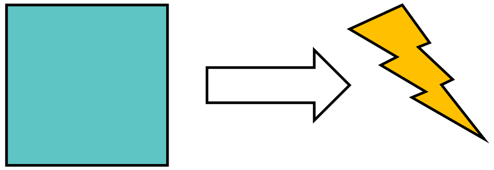
Pragmatic reasoning during associative learning: First attempt at a Bayesian computational model

Dani Navarro

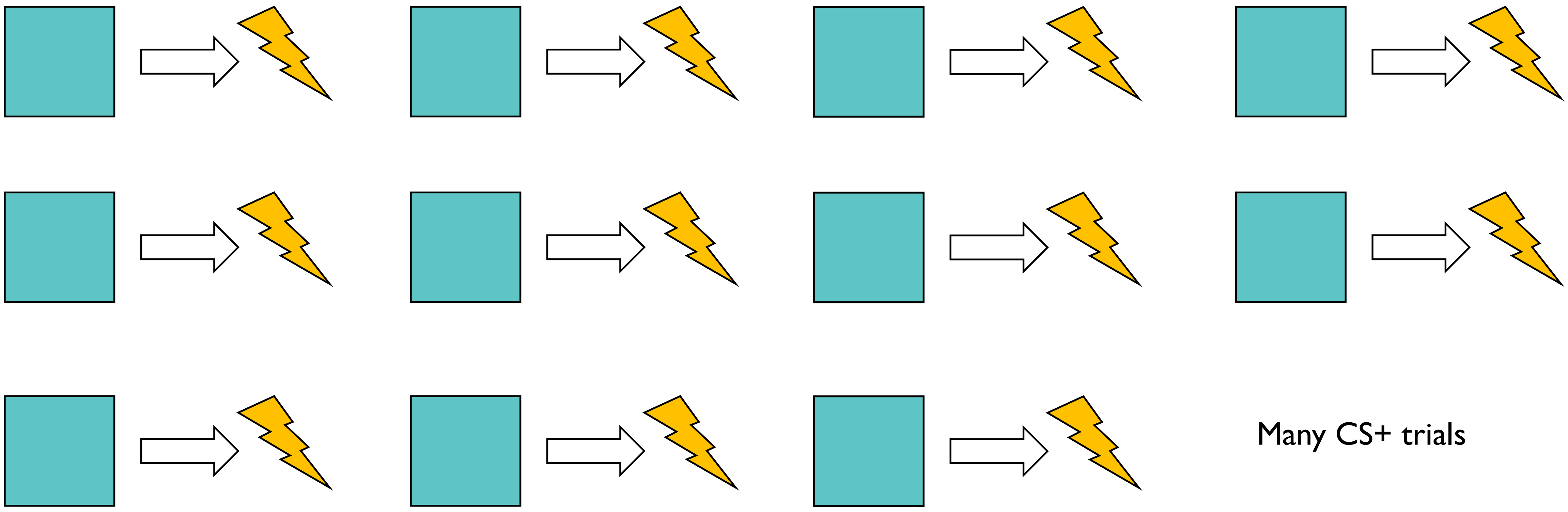
UNSW

The puzzle



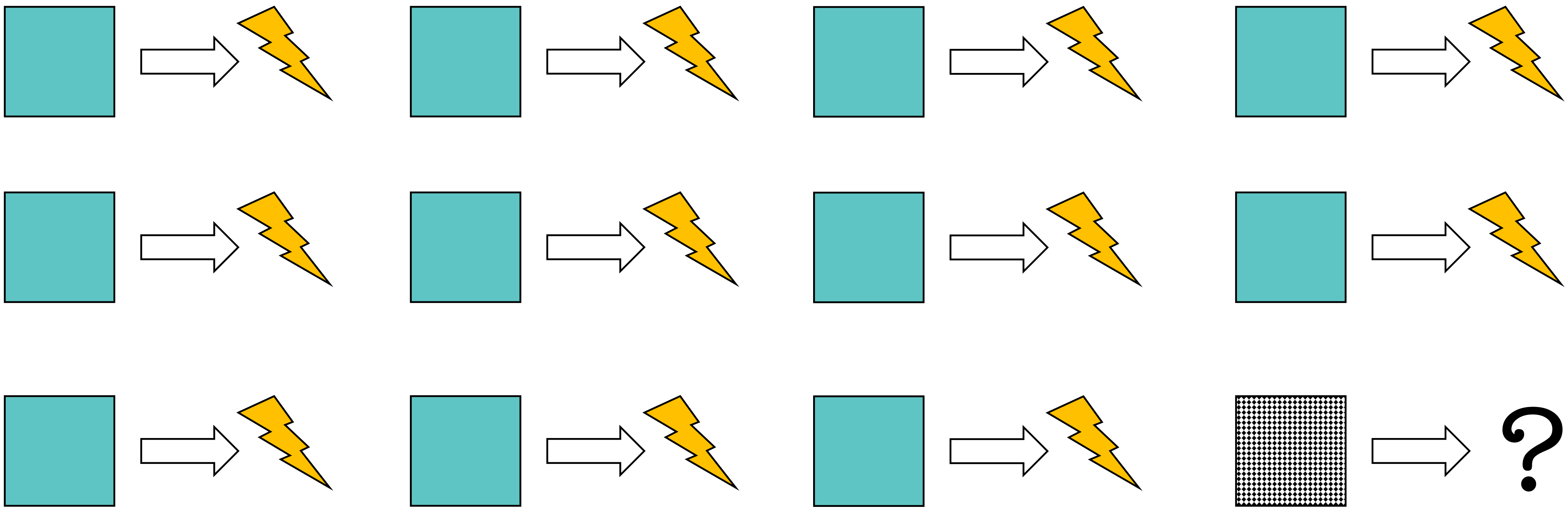
A CS+ trial

The puzzle



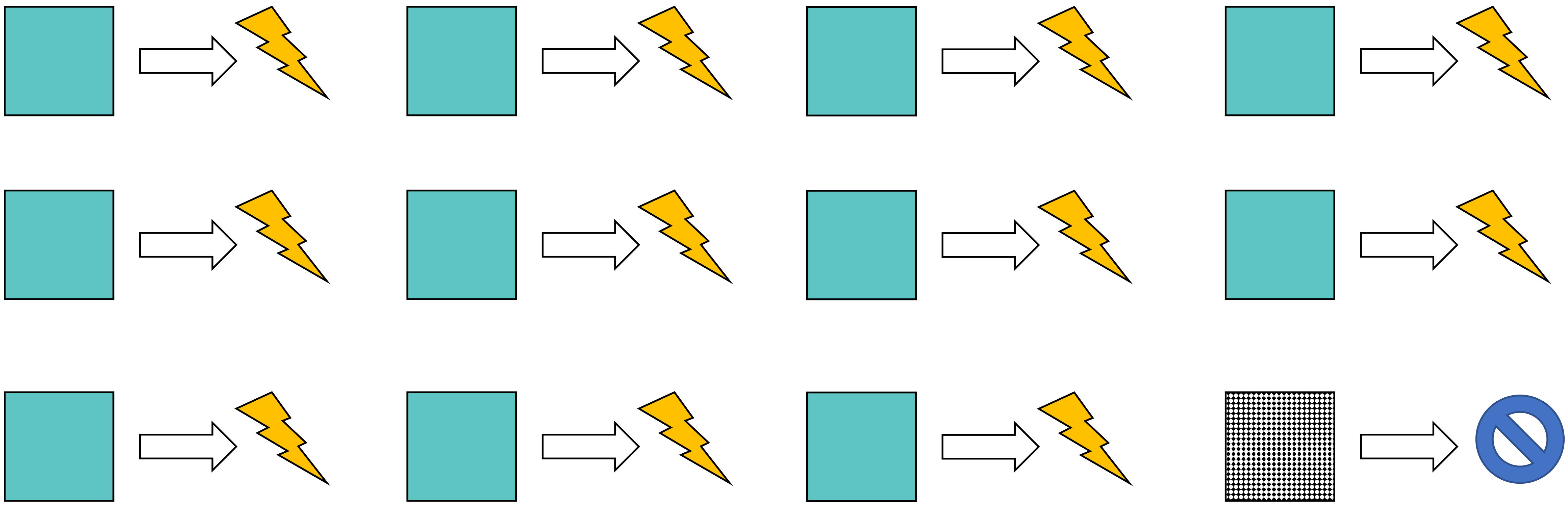
Many CS+ trials

The puzzle



Generalisation trial

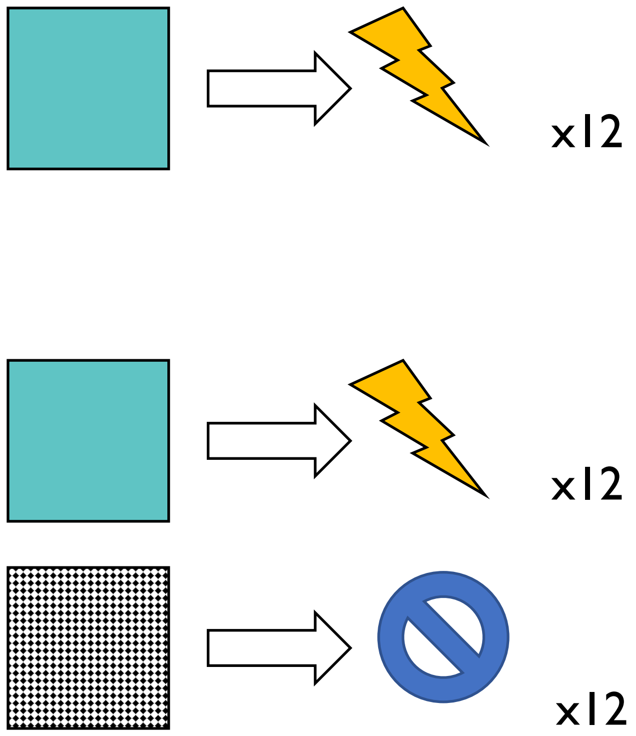
Utterly unsurprising… zero prediction error?



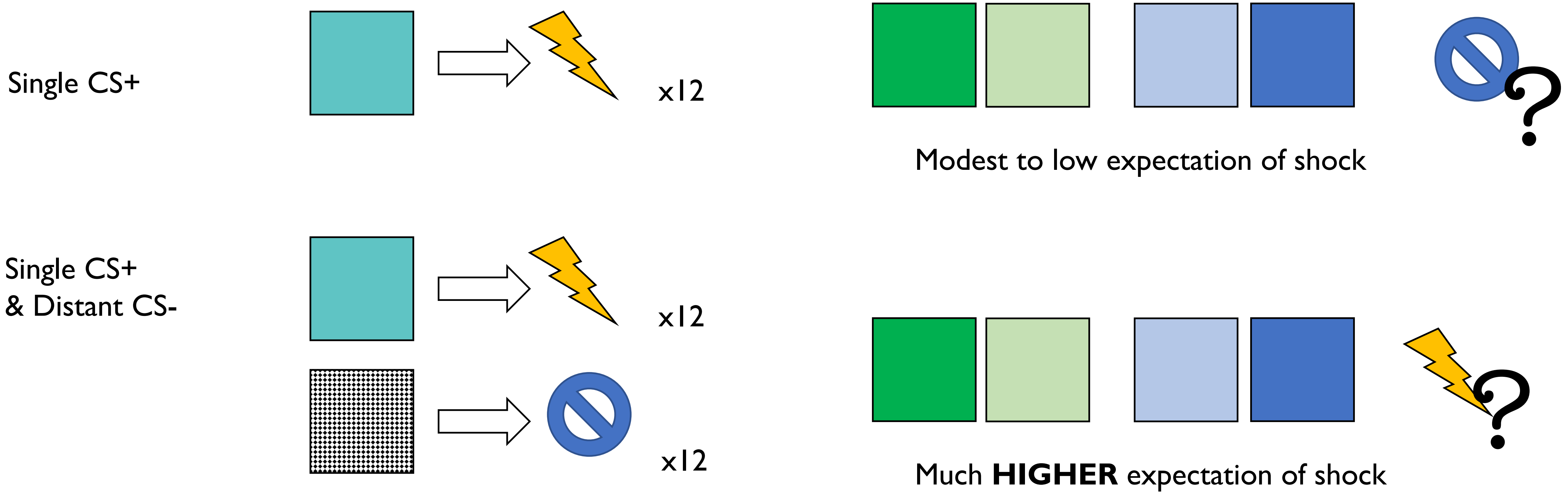
Add no-shock trials for a stimulus you'd never expect to produce shock anyway…

Single CS+

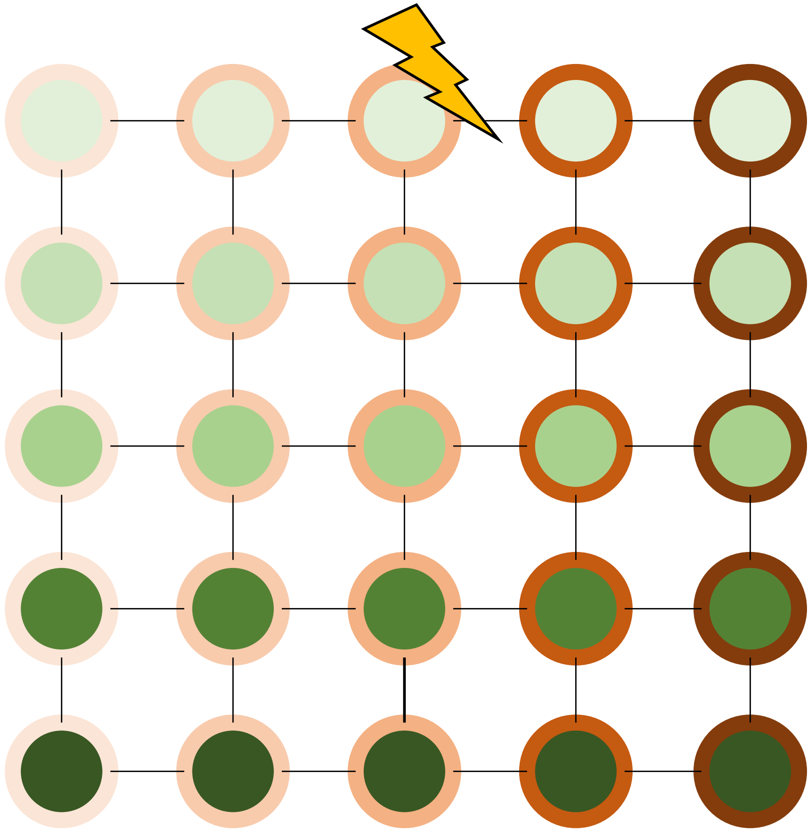
Single CS+
& Distant CS-

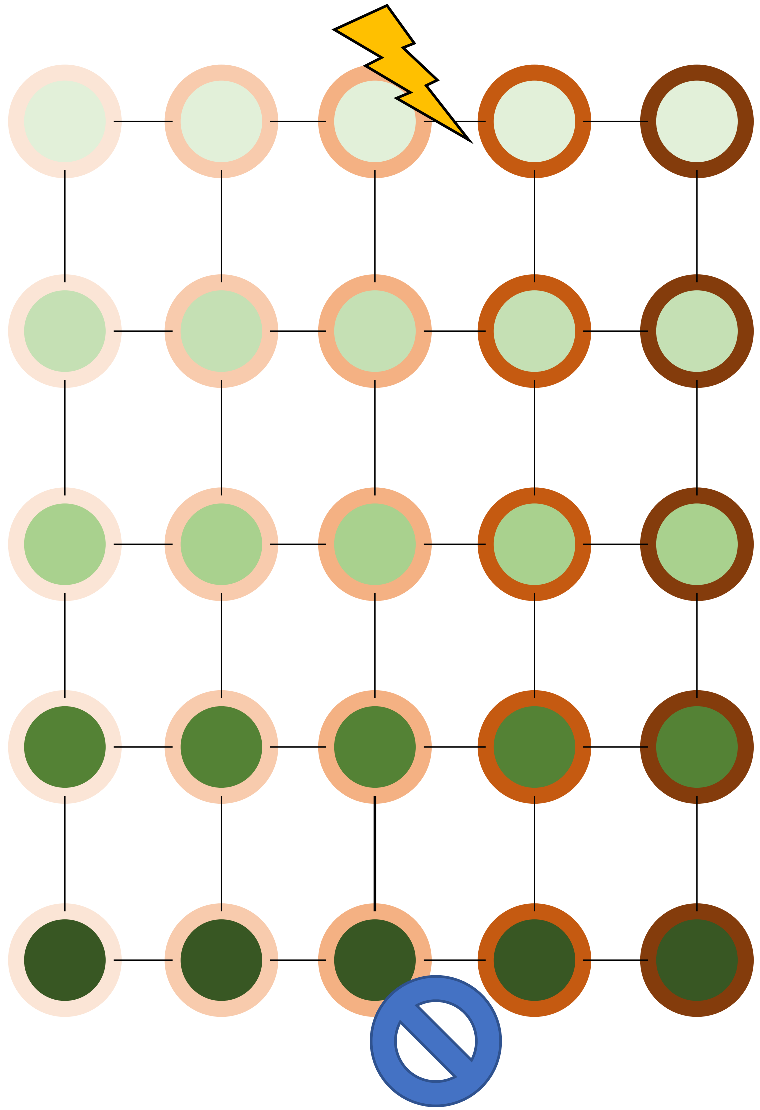


… and expectation of shock to ambiguous items <u>increases</u>???



