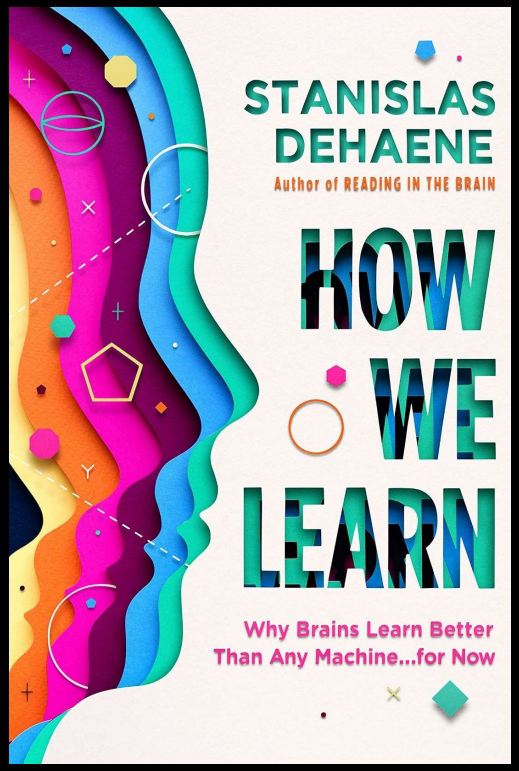
Can current neural networks provide a satisfactory model of the human brain ?

Why a “language of thought” is needed

Stanislas Dehaene
Collège de France
INSERM-CEA
Cognitive Neuroimaging Unit
NeuroSpin Center, Saclay, France
www.unicog.org



Comparing brains with artificial neural networks

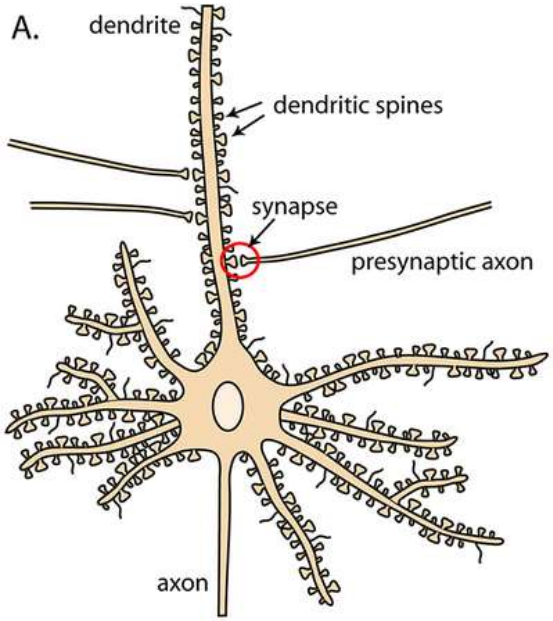
As models of the brain, current artificial neural networks are quite good at

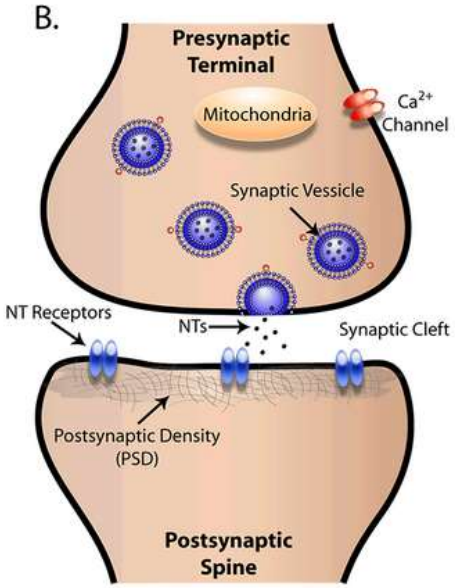
- Capturing how synaptic changes can leading to sophisticated learning
- Modelling the early stages of vision
- Modelling some aspects of language processing

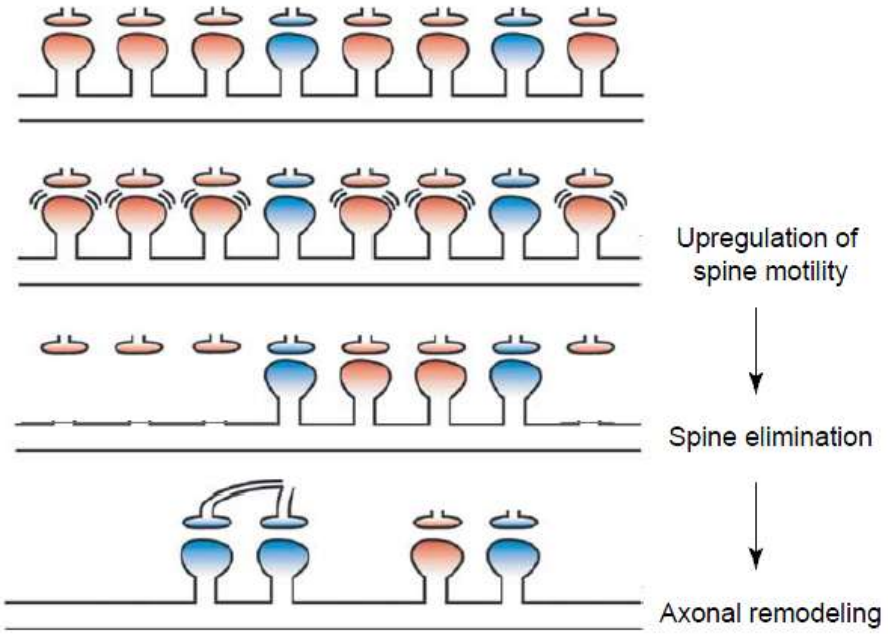
The human brain keeps the upper hand in its ability to

- learn from a very small number of examples, sometimes a single trial, using Bayesian-style reasoning (“the child as a scientist”);
- discover compact, abstract, symbolic, explicit representations of knowledge
- in a form which can be shared with others
- learn from others and learn with others
- learn compositional representations in a “language of thought”.

The main dogma of neuroscience: we learn by modifying synapses







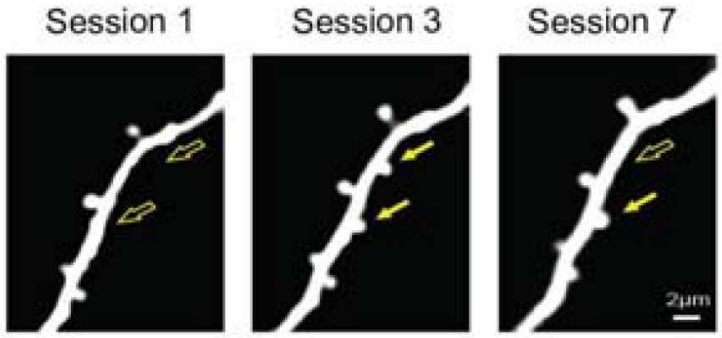
Smrt & Zhao. Frontiers in Biology 2010

Many experiments show that learning rests primarily on the reinforcement and selective elimination of synapses, which form a memory trace of our experiences and affect the tuning of our neurons.

Neuronal activity (or its absence) selectively modulate synapse stability.

Synapses can rearrange on a fast time scale: dendritic spines come and go !

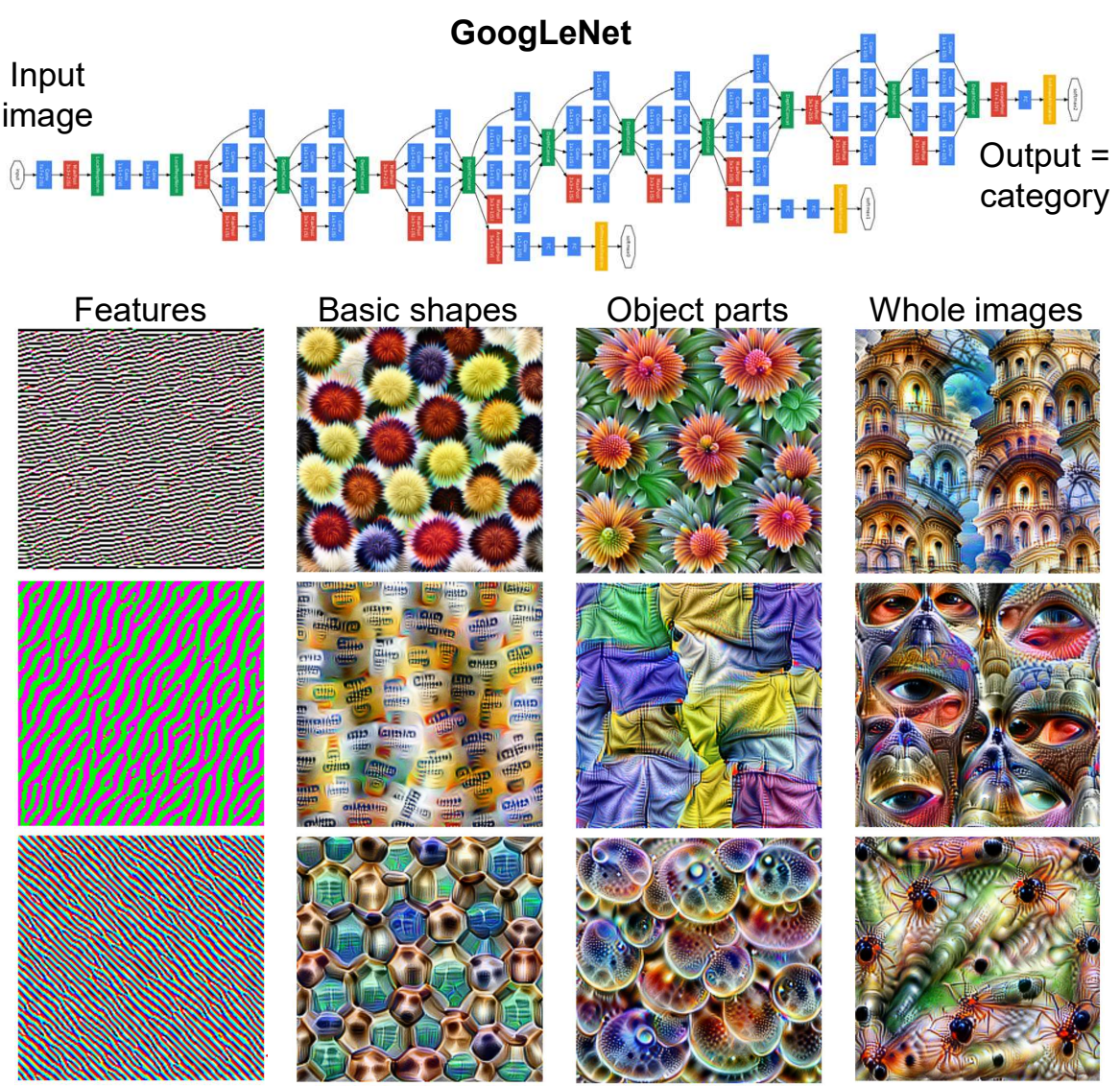
Learning also rests on changes in axonal branching, myelination, and even cell internal parameter changes (e.g. Hesslow's work in Purkinje cells)



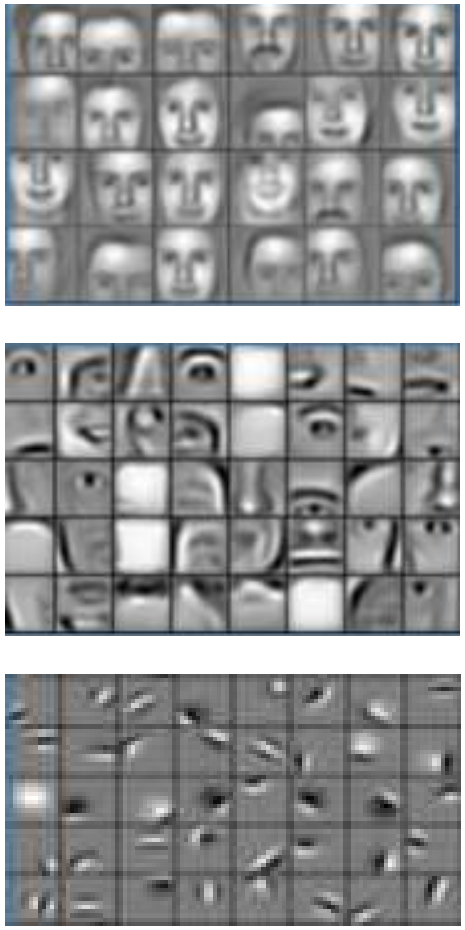
To learn

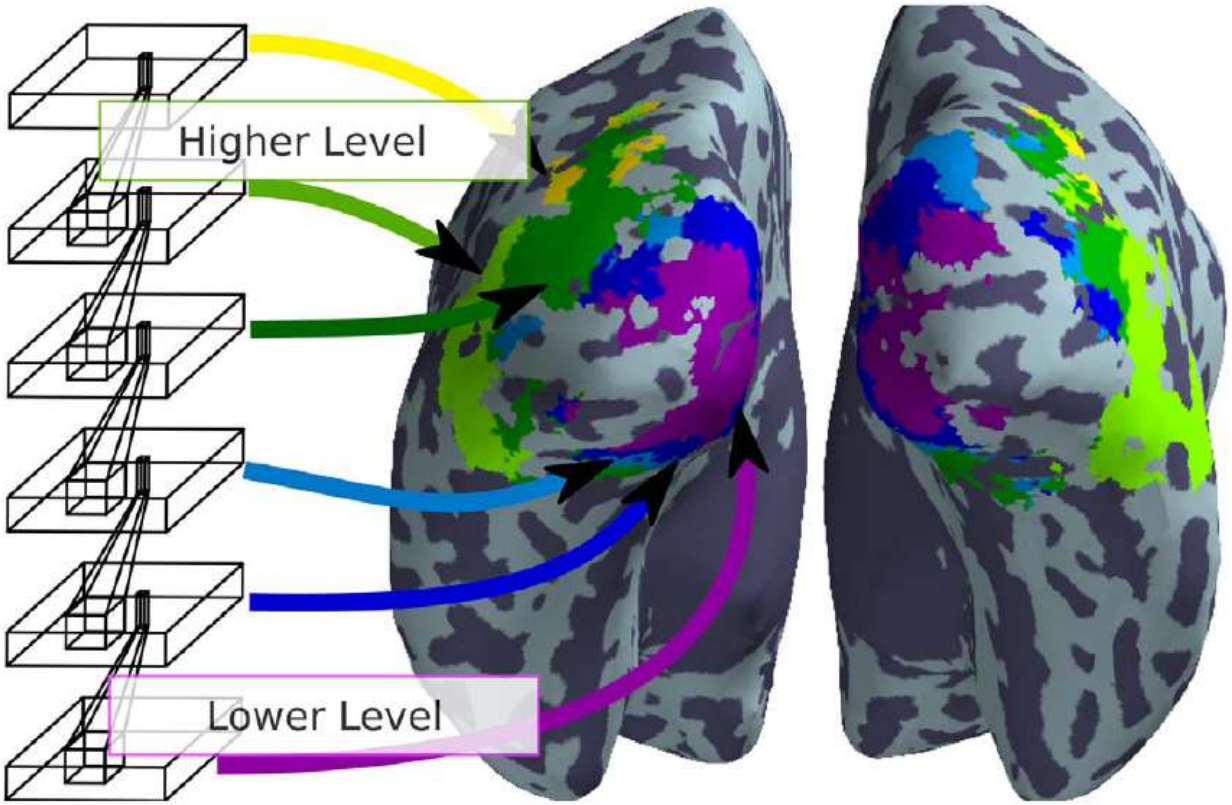
is

to adjust the millions of synapses that allow each neuron to « tune in » to a relevant aspect of the stimulus



Artificial neural networks are beginning to capture the first stages of the hierarchy of primate visual areas



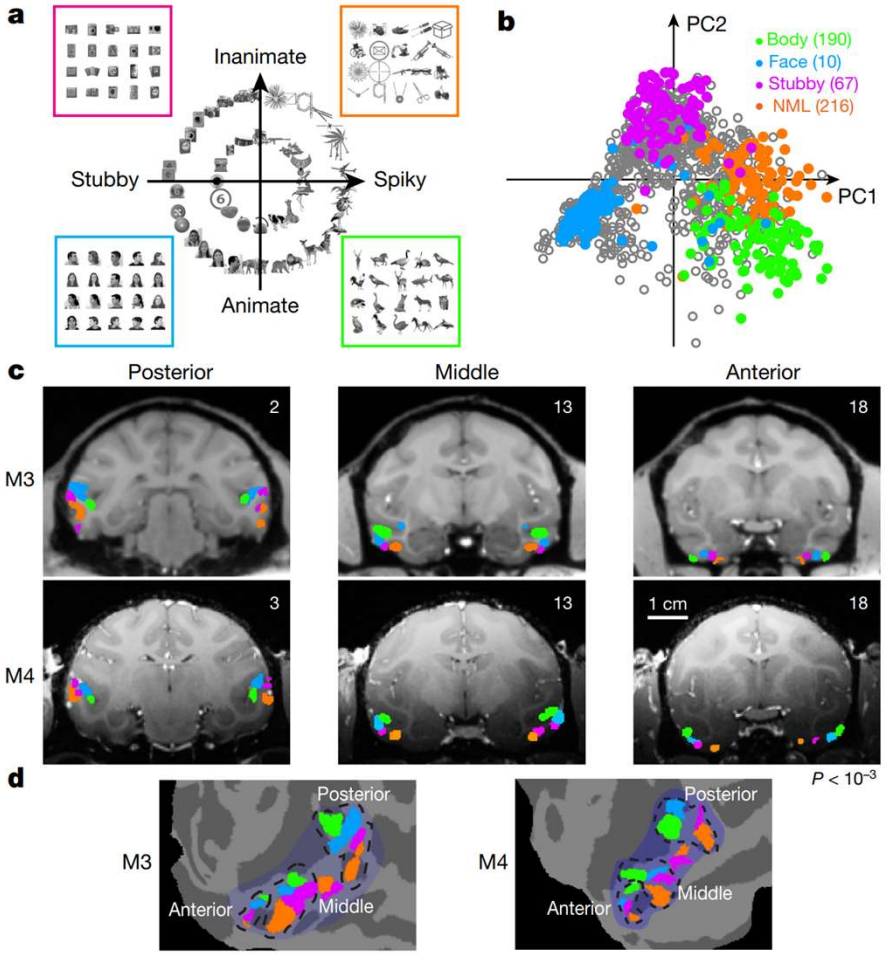


Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all : Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184-194. https://doi.org/10.1016/j.neuroimage.2016.10.001

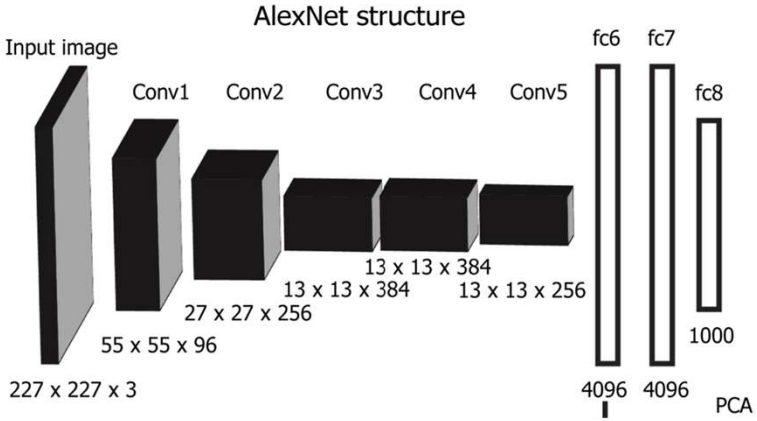
Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, *111*(23), 8619-8624.

