Neural mechanisms underlying visual object recognition:
The convergence of computer vision and biological vision

Center for Brains, Minds, and Machines: Summer School 2015, Woods Hole, MA

James DiCarlo MD, PhD

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

brain+cognitive sciences

“Object recognition” (operationalized)

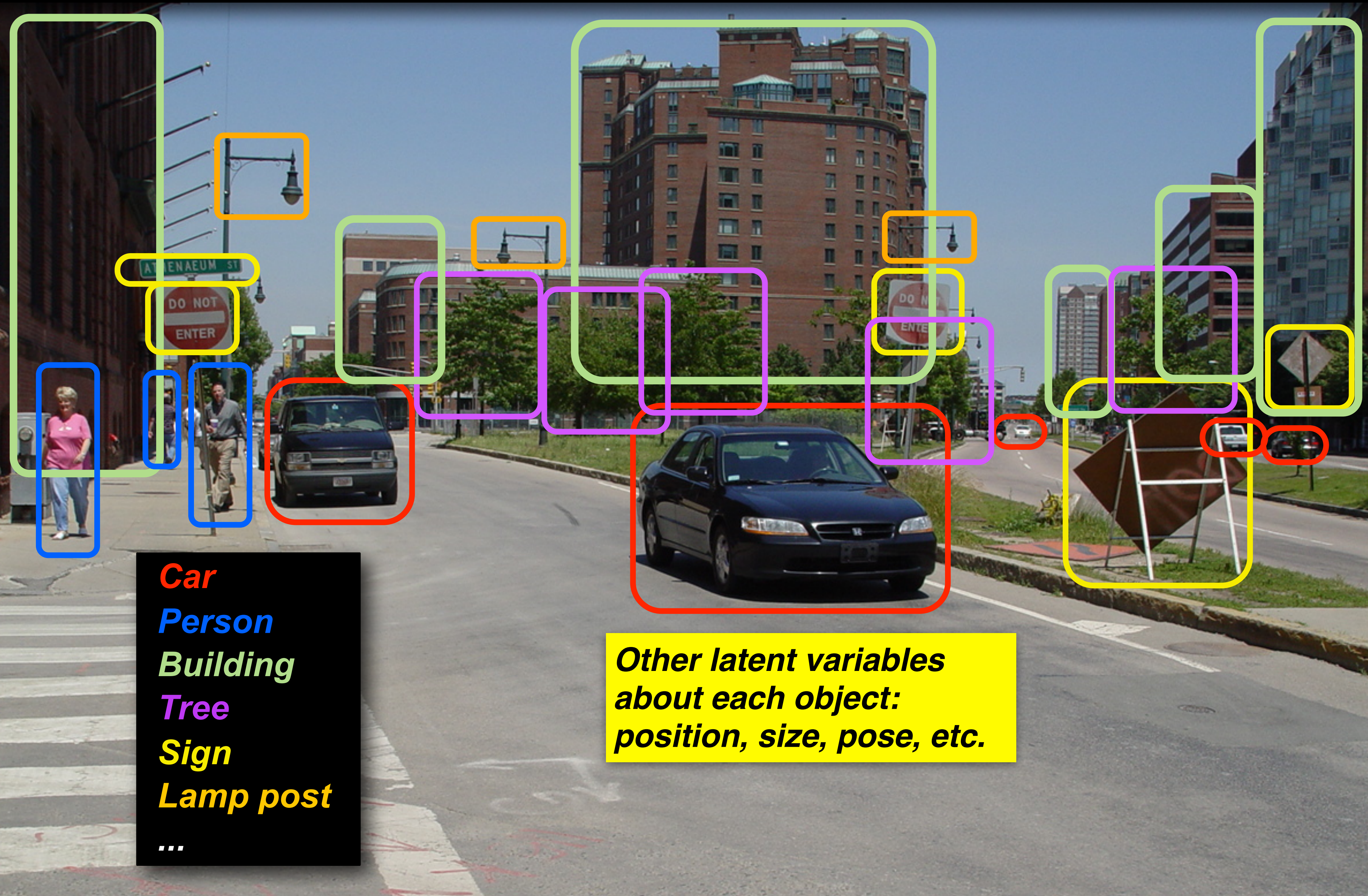


Image adapted from MIT Street Scenes Database (Courtesy of Tommy Poggio)

Why study object recognition in the brain?

The brain's internal representation of objects is the substrate of cognition:

- *memory*
- *value judgements*
- *decisions*
- *actions*

- *Obstacle avoidance*
- *Navigation*
- *Danger avoidance*
- *Resource detection*
- *Social interactions*
- *Mate selection*
- *Threat detection*
- *Reading*
- *...*

The convergence of three fields

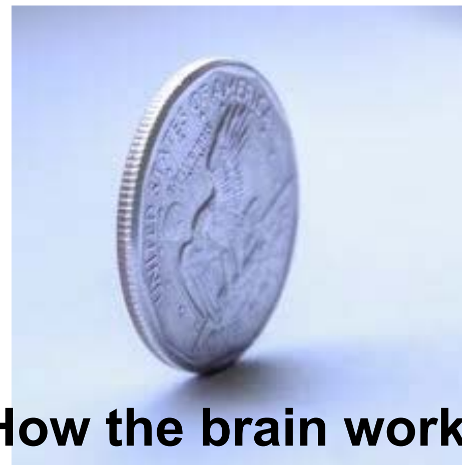
When biological brains perform better than computers

New phenomena

New ideas,
algorithm parameters

psychophysics

neuroscience



How the brain works

computer science

Attempt to test/
falsify those
hypotheses

Falsifiable
hypotheses

When computers perform as well as or better than biological brains

A bit of history…

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966

Vision Memo. No. 100.

THE SUMMER VISION PROJECT

The final goal is OBJECT IDENTIFICATION which will actually name objects by matching them with a vocabulary of known objects.

Goals - Specific

We plan to work by getting a simple form of the system going as soon as possible and then elaborating upon it. To keep the work reasonably coordinated there is a graduated scale of subgoals.

Courtesy of Mike Tarr

- *100 billion computing elements*
- *solves problems not soluble by previous machines*
- *requires only 20 watts of power!*

Key algorithms are ***classified***

An engineer's point of view…

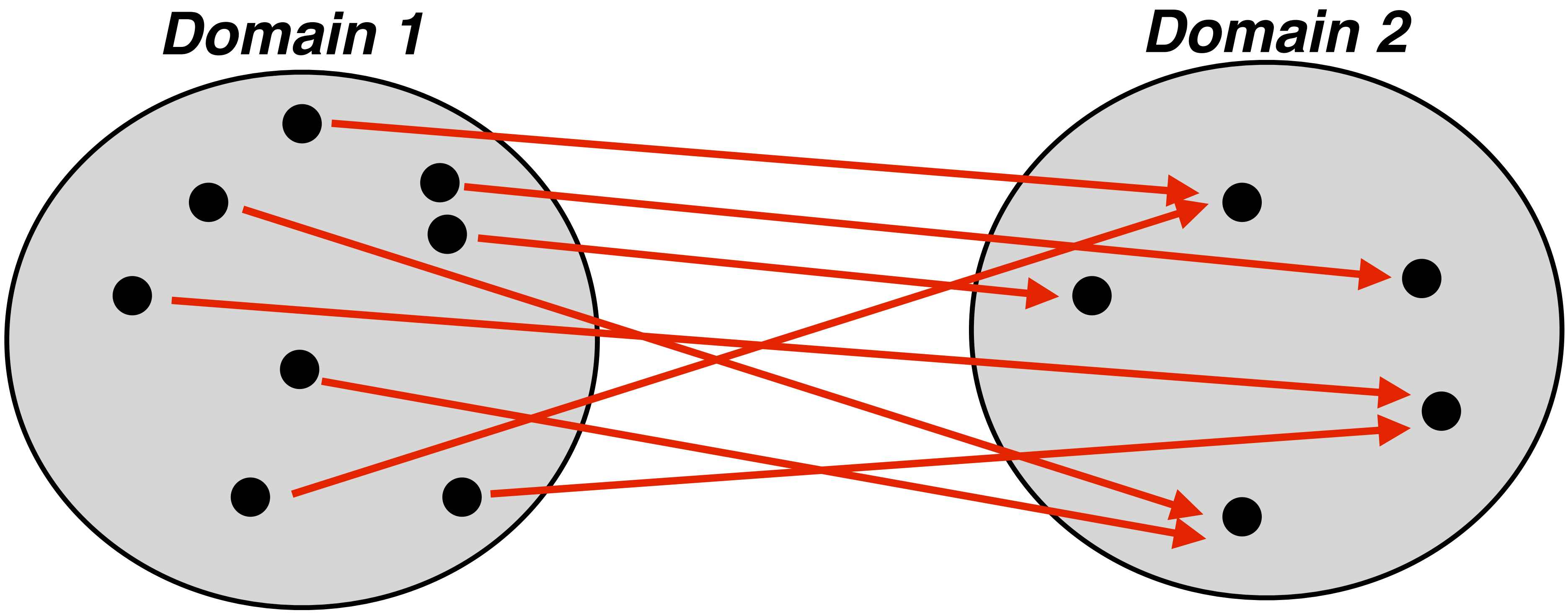
Which system is better?

Problem to solve	***Our brain***	***Machines today*** *(e.g. computers)*
Calculation		**WINNER**
Win at chess		**WINNER**
Win at Jeopardy		**WINNER**
"Memory"		
"Seeing"		
Pattern matching		**WINNER**
Object recognition	**WINNER**	
Scene "understanding"	**WINNER**	
Walking	**WINNER**	

Gateway problem (vision, neocortex)

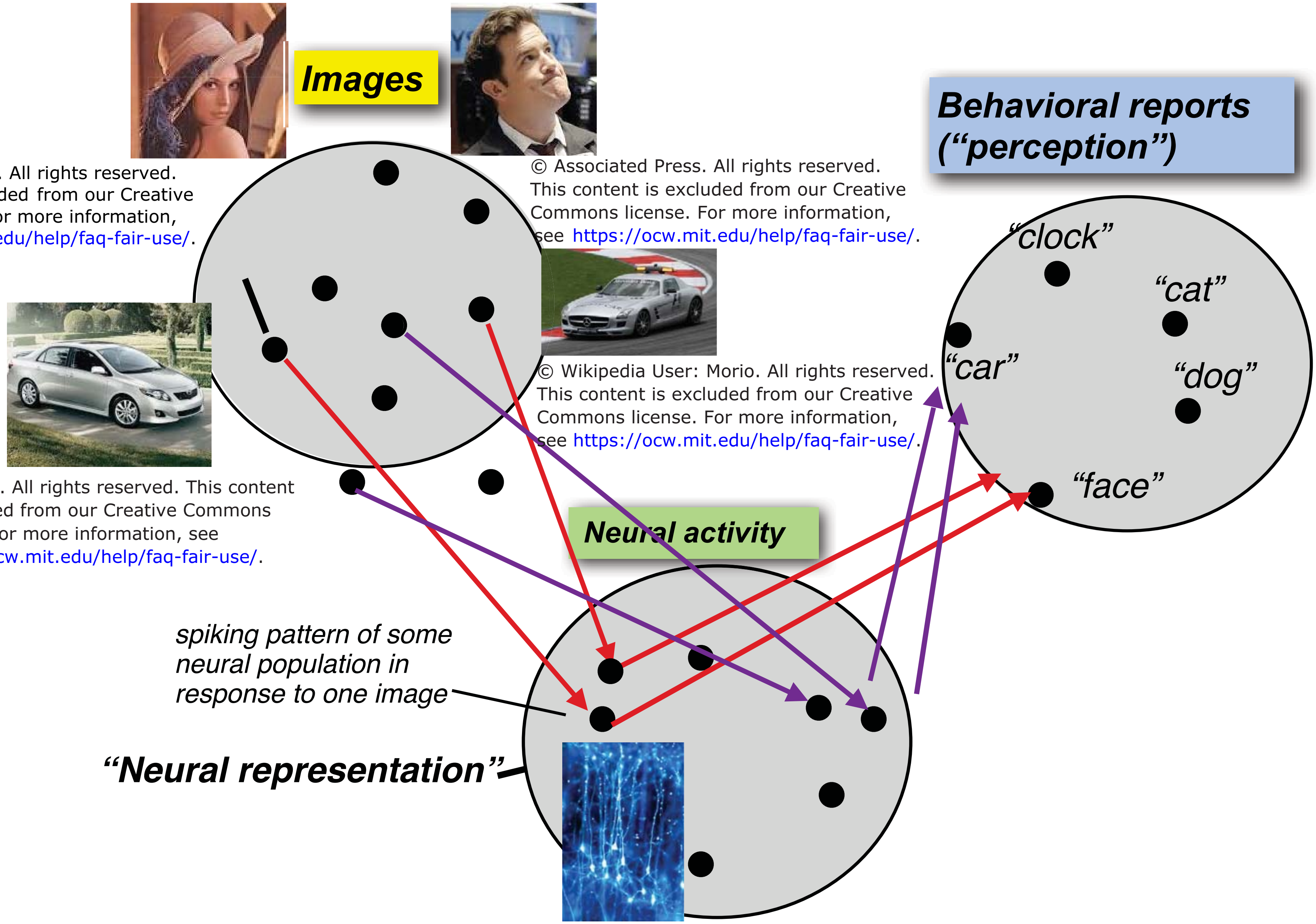
Our goal: Discover how the brain solves object recognition (algorithms)

A scientist's point of view

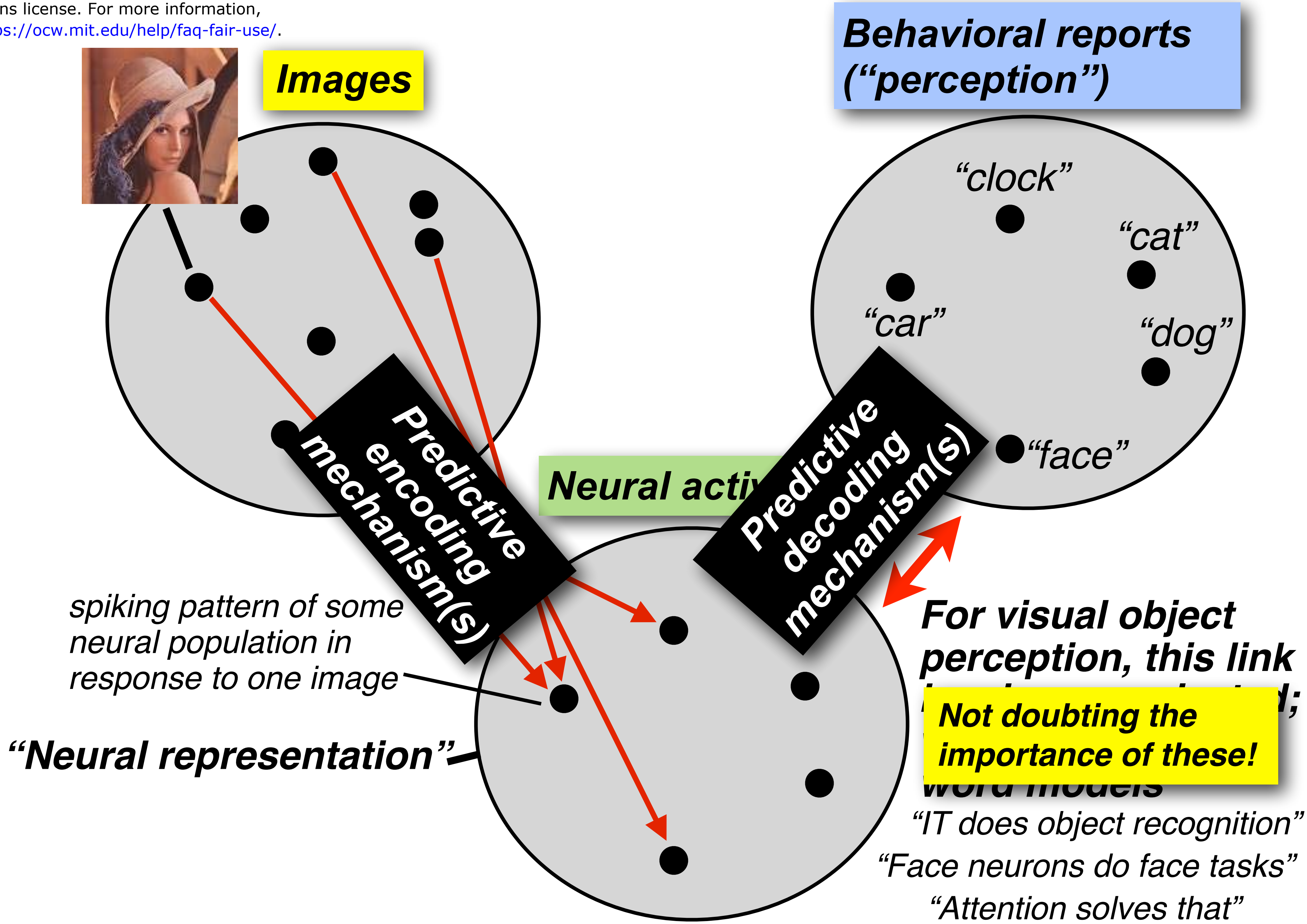


Science: given state of Domain 1, <u>predict</u> state of Domain 2

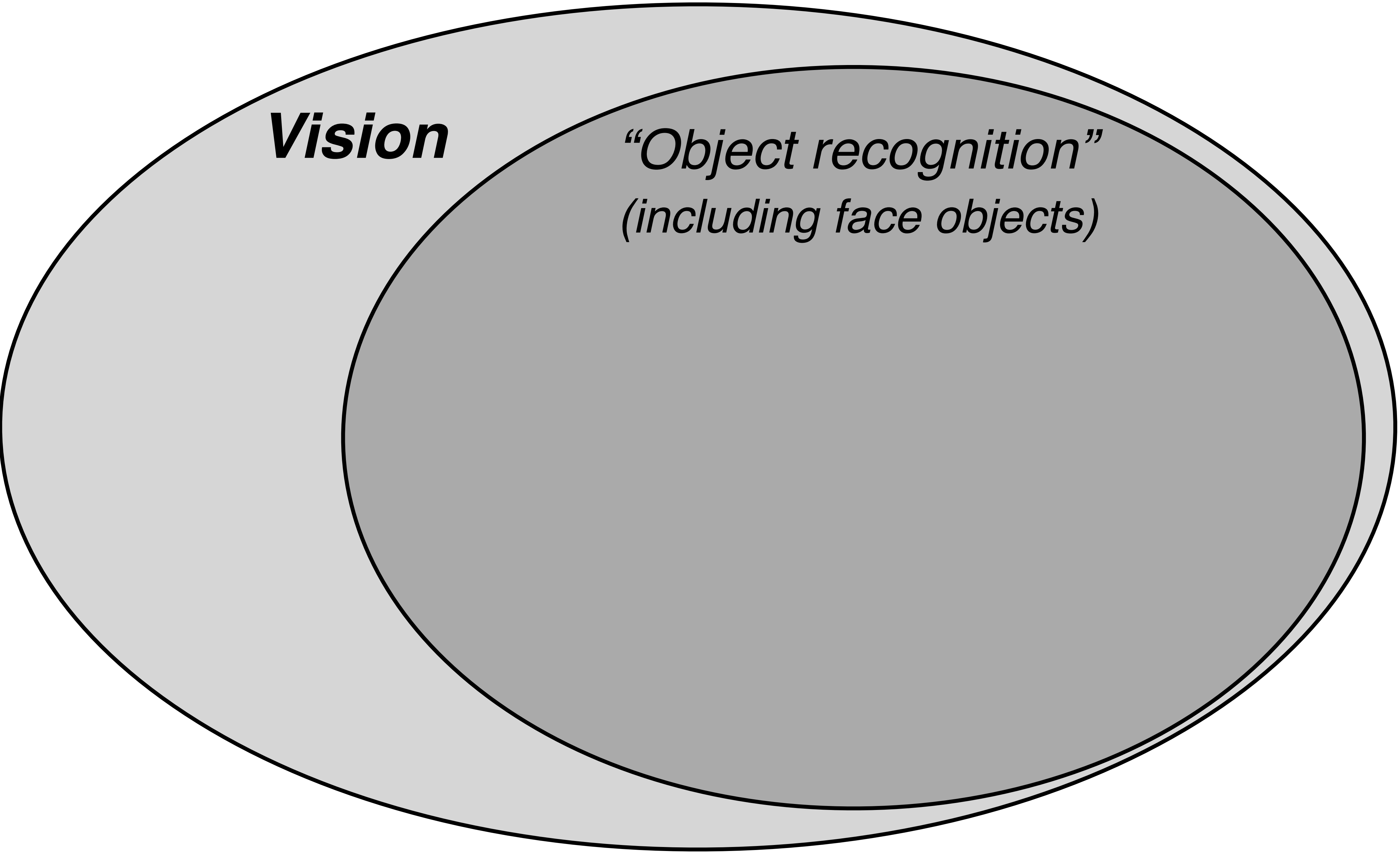
The accuracy of this predictive mapping is a measure of the strength of a scientific field



Accurate predictivity* is the core product of science** → ***Underlies engineer's ability to build, fix, or augment


Images
Behavioral reports
("perception")
"clock"
"cat"
"car"
"dog"
"face"
Predictive encoding mechanism(s)
Neural acti
Predictive decoding mechanism(s)
spiking pattern of some neural population in response to one image
"Neural representation"
For visual object
perception, this link
Not doubting the importance of these!
"IT does object recognition"
"Face neurons do face tasks"
"Attention solves that"

Let's try to define a domain of behavior so that we can gauge/make progress in prediction.



Object recognition as solved by primates

Central ~10 degrees

Object recognition as solved by primates ~200 ms snapshots

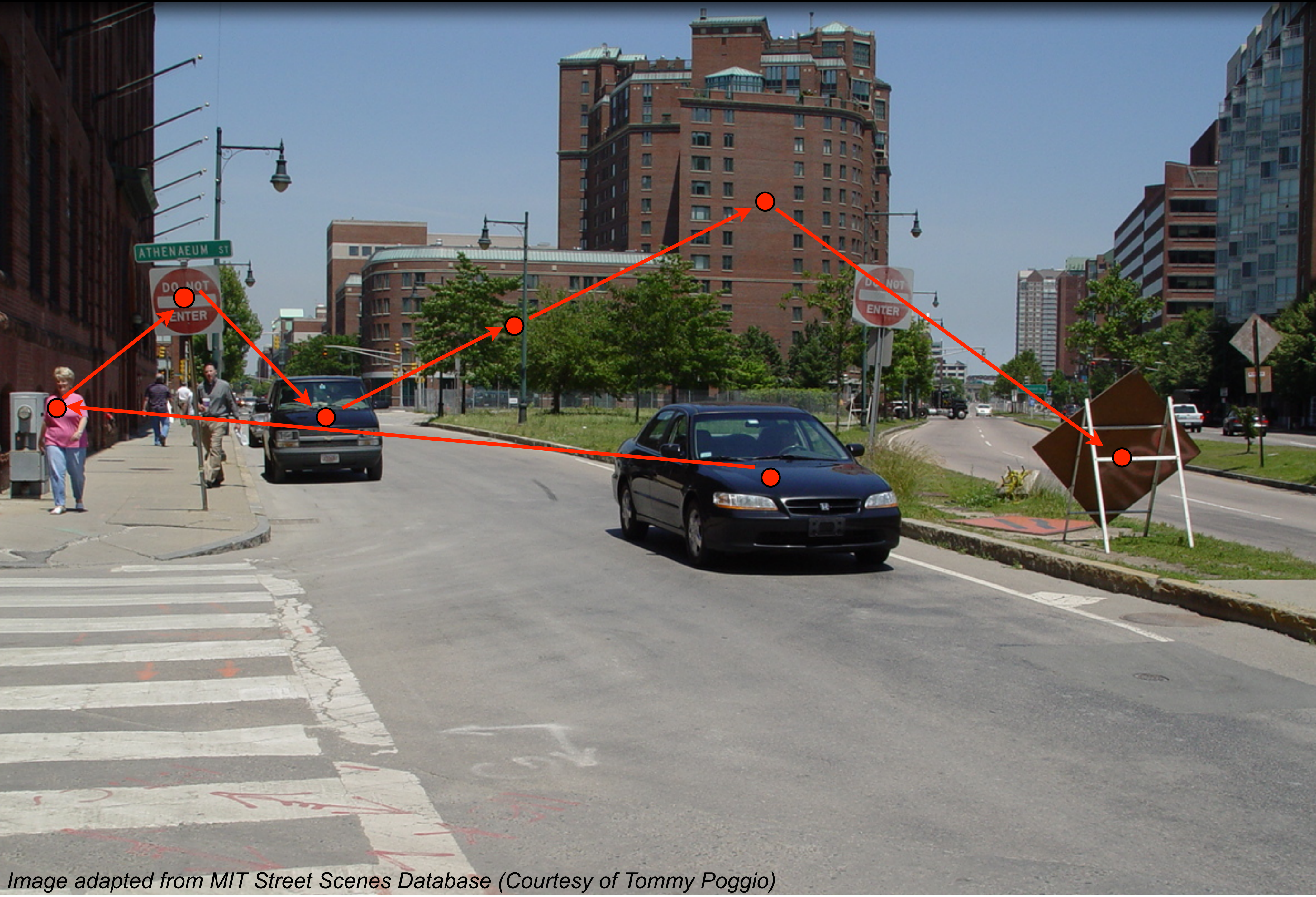


Image adapted from MIT Street Scenes Database (Courtesy of Tommy Poggio)

Object recognition as solved by primates

Core object recognition

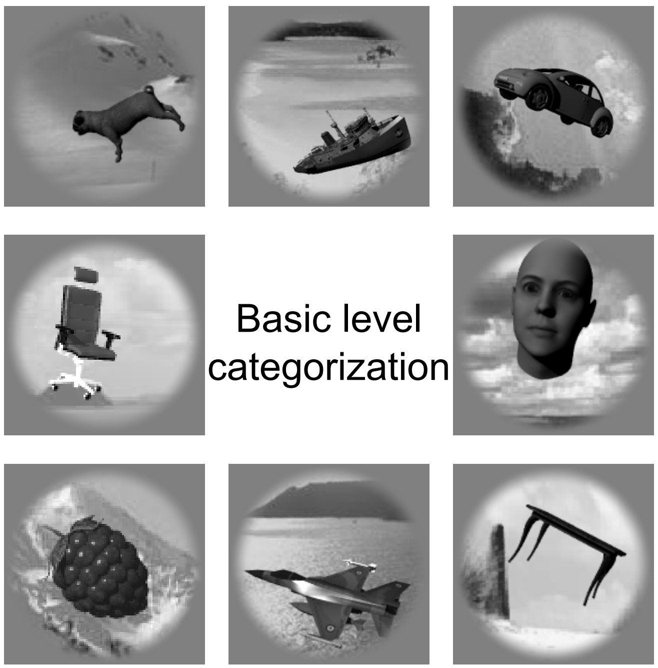
central ~10 deg of visual field
100-200 ms viewing duration

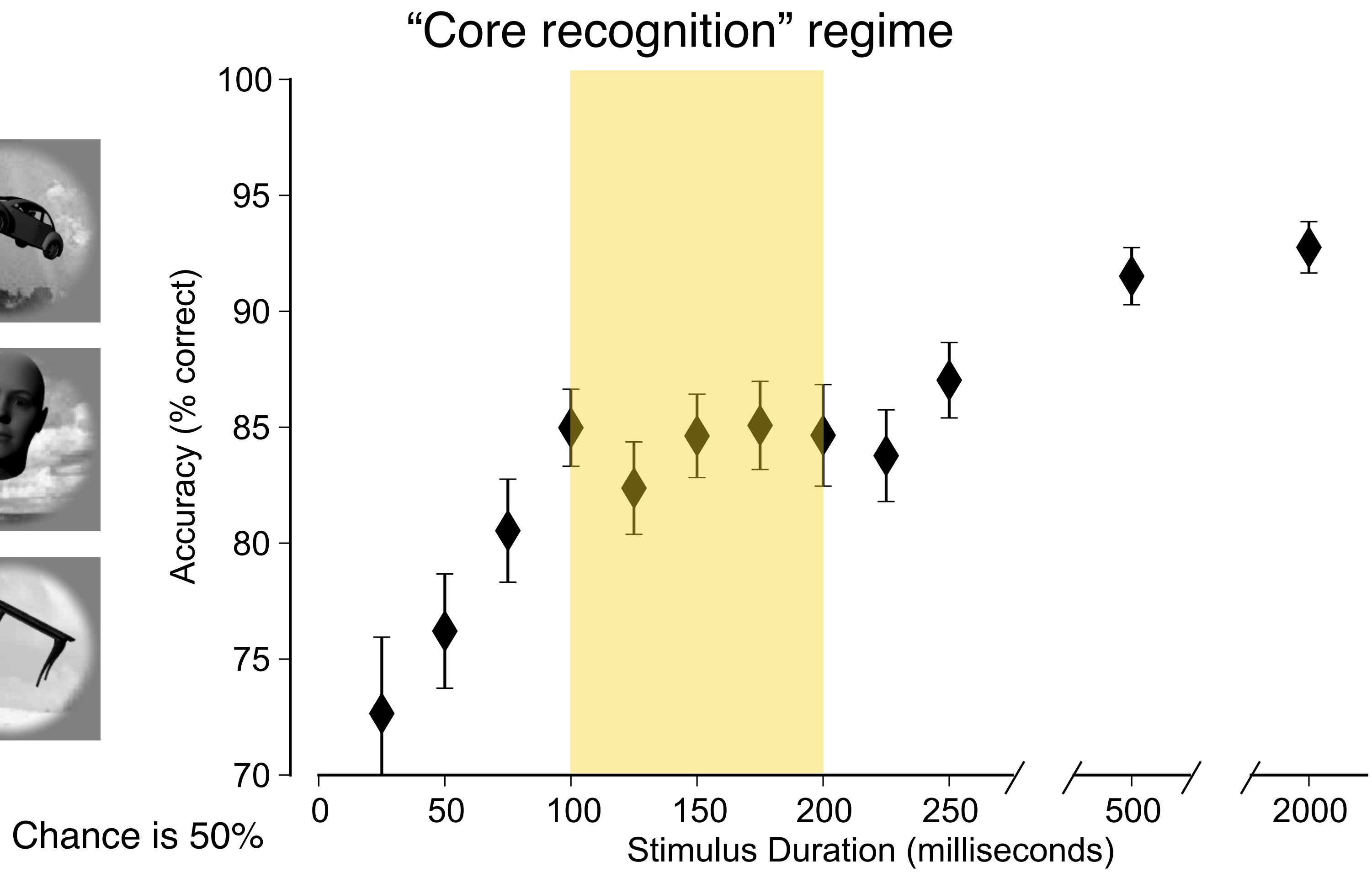
Our visual system excels at core object recognition

Core object recognition

central ~10 deg of visual field
100-200 ms viewing duration

Human object recognition (categorization) accuracy as a function of image viewing time

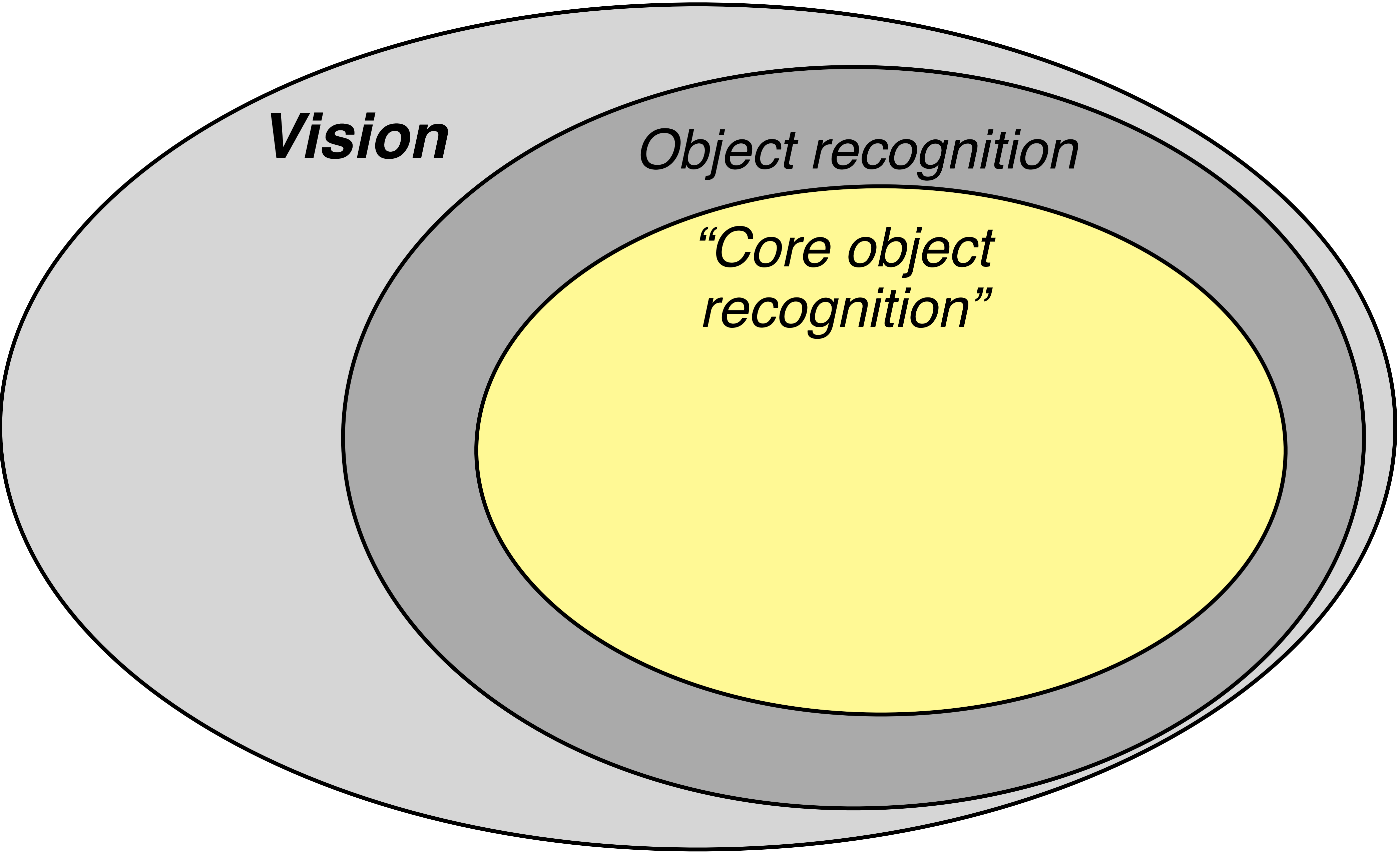




All the data I will show you today

Typical primate fixation duration during natural viewing

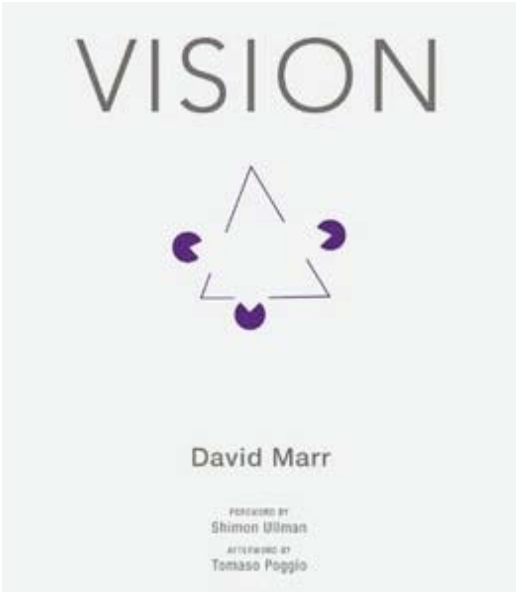
Let's try to define a domain of behavior so that we can gauge/make progress in prediction.



The challenge of level

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

David Courtnay Marr
(1946-1980)

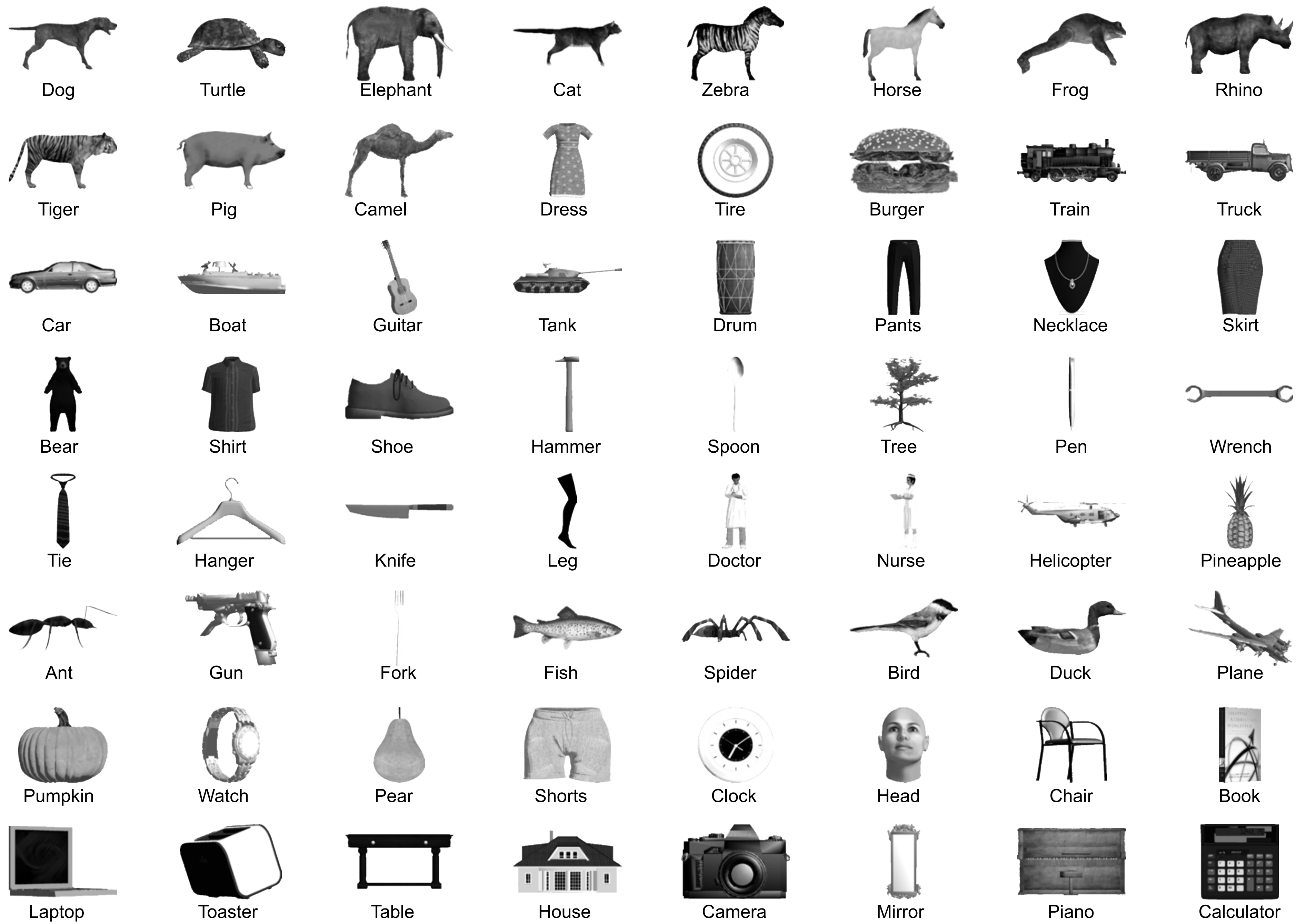


Marr, 1982

Reaching a common language

	Comp vision, Machine learning	Neuroscience, Cognitive Science
1. ***What is the problem we are trying to solve?***	*Benchmarks* *Brain solves "it"*	*"Perception"* *Behavior* *Psychophysics*
2. What do good solutions look like?	*Useful image representations ("features")*	*Explicit neuronal population spiking patterns*
3. How do we instantiate these solutions?	*Algorithms, mechanisms*	*Neuronal wiring / weighting patterns*
4. How do we construct those instantiations?	*Learning rules, initial conditions, training images*	*Plasticity, architecture, experience*

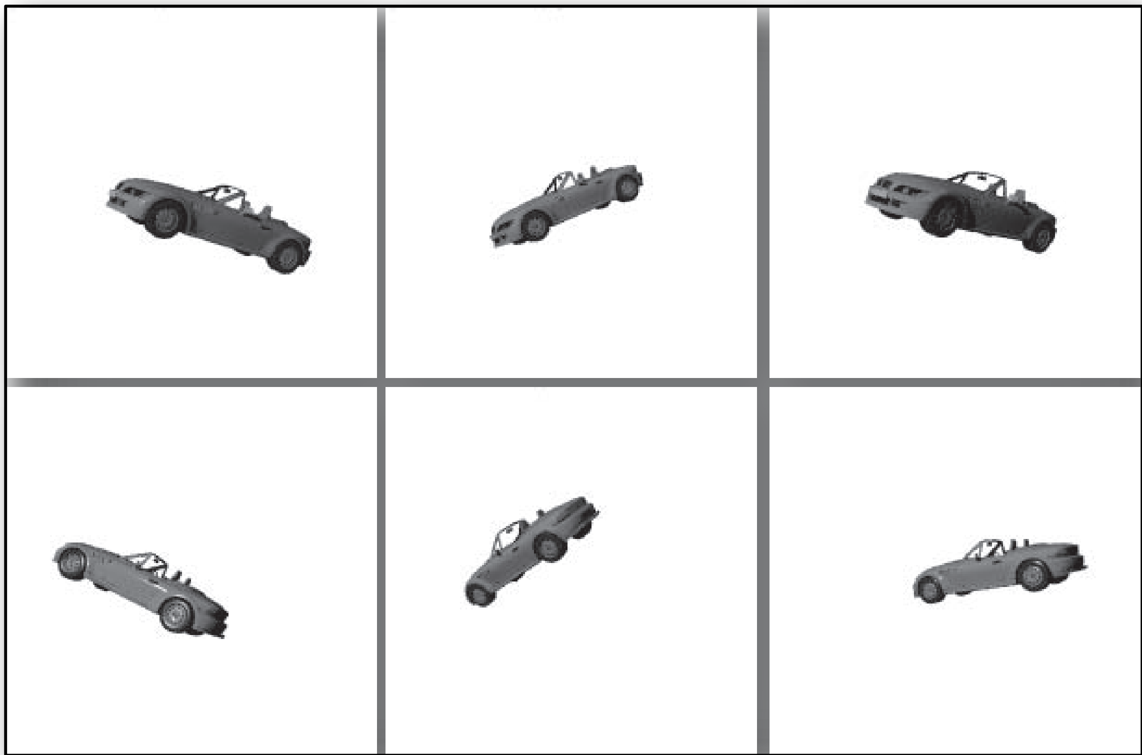
Behavioral challenge 1: Many possible objects



Behavioral challenge 2: Common physical source (object) can produce many images

"Identity preserving image variation"

View: position, size, pose, illumination

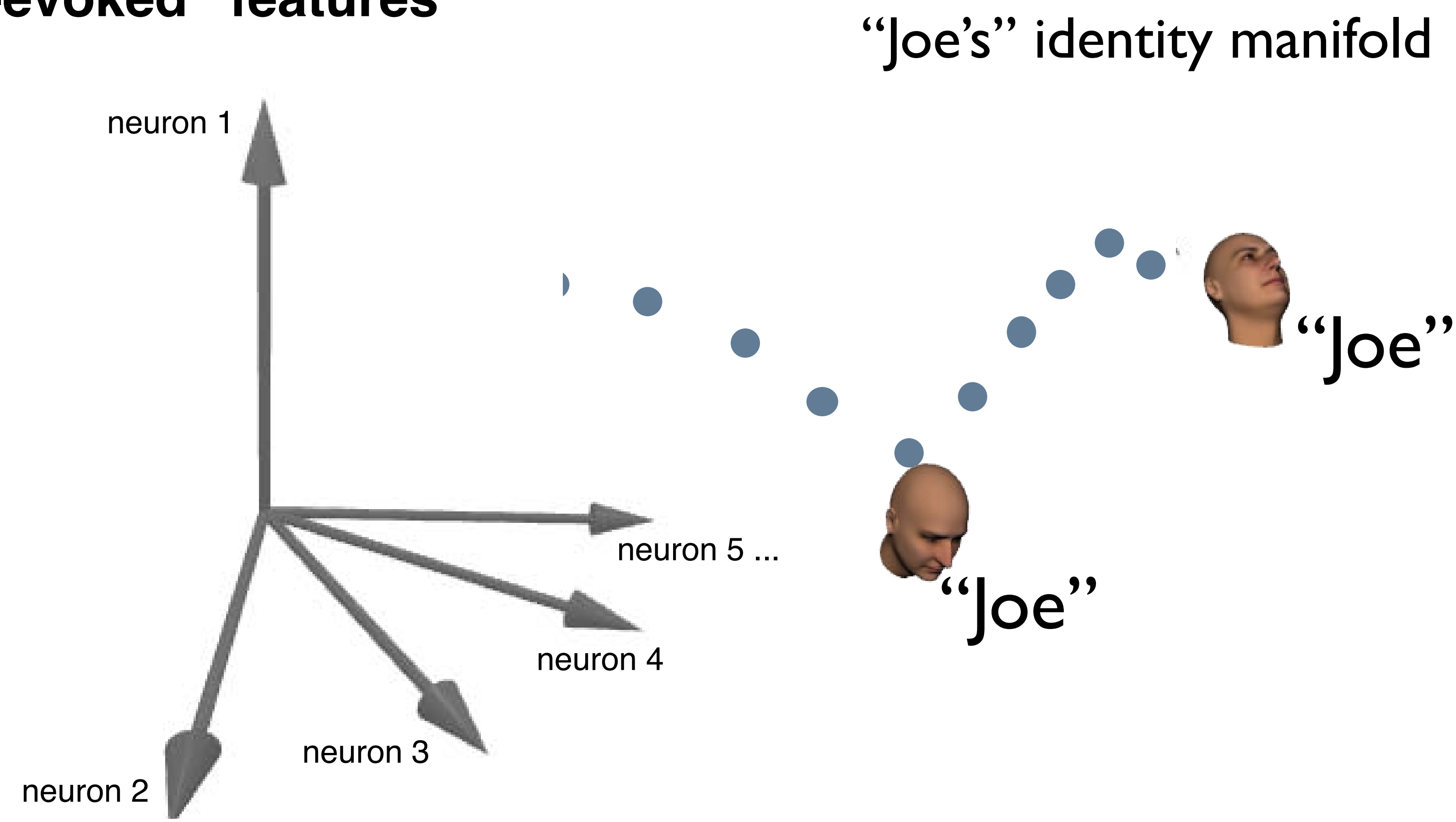


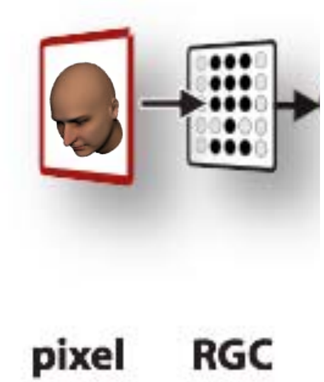
Clutter, occlusion

subordinate level variation

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)*

The brain's "camera" represents the image as populations of visually-evoked "features"





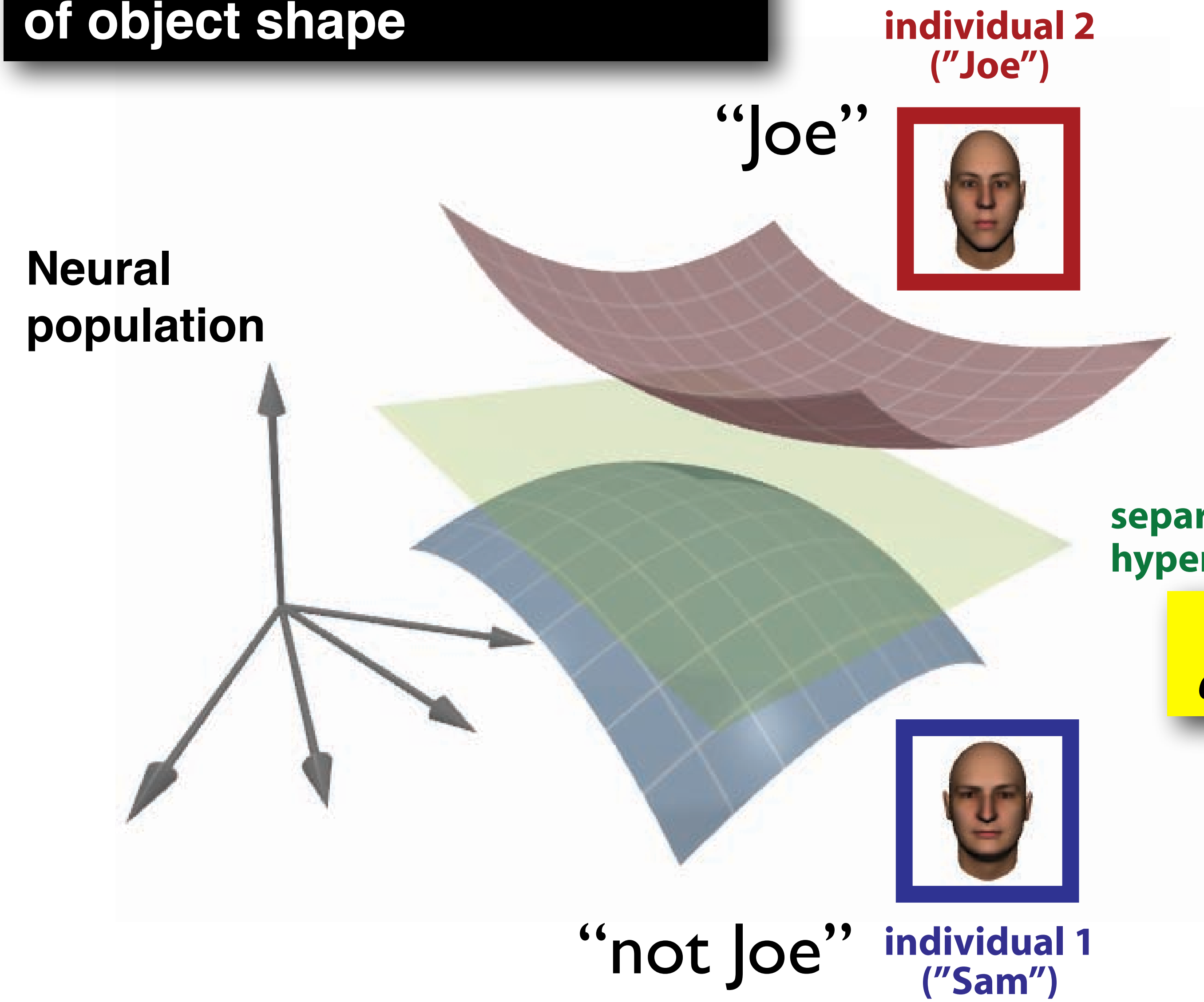
Source: DiCarlo, James J., and David D. Cox. "Untangling invariant object recognition. "Trends in cognitive sciences 11, no. 8 (2007): 333-341; https://doi.org/10.1016/j.tics.2007.06.010.

