Untangling object recognition in the ventral visual processing stream

James DiCarlo

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

brain+cognitive sciences

Visual object and face recognition

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

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

Goals - Specific

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

Courtesy of Mike Tarr

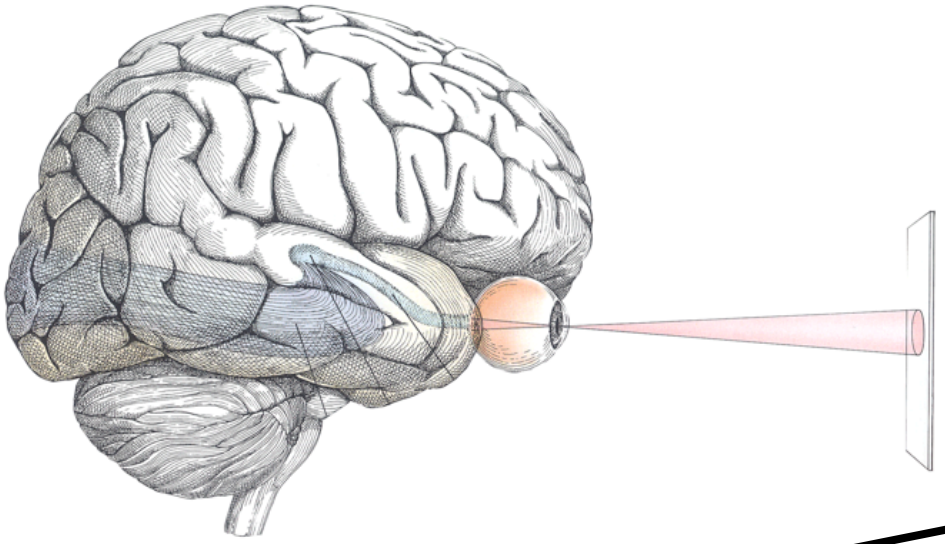
Visual object and face recognition

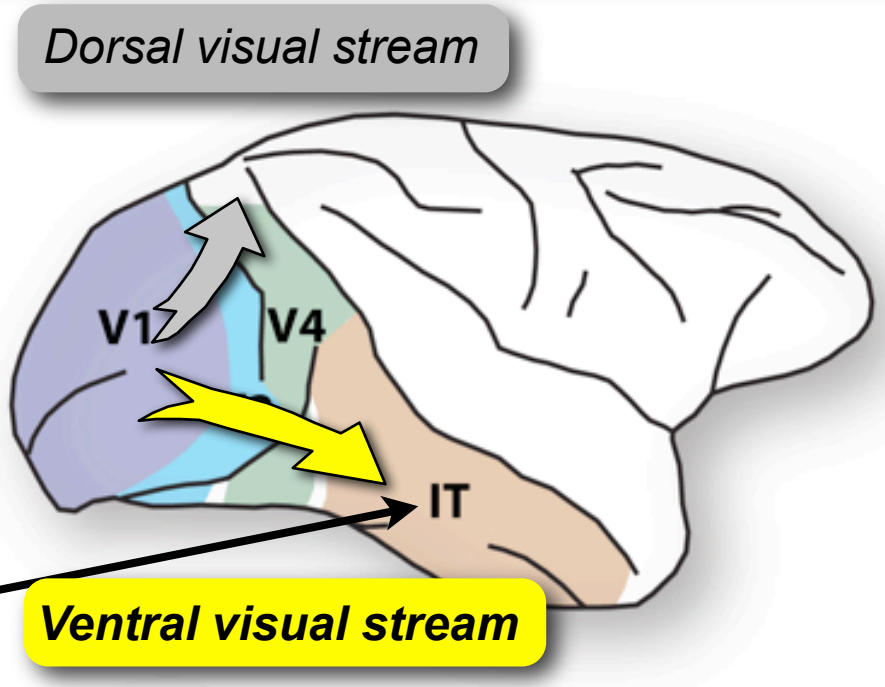
- *Accurate*
- *Fast*
- *Tolerant to variation*
- *Effortless*
- *Critical to survival*

Our mission: Understand how the brain constructs a neuronal representation that underlies object recognition

Focus the problem: core object recognition

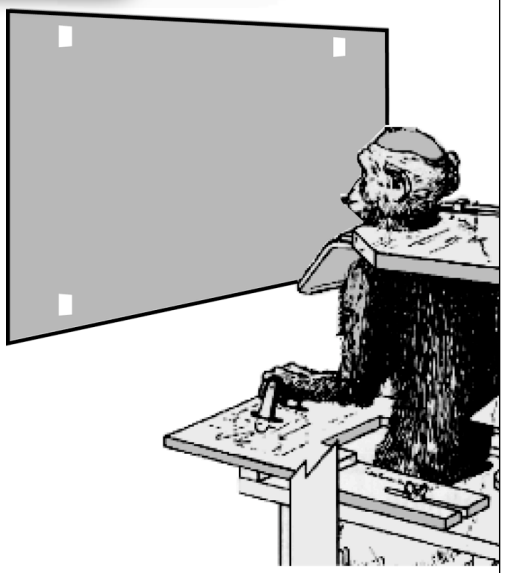
The non human primate model





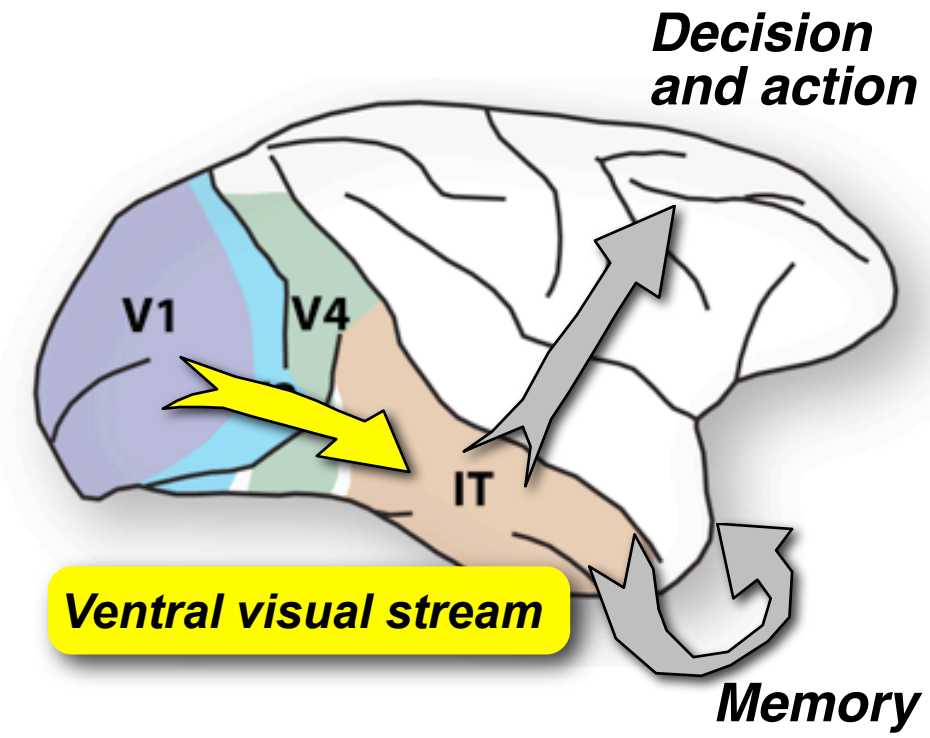
We know where that image representation lives in the primate (Inferior temporal cortex, IT).

We can study that representation (and its precursors) at the level of neuronal spikes.

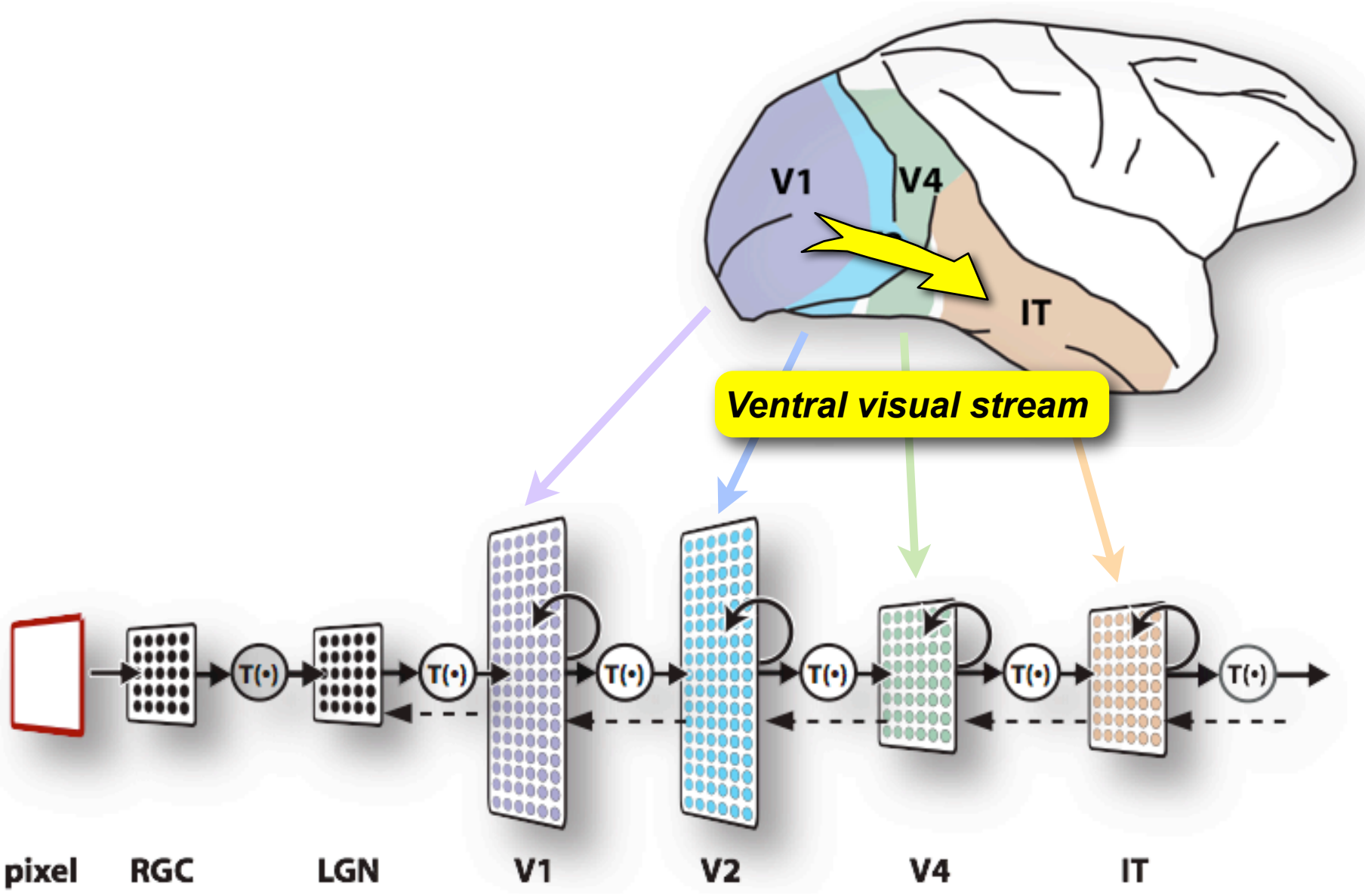


Adapted from Motter and Mountcastle 1981

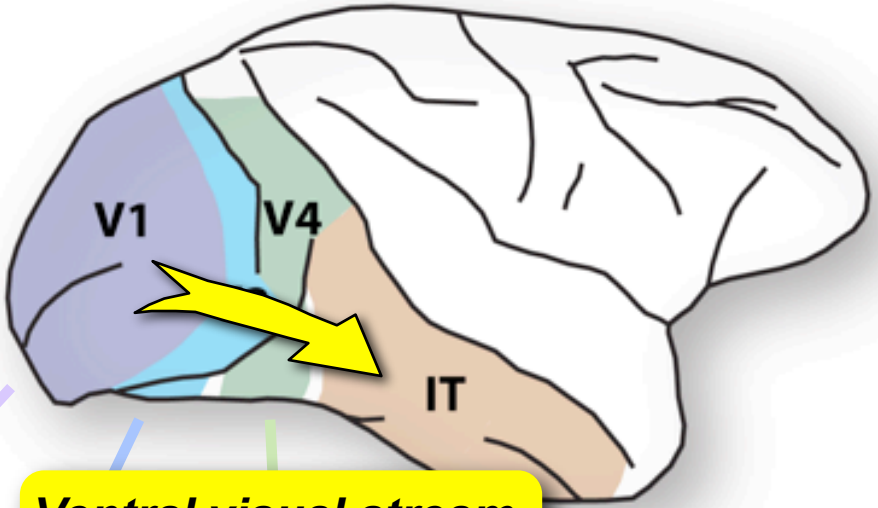
The ventral visual stream



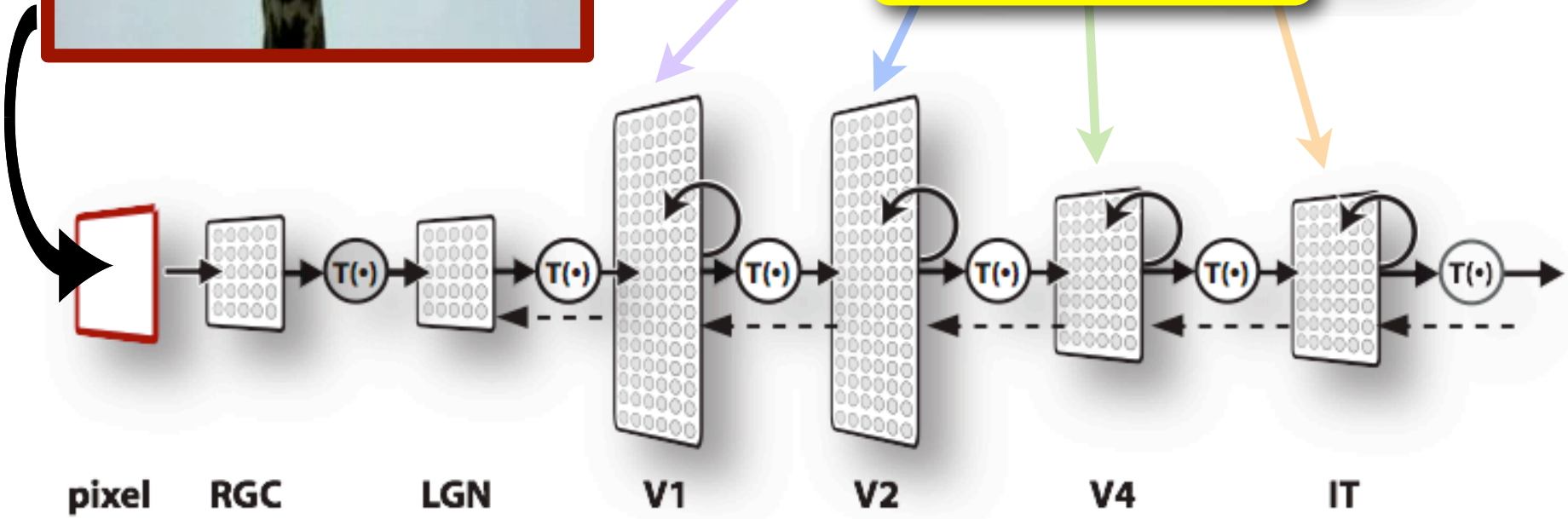
The ventral visual stream



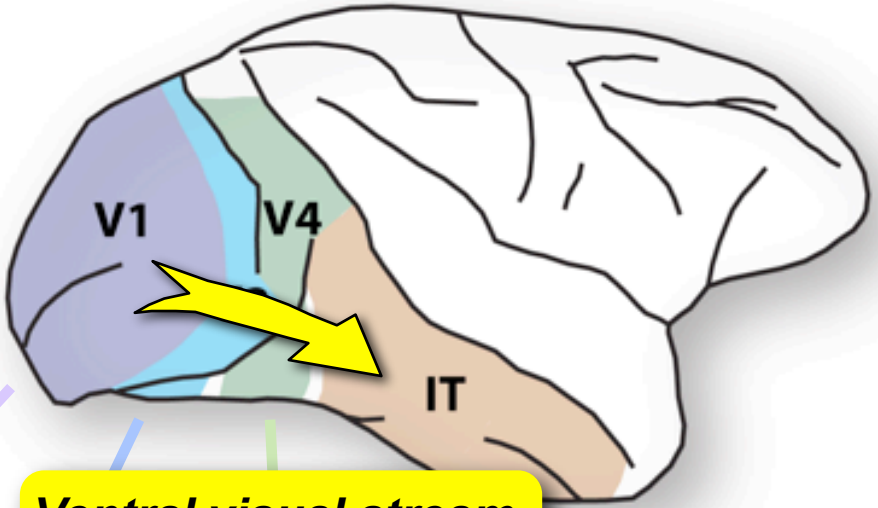
The ventral visual stream



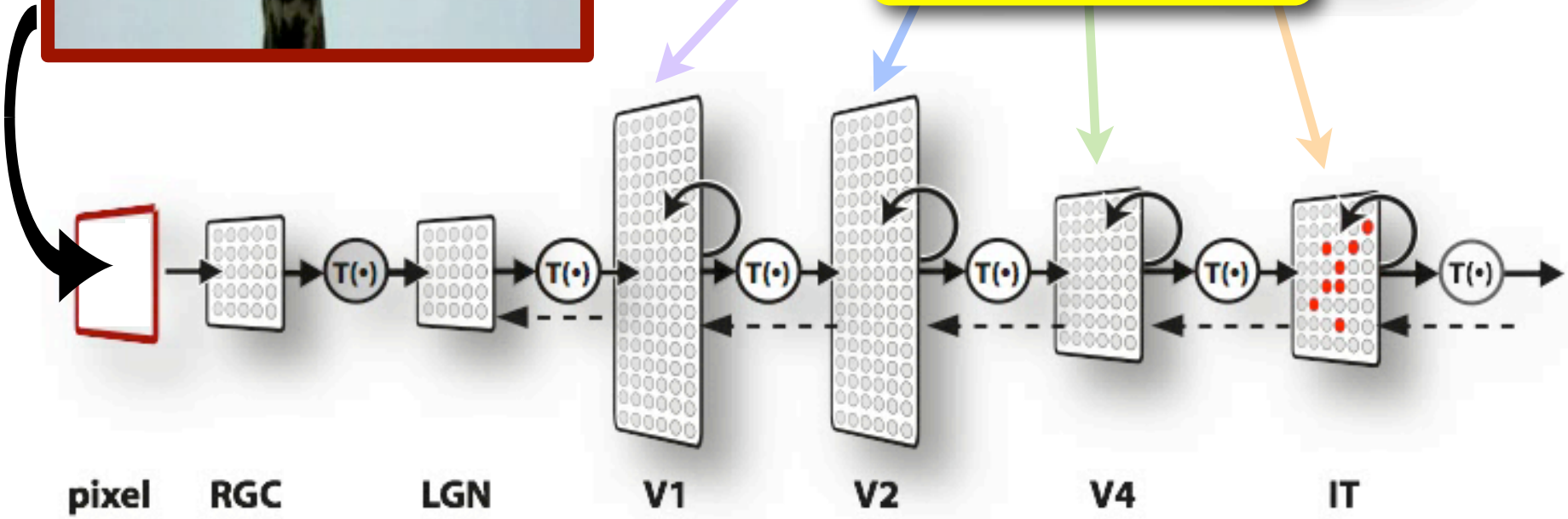
Ventral visual stream



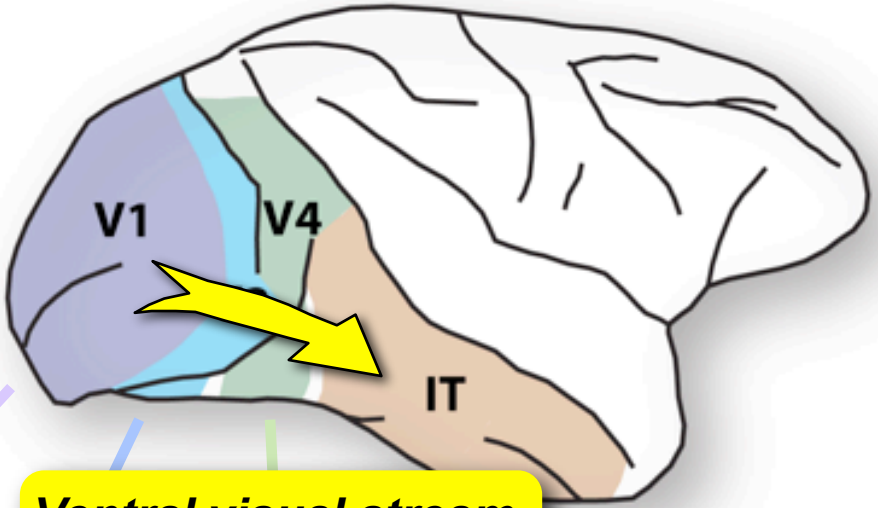
The ventral visual stream



Ventral visual stream



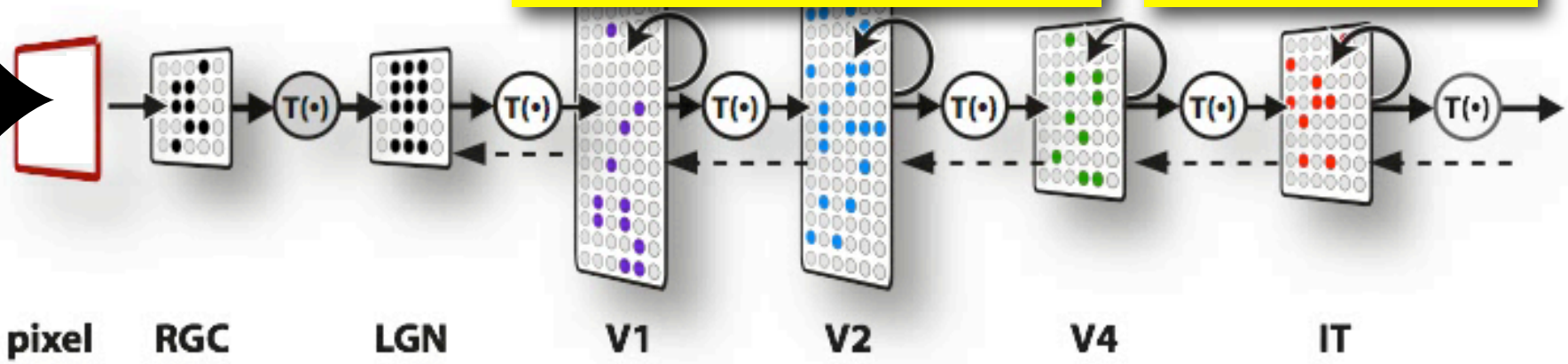
The ventral visual stream



Ventral visual stream

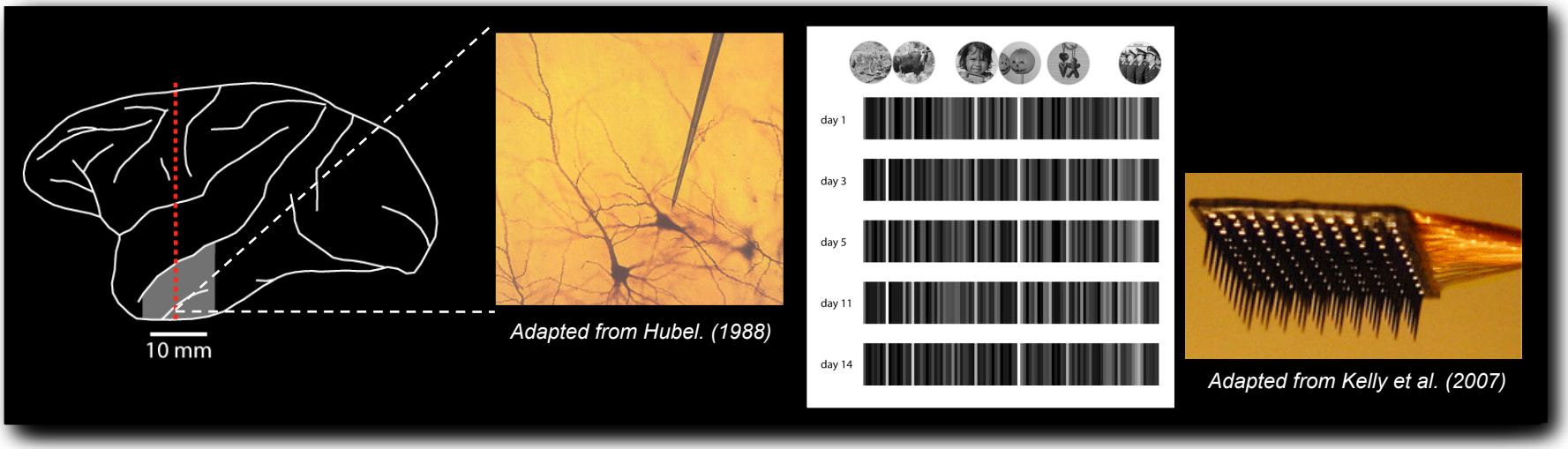
How is that representation constructed ???

Nature of this representation ?



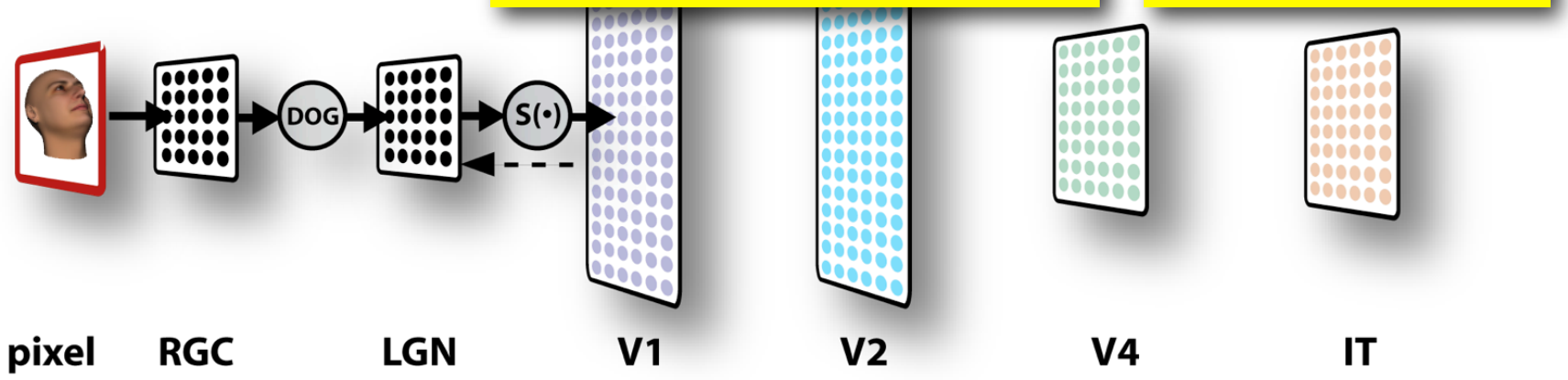
Our primary tools to attack this problem

Behavior ***Neurophysiology*** ***Computation*** ***Imaging***

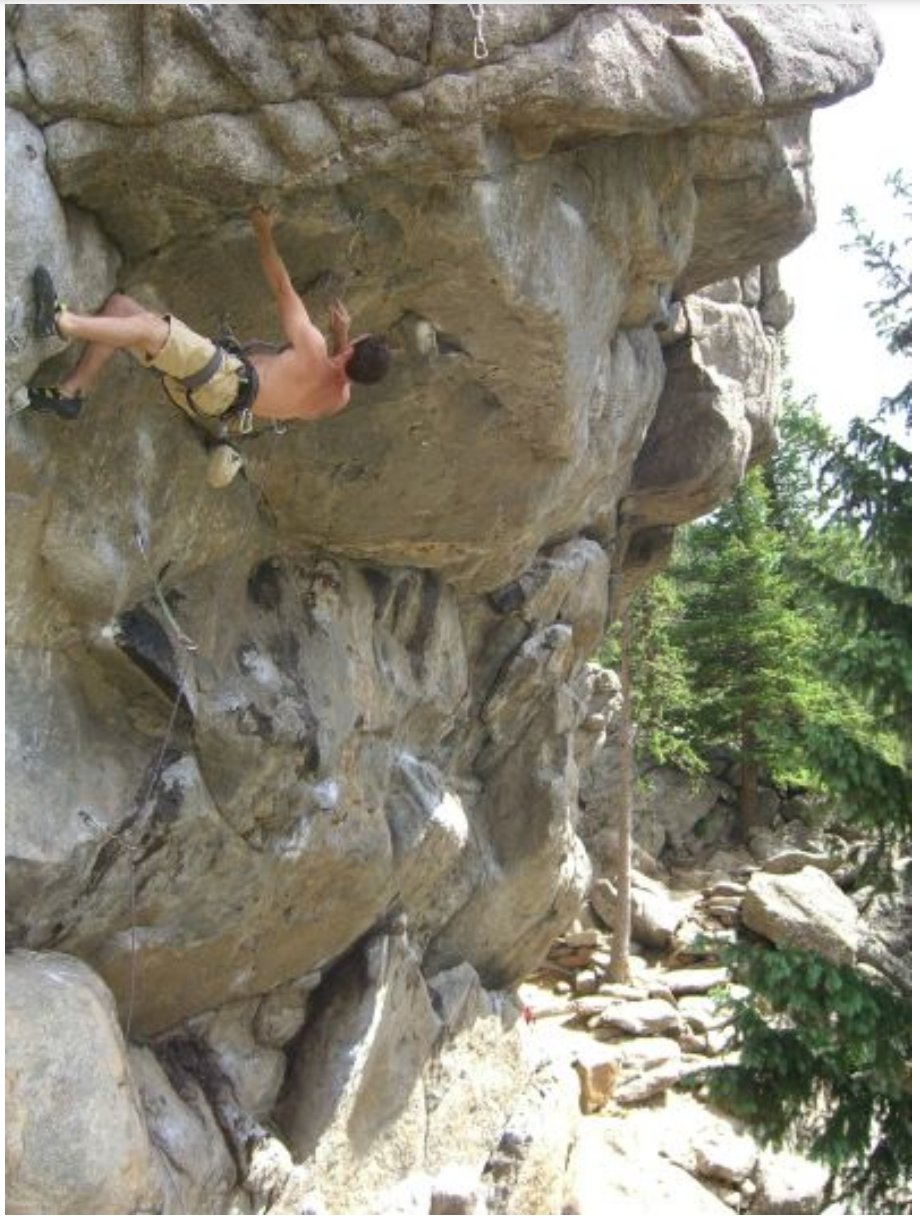


How is that representation constructed ???

