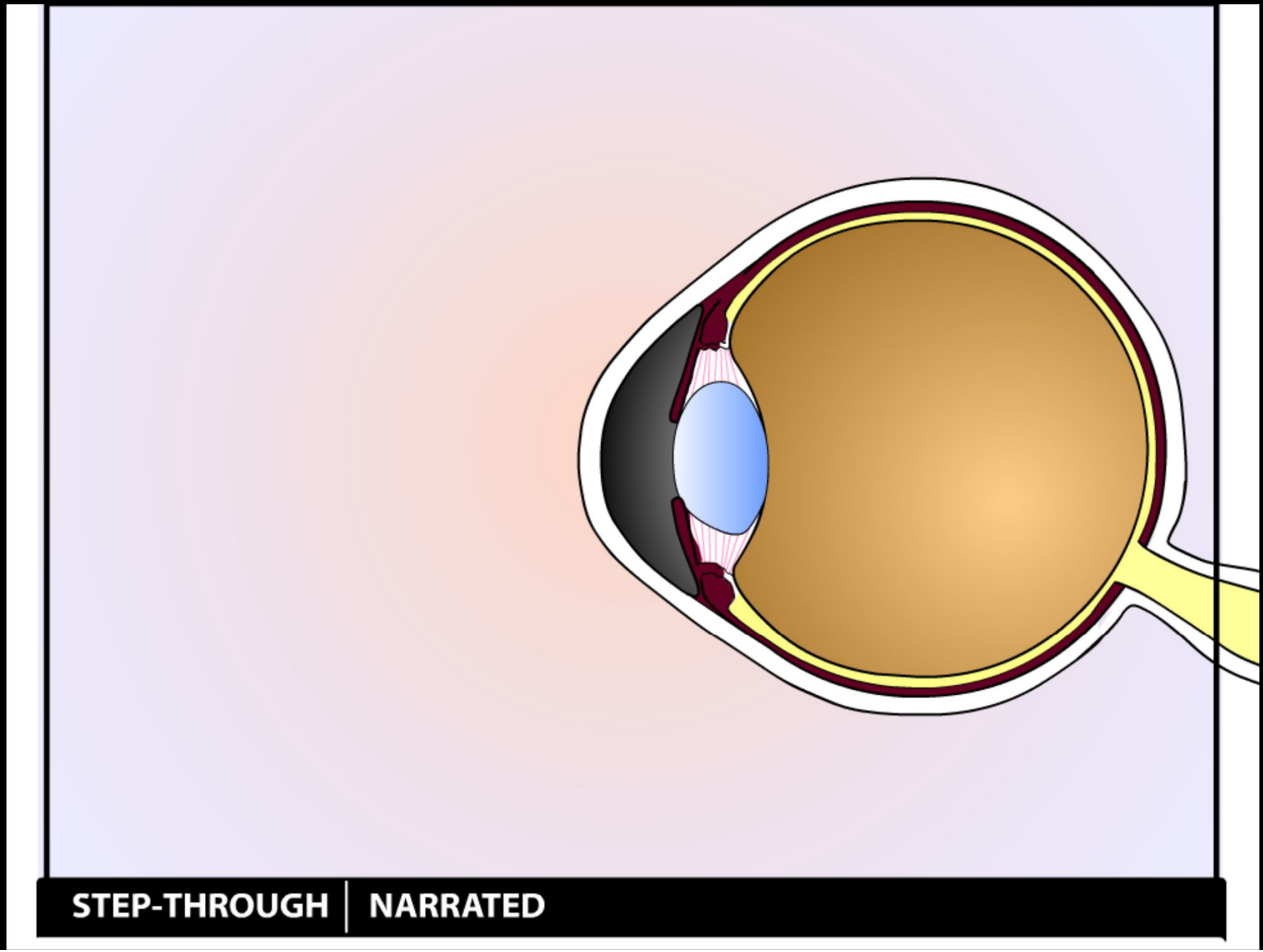
Deep Learning

COGS-Q355

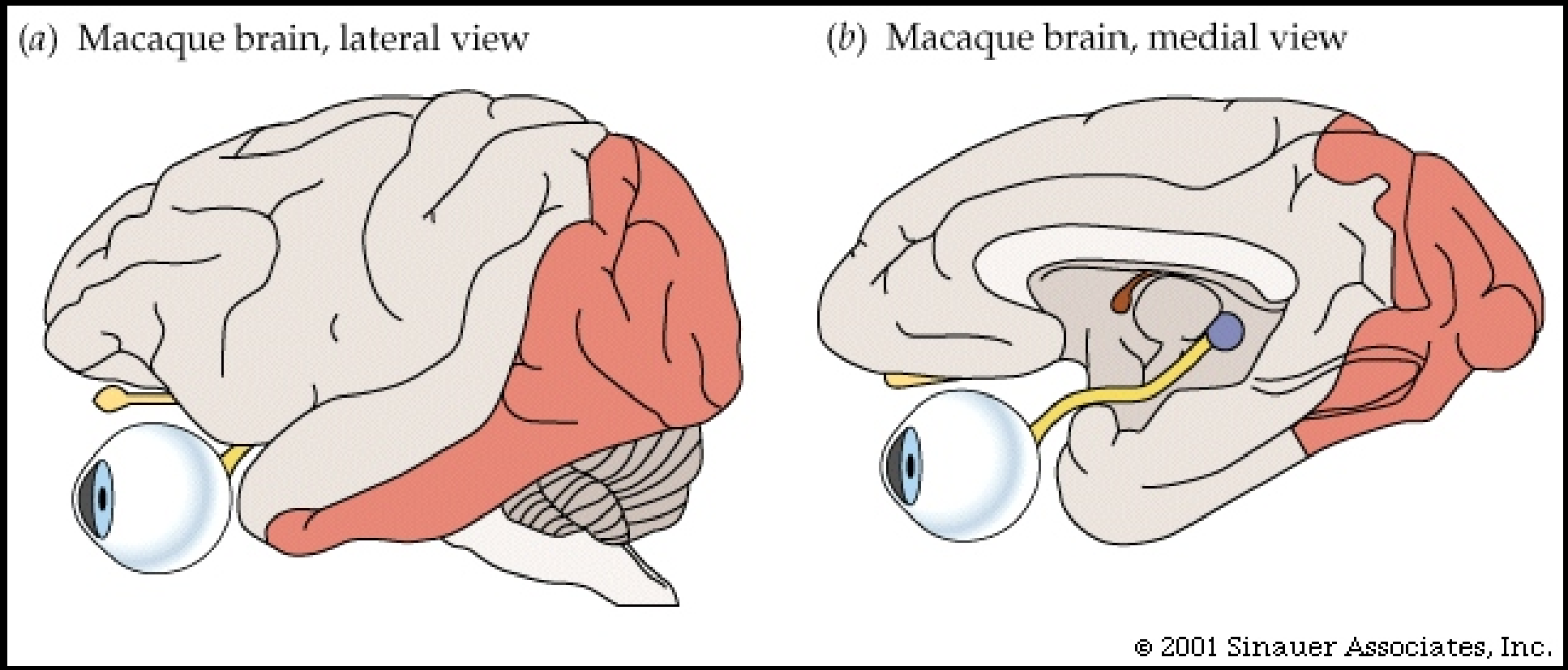
Vision

- The brain's visual system
- Deep learning models of the visual system

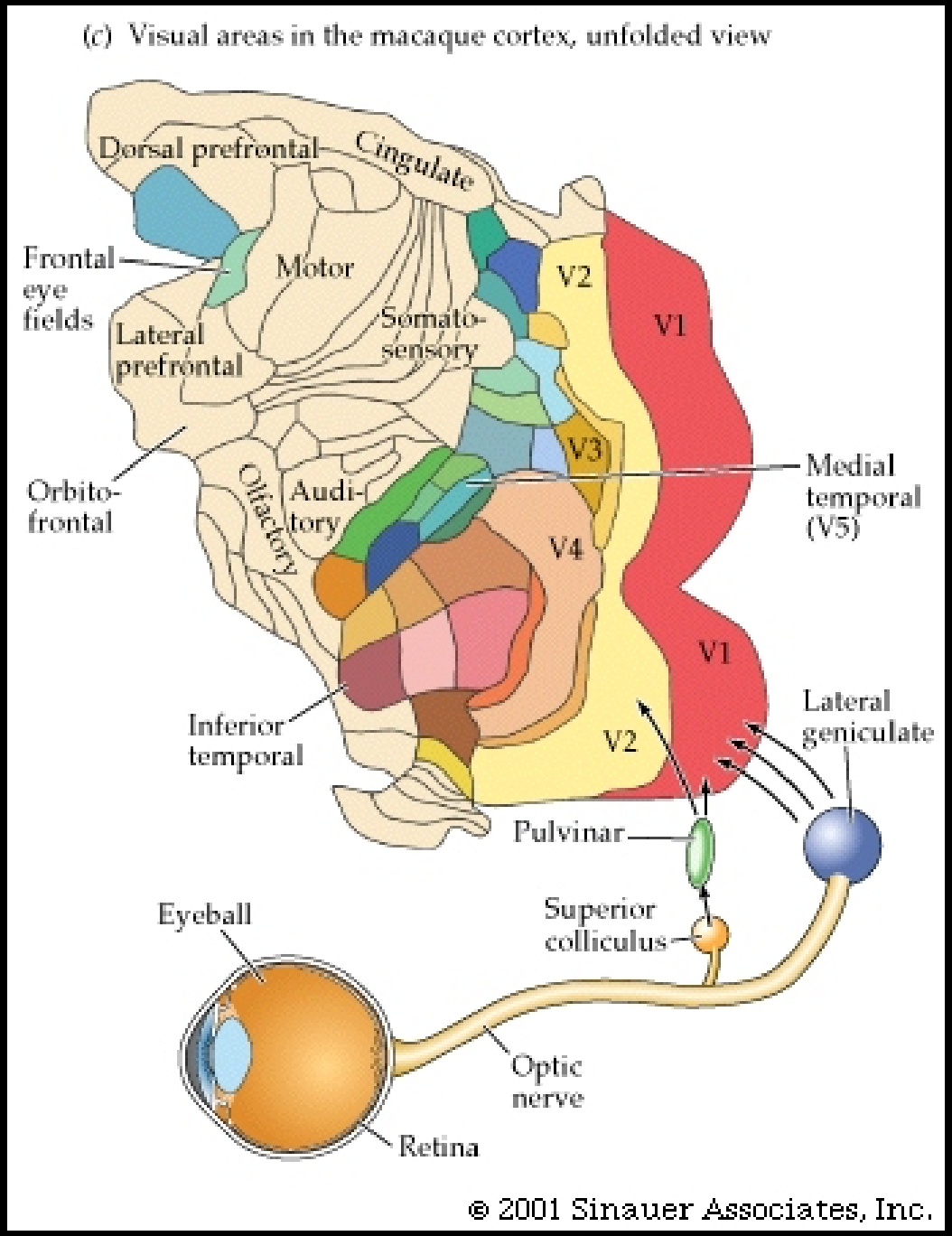
Animation: Light to Brain



Brain Visual Areas



Visual Areas, II



- Lateral geniculate nucleus (LGN) – visual thalamus
- V1 – Primary visual cortex (aka striate cortex)
- Extrastriate cortex – the rest of visual cortex beyond V1
- V2, V4 – elementary shapes
- MT (V5) – Motion
- Inferior Temporal Cortex – cells respond to specific objects (cars, faces, etc.)