The computational crux of object and face recognition

A "good" set of visual features

== "Explicit" representation of object shape

We assume: "shape" maps to "identity" and "category"



Should be able to find it with low* number of training examples

linear classifier ≈ *downstream neuron(s)*

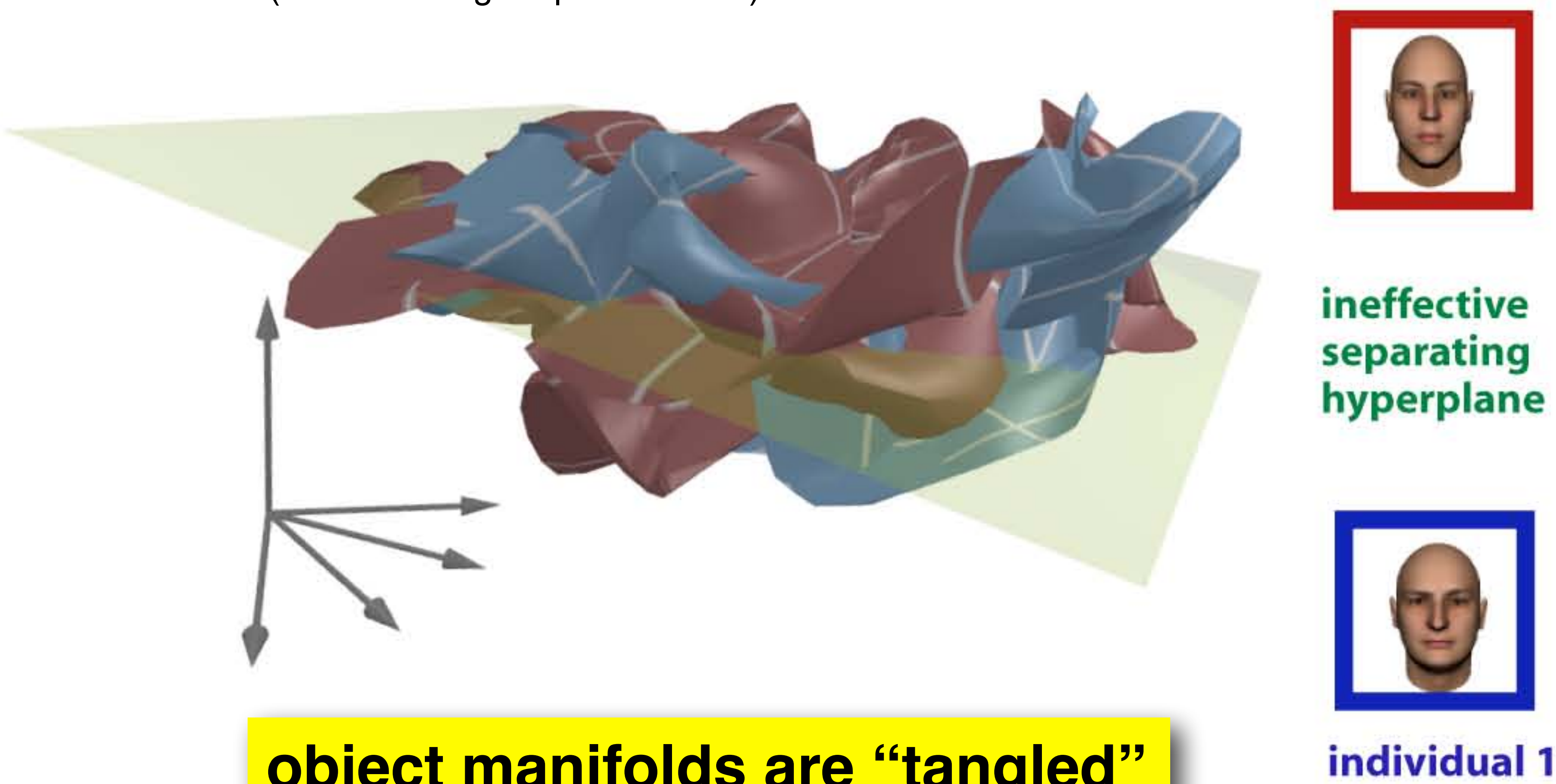
*DiCarlo and Cox, **TICS** (2007)*

Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission. Source: DiCarlo, James J., and David D. Cox. "Untangling invariant object recognition. "Trends in cognitive sciences 11, no. 8 (2007): 333-341; https://doi.org/10.1016/j.tics.2007.06.010.

Invariance is the computational crux of object and face recognition

Pixel population representation

(~ retinal image representation)

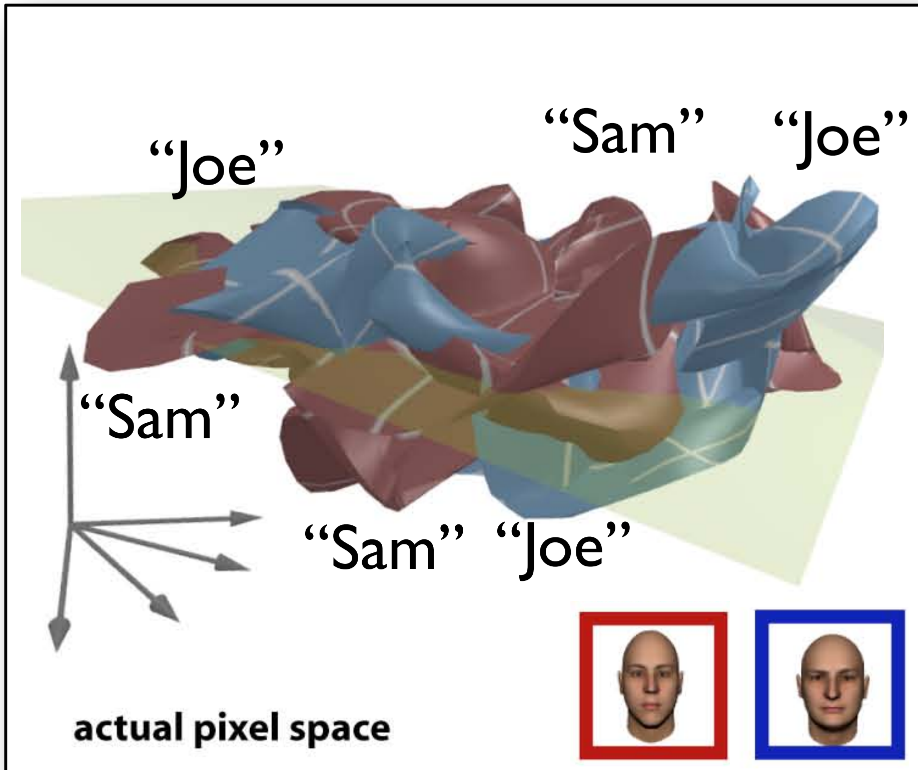


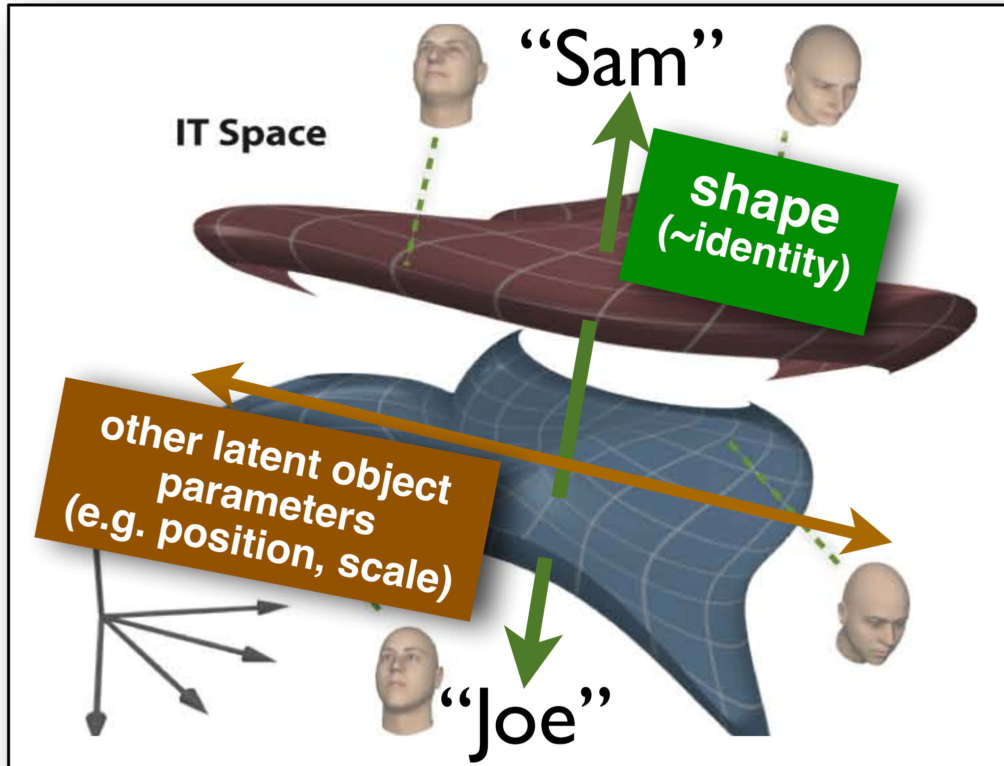
object manifolds are "tangled"

(Due to identity-preserving image variation.)

*DiCarlo and Cox, **TICS** (2007); Pinto, Cox, and DiCarlo, **PLoS Comp Bio** (2008)*

DiCarlo and Cox, ***TICS*** *(2007)*
DiCarlo, Zoccolan and Rust, ***Neuron*** *(2012)*



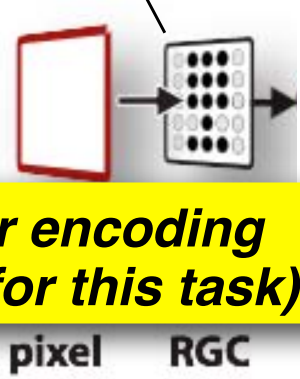


Tangled, implicit object information

Transformation →

This must be non-linear

Untangled, explicit object information

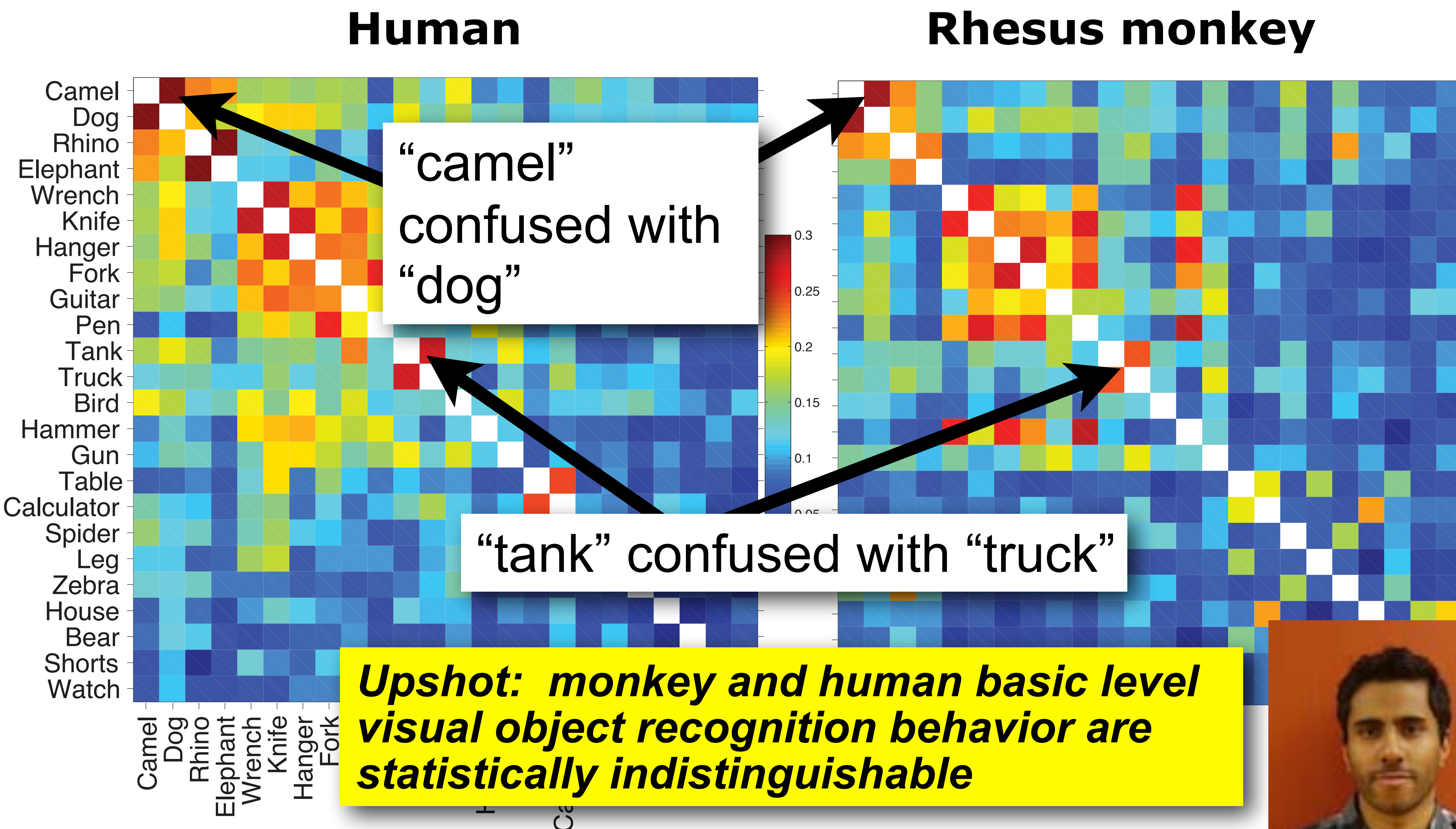


a poor encoding basis (for this task)

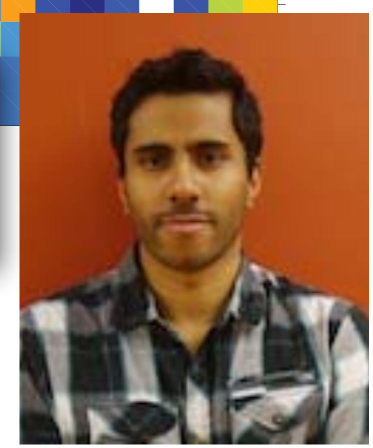
a powerful encoding basis somewhere in the brain

Source: DiCarlo, James J., and David D. Cox. "Untangling invariant object recognition. "Trends in cognitive sciences 11, no. 8 (2007): 333-341; https://doi.org/10.1016/j.tics.2007.06.010.

The ventral visual stream



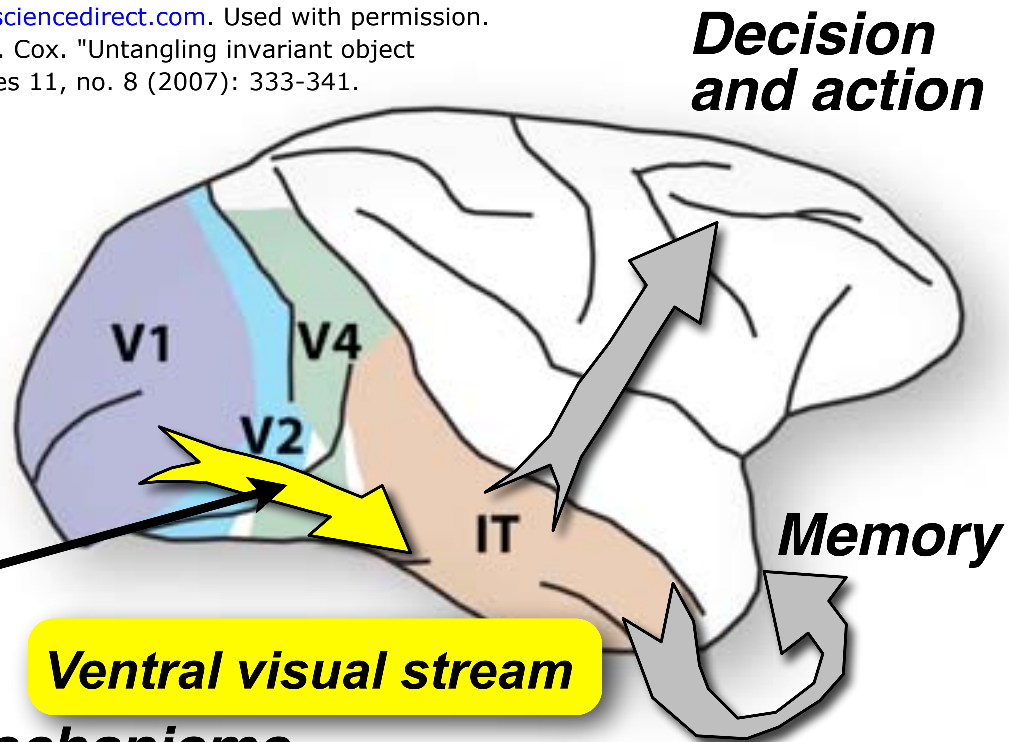
Upshot: monkey and human basic level visual object recognition behavior are statistically indistinguishable



Comparison of Object Recognition Behavior in Human and Monkey
*R. Rajalingham, K Schmidt, J.J. DiCarlo, **Vision Sciences Society** (2014)*
*R. Rajalingham, K Schmidt, J.J. DiCarlo, **J. Neuroscience** (in press)*

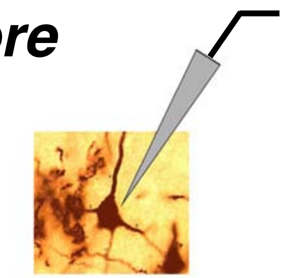
Adapted from Motter and Mountcastle 1981

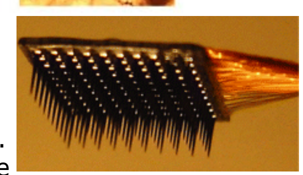
The ventral visual stream

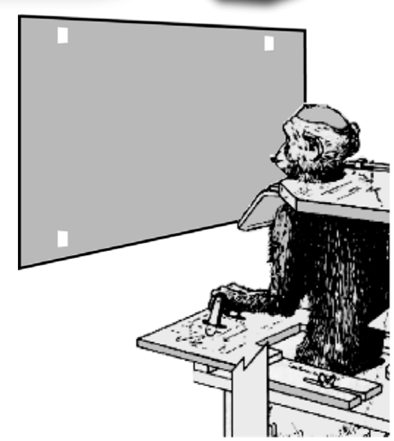


We think we know where the neural mechanisms and resulting representations that solve core object recognition live in the primate brain.

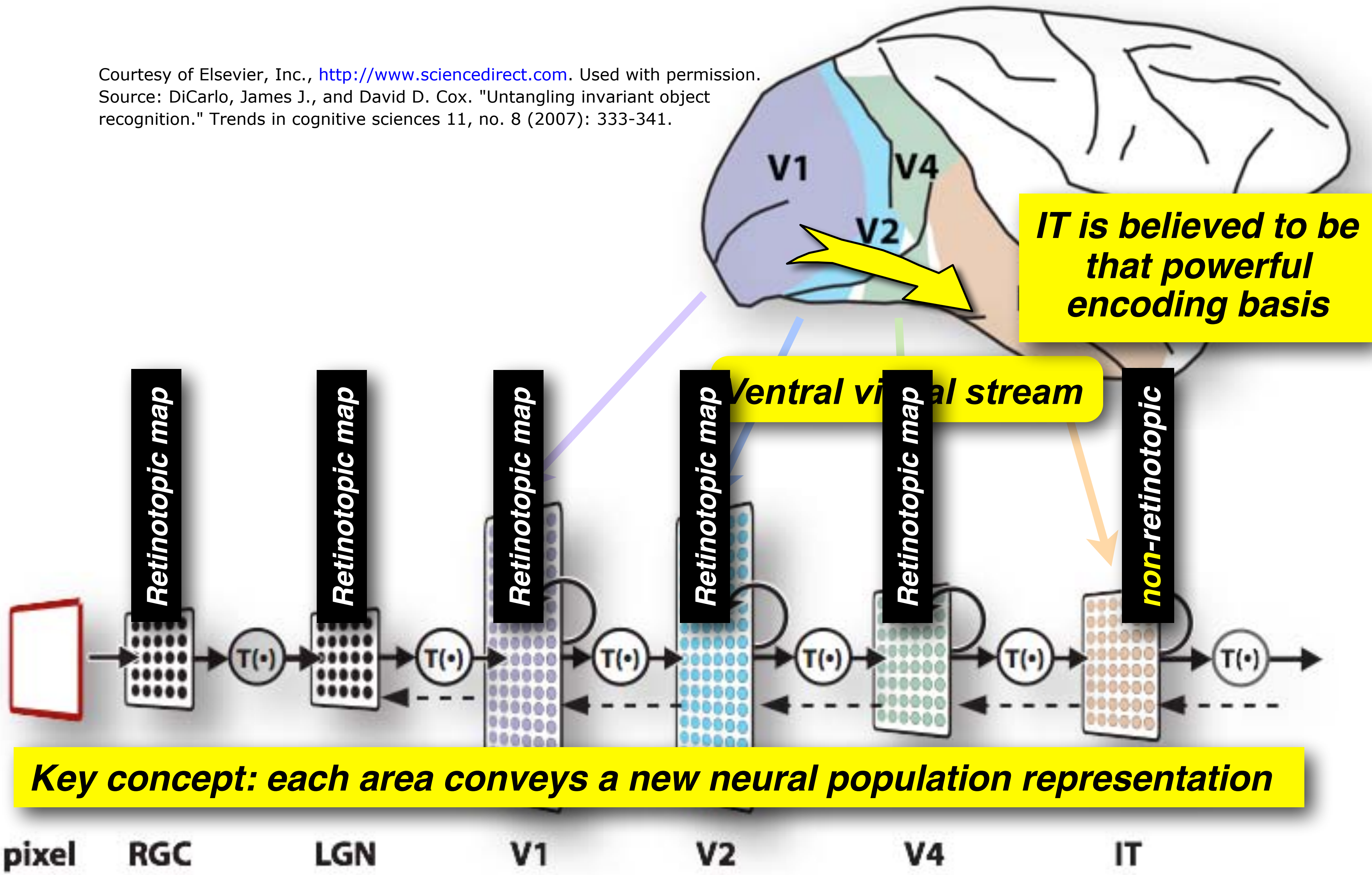
We can measure and manipulate those representations at the level of neuronal spikes.



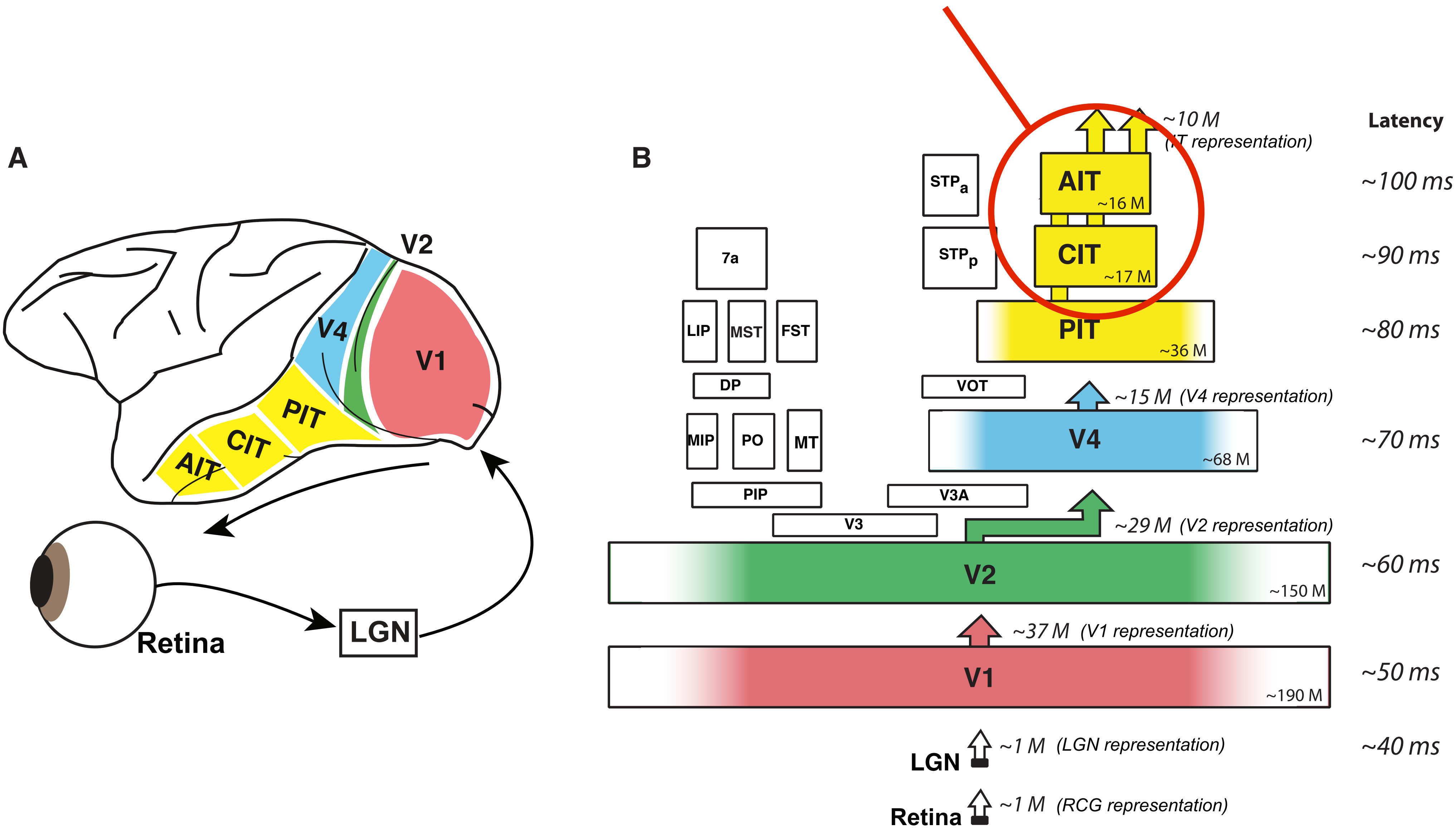



Adapted from Motter and Mountcastle 1981

The ventral visual stream



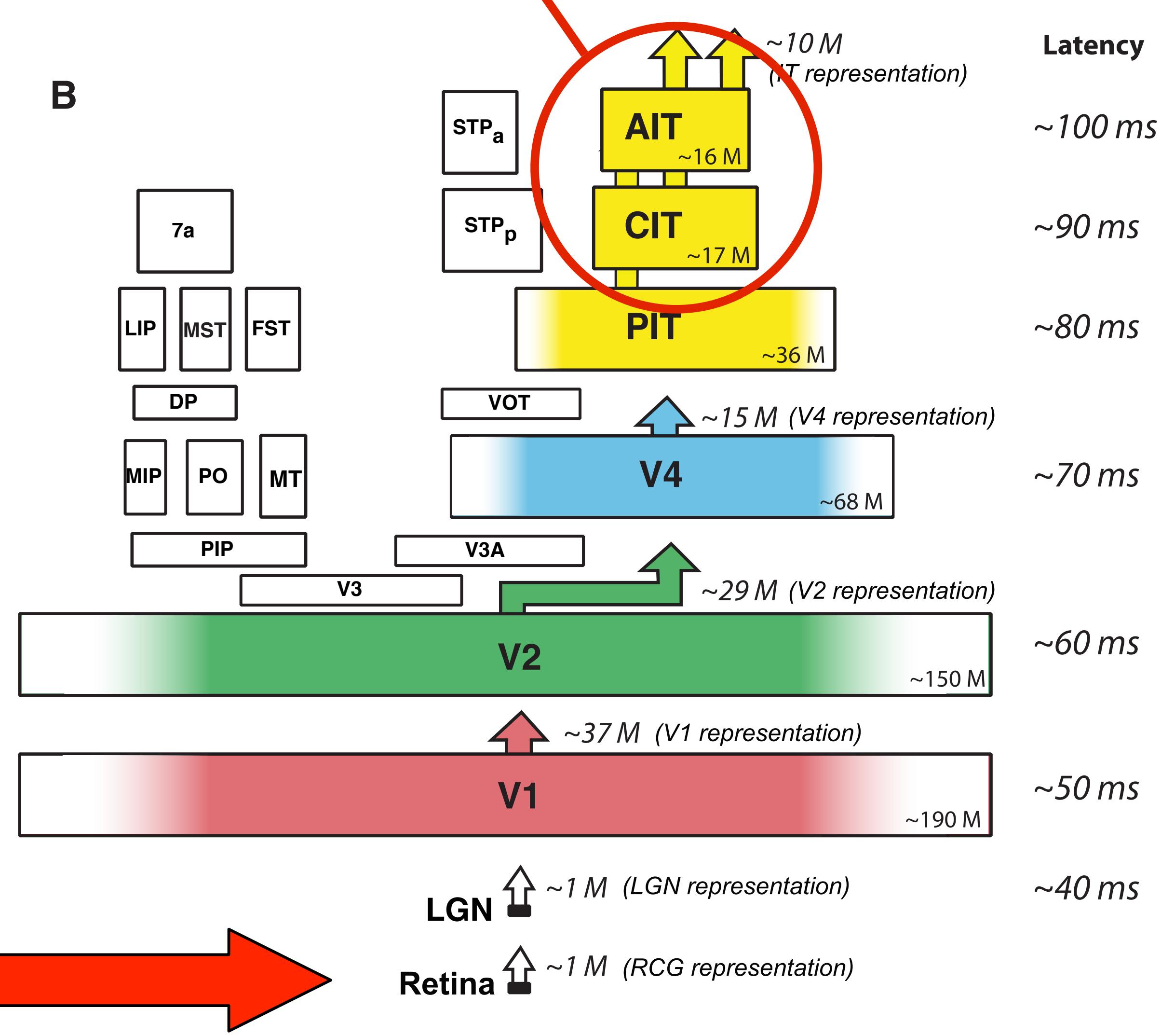
“IT” (Inferior temporal cortex)



Source: DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. "How does the brain solve visual object recognition?" Neuron 73, no. 3 (2012): 415-434.

Adapted from DiCarlo et al. 2012

“IT” (Inferior temporal cortex)



You are here.

Source: DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. "How does the brain solve visual object recognition?" Neuron 73, no. 3 (2012): 415-434.

Adapted from DiCarlo et al. 2012

Retinal ganglion cell RF structure:

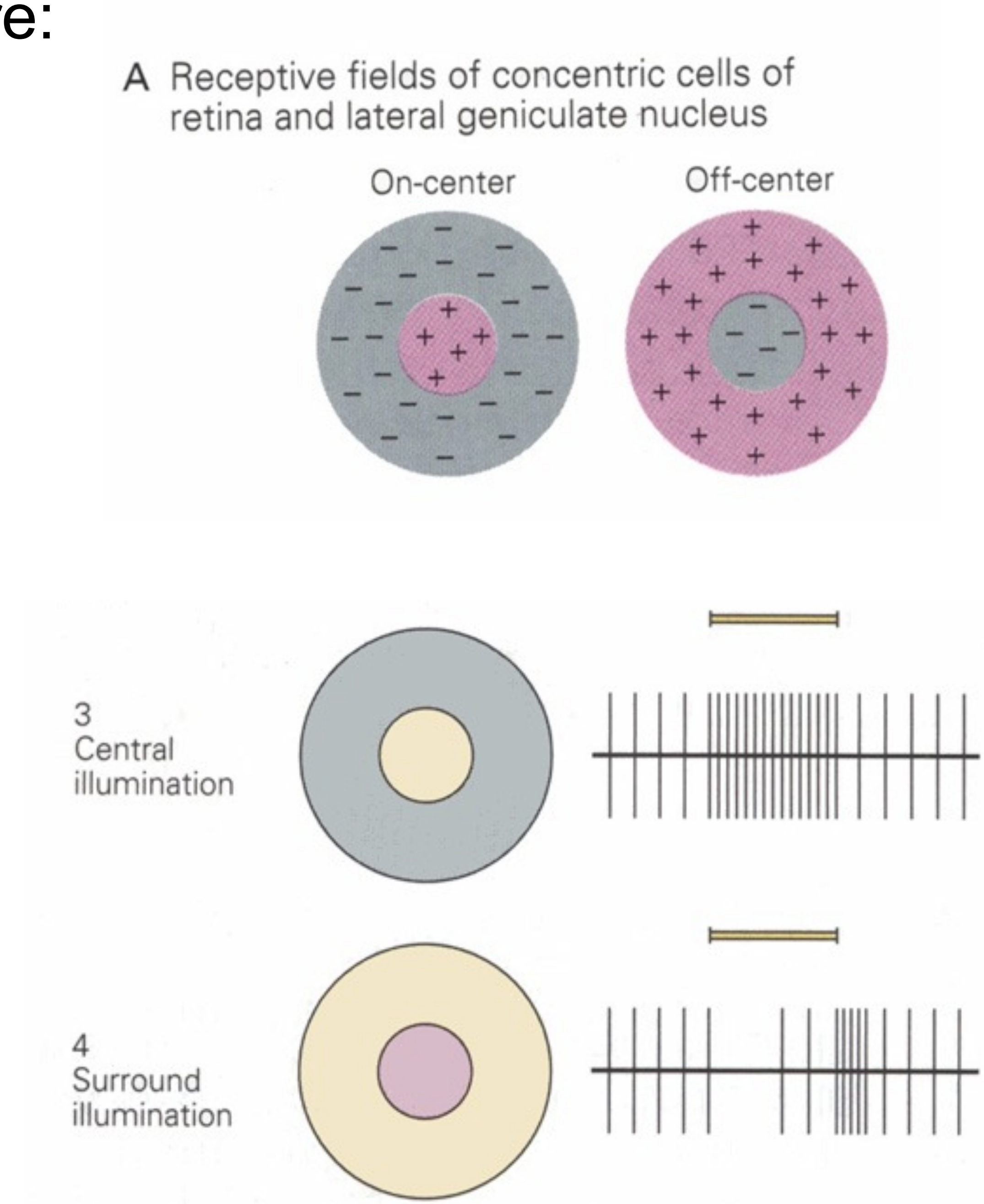


Figure removed due to copyright restrictions. Please see the video.
Source: Eye, Brain, and Vision. David H. Hubel. New York : Scientific American Library : Distributed by W.H. Freeman, c1988. ISBN: 0716750201.

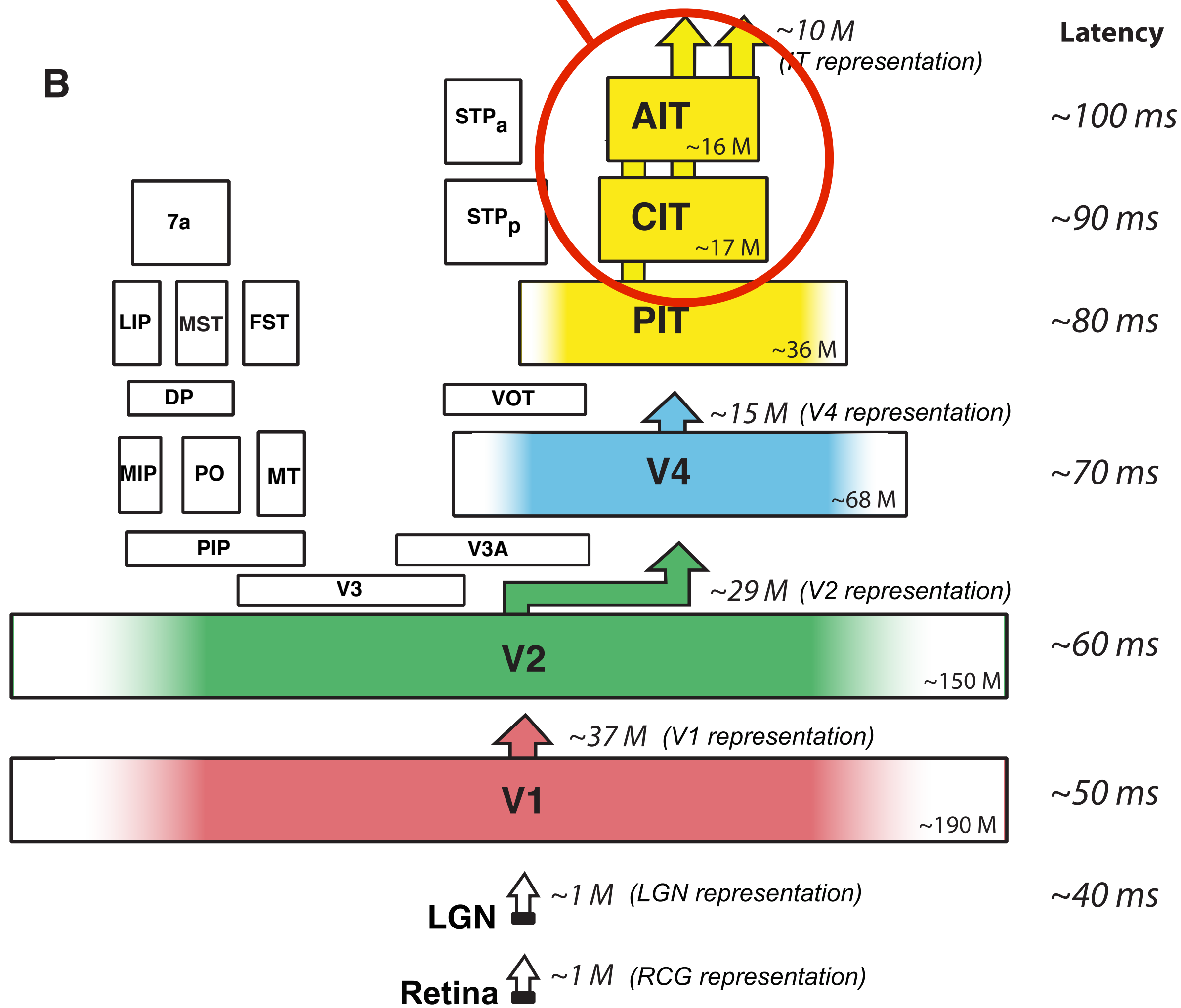
Source: Siegelbaum, Steven A., and A. James Hudspeth. Principles of neural science. Eds. Eric R. Kandel, James H. Schwartz, and ThomasM. Jessell. Vol. 4. New York: McGraw-hill, 2000.

Adapted from Hubel

Adapted from Kandel , Schwartz and Jessell

"IT" (Inferior temporal cortex)

You are here.



Source: DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. "How does the brain solve visual object recognition?" Neuron 73, no. 3 (2012): 415-434.

Adapted from DiCarlo et al. 2012

Primary visual cortex (Area V1):

Orientation selectivity

Figure removed due to copyright restrictions. Please see the video.
Source: Eye, Brain, and Vision. David H. Hubel. New York : Scientific American Library: Distributed by W.H. Freeman, c1988. ISBN: 0716750201.

Orientation selectivity with some position tolerance

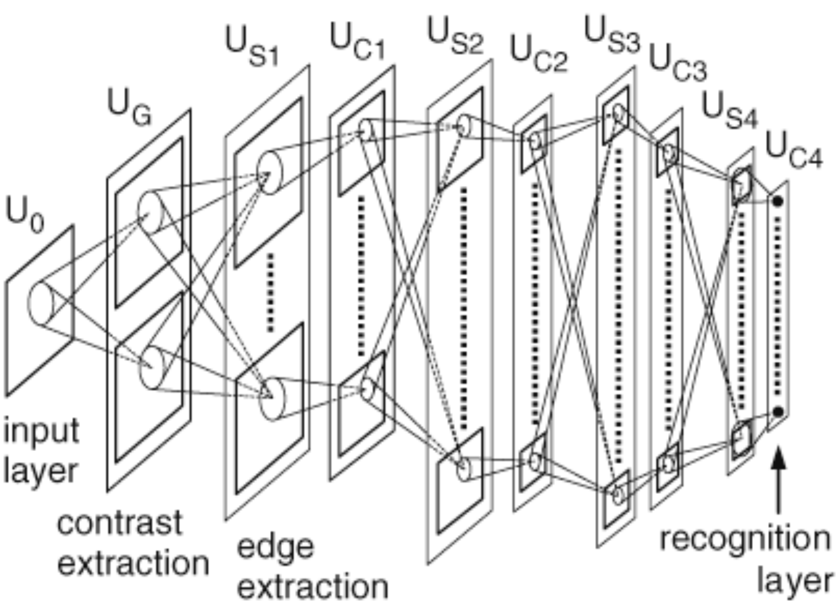
Adapted from Kandel , Schwartz and Jessell

Brain-inspired computer algorithms

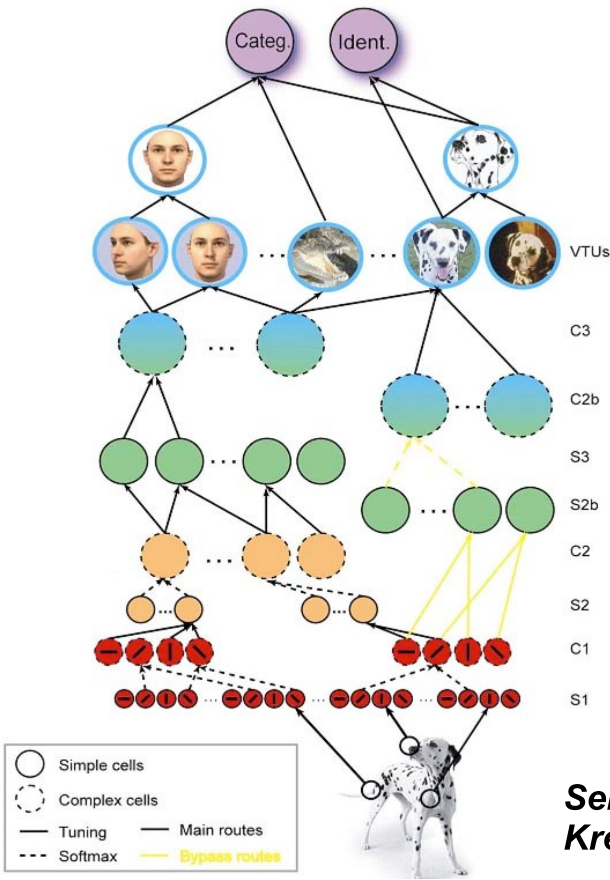
Examples:

- *Hubel & Wiesel (1962)*

Figure removed due to copyright restrictions.
Please see the video. Source: Eye, Brain, and Vision. David H. Hubel. New York: Scientific American Library: Distributed by W.H. Freeman, c1988. ISBN: 0716750201.



