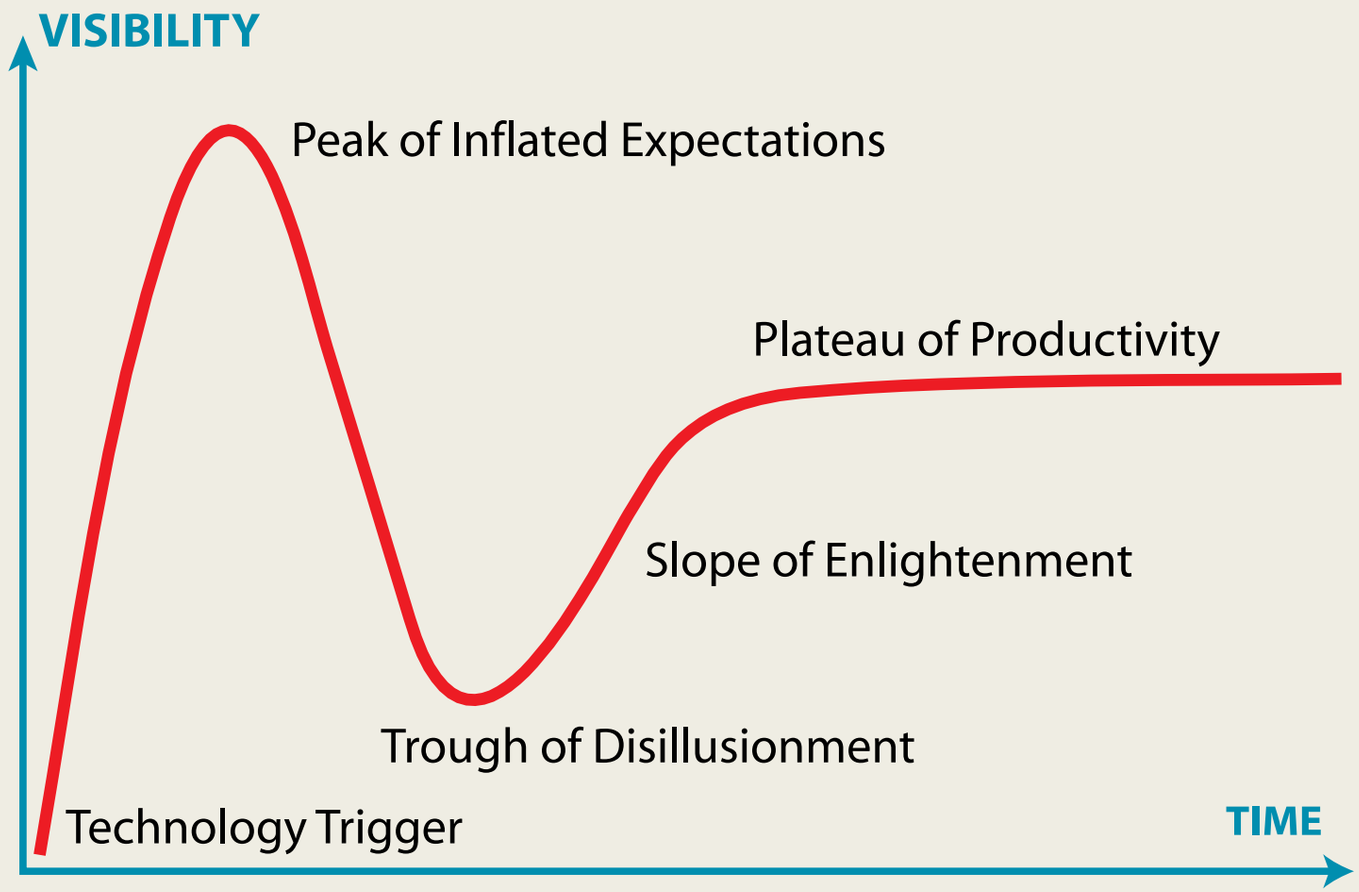
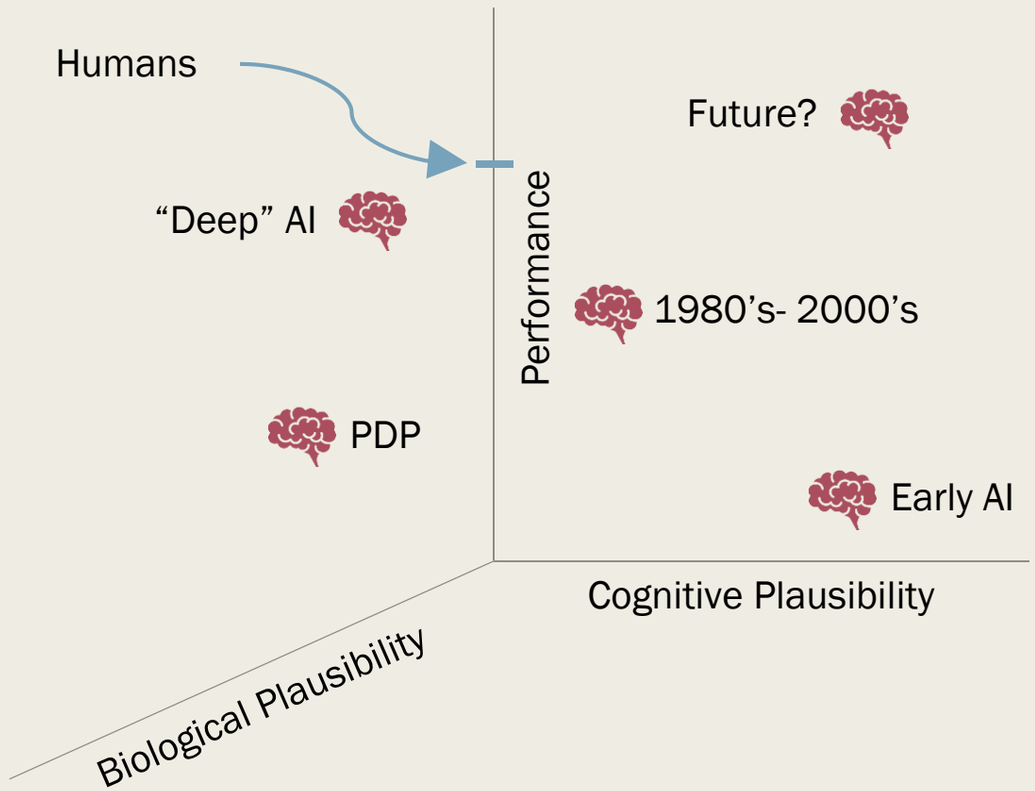
Using CNNs to understand the neural basis of vision

September 2019



AI Space


Humans
Future?
“Deep” AI
1980’s- 2000’s
PDP
Early AI
Performance
Cognitive Plausibility
Biological Plausibility

Different kinds of AI (in practice)

1. AI that maximizes performance
   - *e.g., diagnosing disease – learns and applies knowledge humans might not typically learn/apply – "who cares if it does it like humans or not"*
2. AI that is meant to simulate (to better understand) cognitive or biological processes
   - *e.g., PDP – specifically constructed so as to reveal aspects of how biological systems learn/reason/etc. – understanding at the neural or cognitive levels (or both)*
3. AI that performs well *and* helps understand cognitive or biological processes
   - *e.g., Deep learning models (cf. Yamins/DiCarlo) – "representational learning"*
4. AI that is specifically designed to *predict* human performance/preference
   - *e.g., Google/Netflix/etc. – only useful if it predicts what humans actually do or want*