V1 Properties

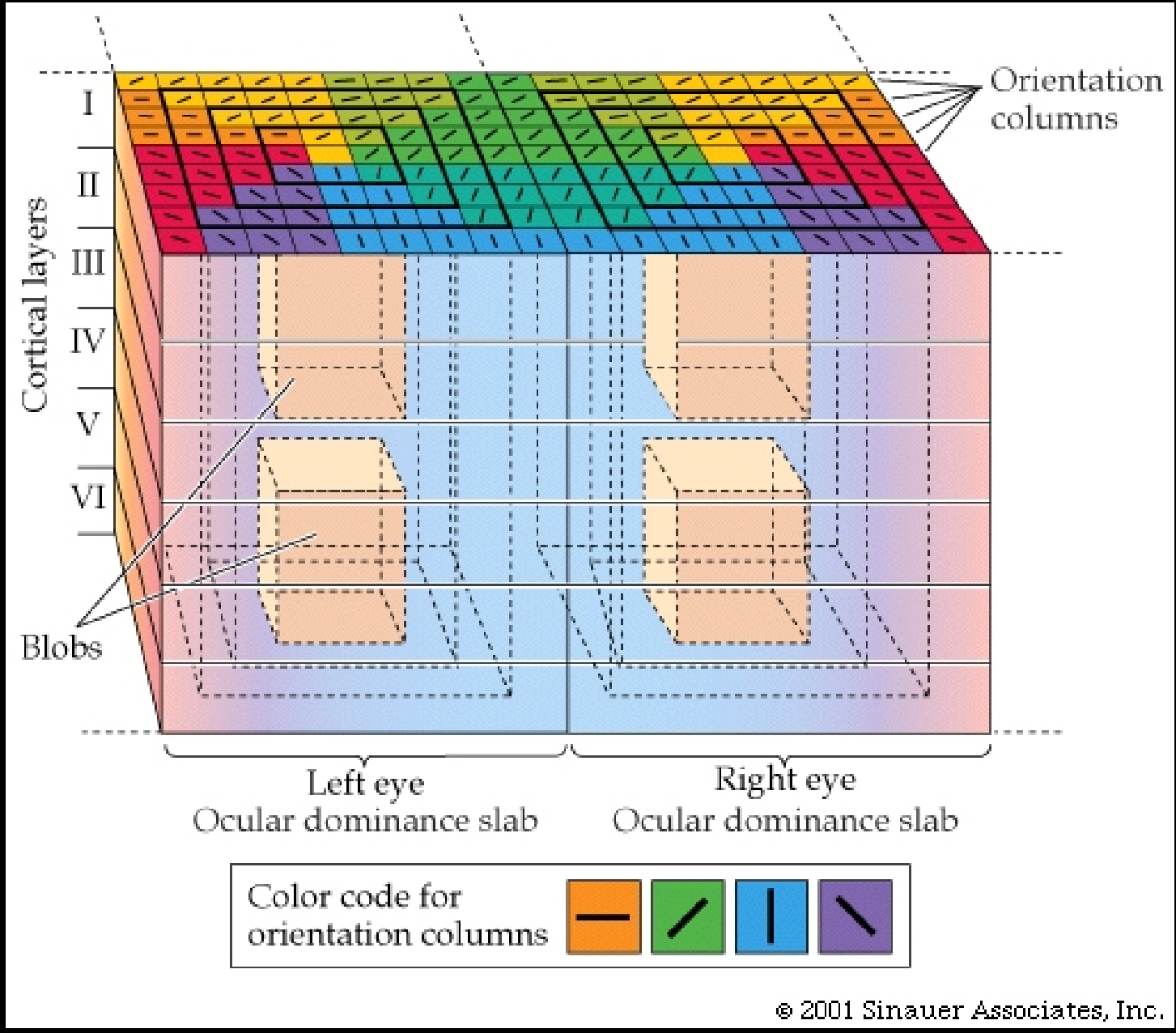
- V1 is primary visual cortex

- Cells are arranged in columns, meaning cells in different layers directly above and below each other have same response properties

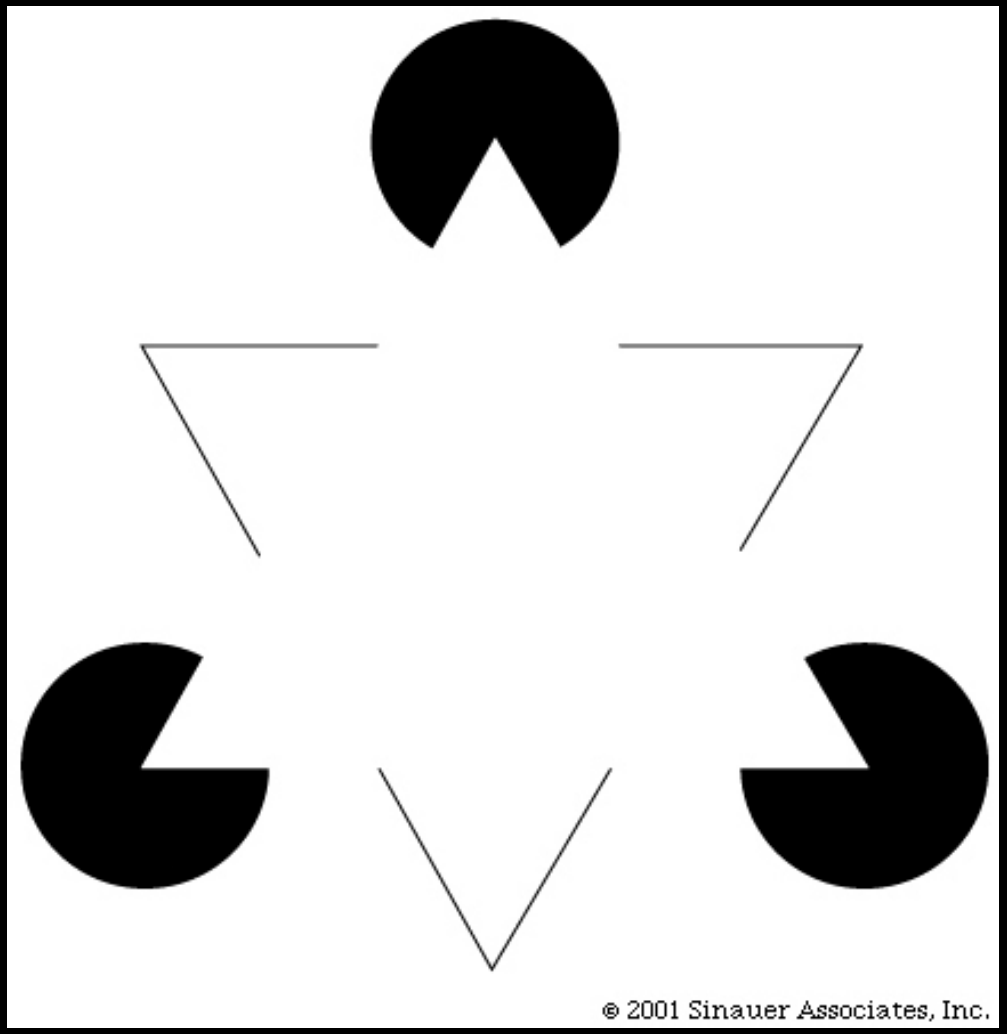
- Cells in V1 respond selectively to:
  - Edges or lines of specific orientations (orientation columns)
  - Input from each of the different eyes (ocular dominance columns)
  - Specific combinations of the above

- Blobs: "The positions of these blobs are closely related to functional organization, particularly ocular dominance, contrast sensitivity, and spatial frequency selectivity. Blobs are centered above ocular dominance columns (Horton 1984) and contain neurons with greater contrast sensitivity and selectivity for lower spatial frequencies than the surrounding interblob regions (Tootell et al 1988a–c, Edwards et al 1995; see also Shoham et al 1997). Although most studies have considered blobs and interblobs to be two distinct compartments, recent evidence suggests that at least some functional properties shift gradually with distance from the blob center. This is reported for contrast sensitivity and spatial frequency selectivity (Edwards et al 1995; but see Shoham et al 1997), but adequate analyses have not been done to assess whether the relationships between selectivity for other stimulus parameters and blobs are binary (blob versus interblob) or shift gradually with distance from blob centers. Regardless, the strong relationship between blobs and functional architecture makes them a useful marker for relating findings from anatomical studies to functional organization." (From Callaway (1998) *Ann. Rev. Neurosci*. 21:47-74)

