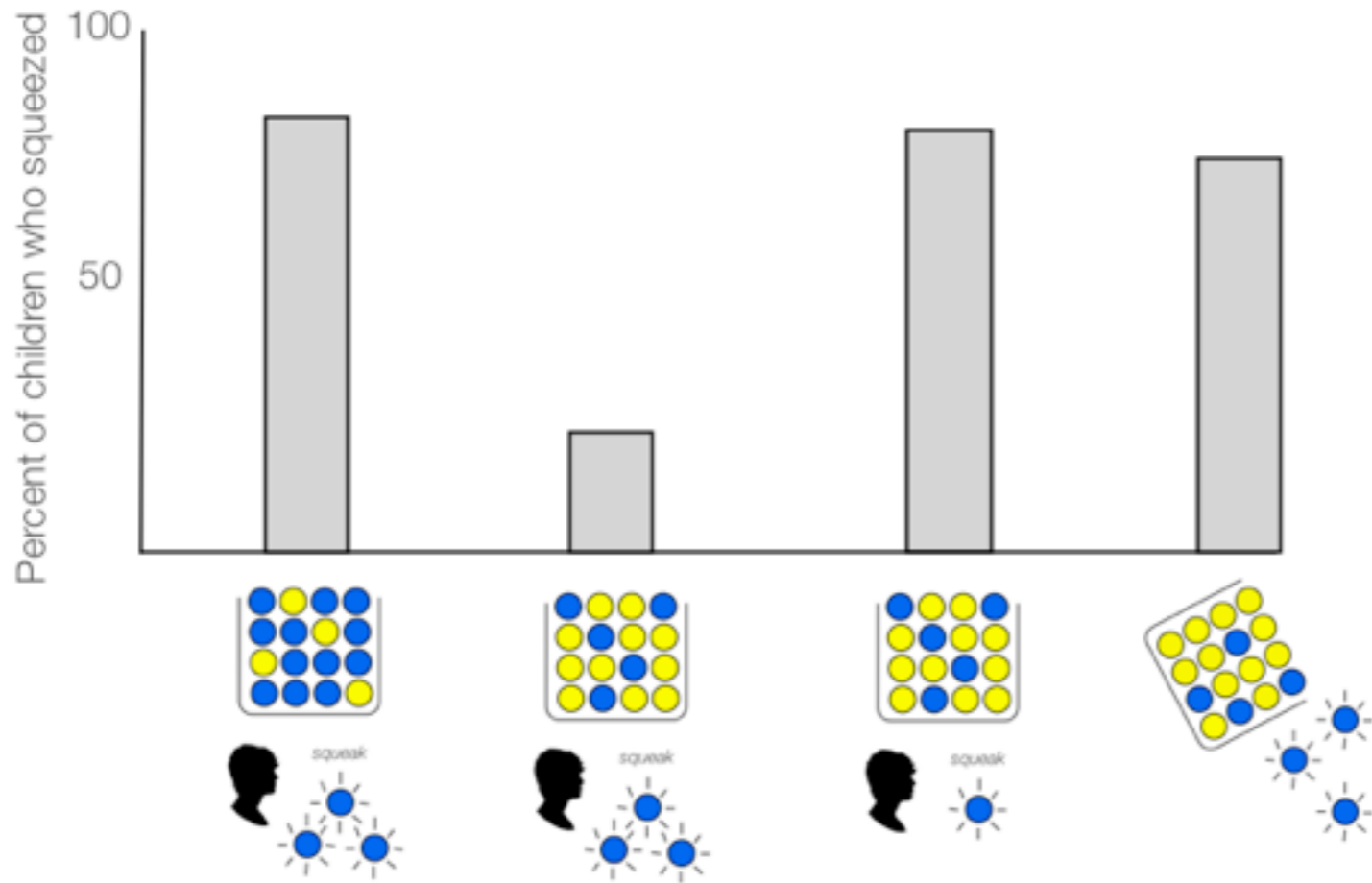
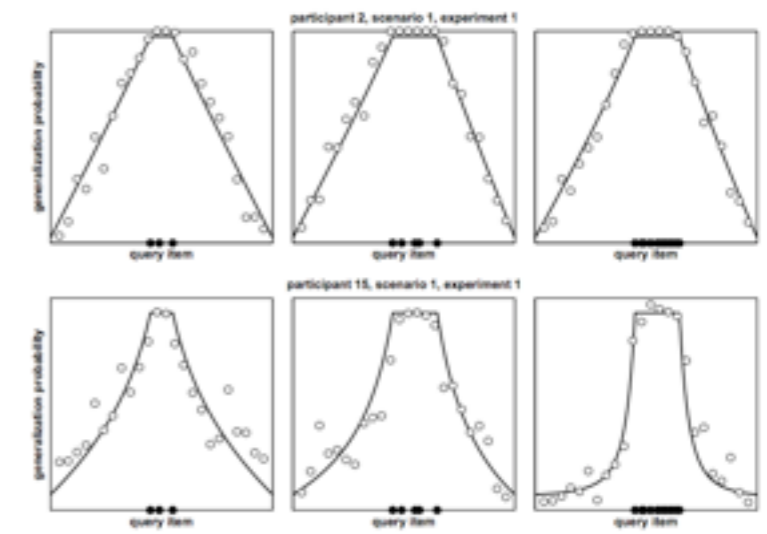
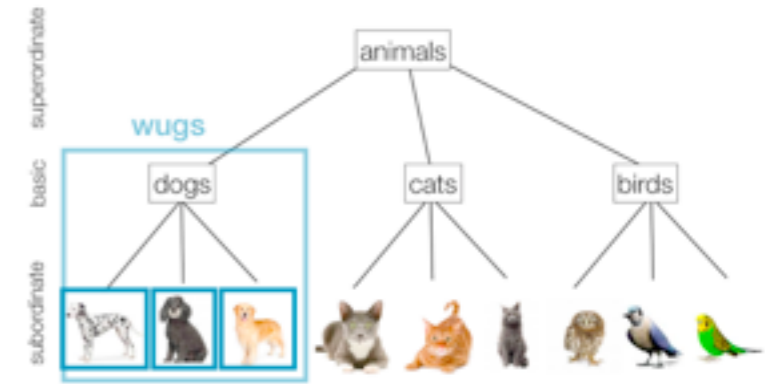


# Computational Cognitive Science



Lecture 20: Strong vs weak sampling



# Where we are

---

- ▶ So far in CCS we've learned about how people (and models) might learn and represent information about sequences of events or words...
- ▶ As well as how they can learn concepts that don't change, or incorporate an element of time
- ▶ Hidden within all of this have been certain implicit assumptions about where and how this data is generated, and what kind of information people get
- ▶ These assumptions have driven the models so far but in the next few lectures we'll revisit them

# Plan

---

- ▶ How is the data generated?
  - Strong vs. weak sampling: the idea
  - People's sensitivity to sampling assumptions
  - Individual differences in sampling sensitivity

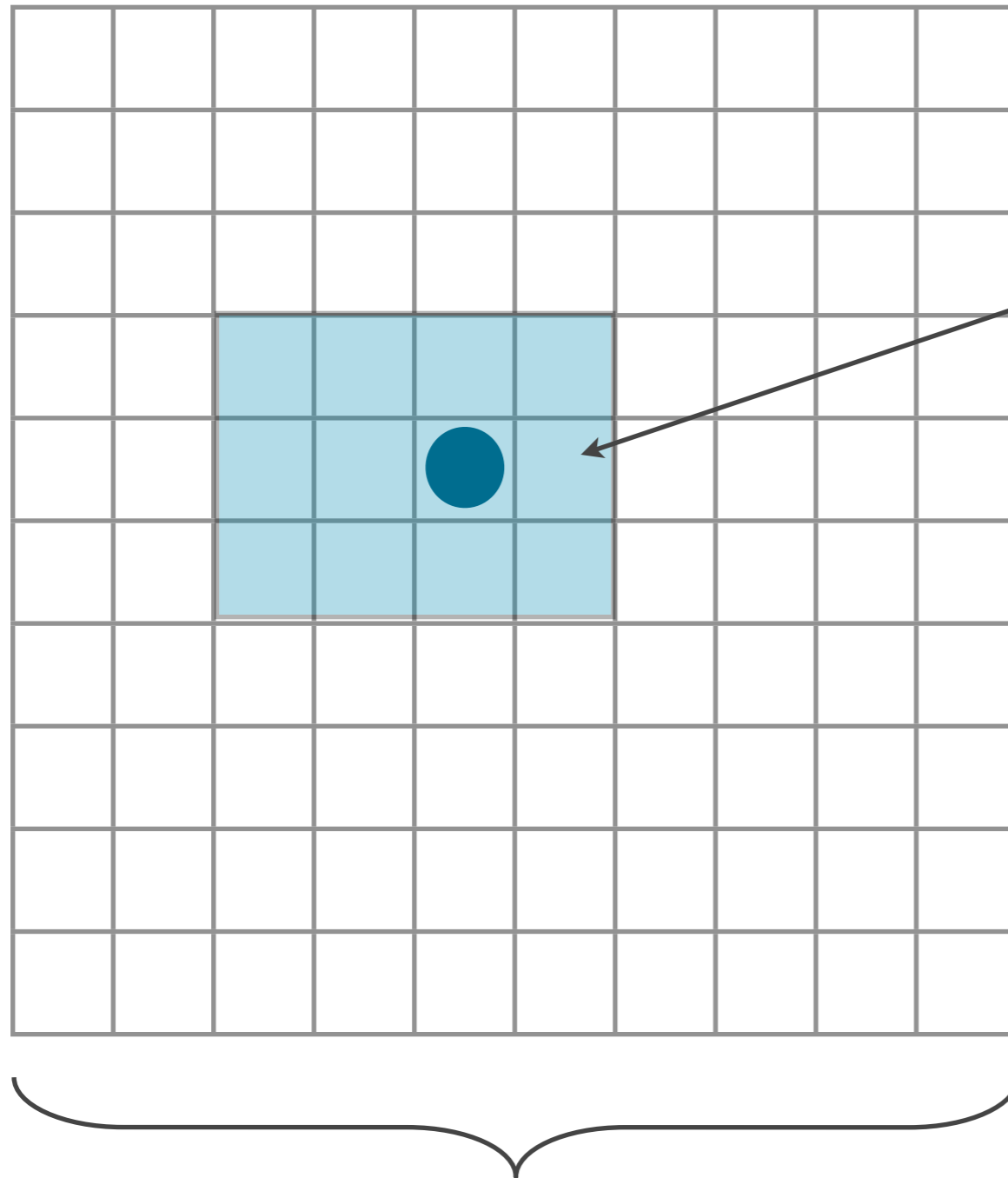
# Plan

---

- ➔ How is the data generated?
  - ➔ Strong vs. weak sampling: the idea
    - People's sensitivity to sampling assumptions
    - Individual differences in sampling sensitivity

# Remember Bayes' Rule...

---



Hypothesis of size  $n$

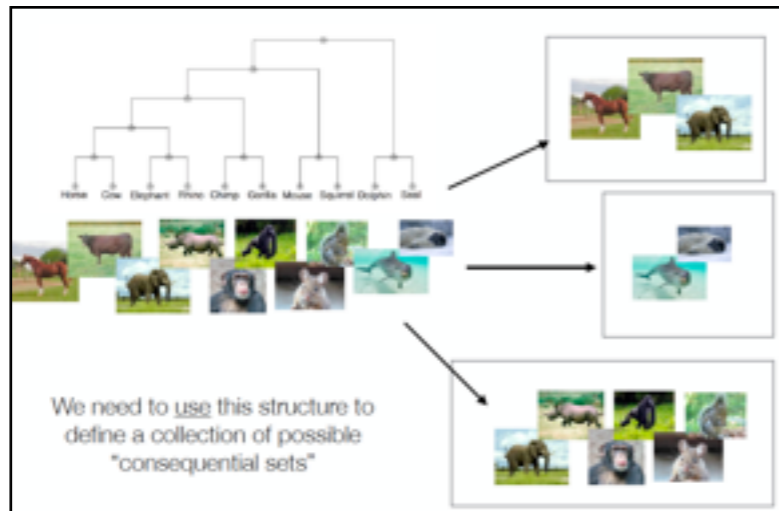
$$p(d|h) = 1/n \\ = 1/12$$

This is known as the  
**size principle**

World

# The size principle has cropped up in many places...

- ▶ Explicitly so when talking about the lotto problem and the problem of generalisation...



The lotto problem  
("this is computer science and not just maths, right?")



## Extending the problem of generalisation

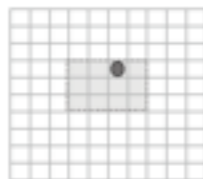
- ▶ Arbitrary representational structure, not just a metric space: Bayesian inference over hypotheses

Hypothesis space  $\mathcal{H}$  is the set of possible consequential regions

$p(h)$  is the prior probability of each hypothesis in the set

$p(x|h)$  is the probability of that hypothesis given data  $x$

by Bayes' Rule, we can calculate the posterior probability as  
 $p(h|x) \propto p(x|h)p(h)$



$$p(h) = 1/|\mathcal{H}|$$

$$p(x|h) = 1/12$$

Use the likelihood to enforce data fit

Prior  
 $P(h) \propto \frac{1}{|\mathcal{H}|}$

Likelihood  
$$P(x|h) \propto \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases}$$

Our "usual" likelihood: every object within the consequential set is equally likely to be "observed" to have the property

## The likelihood

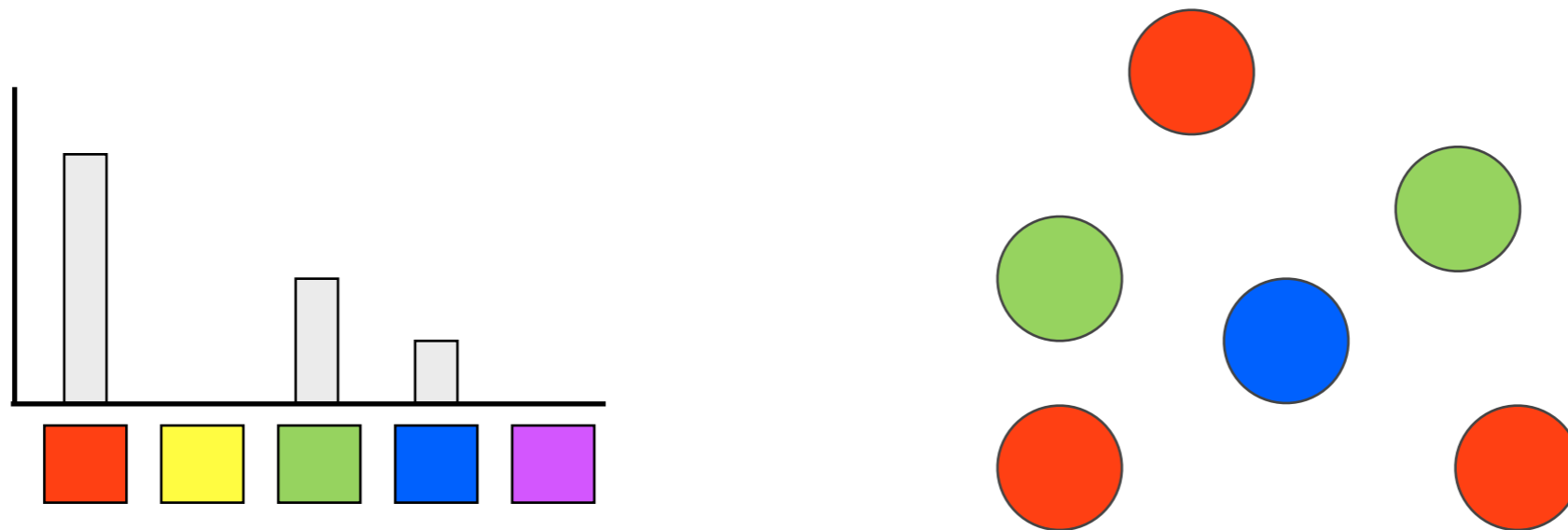
- Each winning number  $x$  is selected uniformly at random from the range  $(l, u)$
- Notation:
  - Let  $|h| = u - l + 1$  be the size of  $h$
  - and  $x \in h$  means  $l \leq x \leq u$
- Likelihood for a single observation:

$$P(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases}$$

# The size principle has cropped up in many places...

---

- ▶ But also much everywhere we have assumed that the data is drawn proportional to the probability distribution that generates it (called **strong sampling**)



- ▶ ... and thus the probability of the data is given by the proportions under that distribution:

$$p(\bullet|h) = 50\%$$


# The size principle has cropped up in many places...

- ▶ But also much everywhere we have assumed that the data is drawn proportional to the probability distribution that generates it

## RMC

$$P(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{k/2} \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

(Multivariate normal)



- $\mathbf{z}|\alpha \sim \text{CRP}(\alpha)$
- $\mu_k \sim \text{Uniform}$
- $\Sigma_k \sim \text{Uniform}$
- $\theta_k|\beta \sim \text{Beta}(\beta, \beta)$
- $\mathbf{x}_i|\mu_k, \Sigma_k, z_i = k \sim \text{Normal}(\mu_k, \Sigma_k)$
- $\ell_i|\theta_k, z_i = k \sim \text{Bernoulli}(\theta_k)$

## Overhypothesis models

### 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  $\mathbf{y}$  with a multinomial distribution. In essence, each multinomial  $\theta$  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

## Iterated learning (most)

### Two steps



- ▶ **Learning step:** learner  $n+1$  sees  $x_n$  (from previous person) and computes a posterior distribution over  $\theta_{n+1}$  according to Bayes' Rule
 
$$P(\theta_{n+1}|\mathbf{x}_n, \mathbf{y}_n) = \frac{P(\mathbf{y}_n|\theta_{n+1})P(\theta_{n+1})}{\sum_{\theta} P(\mathbf{y}_n|\theta)P(\theta)}$$
- ▶ **Production step:** Events are generated independently from  $Q(x)$ . Learner  $n+1$  produces utterances  $y_{n+1}$  according to
 
$$P(\mathbf{y}_{n+1}|\theta_{n+1}, \mathbf{x}_n)$$

▶ Since all learners use the same learning and production steps, we can calculate:

$$P(\theta_{n+1}|\mathbf{x}_n) = \sum_{\theta} \sum_{\mathbf{y}} P(\theta_{n+1}|\mathbf{y}, \mathbf{x}_n)P(\mathbf{y}|\theta)Q(\mathbf{x})$$

## N-gram models

### Predicting the next word: $P(w_n|w_1, \dots, w_{n-1})$

The MLE probability of a word given a previous word or series of words is given by:

$$p(w_n|w_1, \dots, w_{n-1}) = \frac{C(w_1, \dots, w_n)}{C(w_1, \dots, w_{n-1})}$$

Count  $C$  of times there are  $n$  words in a row

Probability of  $w_n$  given previous  $n-1$  words

Count  $C$  of times of previous  $n-1$  words in a row are observed

## Mixture of Gaussians

### Mixture of Gaussians with EM

The **Expectation step** (or E-step) is a direct analogue of the assignment step previously: each datapoint is assigned probabilistically to each cluster


Responsibilities are:

$$r_k^{(n)} = \frac{w_k \frac{1}{\prod_{i=1}^k \sqrt{2\pi\sigma_i^2}} \exp\left(-\sum_{i=1}^k (w_i^{(n)} - \mu_i^{(k)})^2 / 2\sigma_i^{(k)^2}\right)}{\sum_{l=1}^K w_l \frac{1}{\prod_{i=1}^k \sqrt{2\pi\sigma_i^2}} \exp\left(-\sum_{i=1}^k (w_i^{(n)} - \mu_i^{(l)})^2 / 2\sigma_i^{(l)^2}\right)}$$

Equation for a Gaussian - you've seen this in Dan's recent lectures! (so this is exactly the same as calculating the likelihood of that point under the Gaussian distribution with parameters  $\mu, \sigma$ , and  $\pi$ )

### Fitting the data to a structure: Formalisation

Assume that features are independently generated from a Gaussian distribution over the graph



$W$  is a weight matrix, where  $w_{ij} = 1/\epsilon_{ij}$  if nodes  $i$  and  $j$  are joined by an edge of length  $\epsilon_{ij}$  and  $w_{ij} = 0$  otherwise

$$P(f|W) \propto \exp\left(-\frac{1}{4} \sum_{i,j} w_{ij} (f_i - f_j)^2\right)$$

This penalises a feature vector if  $f_i \neq f_j$  and  $i$  and  $j$  are adjacent in the graph. The penalty increases if the edge between them is shorter.

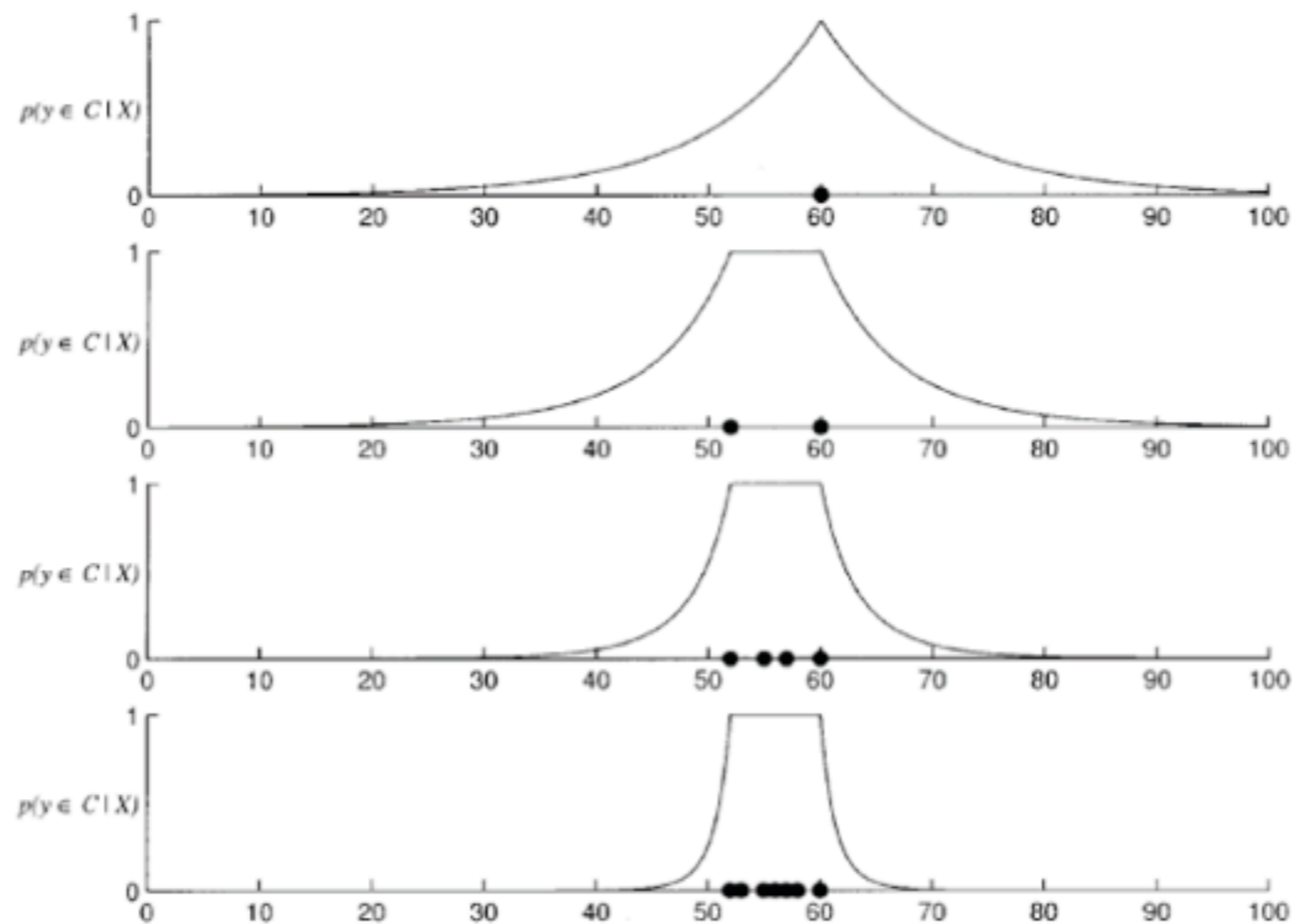
Need to also make assumptions about the variance of the Gaussian for the prior to be proper.



