using deep learning models to understand visual cortex

11-785 Introduction to Deep Learning
Fall 2017

Michael Tarr
Department of Psychology
Center for the Neural Basis of Cognition

this lecture

- A bit out of order…
- Oct 2: Models of Vision, AlexNet, VGG
- Today: Are computer vision models useful for understanding biological vision?

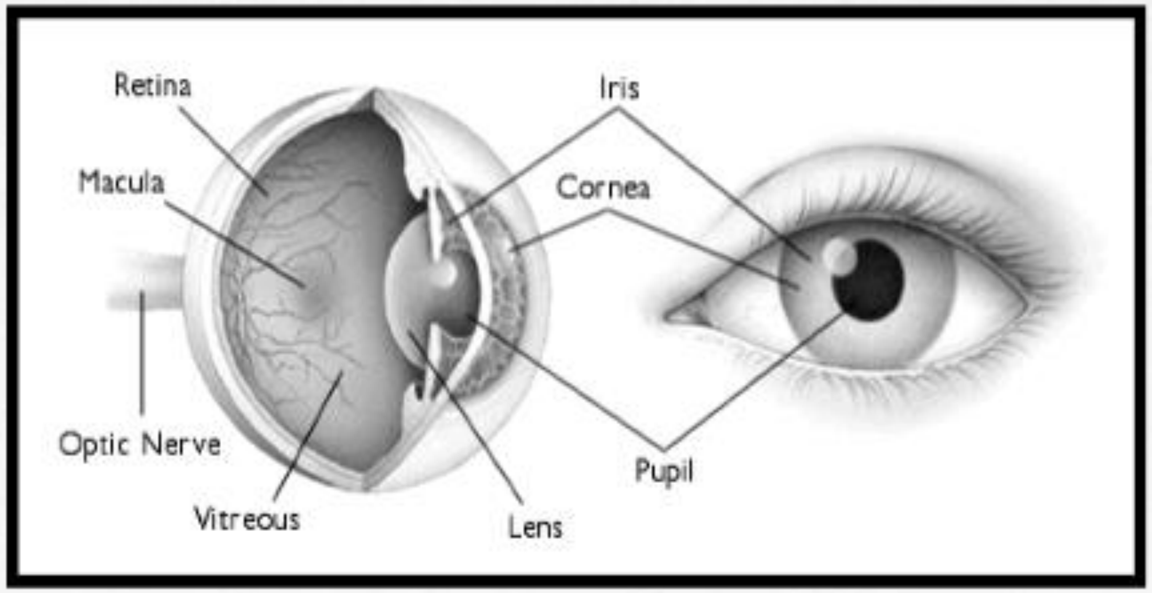
1. Background
   - Biological Vision
   - CNNs
2. Comparisons
3. Models of representation
4. Sandboxes

some numbers (ack)

- Retinal input (~10^8 photoreceptors) undergoes a 100:1 data compression, so that only 10^6 samples are transmitted by the optic nerve to the LGN
- From LGN to V1, there is almost a 400:1 data expansion, followed by some data compression from V1 to V4
- From this point onwards, along the ventral cortical stream, the number of samples increases once again, with at least ~10^9 neurons in so-called "higher-level" visual areas
- Neurophysiology of V1->V4 suggests a feature hierarchy, but even V1 is subject to the influence of feedback circuits – there are ~2x feedback connections as feedforward connections in human visual cortex
- Entire human brain is about ~10^{11} neurons with ~10^{15} synapses

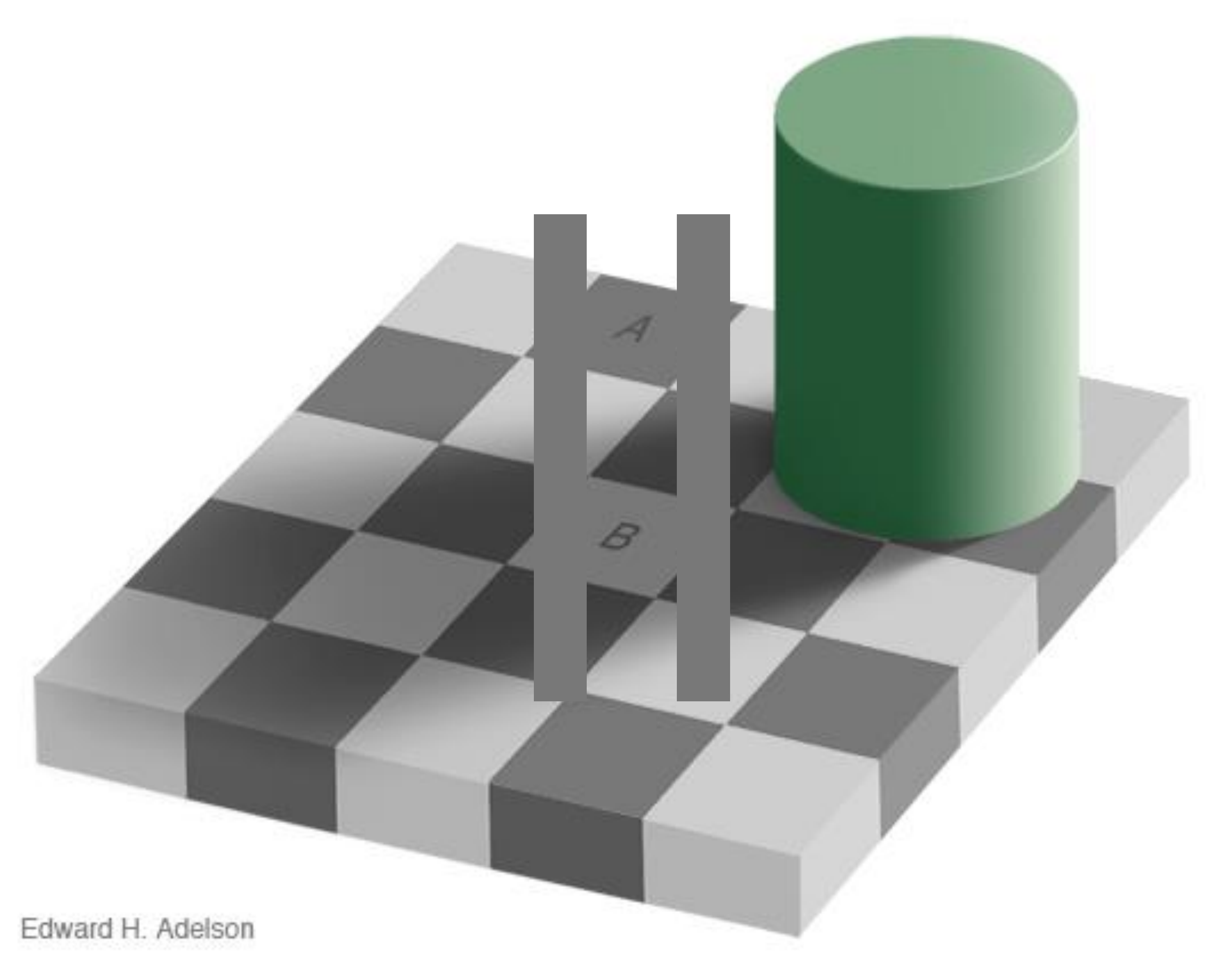
biological vision
the eye is not a camera

- cameras reproduce an image by focusing rays of light on a flat surface
- eyes focus rays of light on our retinae as the first step of visual perception



vision as inference

- we do not reconstruct the 3D world in our heads
- we are presented with a 2D dynamic image of a 3D world and draw inferences about the structure of this world
- most inferences are based on assumptions
- assumptions are simply learned statistics


A
B
Edward H. Adelson

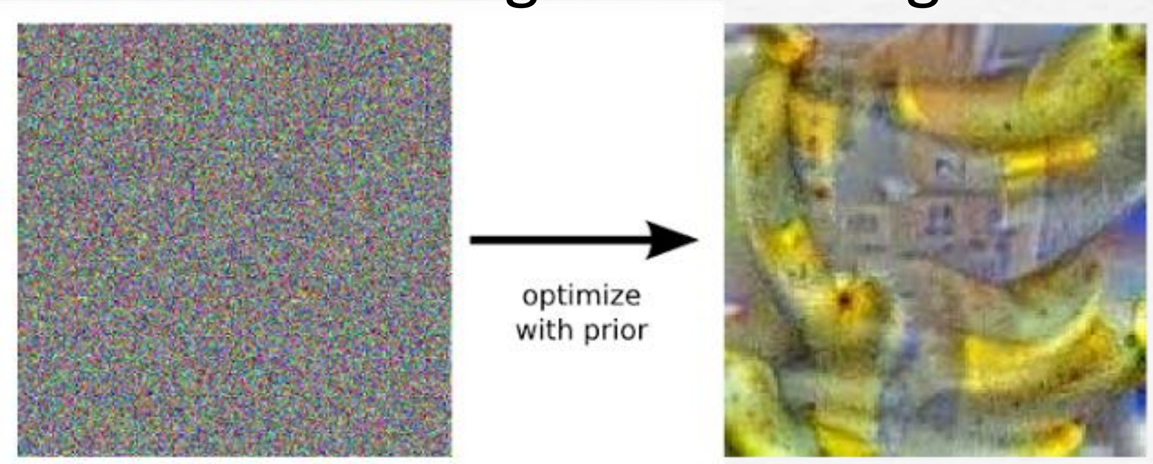
biological vision is fallible

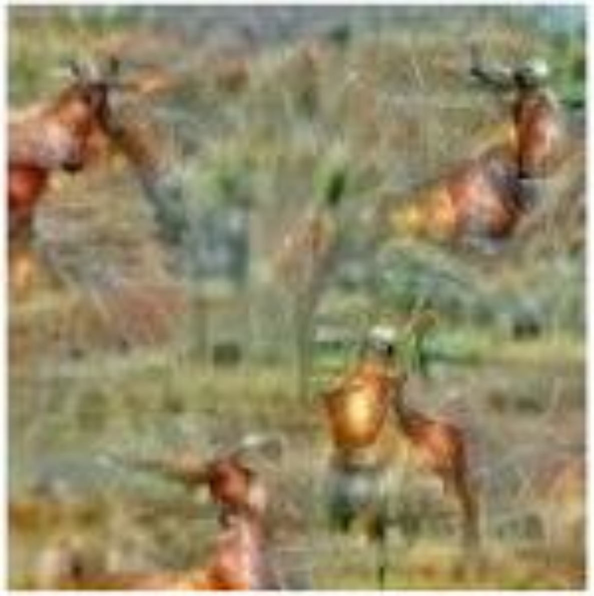
- Our perception of the world rests on assumptions and inference, not veridical measurements
- Context and task play a huge role in this process
- We choose what to treat as signal and what to treat as noise depending on context

- Consequently, we often hallucinate*
- NB. So do CNN’s

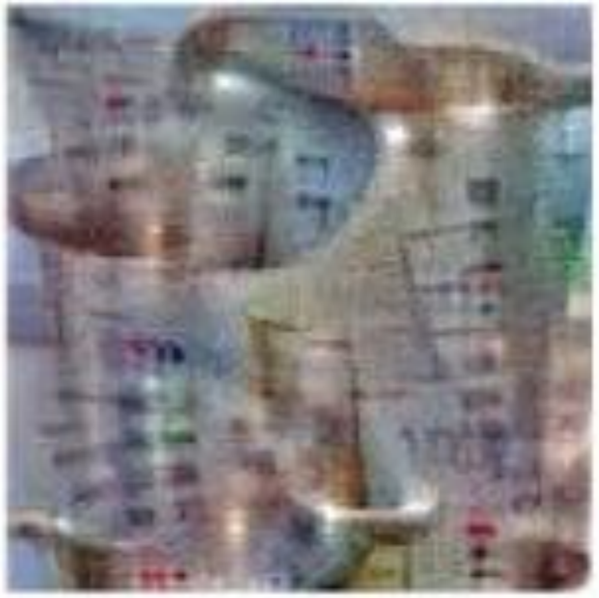
inceptionism | deep dream

- https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
- “One way to visualize what goes on is to turn the network upside down and ask it to enhance an input image in such a way as to elicit a particular interpretation.”
- Need to impose some priors (e.g., neighboring pixels should be correlated)
- “So here’s one surprise: neural networks that were trained to discriminate between different kinds of images have quite a bit of the information needed to generate images too.”

