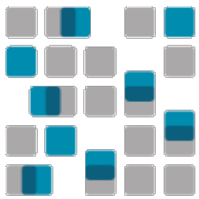
The Science of Natural Intelligence: Reverse engineering primate visual perception

AAAI Symposium, Computational Principles of Natural and Artificial Intelligence, March 2017

James DiCarlo MD, PhD

Peter de Florez Professor of Neuroscience
Head, Department of Brain and Cognitive Sciences
Investigator, The McGovern Institute for Brain Research
Massachusetts Institute of Technology, Cambridge MA, USA



brain+cognitive sciences





The Science of Natural Intelligence: Reverse engineering primate visual perception

AAAI Symposium, Computational Principles of Natural and Artificial Intelligence, March 2017



~3 pounds
~20 watts

Humans have “strong perception.” Can we reverse engineer that?



Image adapted from MIT Street Scenes Database (Courtesy of Tommy Poggio)

We started (~2000) by trying to reverse engineer object detection and categorization

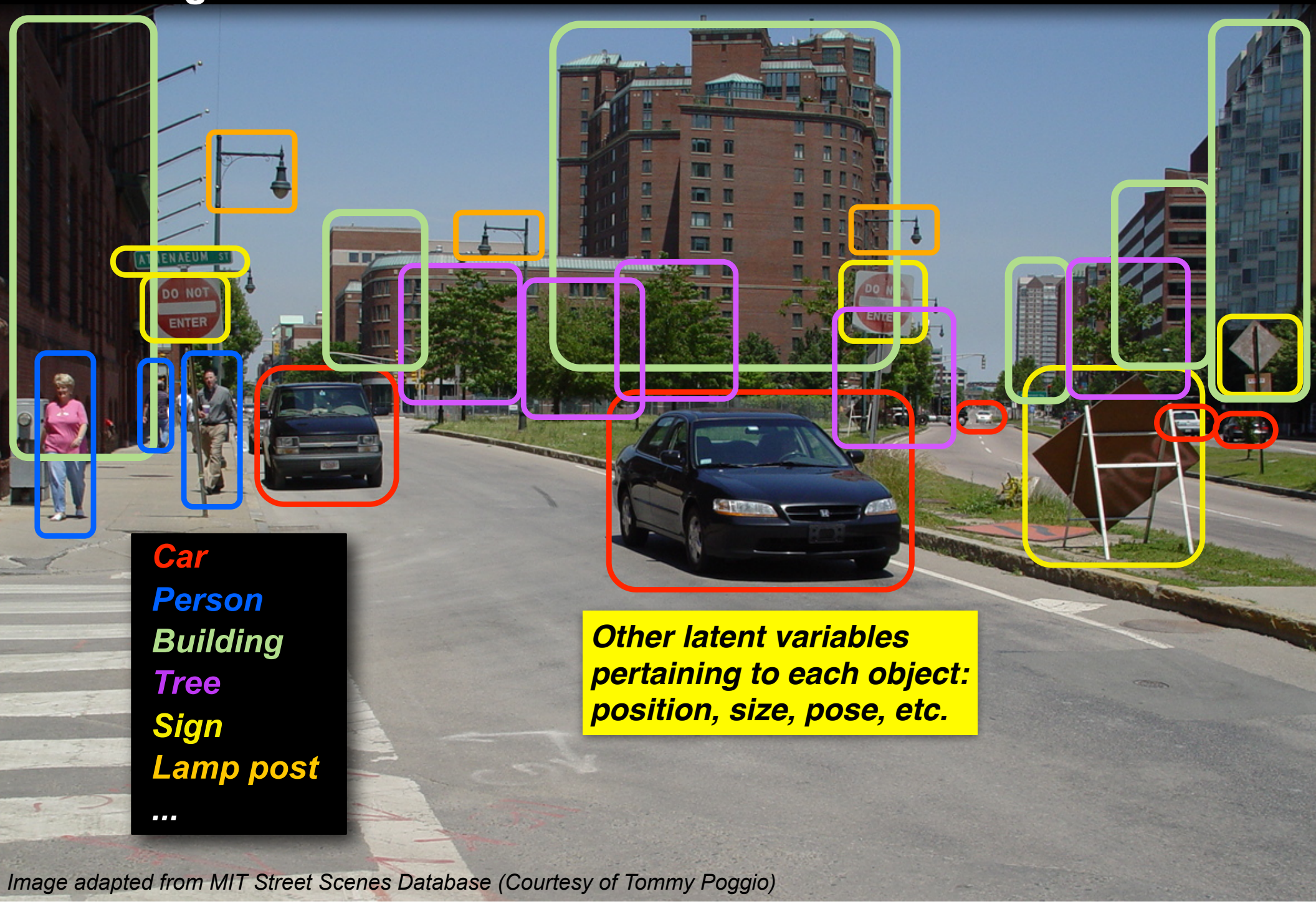


Image adapted from MIT Street Scenes Database (Courtesy of Tommy Poggio)

Constraints from brain and cognitive sciences… Central ~10 degrees



Constraints from brain and cognitive sciences… ~200 ms snapshots

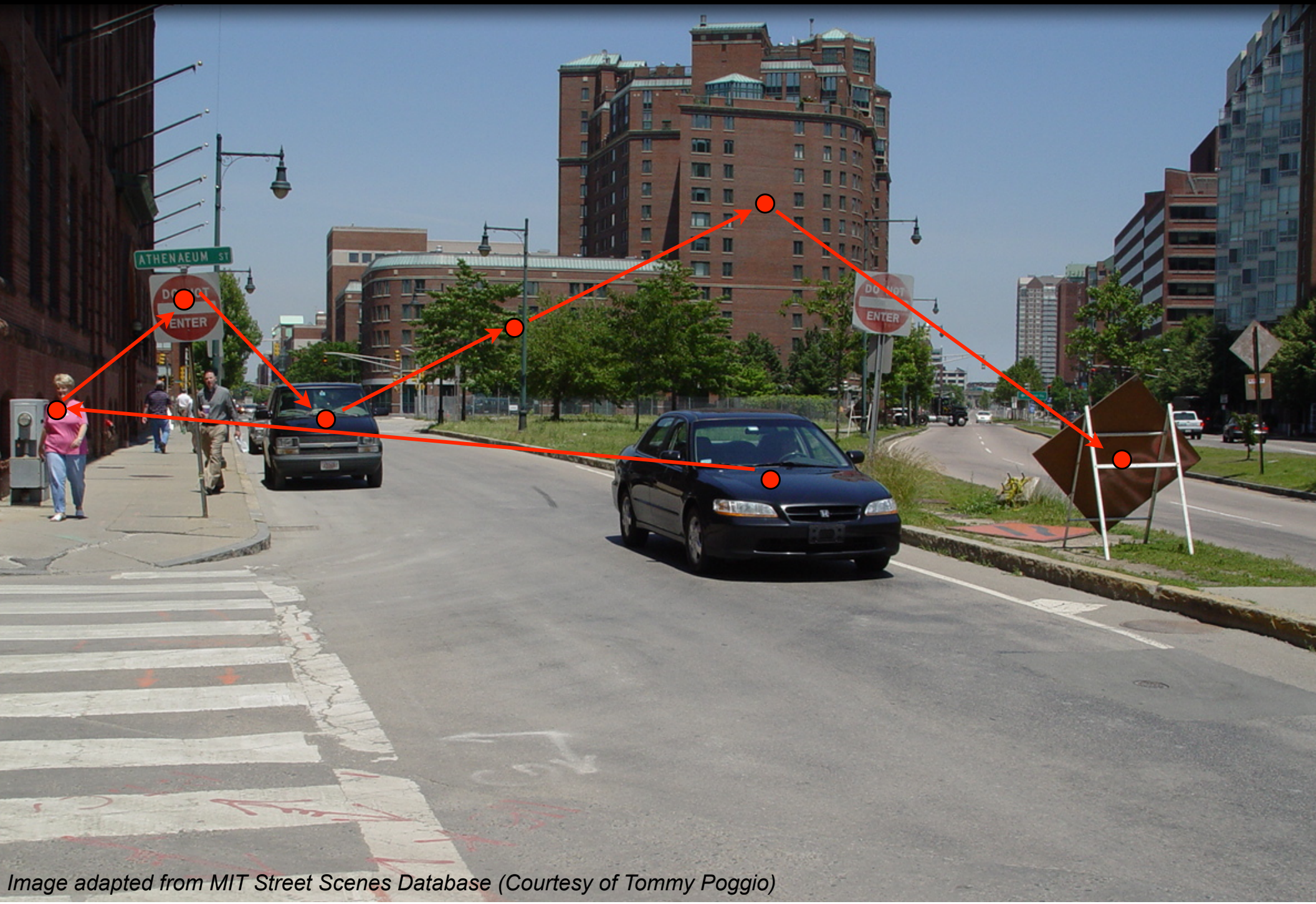


Image adapted from MIT Street Scenes Database (Courtesy of Tommy Poggio)

Object detection/categorization as solved by primates

Core object perception

central ~10 deg of visual field
~200 ms viewing duration

Not surprisingly, the primate brain excels at core object perception

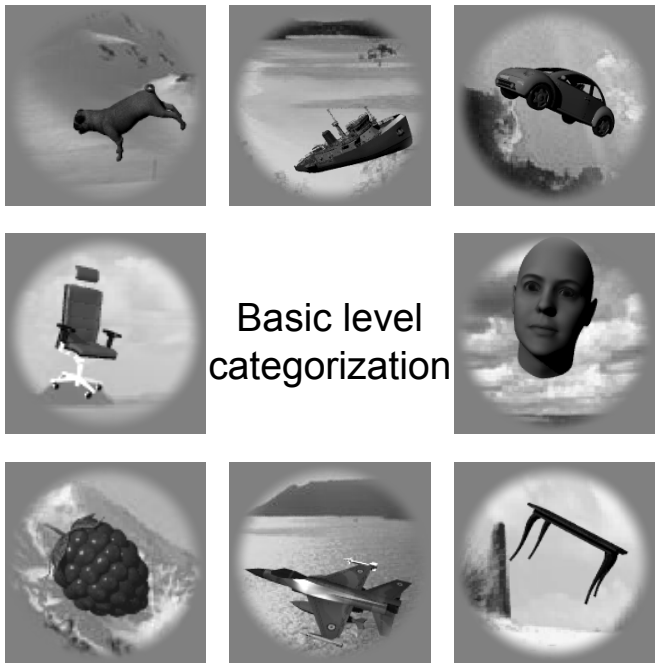
Core object perception

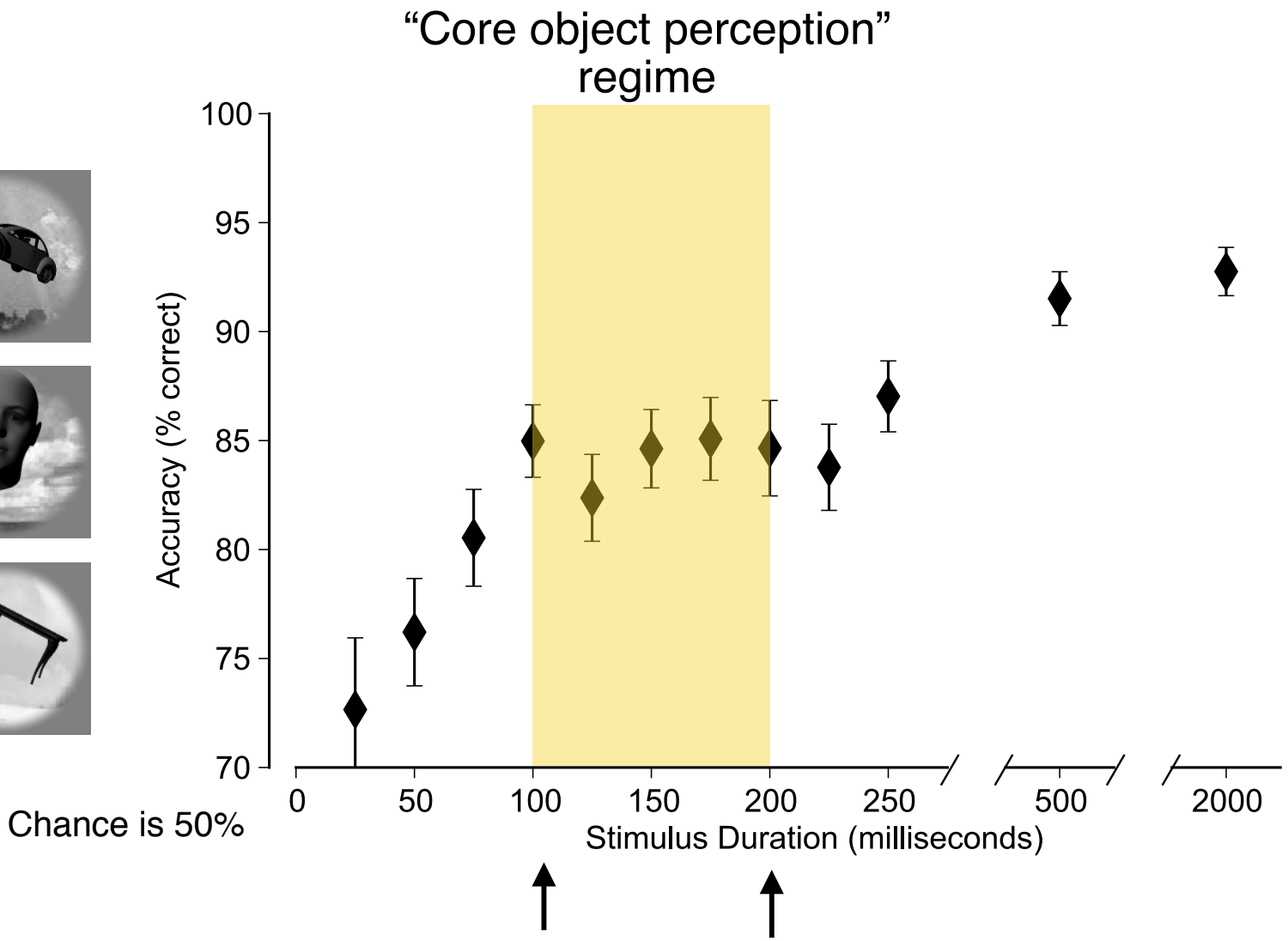
central ~10 deg of visual field
~200 ms viewing duration

8 deg image at center of gaze, 100 ms viewing time

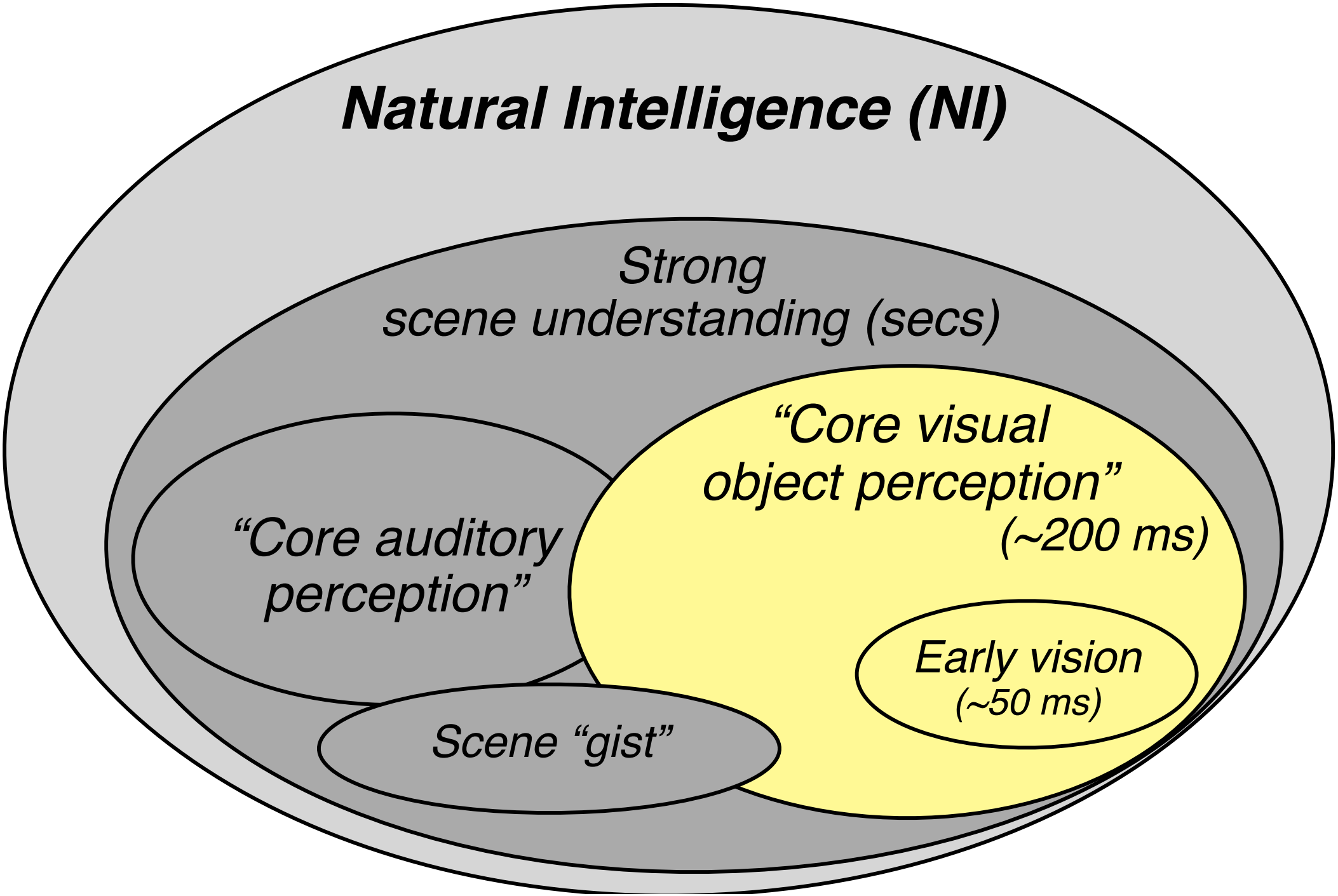


Human object categorization accuracy as a function of image viewing time





A stepwise approach to reverse engineering NI



A stepwise approach to reverse engineering NI

Core visual object perception
(central 10 deg, first ~200 ms)

Q1

Where do currently engineered systems fall down (relative to the human primate)?*

Q2

Which human primate measurements are likely to be most informative?
(behavior, blood flow, neural activity, anatomy, neural perturbation, sub-cellular, genetics, etc.)

Can we get those measurements (during system operation)?

Q3

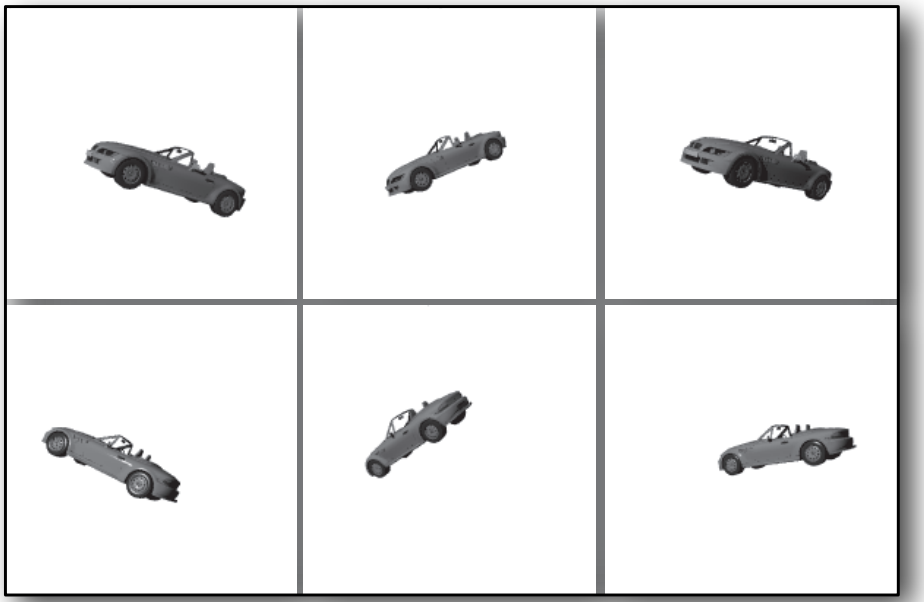
How do we forward engineer from such measurements?

