Using Artificial-Intelligence-Driven Deep Neural Networks to Uncover Principles of Brain Representation and Organization

2017.04.27

Daniel Yamins
Stanford Neurosciences Institute
Stanford Artificial Intelligence Laboratory
Departments of Psychology and Computer Science
Stanford University

The origin of the problem

Understanding complex, noisy data streams is a critical part of cognition.

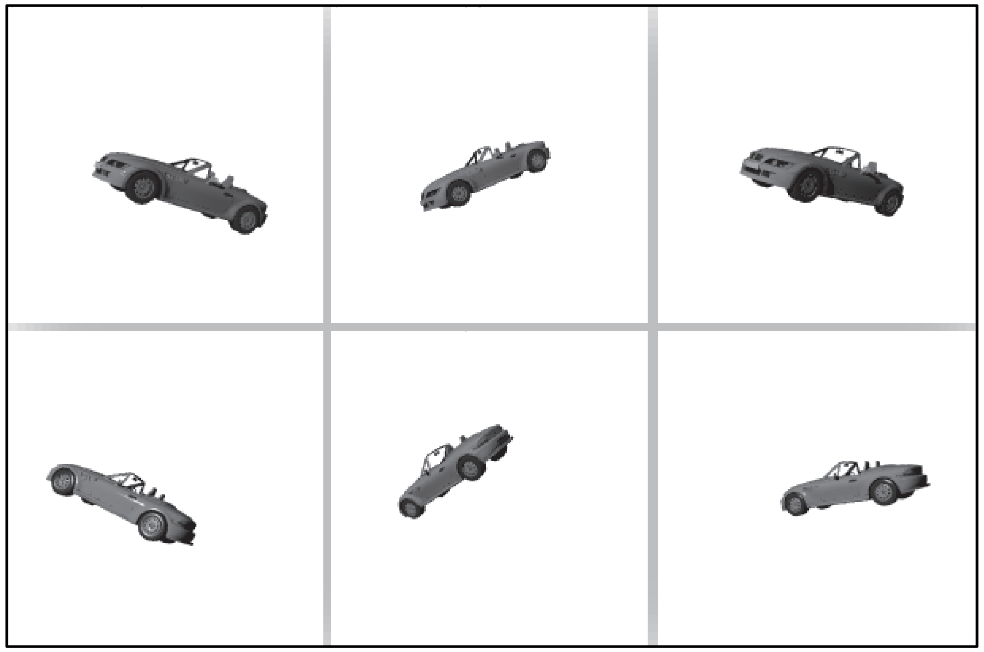
The origin of the problem

Understanding complex, noisy data streams is a critical part of cognition.

"Mercedes behind Lamborghini, on a field in front of mountains."

Variation Makes Object Recognition Challenging

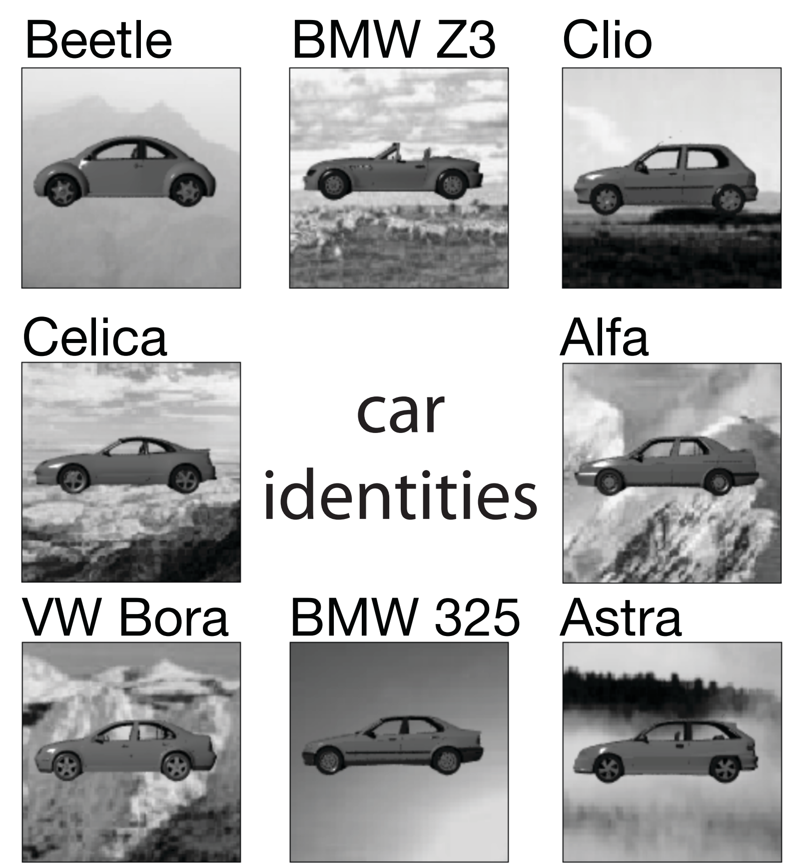
View: position, size, pose, illumination



Background variation

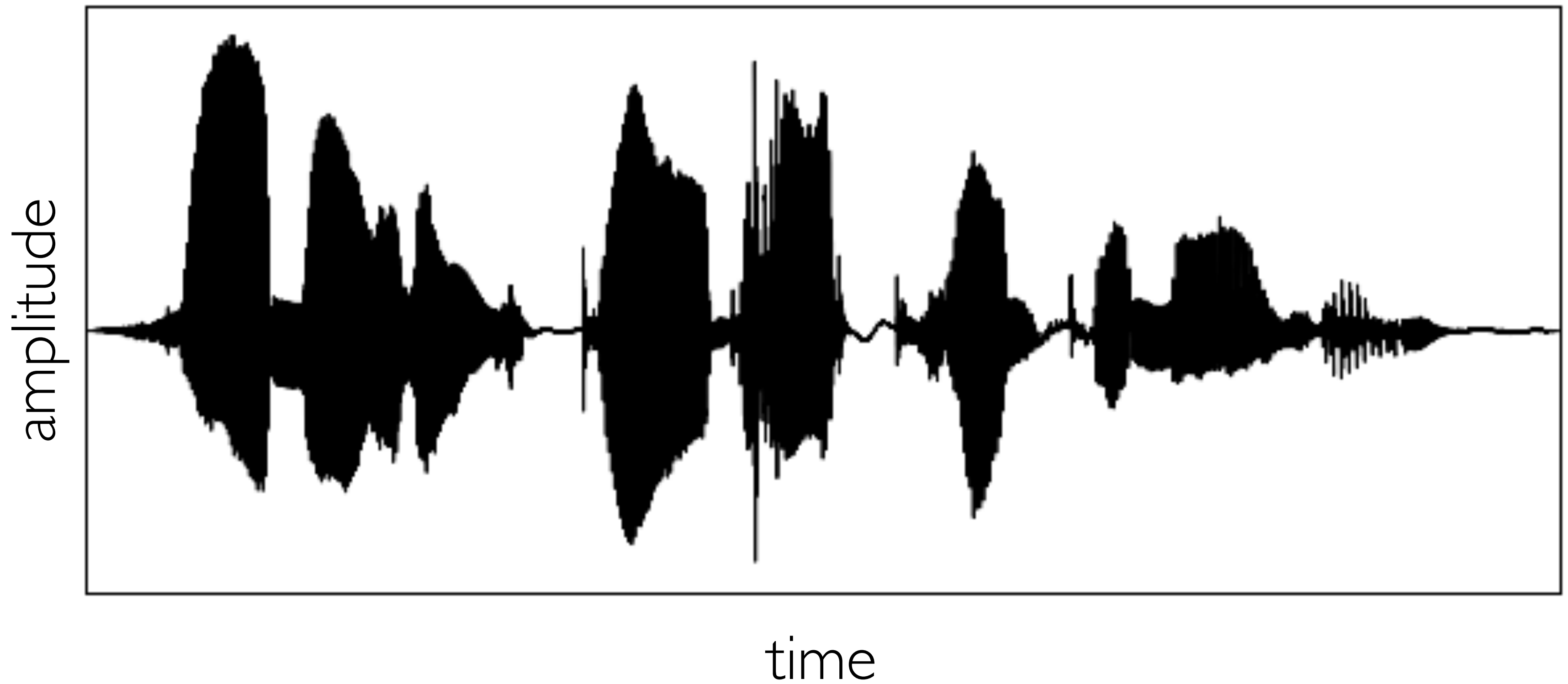
Distortion & Noise

Geometric variation



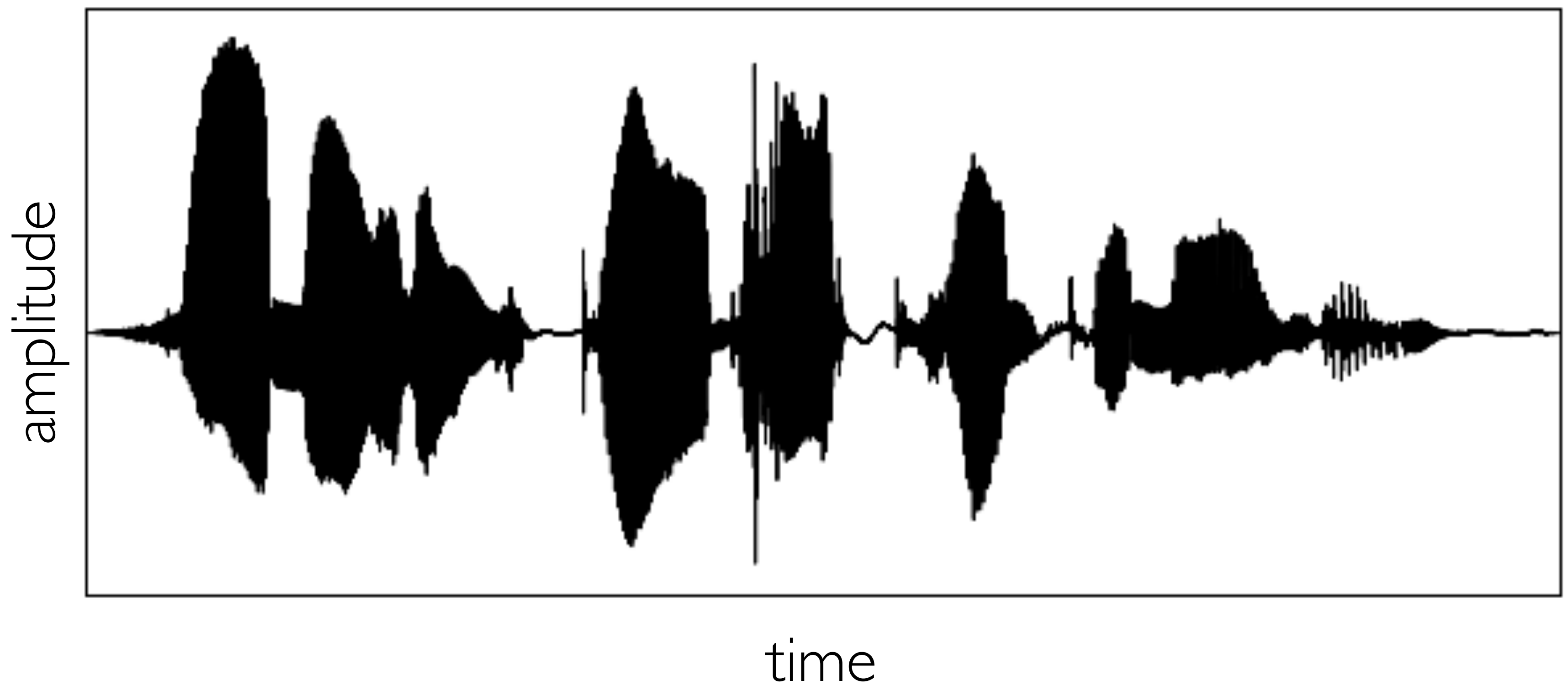
The origin of the problem

Understanding complex, noisy data streams is a critical part of cognition.



The origin of the problem

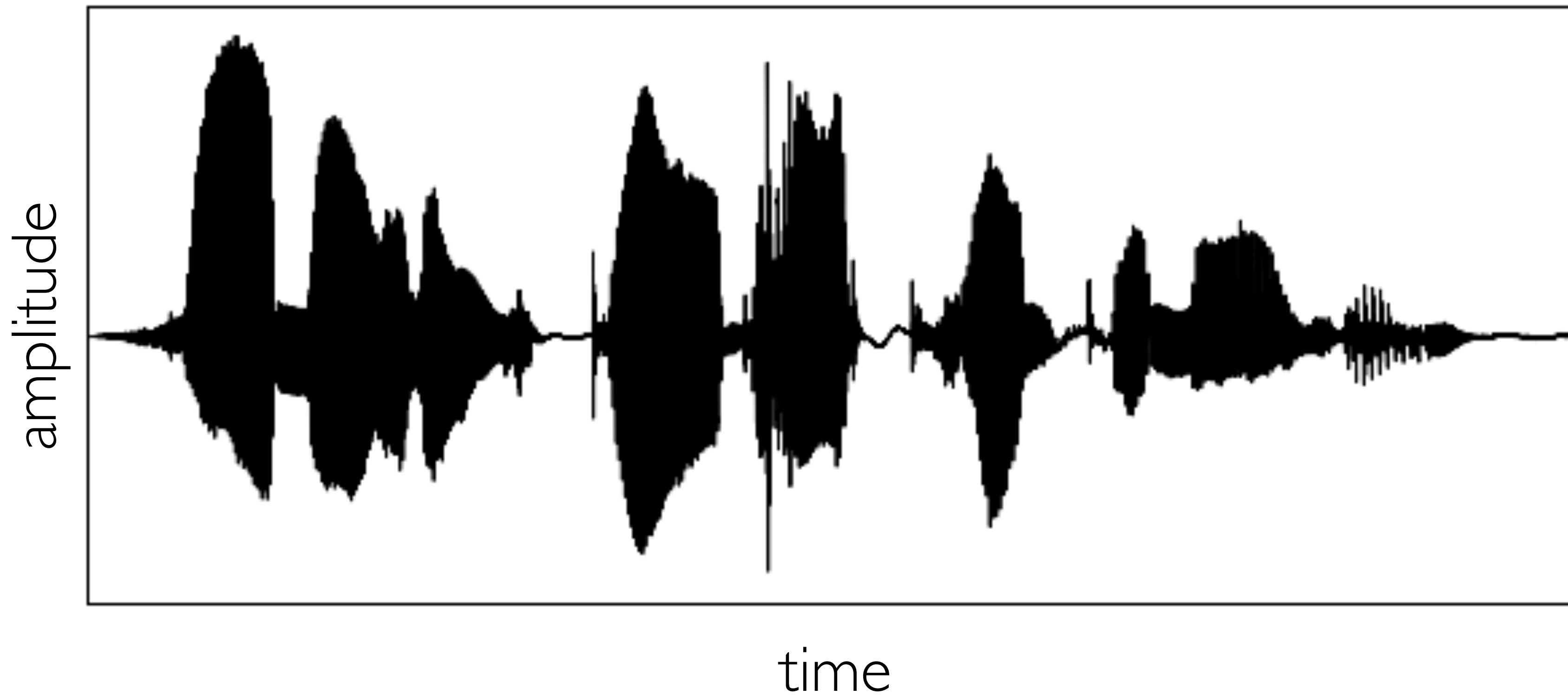
Understanding complex, noisy data streams is critical part of cognition.



"Hannah is good at compromising."

The origin of the problem

Understanding complex, noisy data streams is critical part of cognition.



"Hannah is good at compromising."

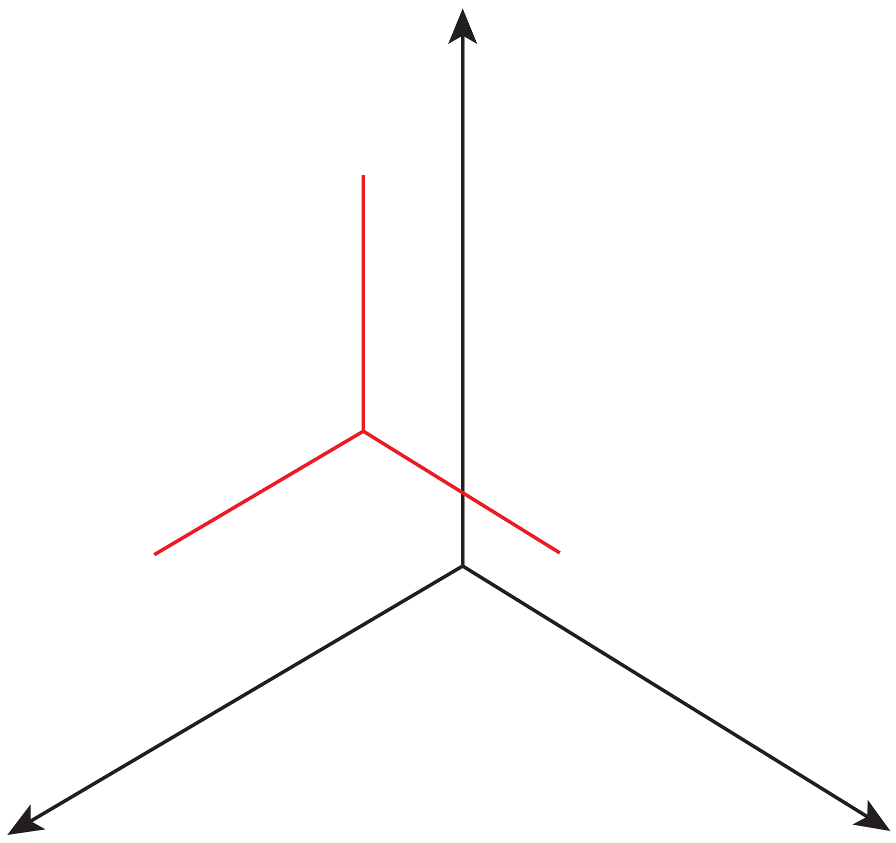
variation sources: speaker identity
background noise
reverberation
...

The origin of the problem

Axes of natural variation of natural
"physics" representation of world

e.g.

retinal photoreceptor voltage
or hair-cell point amplitudes

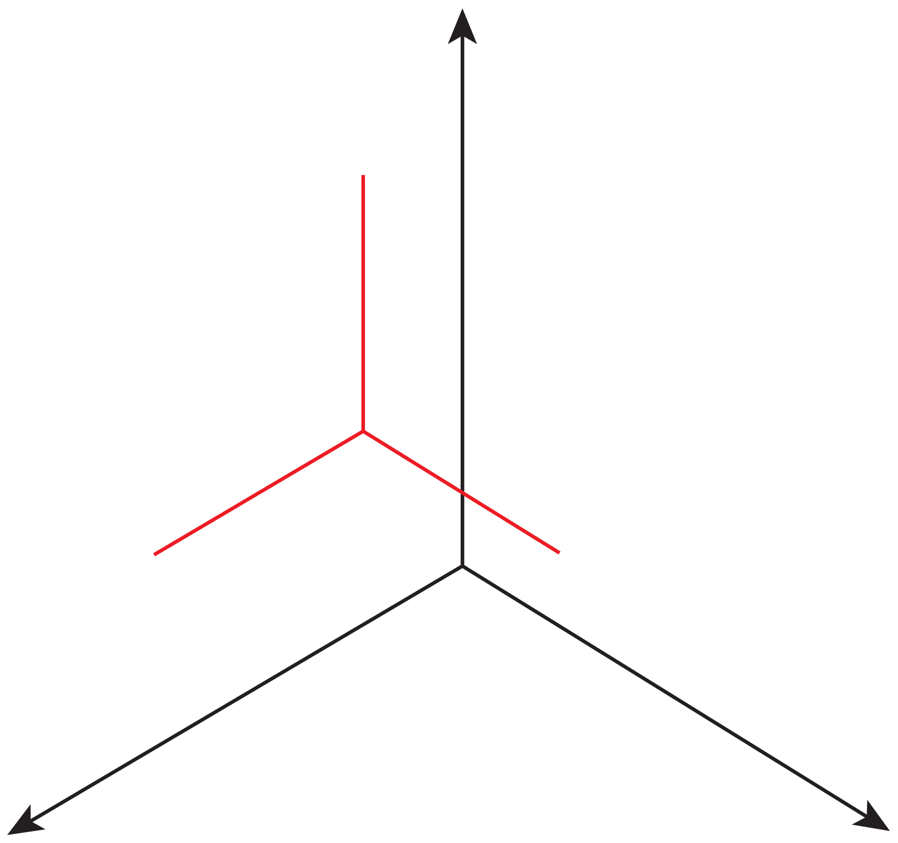


The origin of the problem

Axes of natural variation for natural **behavioral** events

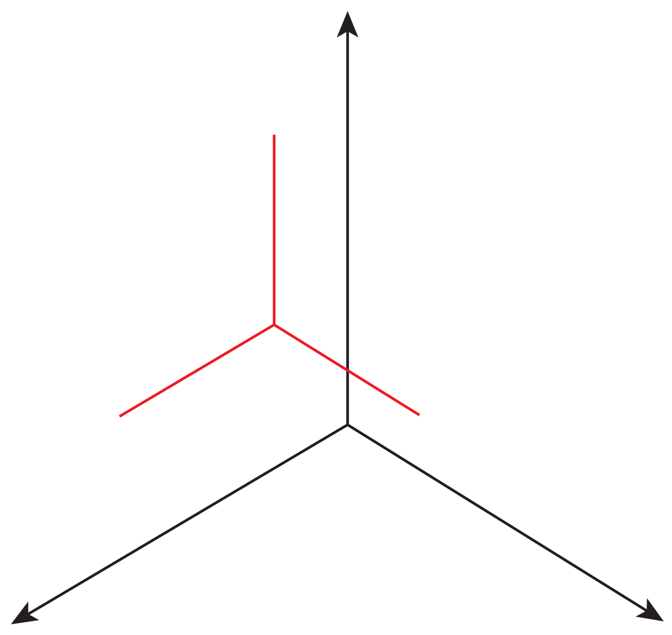
e.g.

deforming face moving in complex-lighted environment



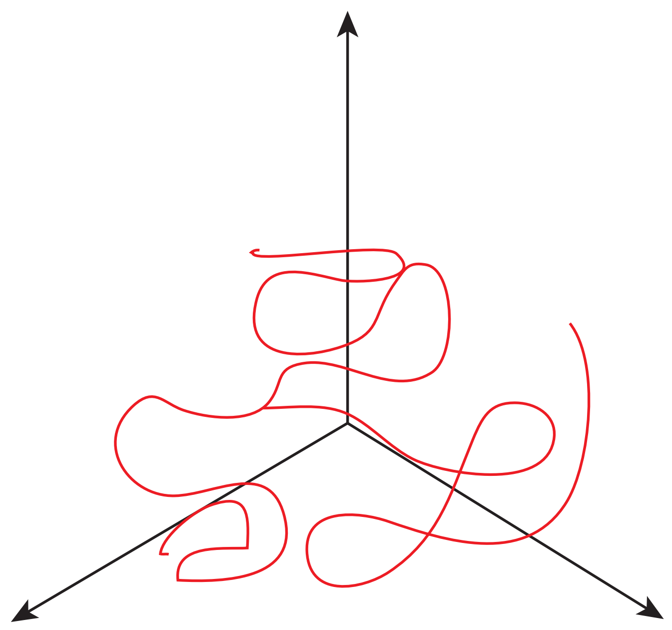
The origin of the problem

Axes of natural variation for natural **behavioral** events (e.g. deforming face moving in complex-lighted environment)

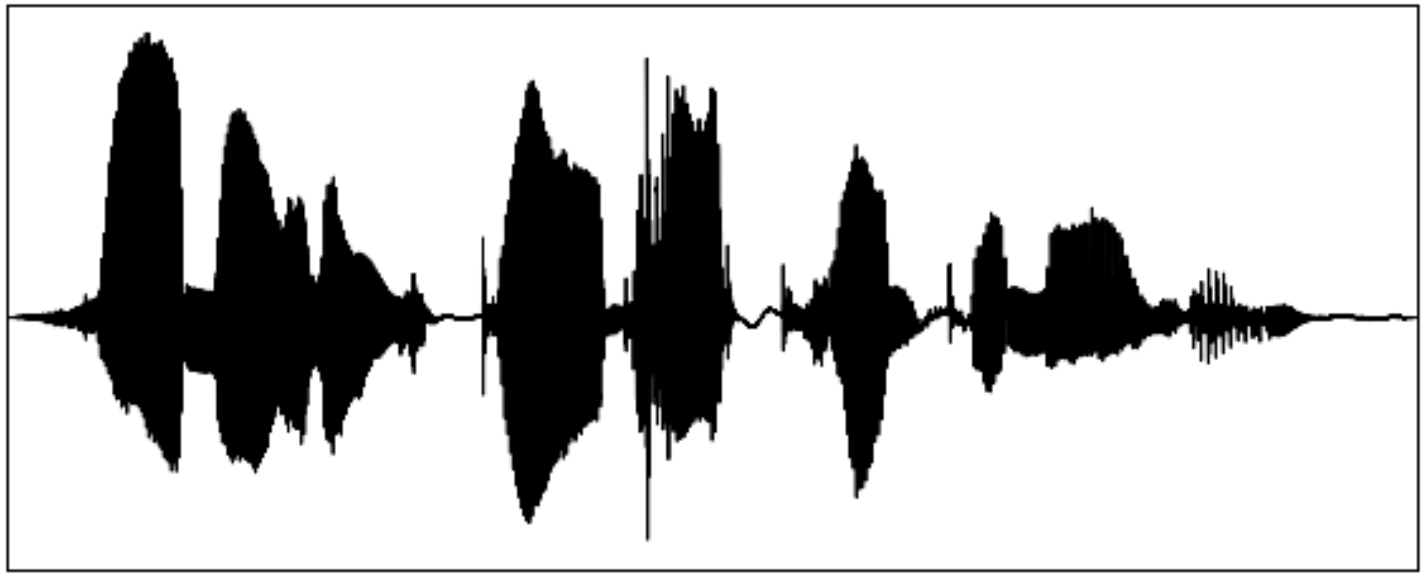


are misaligned with

Axes of natural variation of natural **"physics"** representation of world e.g. retinal photoreceptor voltage or hair-cell point amplitudes



The origin of the problem



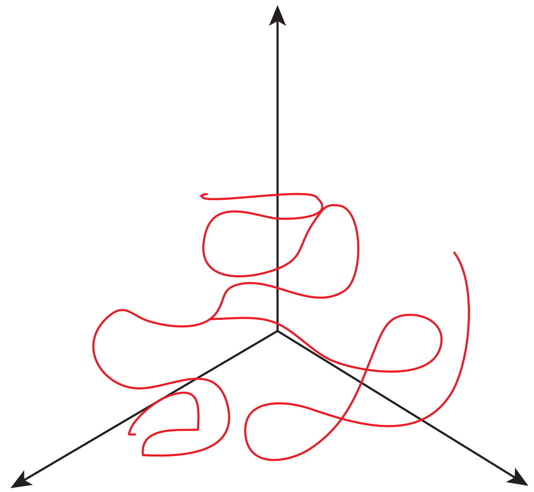
visual cortex

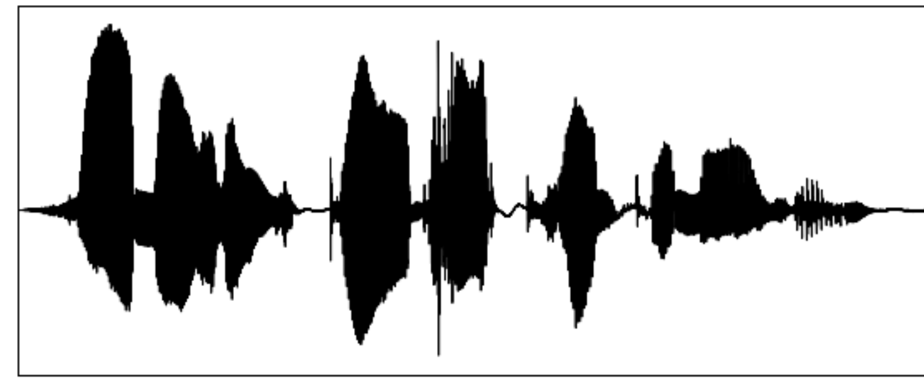
auditory cortex

"Mercedes behind Lamborghini, on a field in front of mountains."

"Hannah is good at compromising"

The origin of the problem

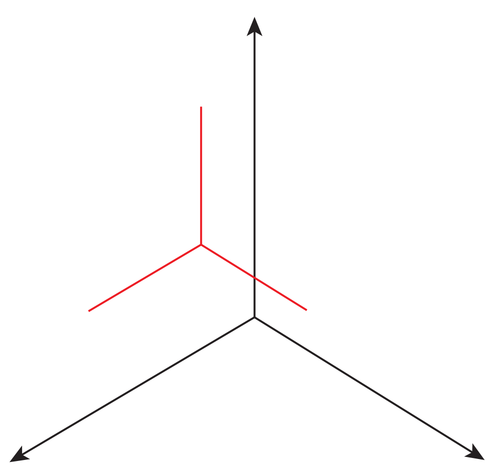




visual cortex

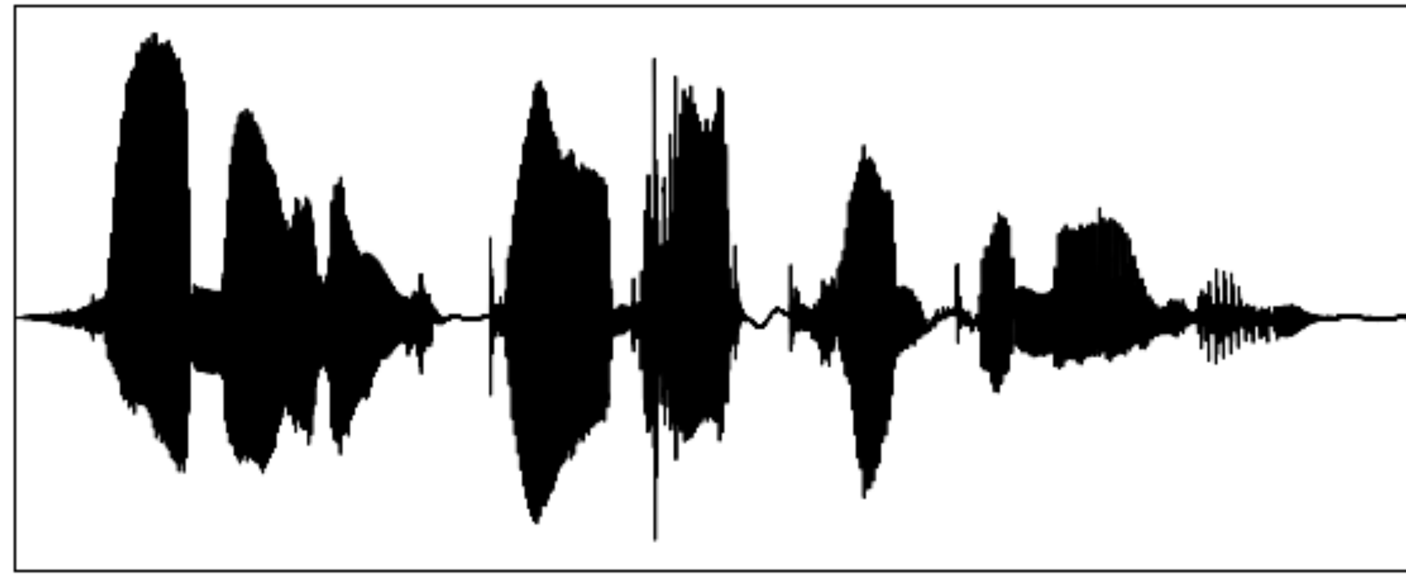
auditory cortex

"Mercedes behind Lamborghini, on a field in front of mountains."

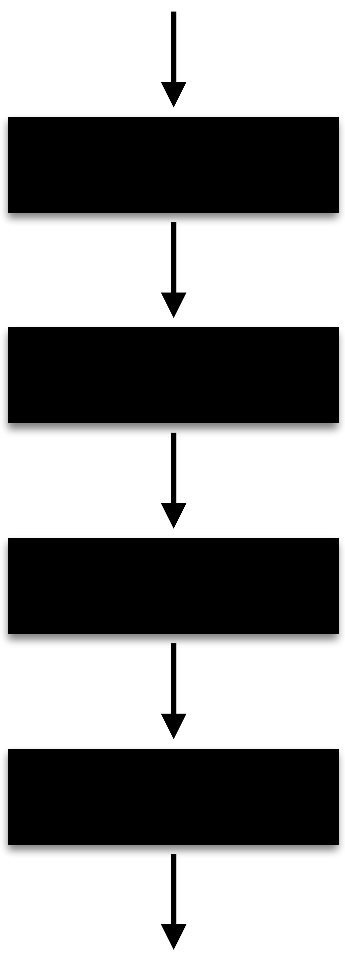


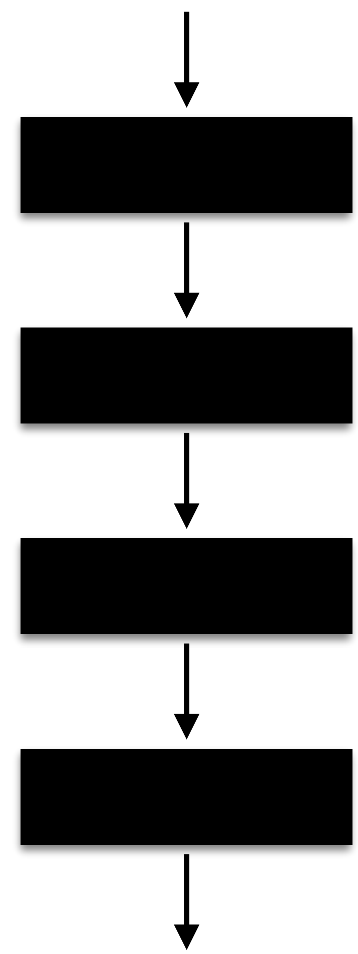
"Hannah is good at compromising"

Sensory cascade



visual cortex





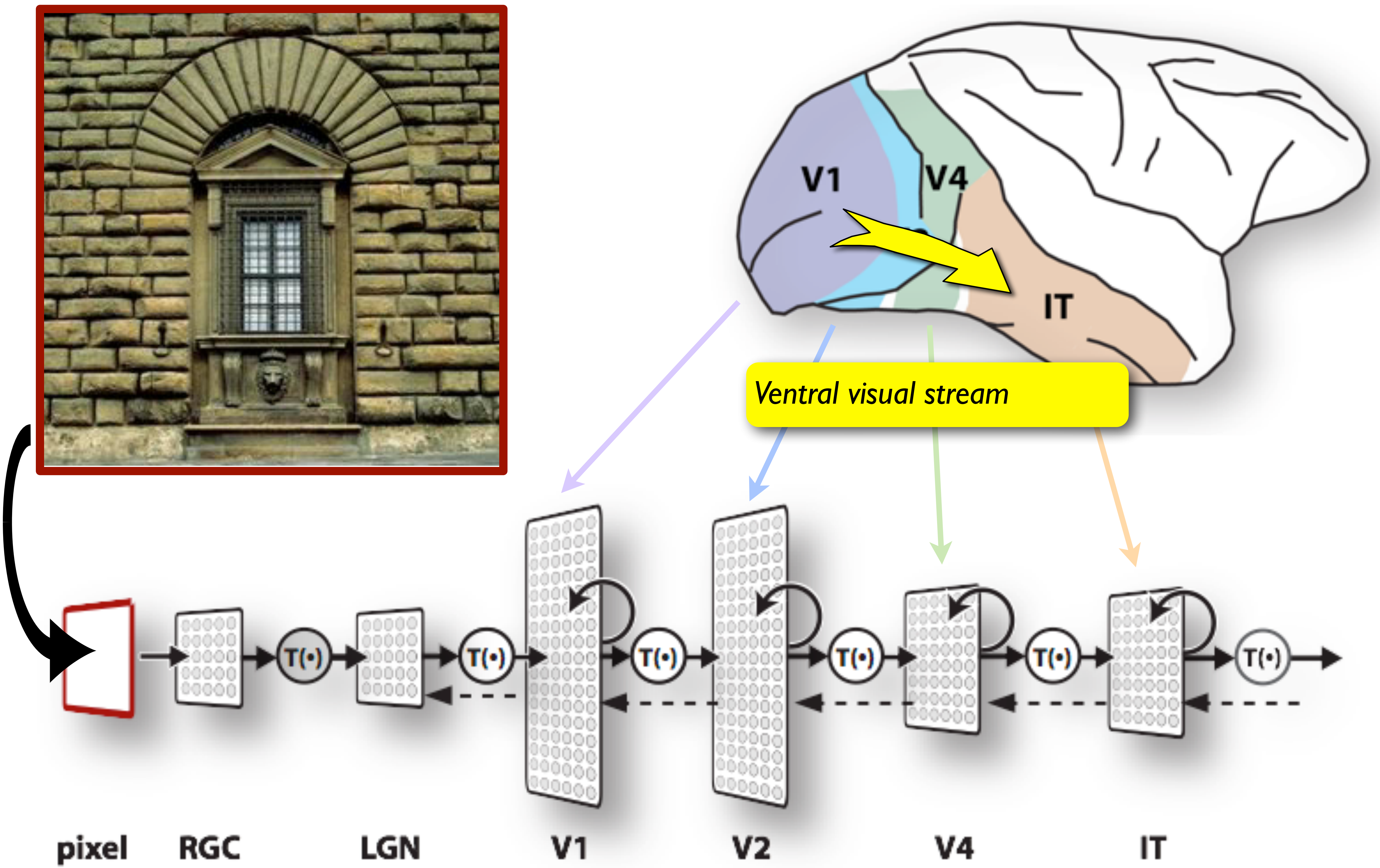
auditory cortex

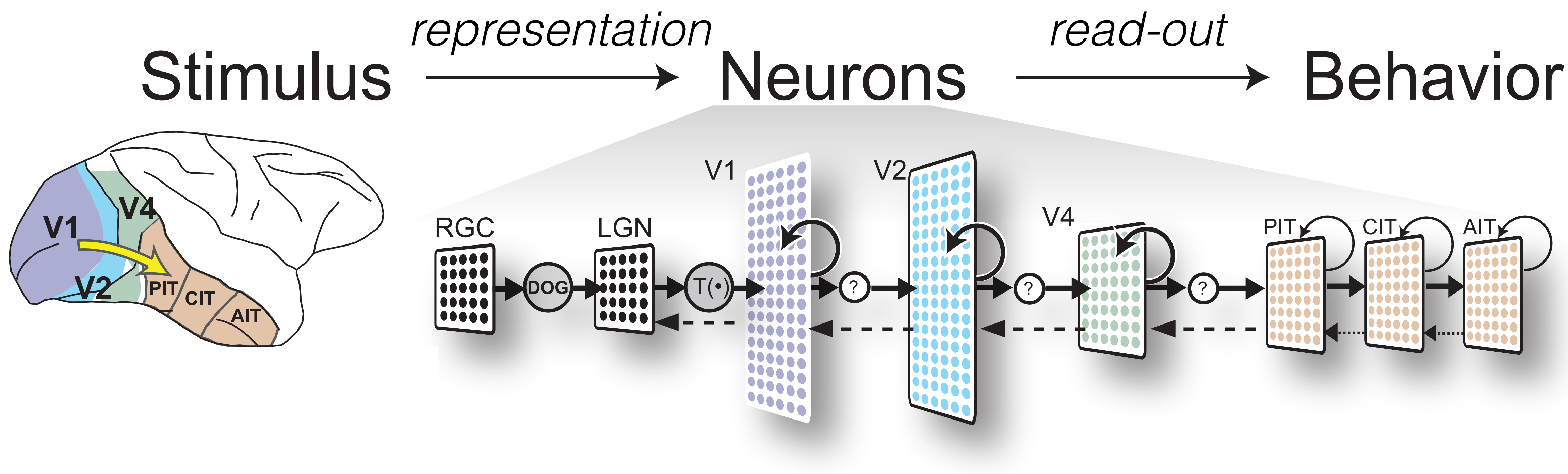
“Mercedes behind Lamborghini, on a field in front of mountains.”

“Hannah is good at compromising”

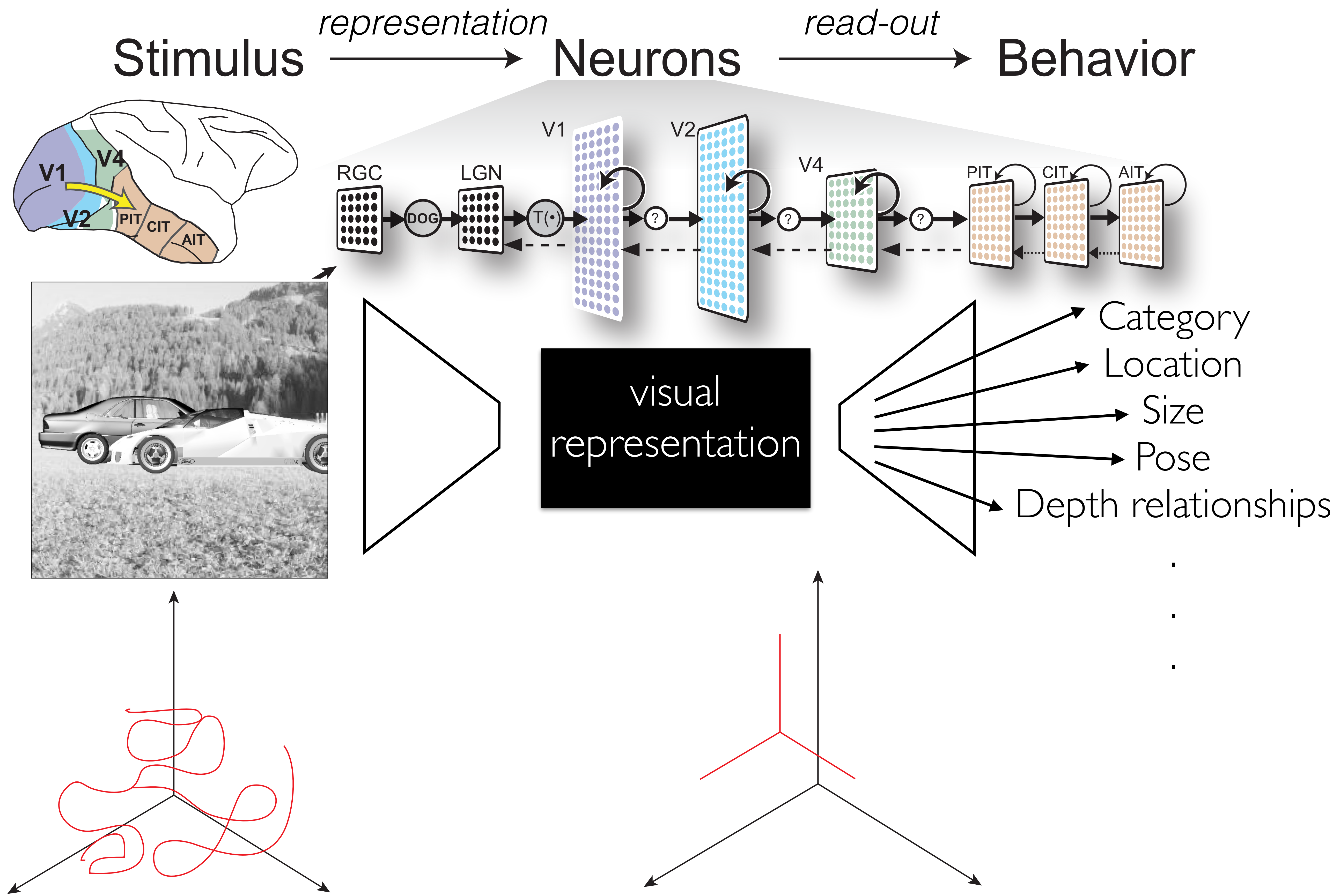
Ventral Stream = Connected series of brain areas

Kaas (2003), Van Essen (2003), Valois and Morgan (1974) Gross (1973), Mishkin and Ungerleider (1983), Holmes and Gross (1984) Horel *et al.* (1987); Freiwald and Tsao (2010), Pitcher, *et al.* (2009) Yaginuma (1982), Holmes (1984), Weiskrantz (1984), Schiller (1995) Afraz (2006), Verhoef (2012) Rust (2010), Freiwald (2010), Lehky (2007) Majaj (2015)





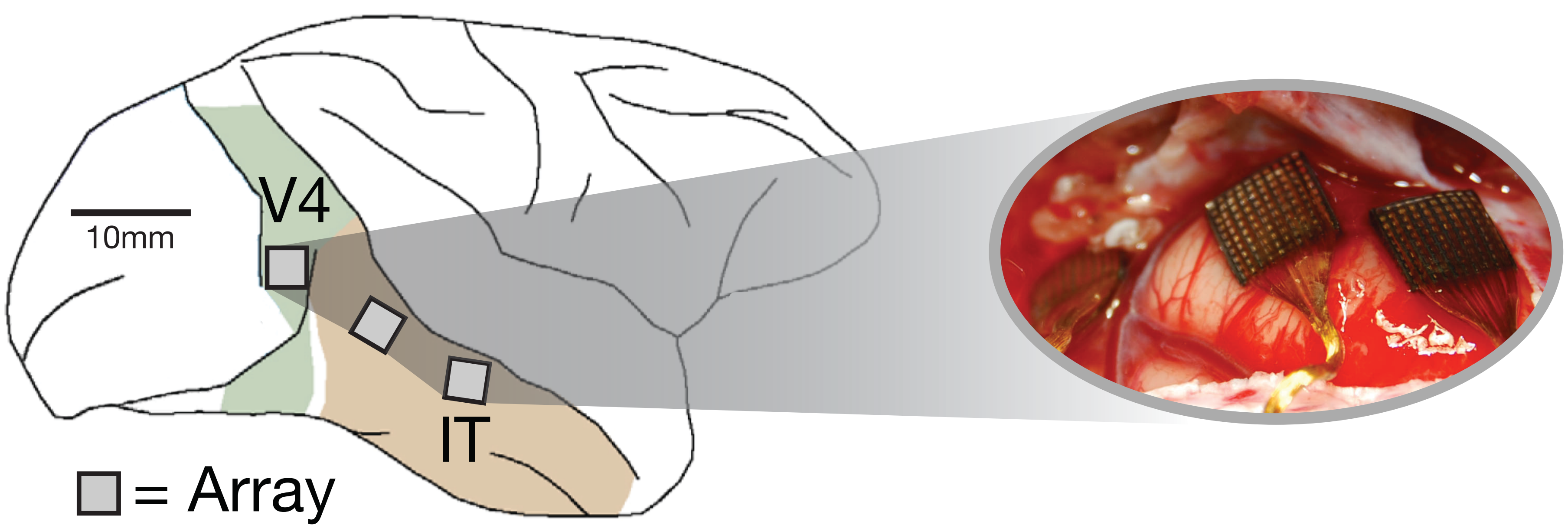
Stimulus
representation
Neurons
read-out
Behavior
V1
V2
V4
PIT
CIT
AIT
RGC
DOG
LGN
T(•)
V1
V2
V4
?
PIT
CIT
AIT



Stimulus
representation
Neurons
read-out
Behavior
V1
V2
V4
PIT
CIT
AIT
RGC
DOG
LGN
T(•)
?
visual representation
Category
Location
Size
Pose
Depth relationships

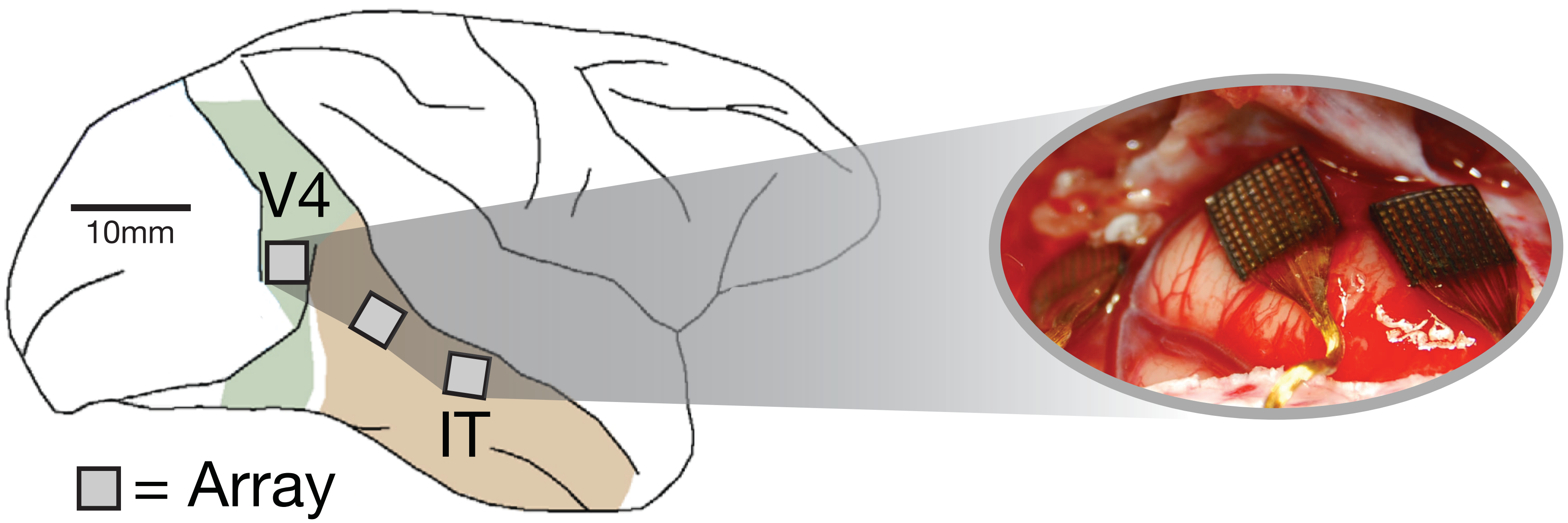
Multi-array Electrophysiology Experiment

Multi-array electrophysiology in macaque V4 and IT.

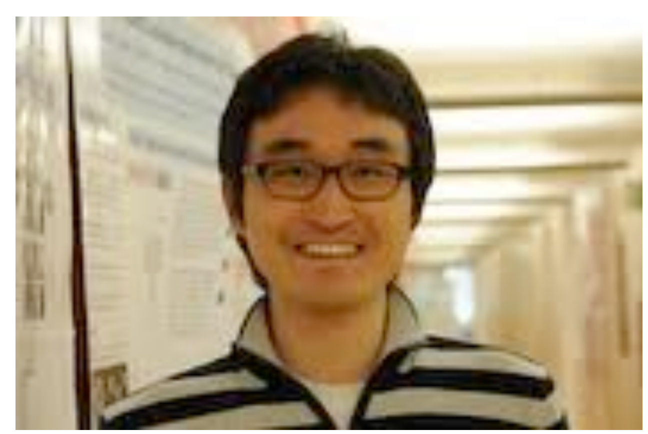


Multi-array Electrophysiology Experiment

Multi-array electrophysiology in macaque V4 and IT.



About 300 total sites



Ha Hong

Jim DiCarlo

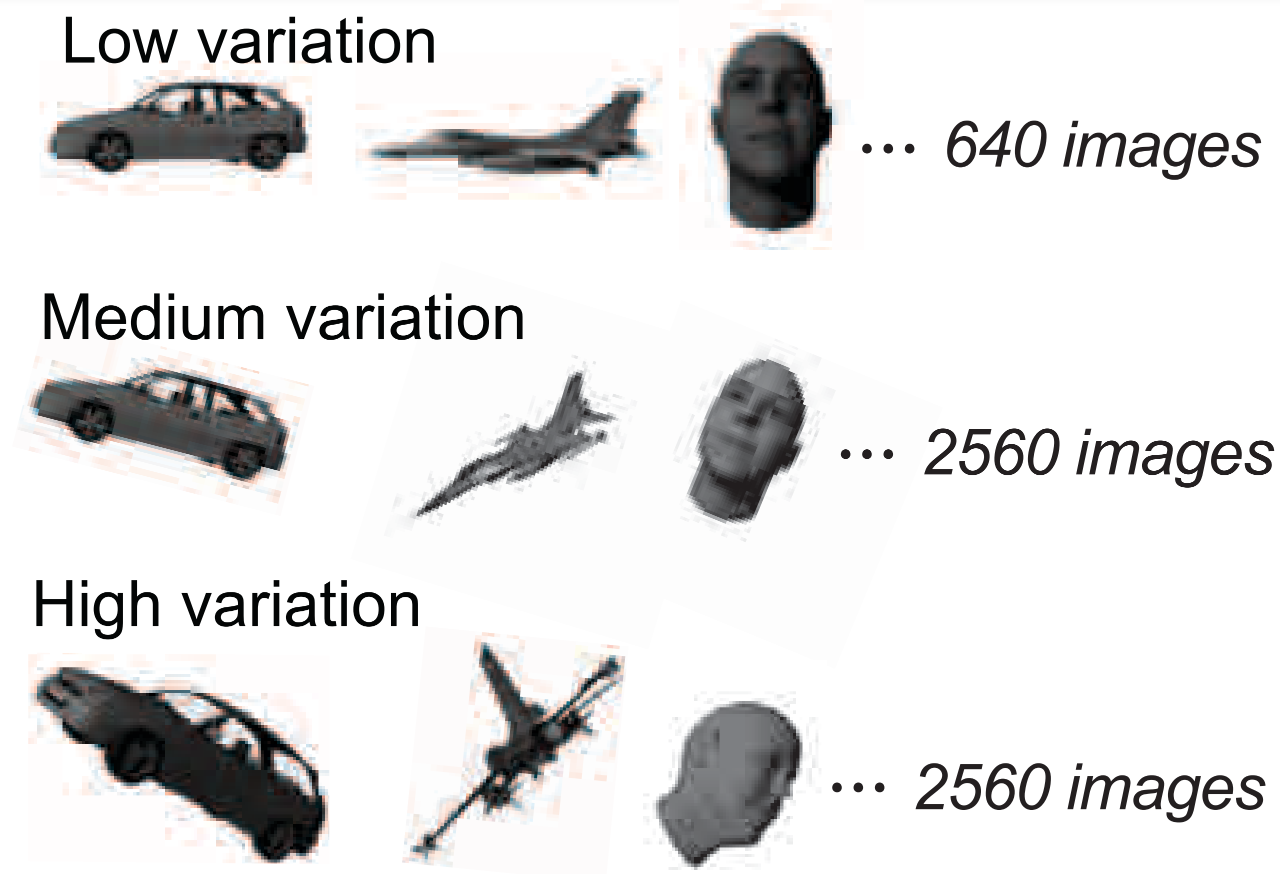
Multi-array Electrophysiology Experiment

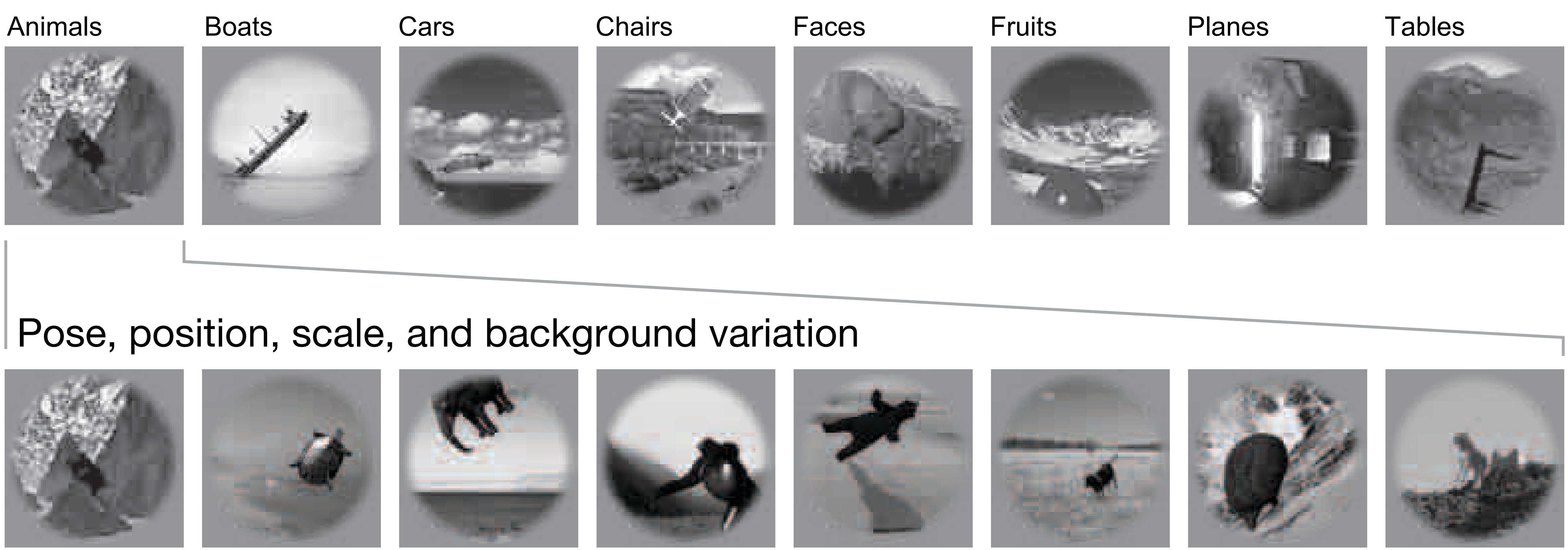
5760 images

64 objects

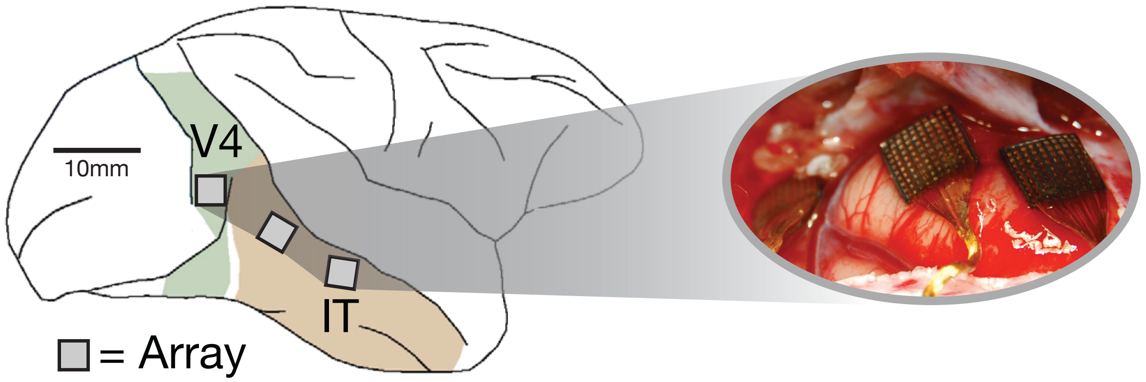
8 categories

uncorrelated photo backgrounds



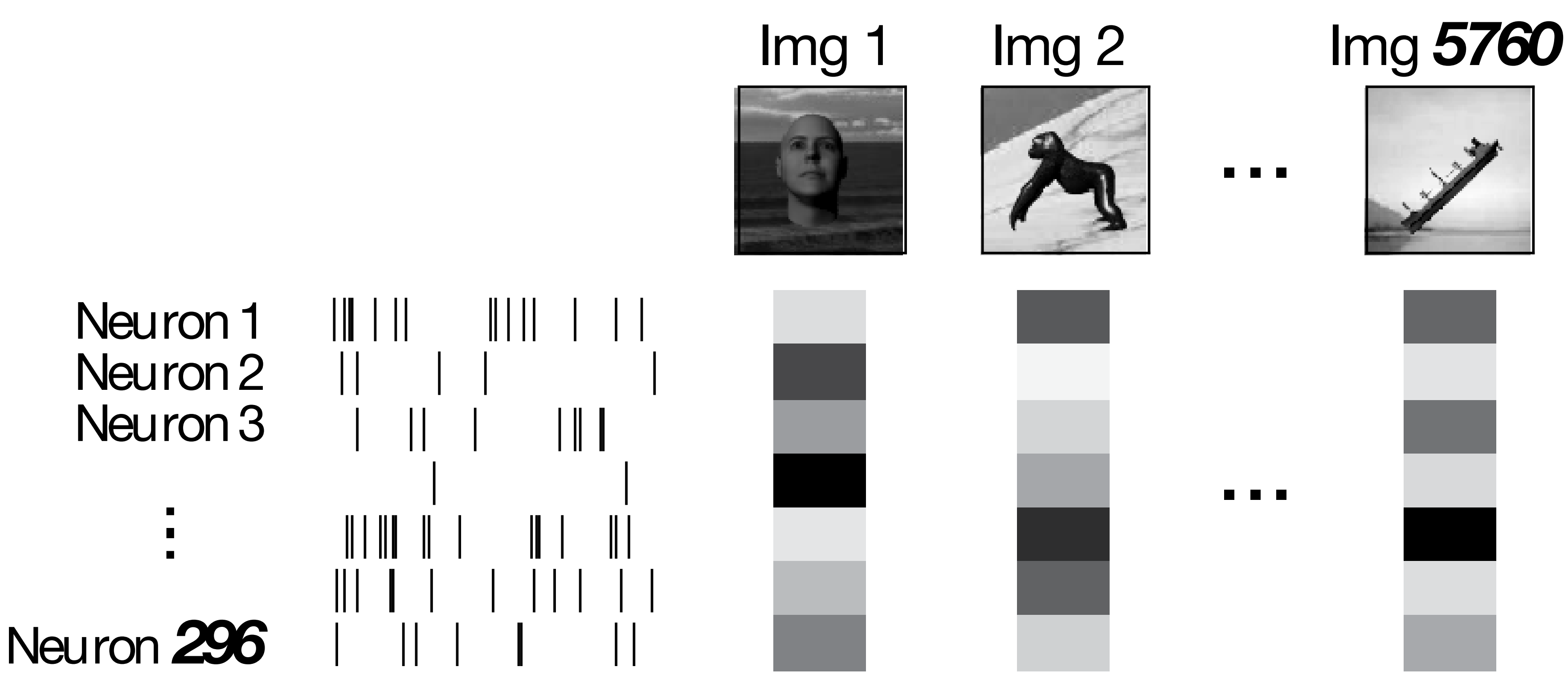


Multi-array Electrophysiology Experiment

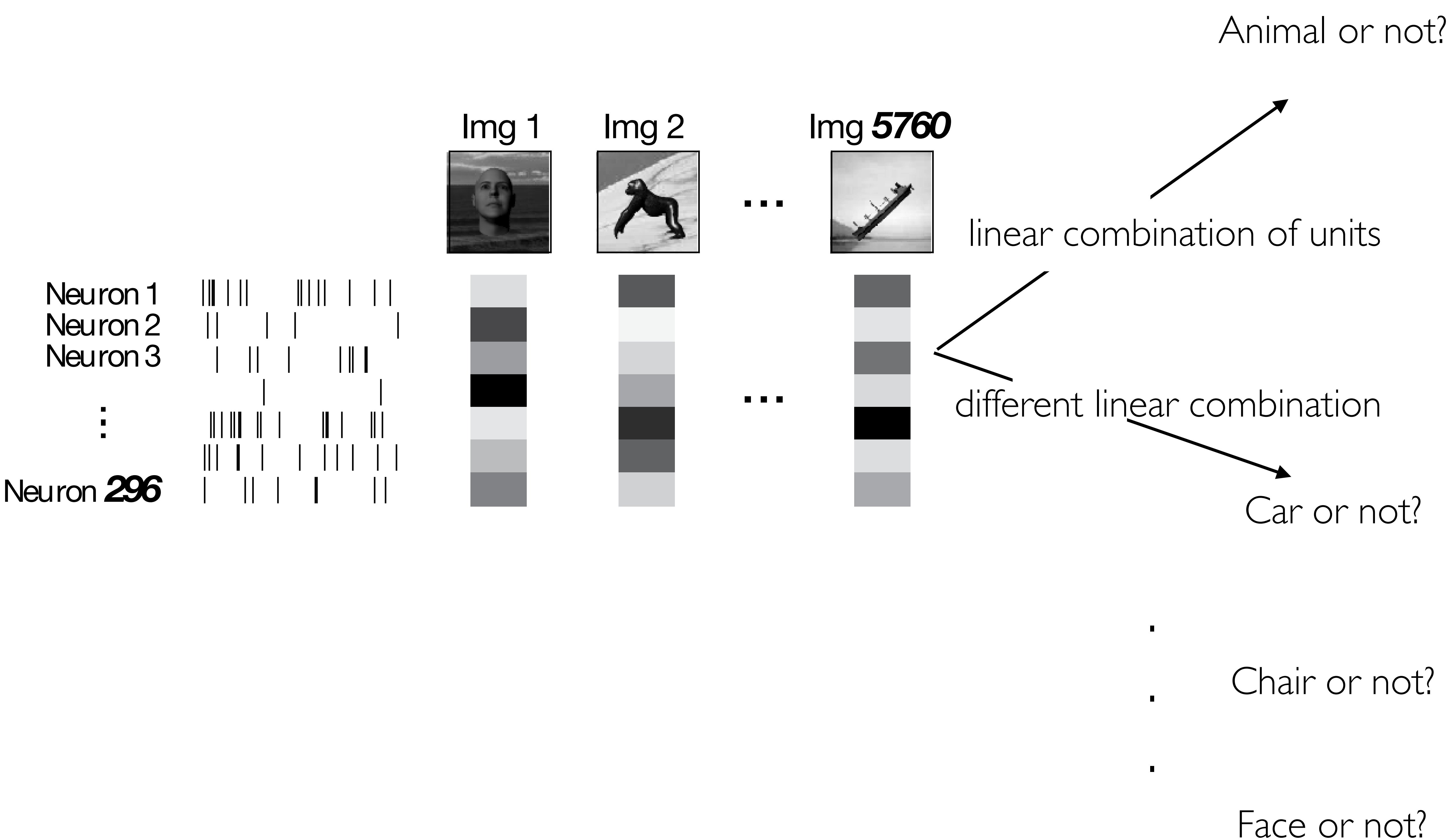


About 300 total sites

Output = Binned spike counts 70ms-170ms post stimulus presentation averaged over 25-50 reps of each image.

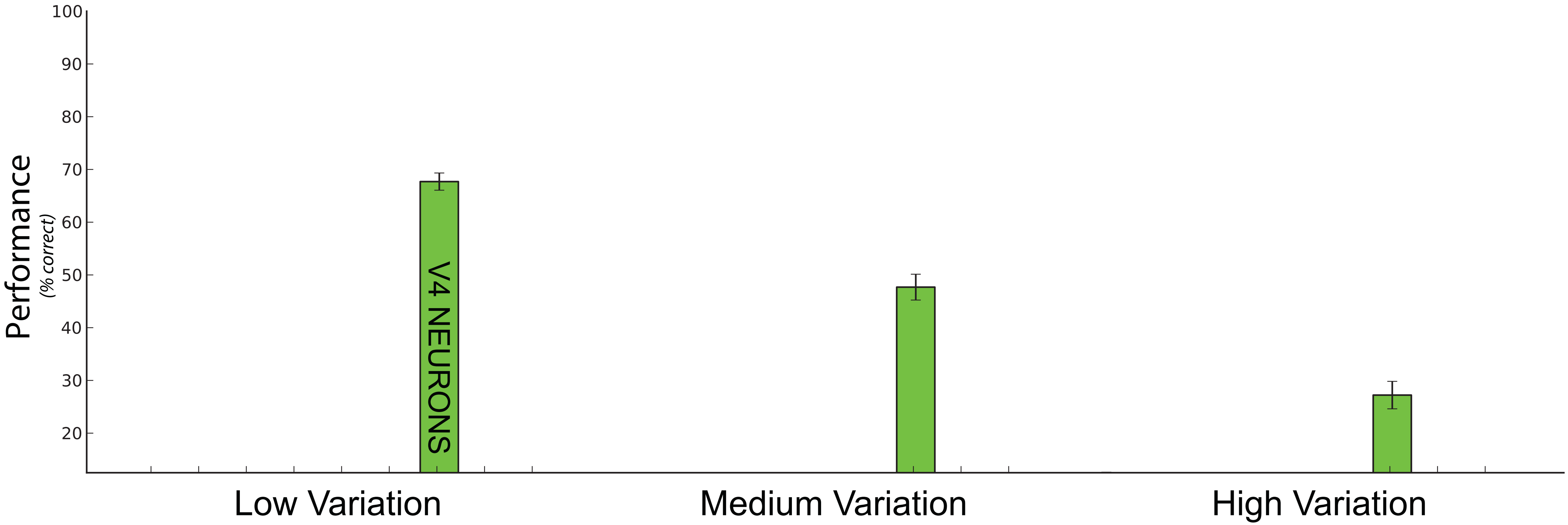


Neural-Behavior Decoding



Decoding Behaviorally Output from Neural Populations

V4 loses out at higher variation:

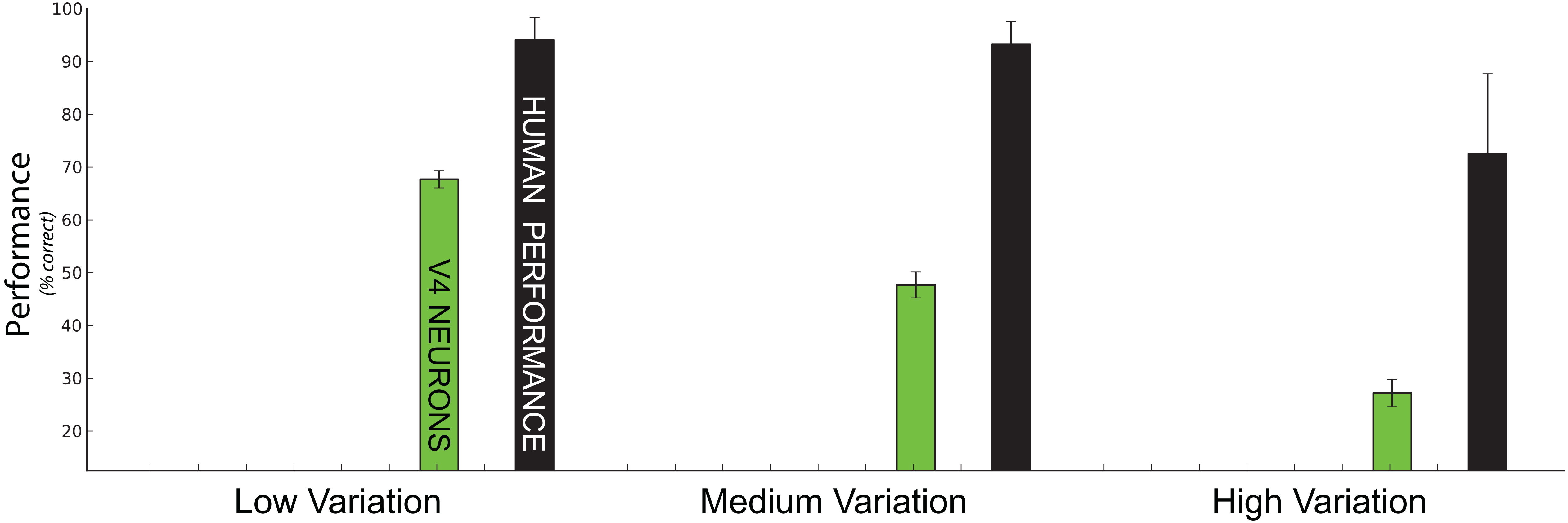


Yamins and Hong* et. al.* ***PNAS*** *(2014)*

Decoding Behaviorally Output from Neural Populations

V4 loses out at higher variation:

… but humans are much less affected.



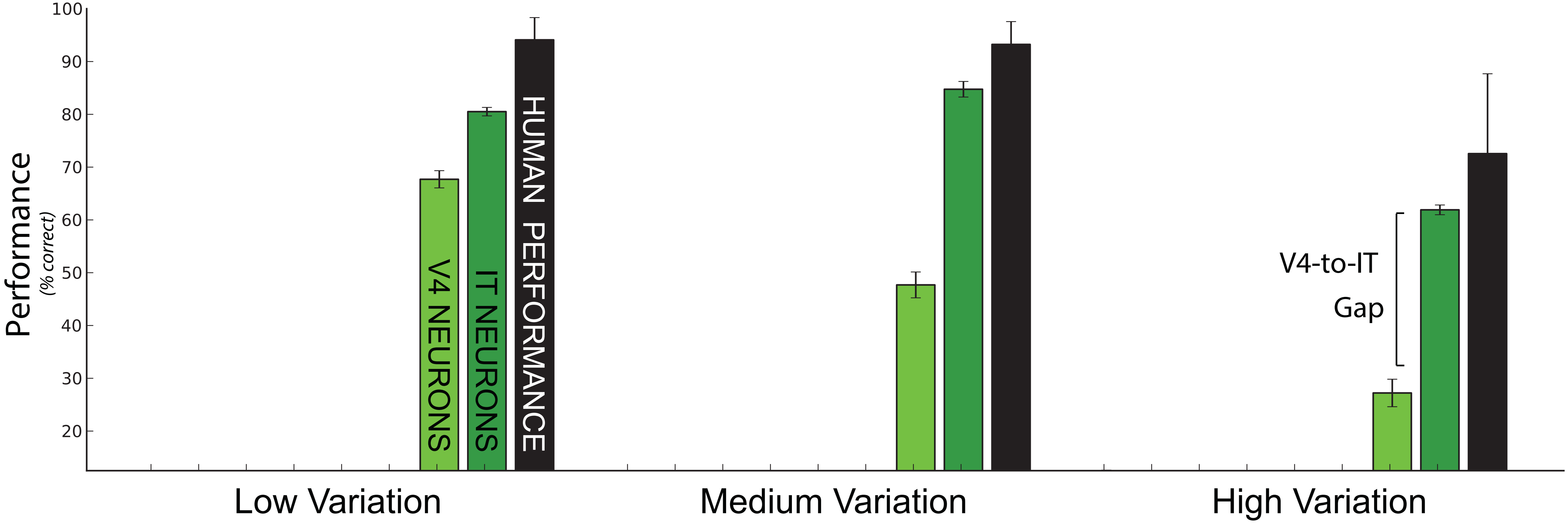
Yamins and Hong* et. al.* ***PNAS*** *(2014)*

IT Neurons Track Human Performance

V4 loses out at higher variation:

… but humans are much less affected.

… as is the IT neural population.



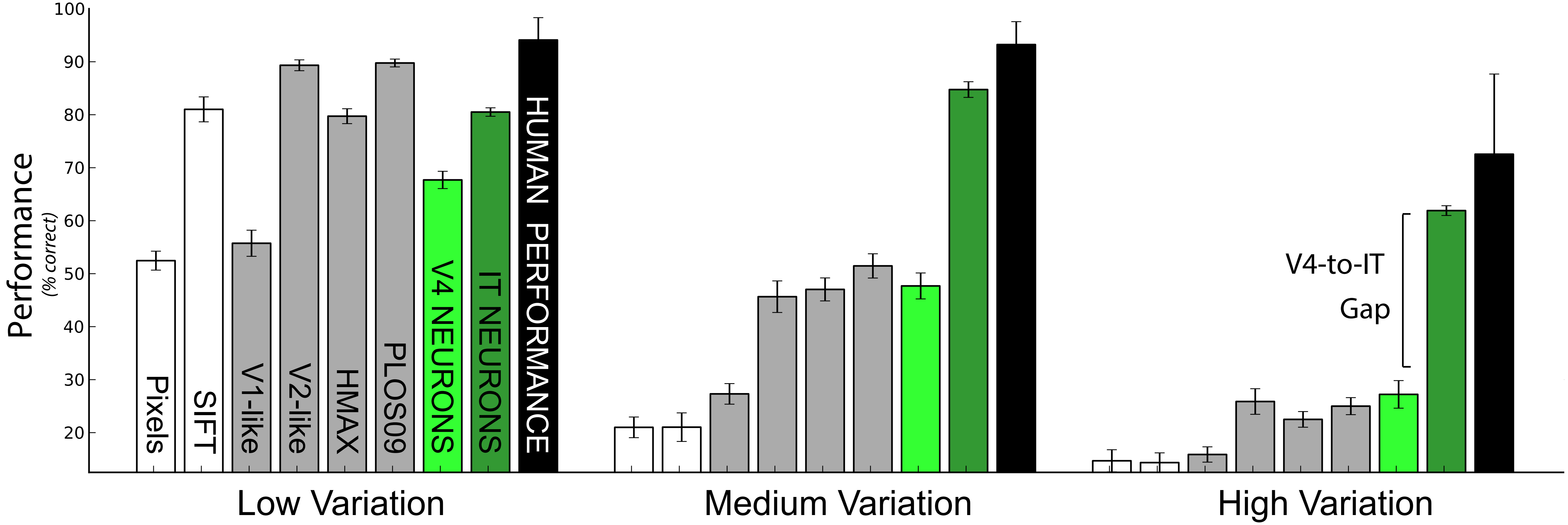
Yamins and Hong* et. al.* ***PNAS*** *(2014)*

IT Neurons Track Human Performance

V4 loses out at higher variation:

… but humans are much less affected.

… as is the IT neural population.



Yamins and Hong* et. al.* ***PNAS*** *(2014)*

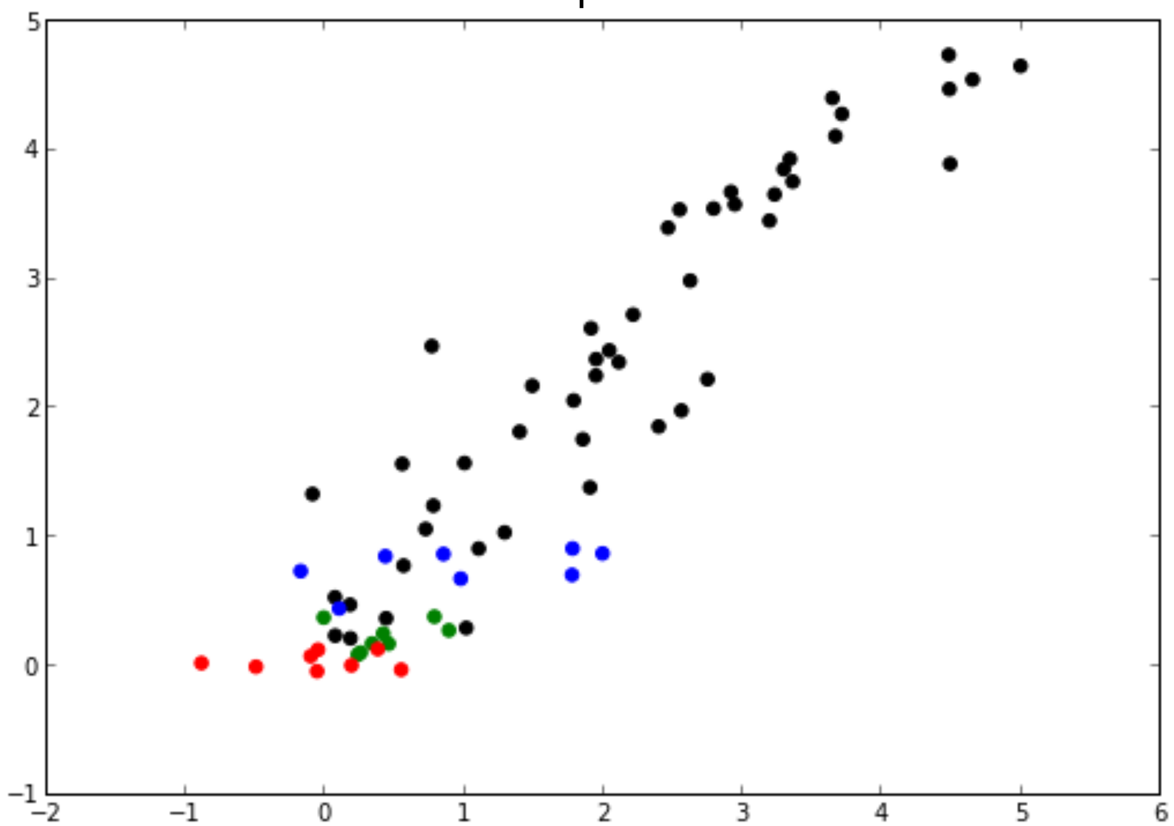
At <u>high variation levels</u>, IT much better than V4 and existing models.

IT Neurons Track Human Performance

IT matches human error patterns as well as raw performance.

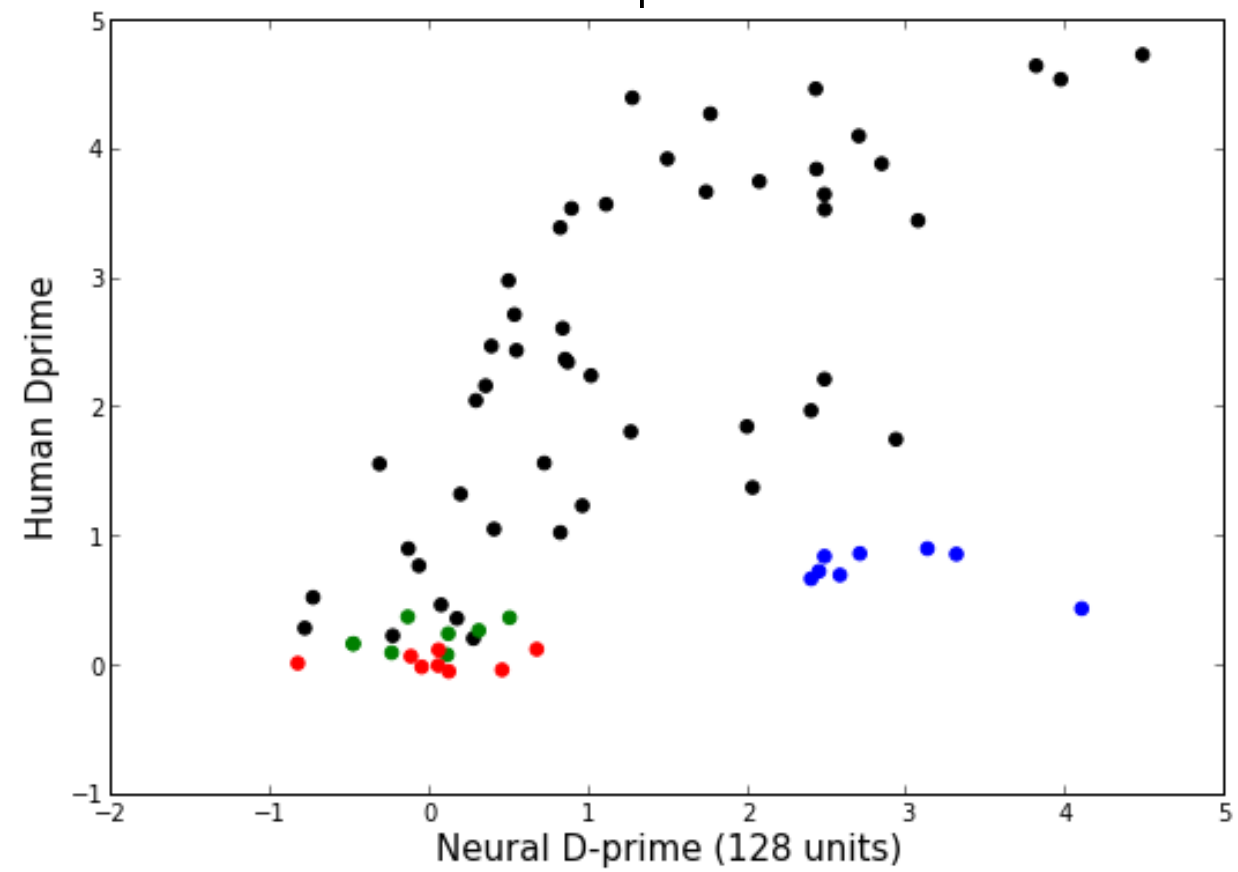
IT Population

Human Performance



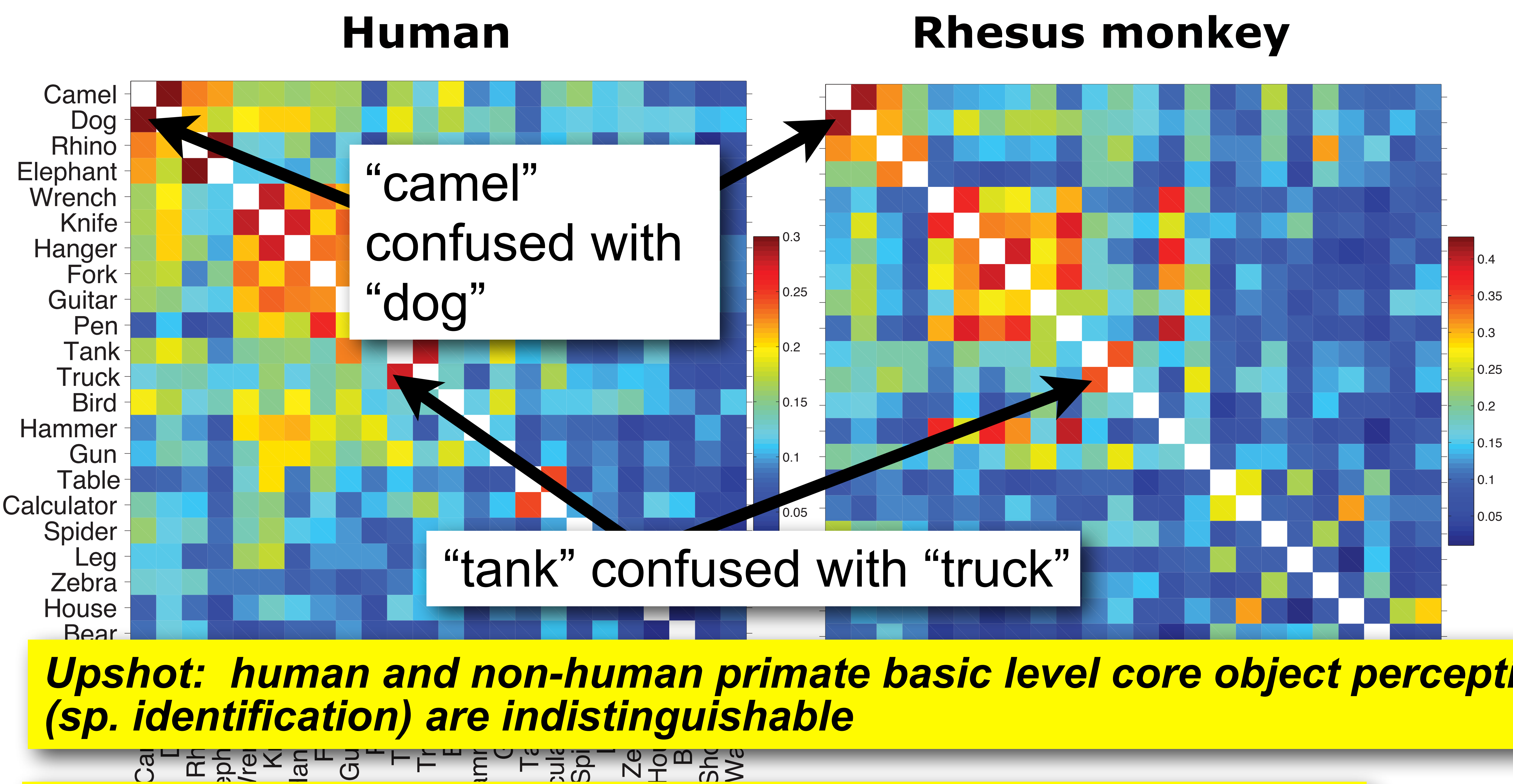
Neural Decode Performance

V4 Population



● Low-Variation Face subordinate tasks.

Human / Monkey similarities



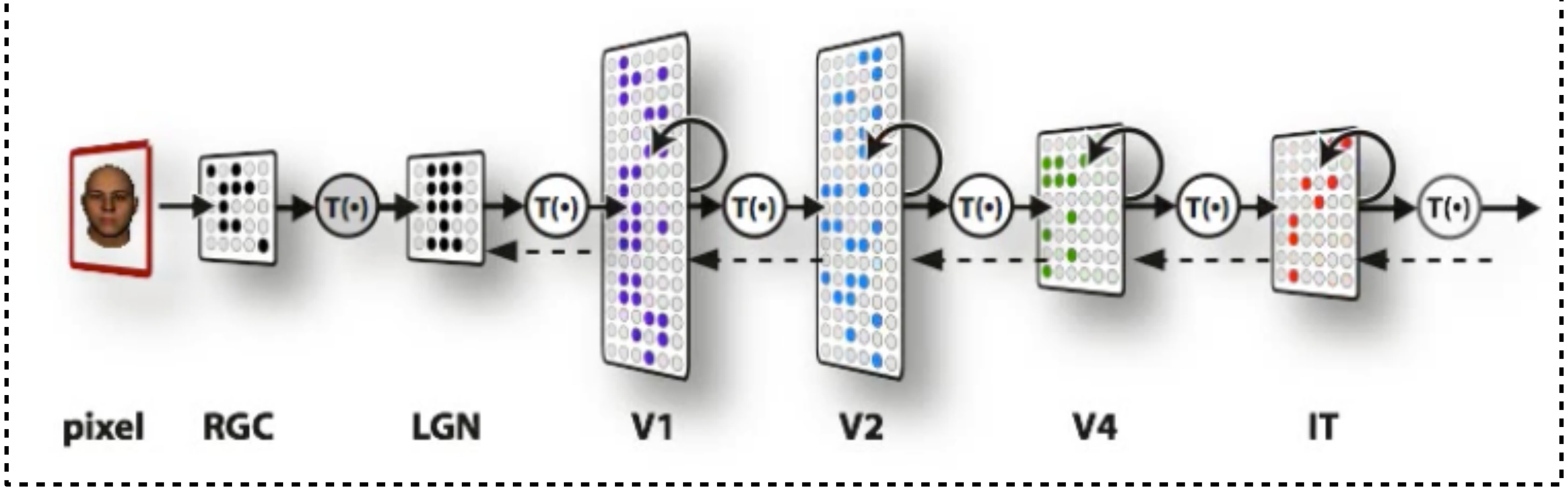
Upshot: human and non-human primate basic level core object percept
(sp. identification) are indistinguishable

Does not depend on reporting effector (touch vs. eye movement)

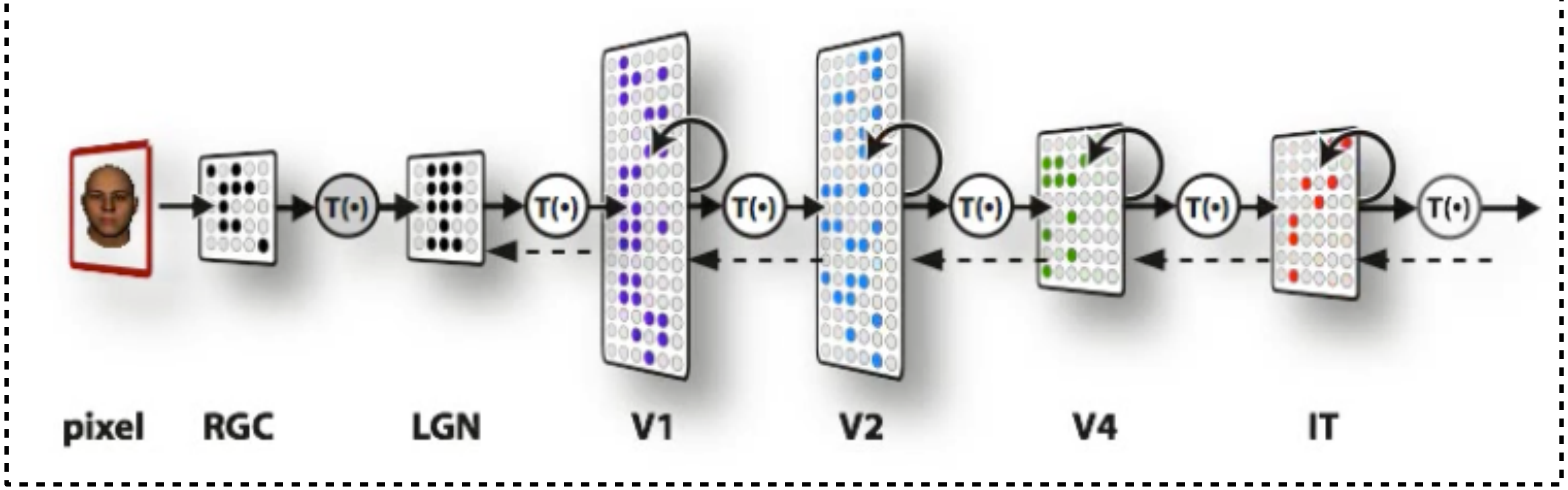
Comparison of Object Recognition Behavior in Human and Monkey
*R. Rajalingham, K Schmidt, J.J. DiCarlo, **Vision Sciences Society** (2014)*
*R. Rajalingham, K Schmidt, J.J. DiCarlo, **J. Neuroscience** (2015)*

Adapted from Motter and Mountcastle 1981

GOAL: Predictive model of single-neuron responses throughout the ventral stream to arbitrary image stimuli.

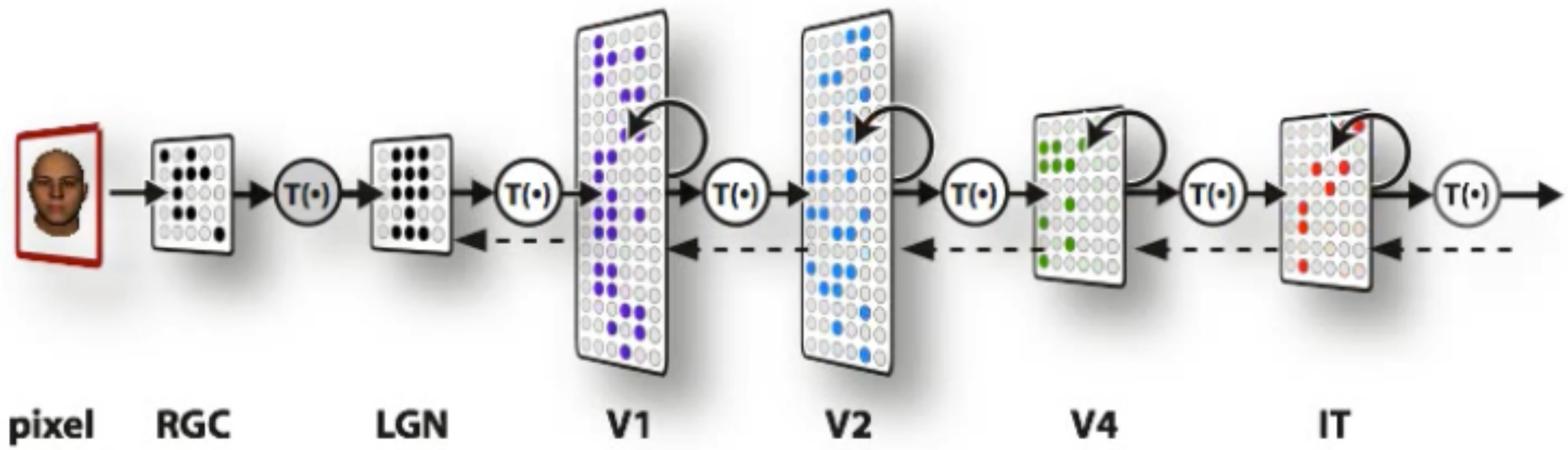


GOAL: Predictive model of single-neuron responses throughout the ventral stream to arbitrary image stimuli.



1. image-computable 2. Predictive 3. Mappable

GOAL: Predictive model of single-neuron responses throughout the ventral stream to arbitrary image stimuli.



1. image-computable

2. Predictive

3. Mappable

→ Convolutional Neural Networks (CNNs)

Fukushima, 1980; Lecun, 1995

Goal: Predictive Model of Ventral Stream

Kunihiko Fukushima!

Tokyo, November 2015

Goal: Predictive Model of Ventral Stream

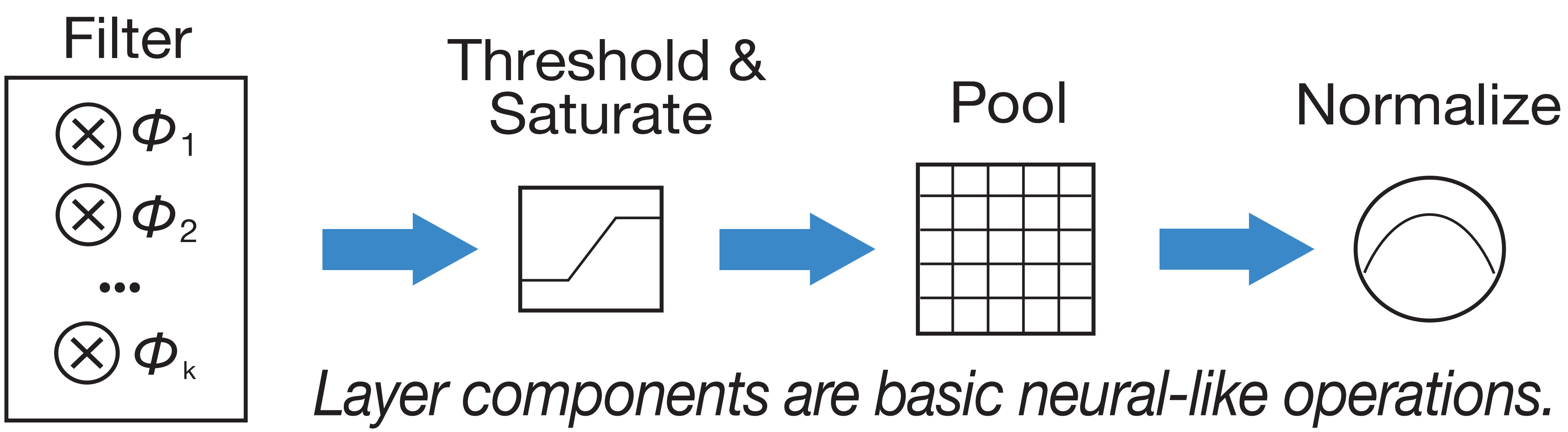
Tokyo, November 2015

Kunihiko Fukushima!

Developed neocognitron while Japan Broadcasting Corporation (NHK) … office directly next door to Keisuke Toyama and Keiji Tanaka

Hierarchical Convolutional Neural Networks

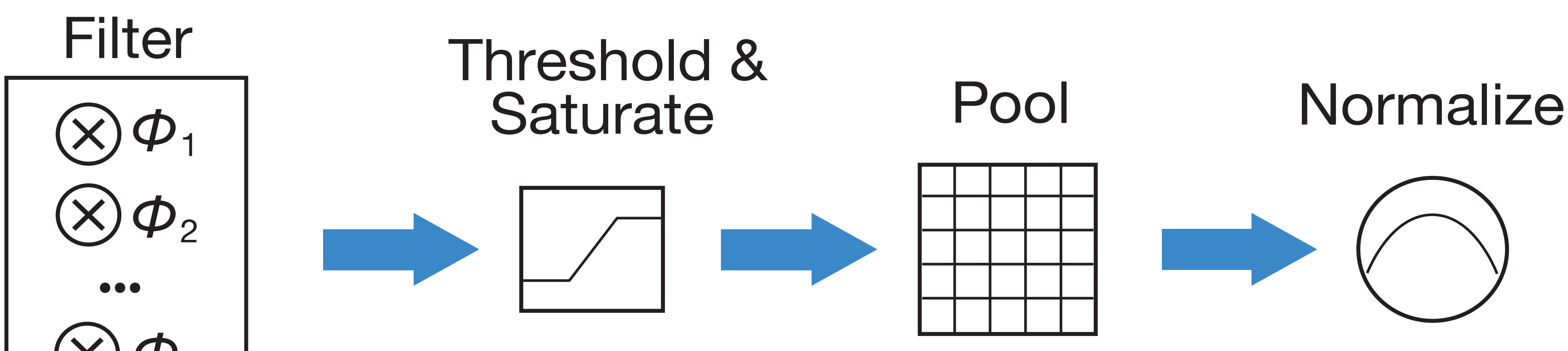
- Individual layers of neurally-plausible **basic operations**



Layer components are basic neural-like operations.

Hierarchical Convolutional Neural Networks

- Individual layers of neurally-plausible **basic operations**



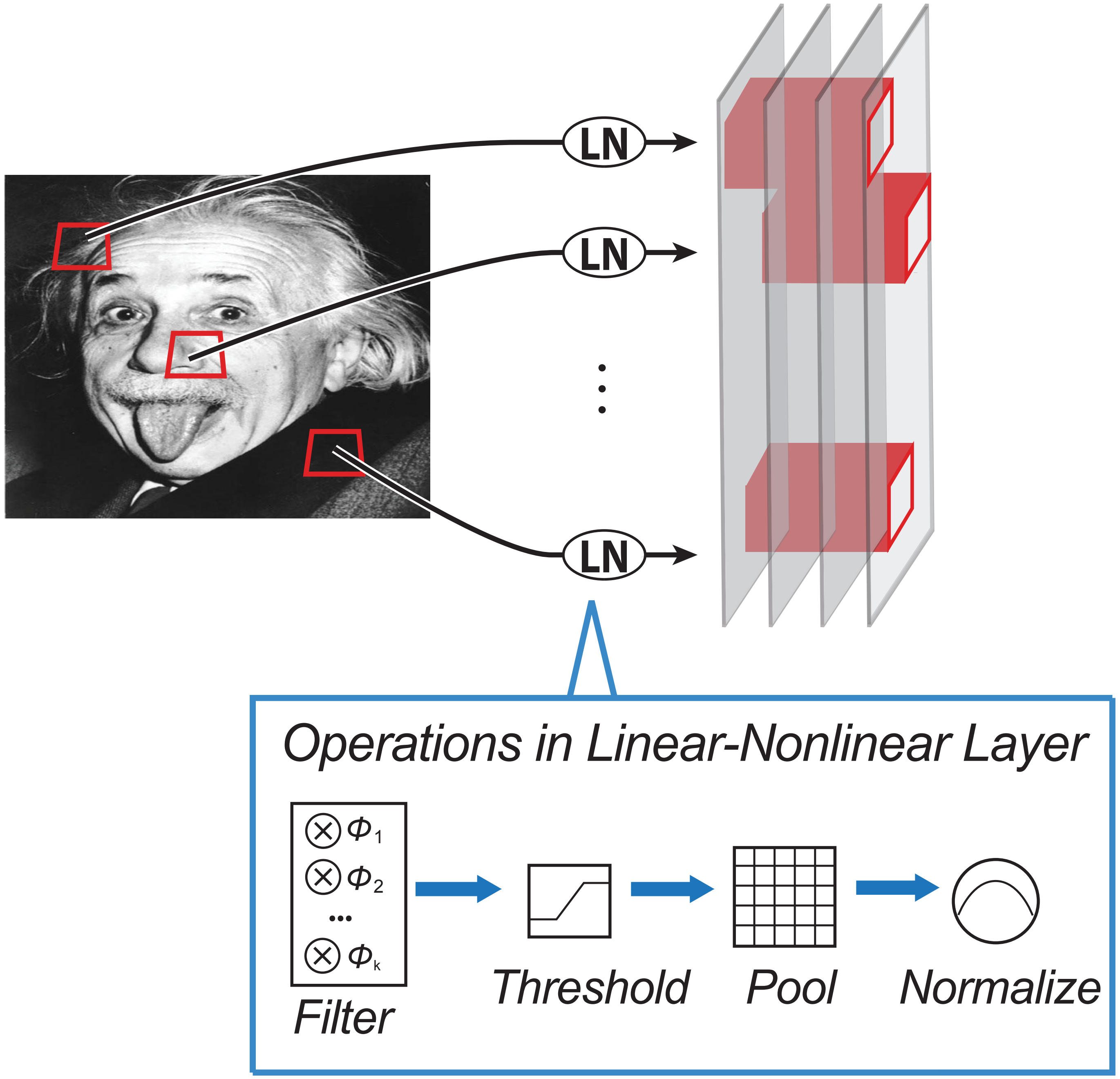
Layer components are basic neural-like operations.

	Filter	Threshold & Saturate	Pool	Normalize
neuro:	synaptic weights patterns	single-unit activations	complex cells	competitive inhibition
data:	untangling through dimension expansion	"AND" operation by limiting dynamic range	adding robustness by dimension reduction	put results back into standard range

Hubel and Wiesel (1965-1975) Lecun (2004), Carandini et. al (2005), Lennie & Movshon (2005) , DiCarlo (2012)

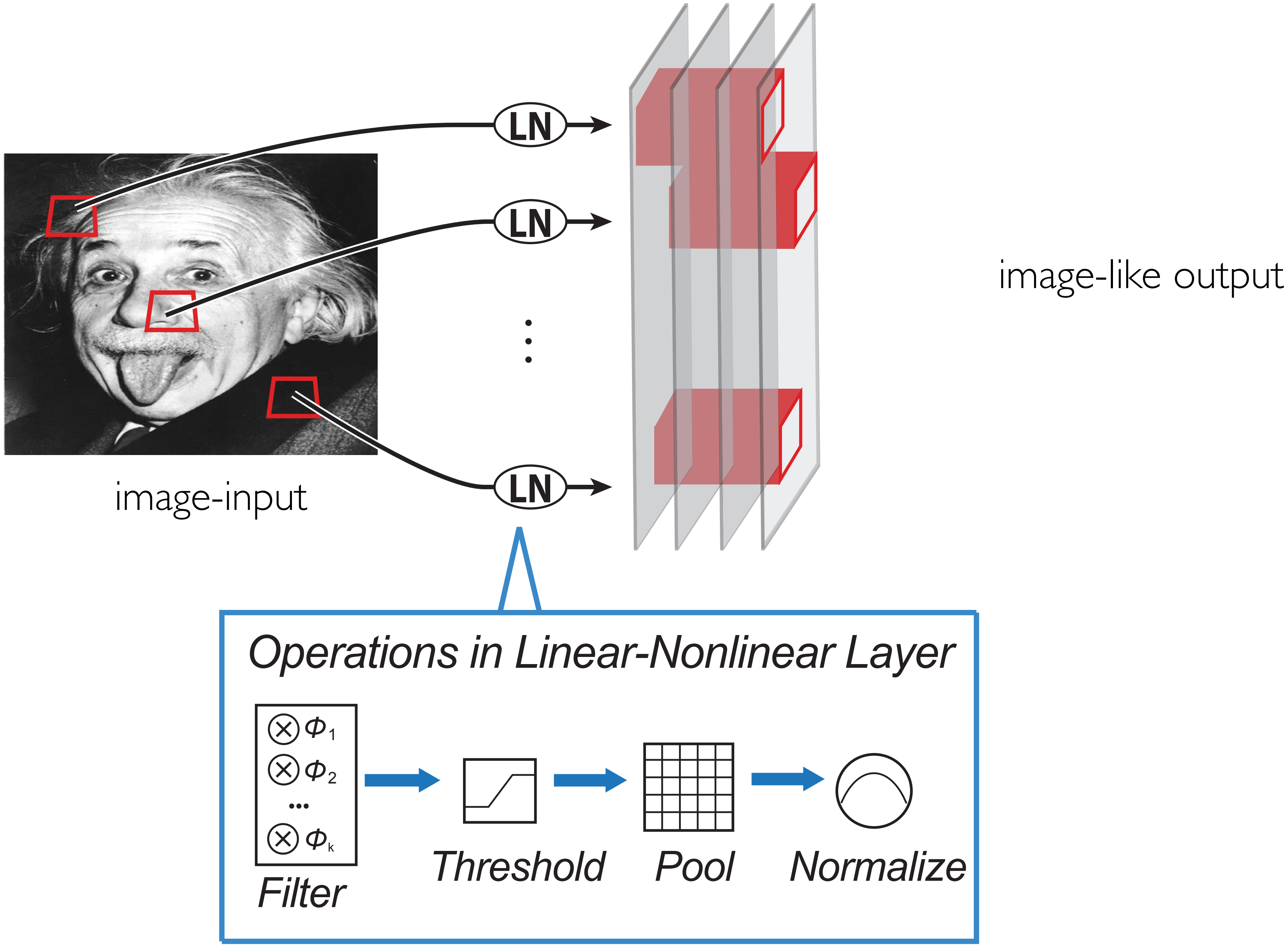
Hierarchical Convolutional Neural Networks

- Individual layers of neurally-plausible **basic operations**
- Applied **convolutionally** — same at all locations: approx. retinopy



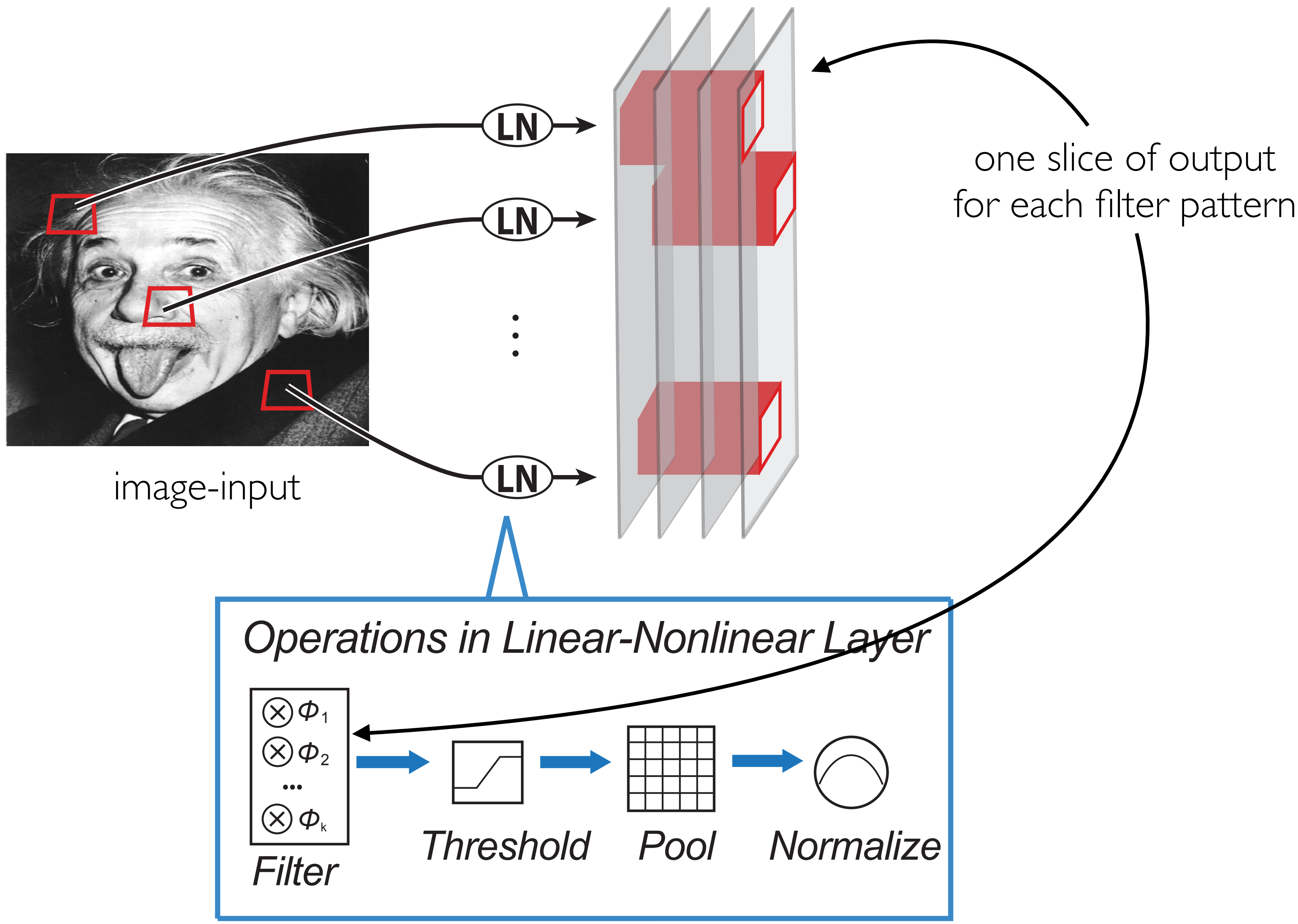
Hierarchical Convolutional Neural Networks

- Individual layers of neurally-plausible **basic operations**
- Applied **convolutionally** — same at all locations: approx. retinopy



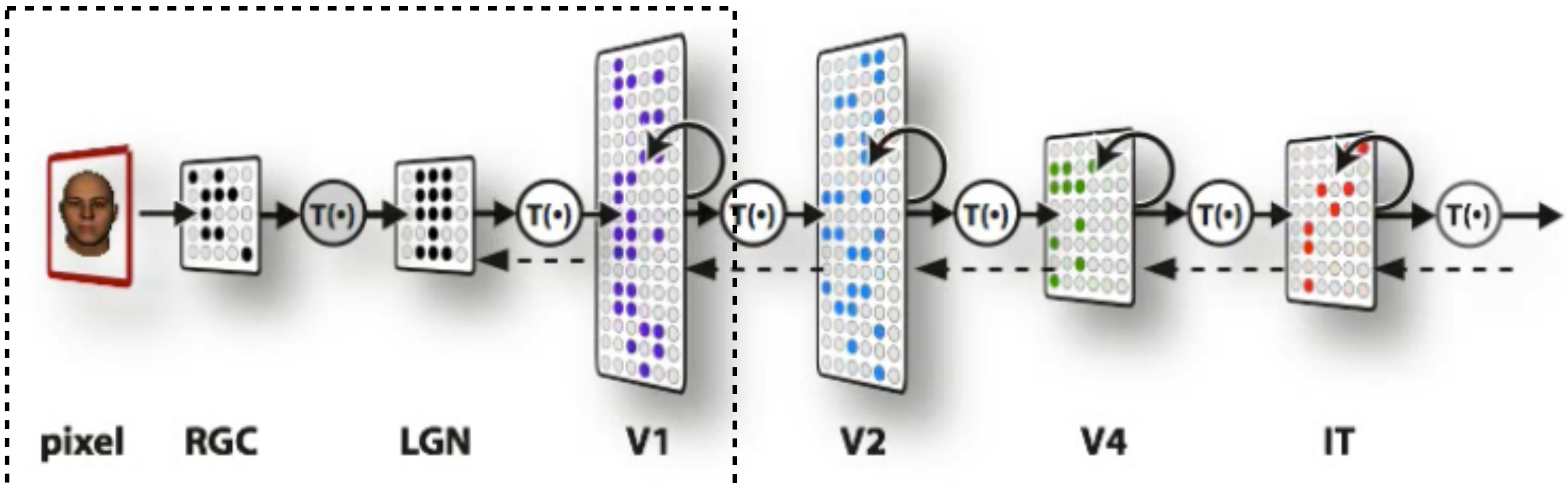
Hierarchical Convolutional Neural Networks

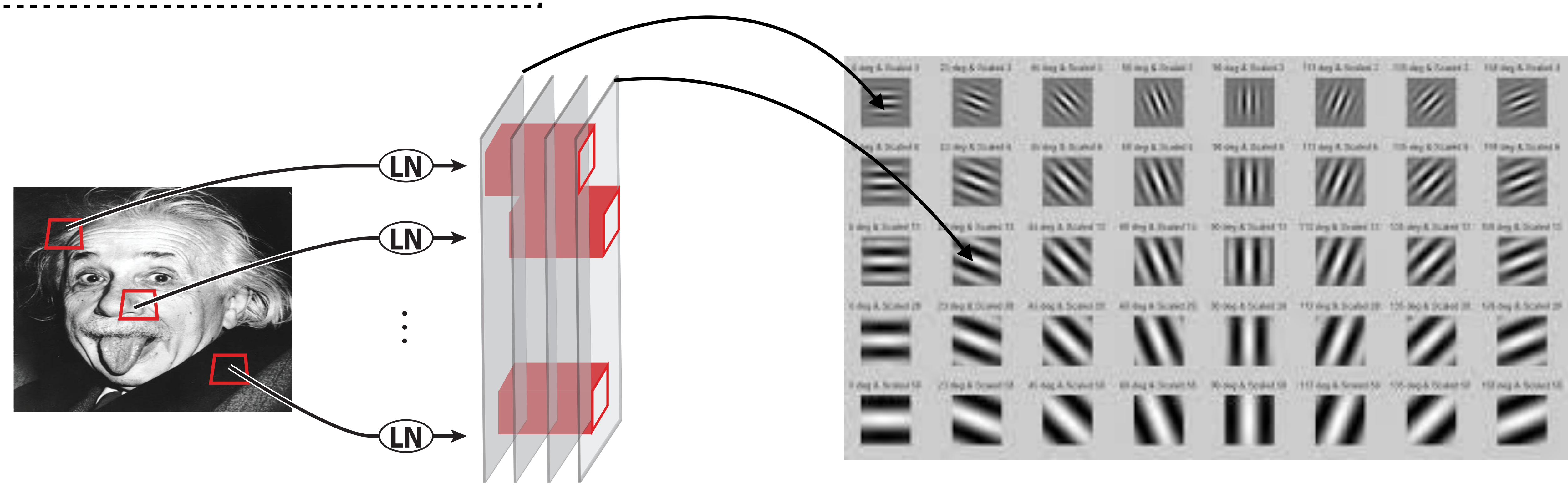
- Individual layers of neurally-plausible **basic operations**
- Applied **convolutionally** — same at all locations: approx. retinopy



Hierarchical Convolutional Neural Networks

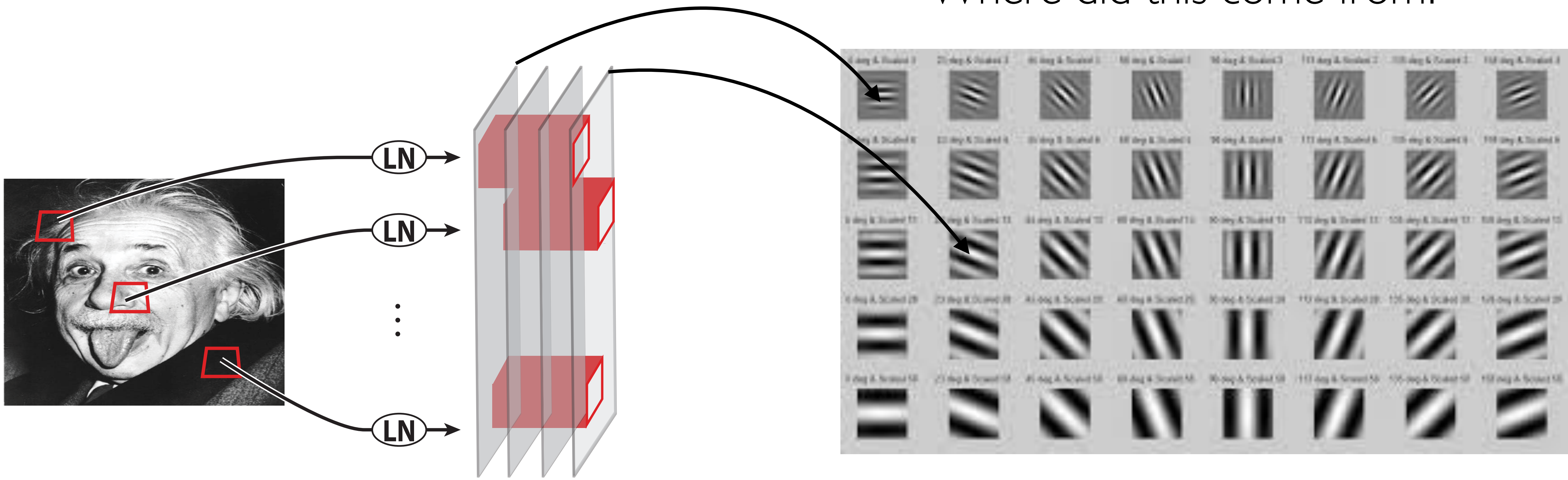
Lower areas, (RGC, LGN, V1) have been reasonably captured by single-layer convolutional model: <u>~50% of variance explained</u>. Carandini et. al (2005), Lennie & Movshon (2005)





Hierarchical Convolutional Neural Networks

Where did this come from?



(1) "Hubel and Wiesel's Intuition"
~1970s and formalized later

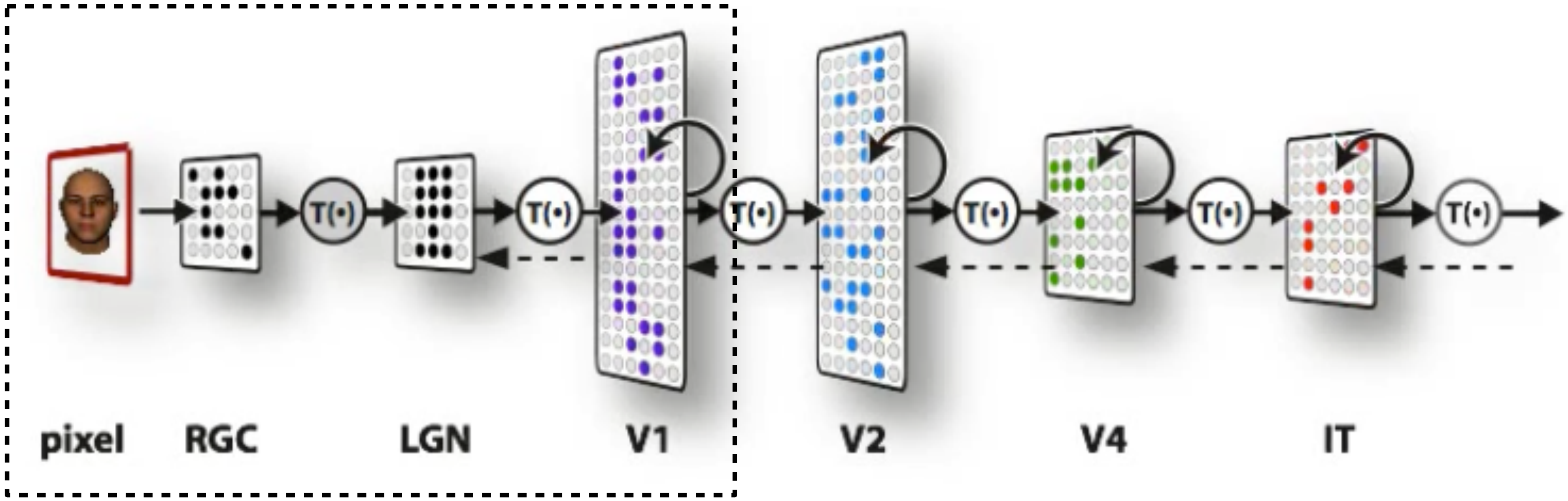
→ e.g. there is a "fixed basis set" that just "makes sense" if we're smart enough

(2) Sparse Coding Foldiak, Olshausen,
mid 1990s

→neurons have to represent their environment, as efficiently as possible

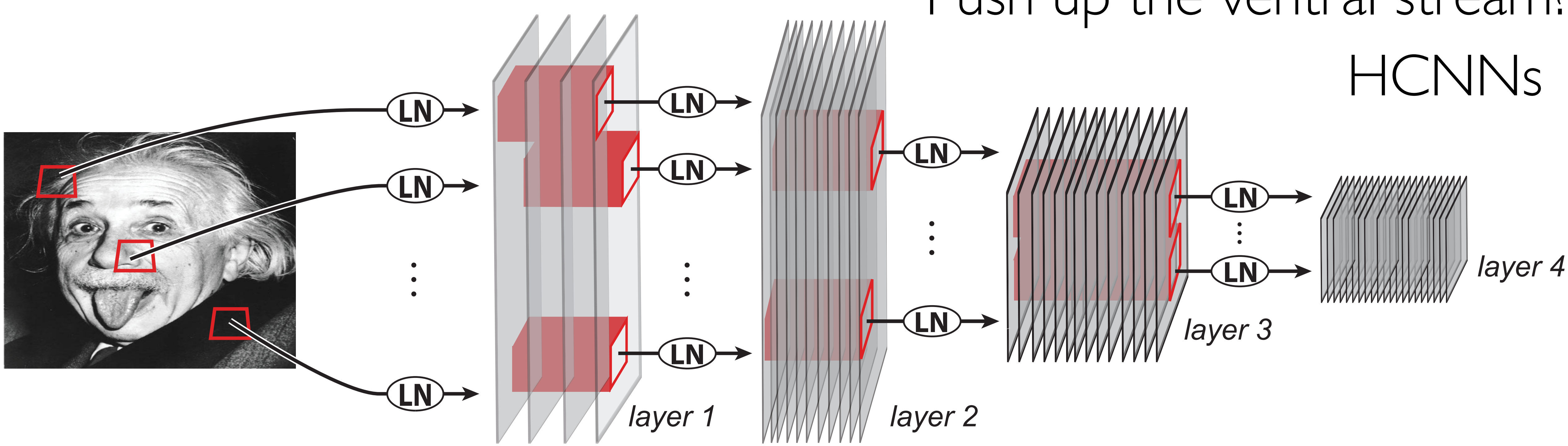
Hierarchical Convolutional Neural Networks

Lower areas, (RGC, LGN, V1) have been reasonably captured by single-layer models: <u>~50% of variance explained</u>. Carandini et. al (2005), Lennie & Movshon (2005)



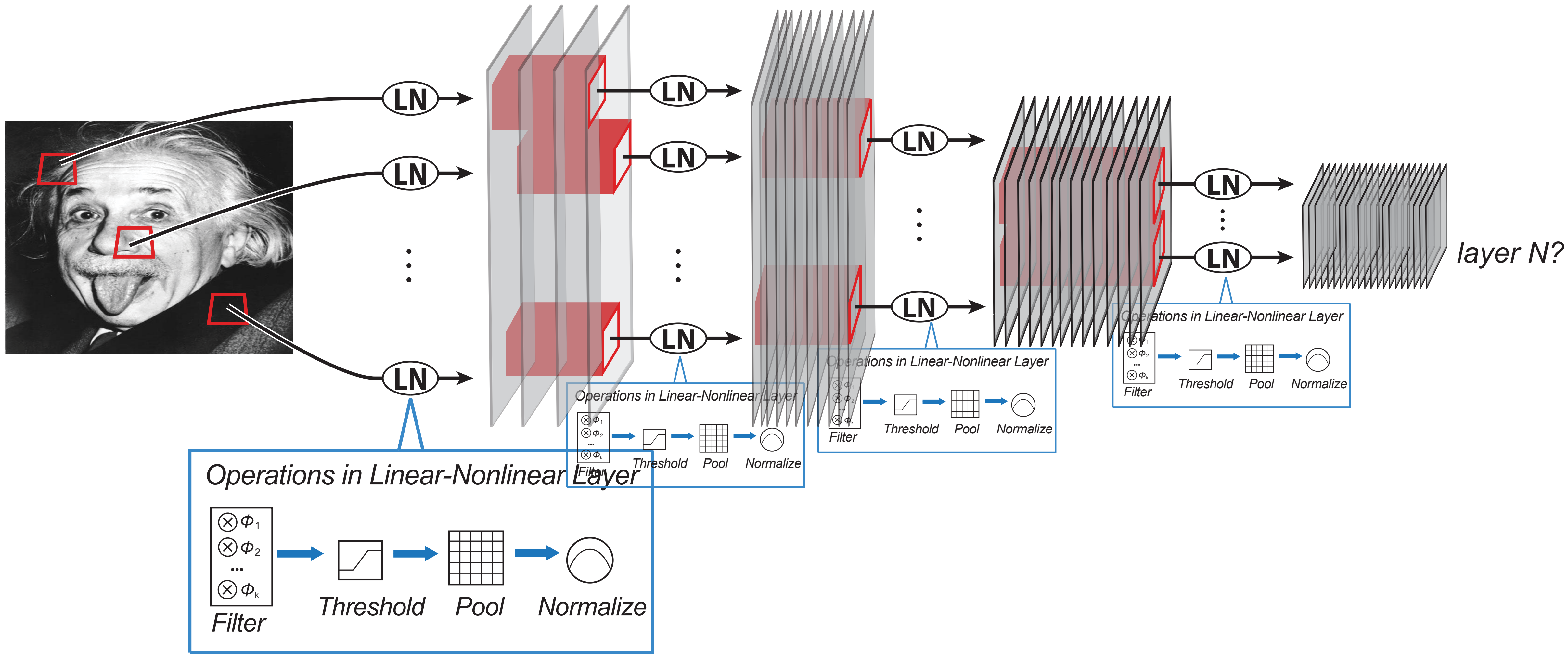
Push up the ventral stream?

HCNNs



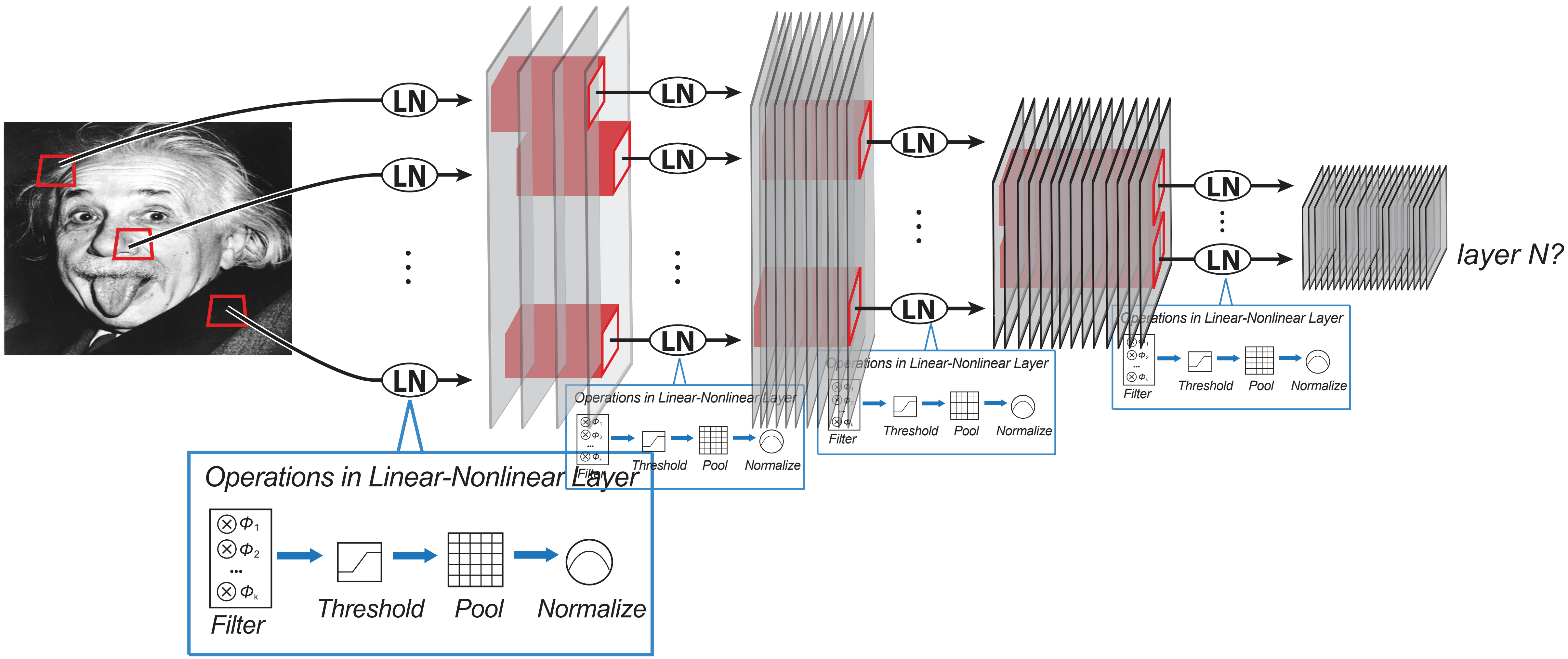
Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.



Neural Fitting Strategy?

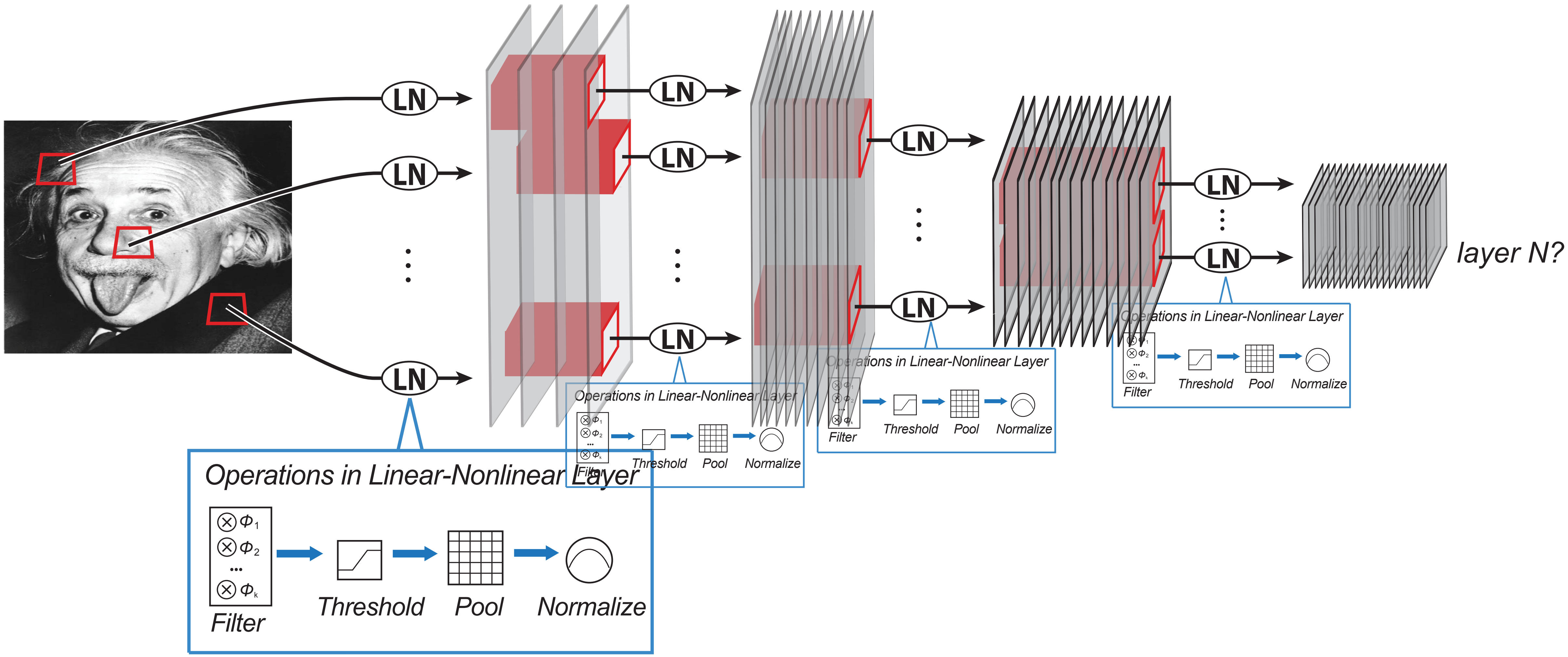
Huge number of parameters consistent with HCNN concept.



i. **architectural** params: (# layers, # filters, receptive field sizes, &c) — "network structure"

Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.

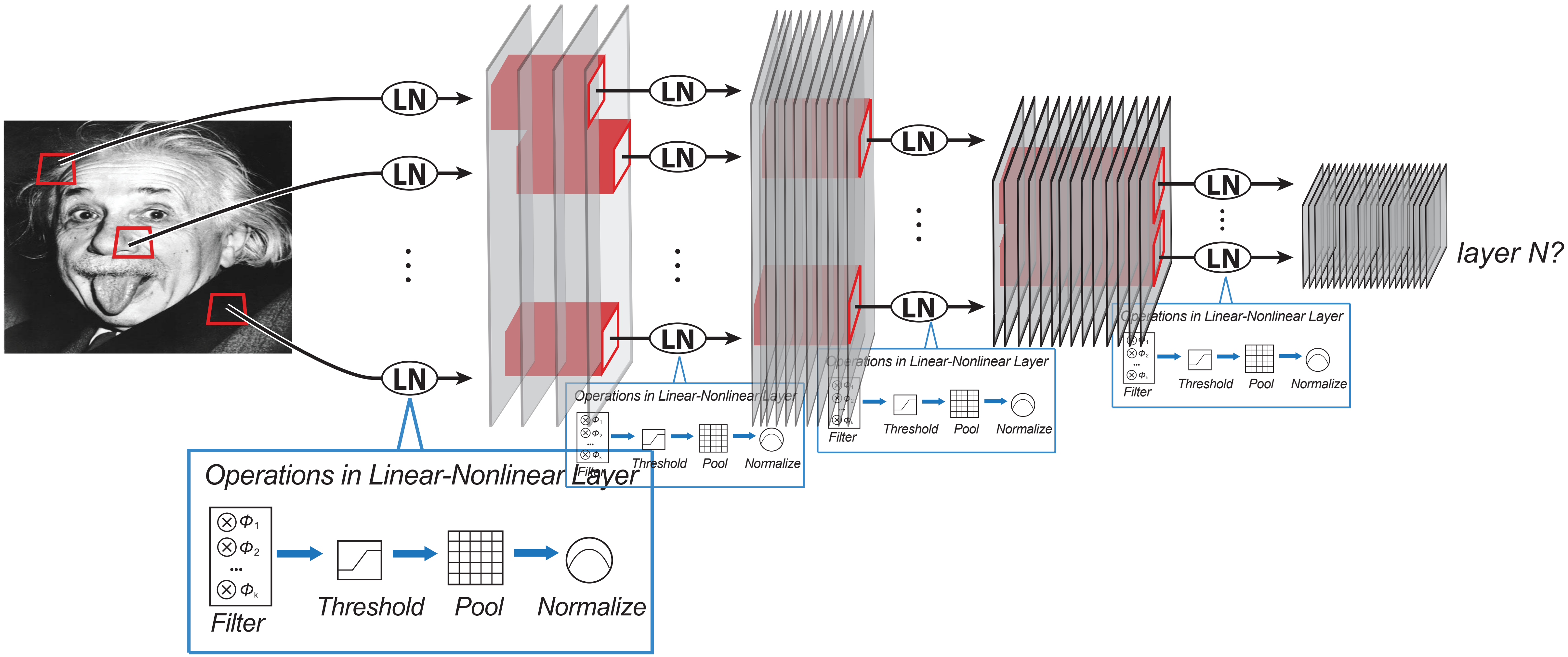


i. **architectural** params: (# layers, # filters, receptive field sizes, &c) — "network structure"

ii. **filter** parameters: continuous valued pattern templates — "network contents"

Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.



i. **architectural** params: (# layers, # filters, receptive field sizes, &c) — "network structure"

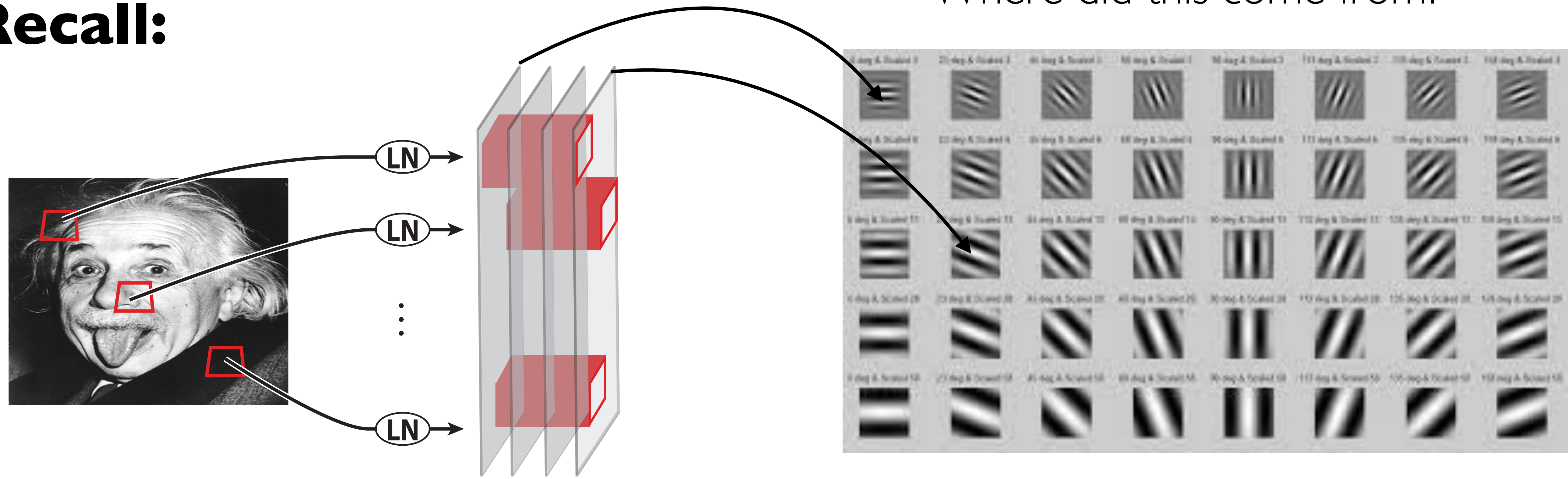
ii. **filter** parameters: continuous valued pattern templates — "network contents"

Q: How to discover the "right" parameters to understand real cortex?

Hierarchical Convolutional Neural Networks

Recall:

Where did this come from?



(1) "Hubel and Wiesel's Intuition"
~1970s and formalized later

→ e.g. there is a "fixed basis set" that just "makes sense" if we're smart enough

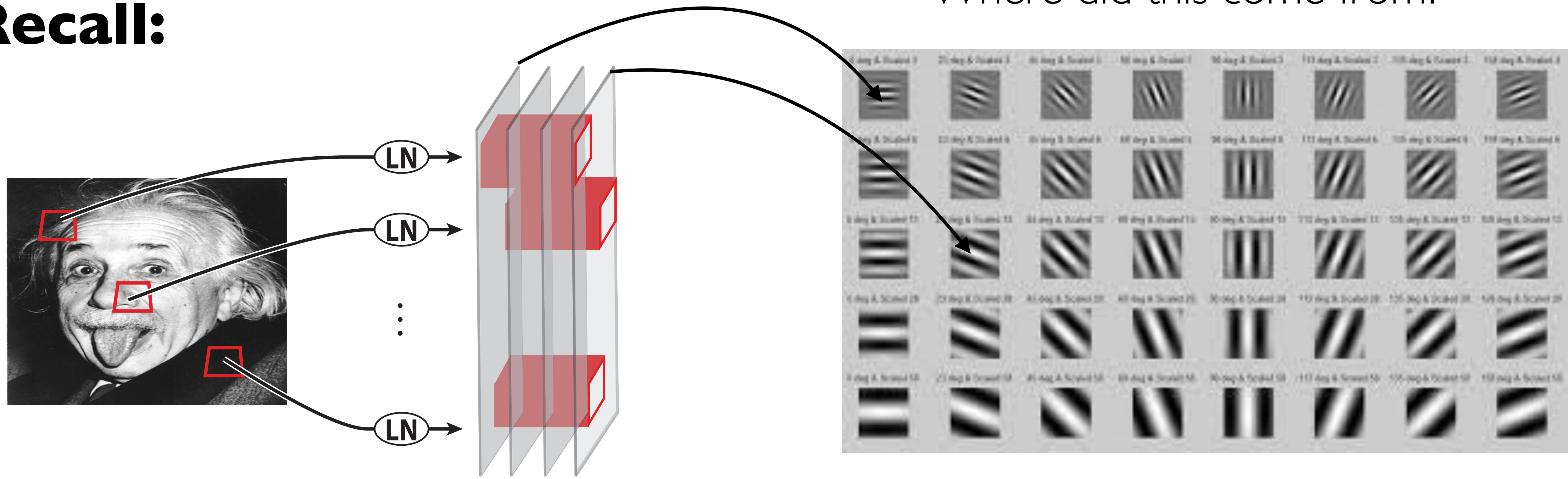
(2) Sparse Coding Foldiak, Olshausen,
mid 1990s

→neurons have to represent their environment, as efficiently as possible

Hierarchical Convolutional Neural Networks

Recall:

Where did this come from?