# Consequence of the size principle

---

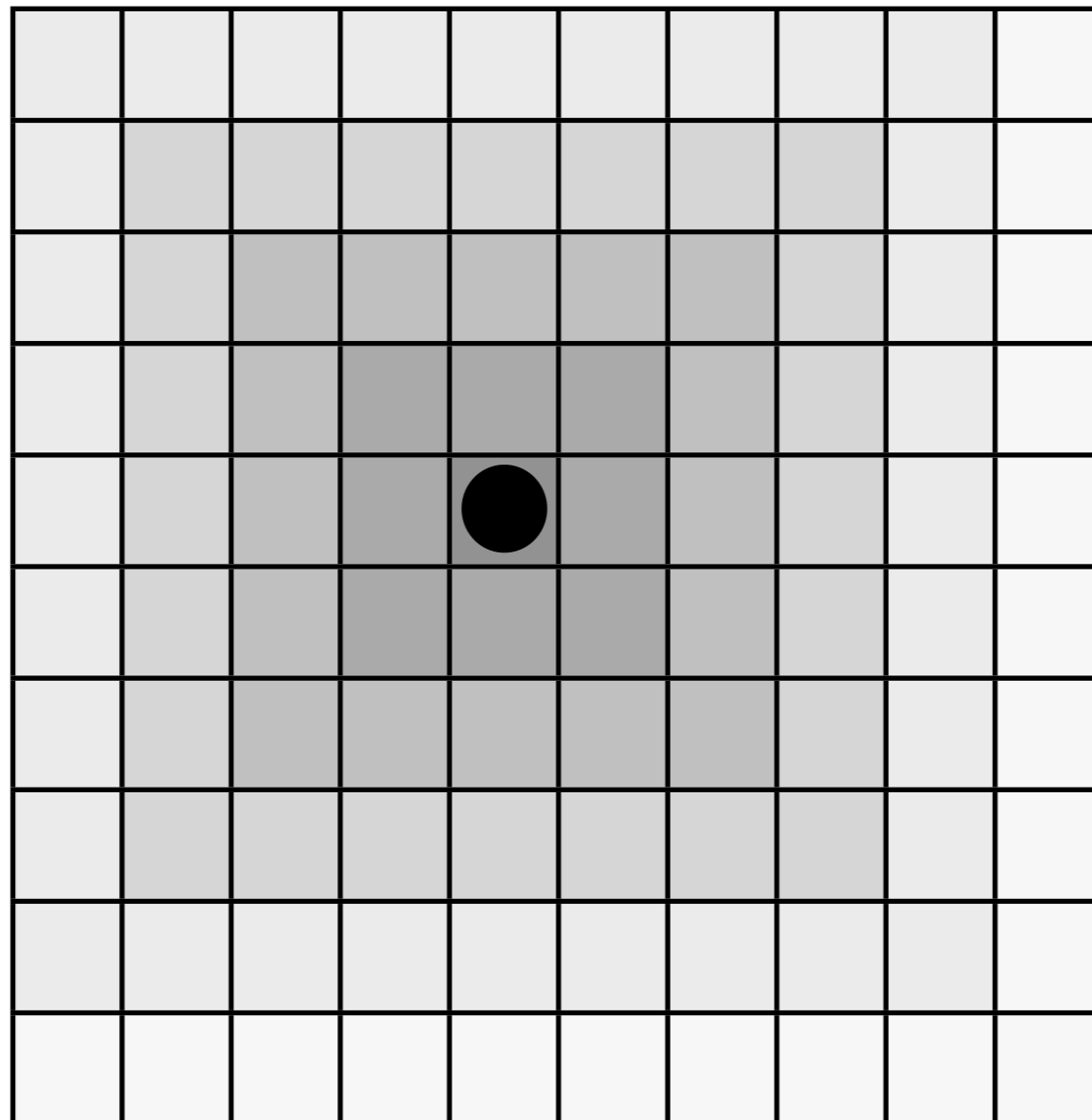
- ▶ It is due to the size principle that additional data points will cause generalisation curves to tighten



# Consequence of the size principle

---

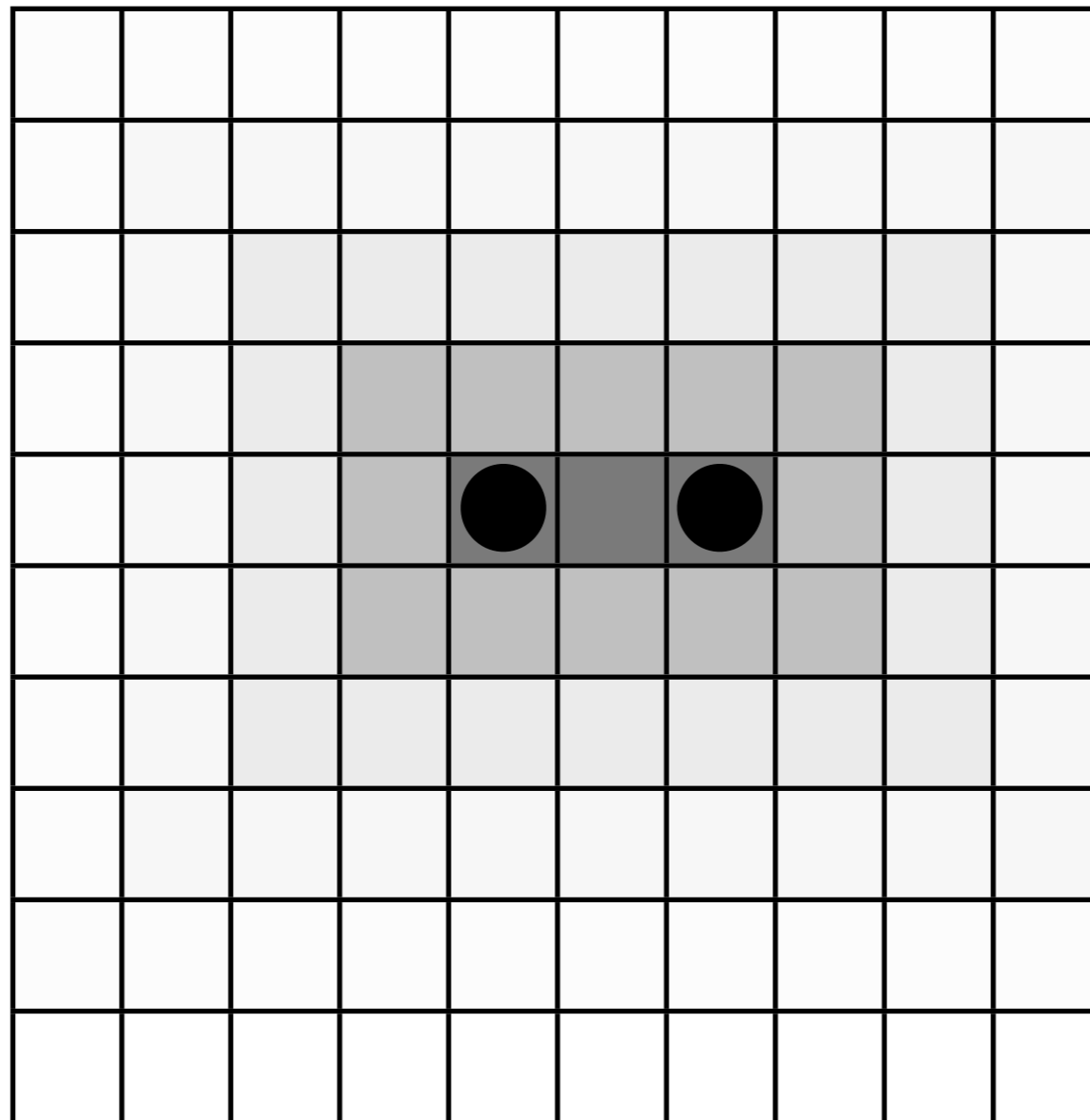
- ▶ It is due to the size principle that additional data points will cause generalisation curves to tighten



# Consequence of the size principle

---

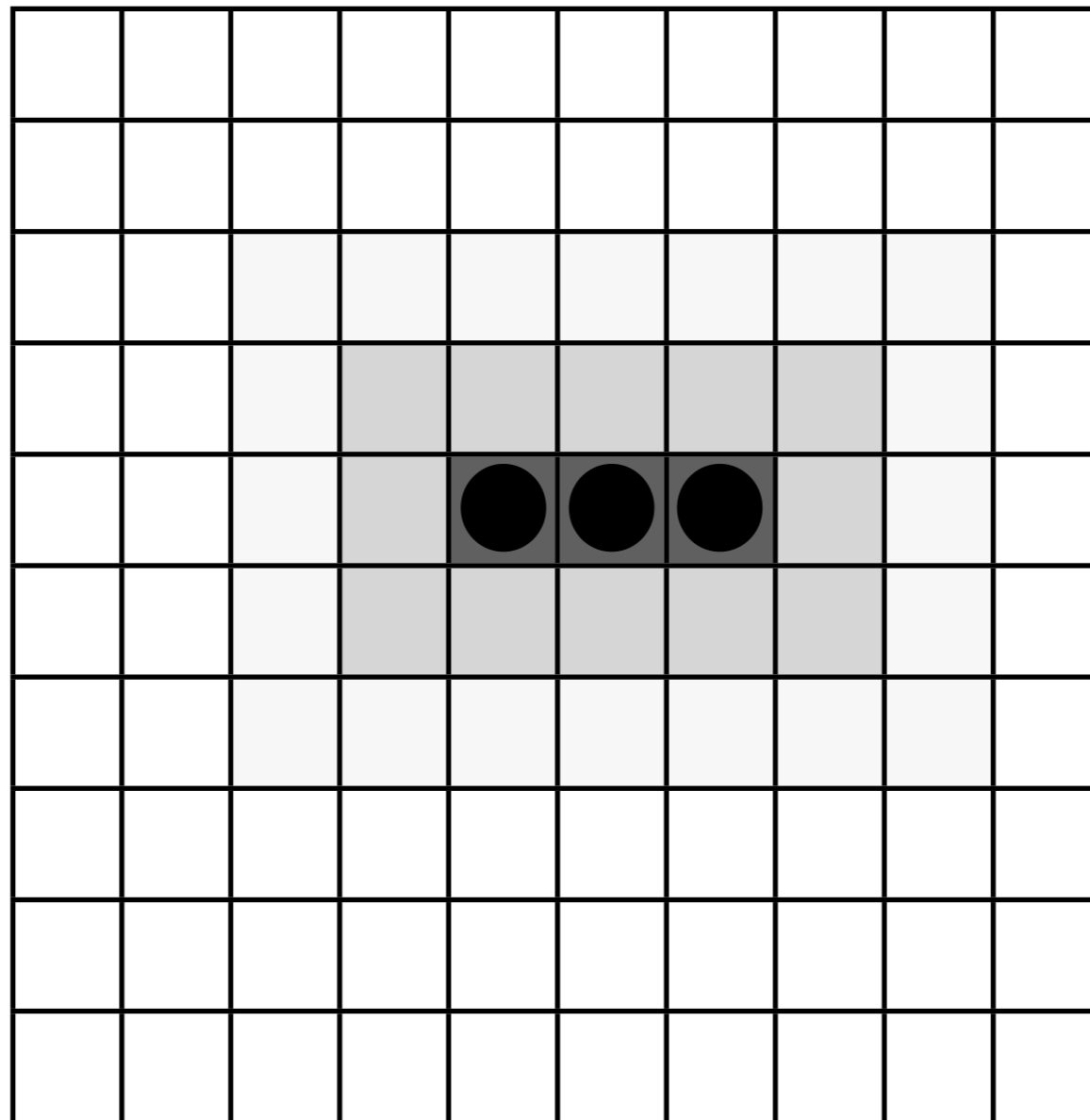
- ▶ It is due to the size principle that additional data points will cause generalisation curves to tighten



# Consequence of the size principle

---

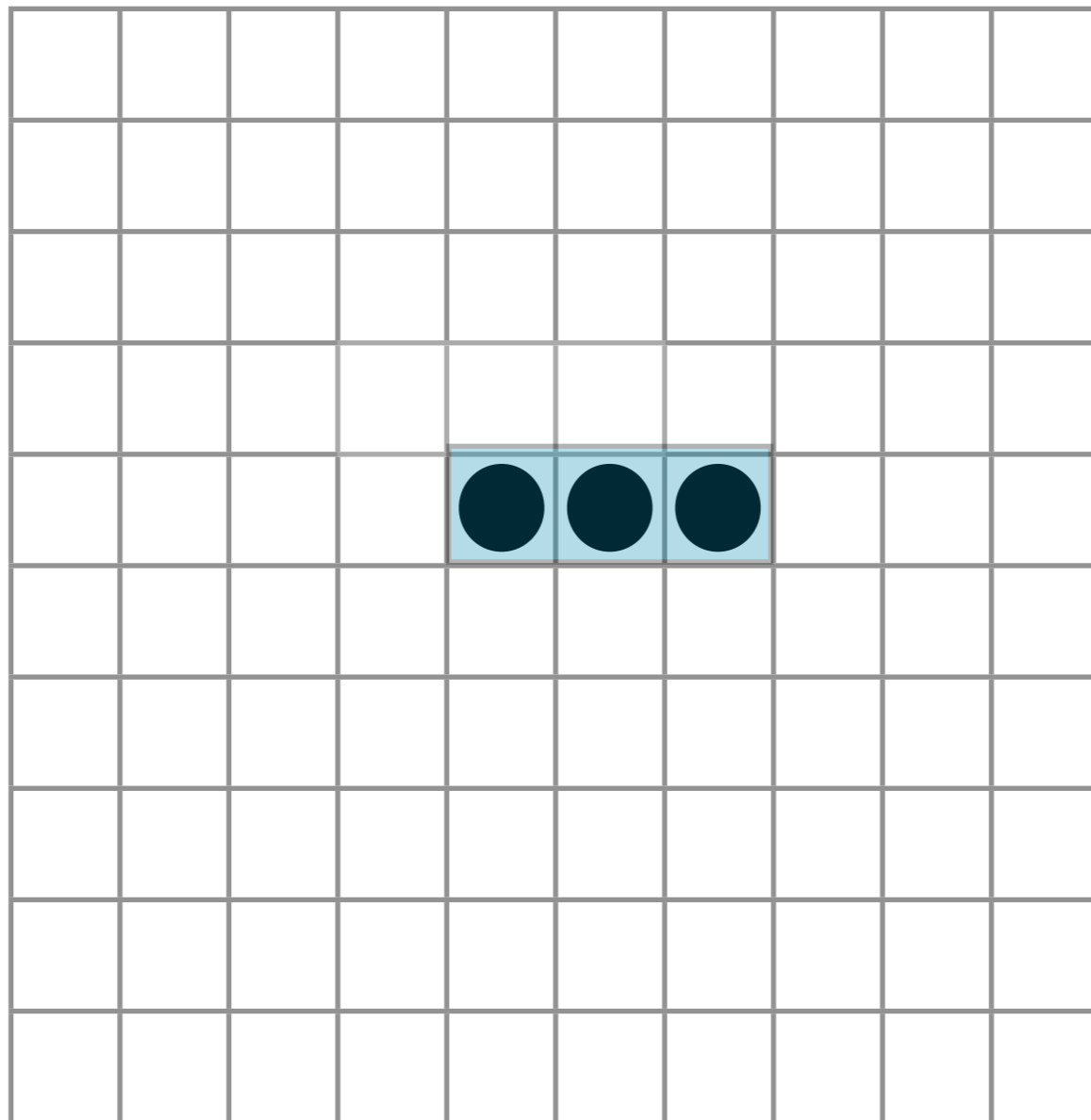
- ▶ It is due to the size principle that additional data points will cause generalisation curves to tighten



# Consequence of the size principle

---

- ▶ It is due to the size principle that additional data points will cause generalisation curves to tighten



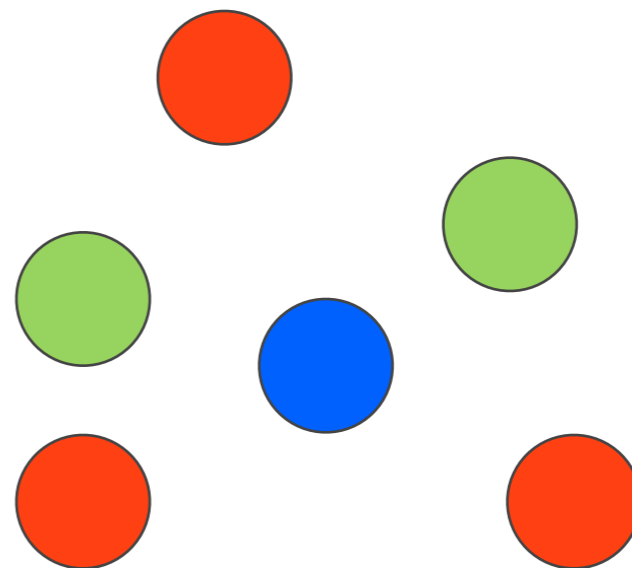
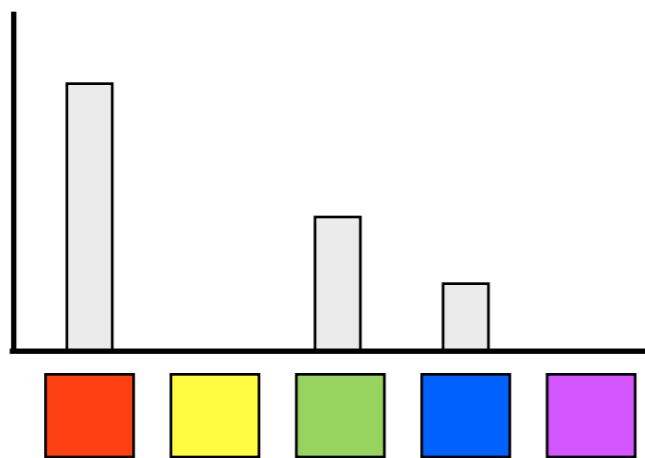
This is because it's quite a suspicious coincidence for these data points to have been generated if the true hypothesis is *not*  $h$

# The size principle is not the only way!

---

- ▶ It is sensible, but it follows from certain assumptions about how data were generated (or sampled)

Each point drawn independently and at random from the hypothesis



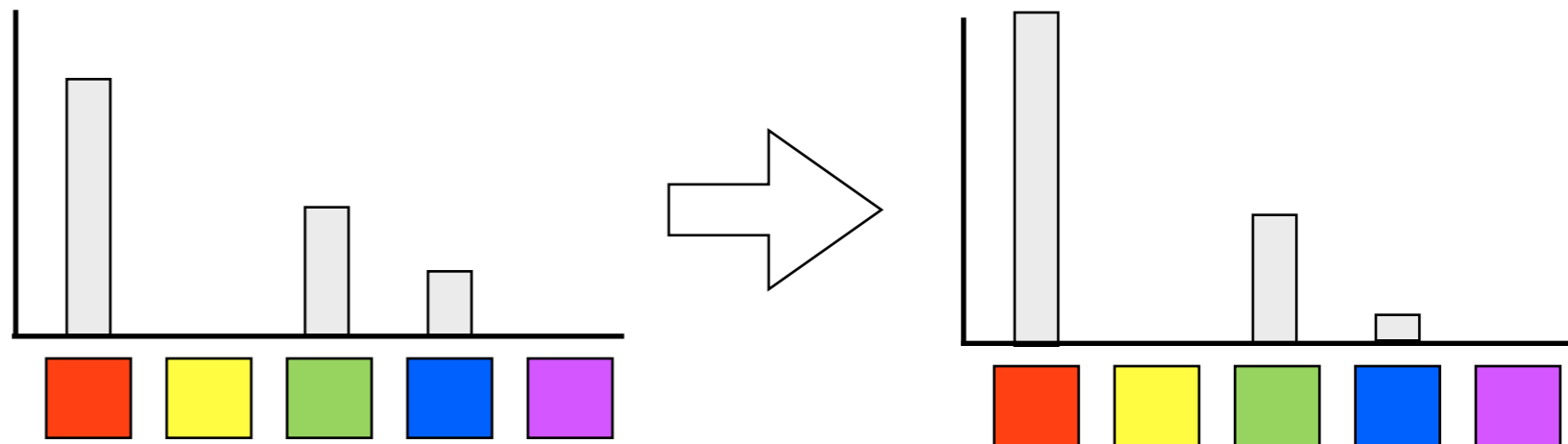
# The size principle is not the only way!

---

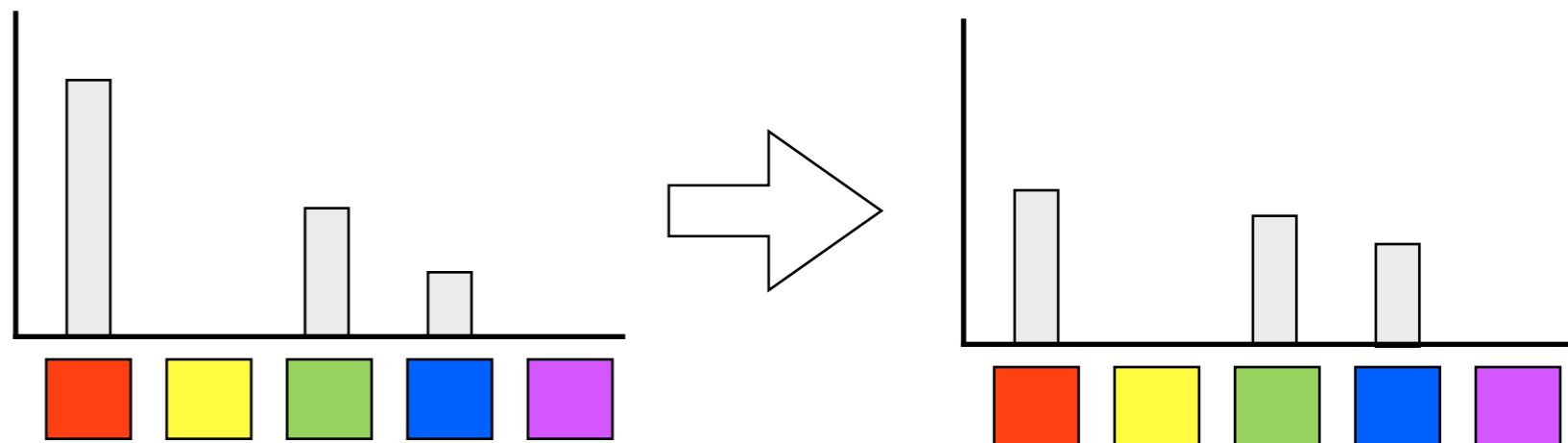
- ▶ It is sensible, but it follows from certain assumptions about how data were generated (or sampled)

Suppose data were non-independent:  
 $p(d_t|h)$  depended on  $p(d_{t-1}|h)$ .