Ice Cube Model of V1



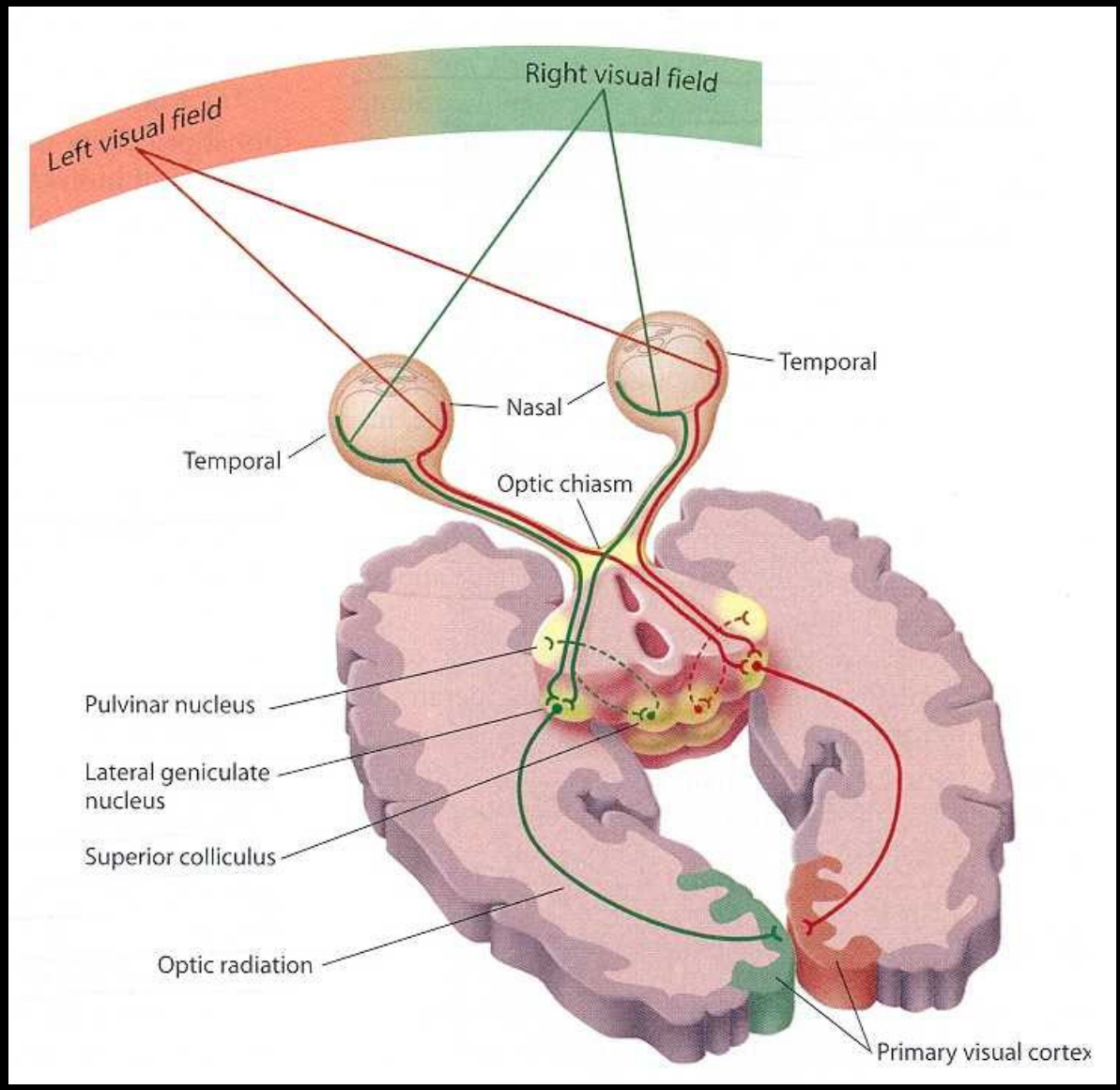
Illusory Contours



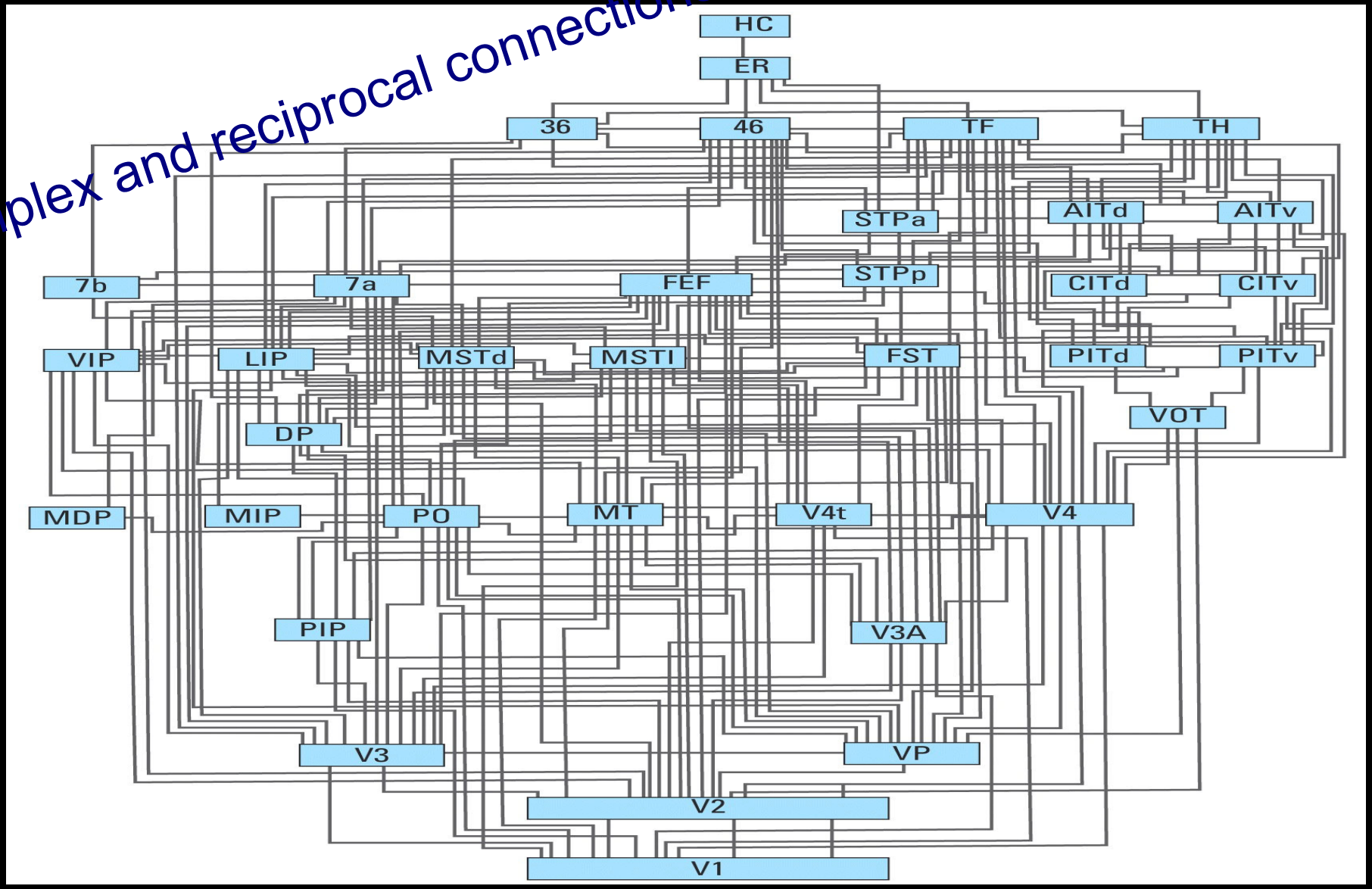
- Some cells in area V2 respond to lines, or even illusions of lines

Disorders of the visual pathways

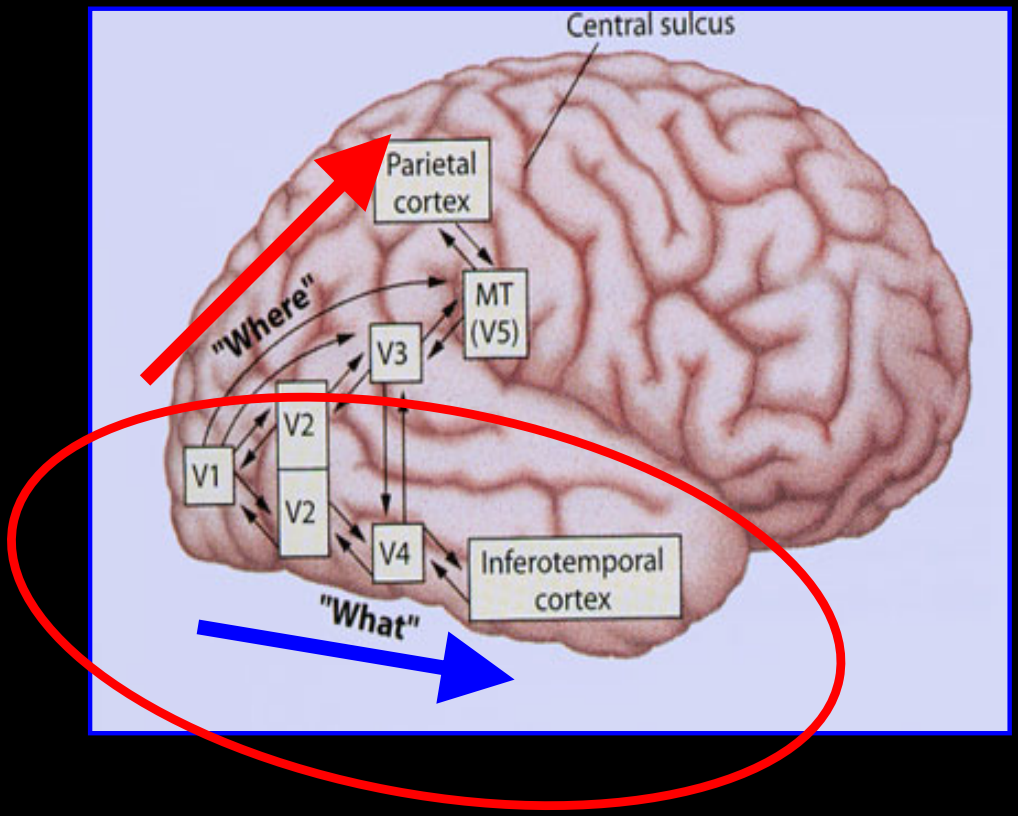
- Right visual field goes to left cortex
- Right VF is represented in BOTH eyes. Temporal left ipsalateral fibers) and nasal right (contralateral fibers).
- Temporal fibers process nasal information.