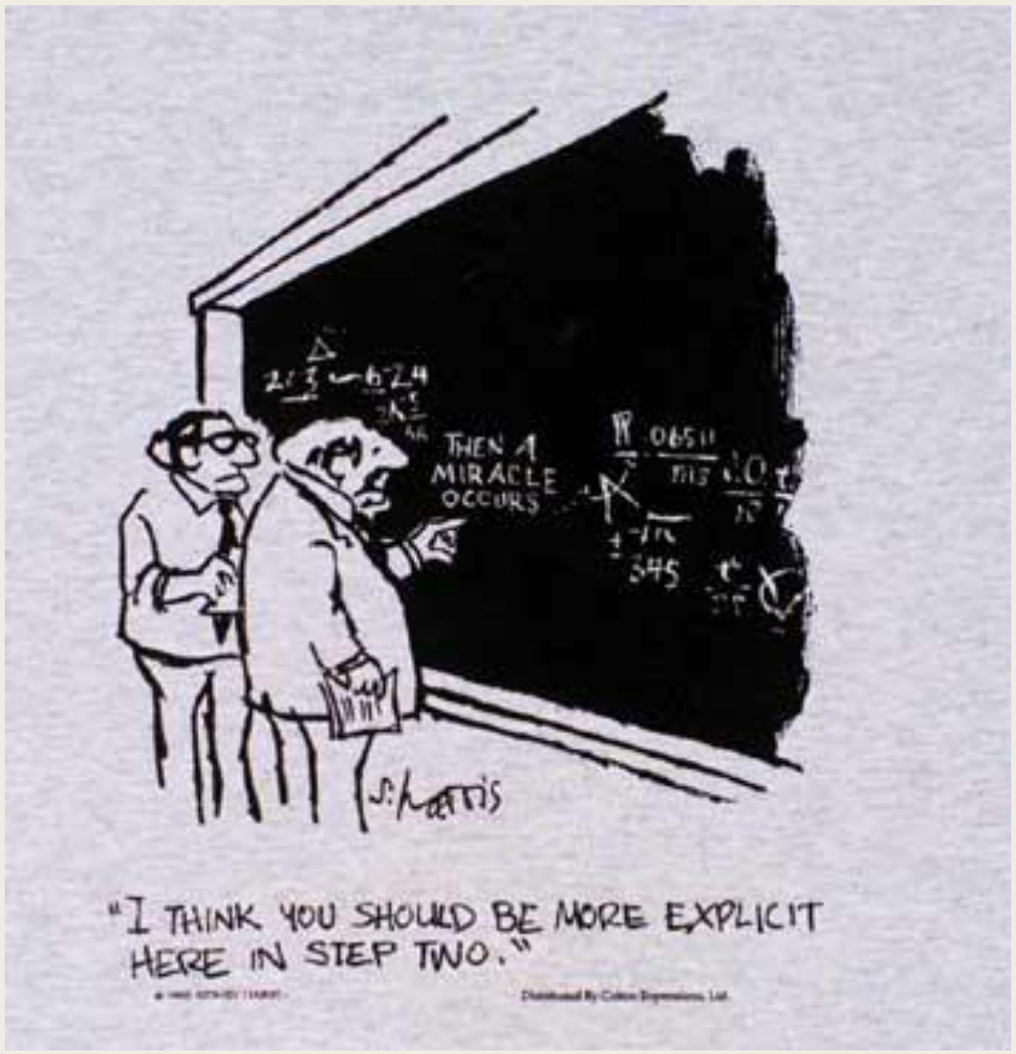
A Bit More on Deep Learning

- Typically relies on *supervised* learning – 1,000,000's of labeled inputs
- Labels are a metric of human performance – so long as the network learns the correct input->label mapping, it will perform "well" by this metric
  - *However, the network can't do better than the labels*
  - *Features might exist in the input that would improve performance, but unless those features are sometimes correctly labeled, the model won't learn that feature to output mapping*
- The network can reduce misses, but it can't discover new mappings unless there existing further correlations between input->labels in the trained data
- So Deep Neural Networks tend to be very good at the kinds of AI that predicts human performance (#4) and that maximize performance (#1), but the jury is still out on AI that performs well and helps us understand biological intelligence (#3); might also be used for simulation of biological intelligence (#2)

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

the problem


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

so how do we fill in the blanks?

- early vision (filters)
  - *image filtering, data reduction*
- mid-level vision (unsupervised)
  - *multiple information channels*
  - *cue combination, binding*
- high-level vision (supervised)
  - *coherent objects, events, and scenes*