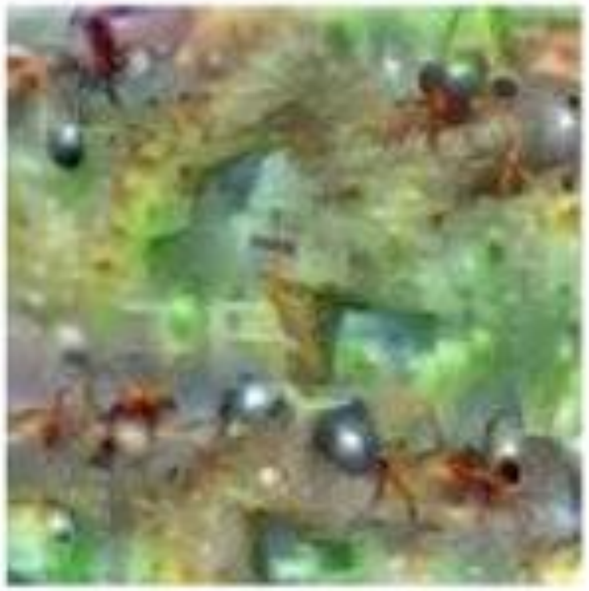




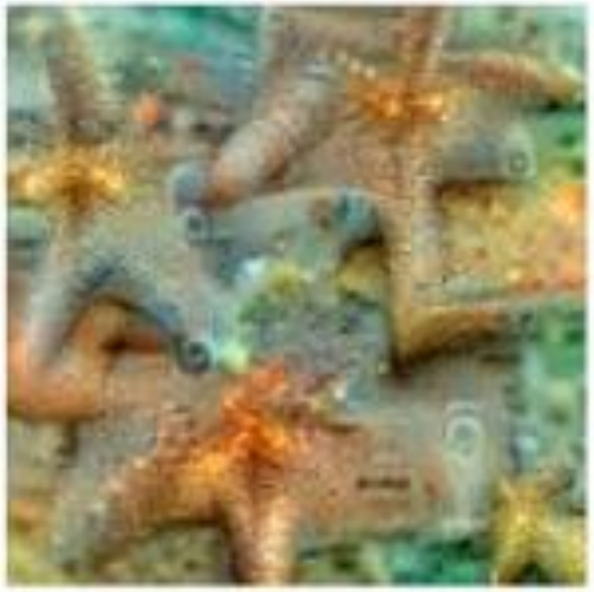
Hartebeest



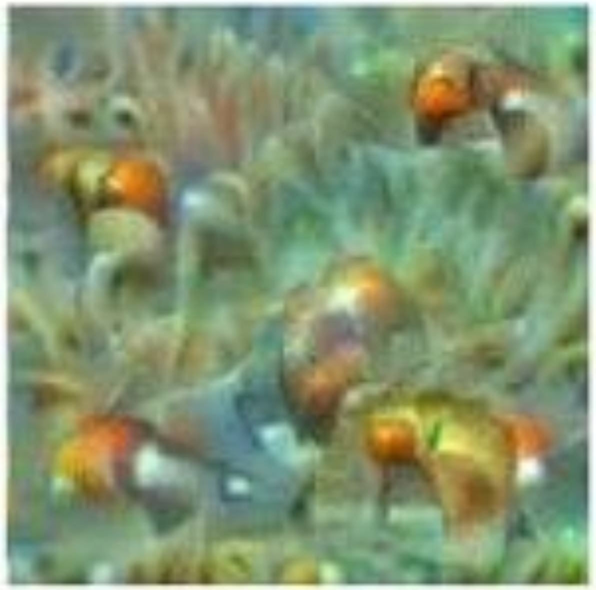
Measuring Cup



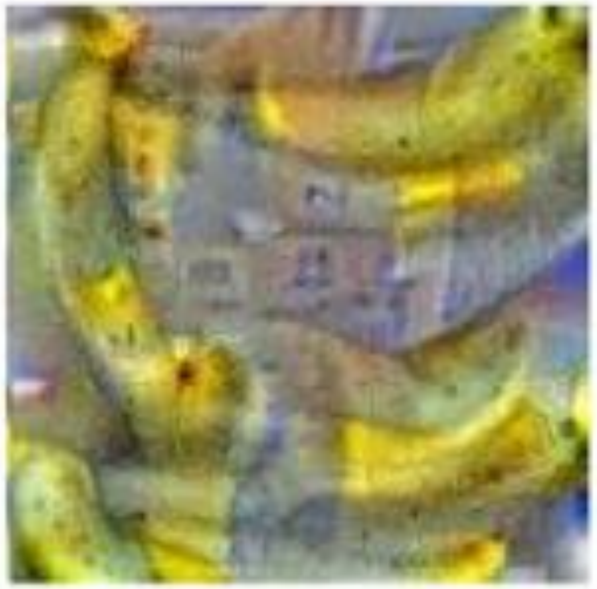
Ant



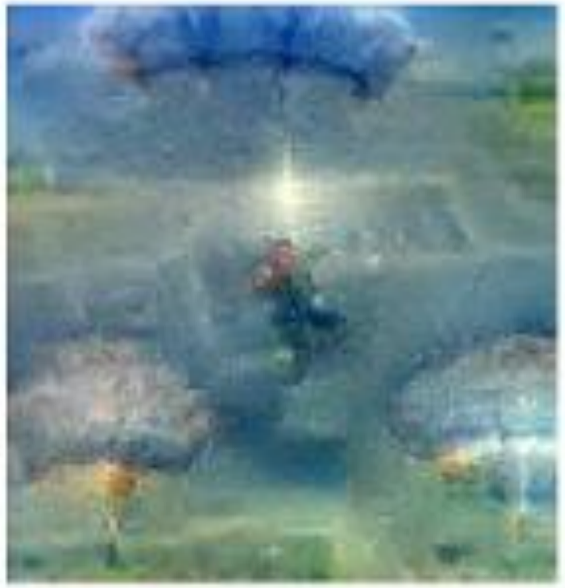
Starfish



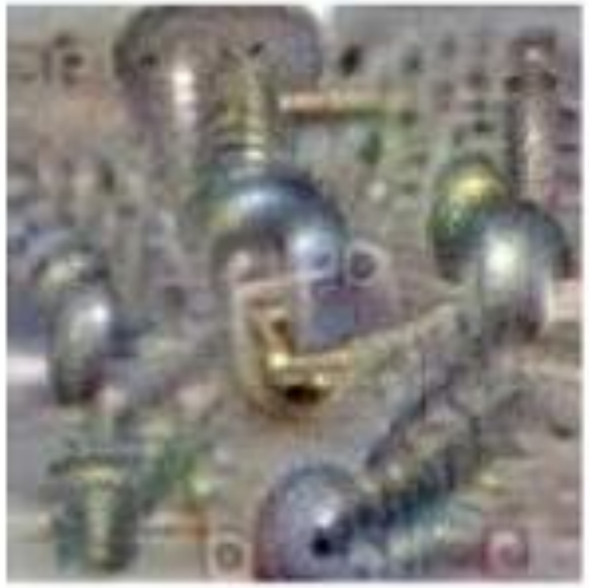
Anemone Fish



Banana



Parachute

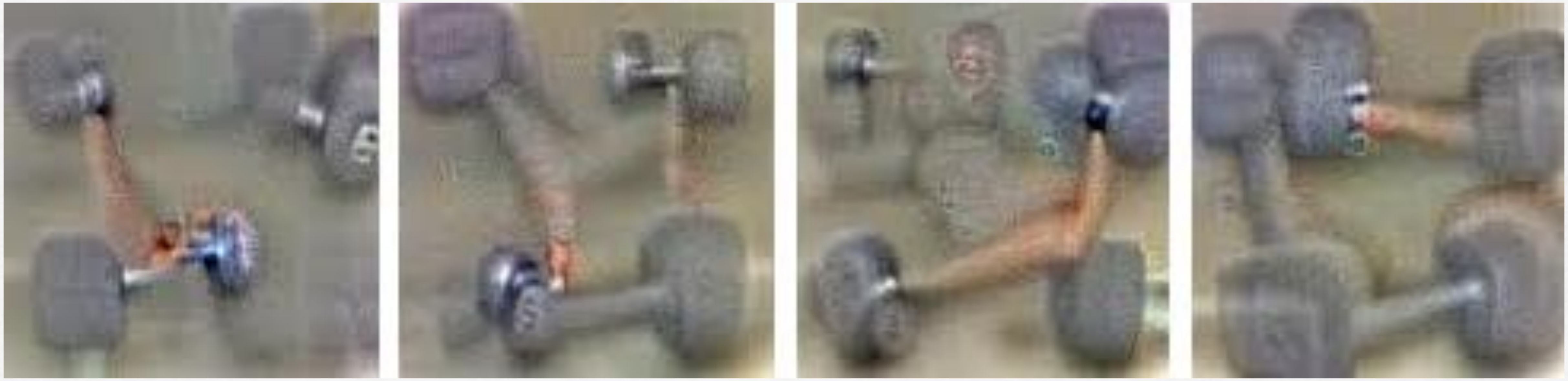


Screw

why do we need assumptions?

- the same image may arise from many different 3D structures/layouts
- vision usually goes with the most plausible, e.g., statistically likely, one
- so assumptions are just built-in high-probability interpretations
- sometimes these are wrong

dumbbells



how can we live this way?

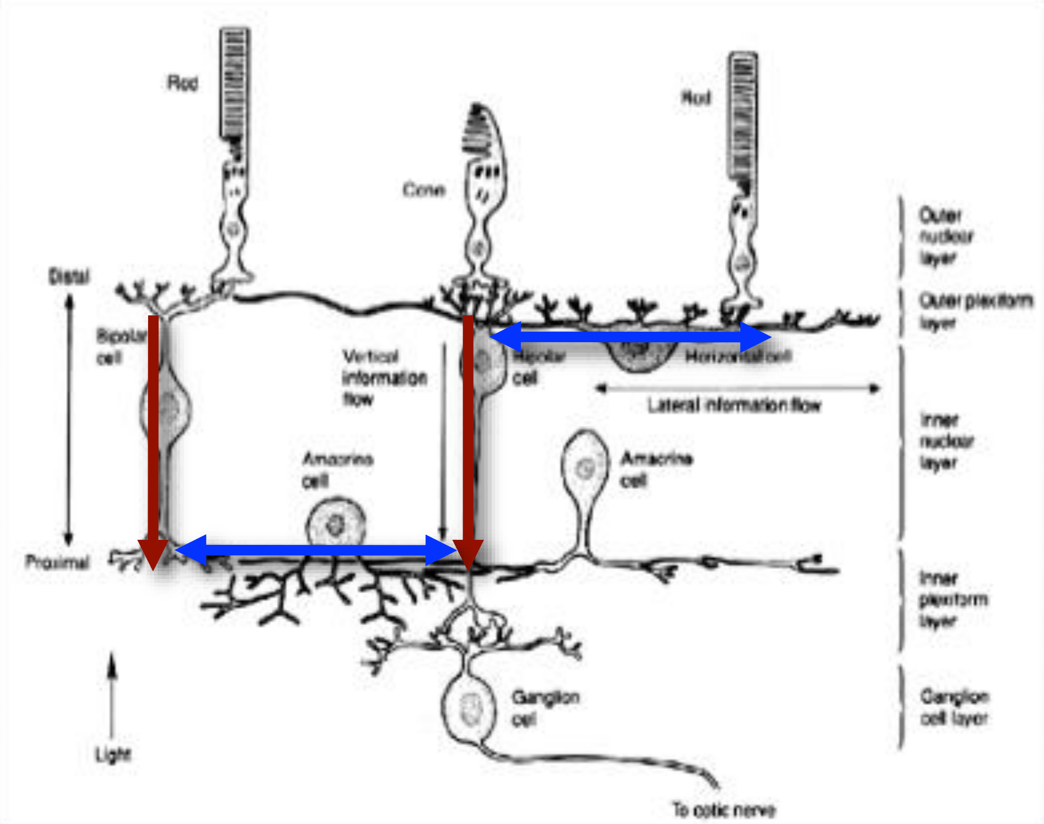
- some decision is better than no decision
- that is, from a survival point of view, make your best guess - if you don’t get eaten or fall off the cliff, it was probably the correct decision
- luckily our ancestors have had lots of time to learn the statistics of the world
- so perhaps the ”goal” for CNNs shouldn’t be “best” performance, but rather optimal given certain survival constraints (amount of training data, time for decision, etc.)

biological vision

- is not a means for describing the world
- is a means for taking in data and then using that data to guide behavior
- we know the structure of the input and we can measure the output - behavior - or these days brain activity

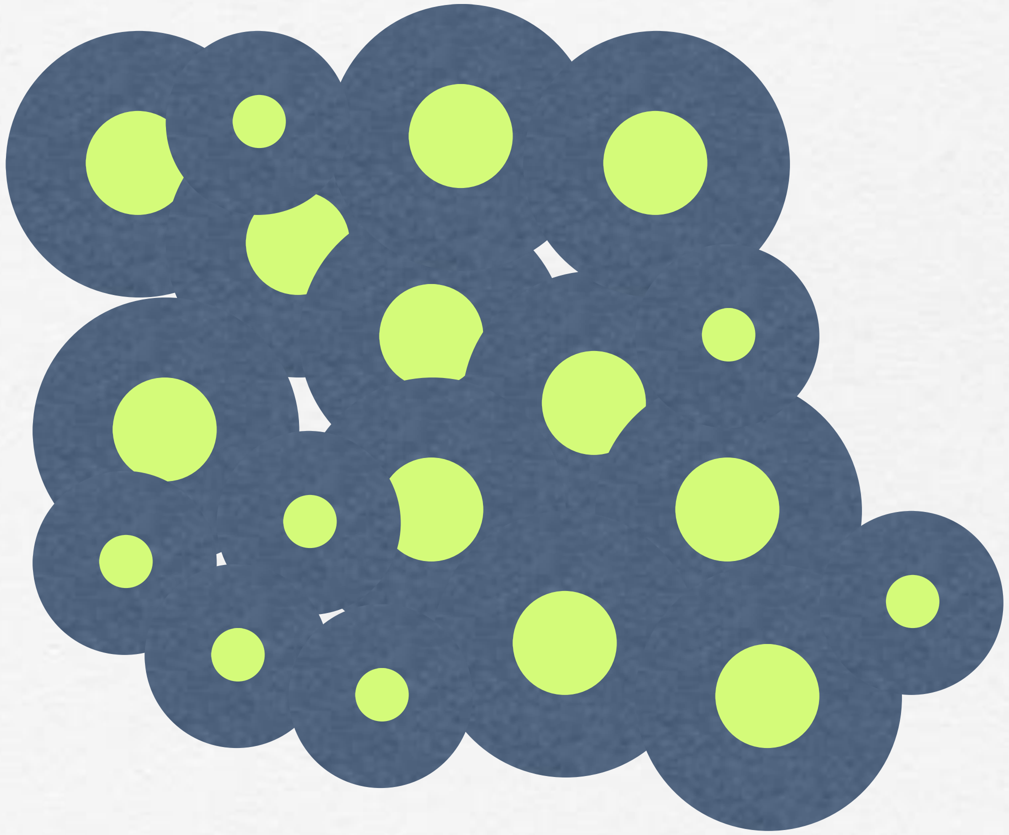
early vision

- begins at the retina
- dramatic data reduction
- center-surround organization appears

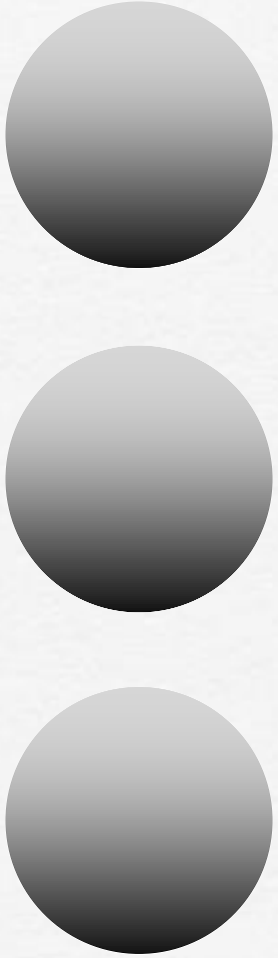


receptive fields

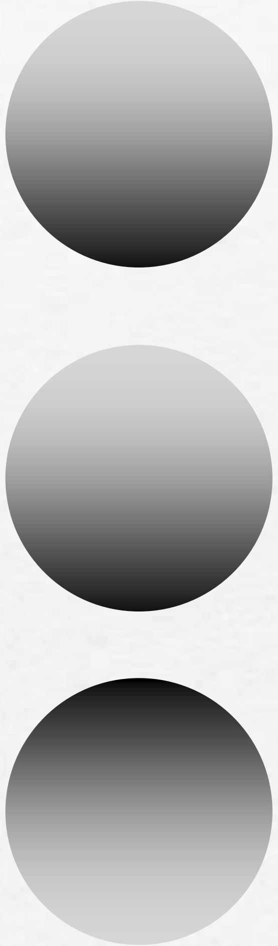
- a receptive field for a given neuron is the area of the retina where the pattern of light affects that cell's firing pattern
- an area of the retina corresponds to a <u>location in space</u>
- great degree of overlap from rf to rf



mid-level vision

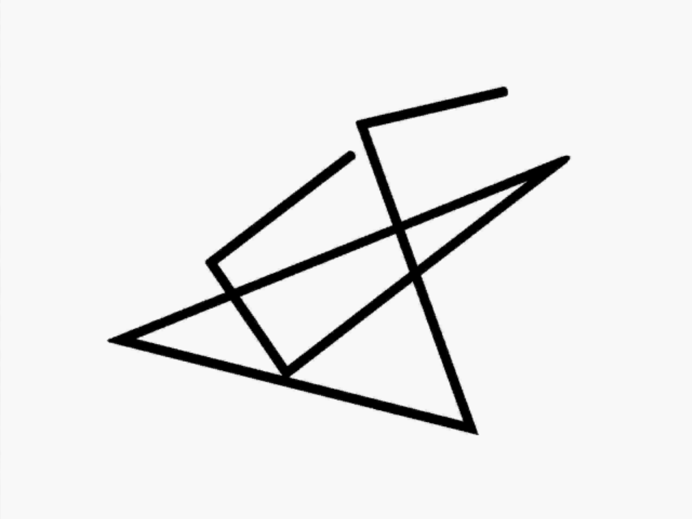


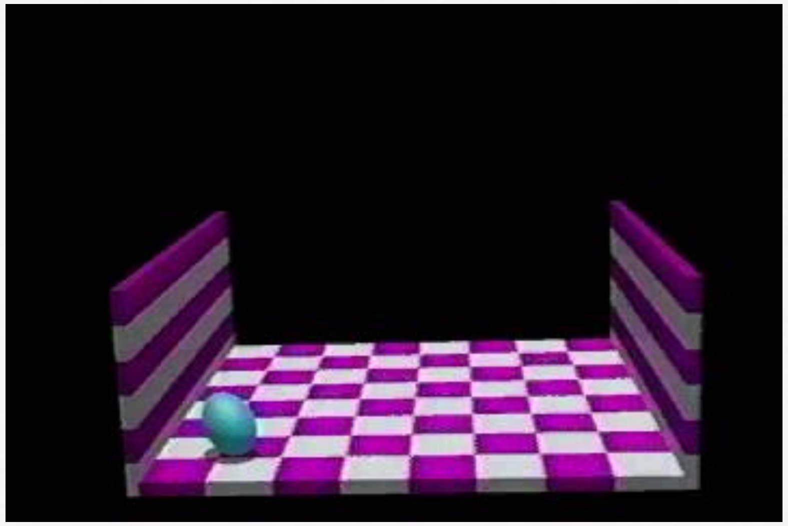
- “cues” to different properties of the scene – lighting, color, depth, texture, shape, etc.
- how do different cues function independently?
- what assumptions are made in interpreting cues?
- how are cues combined to form percepts?
- how do we “explain” different image artifacts?
- constancies