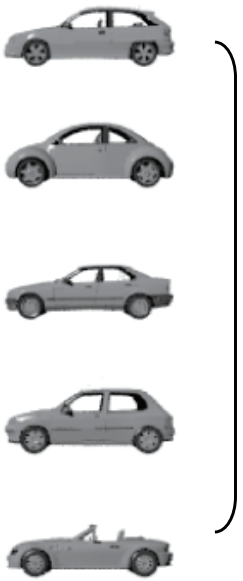
Q1 Where do (did?) engineered systems fall down (relative to humans)?



Identity preserving image variation

View: position, size, pose, illumination



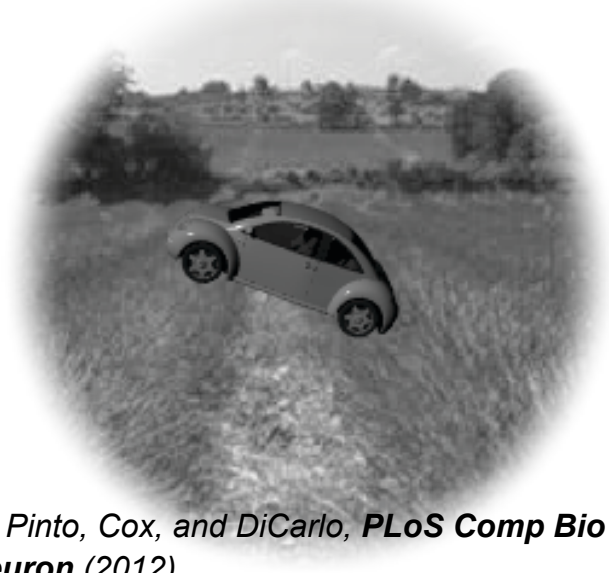


subordinate level variation





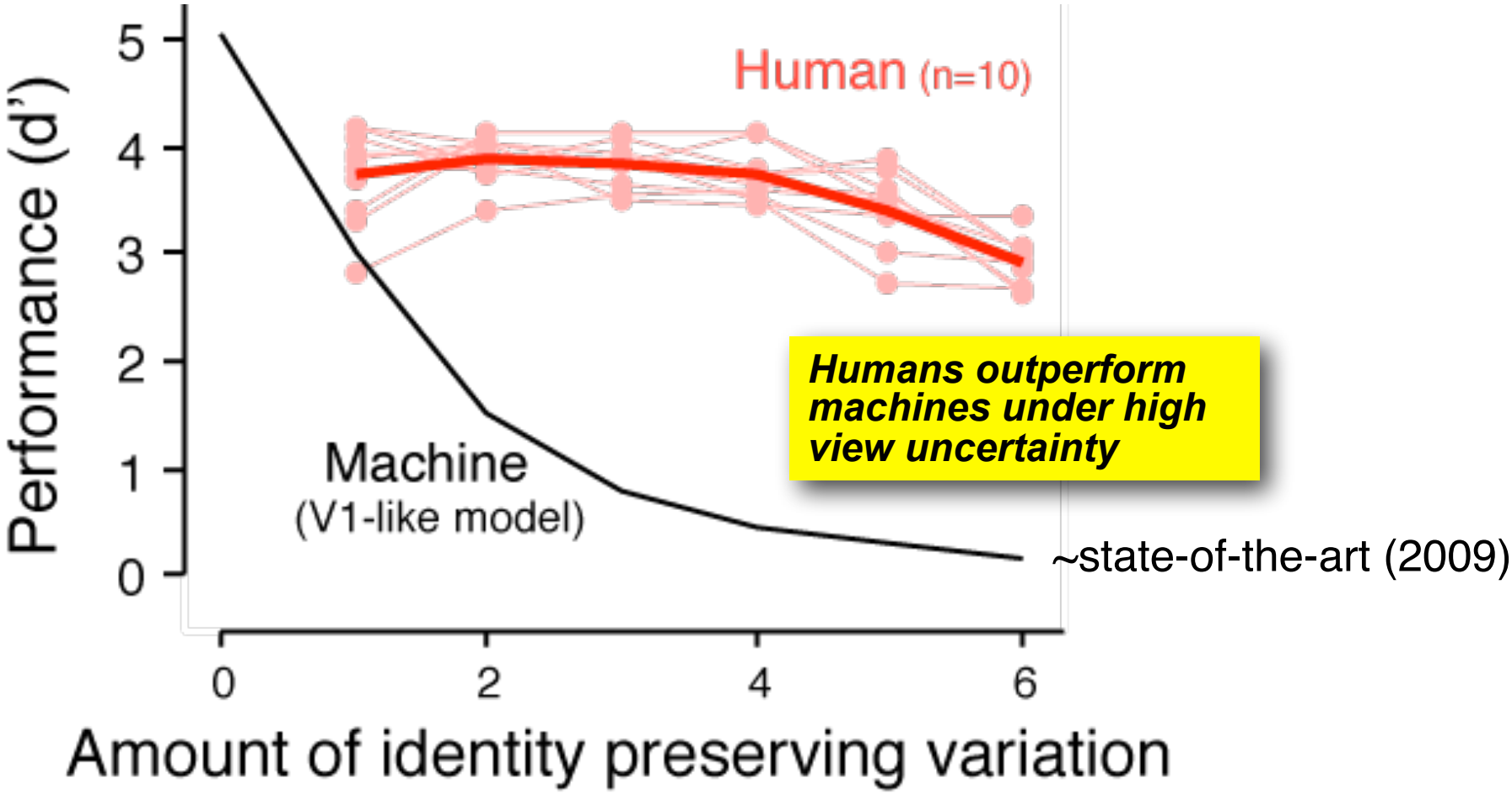
Clutter, occlusion



Poggio, Ullman, Grossberg, Edleman, Biederman, etc.

DiCarlo and Cox, ***TICS*** *(2007), Pinto, Cox, and DiCarlo,* ***PLoS Comp Bio*** *(2008), DiCarlo, Zoccolan and Rust,* ***Neuron*** *(2012)*

Q1 2009: Machines vs. humans on our benchmarks

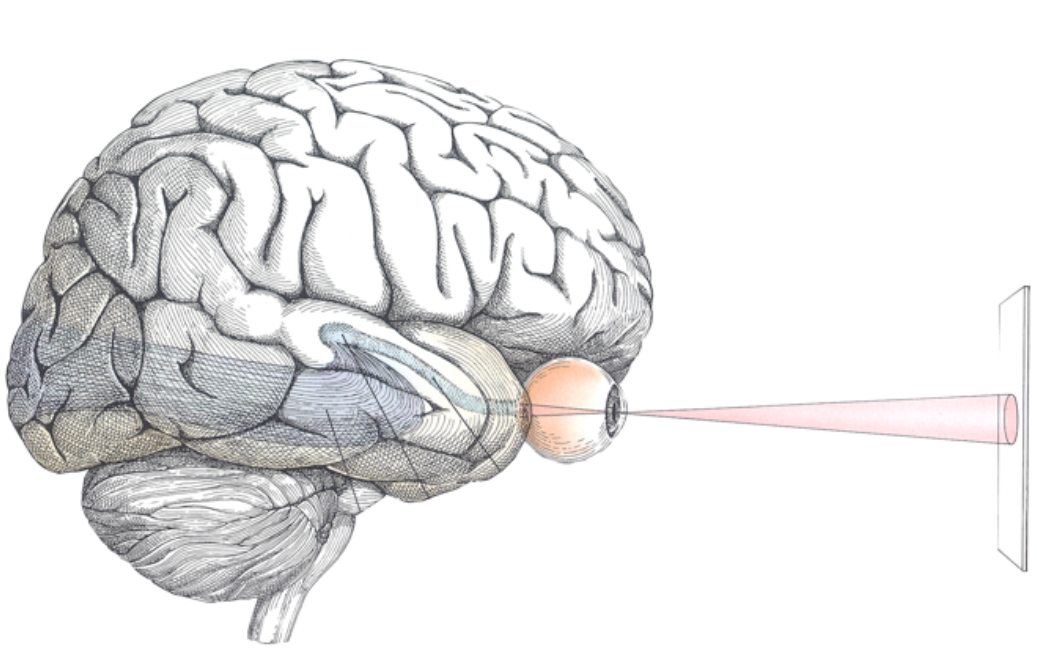


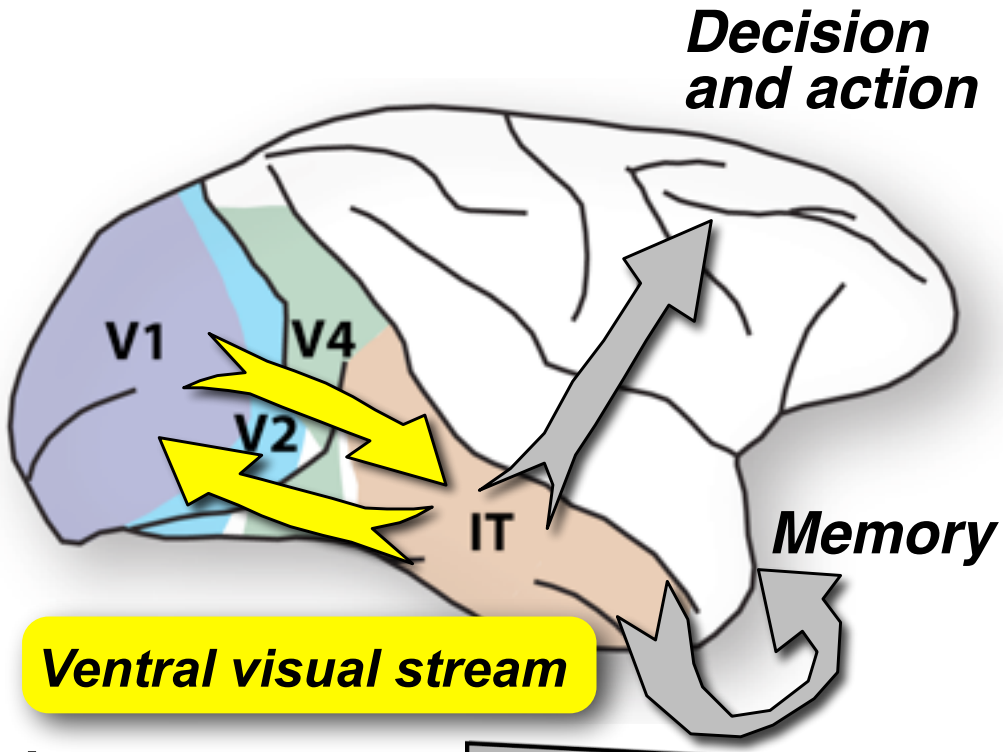
Humans outperform machines under high view uncertainty

~state-of-the-art (2009)

Data merged here: 48 basic-level tasks (8 labels x 6 level of variation)

Q2 Decision to gather data from the non-human primate



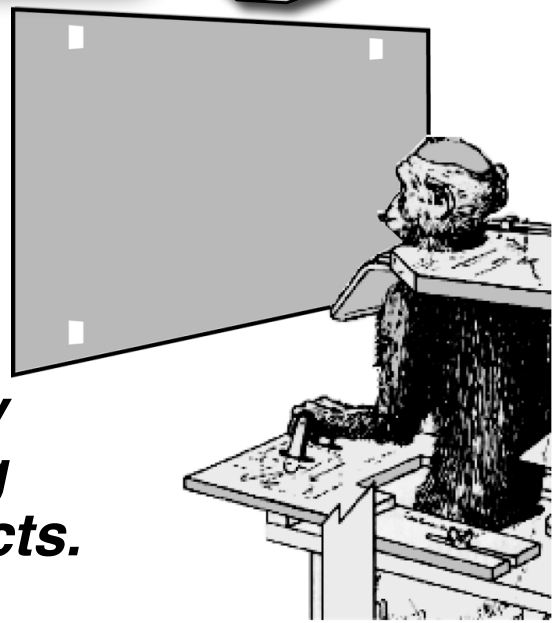


We know which brain regions house the most critical computations.

We know the system anatomy at a course grain.

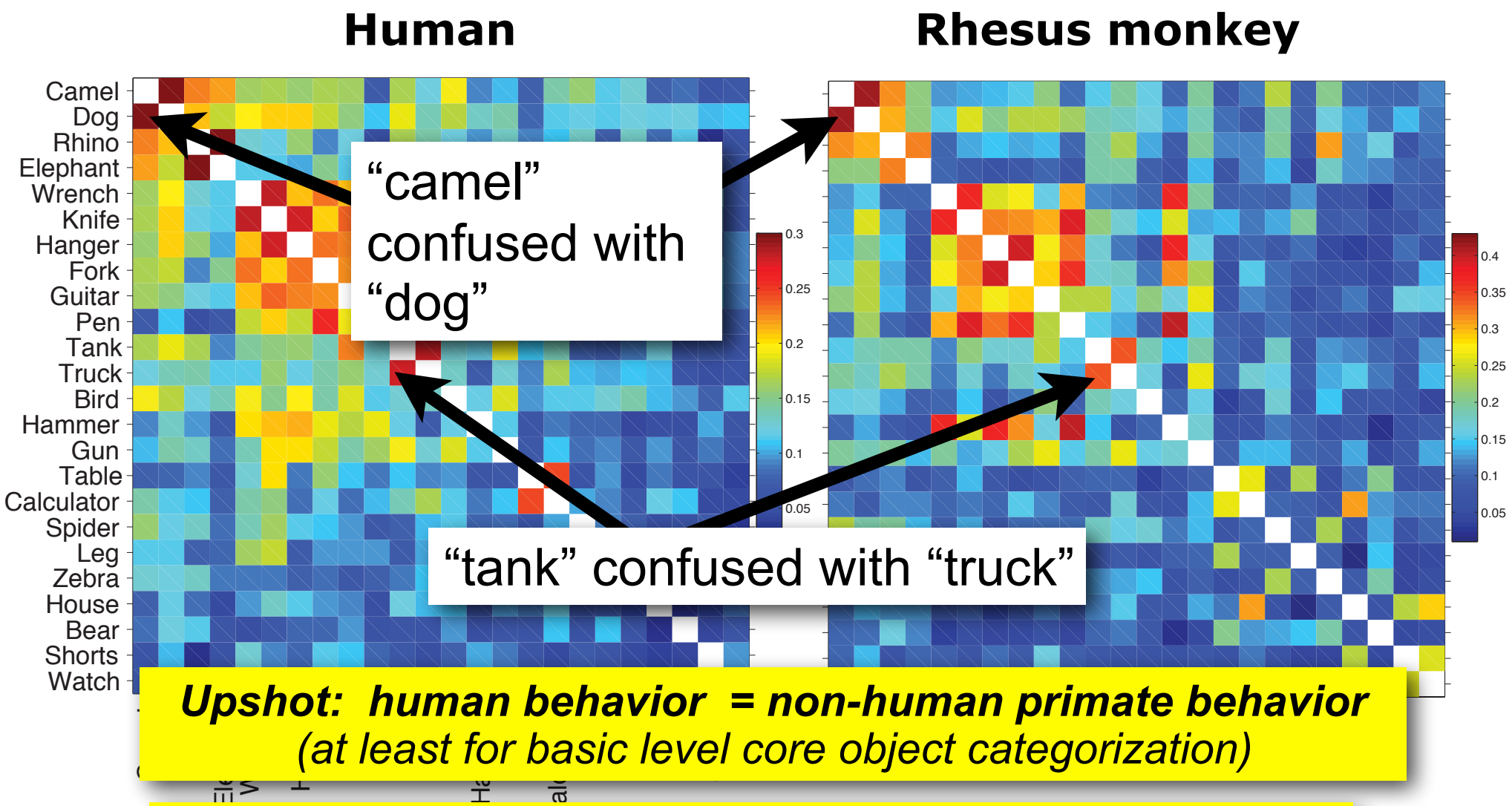
We have models of the elemental computations.

We can systematically measure and manipulate every stage of the processing stream, at the level of spiking neural activity, at msec resolution, in behaving subjects.
(not currently possible in the human brain)



Adapted from Motter and Mountcastle 1981

Q2 Decision to gather data from the non-human primate



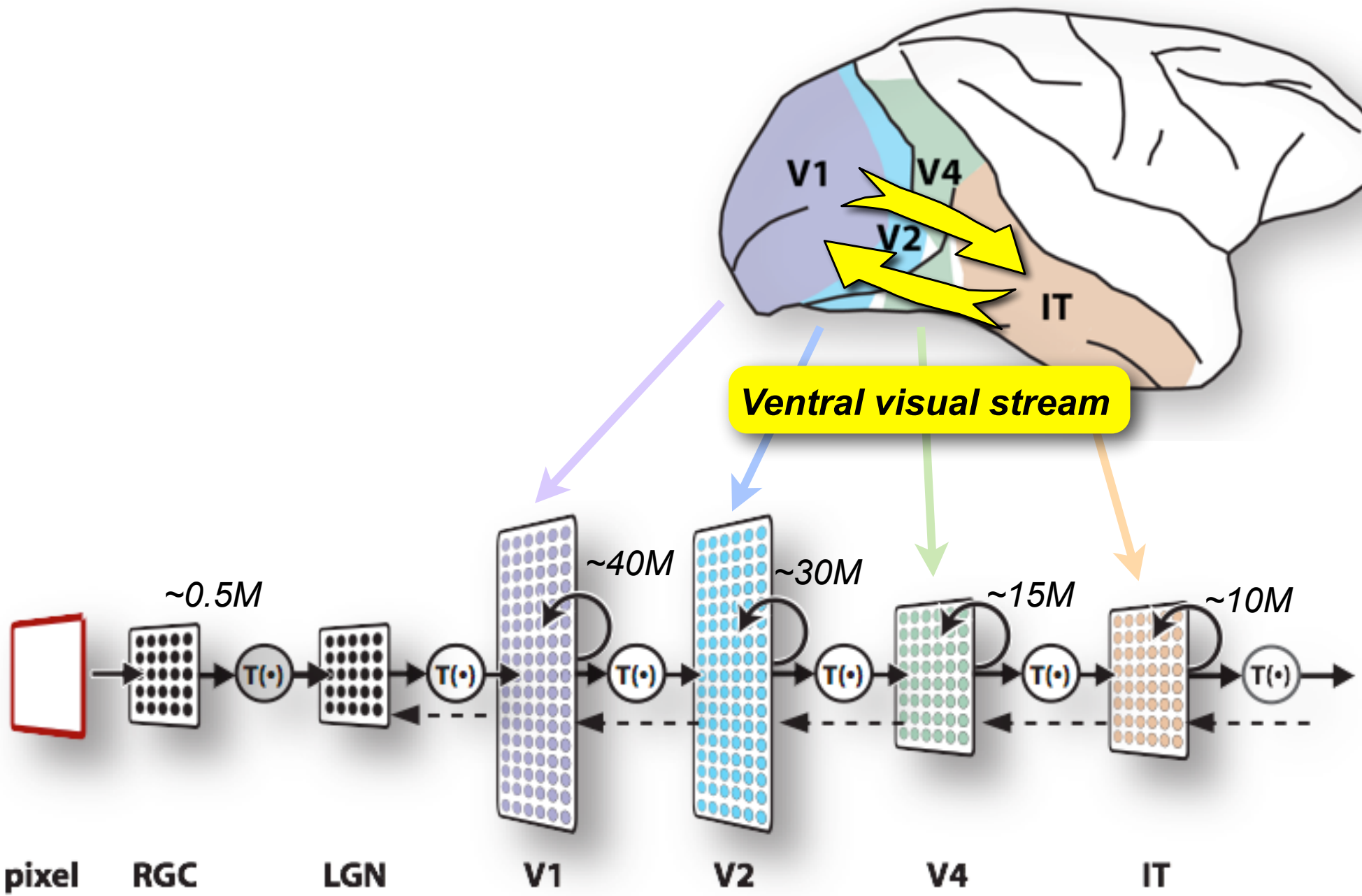
Upshot: human behavior = non-human primate behavior
(at least for basic level core object categorization)