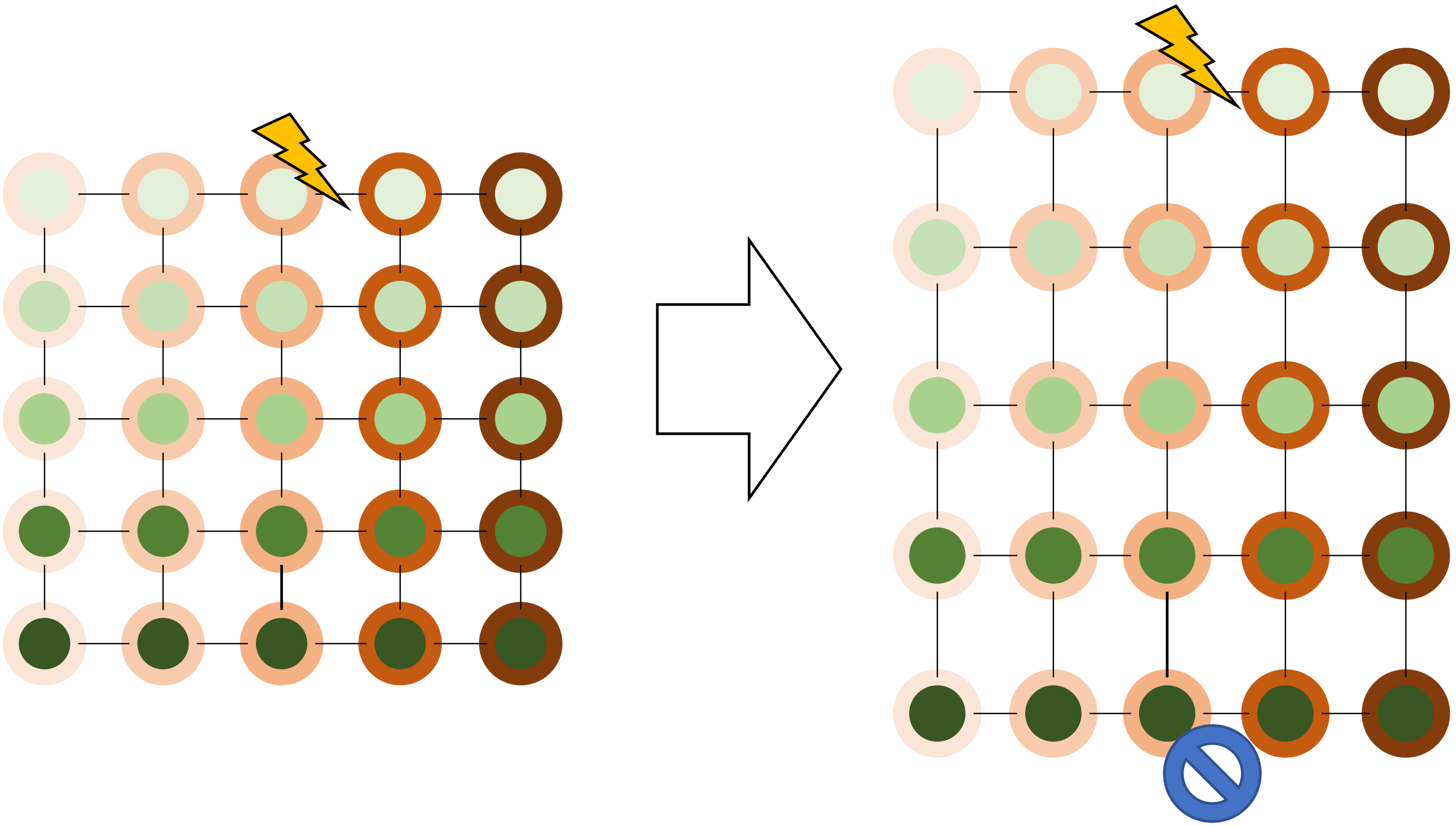
Dimensional attention?





Contraction along this dimension produces more generalisation

Still a puzzle though…



What is the “prediction error” that drives this change?

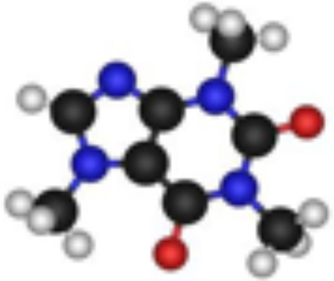
The perspective from the reasoning literature

(cue blatant reuse of slides from a different talk…)

What should we do with
this *sample* of evidence?

These birds have plaxium blood

The problem of inductive generalisation



?????

What factors shape our inductive inferences?

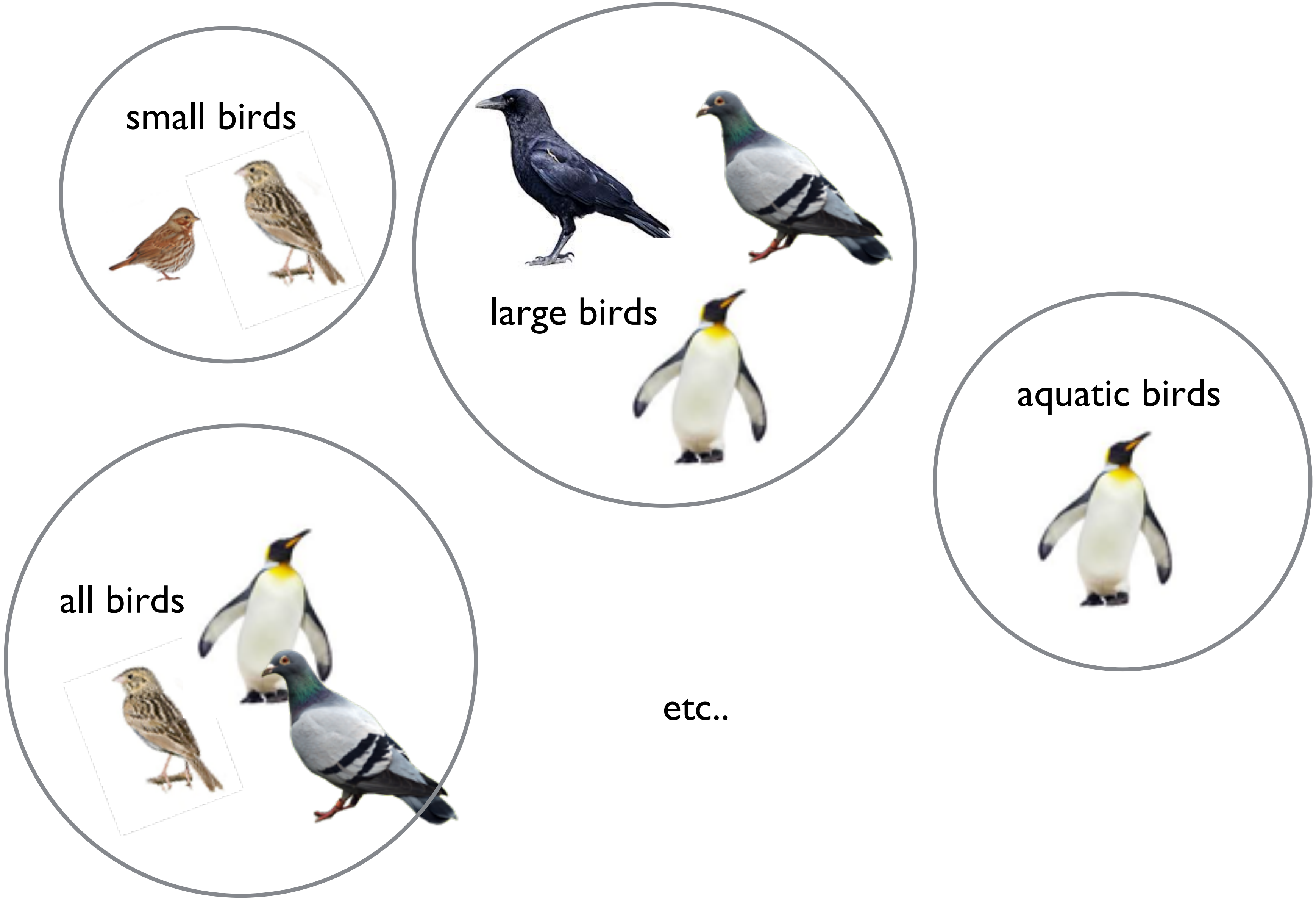
Similarity and typicality of the sample

What factors shape our inductive inferences?

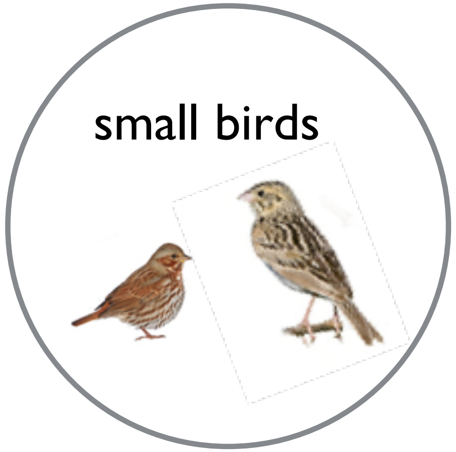
???

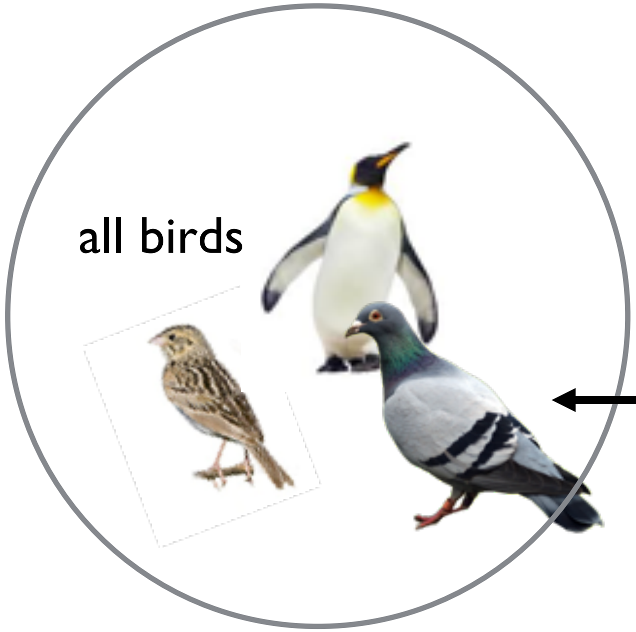
Size and diversity of the sample

Reasoners consider hypotheses



The sample rules out
some and not others…





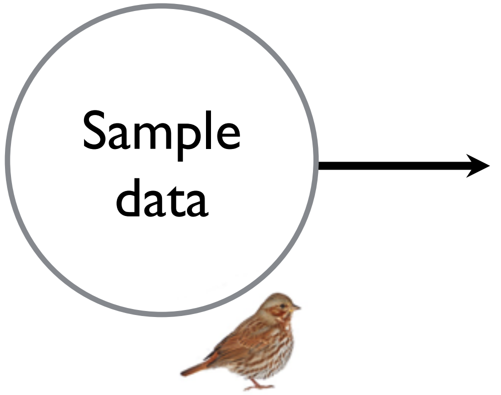
Inductive generalisation is based on hypotheses consistent with the sample

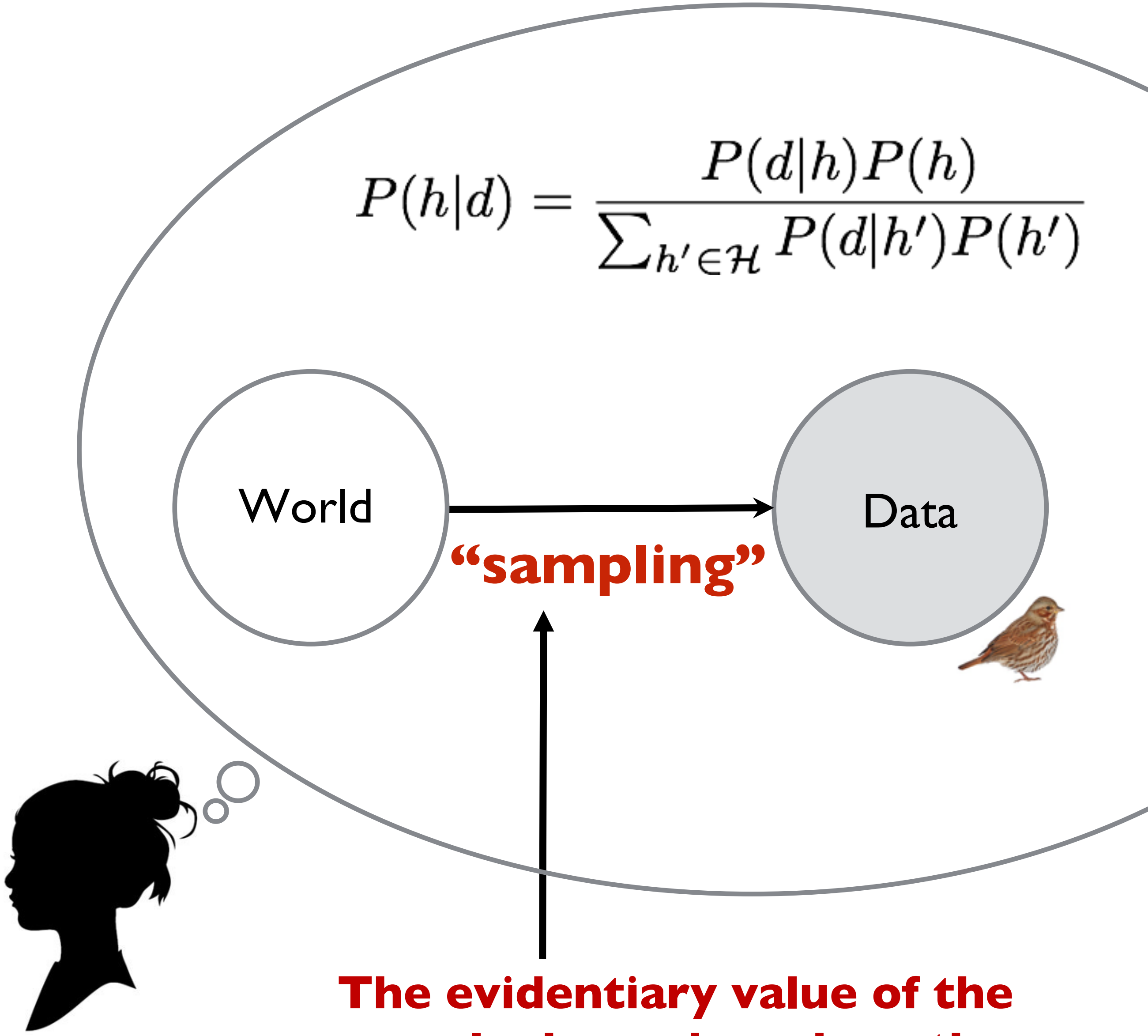
“learning”

Sample data → Belief about the world

Probabilistic perspective…

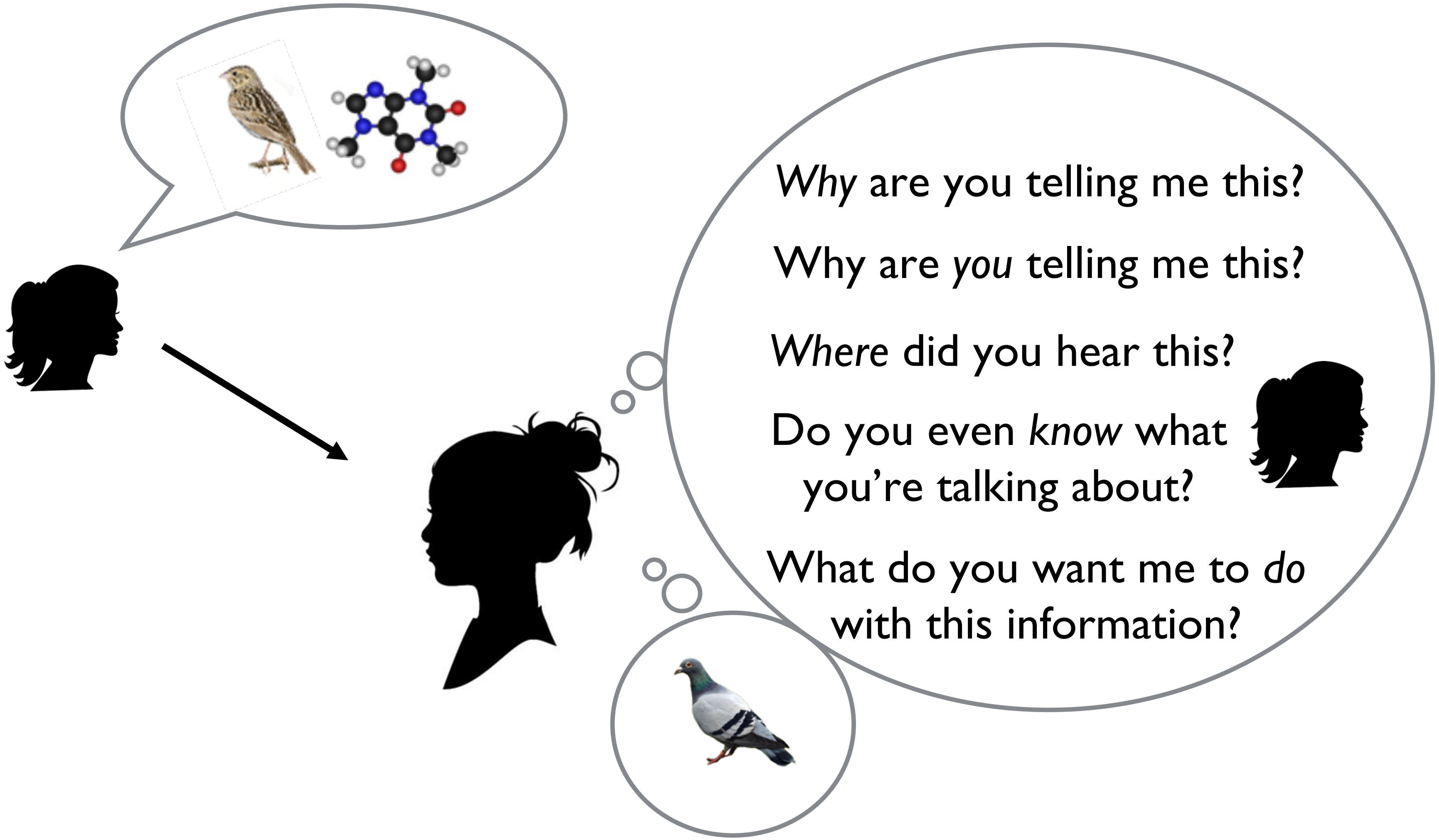
Learning depends on *sampling*





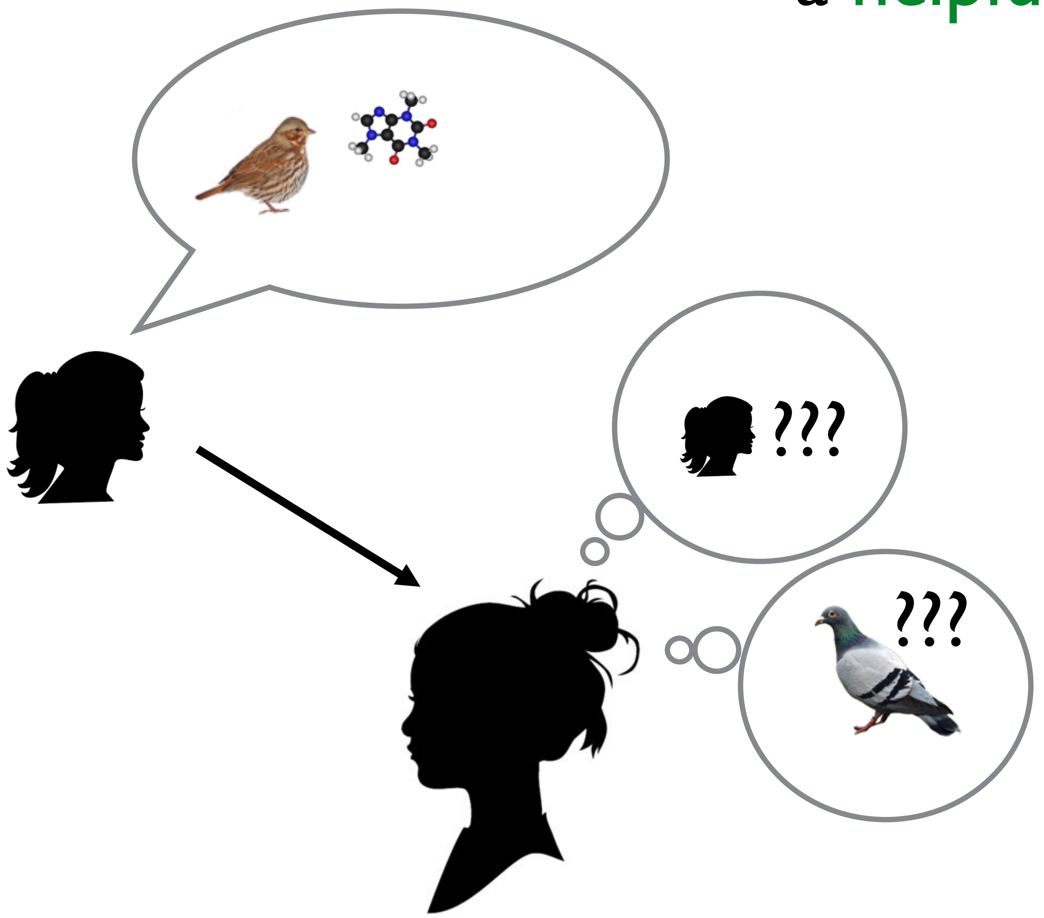
The evidentiary value of the sample depends on how the learner thinks it was generated, or how it came to their attention