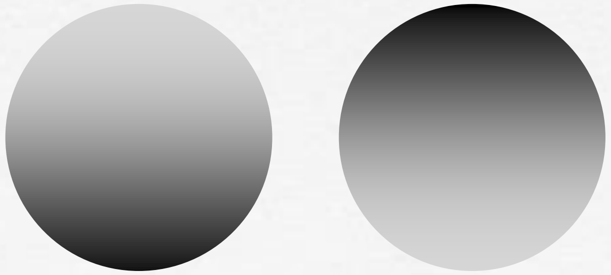


cues to depth/shape

- stereo
- motion
- shading
- shadows
- etc…



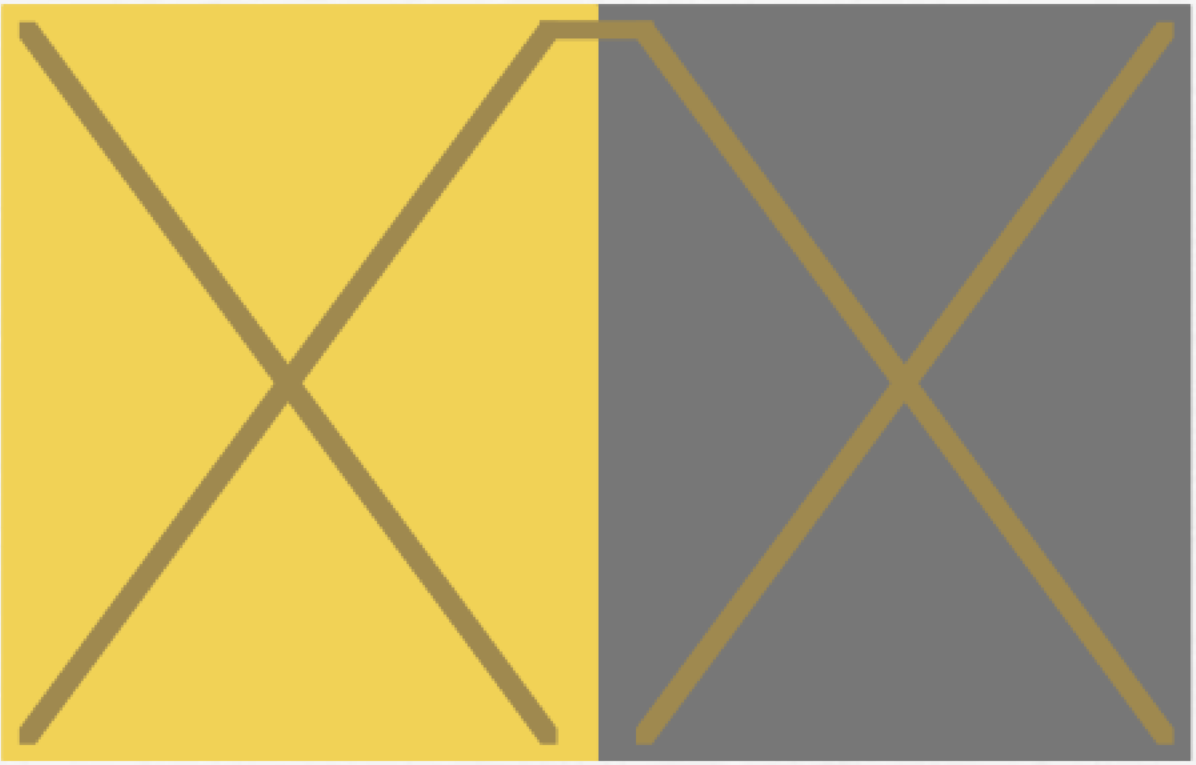




constancies

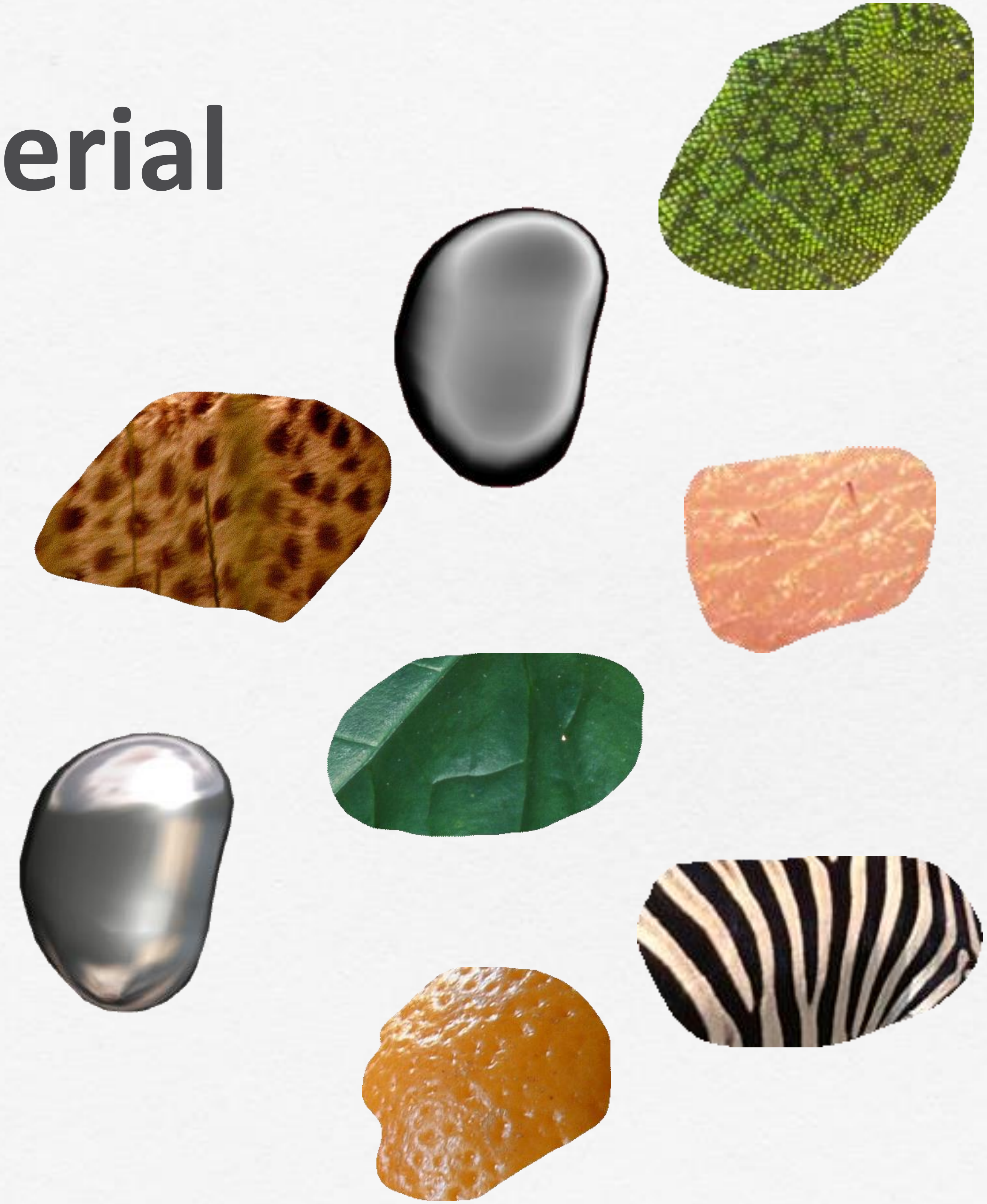
- color* and lightness are not veridical properties of surfaces
- rather, they are perceptual entities that are inferred by taking context into account
- perhaps assumptions about the environment as well

*really interesting



cues to material

- shading
- specularities
- texture
- color
- etc…



high-level vision

- how are objects represented/recognized?
- how are categories formed?
- how do we manipulate visual information?
- how do learn new visual information?

- similar goals to deep networks…
- “**Using goal-driven deep learning models to understand sensory cortex” by Yamins & DiCarlo (2016) ~ similar representations**

Tanaka (2003) used an image reduction method to isolate “critical features” (physiology)

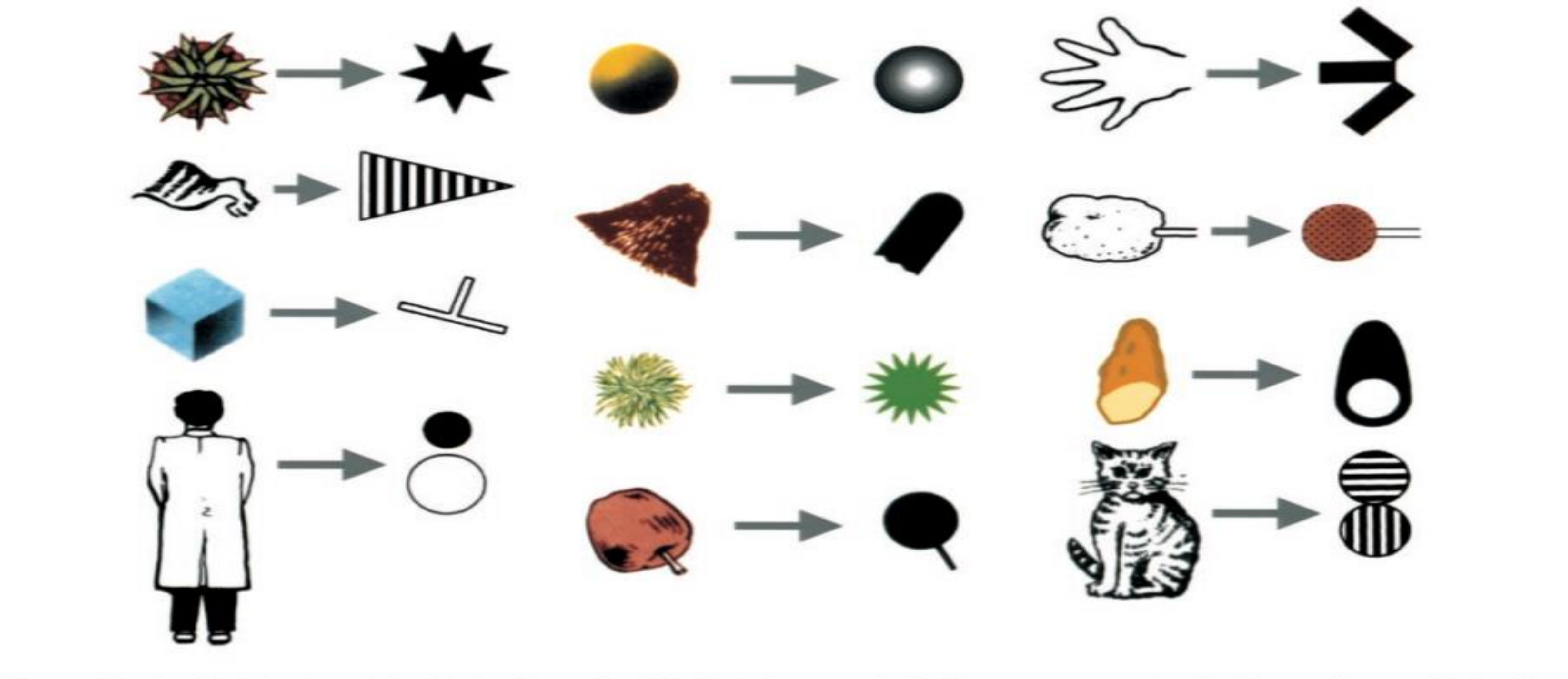
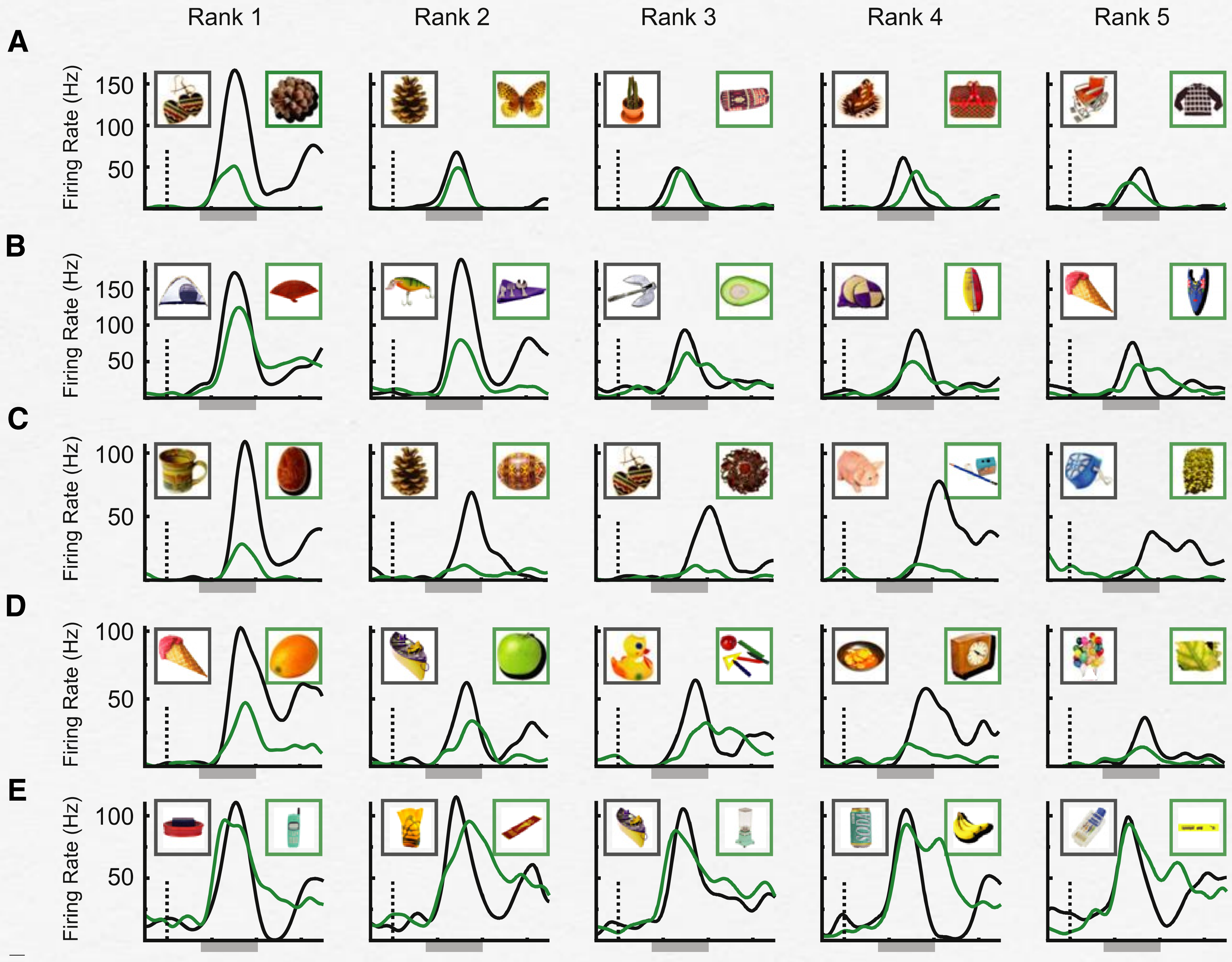
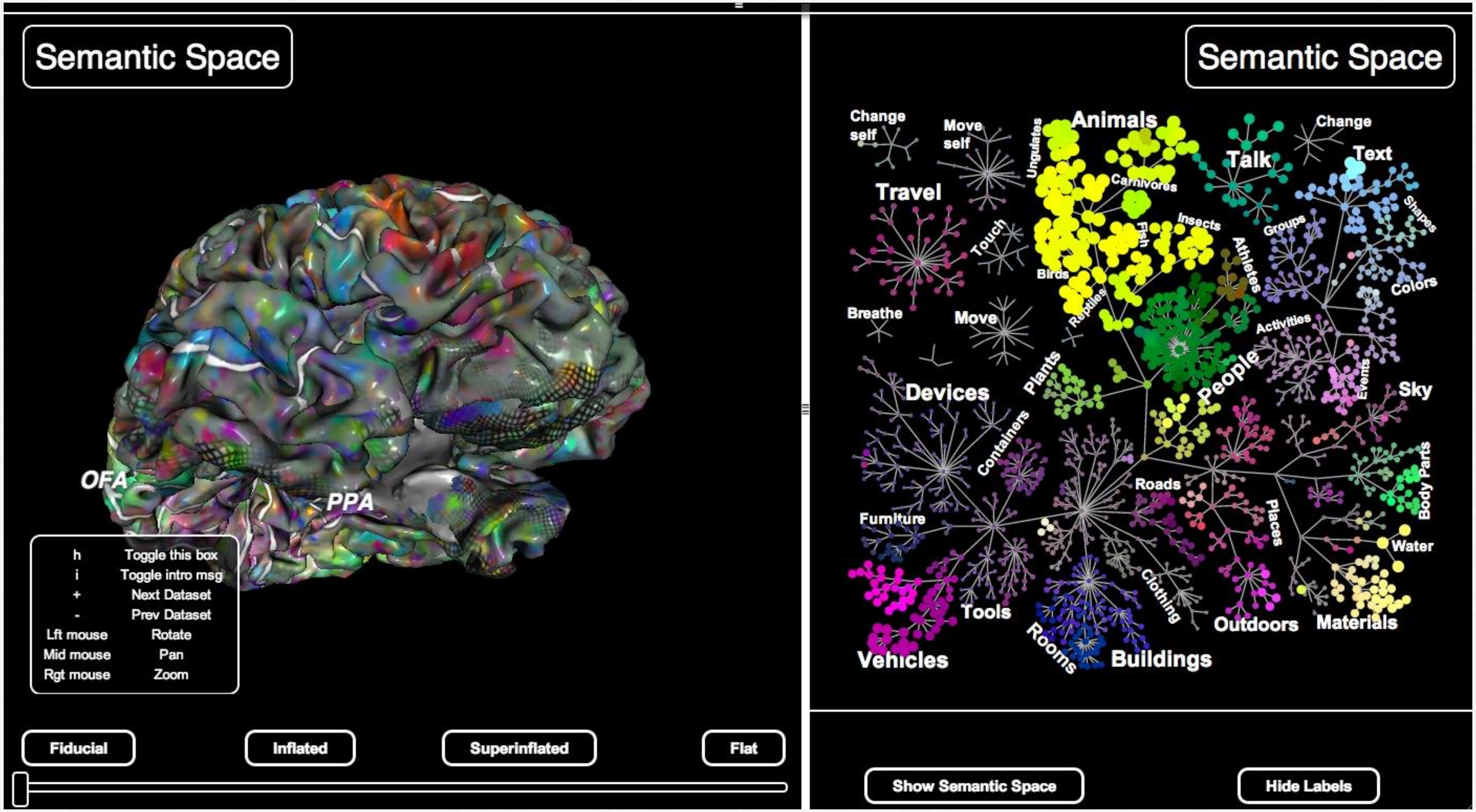