Source: Fukushima, Kunihiko. "Neocognitron for handwritten digit recognition. "Neurocomputing 51 (2003): 161-180.



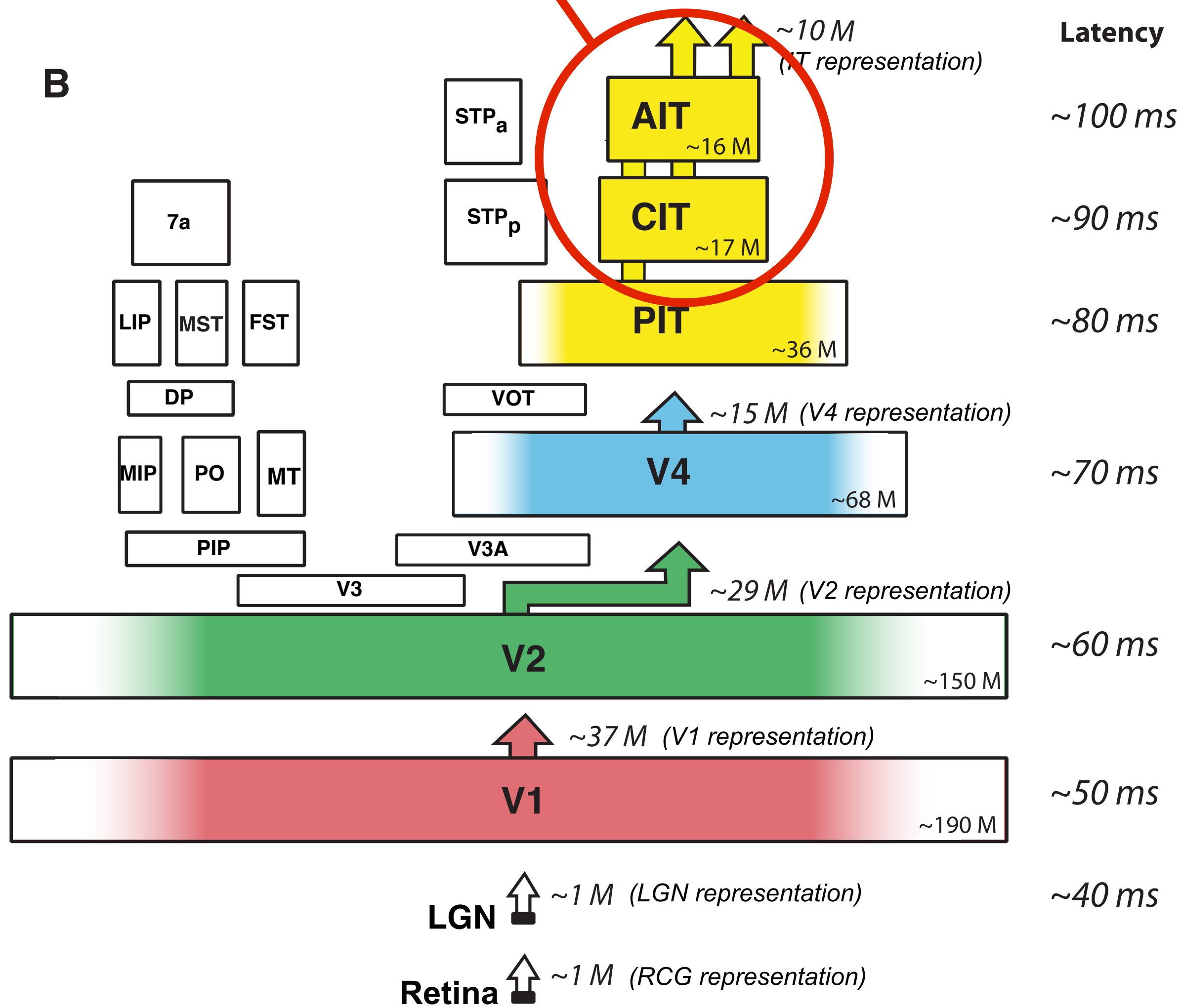
FROM BIOLOGY:

- ***Hierarchy***
- ***Spatially local filters***
- ***Convolution***
- ***Normalization***
- ***Threshold NL***
- ***Unsupervised learning***
- ***...***

Serre, Kouh, Cadieu, Knoblich, Kreiman & Poggio 2005

"IT" (Inferior temporal cortex)

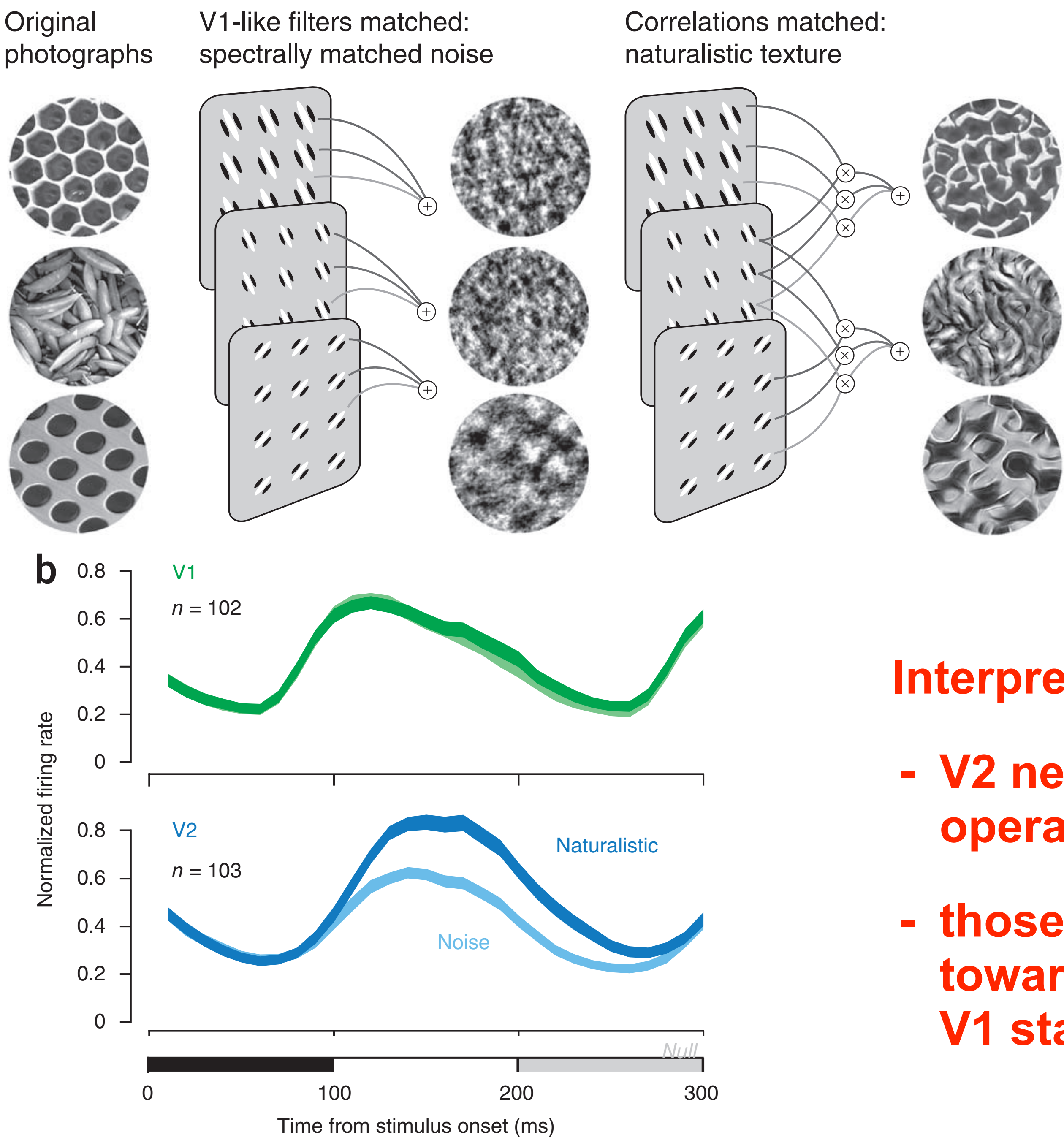
You are here.



Source: DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. "How does the brain solve visual object recognition?" Neuron 73, no. 3 (2012): 415-434.

Adapted from DiCarlo et al. 2012

Area V2 (first cortical area after V1):

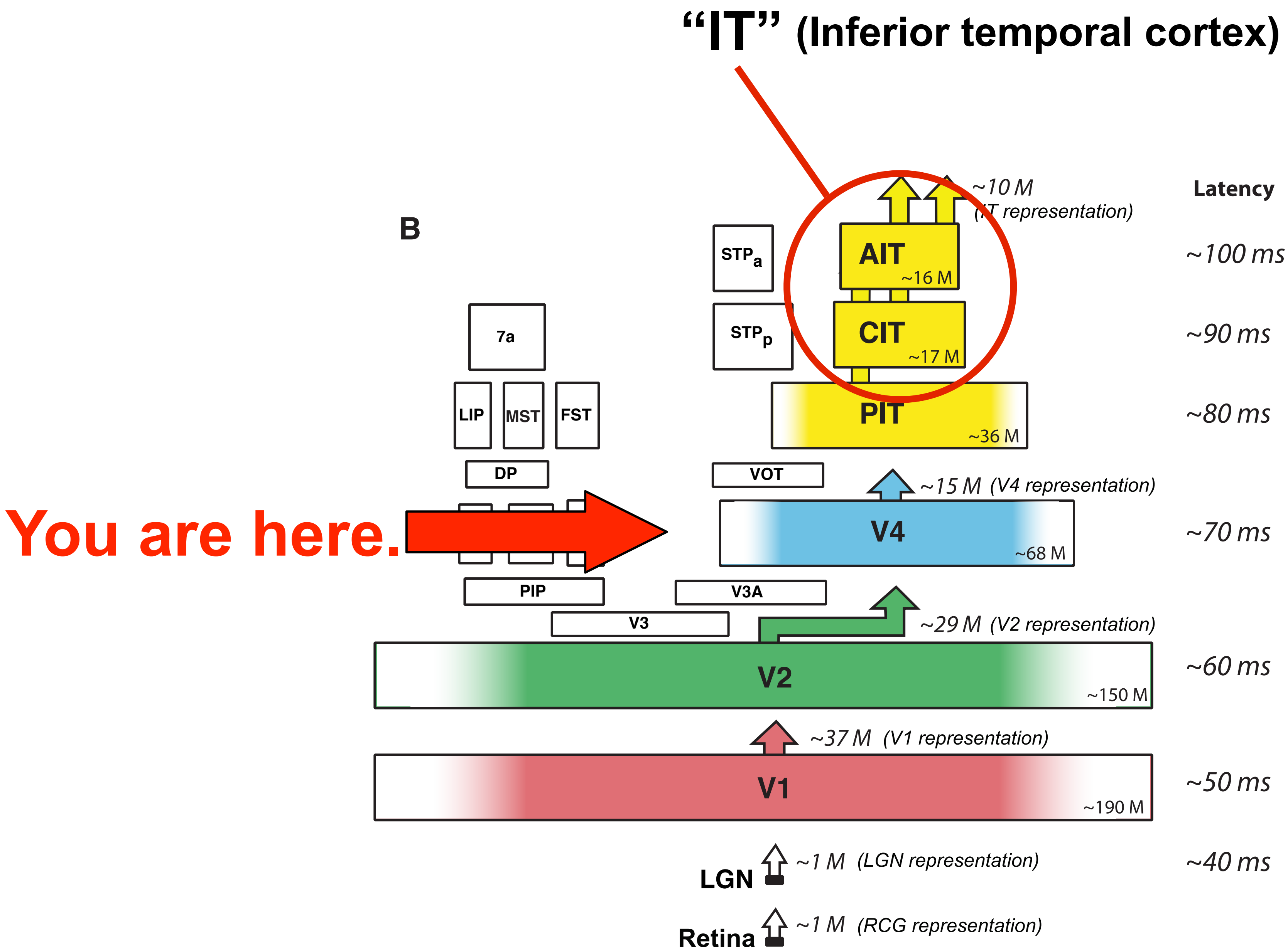


Source: Freeman, Jeremy, Corey M. Ziemba, David J. Heeger, Eero P. Simoncelli, and J. Anthony Movshon. "A functional and perceptual signature of the second visual area in primates. "Nature neuroscience 16, no. 7 (2013): 974-981.

Interpretation:

- **V2 neurons apply "and-like" operators on V1 outputs**
- **those "ands" are tuned toward natural co-occurring V1 statistics**

Adapted from Freeman, Ziemba, Heeger, Simoncelli, & Movshon, Nature Neuro (2013)

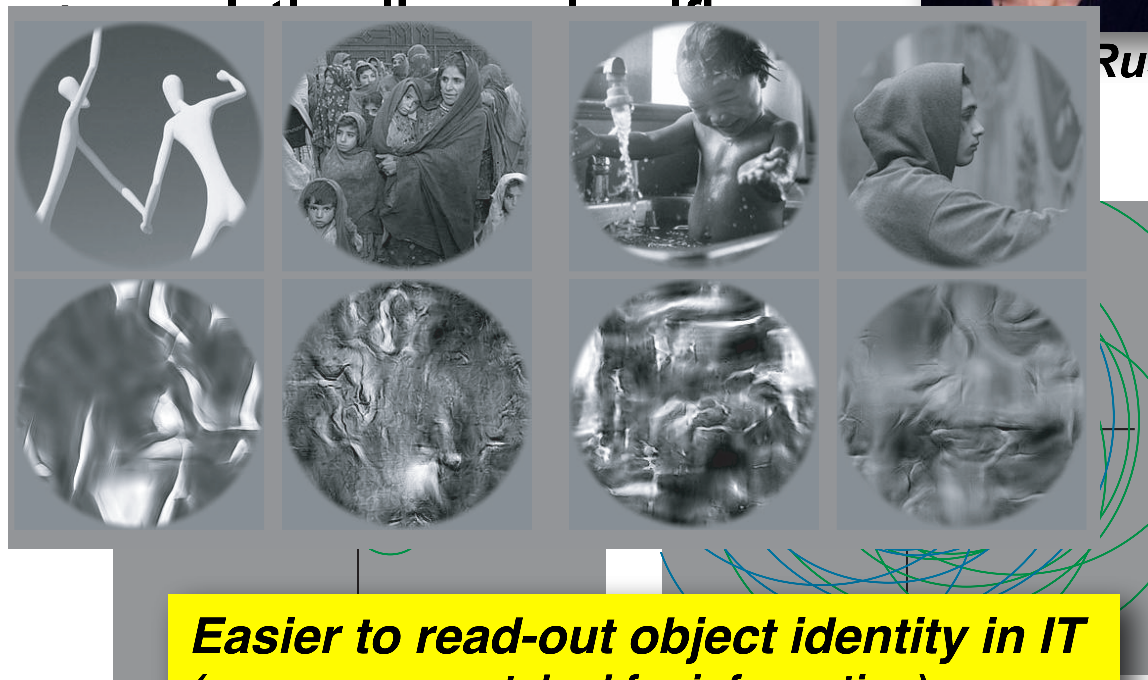


Source: DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. "How does the brain solve visual object recognition?" Neuron 73, no. 3 (2012): 415-434.

Adapted from DiCarlo et al. 2012

What is V4 doing?

Same animal, task, stimuli.

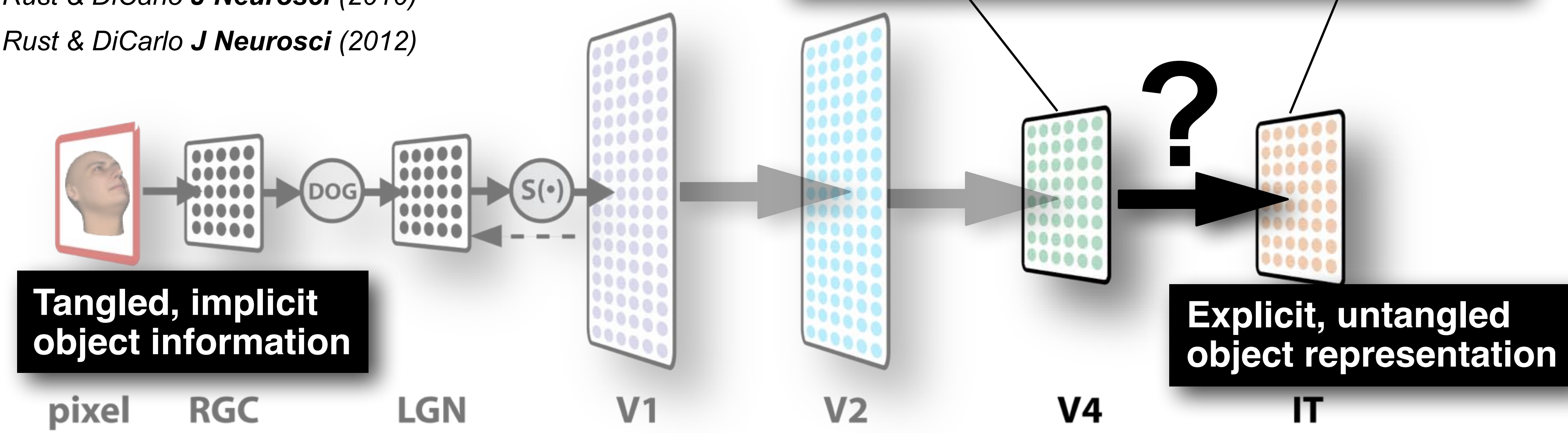


Increased selectivity for conjunction of features that tend to co-occur in natural images

Easier to read-out object identity in IT
(per neuron, matched for information)

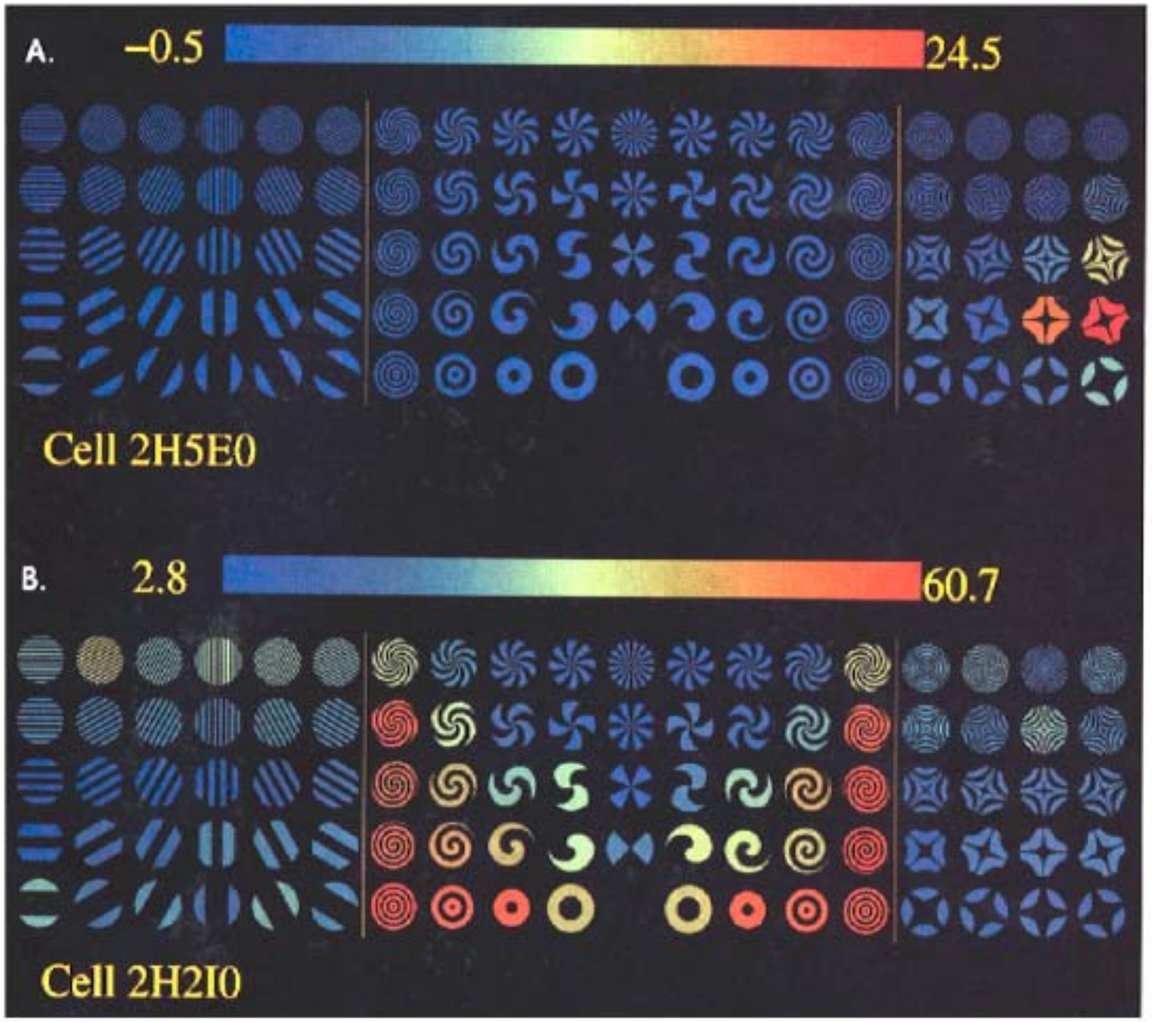
*Rust & DiCarlo **J Neurosci** (2010)*

*Rust & DiCarlo **J Neurosci** (2012)*



What is V4 doing?

V4 Responses to Non-Cartesian Gratings
Gallant et al. 1996



Source: Gallant, Jack L., Charles E. Connor, Subrata Rakshit, James W. Lewis, and DAVID C. Van Essen. "Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey." Journal of neurophysiology 76, no. 4 (1996): 2718-2739.

What shape features drive V4 responses?

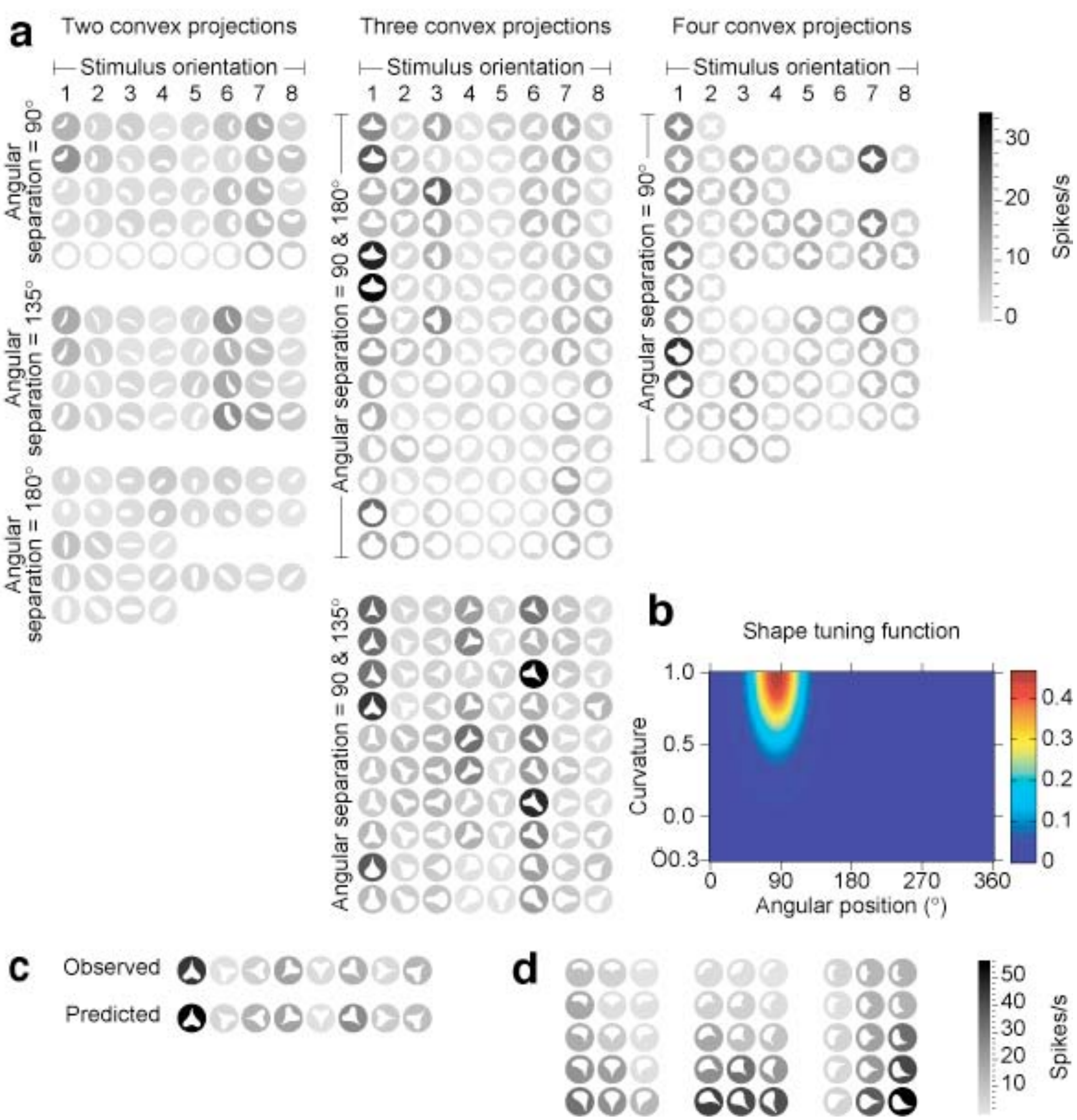
Adapted from C.E. Connor

Make a basis for shapes:
each shape = set of curved elements
each element = (ang position, curvature)

Hypothesis:
V4 neurons are tuned in this basis

What shape features drive V4 responses?

Adapted from C.E. Connor



Reprinted by permission from Macmillan Publishers Ltd: Nature Neuroscience. Source: Pasupathy, Anitha, and Charles E. Connor. "Population coding of shape in area V4." Nature neuroscience 5, no. 12 (2002): 1332-1338.

<u>Make a basis for shapes:</u>
each shape = set of curved elements
each element = (ang position, curvature)

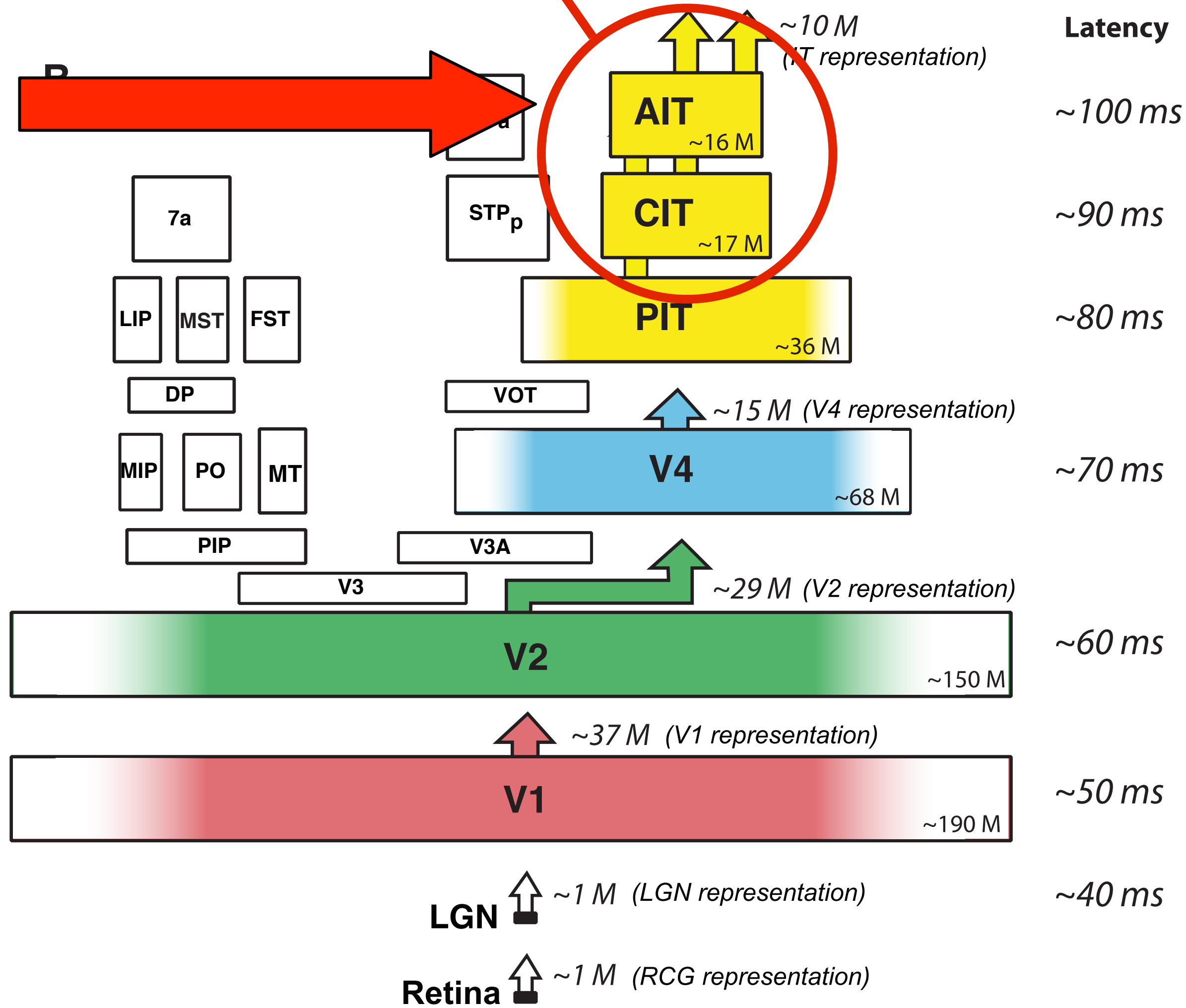
<u>Hypothesis:</u>
V4 neurons are tuned in this basis

<u>Experimental result:</u>
Hypothesis explains ~50% of the explainable response variance

Pasupathy and Connor (V4)
Brincat and Connor (PIT)

“IT” (Inferior temporal cortex)

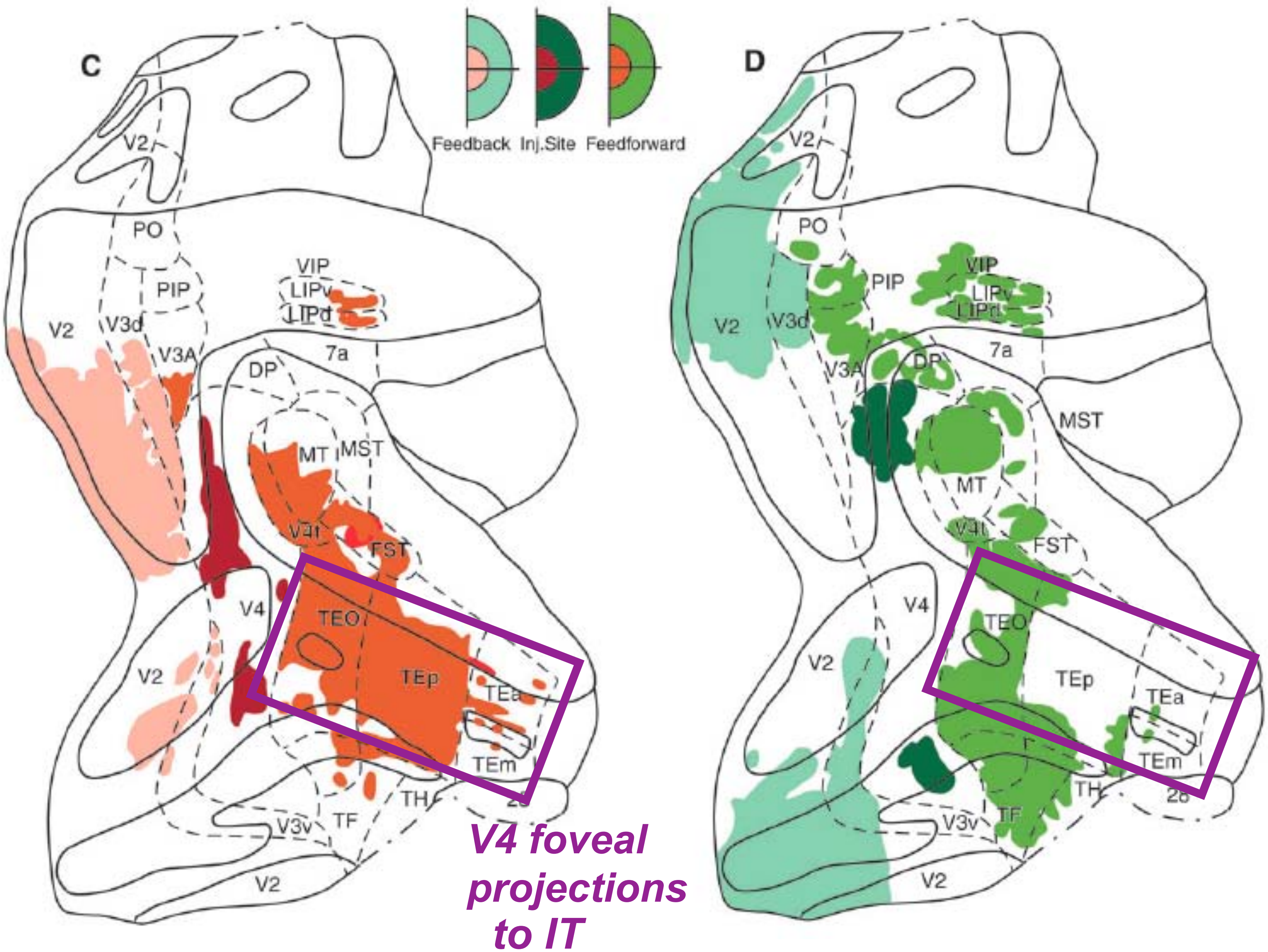
You are here.



Source: DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. "How does the brain solve visual object recognition?" Neuron 73, no. 3 (2012): 415-434.

Adapted from DiCarlo et al. 2012

IT is about central vision



Source: Ungerleider, Leslie G., Thelma W. Galkin, Robert Desimone, and Ricardo Gattass. "Cortical connections of area V4 in the macaque." Cerebral Cortex 18, no. 3 (2008): 477-499.

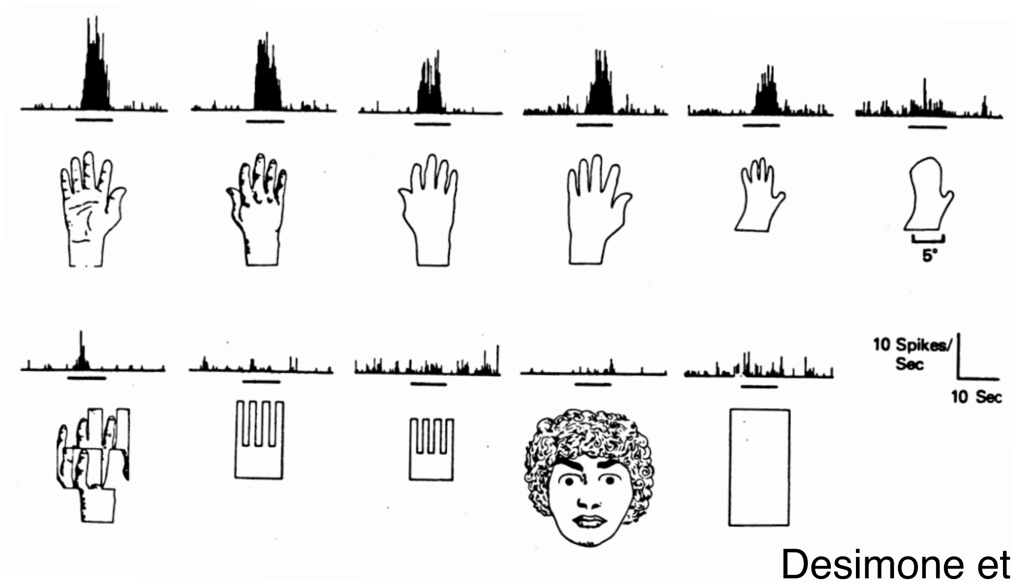
Ungerleider, L. G. et al. Cereb. Cortex 2007

Stimulus selectivity in inferotemporal cortex
Gross, Rocha-Miranda & Bender 1972

The use of [these] stimuli was begun one day when, having failed to drive a unit with any light stimulus, we waved a hand at the stimulus screen and elicited a very vigorous response from the previously unresponsive neuron...

We then spent the next 12 hr testing various paper cutouts in an attempt to find the trigger feature for this unit. When the entire set of stimuli used were ranked according to the strength of the response that they produced, we could not find a simple physical dimension that correlated with this rank order. However, the rank order of adequate stimuli did correlate with similarity (for us) to the shadow of a monkey hand" (Gross et al., 1972).

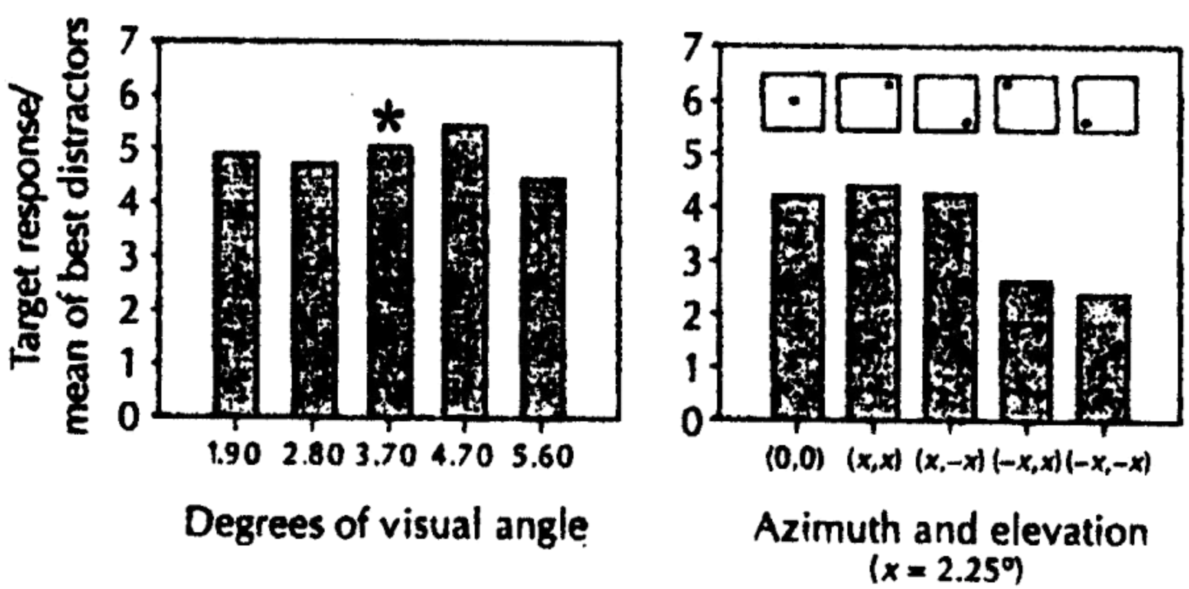
The ventral stream and object recognition



Desimone et al. (1984)

Source: Desimone, Robert, Thomas D. Albright, Charles G. Gross, and Charles Bruce. "Stimulus-selective properties of inferior temporal neurons in the macaque." Journal of Neuroscience 4, no. 8 (1984): 2051-2062.

IT neurons can be tuned to specific combinations of features (high "selectivity")



That selectivity is tolerant to changes in position and size

Source: Castiello, Umberto. "Mechanisms of selection for the control of hand action. Trends in Cognitive Sciences 3, no. 7 (1999): 264-271.

Logothetis et al. (1995)

Primary visual cortex:

Orientation selectivity

Figure removed due to copyright restrictions. Please see the video.
Source: Eye, Brain, and Vision. David H. Hubel. New York : Scientific American Library: Distributed by W.H. Freeman, c1988. ISBN: 0716750201.

Orientation selectivity with some position tolerance

Adapted from Kandel , Schwartz and Jessell

What stimulus feature are IT neurons actually “tuned” to?

Figure removed due to copyright restrictions. Please see the video.
Source: Tanaka, Keiji. "Neuronal mechanisms of object recognition." Science-New York Then Washington 262 (1993): 685-685.

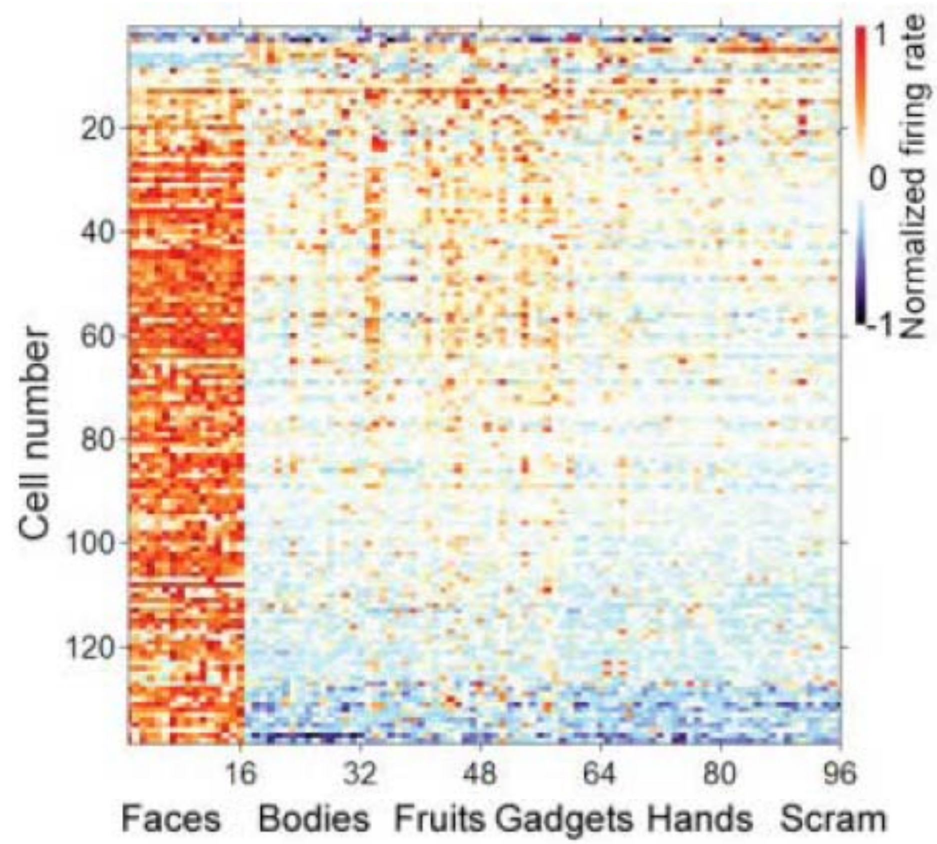
Figure removed due to copyright restrictions. Please see the video.
Source: Tanaka, Keiji. "Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities." Cerebral cortex 13, no. 1 (2003): 90-99. doi: 10.1093/cercor/13.1.90.

IT has spatial organization at 500 um - 1 mm scale

Larger scale (2-6 mm) organization for some image contrasts

Figure removed due to copyright restrictions. Please see the video.

ML



Source: Tsao, Doris Y., Winrich A. Freiwald, Roger BH Tootell, and Margaret S. Livingstone. "A cortical region consisting entirely of face-selective cells." Science 311, no. 5761 (2006): 670-674.