Everyday reasoning *about the world* is intertwined with *social reasoning* about other people



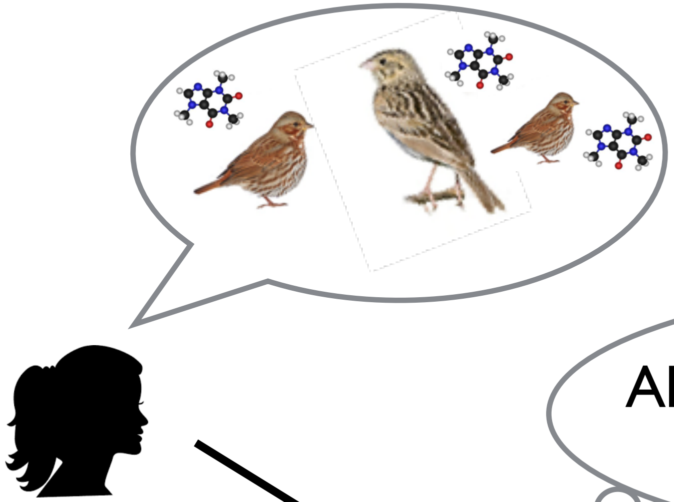
Illustrative example…

Inductive reasoning when a helpful teacher provides the data



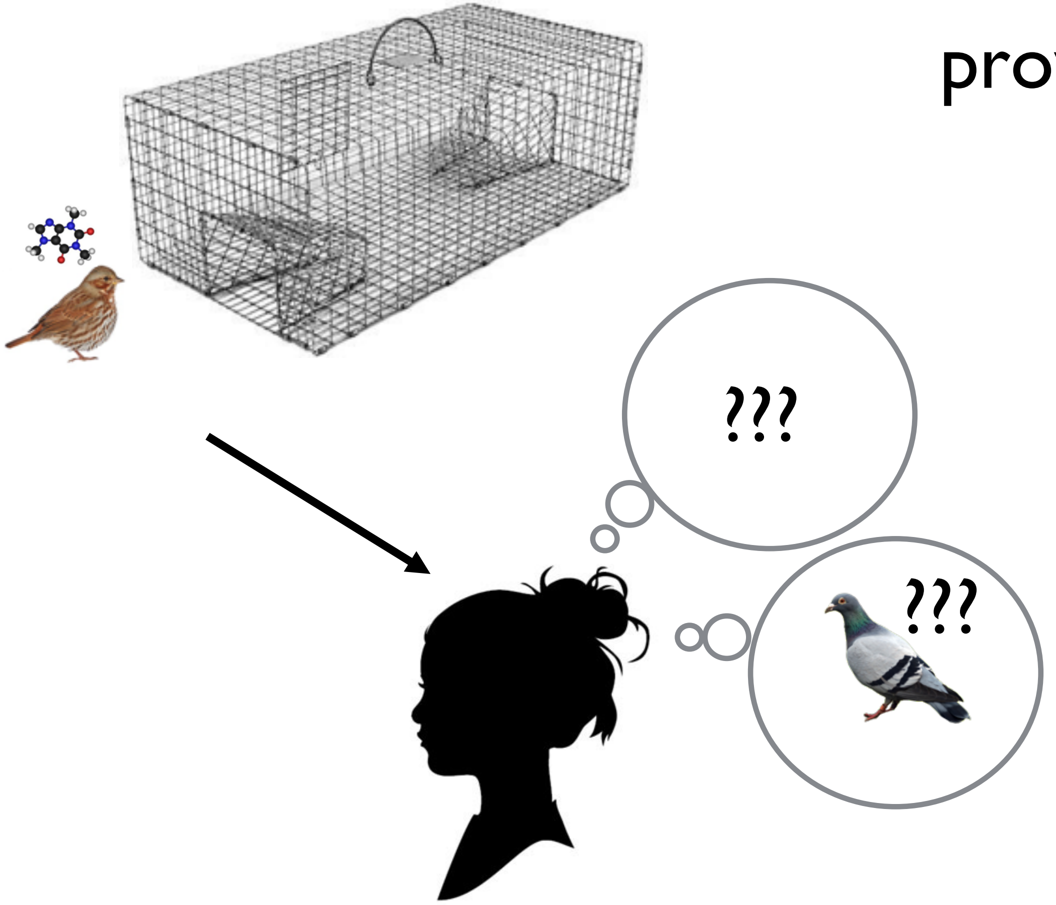
Illustrative example…

Inductive reasoning when a helpful teacher provides the data



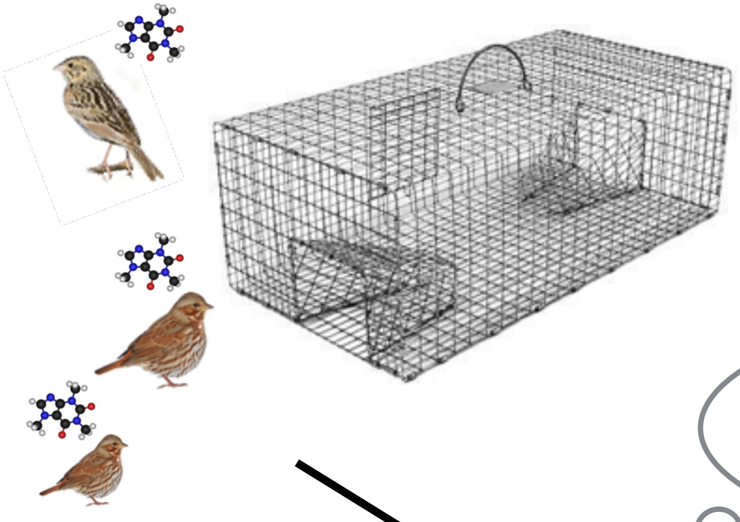
Illustrative example…

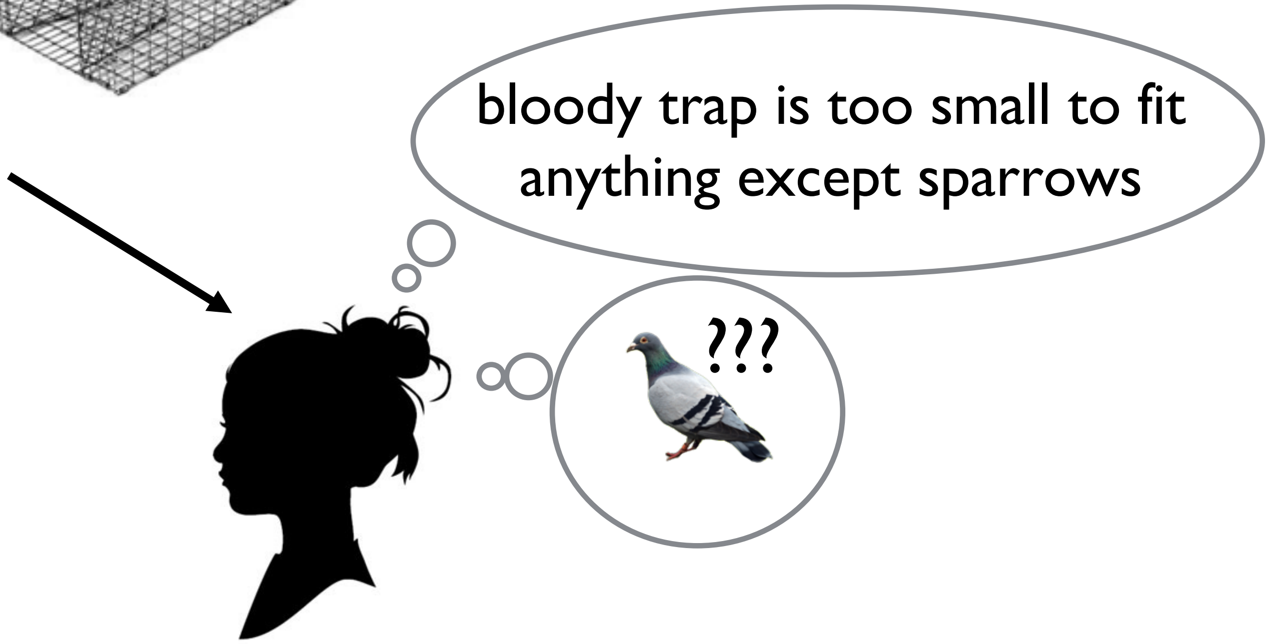
Inductive reasoning when an indifferent world provides the data



Illustrative example…

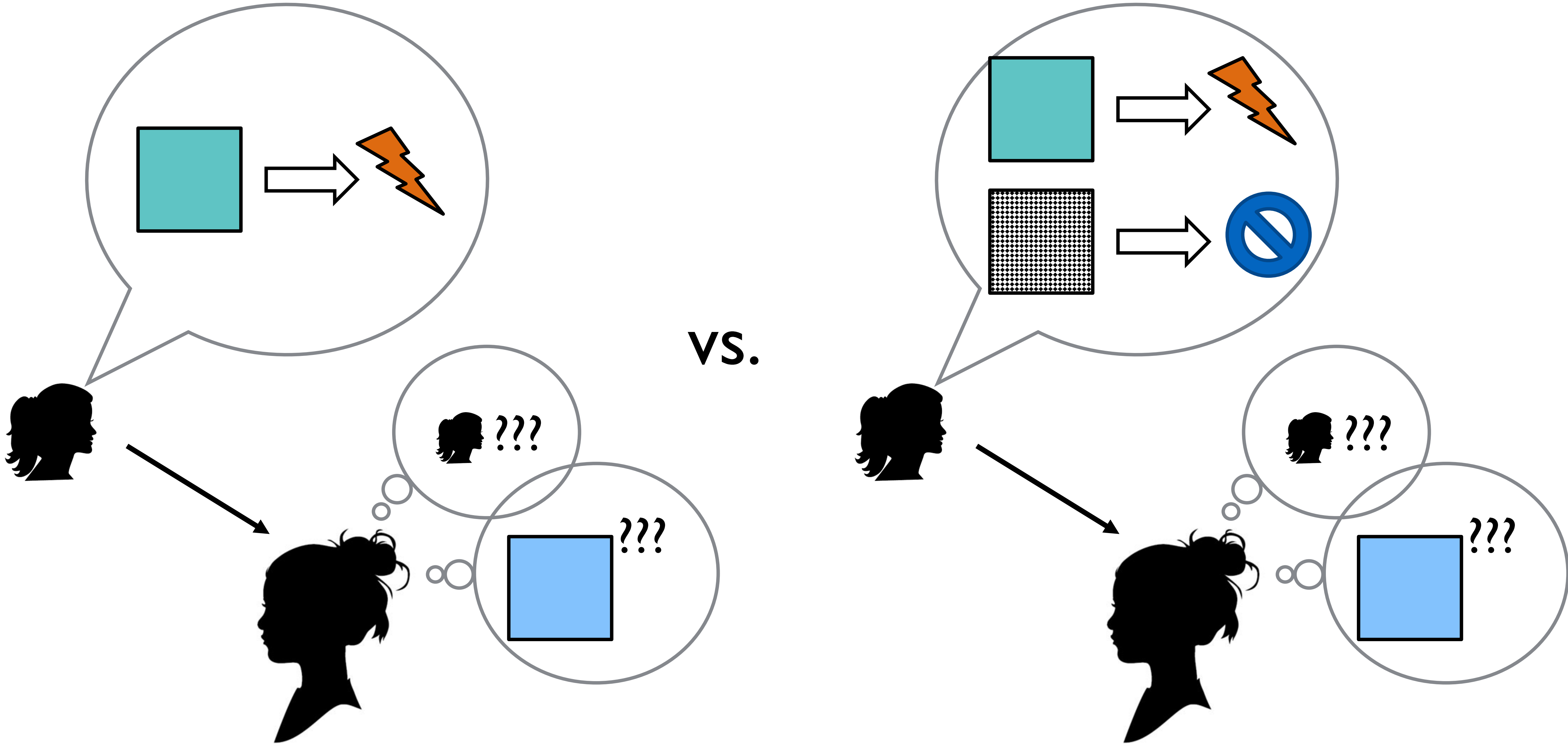
Inductive reasoning when
an indifferent world
provides the data





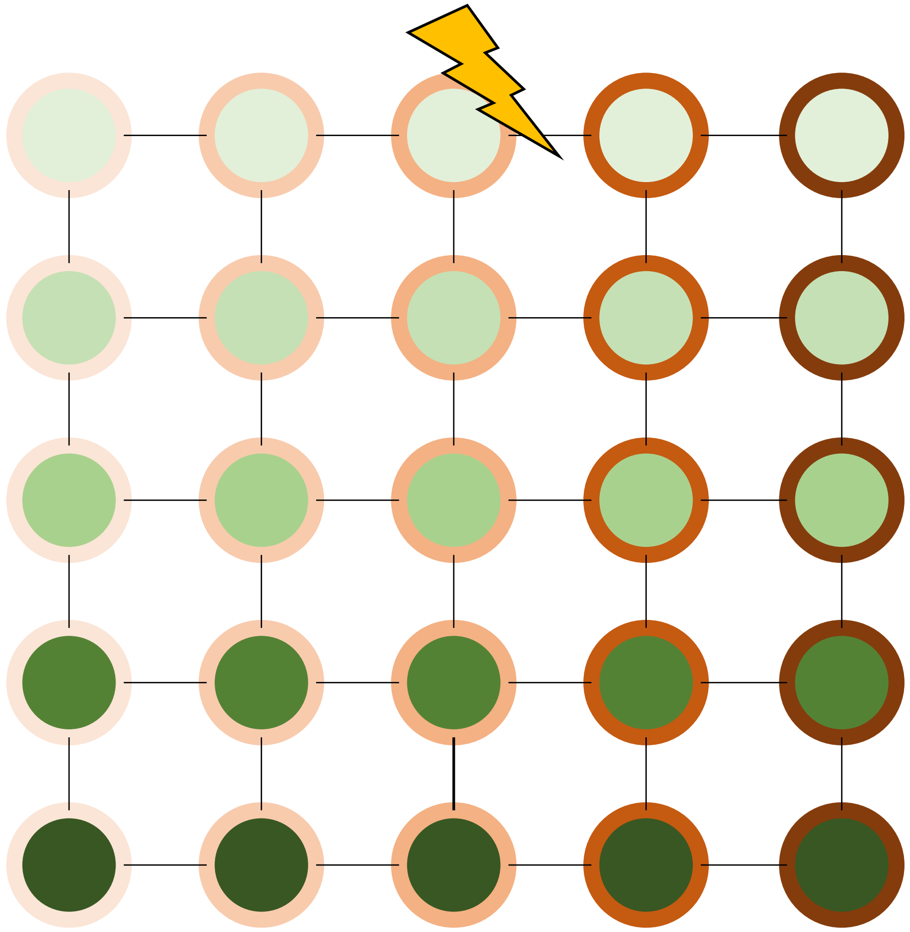
Some empirical examples:

- Ransom, Voorspoels, Perfors & Navarro (2017): the mere suspicion of deceptive informants shapes human (and Bayesian) reasoners
- Ransom, Perfors & Navarro (2016): the evidentiary status of stimulus similarity is different when a human chooses examples or not
- Voorspoels, Navarro, Perfors, Storms & Ransom (2015): ostensibly "irrelevant" negative evidence can be a powerful "hint"
- Hayes, Banner & Navarro (2017): purely <u>*mechanistic*</u> constraints on stimulus selection influence people's willingness to generalise
- Etc.


???
???
vs.
???
???