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

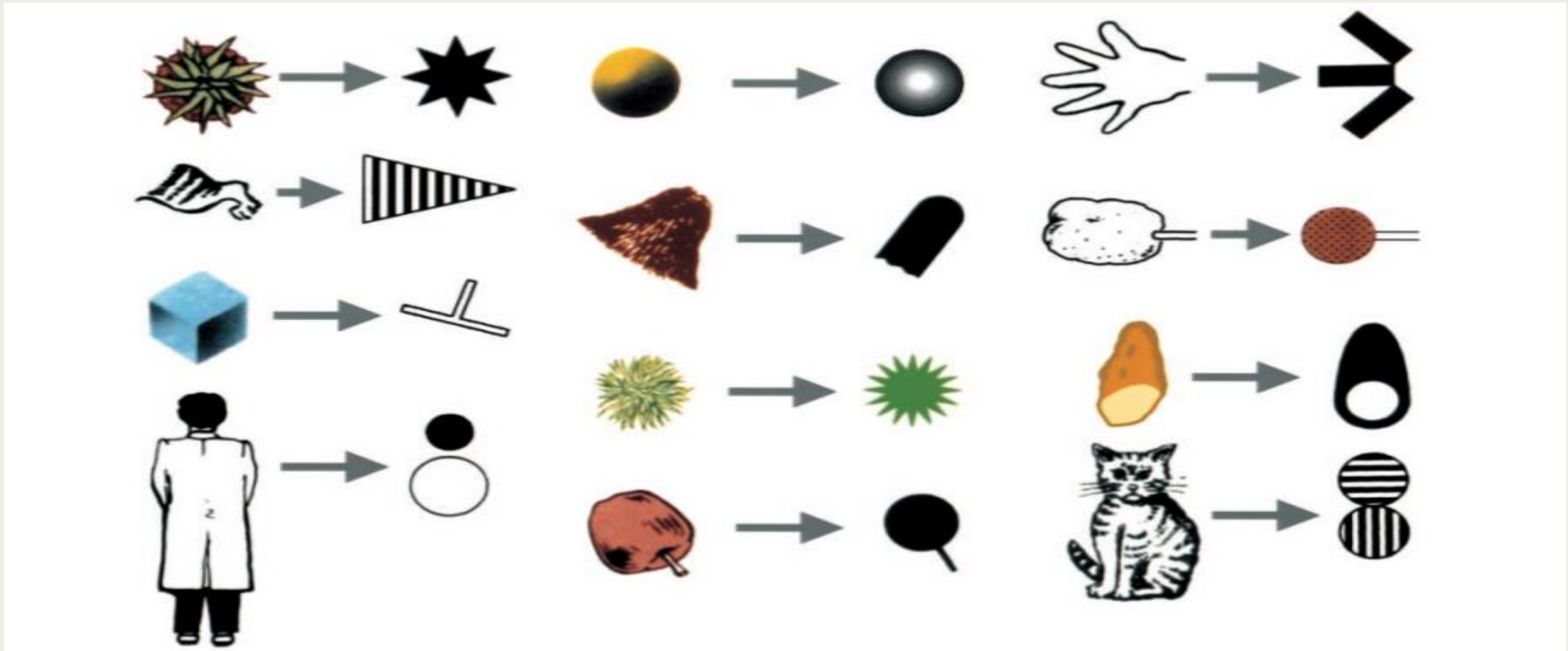
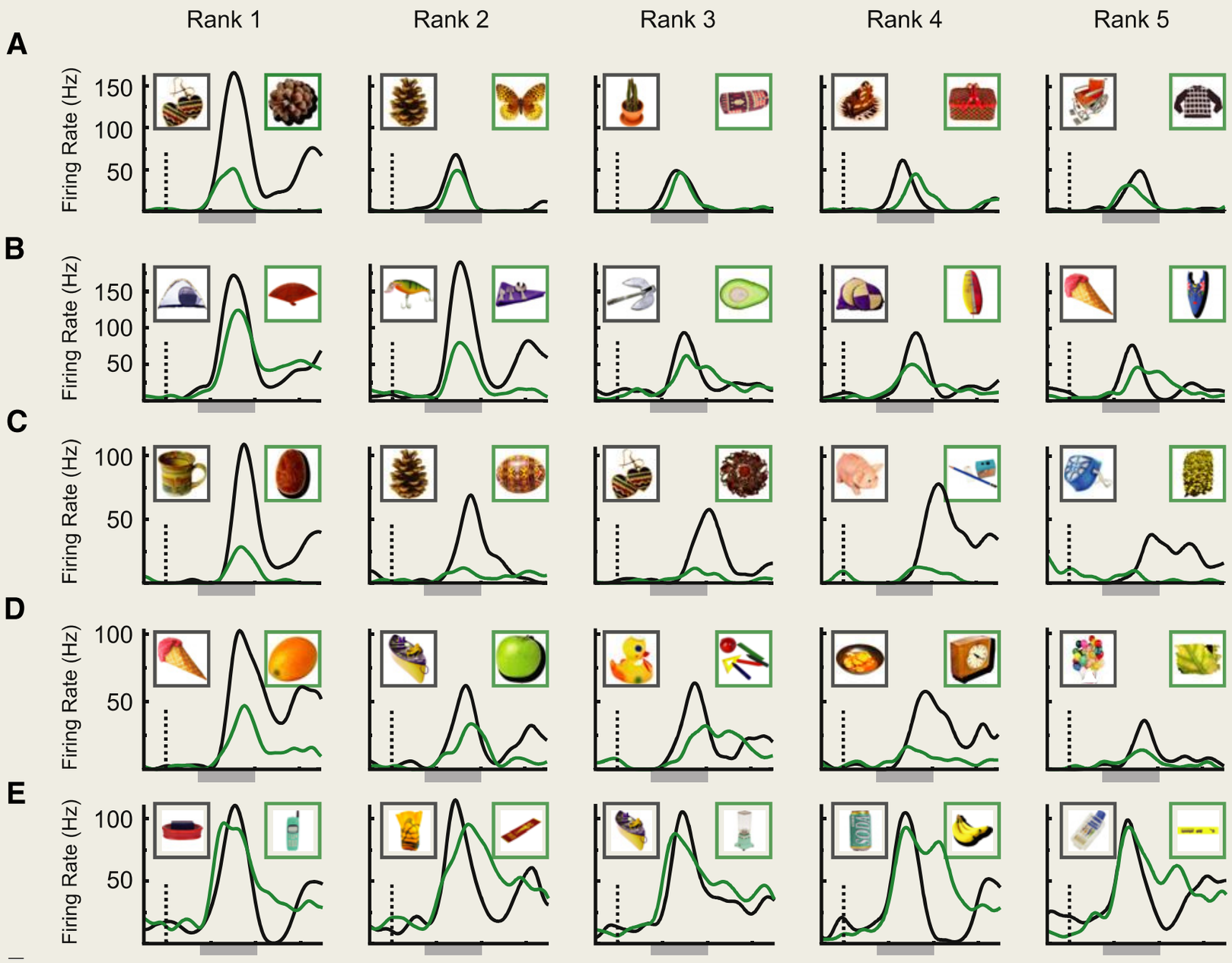


Figure 1. Examples of reductive determination of optimal features for 12 TE cells. The images to the left of the arrows represent the original images of the most effective object stimulus and those to the right of the arrows, the critical features determined by the reduction.

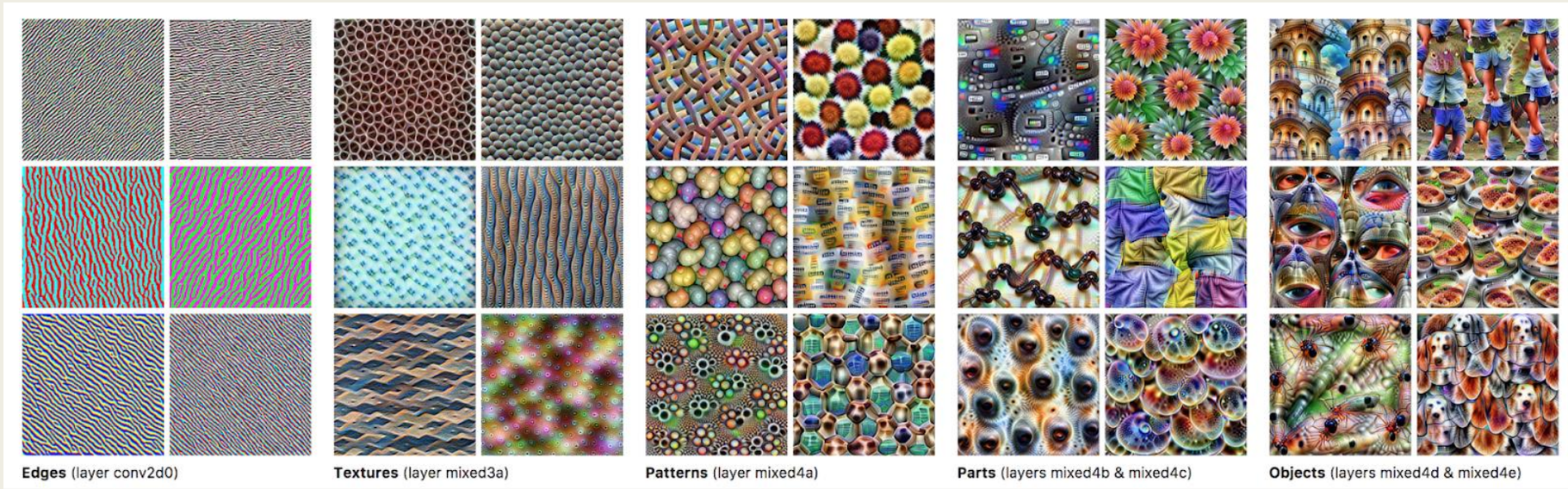
Woloszyn and Sheinberg (2012)