One consequence: we may not need "principles", neurally-mapped models can suffice

Comparis
R. Rajaling
R. Rajalingham, K Schmidt, J.J. DiCarlo, ***J. Neuroscience*** *(2015)*

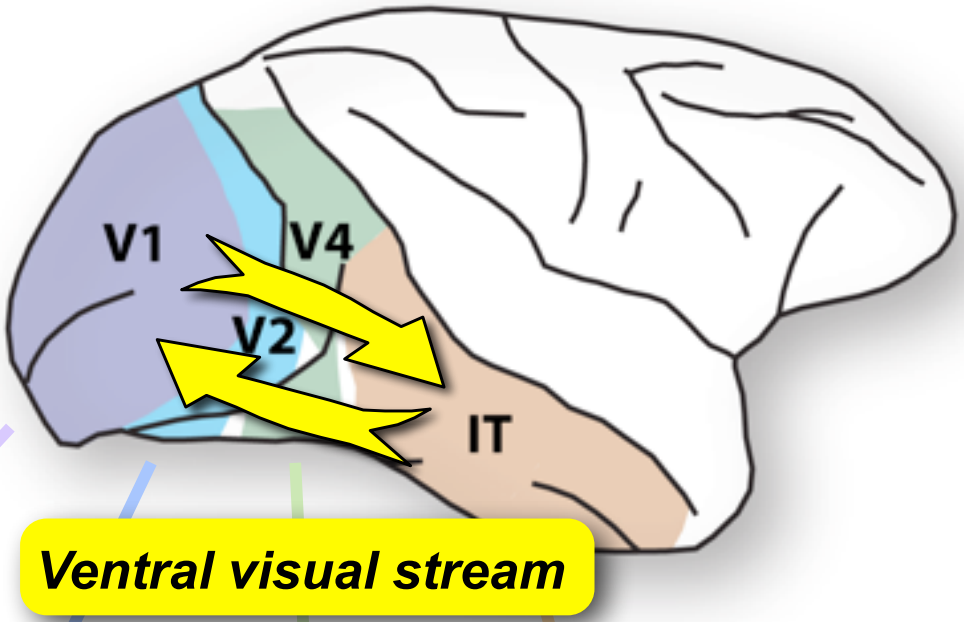
Adapted from Motter and Mountcastle 1981

Q2 Constraint data from the non-human primate



Q2 Constraint data from the non-human primate





Ventral visual stream

~60ms

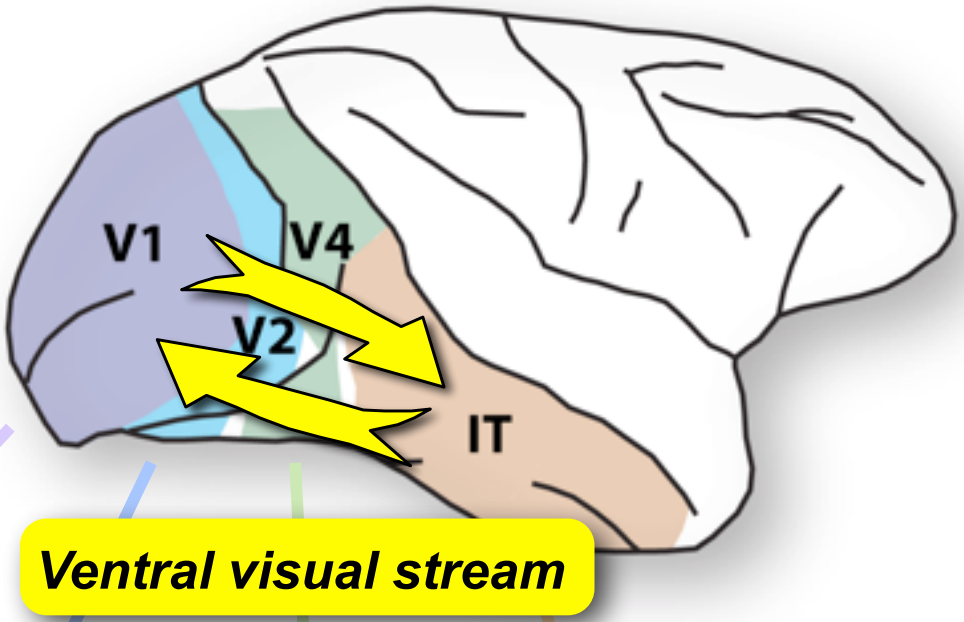
~100ms

T(•)

pixel RGC LGN V1 V2 V4 IT

Q2 Constraint data from the non-human primate

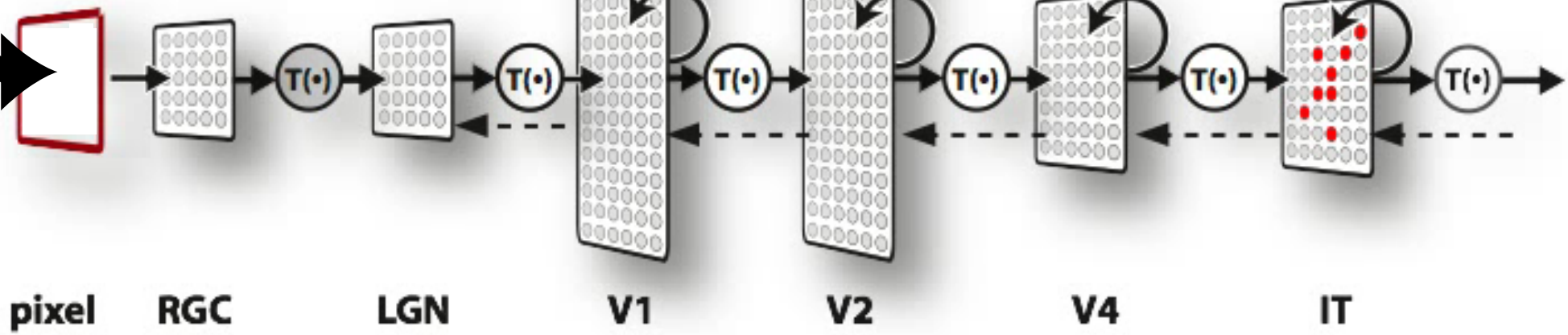




Ventral visual stream

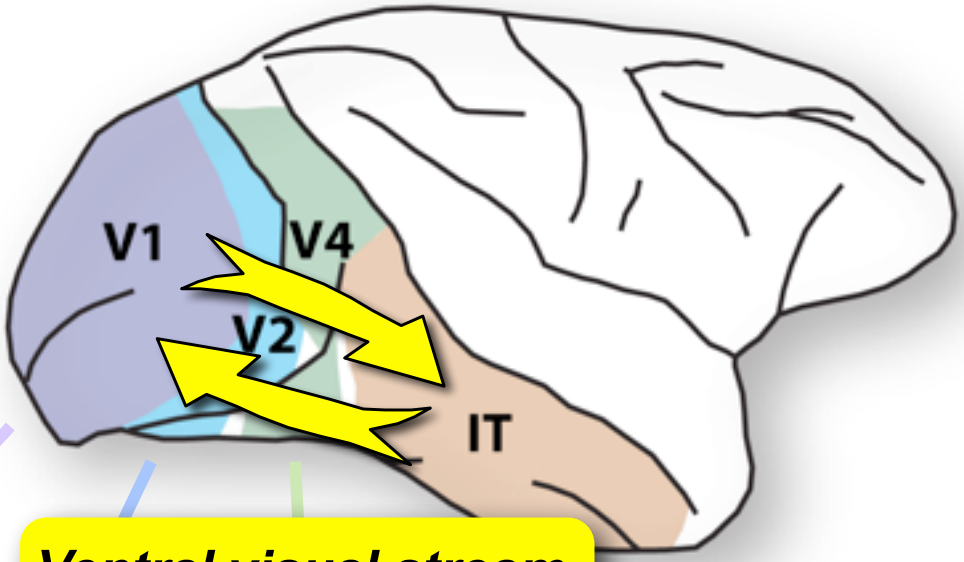
~60ms

~100ms



Q2 Constraint data from the non-human primate

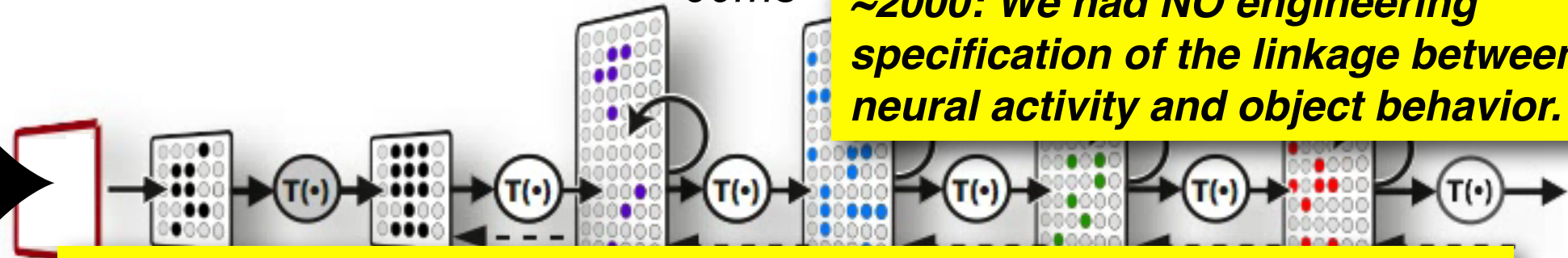




Ventral visual stream

~60ms

~2000: We had NO engineering specification of the linkage between neural activity and object behavior.



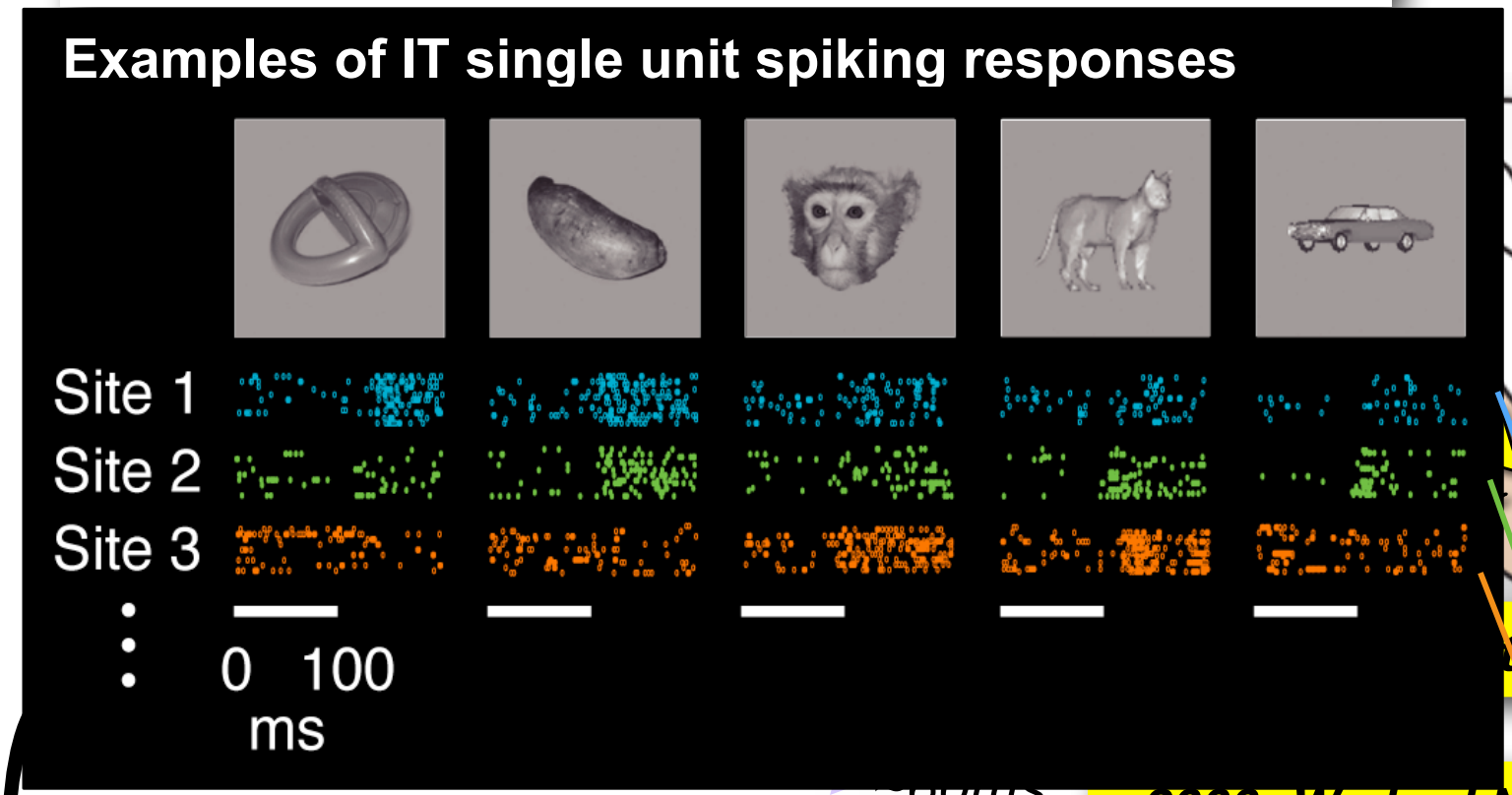
~2000: we had (almost) NO engineering specification of the neural population "image" transformations at work along this pathway

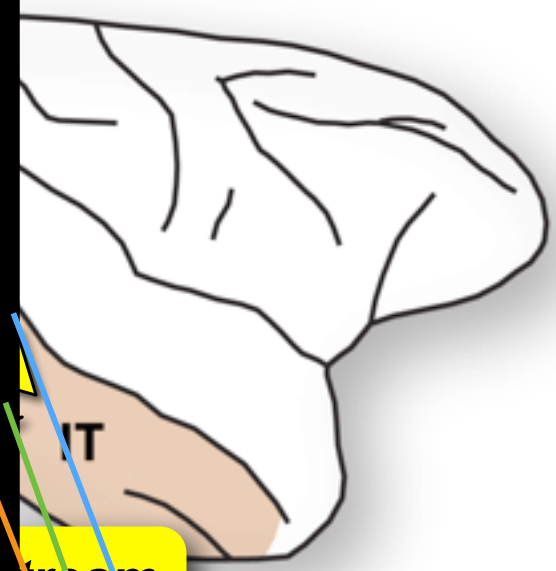
pixel RGC LGN V1 V2 V4 IT

~2005: we discovered that population of IT features (aka "neurons") was much more powerful than state-of-the-art computer vision features

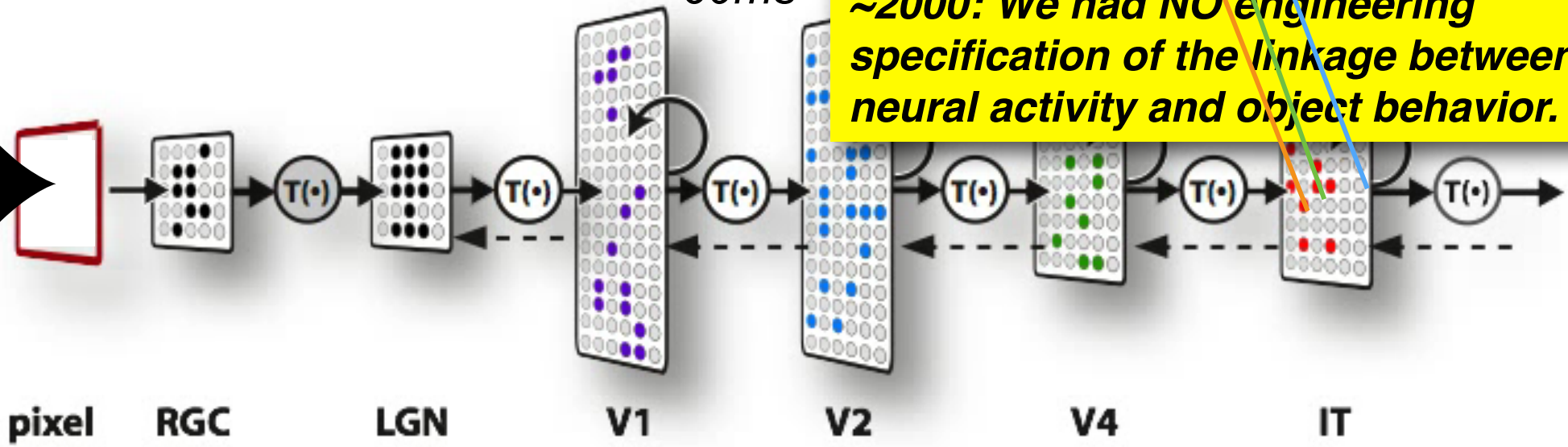
primate

Examples of IT single unit spiking responses

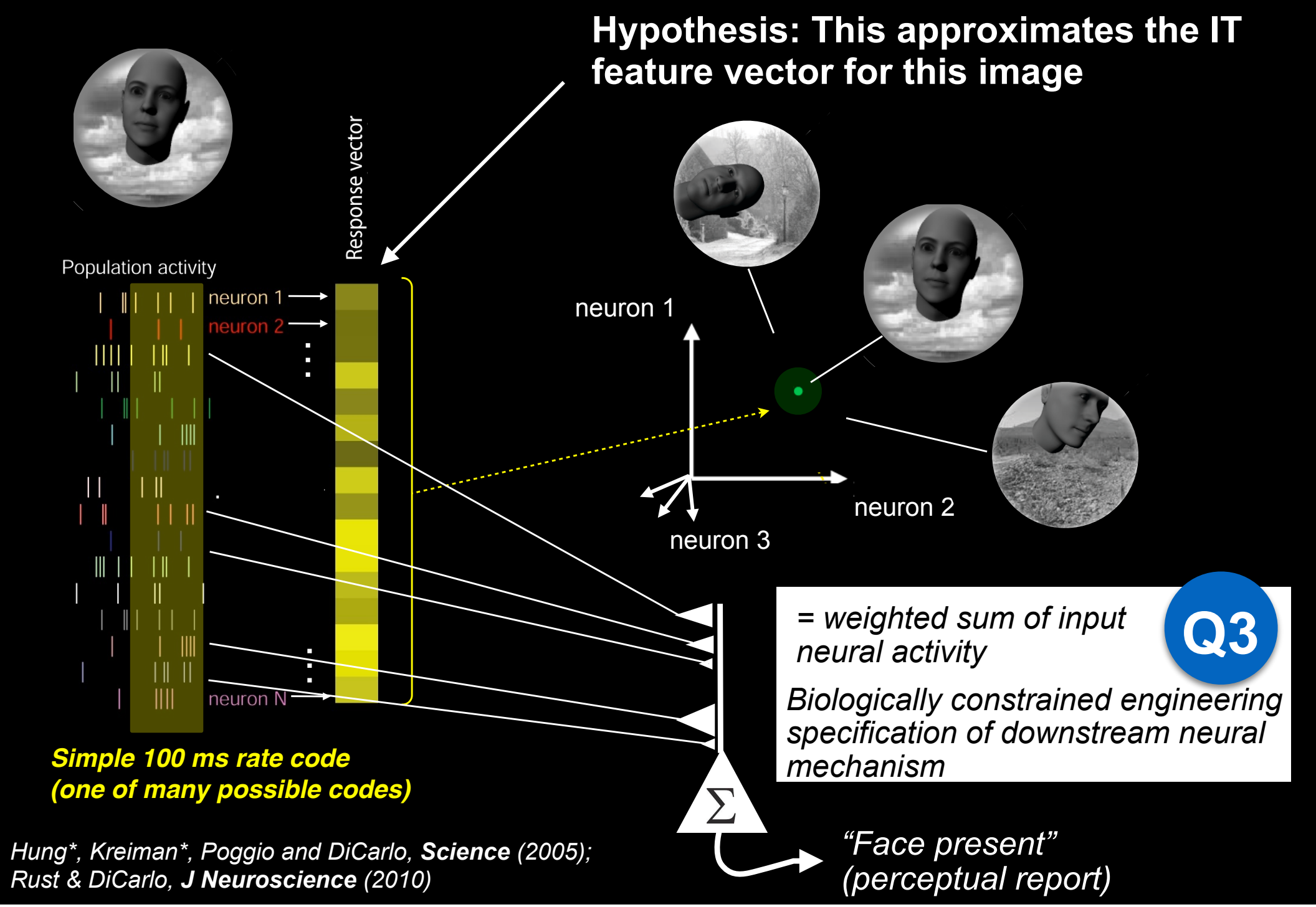




~2000: We had NO engineering specification of the linkage between neural activity and object behavior.