The principal components of AlexNet explain the topography of monkey IT cortex

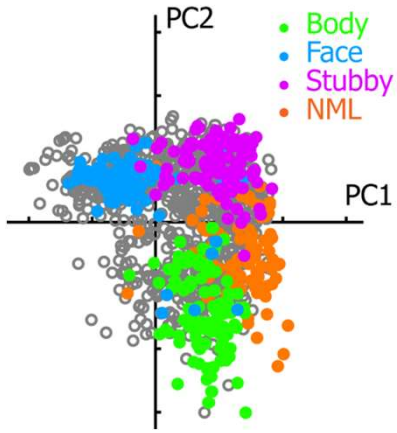
Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 1-6.



The monkey infero-temporal cortex contains several patches of neurons specialized for faces, but also other categories.



This topography can be explained by the first two principal components of the hidden units of a convolutional neural network for object recognition.



W*ithin* the face patches, the principal components *of faces* can explain the responses of *single-neurons* : each neuron is tuned to a small set of PCs.

→ Perhaps the cortex performs principal componant analysis at multiple scales.

Recurrent neural networks explain the dynamics of brain activity during object recognition

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854-21863. https://doi.org/10.1073/pnas.1905544116

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, *16*(10), e1008215. https://doi.org/10.1371/journal.pcbi.1008215

Different CNNs, with or without recurence, are compared.

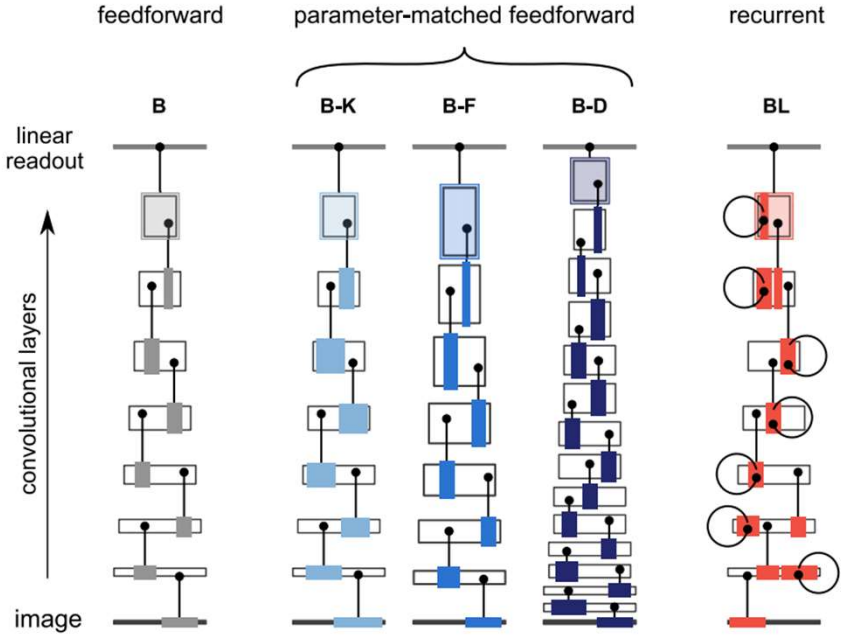
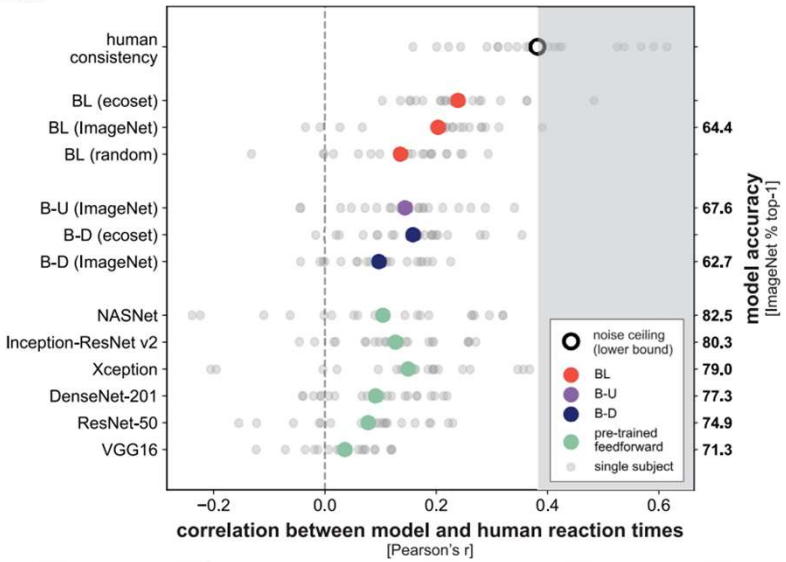
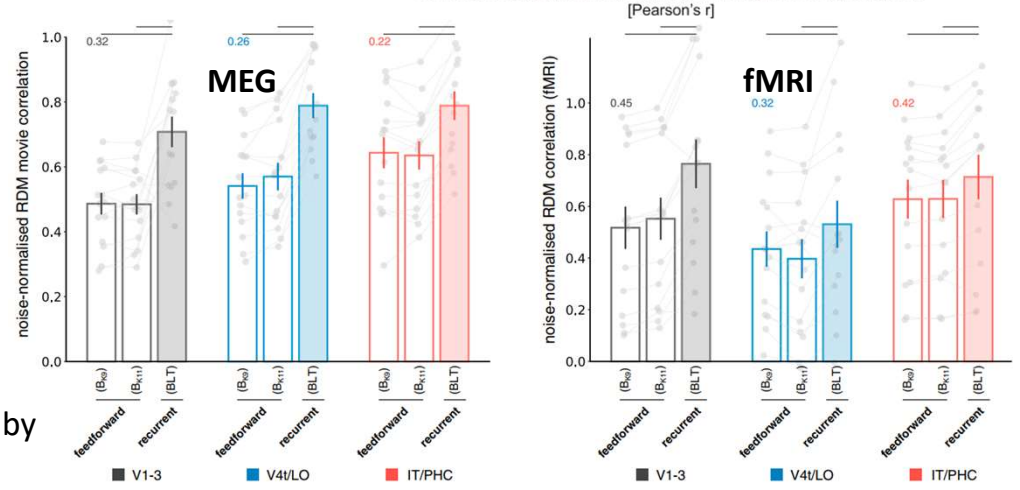


Fig 1. Schematic representation of the parameter-matched networks. White boxes represent convolutional layers,

For the same number of parameters, a recurrent network provides a better predictor of human performance in image recognition

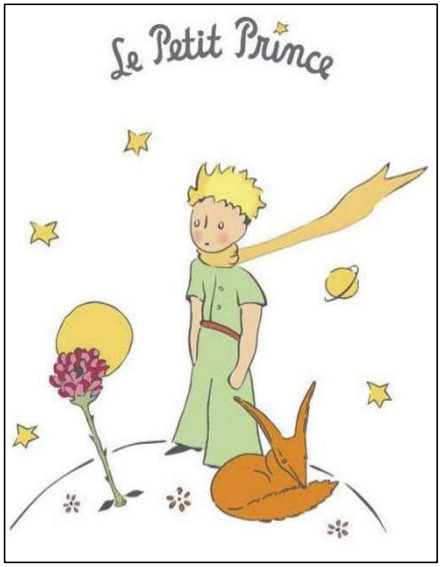




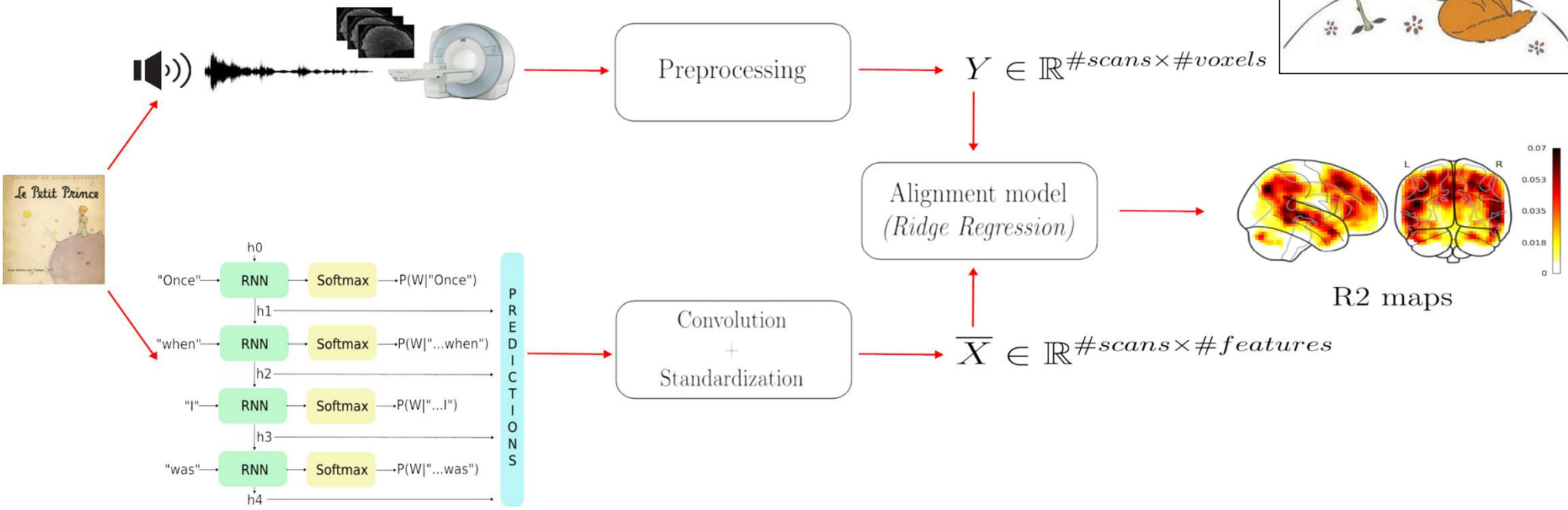
The representational similarity matrices are also better captured by the recurrent CNN, at all levels of the visual system.

The « little prince » project

Christophe Pallier, Alexandre Pasquiou, with Bertrand Thirion et al.

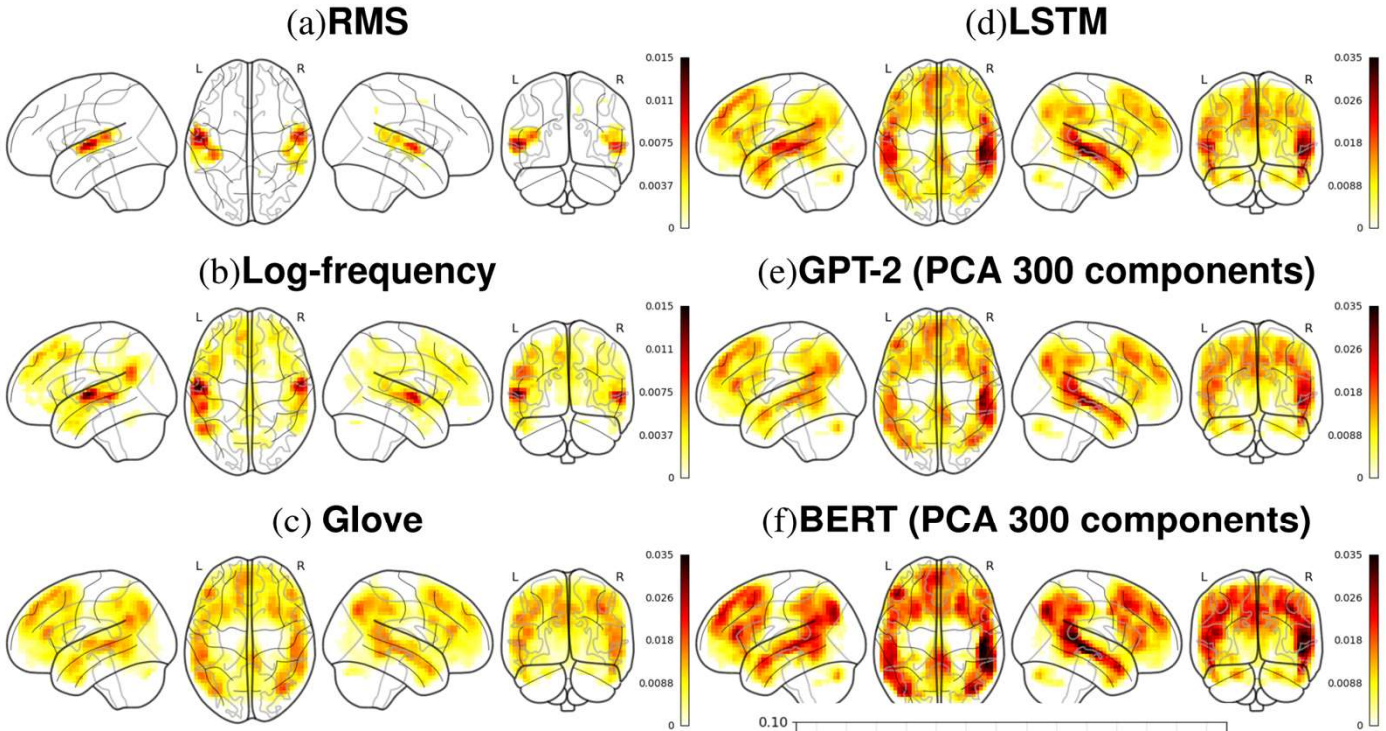


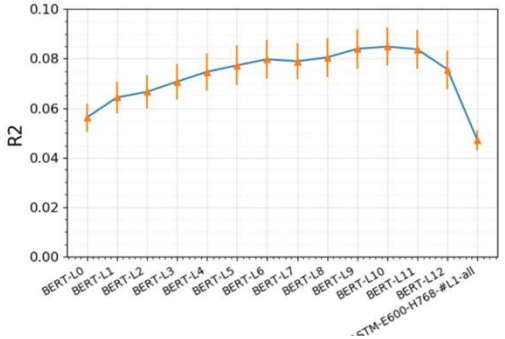
To what extent can the same trick be applied to language processing?
Idea: generate a reference data set that can be compared to various models of language processing.
→ fMRI data on more than a hundred subjects listening to the entire « Petit Prince » in English, French, or Chinese.



The « little prince » project

BERT shows a significant advantage in the prediction of higher-level language areas





Only the deeper layers of BERT can best predict the activation of Broca's area (pars triangularis)

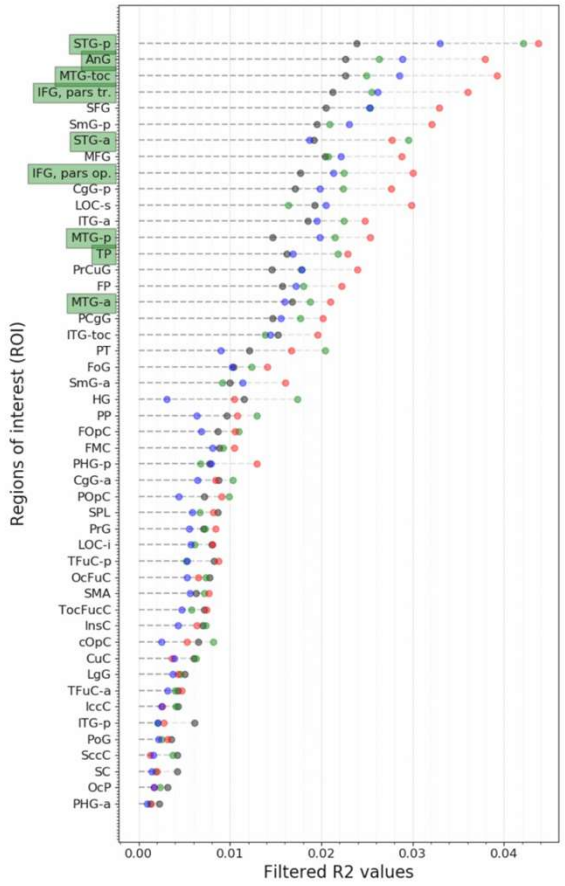
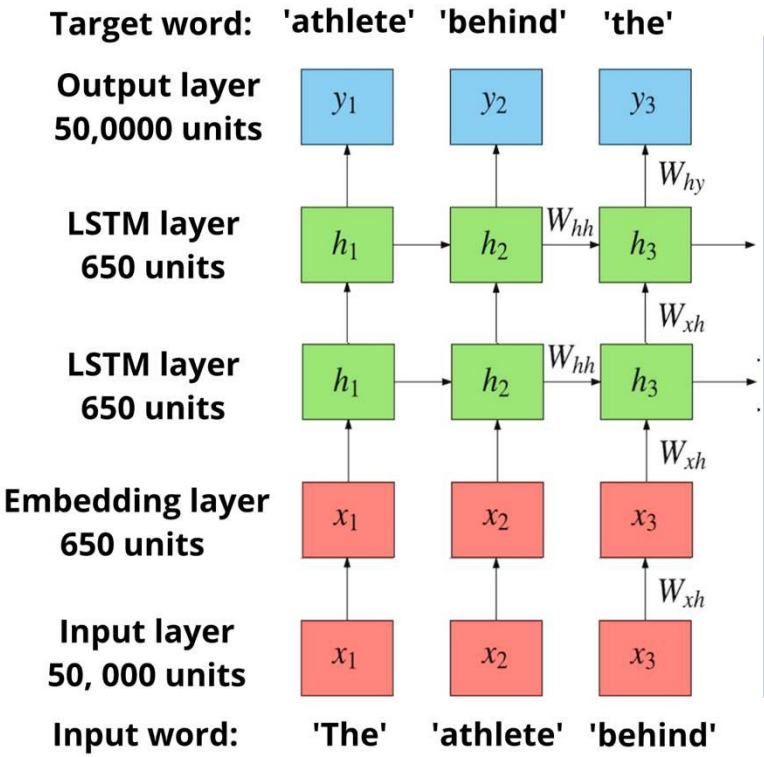


Figure 3: R^2 by Regions of the Harvard-Oxford Atlas, Models: ● BERT (pca) ● GPT-2 (pca) ● GloVe ● LSTM Colored ROIs represent important Language ROIs.