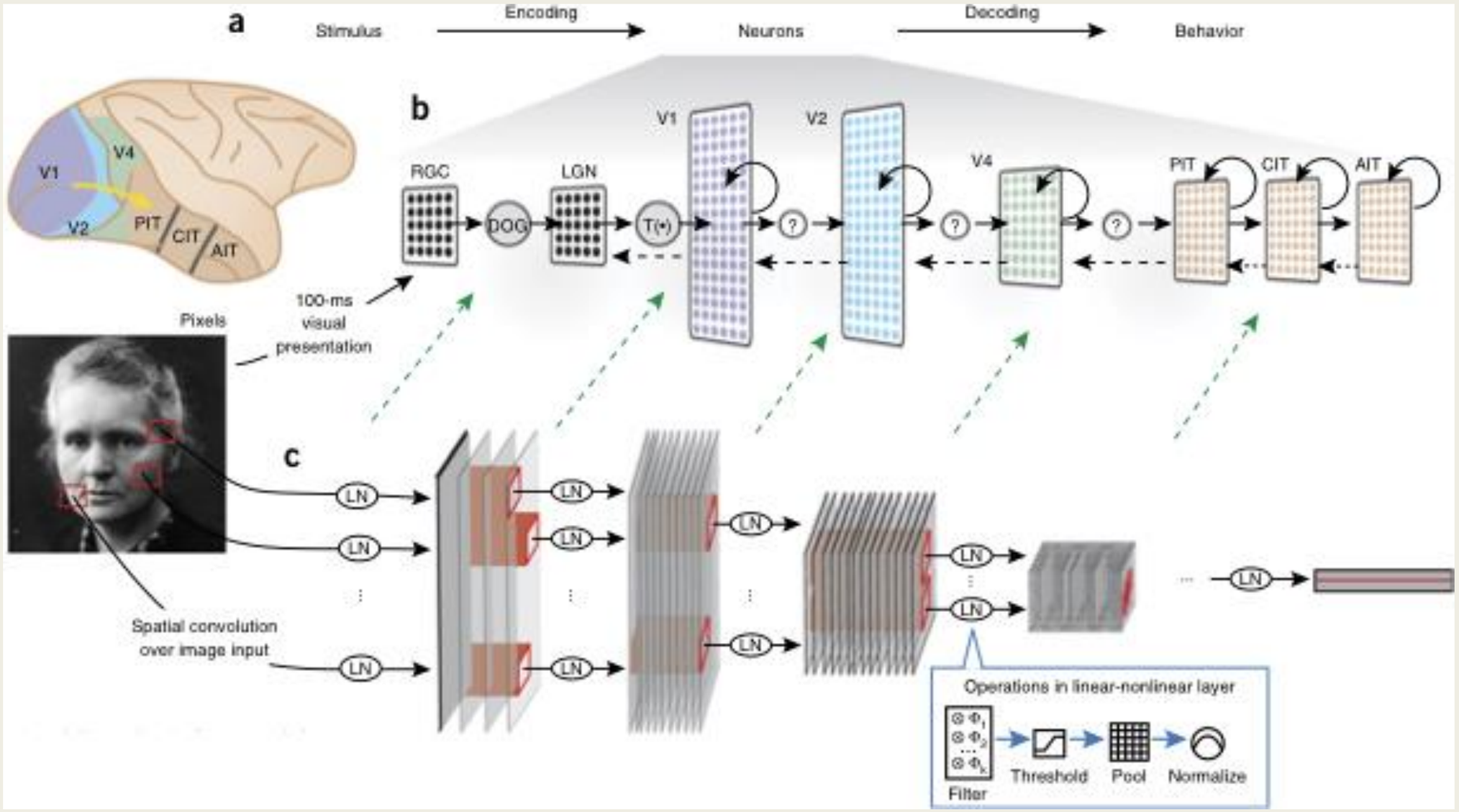
Stupid CNN Tricks

- Hierarchical correspondence
- Visualization of “neurons”

[Digression – is visualization a good metric for evaluating models?]



HCNNs are good candidates for models of the ventral visual pathway



Yamins & DiCarlo

Goal-Driven Networks as Neural Models

- whatever parameters are used, a neural network will have to be effective at solving the behavioral tasks the sensory system supports to be a correct model of a given sensory system
- so… advances in computer vision, etc. that have led to high-performing systems – that solve behavioral tasks nearly as effectively as we do – *could* be correct models of neural mechanisms
- conversely, models that are ineffective at a given task are unlikely to ever do a good job at characterizing neural mechanisms

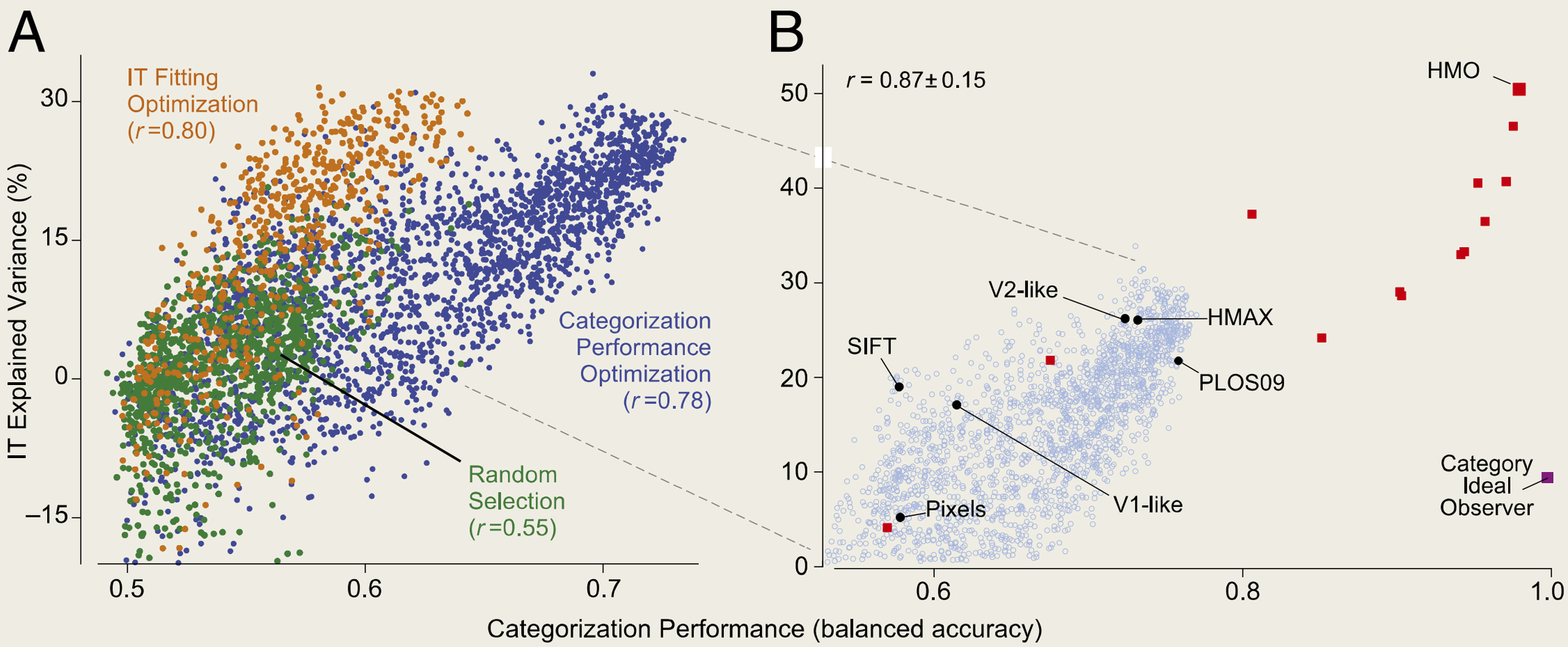
Approach

- Optimize network parameters for performance on a reasonable, ecologically—valid task
- Fix network parameters and compare the network to neural data
- Easier than “pure neural fitting” b/c collecting millions of human-labeled images is easier than obtaining comparable neural data