(1) "Hubel and Wiesel's Intuition"
~1970s and formalized later

→ e.g. there is a "fixed basis set" that just "makes sense" if we're smart enough

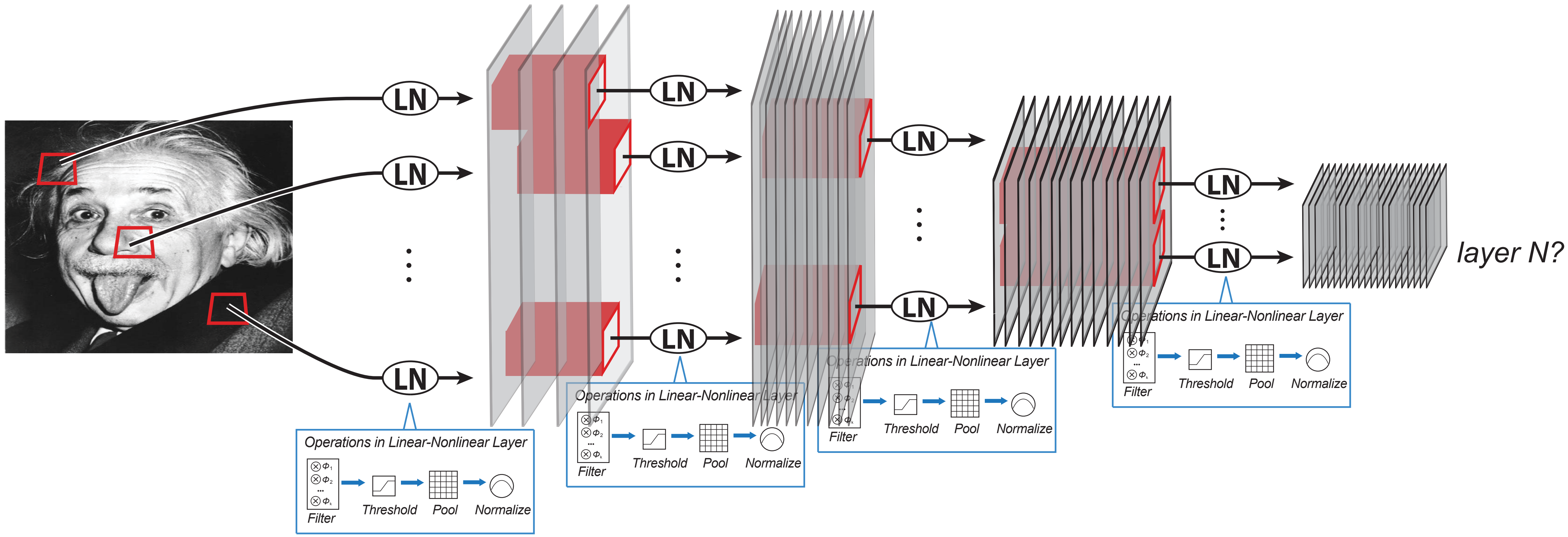
REALLY HARD TO GENERALIZE TO MULTI-LAYER NETWORKS

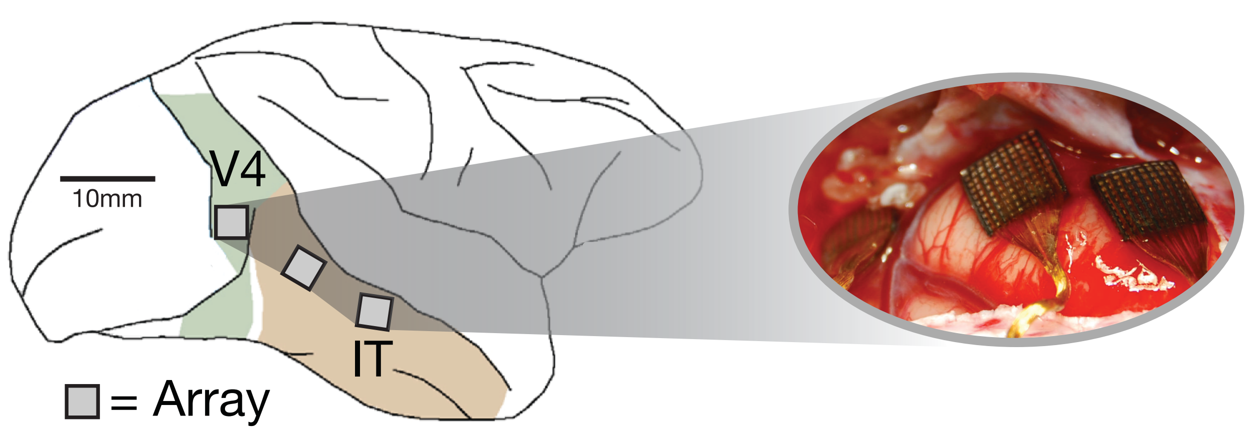
(2) Sparse Coding Foldiak, Olshausen, mid 1990s

→neurons have to represent their environment, as efficiently as possible

Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.

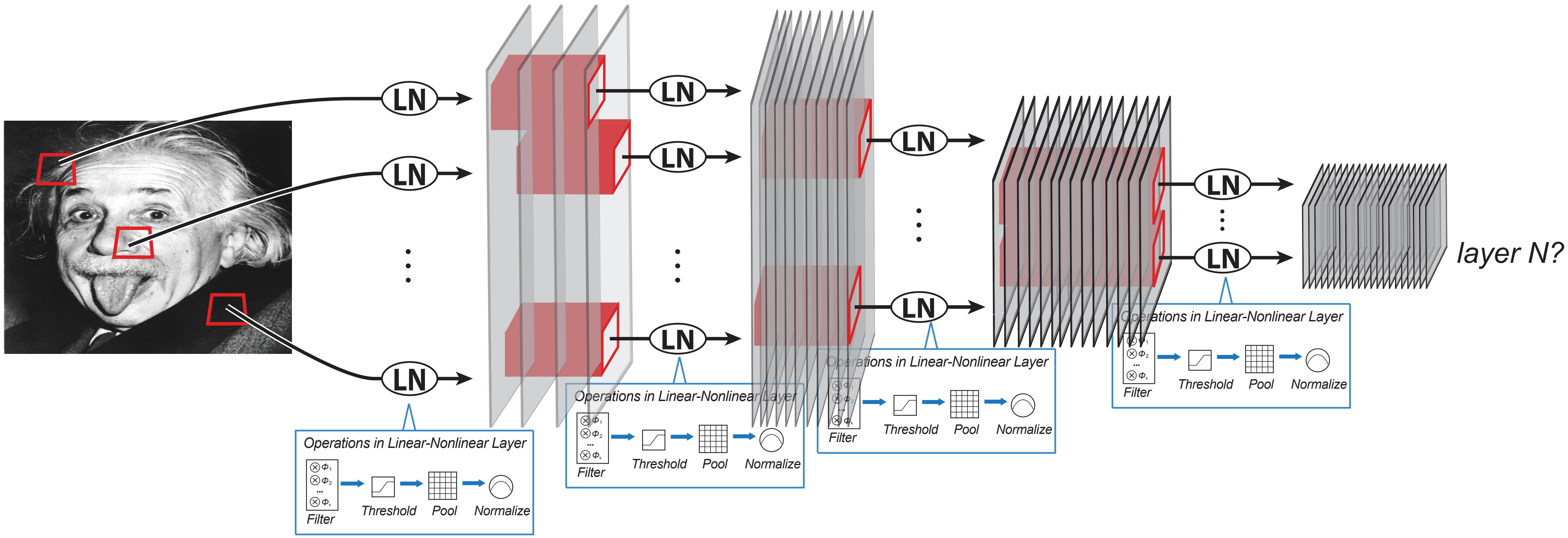


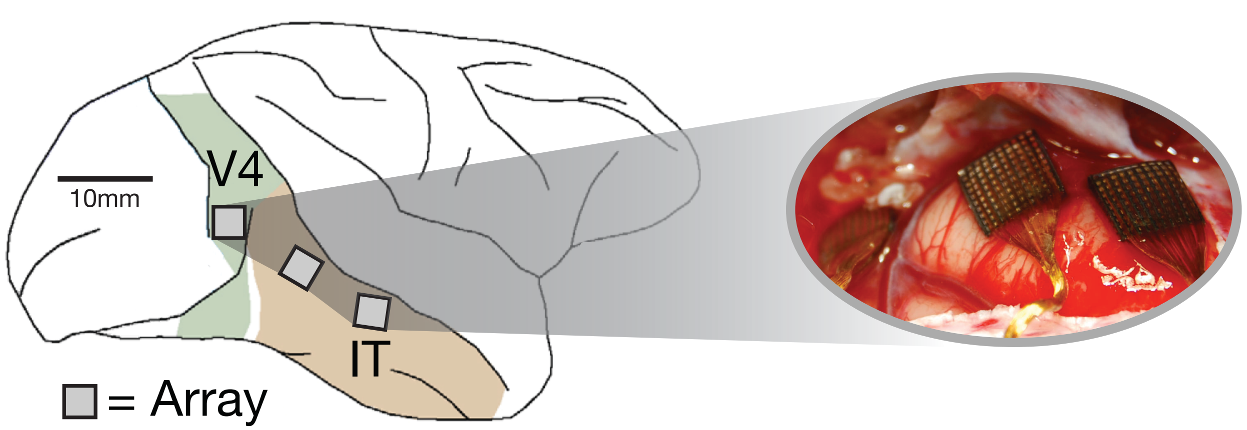


Obvious alternative strategy: fit parameters to neural data.

Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.

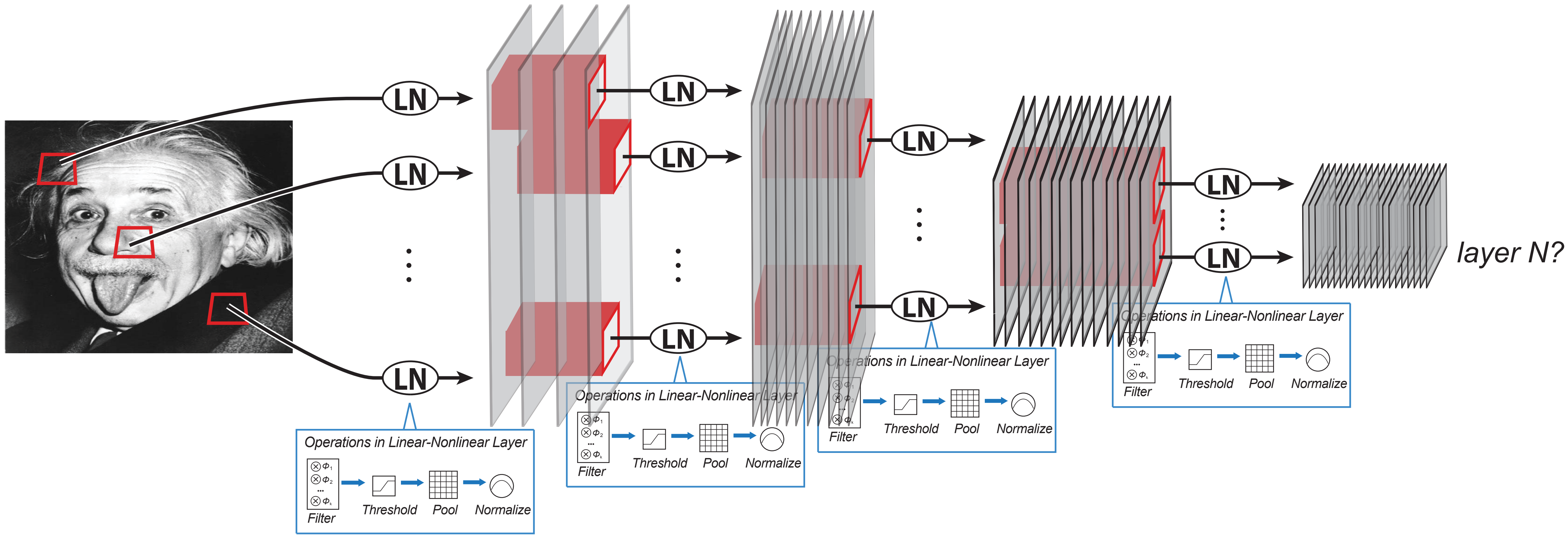


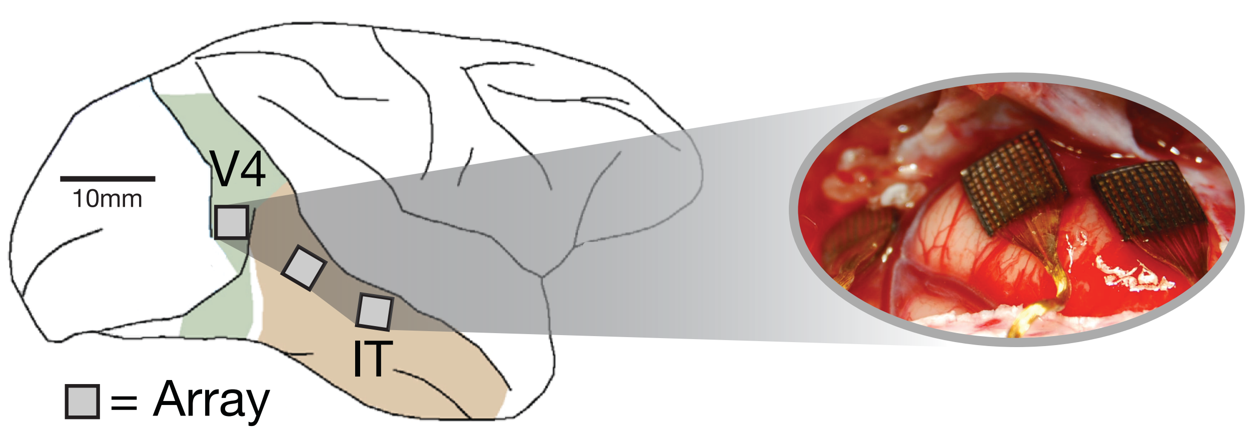


…not enough neural data to constrain model class. Gallant (2007); Rust & Movshon (2006)

Neural Fitting Strategy?

Huge number of parameters consistent with HCNN concept.

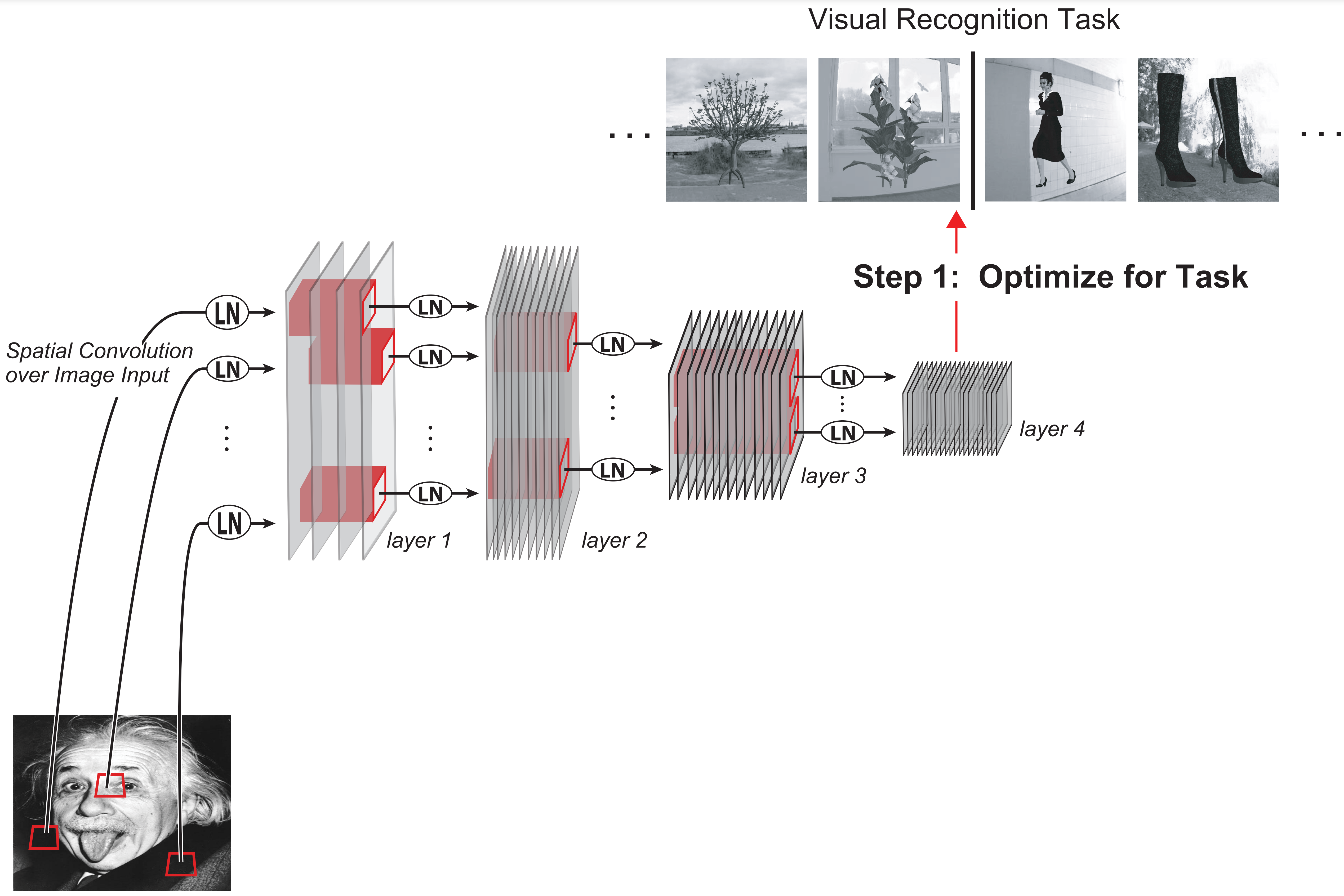




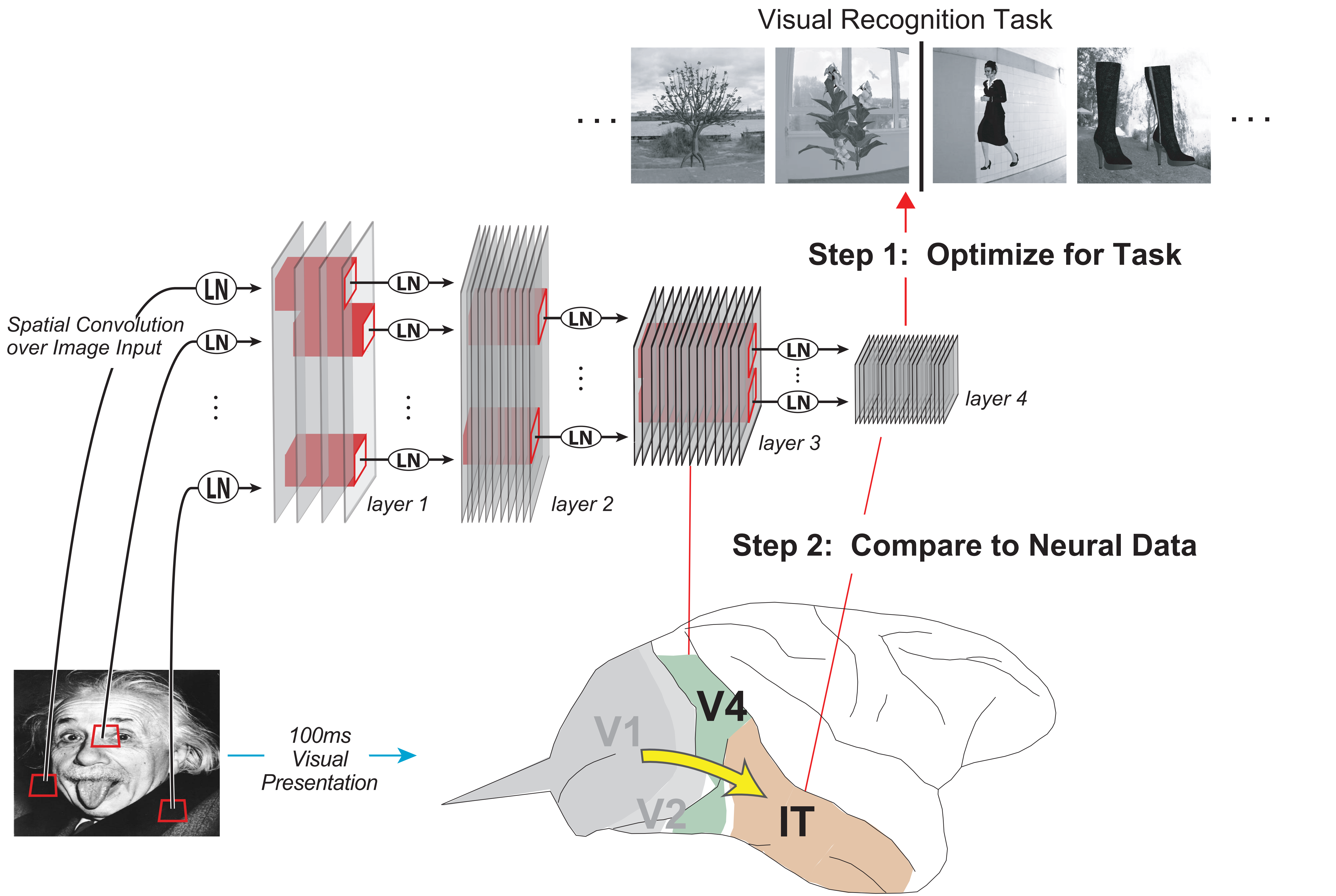
…not enough neural data to constrain model class. Gallant (2007); Rust & Movshon (2006)

Overfitting.

Optimize for Performance, Test Against Neurons



Optimize for Performance, Test Against Neurons



Optimize for Performance, Test Against Neurons

1. **Performance**: accuracy on a <u>challenging, high-variation</u> visual object categorization task.

2. **Neural predictivity**: the ability of model to predict each individual neural site's activity.

Optimize for Performance, Test Against Neurons

1. **Performance**: accuracy on a <u>challenging, high-variation*</u> visual object categorization task.

2. **Neural predictivity**: the ability of model to predict each individual neural site's activity.

<u>*challenging for neural network engineers, not the animal</u>

Optimize for Performance, Test Against Neurons

1. **Performance**: accuracy on a <u>challenging, high-variation*</u> visual object categorization task.

2. **Neural predictivity**: the ability of model to predict each individual neural site's activity.

Our hypothesis: Performance (1) → neural predictivity (2).

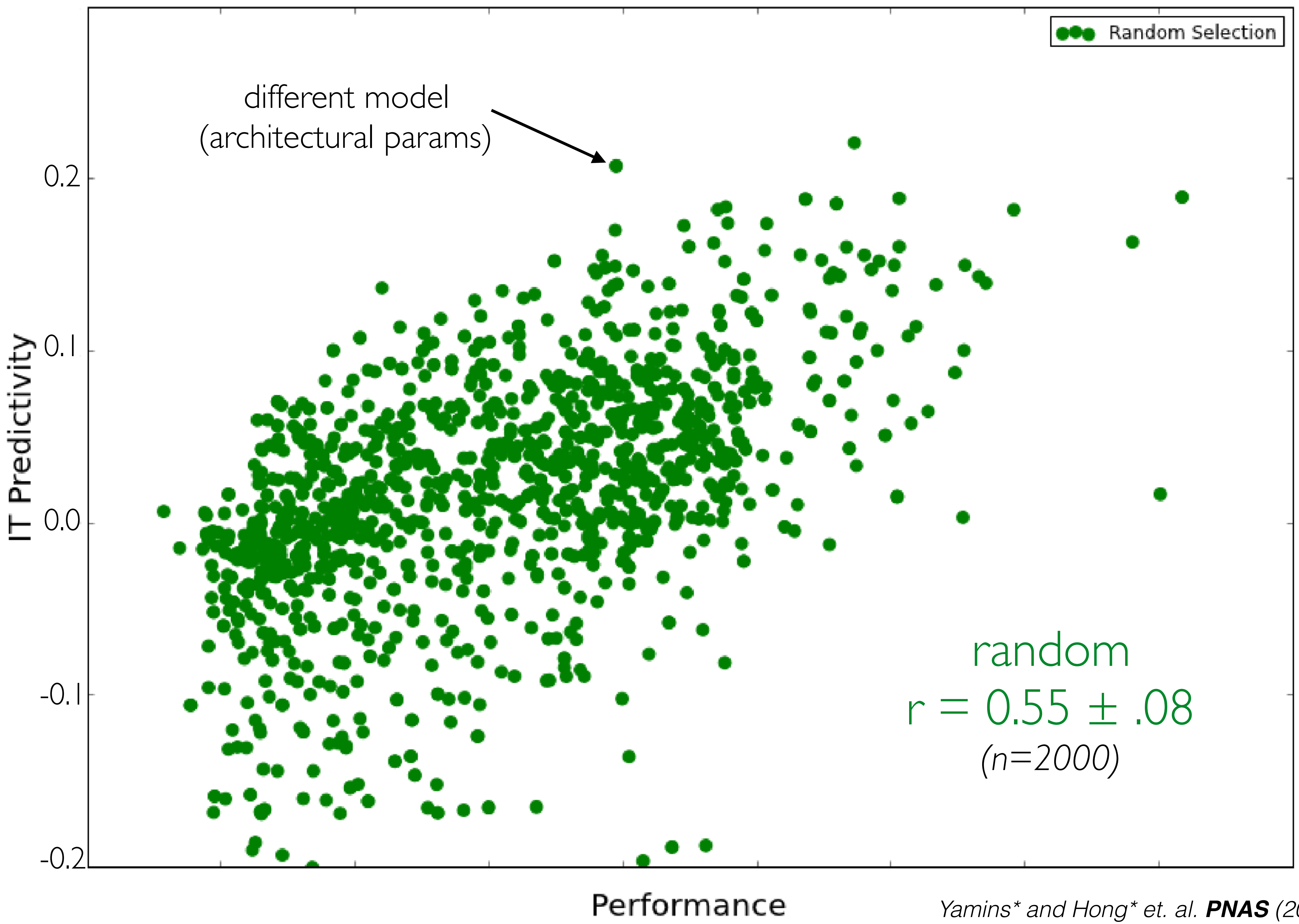
<u>*challenging for neural network engineers, not the animal</u>

Initial Validation of Idea

High-throughput experiments to directly test the relationship between performance and IT neural predictivity.

- Random selection of model parameters; measure performance and neural predictivity Pinto et. al (2008, 2009)

Initial Validation of Idea



Yamins and Hong* et. al.* ***PNAS*** *(2014)*

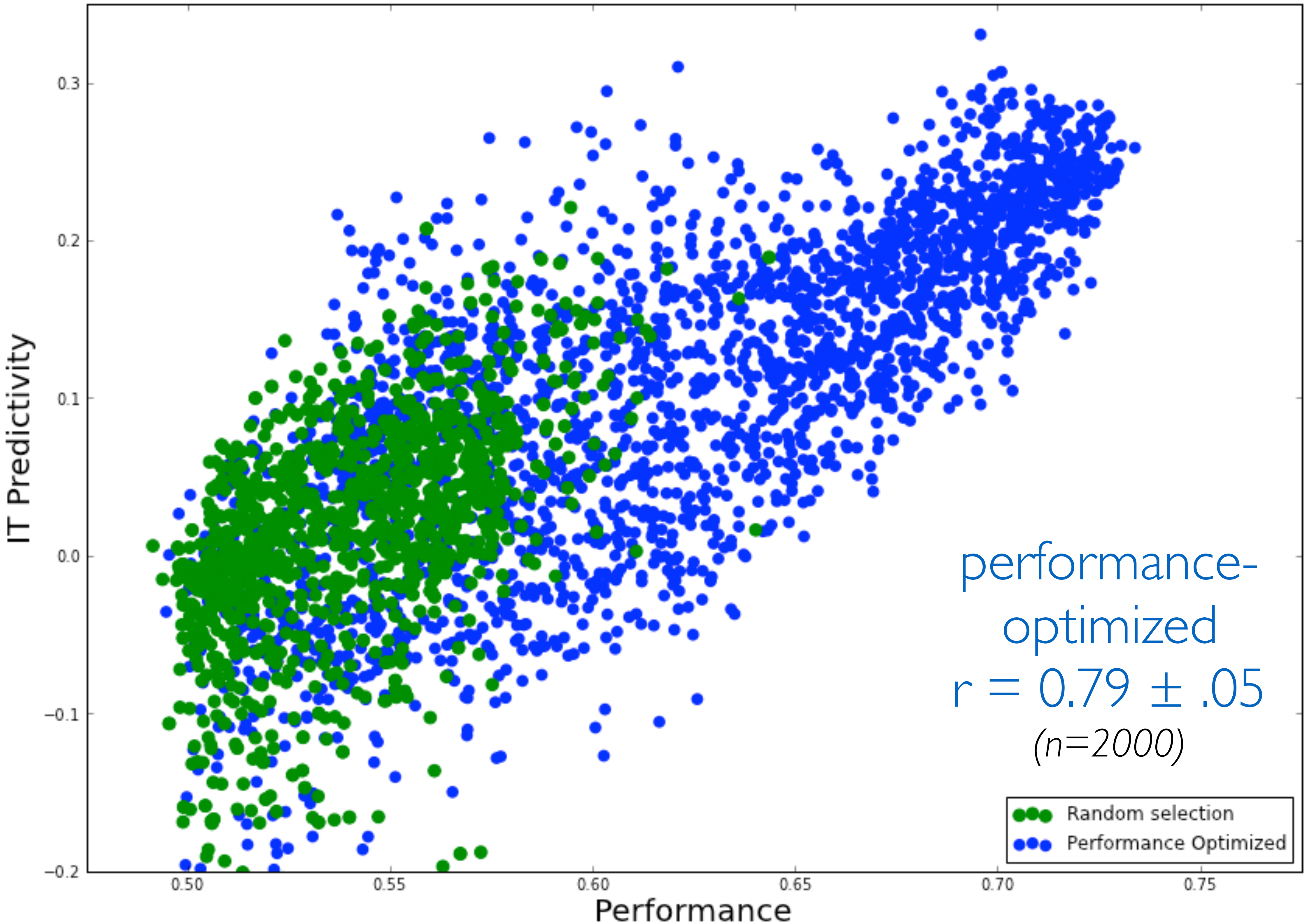
Initial Validation of Idea

High-throughput experiments to directly test the relationship between neural predictivity and performance.

- Random selection of model parameters; measure performance and neural predictivity   Pinto et. al (2008, 2009)

- Optimize parameters for performance; measure neural predictivity. optimization techniques: Bergstra Yamins & Cox (2013)

Initial Validation of Idea



performance-optimized
$r = 0.79 \pm .05$
(n=2000)

Yamins and Hong* et. al.* ***PNAS*** *(2014)*

Initial Validation of Idea

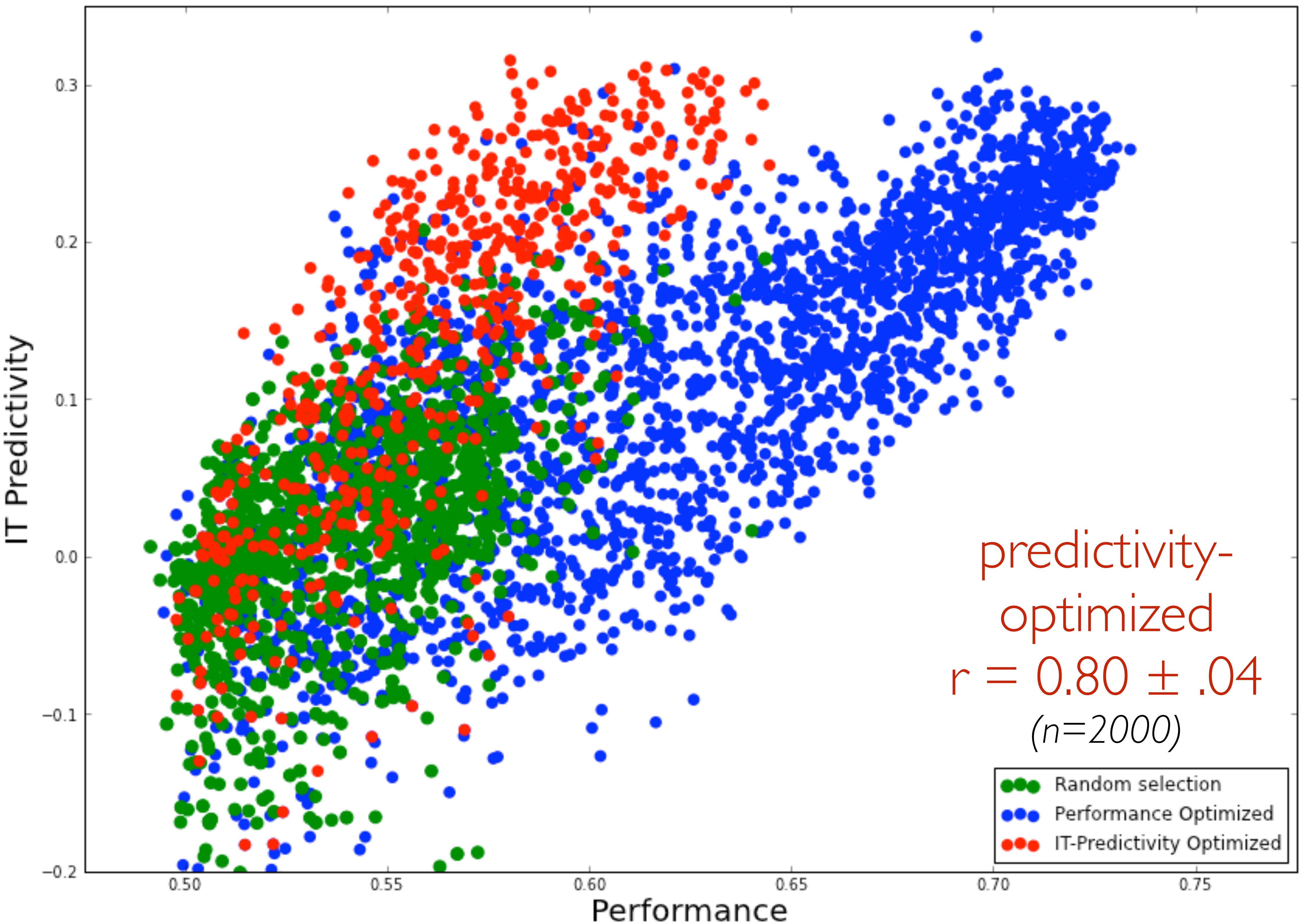
High-throughput experiments to directly test the relationship between neural predictivity and performance.

- Random selection of model parameters; measure performance and neural predictivity Pinto et. al (2008, 2009)

- Optimize parameters for performance; measure neural predictivity optimization techniques: Bergstra Yamins & Cox (2013)

- Optimize parameters for neural predictivity; measure performance

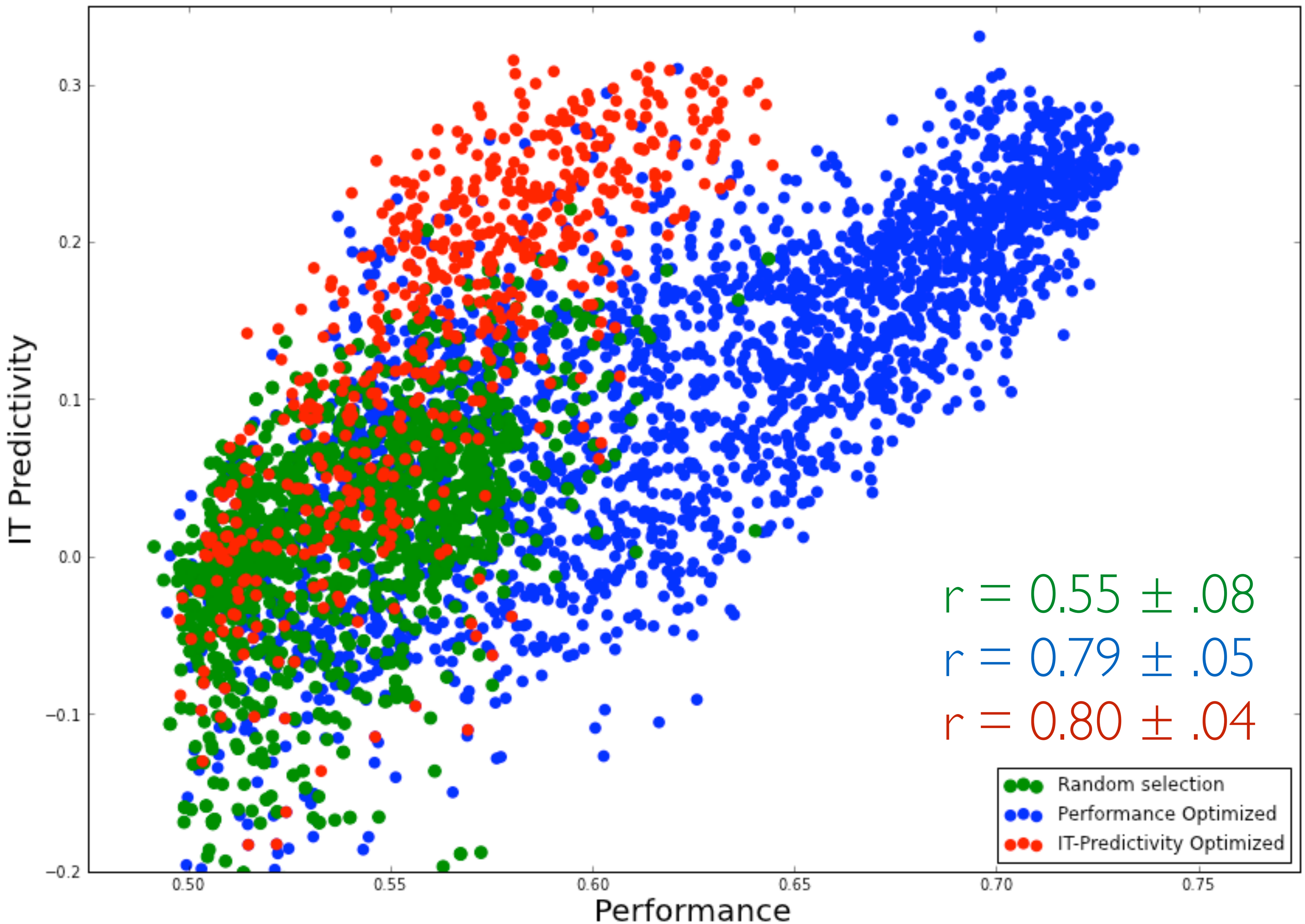
Performance vs IT predictivity: Predictivity-Optimized



Yamins and Hong* et. al.* ***PNAS*** *(2014)*

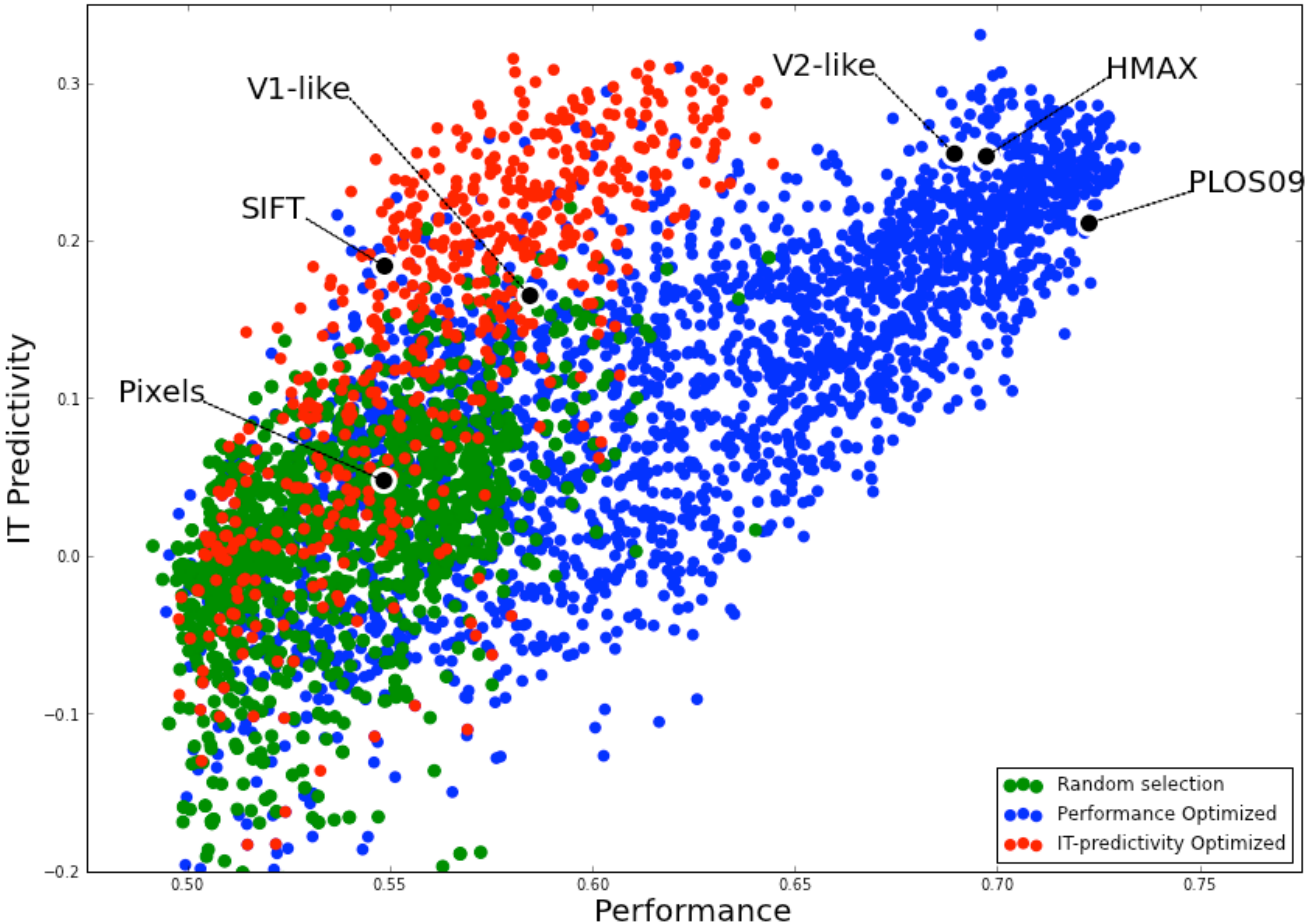
Performance vs IT predictivity: Predictivity-Optimized

Performance is a potentially very good driver of neural prediction.



Yamins and Hong* et. al.* ***PNAS*** *(2014)*

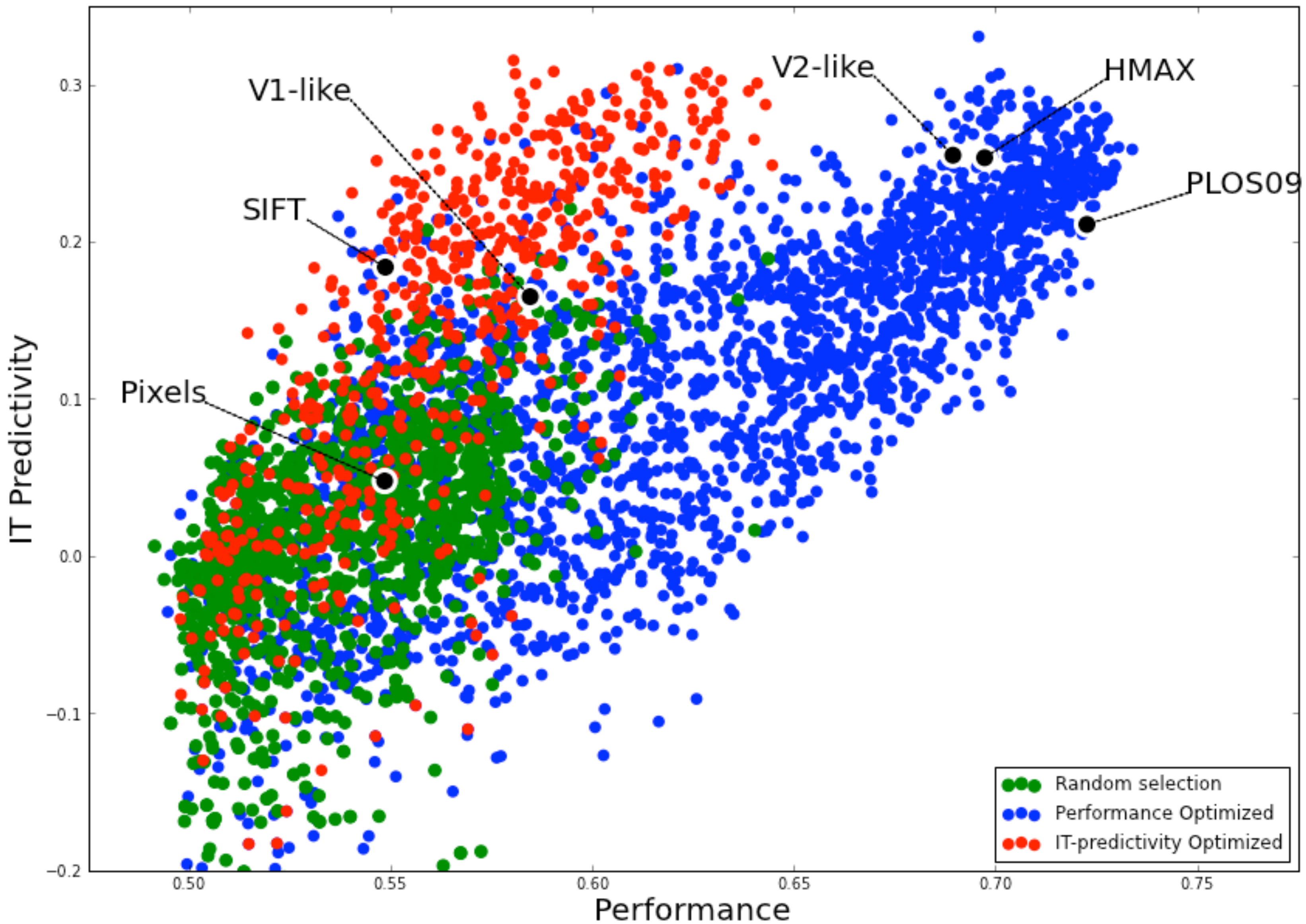
Performance vs IT predictivity



Yamins and Hong* et. al.* ***PNAS*** *(2014)*

Performance vs IT predictivity

But, not doing that well.   Really want to be here:



Optimization Strategy

i. **architectural** params: (# layers, # filters, receptive field sizes, &c) — "network structure"

→ Automated meta-parameter optimization in high-dimensional discrete parameter spaces Bergstra Yamins & Cox (2013)

→ Ensembles of models chosen through modified boosting Yamins et. al (2013, 2014)

Optimization Strategy

i. **architectural** params: (# layers, # filters, receptive field sizes, &c) — "network structure"

→ Automated meta-parameter optimization in high-dimensional discrete parameter spaces
Bergstra Yamins & Cox (2013)

→ Ensembles of models chosen through modified boosting Yamins et. al (2013, 2014)

ii. **filter** parameters: continuous valued pattern templates — "network contents"

→GPU-accelerated stochastic gradient descent Pinto et. al., (2009), Krizhevsky et. al. (2012)

Gradient descent eq: $$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

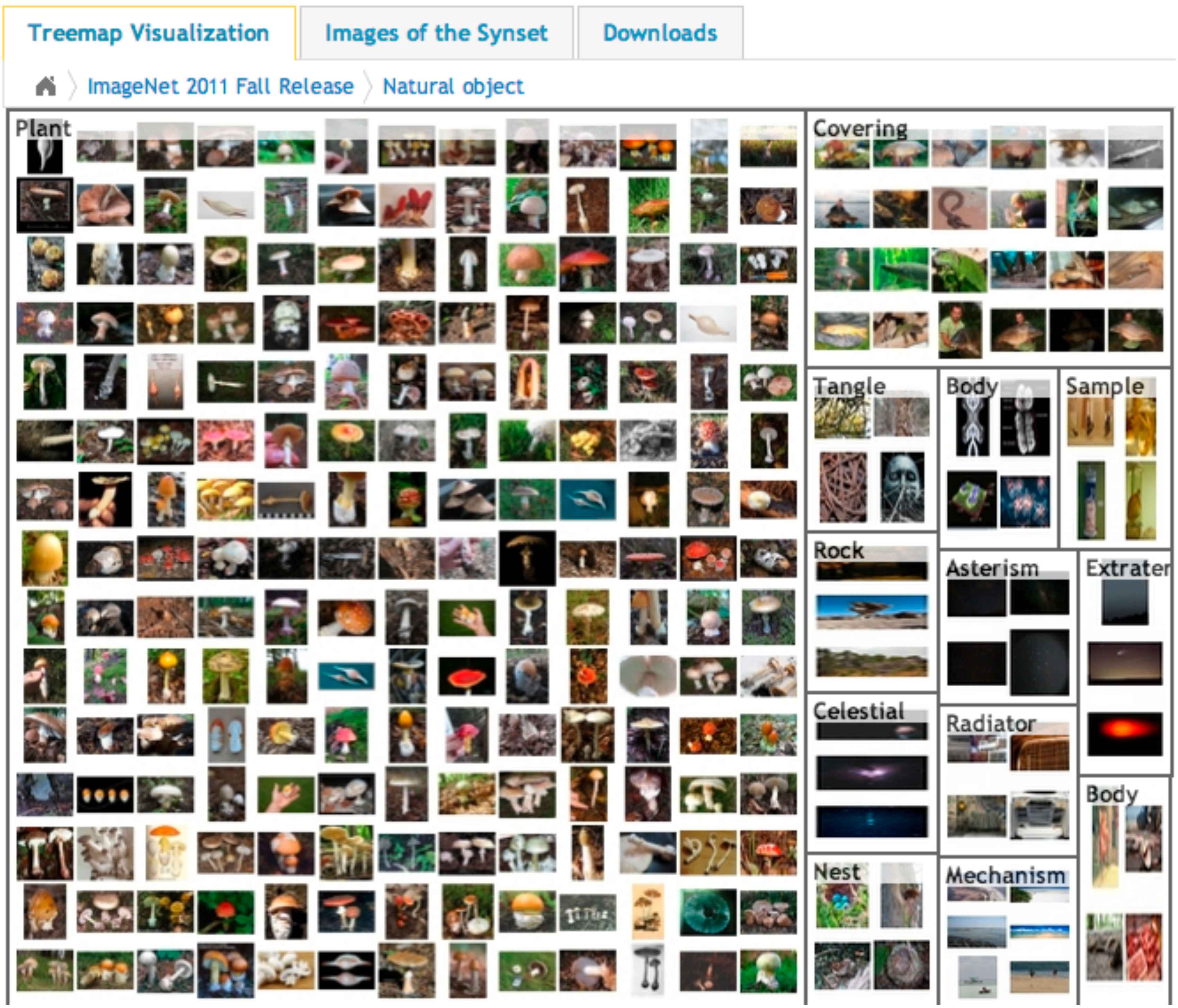
L = loss function
λ = learning rate
D = dataset

In current practice:

L = loss computed from **large numbers of externally-provided object category labels.**

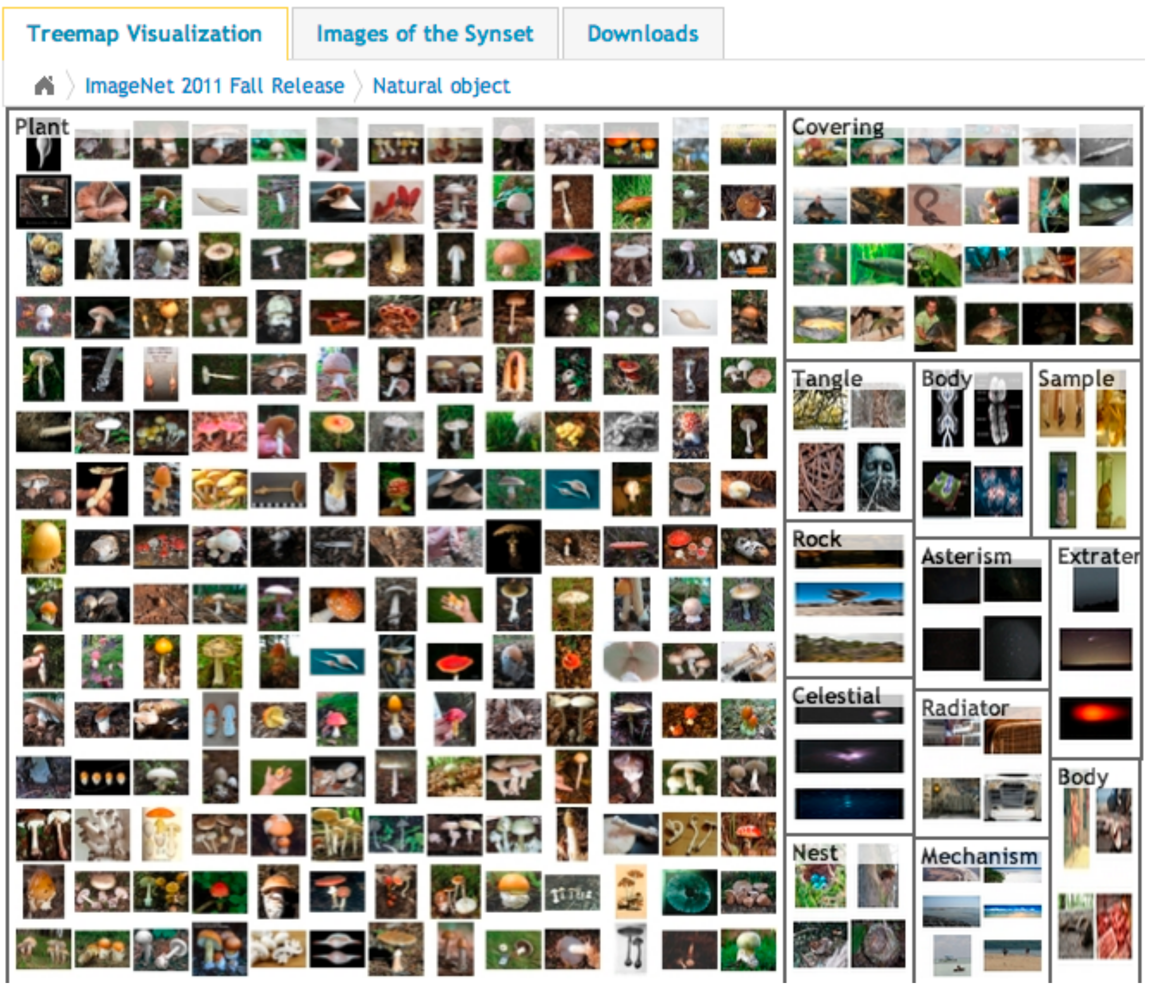
Model Training Regimen

ImageNet (2012). Thousands of images in thousands of categories.



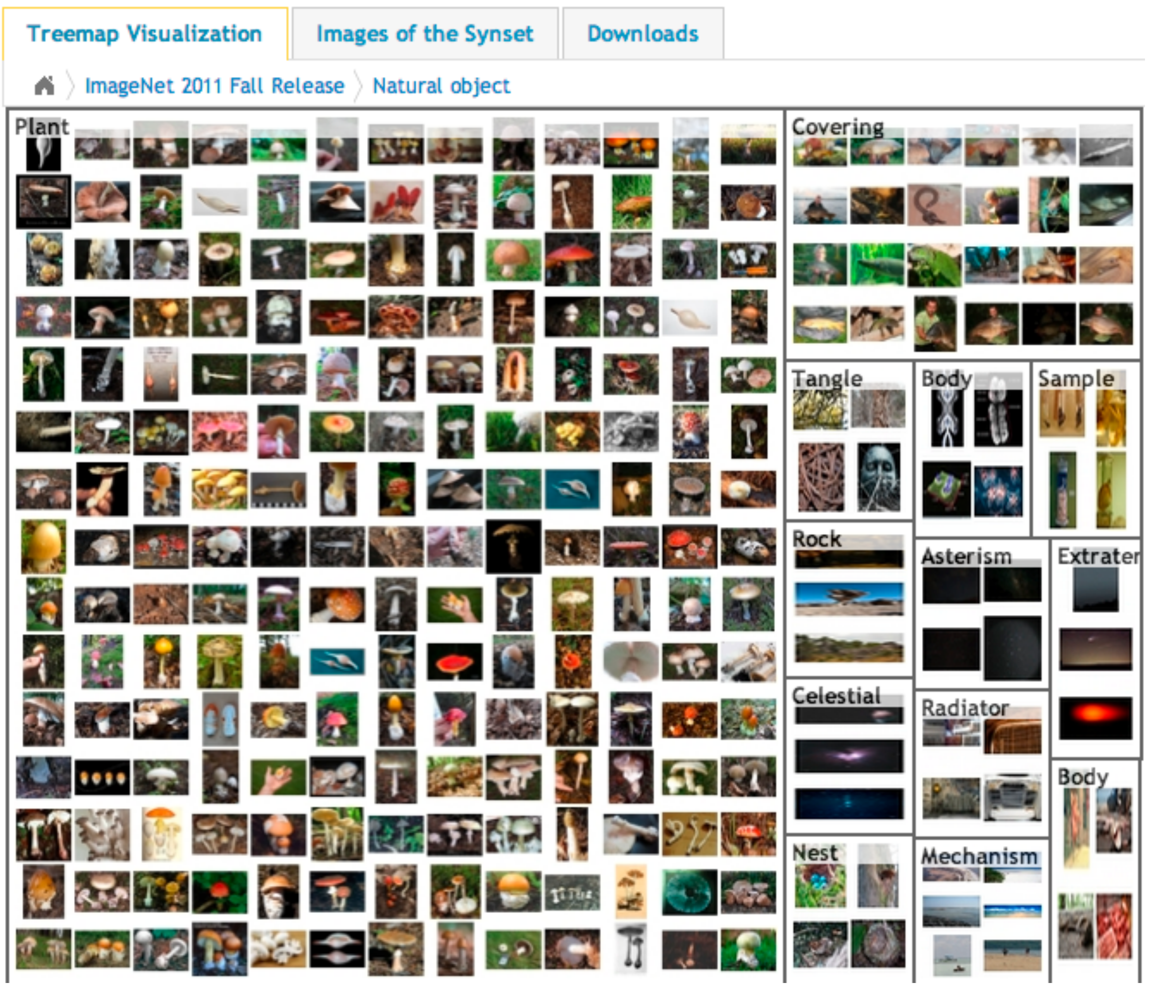
Model Training Regimen

train: real photos



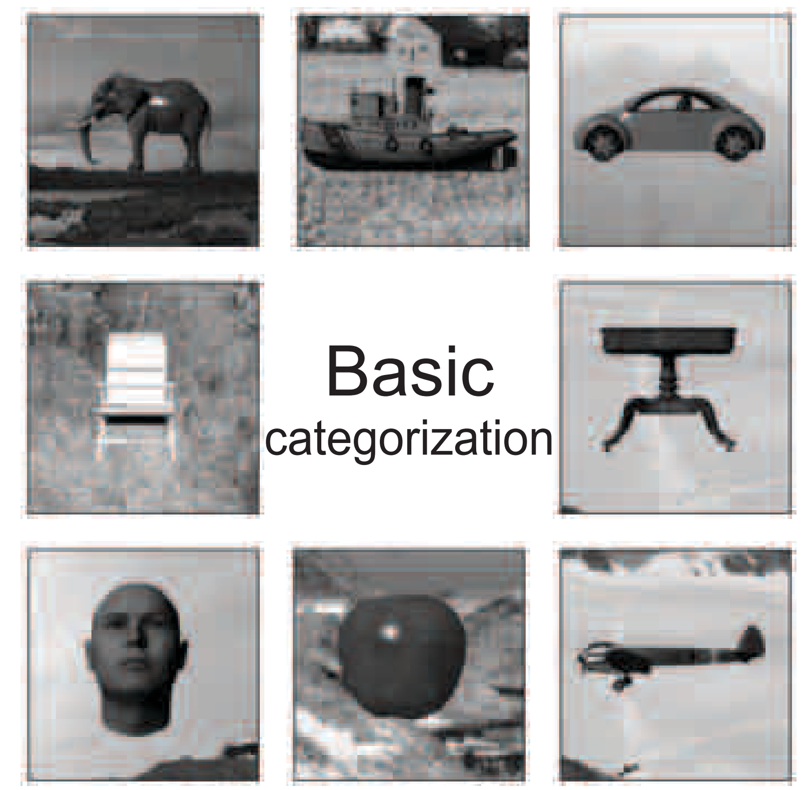
Model Training Regimen

train: real photos



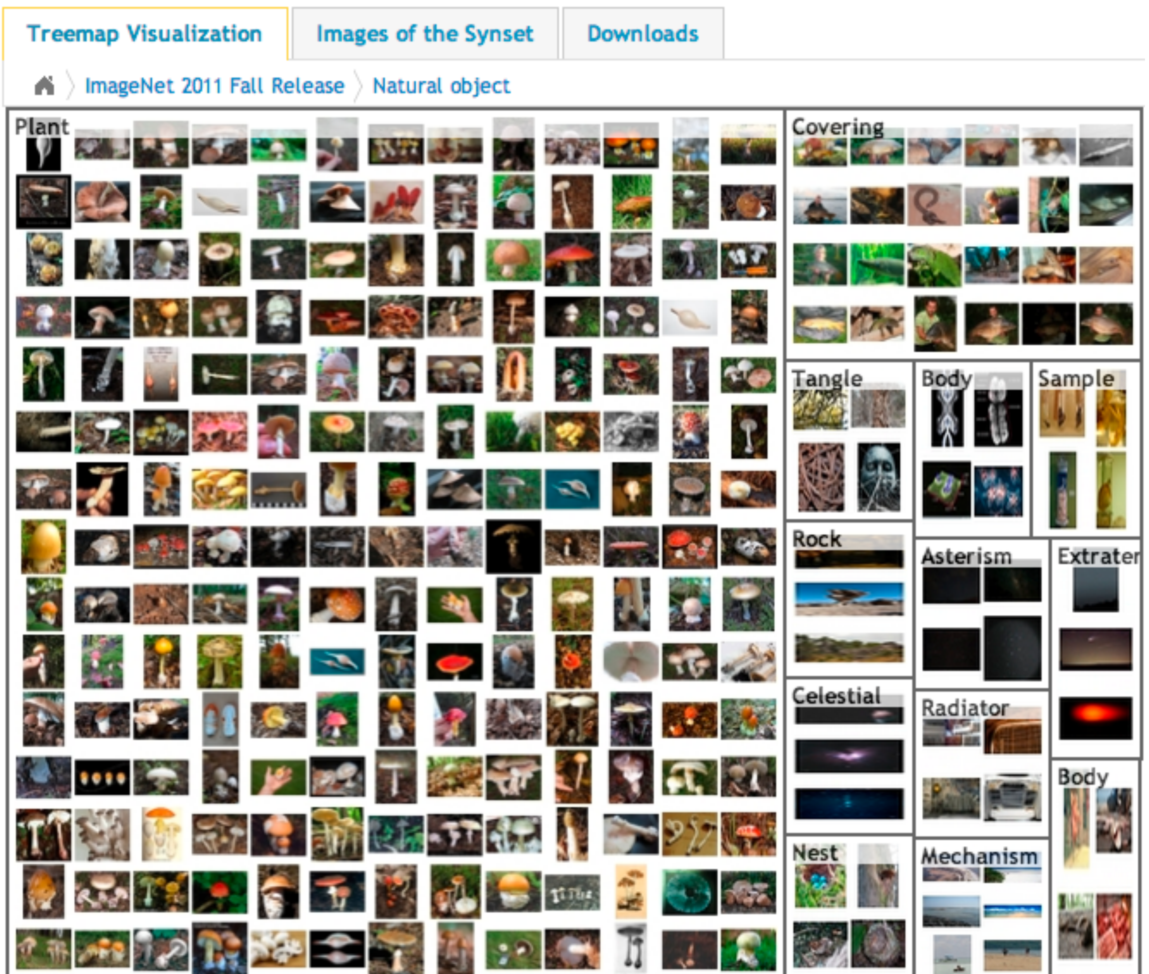
generalize?

test: neural stimuli



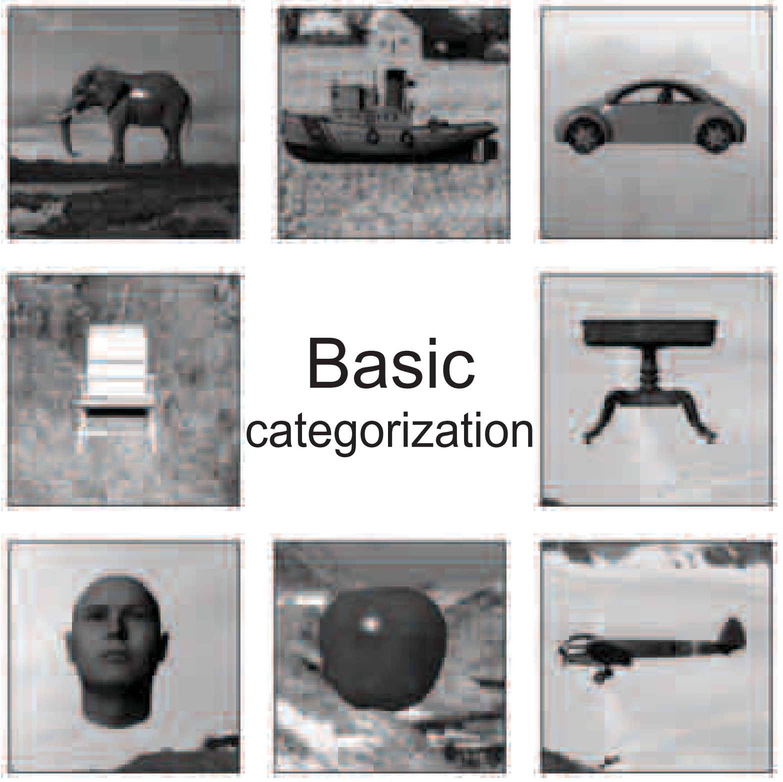
Model Training Regimen

train: real photos



generalize?

test: neural stimuli

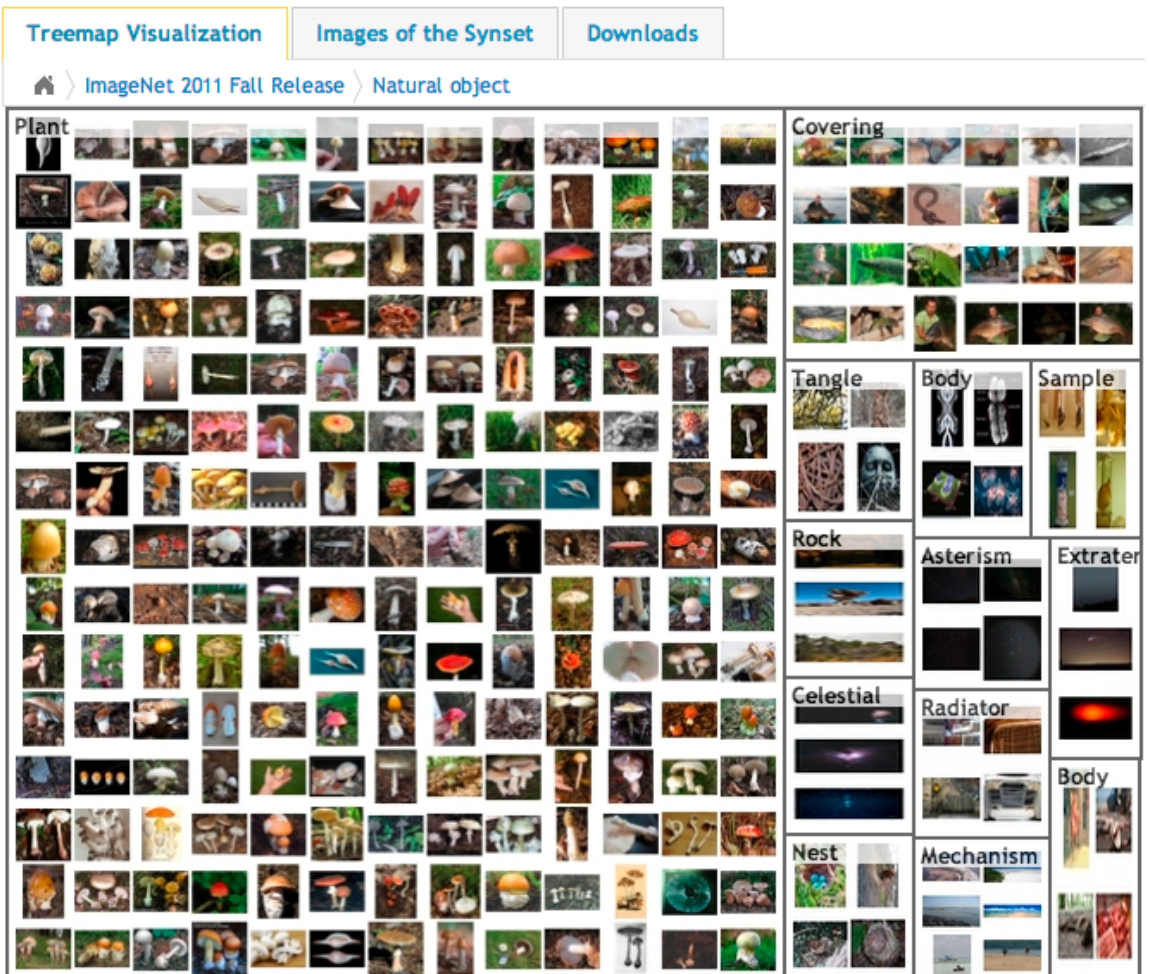


removed categories of photos that appeared in the test stimuli

(animals, boats, cars, chairs, faces, fruits, planes, tables)

Model Training Regimen

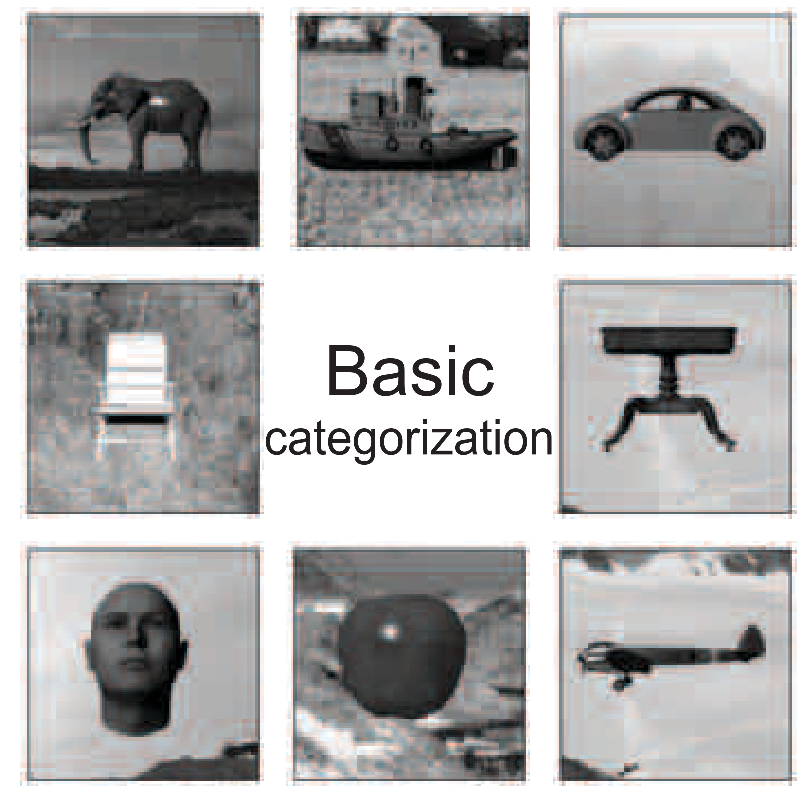
train: real photos



generalize?

→

test: neural stimuli

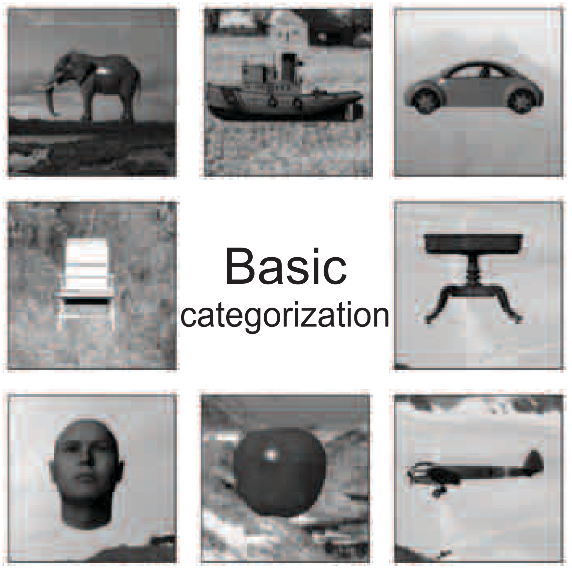


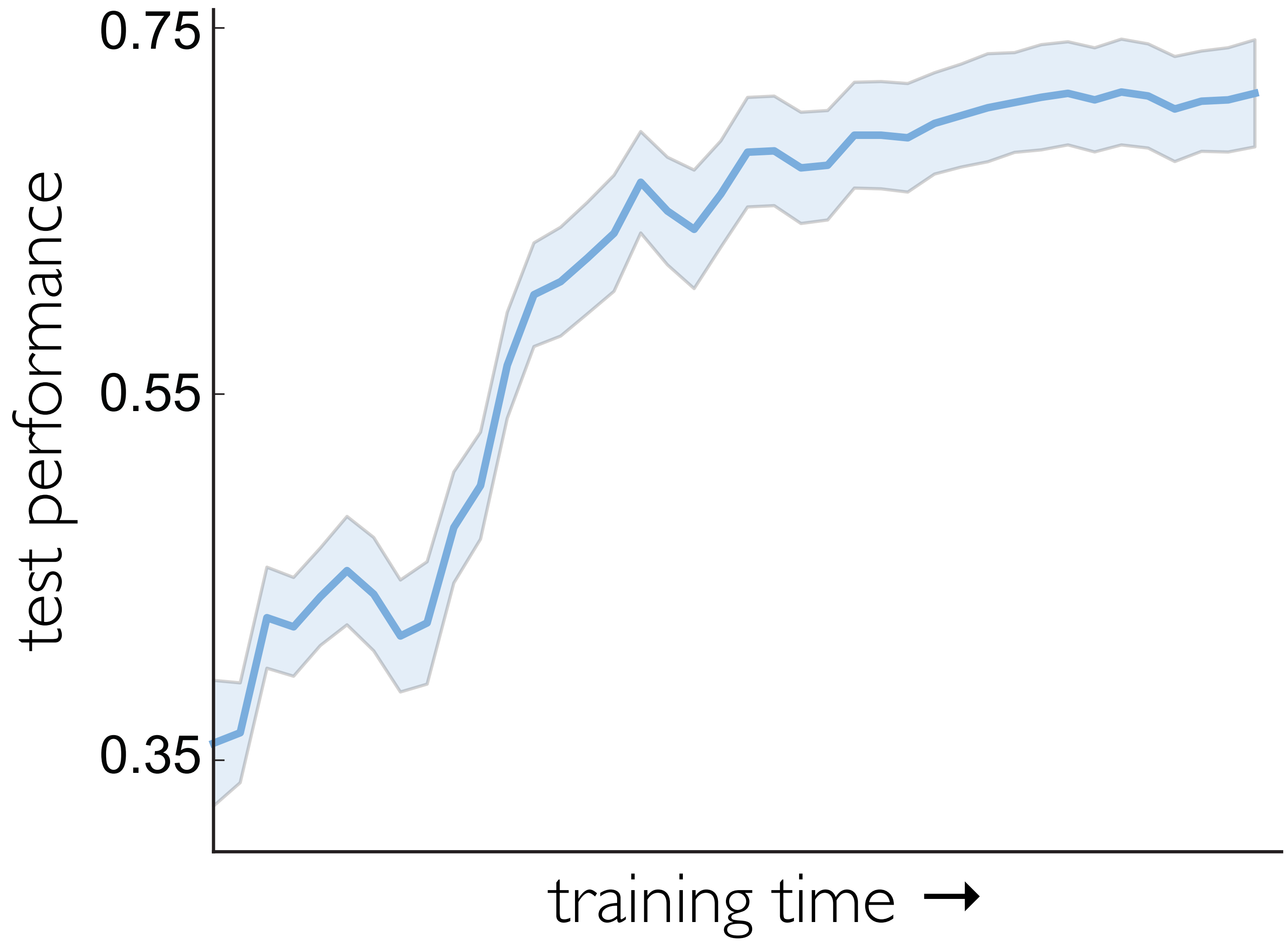
removed categories of photos that appeared in the test stimuli

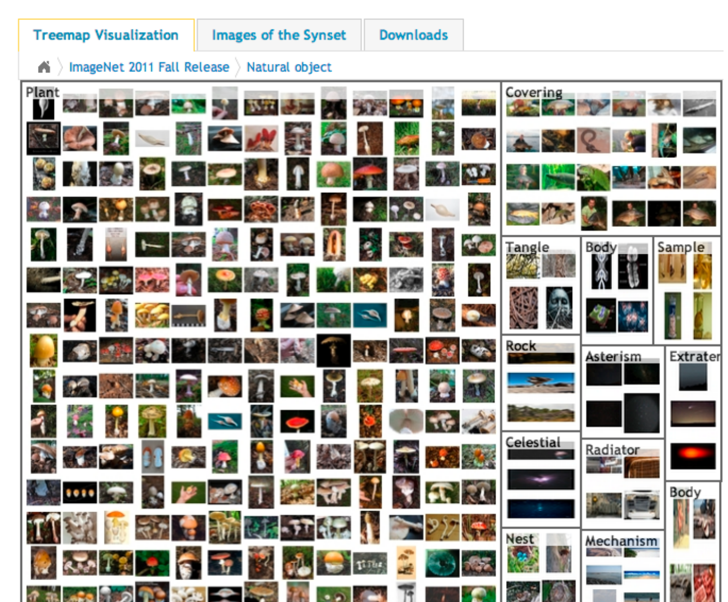
(animals, boats, cars, chairs, faces, fruits, planes, tables)

→ Specific 4-layer model that achieved high recognition performance.

Performance Generalization





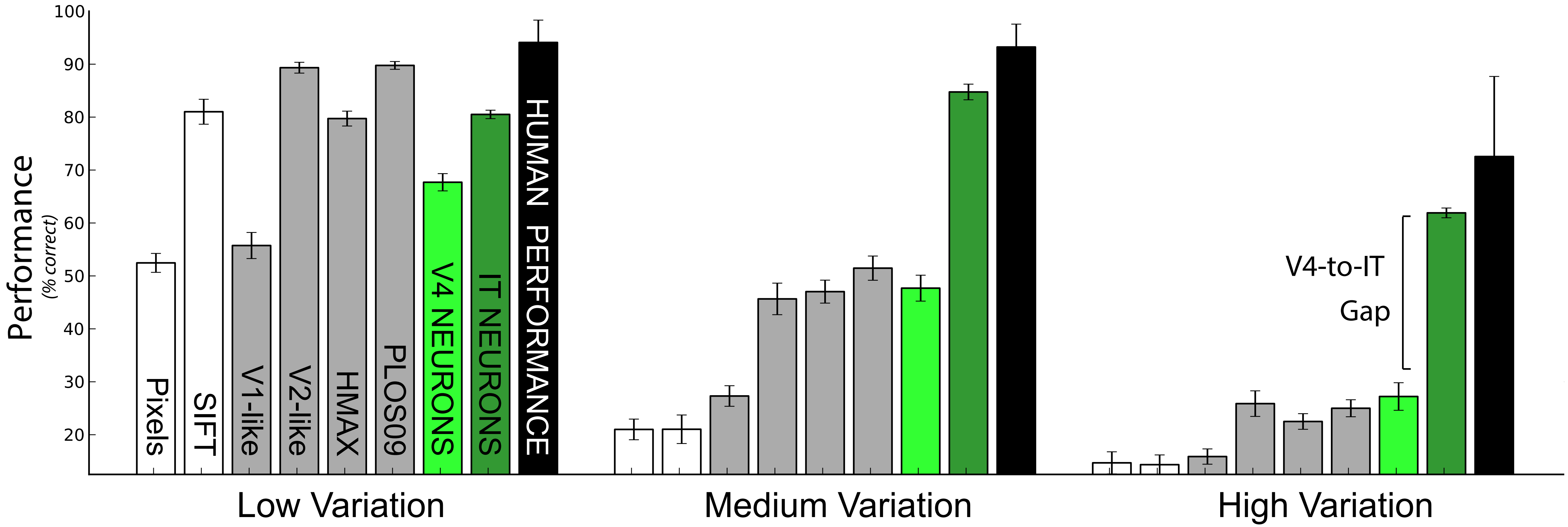


IT Neurons Track Human Performance

V4 loses out at higher variation:

… but humans are much less affected.

… as is the IT neural population.

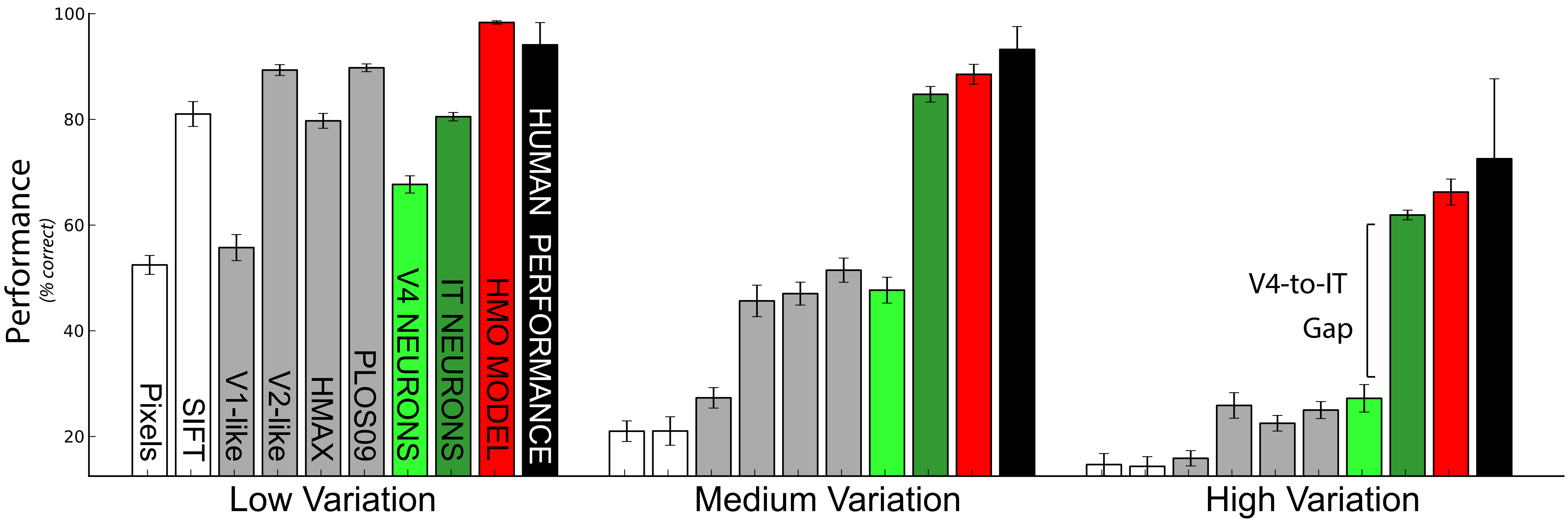


Yamins and Hong* et. al.* ***PNAS*** *(2014)*

At <u>high variation levels</u>, IT much better than V4 and existing models.

Performance Comparison

At high variation levels, IT much better than V4 and existing models

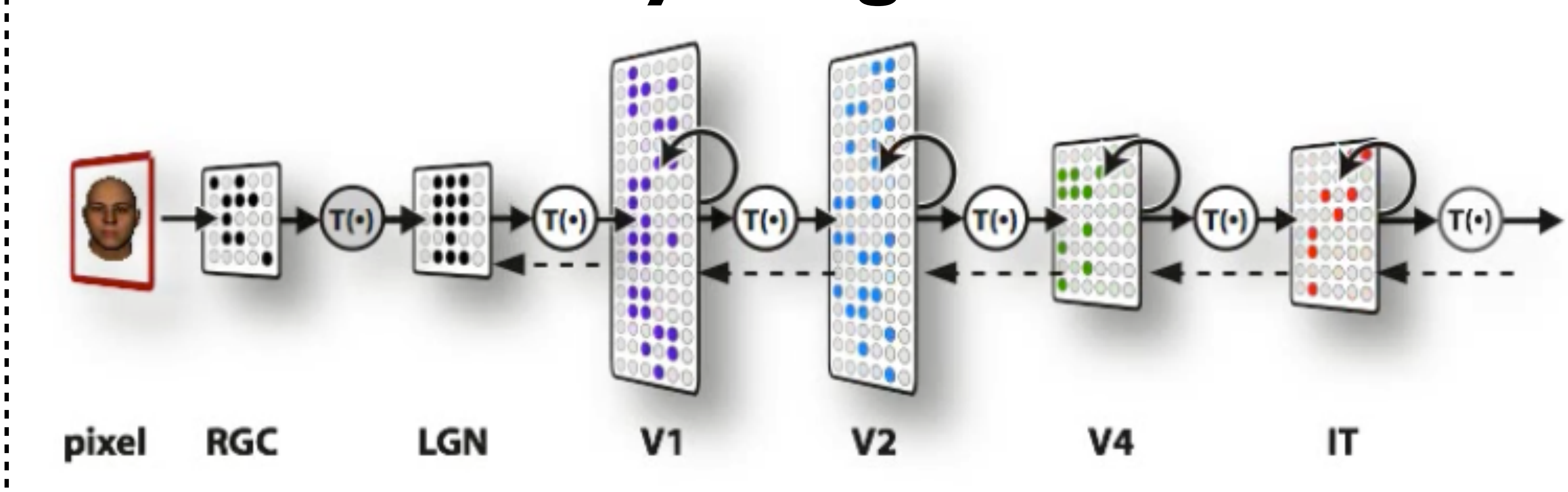


Yamins and Hong* et. al.* ***PNAS*** *(2014)*

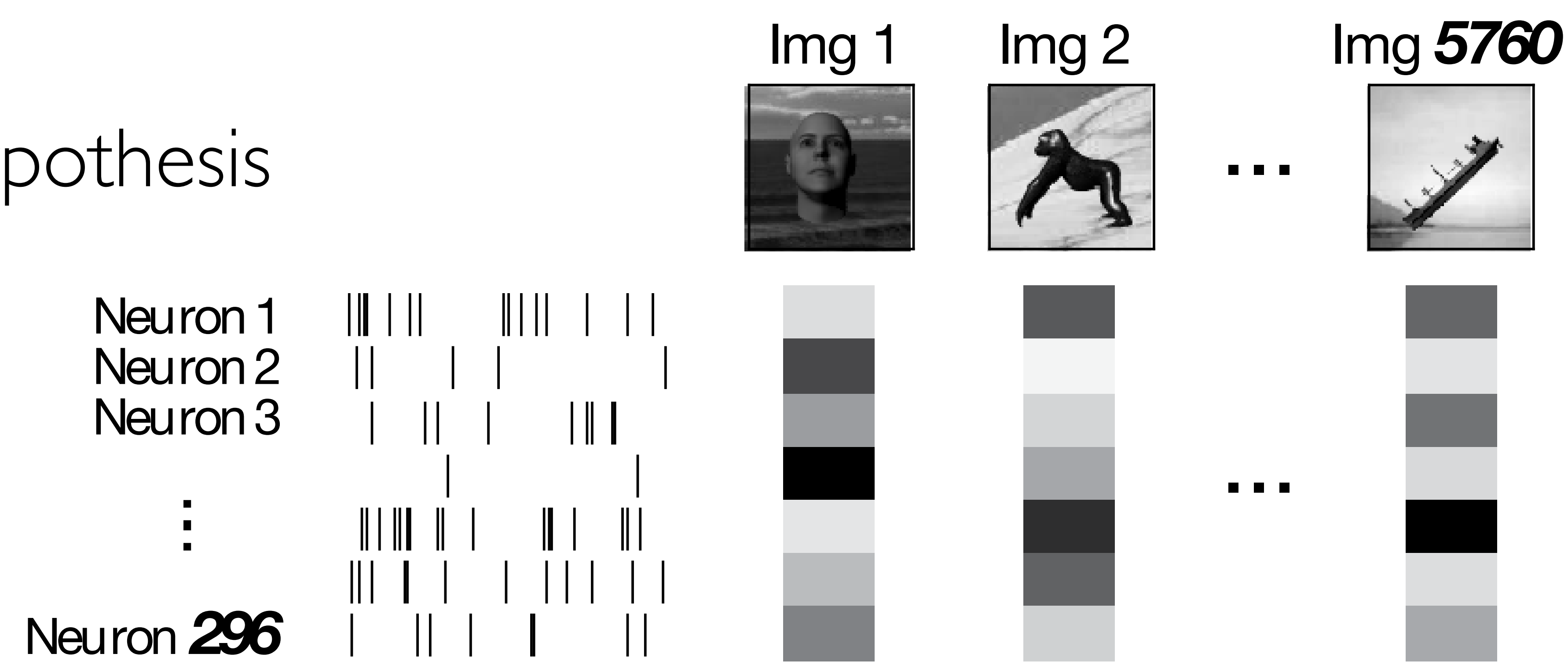
New model comparable to IT / human performance levels.

Ventral Stream as Stacked Cortex

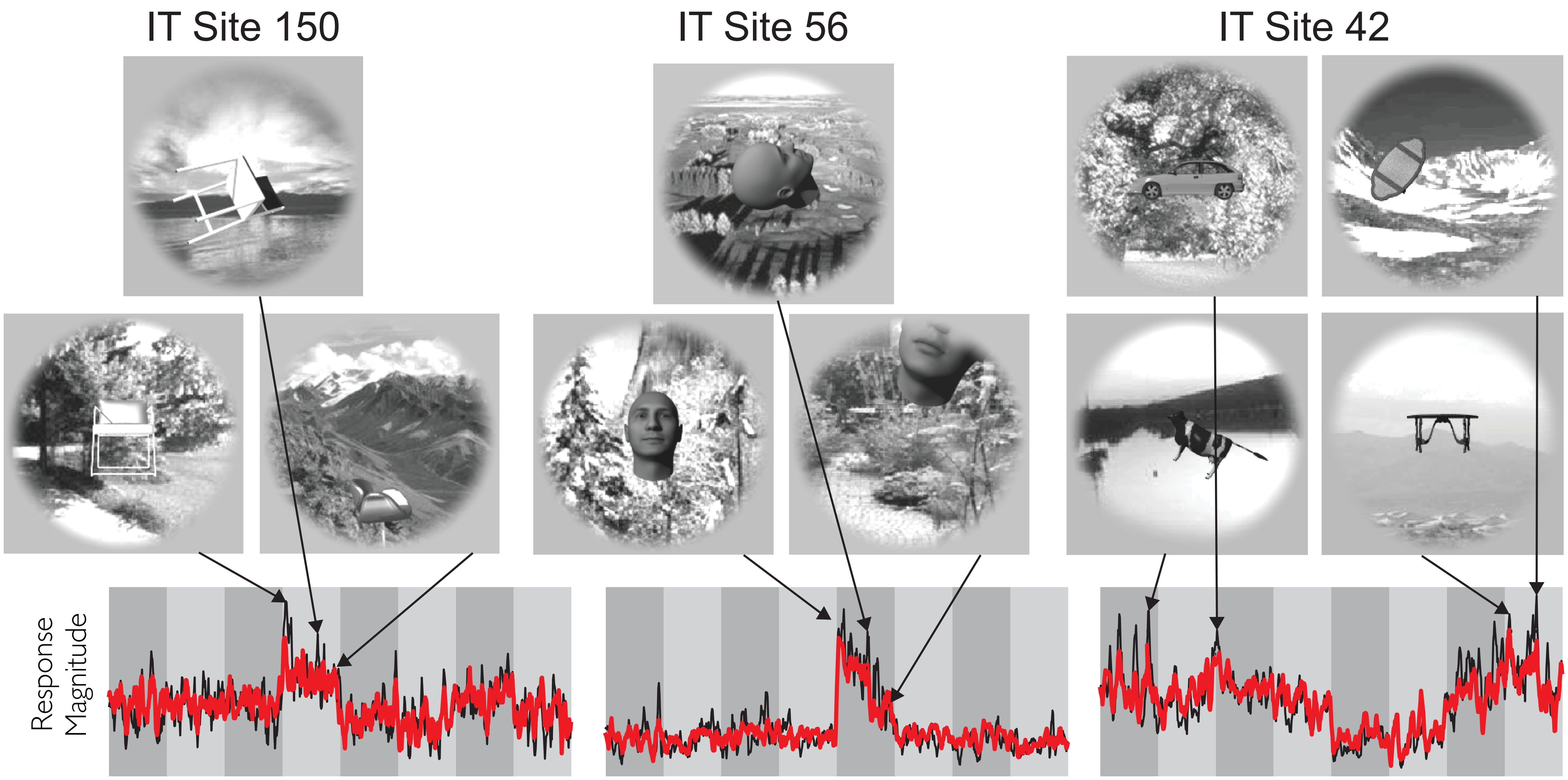
<u>GOAL: Predictive model of single-neuron responses throughout the ventral stream to arbitrary image stimuli.</u>



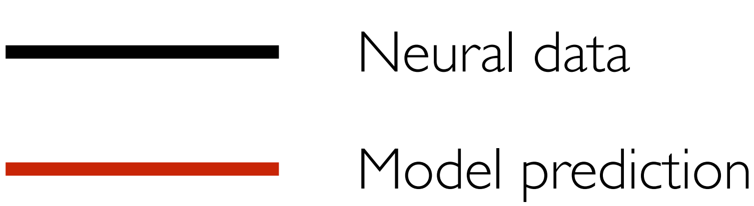
Get quantitative hypothesis for the network that generated this data —>



Predicting IT Neural Responses

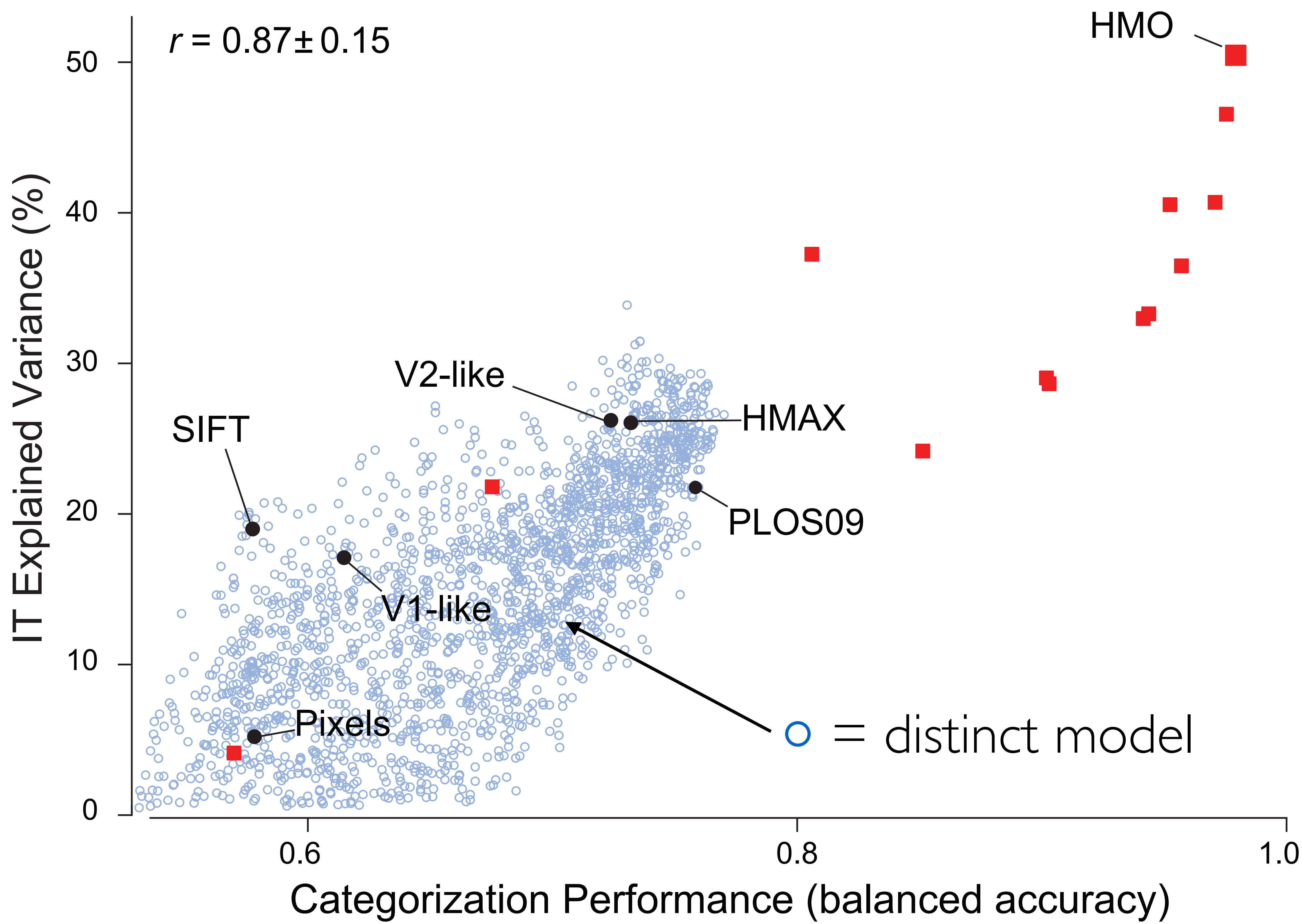


Images sorted first by **category**, then **variation level**.



Key Underlying Principle

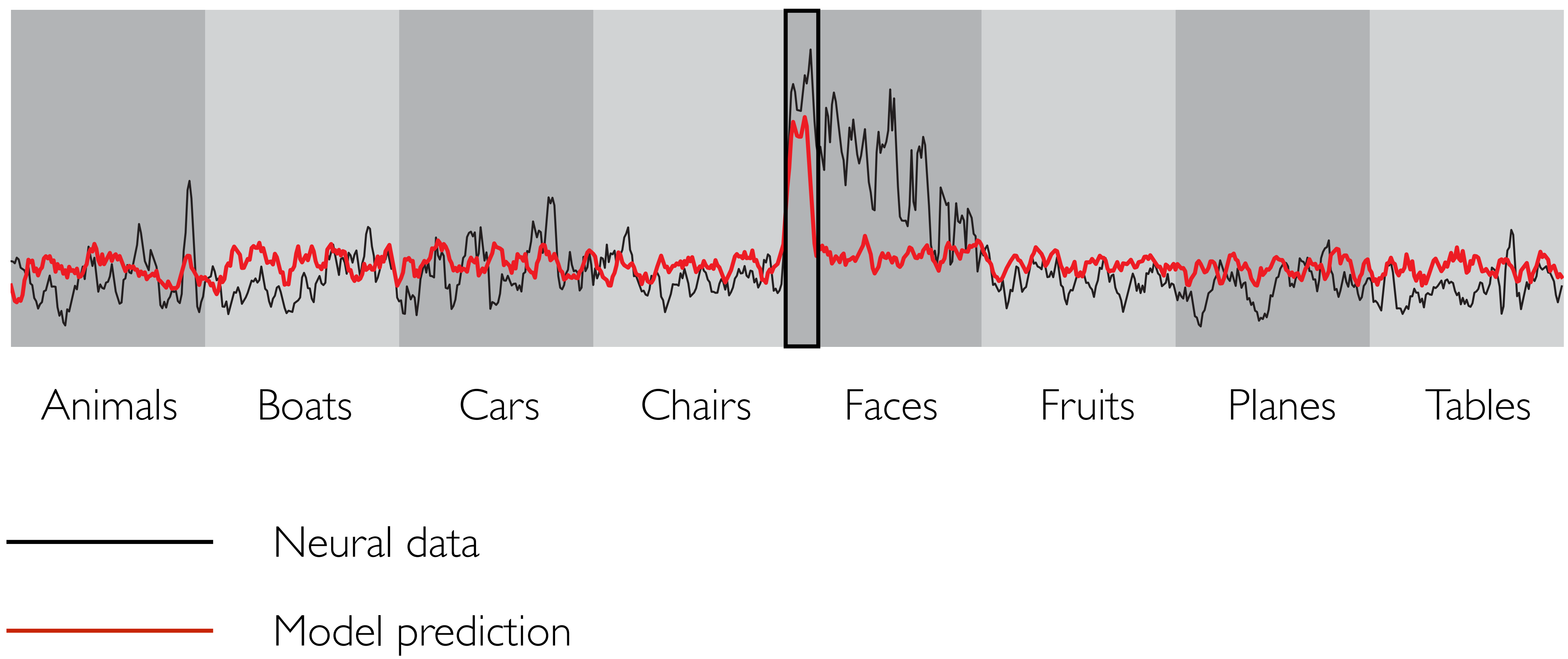
Yamins and Hong* et. al.* ***PNAS*** *(2014)*

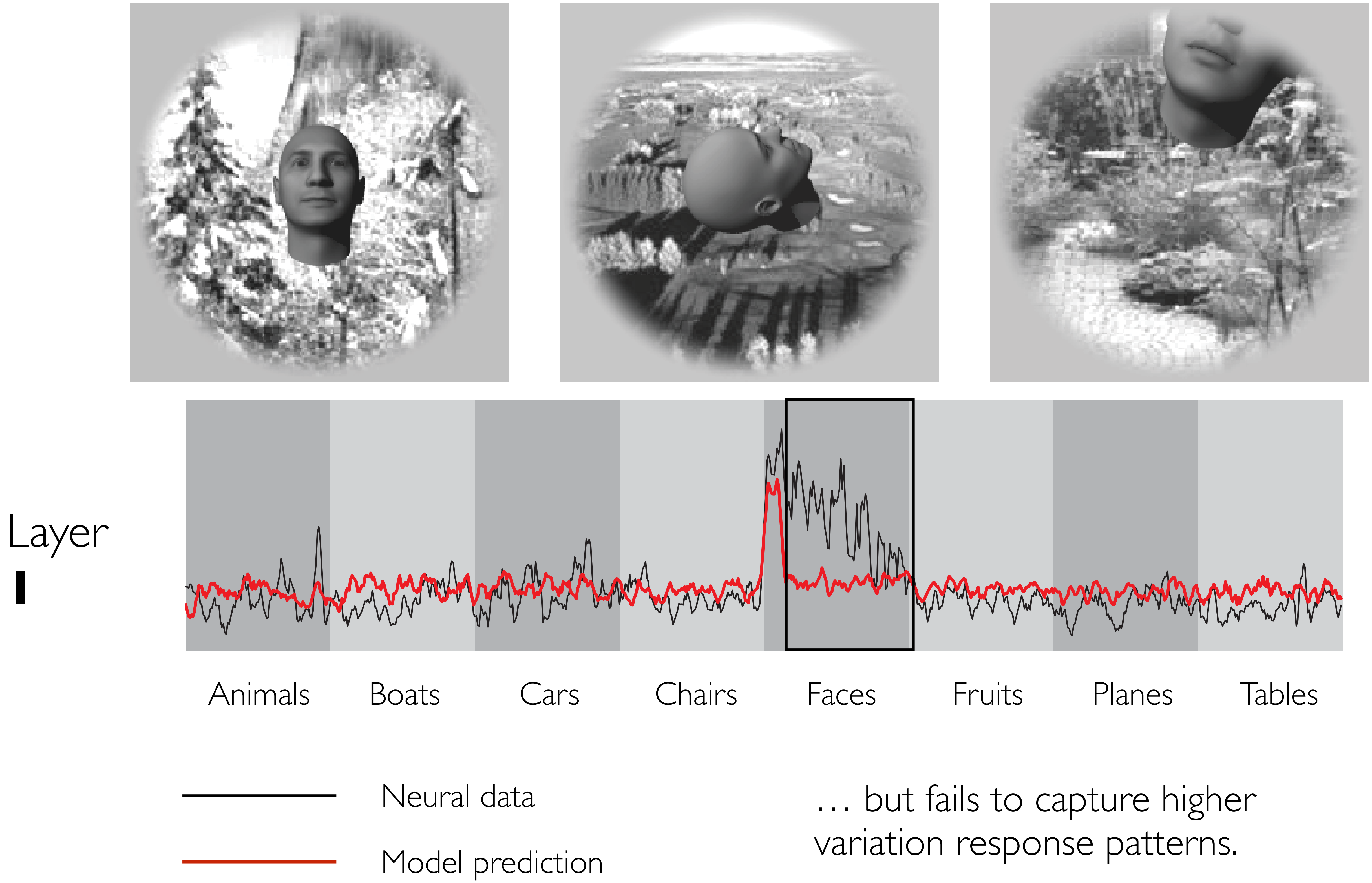


Captures low variation image response patterns …

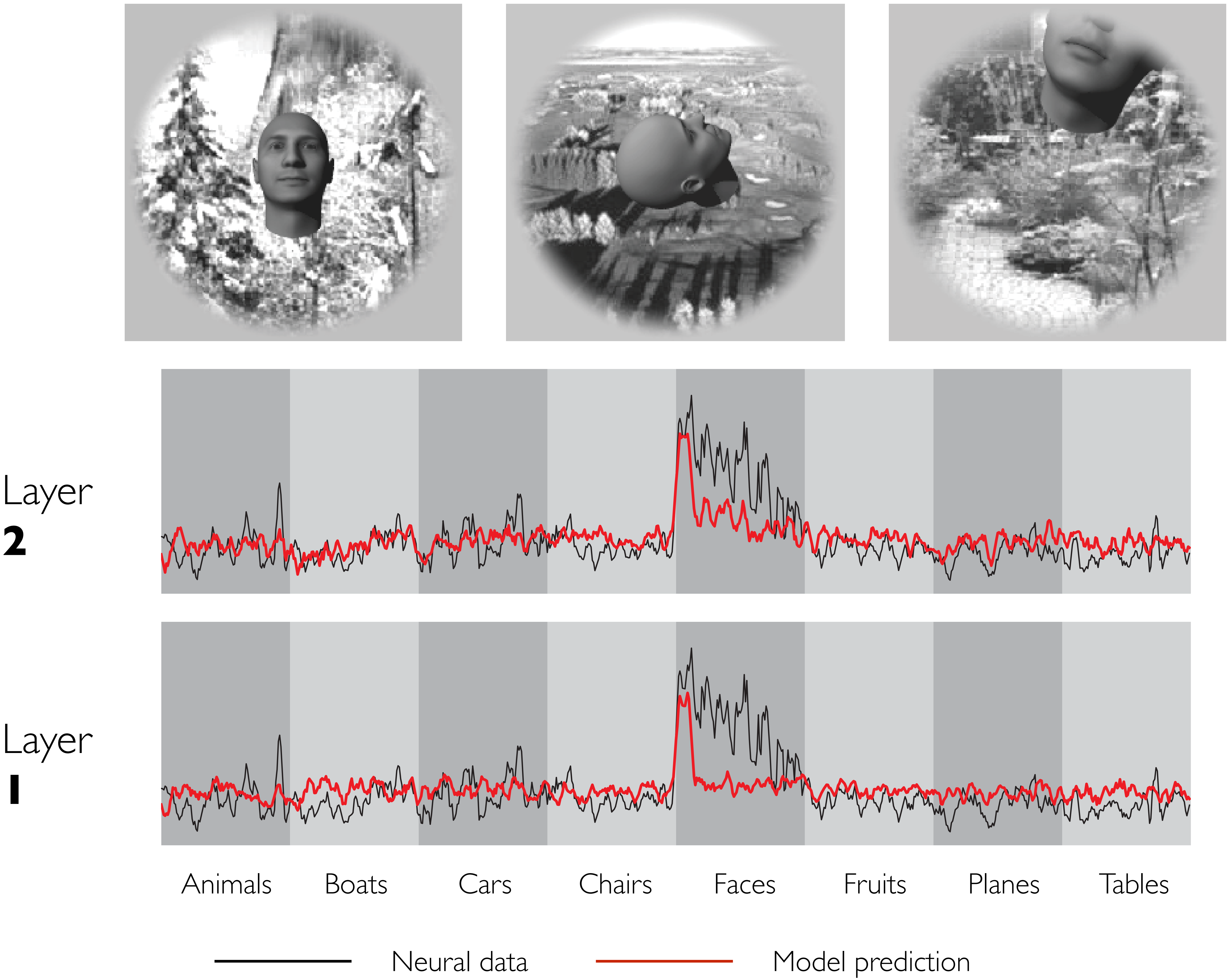
Layer

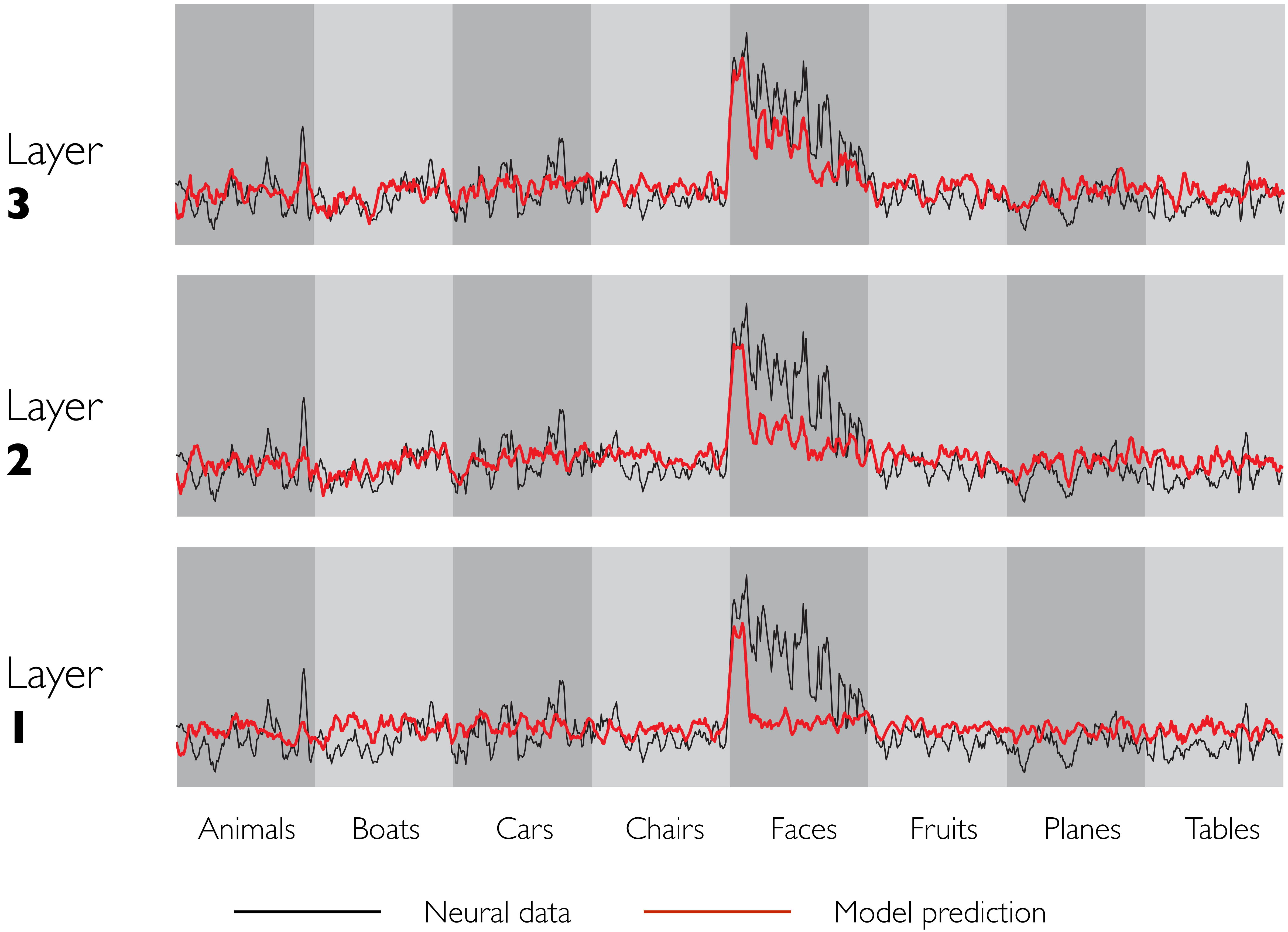
I



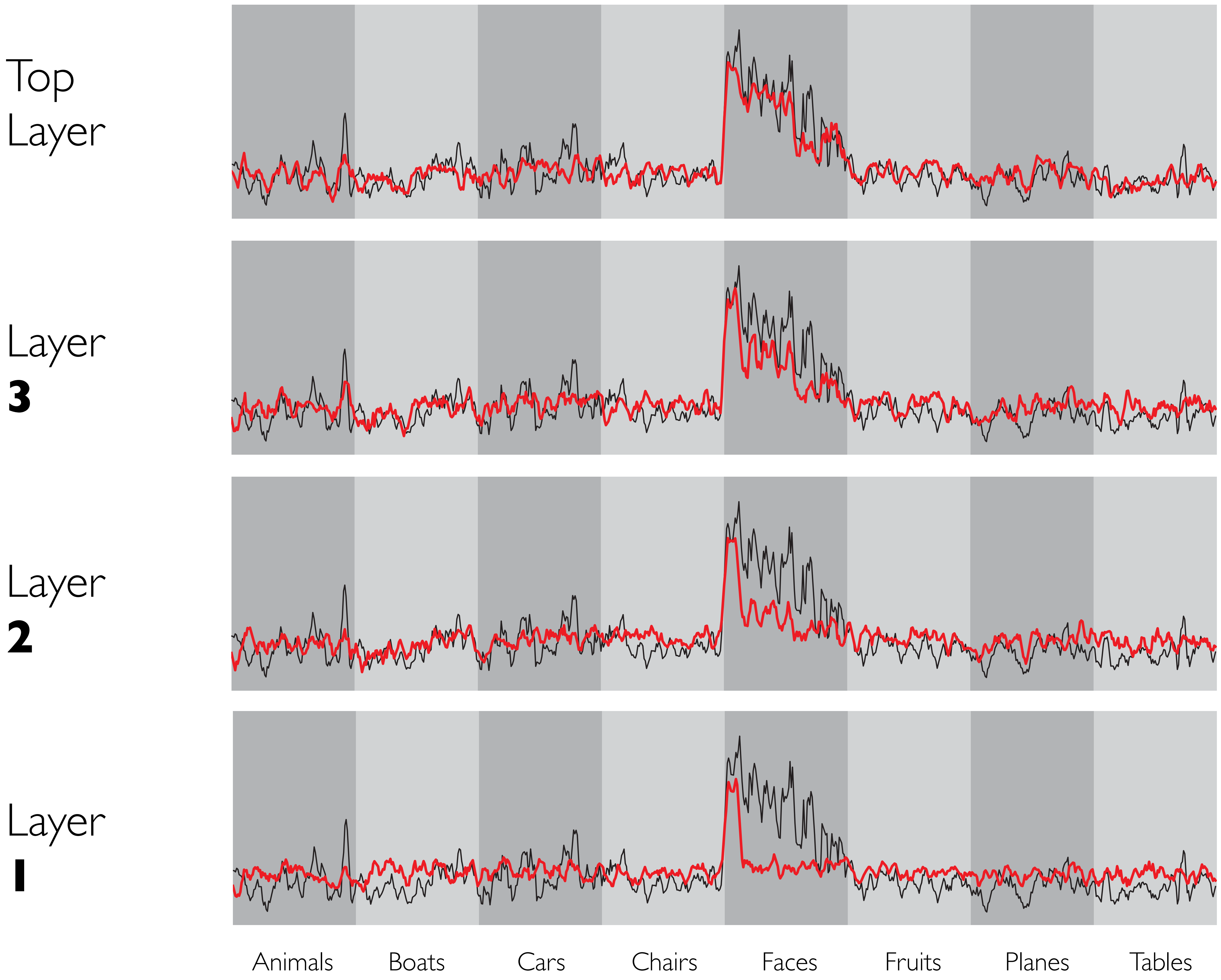


… but fails to capture higher variation response patterns.


Layer 2
Layer 1
Animals
Boats
Cars
Chairs
Faces
Fruits
Planes
Tables
Neural data
Model prediction

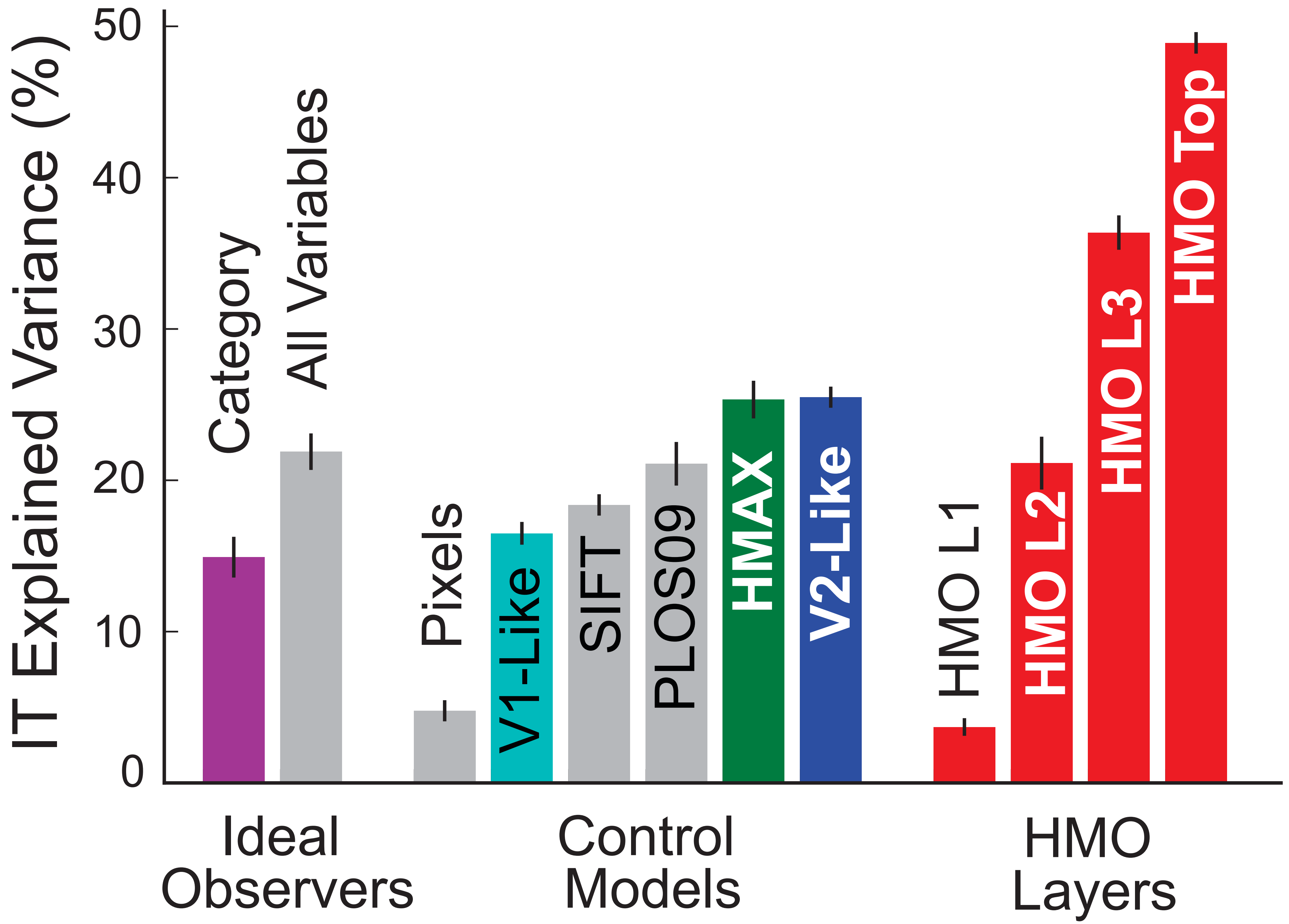

Layer
3
Layer
2
Layer
1
Animals
Boats
Cars
Chairs
Faces
Fruits
Planes
Tables
Neural data
Model prediction

Building tolerance while maintaining selectivity



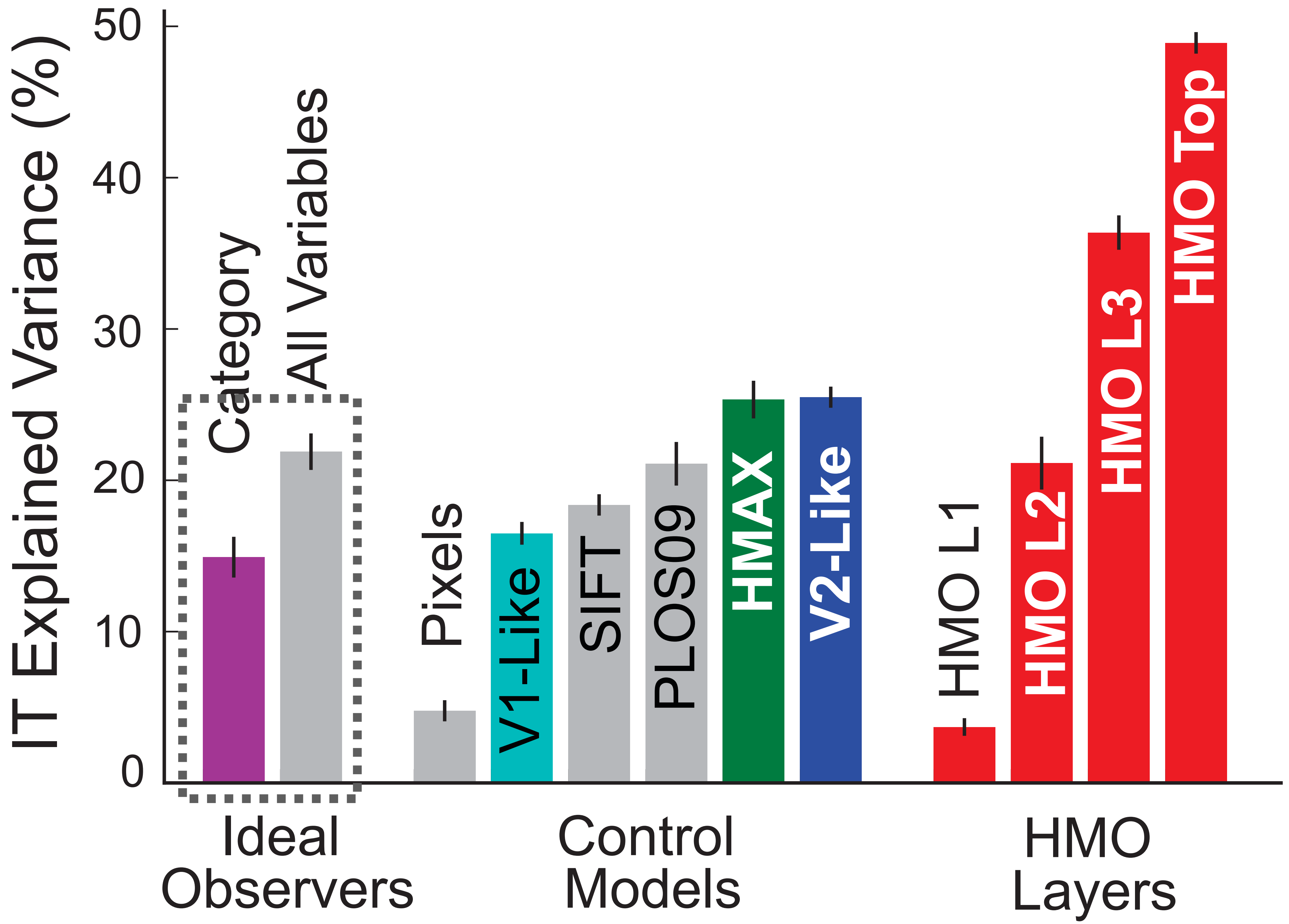
Predicting IT Neural Responses

Yamins and Hong* et. al.* ***PNAS*** *(2014)*



Predicting IT Neural Responses

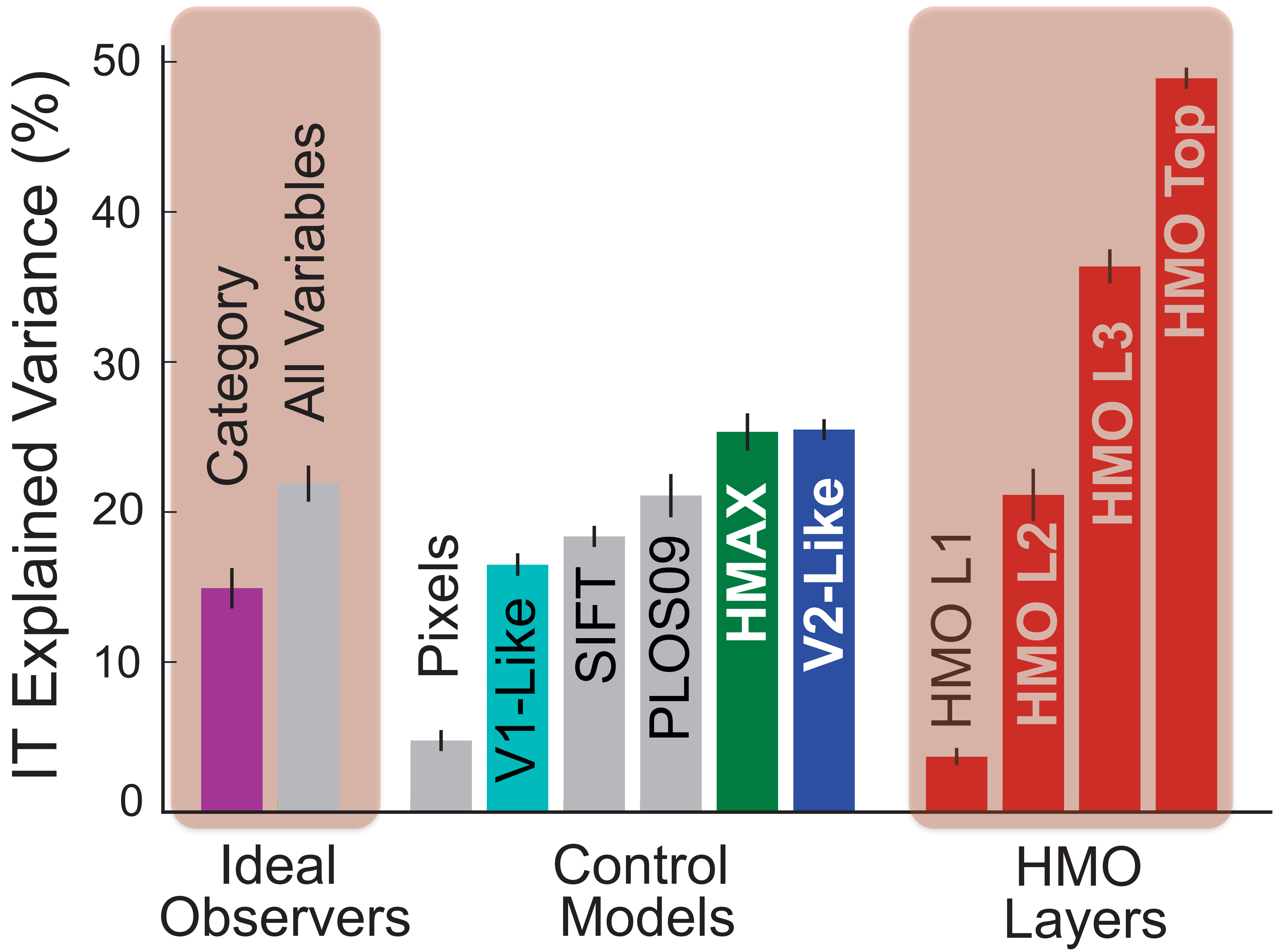
Yamins and Hong* et. al.* ***PNAS*** *(2014)*



Predicting IT Neural Responses

Yamins and Hong* et. al.* ***PNAS*** *(2014)*

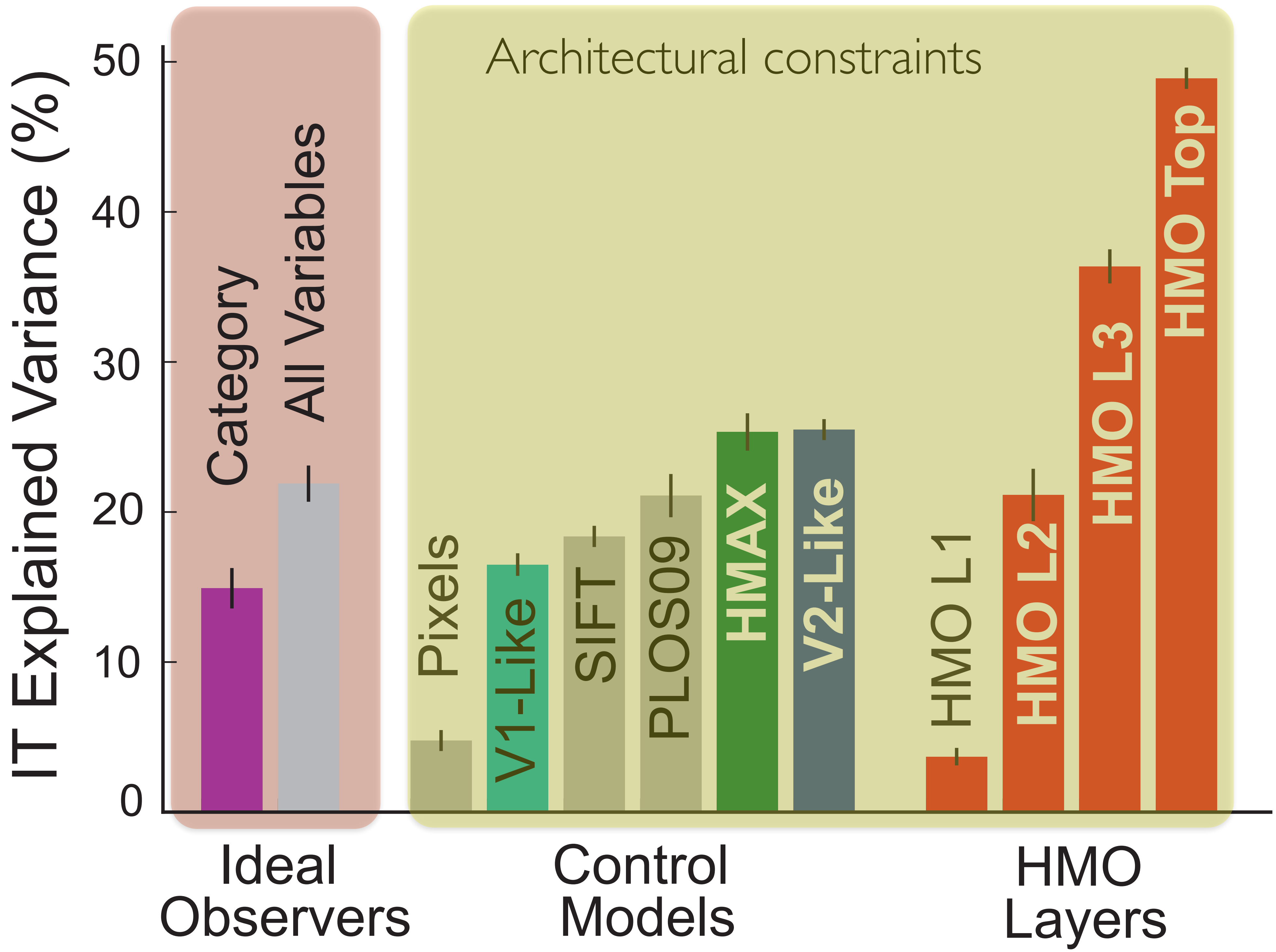
Performance constraints



Predicting IT Neural Responses

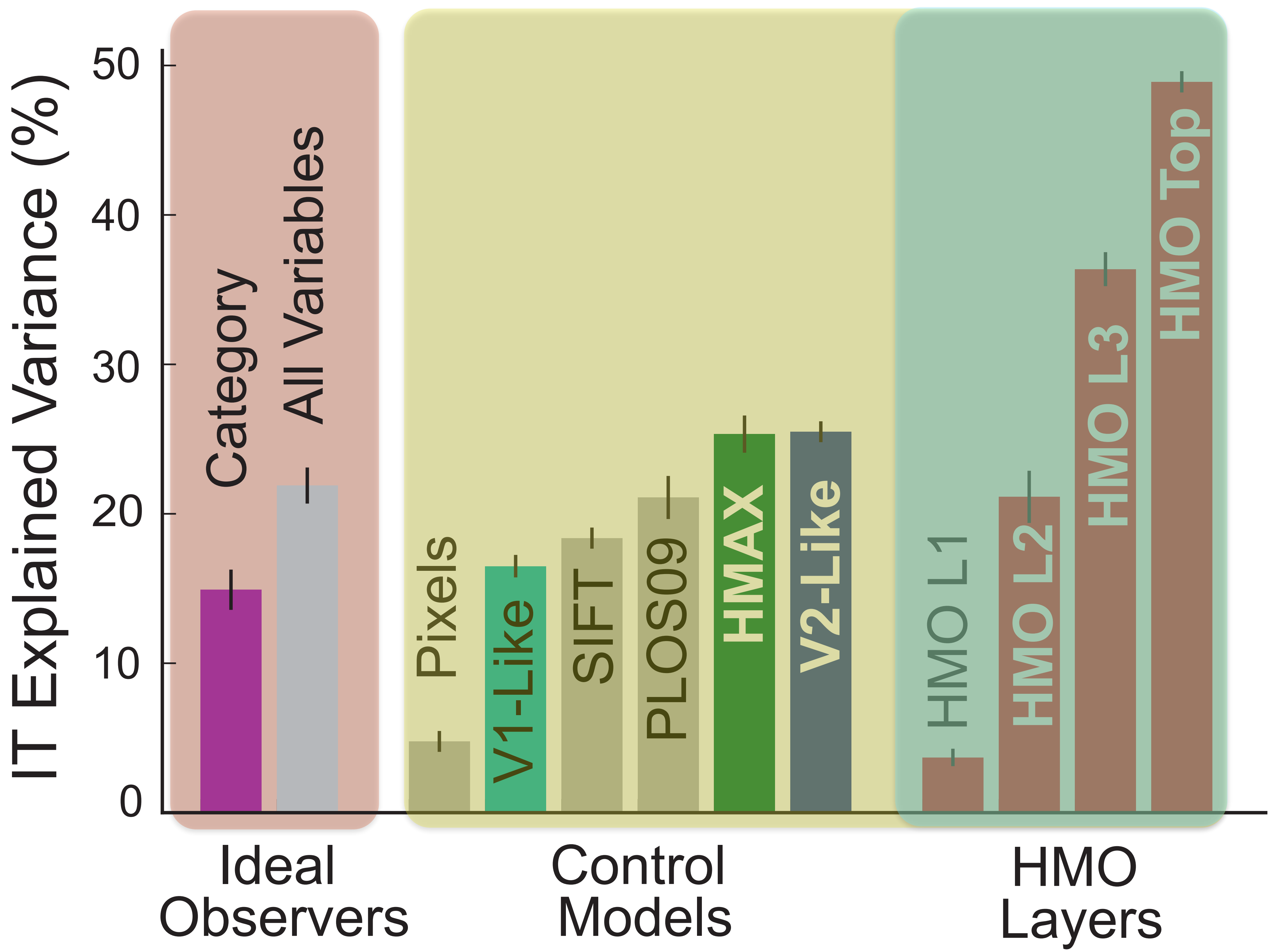
Yamins and Hong* et. al.* ***PNAS*** *(2014)*

Performance constraints



Predicting IT Neural Responses

Performance constraints + architectural constraints → better neural prediction



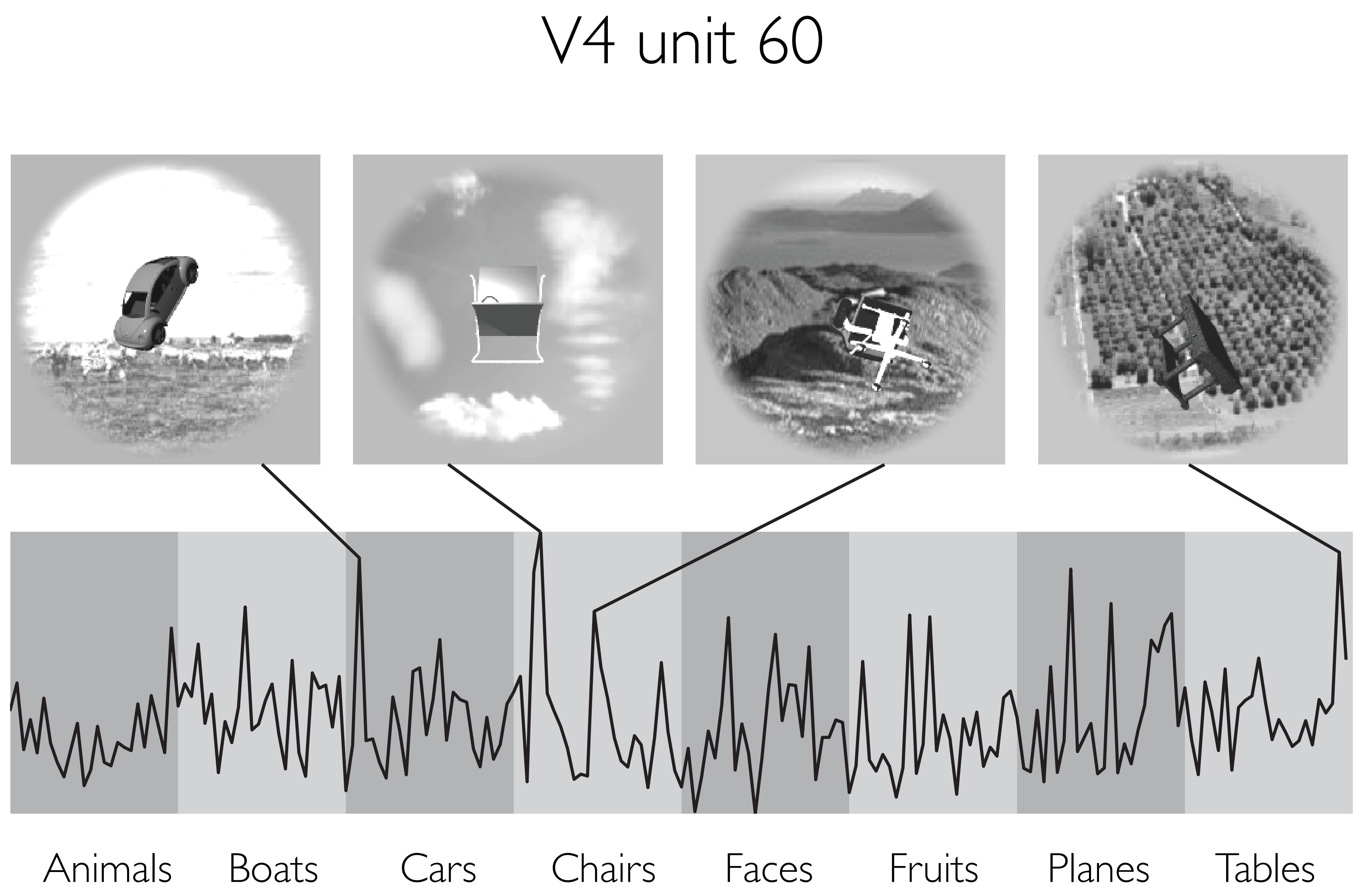
Predicting IT Neural Responses

What about intermediate layers?

