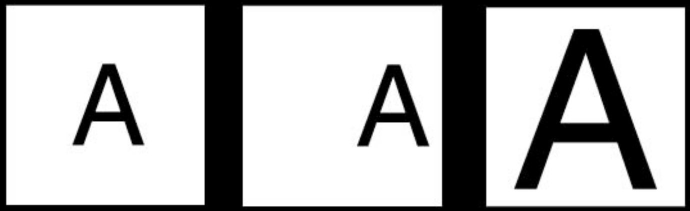
From feedforward vision to natural vision:

The impact of free viewing and clutter on monkey inferior temporal object representations

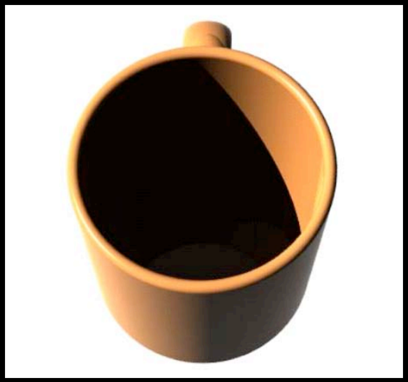
James DiCarlo
The McGovern Institute for Brain Research
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge MA

department of brain and cognitive sciences at MIT

The core problem of object recognition



- Position
- Size
- Pose
- Illumination
- “Clutter”
  - •□ Background scene
  - •□ Other objects



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

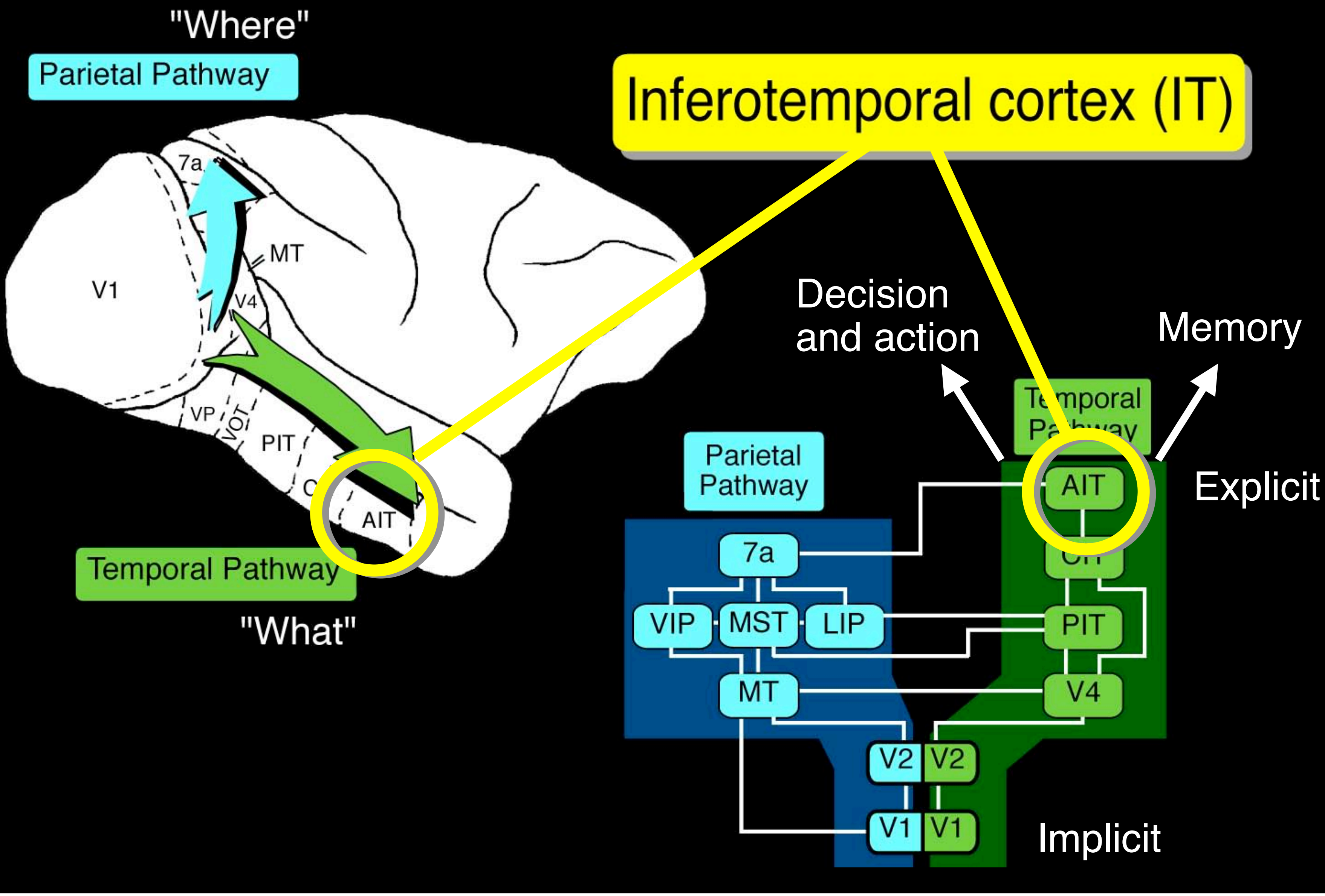
One needs an image representation that is selective for object identity, yet tolerant to such transformations.

Rhesus monkey model

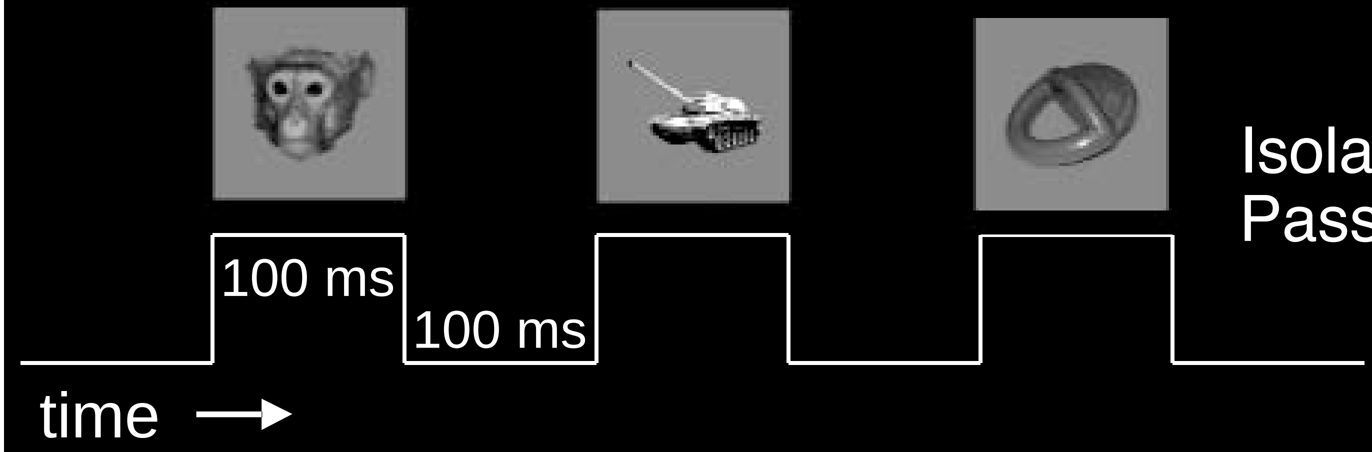
We have some idea of where we can find such an image representation (IT).

We can study it at the most appropriate level of abstraction (neuronal spikes).