Connections between visual areas

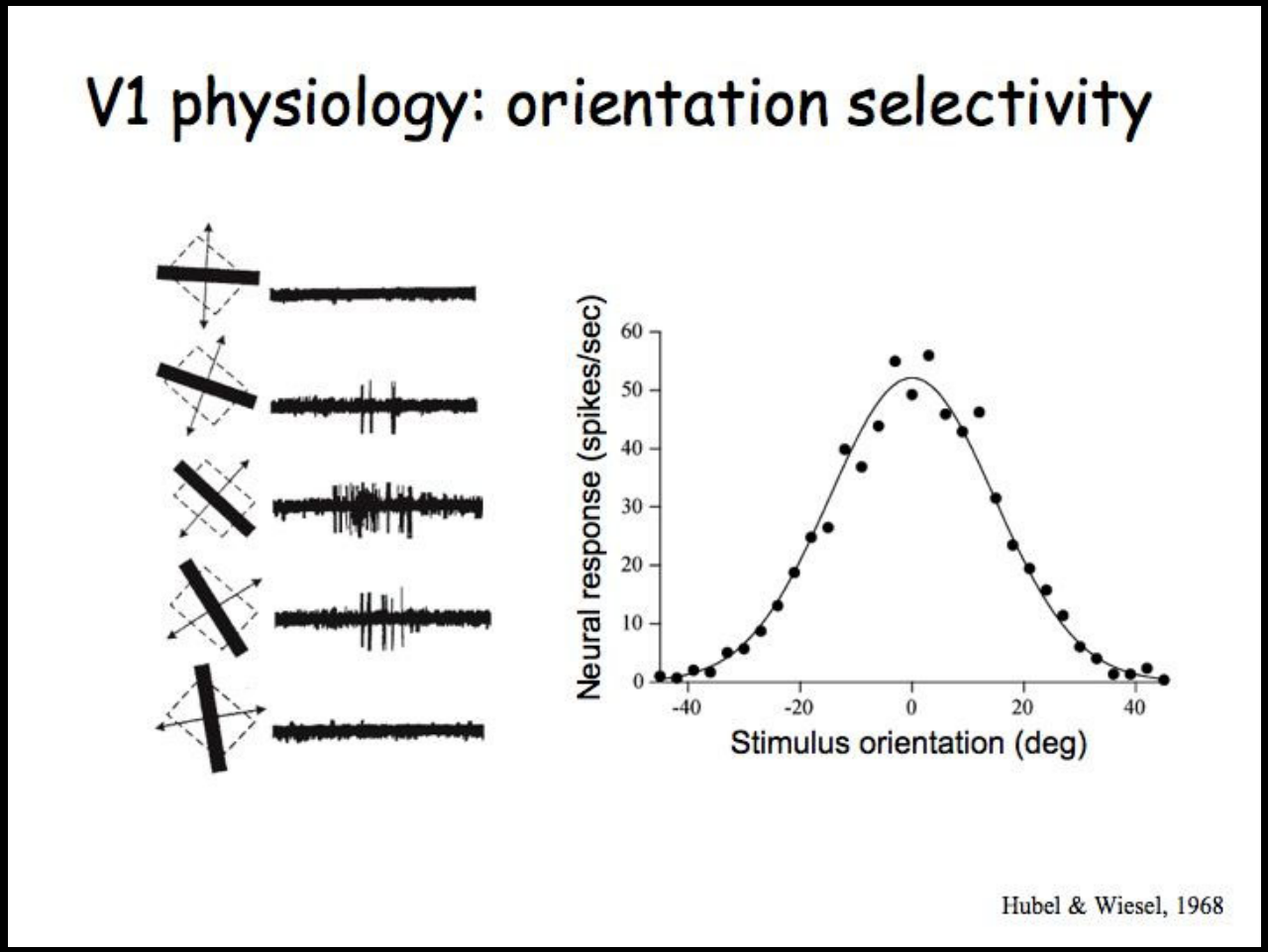


Two visual streams after LGN

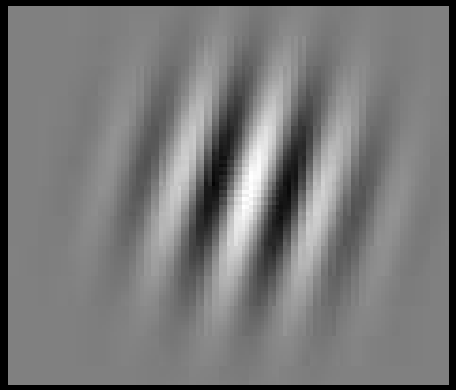


V1 cell receptive fields

- Early visual cells respond particularly to edges

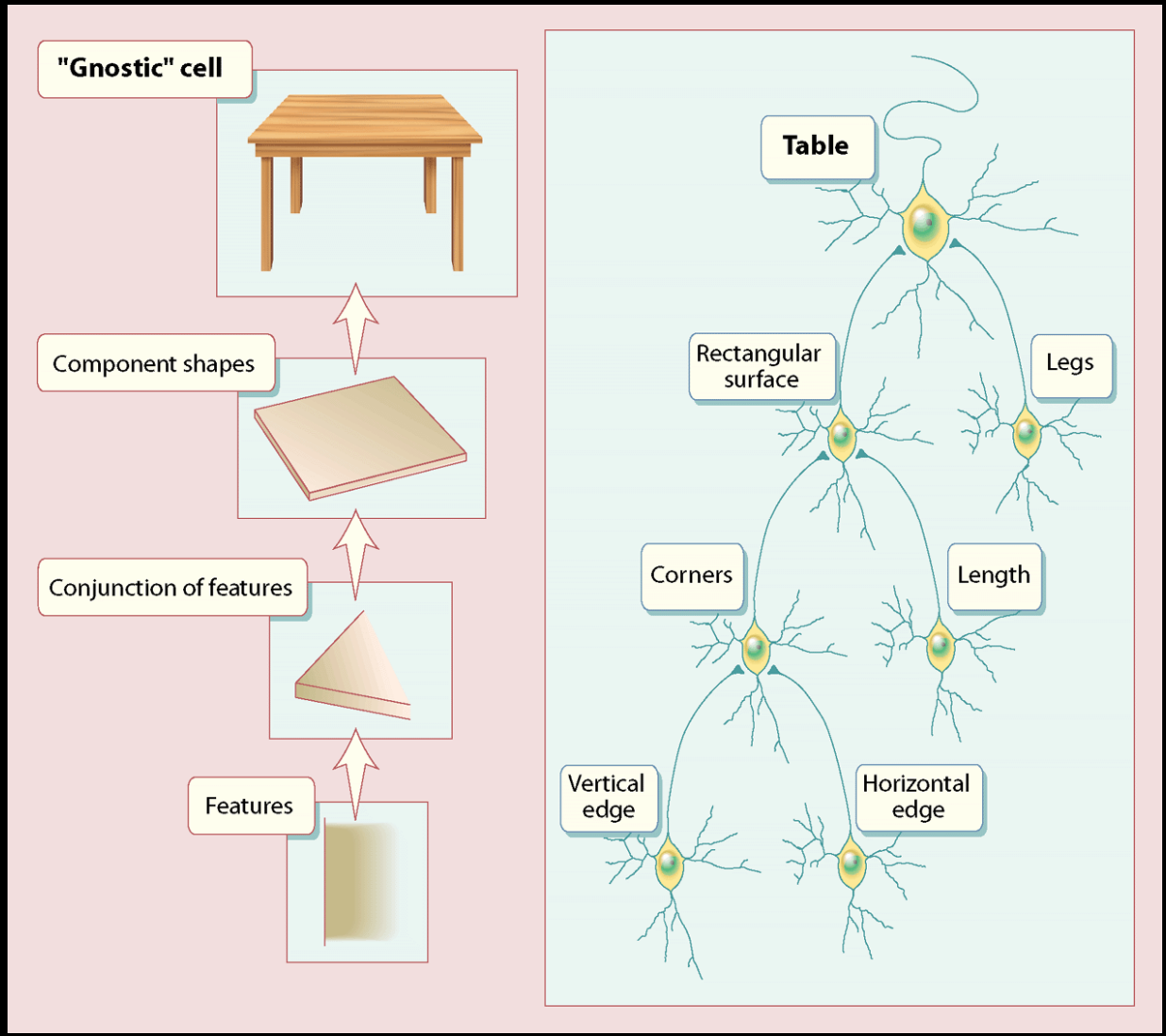


Receptive field



Hierarchical coding hypothesis

- Gnostic cells respond to specific known objects
- In hierarchical coding, gnostic cells respond to complex combinations of shape representations, which in turn respond to combinations of lower level features. There is a hierarchical coding of object properties.
- The "grandmother cell" is a hypothetical cell that would respond only when grandmother comes into view

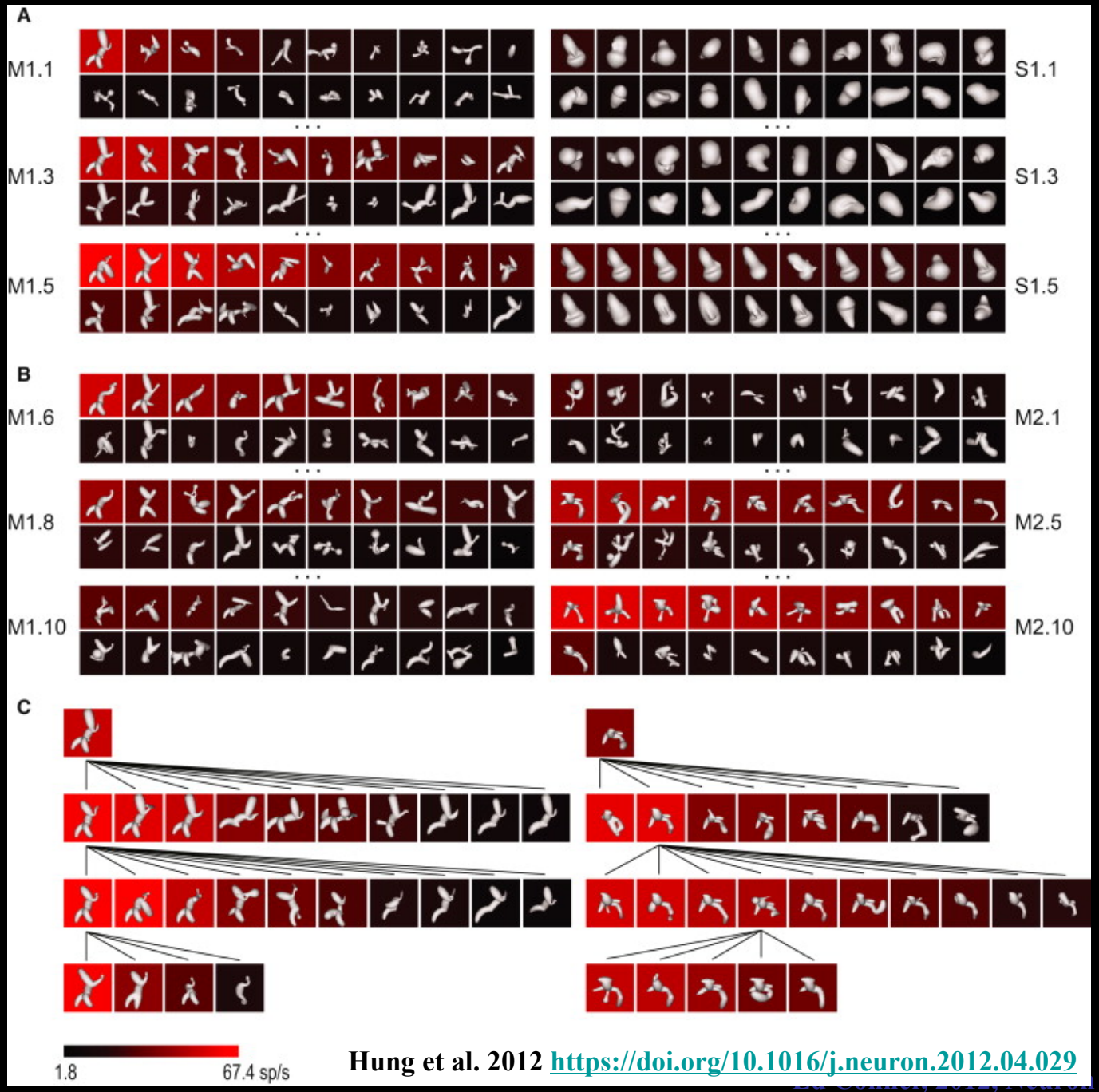


Grandmother cell

- The grandmother cell becomes active only when the specific combination of features unique to grandmother is present.