Tsao, Freiwald, and Livingstone used fMRI to reveal a set of face selective regions in IT (aka "face patches")

Most of the single neurons in these regions showed a preference for frontal faces

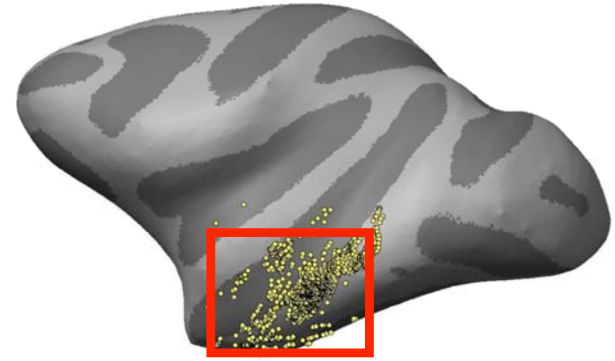
Tsao et al., *Science* 2006

IT selectivity is particularly clustered for some image contrasts

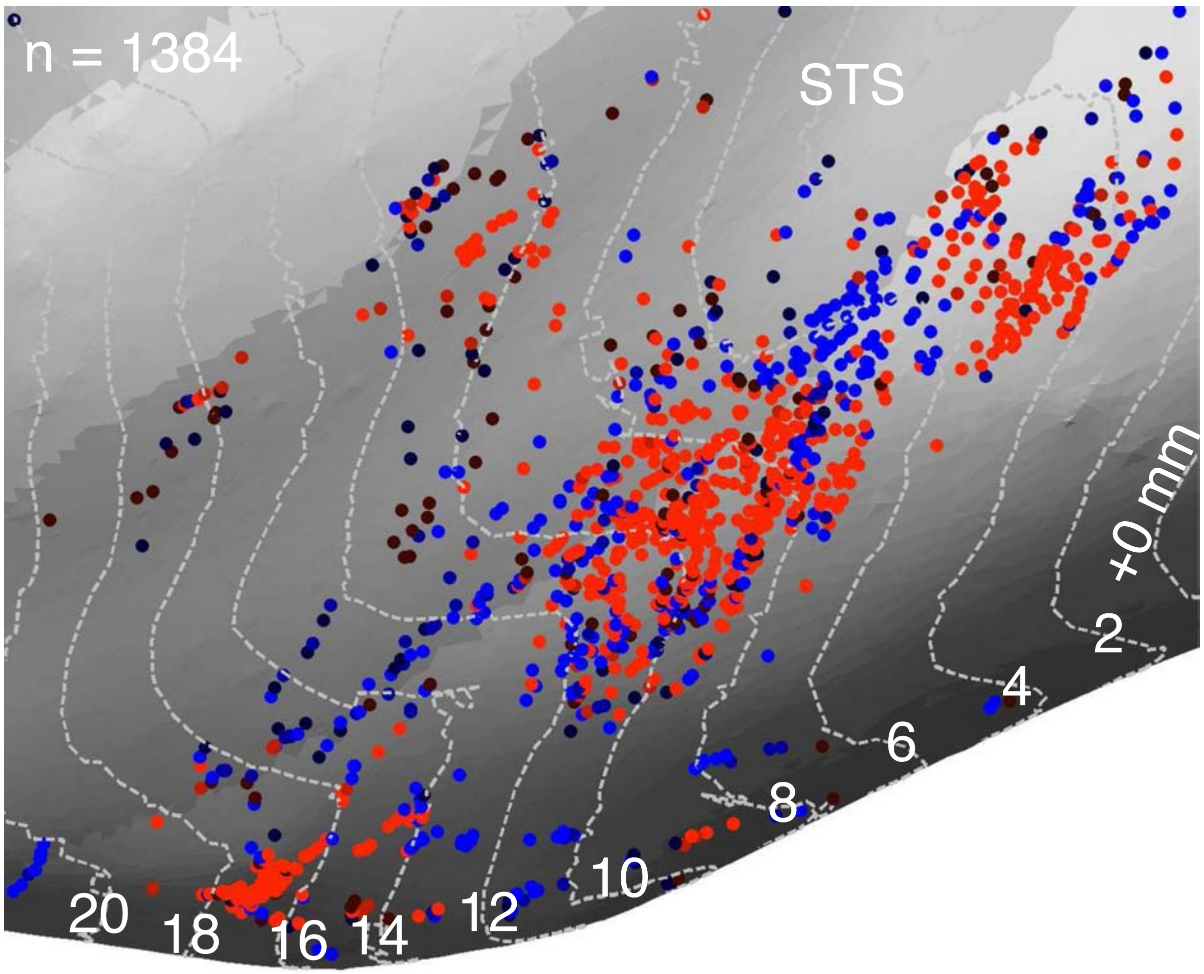
vs

face objects

non-face objects



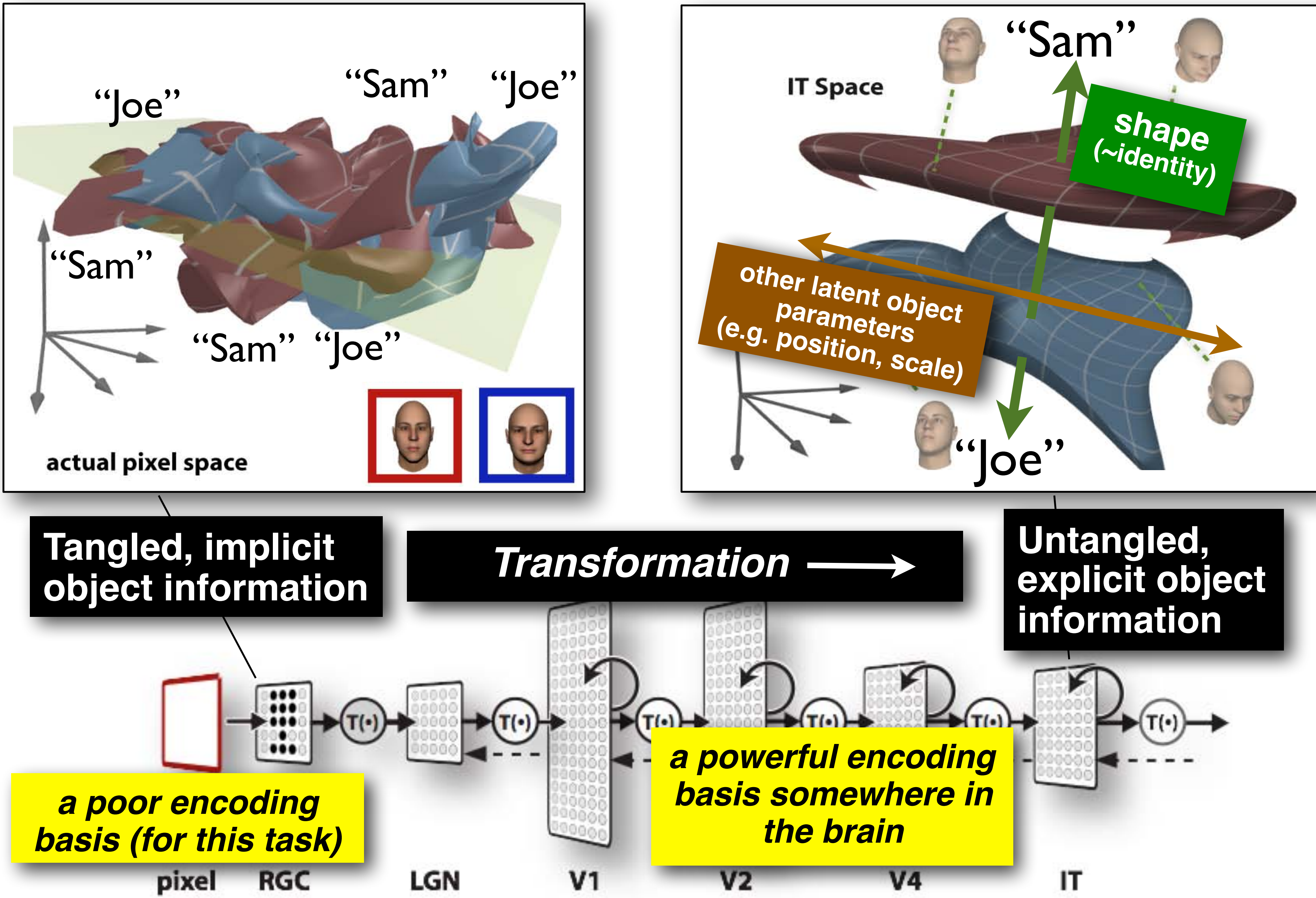
MUA



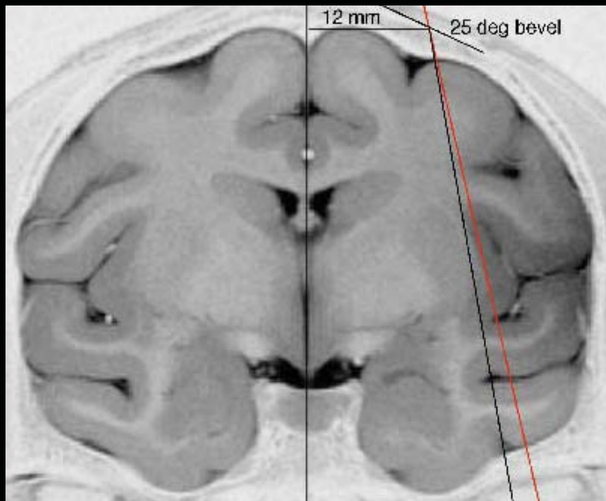
Source: Issa, Elias B., Alex M. Papanastassiou, and James J. DiCarlo. "Large-scale, high-resolution neurophysiological maps underlying FMRI of macaque temporal lobe." Journal of Neuroscience 33, no. 38 (2013): 15207-15219.

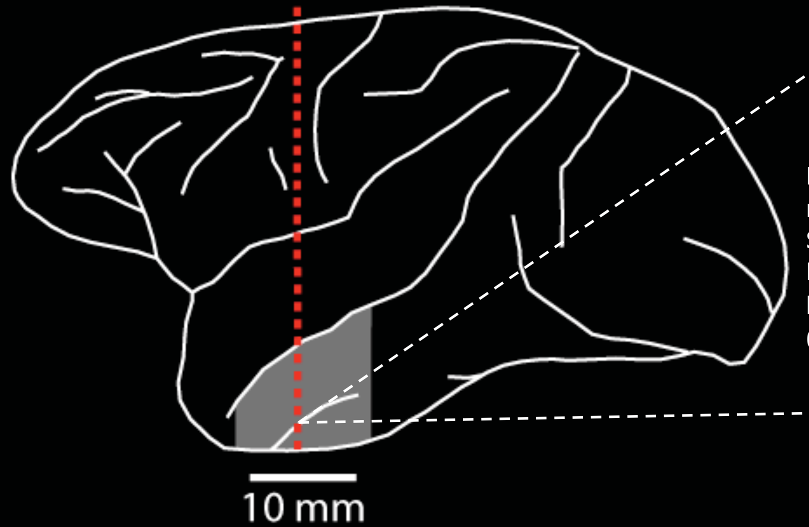
Issa et al., *J Neurosci 2013*
Aparacio*, Issa*, DiCarlo *(In prep)*

*DiCarlo and Cox, **TICS** (2007)*
*DiCarlo, Zoccolan and Rust, **Neuron** (2012)*



Example spiking activity in IT



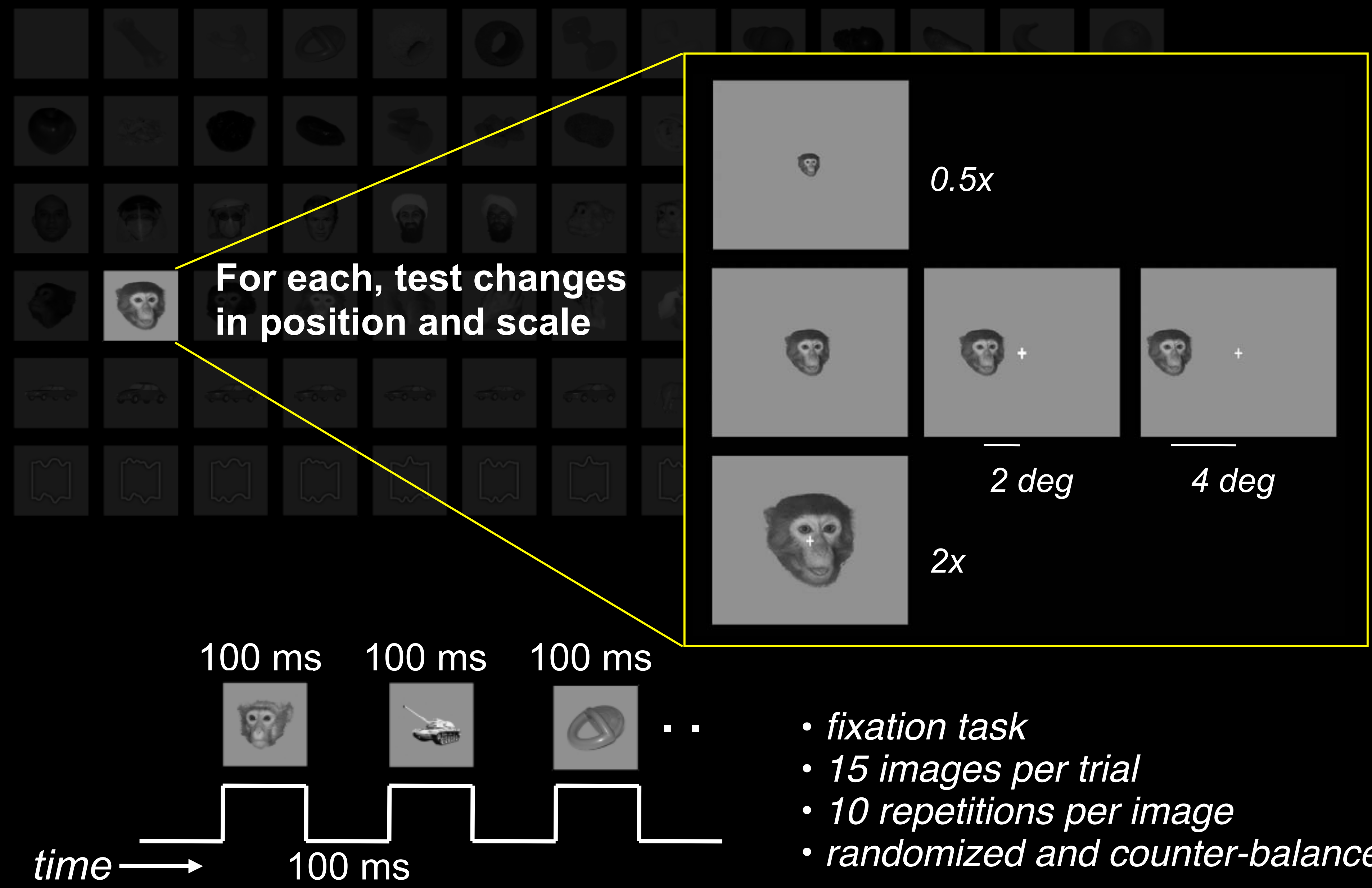


Site 1

0 100
ms

An early test of the IT population

A broad set of 78 test objects from eight categories …

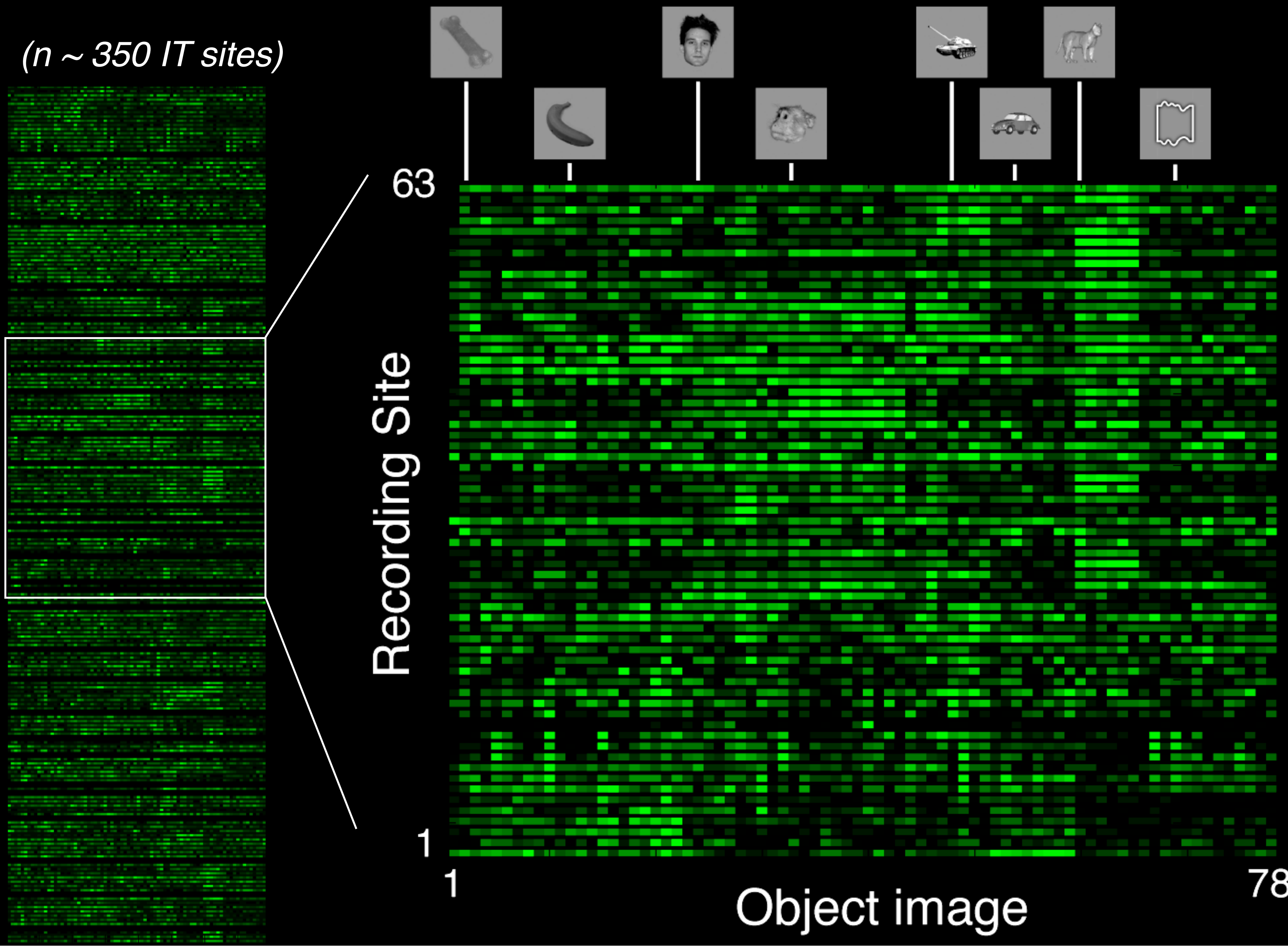


- *fixation task*
- *15 images per trial*
- *10 repetitions per image*
- *randomized and counter-balanced*

Hung, Kreiman*, Poggio and DiCarlo,* ***Science*** *(2005)*

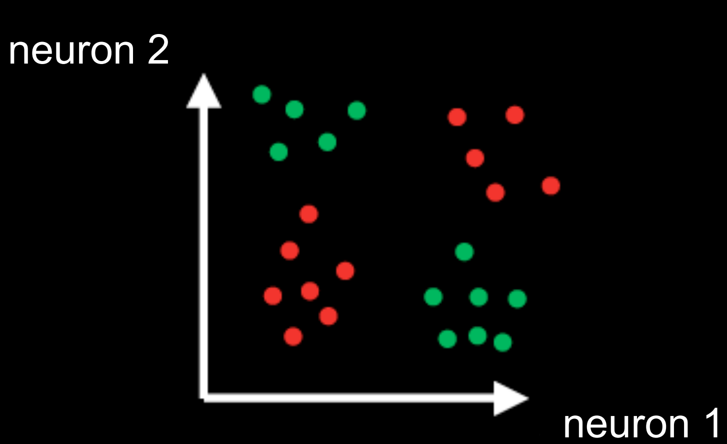
Source: Hung, Chou P., Gabriel Kreiman, Tomaso Poggio, and James J. DiCarlo. "Fast readout of object identity from macaque inferior temporal cortex. "Science 310, no. 5749 (2005): 863-866.

The "mean" IT population



How do we test if the population image is "good"?

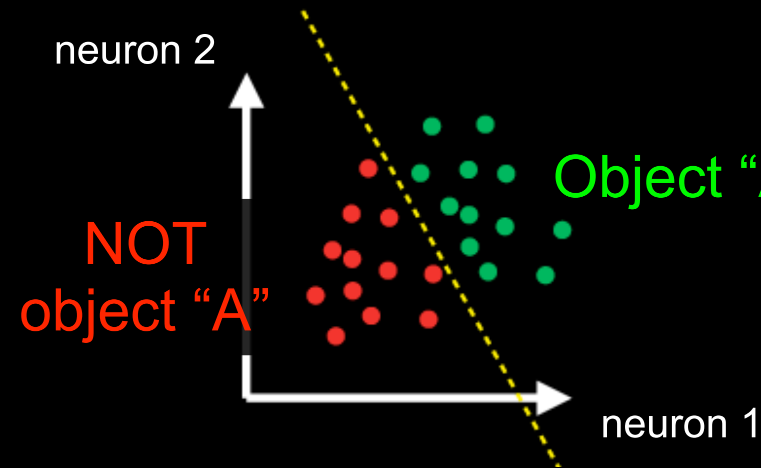
Implicit representation



*"**inaccessible**"*
object information

BAD

Explicit representation

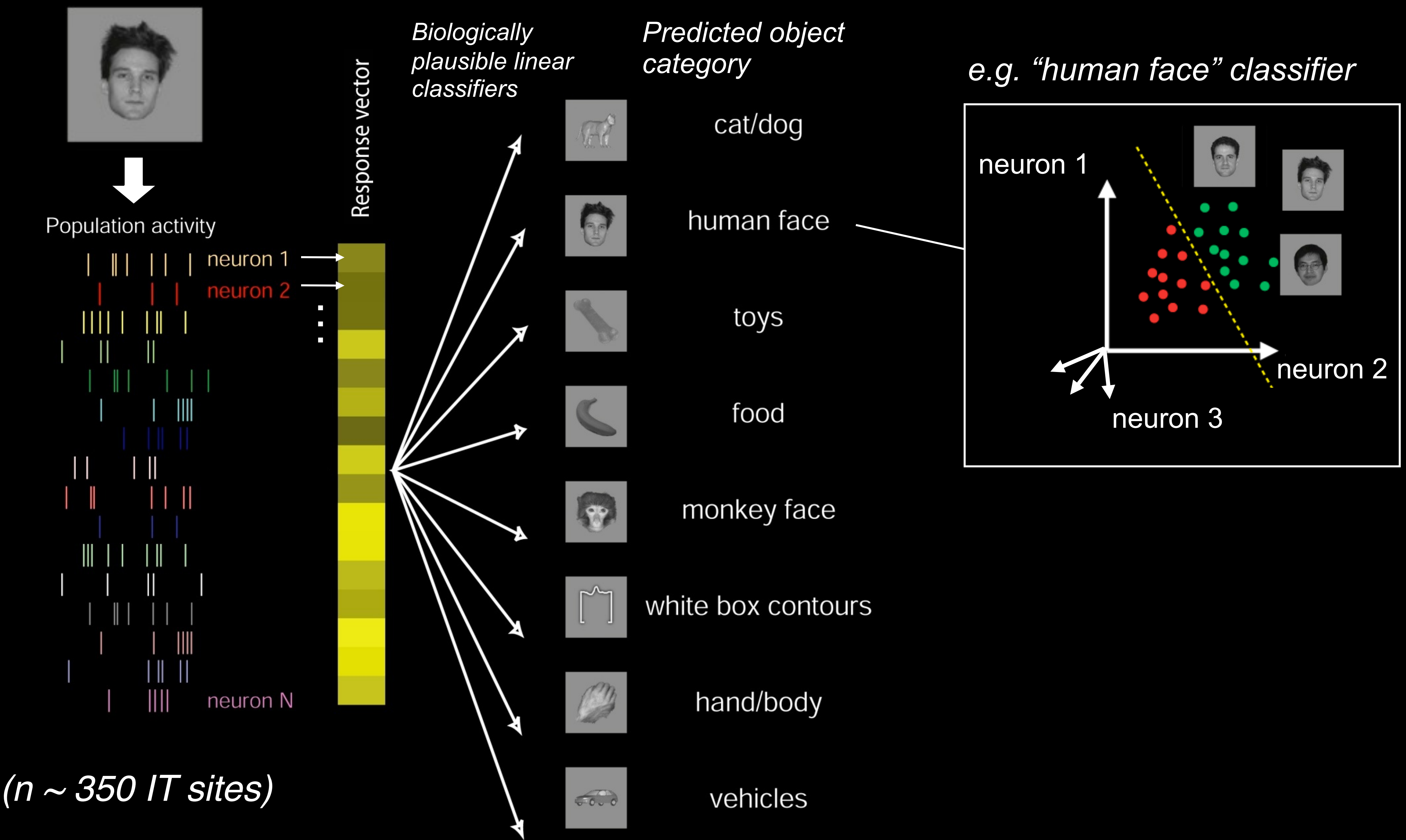


Linearly separable

*"**accessible**"*
object information

GOOD

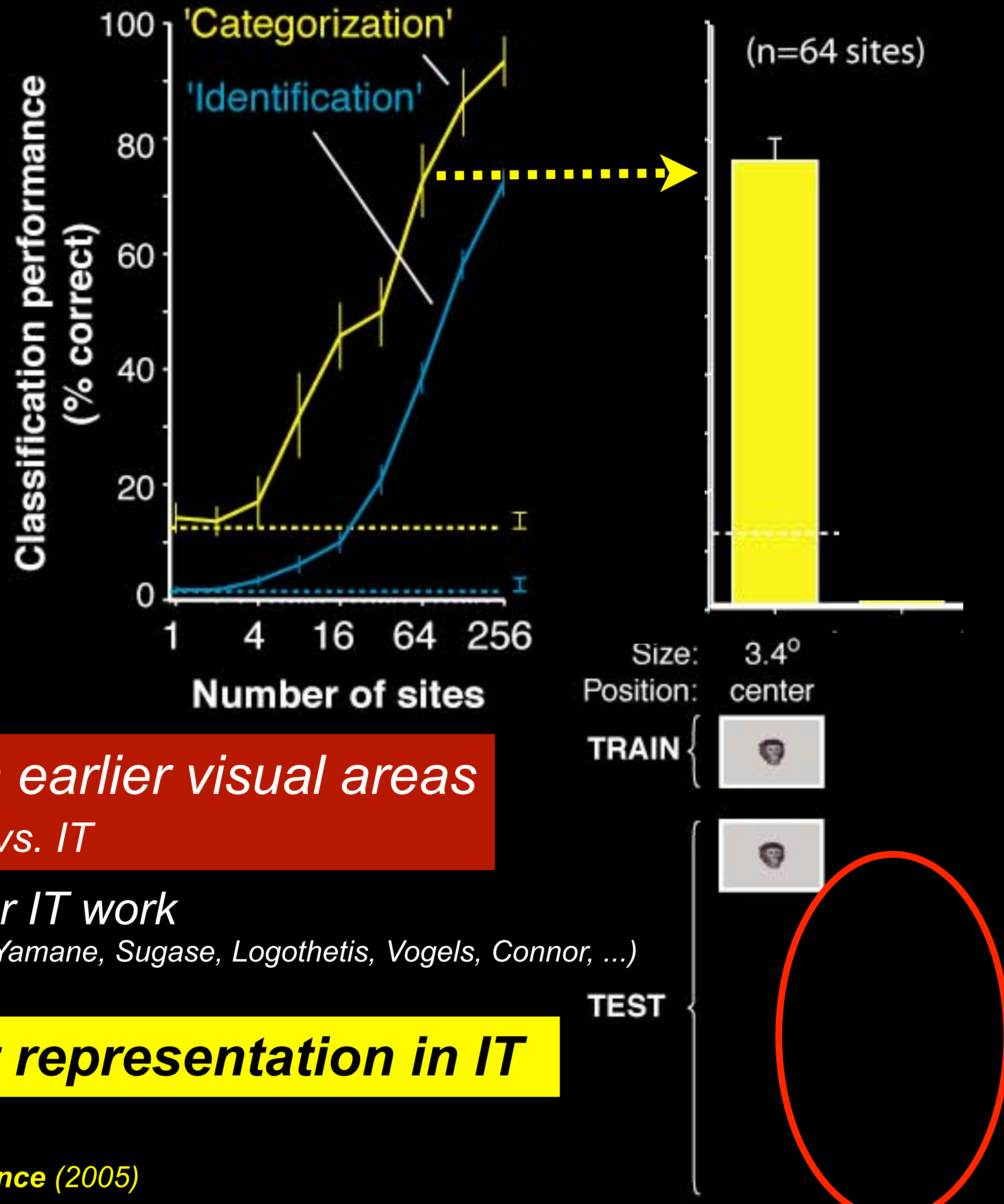
How explicit ("good") is object information in IT?



(n ~ 350 IT sites)

Hung, Kreiman*, Poggio and DiCarlo,* ***Science*** *(2005)*

Explicit object information in IT ?



Does not work in earlier visual areas
e.g. V1 vs. IT or V4 vs. IT

- *Consistent with other IT work*
  (e.g. Rolls, Tanaka, Miyashita, Yamane, Sugase, Logothetis, Vogels, Connor, ...)

Rapid, explicit object representation in IT

Hung, Kreiman*, Poggio and DiCarlo,* ***Science*** *(2005)*

Summary so far:

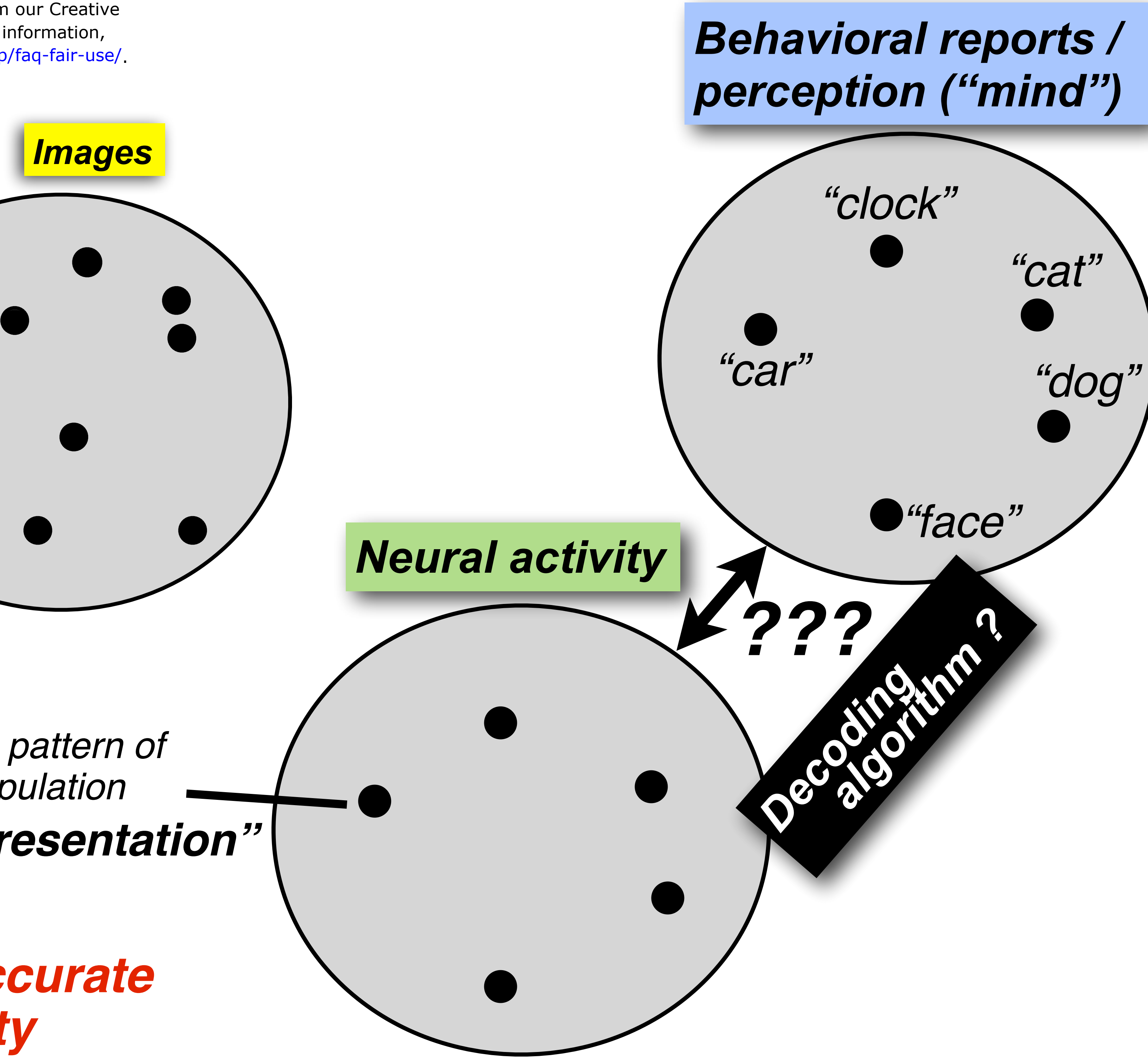
the problem of visual object recognition

a tour of the ventral stream

IT population seems to have solved a key problem

Over the last 40 years. we (the field) have largely described important phenomenology

Next phase of this field: developing and testing predictive models



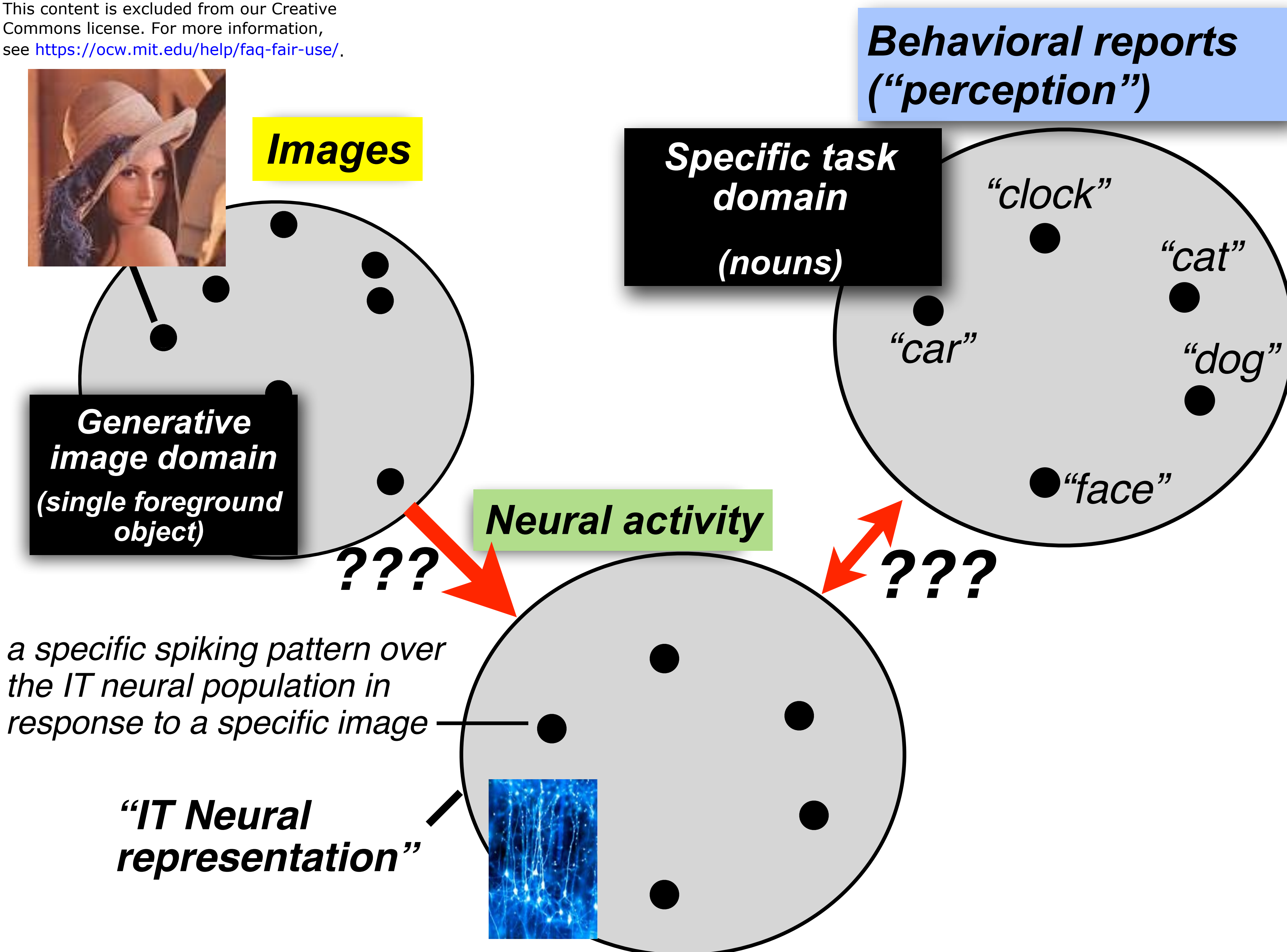
Goal is accurate predictivity

(Domain: core object recognition)

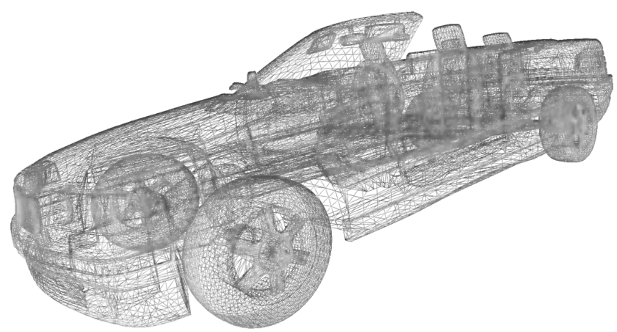
Goal: end-to-end understanding

1. Can we infer the precise decoding mechanism(s) that the brain uses to support perceptual reports about visually presented objects?

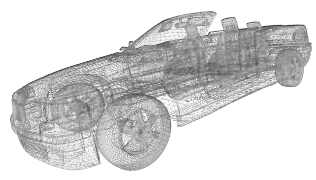
2. Can we infer the encoding mechanism(s) that accurately predicts the relevant ventral stream population patterns of neural activity from each image?


Images
Generative image domain
(single foreground object)
???
Behavioral reports ("perception")
Specific task domain
(nouns)
"clock"
"cat"
"car"
"dog"
"face"
Neural activity
???
a specific spiking pattern over the IT neural population in response to a specific image
"IT Neural representation"

3-d object Models
(e.g. “car”)



experimenter-chosen view parameters



+

Position
Size
Pose

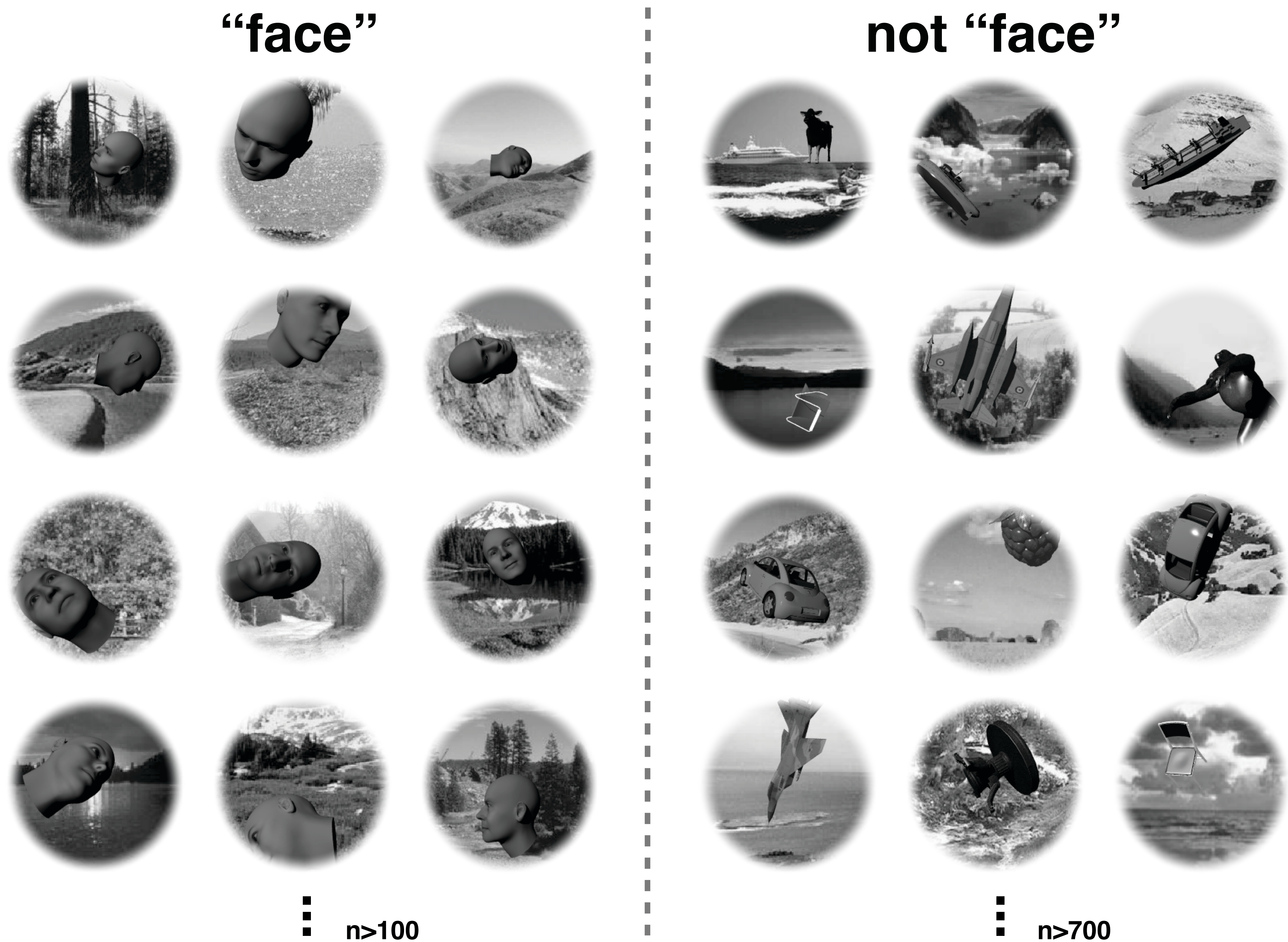
ray-trace render

place on a randomly-chosen background image

- generative space of images, each with a single foreground object and experimenter-known viewing parameters.
- uncorrelated, new background every image
  ==> challenging for computer vision, doable by humans

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

One example core object recognition test:



Another example core object recognition test:

“Beetle”

n>100

Not “Beetle”

n>700

(Domain: core object recognition)

Goal: end-to-end understanding

1. Can we infer the decoding mechanism(s) that the brain uses to support perceptual reports about visually presented objects?

> **Note: this must predict behavioral report and it must include a falsifiable statement of the relevant aspects of neural activity (aka "neural code")**

2. Can we infer the encoding mechanism(s) that accurately predicts the relevant ventral stream population patterns of neural activity from each image?

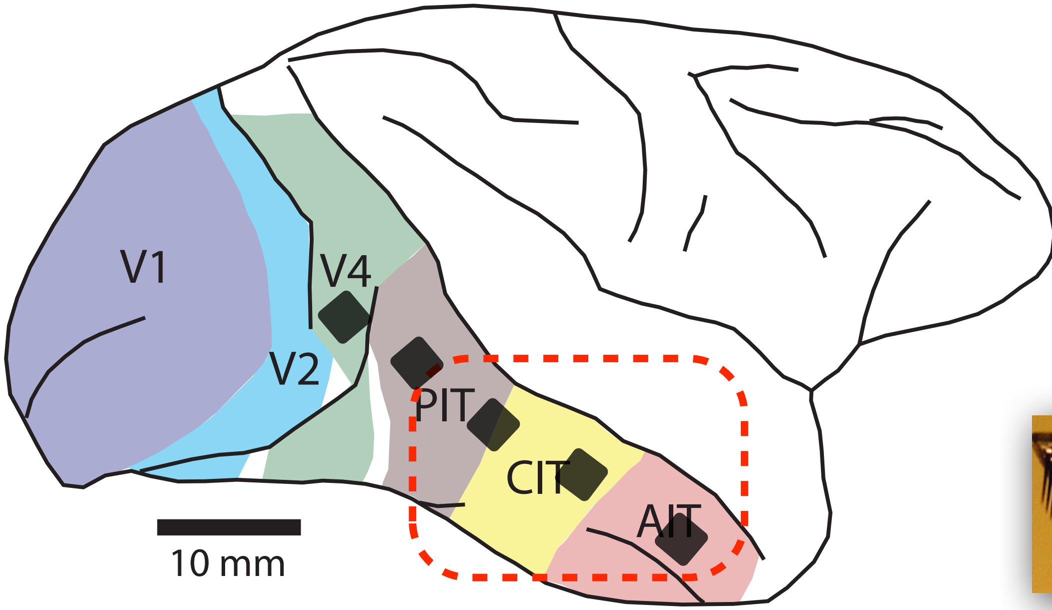
(Domain: core object recognition)

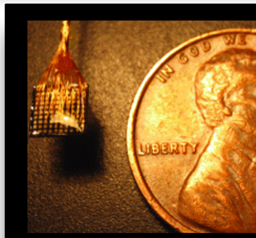
Goal: end-to-end understanding

1. Can we infer the decoding mechanism(s) that the brain uses to support perceptual reports about visually presented objects?

Note: this must predict behavioral report and it must include a falsifiable statement of the relevant aspects of neural activity (aka “neural code”)

Simultaneous recording of hundreds of neural sites along the ventral stream

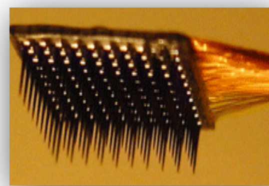




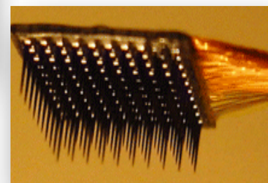
Adapted from Kelly et al. J. Neurosci (2007)

Three, 96-electrode arrays

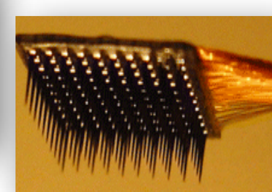
Source: Kelly, Ryan C., Matthew A. Smith, Jason M. Samonds, Adam Kohn, A. B. Bonds, J. Anthony Movshon, and Tai Sing Lee. "Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex." Journal of Neuroscience 27, no. 2 (2007): 261-264.