Figure 1. Examples of reductive determination of optimal features for 12 TE cells. The images to the left of the arrows represent the original images of the most effective object stimulus and those to the right of the arrows, the critical features determined by the reduction.

Woloszyn and Sheinberg (2012)

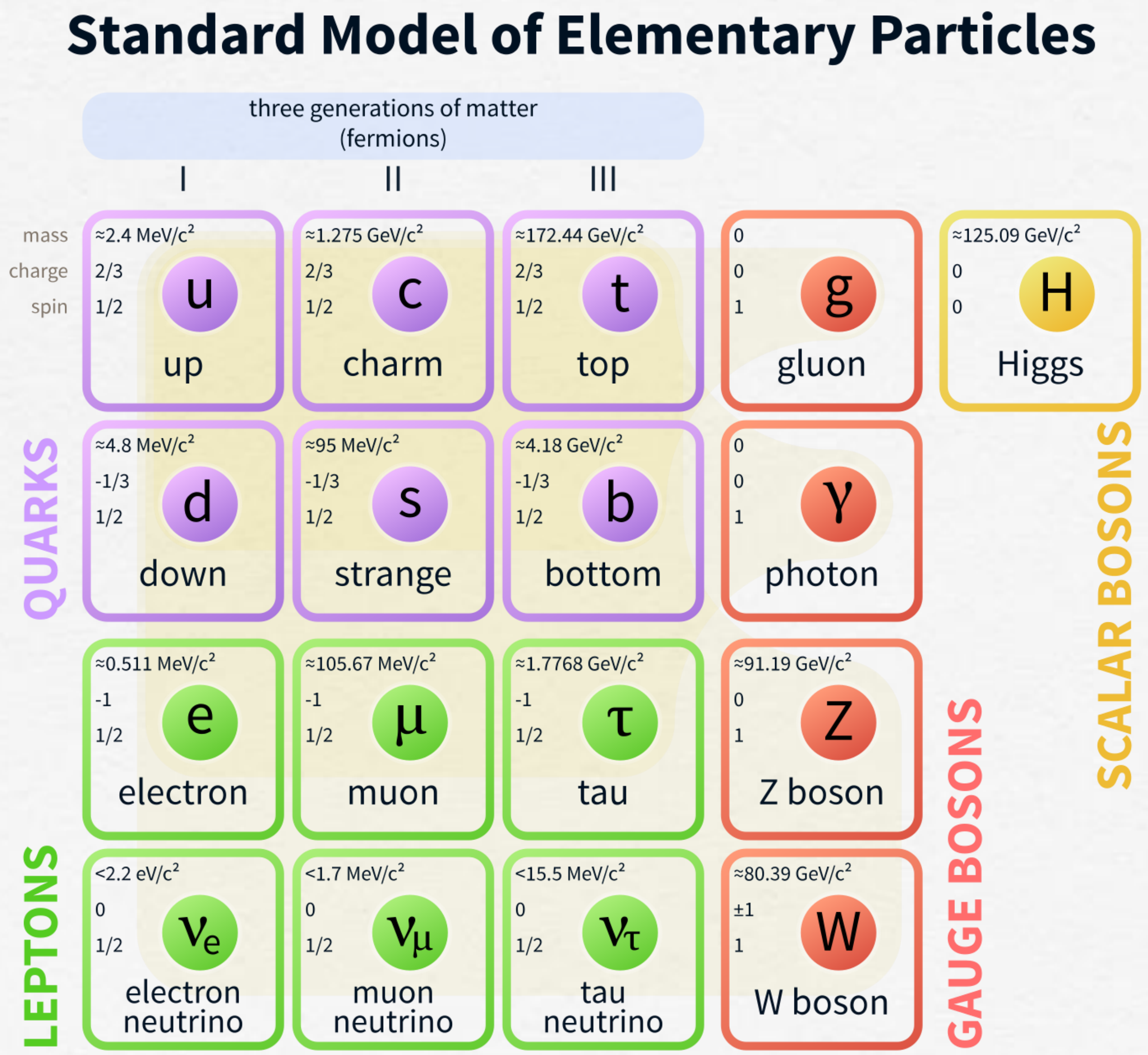


Gallant (2012) constructed a “semantic” map across visual cortex (fMRI)



h	Toggle this box
i	Toggle intro msg
+	Next Dataset
-	Prev Dataset
Lft mouse	Rotate
Mid mouse	Pan
Rgt mouse	Zoom

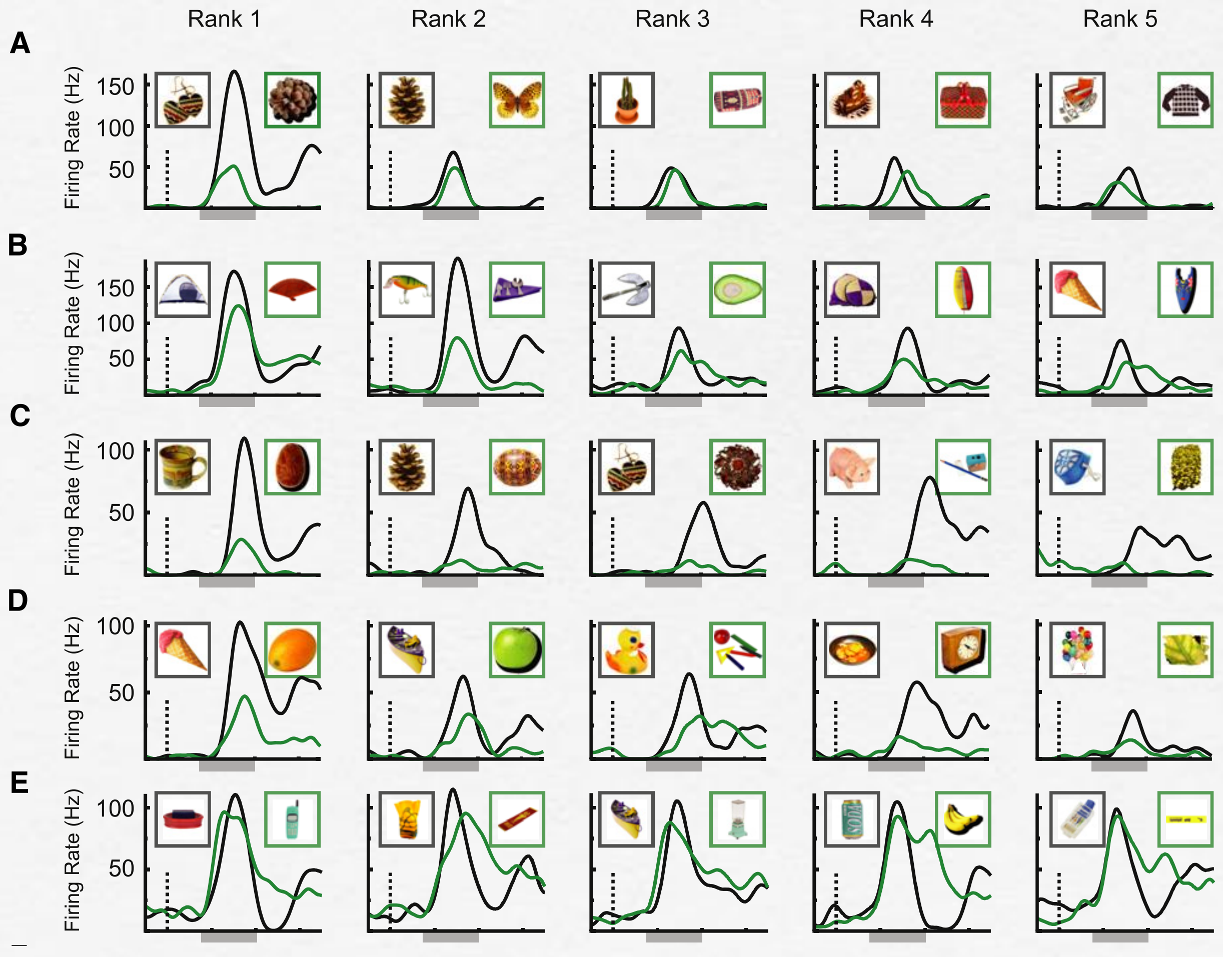
is there a “vocabulary” of high-level features?