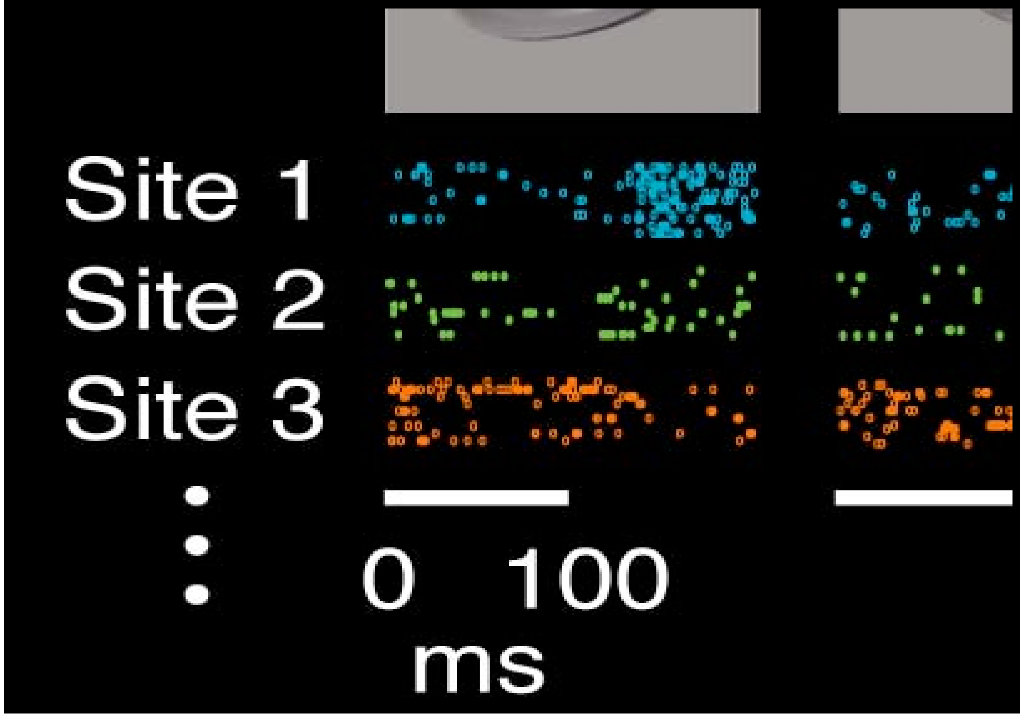
Monkey visual system



AIT contains a rapidly evoked, explicit object representation



Isolated, single objects.
Passive viewing.



Object identity or category is directly* available in the population response, regardless of (e.g.) object position and scale.

Hung, Kreiman, Poggio and DiCarlo Science (2005)

Gross et al. (1972), Perret et al. (1982), Desimone et al.. (1984), Tovee et al. (1984), Schwartz et al (1985), Ito et al. (1995), Logethetis et al. (1996), Op de Beeck and Vogles (2000), DiCarlo and Maunsell (2000), etc.

Feedforward* representation (The Core)

* First evoked pattern of IT activity
when an image is presented to the eye

The Core is fast.

The Core is powerful.

The Core is not yet understood.

Mechanisms ?

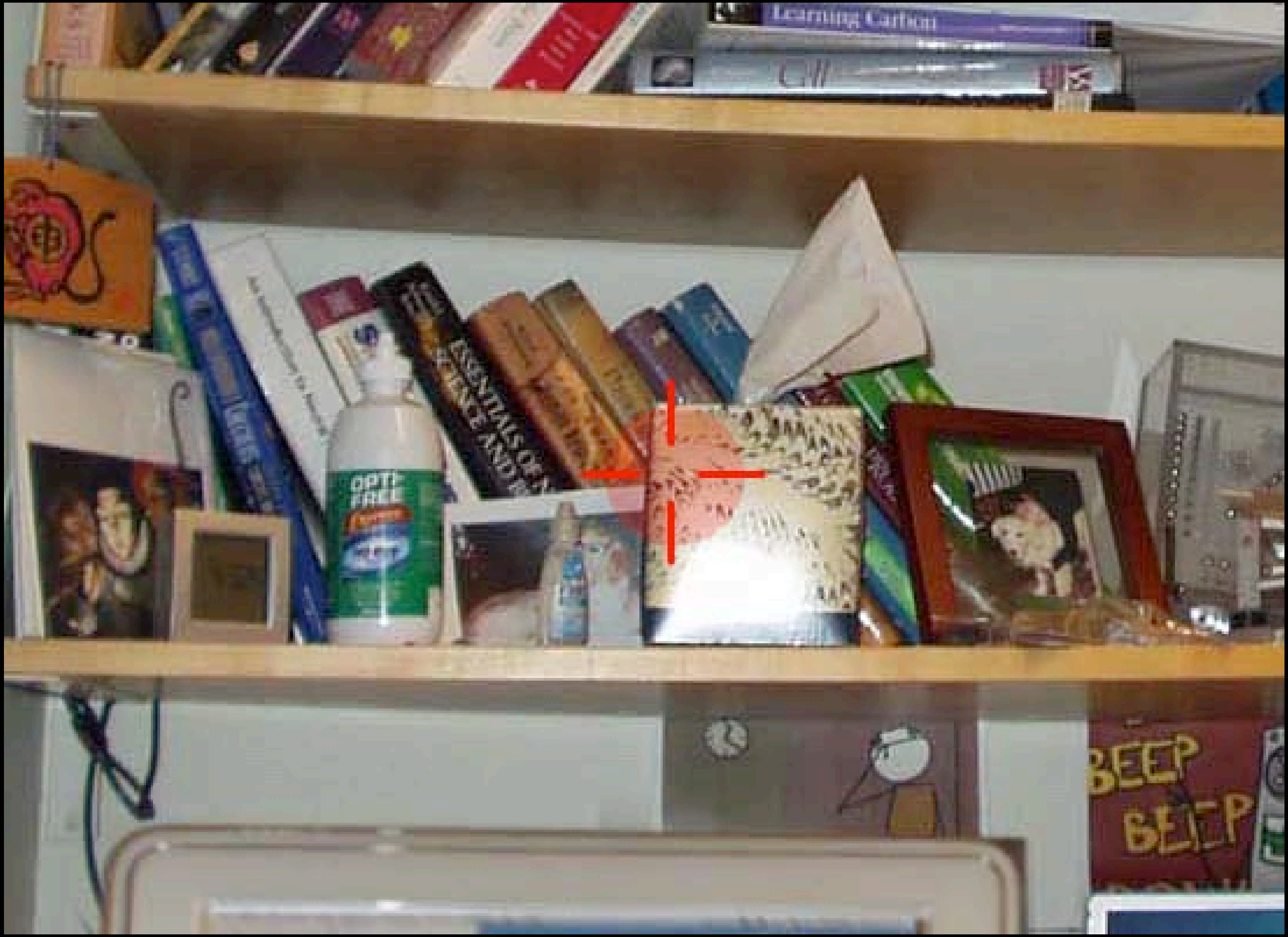
Role in “natural vision” ?
(Is it generalizable?)

The Core and “natural vision”

What is “natural vision” ?

“You know it when you see it.”

The Core and “natural vision”



The Core and “natural vision”

How does “natural vision” challenge the basic model of core vision?

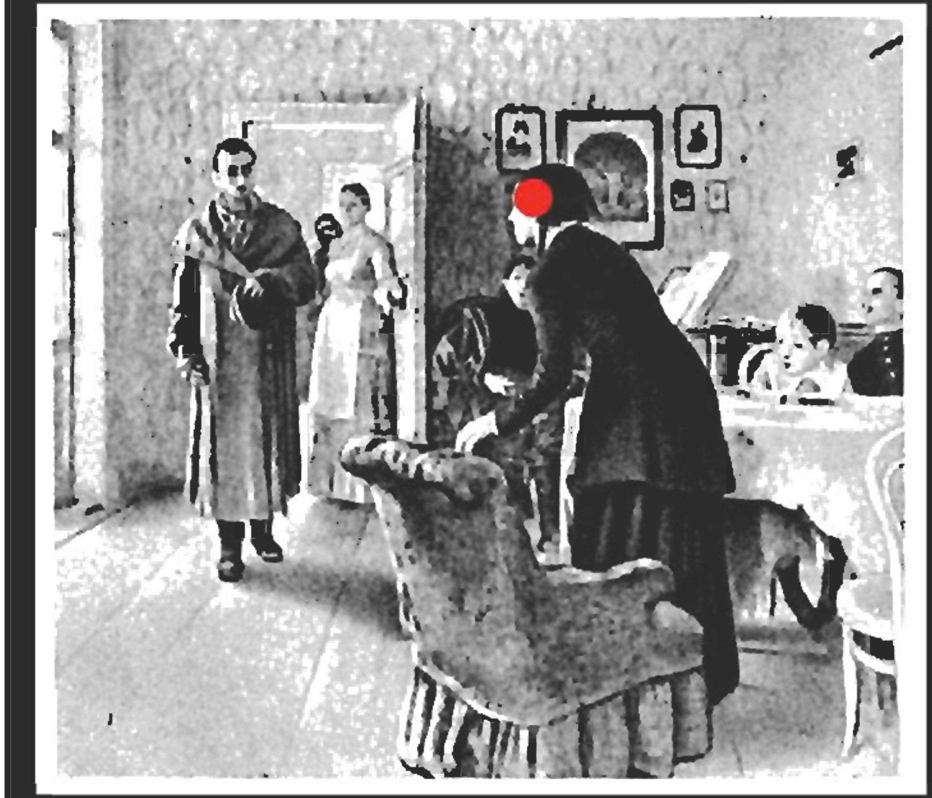
1) Eye movements (“free viewing”)