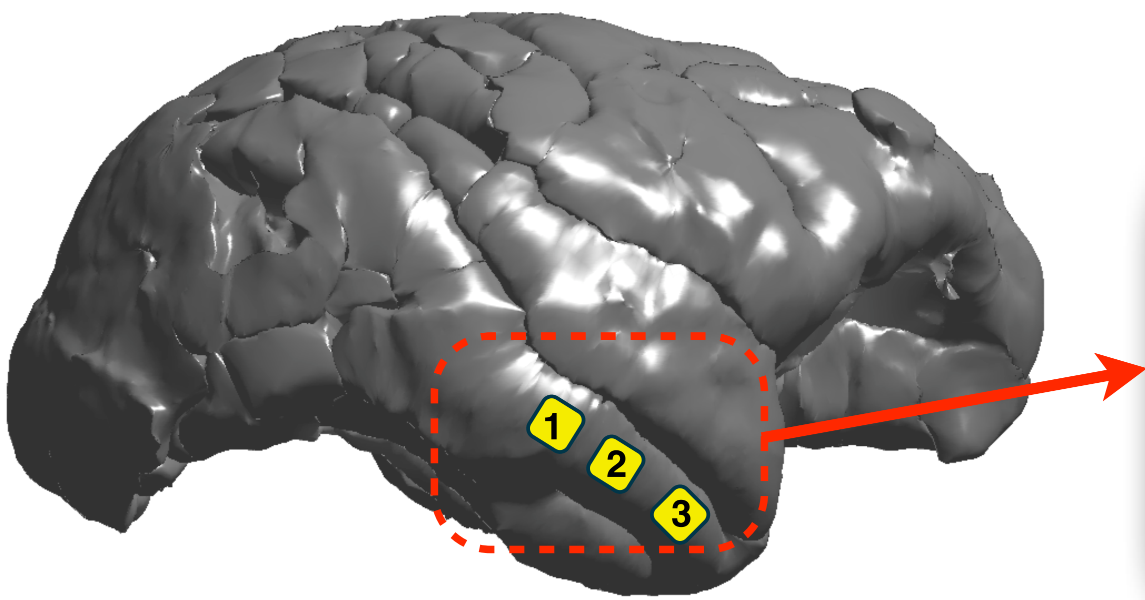
Array 1
location

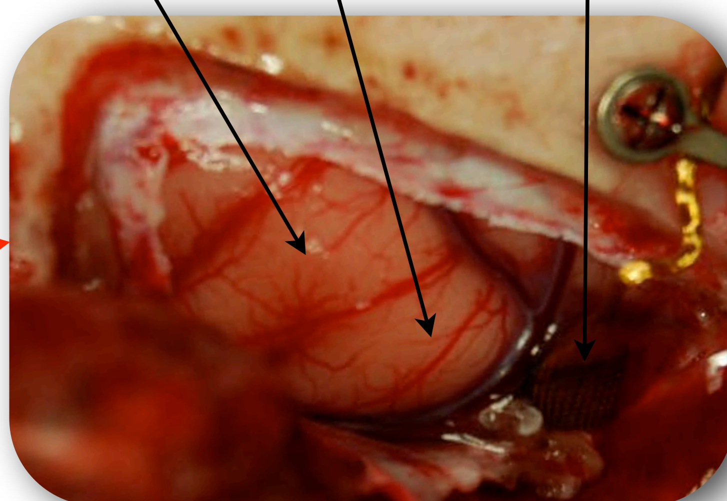


Array 2
location

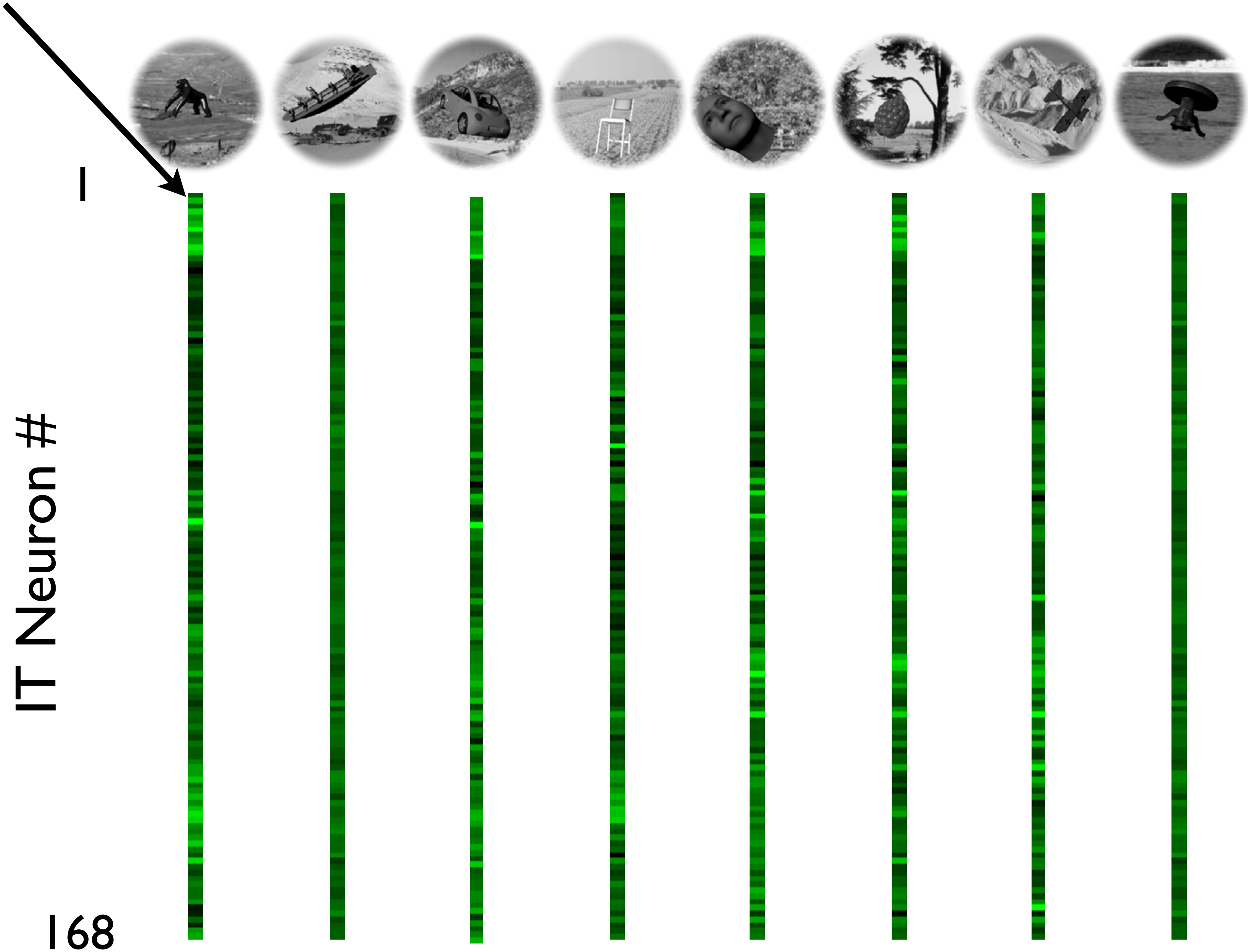


Array 3
(in place)

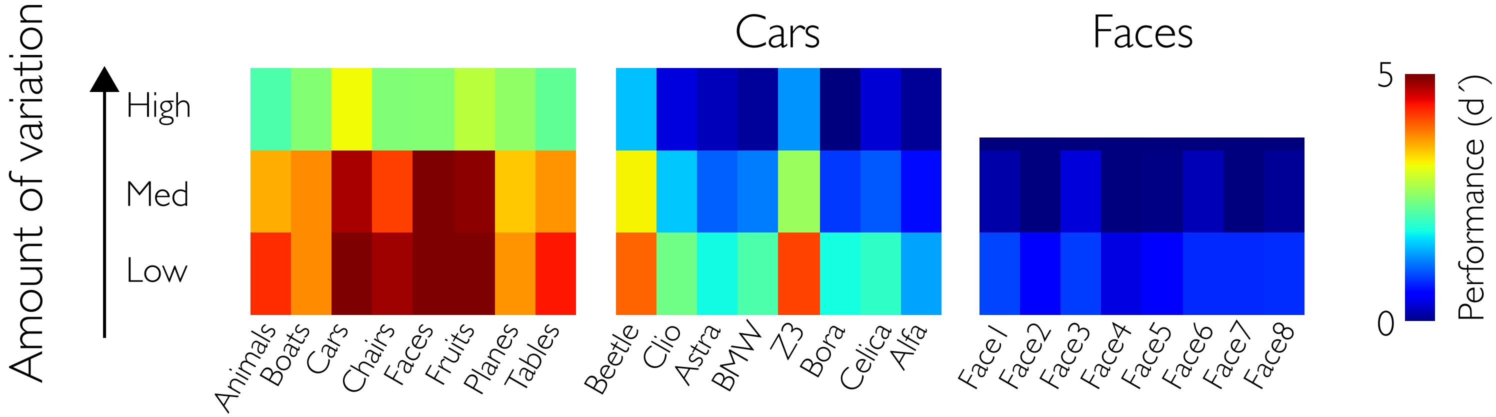




e.g. "response" = mean firing rate 70-170 ms after image onset

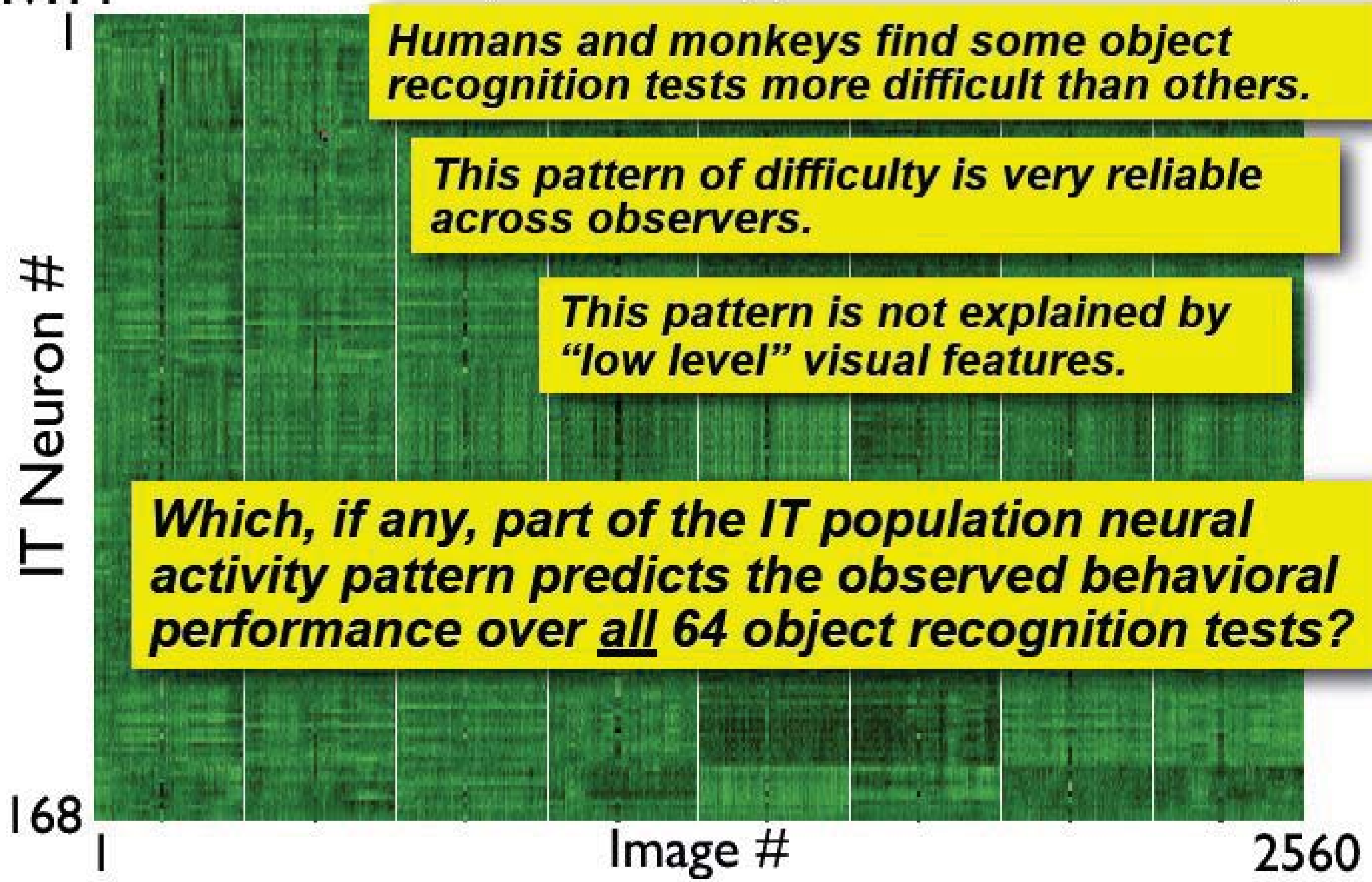


BEHAVIOR
(64 object recognition tests using same images)



Source: Majaj, Najib J., Ha Hong, Ethan A. Solomon, and James J. DiCarlo. "Simple earned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance." Journal of Neuroscience 35, no. 39 (2015): 13402-13418.

NEURAL ACTIVITY



Humans and monkeys find some object recognition tests more difficult than others.

This pattern of difficulty is very reliable across observers.

This pattern is not explained by "low level" visual features.

Which, if any, part of the IT population neural activity pattern predicts the observed behavioral performance over <u>all</u> 64 object recognition tests?

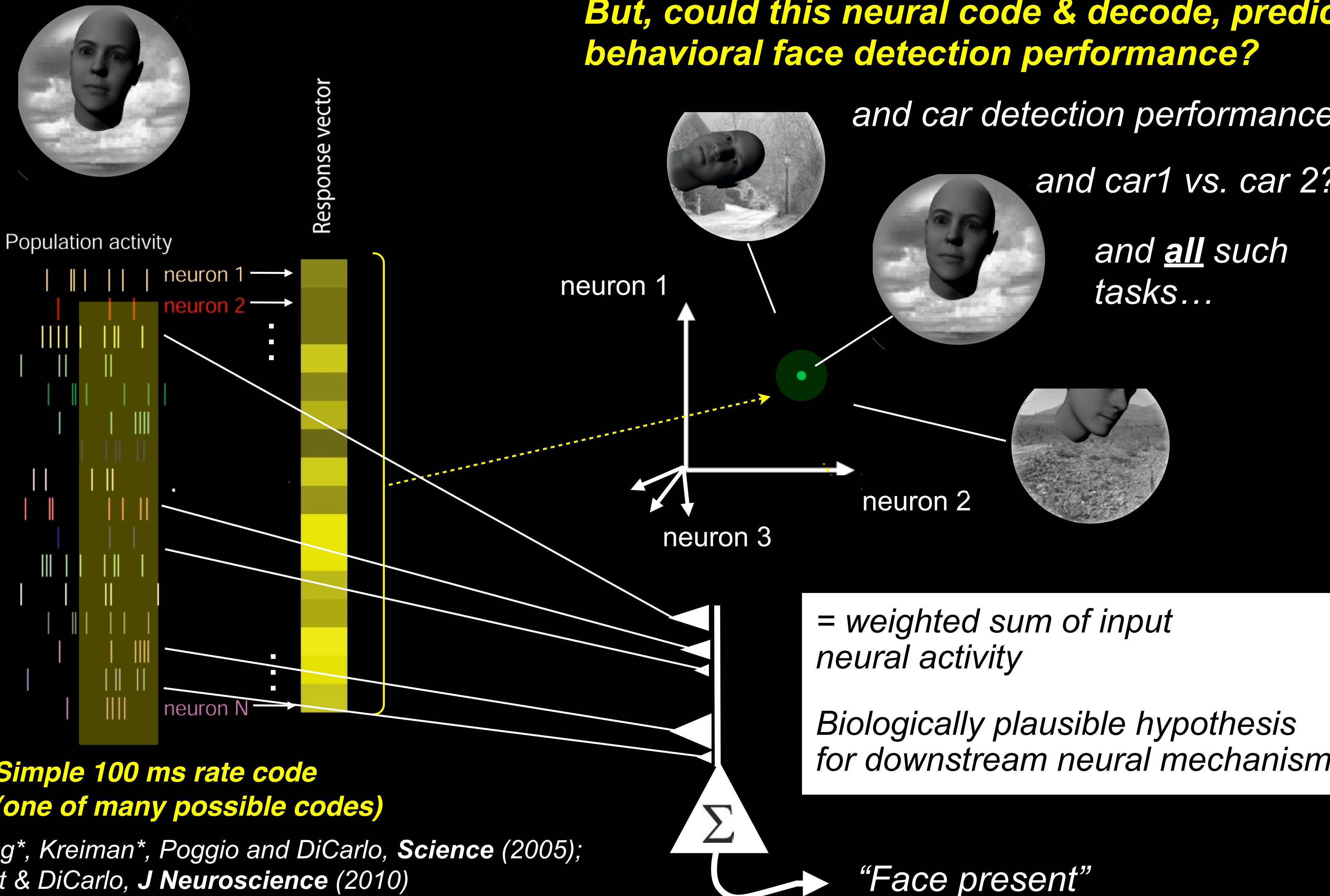
We had previously shown that simple weighted sums of IT population responses have high performance in recognition tasks

But, could this neural code & decode, predict behavioral face detection performance?

and car detection performance?

and car1 vs. car 2?

*and **all** such tasks…*



= weighted sum of input neural activity

Biologically plausible hypothesis for downstream neural mechanism

Simple 100 ms rate code (one of many possible codes)

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

Source: Majaj, Najib J., Ha Hong, Ethan A. Solomon, and James J. DiCarlo. "Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance." Journal of Neuroscience 35, no. 39 (2015): 13402-13418; DOI: https://doi.org/10.1523/JNEUROSCI.5181-14.2015.

What code & decoding mechanism explains object recognition?

Our working hypothesis from previous work:

<u>Passively-evoked</u> spike <u>rate codes (using a single, fixed time scale)</u> that are <u>spatially distributed</u> over a <u>single, fixed number</u> of non-human primate <u>IT cortex</u> neurons and <u>learned from a reasonable number of examples.</u>

If correct, this code/decode should predict monkey <u>and human</u> reports about object category and object identity for all tasks.

Other possibilities:

Attentional and/or arousal mechanisms are needed to “activate” IT

Trial-by-trial coordinated spike timing patterns are crucial

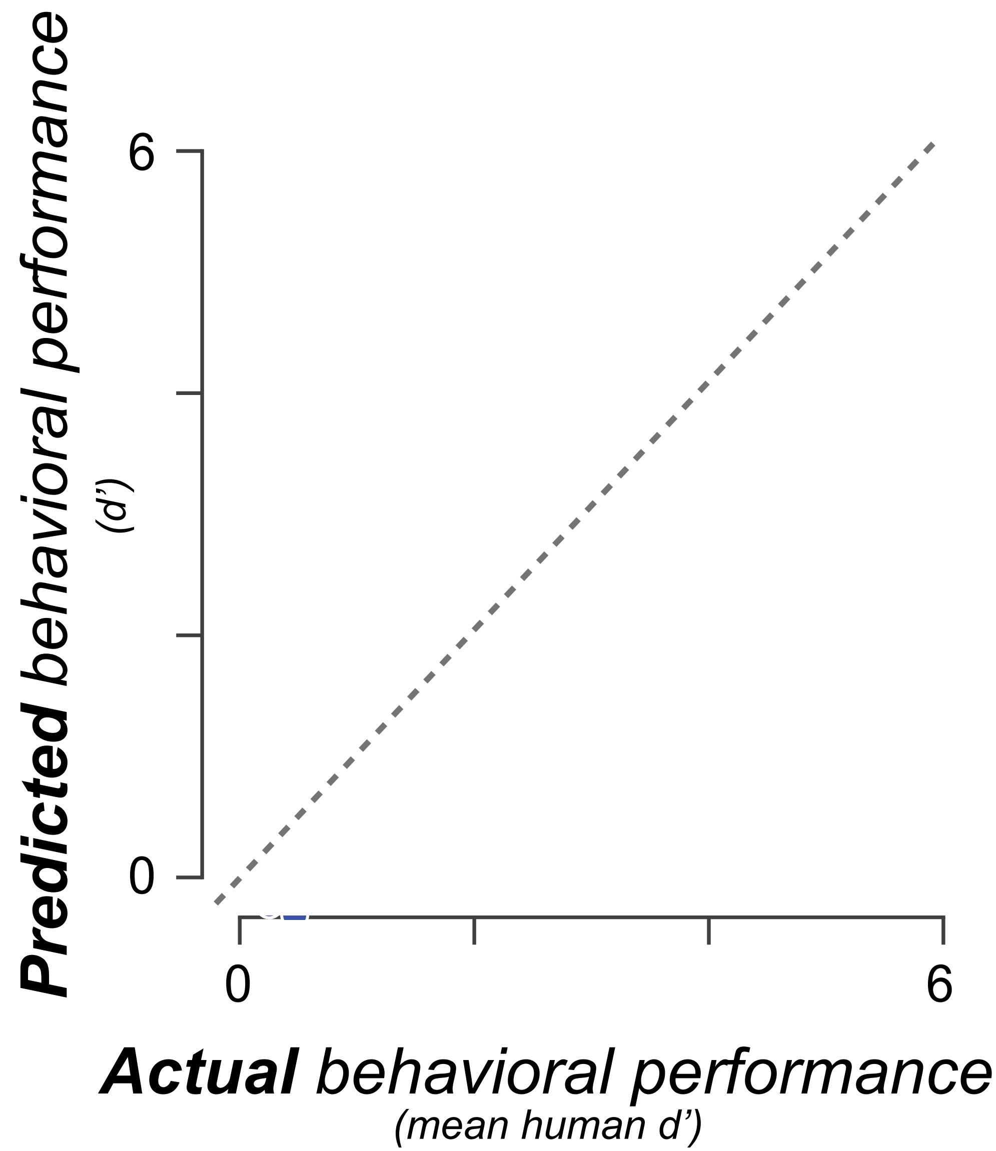
Compartments within IT must be carefully considered
(e.g. tasks related to faces handled exclusively by “face patch” network)

IT does not directly underlie object recognition

Performance requires too many training examples

Monkey neuronal codes cannot explain human behavior

Our first decoder (based on previous work), with number of neurons chosen (once) to match human performance



Majaj, Hong, Solomon, and DiCarlo, ***Cosyne 2012***
Majaj, Hong, Solomon, and DiCarlo, ***Under Review***

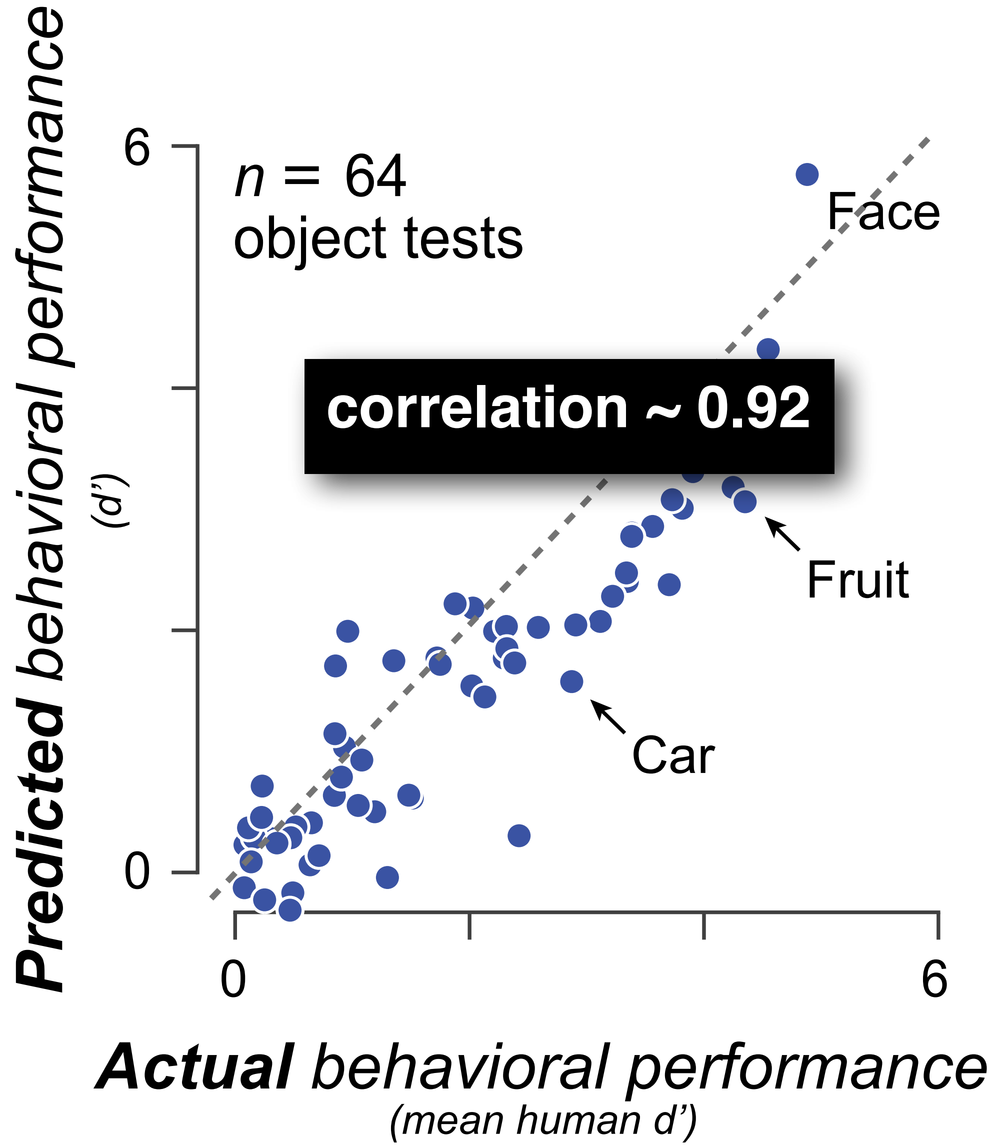
Take home: simple, learned weighted sums of IT firing rates accurately predict the pattern of PERFORMANCE over all object recognition tests

Parameters of inferred neural code/decoding mechanism:

- for each new object, randomly sample ~50,000 single neurons spatially distributed over IT

- "listen" to each IT site's average spiking response (ave over 100 ms)

- learn an appropriately weighted sum of those IT spiking outputs, and then use ~10% of them to judge the likelihood of the object being present

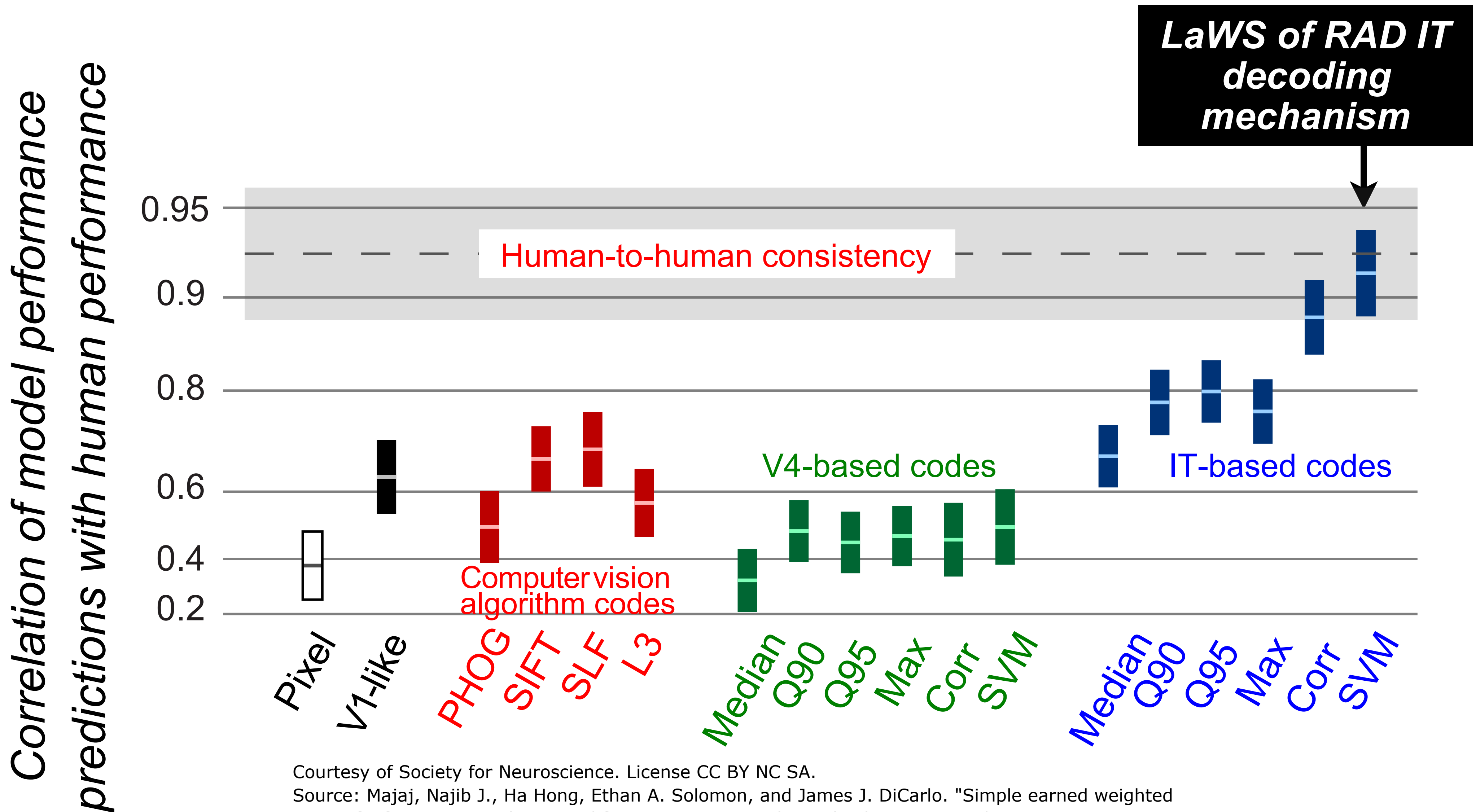


*Learned **W**eighted **S**ums of (~50,000) **R**andom **A**verage (100 ms) single unit responses **D**istributed over **IT***

"LaWS of RAD IT" decoding mechanism

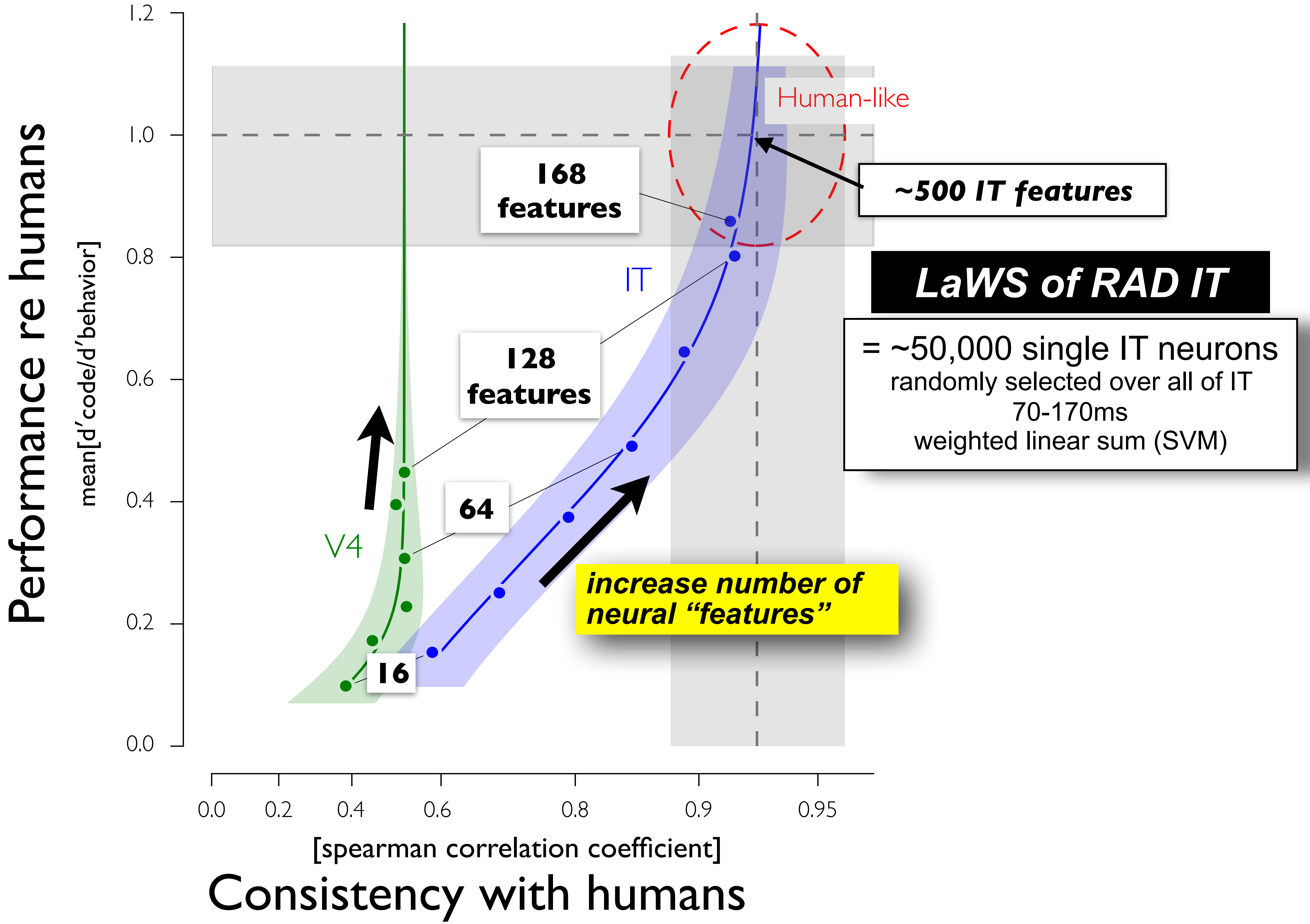
Majaj, Hong, Solomon, and DiCarlo, ***Cosyne 2012***
Majaj, Hong, Solomon, and DiCarlo, ***Under Review***

Some controls…
Most alternative codes/decoding mechanisms are not even close.



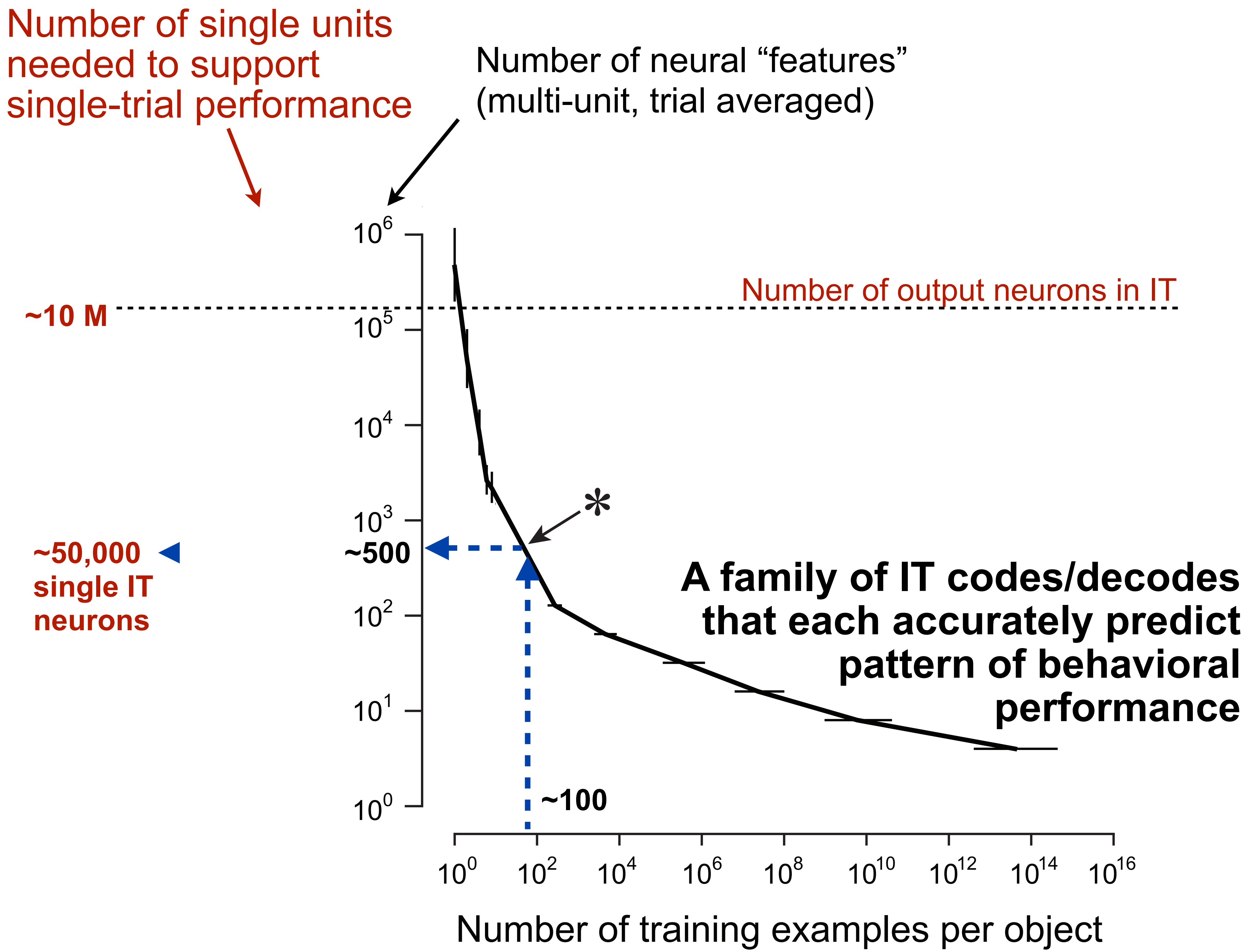
Source: Majaj, Najib J., Ha Hong, Ethan A. Solomon, and James J. DiCarlo. "Simple earned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance." Journal of Neuroscience 35, no. 39 (2015): 13402-13418.

Majaj, Hong, Solomon, and DiCarlo, ***Cosyne 2012***
Majaj, Hong, Solomon, and DiCarlo, ***Under Review***



Source: Majaj, Najib J., Ha Hong, Ethan A. Solomon, and James J. DiCarlo. "Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance." Journal of Neuroscience 35, no. 39 (2015): 13402-13418.

*Majaj, Hong, Solomon, and DiCarlo, **Cosyne 2012***
*Majaj, Hong, Solomon, and DiCarlo, **Under Review***

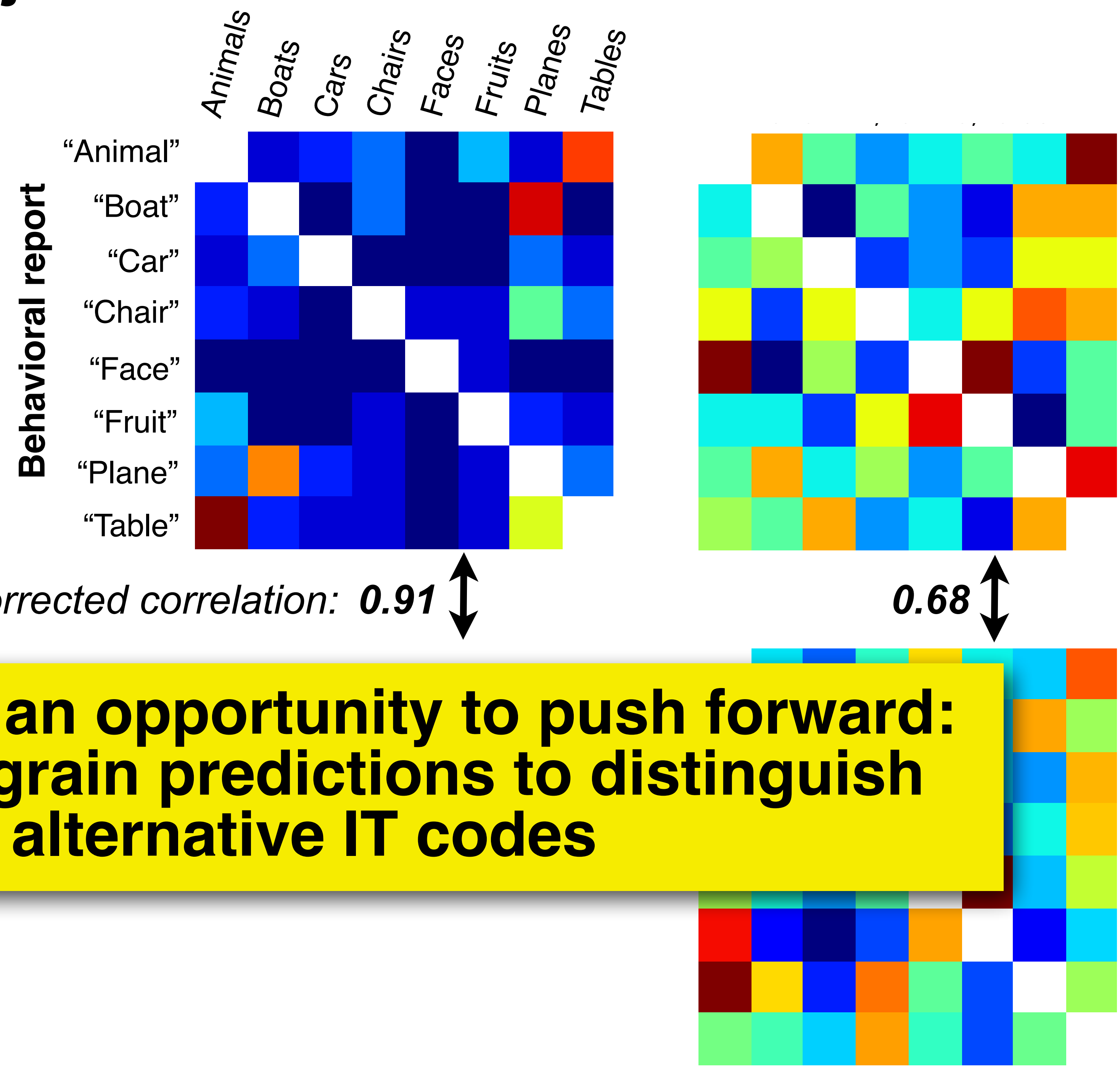


Number of single units needed to support single-trial performance
Number of neural "features" (multi-unit, trial averaged)
Number of output neurons in IT
~10 M
~50,000 single IT neurons
~500
~100
*
A family of IT codes/decodes that each accurately predict pattern of behavioral performance
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^0 10^2 10^4 10^6 10^8 10^{10} 10^{12} 10^{14} 10^{16}
Number of training examples per object

Behavioral object confusions

Predicted:

LaWS of RAD IT decoding mechanism



Noise-corrected correlation: ***0.91***

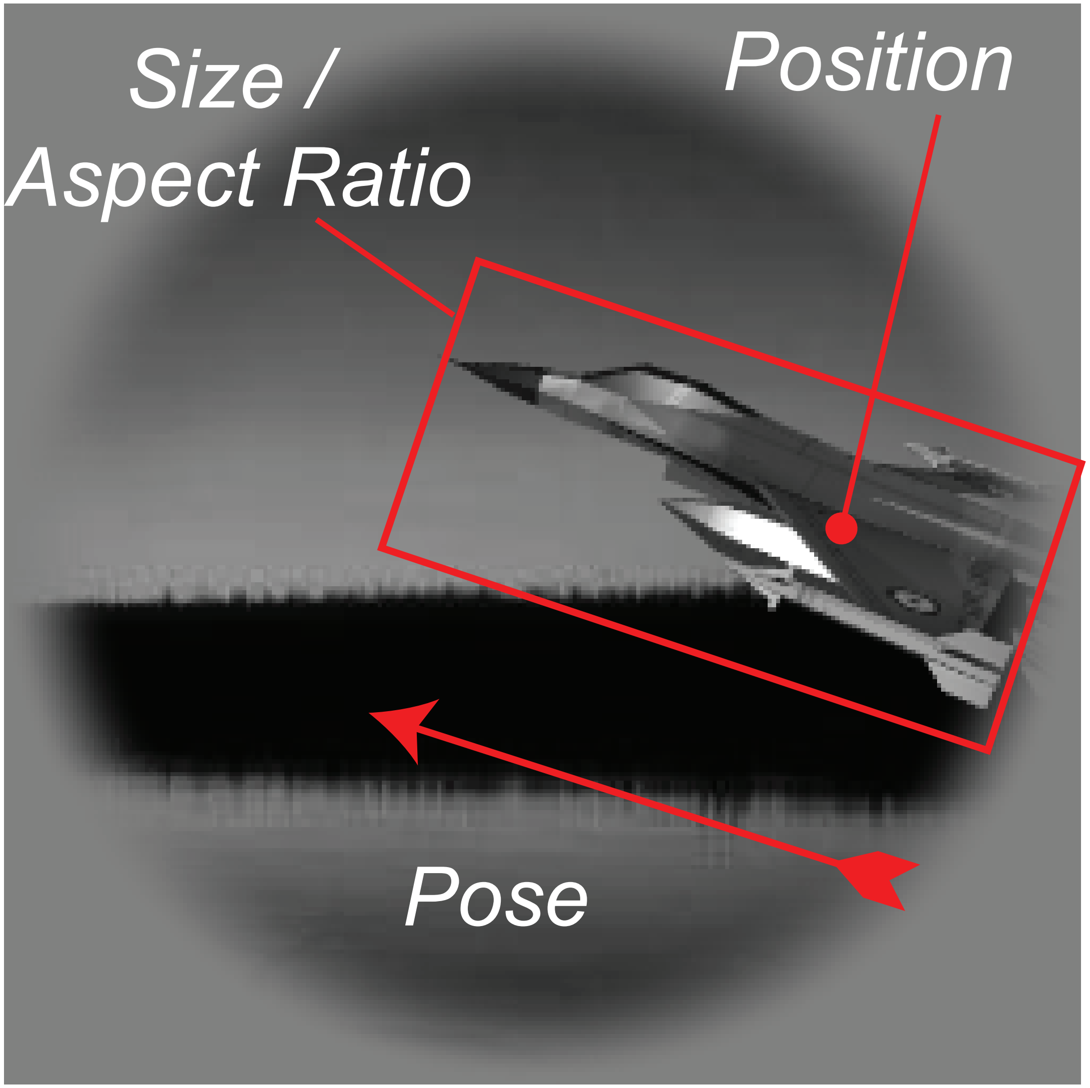
0.68

This is an opportunity to push forward: image grain predictions to distinguish among alternative IT codes

High variation

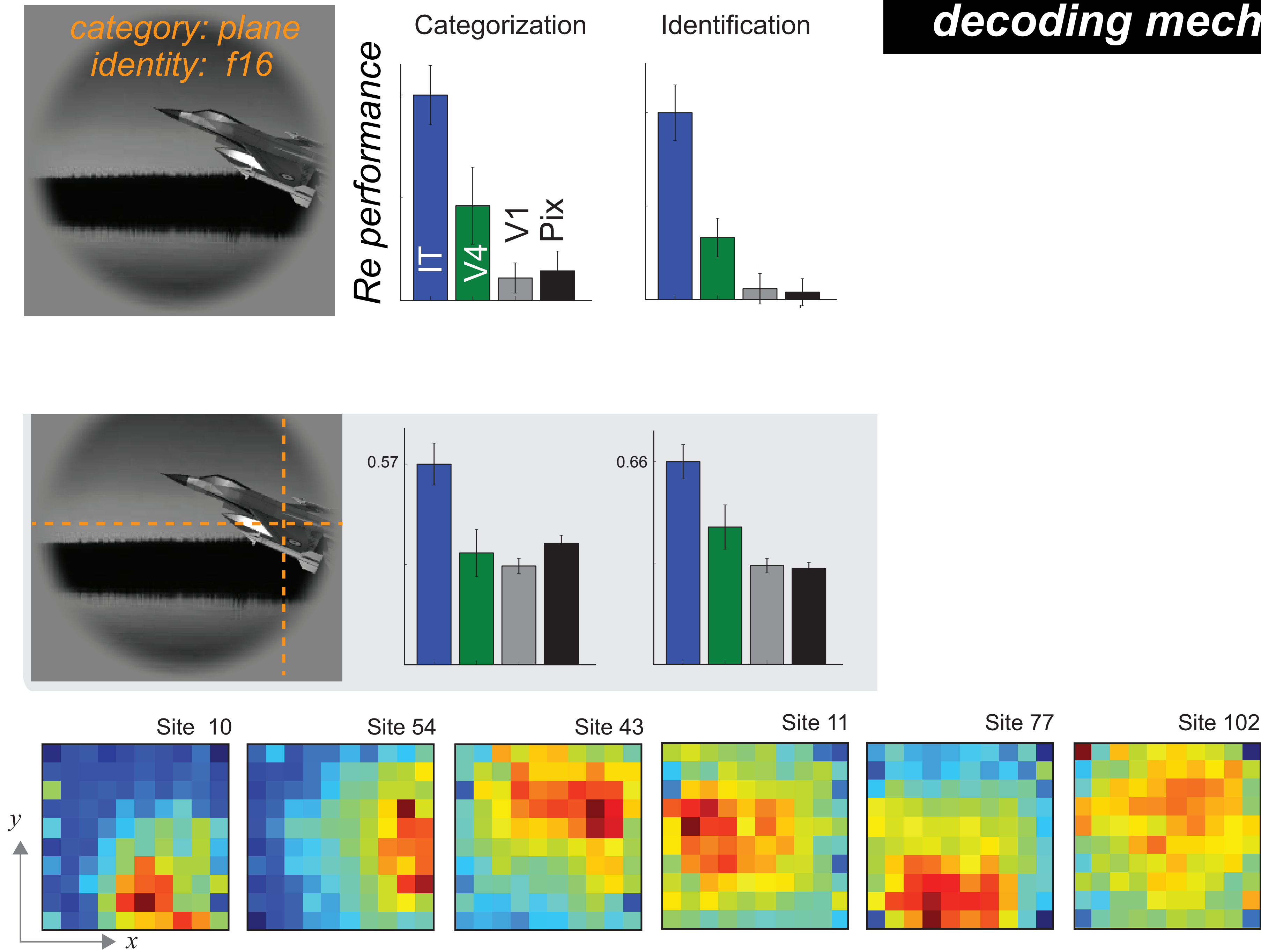
Other object latent variables

Category: plane
Identity: f16



Source: Hong, Ha, Daniel LK Yamins, Najib J. Majaj, and James J. DiCarlo. "Explicit information for category-orthogonal object properties increases along the ventral stream." Nature neuroscience 19, no. 4 (2016): 613-622.