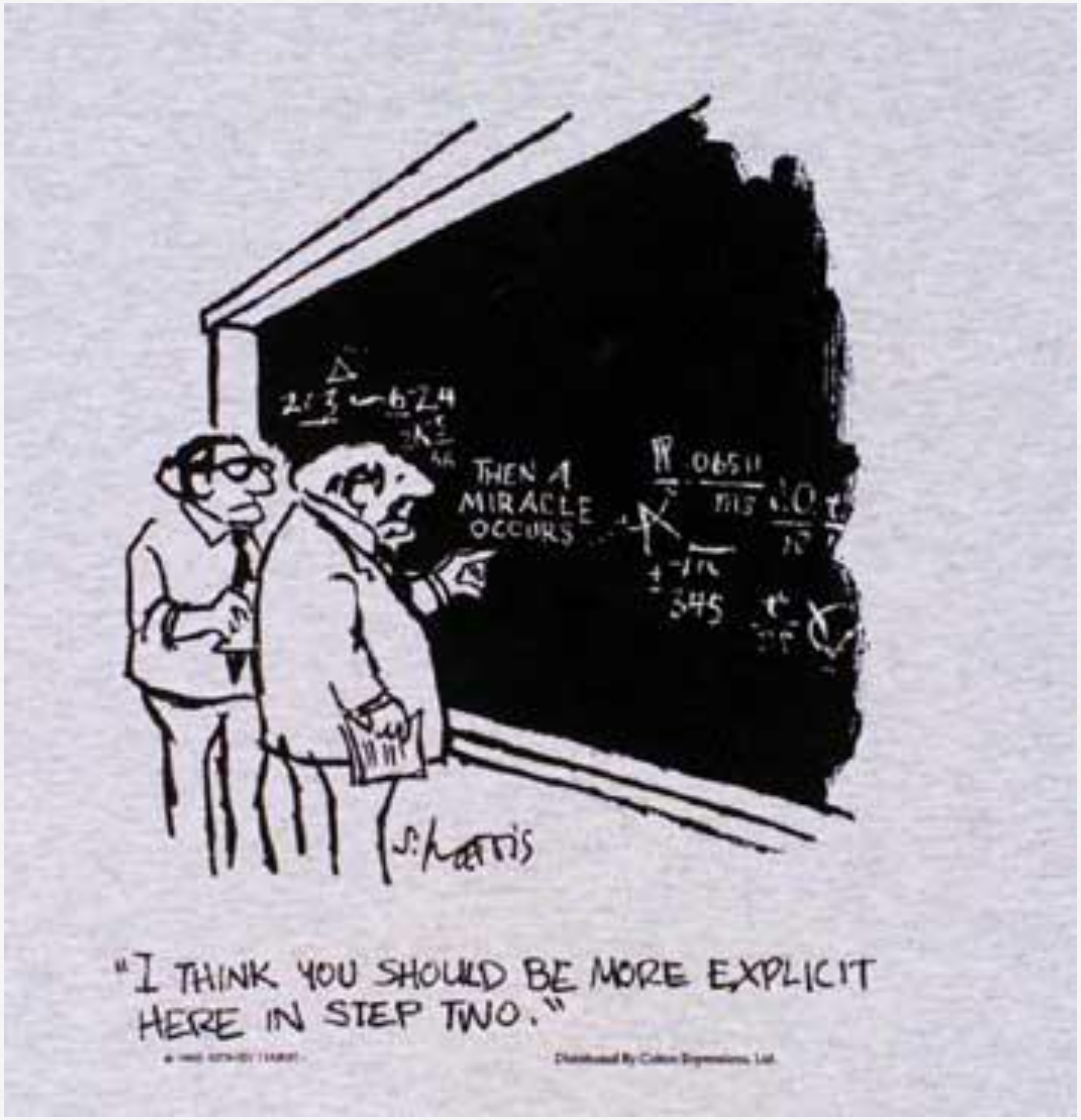
Standard Model of Elementary Particles

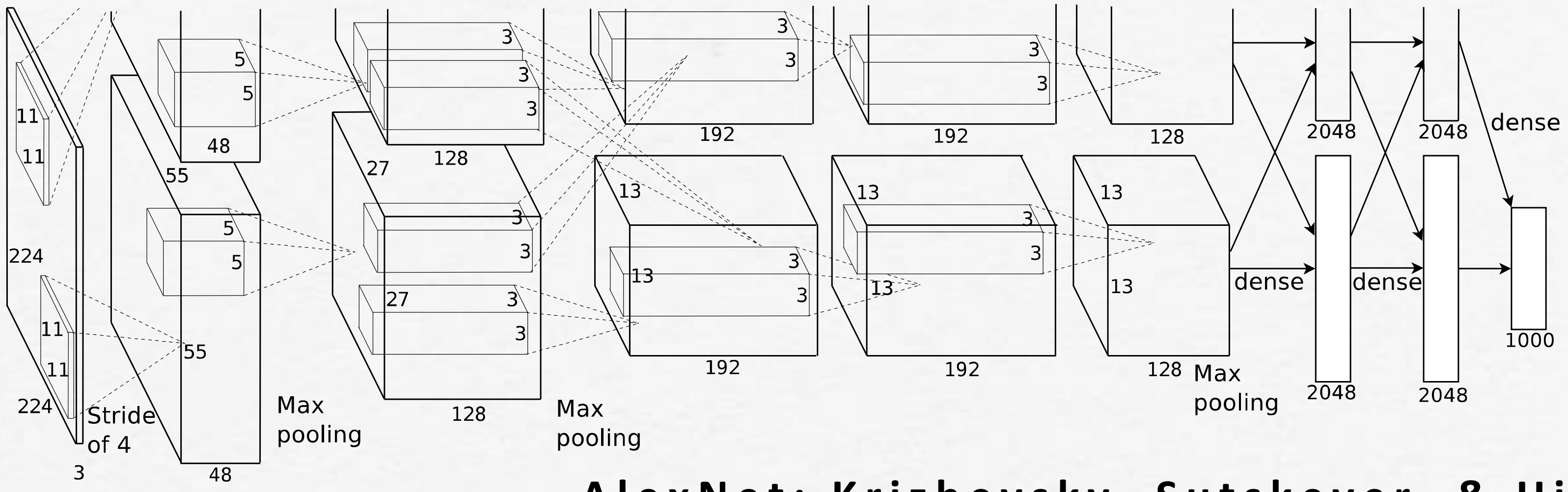
three generations of matter (fermions)

	I	II	III	GAUGE BOSONS	SCALAR BOSONS
mass charge spin					
QUARKS	≈2.4 MeV/c^2 2/3 1/2 u up	≈1.275 GeV/c^2 2/3 1/2 c charm	≈172.44 GeV/c^2 2/3 1/2 t top	0 0 1 g gluon	≈125.09 GeV/c^2 0 0 H Higgs
QUARKS	≈4.8 MeV/c^2 -1/3 1/2 d down	≈95 MeV/c^2 -1/3 1/2 s strange	≈4.18 GeV/c^2 -1/3 1/2 b bottom	0 0 1 γ photon	
LEPTONS	≈0.511 MeV/c^2 -1 1/2 e electron	≈105.67 MeV/c^2 -1 1/2 μ muon	≈1.7768 GeV/c^2 -1 1/2 τ tau	≈91.19 GeV/c^2 0 1 Z Z boson	
LEPTONS	<2.2 eV/c^2 0 1/2 ν_e electron neutrino	<1.7 MeV/c^2 0 1/2 ν_μ muon neutrino	<15.5 MeV/c^2 0 1/2 ν_τ tau neutrino	≈80.39 GeV/c^2 ±1 1 W W boson	

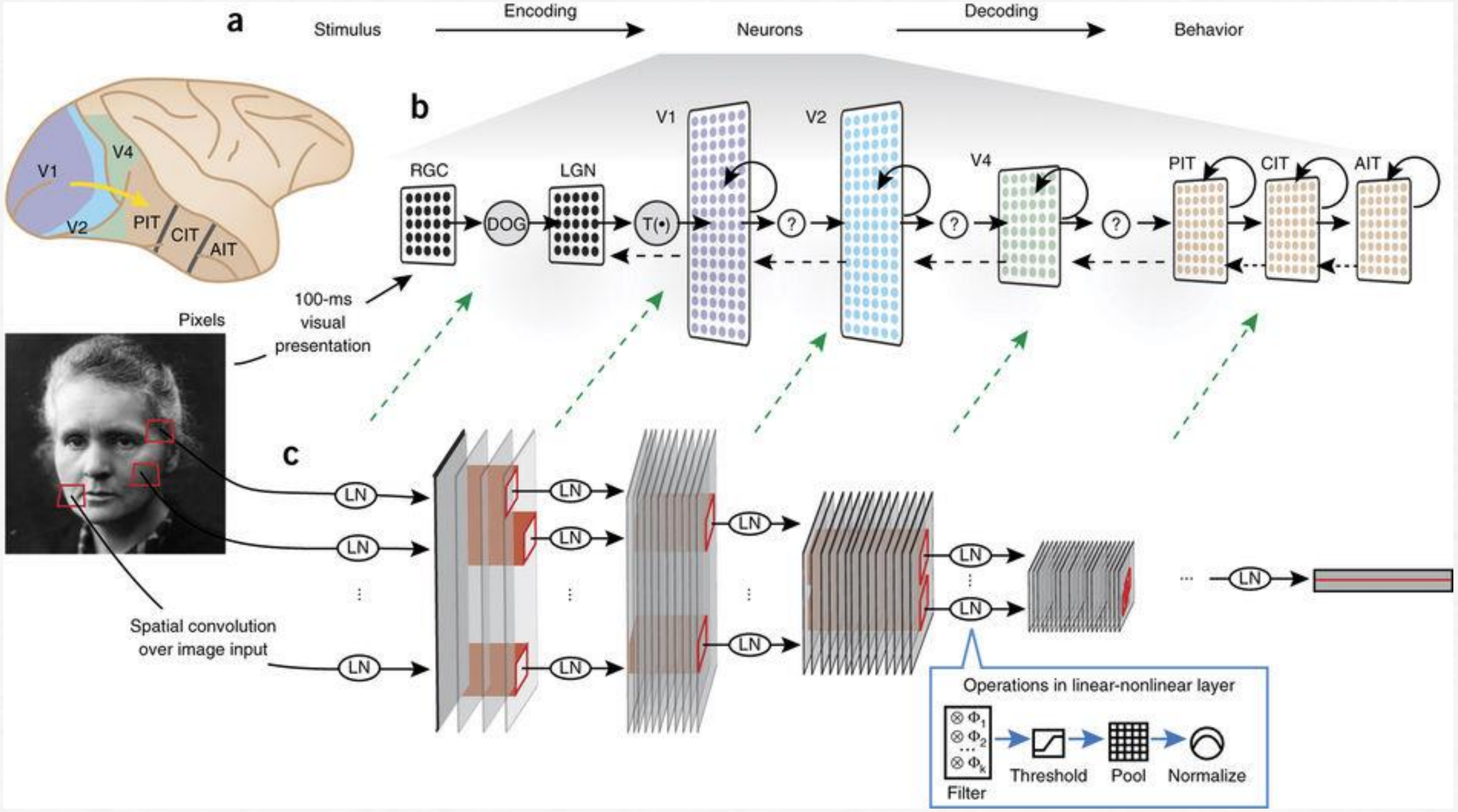


CNNs


THEN A MIRACLE OCCURS . . .
S. Harris
"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

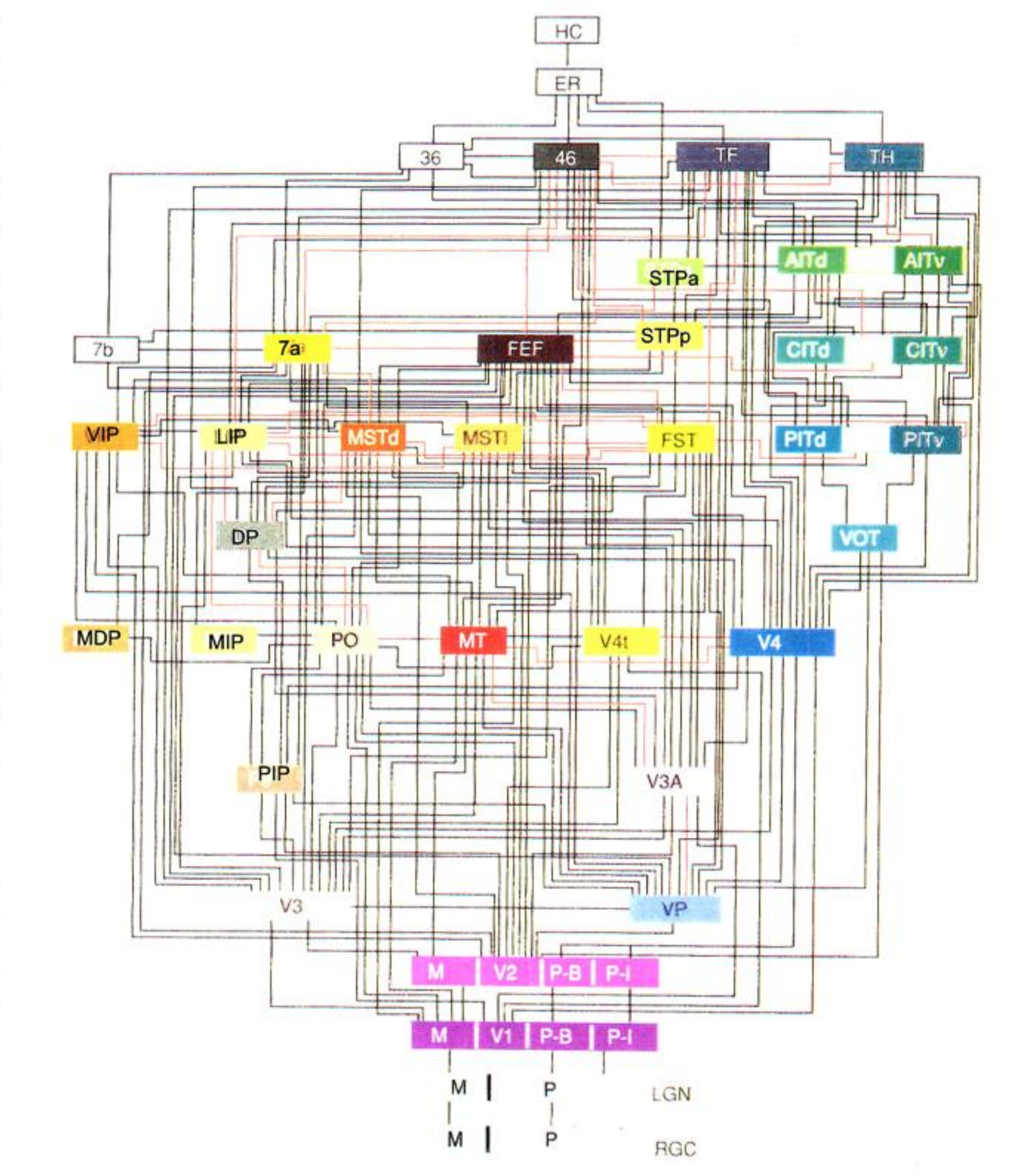


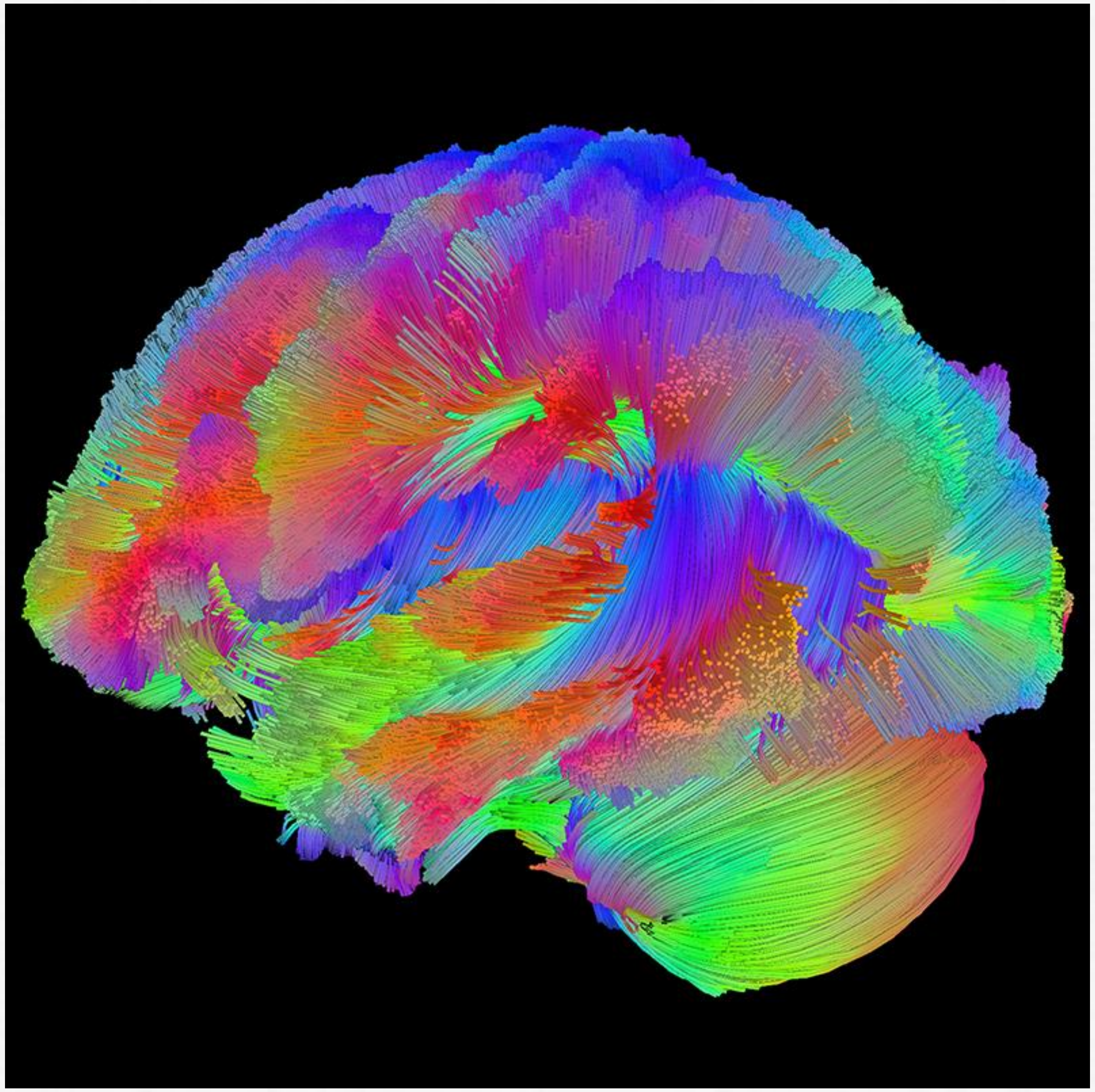
AlexNet: Krizhevsky, Sutskever, & Hinton, *NIP (*2012)

