Microsoft Research

Each year Microsoft Research hosts hundreds of influential speakers from around the world including leading scientists, renowned experts in technology, book authors, and leading academics, and makes videos of these lectures freely available.

Mechanisms underlying visual object recognition:

humans vs. monkeys vs. neurons vs. machines

Neural Information Processing Systems (NIPS), Lake Tahoe, CA

December, 2013

James DiCarlo MD, PhD

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

brain cognitive sciences

Why study object recognition in the brain?

Why study object recognition in the brain?

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

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

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

Why study object recognition in the brain?

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

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

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

Brains vs. Machines

Which system is better?

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

Brains vs. Machines

Which system is better?

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

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

Brains vs. Machines

Which system is better?

Problem to solve	*Our brain*	*Machines today*
Calculation		WINNER
Win at chess		WINNER
Win at Jeopard		
"Memory"		
"Seeing"		
Pattern matching		WINNER
Object recognition		WINNER
Scene "understanding"	WINNER	
Walking	WINNER	

Gateway problem for understanding neocortex

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

Object recognition ("detection") as viewed by computer vision ...

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

Object recognition ("detection") as viewed by computer vision ...

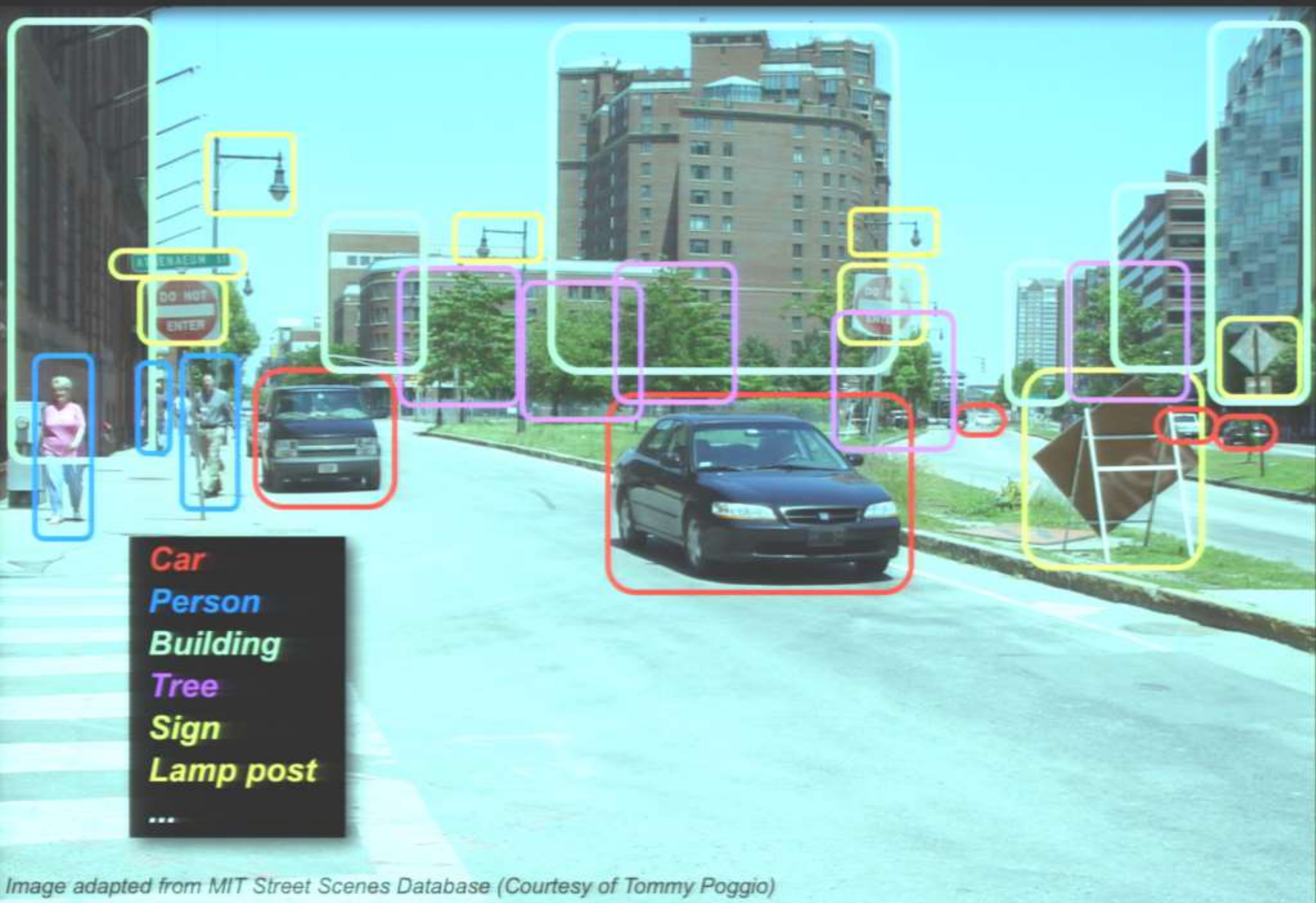


Image adapted from MIT Street Scenes Database (Courtesy of Tommy Poggio)

Object recognition as solved by primates

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

Object recognition as solved by primates

Central ~10 degrees

Object recognition as solved by primates

~200 ms snapshots

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

Object recognition as solved by primates

~200 ms snapshots

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

Object recognition as solved by primates

DiCarlo, Zoccolan and Rust. ***Neuron*** *(2012)*

Core object recognition

central ~10 deg of visual field

< 200 ms viewing duration

Our brain is very good at core object recognition

DiCarlo, Zoccolan and Rust. **Neuron** (2012)

Core object recognition

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

Our brain is very good at core object recognition

DiCarlo, Zoccolan and Rust. ***Neuron*** *(2012)*

Core object recognition

central ~10 deg of visual field

< 200 ms viewing duration

Our brain is very good at core object recognition

DiCarlo, Zoccolan and Rust. **Neuron** (2012)

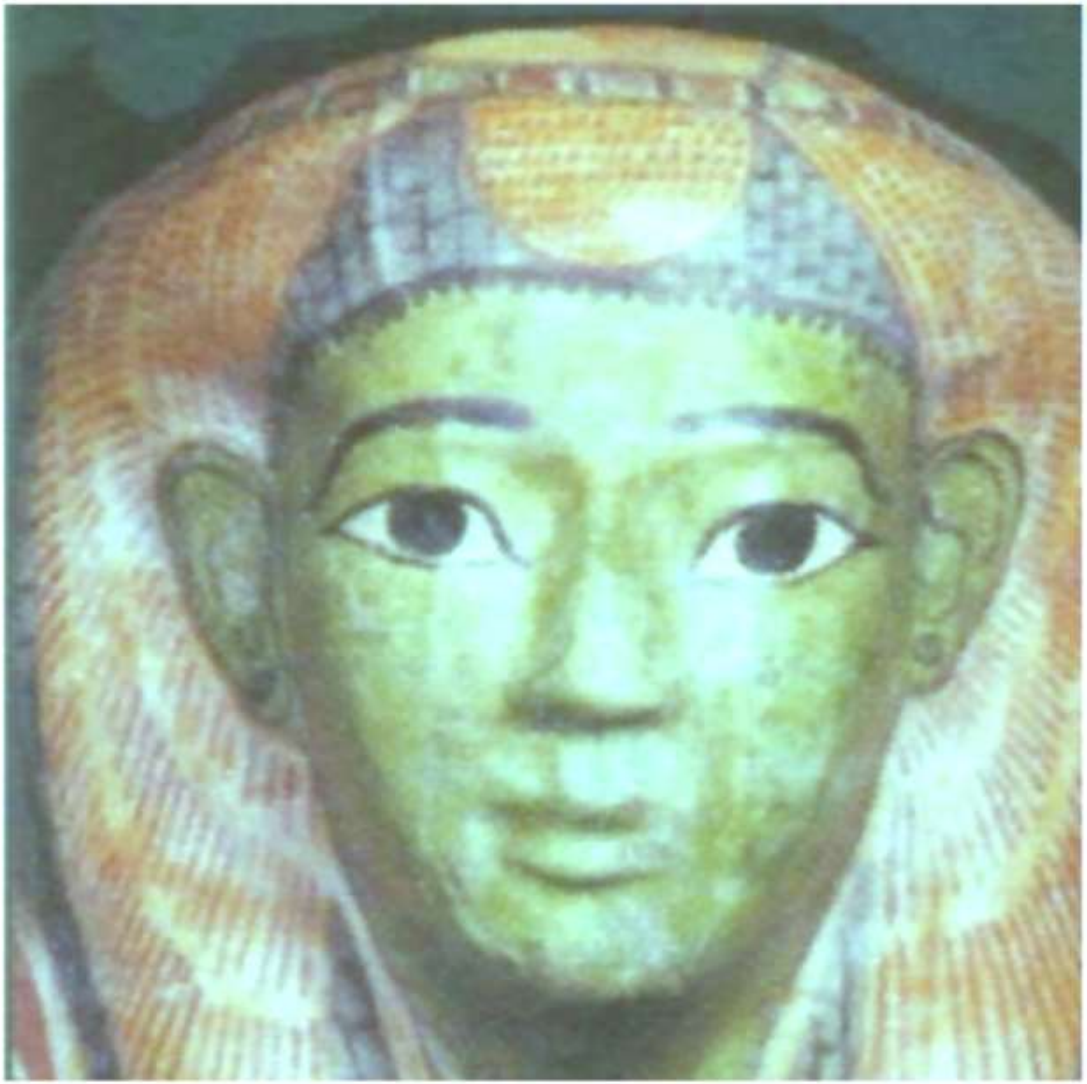
Core object recognition

central ~10 deg of visual field

< 200 ms viewing duration

Our brain is very good at core object recognition

DiCarlo, Zoccolan and Rust. **Neuron** (2012)



Core object recognition

central ~10 deg of visual field

< 200 ms viewing duration

Our brain is very good at core object recognition

DiCarlo, Zoccolan and Rust. ***Neuron*** *(2012)*

Core object recognition

central ~10 deg of visual field

< 200 ms viewing duration

Our brain is very good at core object recognition

DiCarlo, Zoccolan and Rust. **Neuron** (2012)

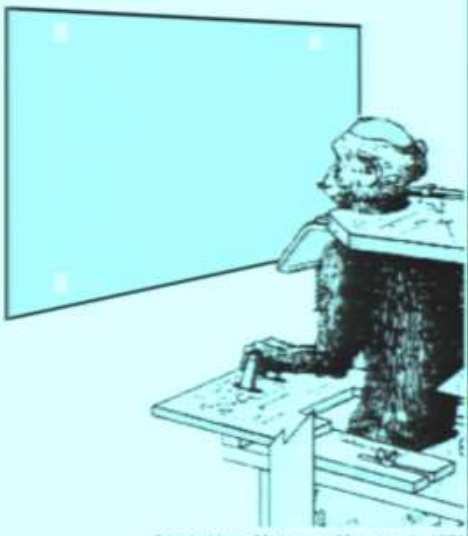
Core object recognition

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

- *Fast*
- *Feels effortless*
- *No pre-cueing needed*
- *Entertain many objects*
- *Tolerant to variation*

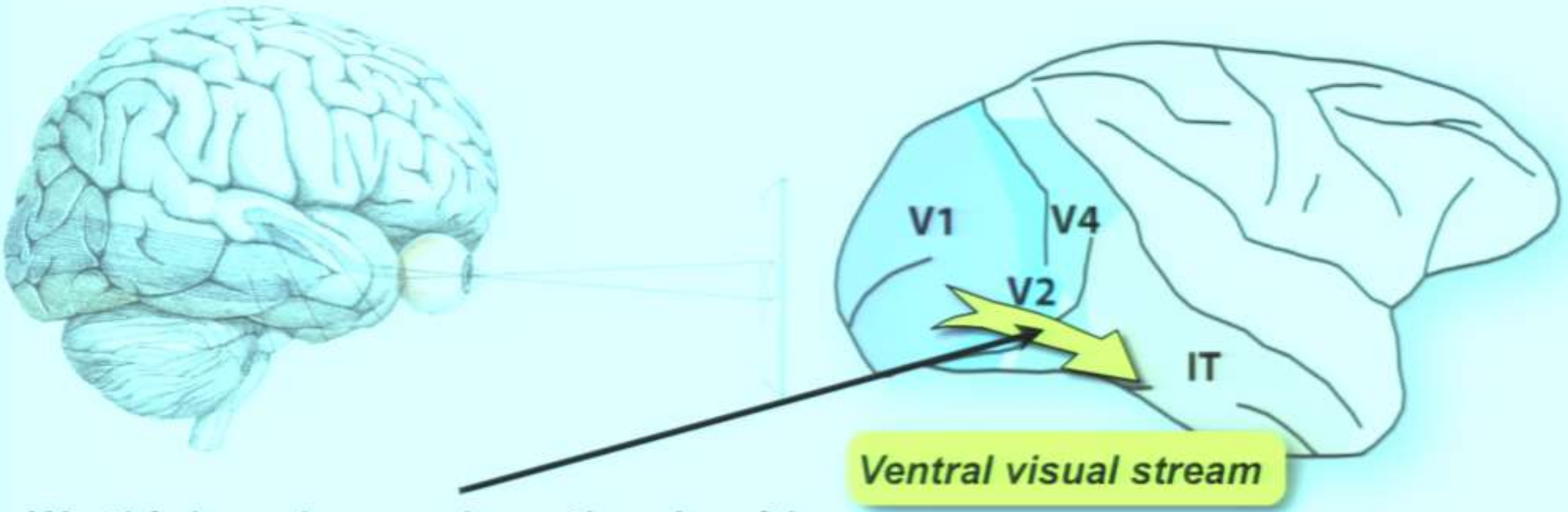
Systems neuroscience: the non human primate model

Systems neuroscience: the non human primate model



Adapted from Mottet and Mountcastle 1981

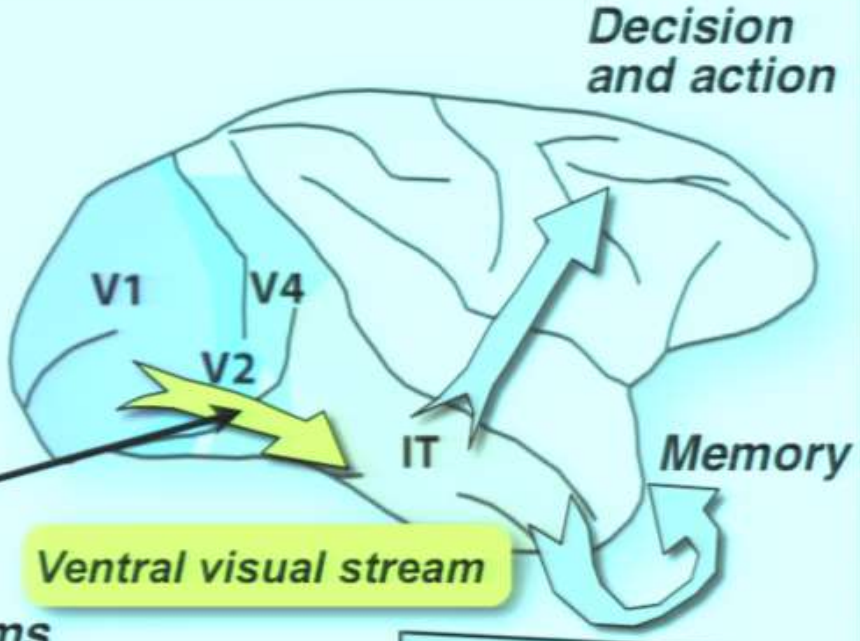
Systems neuroscience: the non human primate model



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

Adapted from Mishkin and Mountcastle 1983

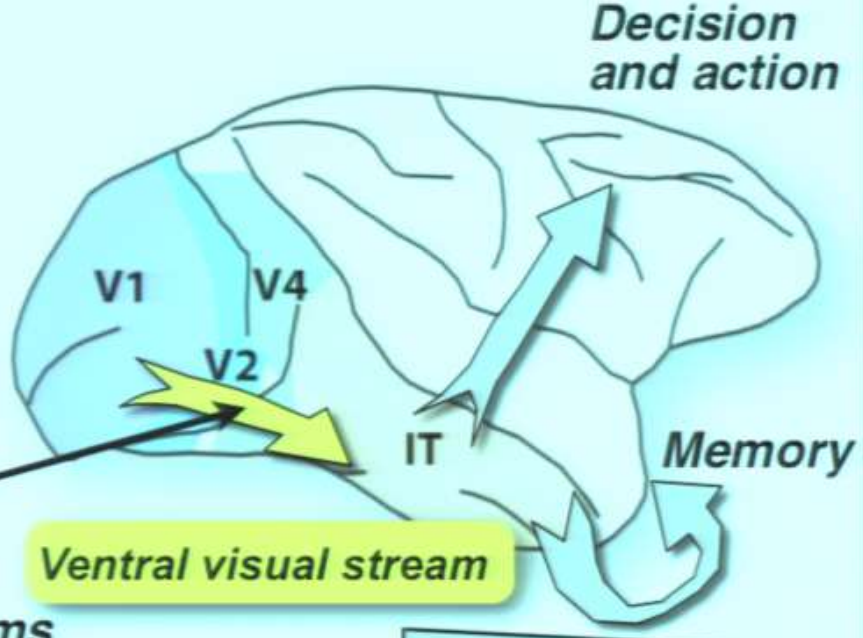
Systems neuroscience: the non human primate model



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

Adapted from Mottler and Mountcastle 1981

Systems neuroscience: the non human primate model

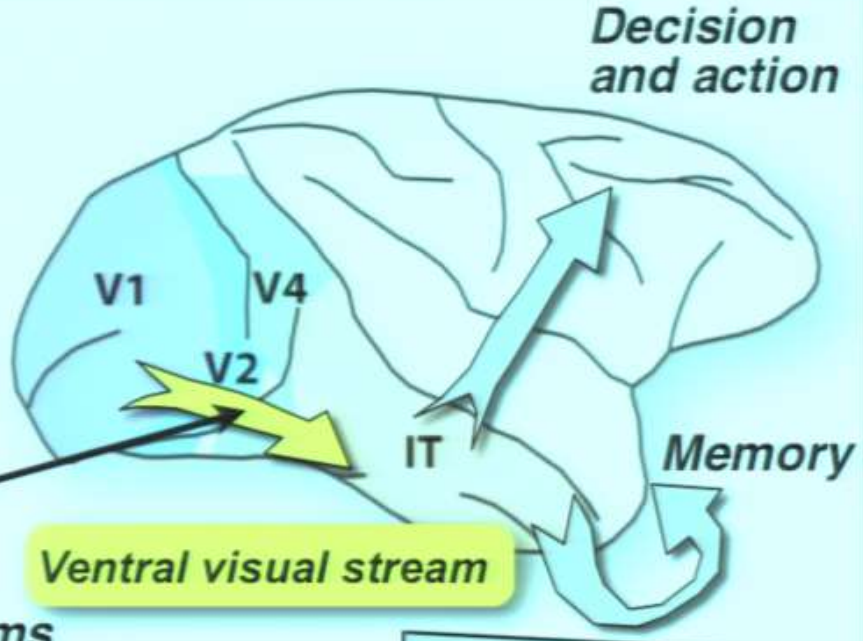


We think we know where the algorithms and representations that solve core object recognition live in the primate brain.

We can study those representations at the level of neuronal spikes in a model system with comparable behavioral abilities.

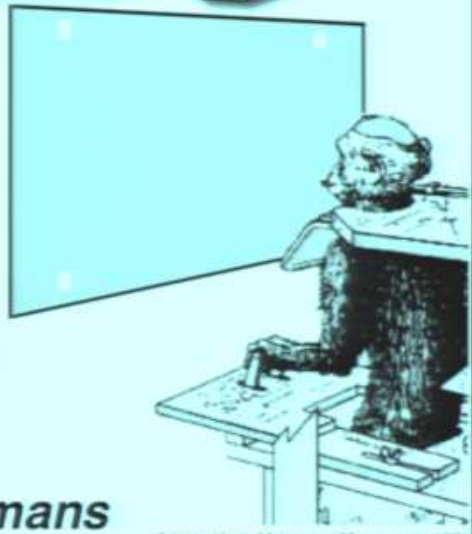
Adapted from Mother and Mountcastle 1981

Systems neuroscience: the non human primate model



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

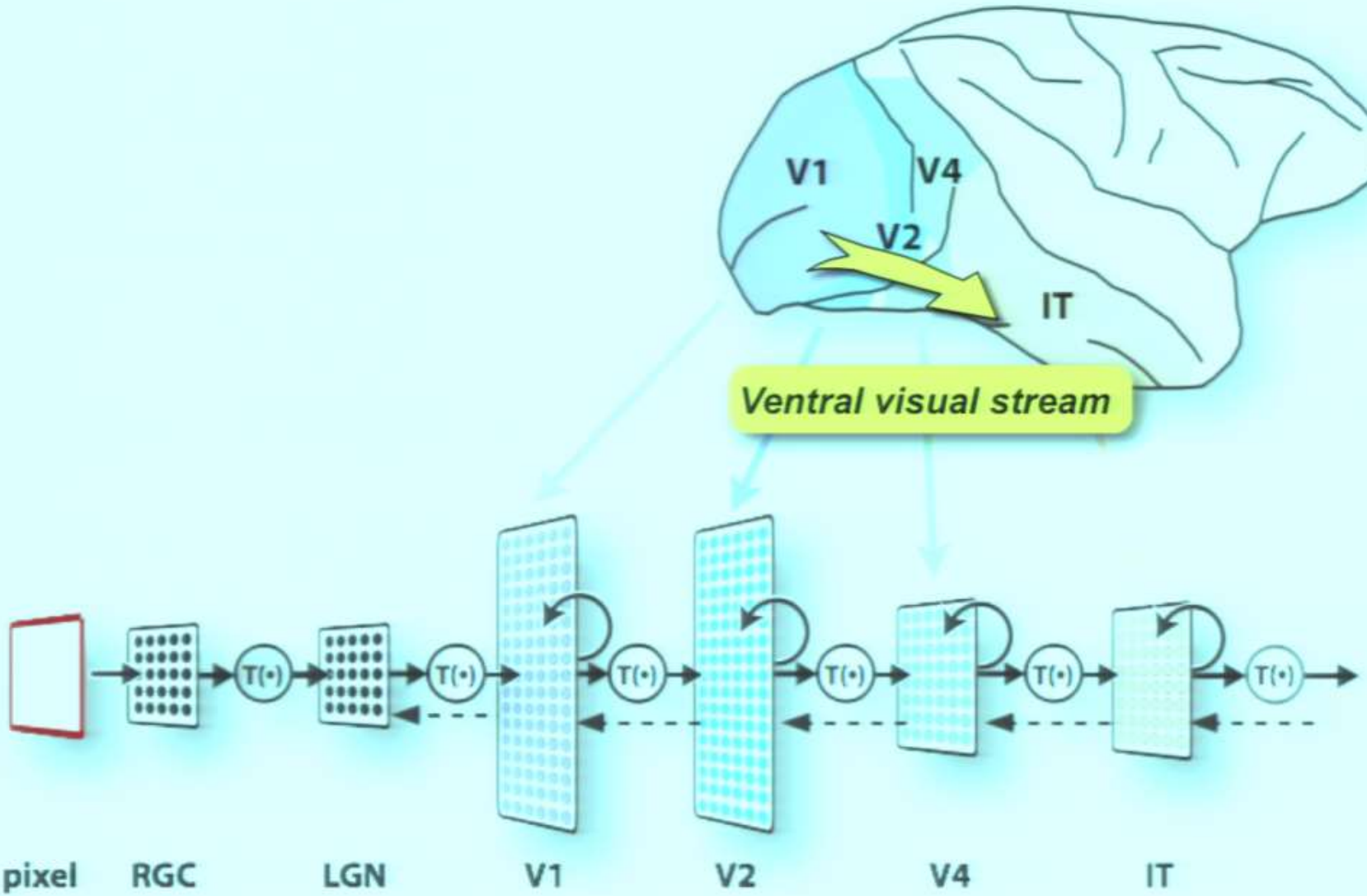
We can study those representations at the level of neuronal spikes in a model system with comparable behavioral abilities.



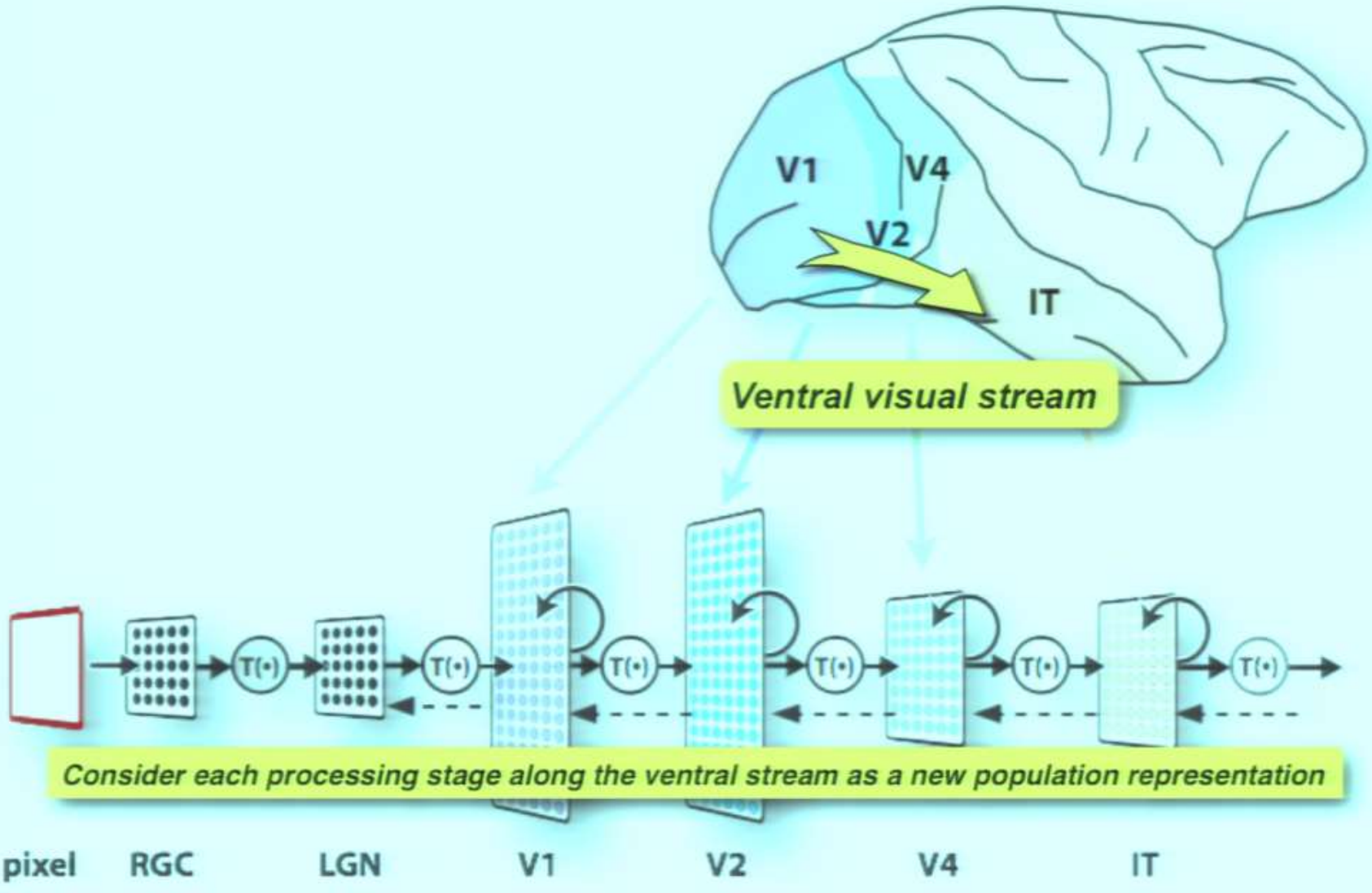
We can directly compare the properties of those representations with likely homologous regions in humans

Adapted from Mother and Mountcastle 1981

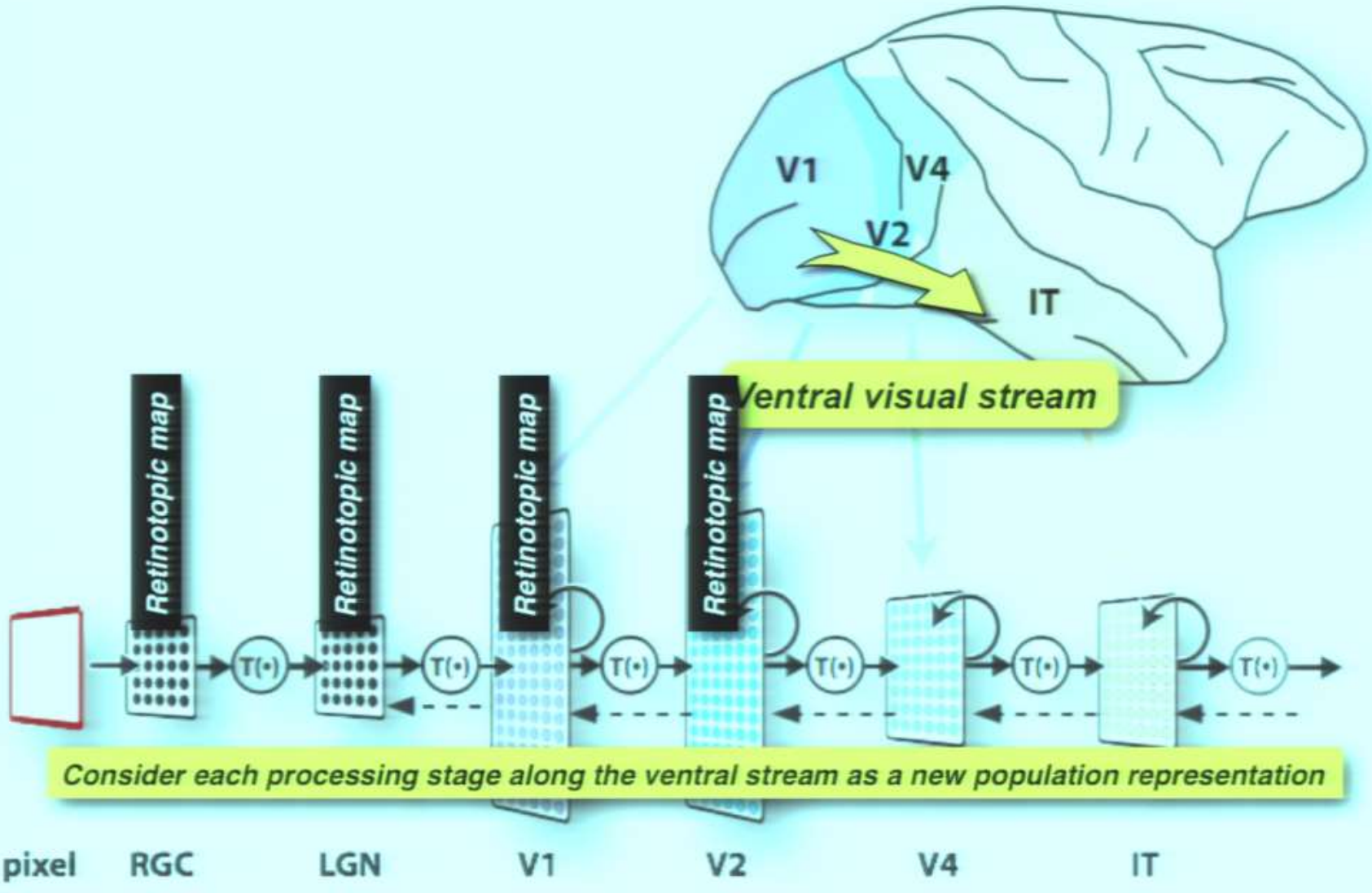
The ventral visual processing stream



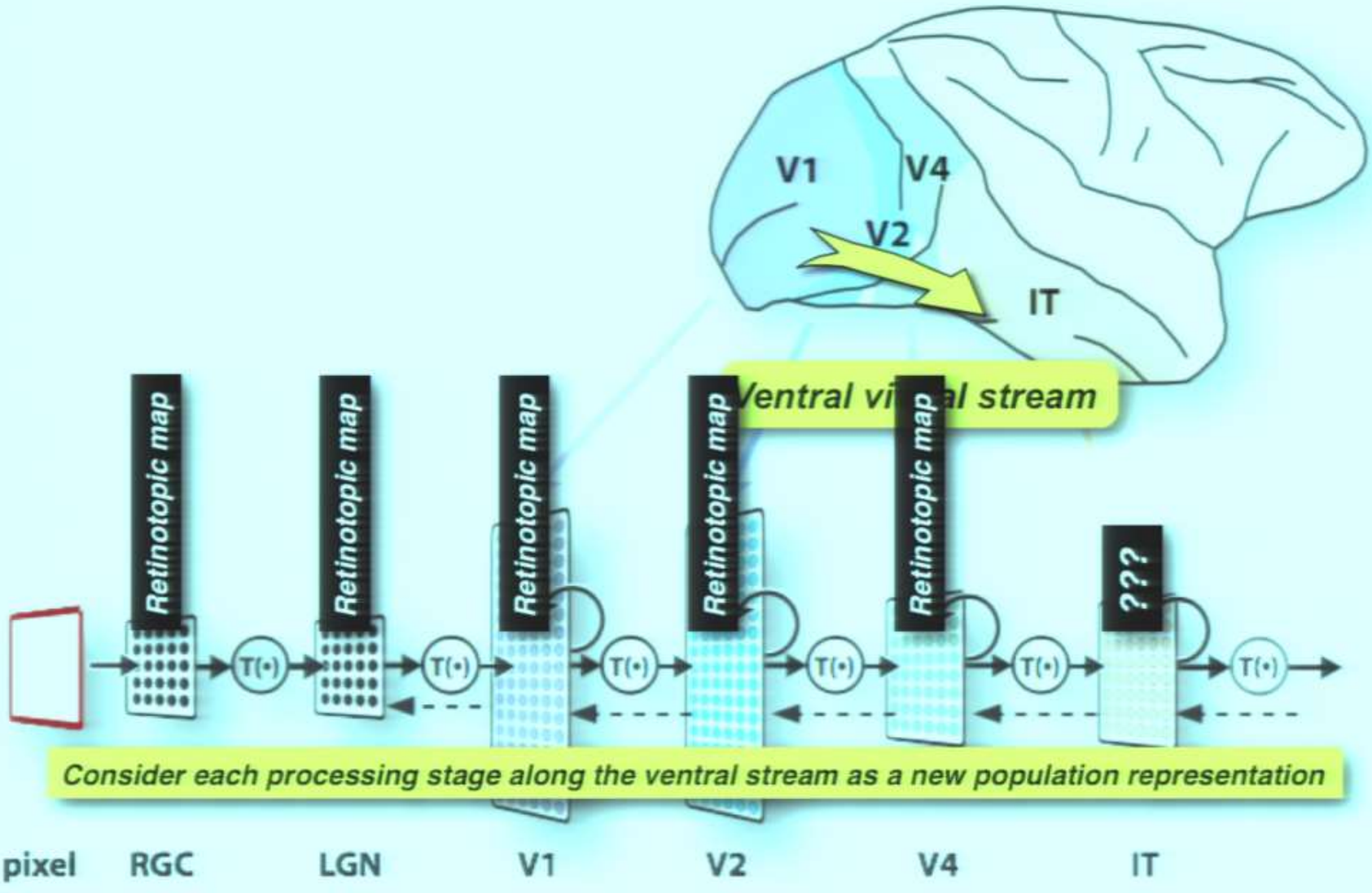
The ventral visual processing stream



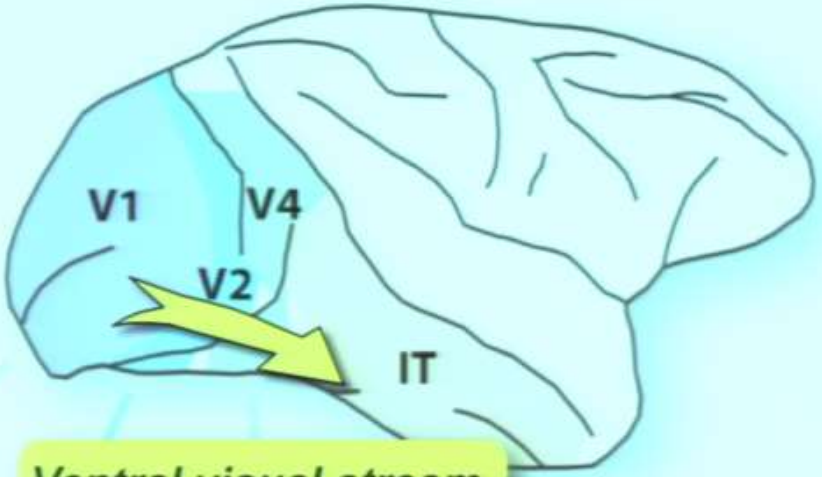
The ventral visual processing stream

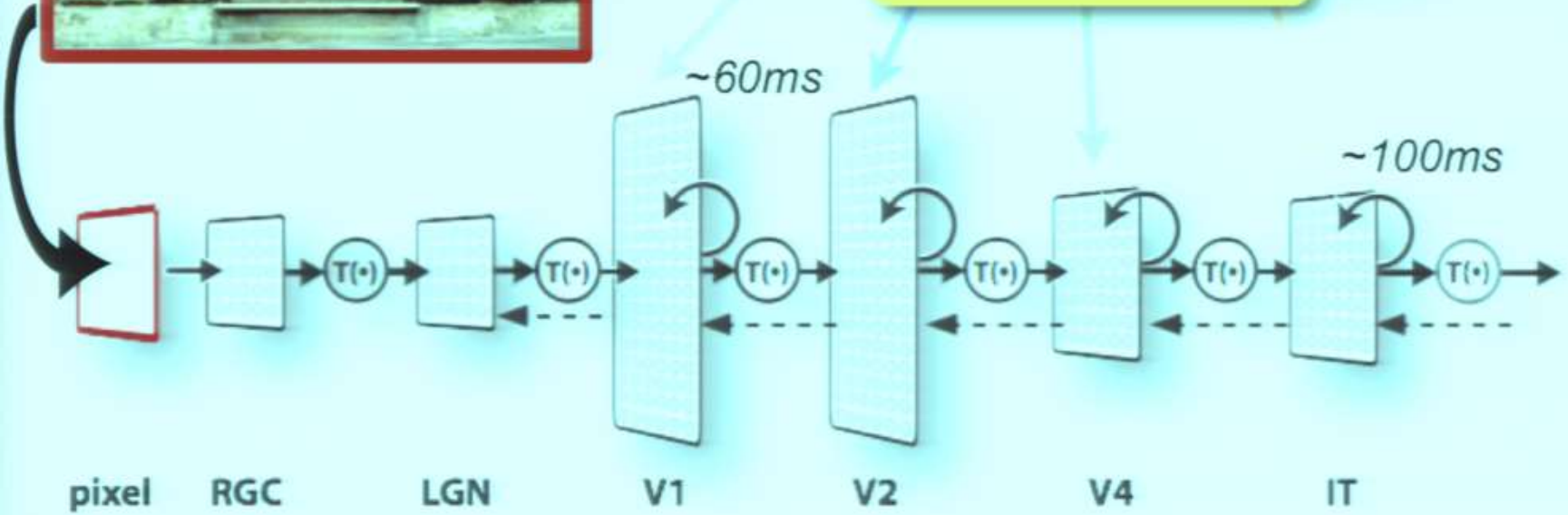


The ventral visual processing stream

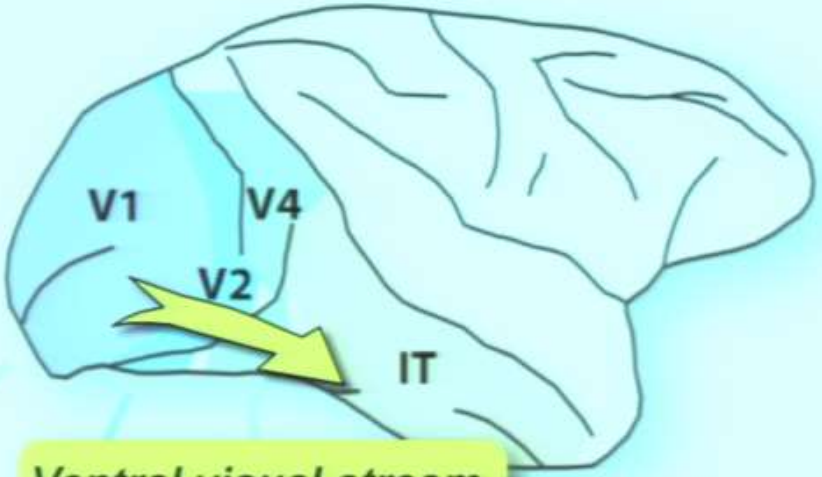


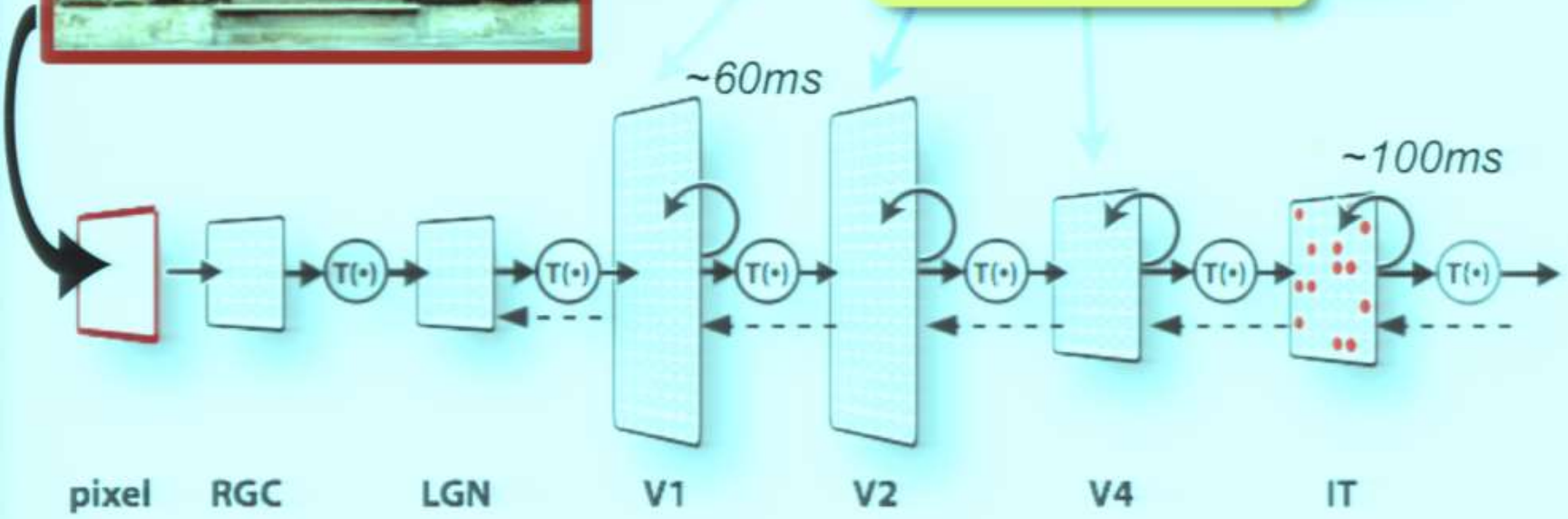
The ventral visual processing stream



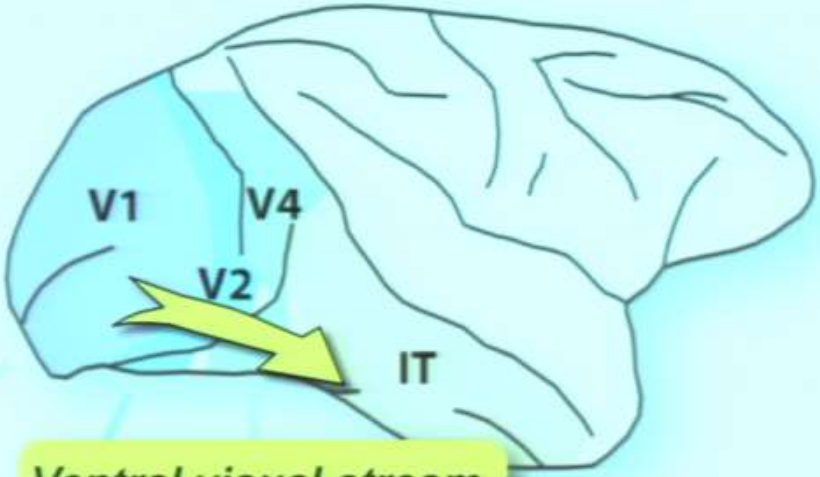


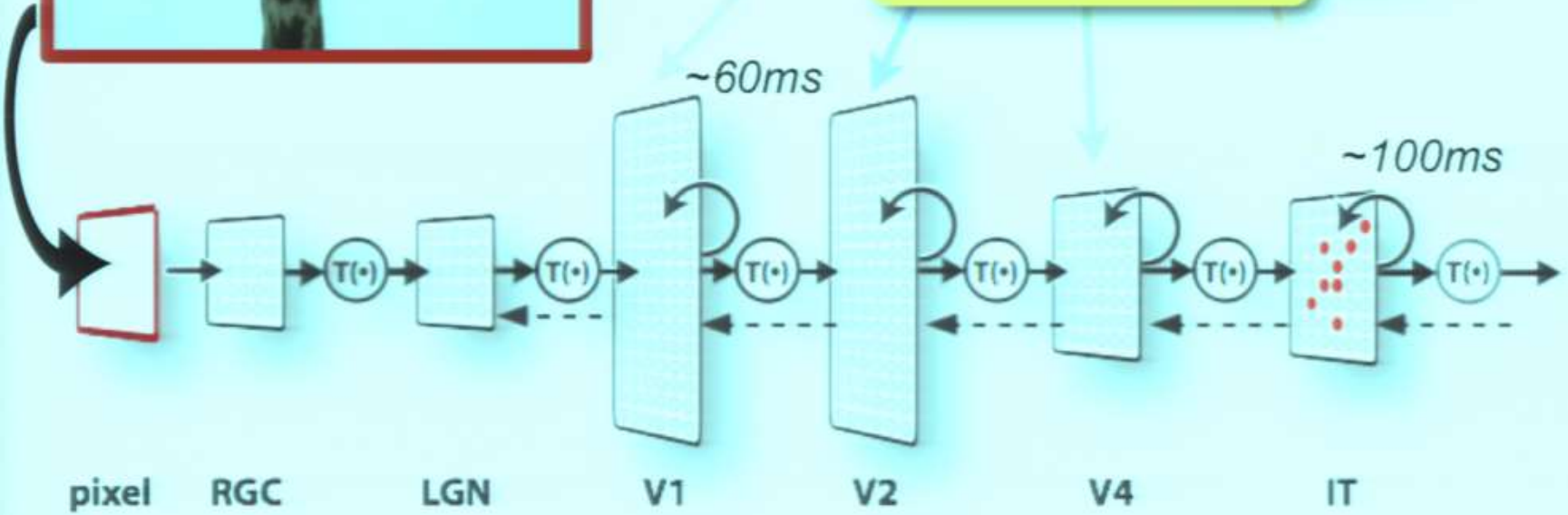
The ventral visual processing stream



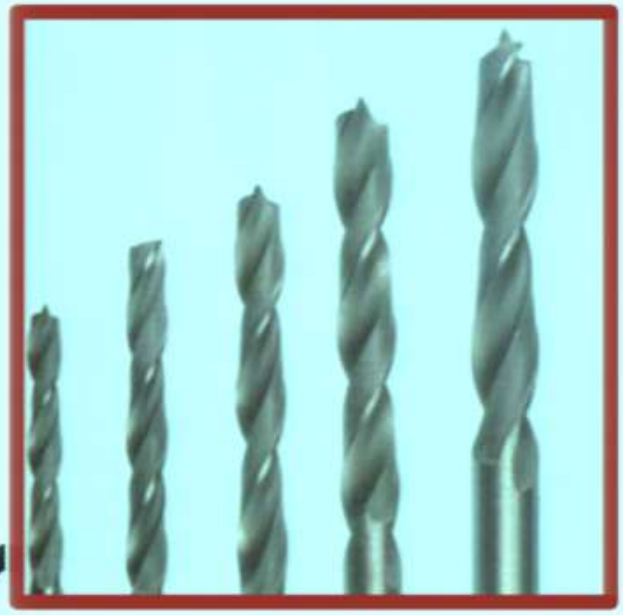


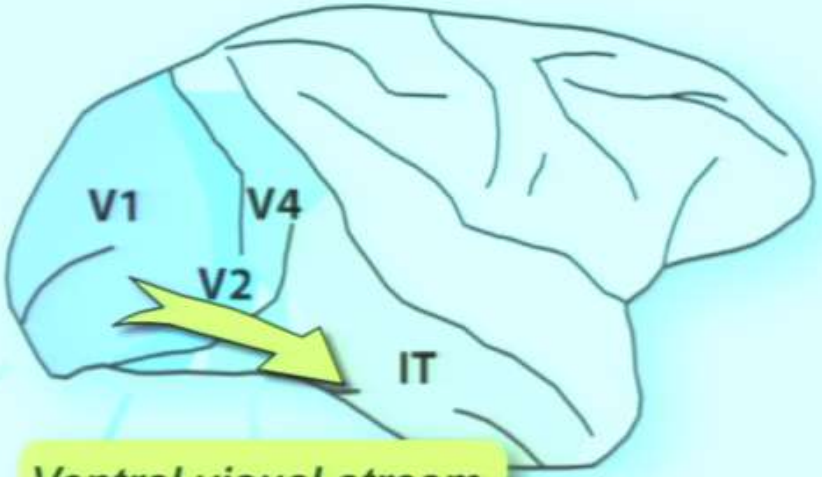
The ventral visual processing stream

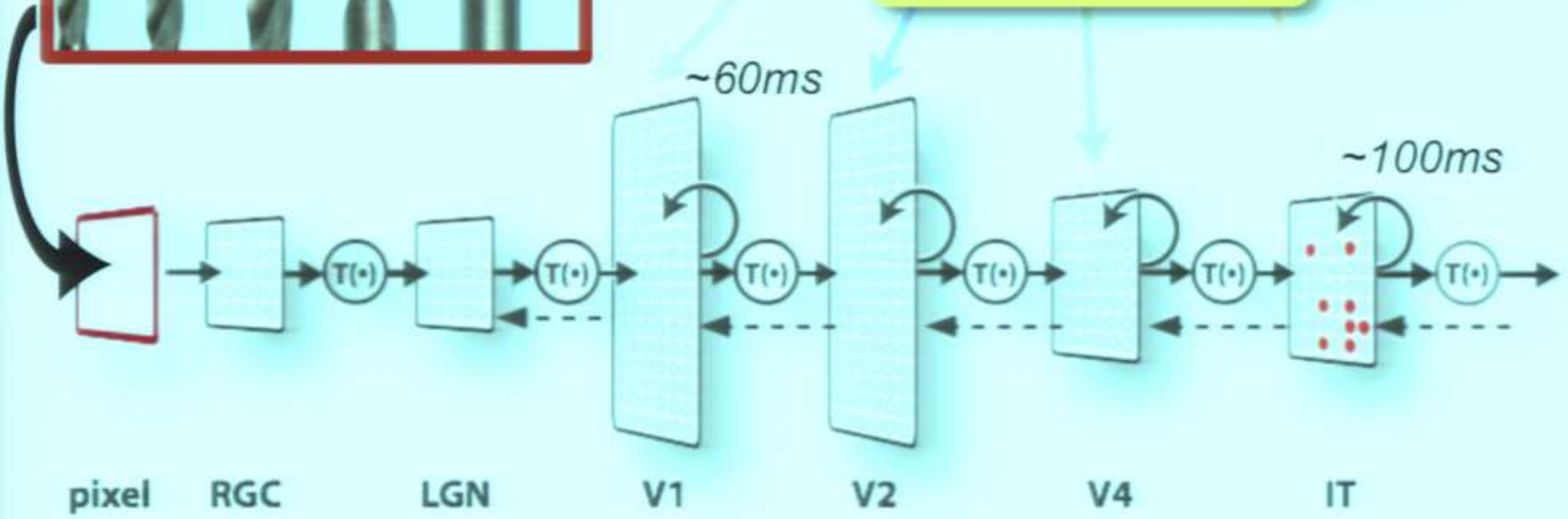




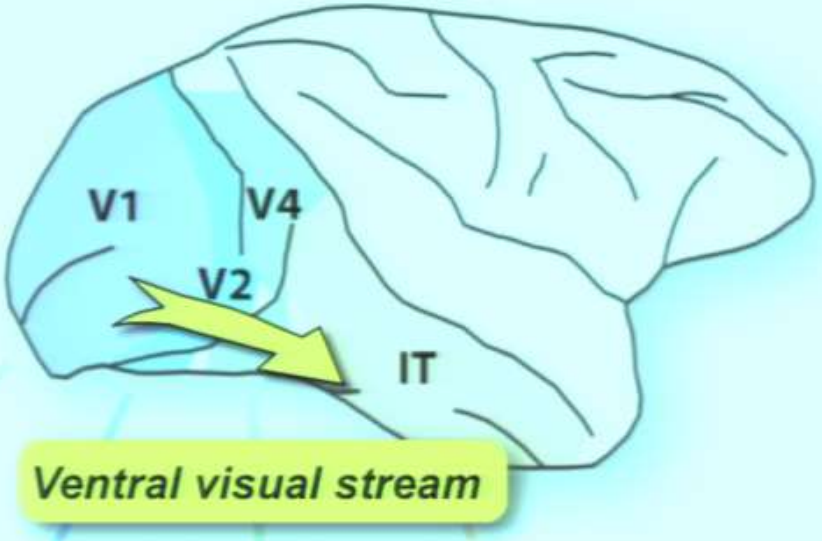
The ventral visual processing stream

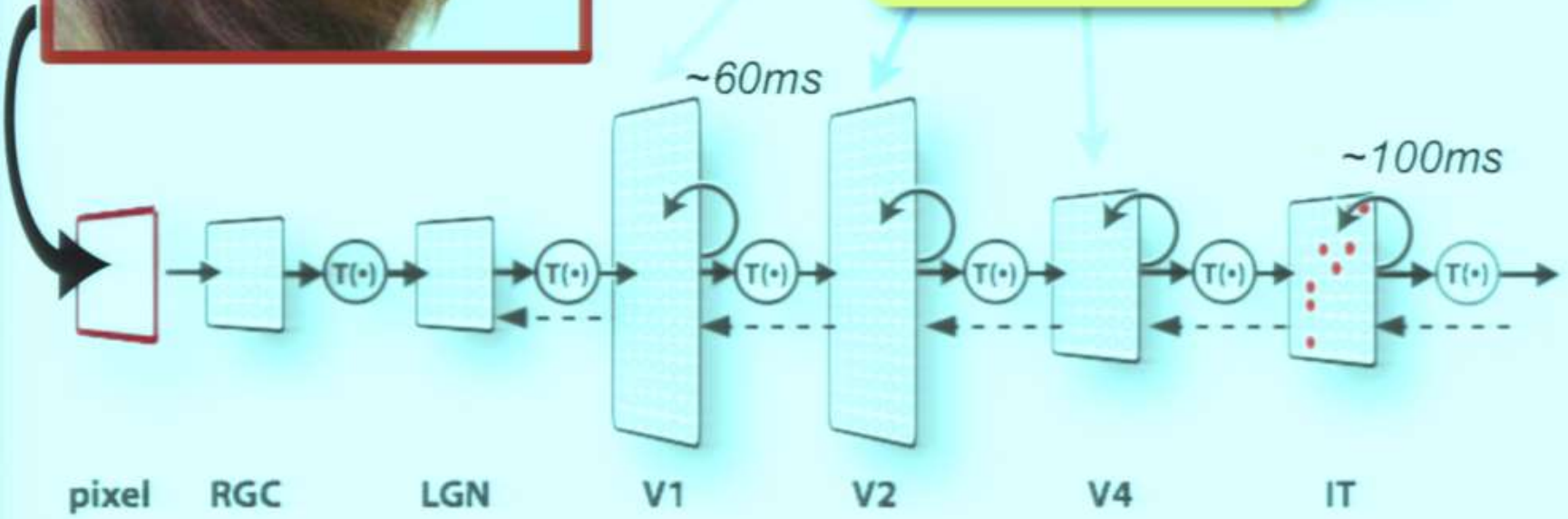




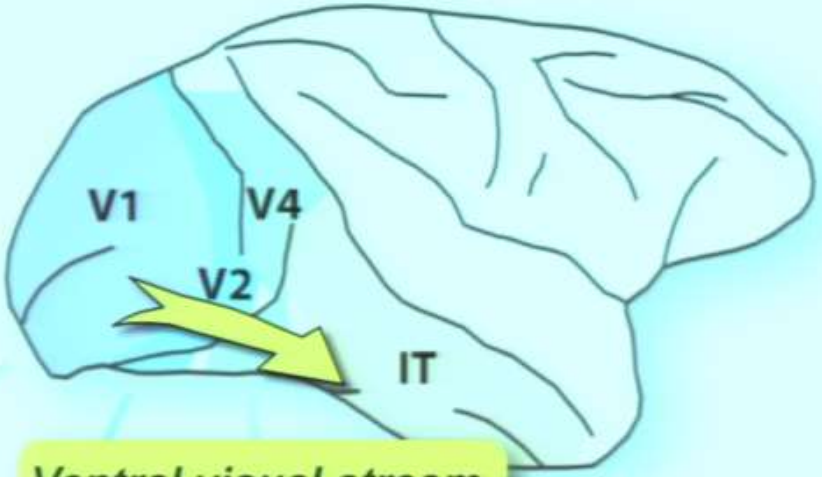


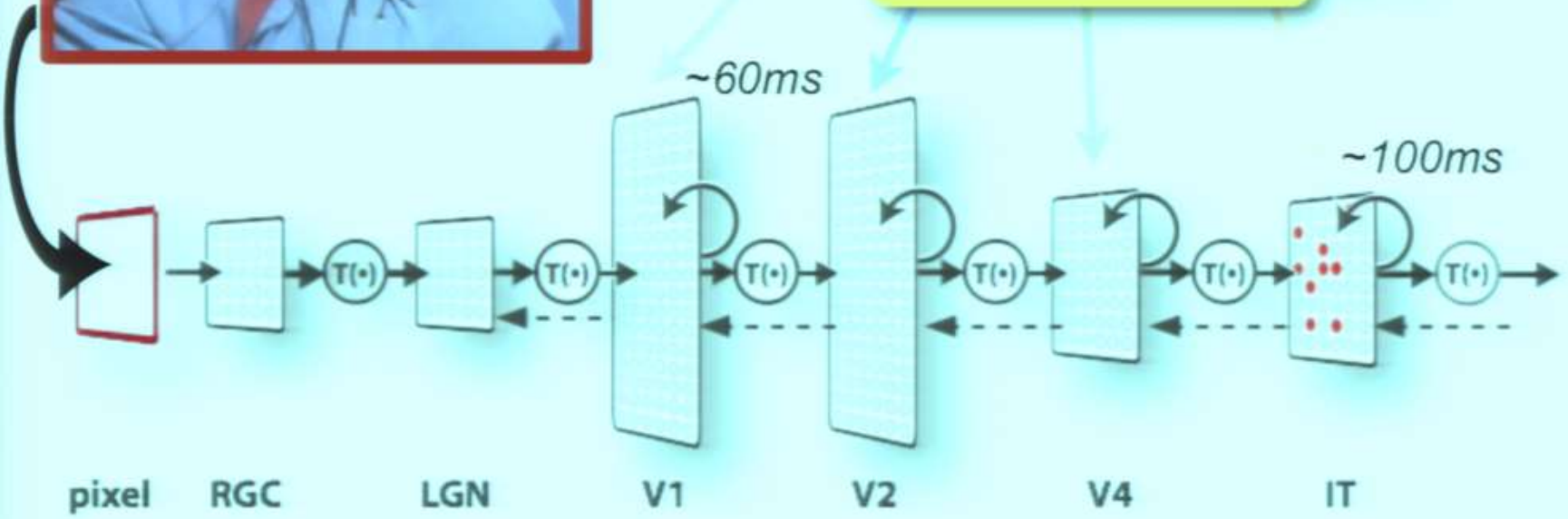
The ventral visual processing stream



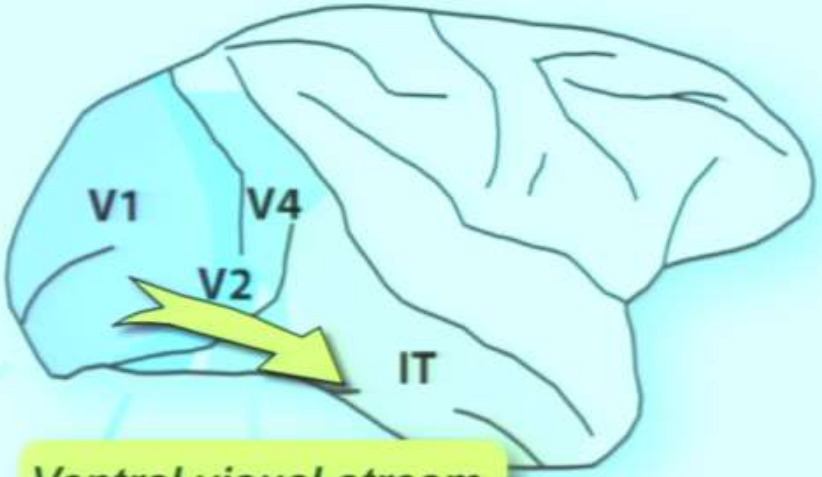


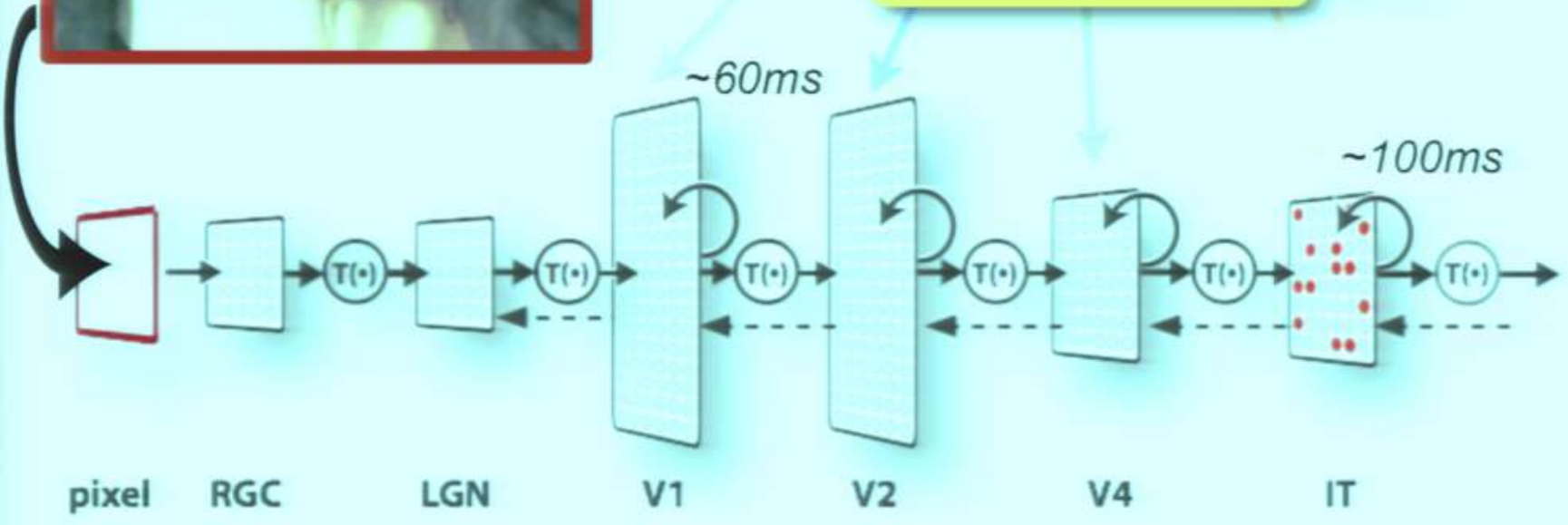
The ventral visual processing stream



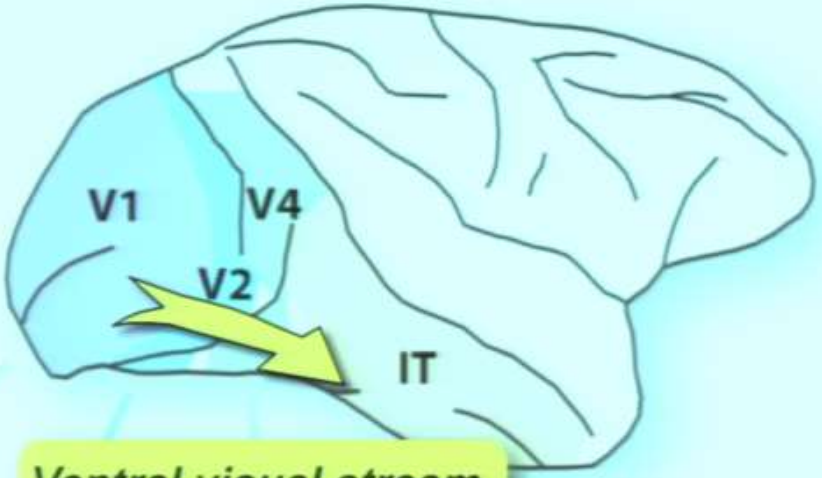


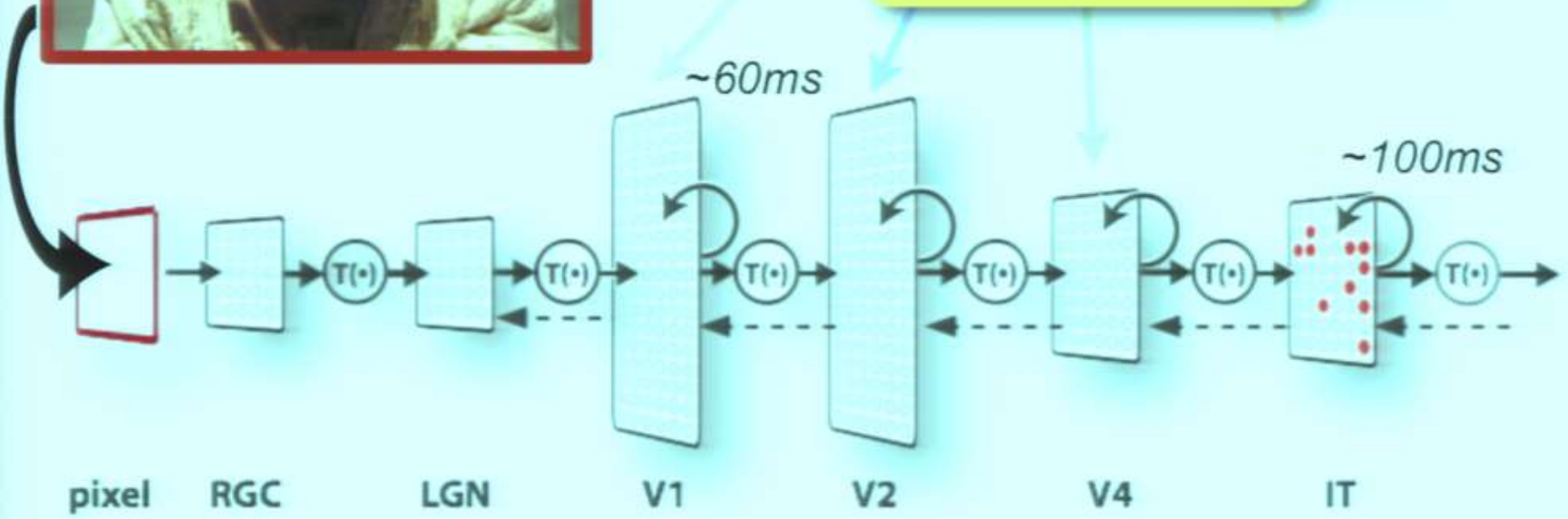
The ventral visual processing stream



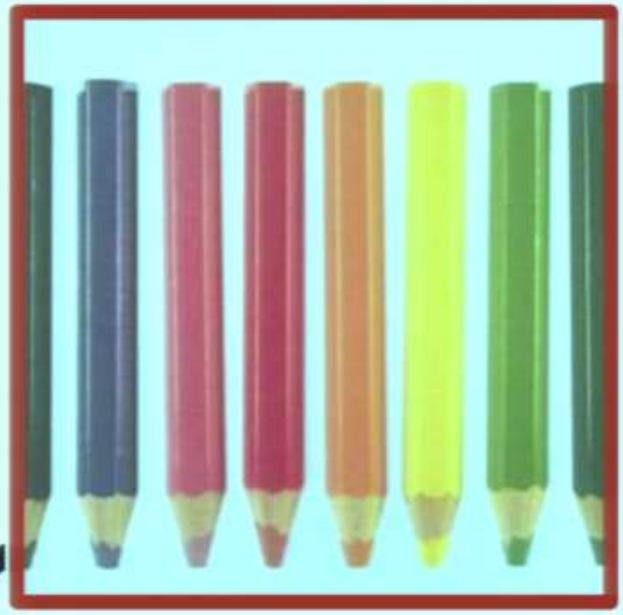


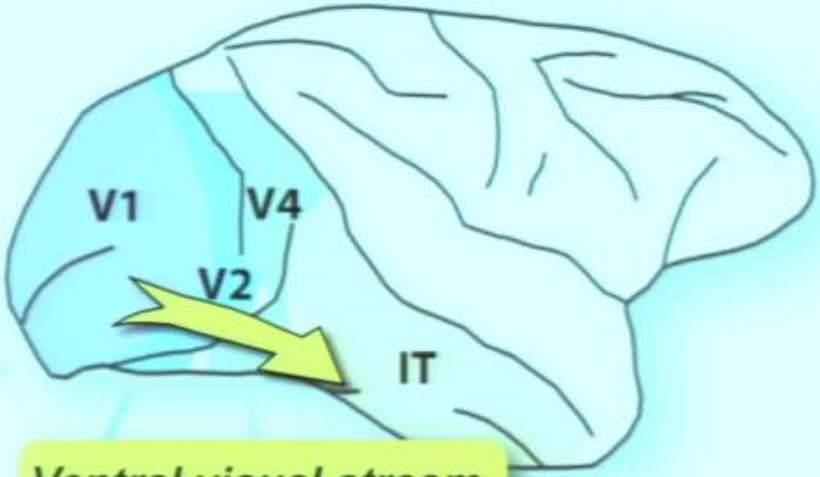
The ventral visual processing stream



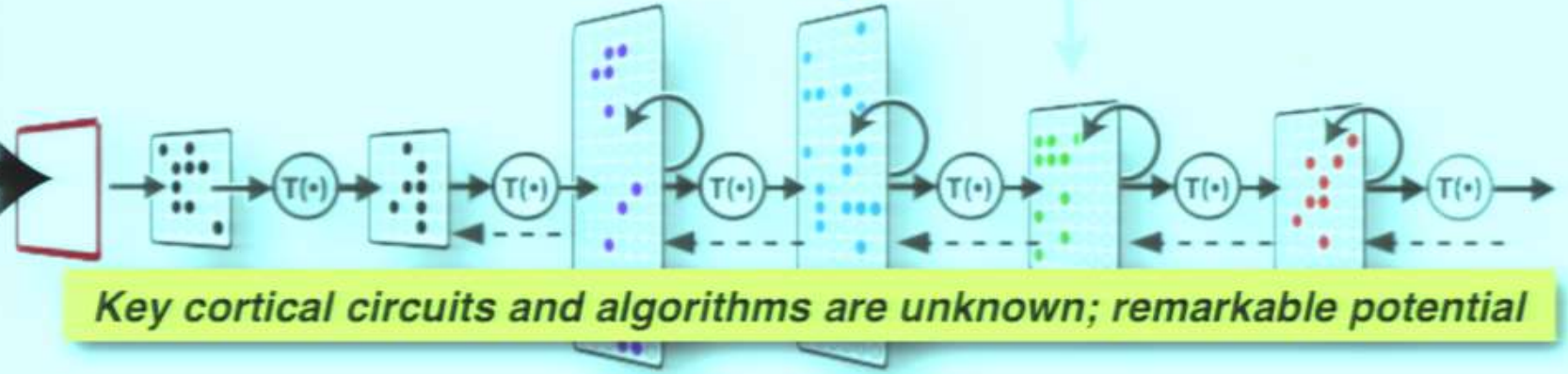


The ventral visual processing stream





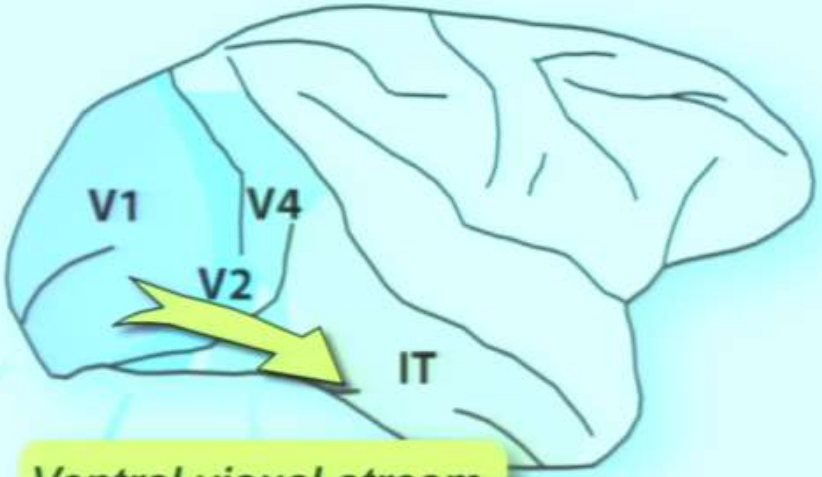
Ventral visual stream



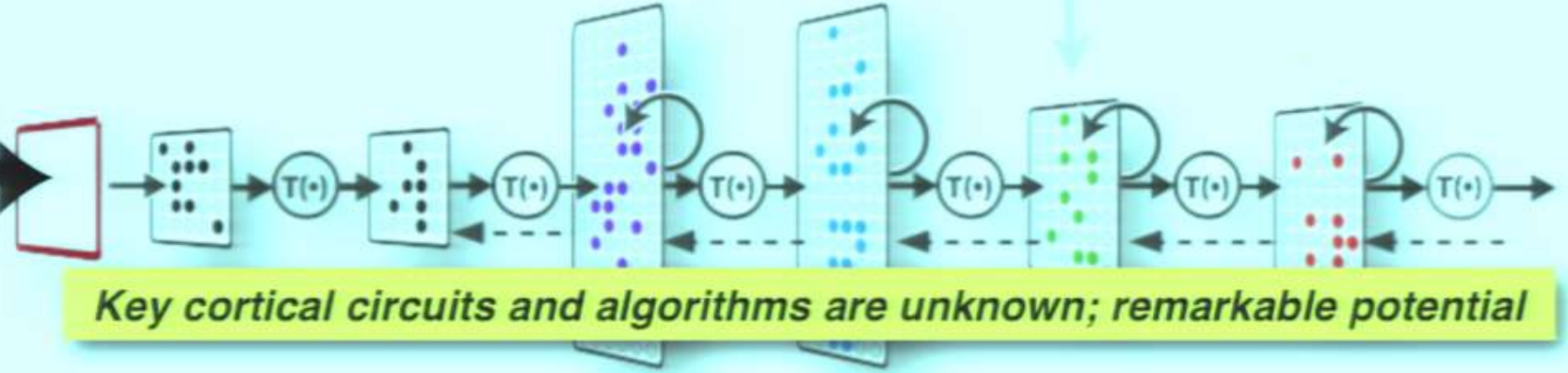
Key cortical circuits and algorithms are unknown; remarkable potential

