01 0.98]

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

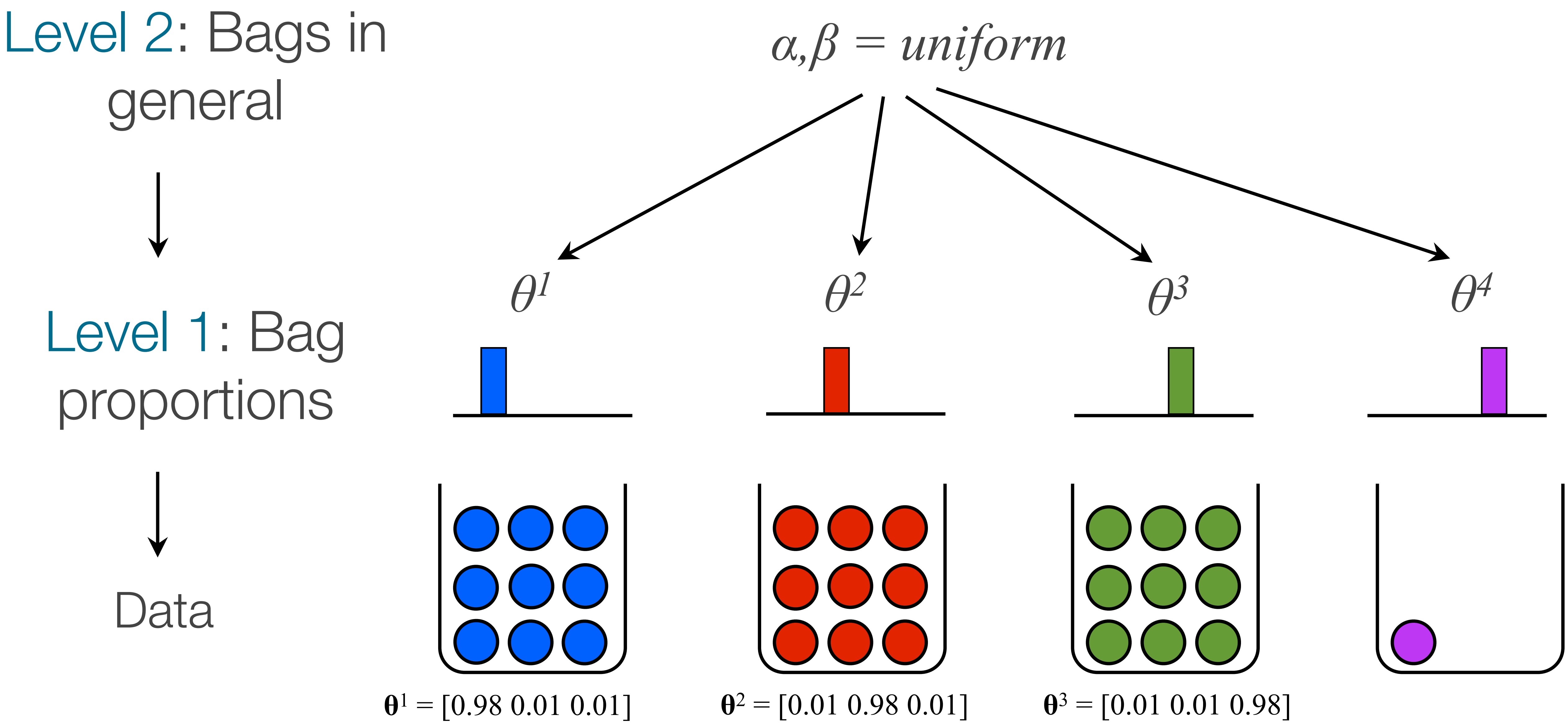
Level 1: Bag proportions

Data

α,β
θ^1
θ^2
θ^3
θ^4
???
$\boldsymbol{\theta}^1$ = [0.98 0.01 0.01]
$\boldsymbol{\theta}^2$ = [0.01 0.98 0.01]
$\boldsymbol{\theta}^3$ = [0.01 0.01 0.98]

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

$\alpha,\beta = uniform$

Level 1: Bag proportions

θ^1 θ^2 θ^3 θ^4

Data

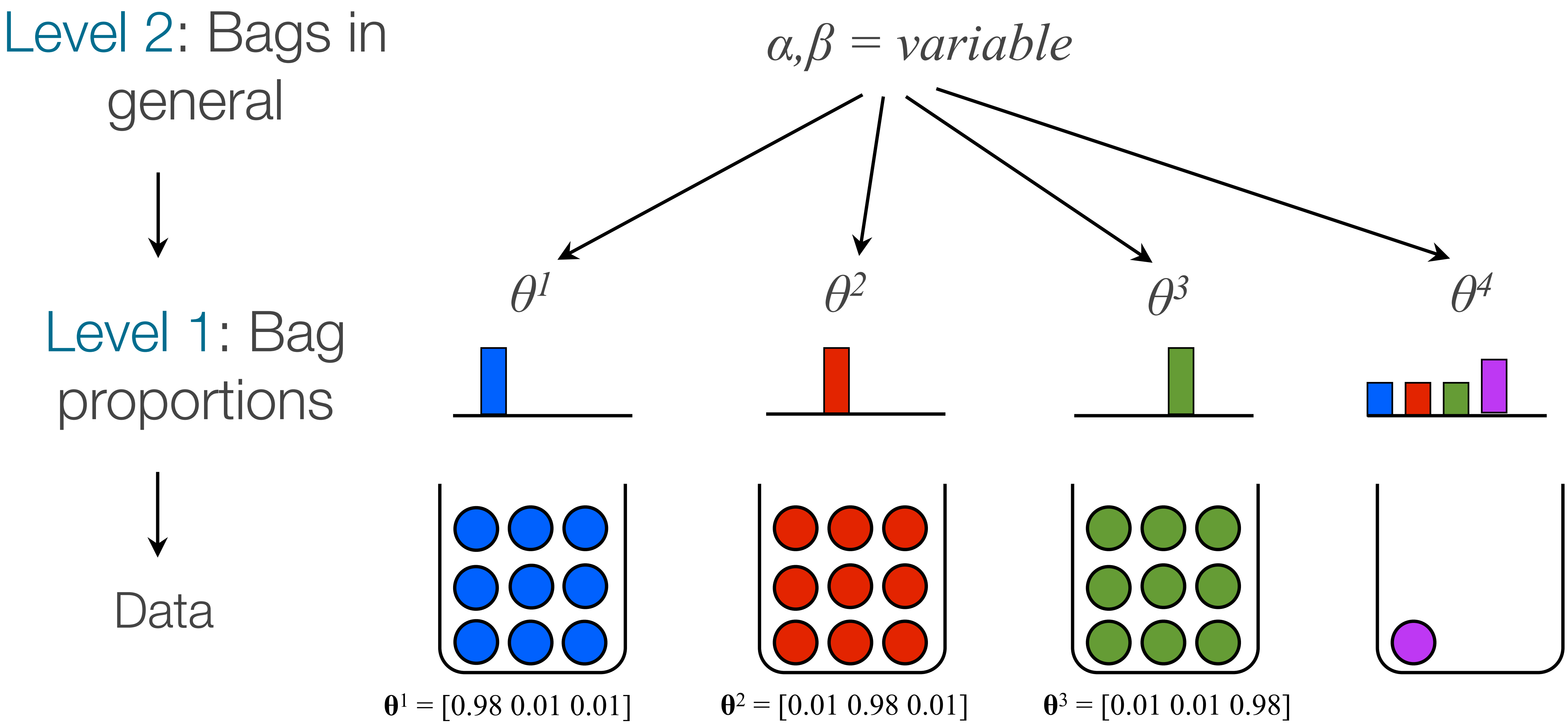
$\boldsymbol{\theta}^1 = [0.98\ 0.01\ 0.01]$