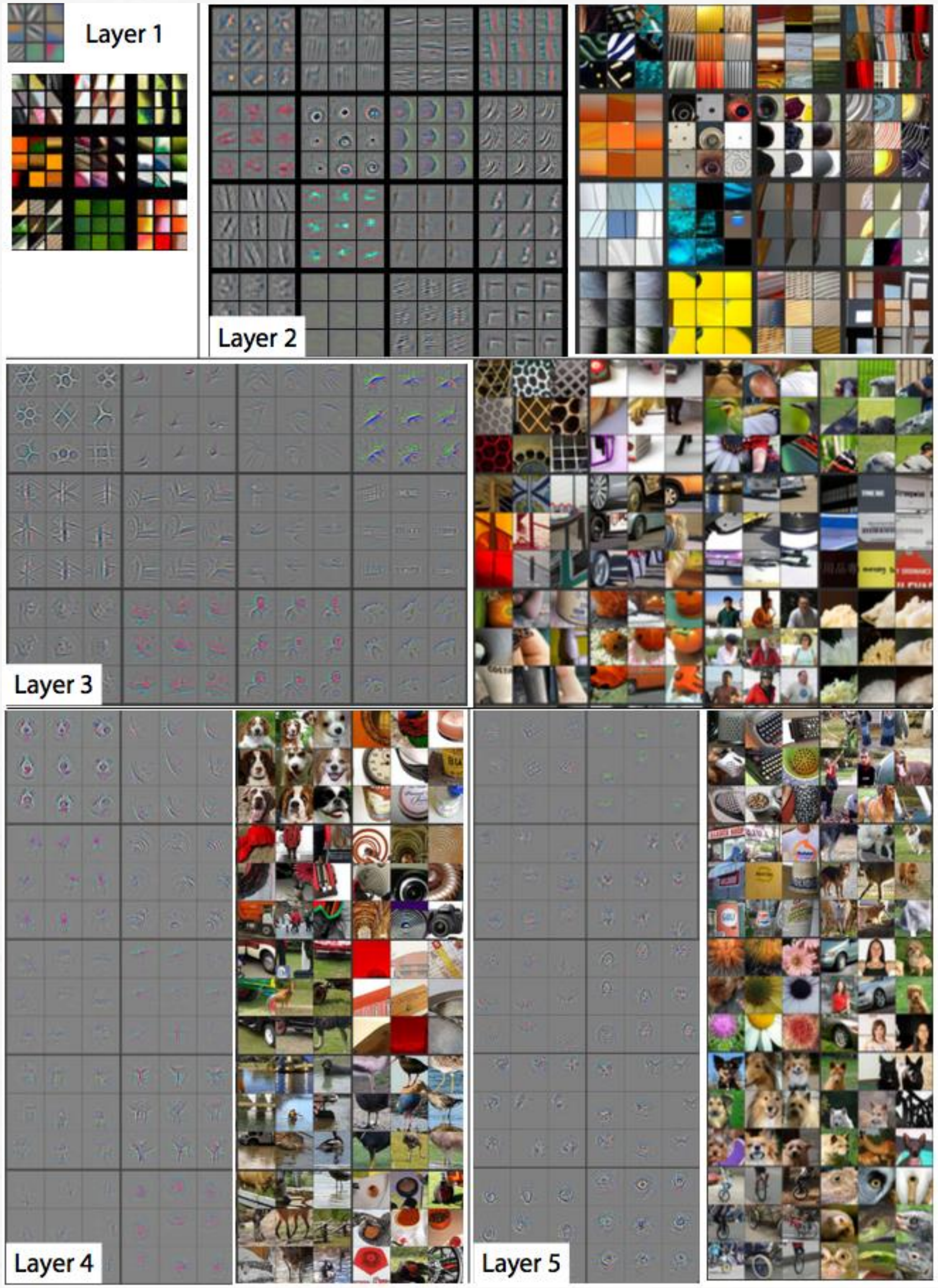


Yamins & DiCarlo (2016)

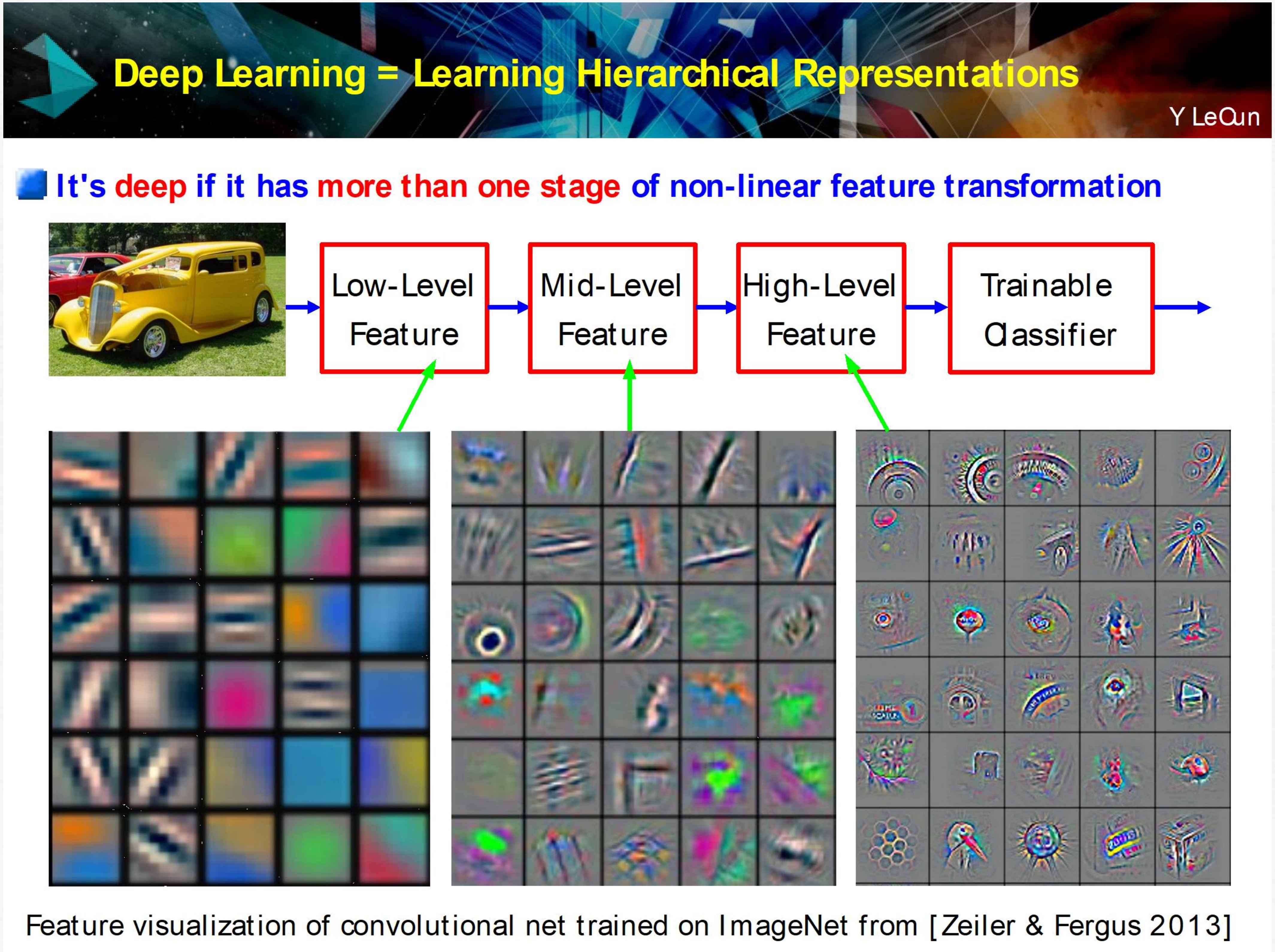
Primate visual cortex







Zeiler & Fergus (2103)



Deep Learning = Learning Hierarchical Representations

Y LeCun

- It's deep if it has more than one stage of non-linear feature transformation

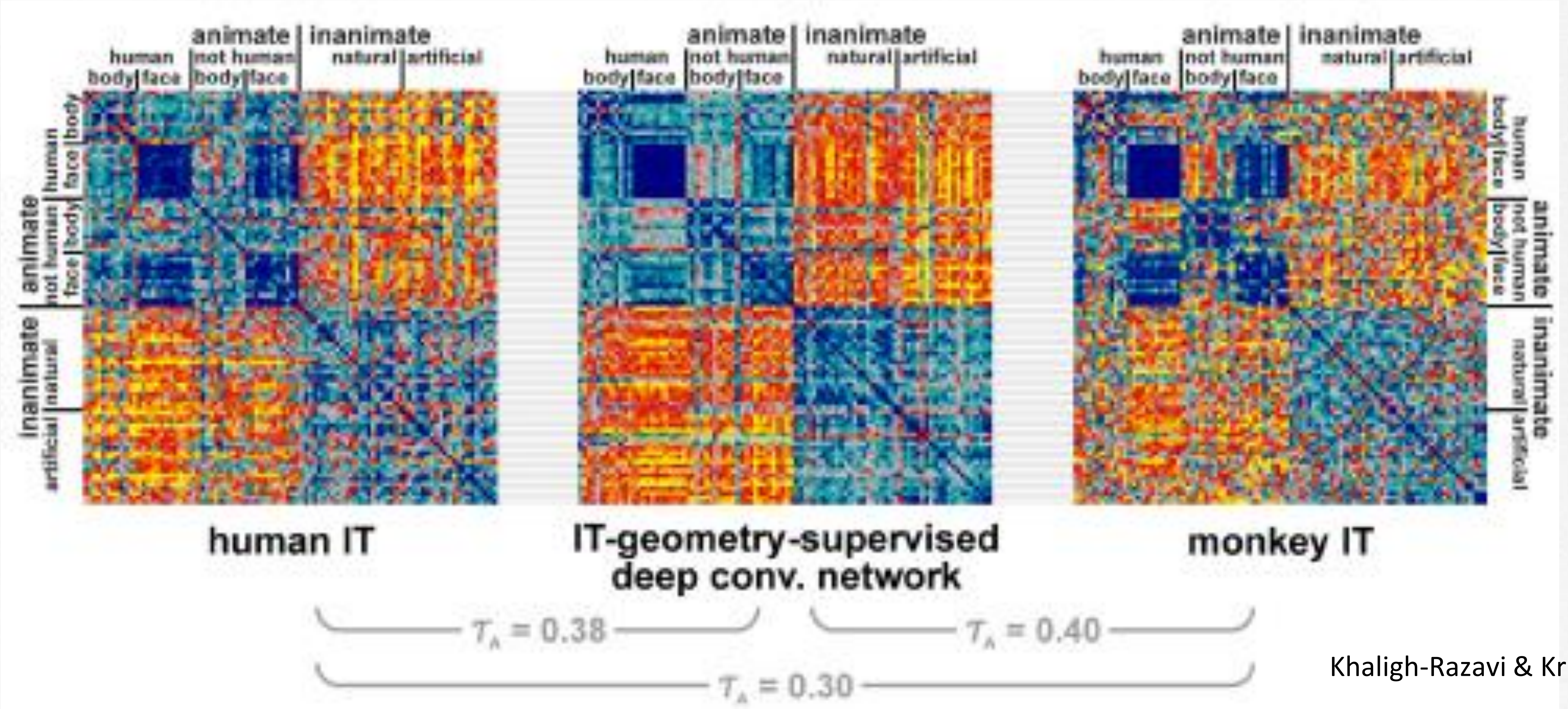
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Comparing models to data – RDMs

For a given set of images, compute pairwise similarities within each model

Compute neurally-derived similarities for the same images within each brain region

