Statistical learning

http://compcogscisydney.org/psyc3211/

A/Prof Danielle Navarro
d.navarro@unsw.edu.au
compcogscisydney.org

Where are we?

- L1: Connectionism
- L2: Statistical learning
- L3: Semantic networks
- L4: Wisdom of crowds
- L5: Cultural transmission
- L6: Summary

Why do networks get this wrong?

A goat being held by a child is labelled a "dog"

Goats in trees become birds or giraffes

http://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep

Why do networks get this wrong?

a woman riding a horse on a dirt road

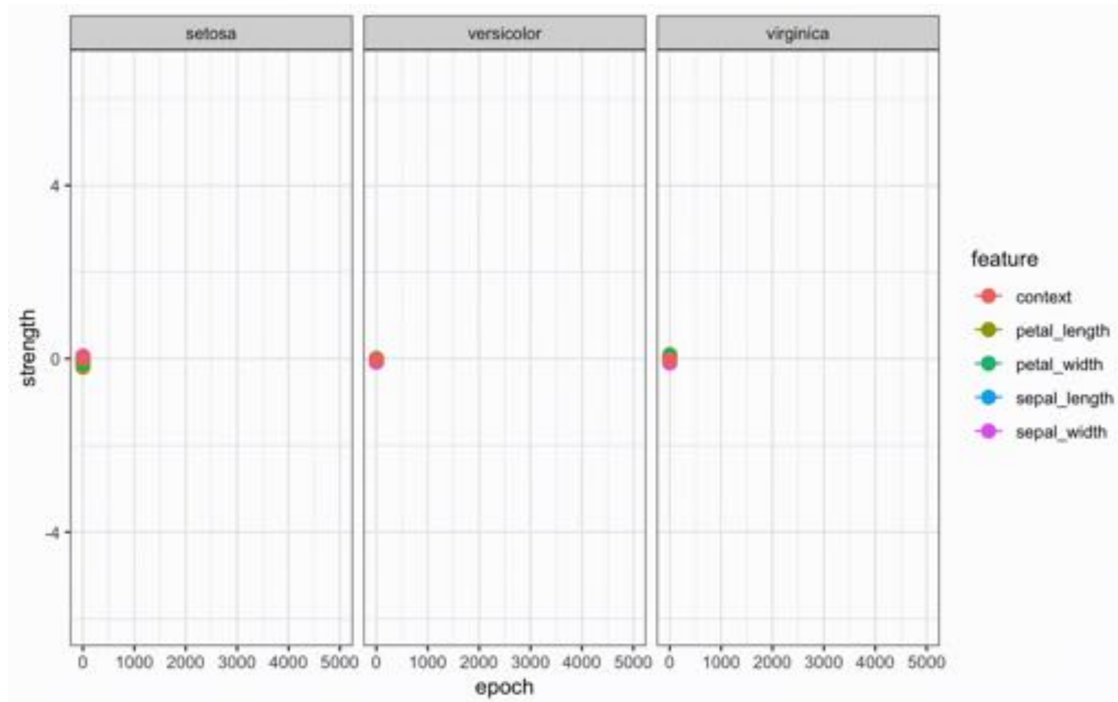
an airplane is parked on the tarmac at an airport

a group of people standing on top of a beach

Figure 6. Perceiving scenes without intuitive physics, intuitive psychology, compositionality, and causality. Image captions are generated by a deep neural network (Karpathy & Fei-Fei 2017) using code from github.com/karpathy/neuraltalk2. Image credits: Gabriel Villena Fernández (left), TVBS Taiwan/Agence France-Presse (middle), and AP Photo/Dave Martin (right). Similar examples using images from Reuters news can be found at twitter.com/interesting_jpg.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). [Building machines that learn and think like people.] Behavioral and Brain Sciences, 40.

Learning slow…



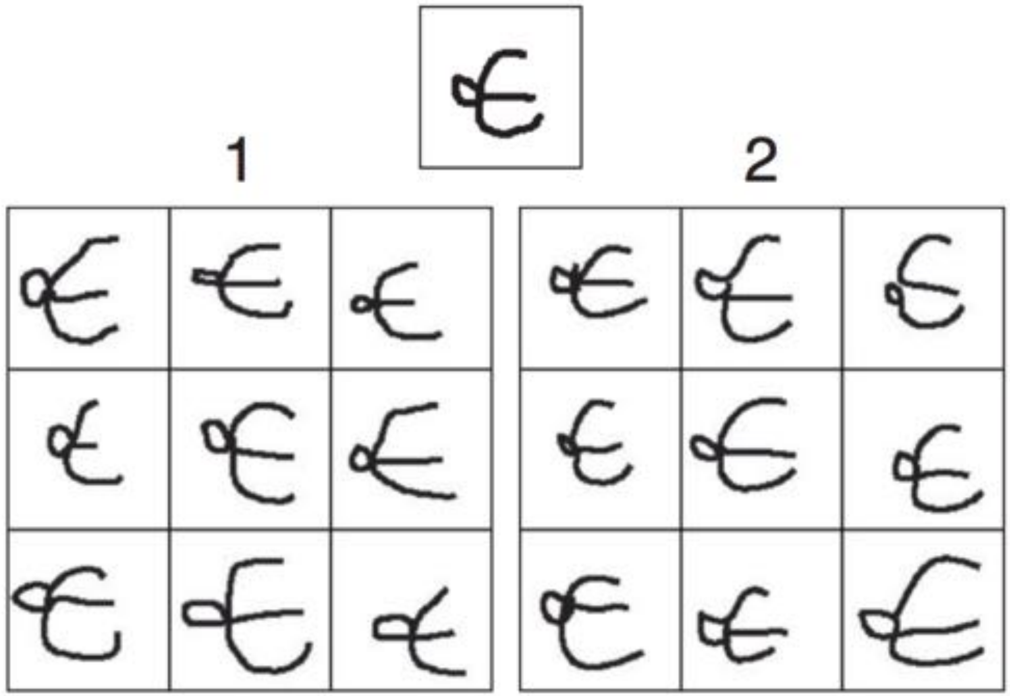
Each epoch is about 150 trials
This learning unfolds over 750,000 episodes

Learning fast…

Here is a letter written in an alien alphabet

Please write down nine more examples

A “Turing test”: Which is the human and which is the machine?

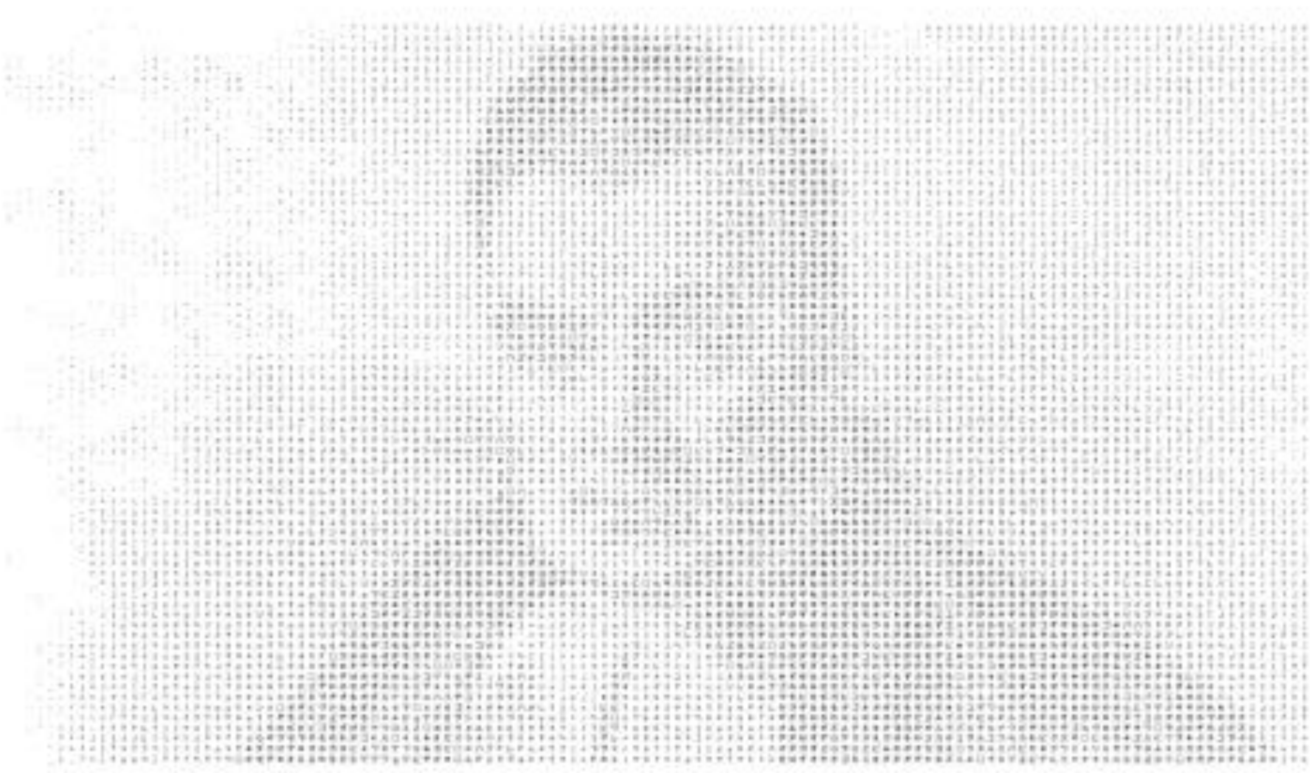


The puzzle: How does a human (or machine) do this “one-shot generalization” if learning is slow???

Structure of the lecture

- What is Bayesian reasoning?
- Two examples of psychological models
  - Coincidence detection
  - Perceptual magnet effect
- Linking Bayesian cognitive models with Bayesian machine learning

Learning with Bayes' rule



P(d|h) : the likelihood of observing d if h is true

P(h) : the prior probability that h is true

P(h|d) : the posterior probability that h is true

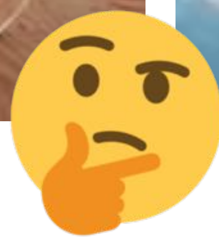
$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

P(d) : the probability of the data

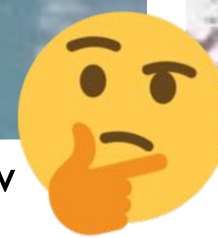
But what does this any of this gibberish *mean*?????