Neuroscience – anterior inferotemporal cortex

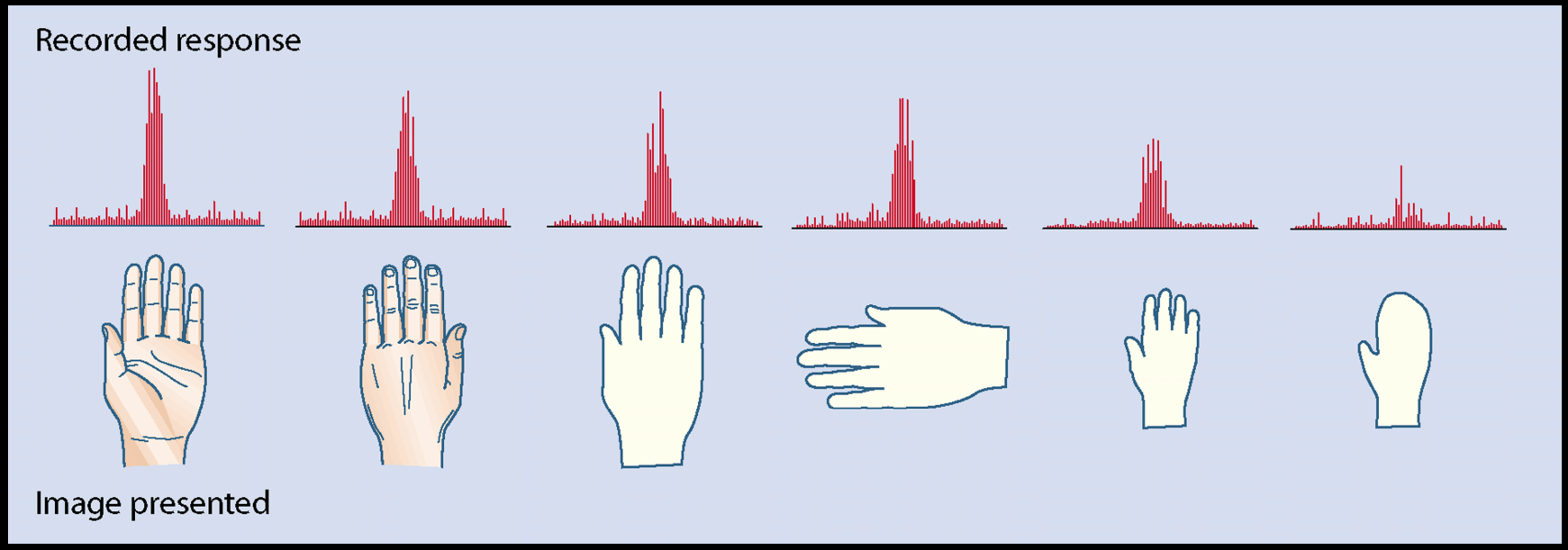
- Recent finding that IT neurons code for all sorts of abstract shapes
- Surfaces
- Skeletal shapes (medial axis transform, orientation invariant)

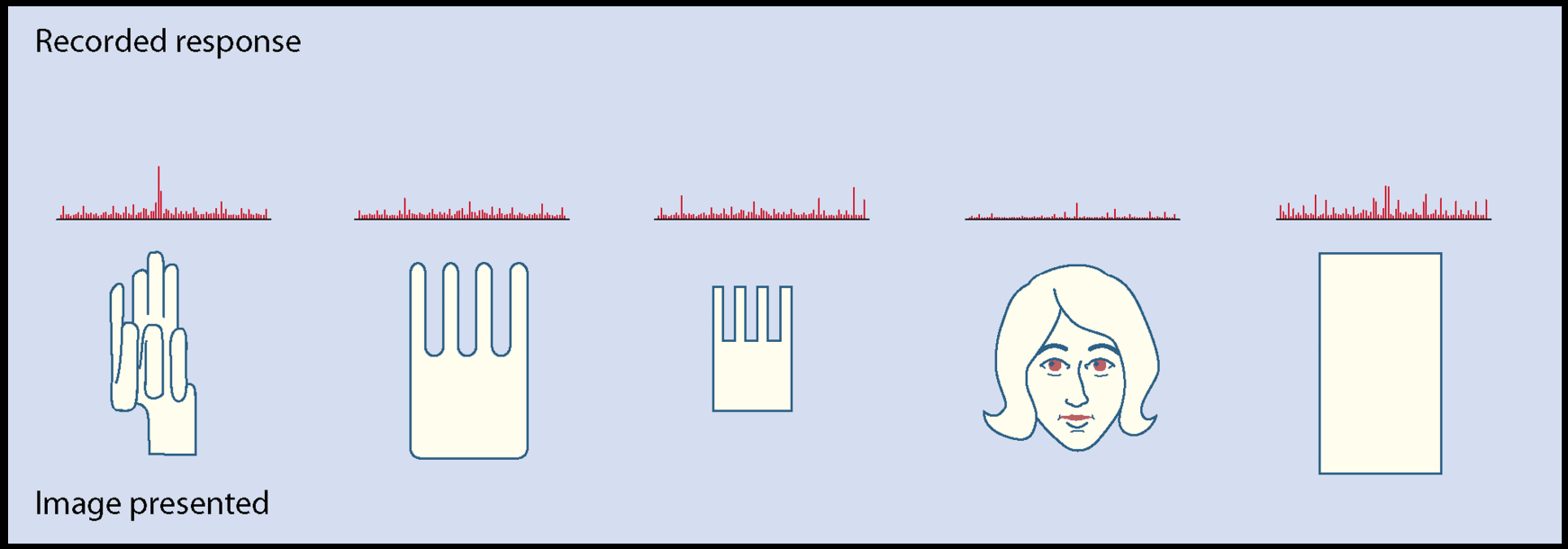


Hung et al. 2012 https://doi.org/10.1016/j.neuron.2012.04.029

Anterior Inferior temporal cortex

- Monkey Inferior temporal cortex cell responds to hands but not other objects





Deep convolutional neural net

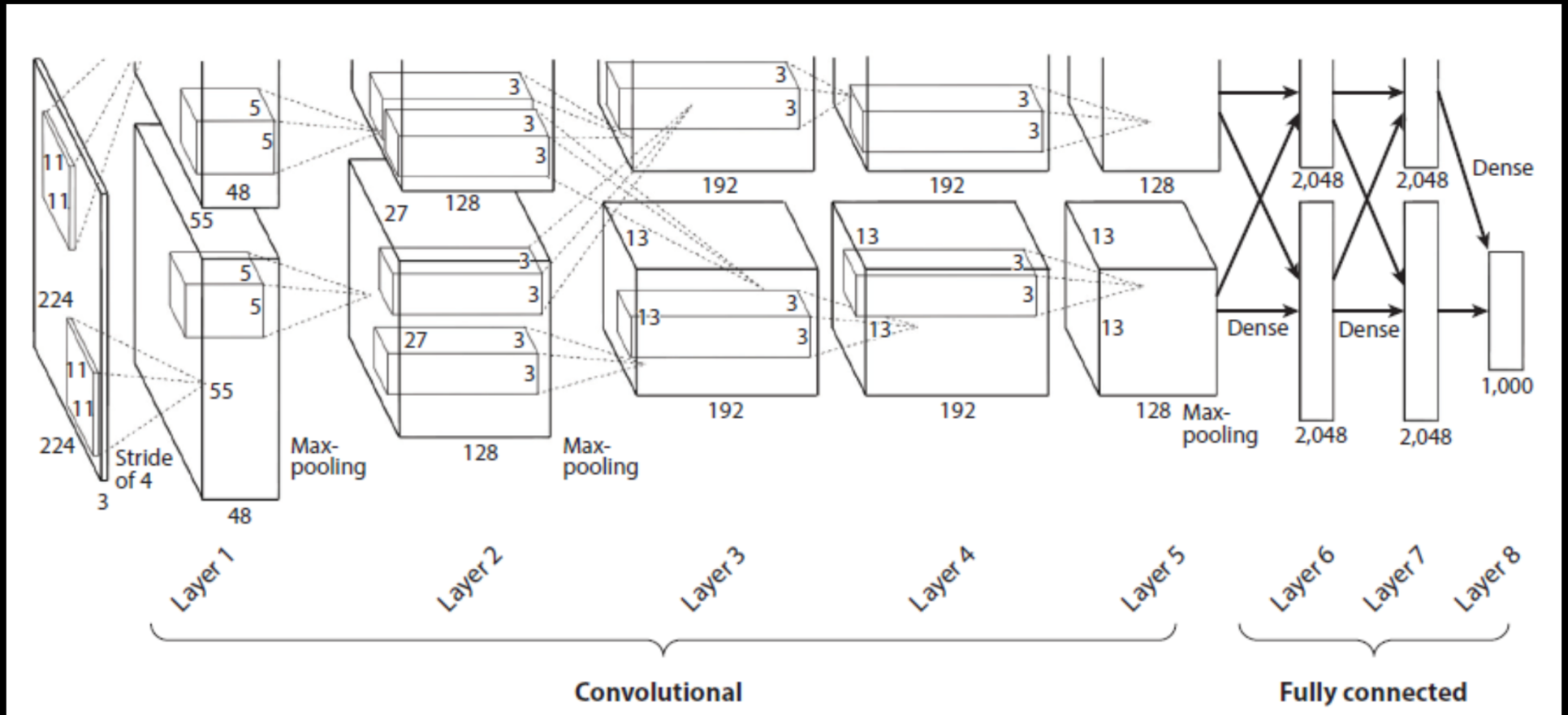
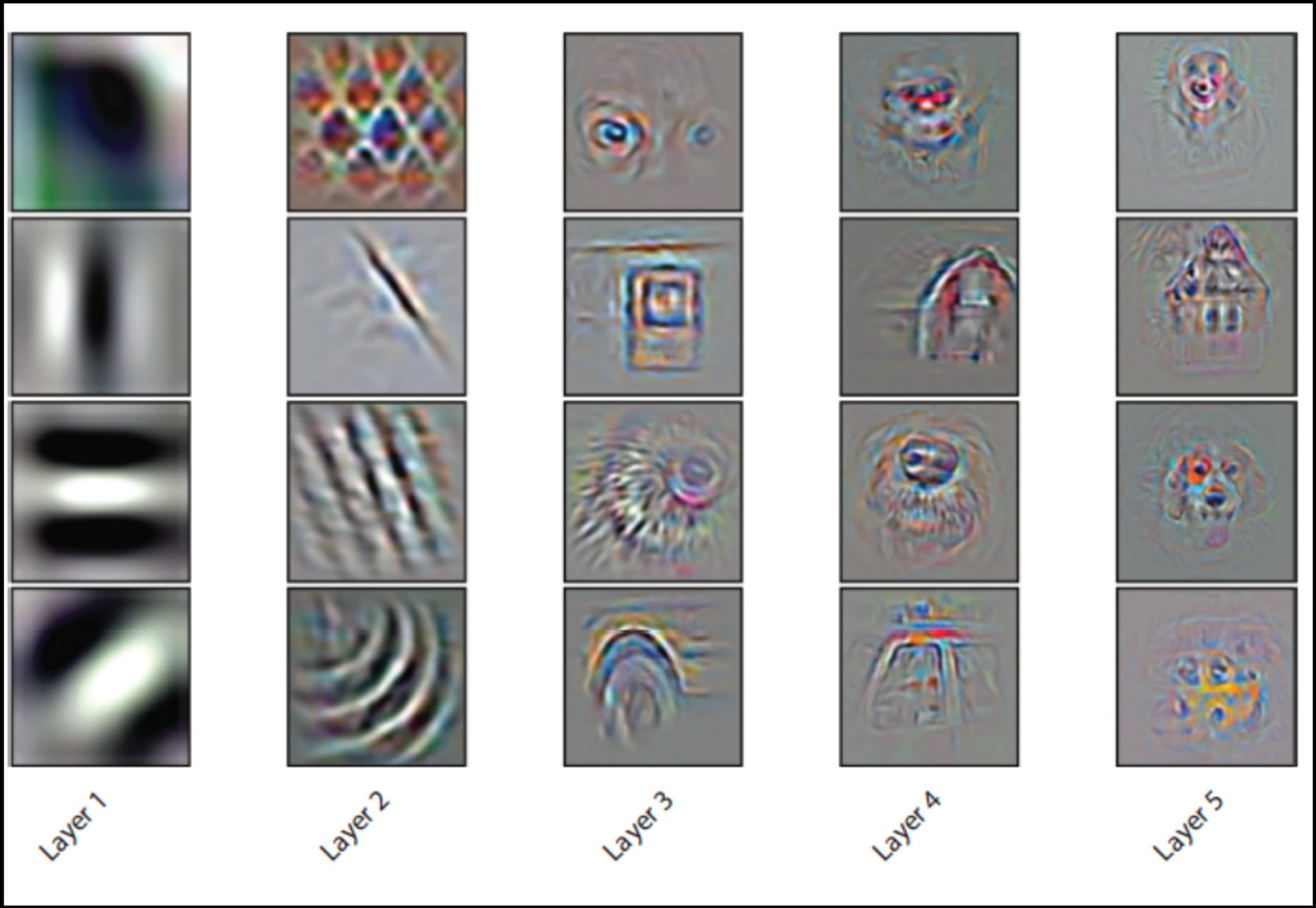