Nature of this representation ?



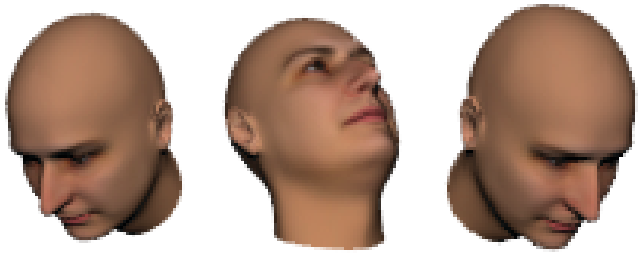
The computational crux of object and face recognition



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

The computational crux of object and face recognition



- ***Position***
- ***Size***
- ***Pose***
- ***Illumination***
- ***Clutter***
  - *Other objects*
  - *Background scene*

Natural variability

How does the brain recognize each object across this wide range of natural variability?

*It must create an **image representation** that is selective for object identity, yet tolerant ("invariant") to this variability.*

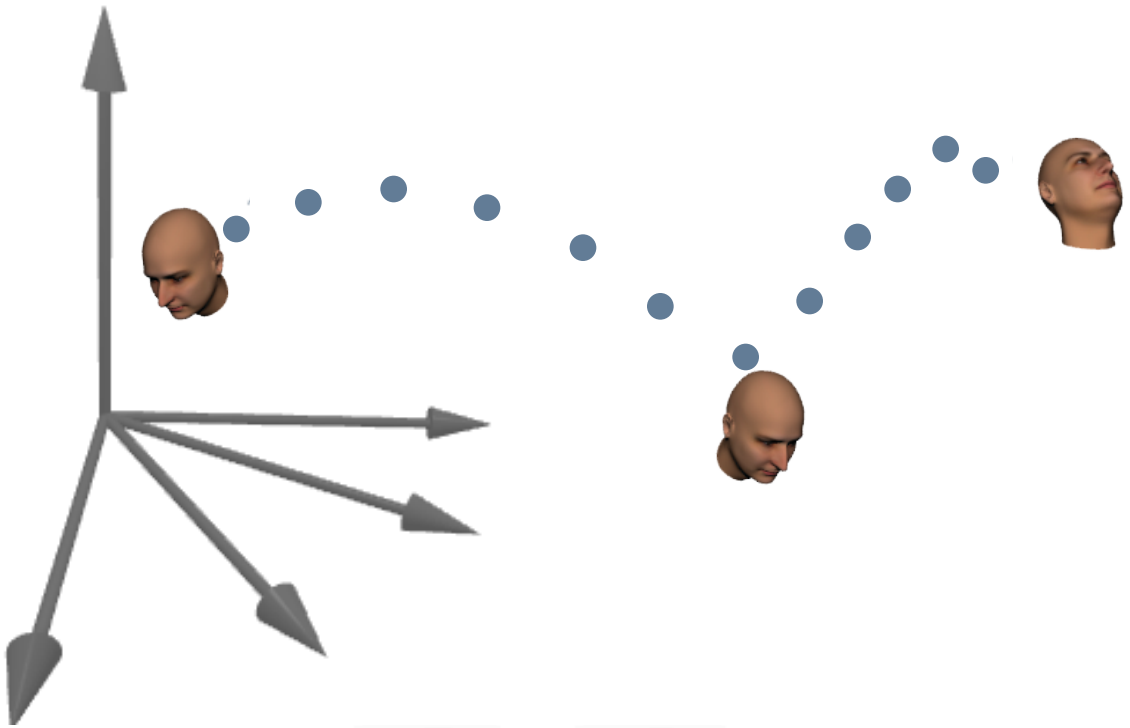
Explicit object representation

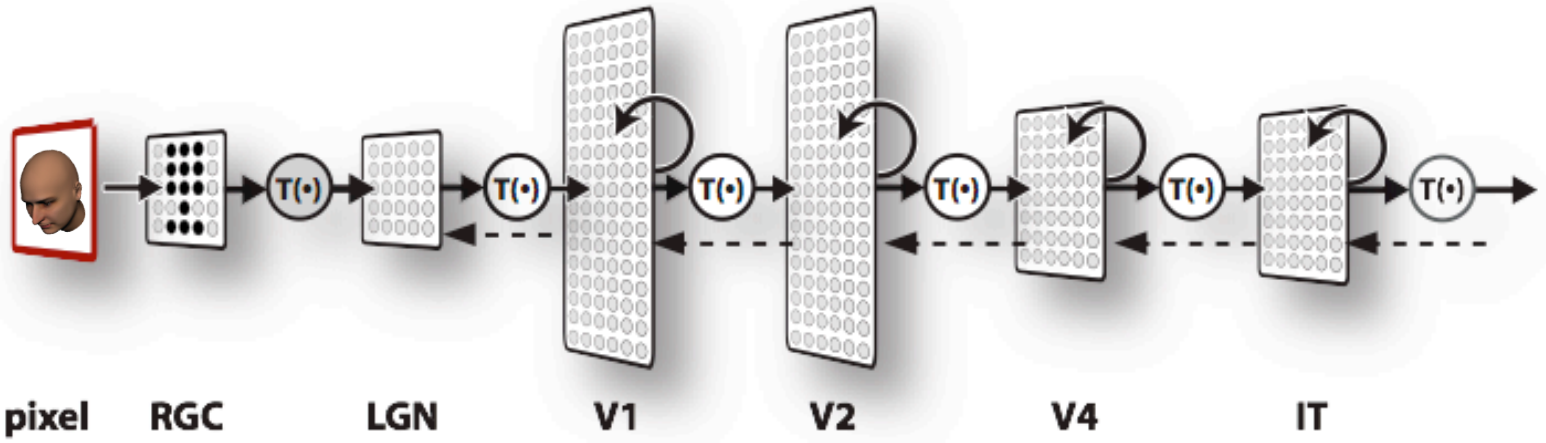
Poggio, Ullman, Grossberg, Edleman, Biederman, etc.
*DiCarlo and Cox, **TICS** (2007); Pinto, Cox, and DiCarlo, **PLoS Comp Bio** (2008)*

The computational crux of object and face recognition

an object identity manifold

Any population representation

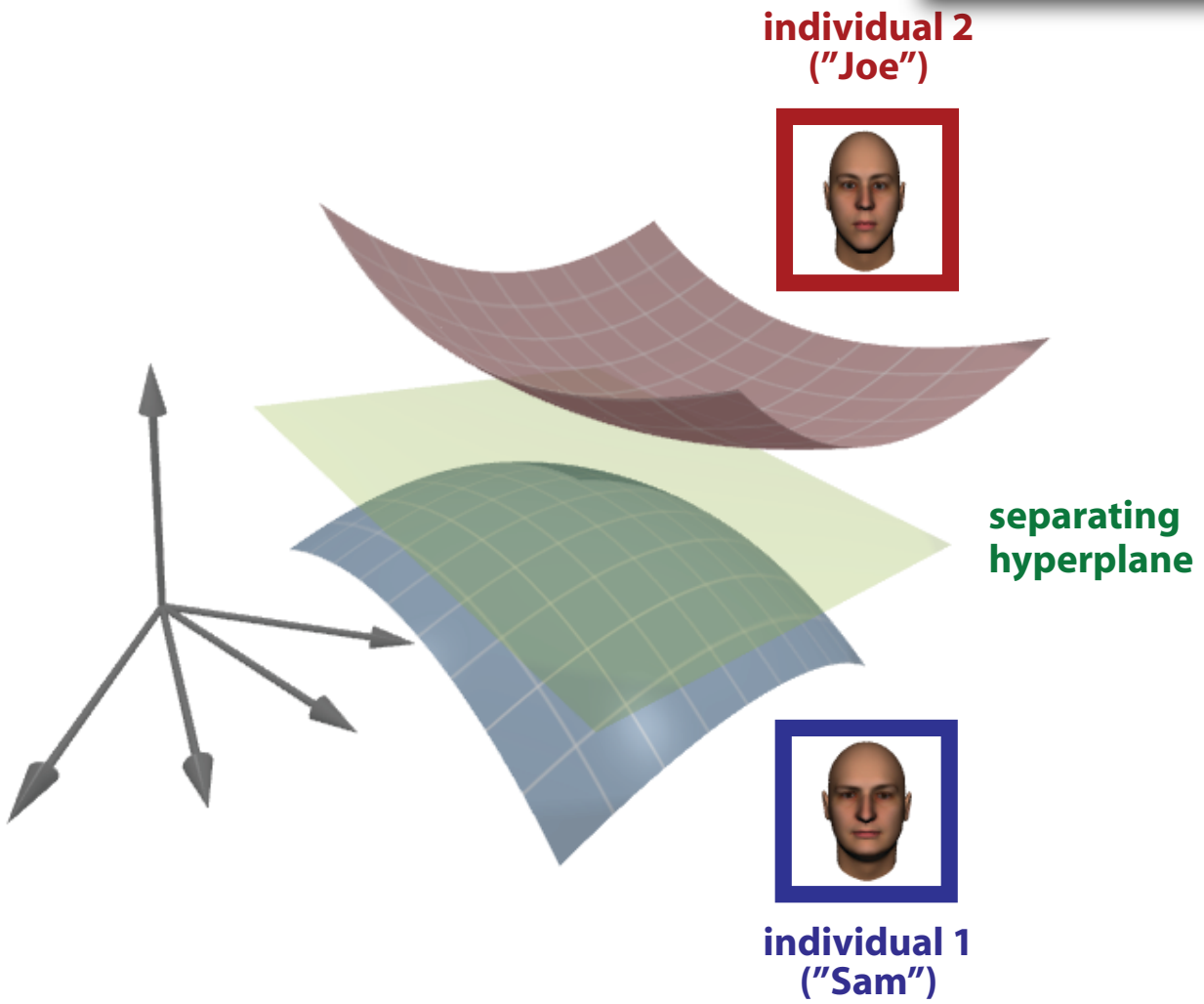




The computational crux of object and face recognition

A "good" population representation

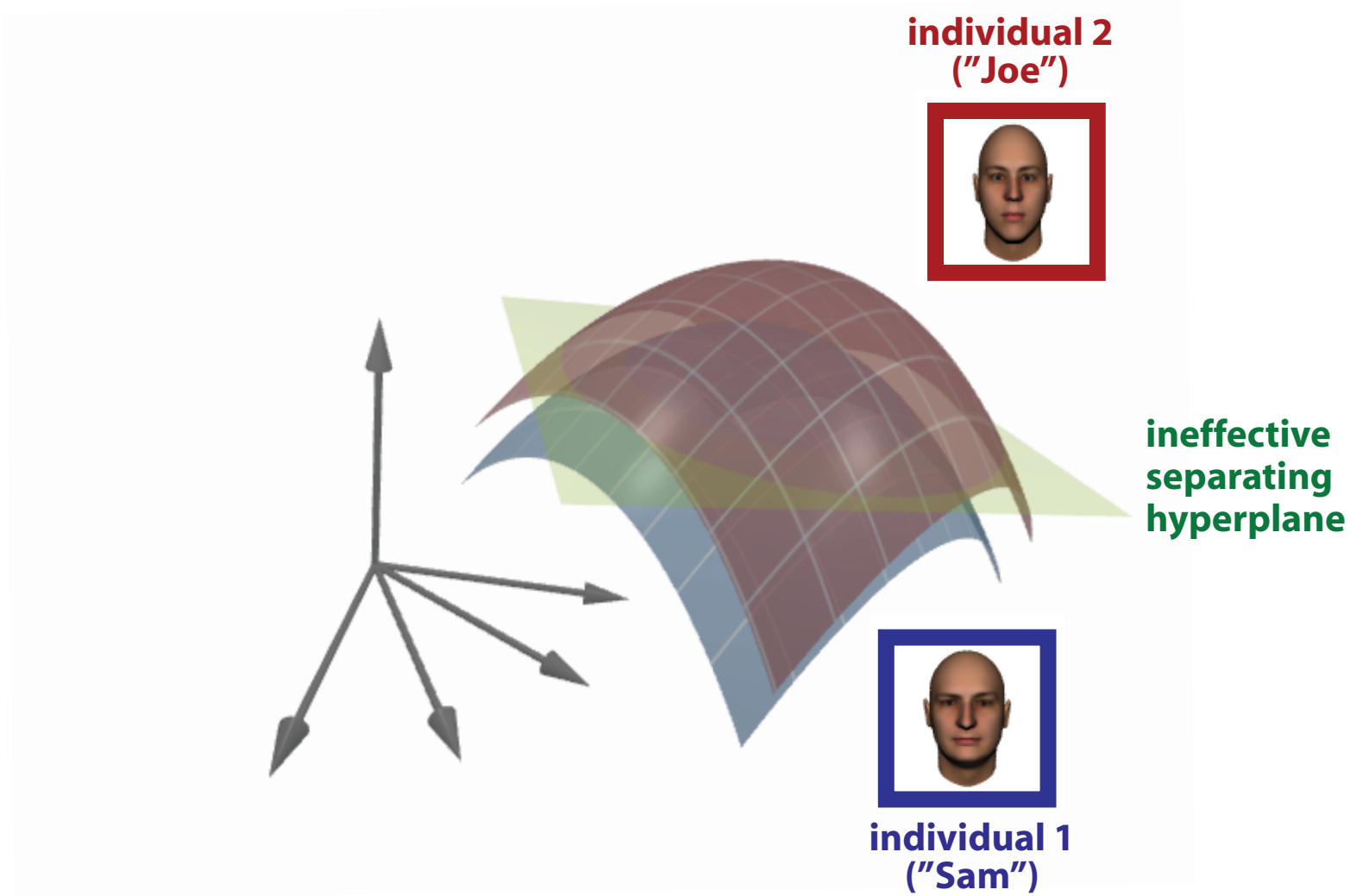
Explicit object representation



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

The computational crux of object and face recognition

A "bad" population representation

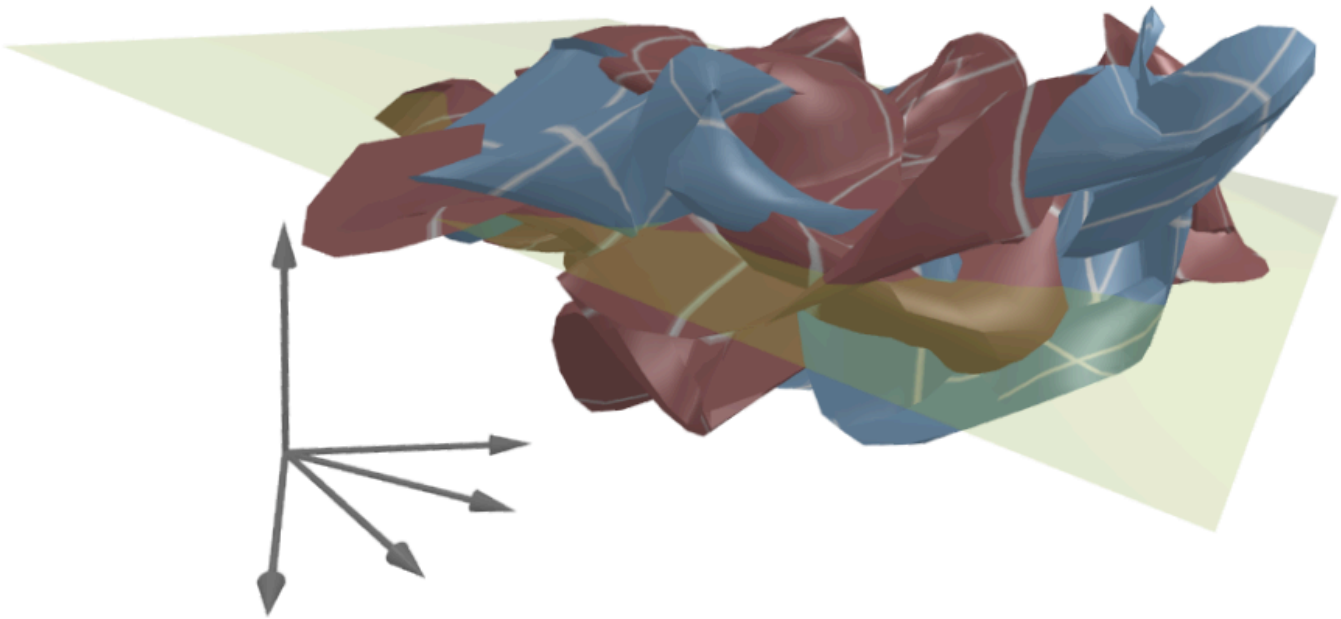


DiCarlo and Cox, ***TICS*** *(2007)*

The computational crux of object and face recognition

Actual pixel representation

(~ retinal image representation)



individual 2

ineffective separating hyperplane

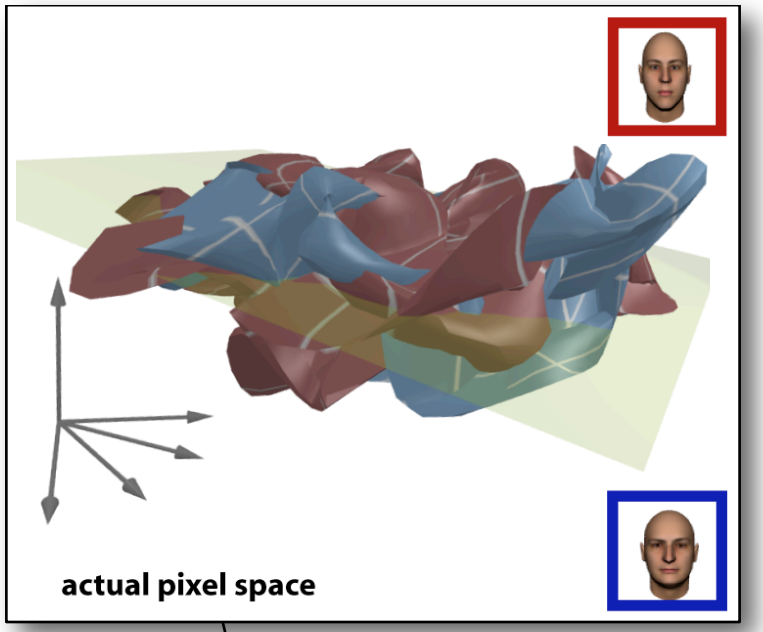
individual 1

object manifolds are "tangled"

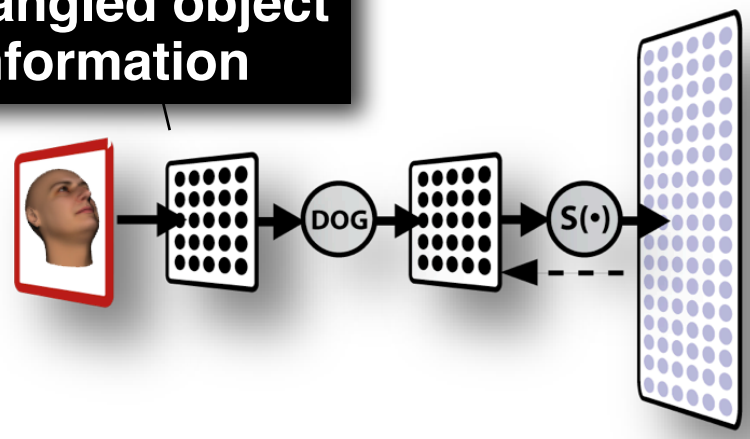
(Due to identity-preserving image variation.)

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

Understanding how the brain solves object recognition



Tangled object information

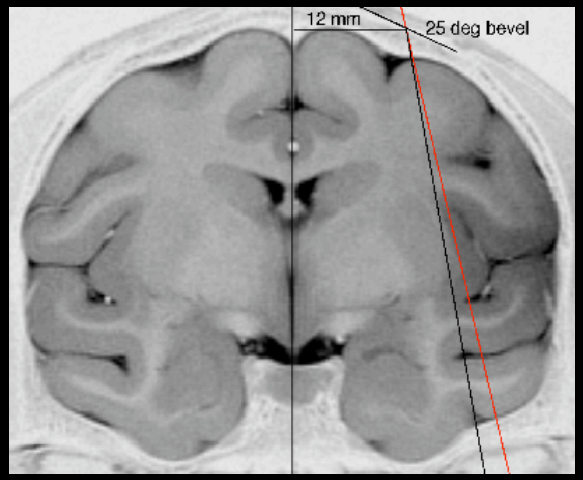


Nature of this representation ?

pixel RGC LGN V1 V2 V4 IT

Background: IT neurons are rapidly selective (among images)

e.g. Gross, Desimone, Albright, Rolls, Tanaka, Logothetis, Miyashita, Sheinberg, Connor, etc. etc.....



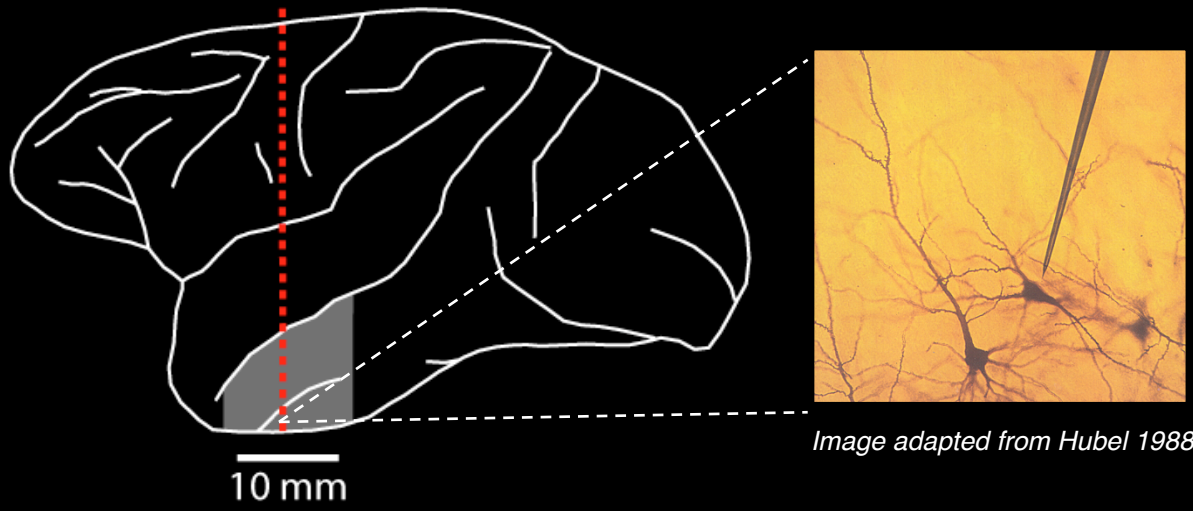
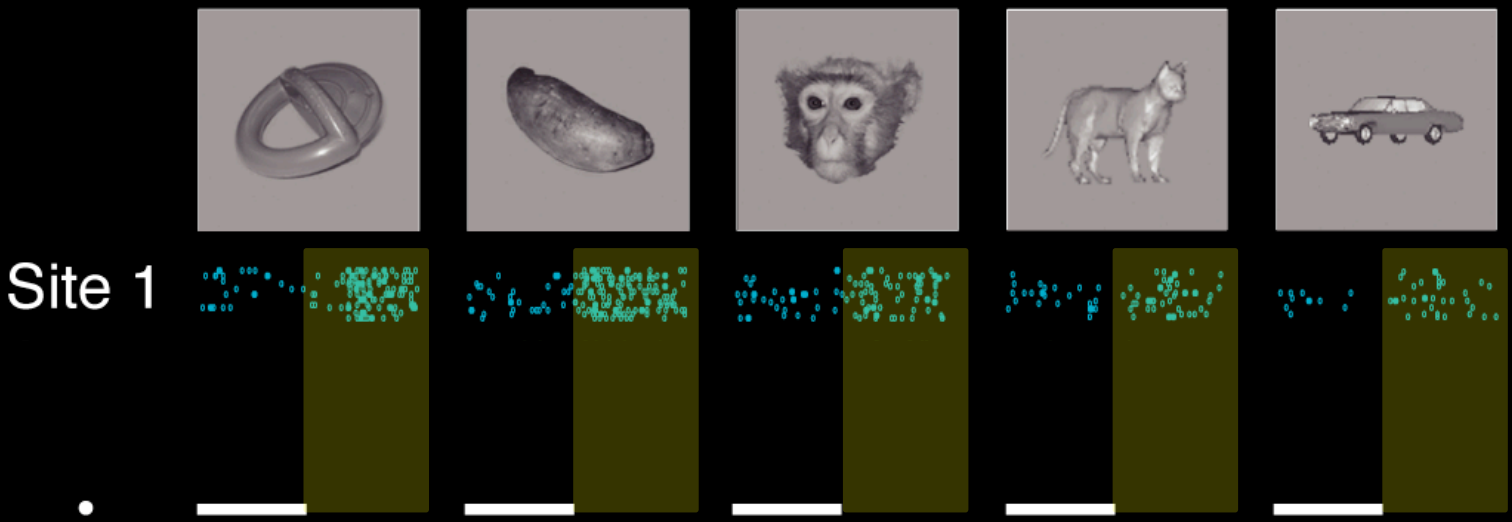