► 2) Clutter / Scene / Context: objects appear among other objects and on backgrounds

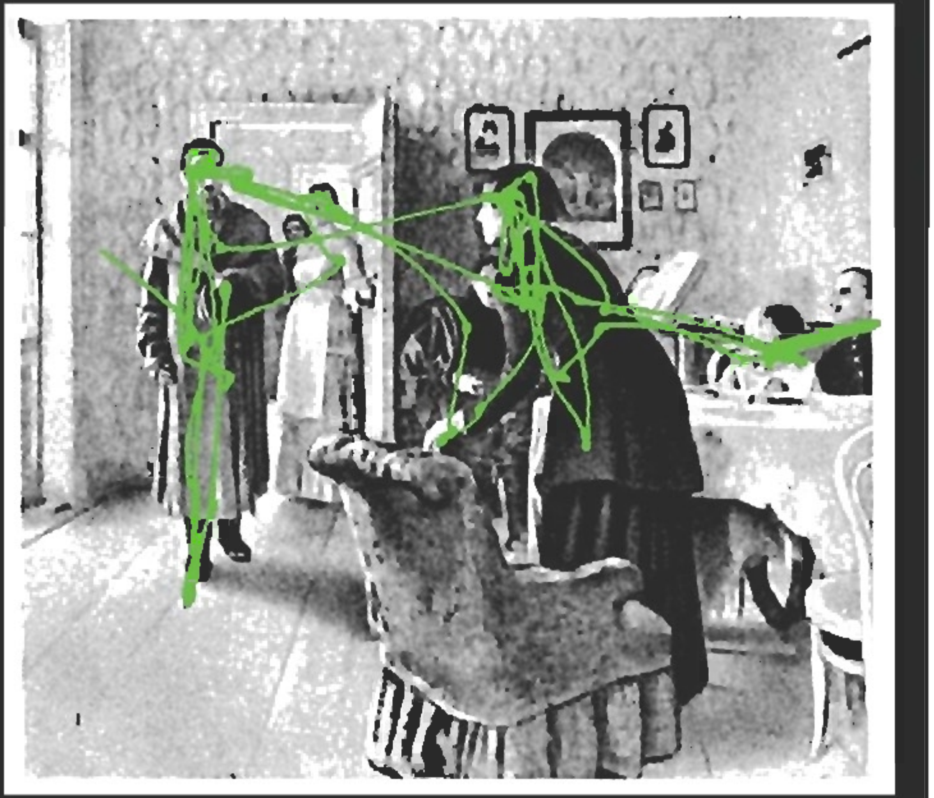
3) Goal directed (e.g. feature and spatial attention, motor preparation to act, arousal)