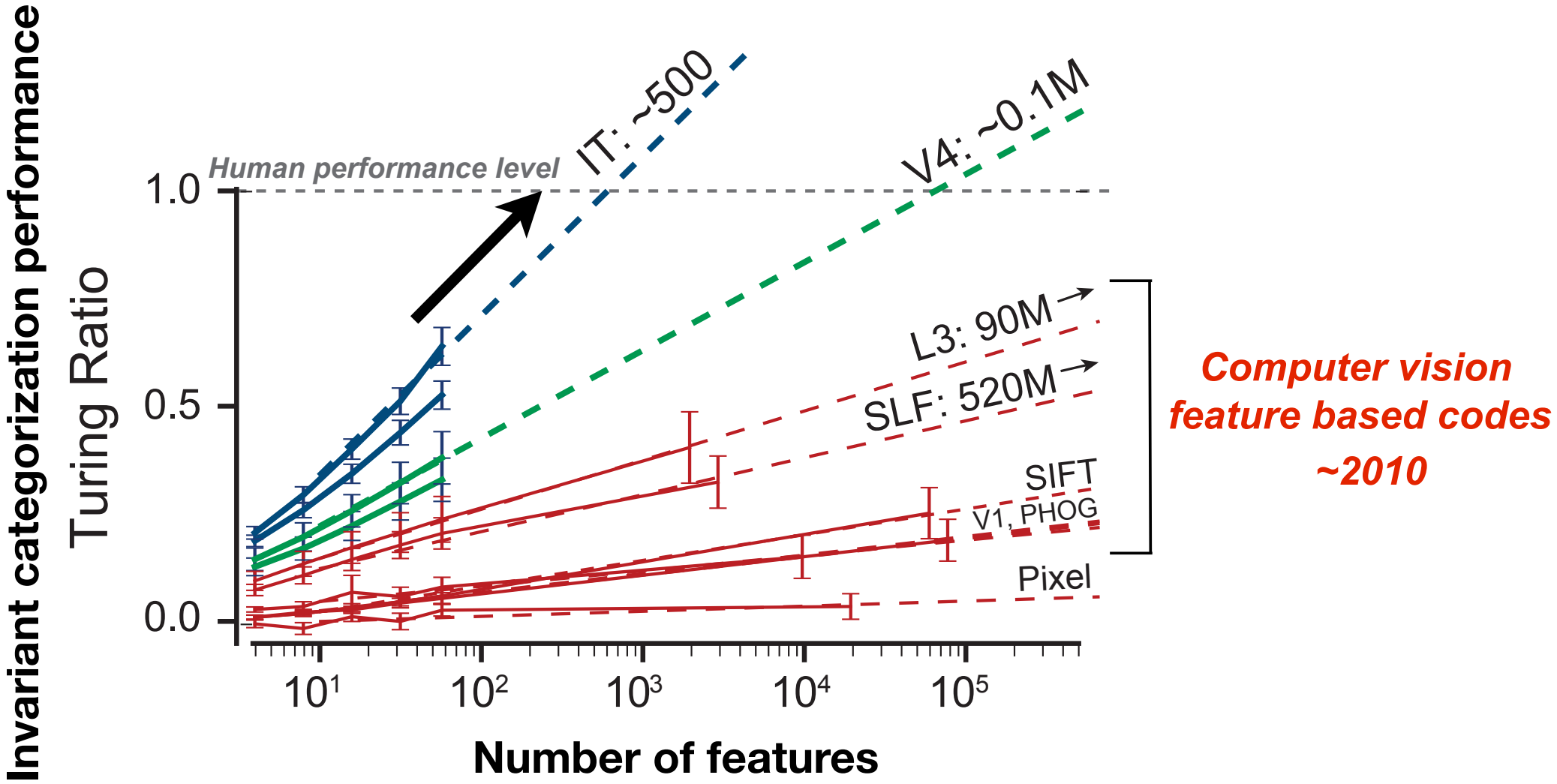


Upshot: learned weighted sums of IT features achieve high performance.

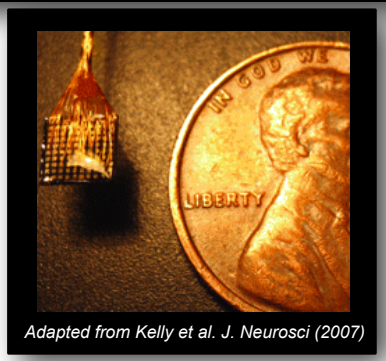


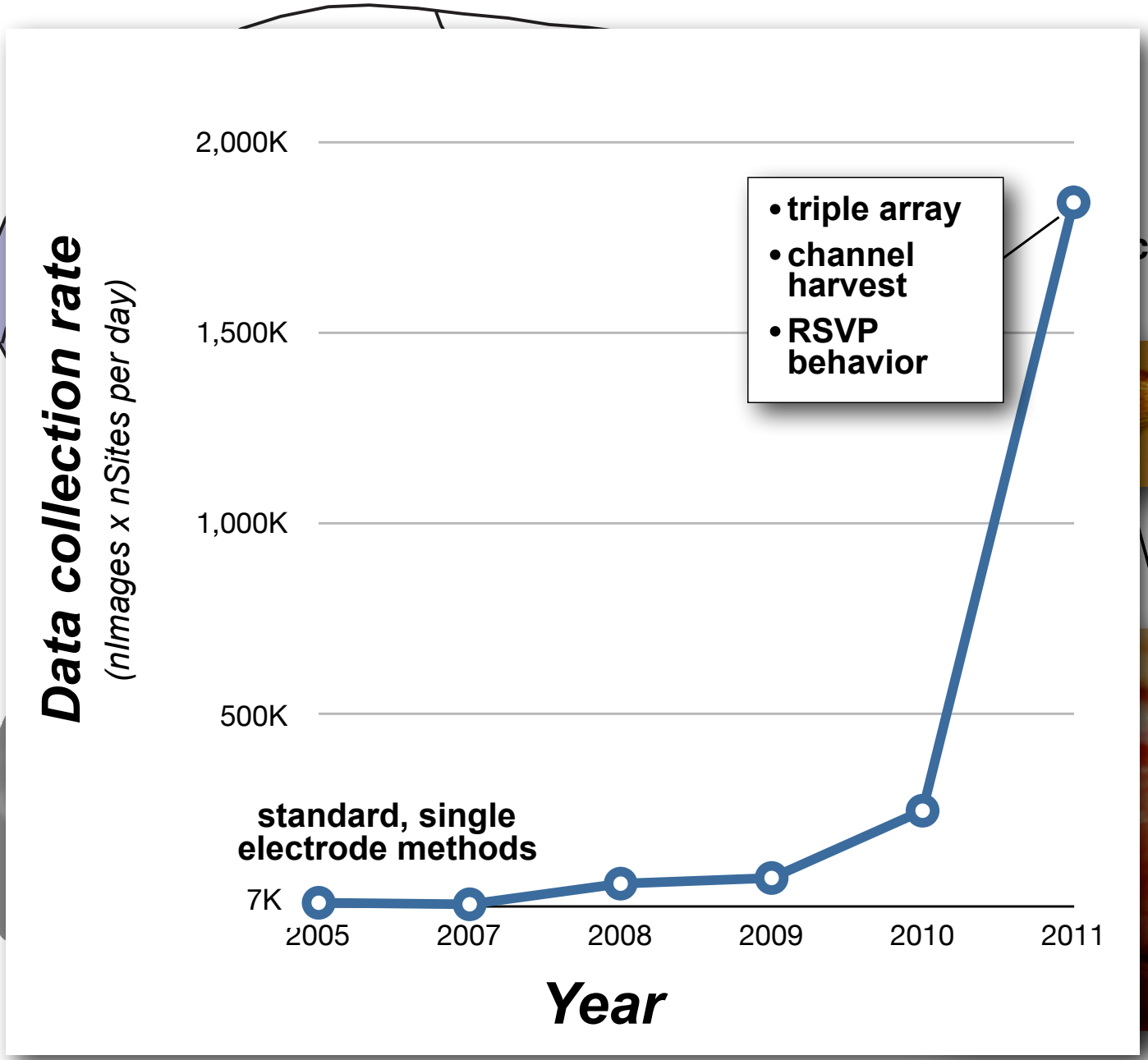
Hung, Kreiman*, Poggio and DiCarlo,* ***Science*** *(2005);*
Rust & DiCarlo, ***J Neuroscience*** *(2010)*

In the language of machine learning: the IT neural population was (and still is) a remarkably powerful set of features.

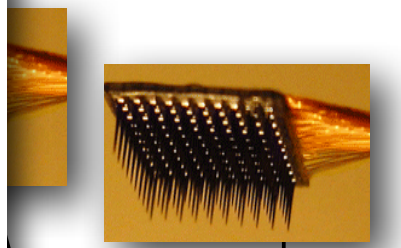


Methods advance: large scale neuronal recording along the ventral stream

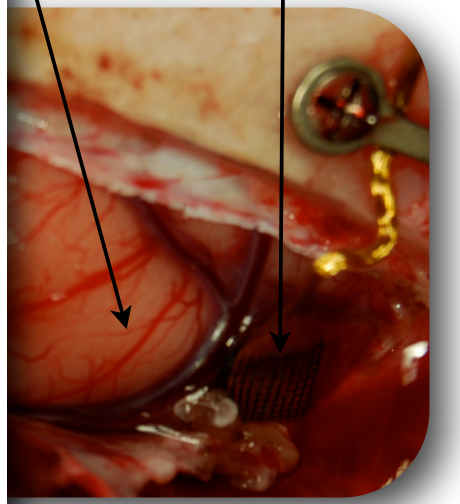




ctrode arrays



Array 3
(in place)



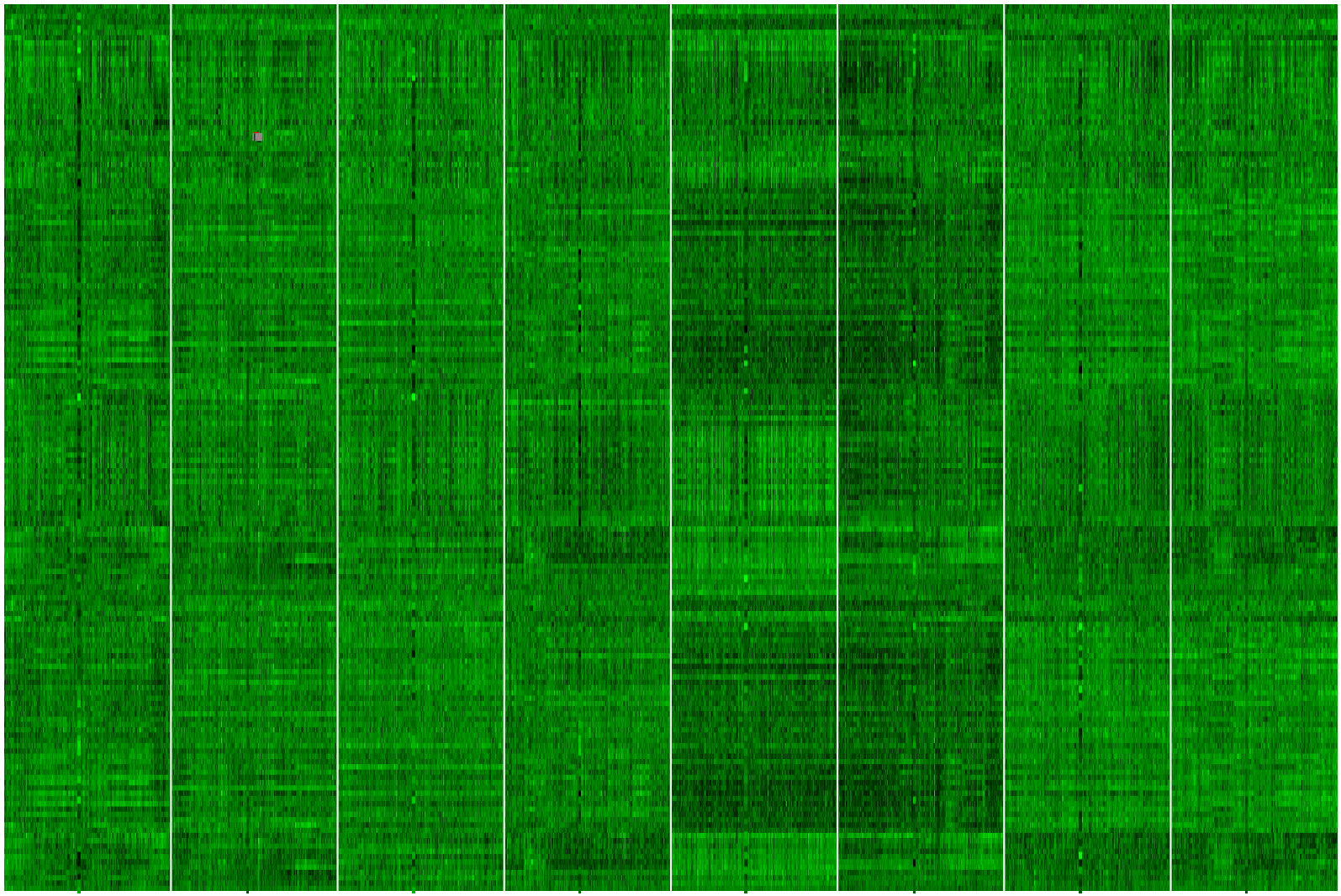


1

IT Neuron #

100-1000



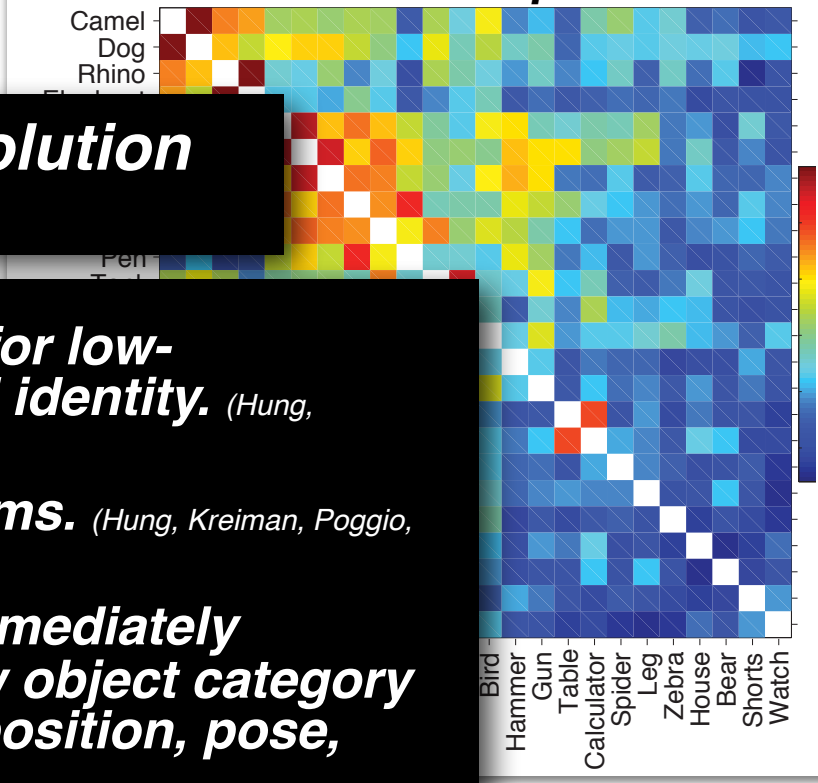


(All at high SNR: ~50 repetitions)

From IT features to performance (behavior). This is easy!

IT neural population patterns

Performance patterns



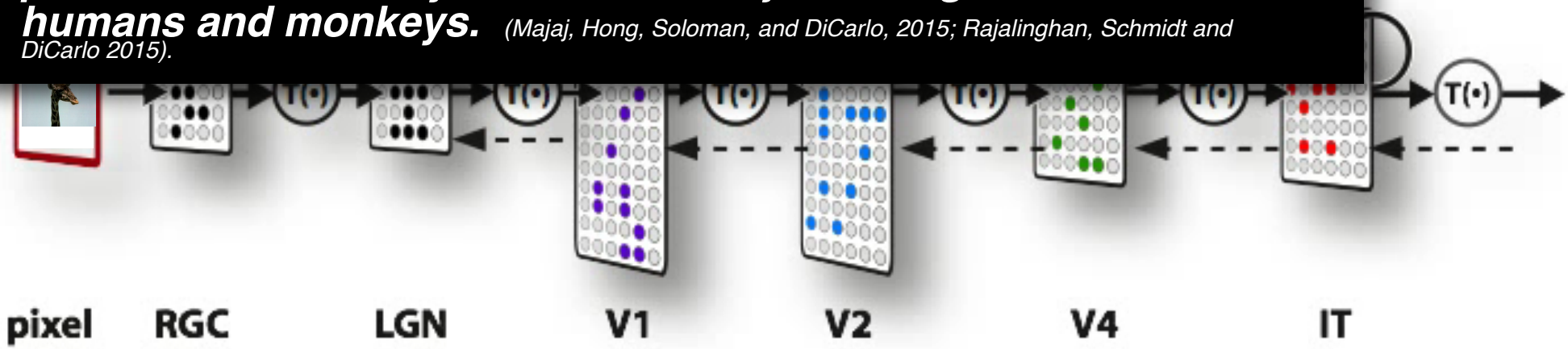
The IT feature set convey's biology's solution to intelligent object sensing.