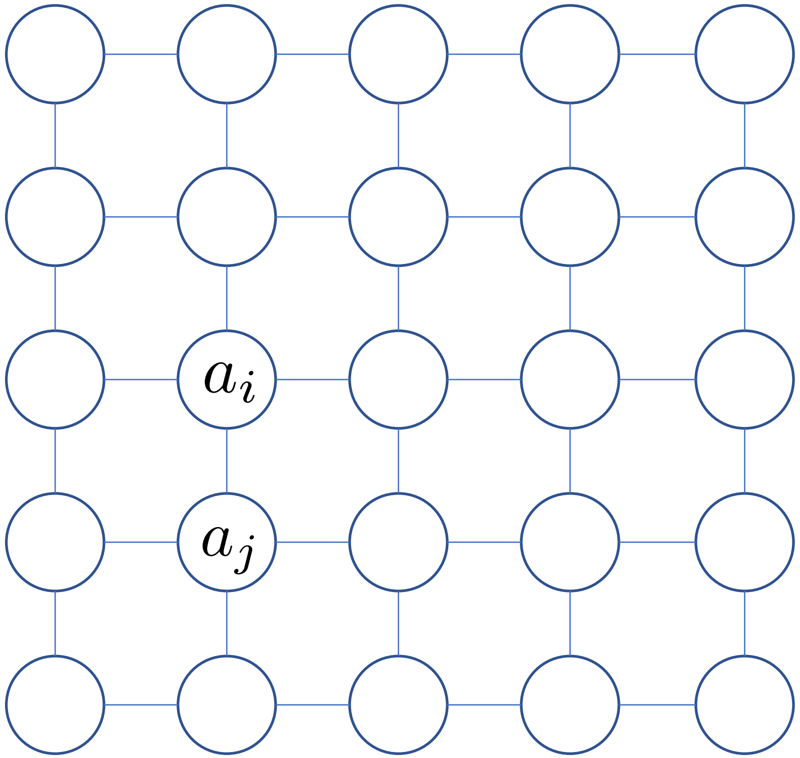
Initial attempt at a Bayesian model

The learning problem?



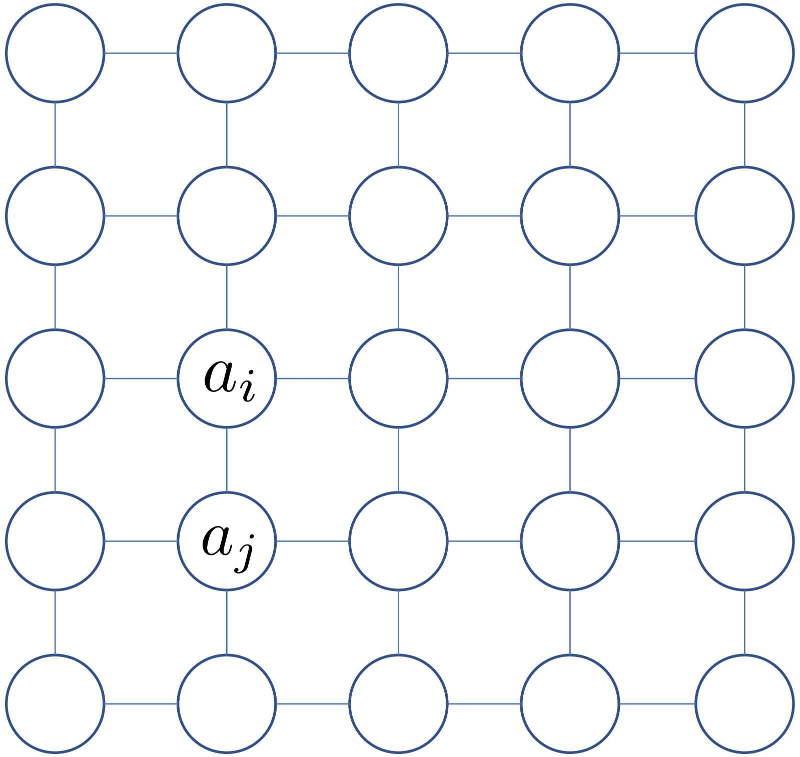
Given the training data, infer the probability of shock $P(o|x)$ across the whole stimulus space

Associative maps as Markov random fields



Associative strength for the i-th and j-th items in the map

Associative maps as Markov random fields

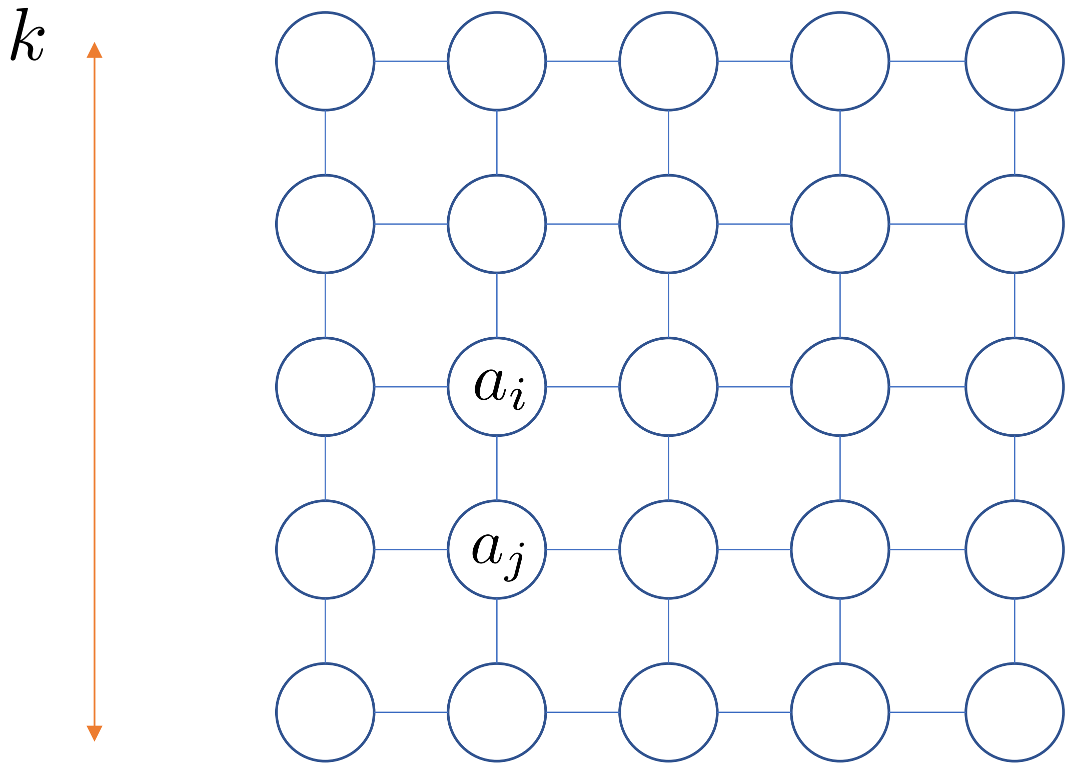


Smoothness of the map at this edge is governed by lambda

$$P(a_i, a_j) \propto (|a_i - a_j|)^{\lambda_{ij}}$$

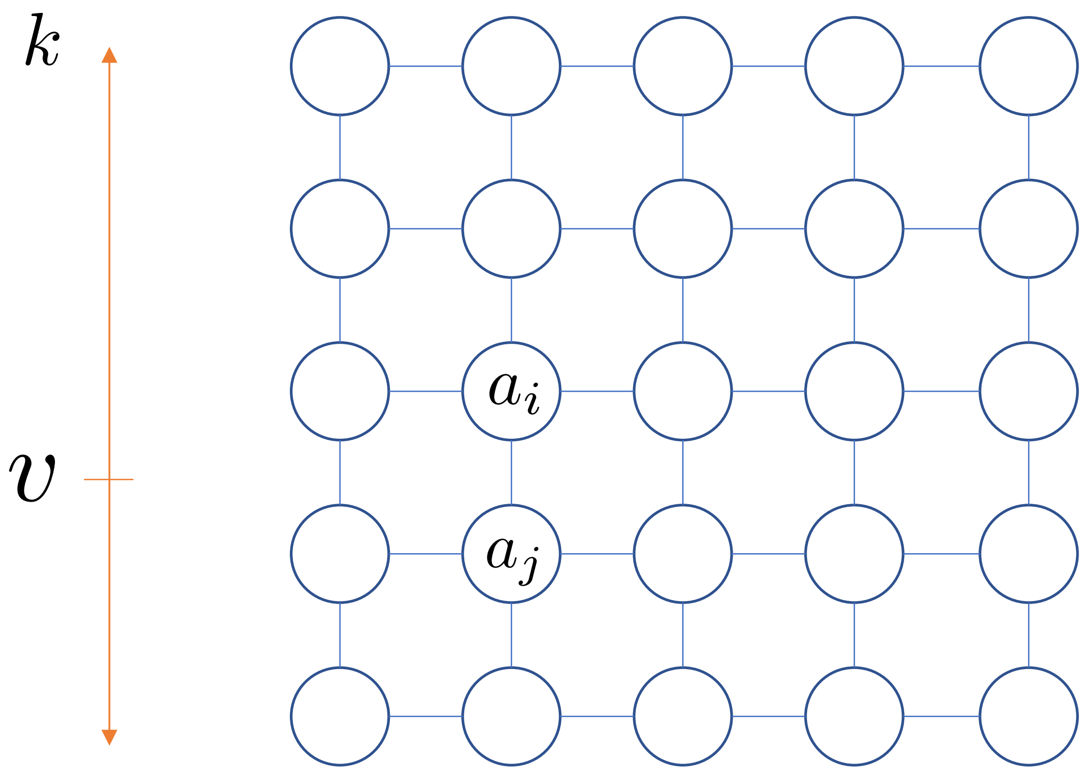
Associative maps as Markov random fields

They are connected because they have the same value on every stimulus dimension except dimension k, and differ only by a single unit along that dimension



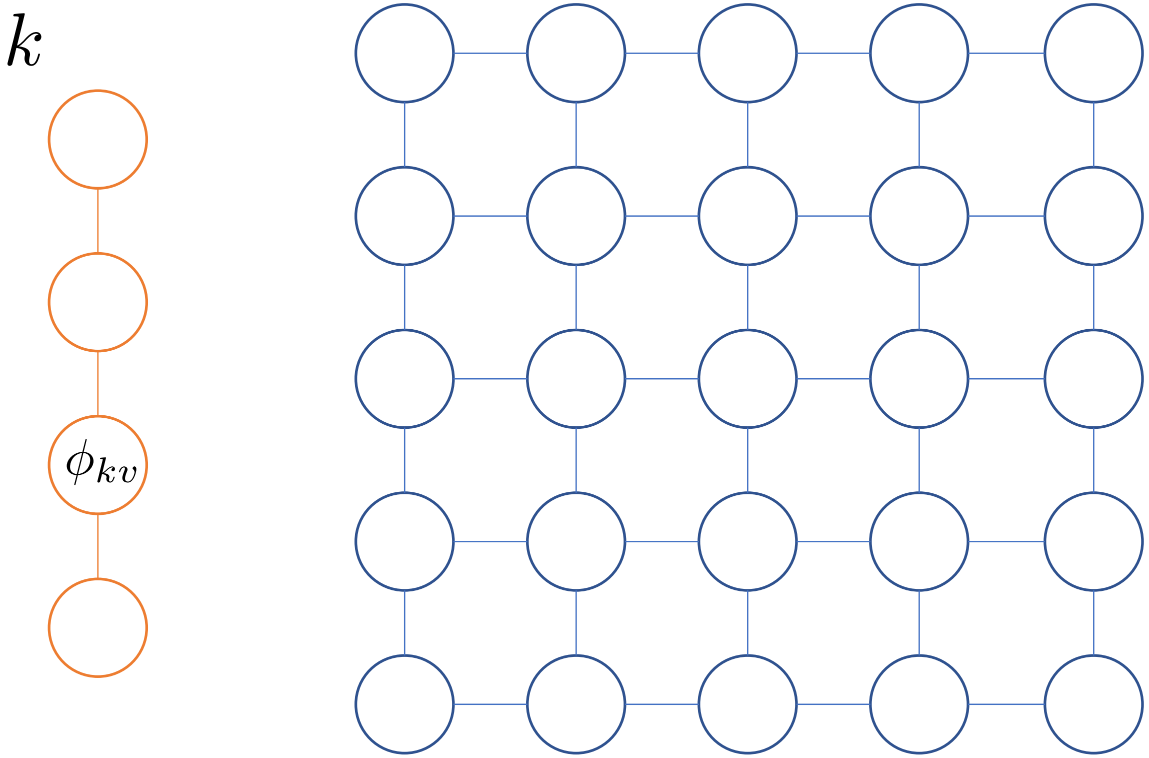
Associative maps as Markov random fields

… and the pair is located either side of position v on dimension k



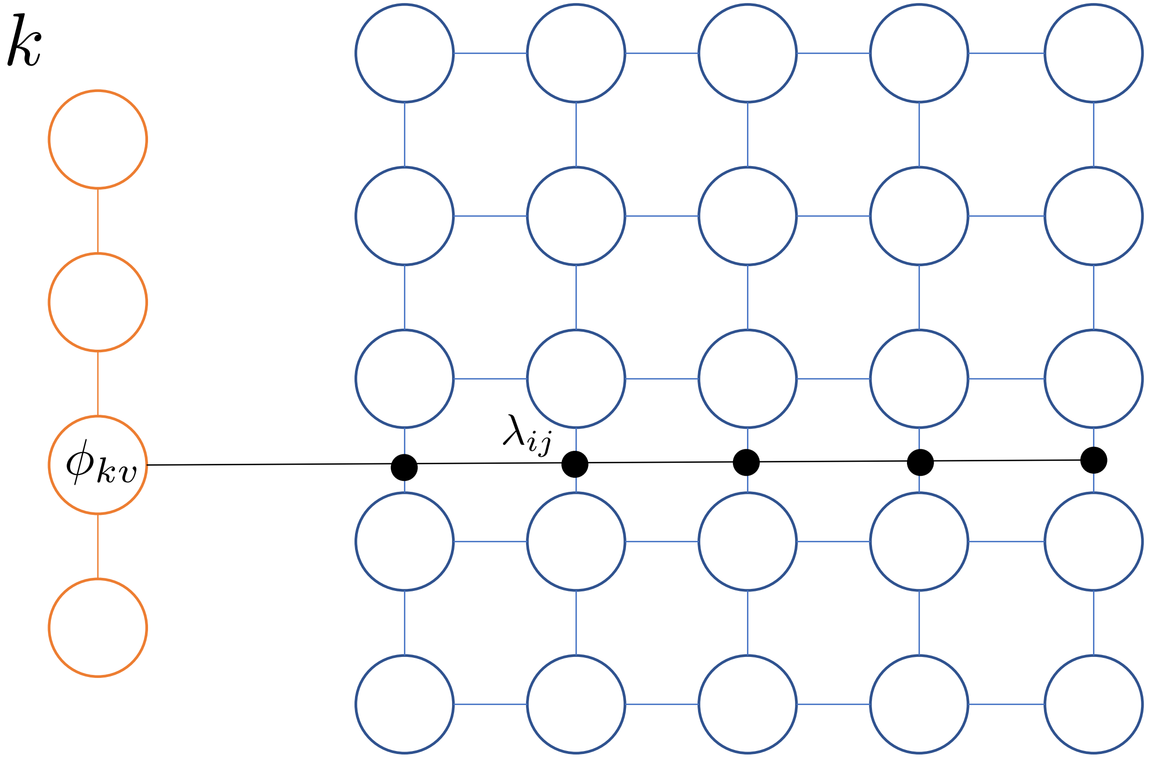
Associative maps as Markov random fields

Smoothness of this *dimension* at this location is governed by phi



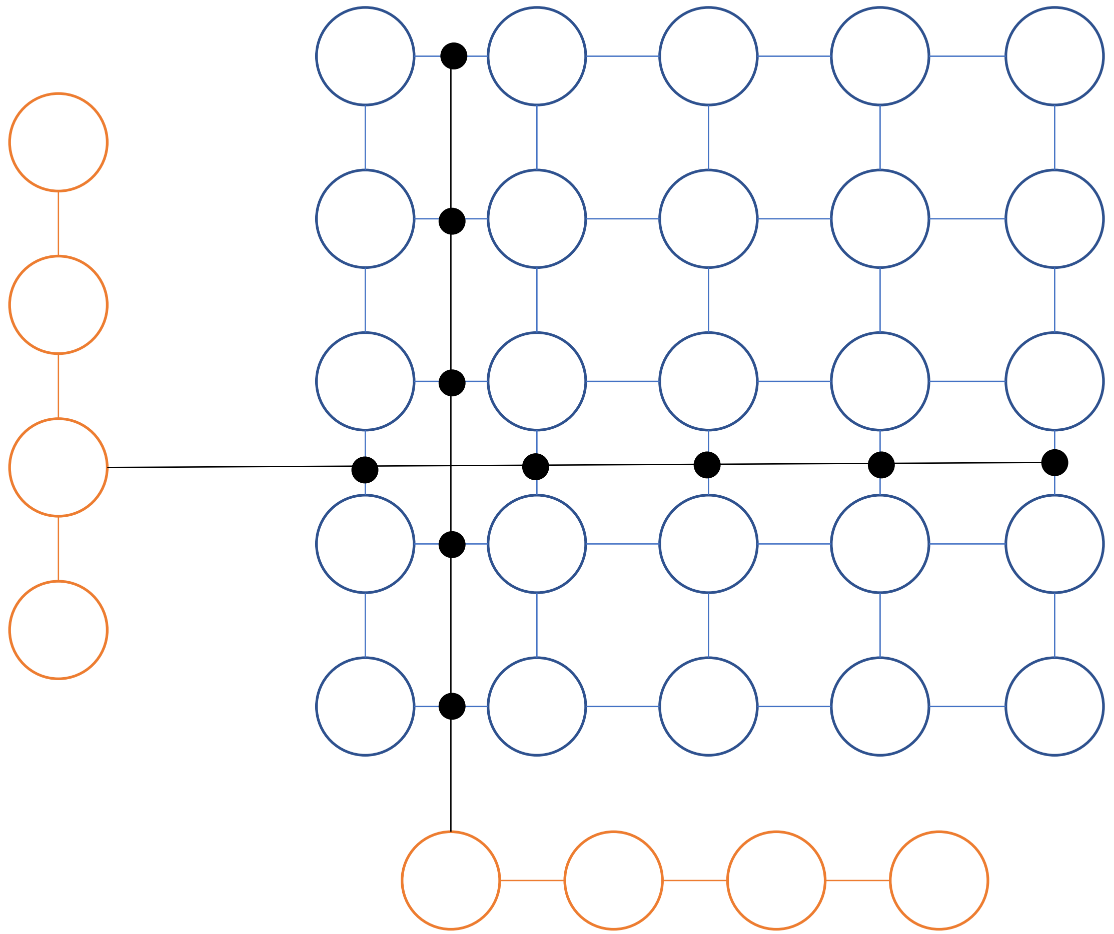
Associative maps as Markov random fields

This dimensional smoothness affects the local smoothness of every relevant edge in the lattice