pixel RGC LGN V1 V2 V4 IT

The ventral visual processing stream



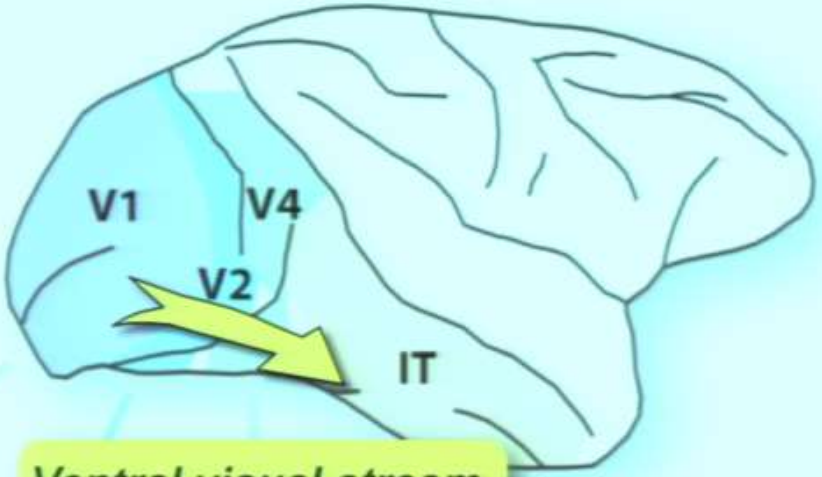
Ventral visual stream



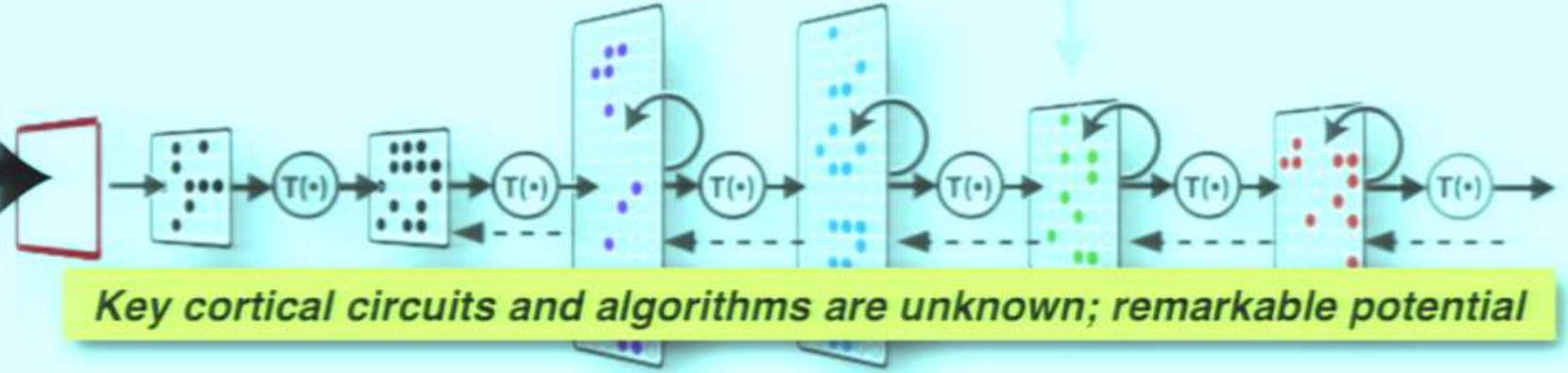
Key cortical circuits and algorithms are unknown; remarkable potential

pixel RGC LGN V1 V2 V4 IT

The ventral visual processing stream



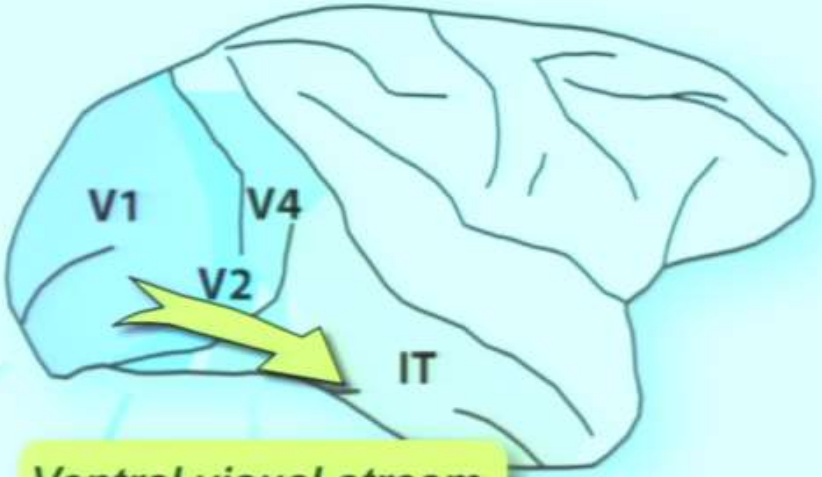
Ventral visual stream



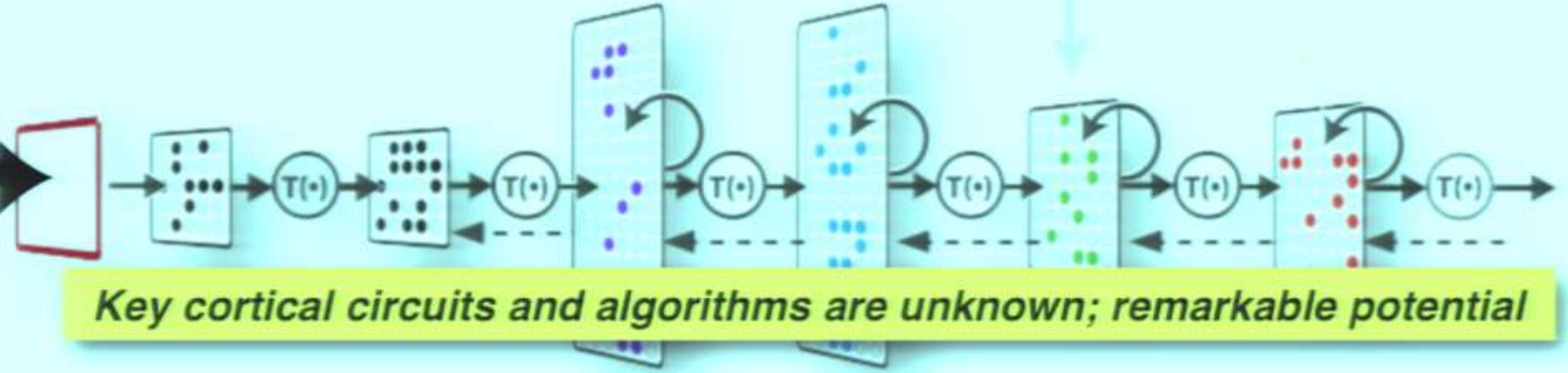
Key cortical circuits and algorithms are unknown; remarkable potential

pixel RGC LGN V1 V2 V4 IT

The ventral visual processing stream



Ventral visual stream

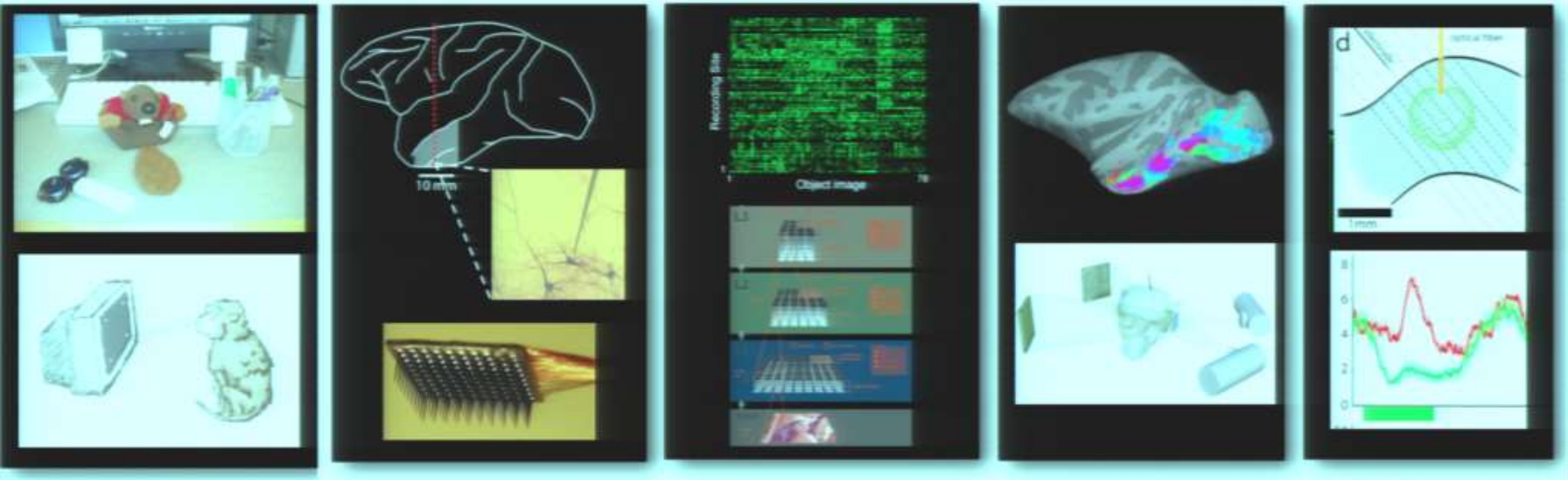


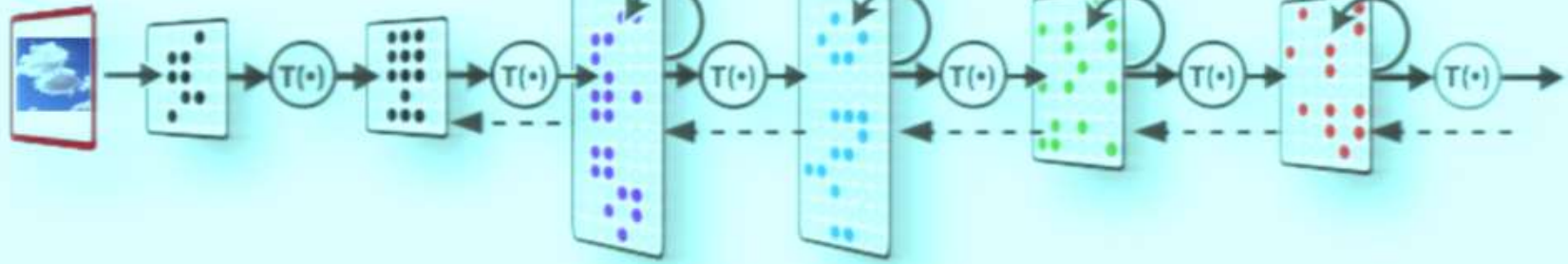
Key cortical circuits and algorithms are unknown; remarkable potential

pixel RGC LGN V1 V2 V4 IT

Our primary tools

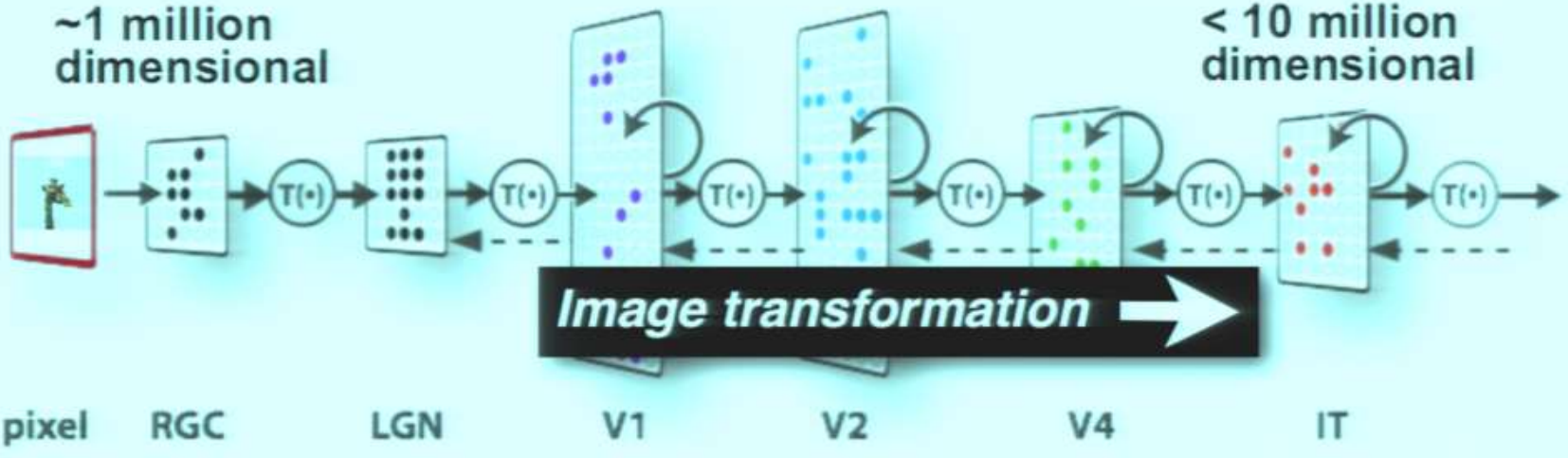
Psychophysics *Neurophysiology* *Computation* *Imaging* *Intervention*





Our primary question:

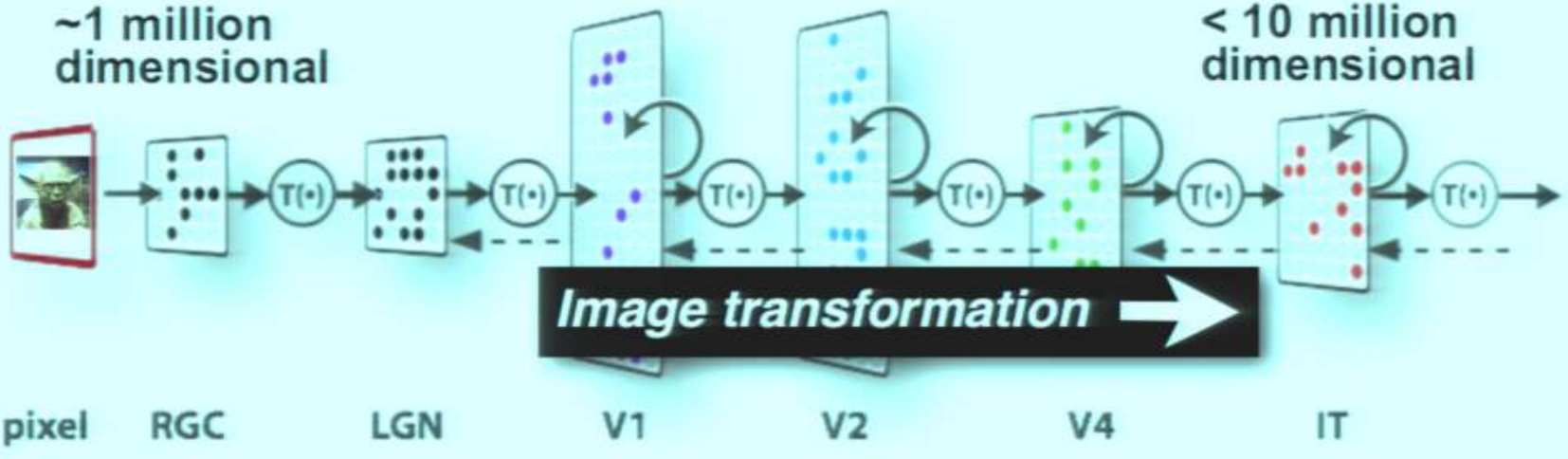
How do the circuits of the ventral stream transform the pixel image to solve object recognition ?



Our primary question:

How do the circuits of the ventral stream transform the pixel image to solve object recognition ?

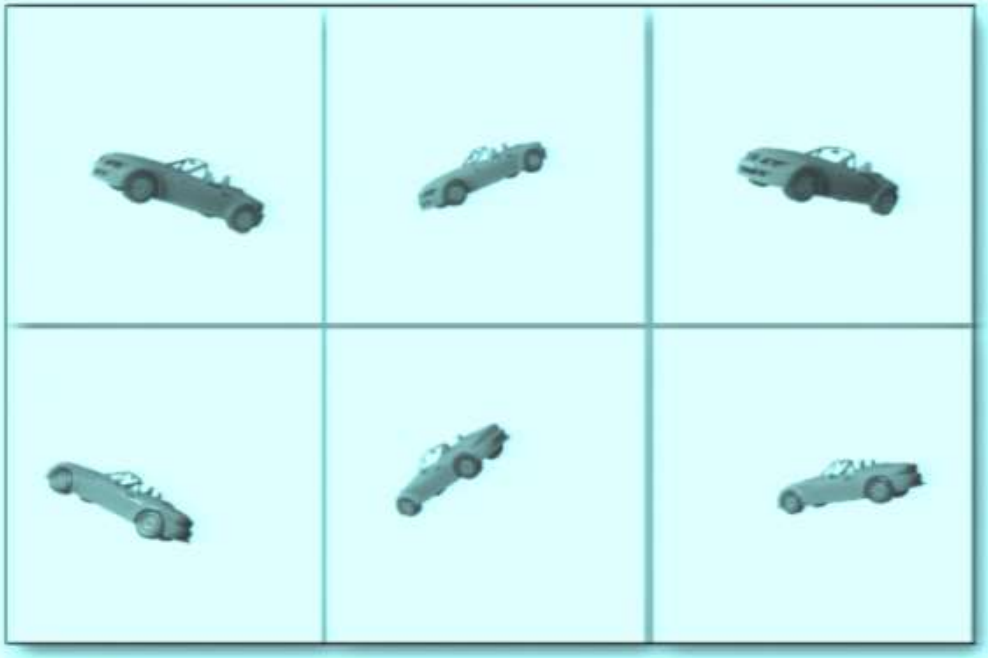
Why does the brain need to transform the pixel image ?



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

"Identity preserving image variation"

View: position, size, pose, illumination



Clutter, occlusion

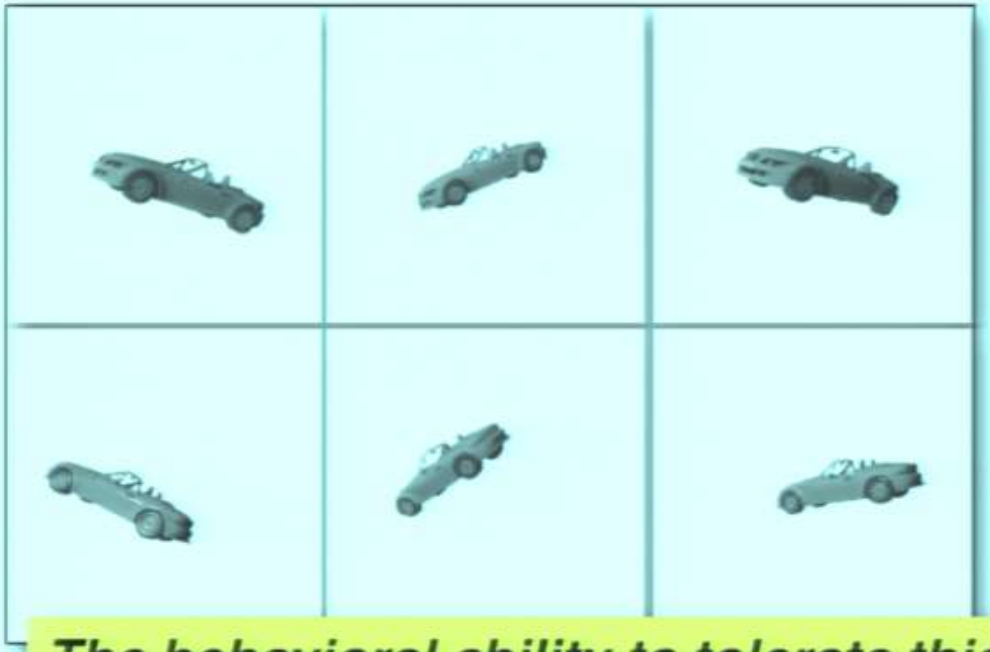
Intraclass

Poggio, Ullman, Grossberg, Edleman, Biederman, etc.
*DiCarlo and Cox, **TICS** (2007), Pinto, Cox, and DiCarlo, **PLoS Comp Bio** (2008),*
*DiCarlo, Zoccolan and Rust, **Neuron** (2012)*

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

"Identity preserving image variation"

View: position, size, pose, illumination



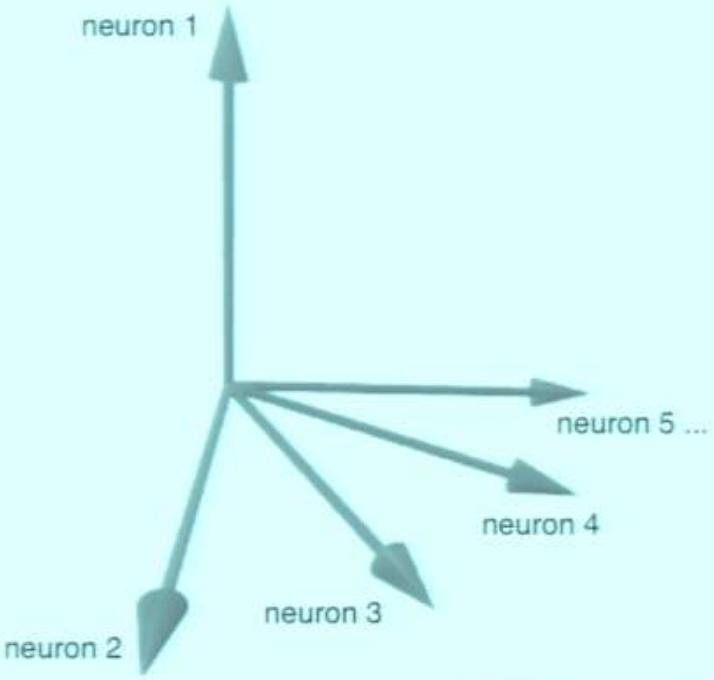
Clutter, occlusion

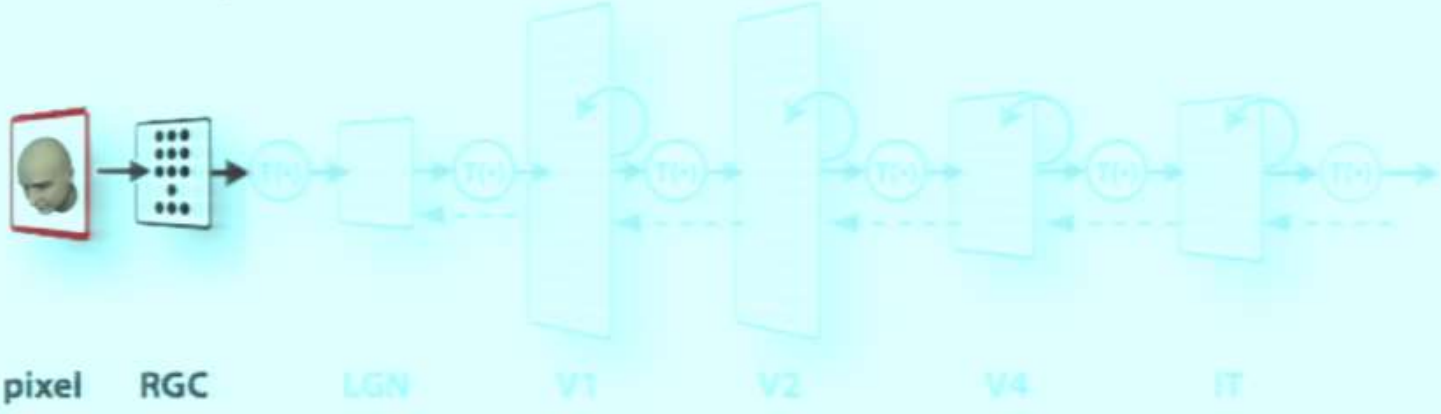
The behavioral ability to tolerate this is called "invariant" object recognition

Intraclass

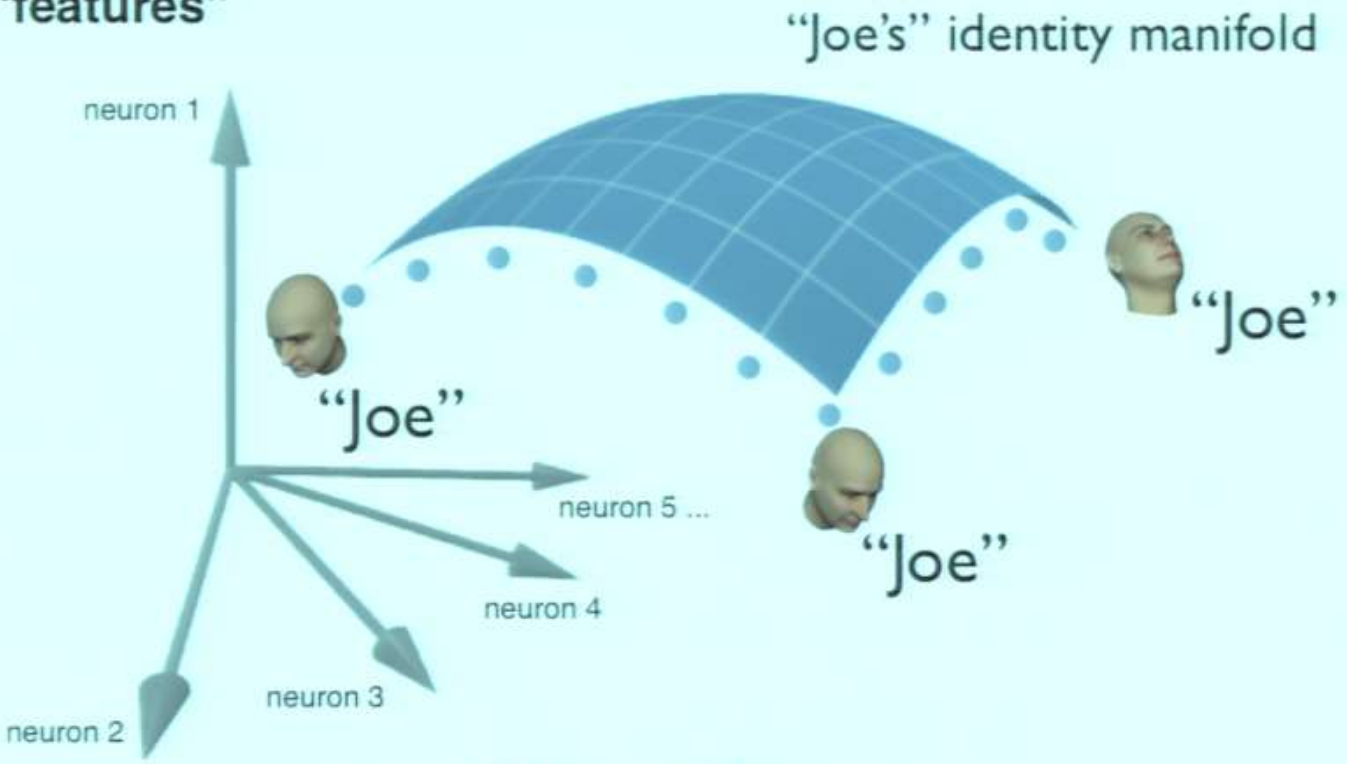
Poggio, Ullman, Grossberg, Edleman, Biederman, etc.
*DiCarlo and Cox, **TICS** (2007). Pinto, Cox, and DiCarlo, **PLoS Comp Bio** (2008).*
*DiCarlo, Zoccolan and Rust, **Neuron** (2012)*

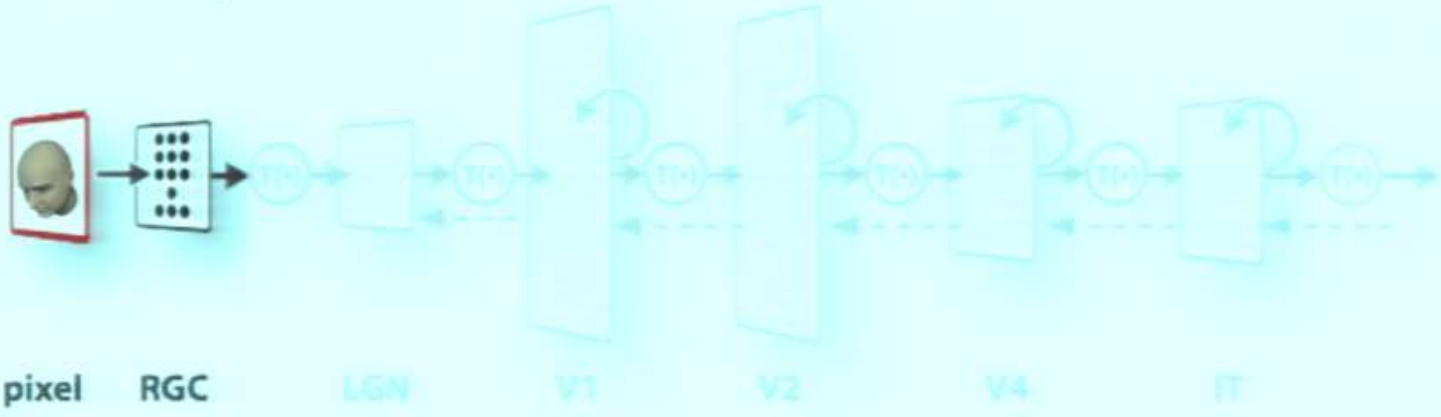
Neurons represent information as populations of visually-evoked "features"





Neurons represent information as populations of visually-evoked "features"

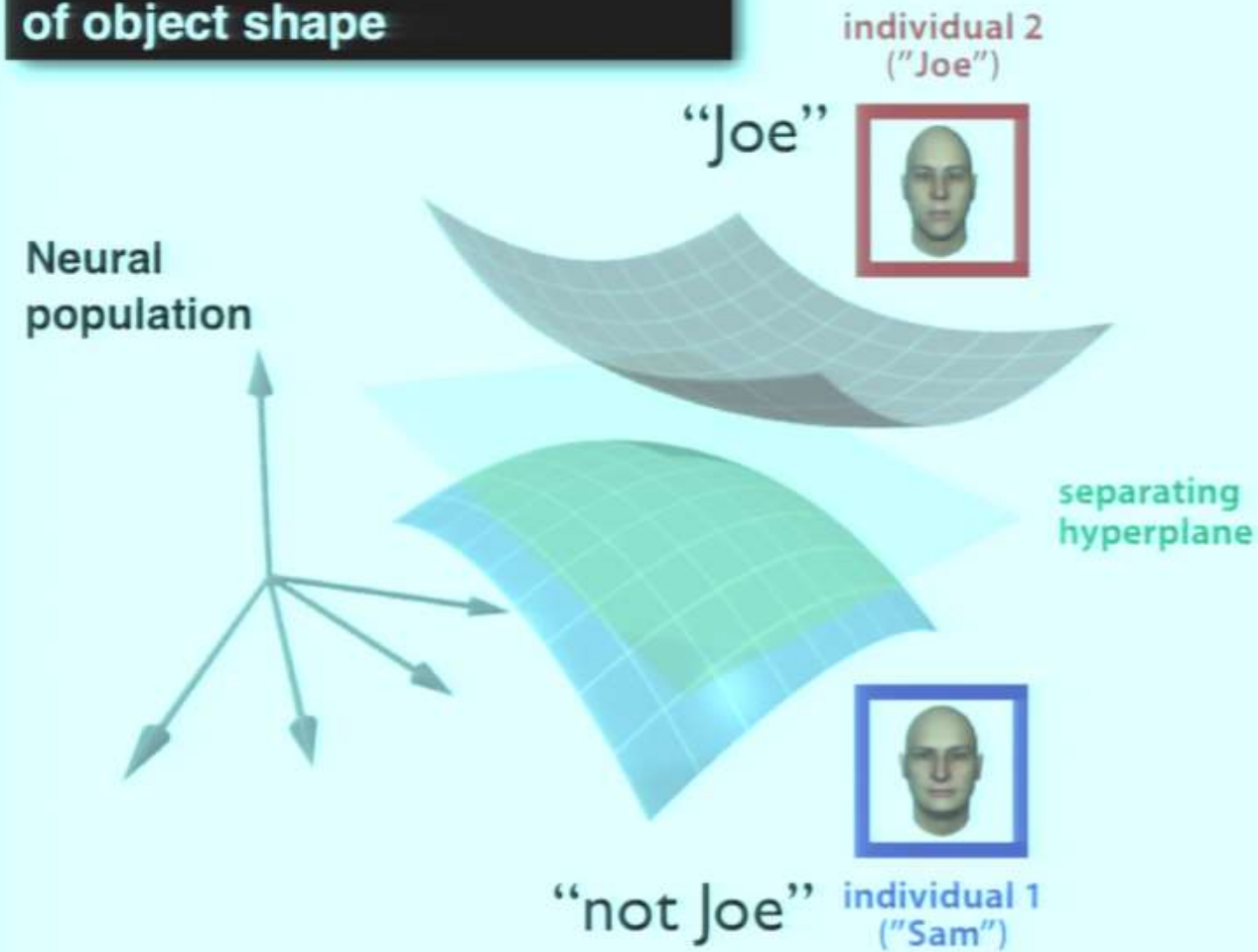




The computational crux of object and face recognition

A "good" set of visual features

== "Explicit" representation of object shape

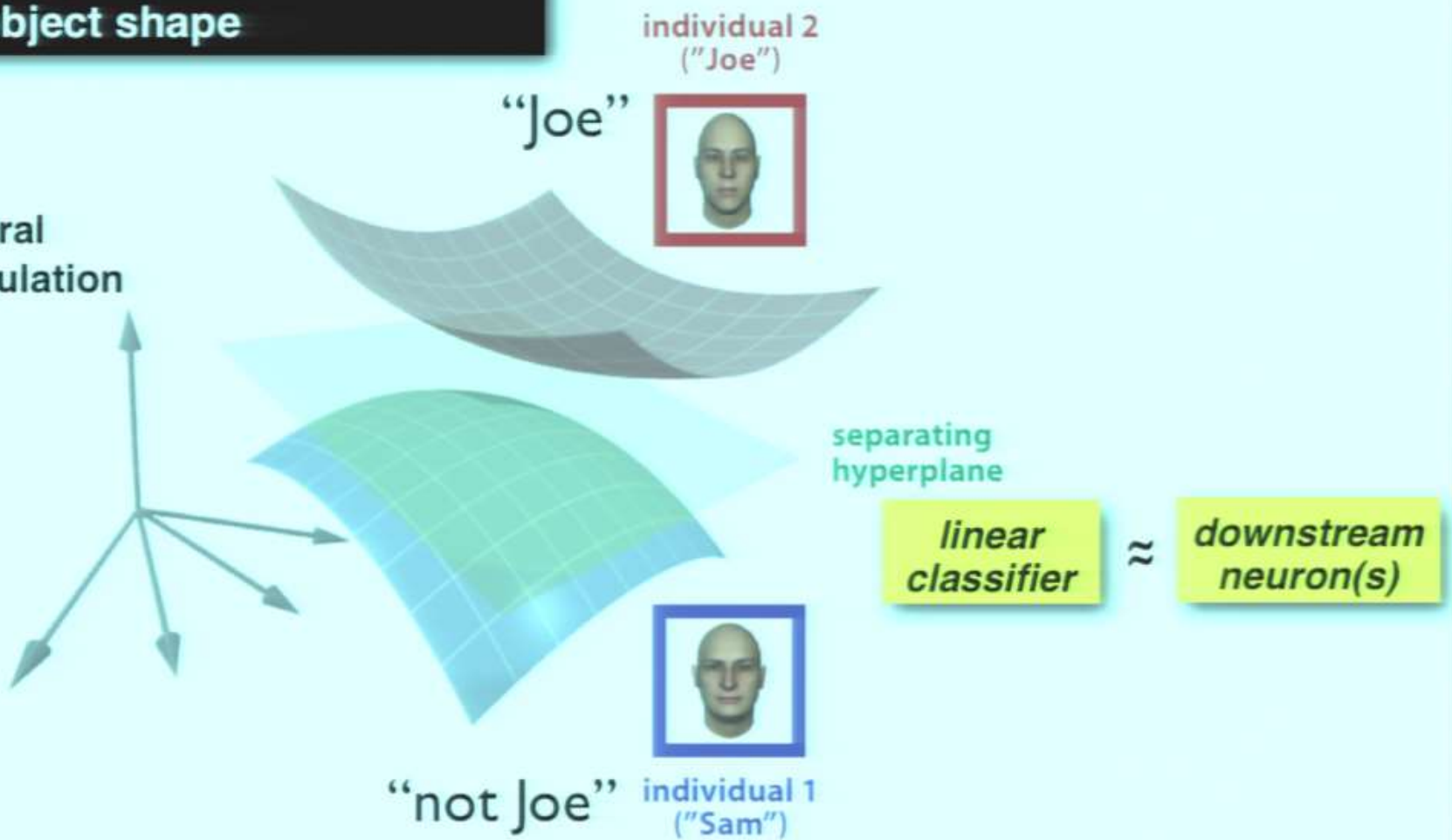


*DiCarlo and Cox, **TICS** (2007)*

The computational crux of object and face recognition

A "good" set of visual features

== "Explicit" representation of object shape



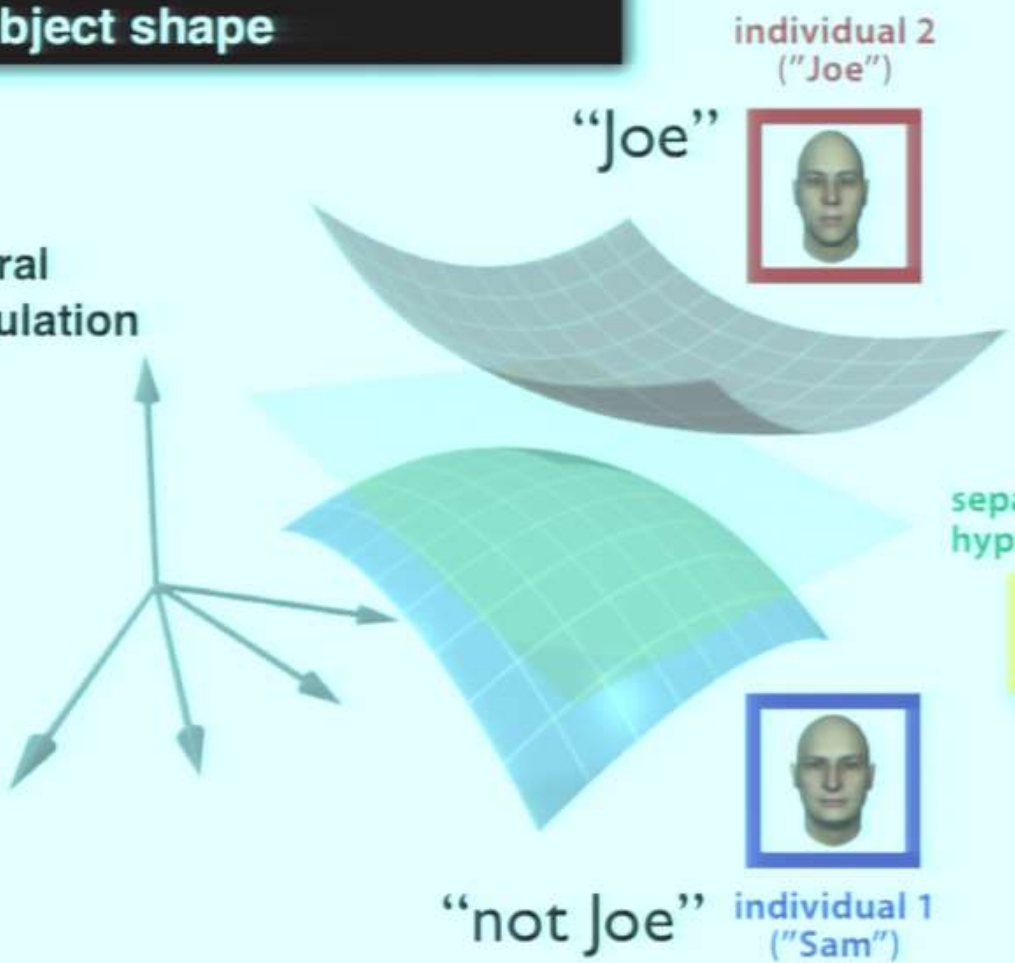
*DiCarlo and Cox, **TICS** (2007)*

The computational crux of object and face recognition

A "good" set of visual features

== "Explicit" representation of object shape

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



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

linear classifier ≈ *downstream neuron(s)*

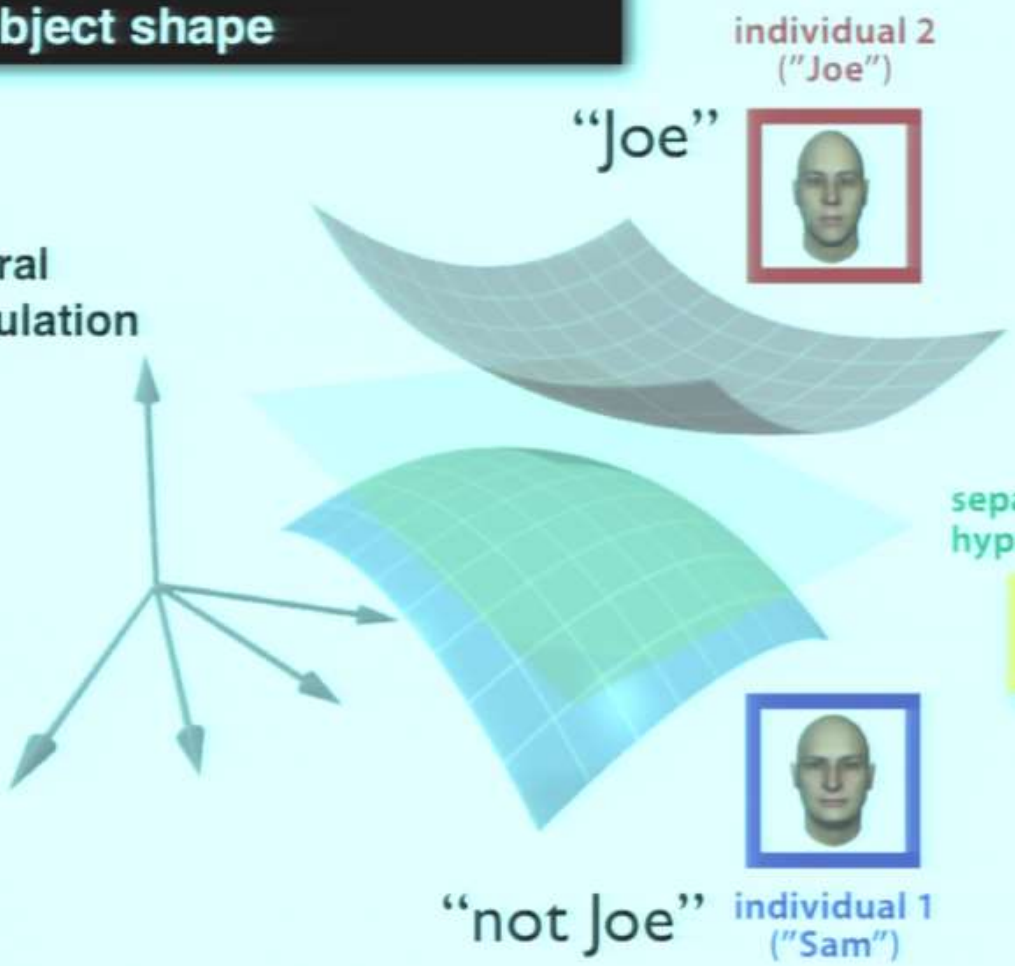
DiCarlo and Cox, **TICS** (2007)

The computational crux of object and face recognition

A "good" set of visual features

== "Explicit" representation of object shape

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

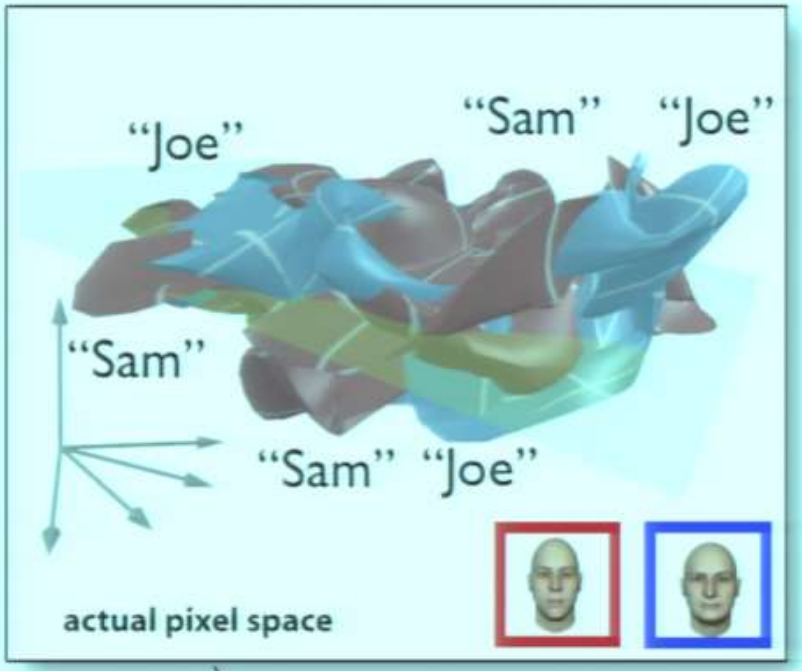


Should be able to find it with low* number of training examples

linear classifier ≈ *downstream neuron(s)*

Manifolds NOT required

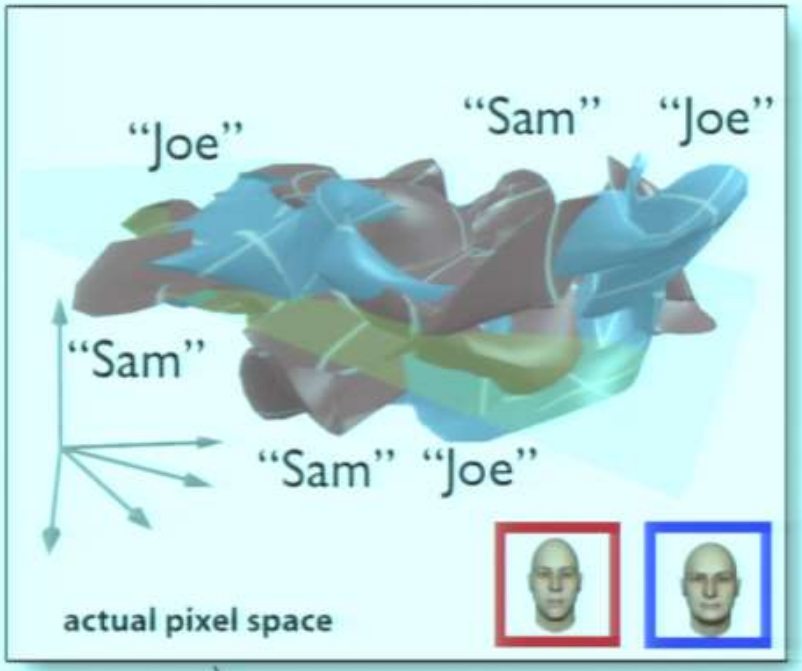
DiCarlo and Cox, **TICS** (2007)

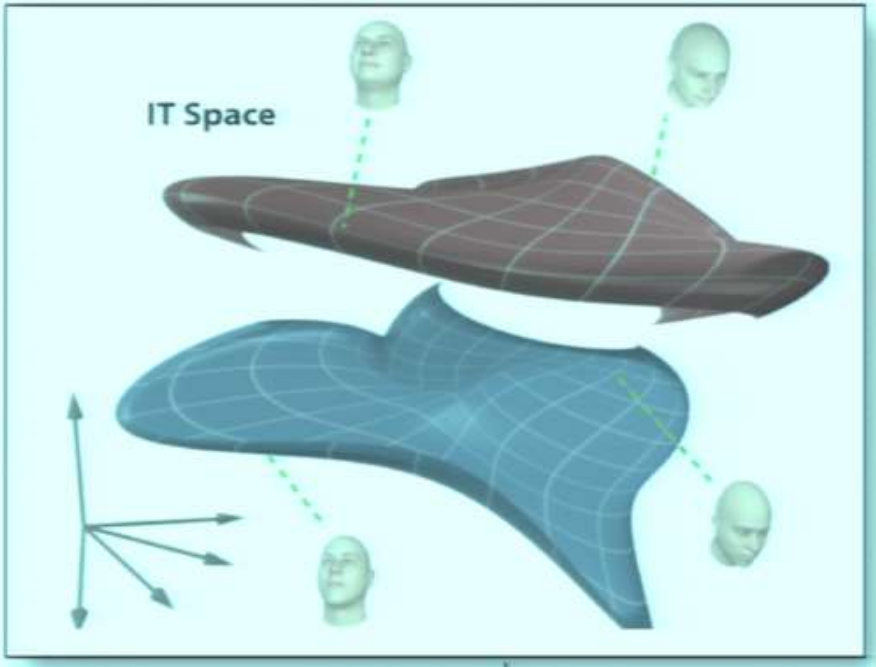


Tangled, implicit object information

pixel RGC LGN V1 V2 V4 IT

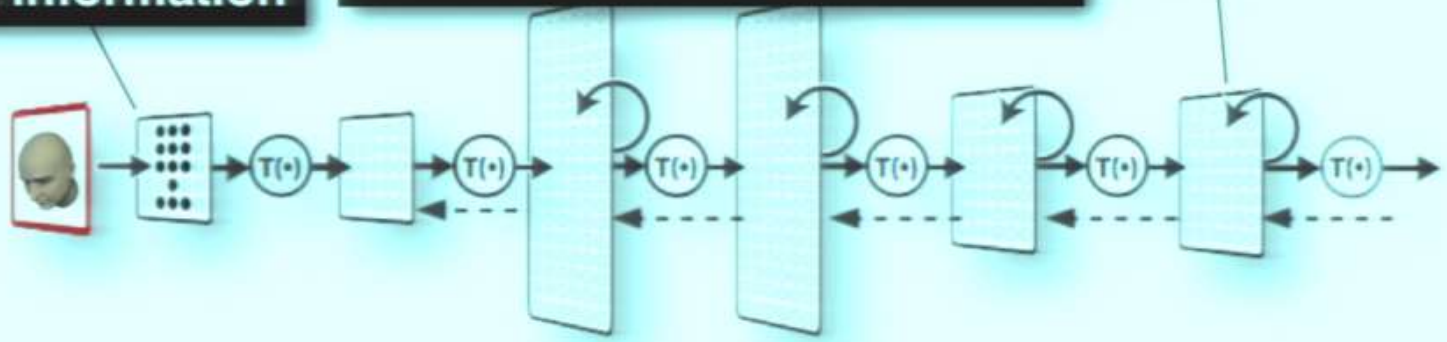
DiCarlo and Cox, **TICS** (2007)
DiCarlo, Zoccolan and Rust, **Neuron** (2012)



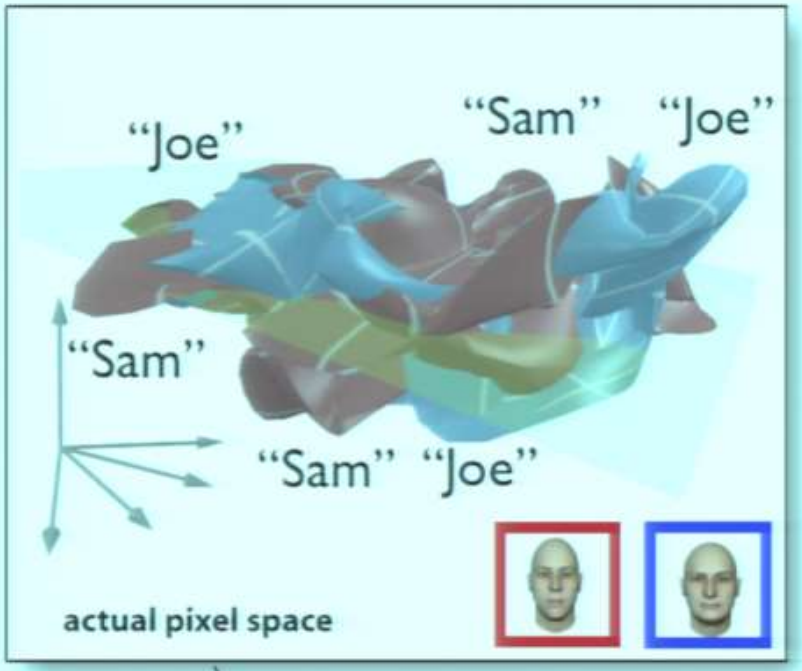


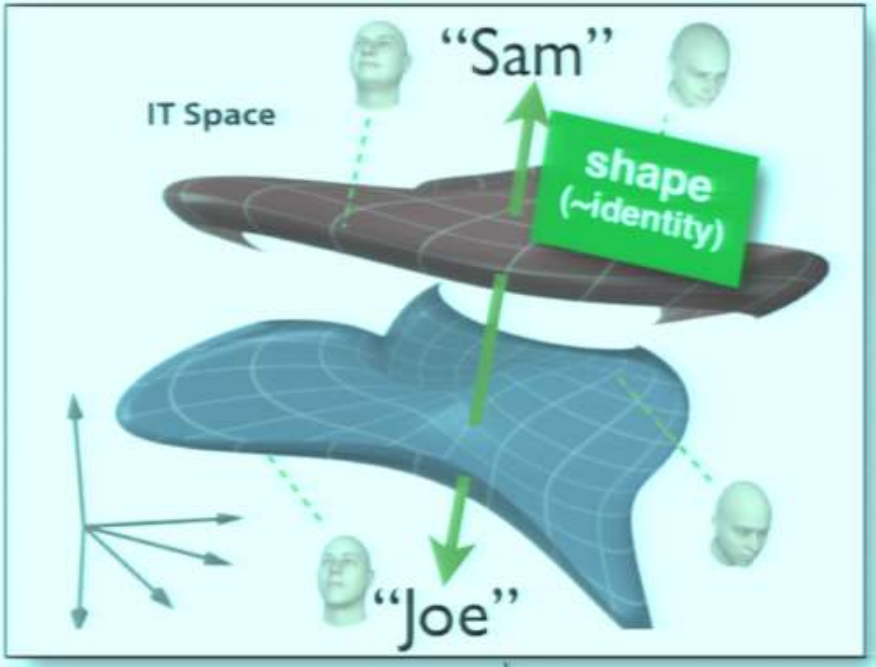
Tangled, implicit object information

Transformation ⟶



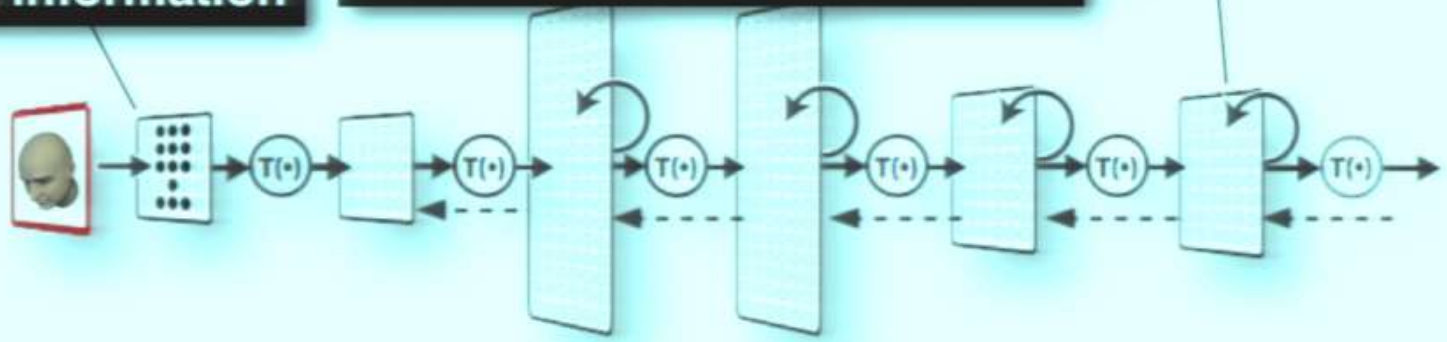
DiCarlo and Cox, **TICS** (2007)
DiCarlo, Zoccolan and Rust, **Neuron** (2012)



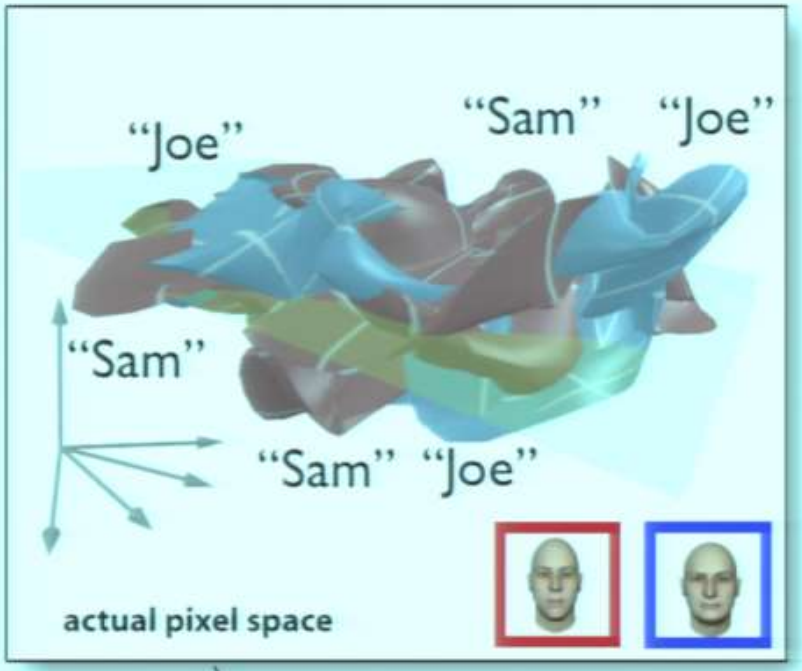


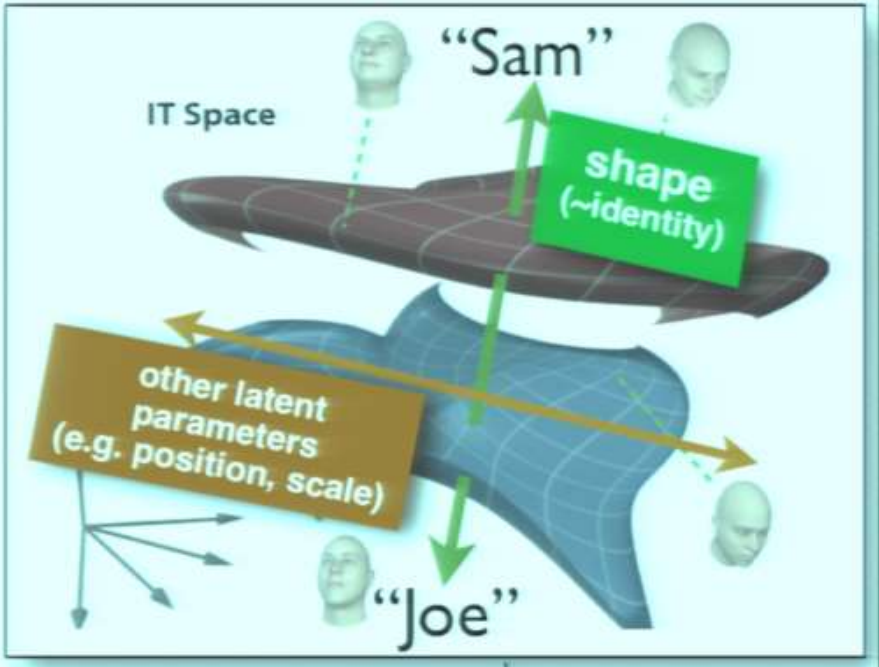
Tangled, implicit object information

Transformation ⟶



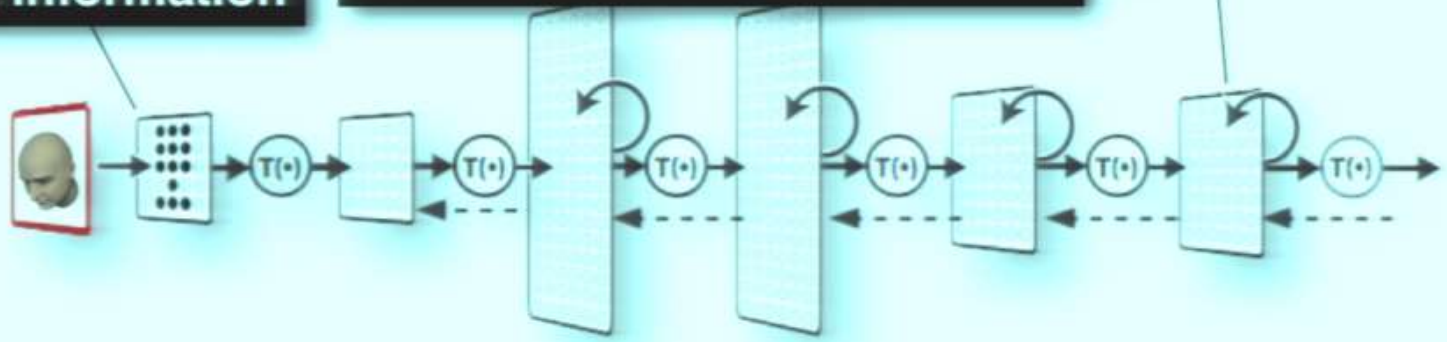
DiCarlo and Cox, **TICS** (2007)
DiCarlo, Zoccolan and Rust, **Neuron** (2012)



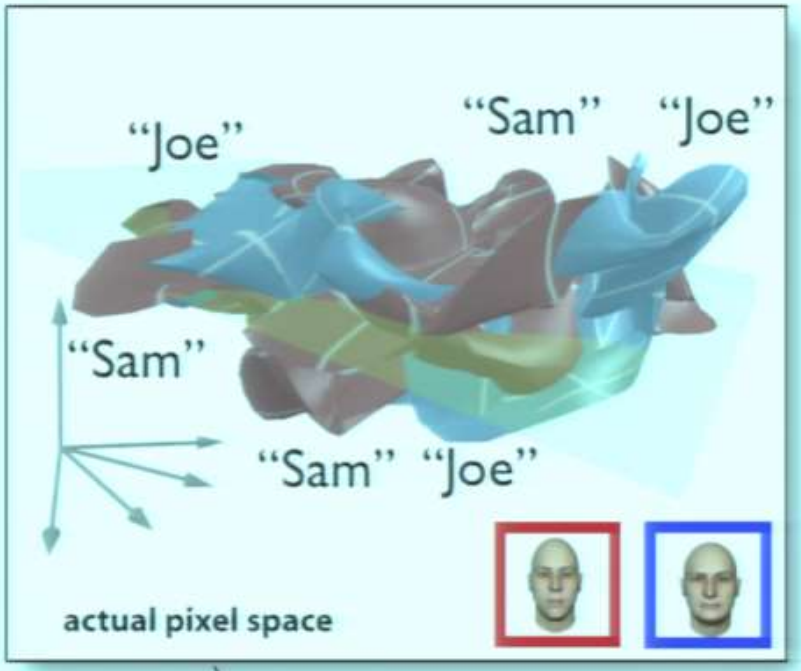


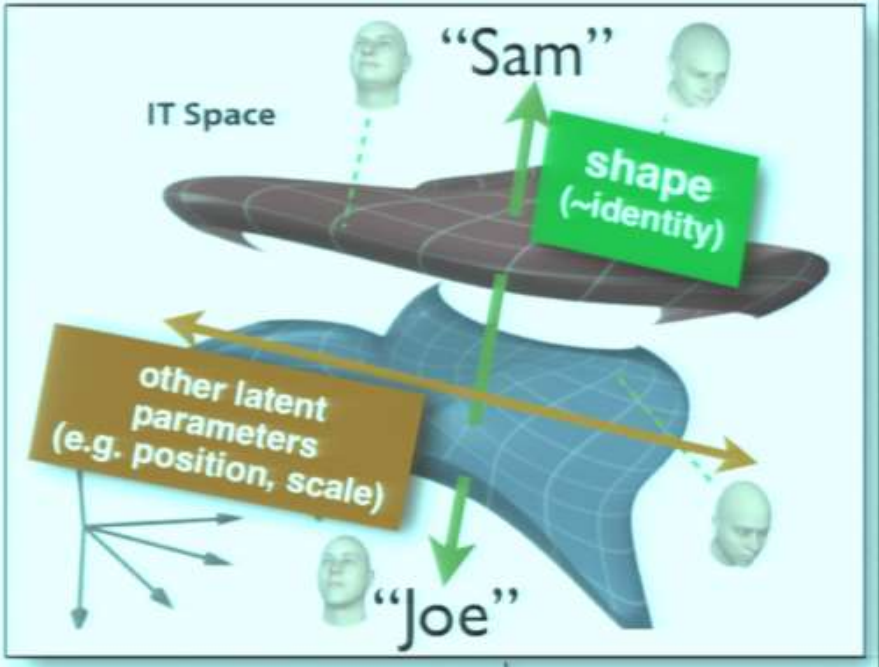
Tangled, implicit object information

Transformation ⟶



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

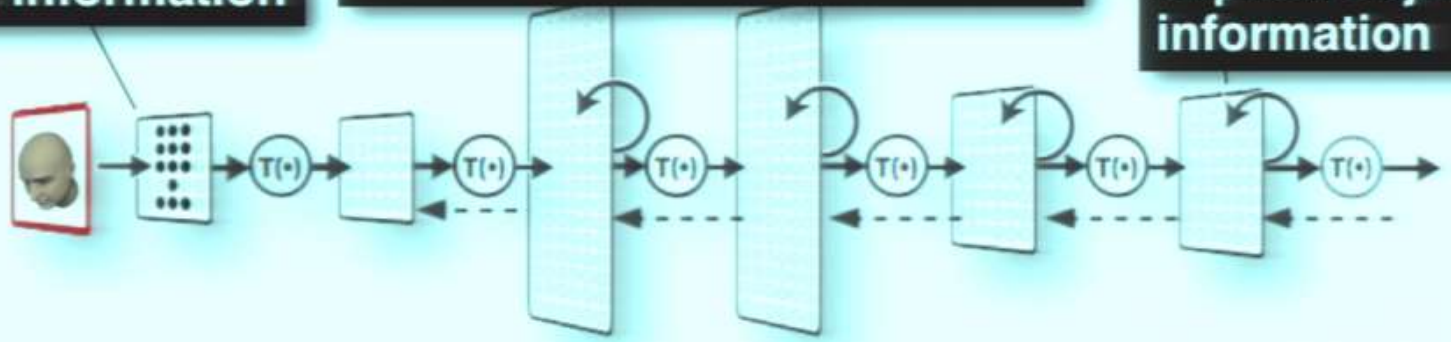




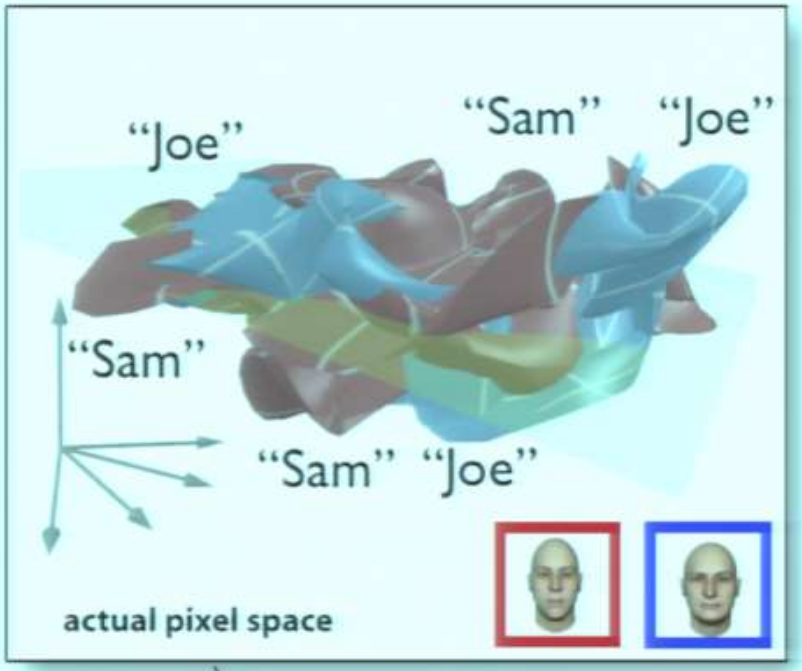
Tangled, implicit object information

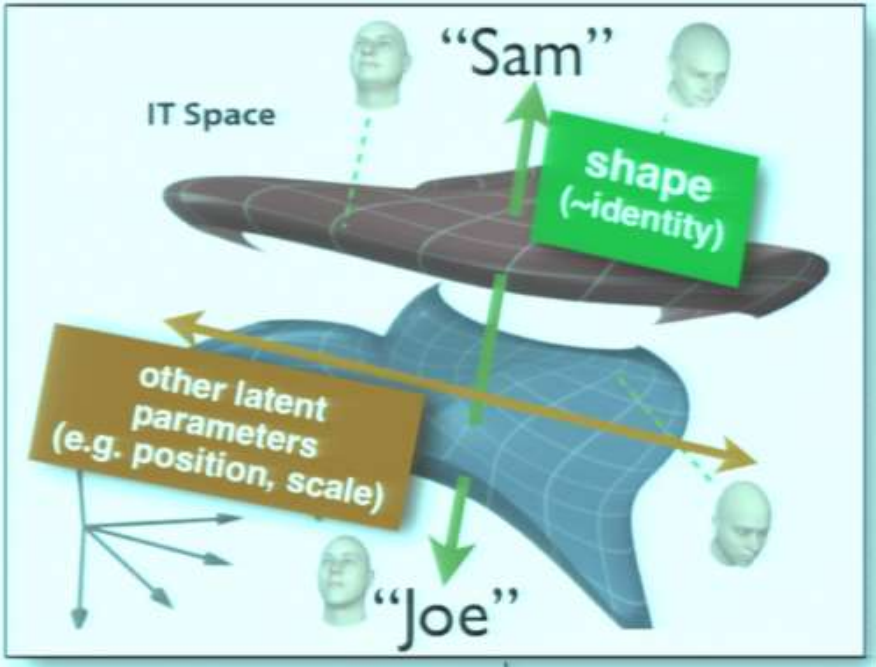
Transformation ⟶

Untangled, explicit object information



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

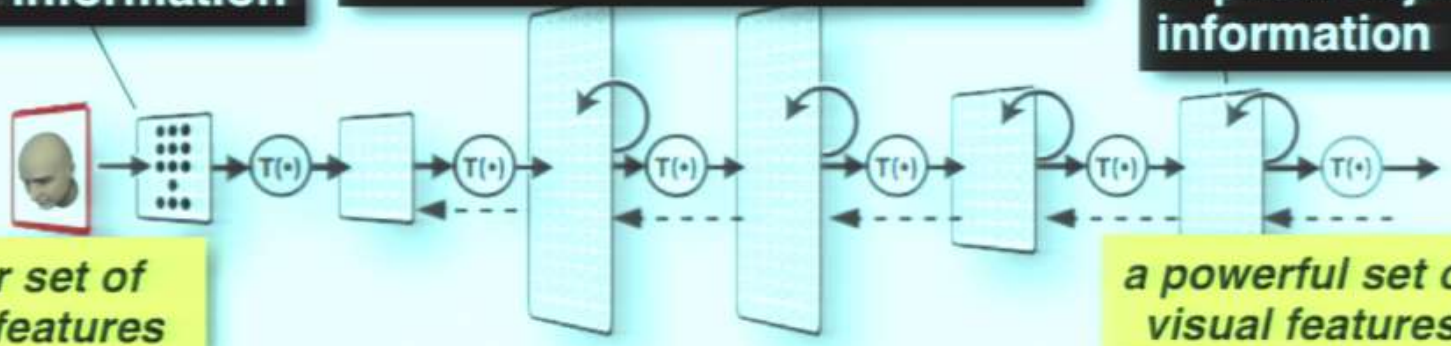




Tangled, implicit object information

Transformation ⟶

Untangled, explicit object information



a poor set of visual features

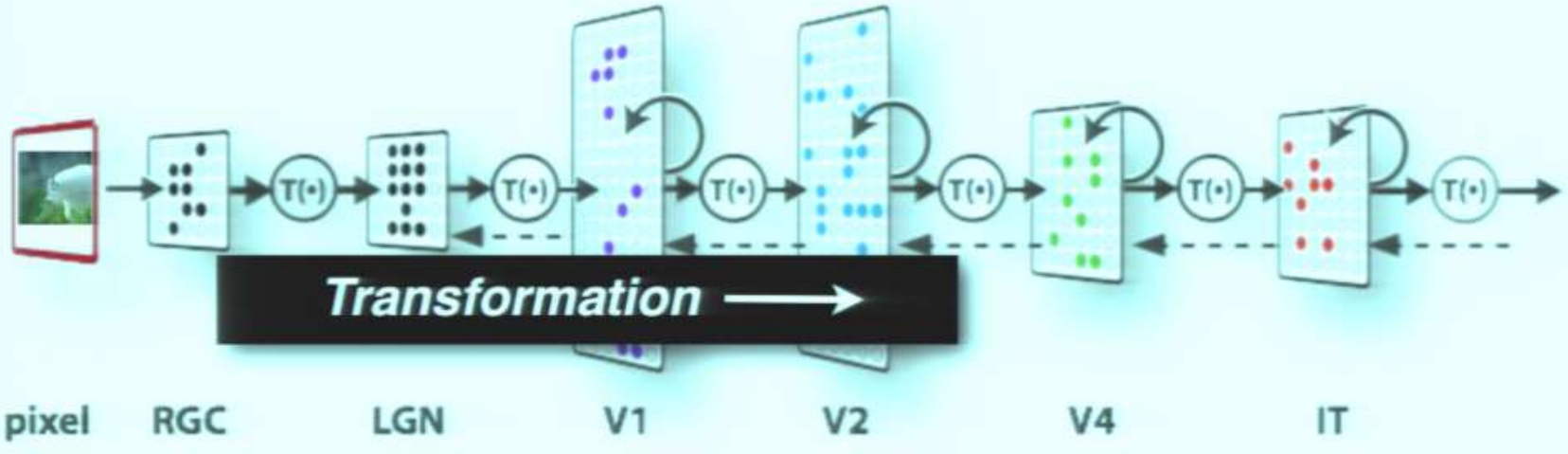
a powerful set of visual features

pixel RGC LGN V1 V2 V4 IT

Our primary questions:

How do the circuits of the ventral stream transform the pixel image to solve recognition ?

✓ **Why does the brain need to transform the pixel image ?**



Comparisons I will present today:

Comparisons I will present today:

1. Monkey neurons vs. Human Behavior

   Suggests that IT population codes are one simple step from object recognition behavior

Comparisons I will present today:

1. Monkey neurons vs. Human Behavior

   Suggests that IT population codes are one simple step from object recognition behavior

2. Machines vs. Monkey neurons

   Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

Comparisons I will present today:

1. Monkey neurons vs. Human Behavior

   Suggests that IT population codes are one simple step from object recognition behavior

2. Machines vs. Monkey neurons

   Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

3. Machines vs. Monkey neurons/Human behavior

   Demonstrates the recent bio-inspired models rival the brain in object recognition

Comparisons I will present today:

1. **Monkey neurons vs. Human Behavior**

   Suggests that IT population codes are one simple step from object recognition behavior

2. **Machines vs. Monkey neurons**

   Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

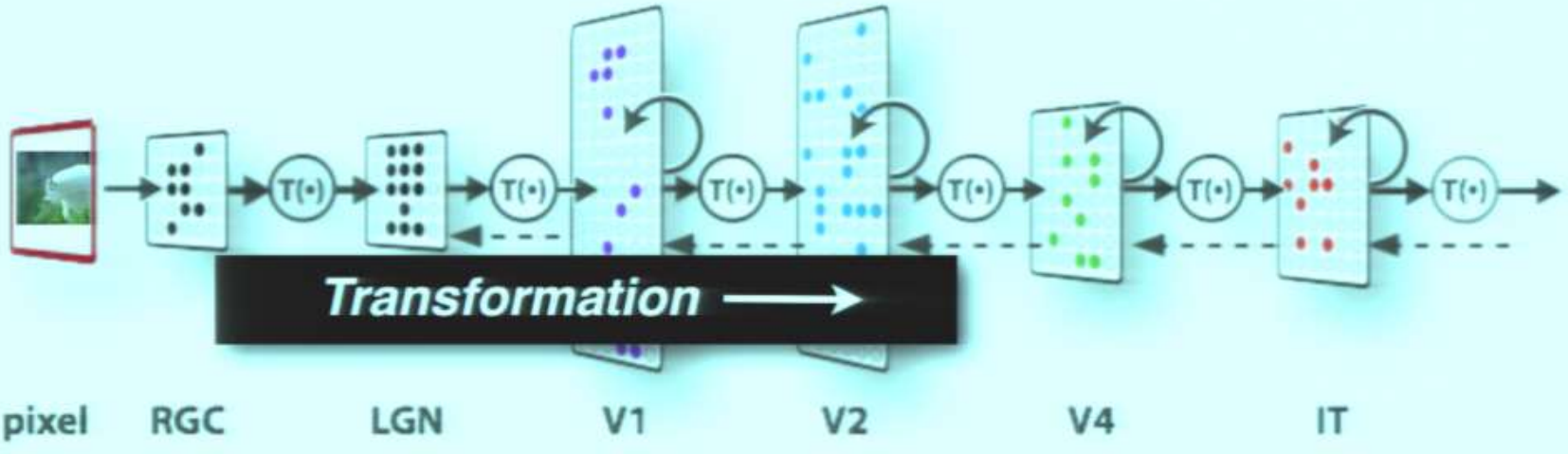
3. **Machines vs. Monkey neurons/Human behavior**

   Demonstrates the recent bio-inspired models rival the brain in object recognition

Our primary questions:

How do the circuits of the ventral stream transform the pixel image to solve recognition ?