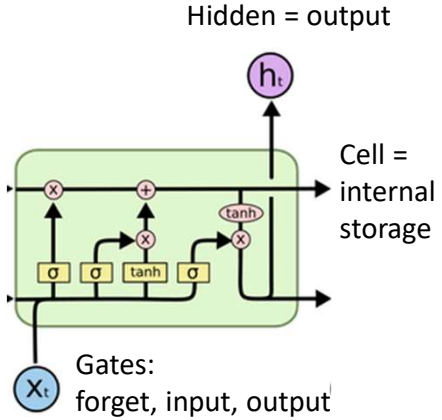
How does a recurrent neural network encode syntax ?

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019).
The emergence of number and syntax units in LSTM language models. NAACL-2019.

We analyzed a state-of-the-art long-short-term memory (LSTM) artificial network, trained to predict the next word in the Wikipedia English corpus (classical « Language model »).



Each unit of the LSTM correspond to a micro-circuit that learns to hold information on line



We test the network with a long-distance agreement task:

e.g. « The keys to the cabinet are blue »
« The cars that pass the truck are blue »

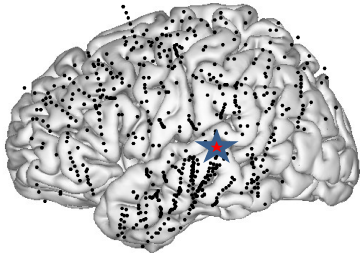
This task requires encoding
1. grammatical number information
2. enough syntactic structure to skip over intervening items (prepositional phrases, relatives, etc)

→ **Capture long-range syntactic dependencies**

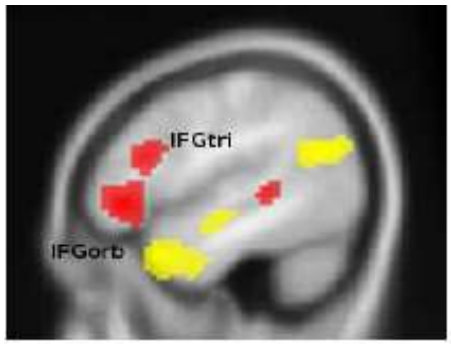
LSTM networks do relatively well in such tasks (Linzen et al., 2016; Gulordava et al., 2018)

but how?

Brain activity closely tracks phrase structures

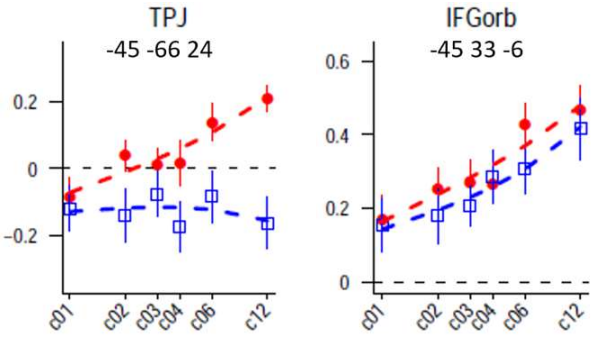


fMRI:
Monotonic increase
with constituent size
(Pallier, Devauchelle & Dehaene, 2011)

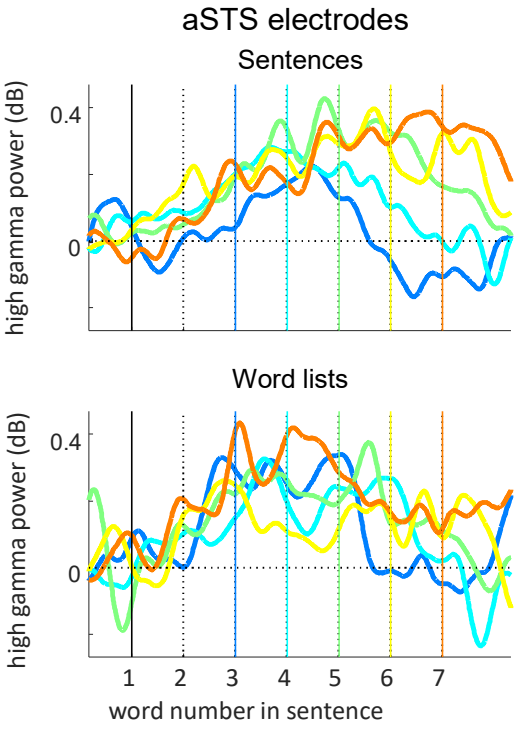


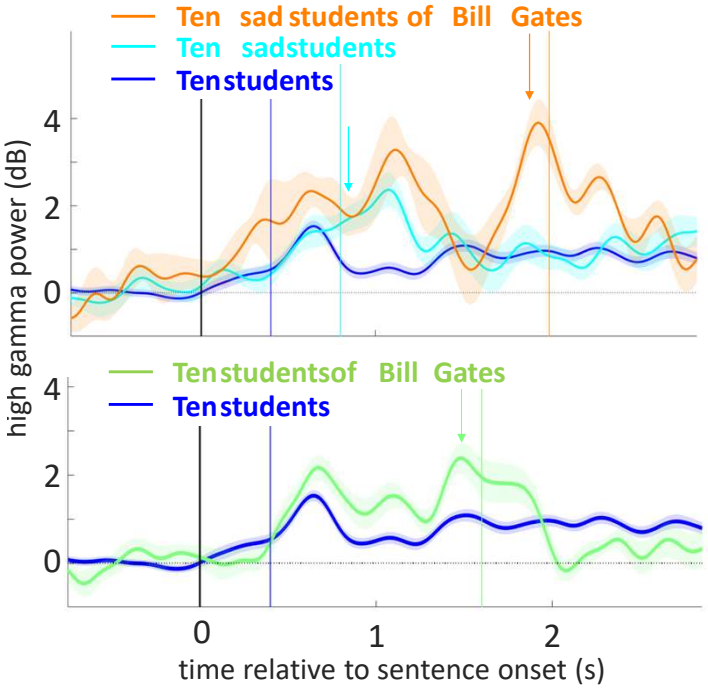
Areas in yellow increase only for sentences

Areas in red increase for both sentences and Jabberwocky

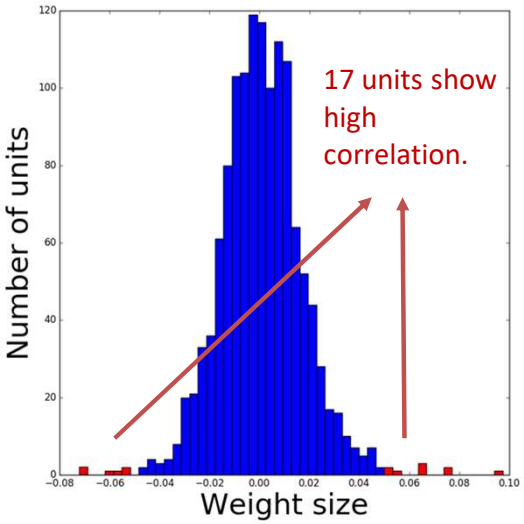


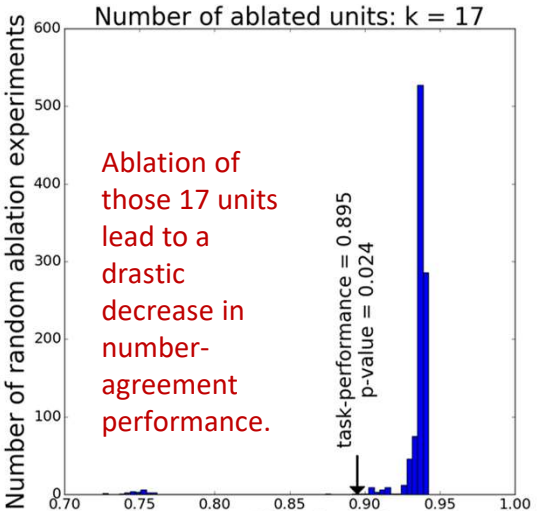
Intracranial recordings:
Monotonic increase with sentence length,
and tracking of constituent size
(Nelson… and Dehaene, PNAS, 2017)

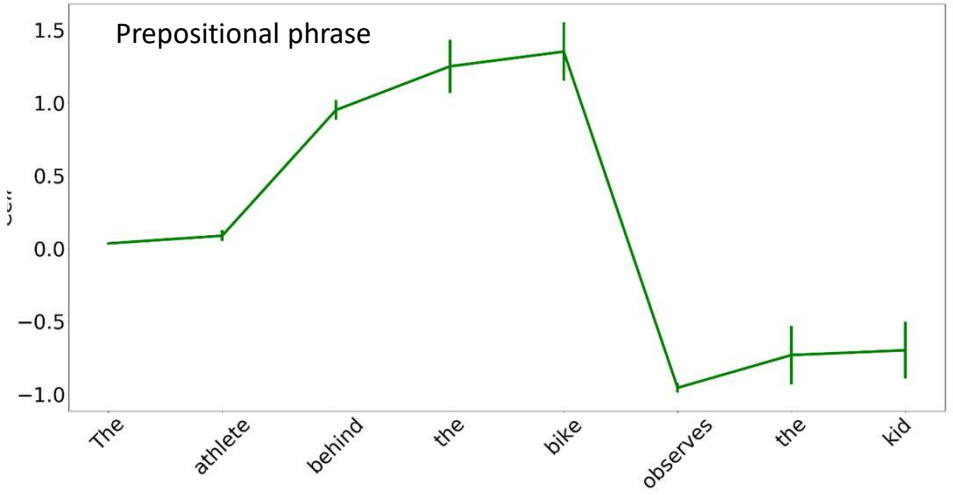




Identifying syntax units

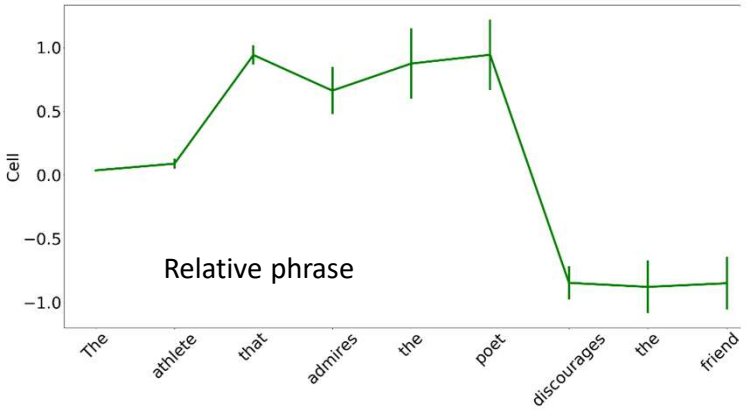


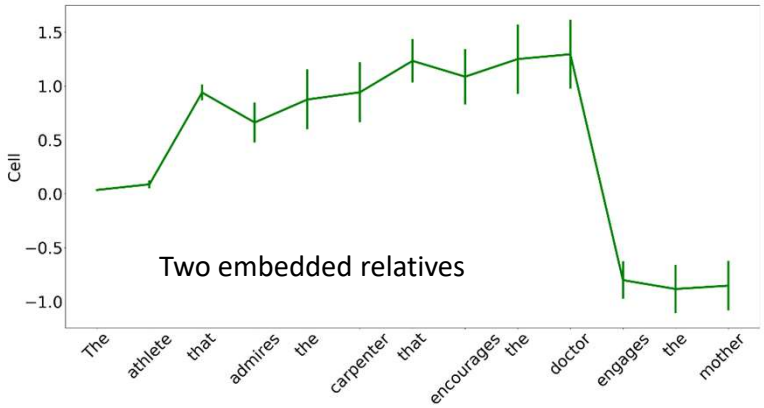




Such units signal embedded phrases and their complexity.

Search for units whose activity correlates with the number of open nodes, our proxy for syntactic complexity

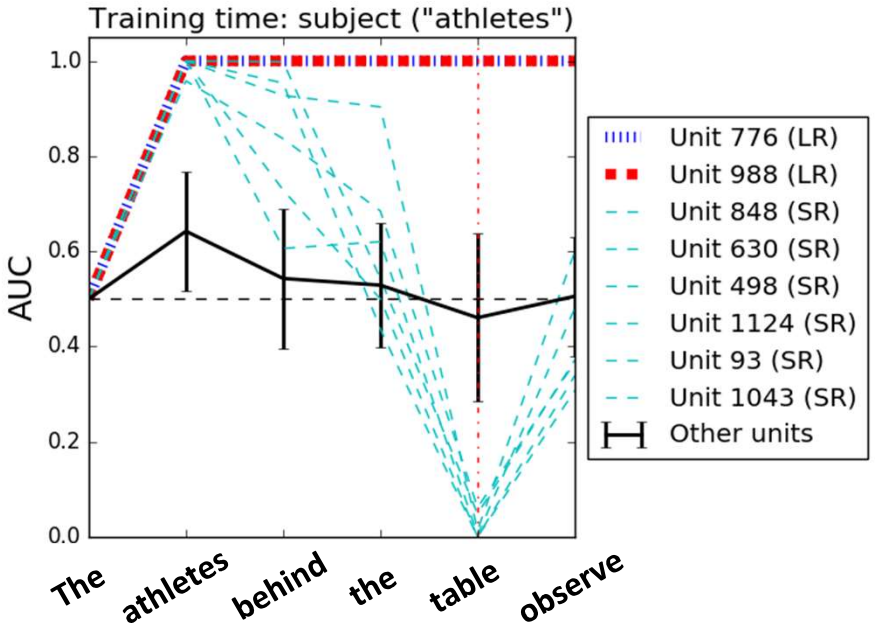




Identifying number units (singular vs plural)

Step 1. Find units whose activity decodes singular vs plural.

- Several units encode the number of the **current** noun.
- Most generalize over time, but only over a short period, and they **refresh** when a new noun is presented.
- Two units, however, show **sustained number coding**.



Step 2. Lesioning. Only the lesion of those two units brings long-distance agreement performance to chance level (one for singular nouns, one for plural nouns)

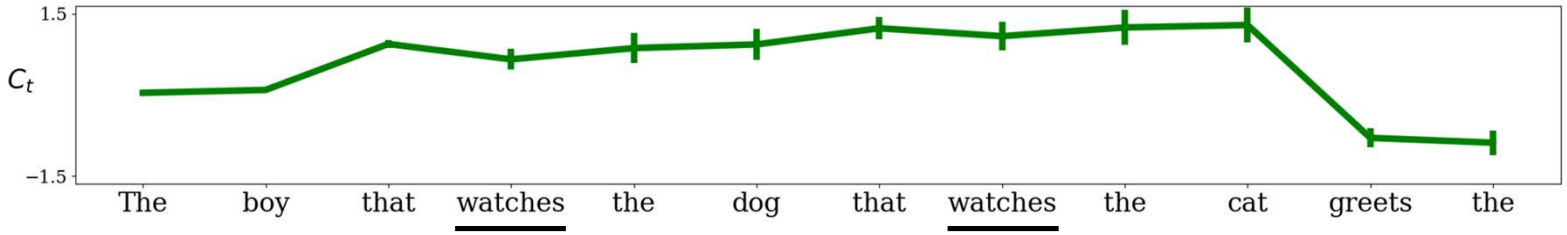
Step 3. Physiology. Those two units memorize the past number information, across intermediate words.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. NAACL-2019.

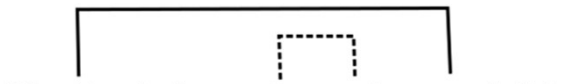
The LSTM language model is structure-sensitive, but not recursive

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2020). Exploring Processing of Nested Dependencies in Neural-Network Language Models and Humans. *arXiv*

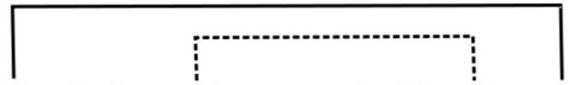
The structure unit correctly tracks nested relatives.
However, the singular and plural units can only store a single noun.



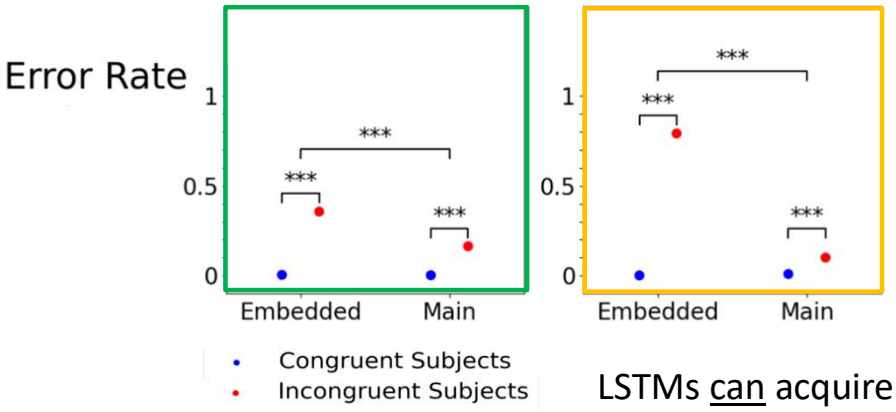
Long + short = ~ok



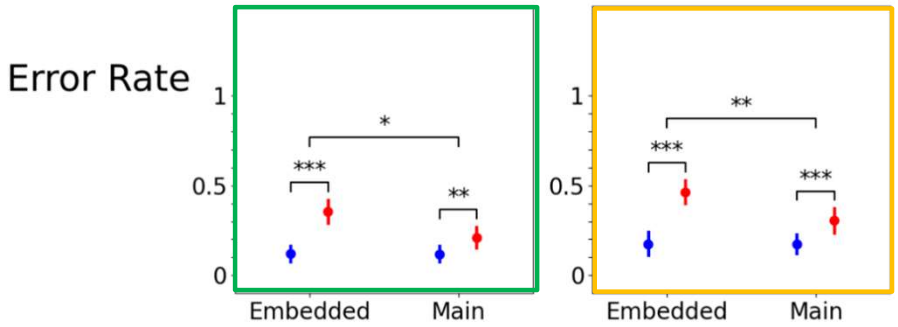
Two long-distance dependencies = worse than chance!