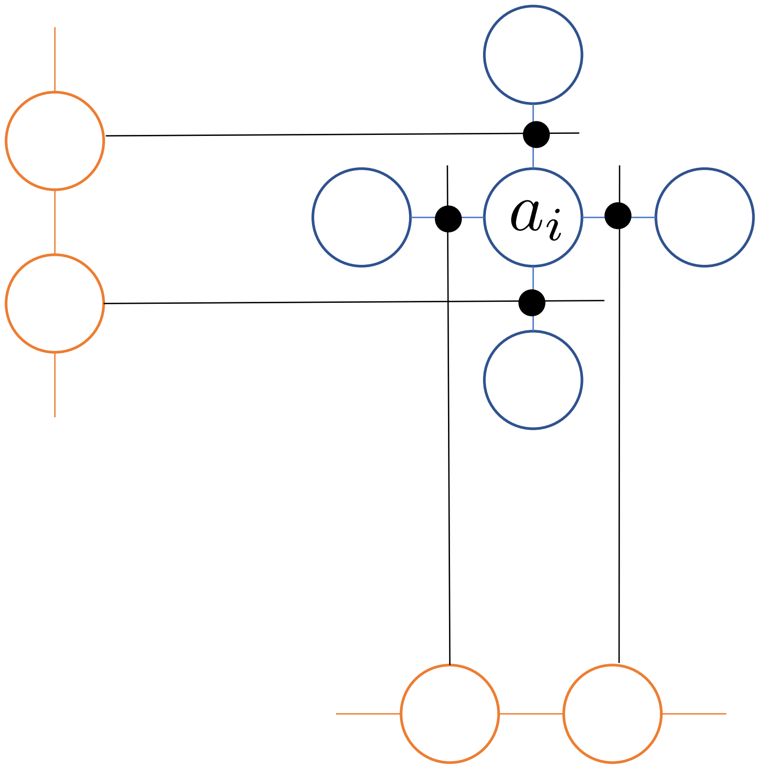
$$P(\lambda_{ij}) \propto \exp(-\phi_{kv}\lambda_{ij})$$

Associative maps as Markov random fields



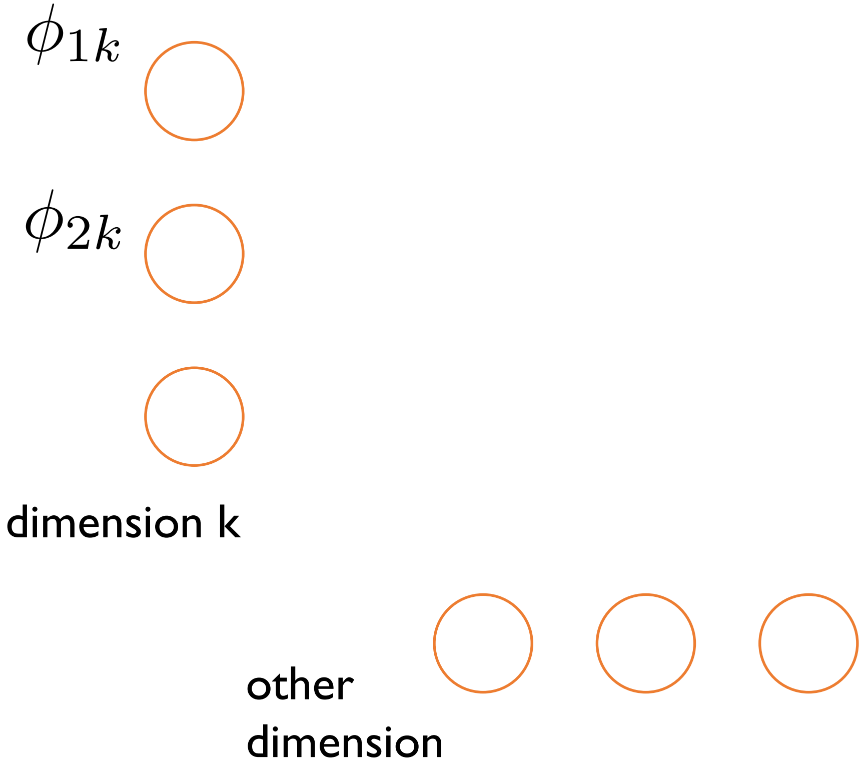
Every stimulus feature has its own dimensional representation and its own pattern of influence on the map

Associative maps as Markov random fields

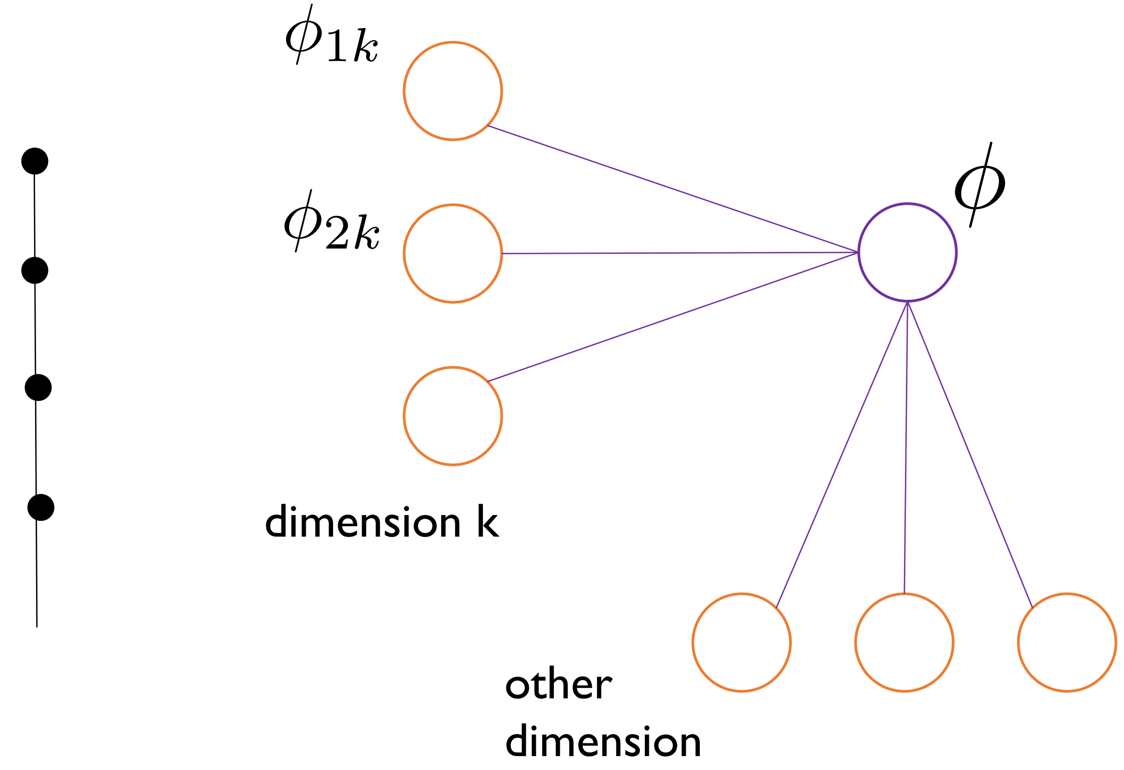


The point of this representation is to allow the associative strength of each item to be influenced by all its neighbours, in a way that respects the relative homogeneity of all dimensions

Stimulus dimensions

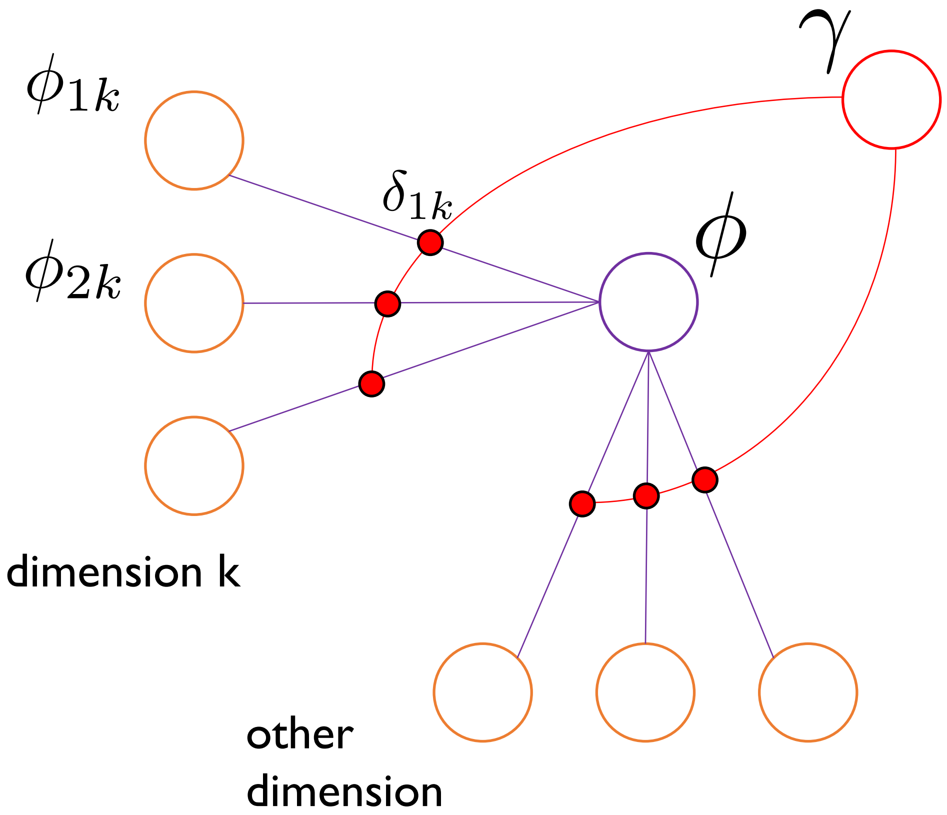


Stimulus dimensions



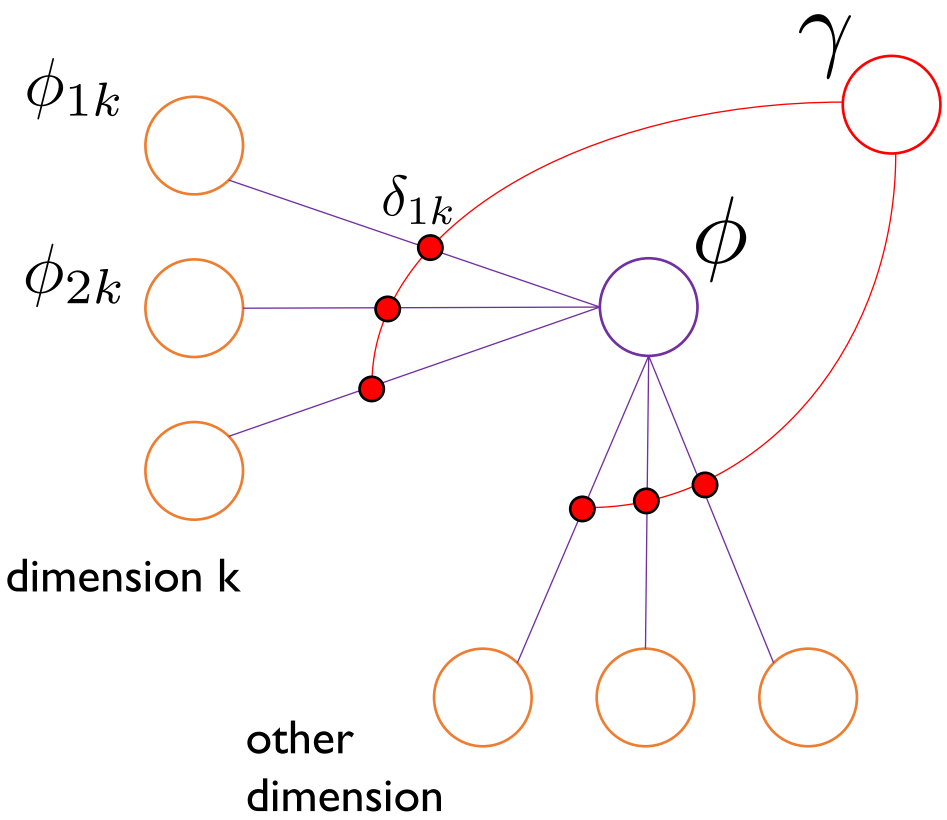
The global smoothing parameter phi influences the entire map: it acts as a tuning parameter for the learner's overall willingness to generalise

Stimulus dimensions



We allow for the possibility of random mutations, points on the dimension where there are sharp changes in association strength

Stimulus dimensions



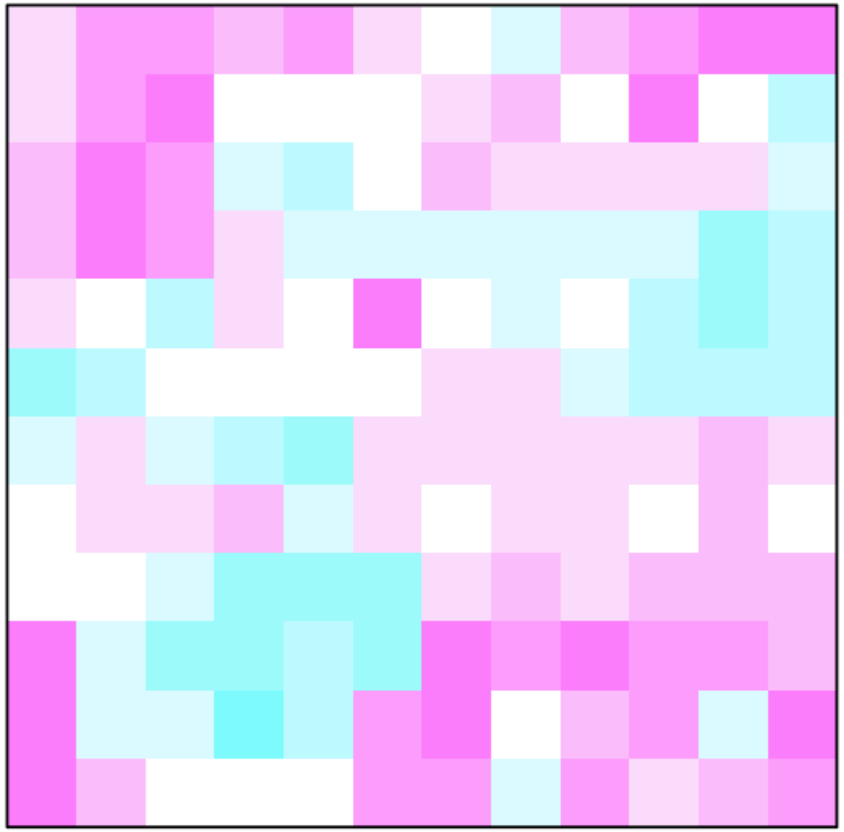
We allow for the possibility of random mutations, points on the dimension where there are sharp changes in association strength

$$\phi_{vk} = \begin{cases} \phi & \text{if } \delta_{vk} = 0 \\ \gamma\phi & \text{if } \delta_{vk} = 1 \end{cases}$$

$$\begin{aligned} P(\delta_{vk} = 1) &= \theta_{vk} \\ P(\theta_{vk}) &\propto 1 \end{aligned}$$

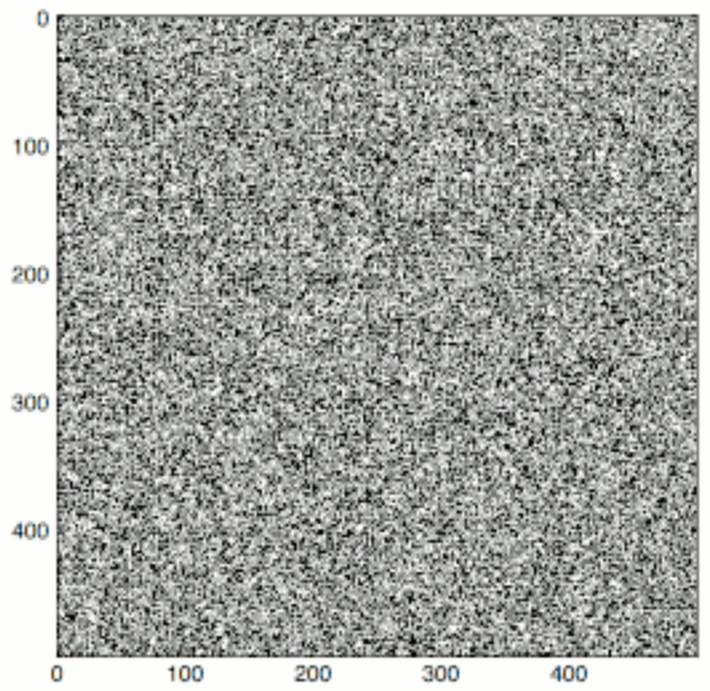
Set gamma = .5 and phi = 15.