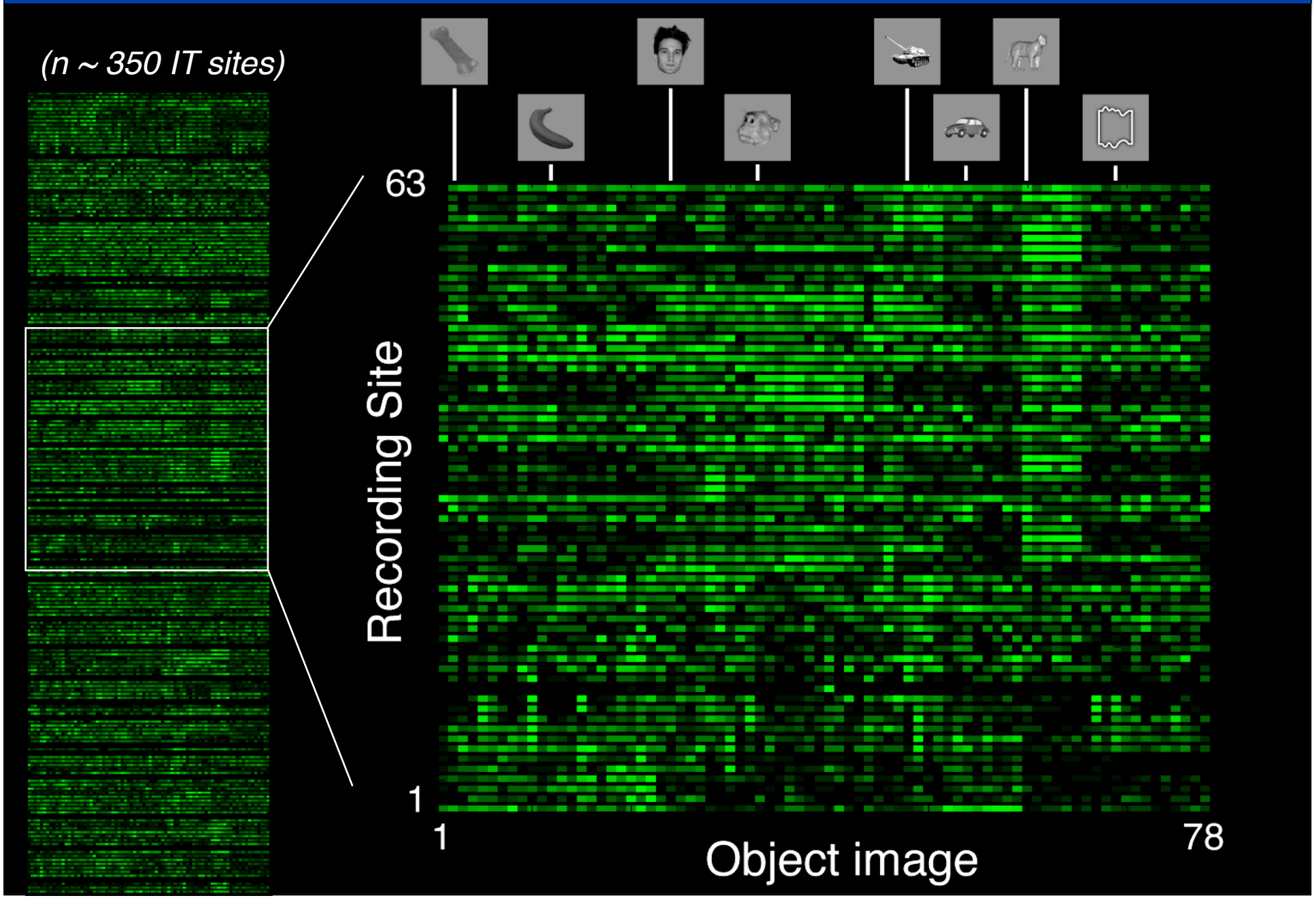
Image adapted from Hubel 1988



Behaviorally relevant analysis window

DiCarlo and Maunsell, Nature Neuroscience (2000)
DiCarlo and Maunsell, J Neurophysiology (2005)

Background: object information is explicit in IT



Background: object information is explicit in IT

Population activity

neuron 1

neuron 2

neuron N

Response vector

Biologically plausible linear classifiers

Predicted object category

cat/dog

toys

food

monkey face

white box contours

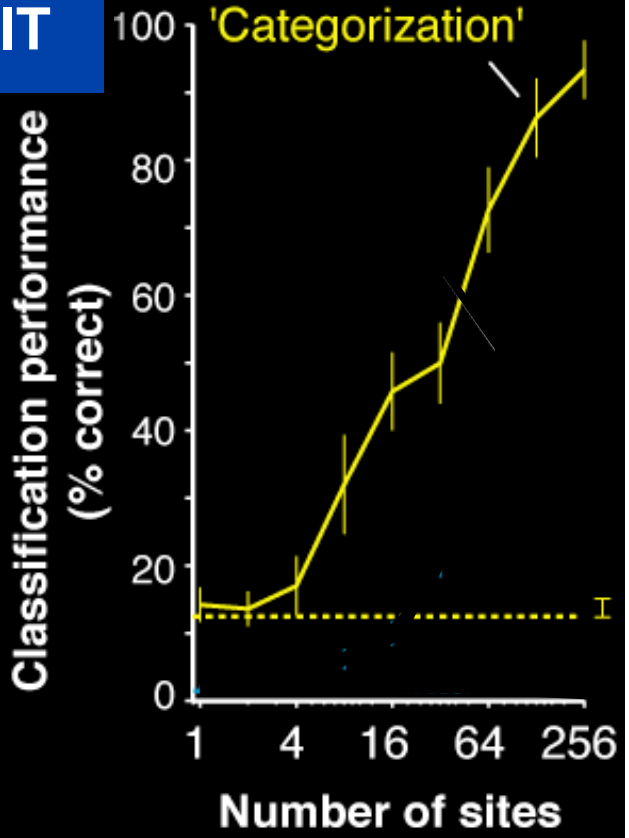
hand/body

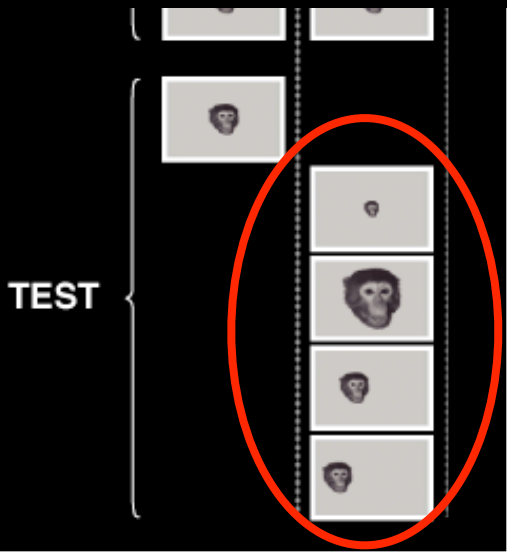
Only a small number of IT neurons needed for real-time performance

(n ~ 350 IT sites)

Generalization over object position, scale, and clutter

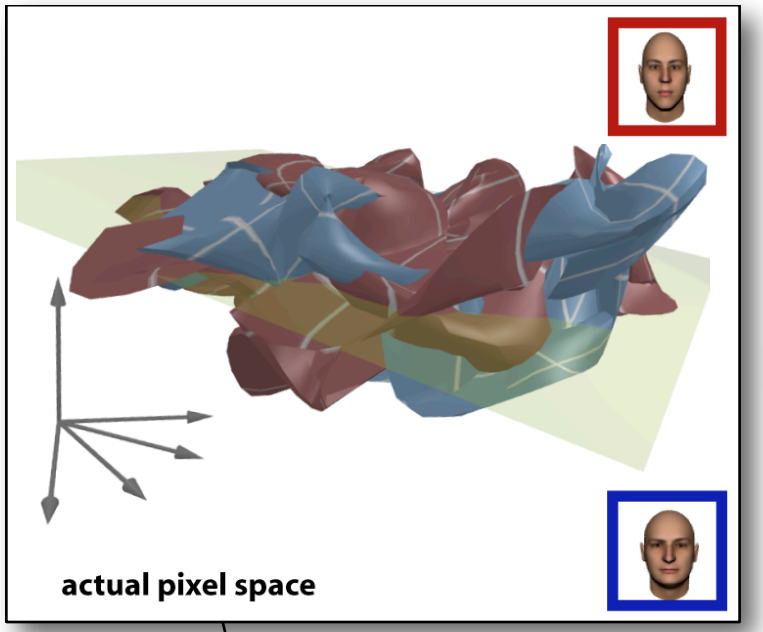
(not possible in early visual cortex)



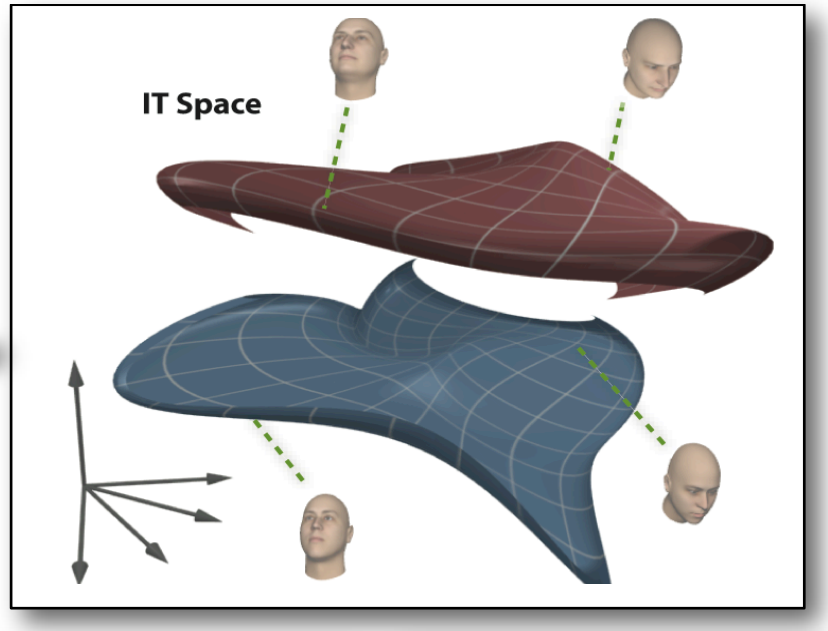


Hung, Kreiman*, Poggio and DiCarlo, **Science** (2005);*
*Li, Cox, Zoccolan & DiCarlo, **J Neurophys** (2009)*

Understanding how the brain solves object recognition



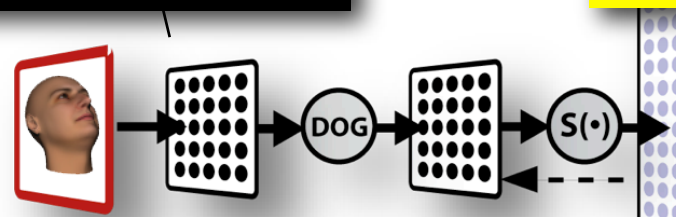
?

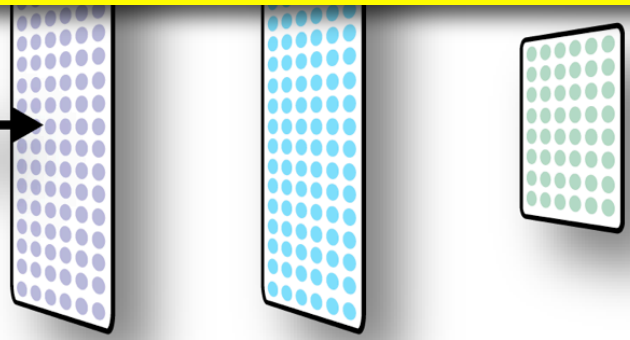


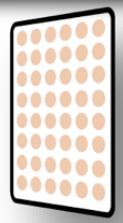
Tangled object information

How is the IT representation constructed ???

Explicit object representation







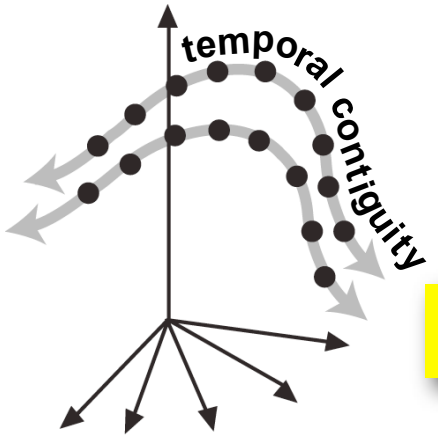
- Focus.
- First spikes.
- No binding

pixel RGC LGN V1 V2 V4 IT

Understanding how the brain solves object recognition

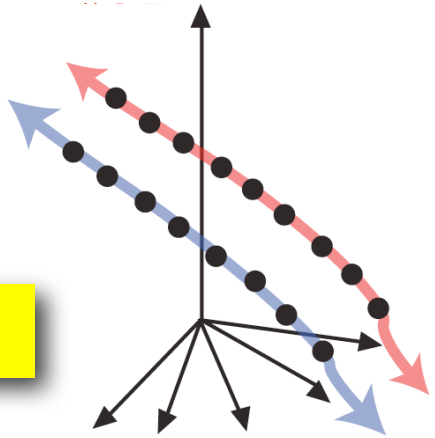
image data in input representation

image data transformed to new representation



?

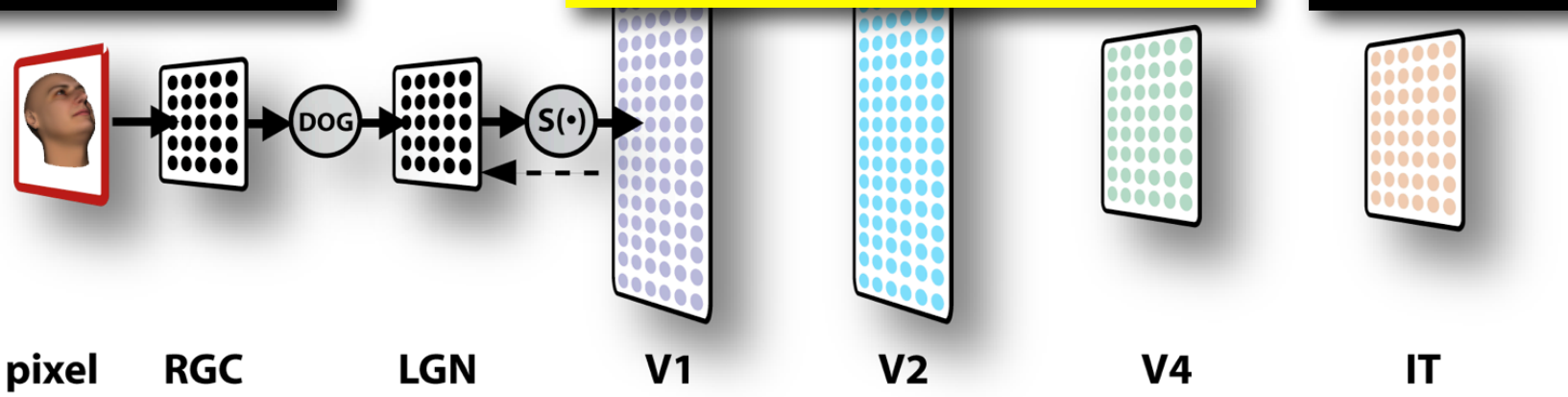
Innate or learned ?



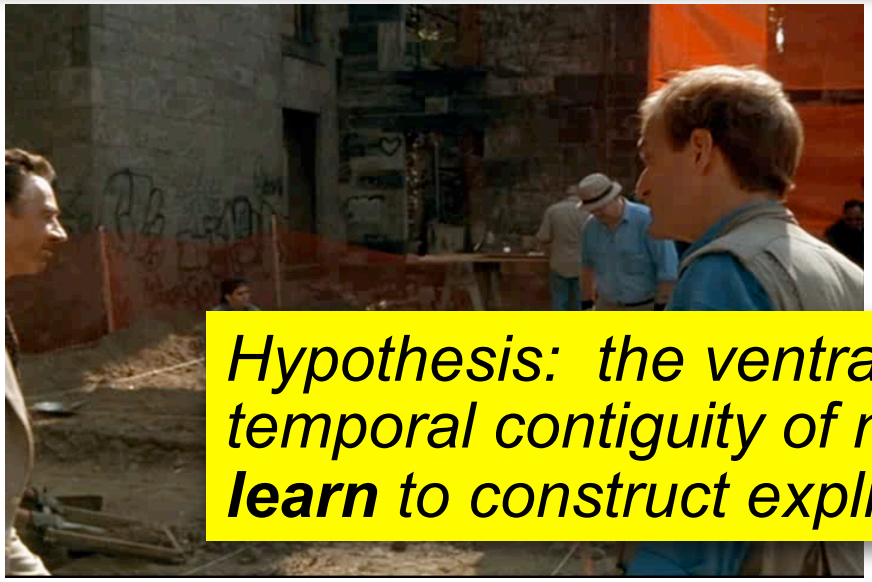
Tangled object information

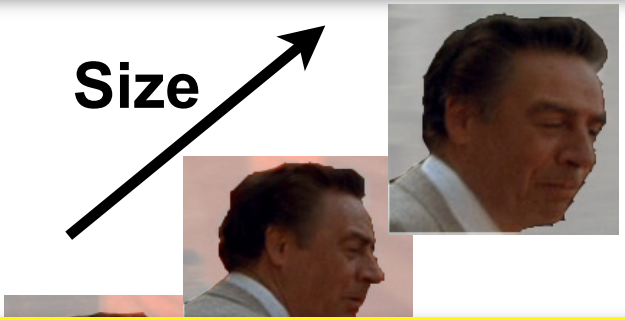
How is the IT representation constructed ???

Explicit object representation



The temporal contiguity of natural experience is a gold mine for learning

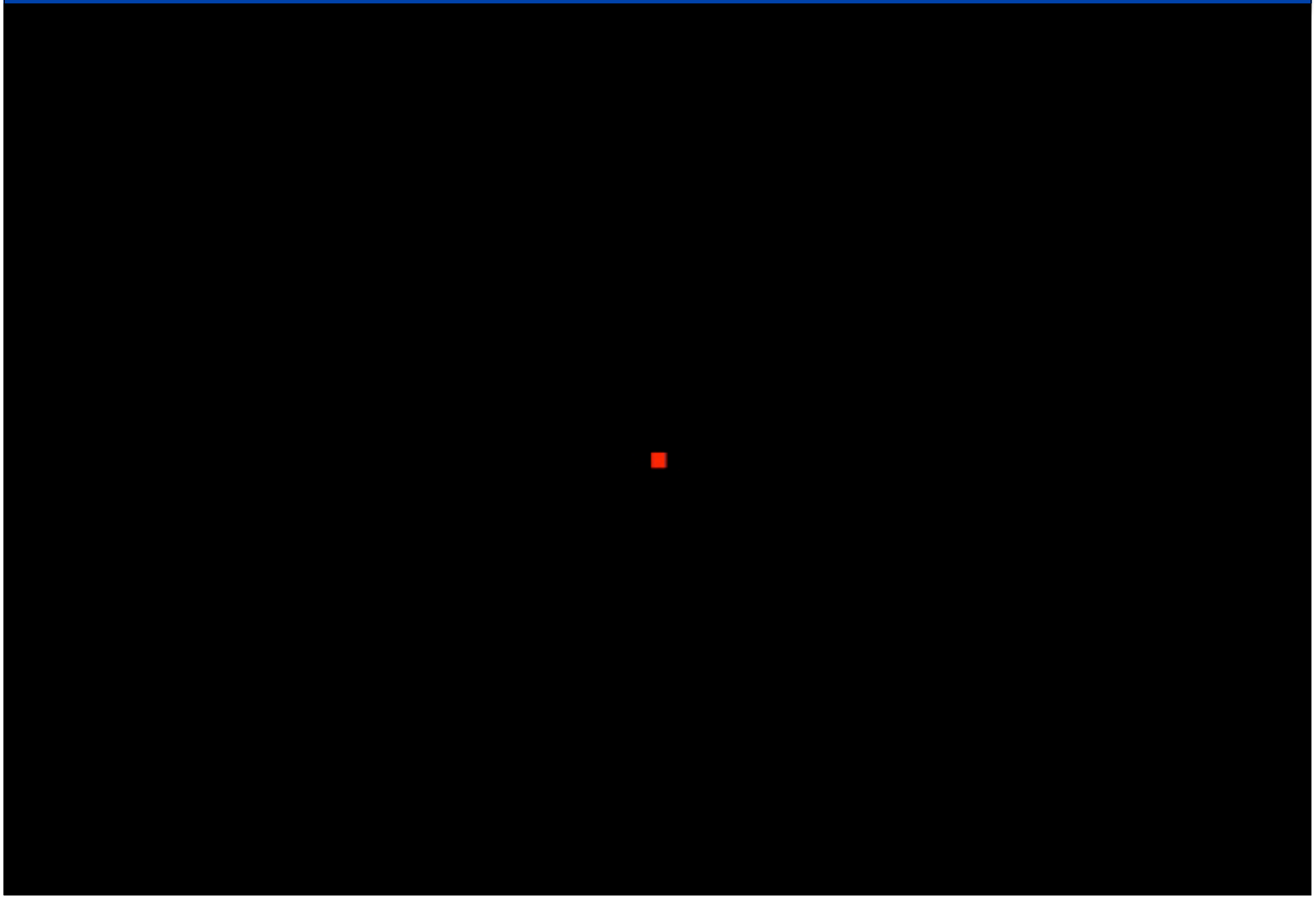




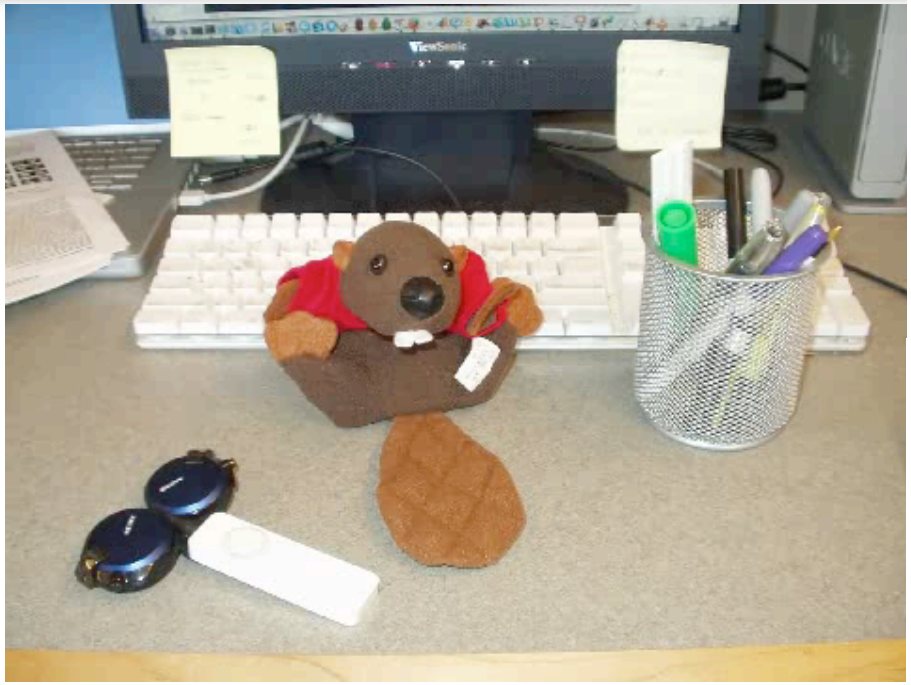
*Hypothesis: the ventral visual stream uses the temporal contiguity of natural visual experience to **learn** to construct explicit object representation*

Foldiak, 1991;
Wallis & Rolls, 1997;
Wallis & Bulthoff, 2001;
Wiskott & Sejnowski, 2002;
Fiser and Aslin, 2002;
Perry, Rolls, Stringer, 2006;
Wyss, Konig, Verschure, 2006;
Sprekeler, Michaelis, Wiskott, 2007;
Masquelier & Thorpe, 2007;
Masquelier, Serre, Thorpe, Poggio, 2008
etc.

Start simple: focus on position tolerance ("invariance")



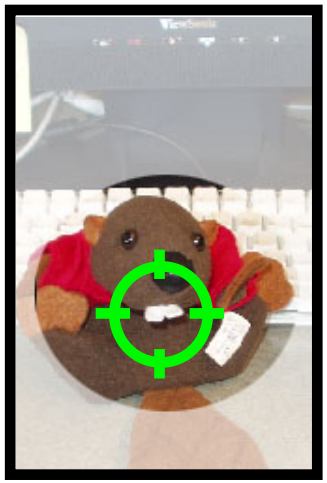
Natural visual experience and position tolerance ("invariance")



Retinal image:

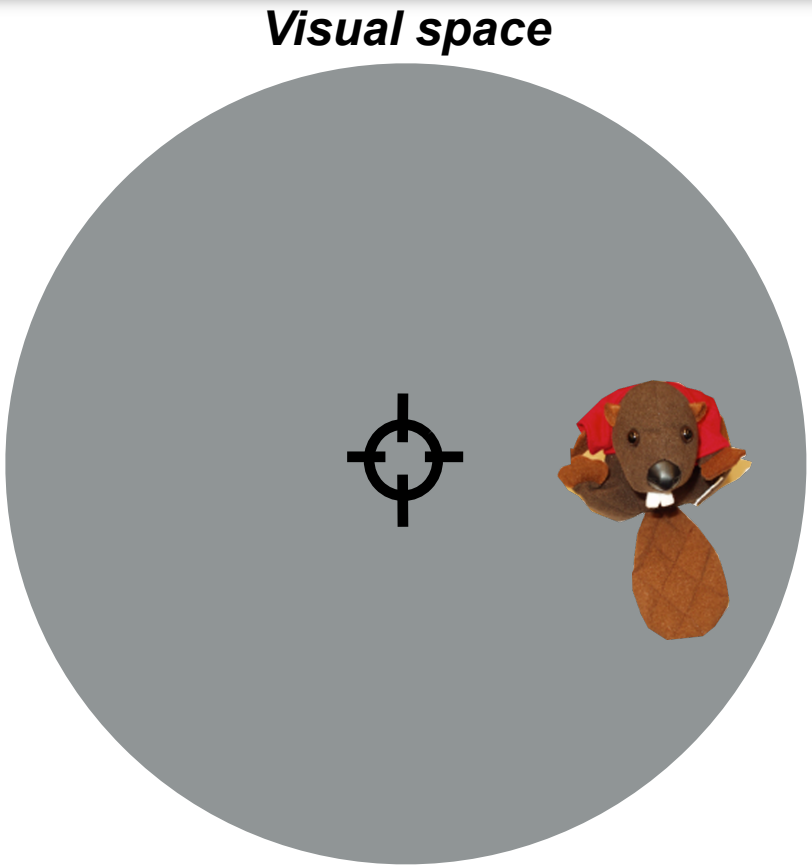
Unsupervised retinal experience:

short time



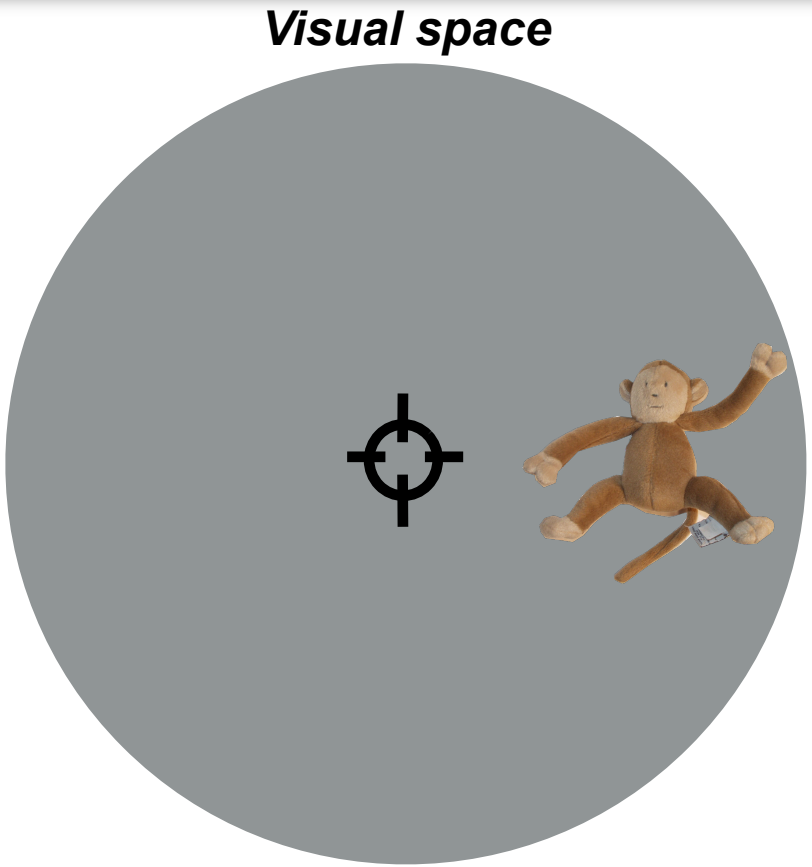
Learn position tolerance?