If  $p(d_t=X|h)$  is larger if  
 $p(d_{t-1}=X|h)$ , the  
distribution skews



If  $p(d_t=X|h)$  is smaller if  
 $p(d_{t-1}=X|h)$ , the  
distribution flattens

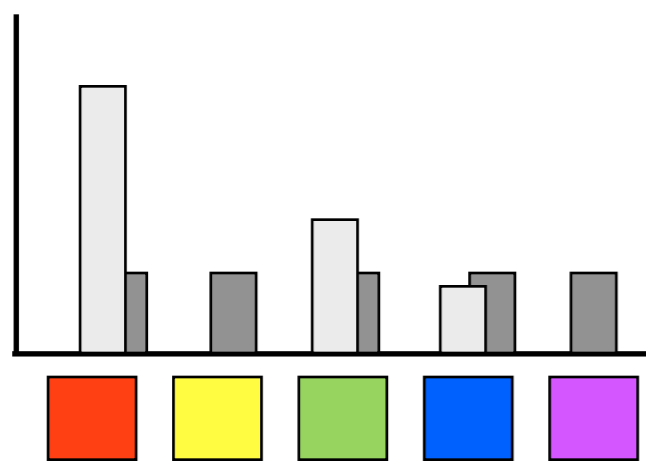


# The size principle is not the only way!

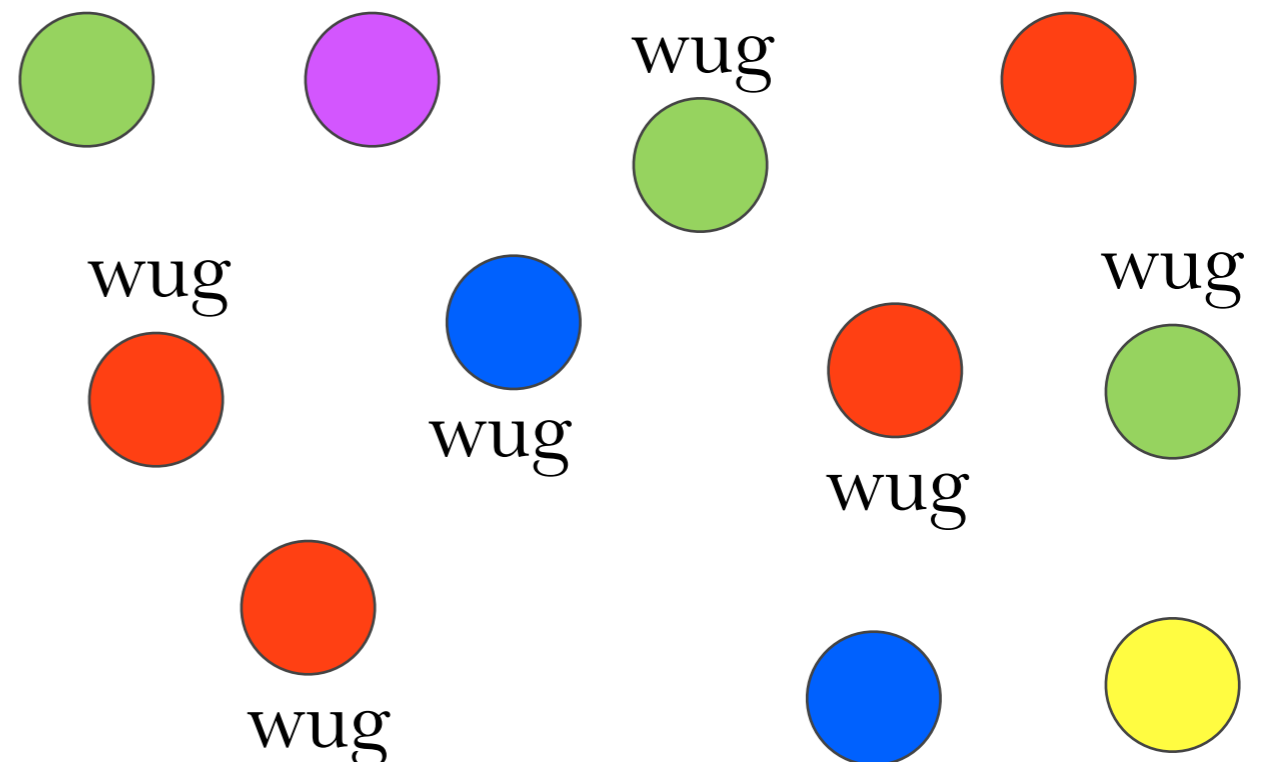
---

- ▶ It is sensible, but it follows from certain assumptions about how data were generated (or sampled)

Suppose data could have been generated from the world in general, and only then labelled as belonging to the hypothesis (or not)



□ Wugs  
□ Things in the world



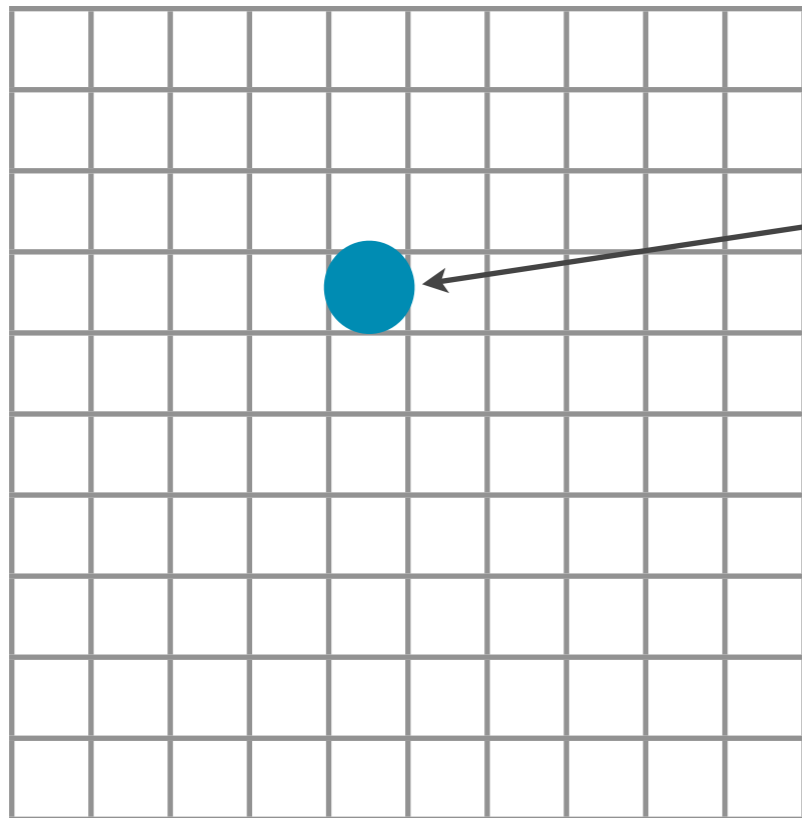


# The size principle is not the only way!

---

- ▶ It is sensible, but it follows from certain assumptions about how data were generated (or sampled)

Suppose data could have been generated from the world in general, and only then labelled as belonging to the hypothesis (or not)



Data sampled from the world at random

Then labelled as in the hypothesis or not

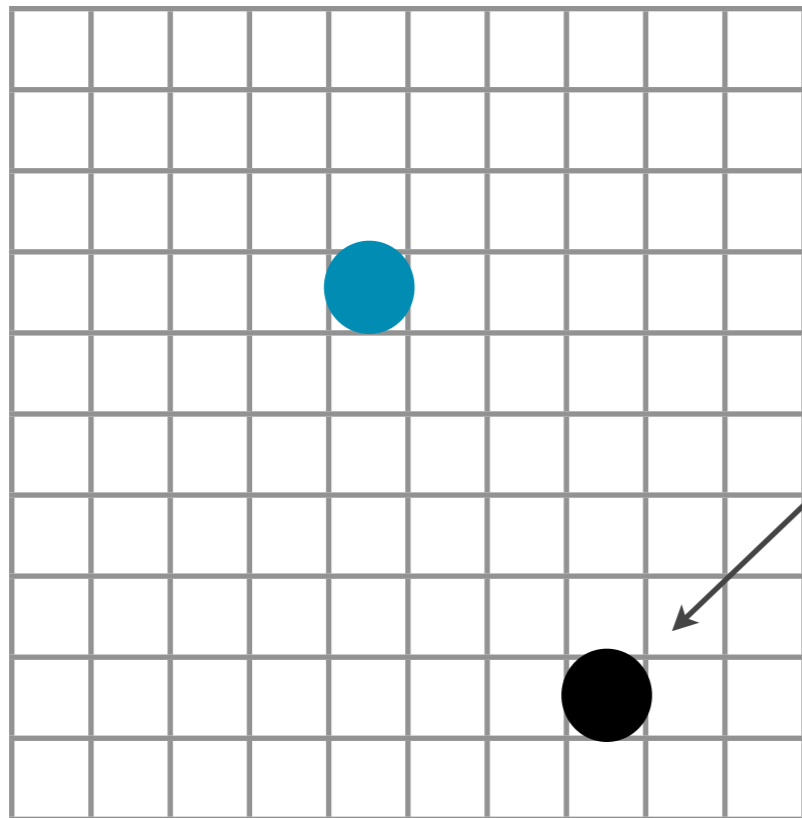
$$p(d=\bullet|h) = 1 \text{ if in the hypothesis} \\ 0 \text{ if not}$$

# The size principle is not the only way!

---

- ▶ It is sensible, but it follows from certain assumptions about how data were generated (or sampled)

Suppose data could have been generated from the world in general, and only then labelled as belonging to the hypothesis (or not)



Data sampled from the world at random

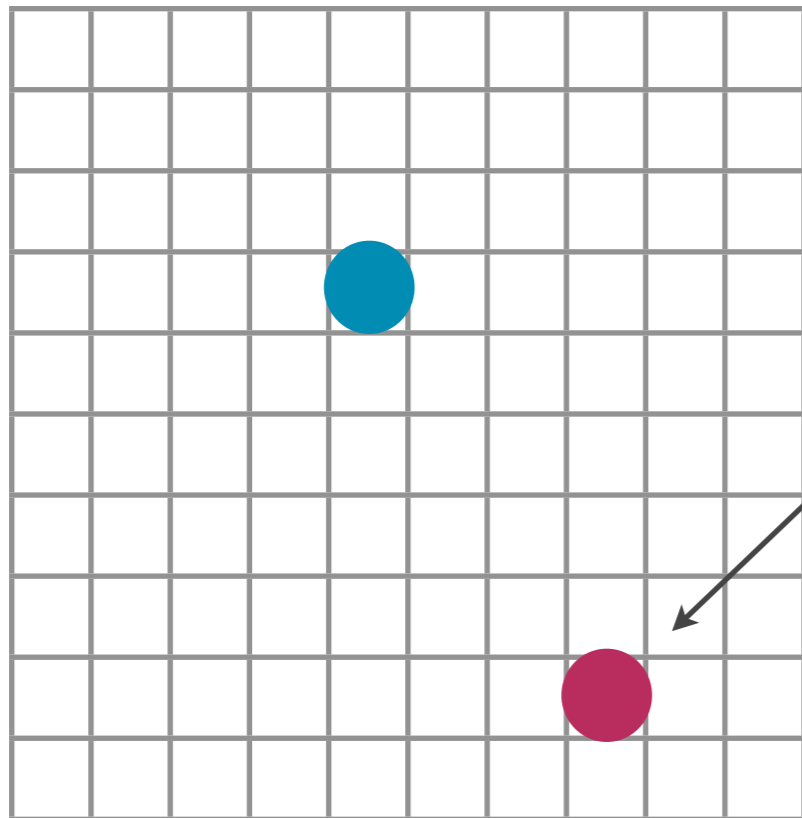
$$p(d=\bullet|h) = 1 \text{ if in the hypothesis} \\ 0 \text{ if not}$$

# The size principle is not the only way!

---

- ▶ It is sensible, but it follows from certain assumptions about how data were generated (or sampled)

Suppose data could have been generated from the world in general, and only then labelled as belonging to the hypothesis (or not)



Data sampled from the world at random

Then labelled as in the hypothesis or not

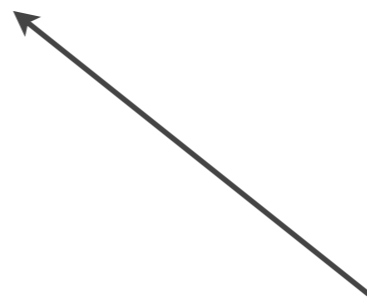
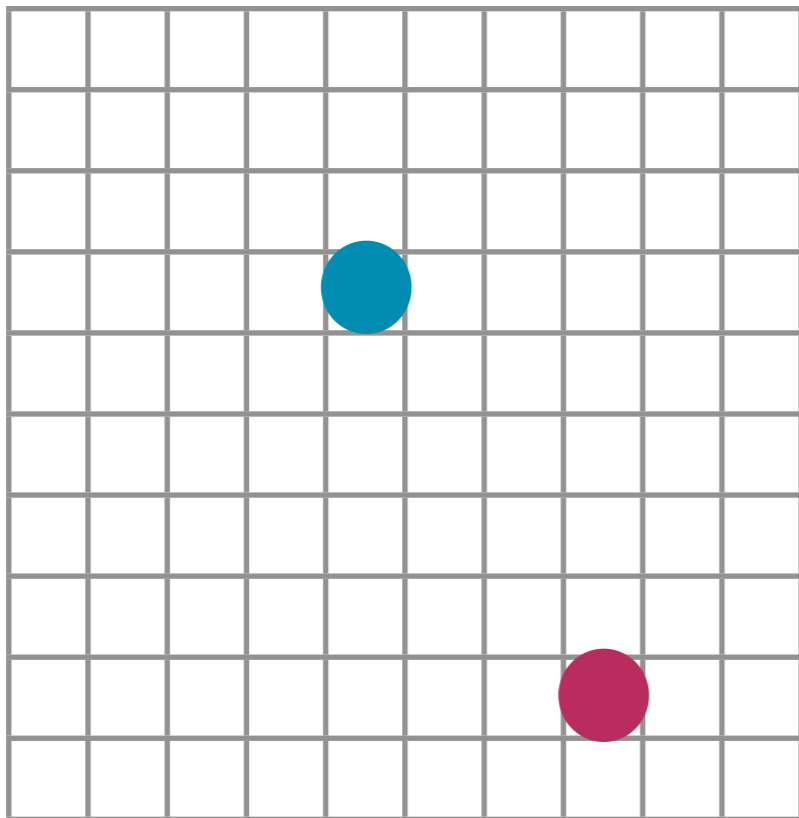
$$p(d=\bullet|h) = 1 \text{ if in the hypothesis} \\ 0 \text{ if not}$$

# The size principle is not the only way!

---

- ▶ It is sensible, but it follows from certain assumptions about how data were generated (or sampled)

Suppose data could have been generated from the world in general, and only then labelled as belonging to the hypothesis (or not)



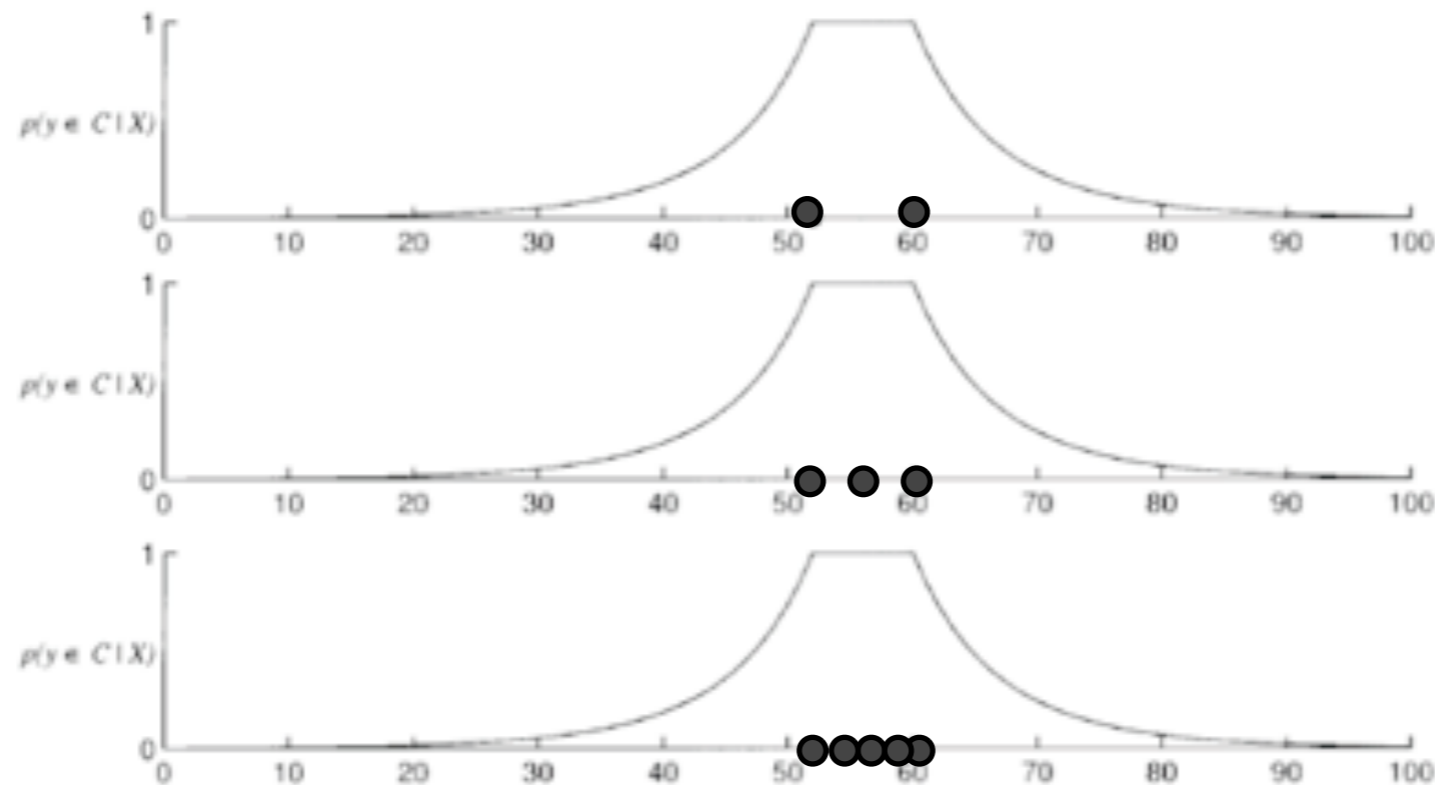
This is called **weak sampling**

$$p(d=\bullet|h) = 1 \text{ if in the hypothesis} \\ 0 \text{ if not}$$

# The size principle is not the only way!

---

- ▶ If data are weakly sampled, the generalisation curves should not tighten -- there is no suspicious coincidence since the data were generated by the *world*, and not from the hypothesis



# Are people sensitive to sampling assumptions?

---

Do people change their generalisations if the data have been sampled differently?

# Plan

---

➔ How is the data generated?

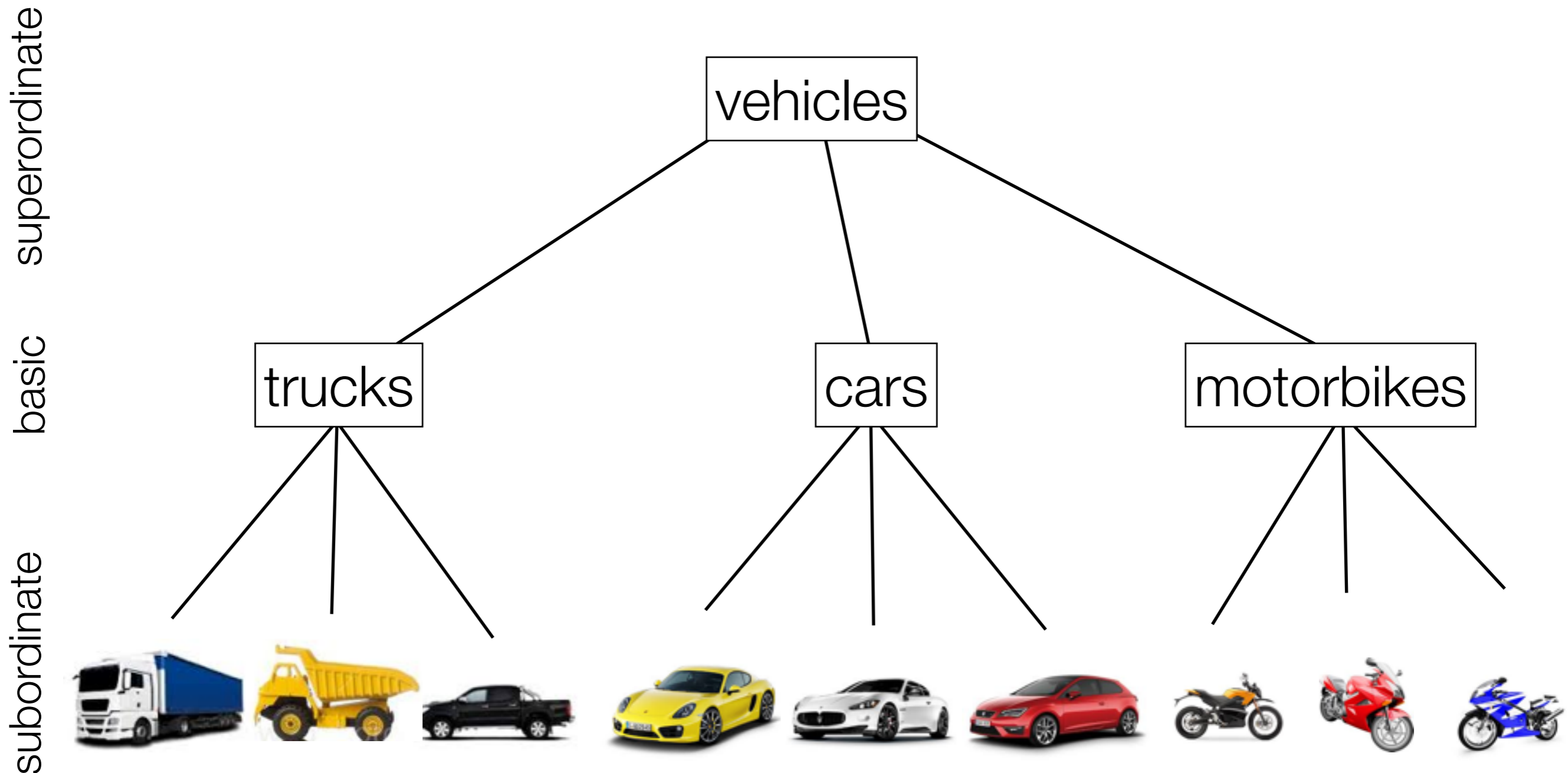
- Strong vs. weak sampling: the idea

➔ People's sensitivity to sampling assumptions

- Individual differences in sampling sensitivity

# Word learning

- ▶ We've already seen that many domains have a hierarchical or tree-based conceptual structure