The IT feature set is a high performance basis for low-dimensional, linear read of object category and identity. *(Hung, Kreiman, Poggio, DiCarlo 2005; Rust and DiCarlo 2010)*

The IT feature set is computed in less that 200 ms. *(Hung, Kreiman, Poggio, DiCarlo 2005; Rust and DiCarlo 2010; Majaj, Hong, Soloman and DiCarlo, 2015)*

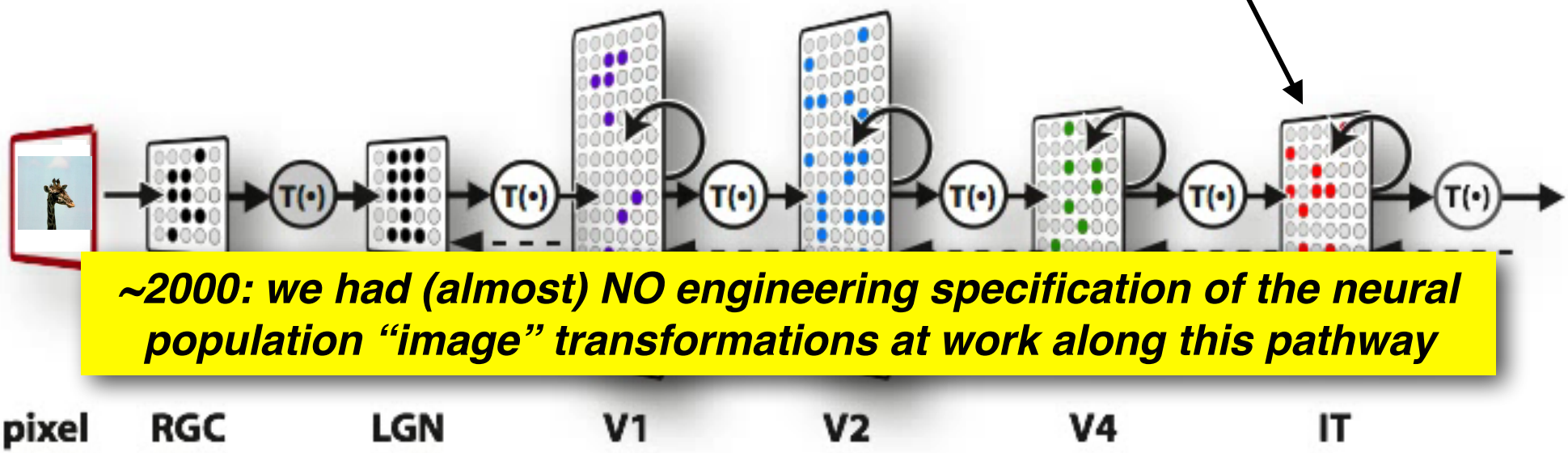
The IT feature set is a GENERAL BASIS — it immediately supports rapid, linear-read, learning of any new object category and identity. AND it supports report of object position, pose, scale, etc. *(Hong, Yamins, Majaj, DiCarlo 2016)*

An engineering spec of the above (LAWS of RAD IT) accurately predicts the difficulty of all tested object recognition tasks in humans and monkeys. *(Majaj, Hong, Soloman, and DiCarlo, 2015; Rajalinghan, Schmidt and DiCarlo 2015).*



The IT feature set convey's biology's solution to intelligent object sensing.

Q3 *Forward engineering within biological constraints.*



~2000: we had (almost) NO engineering specification of the neural population "image" transformations at work along this pathway

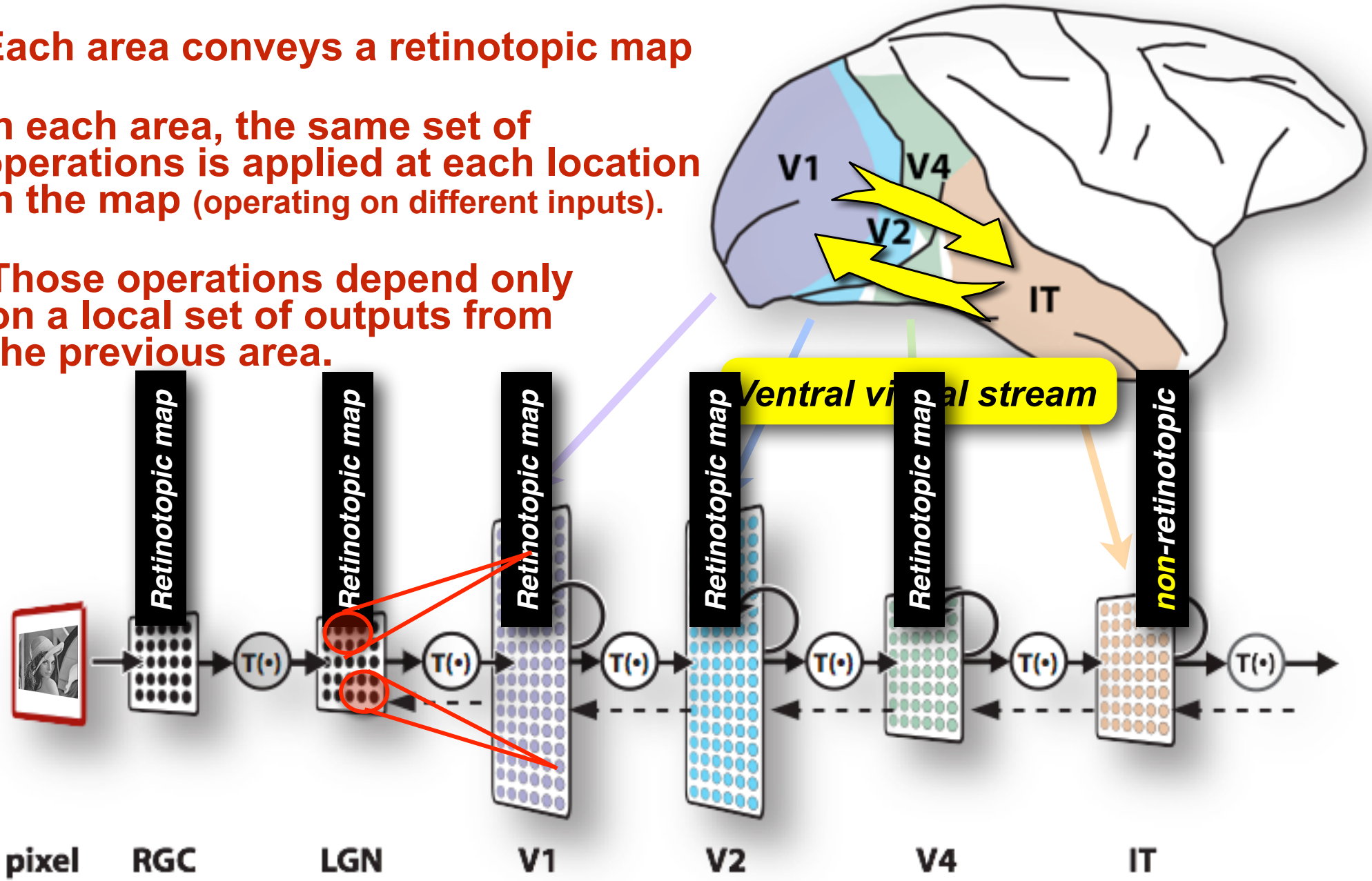
Q3 *Forward engineering (model building) within known constraints on the ventral stream*

Ventral stream is a "deep" stack of areas

Each area conveys a retinotopic map

In each area, the same set of operations is applied at each location in the map (operating on different inputs).

Those operations depend only on a local set of outputs from the previous area.



Q3 *Forward engineering (model building) within known constraints on the ventral stream*

Ventral stream is a "deep" stack of areas

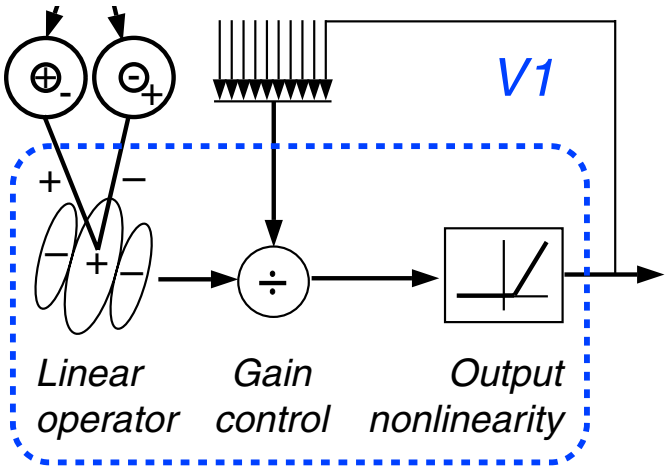
Each area conveys a retinotopic map

In each area, the same set of operations is applied at each location in the map (operating on different inputs).

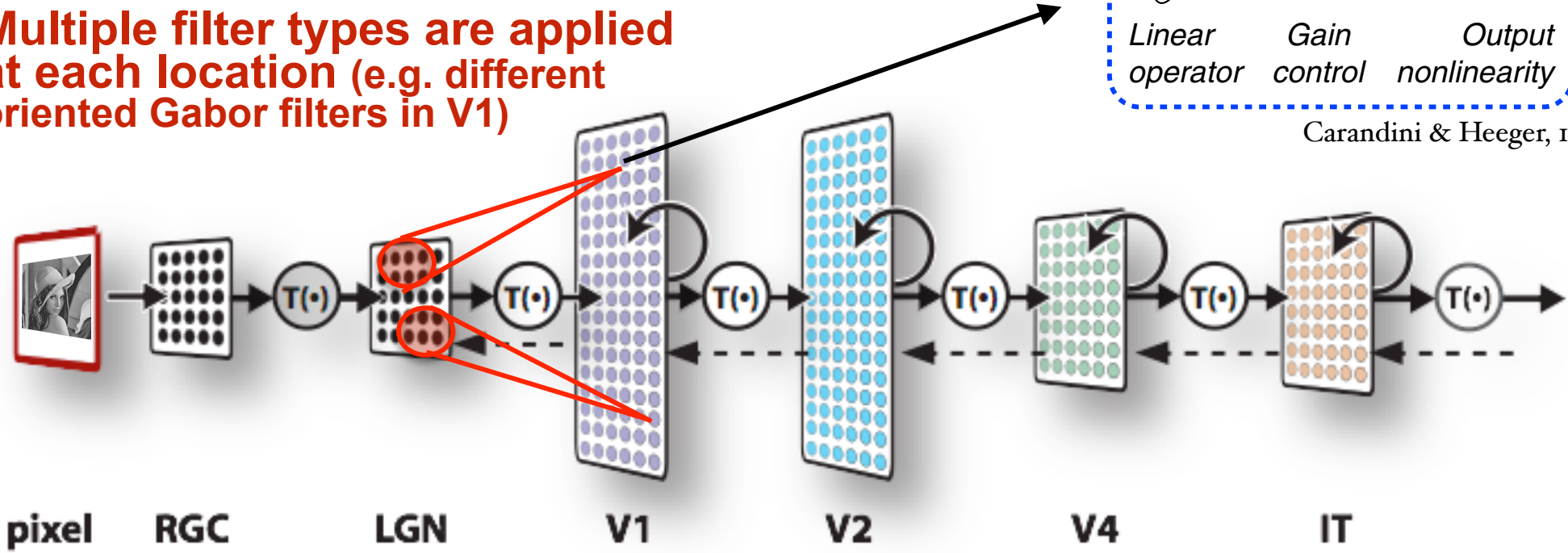
Those operations depend only on a local set of outputs from the previous area.

Multiple filter types are applied at each location (e.g. different oriented Gabor filters in V1)

Each neuron is well approximated as a linear filter + output non-linearity + pooling normalization



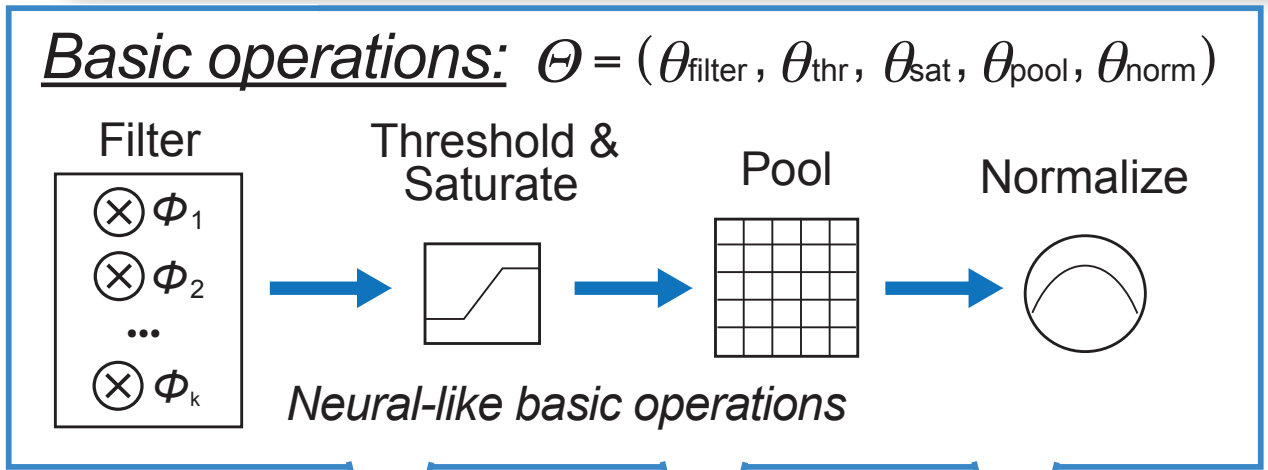
Carandini & Heeger, 1994



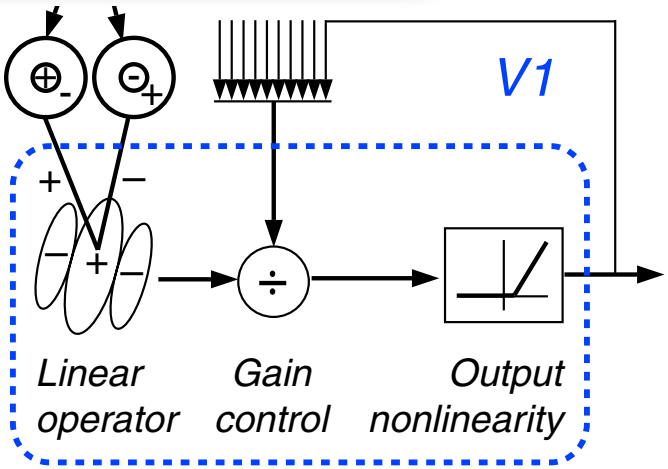
Q3 *Forward engineering (model building) within known constraints on the ventral stream*

Hubel & Wiesel (1962), Fukushima (1980); Perrett & Oram (1993); Wallis & Rolls (1997); LeCun et al. (1998); Riesenhuber & Poggio (1999); Serre, Kouh, et al. (2005), etc....

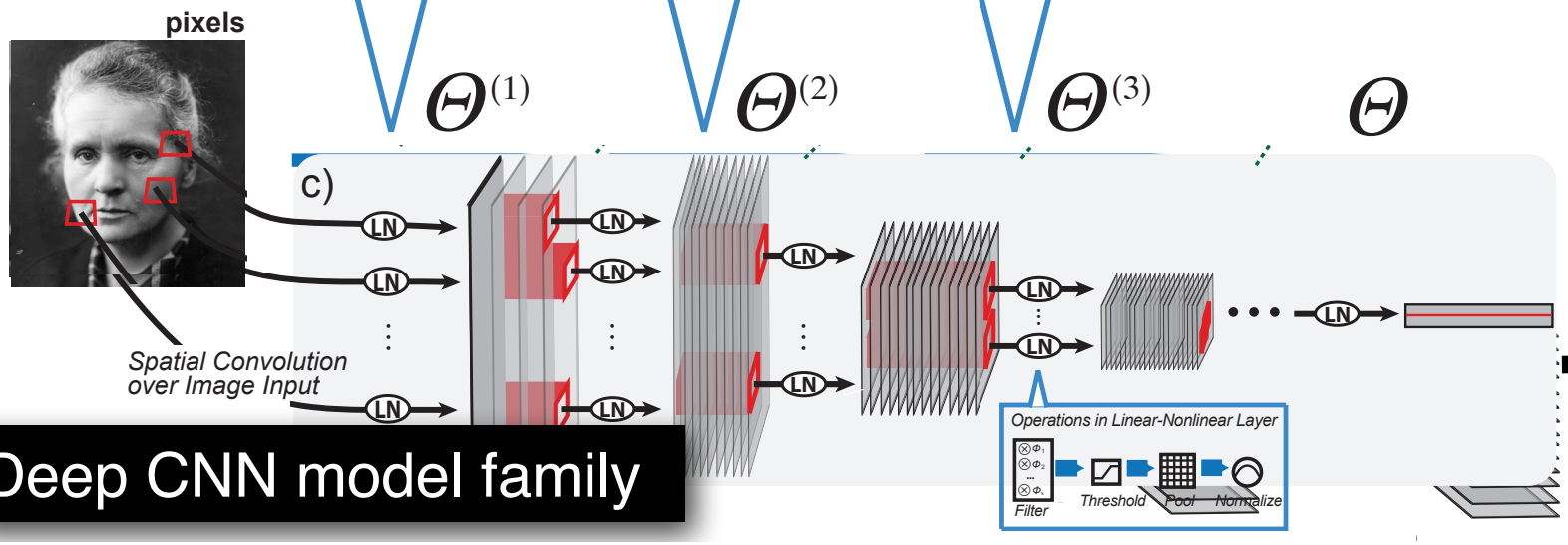
Today, this family of models is called "Deep convolutional neural networks" (Deep CNN's)



≅



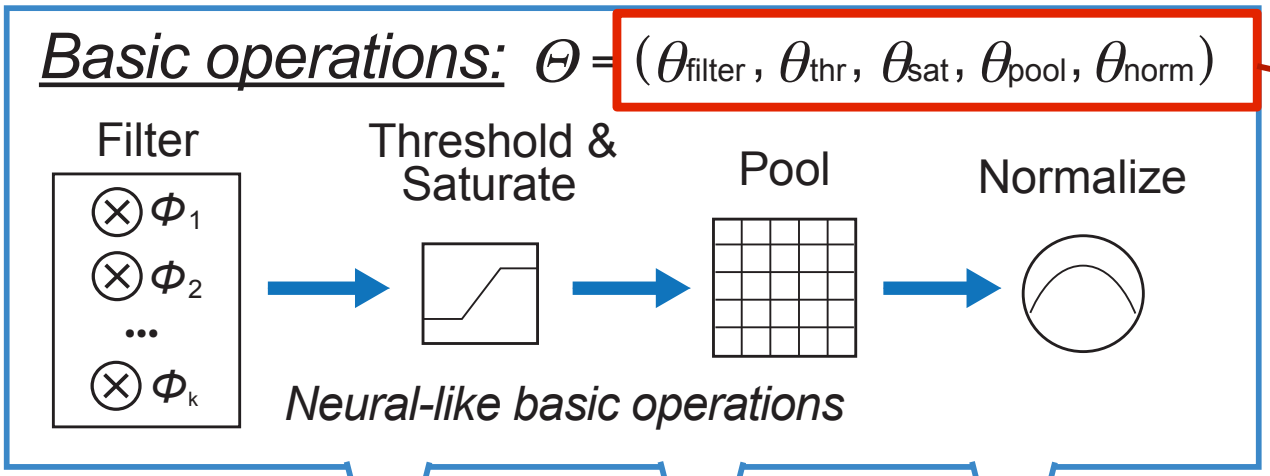
Carandini & Heeger, 1994



Deep CNN model family

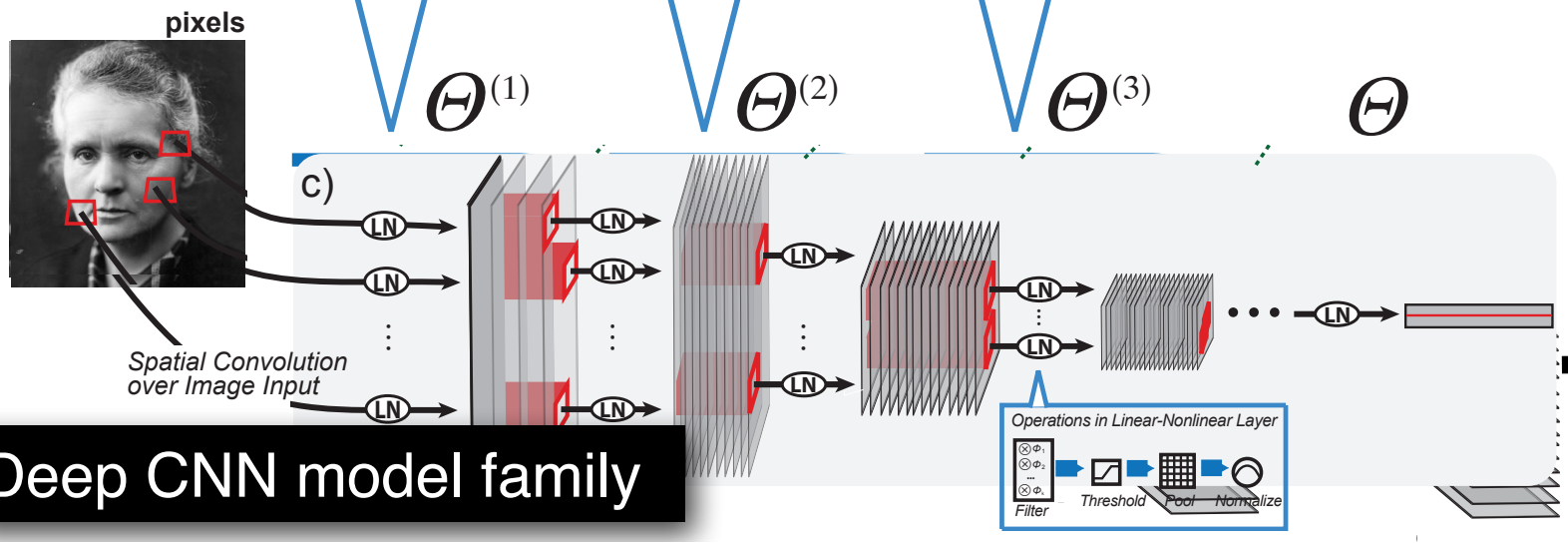
Top layer has thousands of visual "neurons" (aka "features")

Q3 *Forward engineering (model building) within known constraints on the ventral stream*



Problem: Thousands of unknown* parameters in each layer

(* not directly determined by neurobiology)



Top layer has thousands of visual "neurons" (aka "features")

Deep CNN model family

Q3 *Forward engineering (model building) within known constraints on the ventral stream*

How do we determine which model in the deep CNN family, if any, is the actual mechanism of the ventral stream?