NLMs

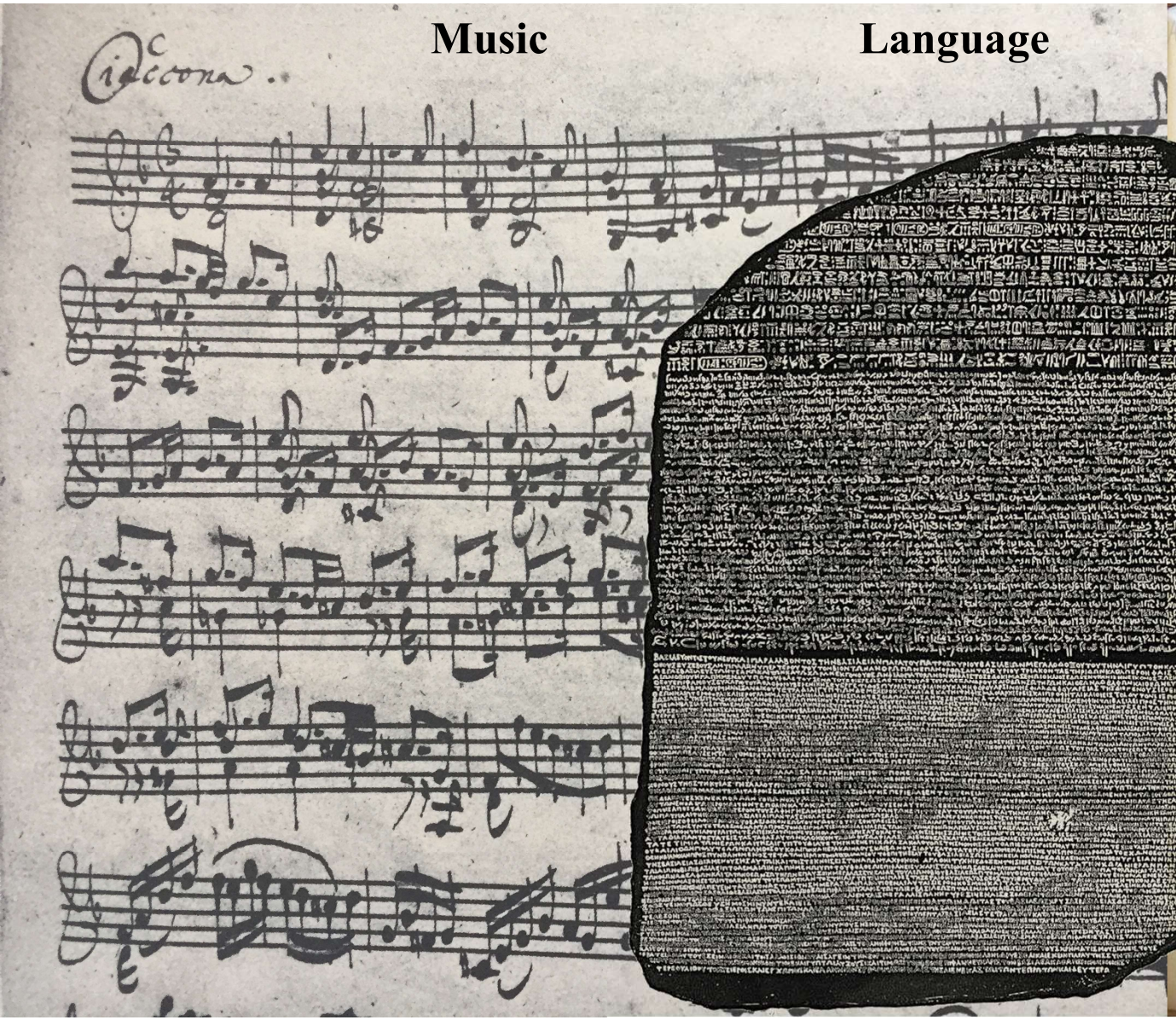


Humans



LSTMs <u>can</u> acquire nested dependencies in artificial grammars, but they only do so up to the trained depth: a stack is implemented by a rotation of the encoding vectors.

Music **Language** **Mathematics**



(7) [22] 18

VIII Theorems on approximate integrations and summation of series.

(1). $1^2\log 1 + 2^2\log 2 + 3^2\log 3 + \ldots + x^2\log x$

$= \frac{x(x+1)(2x+1)}{6}\log x - \frac{x^3}{9} + \frac{1}{4\pi^2}\left(\frac{1}{1^3}+\frac{1}{2^3}+\frac{1}{3^3}+\&c\right)$

$+ \frac{x}{12} - \frac{1}{360x} + \&c.$

) $1 + \frac{x}{11} + \frac{x^2}{12} + \frac{x^3}{13} + \ldots + \frac{x^x}{1^x}\theta = \frac{e^x}{2}.$

where $\theta = \frac{1}{3} + \frac{4}{135(x+k)}$ where k lies between

$\frac{8}{45}$ and $\frac{3}{21}$.

$+ \left(\frac{x}{11}\right)^5 + \left(\frac{x^2}{12}\right)^5 + \left(\frac{x^3}{12}\right)^5 + \&c$

$\frac{\sqrt{5}}{4\pi^2}\cdot\frac{e^{5x}}{5x^2-x+\theta}$ where θ vanishes when $x=\infty$.

$\frac{1^2}{x_1} + \frac{2^2}{e^{2x}-1} + \frac{3^2}{e^{3x}-1} + \frac{4^2}{e^{4x}-1} + \&c$

$\frac{2}{x^3}\left(\frac{1}{1^3}+\frac{1}{2^3}+\frac{1}{3^3}+\&c\right) - \frac{1}{12x} + \frac{x}{1440} + \frac{x^3}{181440}$

$\frac{x^5}{7257600} + \frac{x^7}{159667200} + \&c.$ when x is small

x may be given values from 0 to 2).

$+ \frac{1}{1002^2} + \frac{3}{1003^3} + \frac{4^2}{1004^4} + \frac{5^3}{1005^5} + \&c$

$00 - 10^{-440} \times 1.0135$ nearly.

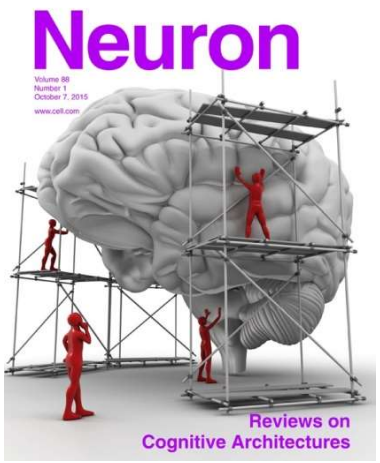
$x^n dx = \frac{\sqrt{\pi}}{2} - \frac{e^{-a^2}}{2a} + \frac{1}{a} + \frac{2}{2a} + \frac{3}{a} + \frac{4}{2a} + \&c.$

The Coeffic. of x^n in $\frac{1}{1-2x+2x^4-2x^9+2x^{16}-\&c}$

= The nearest integer to $\frac{1}{4n}\left\{\cosh(\pi\sqrt{n}) - \frac{\sinh(\pi\sqrt{n})}{\pi\sqrt{n}}\right\}$

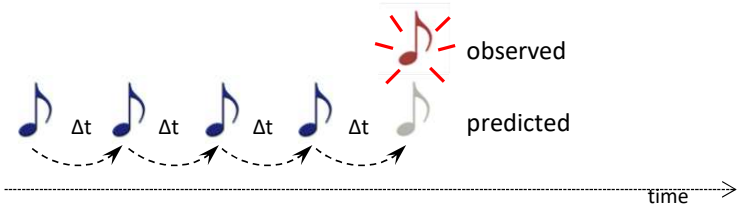
A hypothesis: The singularity of the human brain may lie in the ability to construct nested symbolic tree-like representations

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. Neuron, 88(1), 2–19.



Shared with other primates

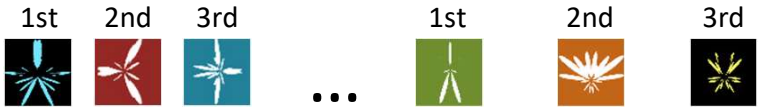
Transitions and timing



Chunking

tokibugikobagopilagikobatokibugopila ...

Ordinal knowledge



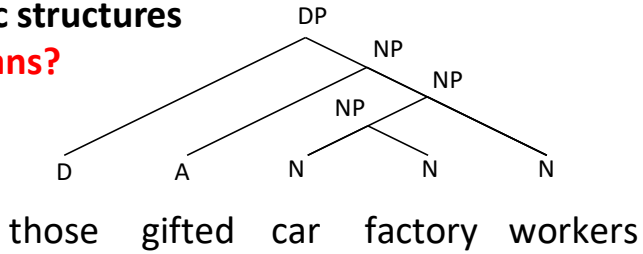
Algebraic patterns

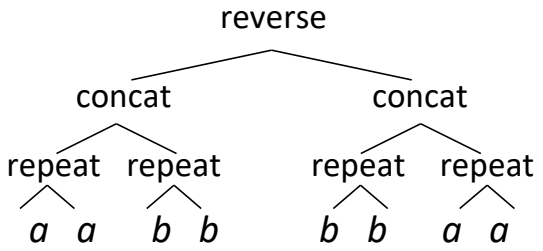
A A B *A A B* *A A B* *A B A (violation)*

totobu ... mimitu ... gagari ... pesipe ...

Nested symbolic structures

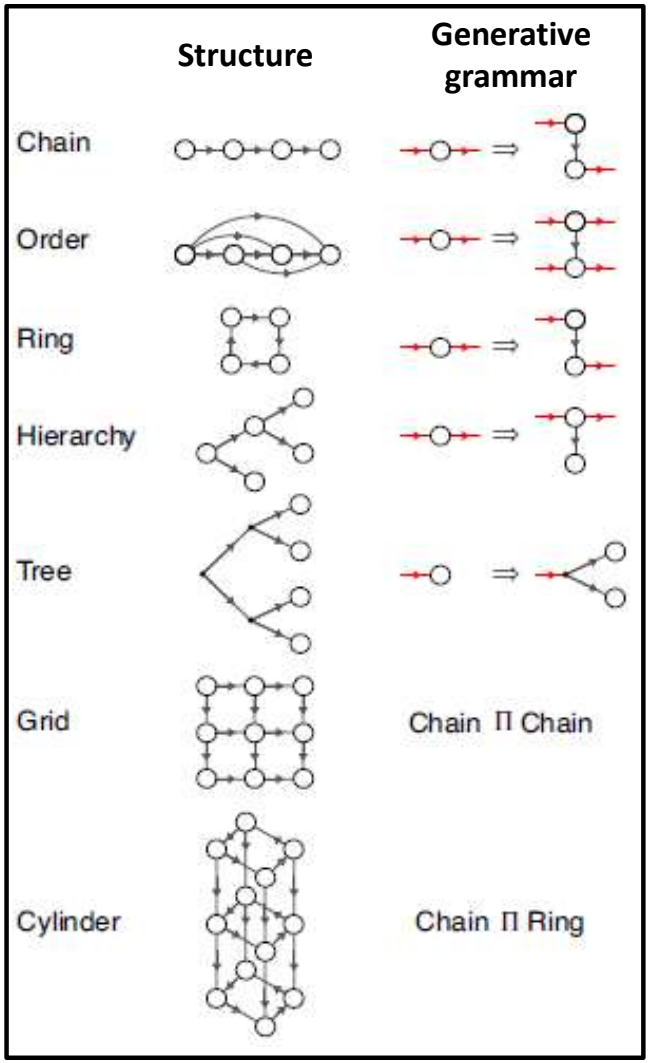
Unique to humans?





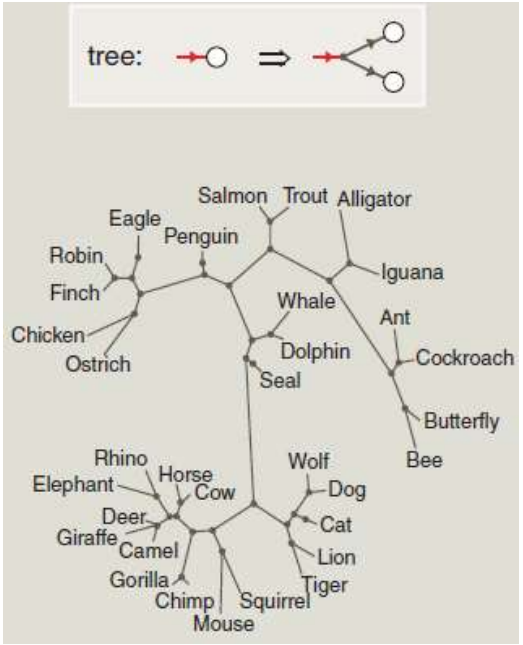
Key hypothesis: the human <u>compresses</u> information using nested structures

The role of a « language of thought » in human conceptual growth

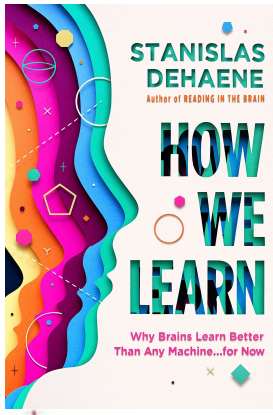


Human learning is based on efficient **Bayesian algorithms** operating on expressions in a mental **language of thought**

Automatic discovery of the tree of species



Kemp & Tenenbaum PNAS, 2008



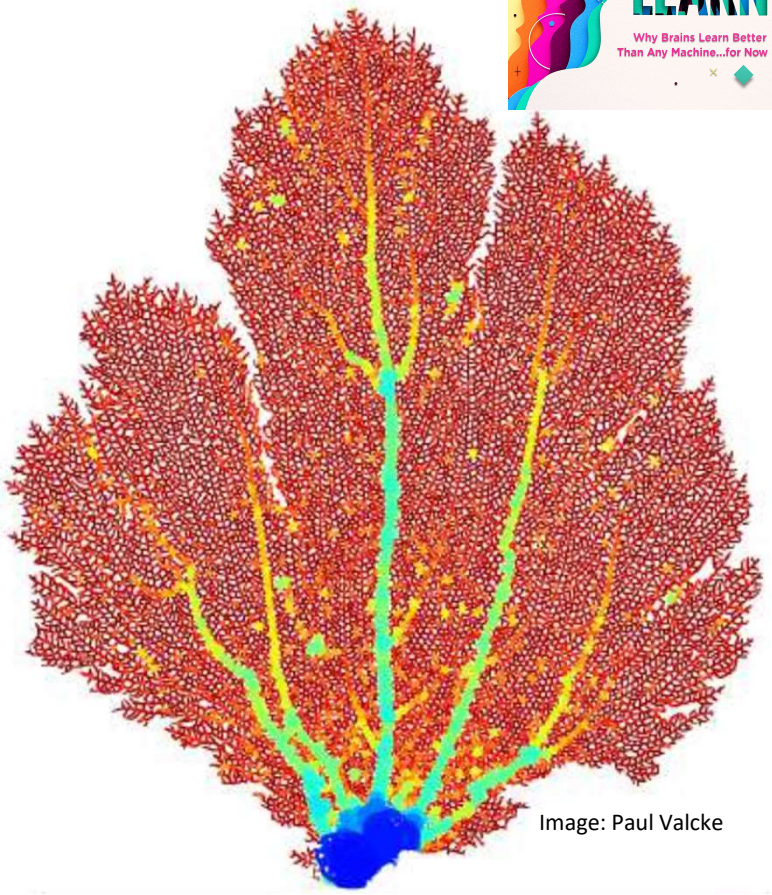
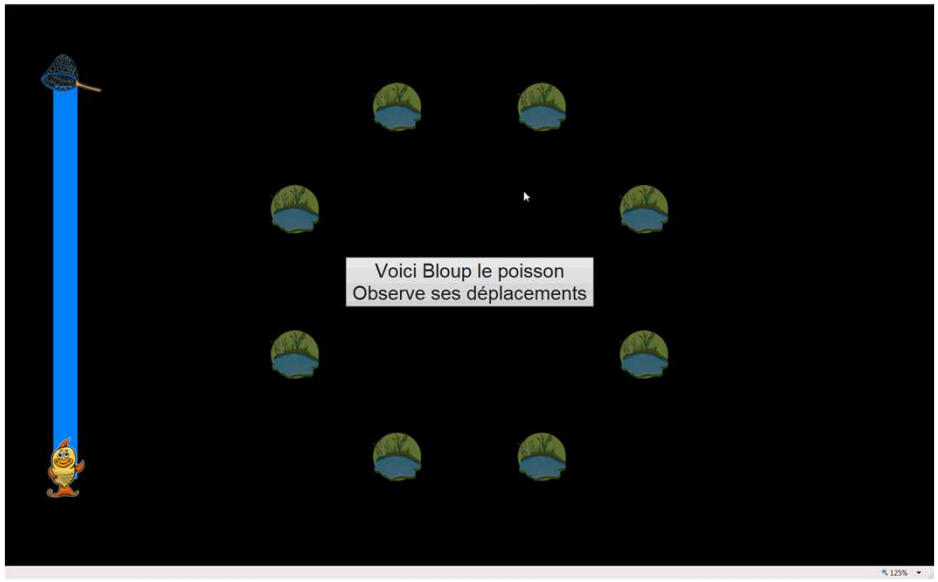


Image: Paul Valcke

A simplified “language of geometry”

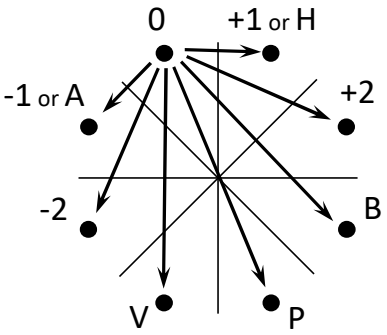
Amalric, M., Wang, L., Pica, P., Figueira, S., Sigman, M., & Dehaene, S. (2017). The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. PLoS Computational Biology, 13(1)



Subjects see a sequence and are asked to anticipate the next location.

A mini « language of geometry » captures the observed regularities.