# Word learning

---

- ▶ We've already seen that many domains have a hierarchical or tree-based conceptual structure

superordinate

vegetables

capsicums

potatoes

eggplants

basic

subordinate



# Word learning

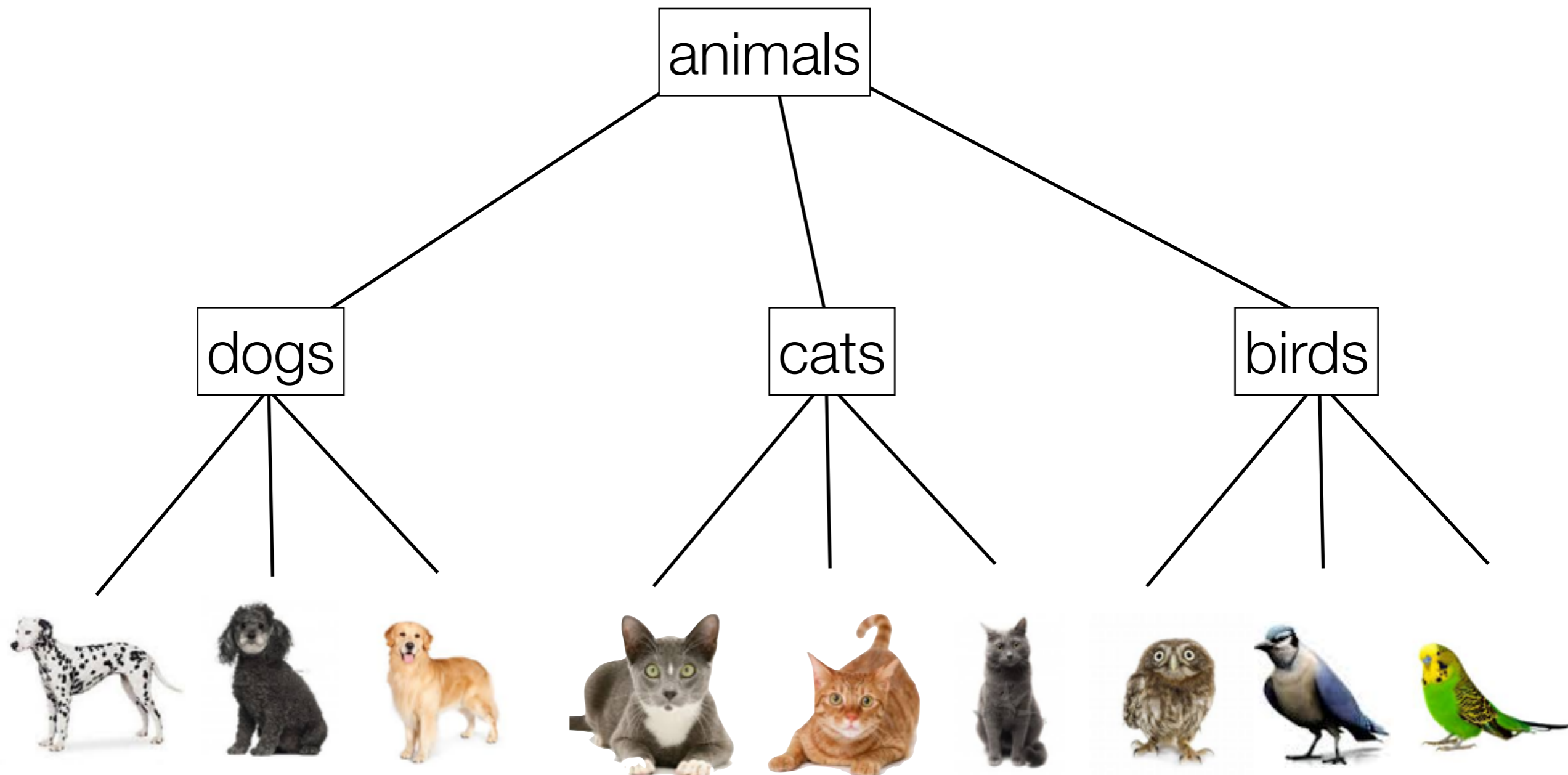
---

- ▶ We've already seen that many domains have a hierarchical or tree-based conceptual structure

superordinate

basic

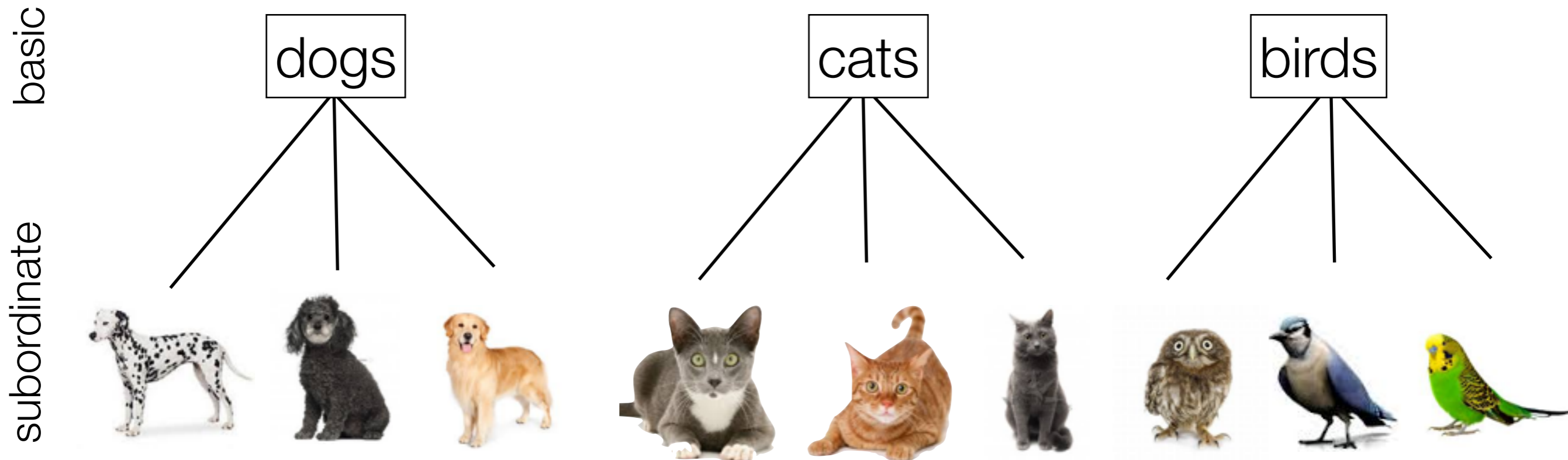
subordinate



# Word learning

---

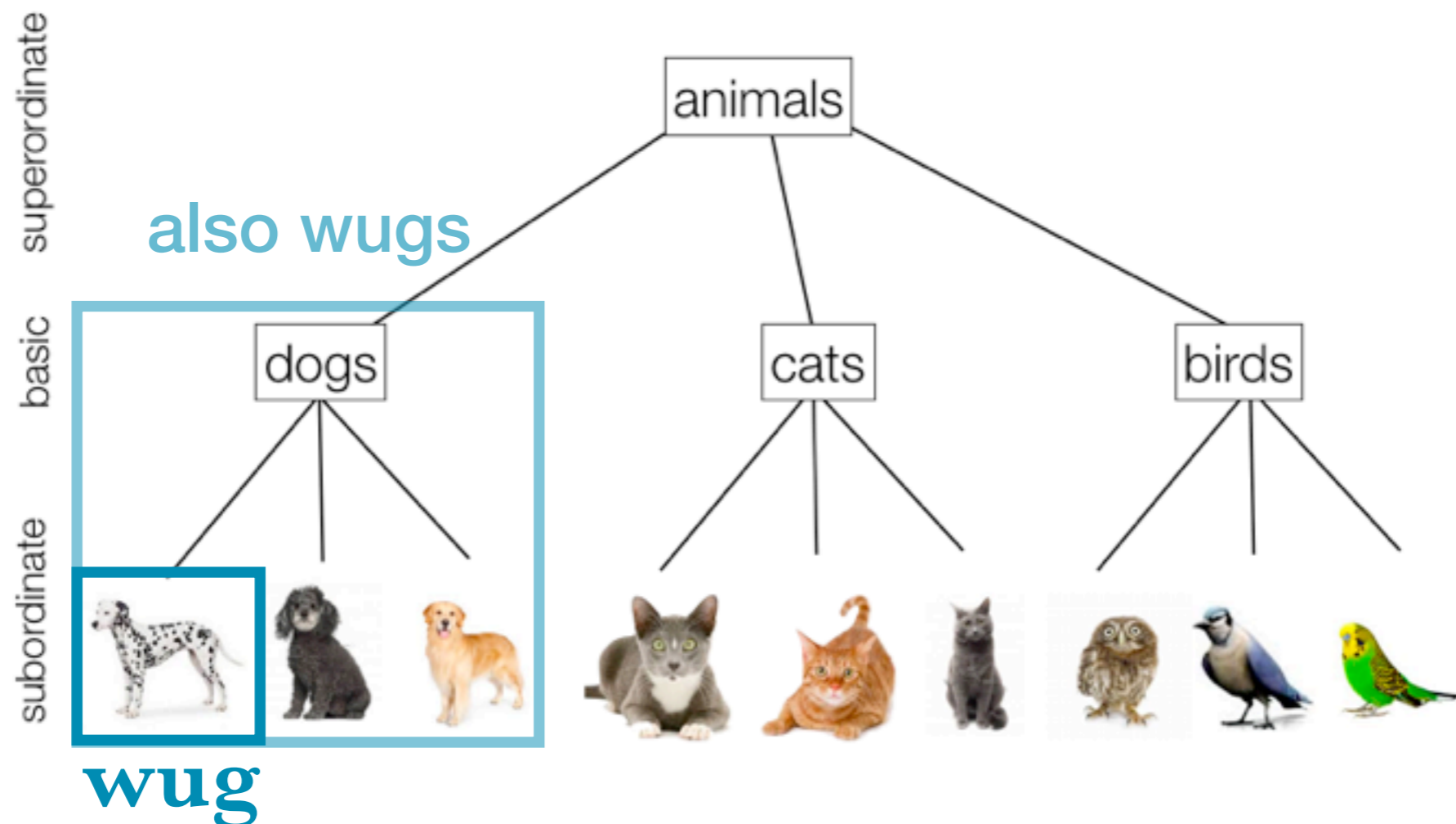
- ▶ There is lots of independent evidence that the basic level is privileged: it is what people default to when using names, it has the highest inductive power, etc



# Word learning

---

- ▶ We would therefore expect that if people were told that *one* item was a wug, people would guess that all other items at the basic level are wugs too



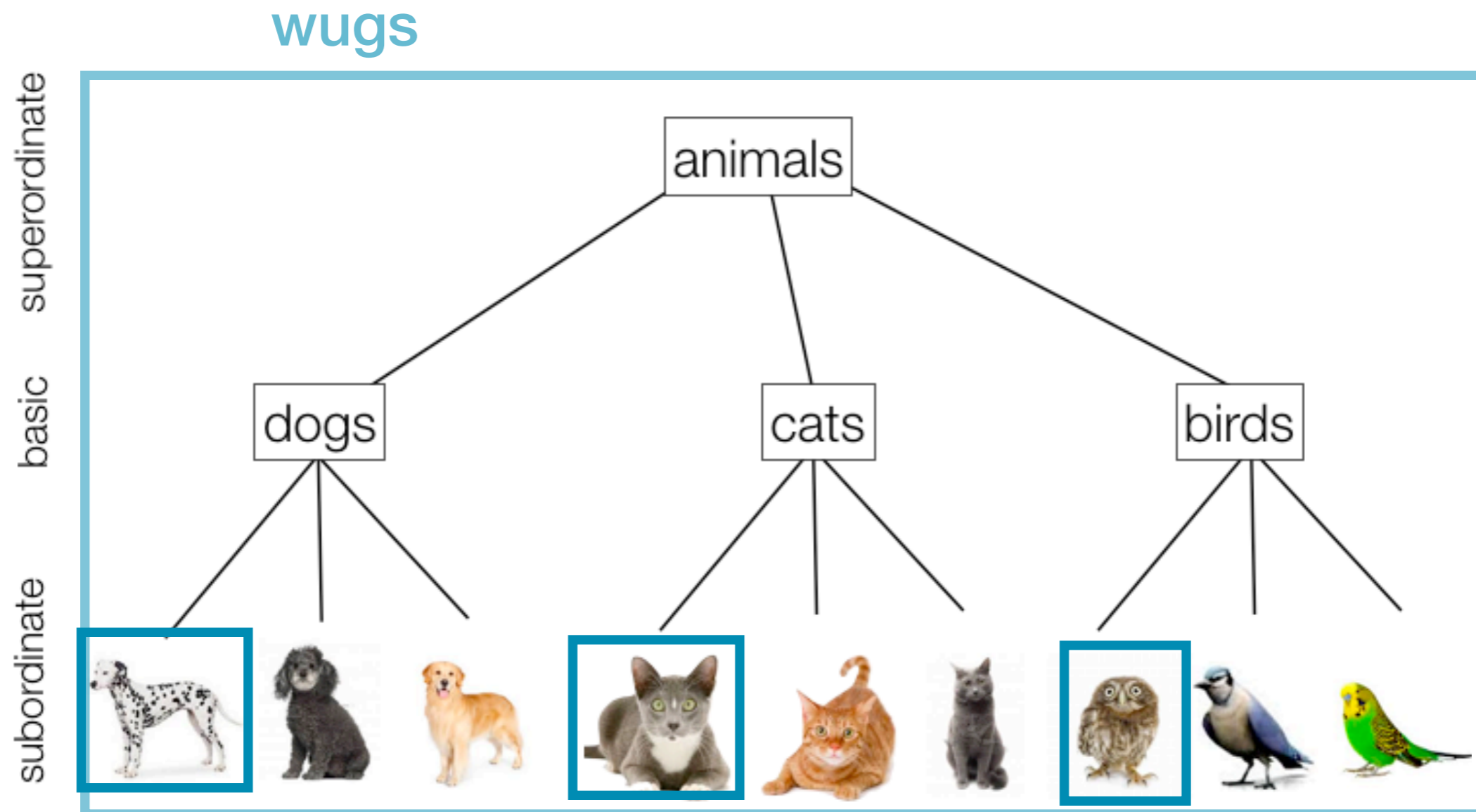
# Word learning

---

- ▶ But what if we are given *three* examples of wugs?
- ▶ Then it depends on which three examples, and whether people are reasoning based on the size principle...

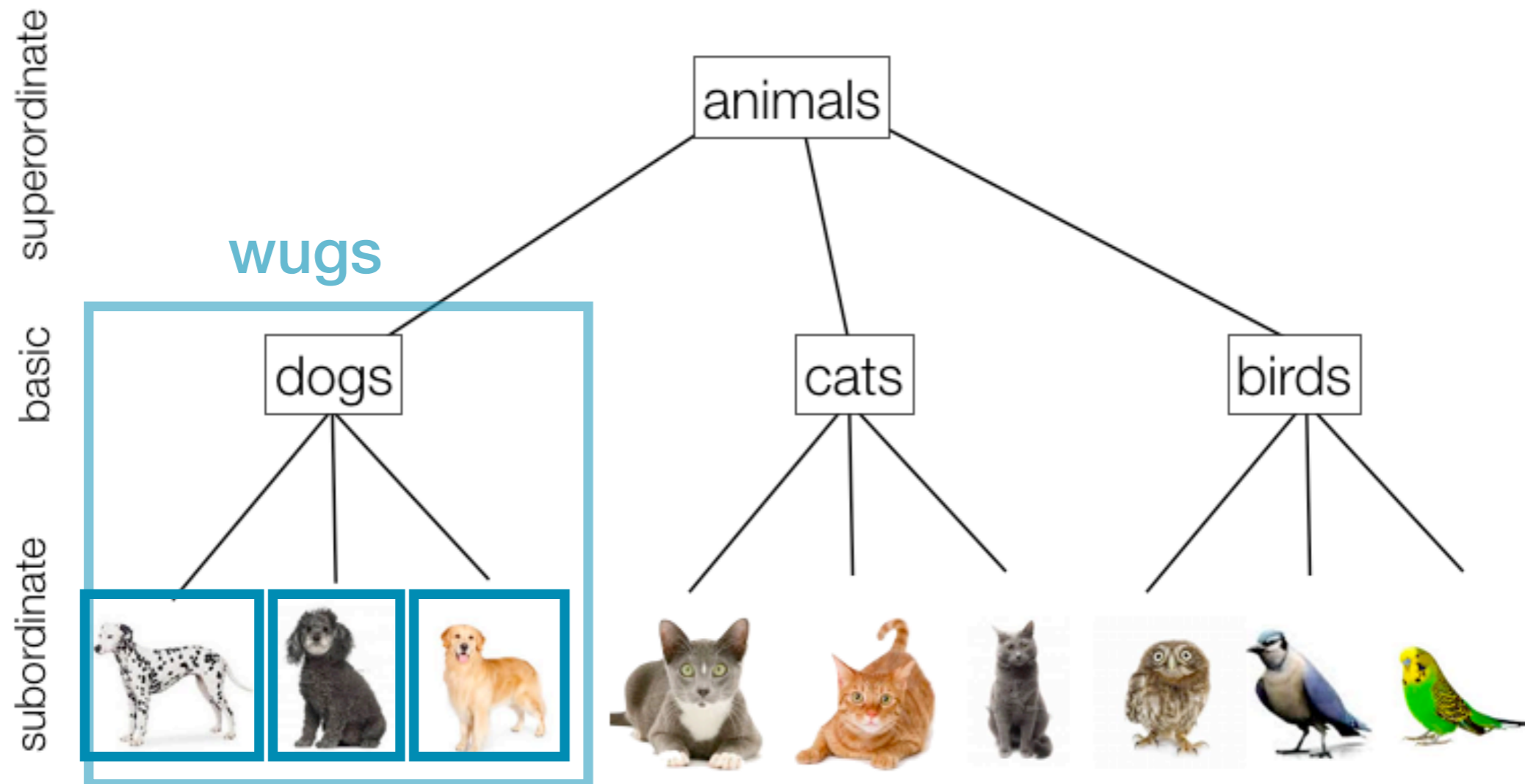
# If people use the size principle

- ▶ Then they should make the tightest possible generalisation



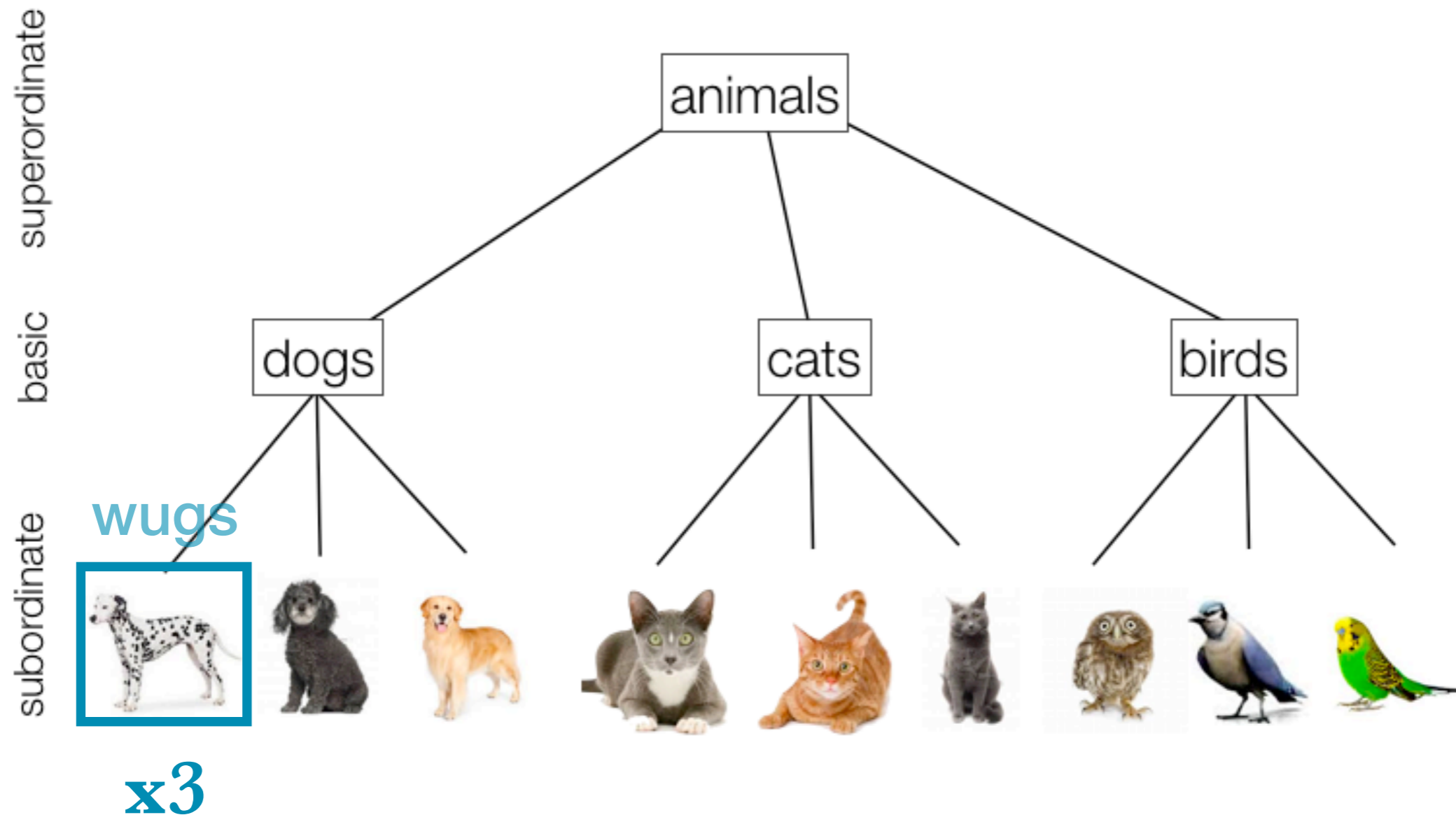
# If people use the size principle

- ▶ Then they should make the tightest possible generalisation



# If people use the size principle

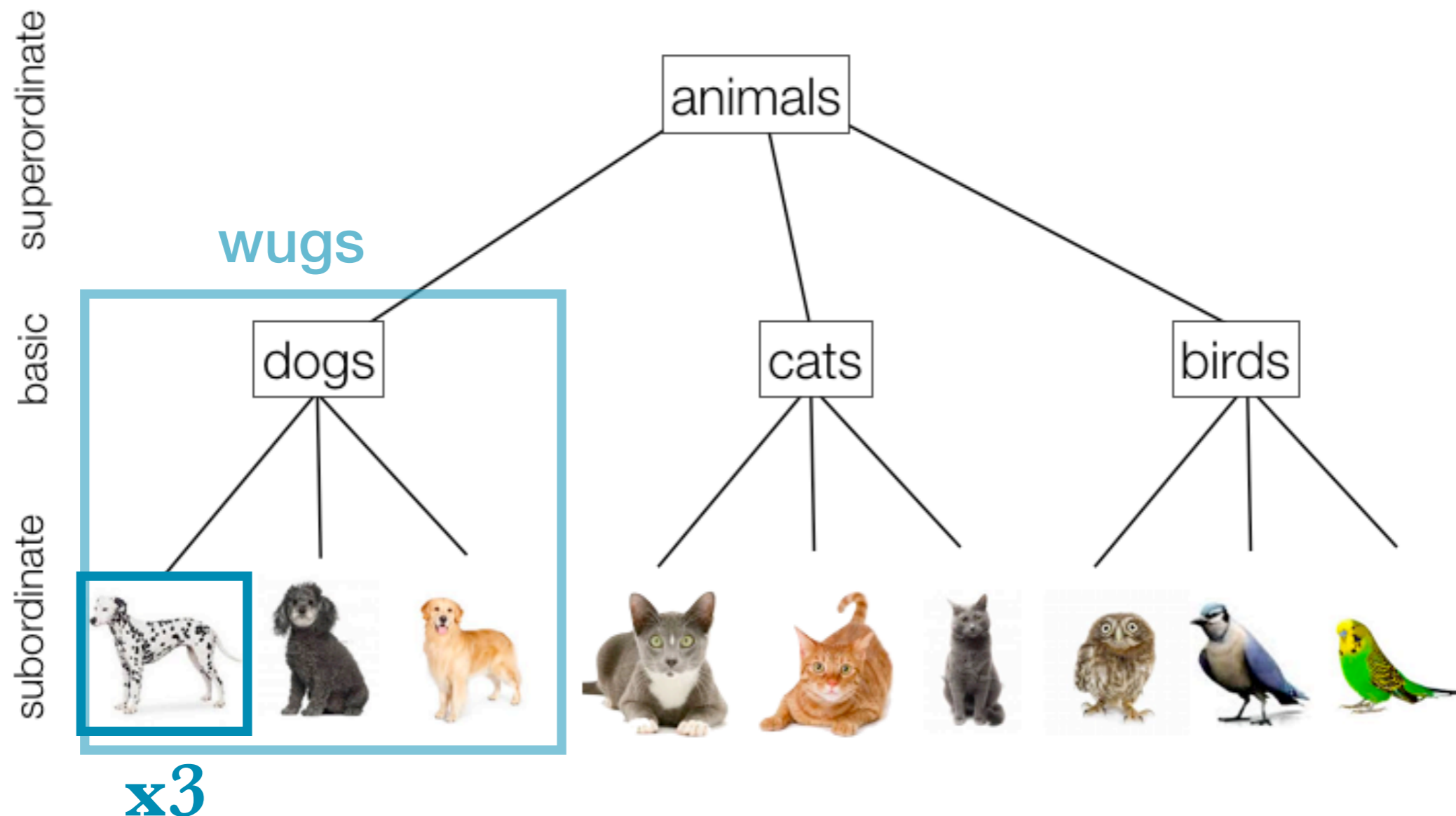
- ▶ Then they should make the tightest possible generalisation

