Experimental rationale to test temporal contiguity* theories of invariance



Cox, Meier, Oertelt and DiCarlo. ***Nature Neuroscience*** *(2005)*

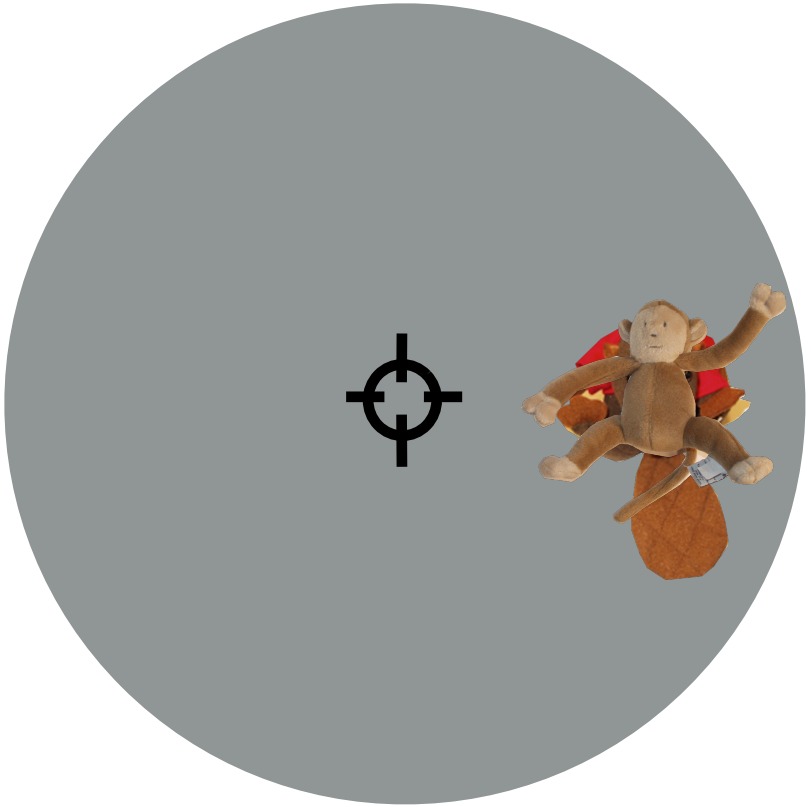
Experimental rationale to test temporal contiguity* theories of invariance



Cox, Meier, Oertelt and DiCarlo. ***Nature Neuroscience*** *(2005)*

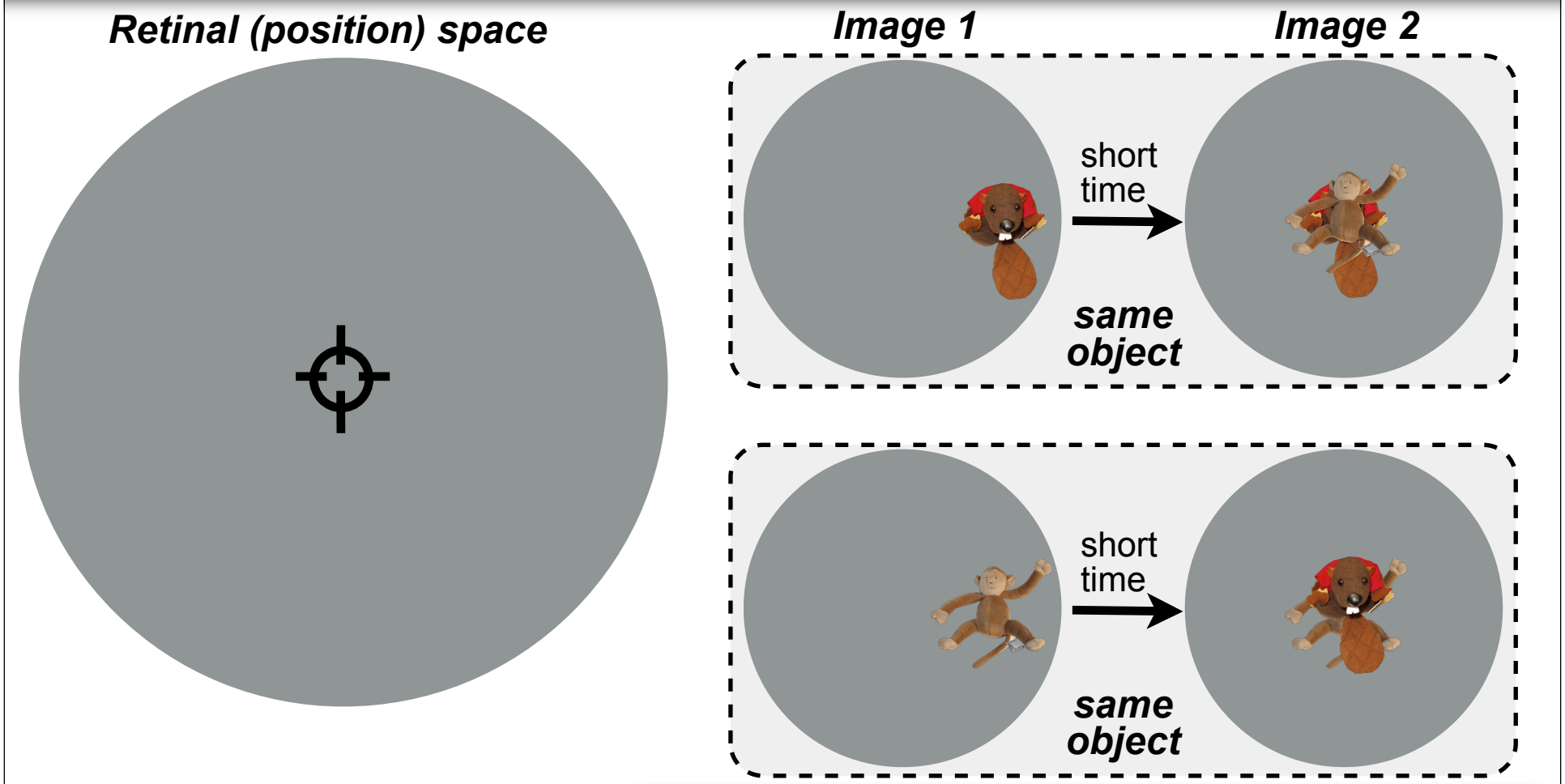
Experimental rationale to test temporal contiguity* theories of invariance

Retinal (position) space



Cox, Meier, Oertelt and DiCarlo. ***Nature Neuroscience*** *(2005)*

Experimental rationale to test temporal contiguity* theories of invariance



What if we altered the visual world?

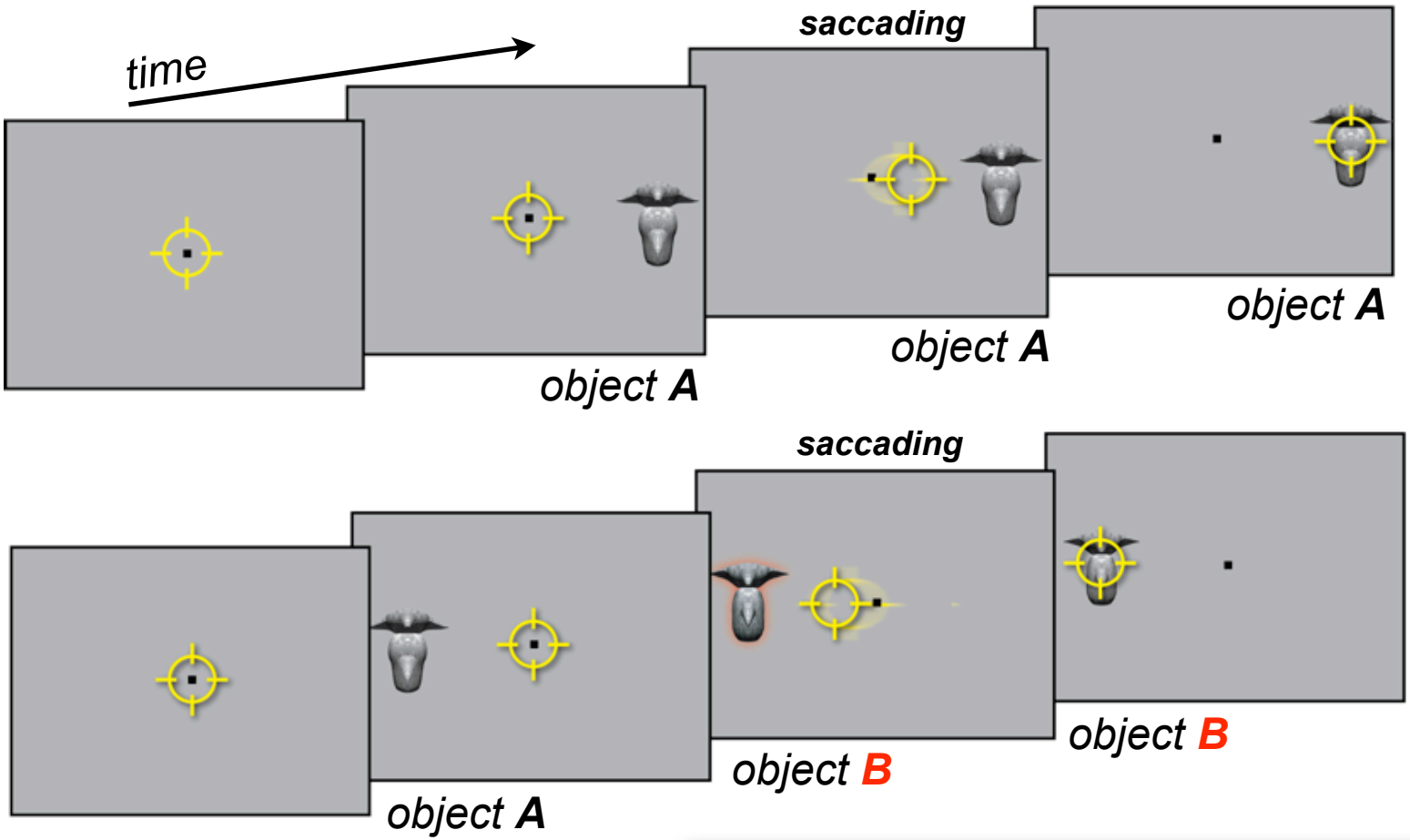
Neuronal prediction:

IT neurons will tend to confuse these objects across these positions

Perceptual prediction:

subjects will tend to confuse these objects across these positions

Experiment 1 (human subject, within subject design)



Normal exposure

Swapped exposure

What if we altered the visual world?

Perceptual prediction:

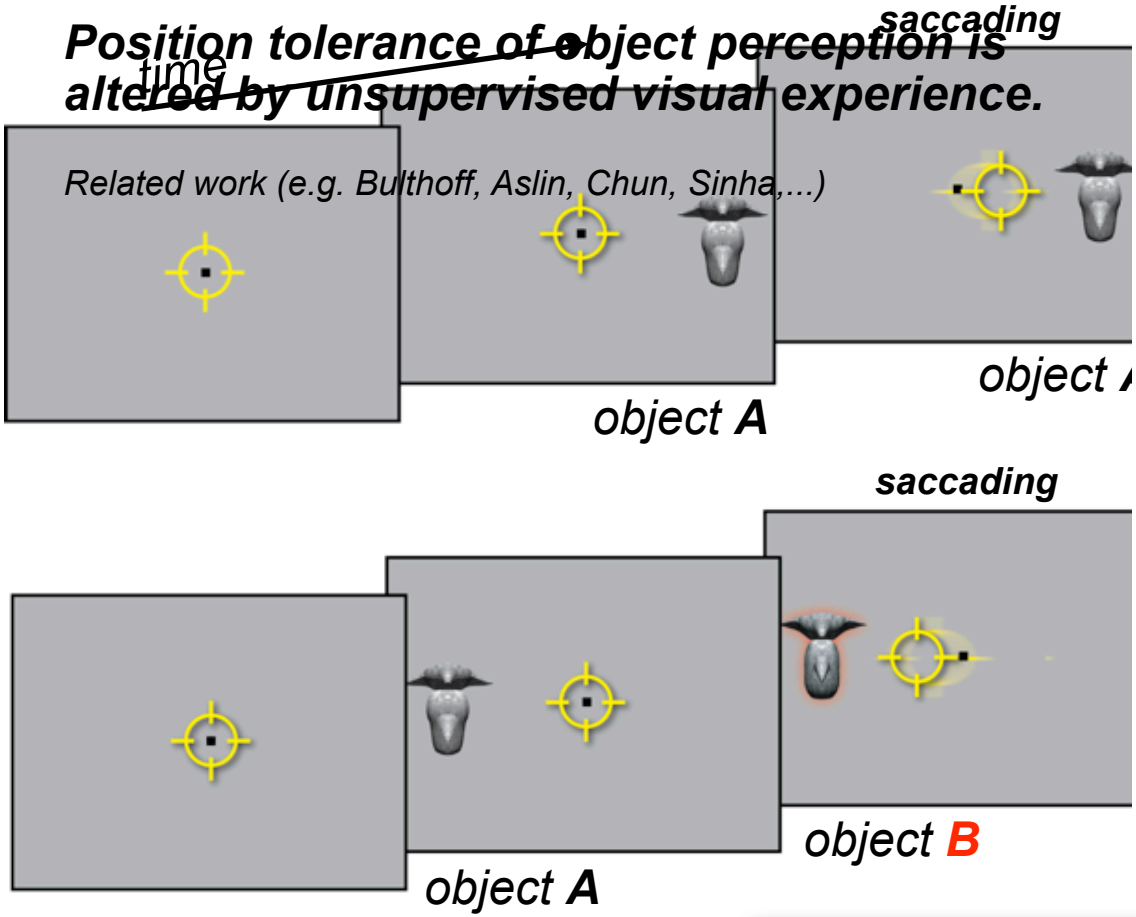
subjects will tend to confuse these objects across these positions

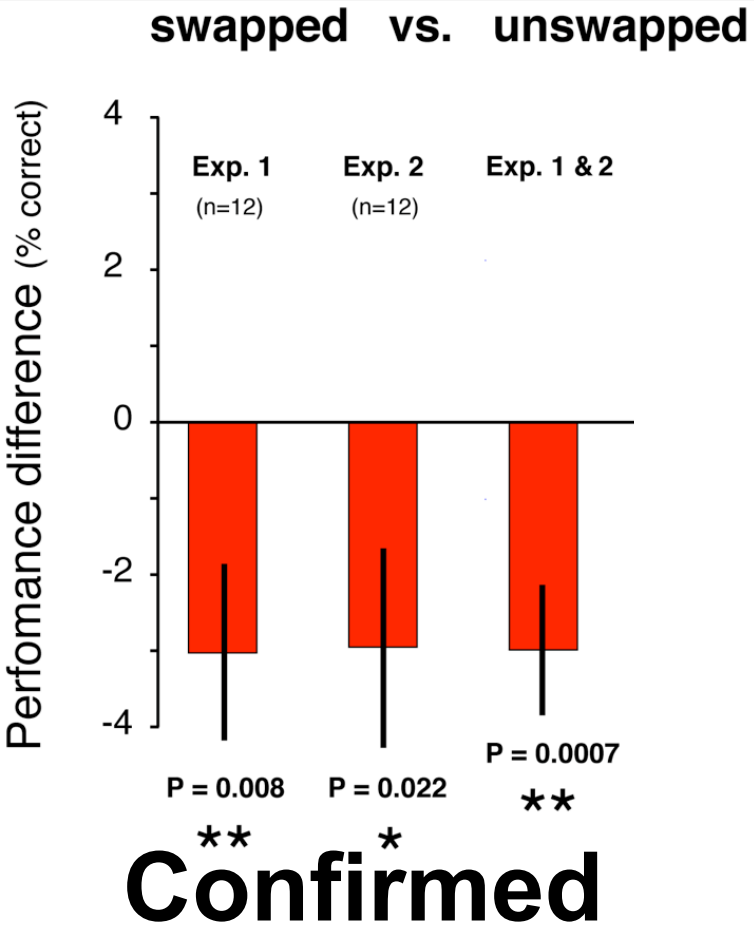
Cox, Meier, Oertelt and DiCarlo. ***Nature Neuroscience*** *(2005)*

Experiment 1 (human subject, within subject design)

Position tolerance of object perception is altered by unsupervised visual experience.

Related work (e.g. Bulthoff, Aslin, Chun, Sinha,...)





Confirmed

What if we altered the visual world?

Exposure phase
200-400 swaps

→

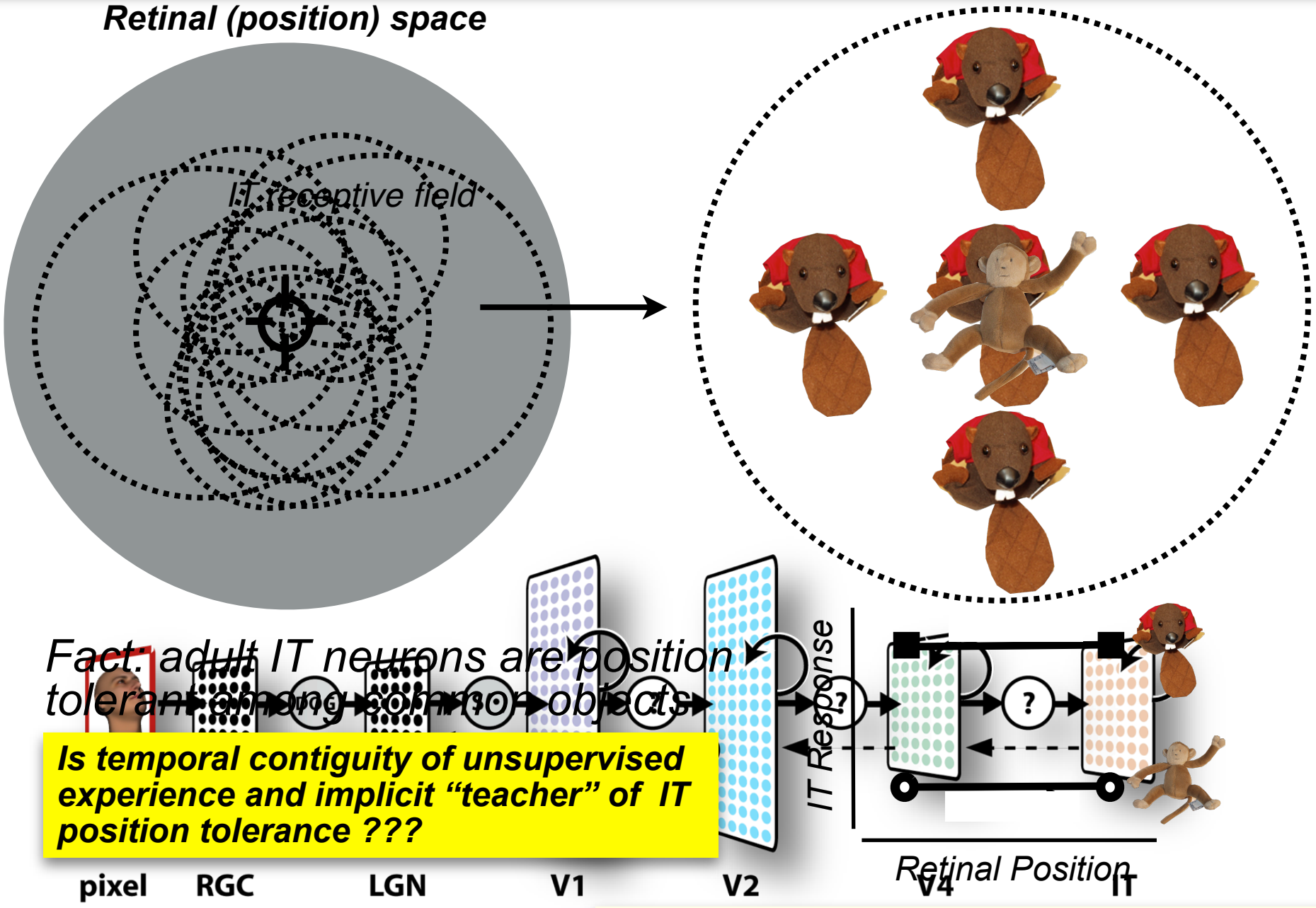
Test phase
"same" / "different" task

Perceptual prediction:

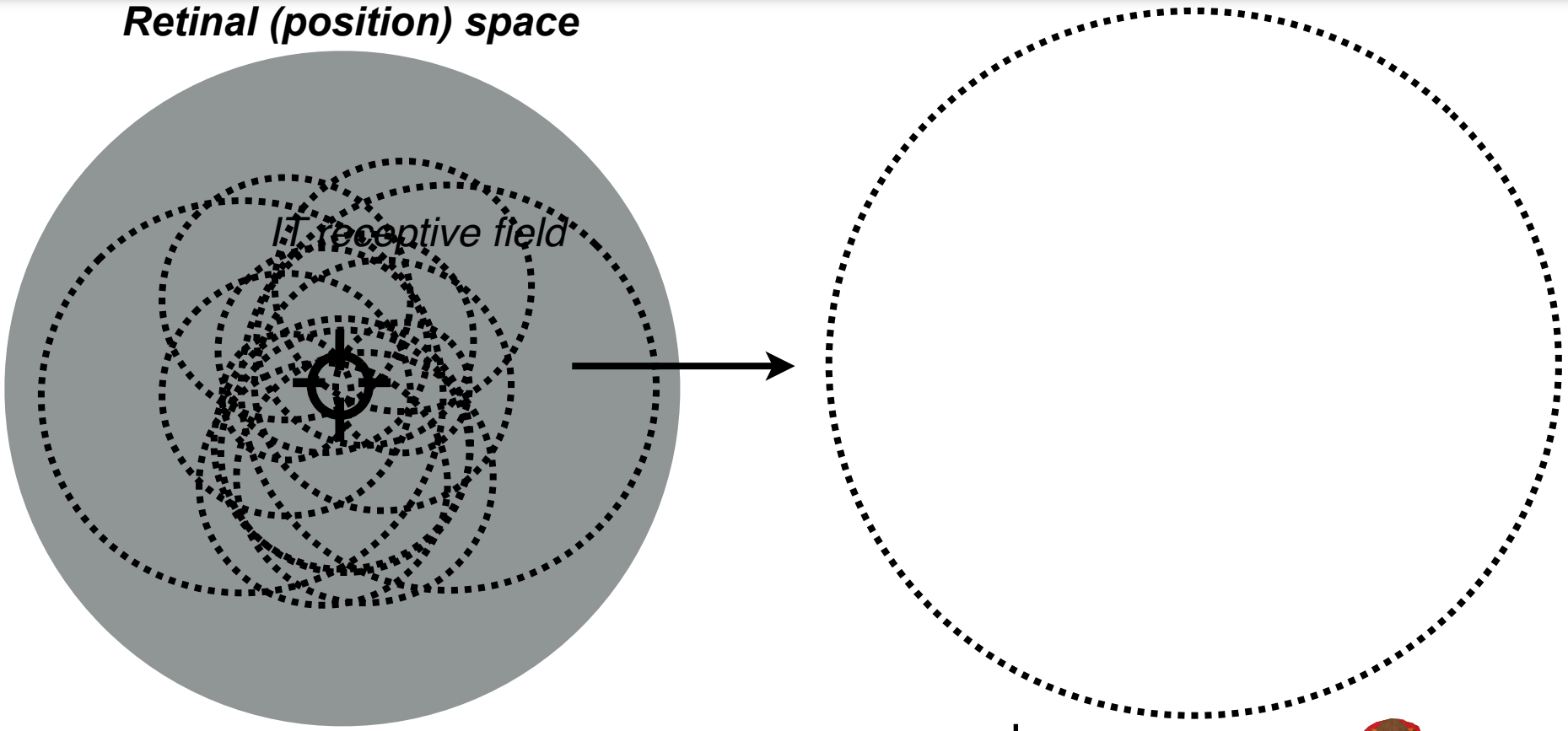
subjects will tend to confuse these objects across these positions

Cox, Meier, Oertelt and DiCarlo. ***Nature Neuroscience*** *(2005)*

Background: Role of IT neurons in position invariant object representation

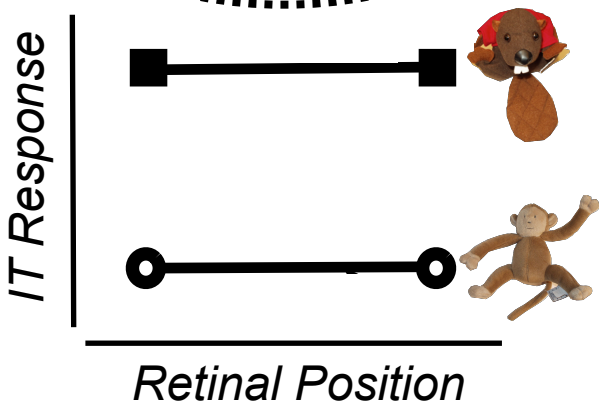


Background: Role of IT neurons in position invariant object representation

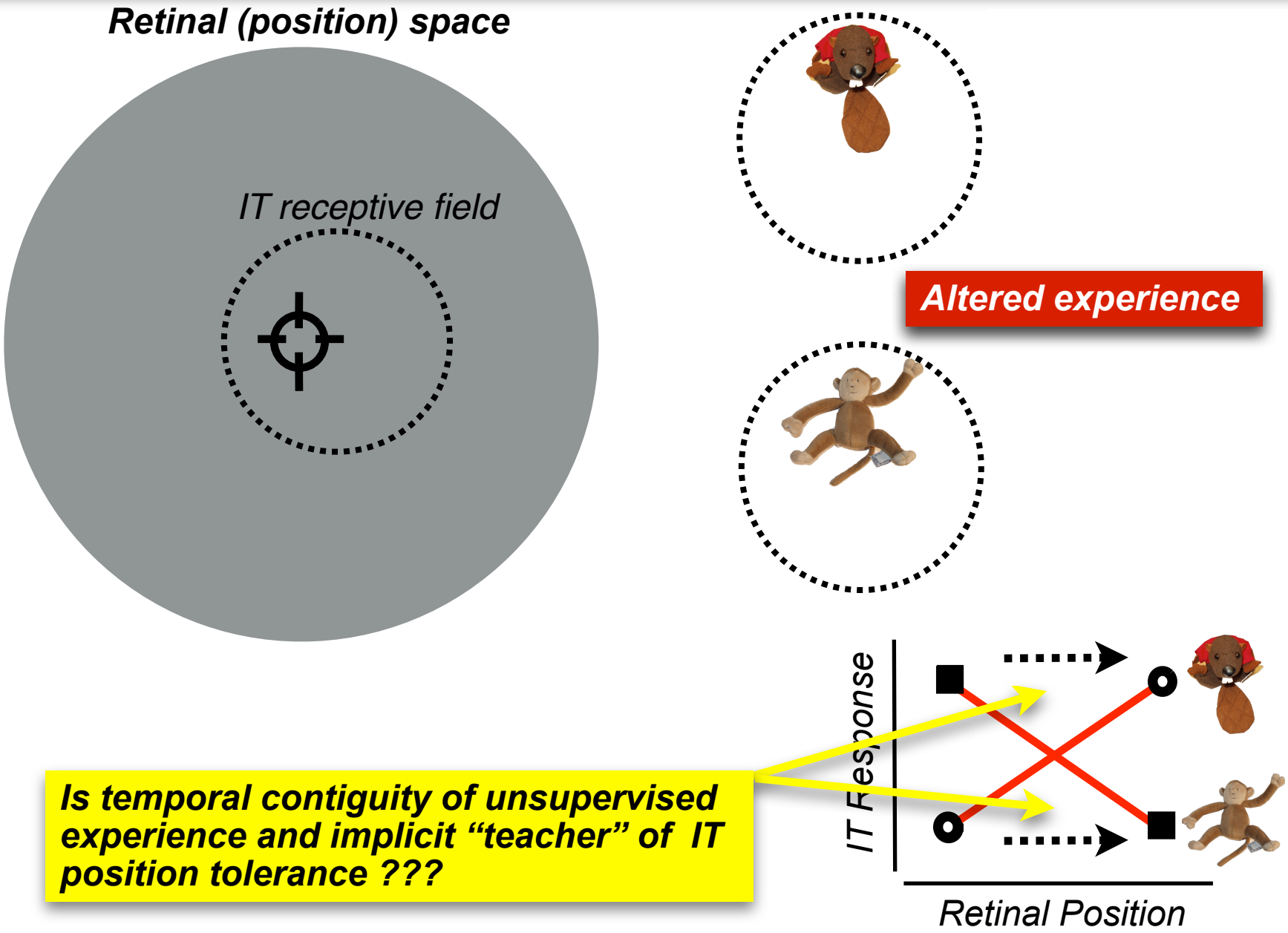


Fact: adult IT neurons are position tolerant among common objects

Is temporal contiguity of unsupervised experience and implicit "teacher" of IT position tolerance ???