Primitives

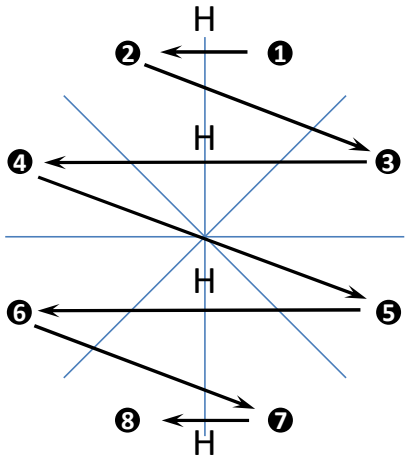


and rules :

e.g. repeat n times []^n

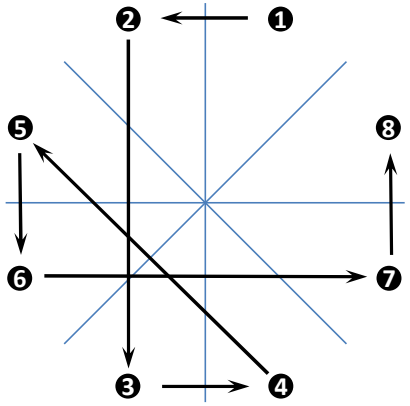
Example 1: “four segments”

Formula = [H^2]^4{+1}



Example 2: “two rectangles”

Formula = [[-1,-3]^2]^2<+2>



Minimal description length in our « language of geometry » predicts error rates

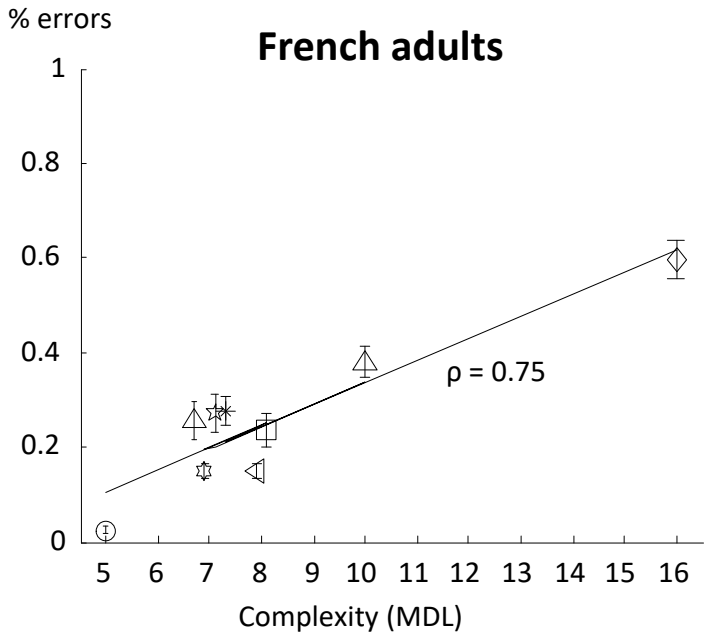
Amalric et al., PLoS Computational Biology 2017; Wang et al., Neuroimage 2019; Al Roumi et al. submitted

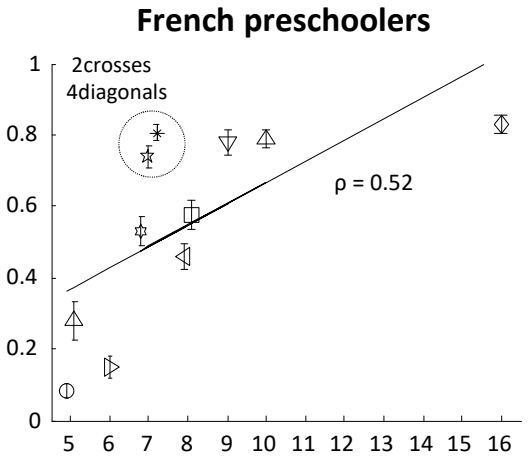
All sequences are of the same length (8 items).

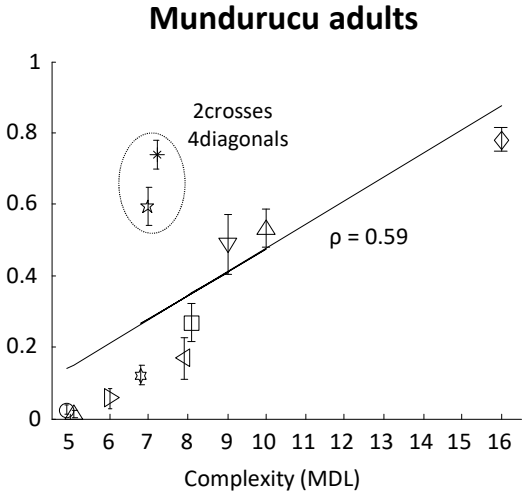
What predicts memory is not actual length, but **minimal description length** (a.k.a Kolmogorov complexity), the length of the shortest expression that can **compress** the sequence.

Ongoing work by Liping Wang:

Monkeys do not seem to care about temporal or geometrical regularities. They simply store each **location** in working memory, without seeming caring for the **structure of their transitions.**





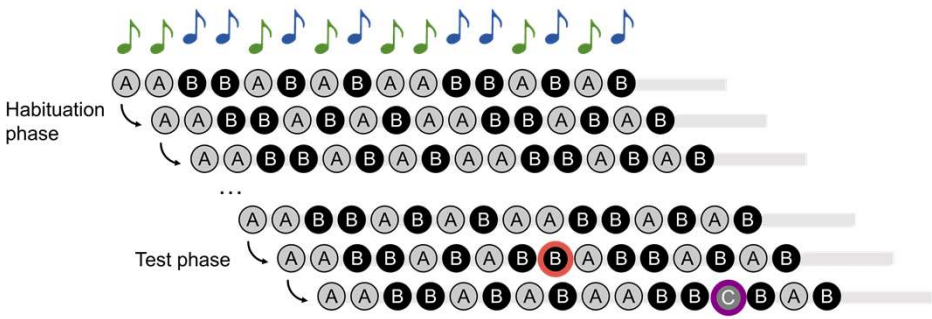


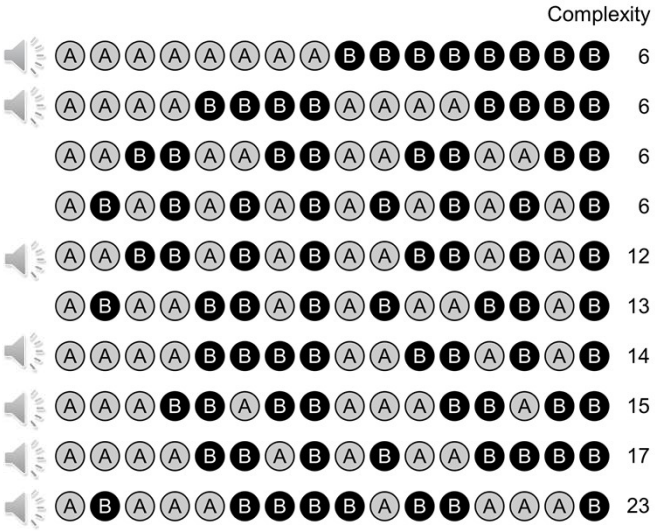
The same language is needed to account for *auditory* sequence complexity

Planton et al., PLoS Computational Biology, in press

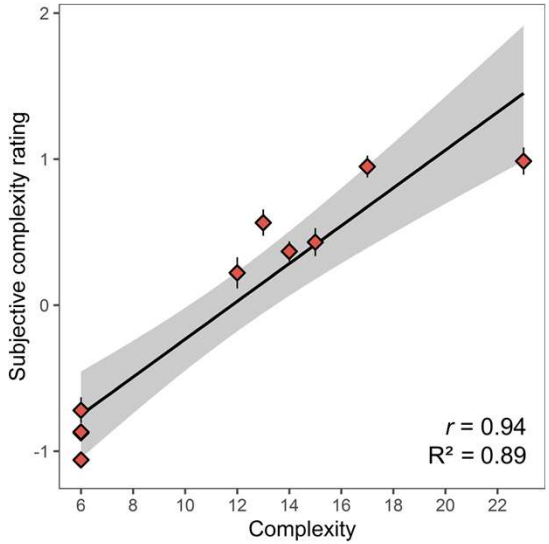
Our language of geometry, **unchanged**, predicts the subjective and objective complexity of a binary **auditory** sequence by its « minimal description length ».

- Subjective complexity ratings of heard auditory sequences of tones are highly correlated with minimal description length.
- Performance (response time and accuracy) in the detection of a deviant sound is also well predicted.

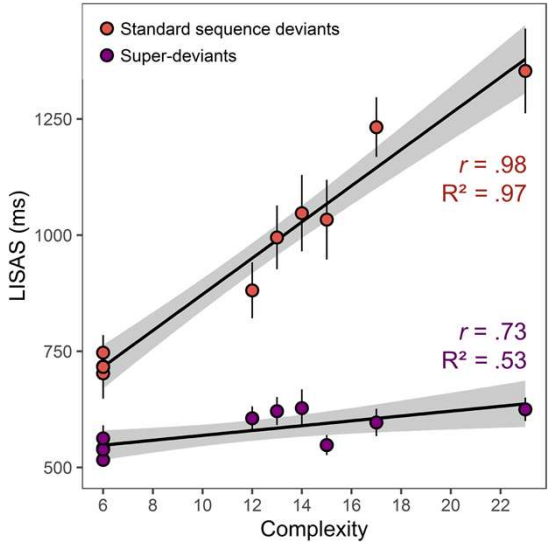




Subjective complexity (Ratings)

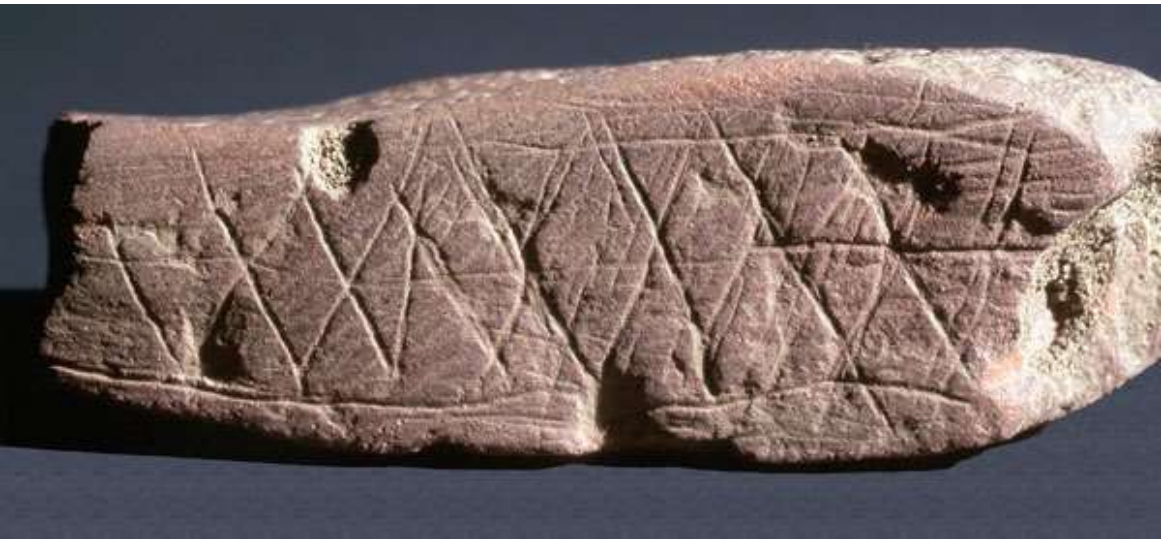


Objective complexity (Violation detection)

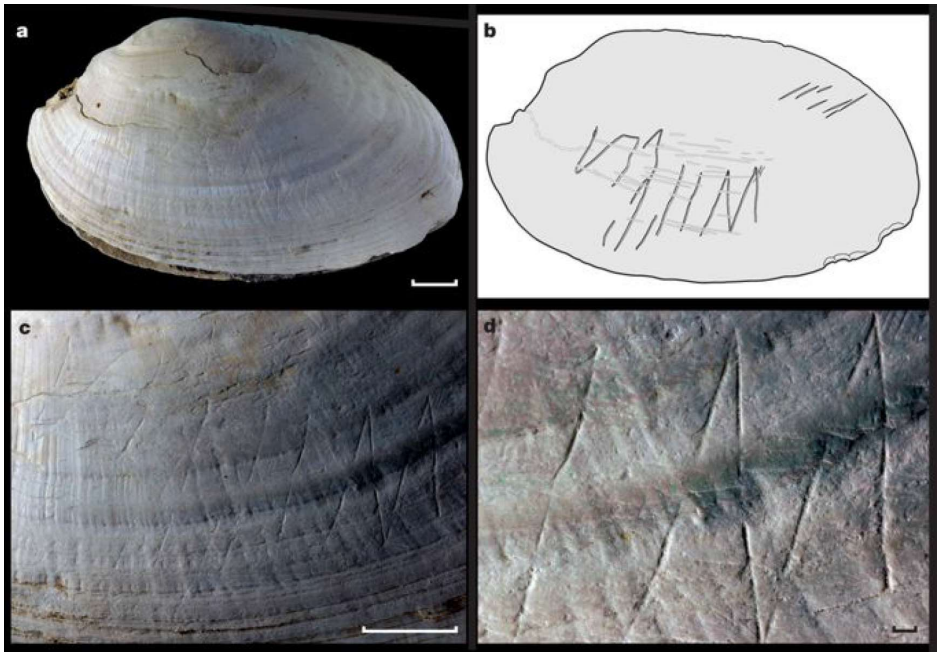


Understanding the human sense of geometrical patterns

How old is the sense of geometry?



Parallels and equilateral triangles, ~70,000 years old (Sapiens)

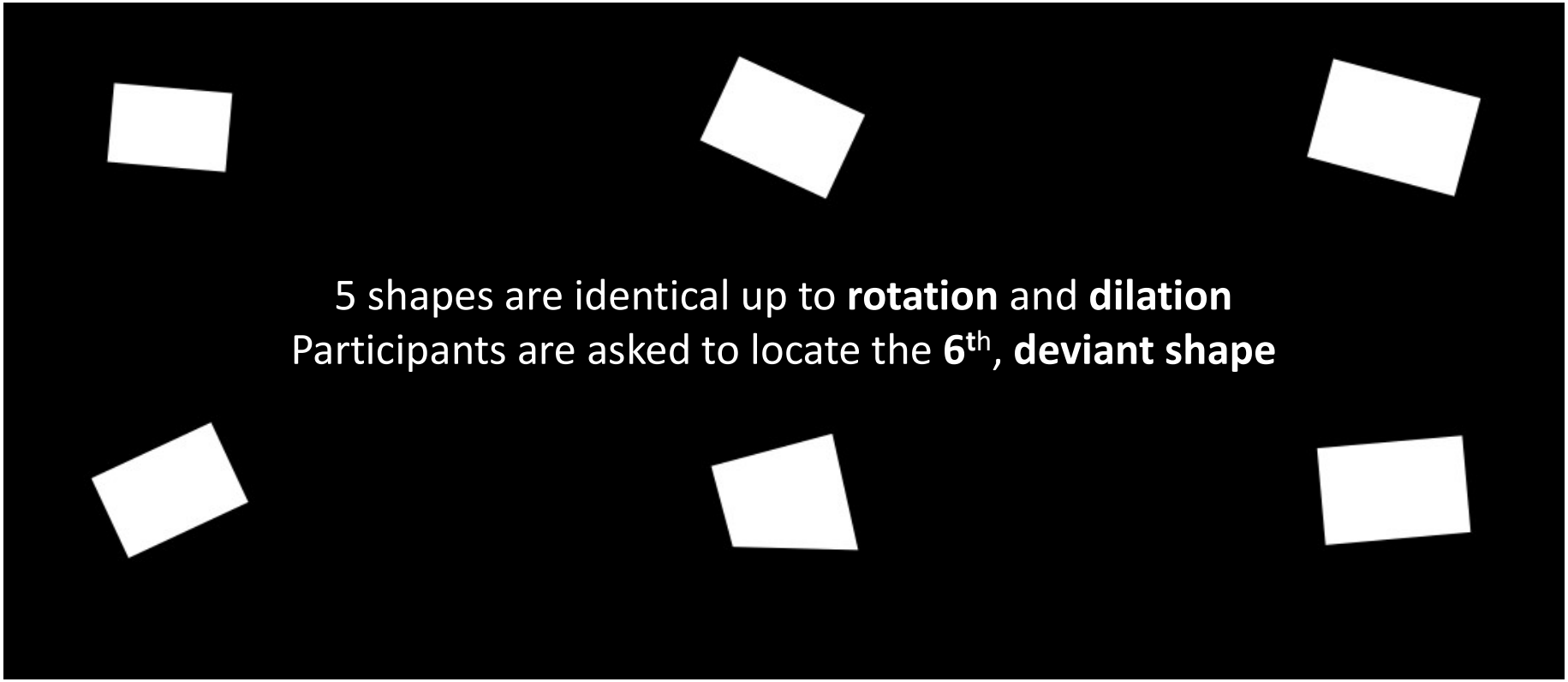


Zig-zag, ~540,00 years old (Erectus)

Bifaces and spheres may be
1,8-2 million years old
(Homo ergaster or archaic erectus)

How do human and non-human primates perceive quadrilaterals?

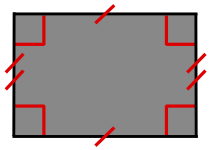
Studies of quadrilaterals (Mathias Sablé-Meyer, ongoing PhD)



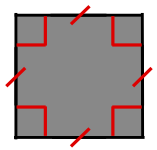
Does shape regularity predict perceptual complexity?

We used 11 quadrilaterals ranging from highly regular (square) to fully irregular

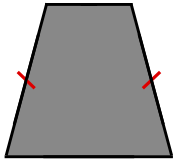
Rectangle



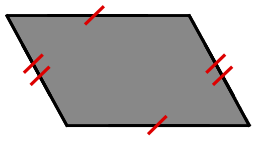
Square



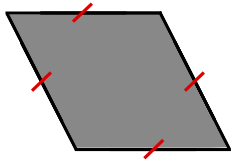
Iso-Trapezoid



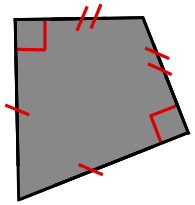
Parallelogram



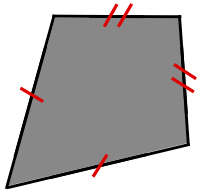
Rhombus



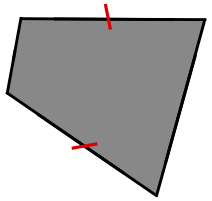
Right Kite



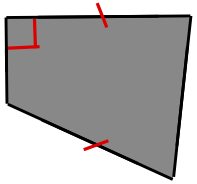
Kite



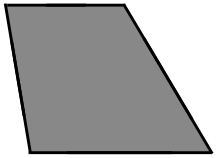
Hinge



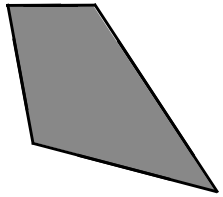
Right Hinge



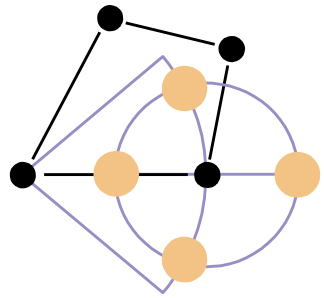
Trapezoid



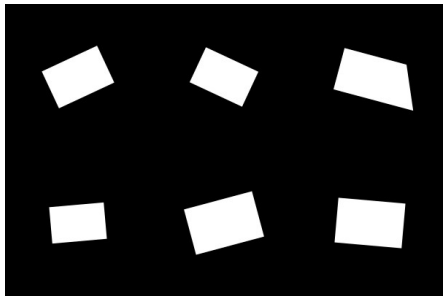
Irregular



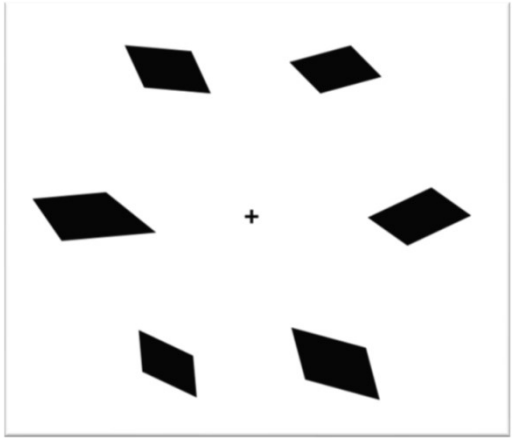
Deviants involve a displacement of the bottom right vertex.