# If people DON'T use the size principle

- ▶ Then they should not tighten their generalisation when given three of the same item - there is no “suspicious coincidence” to explain



# Test

- ▶ Four conditions, in each of three domains

	Vegetables	Vehicles	Animals
1 example			
3 subordinate examples			
3 basic-level examples			
3 superordinate examples			

# Test

---

- ▶ Four conditions, in each of three domains

## Adults

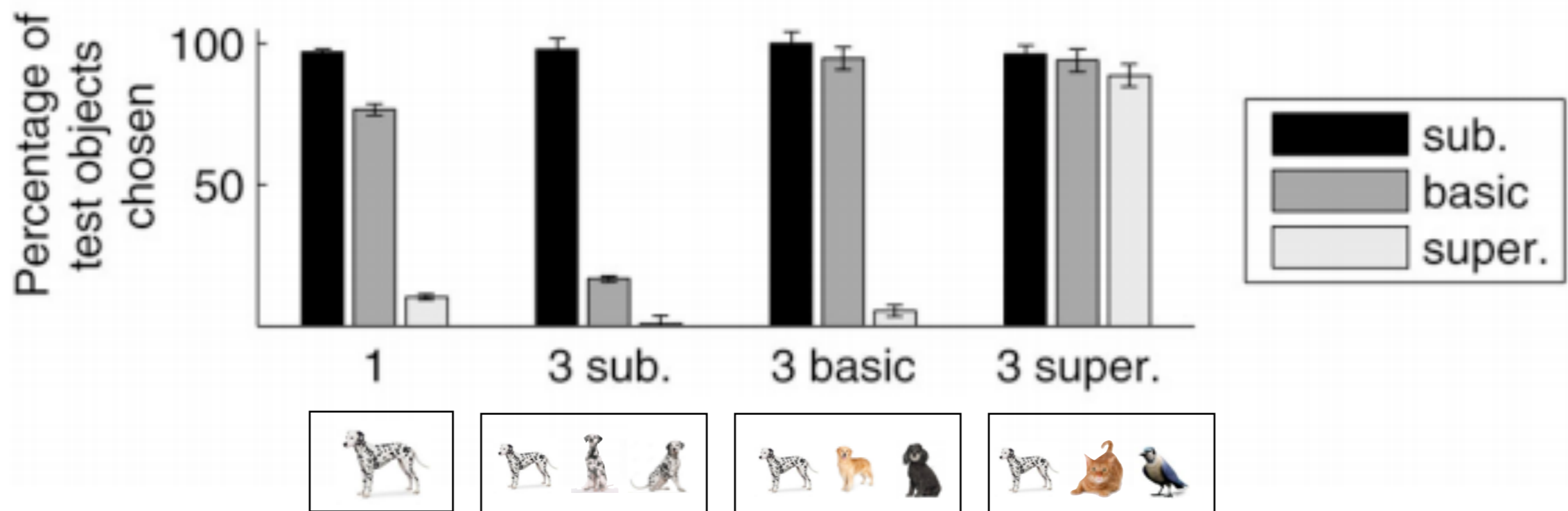
Please select the other objects that this word applies to (i.e., the other wugs).

## Four-year-old children

Mr. Frog speaks a different language, and he has different names than we do for his toys. He is going to pick out some of them, and he would like you to help him pick out the others like he has picked out, OK?

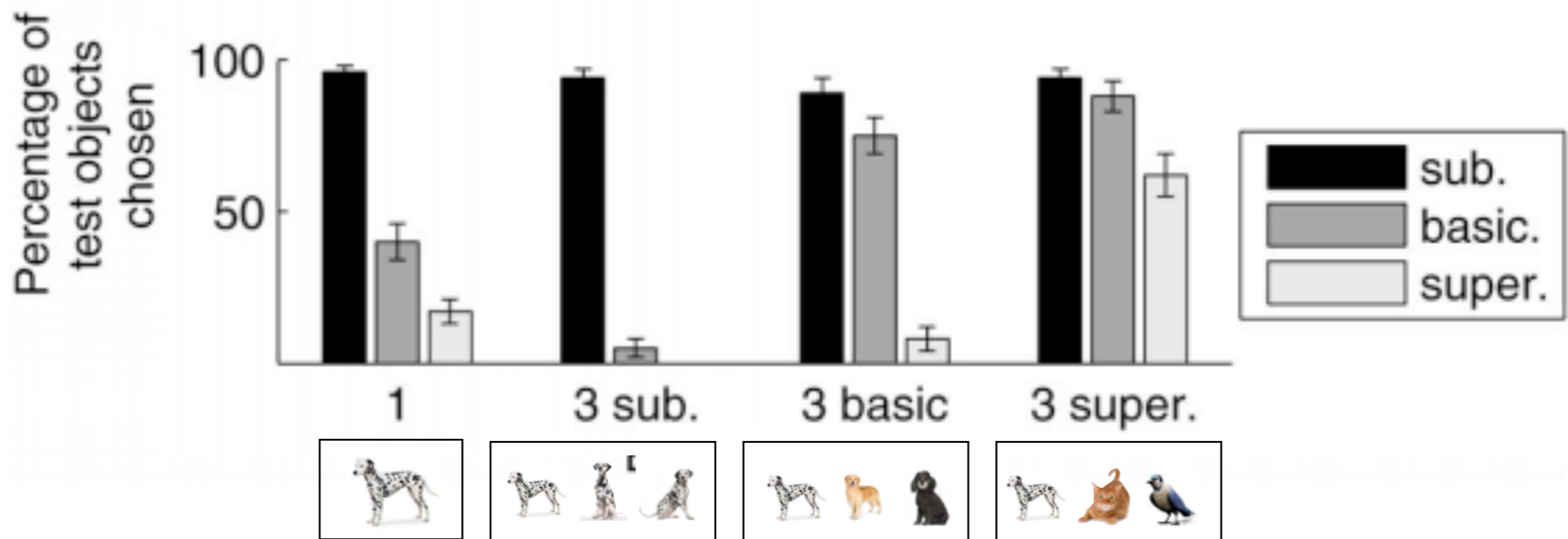
# Test

- ▶ Adults generalise as predicted by the size principle

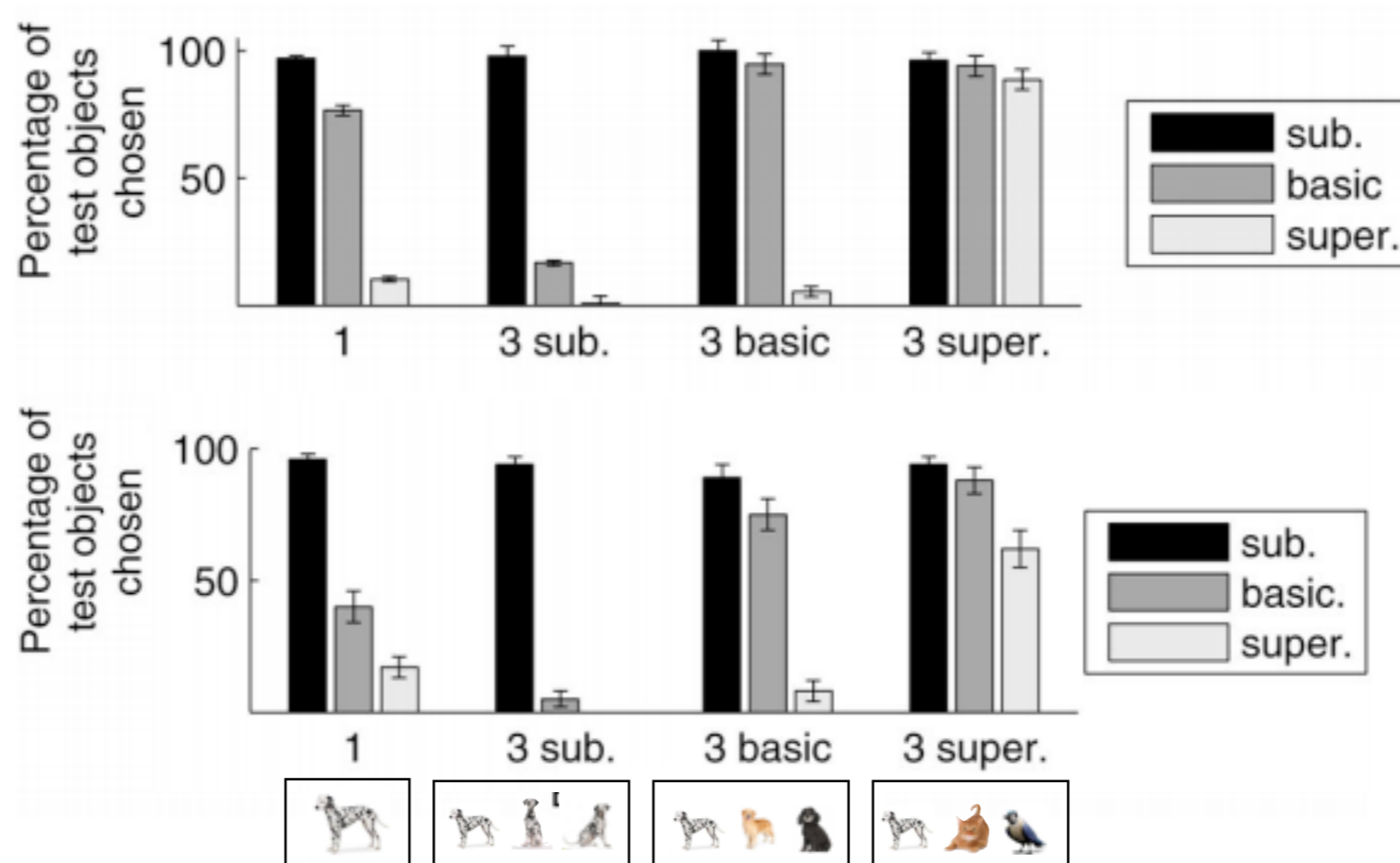


# Test

- ▶ Four-year old children do the same thing!



# Test

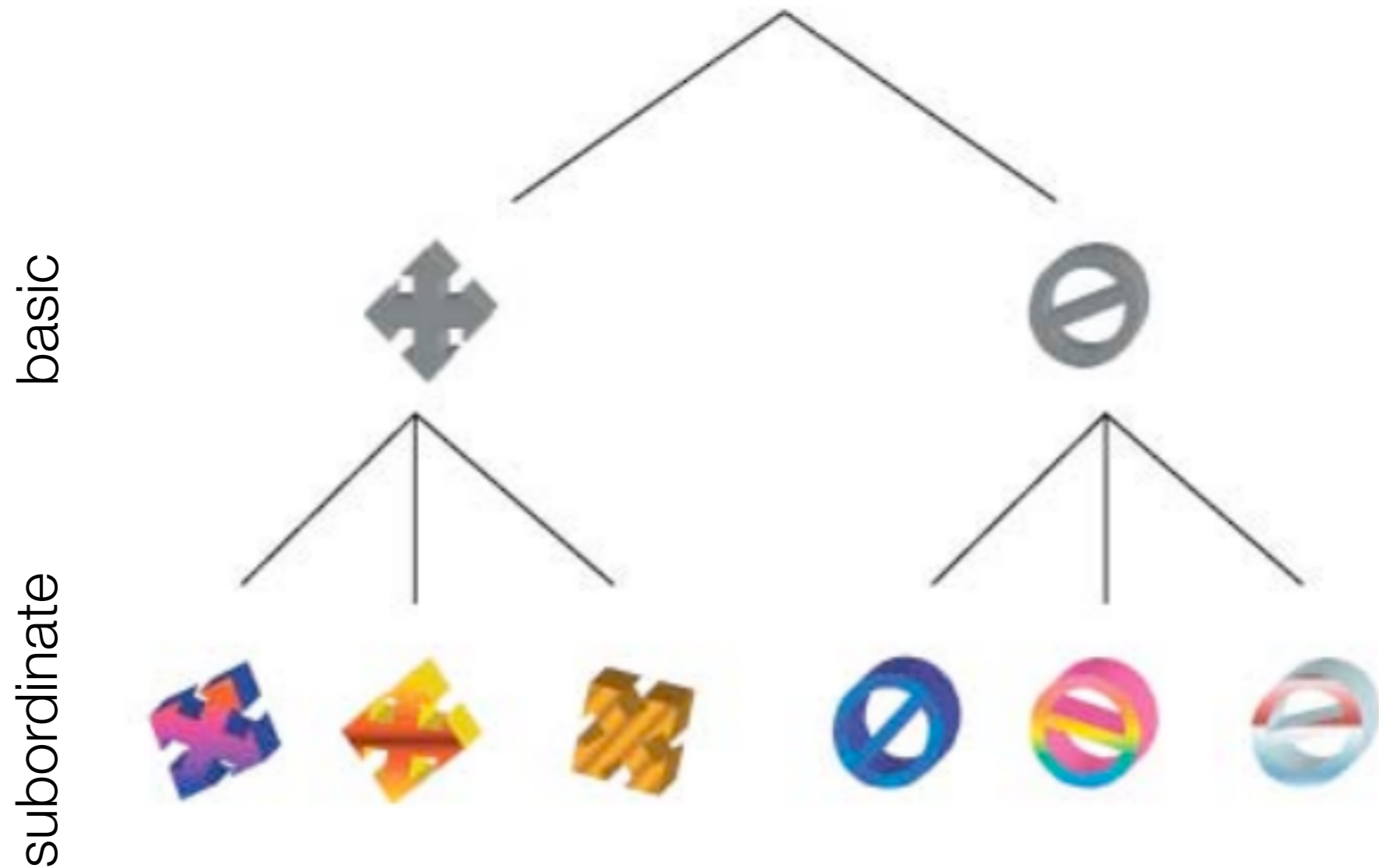


- ▶ But so far this just shows that people follow the qualitative pattern predicted by the size principle. It does not imply that they are sensitive to sampling assumptions -- perhaps they would tighten generalisations no matter what

# Changing sampling assumptions

---

- ▶ This time we vary how data are sampled (also make the objects novel)

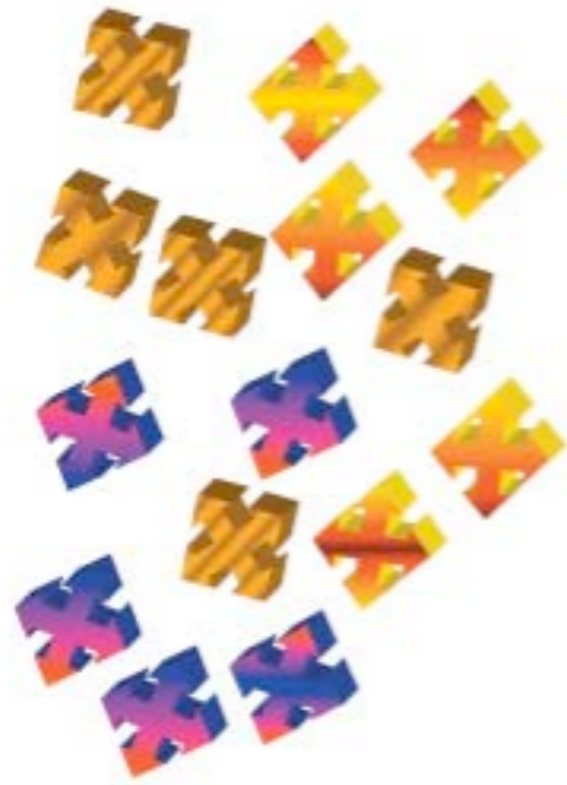


# Changing sampling assumptions

---

- ▶ This time we vary how data are sampled (also make the objects novel)

## Teacher-driven





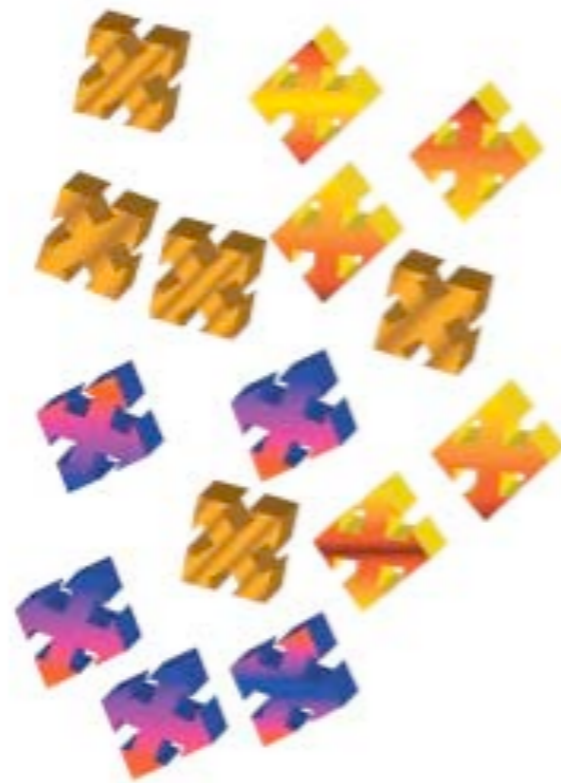
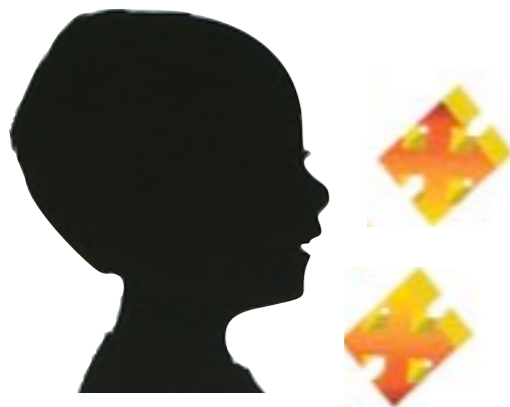
# Changing sampling assumptions

---

- ▶ This time we vary how data are sampled (also make the objects novel)

## Learner-driven

All participants chose two items from the same subordinate category



# Changing sampling assumptions

---

- ▶ This time we vary how data are sampled (also make the objects novel)

## Learner-driven

So in this condition people always saw items from the subordinate category, but the 3 items were not chosen by the teacher



## Teacher-driven

People saw 3 subordinate items, always chosen by the teacher

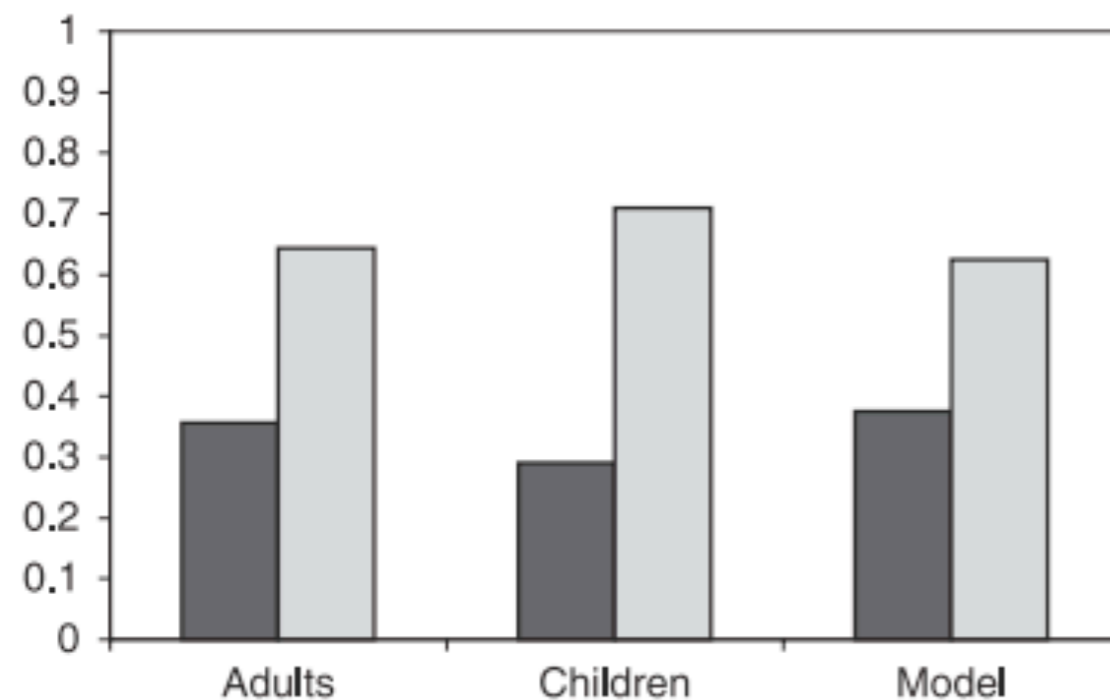


# Changing sampling assumptions

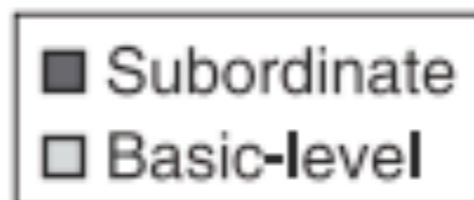
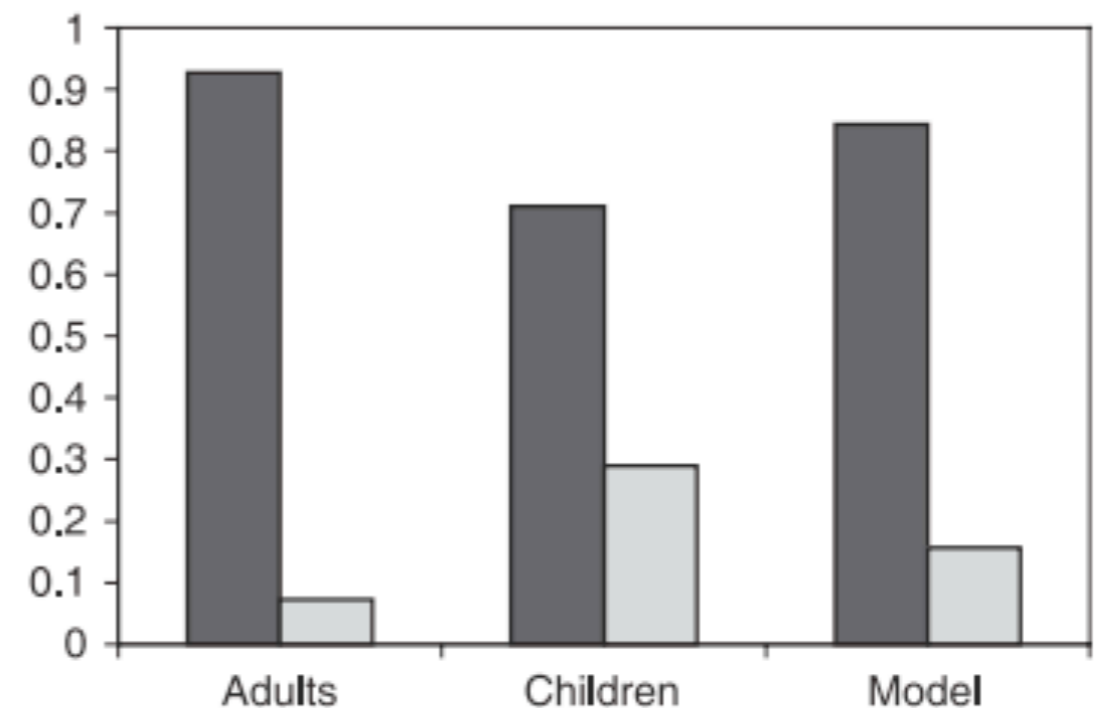
---

- ▶ People generalise tightly only when the teacher sampled the data

## Learner-driven



## Teacher-driven



# Changing sampling assumptions

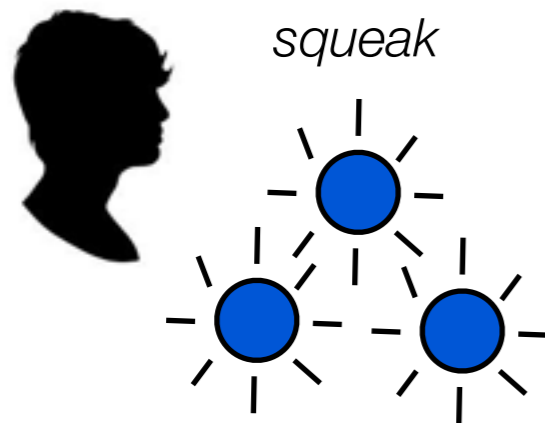
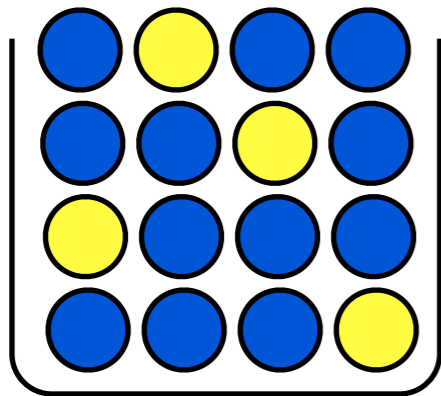
---

This shows that people generalise word labels differently based on how the data was sampled. How about generalising properties? And what about very young children?