Correlate the similarity matrices

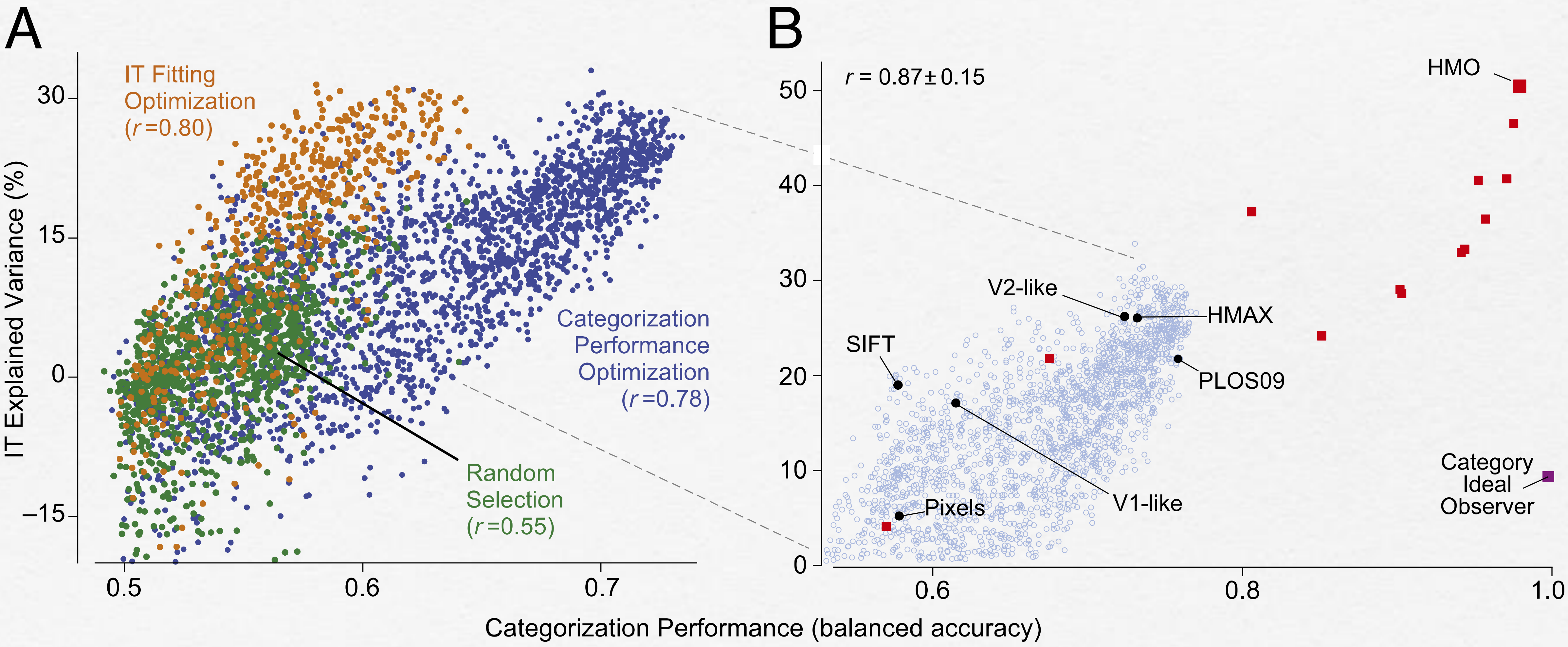


Khaligh-Razavi & Kriegeskorte (2014)

applying proxy models

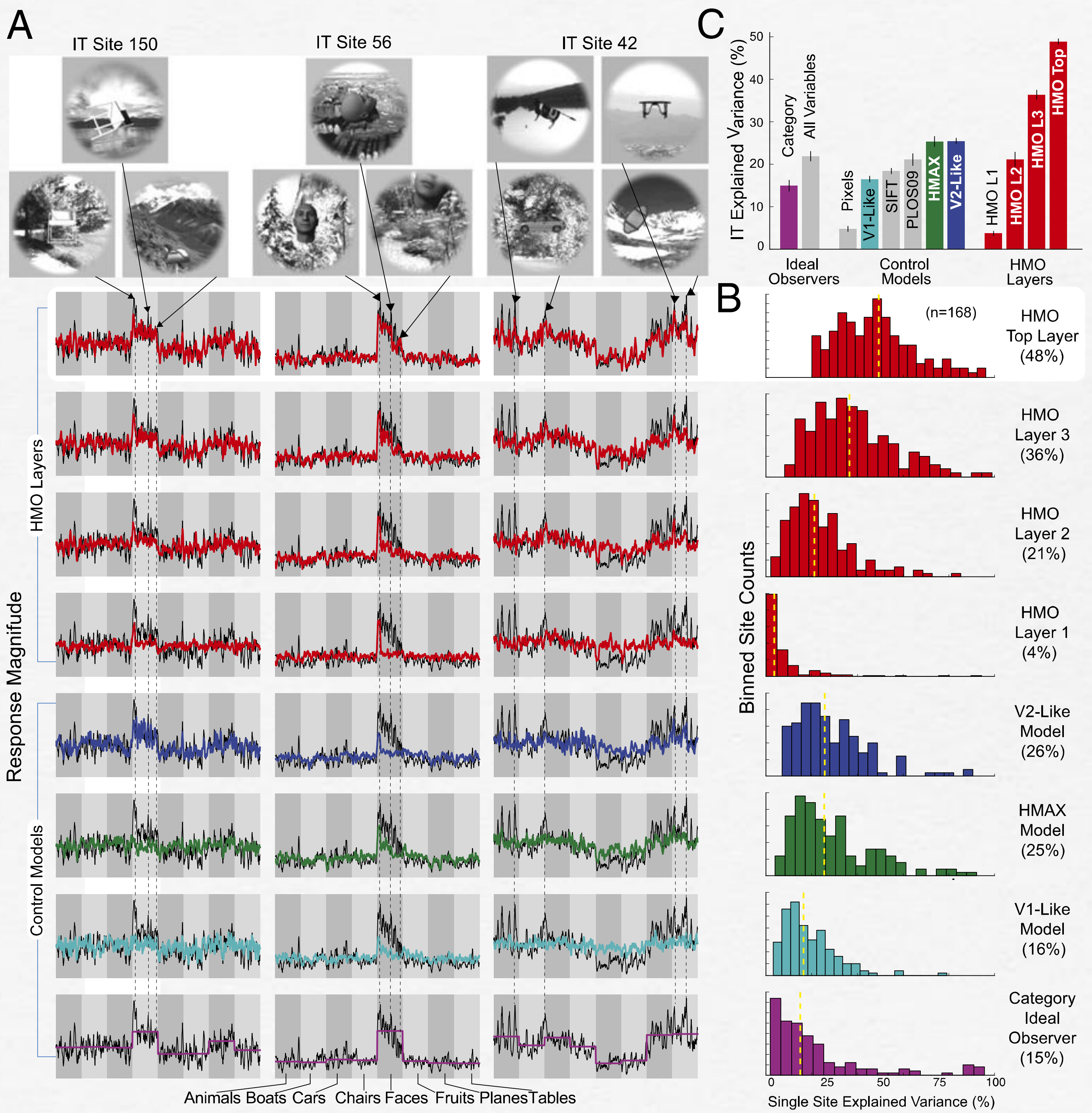
- Models and visual system use same input – images – so early layers will tend to show high similarity to early visual areas
- Models and visual system have similar output goals – object categorization / semantics – so last few layers will tend to show high similarity to IT cortex
- Challenges?
  - overall system performance
    - Categorization
    - Invariant recognition
  - mid-level representation
  - fine-grained similarity not driven by “low-hanging fruit”

optimizing models for similar goals



Yamins et al. (2014)

IT neural predications



Yamins et al. (2014)

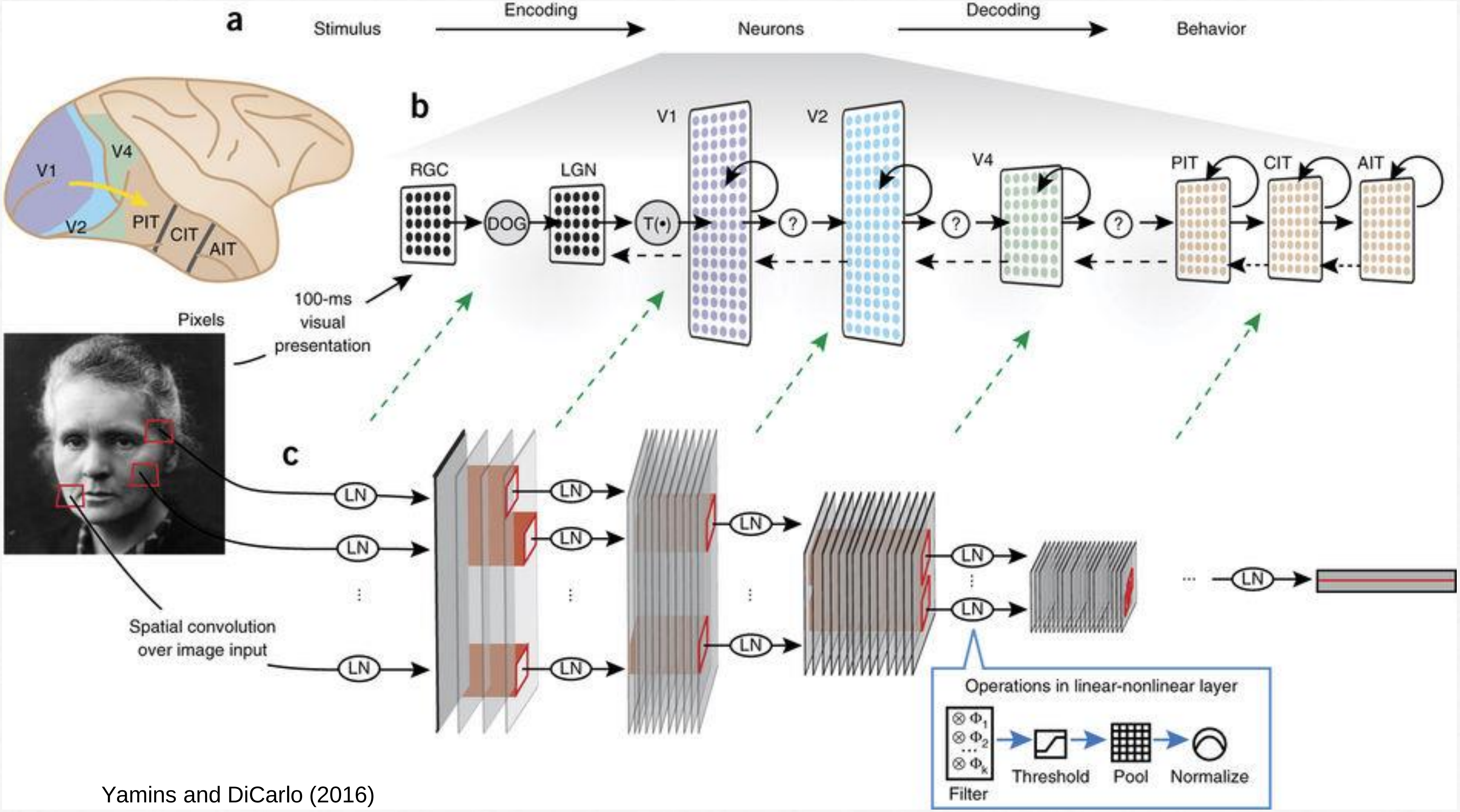
some observations

- Early layers tend to be highly similar irrespective of task
- Higher-level layers are much more task-sensitive, but still correlate
- An off-the-shelf model trained with a relatively small number of examples will typically perform quite well
- How many truly unique tasks are there?
- Fine-grained performance differences will be critical in evaluating CNNs as models of biological vision

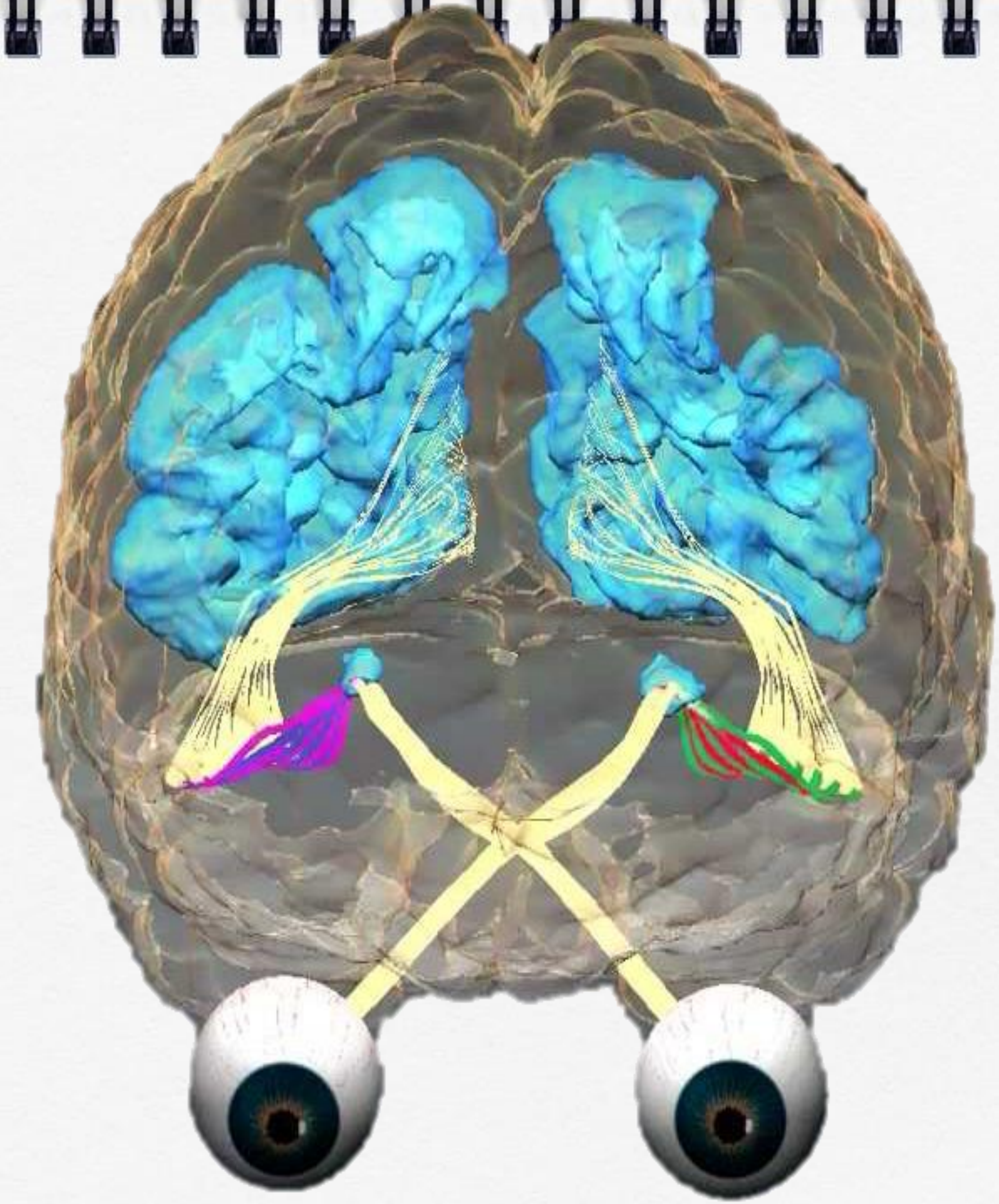
sandboxes

- Explore how high-level functional organization of visual cortex arises
- Push the idea that this complex organization based on category-selectivity can emerge from relatively simple assumptions and <u>minimal</u> starting conditions
- Only add constraints/structures when simpler models fail
- We have some idea of reasonable priors from human and primate neuroimaging/neurophysiology
- Use high-performing visual recognition models inspired by the basic hierarchical architecture of the primate visual system: CNN's as "sandboxes"

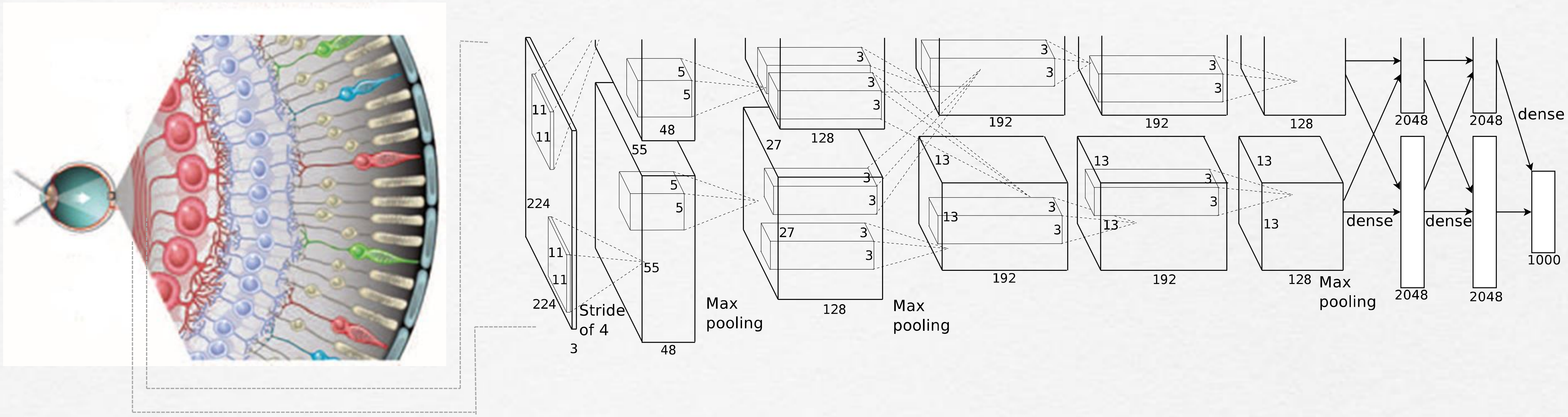
What is missing from this comparison?