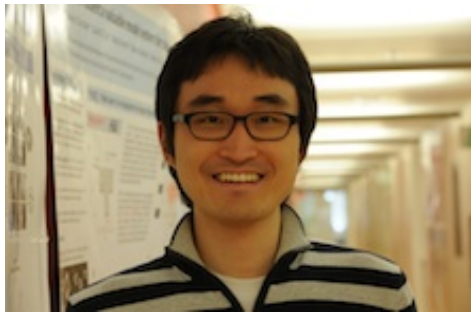
Strategy: Use selection methods to find specific models (i.e. parameter settings) in this model family.

What to select for? Models that are good at tasks that we hypothesize that the ventral stream evolved and/or developed to solve. Neuroscience data suggest: the task is "invariant" object recognition.

How to select? Biology does not yet tell us, so we used engineering optimization methods.



Dan Yamins



Ha Hong

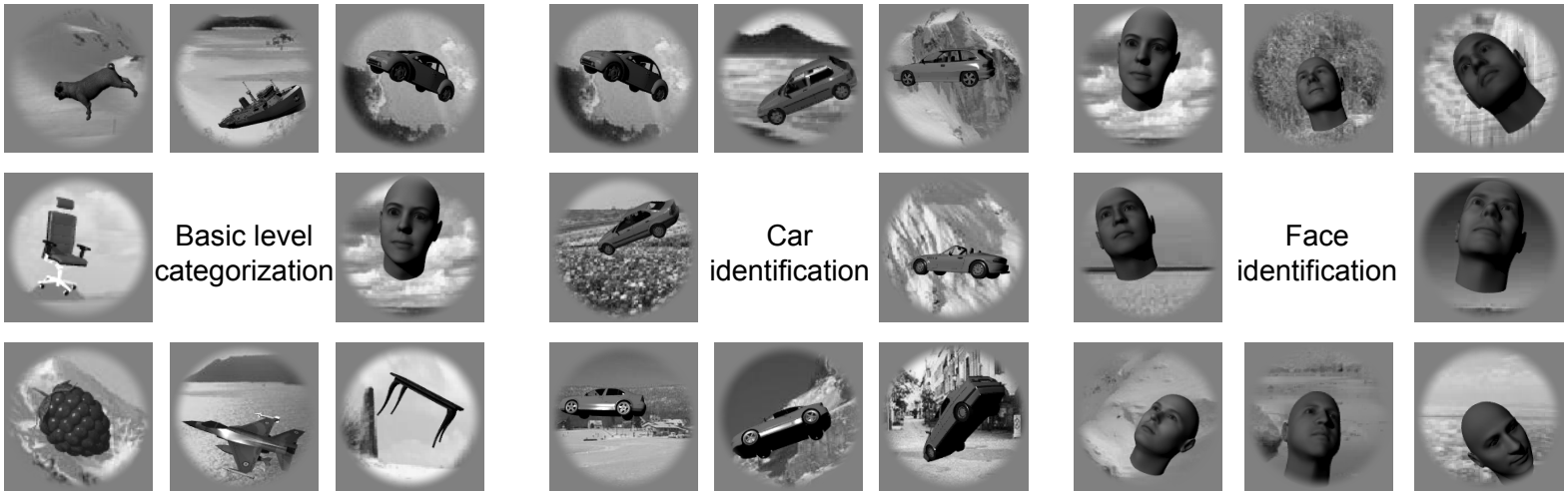
Deep CNN model family

Top layer has thousands of visual "neurons" (aka "features")

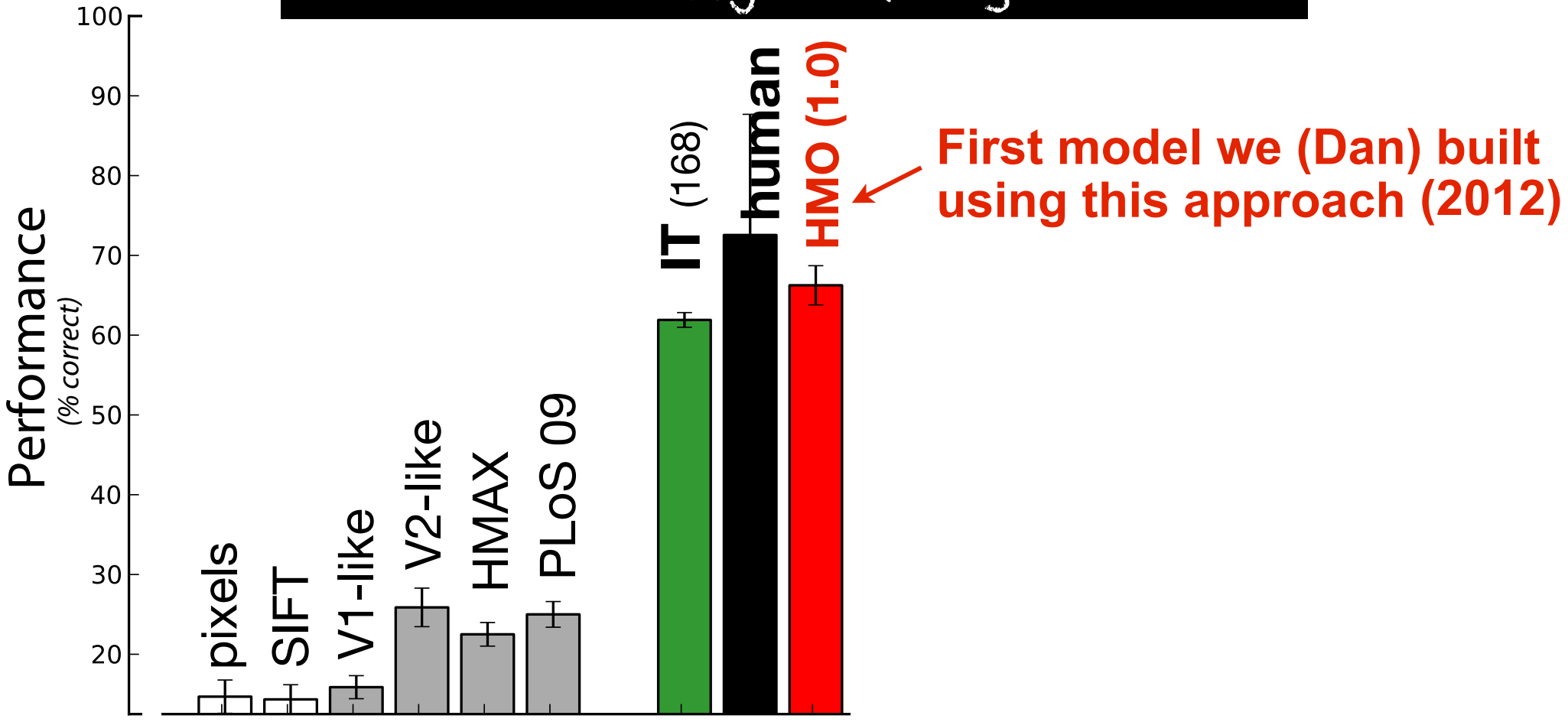
Problem:

ach layer

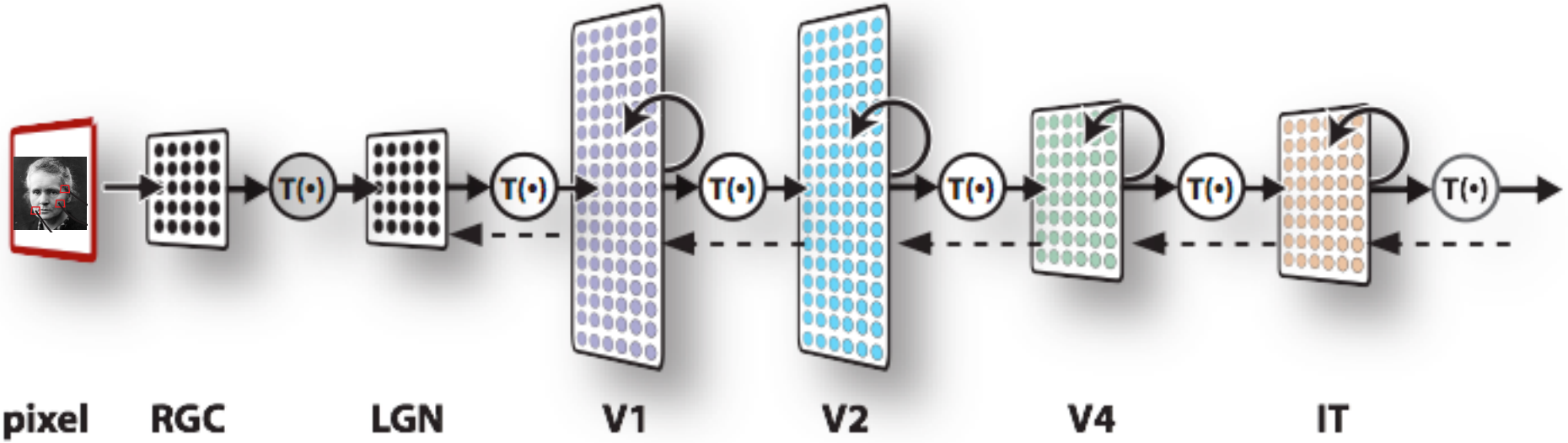
(* not directly determined by neurobiology)



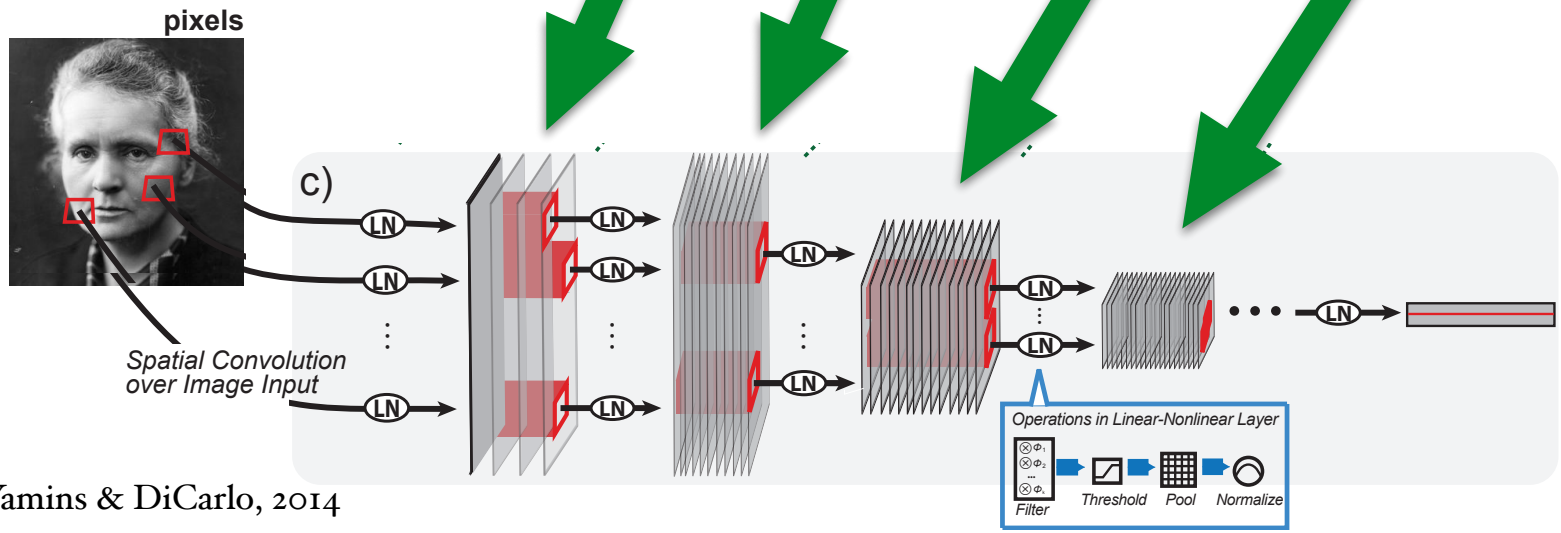
Test on Core Object Recognition 1.0



First model we (Dan) built using this approach (2012)

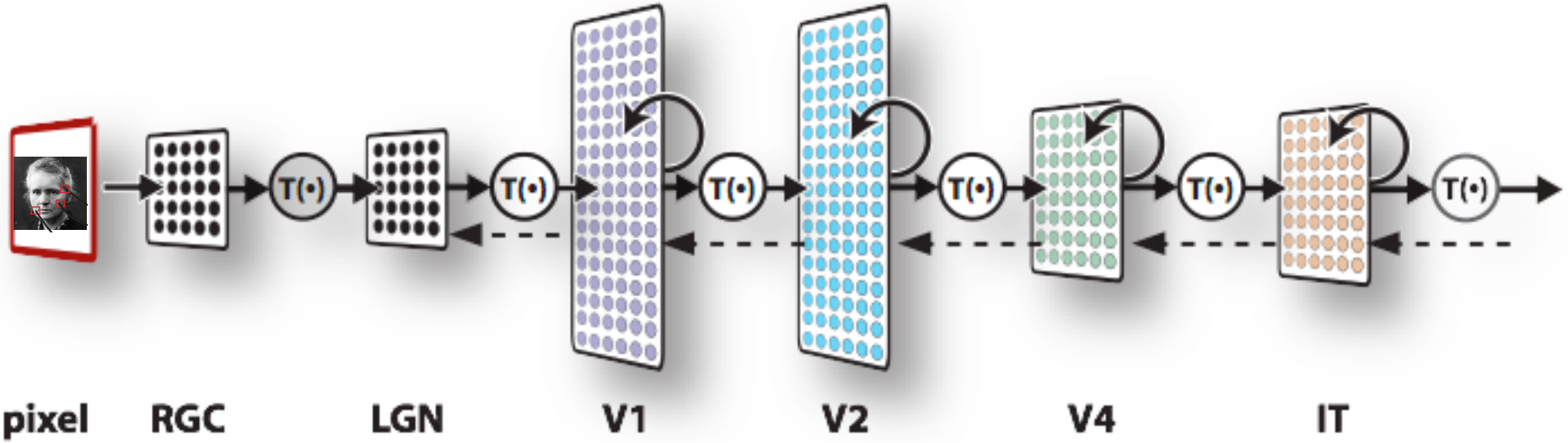


Do each of these match?