✓ **Why** does the brain need to transform the pixel image ?



Our primary questions:

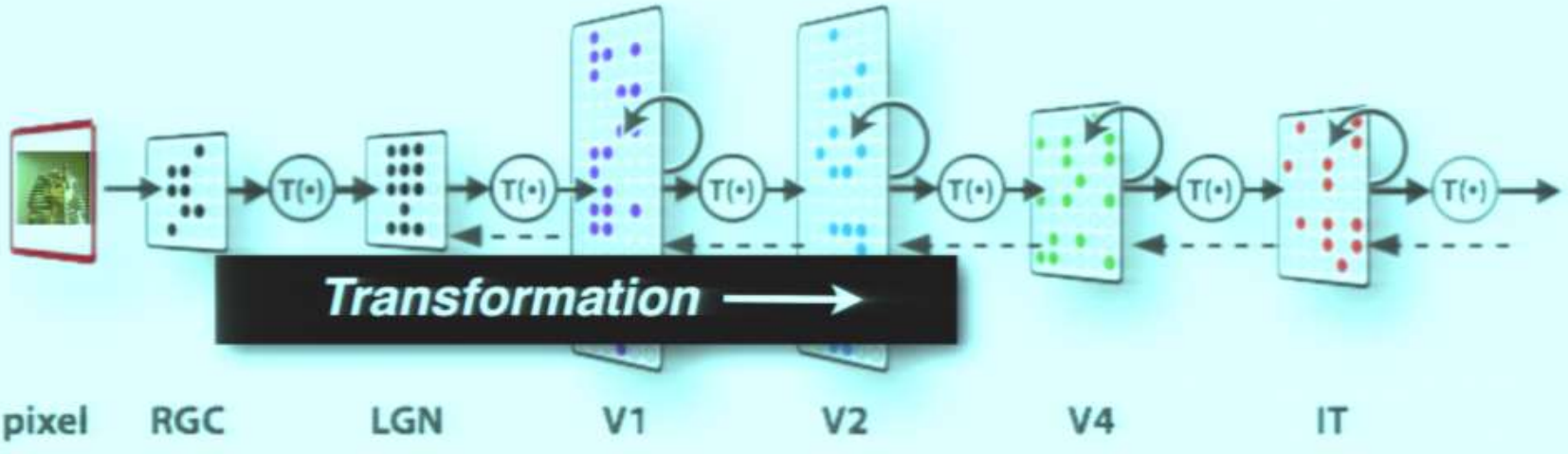
How do the circuits of the ventral stream transform the pixel image to solve recognition ?

✓ **Why does the brain need to transform the pixel image ?**

Where is the solution located, and what form does it take?

Must be sufficient (i.e. perform).

Must quantitatively predict behavior.



Our primary questions:

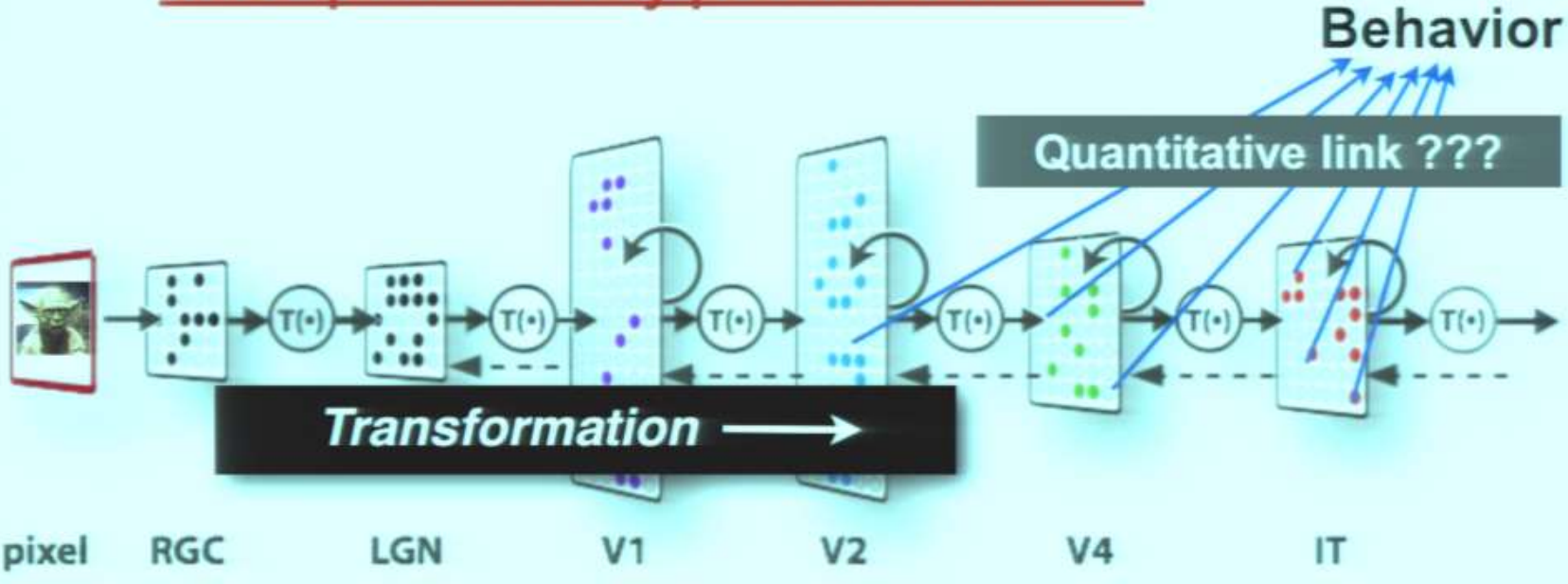
How do the circuits of the ventral stream transform the pixel image to solve recognition ?

✓ **Why** does the brain need to transform the pixel image ?

Where is the solution located, and what form does it take?

Must be sufficient (i.e. perform).

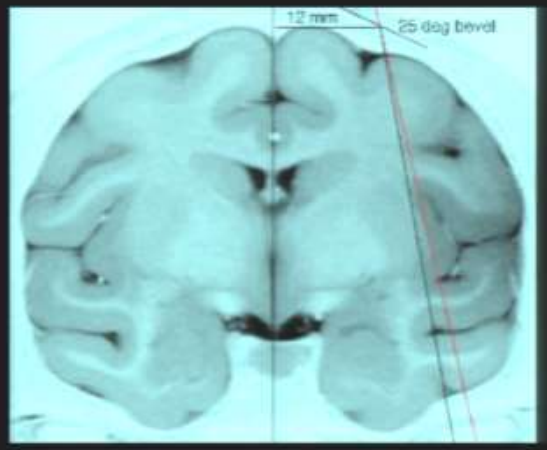
Must quantitatively predict behavior.

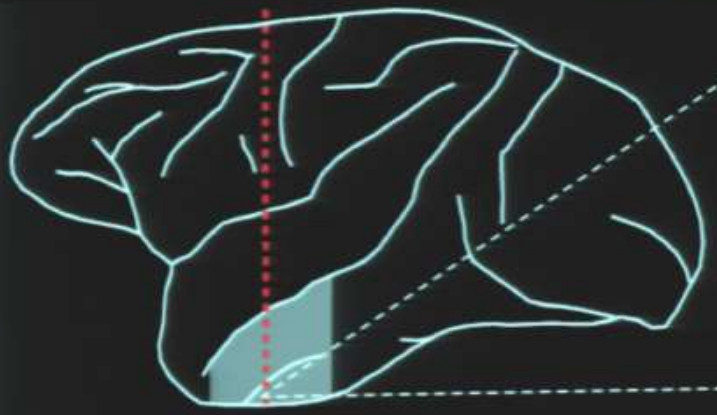


Clue: IT conveys *potentially* powerful visual features



***Gross**, Desimone, Albright, Rolls, Tanaka, Logothetis, Miyashita, Sheinberg, Connor, ...*





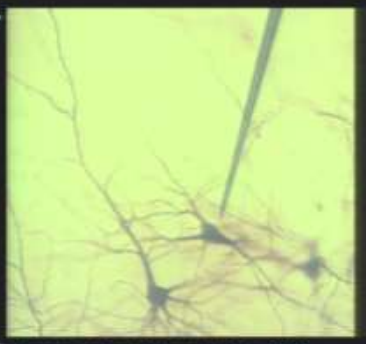


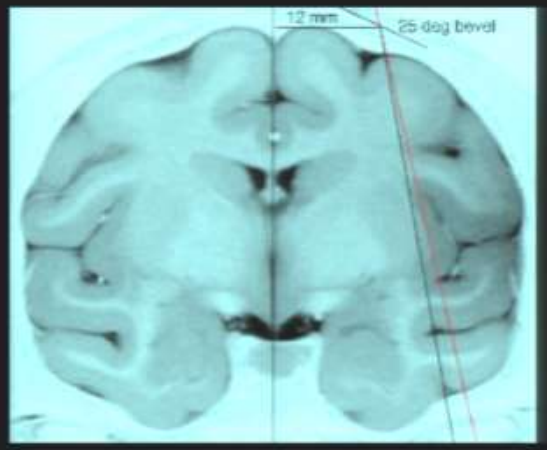
Image adapted from Hubel 1988

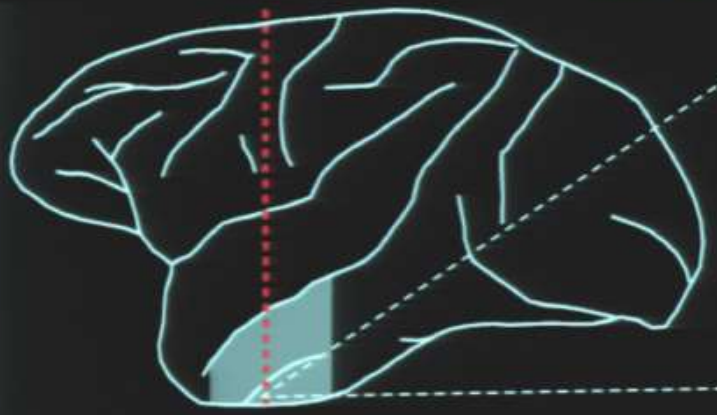
Hung, Kreiman*, Poggio and DiCarlo, **Science** (2005);*

Clue: IT conveys *potentially* powerful visual features



***Gross**, Desimone, Albright, Rolls, Tanaka, Logothetis, Miyashita, Sheinberg, Connor, ...*





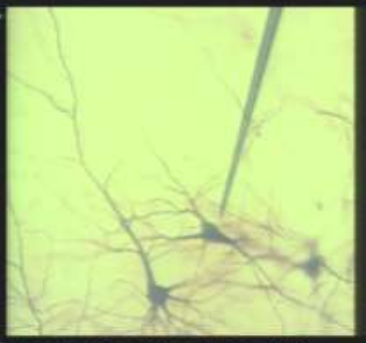


Image adapted from Hubel 1988

10 mm

Awake, fixating monkey

Site 1

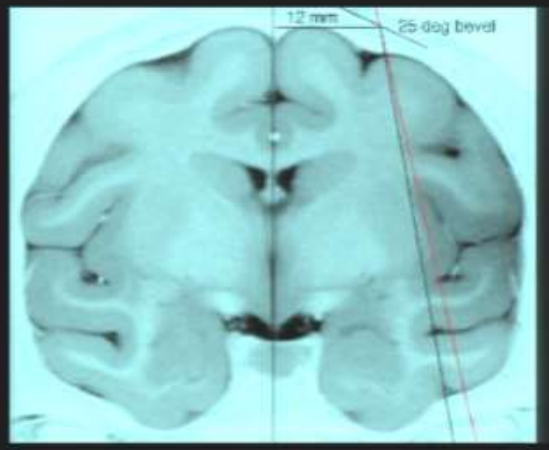
0 100 ms

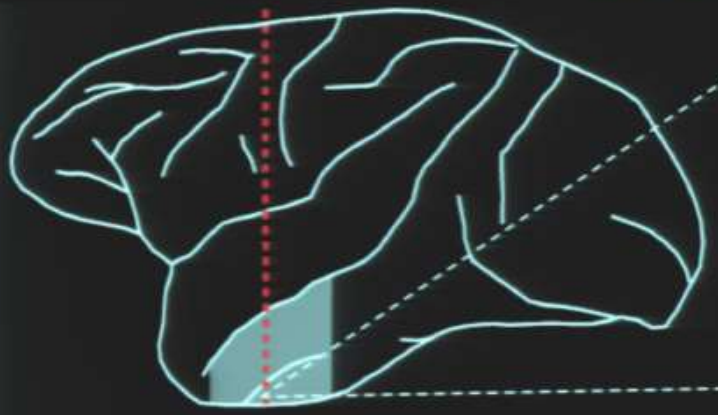
Hung, Kreiman*, Poggio and DiCarlo, **Science** (2005);*

Clue: IT conveys *potentially* powerful visual features



***Gross**, Desimone, Albright, Rolls, Tanaka, Logothetis, Miyashita, Sheinberg, Connor, ...*





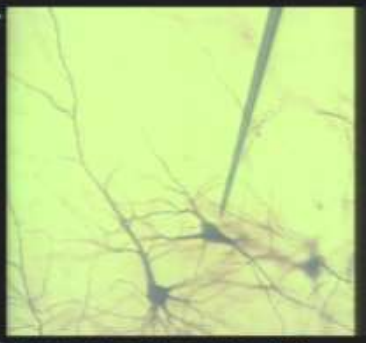


Image adapted from Hubel 1988

10 mm

Awake, fixating monkey

Site 1

Site 2

Site 3

0 100 ms

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

Clue: IT conveys *potentially* powerful visual features

Population activity

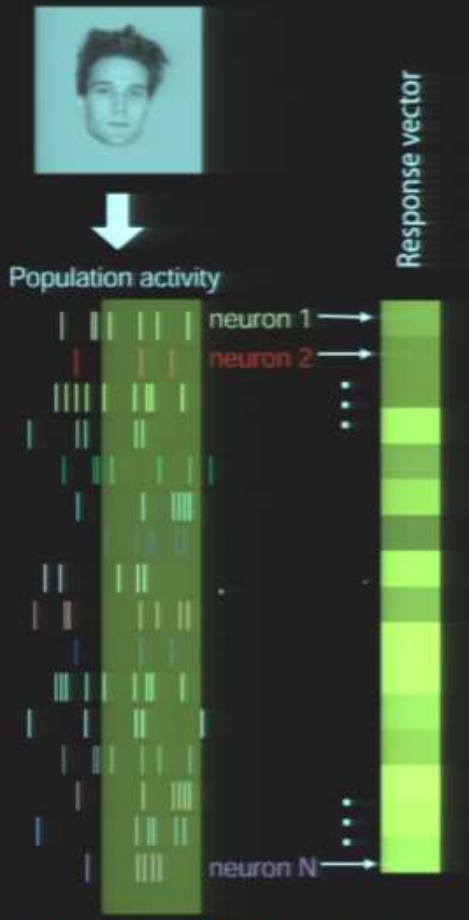
Hung, Kreiman*, Poggio and DiCarlo,* ***Science*** *(2005);*
Li, Cox, Zoccolan & DiCarlo, ***J Neurophys*** *(2009)*

Clue: IT conveys *potentially* powerful visual features

Rate code in behaviorally constrained analysis window

Hung, Kreiman*, Poggio and DiCarlo,* ***Science*** *(2005);*
Li, Cox, Zoccolan & DiCarlo, ***J Neurophys*** *(2009)*

Clue: IT conveys *potentially* powerful visual features



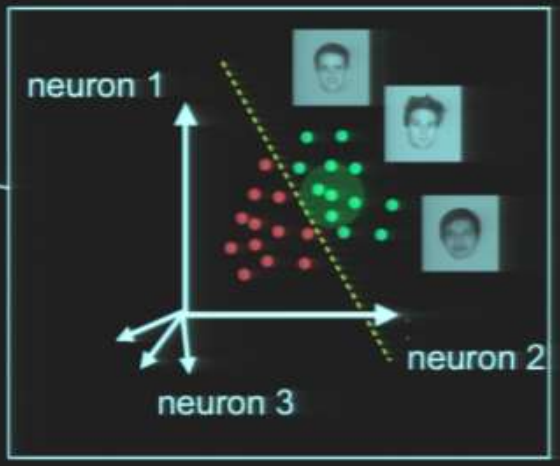
Rate code in behaviorally constrained analysis window

Hung, Kreiman*, Poggio and DiCarlo,* ***Science*** *(2005);*
Li, Cox, Zoccolan & DiCarlo, ***J Neurophys*** *(2009)*

Clue: IT conveys *potentially* powerful visual features

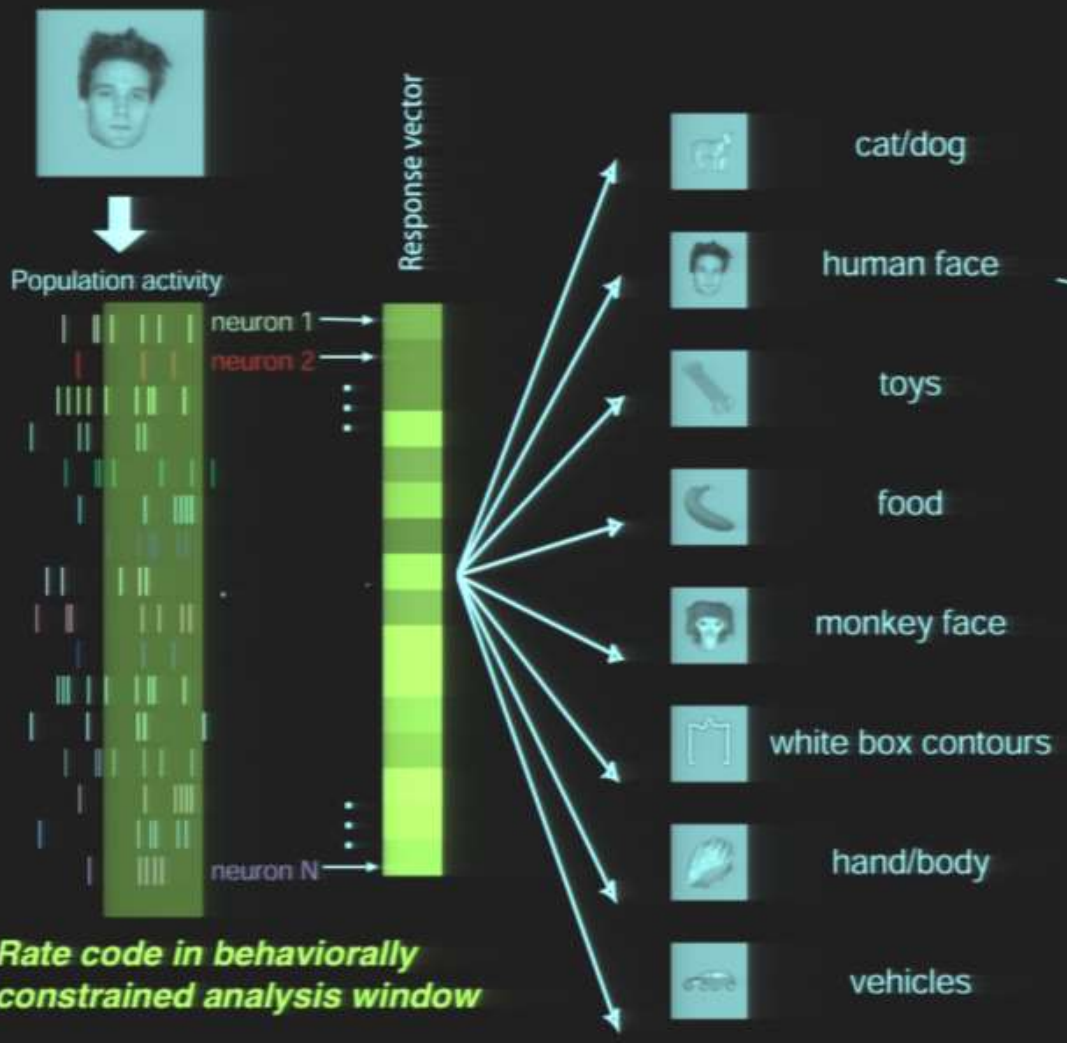
Rate code in behaviorally constrained analysis window

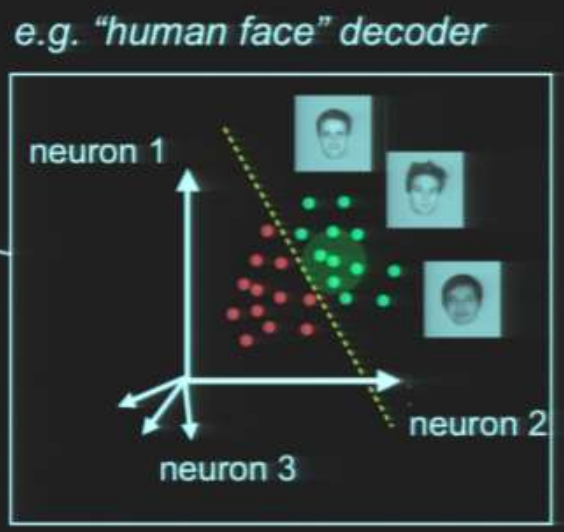
e.g. "human face" decoder



Hung, Kreiman*, Poggio and DiCarlo, **Science** (2005);*
*Li, Cox, Zoccolan & DiCarlo, **J Neurophys** (2009)*

Clue: IT conveys *potentially* powerful visual features





Hung, Kreiman*, Poggio and DiCarlo, **Science** (2005);*
*Li, Cox, Zoccolan & DiCarlo, **J Neurophys** (2009)*

Clue: IT conveys *potentially* powerful visual features

Population activity

neuron 1

neuron 2

neuron N

(vary N, from 1 to 256)

Response vector

IT population supports good real-time performance*

toys

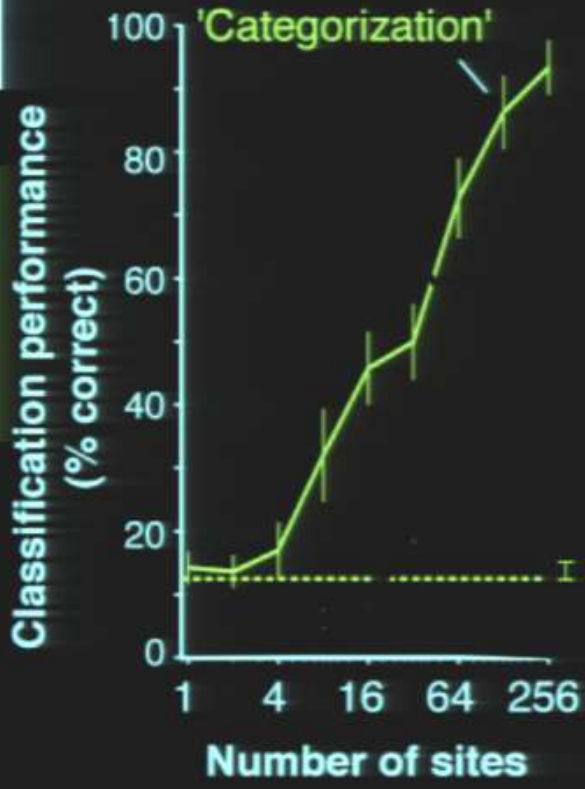
food

monkey face

white box contours

hand/body

vehicles



Hung, Kreiman*, Poggio and DiCarlo, **Science** (2005);*
*Li, Cox, Zoccolan & DiCarlo, **J Neurophys** (2009)*

Clue: IT conveys *potentially* powerful visual features

Population activity

neuron 1

neuron 2

neuron N

(vary N, from 1 to 256)

Response vector

IT population supports good real-time performance*

toys

food

monkey face

white box contours

hand/body

Nice generalization over object position, scale, and clutter

(not found in early visual cortex)

Hung, Kreiman*, Poggio and DiCarlo, **Science** (2005);*
*Li, Cox, Zoccolan & DiCarlo, **J Neurophys** (2009)*

Are any IT neural codes sufficient to explain human object recognition?

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

Alternative, more complex (more attractive?) hypotheses:

IT does not directly underlie object recognition
(i.e. the key neuronal code are elsewhere in the brain, e.g. V4, PFC, LIP, ...)

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

Alternative, more complex (more attractive?) hypotheses:

IT does not directly underlie object recognition
(i.e. the key neuronal code are elsewhere in the brain, e.g. V4, PFC, LIP, ...)

Rate codes in IT are not sufficient
(e.g. coordinated spike timing patterns are the true answer)

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

Alternative, more complex (more attractive?) hypotheses:

IT does not directly underlie object recognition
(i.e. the key neuronal code are elsewhere in the brain, e.g. V4, PFC, LIP, ...)

Rate codes in IT are not sufficient
(e.g. coordinated spike timing patterns are the true answer)

Automatically-evoked spike patterns are not sufficient
(e.g. attentional or arousal mechanisms are critical)

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

Alternative, more complex (more attractive?) hypotheses:

IT does not directly underlie object recognition
(i.e. the key neuronal code are elsewhere in the brain, e.g. V4, PFC, LIP, ...)

Rate codes in IT are not sufficient
(e.g. coordinated spike timing patterns are the true answer)

Automatically-evoked spike patterns are not sufficient
(e.g. attentional or arousal mechanisms are critical)

Compartments within IT must be carefully considered
(e.g. any tasks related to faces are handled by the "face patch" network)

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

Alternative, more complex (more attractive?) hypotheses:

IT does not directly underlie object recognition
(i.e. the key neuronal code are elsewhere in the brain, e.g. V4, PFC, LIP, ...)

Rate codes in IT are not sufficient
(e.g. coordinated spike timing patterns are the true answer)

Automatically-evoked spike patterns are not sufficient
(e.g. attentional or arousal mechanisms are critical)

Compartments within IT must be carefully considered
(e.g. any tasks related to faces are handled by the "face patch" network)

Monkey neuronal codes cannot explain human perception
(e.g. monkeys can't "know" what a chair is; humans must be better)

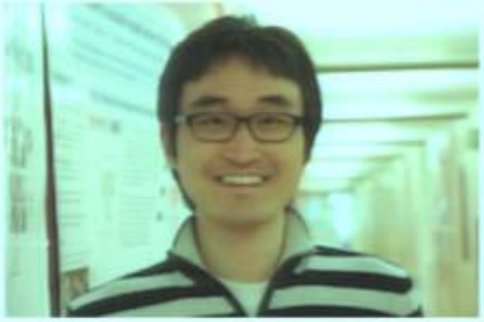
Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

Najib Majaj

(postdoc)



Ha Hong

(graduate student)

Ethan Soloman

(undergraduate student)

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

1. Define a set of challenging object recognition (O.R.) tasks

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

3. Measure large samples of neuronal population spiking responses

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

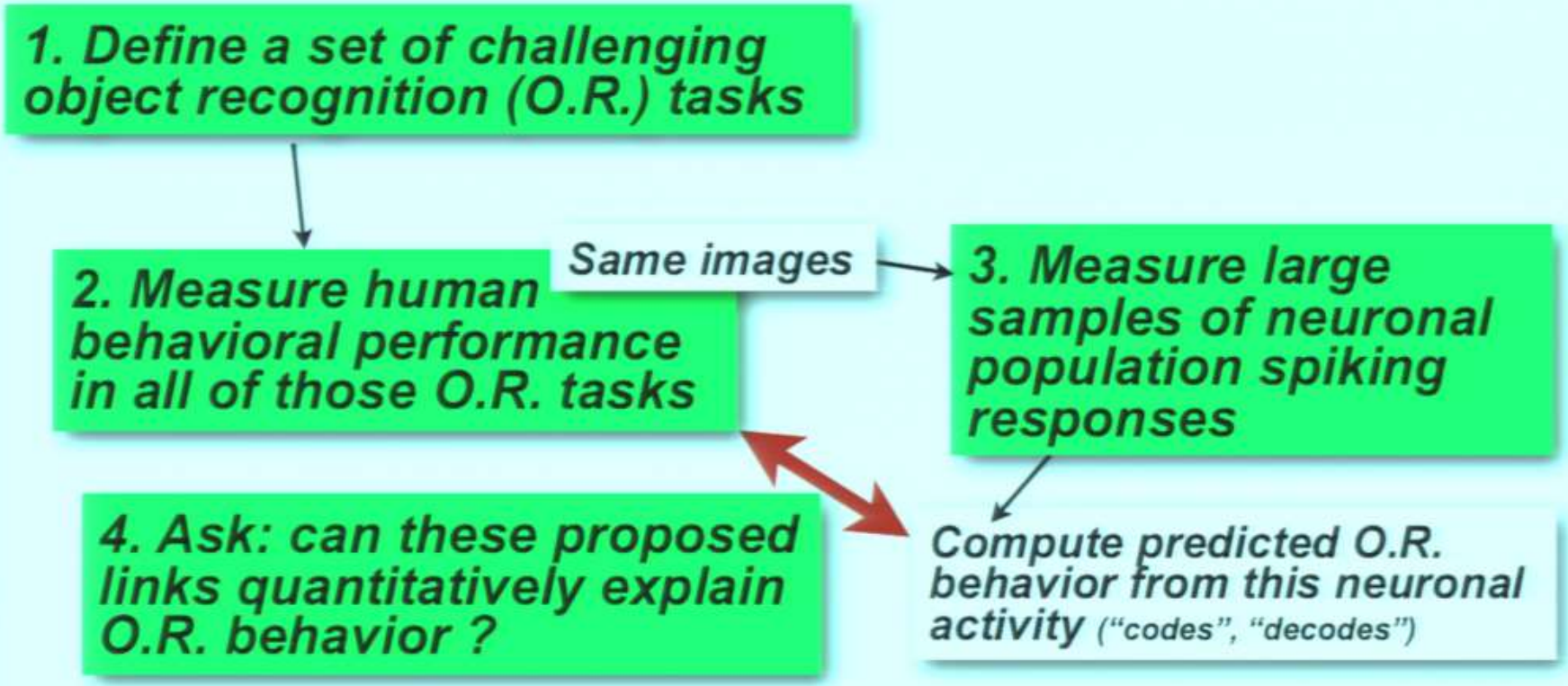
3. Measure large samples of neuronal population spiking responses

Compute predicted O.R. behavior from this neuronal activity ("codes", "decodes")

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition



Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

3. Measure large samples of neuronal population spiking responses

4. Ask: can these proposed links quantitatively explain O.R. behavior ?

Compute predicted O.R. behavior from this neuronal activity ("codes", "decodes")

Strong correlational methods. Causality is our next step.

Are any IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Automatically-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

3. Measure large samples of neuronal population spiking responses

4. Ask: can these proposed links quantitatively explain O.R. behavior ?

Compute predicted O.R. behavior from this neuronal activity ("codes", "decodes")

Strong correlational methods. Causality is our next step.

Our goal is NOT simply "extracting information" from the brain.

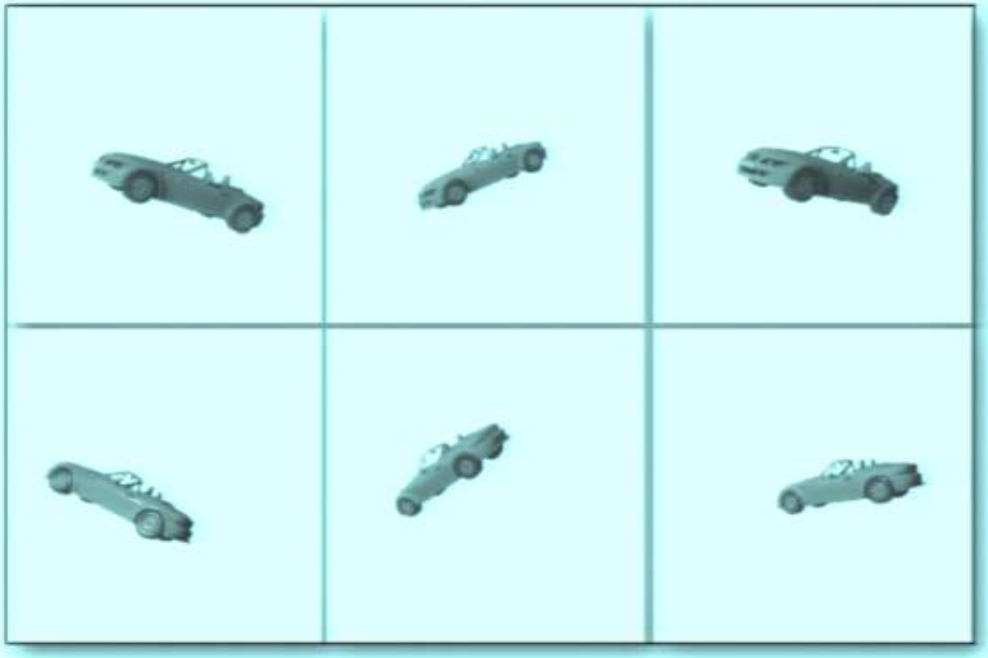
Are any IT neural codes sufficient to explain human object recognition?

1. Define a set of challenging object recognition (O.R.) tasks

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

"Identity preserving image variation"

View: position, size, pose, illumination

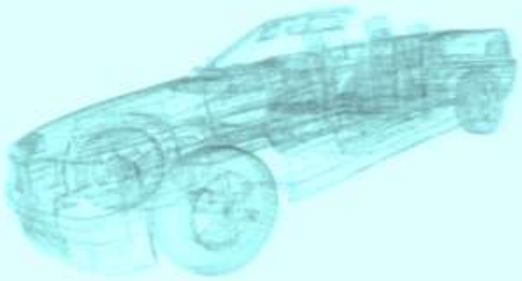


Clutter, occlusion

Intraclass

Poggio, Ullman, Grossberg, Edleman, Biederman, etc.
DiCarlo and Cox, ***TICS*** *(2007);*
Pinto, Cox, and DiCarlo, ***PLoS Comp Bio*** *(2008)*

3-d object Models



add view parameters

+

Position
Size
Pose

use ray tracing to render

add to background

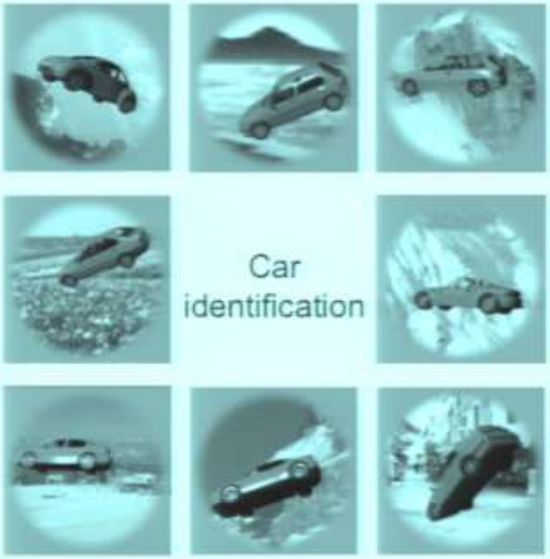
add to background



- 64 objects, can generate as many images as we like
- full parametric control
- "natural" statistics
- uncorrelated, new background every image
- not fully "natural" by design -- challenging for computer vision, doable by humans

Object recognition 1.0 (HVM1.0)

Basic level categorization



Car identification

Face identification

Object recognition 1.0 (HVM1.0)

Are any IT neural codes sufficient to explain human object recognition?

1. Define a set of challenging object recognition (O.R.) tasks

Are any IT neural codes sufficient to explain human object recognition?

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Object recognition 1.0 (HVM1.0)

Three 8-way classification tasks (blocked).

==> 24 binary discriminations, each tested at 6 levels of variation

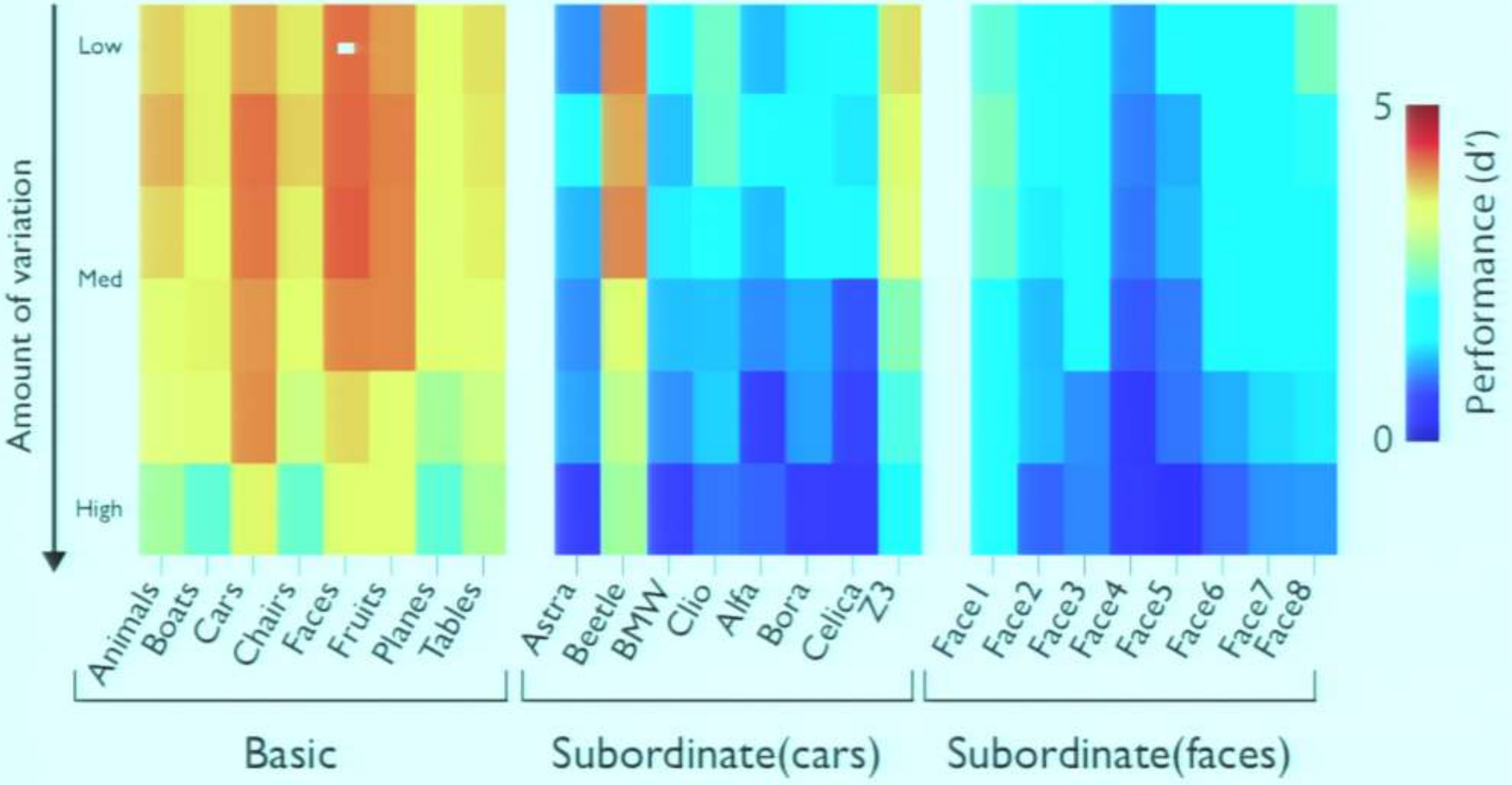
==> (24 x 6) = 144 "tasks" (later, consider only 64 of these "tasks")

8 deg image at center of gaze, 100 ms viewing time (core recognition)

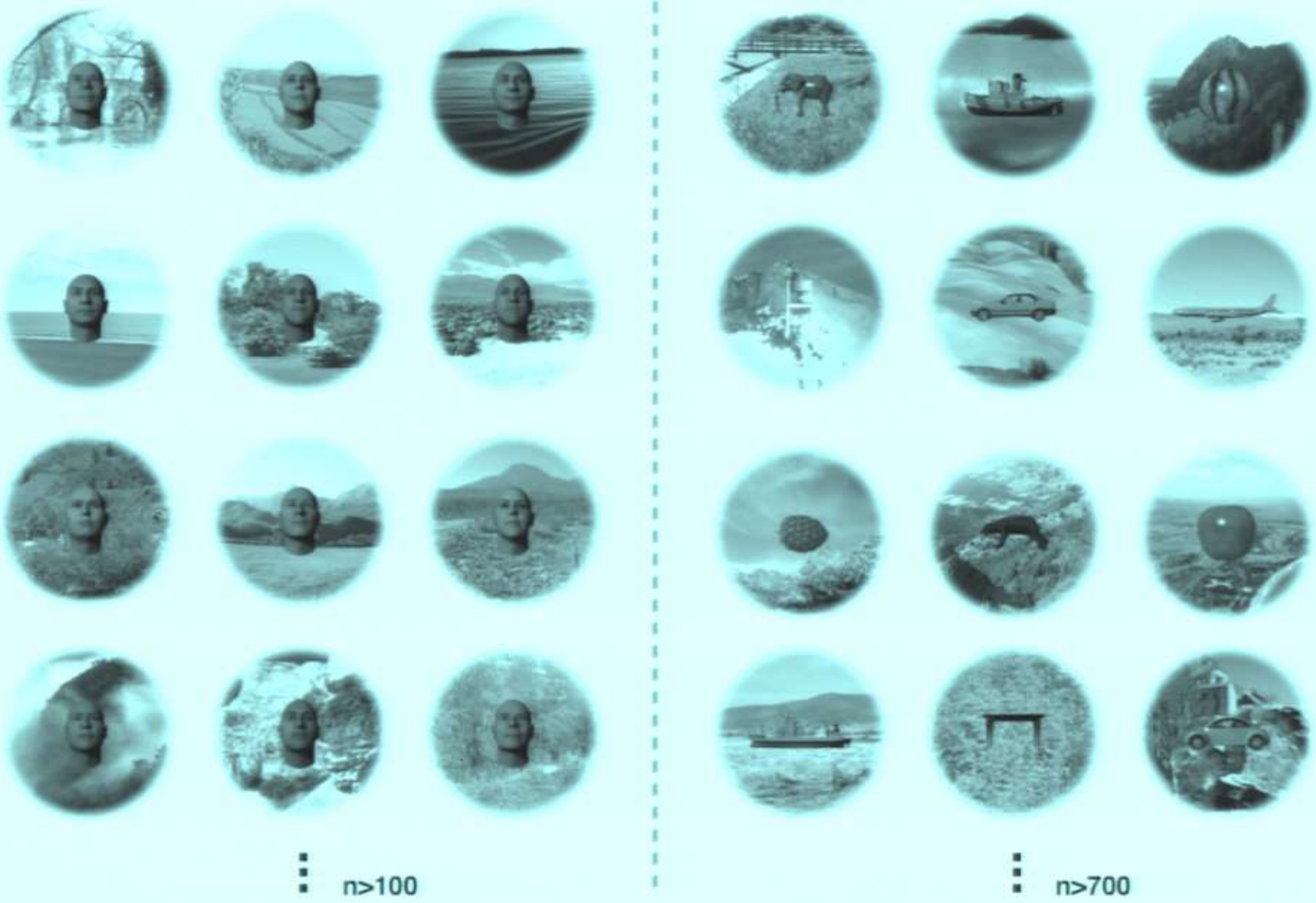
Object recognition 1.0

n=144 tasks

Measurements of human performance (d′)

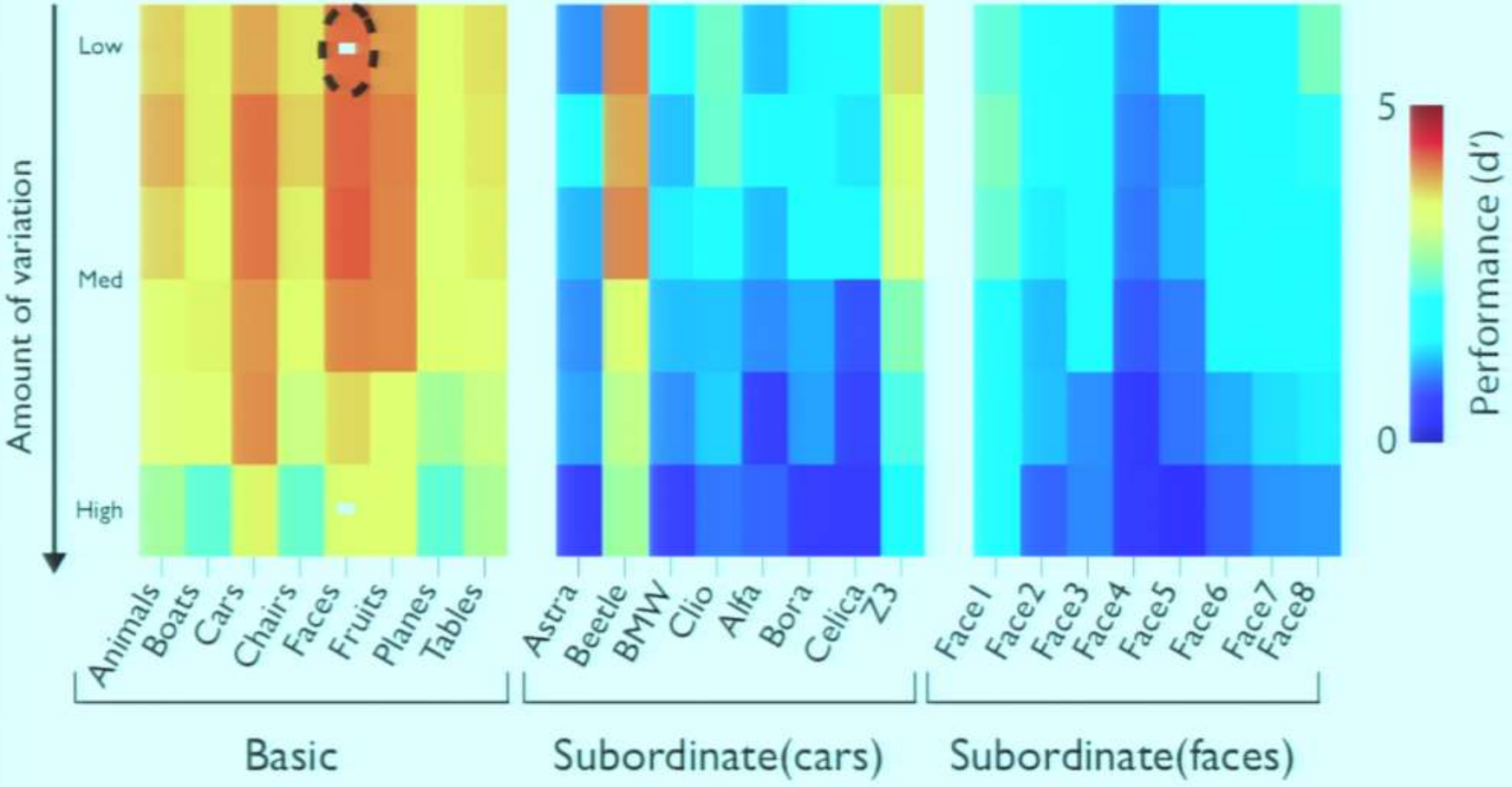


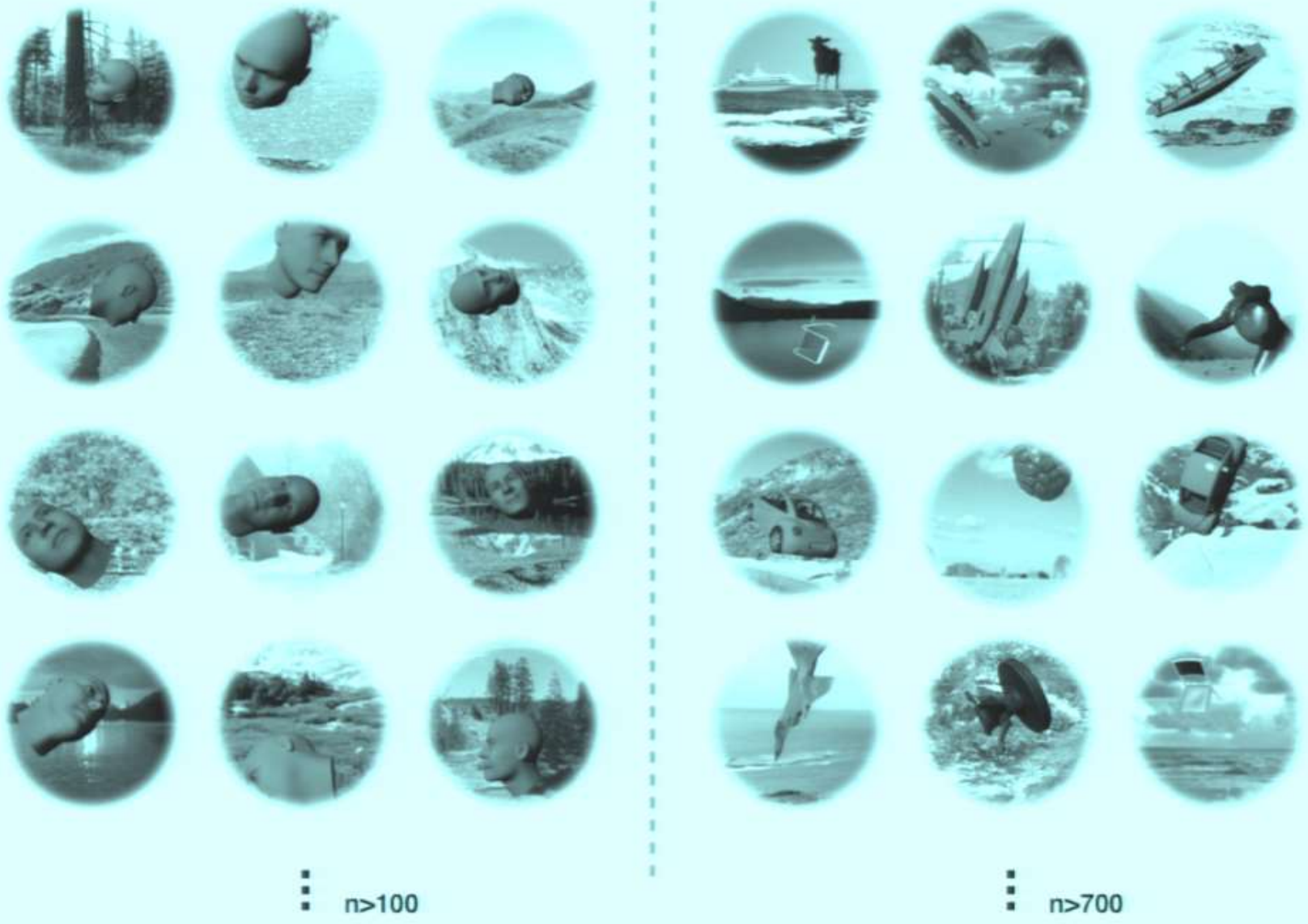
n=39 humans subjects, >23,000 trials


"Face"
Not "Face"
n>100
n>700

Object recognition 1.0

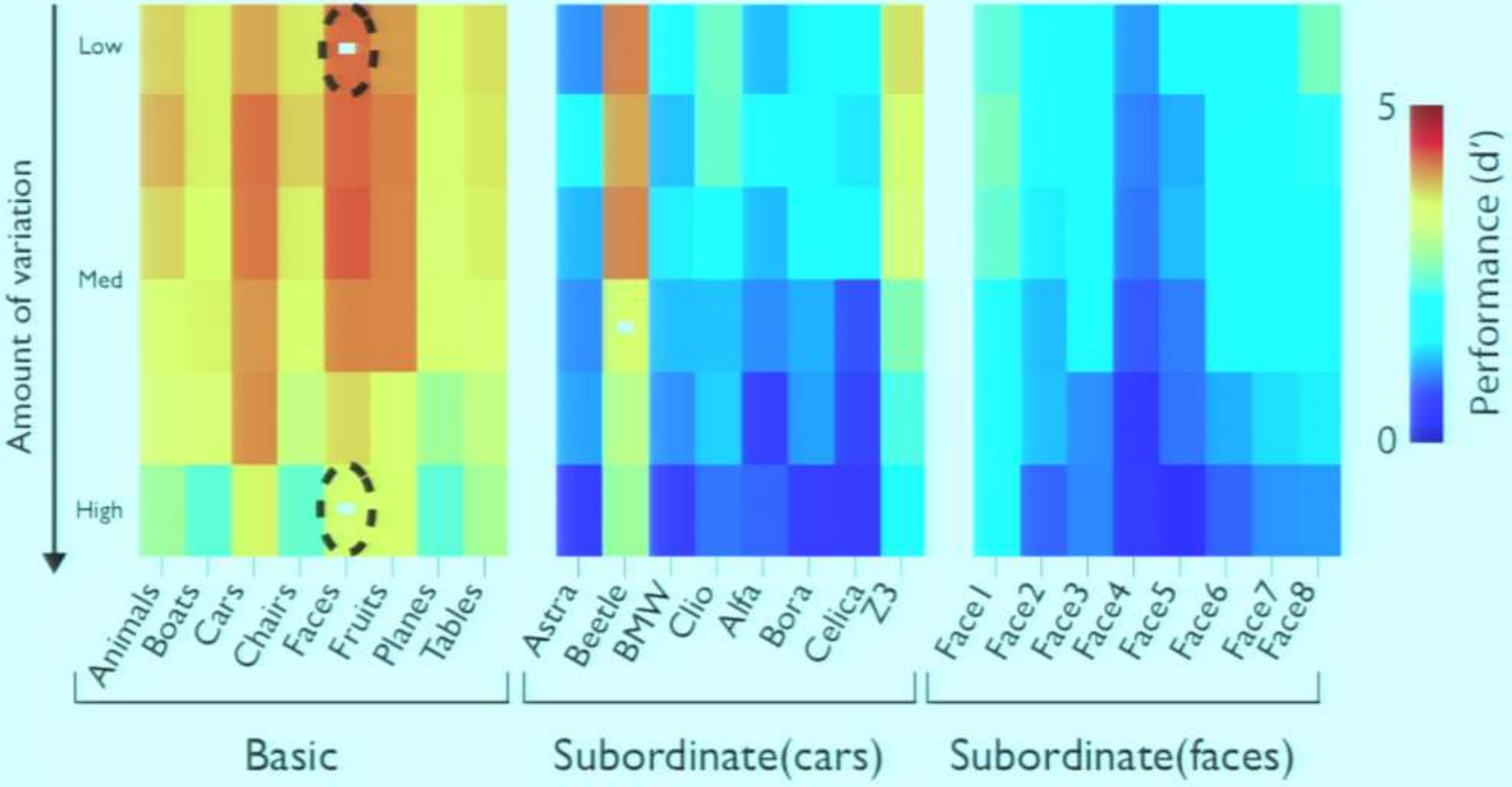
Measurements of human performance (d′)




"face"
not "face"
n>100
n>700

Object recognition 1.0

Measurements of human performance (d′)



“Beetle”

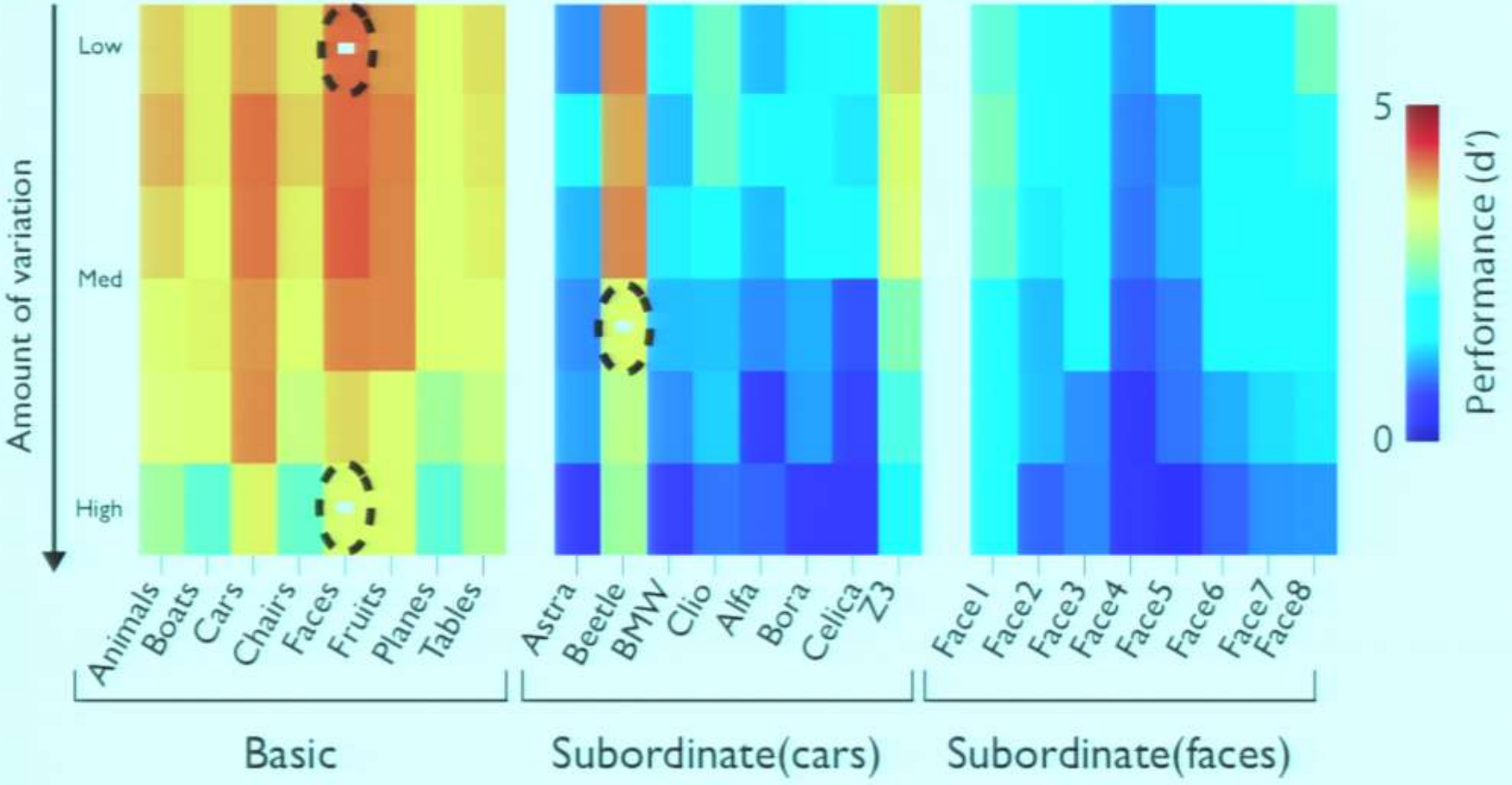
n>100

Not “Beetle”

n>700

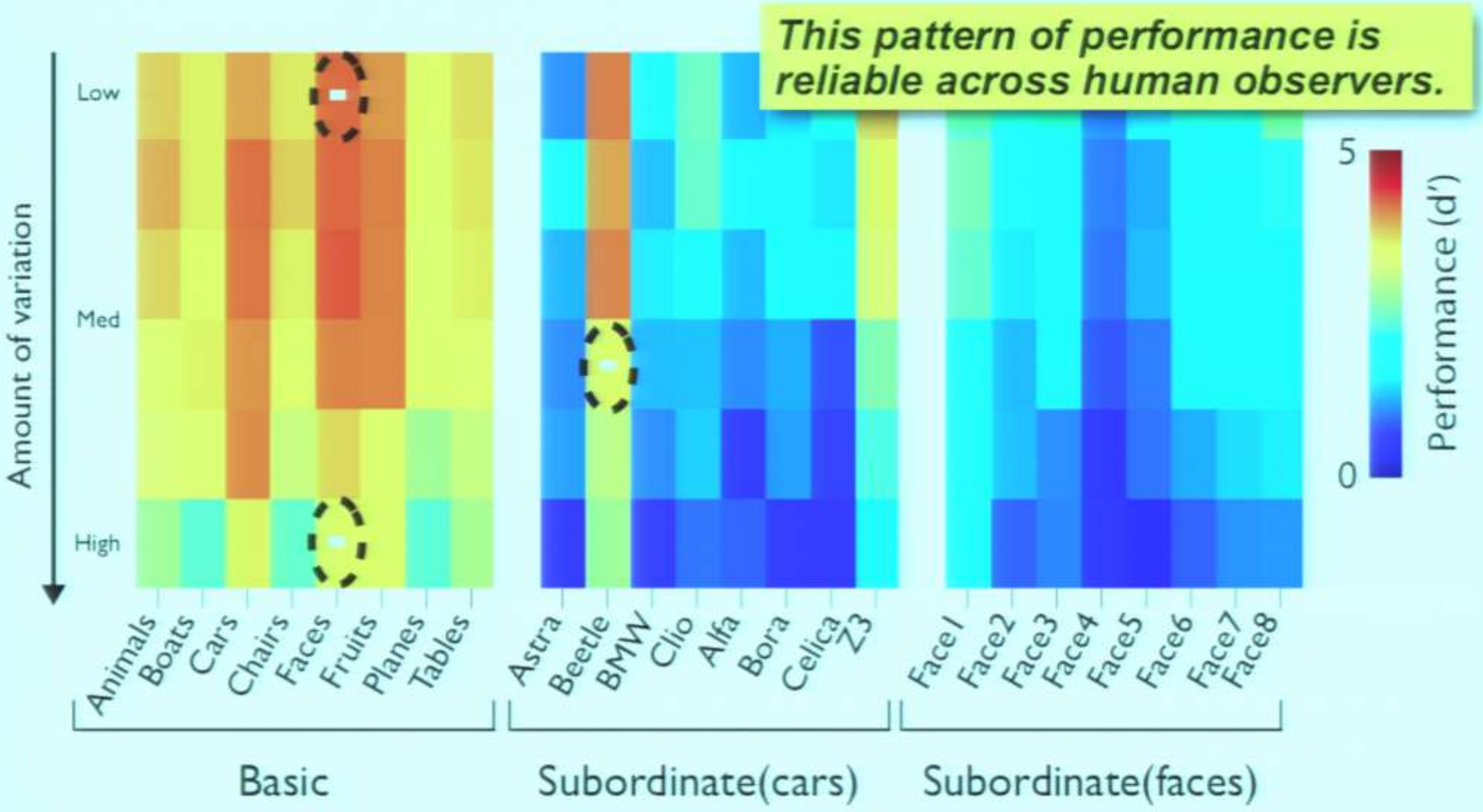
Mosaic of human ability (d′)

Object recognition 1.0



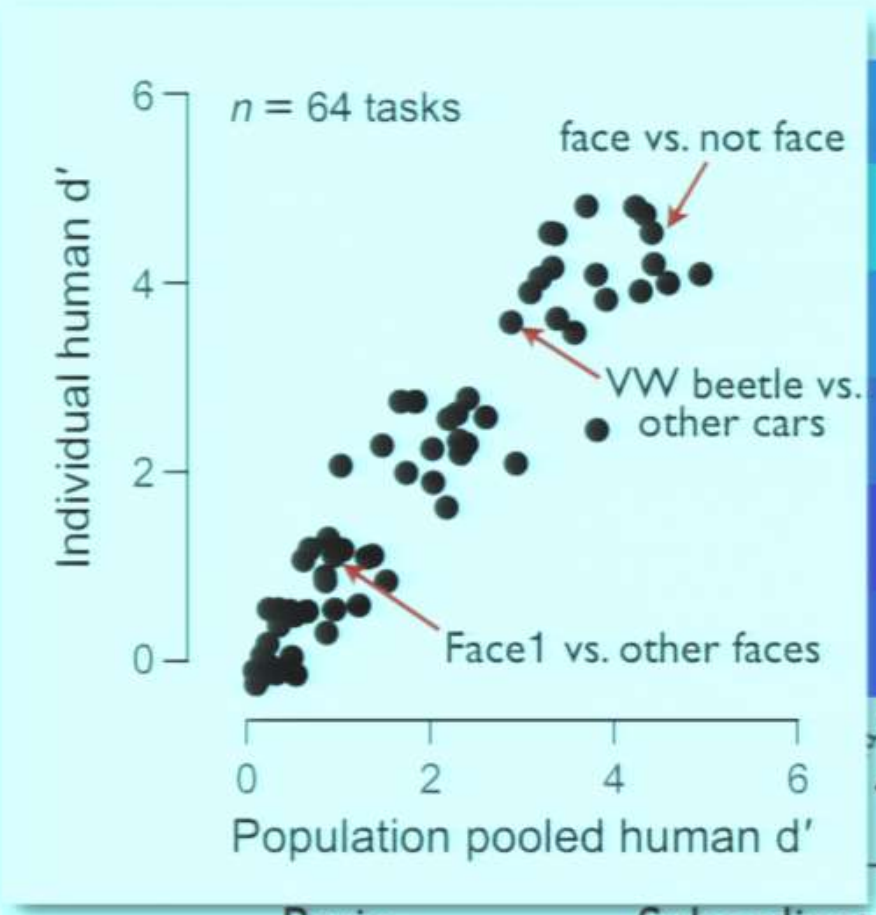
Mosaic of human ability (d′)

Object recognition 1.0

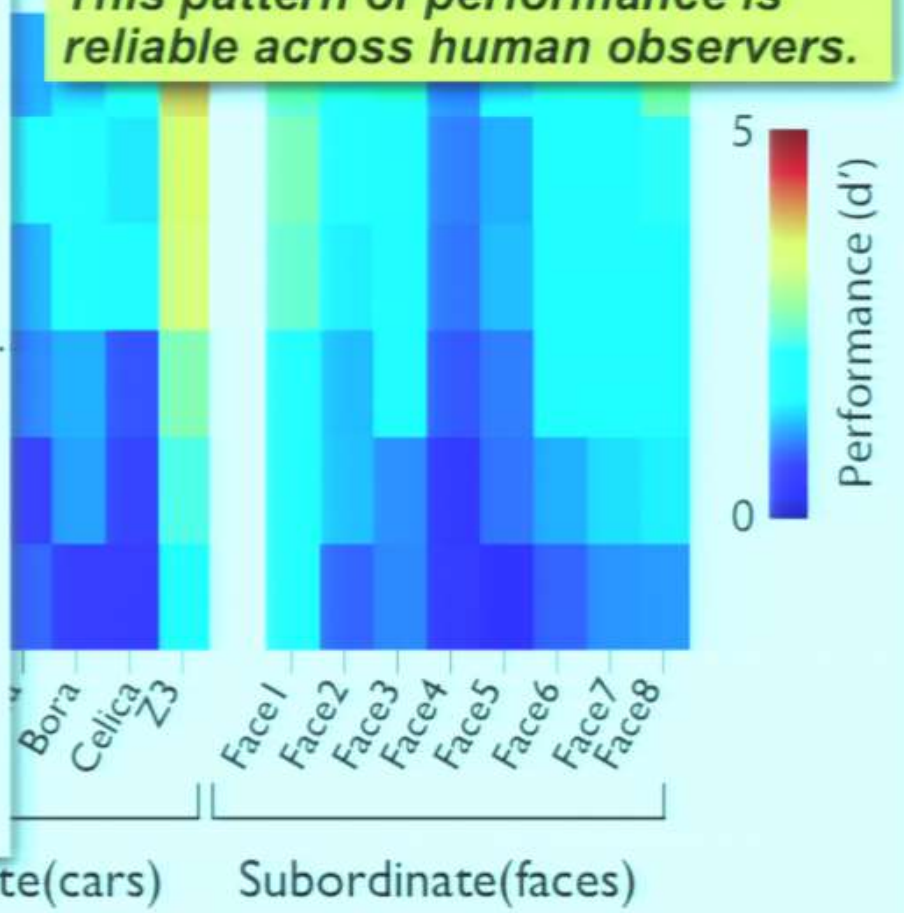


Mosaic of human ability (d′)

Object recognition 1.0

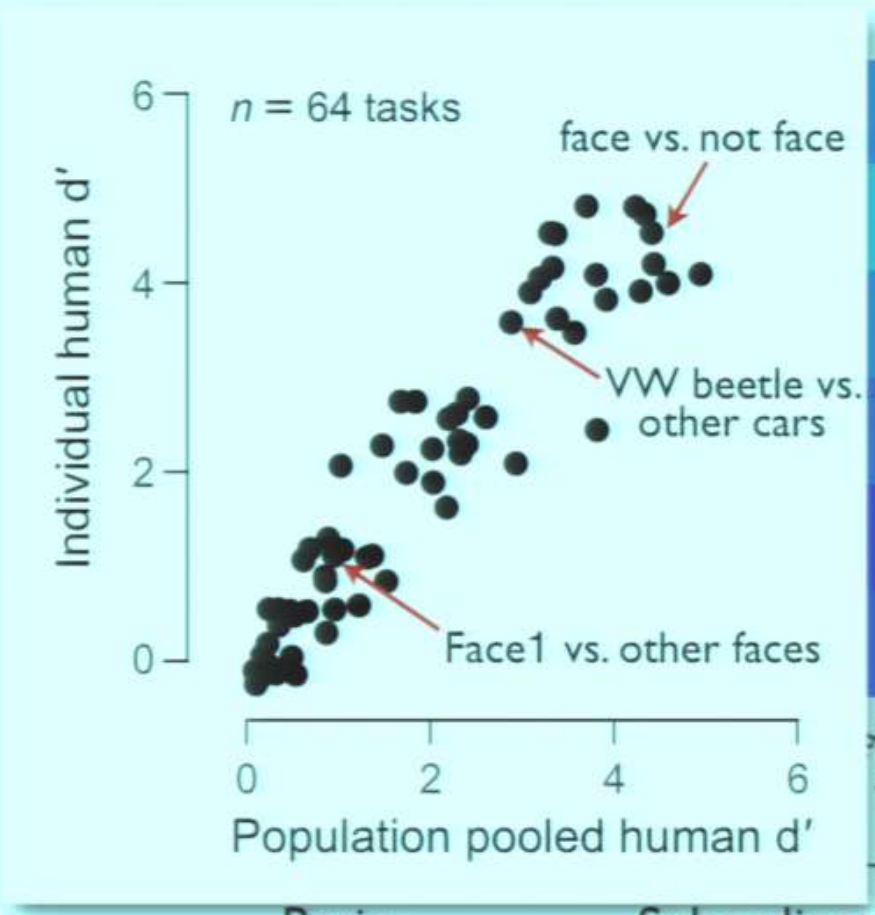


This pattern of performance is reliable across human observers.



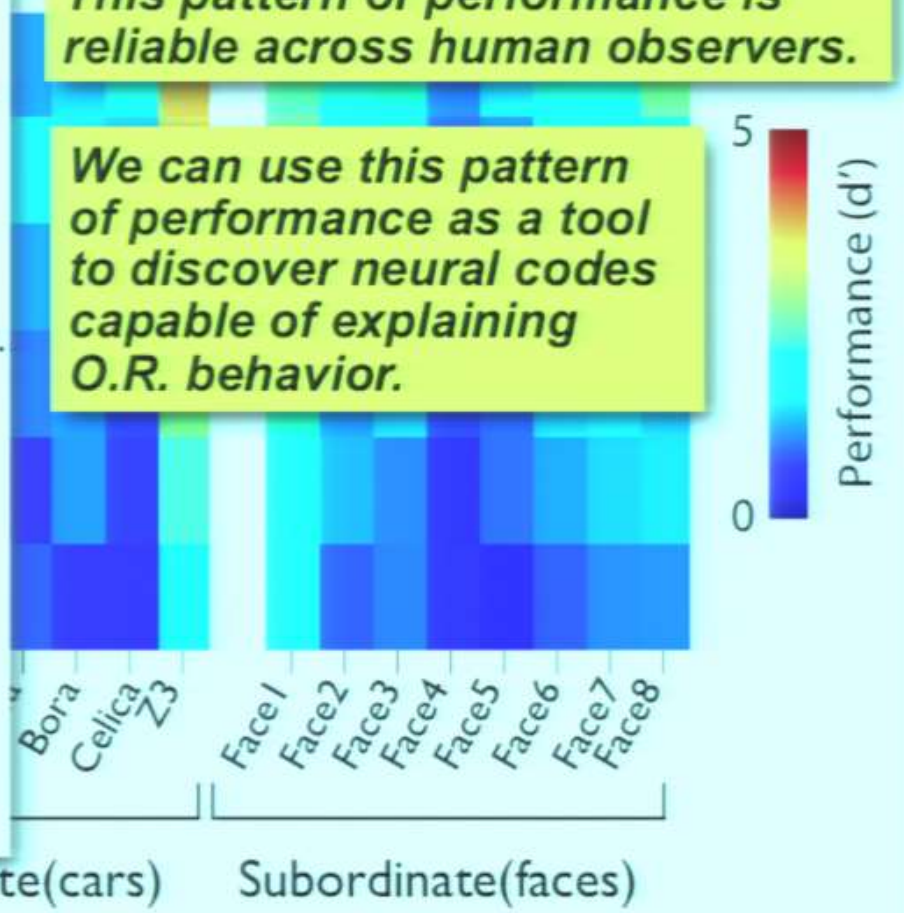
Mosaic of human ability (d′)

Object recognition 1.0



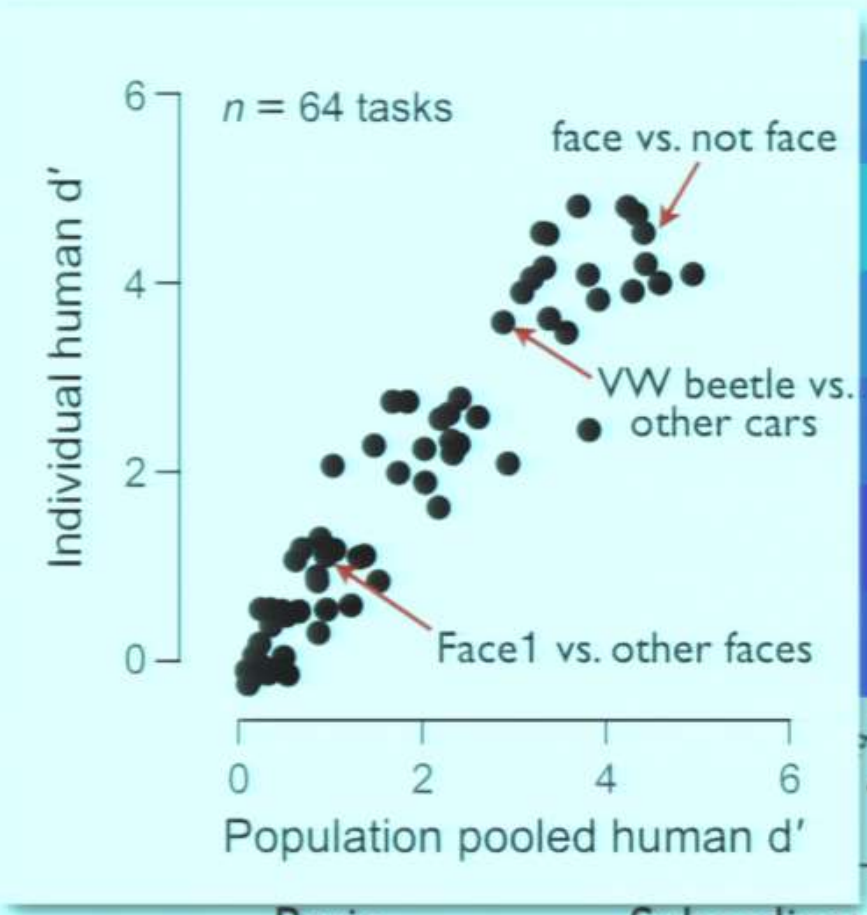
This pattern of performance is reliable across human observers.

We can use this pattern of performance as a tool to discover neural codes capable of explaining O.R. behavior.