Yamins and DiCarlo (2016)



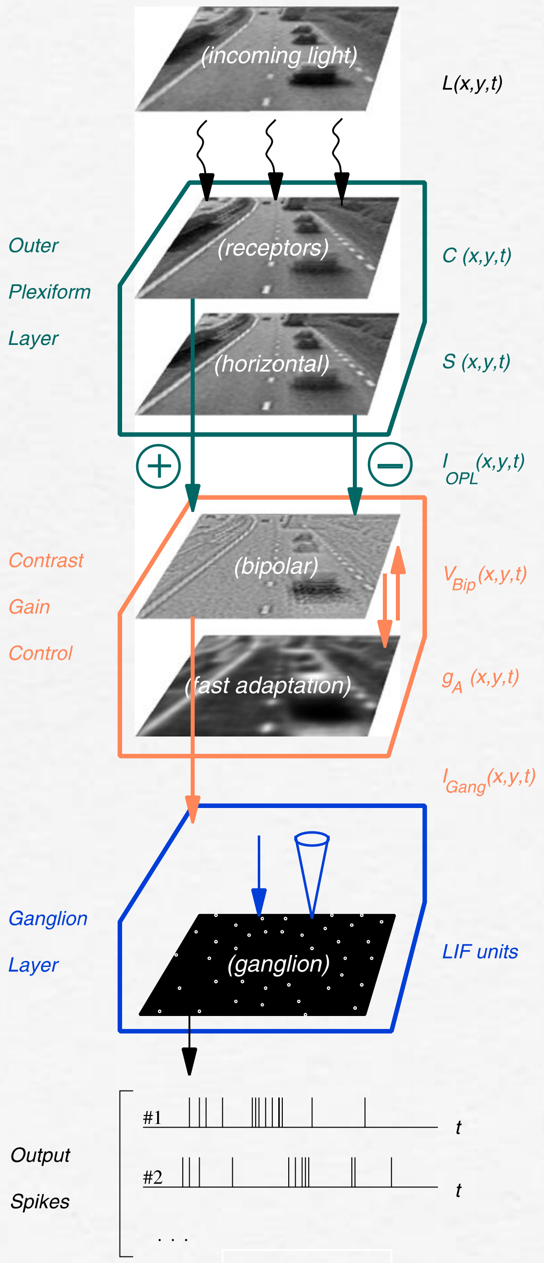
Impact of adding a retina to a CNN

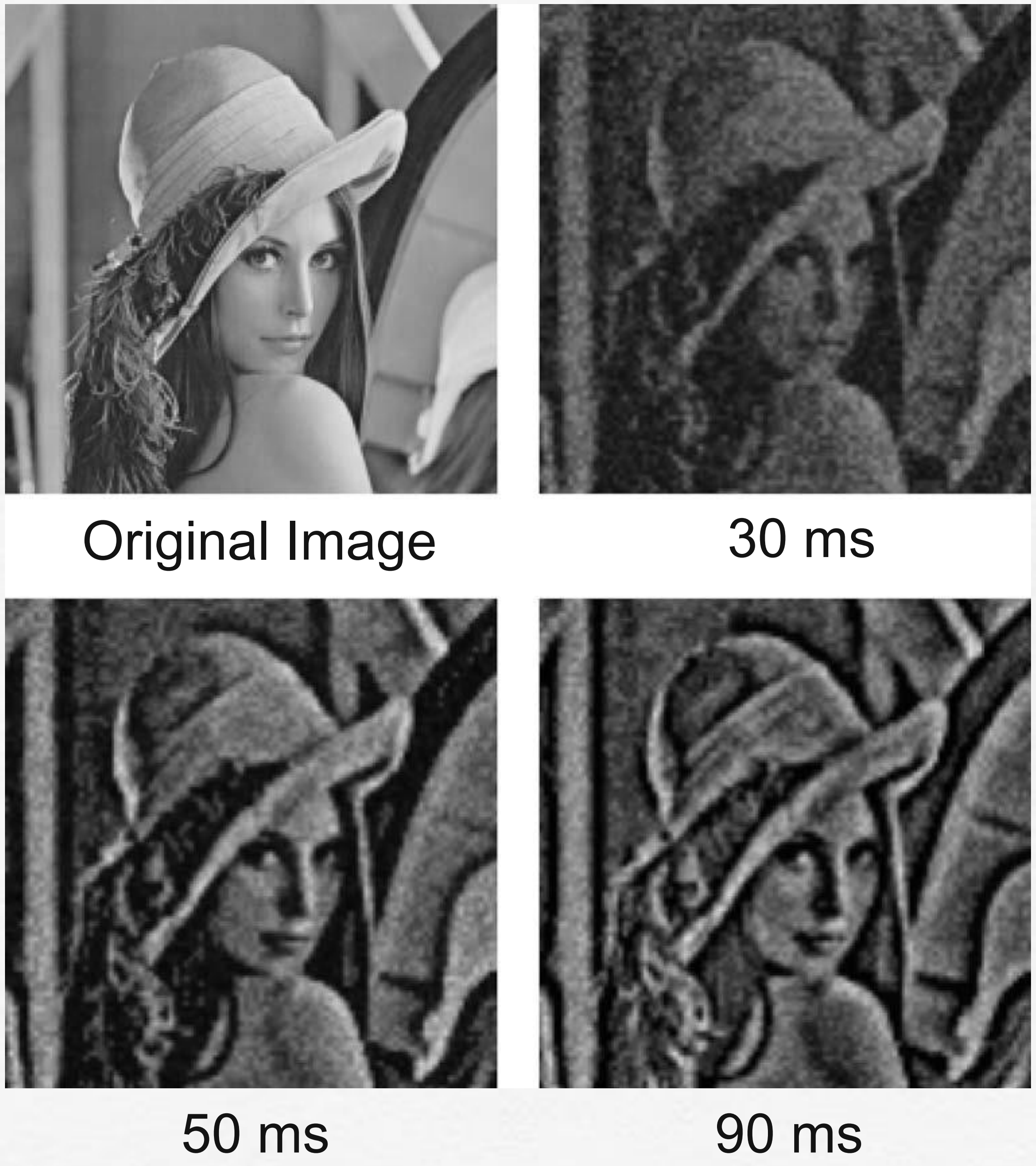


VIRTUAL RETINA

[HTTPS://TEAM.INRIA.FR/BIOVISION/VIRTUALRETINA/](https://team.inria.fr/biovision/virtualretina/)

Virtual Retina




Original Image

30 ms

50 ms

90 ms

Wohrer, A., & Kornprobst, P. (2009). Virtual Retina: A biological retina model and simulator, with contrast gain control. *Journal of Computational Neuroscience*, *26*(2), 219–249. https://doi.org/10.1007/s10827-008-0108-4

Other potential priors

What other priors do we need to incorporate to see a high-level organizational structure similar to that observed in the primate brain?

Connectivity between levels (skip connections)

Connectivity between functional systems (e.g., semantics/language)

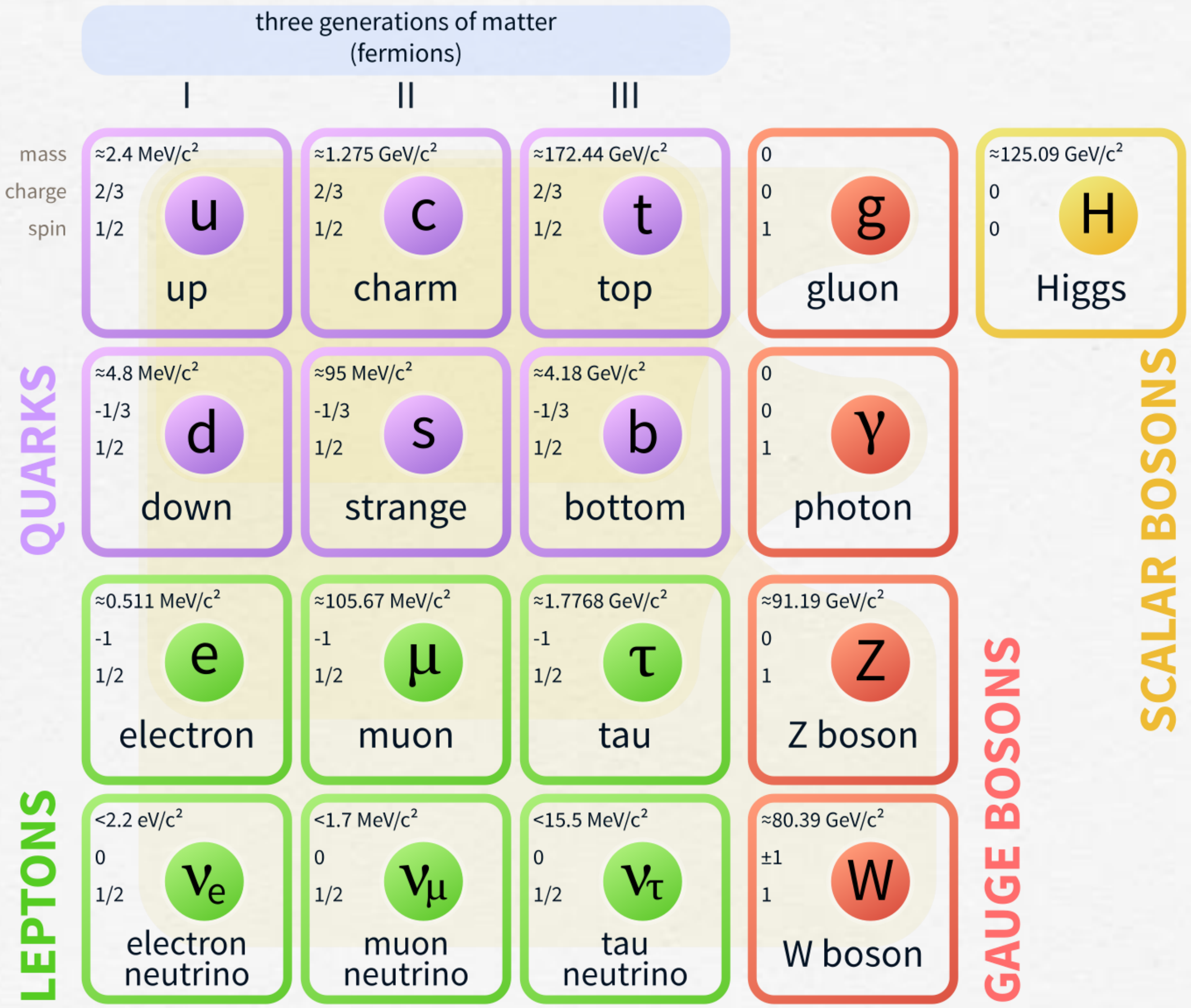
Early attentional preference for face-like images

Developmental contrast-sensitivity function that tracks primate development – importance of “starting small” – may improve learning rate and/or performance maximum

Continue to add constraints only when model fails

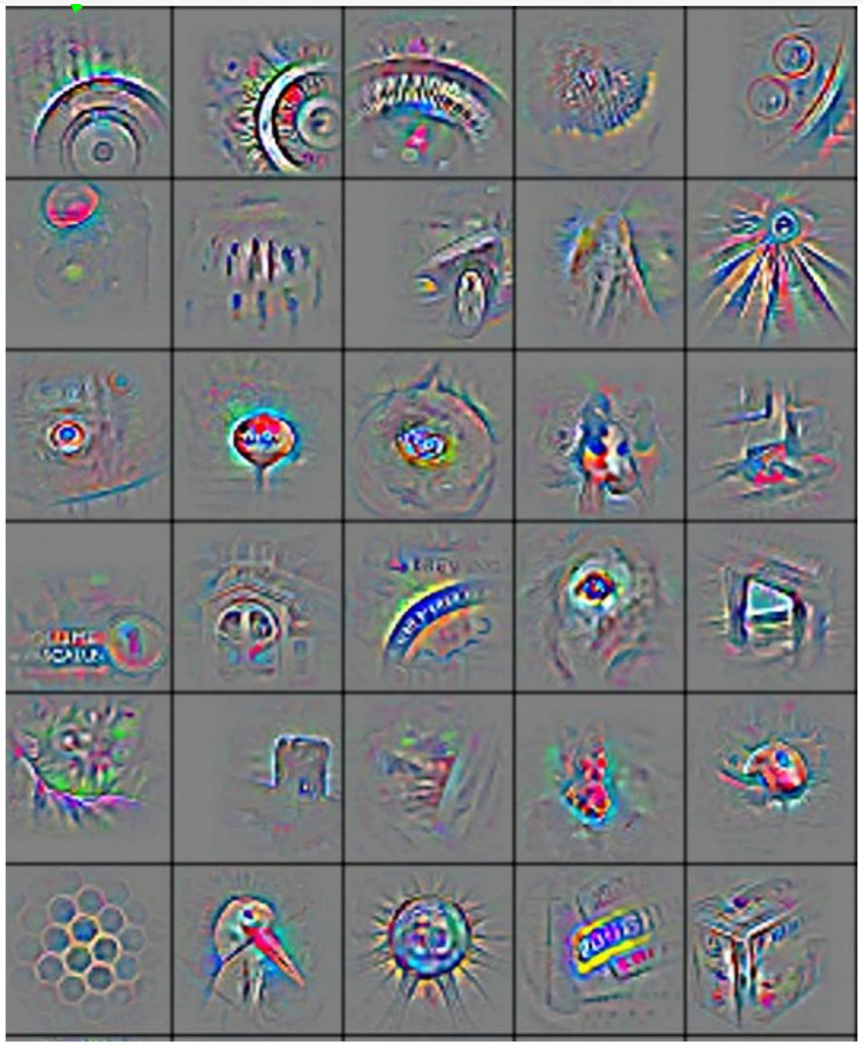
can we have “explainable” AI?

Standard Model of Elementary Particles



three generations of matter (fermions)

	I	II	III		
mass charge spin	≈2.4 MeV/c² 2/3 1/2 u up	≈1.275 GeV/c² 2/3 1/2 c charm	≈172.44 GeV/c² 2/3 1/2 t top	0 0 1 g gluon	≈125.09 GeV/c² 0 0 H Higgs
QUARKS	≈4.8 MeV/c² -1/3 1/2 d down	≈95 MeV/c² -1/3 1/2 s strange	≈4.18 GeV/c² -1/3 1/2 b bottom	0 0 1 γ photon	SCALAR BOSONS
	≈0.511 MeV/c² -1 1/2 e electron	≈105.67 MeV/c² -1 1/2 μ muon	≈1.7768 GeV/c² -1 1/2 τ tau	≈91.19 GeV/c² 0 1 Z Z boson	
LEPTONS	<2.2 eV/c² 0 1/2 ν_e electron neutrino	<1.7 MeV/c² 0 1/2 ν_μ muon neutrino	<15.5 MeV/c² 0 1/2 ν_τ tau neutrino	≈80.39 GeV/c² ±1 1 W W boson	
				GAUGE BOSONS	



compositionality