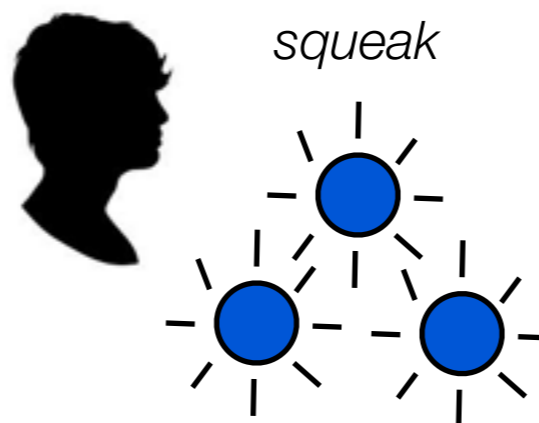
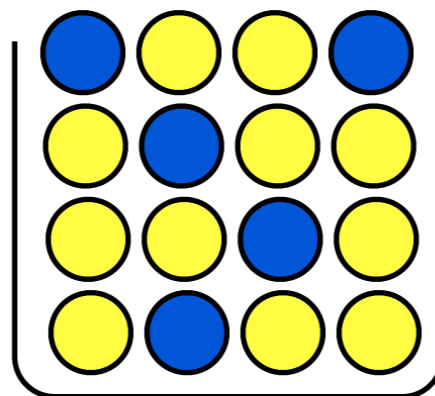
# Infants' use of sampling assumptions

---

Experiment 1



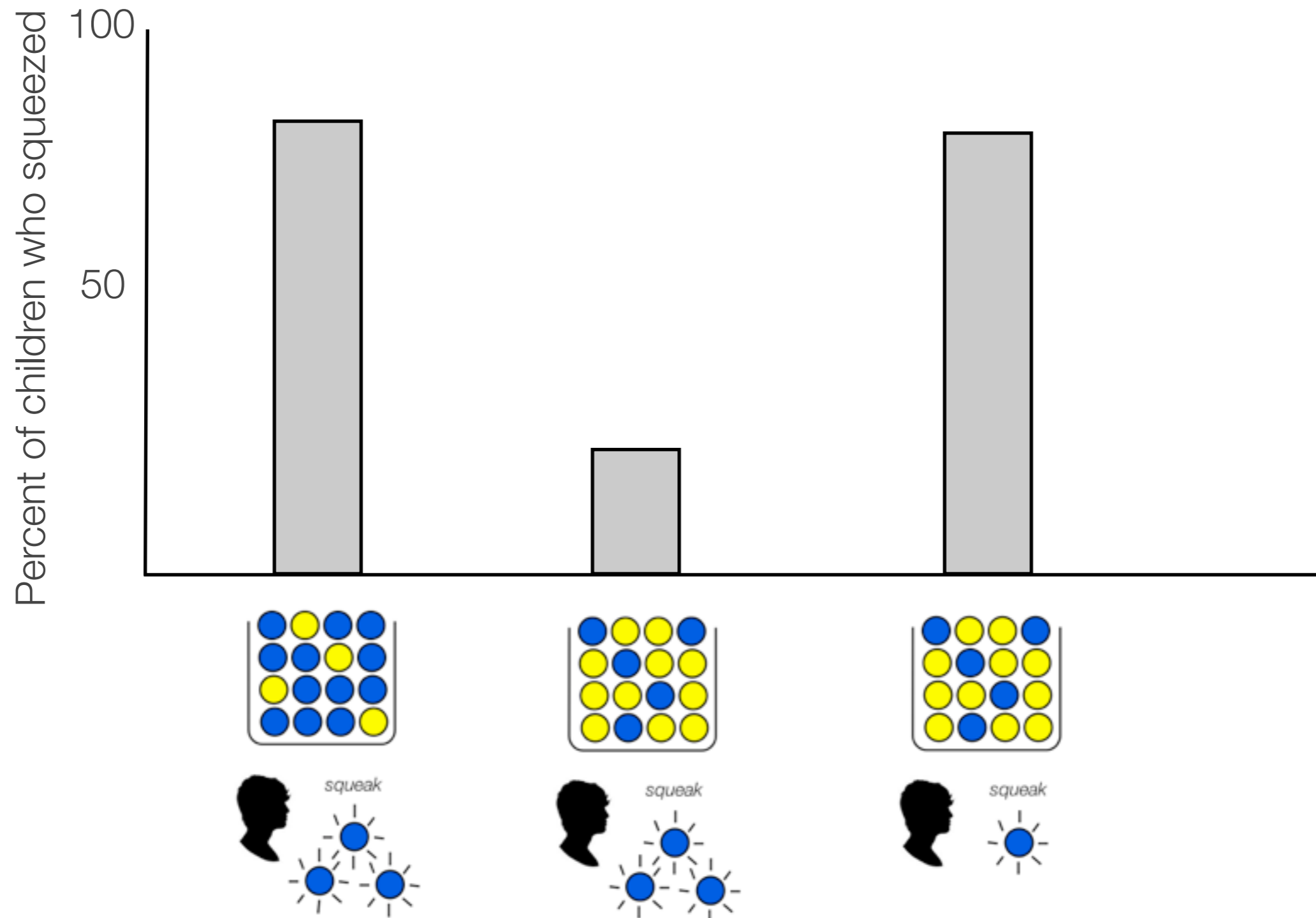
Experiment 2



What do 15-month old infants think about this ball?  
Will it squeak?



# Infants' use of sampling assumptions



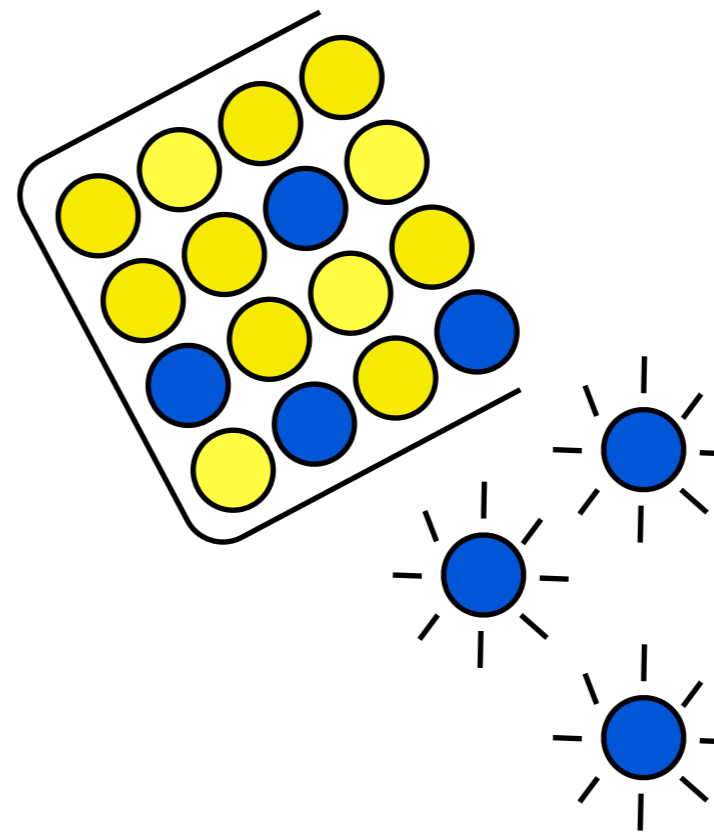
# Infants' use of sampling assumptions

---

This is all consistent with the size principle - but what if data is sampled differently?

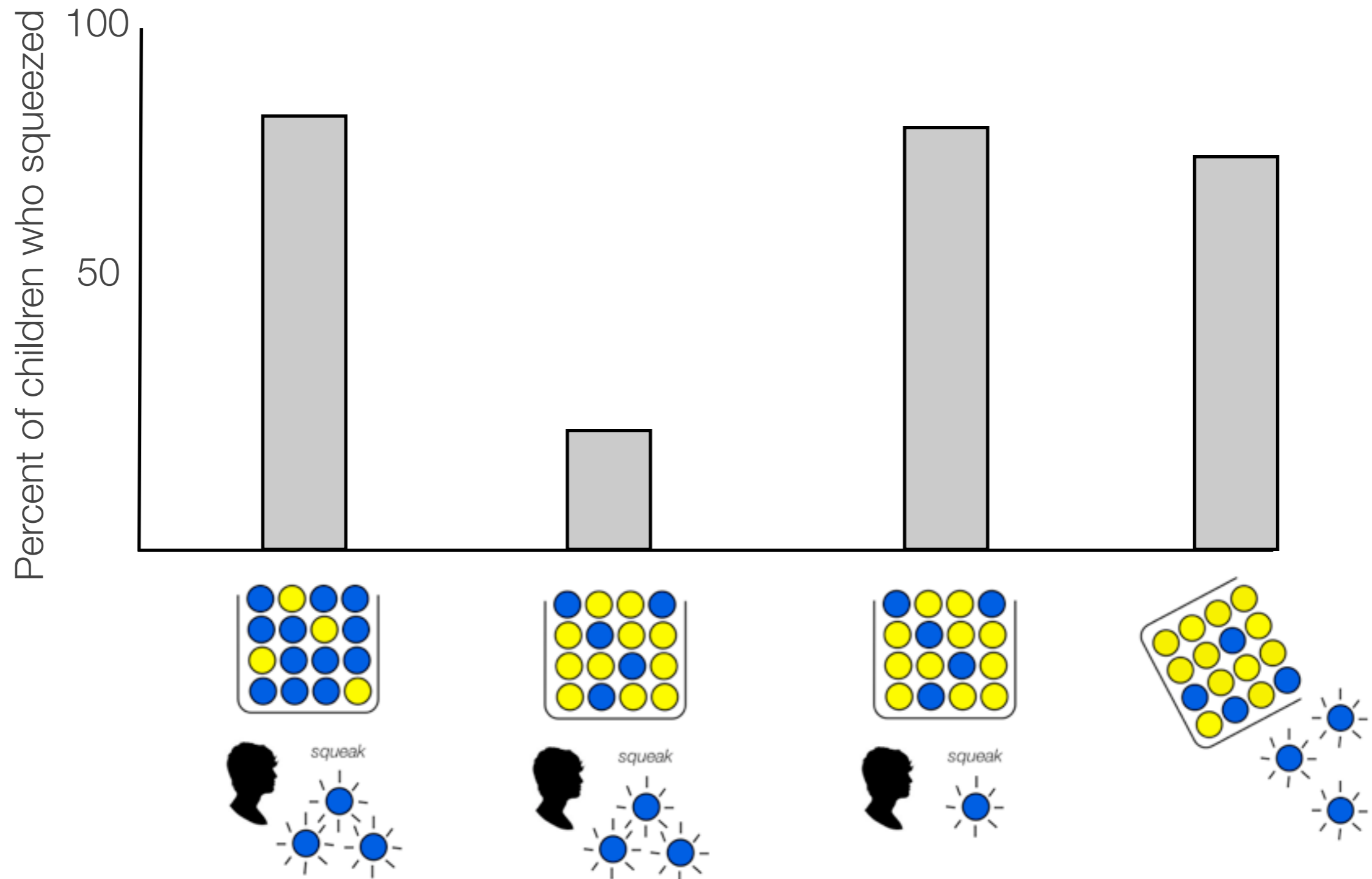


Experiment 4



If infants notice how the data were generated, they should not take this as an indication that yellow balls are not squeaky

# Infants' use of sampling assumptions





# Testing the size principle

---

So far all of this evidence has shown that people (including children) will tighten their generalisations more if they think the examples were generated from the concept/hypothesis directly.

This supports the qualitative ideas, but not necessarily the quantitative ones: people are sensitive to sampling assumptions, but do they tighten their generalisations *as much as* the size principle would predict? Are there individual differences in this?

# Testing the size principle

---

- ▶ We can capture the degree to which people assume that any point was strongly sampled (and the size principle should therefore apply)

$$P(x, x \in h | h, \theta) = \begin{cases} (1 - \theta) \frac{1}{|\mathcal{X}|} + \theta \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise,} \end{cases}$$

generalisation  
probability

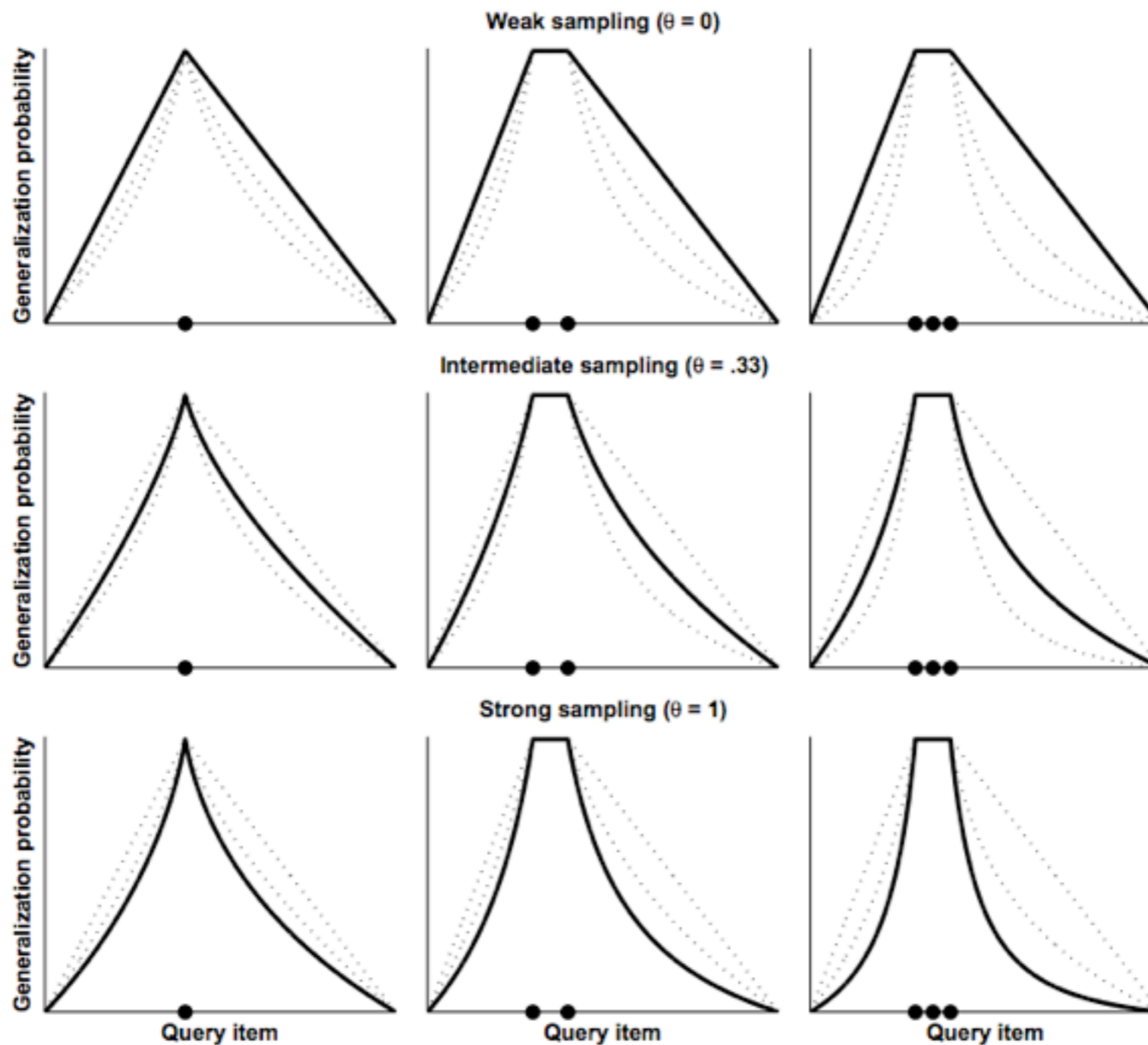
number of total  
possible items in  
the world

probability that  
any observation is  
strongly sampled

size  
principle

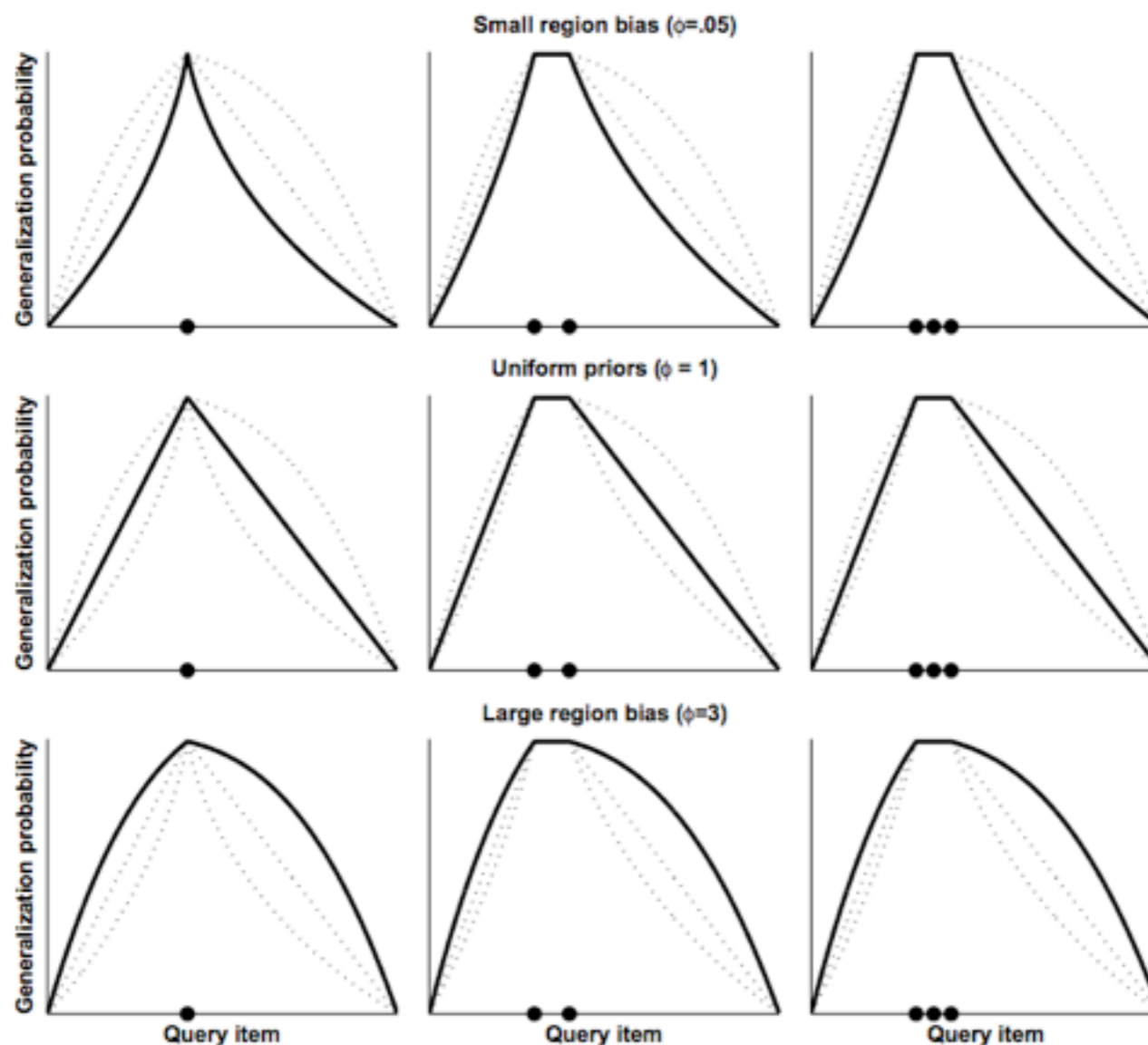
# Testing the size principle

- ▶ Higher  $\theta$  leads to tighter generalisations (model)



# Testing the size principle

- ▶ Note that this is different from a prior; the prior  $\Phi$  guides how large you think the region is,  $\theta$  is how much your generalisation tightens with additional data



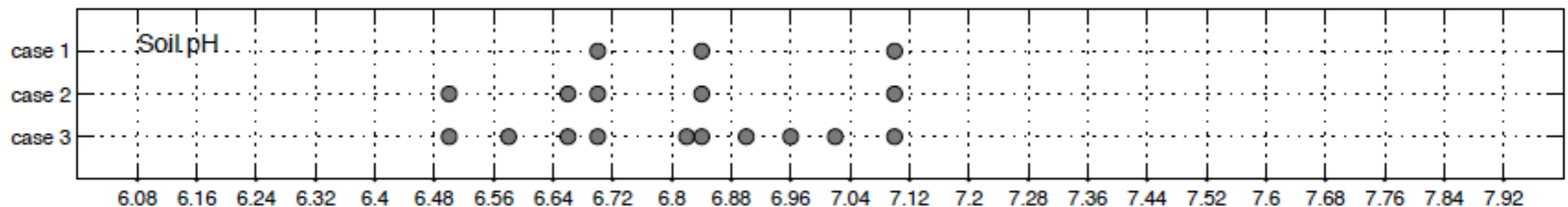
vary prior  $\Phi$ ,  
weak sampling  
( $\theta=0$ )

# Testing the size principle

- ▶ Task: give people data points that vary on a continuum, and look at how their generalisations change with additional data

The colour of the flowers of *Hydrangea macrophylla* change depending upon soil pH levels. Soils with a pH of less than 6 produce blue flowers, and soils with a pH greater than 8 produce pink flowers.