$\boldsymbol{\theta}^2 = [0.01\ 0.98\ 0.01]$

$\boldsymbol{\theta}^3 = [0.01\ 0.01\ 0.98]$

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

$\alpha,\beta = variable$

θ^1 θ^2 θ^3 θ^4

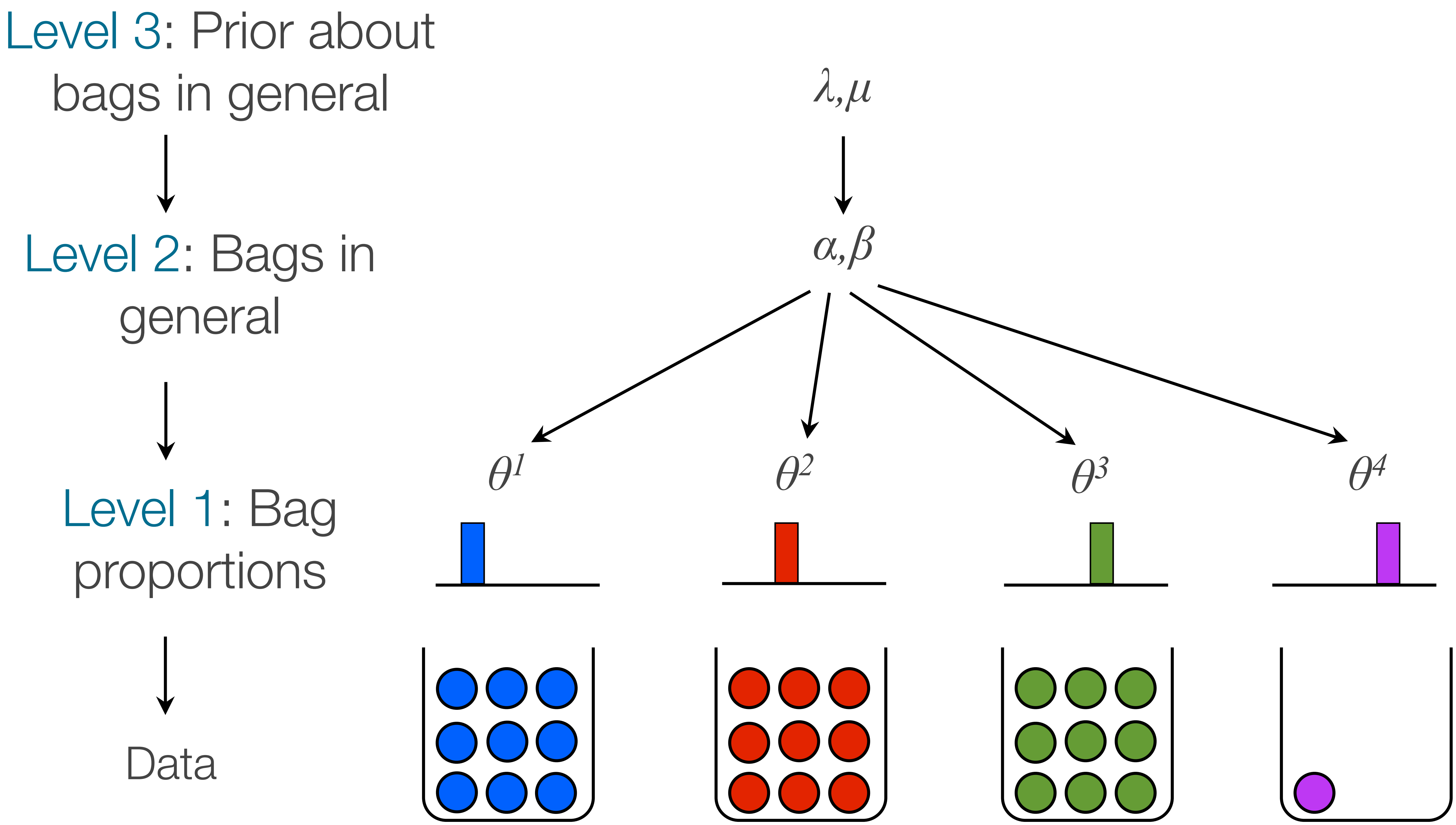
$\boldsymbol{\theta}^1$ = [0.98 0.01 0.01]

$\boldsymbol{\theta}^2$ = [0.01 0.98 0.01]

$\boldsymbol{\theta}^3$ = [0.01 0.01 0.98]

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

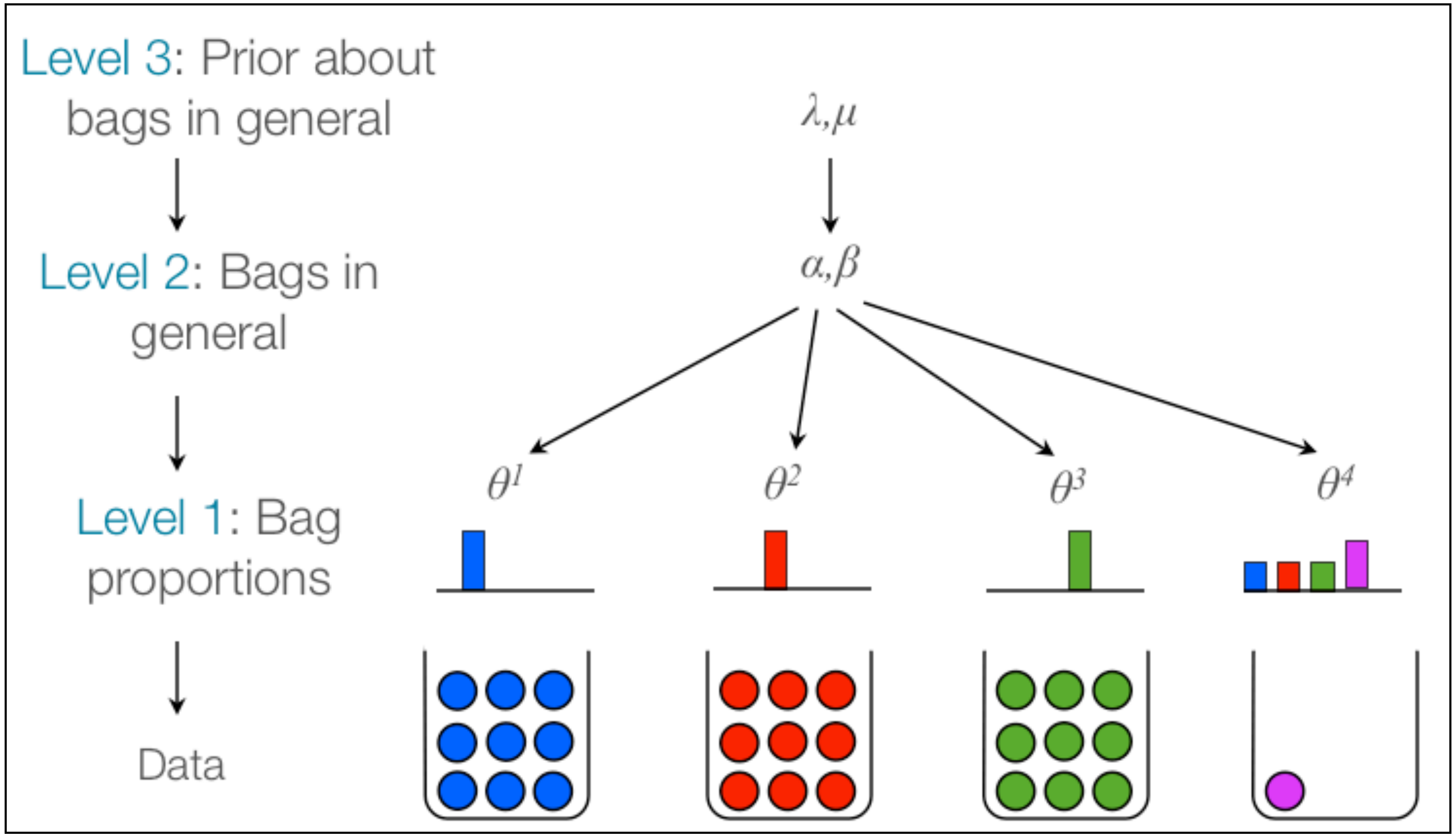
Data

The full model

α is a scalar → $\alpha \sim \mathrm{Exponential}(\lambda)$

The Dirichlet is conjugate to the Dirichlet → $\beta \sim \mathrm{Dirichlet}(\mu)$