Key Questions

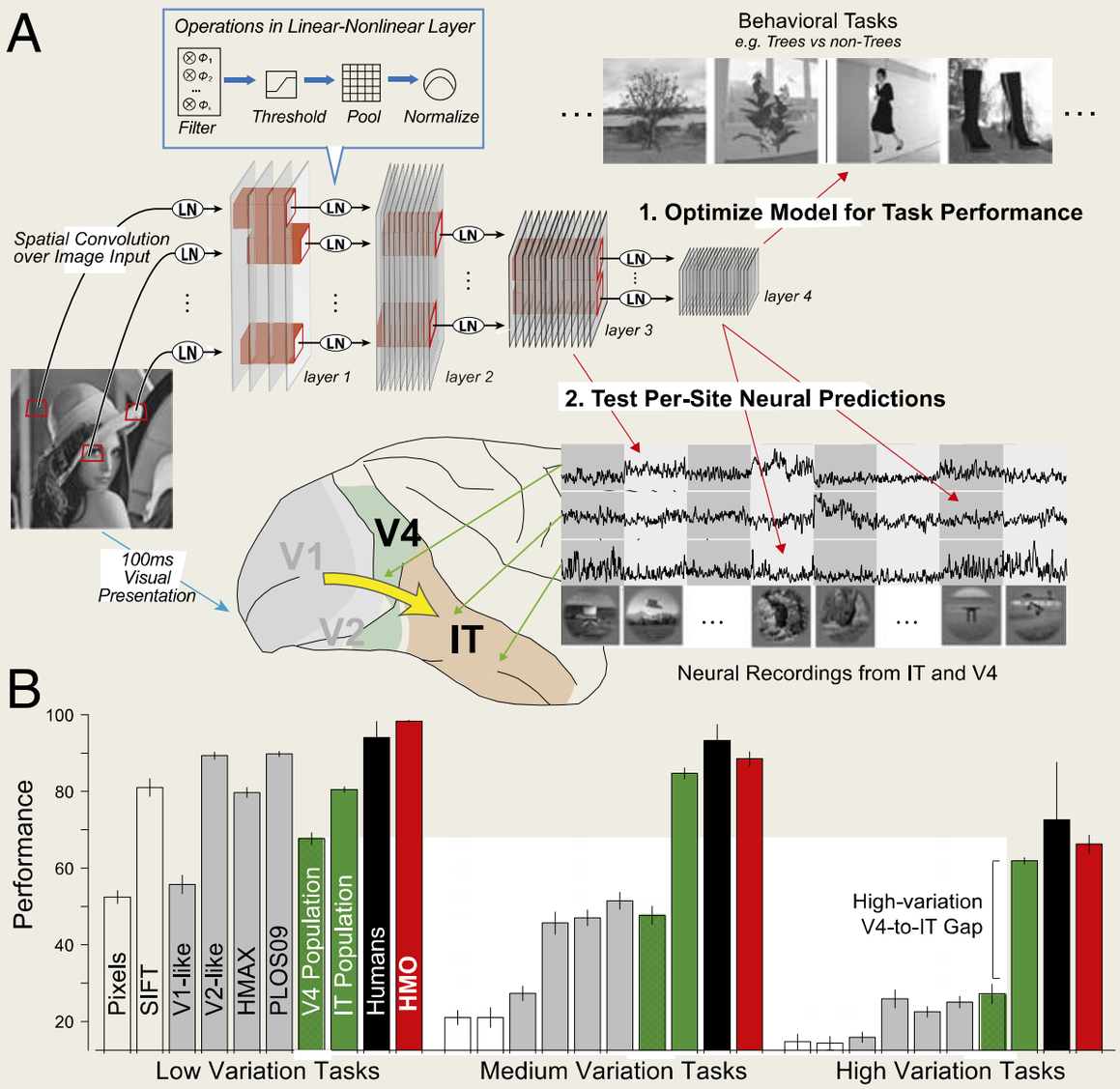
- Do such top-down goals – tasks – constrain biological structure?
- Will performance optimization be sufficient to cause intermediate units in the network to behave like neurons?

Model Performance/IT-Predictivity Correlation



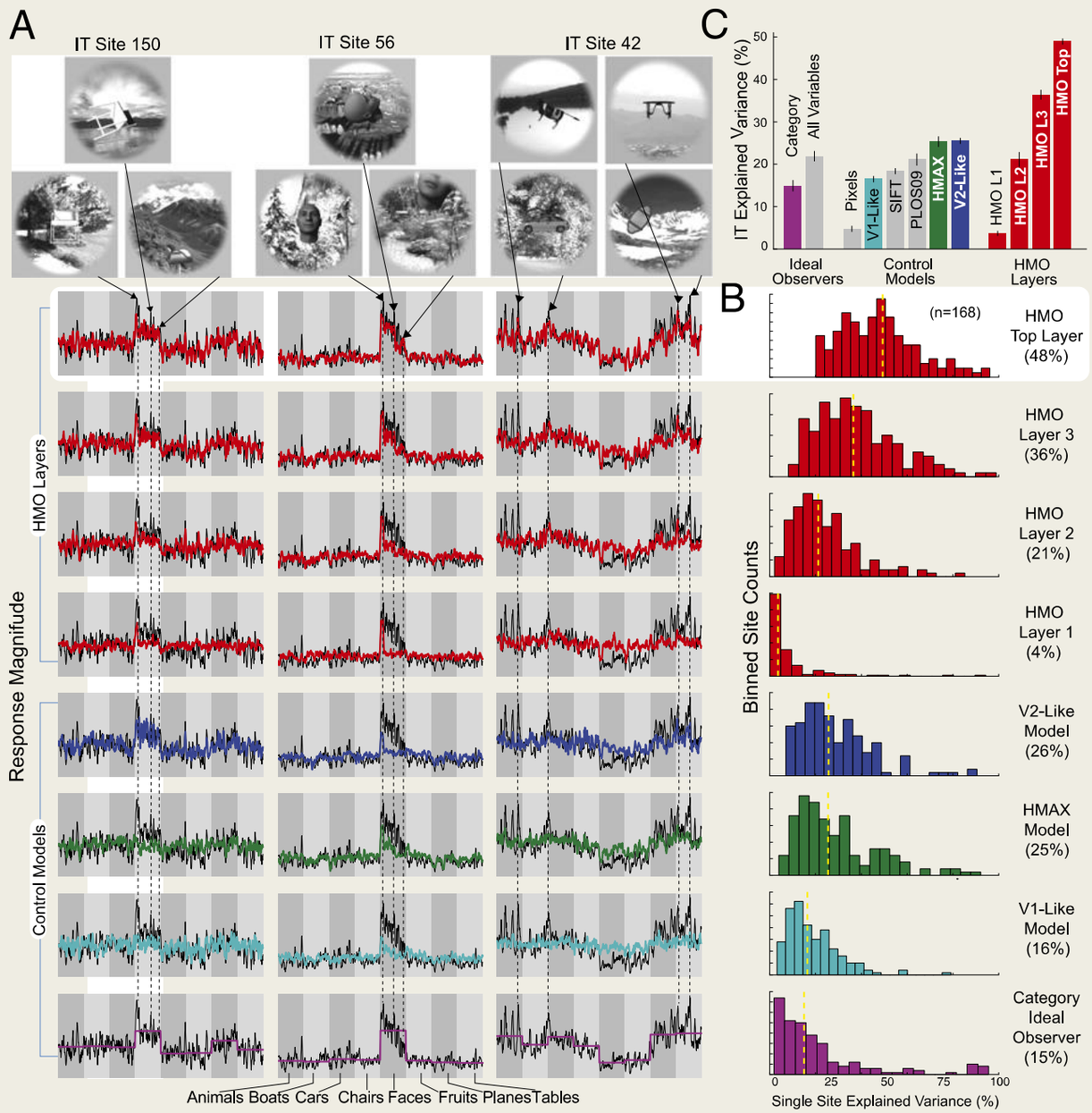
Yamins et al.