Top layer has thousands of visual "neurons" (aka "features")

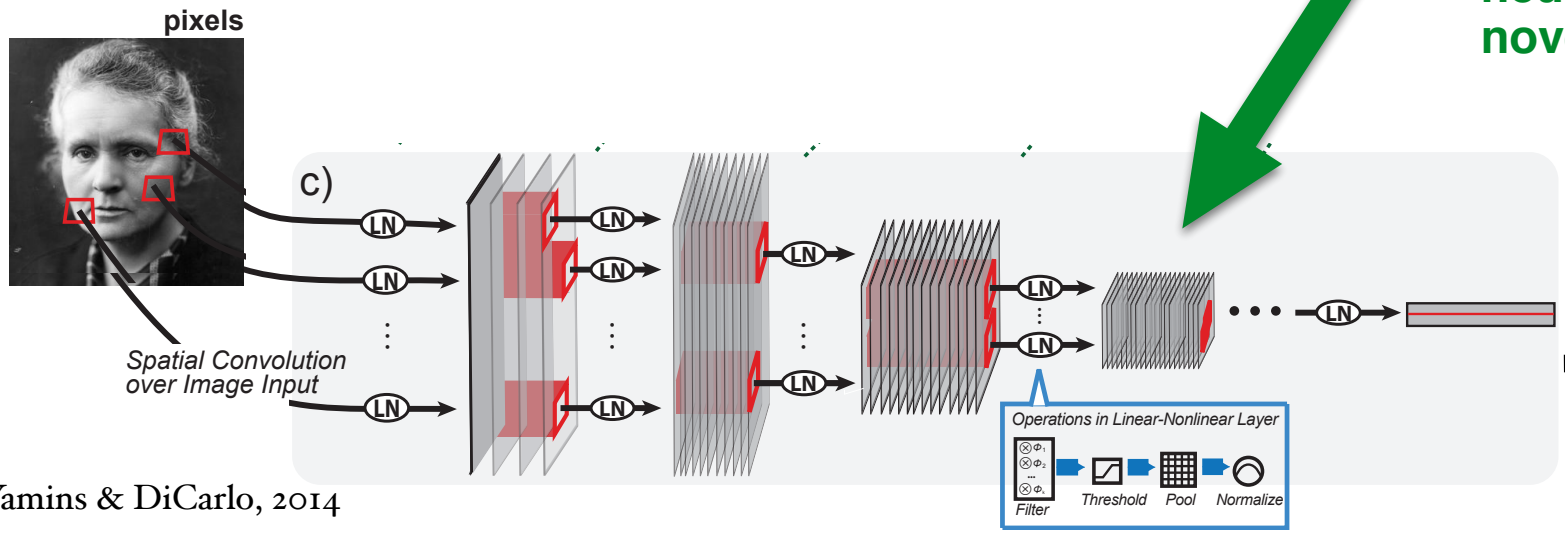
Yamins & DiCarlo, 2014



Metric 1: Single neuron predictivity:

For each IT neuron: Use some of its responses to fit (linear regression) using the model features as a basis set

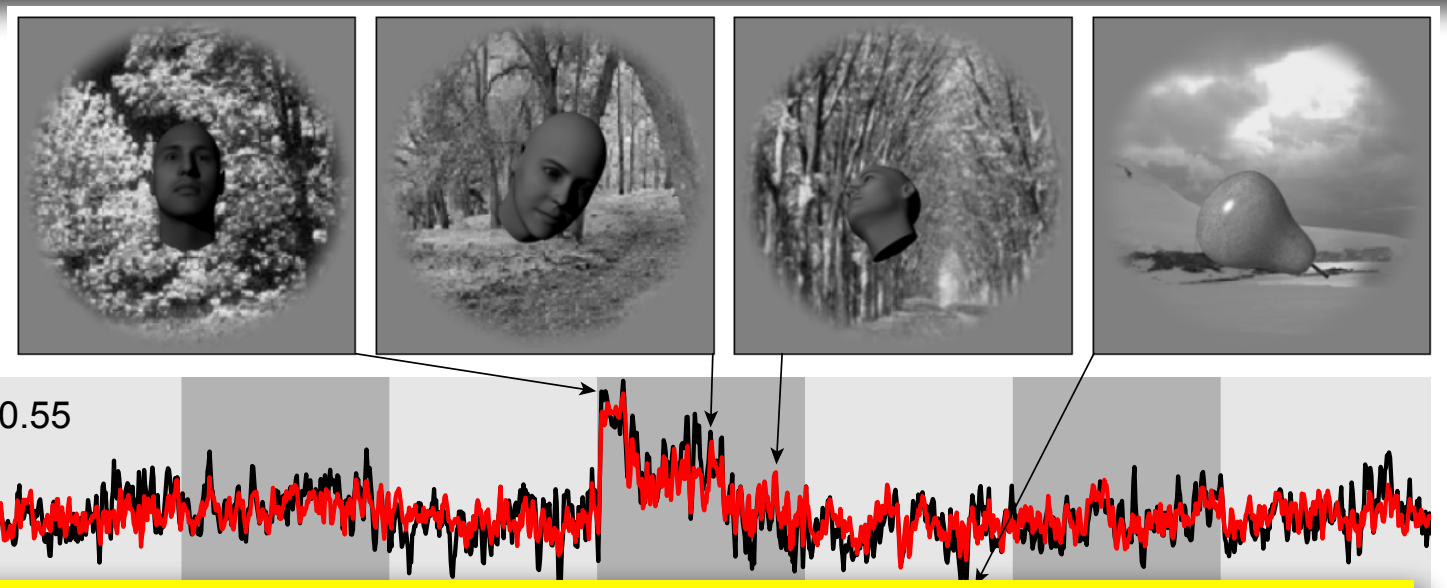
Ask: How well does the model predict the neuron's responses on novel images?



Top layer has thousands of visual "neurons" (aka "features")

Yamins & DiCarlo, 2014

Remarkably ability of these models to explain and predict IT features



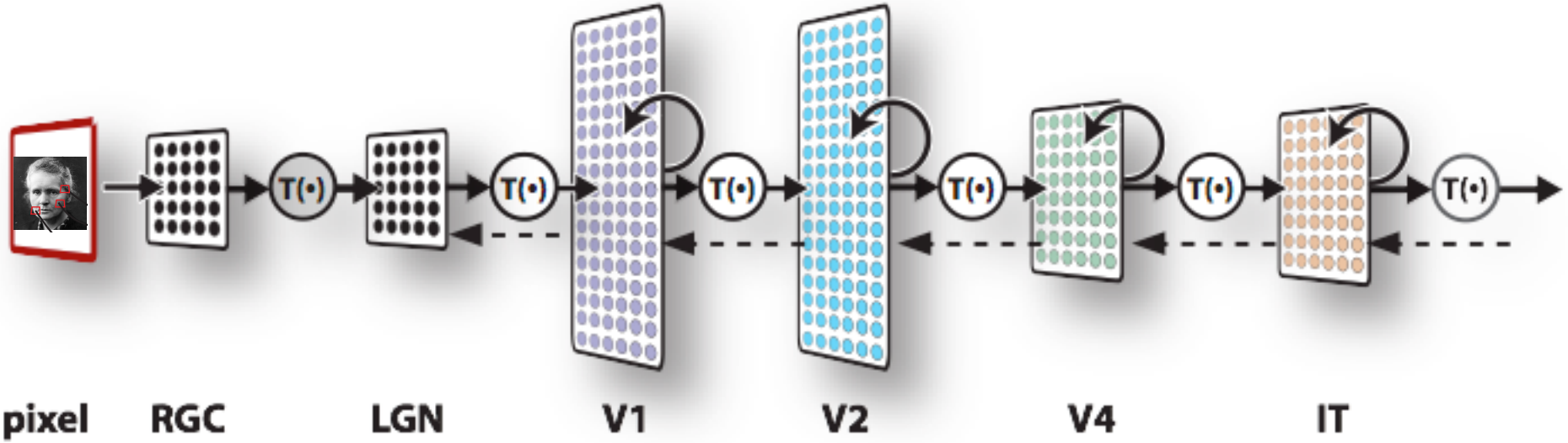
*These are **predictions** from a model feature set that did not see any of these objects during feature learning.*

Remarkably, no ventral stream neural response data were (directly) used to create the model feature sets.

Yet, the hidden units of the model network approximate the hidden units of the brain's network

We take this model network as a viable "mechanism" of the brain's online execution of object categorization

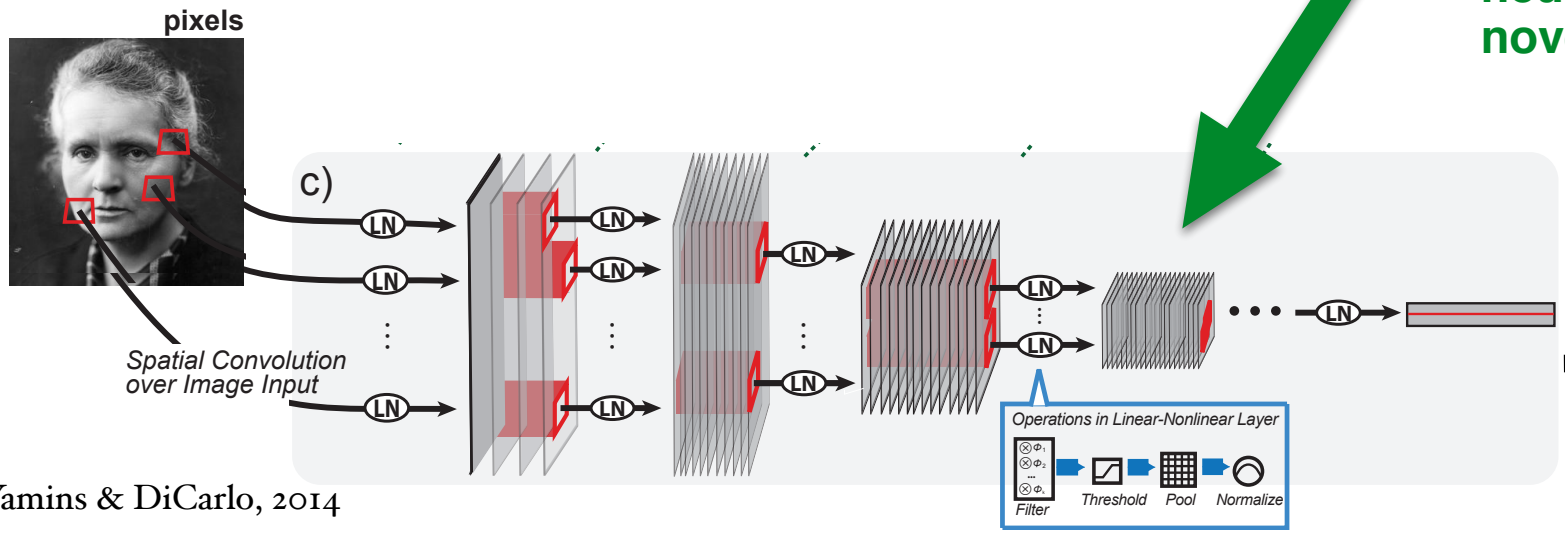
Yamins, Hong, Solomon, Seibert and DiCarlo ***PNAS* (2014)**



Metric 1: Single neuron predictivity:

For each neuron: Use some of its responses to fit (linear regression) using the model features as a basis set

Ask: How well does the model predict the neuron's responses on novel images?



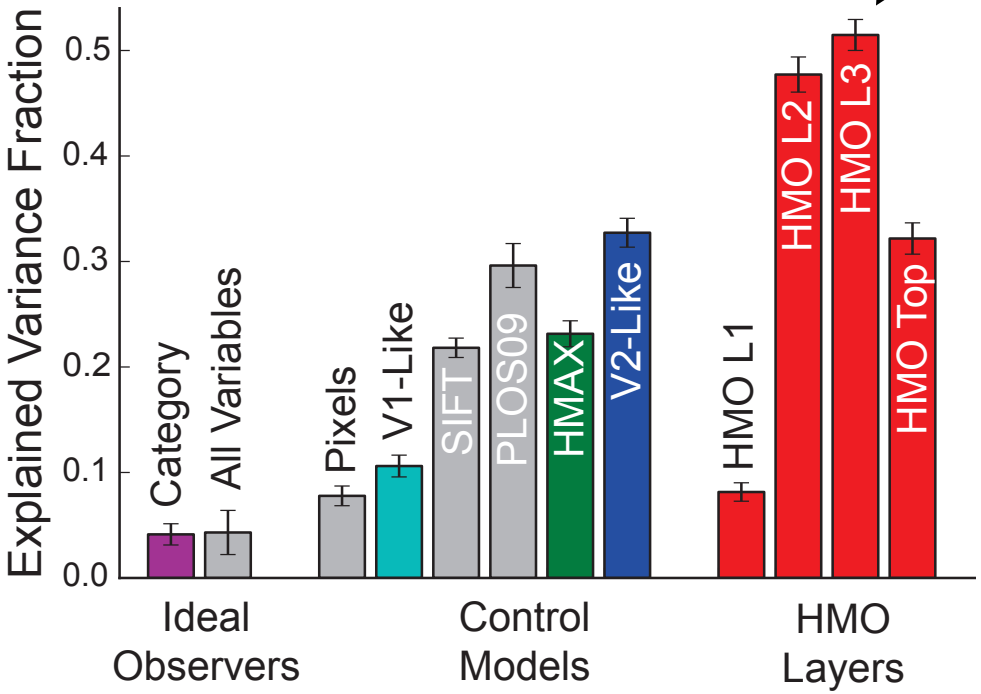
Top layer has thousands of visual "neurons" (aka "features")

Yamins & DiCarlo, 2014

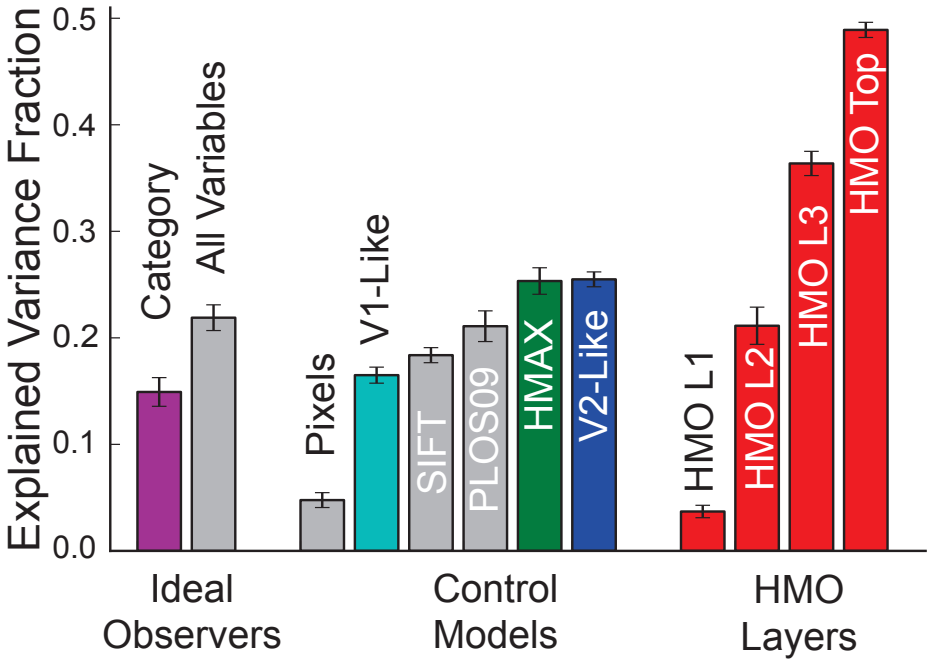
Median IT site: ~50% response variance predicted.
Dramatic improvement over previous models.

But, we are not finished. Likely more to explain.

V4 predictive power
(median over all neurons)



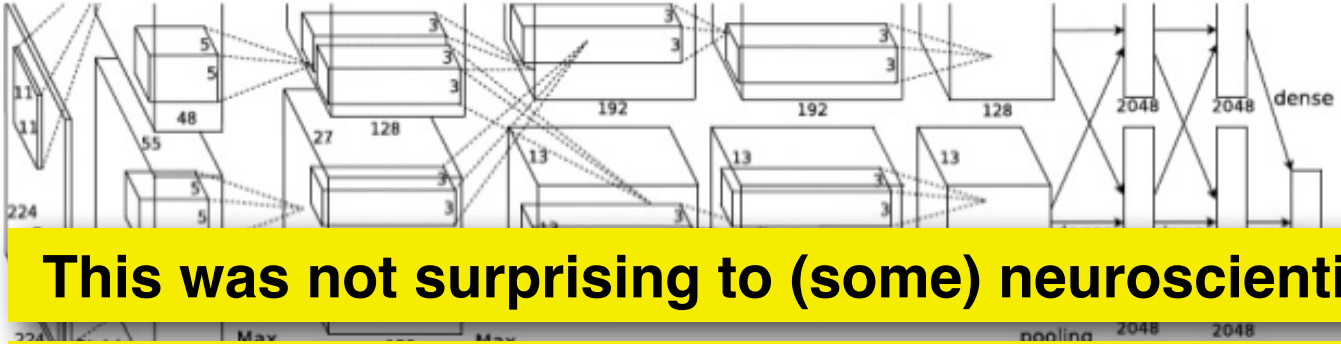
IT predictive power
(median over all neurons)



Yamins, Hong, Solomon, Seibert and DiCarlo ***PNAS (2014)***

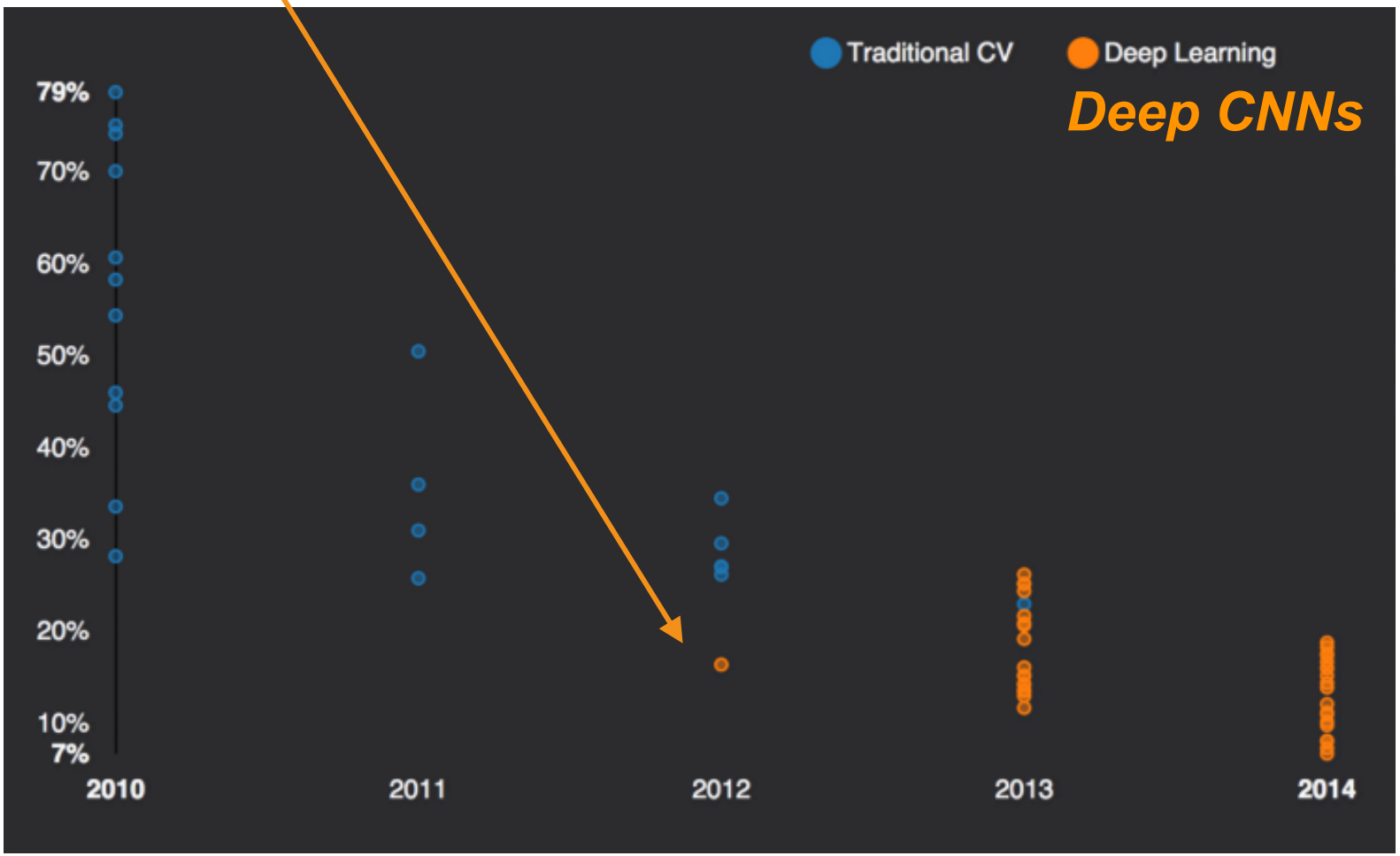
"AlexNet"

In 2012, this model elevated deep CNN models to a leadership role in computer vision



This was not surprising to (some) neuroscientists.

This particular deep CNN predicts V4 and IT better than our previous model!