What happened here?
An example of Bayesian reasoning

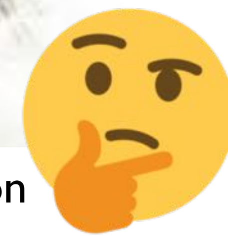

There are many possible explanations

dropped a wine glass

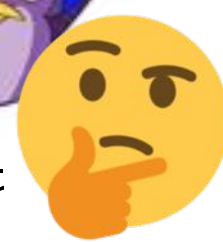



broke a window

psychic explosion


earthquake

a wizard did it

Let’s consider two of them

Someone dropped a wine glass

Kids broke the window

Prior beliefs

P(h) is the **prior**, and refers to the inherent plausibility of h as an explanation, before observing any evidence

$$\frac{P(h_1)}{P(h_2)} = 1/10$$

Relative plausibility of two hypotheses is the ratio between their prior probabilities, the **prior odds**

Before learning anything else I think "wine glass dropping" is 10 times more plausible than "broken window"

Some data

d = there is a cricket ball next to the broken glass

Likelihood of the data

P(d|h) is the **likelihood**, and describes the probability that we would have observed data d <u>if</u> the hypothesis h were true

When I drop a wine glass…

… It's very unlikely that I just happen to do so right next to a cricket ball

P(d|h) = 0.001

Likelihood of the data

P(d|h) is the **likelihood**, and describes the probability that we would have observed data d <u>if</u> the hypothesis h were true

When the kids break a window…

… It's not at all uncommon for a cricket ball to end up near the glass

P(d|h) = 0.15

Likelihood of the data

P(d|h) is the **likelihood**, and describes the probability that we would have observed data d <u>if</u> the hypothesis h were true

$$\frac{P(d|h_1)}{P(d|h_2)} = \frac{0.15}{0.001} = 150$$

Relative probability of the data according to the hypotheses is the evidentiary value of the data, referred to as the **likelihood ratio** (or the **Bayes factor**)

The data (cricket ball) are 150 times more likely under the "broken window" hypothesis

Posterior beliefs

P(h|d) is the **posterior**, and refers to the "updated" plausibility of h as an explanation, <u>after</u> observing the evidence

$$\frac{P(h_1|d)}{P(h_2|d)} = \frac{P(d|h_1)}{P(d|h_2)} \times \frac{P(h_1)}{P(h_2)}$$

Posterior odds

= 15

Likelihood ratio

= 150

Prior odds

= .1

In light of the evidence, I now think that window-breaking is 15 times more plausible than dropped-wine-glass

But I have *many* hypotheses?





…

Prior probabilities for all hypotheses

We have a set of hypotheses *h*, (called a **hypothesis space**) each of which has some degree of prior plausibility

There is a **conservation of belief** rule… if we listed <u>*all*</u> the hypotheses and assessed their prior plausibility, they would have to sum to 1

0.01

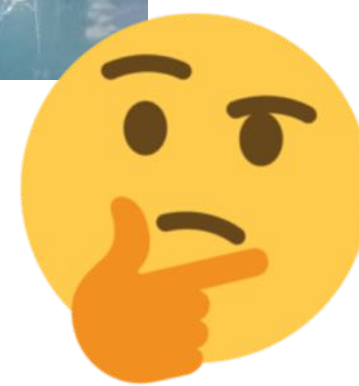
0.01

0.01

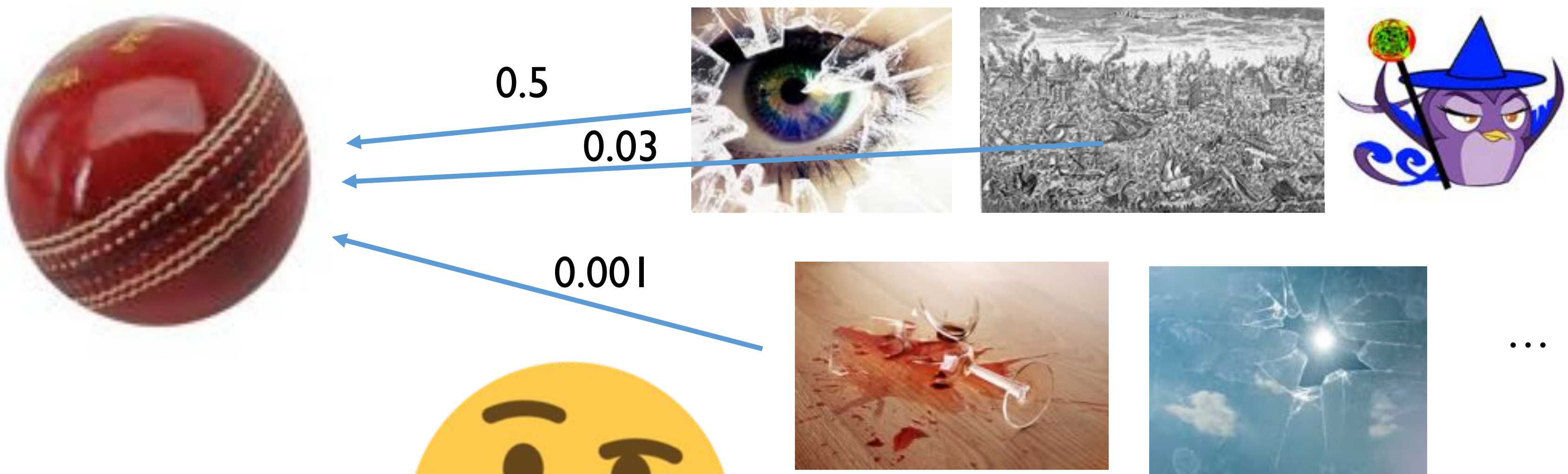
0.80

0.08

…



Likelihoods for the data, according to each hypothesis



Every hypothesis supplies a likelihood… the probability of the data (cricket ball) if that hypothesis is correct

Prior x Likelihood

To calculate posterior plausibility, hypotheses are "scored" by multiplying the prior plausibility by the likelihood of the data

$$P(h|d) \propto P(d|h) \times P(h)$$

My posterior belief P(h|d) in h now that I've seen data d…

… is proportional to …
(we'll come back to that)

… the prior belief P(h) multiplied by the likelihood P(d|h)

Prior x Likelihood

To calculate posterior plausibility, hypotheses are "scored" by multiplying the prior plausibility by the likelihood of the data

$$P(h|d) \propto P(d|h) \times P(h)$$

The posterior must satisfy the conservation of belief, and must sum to 1

The prior must satisfy the conservation of belief, and must sum to 1

Bayes' rule

$$P(h|d) = \frac{P(d|h) \times P(h)}{\sum_{h'} P(d|h') \times P(h')}$$

Conservation of belief means
that we have to divide by the
sum, taken over all hypotheses

Bayes' rule

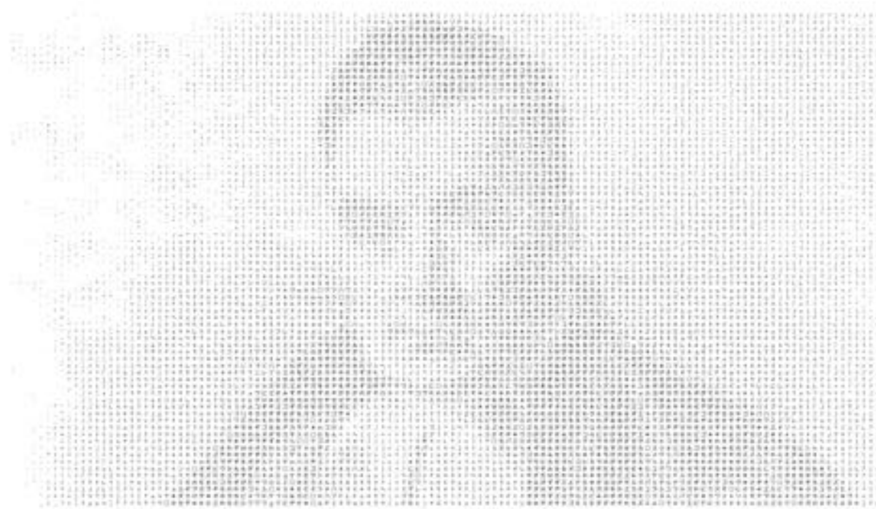
$$P(h|d) = \frac{P(d|h) \times P(h)}{P(d)}$$

That big sum is referred to as the probability of the data P(d)

(still confused? the tutorial exercise will go through this!)

Bayesian models of cognition

Example1: When is a coincidence more than a coincidence?



Mere coincidence? Or something else?

You are travelling overseas and meet your next door neighbor

You flip a coin 10 times and it comes up heads every time

A stage magician flips a coin10 times and it comes up heads every time

Five people are having a conversation and they were all born on a Monday

Coincidences model

(Griffiths & Tenenbaum 2007)

Argues that we evaluate two hypotheses:

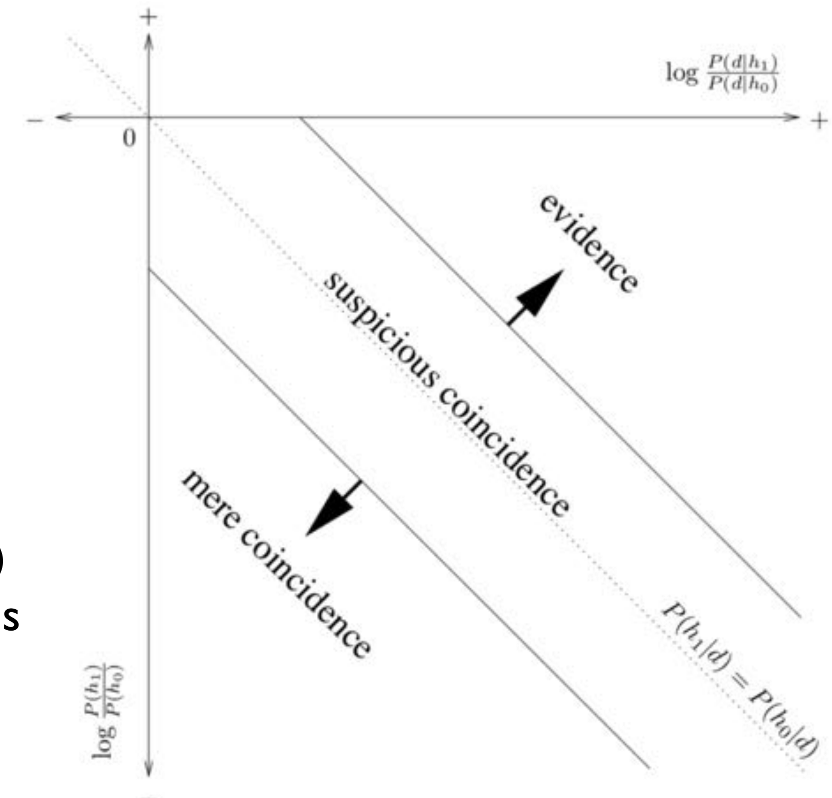
h_1: the observations are due to chance outcomes from an unstructured process

h_2: the observations are the product of a structured process

Coincidences model

(Griffiths & Tenenbaum 2007)