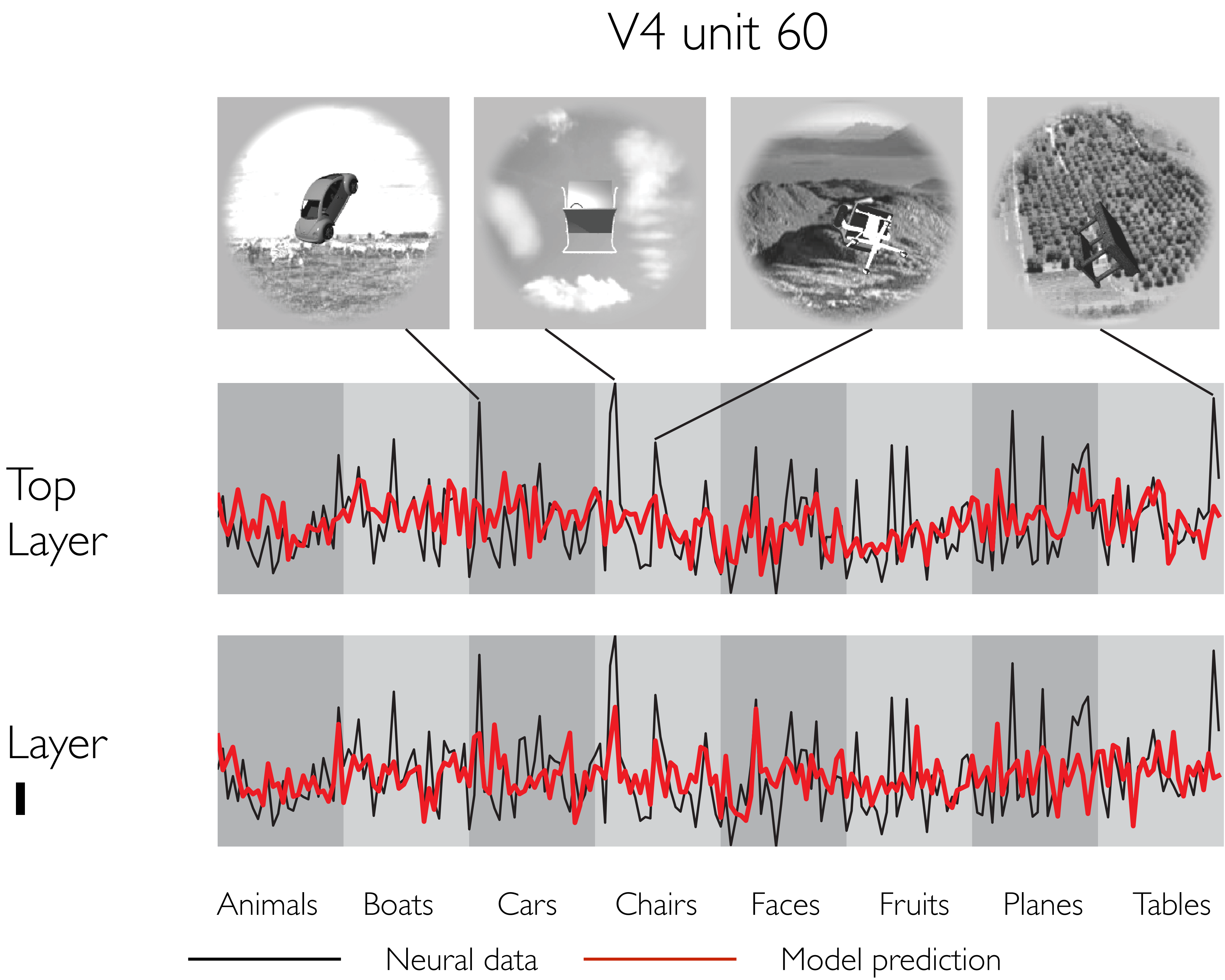
i. compare intermediate model layers to IT neural data

ii. compare all model layers to intermediate visual areas (V4)

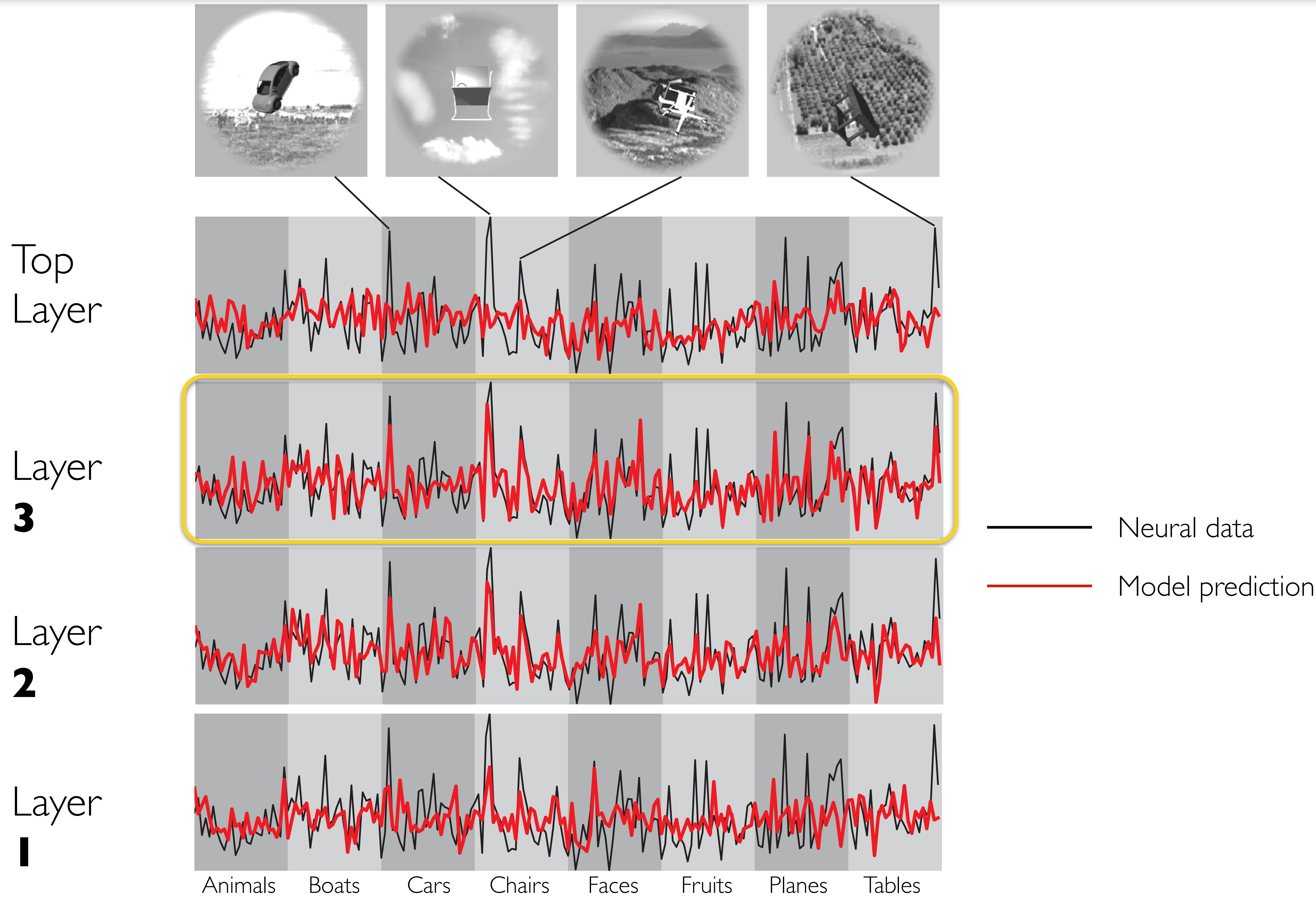
Predicting V4 Neural Responses



Predicting V4 Neural Responses

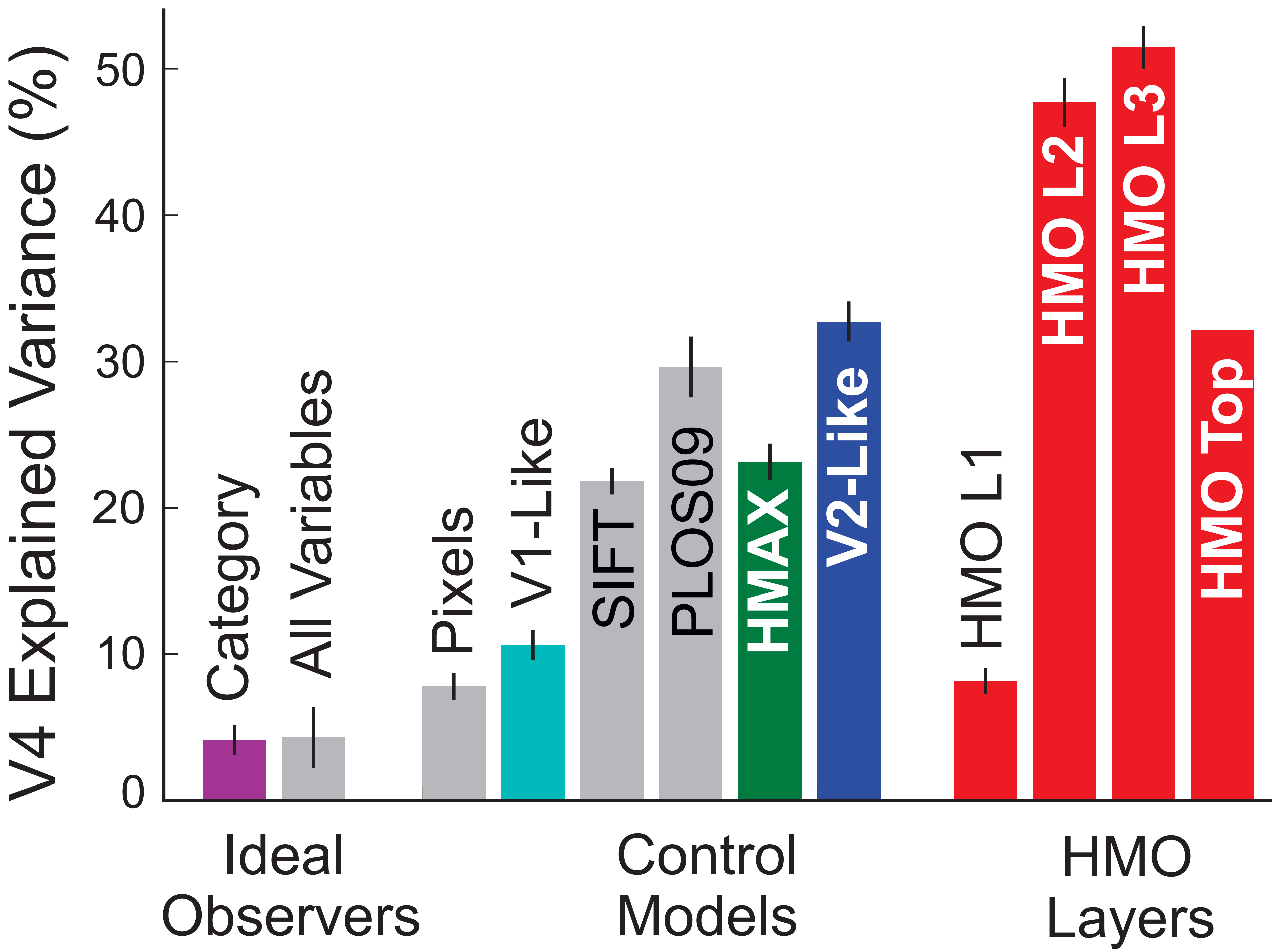


Predicting V4 Neural Responses



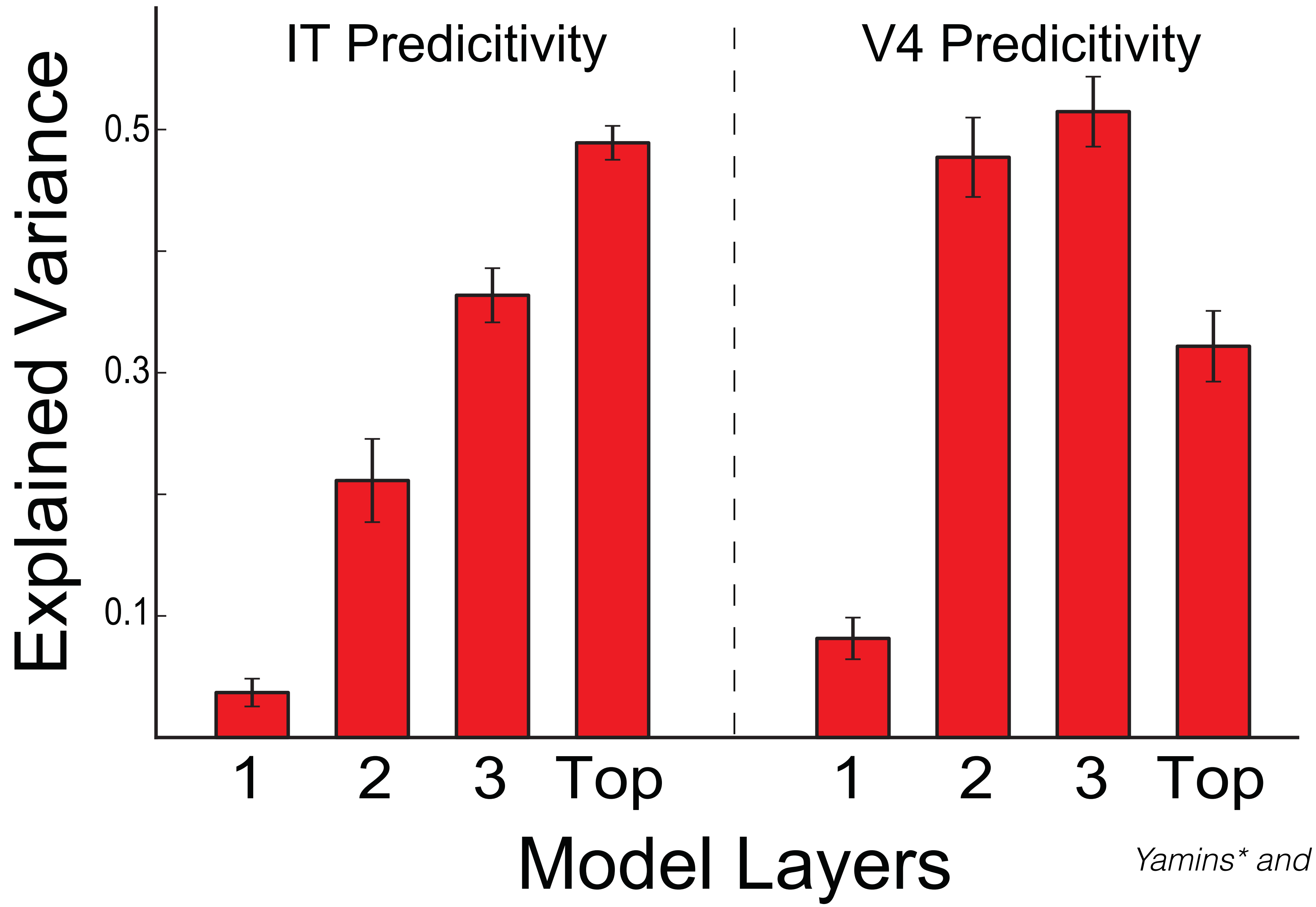
Predicting V4 Neural Responses

Yamins and Hong* et. al.* ***PNAS*** *(2014)*



Layer-area correspondence

Investigating fits as a function of model layer:

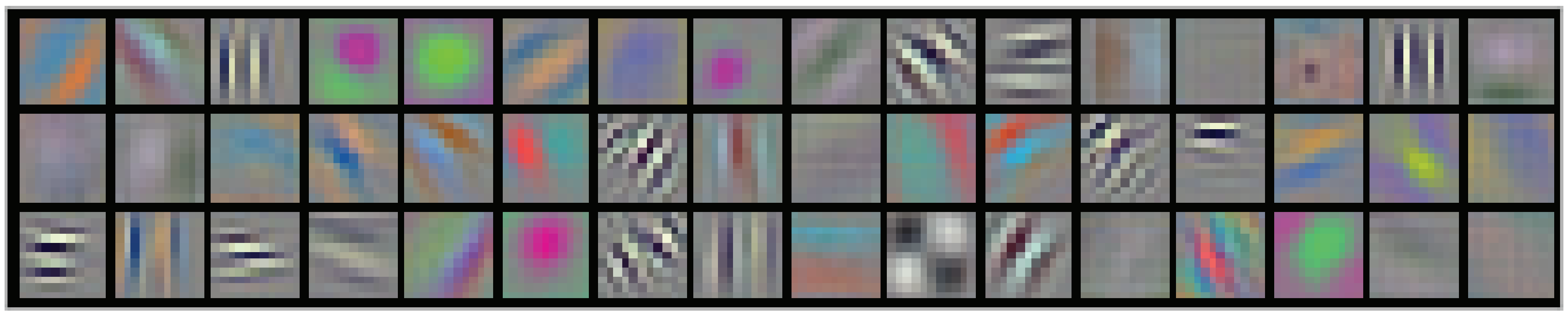


Yamins and Hong* et. al.* ***PNAS*** *(2014)*

IT fit increases at each layer. In contrast, V4 fit peaks and then goes down.

Layer-area correspondence

Model output at lowest layer resembles Gabor wavelets:

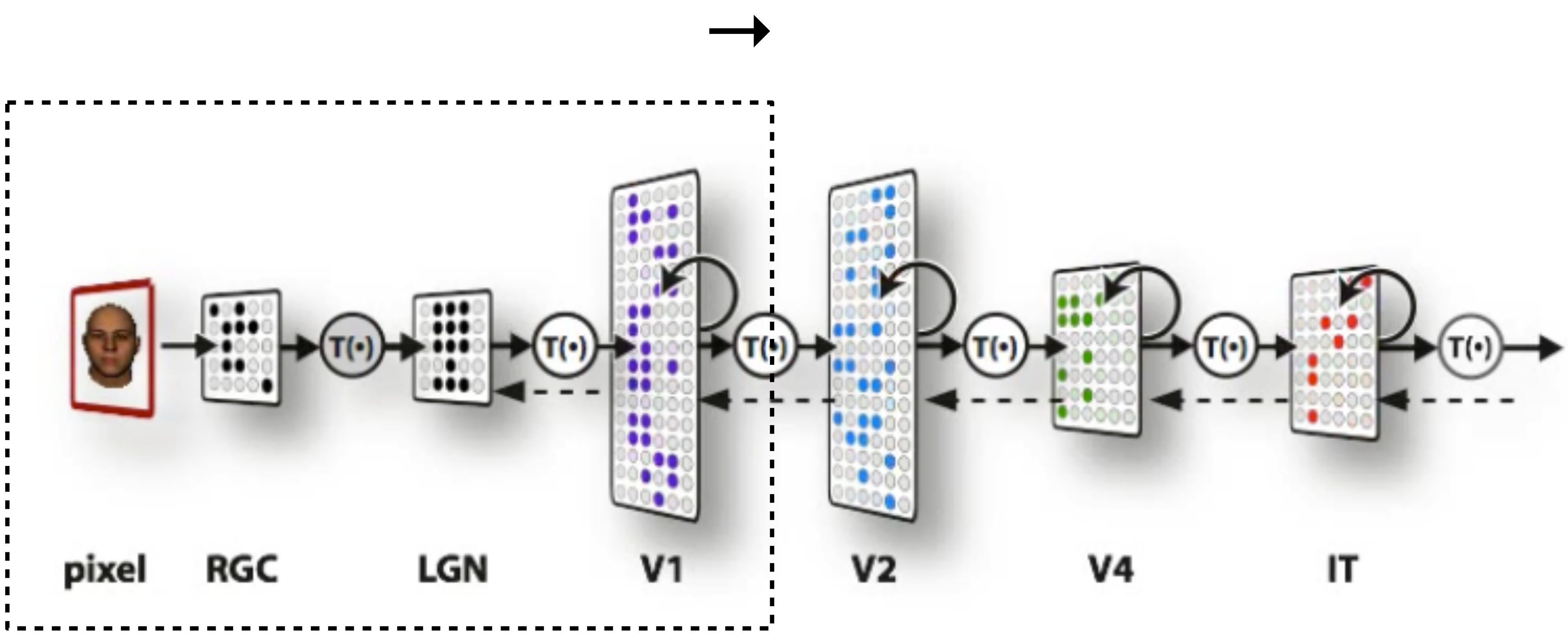


Layer 1 Filters

In submission: model lowest layer is best explanation of imaging data in V1. *(with Darren Seibert and Justin Gardner)*

Behavioral "Top-Down" constraints

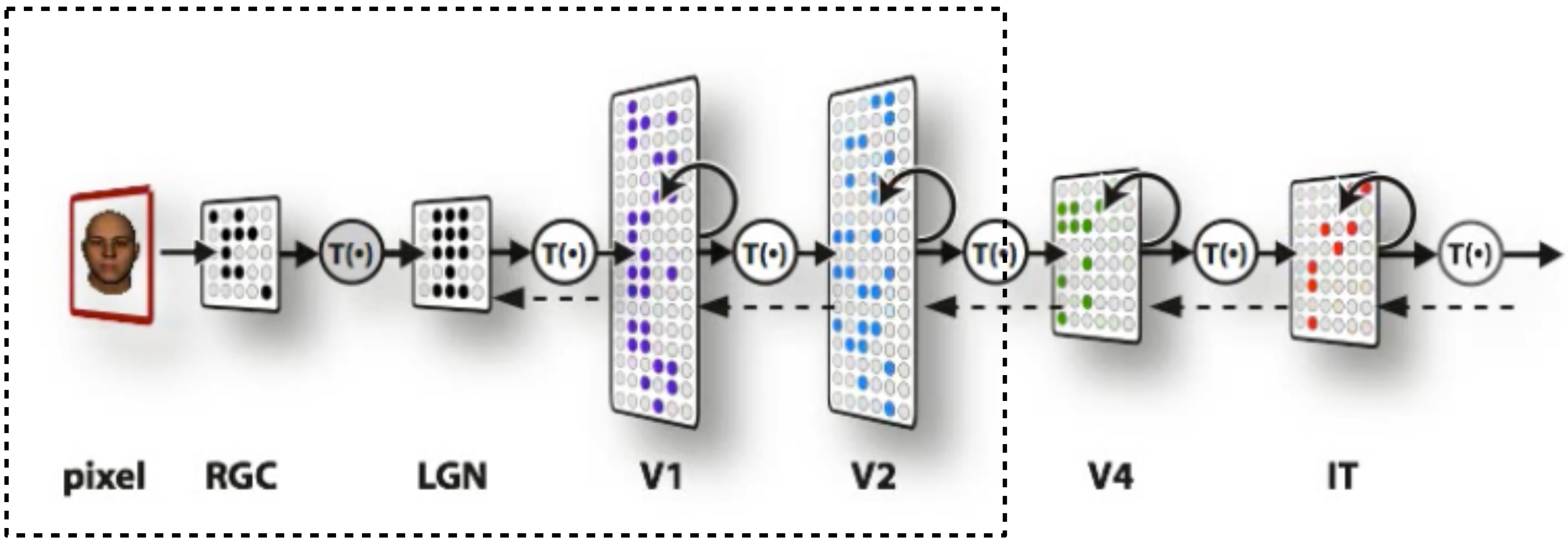
Complement standard "from below" approach …



Behavioral "Top-Down" constraints

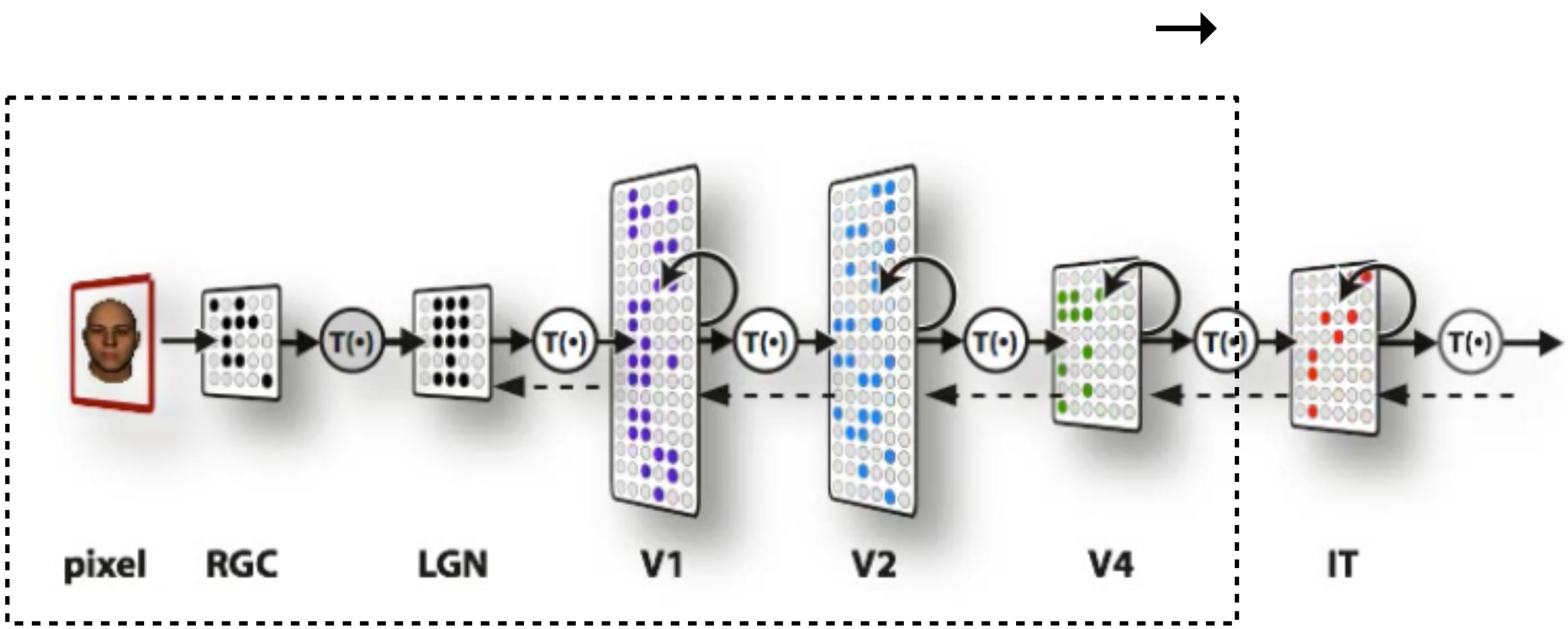
Complement standard "from below" approach …

→



Behavioral "Top-Down" constraints

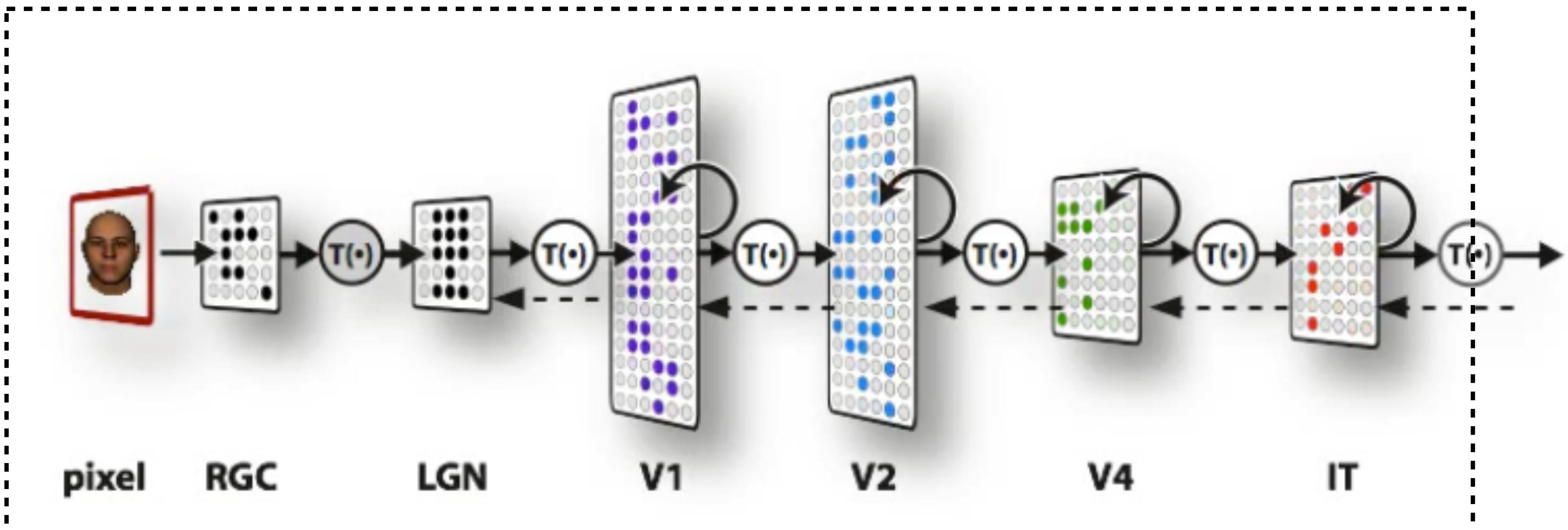
Complement standard "from below" approach …



Behavioral "Top-Down" constraints

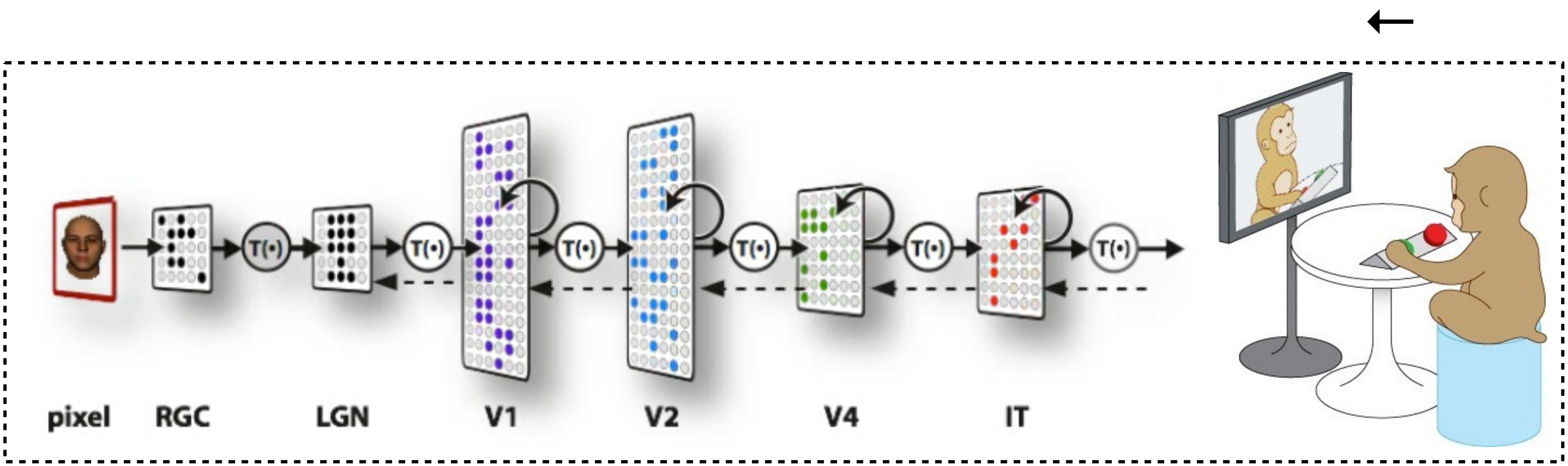
Complement standard "from below" approach …

→



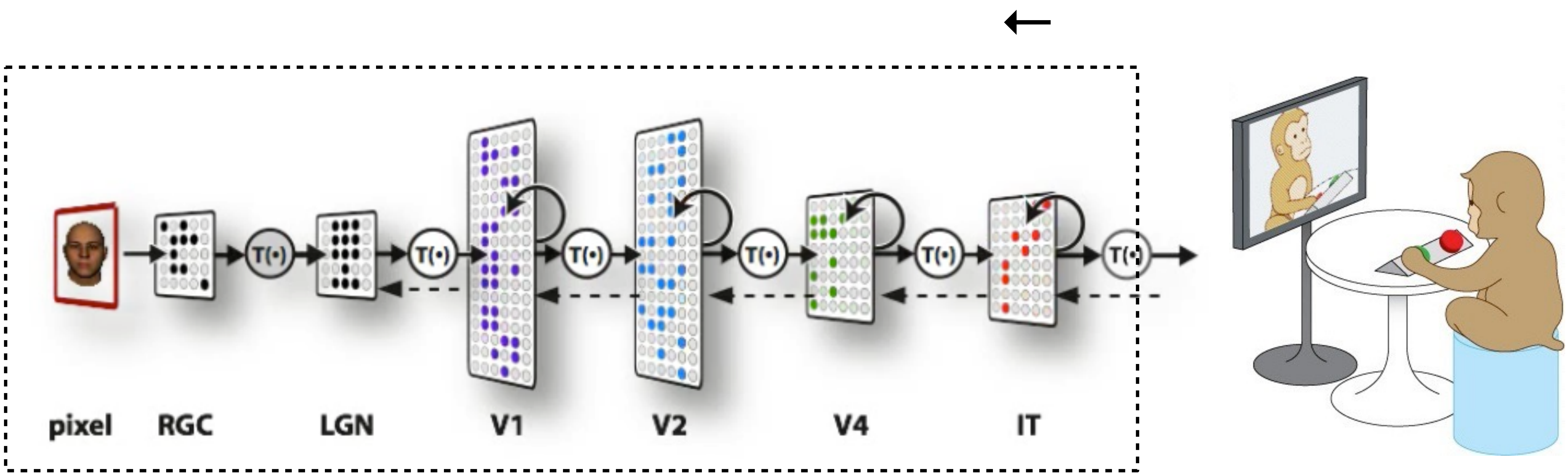
Behavioral "Top-Down" constraints

Complement standard "from below" approach … with behavioral constraints



Behavioral "Top-Down" constraints

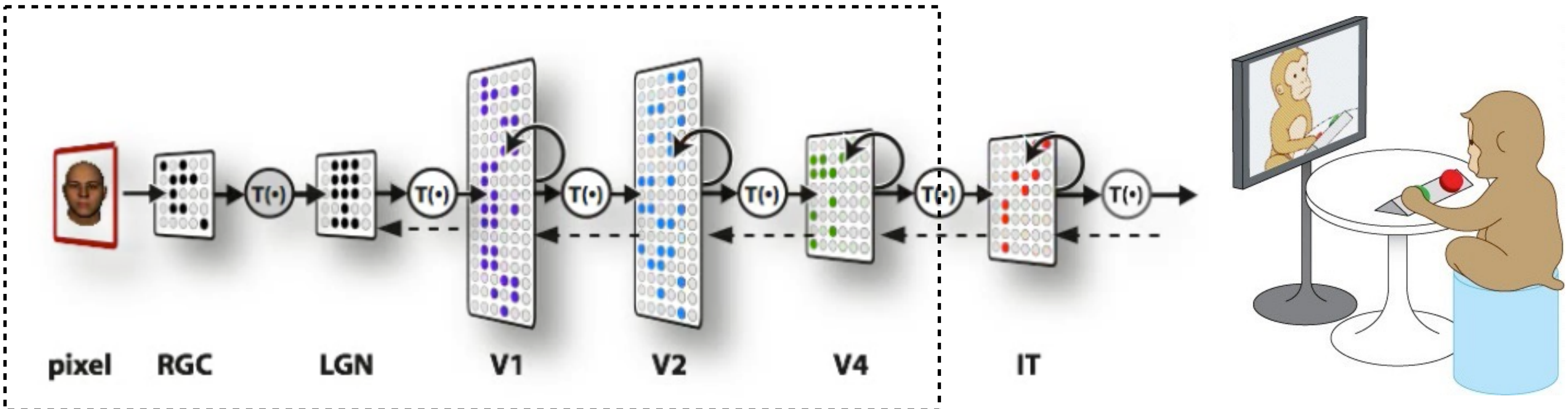
Complement standard "from below" approach … with behavioral constraints



Behavioral "Top-Down" constraints

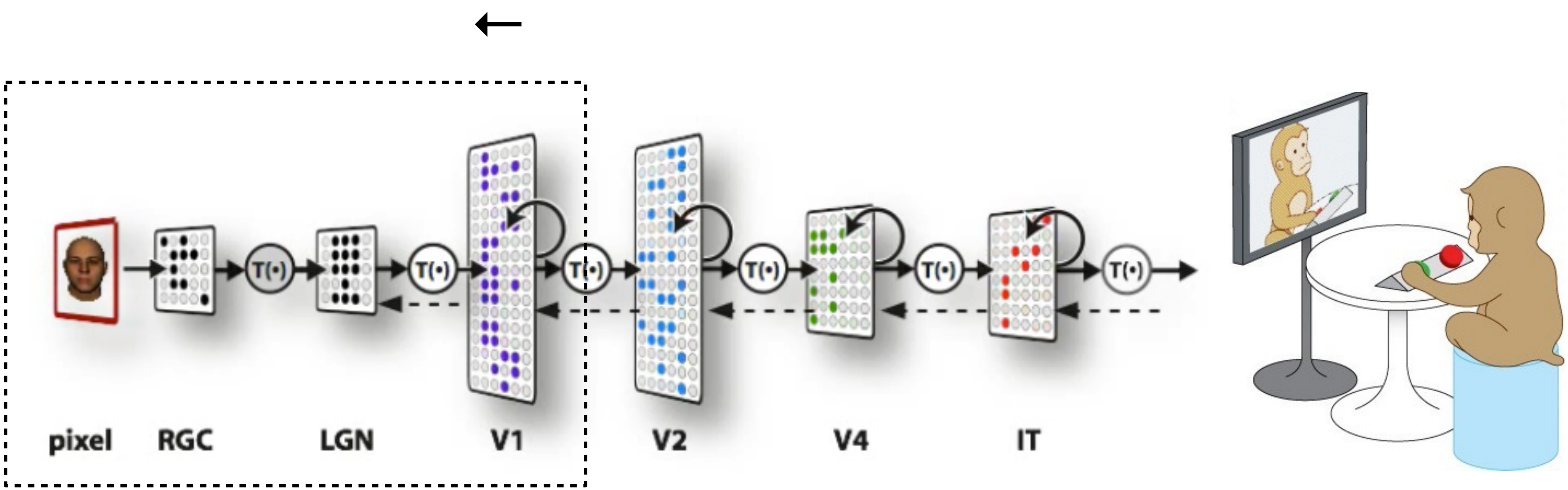
Complement standard "from below" approach … with behavioral constraints

←



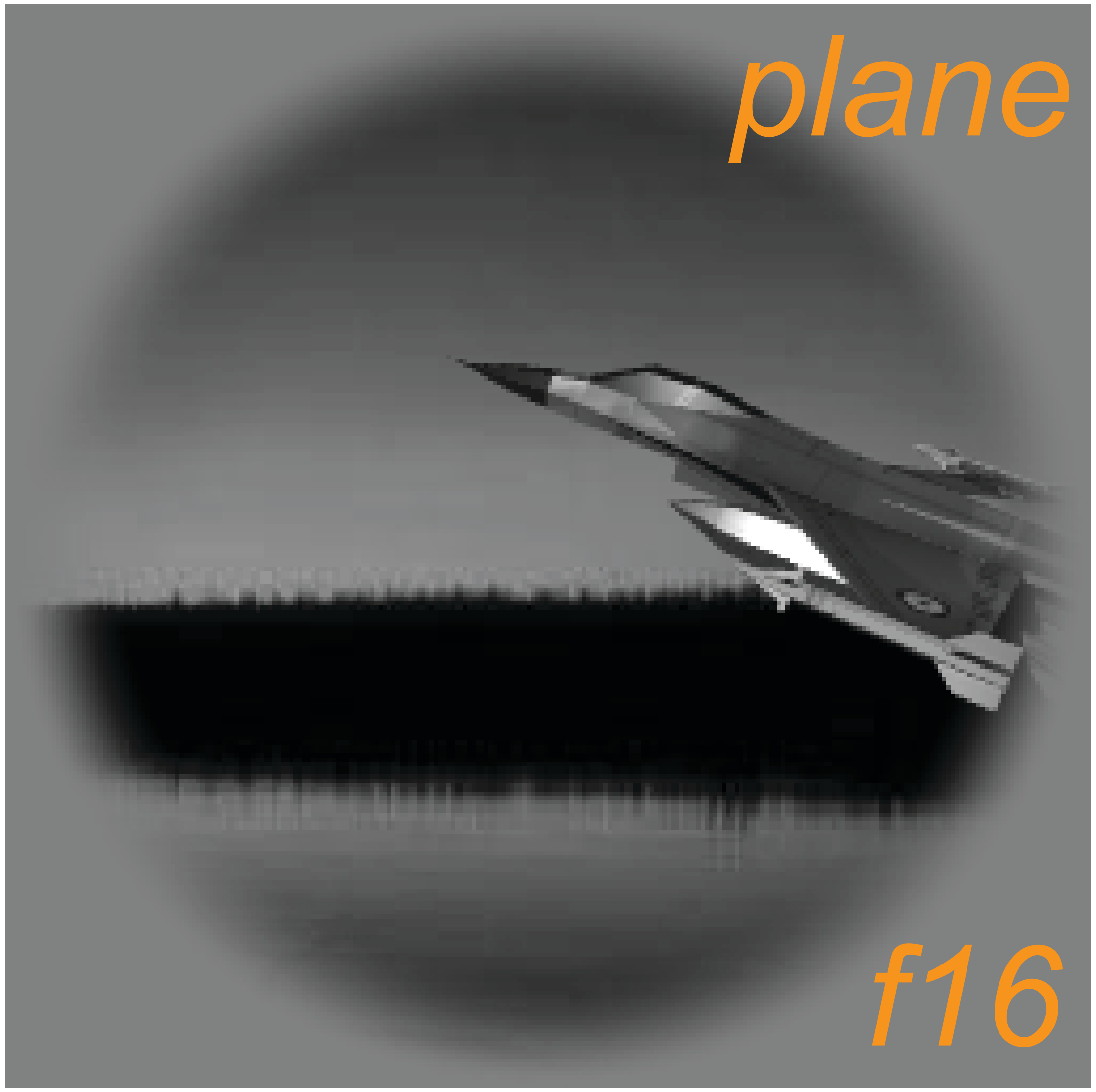
Behavioral "Top-Down" constraints

Complement standard "from below" approach … with behavioral constraints



Similar ideas and results:
Khaligh-Razavi & Kriegeskorte (2014),
Guclu & Van Gerven (2015), Cichy & Oliva (2015)

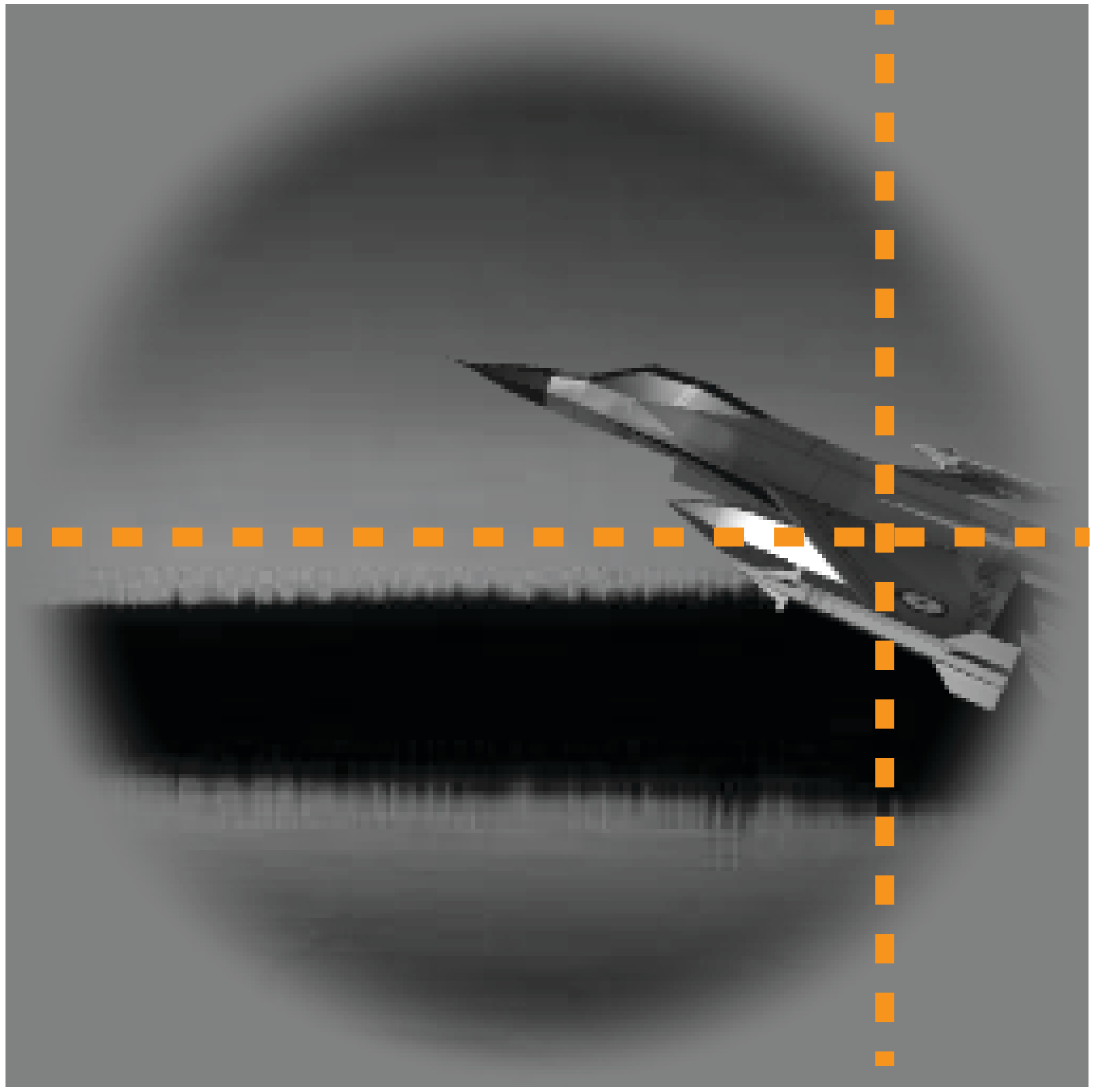
Beyond categorization



Category

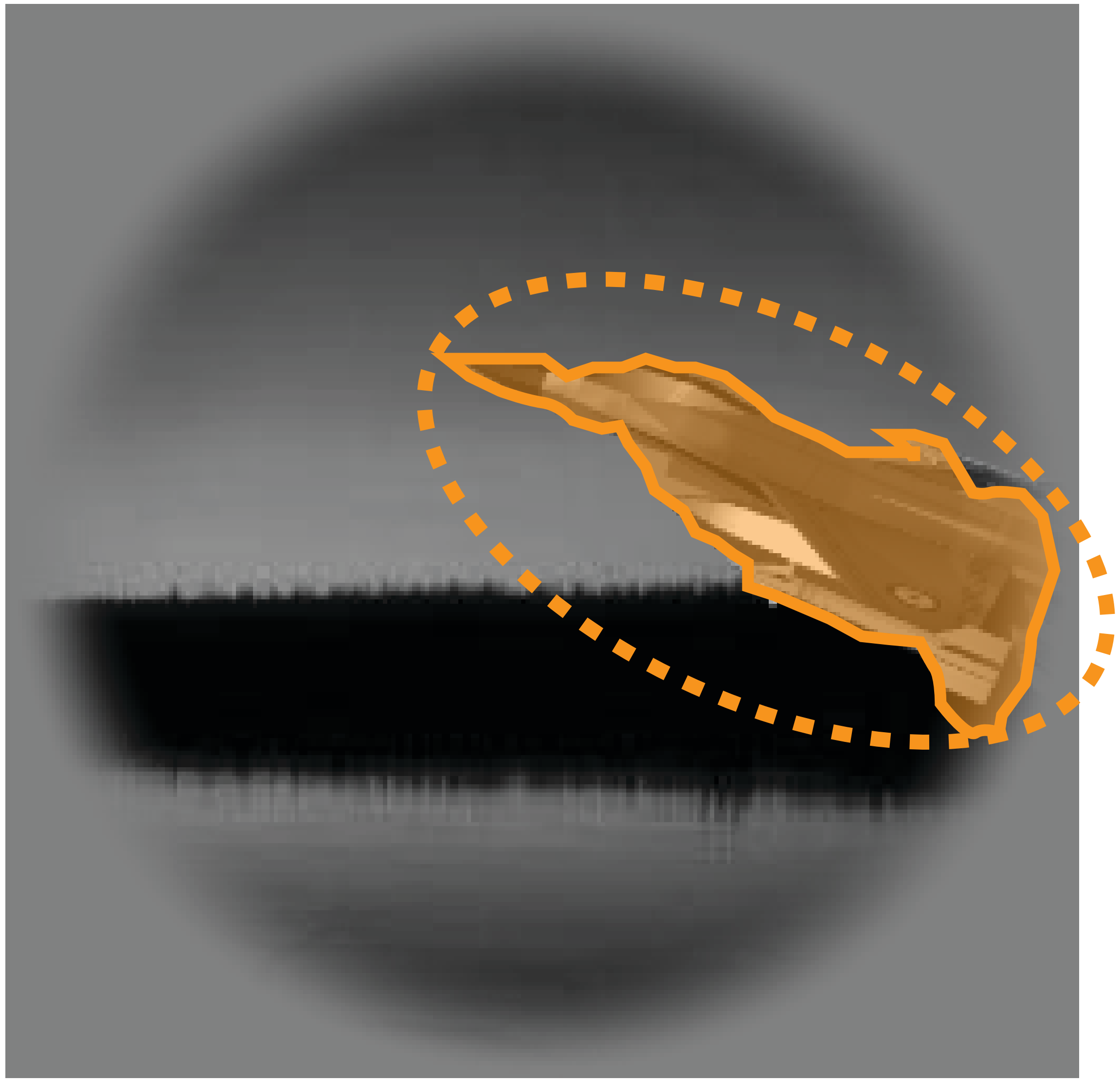
Identity

Beyond categorization



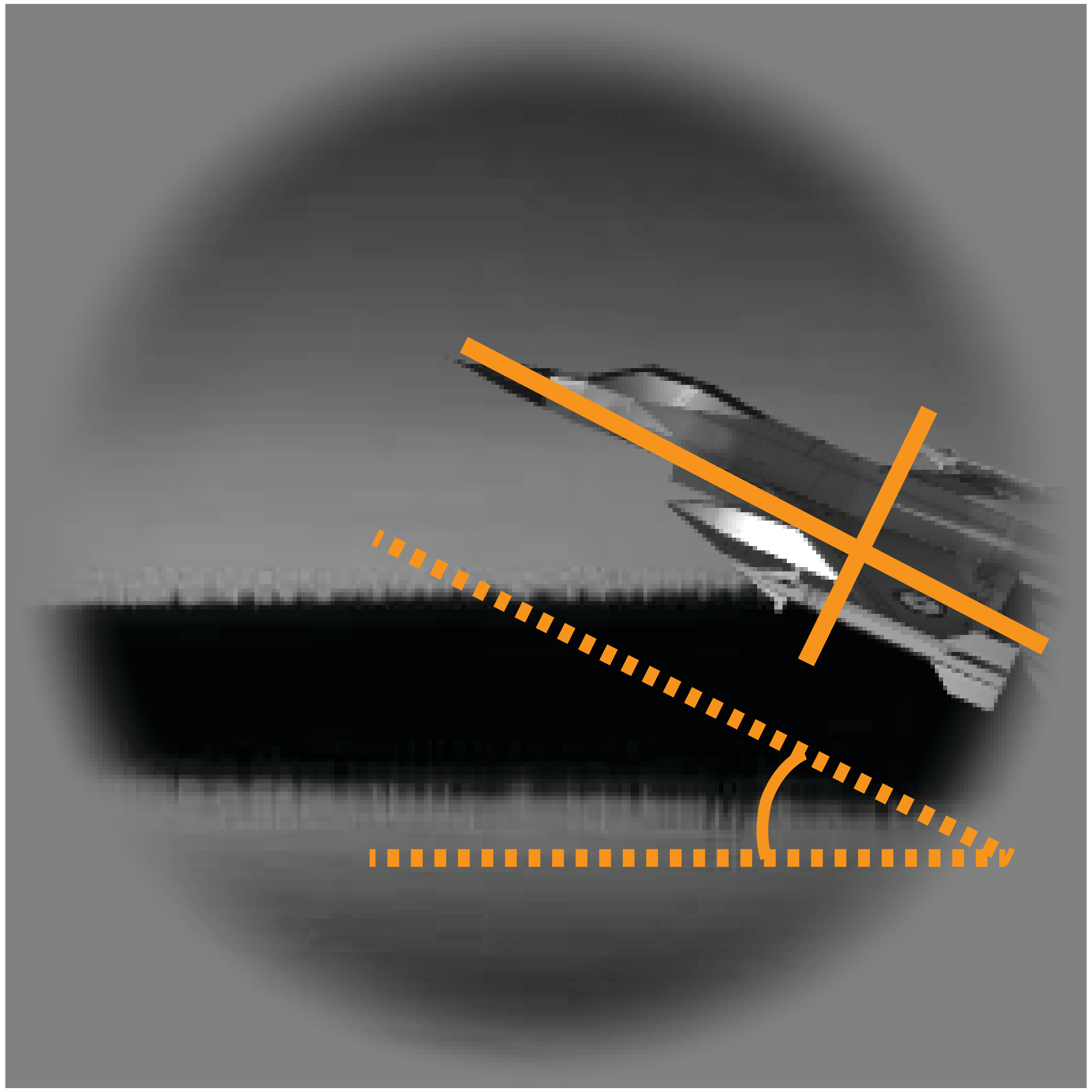
Position

Beyond categorization



Size

Beyond categorization



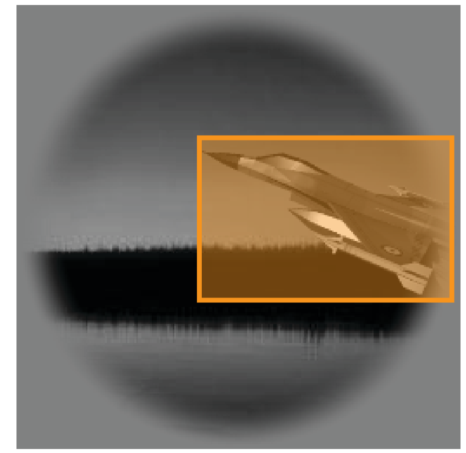
Aspect Ratio and Angle

Beyond categorization

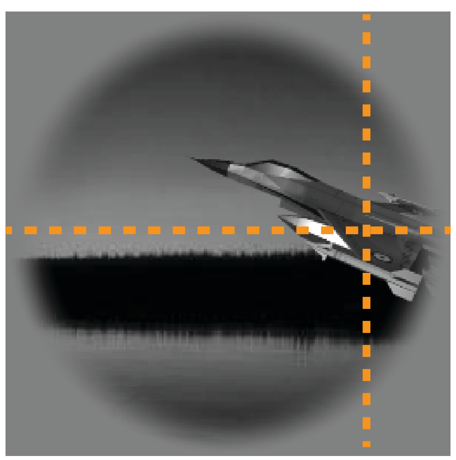
We can quickly assess the scene as a whole.

Category

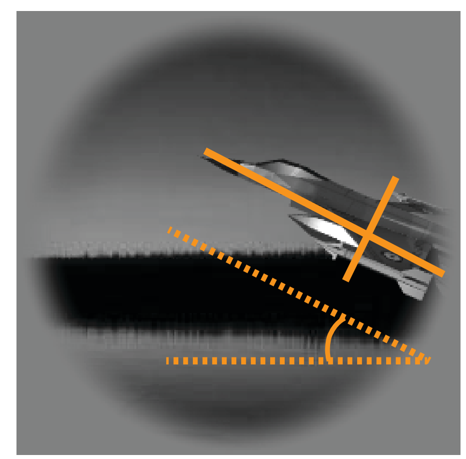
Identity



Bounding Box



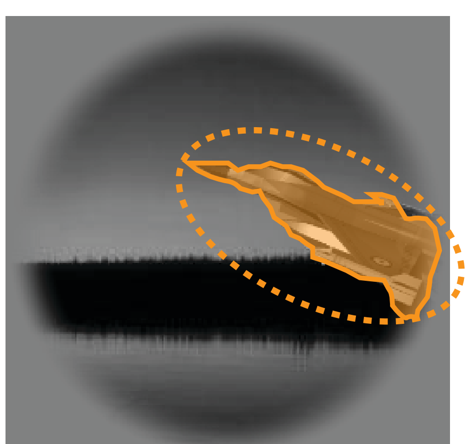
X and Y Axis Position



Aspect Ratio

Major Axis Length

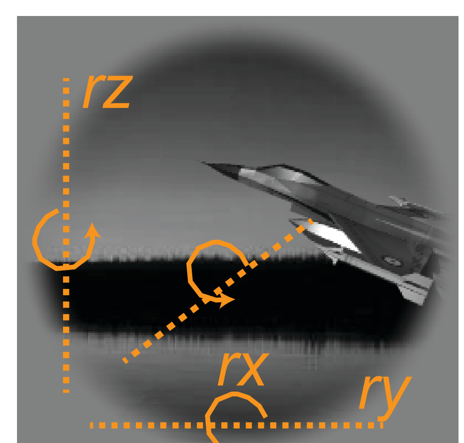
Major Axis Angle



Perimeter

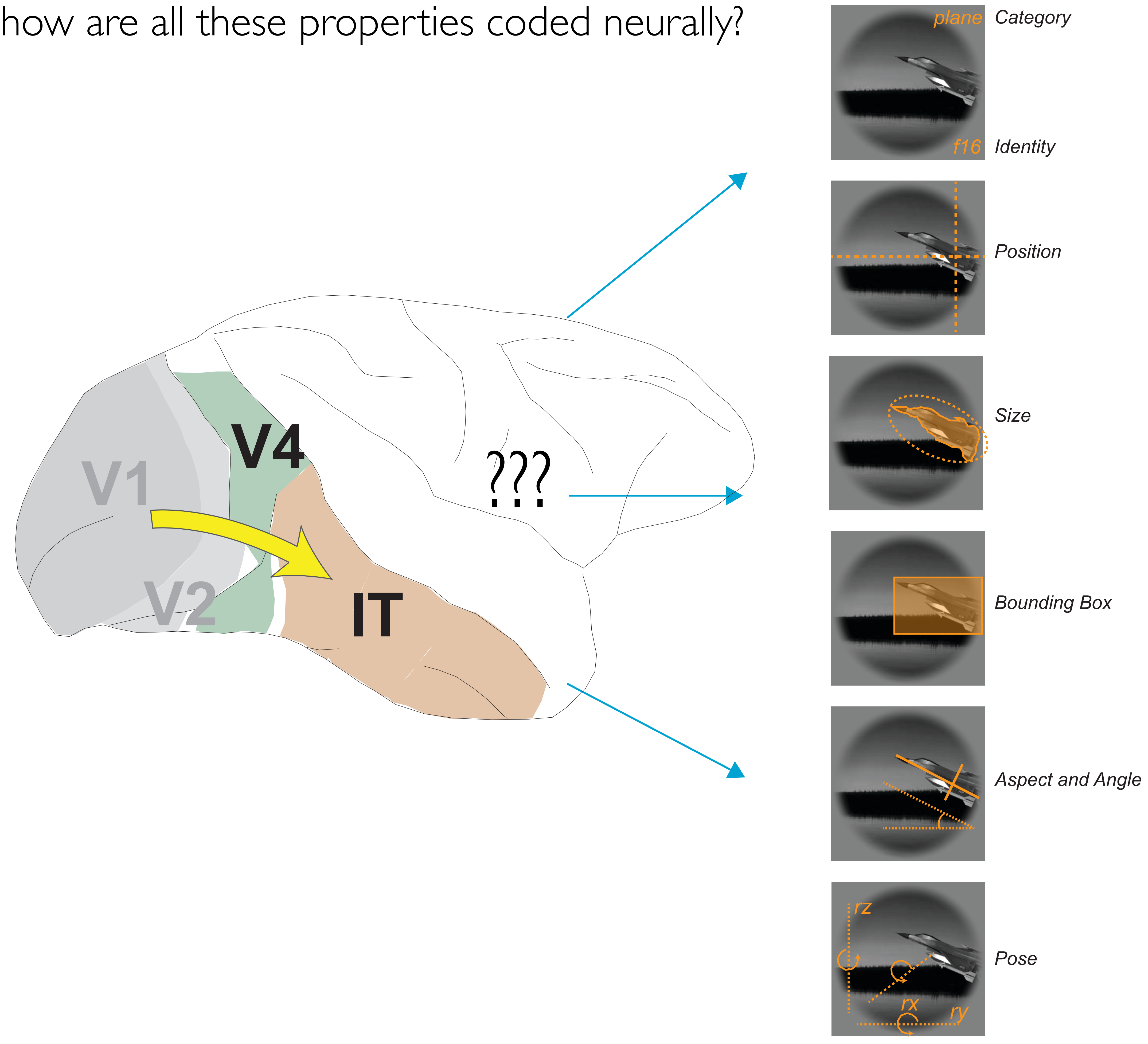
2-D Retinal Area

3-D Object Scale



Pose in each axis

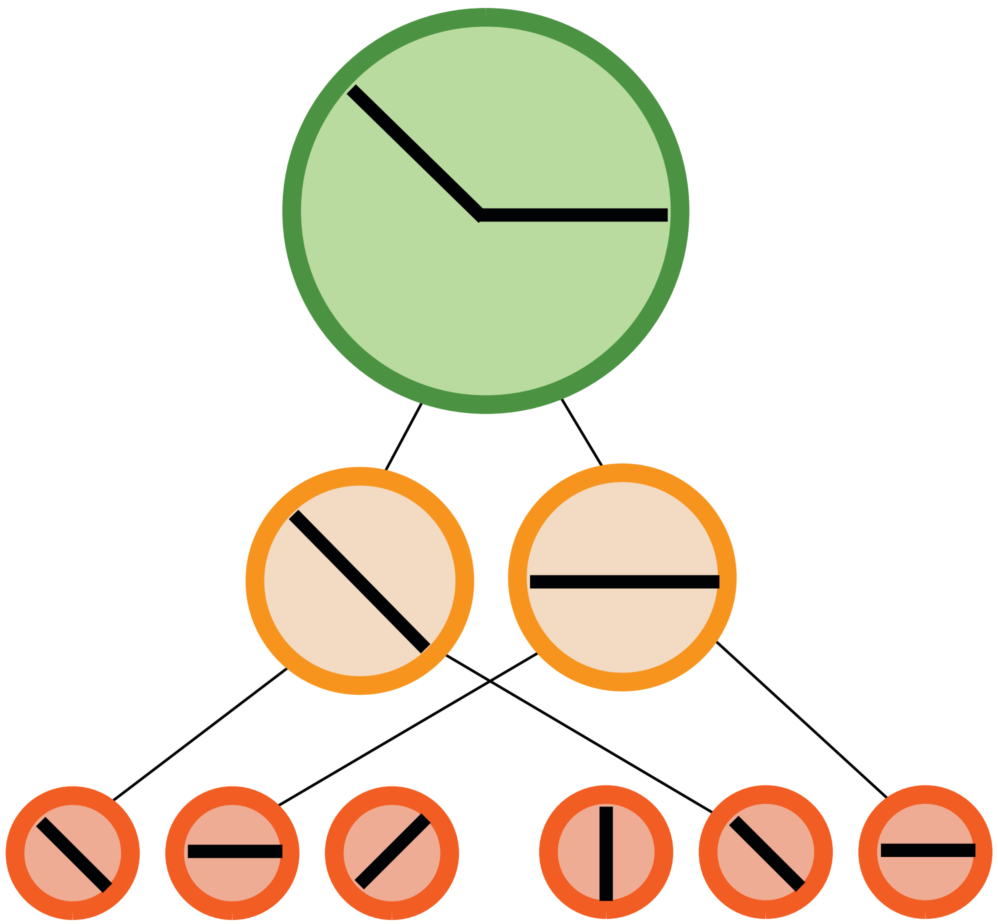
Where and how are all these properties coded neurally?



Beyond categorization

"Standard word model" predicts: **not at the top of the ventral stream**.

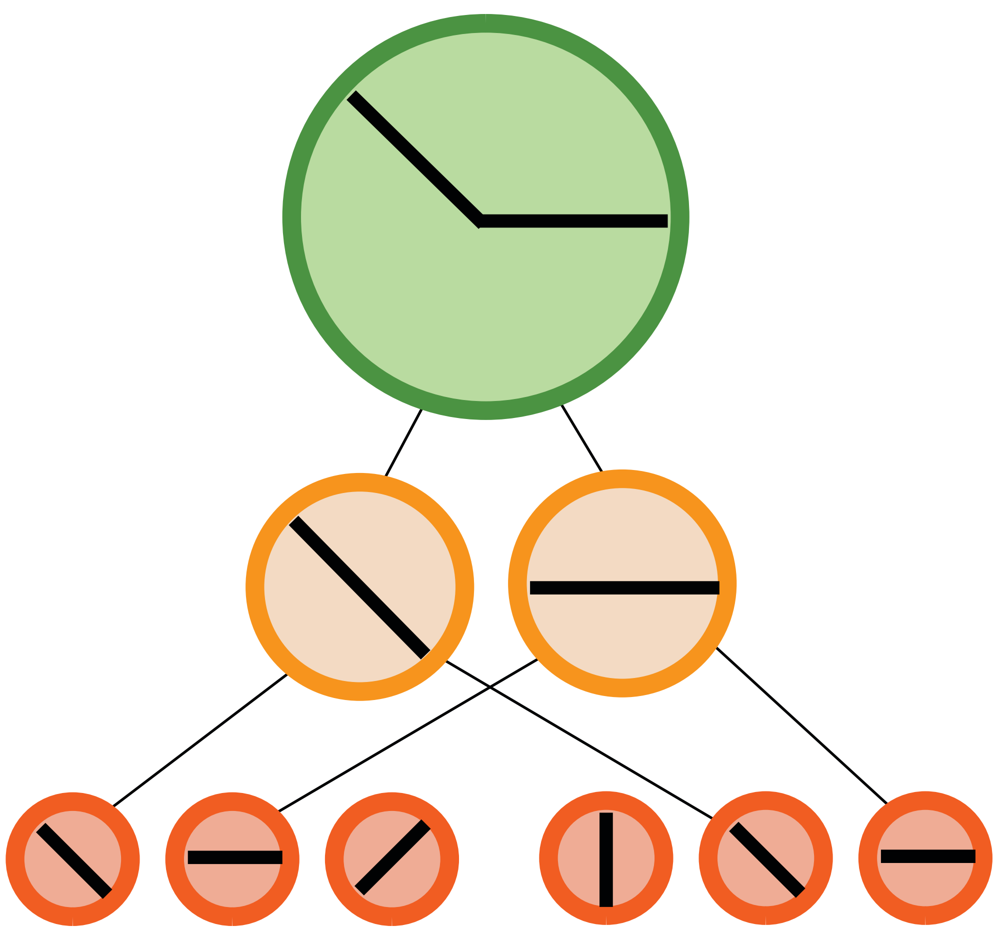
Aggregation over identity-preserving transformations, e.g. translation.



Beyond categorization

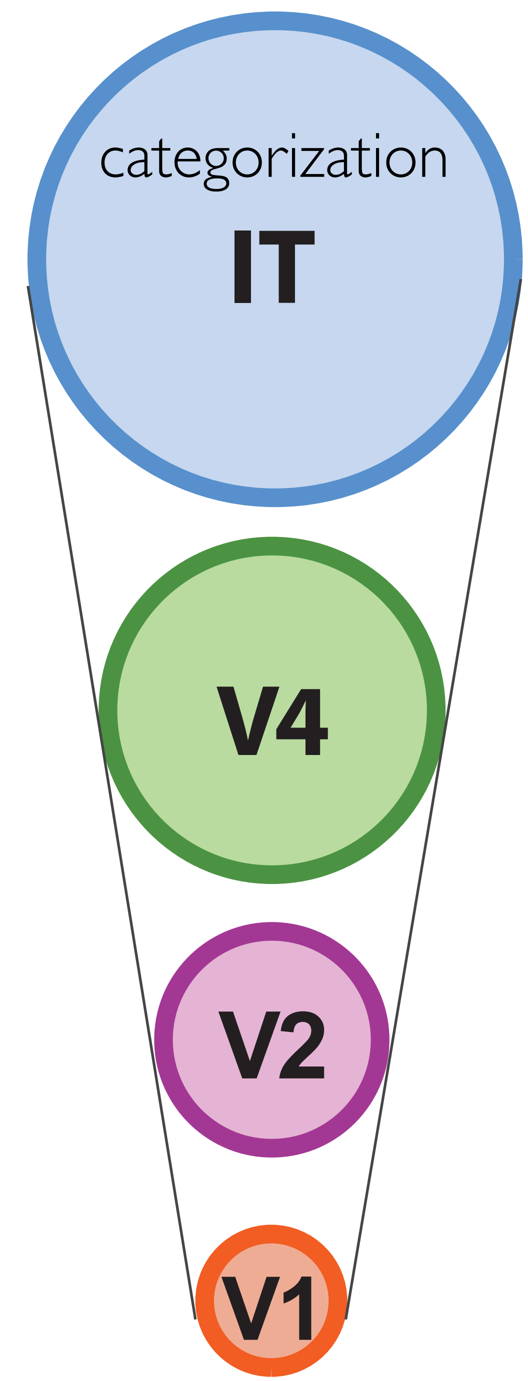
"Standard word model" predicts: **not at the top of the ventral stream**.

Aggregation over identity-preserving transformations, e.g. translation.



Receptive Field Size ↑

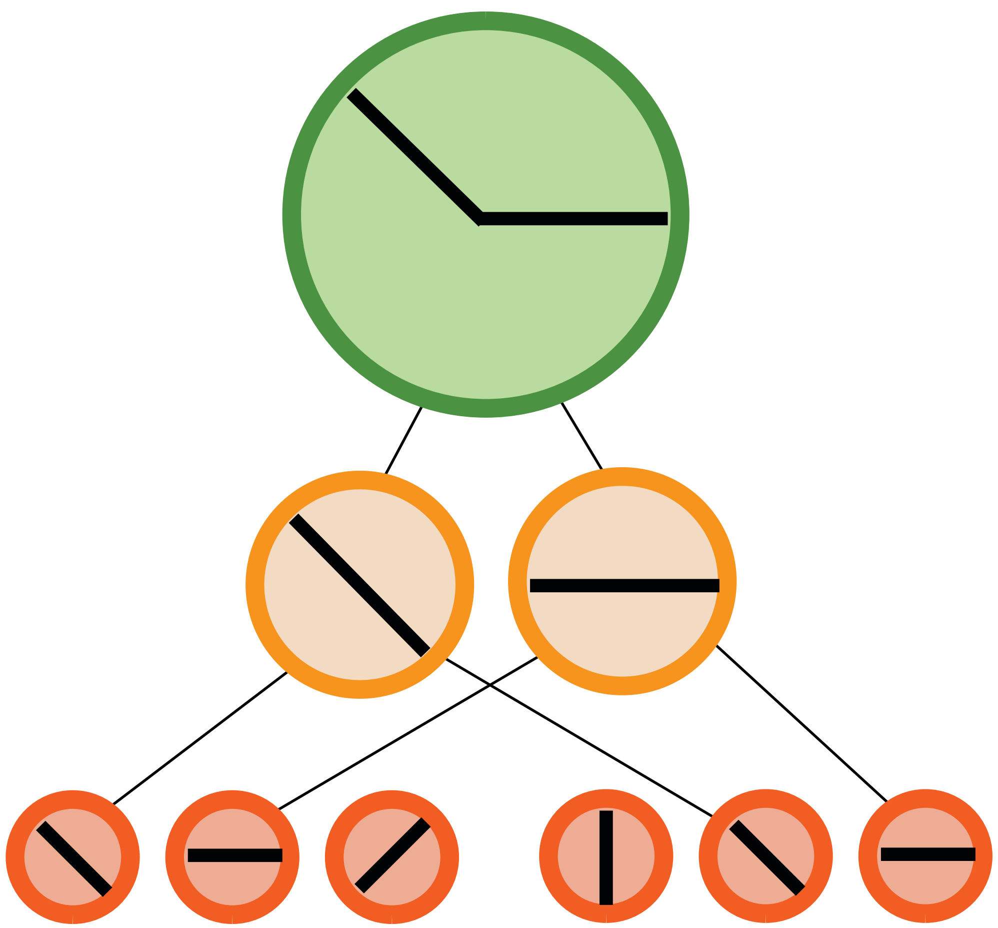
Category Invariance ↑

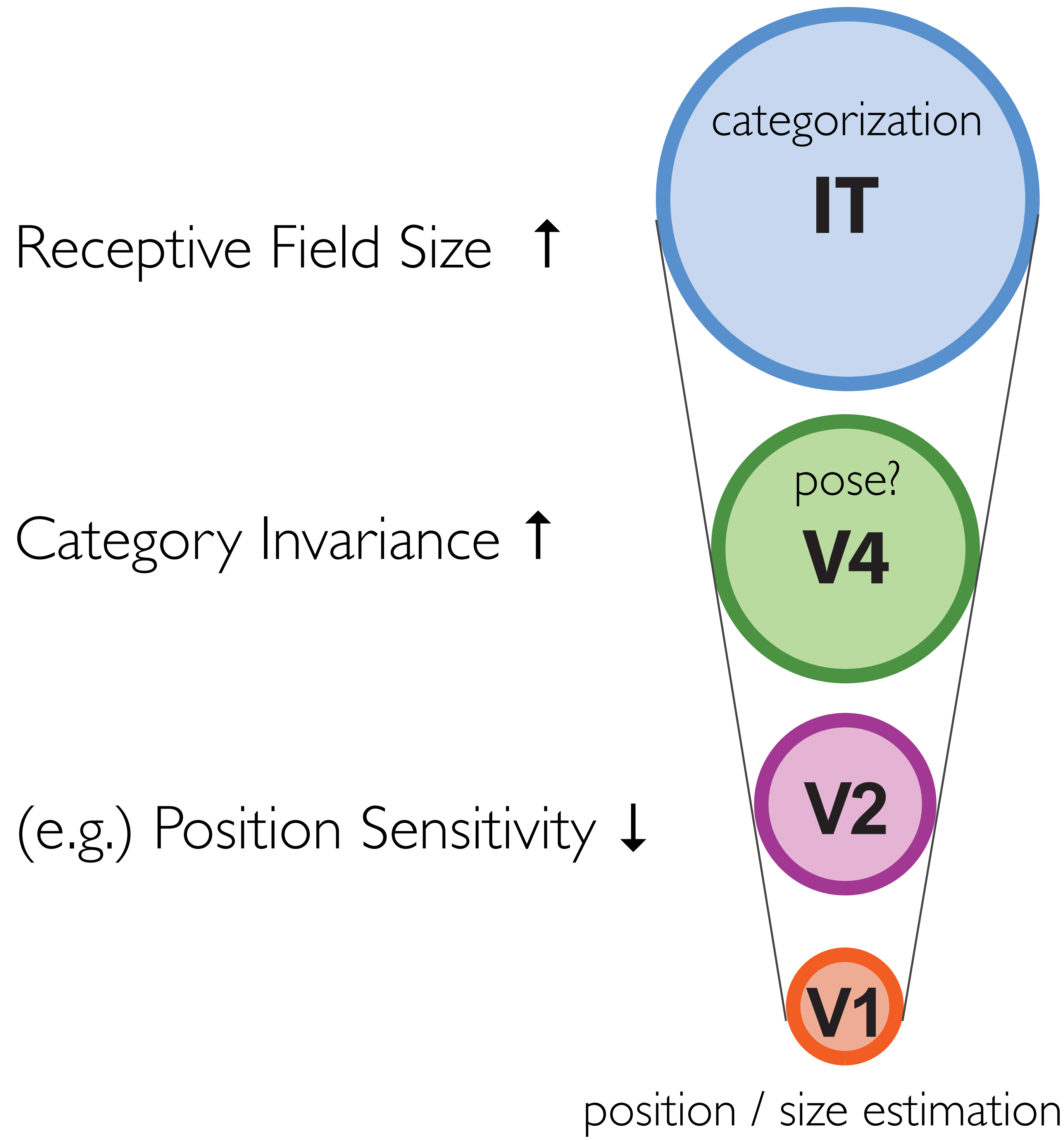


Beyond categorization

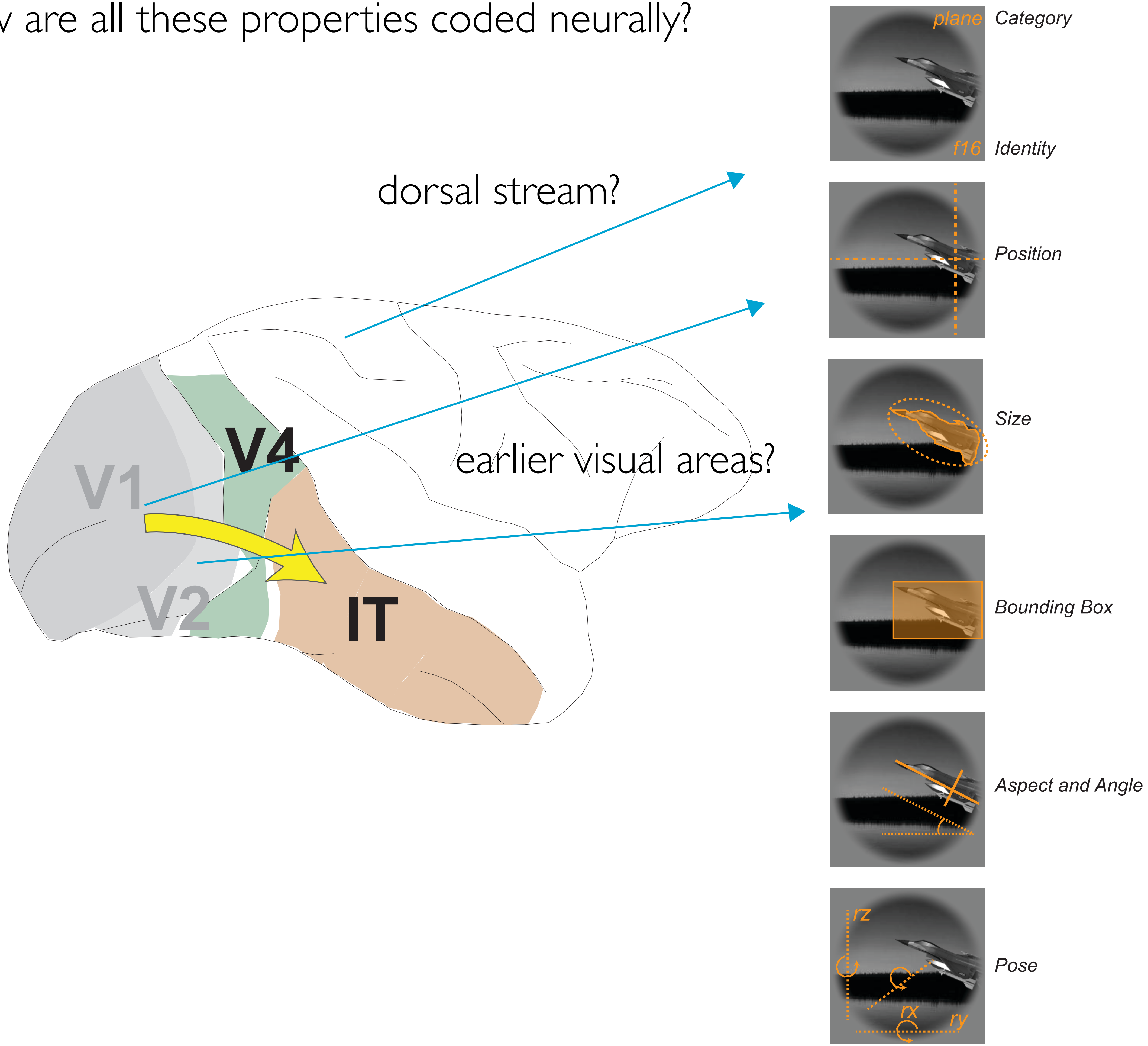
“Standard word model” predicts: **not at the top of the ventral stream**.

Aggregation over identity-preserving transformations, e.g. translation.





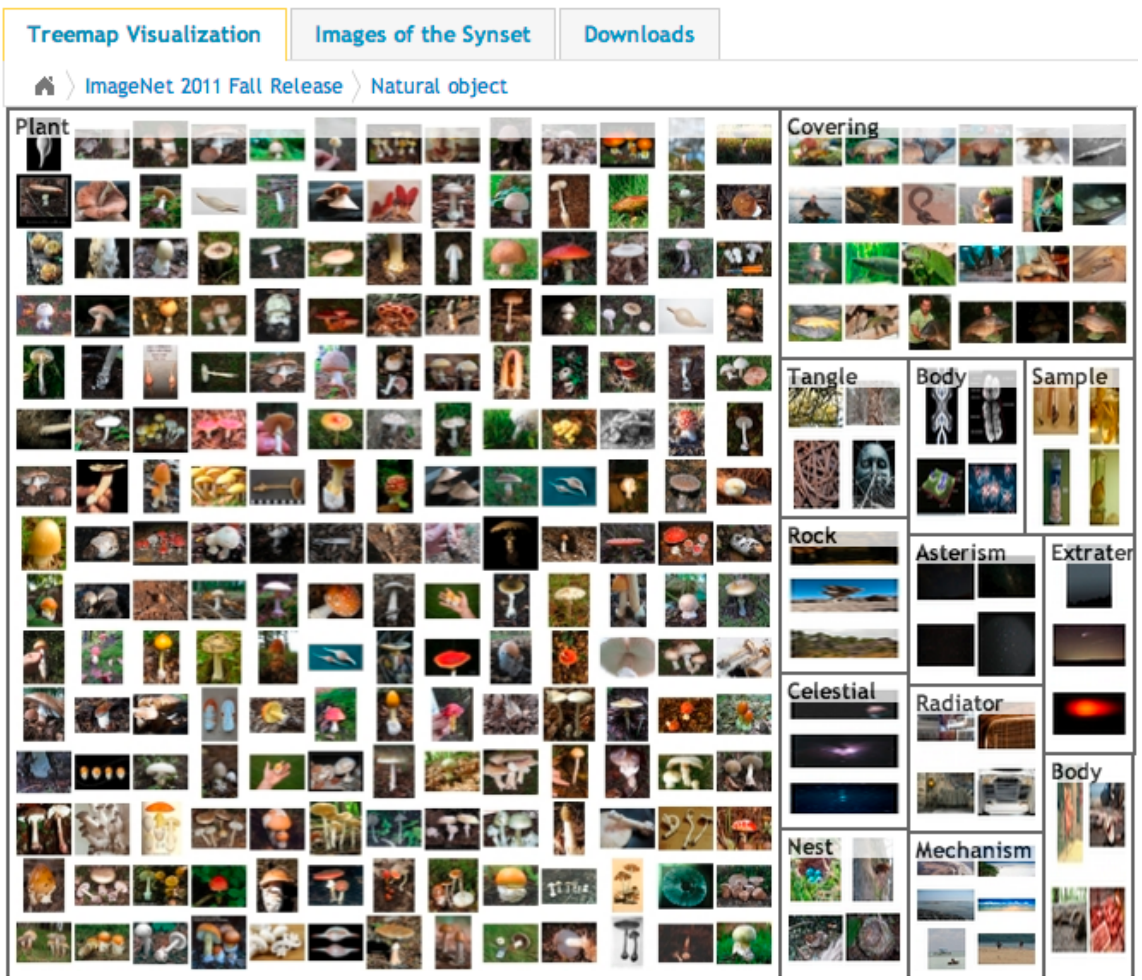
Where and how are all these properties coded neurally?

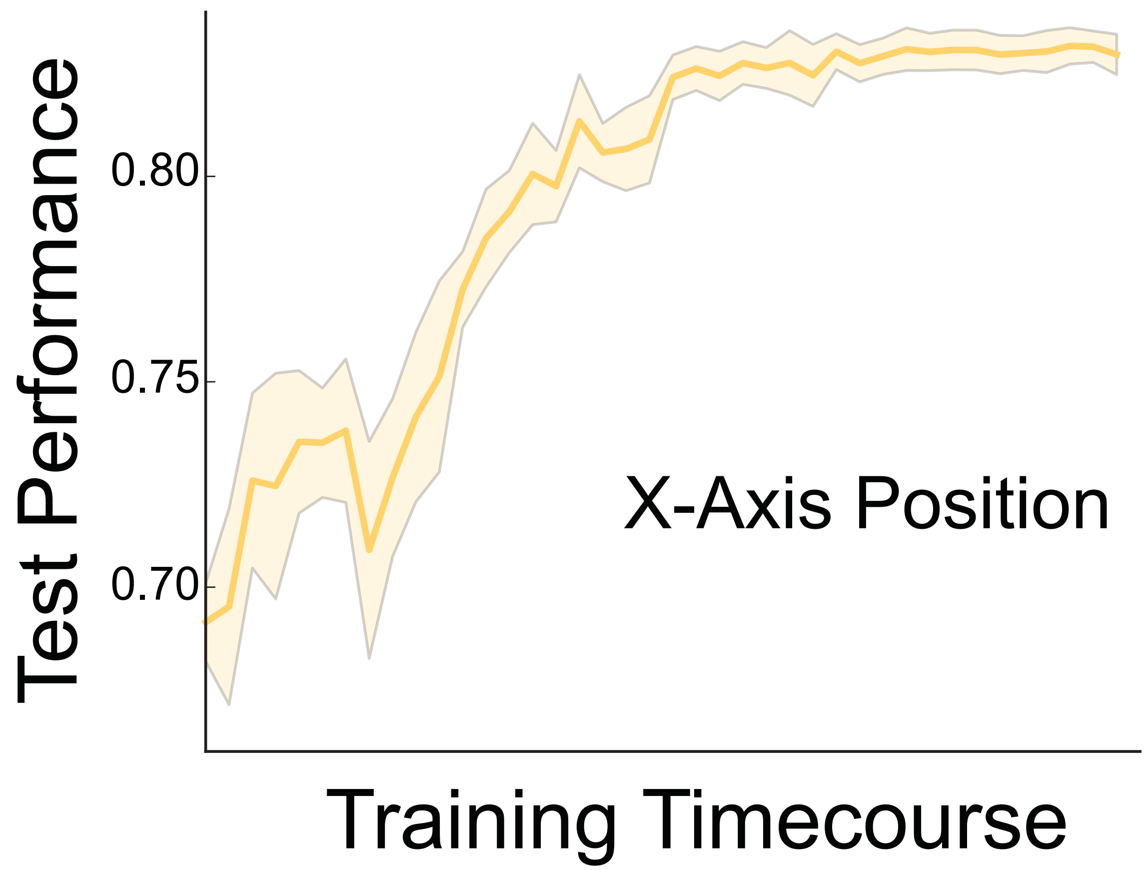


Beyond categorization

Unexpected observation:

Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*





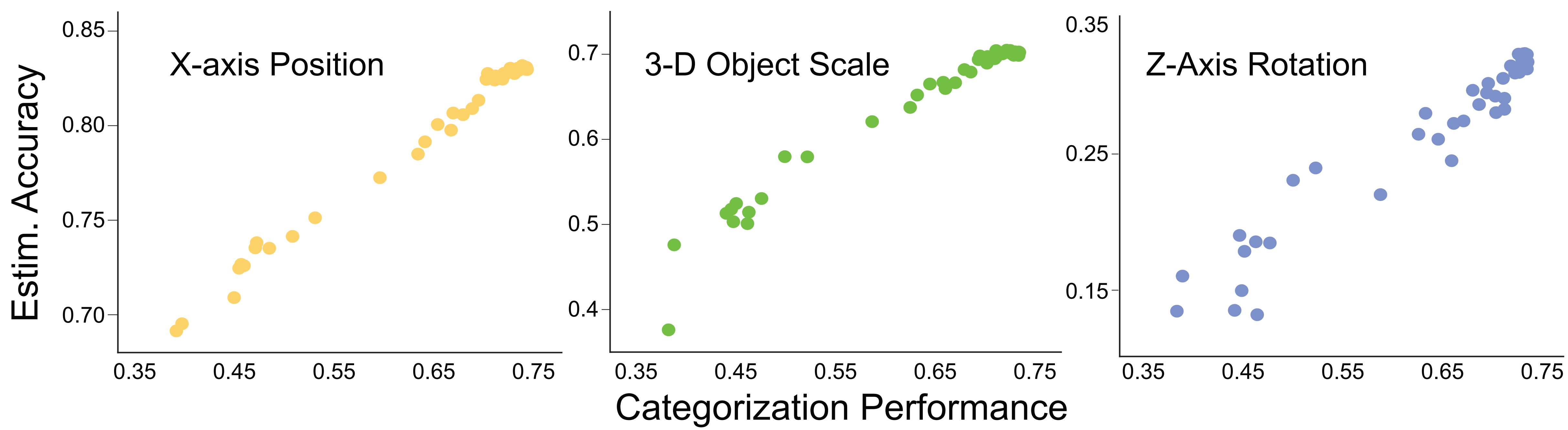
Training on
categorization task

Increased performance on
position estimation task.

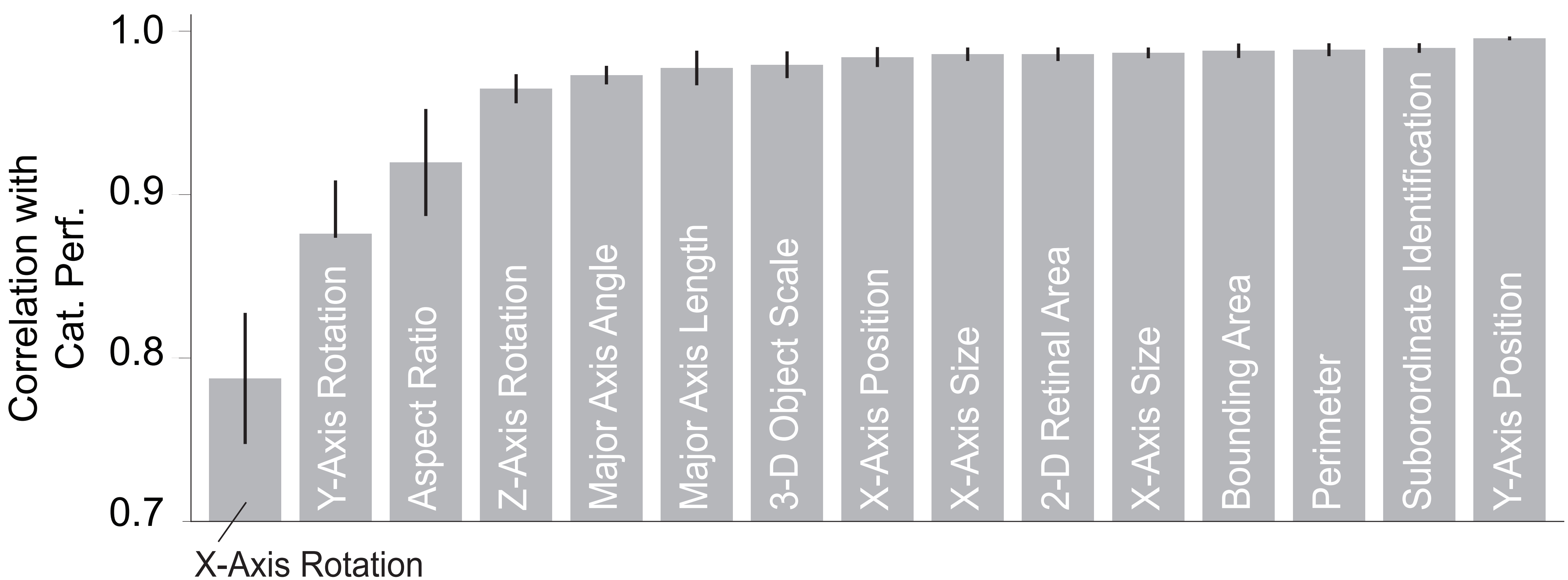
even though the goal was to become INVARIANT to position

Beyond categorization

Category optimization → improved performance on non-categorical tasks.

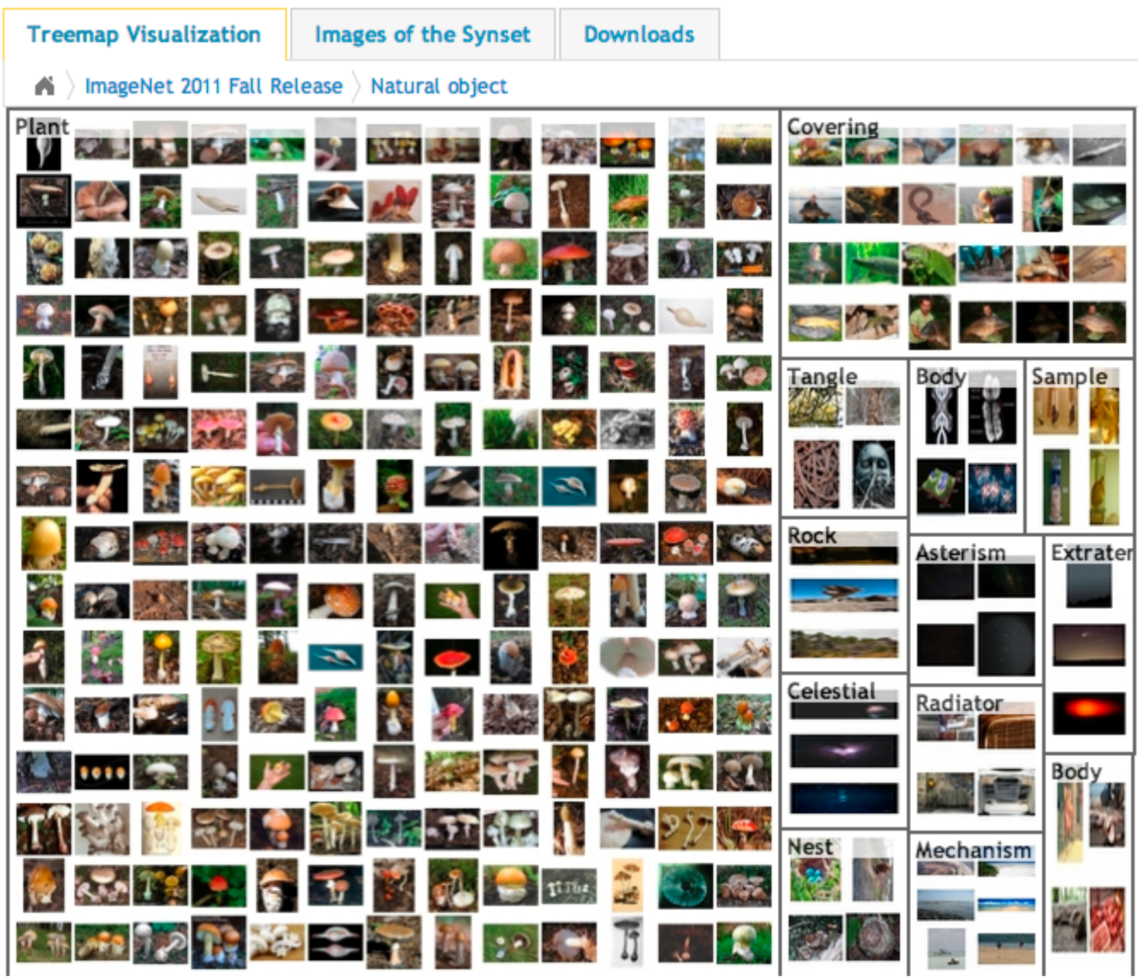


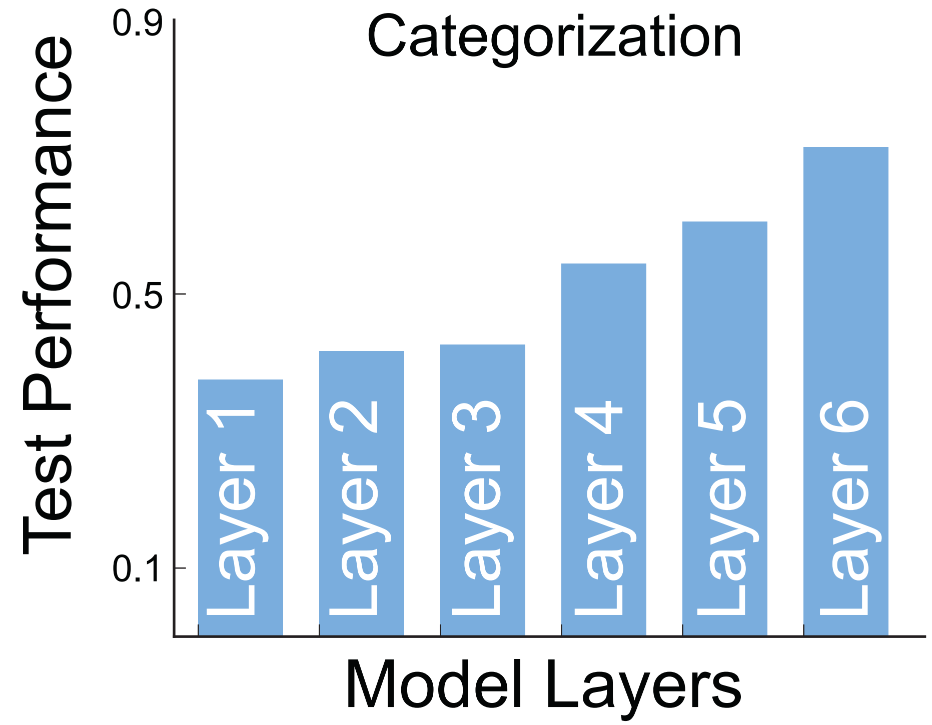
Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*



Beyond categorization

Unexpected observation #2:

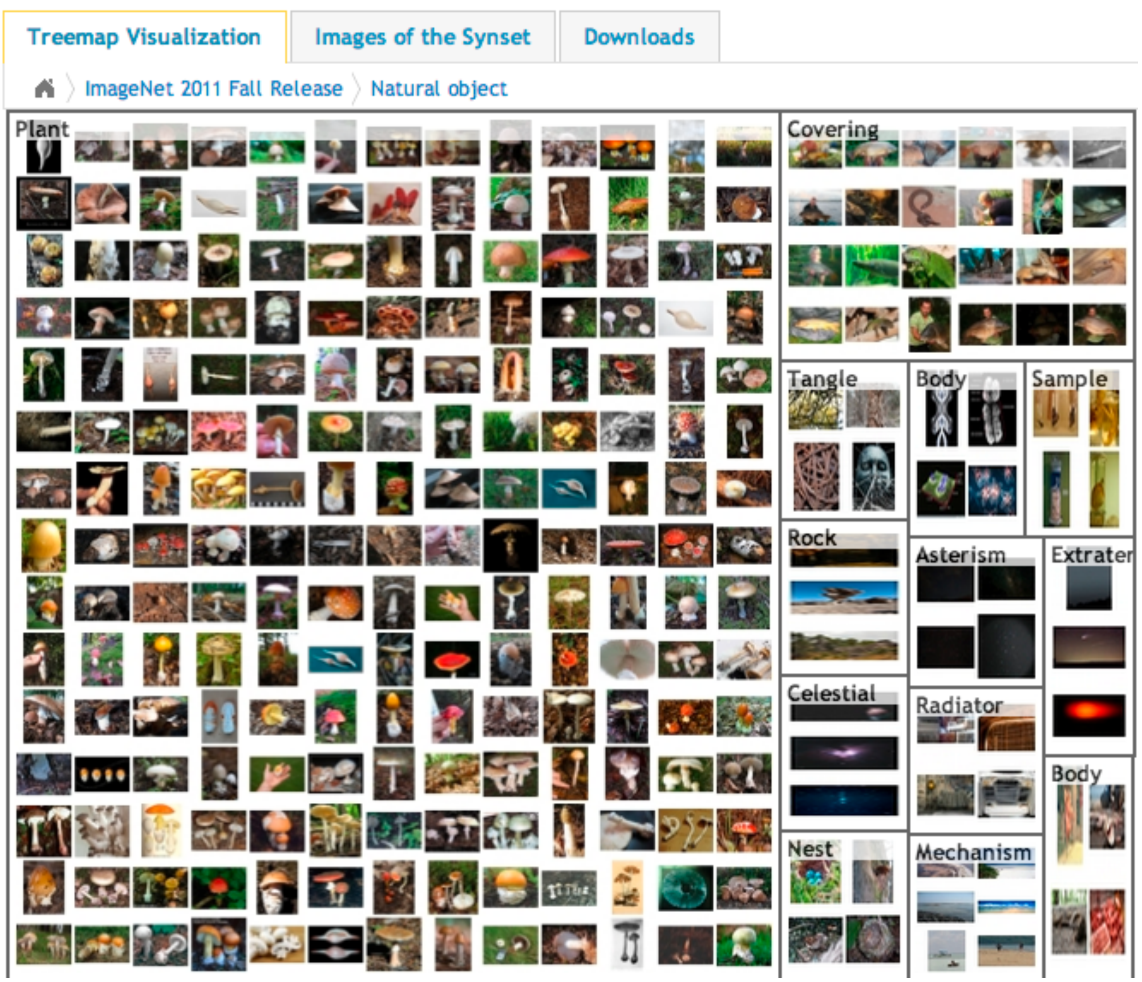


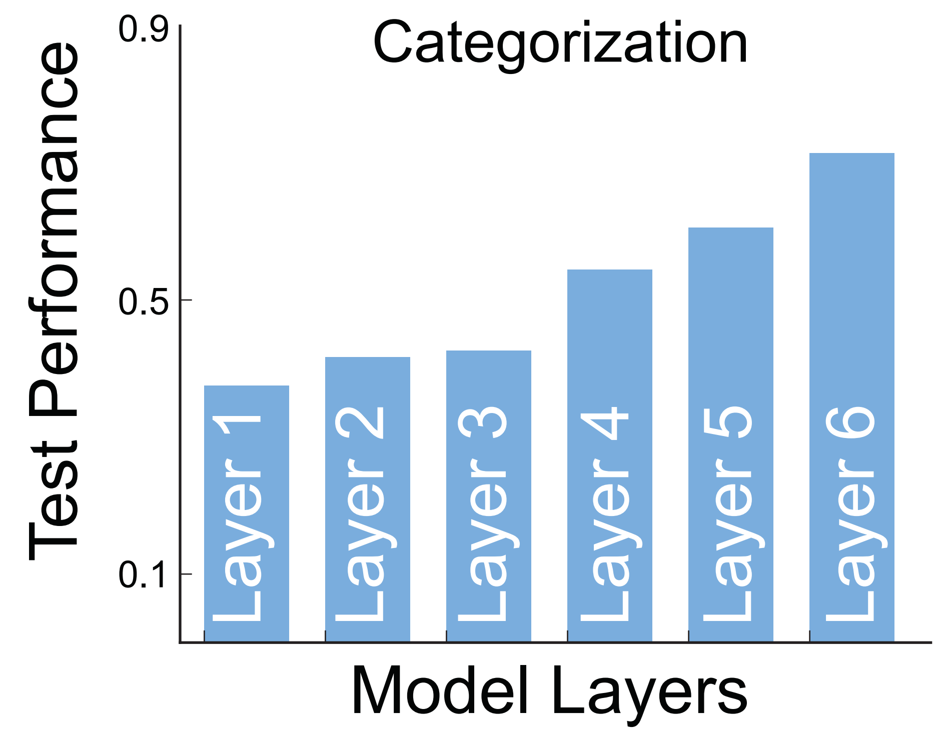


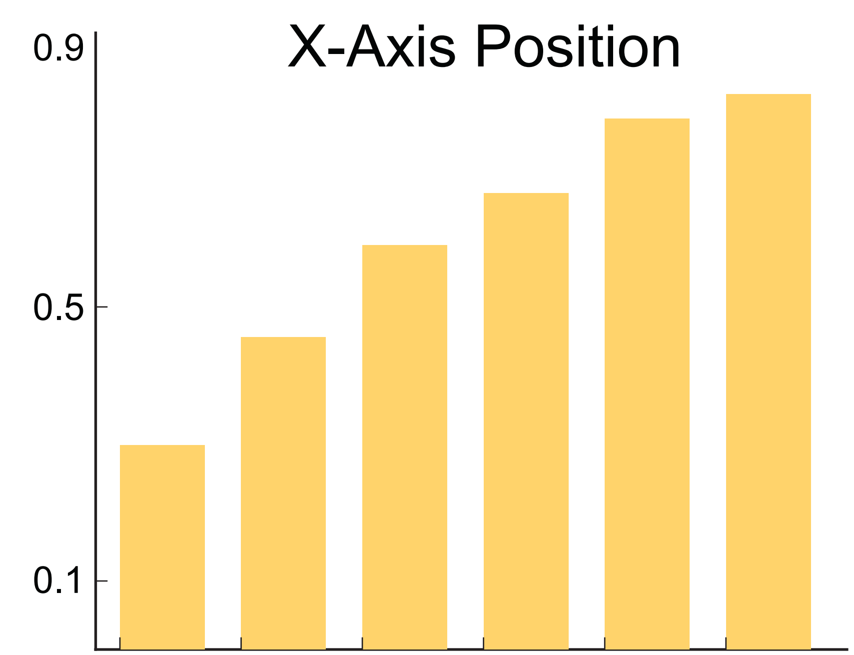
Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*

Beyond categorization

Unexpected observation #2:





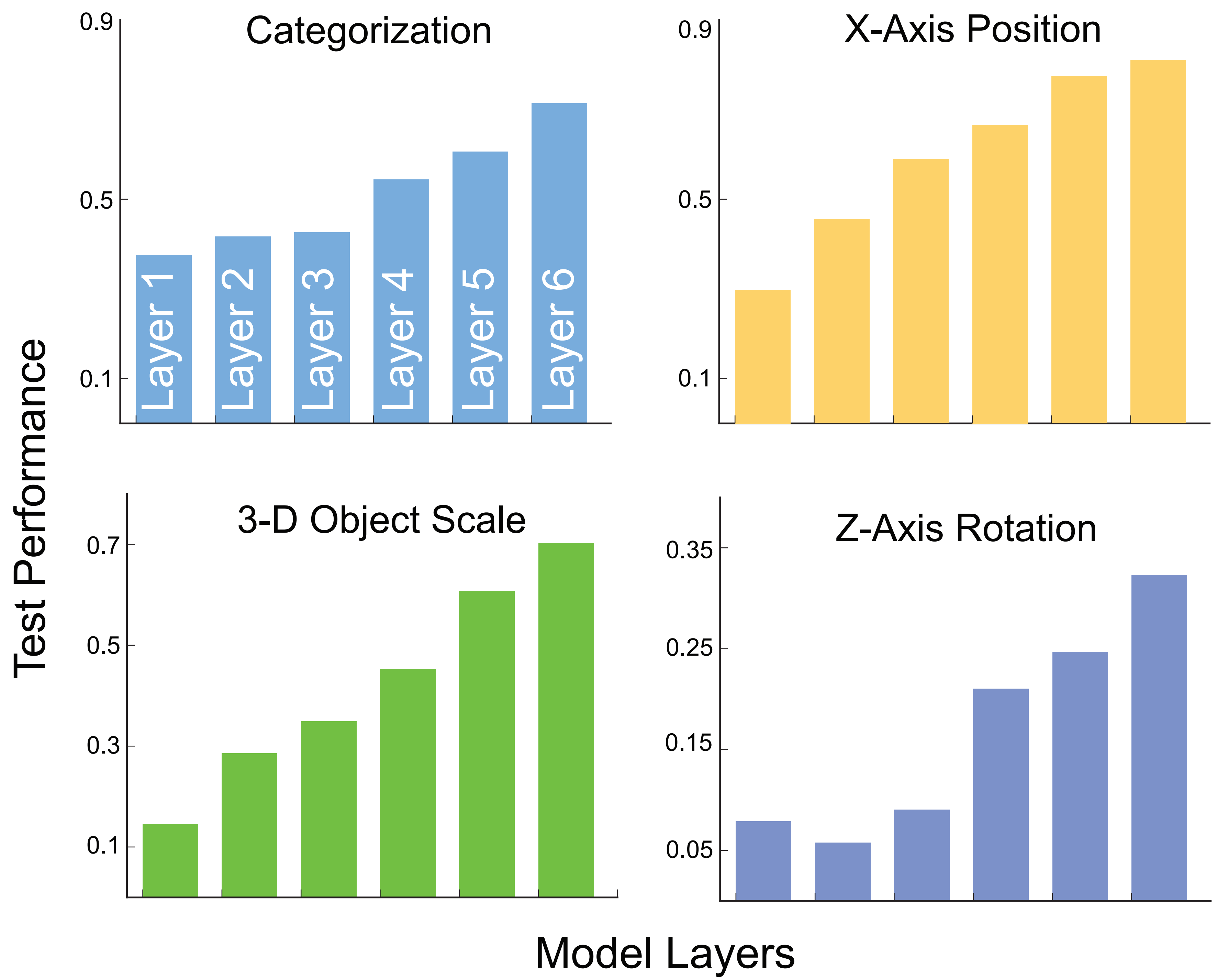


Increased performance on position estimation task at each model layer.

Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*

Beyond categorization

For all tasks of visual interest we could measure in our test dataset:

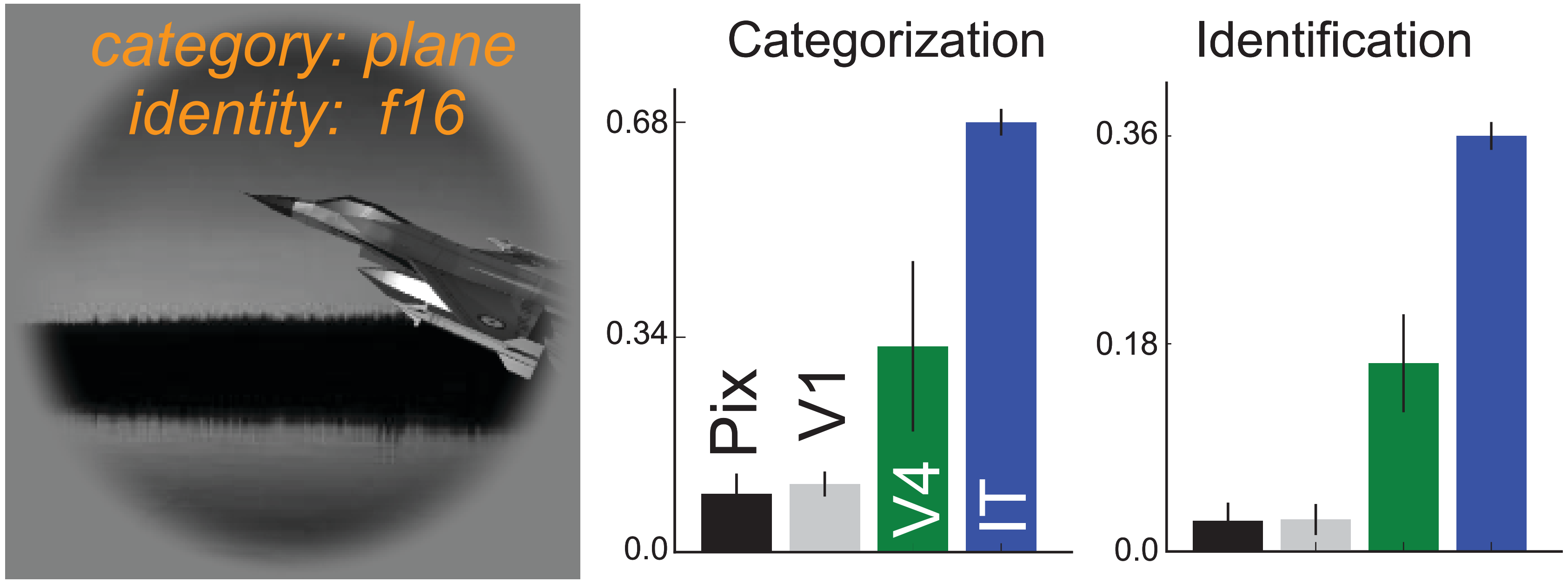


Performance on non-categorical tasks increases at each layer.

Beyond categorization

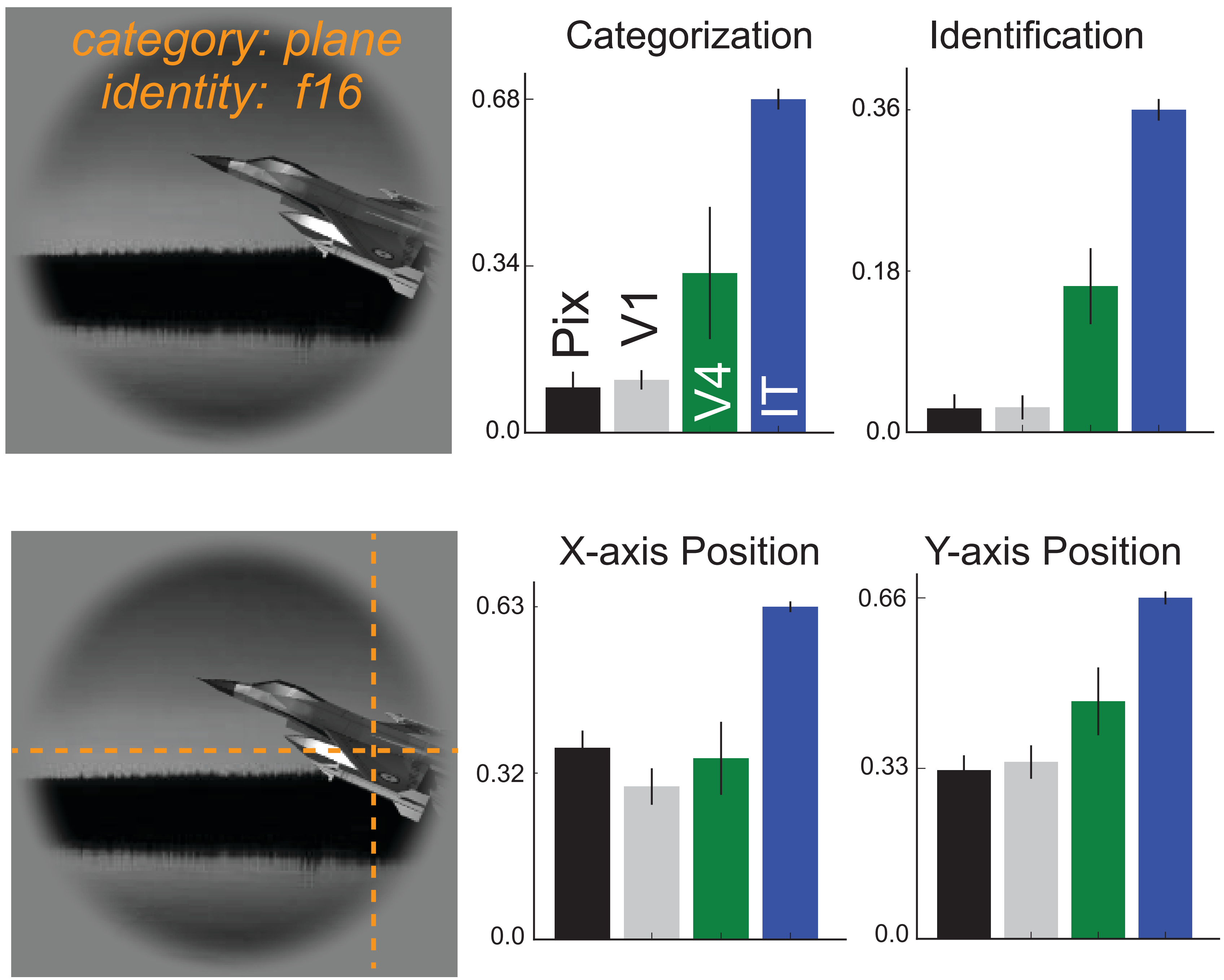
What do the data say?

Population Decoding



Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*

Population Decoding



Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*

IT cortex

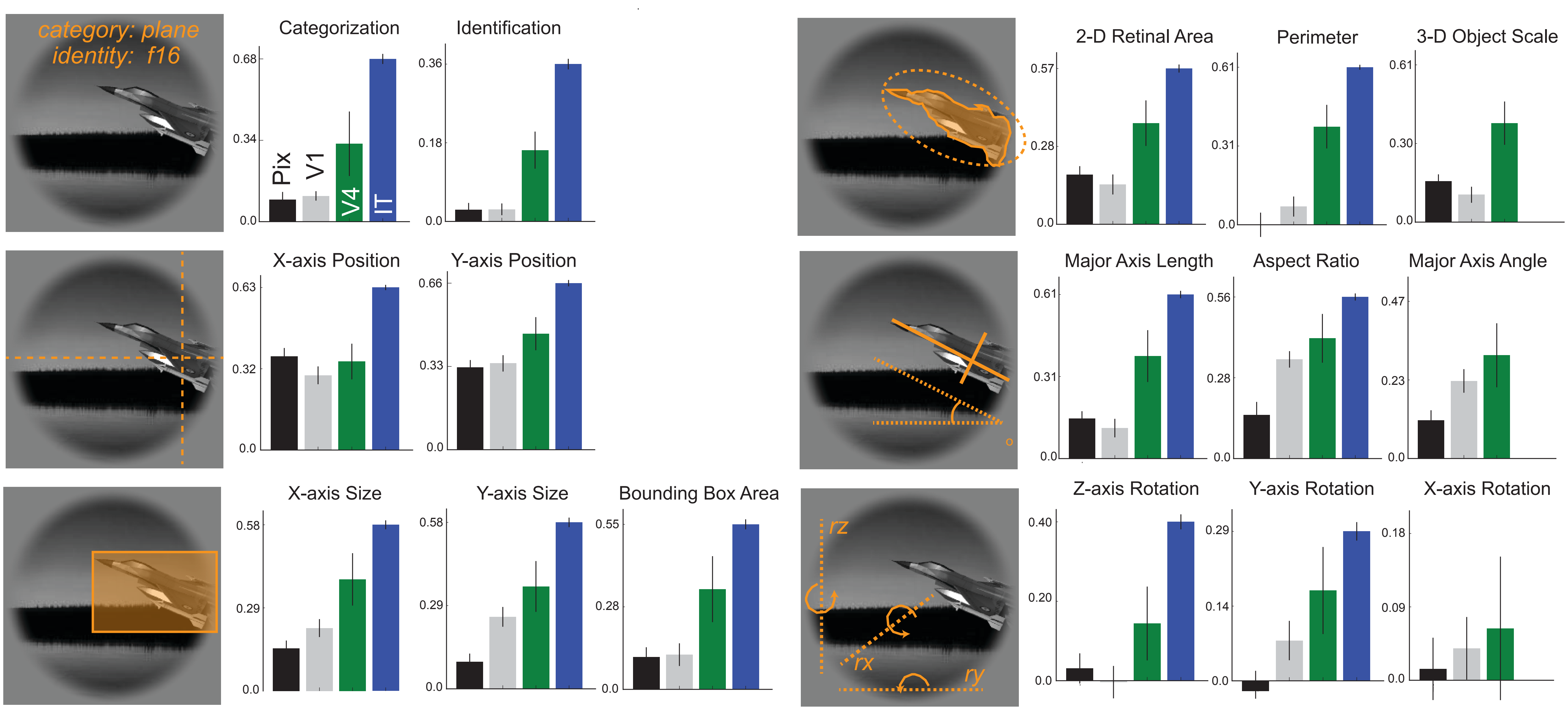
V4 cortex

V1-like model

pixel control

Population Decoding

IT > V4, V1 for all tasks **V4 > V1** for most tasks



Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*

IT cortex

V4 cortex

V1-like model

pixel control

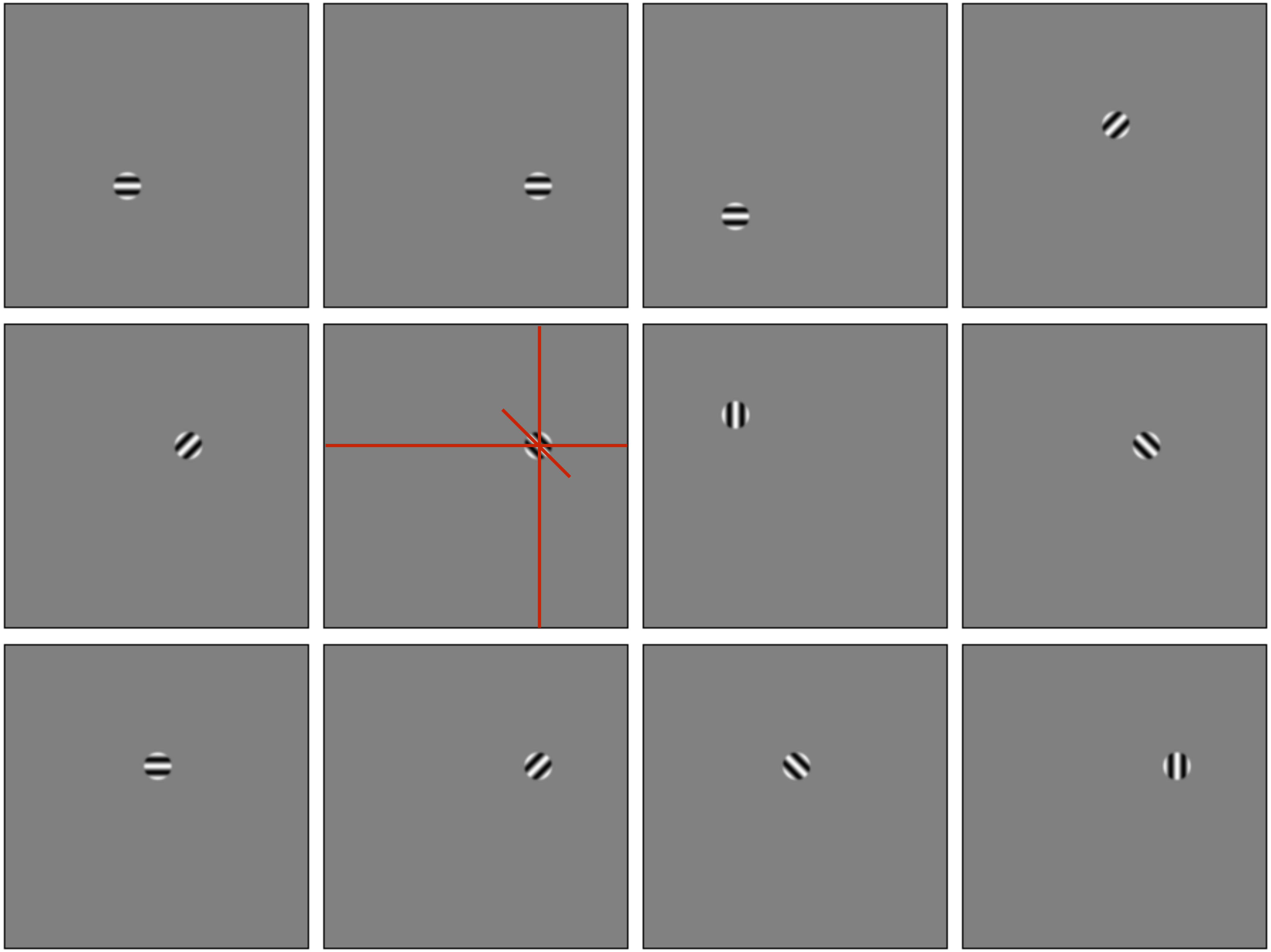
Population Decoding

"Standard" receptive field-mapping stimuli w/ position and orientation variation:

X-position

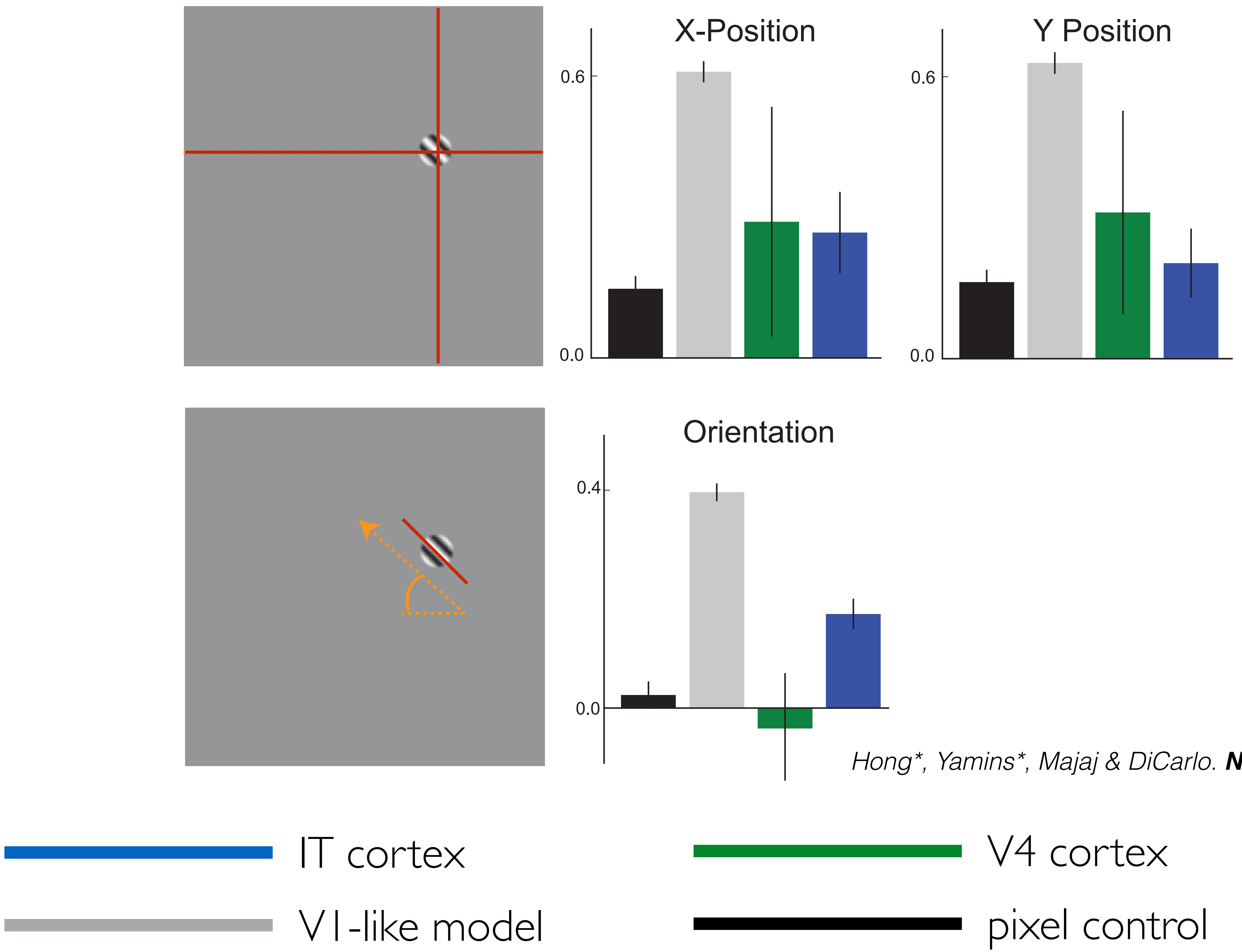
Y-position

Orientation



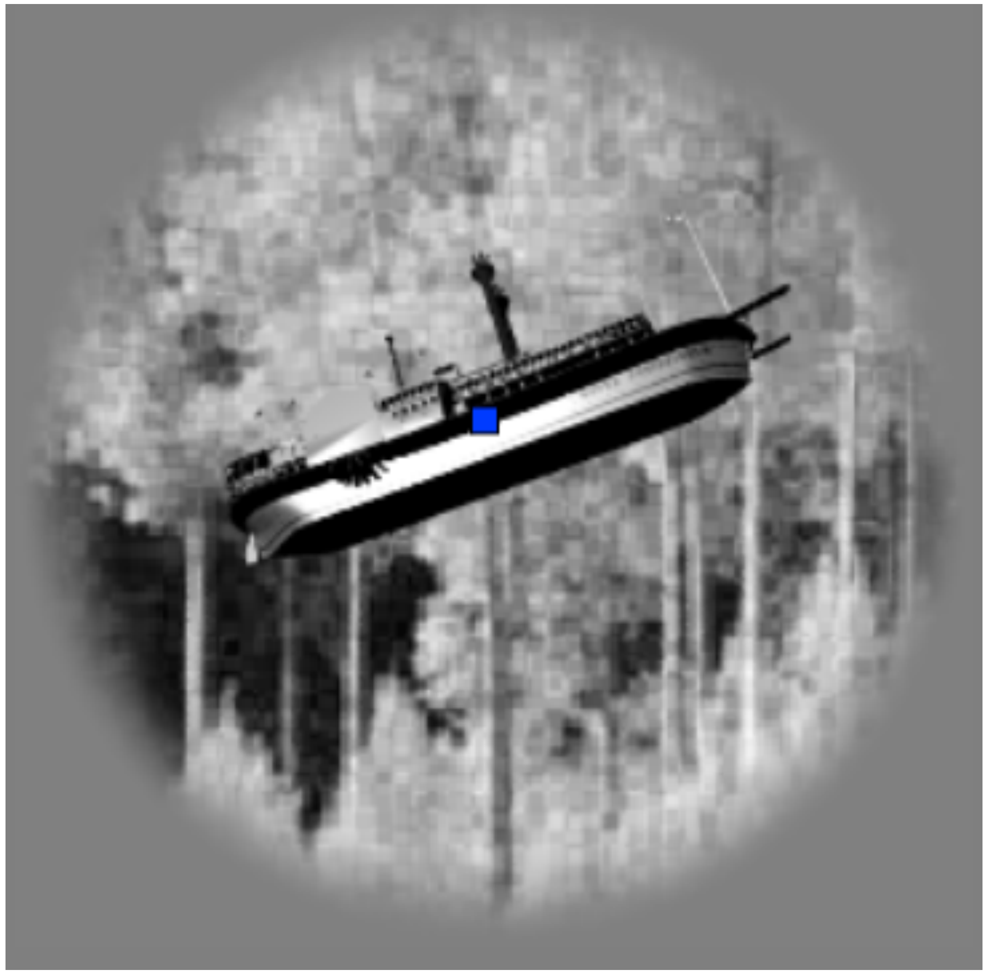
Population Decoding

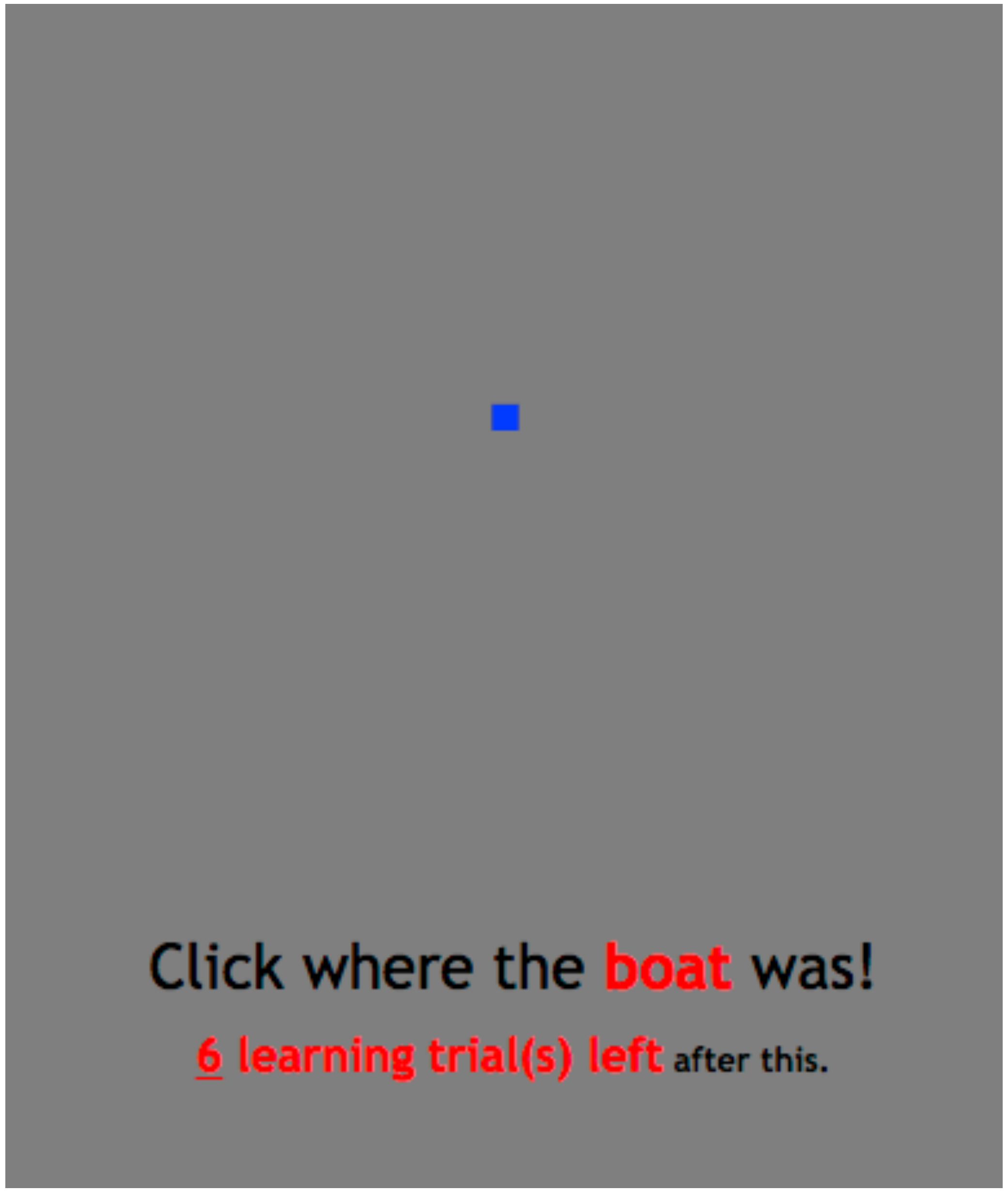
V1 > V4, IT for "standard" tasks



Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*

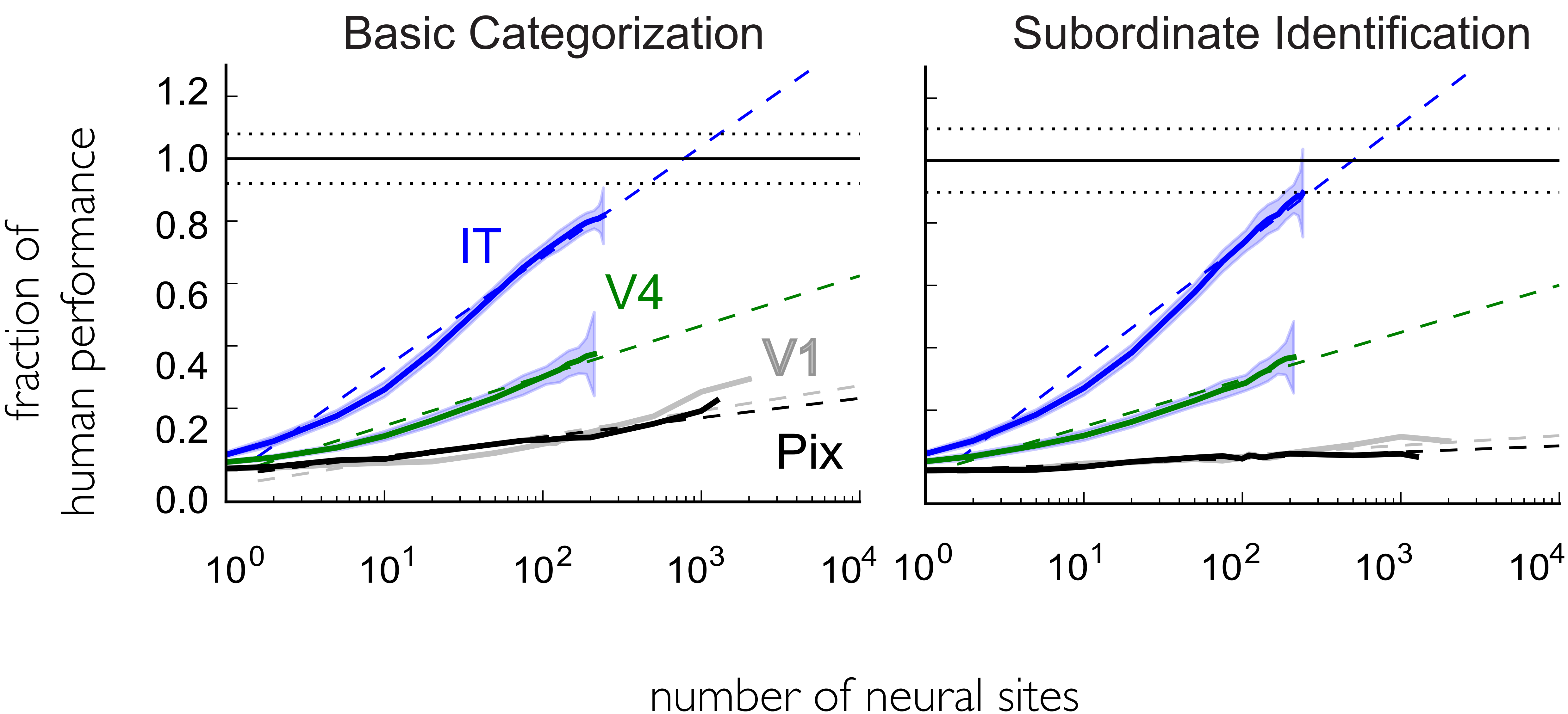
Human Psychophysical Measurements





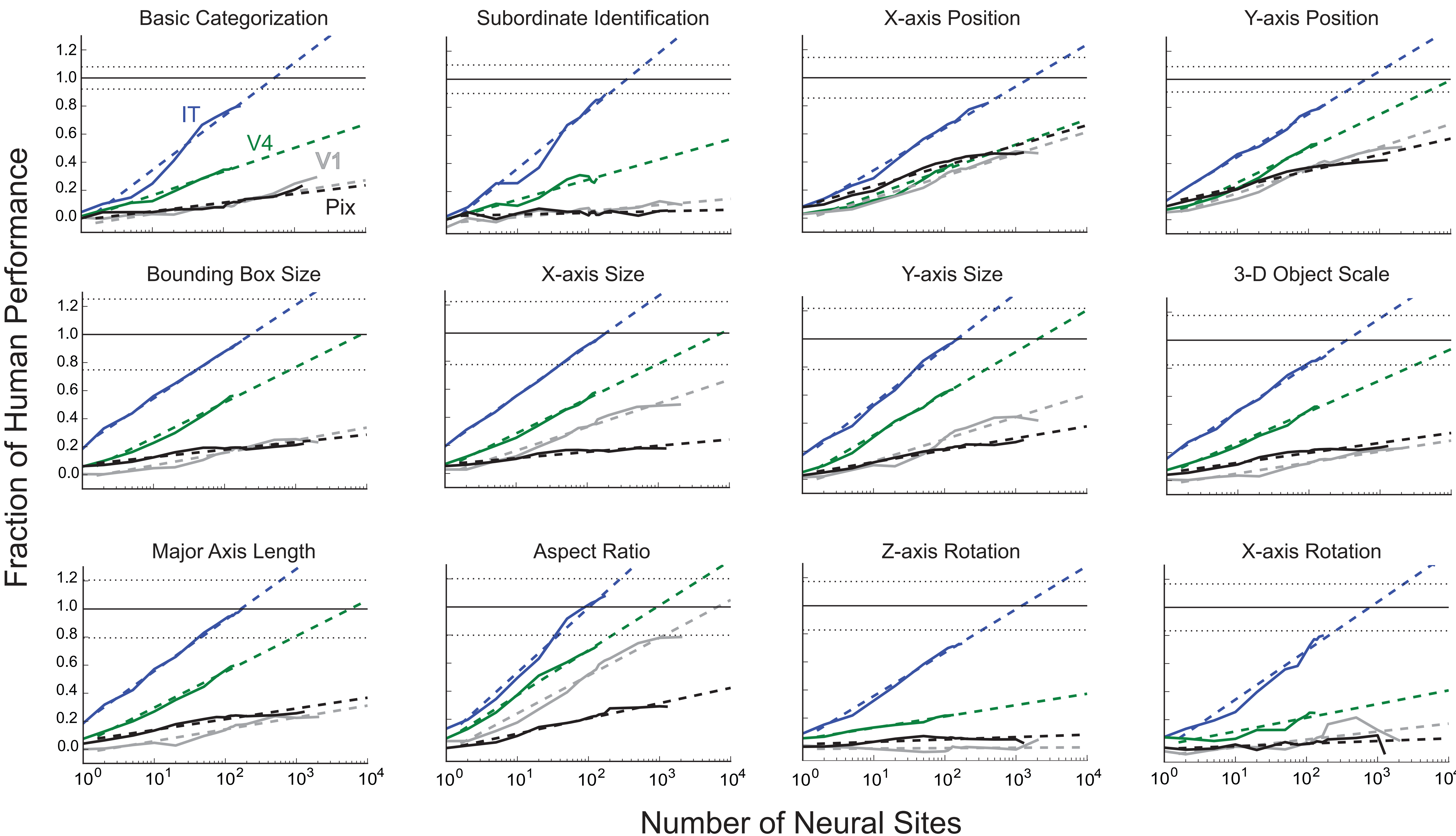
Monkey Neurons vs Humans

performance ~ k * log(N)



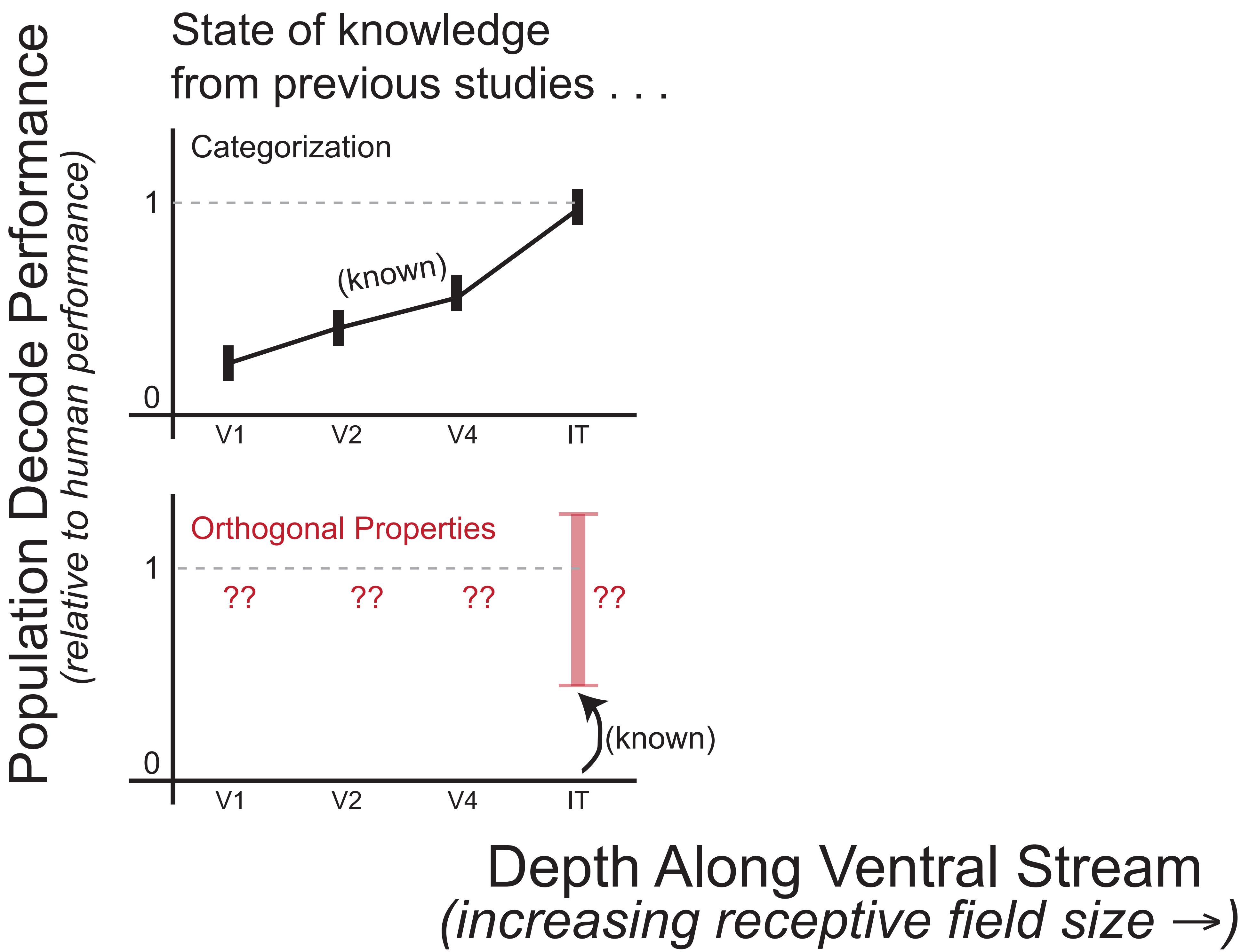
Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*

Monkey Neurons vs Humans

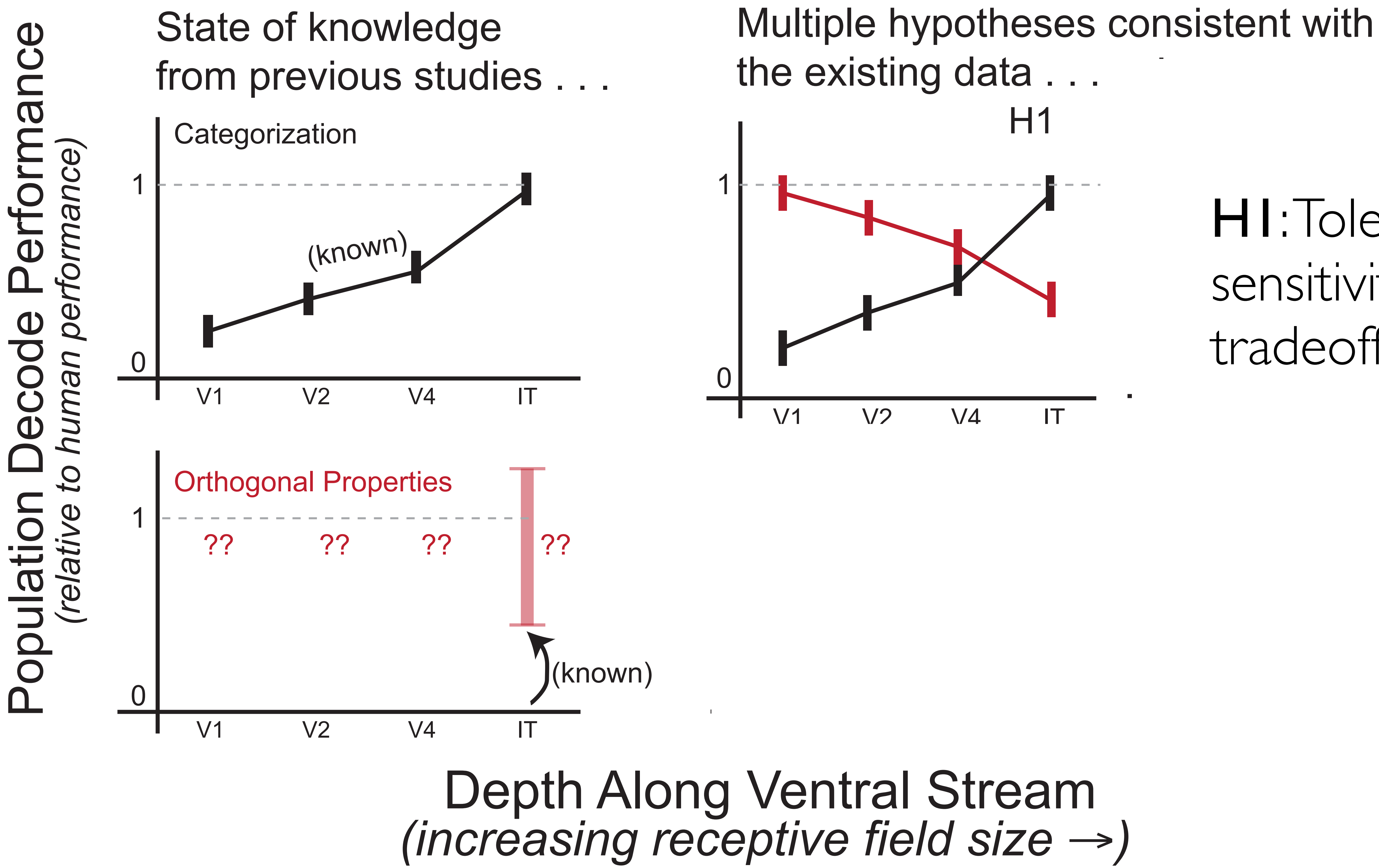


Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(2016)*

Somewhat newish ideas about IT?