Natural vision: Eye movements

In the lab …



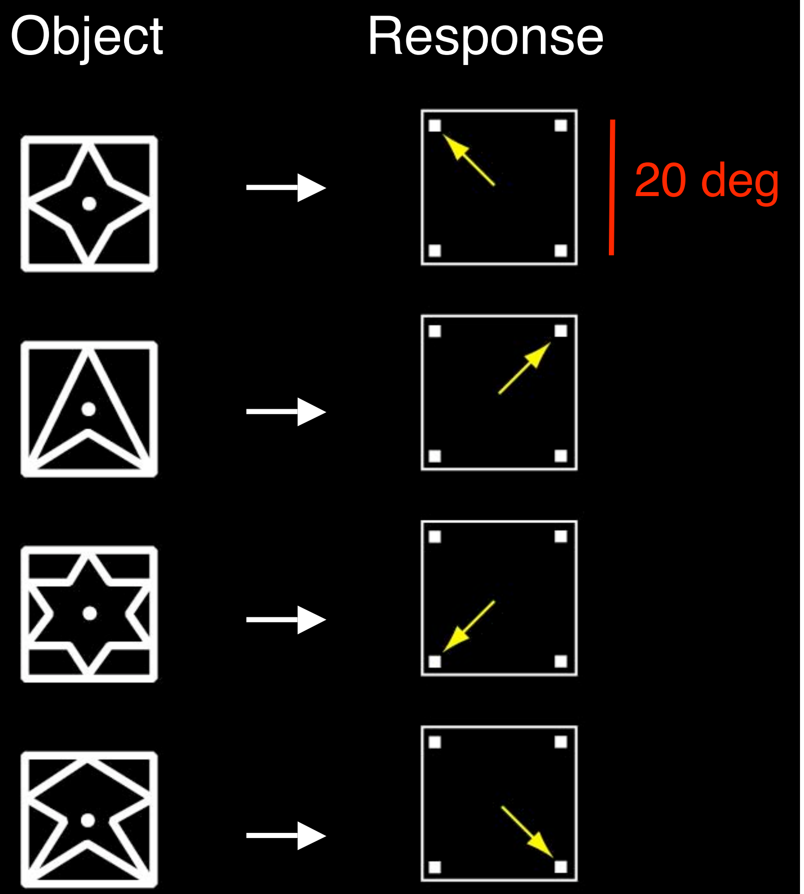
In the real world …

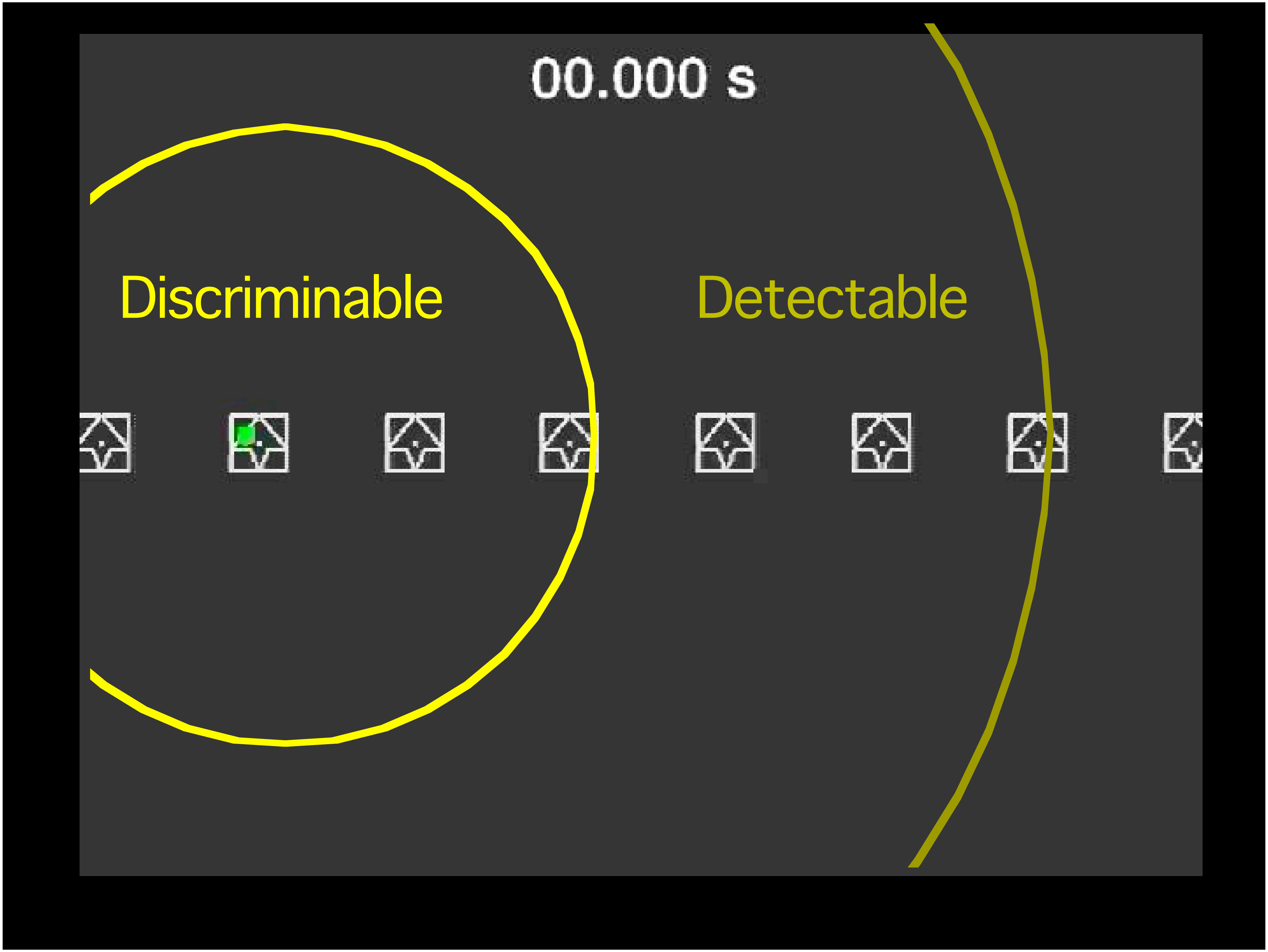


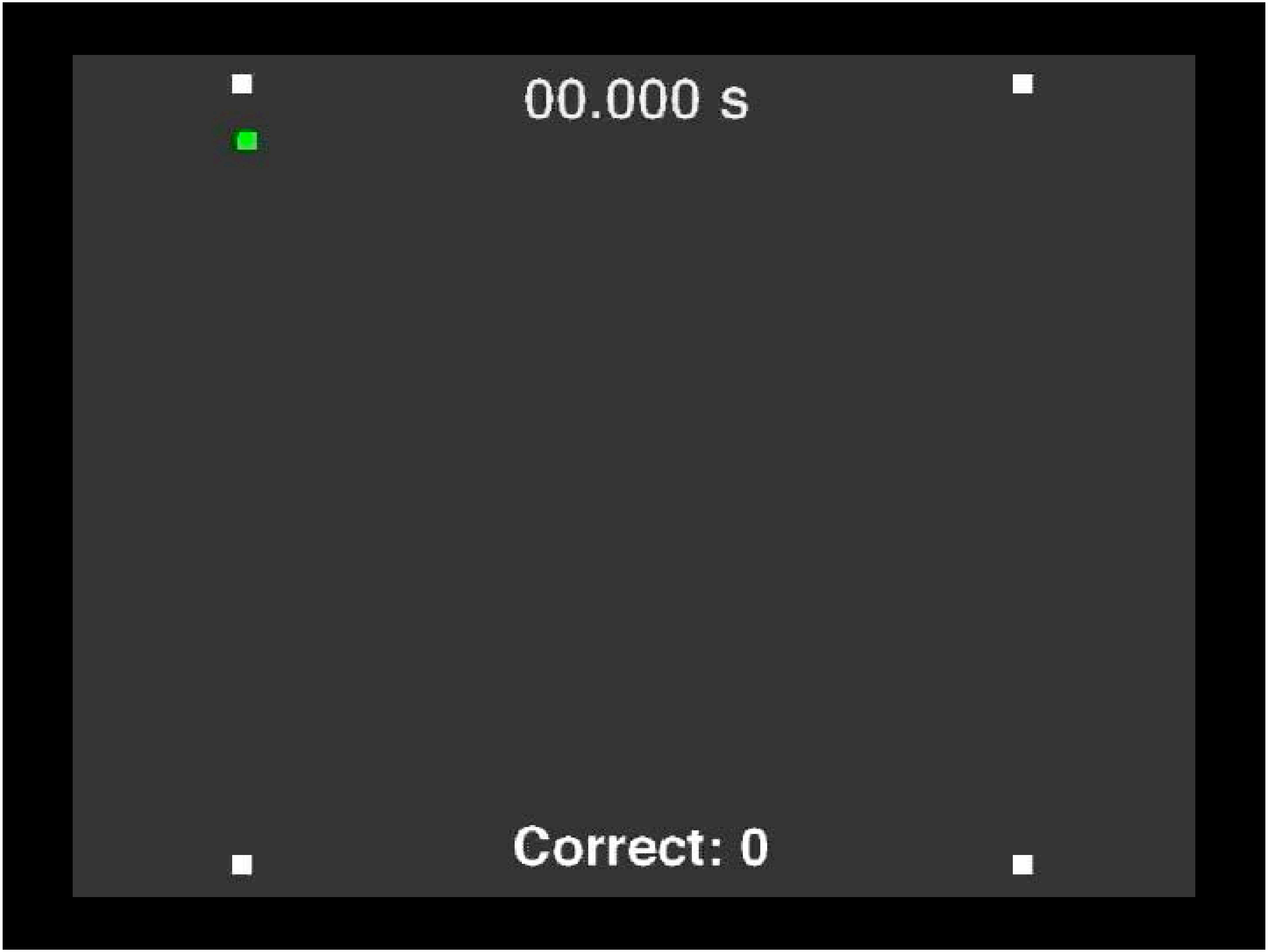
(Adapted from Yarbus, 1967)

Object identification task

Object		Response
A	→	“A”
B	→	“B”
C	→	“C”
D	→	“D”