Figure 3

Deep convolutional feedforward architecture for object recognition. The figure shows the architecture used by Krizhevsky et al. (2012).

Kriegeskorte, 2015

Representation in deep networks



shown) can help us understand its selectivity and tolerances. The deconvolutional visualization technique shown was developed by Zeiler & Fergus (2014). The deep network is from Chatfield et al. (2014). The analysis was performed by Güçlü & van Gerven (2015). Figure adapted with permission from Güçlü & van Gerven (2015).

Kriegeskorte, 2015

DeepDream

- What if we run a CNN backwards given an output? What will the input look like? (i.e. do gradient descent on the input, not the weights)
- Answer: it depends on what the network was trained on.

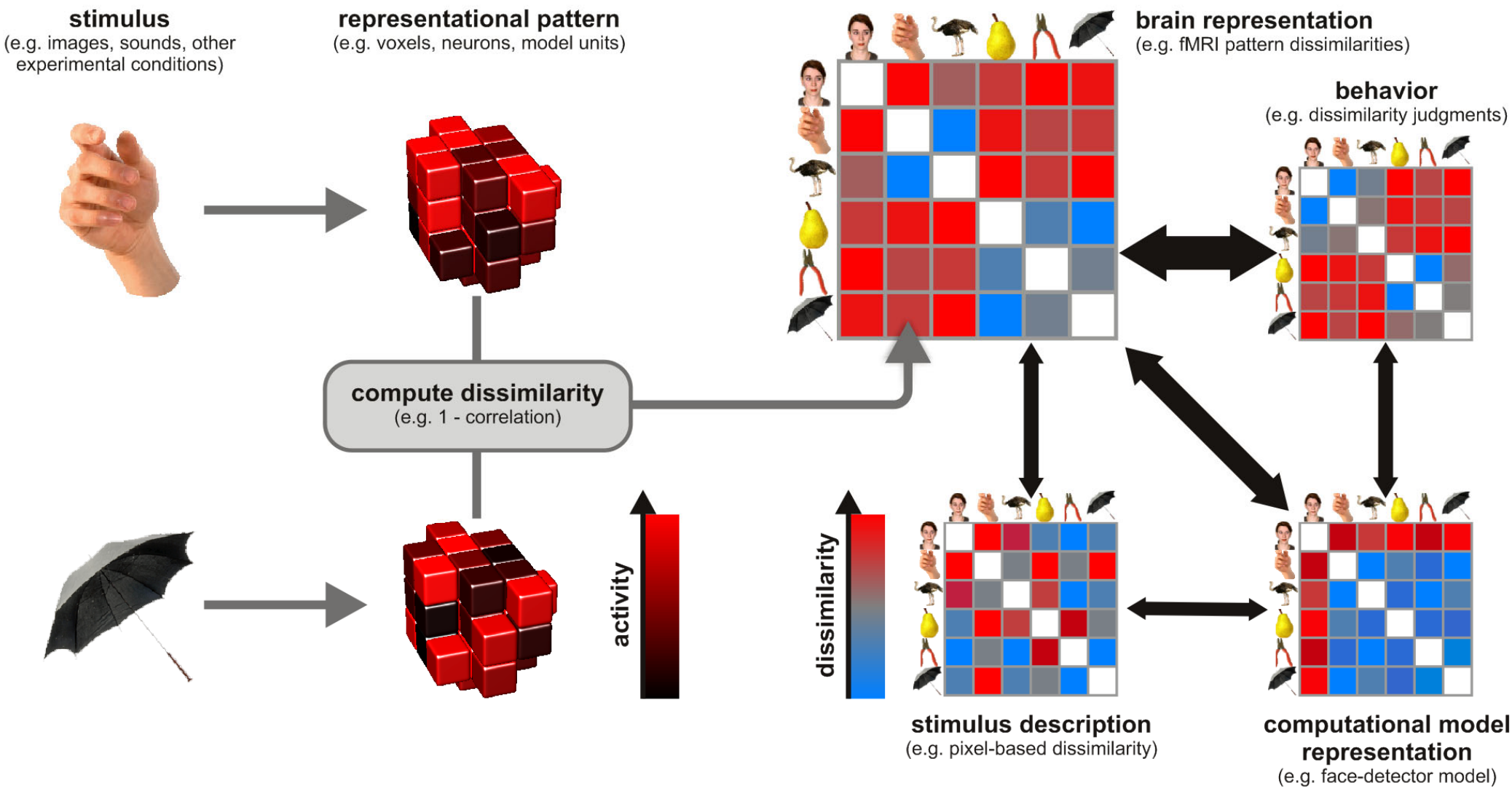
Trained on birds, desired output is night scene

https://deepdreamgenerator.com

DeepDream

https://deepdreamgenerator.com

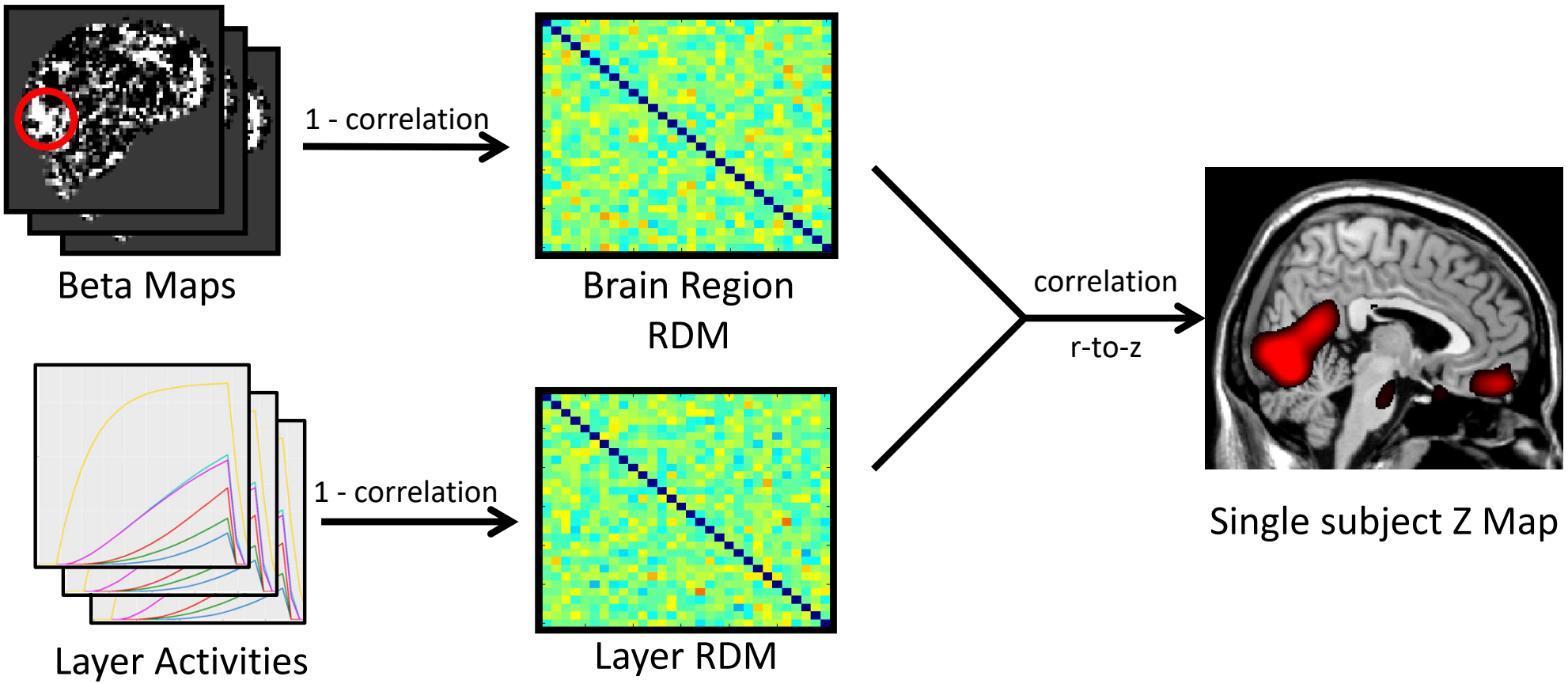
Representational Similarity Analysis



Representational Similarity Analysis

Do any brain regions share pattern similarity structure with model components?