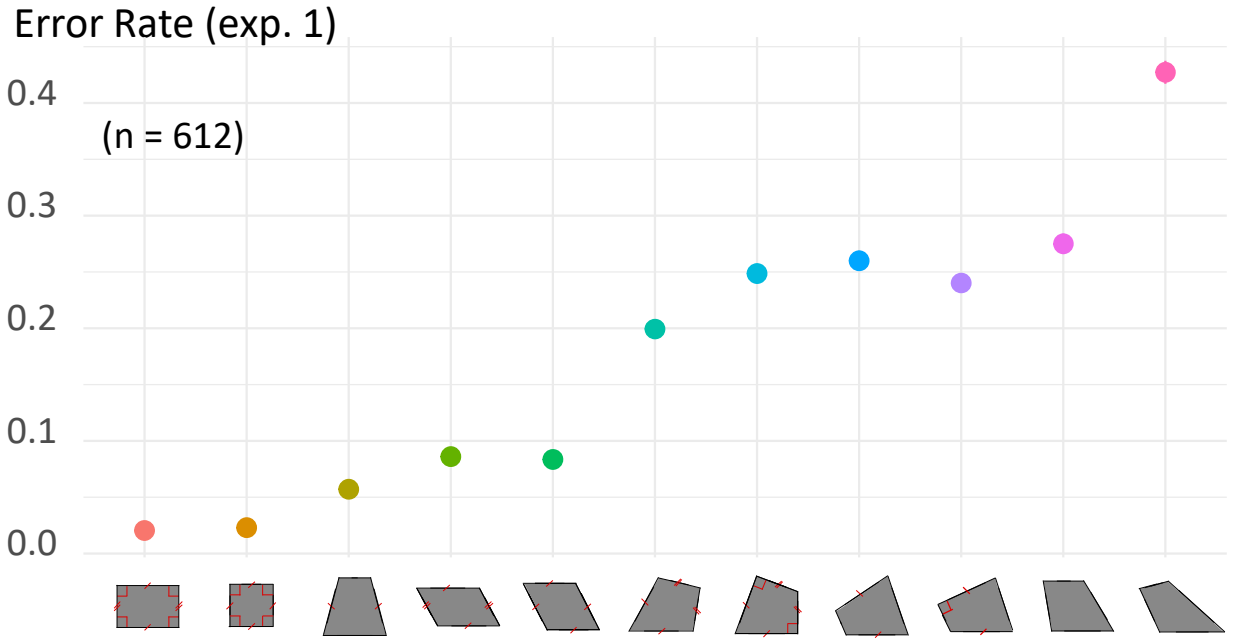


Exemple display :



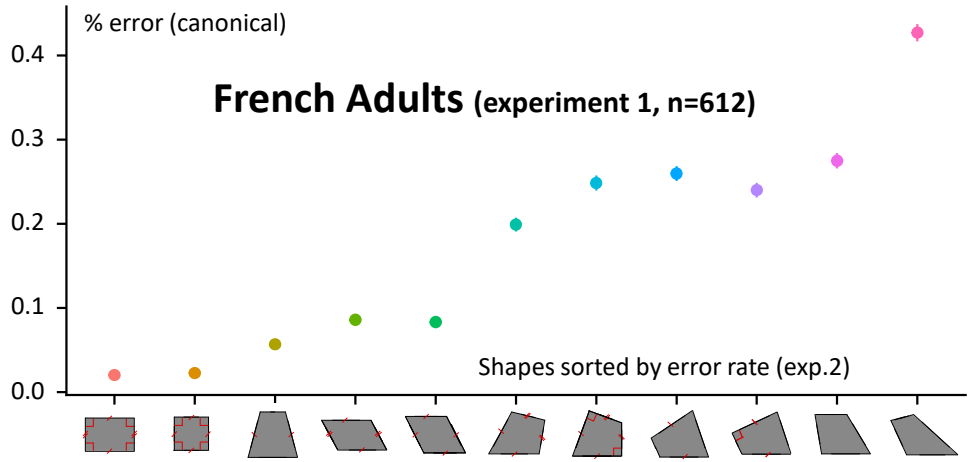
Human adults: a major effect of shape regularity

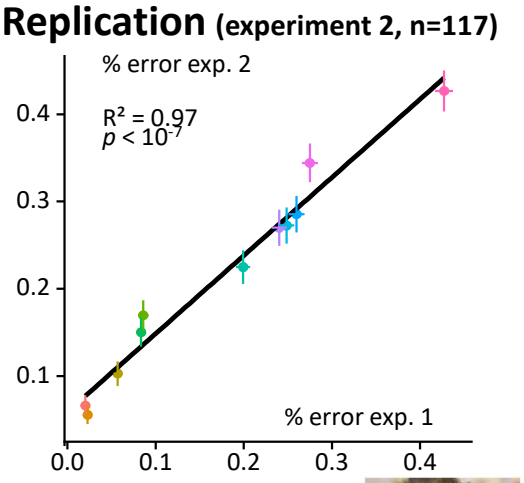


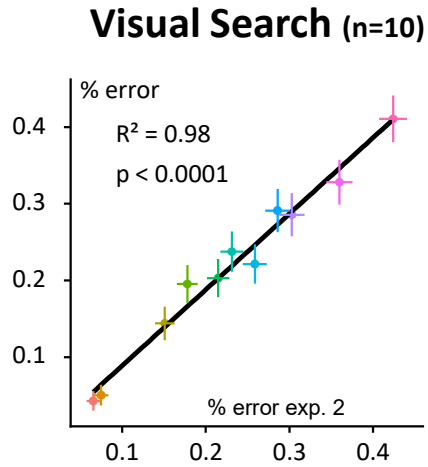


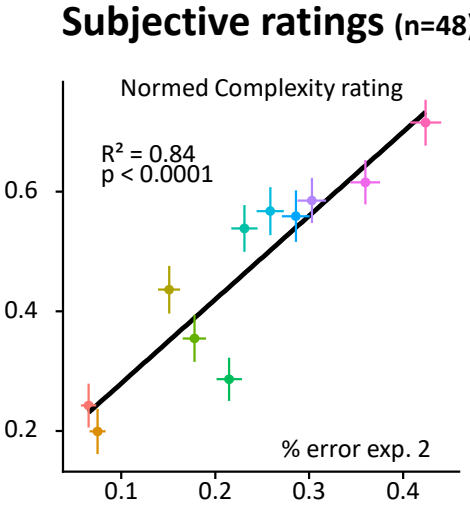
- Performance is above chance for all shapes, but varies from 7% to 42% errors.
- Response time follows the same pattern.
- The position, rotation and size of the outlier have either no significant effect or significant effects with almost no explained variance

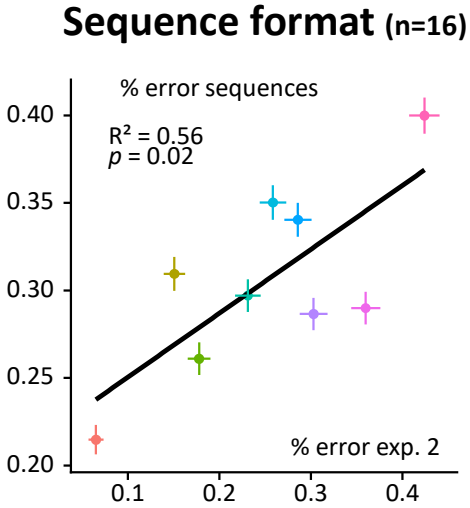
The geometrical regularity effect: a human universal ?

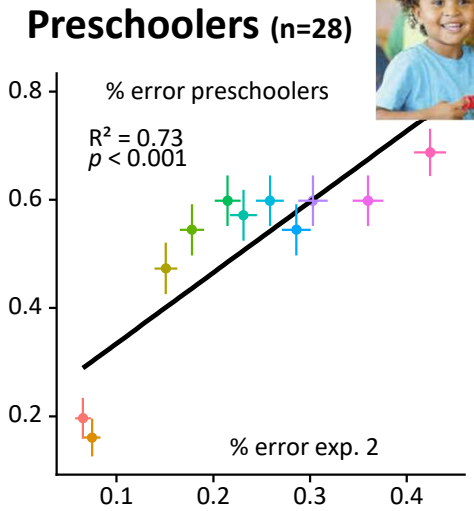


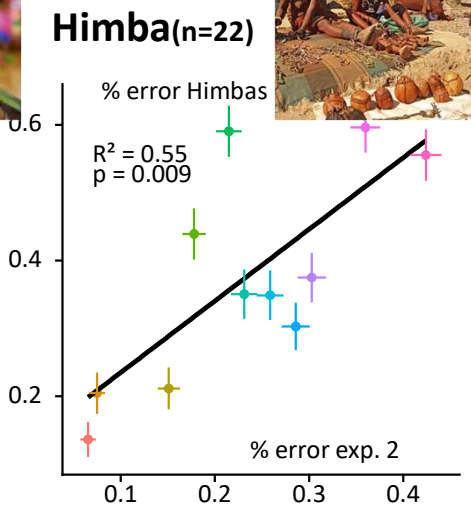








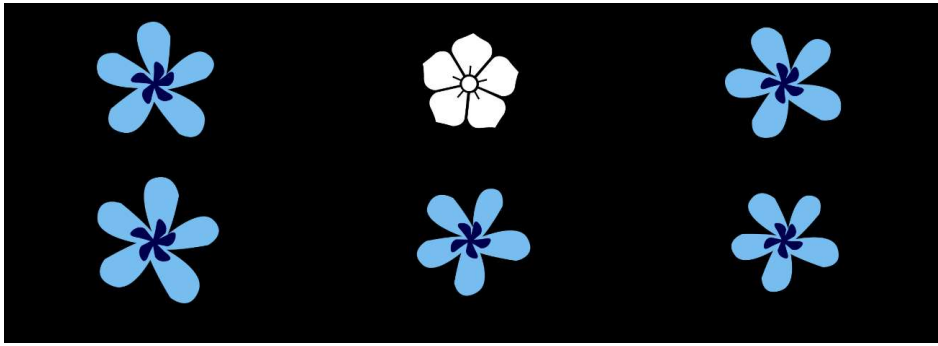




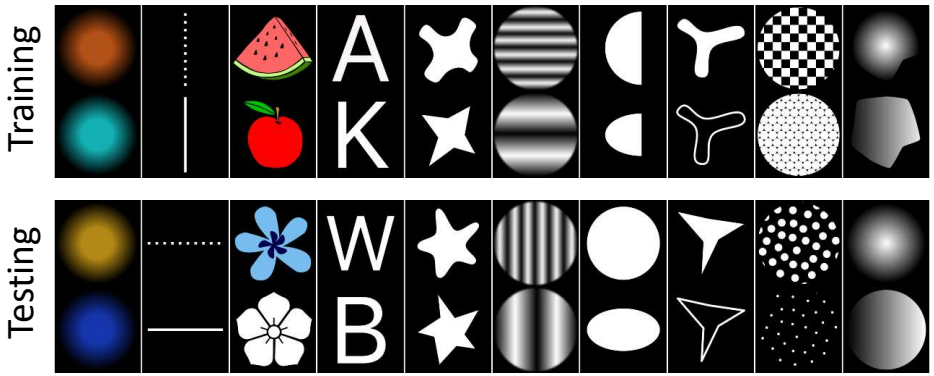
Is the shape regularity effect present in non-human primates? A study in baboons (with Joël Fagot)

Prediction: baboons should fail to show the shape regularity effect

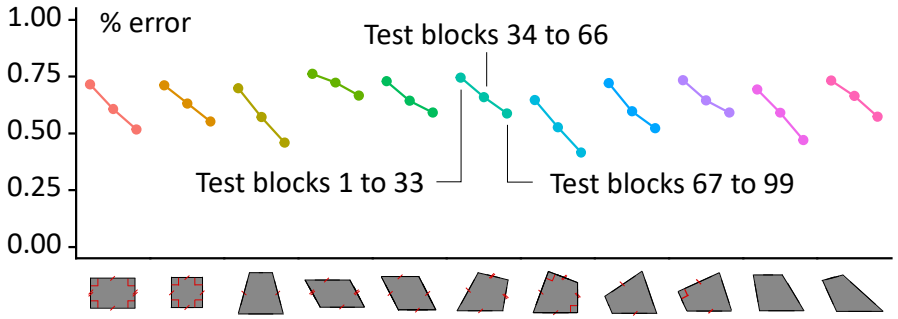
Baboons were first trained to perform the outlier task with simple pictures:



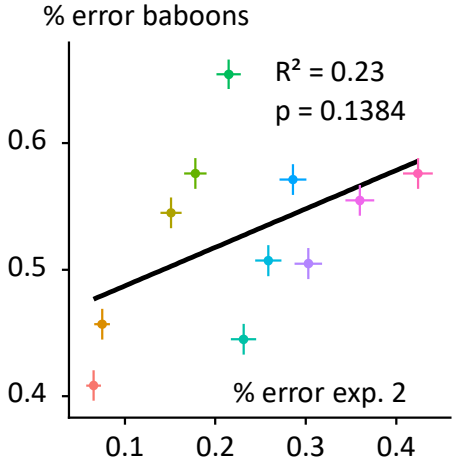
Following training, we tested generalization to novel pictures, and only then to geometrical shapes.

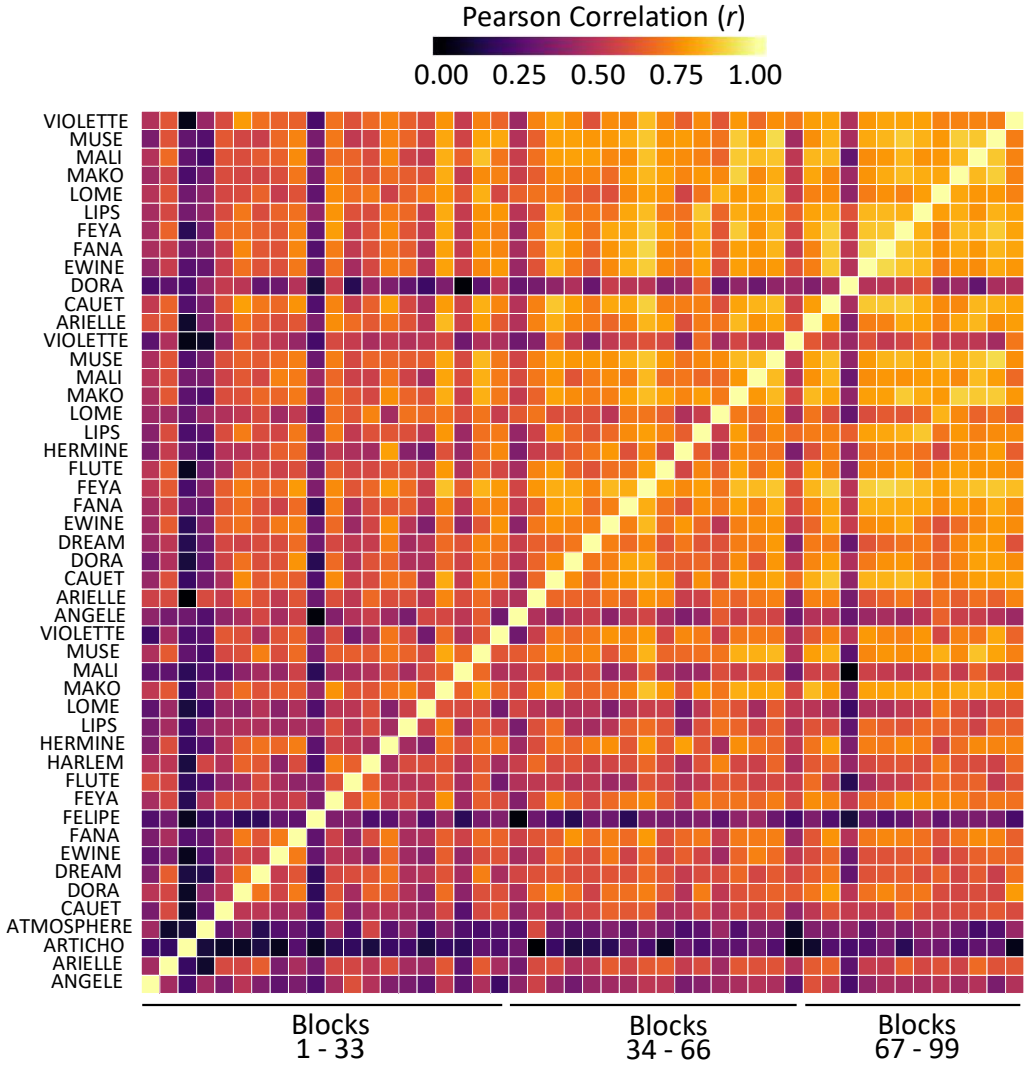


The shape regularity effect is absent in baboons

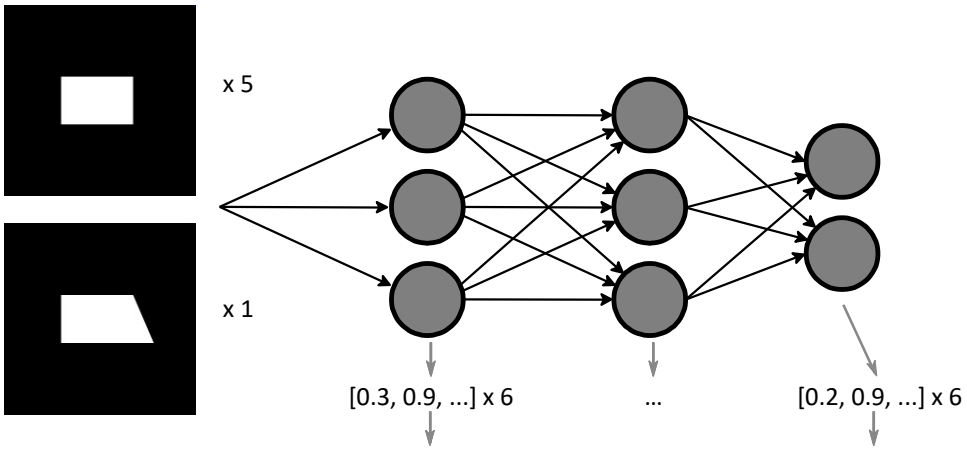


- Baboons are essentially at chance with all shapes on the first block with quadrilaterals
- After 8800 trials, performance improves, but remains poor and uncorrelated with humans.
- Nevertheless, there is a striking consistency of the baboon pattern across time and individuals