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**

Charles Kemp, Amy Perfors and Joshua B. Tenenbaum (2007). Learning overhypotheses with hierarchical Bayesian models DOI: 10.1111/j.1467-7687.2007.00585.x

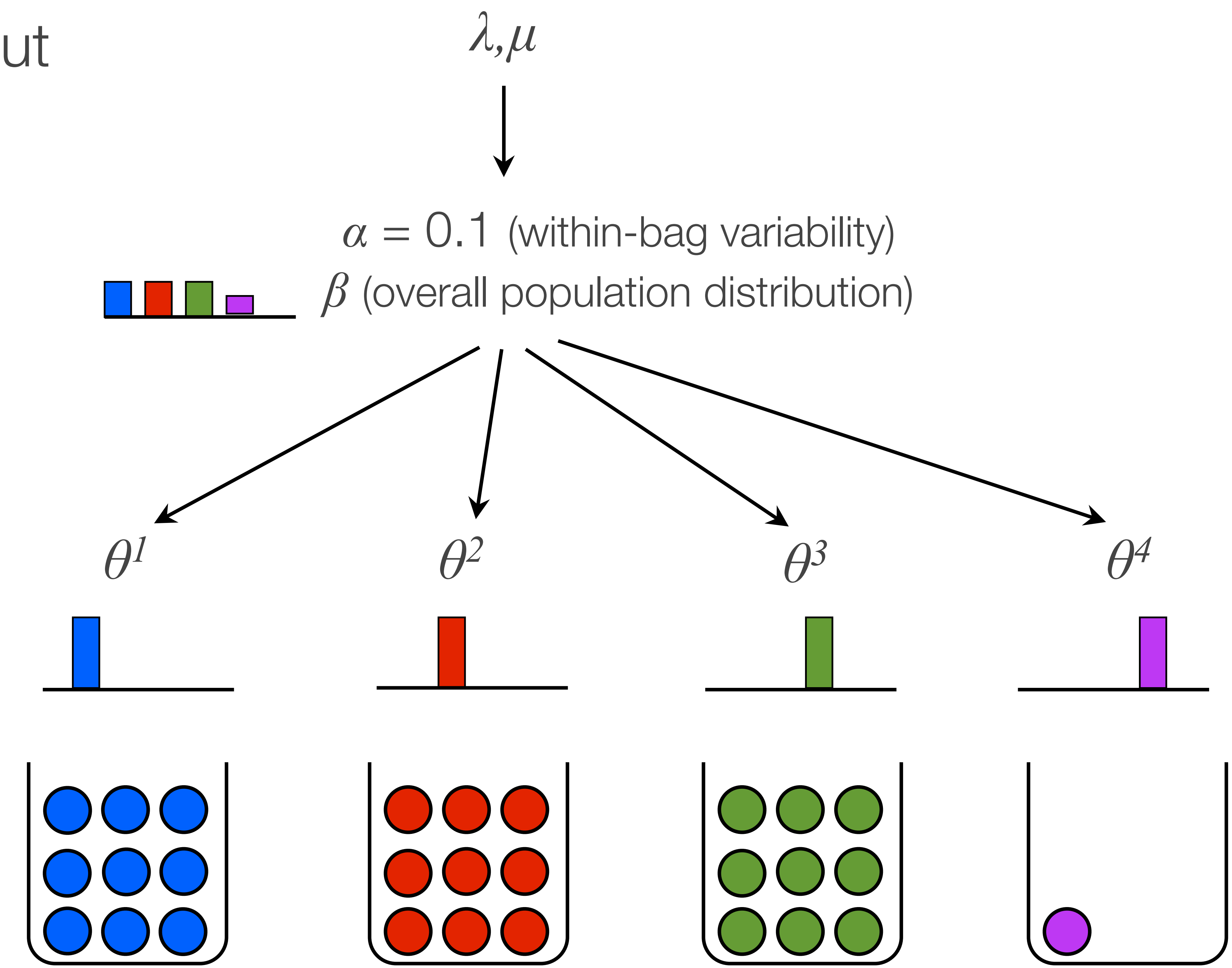
This model can learn which features “matter”

Level 3: Prior about bags in general

Level 2: Bags in general

Level 1: Bag proportions

Data



λ, μ

α = 0.1 (within-bag variability)

β (overall population distribution)

θ^1 θ^2 θ^3 θ^4

This model can learn which features “matter”

Level 3: Prior about bags in general

↓

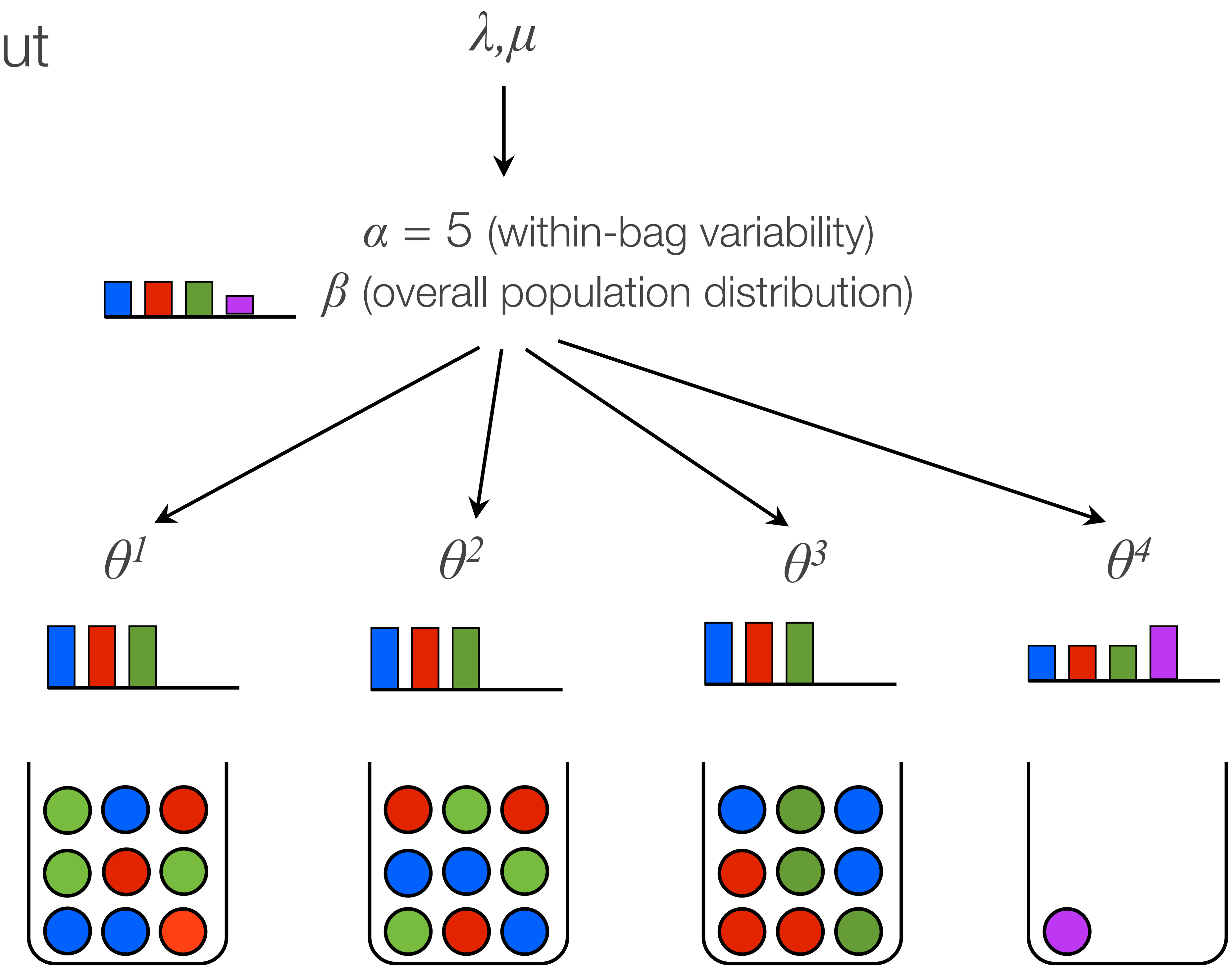
Level 2: Bags in general

↓

Level 1: Bag proportions

↓

Data



λ, μ

$\alpha = 5$ (within-bag variability)

β (overall population distribution)

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.