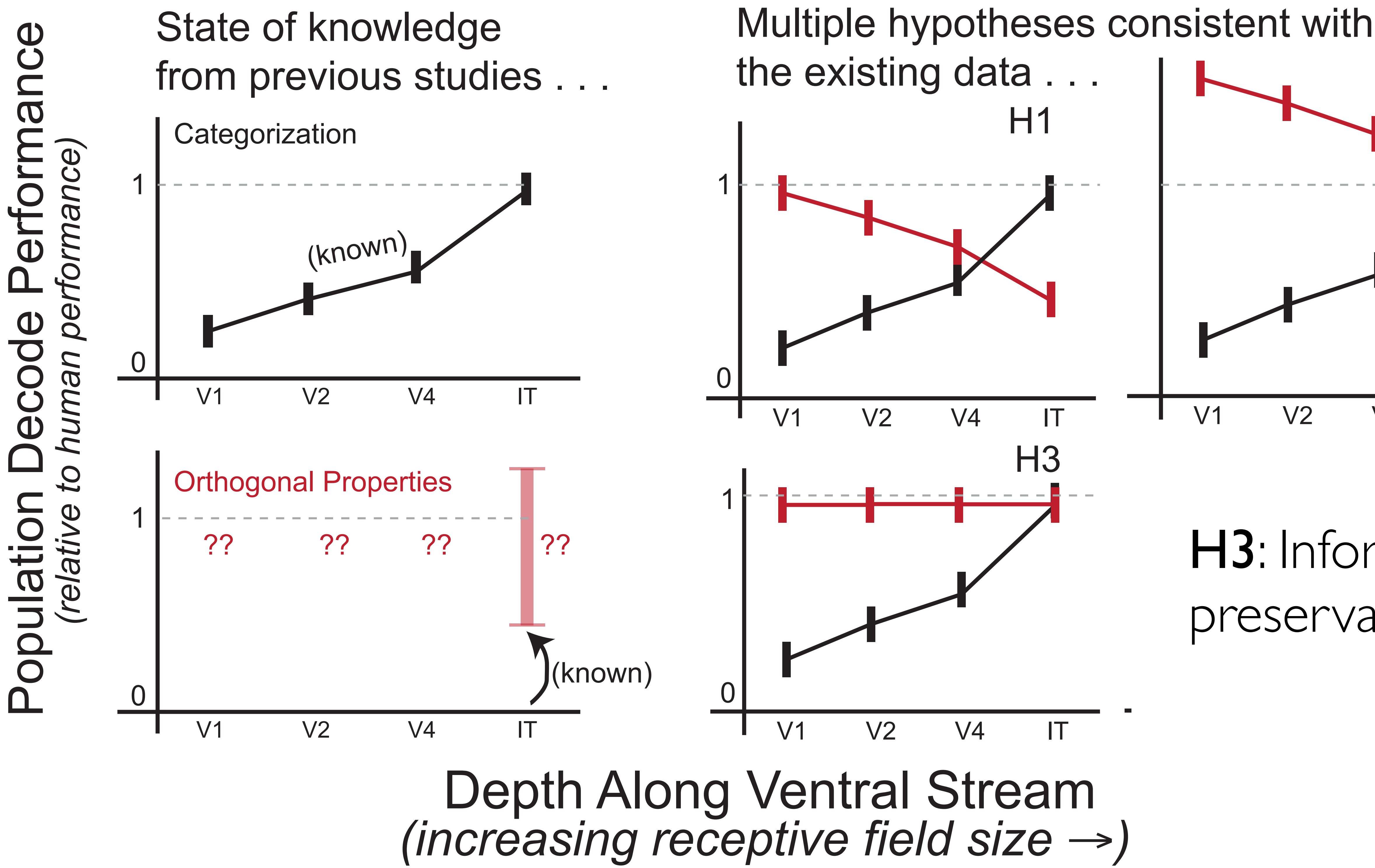


Somewhat newish ideas about IT?



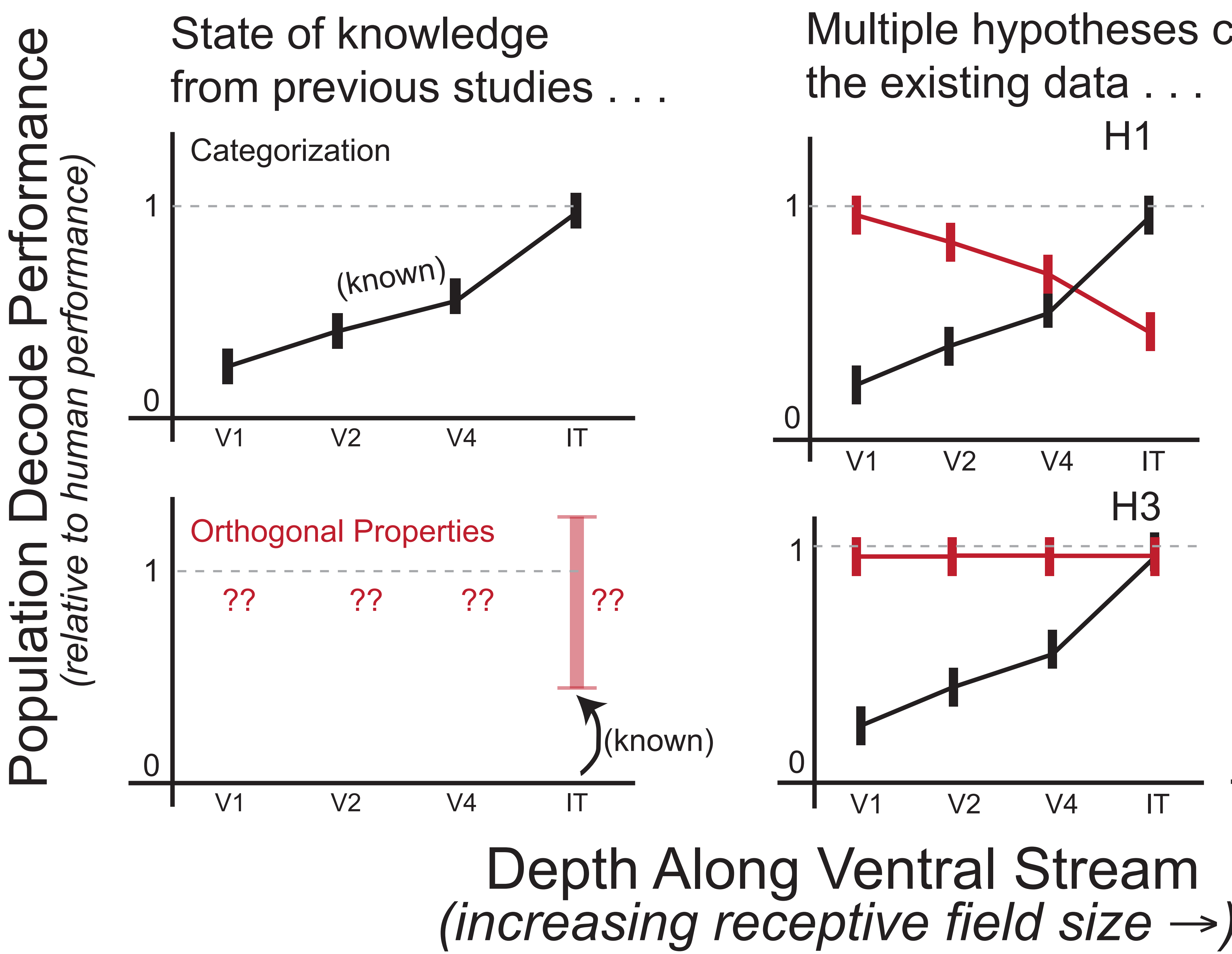
H1: Tolerance / sensitivity tradeoff?

Somewhat newish ideas about IT?

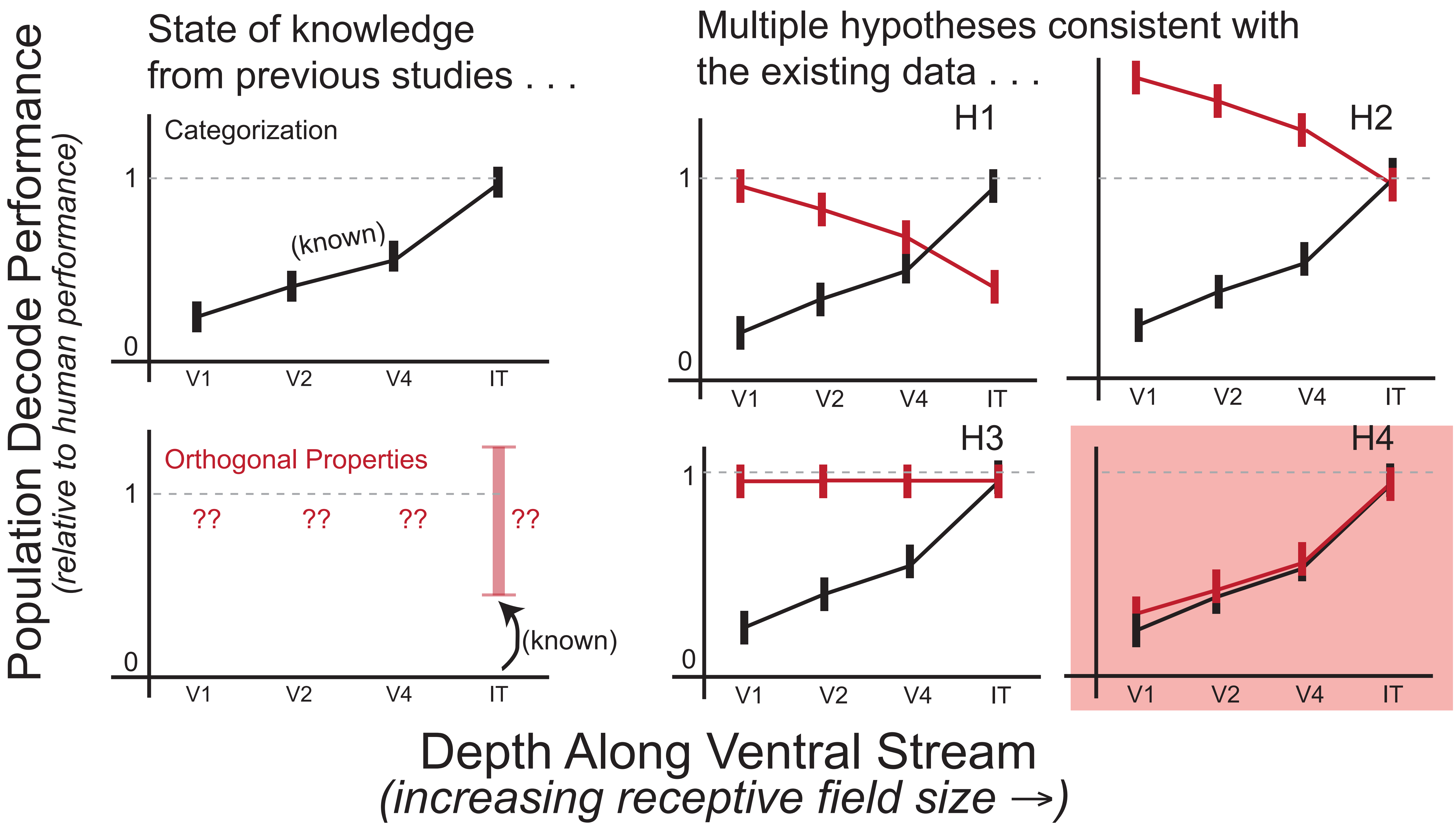


H3: Information preservation?

Somewhat newish ideas about IT?



Somewhat newish ideas about IT?



H4: Simultaneous build-up of encoding

Somewhat newish ideas about IT?

1. IT is *NOT* invariant. Strict generalization of simple-to-complex cells: **no**.

2. "Lower-level" properties are not that low-level — at least, with complex objects and backgrounds.

3. Categorization and non-categorical properties "go together" — *not* just that "not all (e.g.) position information is lost" (MacEvoy 2013, DiCarlo 2003)

Provides support to a hypothesis for what IT does:

"Inverting the generative model of the scene"

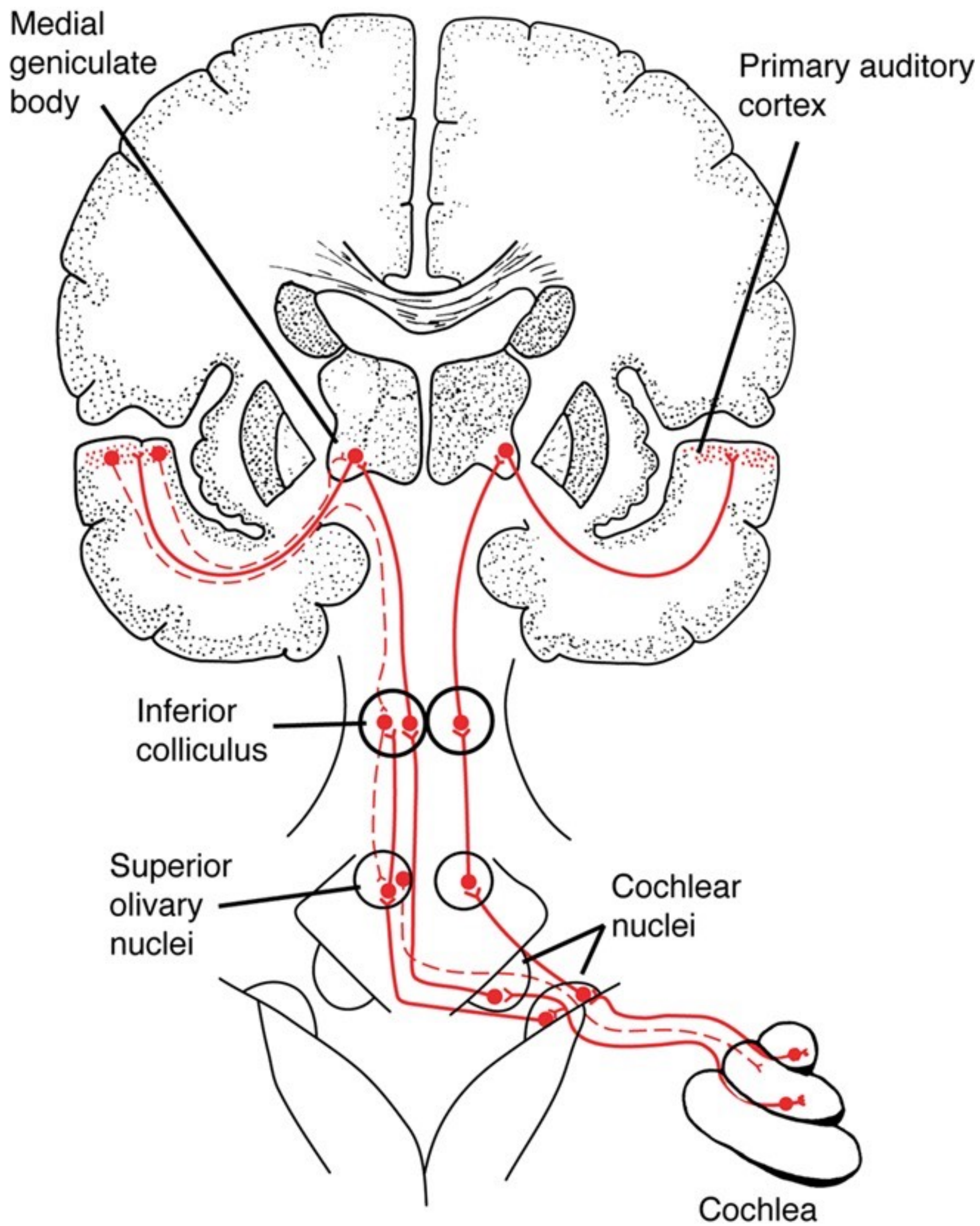
Auditory Cortex

Alex Kell

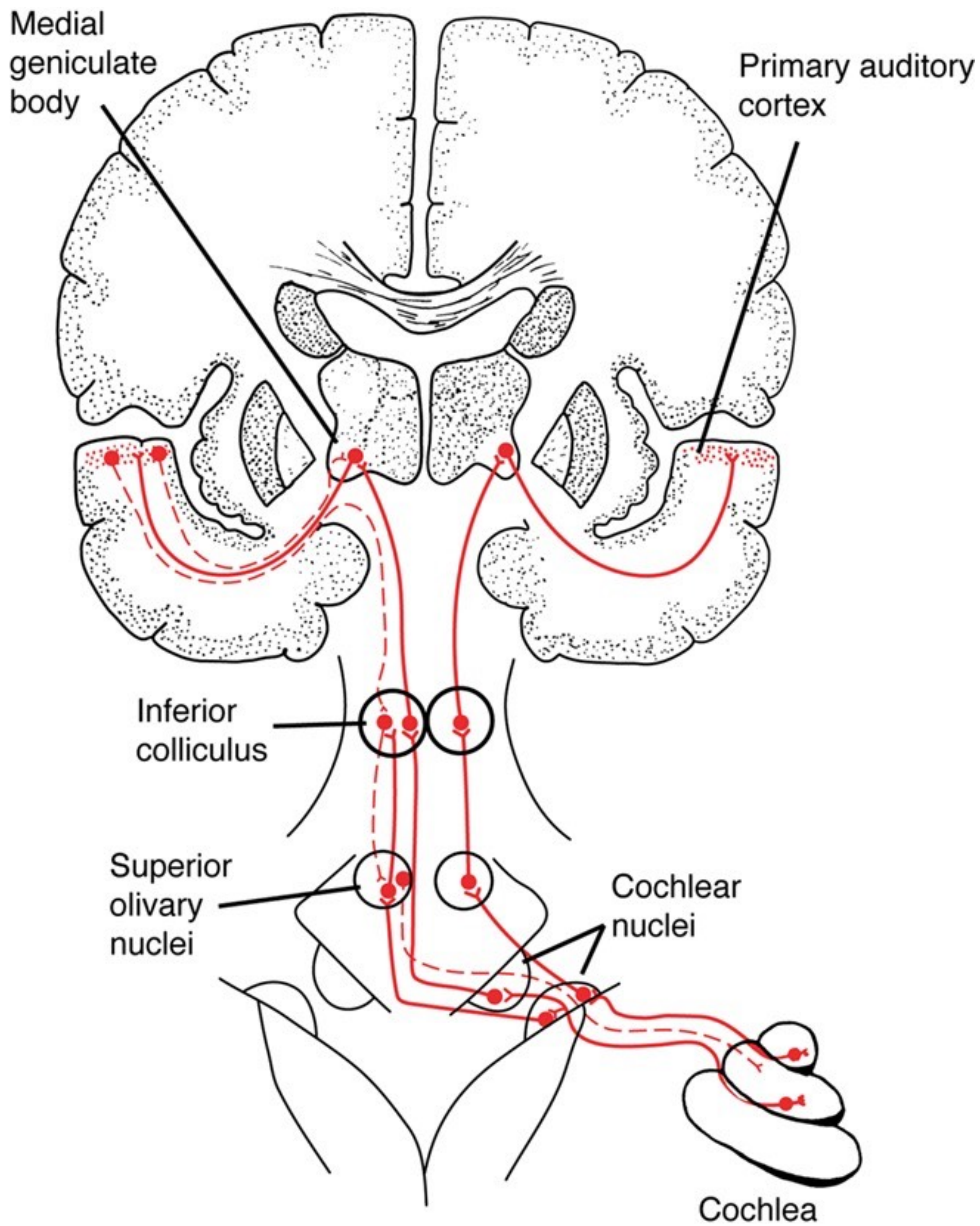
Sam Norman-Haignere

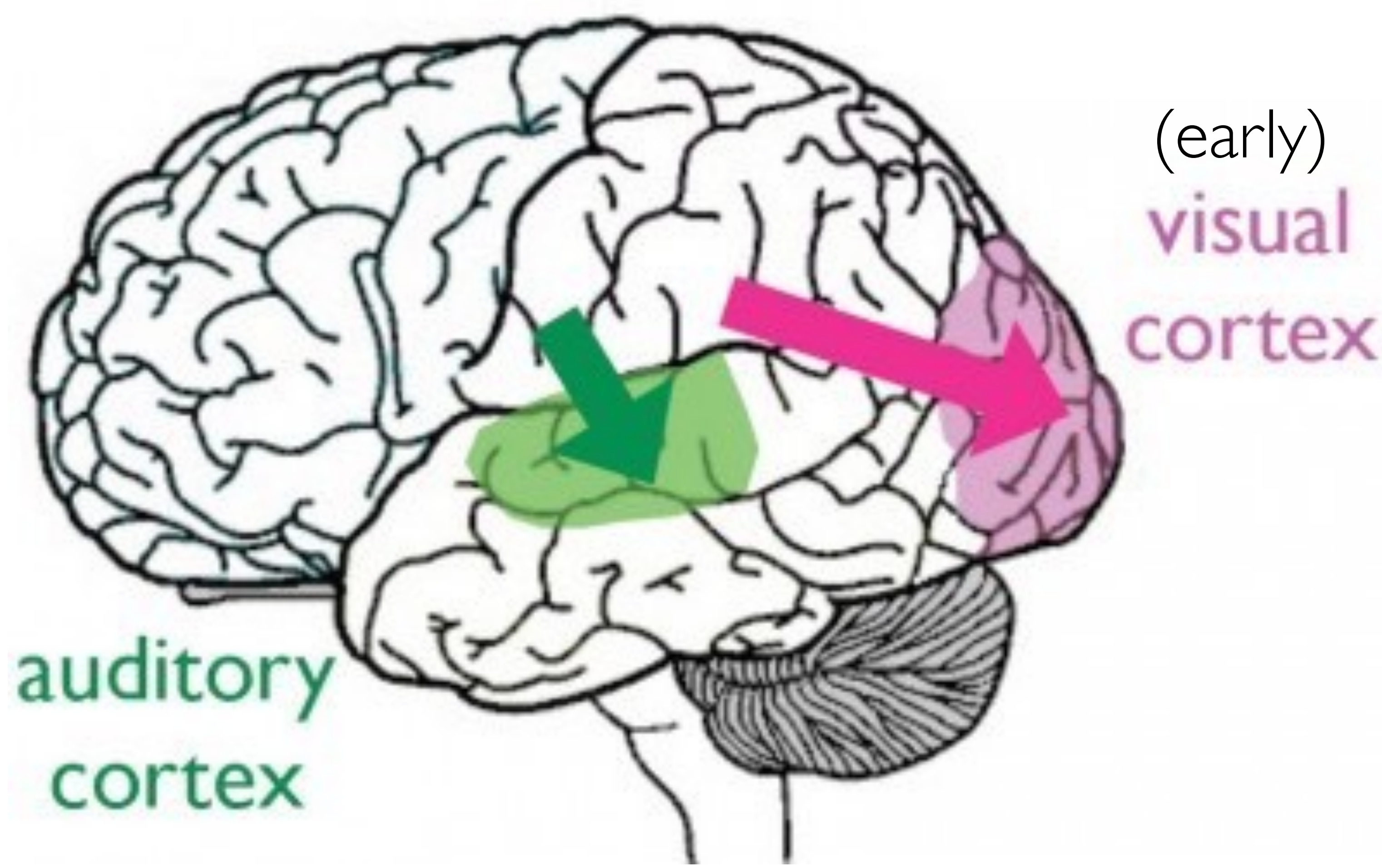
Josh McDermott

Auditory Cortex Background



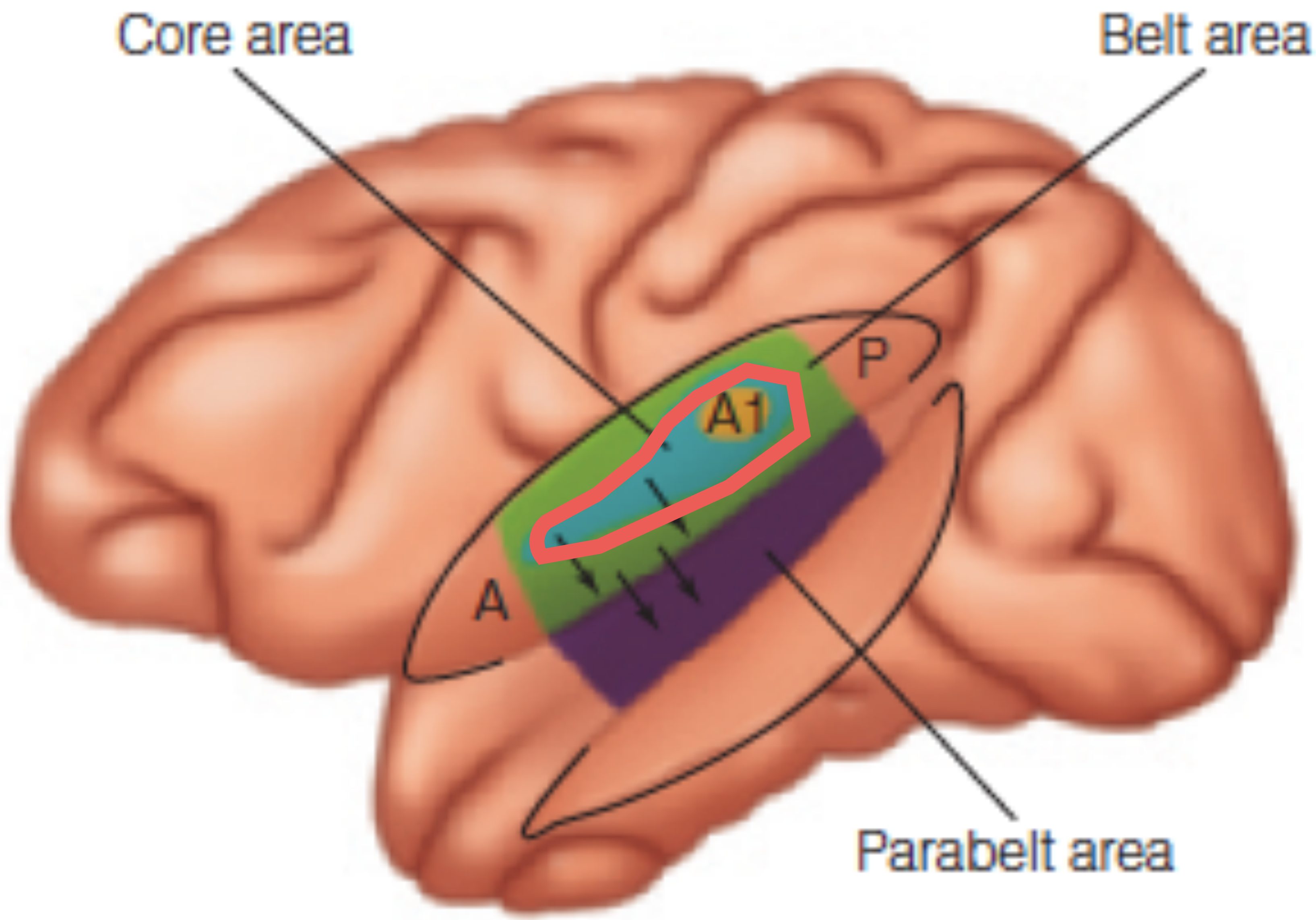
Auditory Cortex Background





How are circuits making sense of complex sound patterns?

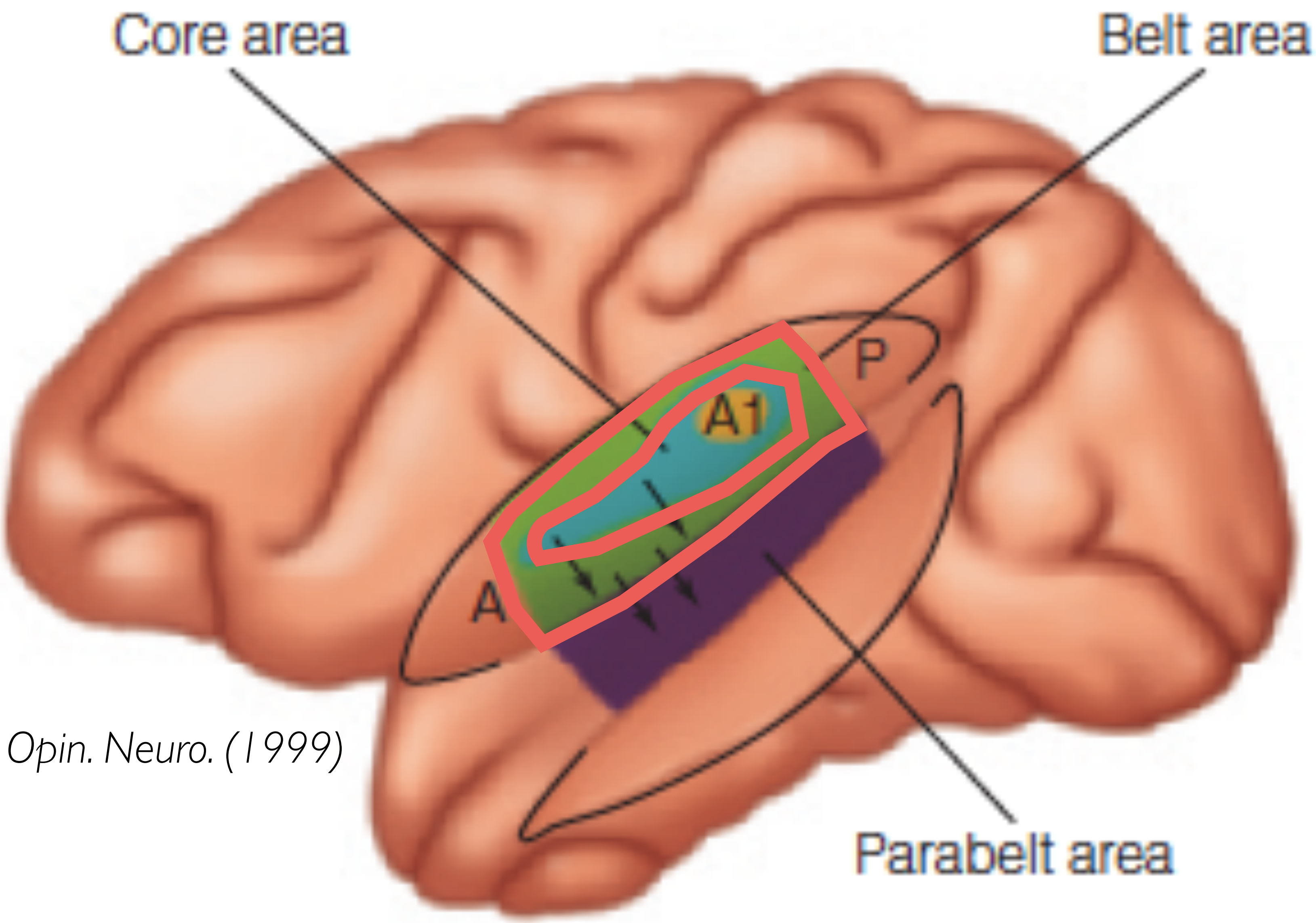
Core / Belt / Parabelt Structure



*monkey
*

Tramo et. al, Curr. Opin. Neuro. (1999)

Core / Belt / Parabelt Structure

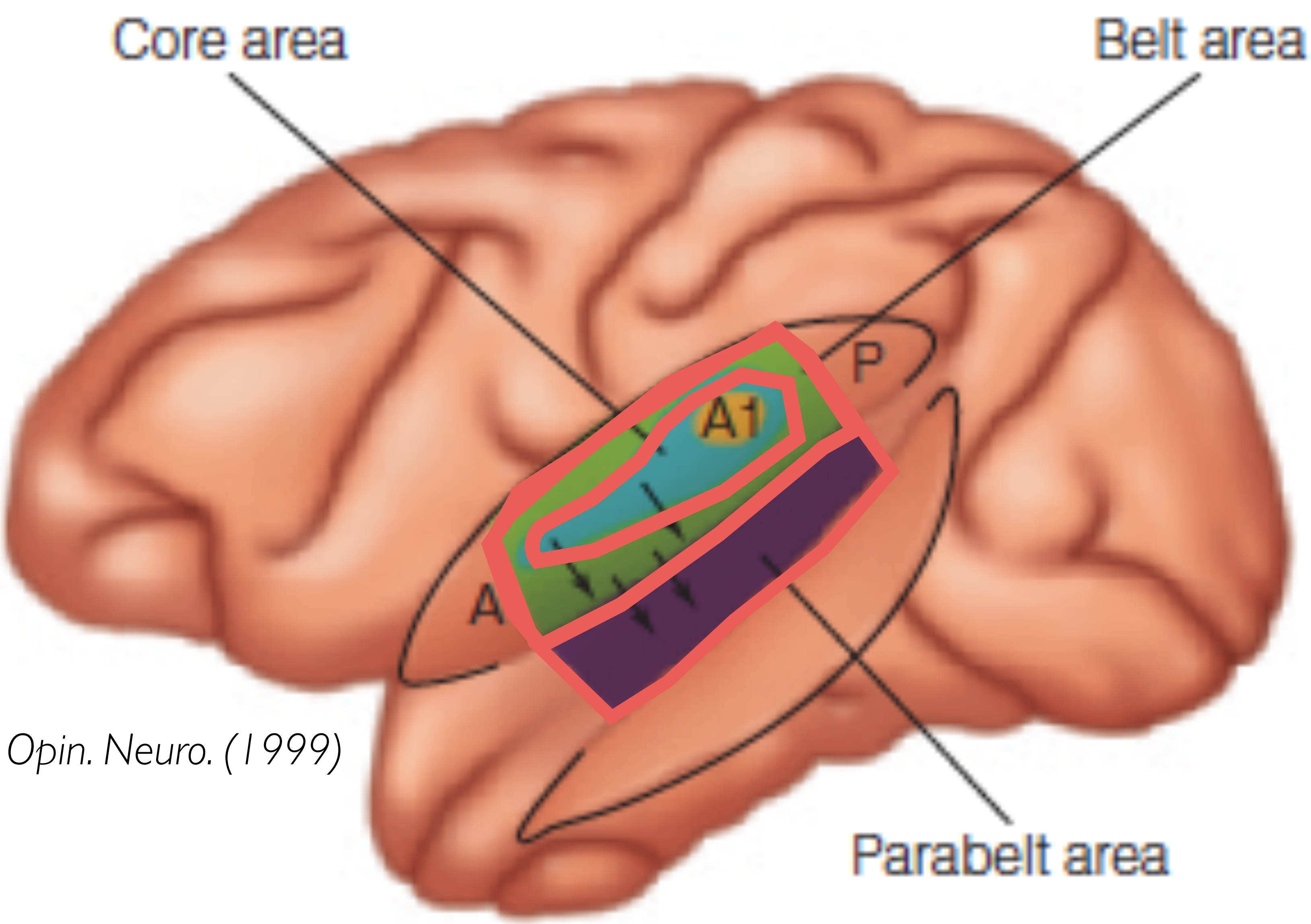


Tramo et. al, Curr. Opin. Neuro. (1999)

*__monkey__

*

Core / Belt / Parabelt Structure

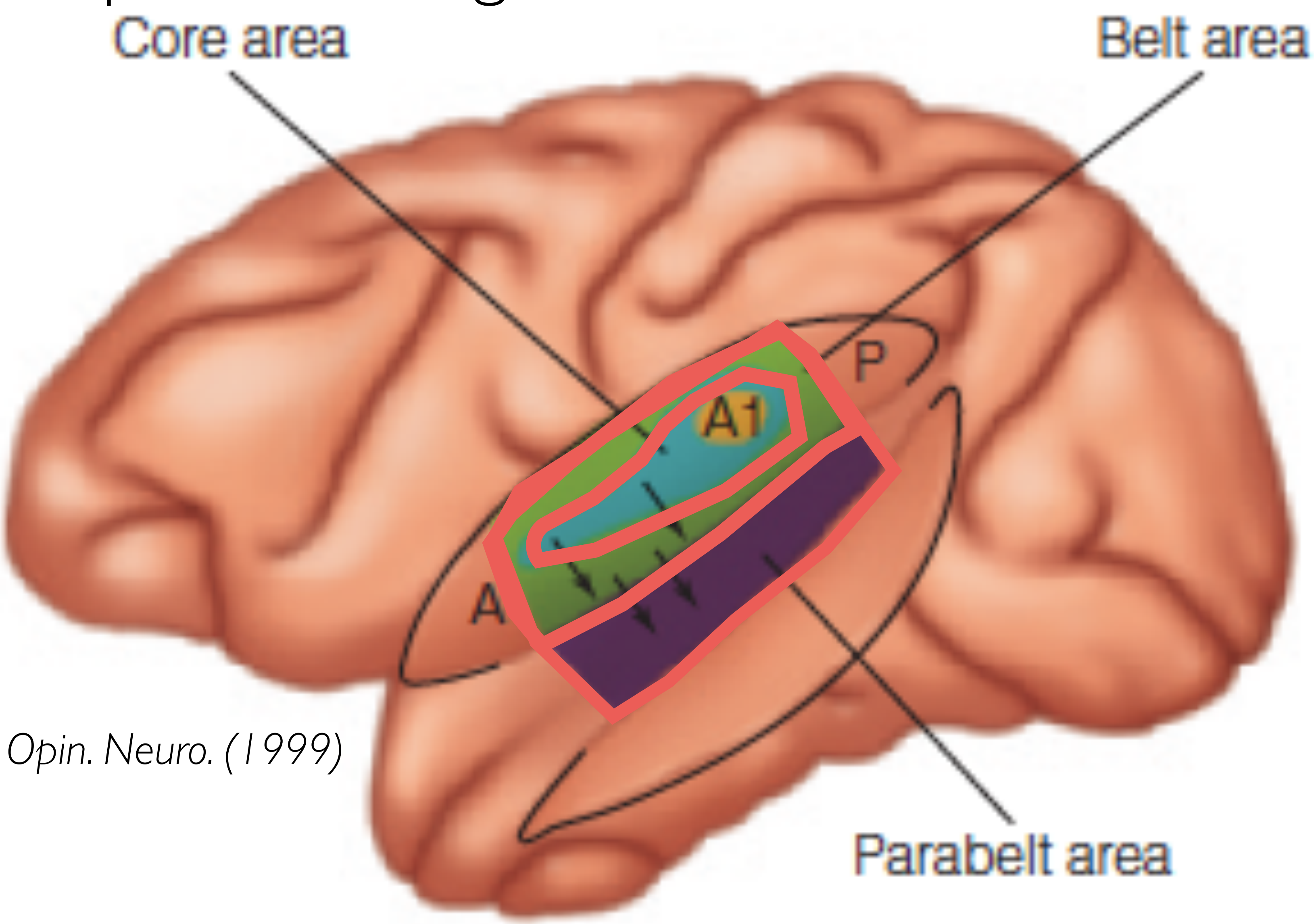


Tramo et. al, Curr. Opin. Neuro. (1999)

* **monkey**
*

Core / Belt / Parabelt Structure

Spatiotemporal filtering? *Shamma, 2005*

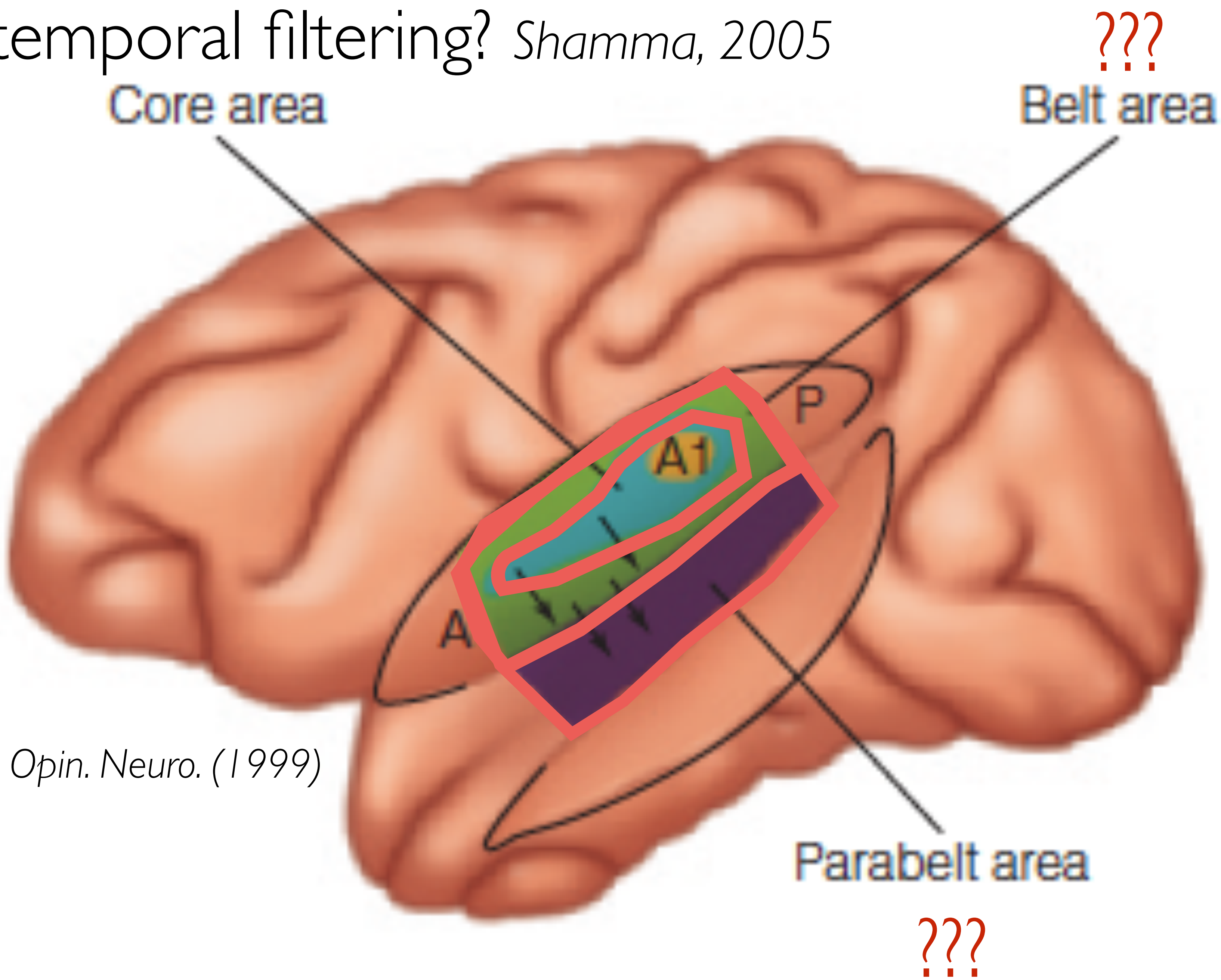


Tramo et. al, Curr. Opin. Neuro. (1999)

* **monkey**
*

Core / Belt / Parabelt Structure

Spatiotemporal filtering? *Shamma, 2005*



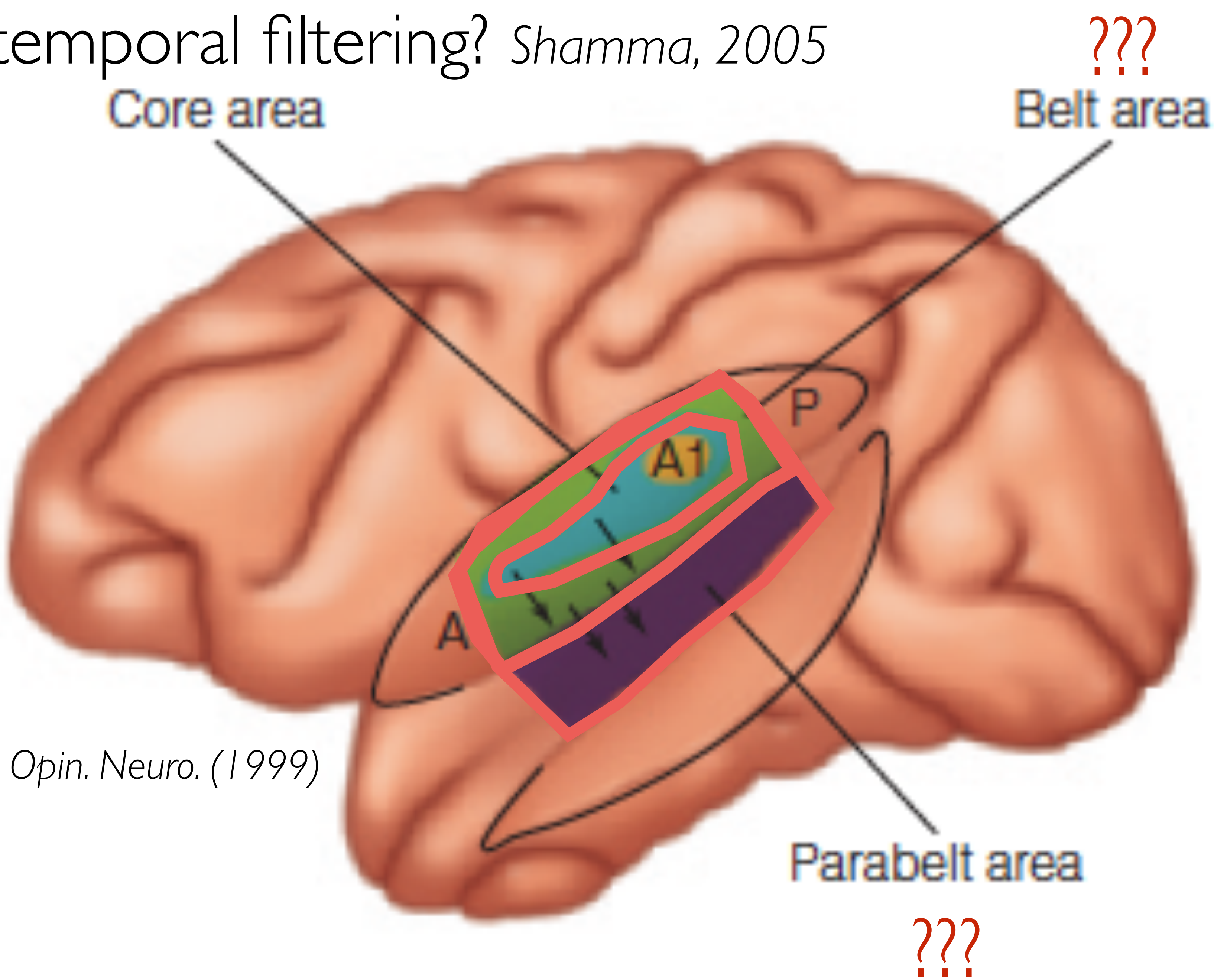
Tramo et. al, Curr. Opin. Neuro. (1999)

*monkey

*

Core / Belt / Parabelt Structure

Spatiotemporal filtering? *Shamma, 2005*



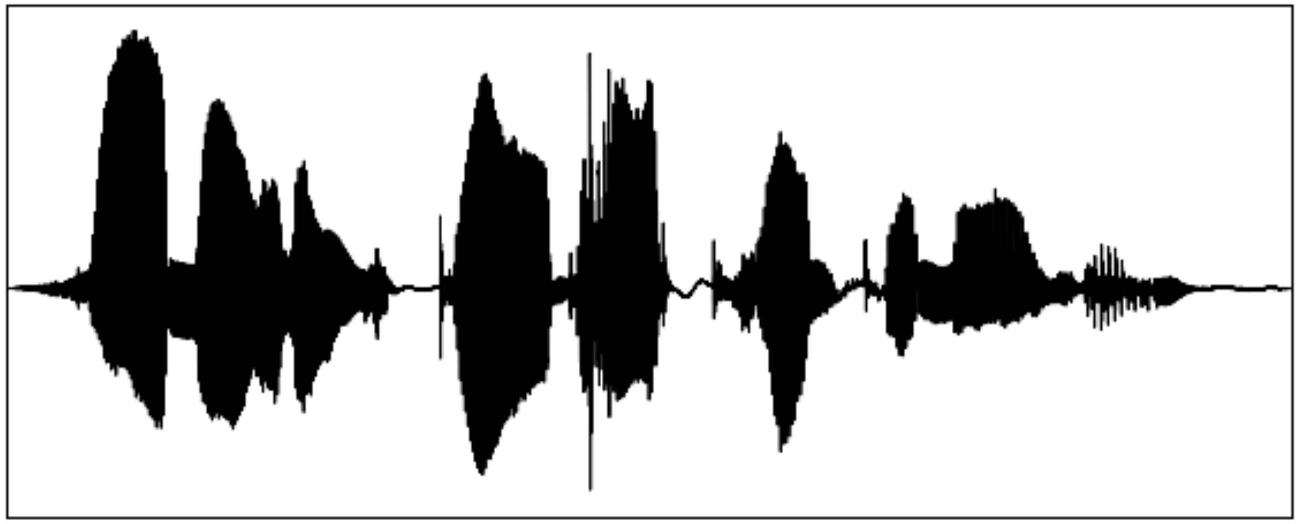
Tramo et. al, Curr. Opin. Neuro. (1999)

*monkey
*

Our goal: use computational models to help deepen understanding of non-primary areas.

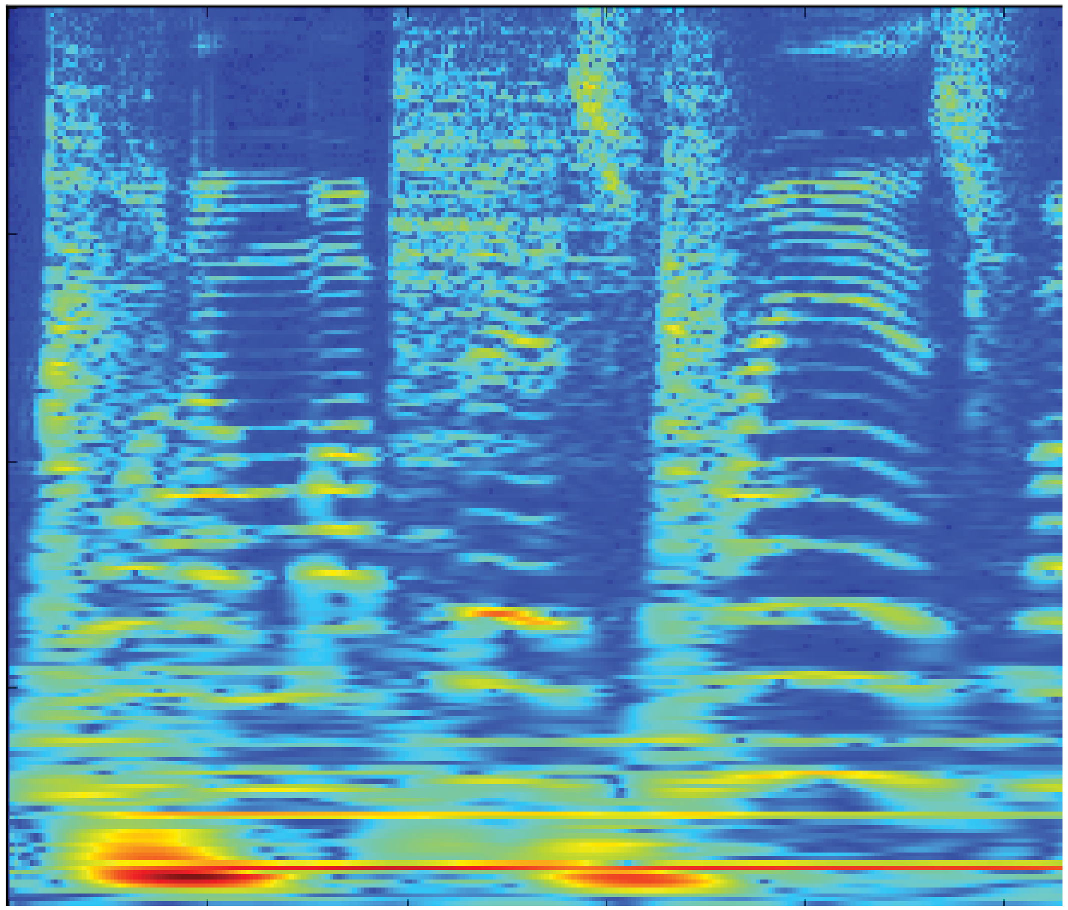
Convolutional Neural Networks

Waveform representation



Time →

Cochleagram representation

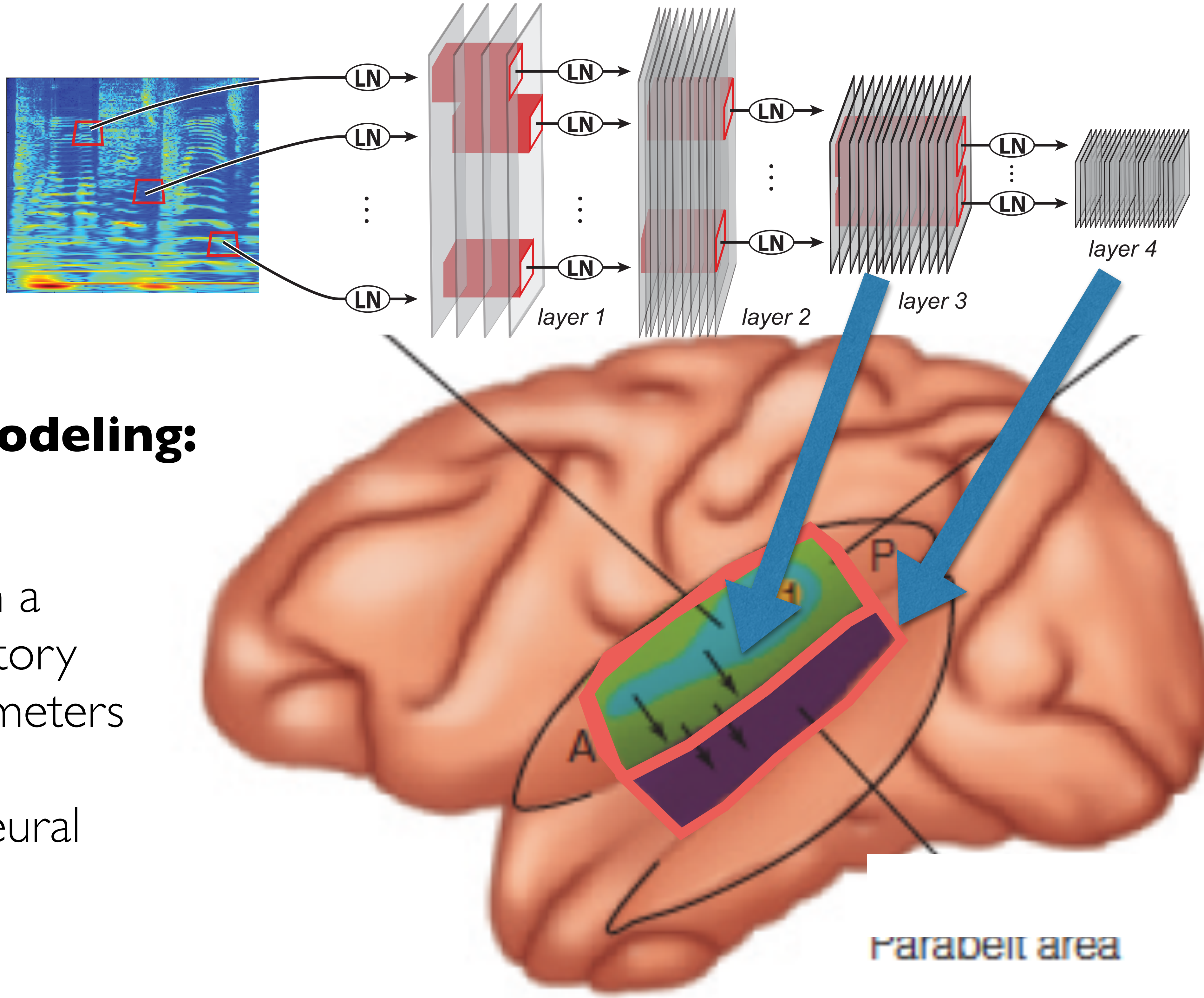


Frequency →

Time →

Coarse model of the cochlea

Core Task-Driven Modeling Idea



Task-Driven Modeling:

1. Optimize for performance on a challenging auditory task, fixing parameters

2. Compare to neural data.

Apply to auditory tasks, where the regions themselves are less well known.

Optimize for Performance: The Task

600-way word-recognition task assembled by:

- Recordings from standard speech recognition databases (TIMIT, WSJ) with words spoken at least 20 times

- Combined with significant background noise

  - auditory scenes
  - speech babble
  - music clips

*"She **had** your*
*dark **suit** in*
*greasy **wash** water*
*all **year** … "*

'had'
'suit'
'wash'
'year'

Optimize for Performance: The Task

600-way word-recognition task assembled by:

- Recordings from standard speech recognition databases (TIMIT, WSJ) with words spoken at least 20 times

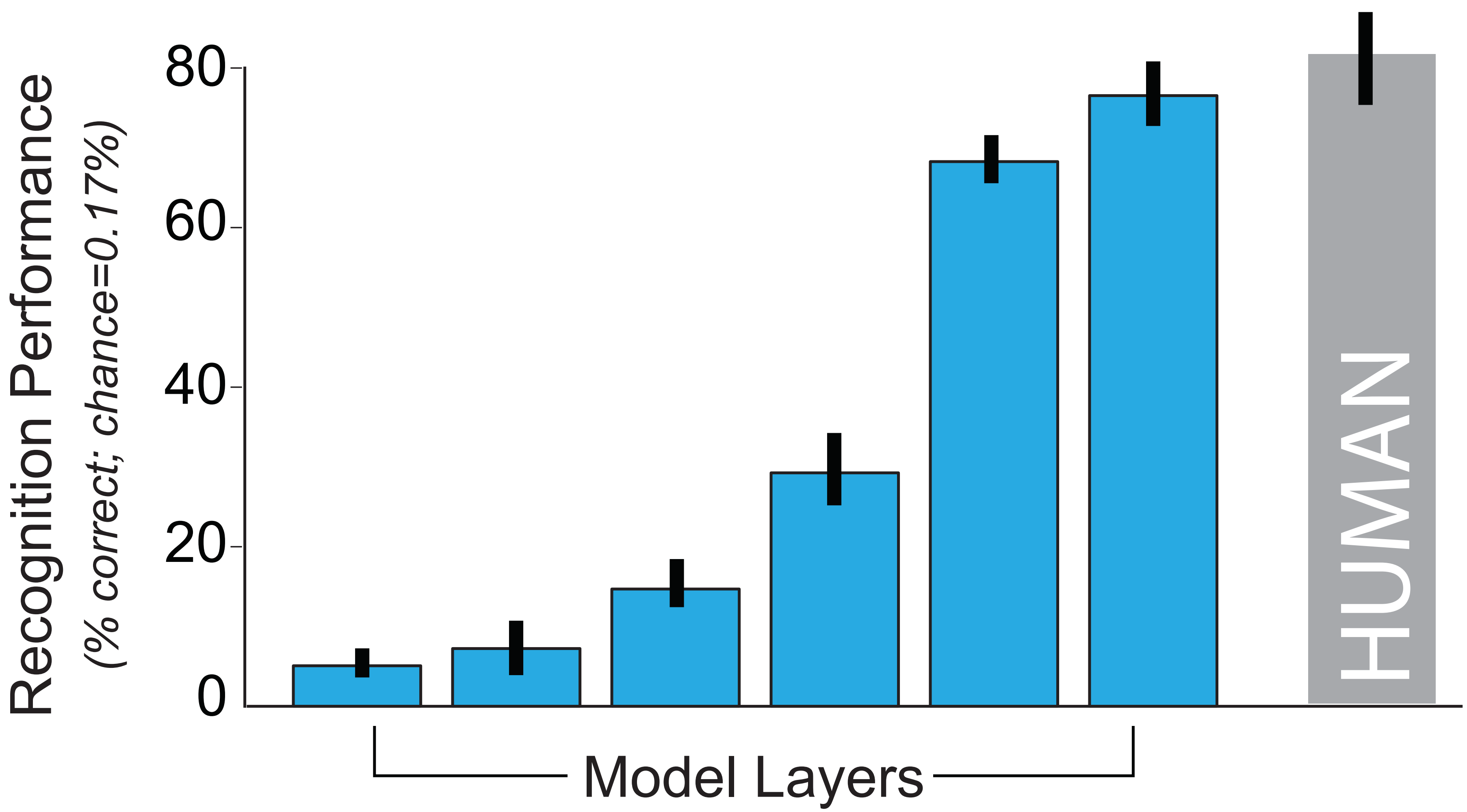
- Combined with significant background noise

  - auditory scenes
  - speech babble
  - music clips

Backgrounds → humans not close to ceiling.

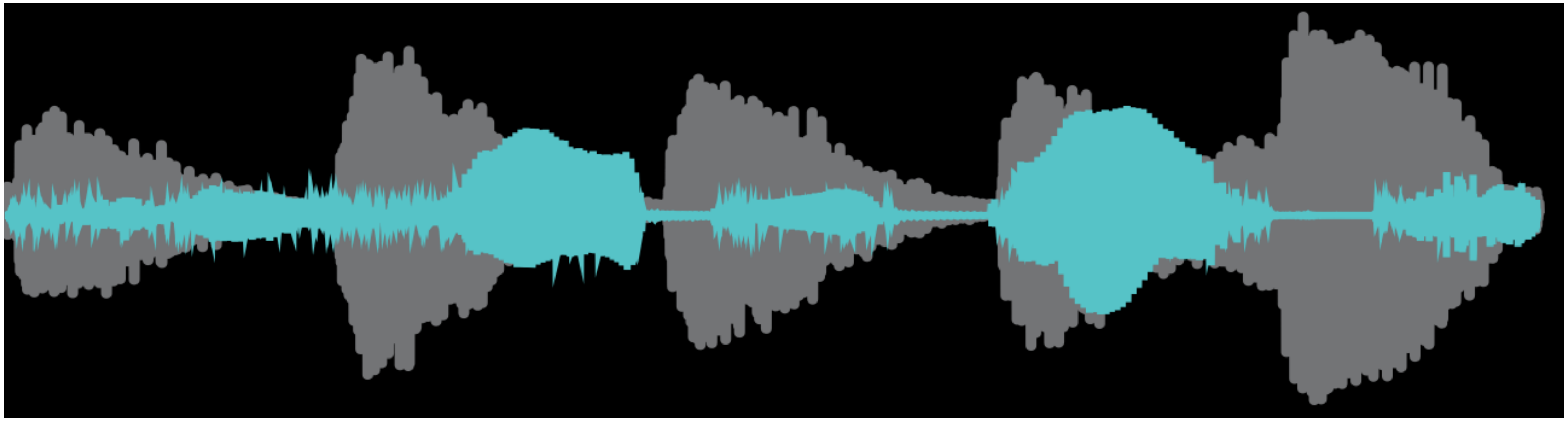
Performance Results

Performance on 600-way word-recognition task



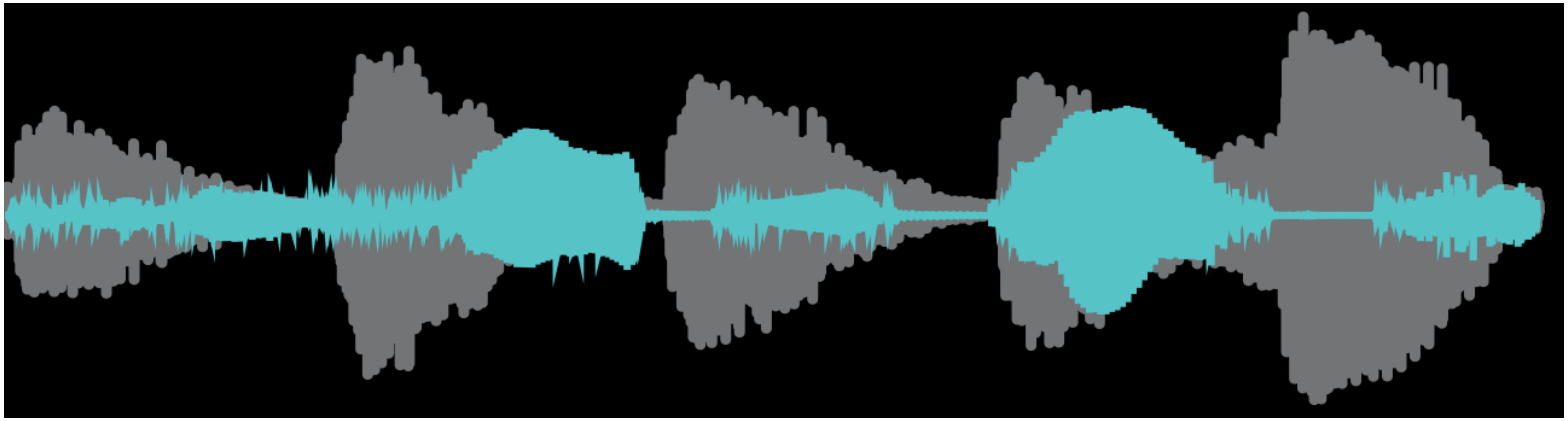
… for model, measured on held-out data with novel speakers and auditory background noise.

Behavioral comparison: CNN & humans on same task



Word recognition in complex backgrounds

Behavioral comparison: CNN & humans on same task



Word recognition in complex backgrounds

21 conditions:

dry

+

4 different background types at 5 SNR levels:

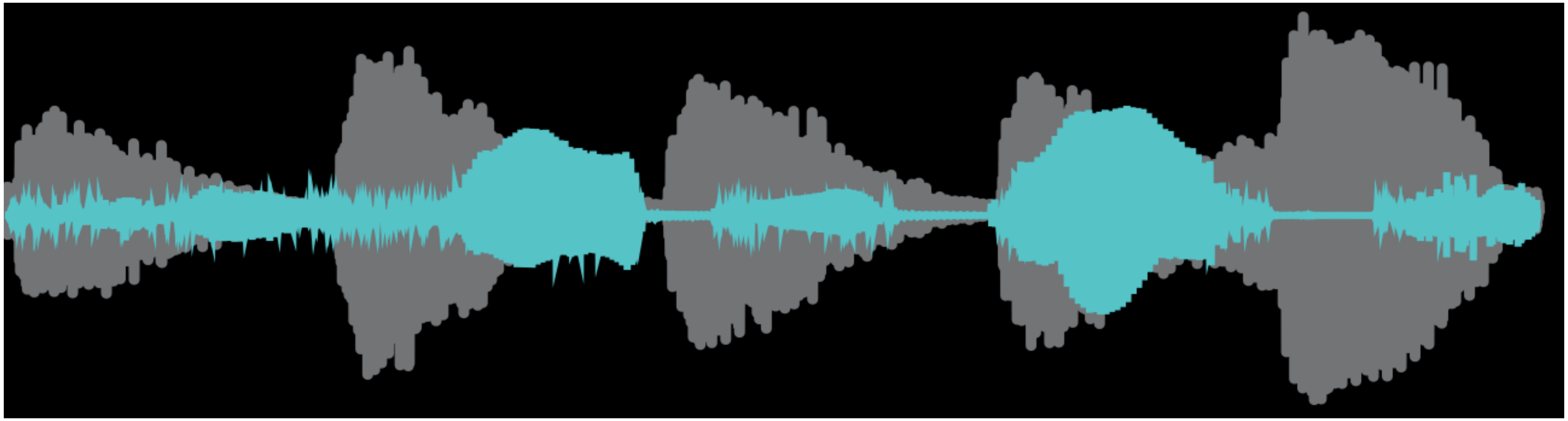
Auditory scenes

Speech babble

Music

Speech-shaped noise

Behavioral comparison:
CNN & humans on same task



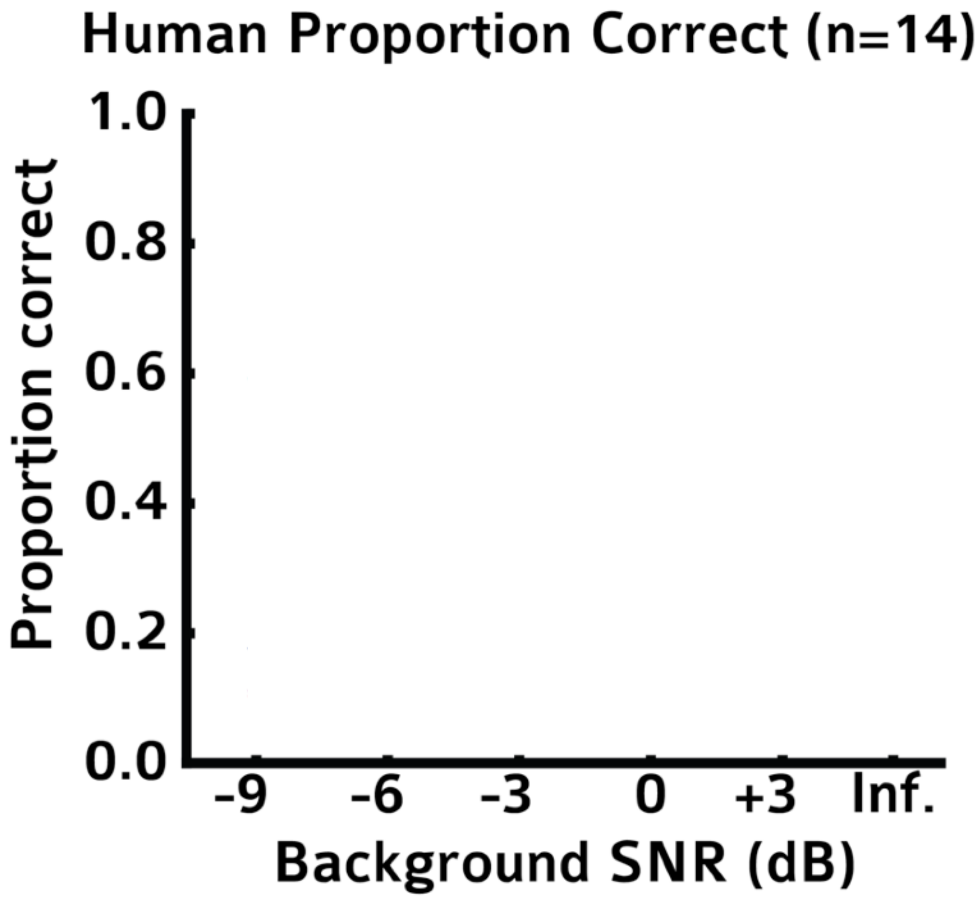
Word recognition in complex backgrounds

<u>21 conditions:</u>
dry
+
4 different background types at 5 SNR levels:

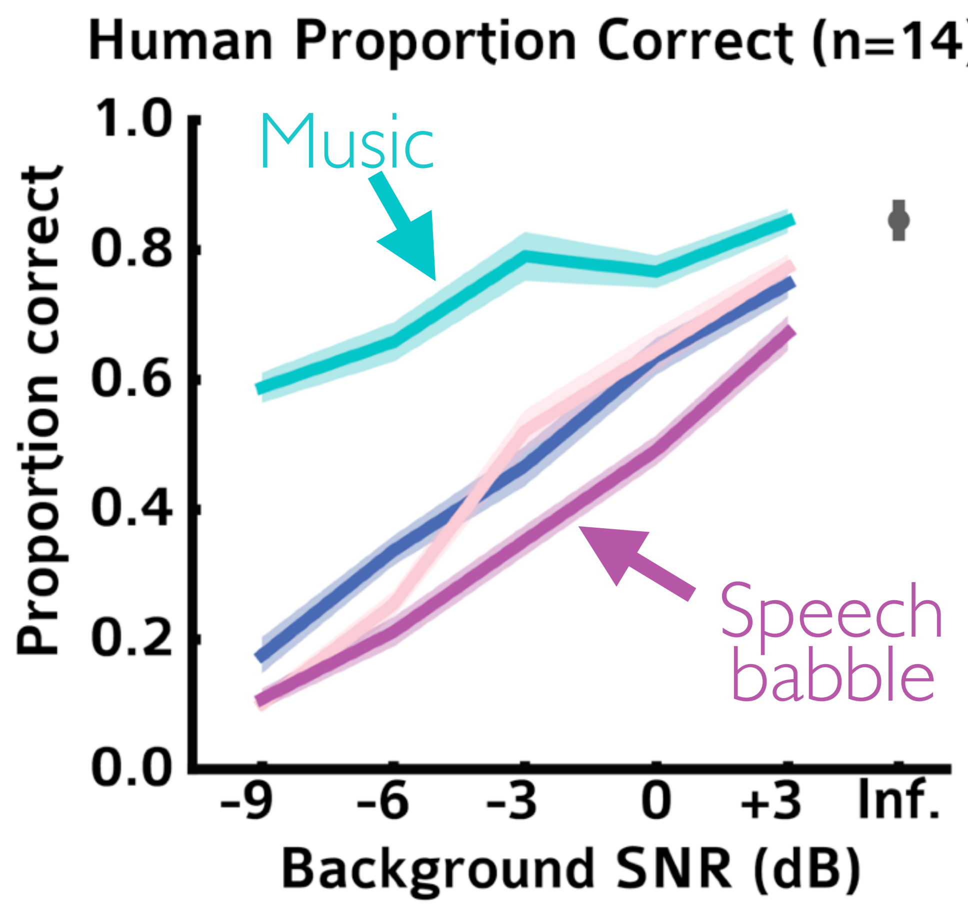
600
AFC

Auditory scenes   Speech babble
Music   Speech-shaped noise

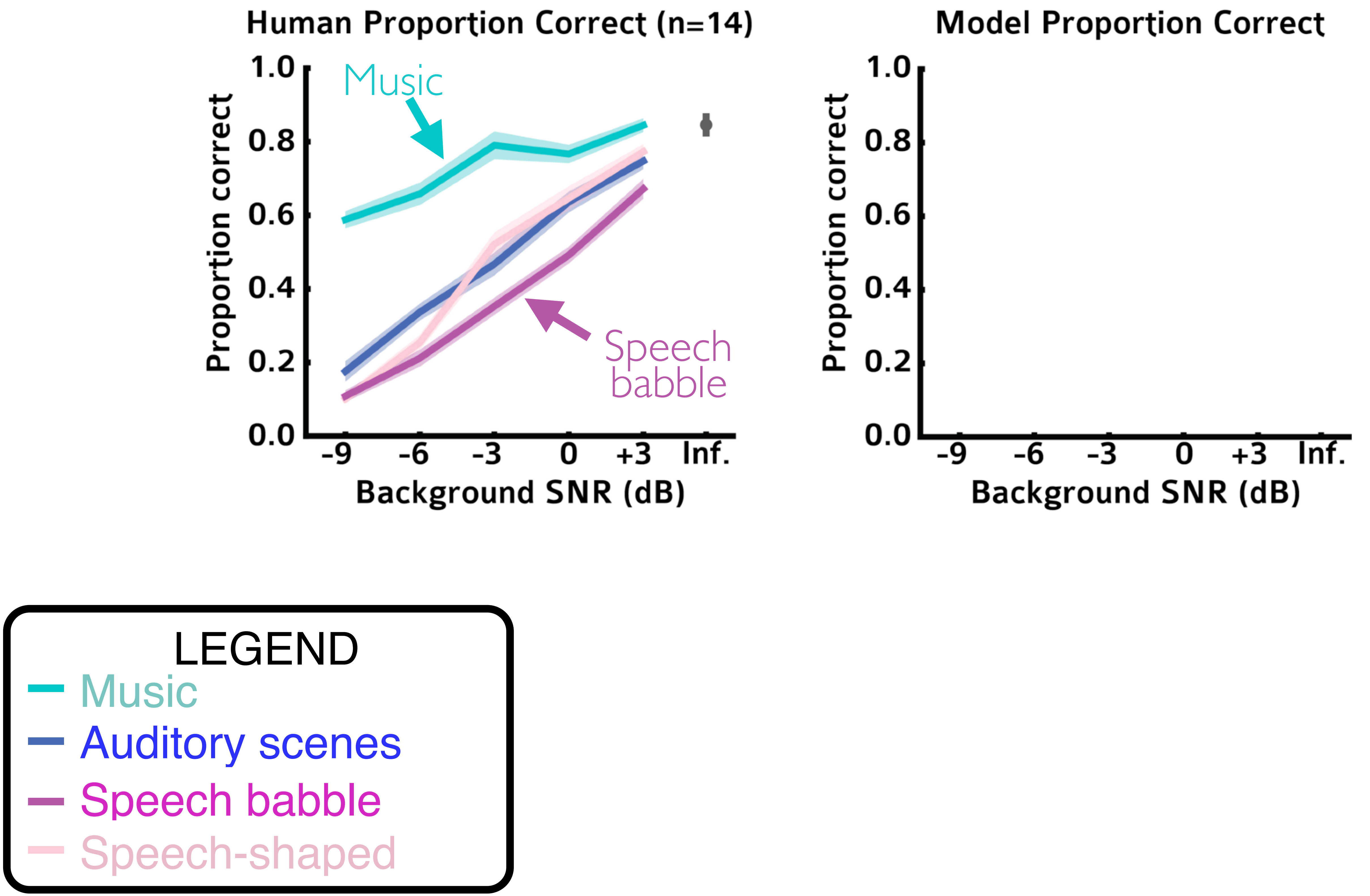
Behavioral comparison: CNN & humans on same task



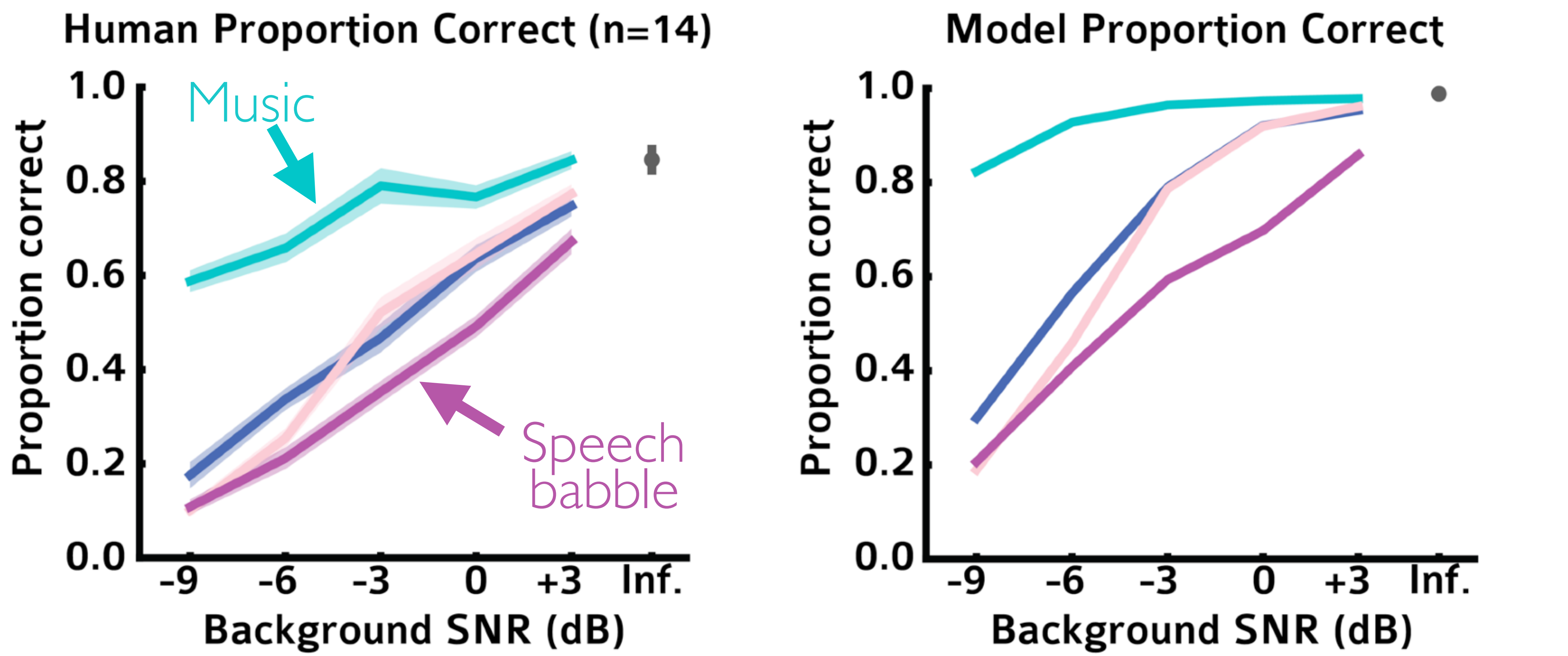
Behavioral comparison: CNN & humans on same task



Behavioral comparison: CNN & humans on same task

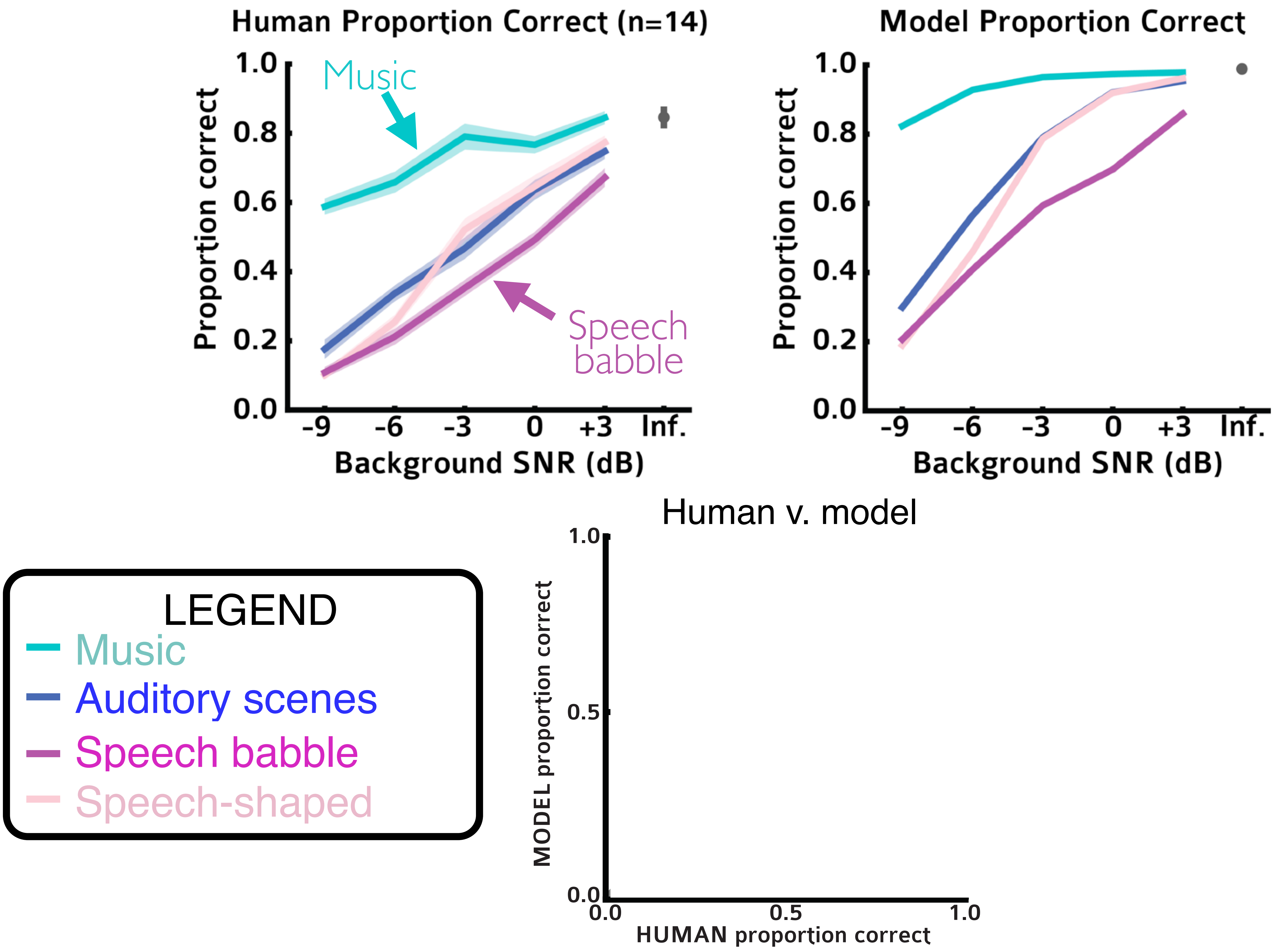


Behavioral comparison: CNN & humans on same task

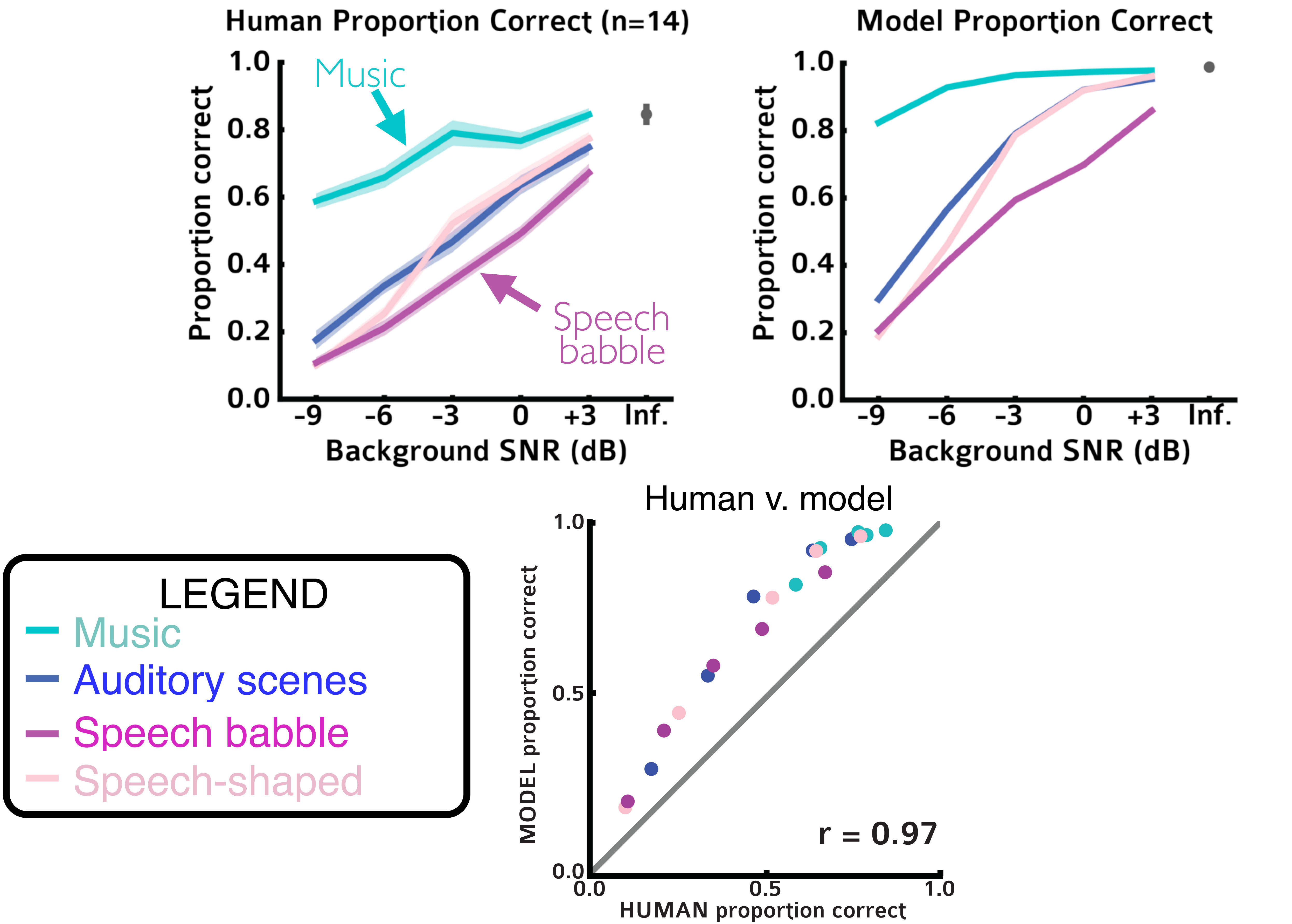


LEGEND
- Music
- Auditory scenes
- Speech babble
- Speech-shaped

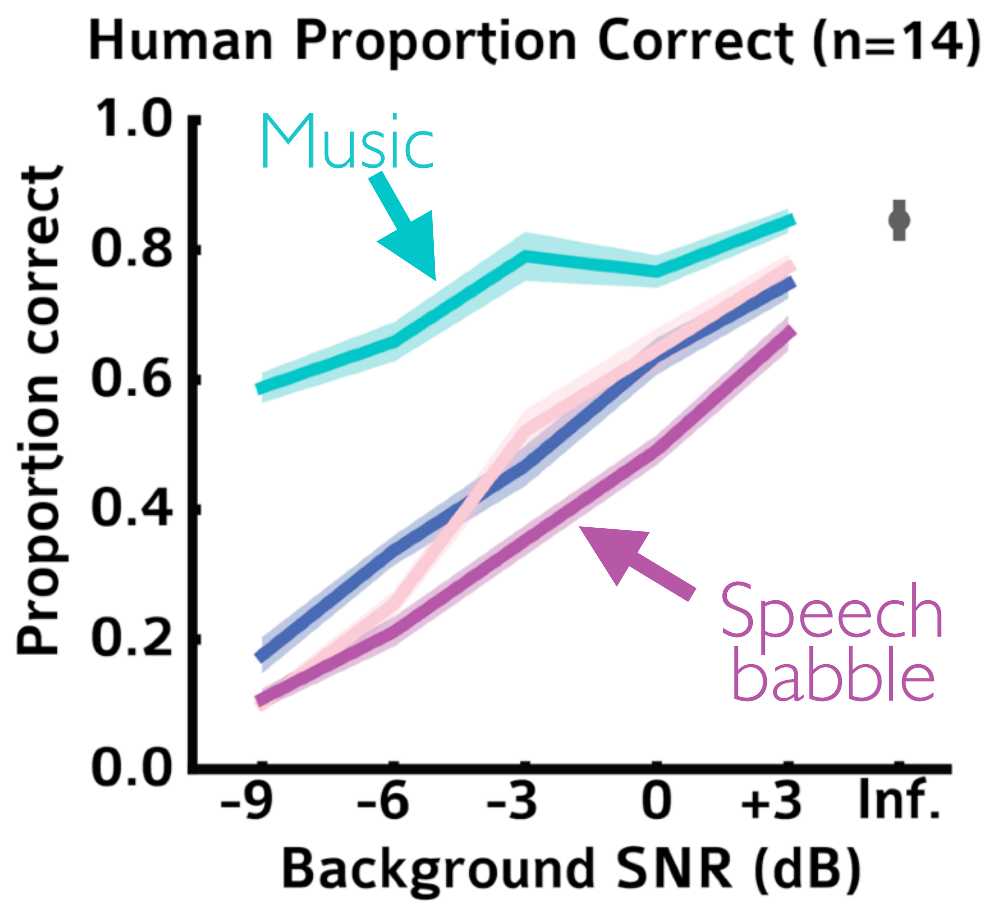
Behavioral comparison: CNN & humans on same task

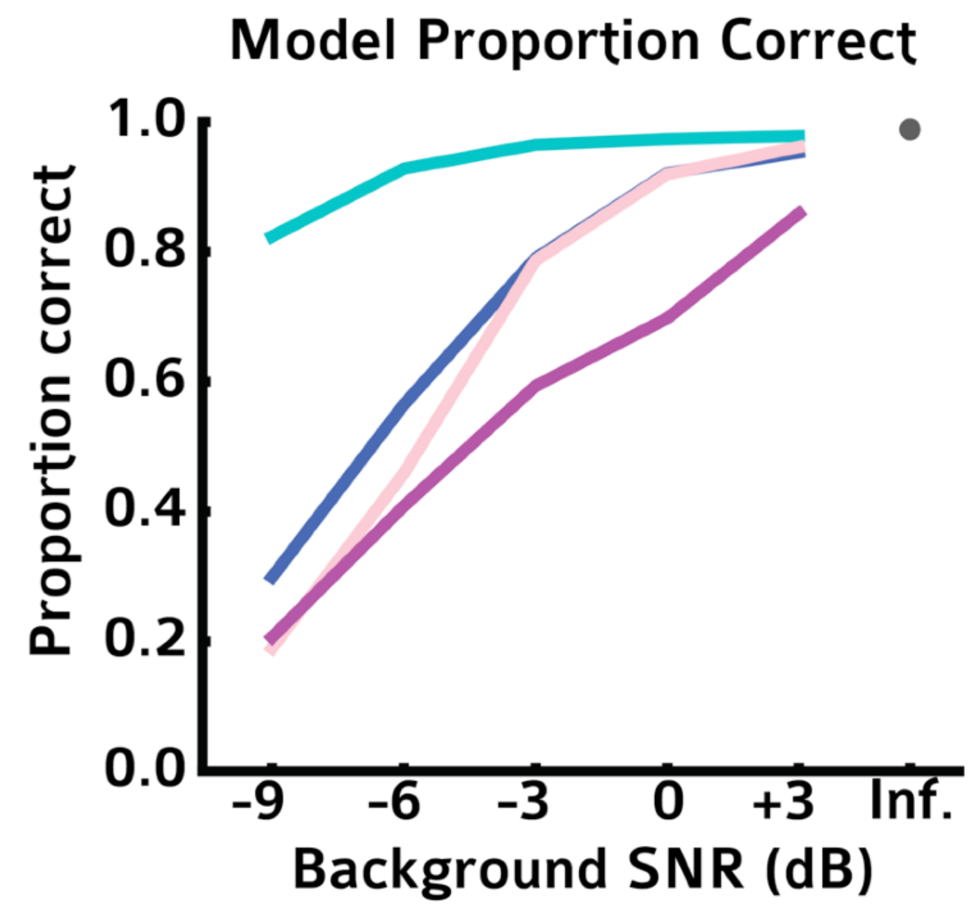


Behavioral comparison: CNN & humans on same task

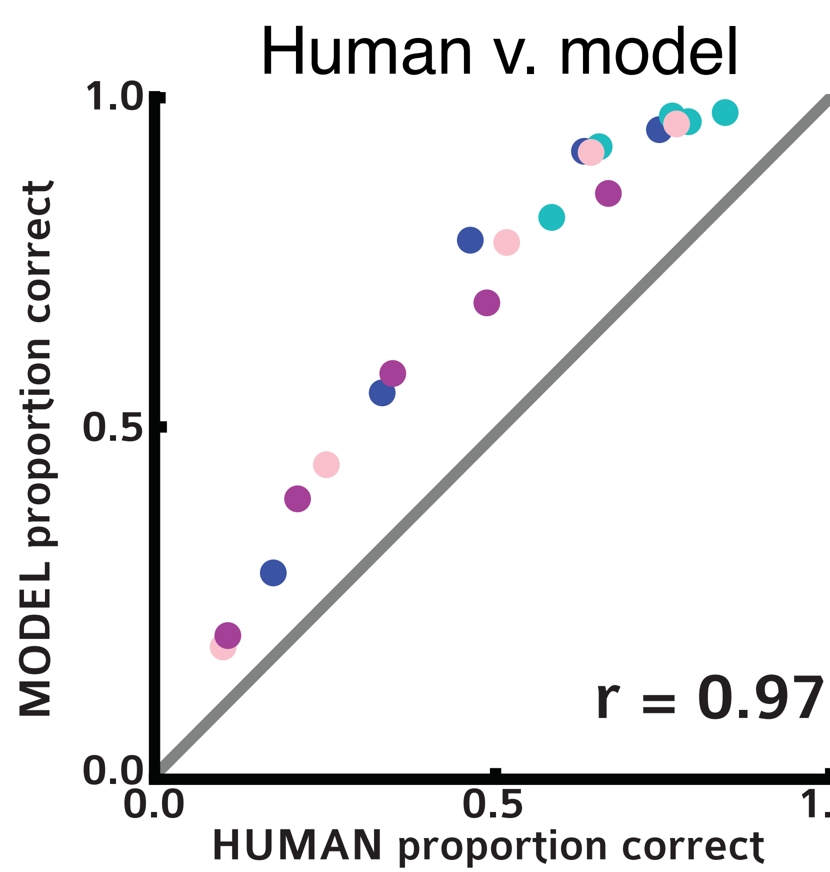


Behavioral comparison: CNN & humans on same task





LEGEND
- Music
- Auditory scenes
- Speech babble
- Speech-shaped

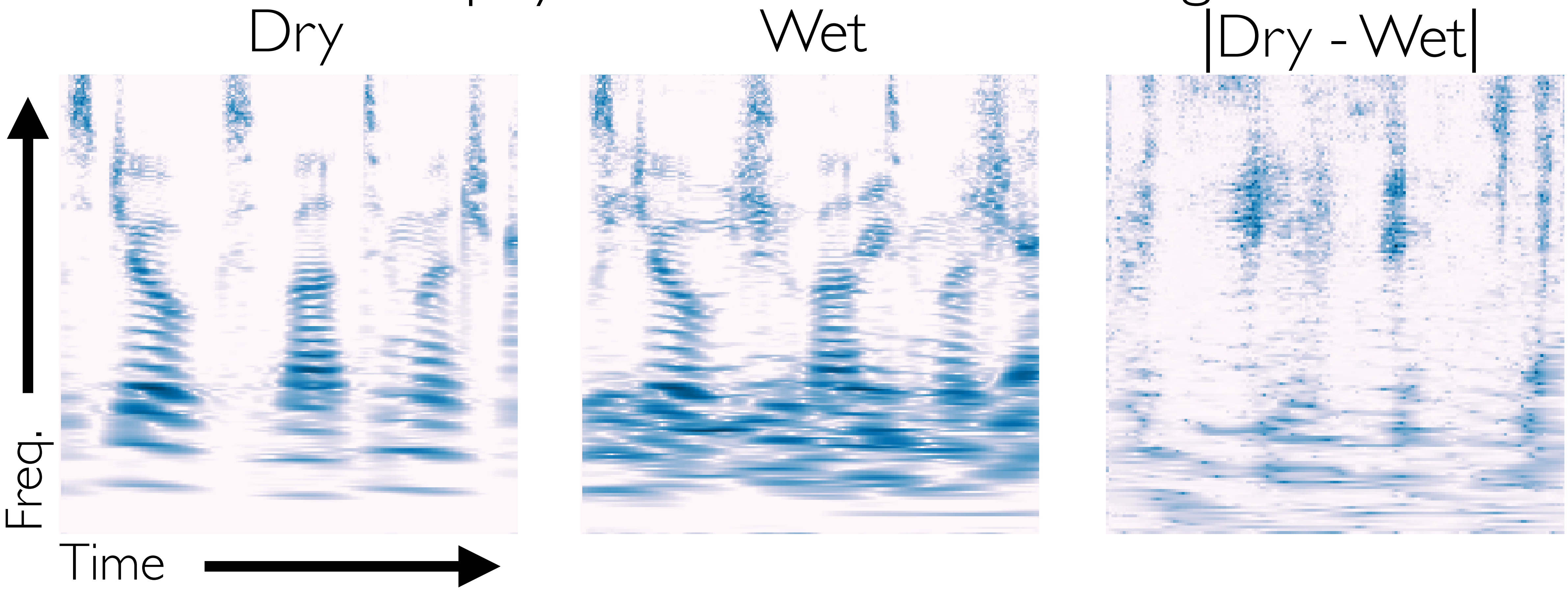


NB:

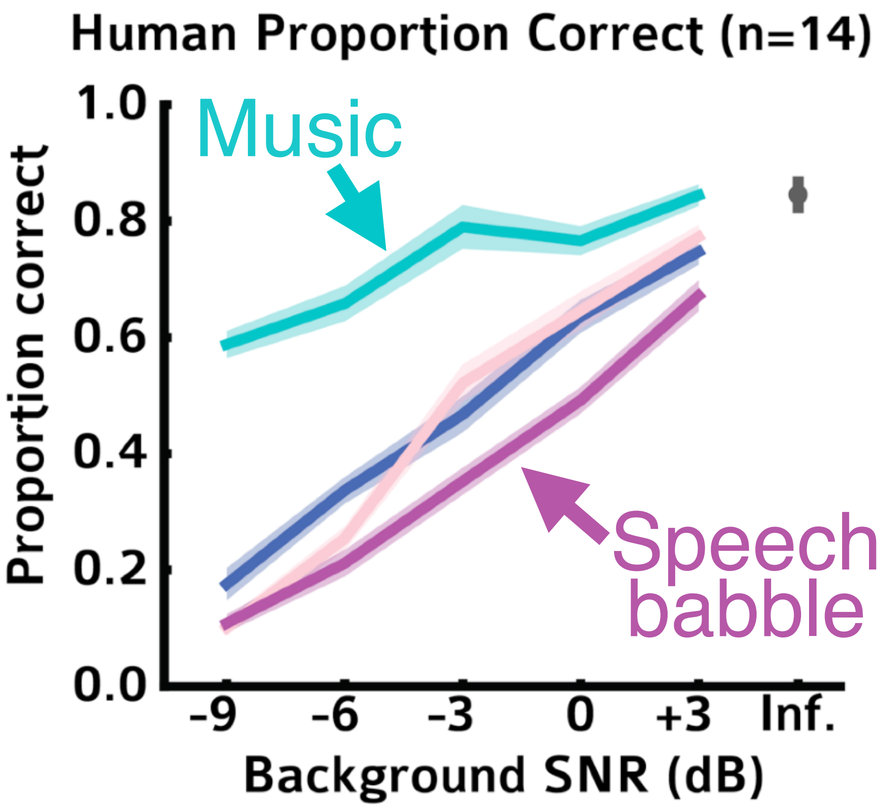
CNN optimized for task performance ***not*** for human behavior match

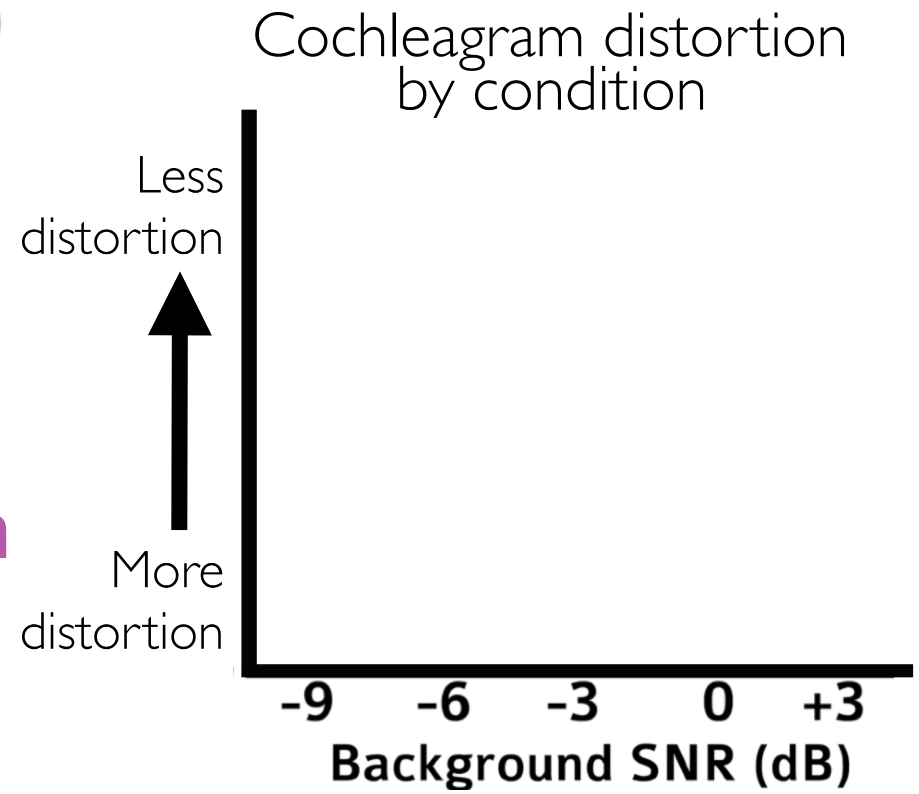
Does distortion in a periphery-like representation explain pattern of performance?

Measure physical distortion of background noise



Distortion and pattern of performance.

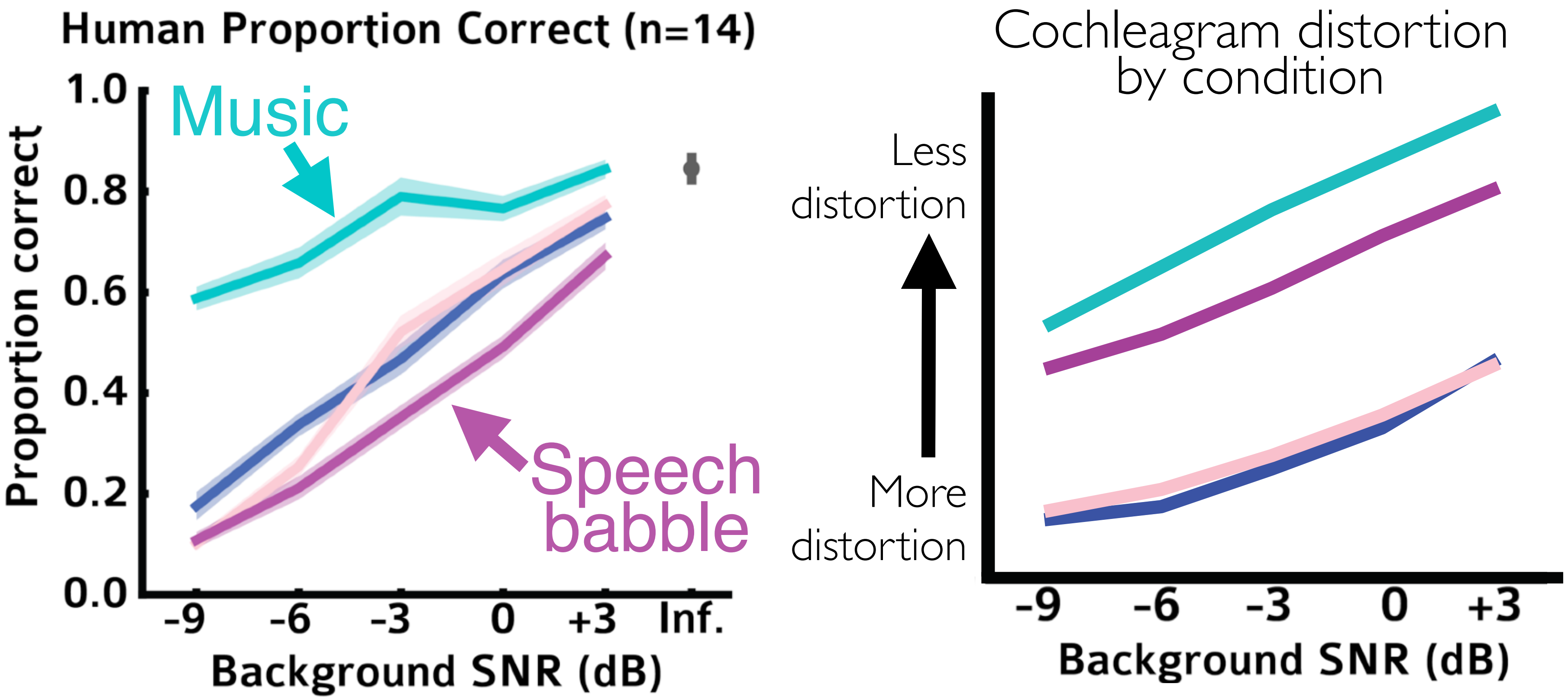




LEGEND

Music   Auditory scenes   Speech babble   Speech-shaped noise

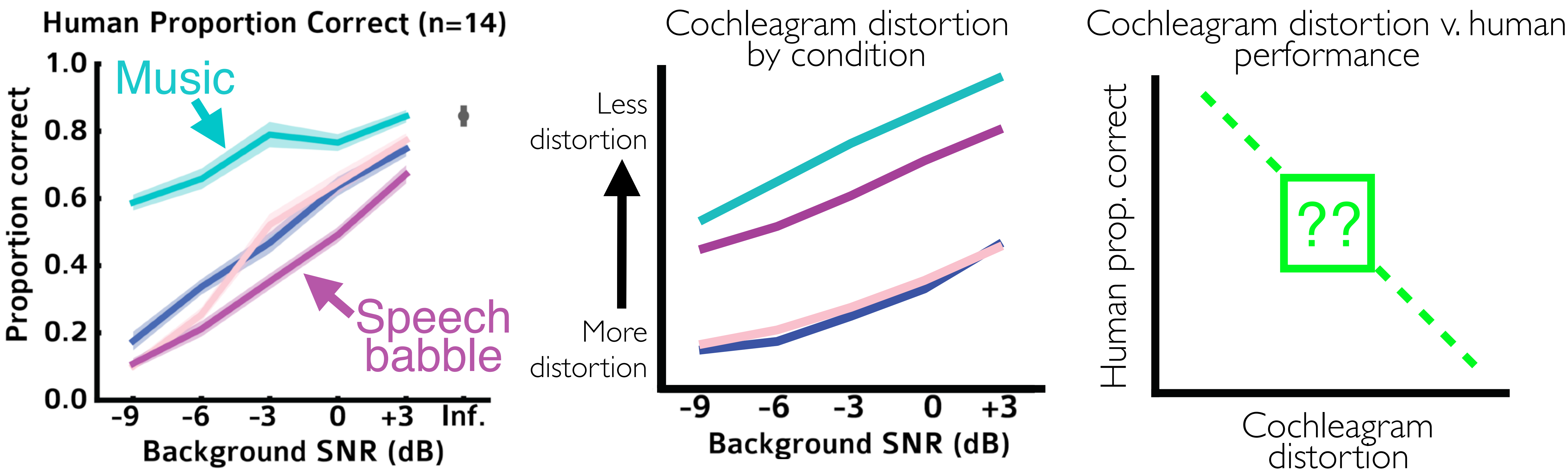
Distortion and pattern of performance.



LEGEND

Music Auditory scenes Speech babble Speech-shaped noise

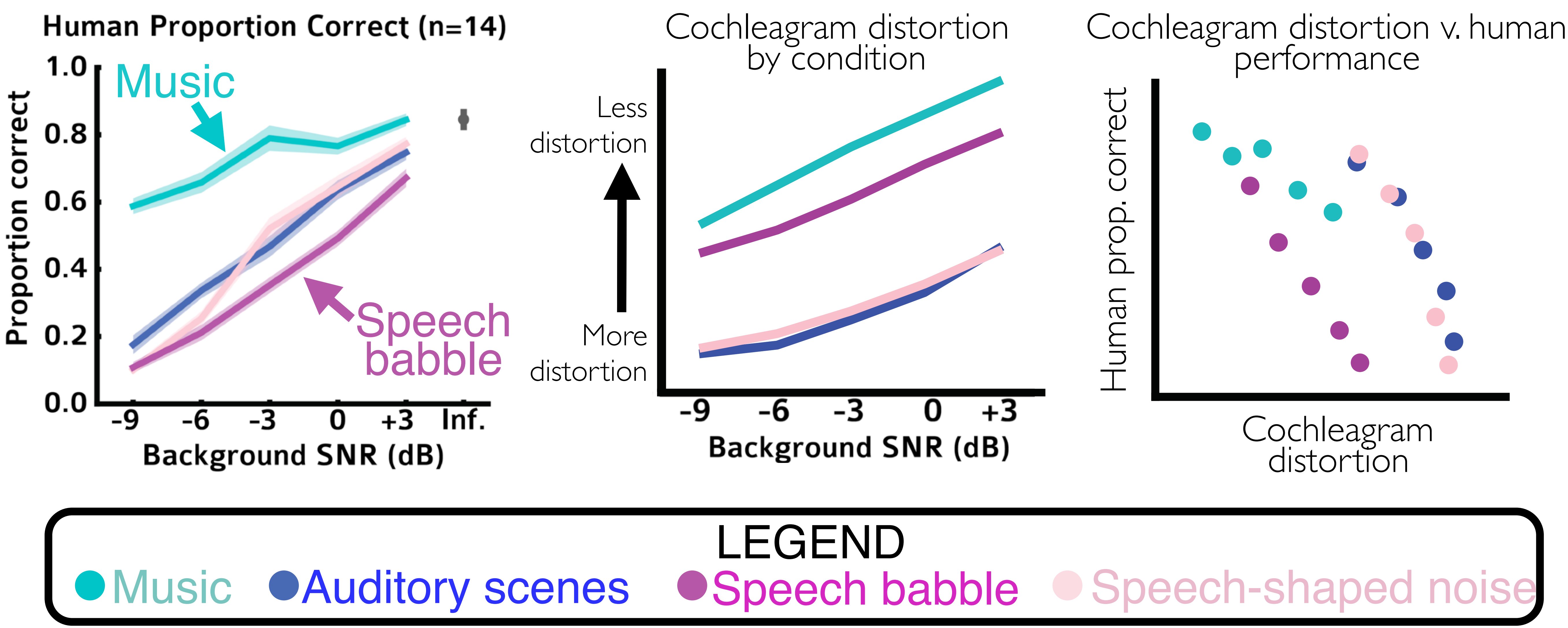
Distortion and pattern of performance.



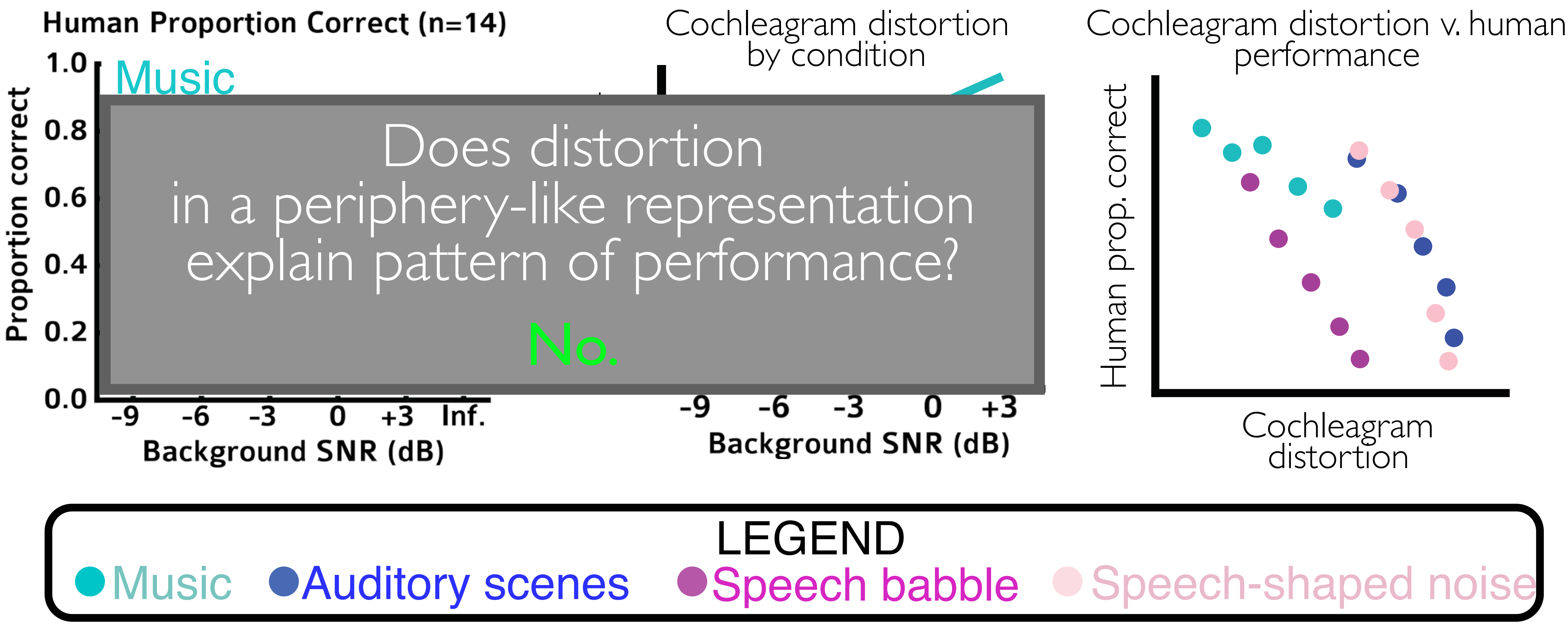
LEGEND

Music Auditory scenes Speech babble Speech-shaped noise

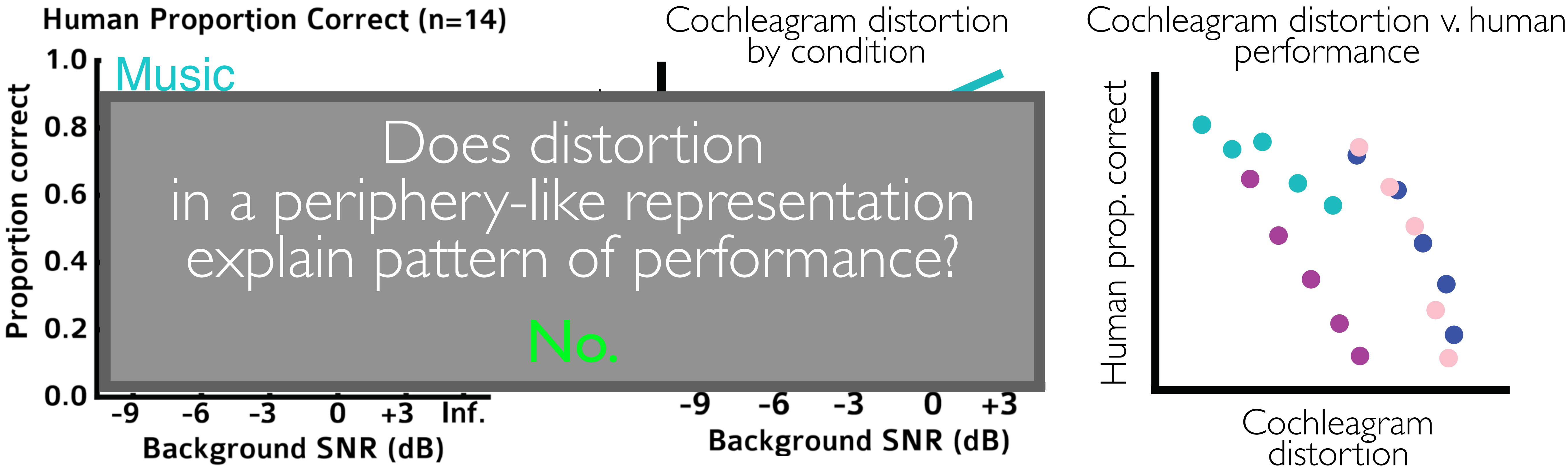
Distortion and pattern of performance.



Distortion and pattern of performance.



Distortion and pattern of performance.

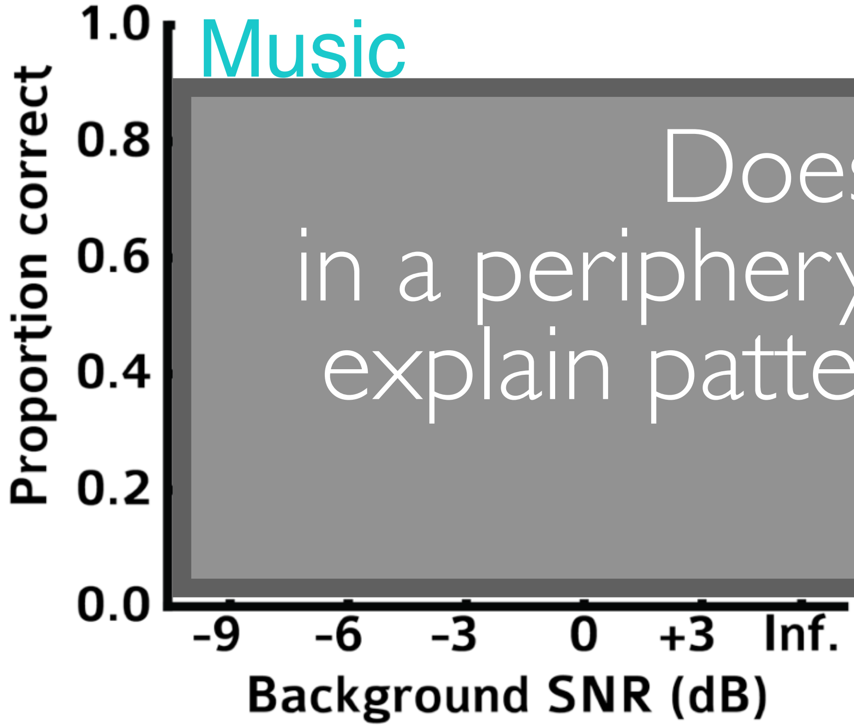


LEGEND

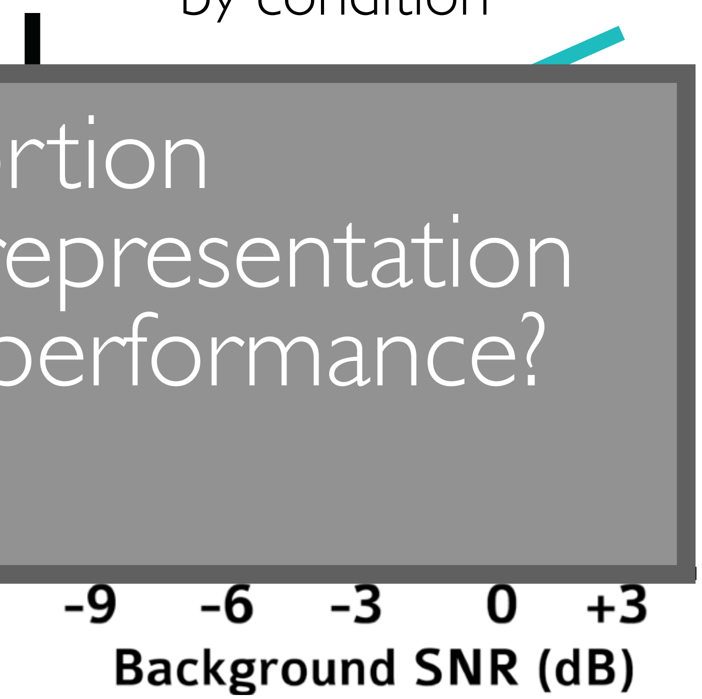
Music Auditory scenes Speech babble Speech-shaped noise

Does distortion of CNN representation explain pattern of performance?

Distortion and pattern of performance.



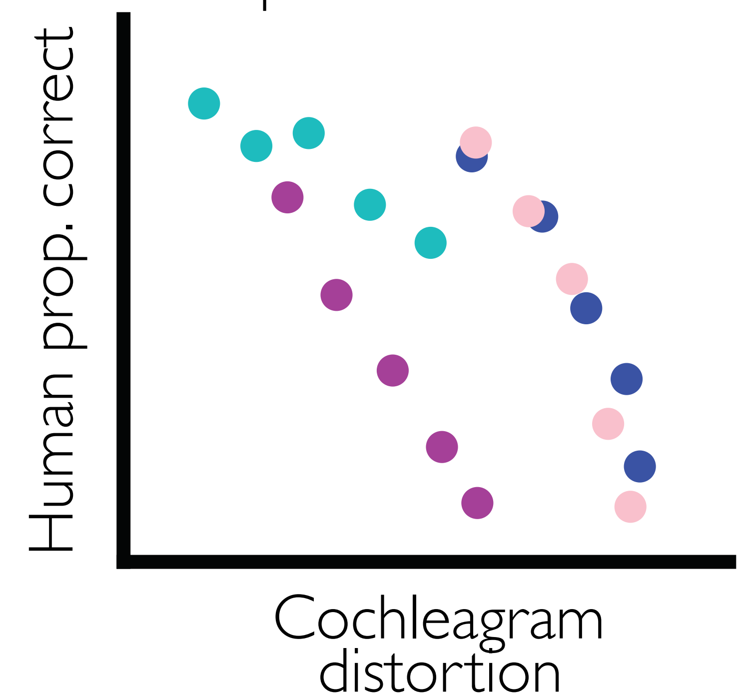
Cochleagram distortion by condition



Does distortion
in a periphery-like representation
explain pattern of performance?

No.

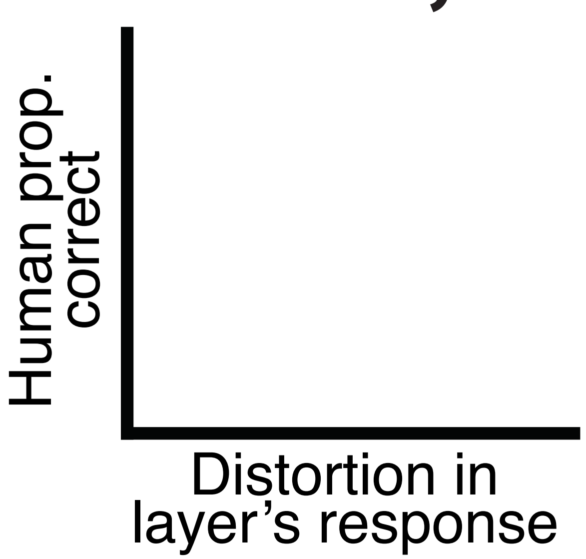
Cochleagram distortion v. human performance

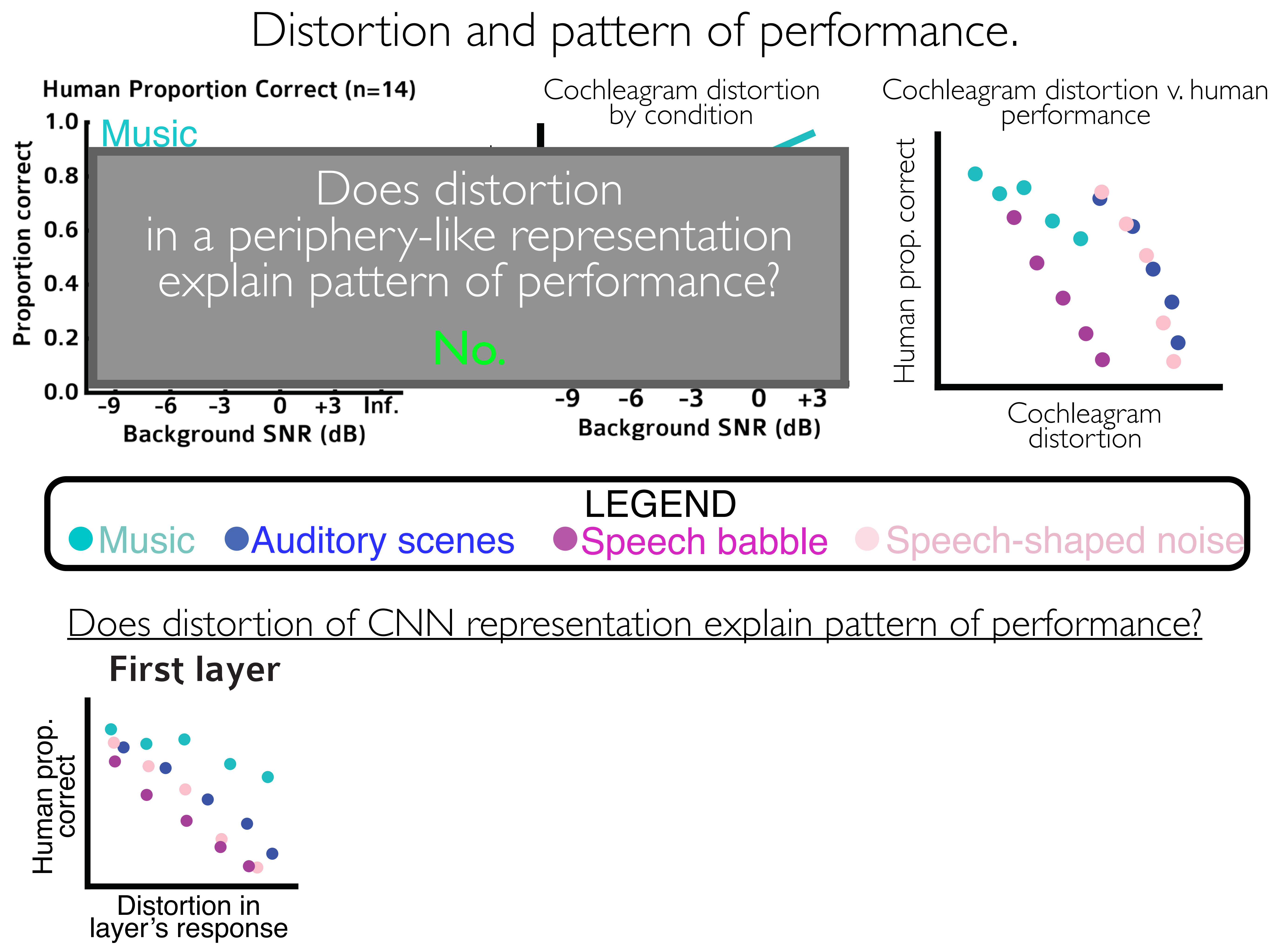


LEGEND
Music   Auditory scenes   Speech babble   Speech-shaped noise

Does distortion of CNN representation explain pattern of performance?

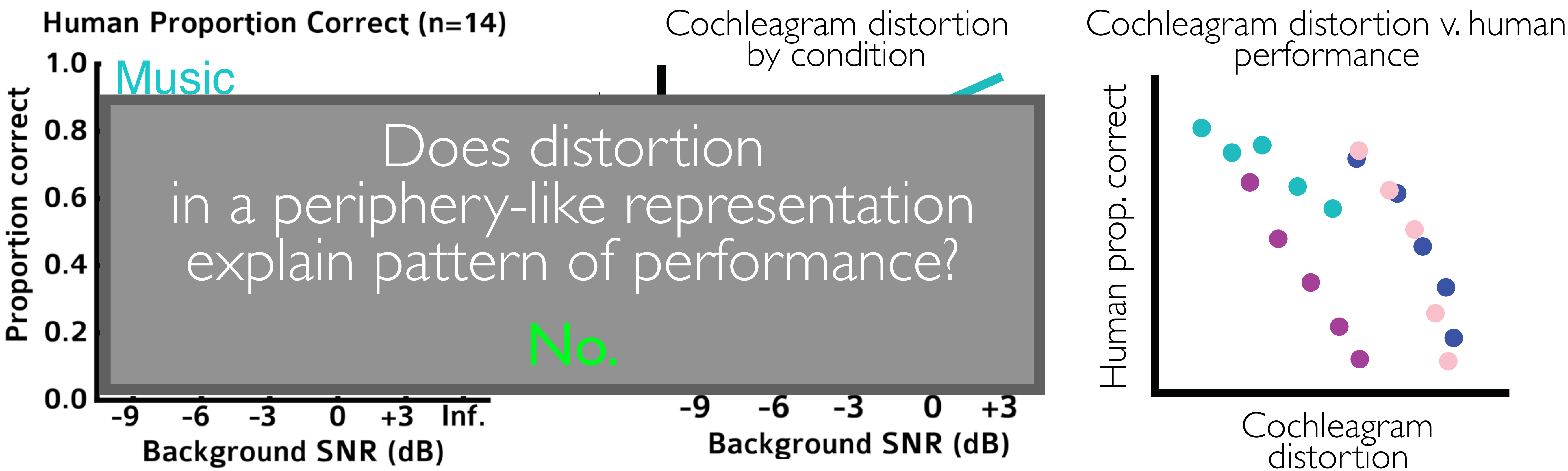
First layer





Distortion and pattern of performance.
Human Proportion Correct (n=14)
Music
Proportion correct
1.0
0.8
0.6
0.4
0.2
0.0
-9
-6
-3
0
+3
Inf.
Background SNR (dB)
Cochleagram distortion by condition
-9
-6
-3
0
+3
Background SNR (dB)
Cochleagram distortion v. human performance
Human prop. correct
Cochleagram distortion
Does distortion
in a periphery-like representation
explain pattern of performance?
No.
LEGEND
Music
Auditory scenes
Speech babble
Speech-shaped noise
Does distortion of CNN representation explain pattern of performance?
First layer
Human prop. correct
Distortion in layer's response

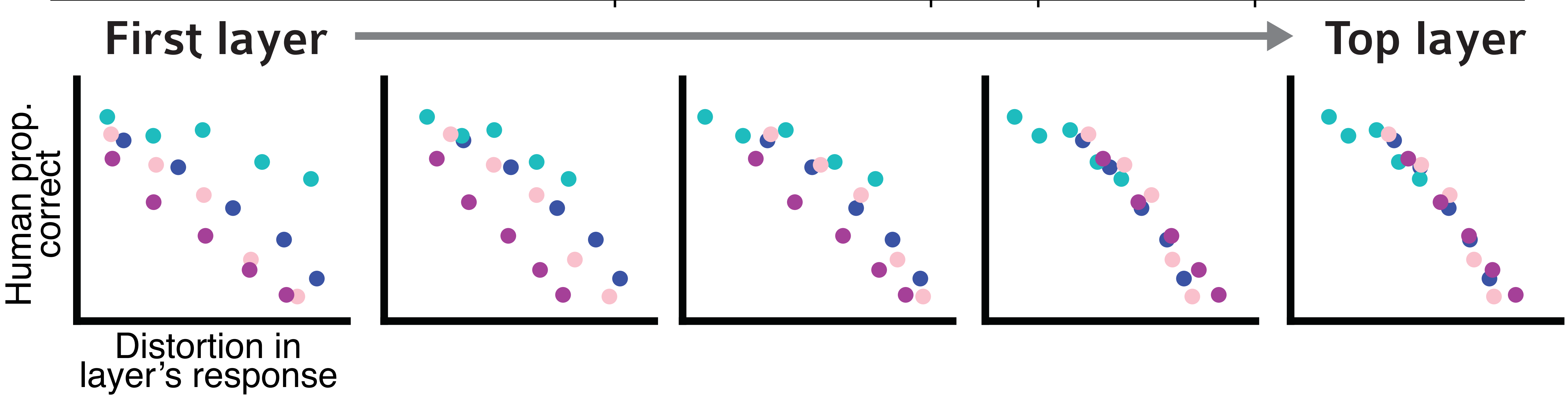
Distortion and pattern of performance.



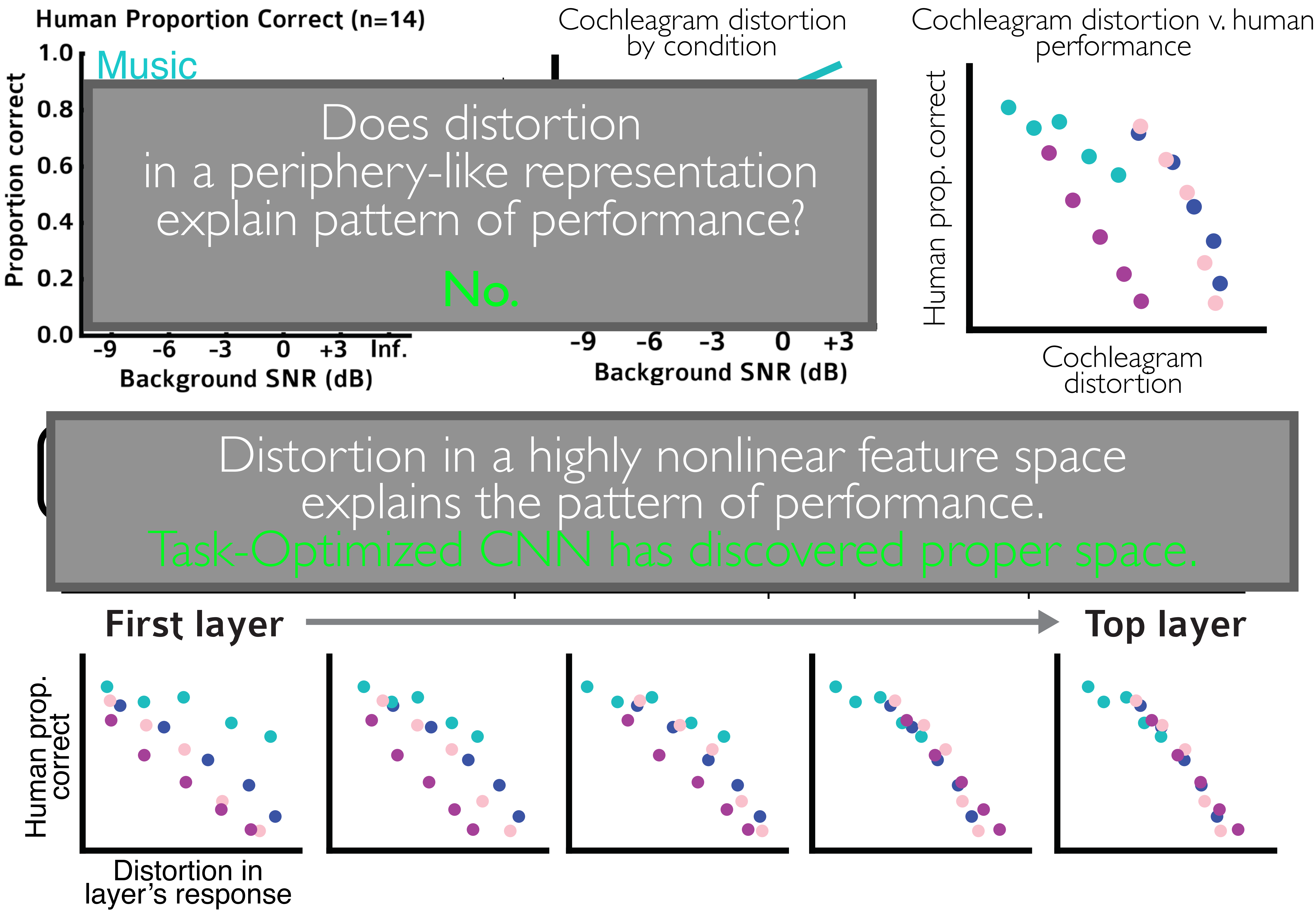
LEGEND

Music · Auditory scenes · Speech babble · Speech-shaped noise

Does distortion of CNN representation explain pattern of performance?



Distortion and pattern of performance.



Imaging Experiment

fMRI response data collected* on 165 commonly heard natural sound stimuli.