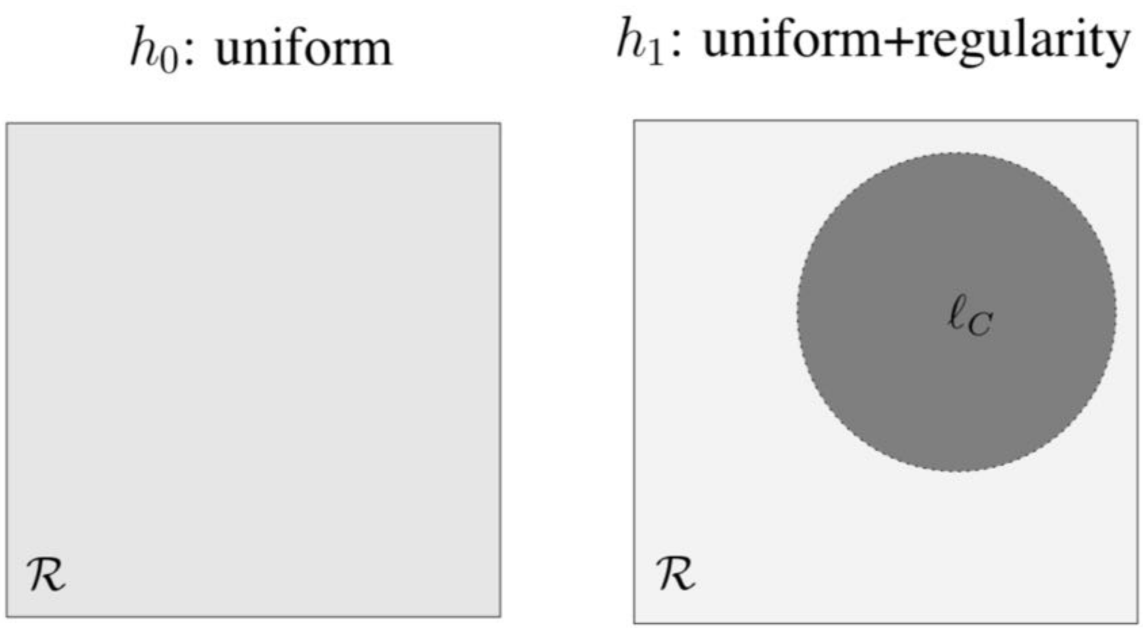
(logarithm of) the posterior odds



(logarithm of) the prior odds

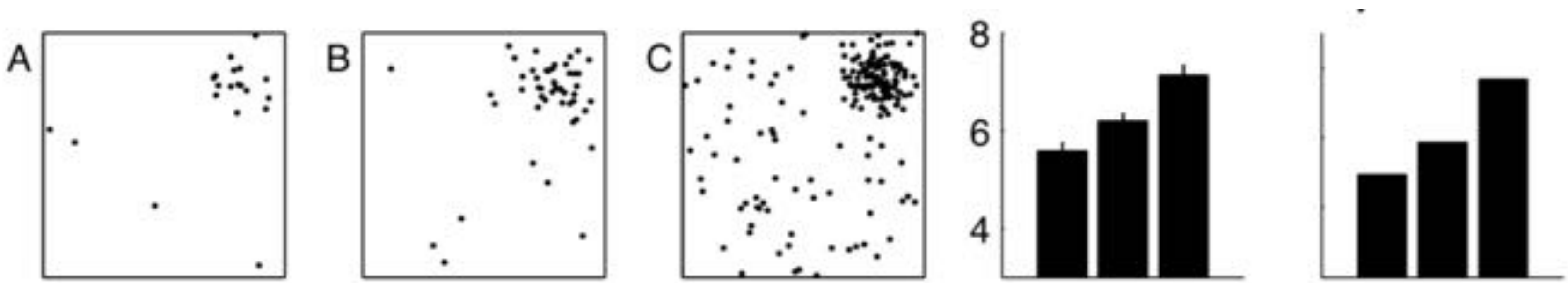
Coincidences in space

When is spatial clustering "mere coincidence"?