Learning multiple kinds

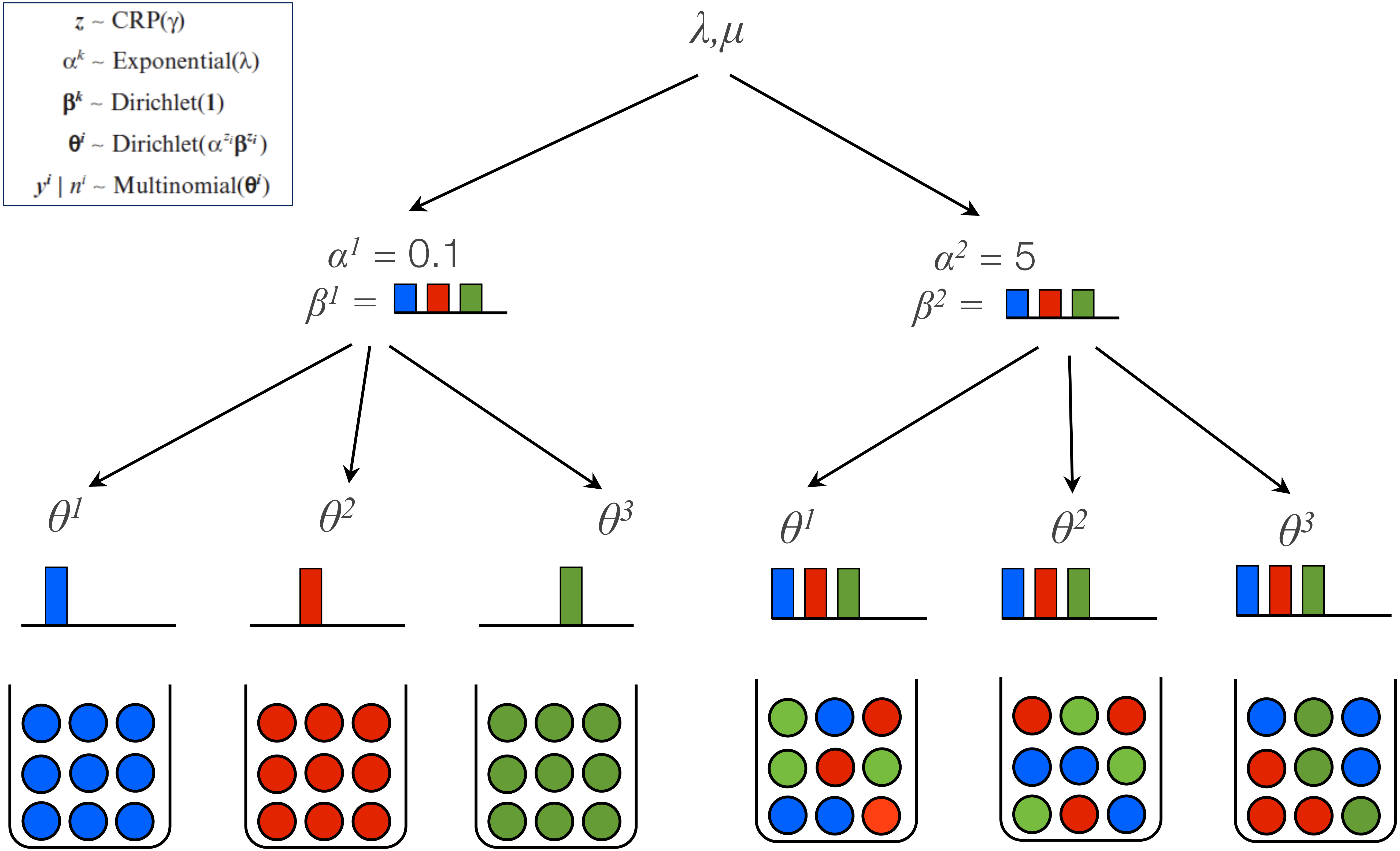
$z \sim \text{CRP}(\gamma)$

$\alpha^k \sim \text{Exponential}(\lambda)$

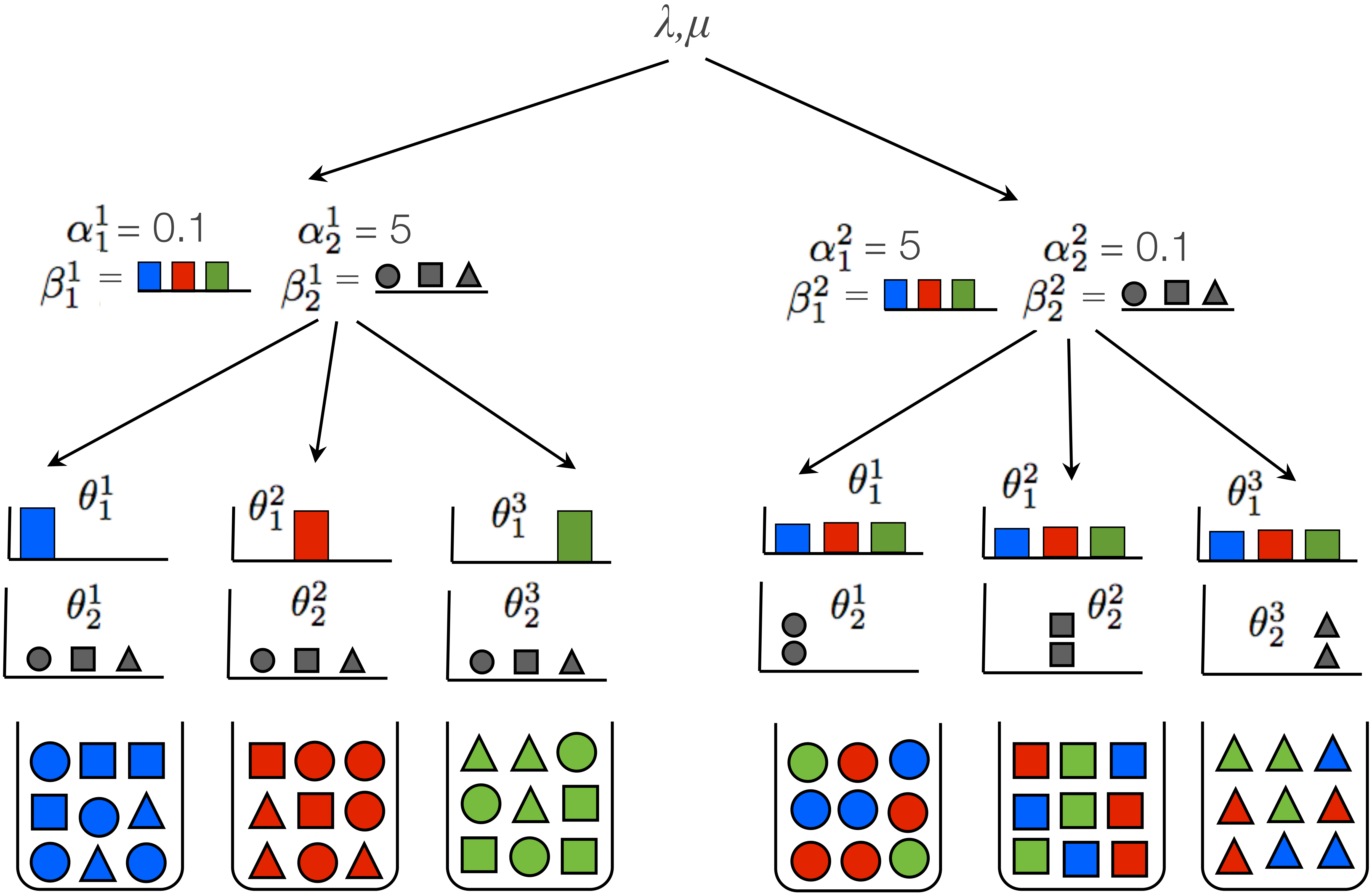
$\boldsymbol{\beta}^k \sim \text{Dirichlet}(\mathbf{1})$