“Neural-like” models via performance optimization



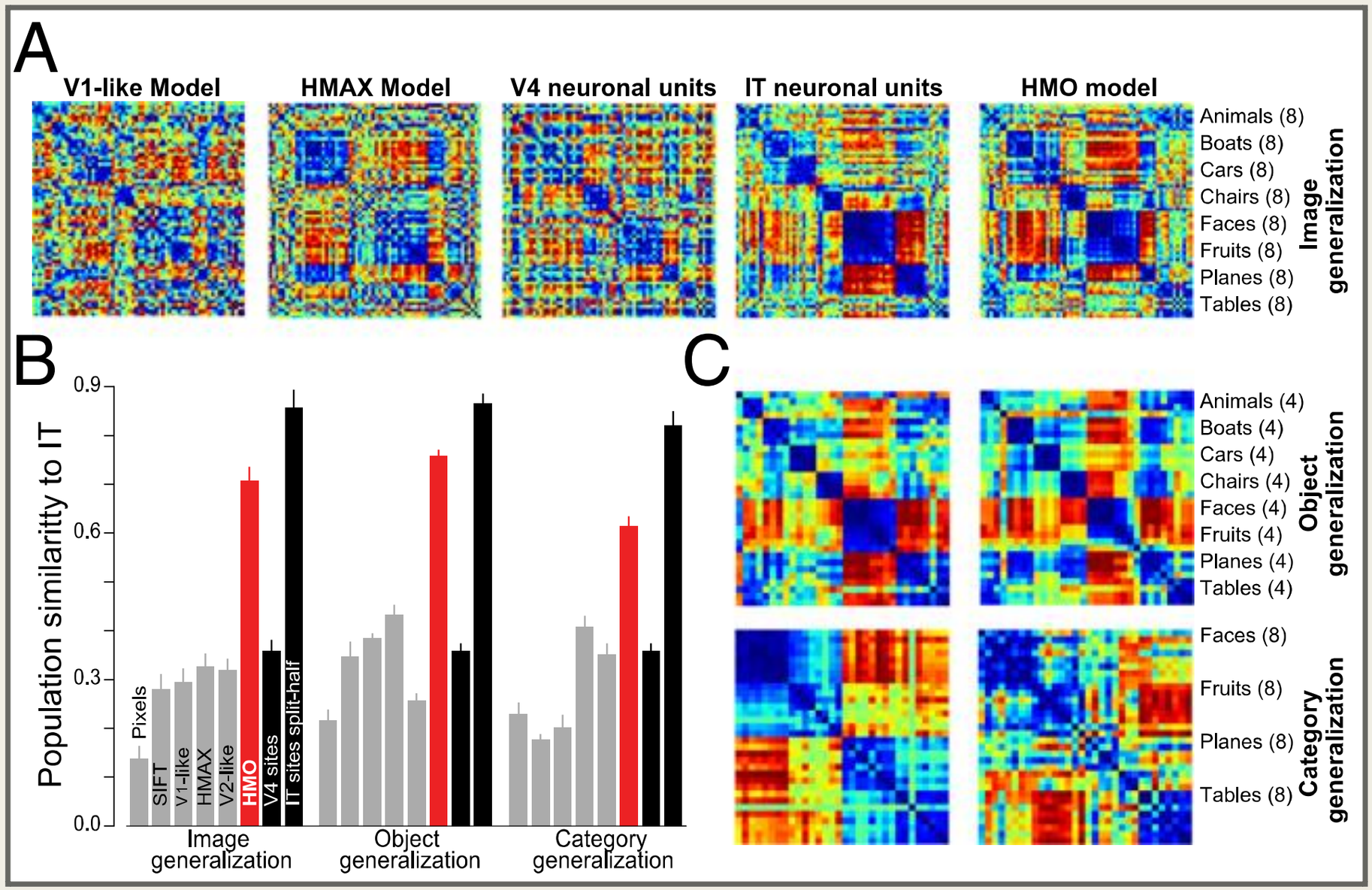
Yamins et al.

IT Neural Predictions



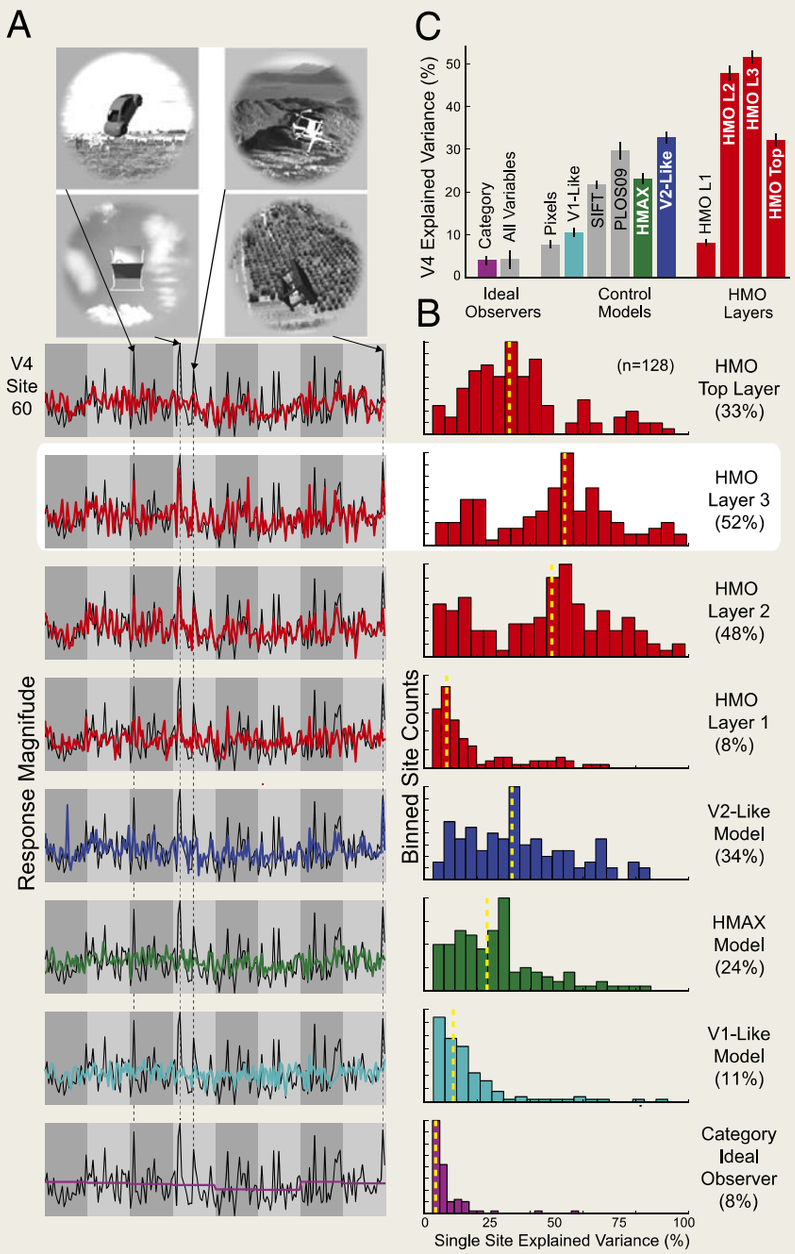
Yamins et al.