Mosaic of human ability (d′)

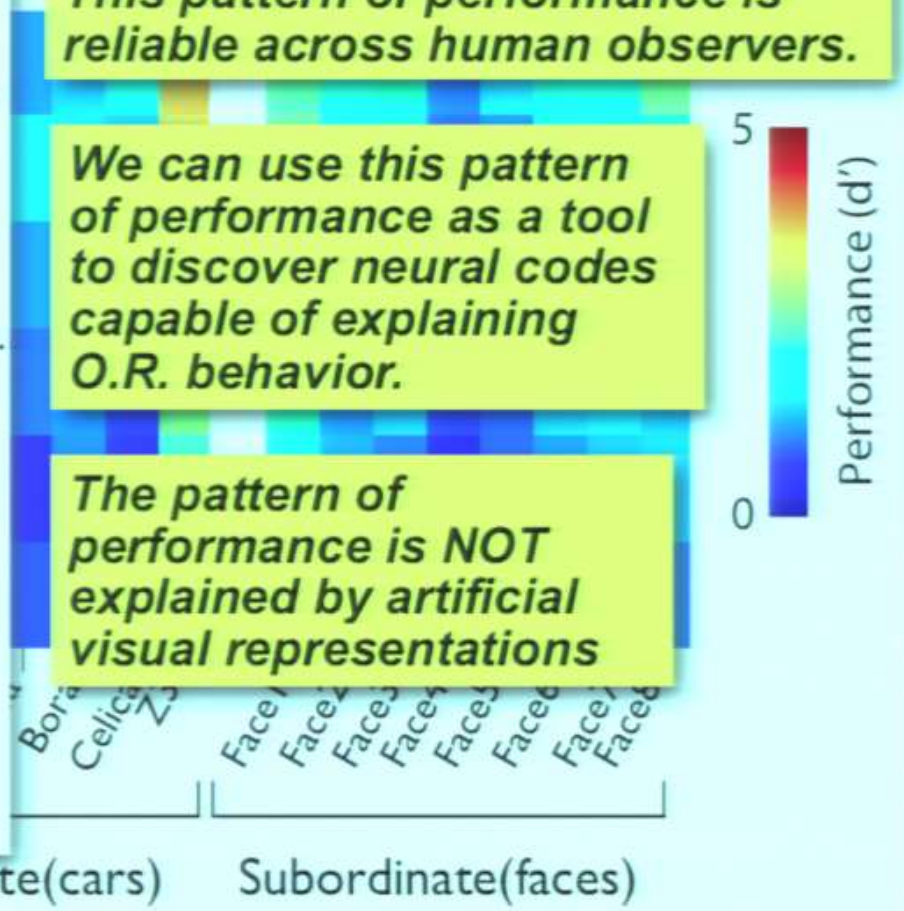
Object recognition 1.0



This pattern of performance is reliable across human observers.

We can use this pattern of performance as a tool to discover neural codes capable of explaining O.R. behavior.

The pattern of performance is NOT explained by artificial visual representations



Are any IT neural codes sufficient to explain human object recognition?

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Are any IT neural codes sufficient to explain human object recognition?

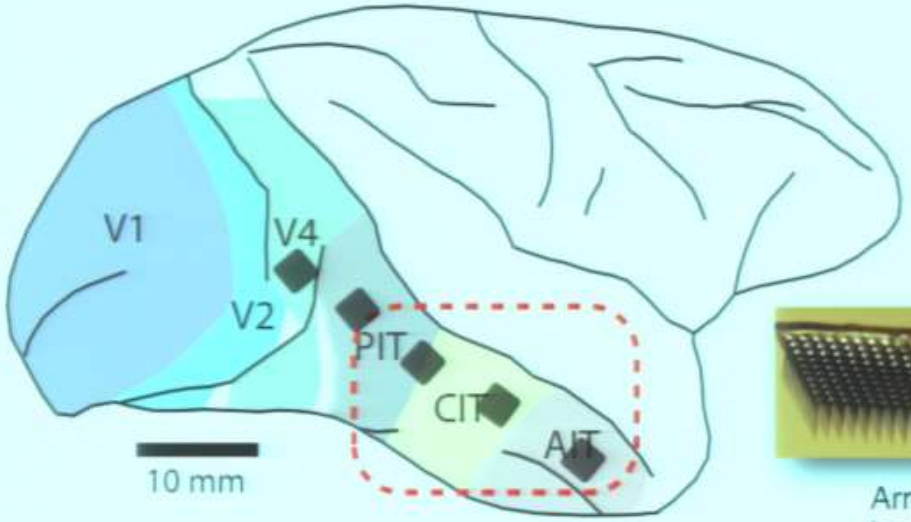
1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

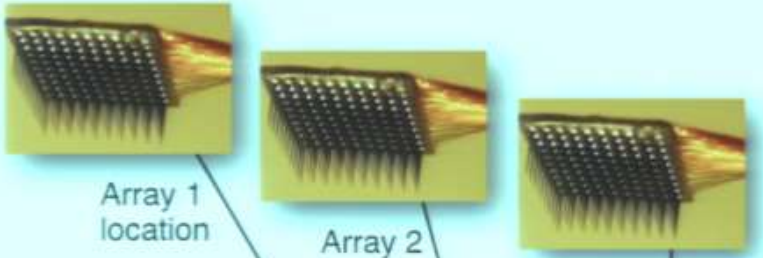
Same images

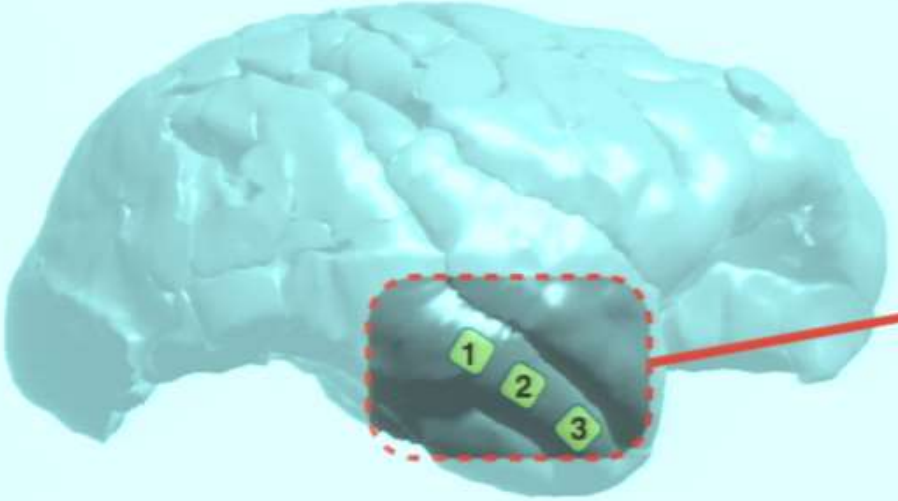
3. Measure large samples of neuronal population spiking responses

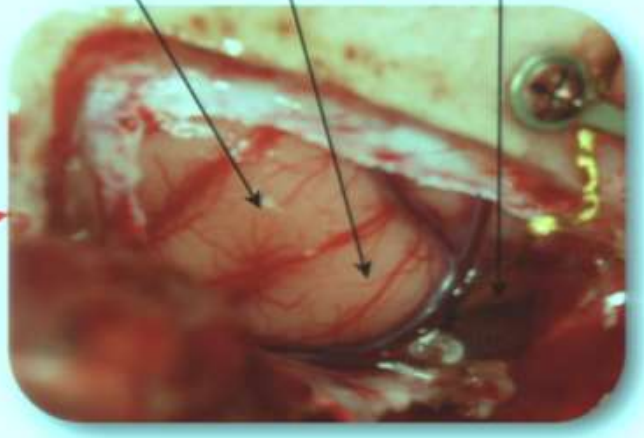
Methods advance: large scale neuronal recording along the ventral stream



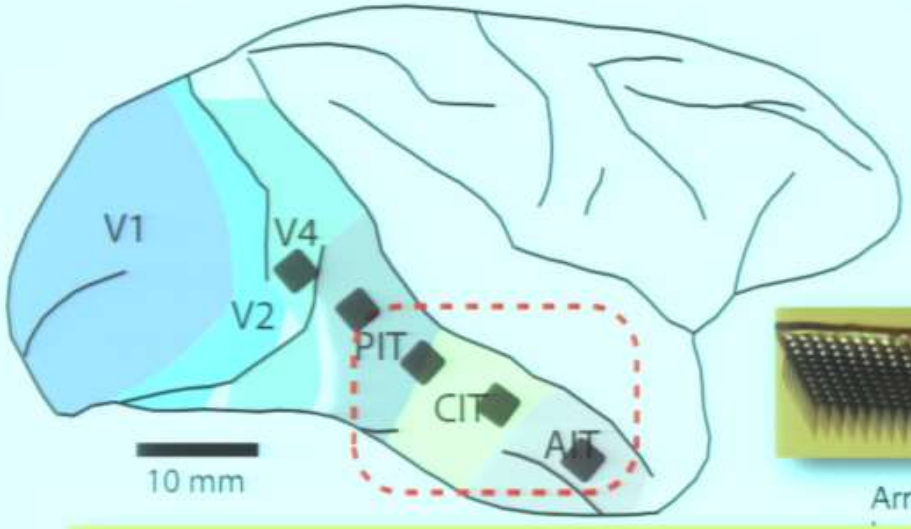
Three, 96-electrode arrays



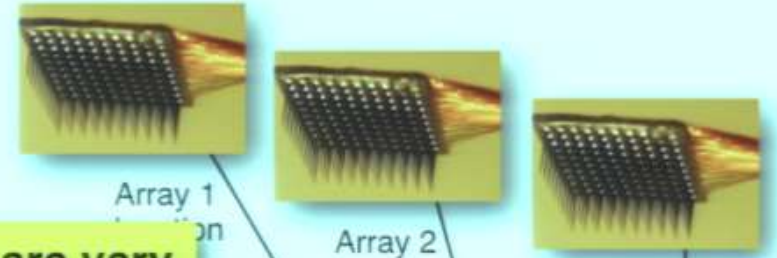




Methods advance: large scale neuronal recording along the ventral stream



Three, 96-electrode arrays

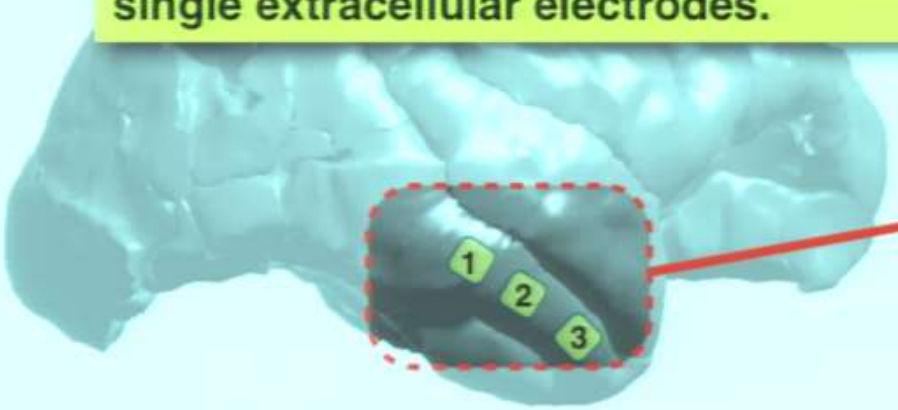


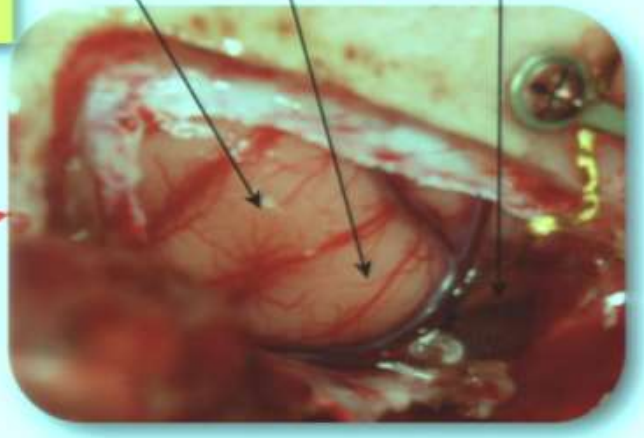
Array 1

Array 2
location

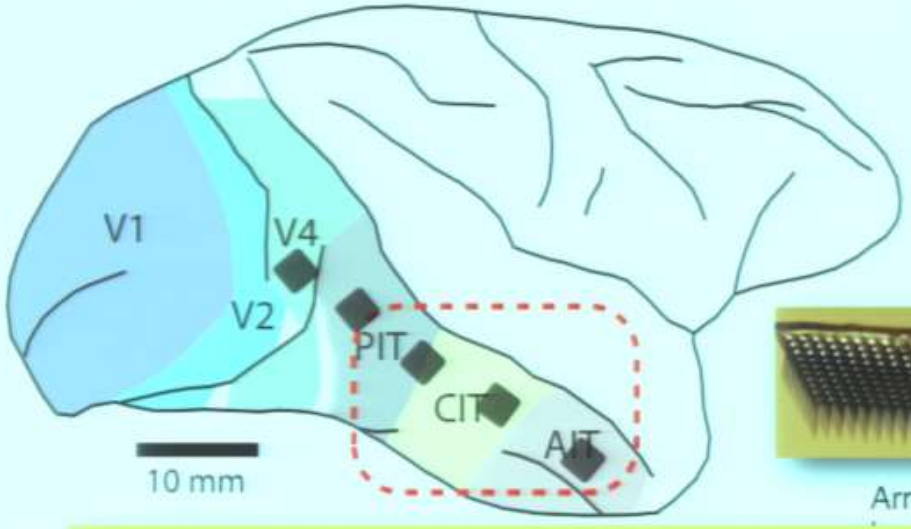
Array 3
(in place)

Neuronal selectivity properties are very comparable to those obtained with single extracellular electrodes.

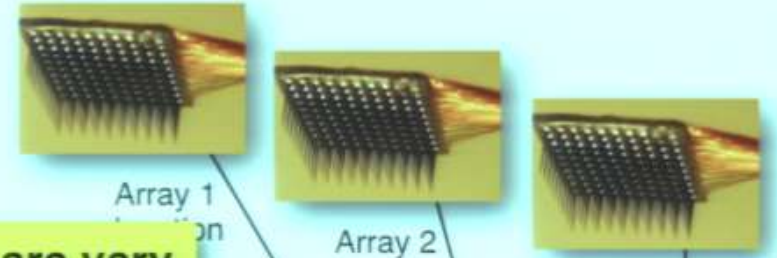




Methods advance: large scale neuronal recording along the ventral stream



Three, 96-electrode arrays



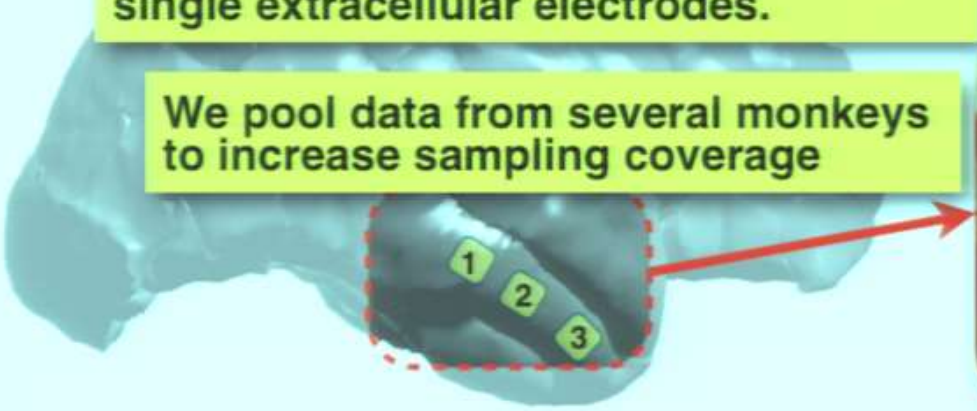
Array 1

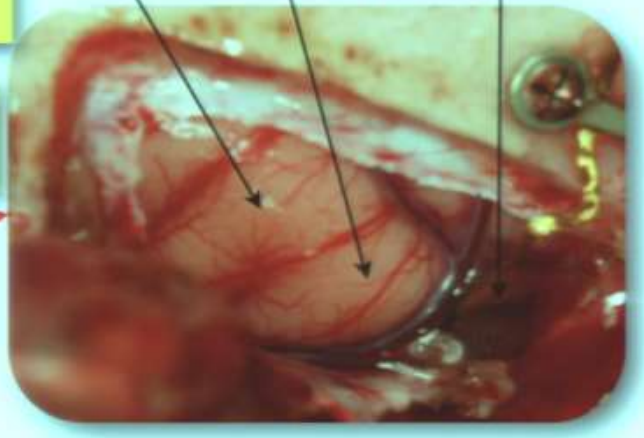
Array 2
location

Array 3
(in place)

Neuronal selectivity properties are very comparable to those obtained with single extracellular electrodes.

We pool data from several monkeys to increase sampling coverage

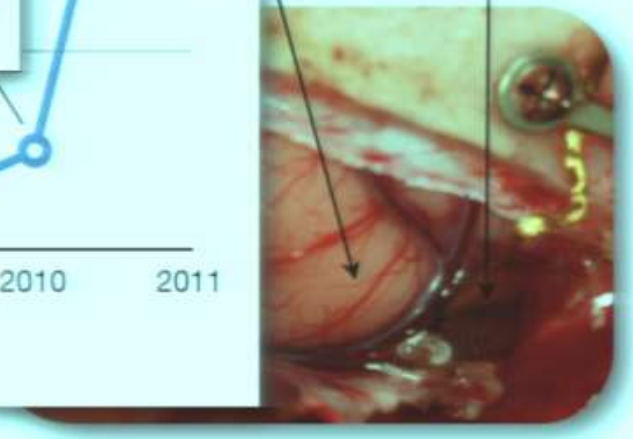


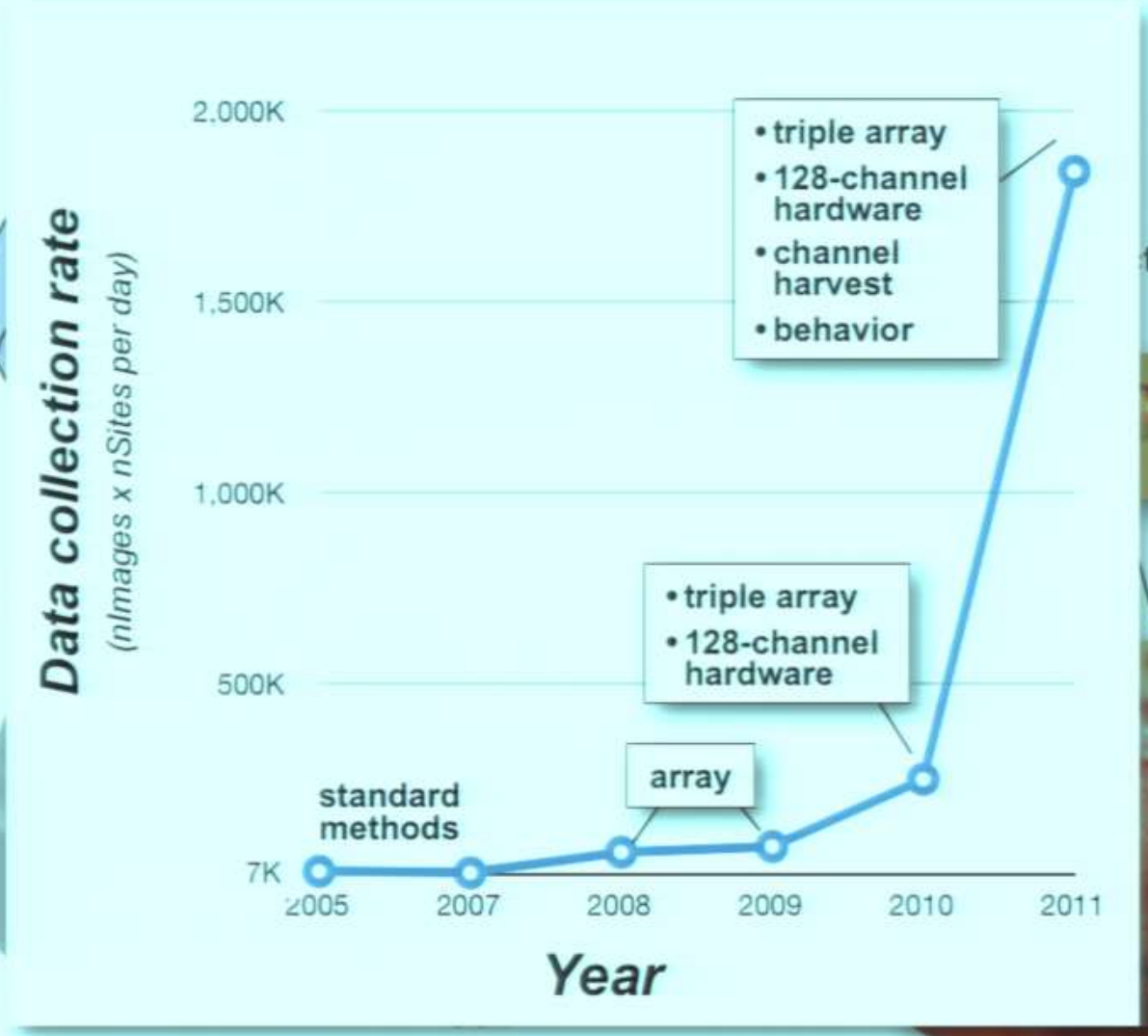


Methods advance: large scale neuronal recording along the ventral stream

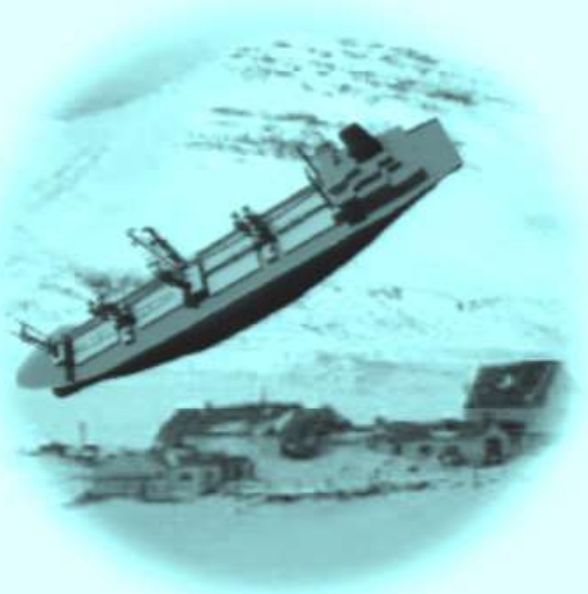
trode arrays

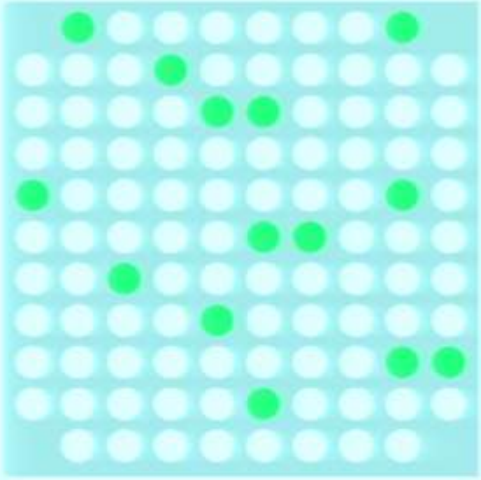
Array 3
(in place)





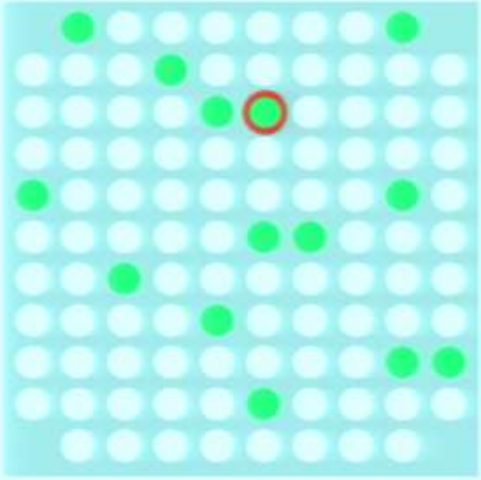
96 electrodes per array





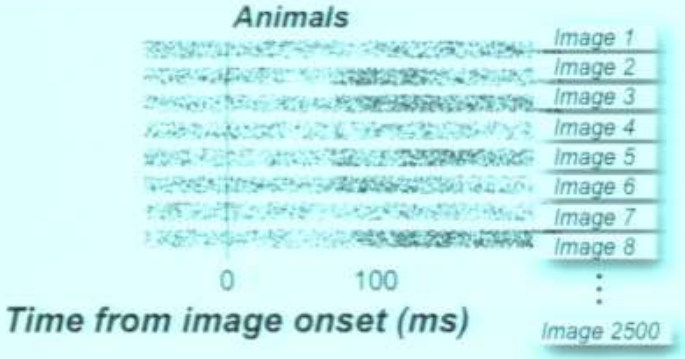
Monkey is simply fixating.
Same retinal images as human data

Example channel

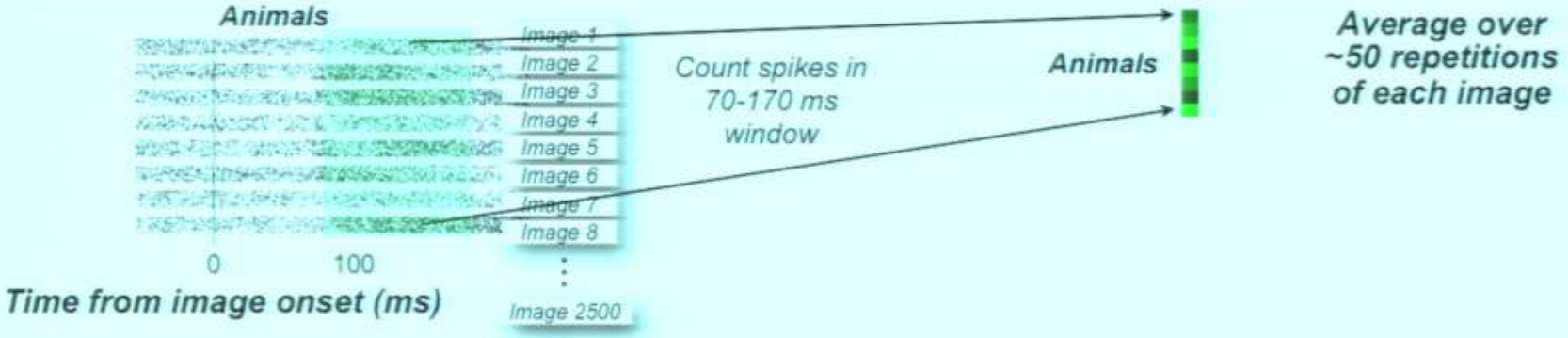


Monkey is simply fixating.
Same retinal images as human data

Example of a neuronal data volume

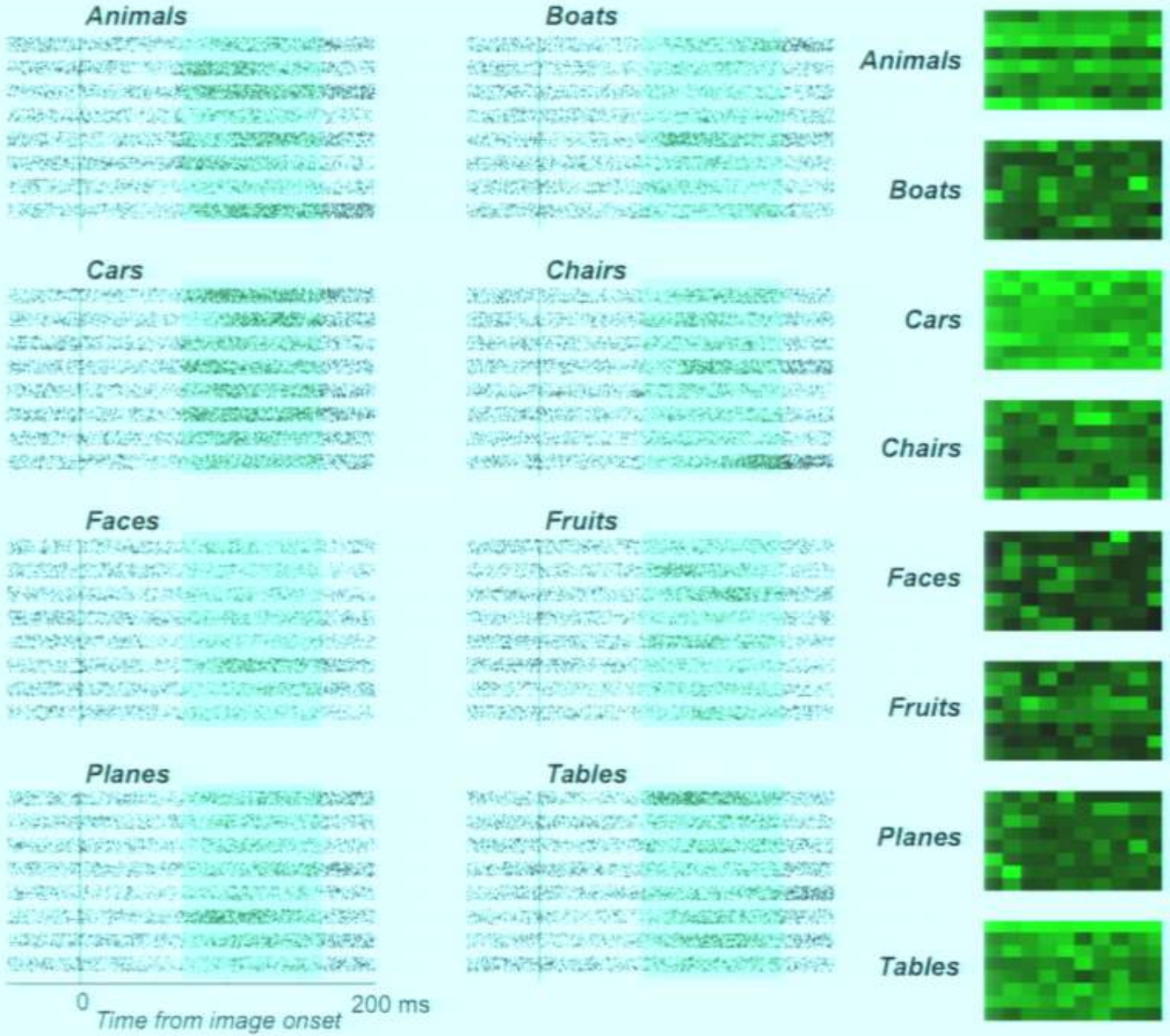


Example of a neuronal data volume

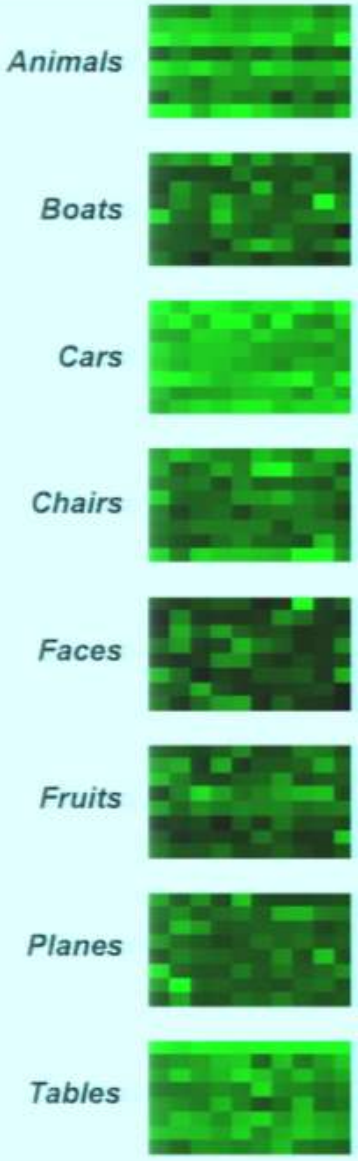


Example of a neuronal data volume

Example of a neuronal data volume



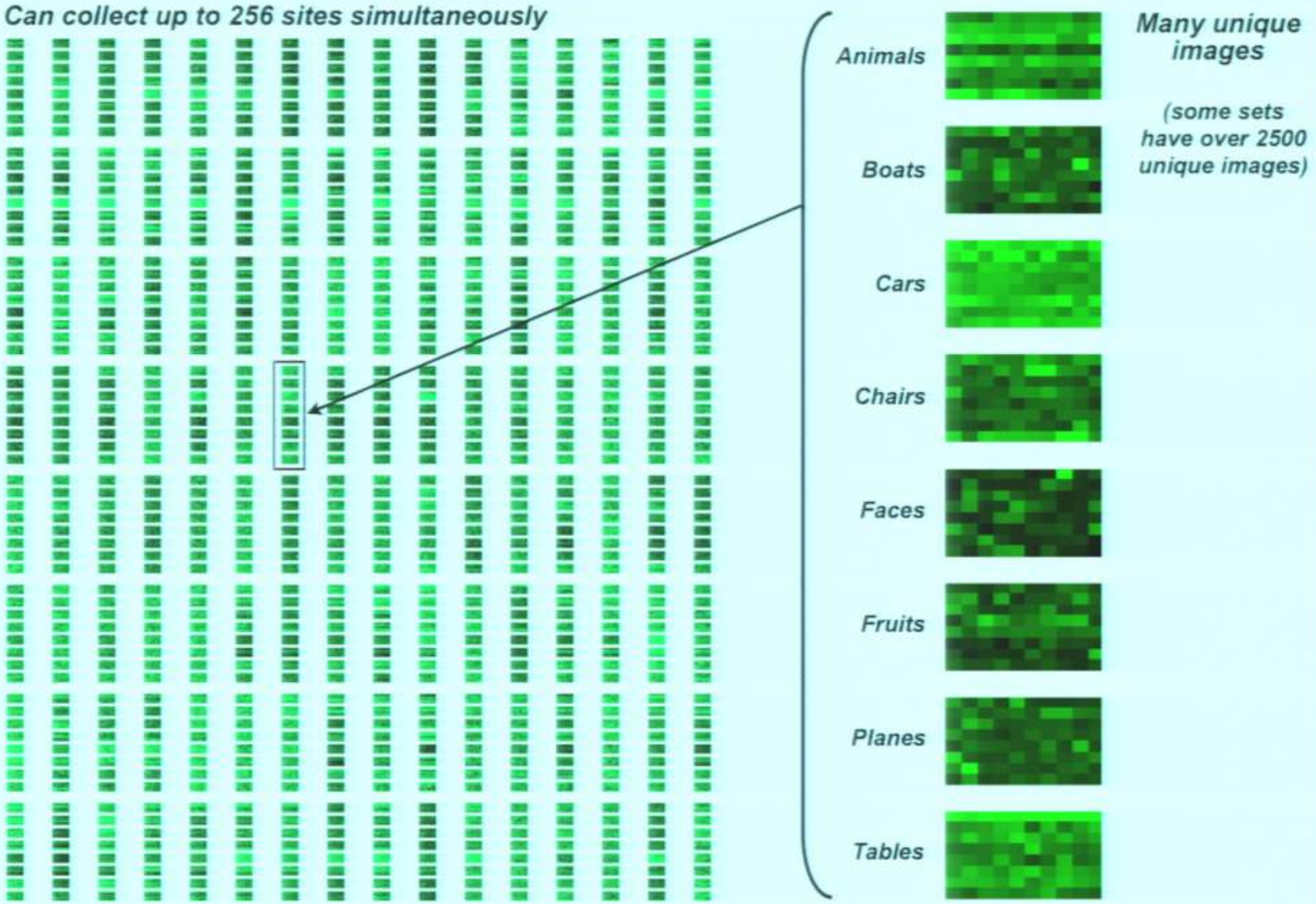
Example of a neuronal data volume



Many unique images

(some sets have over 2500 unique images)

Example of a neuronal data volume



IT Neuron

168

IT Neuron

168

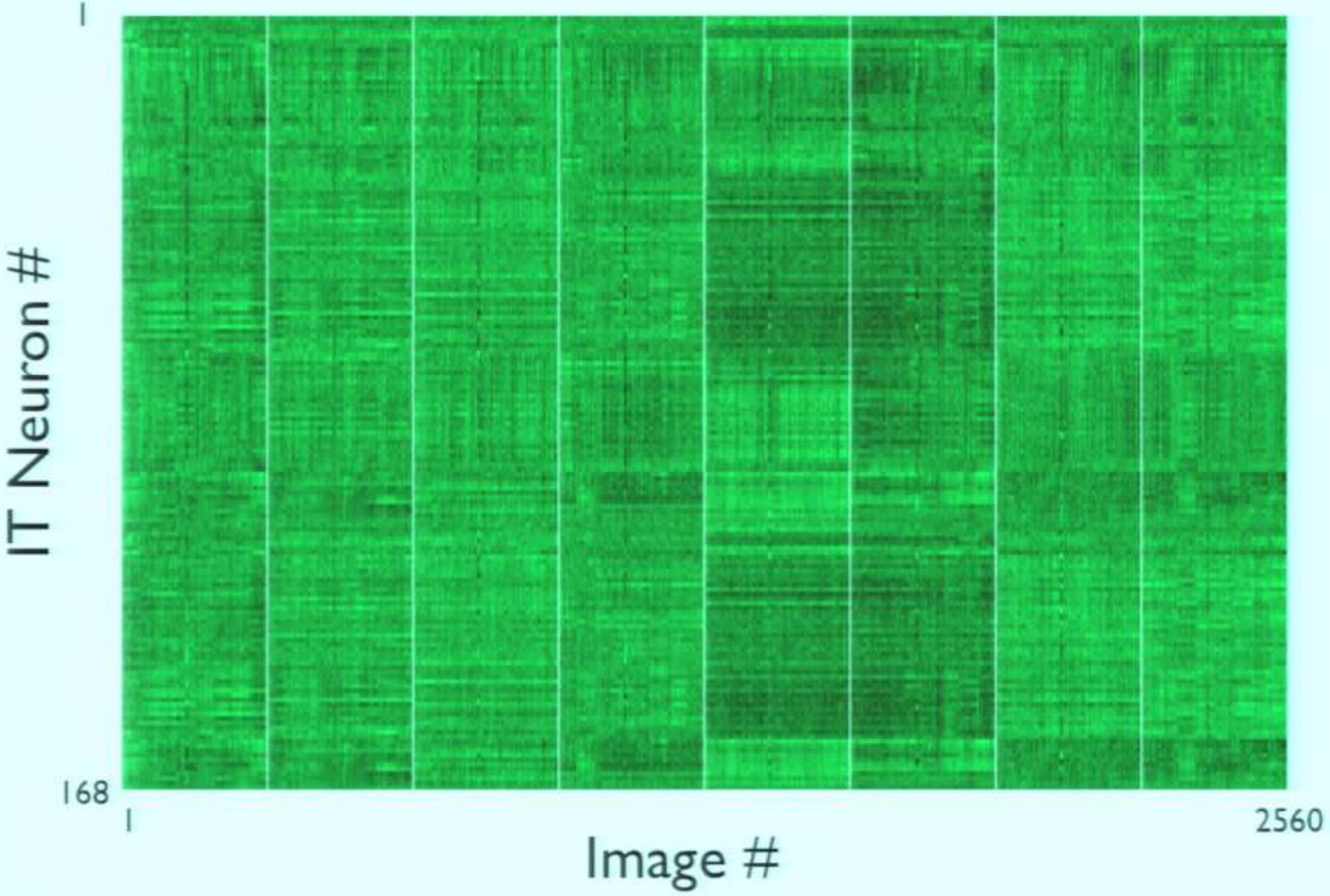
IT Neuron

168









1
IT Neuron #
168
1
2560
Image #

Are IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Passively-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

3. Measure large samples of neuronal population spiking responses

Are IT neural codes sufficient to explain human object recognition?

The simple hypothesis:

Passively-evoked spike rate codes distributed over non-human primate IT cortex can fully explain human object recognition

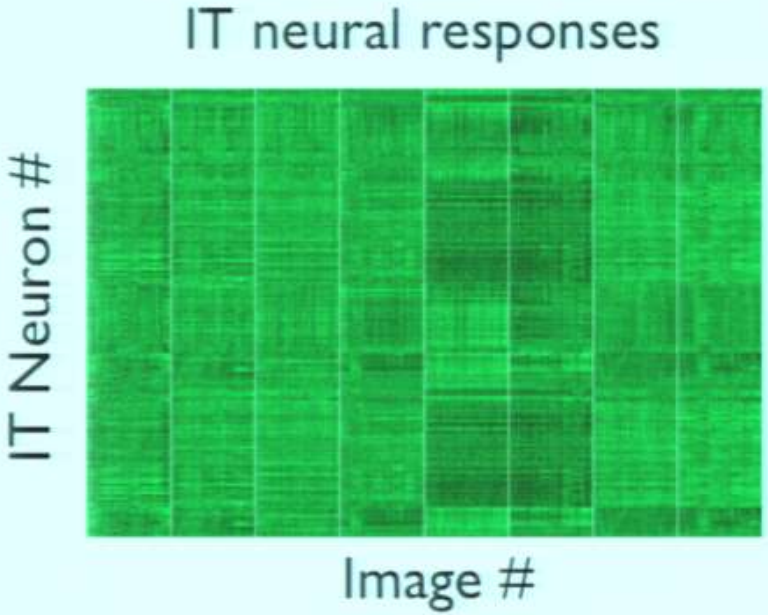
1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

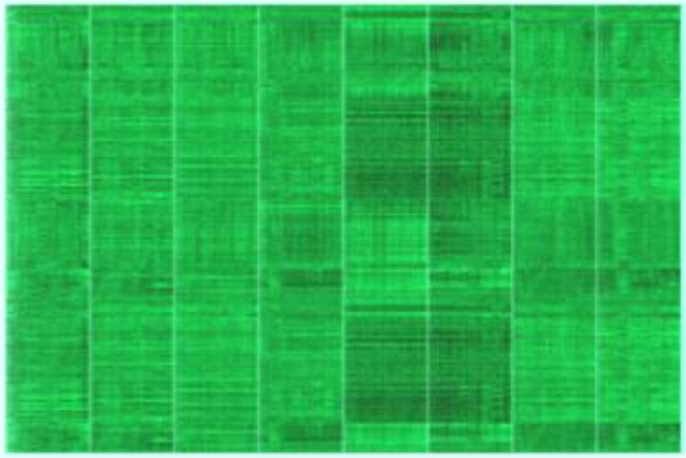
Same images

3. Measure large samples of neuronal population spiking responses

Compute predicted O.R. behavior from this neuronal activity ("codes", "decodes")

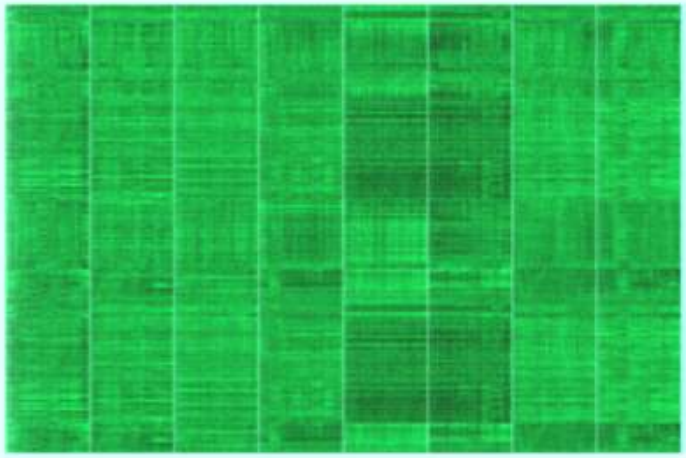

IT neural responses
IT Neuron #
Image #

IT neural responses



⇒ Does this predict performance on all our recognition tasks?

IT neural responses

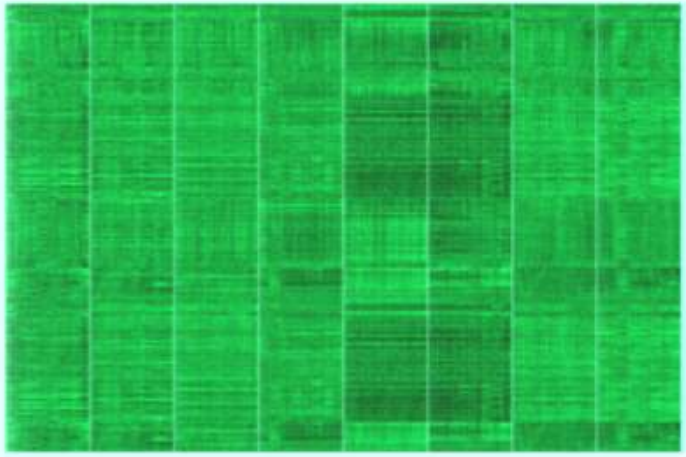


⇒ Need to predict d′ values for all 64 tasks

One decoder for each task

- Linear discriminant ("classifier")
- Learn weights that optimize performance

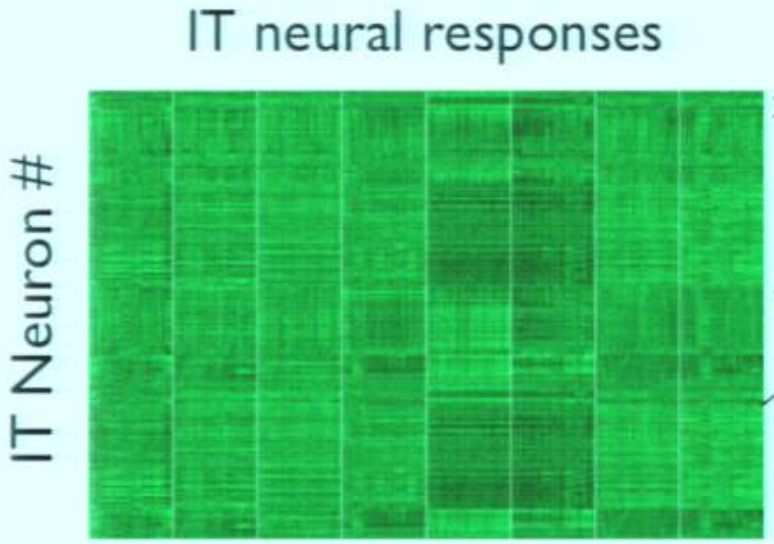
IT neural responses

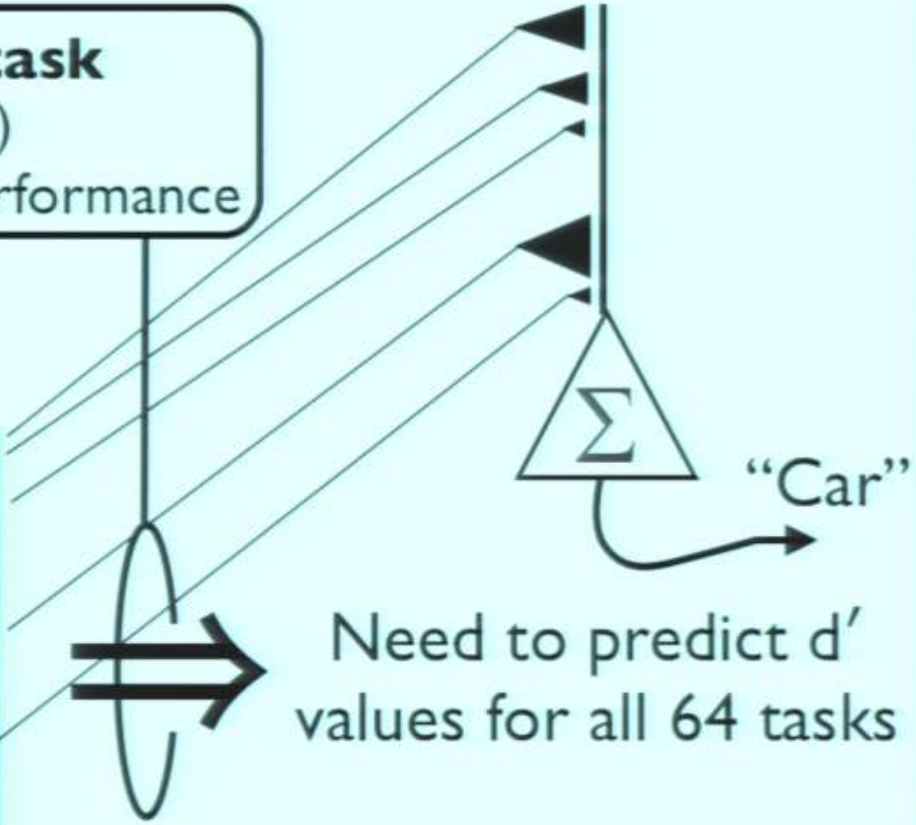


Need to predict d′ values for all 64 tasks

One decoder for each task

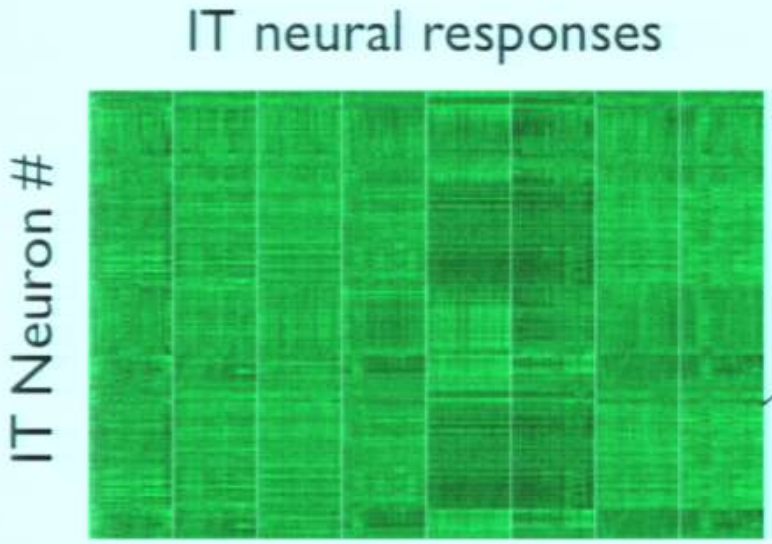
- Linear discriminant ("classifier")
- Learn weights that optimize performance

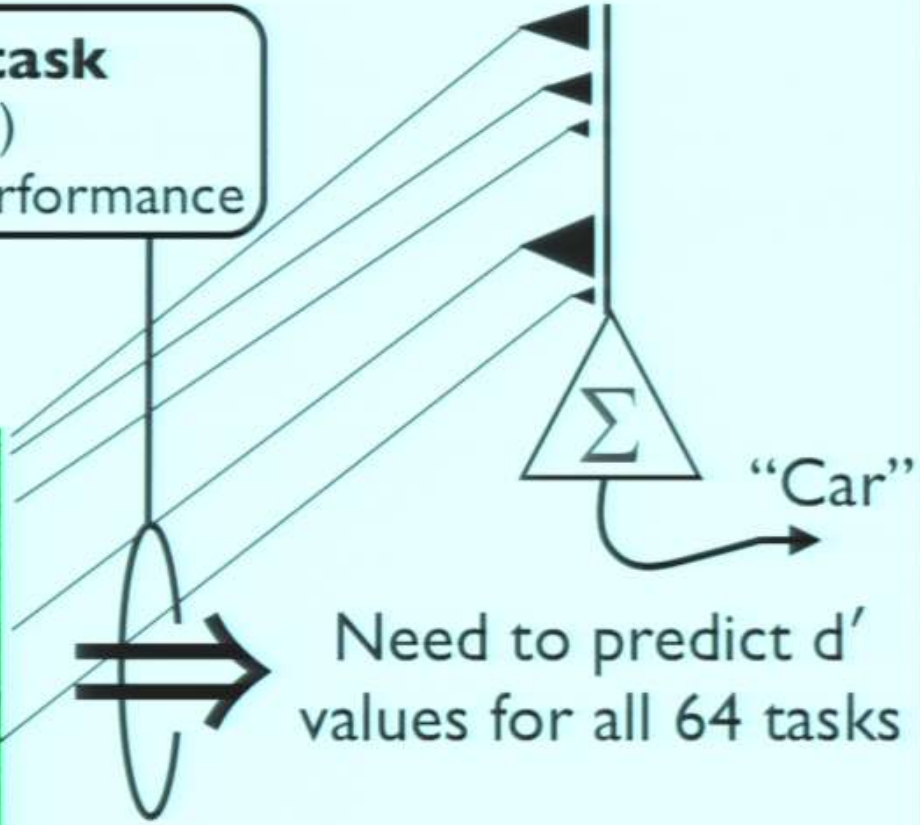




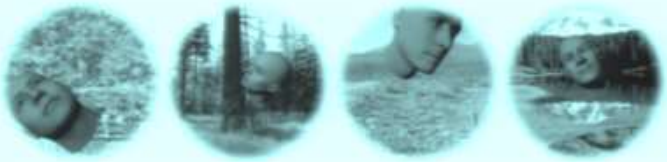
One decoder for each task

- Linear discriminant ("classifier")
- Learn weights that optimize performance





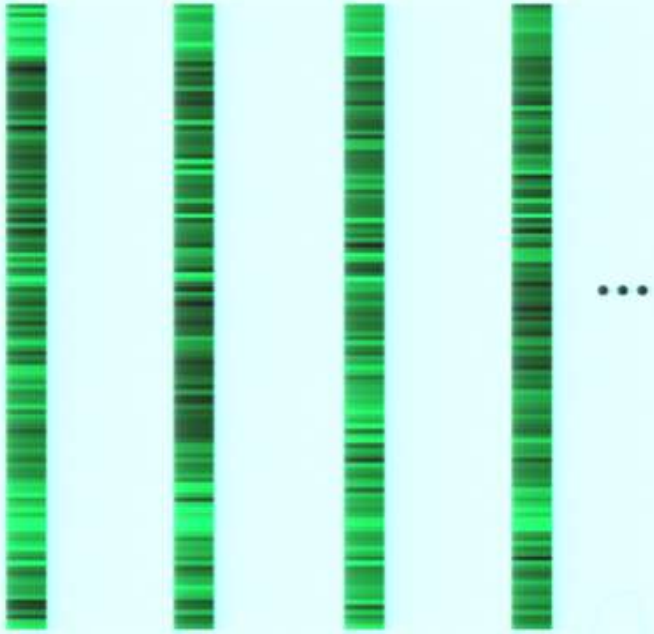
These decoders are simple, specific, instantiated hypotheses about how neuronal activity gives rise to behavior.

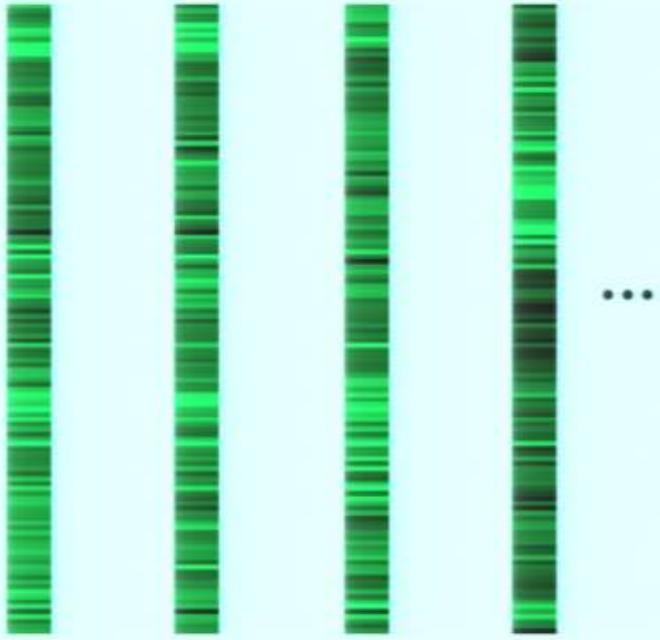


"Face"

"Not Face"

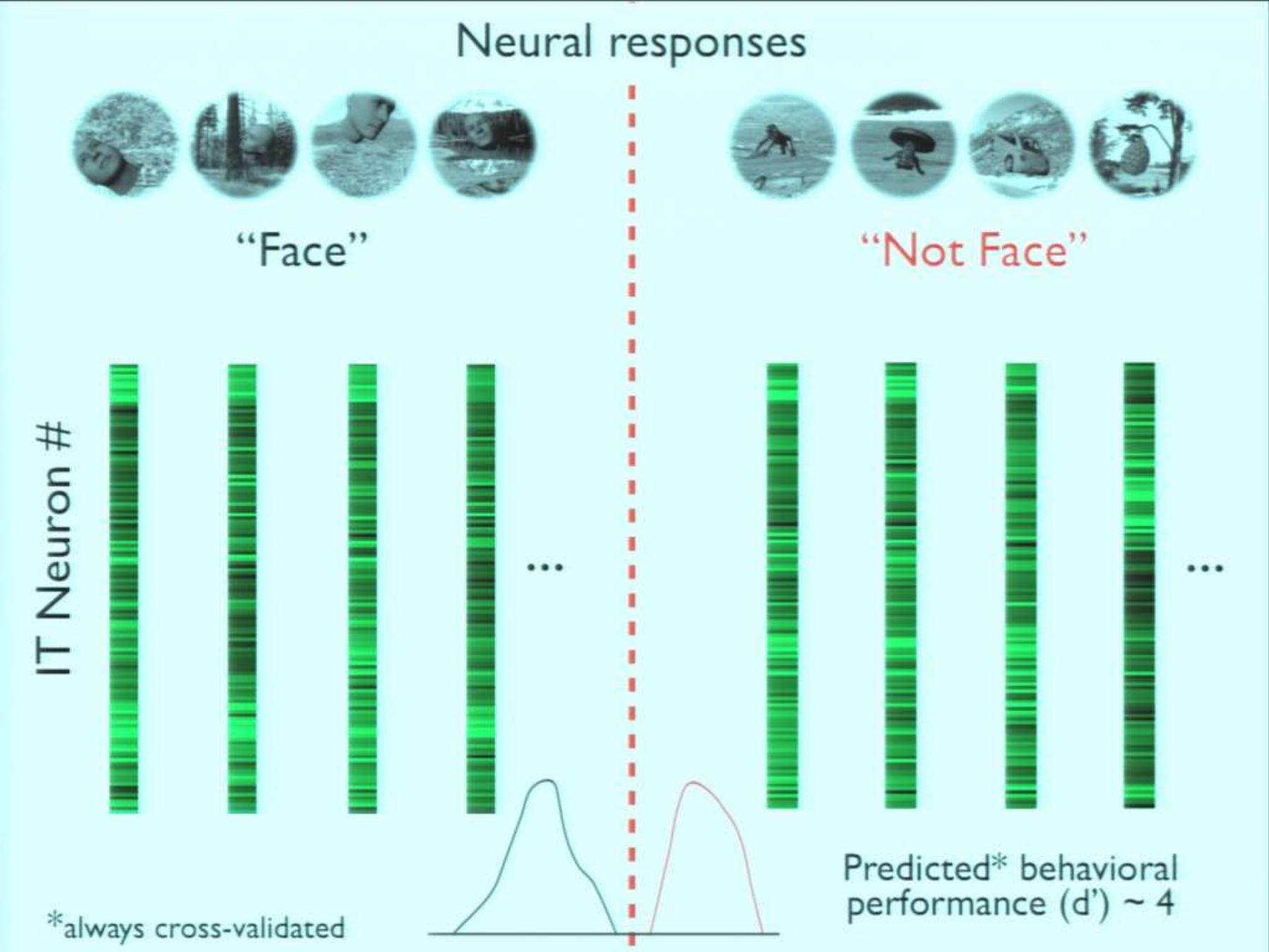
IT Neuron #

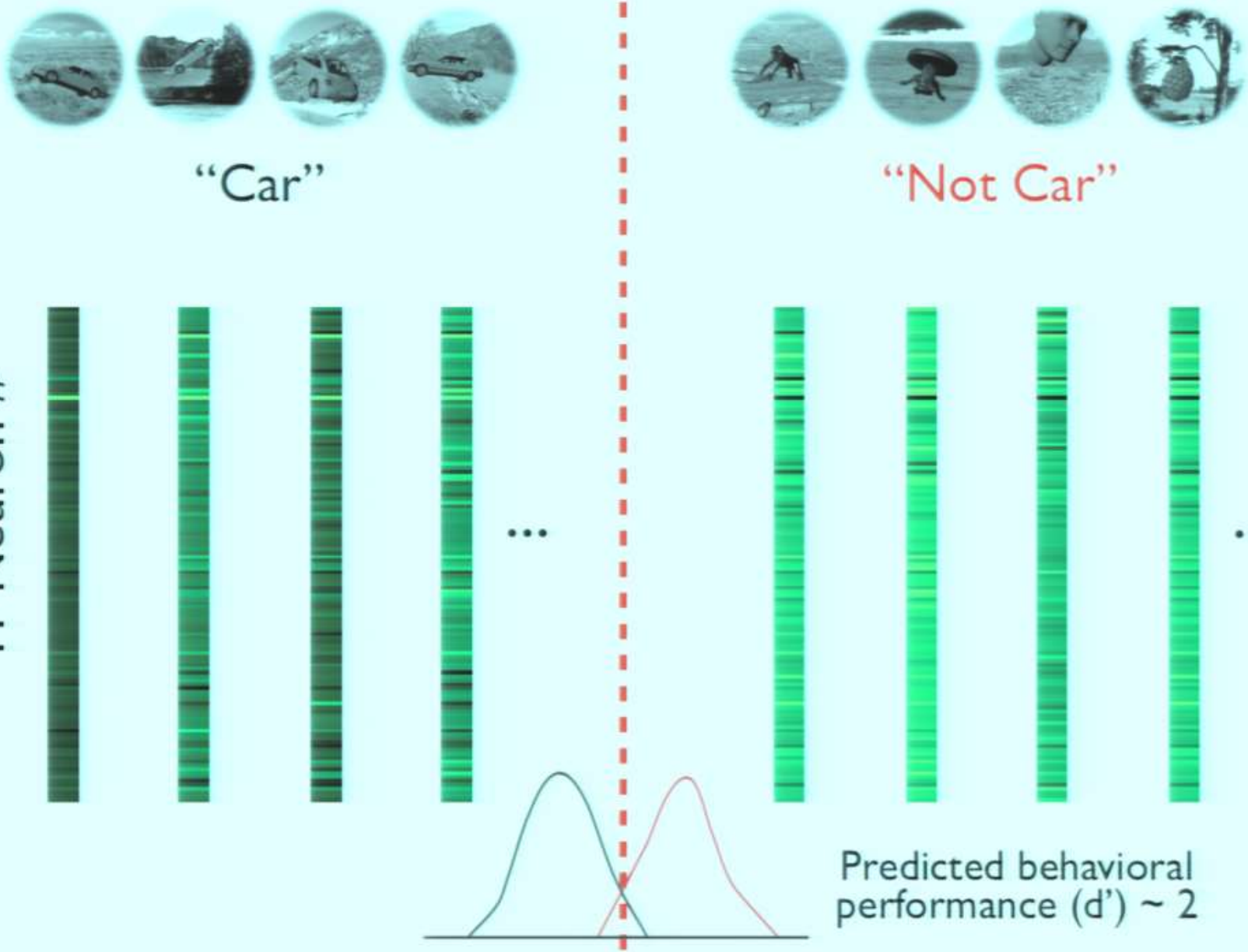




Predicted* behavioral performance (d') ~ 4

*always cross-validated


Neural responses
"Face"
"Not Face"
IT Neuron #
...
...
*always cross-validated
Predicted* behavioral performance (d') ~ 4


"Car"
"Not Car"
IT Neuron #
...
...
Predicted behavioral
performance (d') ~ 2

Are any IT neural codes sufficient to explain human object recognition?

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

3. Measure large samples of neuronal population spiking responses

Compute predicted O.R. behavior from this neuronal activity ("codes", "decodes")

Are any IT neural codes sufficient to explain human object recognition?

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

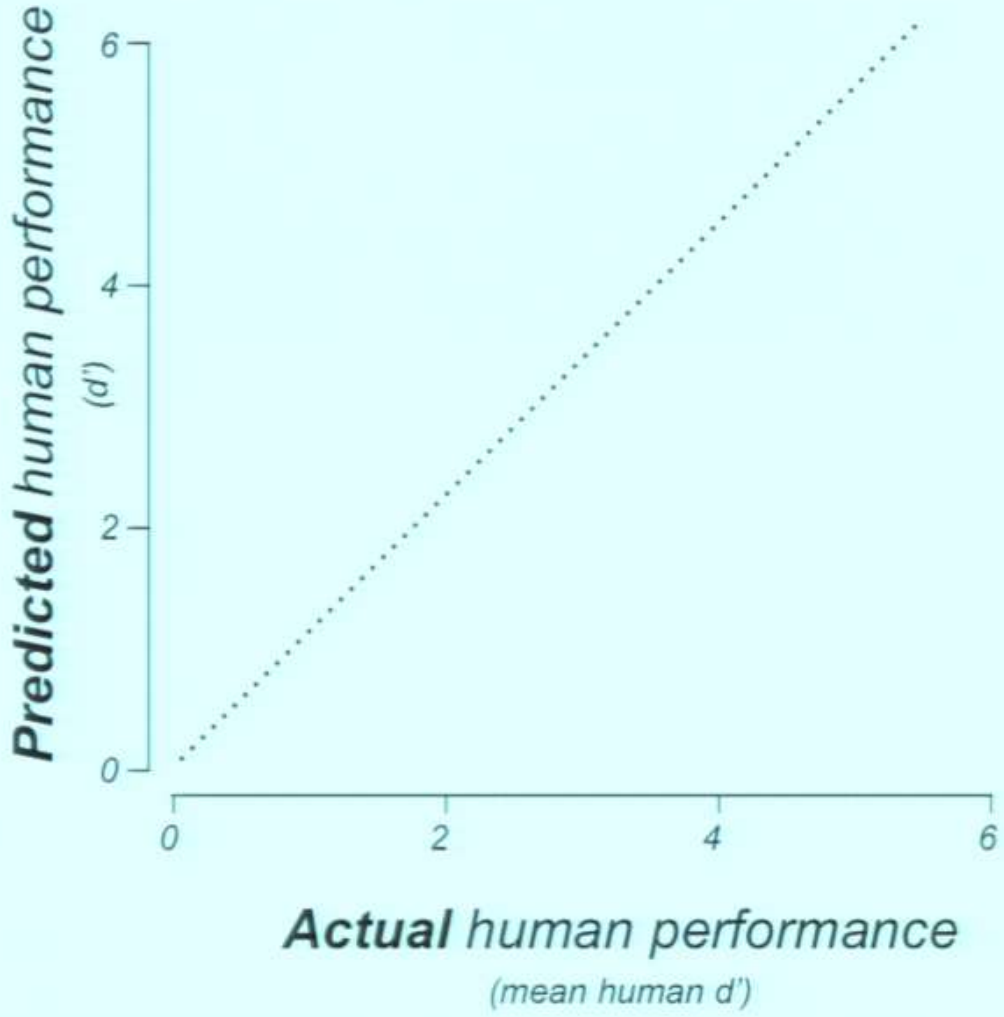
3. Measure large samples of neuronal population spiking responses

?

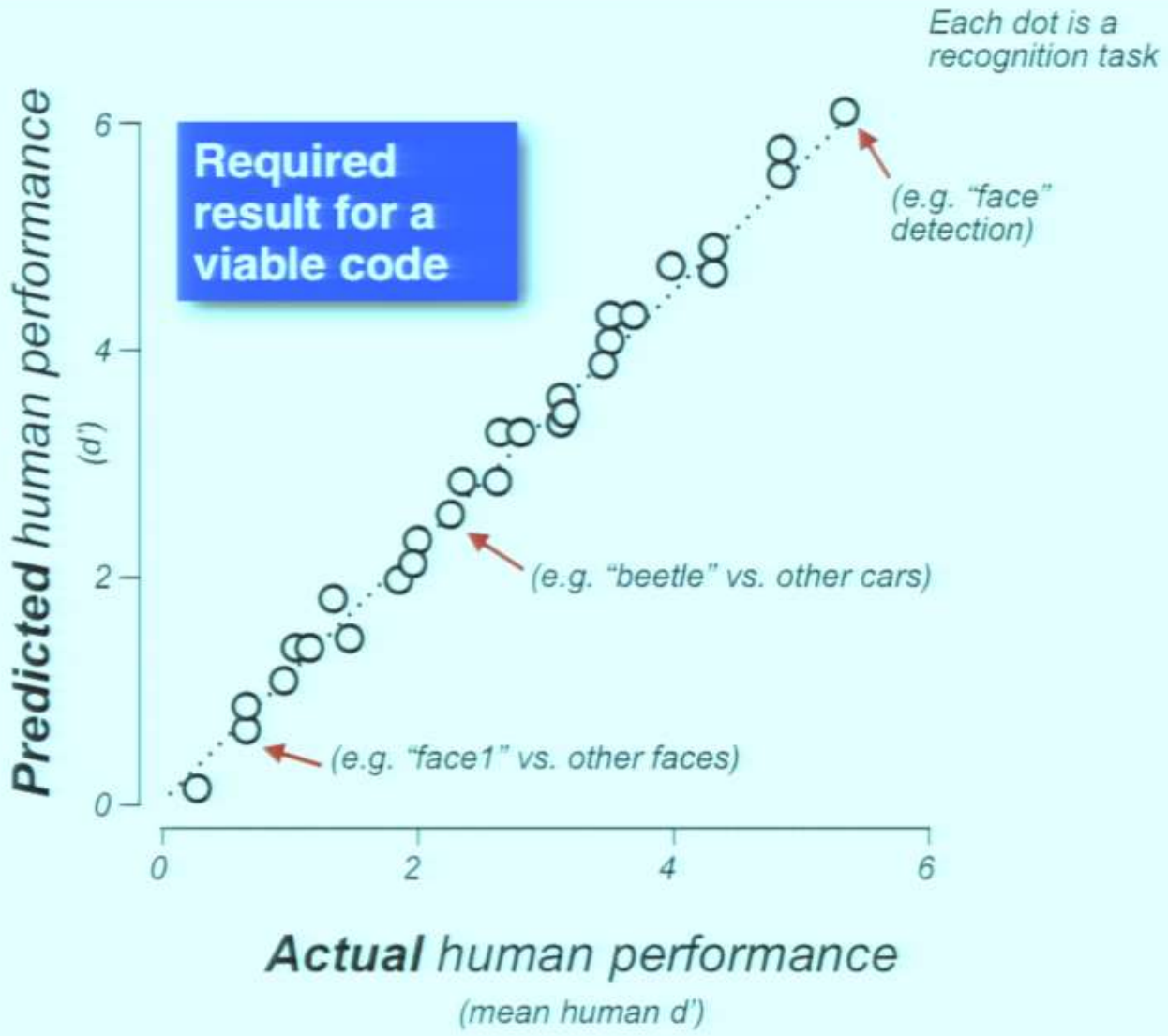
4. Ask: can the proposed link quantitatively predict O.R. behavior ?

Compute predicted O.R. behavior from this neuronal activity ("codes", "decodes")

Test ANY putative visual "code" over a battery of recognition tasks

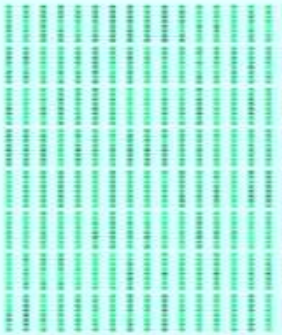


Test ANY putative visual "code" over a battery of recognition tasks

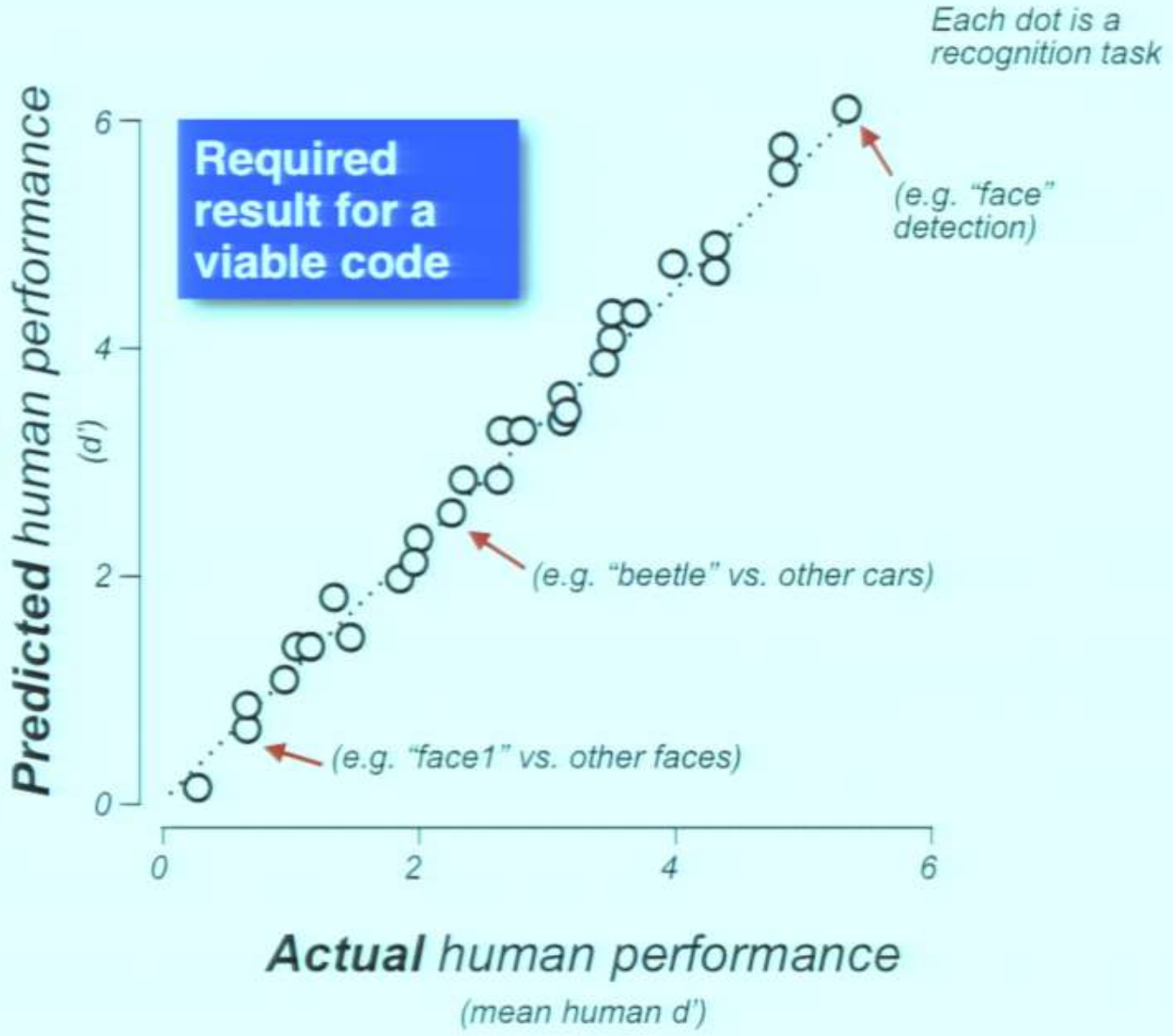


Test ANY putative visual "code" over a battery of recognition tasks

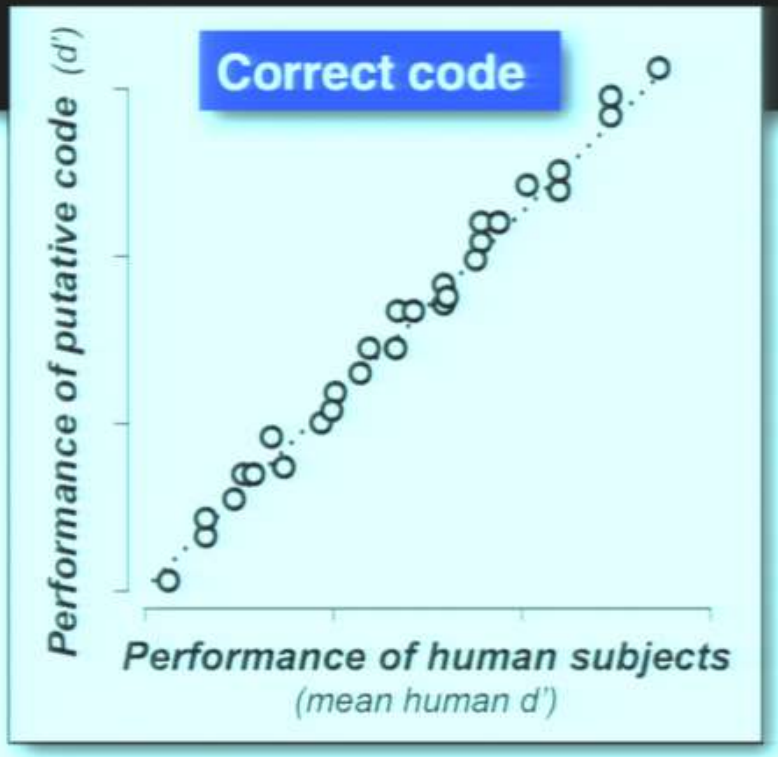
Specific hypothesized code



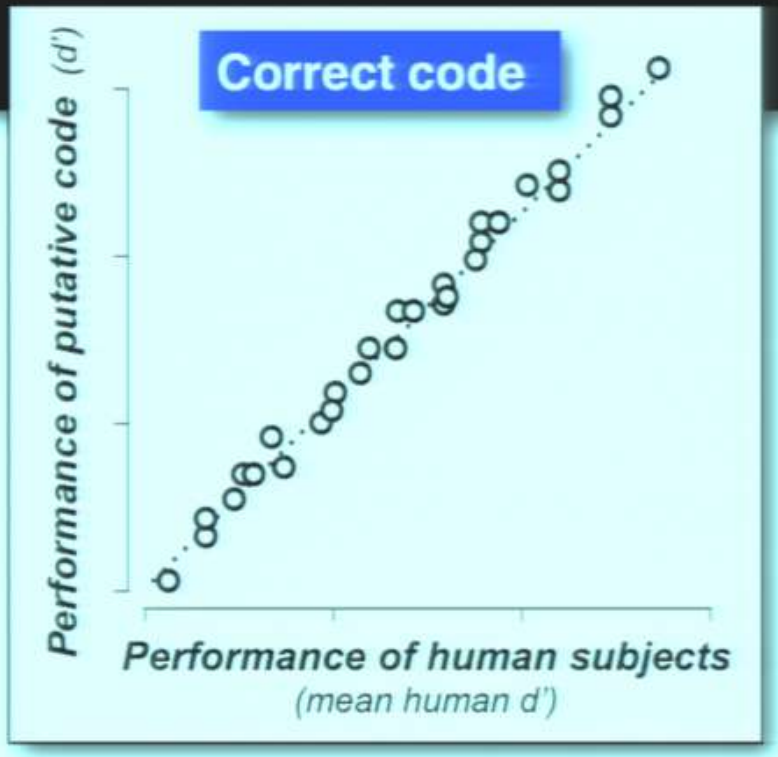
IT neuronal data

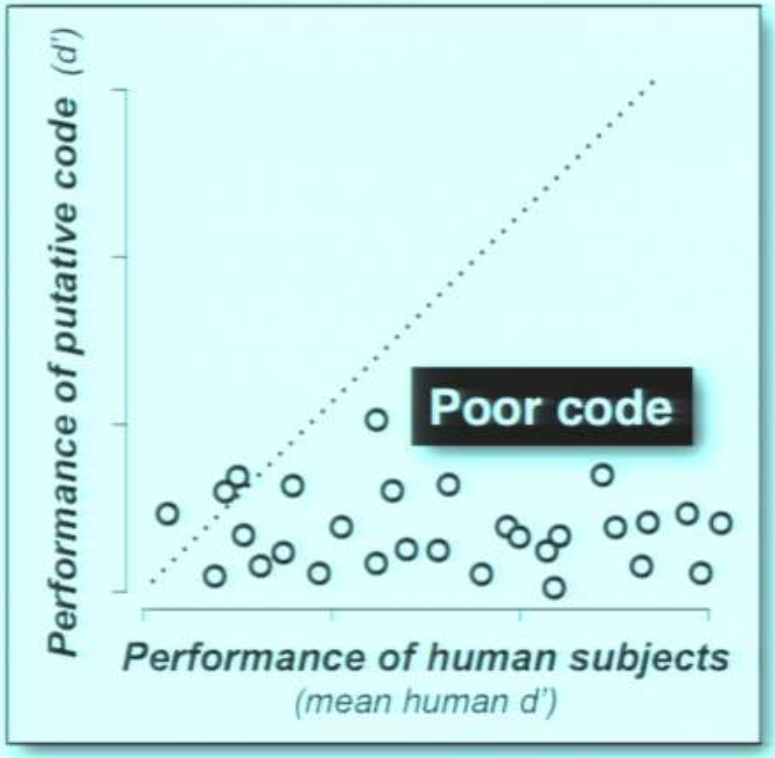


Other possible results we might find.