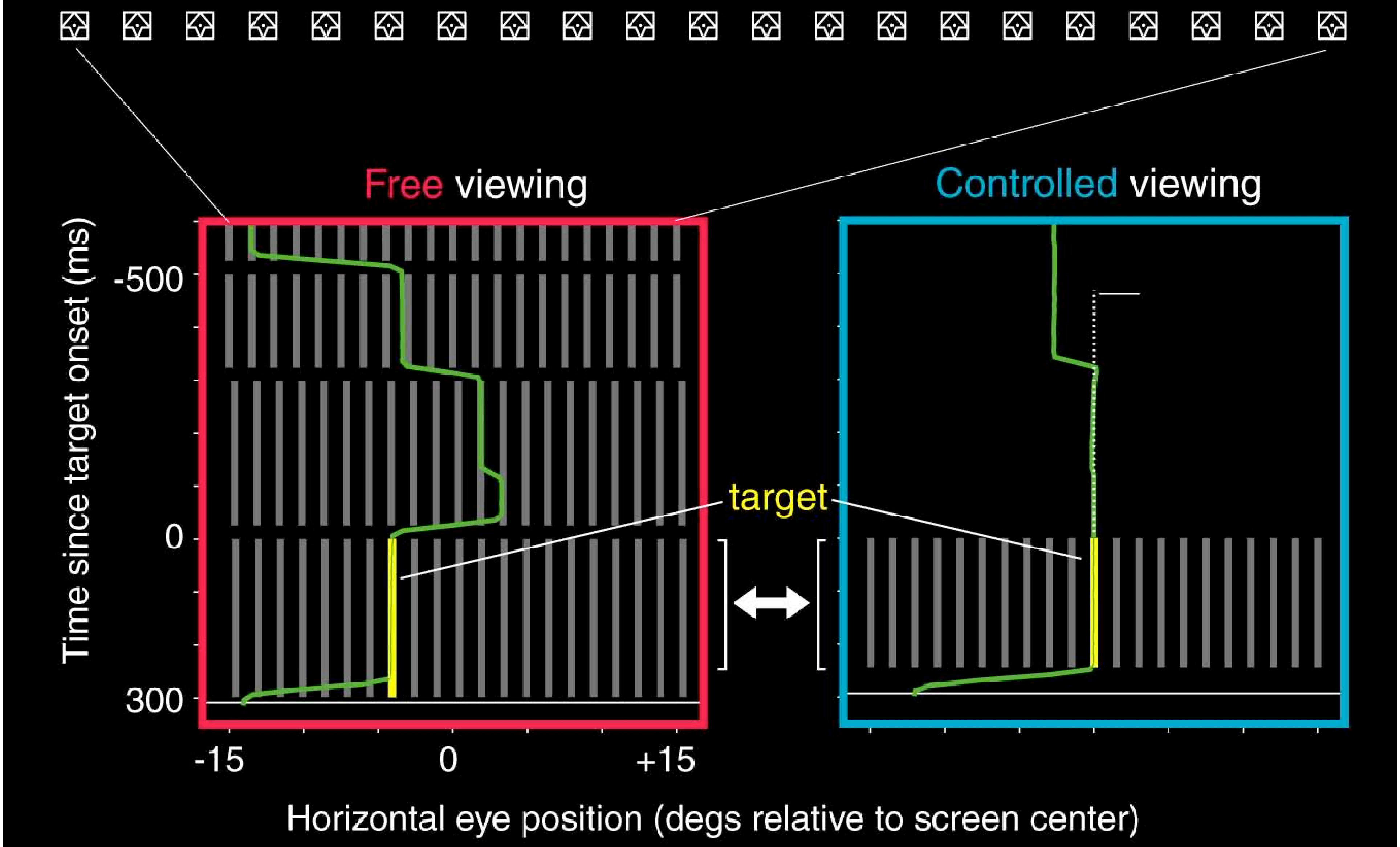



00.000 s
Discriminable
Detectable

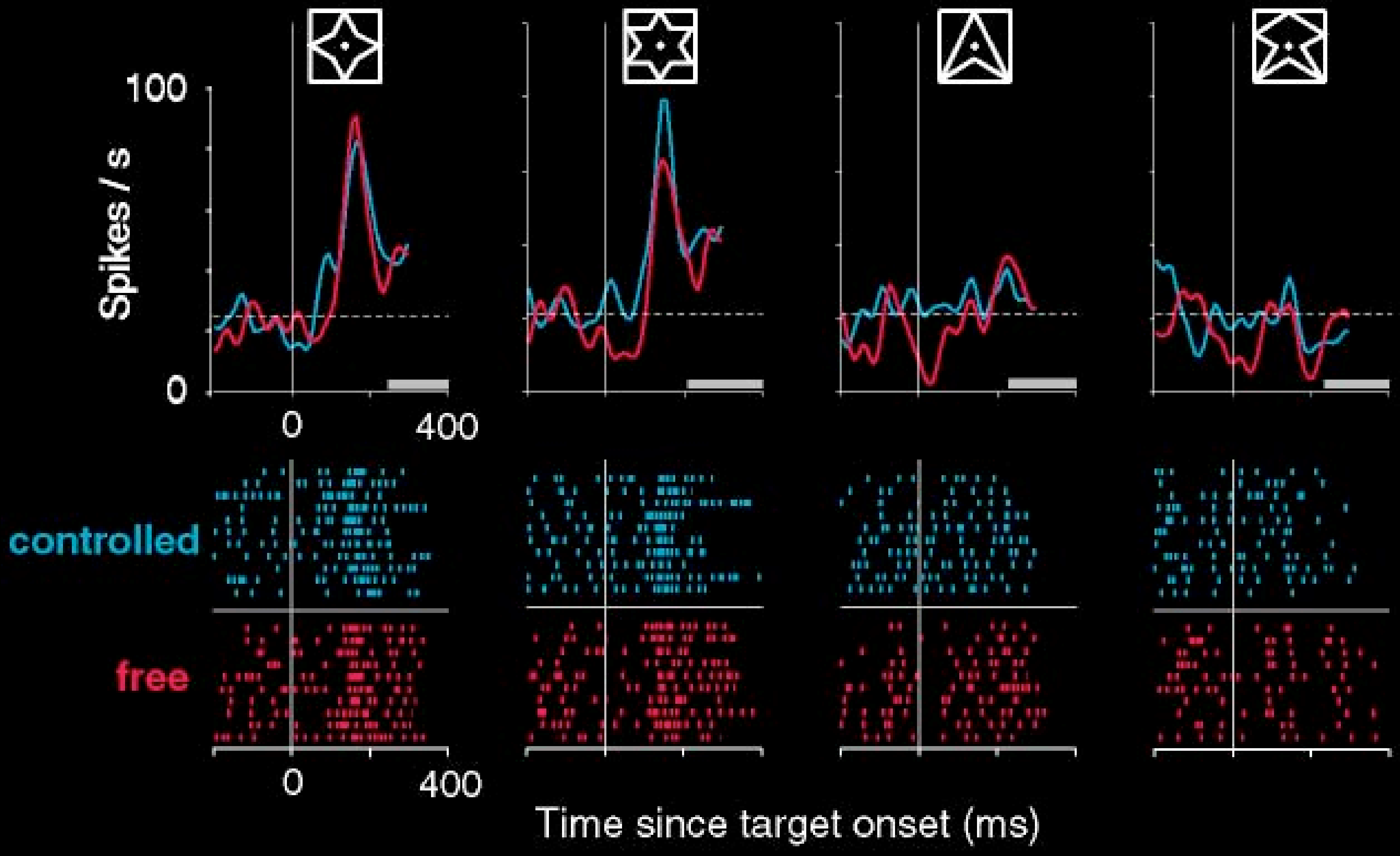


00.000 s

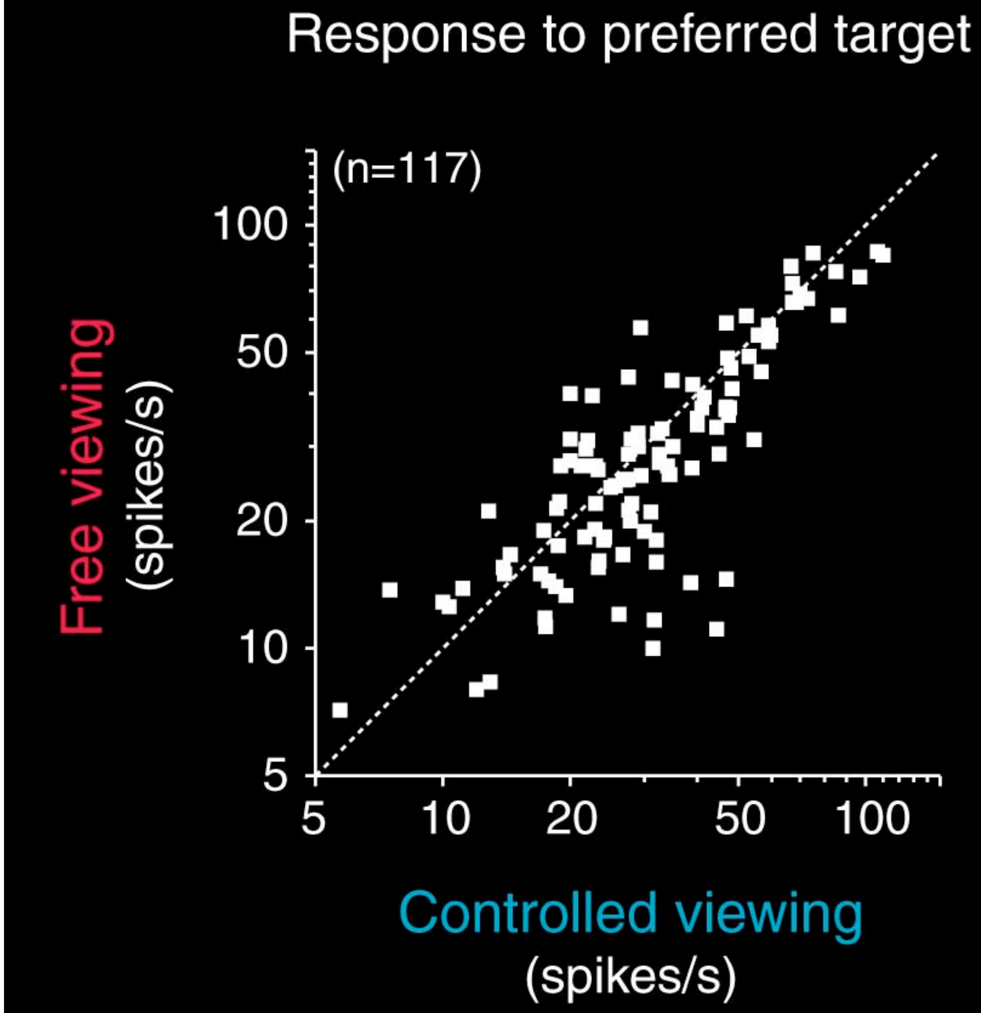
Correct: 0

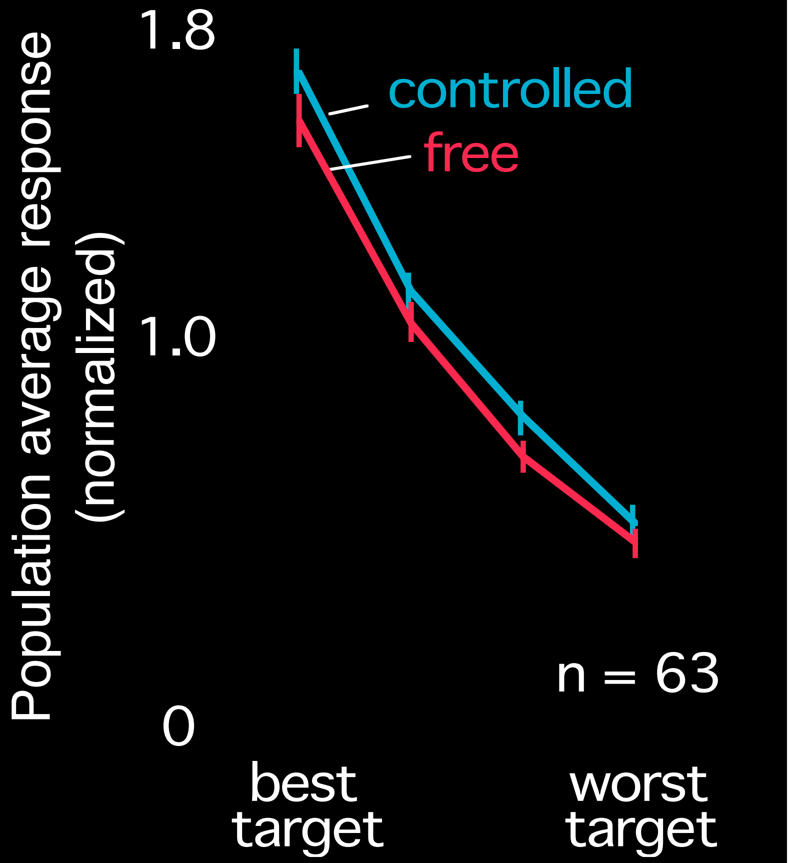

Free viewing
Controlled viewing
Time since target onset (ms)
-500
0
300
target
-15
0
+15
Horizontal eye position (degs relative to screen center)