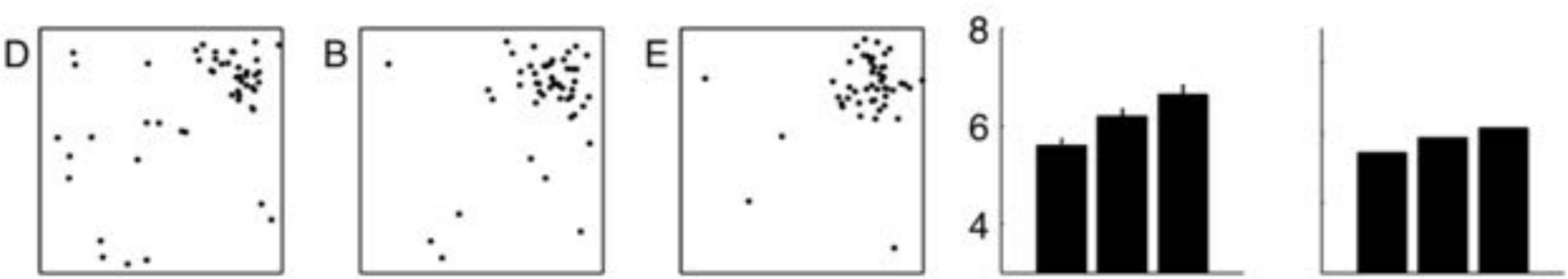


Coincidences in space

Increasing the total number of points…. Human Model

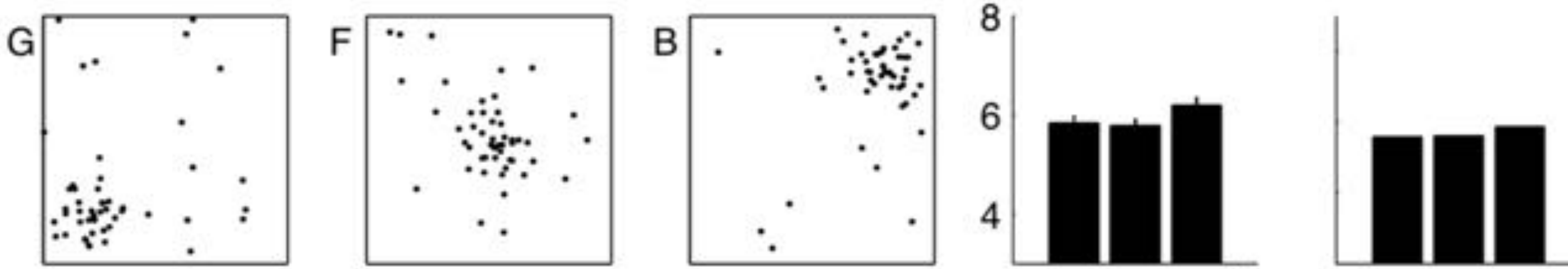


Changing the proportion of points… Human Model

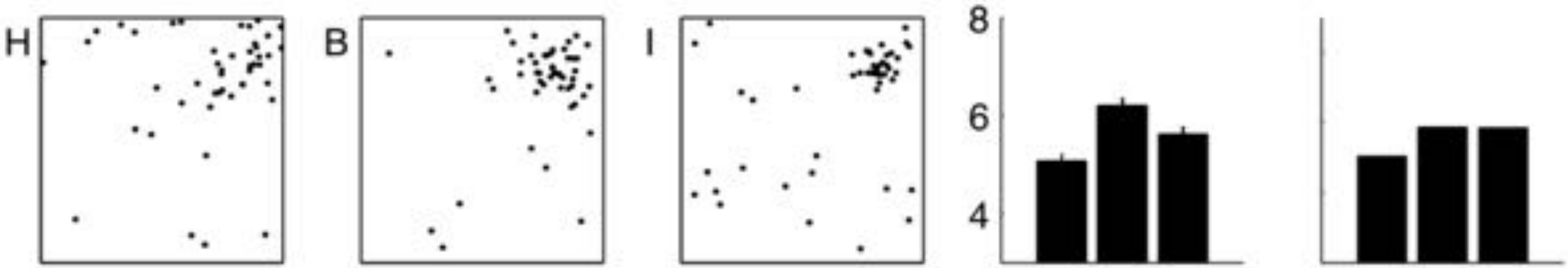


Coincidences in space

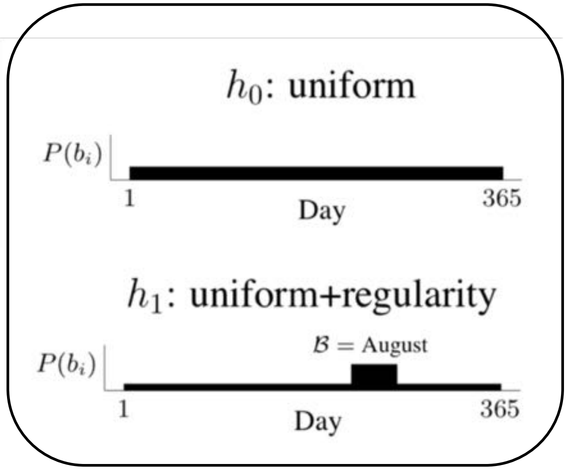
Moving the points around… Human Model

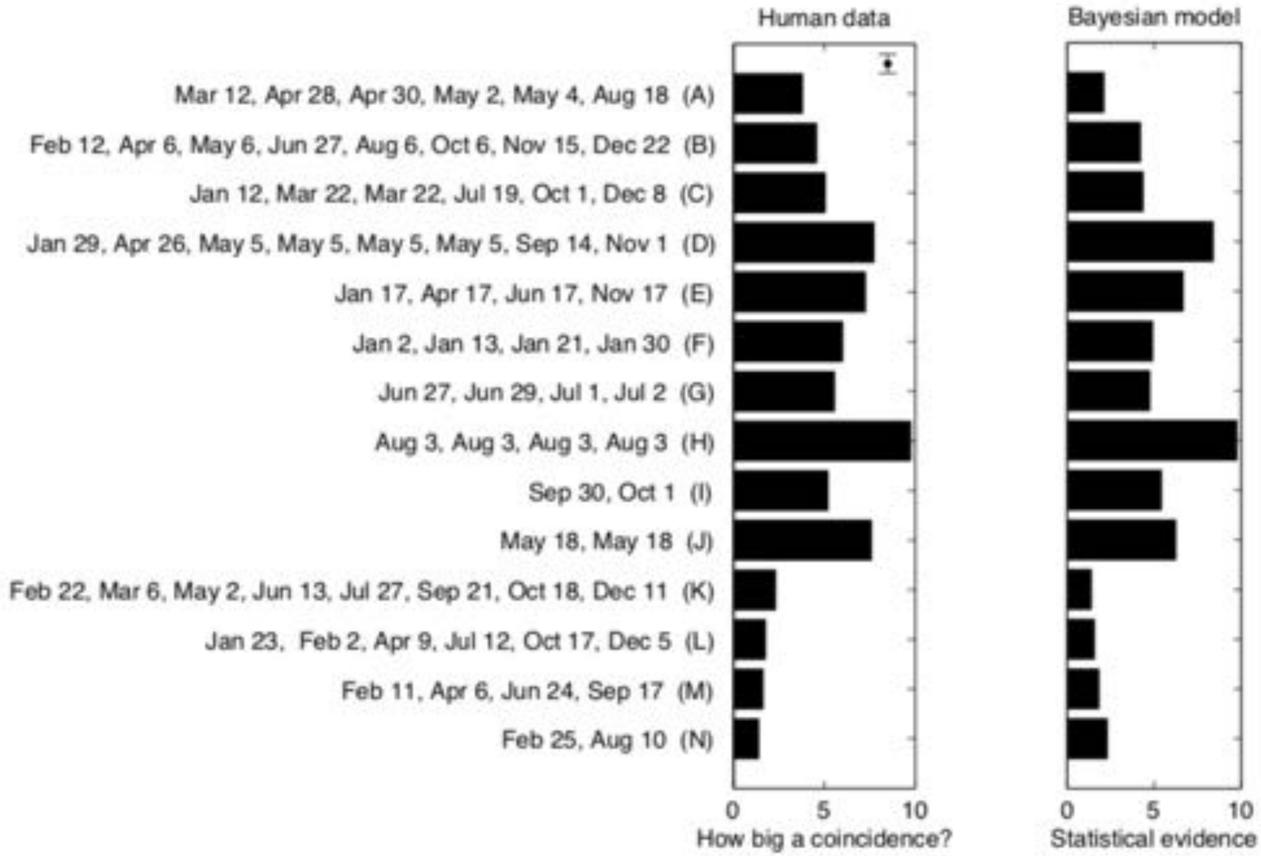


Changing the spread… Human Model



Coincidences in time





But it's complicated…

(Tauber et al 2017)

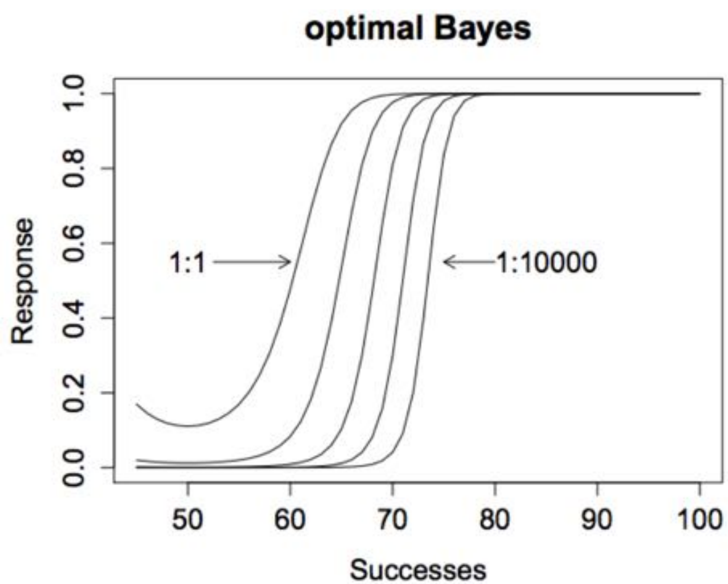
A group of scientists investigating genetic engineering have conducted a series of experiments testing drugs that influence the development of rat fetuses. All of these drugs are supposed to affect the sex chromosome: they are intended to affect whether rats are born male or female. The scientists tested this claim by producing 100 baby rats from mothers treated with the drugs. Under normal circumstances, male and female rats are equally likely to be born. The results of these experiments are shown below: The identities of the drugs are concealed with numbers, but you are given the number of times male or female rats were produced by mothers treated with each drug.

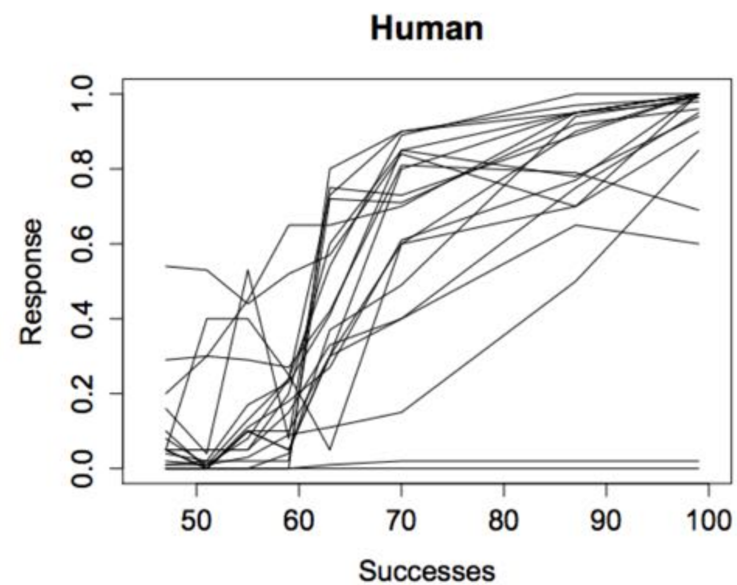
But it's complicated…

(Tauber et al 2017)

If people used the "optimal" statistical model to update data curves should look like this…

Empirical data for individual subjects are systematically flatter… we revise our beliefs <u>*more slowly*</u> when evidence arrives

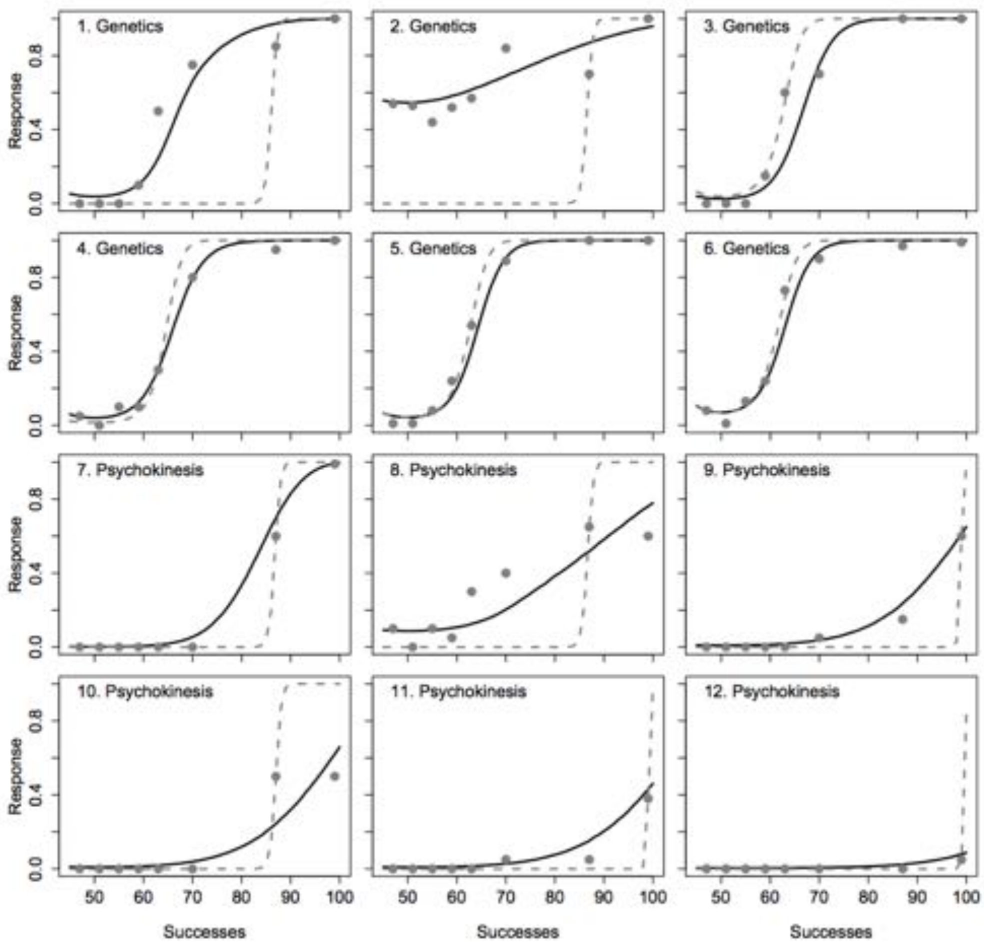




(very old phenomenon… conservatism in belief updating)

But it’s complicated…

(Tauber et al 2017)

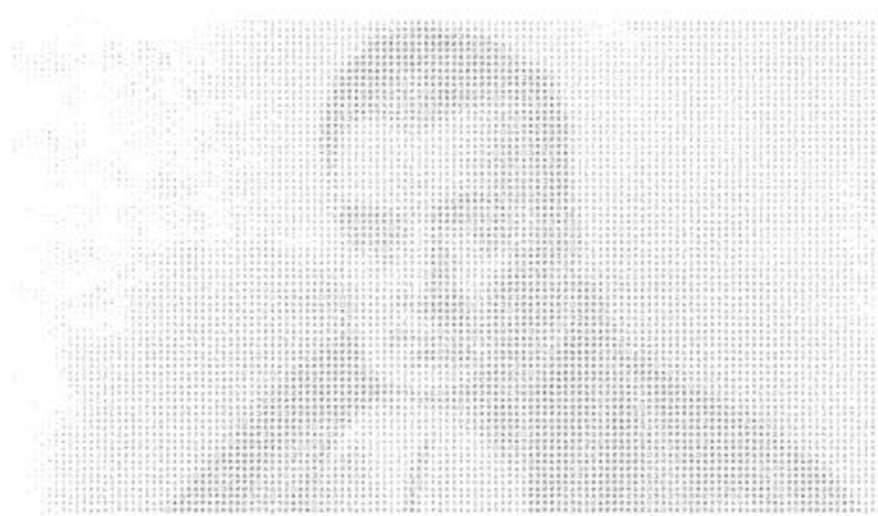


People <u>do</u> have stronger *prior biases* to believe that a “genetic” experiment works (as opposed to “psychokinesis”) but…

… we <u>*also*</u> apply a more conservative Bayesian belief revision rule when the data are at odds with our priors!

Bayesian models of cognition

Example 2: How do categories influence perception?