This is what a sample from $P(A)$ looks like



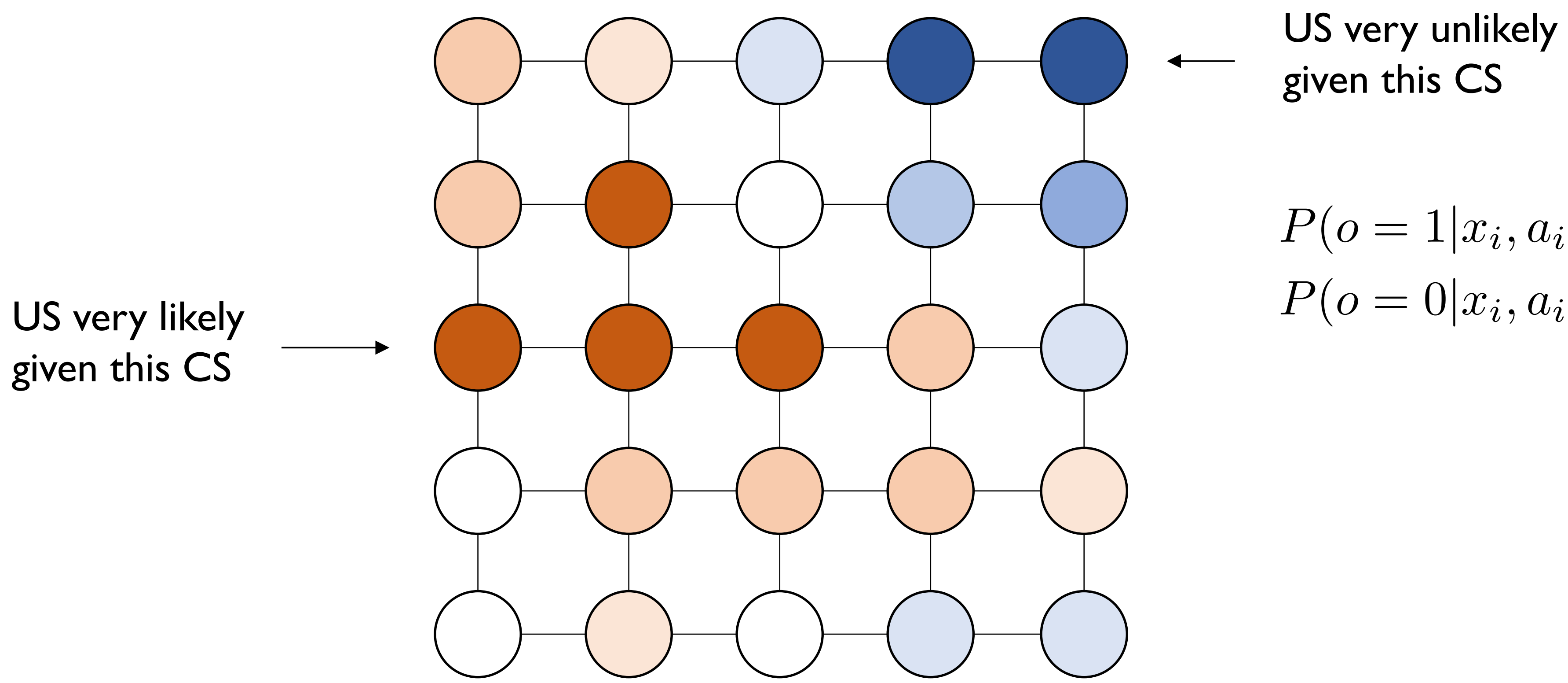
Imposes a weak "local smoothness" constraint

Not as novel as it sounds. This is a slightly fancier version of an old idea in physics and computer science…



(Ising model)

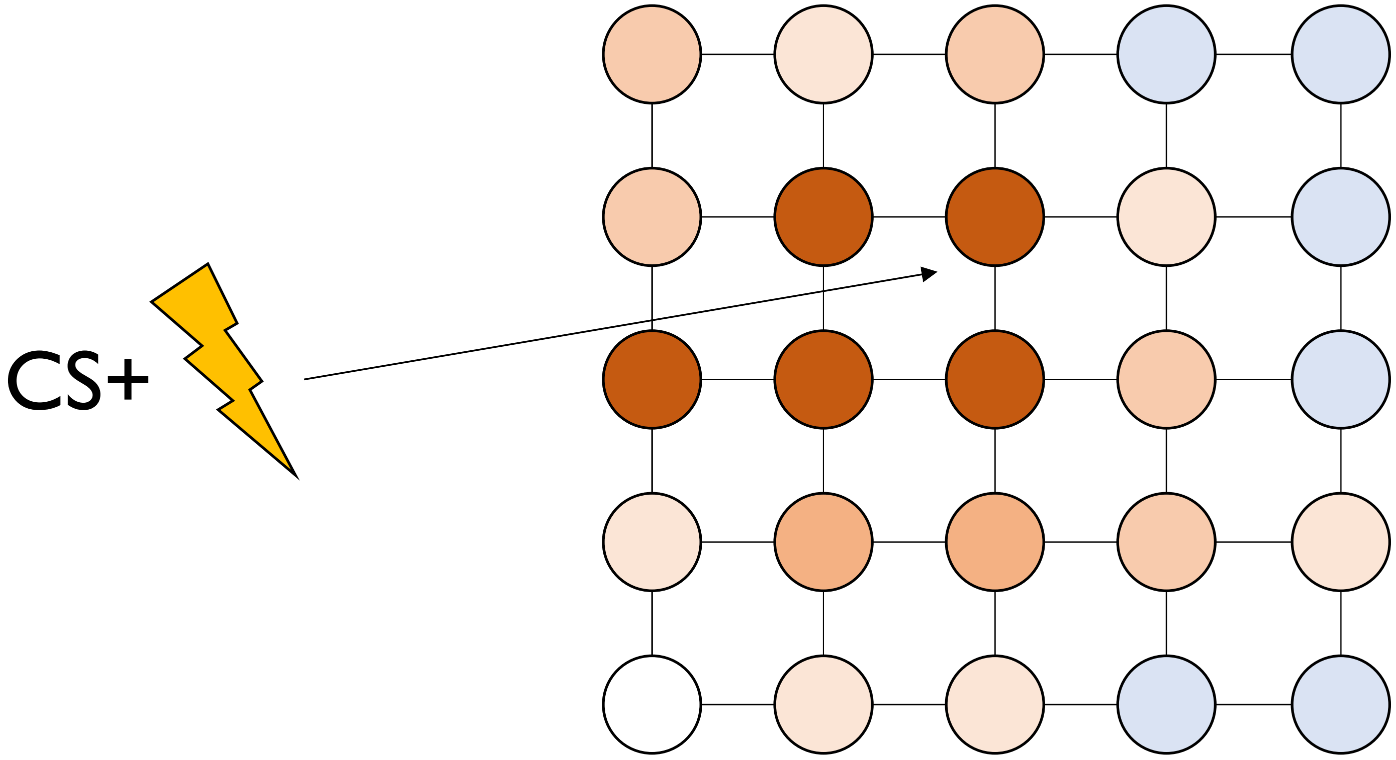
An associative map makes predictions about CS-US contingencies for all items



$$
\begin{aligned}
P(o = 1 | x_i, a_i) &= a_i \\
P(o = 0 | x_i, a_i) &= 1 - a_i
\end{aligned}
$$

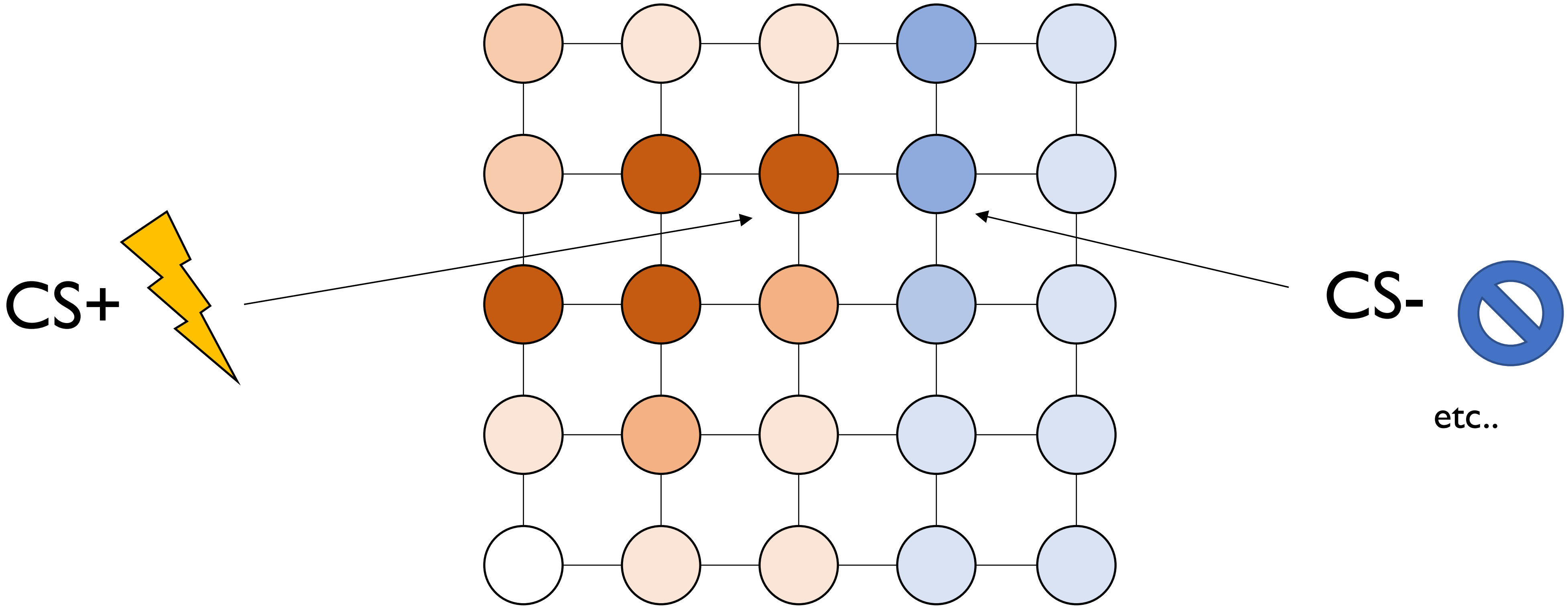
Every training trial causes learning about the presented CS, which propagates through the map

(using MCMC for Bayesian updating, but whatever)



Every training trial causes learning about the presented CS, which propagates through the map

(using MCMC for Bayesian updating, but whatever)



Bayes rule for this problem

$$\begin{aligned} P(a|x,o) &\propto P(x,o|a)P(a) \\ &= P(o|x,a)P(x|a)P(a) \end{aligned}$$

This is the prediction our associative map makes about the outcome when a stimulus is presented

This is our MRF prior over possible associative maps

Bayes rule for this problem

$$
\begin{aligned}
P(a|x,o) &\propto P(x,o|a)P(a) \\
&= P(o|x,a)P(x|a)P(a)
\end{aligned}
$$

What is this????

Bayes rule for this problem

$$
\begin{aligned}
P(a|x,o) &\propto P(x,o|a)P(a) \\
&= P(o|x,a)P(x|a)P(a)
\end{aligned}
$$

The sampling model provides the learner's theory of the situation … P(x|a) is the probability that we would encounter stimulus x if this association map is true

The learner can have many theories

I only encounter things that shock me

Stimuli appear randomly with no connection to shock

Someone is trying to teach me about shock

Someone is trying to protect me from shock

Two important cases

The world is selects the stimuli with no goal and no purpose

The stimulus selection is independent of the associative map, so…

$$P(x|a) \propto 1$$

(weak sampling)

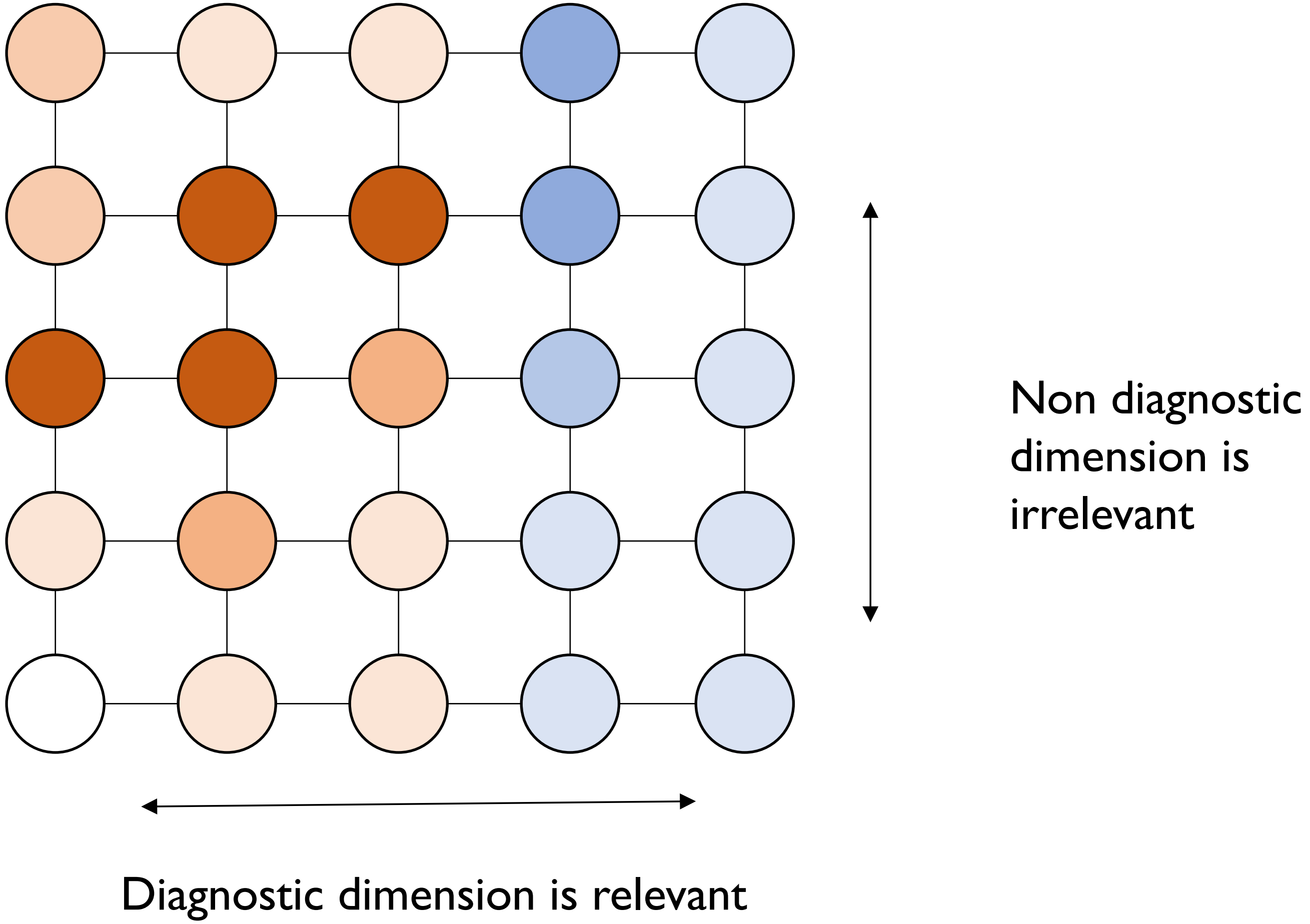
A knowledgeable person is trying to **teach** me the association map

The stimulus selection is designed to be **helpful**..

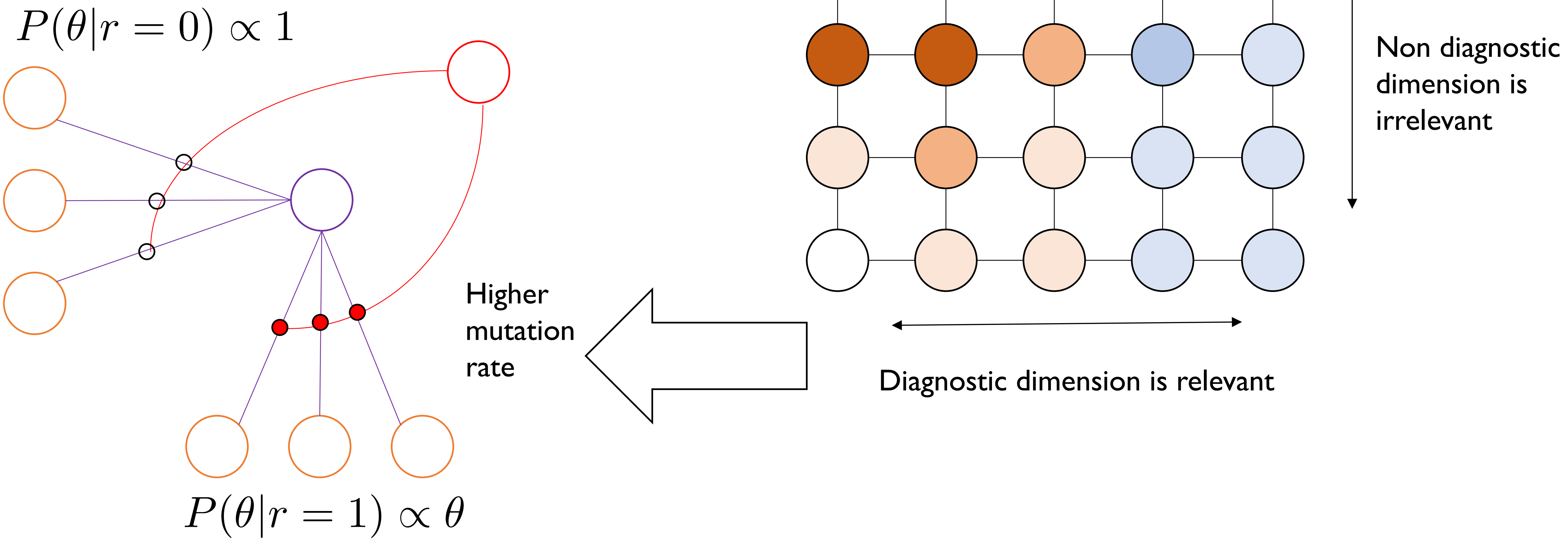
- Gricean maxims
- Pedagogical sampling
- Rational speech act

GOAL #1

Teacher wishes to communicate which stimulus dimensions are relevant and which are irrelevant to the problem

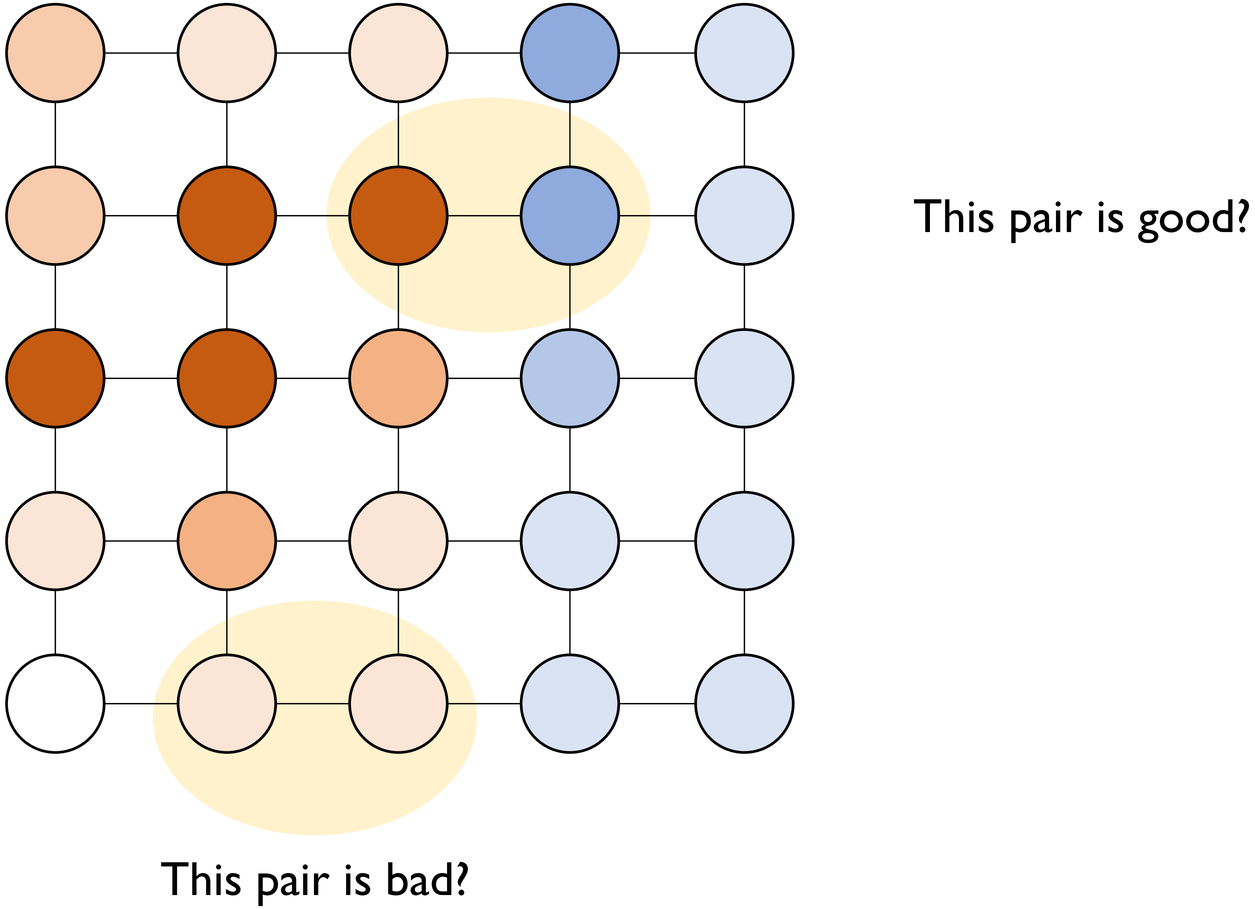


If the teacher successfully communicates relevance, the learner should make finer grained distinctions with respect to relevant dimensions