Example IT neuron



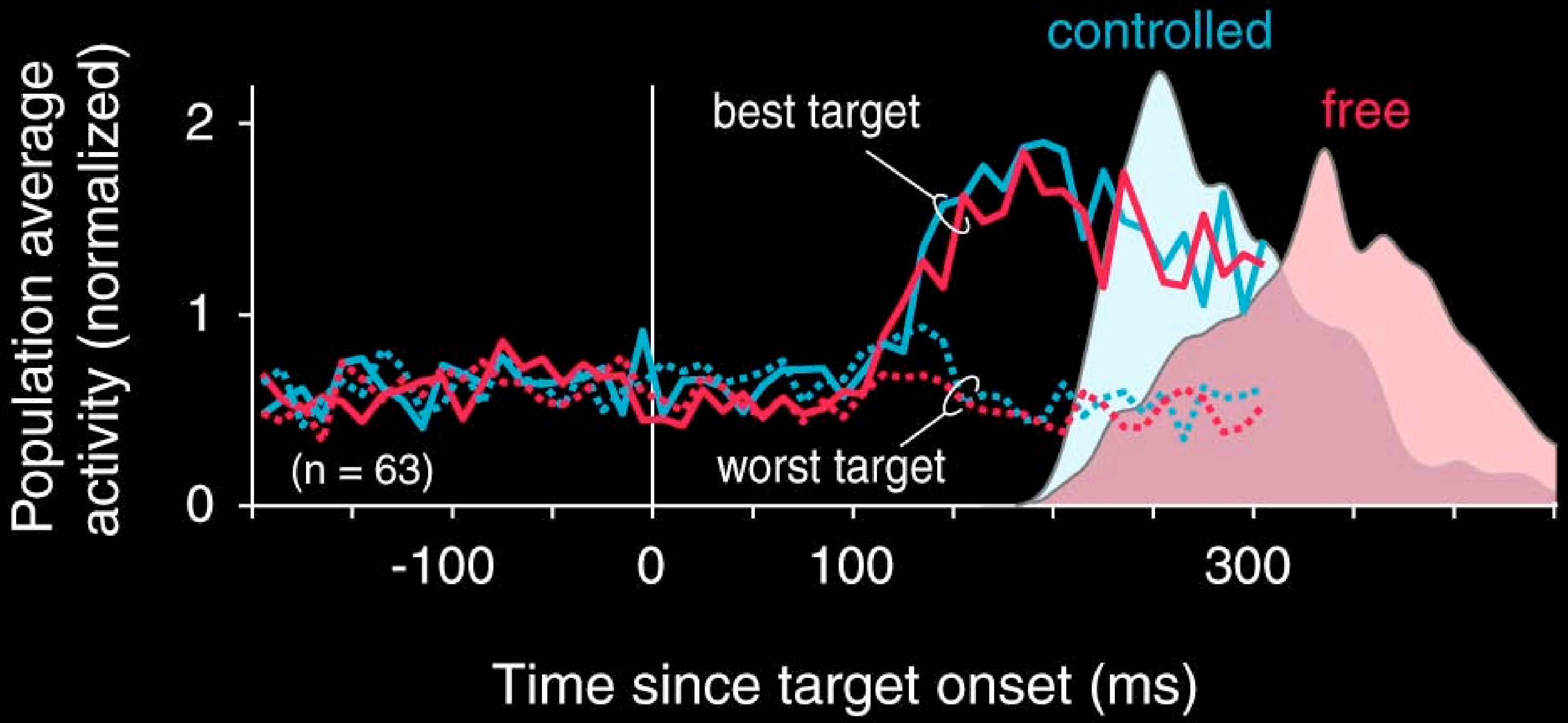
IT Population summary





DiCarlo and Maunsell, Nature Neuroscience, 3: 814-821 (2000)

IT responses are nearly identical in controlled and free viewing conditions



DiCarlo and Maunsell, Nature Neuroscience, 3: 814-821 (2000)

DiCarlo and Maunsell, J Neurophysiology (2005)

The Core and “natural vision”

How does “natural vision” challenge the basic model of core vision?

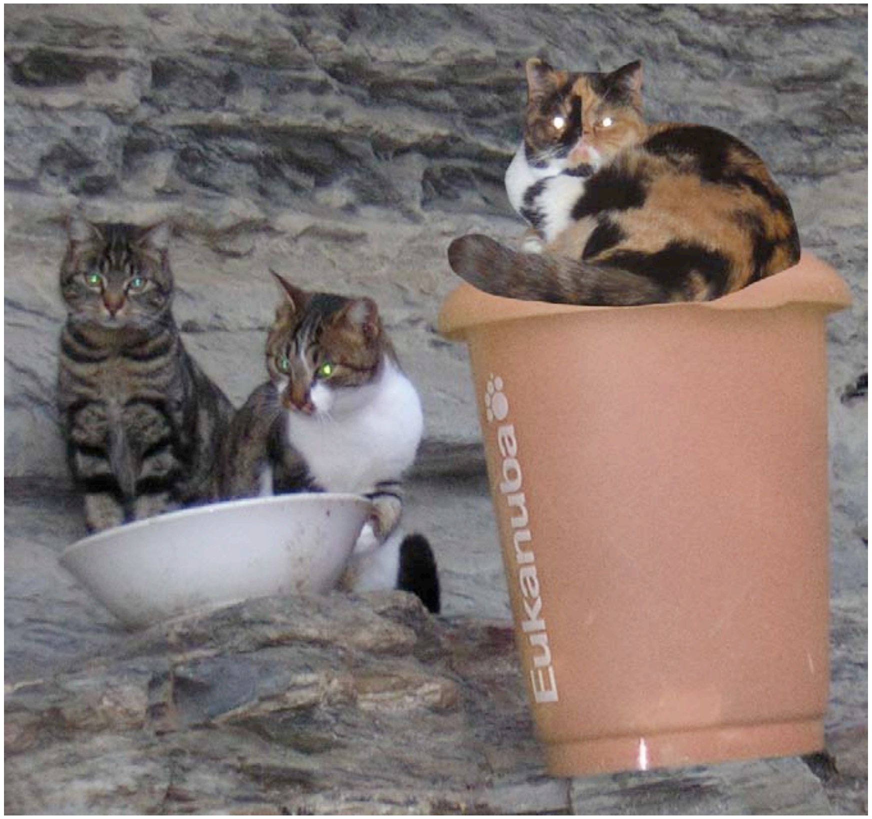
1) Eye movements (“free viewing”)
2) Clutter / Scene / Context: objects appear among other objects and on backgrounds
3) Goal directed (e.g. feature and spatial attention, motor preparation to act, arousal)

Not much to worry about here.