Bayesian perceptual magnets

(Feldman et al 2009)

We have knowledge about the perceptual categories that are used in our language

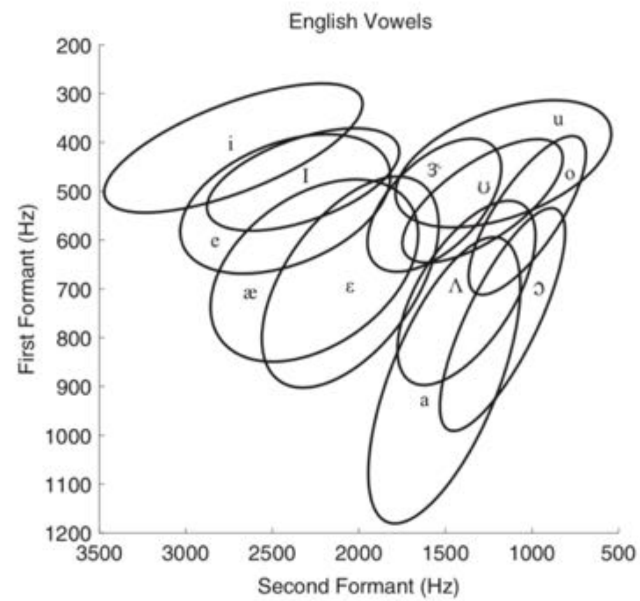
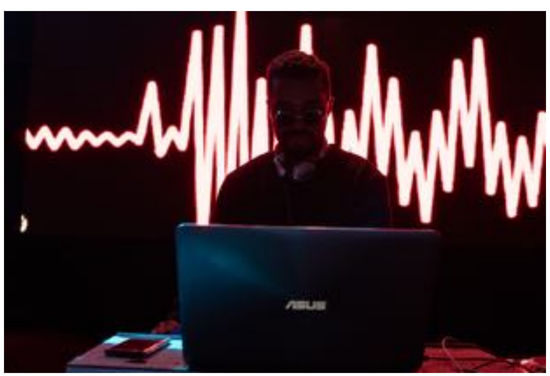


Figure 1. Map of vowel space from Hillenbrand et al.'s (1995) production experiment. Ellipses delimit regions corresponding to approximately 90% of tokens from each vowel category. Adapted from "Acoustic Characteristics of American English Vowels" by J. Hillenbrand, L. A. Getty, M. J. Clark, & K. Wheeler, 1995, *Journal of the Acoustical Society of America*, *97*, p. 3103. Copyright 1995 by the Acoustical Society of America. Reprinted with permission.

Sensory input is noisy, and it's often hard to decode speech sounds



Bayesian perceptual magnets

(Feldman et al 2009)

758 FELDMAN, GRIFFITHS, AND MORGAN

language with one phonetic category. This shrinks perceptual space in areas of unambiguous categorization. If listeners are uncertain about category membership, they should take into account all of the categories that could have generated the speech sound they heard, but they should weight the influence of each category by the probability that the speech sound came from that category. This ensures that under assumptions of equal frequency and variance, nearby categories are weighted more heavily than those farther away. Perception of speech sounds precisely on the border between two categories is pulled simultaneously toward both category means, each canceling out the other's effect. Perception of speech sounds that are near the border between categories is biased toward the most likely category, but the competing category dampens the bias. The resulting pattern for the two-category case is shown in Figure 2b.

The interaction between the categories produces a pattern of perceptual warping that is qualitatively similar to descriptions of the perceptual magnet effect and other categorical effects that have been reported in the literature. Speech sounds near category centers are extremely close together in perceptual space, whereas speech sounds near the edges of a category are much farther apart. This perceptual pattern results from a combination of two factors, both of which were proposed by Liberman et al. (1957) in reference to categorical perception. The first is acquired equivalence within categories due to perceptual bias toward category means; the second is acquired distinctiveness between categories due to the presence of multiple categories. Consistent with these predictions, infants acquiring language have shown both acquired distinctiveness for phonemically distinct sounds and acquired equivalence for members of a single phonemic category over the course of the first year of life (Kuhl et al., 2006).

Mathematical Presentation of the Model

This section formalizes the rational model within the framework of Bayesian inference. The model is potentially applicable to any perceptual problem in which a perceiver needs to recover a target from a noisy stimulus, using knowledge that the target has been sampled from a Gaussian category. We therefore present the mathematics in general terms, referring to a generic stimulus S, target T, category c, category variance σ_c^2, and noise variance σ_S^2. In the specific case of speech perception, S corresponds to the speech sound heard by the listener, T to the phonetic detail of a speaker's intended target production, and c to the language's phonetic categories; the category variance σ_c^2 represents meaningful within-category variability, and the noise variance σ_S^2 represents articulatory, acoustic, and perceptual noise in the speech signal.

The formalization is based on a generative model in which a target T is produced by sampling from a Gaussian category c with mean μ_c and variance σ_c^2. The target T is distributed as

$$T|c \sim N(\mu_c, \sigma_c^2). \quad (1)$$

Perceivers cannot recover T directly, but instead perceive a noisy stimulus S that is normally distributed around the target production with noise variance σ_S^2 such that

$$S|T \sim N(T, \sigma_S^2). \quad (2)$$

Note that integrating over T yields

$$S|c \sim N(\mu_c, \sigma_c^2 + \sigma_S^2), \quad (3)$$

indicating that under these assumptions, the stimuli that perceivers observe are normally distributed around a category mean μ_c, with a variance that is a sum of the category variance and the noise variance.

Given this generative model, perceivers can use Bayesian inference to reconstruct the target from the noisy stimulus. According to Bayes' rule, given a set of hypotheses H and observed data d, the posterior probability of any given hypothesis h is

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h \in H} p(d|h)p(h)}, \quad (4)$$

indicating that it is proportional to both the likelihood $p(d|h)$, which is a measure of how well the hypothesis fits the data, and the prior $p(h)$, which gives the probability assigned to the hypothesis before any data were observed. Here, the stimulus S serves as data d; the hypotheses under consideration are all the possible targets T; and the prior $p(h)$, which gives the probability that any particular target will occur, is specified by category structure. In laying out the solution to this statistical problem, we begin with the case in which there is a single category and then move to the more complex case of multiple categories.

One Category

Perceivers are trying to infer the target T given stimulus S and category c, so they must calculate $p(T|S, c)$. They can use Bayes' rule:

$$p(T|S, c) \propto p(S|T)p(T|c). \quad (5)$$

The likelihood $p(S|T)$, given by the noise process (Equation 2), assigns highest probability to stimulus S, and the prior $p(T|c)$, given by category structure (Equation 1), assigns highest probability to the category mean. As described in Appendix A, the right-hand side of this equation can be simplified to yield a Gaussian distribution

$$p(T|S, c) = N\left(\frac{\sigma_c^2 S + \sigma_S^2 \mu_c}{\sigma_c^2 + \sigma_S^2}, \frac{\sigma_c^2 \sigma_S^2}{\sigma_c^2 + \sigma_S^2}\right) \quad (6)$$

whose mean falls between the stimulus S and the category mean μ_c.

This posterior probability distribution can be summarized by its mean (the expectation of T given S and c),

$$E[T|S, c] = \frac{\sigma_c^2 S + \sigma_S^2 \mu_c}{\sigma_c^2 + \sigma_S^2}. \quad (7)$$

The optimal guess at the target, then, is a weighted average of the observed stimulus and the mean of the category that generated the stimulus, where the weighting is determined by the ratio of cate-

Blah blah blah lots of fancy maths because they are smart

Short version:

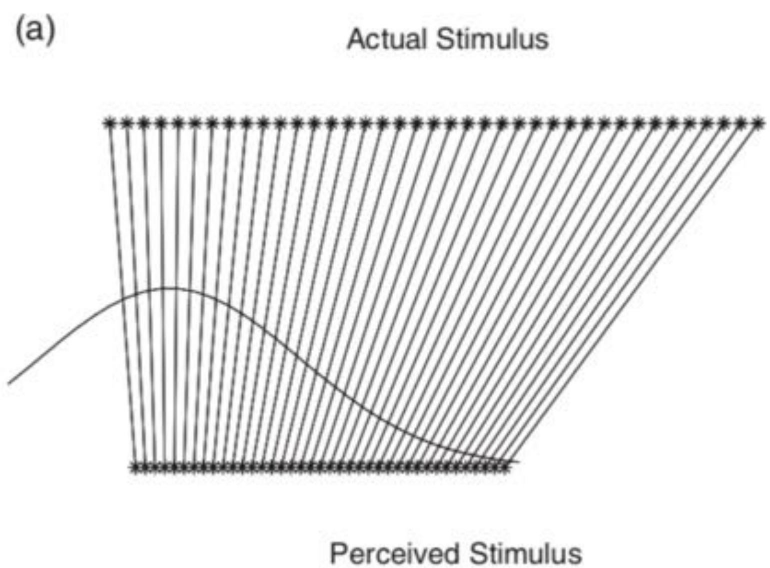
- Knowledge about the perceptual/linguistic categories supplies a prior P(h) for what the possible speech sound could have been
- Sensory system supplies the likelihood P(d|h) that we would receive this input given any speech sound

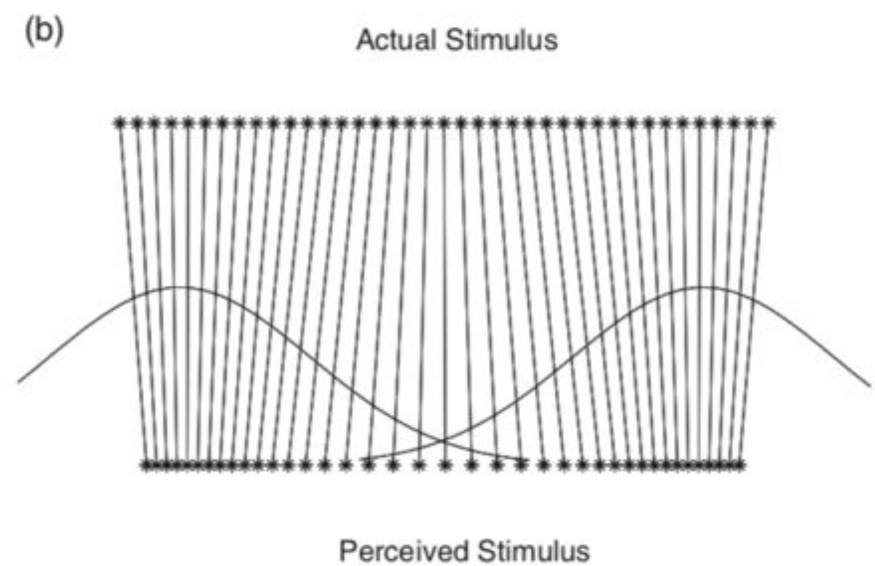
Bayesian perceptual magnets

(Feldman et al 2009)