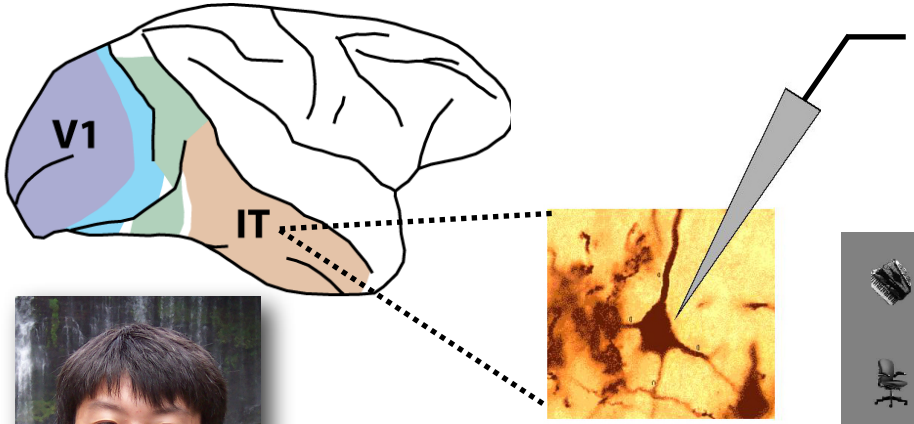


Neuronal (IT) prediction of a temporal contiguity theory of invariance



Is temporal contiguity of unsupervised experience and implicit "teacher" of IT position tolerance ???

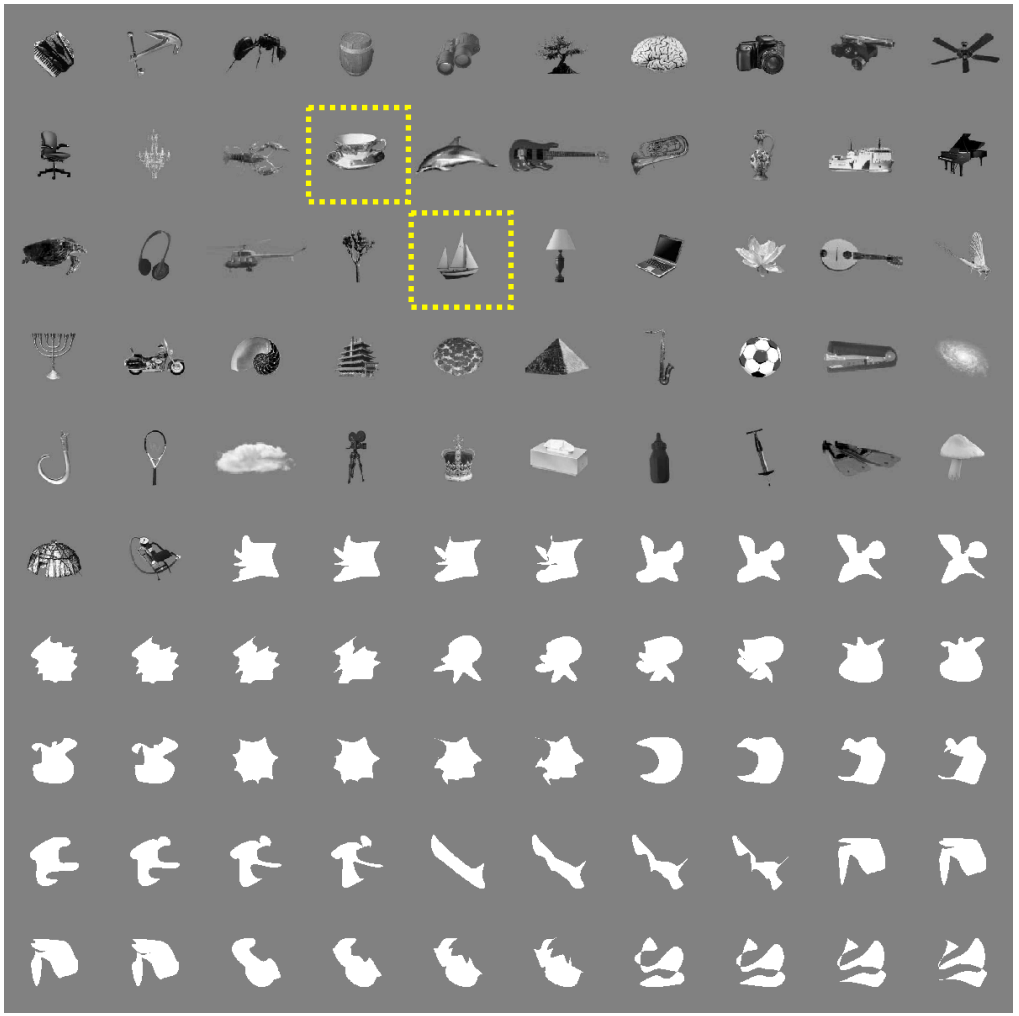
Experimental design (within-neuron, longitudinal design)



Nuo Li

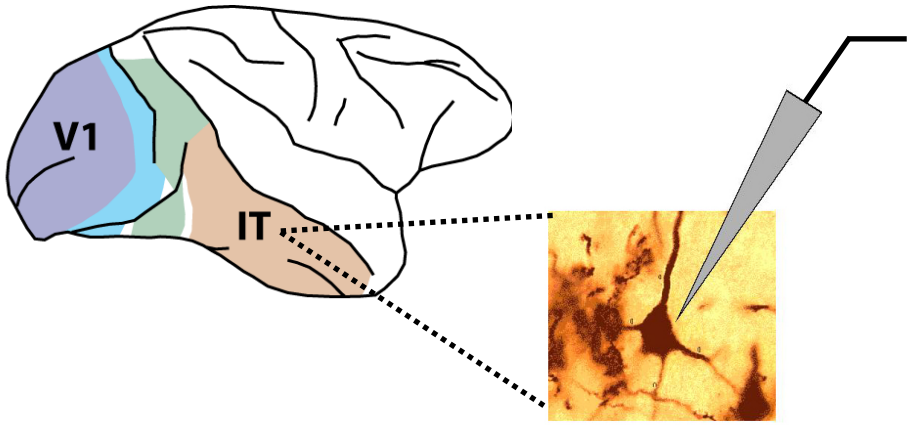
P

N



Screen

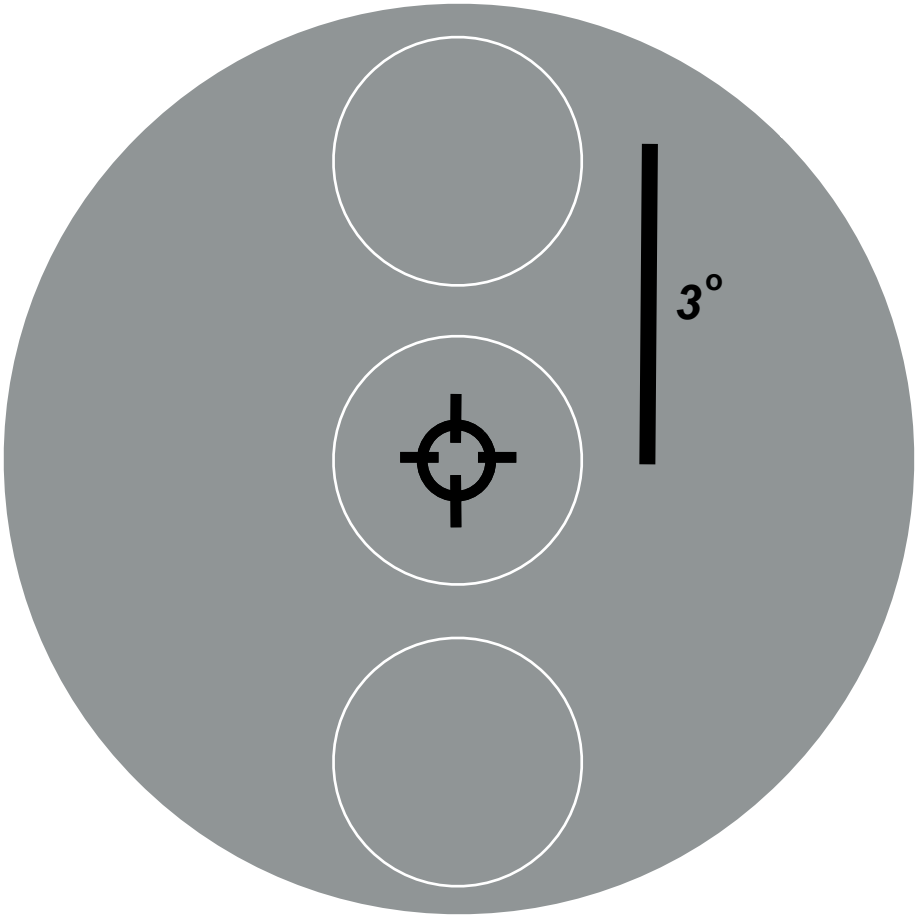
Experimental design (within-neuron, longitudinal design)

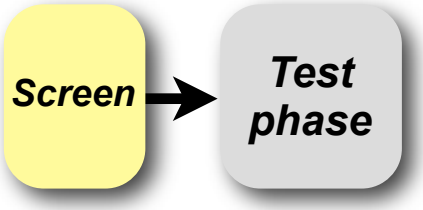


P

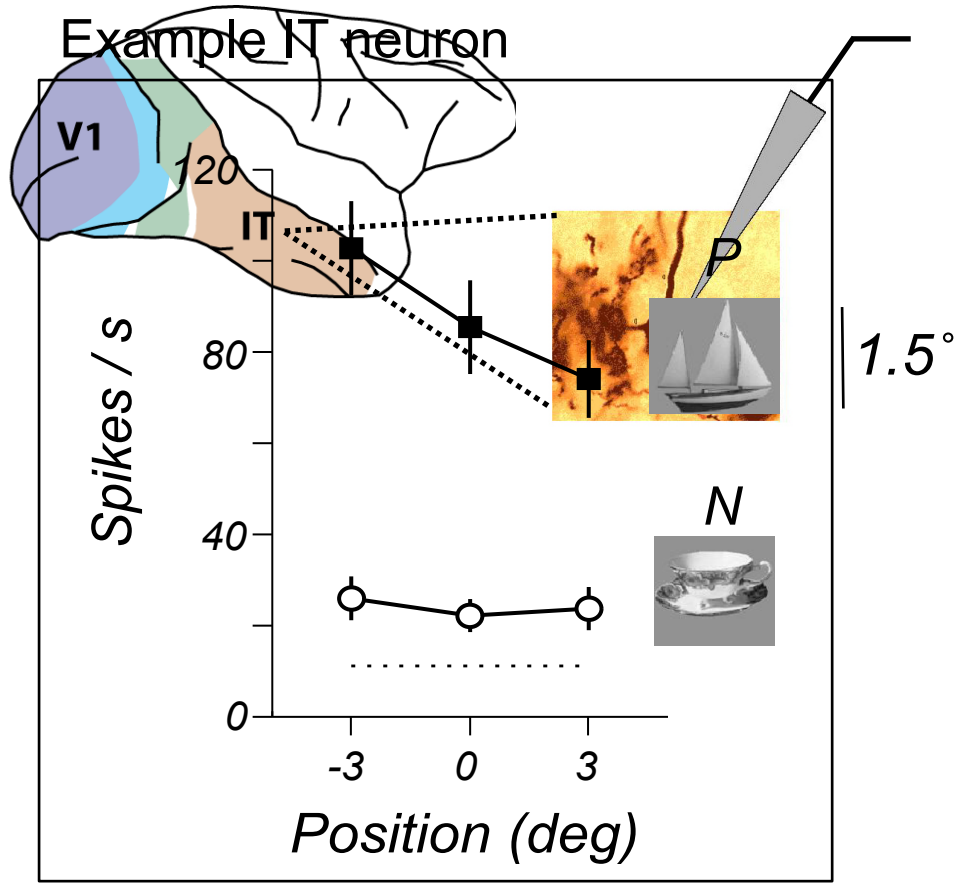
N

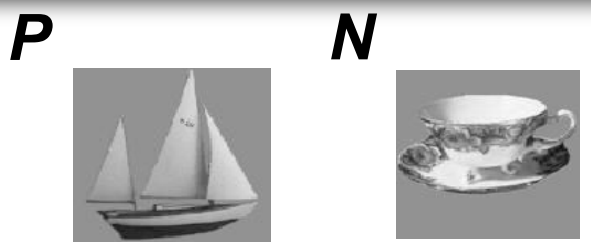
Retinal (position) space



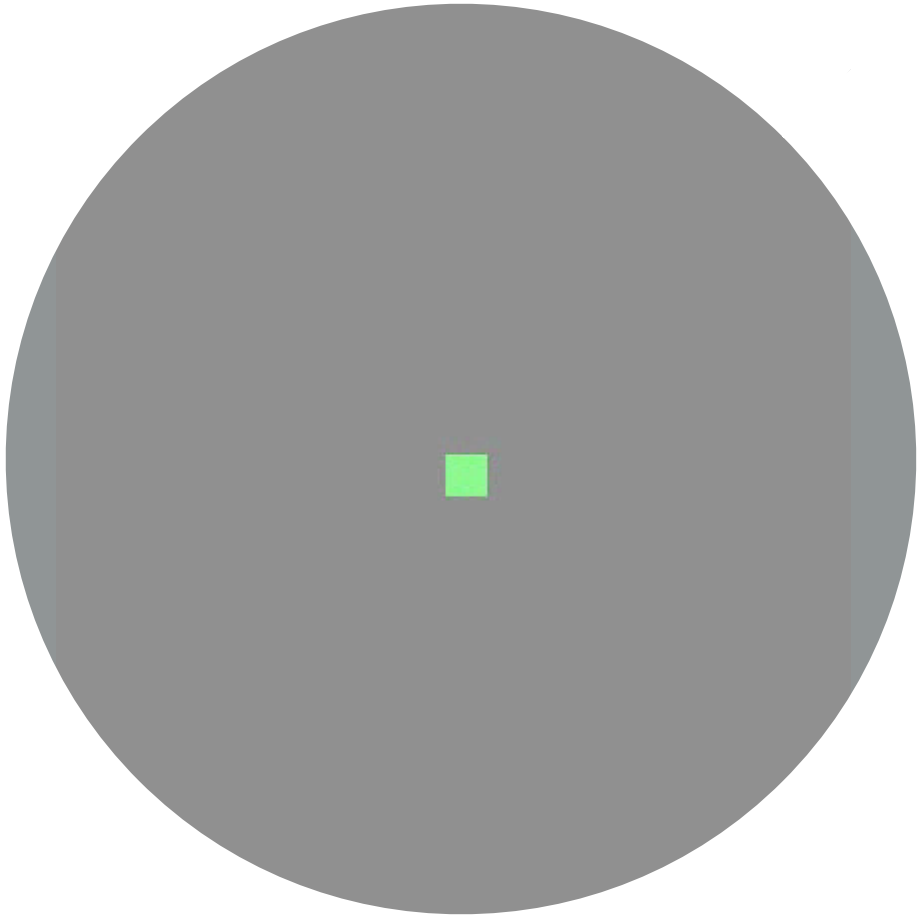


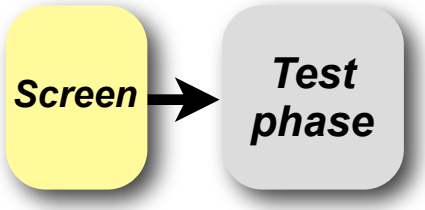
Experimental design (within-neuron, longitudinal design)



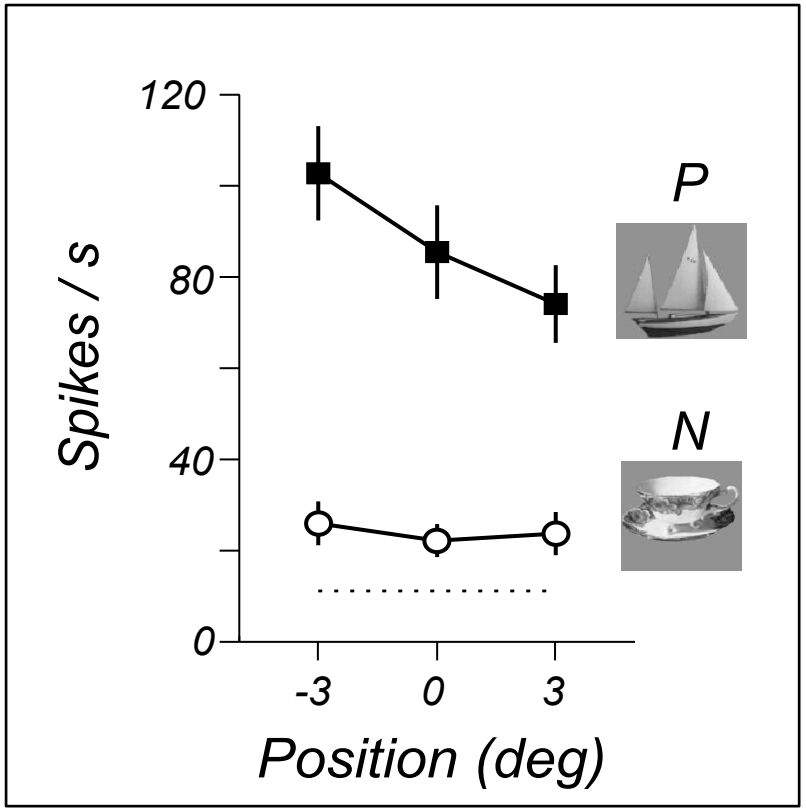


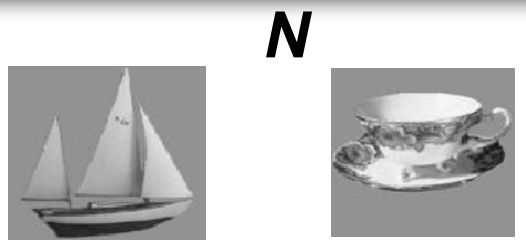
Retinal (position) space

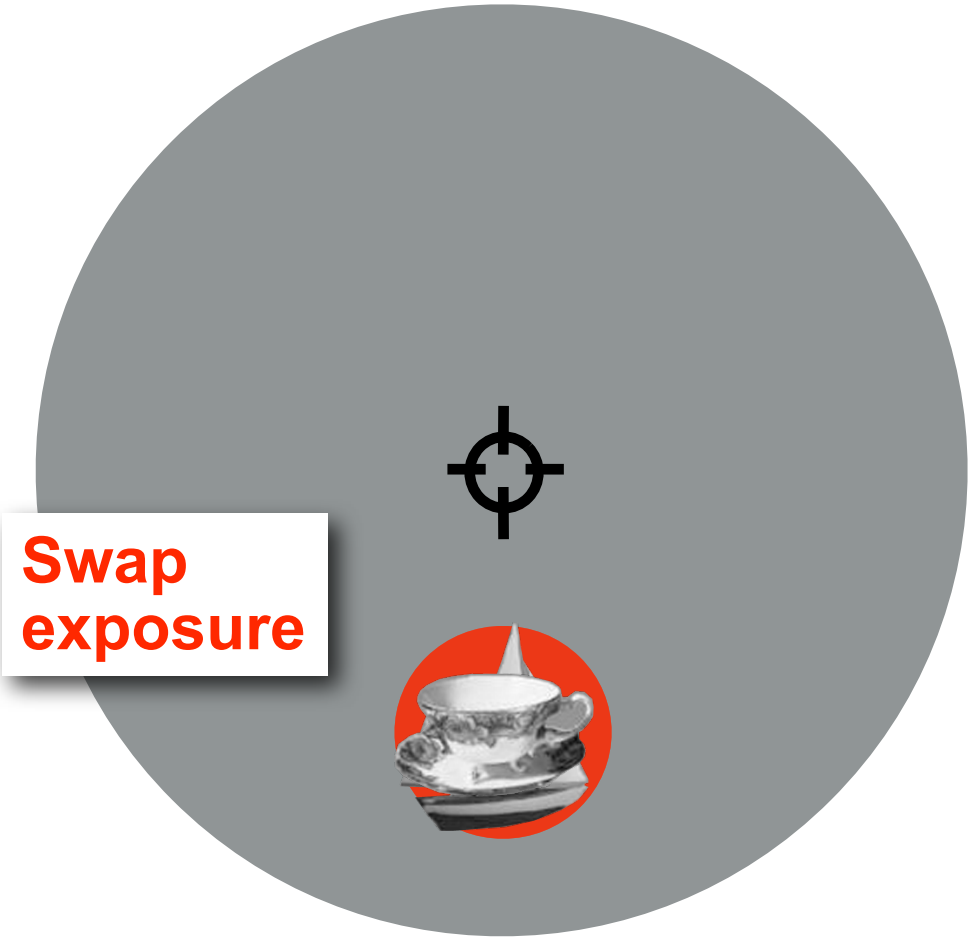


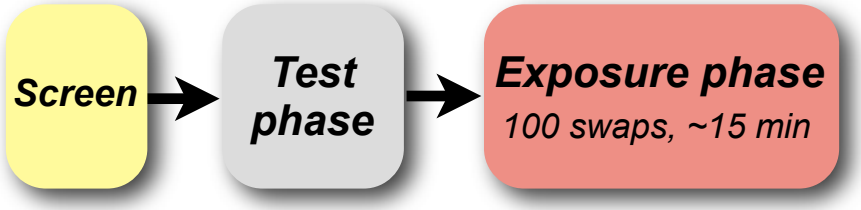


Experimental design (within-neuron, longitudinal design)

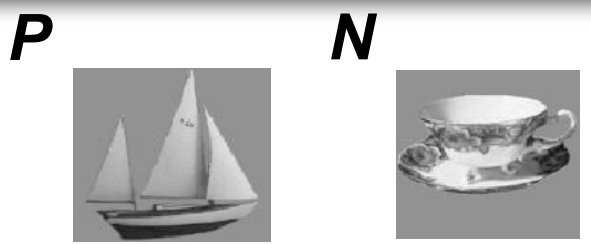


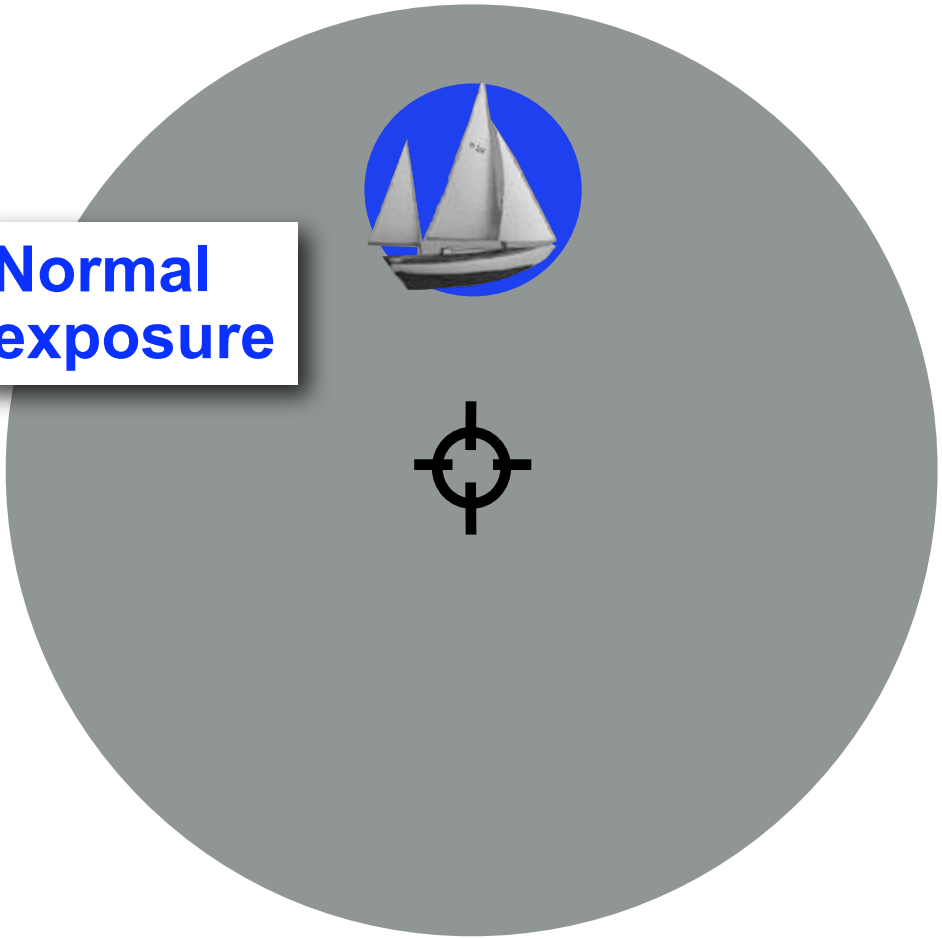


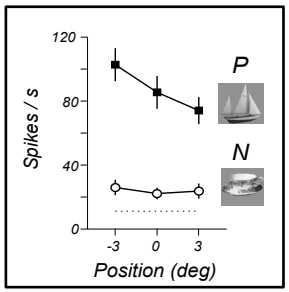


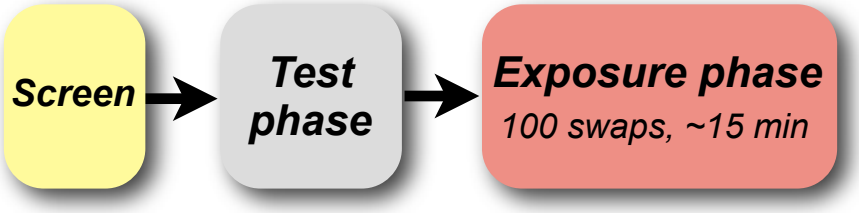


Experimental design (within-neuron, longitudinal design)