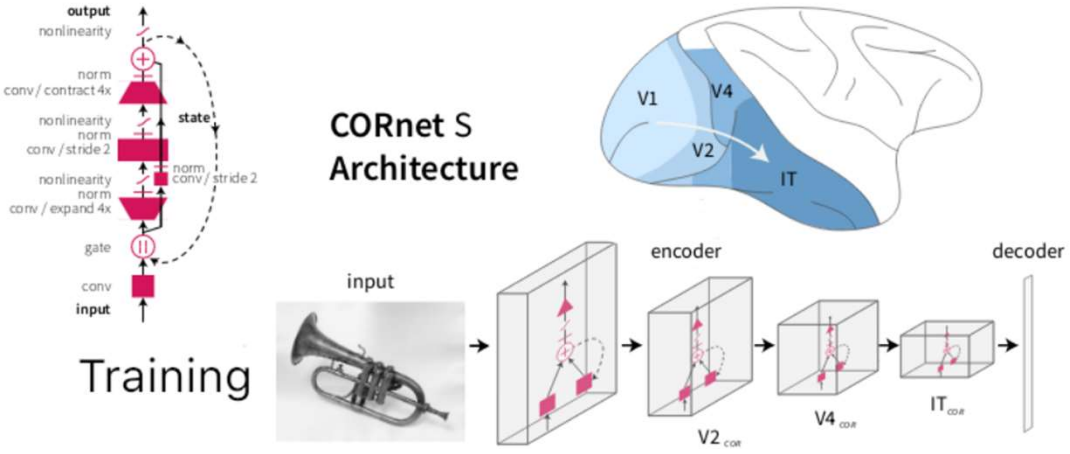


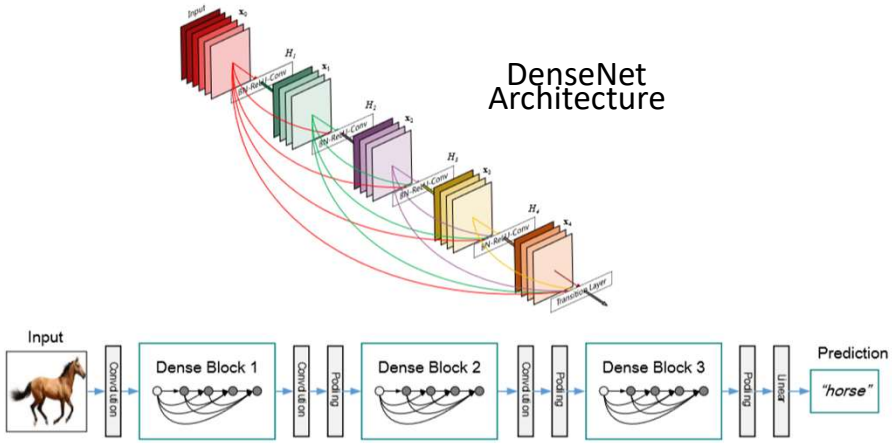
Model 1: shape perception by a convolutional neural network (CNN)



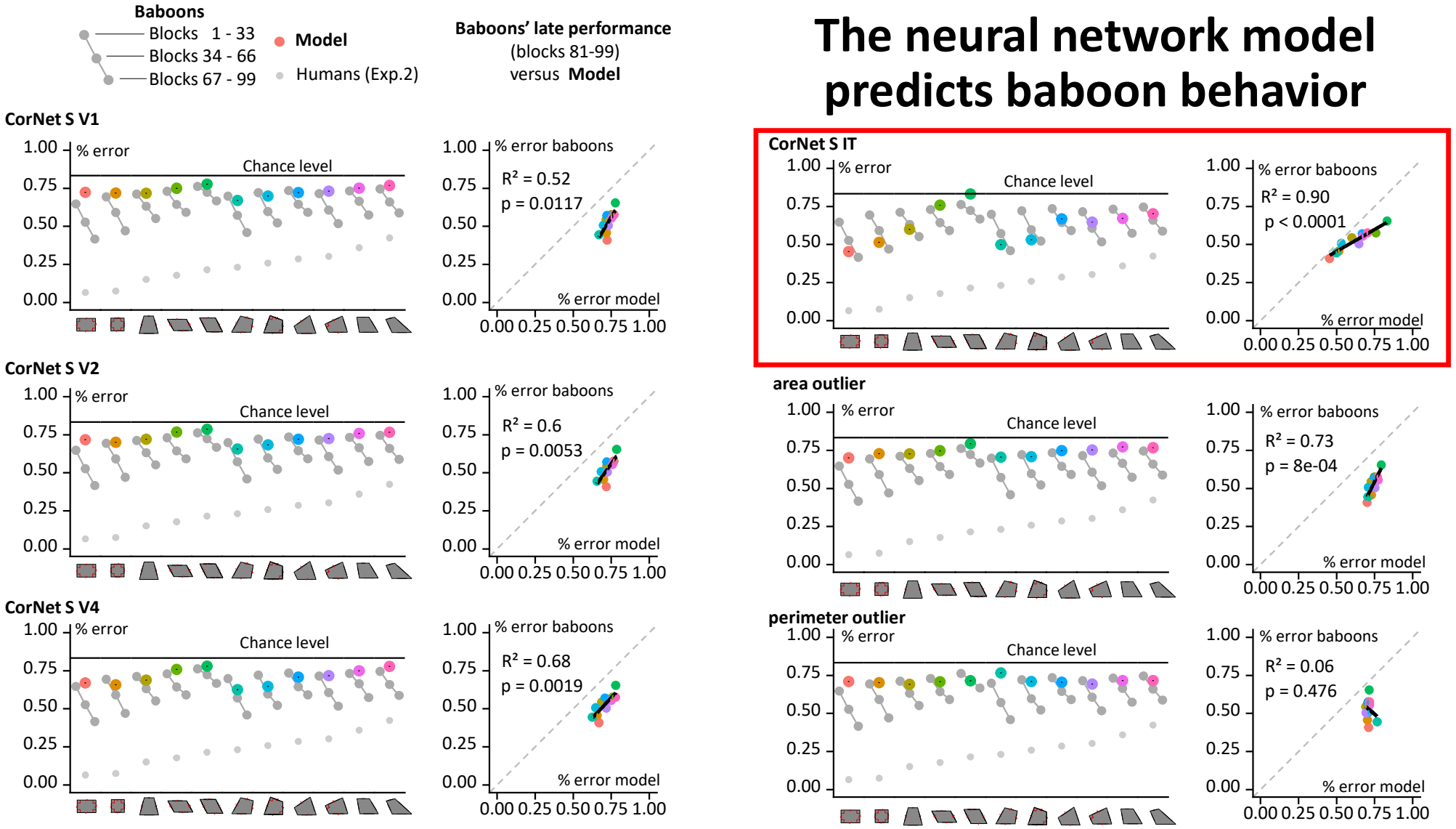
- We presented our stimuli to CoreNet-S, a model trained to categorize natural images and which provides a good match to human performance and inferotemporal neuronal recordings.
- A similar experiment was done with two other CNNs, DenseNet and ResNet, with similar results

Outlier =
Vector most different from the mean of the others (in a given layer)

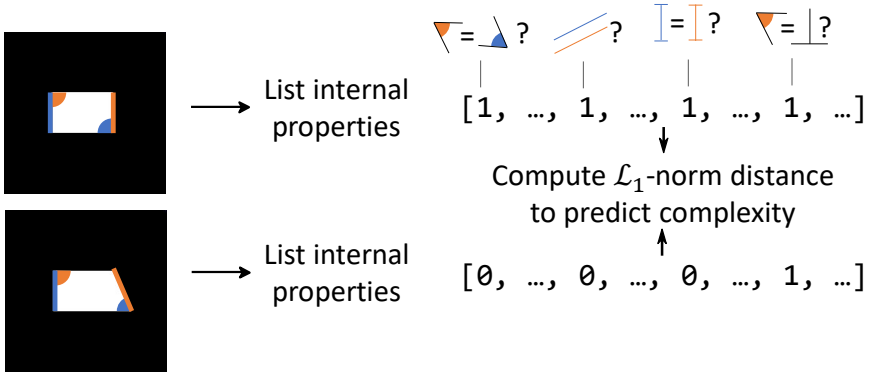




The neural network model predicts baboon behavior



Model 2: A symbolic model with discrete geometrical properties

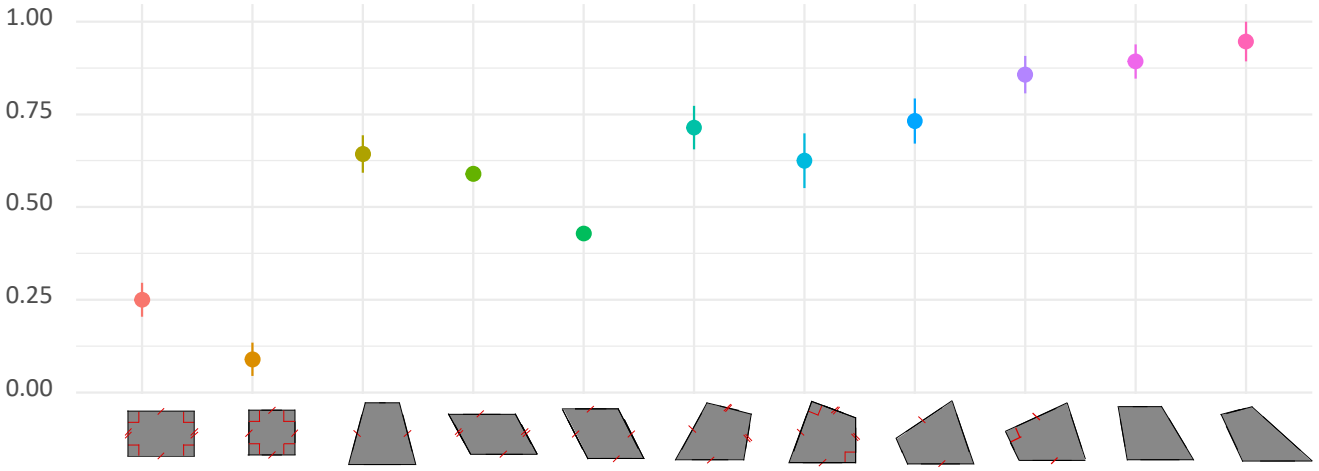


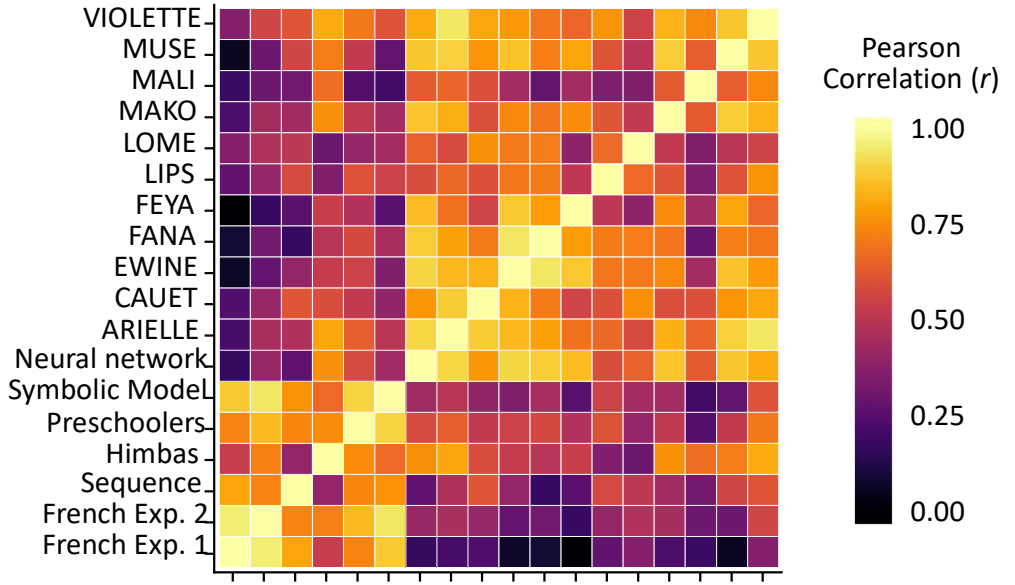
The symbolic model lists the discrete properties of the shapes (within a certain tolerance level)

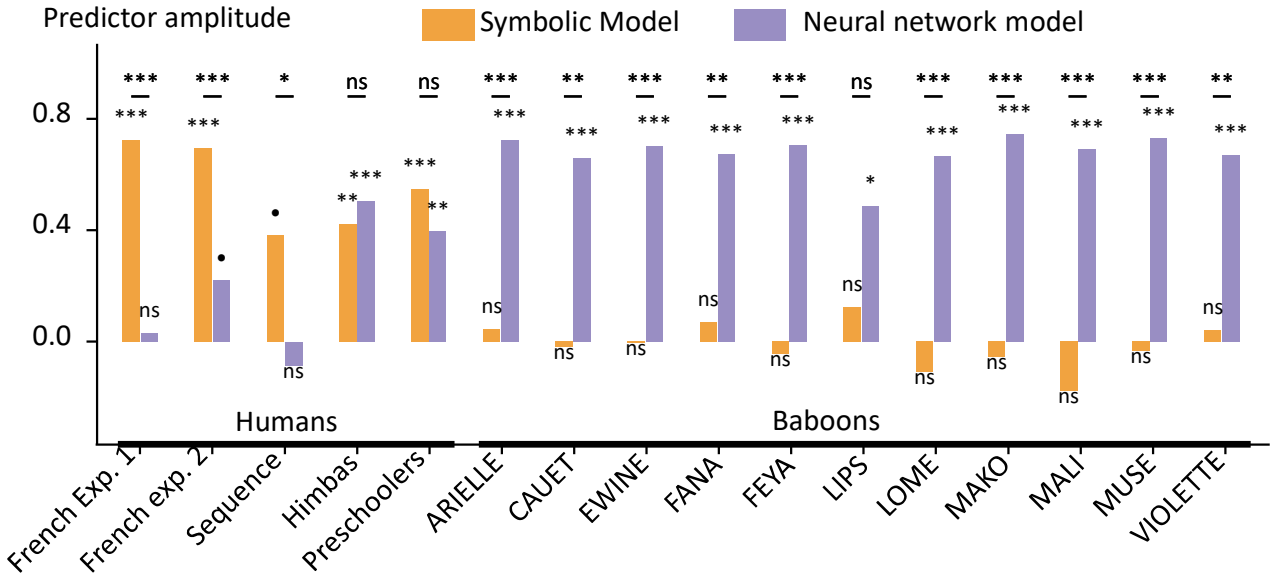
- Equal angles
- Parallelisms
- Equal lengths
- Right angles

The ease of outlier detection is predicted by the number of properties that differ.

This model nicely predicts the shape regularity effect:







A double dissociation between humans and baboons

- In a multiple regression, the neural-network and symbolic models capture respectively the baboon and human data.
- The symbolic model fails to predict of the baboon data even at the individual level
- Himba and preschoolers rely on a mixture of the two strategies

Could experience explain the human pattern?

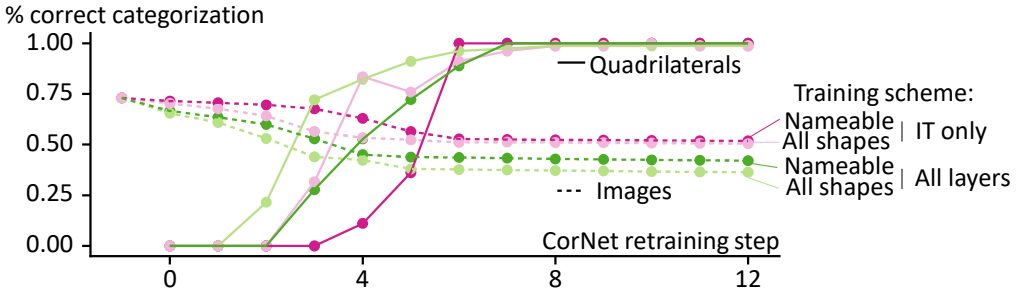
1. Training in a "carpentered world"? No

The baboons live a world which is arguably more "carpentered" than the Himba, yet they have opposite results.

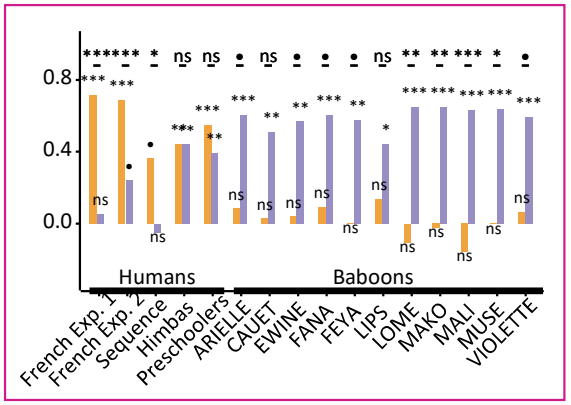
2. **Training with geometrical shapes ? No**

- The baboons received thousands of training trials.
- We trained the network to label our geometrical shapes with additional output units:
  - Either all 11 shapes, or just the shapes with names
  - Either by updating the entire network, or by changing the last layer (IT only)

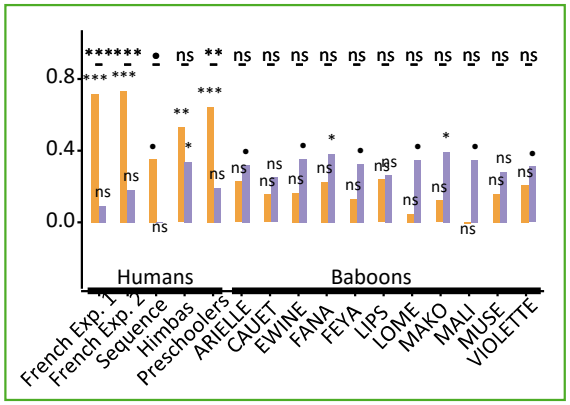
The network reaches perfect scores on novel displays of those shapes, but predictions are unchanged.