The categorical knowledge shapes the perceived sound…

The predicted distortion pattern depends on the locations of the categories…



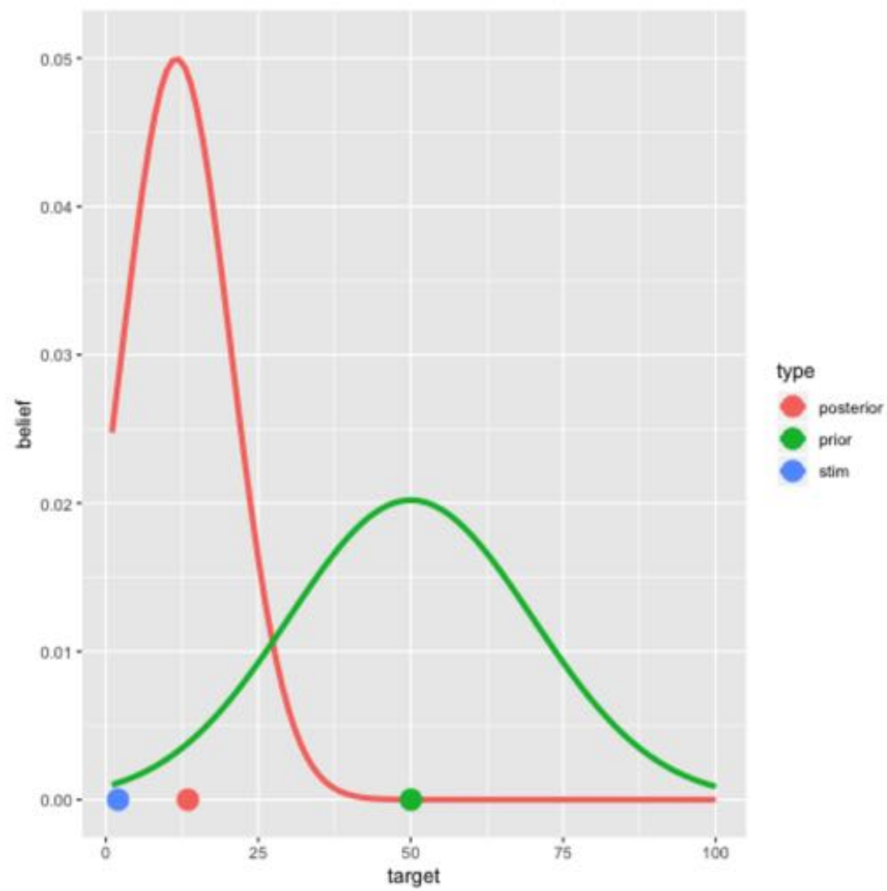


Bayesian perceptual magnets

(Feldman et al 2009)

Example 1:

Moving the stimulus relative to the category

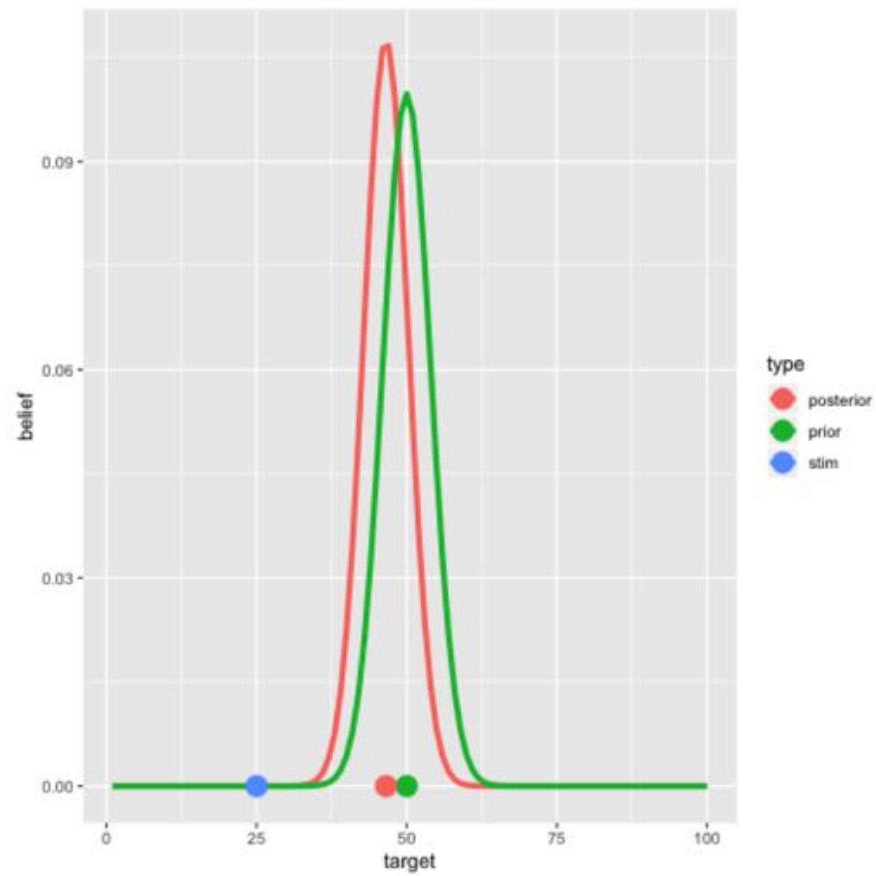


Bayesian perceptual magnets

(Feldman et al 2009)

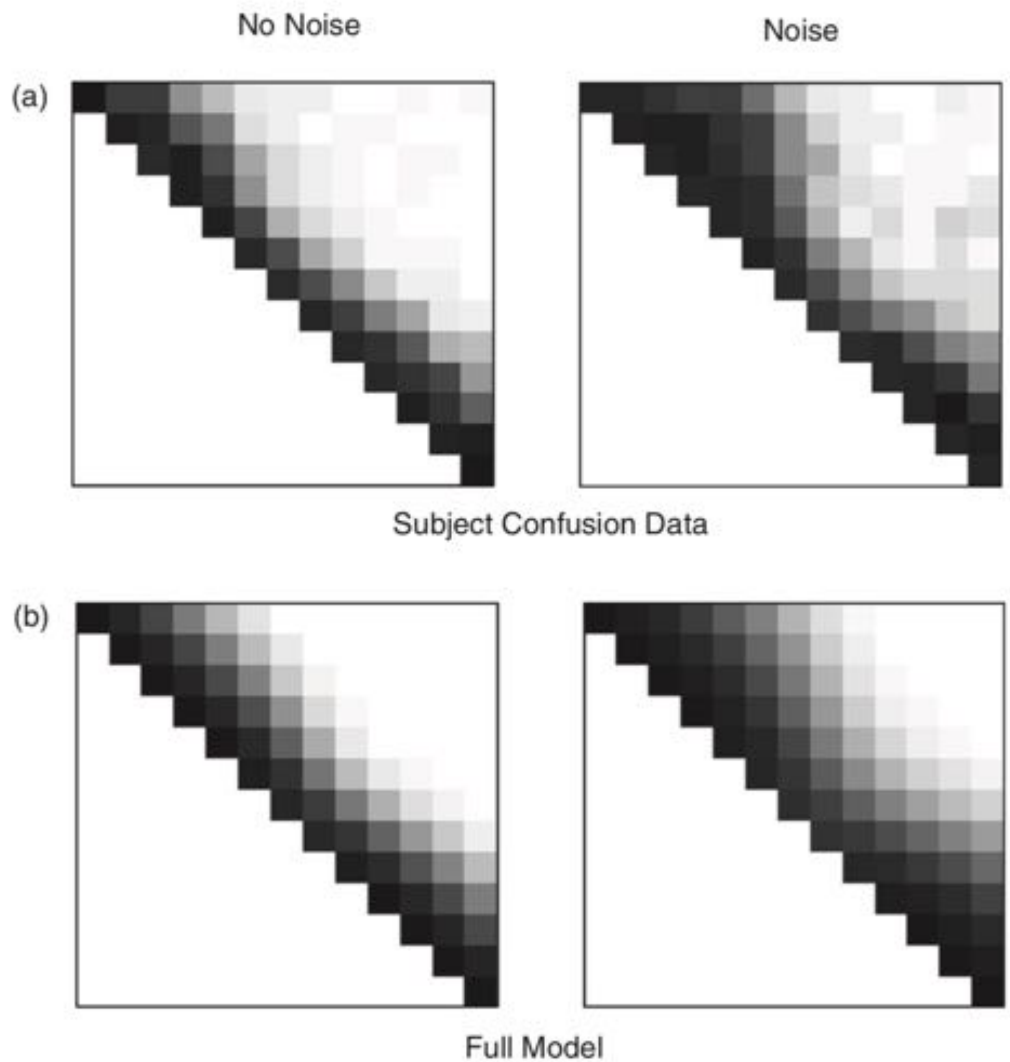
Example 2:

Changing the strength of prior knowledge relative to the noise in the environment



Bayesian perceptual magnets

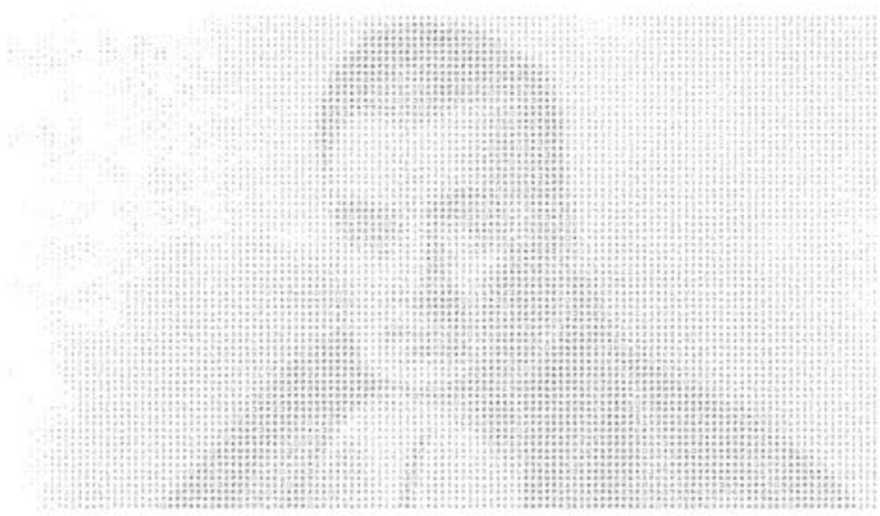
(Feldman et al 2009)



The perceptual magnet effect is strongest in moderately noisy environments, roughly in accordance with model predictions

(Needs to be clean enough that you can work out what the category is supposed to be but not so noisy that you can't hear anything)

Connecting Bayesian cognitive models with Bayesian machine learning



The structure problem

Left: A man is holding a dog in his hand
Right: A woman is holding a dog in her hand
Image: @SouperSarah

NeuralTalk2: A flock of birds flying in the air
Microsoft Azure: A group of giraffe standing next to a tree
Image: Fred Dunn, https://www.flickr.com/photos/gratapictures - CC-BY-NC

A goat being held by a child is labelled a "dog"