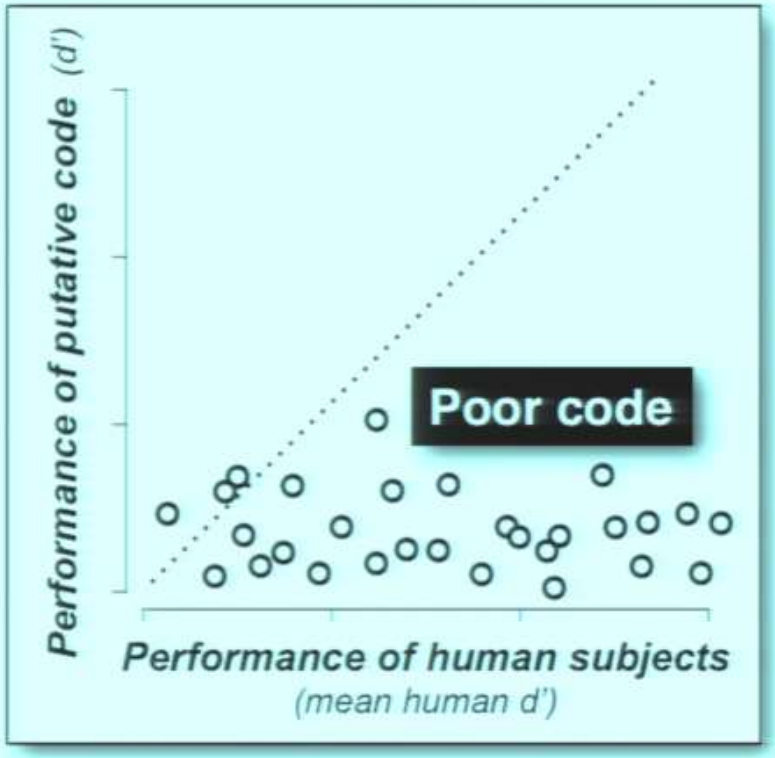


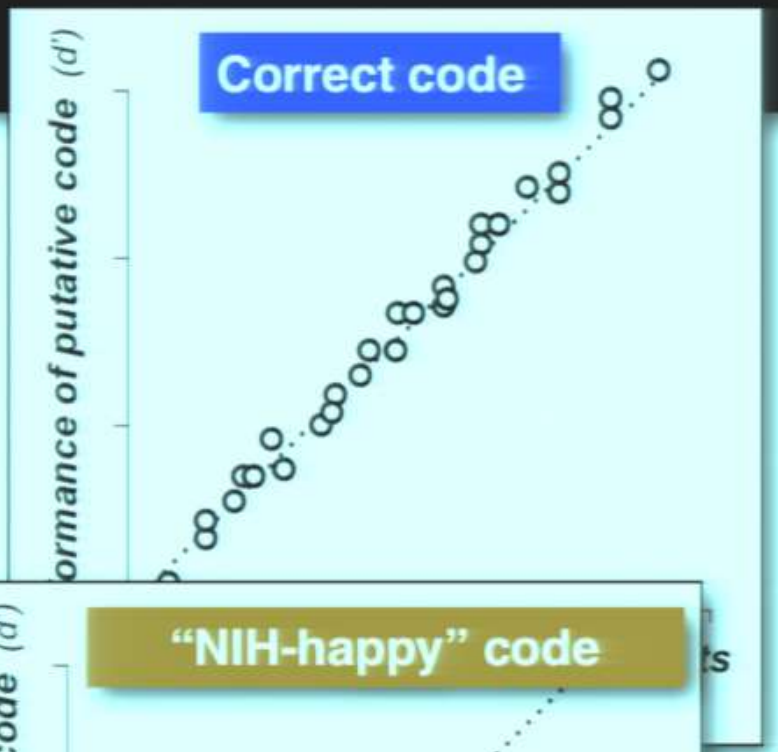
Other possible results we might find.

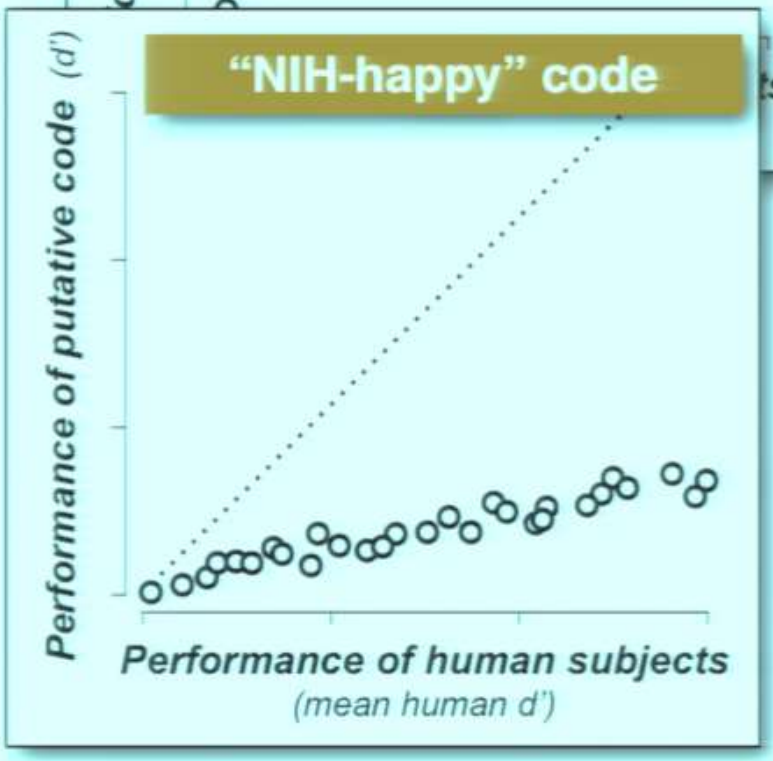




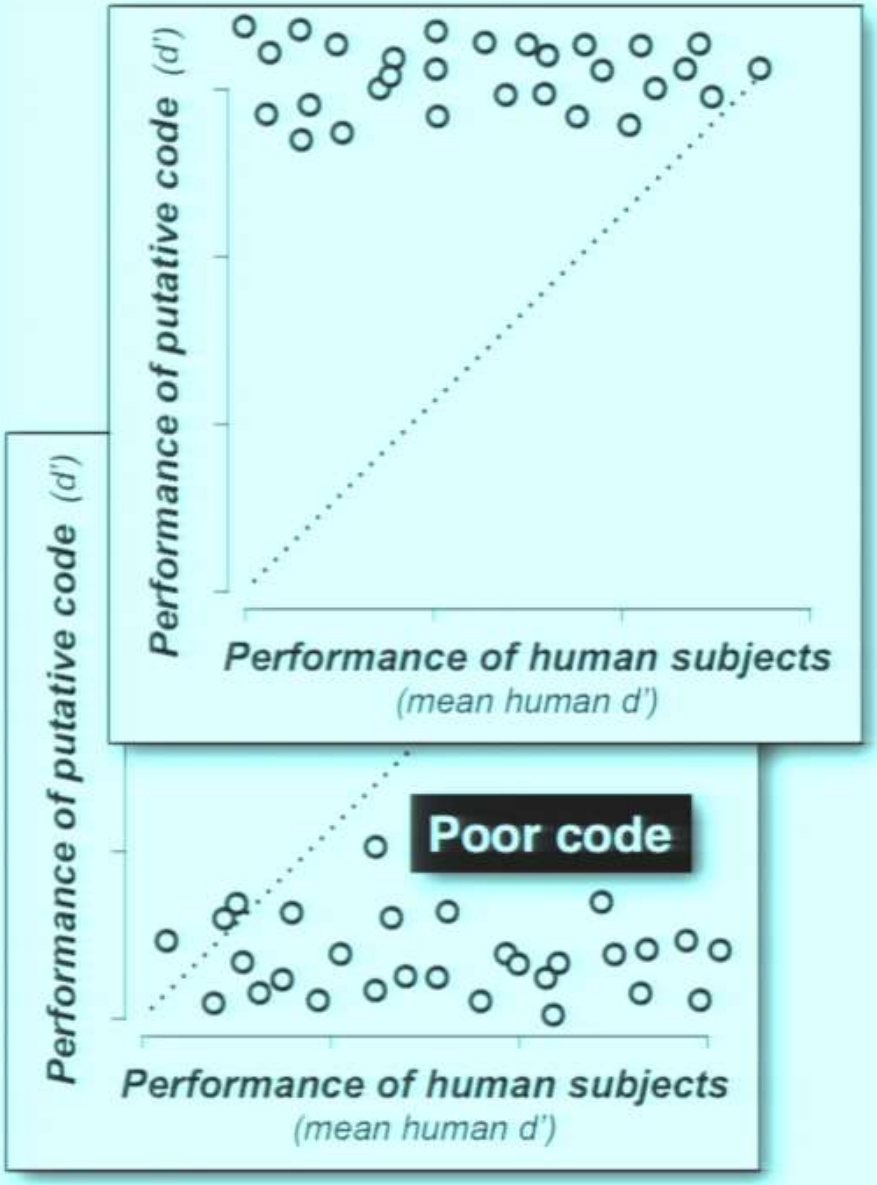
Other possible results we might find.

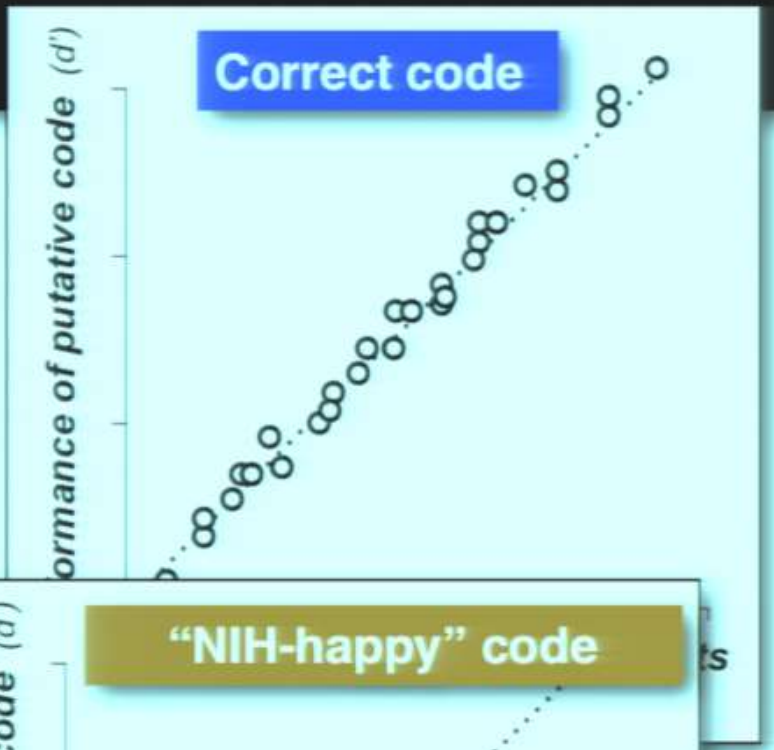


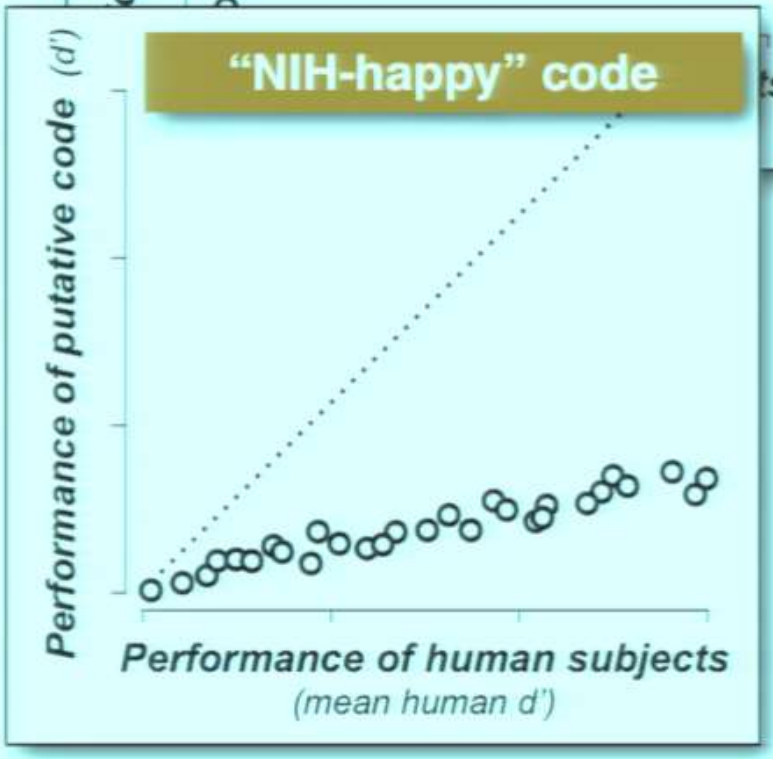




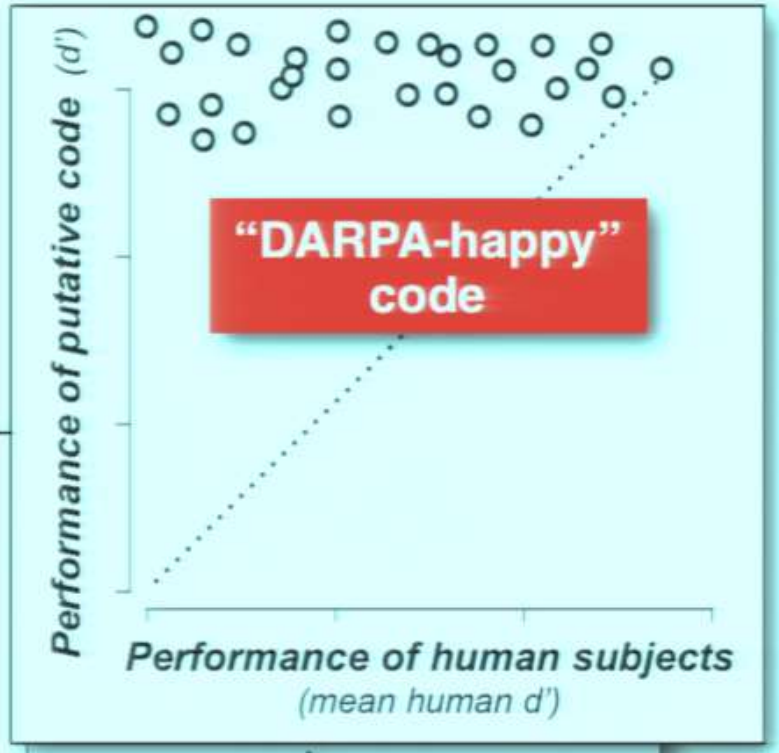
Other possible results we might find.

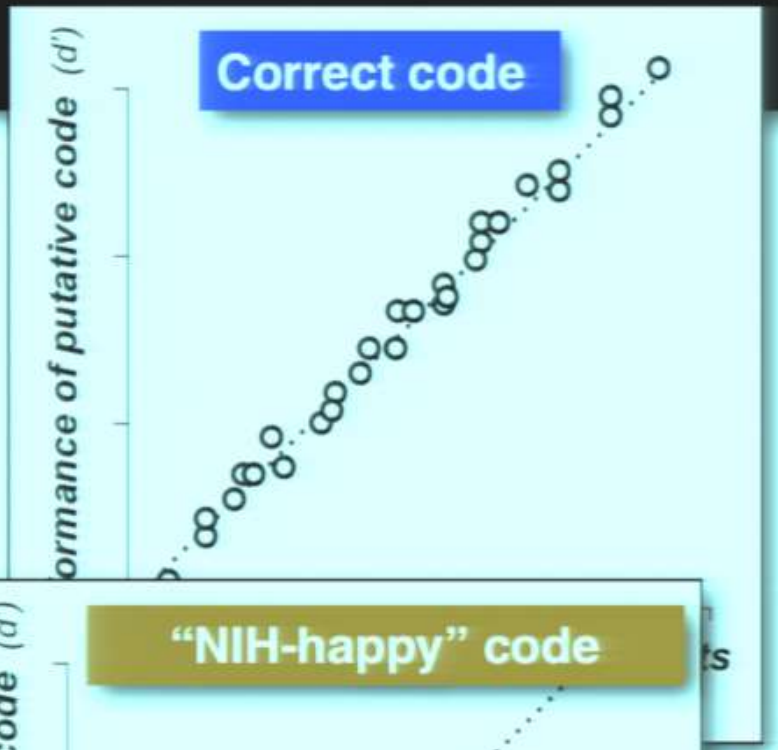


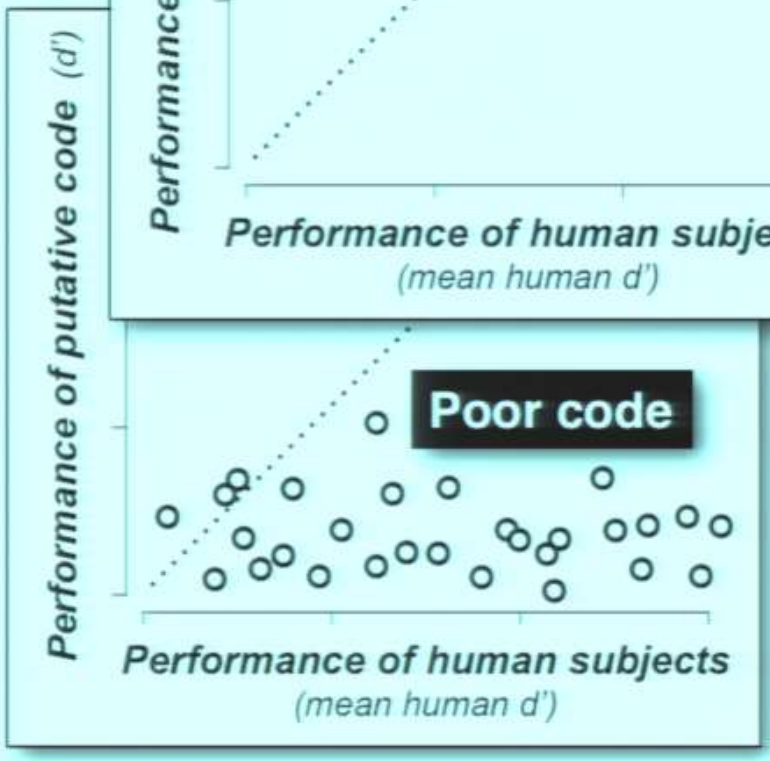


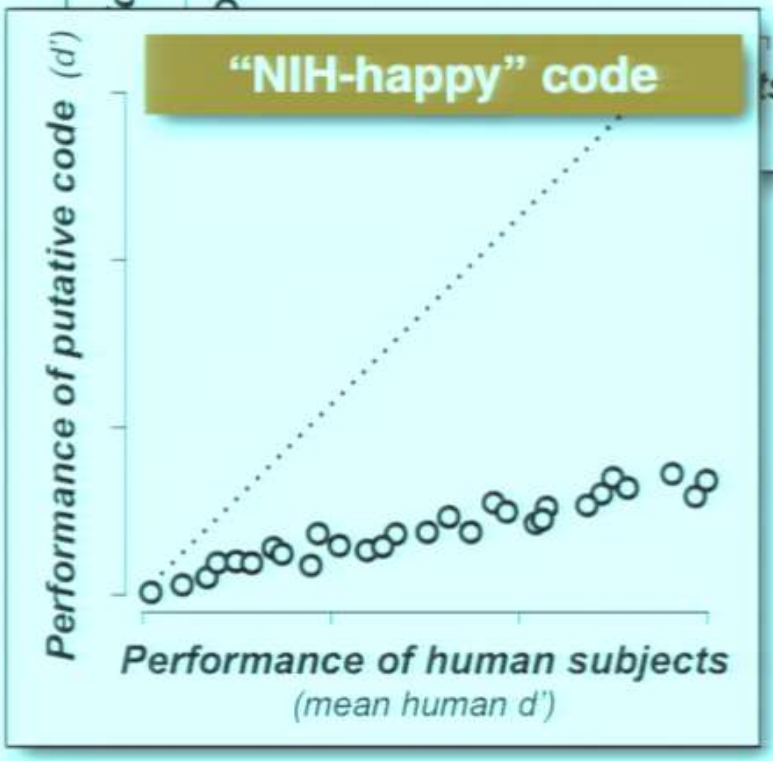


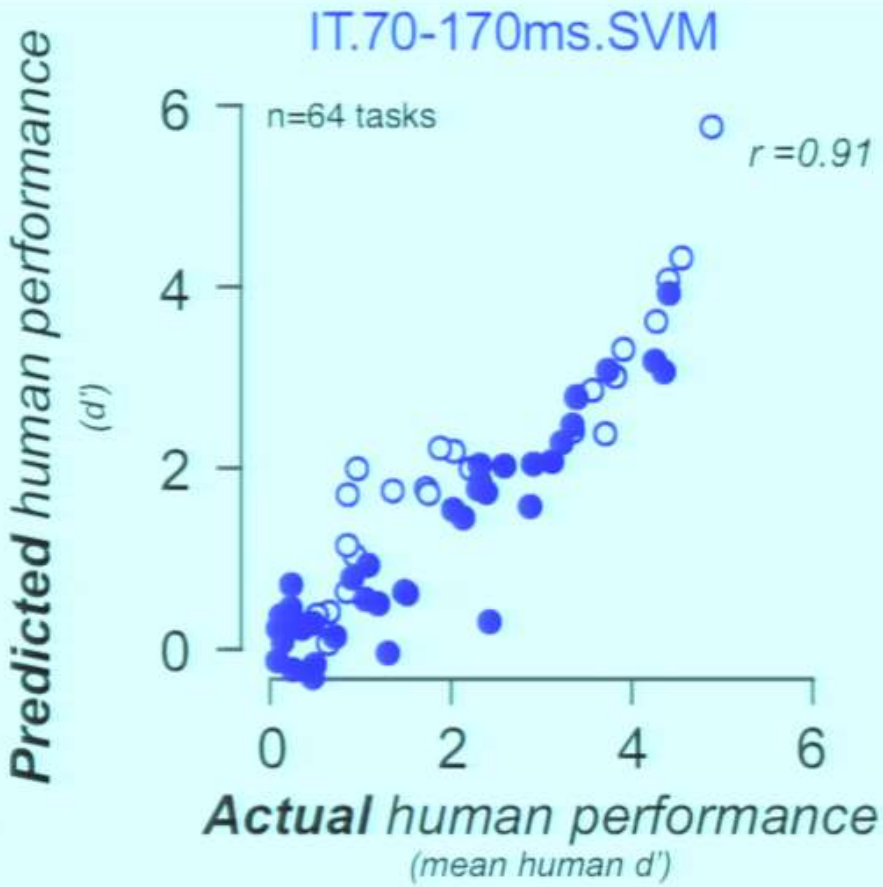
Other possible results we might find.

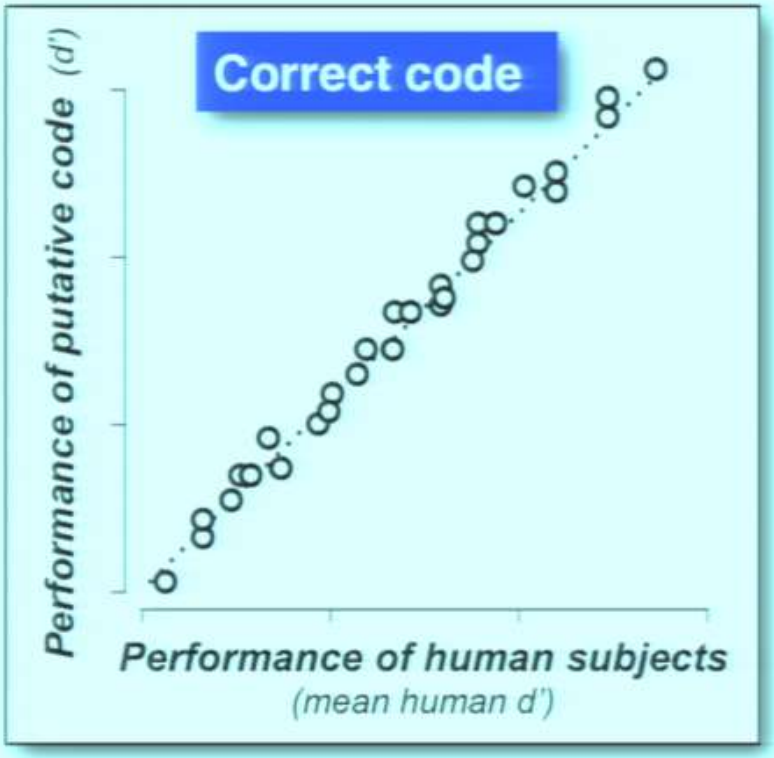


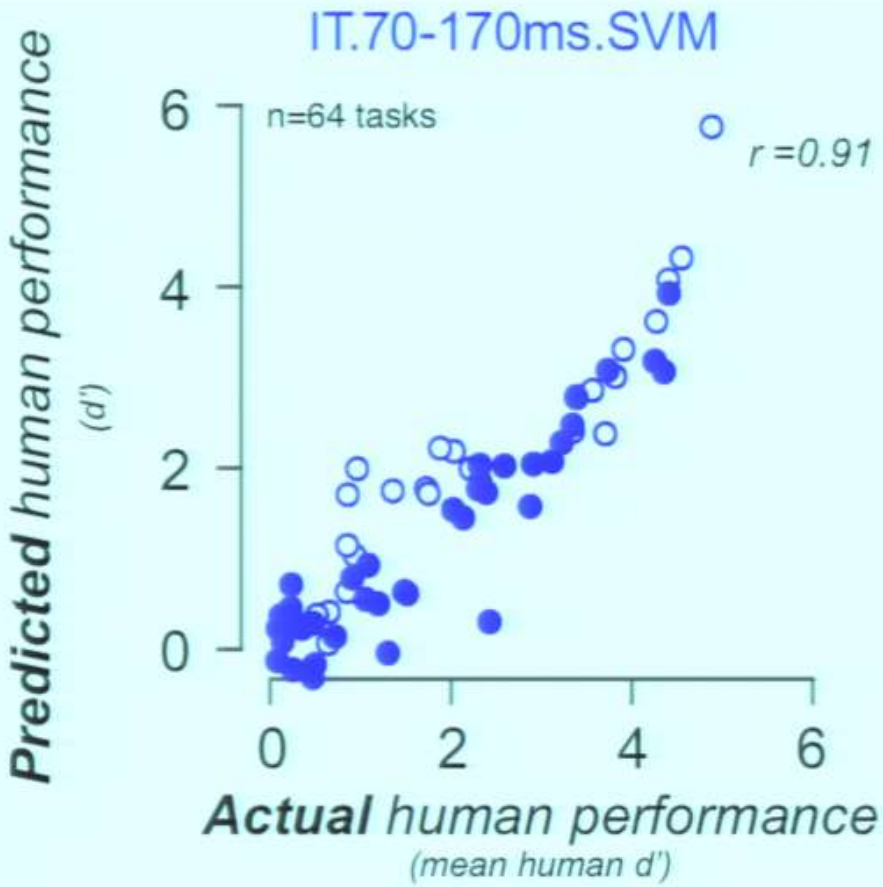


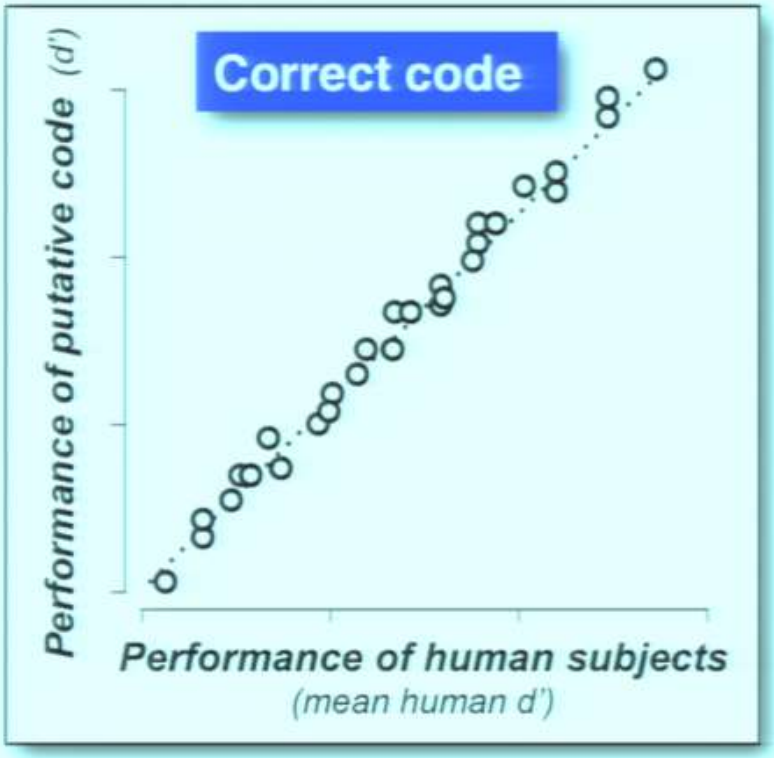


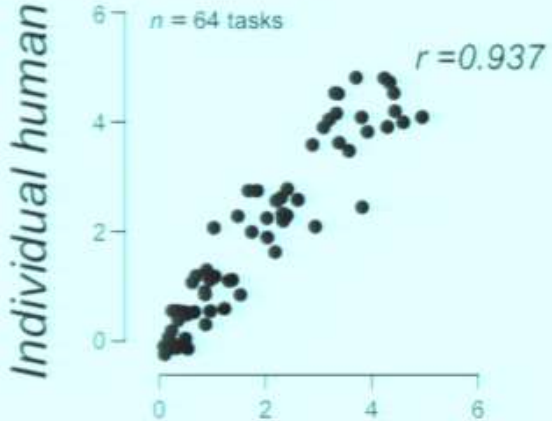


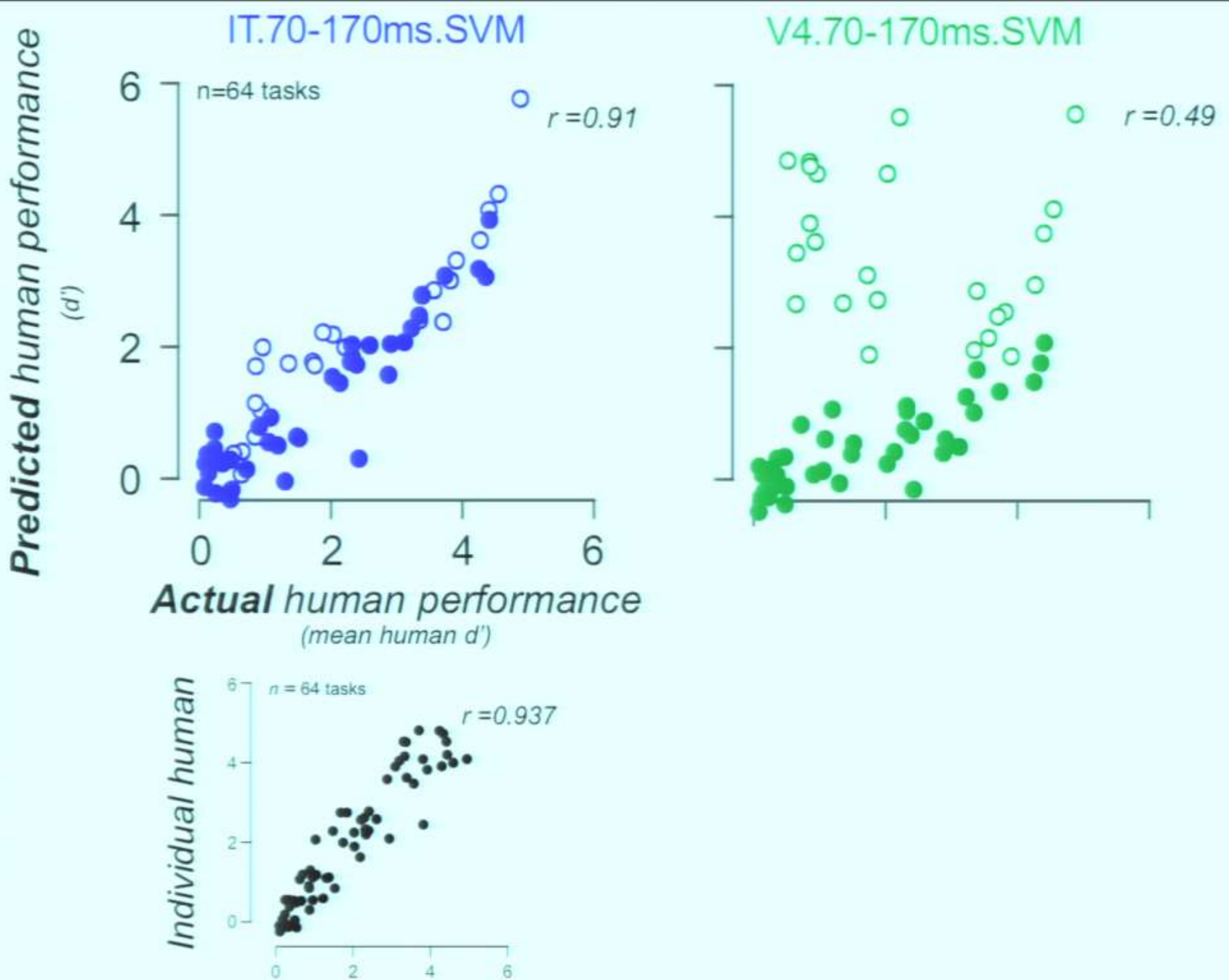

IT.70-170ms.SVM
n=64 tasks
r =0.91
Predicted human performance
(d')
Actual human performance
(mean human d')
0
2
4
6

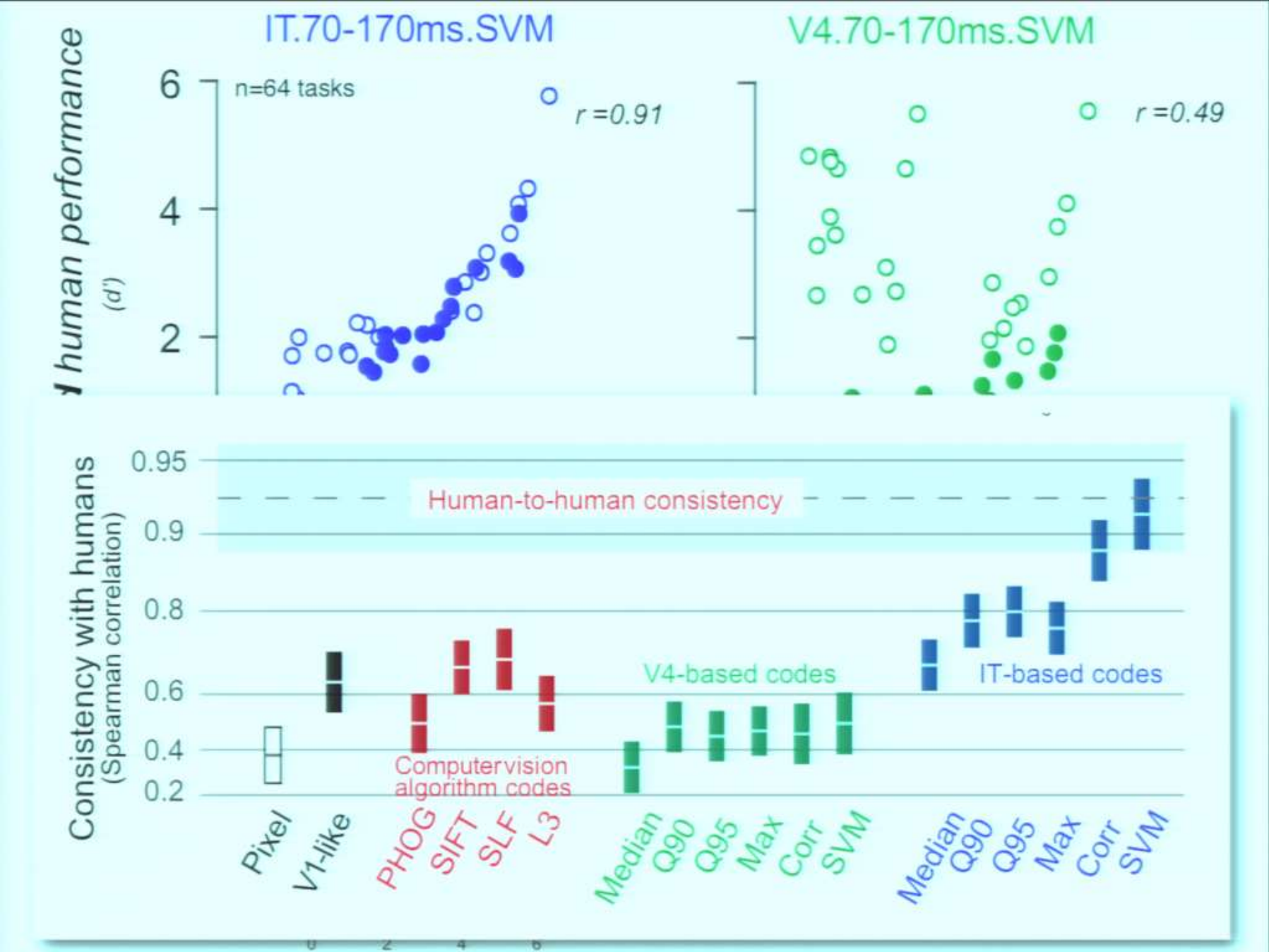

Correct code
Performance of putative code (d')
Performance of human subjects
(mean human d')

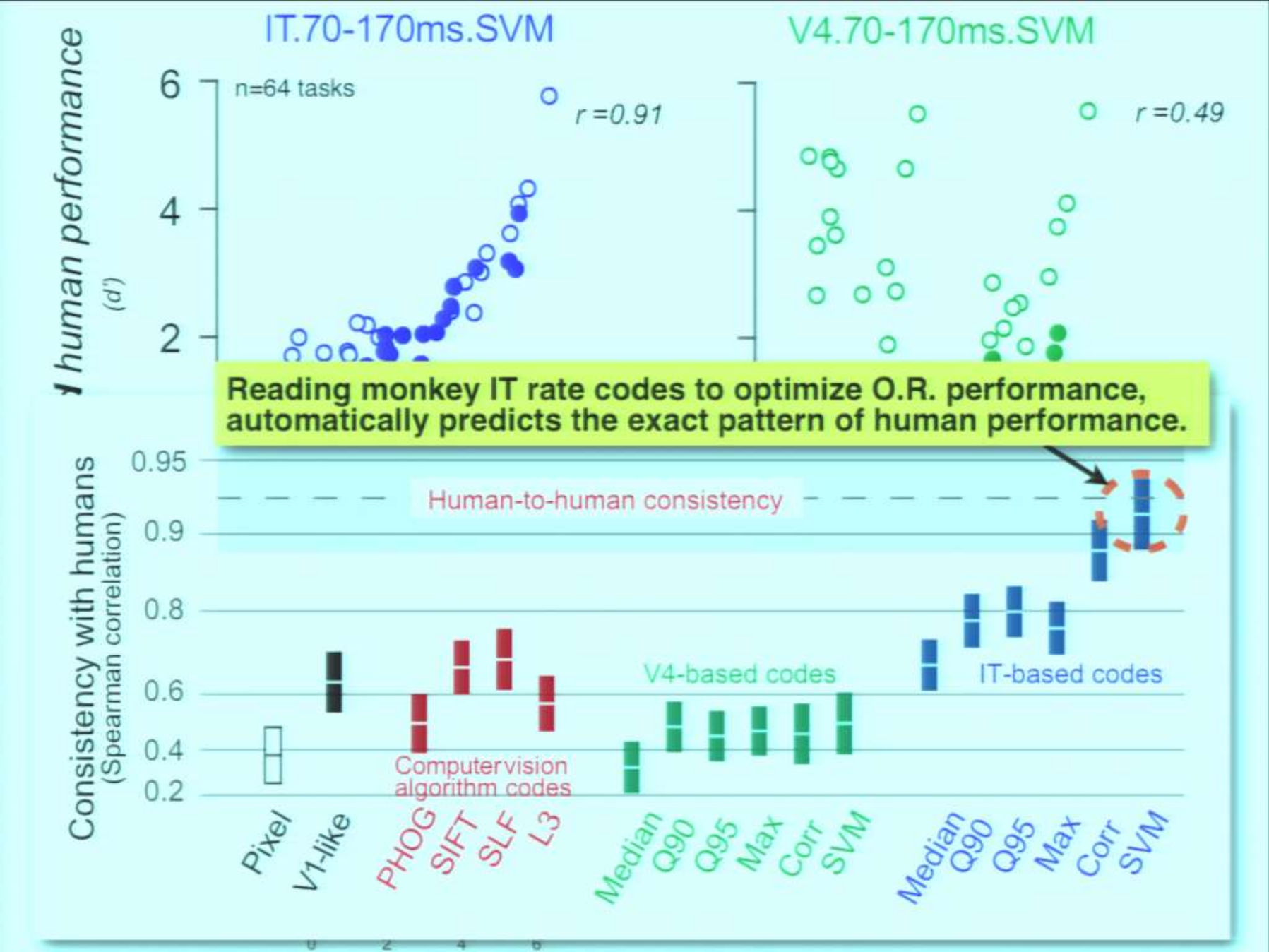

IT.70-170ms.SVM
n=64 tasks
r =0.91
Predicted human performance
(d')
Actual human performance
(mean human d')
0
2
4
6


Correct code
Performance of putative code (d')
Performance of human subjects
(mean human d')

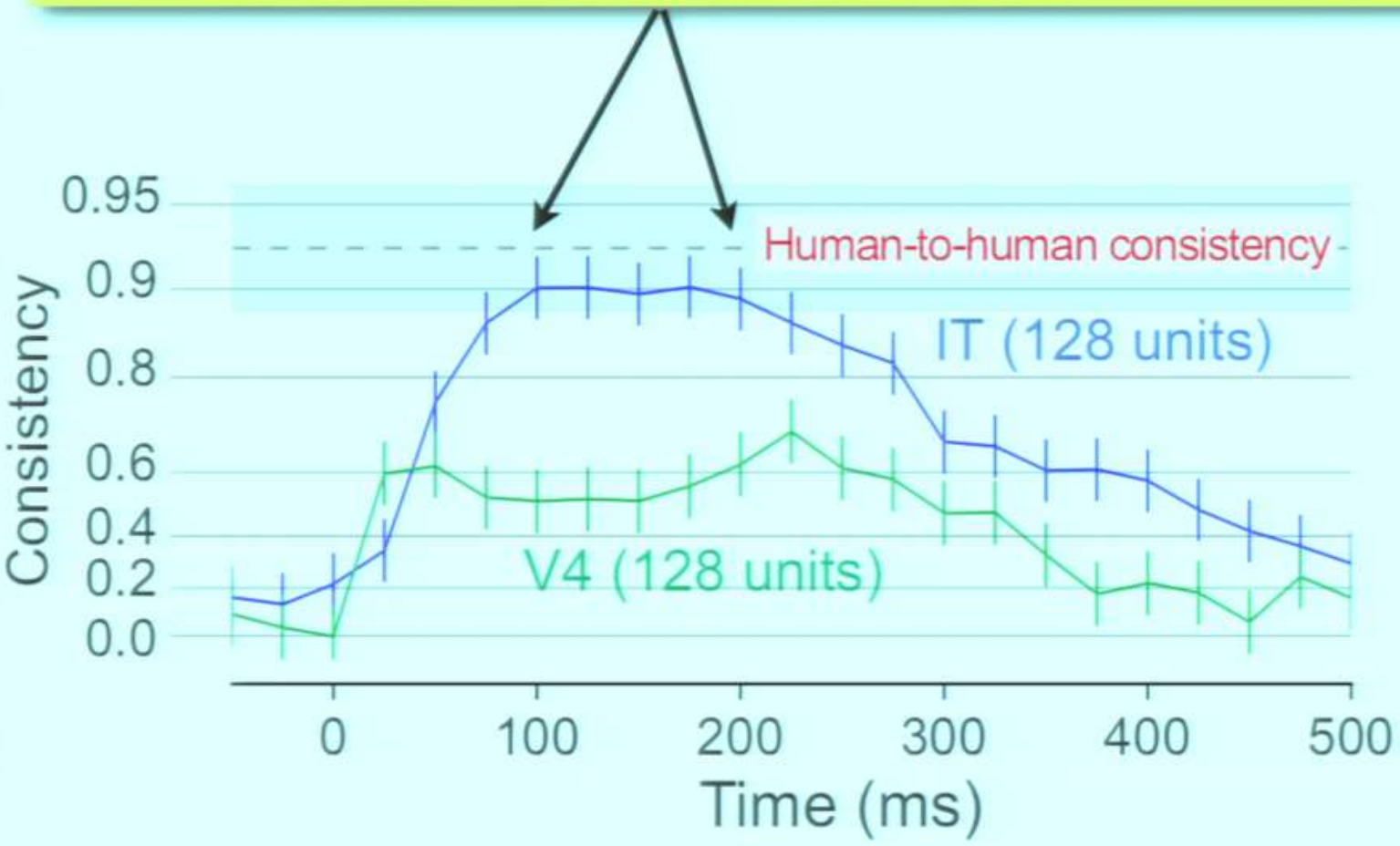

n = 64 tasks
r =0.937
Individual human


IT.70-170ms.SVM
V4.70-170ms.SVM
n=64 tasks
r =0.91
r =0.49
Predicted human performance
(d')
Actual human performance
(mean human d')
0
2
4
6
n = 64 tasks
r =0.937
Individual human

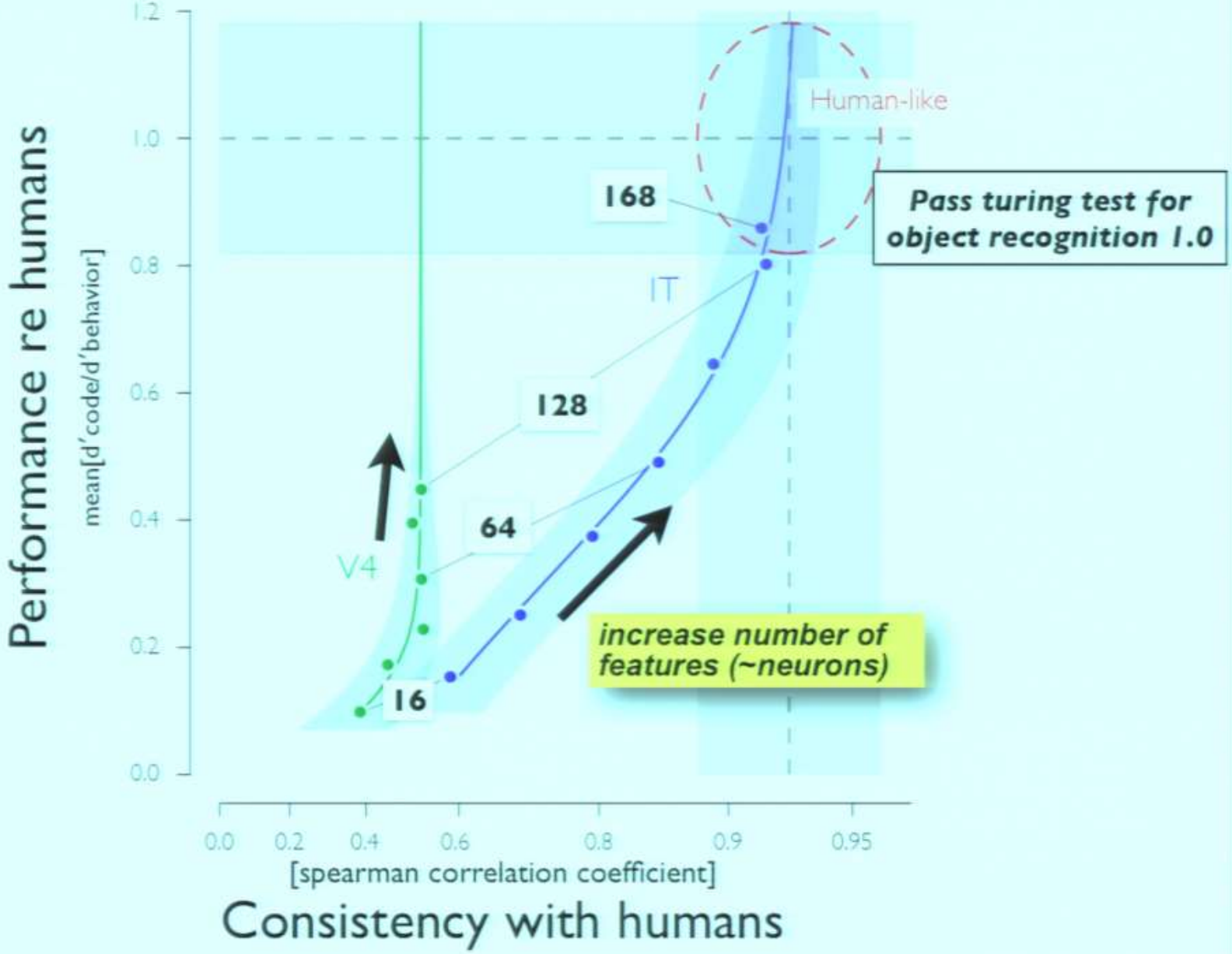

IT.70-170ms.SVM
V4.70-170ms.SVM
n=64 tasks
r =0.91
r =0.49
human performance
(d')
6
4
2
Consistency with humans
(Spearman correlation)
0.95
0.9
0.8
0.6
0.4
0.2
Human-to-human consistency
Computer vision
algorithm codes
V4-based codes
IT-based codes
Pixel
V1-like
PHOG
SIFT
SLF
L3
Median
Q90
Q95
Max
Corr
SVM
Median
Q90
Q95
Max
Corr
SVM


IT.70-170ms.SVM
V4.70-170ms.SVM
human performance
(d')
n=64 tasks
r =0.91
r =0.49
6
4
2
Reading monkey IT rate codes to optimize O.R. performance,
automatically predicts the exact pattern of human performance.
Consistency with humans
(Spearman correlation)
0.95
0.9
0.8
0.6
0.4
0.2
Human-to-human consistency
V4-based codes
IT-based codes
Computer vision
algorithm codes
Pixel
V1-like
PHOG
SIFT
SLF
L3
Median
Q90
Q95
Max
Corr
SVM
Median
Q90
Q95
Max
Corr
SVM

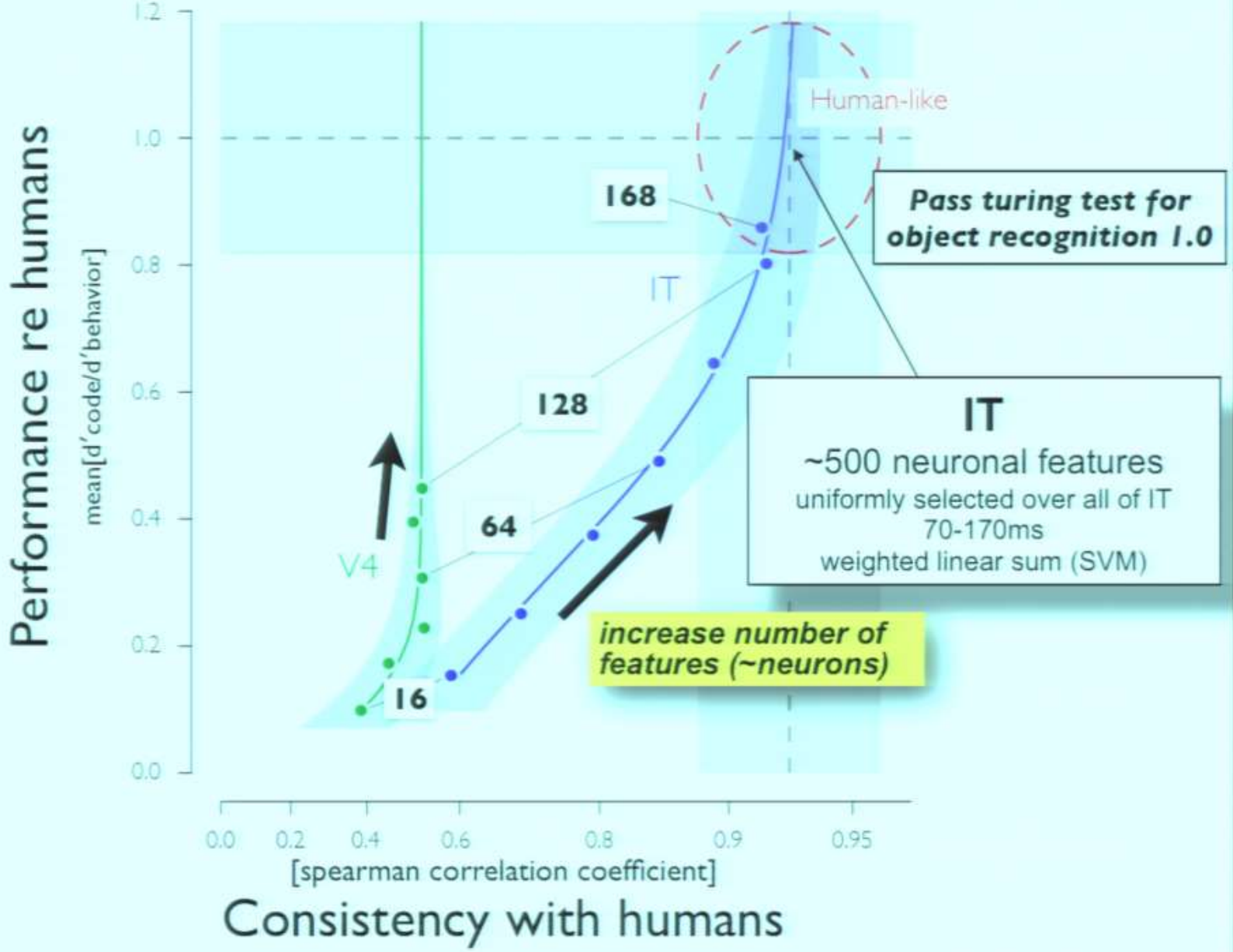
IT population code that predicts behavior is available from 100 to 200 ms after stimulus onset

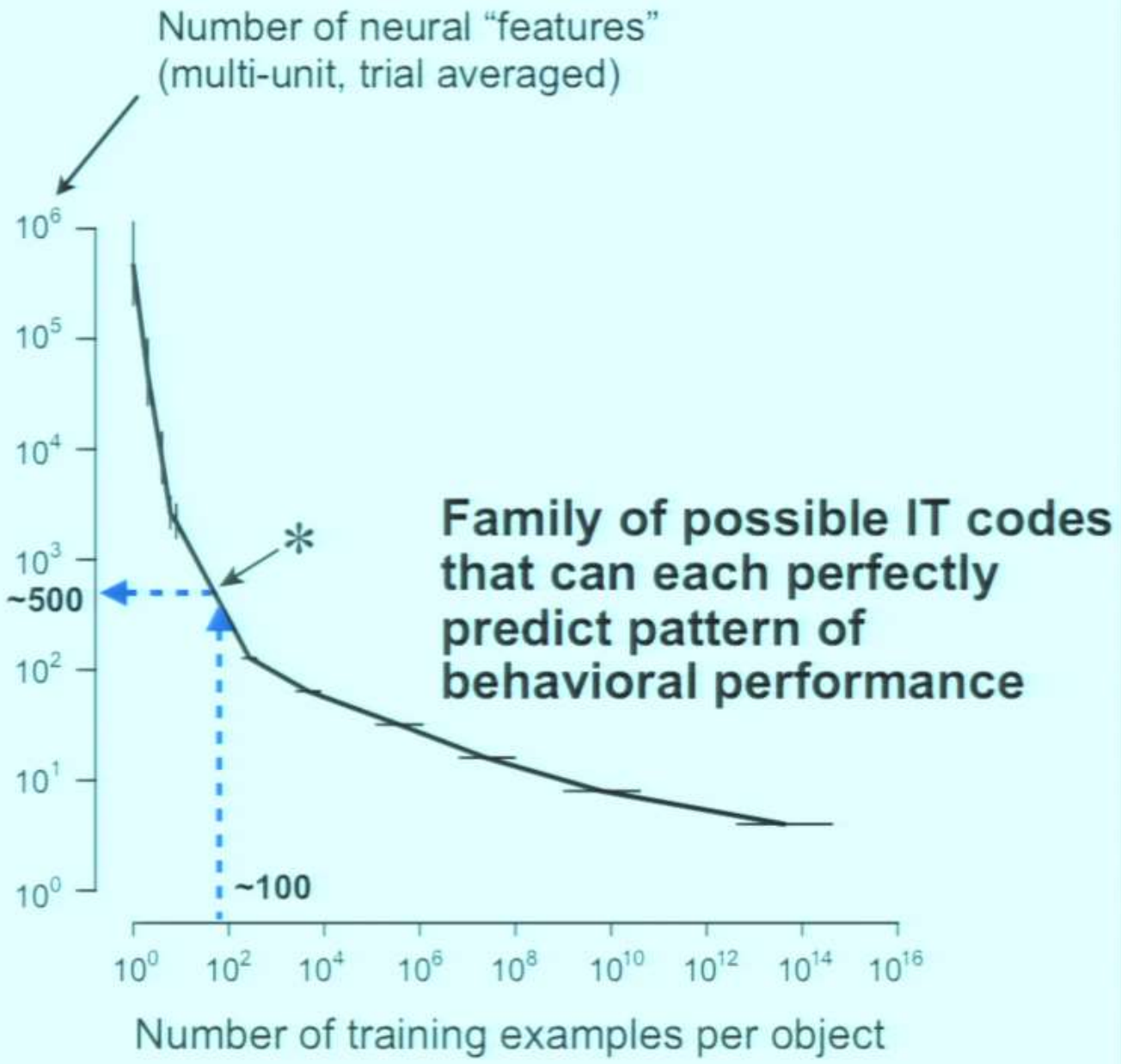


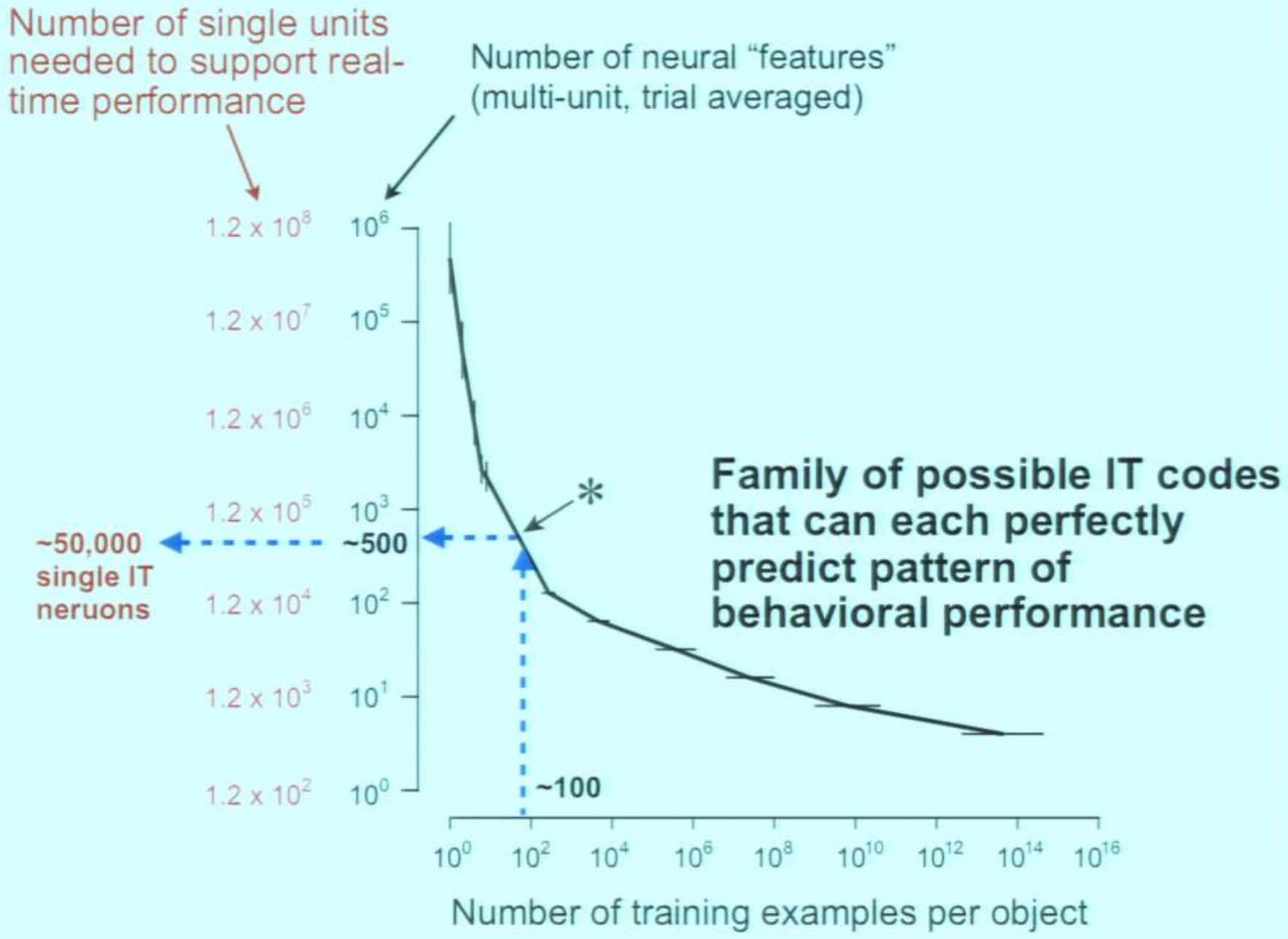
Majaj, Hong, Soloman and DiCarlo, **COSYNE** (2012)
Majaj, Hong, Soloman and DiCarlo (under review)



Majaj, Hong, Soloman and DiCarlo, **COSYNE** (2012)
Majaj, Hong, Soloman and DiCarlo (under review)




Number of neural "features"
(multi-unit, trial averaged)
10^6
10^5
10^4
10^3
~500
10^2
10^1
10^0
*
~100
10^0 10^2 10^4 10^6 10^8 10^{10} 10^{12} 10^{14} 10^{16}
Number of training examples per object
Family of possible IT codes that can each perfectly predict pattern of behavioral performance


Number of single units needed to support real-time performance
Number of neural "features" (multi-unit, trial averaged)
1.2 x 10^8
10^6
1.2 x 10^7
10^5
1.2 x 10^6
10^4
1.2 x 10^5
10^3
~50,000 single IT neruons
~500
1.2 x 10^4
10^2
1.2 x 10^3
10^1
1.2 x 10^2
10^0
*
~100
10^0 10^2 10^4 10^6 10^8 10^10 10^12 10^14 10^16
Number of training examples per object
Family of possible IT codes that can each perfectly predict pattern of behavioral performance

Are any IT neural codes sufficient to explain human object recognition?

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

3. Measure large samples of neuronal population spiking responses

4. Ask: does the proposed link quantitatively predict O.R. behavior ?

Compute predicted O.R. behavior from this neuronal activity ("codes", "decodes")

Are any IT neural codes sufficient to explain human object recognition?

1. Define a set of challenging object recognition (O.R.) tasks

2. Measure human behavioral performance in all of those O.R. tasks

Same images

3. Measure large samples of neuronal population spiking responses

4. Ask: does the proposed link quantitatively predict O.R. behavior ?

Compute predicted O.R. behavior from this neuronal activity ("codes", "decodes")

YES !

What neural codes explain human object recognition?

The simple hypothesis:

- Automatically-evoked spike rate codes distributed over non-human primate IT cortex can explain human object recognition

What neural codes explain human object recognition?

The simple hypothesis:

✓ Automatically-evoked spike rate codes distributed over non-human primate IT cortex can explain human object recognition

Alternative, more complex (more attractive?) hypotheses:

IT does not directly underlie object recognition
(i.e. the key neuronal representations are elsewhere, e.g. V4, LIP, ...)

Rate codes in IT are not sufficient
(e.g. coordinated spike timing patterns are the true object codes)

Passively-evoked spike patterns are not sufficient
(e.g. attentional mechanisms are critical)

Compartments within IT must be carefully considered
(e.g. any tasks related to faces are handles by the "face patch" network)

Monkey neuronal codes cannot explain human perception
(e.g. any tasks related to faces are handles by the "face patch" network)

. . .

What neural codes explain human object recognition?

The simple hypothesis:

- ✓ Automatically-evoked spike rate codes distributed over non-human primate IT cortex can explain human object recognition

Alternative, more complex (more attractive?) hypotheses:

IT does not directly underlie object recognition
(i.e. the key neuronal repre

Rate codes in IT are
(e.g. coordinated spike tim

Passively-evoked spike patterns are not sufficient
(e.g. attentional mechanisms are critical)

Compartments within IT must be carefully considered
(e.g. any tasks related to faces are handles by the "face patch" network)

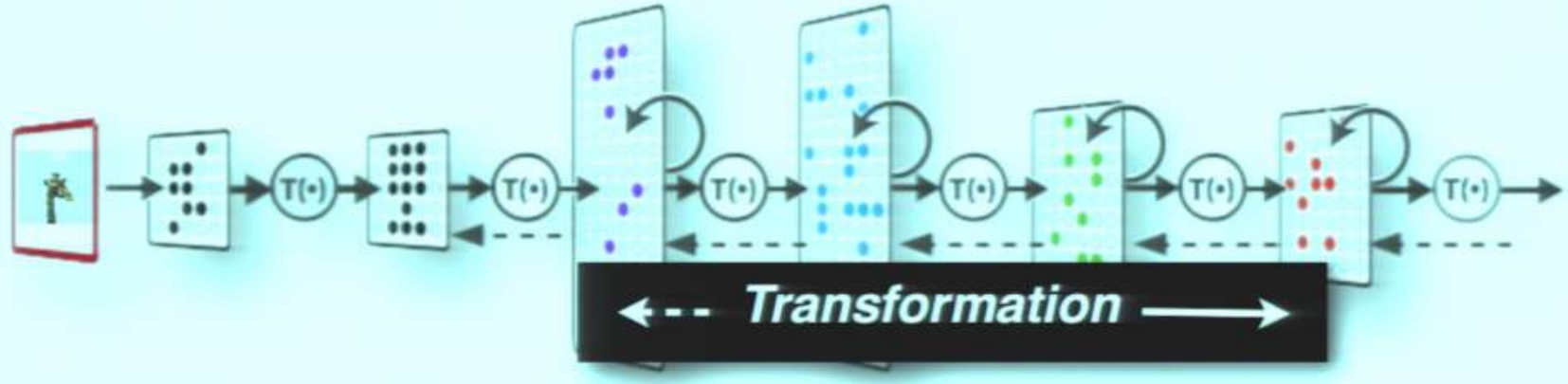
Monkey neuronal codes cannot explain human perception
(e.g. any tasks related to faces are handles by the "face patch" network)

. . .

Parsimony: these more complex alternatives are not (yet) needed to explain object recognition.

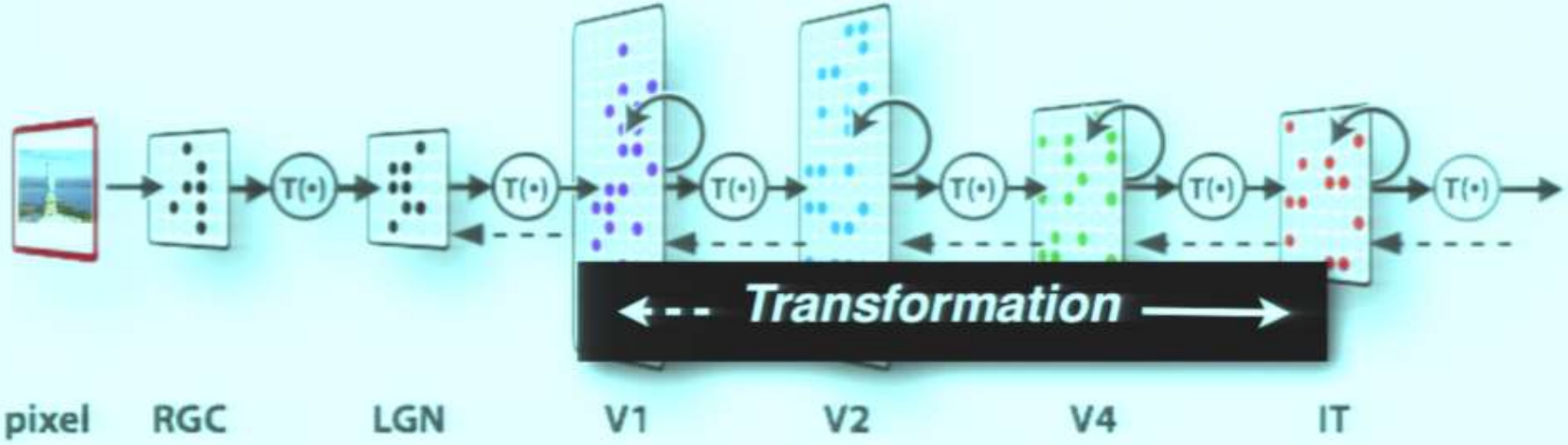
Our primary questions:

Why does the brain need to transform the pixel image ?



Our primary questions:

- Why does the brain need to transform the pixel image ?
- Where is the solution located, and what form does it take?

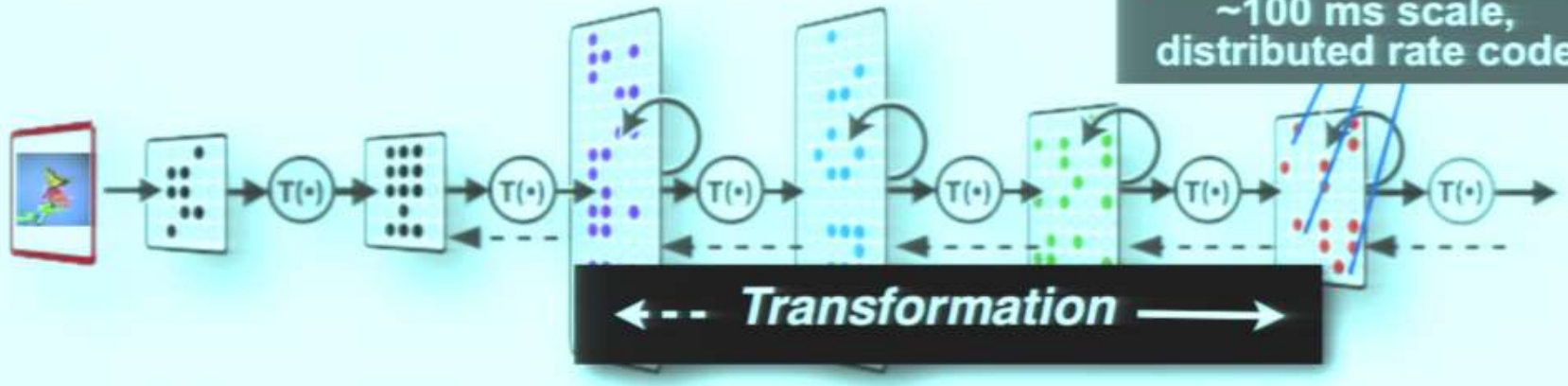


Our primary questions:

- Why does the brain need to transform the pixel image ?
- Where is the solution located, and what form does it take?