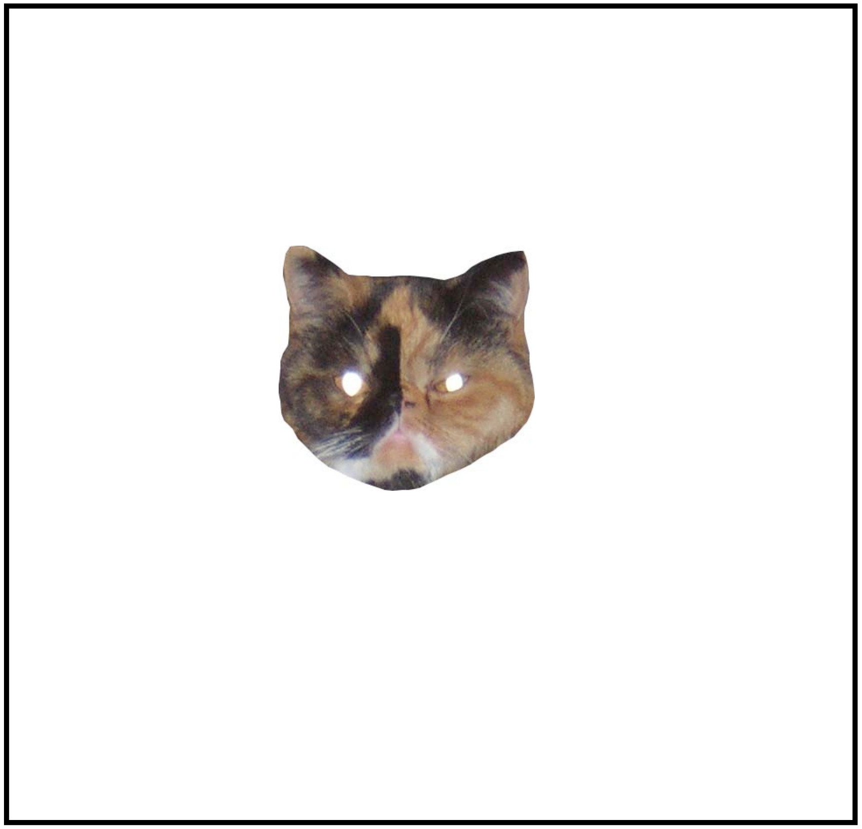
DiCarlo and Maunsell, 2000
Sheinberg and Logothetis, 2001

Natural vision: Clutter, scene, and context

In the real world…

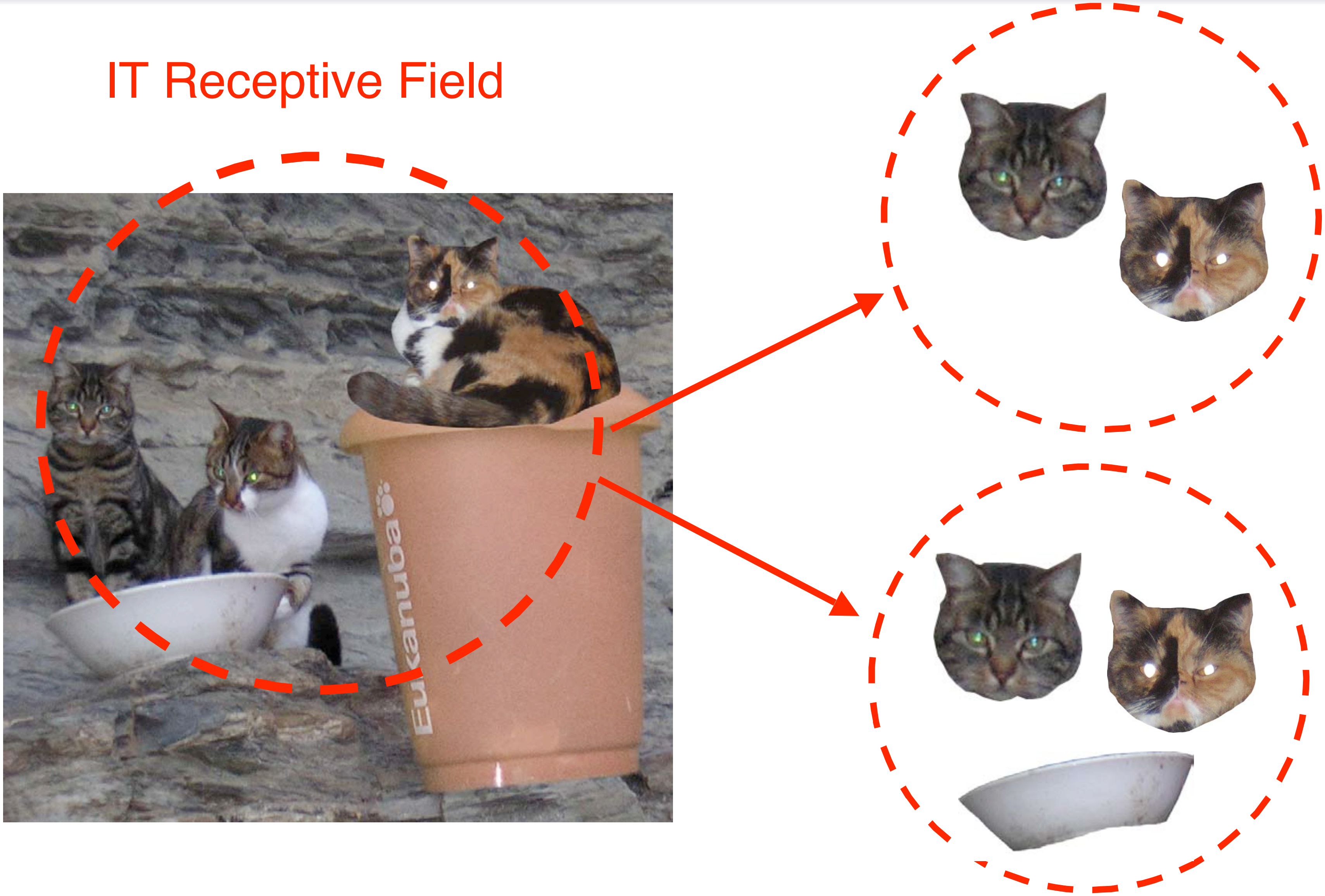


In the lab…



Natural vision: Clutter, scene, and context

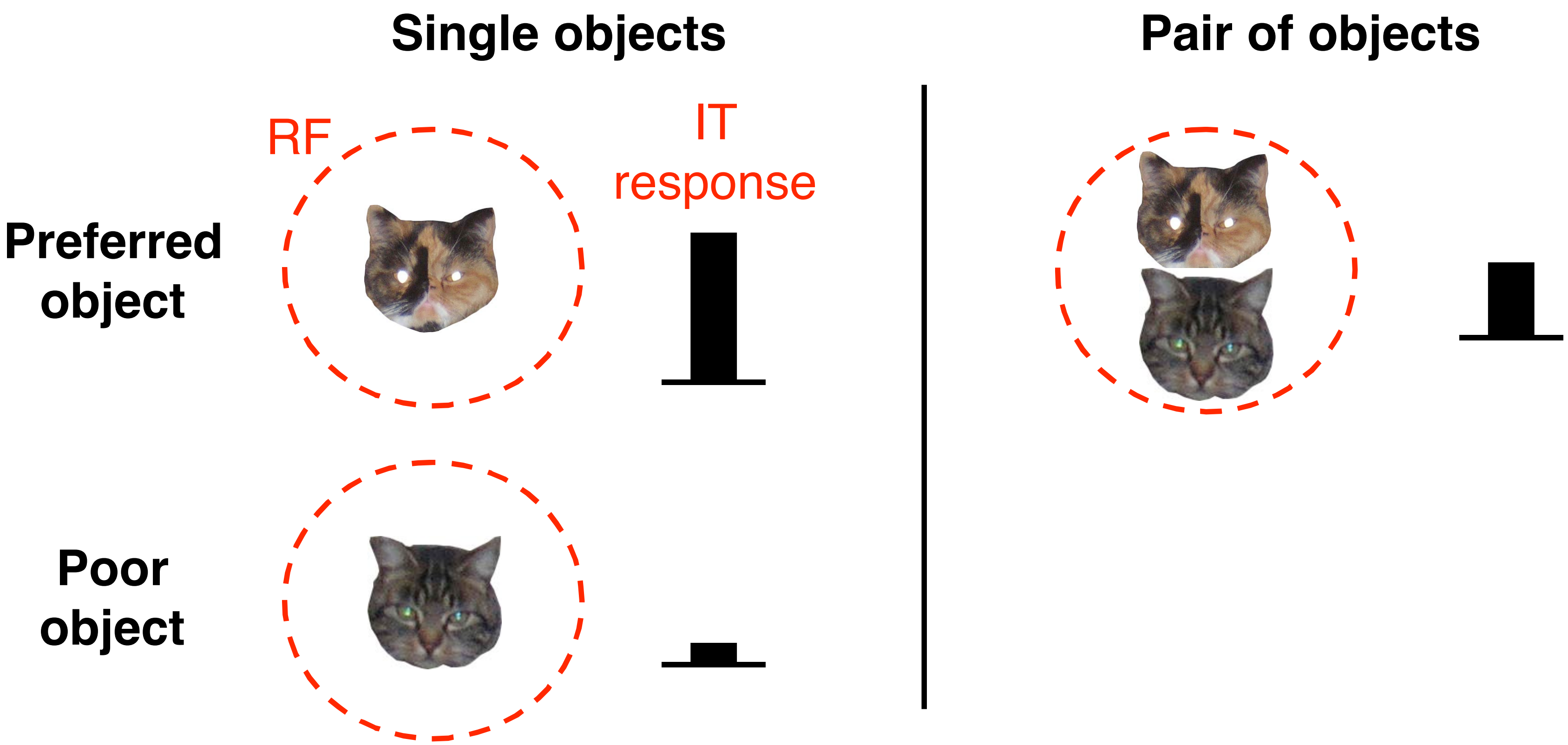
IT Receptive Field



Long term goal: Understand IT in clutter

IT responses to object are typically reduced when additional objects are presented