To test this, the model was augmented with a planning module before completing the goal pursuit task.



Deep network vs. human ventral visual stream

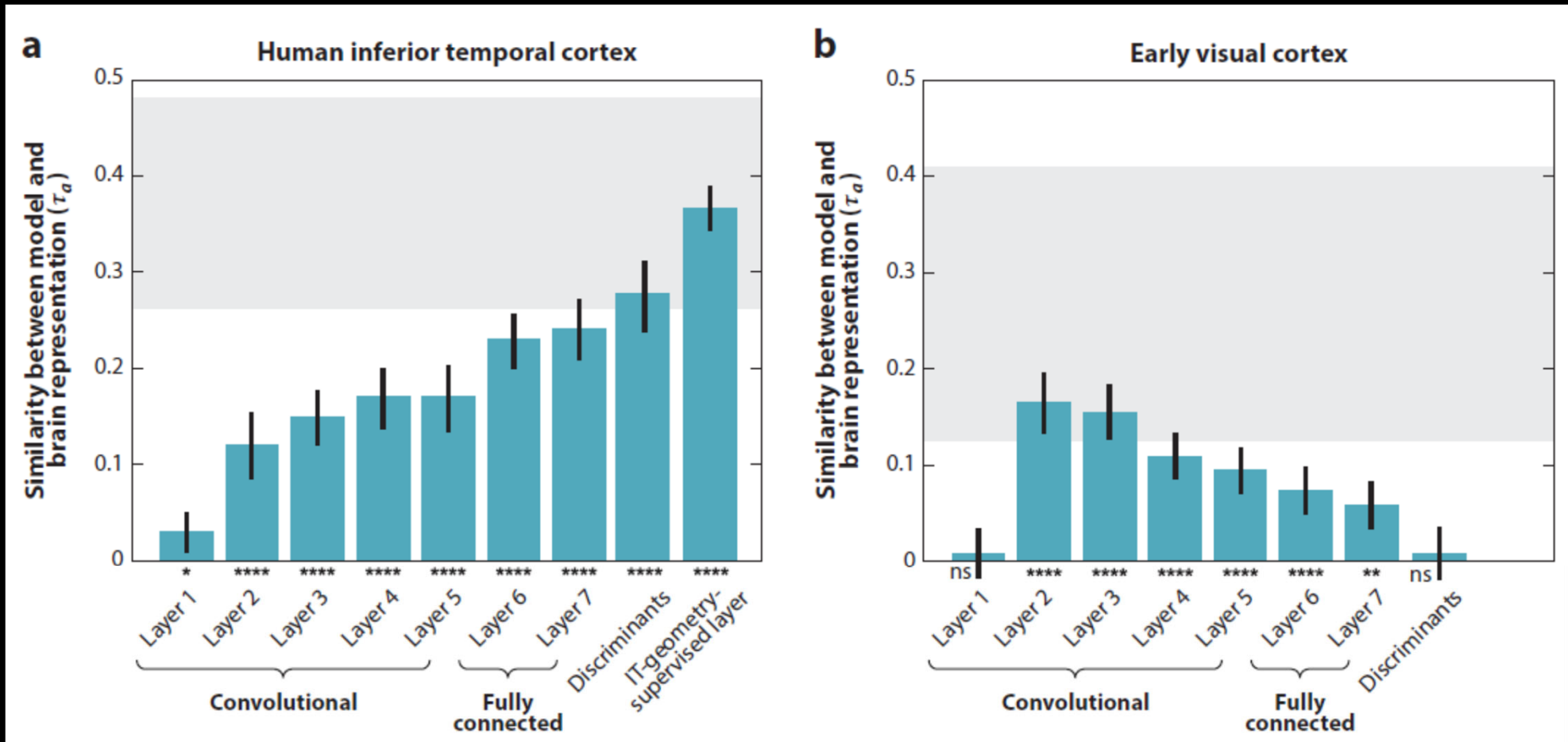


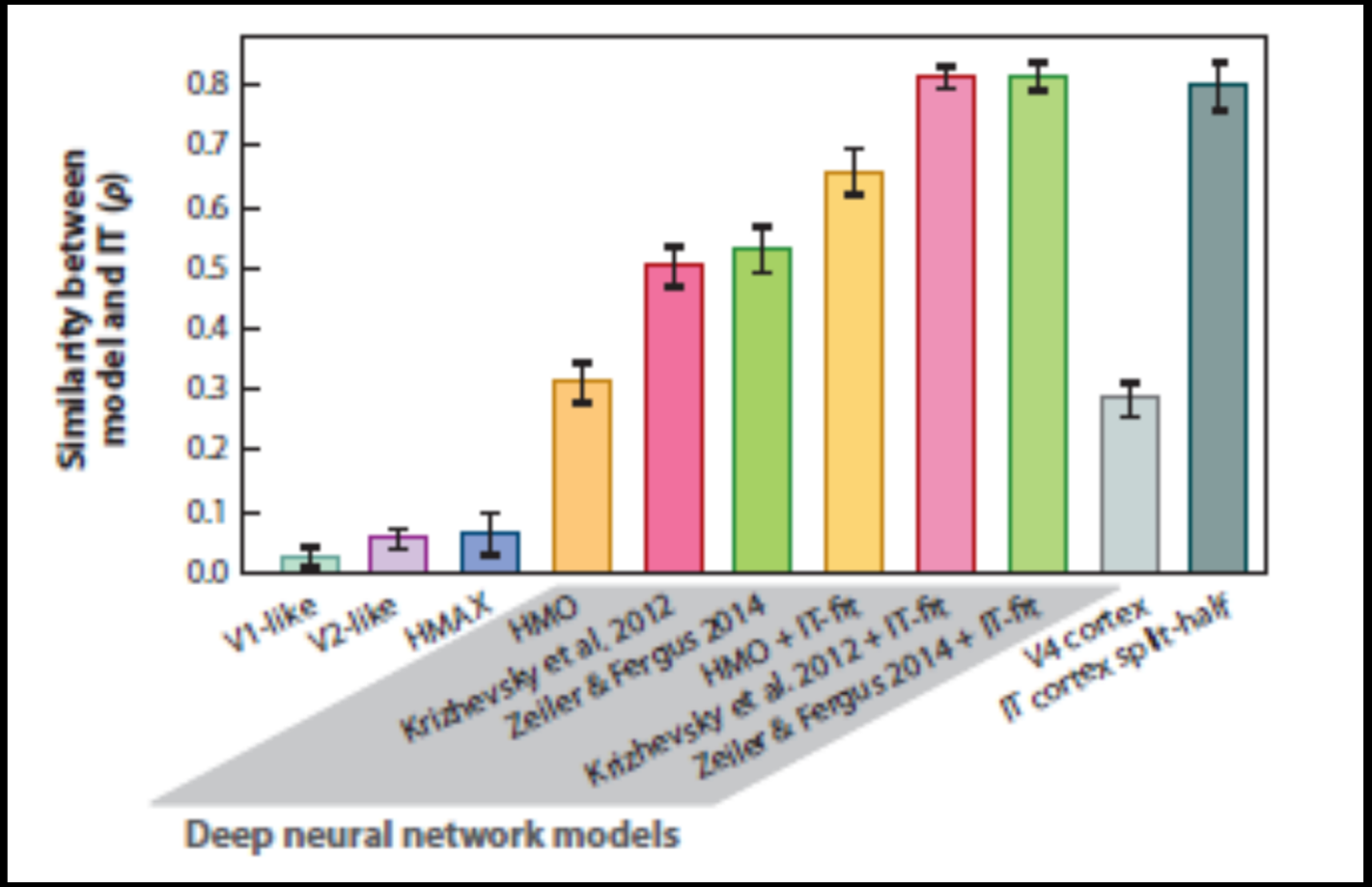
Figure 5

Deep neural network explains early visual and inferior temporal representations of object images. Each representation in model and