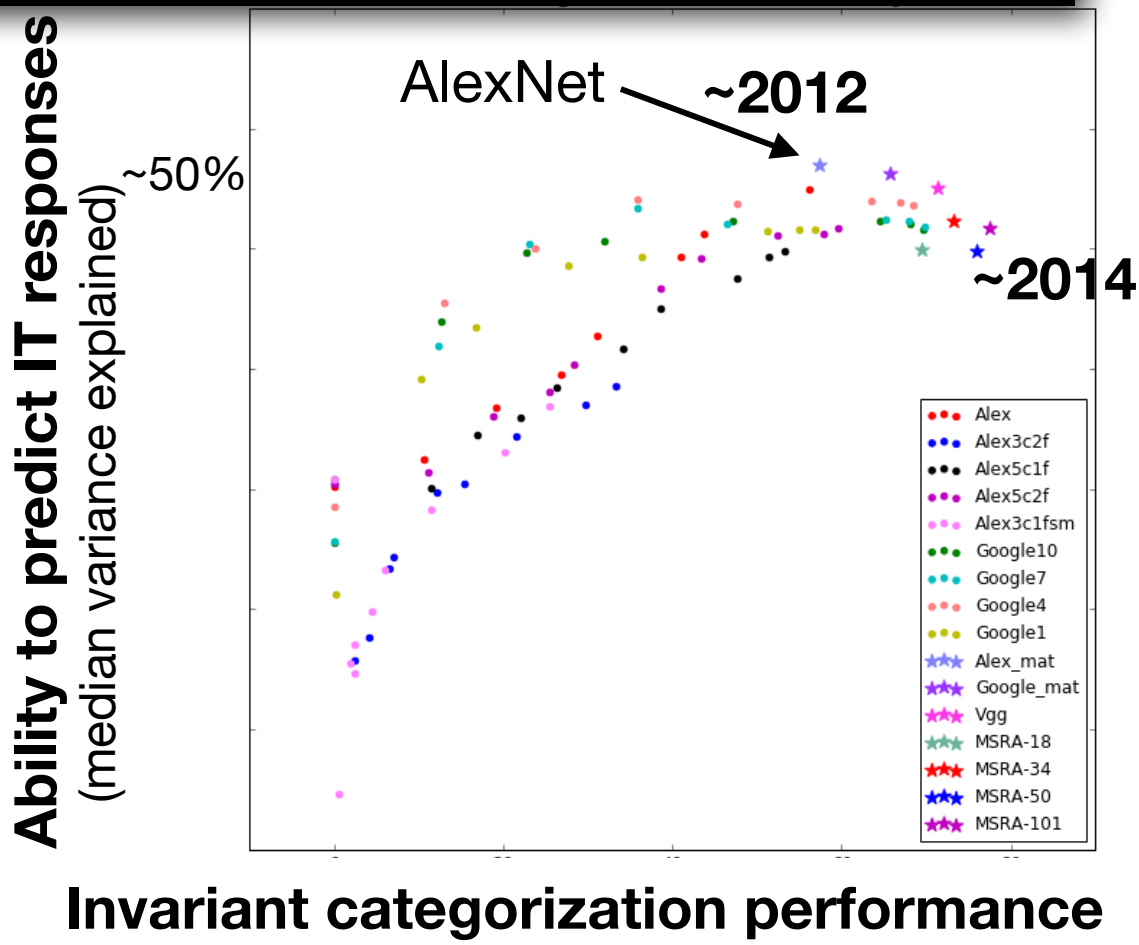


Q3 *Forward engineering summary:*

1) neuroscience-constrained architectural family (deep CNN) **+**
2) cognitive science-derived task (e.g. invariant categorization) **+**
3) engineering optimization (ImageNet, s. gradient descent, etc.)
—> produces a decent approximation of nature's neural mechanism of intelligent object sensing (in primates).

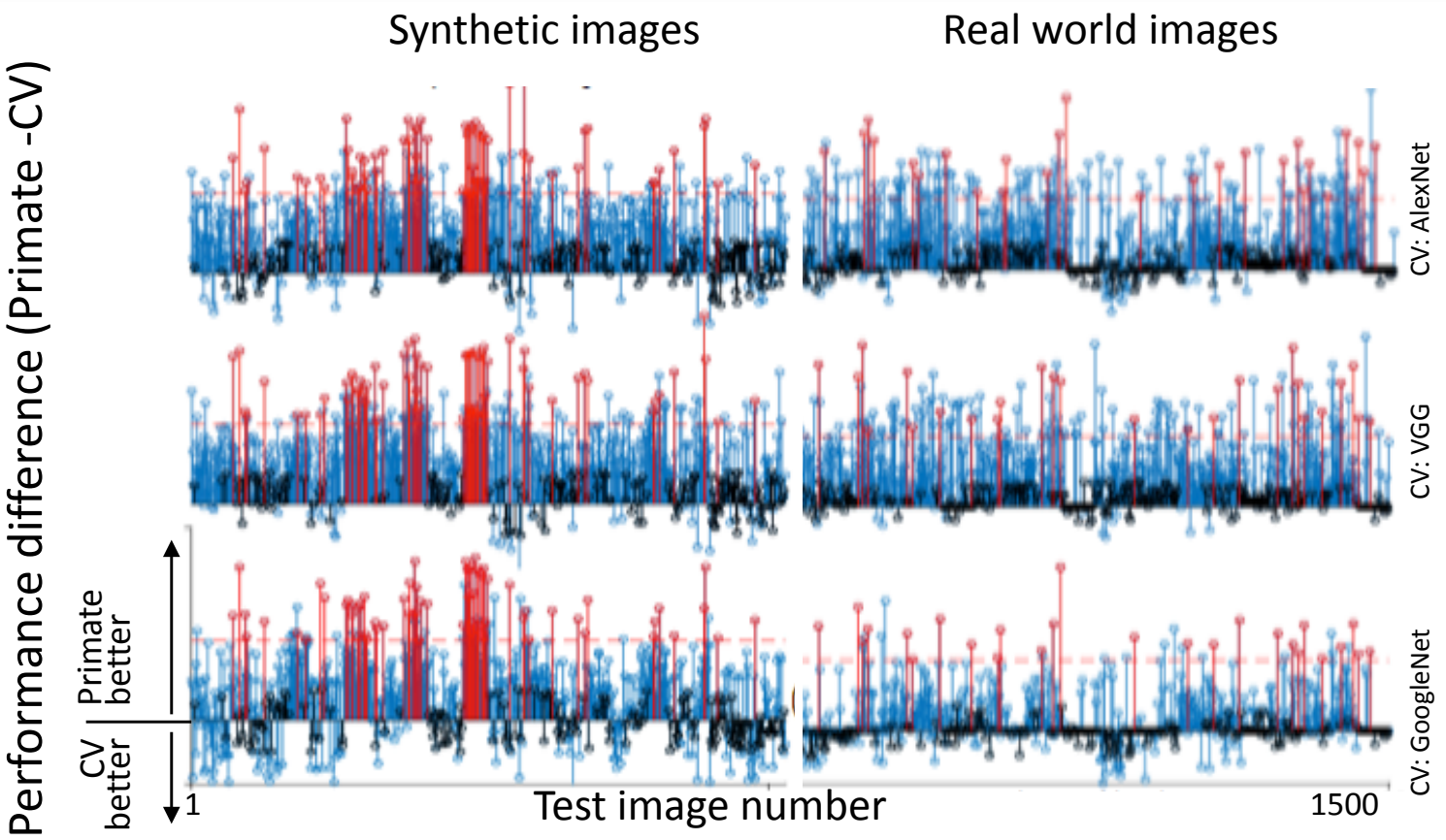
But these models do not fully predict nature's solution.

And primate behavior (and IT) are still better than current deep CNNs



Q1 Differences between state-of-the-art deep CNNs and Primates

Where do currently engineered systems fall down (relative to humans)?





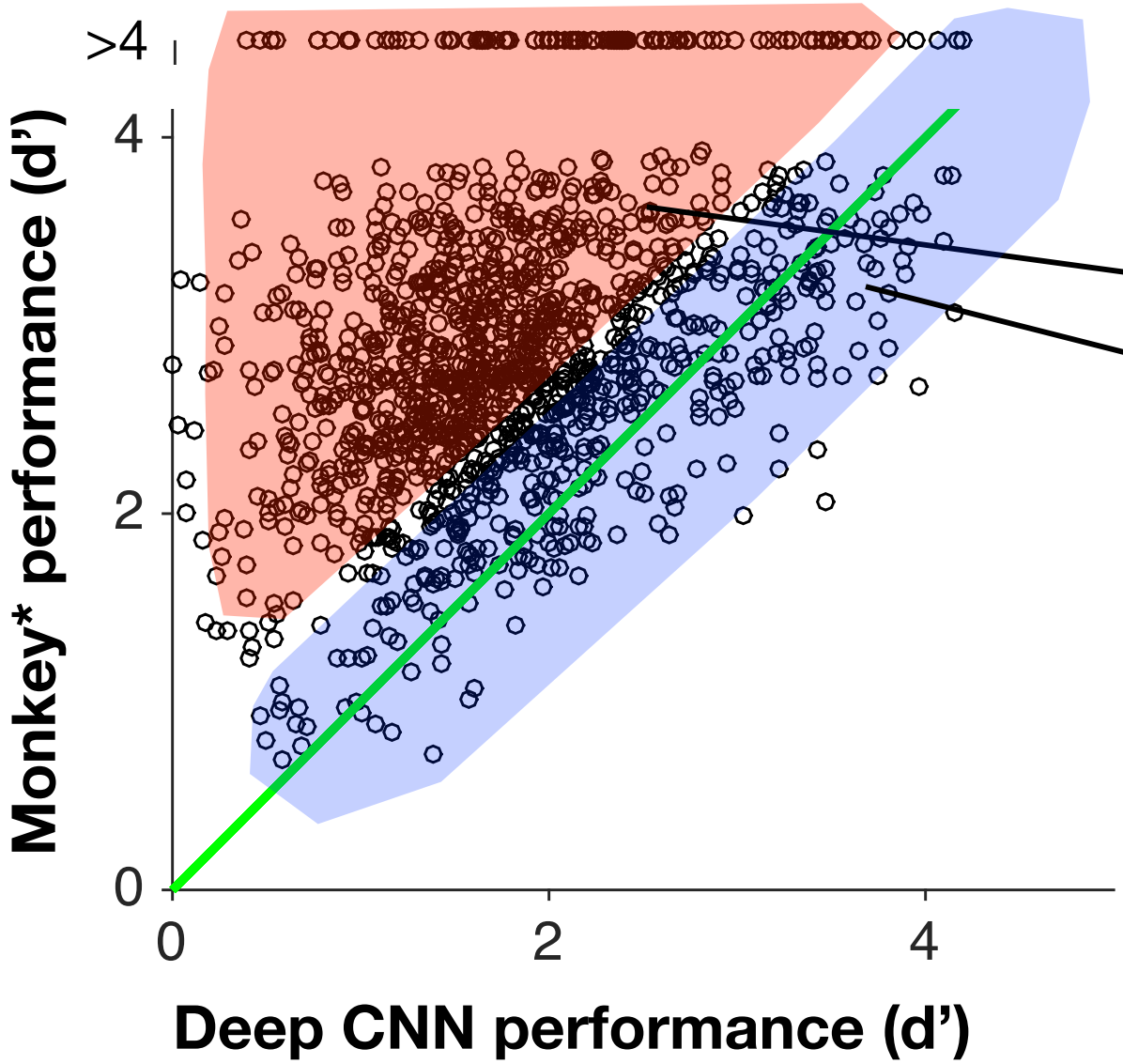
Kohitij Kar

We tested thousands of images and discovered hundreds of images that are reliably solved by (human AND non-human) primate brains with only 100 ms viewing duration, but not solved by current CV systems (red lines).

Kar, Kubilius, Issa, Schmidt & DiCarlo COSYNE 2017

Q1 Differences between state-of-the-art deep CNNs and Primates

Each dot is an image



(*results are very similar for human performance)

Q1 Differences between state-of-the-art deep CNNs and Primates

Challenge images



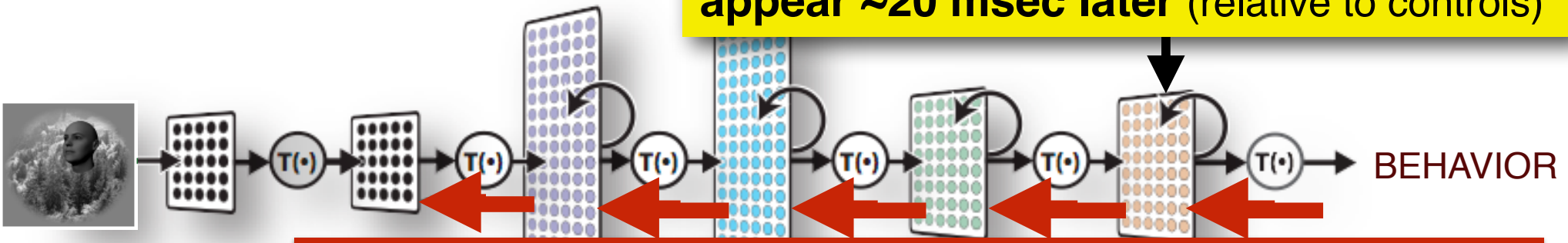
Control images



Q2

We have the experimental resolution to see that the brain's solution to each challenge image is conveyed in the IT feature set

But, we discovered that those solutions appear ~20 msec later (relative to controls)

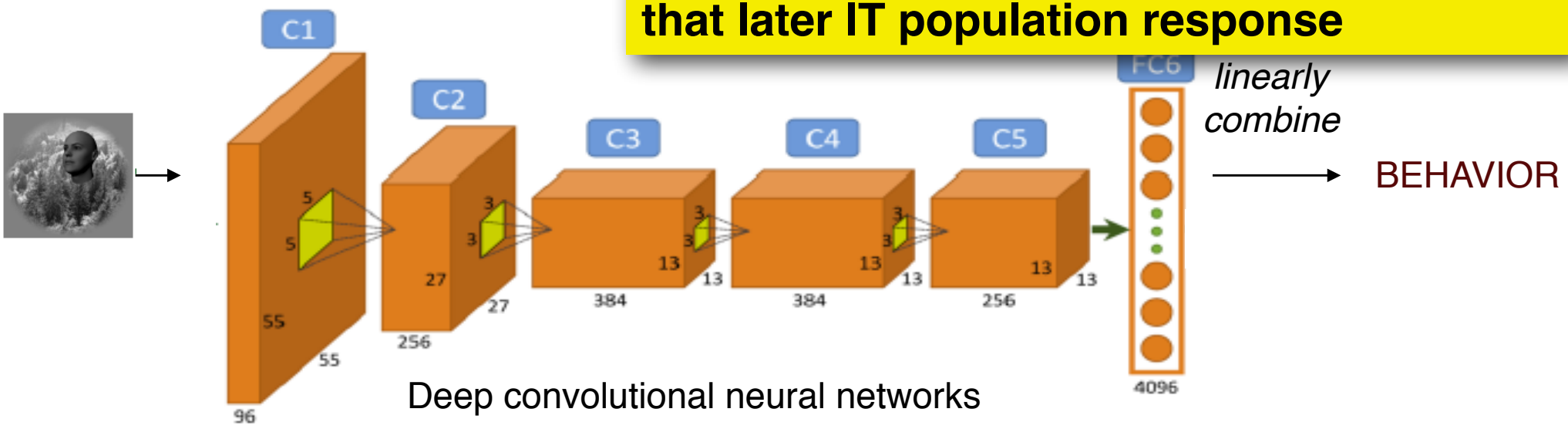


Along with controls. argues that feedback and recurrent connections are providing a big performance boost for these challenge images

pixel RGC LGN V1 V2 V4 IT

Notably, deep CNNs best predict the early IT population response, but poorly predict that later IT population response

Enc

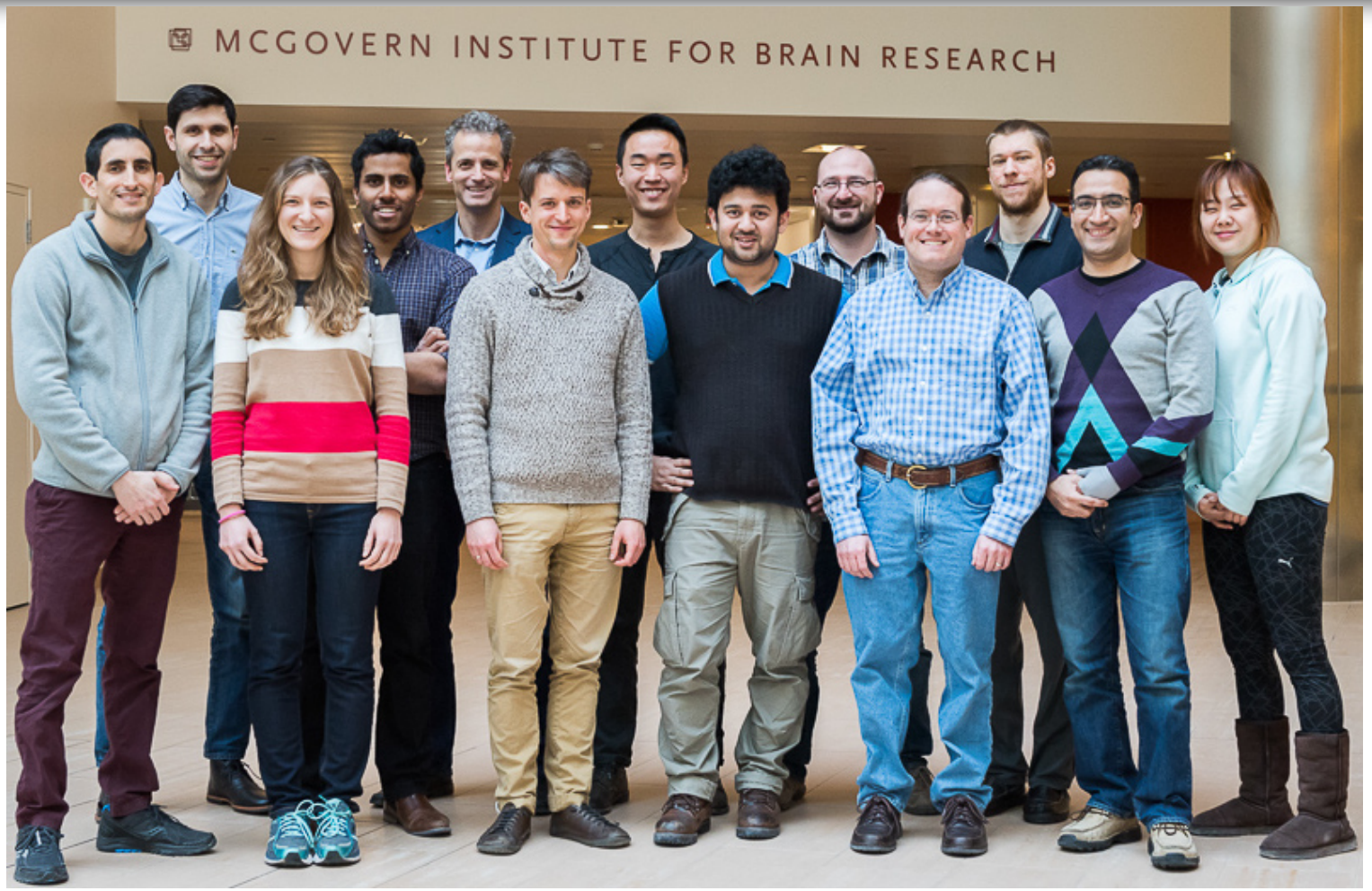


Deep convolutional neural networks

Acknowledgements

Current lab members:

Pouya Bashivan
Elias Issa
Kohitij Kar
Hyodong Lee
Micheal Lee
Shay Ohayon
Jon Prescott-Roy
Rishi Rajalingham
Kailyn Schmidt
Darren Seibert
Christopher Shay
Chris Stawarz



Alumni:

Arash Afraz
David Cox
Chou Hung
Gabriel Kreiman
Nuo Li
Najib Majaj
Nicolas Pinto
Nicole Rust
Ethan Soloman
Dan Yamins
Davide Zoccolan

Key collaborating labs:

Ed Boyden (MIT)
David Cox (Harvard)
Bob Desimone (MIT)
Tomaso Poggio (MIT)
John H.R. Maunsell (U Chicago)
J. Anthony Movshon (NYU)
Nancy Kanwisher (MIT)

- *NIH NEI*
- *Simons Foundation*
- *ONR MURI*
- *IARPA*
- *McGovern Institute*