Behavior

Automatically-evoked
~100 ms scale,
distributed rate code

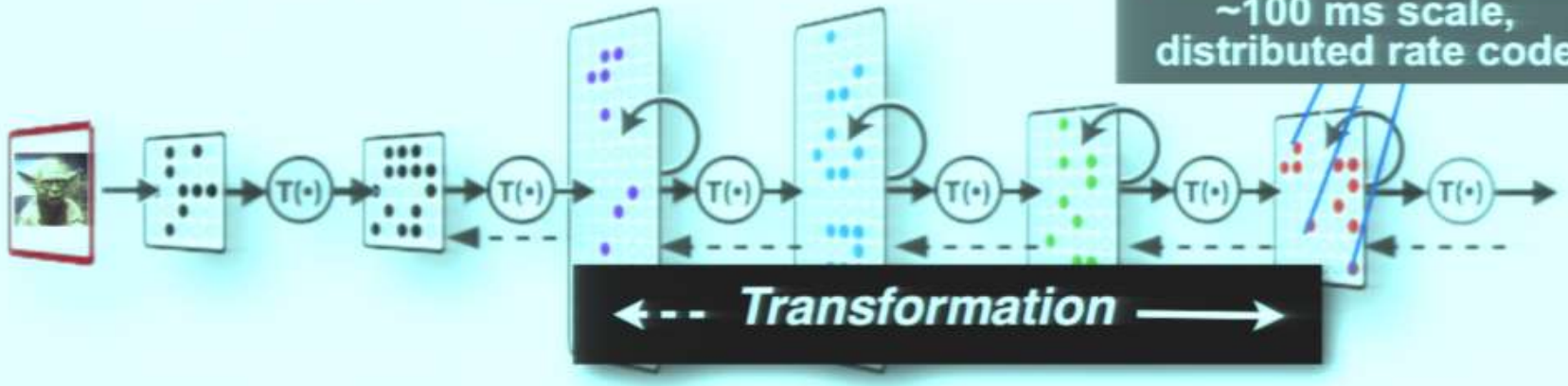


Our primary questions:

- Why does the brain need to transform the pixel image ?
- Where is the solution located, and what form does it take?
  - Sufficient to explain performance

Behavior

Automatically-evoked
~100 ms scale,
distributed rate code



Comparisons I will present today:

1. **Monkey neurons vs. Human Behavior**

   Suggests that IT population codes are one simple step from object recognition behavior

2. **Machines vs. Monkey neurons**

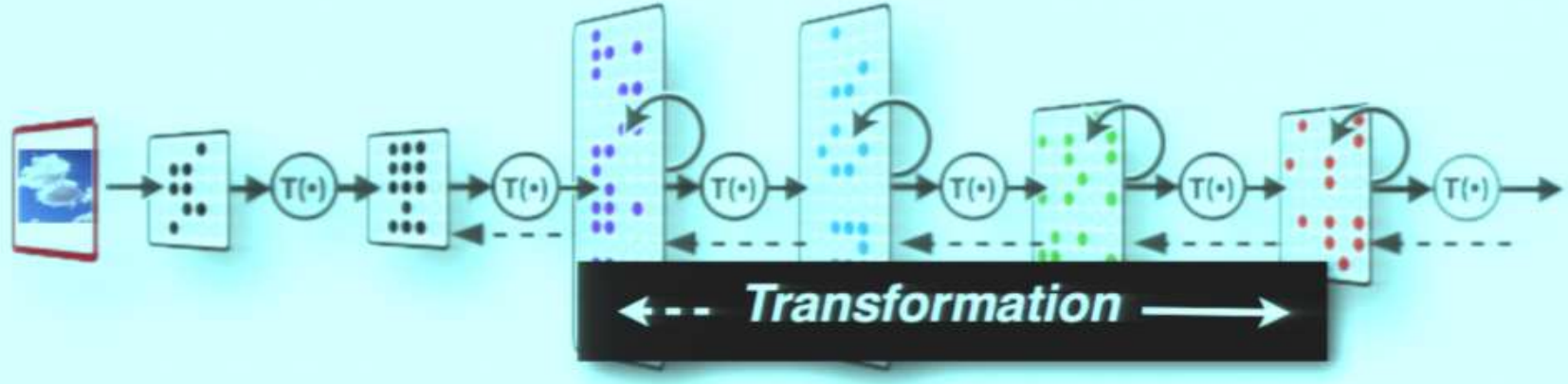
   Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

3. **Machines vs. Monkey neurons/Human behavior**

   Demonstrates the recent bio-inspired models rival the brain in object recognition

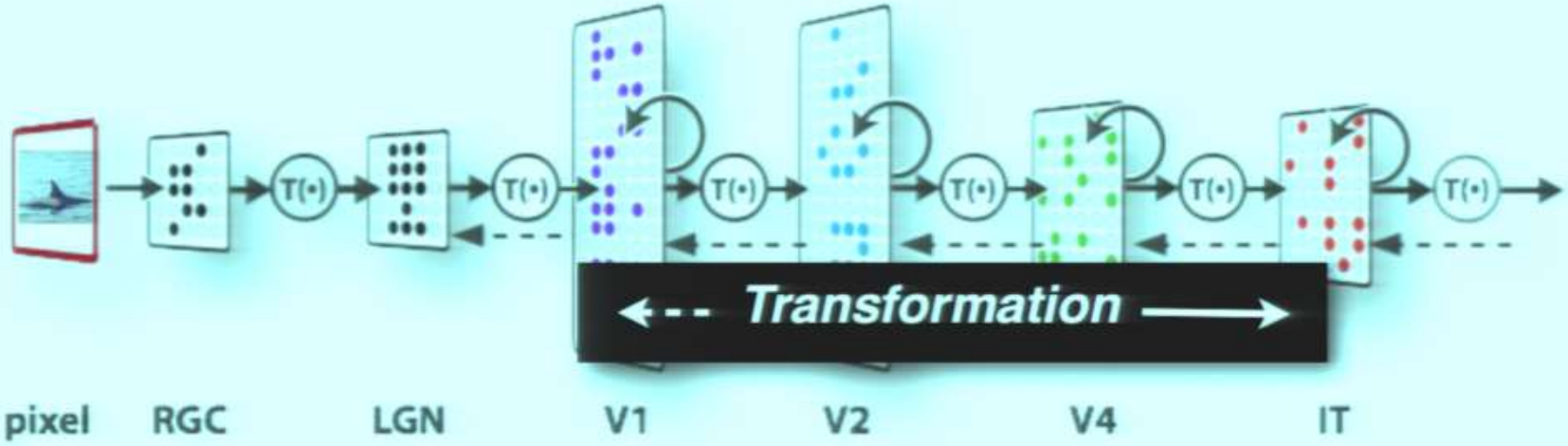
Our primary questions:

✓ Why does the brain need to transform the pixel image ?



Our primary questions:

- Why does the brain need to transform the pixel image ?
- Where is the solution located, and what form does it take?

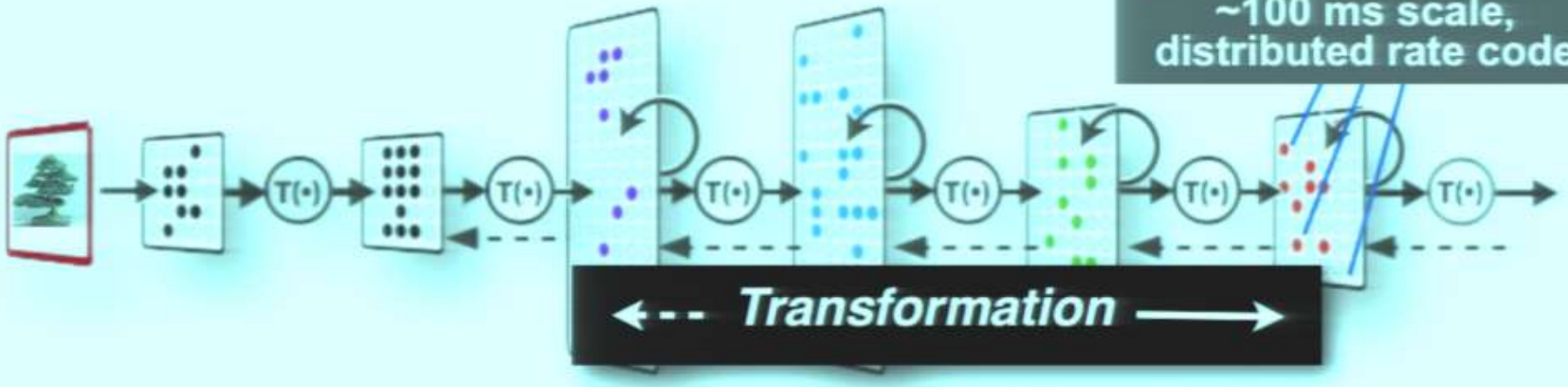


Our primary questions:

- Why does the brain need to transform the pixel image ?
- Where is the solution located, and what form does it take?
  - Sufficient to explain performance

Behavior

Automatically-evoked
~100 ms scale,
distributed rate code

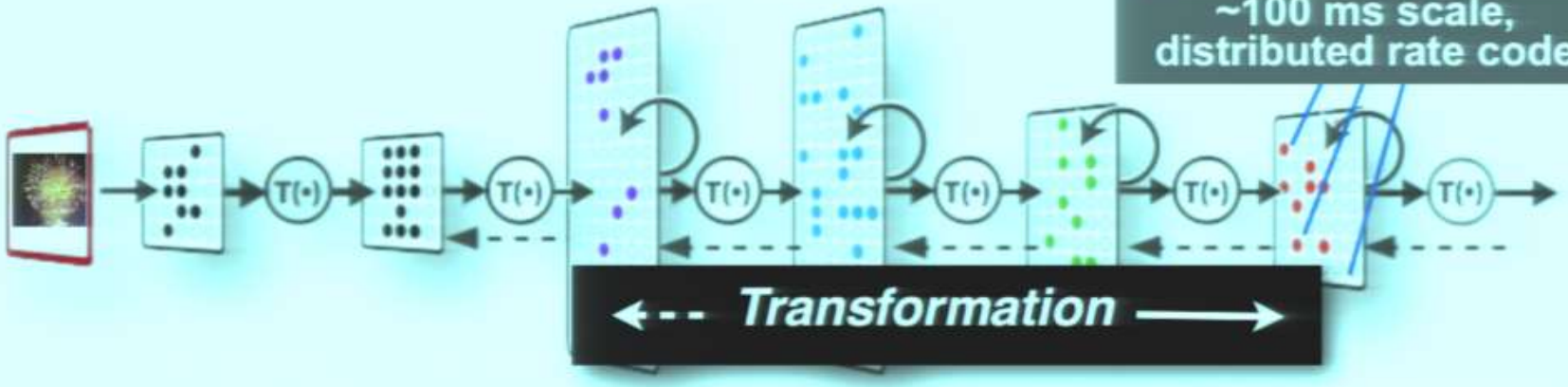


Our primary questions:

- ✓ Why does the brain need to transform the pixel image ?
- ✓ Where is the solution located, and what form does it take?
  - ✓ Sufficient to explain performance
  - ✓ Predicts pattern of human behavior.

Behavior

Automatically-evoked
~100 ms scale,
distributed rate code



Comparisons I will present today:

1. **Monkey neurons vs. Human Behavior**

   Suggests that IT population codes are one simple step from object recognition behavior

2. **Machines vs. Monkey neurons**

   Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

3. **Machines vs. Monkey neurons/Human behavior**

   Demonstrates the recent bio-inspired models rival the brain in object recognition

Comparisons I will present today:

1. **Monkey neurons vs. Human Behavior**

   Suggests that IT population codes are one simple step from object recognition behavior

2. **Machines vs. Monkey neurons**

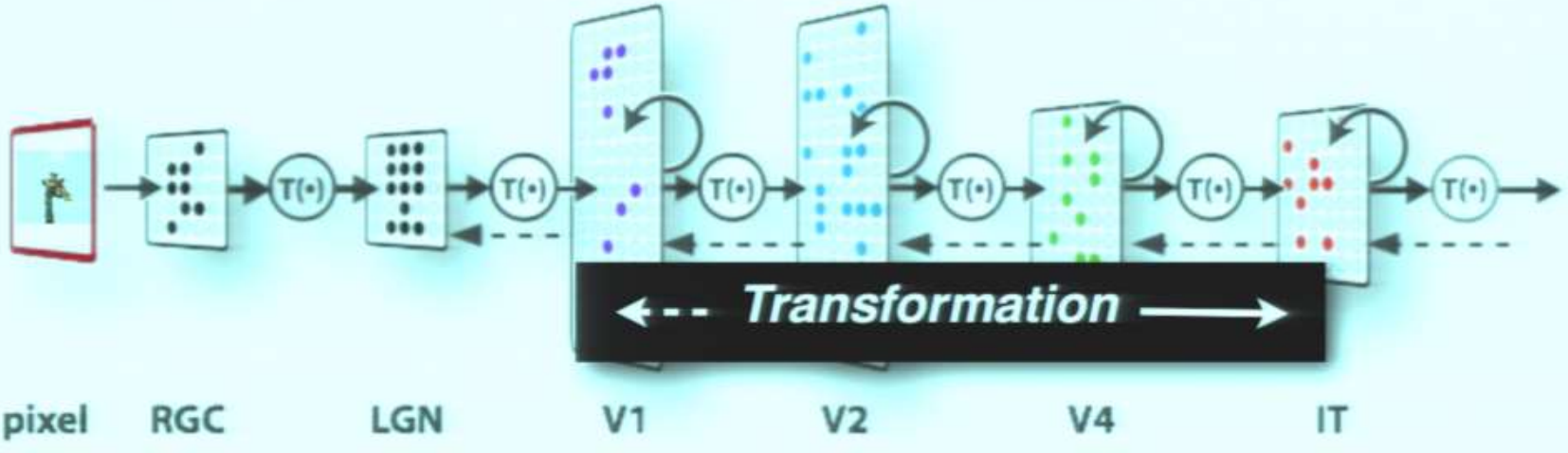
   Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

3. **Machines vs. Monkey neurons/Human behavior**

   Demonstrates the recent bio-inspired models rival the brain in object recognition

Our primary questions:

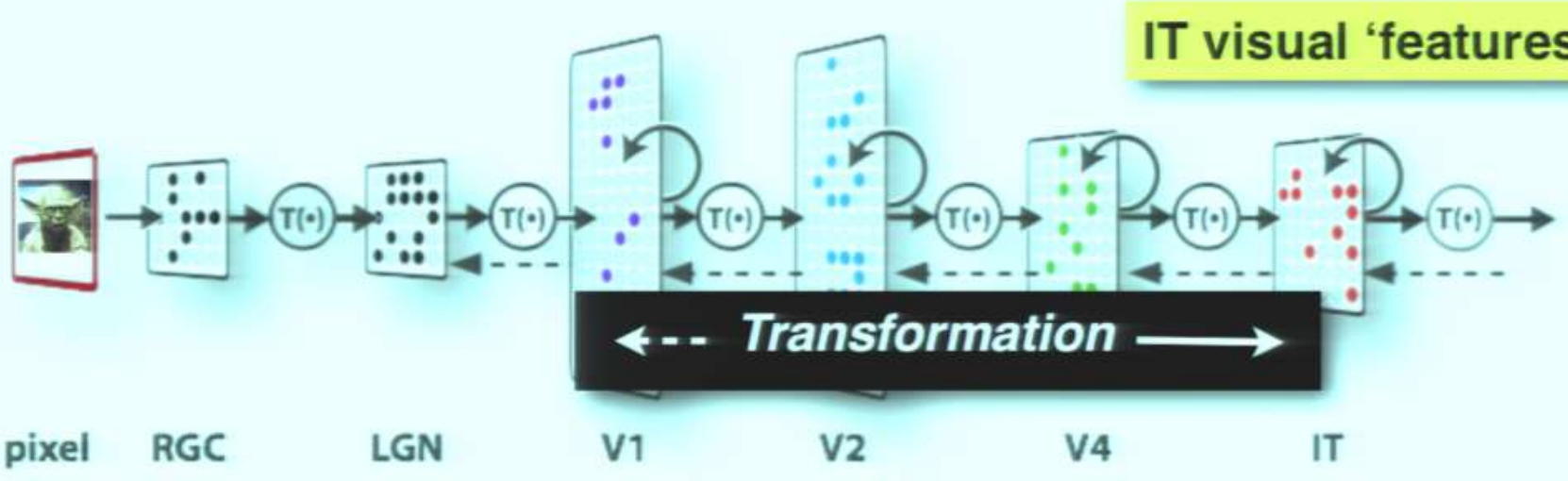
<u>How</u> do the circuits of the ventral stream transform the pixel image to produce the IT representation ?



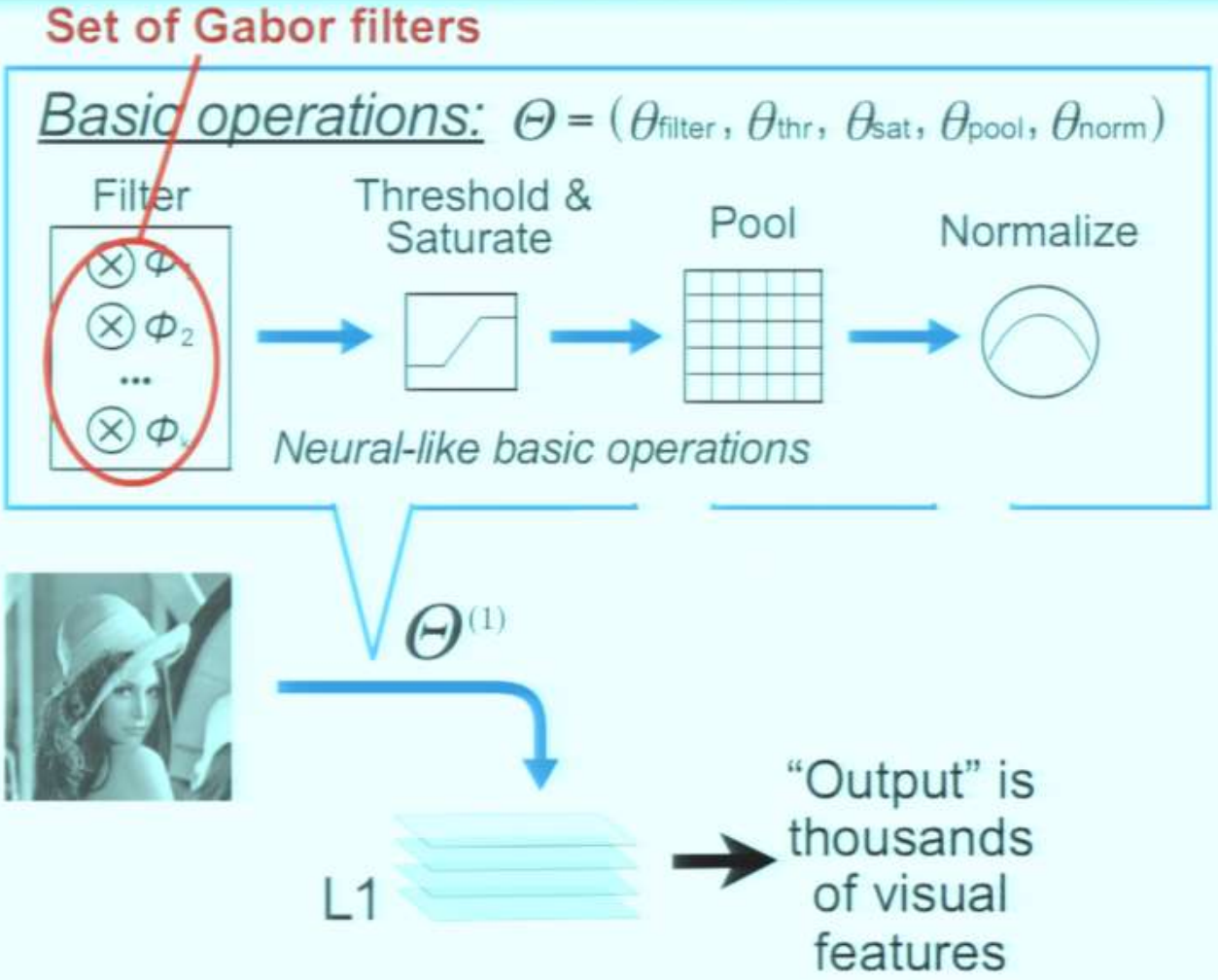
Our primary questions:

How do the circuits of the ventral stream transform the pixel image to produce the IT representation ?

This is where neuroscience meets computer vision, so let's start with those models.

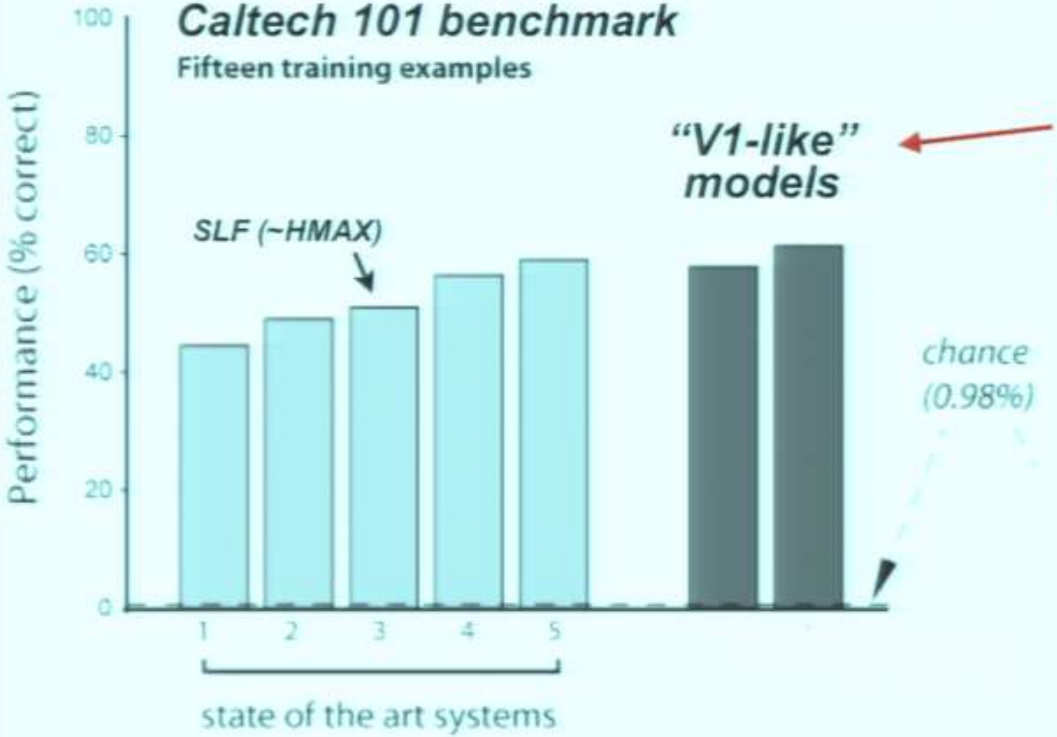


Basic bio-inspired model layer



Pinto, Cox & DiCarlo, ***PLoS Comp Biol*** *(2008)*

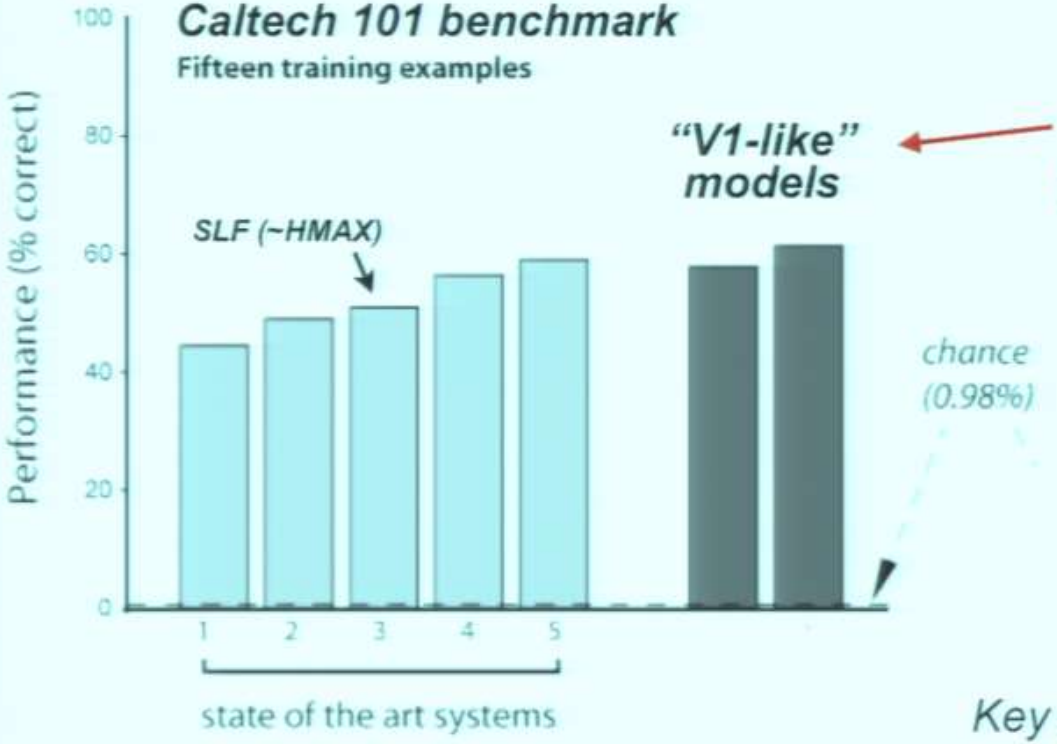
~2008: Tests of performance were not stringent enough.



Neuroscientist null model outperforming computer vision!

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

~2008: Tests of performance were not stringent enough.



Neuroscientist null model outperforming computer vision!

Key problem was insufficient variation in the test sets.

car car

Pinto, Cox. and DiCarlo. **PLoS Comp Bio** (2008)

2009: We made more stringent, but compact benchmarks

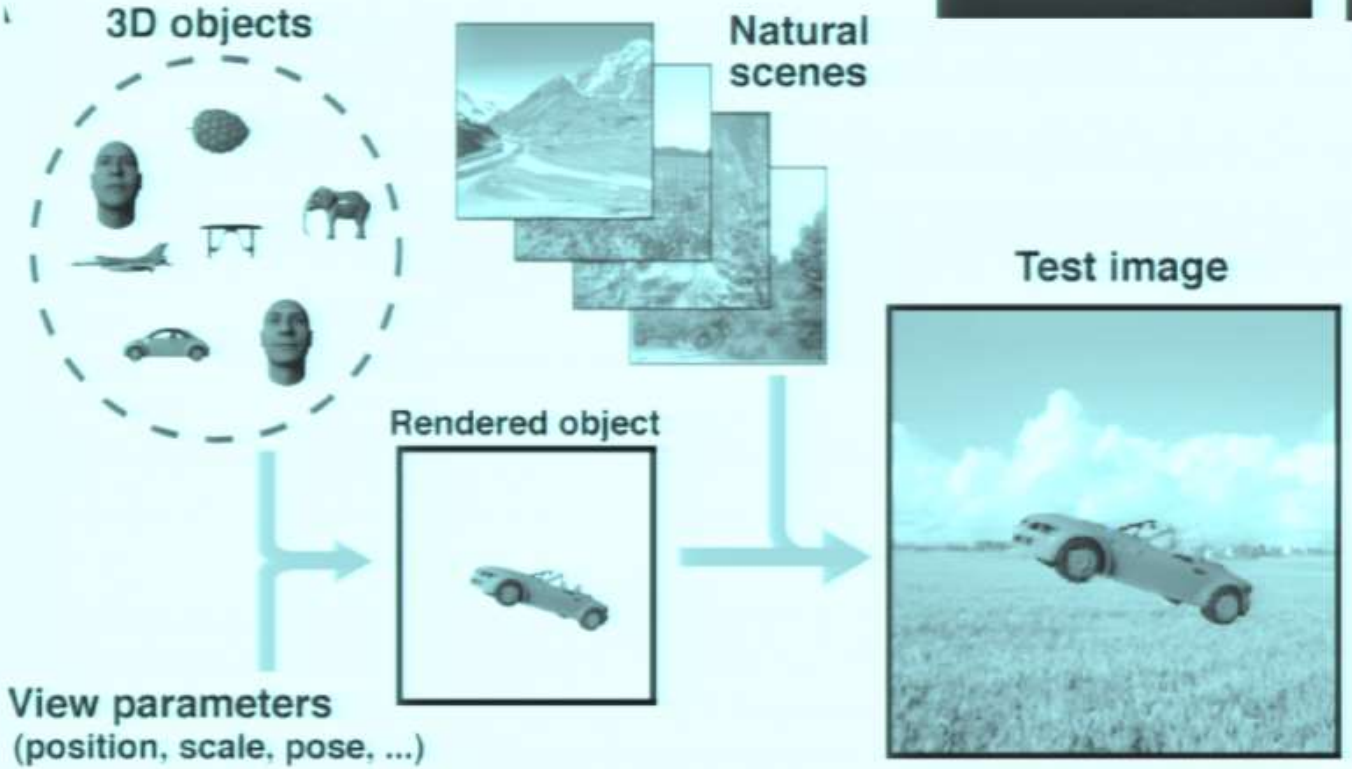
Example object recognition task: "car detection"

Pinto, Cox & DiCarlo, PLoS Comp Bol (2008). Pinto, DiCarlo and Cox, ECCV (2008); Pinto, Doukan, DiCarlo & Cox, PLoS Comp Biol (2009)

2009: We made more stringent, but compact benchmarks

Example object recognition task: "car detection"

Image generation strategy:



Pinto, Cox & DiCarlo, PLoS Comp Bol (2008). Pinto, DiCarlo and Cox, ECCV (2008); Pinto, Doukan, DiCarlo & Cox, PLoS Comp Biol (2009)

2009: We made more stringent, but compact benchmarks

Example object recognition task: "car detection"

Image generation strategy:

- Parametric control of task demand (esp. invariance)

no variation → more variation → lots of variation

Pinto, Cox & DiCarlo, PLoS Comp Bol (2008). Pinto, DiCarlo and Cox, ECCV (2008); Pinto, Doukan, DiCarlo & Cox, PLoS Comp Biol (2009)

2009: We made more stringent, but compact benchmarks

Basic car task,
variation level: 3

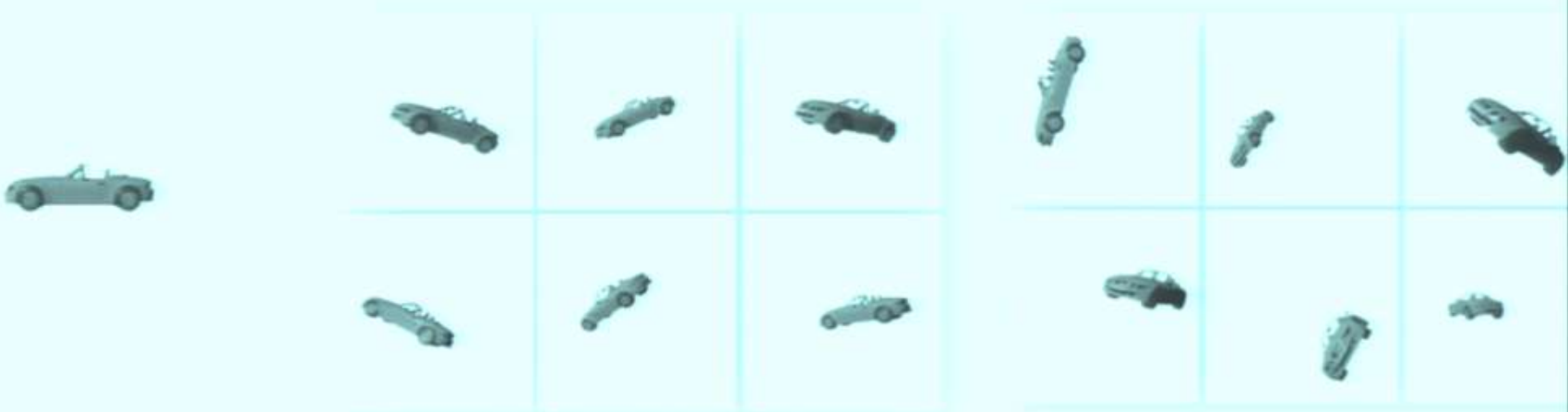
Pinto, Cox & DiCarlo, PLoS Comp Bol (2008). Pinto, DiCarlo and Cox, ECCV (2008); Pinto, Doukan, DiCarlo & Cox, PLoS Comp Biol (2009)

2009: We made more stringent, but compact benchmarks

Example object recognition task: "car detection"

Image generation strategy:

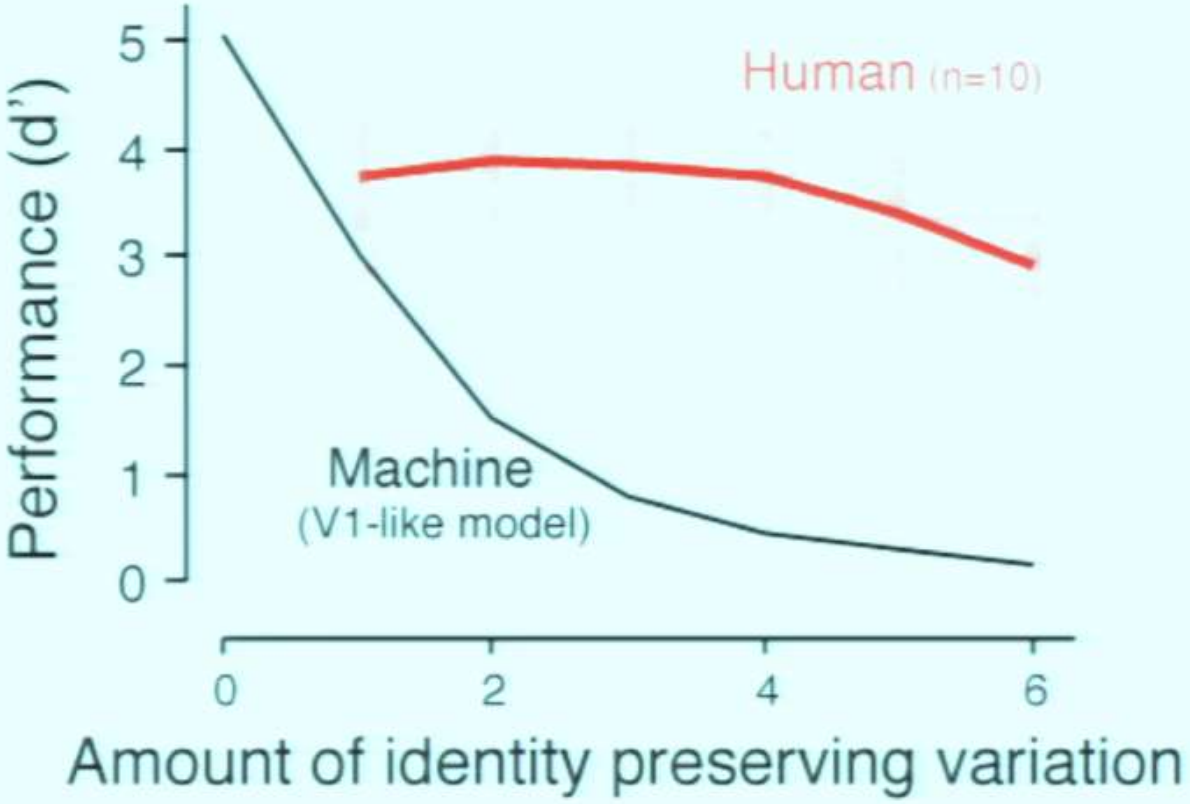
- Parametric control of task demand (esp. invariance)



no variation → more variation → lots of variation

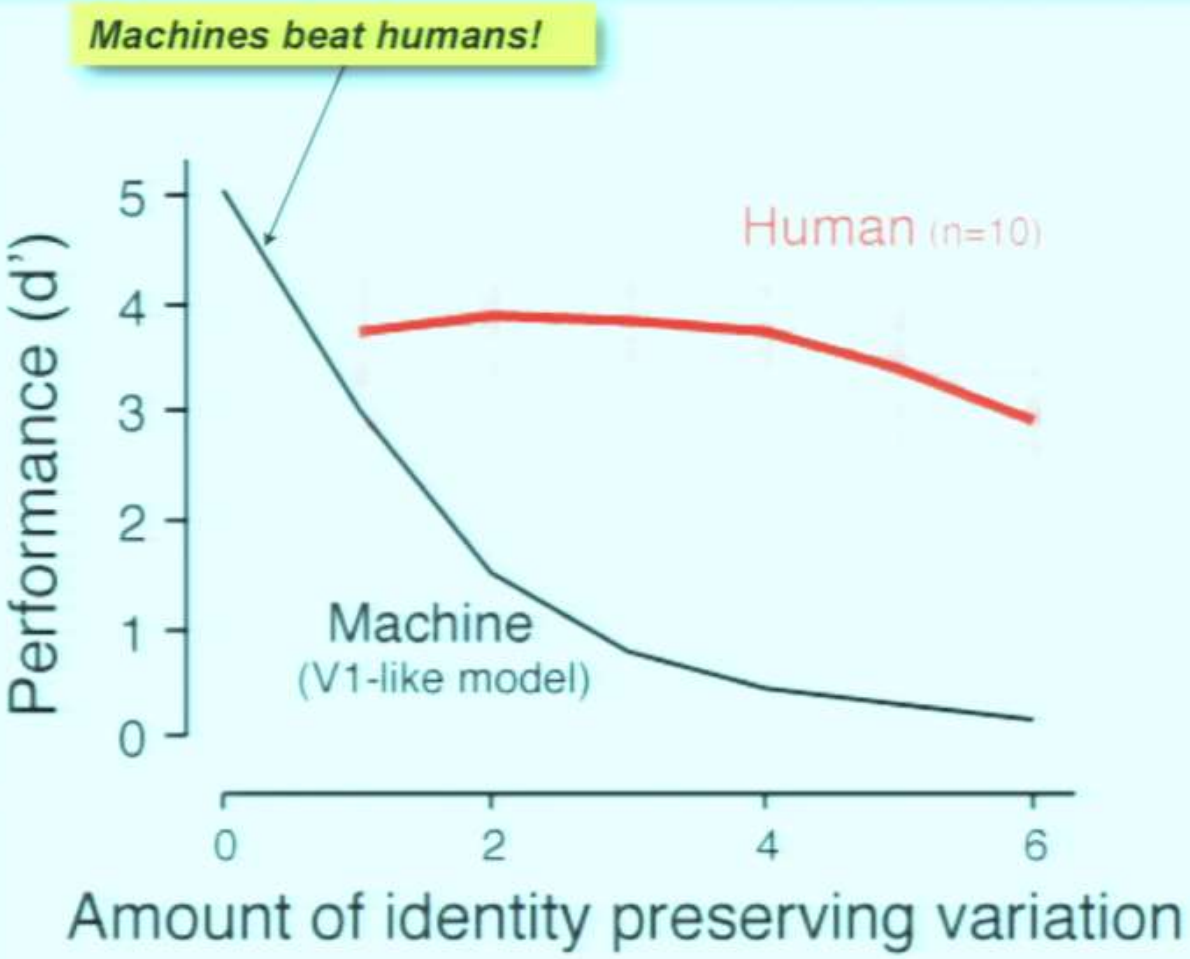
Pinto, Cox & DiCarlo, PLoS Comp Bol (2008). Pinto, DiCarlo and Cox, ECCV (2008); Pinto, Doukan, DiCarlo & Cox, PLoS Comp Biol (2009)

2010: Machines vs. human brains on these benchmarks



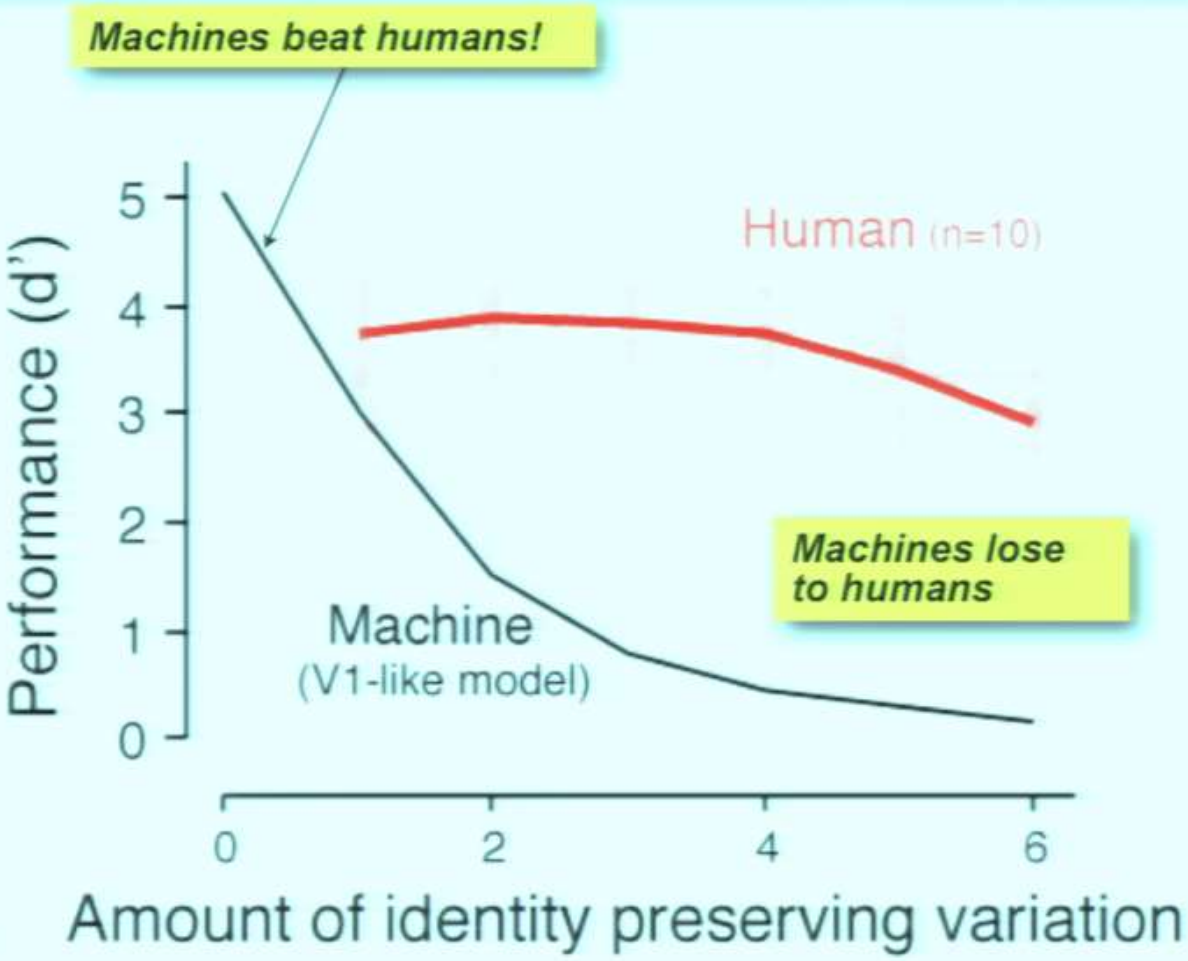
Data merged here: 48 basic-level tasks (8 labels x 6 level of variation)

2010: Machines vs. human brains on these benchmarks

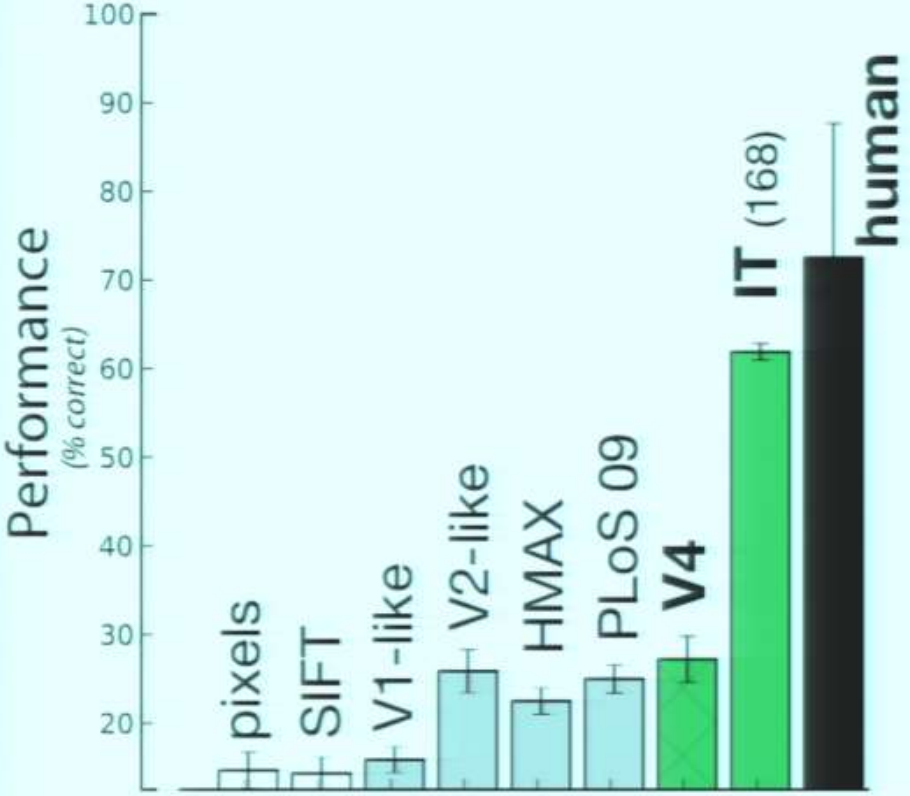


Data merged here: 48 basic-level tasks (8 labels x 6 level of variation)

2010: Machines vs. human brains on these benchmarks



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



Object recognition 1.0
(HVM 1.0 test set)

~2009: Computer vision systems were not viable hypotheses of the ventral stream

And did not need 1,000,000 images to tell us this

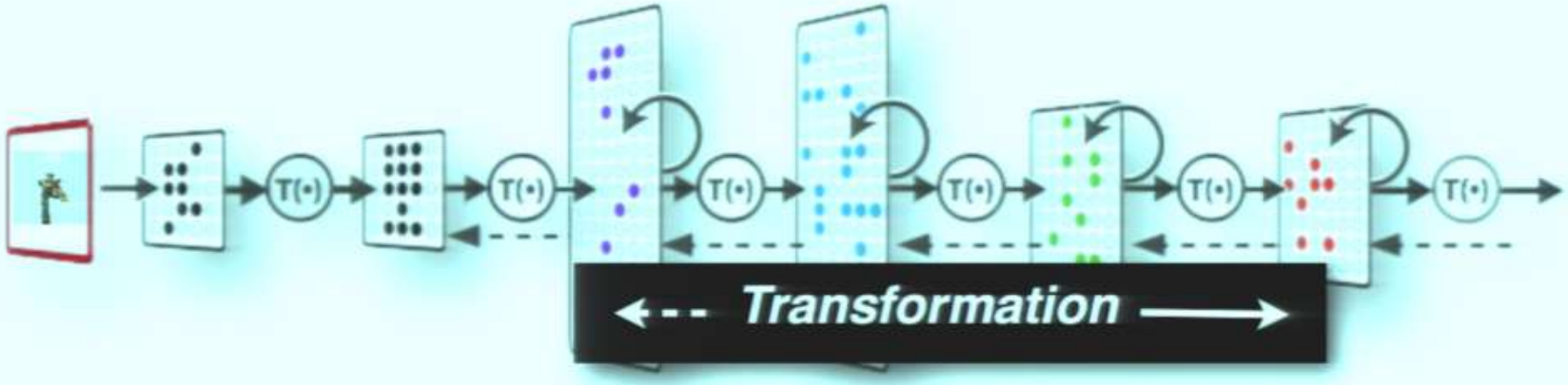
What was the problem?

Possibility A. This largely feedforward hypothesis is deeply lacking.

Theoretical models (e.g. HMAX) don't perform well or capture observed neural responses.
Pinto et. al. (2008), Kriegeskorte (2009)

Possibility B. We just don't know how to find the parameters.

Direct fits to V4 and IT neural data either explain low amounts of variance (<20%) or aren't image-driven. Gallant (2007), Pasupathy & Connor (2004), Brincat & Connor (2008)



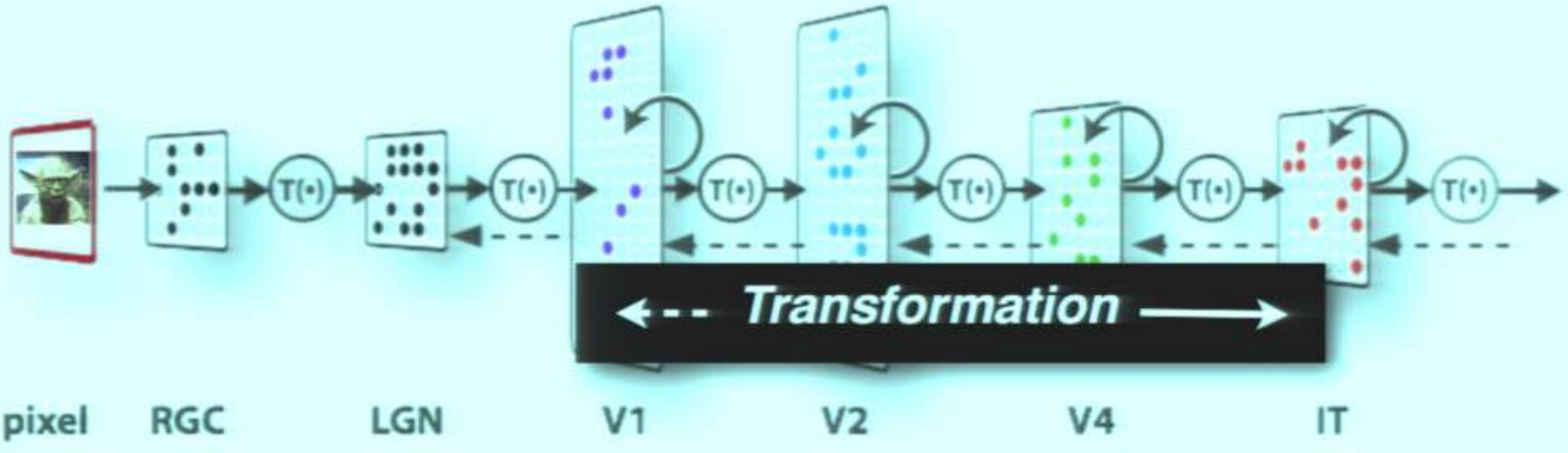
What was the problem?

Possibility A. This largely feedforward hypothesis is deeply lacking.

Theoretical models (e.g. HMAX) don't perform well or capture observed neural responses.
Pinto et. al. (2008). Kriegeskorte (2009)

Possibility B. We just don't know how to find the parameters.

Direct fits to V4 and IT neural data either explain low amounts of variance (<20%) or aren't image-driven. Gallant (2007). Pasupathy & Connor (2004). Brincat & Connor (2008)



What was the problem?

Possibility A. This largely feedforward hypothesis is deeply lacking.

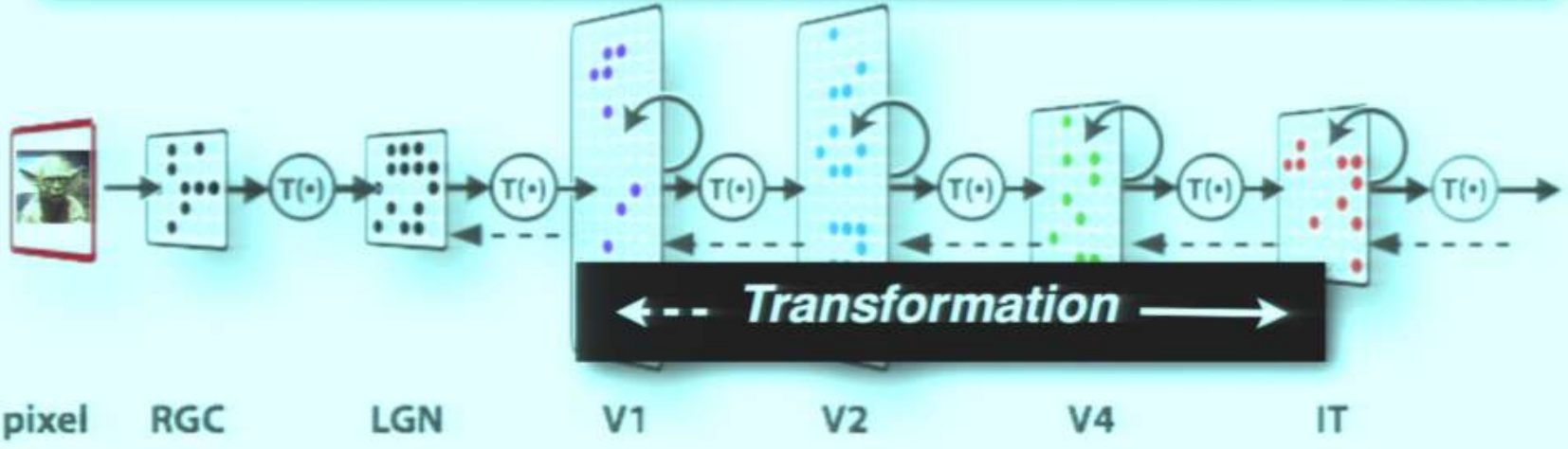
Theoretical models (e.g. HMAX) don't perform well or capture observed neural responses.
Pinto et. al. (2008). Kriegeskorte (2009)

Possibility B. We just don't know how to find the parameters.

Direct fits to V4 and IT neural data either explain low amounts of variance (<20%) or aren't image-driven. Gallant (2007). Pasupathy & Connor (2004). Brincat & Connor (2008)

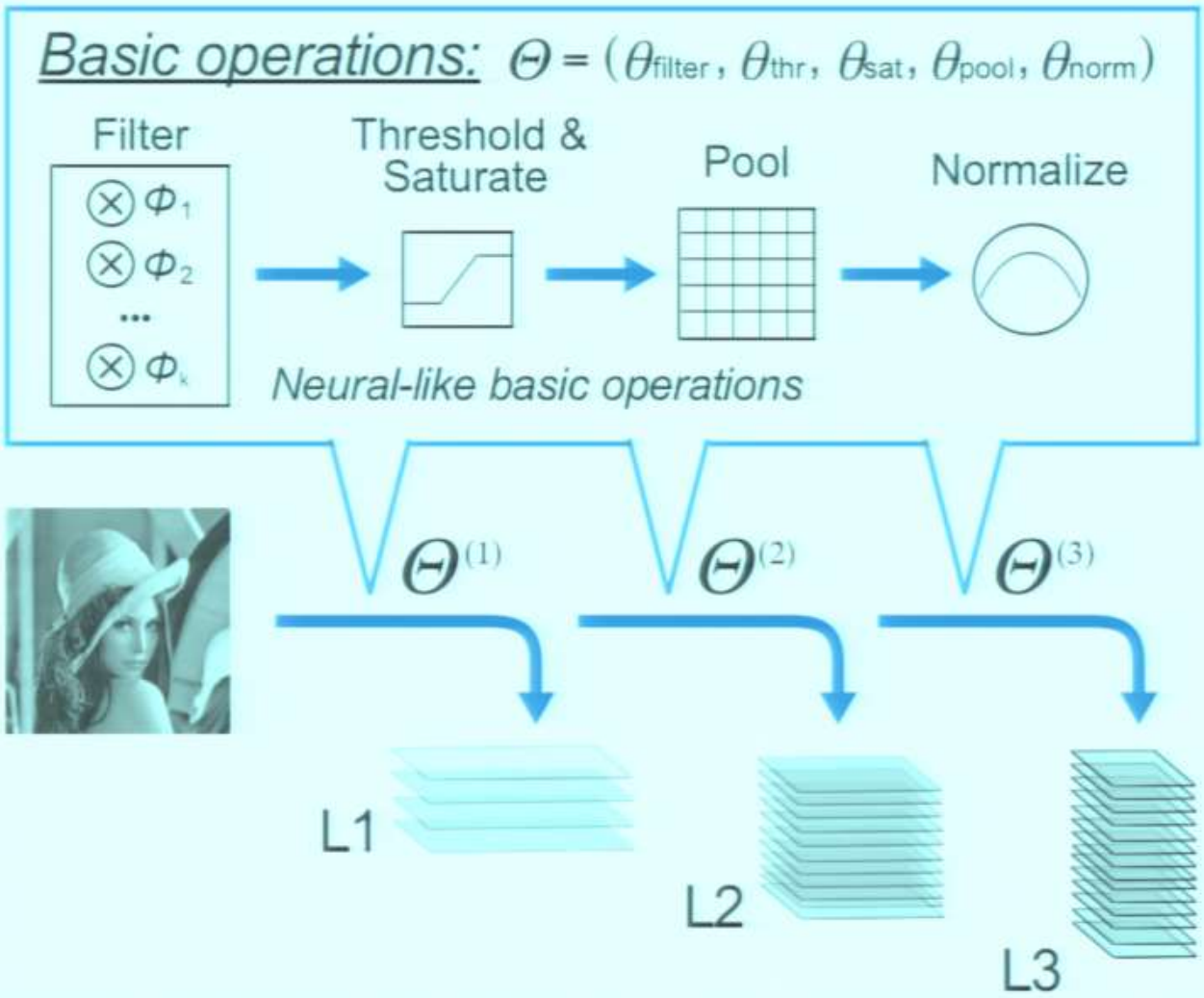
Our approach: work on **B** before introducing the complexity of **A**.

0. Start with largely feedforward, bio-inspired model class
1. Optimize performance on tasks the brain is (re.) good at
2. Ask: do model features looks like the brains features?



Basic bio-inspired (deep) model

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



Deep hierarchy

Convolutional

LNN

Limited feedback

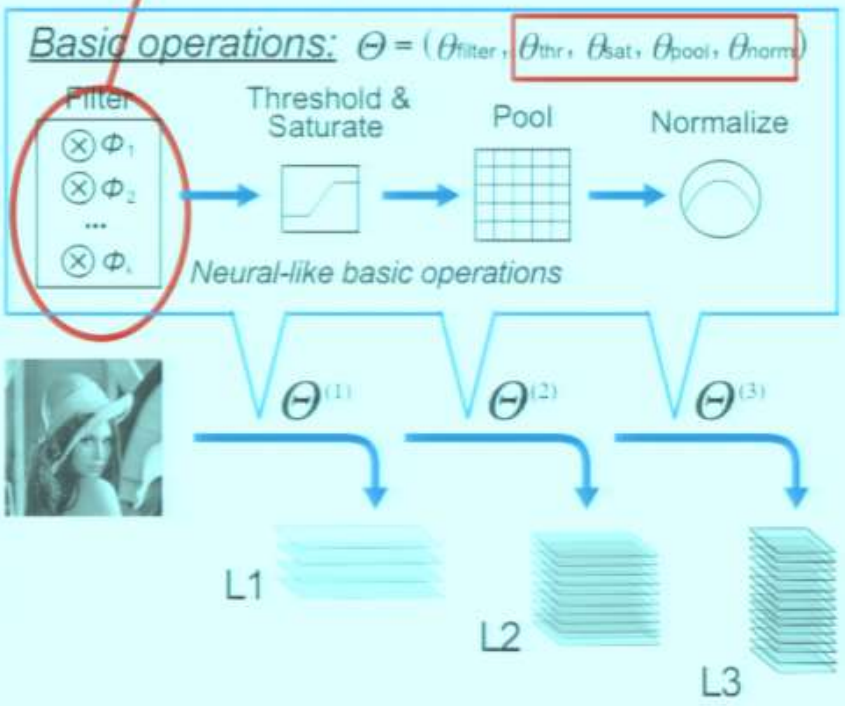
"Output" is thousands of visual features

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

Basic bio-inspired (deep) model

Random filter params

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



We saw large performance gains by optimizing* the architectural parameters (a.k.a. hyperparameters)

David Cox

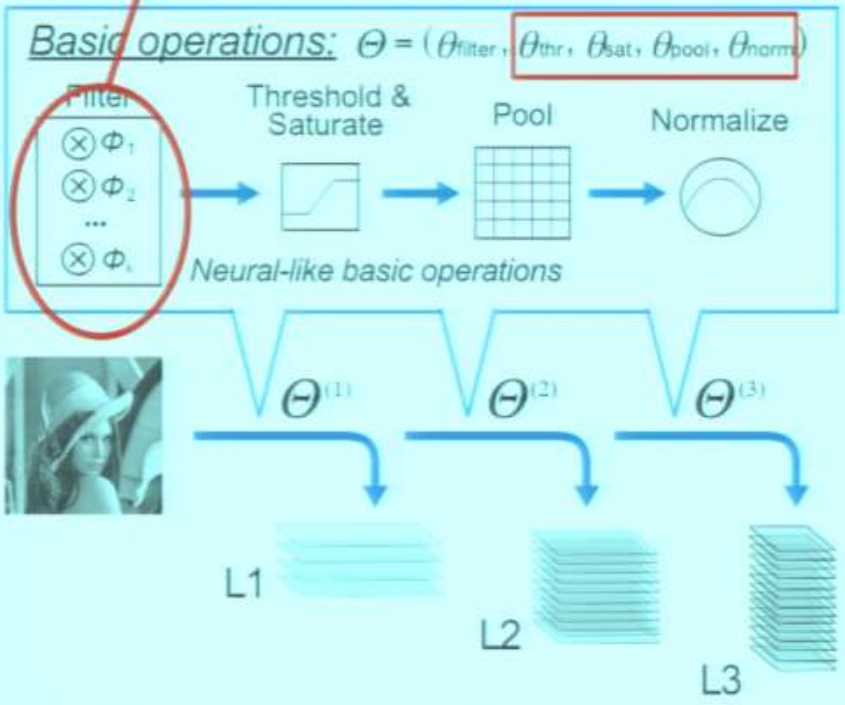
Nicolas Pinto

Nicolas Pinto, David Doukhan, James J. DiCarlo, **David D. Cox** (2009)
A High-Throughput Screening Approach to Discovering Good Forms of Biologically Inspired Visual Representation
PLoS Computational Biology 5 (11)

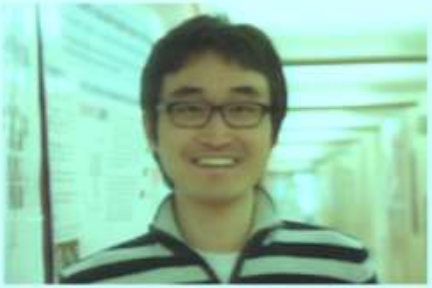
Nicolas Pinto, James J. DiCarlo, **David D. Cox** (2009)
How far can you get with a modern face recognition test set using only simple features?
IEEE Computer Vision and Pattern Recognition

Basic bio-inspired (deep) model

Random filter params

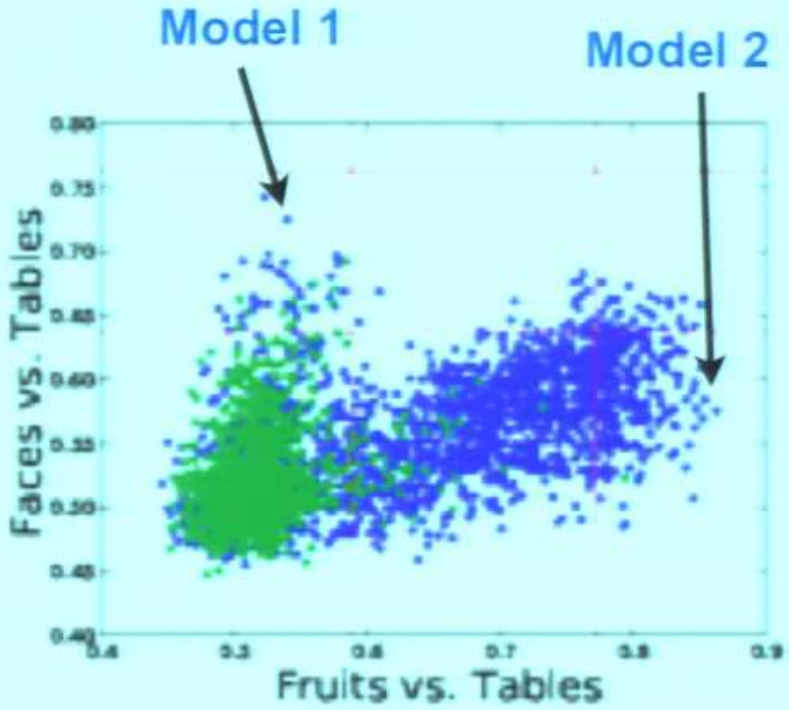


Dan Yamins



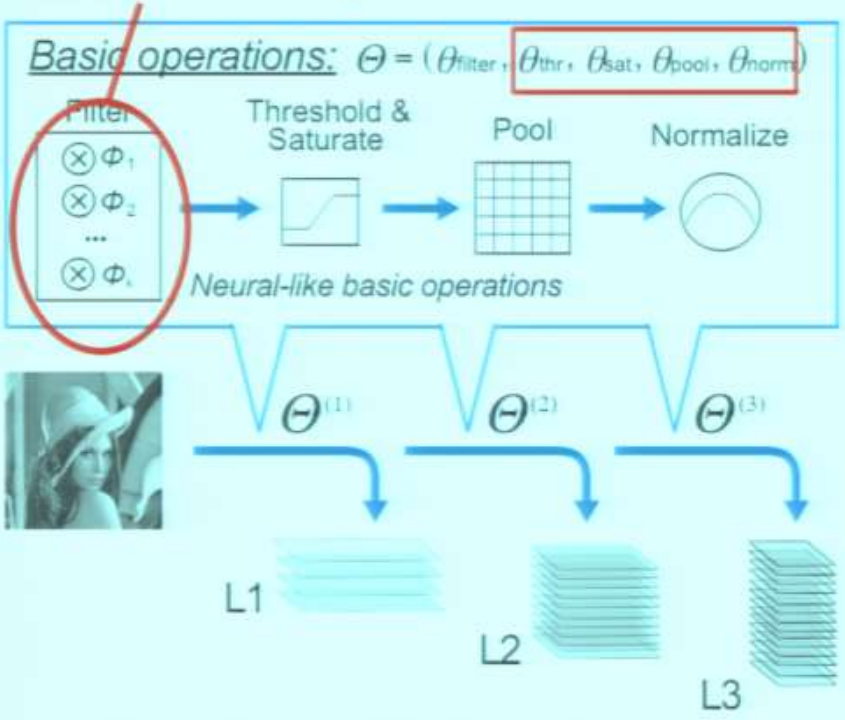
Ha Hong

Noticed that different types of object recognition tasks were best solved by different choices of architectural parameters



Basic bio-inspired (deep) model

Random filter params

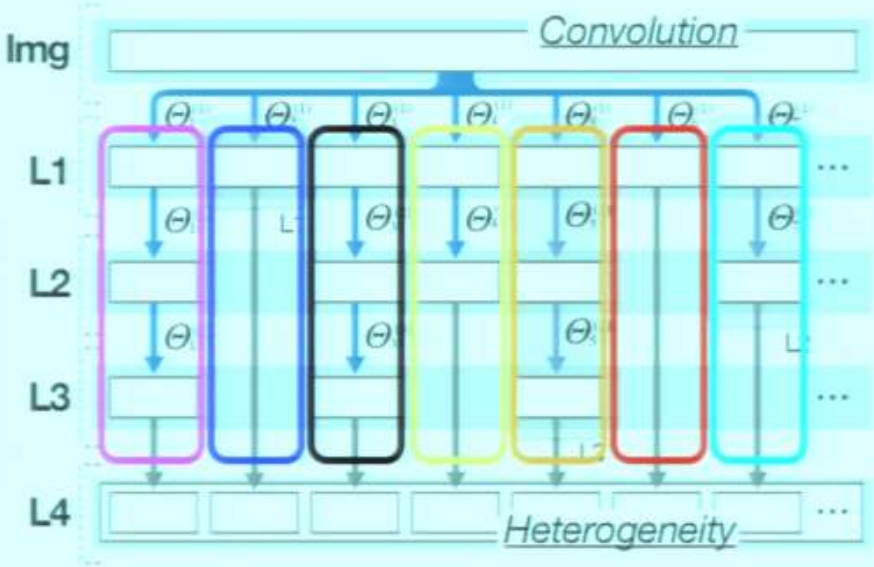


Dan Yamins

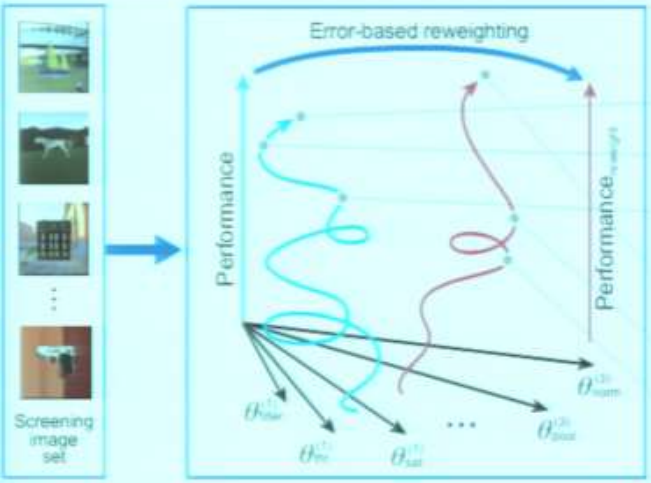
Ha Hong

Noticed that different types of object recognition tasks were best solved by different choices of architectural parameters

Suggested deep mixture model:



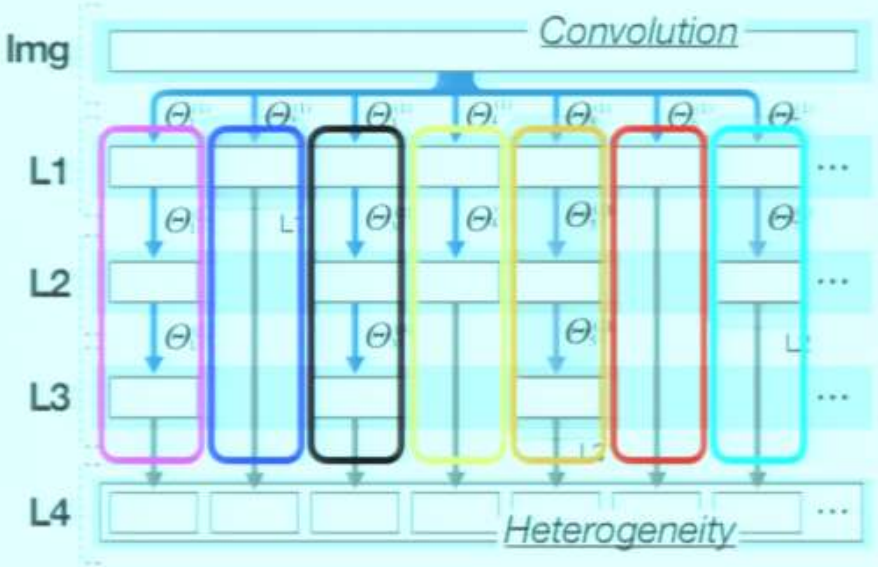
Extended bio-inspired (deep) model



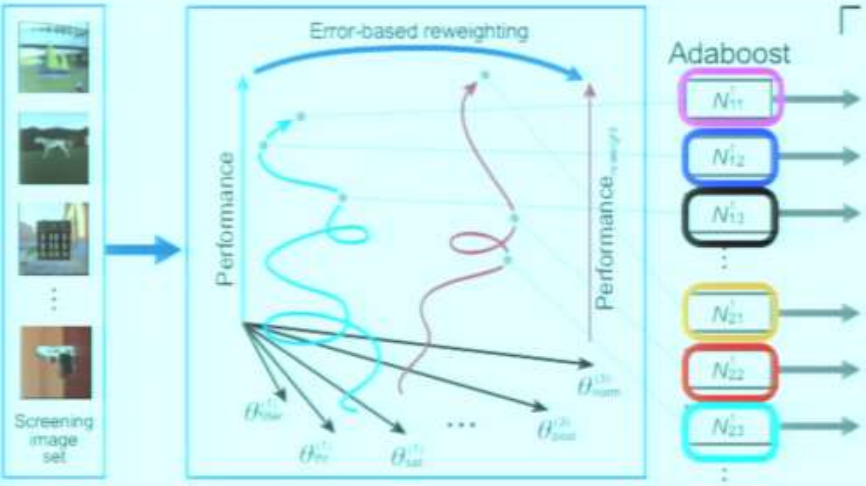
James Bergstra, Daniel Yamins, **David Cox**
Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms (2013)

Dan Yamins

Ha Hong

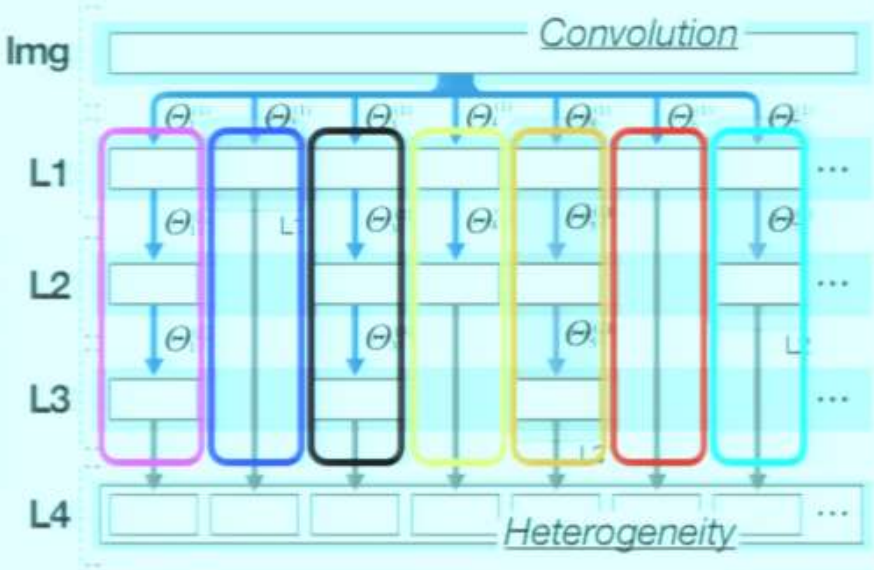


Extended bio-inspired (deep) model

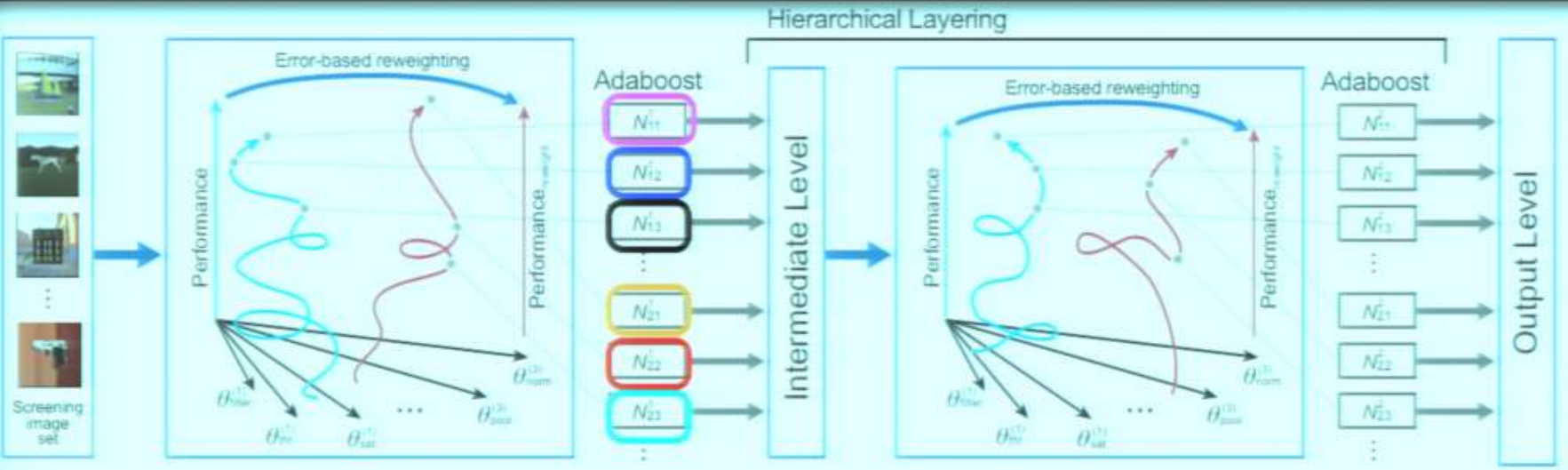


Dan Yamins

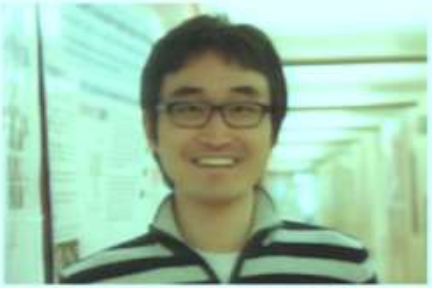
Ha Hong



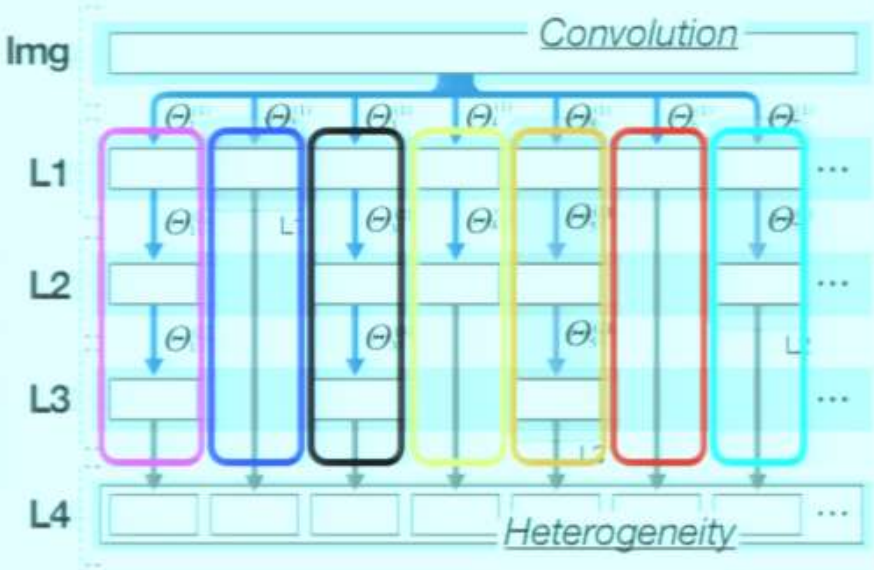
Extended bio-inspired (deep) model



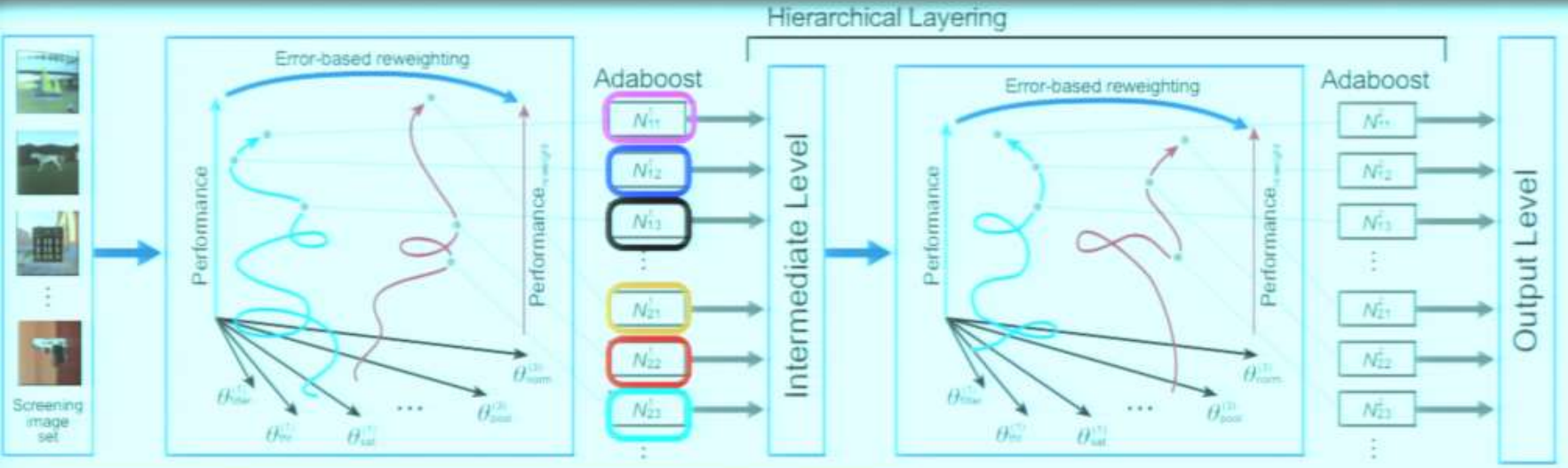
Dan Yamins



Ha Hong



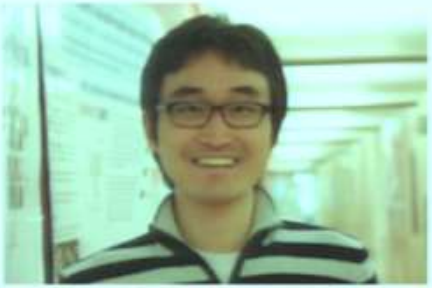
Extended bio-inspired (deep) model



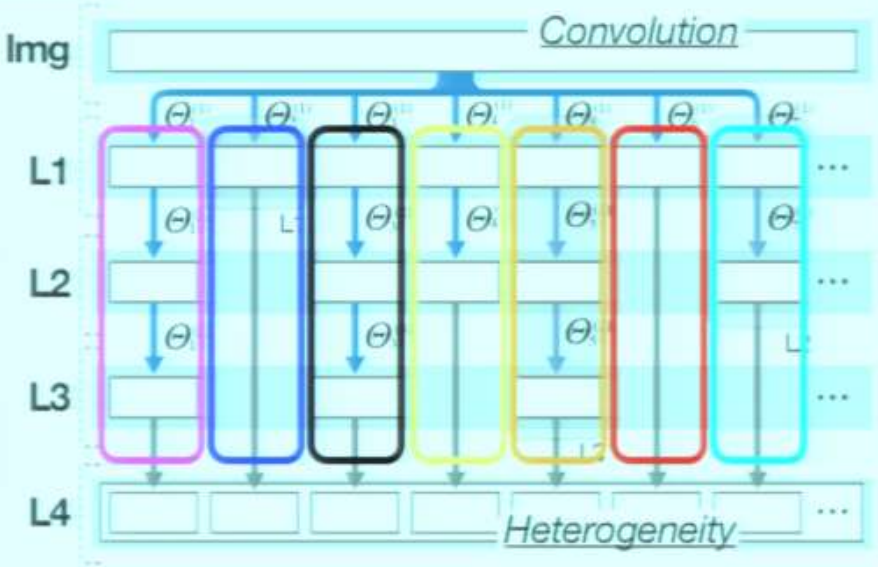
We are not wedded to this optimization.