Man speaking
Flushing toilet
Pouring liquid
Tooth-brushing
Woman speaking
Car accelerating
Biting and chewing
Laughing
Typing
Car engine starting
Running water
Breathing
Keys jangling
Dishes clanking
Ringtone
Microwave
Dog barking

Road traffic
Zipper
Cellphone vibrating
Water dripping
Scratching
Car windows
Telephone ringing
Chopping food
Telephone dialing
Girl speaking
Car horn
Writing
Computer startup sound
Background speech
Songbird
Pouring water
Pop song
Water boiling

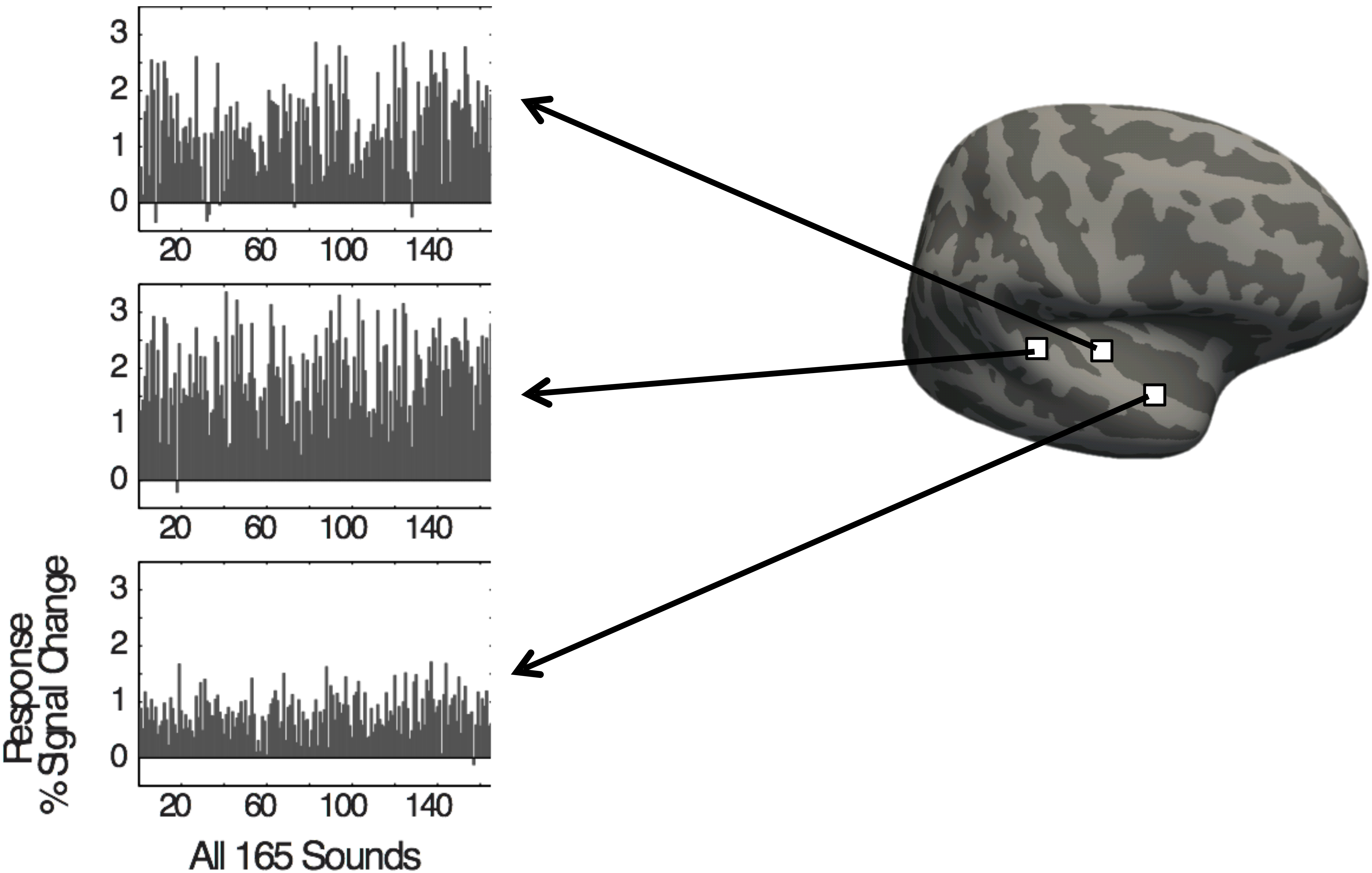
Guitar
Coughing
Crumpling paper
Siren
Splashing water
Computer speech
Alarm clock
Walking with heels
Vacuum
Wind
Boy speaking
Chair rolling
Rock song
Door knocking

•
•
•

*Sam Norman-Haignere, Nancy Kanwisher, and Josh McDermott

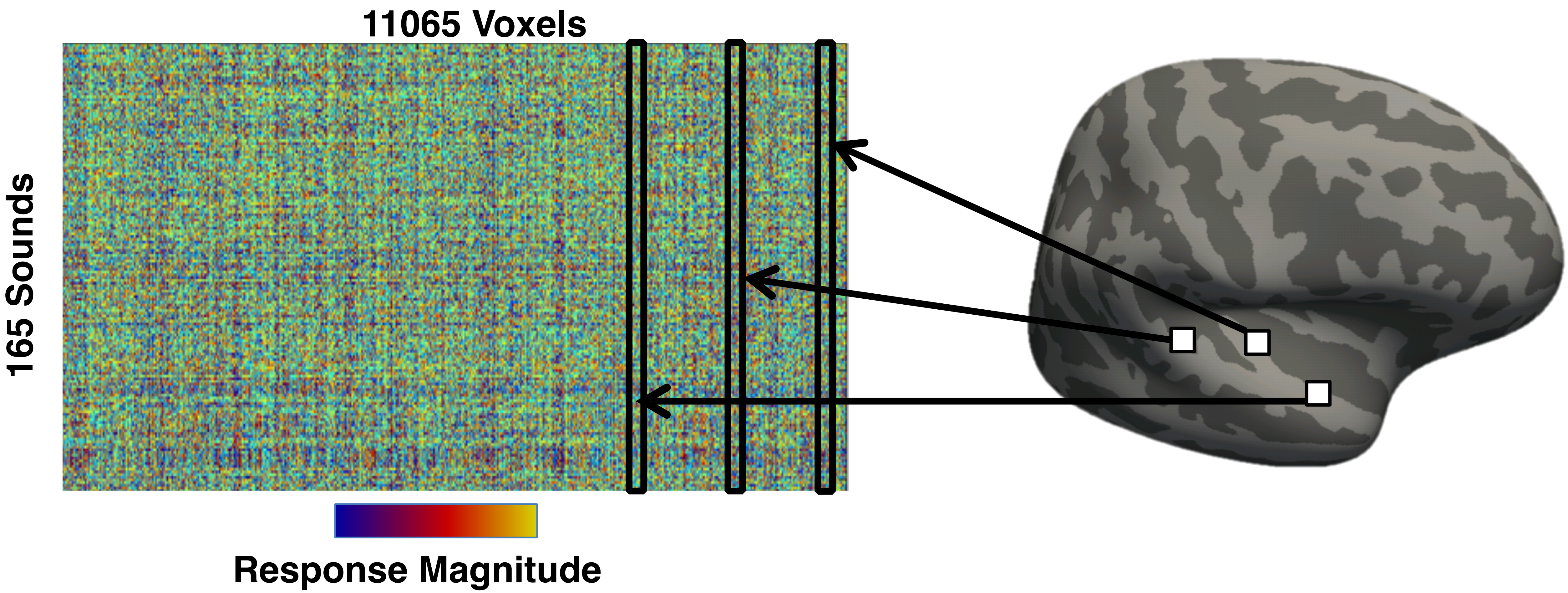
Imaging Experiment

For each voxel, measured average response to each sound:



Imaging Experiment

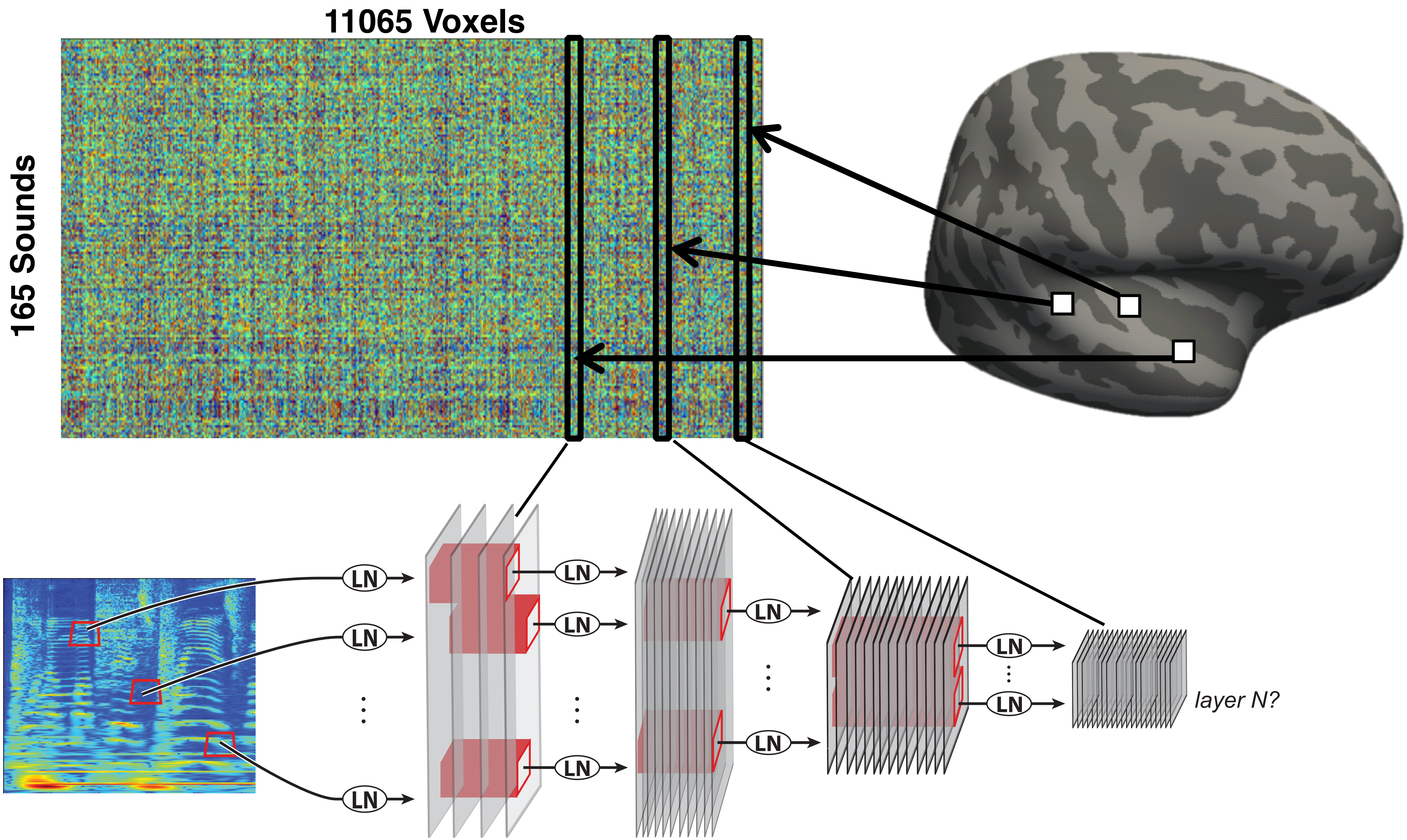
For each voxel, measured average response to each sound:



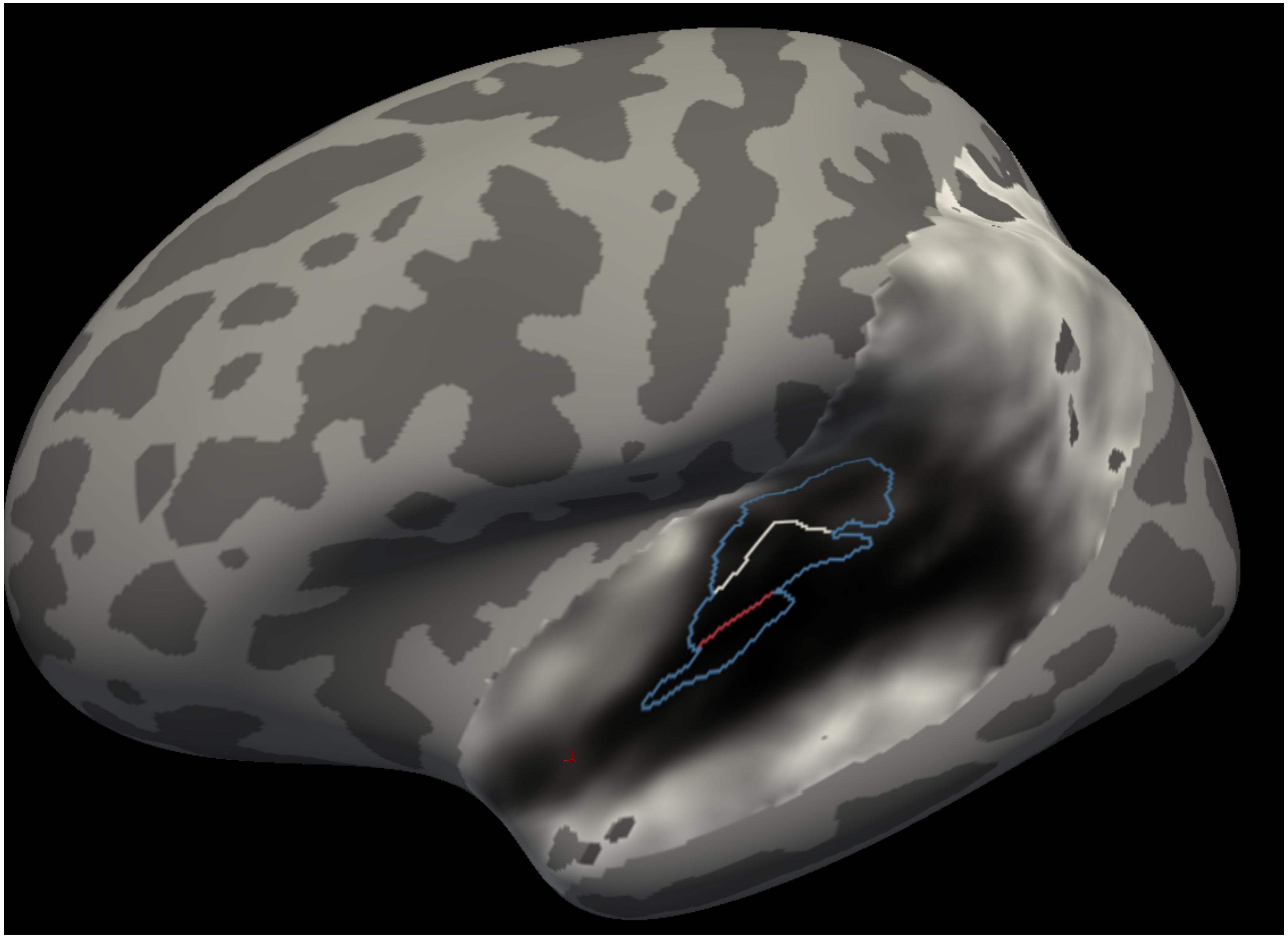
Data matrix: voxels X sounds.

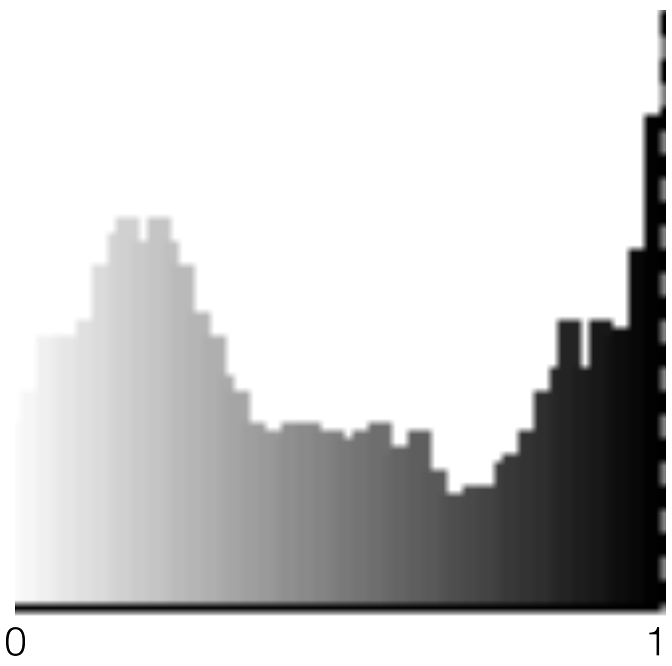
Imaging Experiment

Neural predictivity: the ability of model to predict each individual voxel's activity using linear regression.

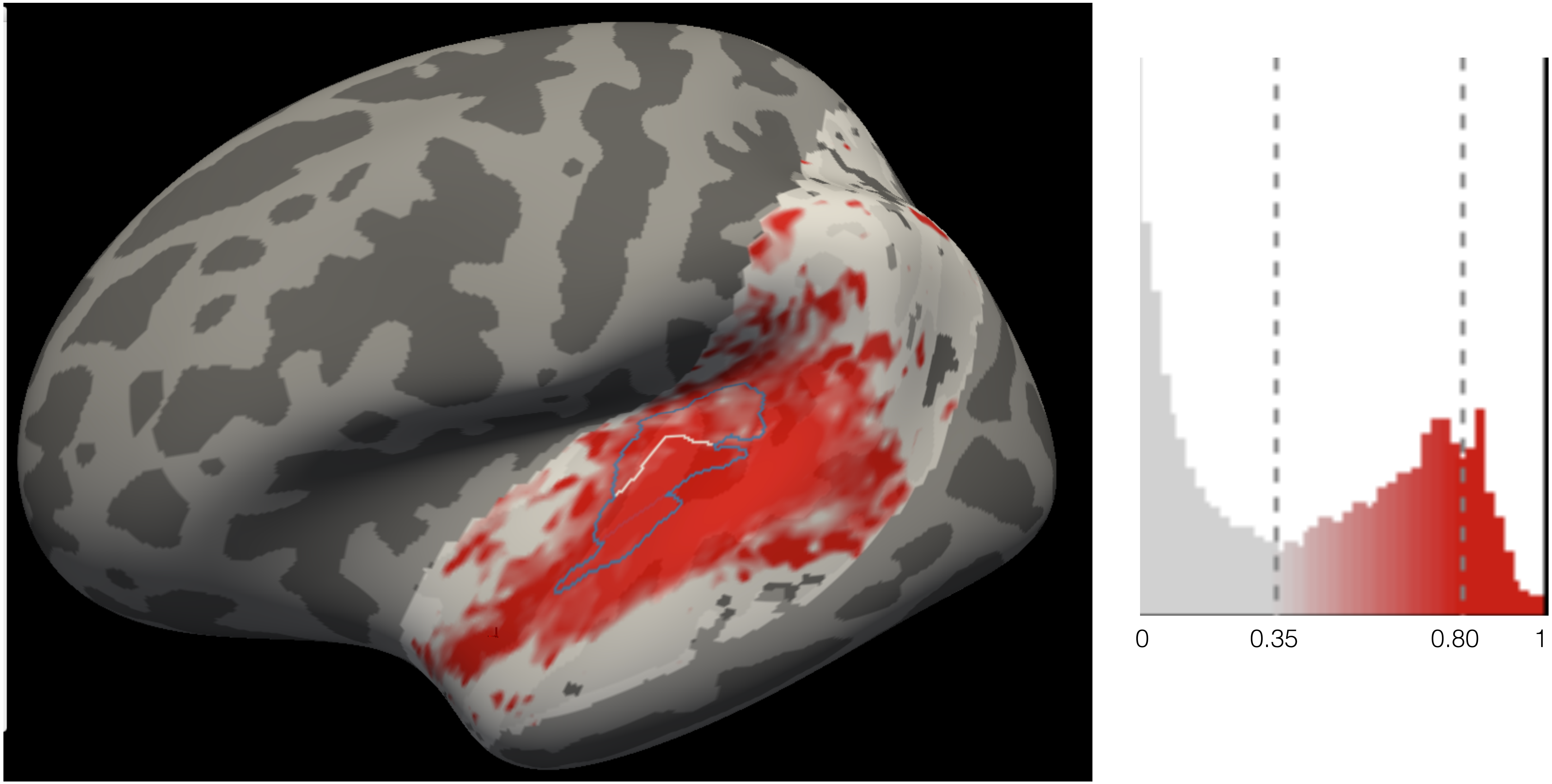


Response Reliability at Voxel Level

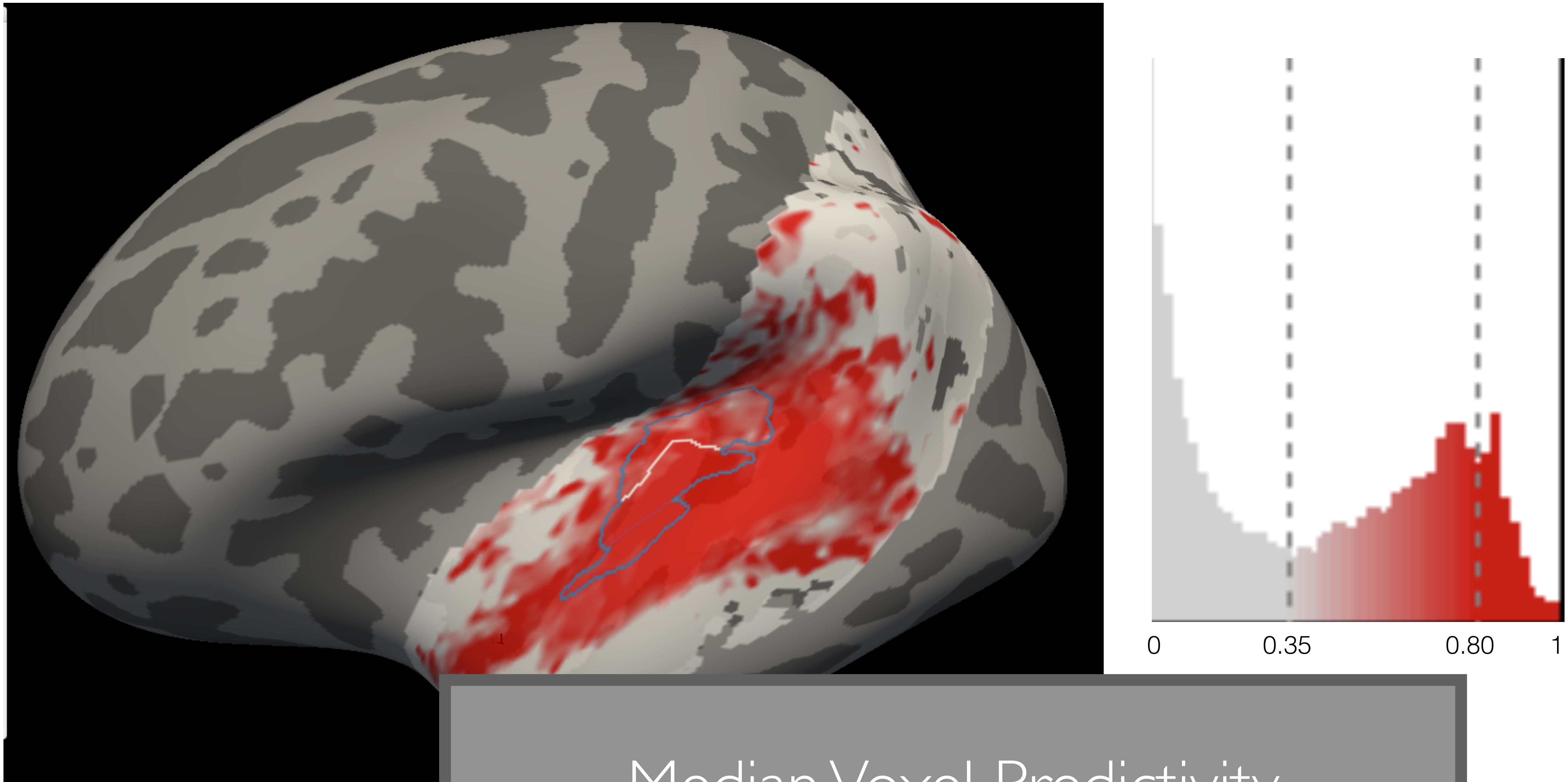




Model Productivity at Best Layer



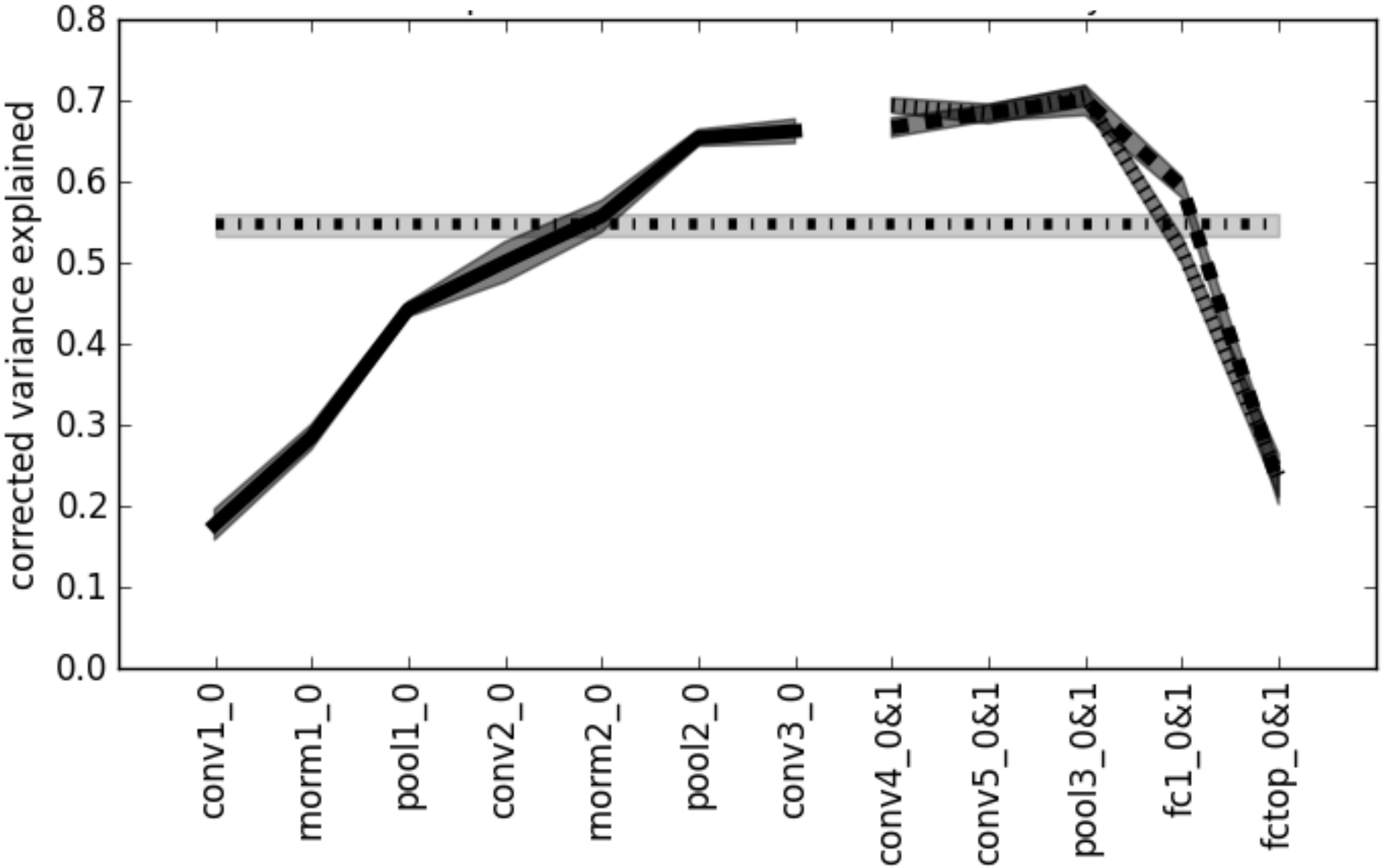
Model Productivity at Best Layer



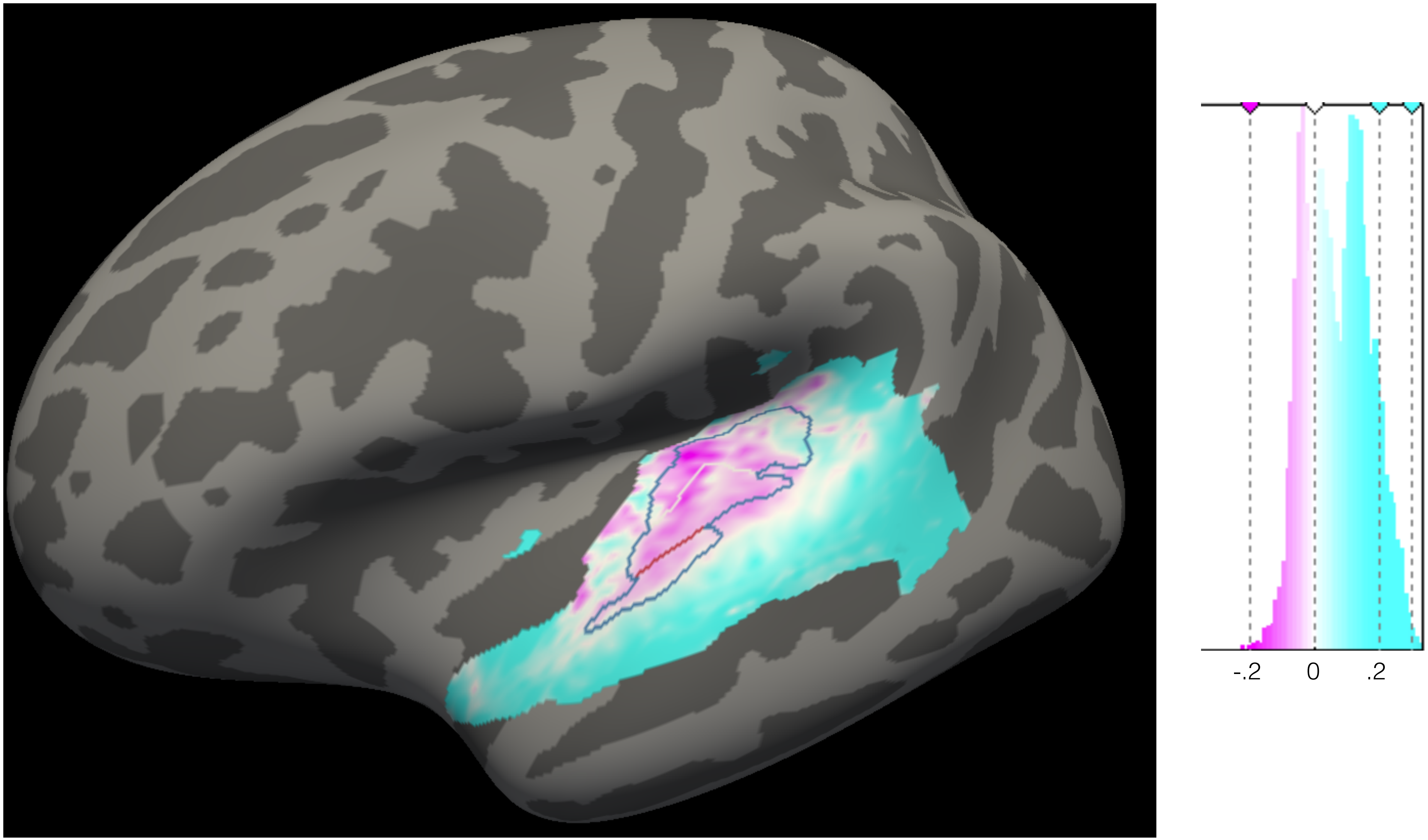
Median Voxel Predictivity

~80%.

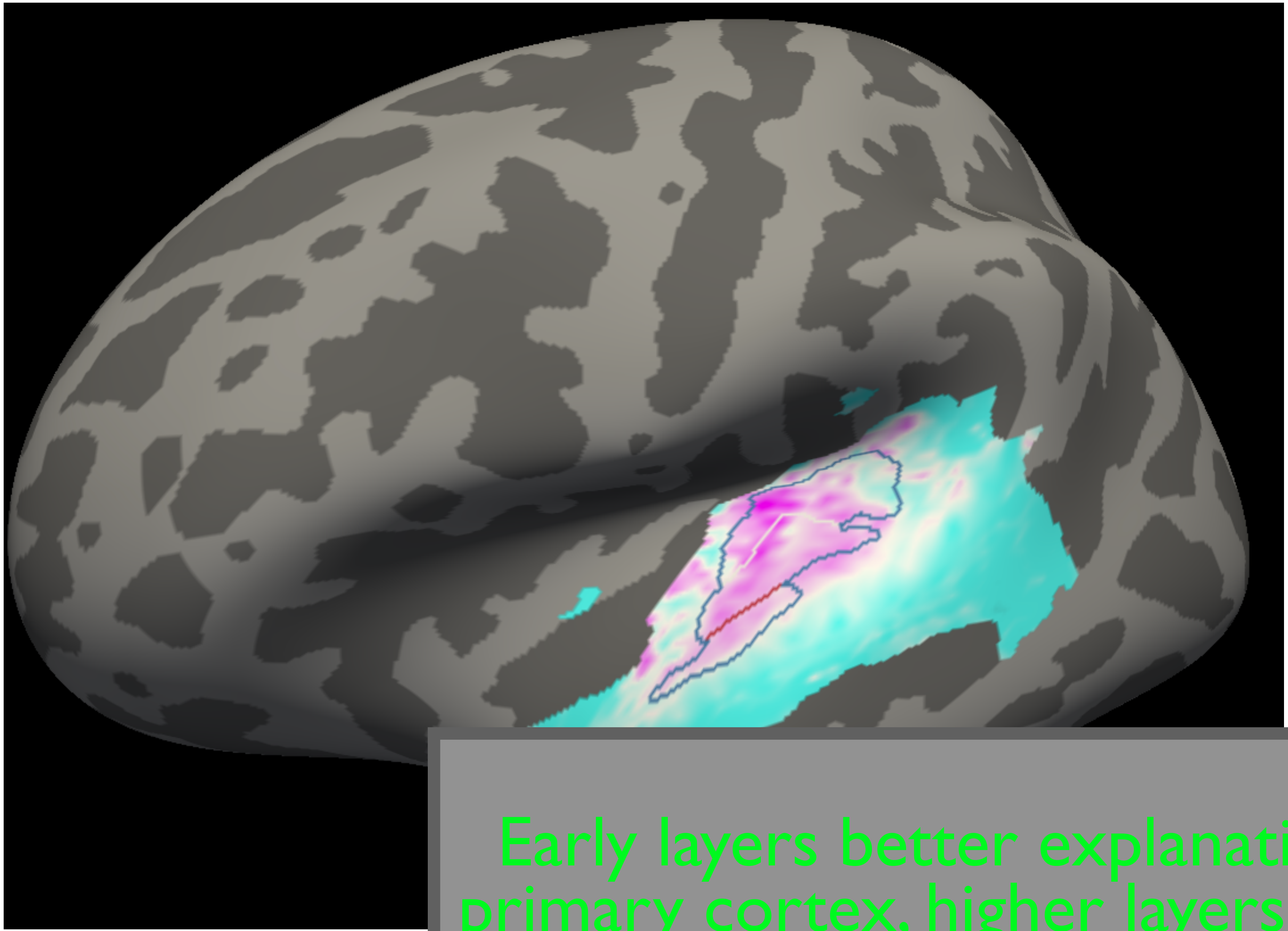
Median Predictivity as a Function of Model Layer

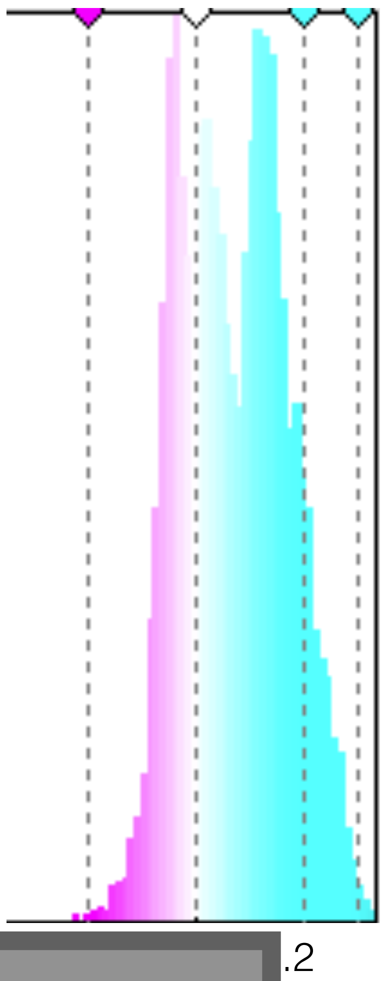


Predictivity Difference Between High and Low Model Layers



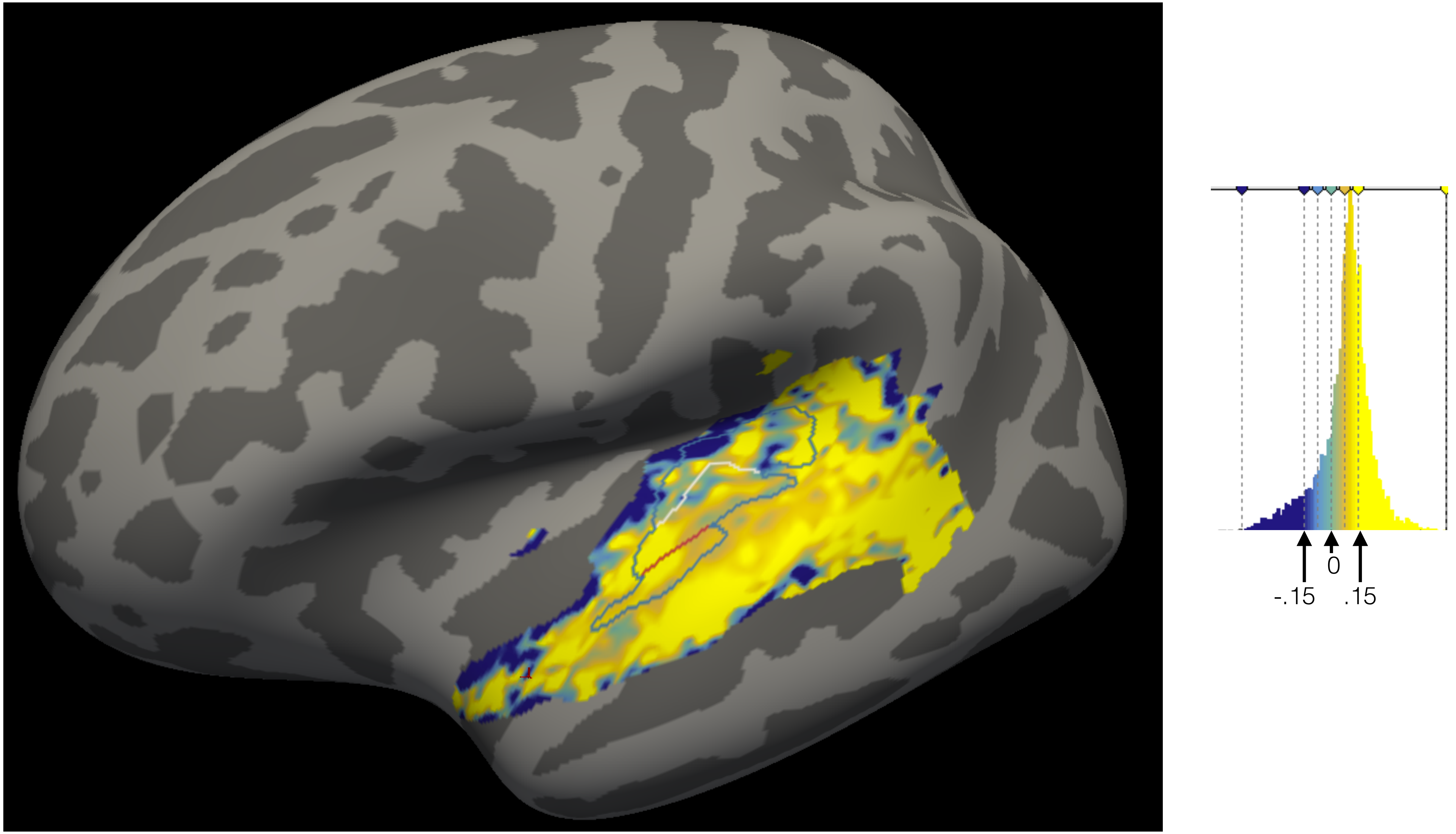
Predictivity Difference Between High and Low Model Layers



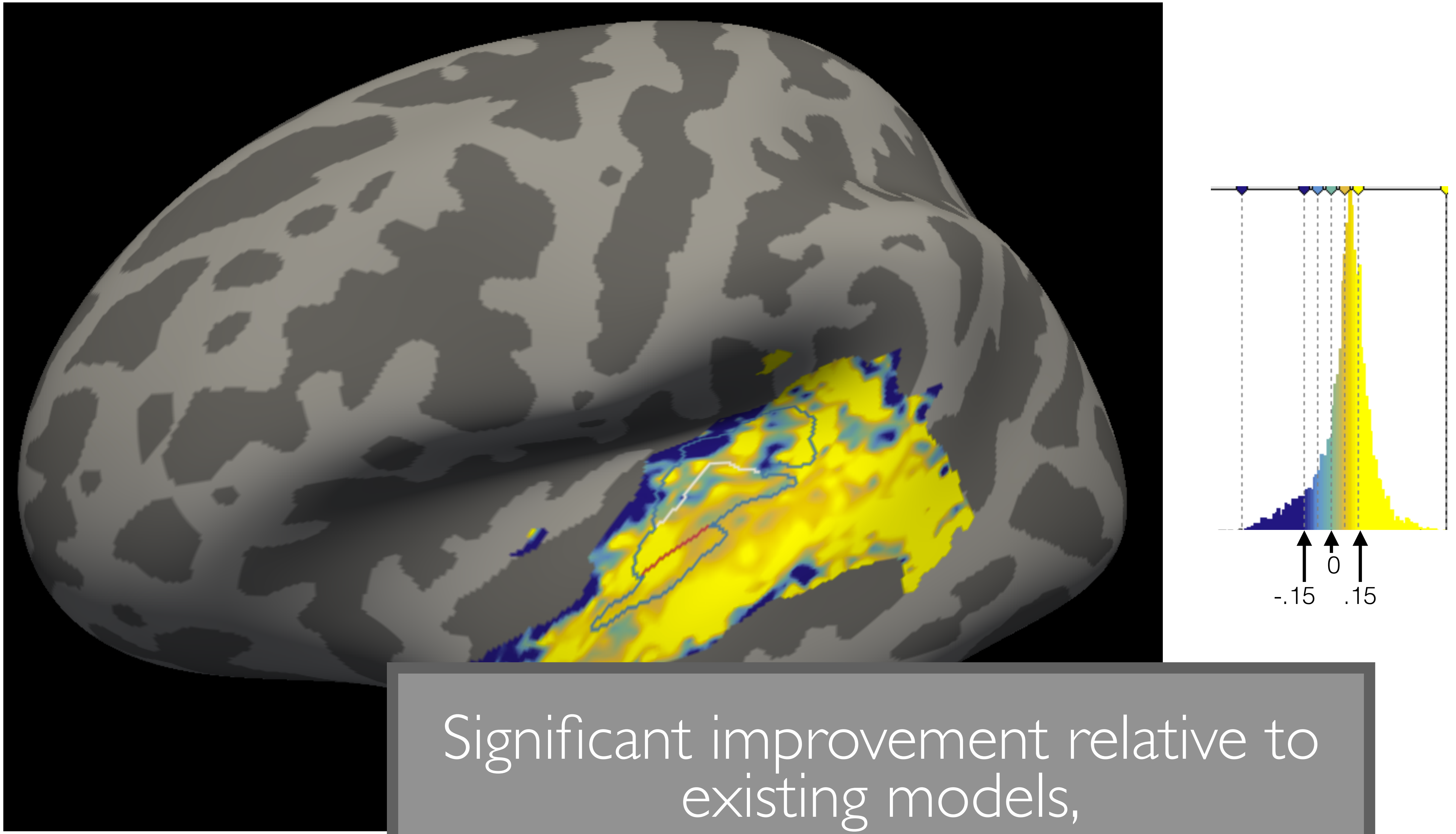


Early layers better explanation of primary cortex, higher layers better explanation of non-primary cortex.

Comparison to Spectrotemporal Filtering Model



Comparison to Spectrotemporal Filtering Model

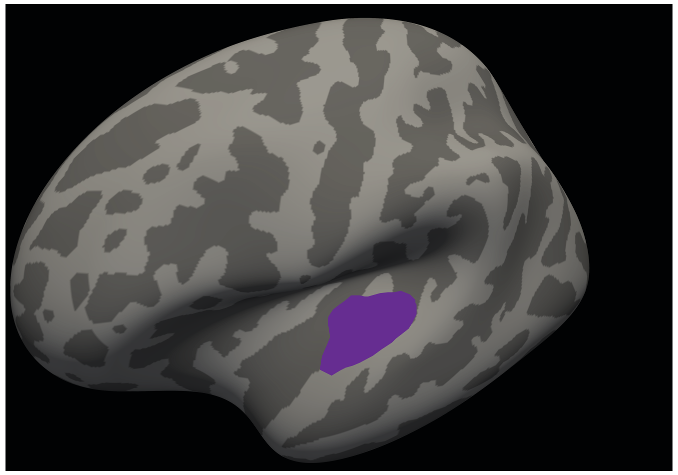


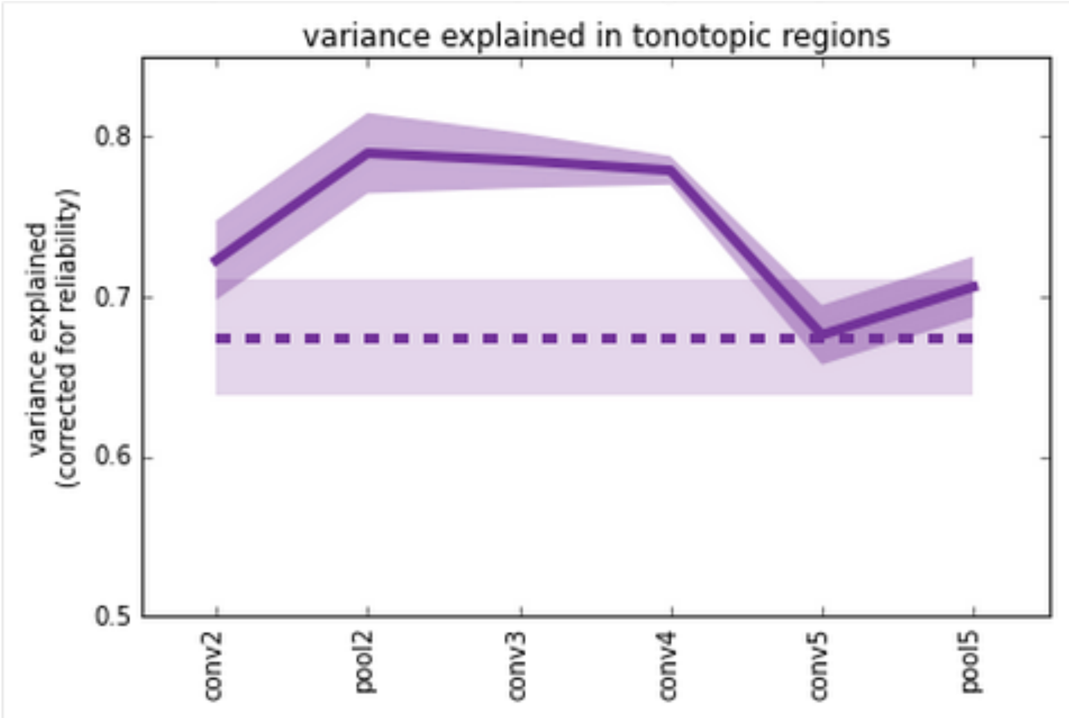
Significant improvement relative to existing models,

but especially in non-primary areas.

Differentiation by Region of Interest

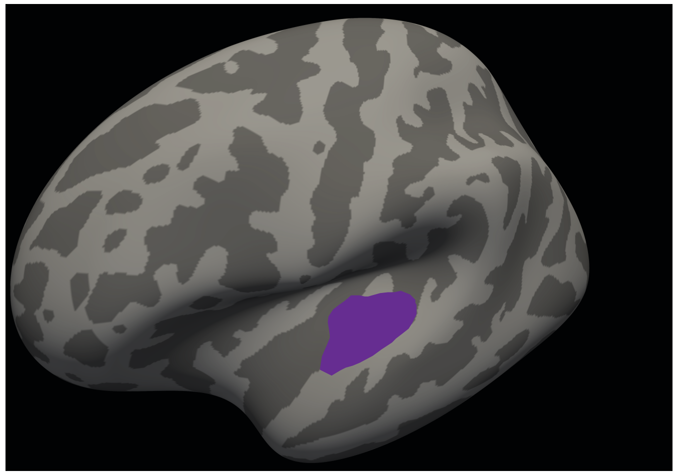
Tonotopic
(⊂ Primary)

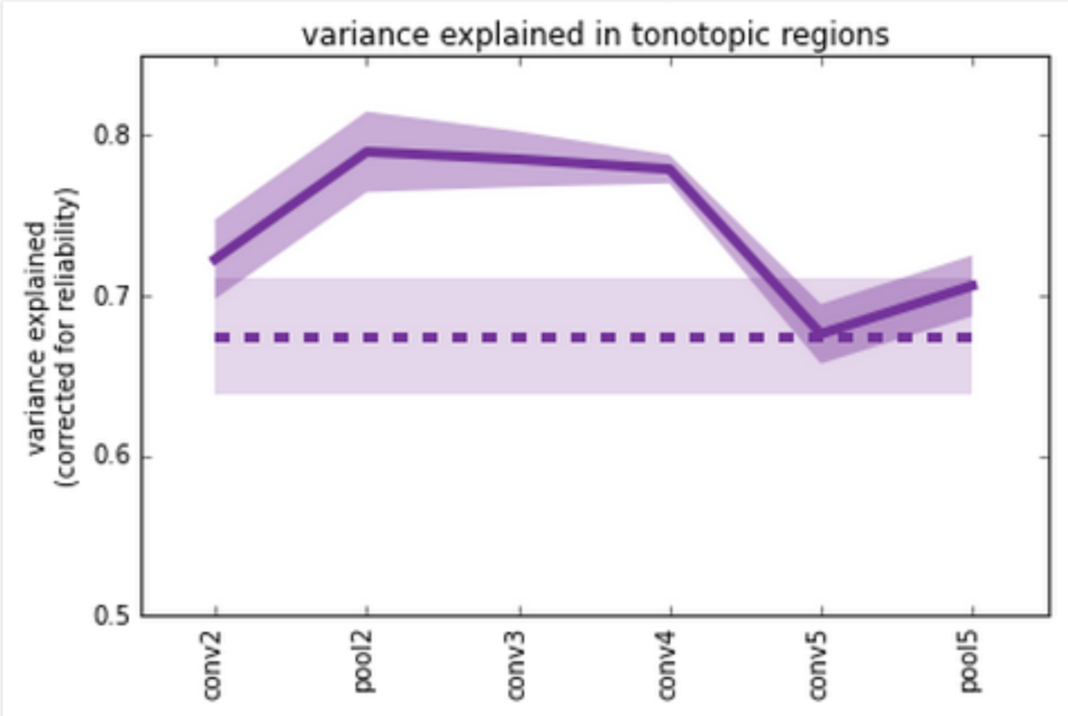




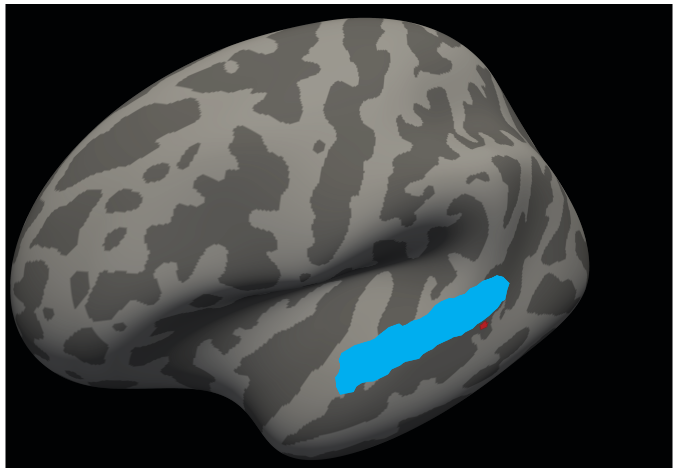
Differentiation by Region of Interest

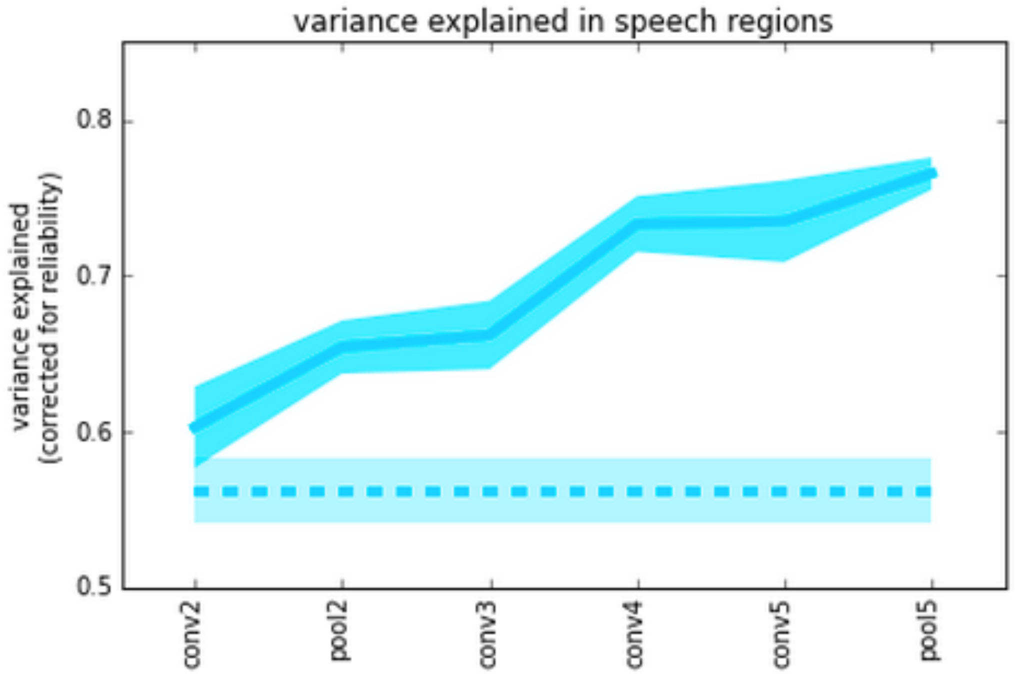
Tonotopic
(⊂ Primary)



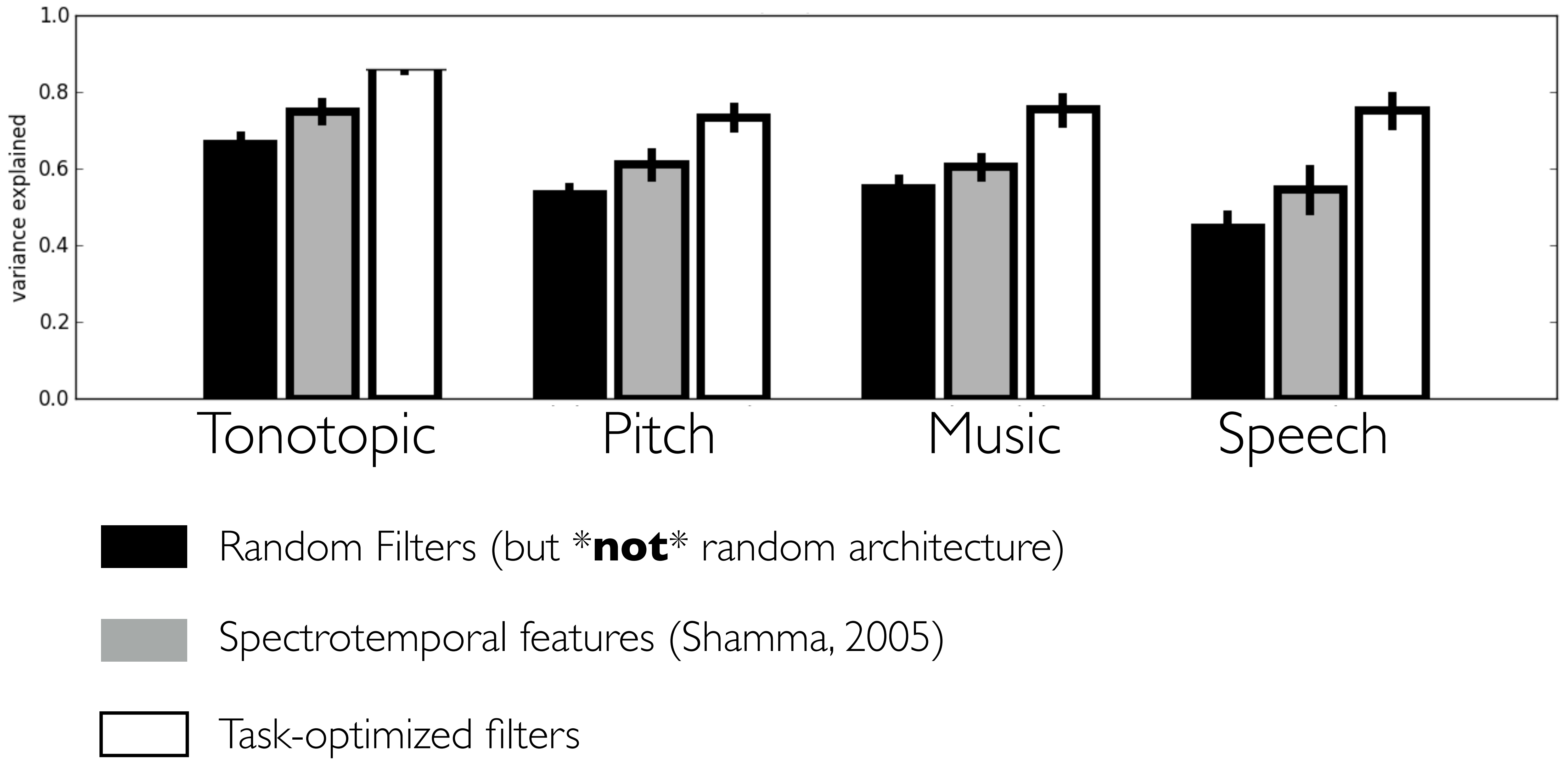


Speech-selective
(⊂ Non-primary)

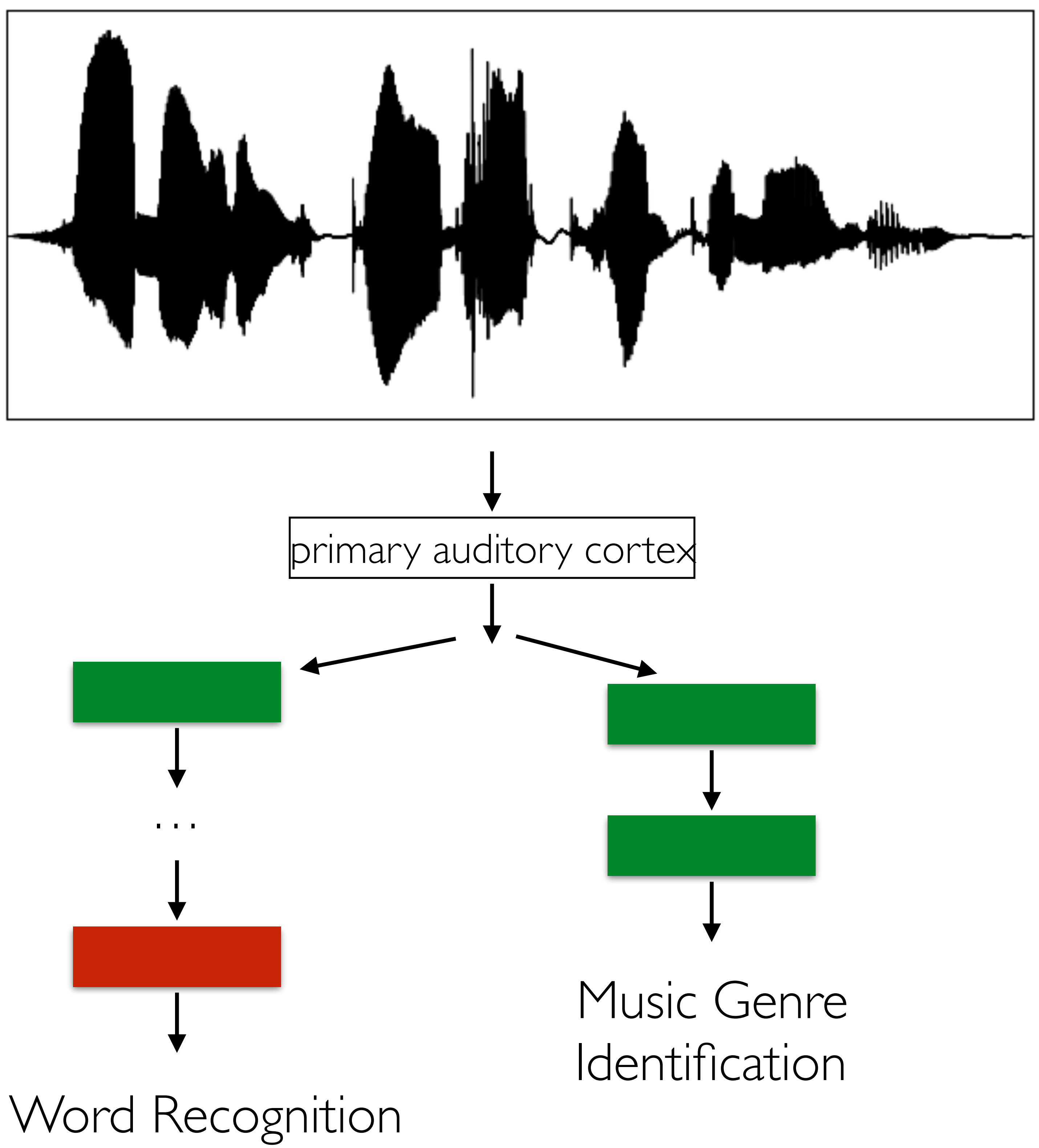




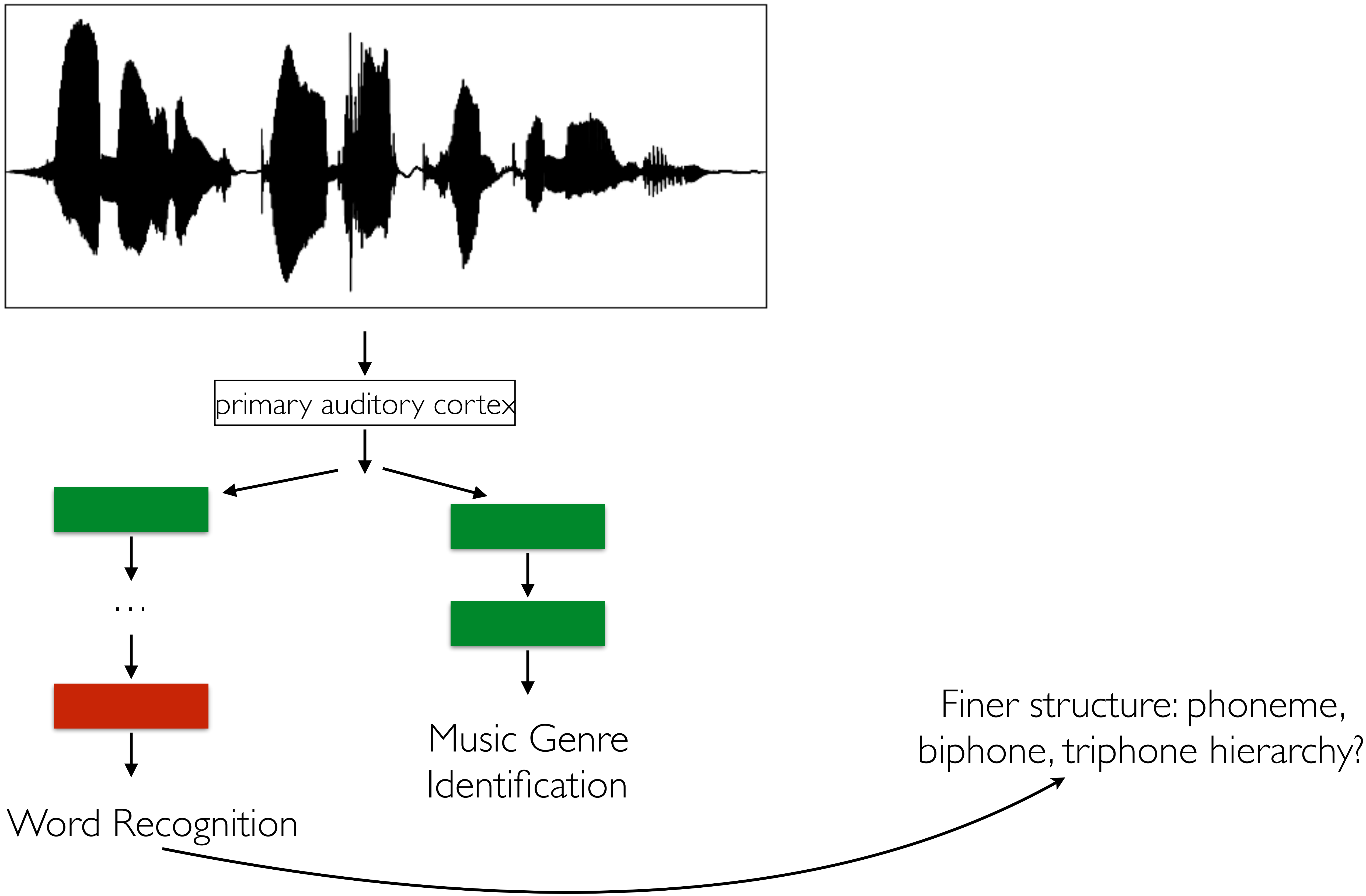
Comparison of Predictivity by RoI



Ongoing: Functionality Organization by Task

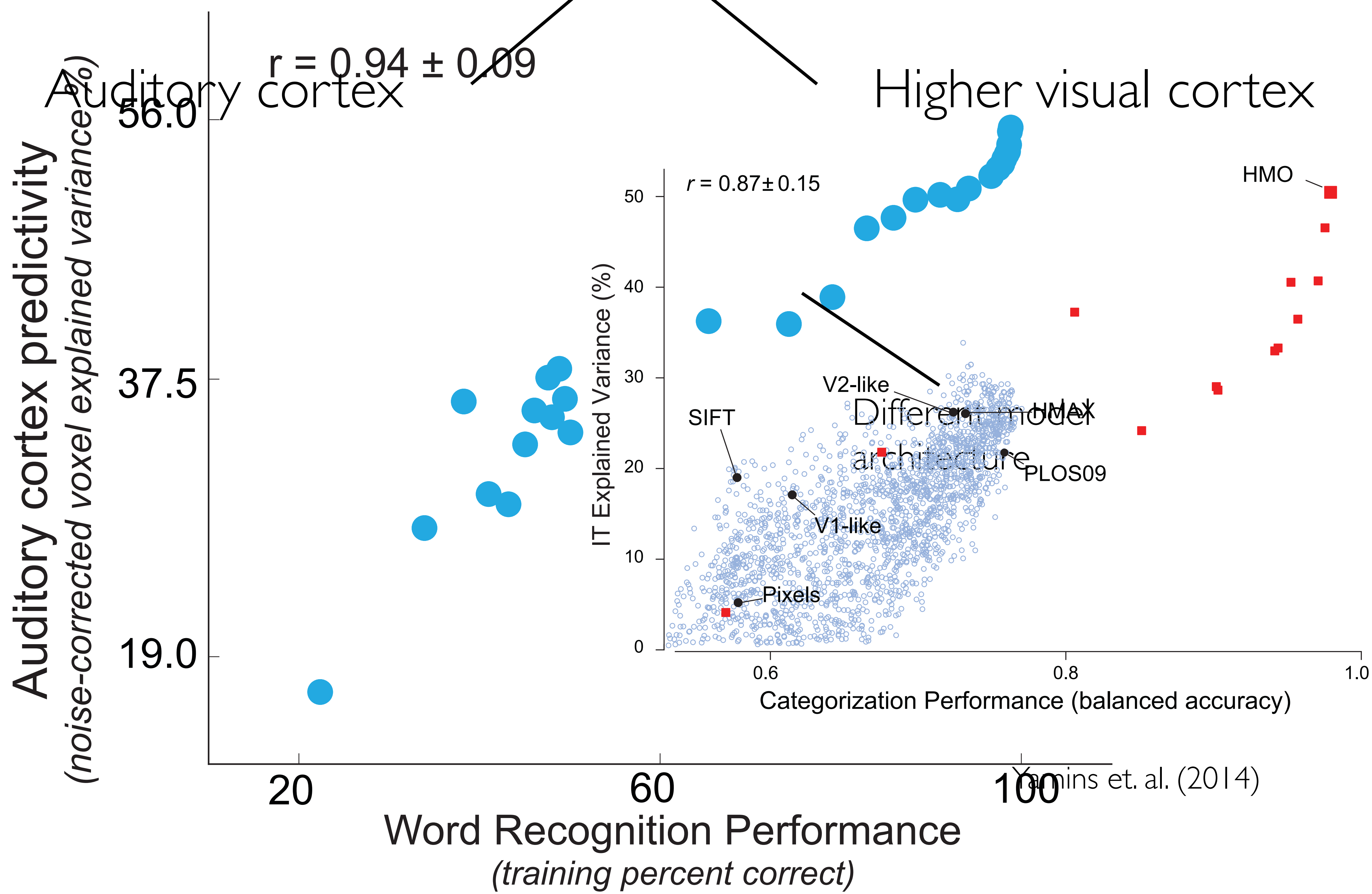


Ongoing: Functionality Organization by Task



Analysis of Model Architectures

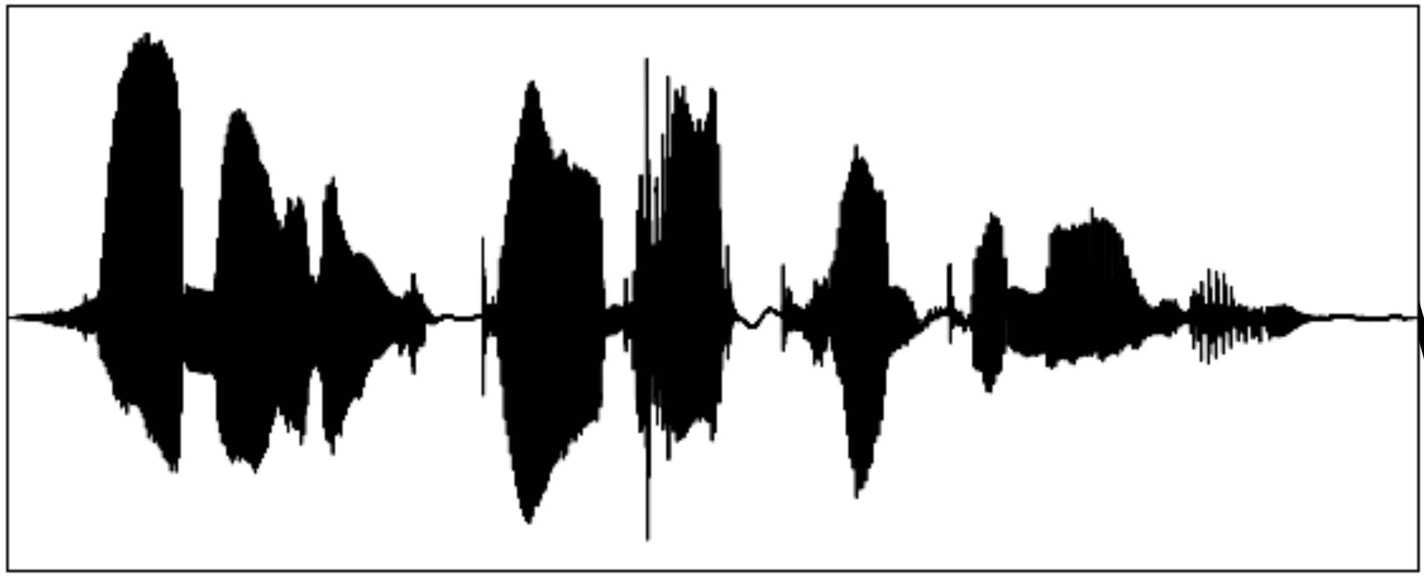
High-variation task performance vs:



Yamins et. al. (2014)

Principle of "Goal-Driven Modeling"

Heuristic of "Goal-Driven Modeling"



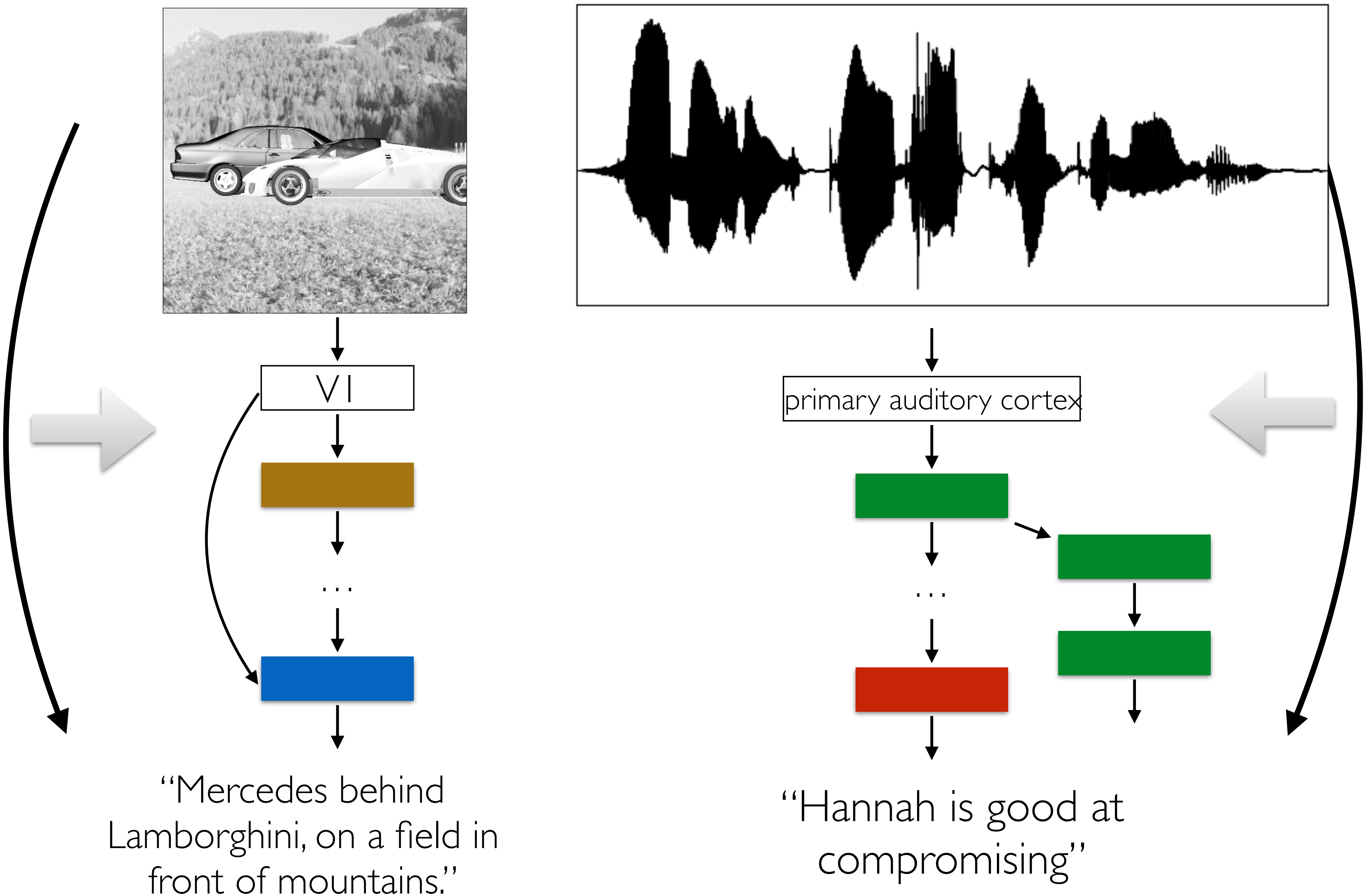
visual cortex

auditory cortex

"Mercedes behind Lamborghini, on a field in front of mountains."

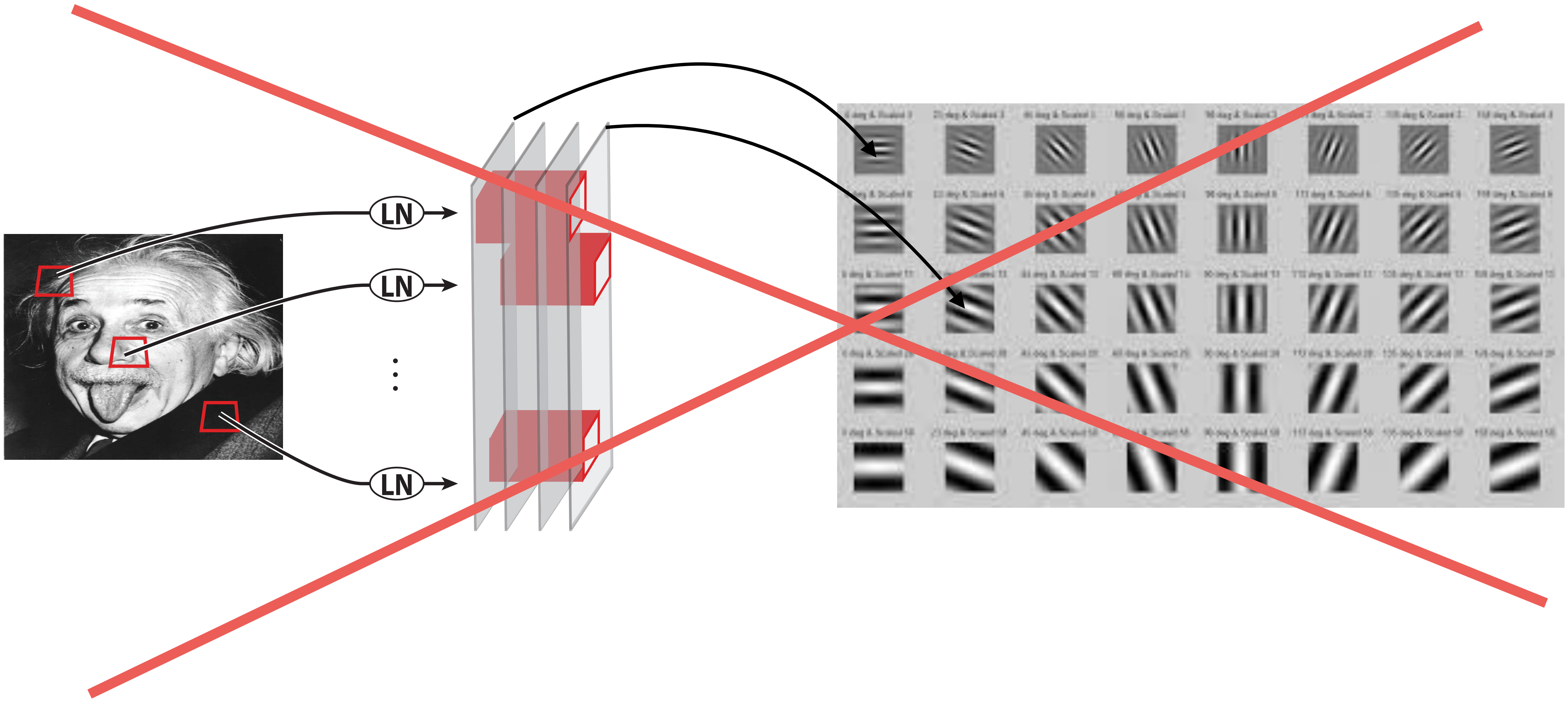
"Hannah is good at compromising"

Heuristic of "Goal-Driven Modeling"



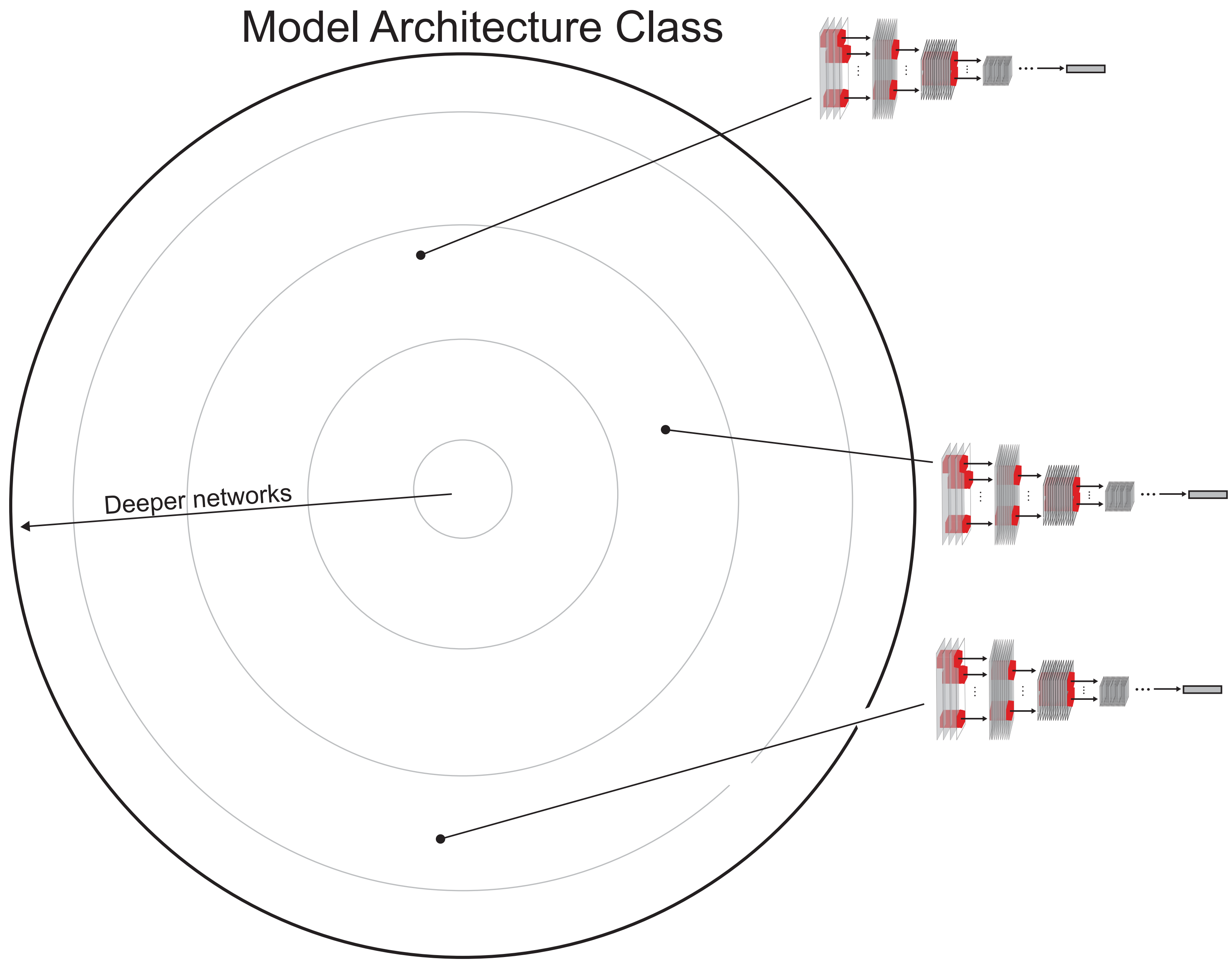
But what type of understanding is this?

But what type of understanding is this?



not saying this type of understanding is impossible …

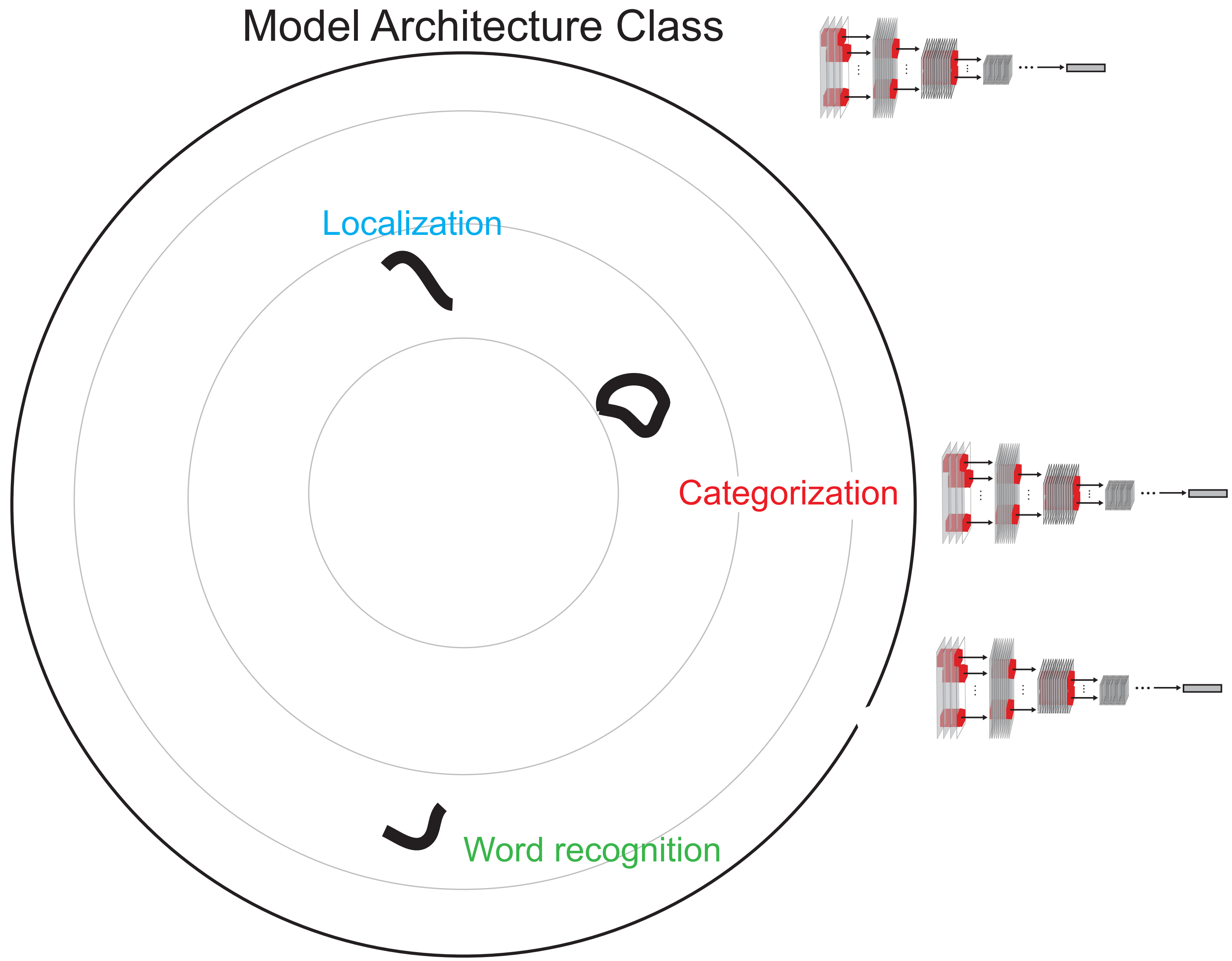
1. Formulate comprehensive model class (**CNNs**)



Yamins & DiCarlo.
***Nat. Neuro.** (2016)*

1. Formulate comprehensive model class (**CNNs**)

2. Choose challenging, ethologically-valid tasks (**categorization**)

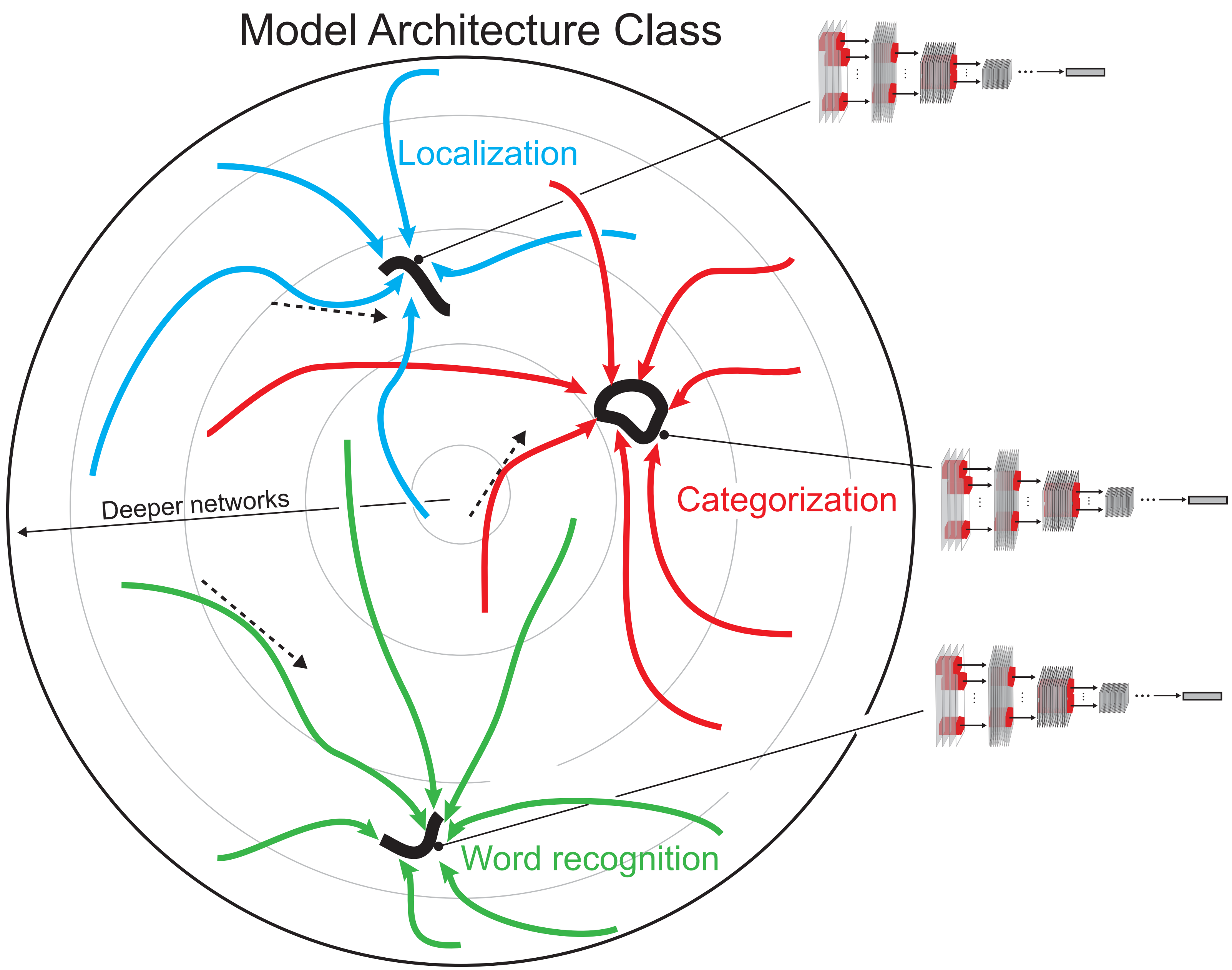


Yamins & DiCarlo.
Nat. Neuro. *(2016)*

1. Formulate comprehensive model class (**CNNs**)

2. Choose challenging, ethologically-valid tasks (**categorization**)

3. Implement generic learning rules (**gradient descent**)



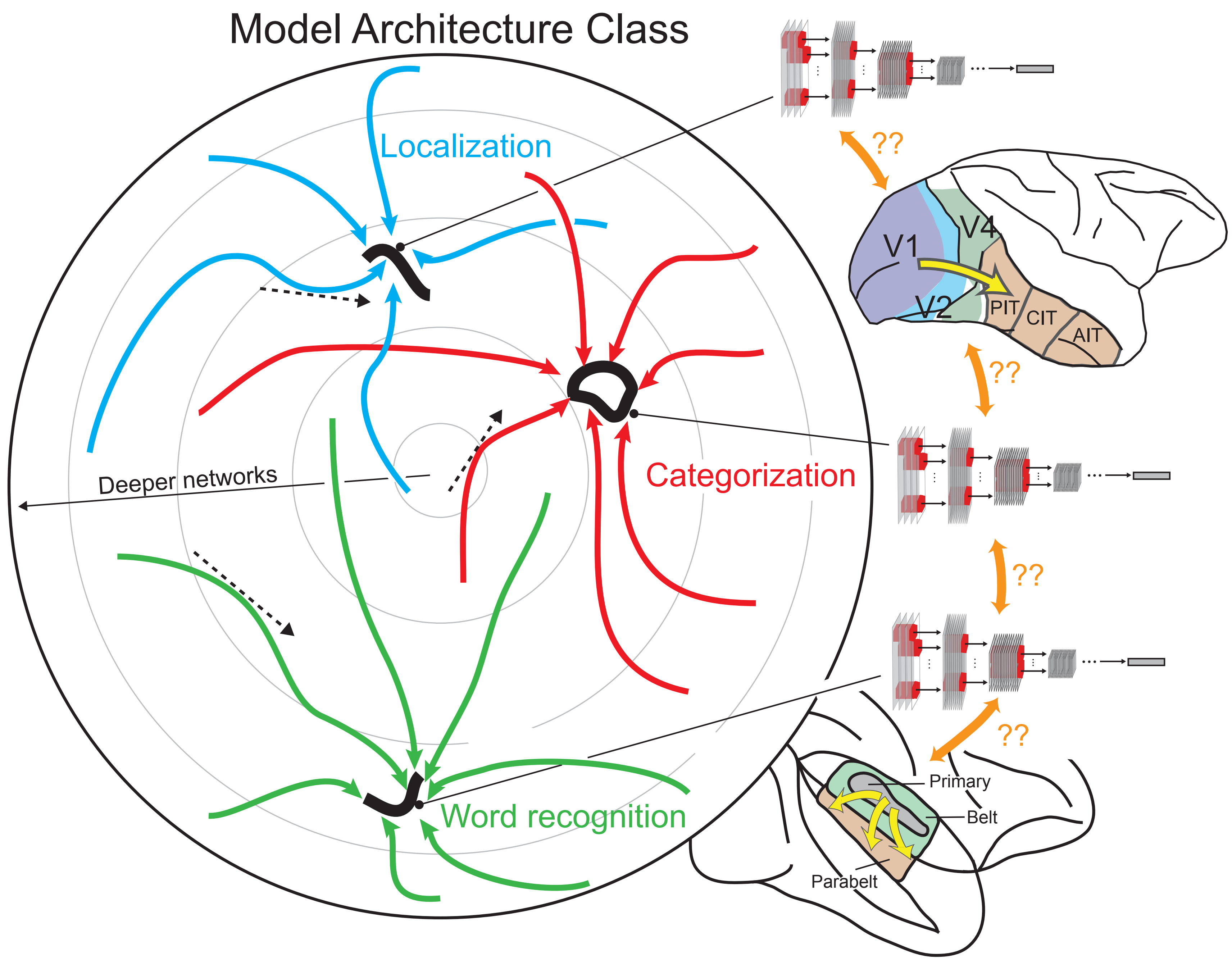
Yamins & DiCarlo.
Nat. Neuro. *(2016)*

1. Formulate comprehensive model class (**CNNs**)

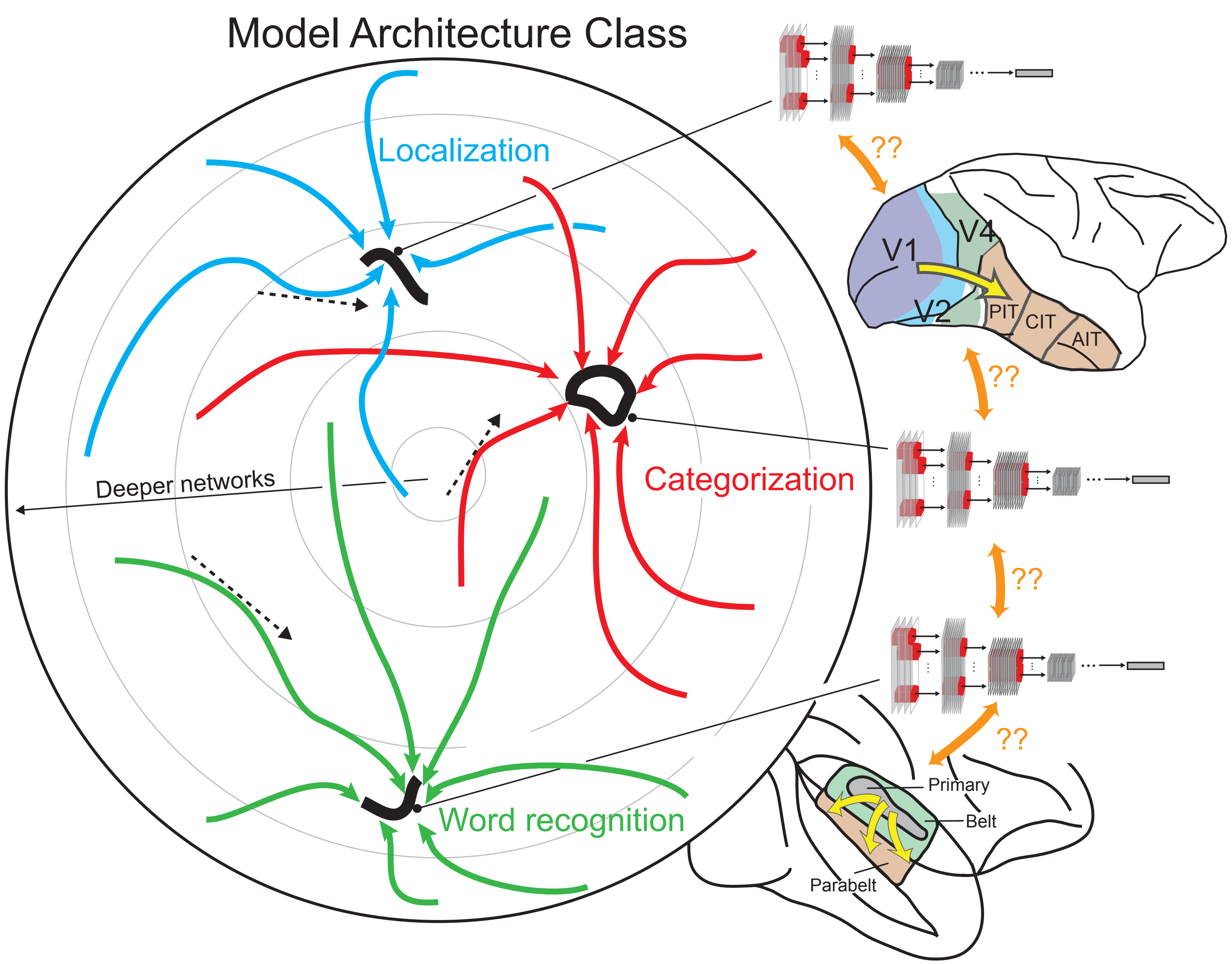
2. Choose challenging, ethologically-valid tasks (**categorization**)

3. Implement generic learning rules (**gradient descent**)

> Map to brain data. (**ventral stream**)



Yamins & DiCarlo.
***Nat. Neuro.** (2016)*



$$\underset{a \in \mathcal{A}}{\operatorname{argmin}}[L(p_a^*)]$$

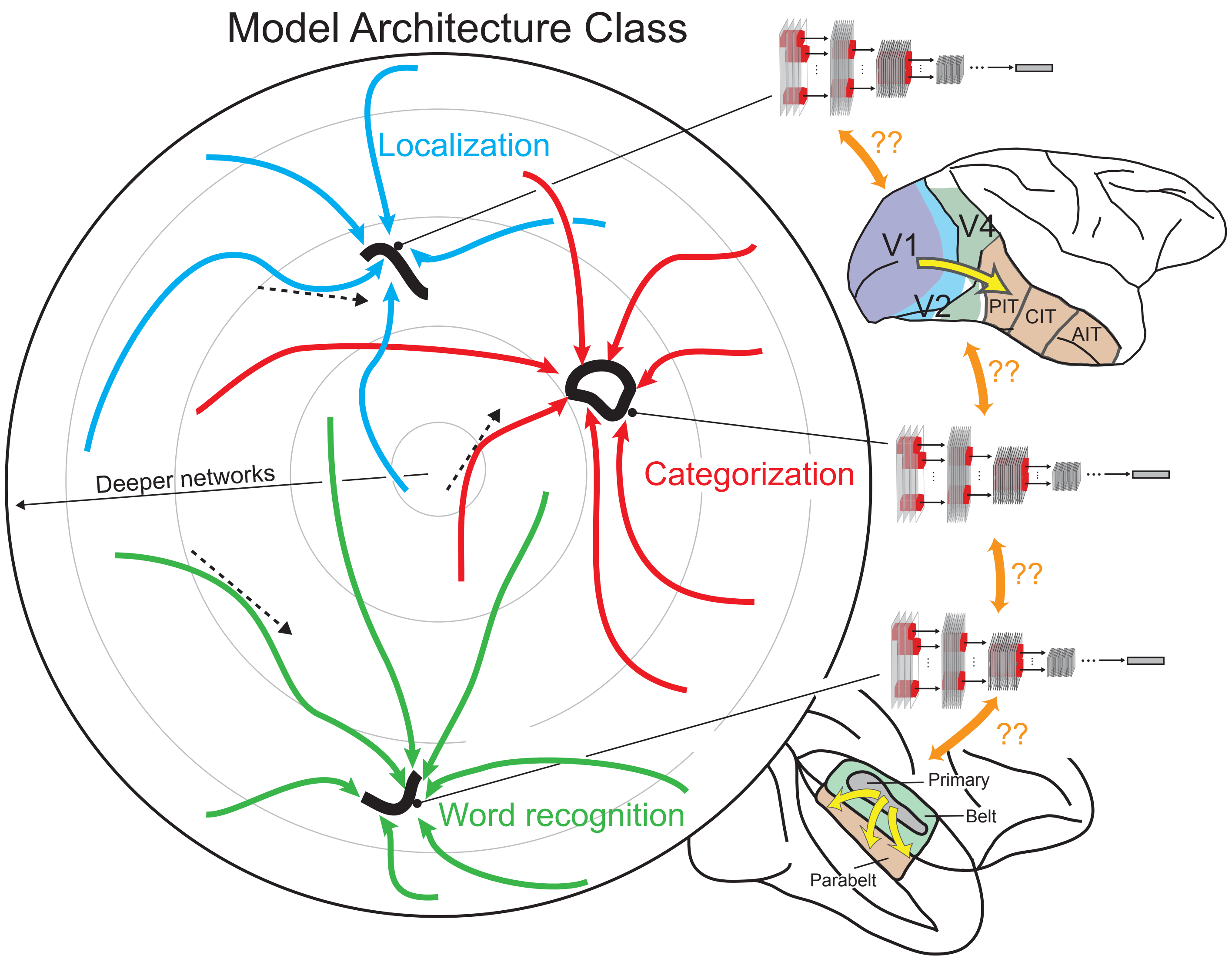
where p^* is result of

$$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

A = *architecture class*

L = *loss function*

D = *dataset*



3.

$$\underset{a\in\mathcal{A}}{\operatorname{argmin}}[L(p_a^*)]$$

where p^* is result of

$$\frac{dp_a}{dt} = -\lambda(t)\cdot\langle\nabla_{p_a}L(x)\rangle_{x\in\mathcal{D}}$$

"learning rule"

1.

A = *architecture class*

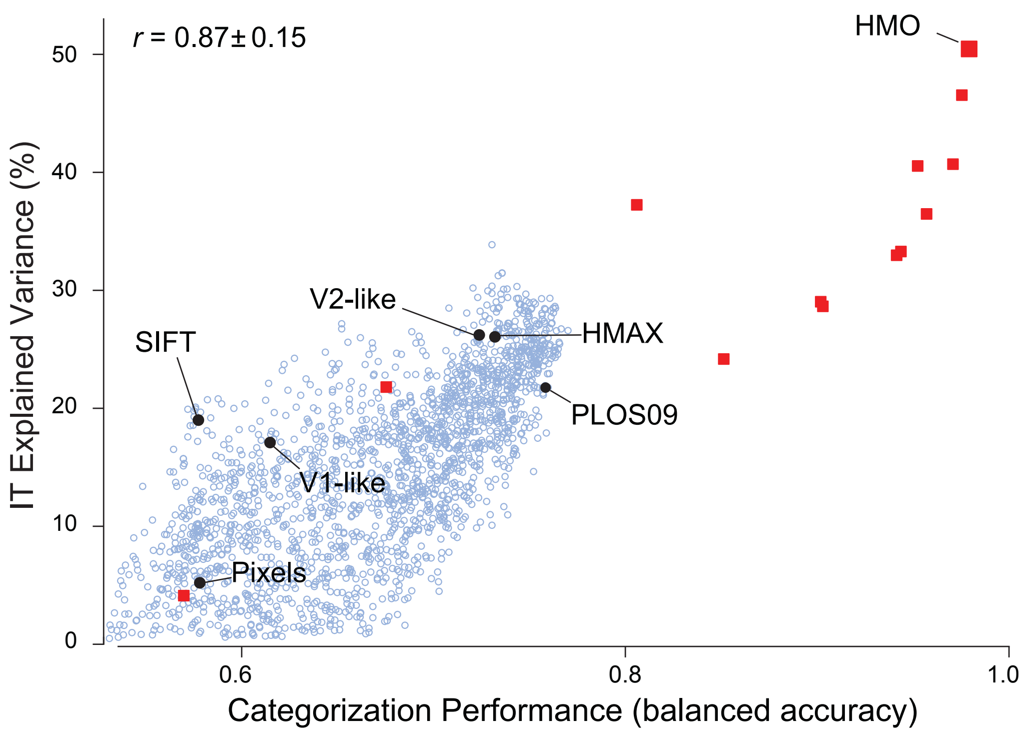
2.

L = *loss function* ***D*** = *dataset*

"task"

~~Principle~~ of "Goal-Driven Modeling"

Heuristic of "Goal-Driven Modeling"



res-net?

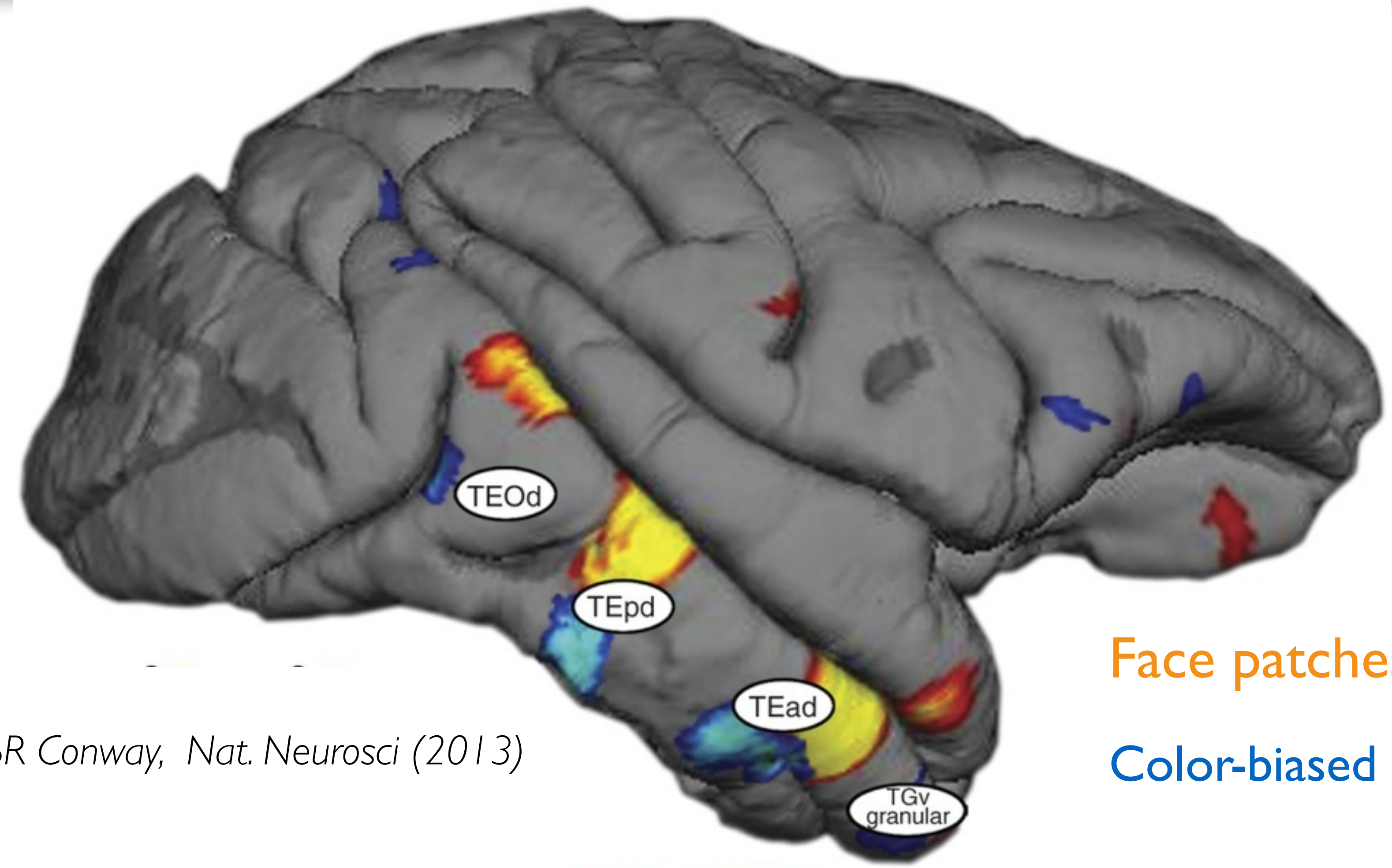
… after all at some point, for any given task, you'll probably "go over the hump" … perhaps when you exceed human performance or overfit on that task

New Lab: Sep 2016

dicarlolab.mit.edu → neuroailab.stanford.edu

MIT / BCS → Stanford

Q1: Functional Organization in Higher Visual Cortex



R. Lafer-Sousa and BR Conway, Nat. Neurosci (2013)

Face patches

Color-biased regions

Regions selective for:

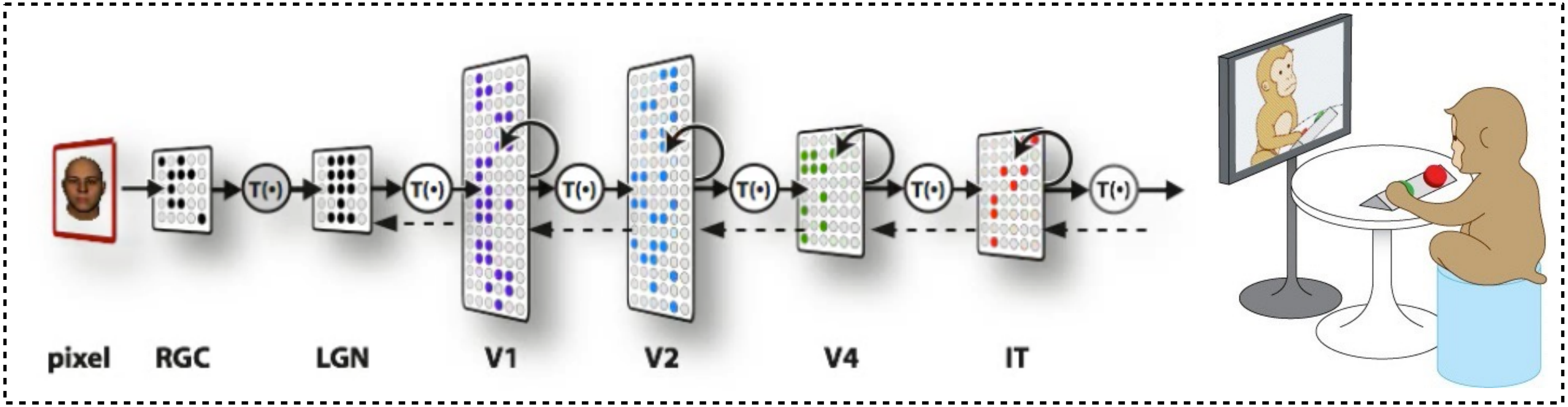
- faces
- places
- bodies
- color

Where do these patches come from?

- In-born built-in structure??
- or developmentally determined by domain-specific experience?

Q2: Intermediate Visual Tasks/Properties?

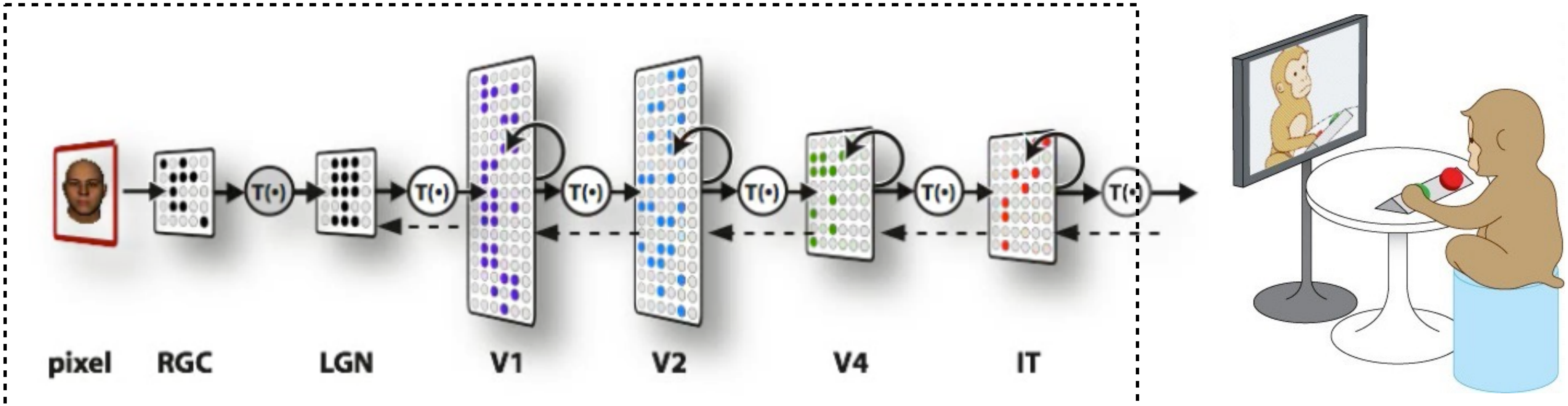
←



Categorization

Q2: Intermediate Visual Tasks/Properties?

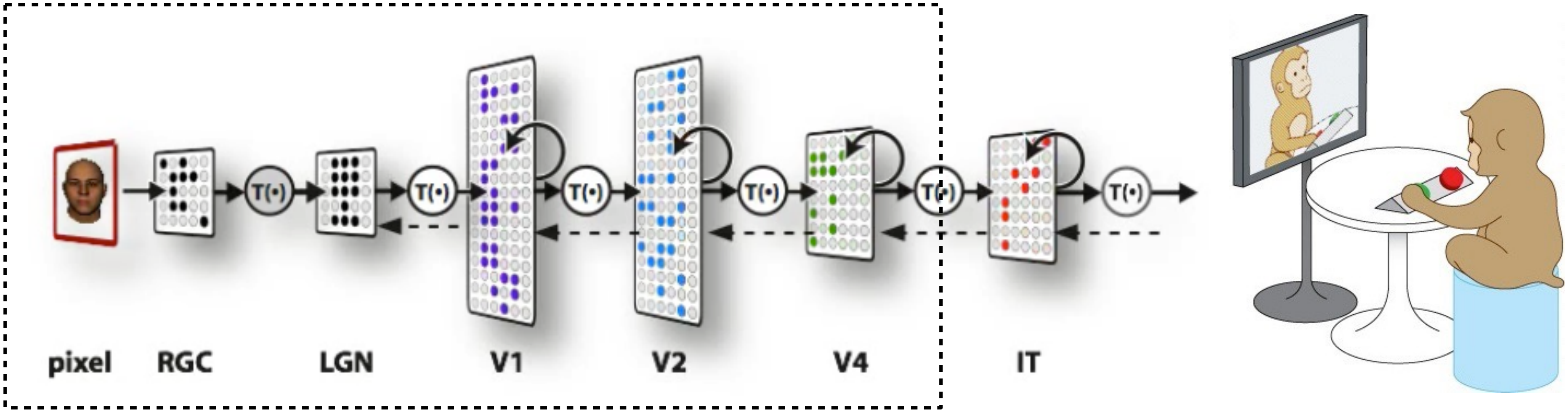
←



Categorization

Q2: Intermediate Visual Tasks/Properties?

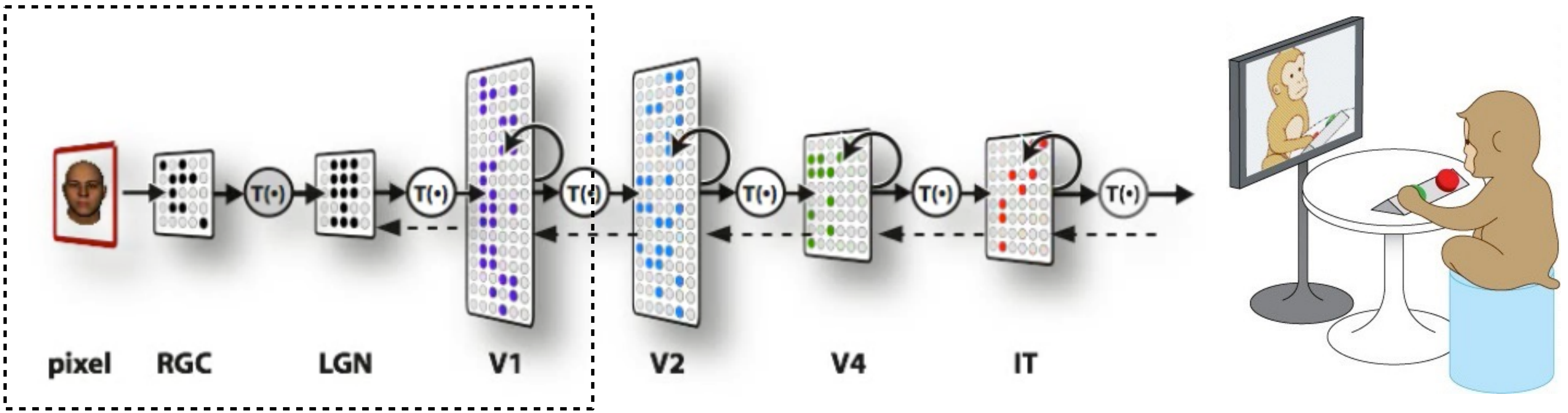
←



Categorization

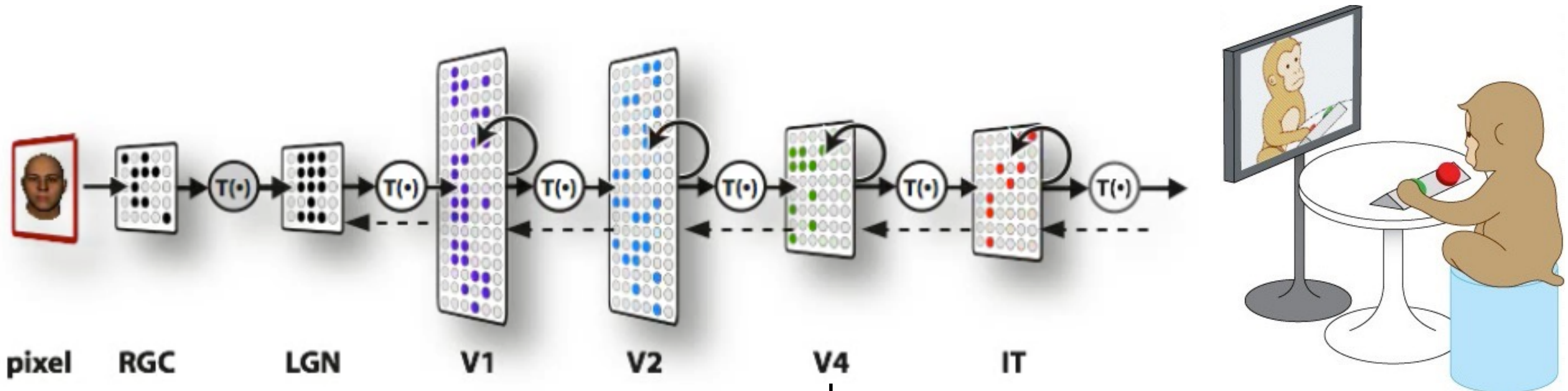
Q2: Intermediate Visual Tasks/Properties?

←



Categorization

Q2: Intermediate Visual Tasks/Properties?



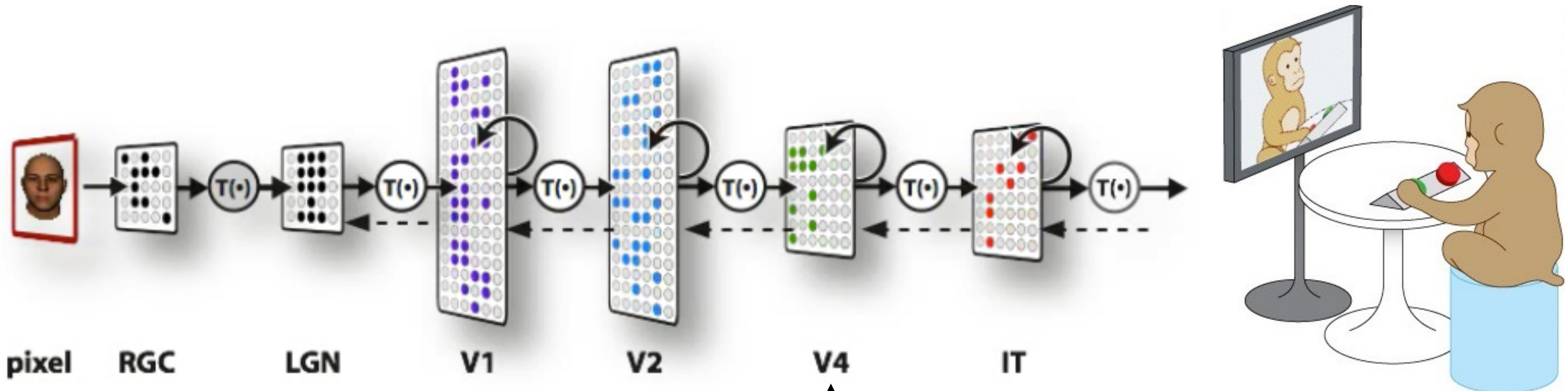
Categorization

virtual electrophysiology
in models

Shape / curvature?

"Second-order"
conjunctions

Q2: Intermediate Visual Tasks/Properties?



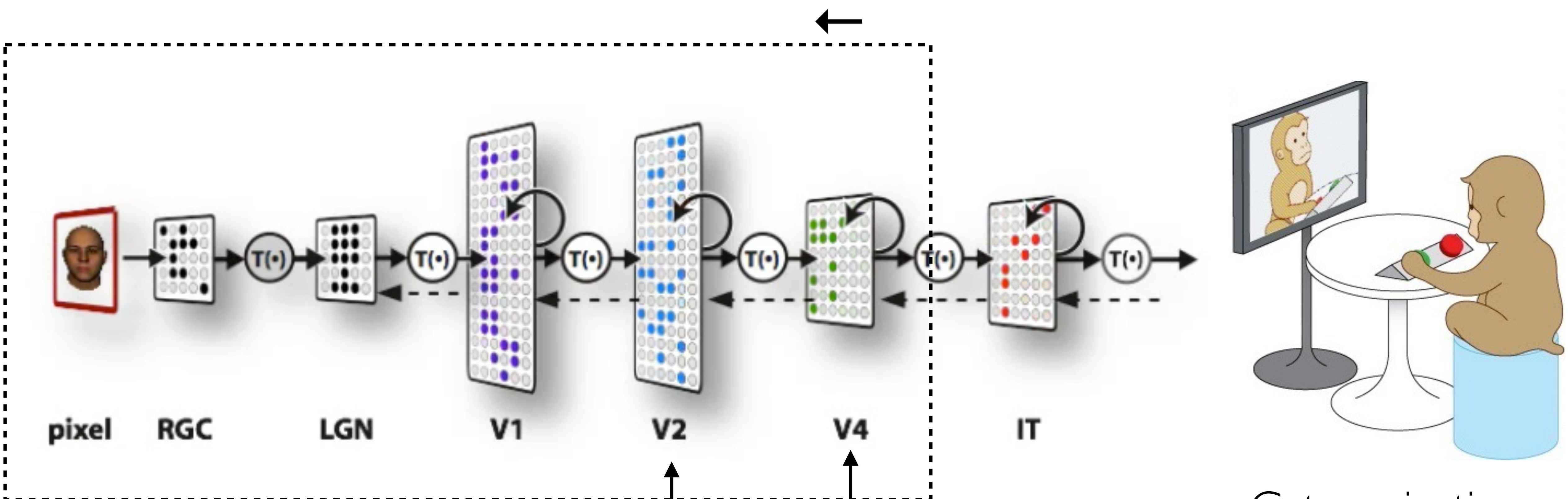
Categorization

independent goal-driven constraint?

Shape / curvature?

"Second-order" conjunctions

Q2: Intermediate Visual Tasks/Properties?



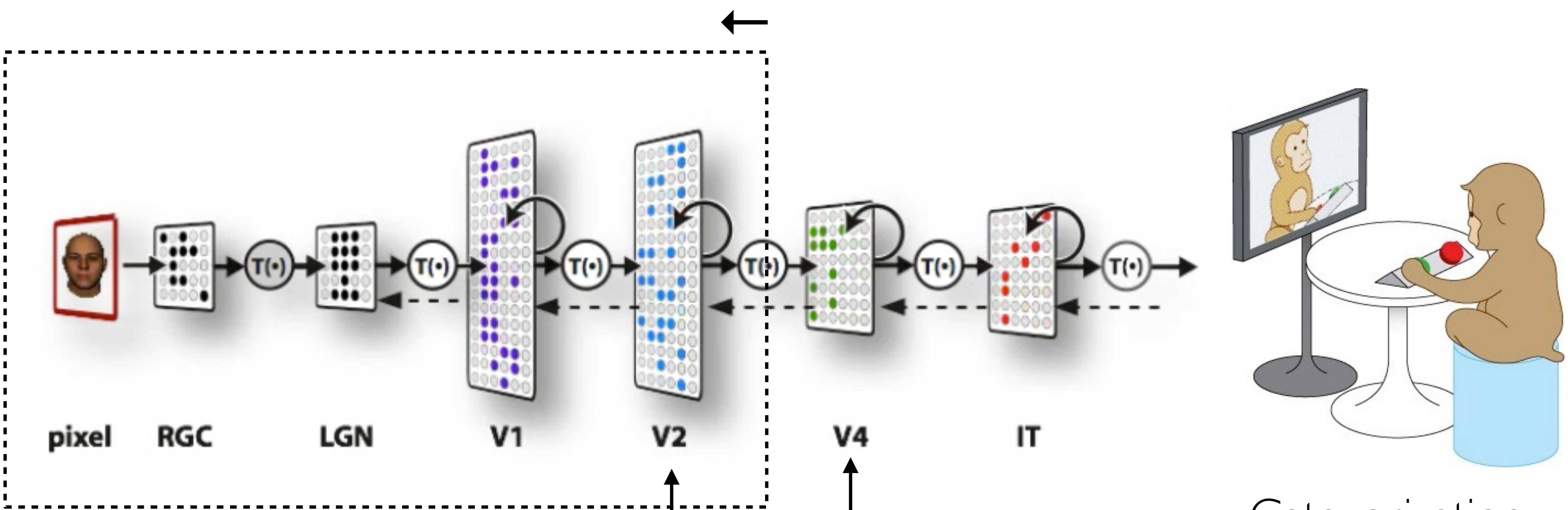
Categorization

independent goal-driven constraint?

Shape / curvature?

"Second-order" conjunctions

Q2: Intermediate Visual Tasks/Properties?



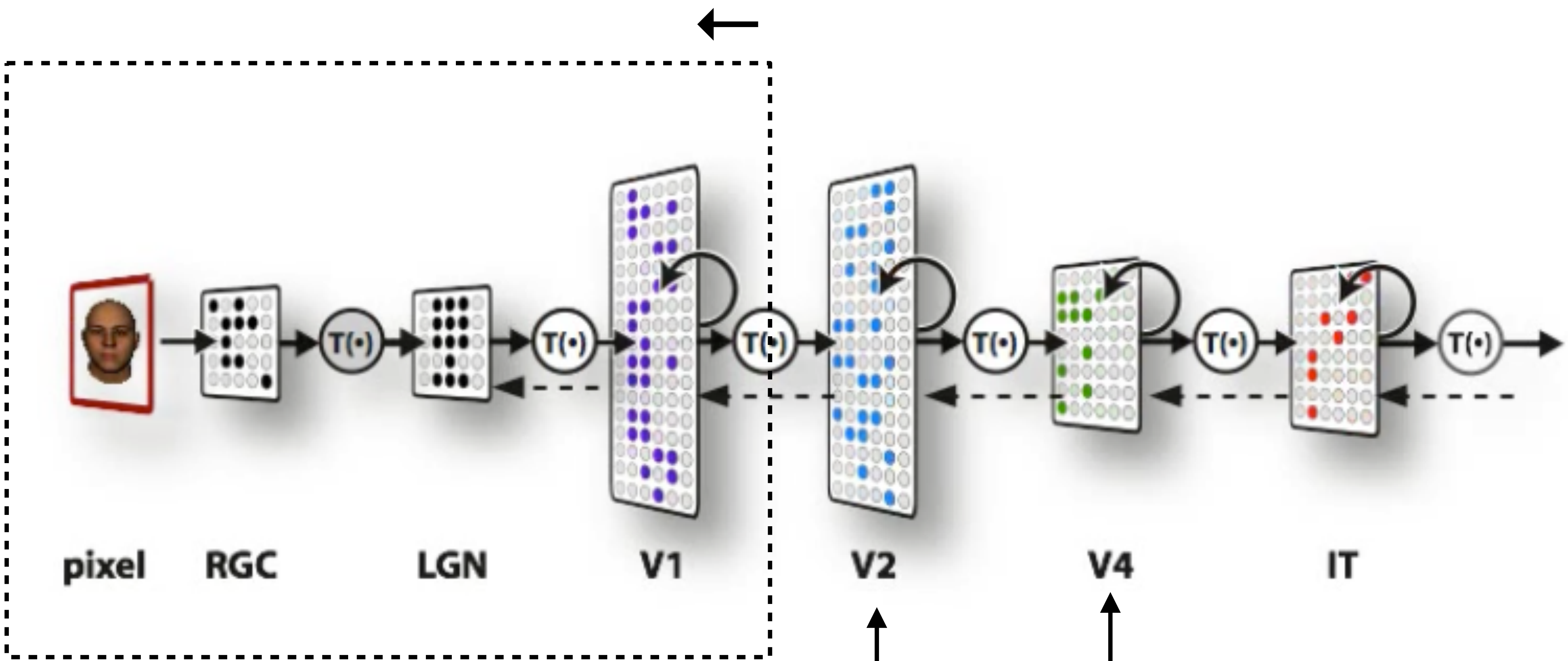
Categorization

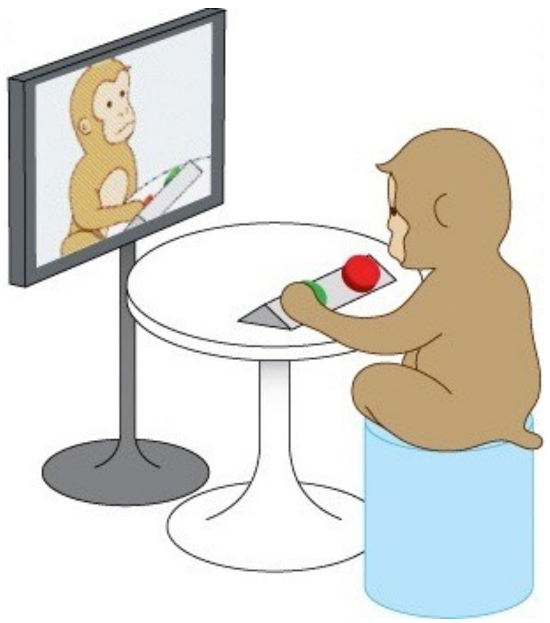
independent goal-driven constraint?

Shape / curvature?

"Second-order" conjunctions

Q2: Intermediate Visual Tasks/Properties?





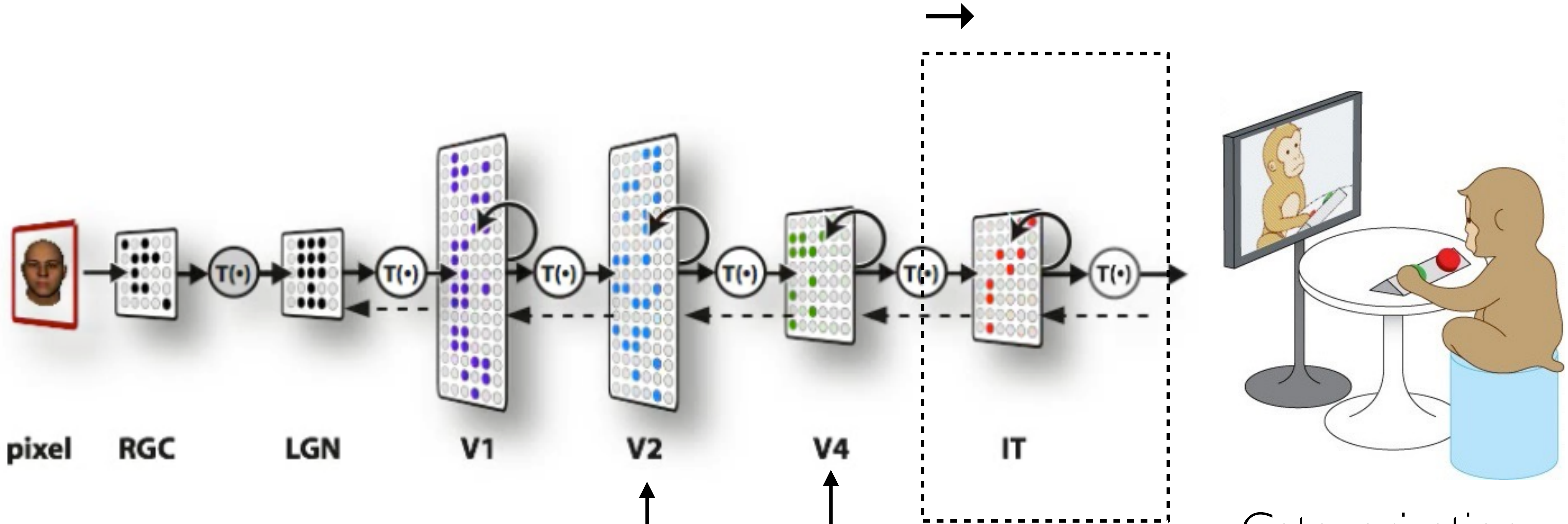
Categorization

independent goal-driven constraint?

Shape / curvature?

"Second-order" conjunctions

Q2: Intermediate Visual Tasks/Properties?



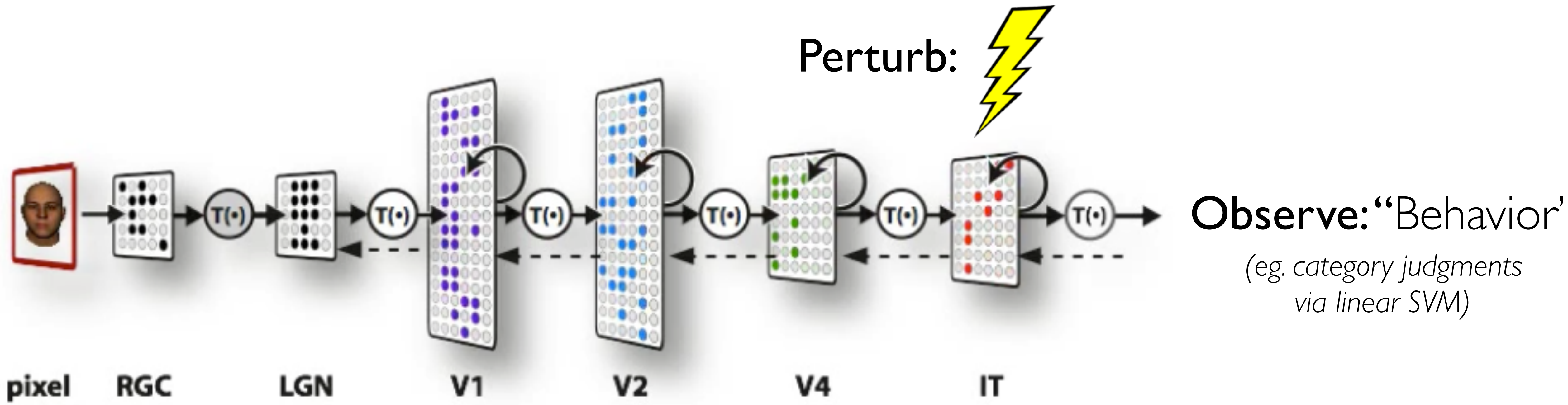
Categorization

independent goal-driven constraint?

Shape / curvature?