$\boldsymbol{\theta}^i \sim \text{Dirichlet}(\alpha^{z_i}\boldsymbol{\beta}^{z_i})$

$y^i \mid n^i \sim \text{Multinomial}(\boldsymbol{\theta}^i)$



Extendible to having multiple features



Hierarchical Bayesian Models, more generally

Ability to capture inferences at multiple levels is really powerful, also outside cognitive modelling:

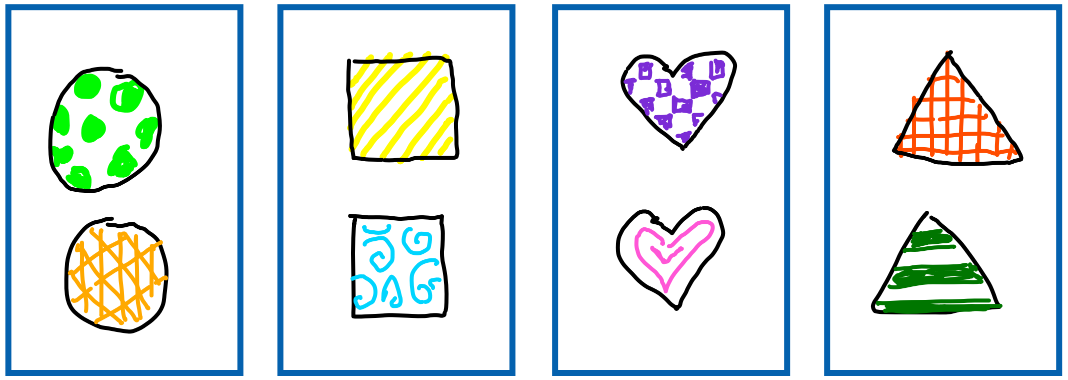
- Domain adaptation
- Hierarchical Bayesian Language Model (Teh, 2006): n-gram order (trigrams vs bigram) corresponds to level in hierarchy
- Or multilevel/hierarchical modelling of group vs. individuals, Chapters 5 and 9 in F&L

Experiments

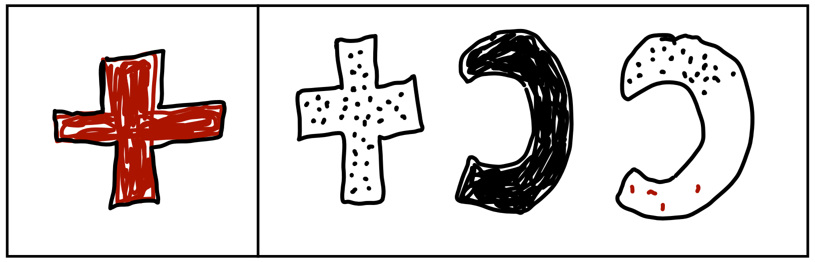
Captures the acquisition of the shape bias

Classifies by shape based on four training categories

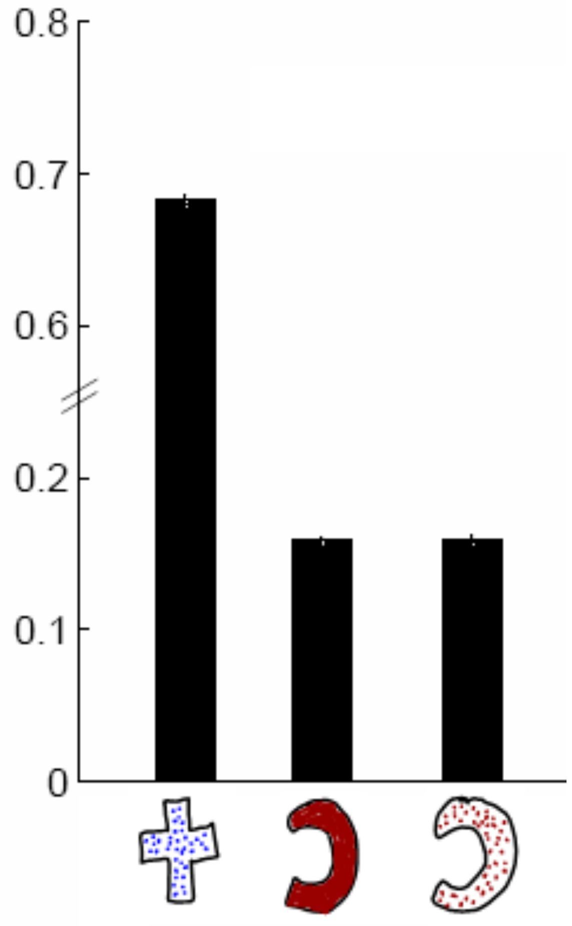
Training



Testing



Probability that object belongs to the same category as ✚



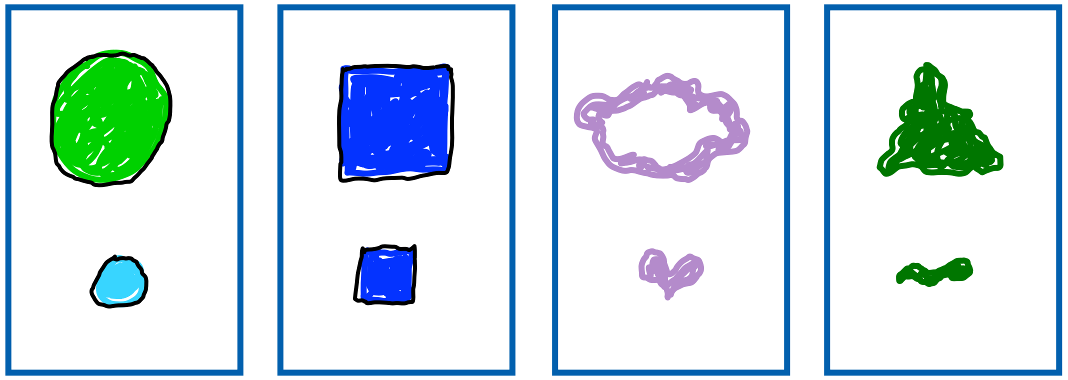
Based on Smith (2002)