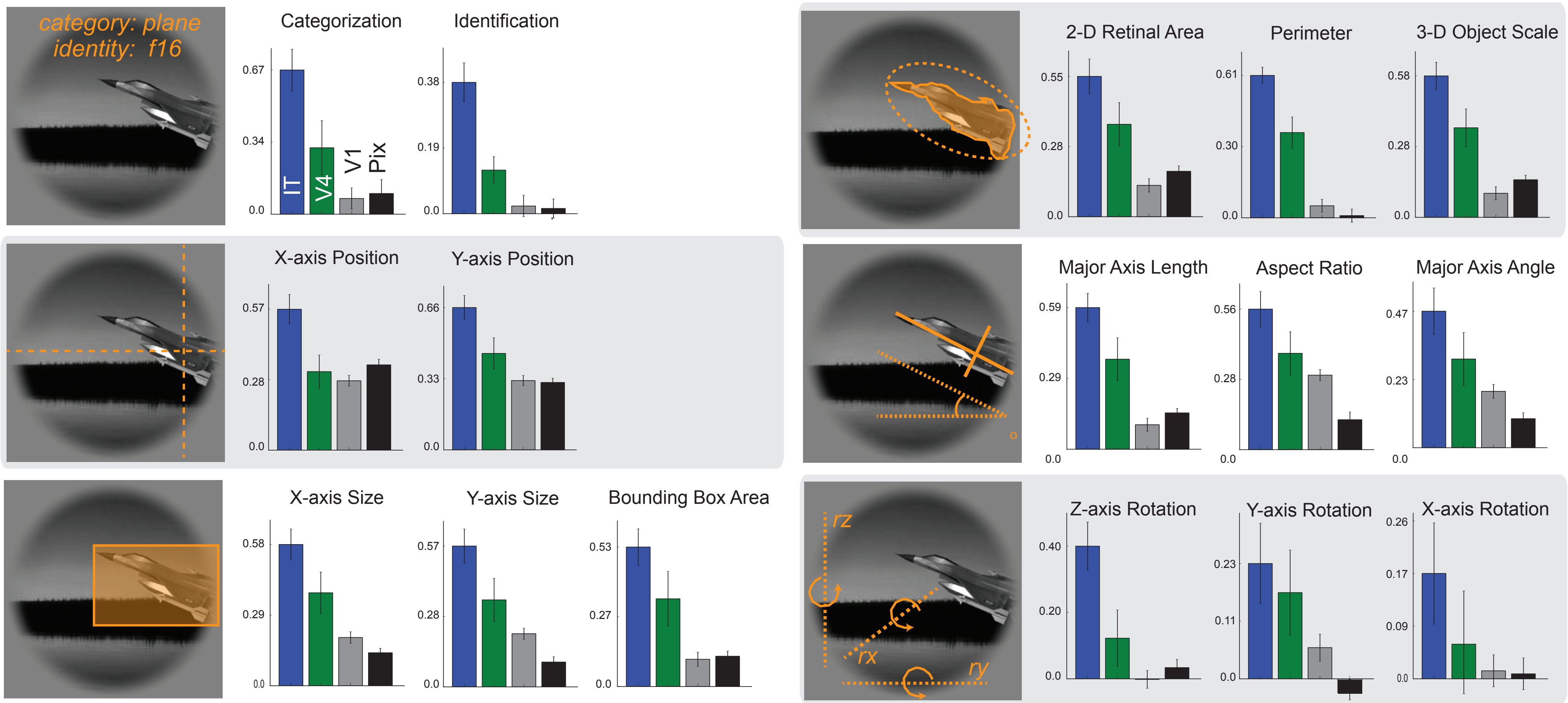
LaWS of RAD IT decoding mechanism



Source: Hong, Ha, Daniel LK Yamins, Najib J. Majaj, and James J. DiCarlo. "Explicit information for category-orthogonal object properties increases along the ventral stream." Nature neuroscience 19, no. 4 (2016): 613-622.

Sum: LaWS of RAD IT performs better than other codes/decodes.

LaWS of RAD IT decoding mechanism

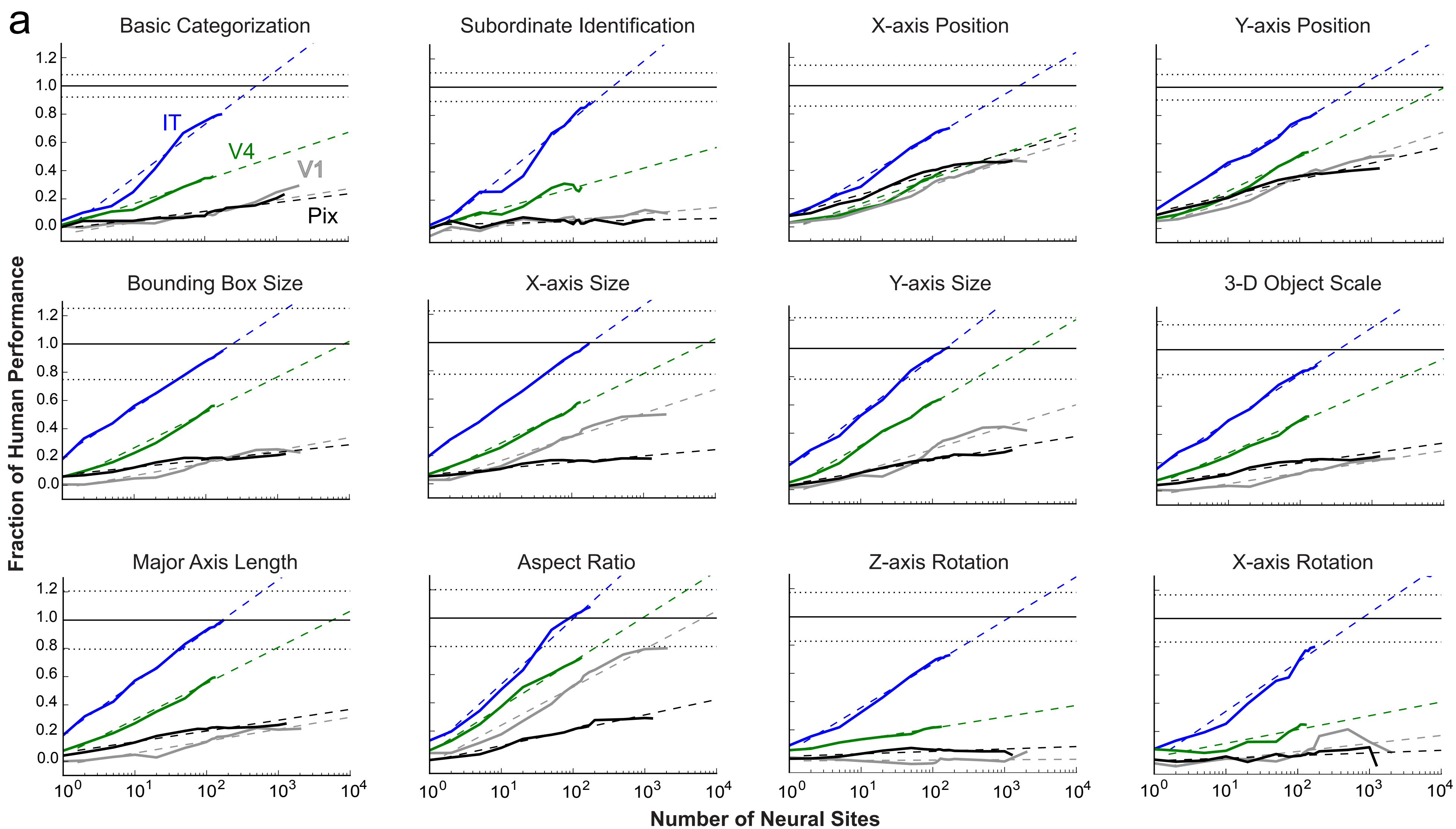


Source: Hong, Ha, Daniel LK Yamins, Najib J. Majaj, and James J. DiCarlo. "Explicit information for category-orthogonal object properties increases along the ventral stream." Nature neuroscience 19, no. 4 (2016): 613-622.

But these tasks are not all equally difficult for humans. Does this decoding mechanism predict that pattern of difficulty?

To test this, we collected human performance data on these images/tasks.

LaWS of RAD IT decoding mechanism



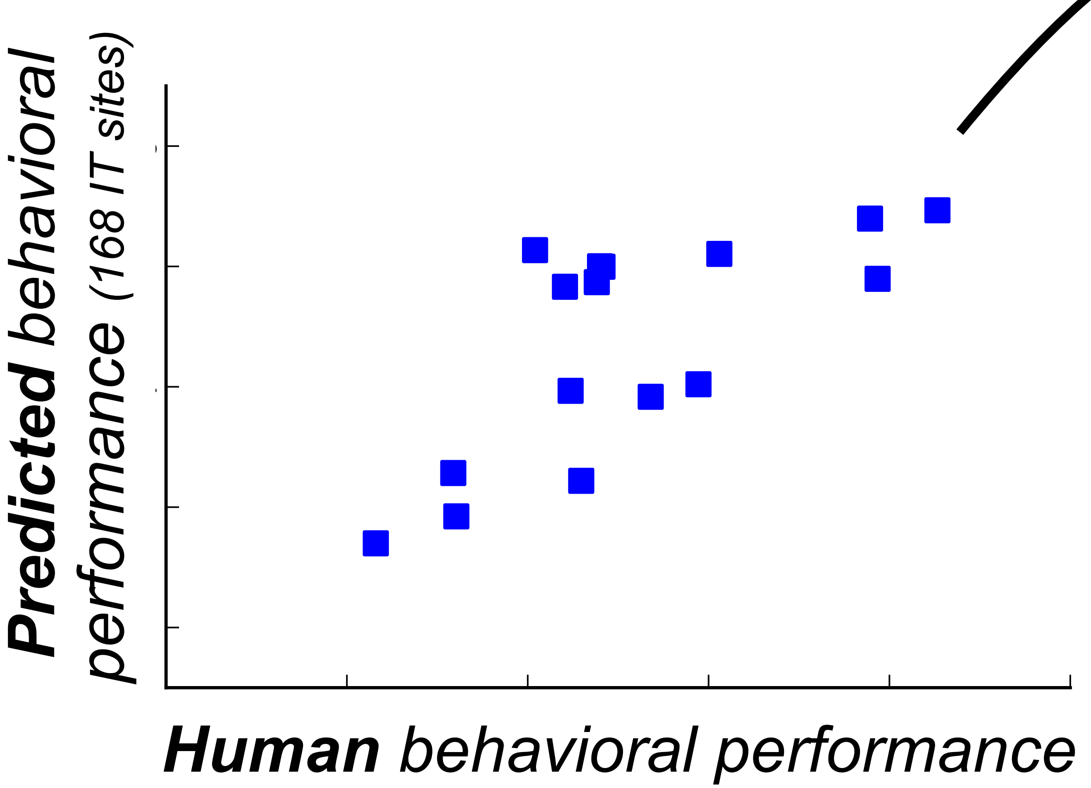
Source: Hong, Ha, Daniel LK Yamins, Najib J. Majaj, and James J. DiCarlo. "Explicit information for category-orthogonal object properties increases along the ventral stream." Nature neuroscience 19, no. 4 (2016): 613-622.

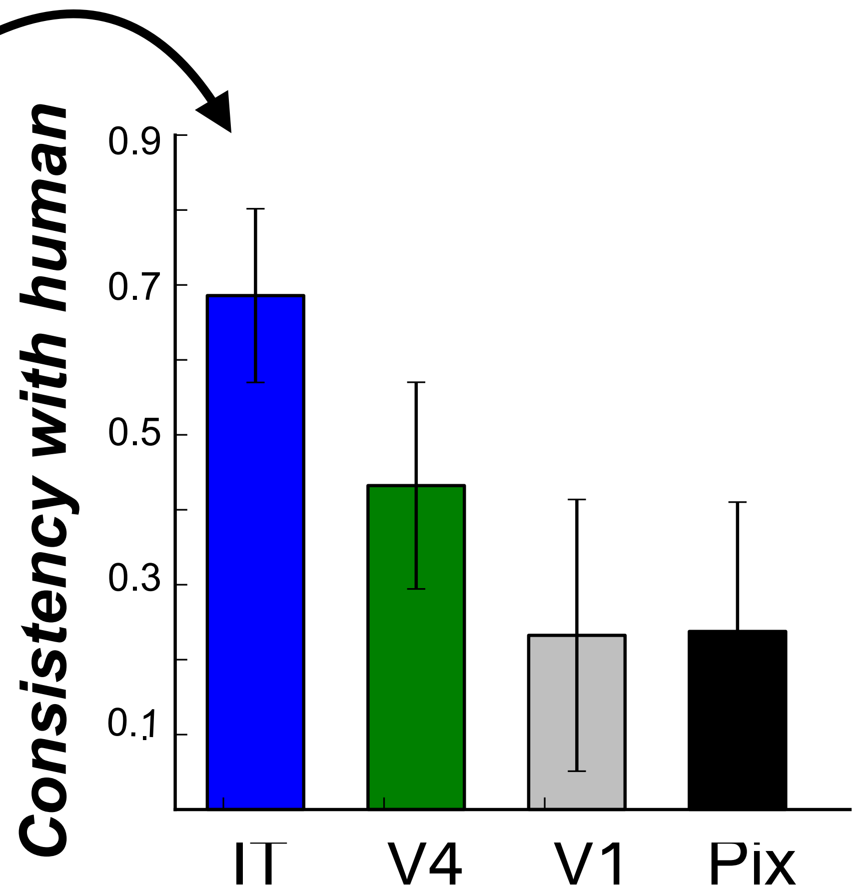
Number of IT sites needed to match human performance

LaWS of RAD IT decoding mechanism

	IT	V4	V1	Pix
Basic Categorization	520 +/- 165	8.84 x 10^5	---	---
Subordinate Identification	444 +/- 61	9.15 x 10^6	---	---
X-axis Position	1624 +/- 44	4.5 x 10^6	3 x 10^7	---
Y-axis Position	647 +/- 215	1.1 x 10^5	8.7 x 10^6	---
Bounding Box Size	234 +/- 91	8.4 x 10^3	---	---
X-axis Size	150 +/- 55	2.1 x 10^3	3.4 x 10^7	---
Y-axis Size	182 +/- 62	7.8 x 10^3	9.5 x 10^6	---

	IT	V4	V1	Pix
3-D Object Scale	339 +/- 79	1.9 x 10^5	---	---
Major Axis Length	165 +/- 59	5.7 x 10^3	---	---
Aspect Ratio	103 +- 37	922 +/- 59	6.5 x 10^3	---
Major Axis Angle	520 +/- 165	520 +/- 165	---	---
Z-axis Rotation	1206 +/- 473	---	---	---
Y-axis Rotation	1317 +/- 459	1.1 x 10^5	---	---
X-axis Rotation	775 +/- 248	---	---	---

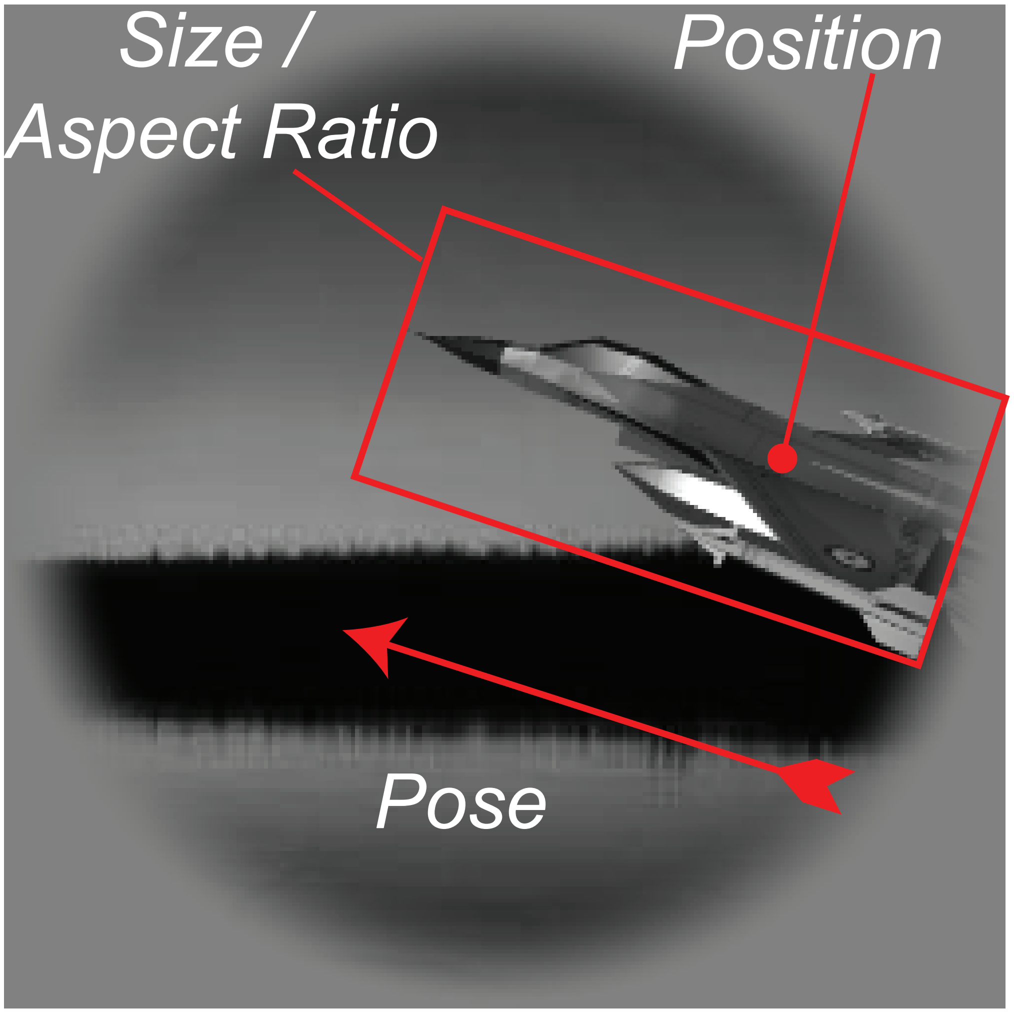




Source: Hong, Ha, Daniel LK Yamins, Najib J. Majaj, and James J. DiCarlo. "Explicit information for category-orthogonal object properties increases along the ventral stream." Nature neuroscience 19, no. 4 (2016): 613-622.

LaWS of RAD IT decoding mechanism

Category: plane
Identity: f16



Source: Hong, Ha, Daniel LK Yamins, Najib J. Majaj, and James J. DiCarlo. "Explicit information for category-orthogonal object properties increases along the ventral stream." Nature neuroscience 19, no. 4 (2016): 613-622.

Summary: This ventral stream code/decoding mechanism also predicts human patterns of performance for other object latent variables.

This suggests that:

- ***the IT population conveys a general purpose object representation***
- ***the job of the ventral stream is not to produce category "invariant" representations***

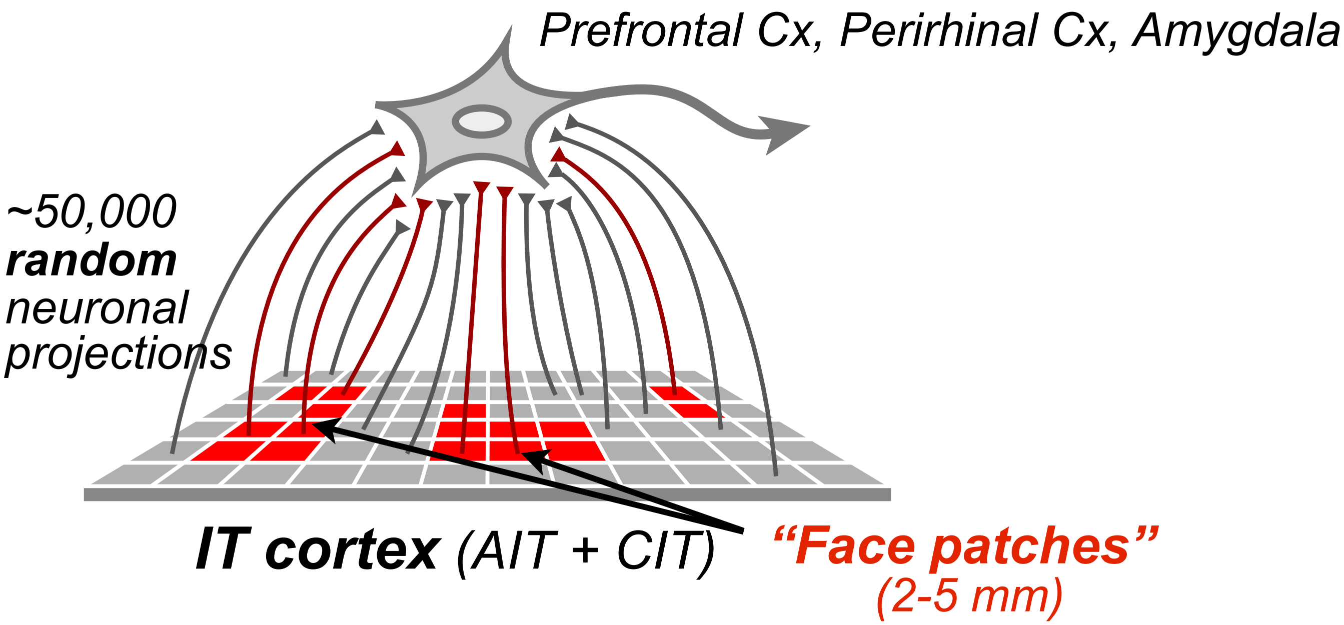
Edelman (1998), DiCarlo and Cox (2007), Li et al. (2009), etc.

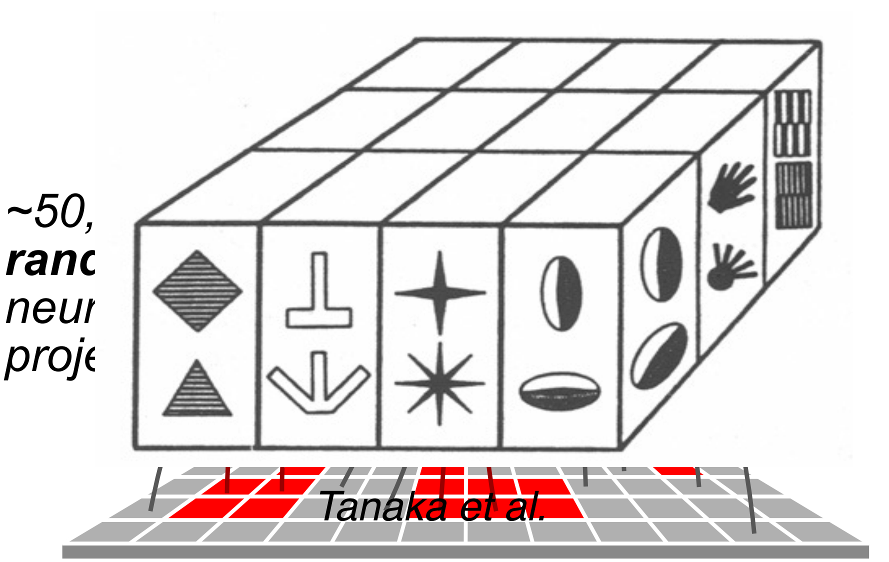
Hong, Yamins, Majaj, and DiCarlo, ***Cosyne 2014***

Hong, Yamins, Majaj, and DiCarlo, ***(in prep)***

Sketch of the inferred anatomy:

LaWS of RAD IT *[70-170ms, 50,000n, 100t]*





Source: Tanaka, Keiji. "Neuronal mechanisms of object recognition." Science-New York Then Washington 262 (1993): 685-685.

Causal tests of this model

LaWS of RAD IT *[70-170ms, 50,000n, 100t]*

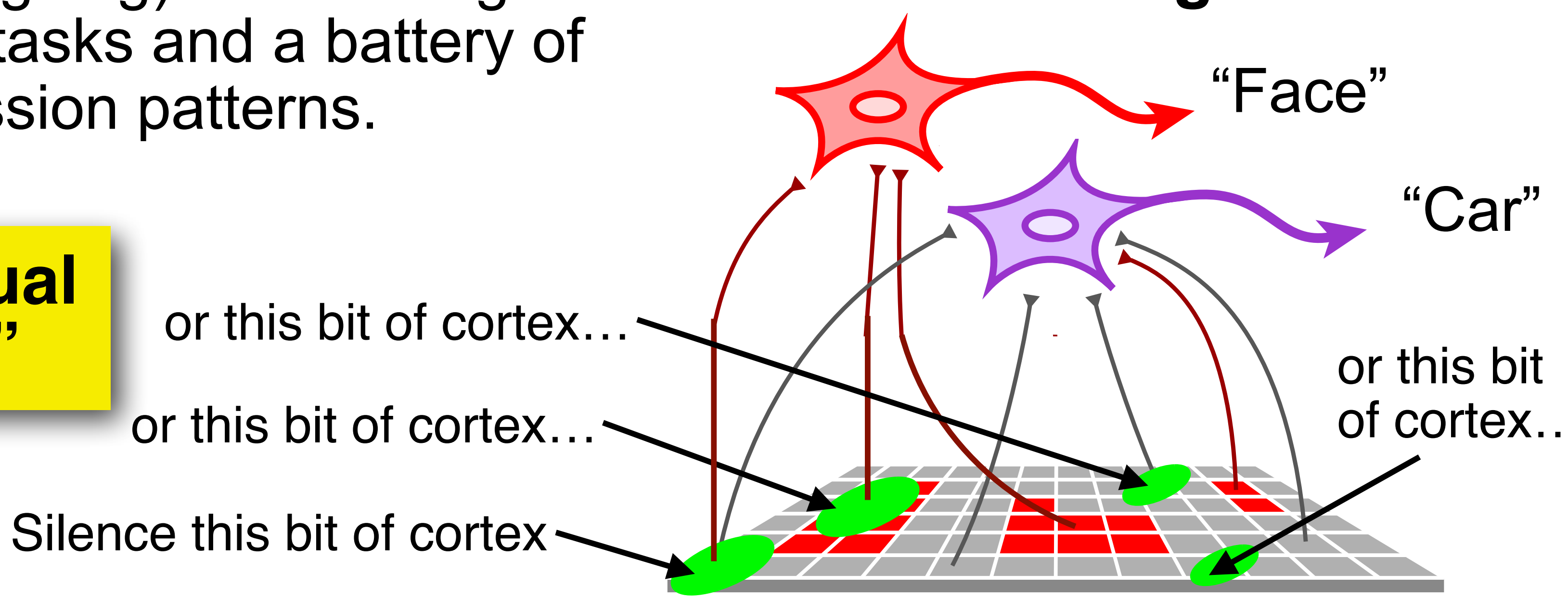
The model allows us to predict how much any object recognition task will be disrupted by direct suppression of IT neurons.

Step 1: (done) Tool building and testing: Can we reliably disrupt performance of a recognition task by directly suppressing the activity of ~1mm IT neural sub-populations?

Step 2 (ongoing): Test a large battery of tasks and a battery of IT suppression patterns.

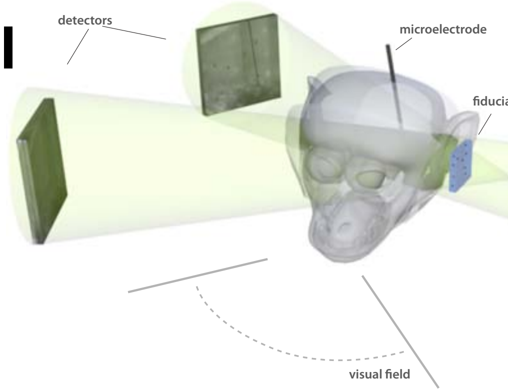
Post-learning:

Towards actual "inception"

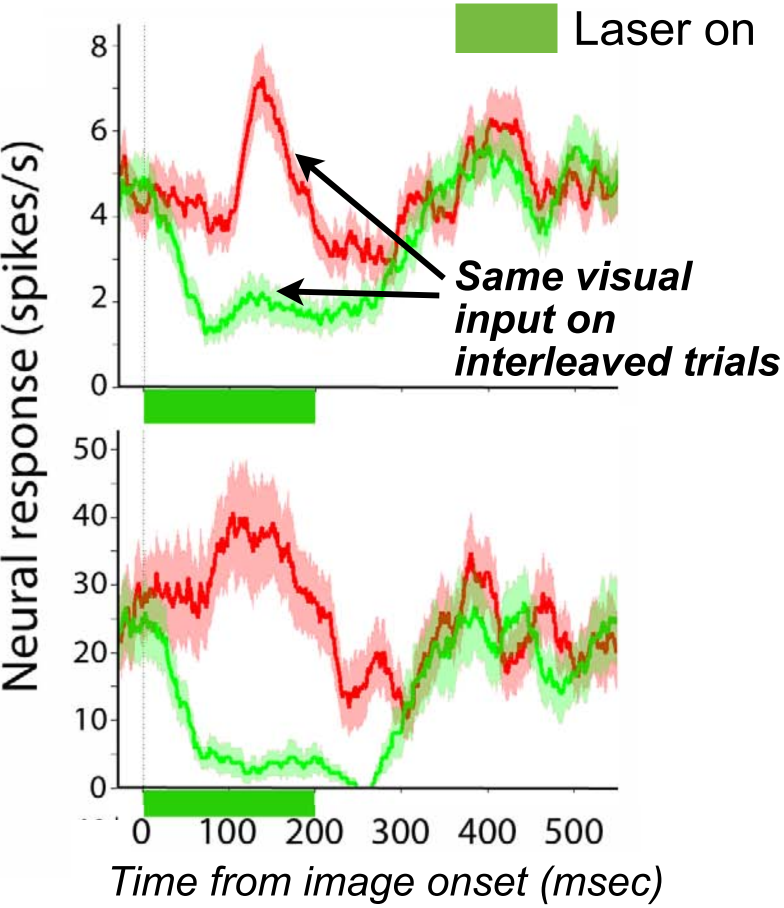


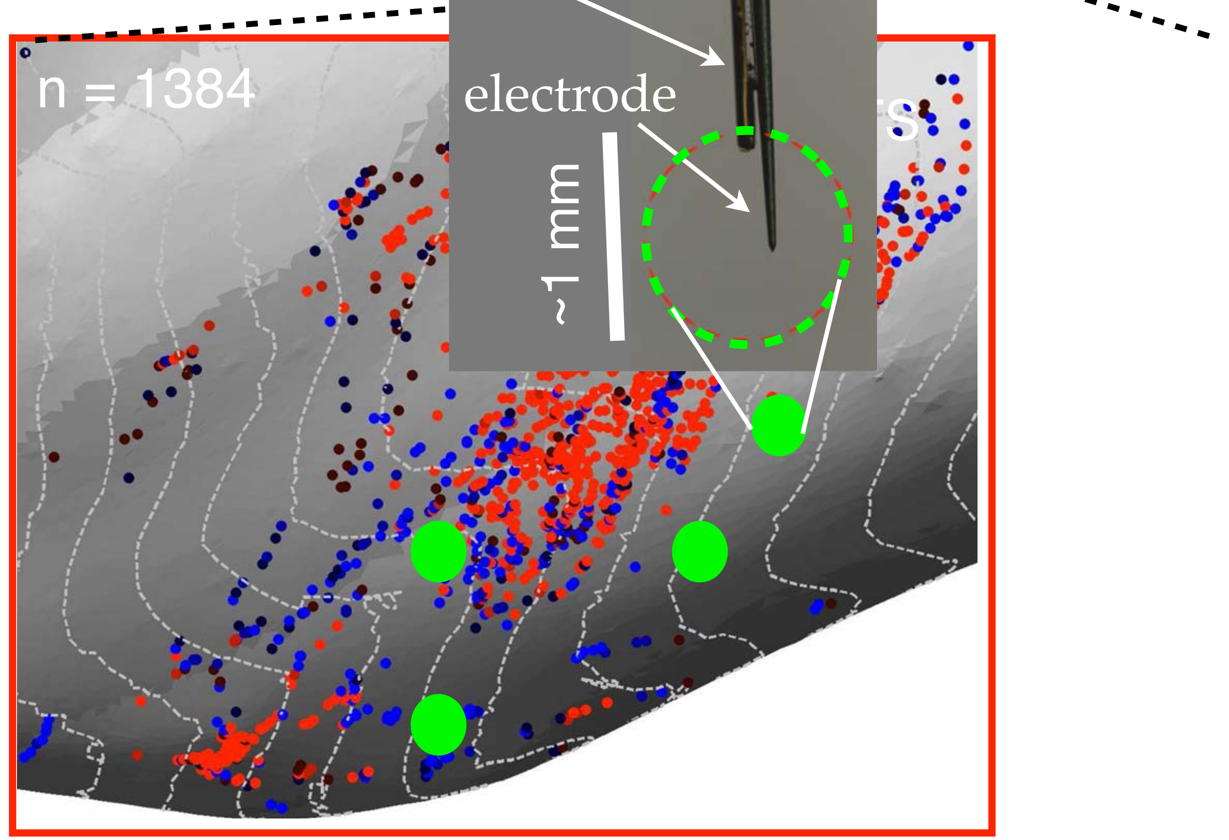
IT cortex *(AIT + CIT)* ***~150 IT sub-regions, each ~1 mm in scale***

Stereo, microfocal x-ray system



Optogenetic (ArchT, CAG, AAV) suppression of visually-driven IT activity



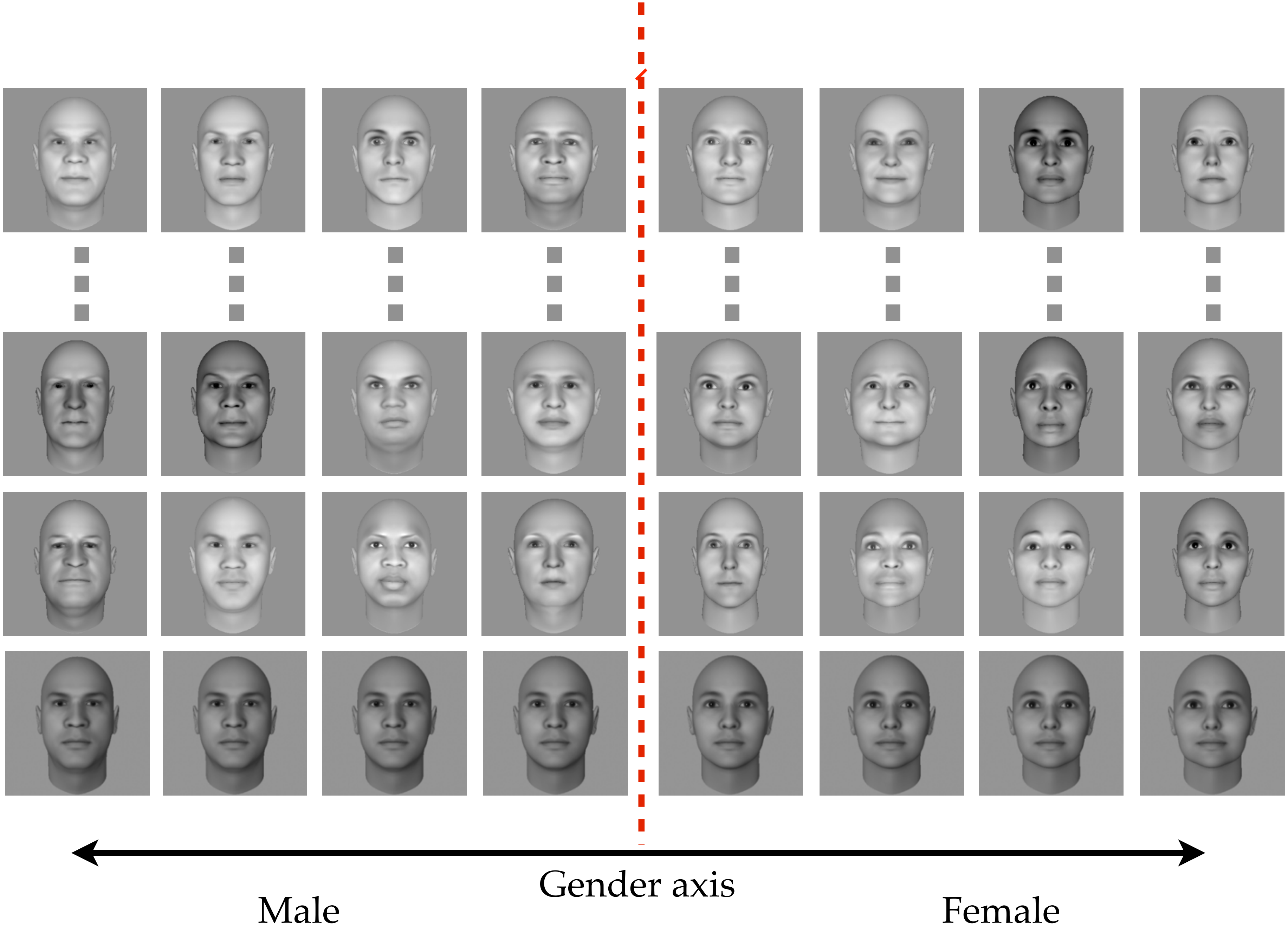


Source: Issa, Elias B., and James J. DiCarlo. "Precedence of the eye region in neural processing of faces." Journal of Neuroscience 32, no. 47 (2012): 16666-16682.

*Afraz, Boyden and DiCarlo, **SFN** (2013)*

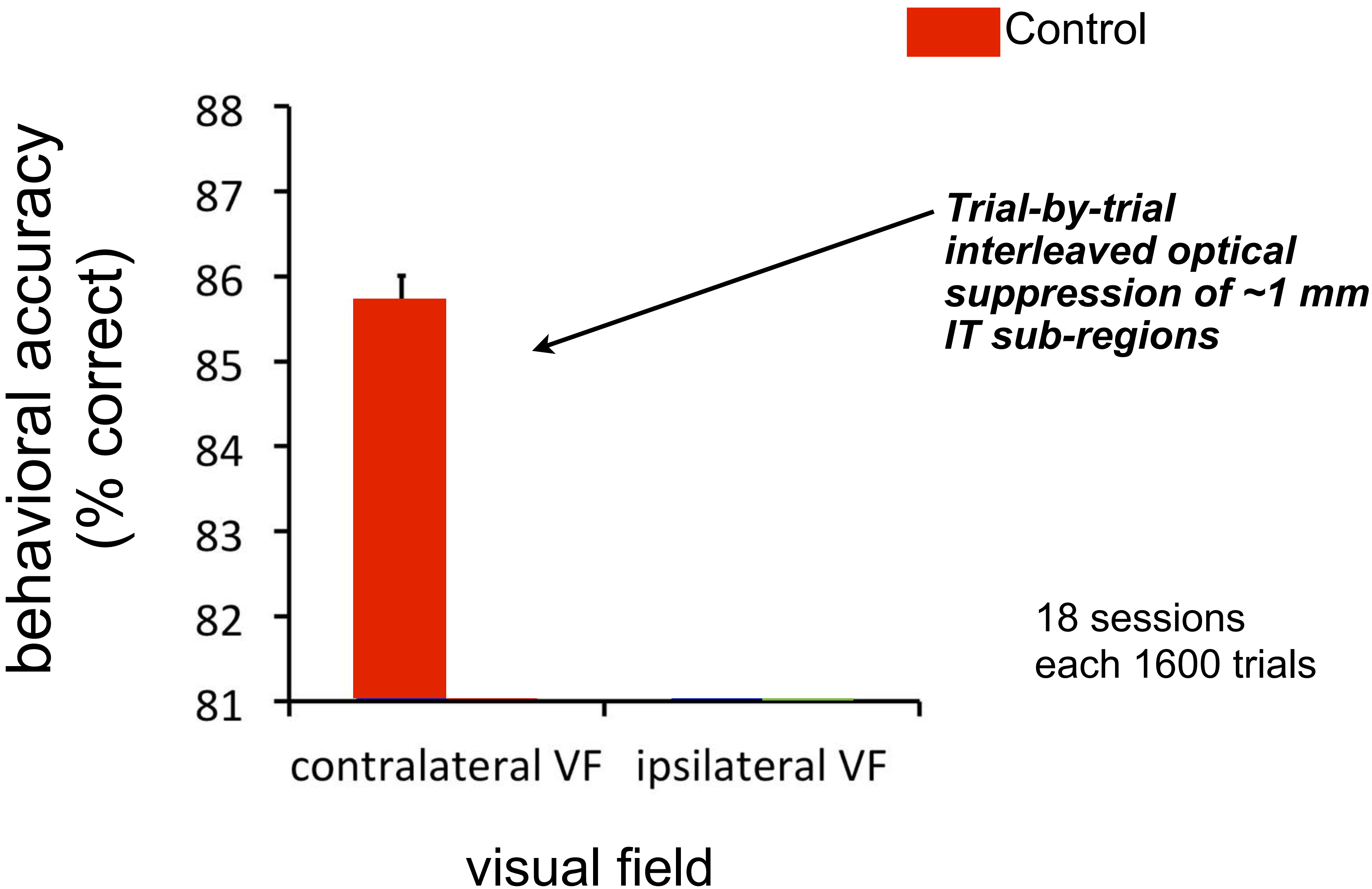
*Issa and DiCarlo, **J Neurosci** (2012)*

Monkey task: face gender discrimination



Source: Afraz, Arash, Edward S. Boyden, and James J. DiCarlo."Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination." Proceedings of the National Academy of Sciences 112, no. 21 (2015): 6730-6735.