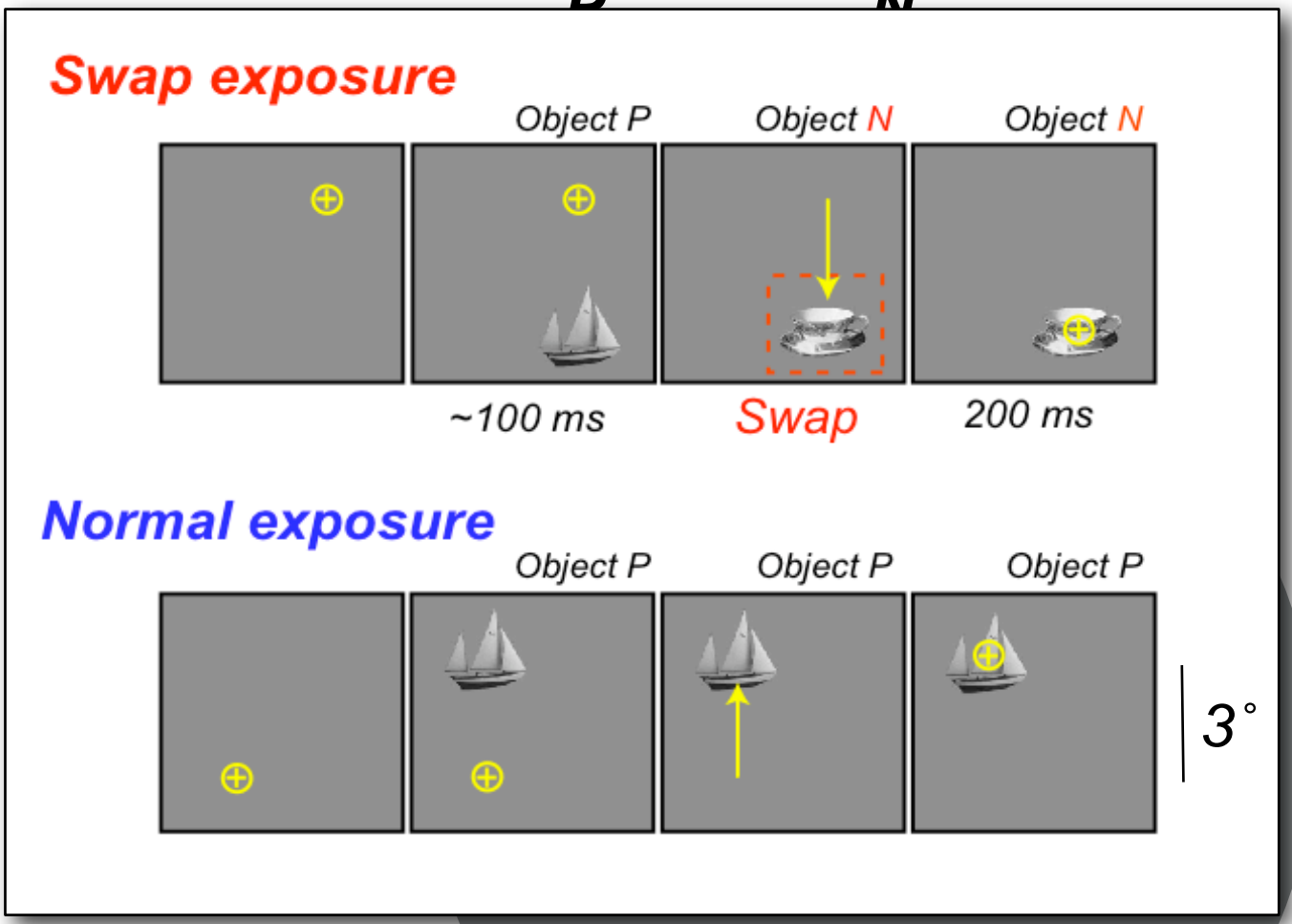


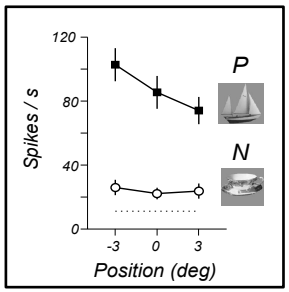






Experimental design (within-neuron, longitudinal design)



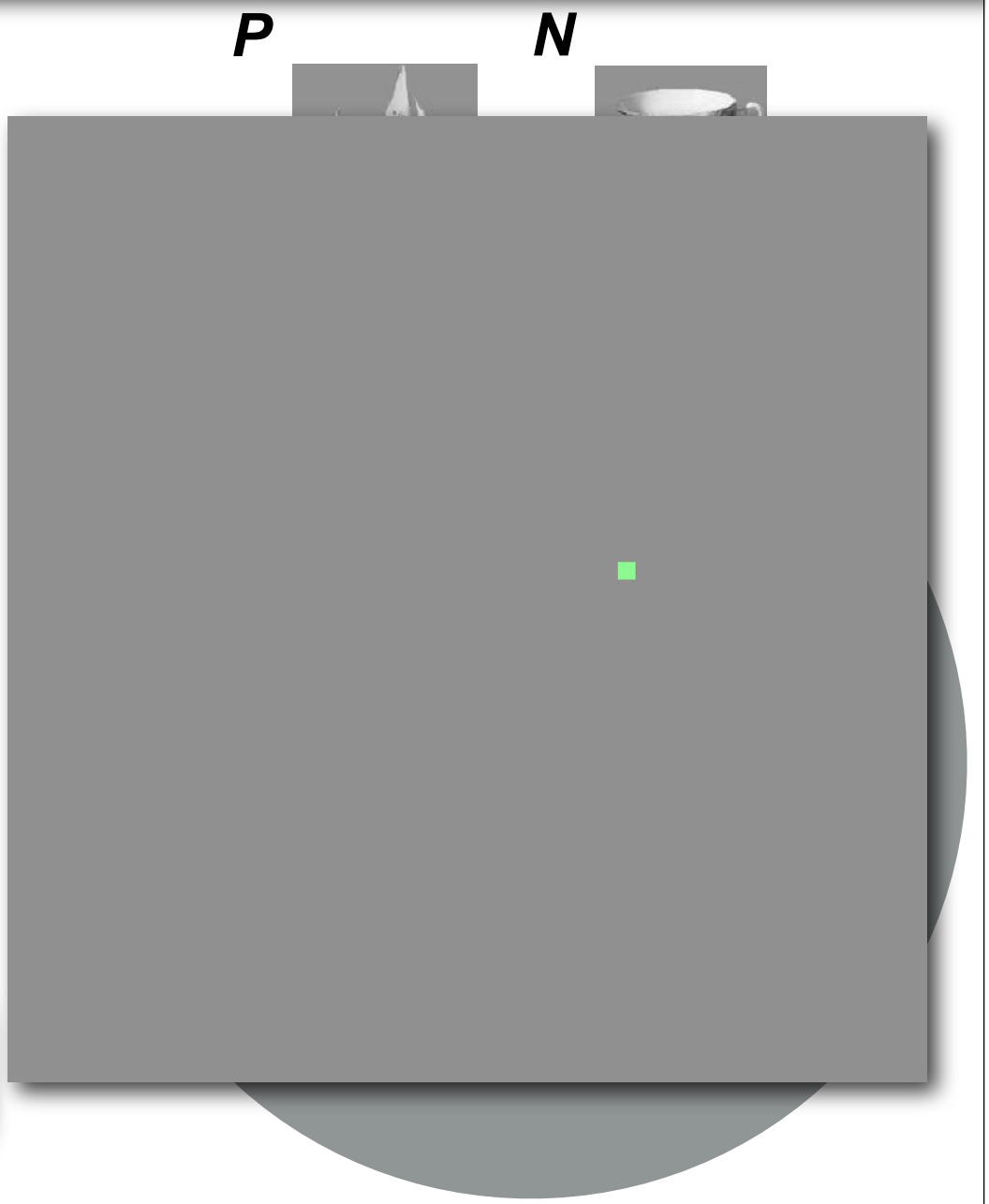


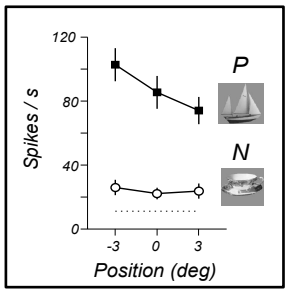
Screen → **Test phase** → **Exposure phase**
100 swaps, ~15 min

Experimental design (within-neuron, longitudinal design)

Real-time eye tracking and stimulus control

Unsupervised !

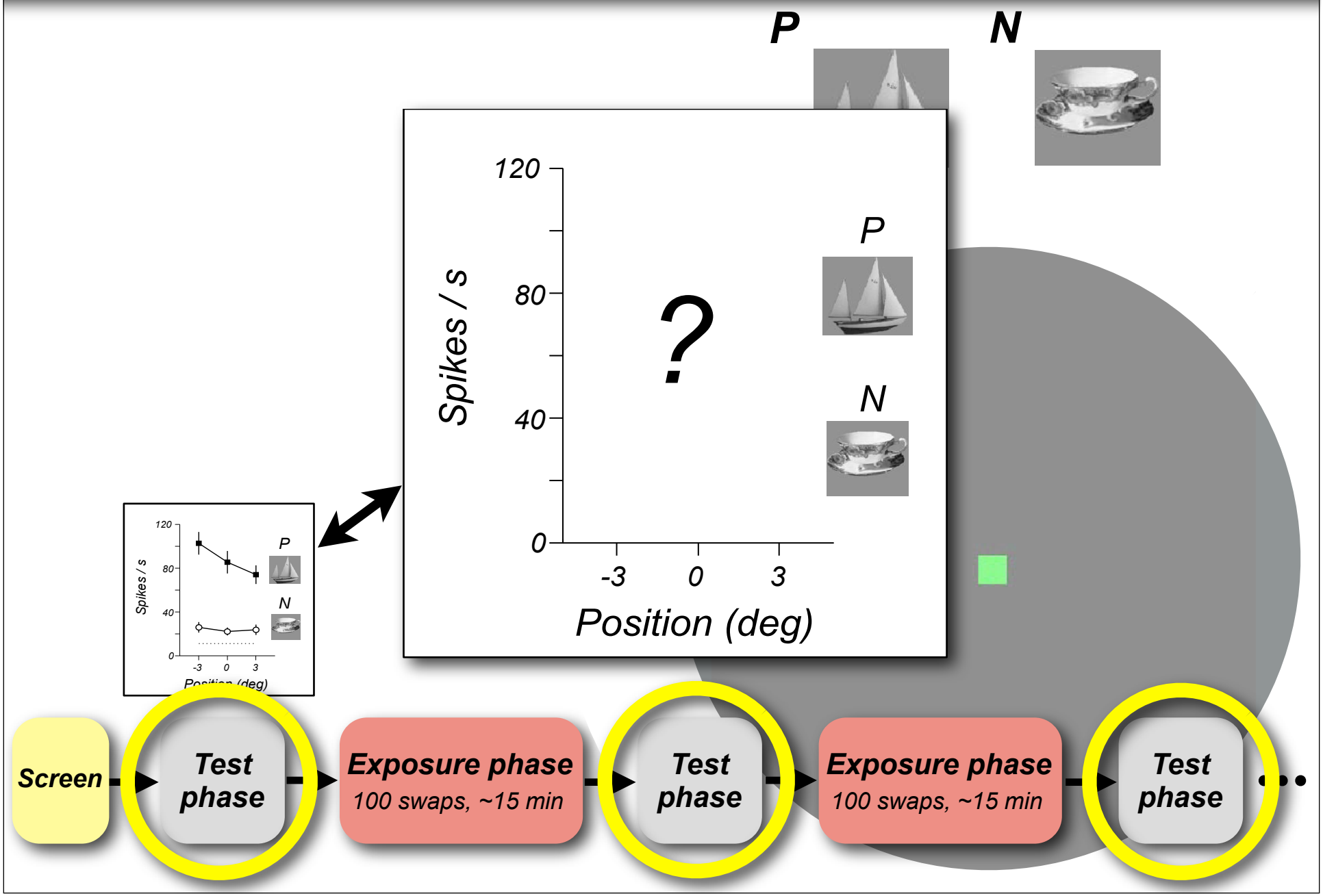




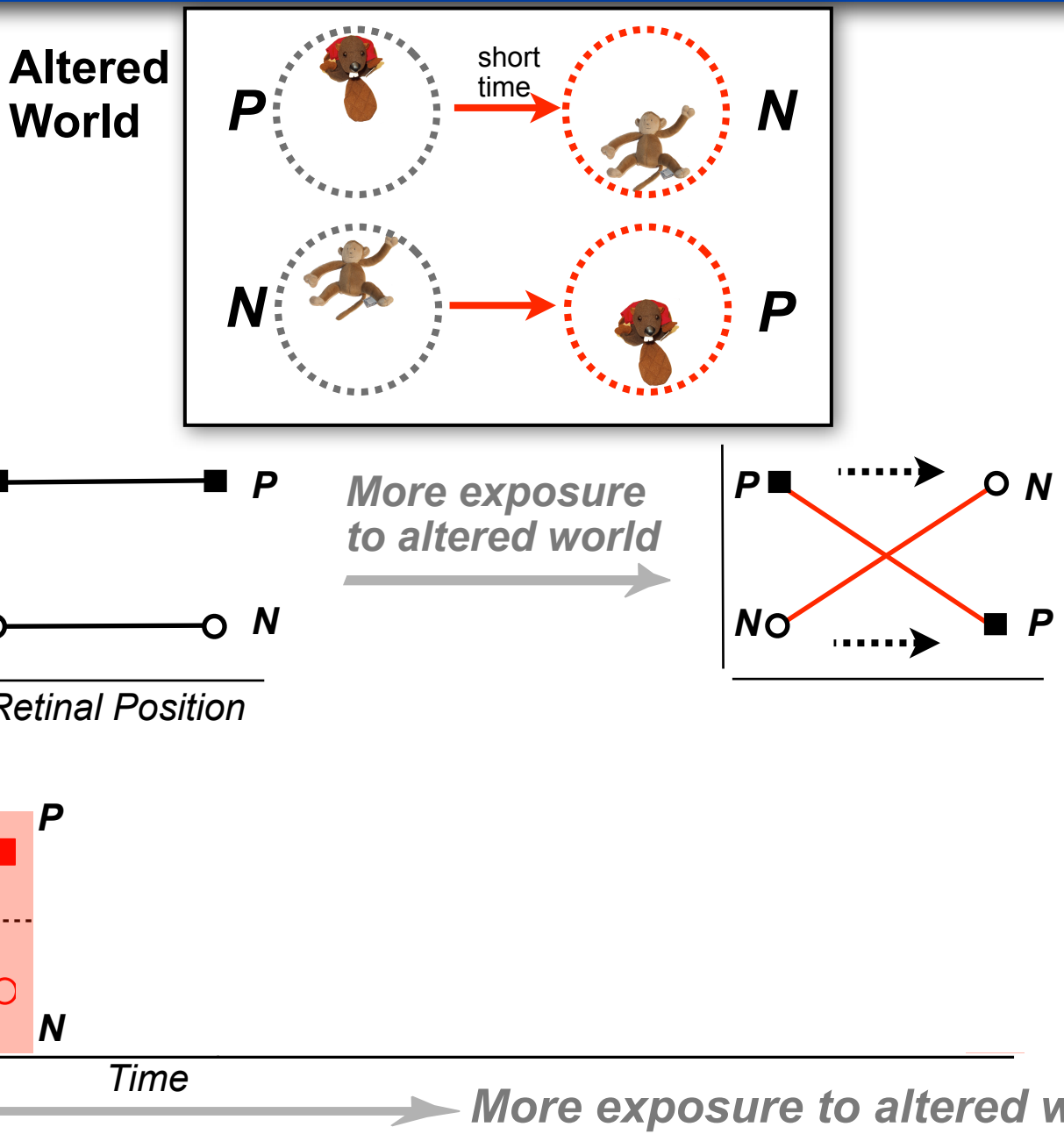
Screen → ***Test phase*** → ***Exposure phase***

100 swaps, ~15 min

Experimental design (within-neuron, longitudinal design)

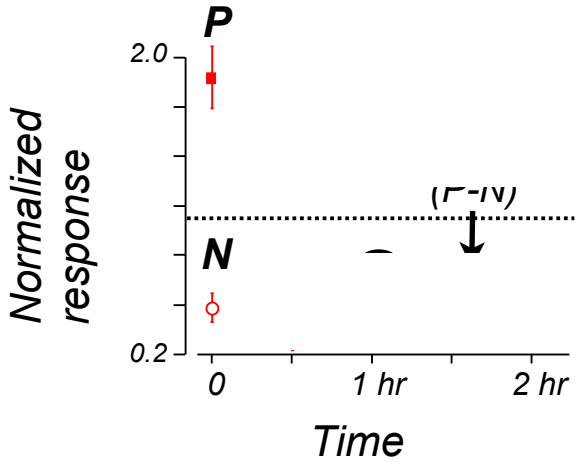


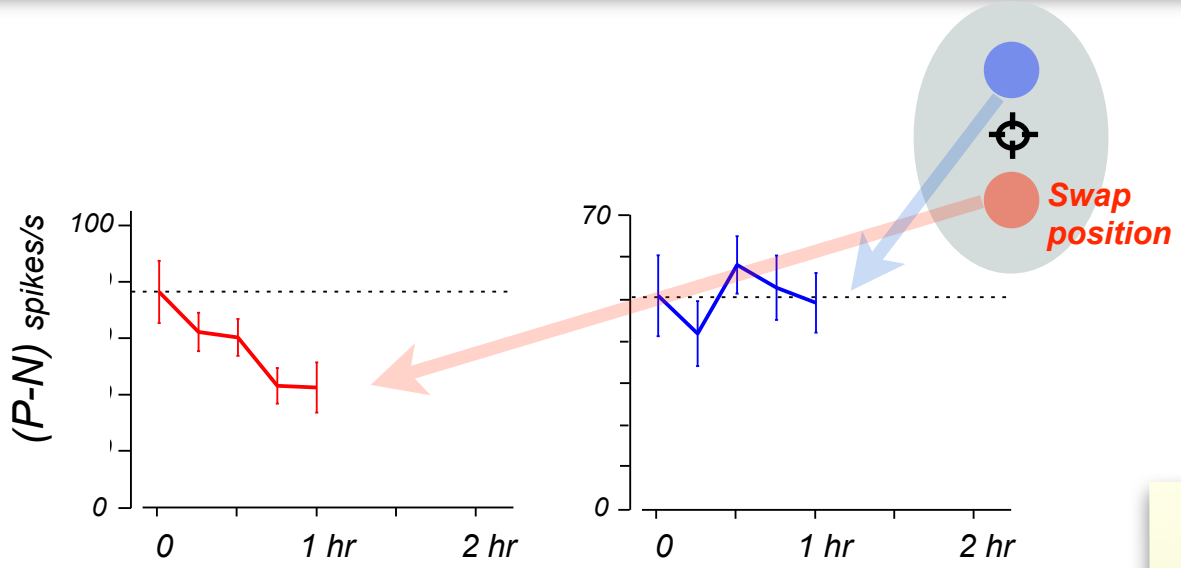
Prediction


Altered World
short time
P
N
N
P
IT Response
P
N
Retinal Position
More exposure to altered world
P
N
N
P
Swap position
P
N
Time
More exposure to altered world

Results: IT single unit activity

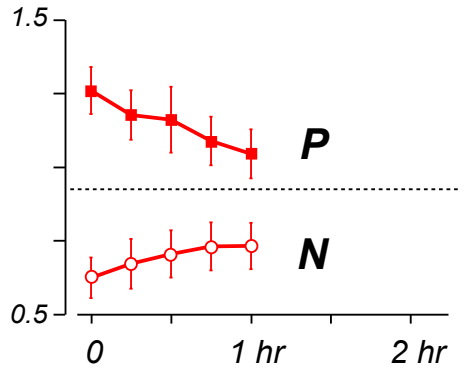
Single IT neuron

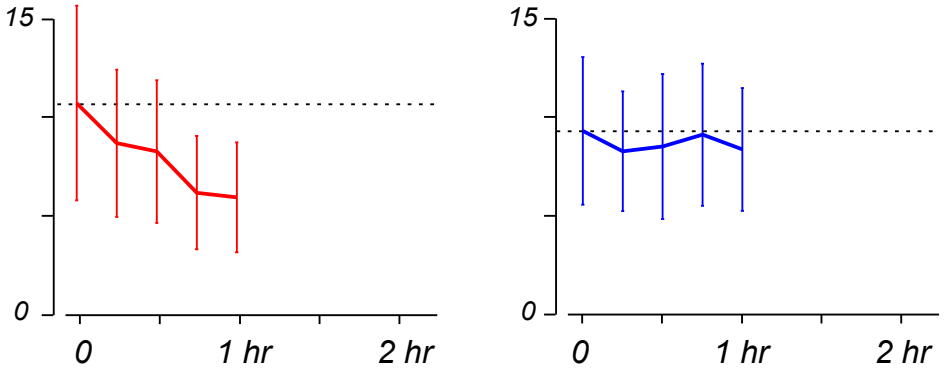




p=0.007 (n=101**, "Position x Exposure" interaction)

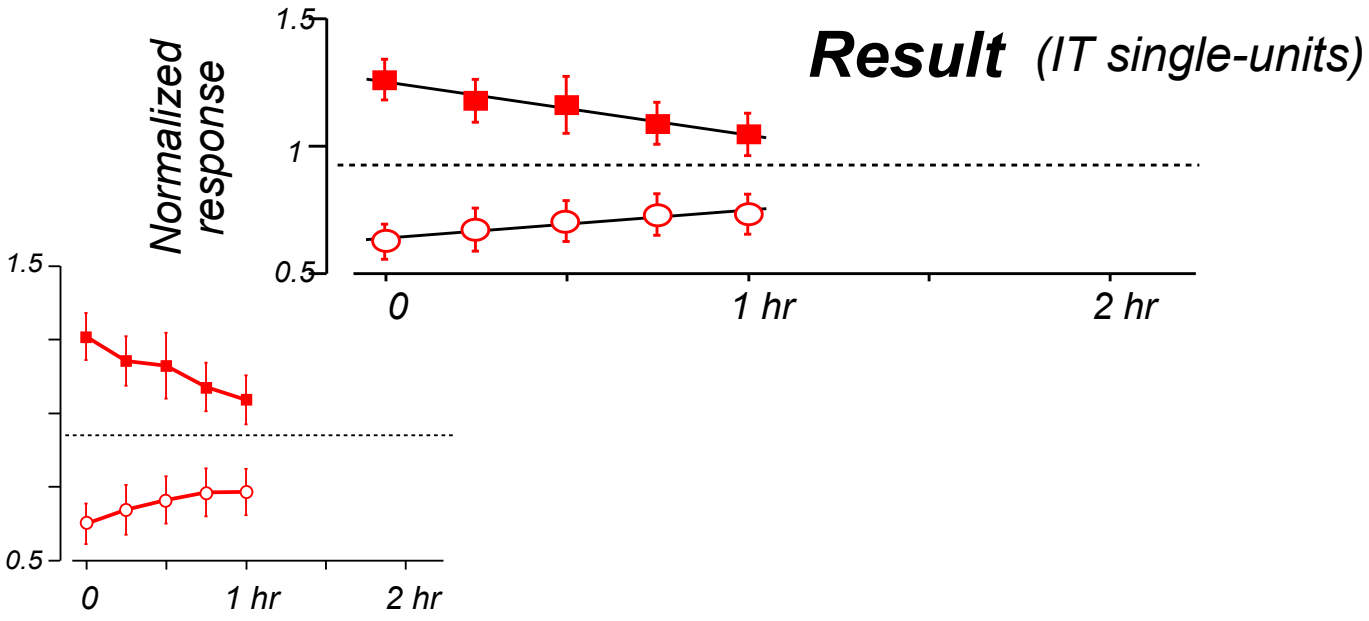
IT Population (pooled)





*Li and DiCarlo, **Science** (2008)*

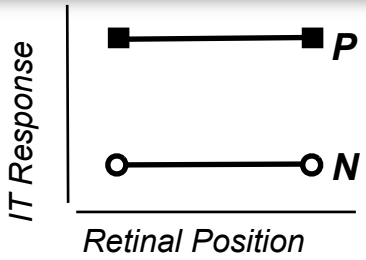
Results: IT single unit activity



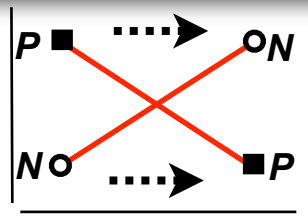
Result *(IT single-units)*

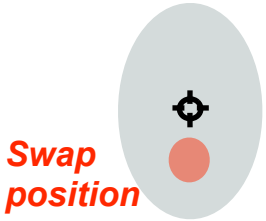
Li and DiCarlo, ***Science*** *(2008)*

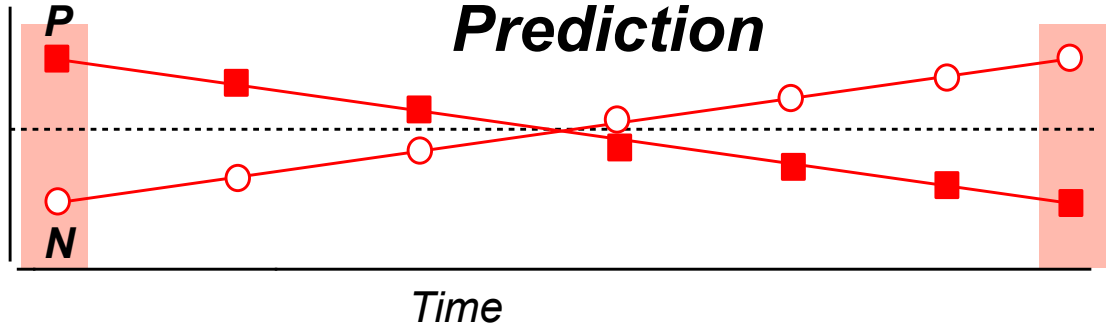
Results: pooled population data

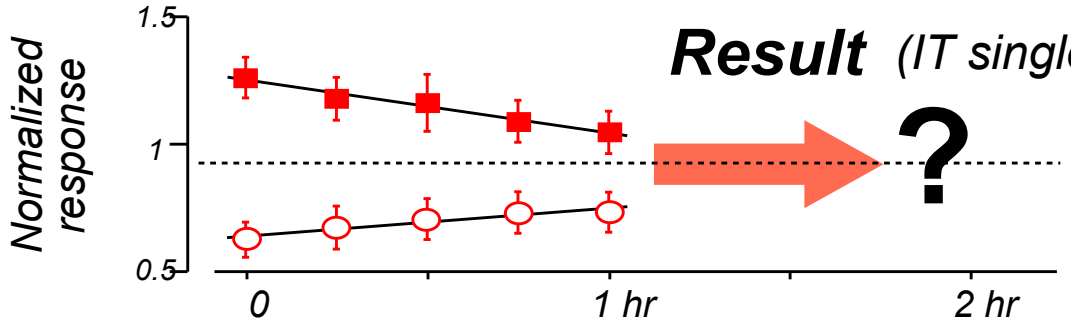


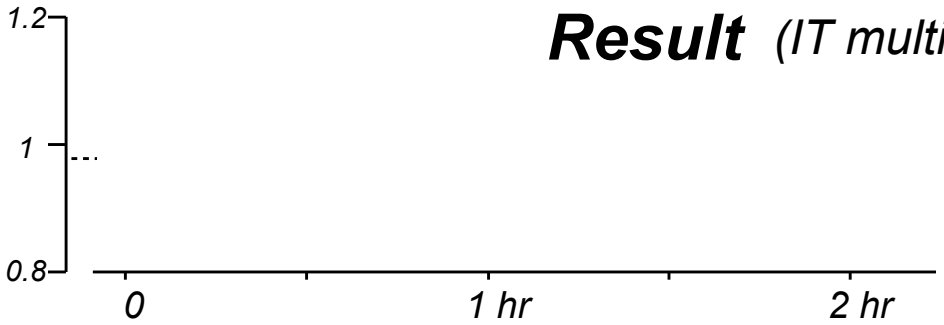
More exposure to altered world











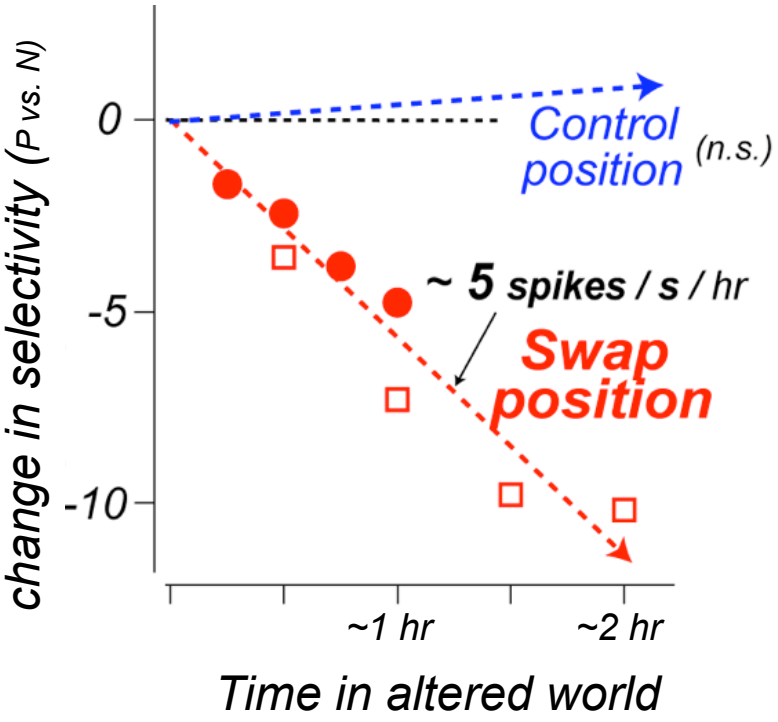
Li and DiCarlo, **Science** (2008)

Effects of experience in visual worlds with specific alterations in temporal contiguity of object images

- ***Monkey IT neurons:***

  highly specific change in position tolerant object selectivity

  Li and DiCarlo, ***Science*** *(2008)*



- ***Very strong:***

  - effect size is as large or larger than long-term IT plasticity studies

  (e.g. Baker et al 2003; Sigala & Logothetis 2000; Kobetake et al 1998; Cox and DiCarlo, 2008)

  - significant at individual IT sites

Can study learning "online"

More important facts about this result ...