Captures the acquisition of the shape bias

Can also learn multiple different kinds

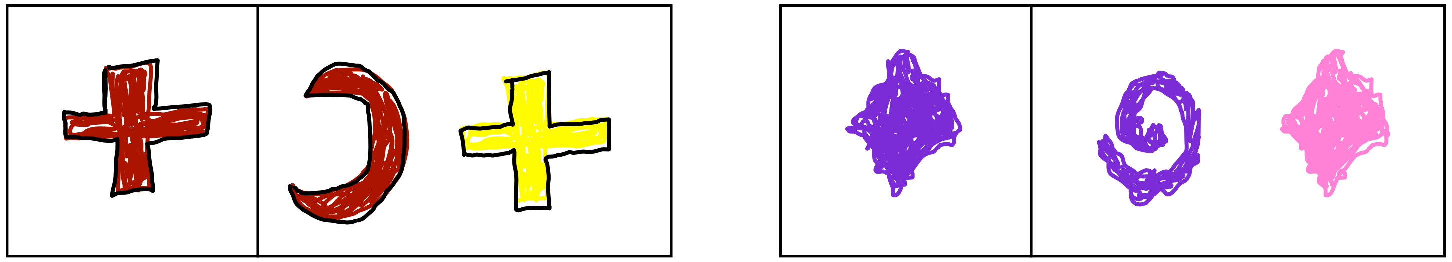
Training

Solids Nonsolids

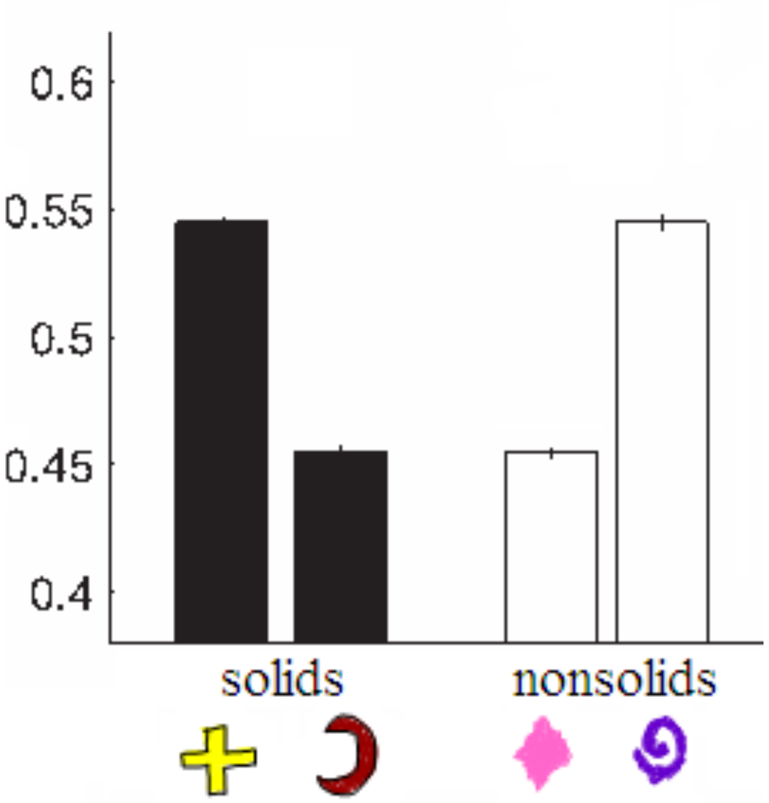


Testing

Solids Nonsolids



Probability that object belongs to the same category as the test



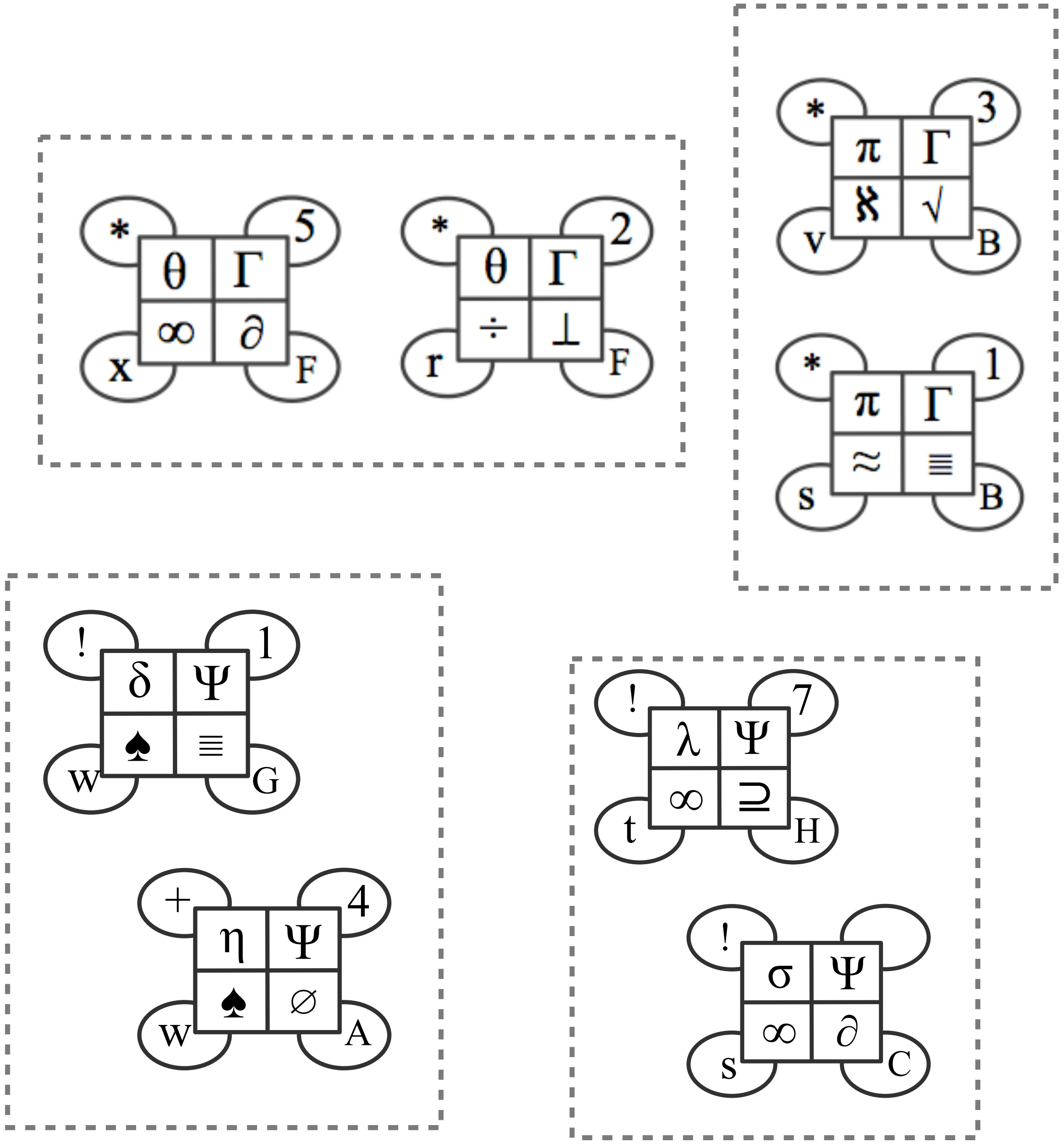
Based on Jones & Smith (2002)

In other cases, in more complicated experiments
(larger feature set, unnatural features, larger memory load)
humans fail to learn, while the model does well.