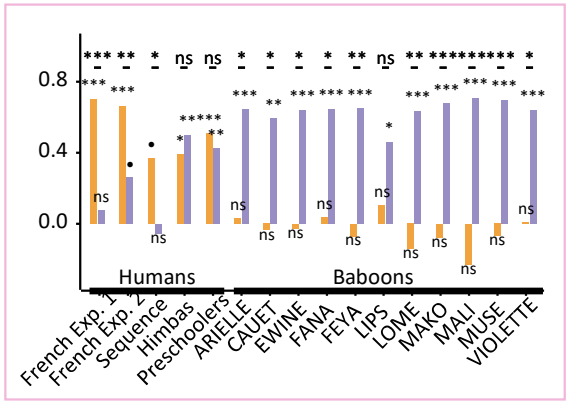
IT only ; nameable



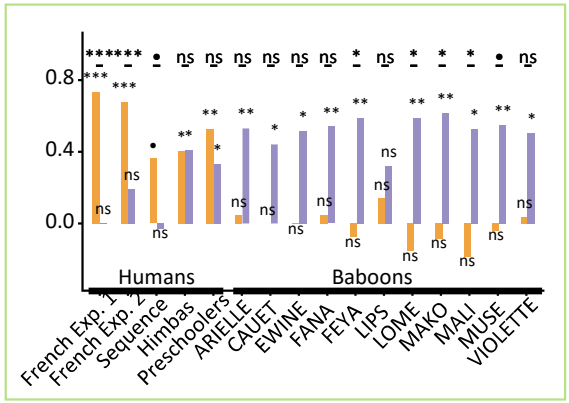
All layers ; nameable

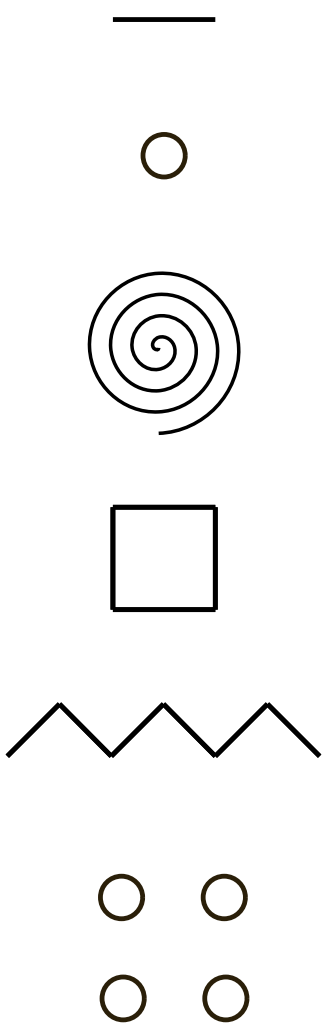


IT only ; all shapes



All layers ; all shapes





Generalization beyond the quadrilaterals:

A generative language for geometrical shapes

Goal: propose an **actual programming language** that can account for the basic geometrical shapes used in human cultures throughout the world.

The language contains a few key primitives:

- ***Number:*** *1,2,3, successor, half, double*
- ***Geometry:*** *Move, Turn Trace*
- ***Control:*** *Repeat, Concatenate, Embed*

For instance a simple square is:

```
Repeat (4)
   { Concatenate ( Trace(1) , Turn(90°) }
```

Shape perception is program induction!

A prediction about shape complexity

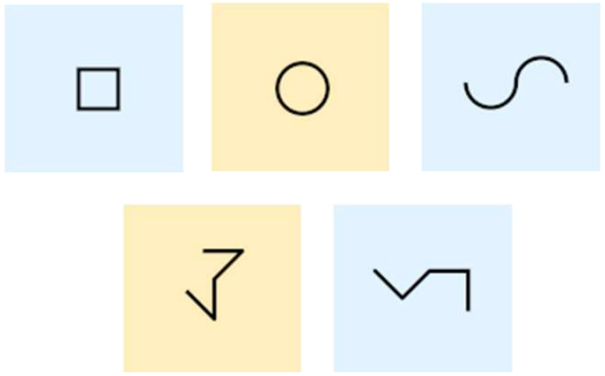
Prediction: Shape complexity should be determined by the **length of the shortest program capable of reproducing it.**

Perceptually rich drawings can be generated by **a single instruction**: *repeat*, *concat*, or *embed.*

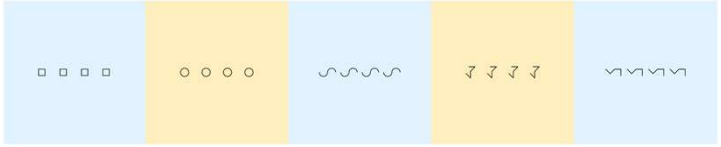
Complexity should follow **additive** rules:

Complexity (***Repeat*****(x)**) = Complexity (x) + constant
Complexity (***Concat*****(x,y)**) = Complexity (x) + Complexity (y) + constant'
Complexity (***Embed*****(x,y)**) = Complexity (x) + Complexity (y) + constant''

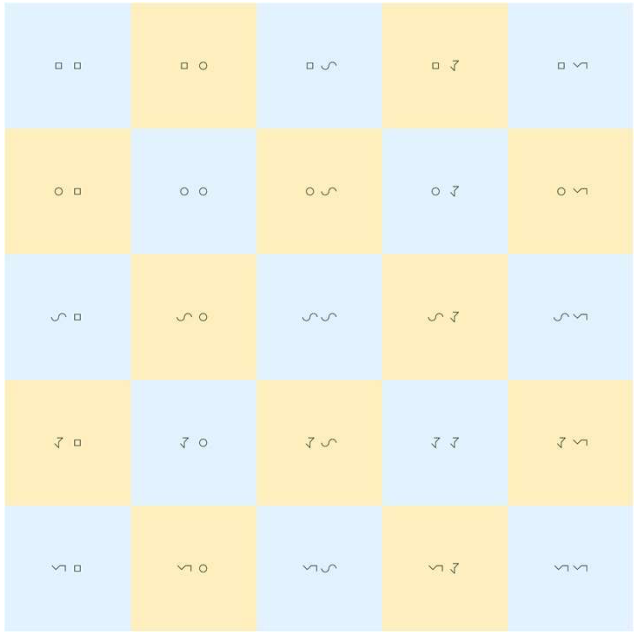
We selected 5 base shapes with increasing complexities



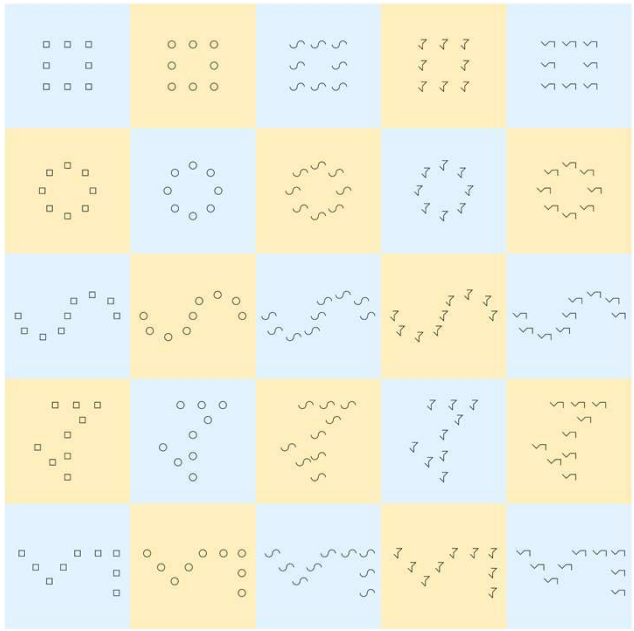
… and used them into programs:



Repeat

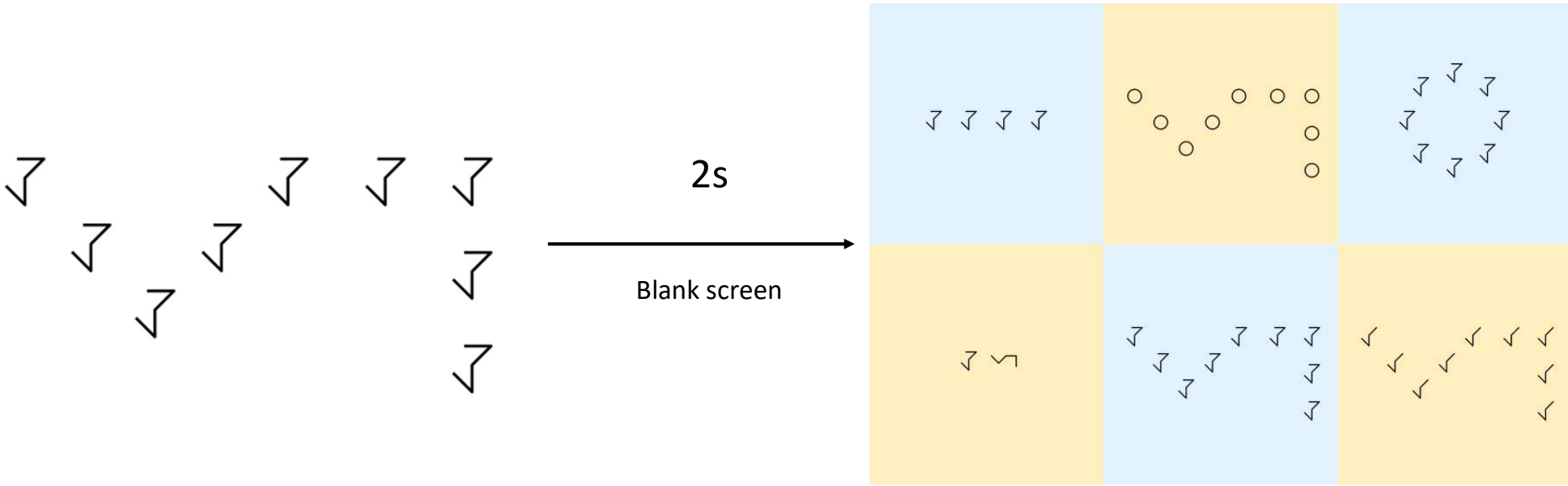


Concatenate



Embed

Two behavioral measures of shape complexity in humans



Phase 1 : Encoding

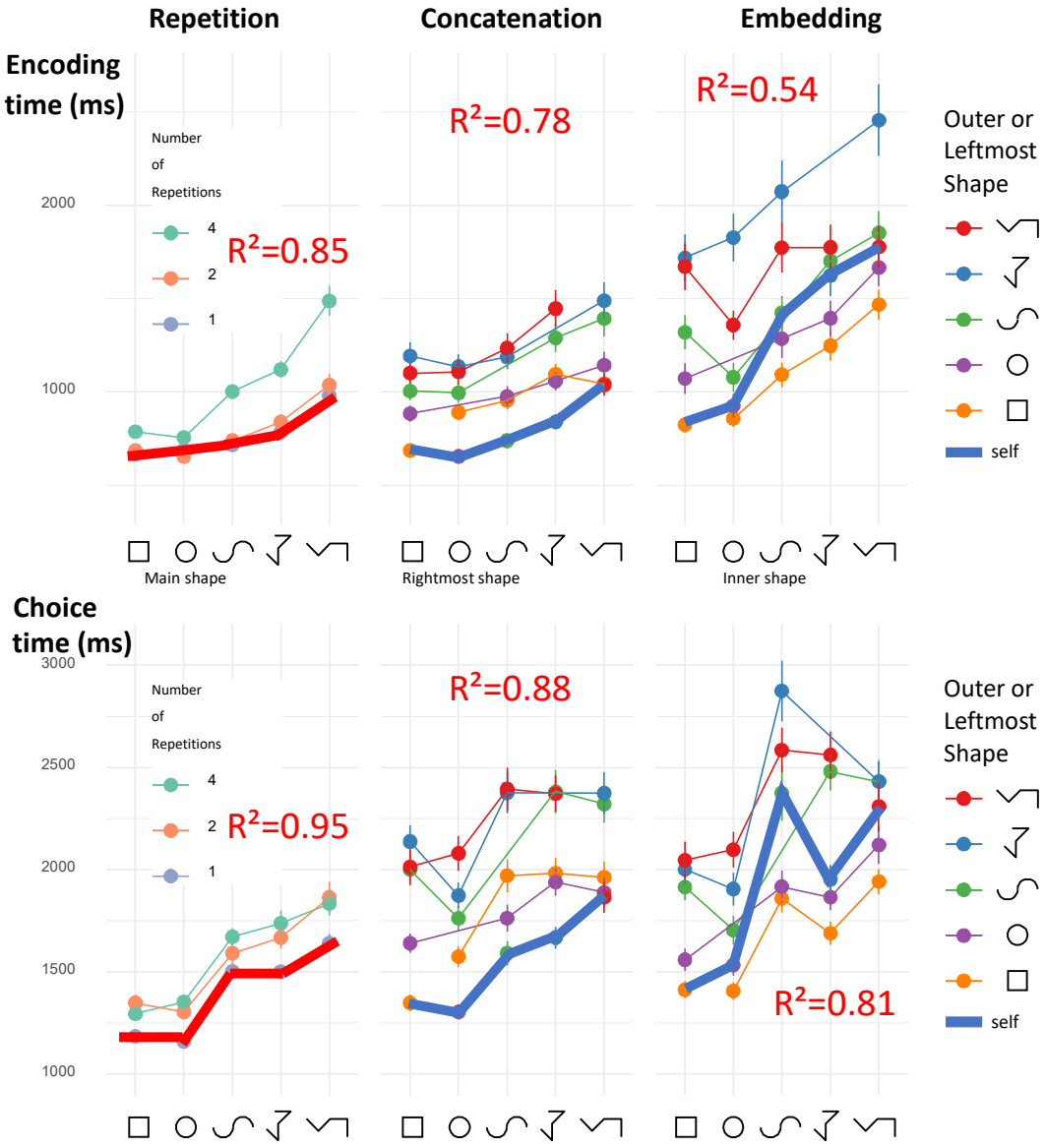
Subjects press a bar, then lift it when they feel they remember the pattern

measure = **encoding time**

Phase 2: Multiple choice

Subjects select the previous pattern among many mutants

measure = **choice time** (and errors)



Testing the predicted additive relationships

There is an effect of shape complexity even for individual shapes → different "programs"

This effect predicts what happens in other conditions:

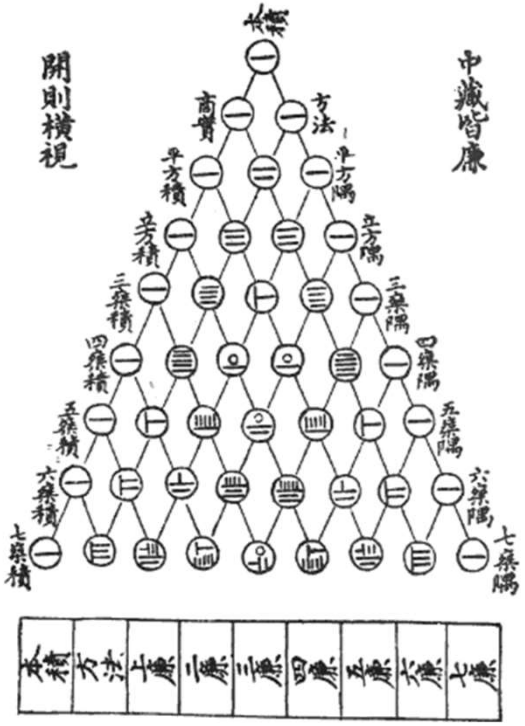
- **Repetition** of a shape n times
  = addition of a term roughly proportional to log(n)

- **Concatenation** of two shapes
  = addition of the two complexities

no interaction term, once we remove the special case of two identical shapes

- **Embedding** of two shapes (e.g. a circle of squares)
  = addition of the two complexities, with steeper slopes

Again, no interaction term, but a special savings when the same program is used twice (e.g. a circle of circles)



Conclusions

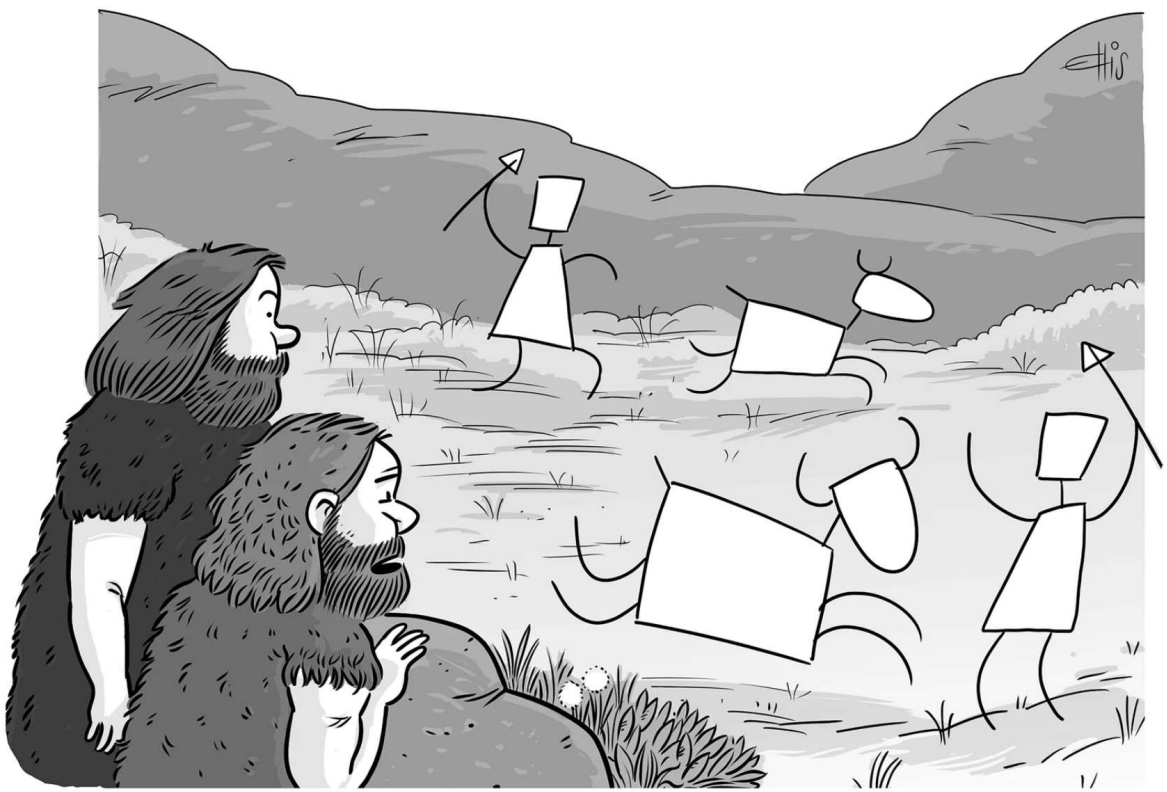
Humans, unlike other non-human primates, possess **mental programs in a language of thought** that

- **Discretizes** concepts :
  numbers, lines, lengths, angles

- Assigns them **symbols that combine** recursively

All human consider the same programs as **simple** (because they are short) → **cross-cultural convergence** towards the same math concepts.

Consequences for AI: we still do not have satisfactory **neural models of symbolic learning**.



"We have to record this, or no one's going to believe us."

Thank you for your attention!

Mathias Sablé Meyer

NeuroSpin
European Research Council
cea
Inserm
PARIS-SACLAY
UNIVERSITÉ PARIS SUD
COLLÈGE DE FRANCE 1530
fondation ROGER DE SPOELBERCH
FONDATION BETTENCOURT SCHUELLER