Hierarchical modular optimization (HMO)

Dan Yamins



Ha Hong



Model screening images/tasks:

- variety of objects (36) with some semantic breadth (e.g. not all faces)
- no background/object correlation confounds
- rendered with large amount of variation ==> 4500 images

Bodies

Buildings

Flowers

Guns

Instruments

Jewelry

Shoes

Tools

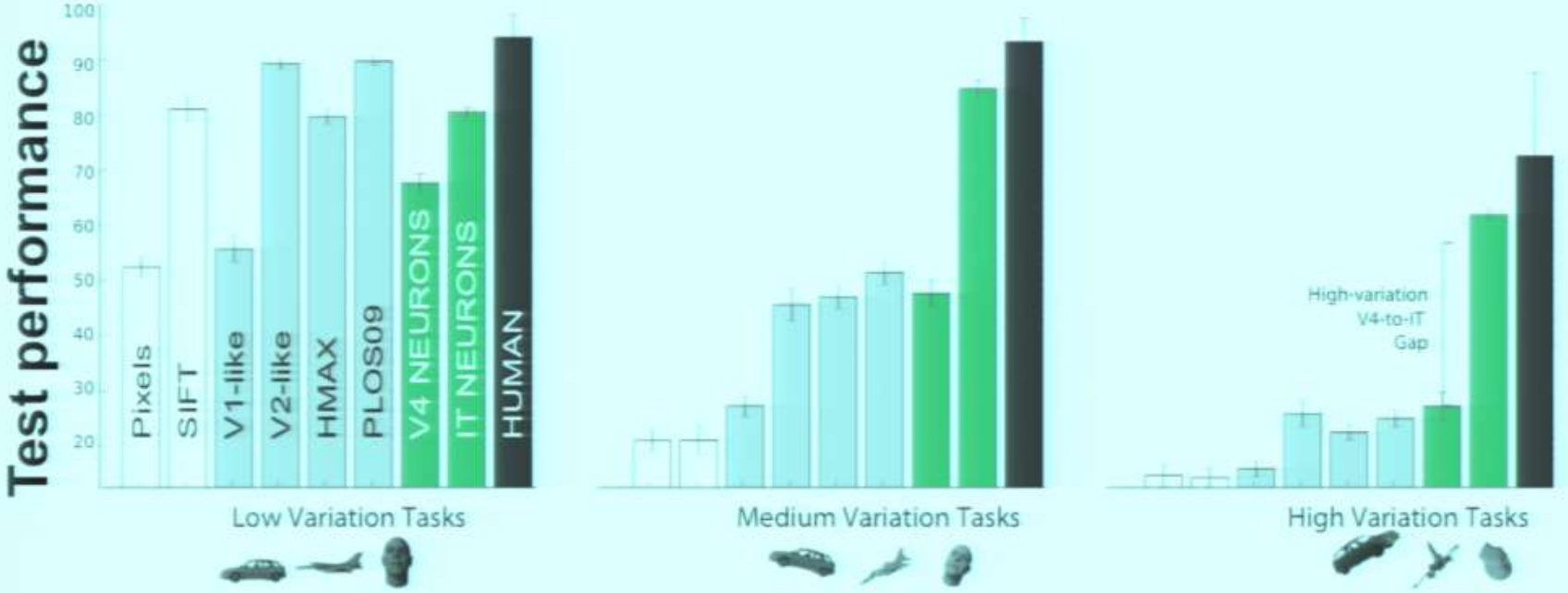
Trees

Model test images/tasks:

Object recognition
(HVM 1.0)

Model test images/tasks:

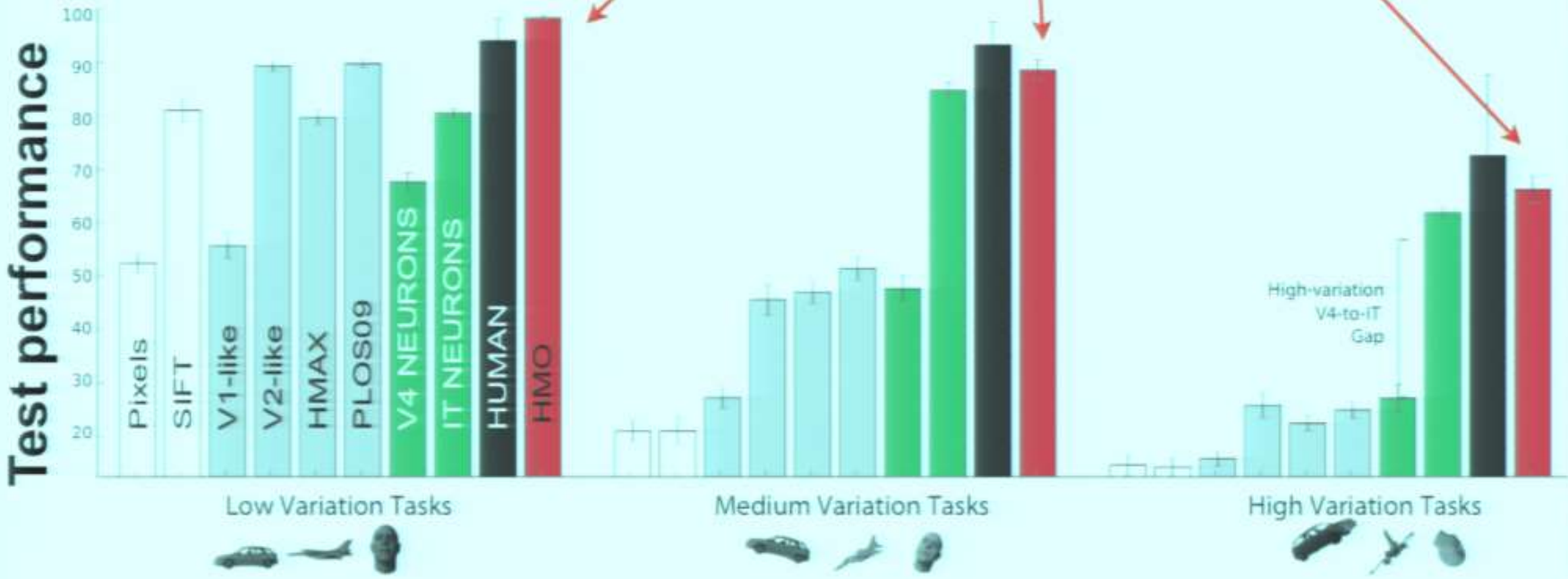
Object recognition (HVM 1.0)



Model test images/tasks:

Object recognition (HVM 1.0)

First model produced by the HMO procedure (HMO 1.0)

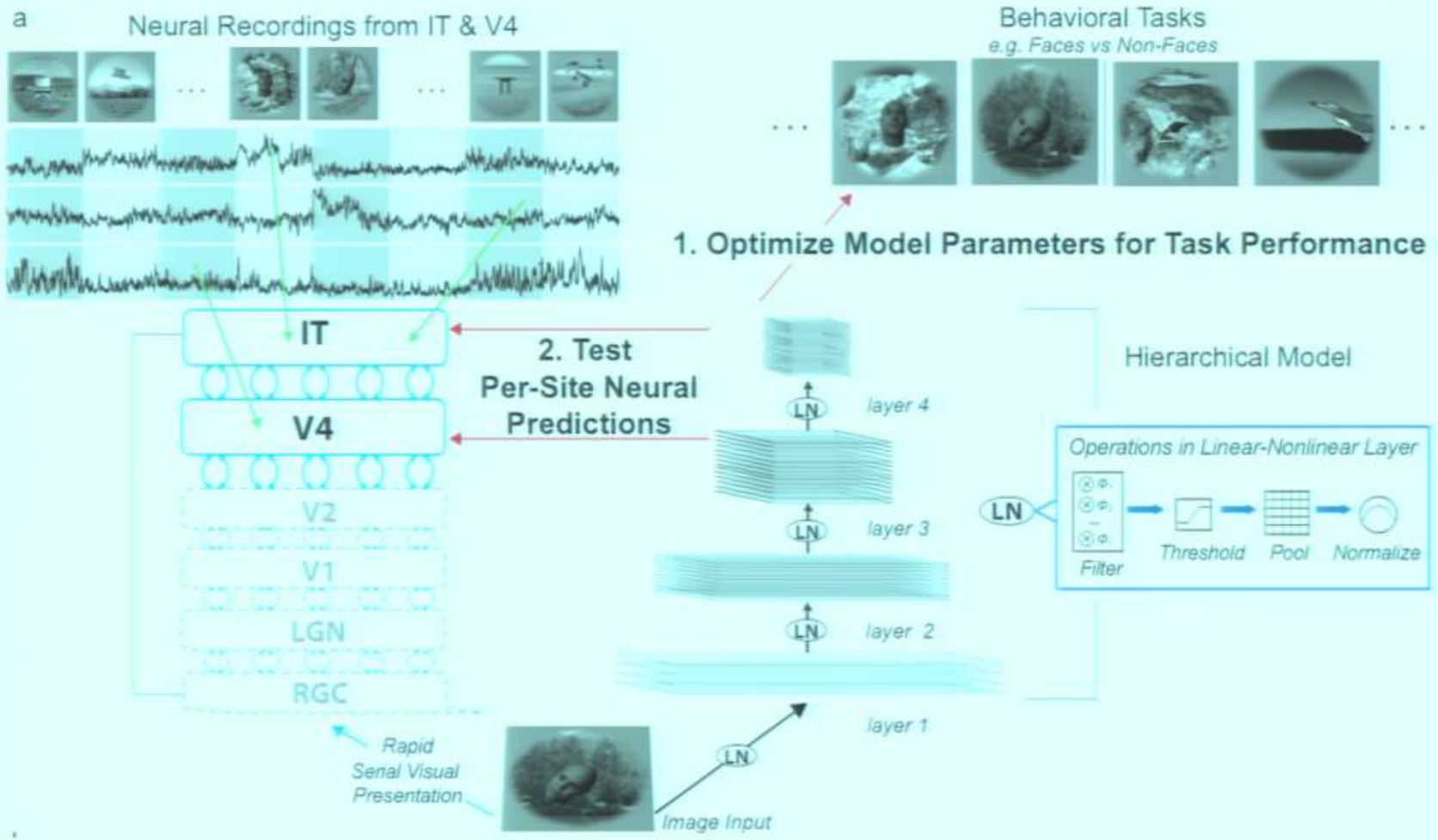


Our overarching strategy:

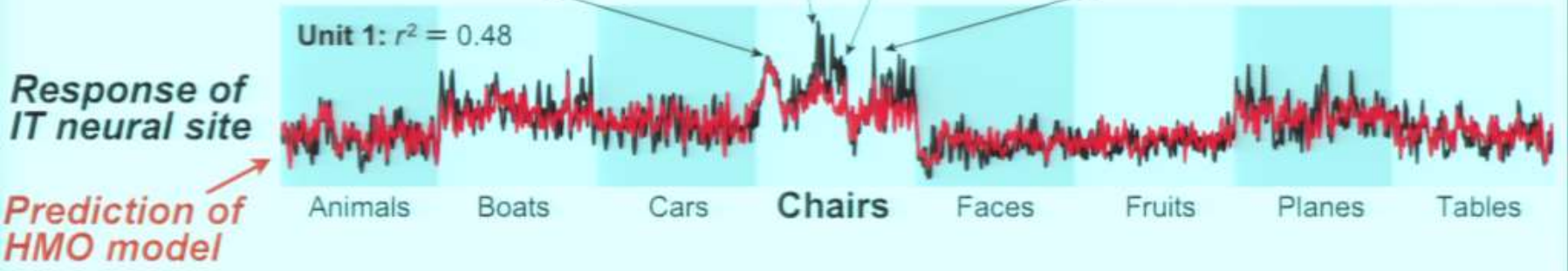
0. Bio-inspired model class

1. Optimize performance on tasks the brain is re. good at

2. Ask: do model features looks like the brains?

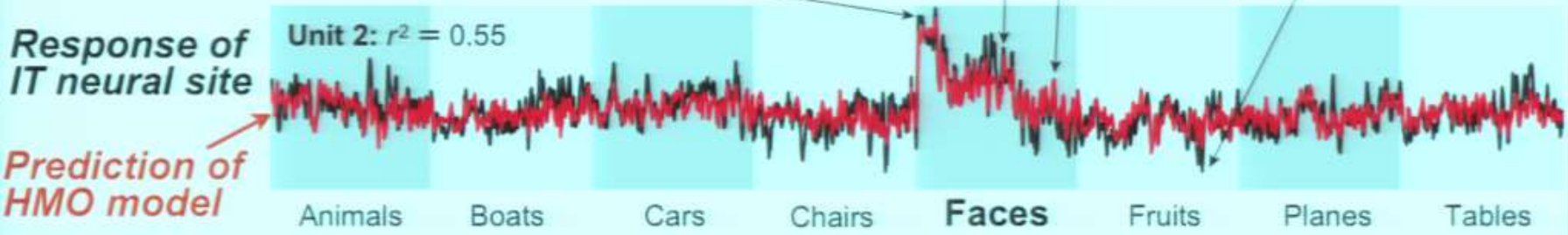


Predictions of single site IT responses from HMO 1.0 model



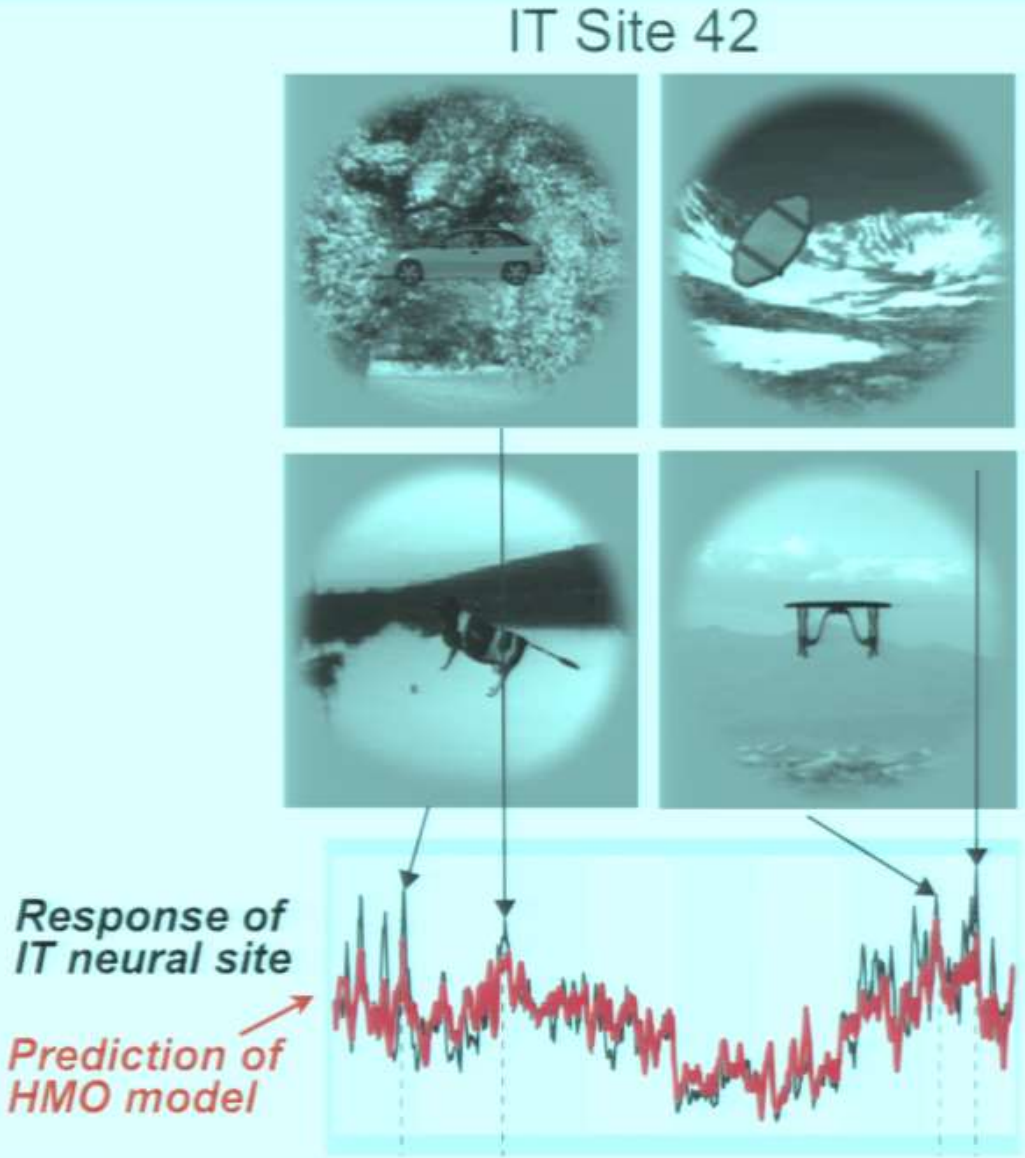
Yamins, Hong, Solomon, Seibert and DiCarlo (under review)

Predictions of single site IT responses from HMO 1.0 model

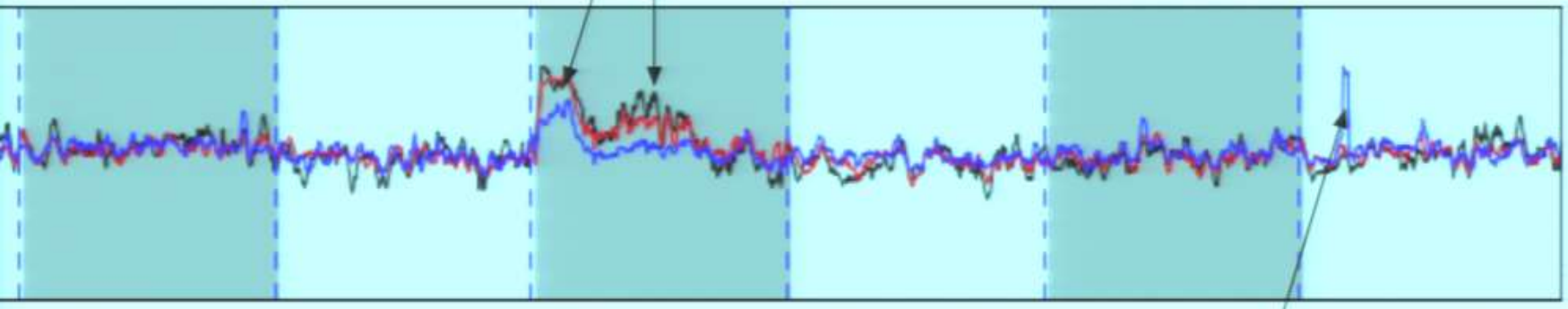


Yamins, Hong, Solomon, Seibert and DiCarlo (under review)

Predictions of single site IT responses from HMO 1.0 model



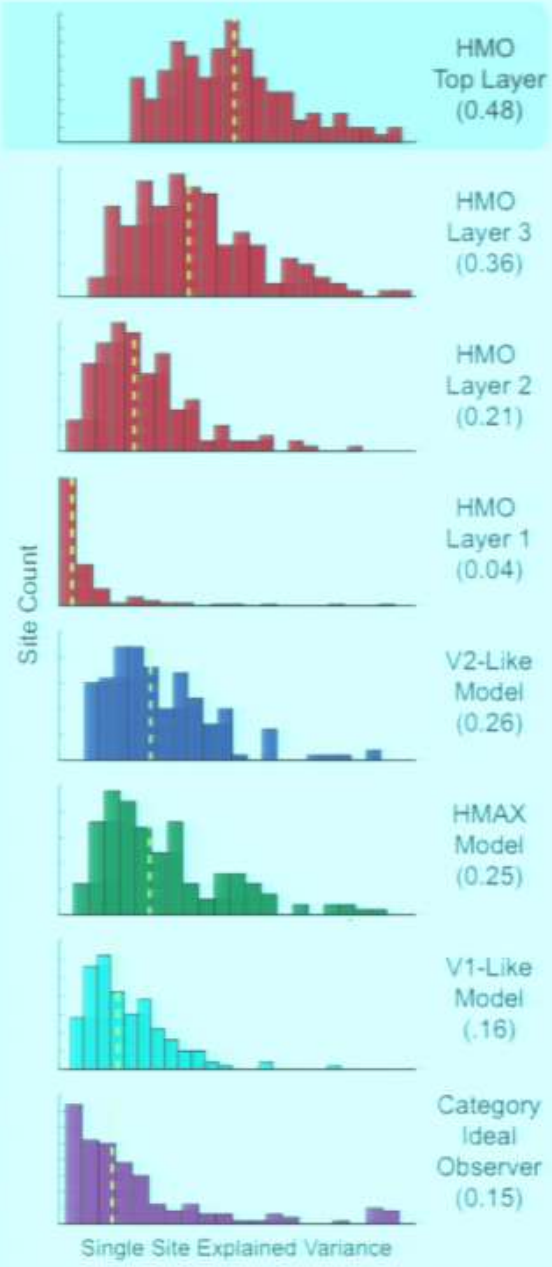
Yamins, Hong, Solomon, Seibert and DiCarlo (under review)



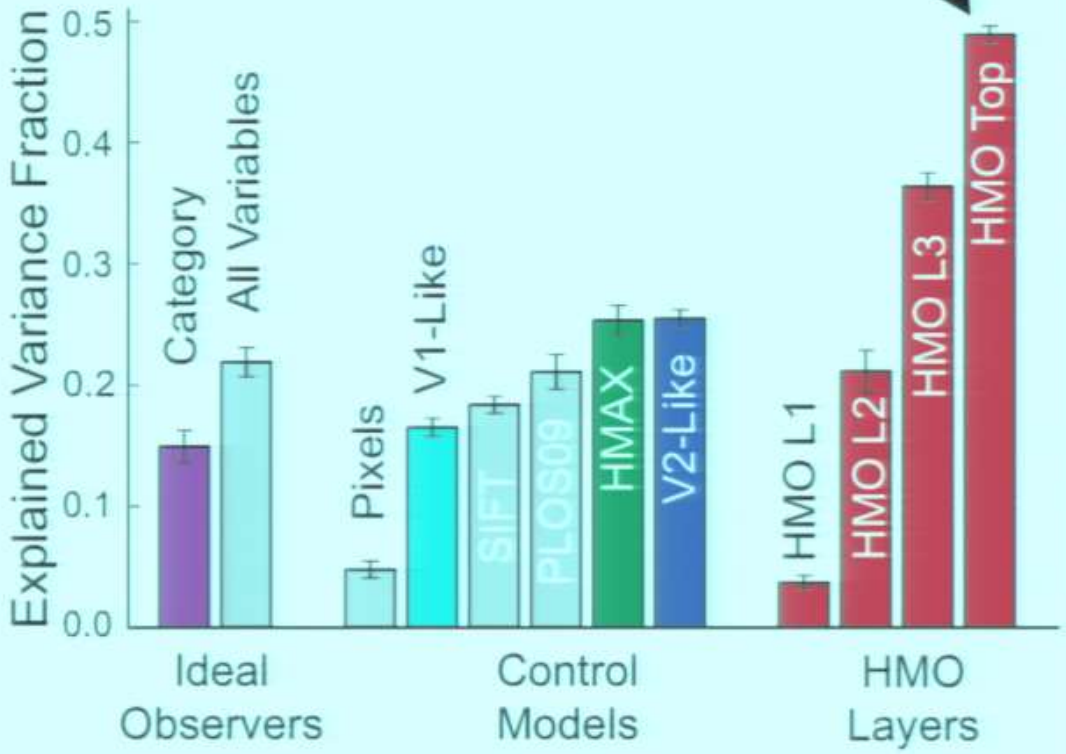
Cars Chairs **Faces** Fruits Planes Tables

— HMO prediction

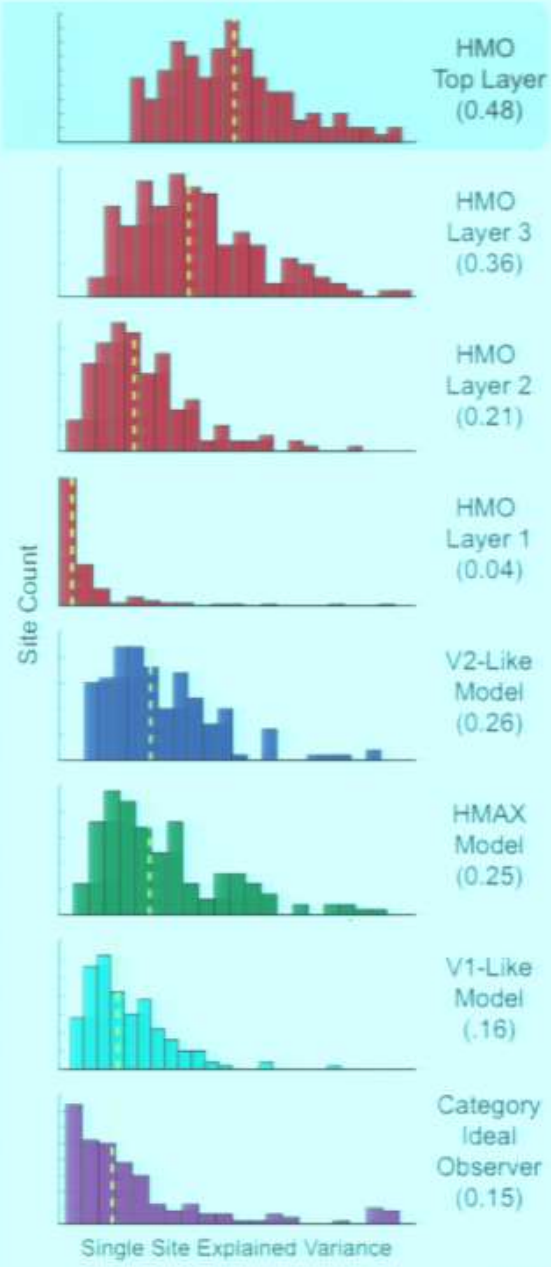
— V2-like prediction



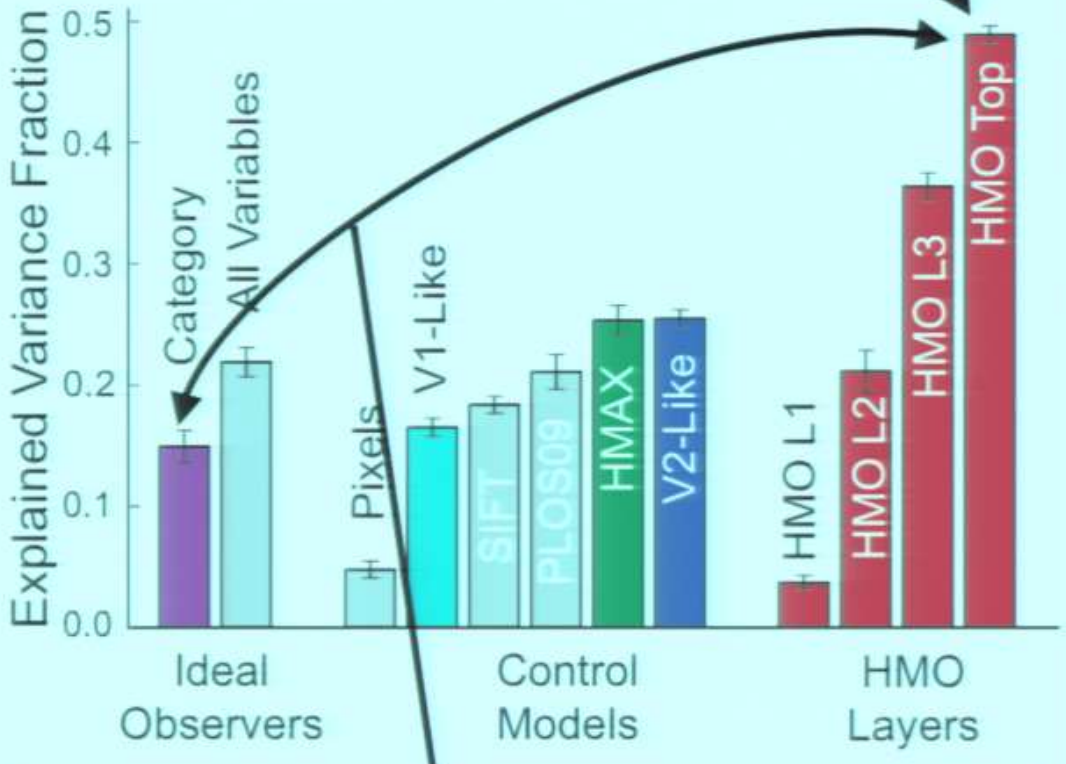
~50% of IT single unit response variance explained.
Dramatic improvement over previous models.



Yamins, Hong, Solomon, Seibert and DiCarlo (under review)



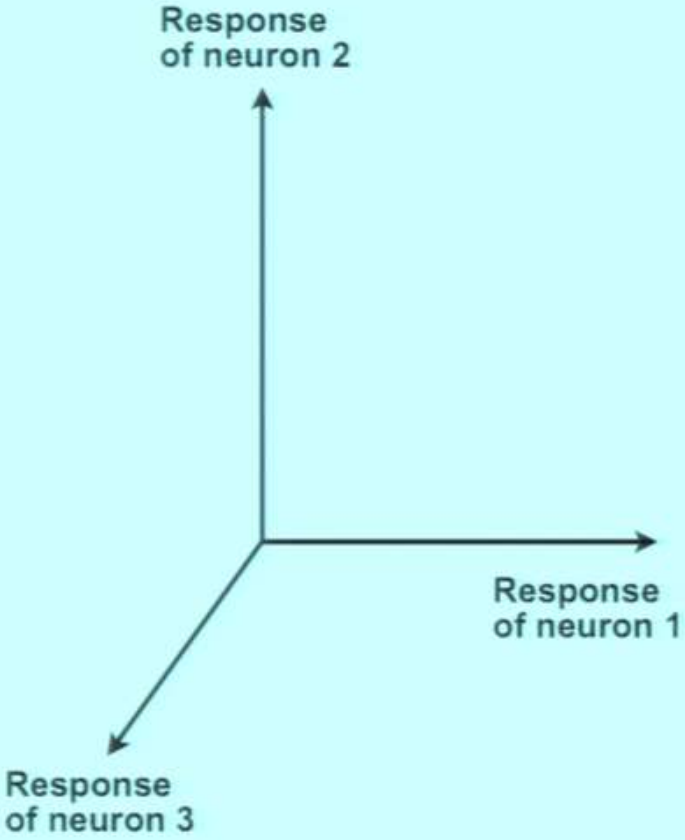
~50% of IT single unit response variance explained. Dramatic improvement over previous models.



Model class plus optimization criteria are inducing brain-like structure beyond that induced by the task

Yamins, Hong, Solomon, Seibert and DiCarlo (under review)

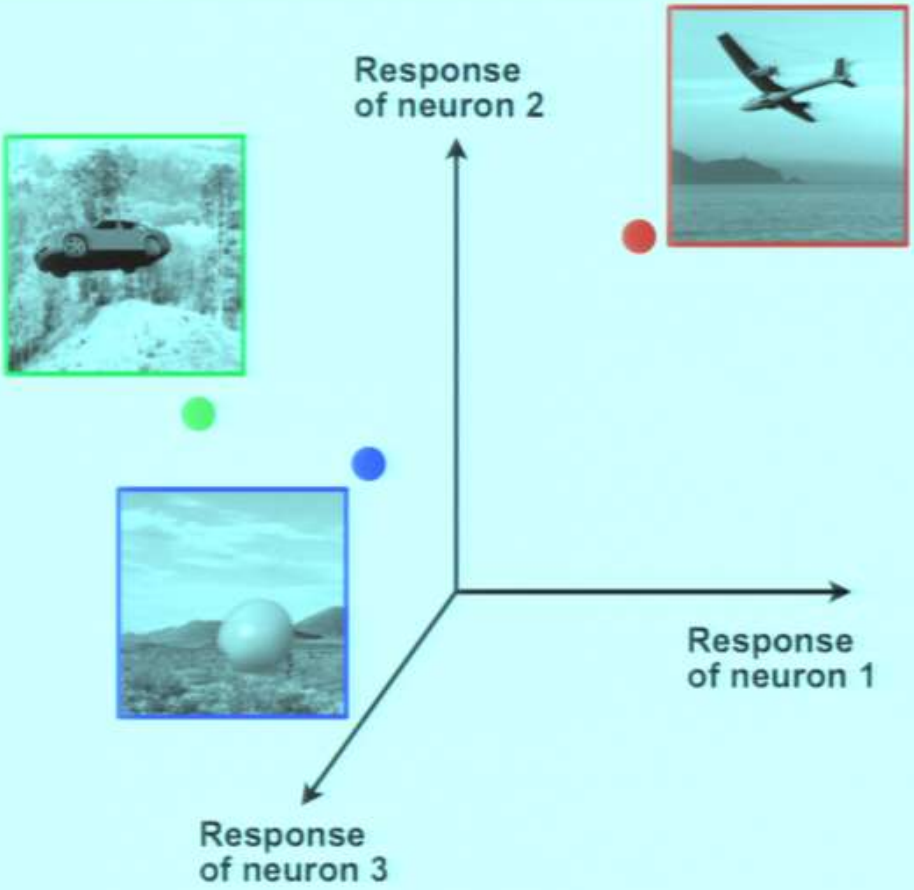
Comparing two population representations



Layout of images in
neuronal space (e.g. IT)

see Kriegeskorte et al. (2008)

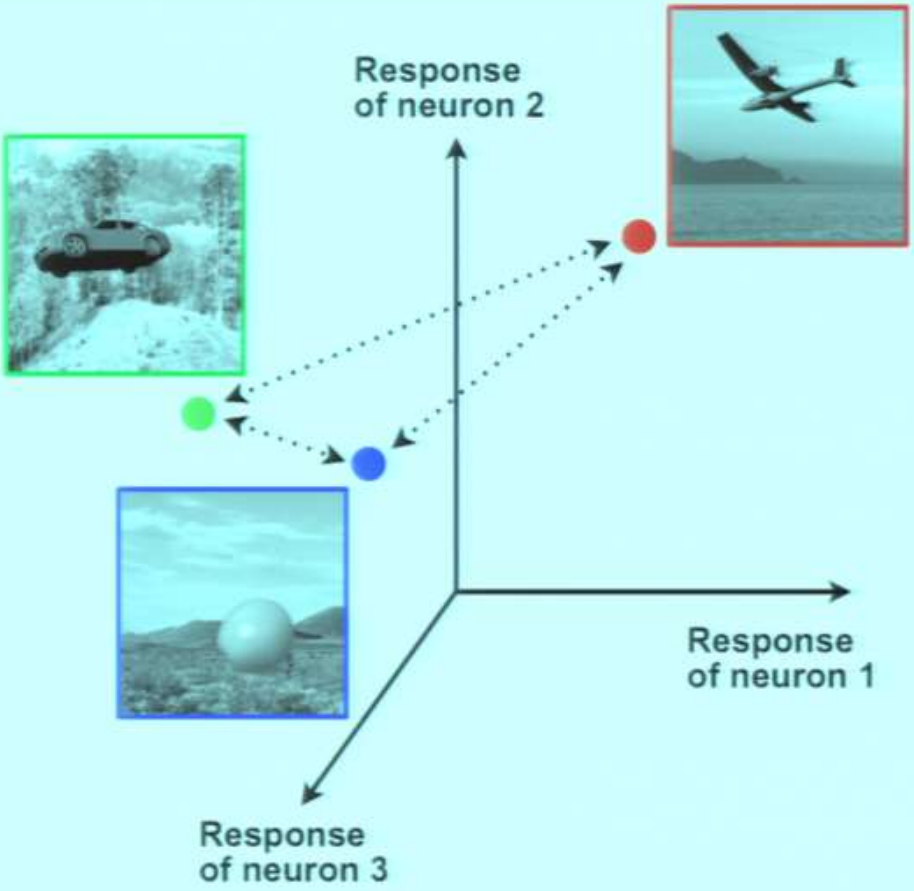
Comparing two population representations



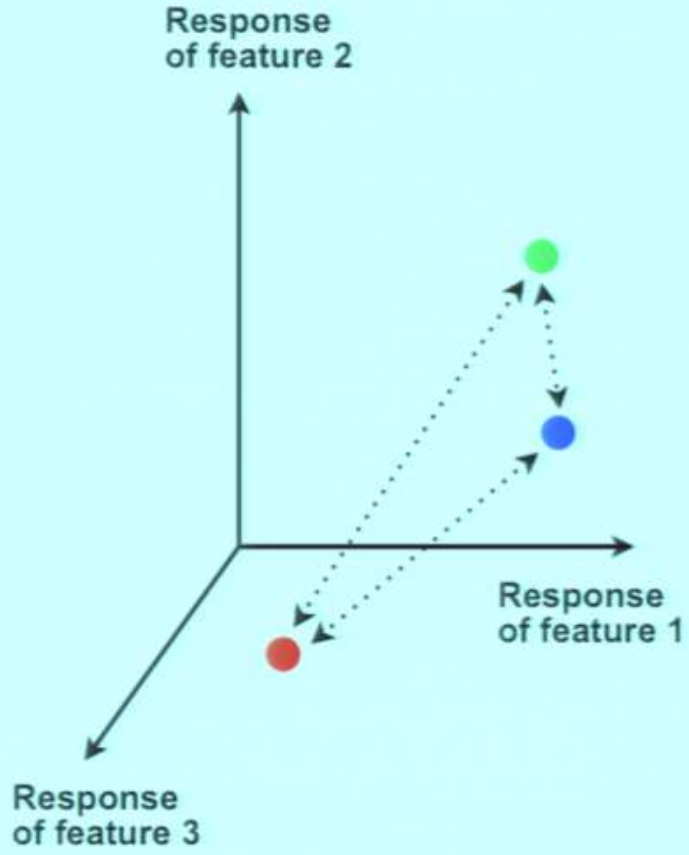
Layout of images in
neuronal space (e.g. IT)

see Kriegeskorte et al. (2008)

Comparing two population representations



Layout of images in neuronal space (e.g. IT)

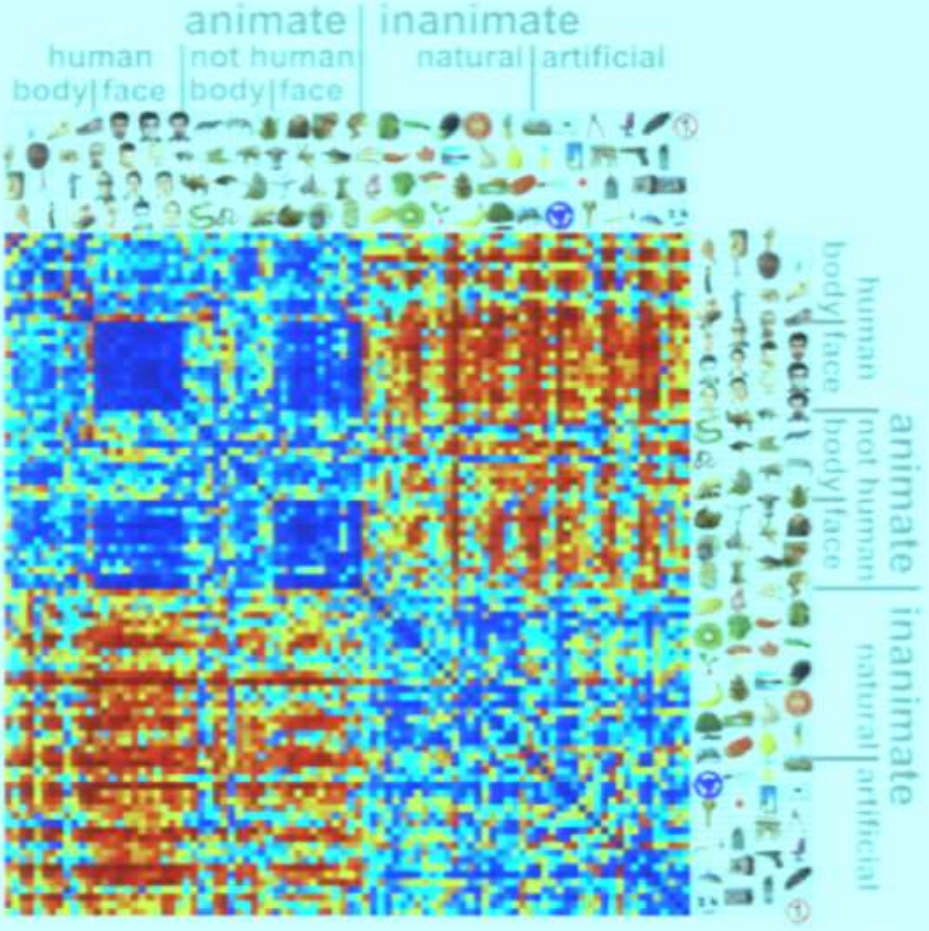


Layout of same images in feature representation of any putative model of the ventral stream

see Kriegeskorte et al. (2008)

Representation Dissimilarity Matrices (Kriegeskorte, 2008)

$$M_{ij} = 1 - correlation(r_i, r_j)$$



Low M_{ij} (blue) means the two stimuli are close in feature/neural population response space

High M_{ij} (red) means the two stimuli are far in feature/neural space

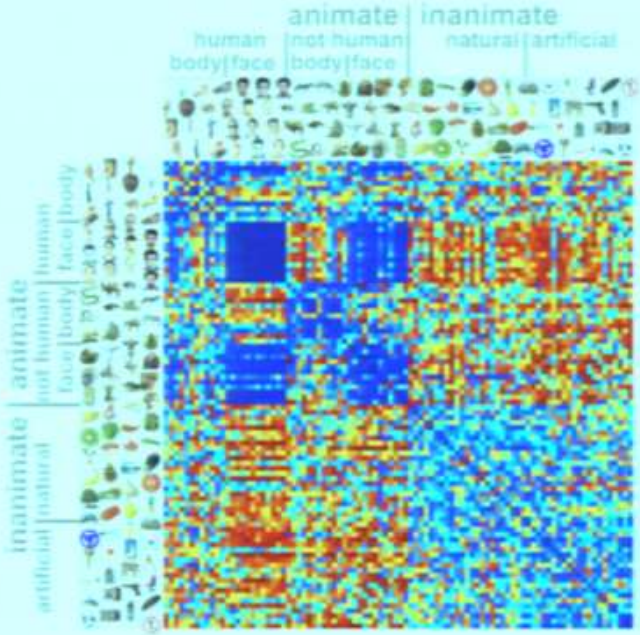
Representation Dissimilarity Matrices

- RDMs allow comparison of any two feature representations on a common stimulus set

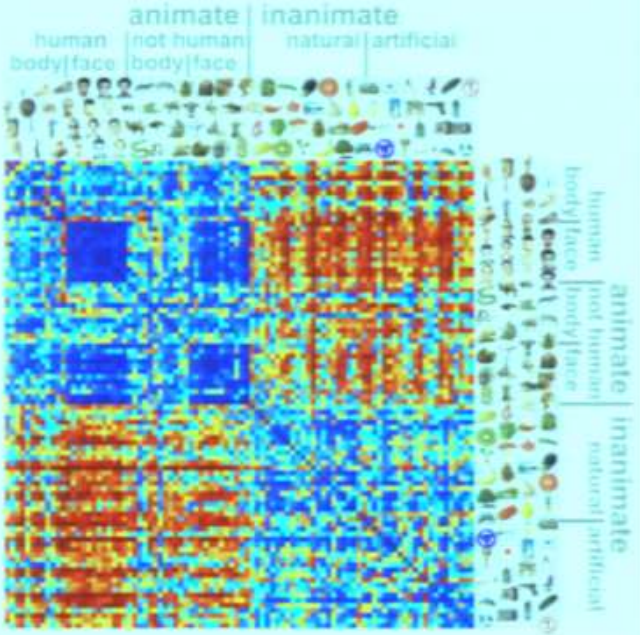
Representation Dissimilarity Matrices

- RDMs allow comparison of any two feature representations on a common stimulus set
  - IT (Monkey 1) vs. IT (Monkey2)
  - IT vs. V4
  - IT vs. Model X
  - Monkey IT vs. Human "IT"

Monkey IT
(neural spiking responses)



Human LOC ("IT")
(fMRI voxel responses)



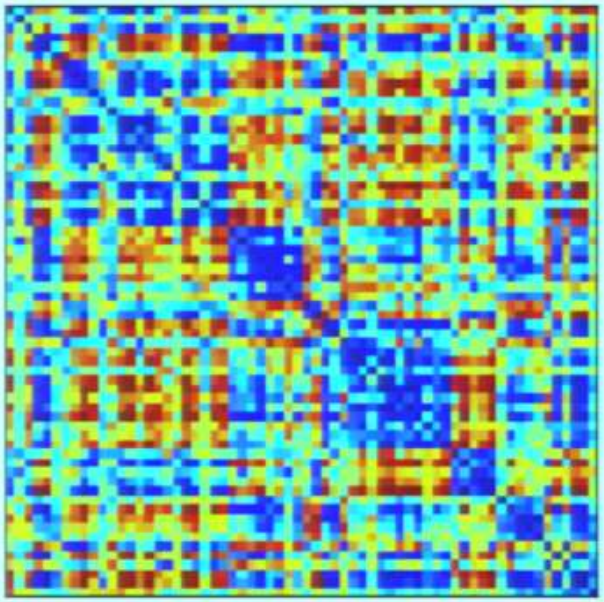
Kriegeskorte, 2008

Representation Dissimilarity Matrices

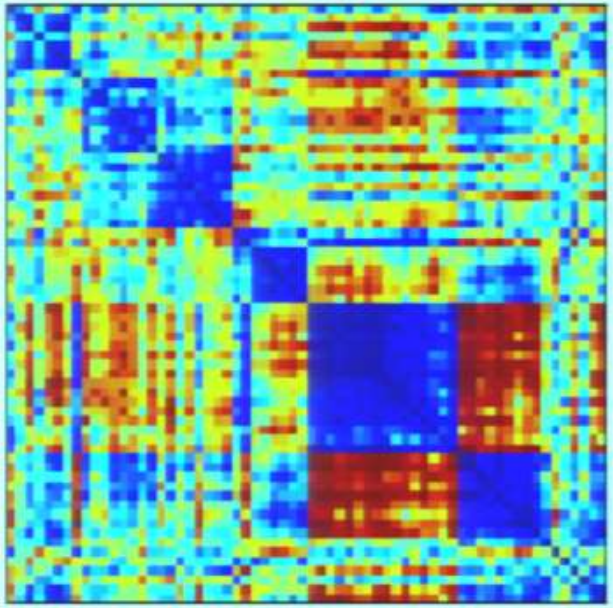
- RDM structure echoes the performance of the population code

e.g. Images for Object recognition 1.0 (HVM 1.0)

Monkey V4



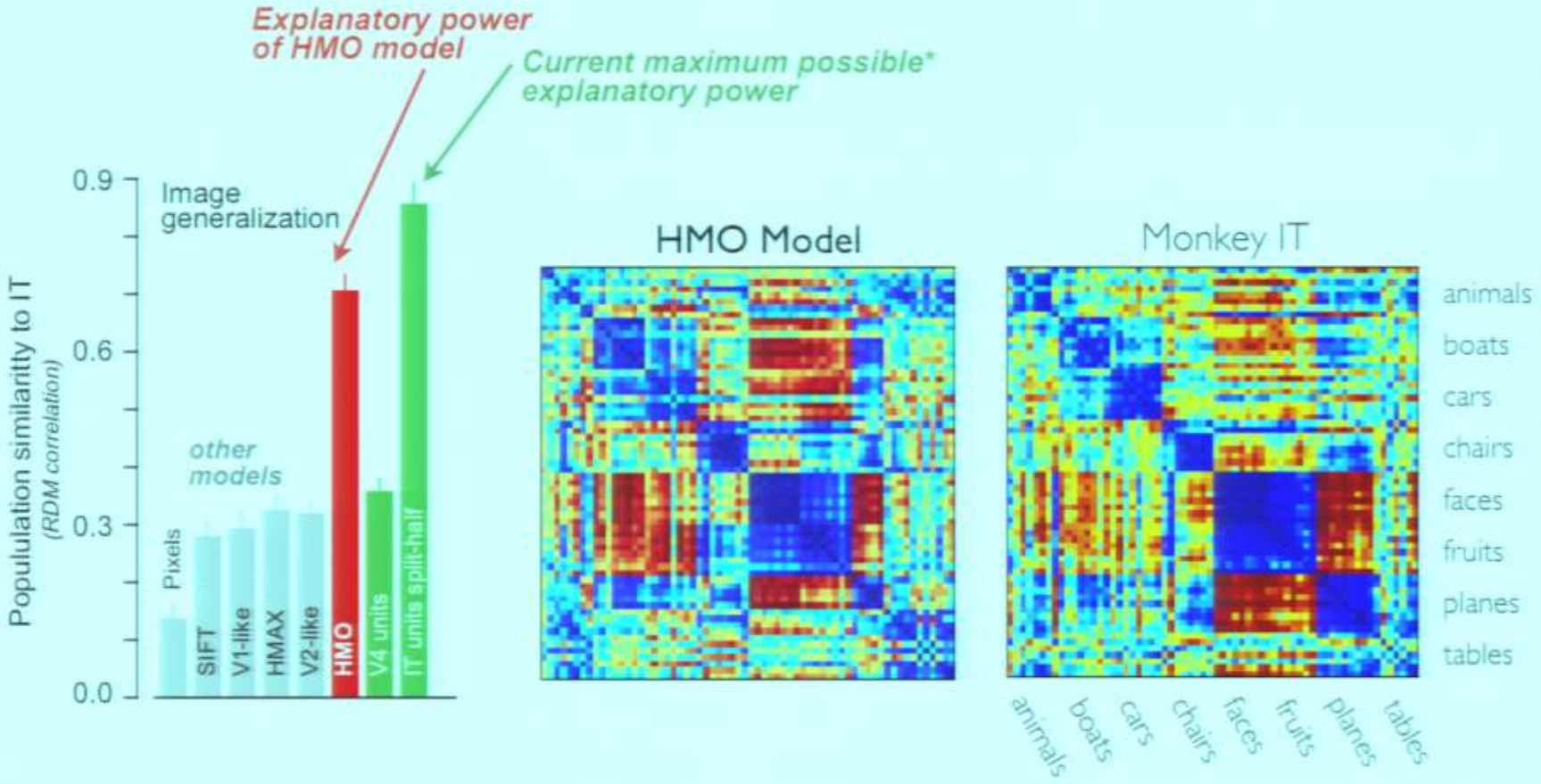
Monkey IT



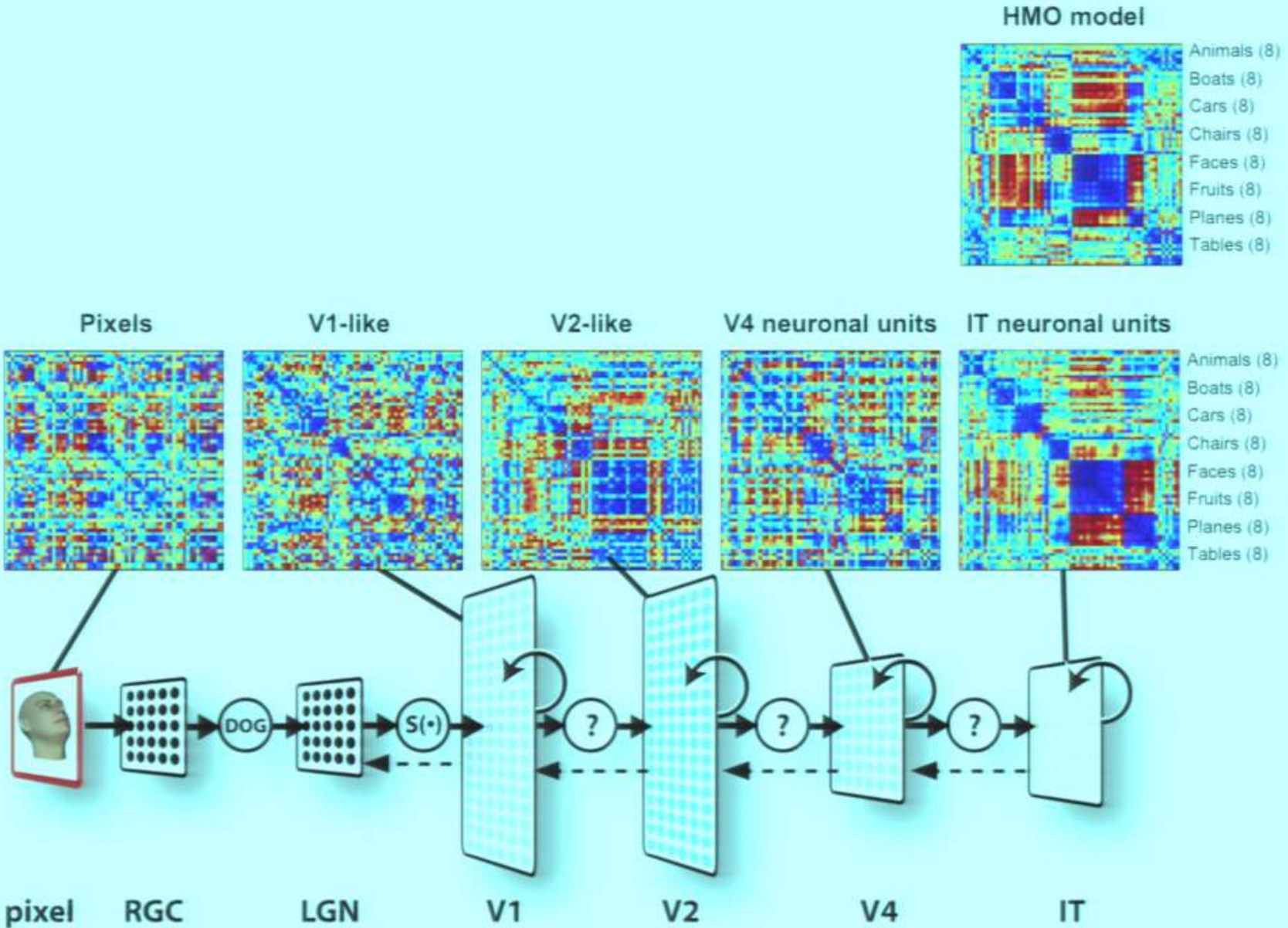
animals
boats
cars
chairs
faces
fruits
planes
tables

Representation Dissimilarity Matrices: models vs. IT

Model captures diagonal and off-diagonal RDM structure effectively.



Yamins, Hong, Solomon, Seibert and DiCarlo (under review)

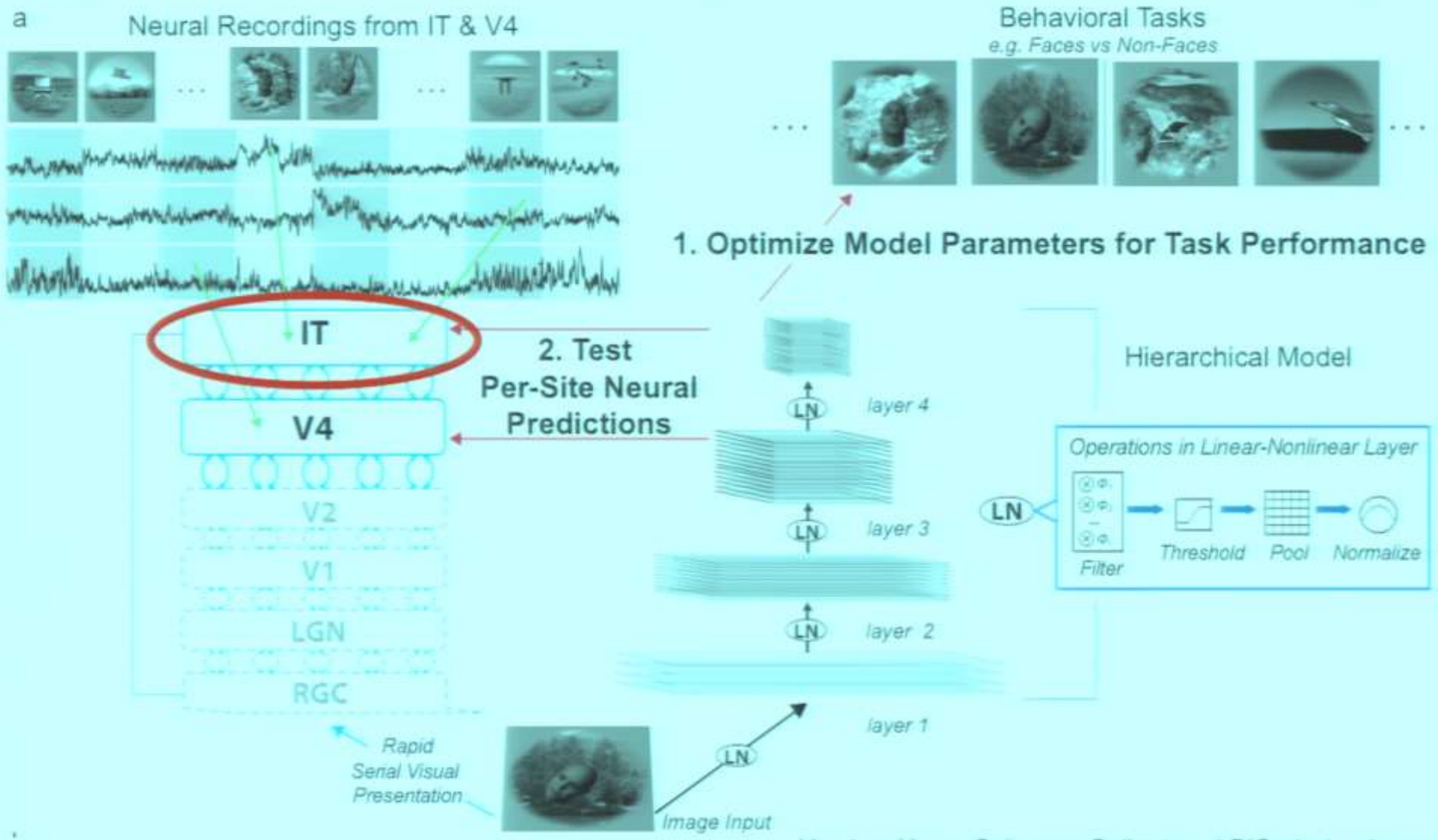

HMO model
Animals (8)
Boats (8)
Cars (8)
Chairs (8)
Faces (8)
Fruits (8)
Planes (8)
Tables (8)
Pixels
V1-like
V2-like
V4 neuronal units
IT neuronal units
Animals (8)
Boats (8)
Cars (8)
Chairs (8)
Faces (8)
Fruits (8)
Planes (8)
Tables (8)
DOG
S(•)
?
?
?
pixel
RGC
LGN
V1
V2
V4
IT

Our overarching strategy:

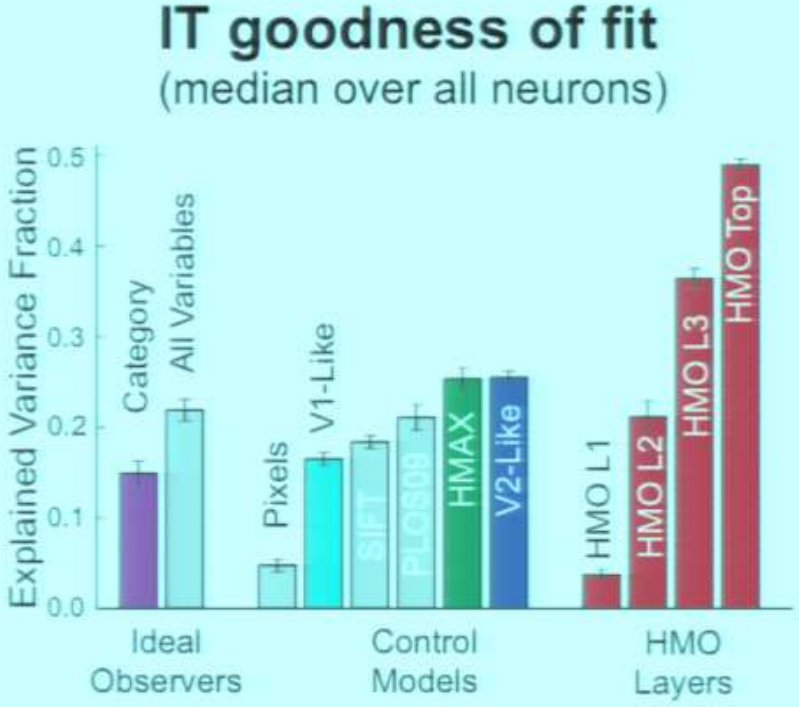
0. Bio-inspired model class

1. Optimize performance on tasks the brain is re. good at

2. Ask: do model features looks like the brains?

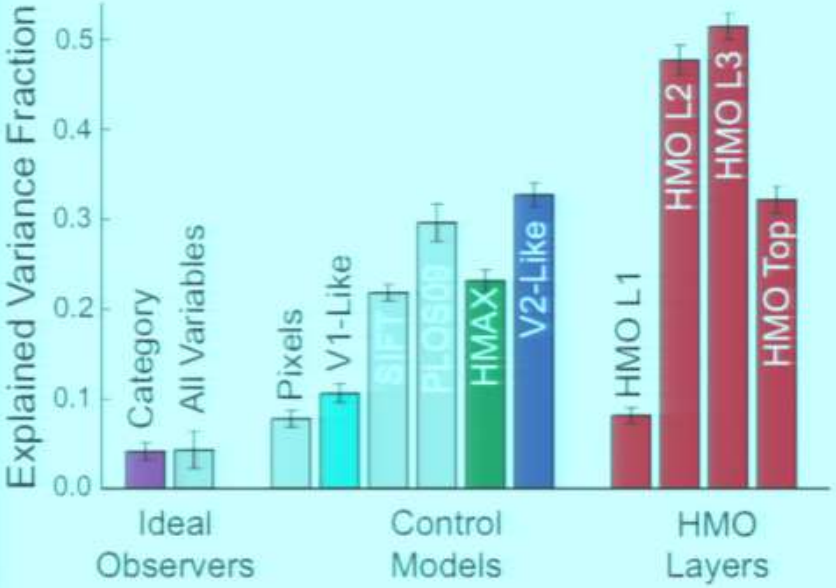


Yamins, Hong, Soloman, Seibert and DiCarlo (under review)

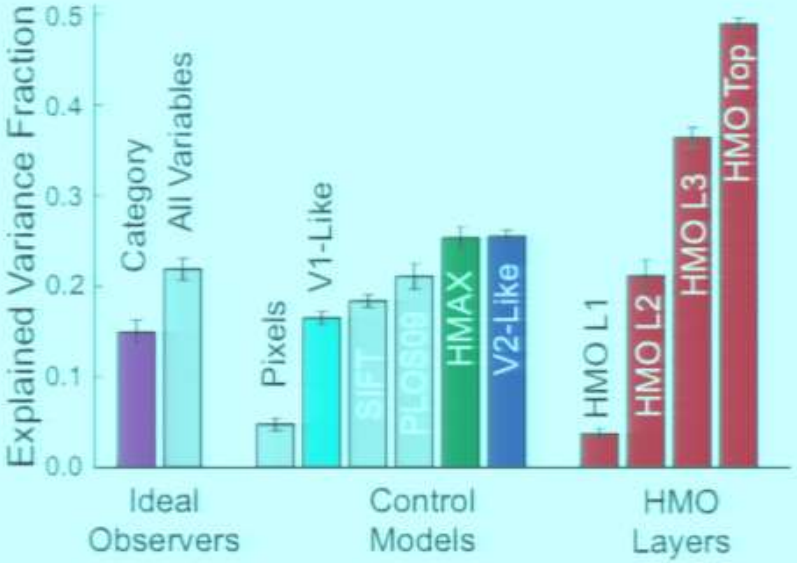


Yamins, Hong, Solomon, Seibert and DiCarlo (under review)

V4 goodness of fit
(median over all neurons)



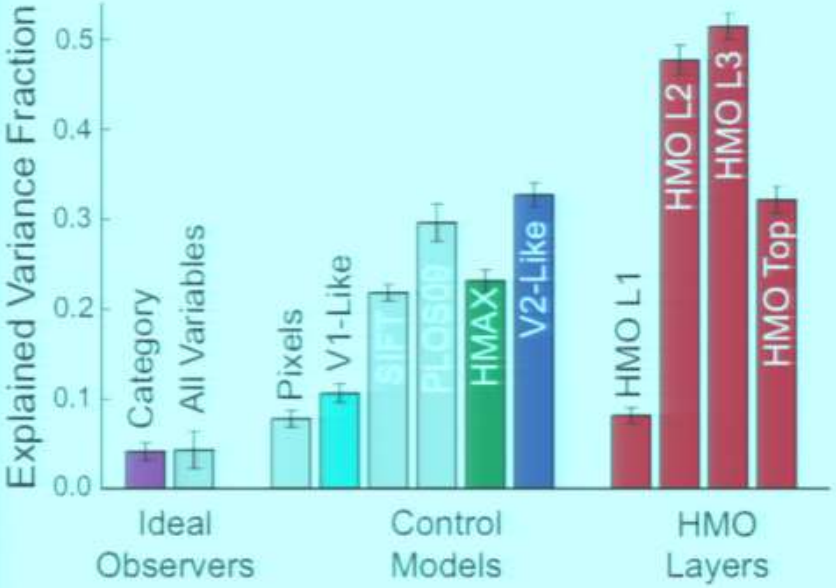
IT goodness of fit
(median over all neurons)



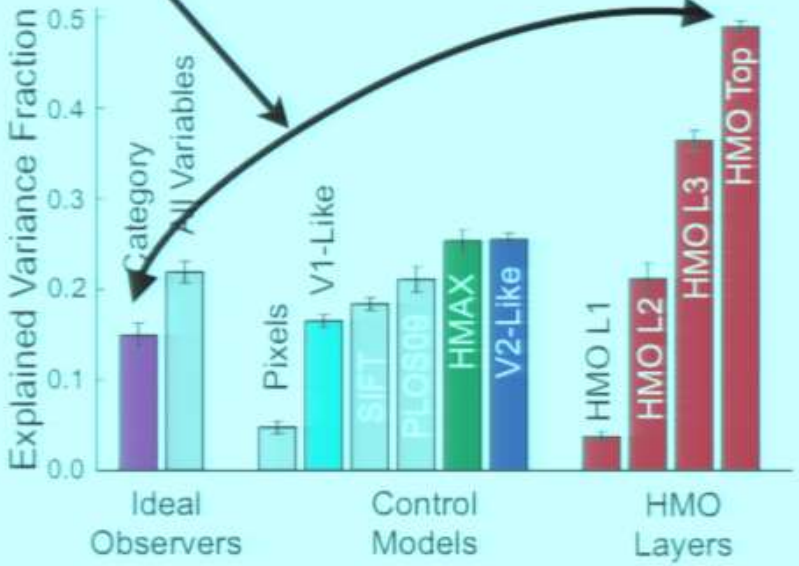
Yamins, Hong, Solomon, Seibert and DiCarlo (under review)

Model class plus optimization criteria is inducing brain-like structure beyond that induced by the task

V4 goodness of fit
(median over all neurons)

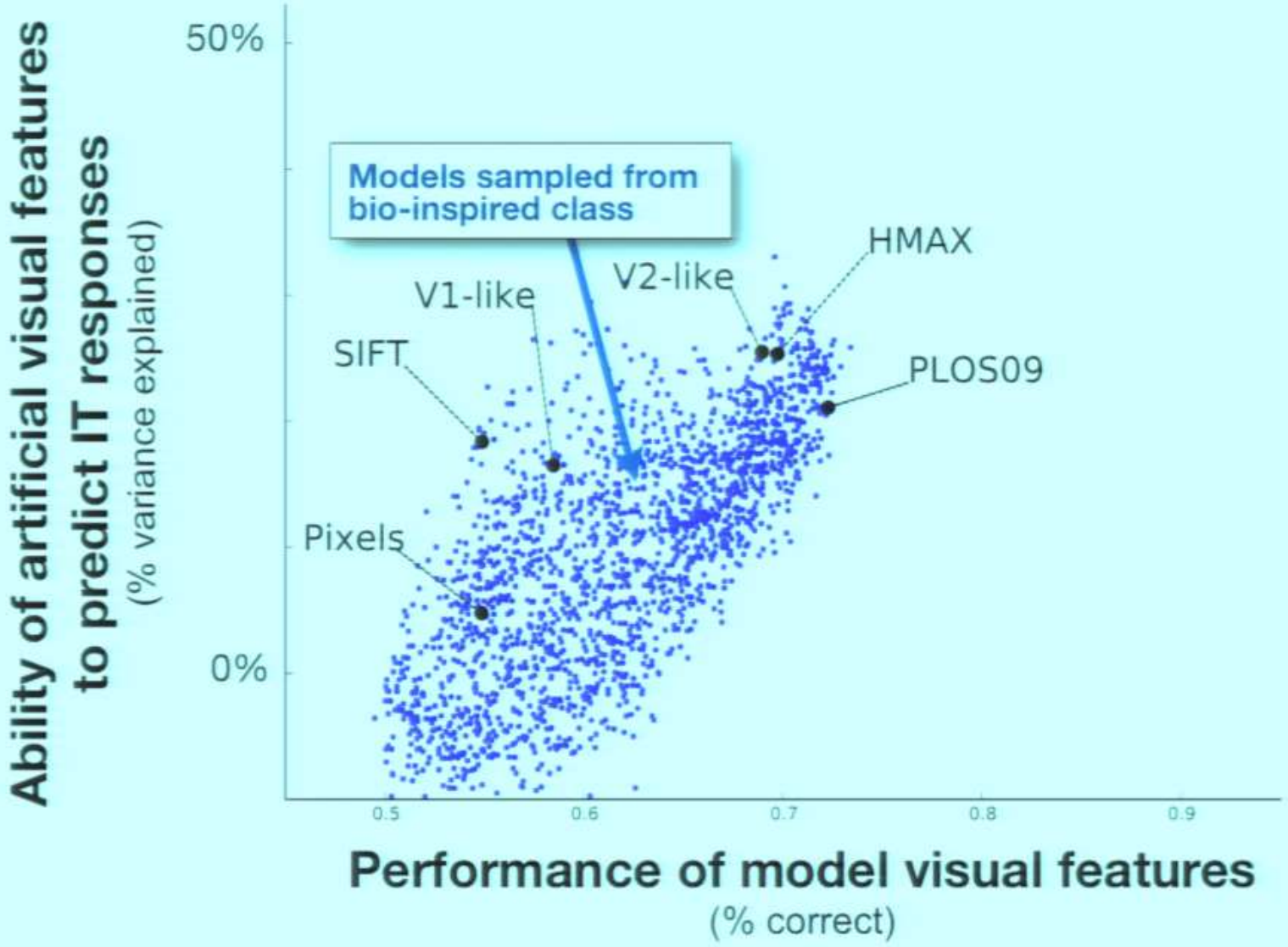


IT goodness of fit
(median over all neurons)



Yamins, Hong, Solomon, Seibert and DiCarlo (under review)