Population-level Similarity



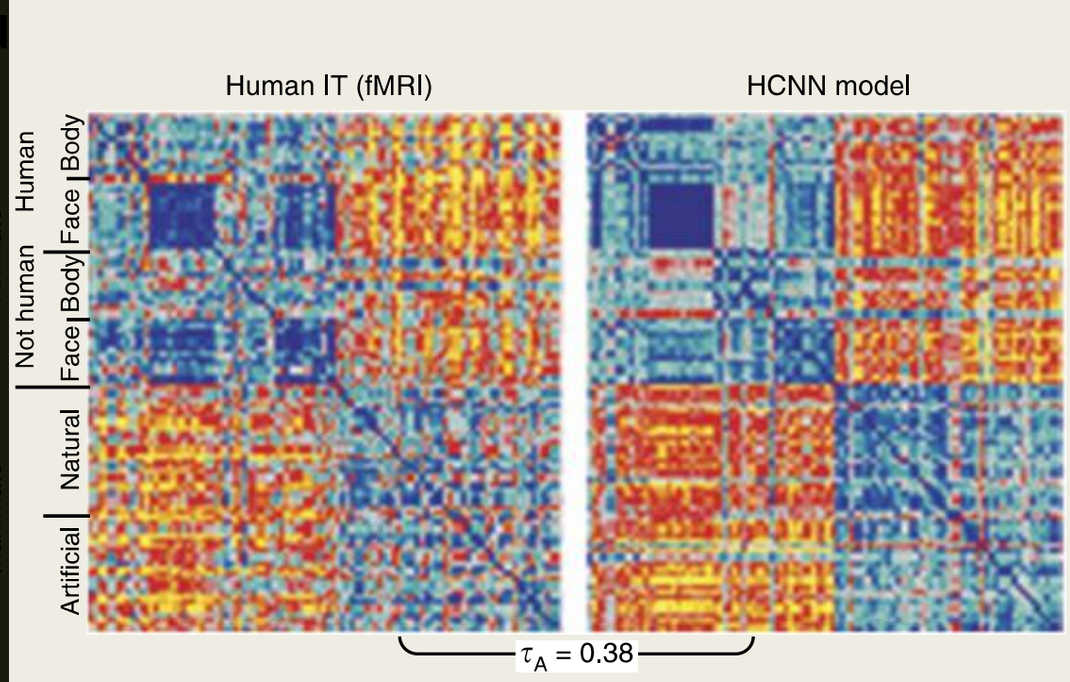
Yamins et al.

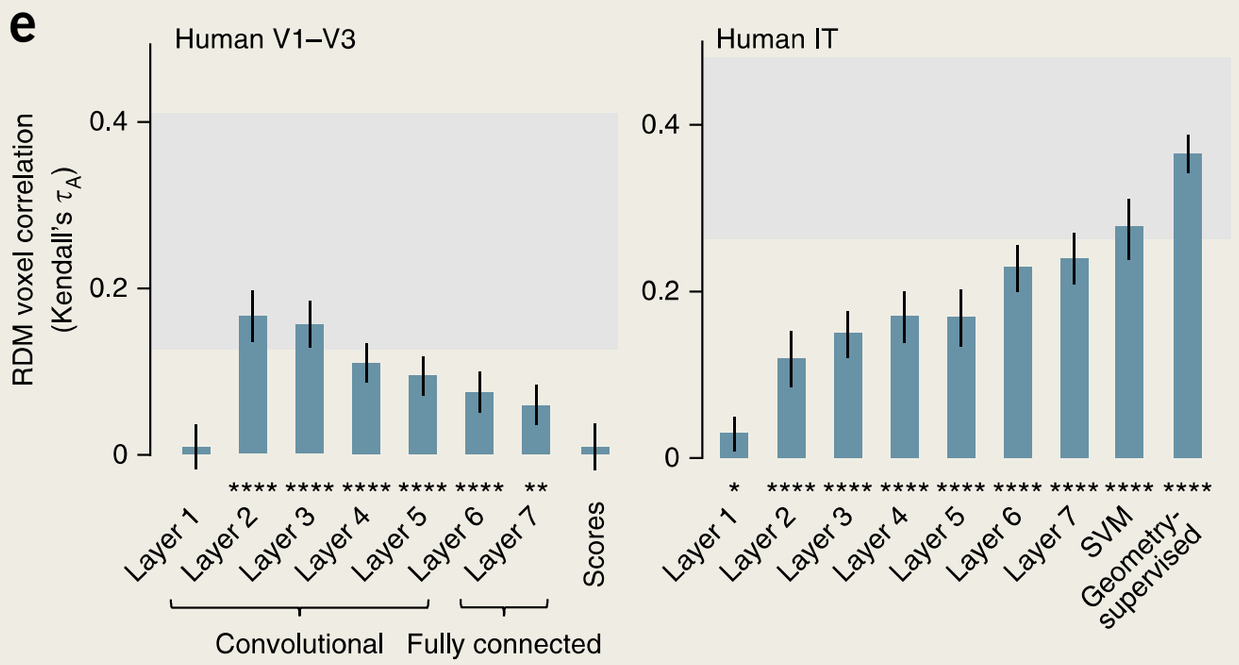
V4 Neural Predictions



Yamins et al.