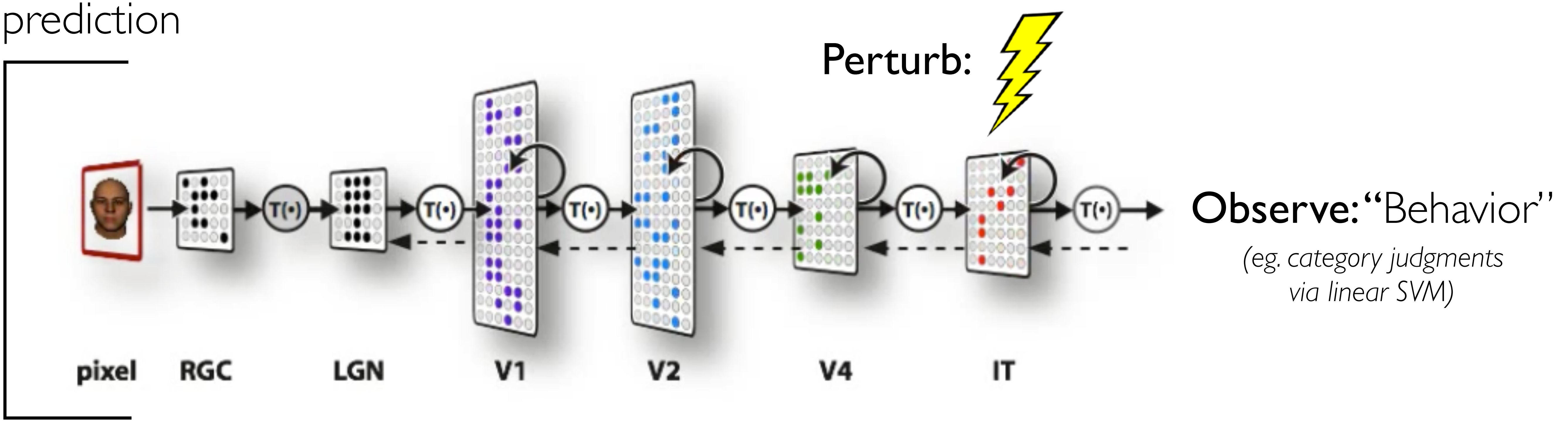
"Second-order" conjunctions

Q3: Predicting "In-Head" Neural Perturbations



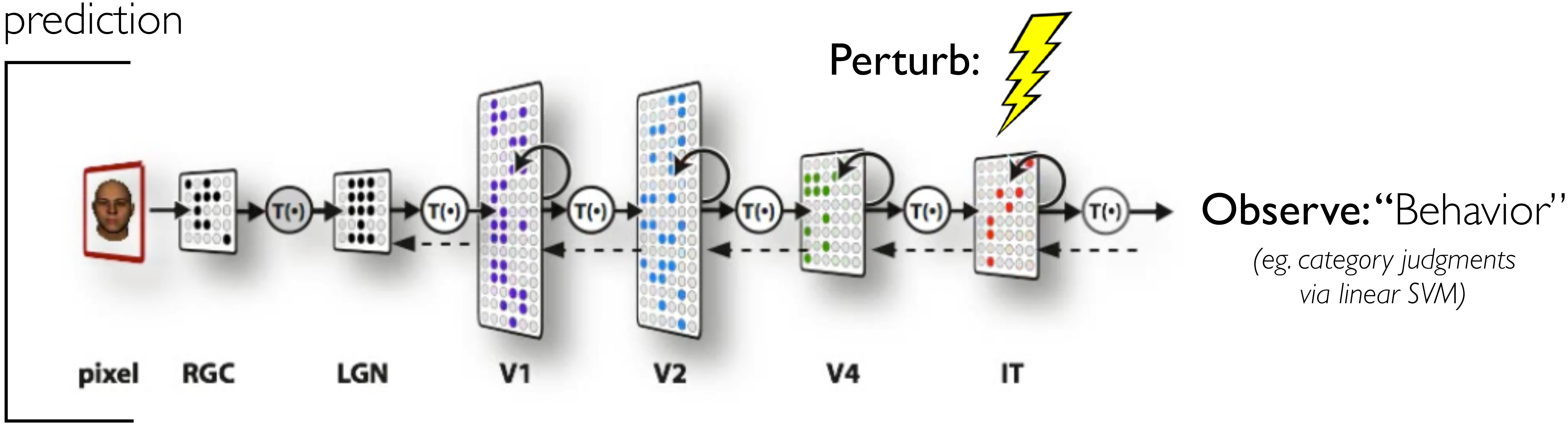
Q3: Predicting "In-Head" Neural Perturbations

prediction

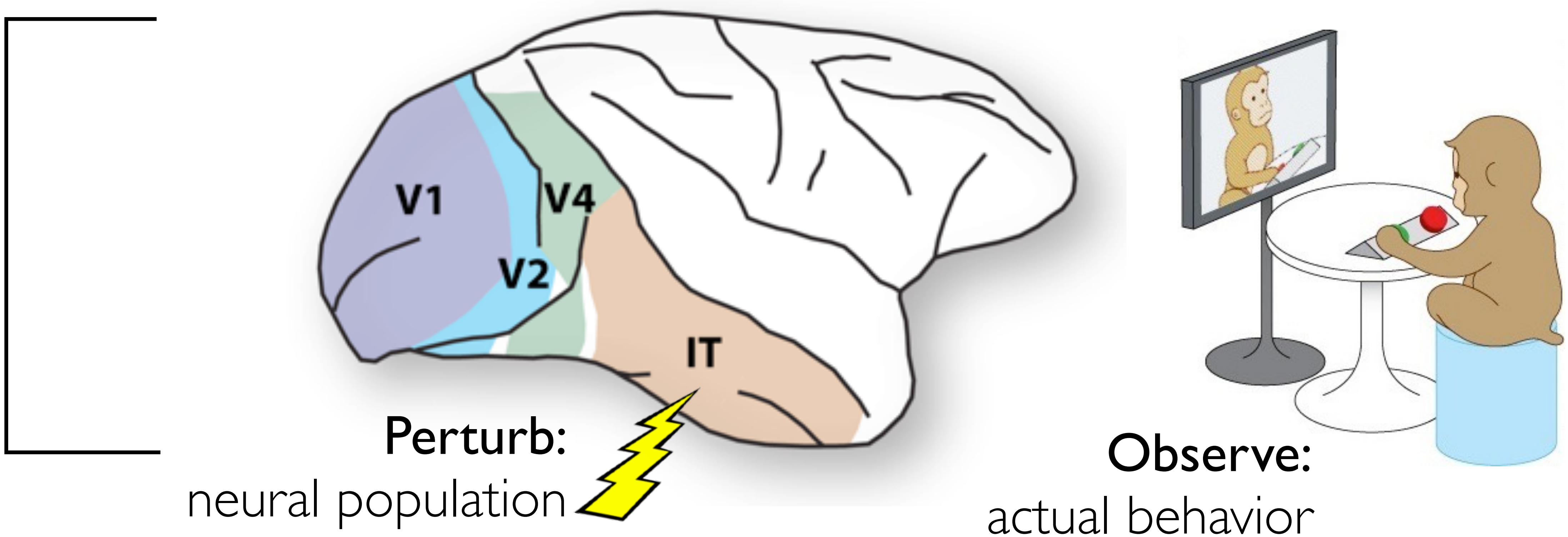


Q3: Predicting "In-Head" Neural Perturbations

prediction

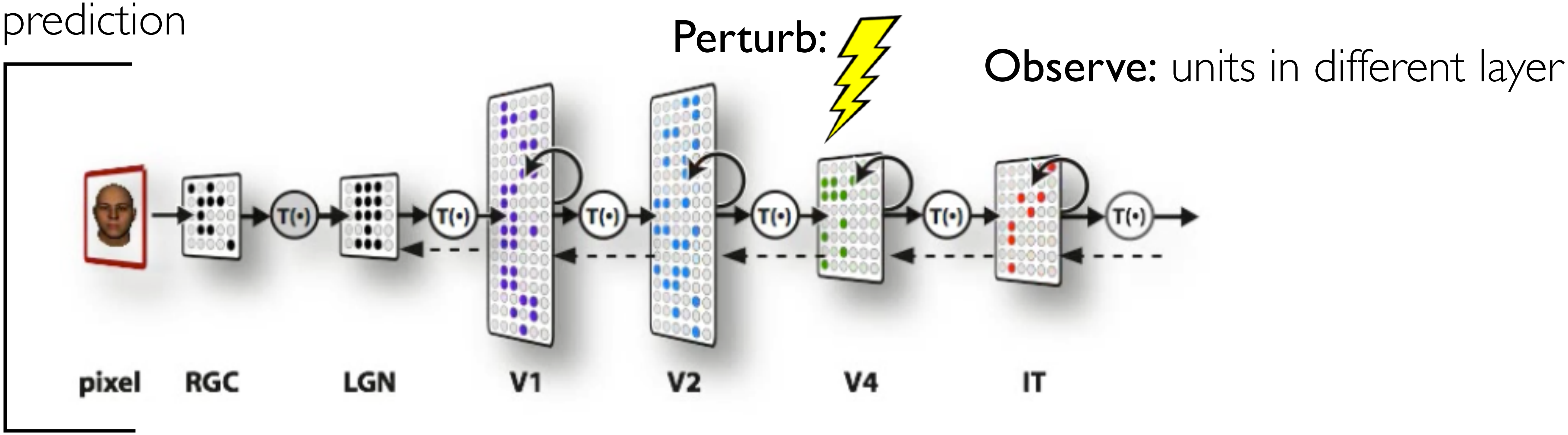


measurement

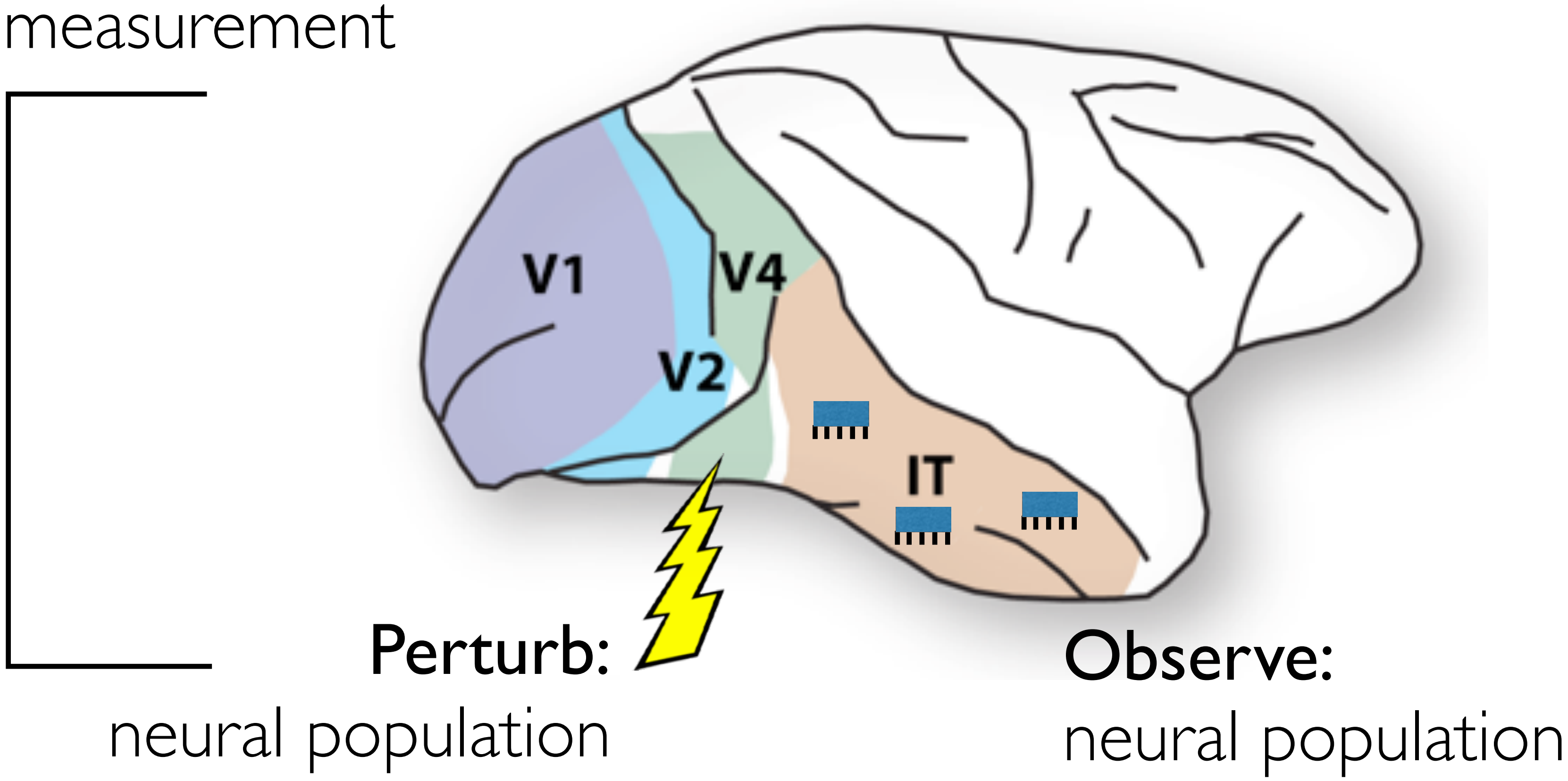


Q3: Predicting "In-Head" Neural Perturbations

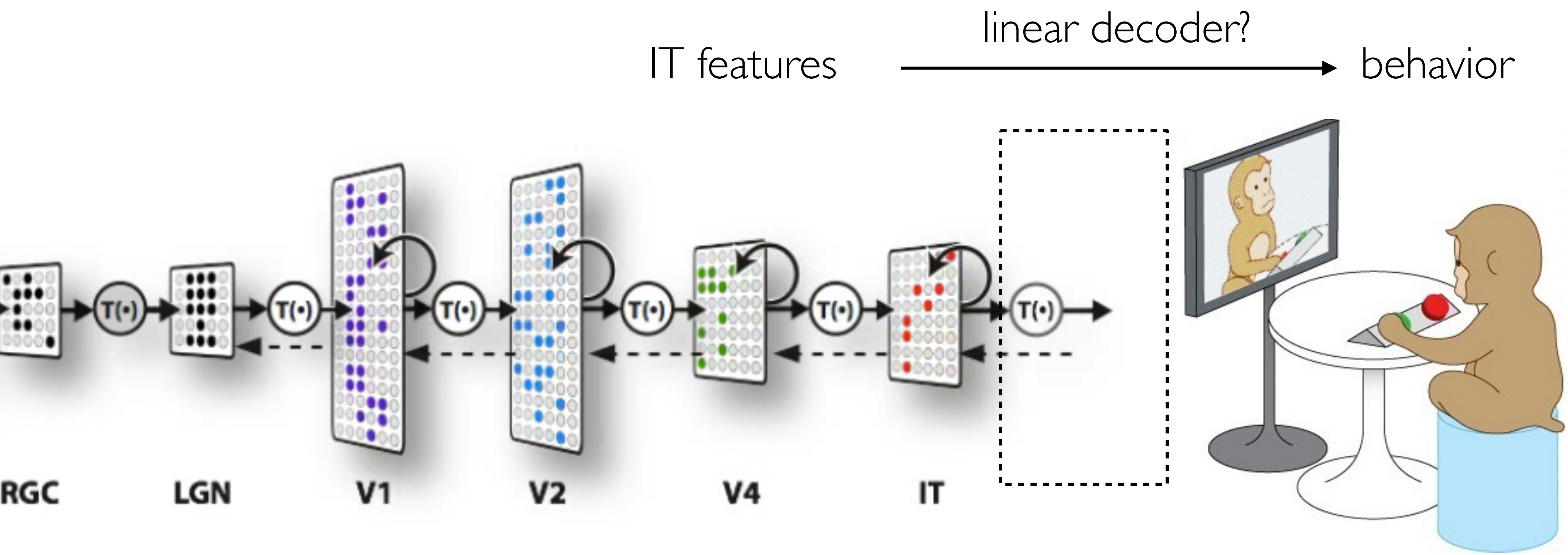
prediction



measurement

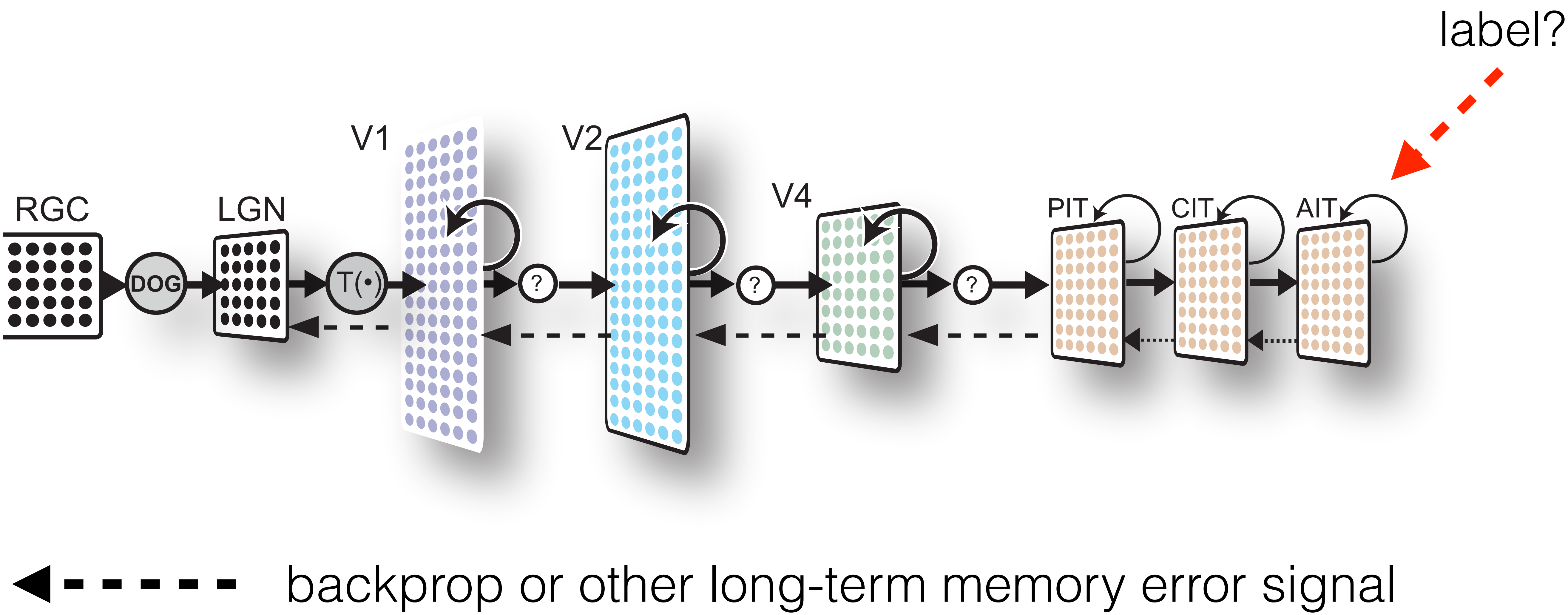


Q4: Linear Readout as a Model of Neural Decoding



- Nonlinear? Temporal instead of rate code?
- View the process realtime?
    - XX Put in Rishi learning slides
- How are they learned? What types of (e.g.) regularizations are implemented?
- Default readouts and task switching (“hyperplane management”)

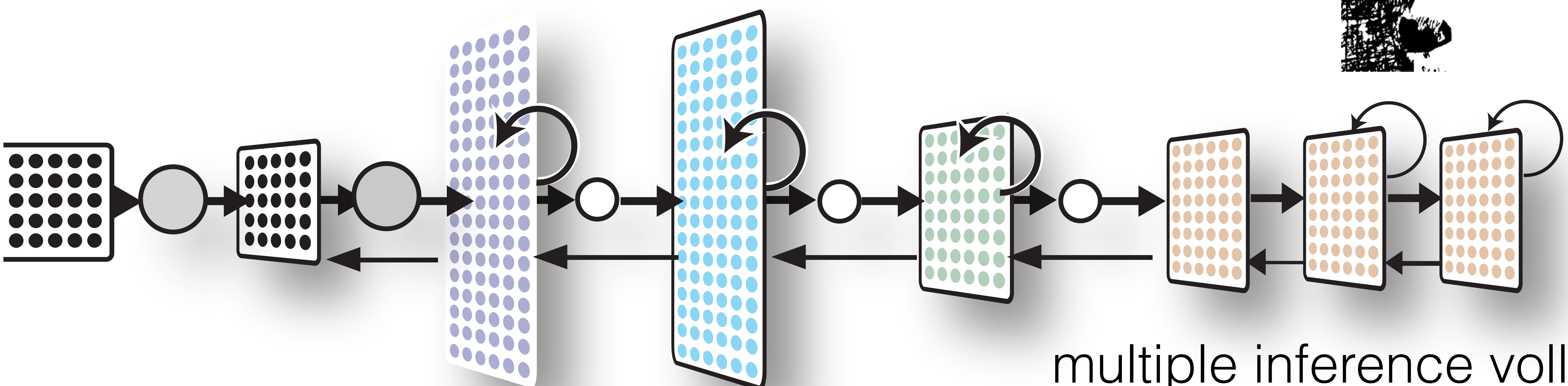
H1: Feedback implements learning of filters — form of long-term memory — not online



Q5: Role of Top-Down, Recurrent Connections

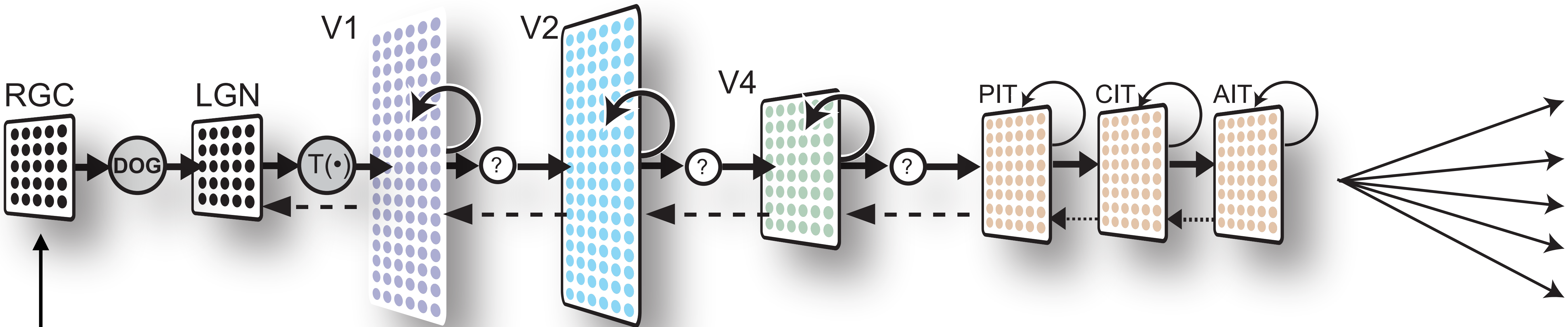
H2: Feedback solves <u>hard cases</u> that aren't embedded in single feedforward volley …. like ambiguity. Online inference in the ventral stream.

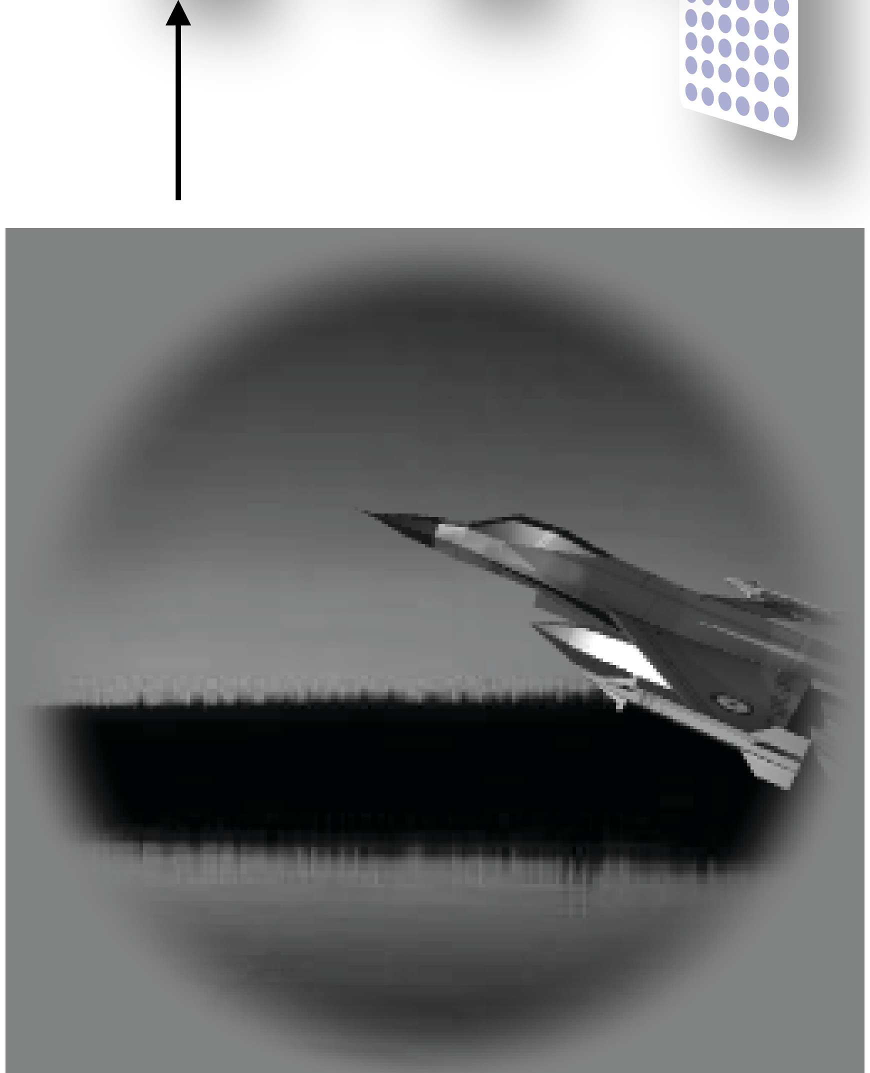
(Dallenbach's cow)



multiple inference volleys

H3a: Downstream, online feedback helps solve dynamic problems like (e.g.) task switching.

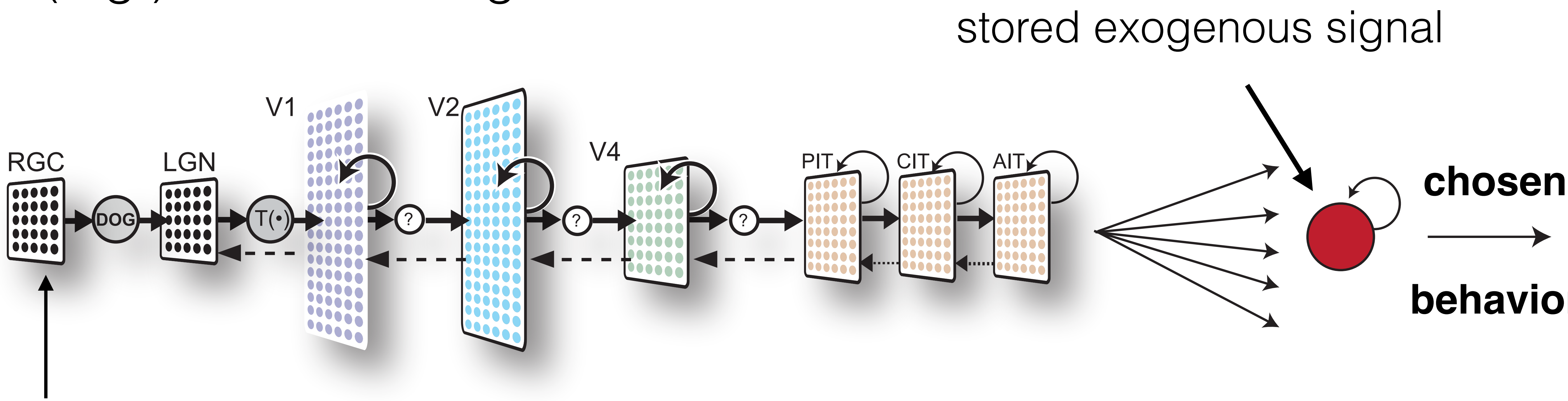




- category
- identity
- position
- size
- pose
- …

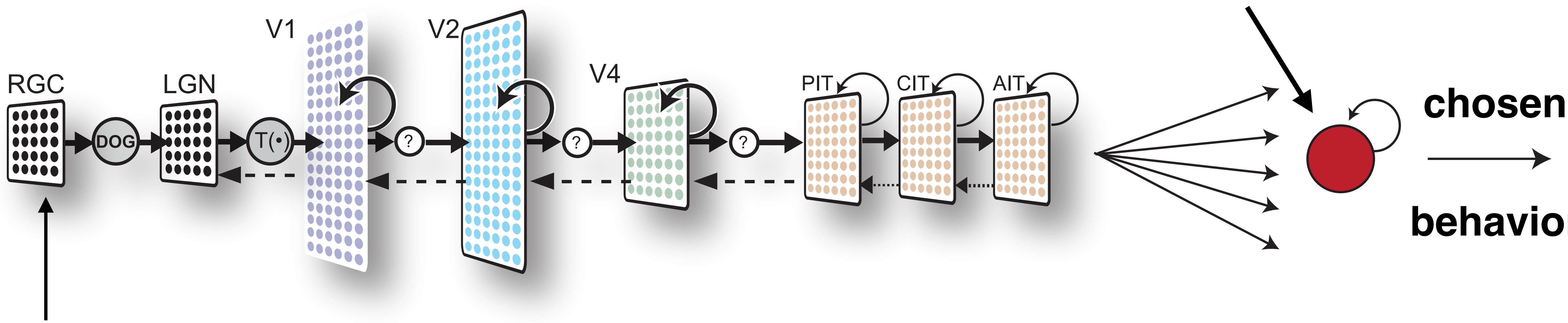
Q5: Role of Top-Down, Recurrent Connections

H3a: Downstream, online feedback helps solve dynamic problems like (e.g.) task switching.



- category
- identity
- position
- size
- pose
- …

H3a: Downstream, online feedback helps solve dynamic problems like (e.g.) task switching.

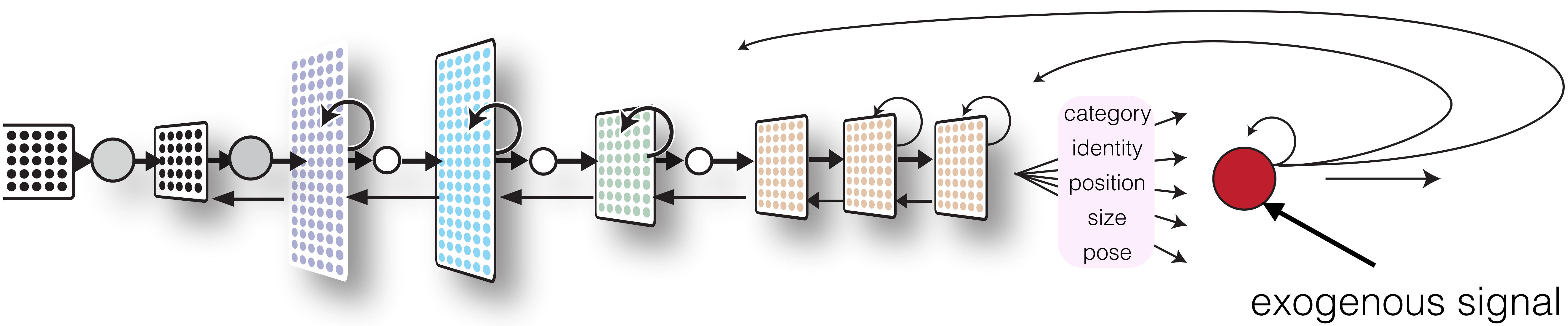


"Hyperplane Management"

- category
- identity
- position
- size
- pose
- …

Q5: Role of Top-Down, Recurrent Connections

H3b: Task switching algorithms reach down into ventral stream — benefit from nonlinear combinations of gating variable and IT features — e.g., attentional effects



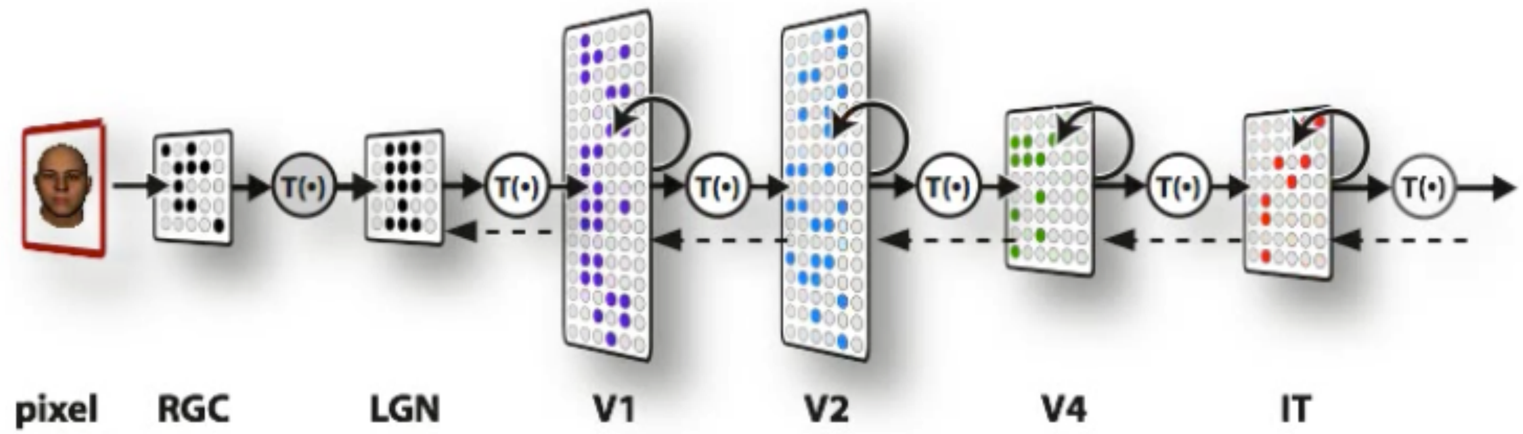
Distinguished ventrally from **H2** (“hard cases”) by the nature of the task that elicits it — (e.g.) pre-cuing (volitional control) instead of (e.g.) passive viewing.

Q5: Role of Top-Down, Recurrent Connections

Something about relationship to generative models

Q6: How Is Visual Learning Implemented?

Many parameters, **P**



How are they learned?

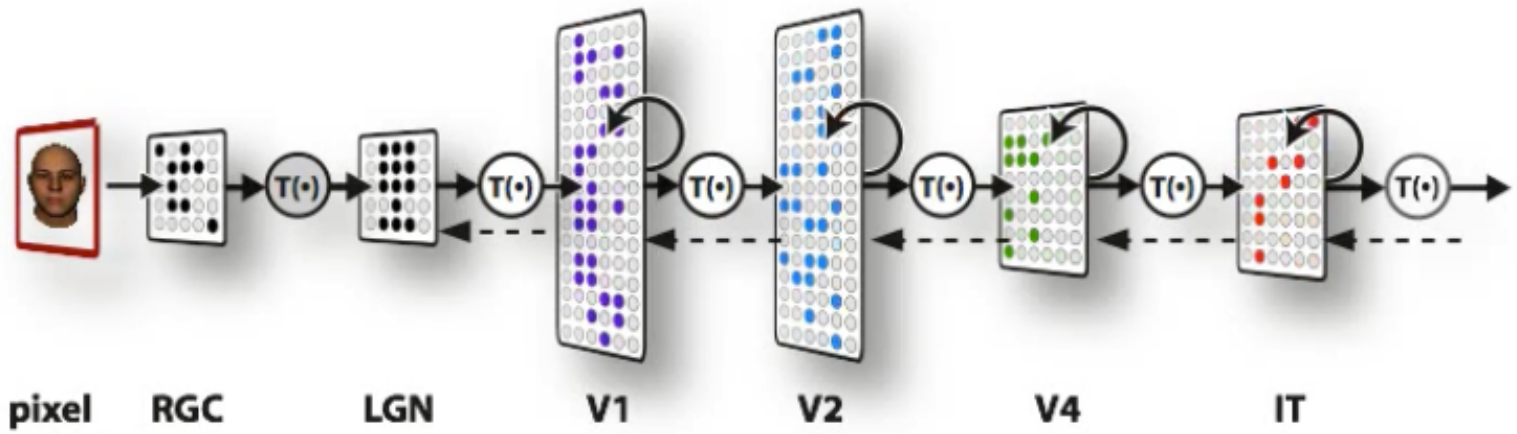
Gradient descent eq: $$\frac{dp}{dt} = -\lambda(t) \cdot \frac{\partial L}{\partial P}$$

L = loss function

λ = learning rate

Q6: How Is Visual Learning Implemented?

Many parameters, **P**



How are they learned?

Gradient descent eq: $$\frac{dp}{dt} = -\lambda(t) \cdot \frac{\partial L}{\partial P}$$

L = loss function

λ = learning rate

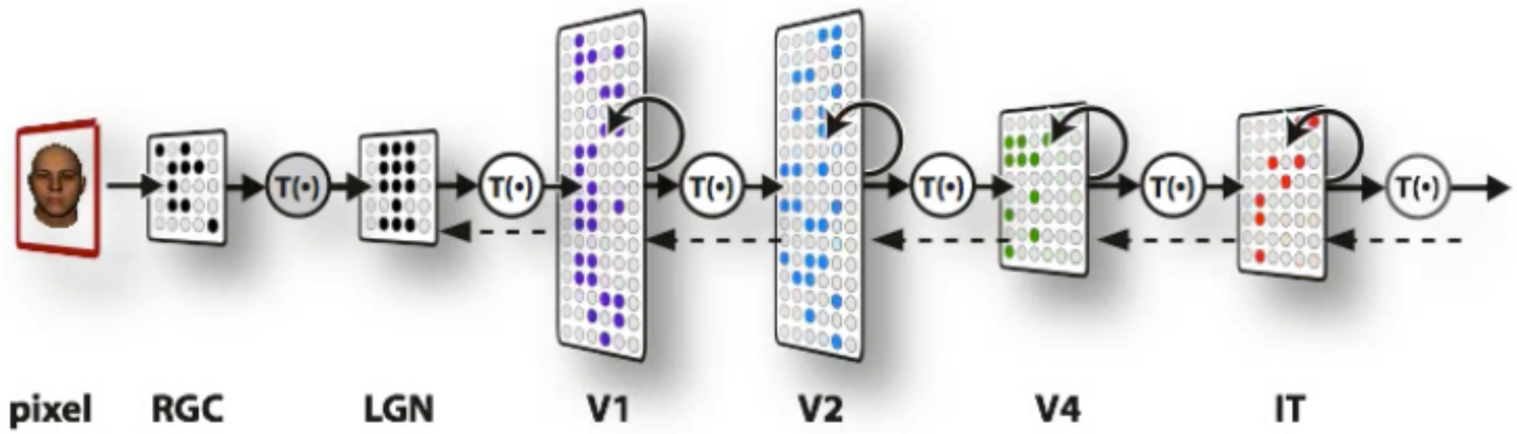
In current standard practice:

L = soft-max loss computed relative large numbers of externally-provided semantic labels.

Q6: How Is Visual Learning Implemented?

Many parameters, **P**

How are they learned?



Gradient descent eq: $$\frac{dp}{dt} = -\lambda(t) \cdot \frac{\partial L}{\partial P}$$

L = loss function

λ = learning rate

In current standard practice:

L = loss computed via large numbers of externally-provided semantic labels.

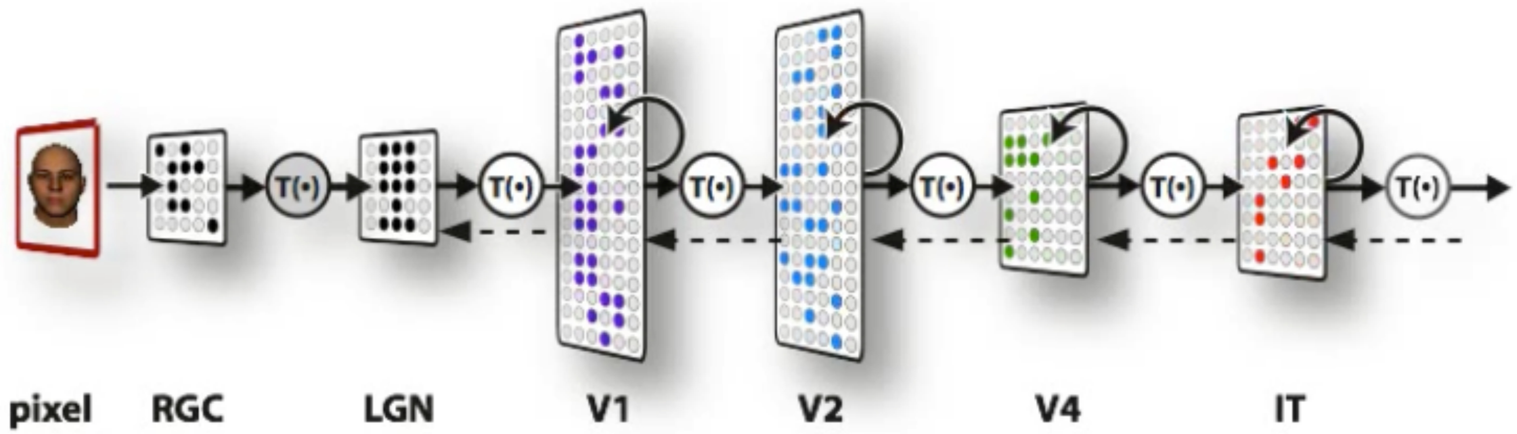
Ideally:

L = un- or semi-supervised function computable from easily accessible data about agent's environment

Q6: How Is Visual Learning Implemented?

Many parameters, **P**

How are they learned?



Gradient descent eq: $$\frac{dp}{dt} = -\lambda(t) \cdot \frac{\partial L}{\partial P}$$

L = loss function

λ = learning rate

(1) Which parameters are learned vs developed or evolved?

(2) What are the right loss functions(s)?

(3) How are the loss functions and the GDE implemented/ approximated via neural circuits?

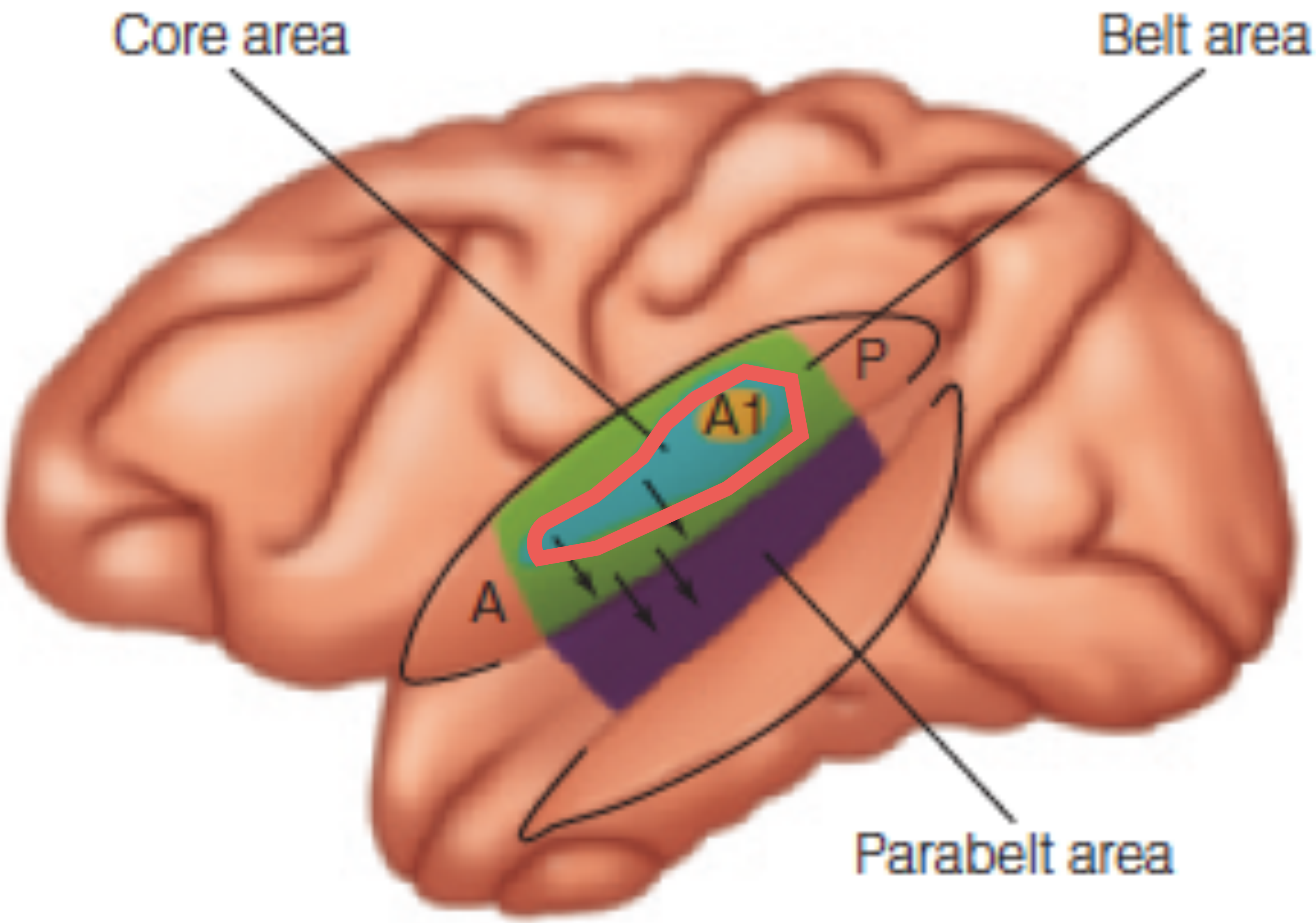
Q7: Other Sensory Domains?

Can goal-drive modeling approaches generalize to other areas?

For example, in auditory cortex:

- can HCNN models explain higher auditory cortex?
- which tasks best explain functional organization of AC?
- how to auditory-optimized architectures related to visual ones?

Core / Belt / Parabelt Structure

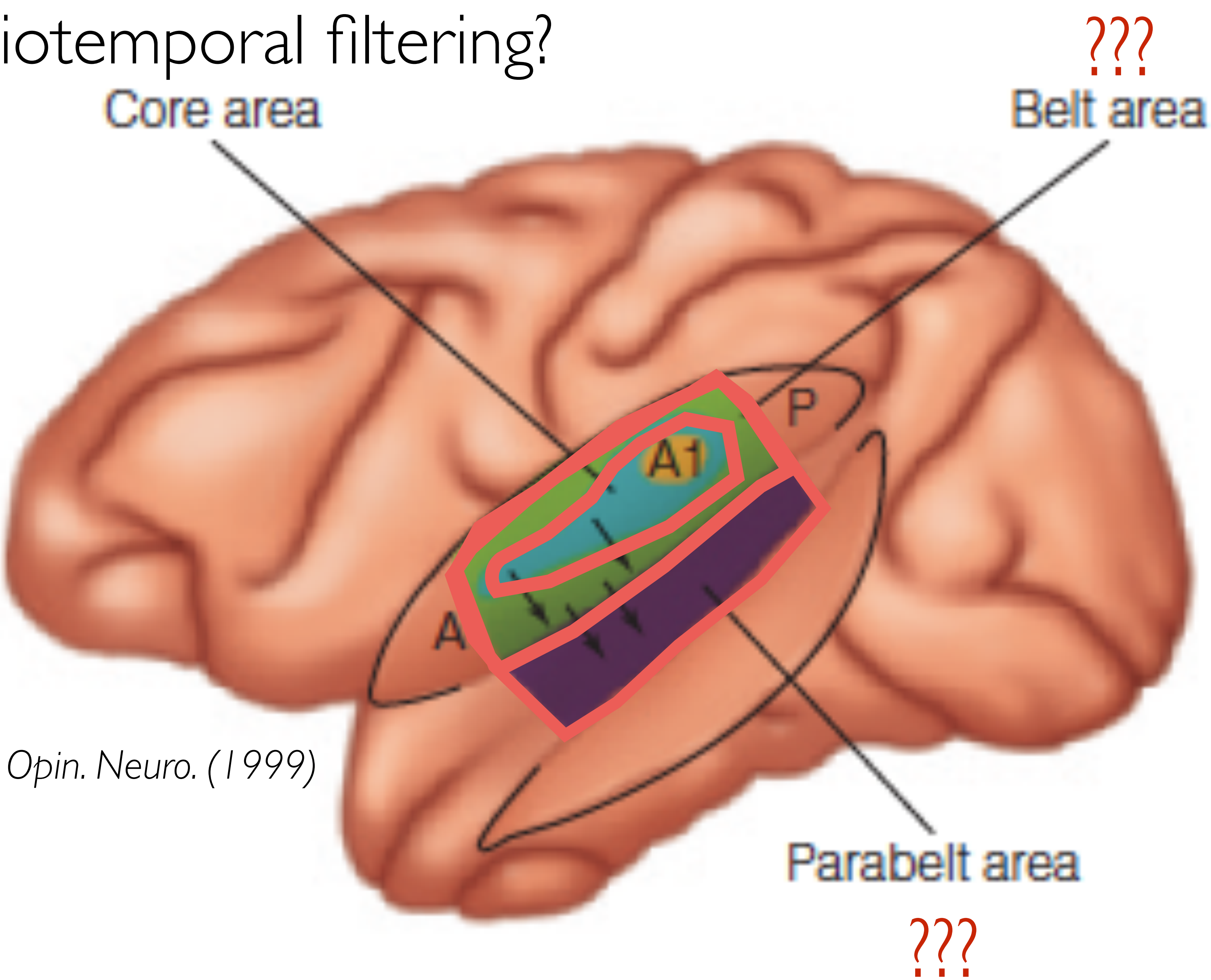


***monkey**
*

Tramo et. al, Curr. Opin. Neuro. (1999)

Core / Belt / Parabelt Structure

Spatiotemporal filtering?

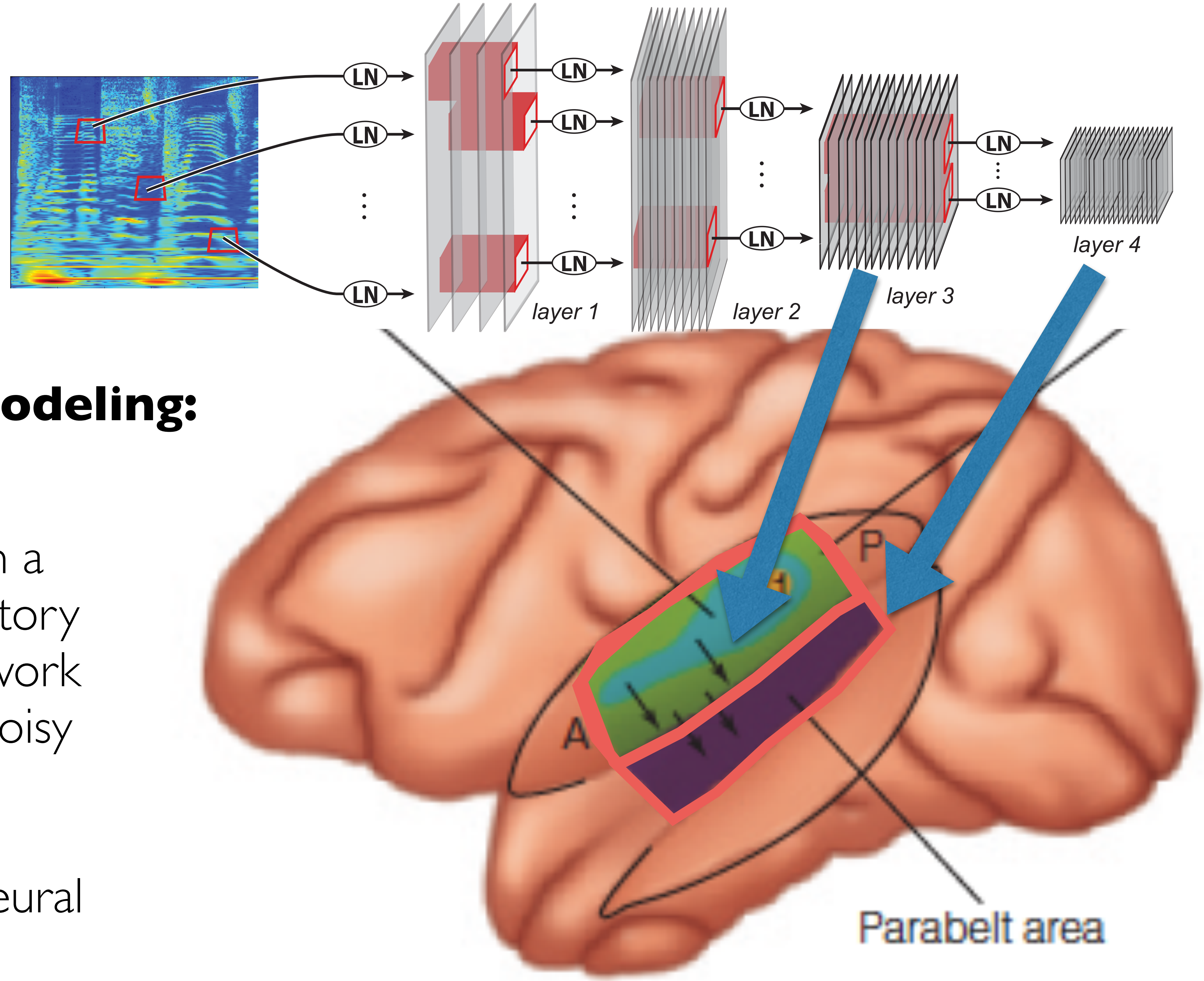


Tramo et. al, Curr. Opin. Neuro. (1999)

*__monkey__
*

Example: use computational models to help deepen understanding of non-primary areas.

Core Task-Driven Modeling Idea



Task-Driven Modeling:

1. Optimize for performance on a challenging auditory task (600-way work recognition in noisy speech)

2. Compare to neural data.

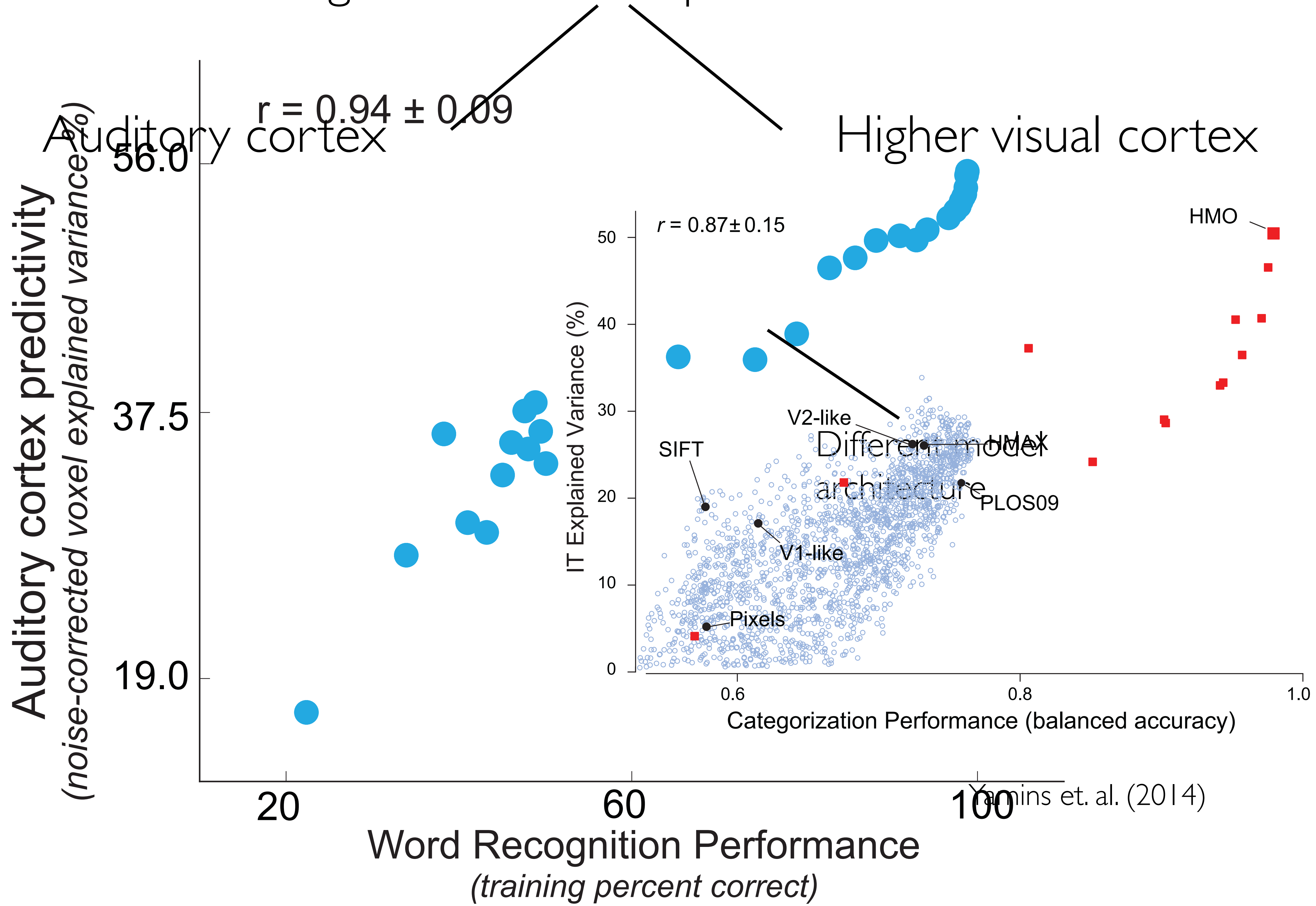
Apply to auditory tasks, where the regions themselves are less well known.

Which layer best predicts each voxel's responses?

Primary auditory cortex: predicted by **lower CNN layers**.
Non-primary auditory cortex: predicted by **higher CNN layers**.

Analysis of Model Architectures

High-variation task performance vs:



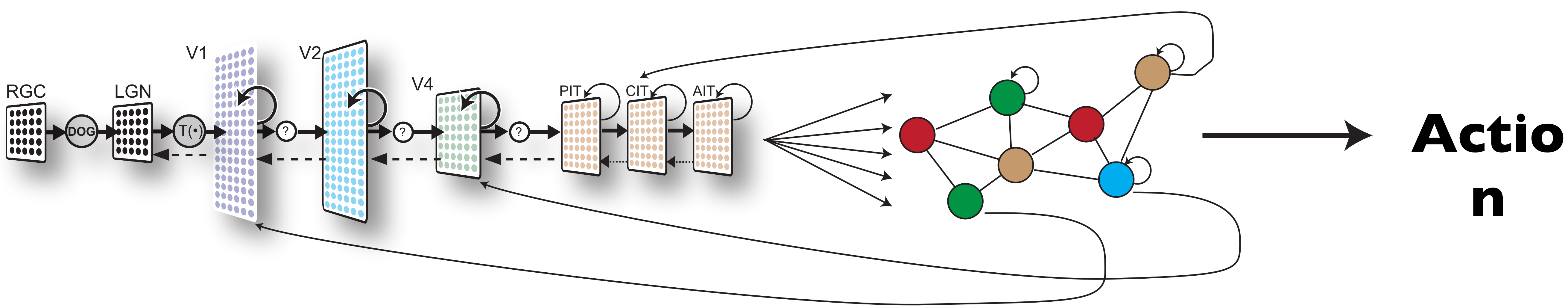
Yamins et. al. (2014)

Q8: Integration of Working Memory

Many visual behaviors beyond vision at a glance (e.g.):

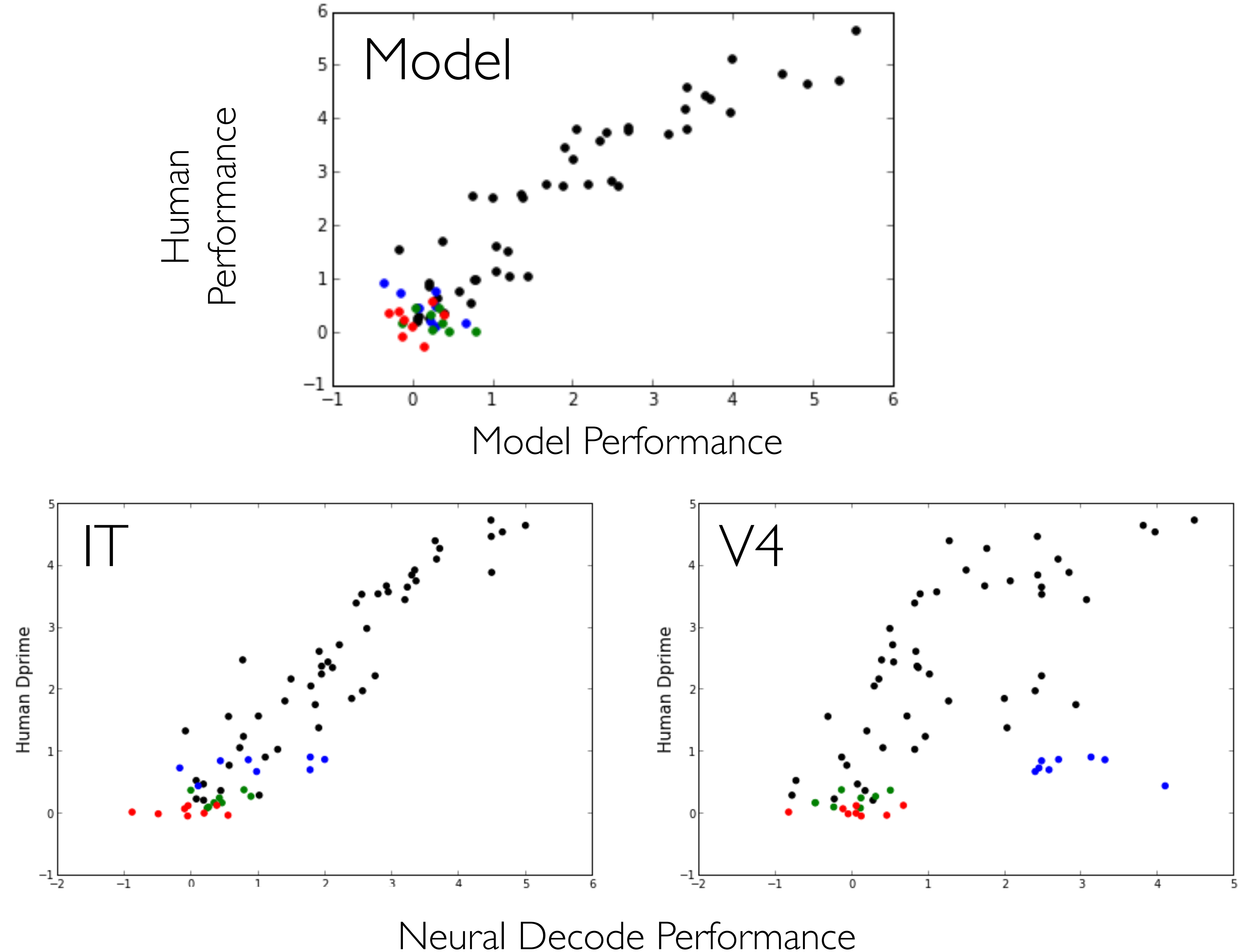
- Scene understanding over multiple saccades
- Strategic decision-making in complex environments

Involve integration of working memory, likely via RNNs.



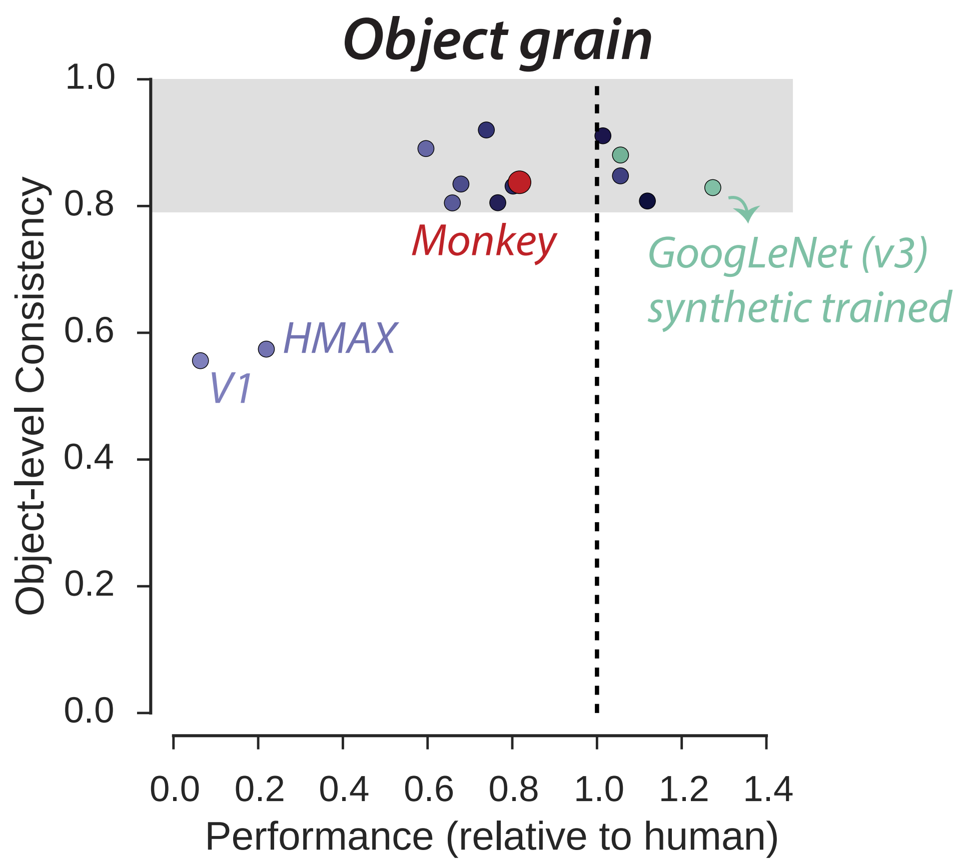
Digging deeper into understanding visual cortex

Match between models and data at category confusion level is pretty good …

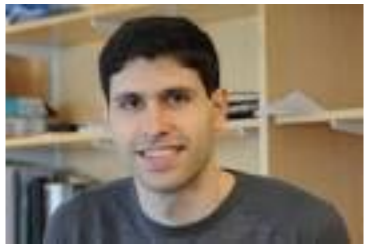


Digging deeper into understanding visual cortex

Match between models and data at category confusion level is pretty good …



work of:



Elias
Issa

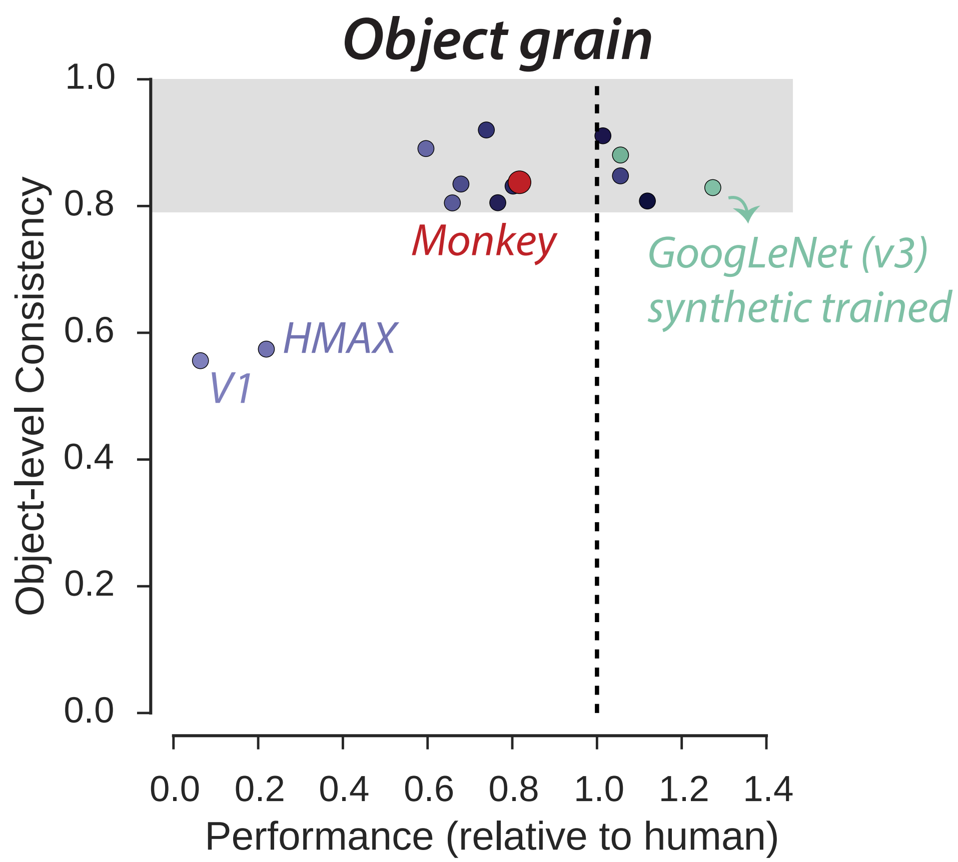
Rishi
Rajalingham

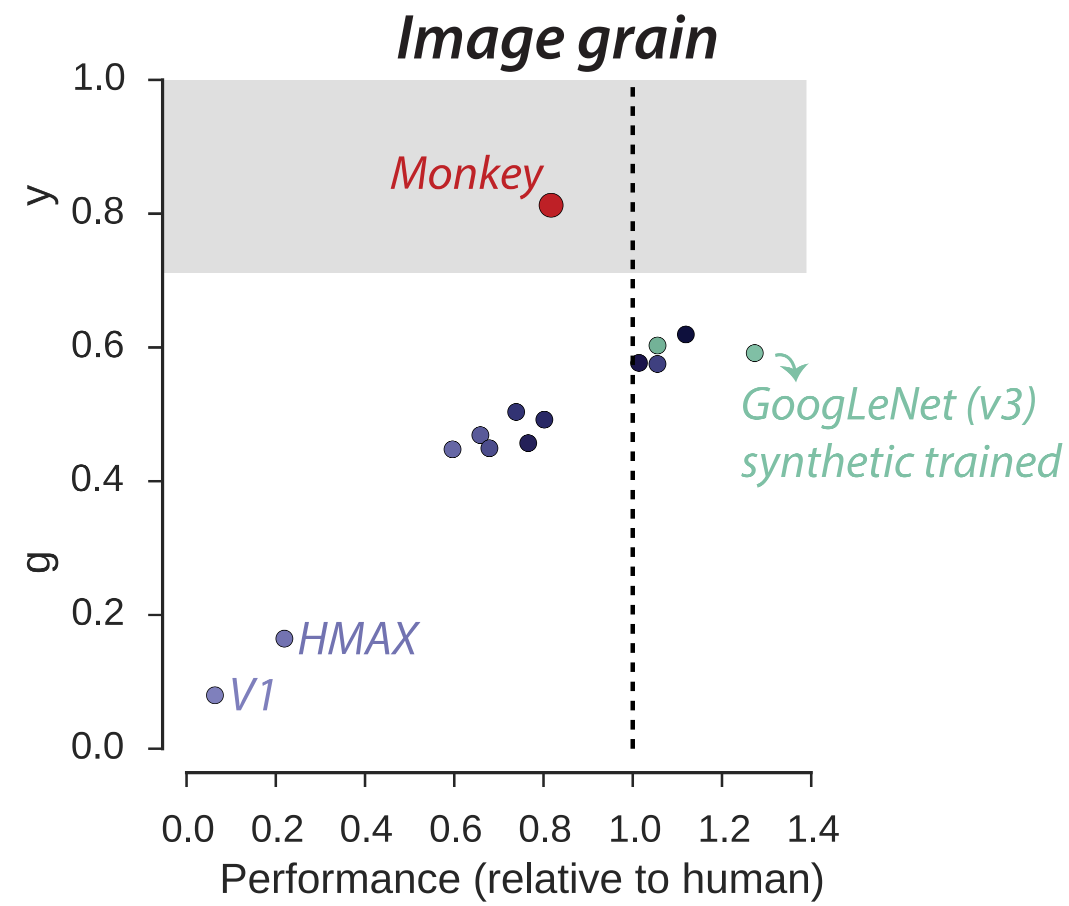
Kohitij
Kar

Kailyn
Schmidt

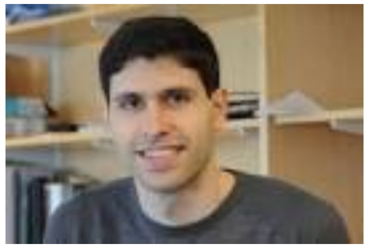
Jim
DiCarlo

Match between models and data at category confusion level is pretty good … but less good at *image* grain:





work of:



Elias
Issa

Rishi
Rajalingham

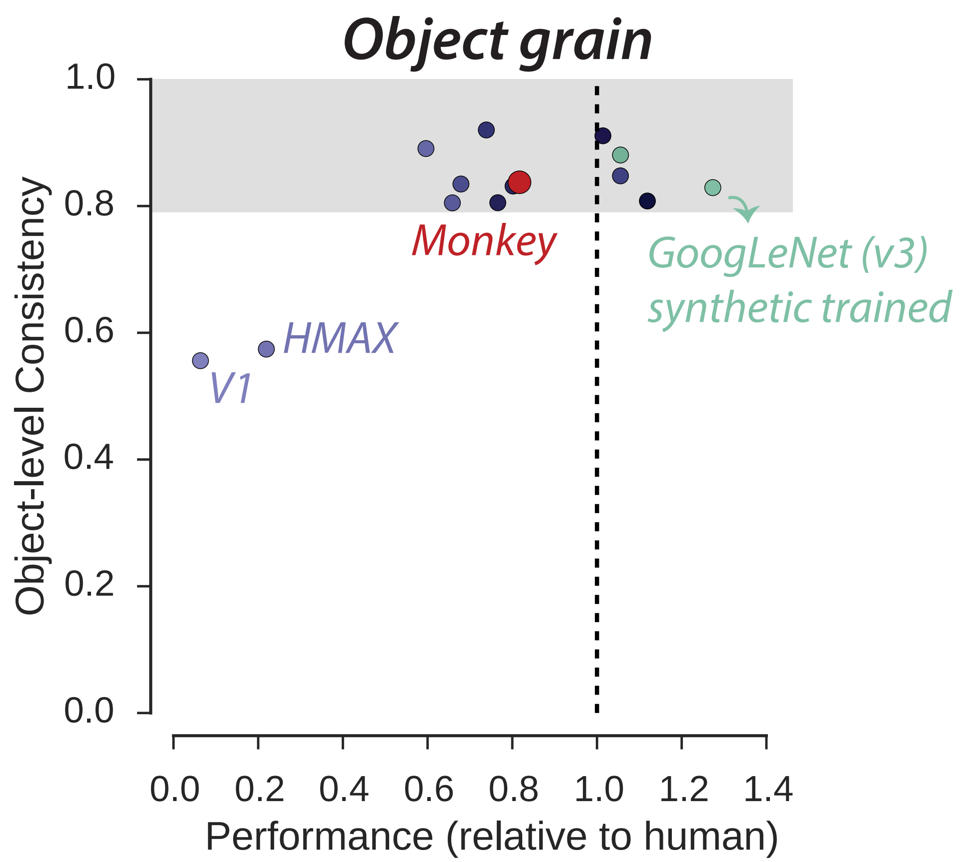
Kohitij
Kar

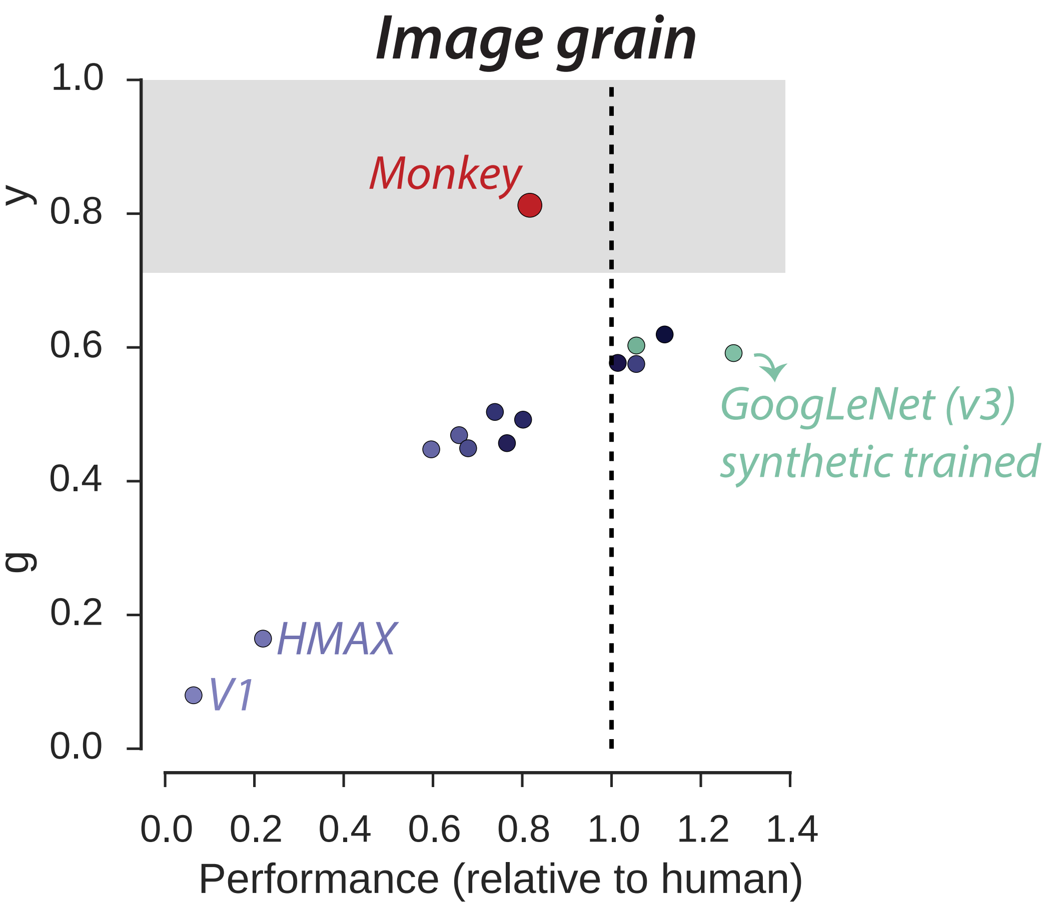
Kailyn
Schmidt

Jim
DiCarlo

Digging deeper into understanding visual cortex

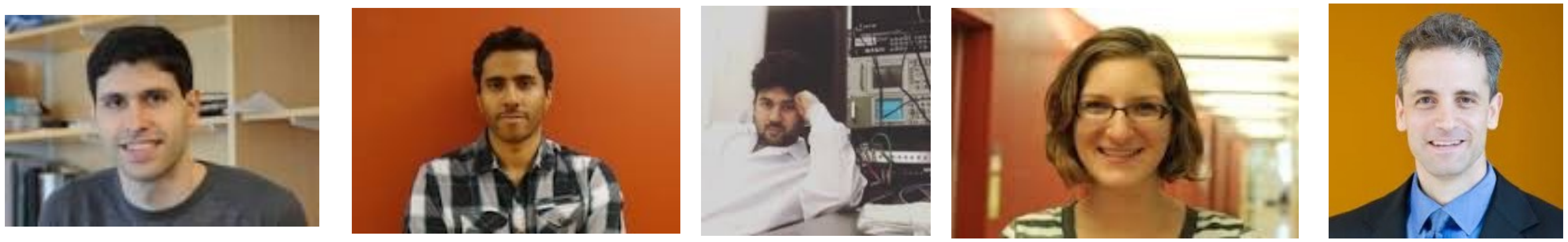
Match between models and data at category confusion level is pretty good … but less good at *image* grain:



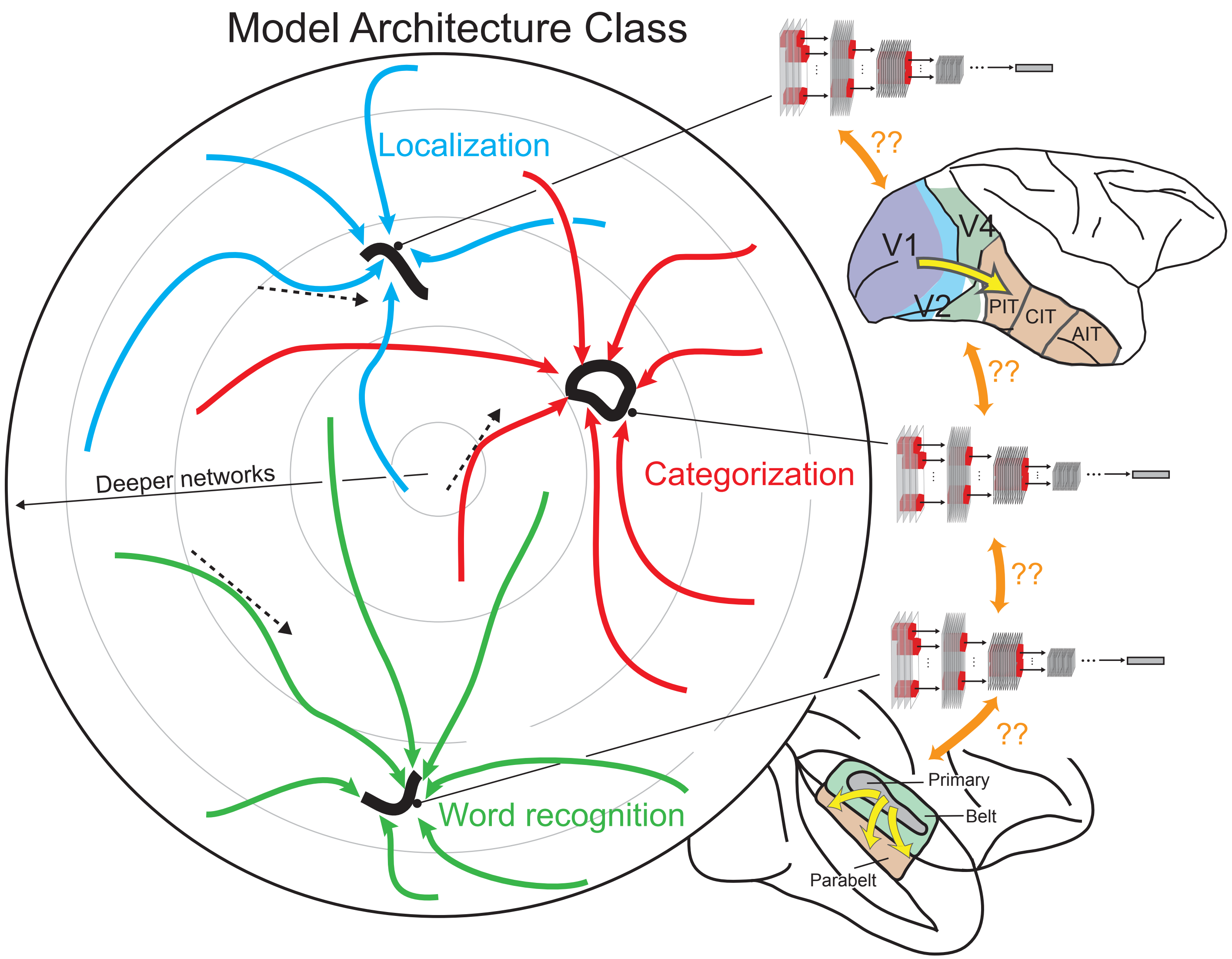


(remember, neural fits only ~50%)…

work of:



Elias Issa | Rishi Rajalingham | Kohitij Kar | Kailyn Schmidt | Jim DiCarlo



3.

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}}[L(p_a^*)]$$

where p^* is result of

$$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

"learning rule"

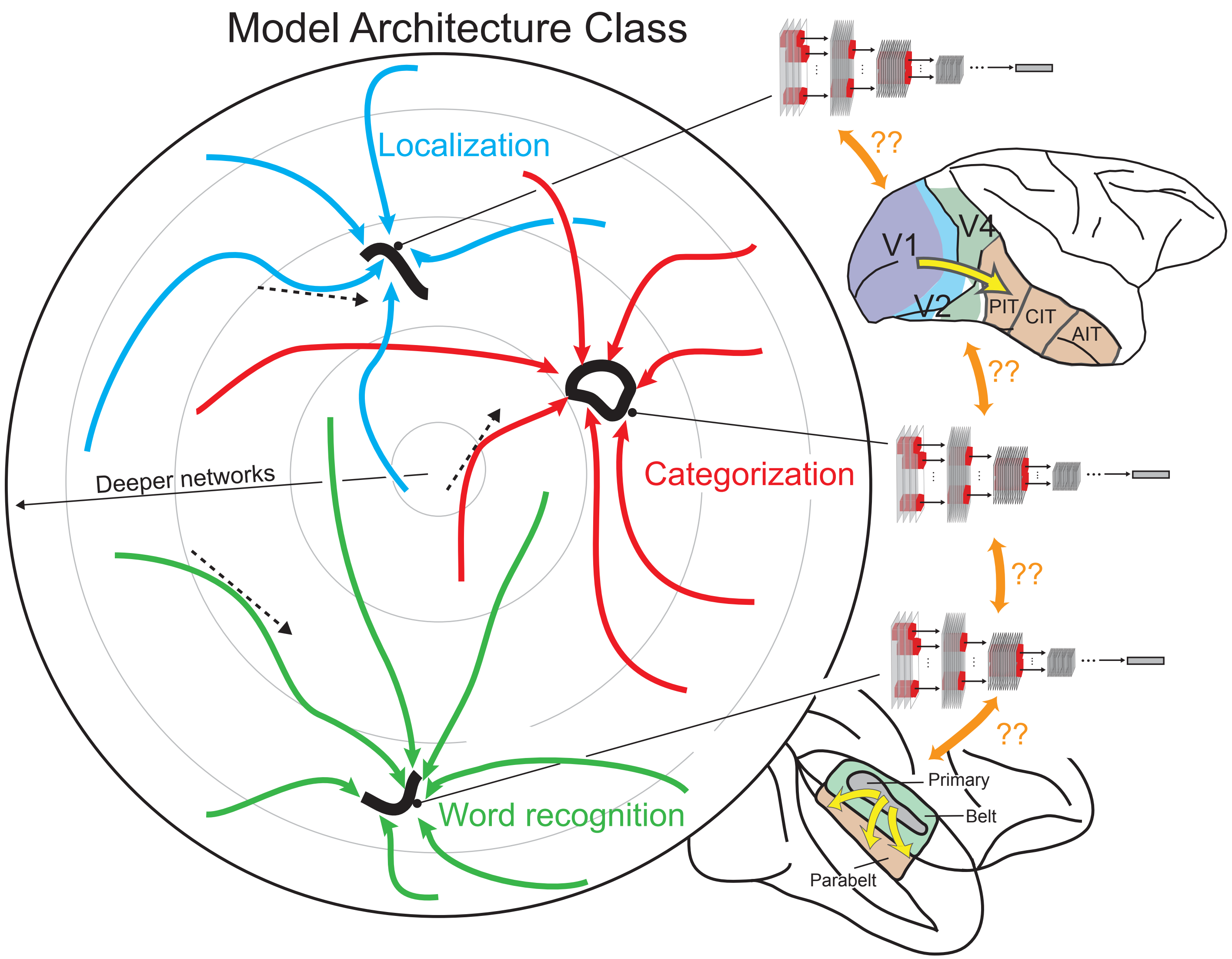
1.

A = *architecture class*

2.

L = *loss function* ***D*** = *dataset*

"task"



3.

$$\underset{a \in \mathcal{A}}{\operatorname{argmin}}[L(p_a^*)]$$

where p^* is result of

$$\frac{dp_a}{dt} = -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

"learning rule"

1.

A = *architecture class*

2.

L = *loss function* ***D*** = *dataset*

"task"

Digging deeper into understanding visual cortex

Three hypotheses:

1) the task (loss function **L** or dataset **D**) is wrong

2) the architecture class (**A**) is wrong

~~3) the learning rule (argmin, SGD, &c) rule is wrong~~

bet: some version of approximate backprop-like error correction is reasonable

Better tasks (loss functions)

Optimize models of the current structure to directly match the neural data …

Better tasks (loss functions)

Optimize models of the current structure to directly match the neural data …

no

yes

Model class is wrong

Task is wrong

Better tasks (loss functions)

Optimize models of the current structure to directly match the neural data …

no

yes

Model class is wrong

Task is wrong

… but not enough neural data?

Better tasks (loss functions)

Optimize models of the current structure to directly match the behavioral data … then check against neural data.

no

yes

Model class is wrong

Task is wrong

Better tasks (loss functions)

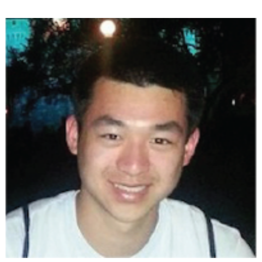
Optimize models of the current structure to directly match the behavioral data … then check against neural data.

no

yes

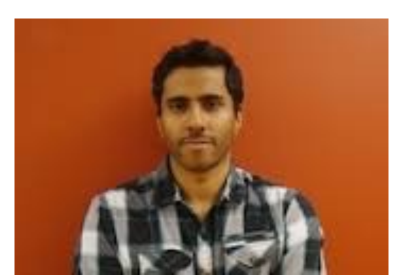
Model class is wrong

Task is wrong

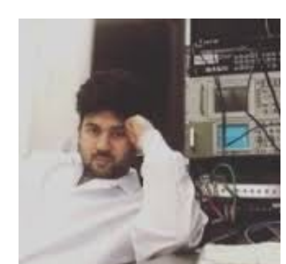


Eli Wang

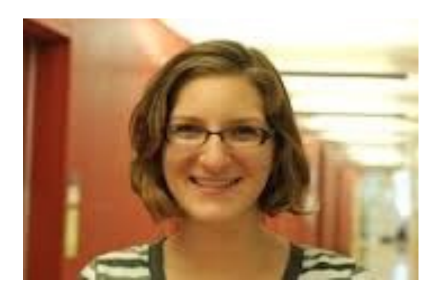
Elias Issa



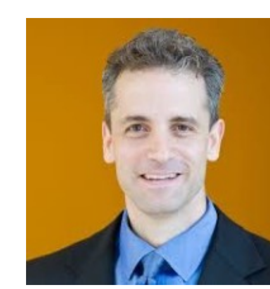
Rishi Rajalingham



Kohitij Kar



Kailyn Schmidt

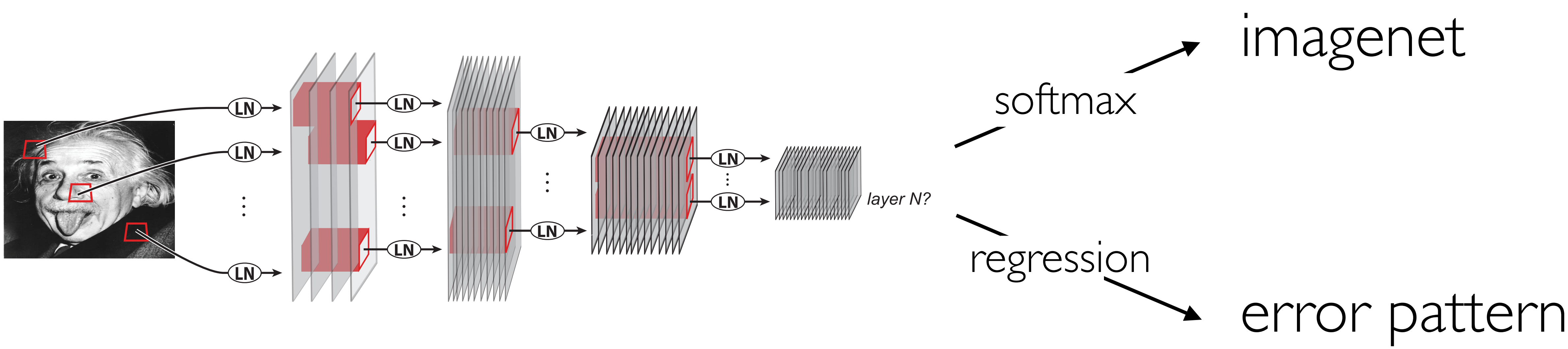


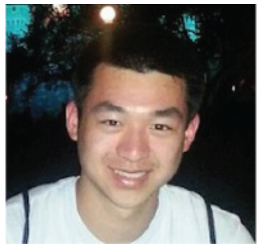
Jim DiCarlo

Better tasks (loss functions)

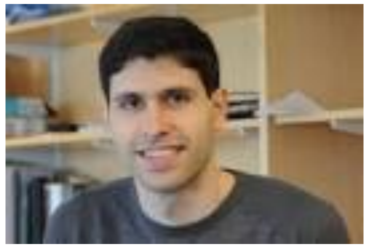
Optimize models of the current structure to directly match the behavioral data … then check against neural data.

(i) predict vector of errors





Eli
Wang



Elias
Issa

Rishi
Rajalingham

Kohitij
Kar

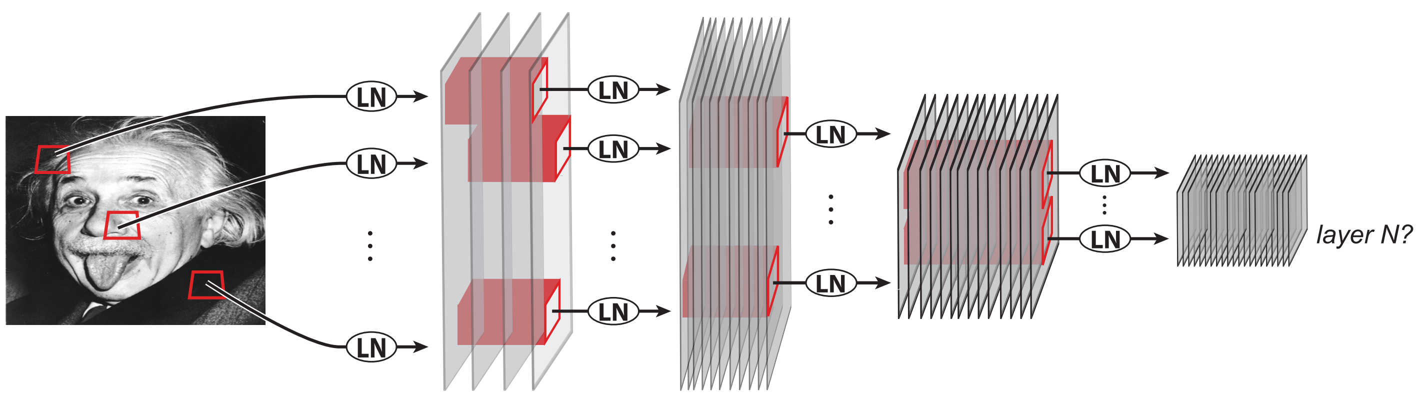
Kailyn
Schmidt

Jim
DiCarlo

Better tasks (loss functions)

Optimize models of the current structure to directly match the behavioral data … then check against neural data.

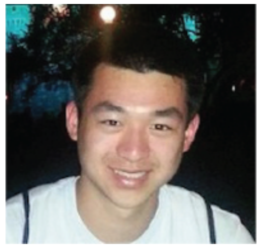
(ii) as actual error pattern



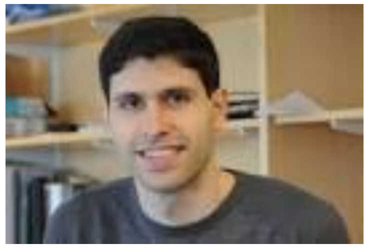
softmax

imagenet

error pattern



Eli
Wang



Elias
Issa

Rishi
Rajalingham

Kohitij
Kar

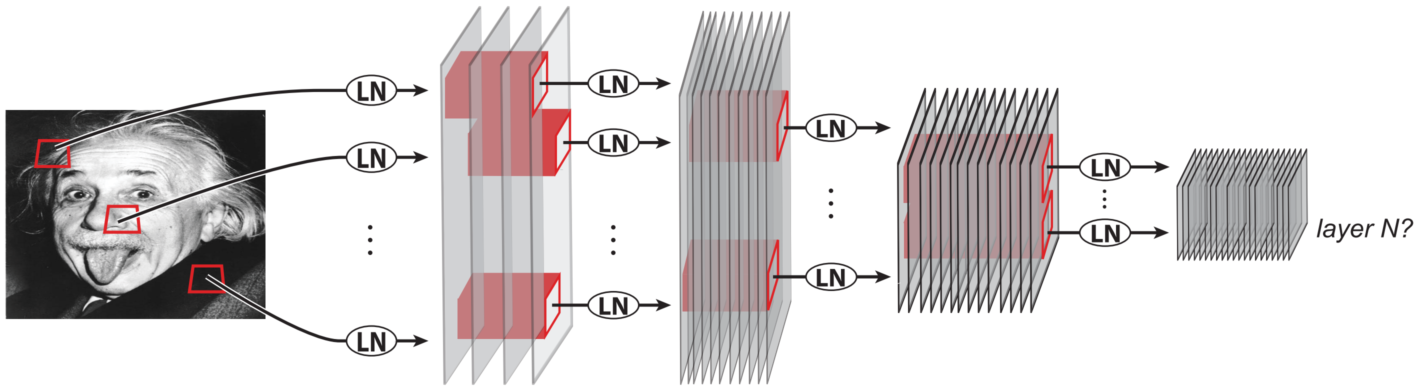
Kailyn
Schmidt

Jim
DiCarlo

Better tasks (loss functions)

Optimize models of the current structure to directly match the behavioral data … then check against neural data.

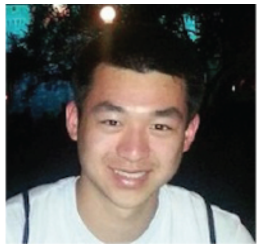
(iii) as multiplier — indicator of niche?



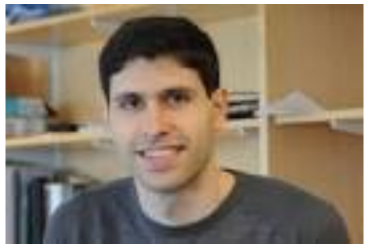
softmax

imagenet

error-inverse weighting



Eli
Wang



Elias
Issa

Rishi
Rajalingham

Kohitij
Kar

Kailyn
Schmidt

Jim
DiCarlo

Better tasks (loss functions)

less normative "task" *more normative "task"*

Fit neural data

Better tasks (loss functions)

less normative "task" *more normative "task"*

Fit neural data

Fit categorization error pattern
…
check against neural data

Better tasks (loss functions)

less normative "task" ⟷ *more normative "task"*

Fit neural data	Fit categorization error pattern … check against neural data	Solve non-categorization tasks … check against neural data

Better tasks (loss functions)

less normative "task" ⟷ *more normative "task"*

Fit neural data

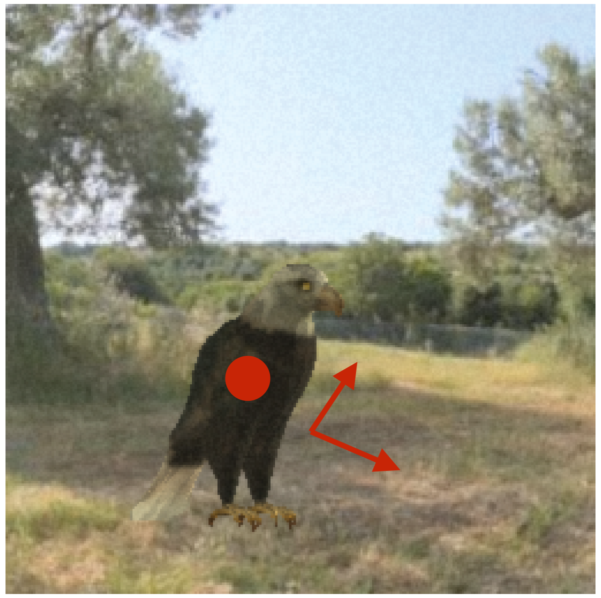
Fit categorization error pattern
…
check against neural data

Solve non-categorization tasks
…
check against neural data

But which non-categorical tasks?

Better tasks (loss functions)

Pose / position estimation

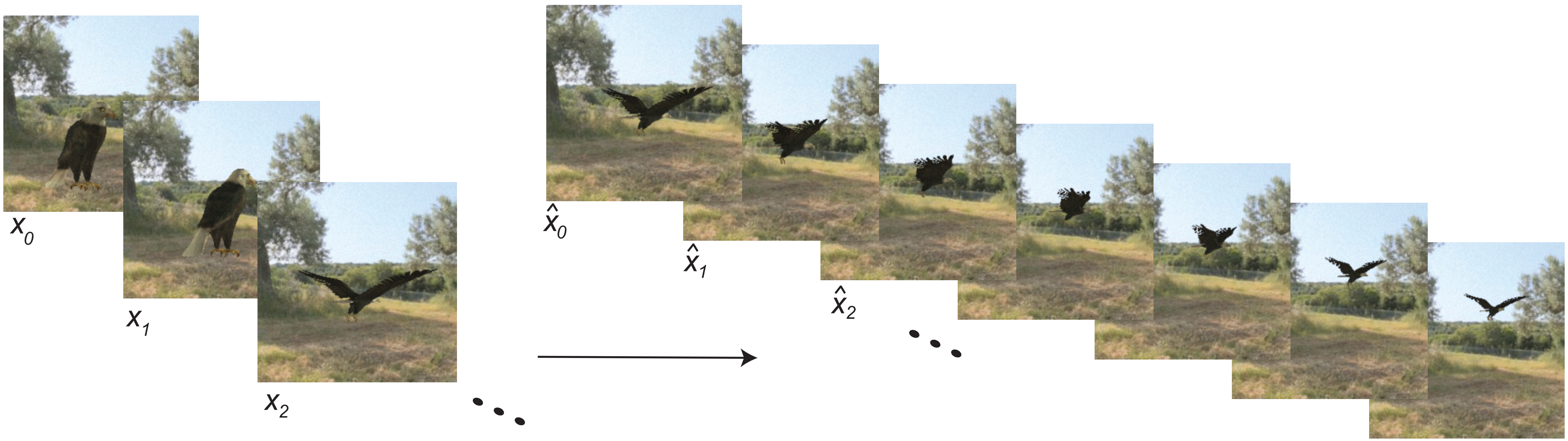


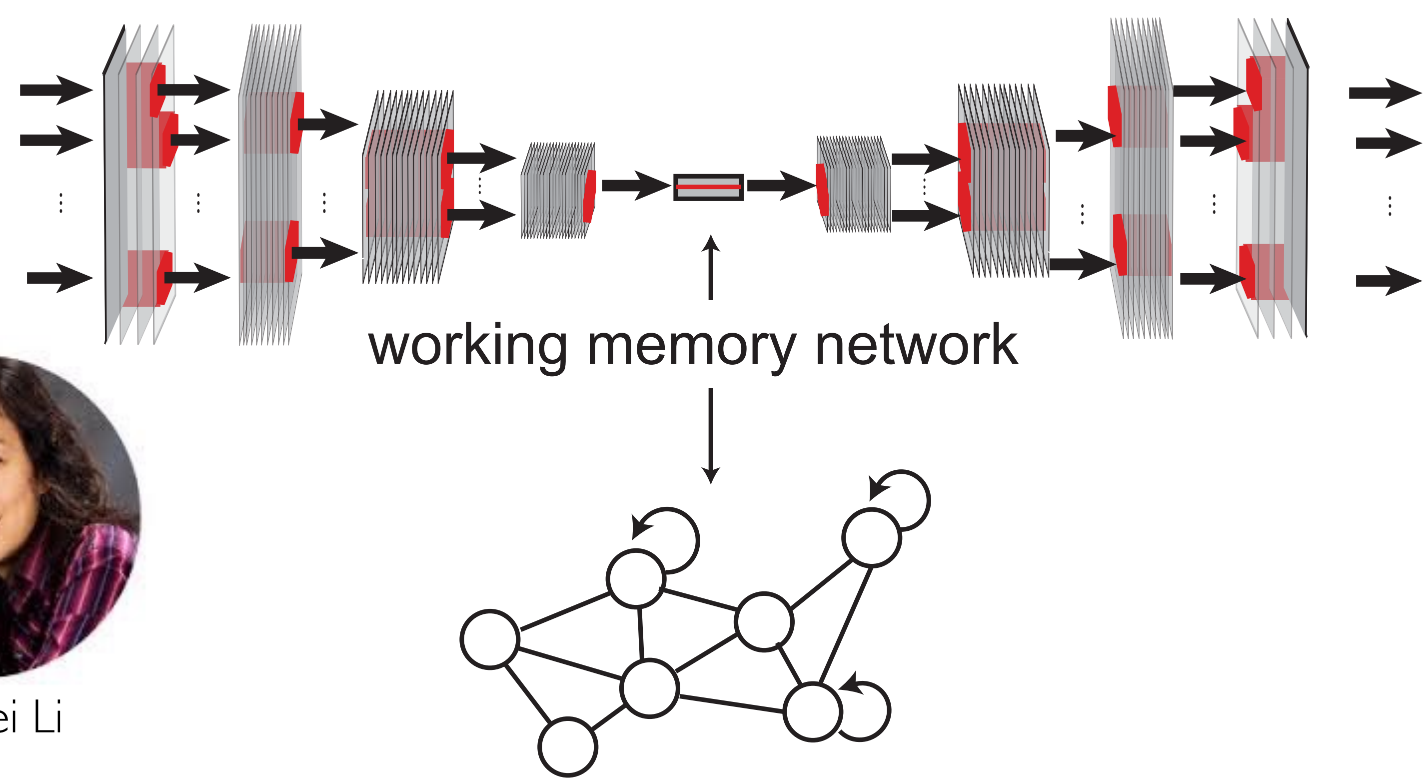
Normal/Depth estimation

Segmentation

Chengxu Zhuang

Future prediction under agent-controlled actions





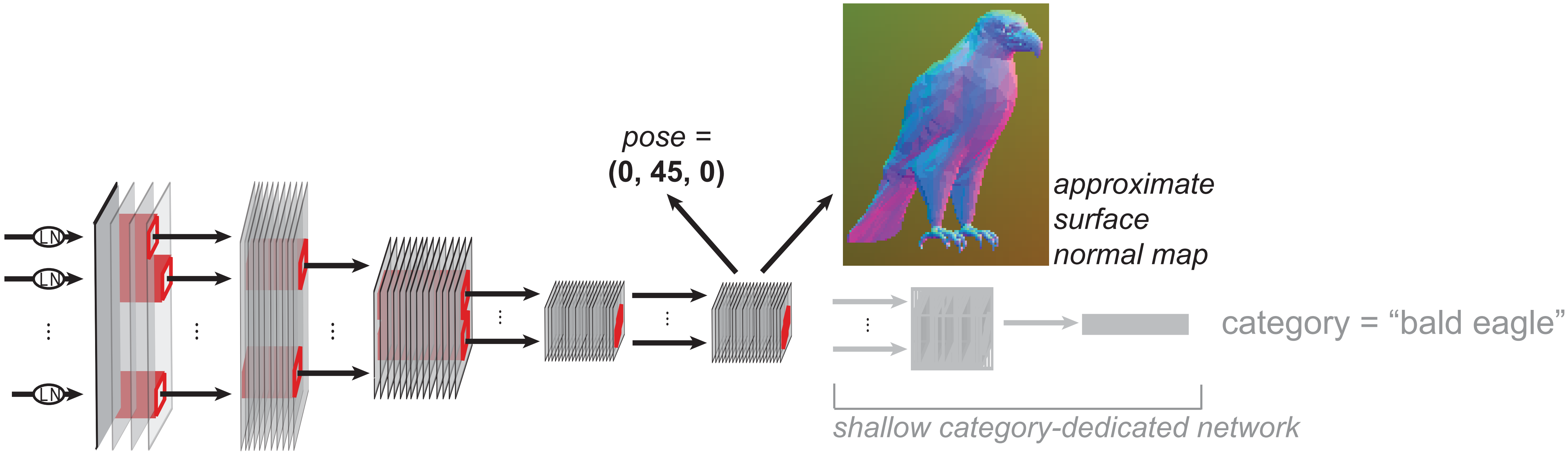

Nick Haber


Damian Mrowca


Fei-Fei Li

Better tasks (loss functions)

Where should the tasks be imposed? (intermediate?)



Strategy: optimize over architectures for solving joint tasks, compare to neural data

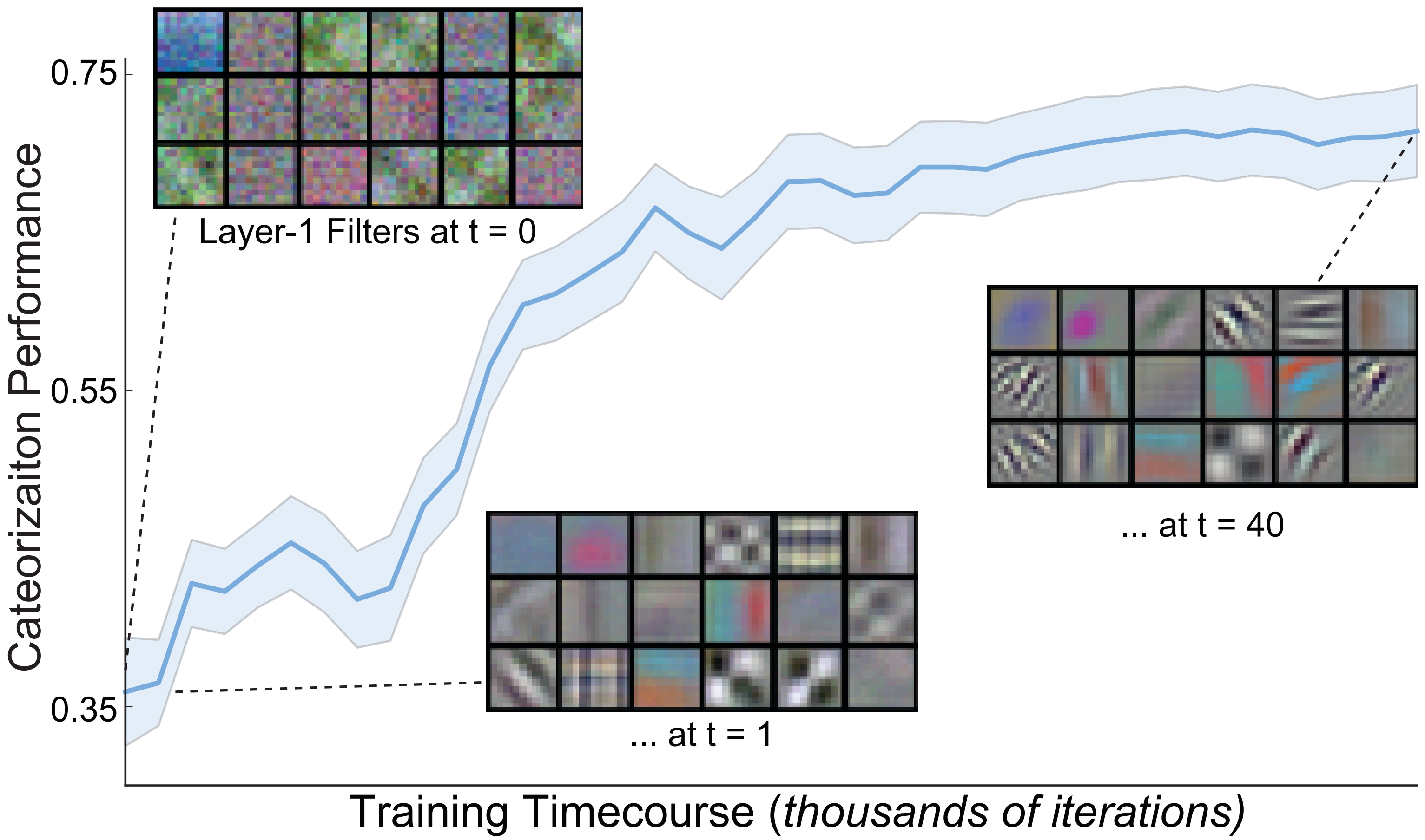
How much work can less heavily supervised tasks do?

Comparing to Neural Data

- Various second-order metrics: encoding regression, RSAs, etc

- Behavioral consistency — pattern of errors at various grains of detail

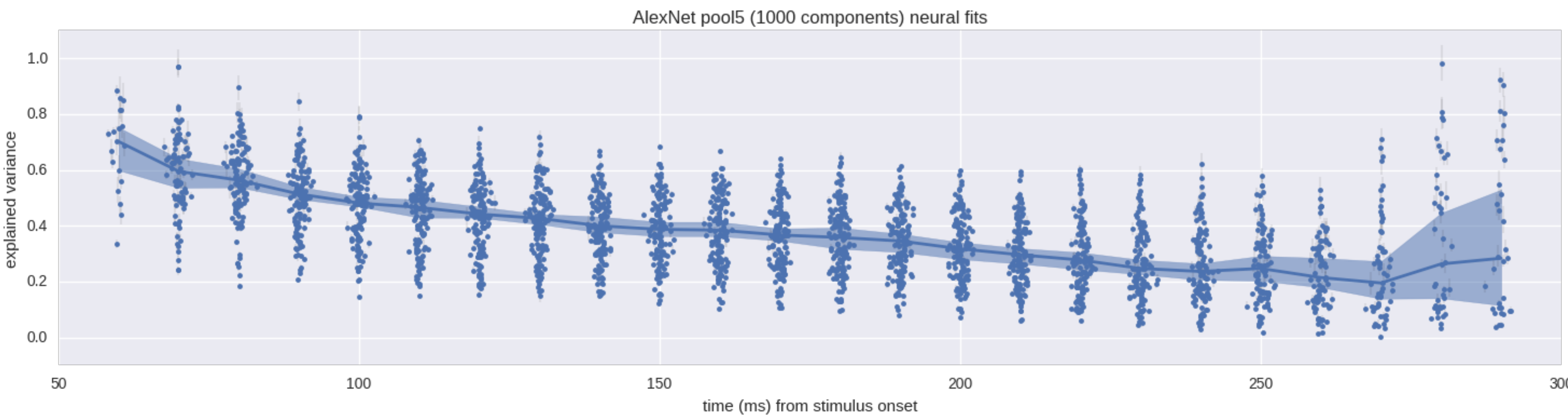
- But really, there is a developmental hypothesis implicit in these models. Time course of all metrics should be matched:



Use developmental data separate more biologically correct loss functions from less correct ones?

Better architectures

Better architectures



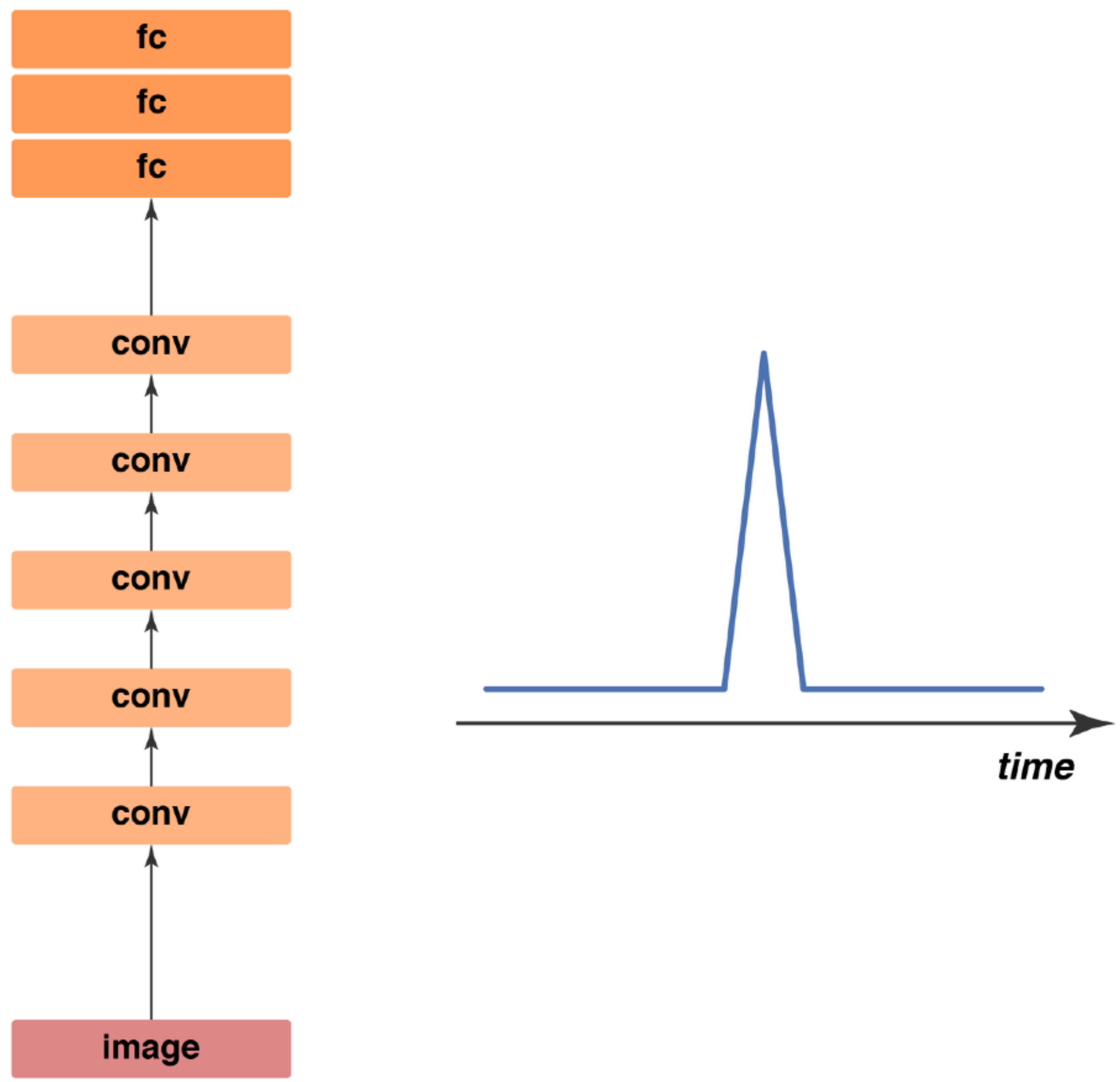
work of

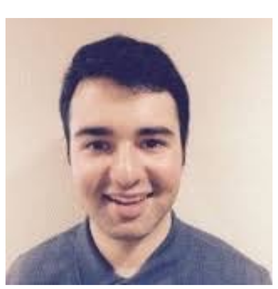
Kohitij
Kar

Jonas
Kubilius

Jim
DiCarlo

Better architectures

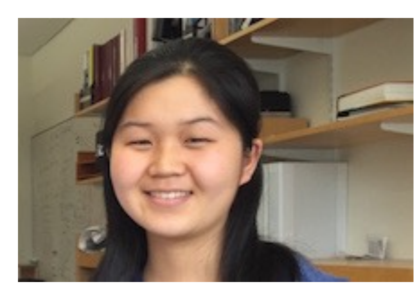




Aran Nayebi

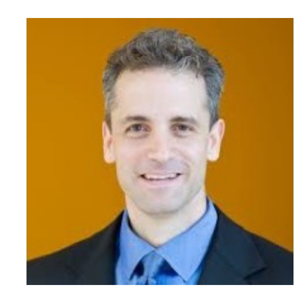
Surya Ganguli

Jonas Kubilius



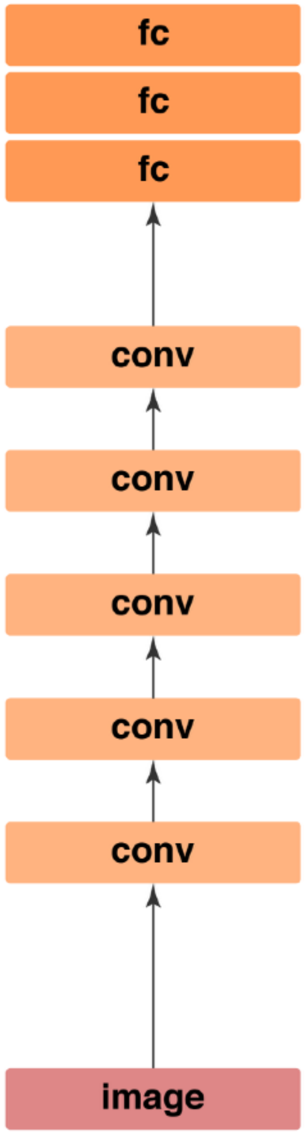
Maryann Rui

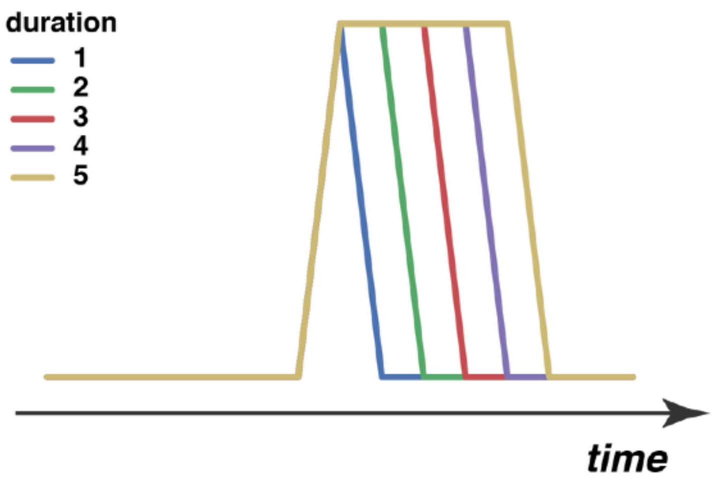
Kohitij Kar



Jim DiCarlo

Better architectures

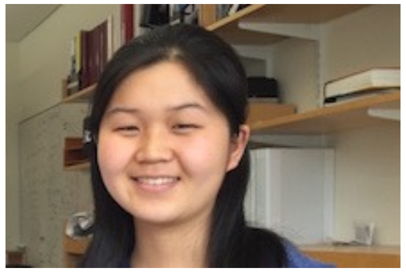




Aran Nayebi

Surya Ganguli

Jonas Kubilius

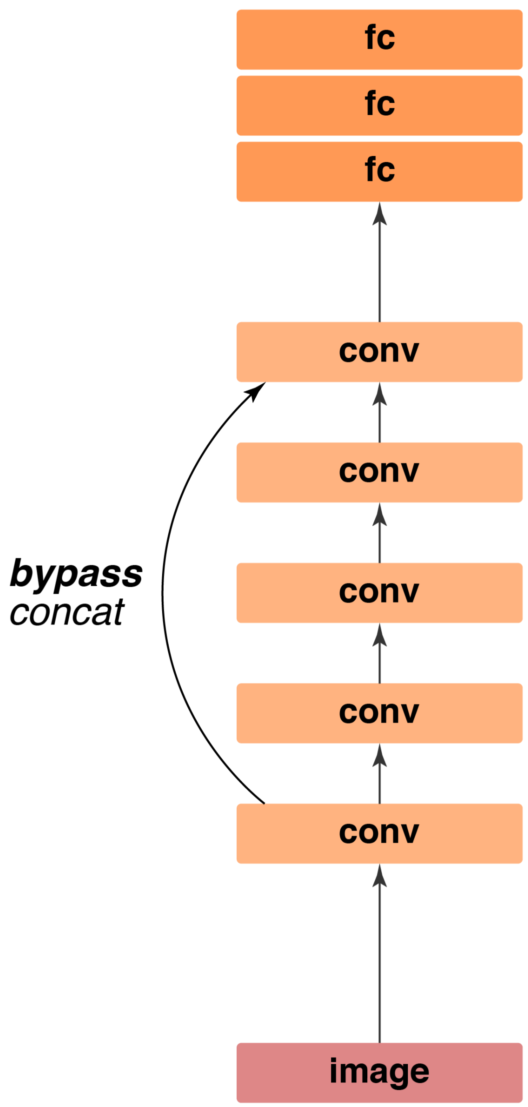


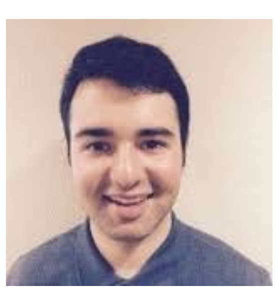
Maryann Rui

Kohitij Kar

Jim DiCarlo

Better architectures

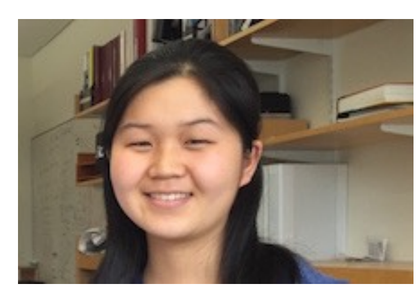




Aran Nayebi

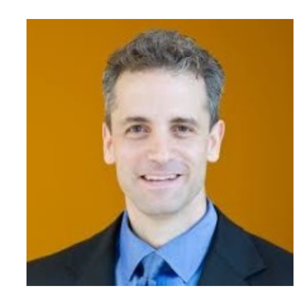
Surya Ganguli

Jonas Kubilius



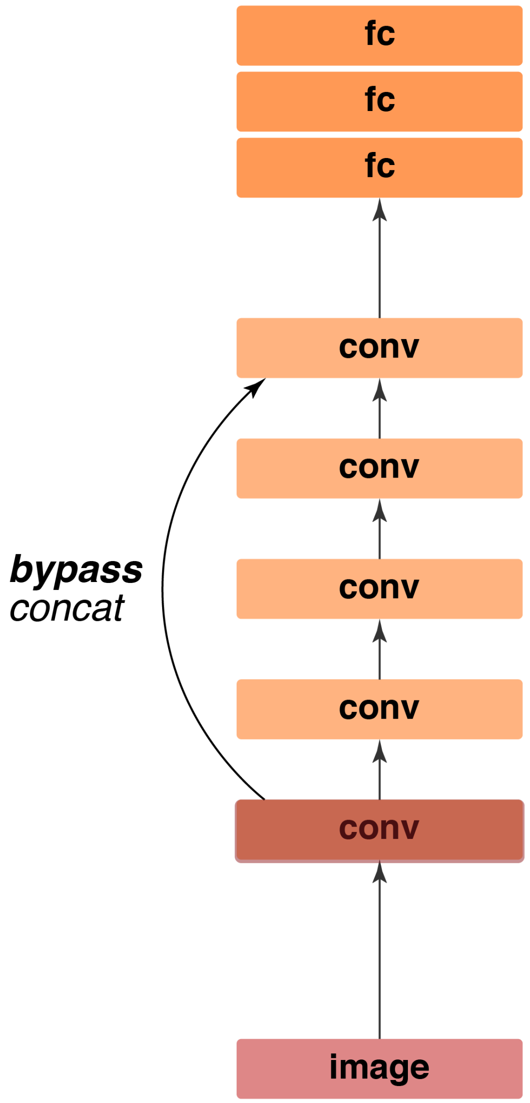
Maryann Rui

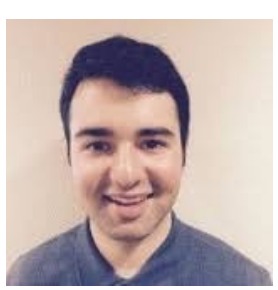
Kohitij Kar



Jim DiCarlo

Better architectures

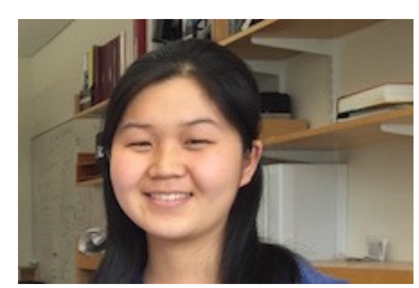




Aran Nayebi

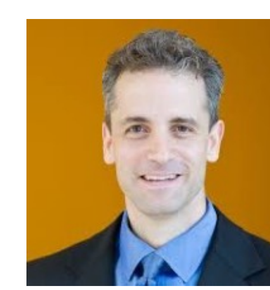
Surya Ganguli

Jonas Kubilius



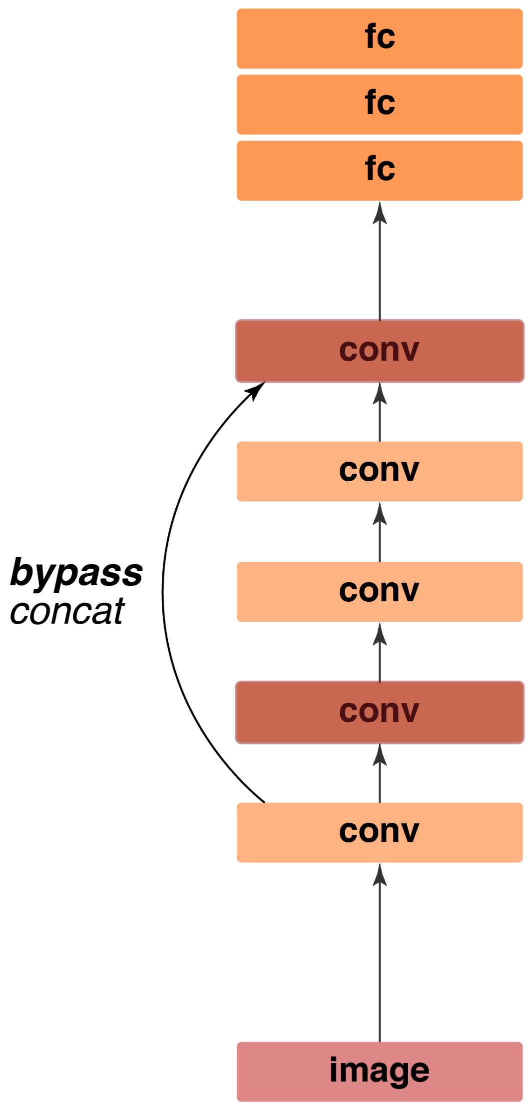
Maryann Rui

Kohitij Kar



Jim DiCarlo

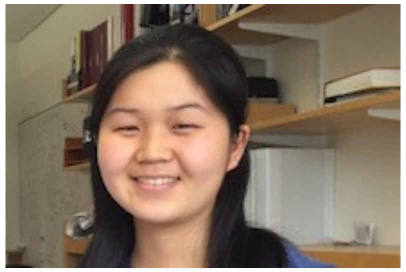
Better architectures



Aran Nayebi

Surya Ganguli

Jonas Kubilius

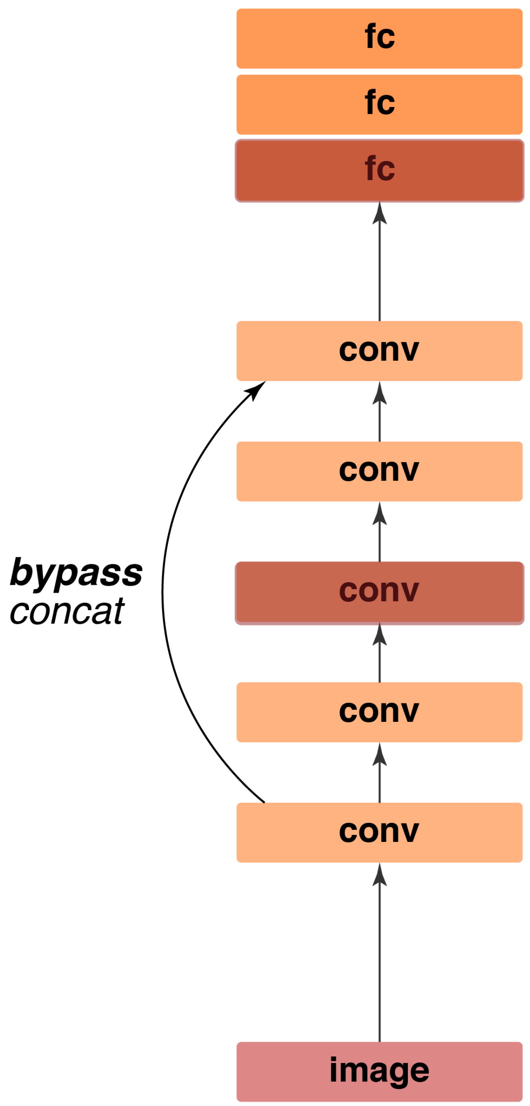


Maryann Rui

Kohitij Kar

Jim DiCarlo

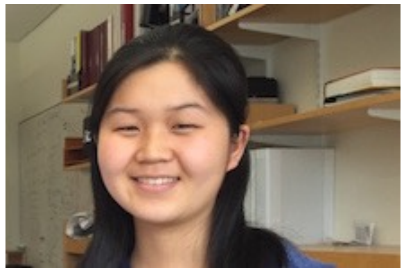
Better architectures



Aran Nayebi

Surya Ganguli

Jonas Kubilius

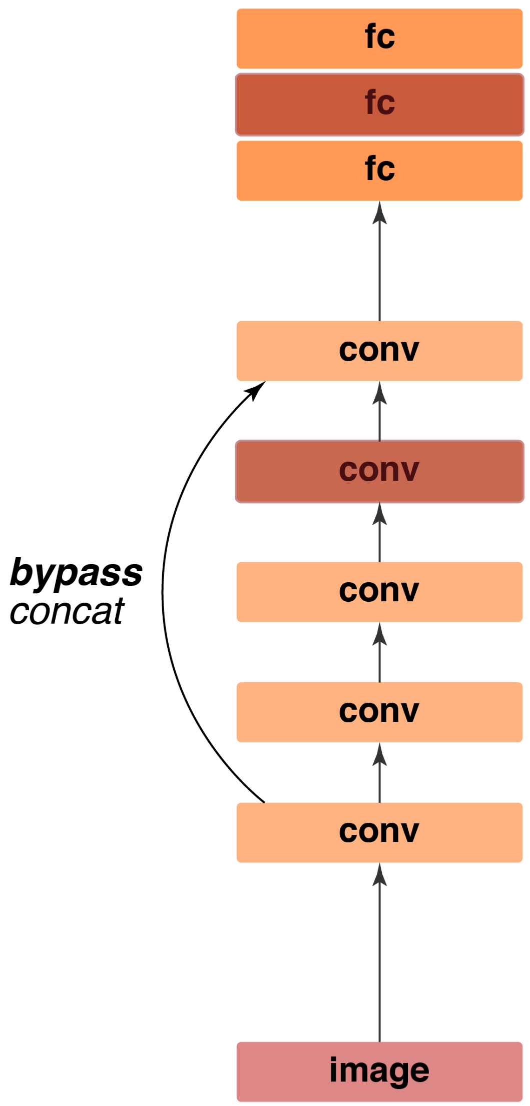


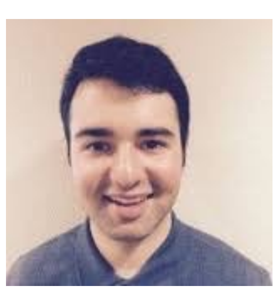
Maryann Rui

Kohitij Kar

Jim DiCarlo

Better architectures

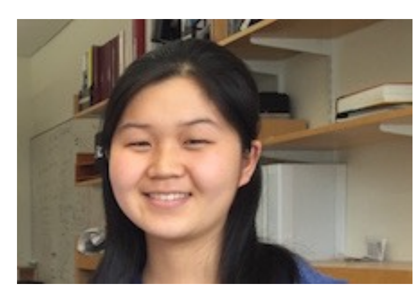




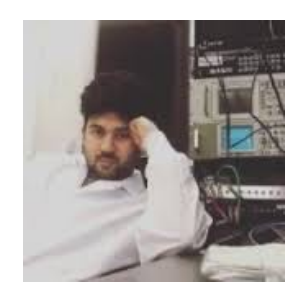
Aran Nayebi

Surya Ganguli

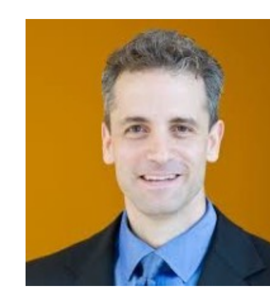
Jonas Kubilius



Maryann Rui

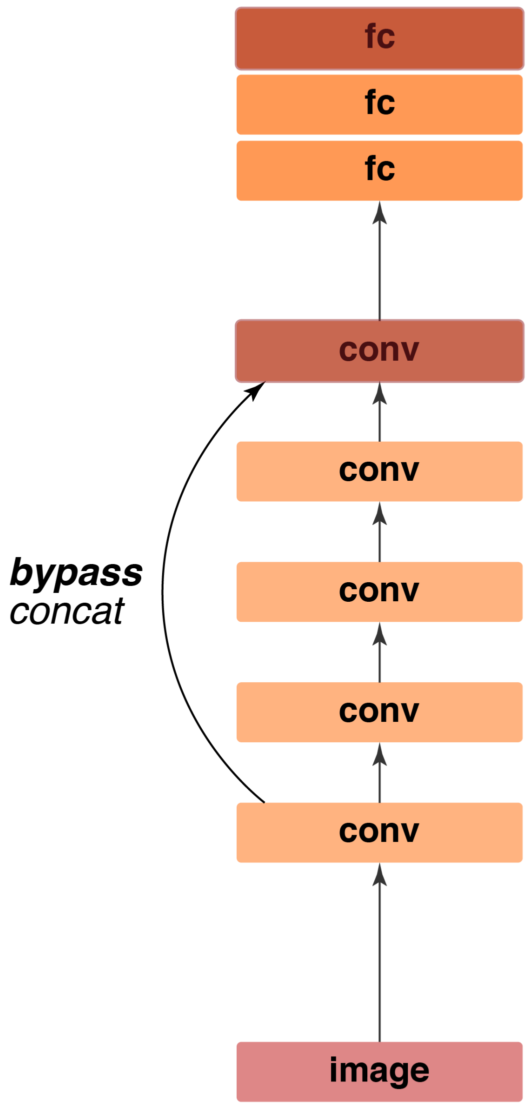


Kohitij Kar



Jim DiCarlo

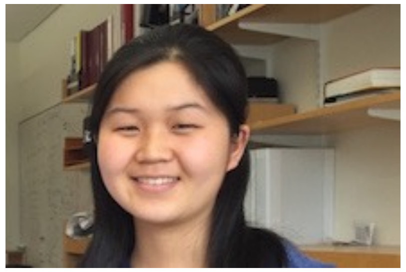
Better architectures



Aran Nayebi

Surya Ganguli

Jonas Kubilius

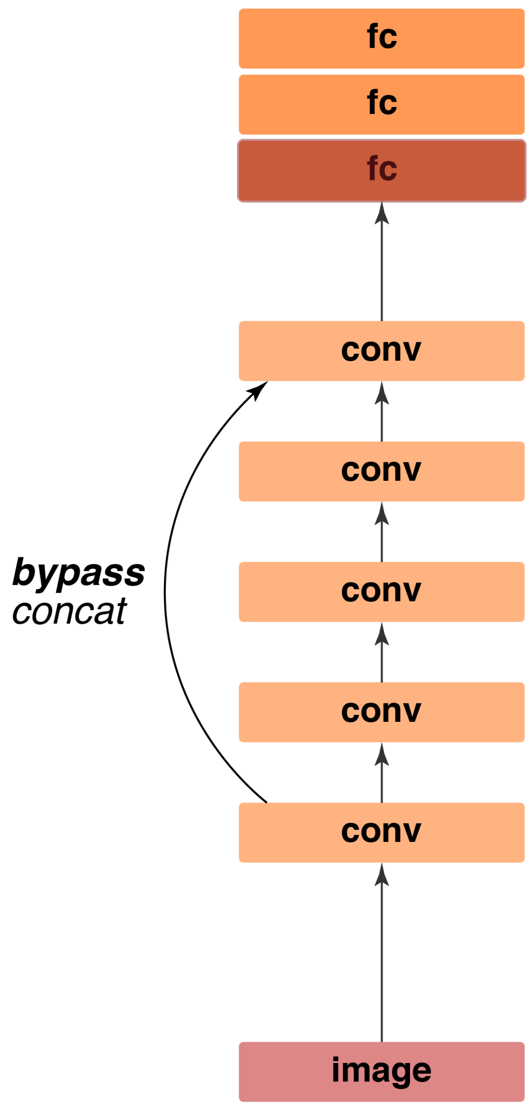


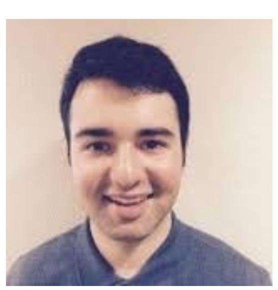
Maryann Rui

Kohitij Kar

Jim DiCarlo

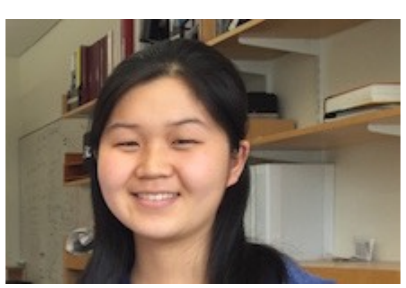
Better architectures



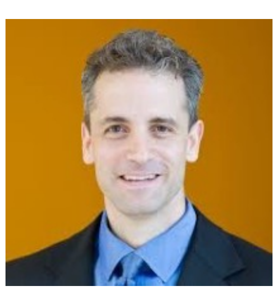

Aran Nayebi


Surya Ganguli

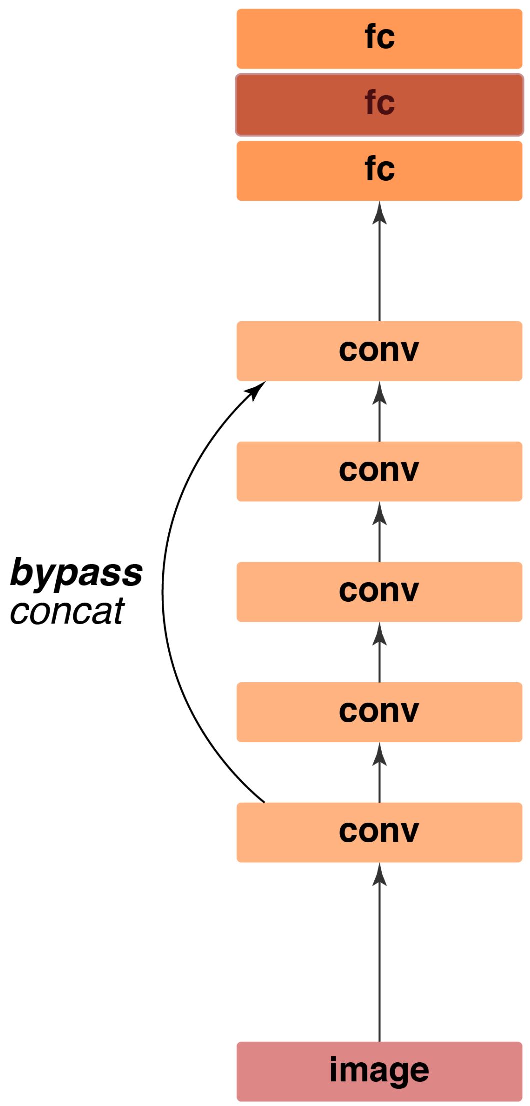

Jonas Kubilius

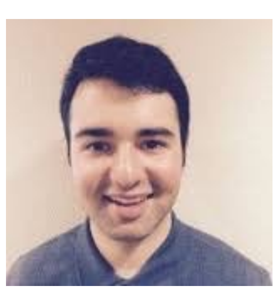

Maryann Rui


Kohitij Kar


Jim DiCarlo

Better architectures

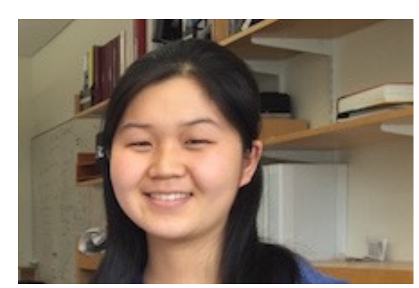




Aran Nayebi

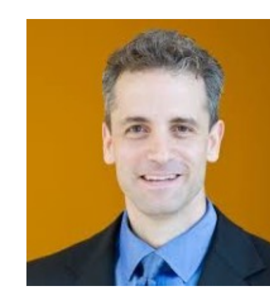
Surya Ganguli

Jonas Kubilius



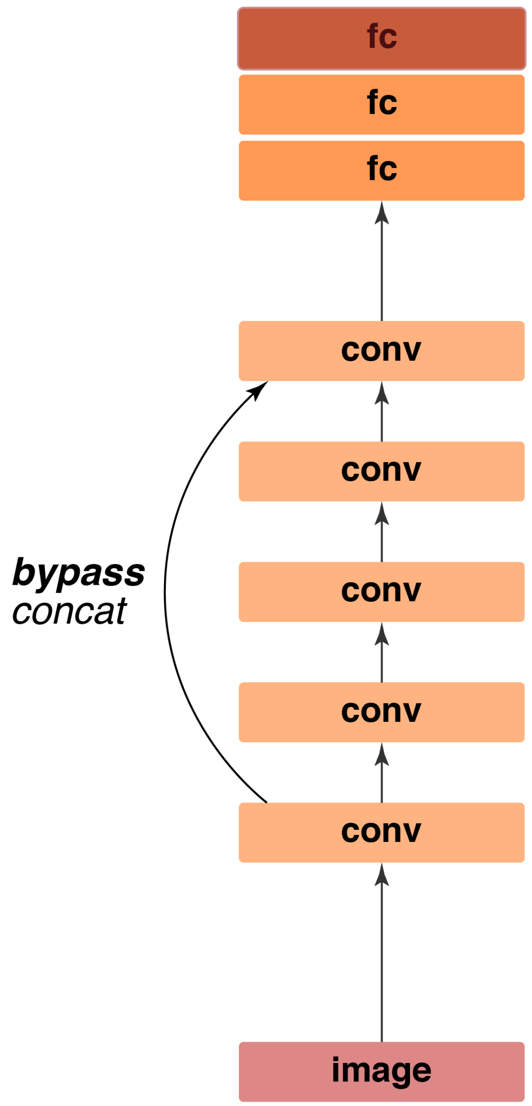
Maryann Rui

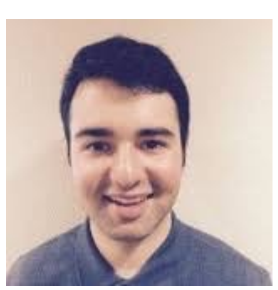
Kohitij Kar



Jim DiCarlo

Better architectures

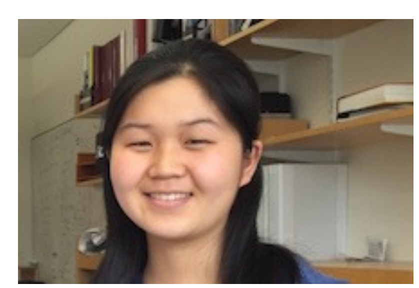




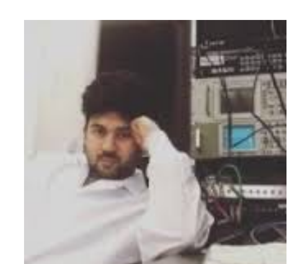
Aran Nayebi

Surya Ganguli

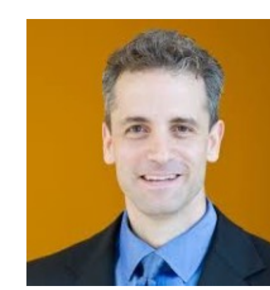
Jonas Kubilius



Maryann Rui

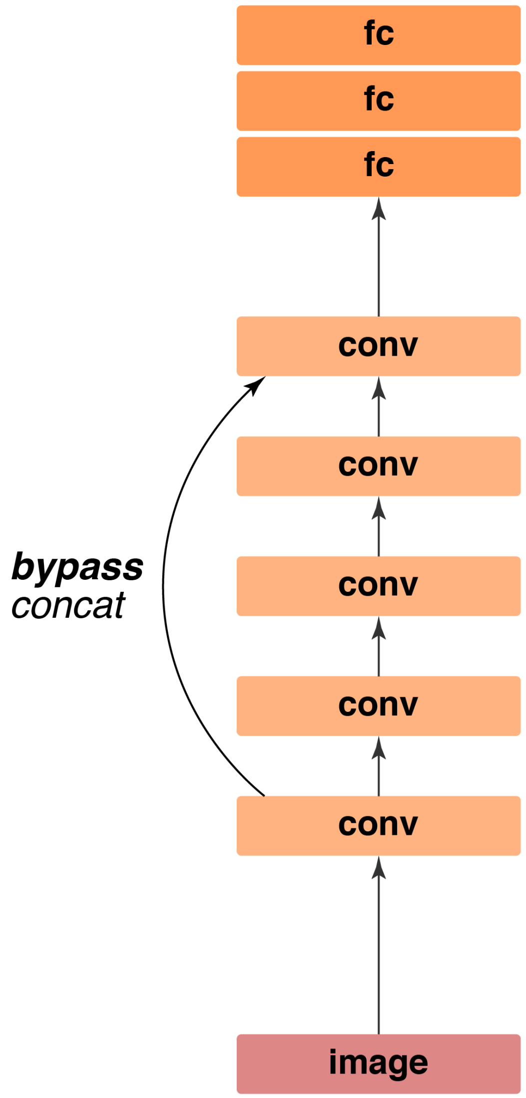


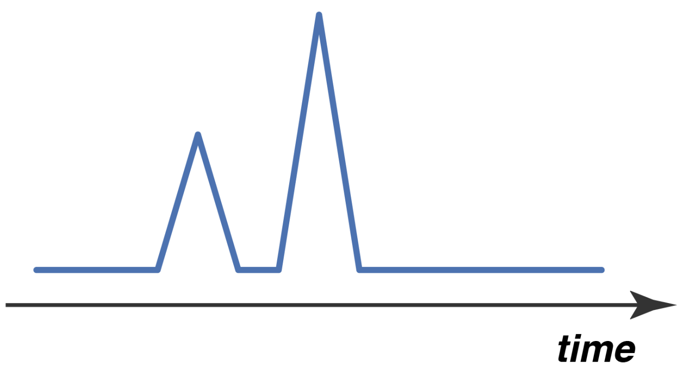
Kohitij Kar



Jim DiCarlo

Better architectures

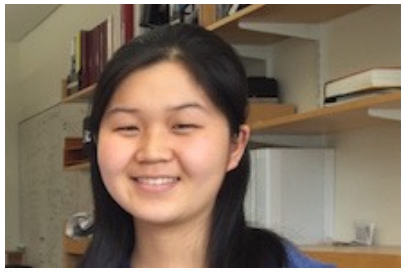




Aran Nayebi

Surya Ganguli

Jonas Kubilius

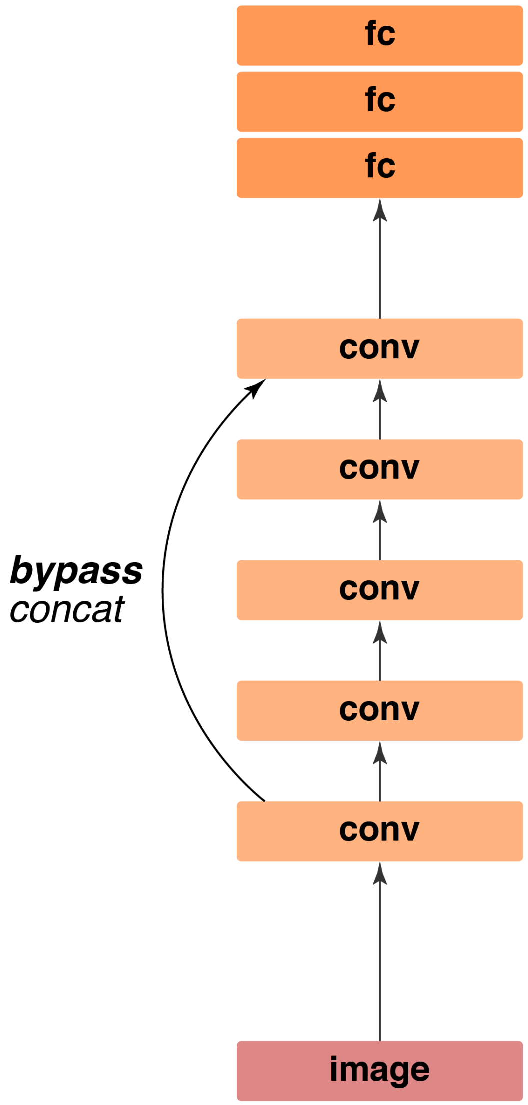


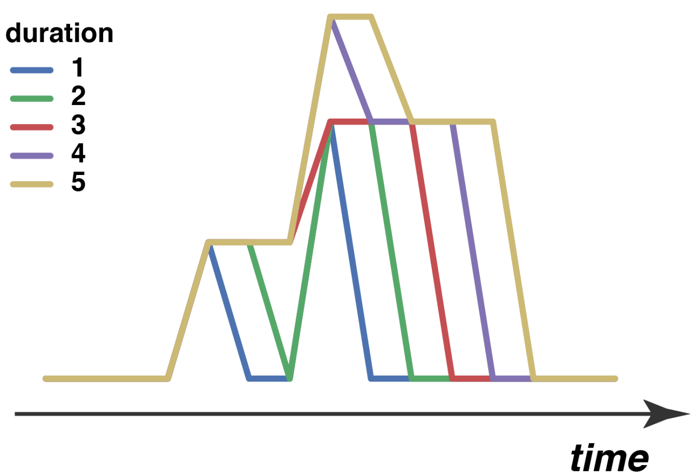
Maryann Rui

Kohitij Kar

Jim DiCarlo

Better architectures

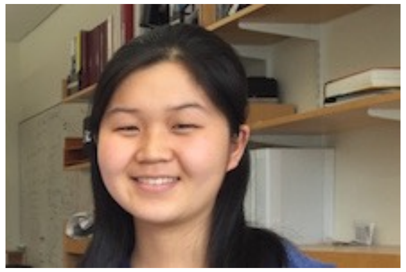




Aran Nayebi

Surya Ganguli

Jonas Kubilius

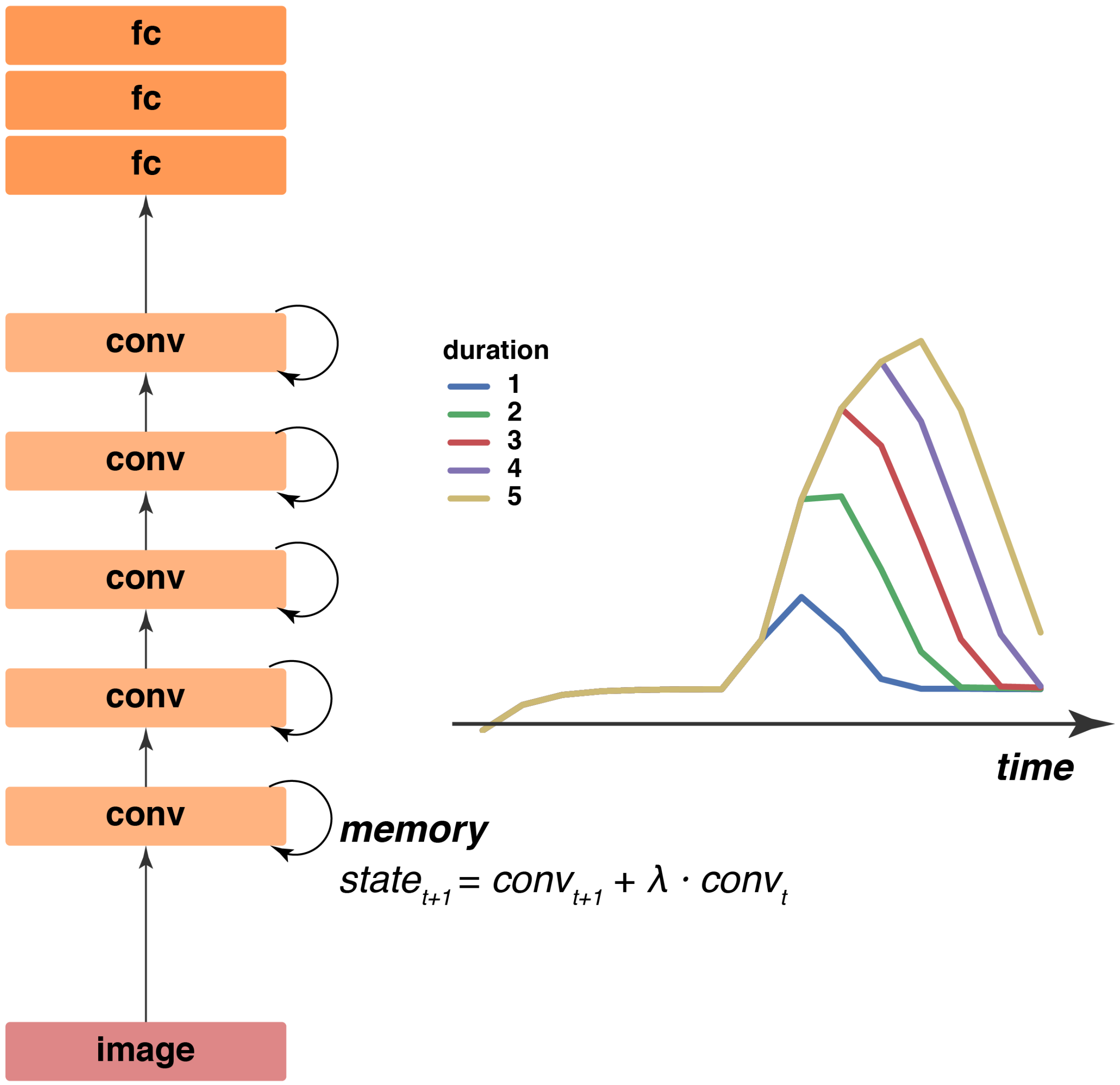


Maryann Rui

Kohitij Kar

Jim DiCarlo

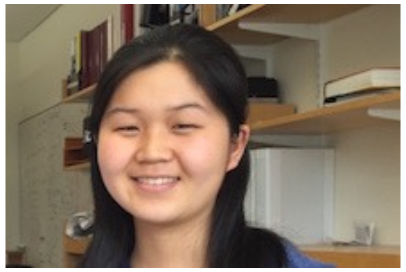
Better architectures




Aran Nayebi


Surya Ganguli

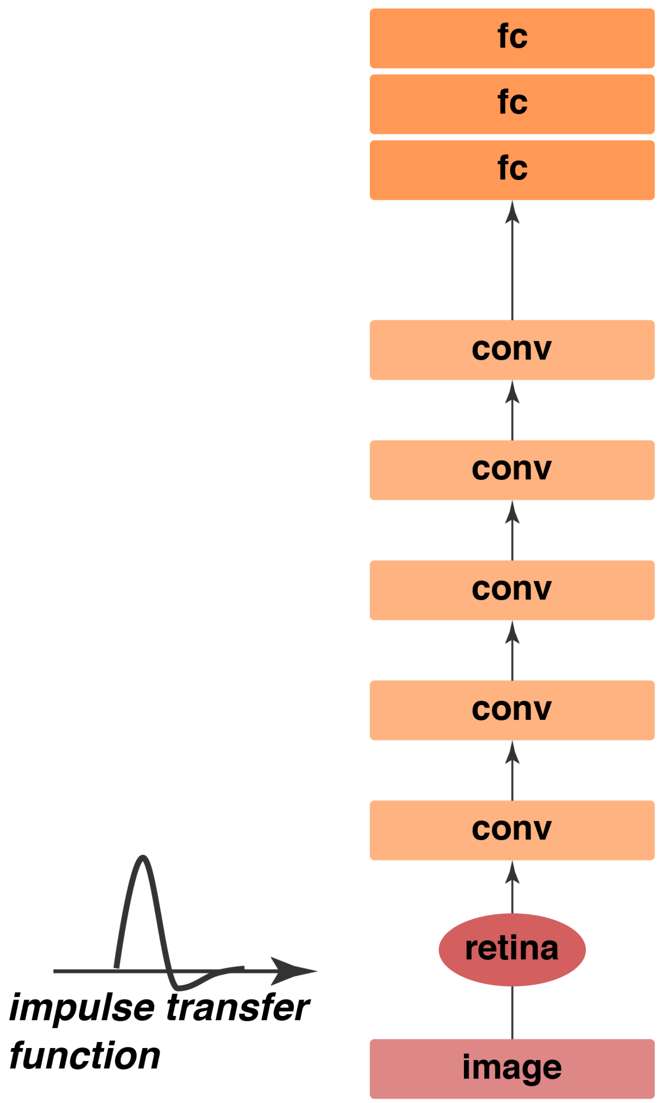

Jonas Kubilius

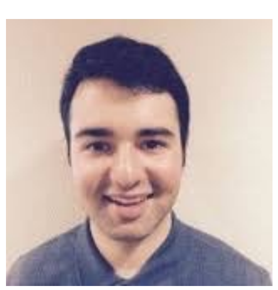

Maryann Rui


Kohitij Kar


Jim DiCarlo

Better architectures

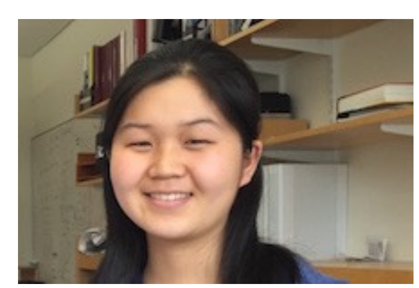




Aran Nayebi

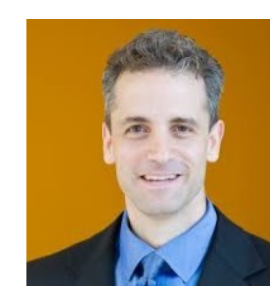
Surya Ganguli

Jonas Kubilius



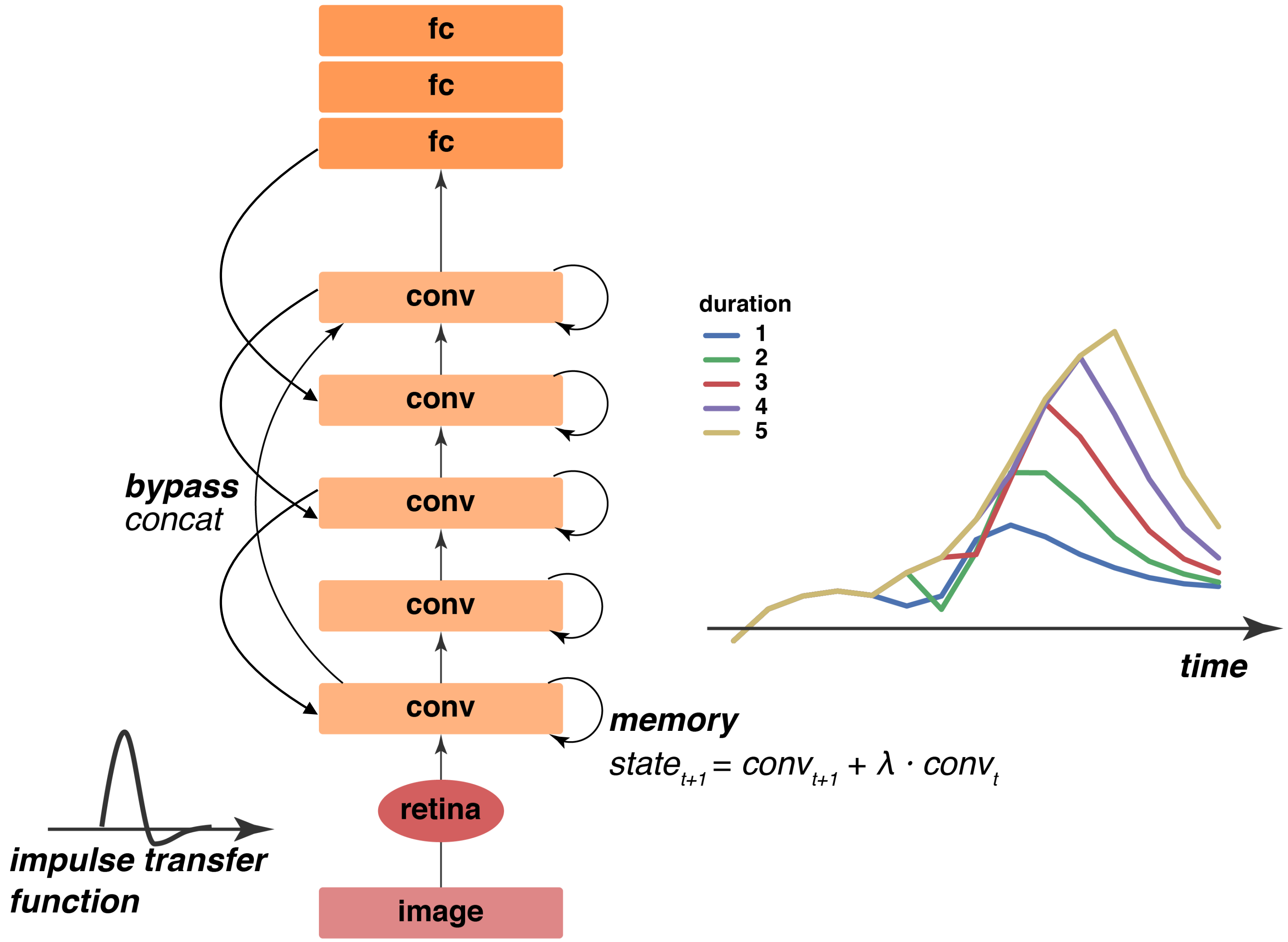
Maryann Rui

Kohitij Kar



Jim DiCarlo

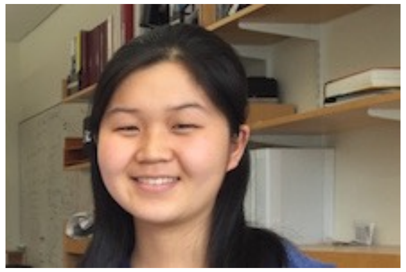
Better architectures



Aran Nayebi

Surya Ganguli

Jonas Kubilius

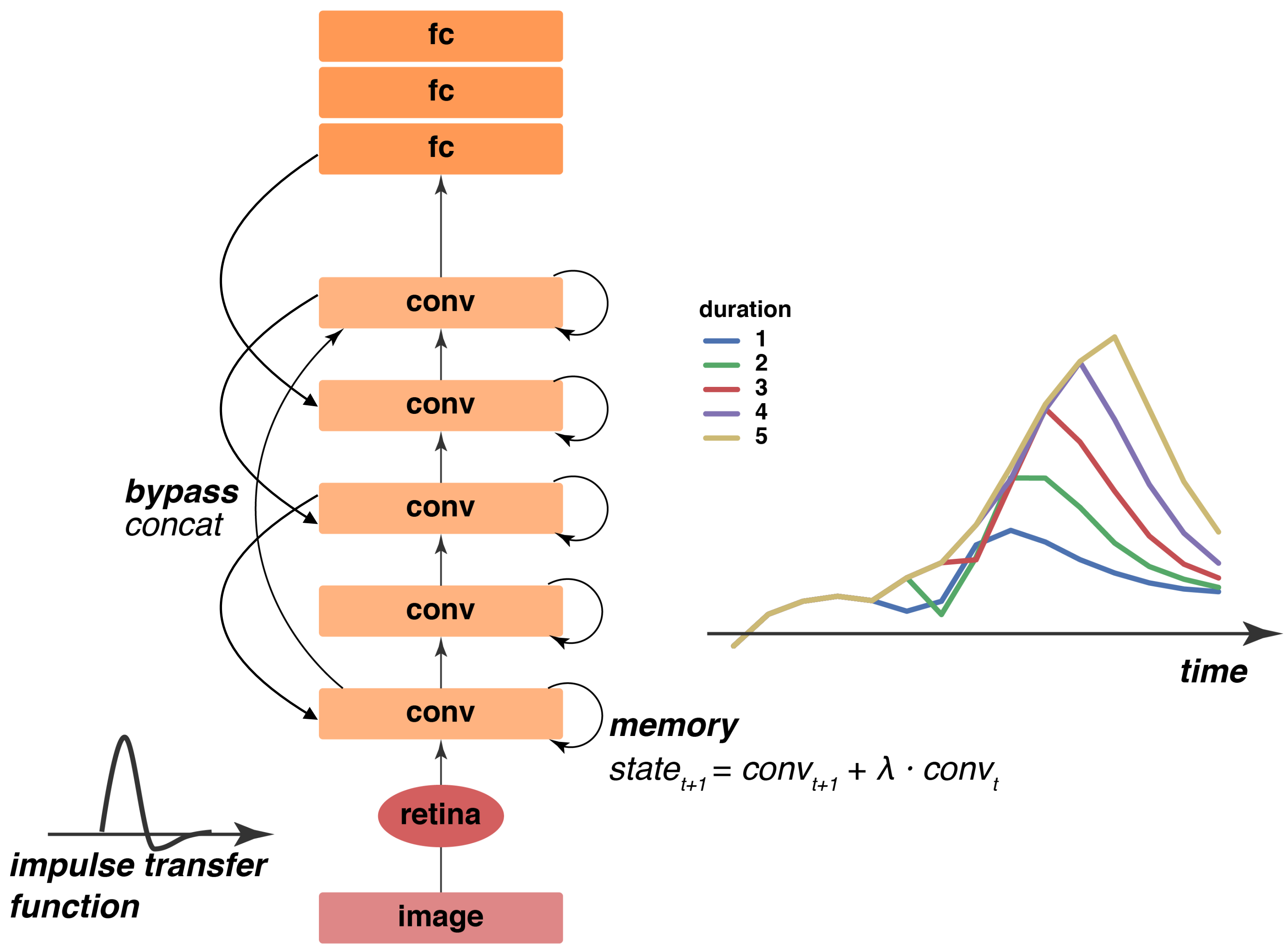


Maryann Rui

Kohitij Kar

Jim DiCarlo

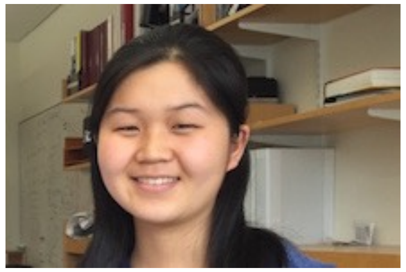
Better architectures



Aran Nayebi

Surya Ganguli

Jonas Kubilius

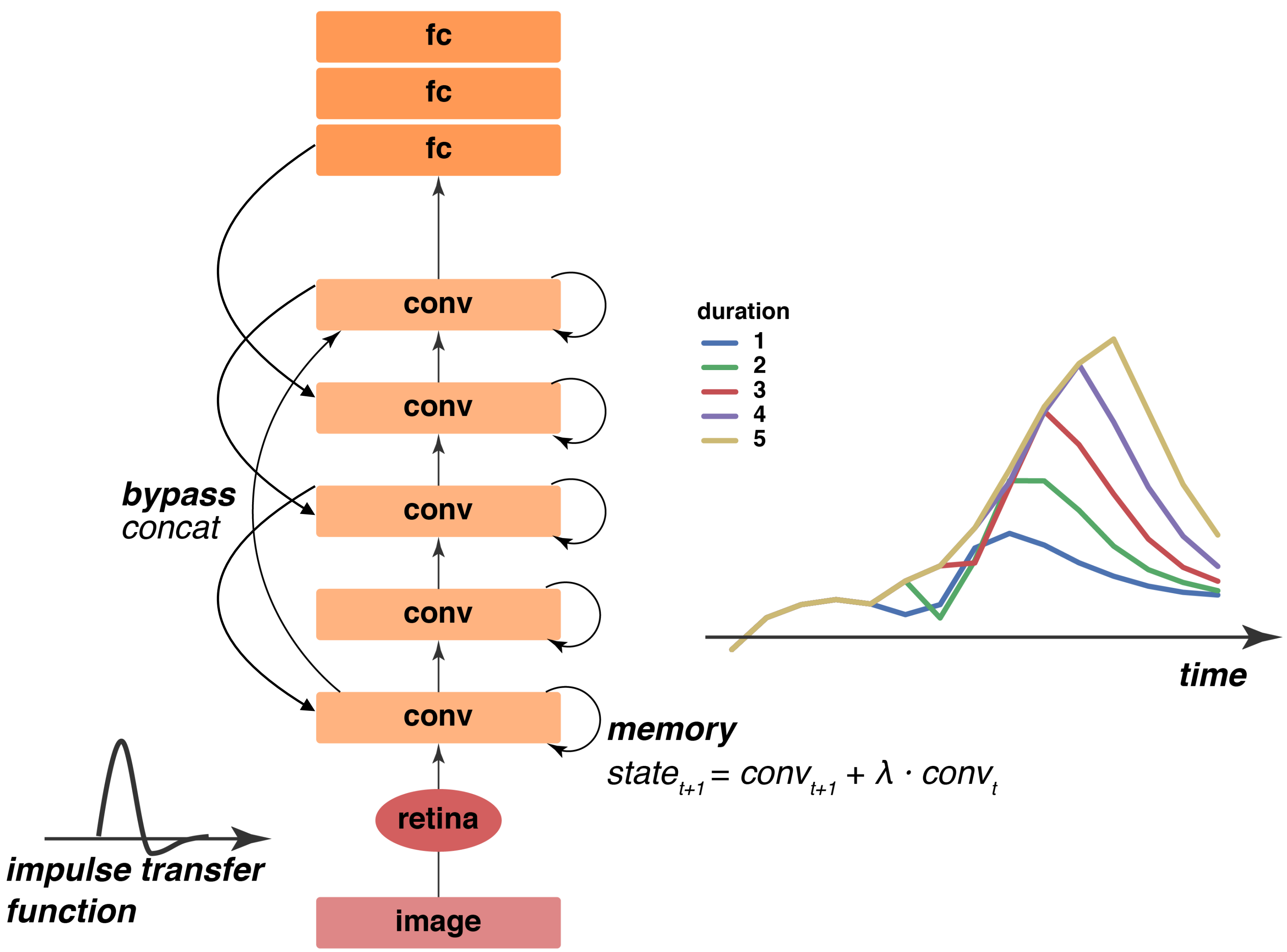


Maryann Rui

Kohitij Kar

Jim DiCarlo

Better architectures



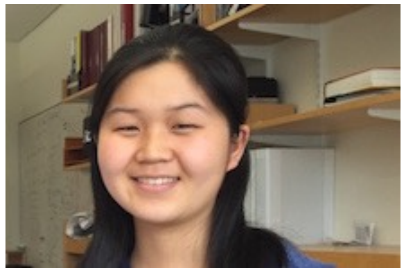
What task(s)?