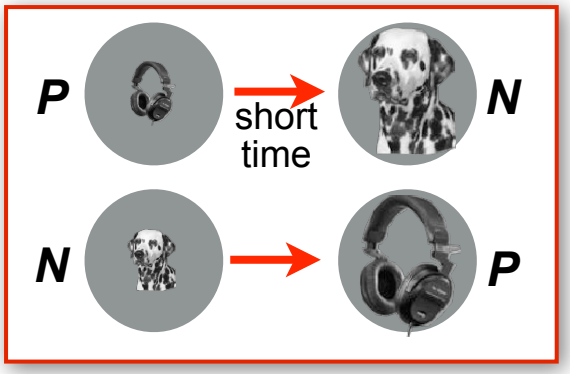
- *Cannot be explained as attention or adaptation*
- *Likely reflects changes in feed-forward processing*
- *Is object/feature specific*
- *IT changes last at least weeks (indirect, preliminary)*
- *May share underlying mechanism with "paired associates" learning* (Miyashita 1988, 1991; Erikson et al 1999; Messinger et al. 2001)
- *... but is unsupervised and on time scale of natural vision*

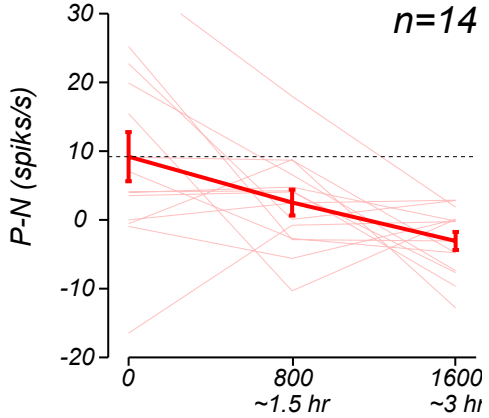
"Unsupervised temporal tolerance learning" (UTL)

Inference: temporal contiguity of visual experience "instructs" explicit ("invariant") object representation

- ***Future experiments: characterization of UTL***
  - *role of saccades? size, pose tolerance ? animal perception? stability? development?*

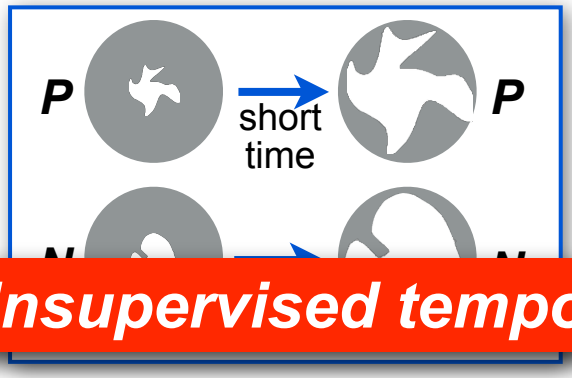
Swap exposure

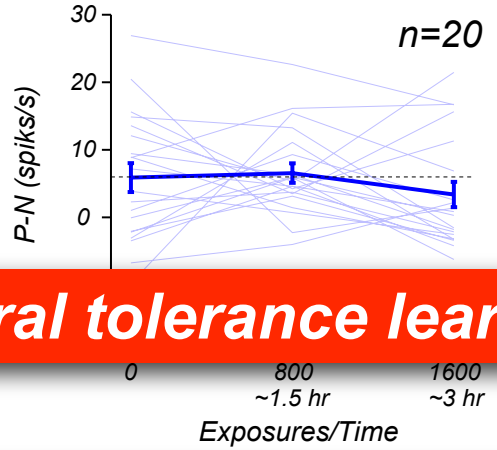




UTL appears to extend to size tolerance of IT neurons

Control



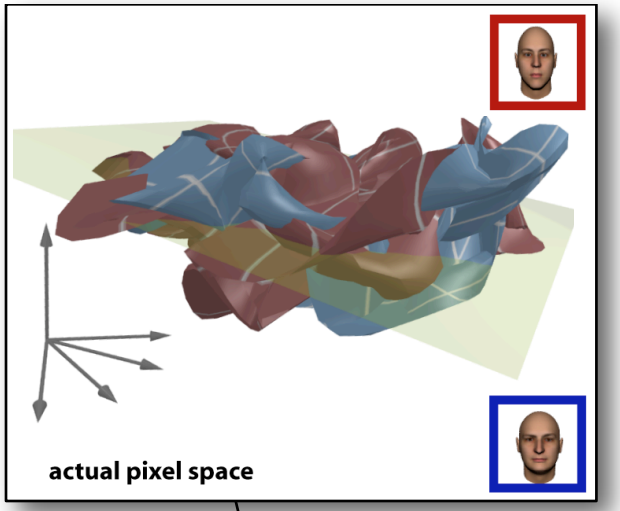


"Unsupervised temporal tolerance learning" (UTL)

Inference: temporal contiguity of visual experience "instructs" explicit ("invariant") object representation

Summary

Adult learning --> same mechanisms as infant?

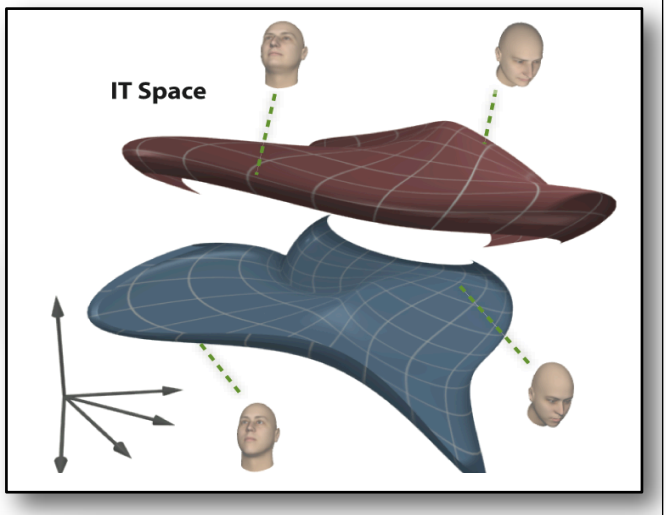


?

temporal contiguity

unsupervised

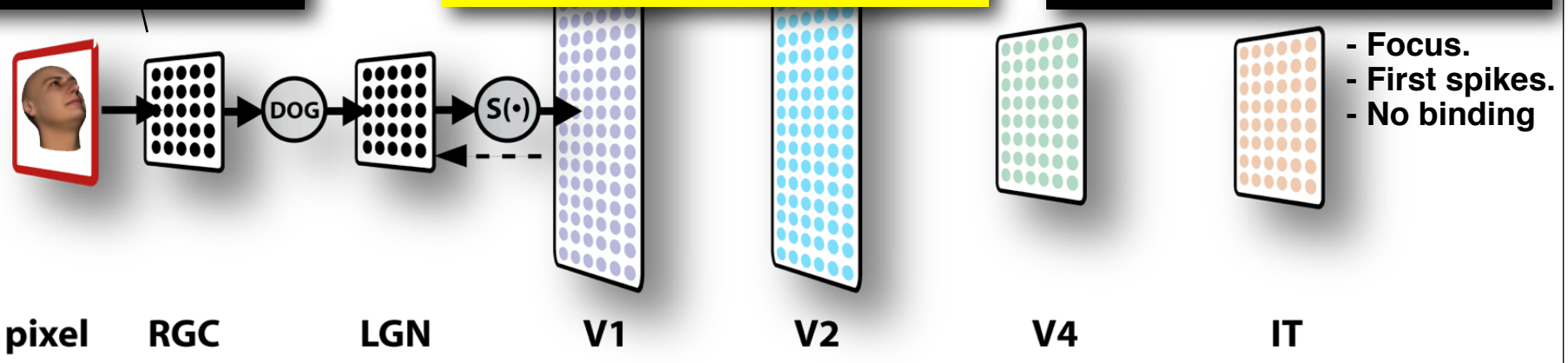
learned



Tangled object information

How does this untangling happen ???

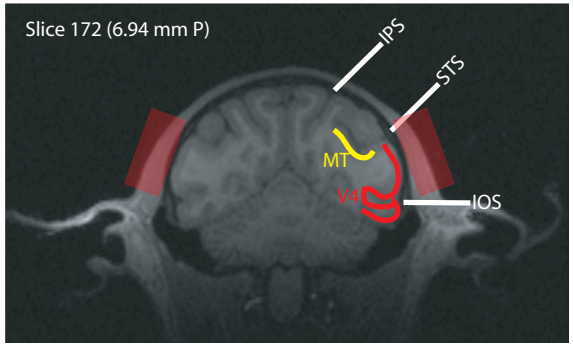
Explicit, untangled object representation



- Focus.
- First spikes.
- No binding

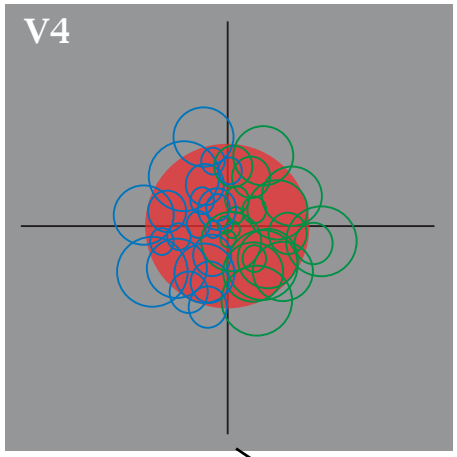
Direct population comparison of the V4 and IT representation

Nicole Rust

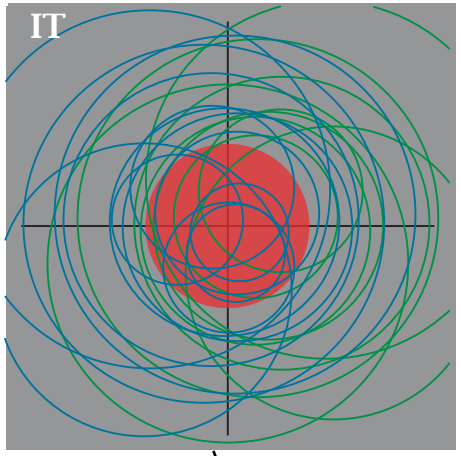


Same animal, task, stimuli.

140 V4 neurons

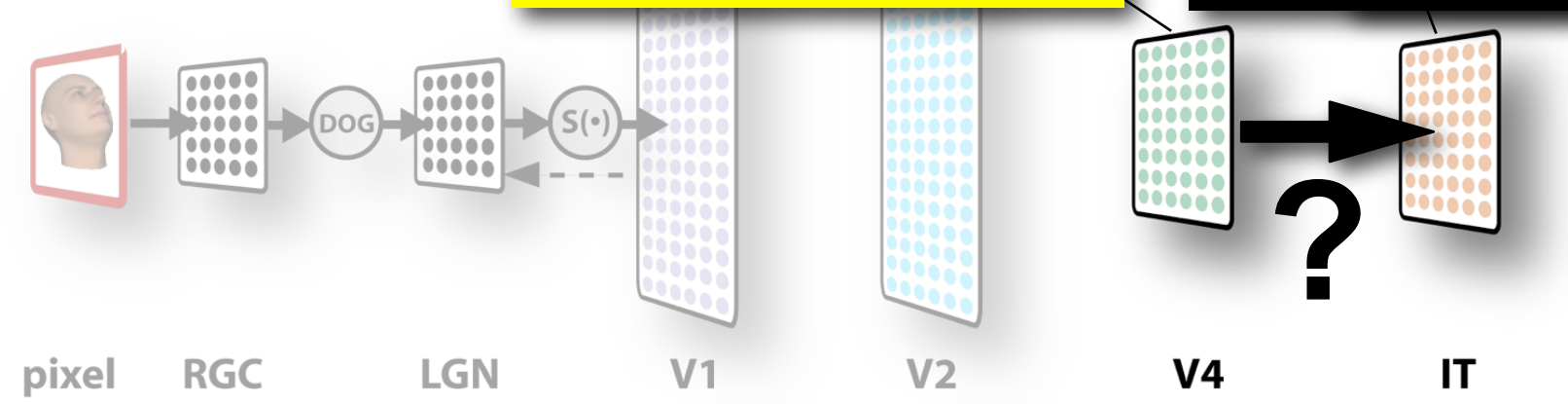


140 IT neurons

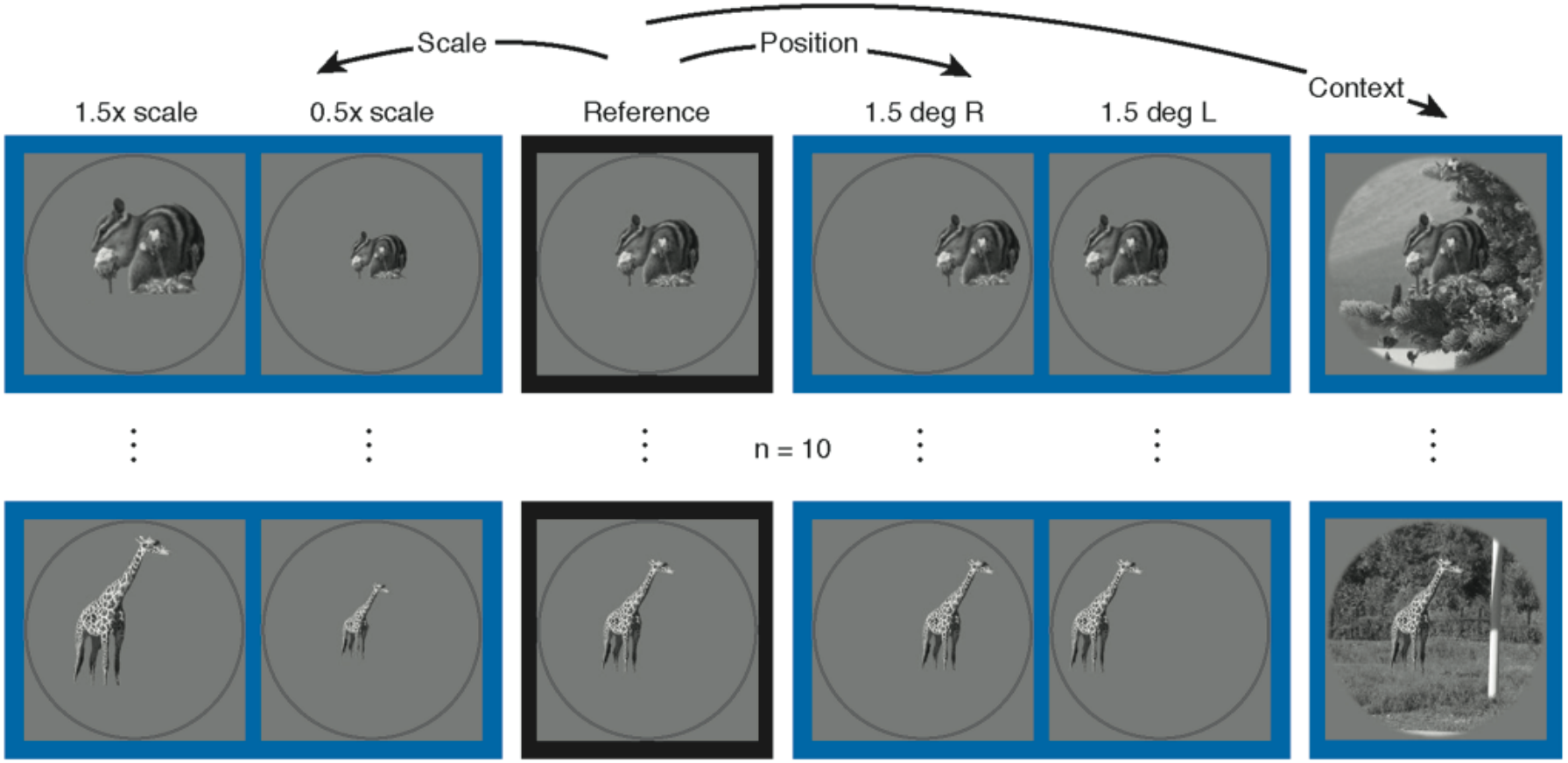


How does this untangling happen ???

Explicit, untangled object representation



How well can each population (V4 vs. IT) retain object selectivity?



Rust & DiCarlo (in prep)

Result: the IT representation improves upon the V4 representation

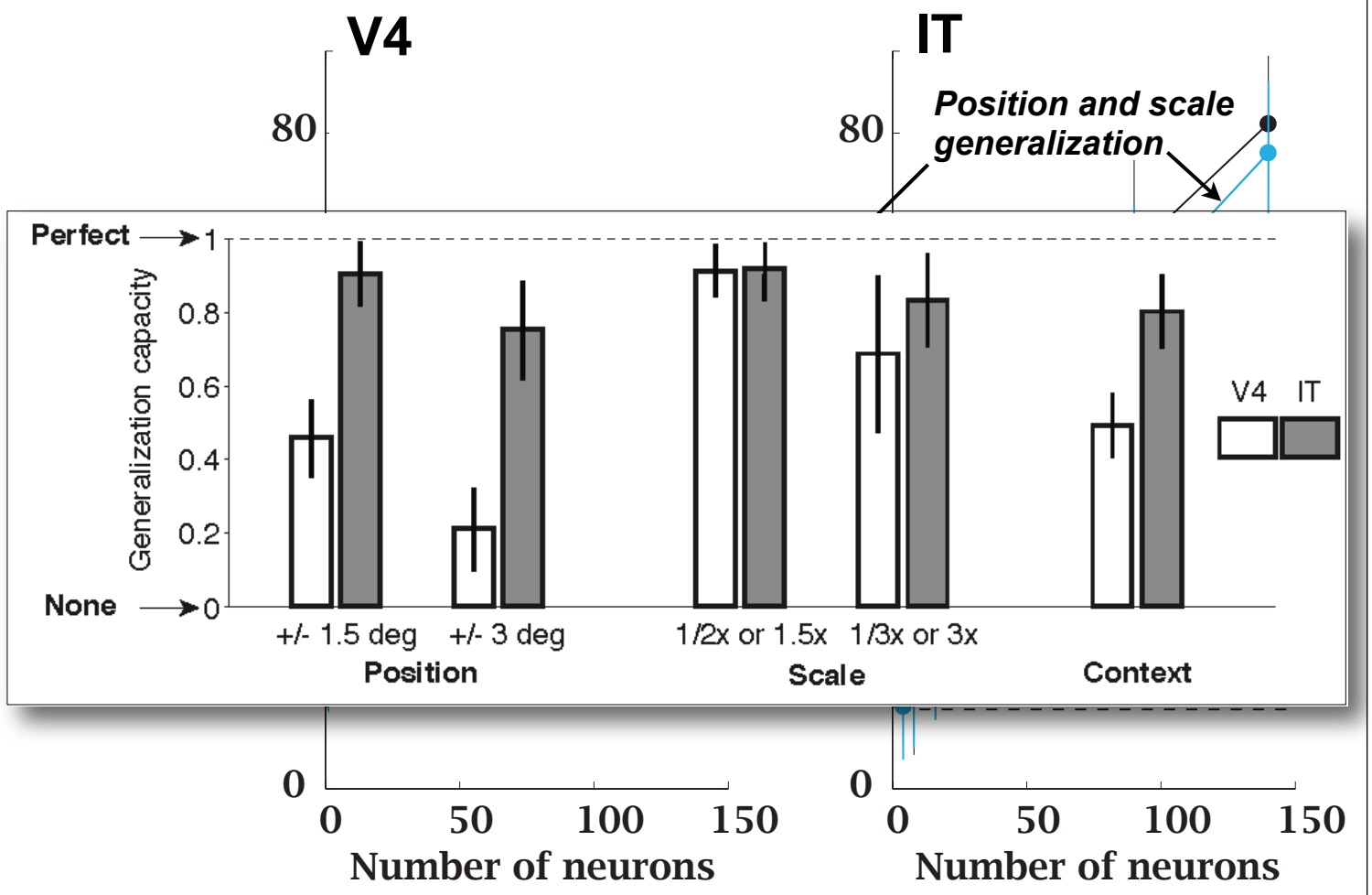
Ref.

Big

Small

Right

Left

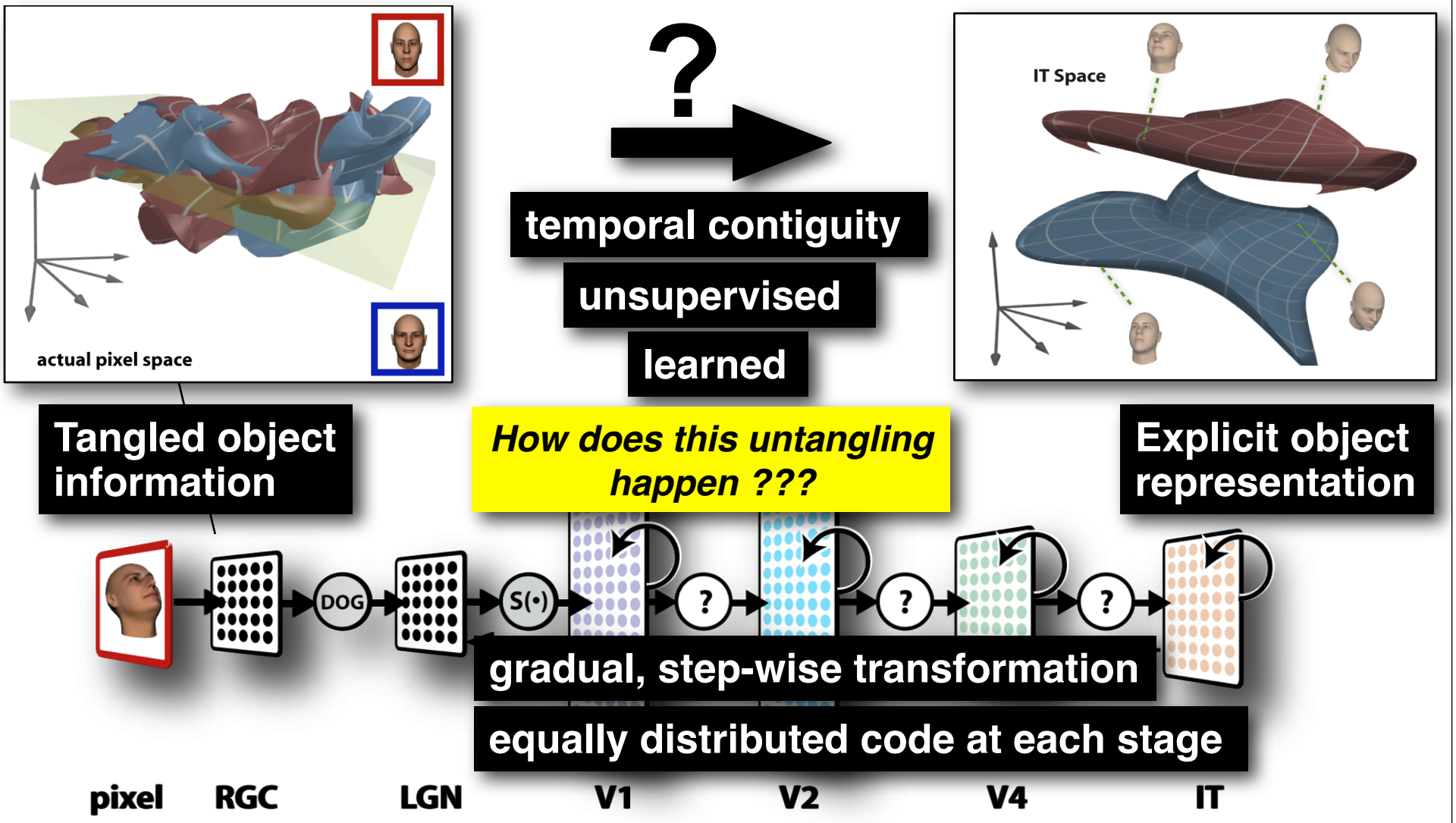


***Not** due to lack of information*
(IT and V4 convey roughly the same amount of natural-image information)

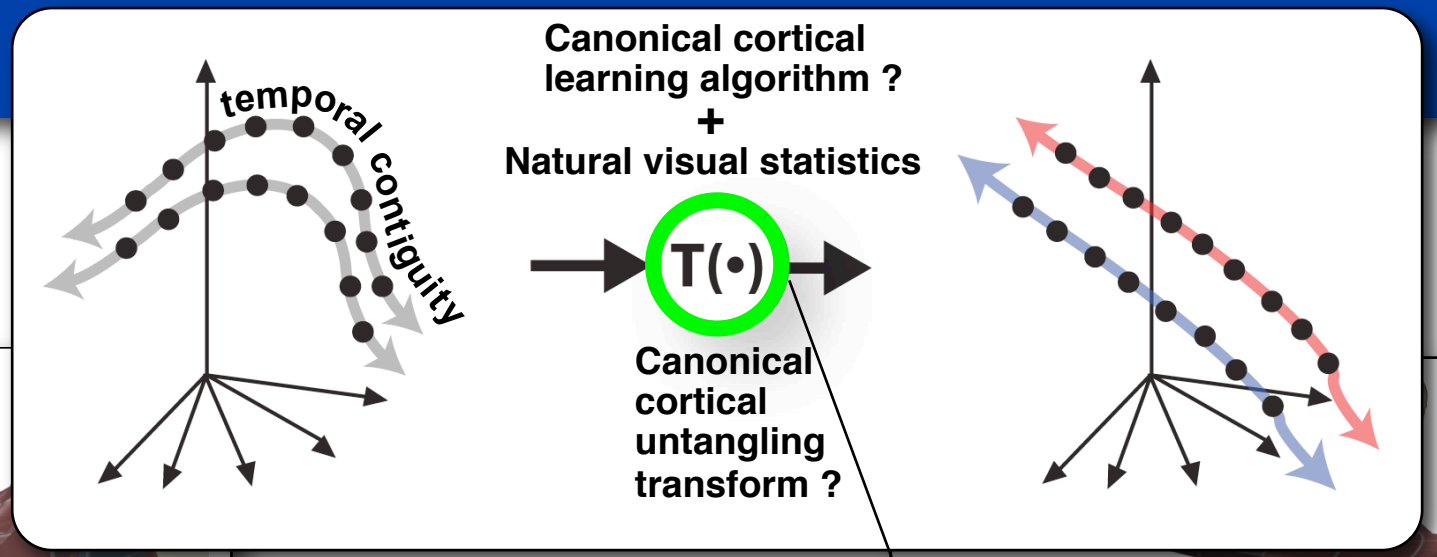
***Not** simply explained by larger RF sizes in IT*

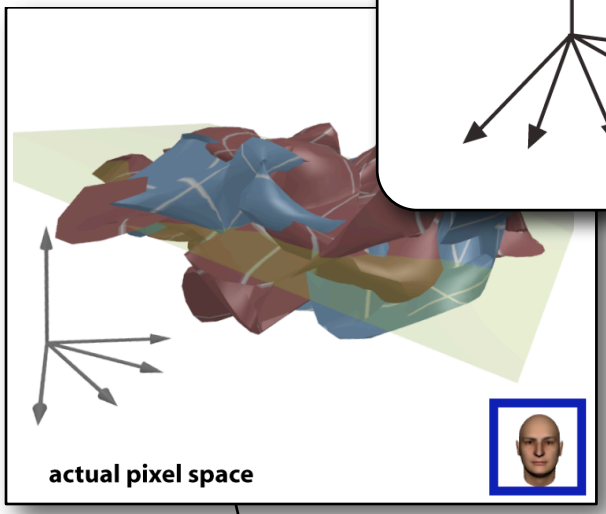
Rust & DiCarlo (in prep)

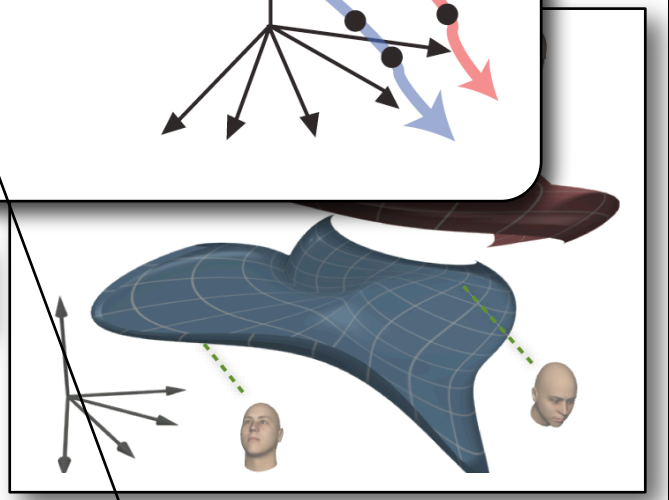
Summary


actual pixel space
?
temporal contiguity
unsupervised
learned
IT Space
Tangled object information
How does this untangling happen ???
Explicit object representation
DOG
S(•)
?
?
?
gradual, step-wise transformation
equally distributed code at each stage
pixel
RGC
LGN
V1
V2
V4
IT

Summary







temporal contiguity

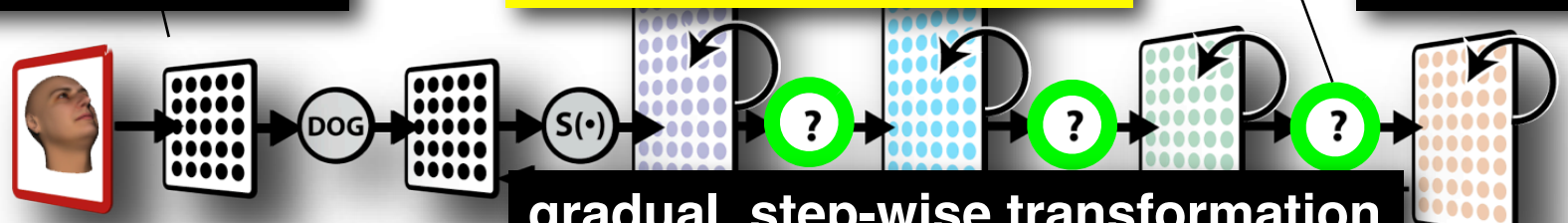
unsupervised

learned

Tangled object information

How does this untangling happen ???