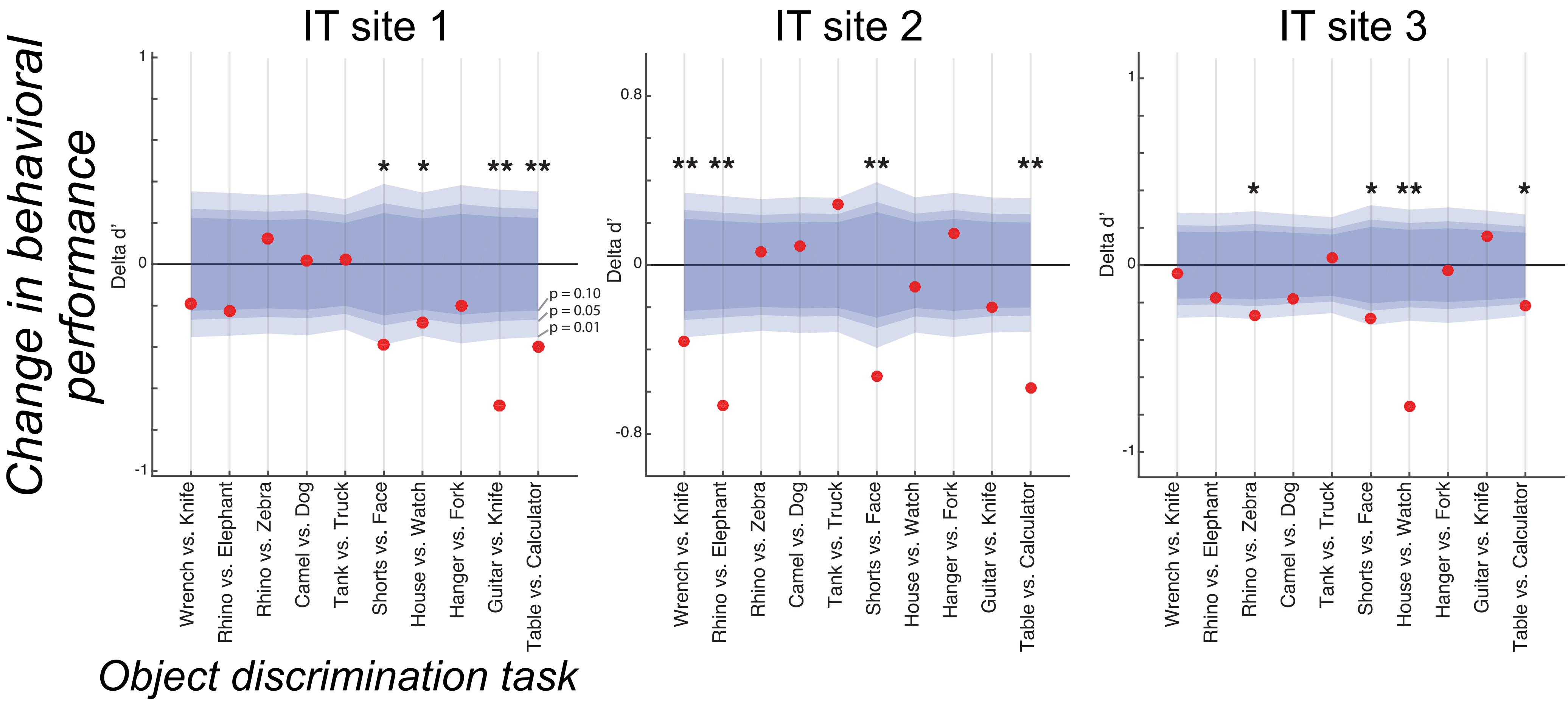
We found a spatially-specific behavioral effect on this object discrimination task



18 sessions
each 1600 trials

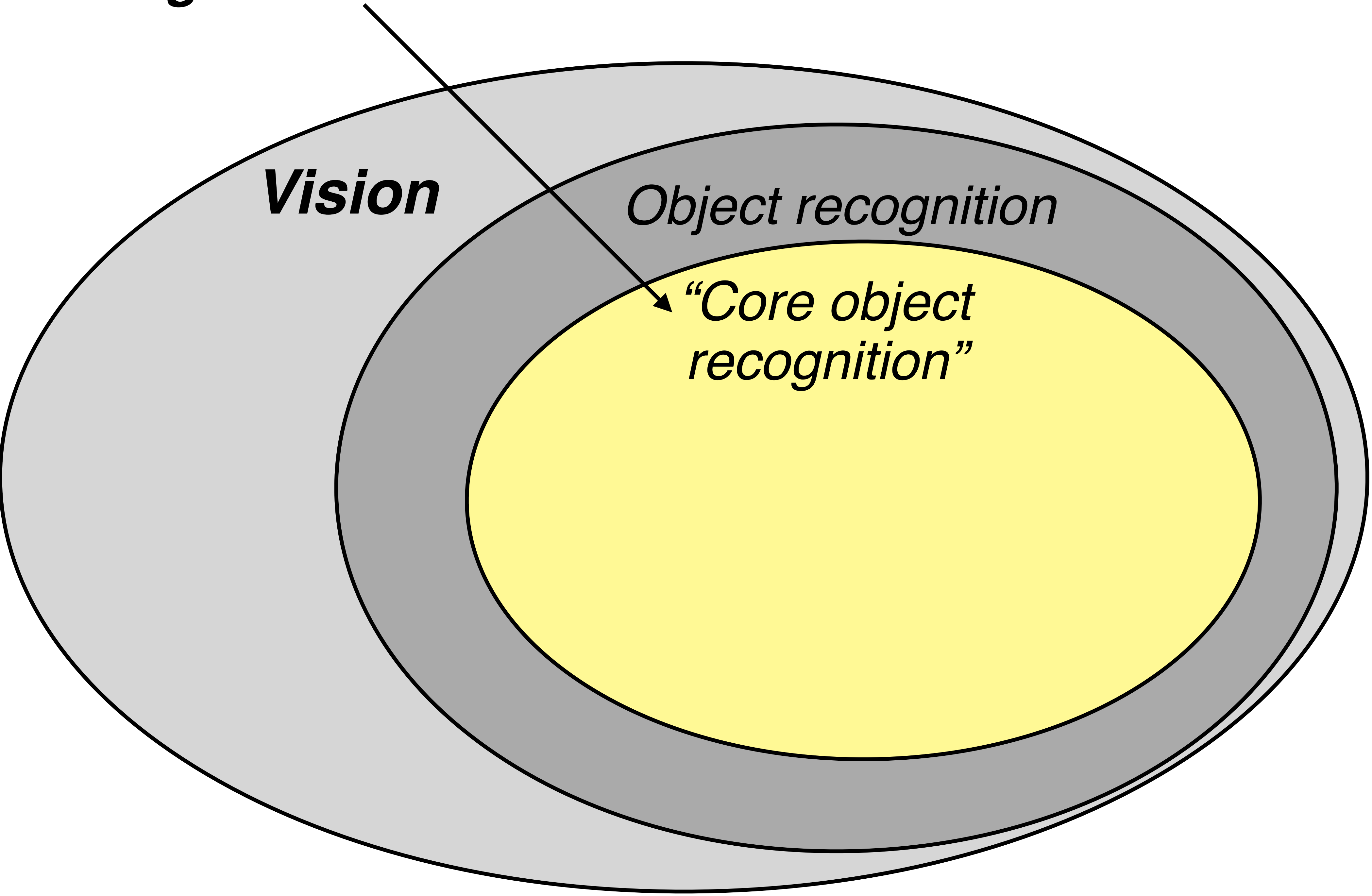
Afraz, Boyden and DiCarlo, ***SFN*** *(2013),* ***VSS*** *(2014);* ***PNAS*** *(2015)*

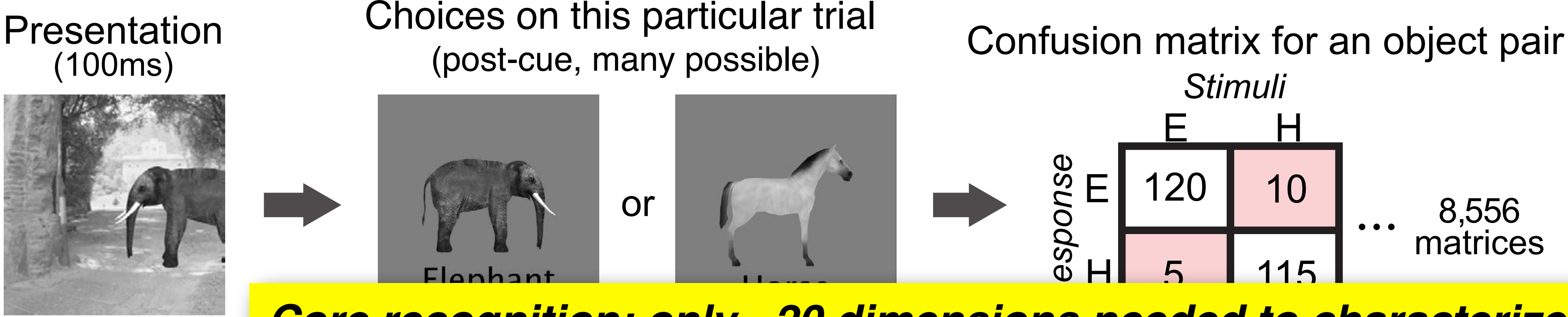
Pharmacological suppression of different IT sub-regions results in different patterns of deficit in basic level object tasks



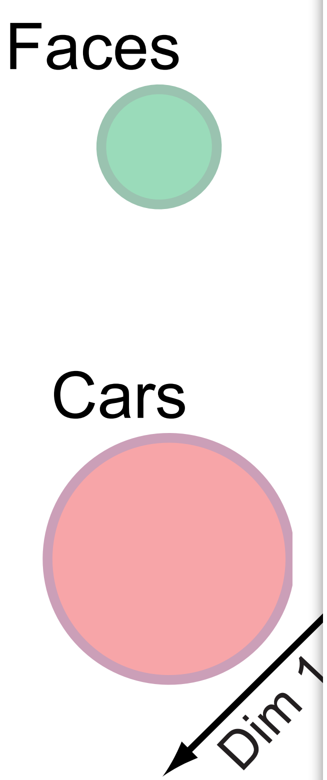
Our current aim is to systematically measure the specific pattern of behavioral change induced by suppression of each IT sub-region (~100) and compare with model predictions

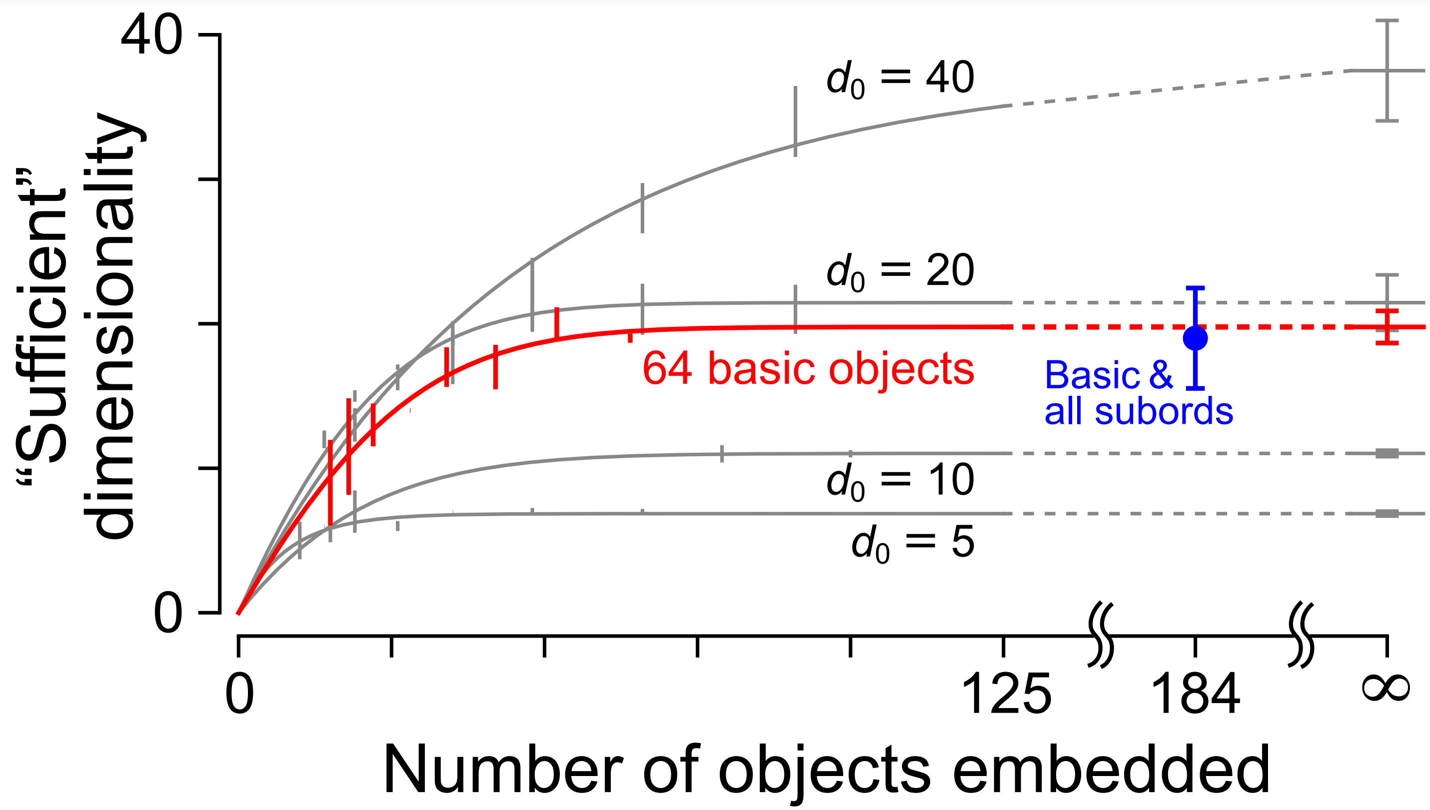
Can we span the entire domain of core recognition tasks? How?





Core recognition: only ~20 dimensions needed to characterize confusions among all basic and subordinate-level objects



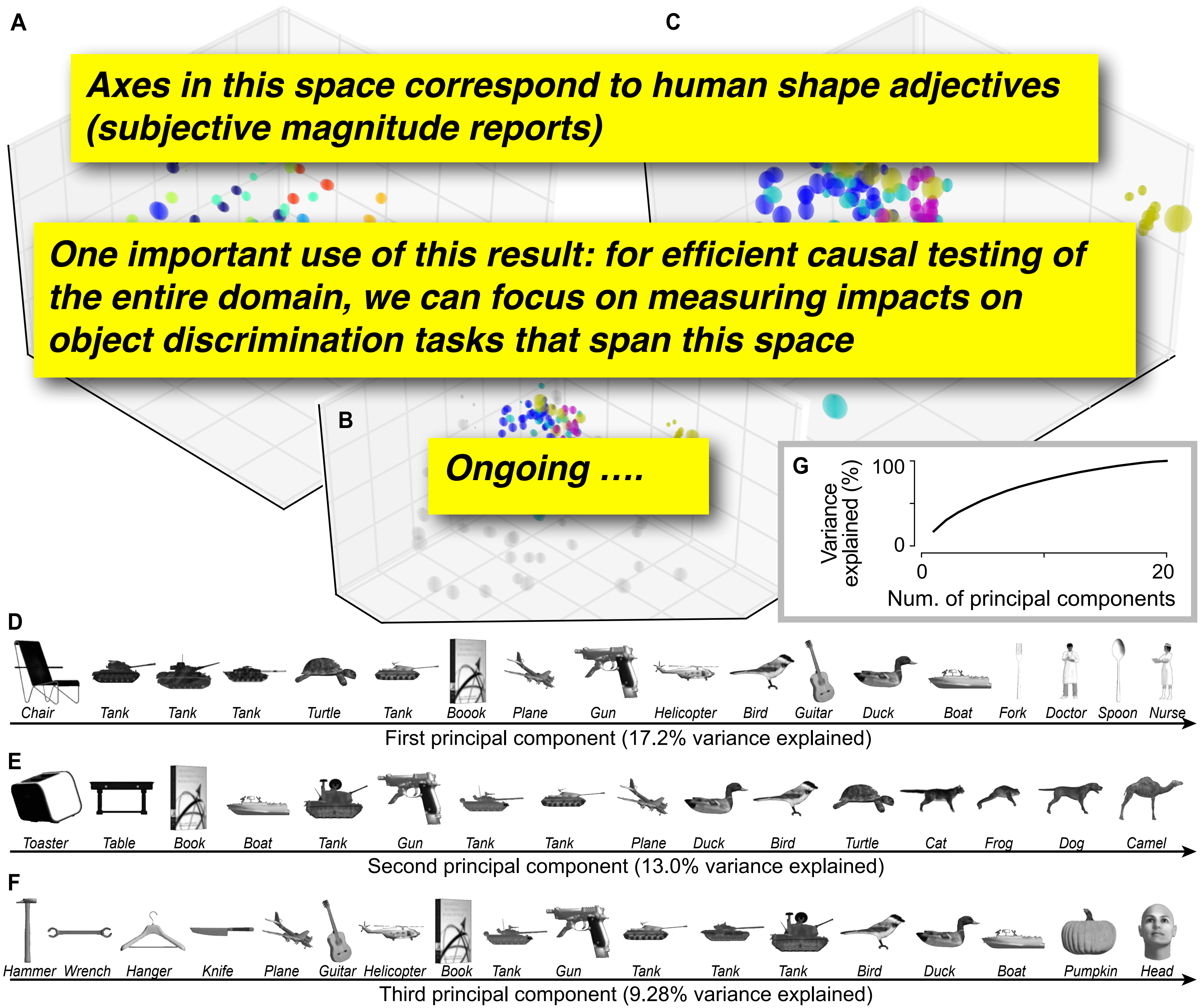


Hong*, Solomon*, Yamins*, and DiCarlo. Large-scale Characterization of a Universal and Compact Visual Perceptual Space. VSS, 2014; in prep

Axes in this space correspond to human shape adjectives (subjective magnitude reports)

One important use of this result: for efficient causal testing of the entire domain, we can focus on measuring impacts on object discrimination tasks that span this space

Ongoing ….



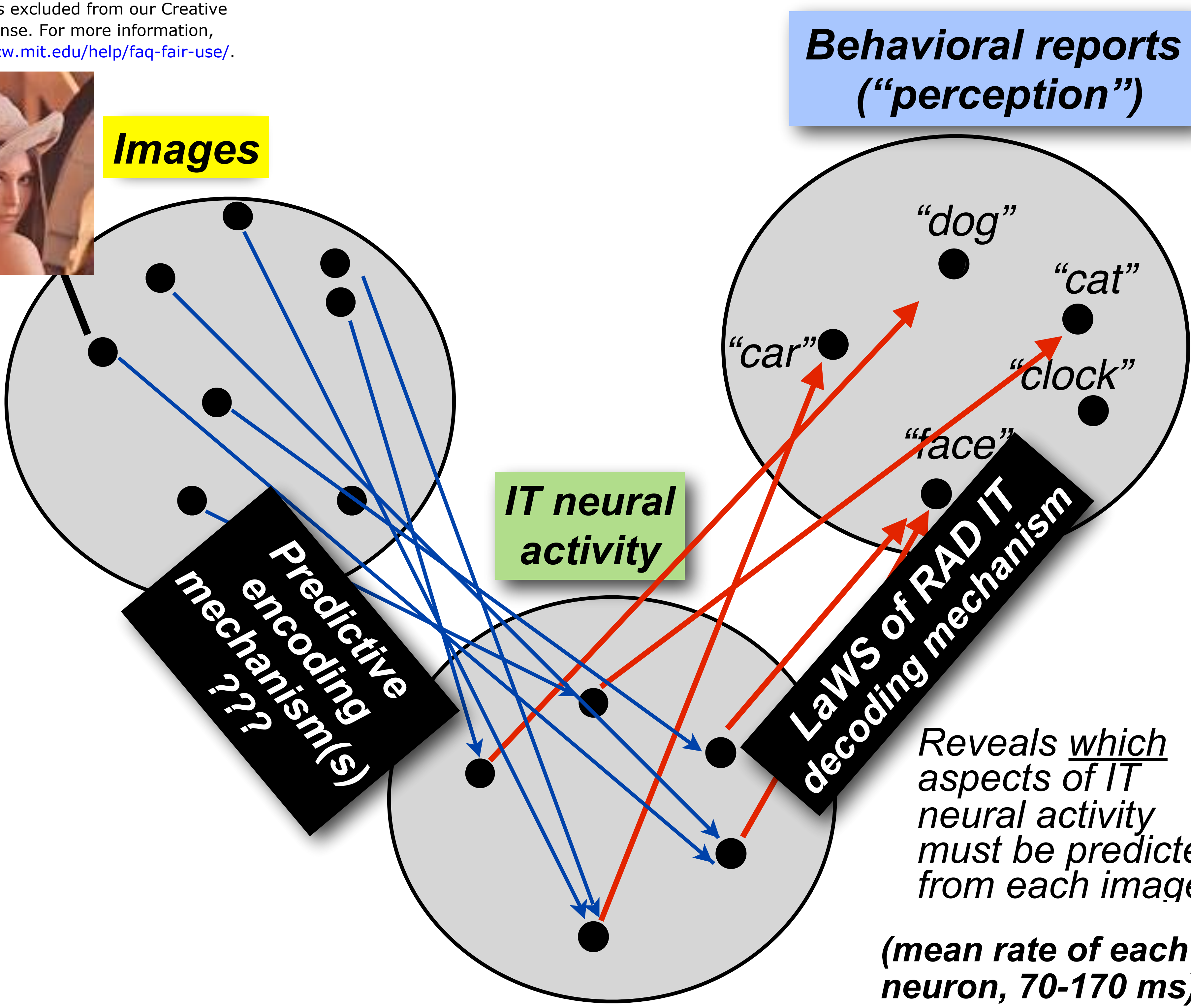
(Domain: core object recognition)

Goal: end-to-end understanding

1. Can we infer the decoding mechanism that the brain uses to support perceptual reports about visually presented object?

   Note: this must predict behavioral report and it must include a falsifiable statement of the relevant aspects of neural activity (aka "neural code")

2. Can we infer the encoding mechanism(s) that accurately predict the relevant ventral stream population patterns of neural activity from each image?

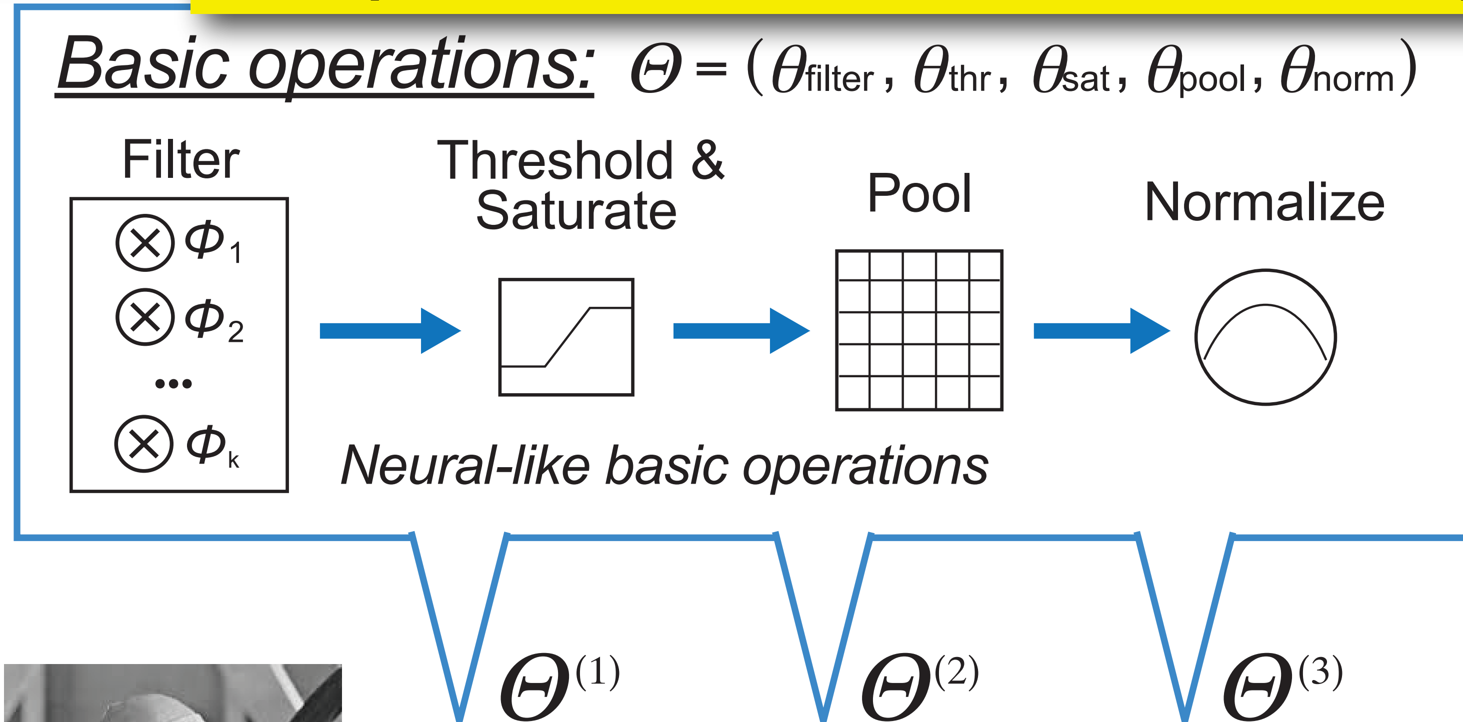


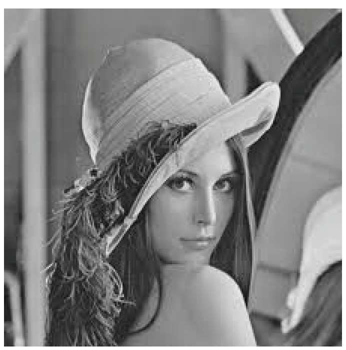
Reveals which aspects of IT neural activity must be predicted from each image

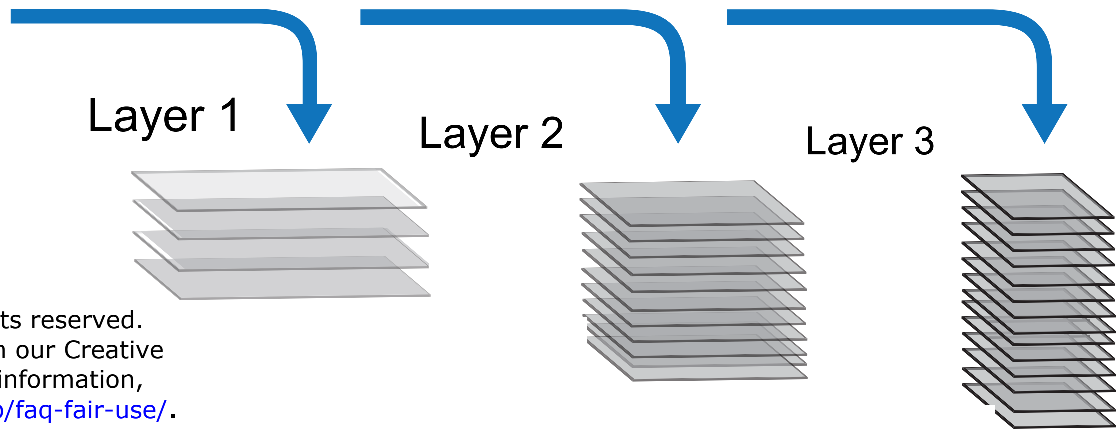
(mean rate of each IT neuron, 70-170 ms)

Our goal (2008): explore a family of possible encoding mechanisms

"Deep convolutional neural networks" (Deep CNN's)







Elements ("neurons") **have large fan-in**

Simple, bio-known non-linearities

Each layer:

is convolutional (i.e. retinotopy)

has many types of tuning functions

Deep stack of layers

Top layer has thousands of visual "neurons"

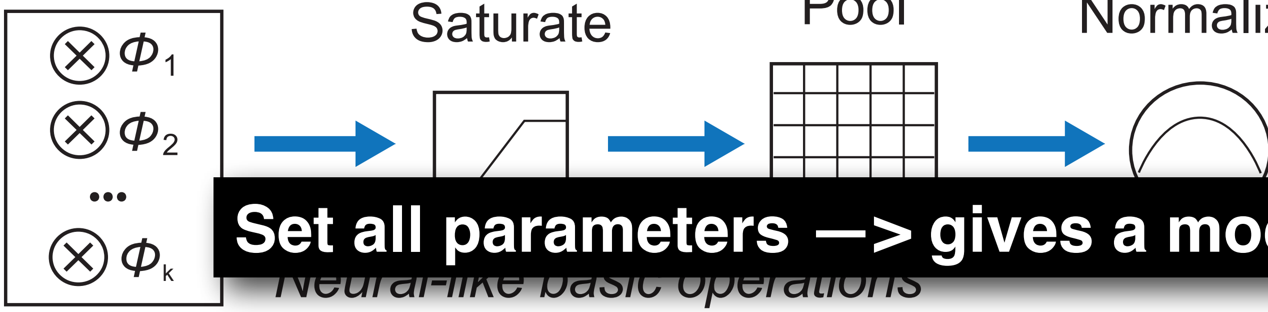
*Pinto, Doukan, DiCarlo & Cox, **PLoS Comp Biol** (2009)*

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

Our goal (2008): explore a family of possible encoding mechanisms

"Deep convolutional neural networks" (Deep CNN's)

Basic operations: $\Theta = (\theta_{filter}, \theta_{thr}, \theta_{sat}, \theta_{pool}, \theta_{norm})$

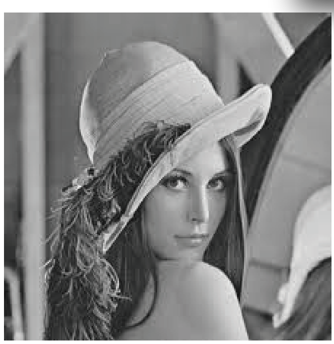


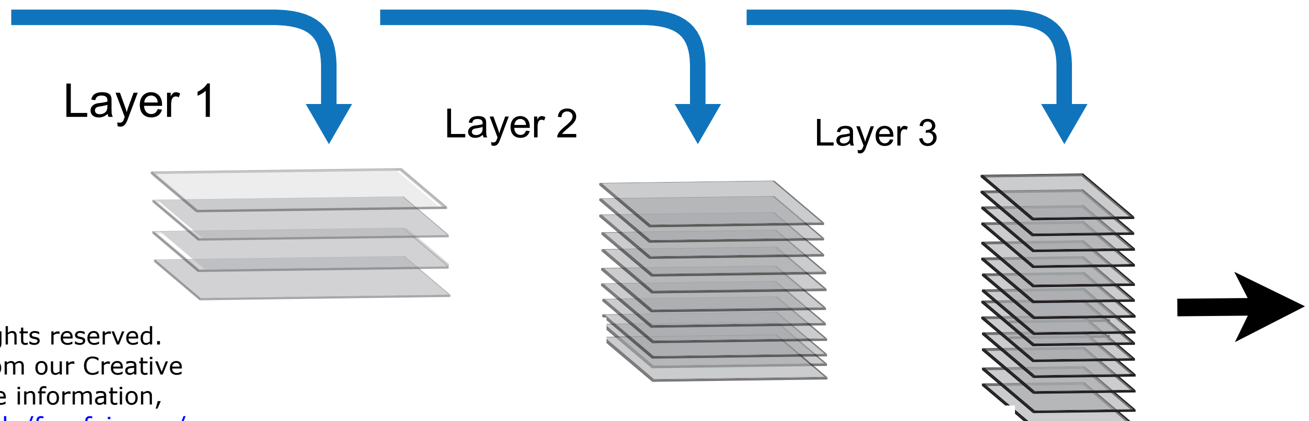
Thousands of unknown parameters

(i.e. not directly determined by neurobiology)

Set all parameters —> gives a model

That model PREDICTS the entire neural population response to ANY image, in each successive visual area





Top layer has thousands of visual "neurons"

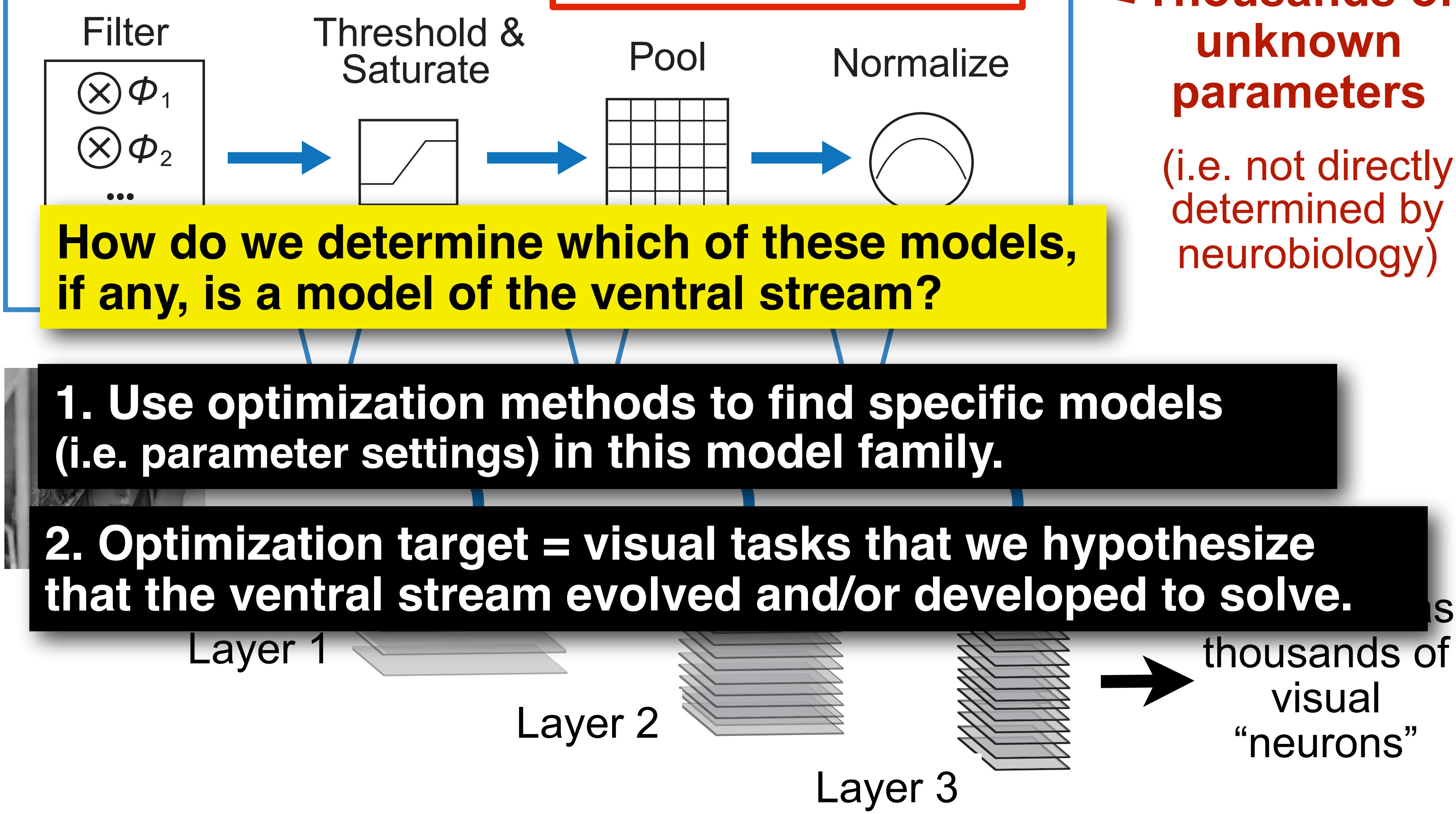
*Pinto, Doukan, DiCarlo & Cox, **PLoS Comp Biol** (2009)*

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

Our goal (2008): explore a family of possible encoding mechanisms

"Deep convolutional neural networks" (Deep CNN's)

Basic operations: $\Theta = (\theta_{filter}, \theta_{thr}, \theta_{sat}, \theta_{pool}, \theta_{norm})$



Thousands of unknown parameters

(i.e. not directly determined by neurobiology)

How do we determine which of these models, if any, is a model of the ventral stream?

1. Use optimization methods to find specific models (i.e. parameter settings) in this model family.

2. Optimization target = visual tasks that we hypothesize that the ventral stream evolved and/or developed to solve.

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

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

2. Optimization target

- **variety of 3D objects (36)** with semantic breadth (e.g. not all faces)
- rendered with large amount of **variation**
- These are **different objects** that those we will use later in testing

Nine example objects:

Bodies

Buildings

Flowers

Guns

Instruments

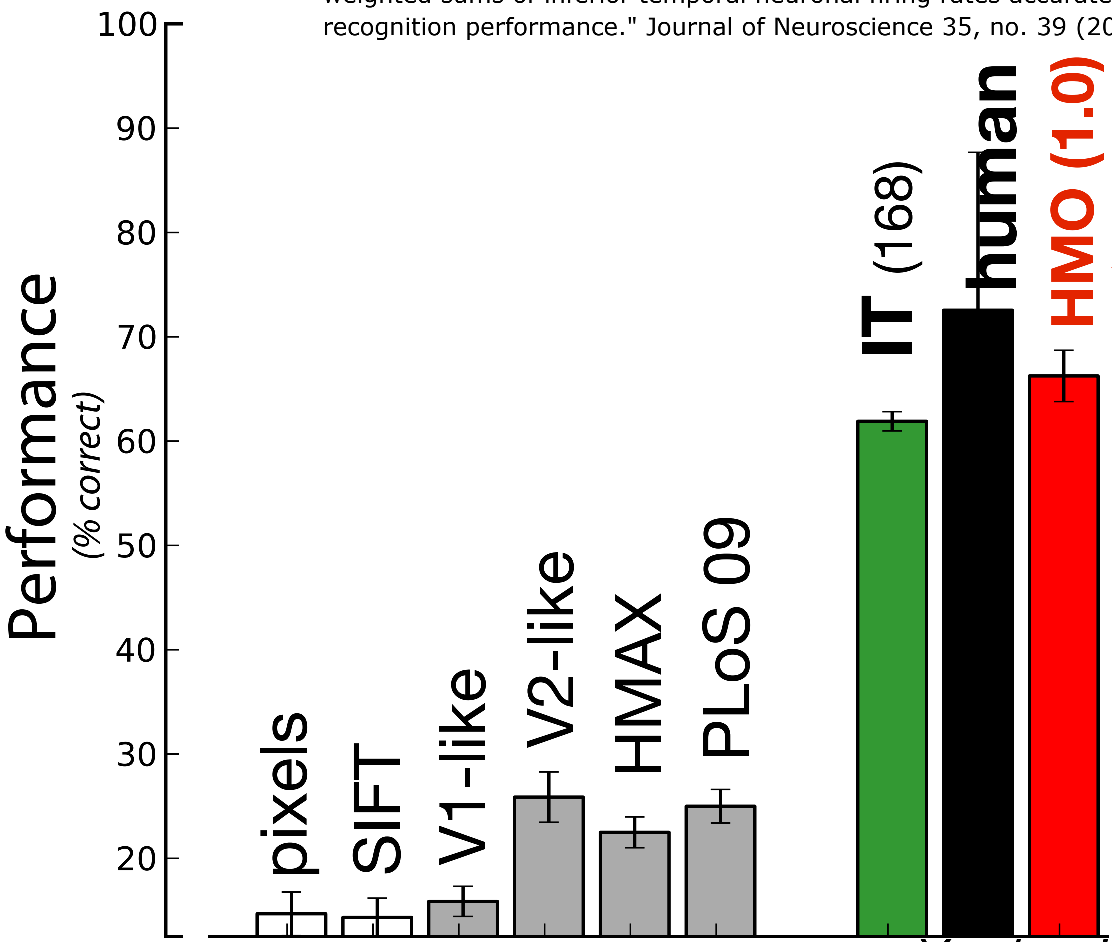
Jewelry

Shoes

Tools

Trees

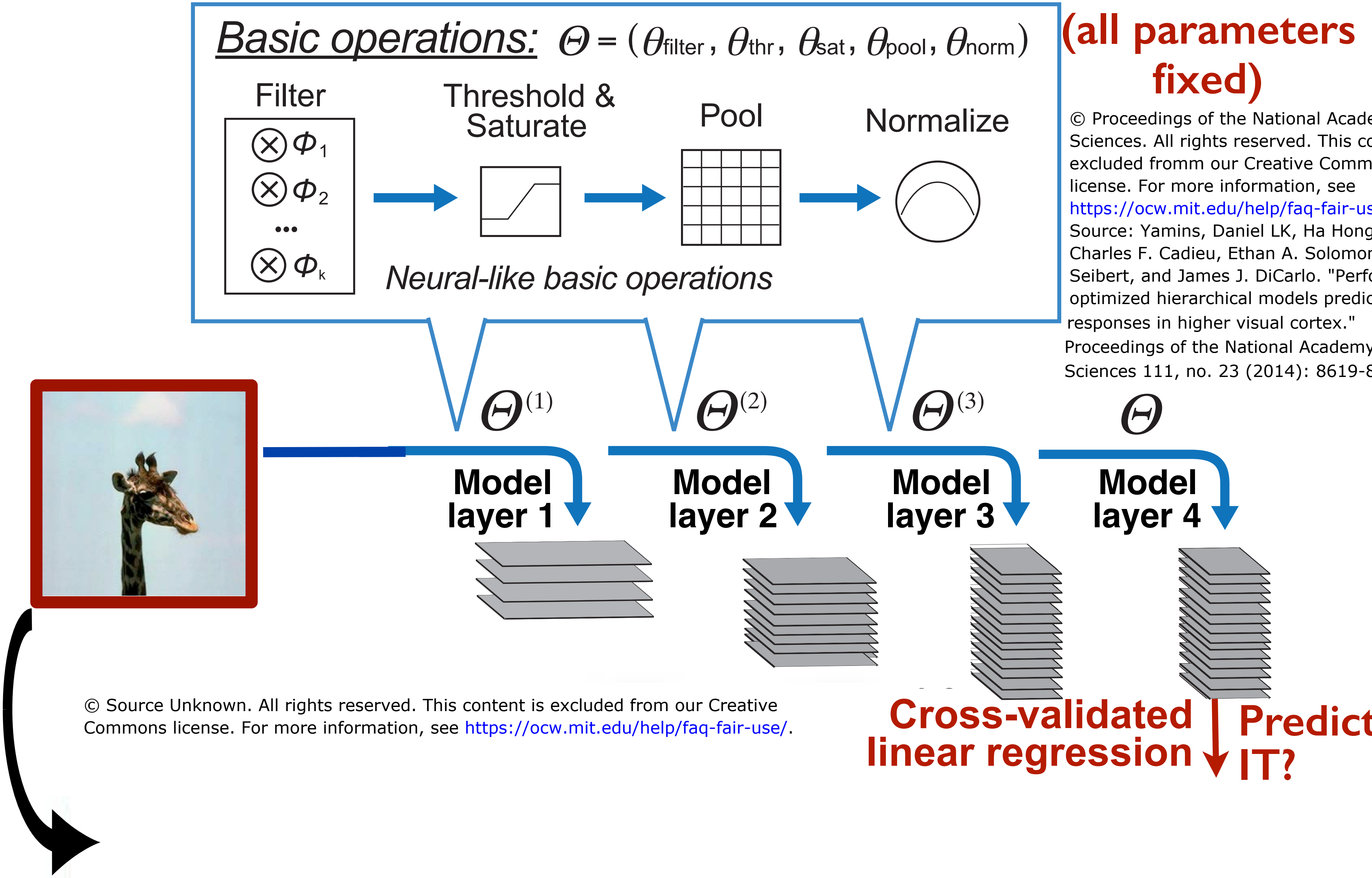
Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.



First algorithm we discovered using this approach (2012)

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

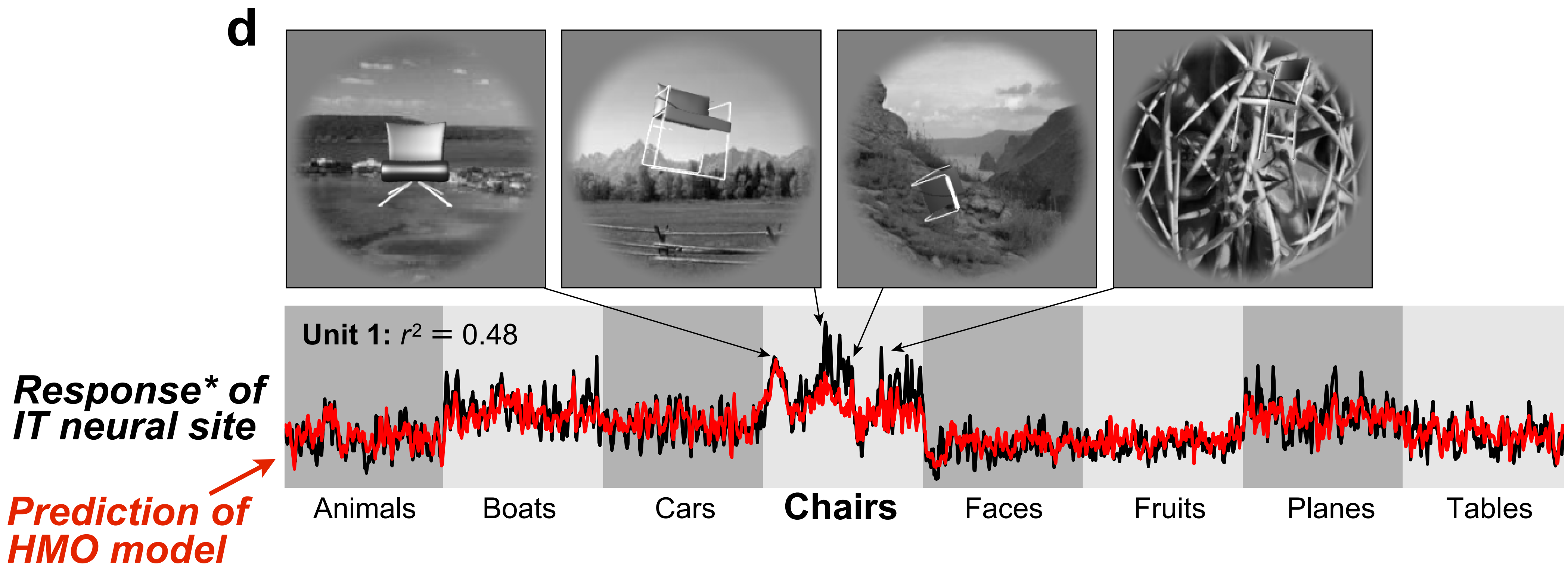
HM0 1.0 (all parameters fixed)



Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon,Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.

Predictions of single site IT responses from layer 4 of HMO 1.0 model

These are PREDICTIONS: All of these objects and images were never previously seen by the HMO model



Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.

(* mean rate 70-170 ms after image onset)

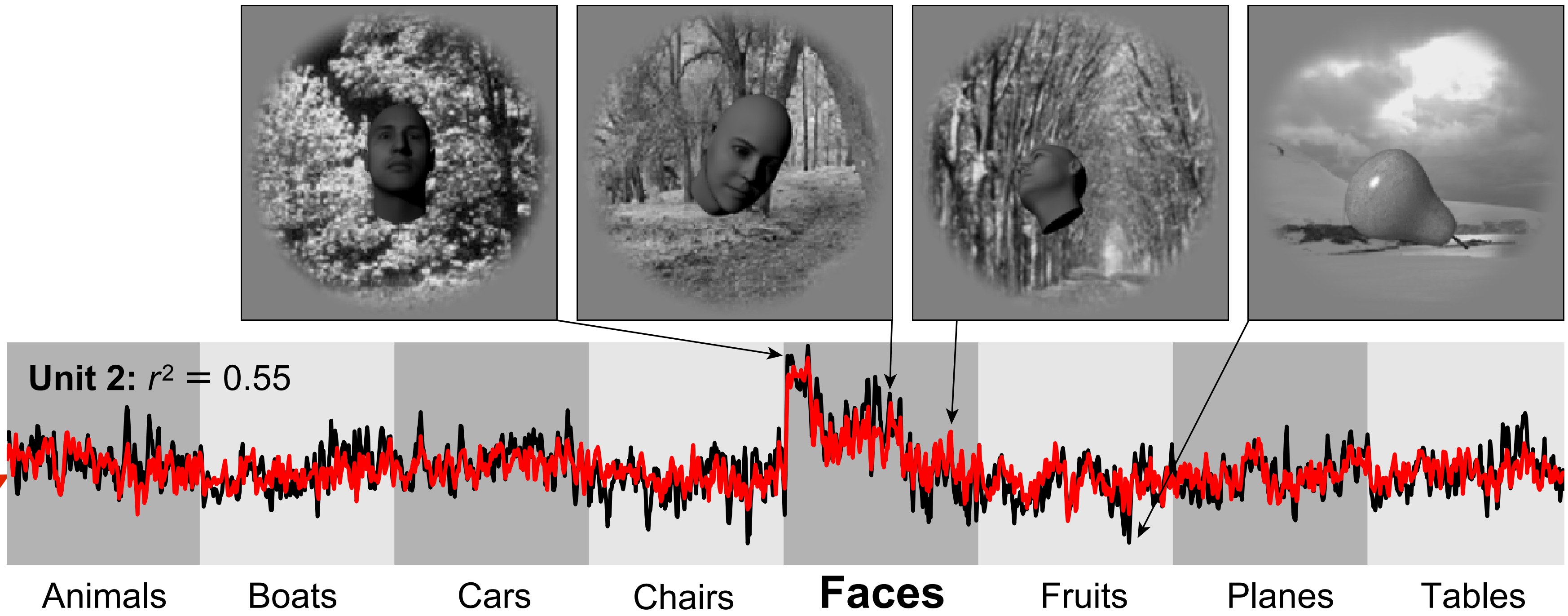
Yamins, Hong, Solomon, Seibert and DiCarlo ***PNAS (2014)***

Predictions of single site IT responses from layer 4 of HMO 1.0 model

These are PREDICTIONS: All of these objects and images were never previously seen by the HMO model

Response of IT neural site

Prediction of HMO model



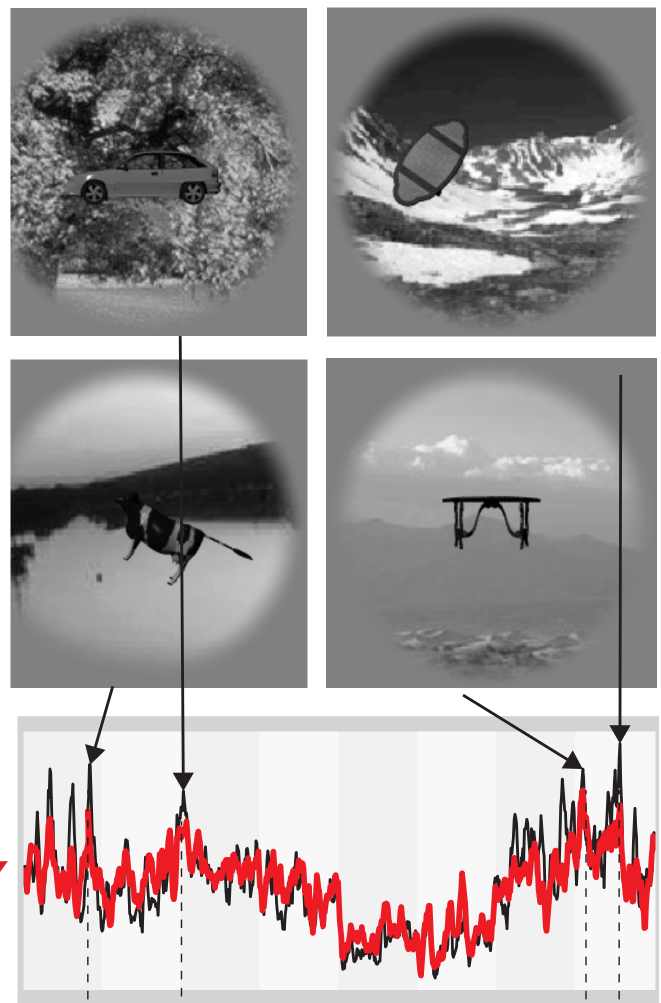
Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.