a) vanilla categorization

Aran Nayebi

Surya Ganguli

Jonas Kubilius

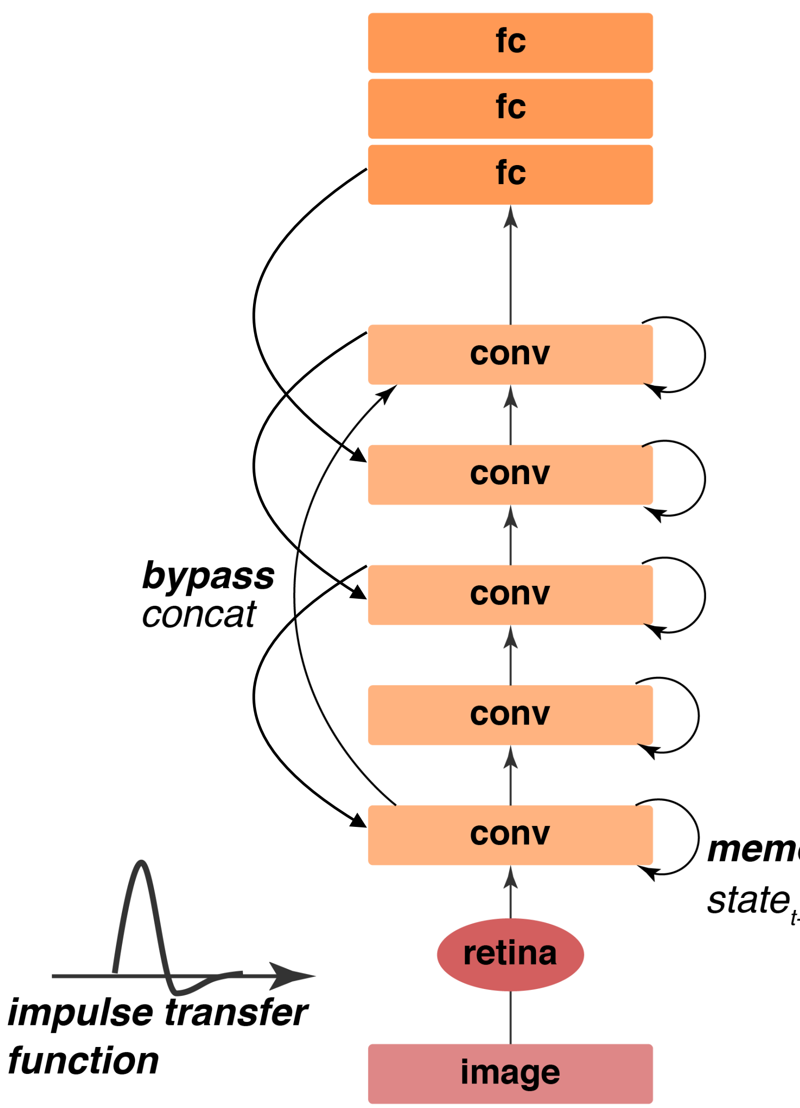


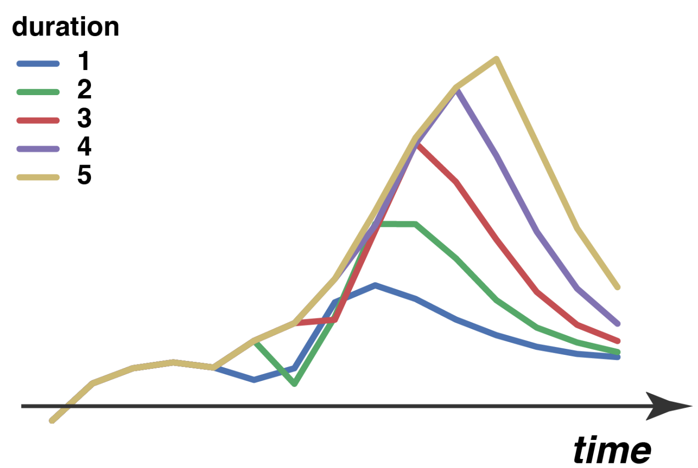
Maryann Rui

Kohitij Kar

Jim DiCarlo

Better architectures





What task(s)?

a) vanilla categorization

b) time-discounting

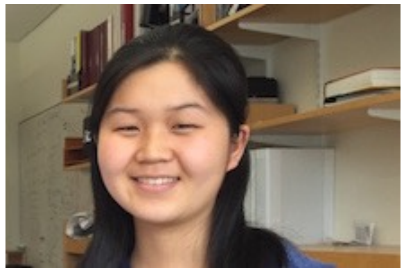
$$L = \sum_t \gamma^t \cdot L_t$$

be accurate but also fast

Aran Nayebi

Surya Ganguli

Jonas Kubilius

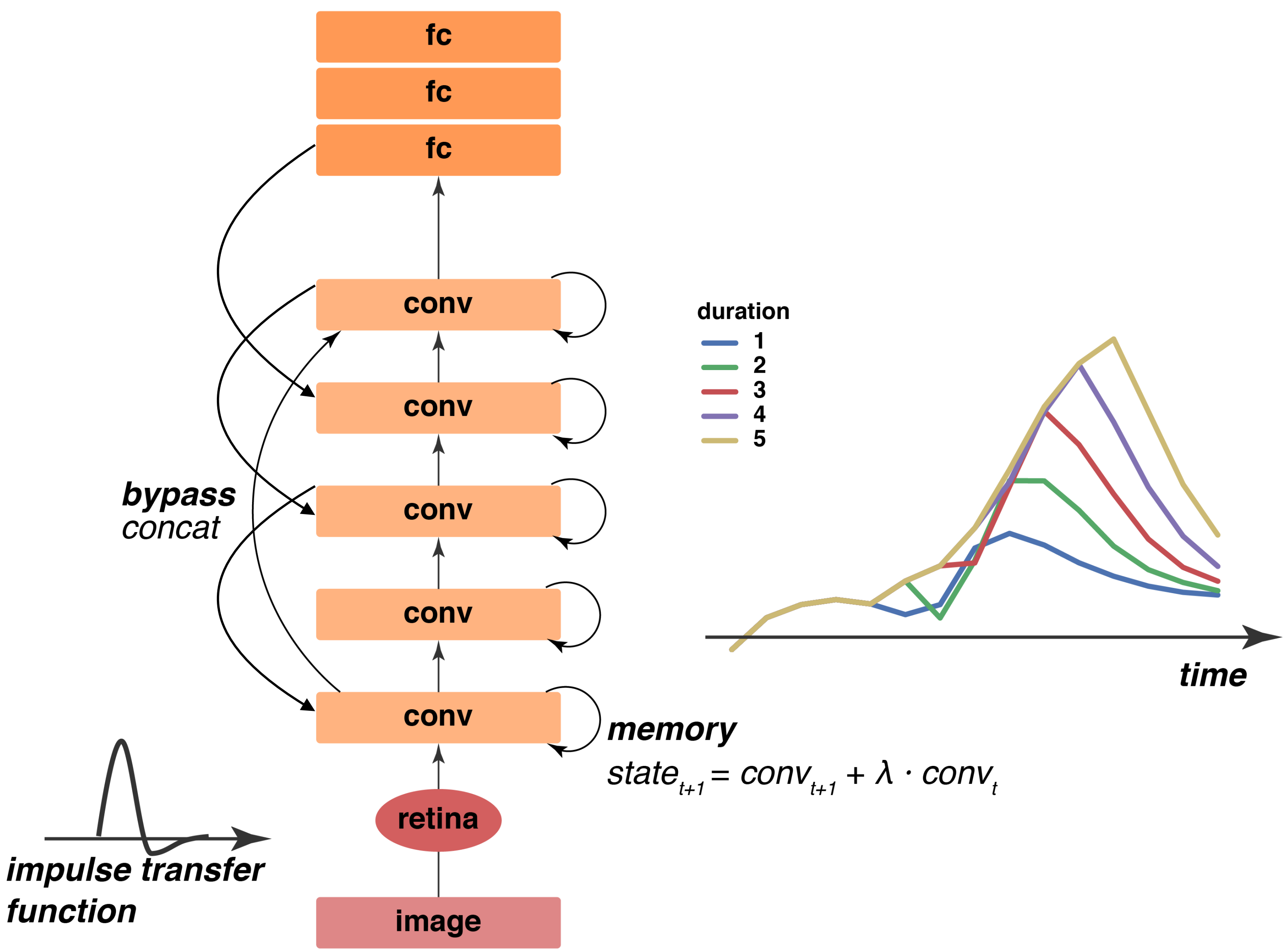


Maryann Rui

Kohitij Kar

Jim DiCarlo

Better architectures



What task(s)?

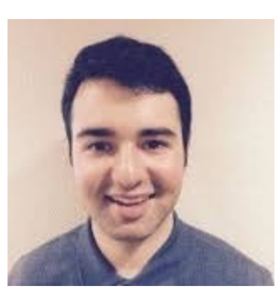
a) vanilla categorization

b) time-discounting

$$L = \sum_t \gamma^t \cdot L_t$$

<u>*be accurate but also fast*</u>

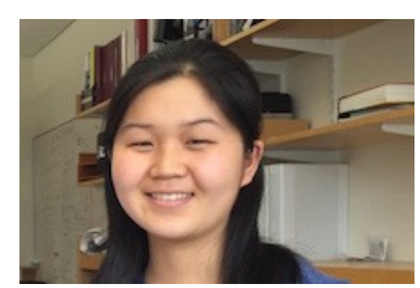
c) heavy occlusion &c



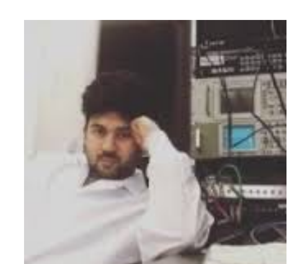
Aran Nayebi

Surya Ganguli

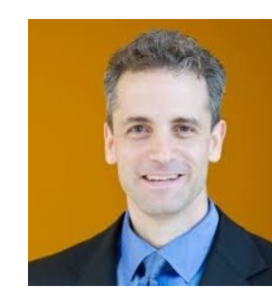
Jonas Kubilius



Maryann Rui

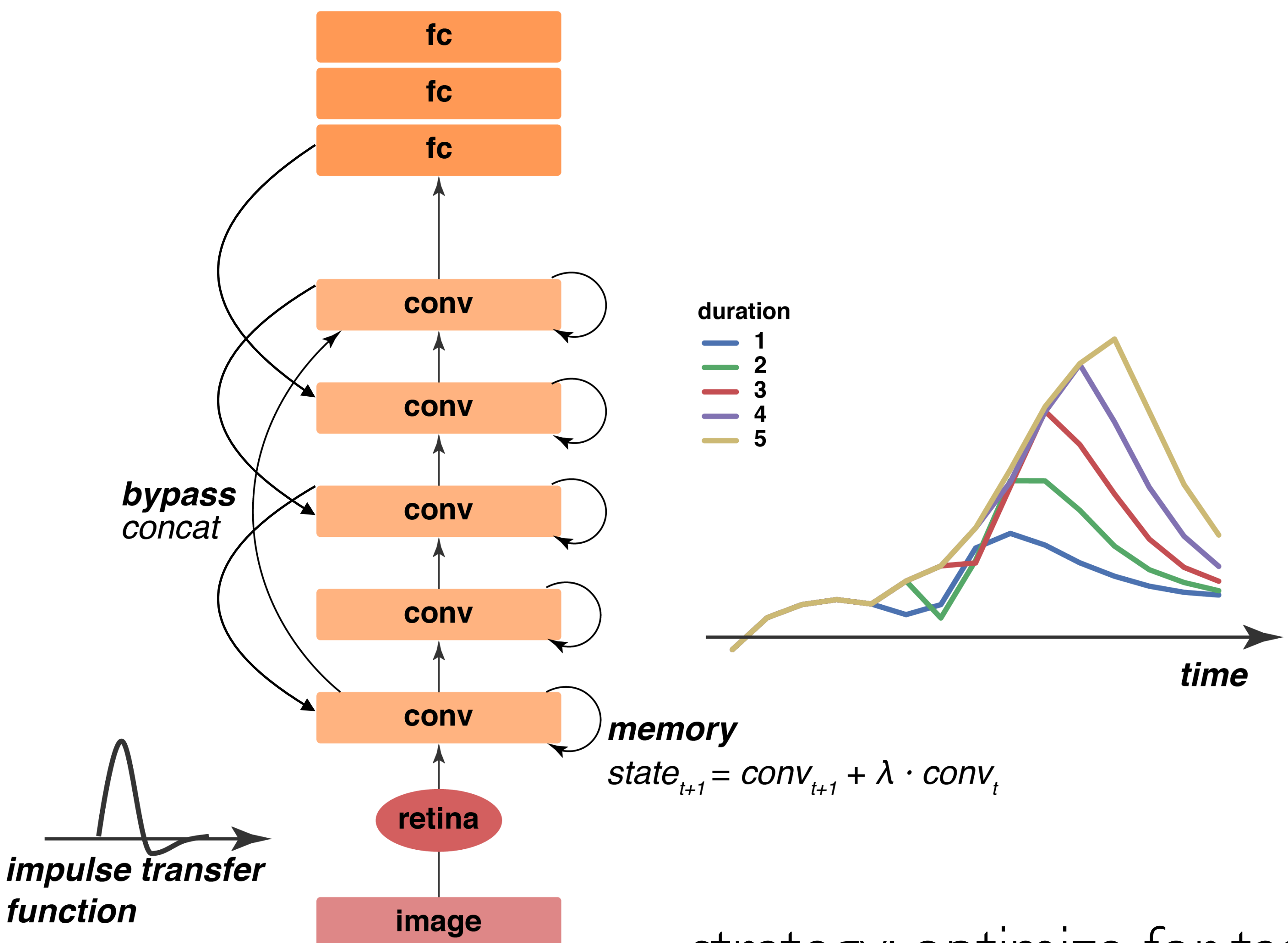


Kohitij Kar



Jim DiCarlo

Better architectures



What task(s)?

a) vanilla categorization

b) time-discounting

$$L = \sum_t \gamma^t \cdot L_t$$

be accurate but also fast

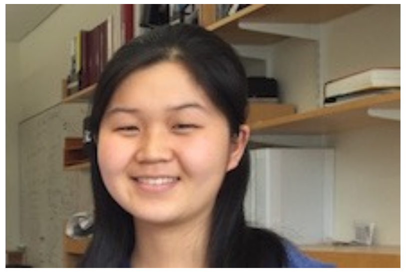
c) heavy occlusion &c

strategy: optimize for tasks check against static & dynamic data

Aran Nayebi

Surya Ganguli

Jonas Kubilius



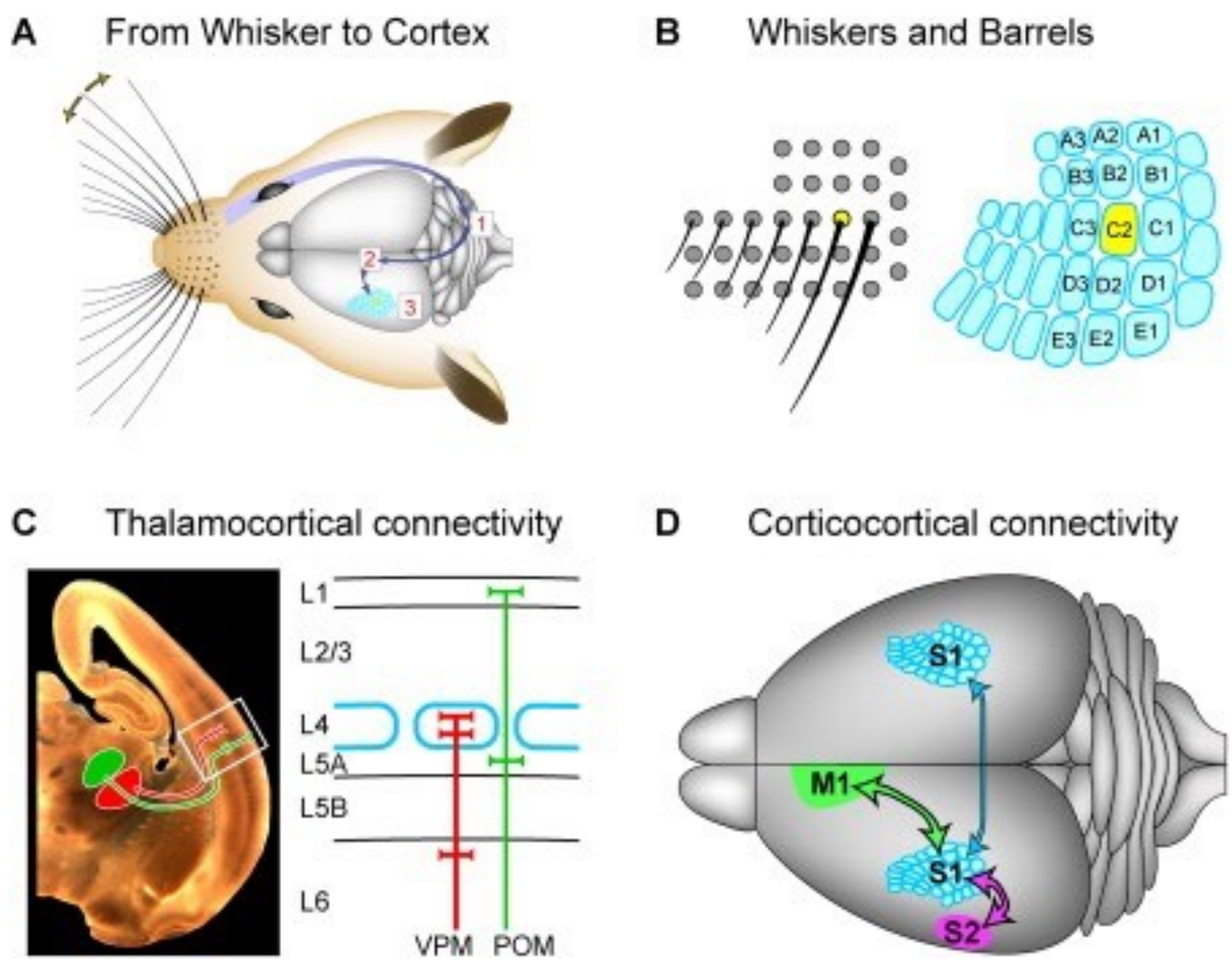
Maryann Rui

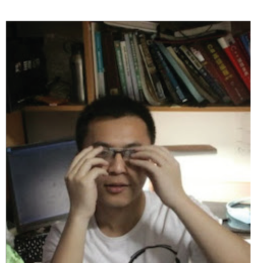
Kohitij Kar

Jim DiCarlo

Rodent Somatosensory Cortex

Petersen, 2007



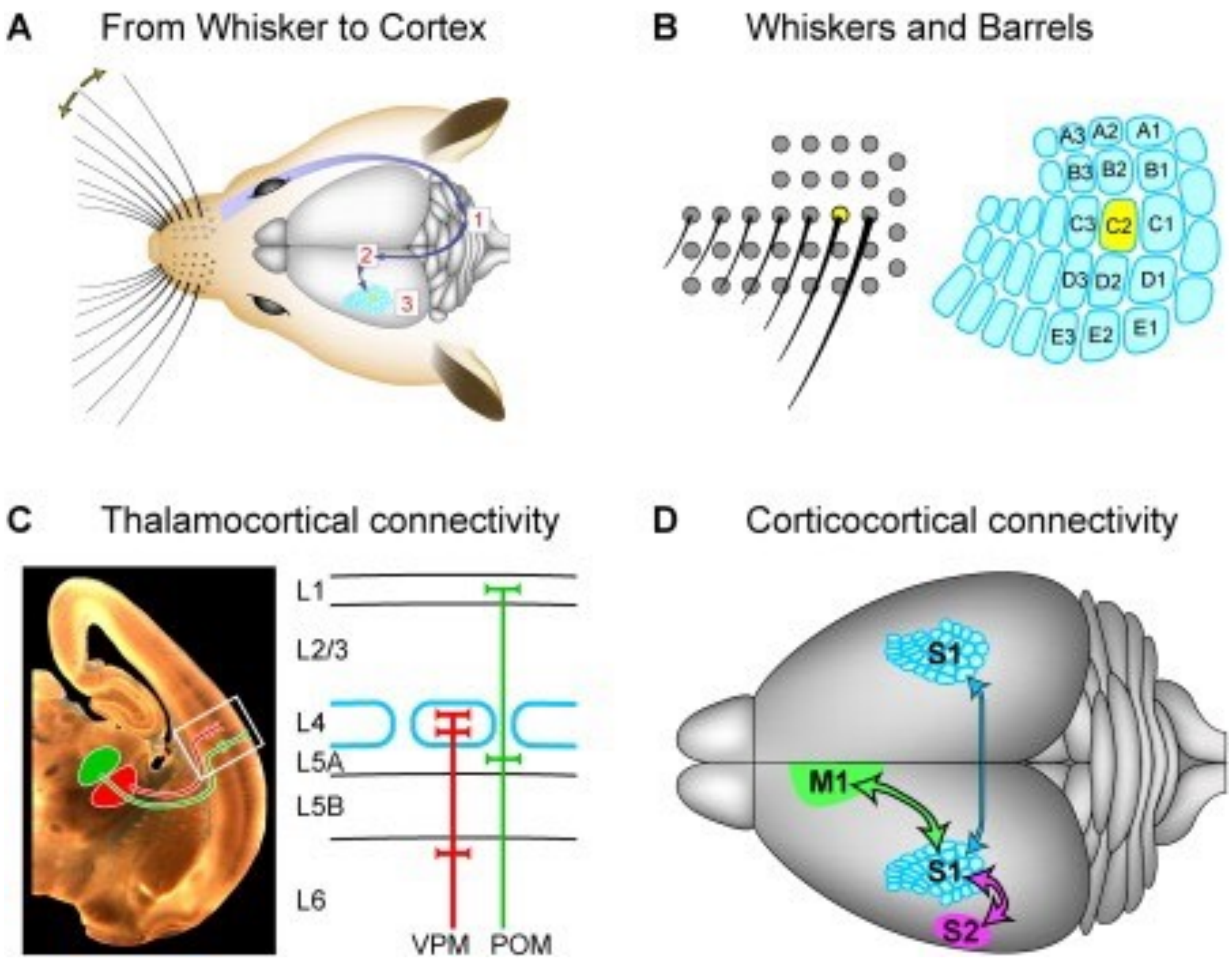


Chengxu Zhuang

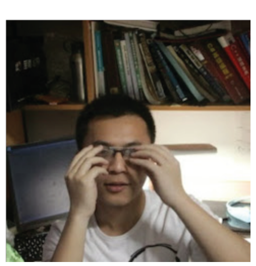
Mitra Hartmann & Lab

Rodent Somatosensory Cortex

Petersen, 2007



- Spatially-structured input data

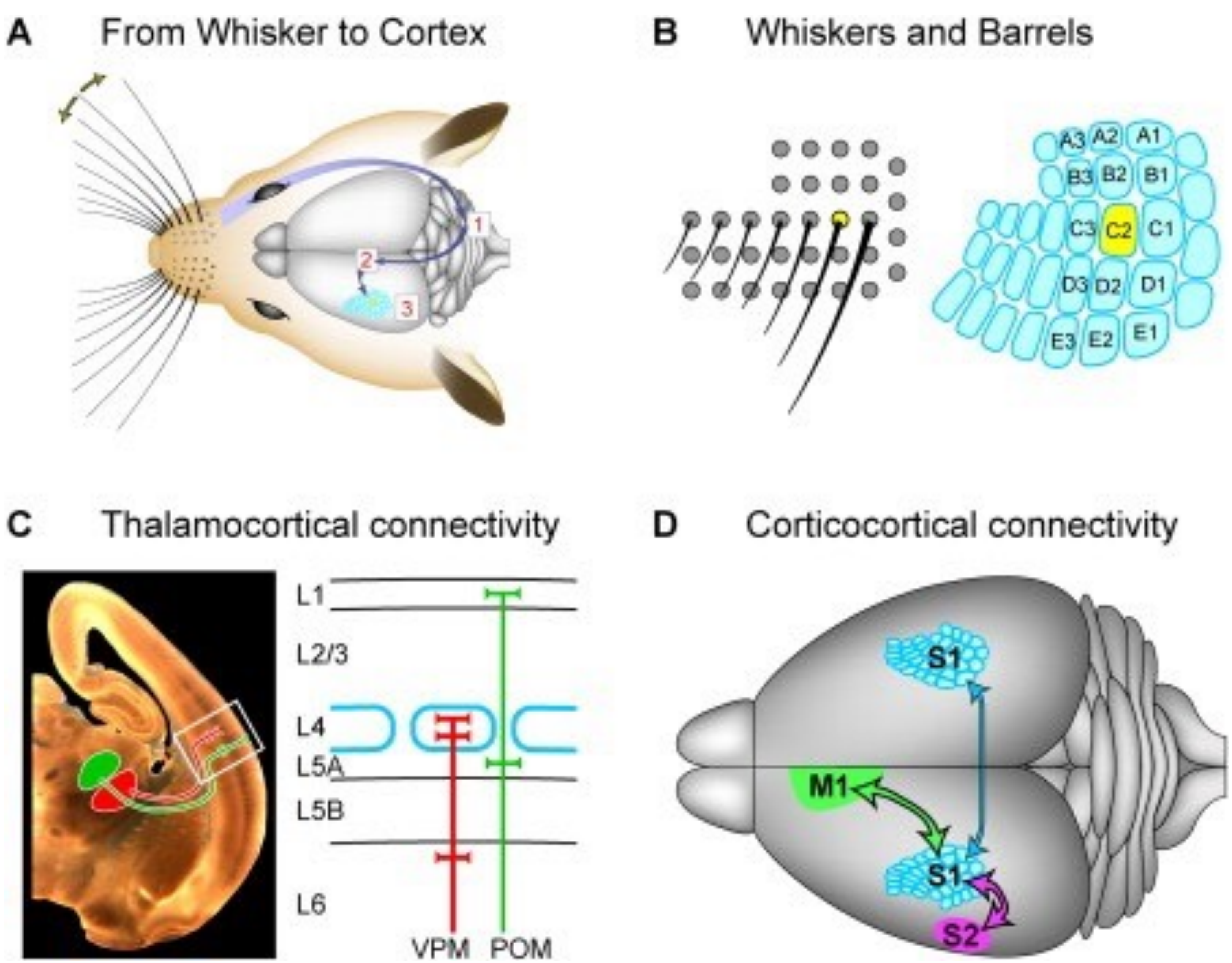


Chengxu Zhuang

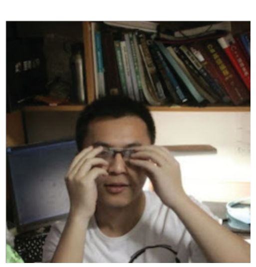
Mitra Hartmann & Lab

Rodent Somatosensory Cortex

Petersen, 2007



- Spatially-structured input data
- Spatiotopic sensor

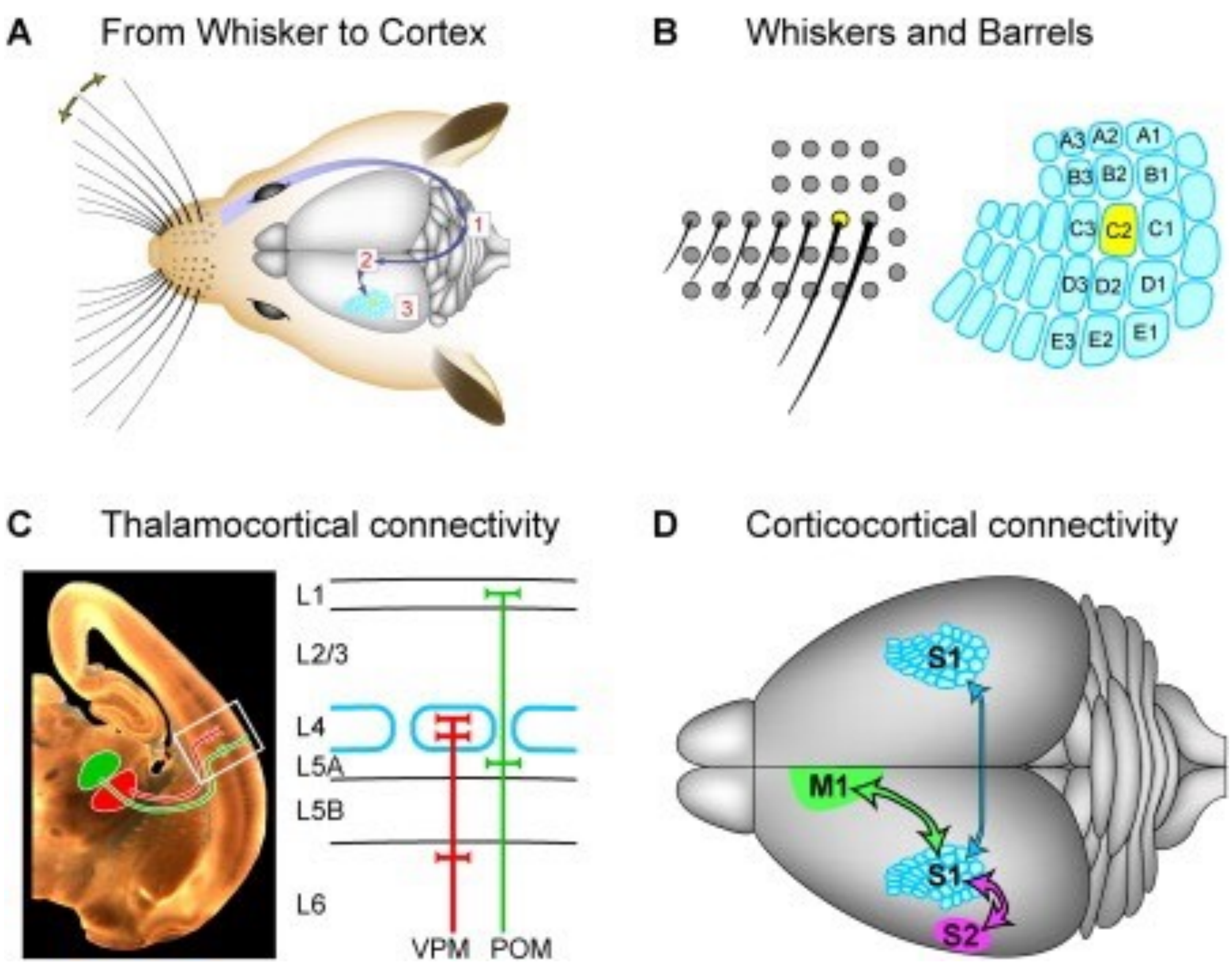


Chengxu Zhuang

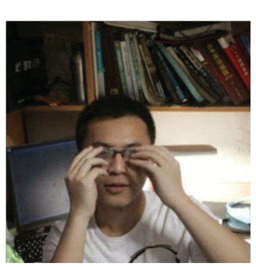
Mitra Hartmann & Lab

Rodent Somatosensory Cortex

Petersen, 2007



- ✳ Spatially-structured input data
- ✳ Spatiotopic sensor
- ✳ Potentially hierarchical structure

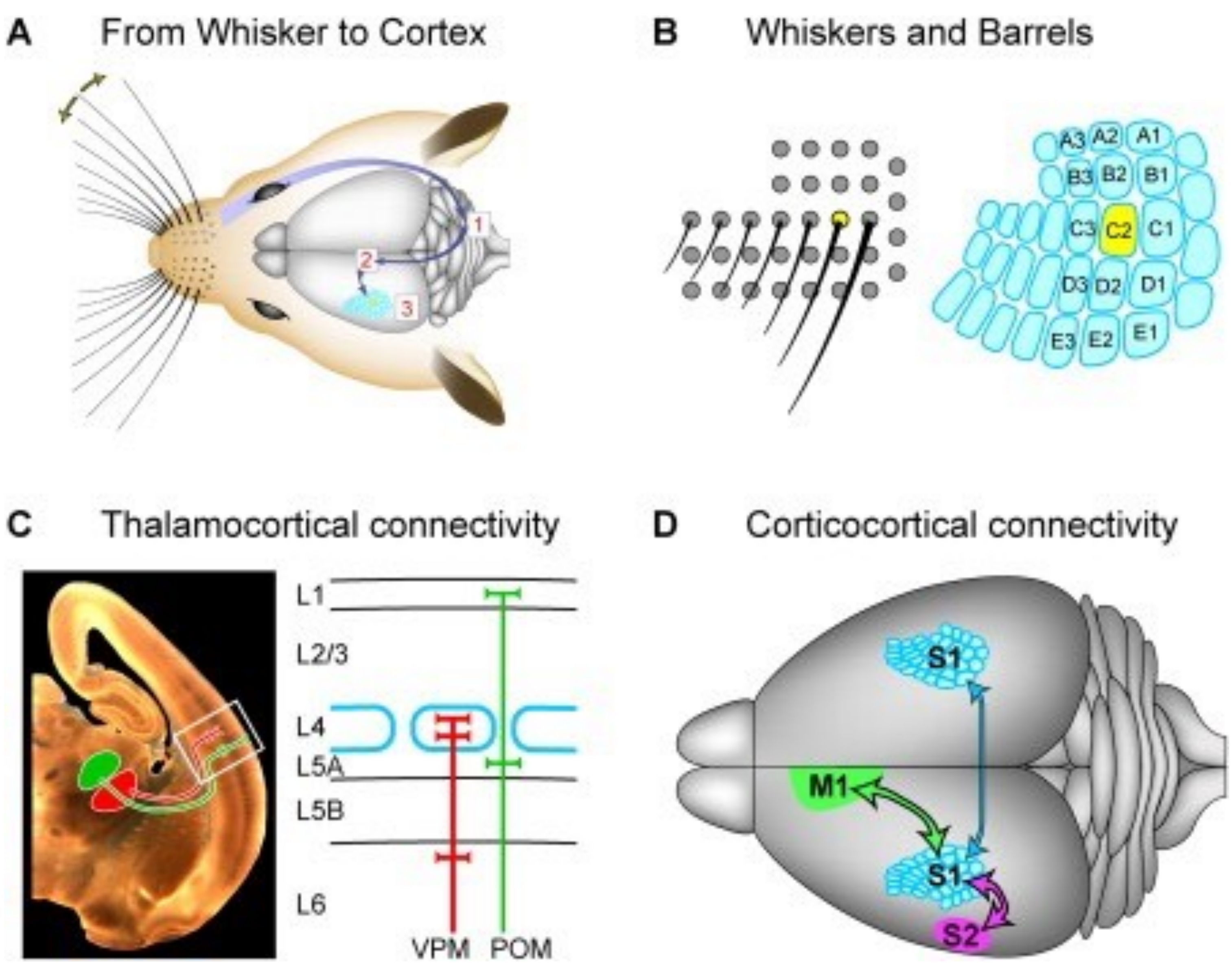


Chengxu Zhuang

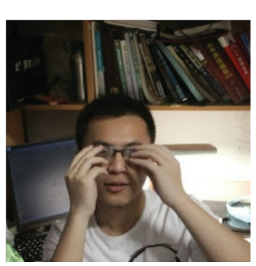
Mitra Hartmann & Lab

Rodent Somatosensory Cortex

Petersen, 2007



* Spatially-structured input data
* Spatiotopic sensor
* Potentially hierarchical structure
* Poorly understood higher cortical areas

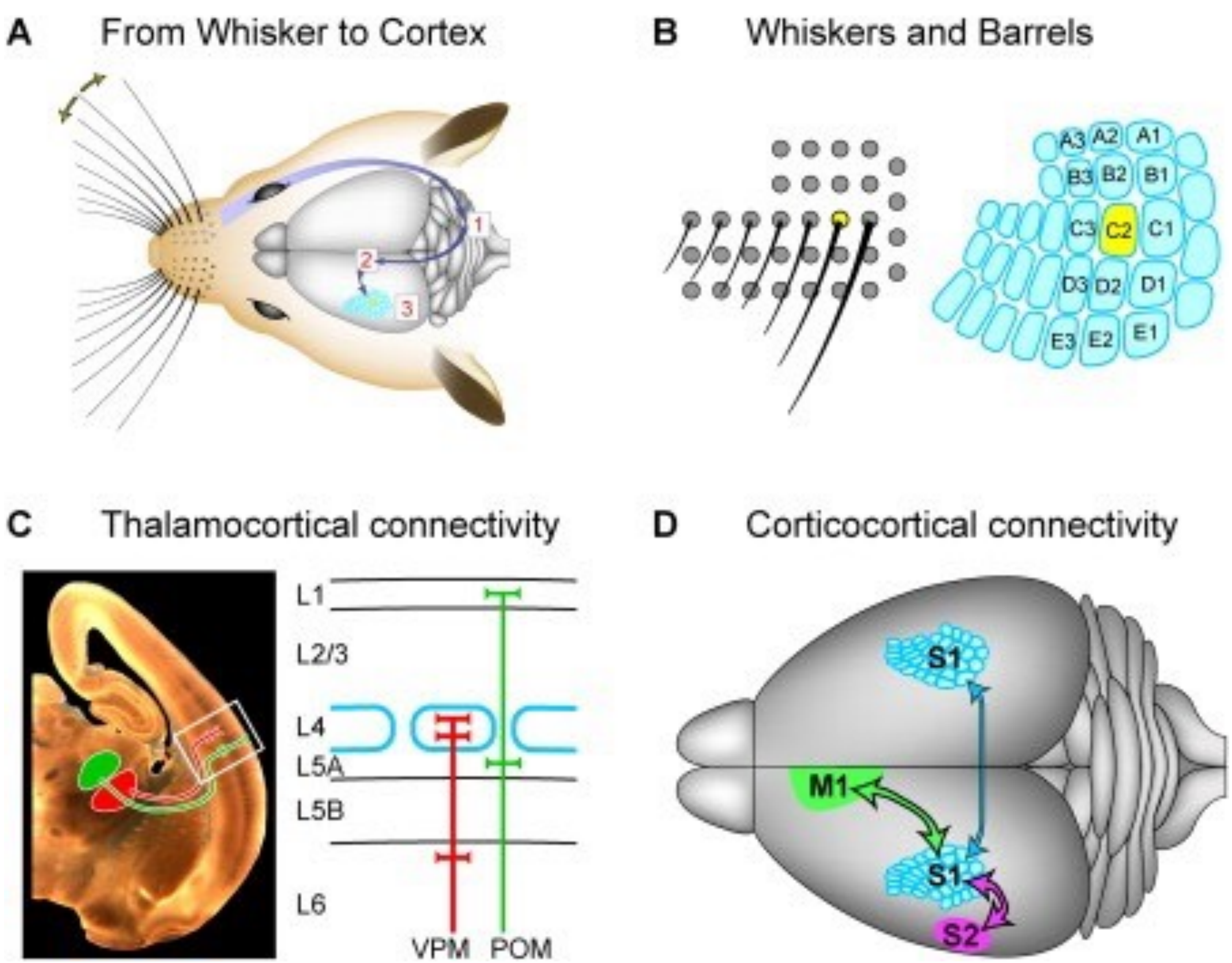


Chengxu Zhuang

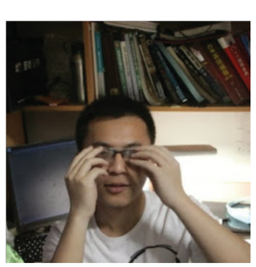
Mitra Hartmann & Lab

Rodent Somatosensory Cortex

Petersen, 2007



- ✳ Spatially-structured input data
- ✳ Spatiotopic sensor
- ✳ Potentially hierarchical structure
- ✳ Poorly understood higher cortical areas



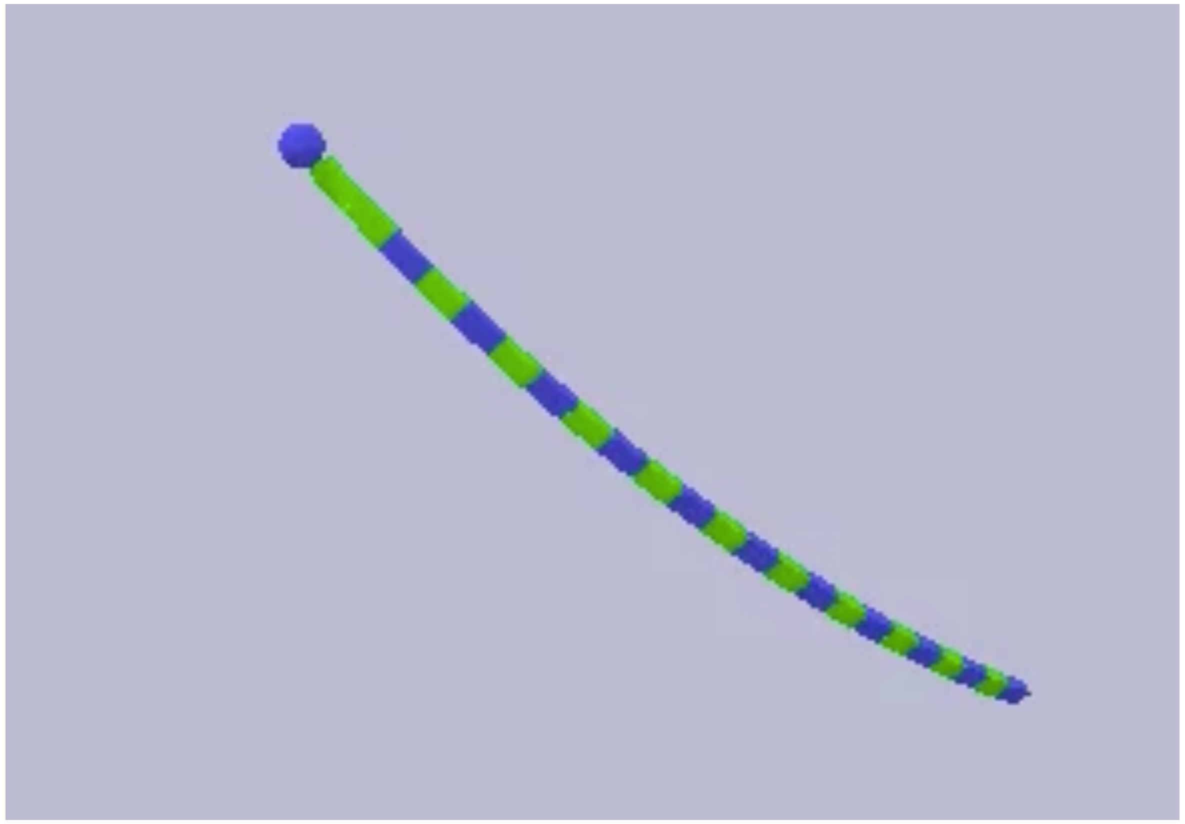
Chengxu Zhuang

Mitra Hartmann & Lab

Hypothesis: can get a model for this cortical cascade by optimizing properly-sized CNN with whisker-like sensor input for some ethologically relevant somatosensory task.

Rodent Somatosensory Cortex

First have to build a model of the sensory to gather data.



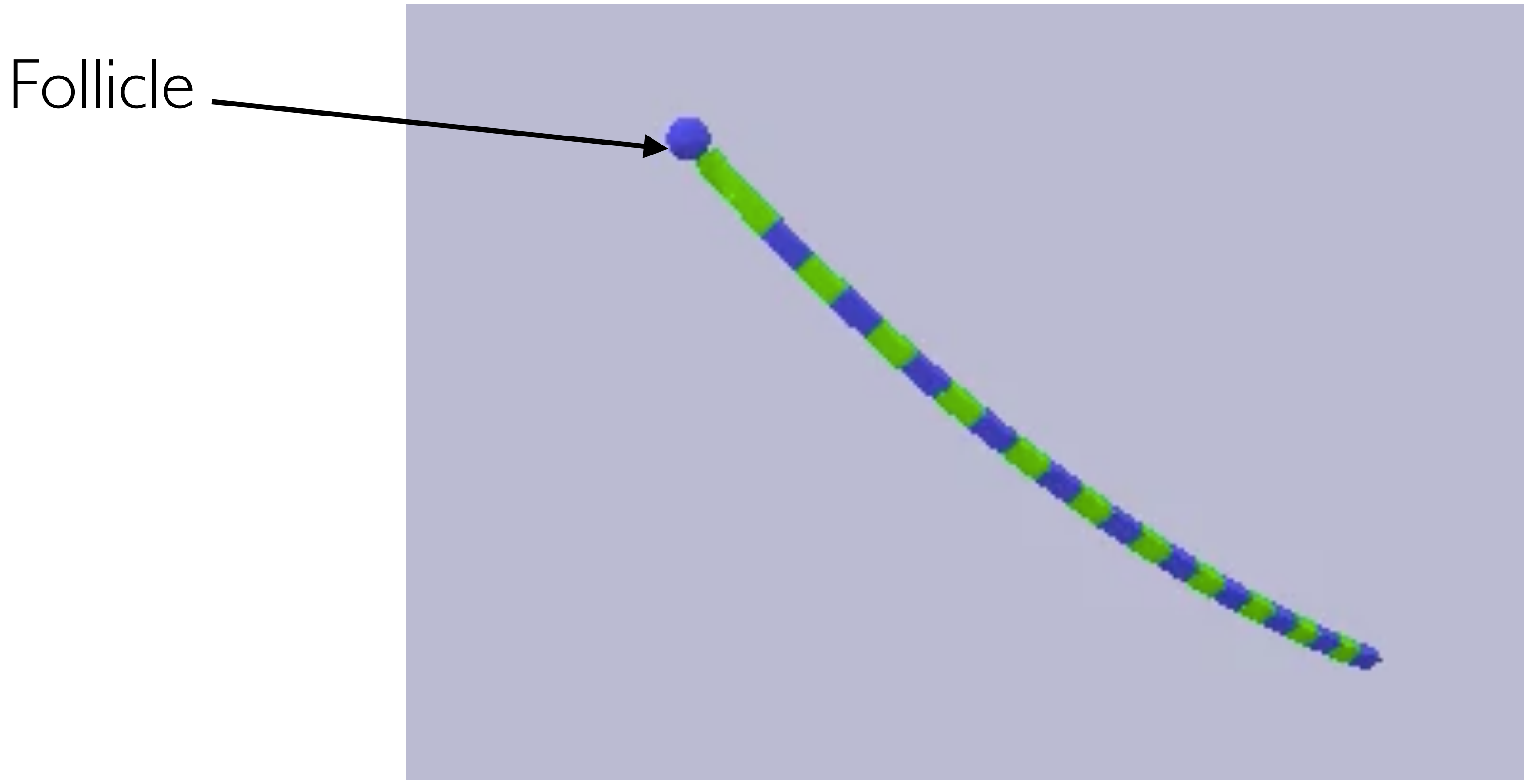
Using published data from Mitra Hartmann's group

Chengxu Zhuang

Mitra Hartmann & Lab

Rodent Somatosensory Cortex

First have to build a model of the sensory to gather data.



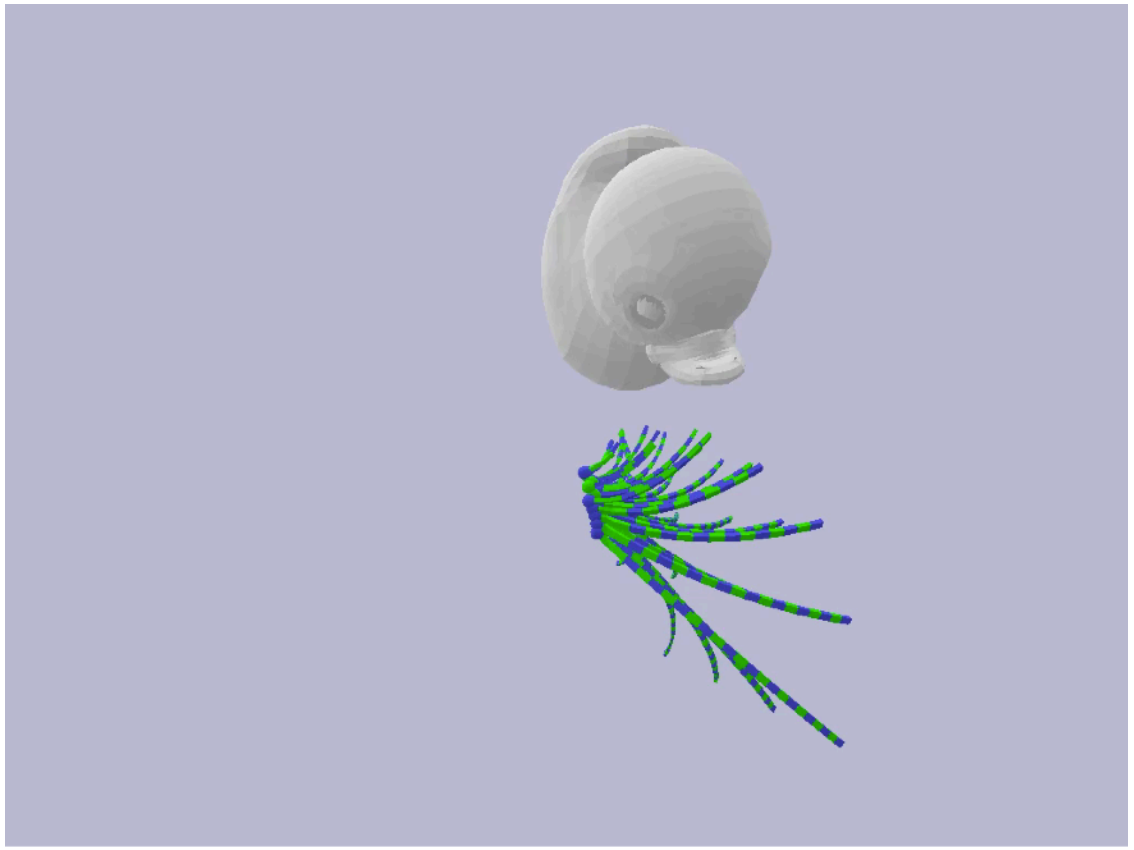
Using published data from Mitra Hartmann's group

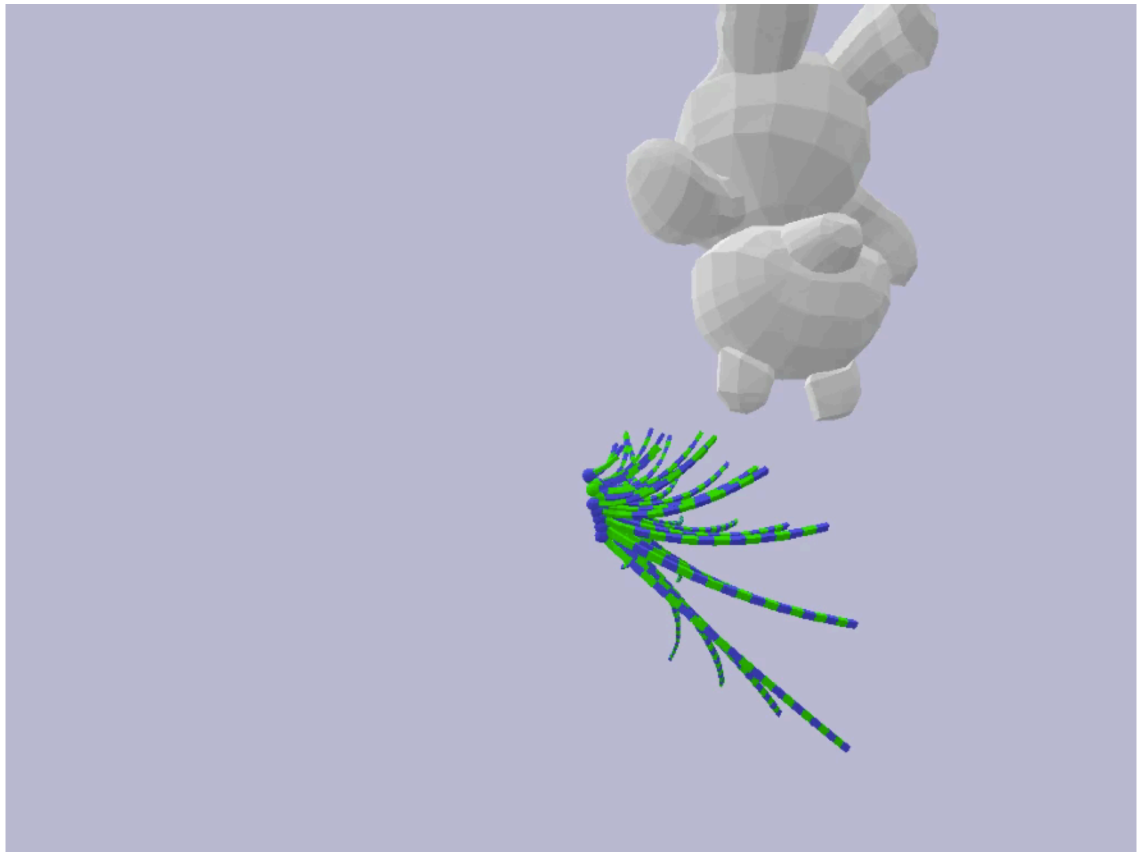
Chengxu Zhuang

Mitra Hartmann & Lab

Rodent Somatosensory Cortex

First have to build a model of the sensory to gather data.



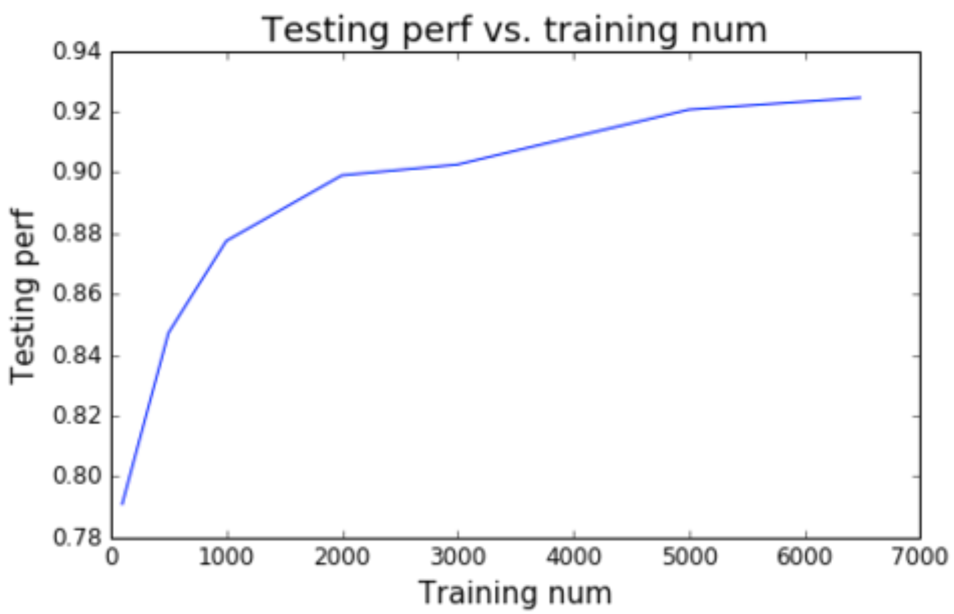


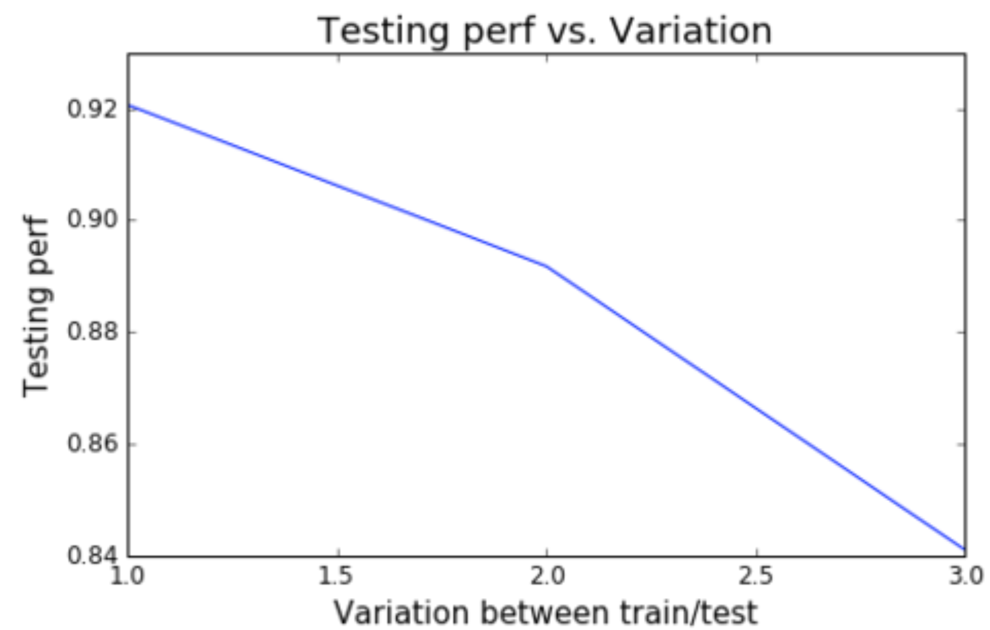
Chengxu Zhuang

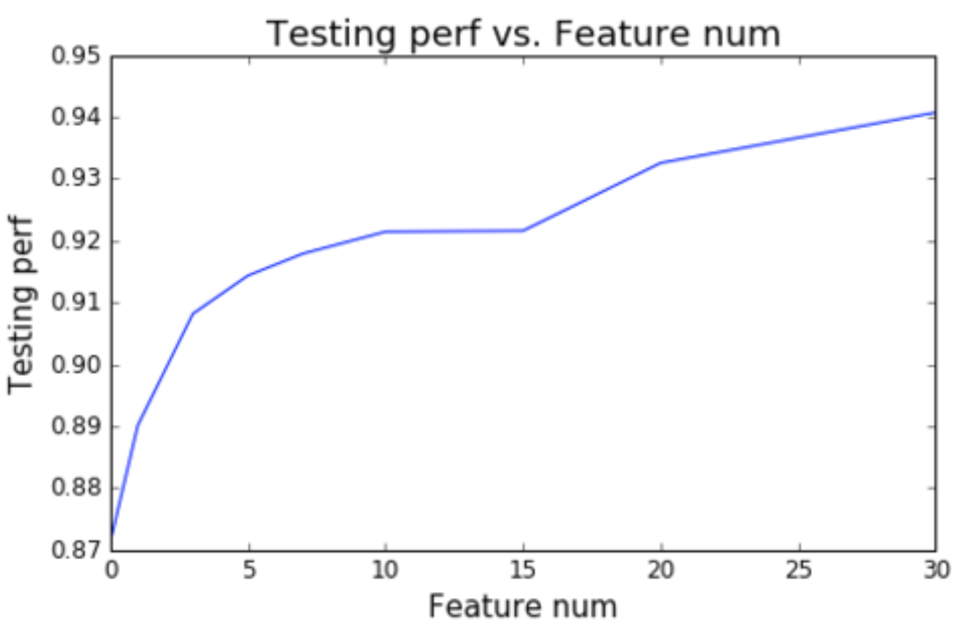
Mitra Hartmann & Lab

Rodent Somatosensory Cortex

Validate sensor on one-object tasks … (teddy vs. duck)







train/test splits with different:

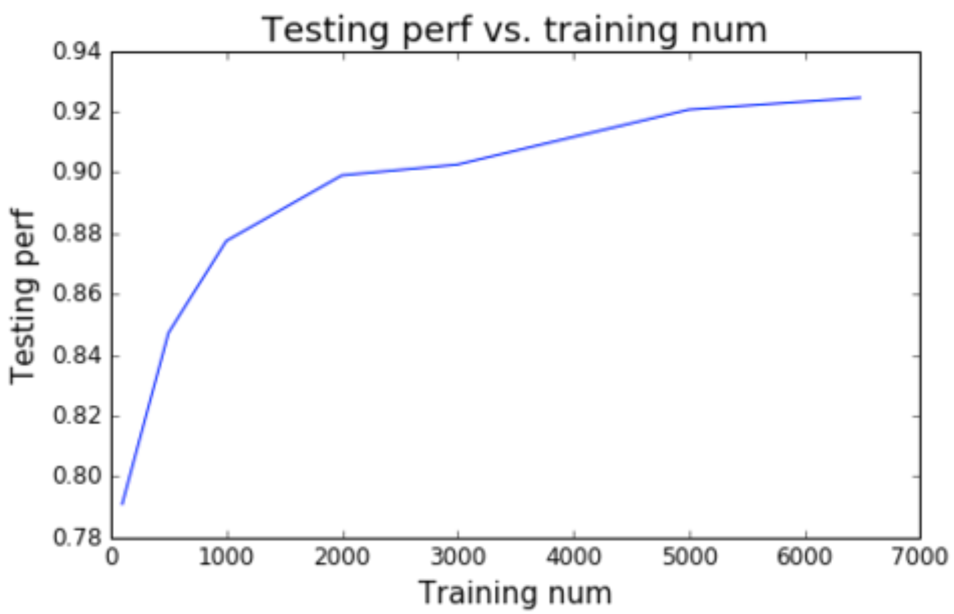
- ✴ attack vectors
- ✴ attack speed
- ✴ object rotations
- ✴ object size

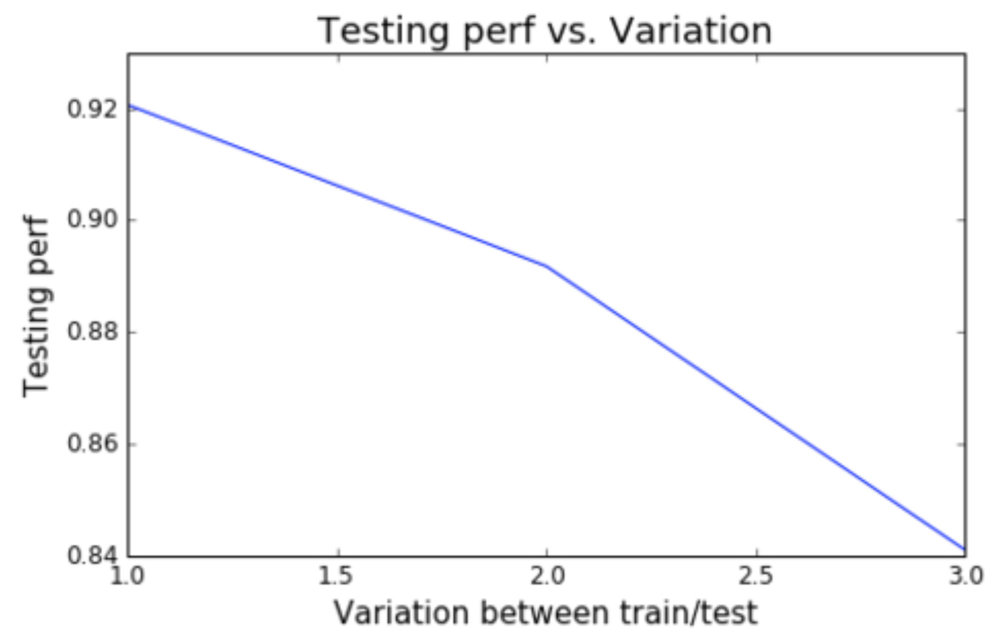
Chengxu Zhuang

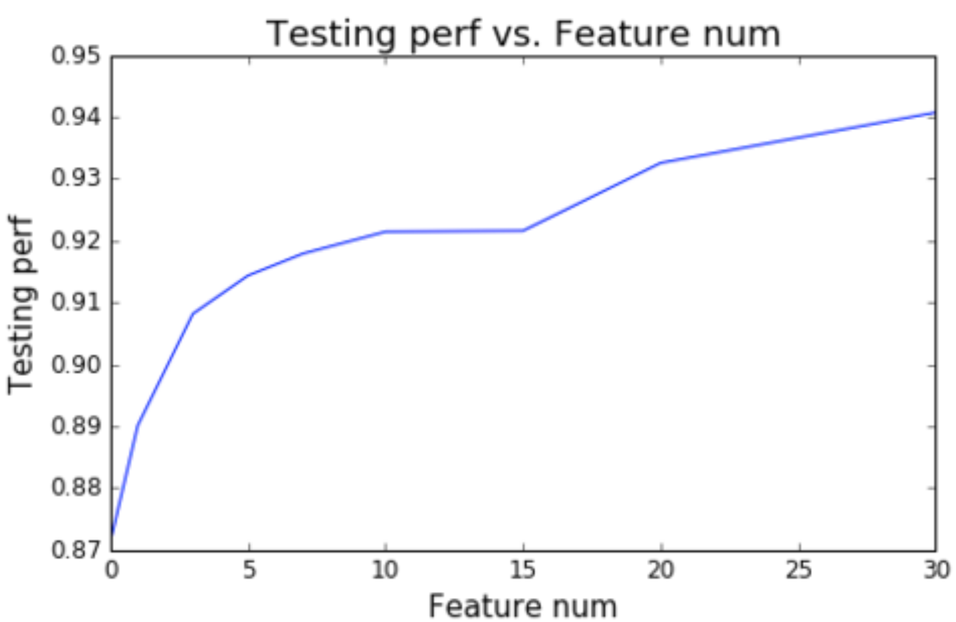
Mitra Hartmann & Lab

Rodent Somatosensory Cortex

Validate sensor on one-object tasks … (teddy vs. duck)

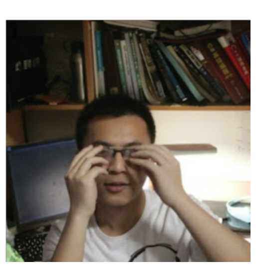






train/test splits with different:

- ✳ attack vectors
- ✳ attack speed
- ✳ object rotations
- ✳ object size



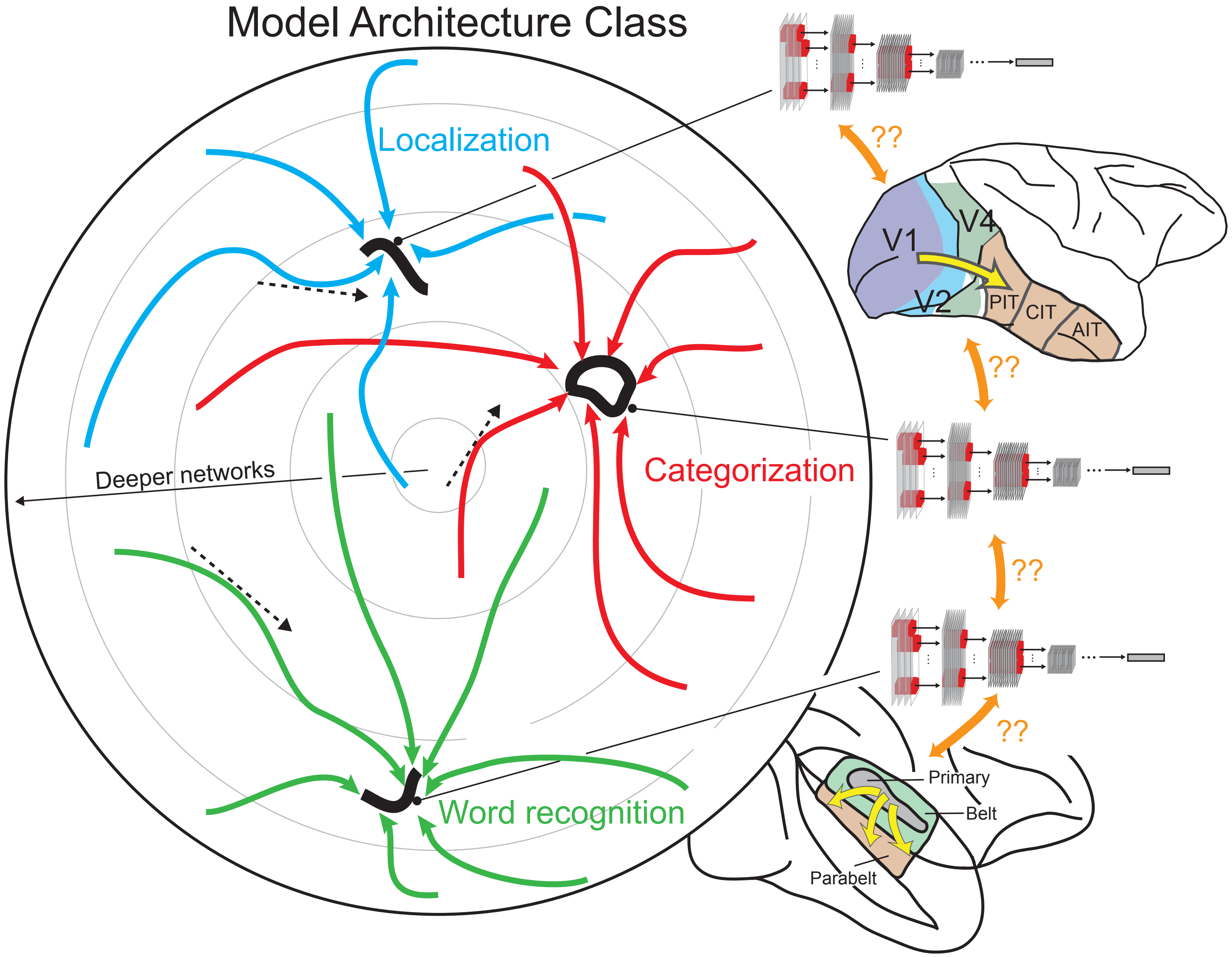
Chengxu Zhuang

Mitra Hartmann & Lab

Train on shape recognition and/or normal estimation task, compare to neural data

Rodent Somatosensory Cortex

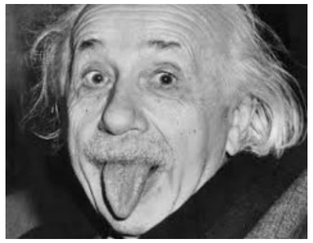
***If** successful:*



somatosensation

vision

audition

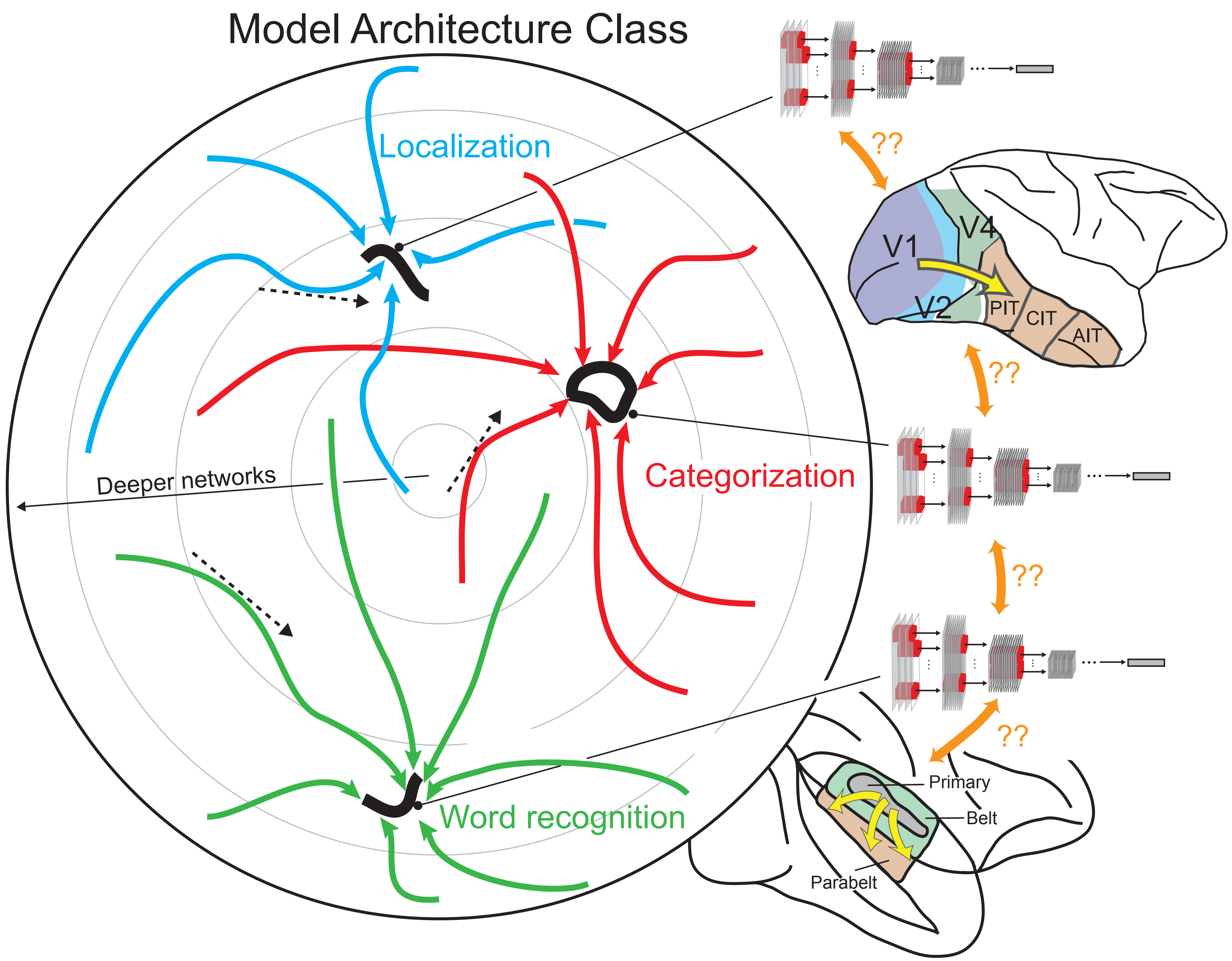


> Formulate comprehensive model class (**CNNs**)

> Choose challenging, ethologically-valid tasks (**categorization**)

> Implement generic learning rules (**gradient descent**)

> Map to brain data. (**ventral stream**)

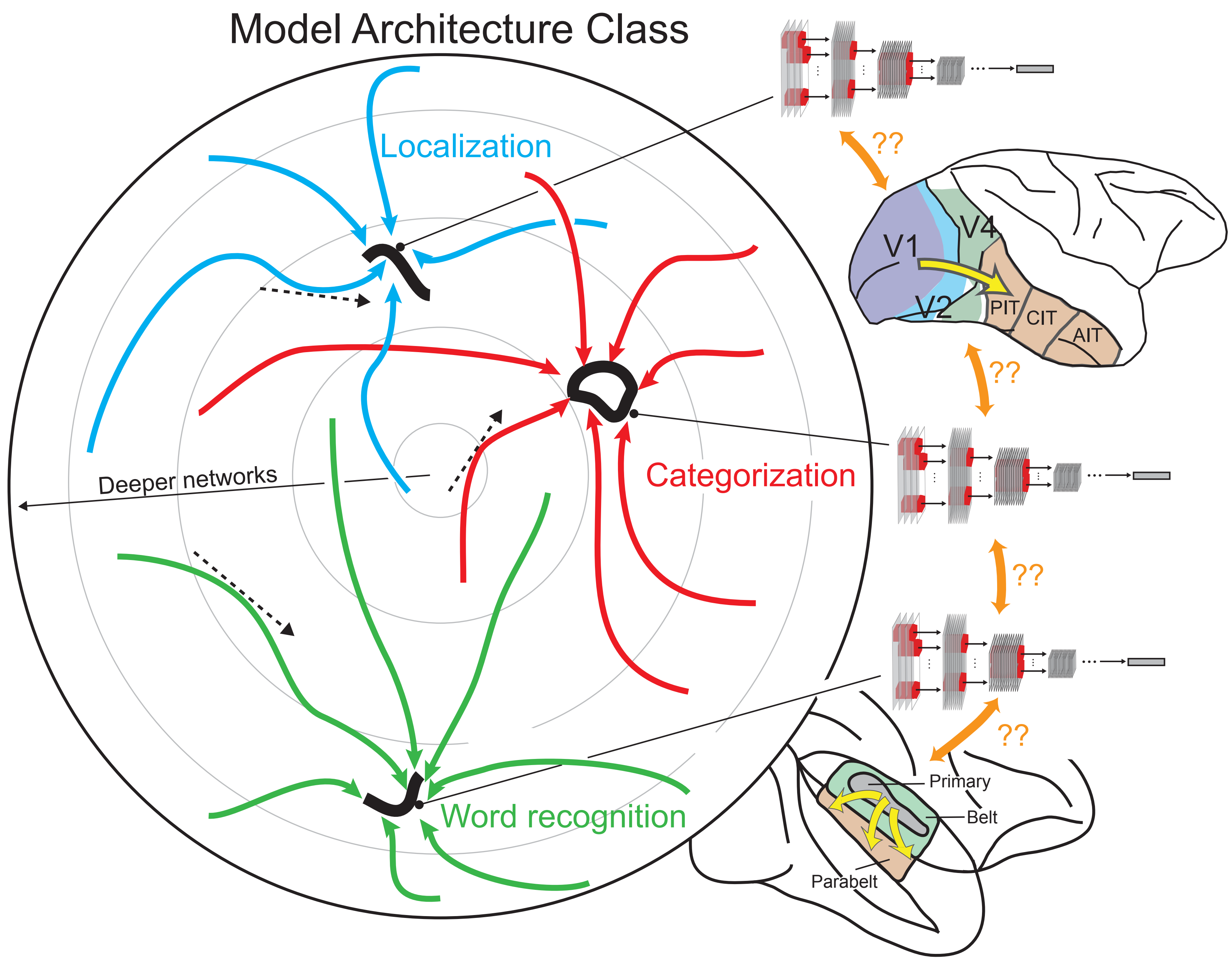


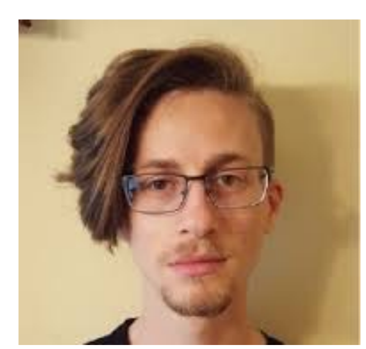
> Formulate comprehensive model class (**CNNs + RNs**)

> Choose challenging, ethologically-valid tasks (**task switching/ memory**)

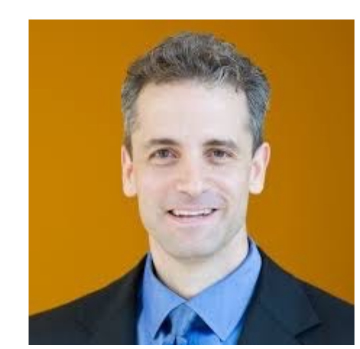
> Implement generic learning **and expansion** rules

> Map to brain data. (**PFC, Hippocampus, &c**)




Kevin Feigelis


Mark Schnitzer


Jim DiCarlo

Q&A

Key Results

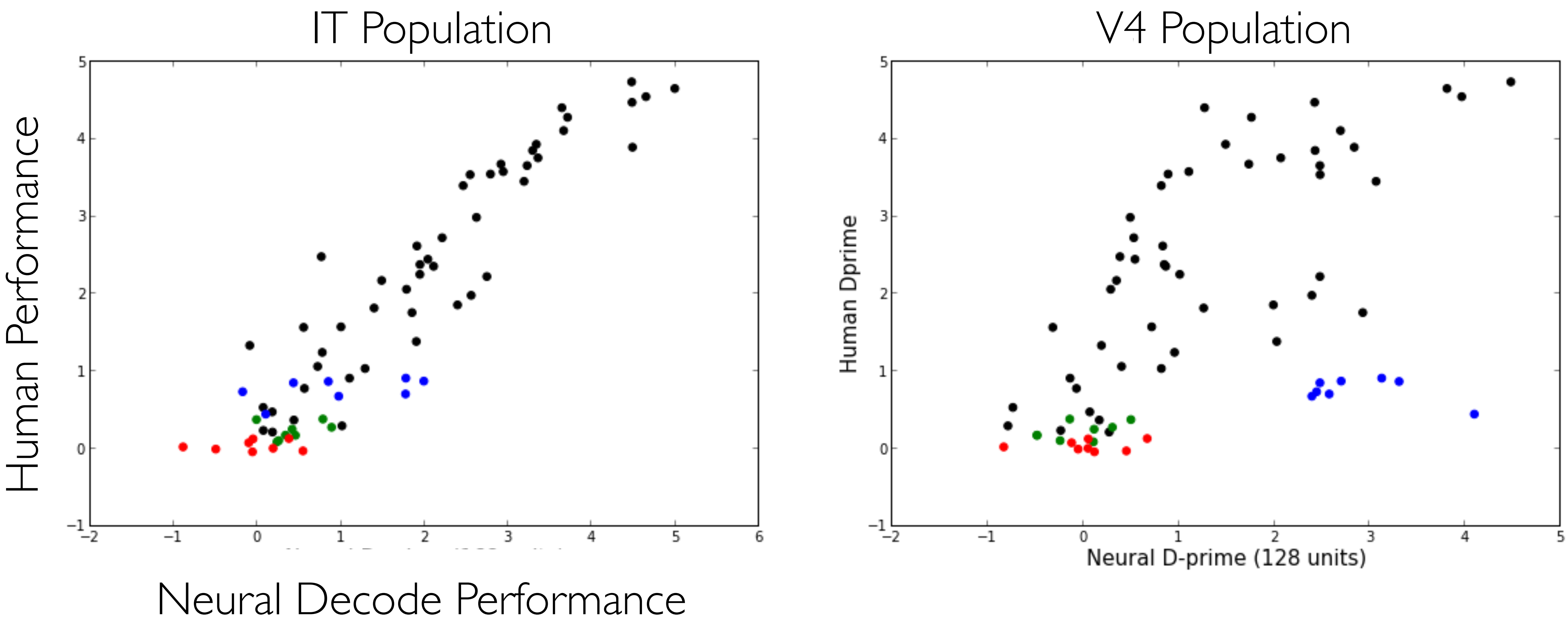
Task-driven modeling can make greatly improved quantitative models of high-level cortical areas.

These models can lead to new qualitative insight about how the brain solves sensory tasks.

These concepts are useful across multiple sensory modalities.

IT Neurons Track Human Performance

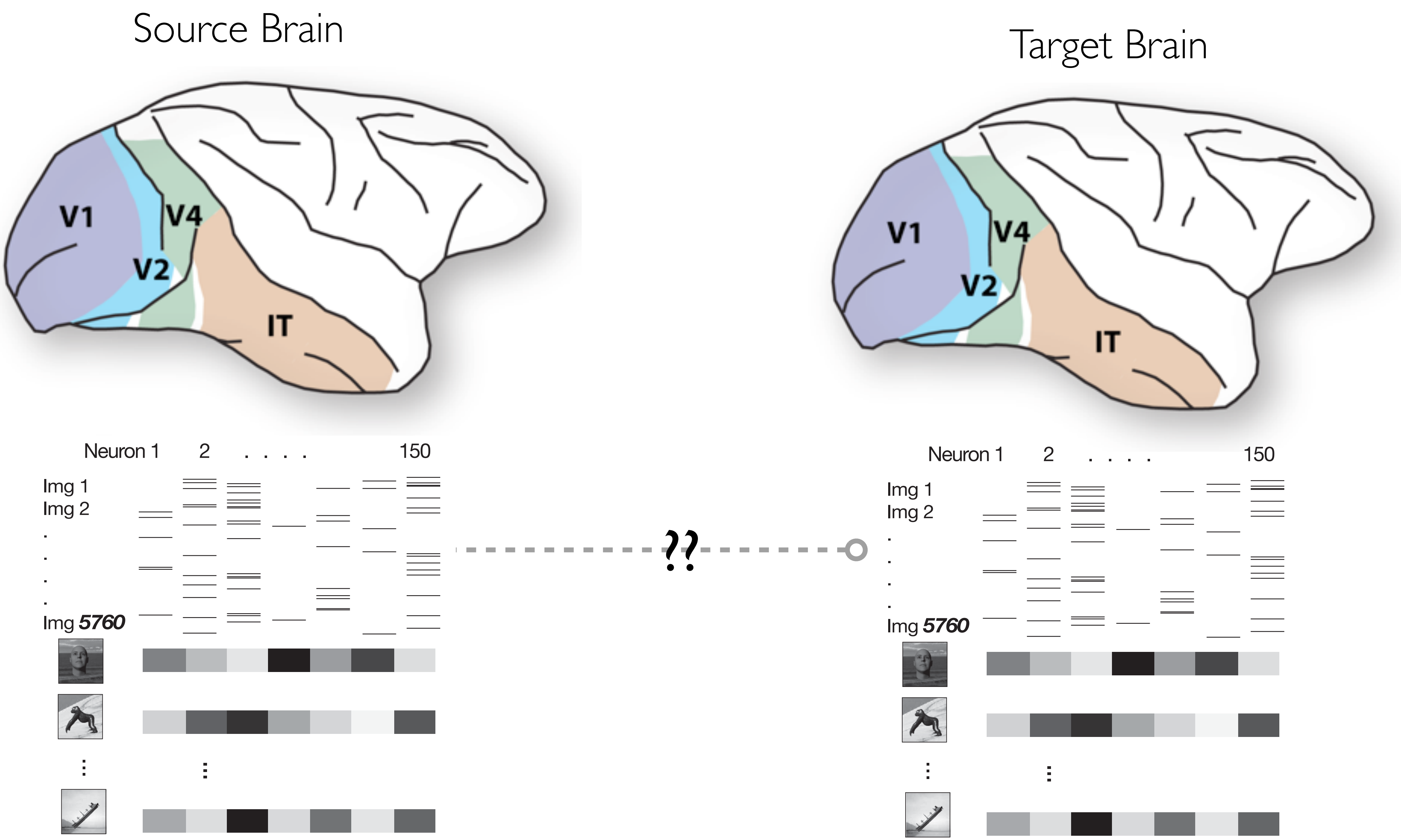
IT matches human error patterns as well as raw performance.



- Low-Variation Face subordinate tasks.

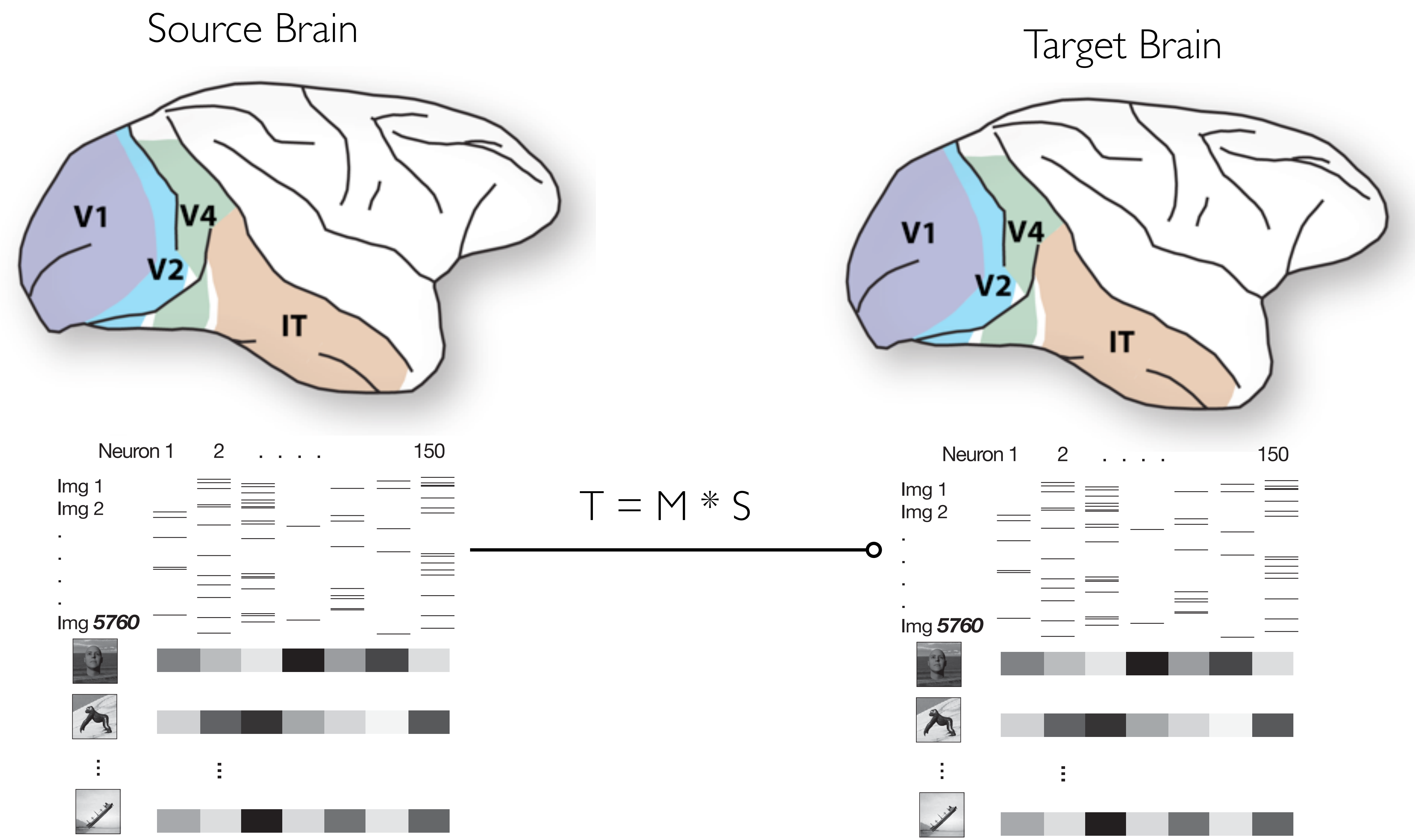
Neural Response Prediction

Some kind of mapping is necessary.



Neural Response Prediction

Here, we use linear regression.



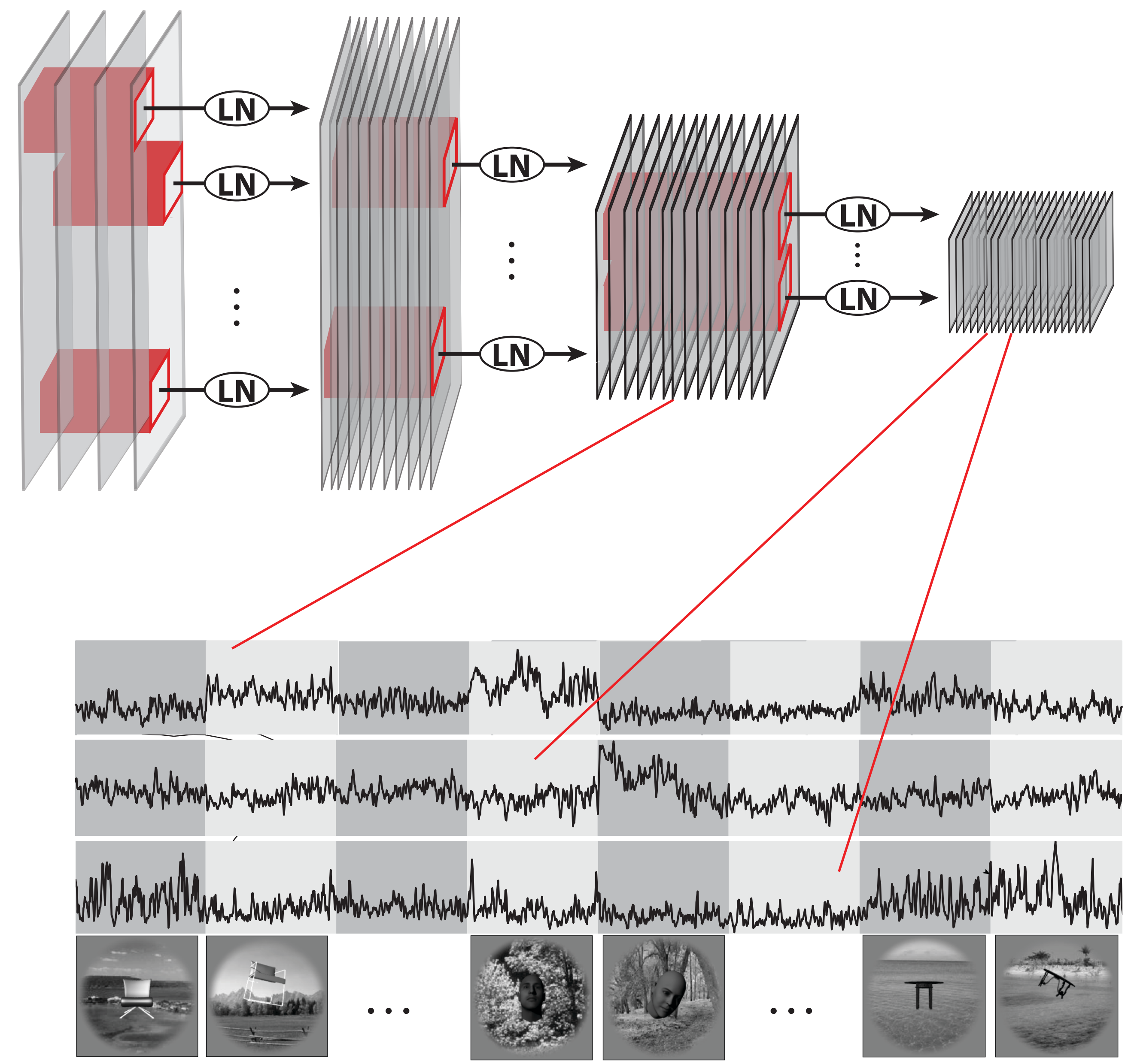
Neural predictivity: the ability of model to predict each individual neural site's activity.

Neural site unit ~ sparse linear combination of model units

Linear regression with fixed training images.

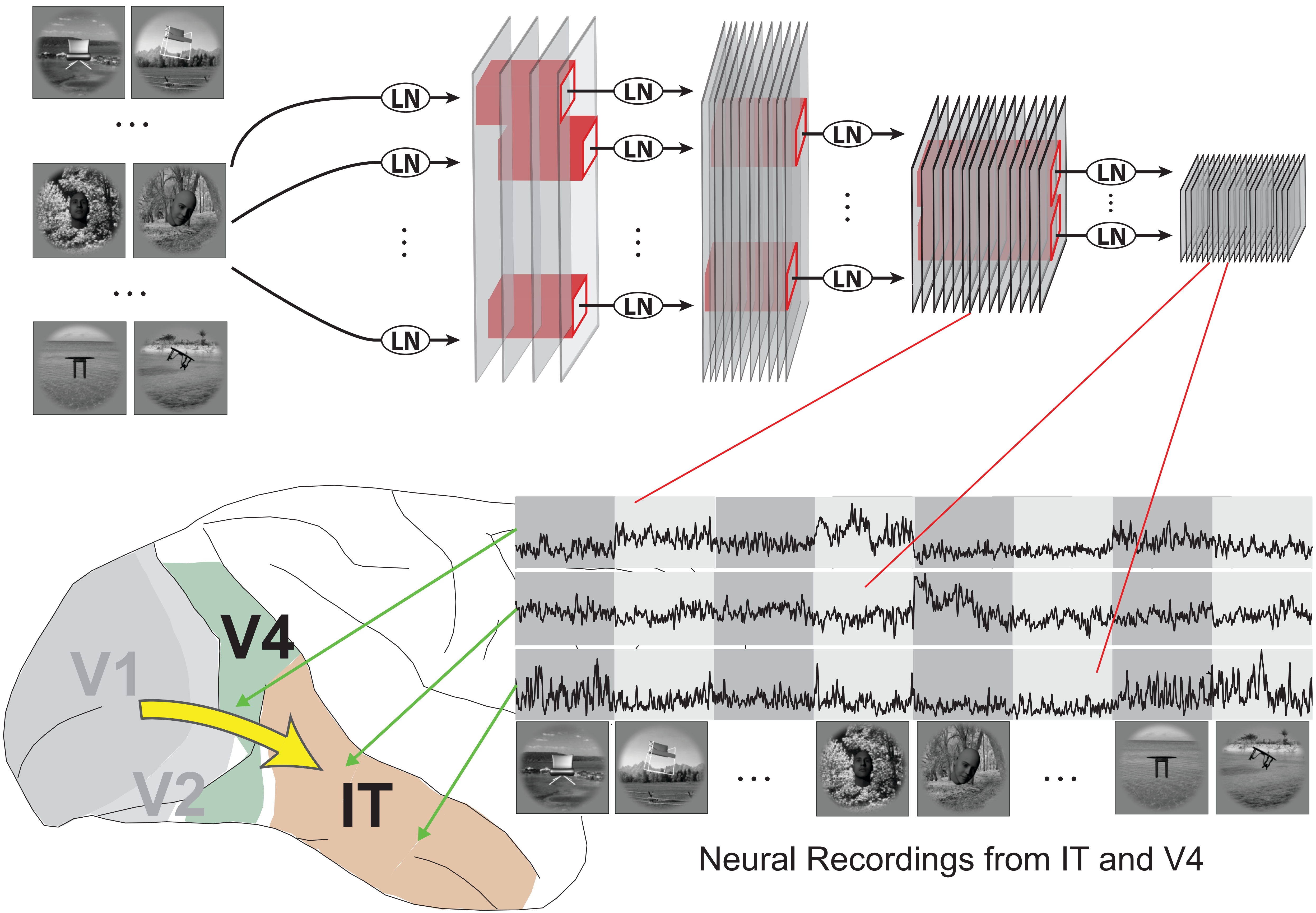
Accuracy = goodness-of-fit on held-out testing images (Cross validated)

Neural predictivity = median accuracy over all units.



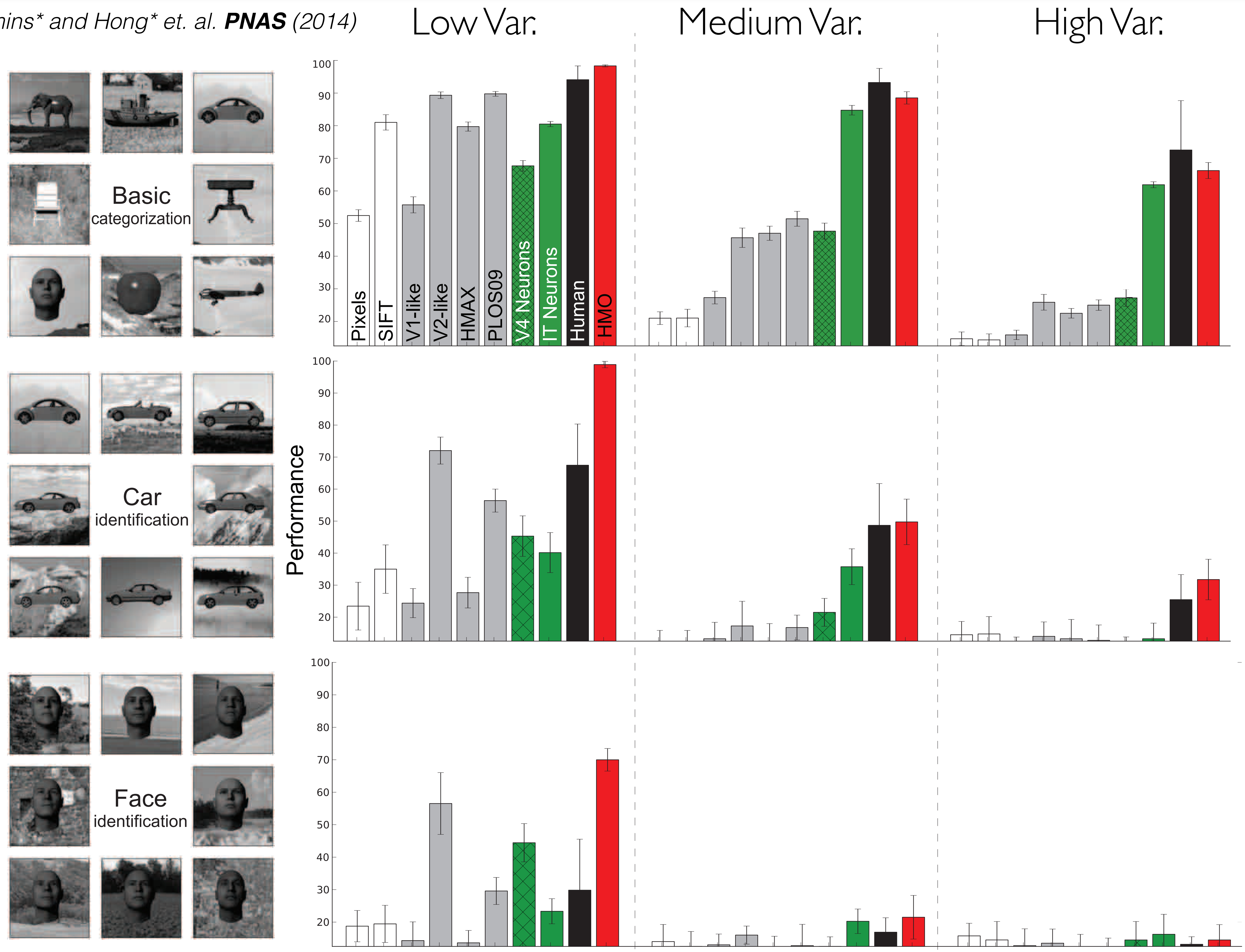
Neural Recordings from IT and V4

Neural predictivity: the ability of model to predict each individual neural site's activity.



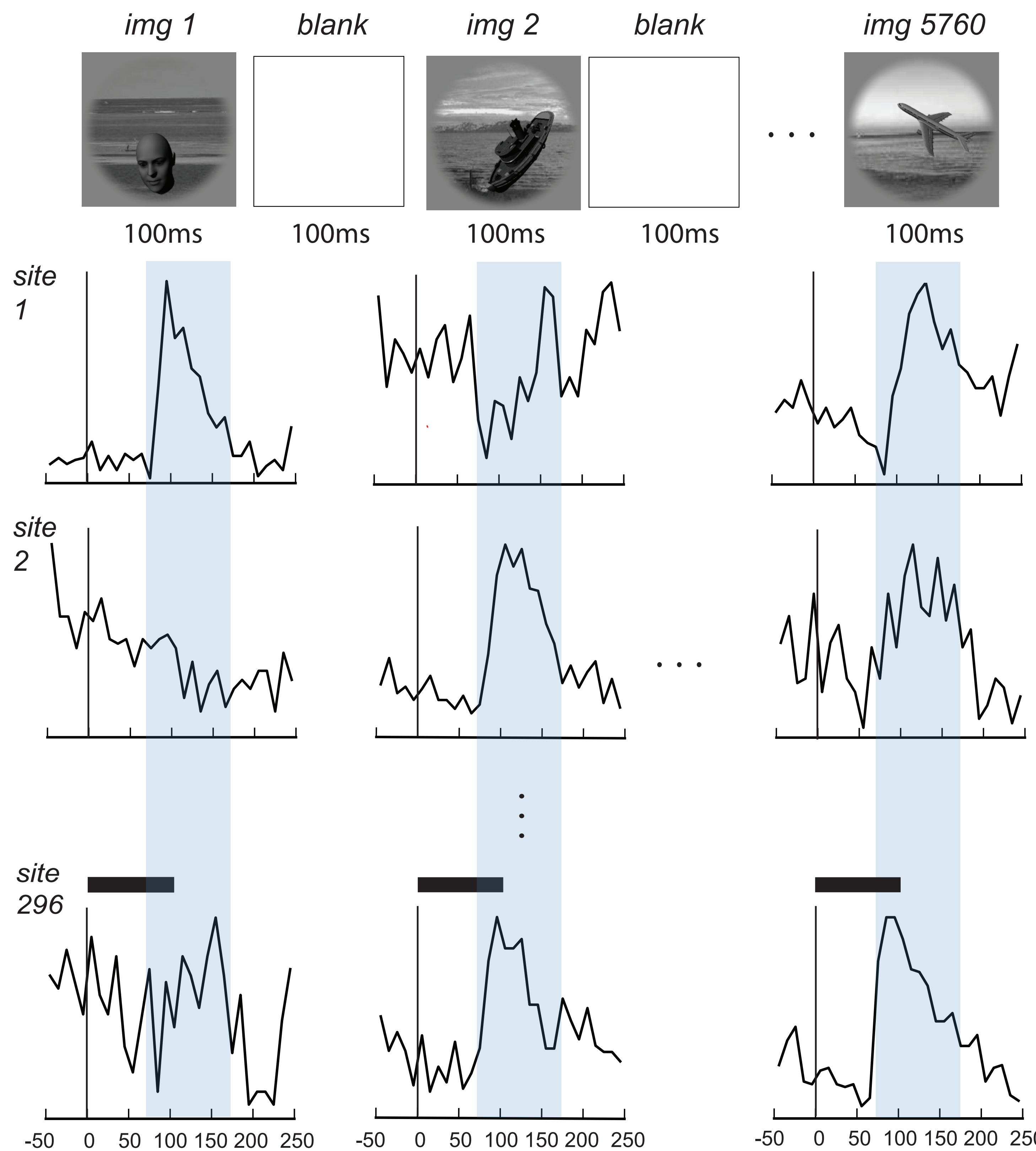
Performance Comparison

Yamins and Hong* et. al.* ***PNAS*** *(2014)*

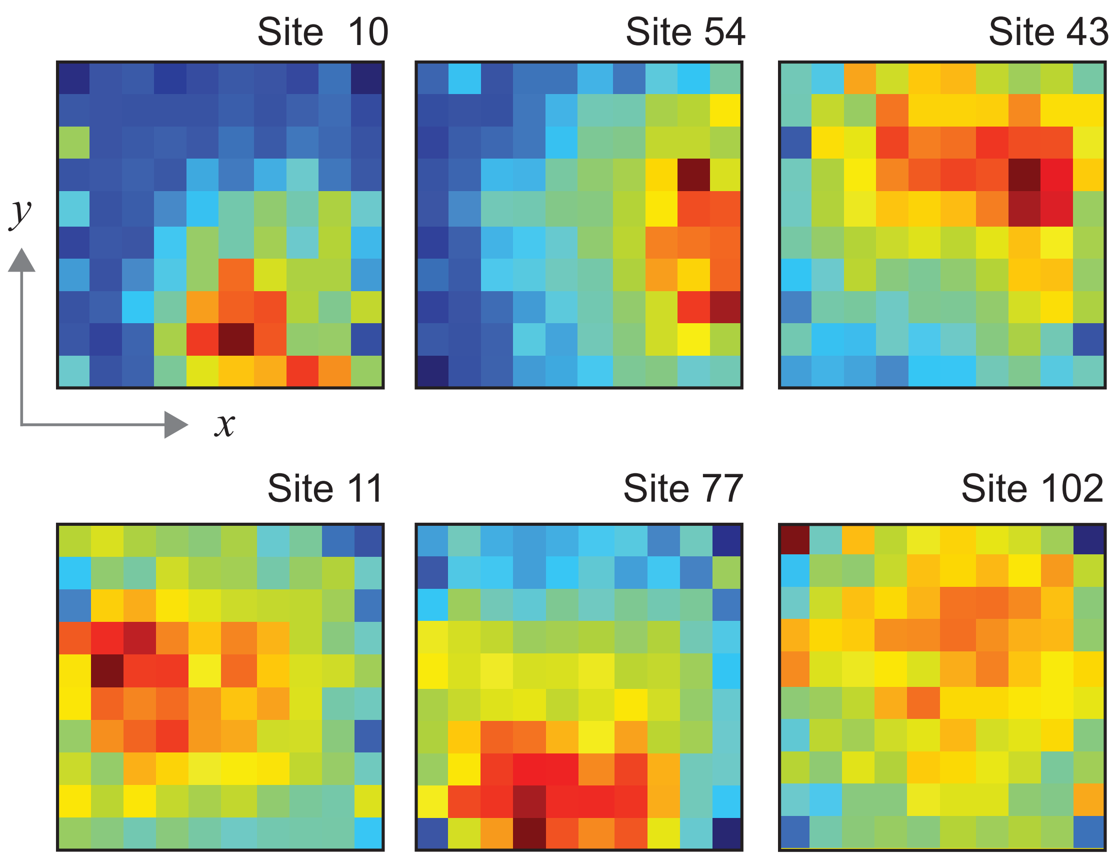


Neural Data Recording

Output = Binned spike counts in 70ms-170ms post stimulus presentation; averaged over 25-50 reps of each image.



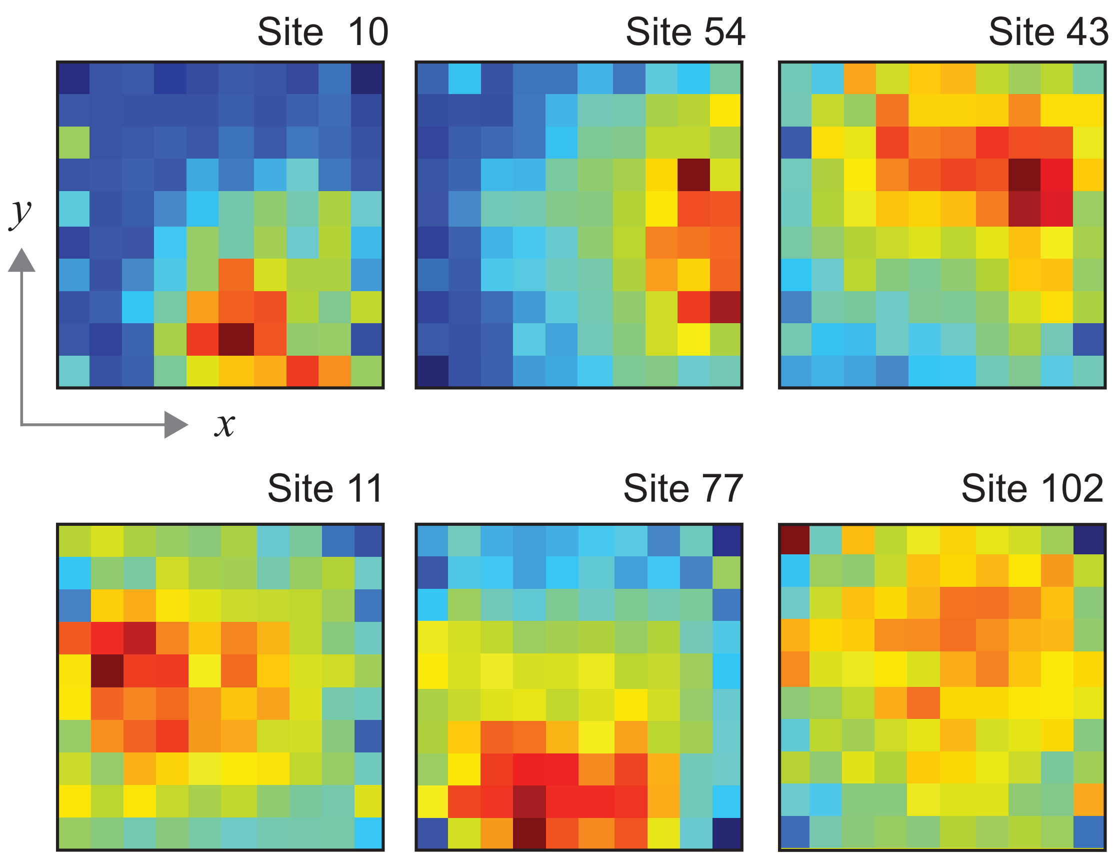
Single Site Responses



Best single position-encoding sites.

heat map value at x, y =
response averaged over all images where object center is in position x, y

Single Site Responses

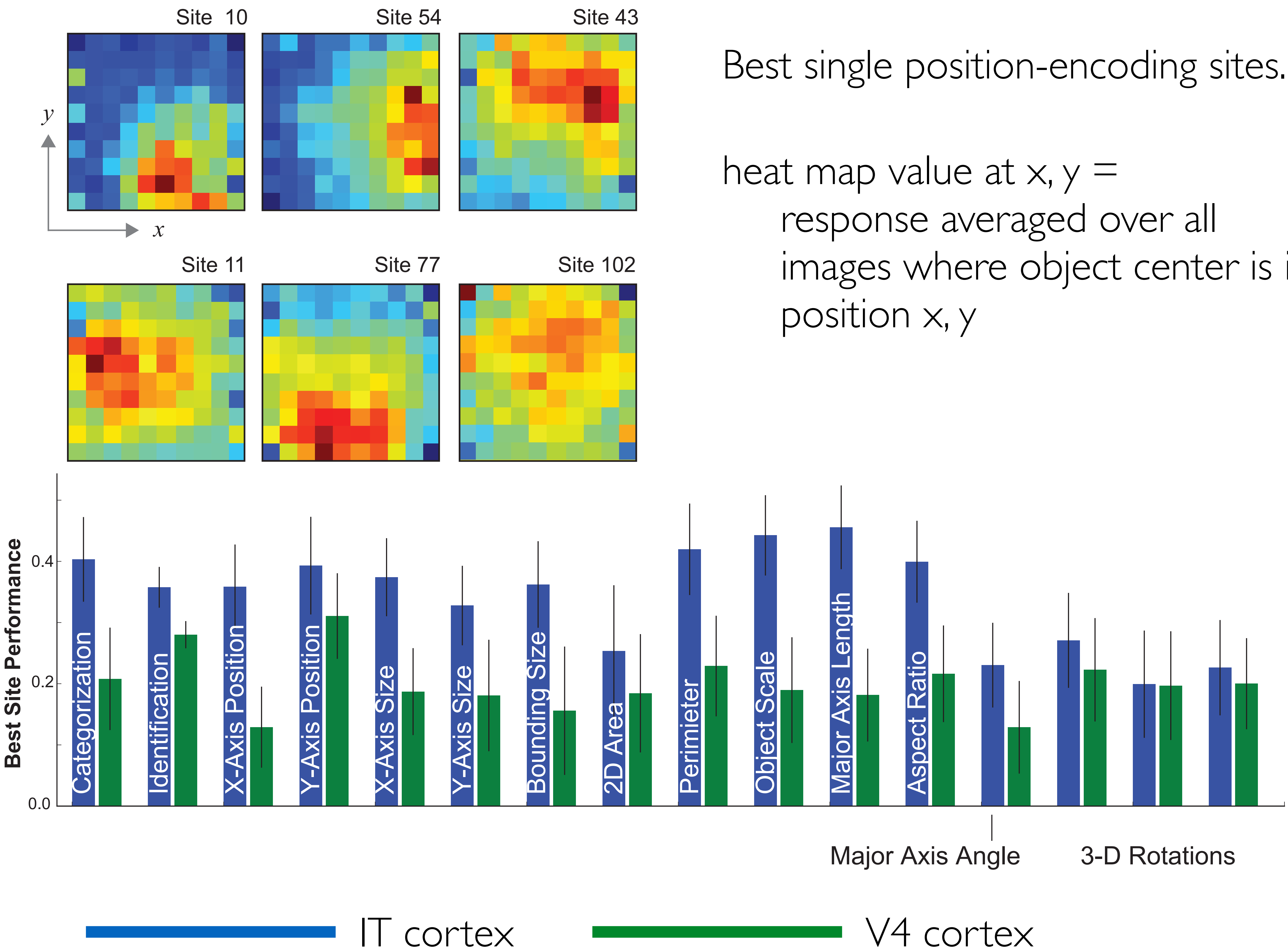


Best single position-encoding sites.

heat map value at x, y =
response averaged over all images where object center is in position x, y

Similar to MacEvoy (2013) and DiCarlo(2003) except — dramatically more variation.

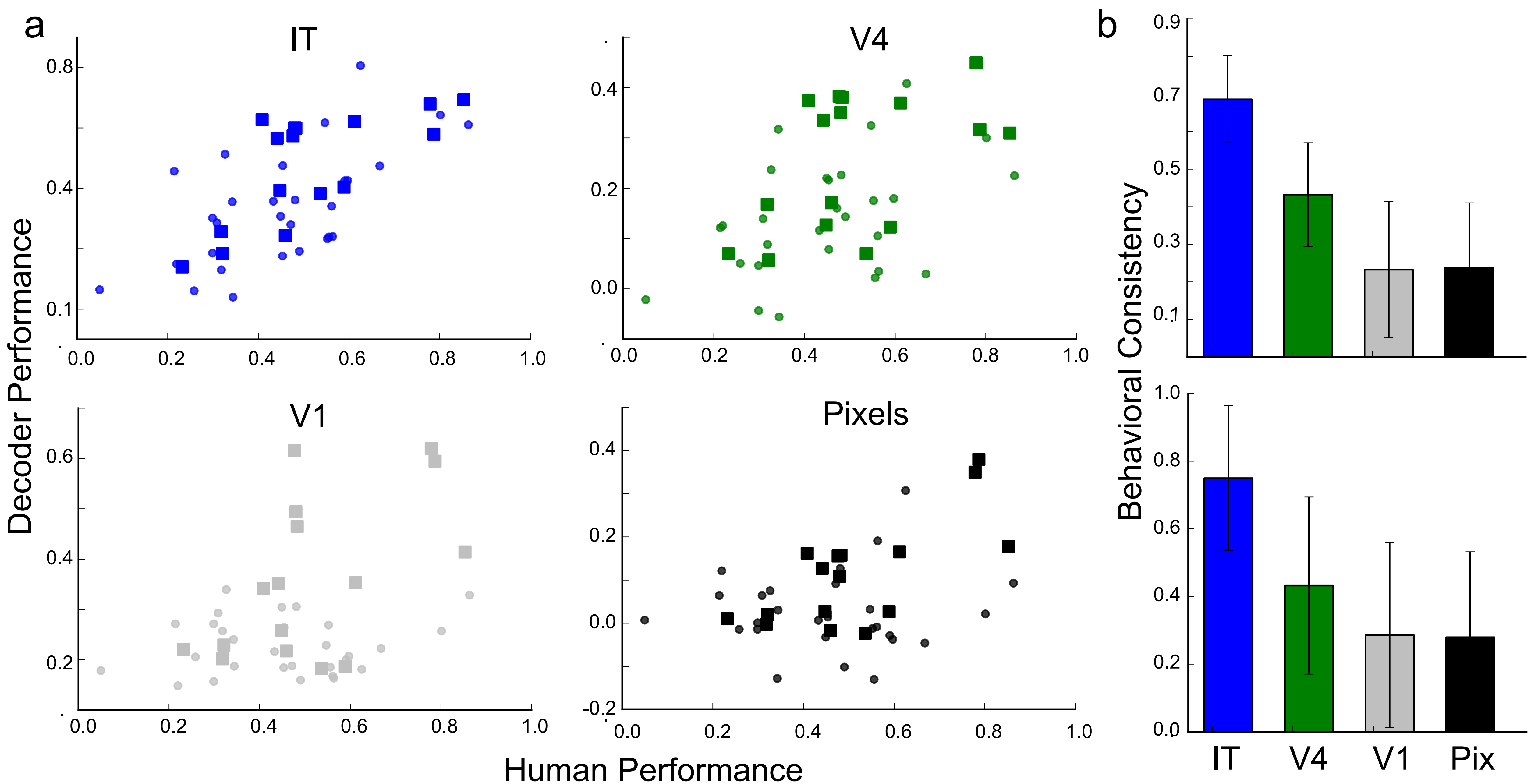
Single Site Responses



Best single position-encoding sites.

heat map value at x, y =
response averaged over all images where object center is in position x, y

Monkey Neurons vs Humans



Monkey Neurons vs Humans

	IT	V4	V1	Pix
Basic Categorization	773 ± 185	2.2×10^6	—	—
Subordinate Identification	496 ± 93	4.4×10^6	—	—
X-axis Position	1414 ± 403	5.2×10^5	3.0×10^7	—
Y-axis Position	918 ± 309	2.5×10^4	8.7×10^6	—
Bounding Box Size	322 ± 90	1.7×10^4	—	—
X-axis Size	256 ± 87	9.8×10^3	3.4×10^7	—
Y-axis Size	237 ± 87	3.8×10^3	9.5×10^6	—
3-D Object Scale	401 ± 90	3.2×10^4	—	—
Major Axis Length	201 ± 70	1.1×10^4	—	—
Aspect Ratio	163 ± 61	951 ± 59	6.5×10^3	—
Major Axis Angle	804 ± 136	3.2×10^6	—	—
Z-axis Rotation	1932 ± 1061	—	—	—
Y-axis Rotation	369 ± 115	2.8×10^5	—	—
X-axis Rotation	1570 ± 530	—	—	—

— = more than 10 billion sites required

Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(in press)*

Mean over tasks, human-parity for IT is at ~**700** multi-unit trial-averaged sites.