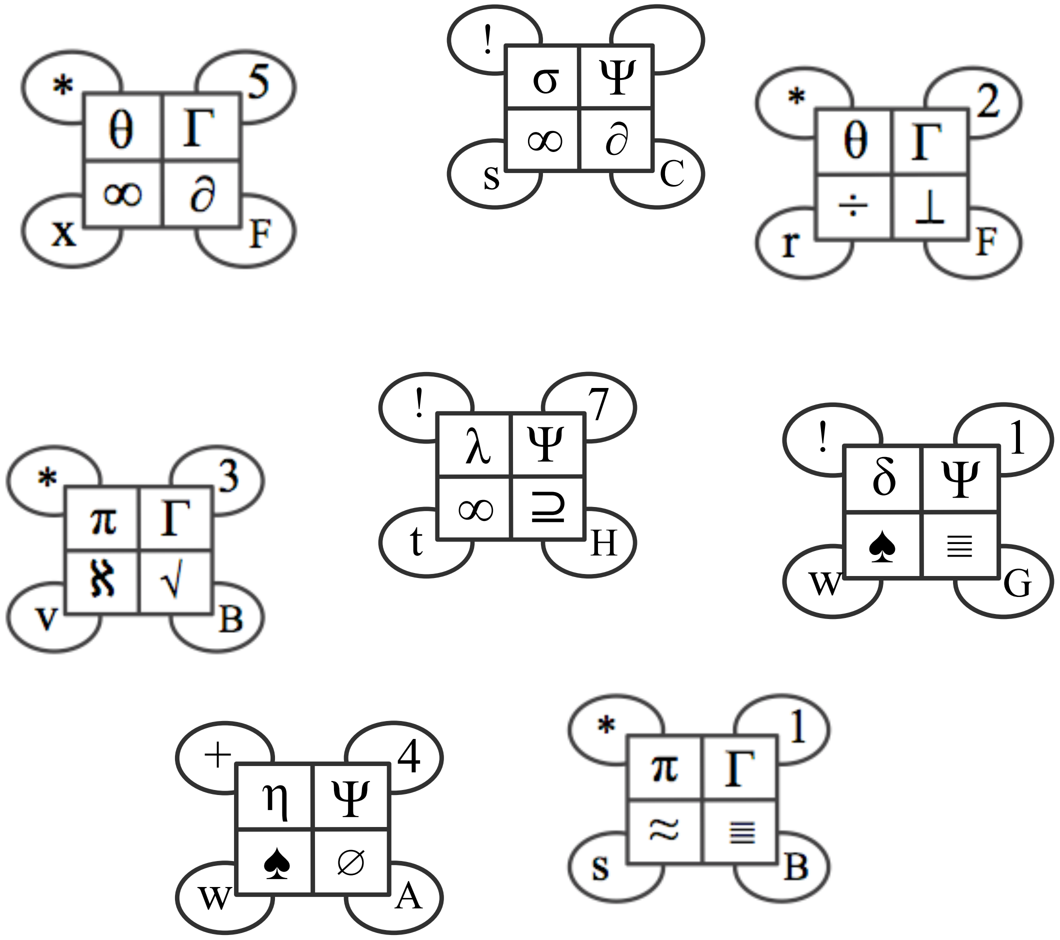
Test: Can people learn arbitrary overhypotheses?

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

Supervised



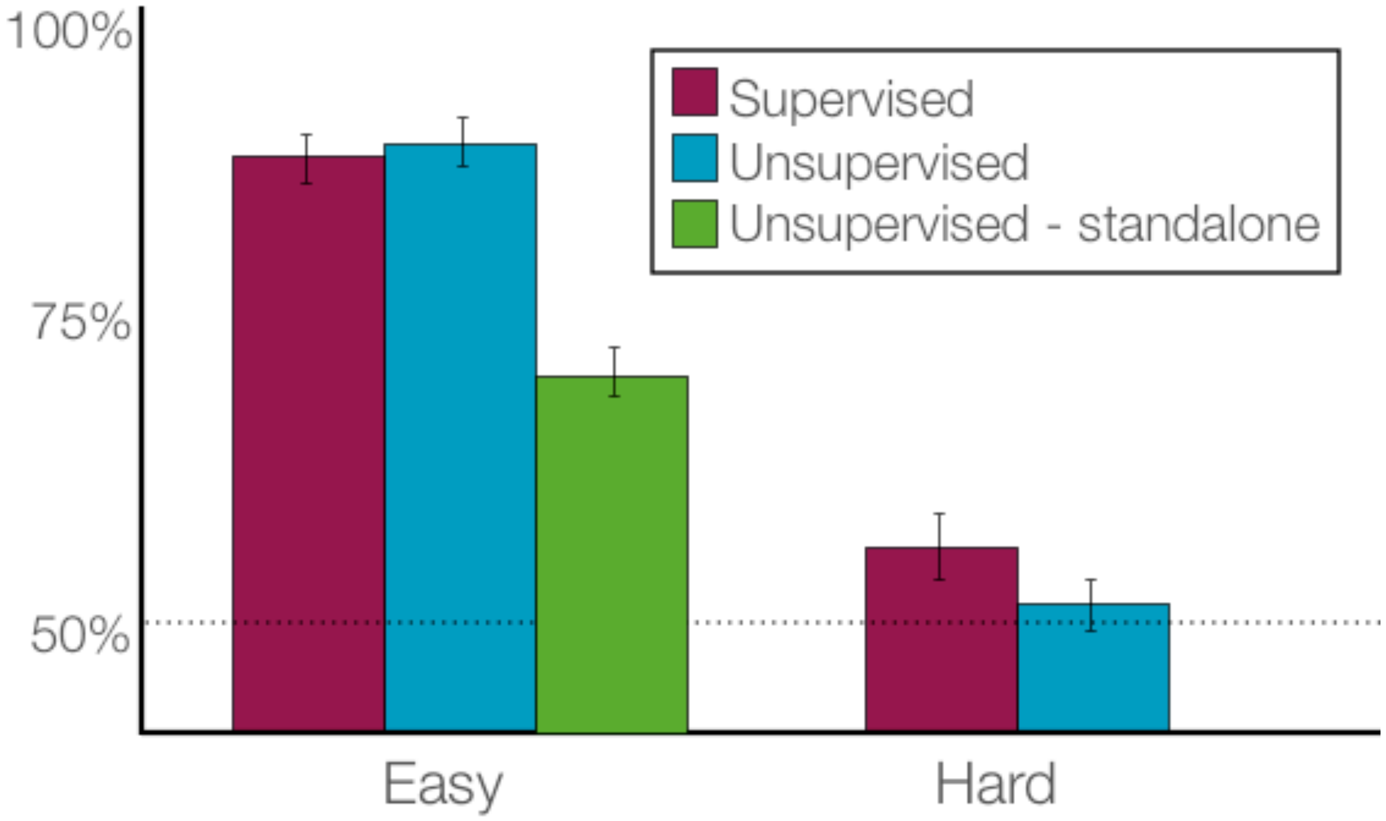
Unsupervised



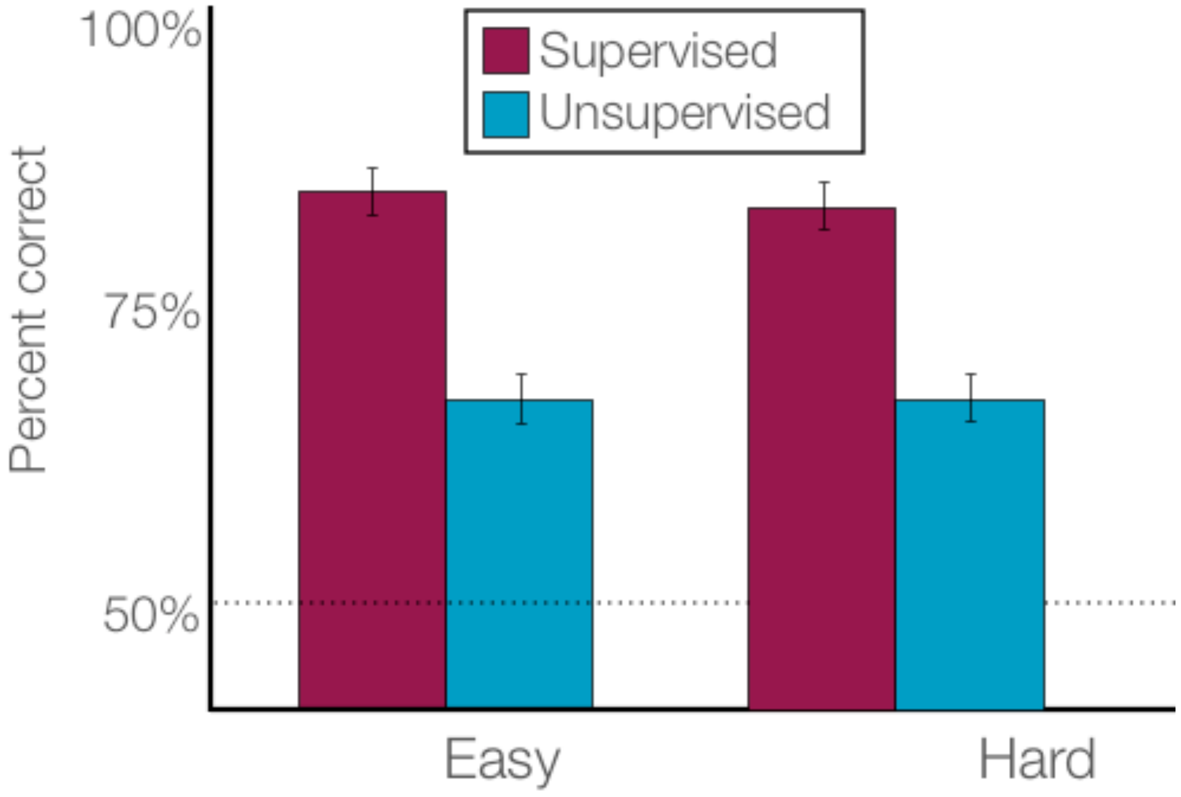
Model performance

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

Humans



Model



In more complicated experiments
(larger feature set, unnatural features, larger memory load)
humans fail to learn, while the model does well.