Explicit object representation



gradual, step-wise transformation

equally distributed code at each stage

pixel RGC LGN V1 V2 V4 IT

Our current directions...

"Unsupervised temporal tolerance learning" (UTL)

- ***Experiments: characterization of UTL***
  - *size, pose tolerance ? saccade? attention? animal perception? stability? development?*

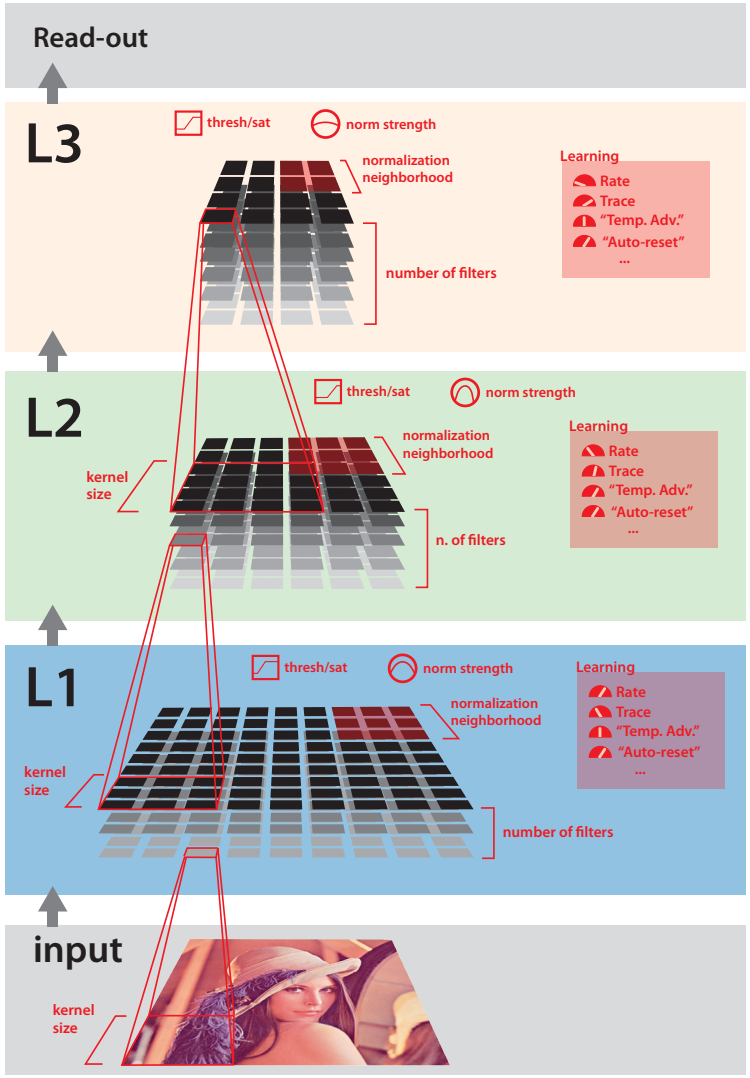
Our current directions...

"Unsupervised temporal tolerance learning" (UTL)

- ***Experiments: characterization of UTL***
  - *size, pose tolerance ? saccade? attention? animal perception? stability? development?*
- ***Search for cortical untangling transforms***

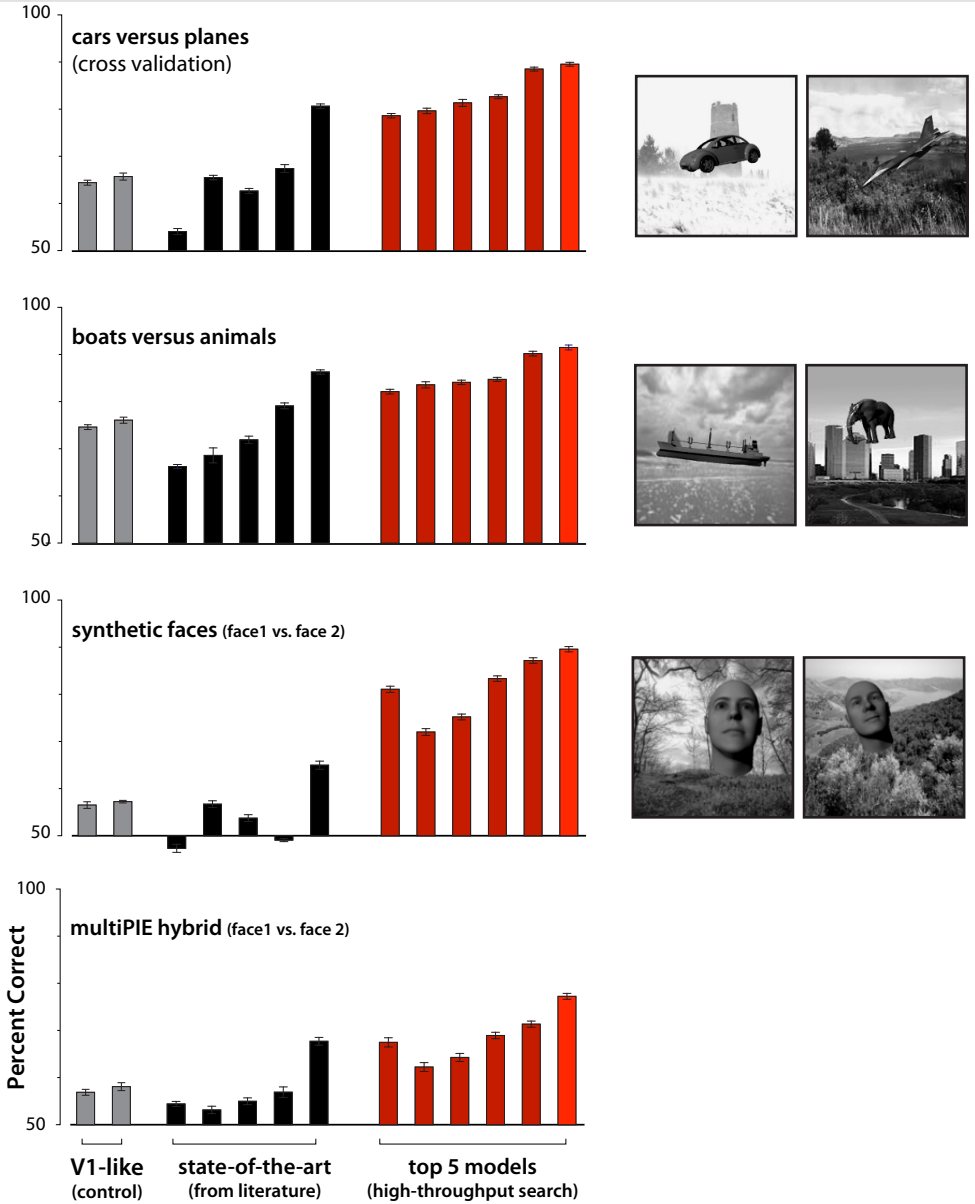
Dave Cox

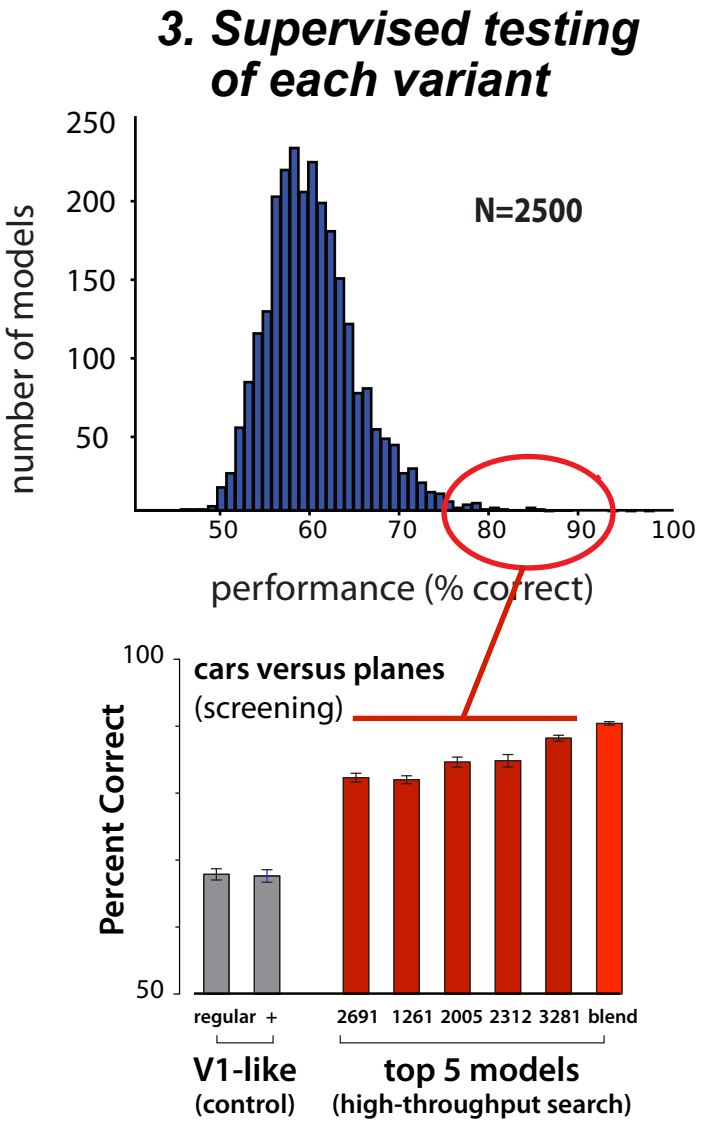
Nicolas Pinto



Pinto, Cox & DiCarlo, PLoS Comp Bol (2008), COSYNE (2008)
Pinto, DiCarlo and Cox, ECCV (2008); Pinto, DiCarlo & Cox (submitted)

High-throughput search for a cortical untangling transform





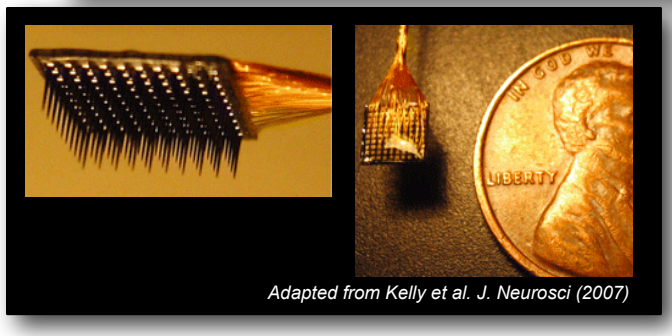
Pinto, Cox & DiCarlo, PLoS Comp Bol (2008), COSYNE (2008)
Pinto, DiCarlo and Cox, ECCV (2008); Pinto, DiCarlo & Cox (submitted)

Our current directions...

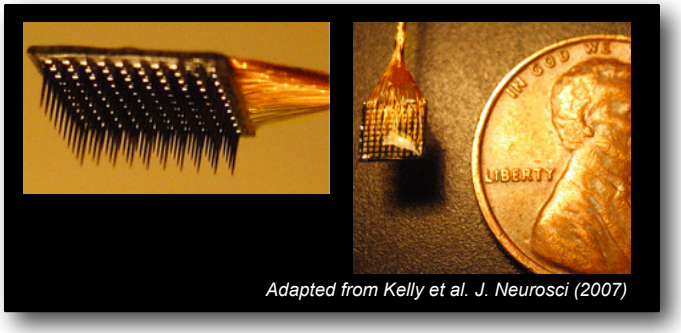
"Unsupervised temporal tolerance learning" (UTL)

- ***Experiments: characterization of UTL***
  - *size, pose tolerance ? saccade? attention? animal perception? stability? development?*
- ***Search for cortical untangling transforms***

Focus on understanding the building of invariance over smaller cortical steps



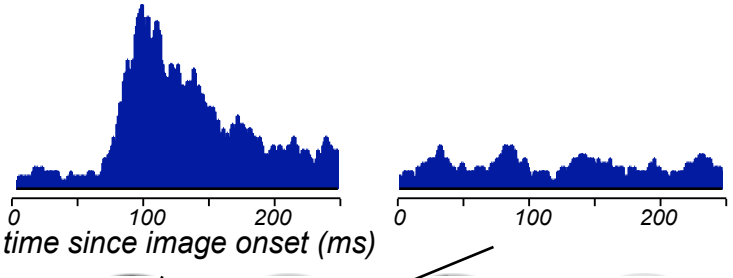
Our current directions...

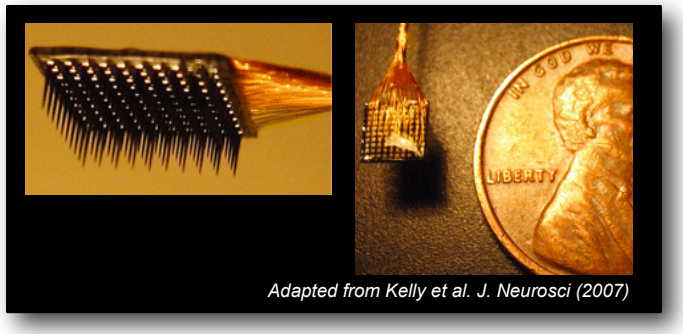


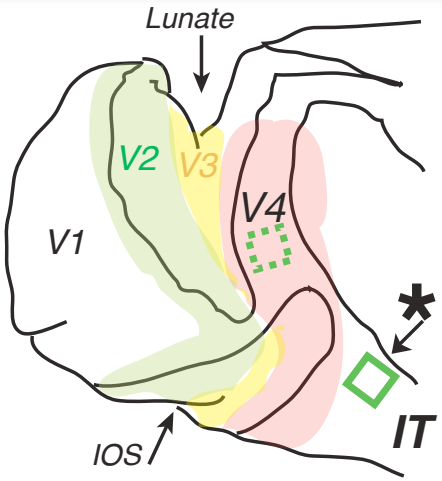
Adapted from Kelly et al. J. Neurosci (2007)

DiCarlo Lab (unpublished)

Our current directions...

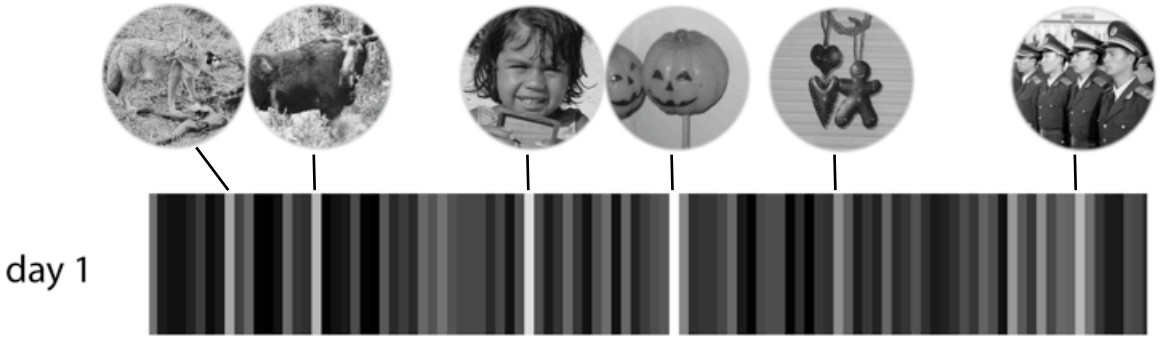


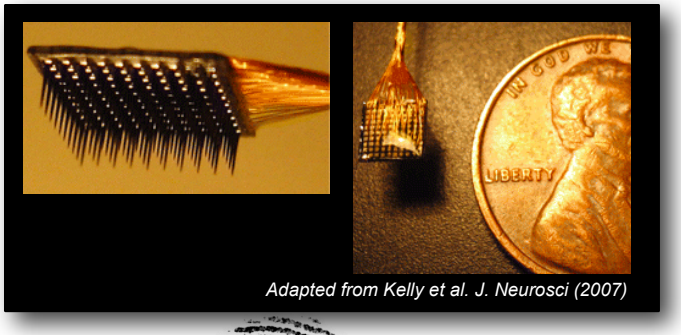




DiCarlo Lab (unpublished)

Our current directions...





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

DiCarlo Lab (unpublished)

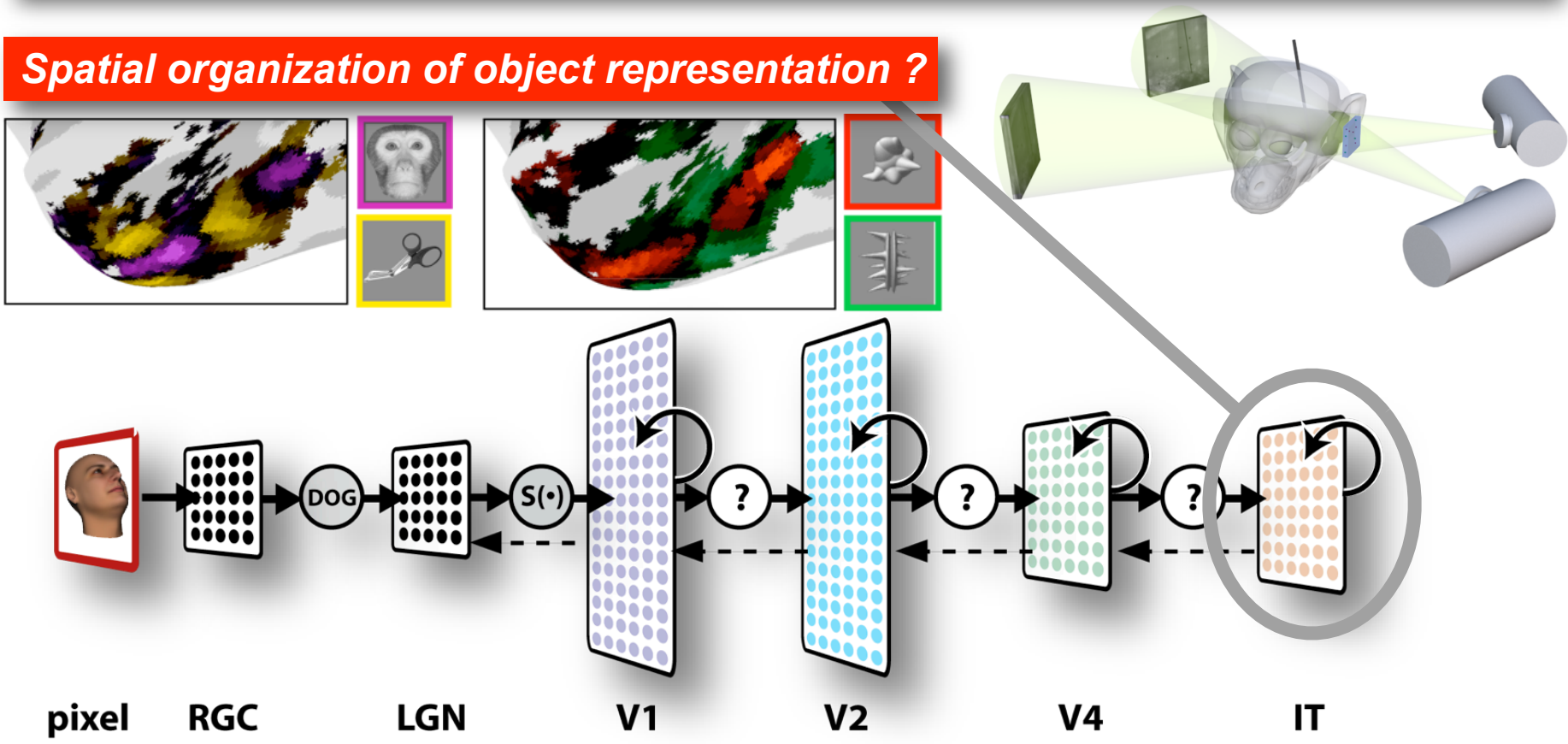
Our current directions...

"Unsupervised temporal tolerance learning" (UTL)

- ***Experiments: characterization of UTL***
  - *size, pose tolerance ? saccade? attention? animal perception? stability? development?*
- ***Test real-world performance of these ideas***

Focus on understanding the building of invariance over smaller cortical steps

Spatial organization of object representation ?

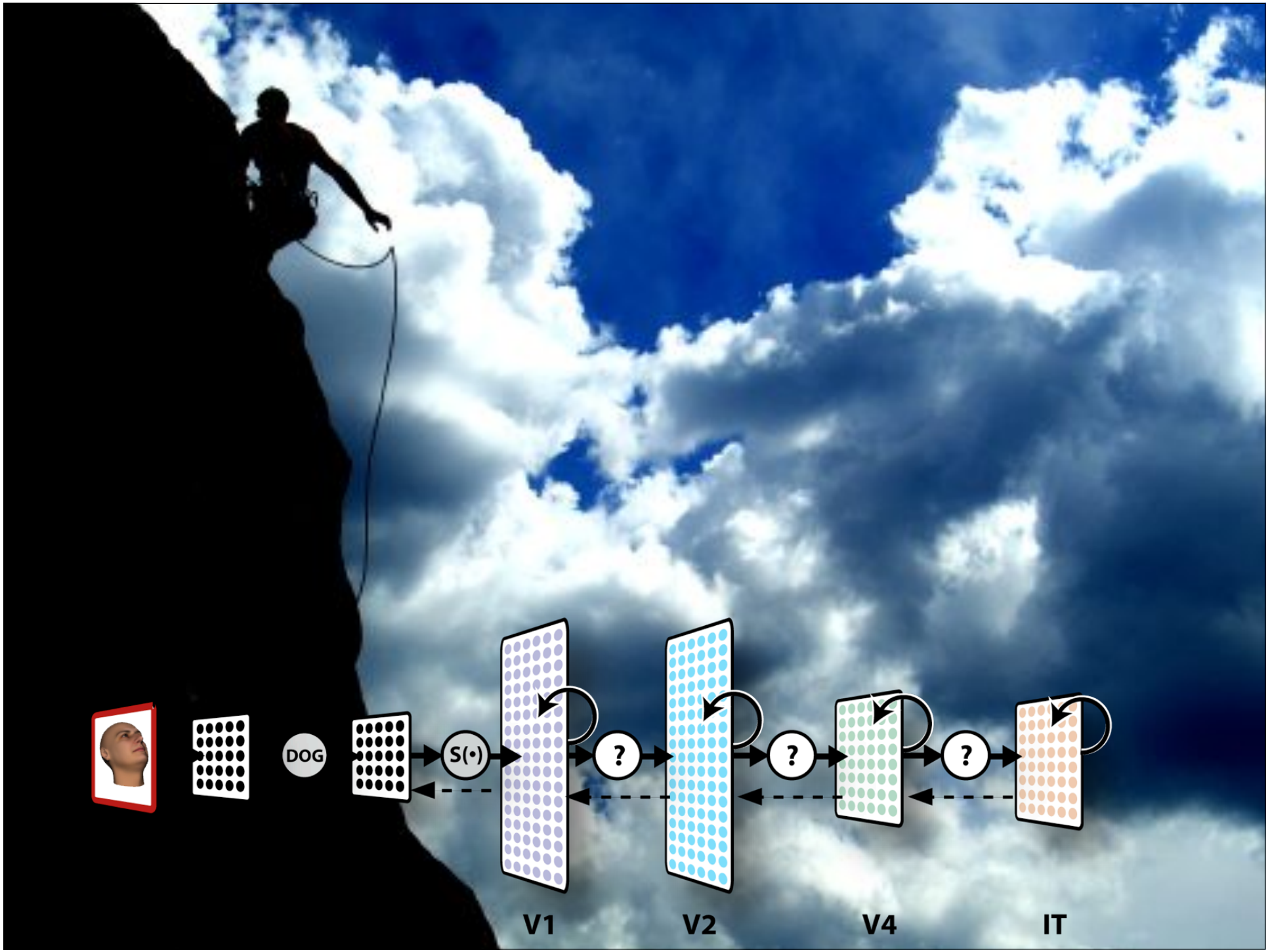


4. View of the solution

3. New empirical results

2. Data transformations

1. Focus on the crux problem


DOG
S(•)
?
?
?
V1
V2
V4
IT

Acknowledgements

DiCarlo Lab

Chou Hung

Gabriel Kreiman

David Cox

Nuo Li

Nicolas Pinto

Nicole Rust

MIT:

Tomaso Poggio
Nancy Kanwisher
--
Paul Aparicio
Jennifer Deutsch
Elias Issa
Najib Majaj
Marie Maloof
Davide Zoccolan
--
Ben Andken
Philip Meier
Nadja Oertelt
Hans Op de Beeck
Dan Oreper
Alex Papanastassiou

MGH:

Wim Vanduffel

Caltech:

Christof Koch
Rodrigo Quiroga

Harvard/HHMI:

John Maunsell

- *NIH NEI*
- *NIH NIMH*
- *DARPA / ONR*
- *Pew Trust*
- *McKnight Foundation*
- *McGovern Institute*