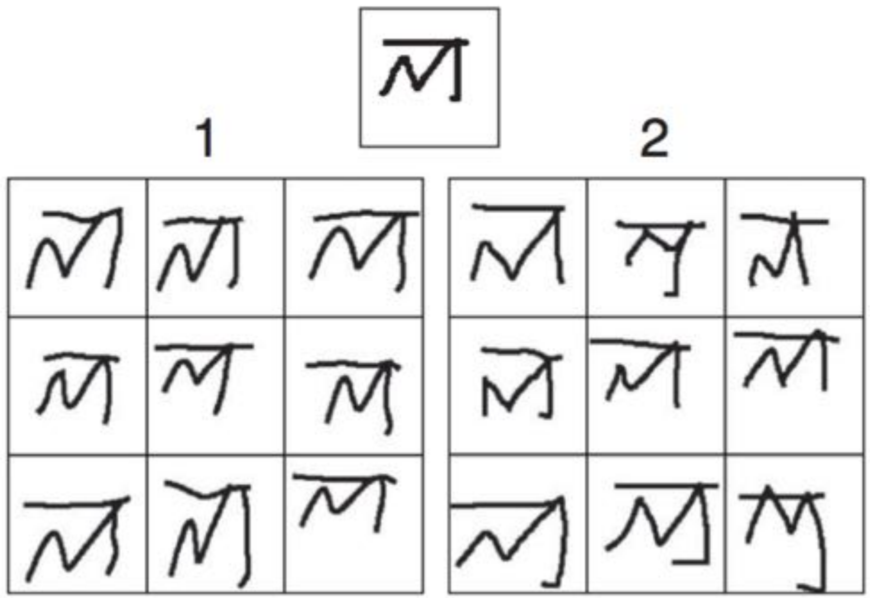
Goats in trees become birds or giraffes

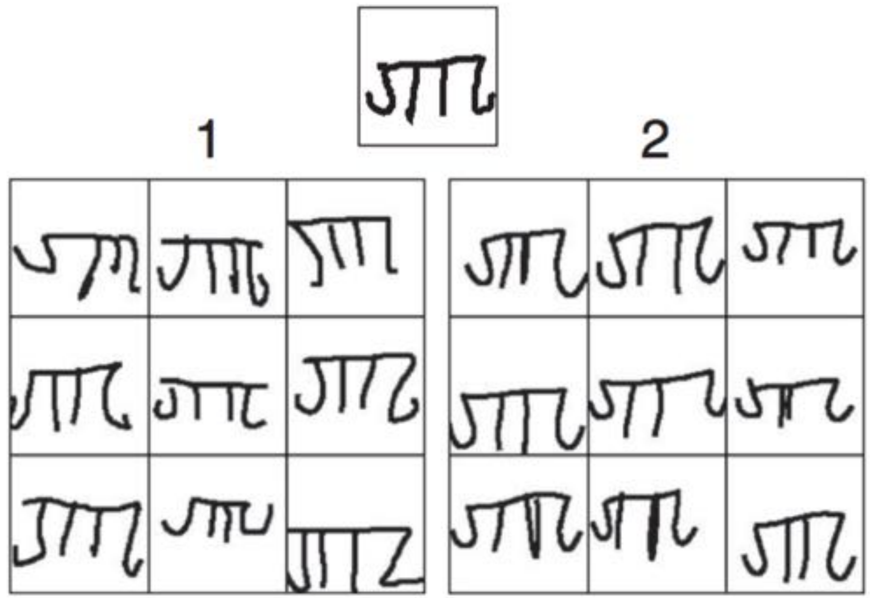
The structure problem

- Even though it is comparatively simple, this is still a structured object.
- It has distinct parts, they are related to one another
- There is a production method (writing) that tells you what the relations are
- Human reasoning about these concepts exploits this knowledge
- How do we build theories that do that?

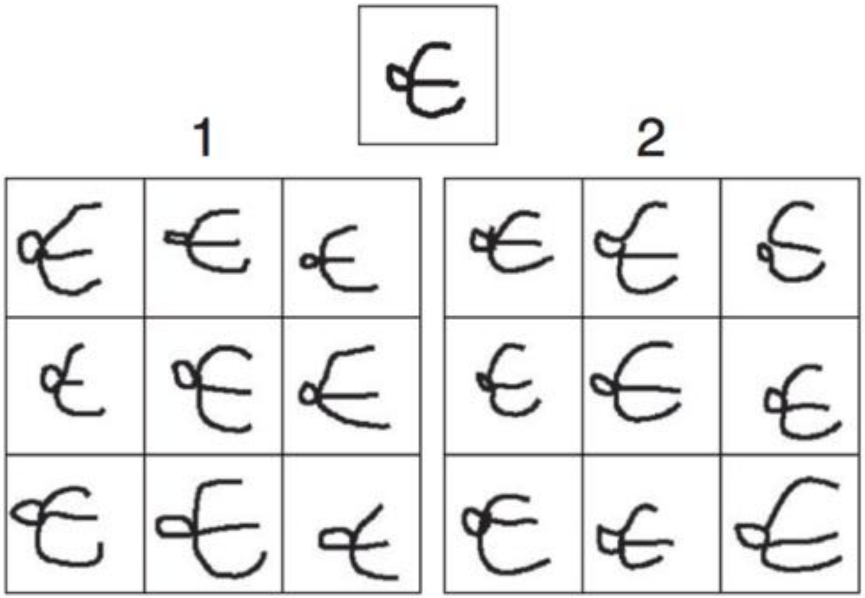
Human level concept learning with "Bayesian program induction"

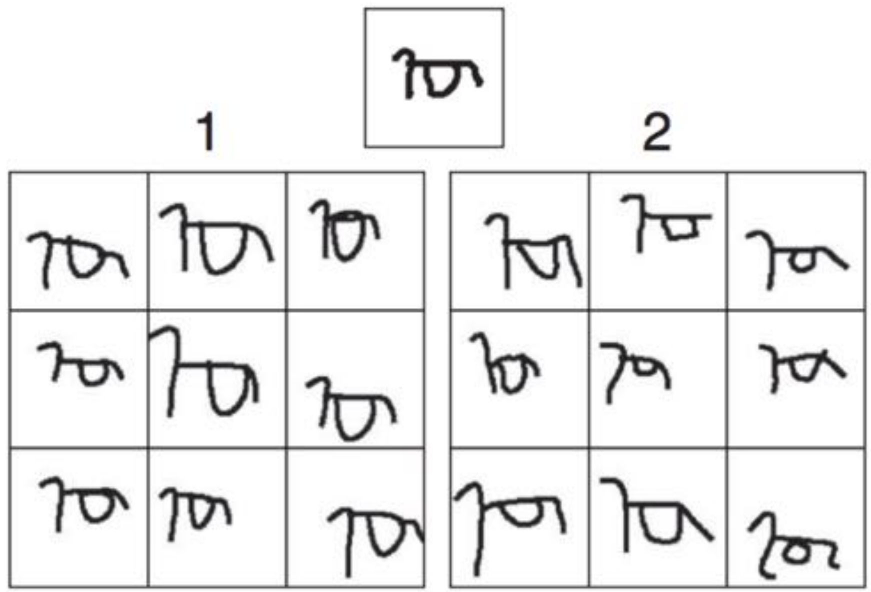
(Lake et al 2015)

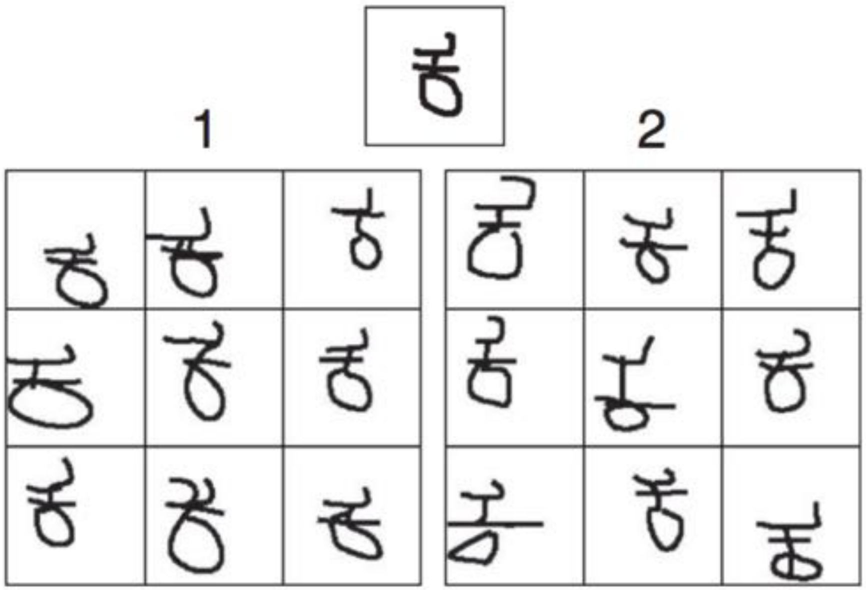


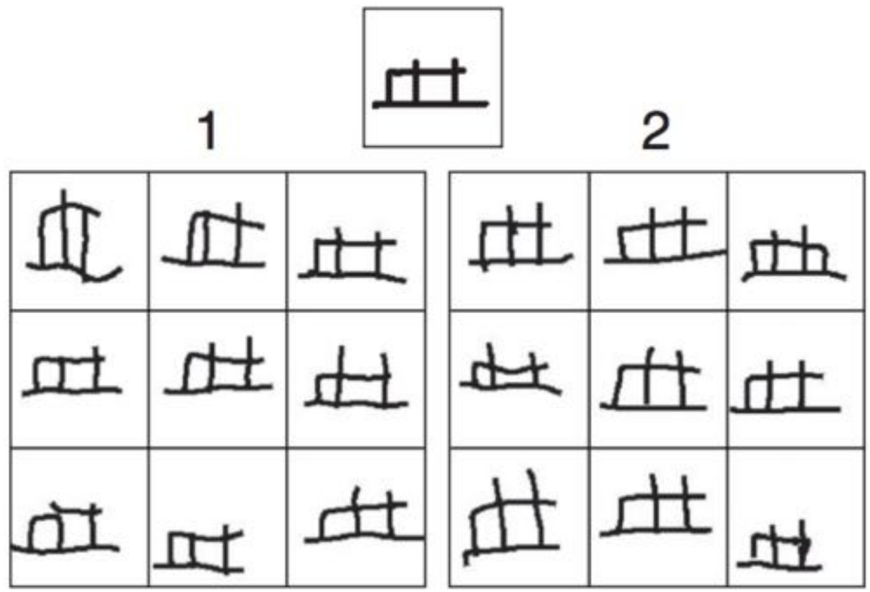


1
2


1
2


1
2


1
2


1
2

A library of visual concepts

Fig. 2. Simple visual concepts for comparing human and machine learning. 525 (out of 1623) character concepts, shown with one example each.