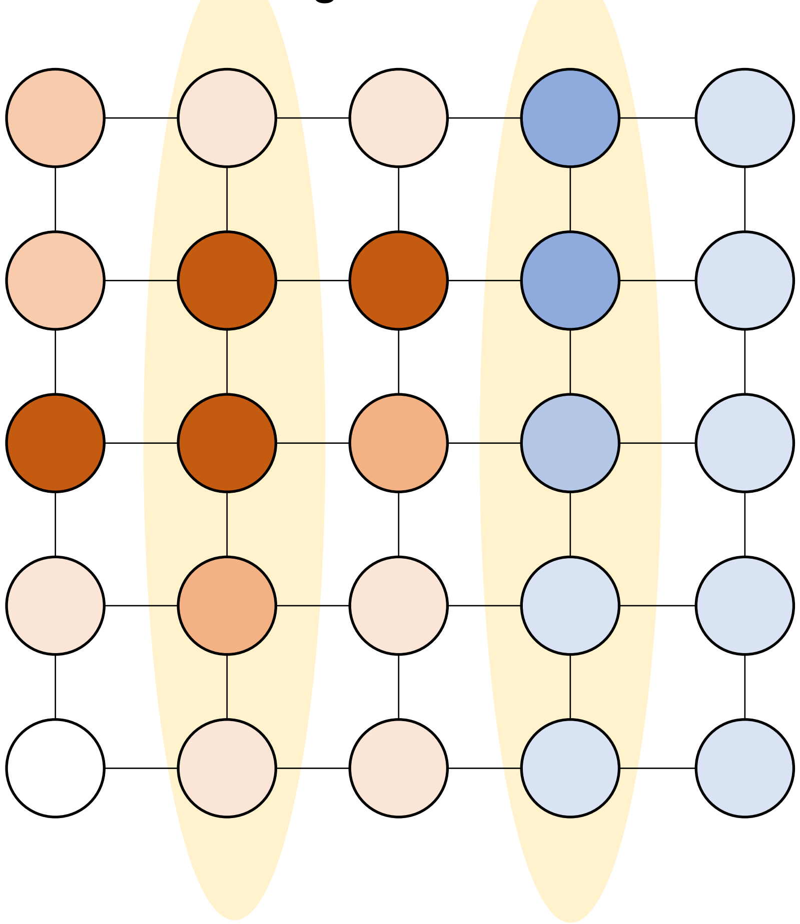
GOAL #2

Teacher wishes to select items that provide unambiguous evidence about the relevant distinction?



These items have the highest average associative strength

These items have the lowest average associative strength



For a CS+ and CS- design, these are the best dimensional values to communicate

Learner assumes that the teacher selected CS+ probability proportional to the average associative strength of items that share the relevant value

$$
\begin{aligned}
u_{o=1}(x|r) &= \bar{a}(x,r) \\
u_{o=0}(x|r) &= 1-\bar{a}(x,r)
\end{aligned}
$$

What behaviour do these models produce?

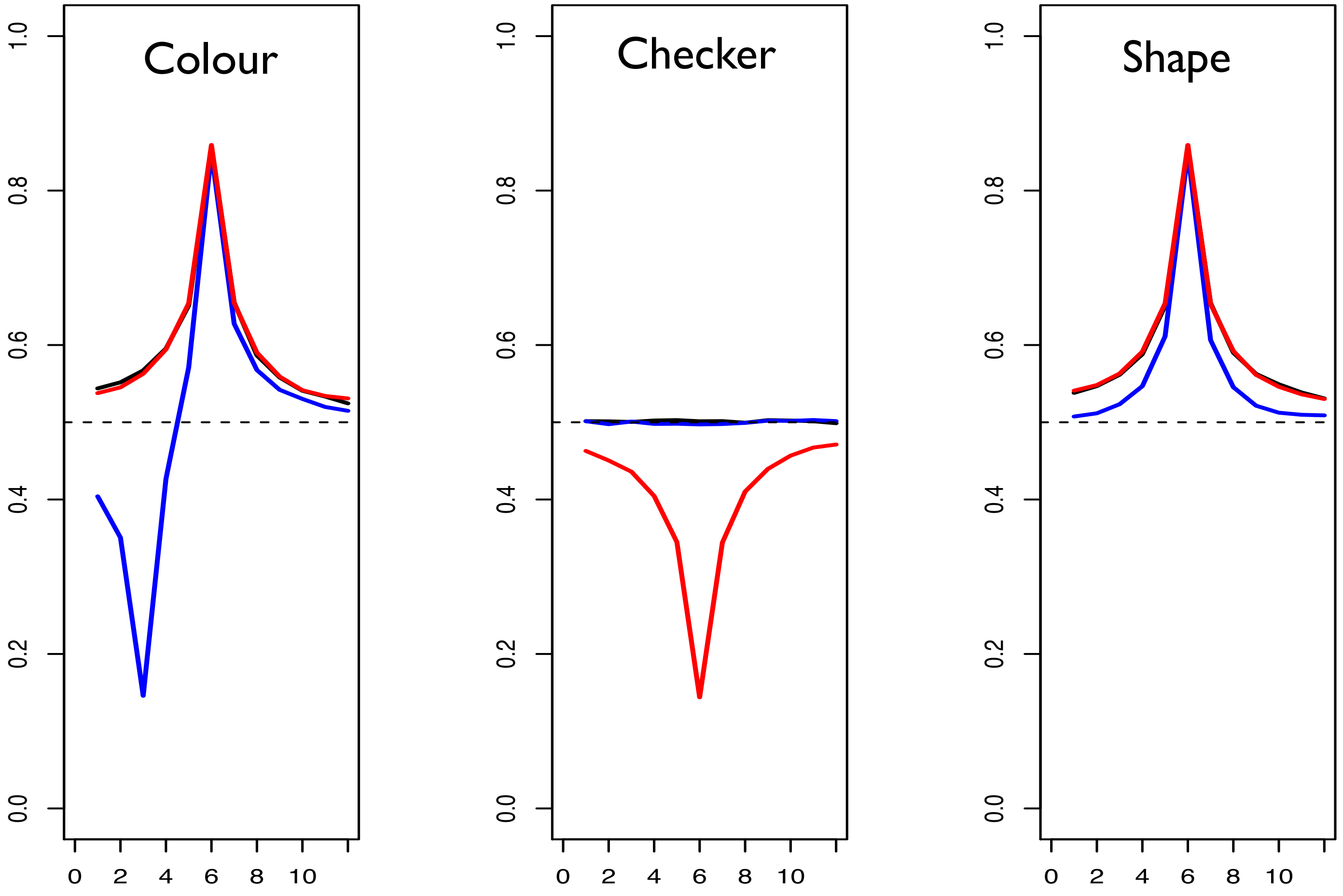
Weak sampling

```
> opt$relevance_weak
       TT SZ BG CH
single  0  0  0  0
near    0  0  0  0
far     0  0  0  0
```

We "hard code" a model in which nothing is deemed relevant and no communicative intentions exist

Generalisation patterns under weak sampling

```
> opt$relevance_weak
       TT SZ BG CH
single  0  0  0  0
near    0  0  0  0
far     0  0  0  0
```



What if relevance has been communicated?

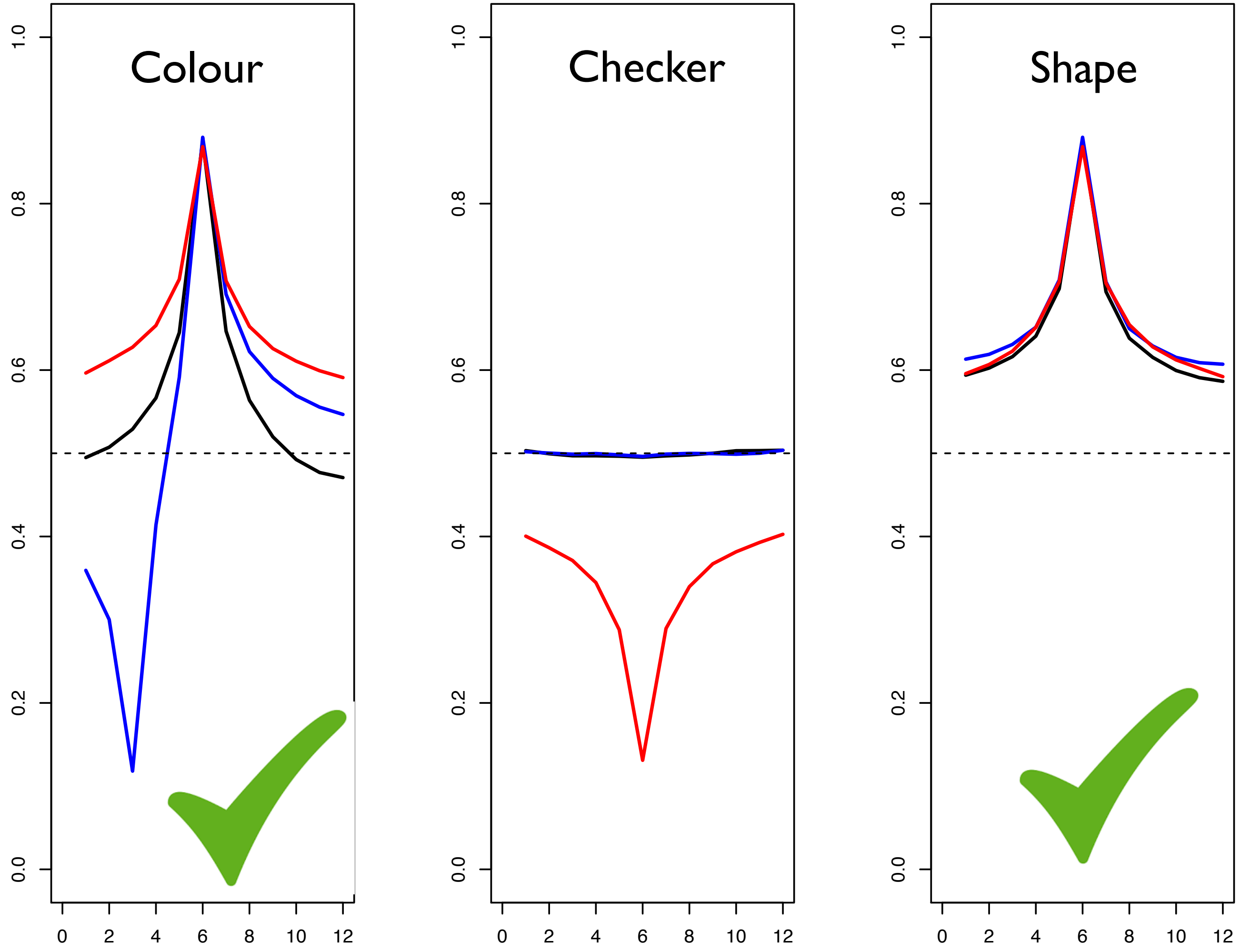
```
> opt$relevance_texture
       TT SZ BG CH
single  0  0  1  0
near    0  0  1  0
far     1  0  0  0
```

We “hard code” a model in which the learner has mysteriously worked out that colour is relevant in the single and near conditions; whereas the texture type (checkered vs solid) is relevant in the far condition

Generalisation when a single relevant dimension is communicated

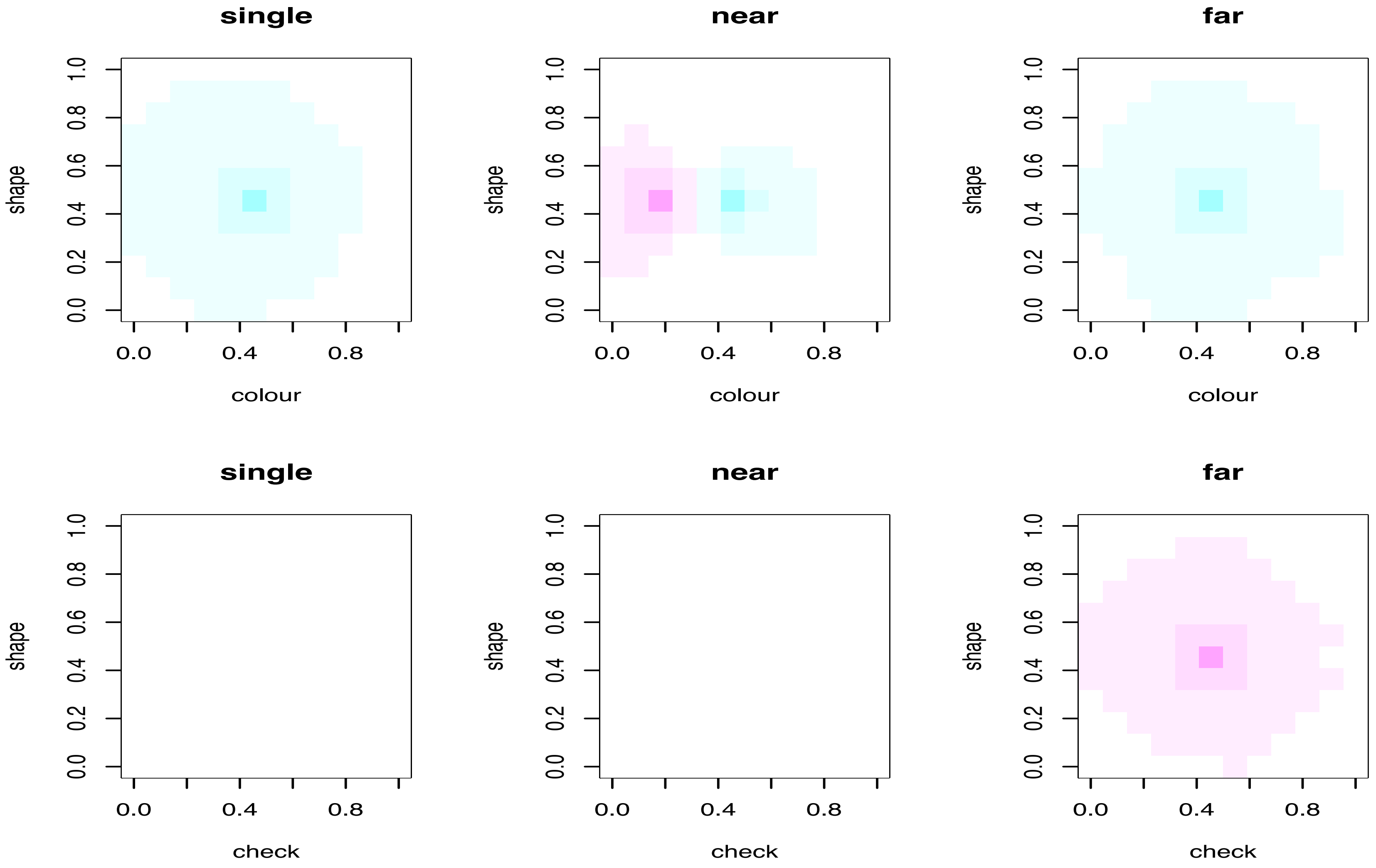
```
> opt$relevance_texture
       TT SZ BG CH
single  0  0  1  0
near    0  0  1  0
far     1  0  0  0
```

Yay!