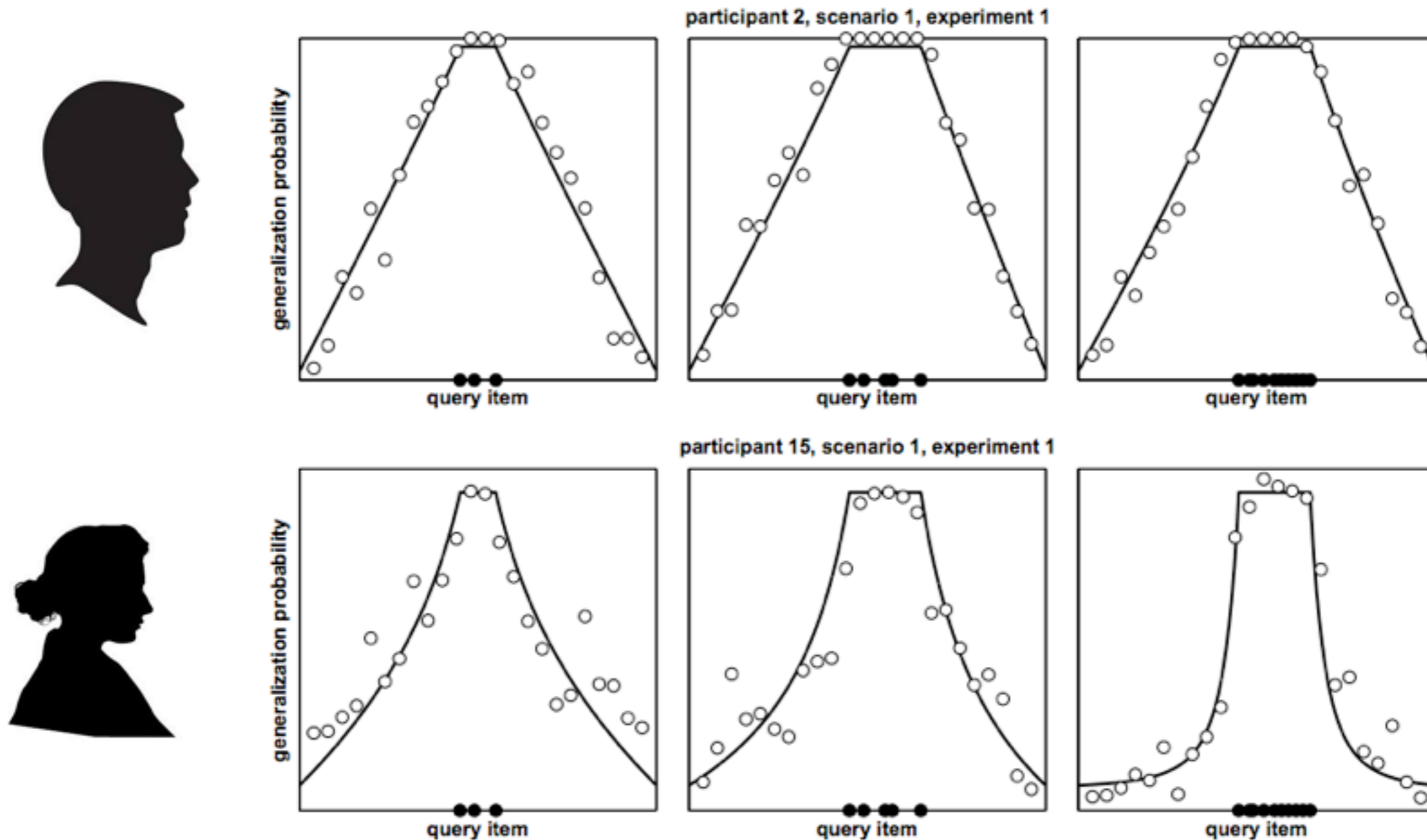
Neutral soils tend to produce very pale cream flowers. Given that cream flowers were produced by *Hydrangeas* growing in soils with the pH levels shown as black dots below, what is the probability that *Hydrangeas* would also produce cream flowers if they were grown in soil with the pH level specified by the red question mark?



Three tasks: bandicoot foraging hours, bacteria temperatures, and flowers growing in soils of different pH

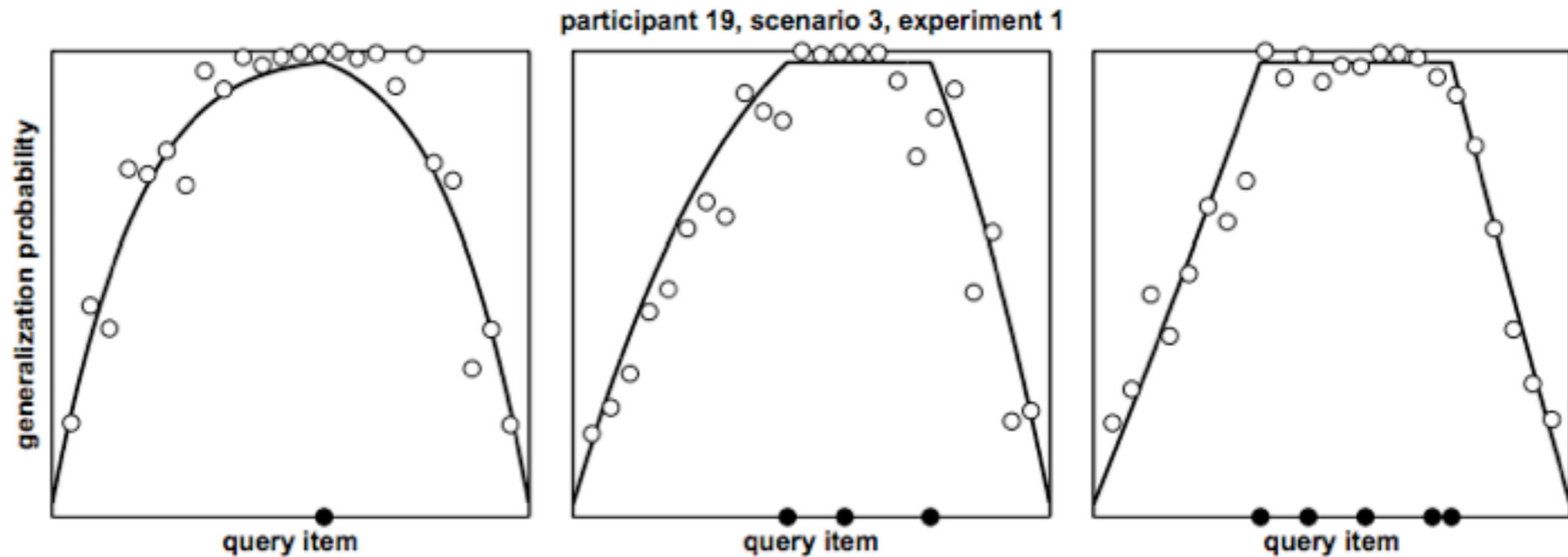
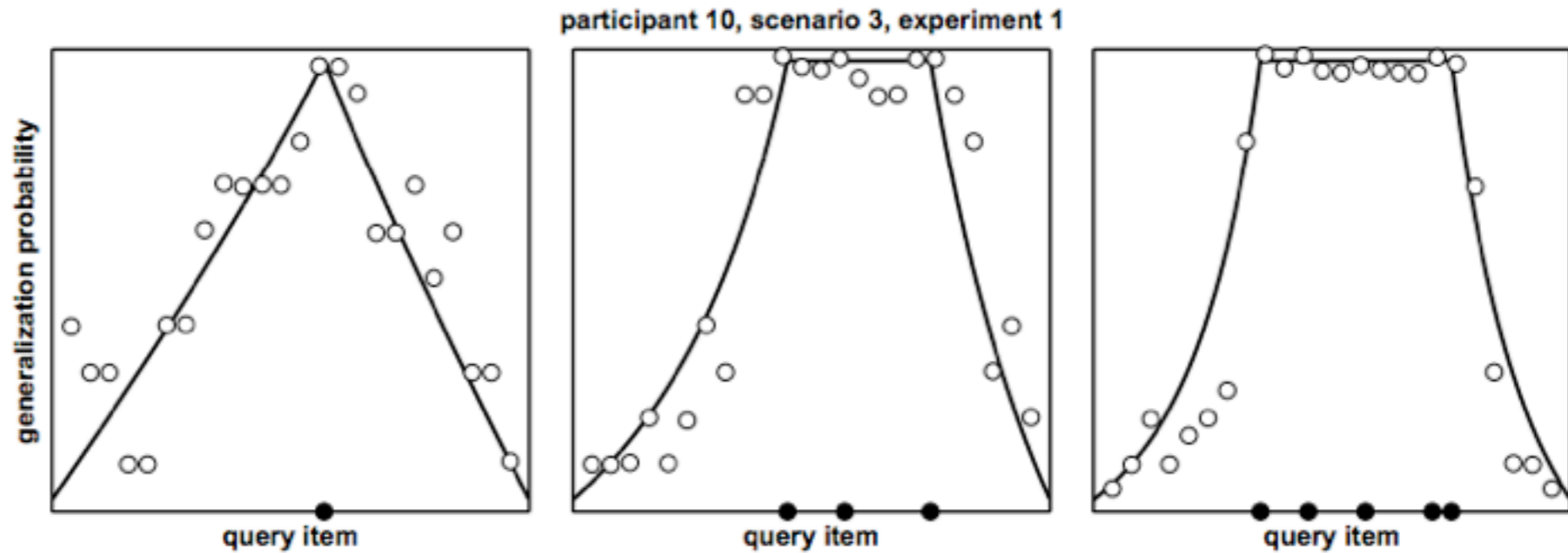
# Testing the size principle

- ▶ People appear to vary in the degree to which they assume strong sampling



# Testing the size principle

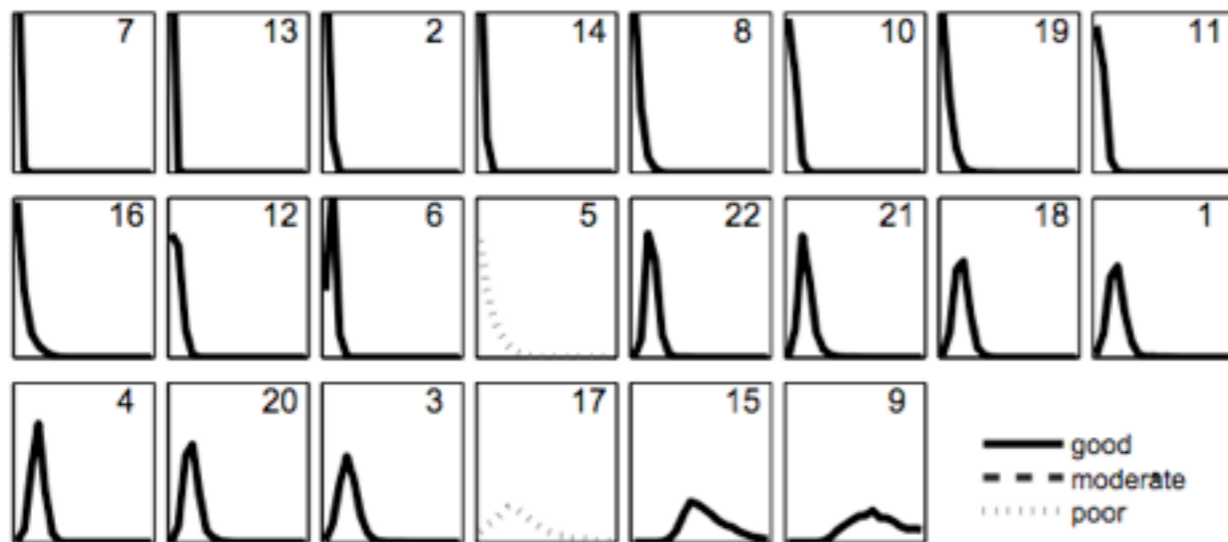
- ▶ They also vary in their priors



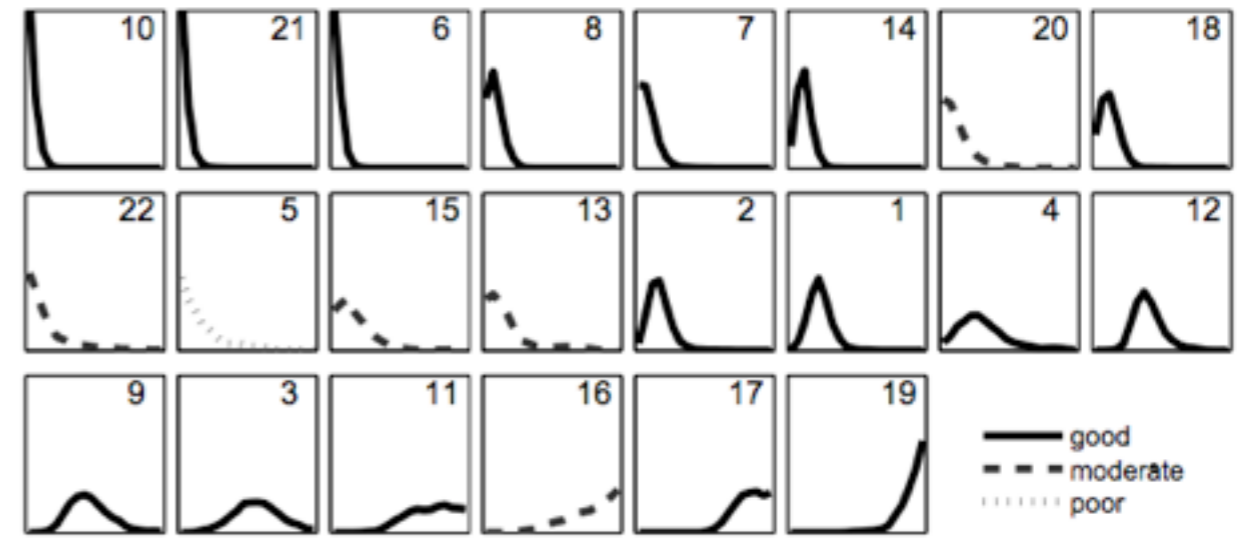
# Testing the size principle

- ▶ Individual differences in the degree of sampling assumptions

## Bacteria temperatures

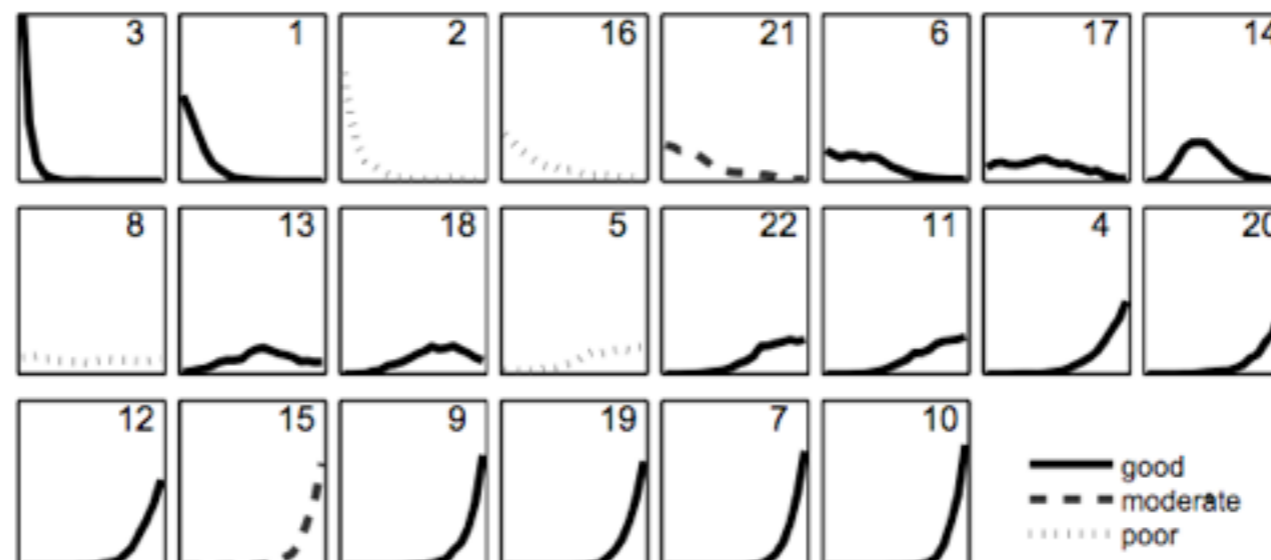


## Soil pH



varying  $\theta$   
(0 to 1)

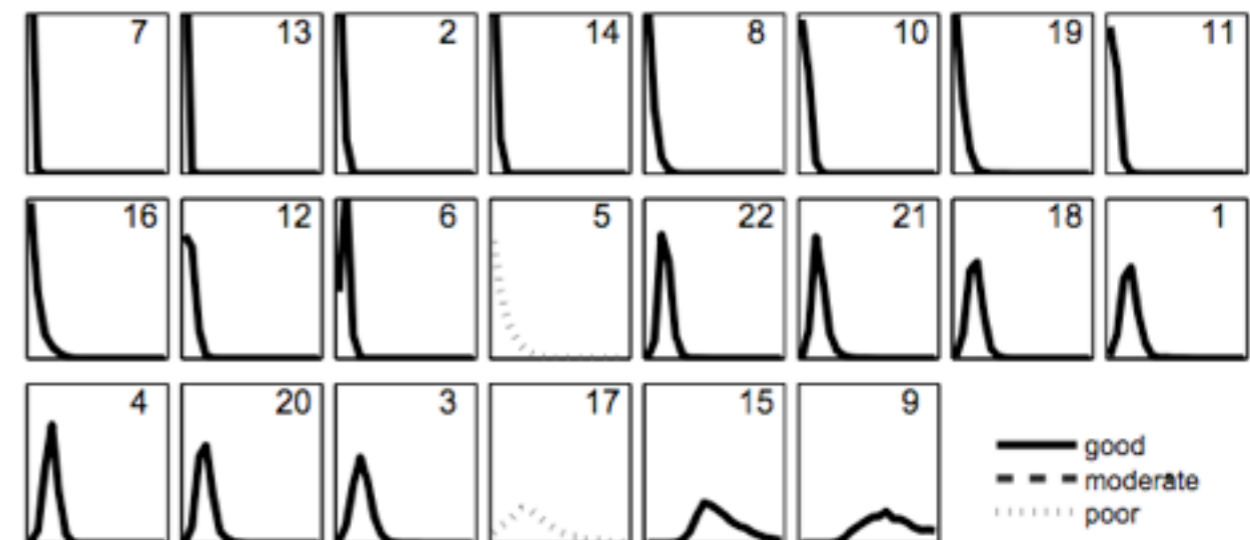
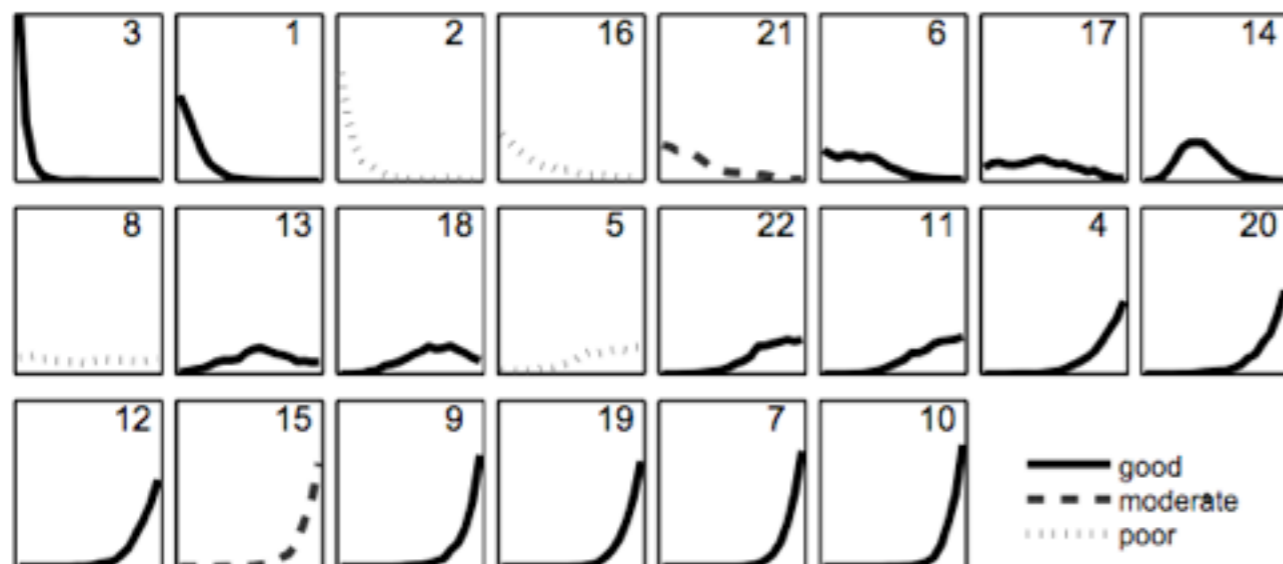
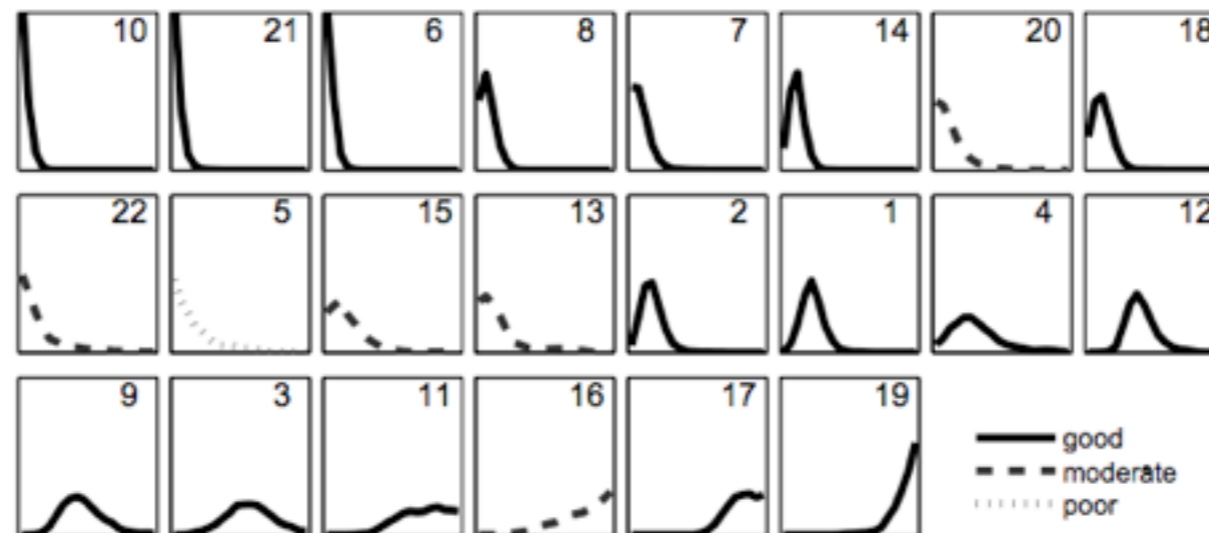
## Bandicoot foraging