Human fMRI

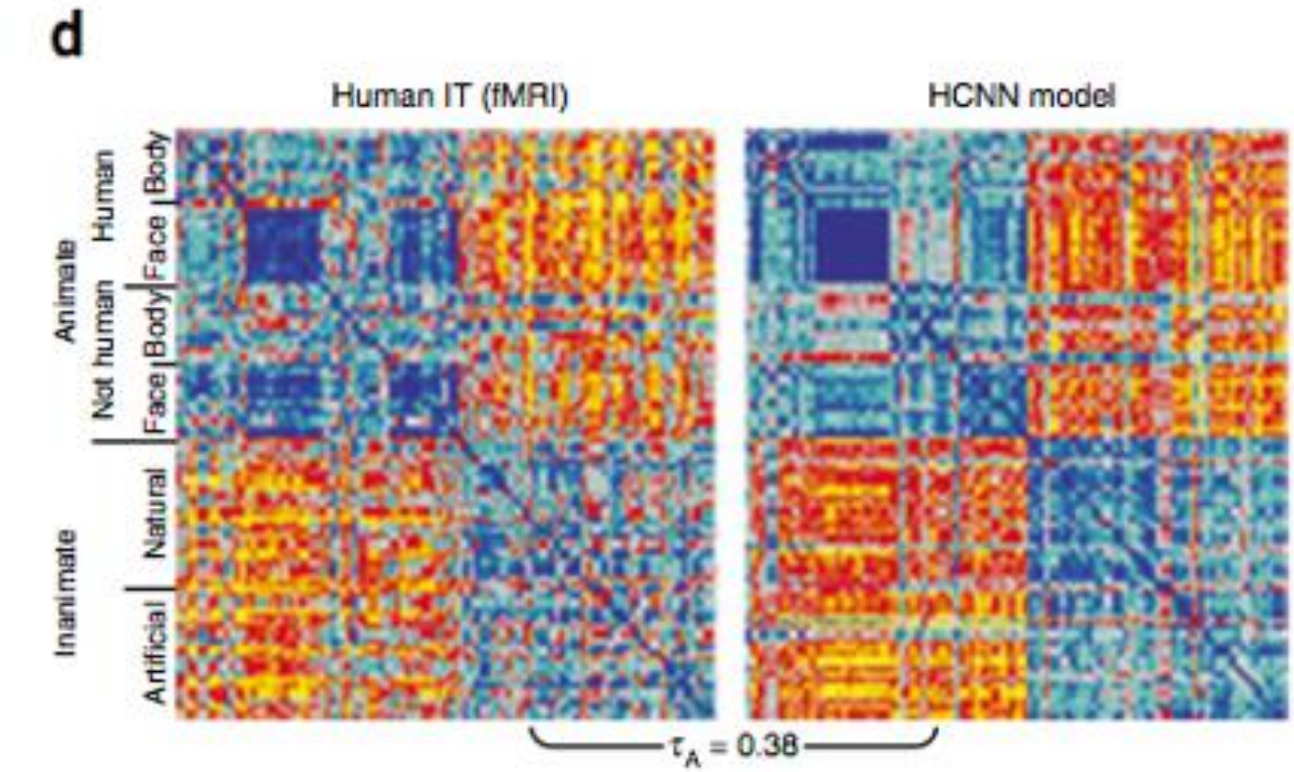




Khaligh-Razavi & Kriegeskorte

RSA – Baselines?

Representational similarity of human IT with DCNN higher layers (distance between population codes of stimuli)

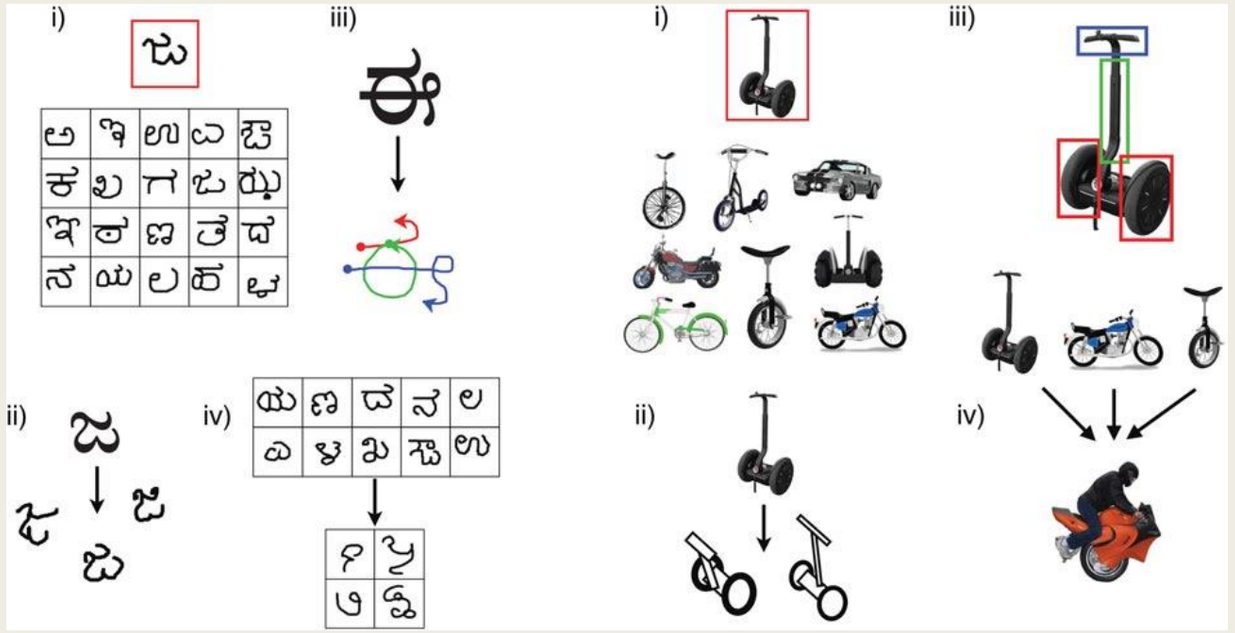


Semantic clustering in IT

Yamins and DiCarlos. Nature Neuroscience (2016)

Challenges

- Lake et al. articulate two challenge problems to elucidate the role of early inductive biases and the ability to learn based on small amounts of data
  - *Learning simple visual concepts*
  - *Playing a video game*



Do deep networks and humans perform this sort of task in the same way?

- Two important differences:

1. People learn from fewer examples
2. People learn “richer” representations
    - *Decomposable into parts*
    - *Learn a concept that can be flexibly applied*
        - Generate new examples
        - Parse an object into parts and their relations
        - Generalize to new instances of the overall class

"This richness and flexibility suggest that learning as model building is a better metaphor than learning as pattern recognition. Furthermore, the human capacity for one-shot learning suggests that these models are built upon rich domain knowledge rather than starting from a blank slate."

- Two (non) issues:
  - *Generative capacities*
    - Generative Adversarial Networks (GANs) are capable of learning and generating new exemplars within categories
  - *Few-shot learning*
    - There are many recent implementations of learning from a small number of examples; moreover, the fact that humans can do this (sometimes*) isn't strong evidence for model-driven learning in and of itself

*I think they overestimate both the amount data we learn from and how effective humans are at this

Duh. The particular model being tested did not have general world knowledge/context – it only was intended to perform captioning using simple object and scene labeling (~semantics)

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

BOLD5000

TSNE

Encoding Models

To explore how and where visual features are represented in human scene processing, we extracted different features spaces describing each of the stimulus images and used them in an encoding model to predict brain responses.

Leveraging Brain Data

My own work: what are the *minimal* assumptions needed to give rise to high-level structure/concepts?

Three current projects:

- How does the basic spatiotopic and processing hierarchy of the primate visual system arise?
  - *Arcaro: "proto-structure" present in newborn monkeys*
  - *Testing whether this can be learned* in vivo *given only retinal structure*
- Can deep network architectures rapidly learn new categories using only a few examples?
  - *Leverage the natural "clumpiness" of almost all visual categories and simple "nearest neighbor" visual reasoning*
- Can visual category learning be accelerated by early developmental constraints?
  - *Infant vision is high contrast and blurry, yielding inputs of reduced dimensionality (relative to adult vision)*