Monkey Neurons vs Humans

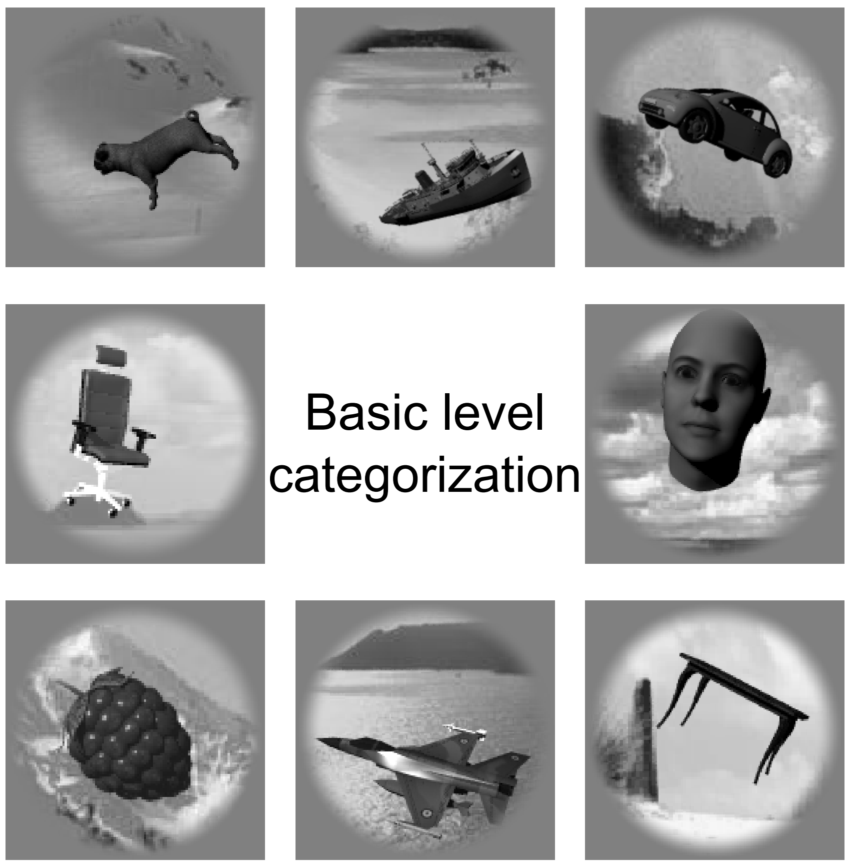
	IT	V4	V1	Pix
Basic Categorization	773 ± 185	2.2×10^6	—	—
Subordinate Identification	496 ± 93	4.4×10^6	—	—
X-axis Position	1414 ± 403	5.2×10^5	3.0×10^7	—
Y-axis Position	918 ± 309	2.5×10^4	8.7×10^6	—
Bounding Box Size	322 ± 90	1.7×10^4	—	—
X-axis Size	256 ± 87	9.8×10^3	3.4×10^7	—
Y-axis Size	237 ± 87	3.8×10^3	9.5×10^6	—
3-D Object Scale	401 ± 90	3.2×10^4	—	—
Major Axis Length	201 ± 70	1.1×10^4	—	—
Aspect Ratio	163 ± 61	951 ± 59	6.5×10^3	—
Major Axis Angle	804 ± 136	3.2×10^6	—	—
Z-axis Rotation	1932 ± 1061	—	—	—
Y-axis Rotation	369 ± 115	2.8×10^5	—	—
X-axis Rotation	1570 ± 530	—	—	—

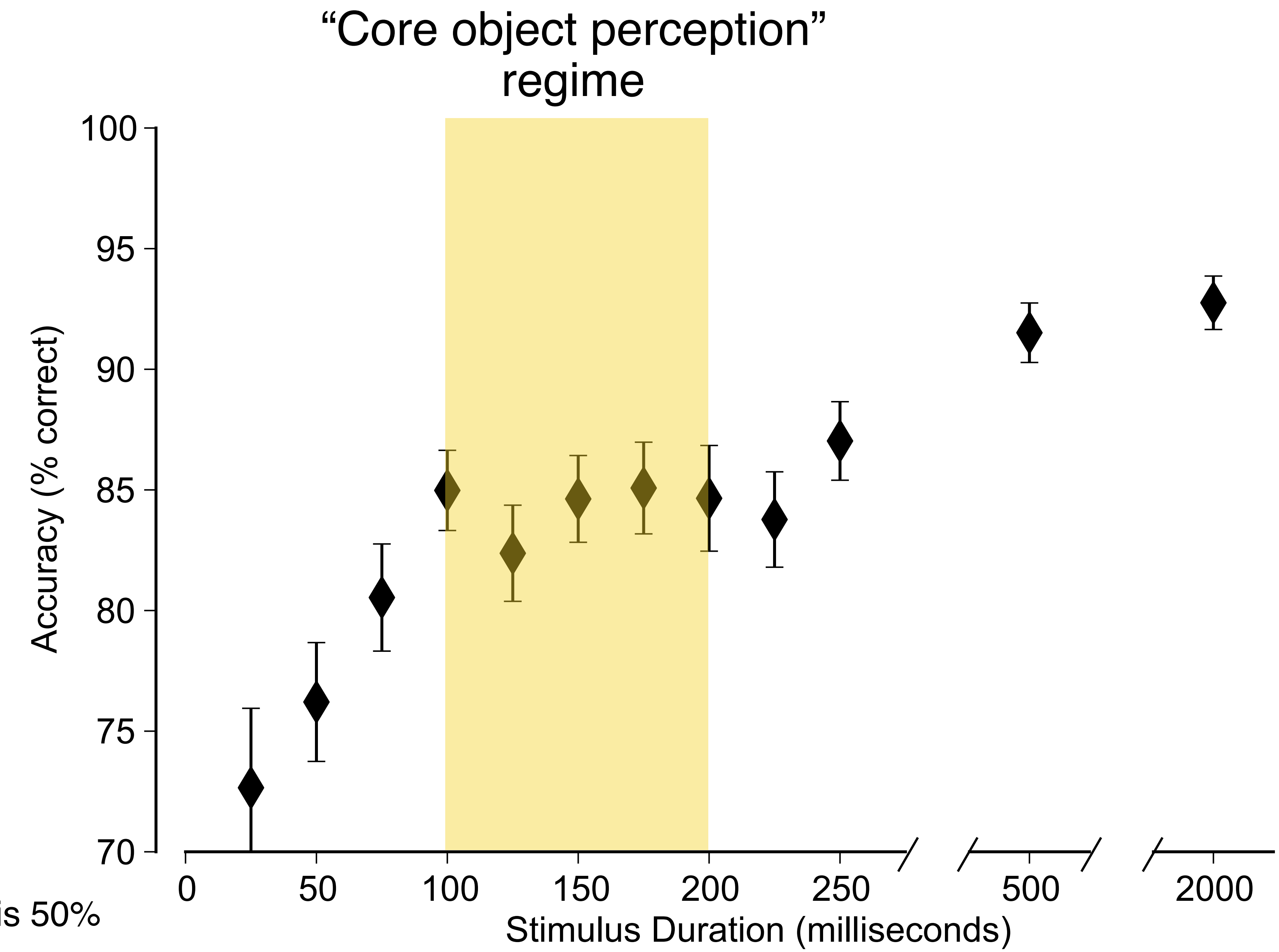
— = more than 10 billion sites required

Hong, Yamins*, Majaj & DiCarlo.* ***Nat. Neuro.*** *(in press)*

Mean over tasks, human-parity for IT is at ~**350000** single-unit single-trial neurons.

Example: Human object categorization accuracy as a function of image viewing time



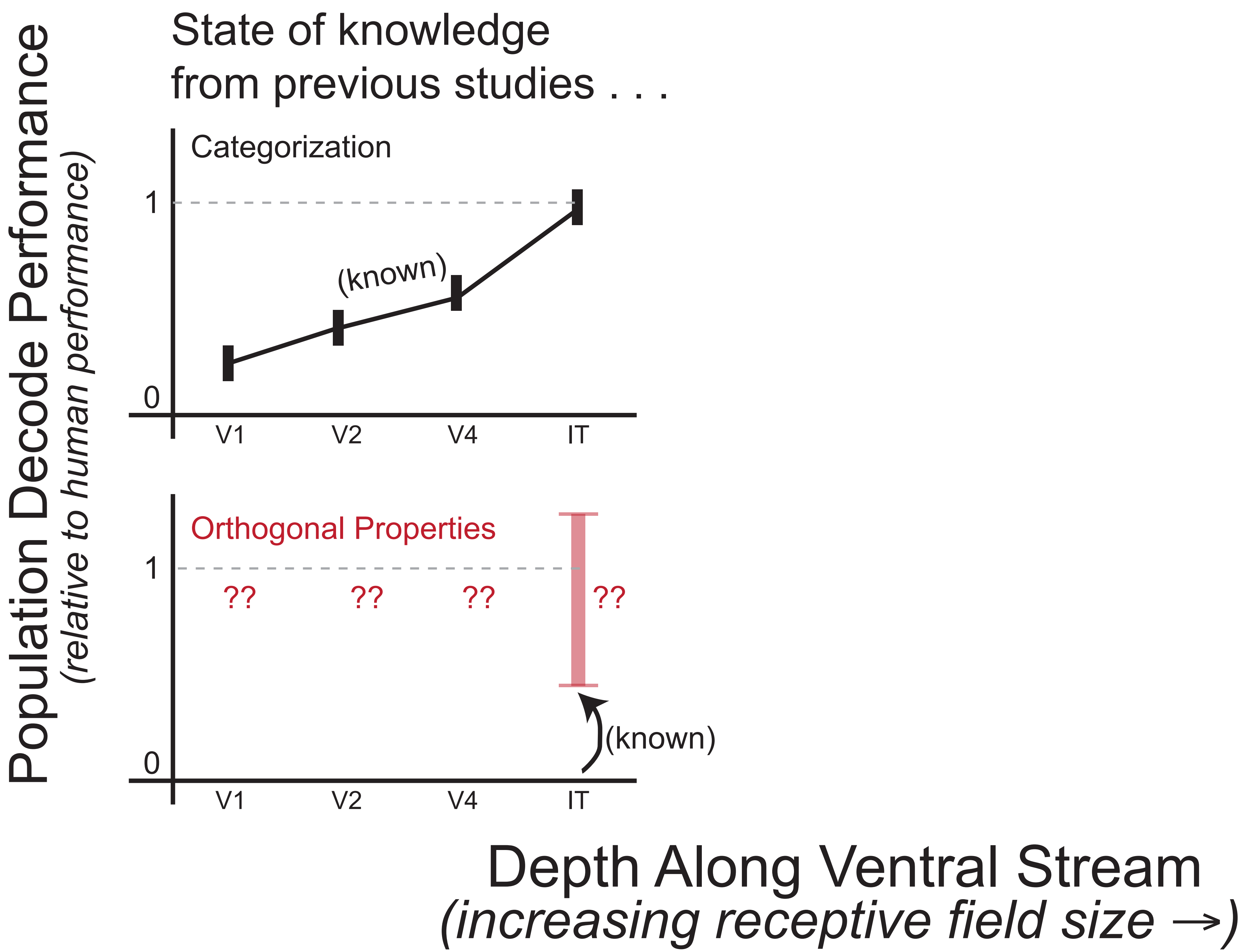


Chance is 50%

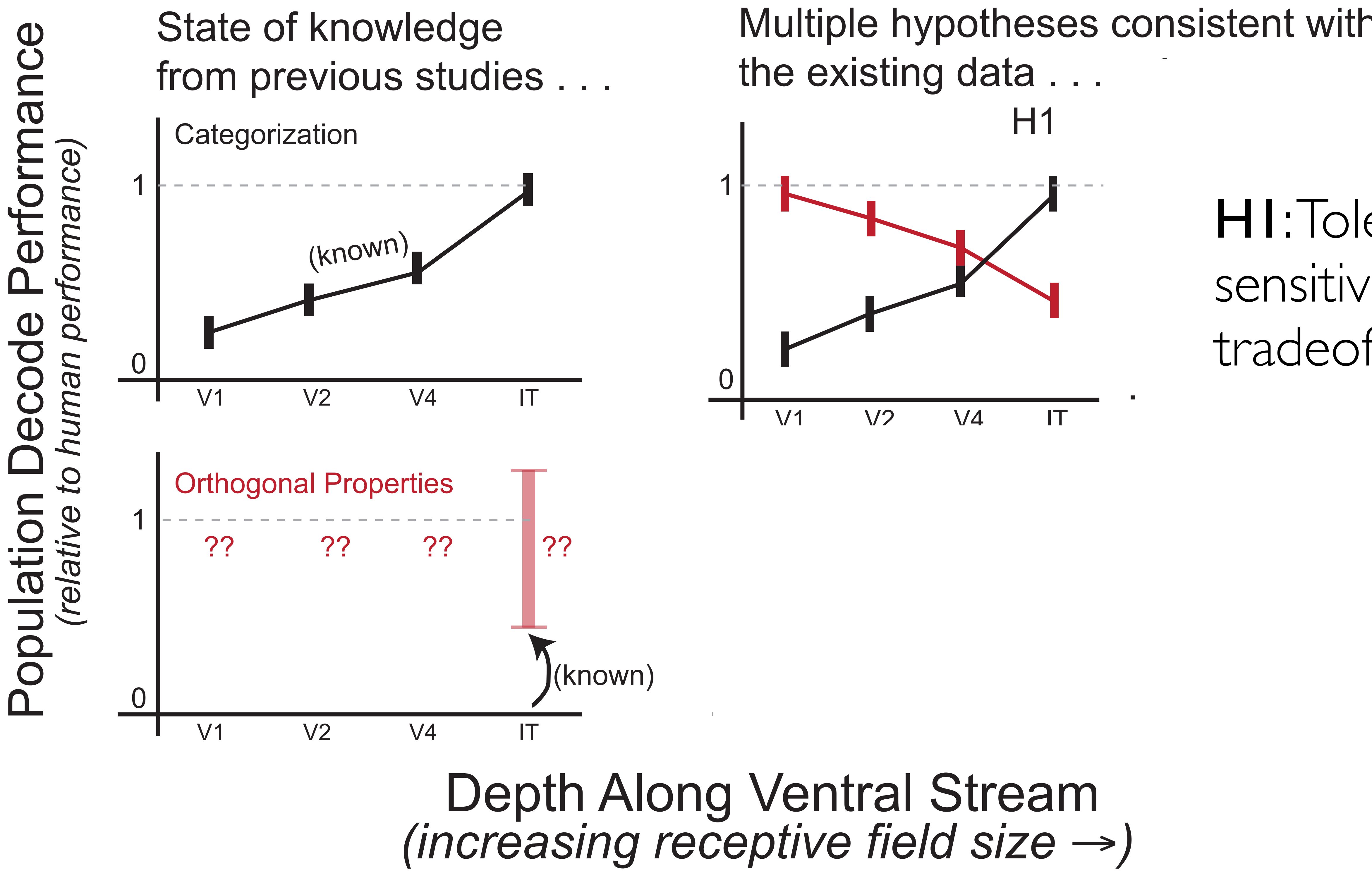
All the data I will show you today

Typical primate fixation duration during natural viewing

Somewhat newish ideas about IT?

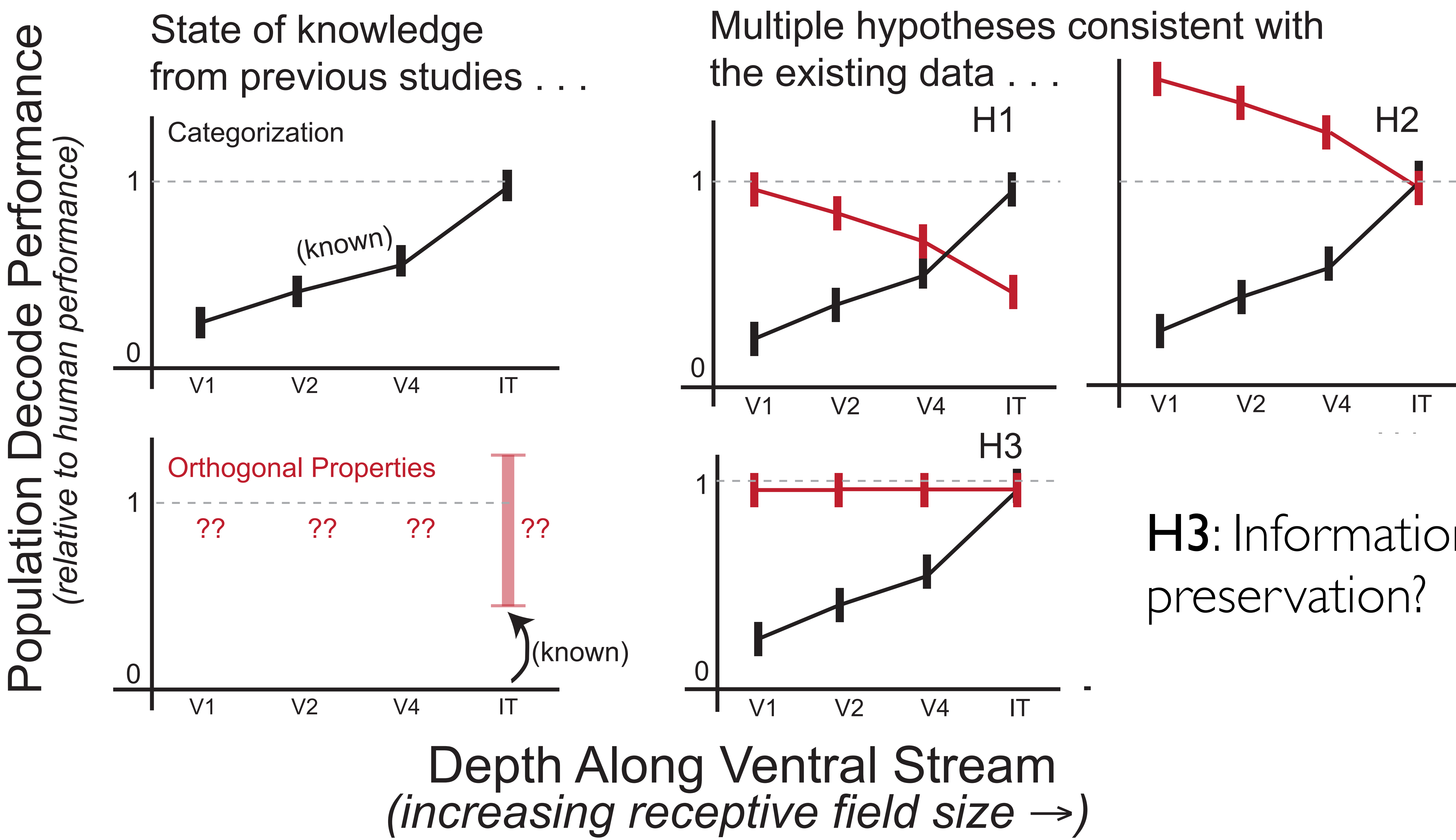


Somewhat newish ideas about IT?



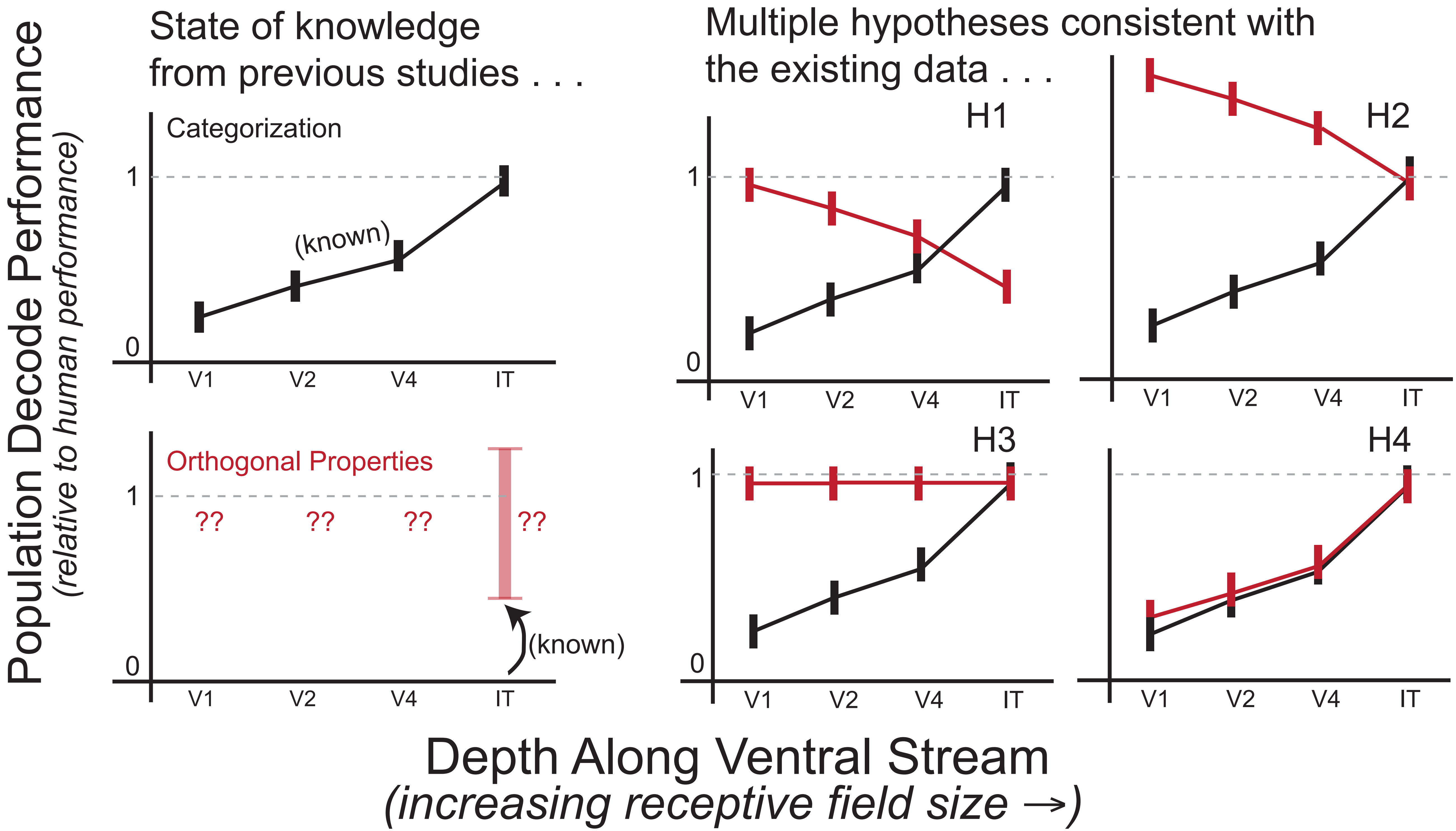
H1: Tolerance / sensitivity tradeoff?

Somewhat newish ideas about IT?

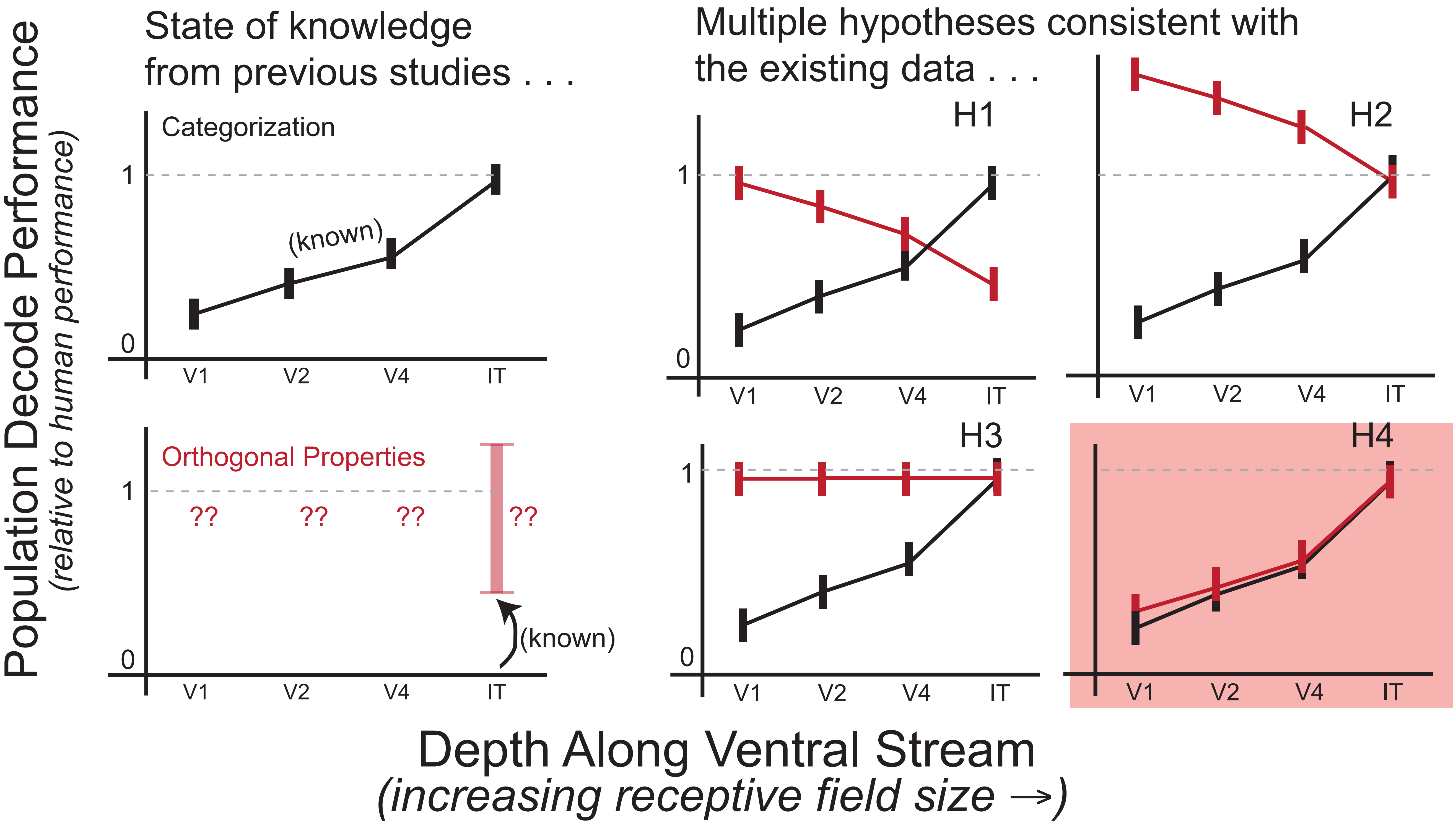


H3: Information preservation?

Somewhat newish ideas about IT?

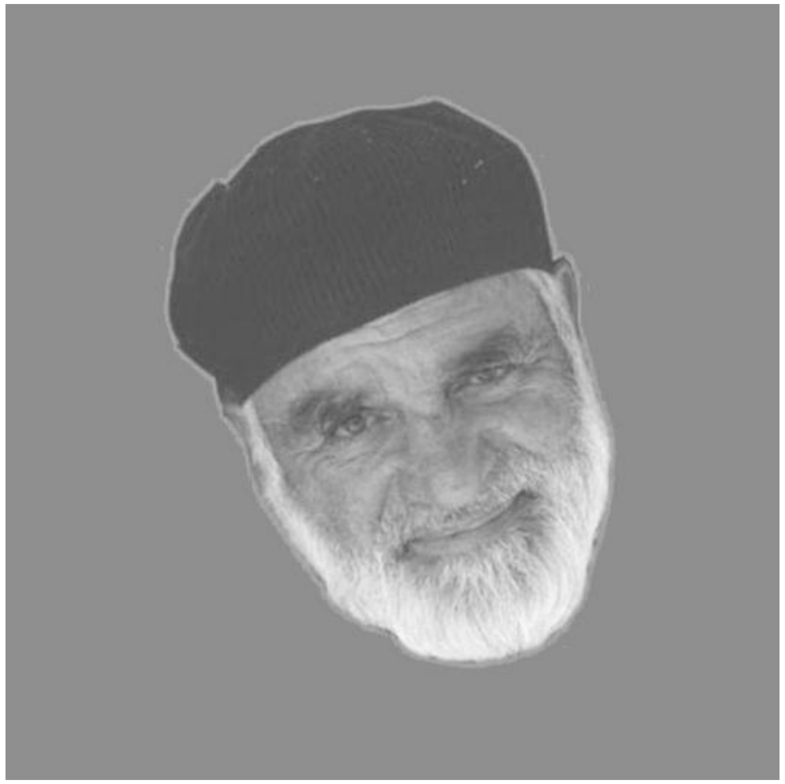


Somewhat newish ideas about IT?

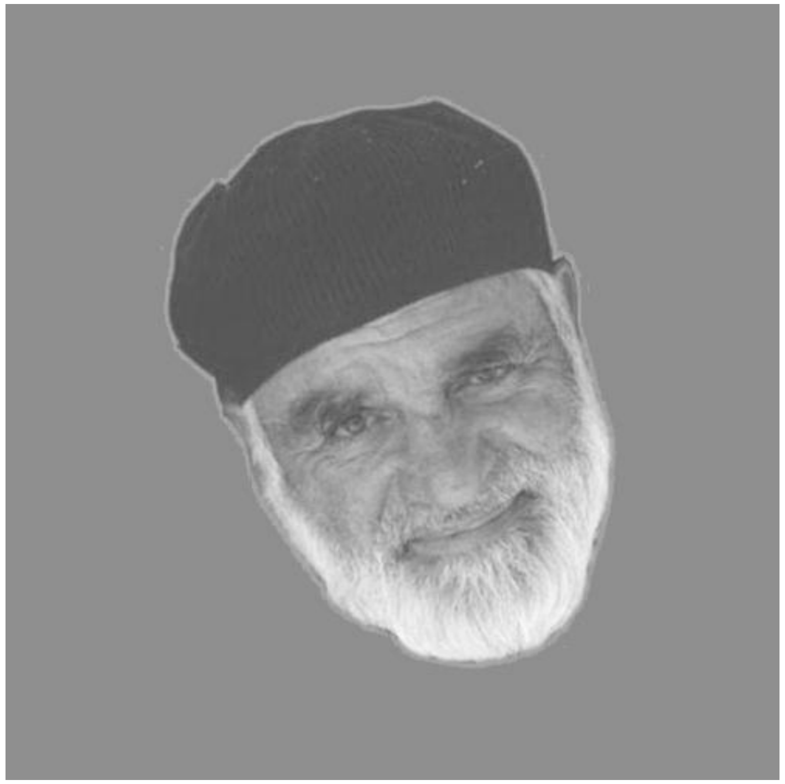


H4: Simultaneous build-up of encoding

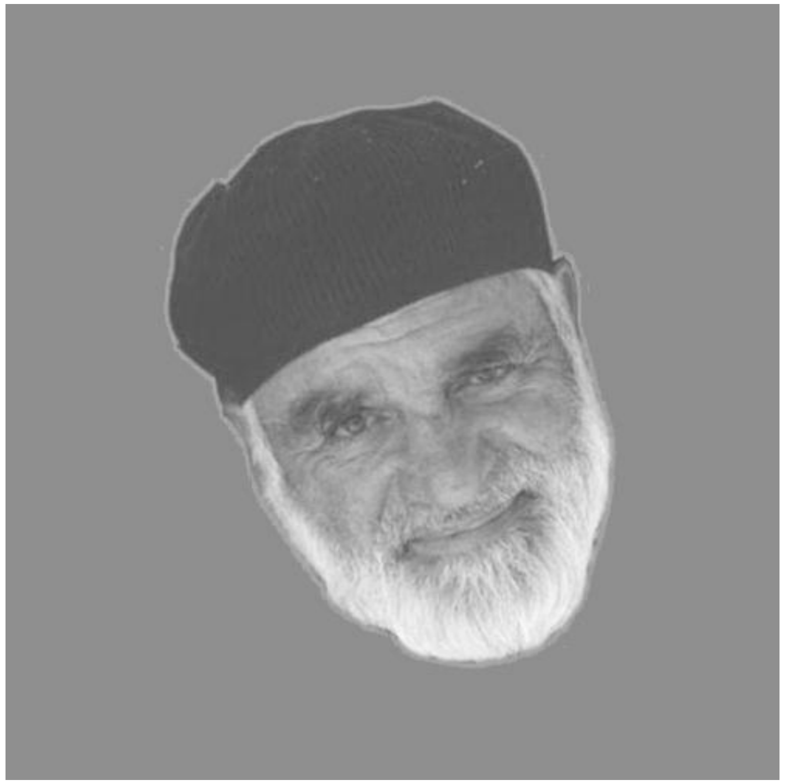
Face Perception is Fast, Robust, and Accurate

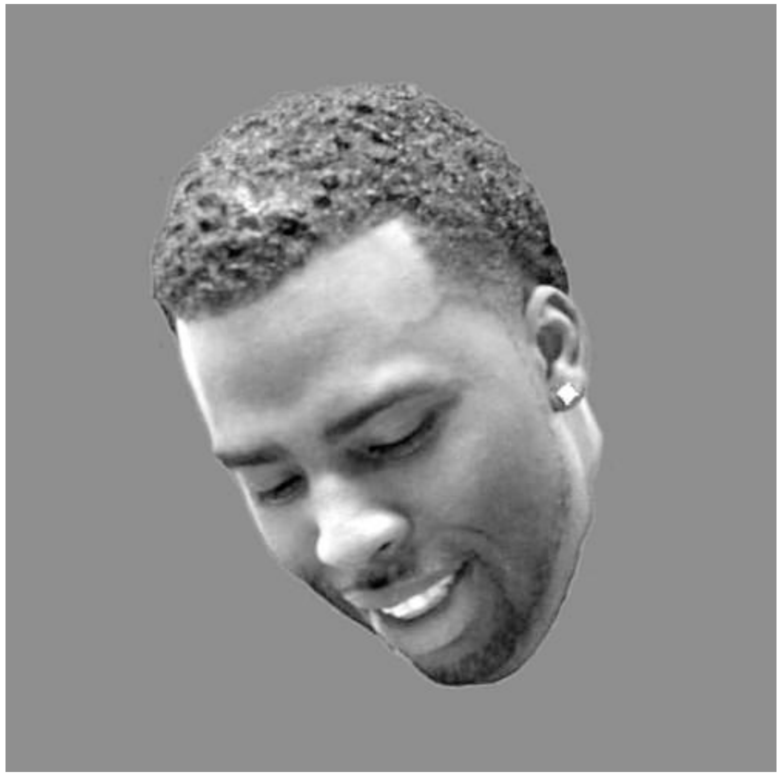


Face Perception is Fast, Robust, and Accurate

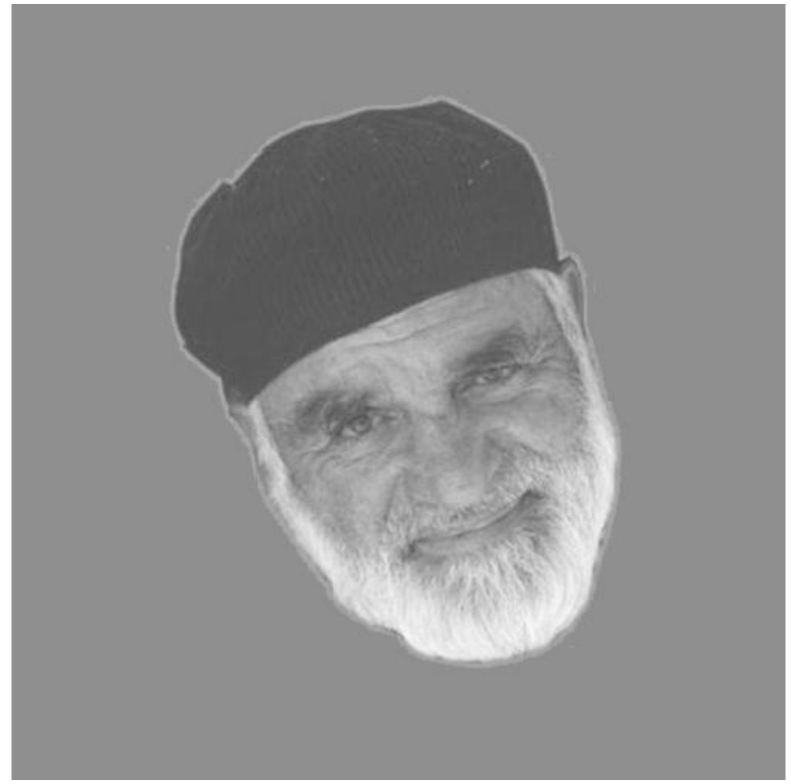


Face Perception is Fast, Robust, and Accurate

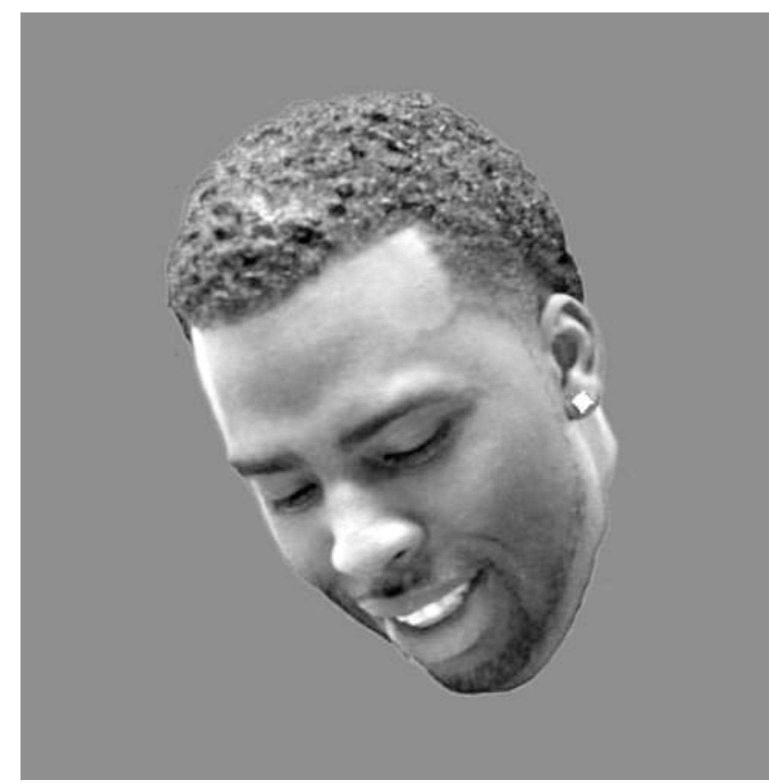




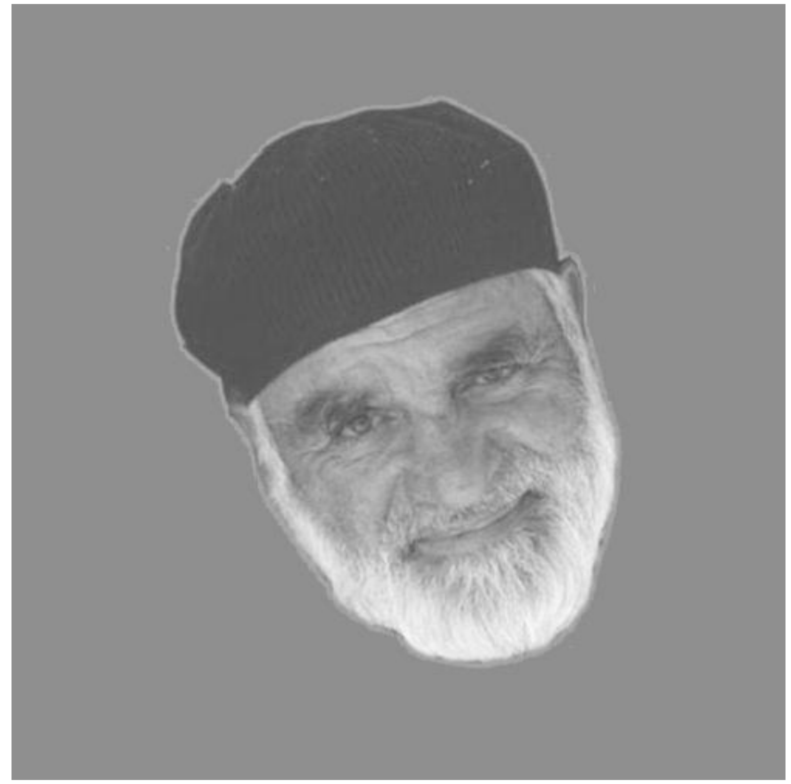
Face Perception is Fast, Robust, and Accurate

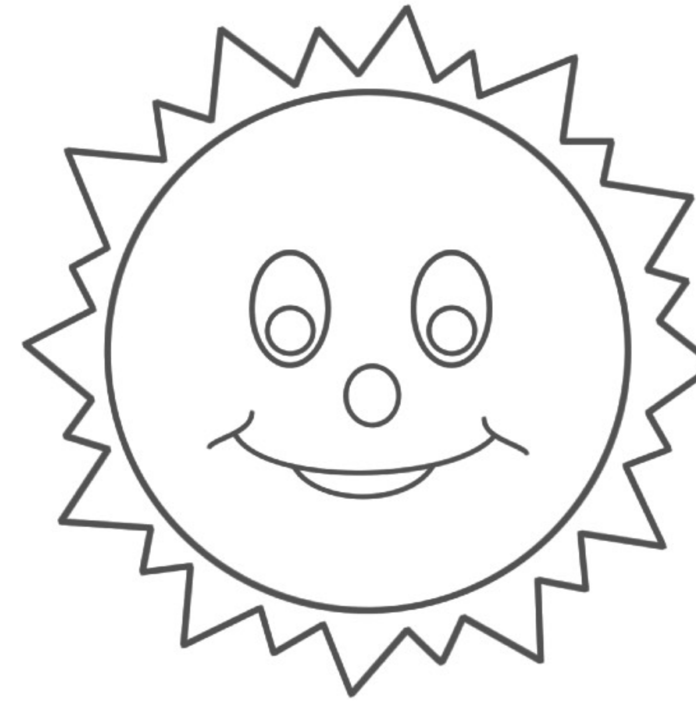



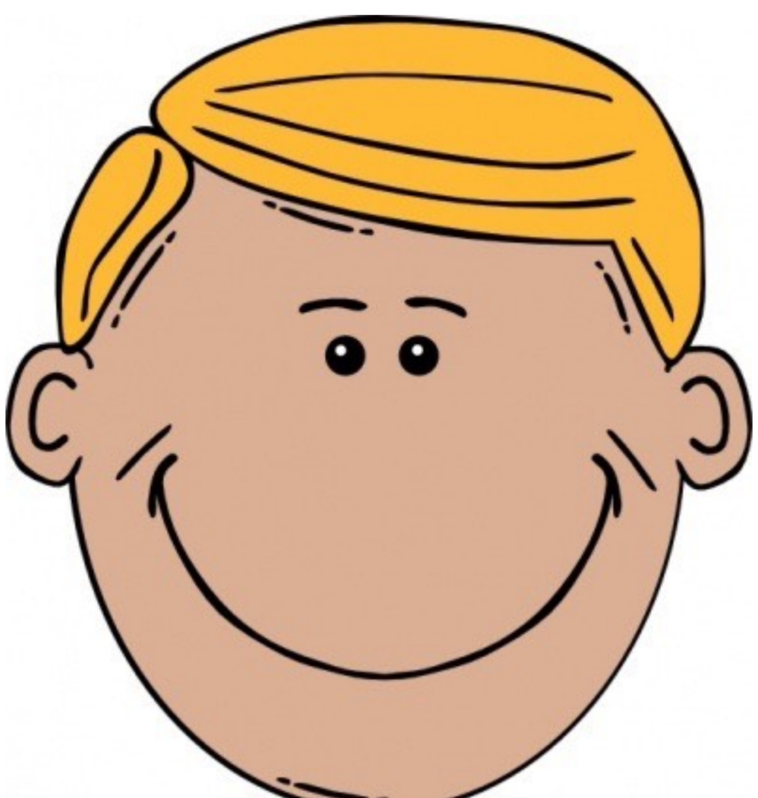




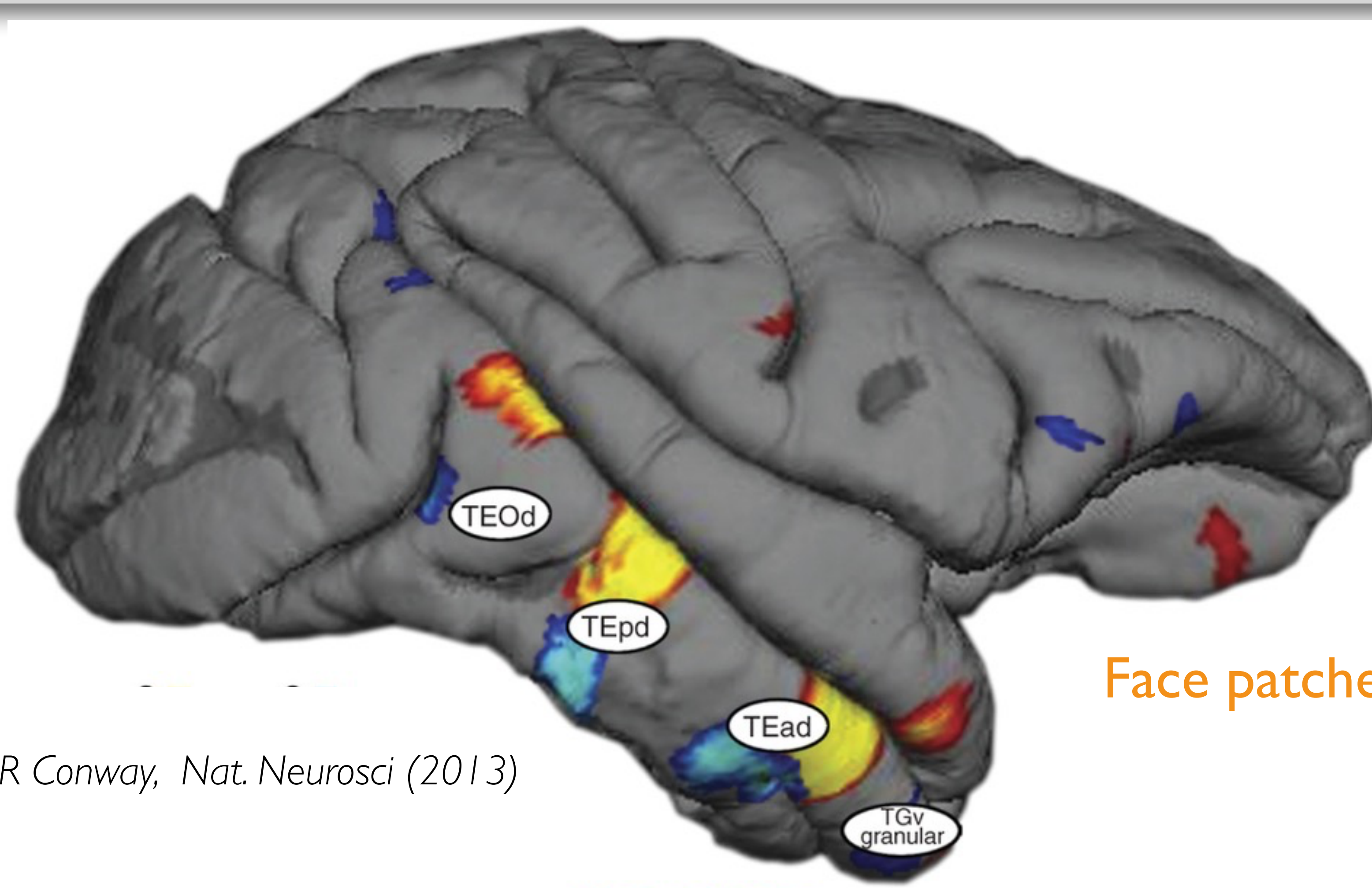
Face Perception is Fast, Robust, and Accurate







Selective Patches in Higher Visual Cortex



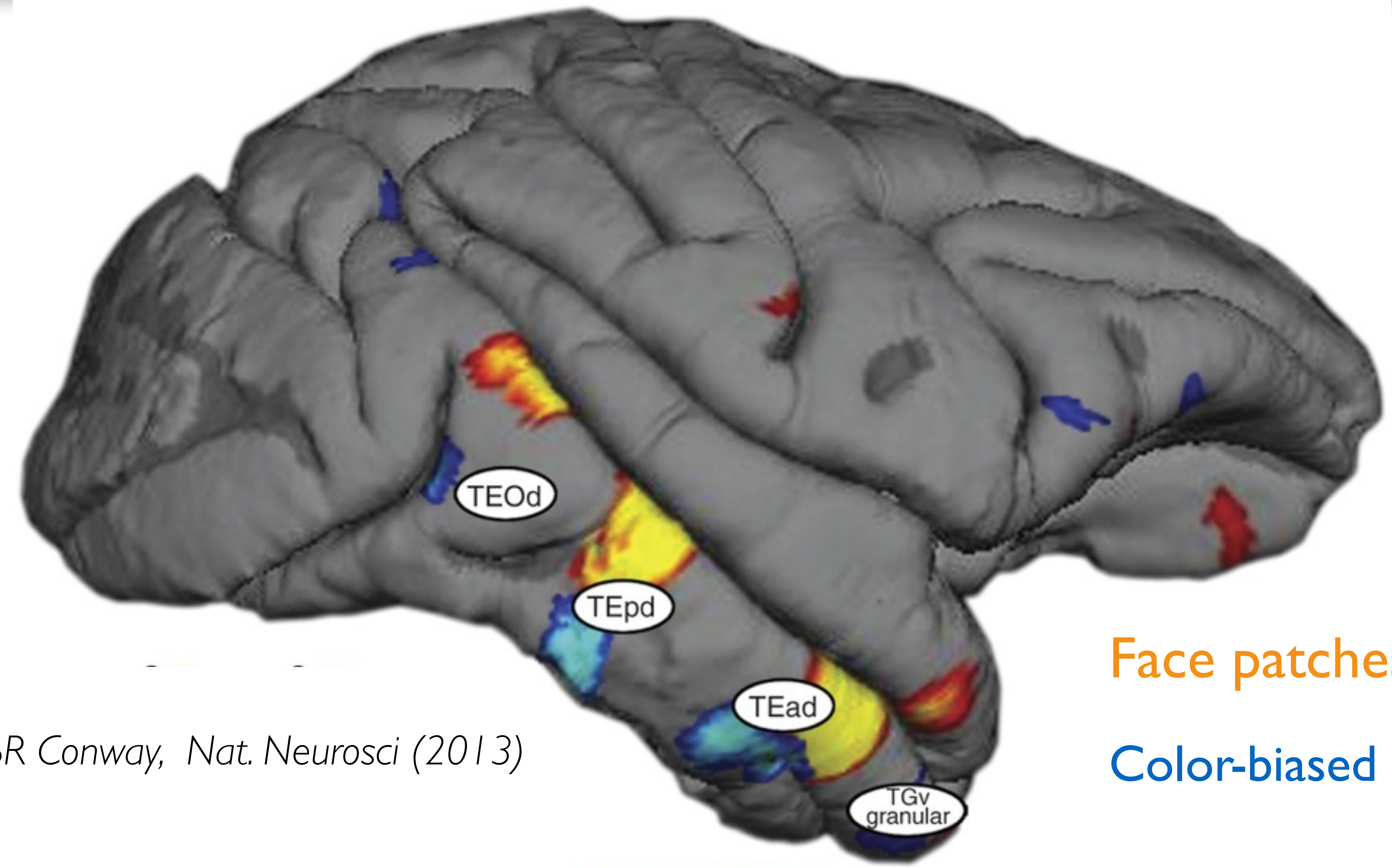
Face patches

R. Lafer-Sousa and BR Conway, Nat. Neurosci (2013)

Regions selective for:

- faces

Selective Patches in Higher Visual Cortex



R. Lafer-Sousa and BR Conway, Nat. Neurosci (2013)

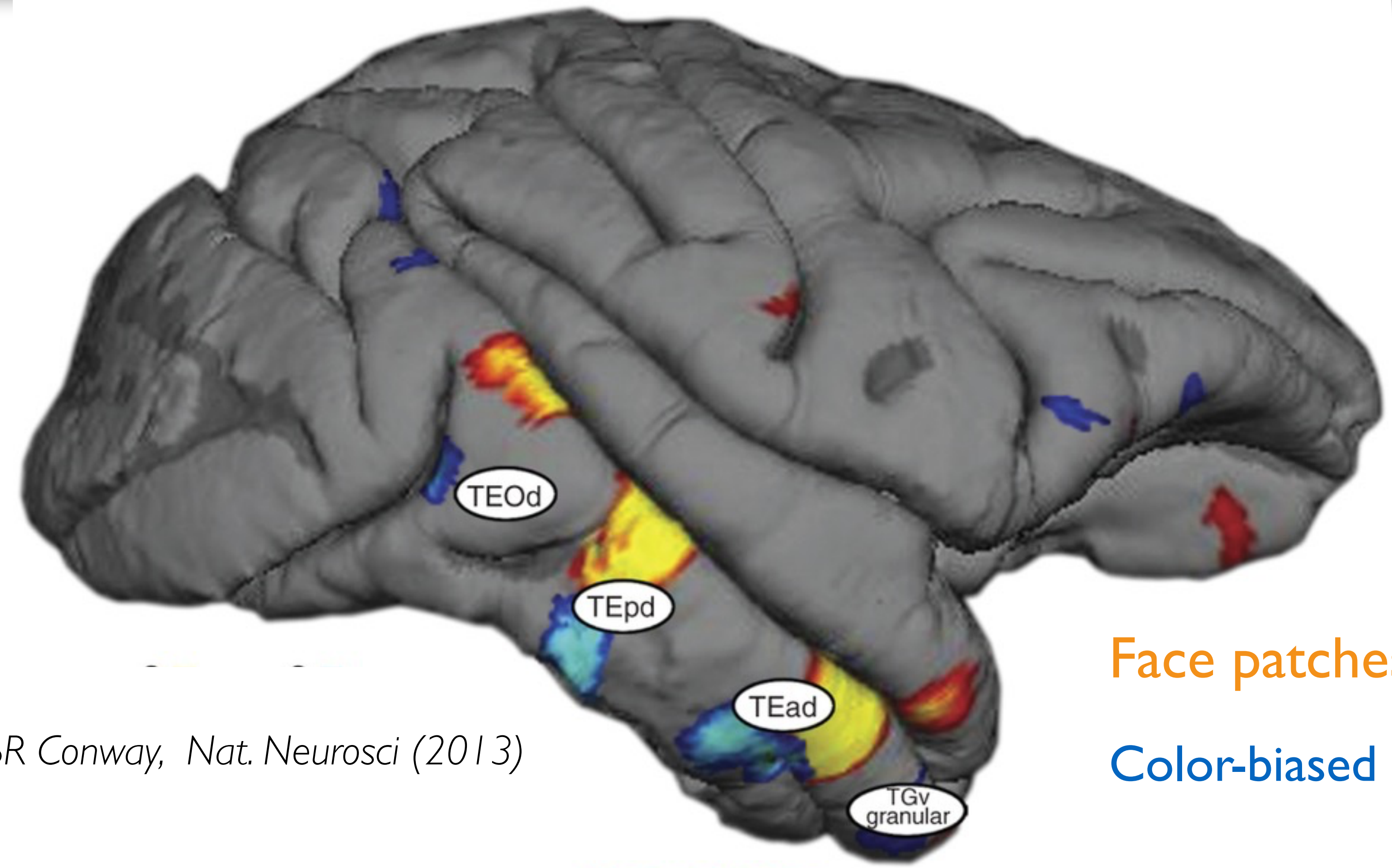
Face patches

Color-biased regions

Regions selective for:

- faces
- places
- bodies
- color

Selective Patches in Higher Visual Cortex



Face patches

Color-biased regions

R. Lafer-Sousa and BR Conway, Nat. Neurosci (2013)

Regions selective for:

- faces
- places
- bodies
- color

Where do these patches come from?

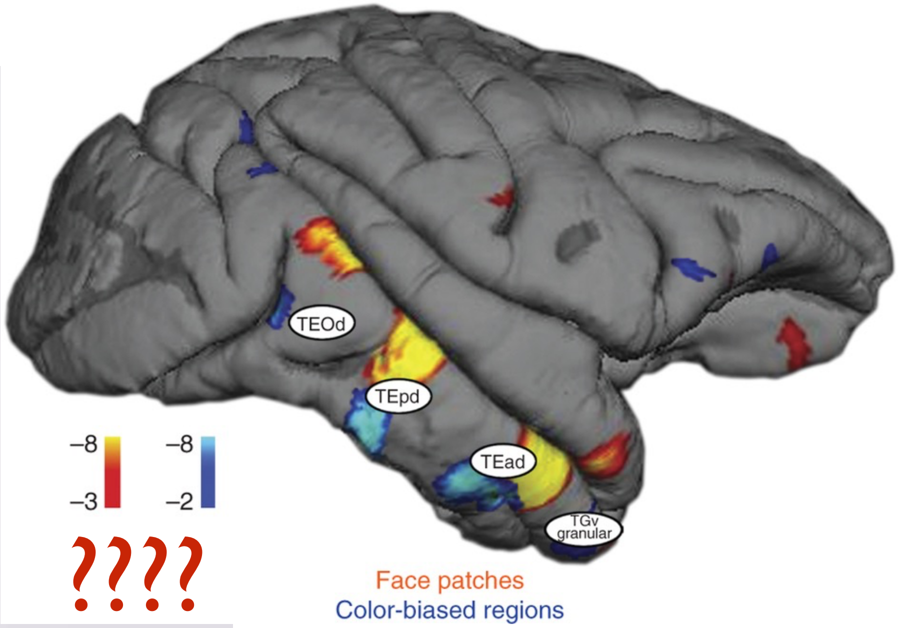
- In-born built-in structure??
- or developmentally determined by domain-specific experience?

Selective Patches in Higher Visual Cortex

controlled rearing

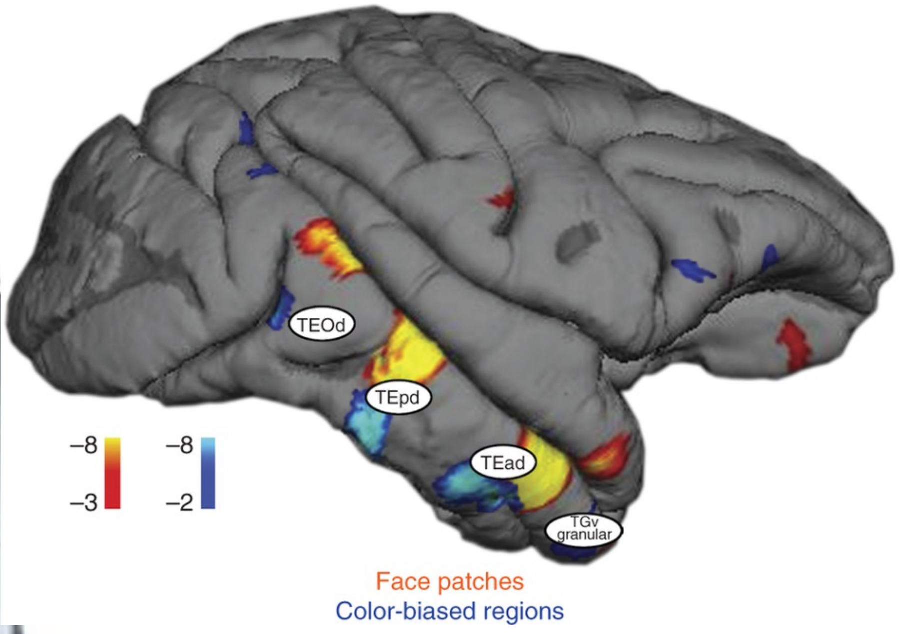
Selective Patches in Higher Visual Cortex

controlled rearing



Selective Patches in Higher Visual Cortex

controlled rearing

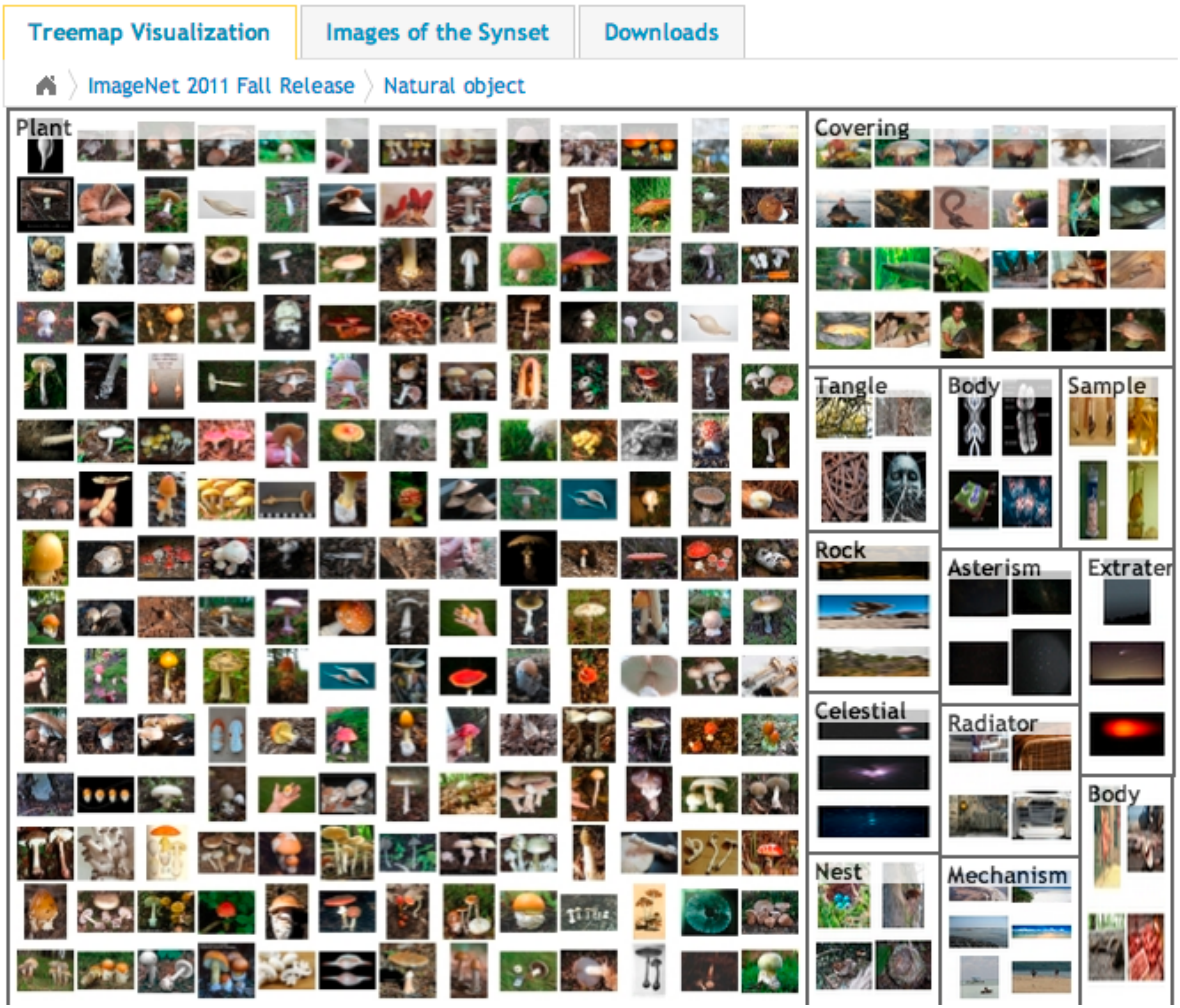


????

… in a computational model

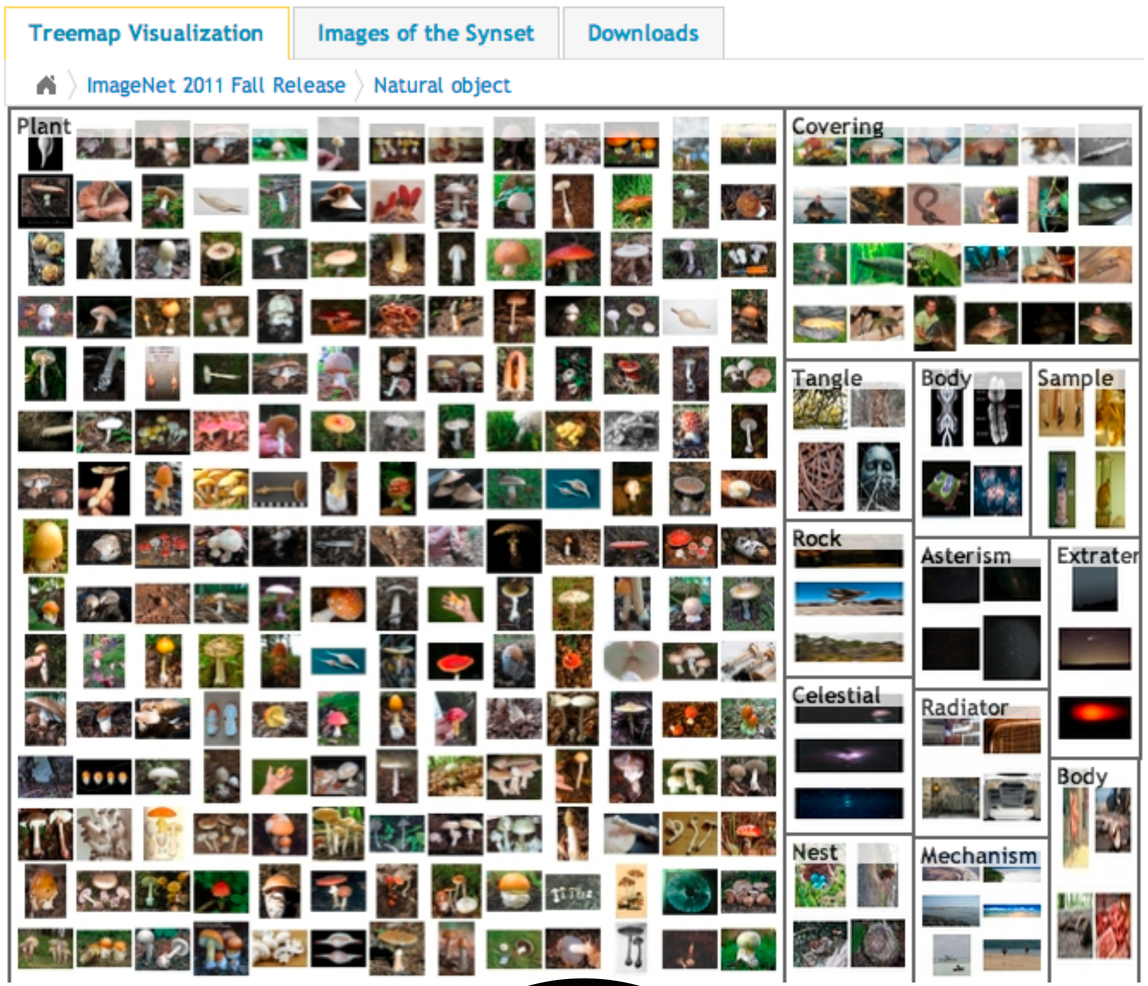
Model Training Regimen

ImageNet (2012). Thousands of images in thousands of categories.



Model Training Regimen

controlled rearing



remove all images containing faces
as well as all categories of photos containing animate objects

question: how does removing this content affect the model?

revolver

cheeseburger

tiger

german shepherd

bolete

sycamore

laptop

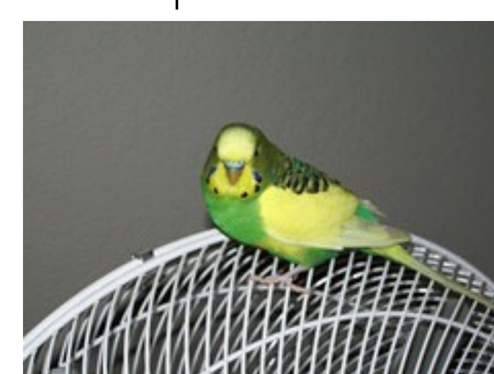
beeker

television

catamaran

thatched roof

parakeet



revolver

cheeseburger

tiger

german shepherd

bolete

sycamore

laptop

beeker

television

catamaran

thatched roof

parakeet

revolver

german shepherd

laptop

catamaran

cheeseburger

bolete

beeker

thatched roof

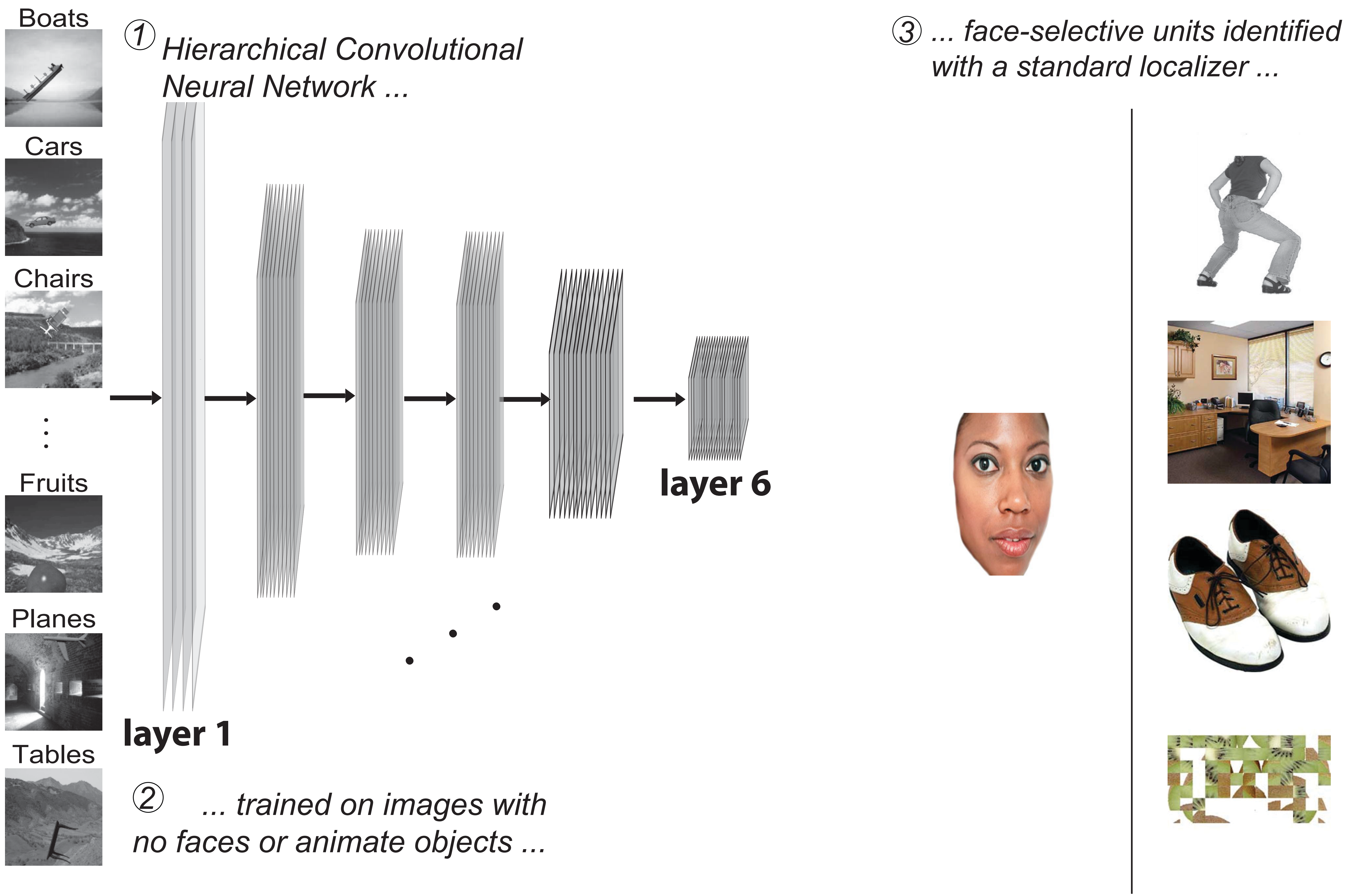
tiger

sycamore

television

parakeet

Testing the base-line non-face model.



Testing the base-line non-face model.

Original speculation: we won't find any (or statistically significantly many) face selective units because:

Hypotheses for existing of face-selective units:

~~i. face processing machinery is in-born~~

Testing the base-line non-face model.

Original speculation: we won't find any (or statistically significantly many) face selective units because:

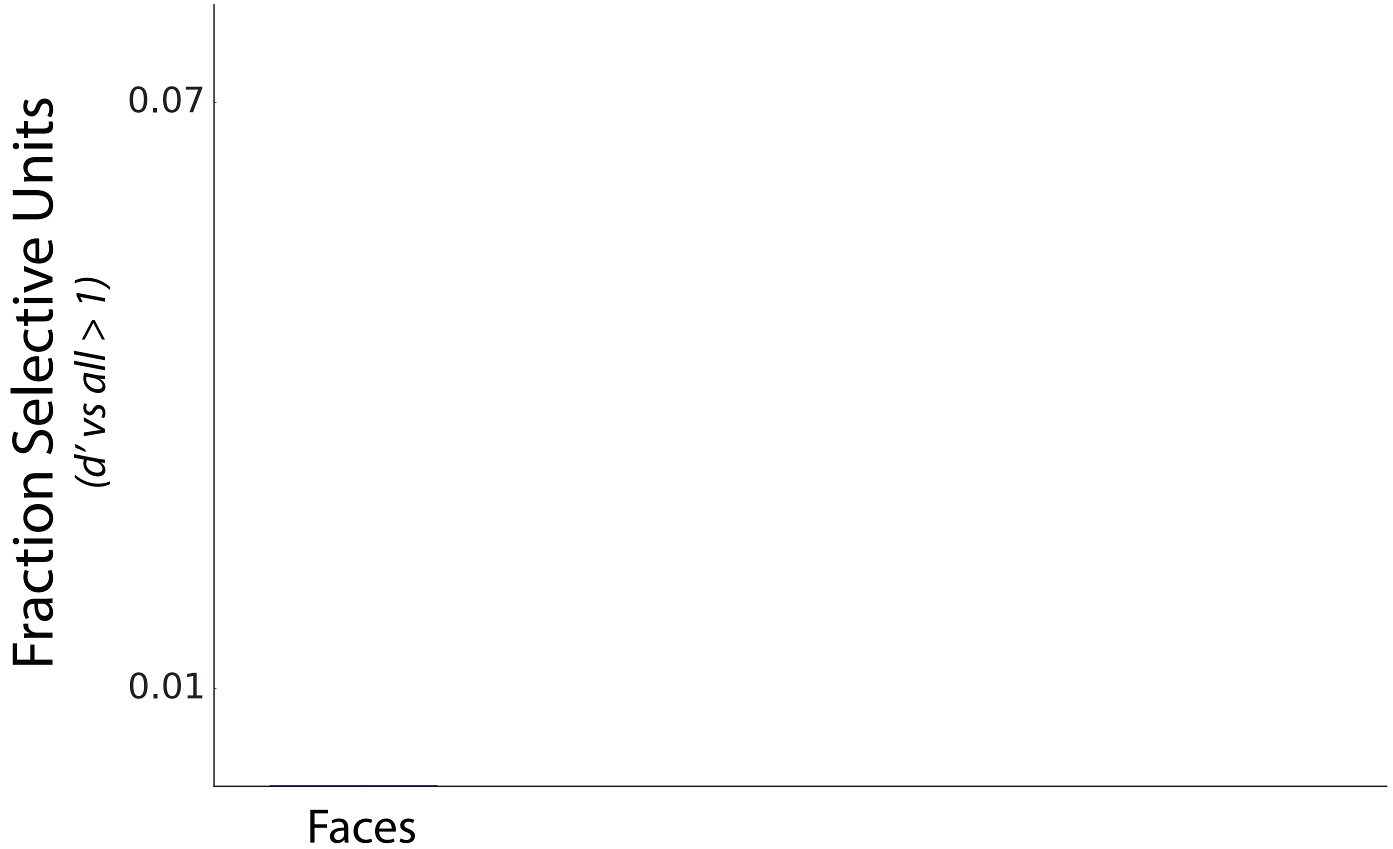
Hypotheses for existing of face-selective units:

i. ~~face processing machinery is in-born~~

ii. ~~it is due to extensive post-natal experience with faces~~

Testing the base-line non-face model.

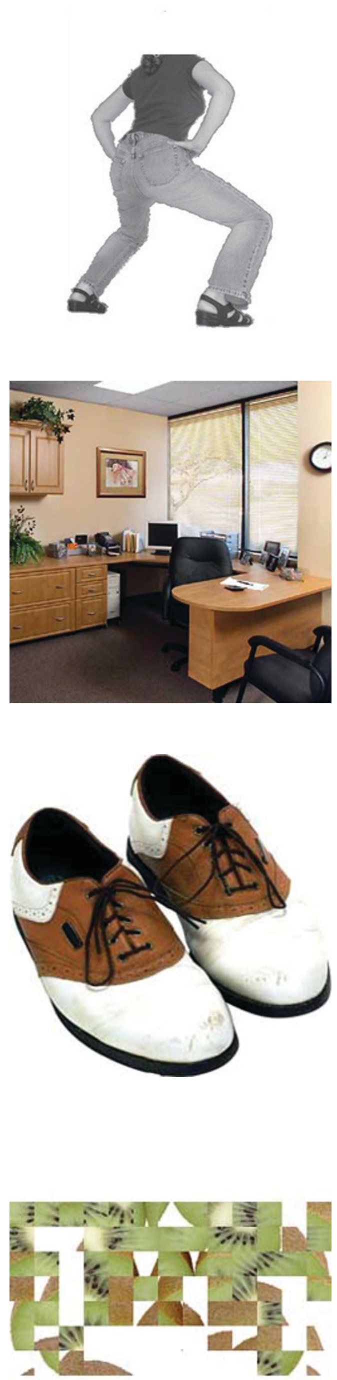
face selective ⇒ d' vs non-face > 1

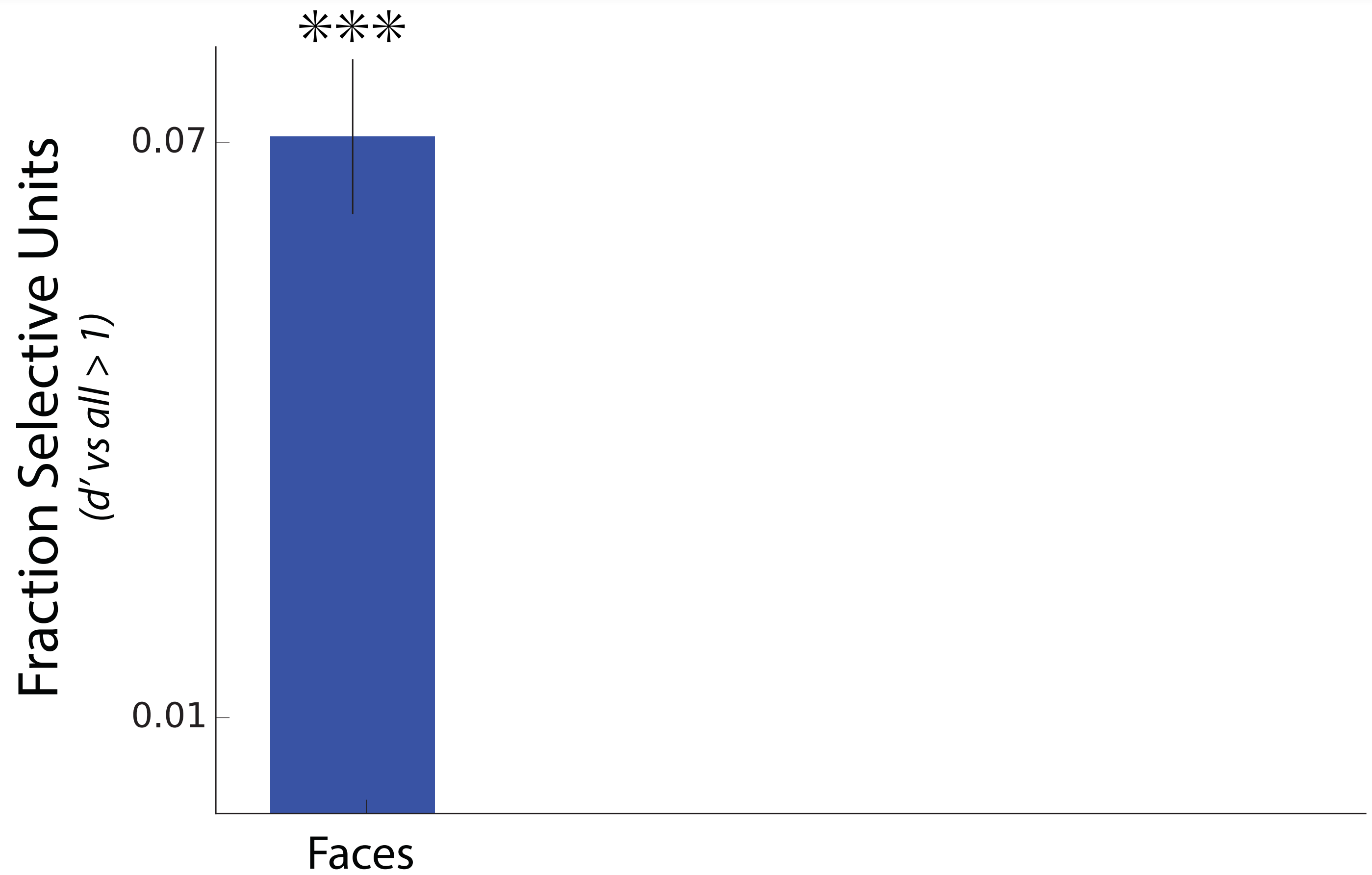


Kanwisher, 1997

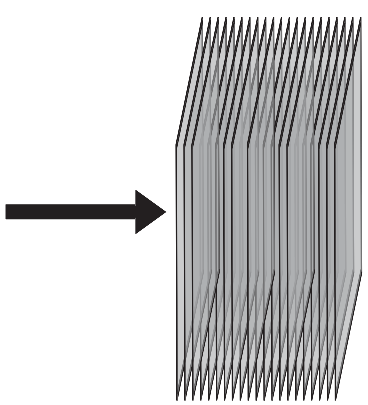
Testing the base-line non-face model.

face selective ⇒ d' vs non-face > 1





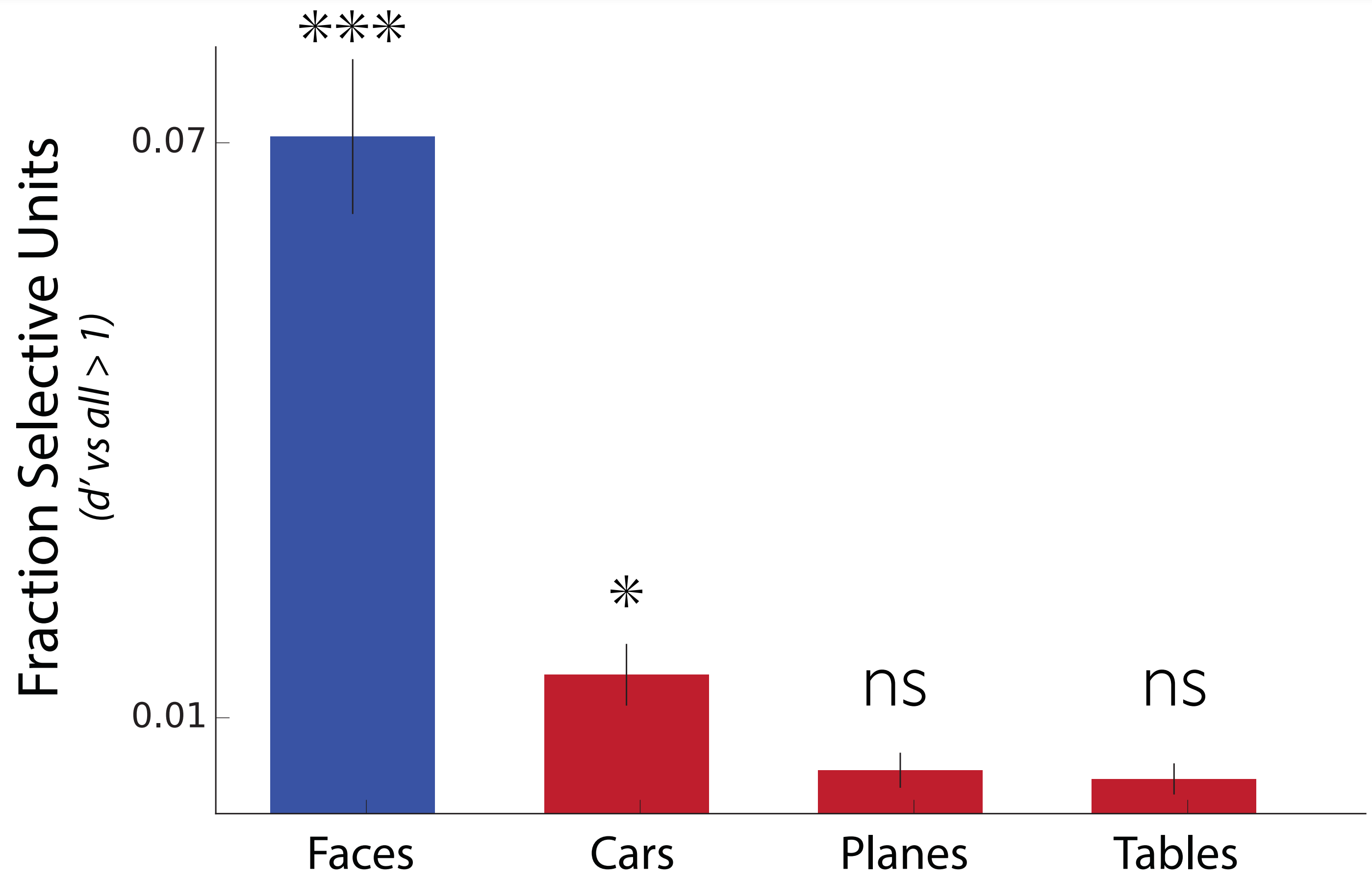
. . .



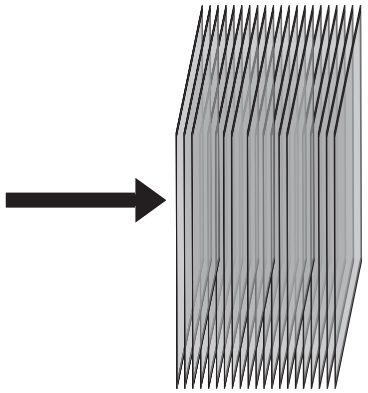
~7% of units in model were face selective

Testing the base-line non-face model.

face selective ⇒ d' vs non-face > 1



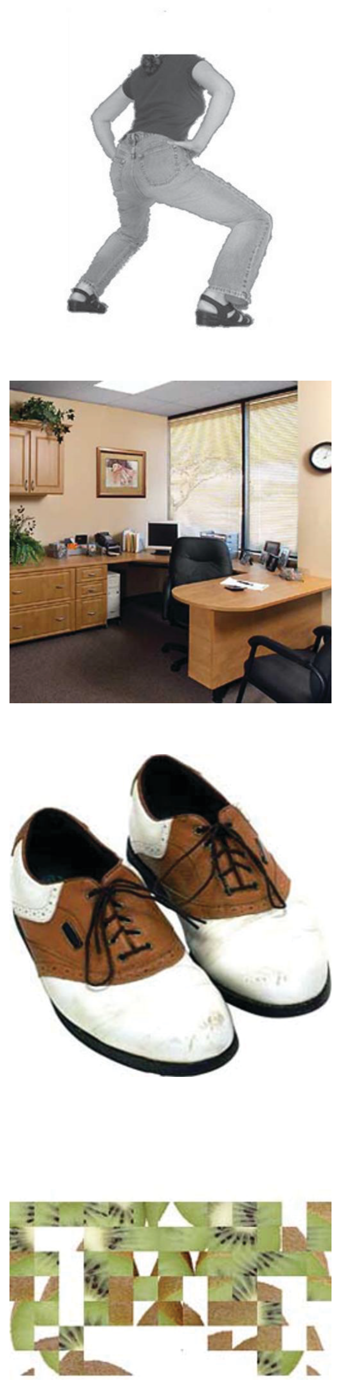
. . .

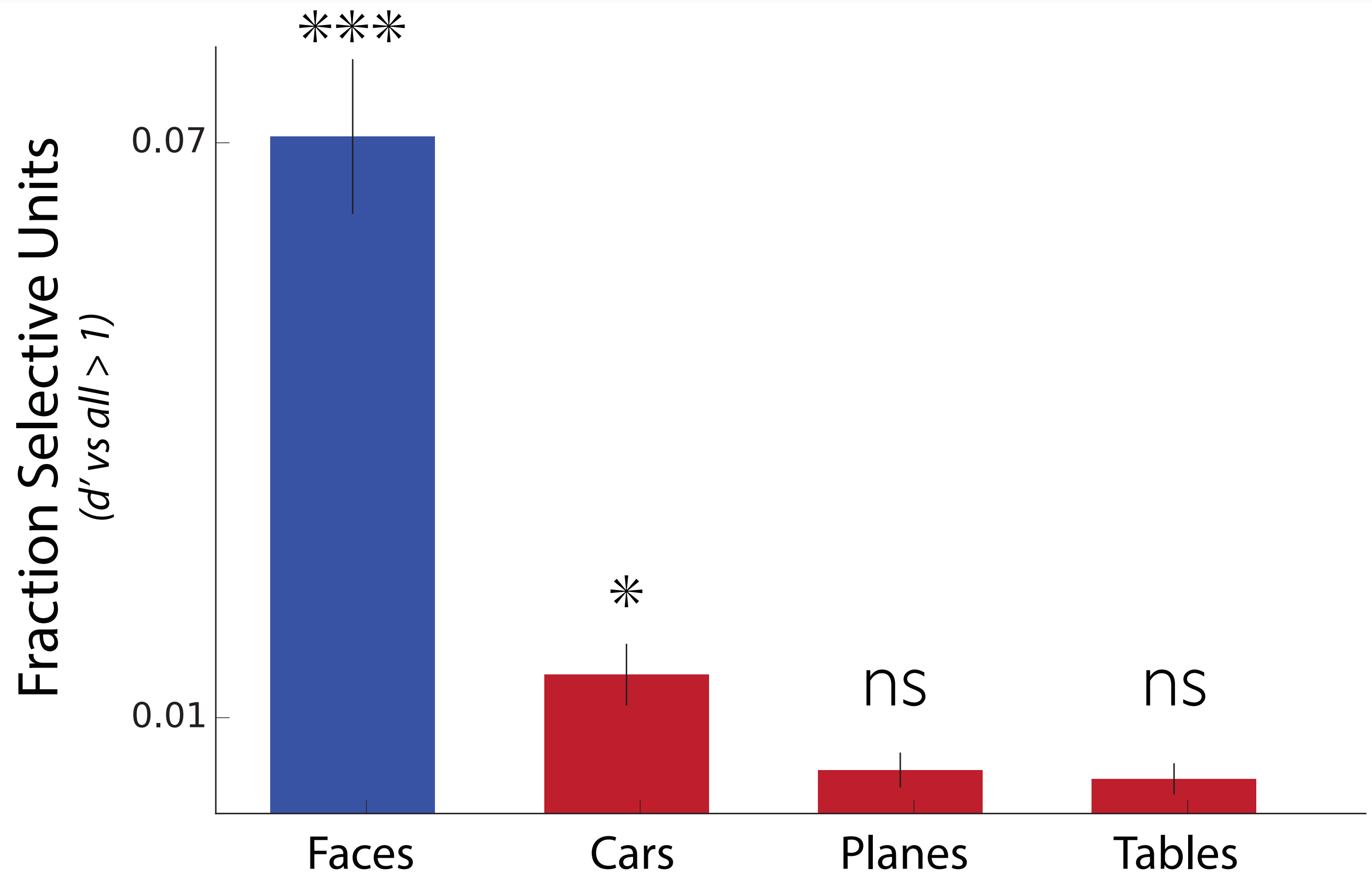


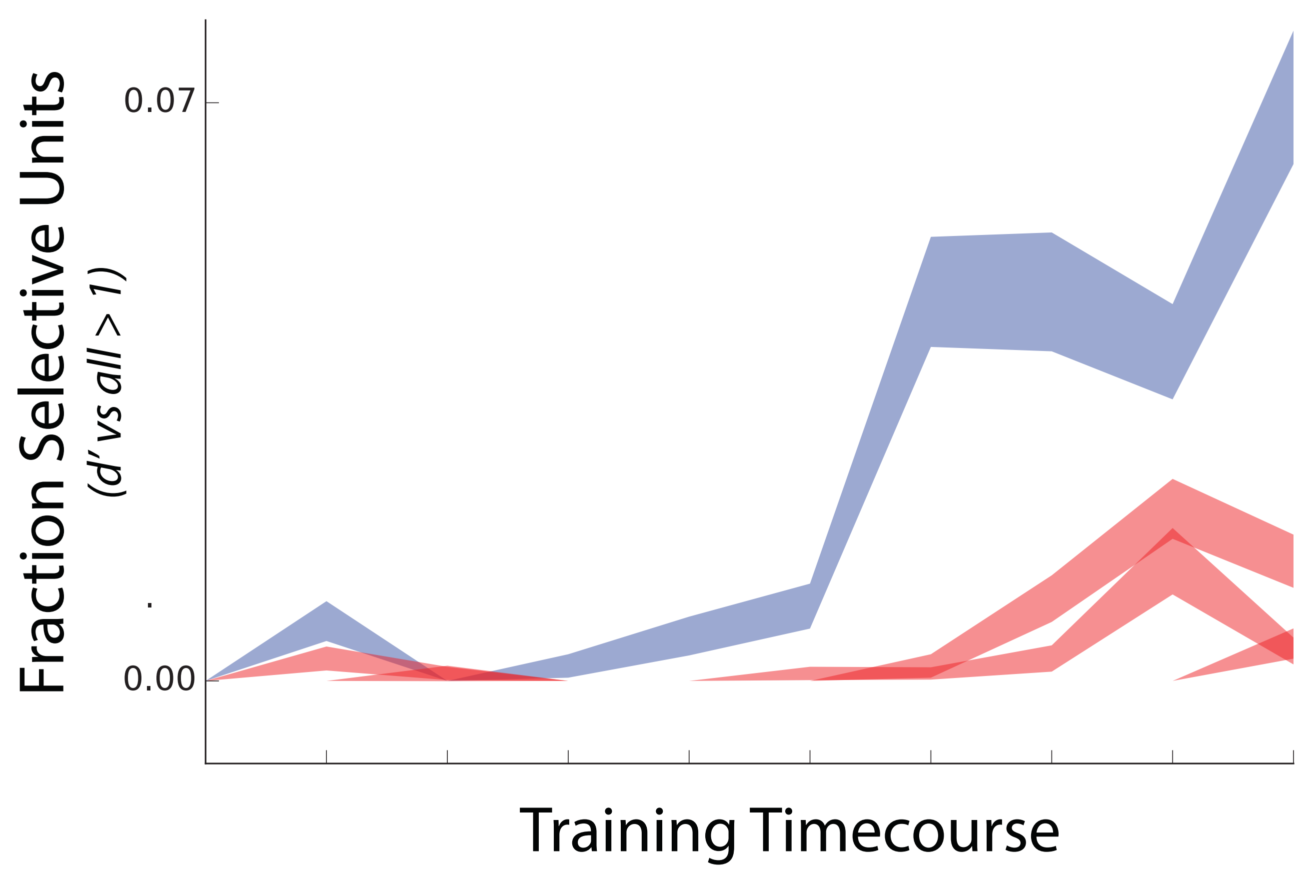
lower (or n.s.)
numbers for several
other tested categories

Testing the base-line non-face model.

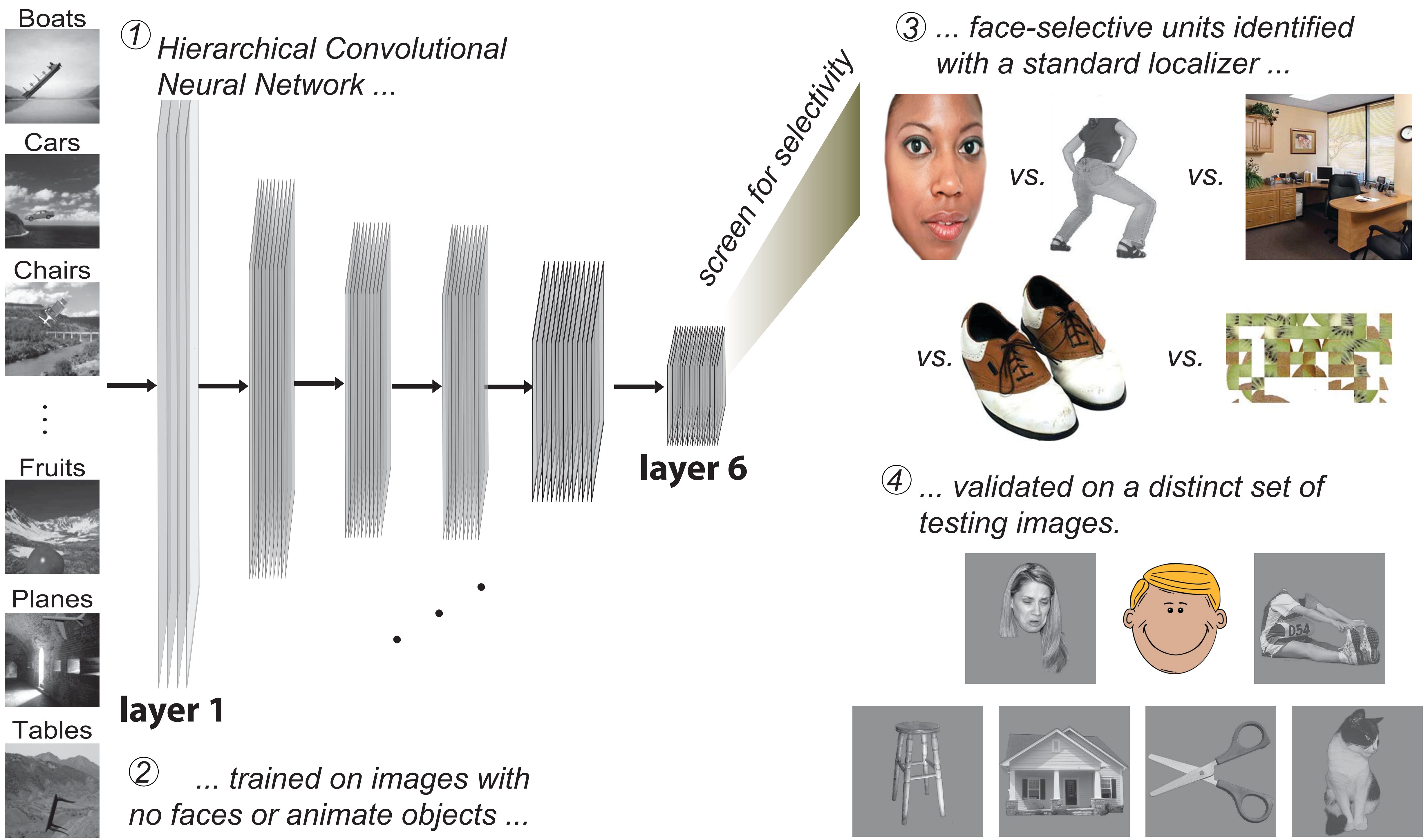
face selective ⇒ d' vs non-face > 1





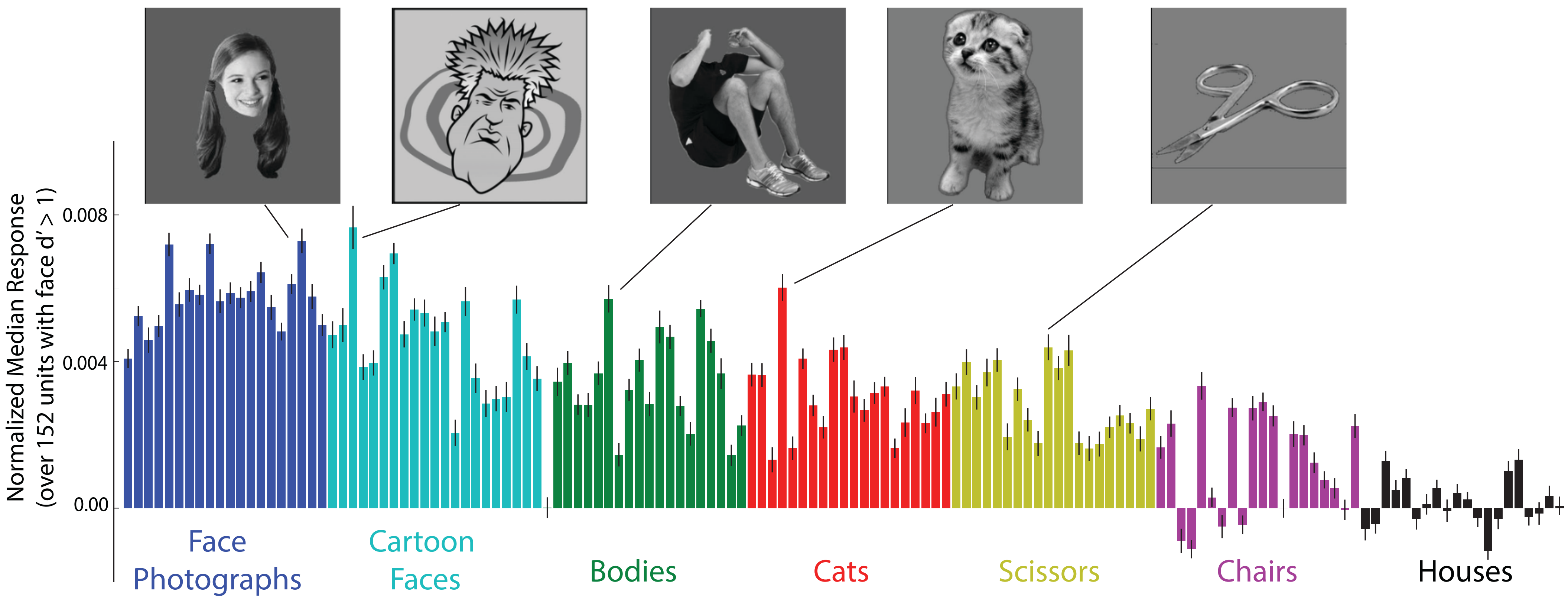


Validating the face-selective units

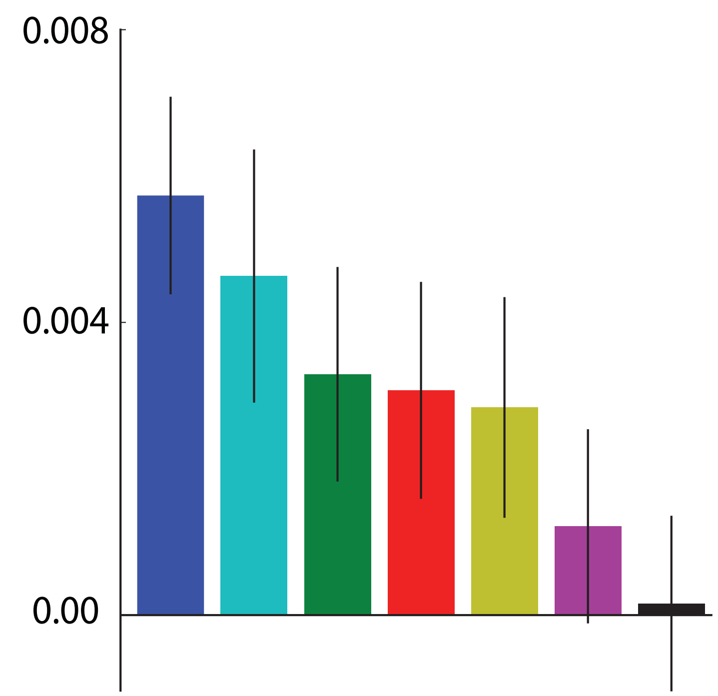


Validating the face-selective units

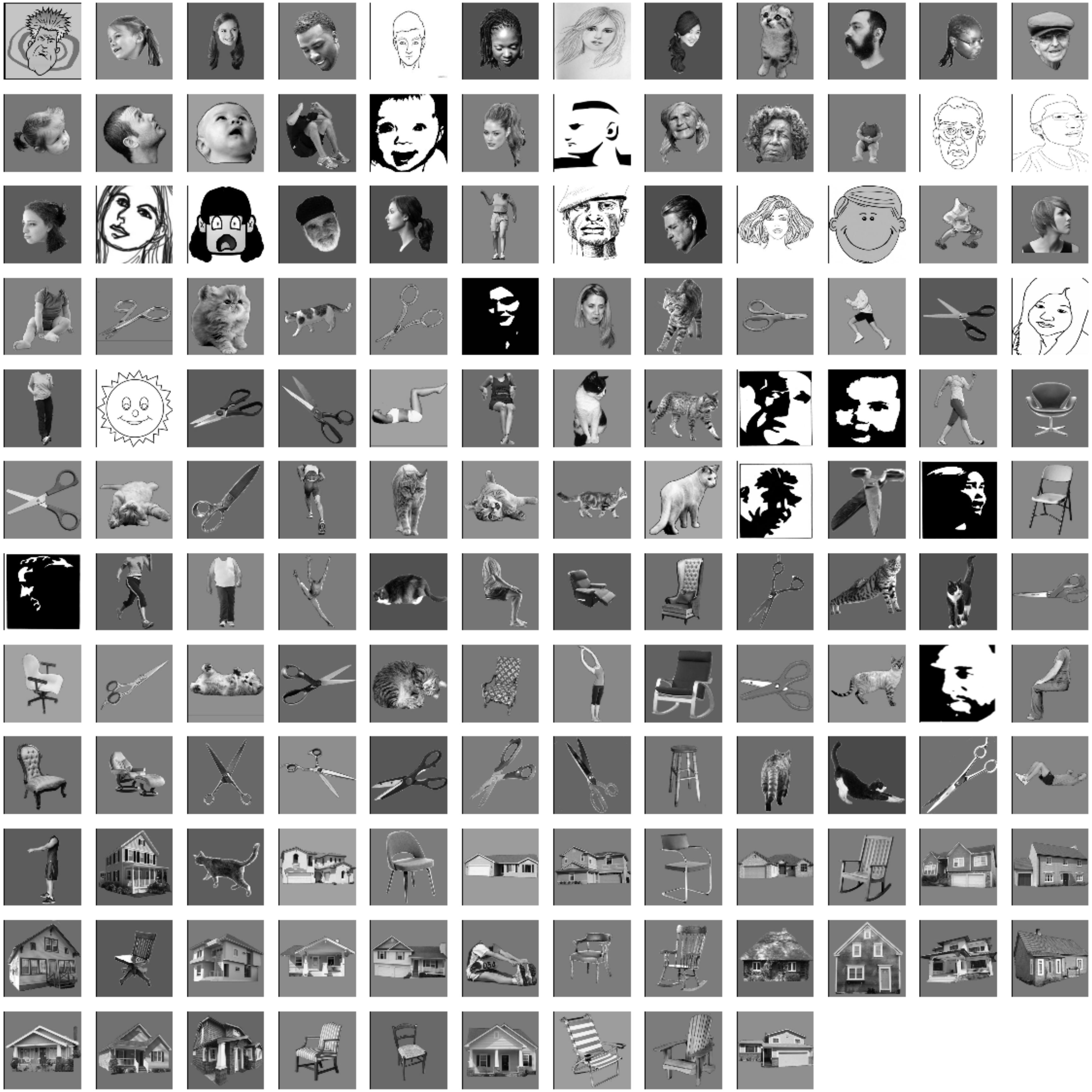
average ranked response over all face-selective units



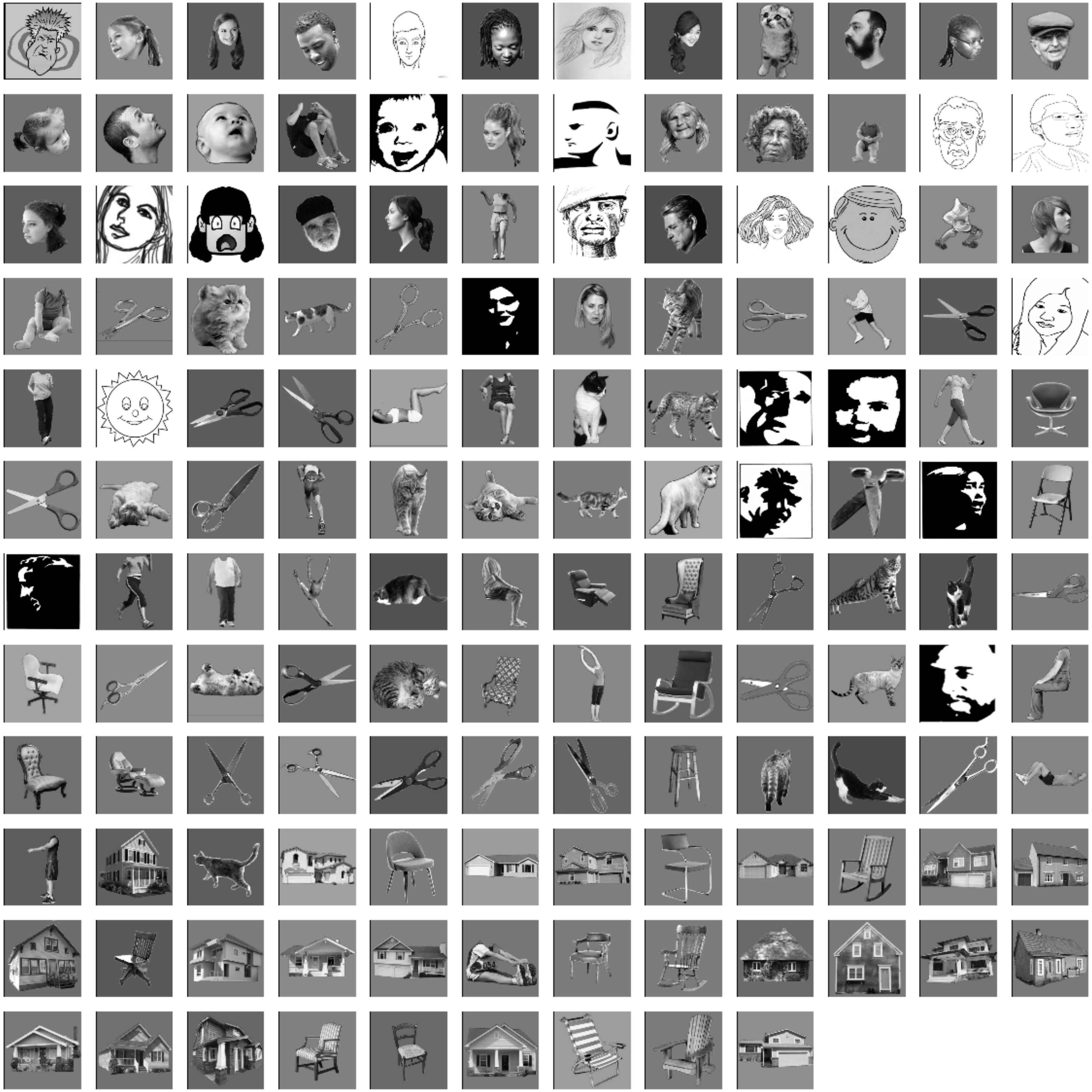
averaged over images within category →



Validating the face-selective units



Validating the face-selective units

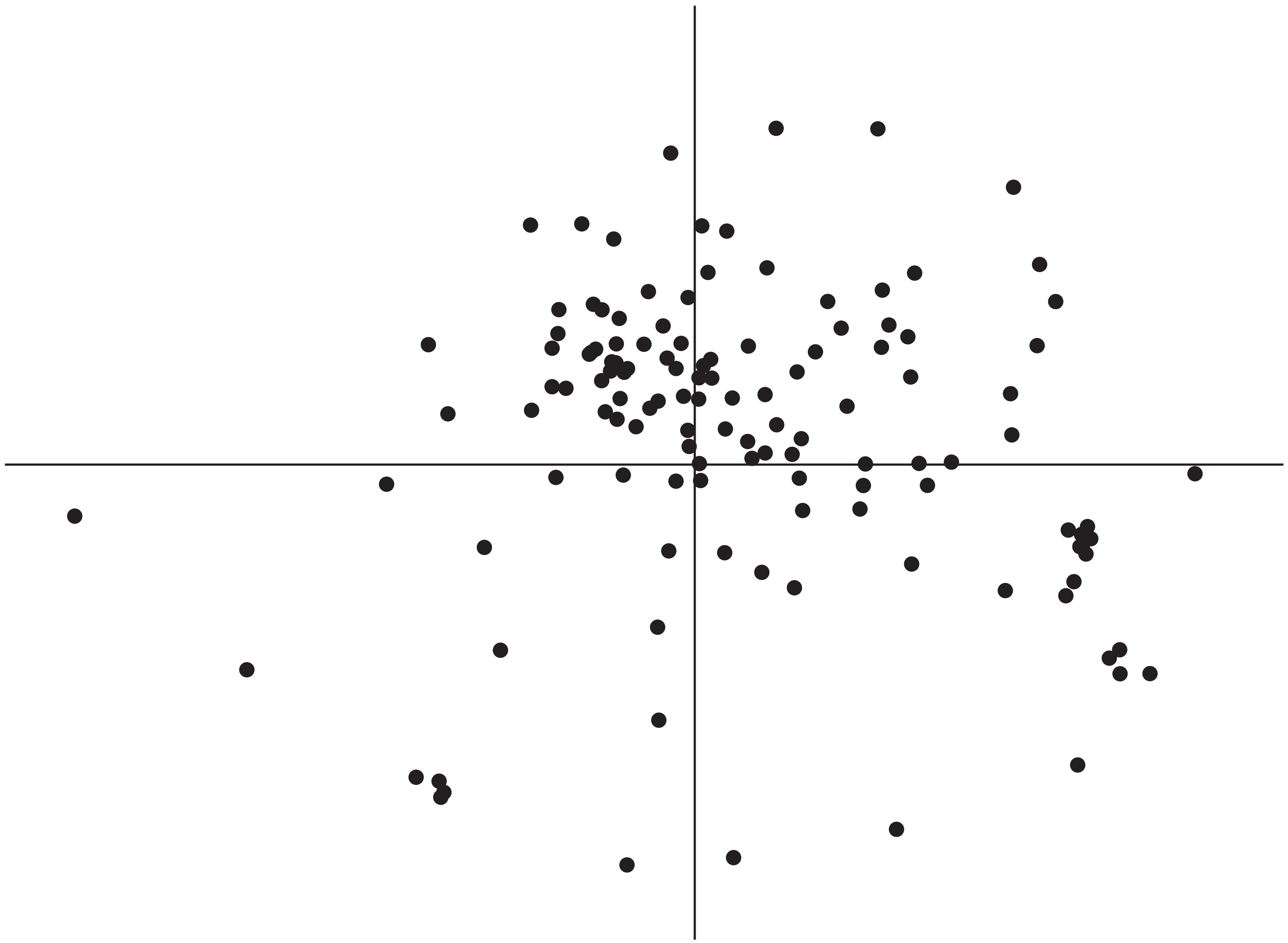


Possible explanation

How could this result be true?

Possible explanation

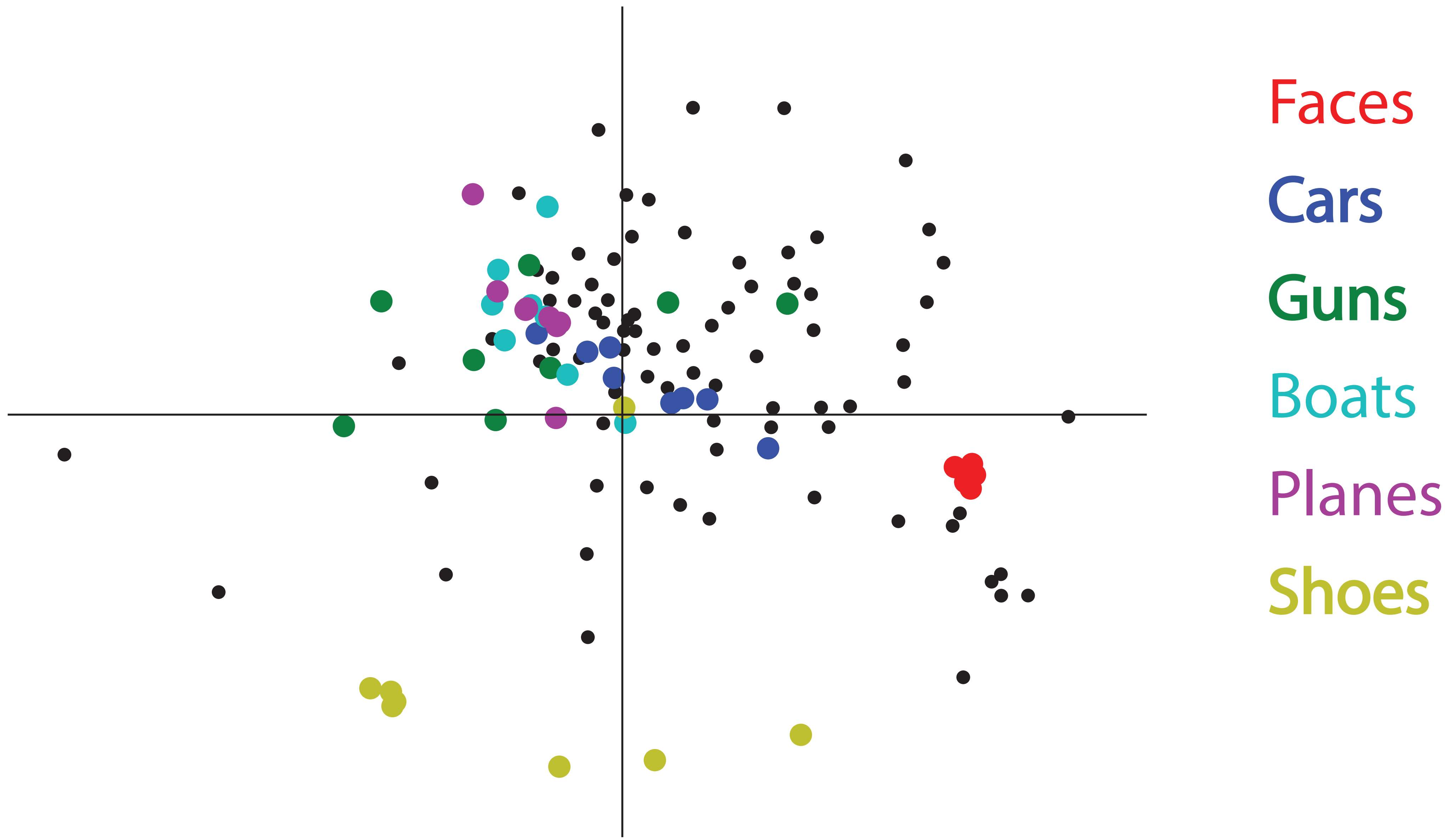
2-d MDS of three-d mesh distances for 128 objects in 16 categories.



How could this result be true?

Possible explanation

2-d MDS of three-3 mesh distances for 128 objects in 16 categories

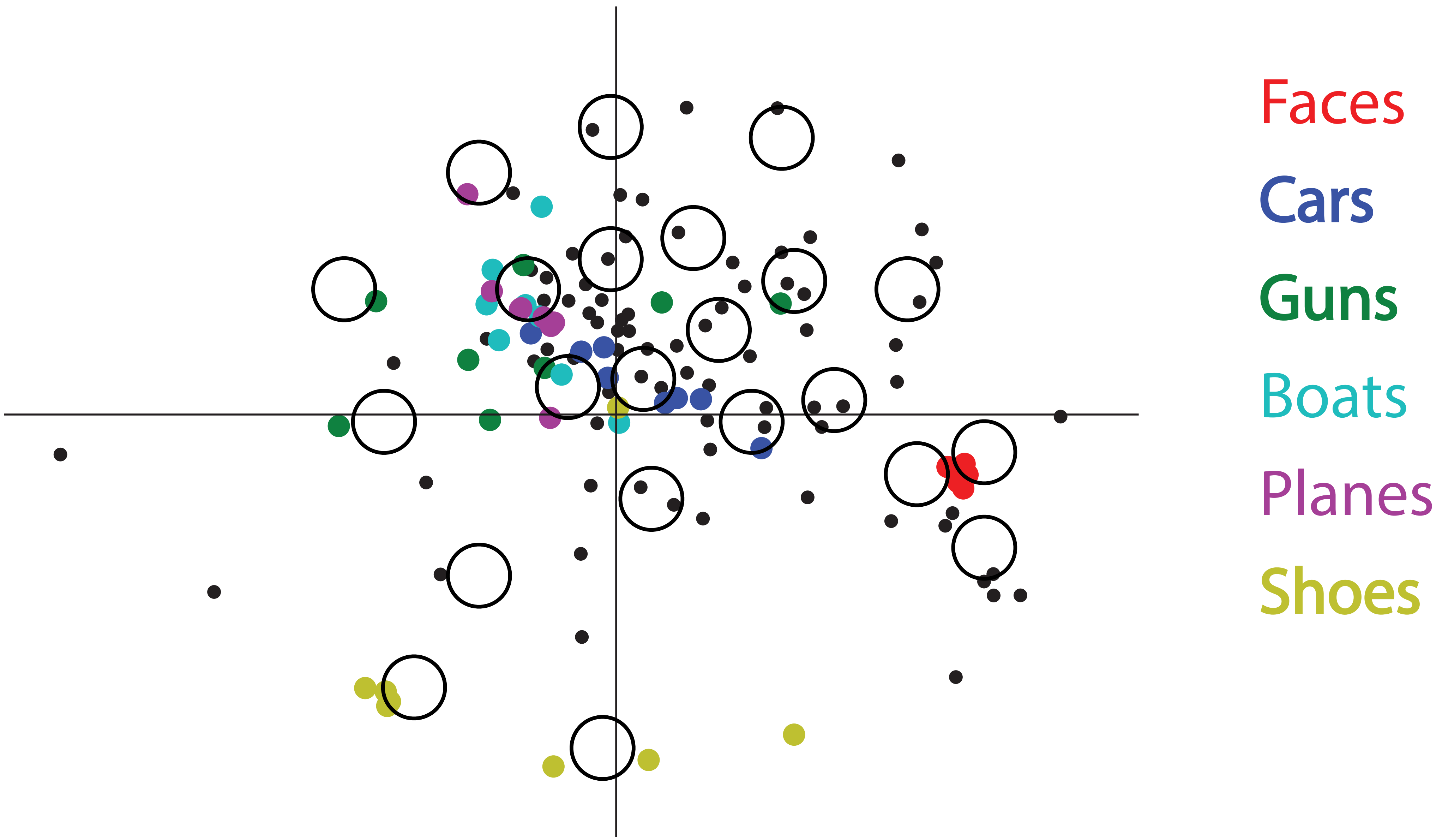


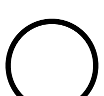
1) Faces are more clustered in shape space than most other categories

2) but they're not totally isolated in shape space.

Possible explanation

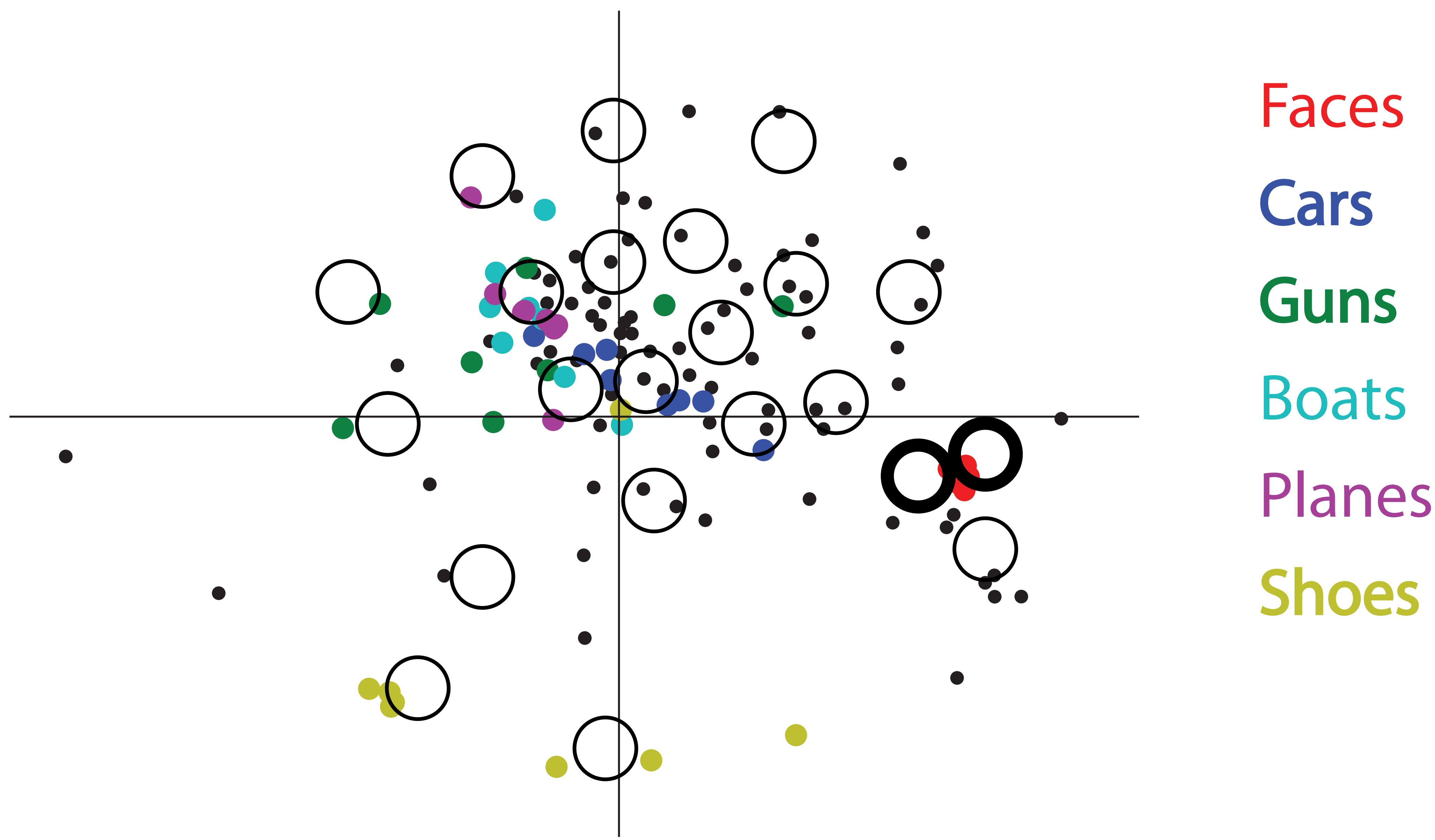
2-d MDS of three-3 mesh distances for 128 objects in 16 categories



 = unit as gaussian blob in shape space.

Possible explanation

2-d MDS of three-3 mesh distances for 128 objects in 16 categories



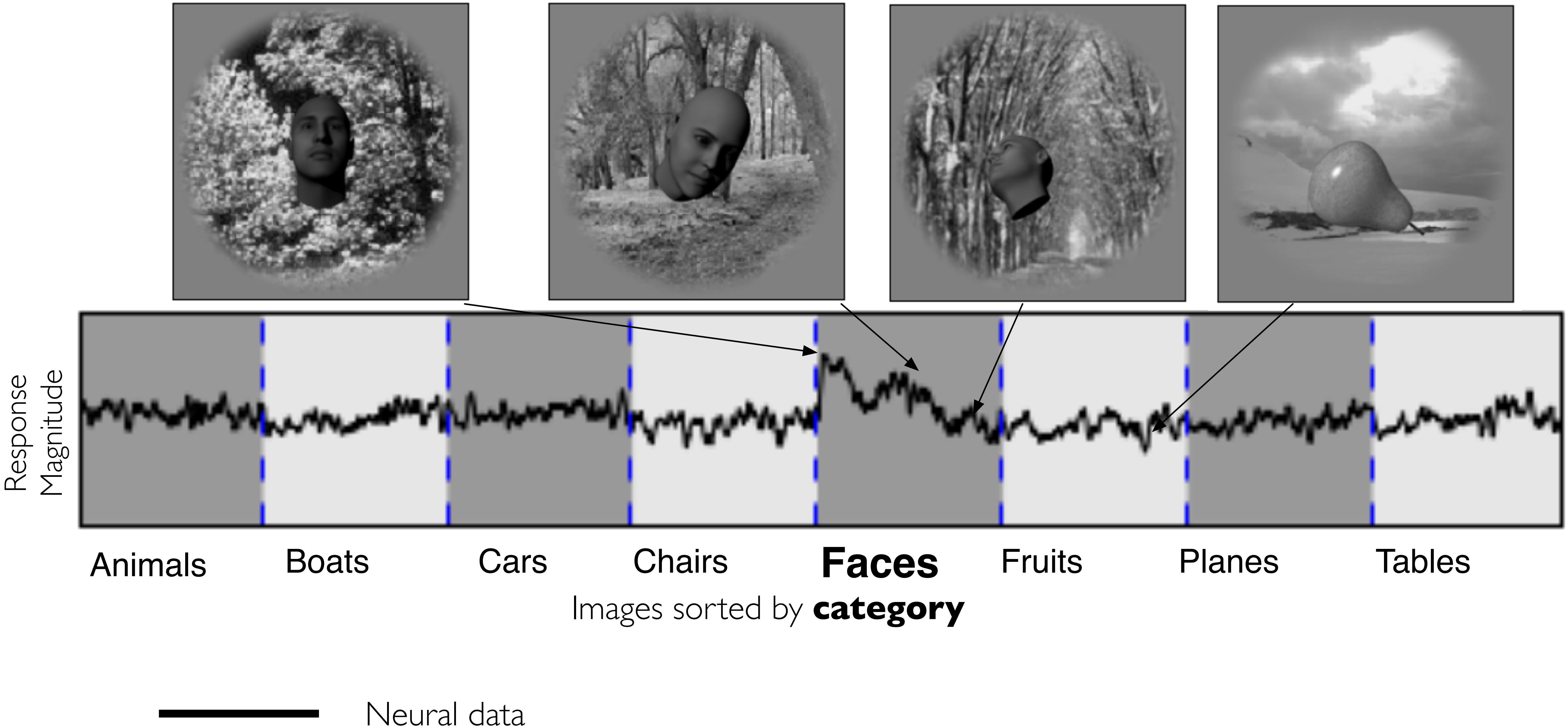
= unit as gaussian blob in shape space.

Multi-array Electrophysiology Experiment

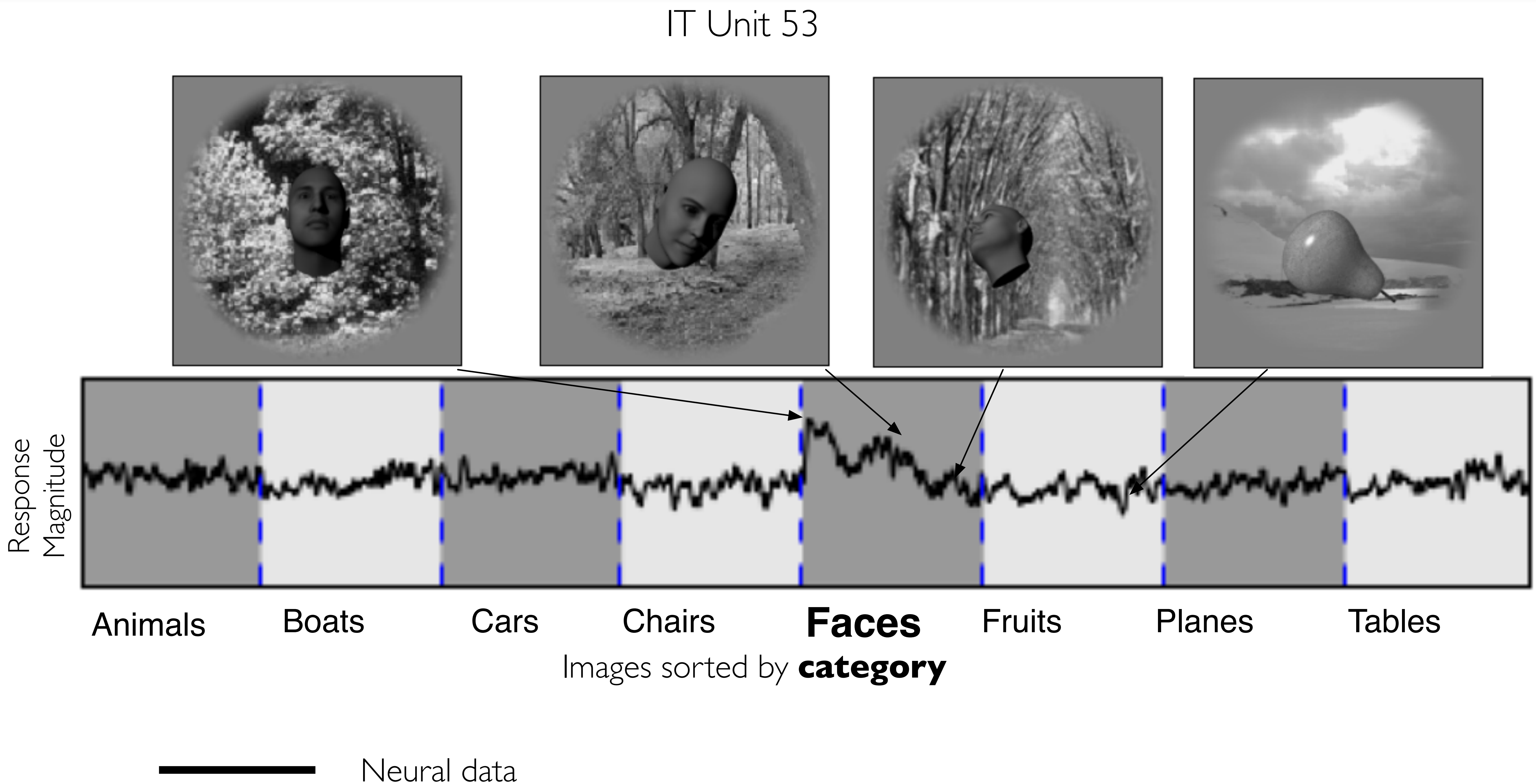
detailed comparison to face neurons

Predictions of Face-Selective Neural Responses

IT Unit 53



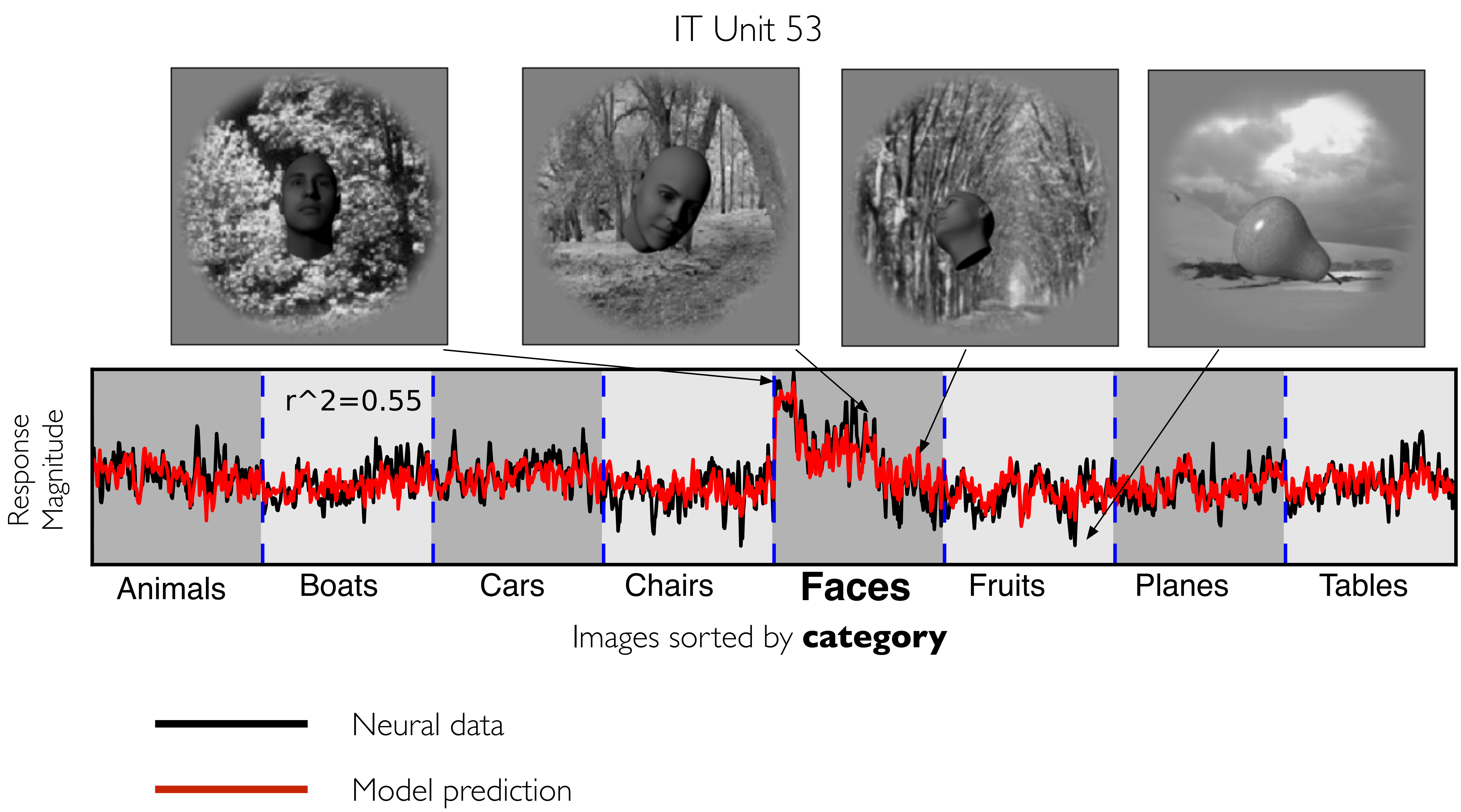
Predictions of Face-Selective Neural Responses



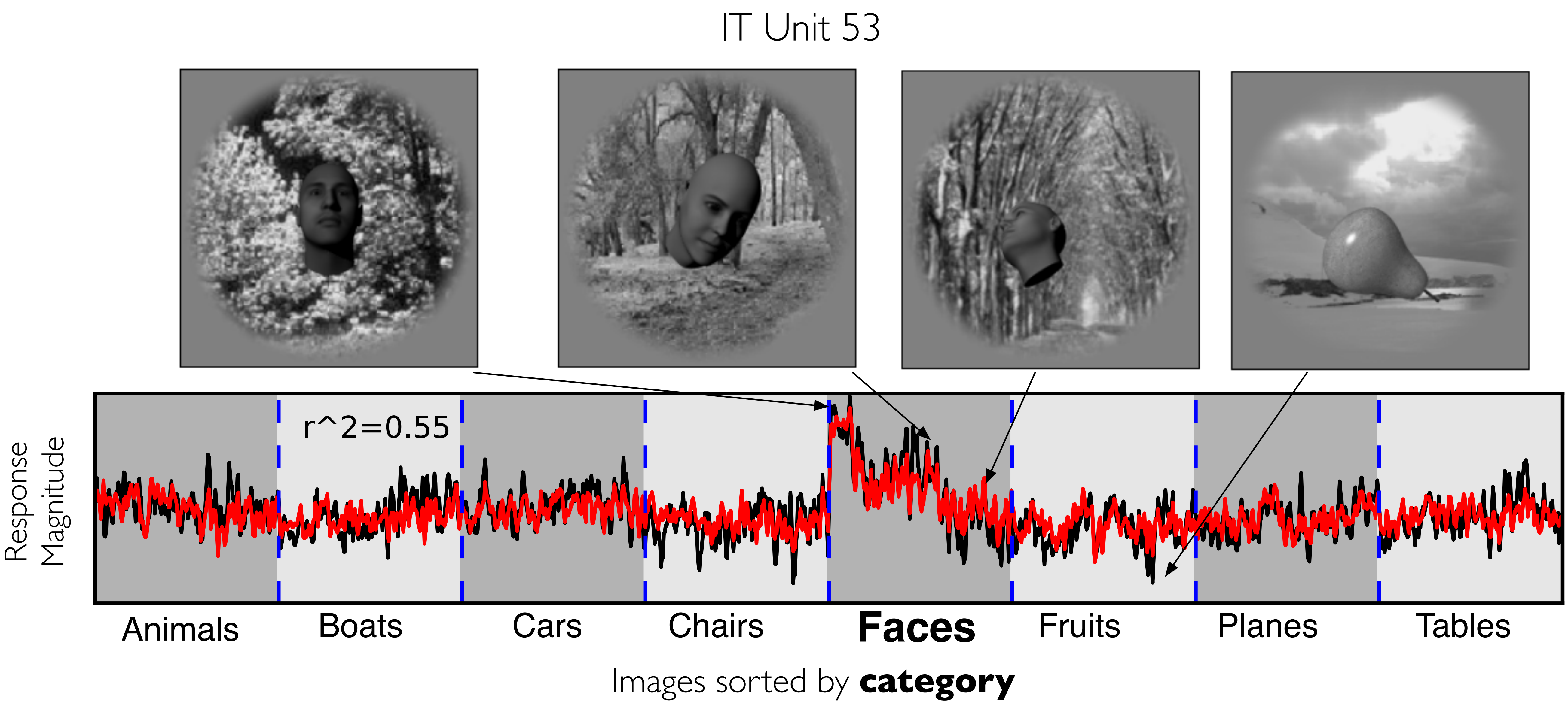
Regularized linear regression to map model units to neural units,

predictions on held-out testing images.

Predictions of Face-Selective Neural Responses



Predictions of Face-Selective Neural Responses



Neural data

Model prediction

Explained Variance Across All Face Selective Units:

With Faces in Training: **51.5** ± 3.9 %

Without Faces in Training: **50.8** ± 4.4 %

Conclusions

Models "raised" without faces can still have face-selective units.

Consistent with Sugita (2008)

Some aspects of specialized face machinery may be explicable from the "null model" of general object recognition.

A third hypothesis for the development of face (and other) selective regions:

- In-born built-in structure or
- Developmentally determined by particular experience.
- **Developmentally determined by general experience?**

Future questions / limitations

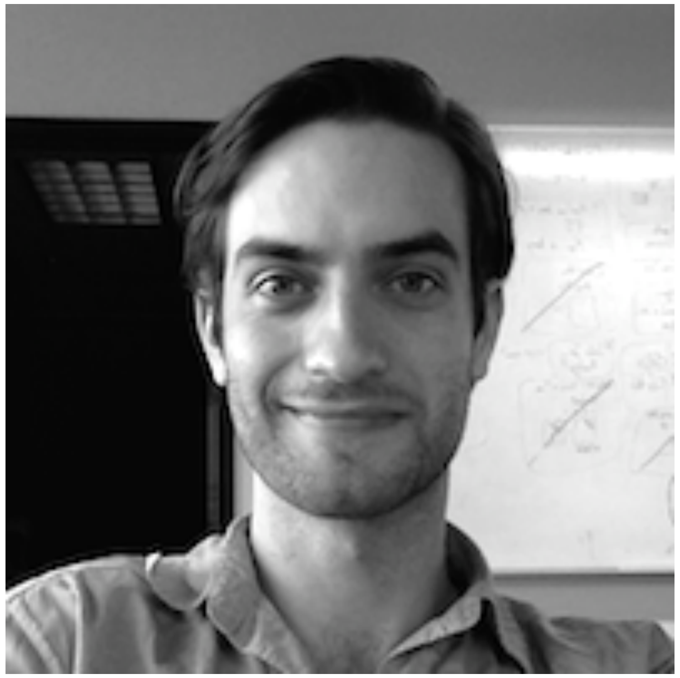
Better exploration of category selective across many categories as a function of contents of training data.

More detailed comparison to neurophysiology of face patch system. Freiwald & Tsao, 2011, Issa & DiCarlo 2014

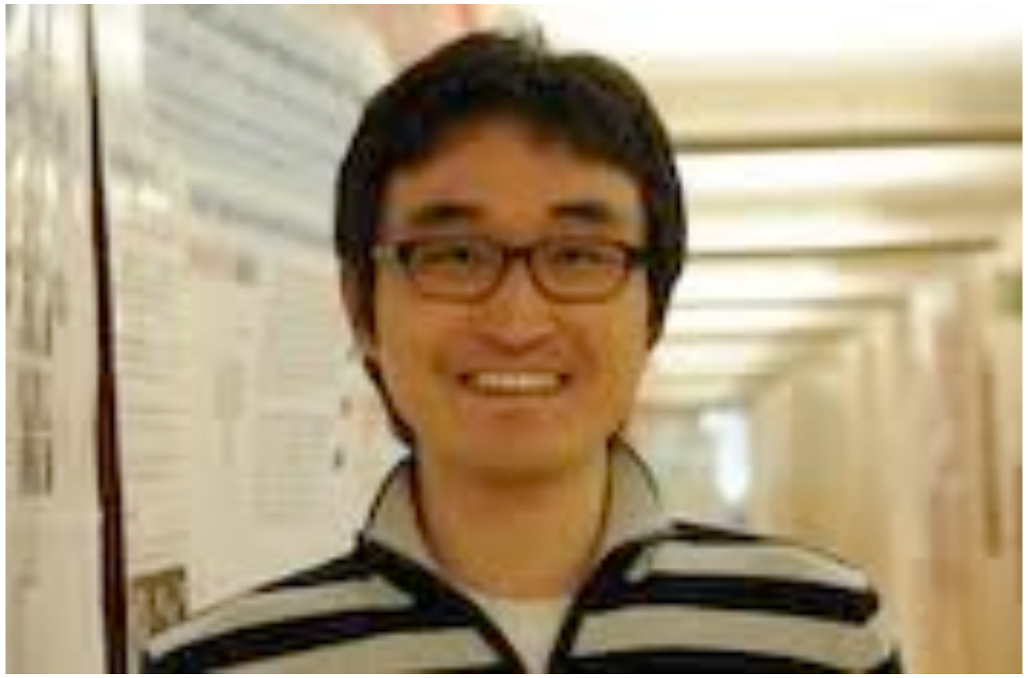
Explicitly address question of spatial layout.

Results here do NOT imply monkeys without face experience will necessarily have a ***patch***.

Thanks to great colleagues!



Michael Cohen



Ha Hong

Nancy Kanwisher

Jim DiCarlo