We now have a new way forward to understanding the ventral stream



See Poster T63 tonight!
Yamins, Hong et al. NIPS 2013

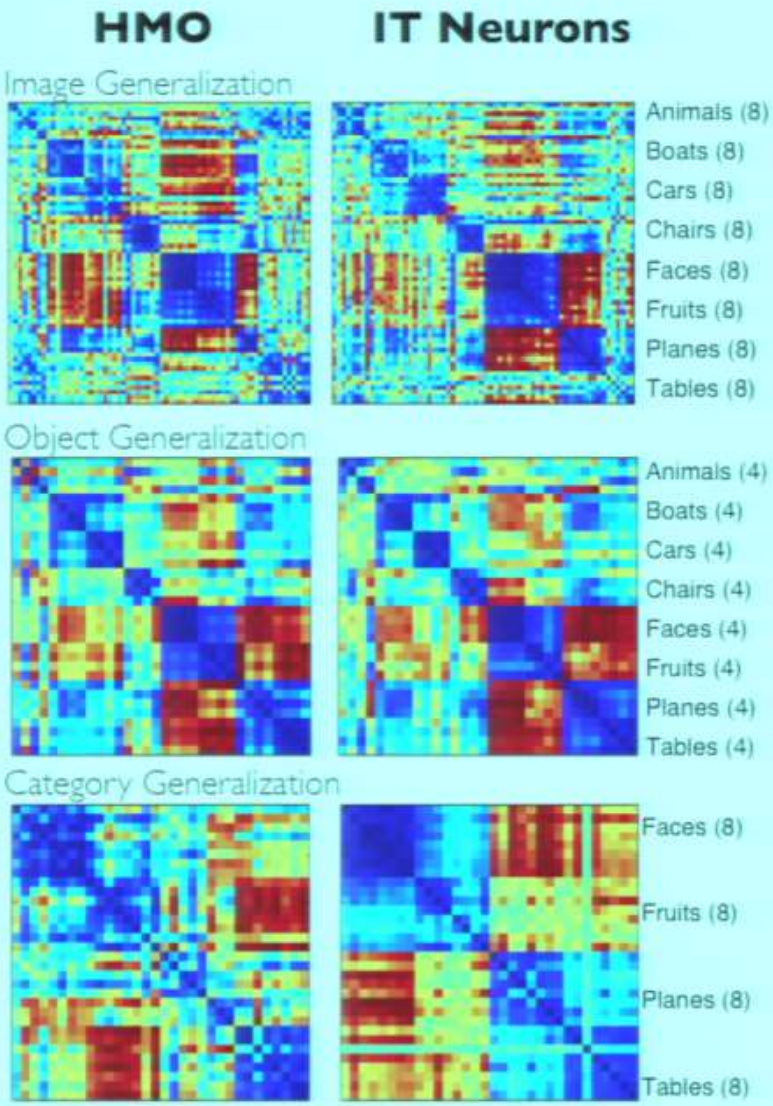
2. Machines vs. Monkey neurons

Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

3. Machines vs. Monkey neurons/Human behavior

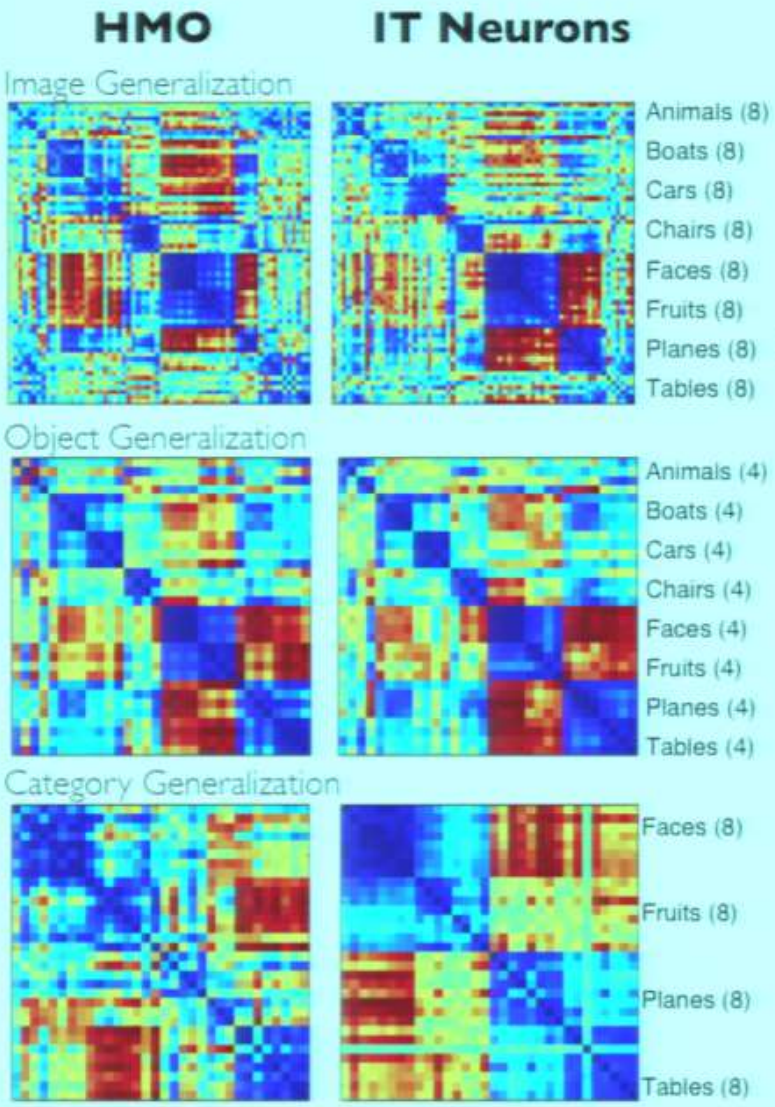
Demonstrates the recent bio-inspired models rival the brain in object recognition

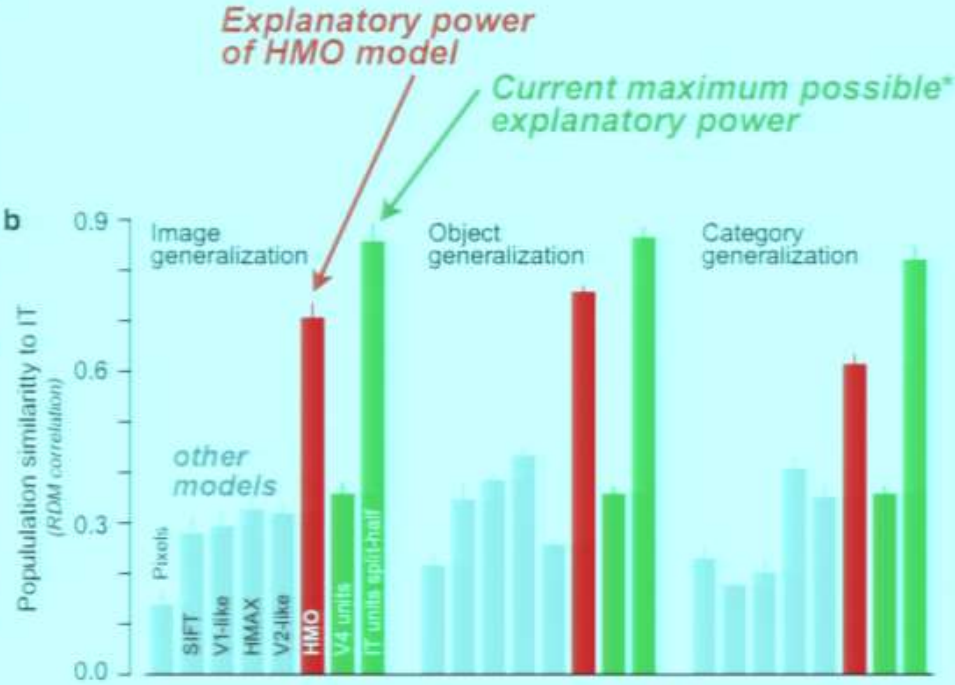
Representation Dissimilarity Matrices: models vs. IT



Yamins, Hong, Soloman, Seibert and DiCarlo (under review)

Representation Dissimilarity Matrices: models vs. IT





Ability of current best performing model to predict IT population is extremely good

Yamins, Hong, Soloman, Seibert and DiCarlo (under review)

1. **Monkey neurons vs. Human Behavior**

   Suggests that IT population codes are one simple step from object recognition behavior

2. **Machines vs. Monkey neurons**

   Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

3. **Machines vs. Monkey neurons/Human behavior**

   Demonstrates the recent bio-inspired models rival the brain in object recognition

What about other networks built for high performing object recognition? (e.g. DNNs)

Charles Cadieu

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

What about other networks built for high performing object recognition? (e.g. DNNs)

Charles Cadieu

Krizhevsky et al. (2012) SuperVision

Zeiler and Fergus (2013)

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

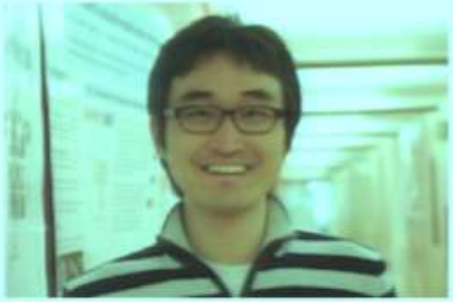
What about other networks built for high performing object recognition? (e.g. DNNs)

Charles Cadieu

Krizhevsky et al. (2012) SuperVision

Zeiler and Fergus (2013)

Najib Majaj



Ha Hong

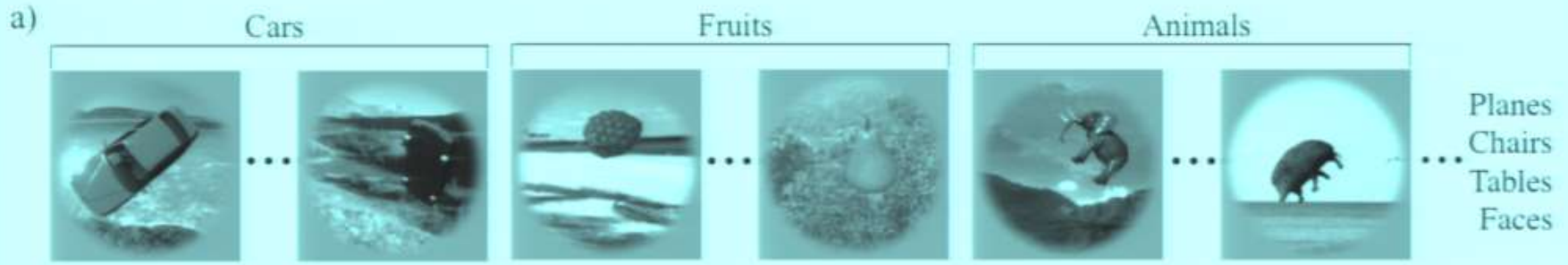
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Brain features vs. Machine features

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Brain features vs. Machine features

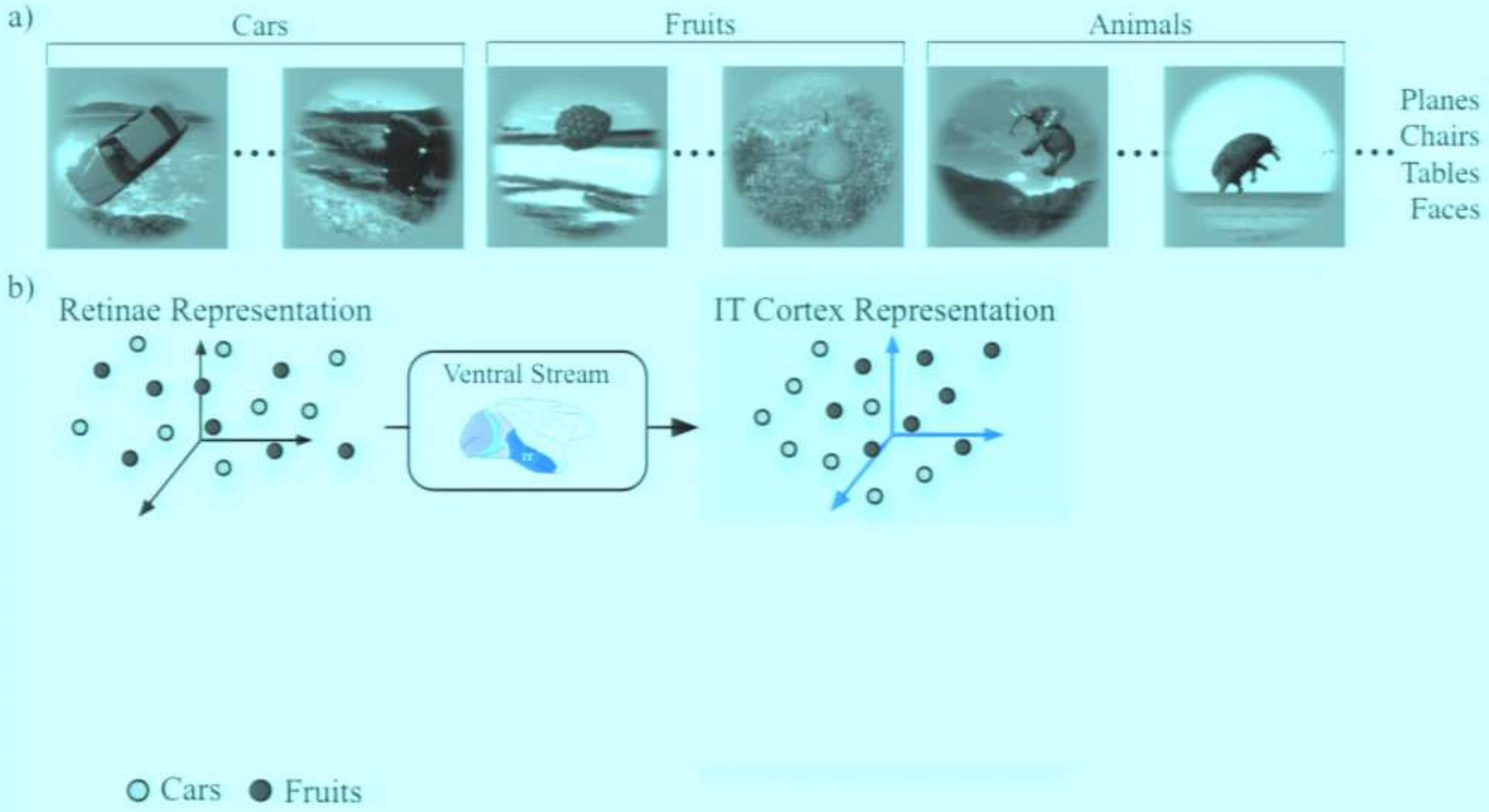
Object recognition 1.0 (HVM1.0)



Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Brain features vs. Machine features

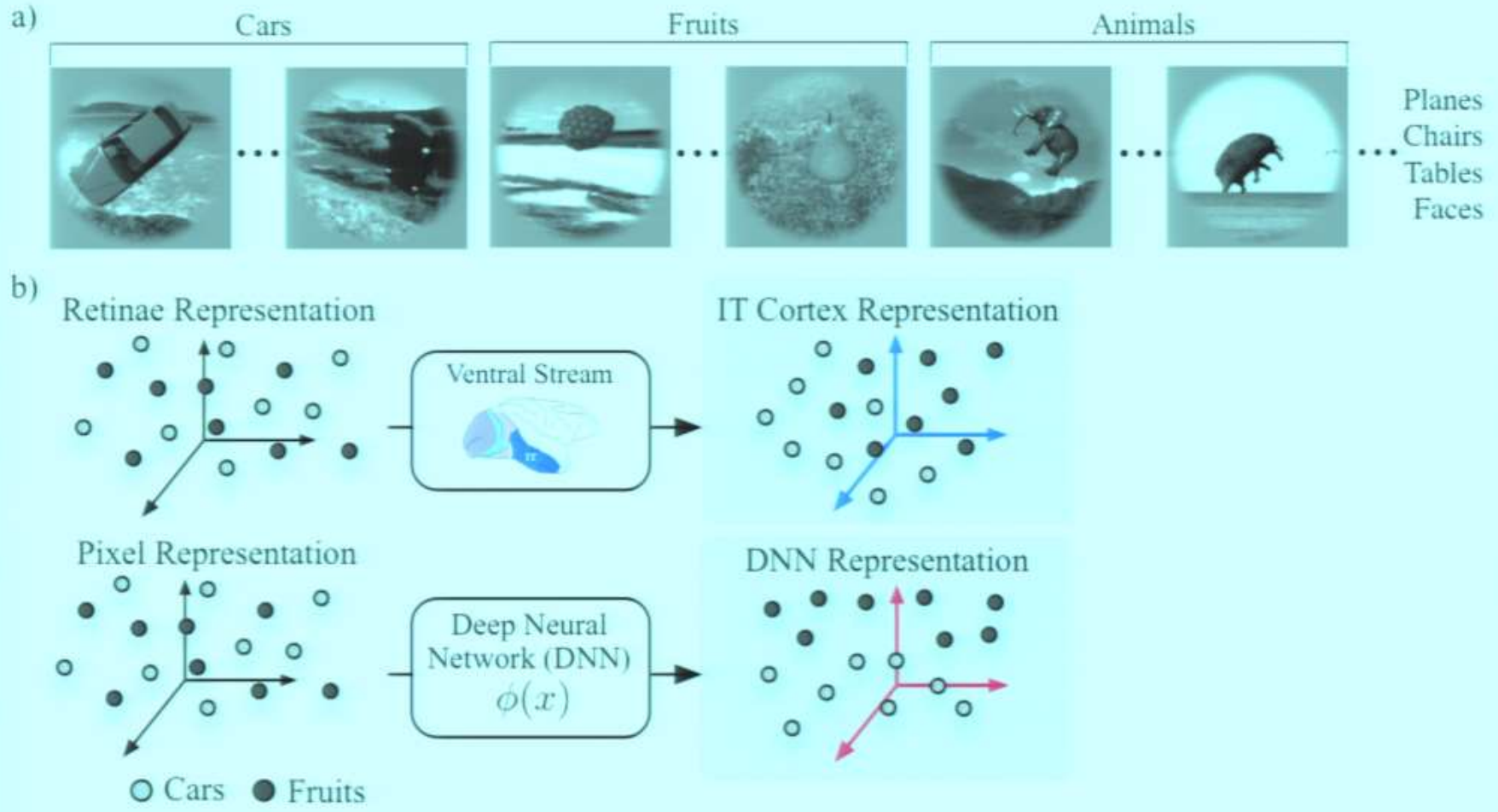
Object recognition 1.0 (HVM1.0)



Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Brain features vs. Machine features

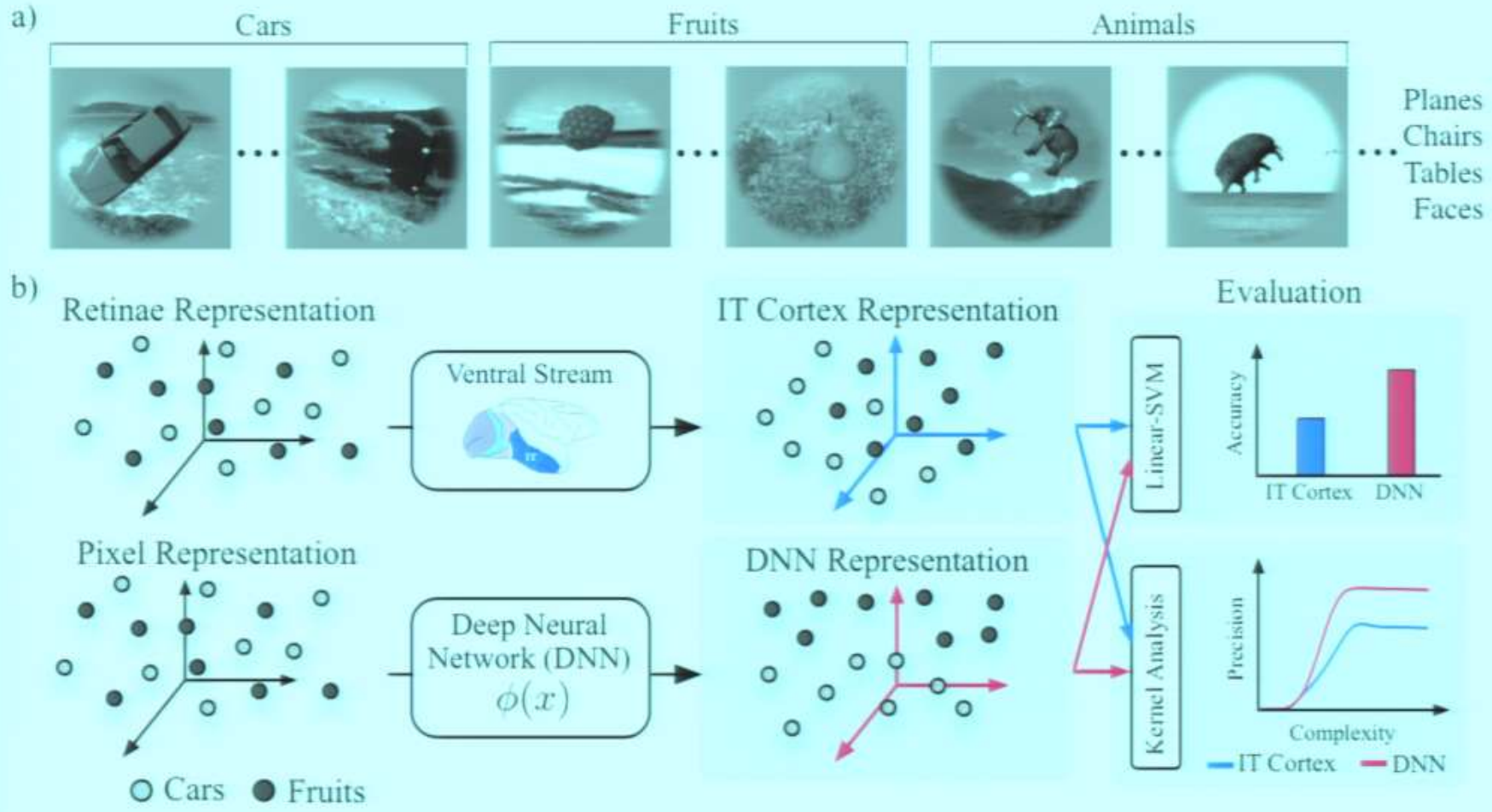
Object recognition 1.0 (HVM1.0)



Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

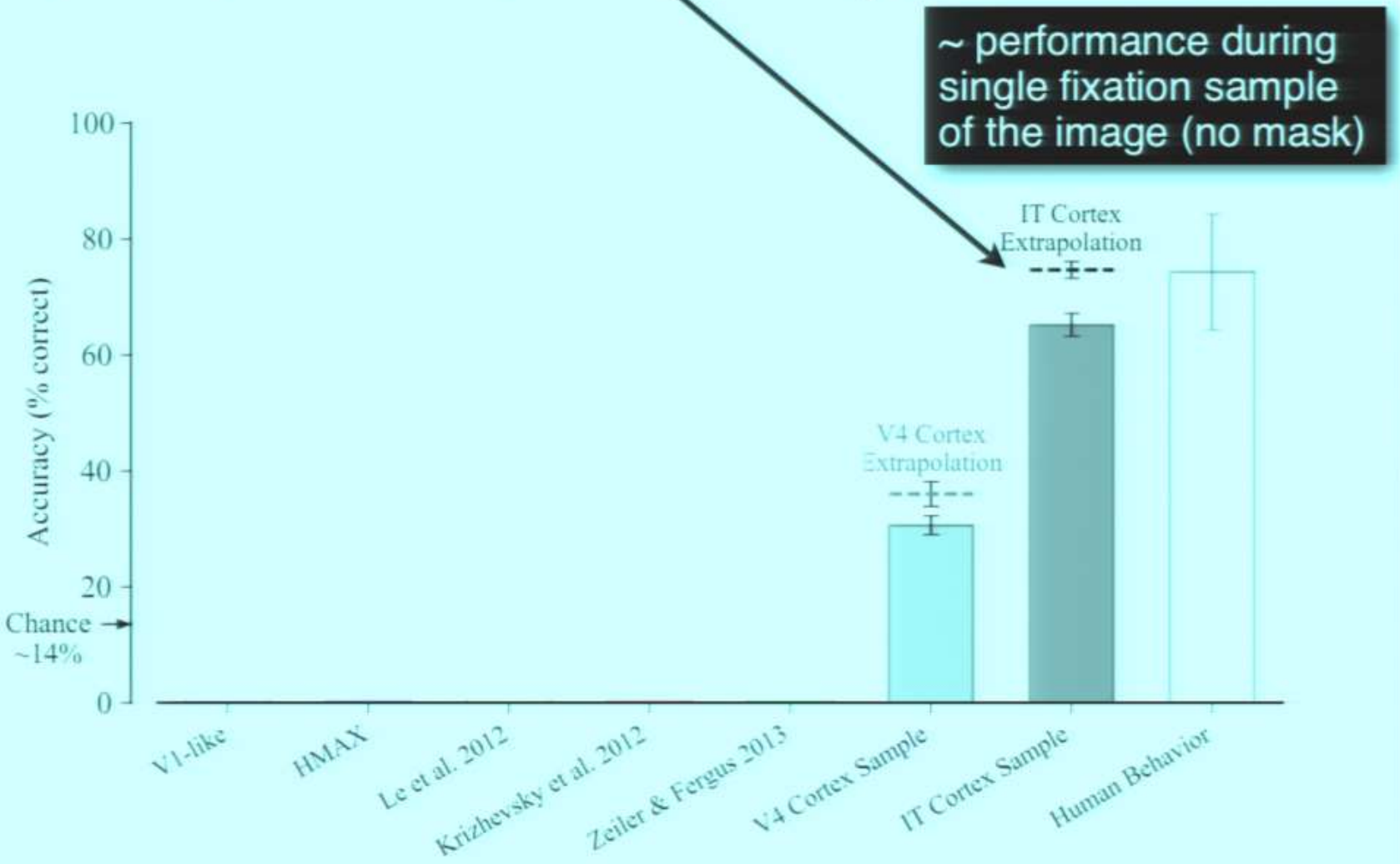
Brain features vs. Machine features

Object recognition 1.0 (HVM1.0)



Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Linear-SVM Generalization Performance of Machine and Neural Representations



Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

~200 ms “snapshot” samples

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

Central 10 deg, 200 ms “snapshot” samples

Core object recognition

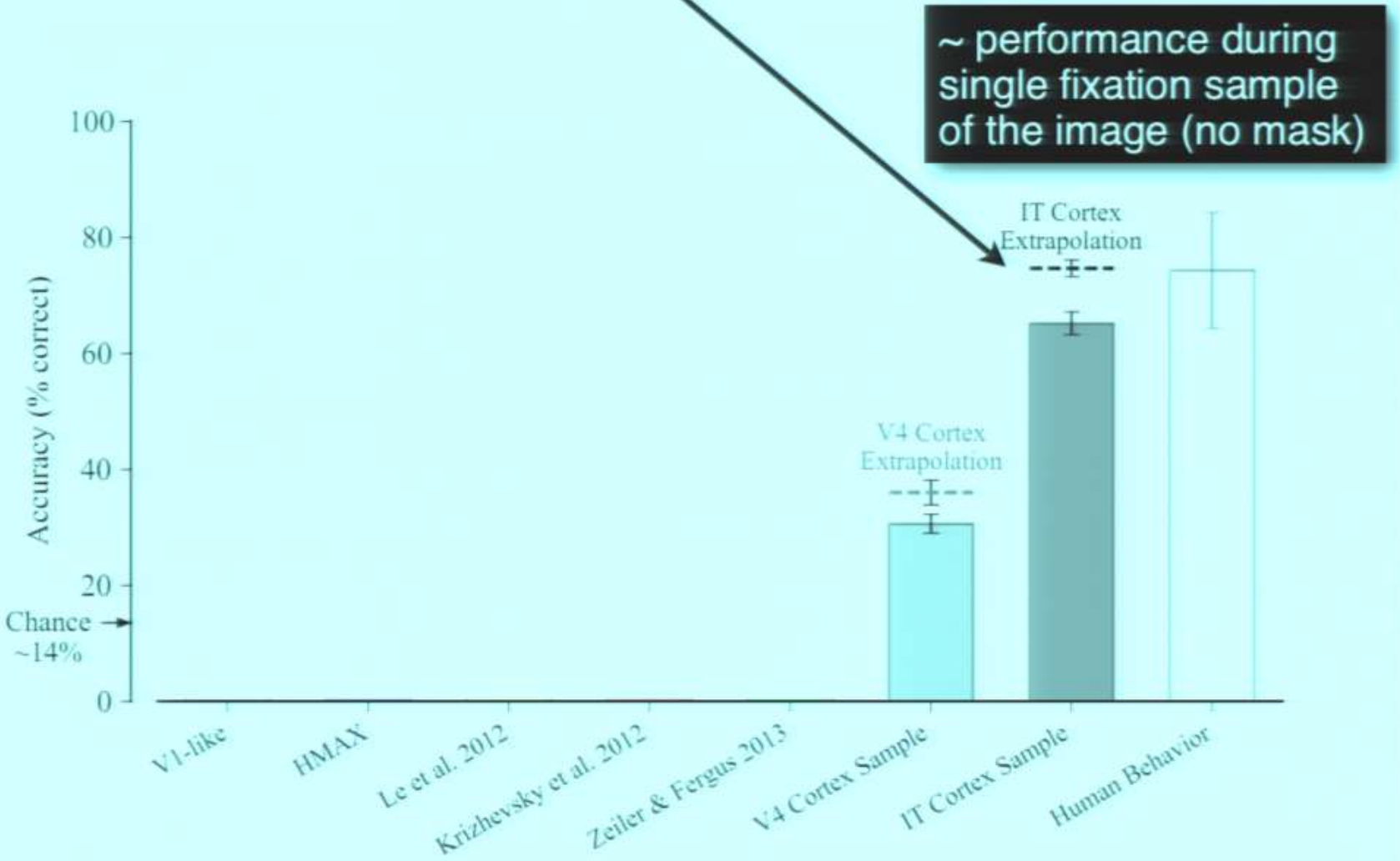
Central 10 deg, 200 ms “snapshot” samples

Core object recognition

Central 10 deg, 200 ms “snapshot” samples

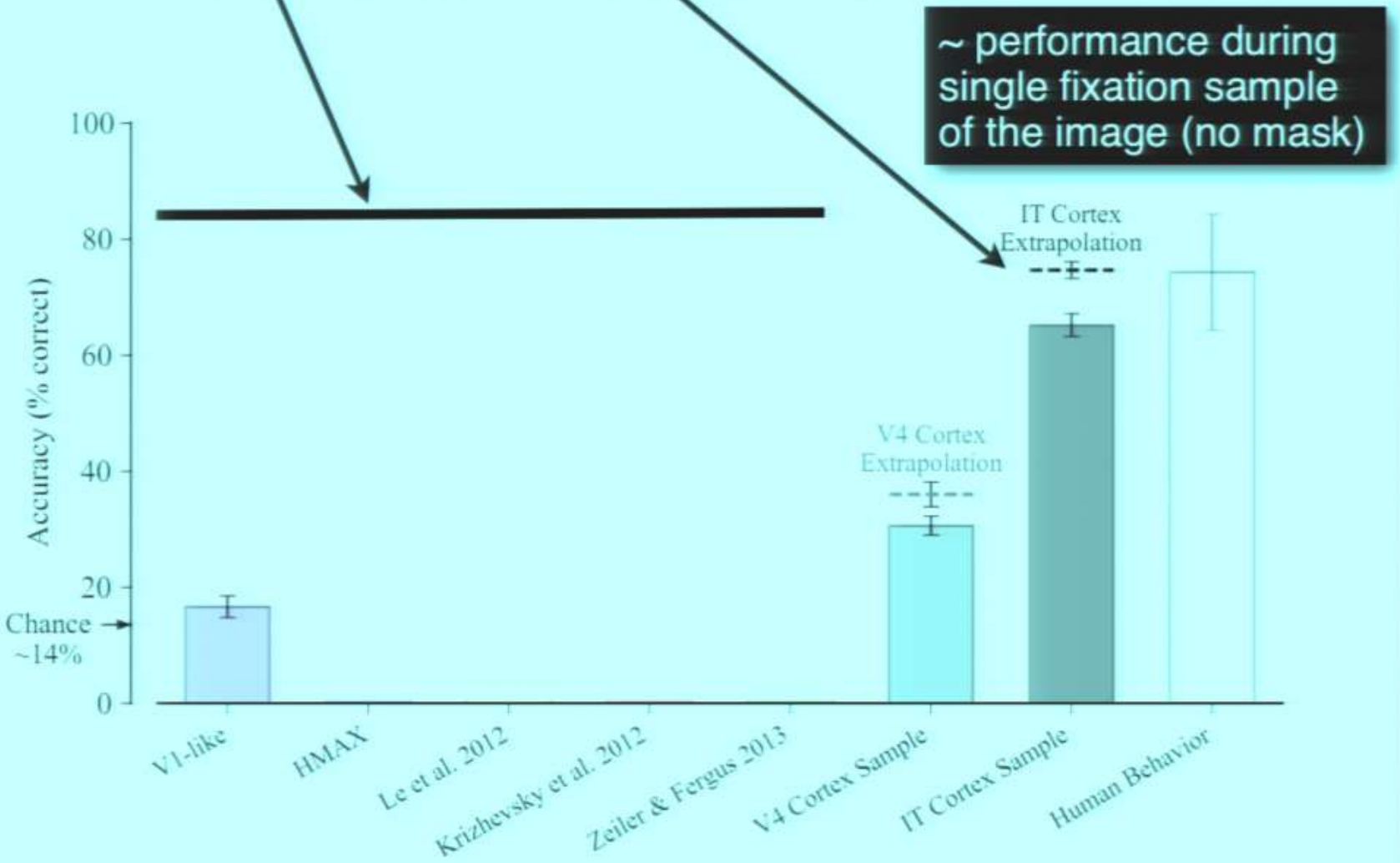
Core object recognition

Linear-SVM Generalization Performance of Machine and Neural Representations



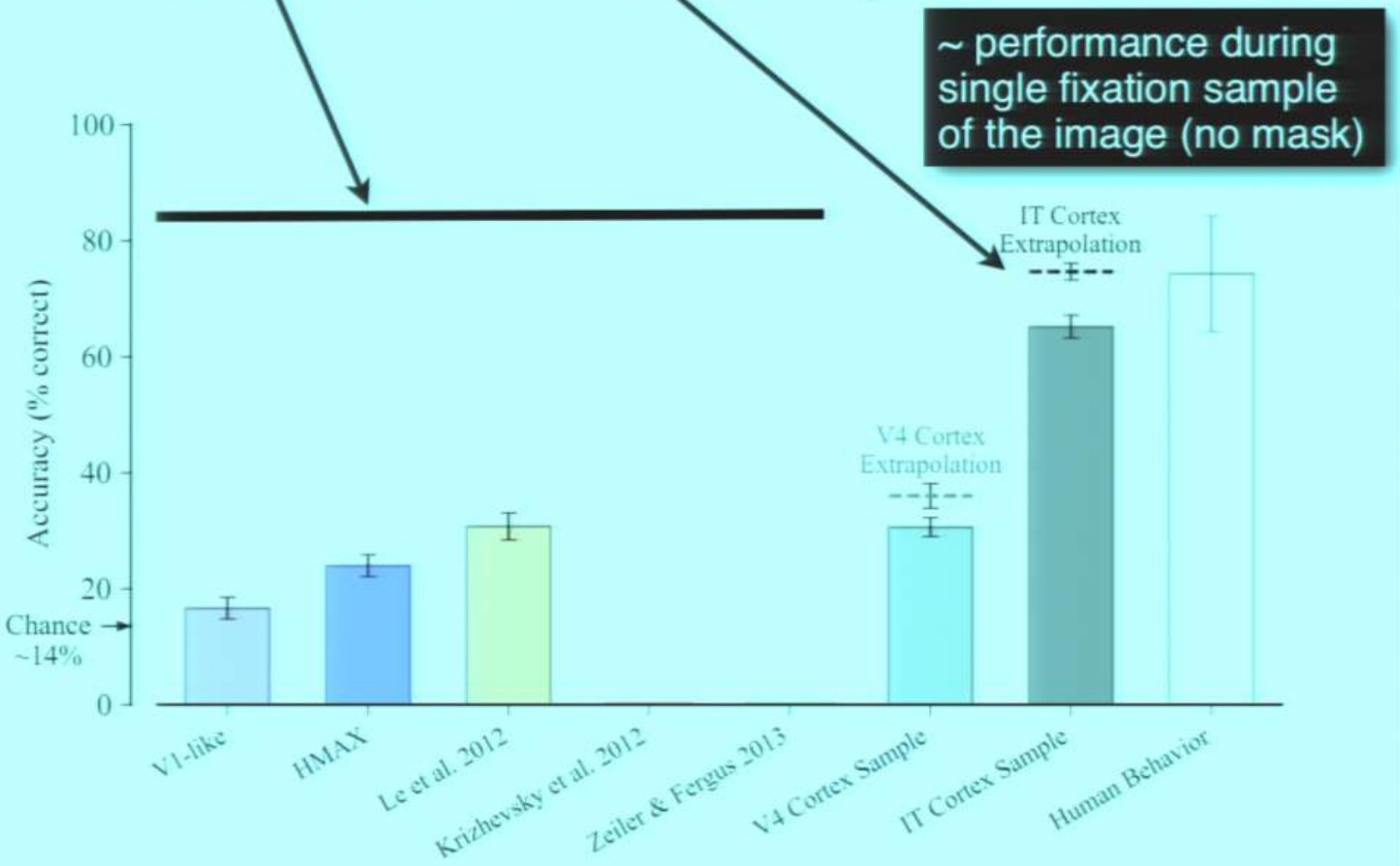
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Linear-SVM Generalization Performance of Machine and Neural Representations



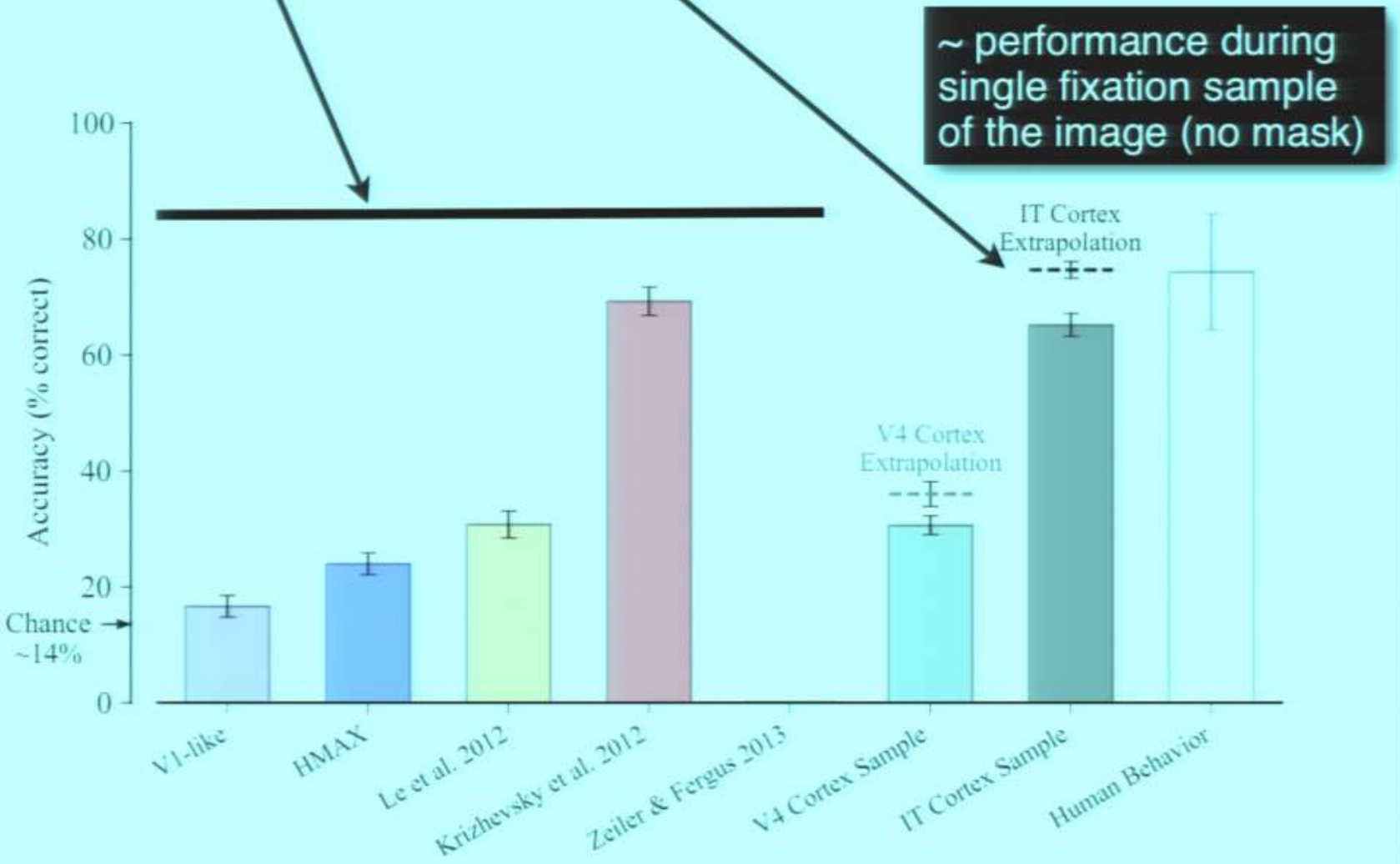
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Linear-SVM Generalization Performance of Machine and Neural Representations



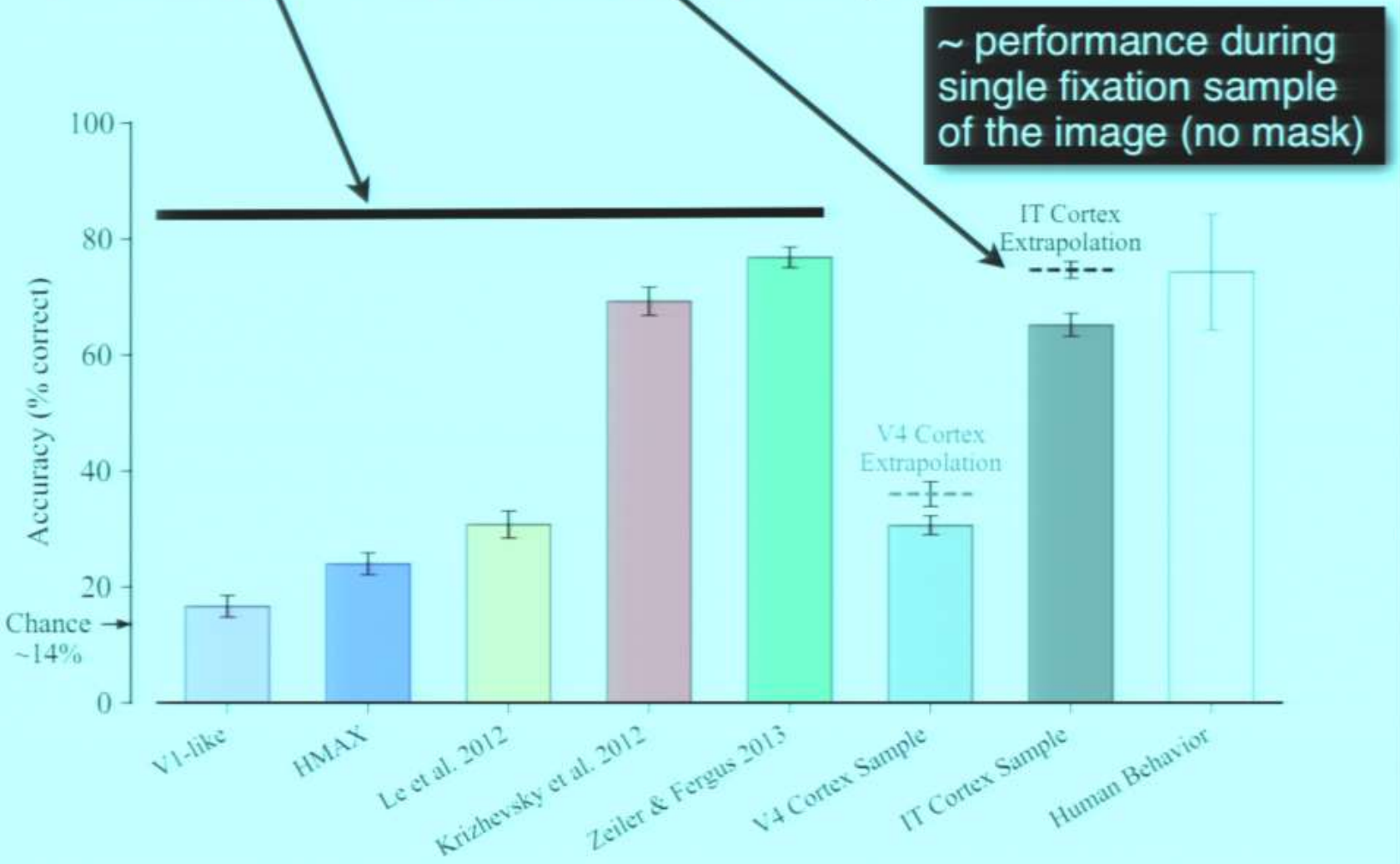
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Linear-SVM Generalization Performance of Machine and Neural Representations



Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

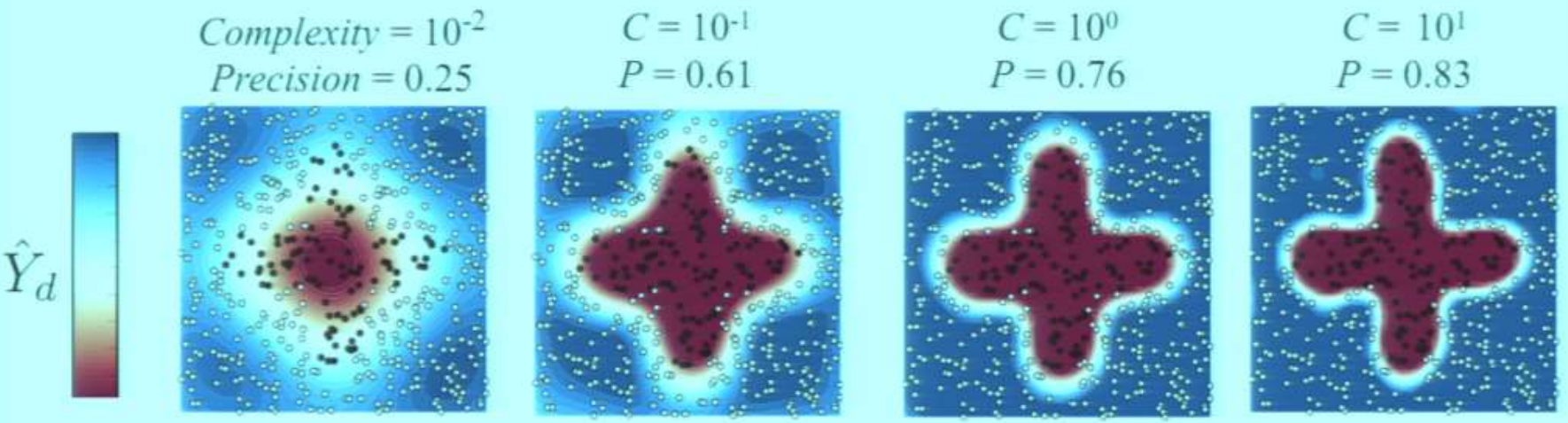
Linear-SVM Generalization Performance of Machine and Neural Representations



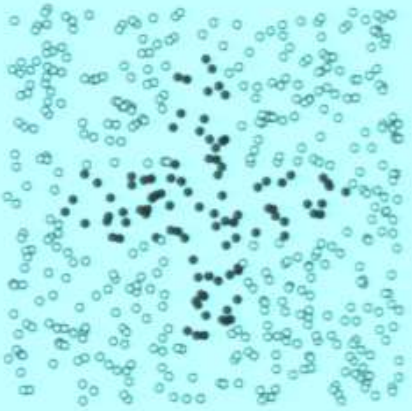
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

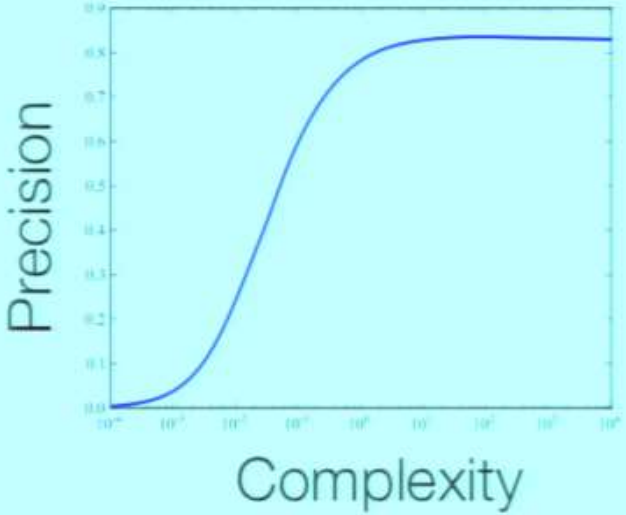
Demonstration of Kernel Analysis

Based on [Braun et al. 2008] and [Montavon et al. 2012]

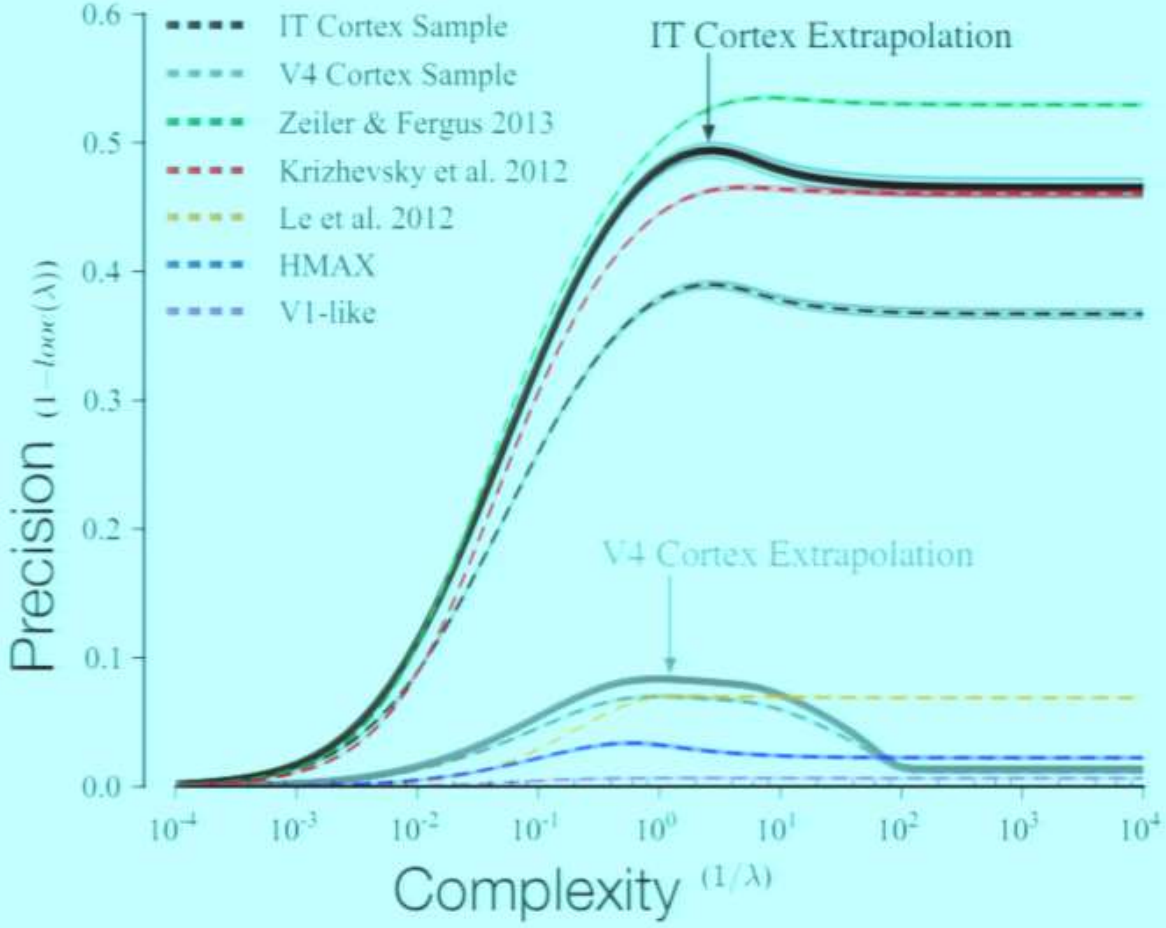


$$x \mapsto \phi(x)$$



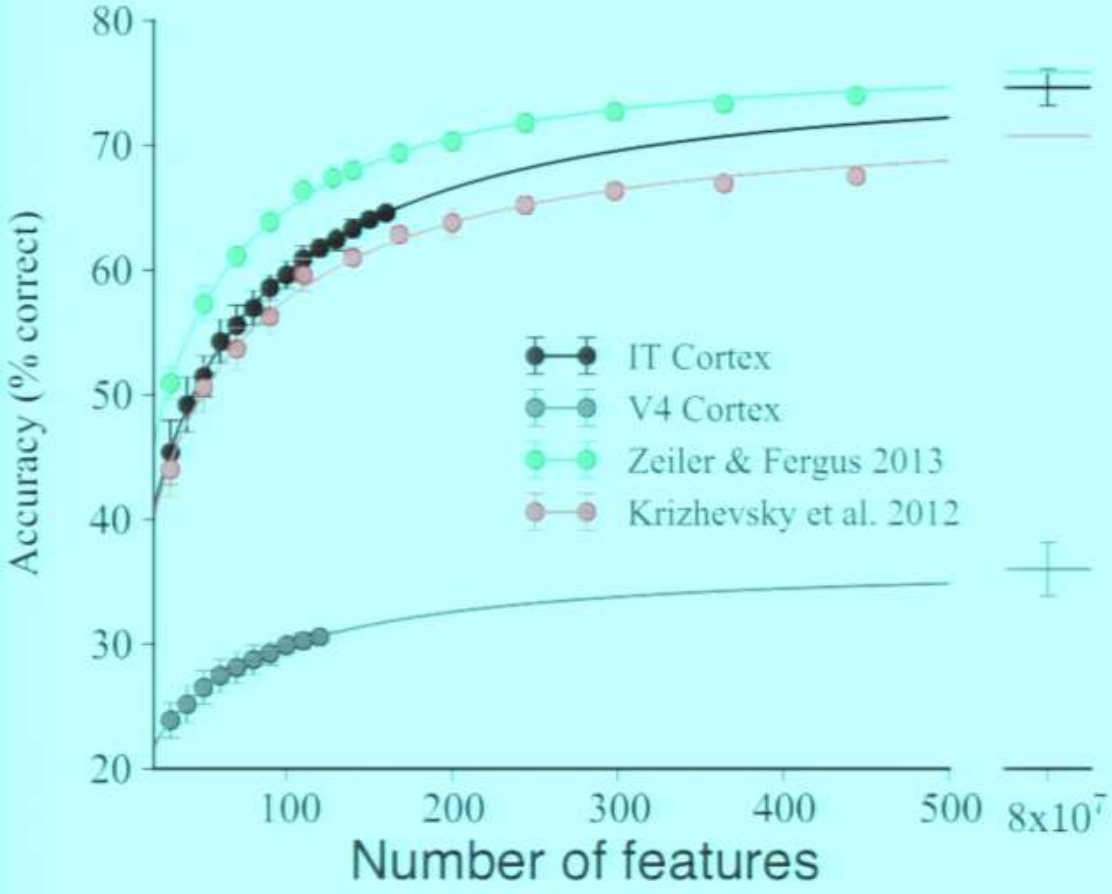


Kernel Analysis Curves of Neural and Machine Representations



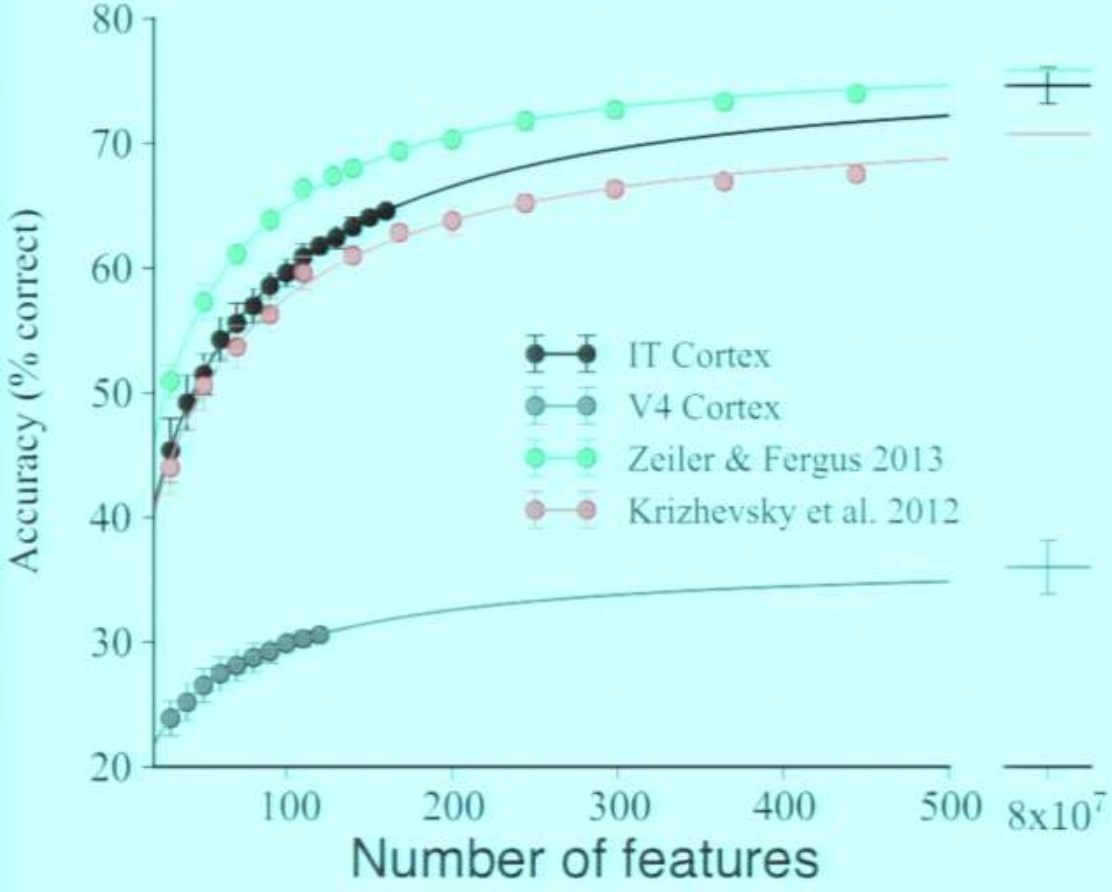
Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

These results hold, regardless of number of features



Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

These results hold, regardless of number of features



Upshot: the field now has at least three candidate hypotheses for the brain's ventral stream mechanisms.

SuperVision

Zeiler&Fergus

HMO

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Run the HVM1.0 benchmark:

http://dicarlolab.mit.edu/neuralbenchmark

Images: for each variation level

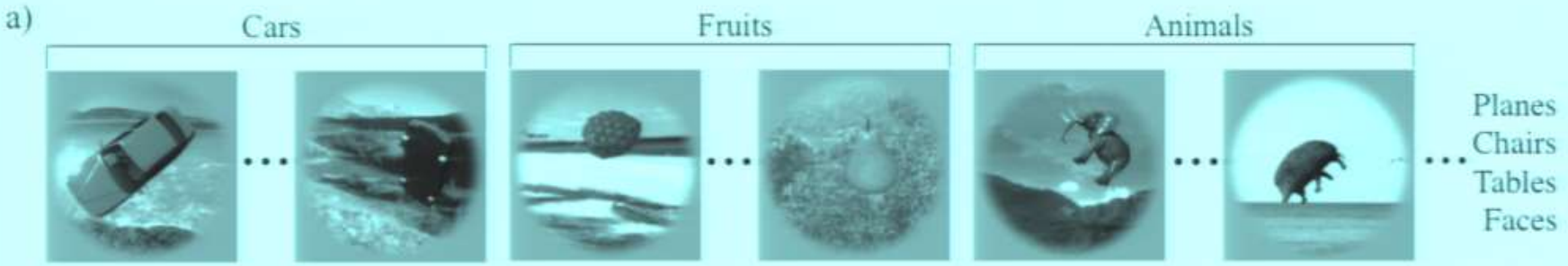
Code: to compute benchmark from your features

Training Set: Independent set to train algorithms

Cadieu CF, Hong H, Yamins D, Pinto N, Majaj N, and DiCarlo JJ. ICLR (2013)

Brain vs. Machine

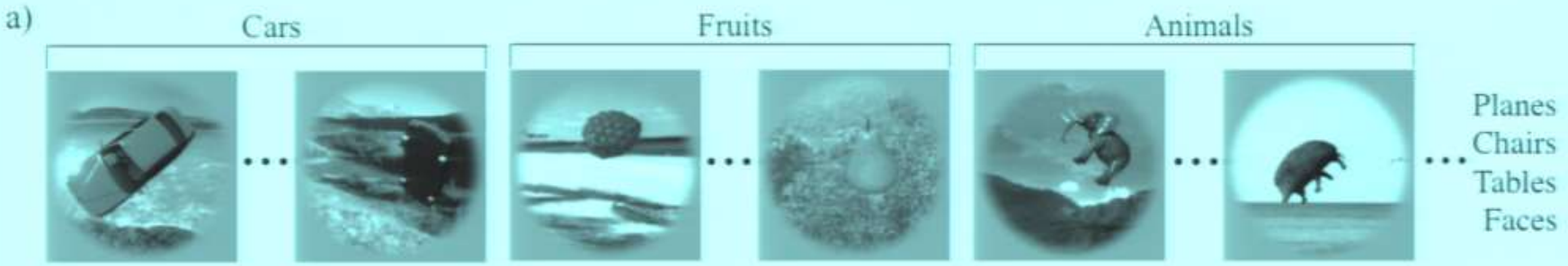
Object recognition 1.0 (HVM1.0)



Too early to declare victory

Brain vs. Machine

Object recognition 1.0 (HVM1.0)



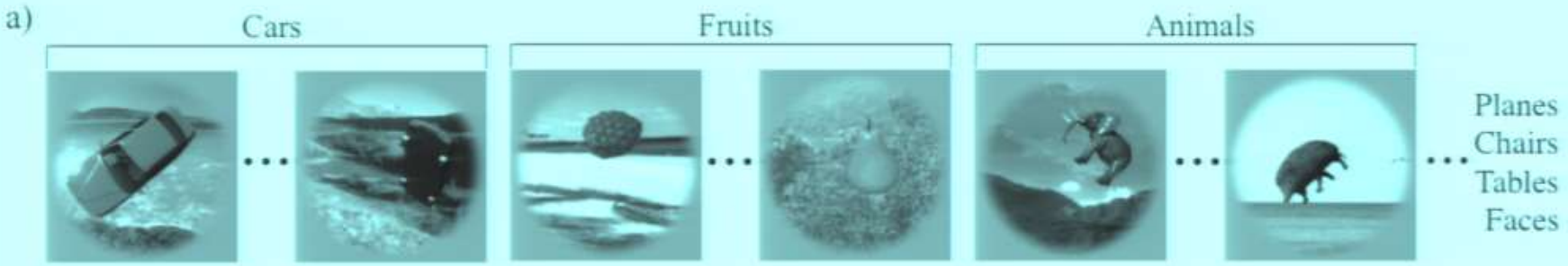
Too early to declare victory

Do models satisfy more stringent predictions on these images?

(e.g. image-by-image patterns of confusion?)

Brain vs. Machine

Object recognition 1.0 (HVM1.0)



Too early to declare victory

Do models satisfy more stringent predictions on these images?

(e.g. image-by-image patterns of confusion?)

Test other task challenges!

E.g. occlusion, illumination, ...

Object recognition 2.0 (HVM2.0)

Comparisons I will present today:

1. Monkey neurons vs. Human Behavior

   Suggests that IT population codes are one simple step from object recognition behavior

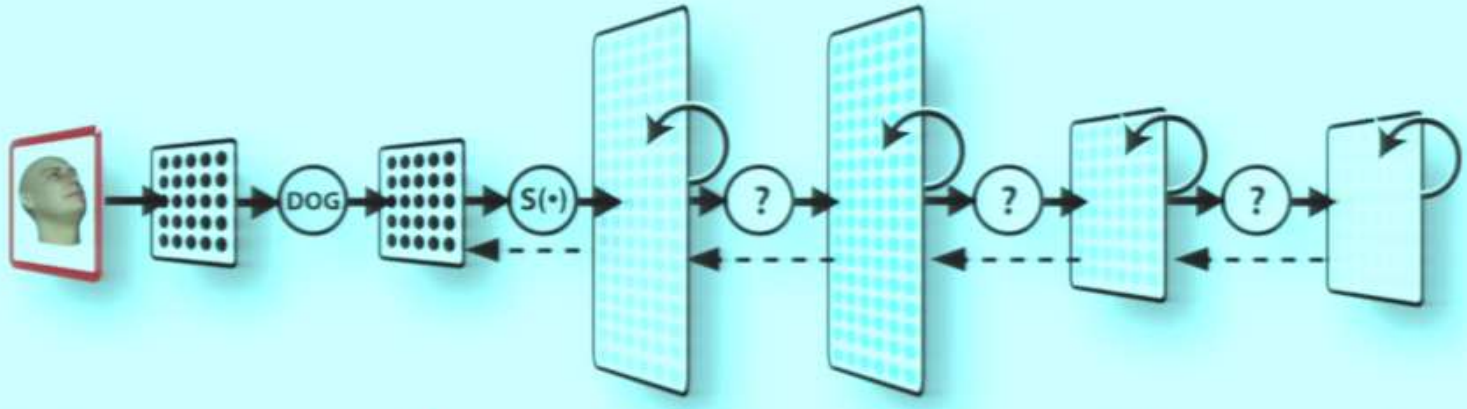
2. Machines vs. Monkey neurons

   Shows that a model focus on the behavioral goal leads to a potential understanding of underlying brain mechanisms.

3. Machines vs. Monkey neurons/Human behavior

   Demonstrates the recent bio-inspired models rival the brain in object recognition

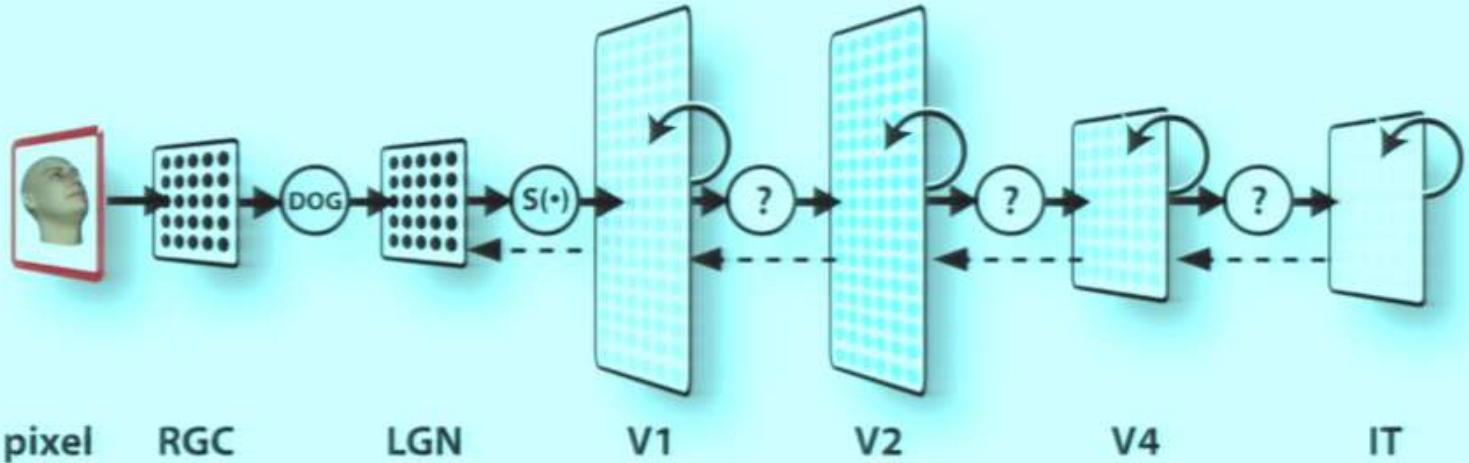
Take home messages



Take home messages

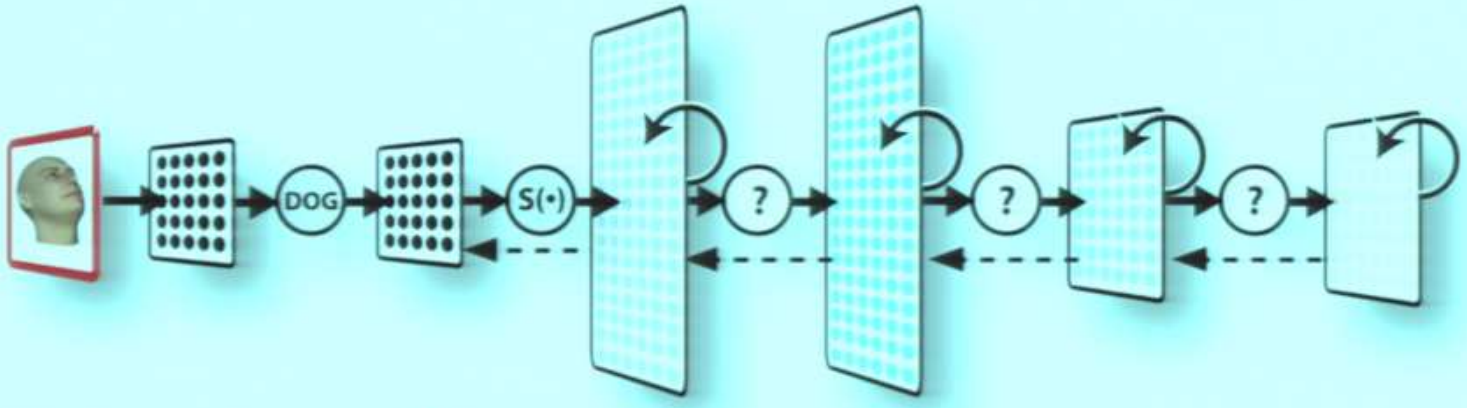
- *Invariance is the crux computational problem*

- ***"Simple" IT population rate codes are sufficient to account for unfettered human object recognition (HVM 1.0).*** *Testing monkey behavior & sharpening tests.*



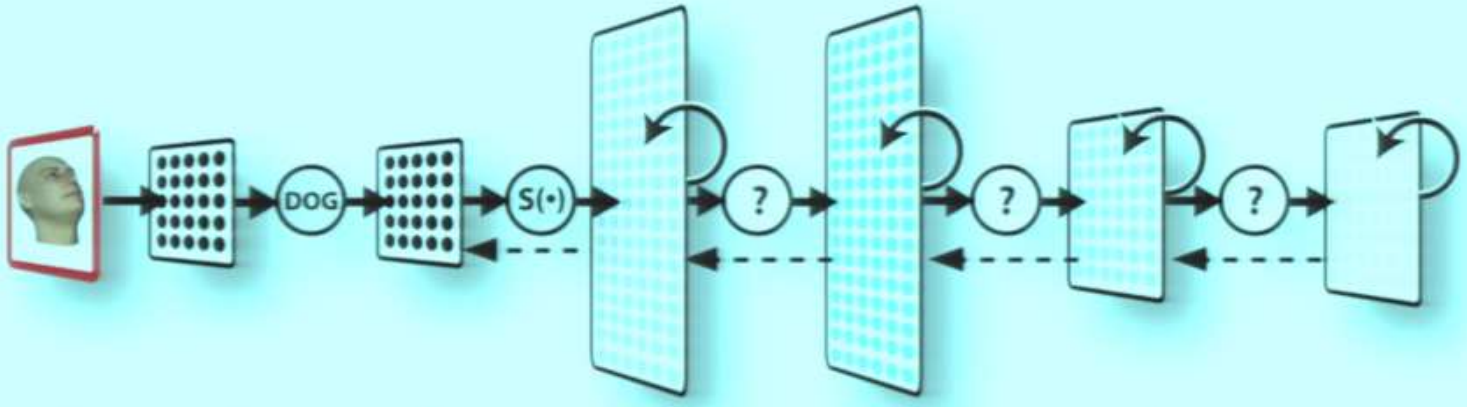
Take home messages

- *Invariance is the crux computational problem*
- ***"Simple" IT population rate codes are sufficient to account for unfettered human object recognition (HVM 1.0).*** *Testing monkey behavior & sharpening tests.*
- ***These IT codes are computed "reflexively" in ~100 ms, and are likely shared by human and non-human primates***



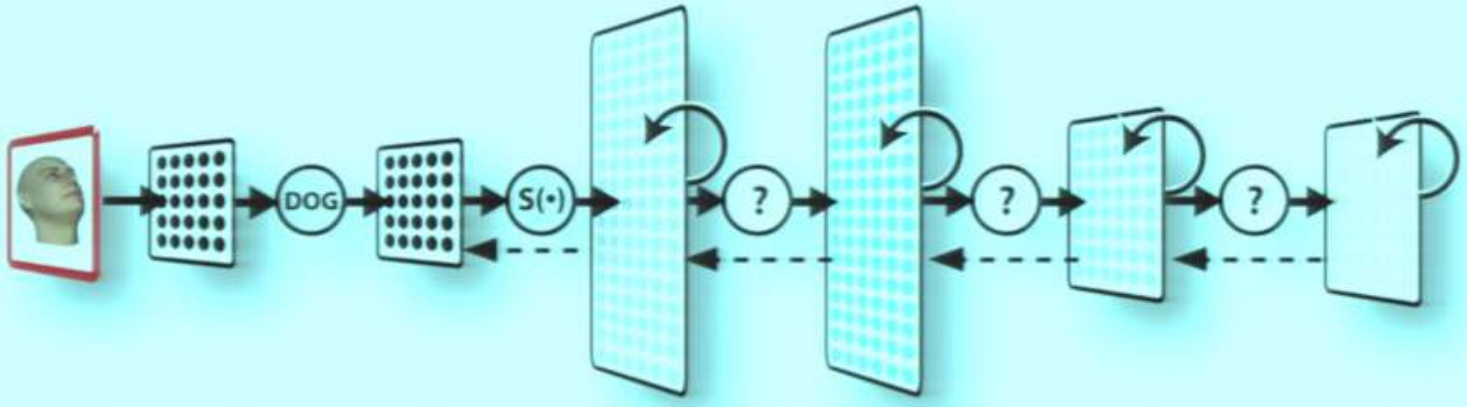
Take home messages

- *Invariance is the crux computational problem*
- ***"Simple" IT population rate codes are sufficient to account for unfettered human object recognition (HVM 1.0).*** *Testing monkey behavior & sharpening tests.*
- ***These IT codes are computed "reflexively" in ~100 ms, and are likely shared by human and non-human primates***
- *The key transformations live between V1 and IT*
- ***We have been searching a large class of bio-constrained models. High performing models can accurately predict IT neuronal responses, and their intermediate layers predict V4 responses.***



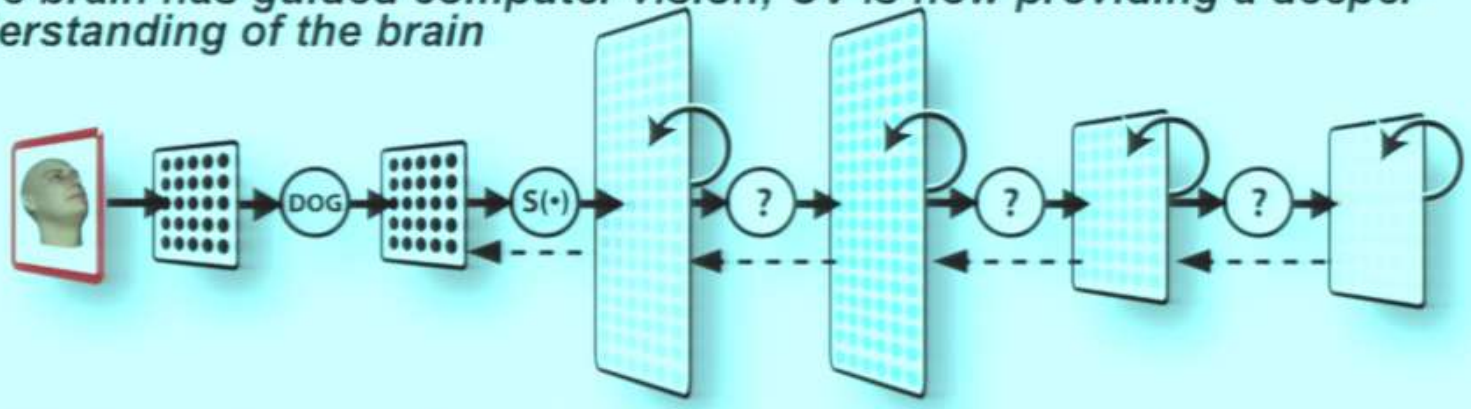
Take home messages

- *Invariance is the crux computational problem*
- ***"Simple" IT population rate codes are sufficient to account for unfettered human object recognition (HVM 1.0).*** *Testing monkey behavior & sharpening tests.*
- ***These IT codes are computed "reflexively" in ~100 ms, and are likely shared by human and non-human primates***
- *The key transformations live between V1 and IT*
- ***We have been searching a large class of bio-constrained models. High performing models can accurately predict IT neuronal responses, and their intermediate layers predict V4 responses.***
- *Other artificial deep conv networks are now rivaling neural and human performance on our (HVM1.0) benchmarks --> viable hypotheses for ventral stream mechanisms.*



Take home messages

- *Invariance is the crux computational problem*
- ***"Simple" IT population rate codes are sufficient to account for unfettered human object recognition (HVM 1.0).*** *Testing monkey behavior & sharpening tests.*
- ***These IT codes are computed "reflexively" in ~100 ms, and are likely shared by human and non-human primates***
- *The key transformations live between V1 and IT*
- ***We have been searching a large class of bio-constrained models. High performing models can accurately predict IT neuronal responses, and their intermediate layers predict V4 responses.***
- *Other artificial deep conv networks are now rivaling neural and human performance on our (HVM1.0) benchmarks --> viable hypotheses for ventral stream mechanisms.*
- ***The brain has guided computer vision; CV is now providing a deeper understanding of the brain***



Acknowledgements

Current lab members:

Arash Afraz
Paul Aparicio
Diego Ardila
Charles Cadieu
Ha Hong
Elias Issa
Xiaoxuan Jia
Rishi Rajalingham
Kailyn Schmidt
Darren Seibert
Chris Stawarz
Dan Yamins

Alumni:

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

Key contributing labs:

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

- NIH NEI
- NSF
- DARPA / ONR
- McGovern Institute