# Testing the size principle

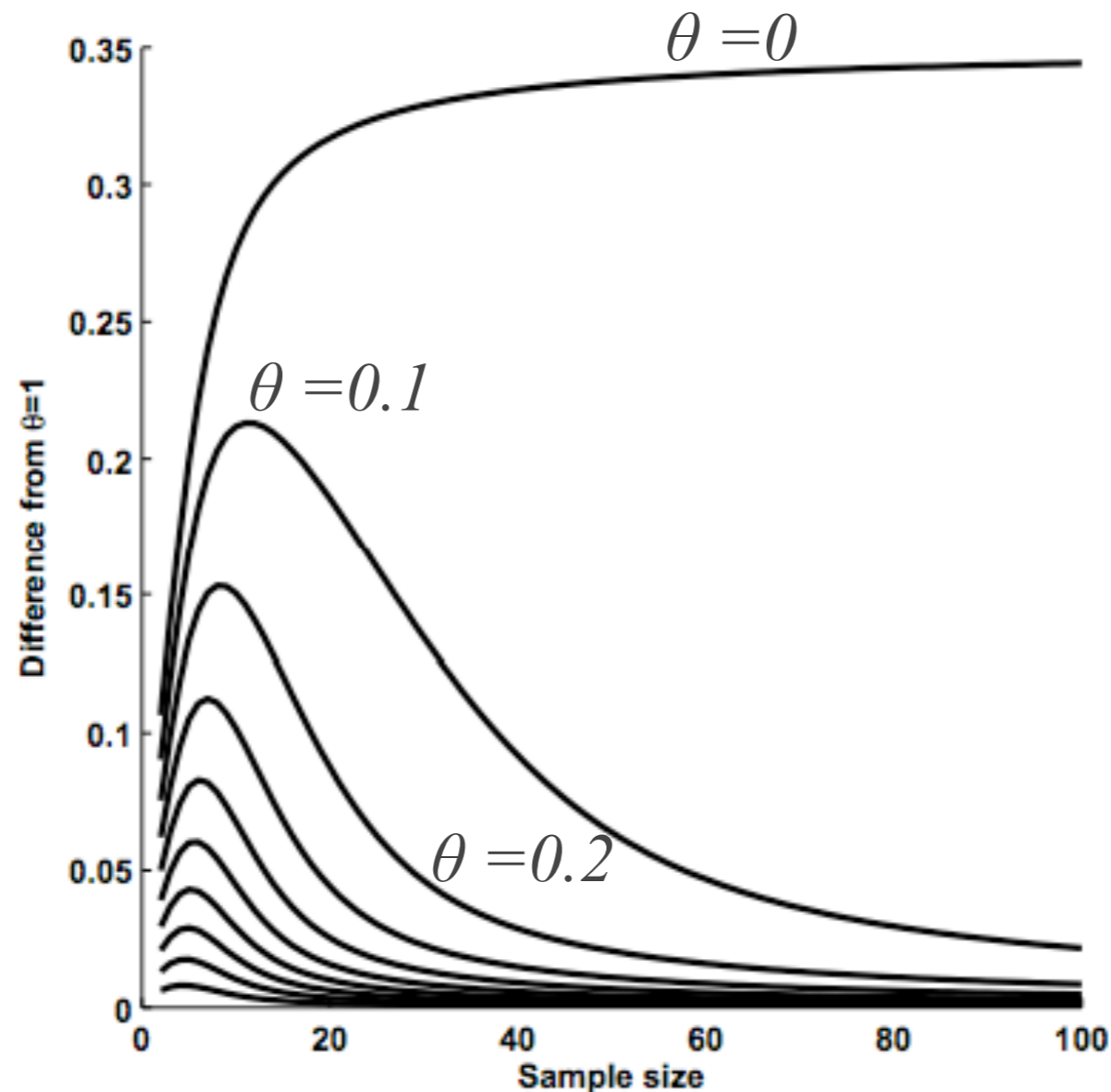
- ▶ Individual differences in the degree of sampling assumptions
- ▶ Note that there is high variance, but very very few don't tighten their generalisations at all (i.e., have  $\theta=0$ )



# Testing the size principle

---

- ▶ Interestingly, in the long run, *any*  $\theta$  greater than zero ends up looking the same: you get the same eventual tightening with enough data



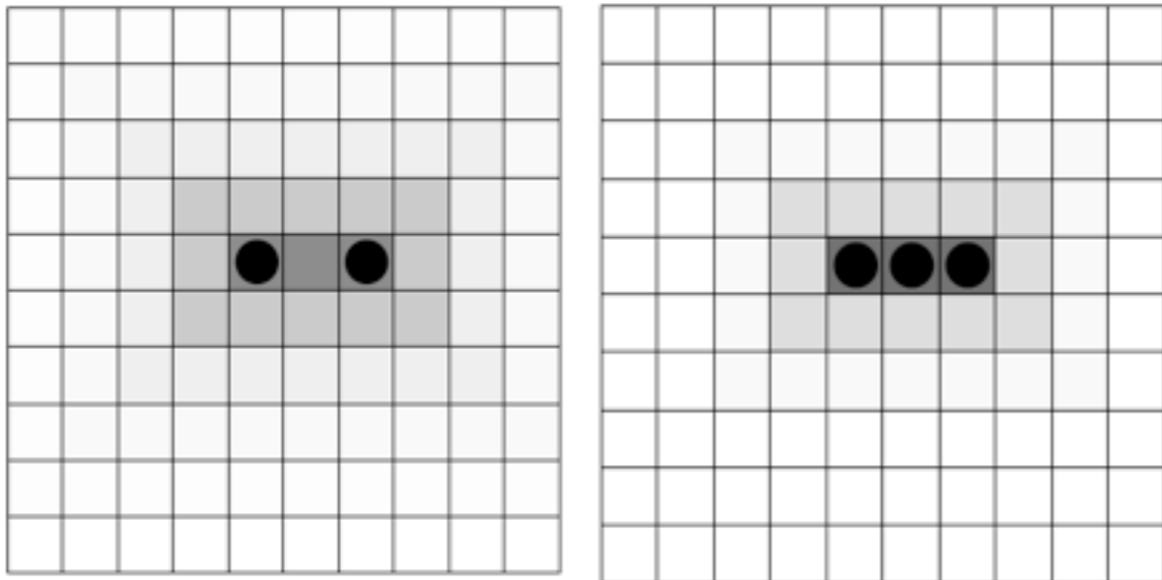
# Summary

---

- ▶ Difference between strong and weak sampling

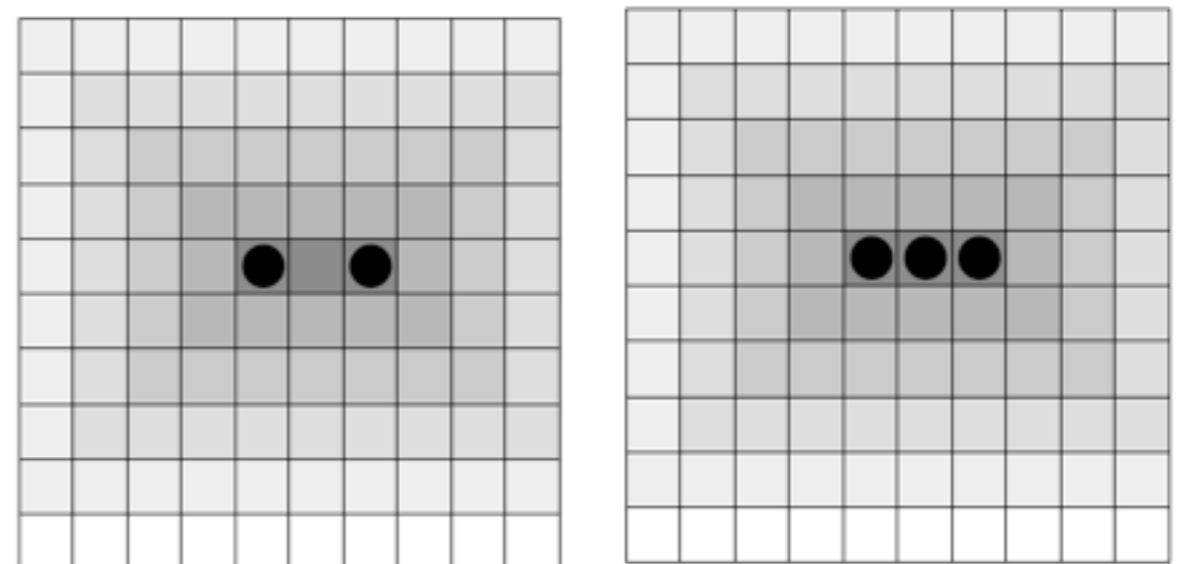
## Strong

- ▶ Items generated from concept
- ▶ Additional items lead to tighter generalisations



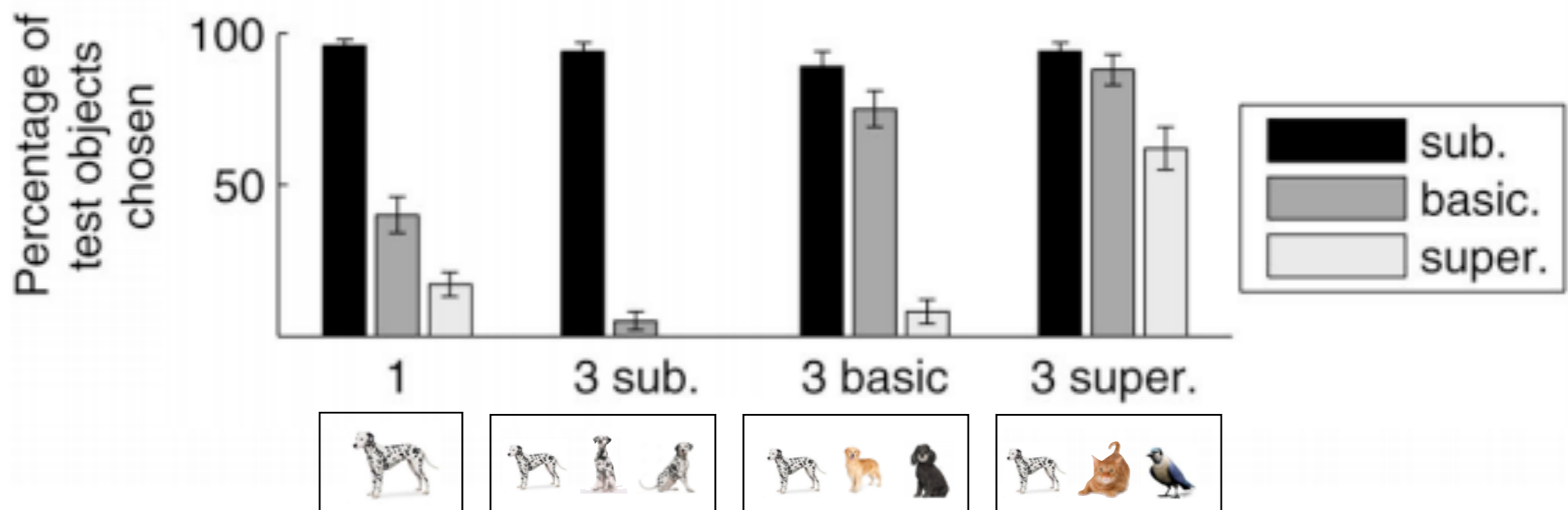
## Weak

- ▶ Items generated from world and then labelled
- ▶ Additional items do not lead to tighter generalisations



# Summary

- ▶ Difference between strong and weak sampling
- ▶ People pay attention to how data were sampled when figuring out how to generalise words

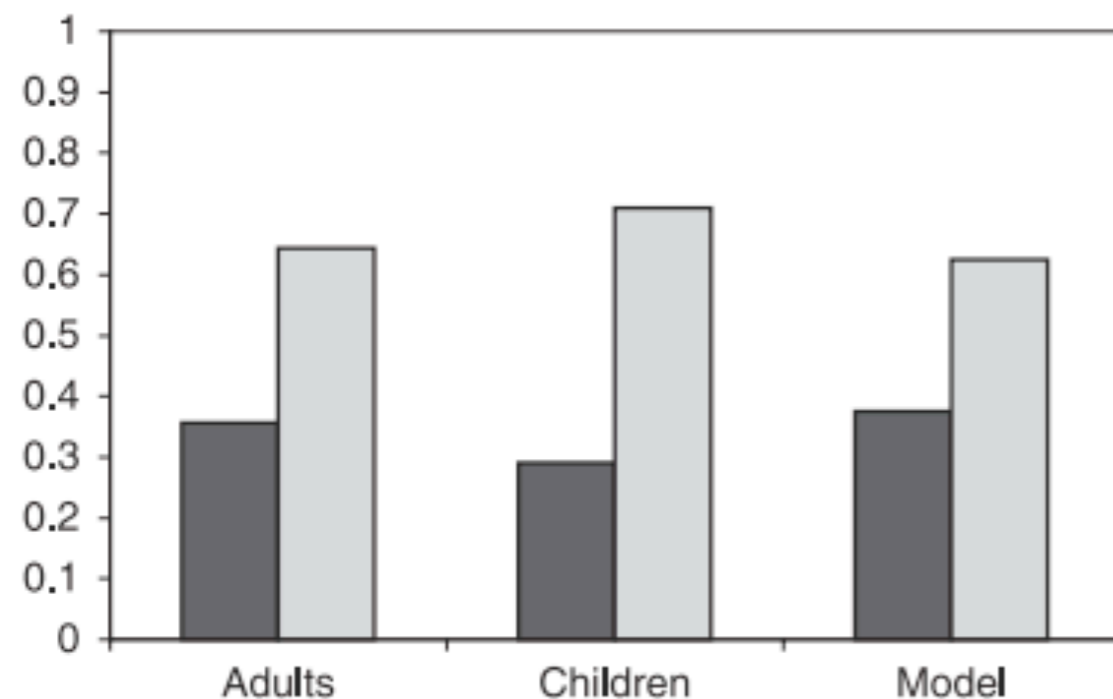


# Summary

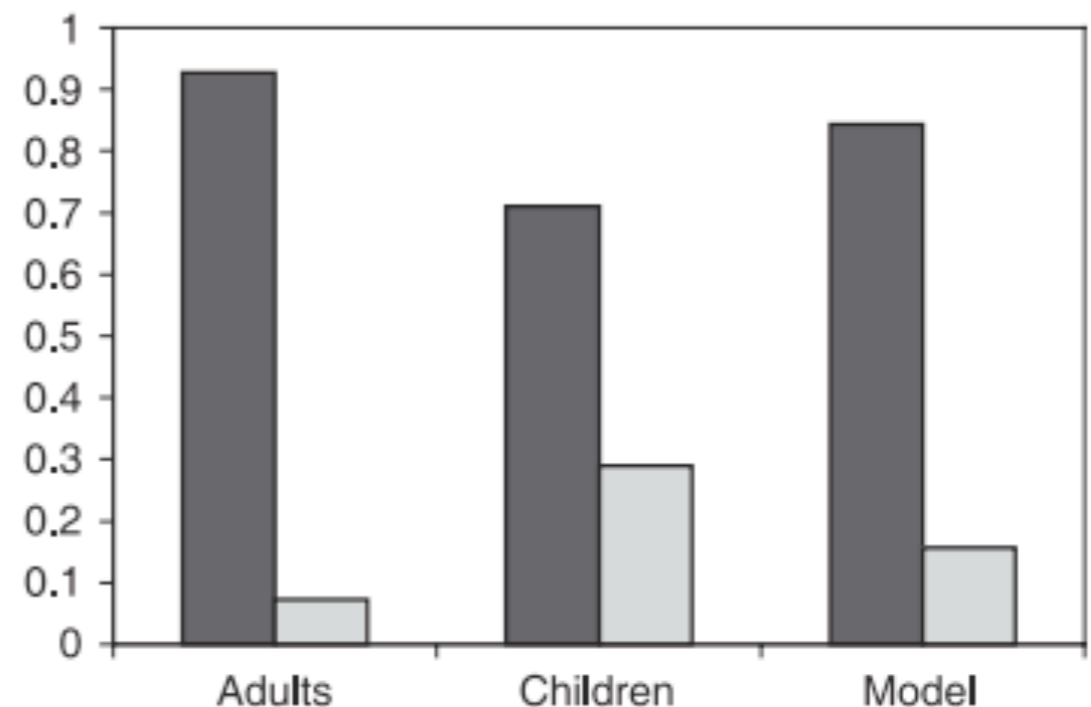
---

- ▶ Difference between strong and weak sampling
- ▶ People pay attention to how data were sampled when figuring out how to generalise words, changing their generalisations if they did not come from the concept

## Learner-driven



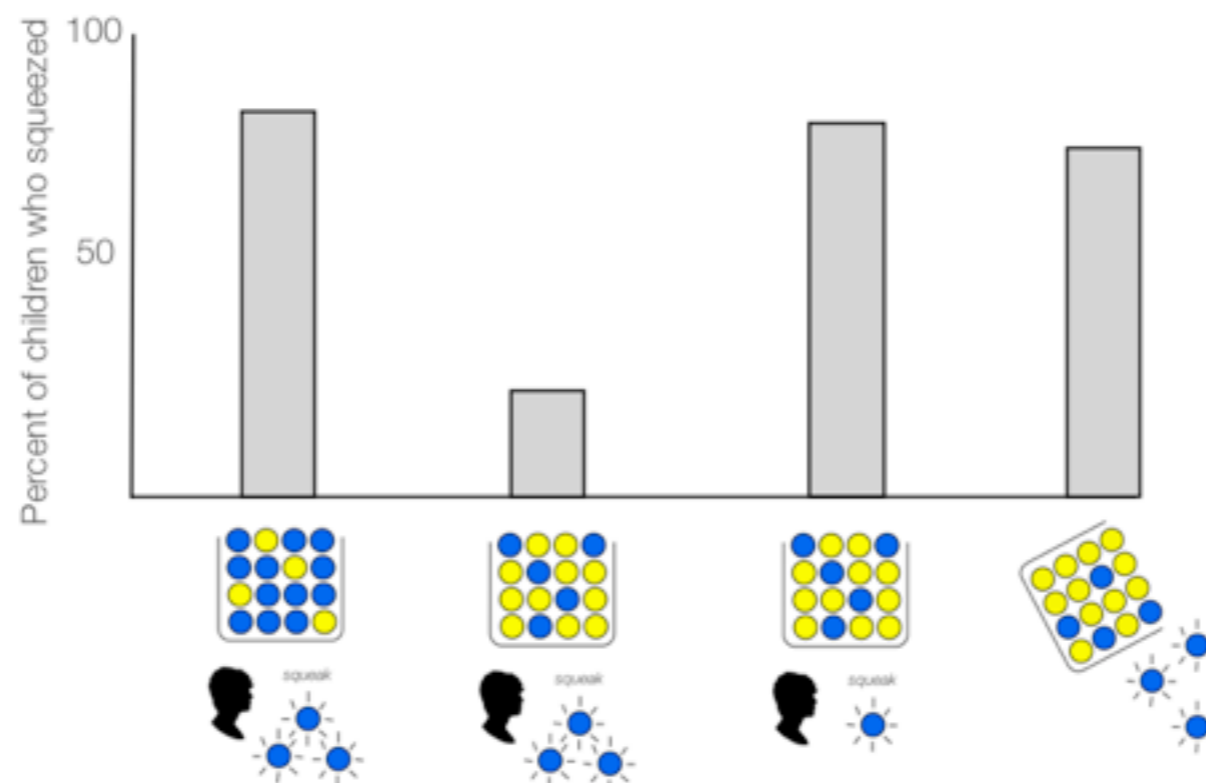
## Teacher-driven



# Summary

---

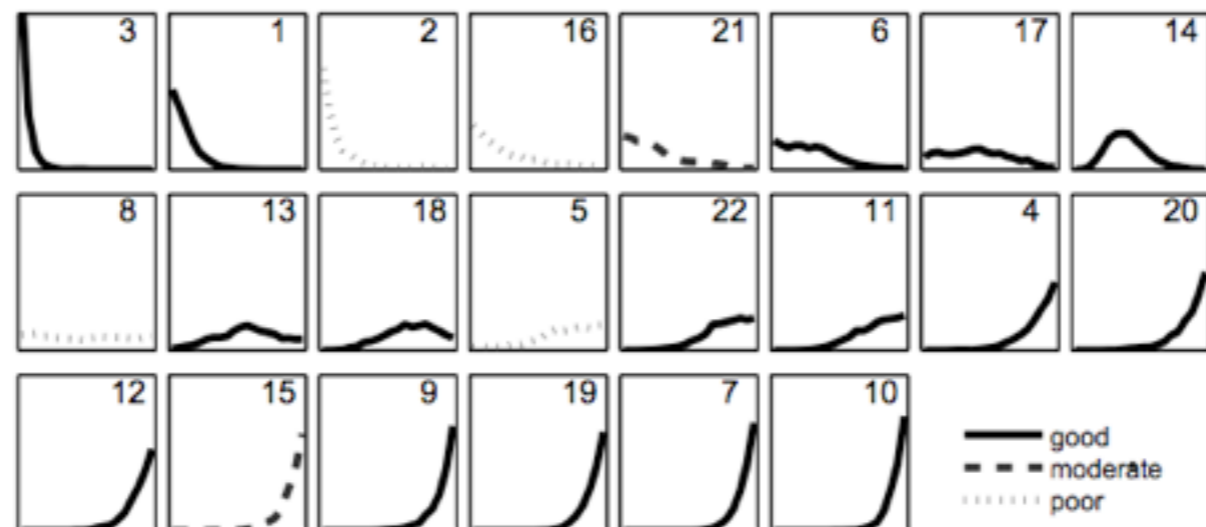
- ▶ Difference between strong and weak sampling
- ▶ People pay attention to how data were sampled when figuring out how to generalise words, changing their generalisations if they did not come from the concept
- ▶ Even infants do this, and with novel features



# Summary

---

- ▶ Difference between strong and weak sampling
- ▶ People pay attention to how data were sampled when figuring out how to generalise words, changing their generalisations if they did not come from the concept
- ▶ Even infants do this, and with novel features
- ▶ People show strong individual differences in the amount to which they assume strong sampling, but almost always they tighten at least somewhat



# Additional references (not required)

---

## Weak and strong sampling

- ▶ Gweon, H., Tenenbaum, J., & Schulz, L. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences* 107(20): 9066-9071
- ▶ Navarro, D., Dry, M., & Lee, M. (2012). Sampling assumptions in inductive generalization. *Cognitive Science* 36: 187-223.
- ▶ Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences* 24: 629-641.
- ▶ Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review* 114: 245-272.
- ▶ Xu, F., & Tenenbaum, J. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science* 10: 288-297.