Kriegeskorte, 2015

Deep network vs. human ventral visual stream

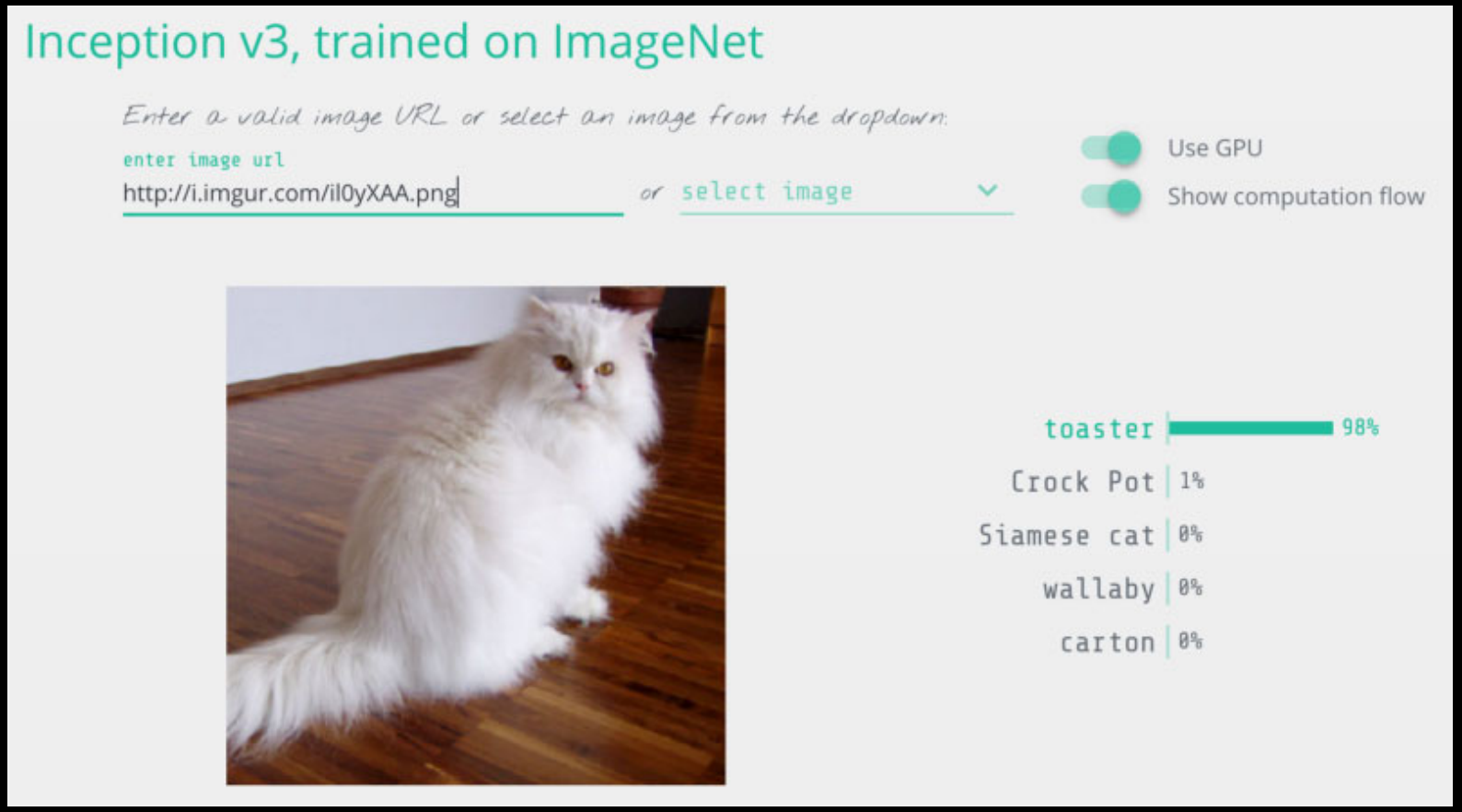
- Deep neural networks



Kriegeskorte, 2015

How are neural nets UNLIKE biology?

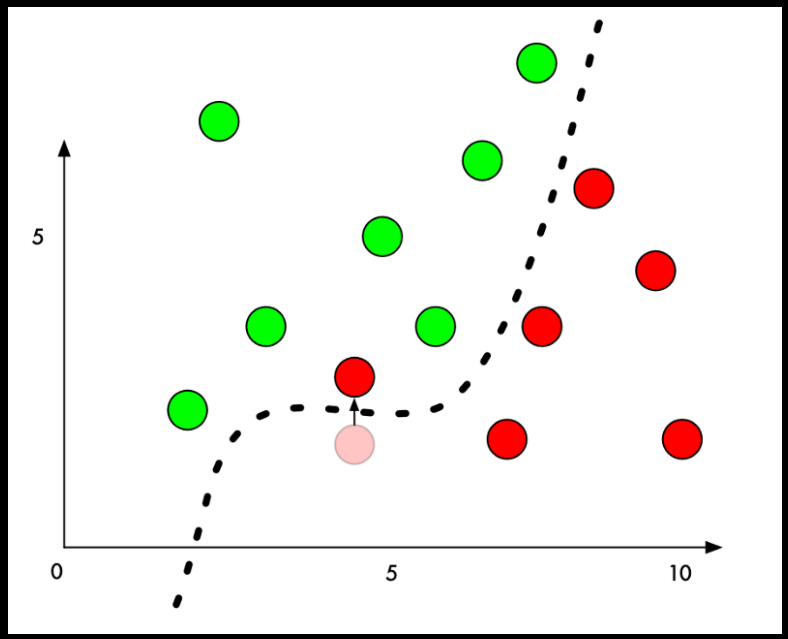
- Biological plausibility issues of backpropagation, already discussed
- Can you “hack” a neural network?
- Surprisingly EASY to trick a deep network image classifier – overfitting?



https://medium.com/@ageitgey/machine-learning-is-fun-part-8-how-to-intentionally-trick-neural-networks-b55da32b7196

How are neural nets UNLIKE biology?

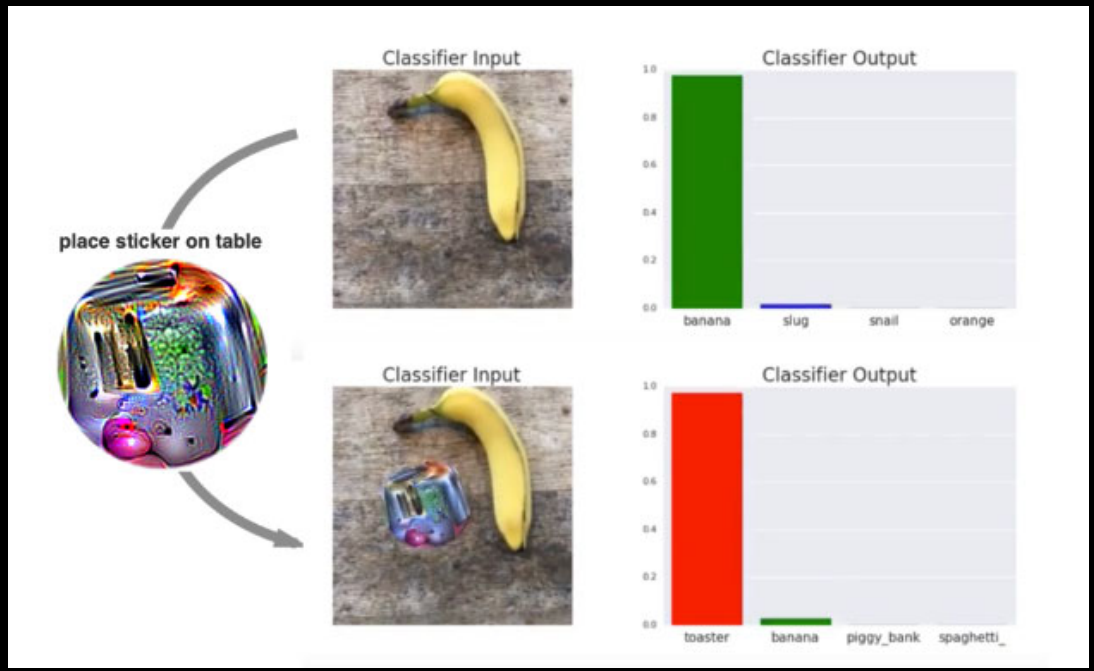
- How does this work?
- Typically, gradient descent adjusts the weights to generate the desired output
- What if instead, we fix the weights, then adjust an input image to generate a desired output
- Constrain the image changes so that no pixel is changed by more than a small percentage
- RESULT: Hacked image looks like original but still gets mis-classified with high confidence

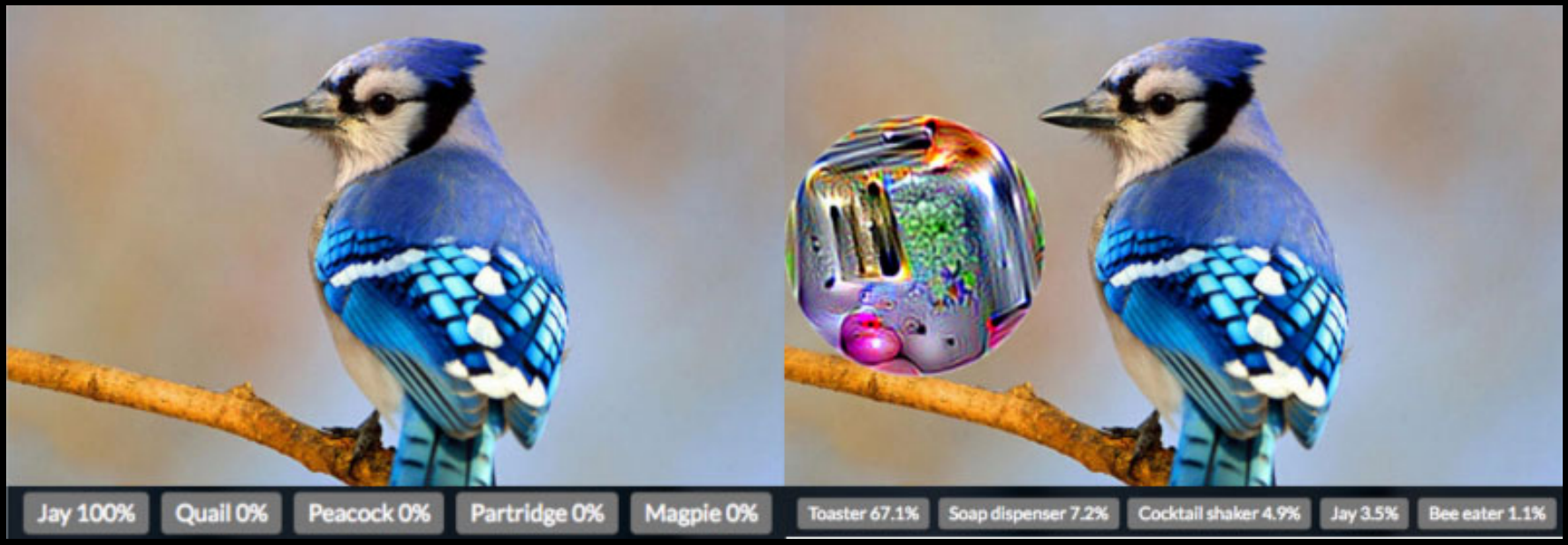


https://medium.com/@ageitgey/machine-learning-is-fun-part-8-how-to-intentionally-trick-neural-networks-b55da32b7196

More ways to fool a visual deep network

- The Adversarial patch
- Use gradient descent on an unconstrained, blank image to generate an image that optimally produces a desired classifier output
- Instead of adjusting weights via dE/dW, fix the weights and adjust the input by dE/dInput





https://towardsdatascience.com/lets-fool-a-neural-network-b1cded8c4c07

How to fix the problem?

- How might we develop networks that are not so easily fooled by “hacked” images?
  - Include “hacked” images in the training set
- What does the susceptibility to hacking tell us about the neural representations?

https://medium.com/@ageitgey/machine-learning-is-fun-part-8-how-to-intentionally-trick-neural-networks-b55da32b7196