A generative "language" for characters

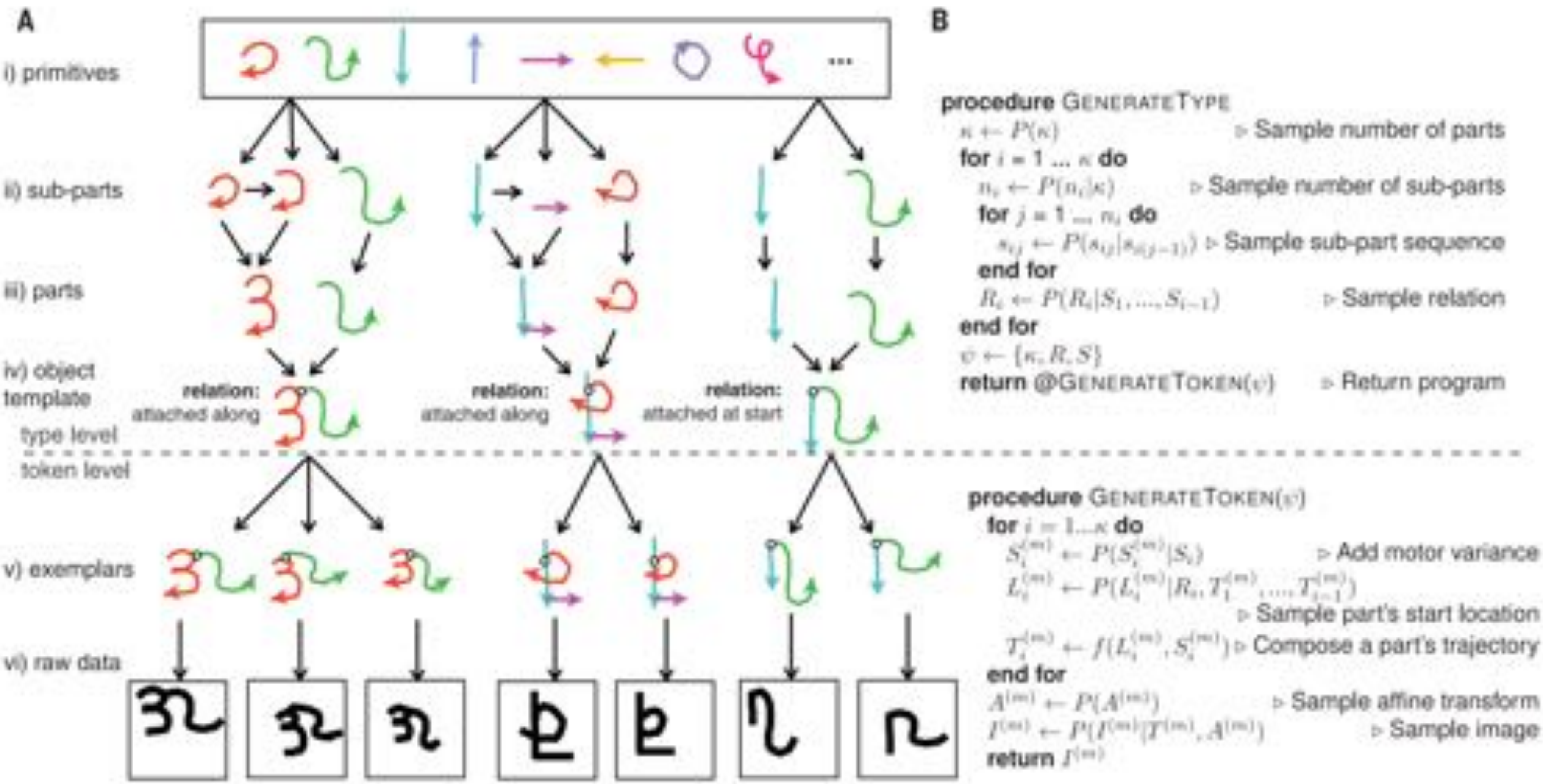
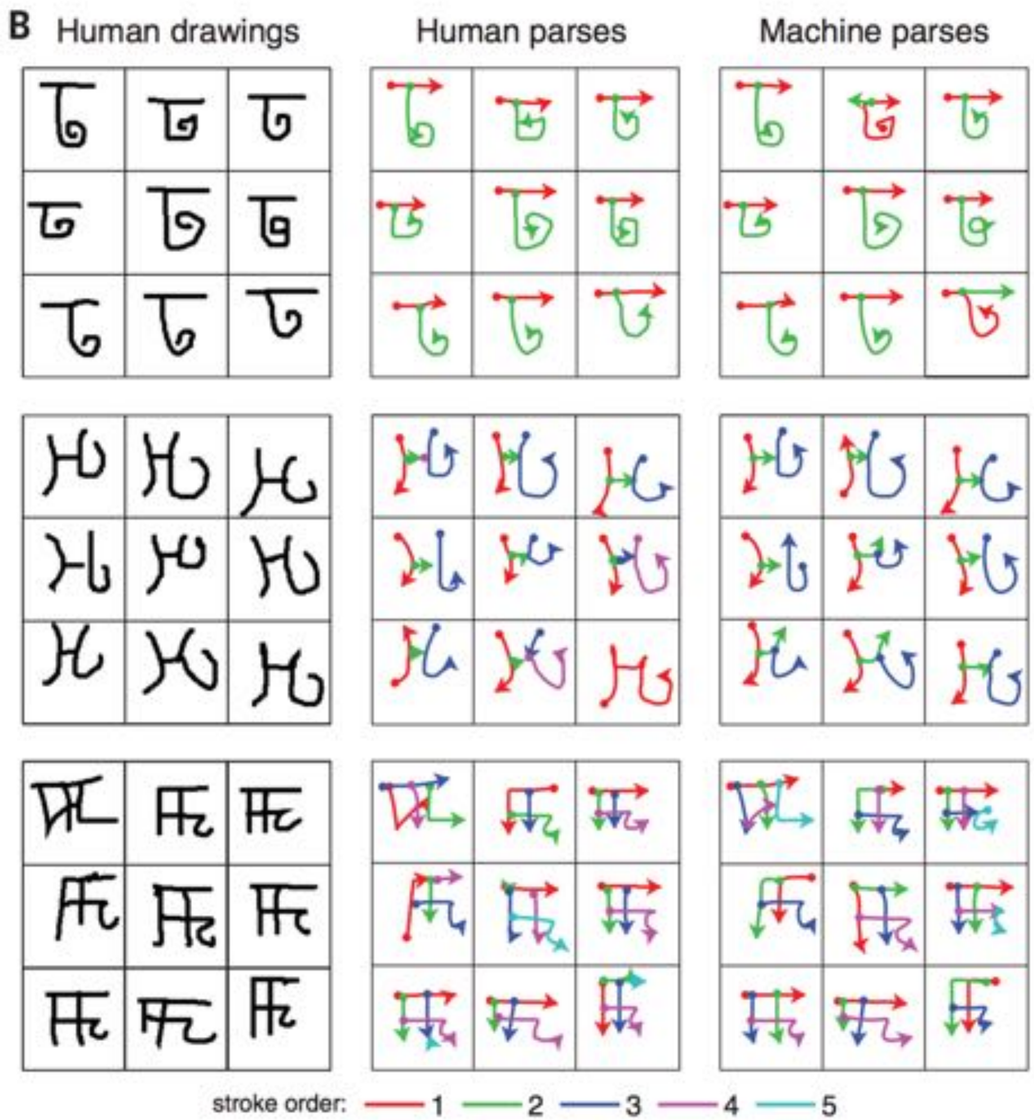


Fig. 3. A generative model of handwritten characters. (**A**) New types are generated by choosing primitive actions (color coded) from a library (i), combining these subparts (ii) to make parts (iii), and combining parts with relations to define simple programs (iv). New tokens are generated by running these programs (v), which are then rendered as raw data (vi). (**B**) Pseudocode for generating new types ψ and new token images $I^{(m)}$ for $m = 1, \ldots, M$. The function $f(\cdot, \cdot)$ transforms a subpart sequence and start location into a trajectory.

Grammar allows structure learning



Structure allows smart generalization!

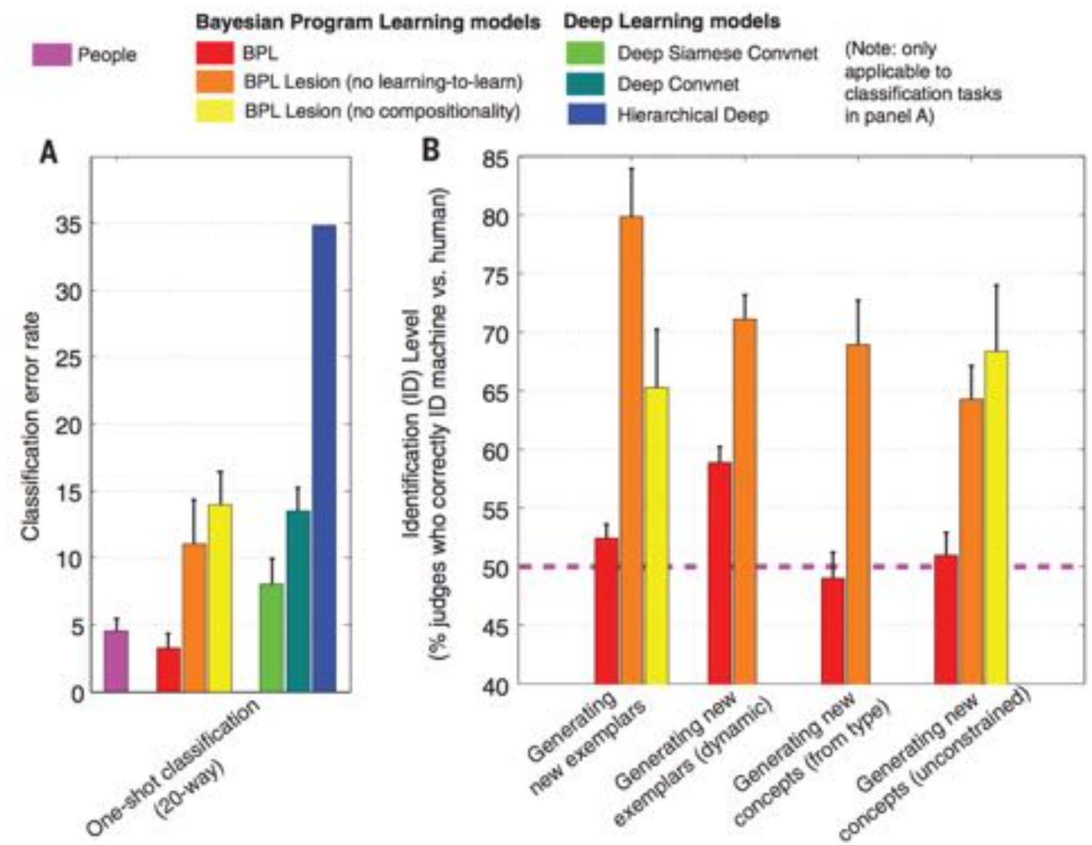


Fig. 6. Human and machine performance was compared on (A) one-shot classification and (B) four generative tasks. The creative outputs for humans and models were compared by the percent of human judges to correctly identify the machine. Ideal performance is 50%, where the machine is perfectly confusable with humans in these two-alternative forced choice tasks (pink dotted line). Bars show the mean ± SEM [N = 10 alphabets in (A)]. The no learning-to-learn lesion is applied at different levels (bars left to right): (A) token; (B) token, stroke order, type, and type.

Thanks!