- Cognitive modelling has to take these mismatches seriously, otherwise we're just doing engineering.

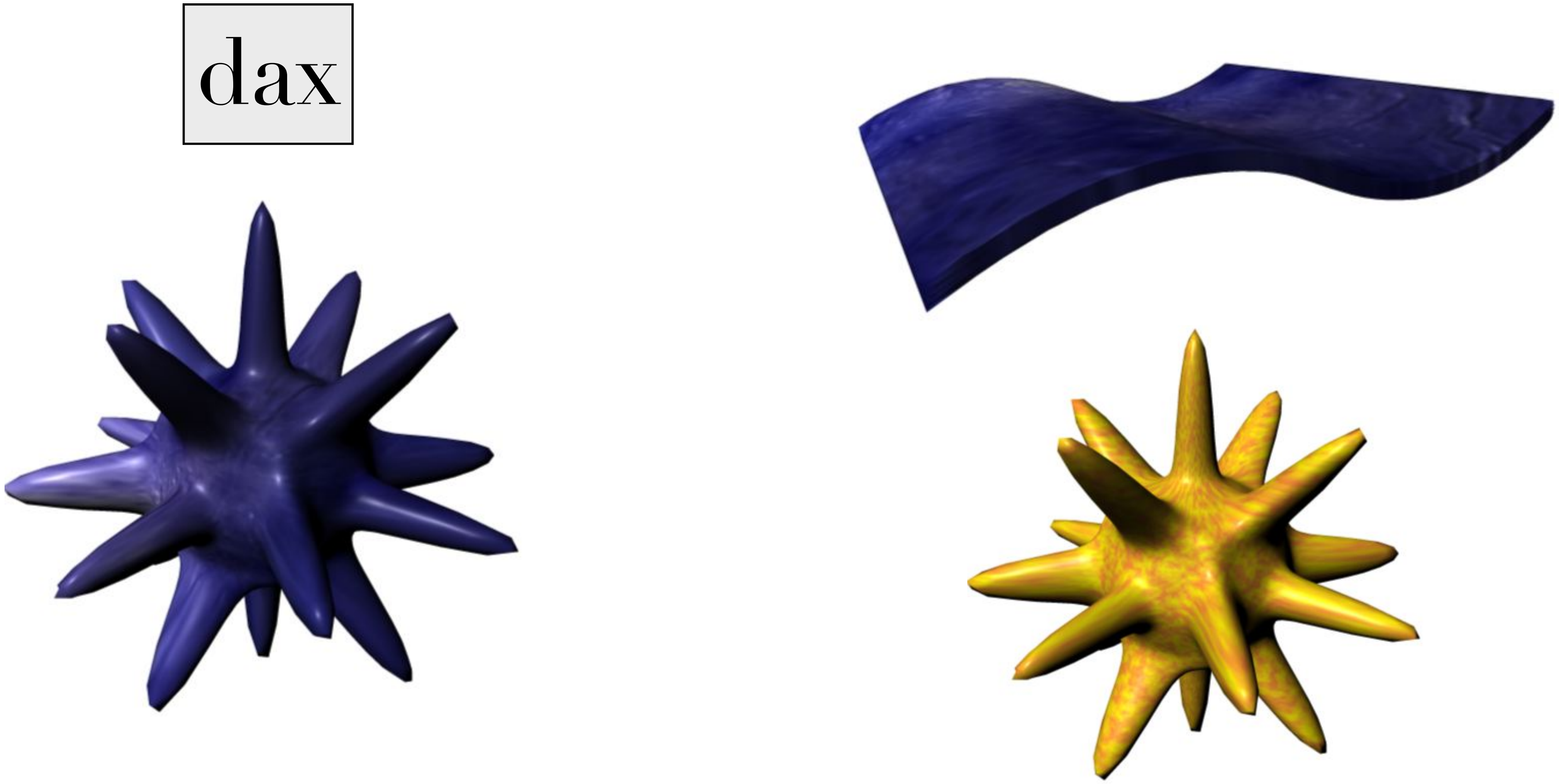
- Bayesian "ideal observer" models can demonstrate learnability in theory, on Marr's computational level;

- But have trouble integrating process-level constraints when humans are less than or different from ideal.

- Failure of model to match human behaviour is still informative!

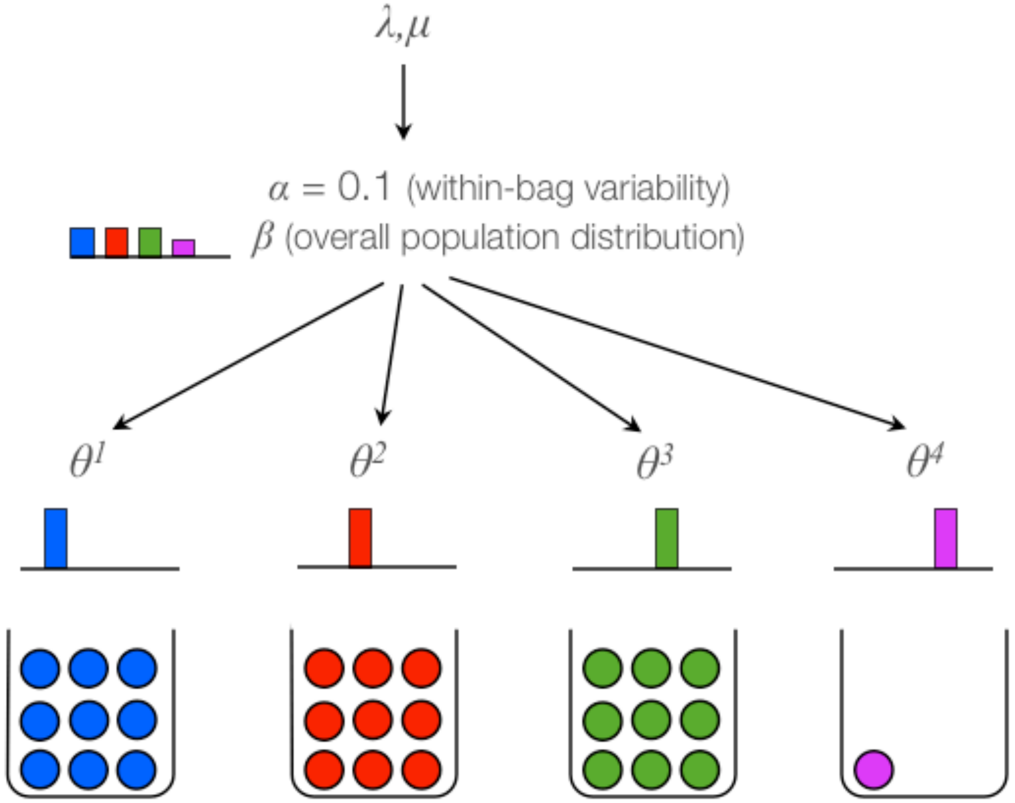
Summary

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



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


λ, μ
α = 0.1 (within-bag variability)
β (overall population distribution)
θ^1
θ^2
θ^3
θ^4

Next time

Perfors, Tenenbaum and Regier (2011). The Learnability of Abstract Syntactic Structures. Cognition

- Q: How can infants use (linear) word sequences to discover that language has a hierarchical structure?

- A: Use overhypotheses!