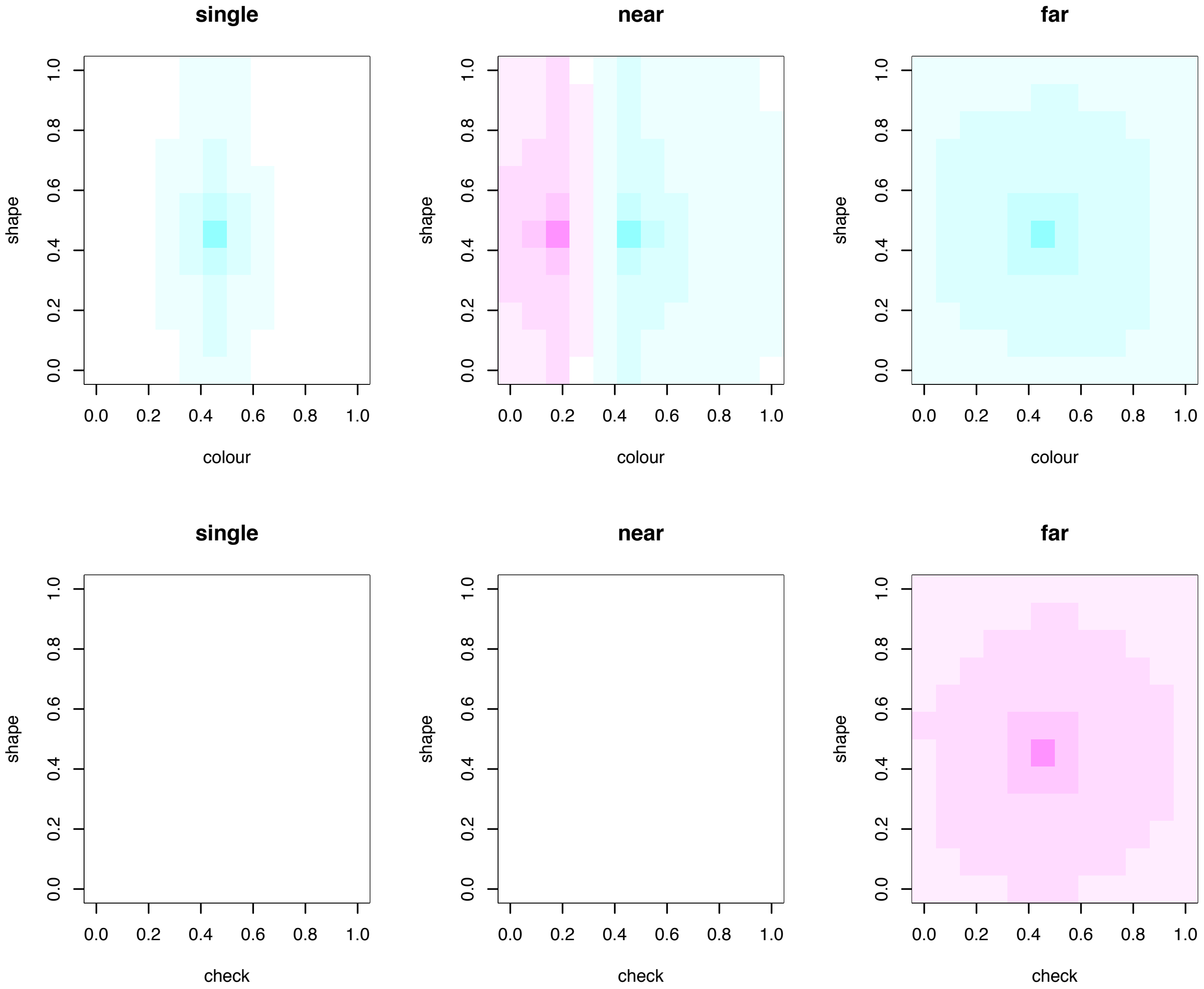
Maps learned via weak sampling

```
> opt$relevance_weak
       TT SZ BG CH
single  0  0  0  0
near    0  0  0  0
far     0  0  0  0
```



Maps learned by communicative model

```
> opt$relevance_texture
        TT SZ BG CH
single   0  0  1  0
near     0  0  1  0
far      1  0  0  0
```



Possible hints as to relevance?

```
> opt$hints
$single
              TT SZ BG CH
exists         0  1  1  0
varies_train   0  0  0  0
varies_test    0  1  1  0

$near
              TT SZ BG CH
exists         0  1  1  0
varies_train   0  0  1  0
varies_test    0  1  1  0

$far
              TT SZ BG CH
exists         1  1  1  1
varies_train   1  0  0  0
varies_test    0  1  1  0
```

Gricean maxims suggest…

(1) The teacher should include features that are relevant
(2) The teacher should not include irrelevant features

(3) The teacher should vary relevant dimensions at training
(4) The teacher should not vary irrelevant dimensions at training

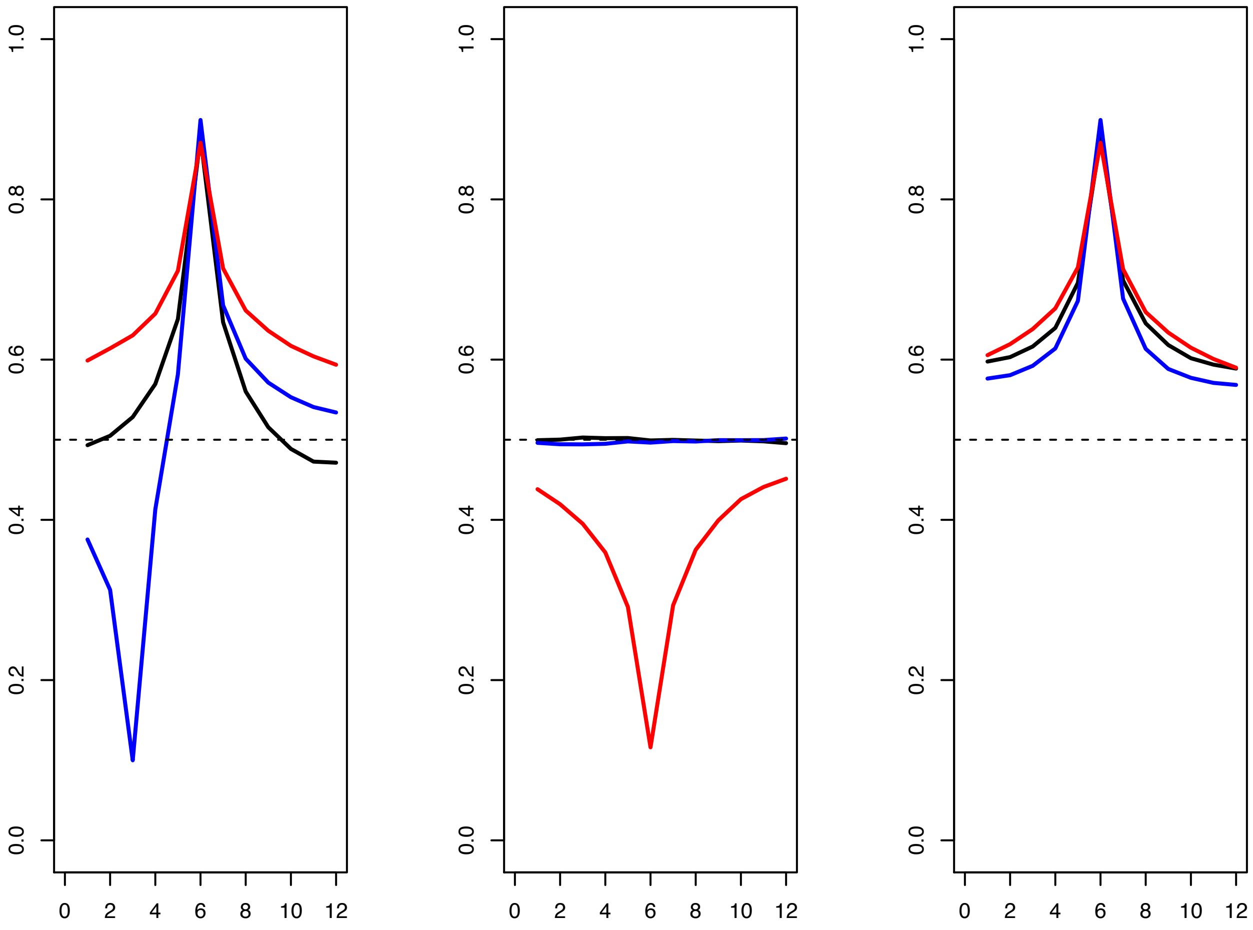
(5) The teacher should make relevant features salient

… not so sure about test trial variability, so I'm ignoring it

It works?

Posterior probability of relevance

	texture	bluegreen	checker	size
single	0	1	0	0.01
near	0	1	0	0.33
far	1	0	1	0.00

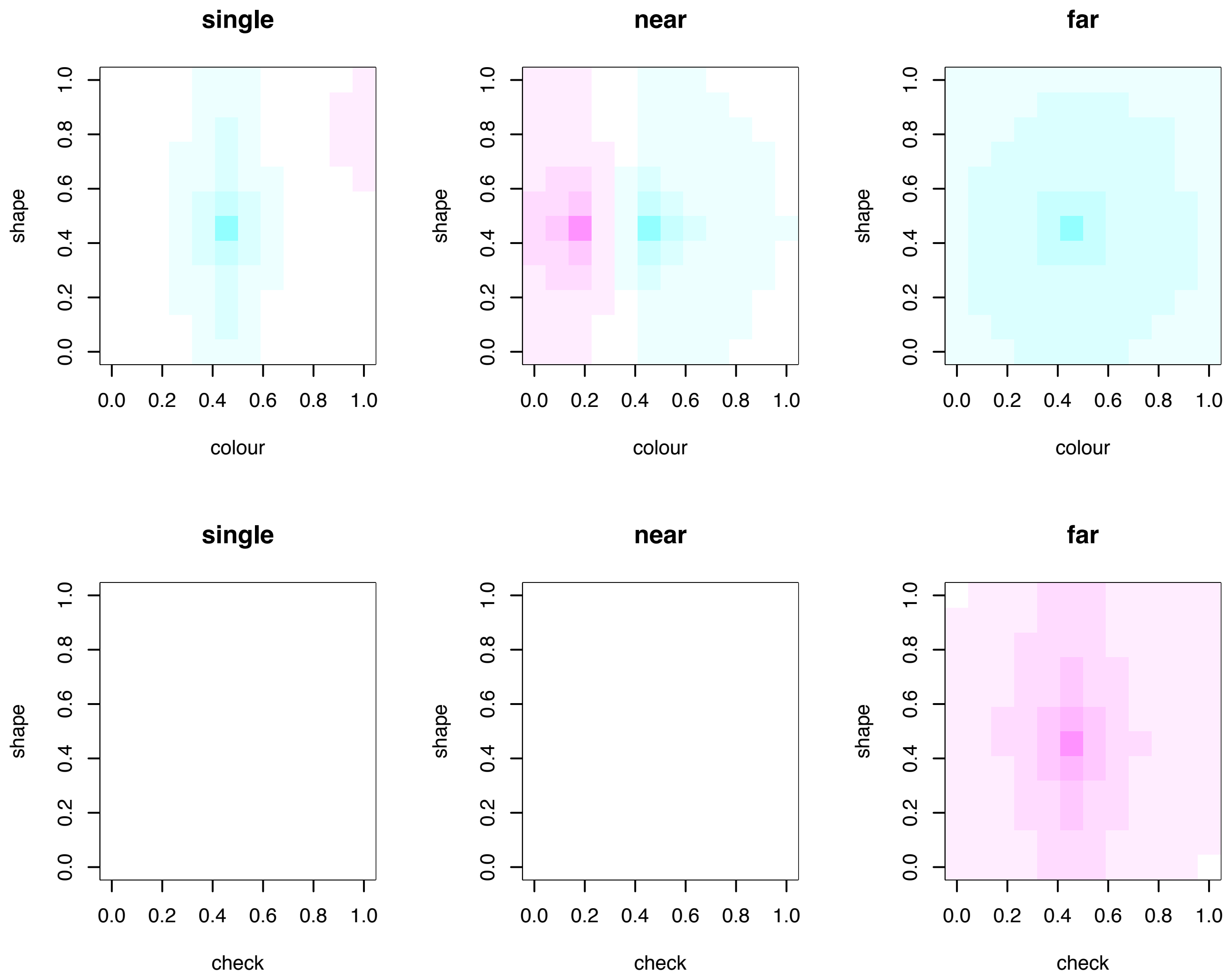


* Take this with a grain of salt. It's pretty post hoc, but still kind of neat I think

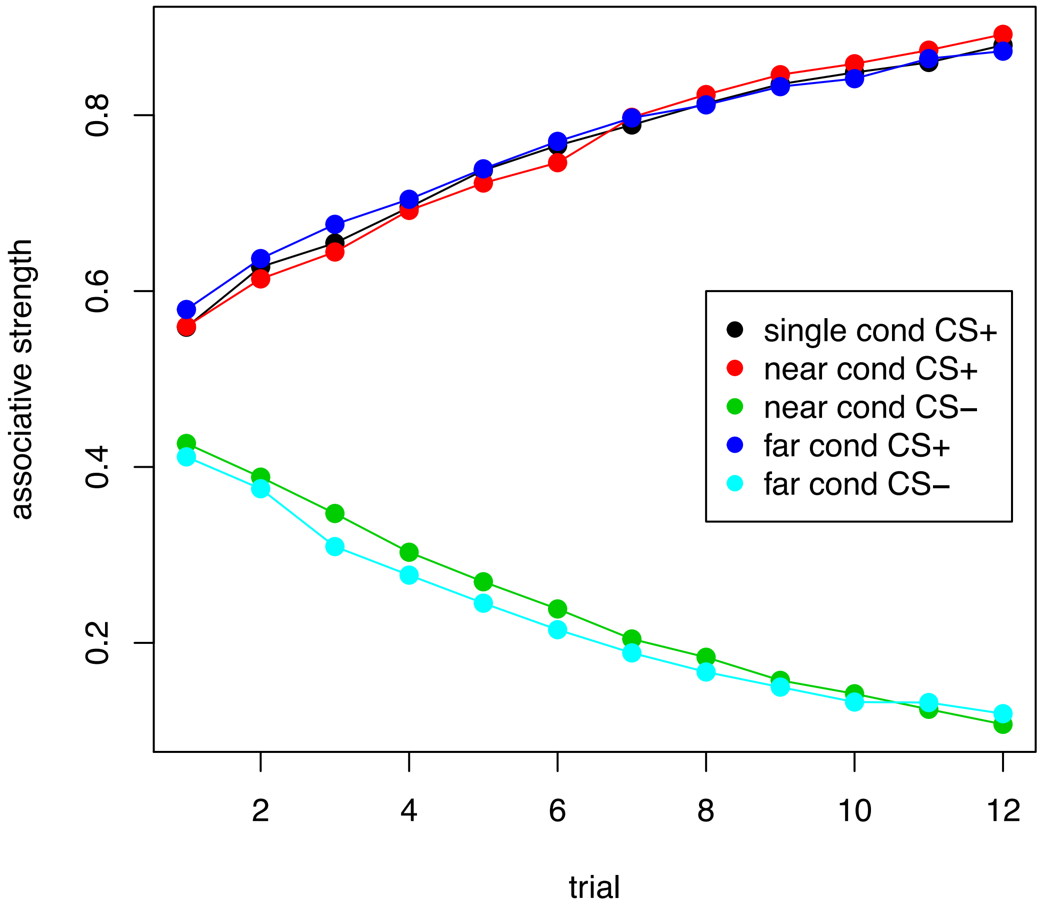
It works?

Posterior probability of relevance

	texture	bluegreen	checker	size
single	0	1	0	0.01
near	0	1	0	0.33
far	1	0	1	0.00

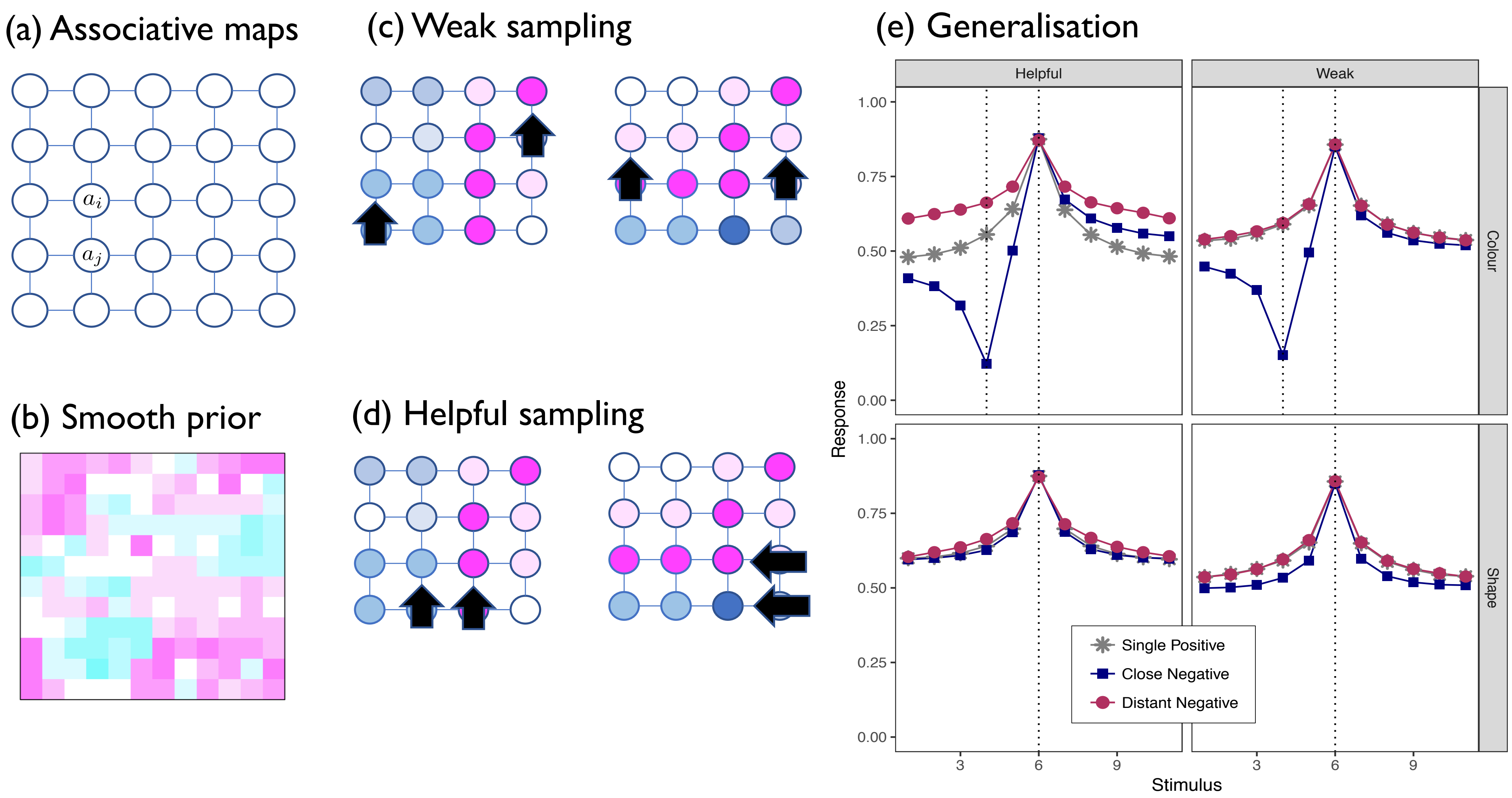


Not perfect… learning curves too shallow



Note, I haven't corrected for stimulus order info (e.g., on trial 1 in near and far conds half the time this item comes first, half the time the other does

Thanks!



(a) Associative maps
a_i
a_j
(b) Smooth prior
(c) Weak sampling
(d) Helpful sampling
(e) Generalisation
Helpful
Weak
Colour
Shape
Response
Stimulus
1.00
0.75
0.50
0.25
0.00
3
6
9
Single Positive
Close Negative
Distant Negative