(* mean rate 70-170 ms after image onset)

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

Predictions of single site IT responses from layer 4 of HMO 1.0 model

IT Site 42



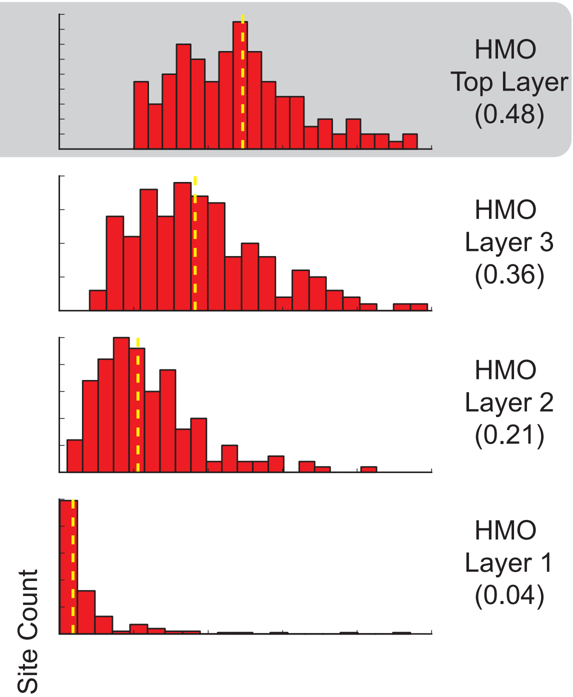
Response
IT neural s

Prediction
HMO mod

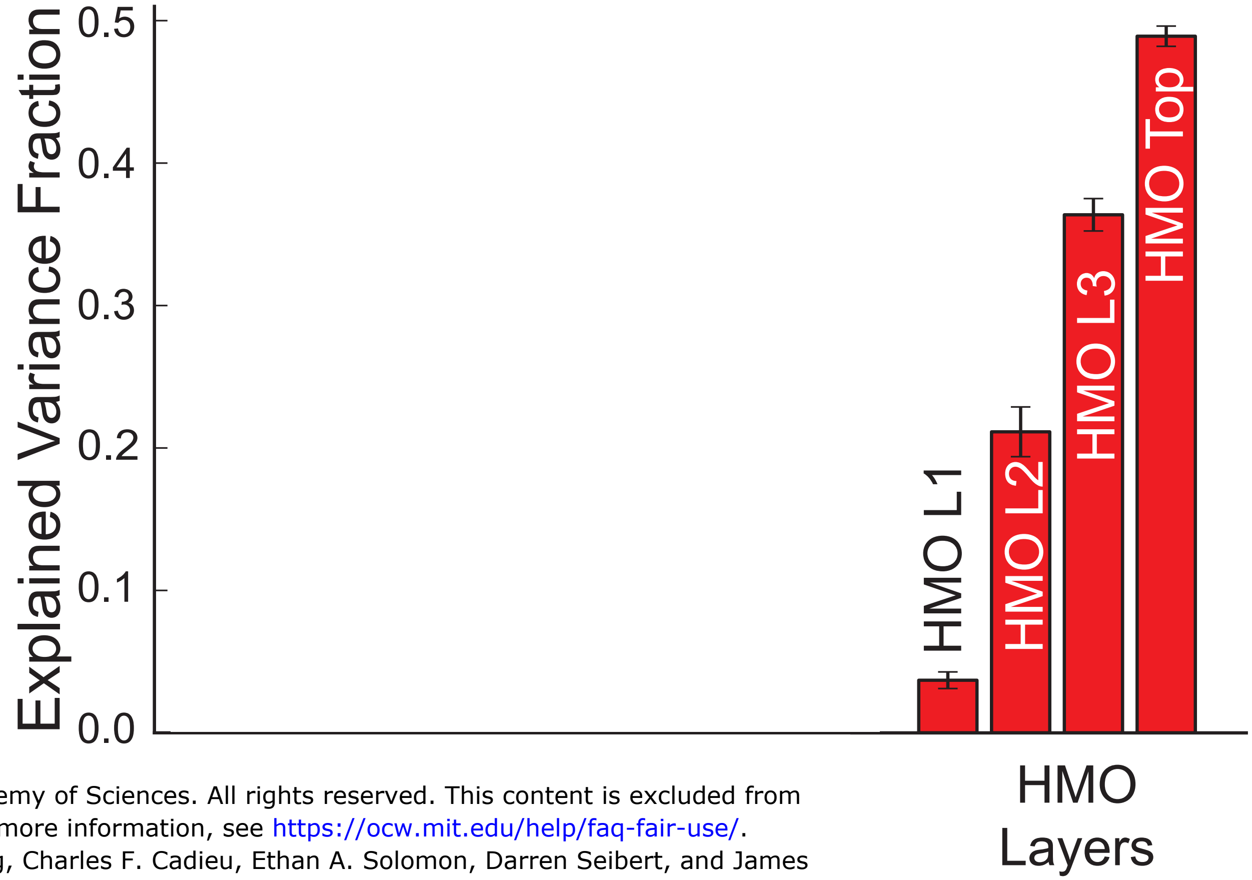
Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.

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

Ability of various encoding mechanisms (specific models) to predict IT responses to naturalistic images



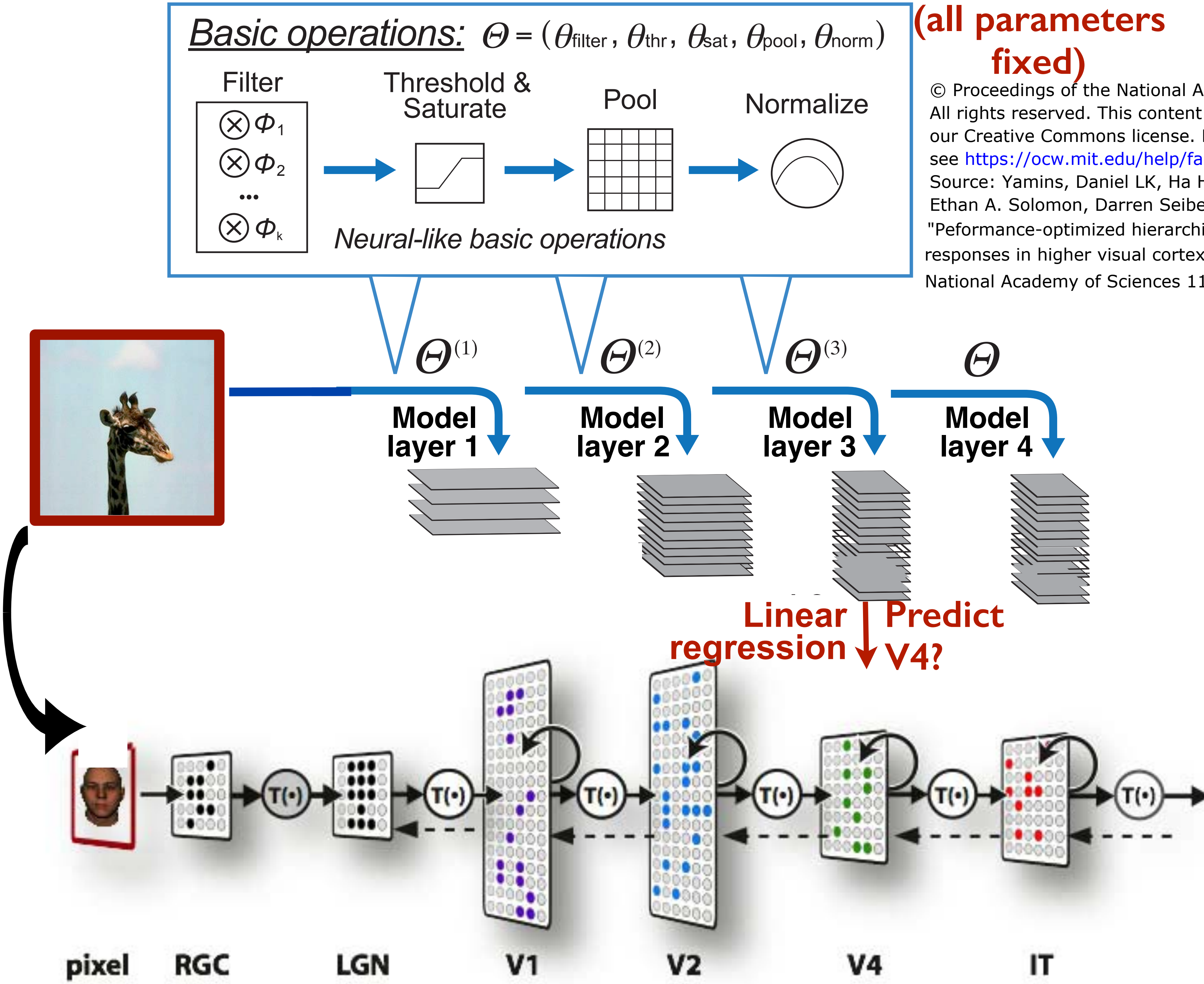
~50% of IT single unit response variance predicted.
Dramatic improvement over previous models.



Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.

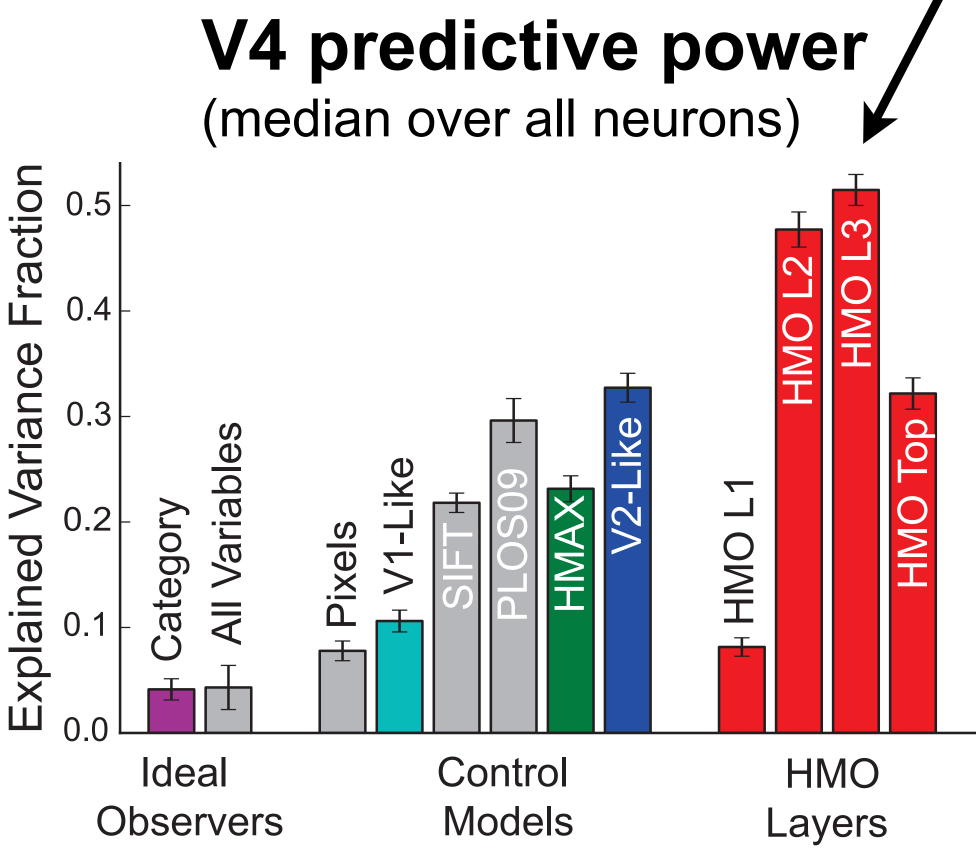
*Yamins, Hong, Solomon, Seibert and DiCarlo **PNAS (2014)***

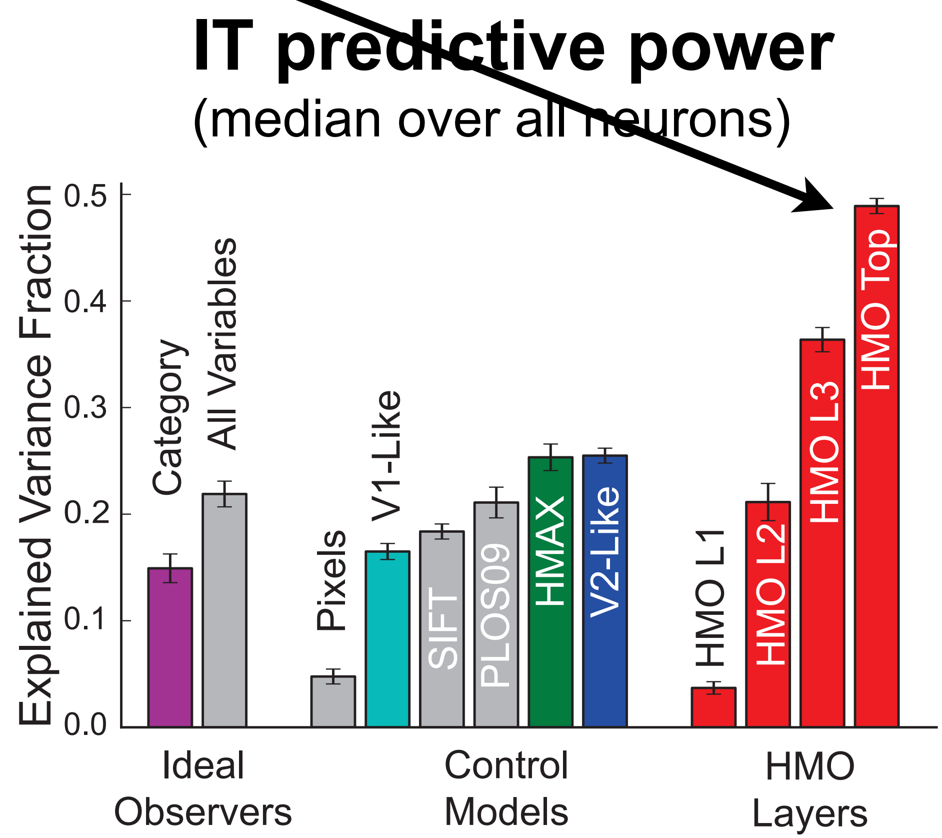
HM0 1.0 (all parameters fixed)



Bio-inspired algorithm class + tasks in domain + optimization ==> neural-like encoding functions!

Even in intermediate layers!

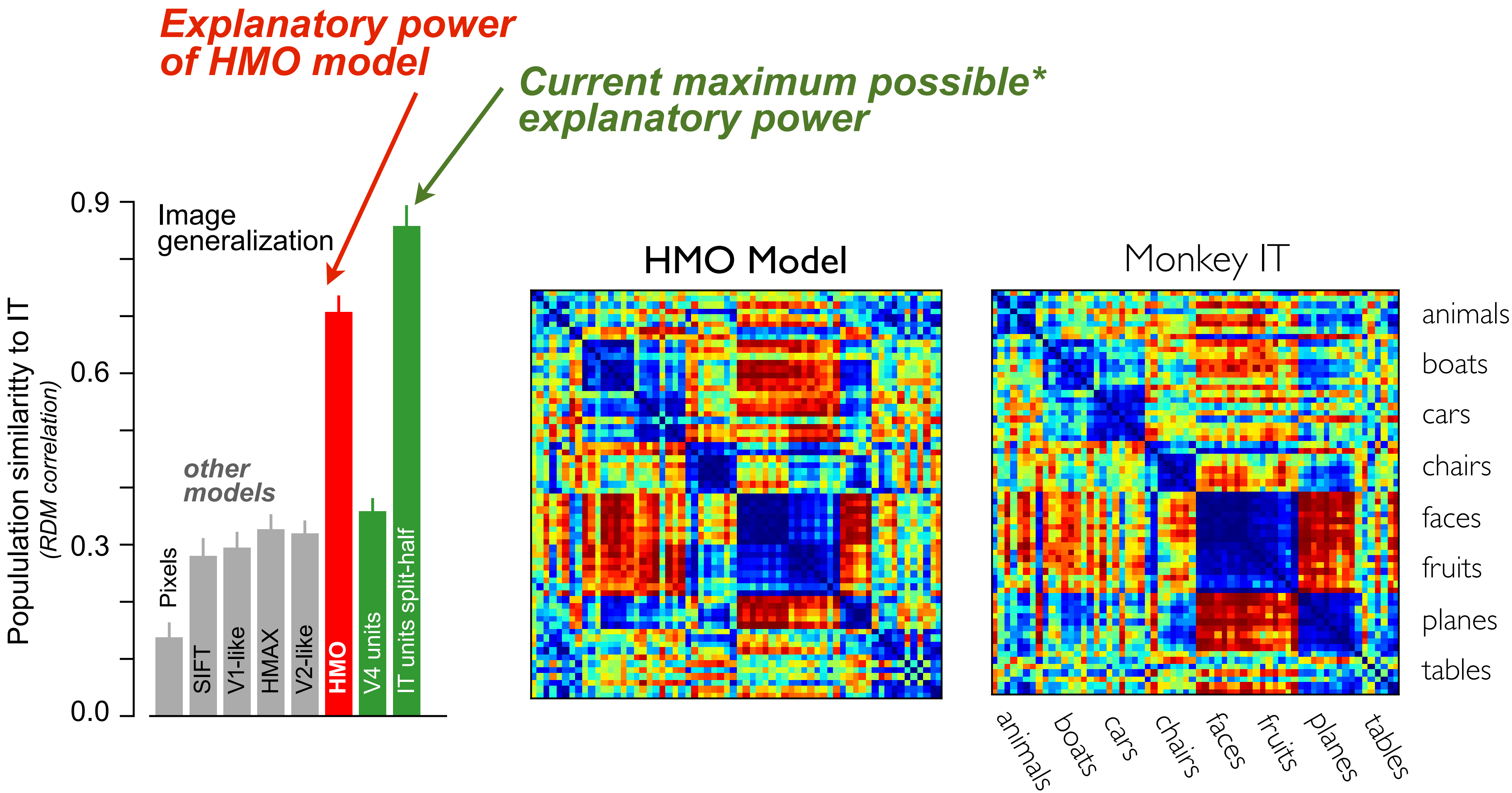




Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.

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

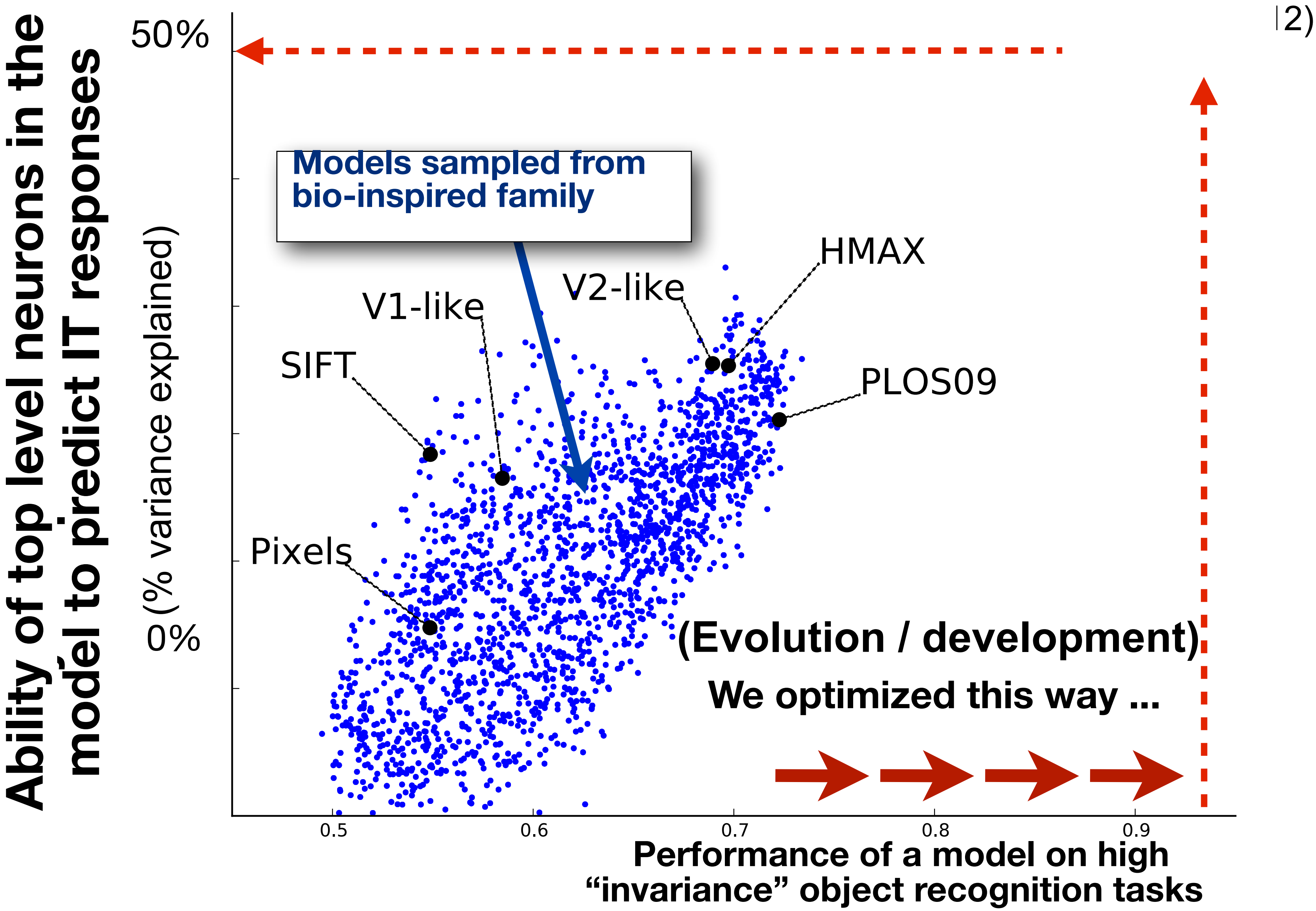
Representation Dissimilarity Matrices (Kriegeskorte, 2008)



Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.

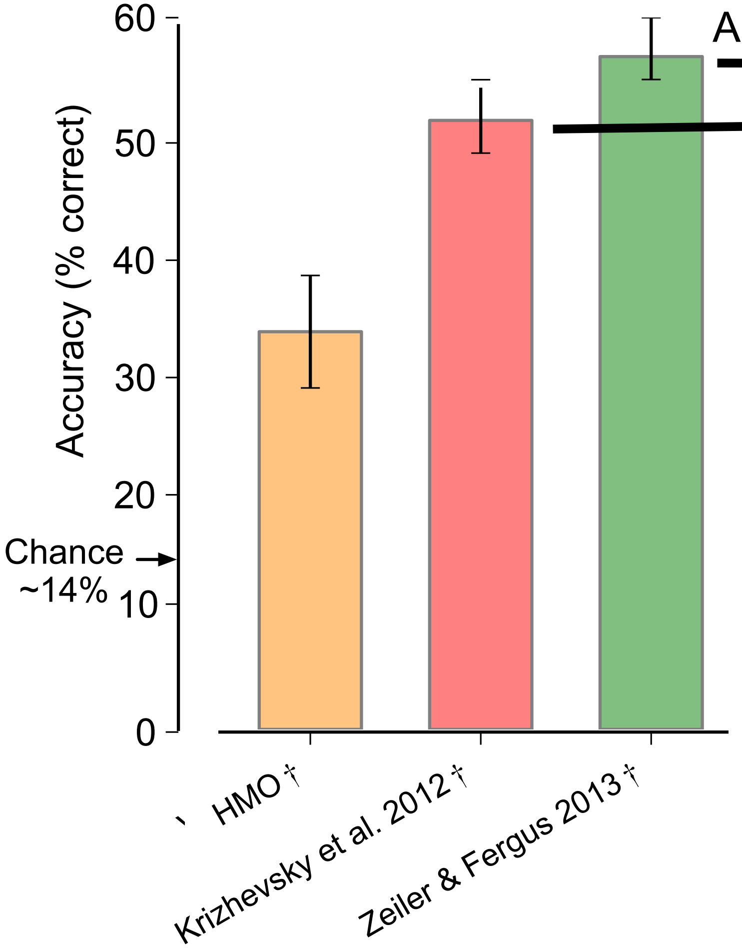
Yamins, Hong, Solomon, Seibert and DiCarlo ***PNAS (2014***)

Suggests that continued optimization within this family of models would lead to even higher neural predictive power.



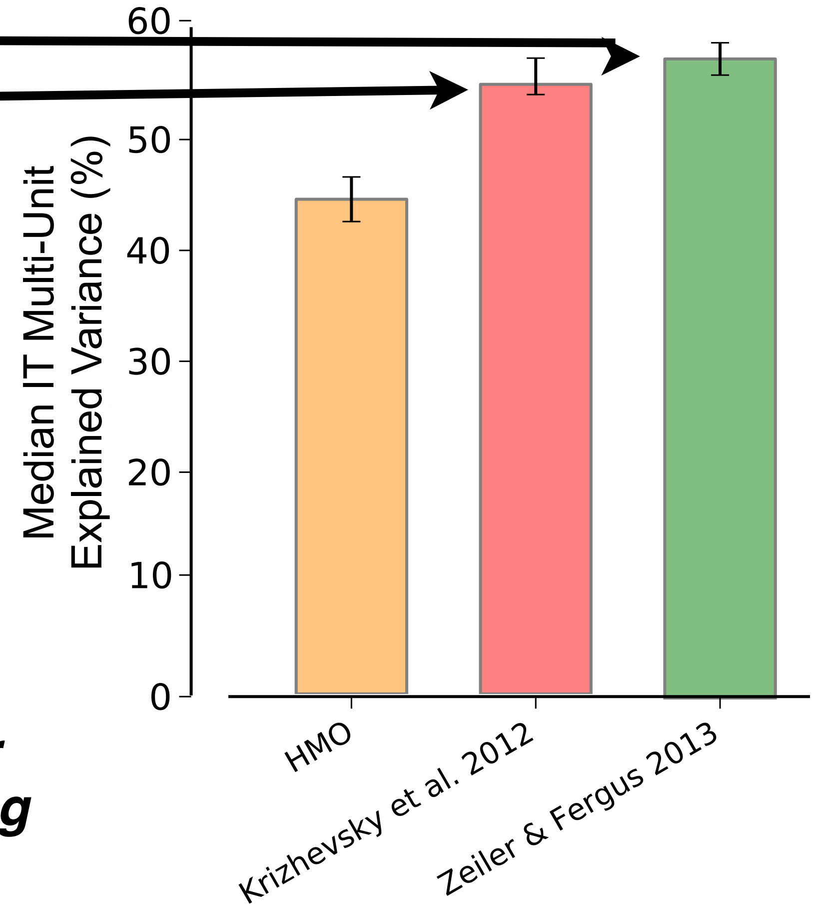
Source: Yamins, Daniel LK, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proceedings of the National Academy of Sciences 111, no. 23 (2014): 8619-8624.

Suggests that continued optimization within this family of models would lead to even higher neural predictive power.



Better performance on our tasks again leads to better neural predictive power

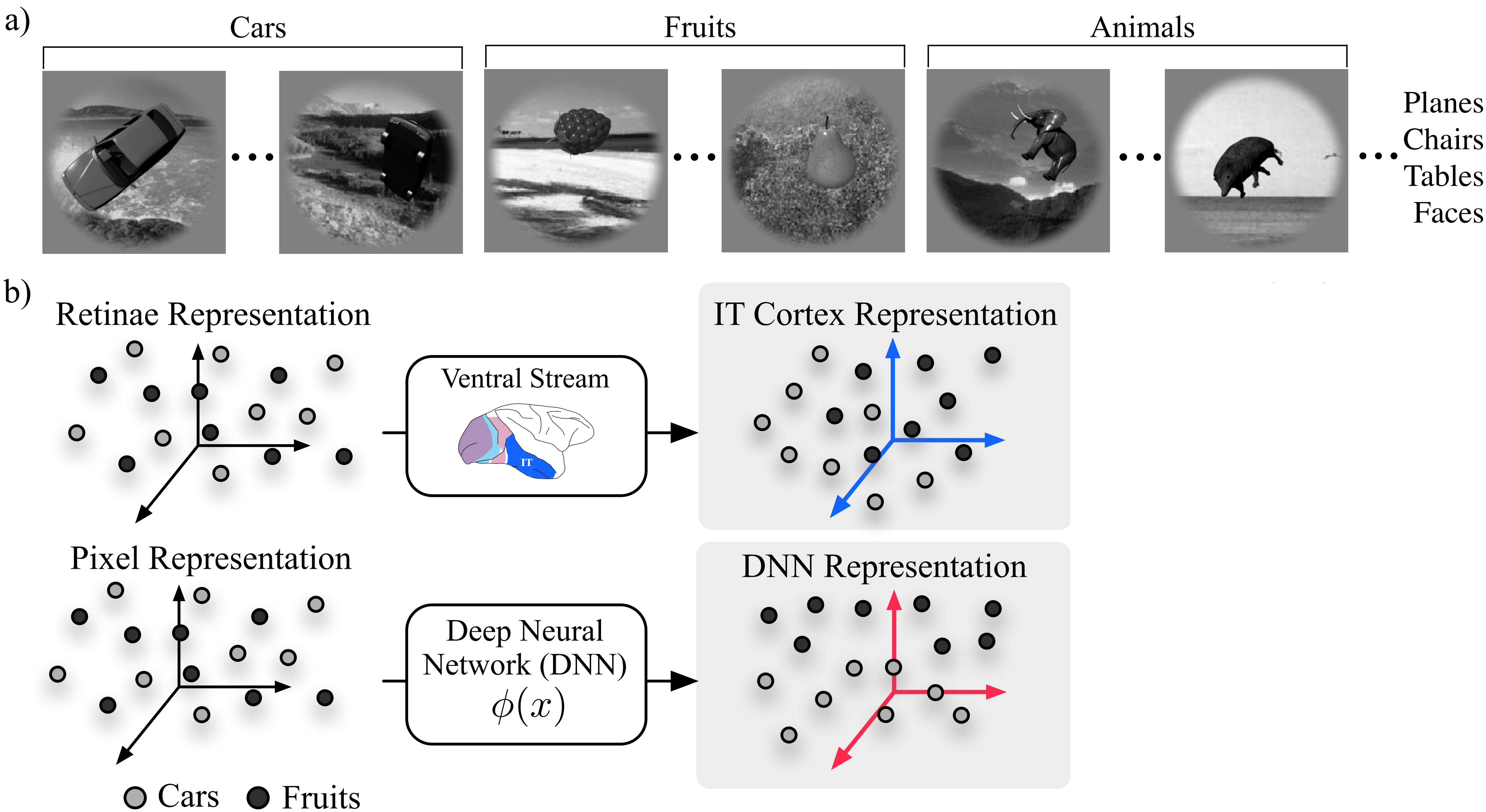
(even when other groups are driving up the model performance!)



Cadieu, Charles F., Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. "Deep neural networks rival the representation of primate IT cortex for core visual object recognition. "PLoS Comput Biol 10, no. 12 (2014): e1003963; https://doi.org/10.1371/journal.pcbi.1003963. License CC BY.

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ***ICLR*** *(2013);*
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ***PLoS Comp Bio*** *(2014)*

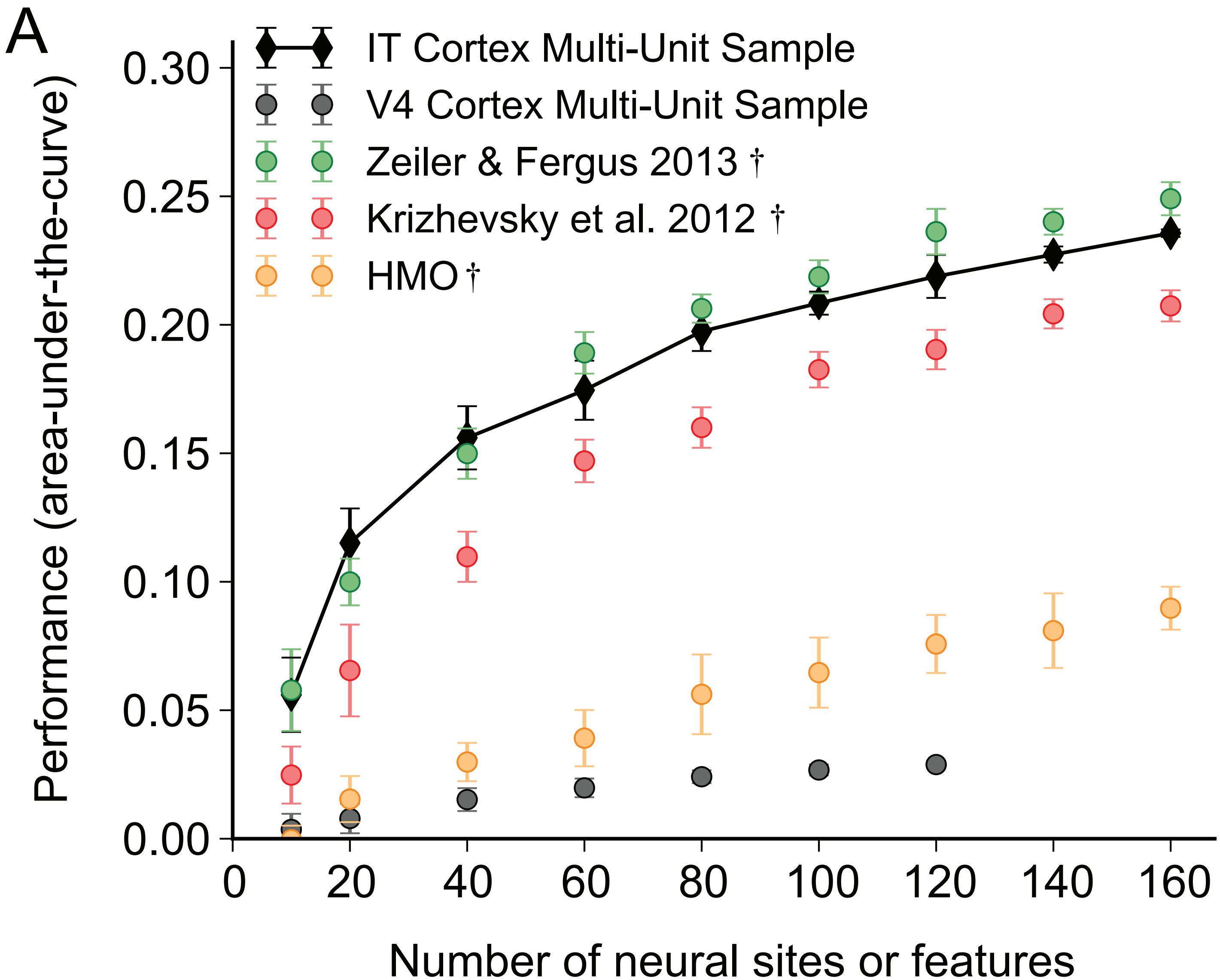
CNN features vs. IT "features"



Cadieu, Charles F., Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. "Deep neural networks rival the representation of primate IT cortex for core visual object recognition. "PLoS Comput Biol 10, no. 12 (2014): e1003963; https://doi.org/10.1371/journal.pcbi.1003963. License CC BY.

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ***ICLR*** *(2013);*
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ***PLoS Comp Bio*** *(2014)*

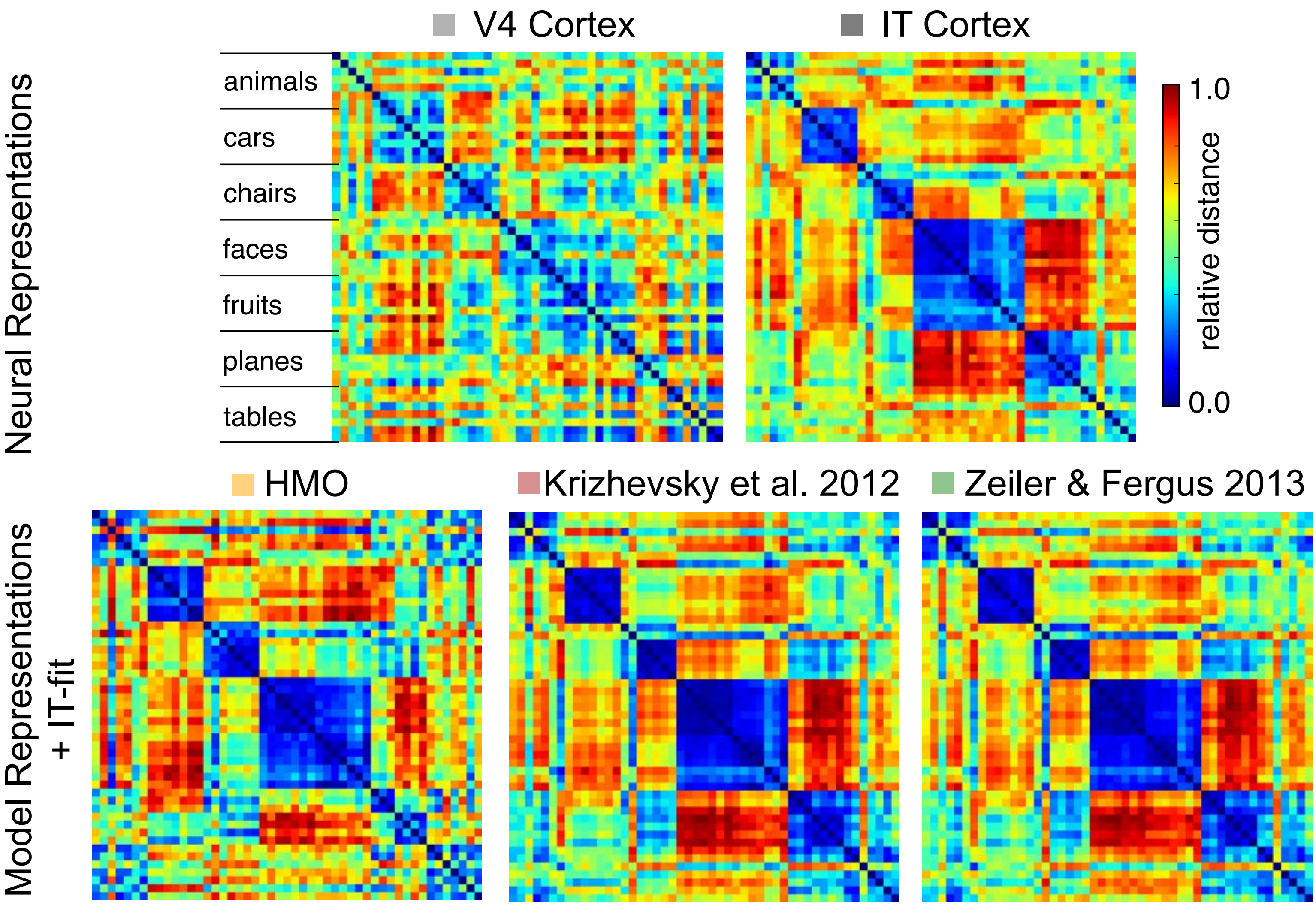
CNN features vs. IT "features"



Cadieu, Charles F., Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. "Deep neural networks rival the representation of primate IT cortex for core visual object recognition. "PLoS Comput Biol 10, no. 12 (2014): e1003963; https://doi.org/10.1371/journal.pcbi.1003963.

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ***ICLR*** *(2013);*
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ***PLoS Comp Bio*** *(2014)*

Better performing deep CNN networks also better predict the patterns of IT neural responses



Cadieu, Charles F., Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. "Deep neural networks rival the representation of primate IT cortex for core visual object recognition. "PLoS Comput Biol 10, no. 12 (2014): e1003963; https://doi.org/10.1371/journal.pcbi.1003963.

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ***ICLR*** *(2013);*
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ***PLoS Comp Bio*** *(2014)*

Summary of what I presented today (Domain: Core recognition)

1. Showed that IT firing rates are a feature basis on which learned object judgements naturally predict human/monkey performance; defined parameters.

LaWS of RAD IT
[70-170ms, 50,000n, 100t]

Inference: this might be the specific neural code and decoding mechanism that the brain uses to support these tasks.

Systematic causal tests of this model ongoing, but results thus far are as predicted by the model …

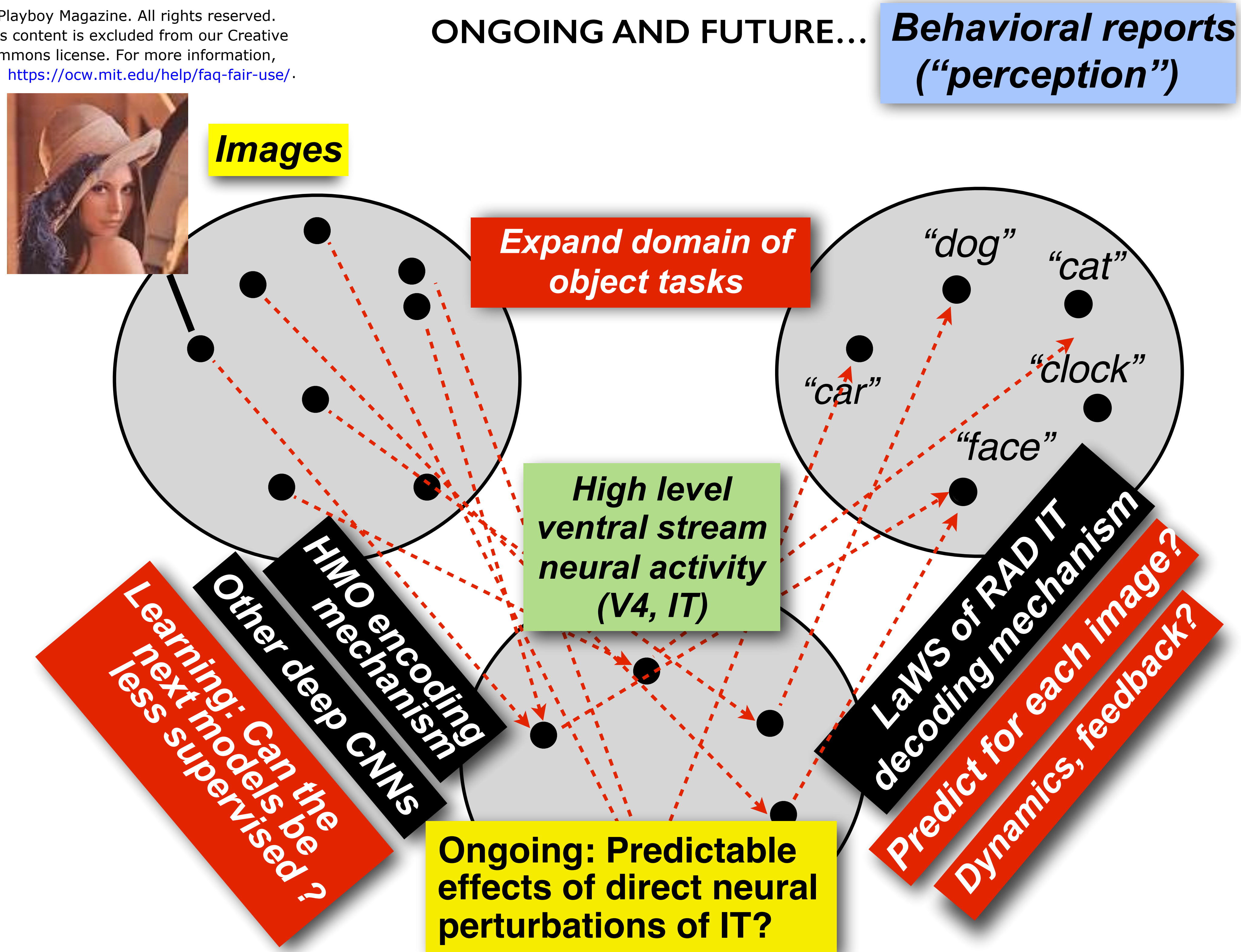
2. Showed that optimization of deep CNNs (models) for invariant object recognition tasks led to dramatic improvements in our ability to predict IT and V4 neural responses.

HMO 1.0, CNN 2.0

Inference: the encoding mechanisms in these models are similar to those at work in the ventral stream.

This is allowing the field to design experiments to explore what remains unique and powerful about primate object perception.

ONGOING AND FUTURE…



Acknowledgements

Current lab members:

Arash Afraz
Diego Ardila
Ha Hong
Elias Issa
Xiaoxuan Jia
Hyodong Lee
Shay Ohayon
Rishi Rajalingham
Kailyn Schmidt
Darren Seibert
Chris Stawarz
Dan Yamins

Key alumni:

Charles Cadieu
Najib Majaj
Ethan Soloman

Contributing labs:

Ed Boyden (MIT)
David Cox (Harvard)
Bob Desimone (MIT)
Tomaso Poggio (MIT)
Nancy Kanwisher (MIT)
Wim Vanduffel (MGH, KU L.)

- *NIH NEI*
- *NSF*
- *DARPA / ONR*
- *Simons Foundation*
- *McGovern Institute*

Resource: Brains, Minds and Machines Summer Course
Tomaso Poggio and Gabriel Kreiman