(Sato, 1989; Miller et al., 1993; Rolls and Tovee, 1995; Chelazzi et al., 1998; Missal et al., 1999)



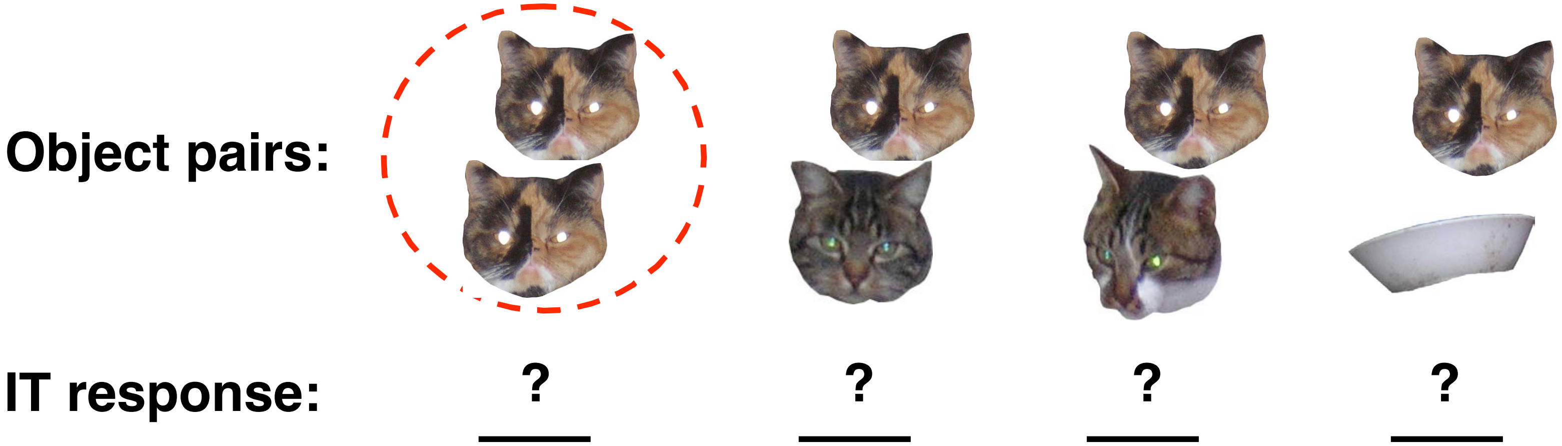
Single objects

Pair of objects

Preferred object

Poor object

First open questions …



Object pairs:

IT response: ? ? ? ?

- Any **systematic relationship** between:
  - response to an object pair
  - responses to the constituent objects?

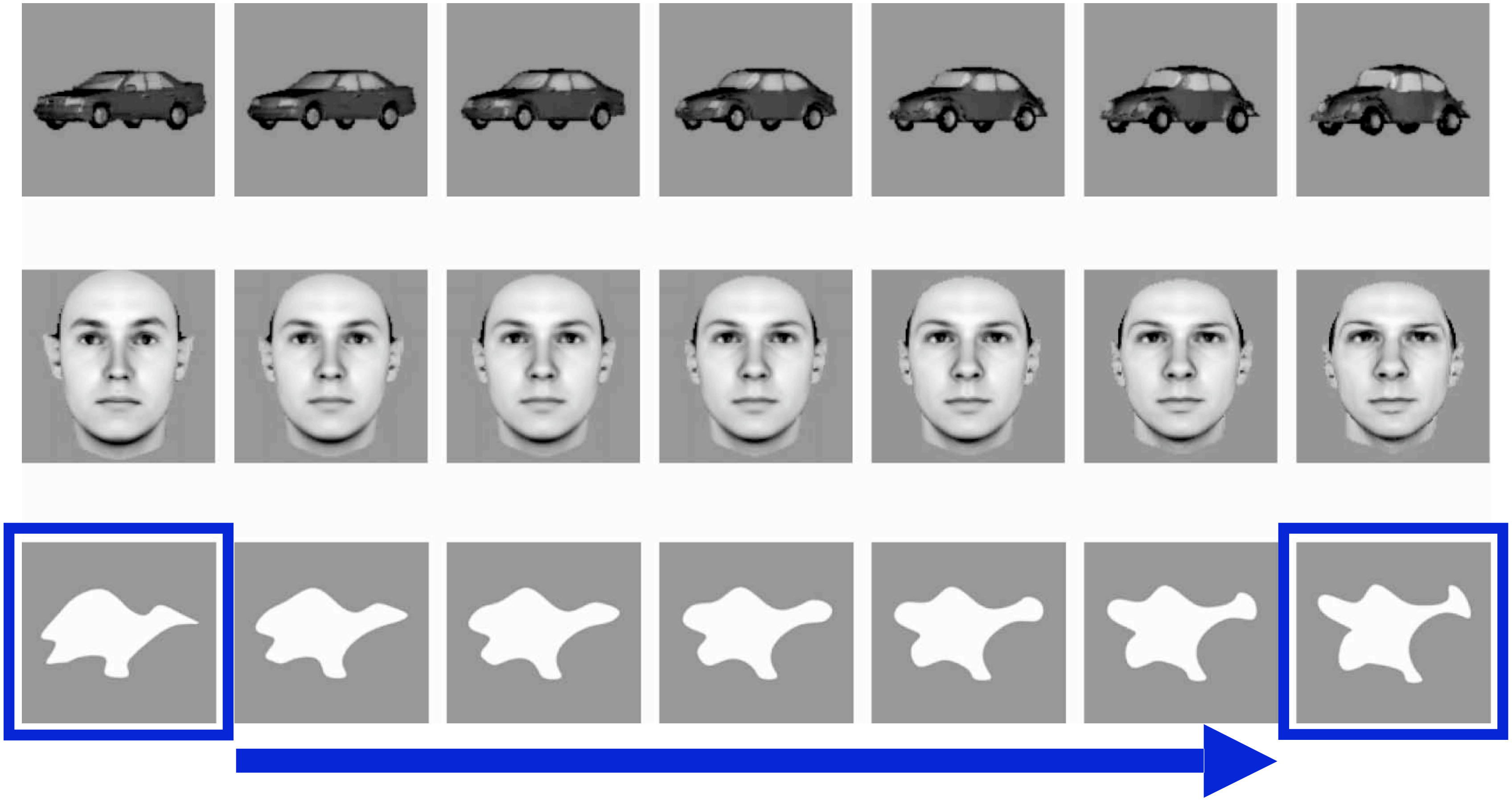
Experimental design
overview

- Davide Zoccolan and David Cox
- Recorded IT neuronal responses to the presentation of:
  - **Single** objects
  - **Pairs** of objects
  - **Triplets** of objects

  - In **three monkeys**
  - Using **two complementary experiments**

Experiment 1

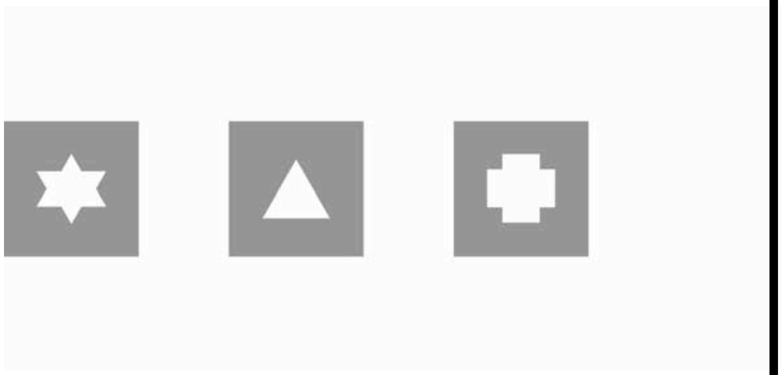
Experiment 2

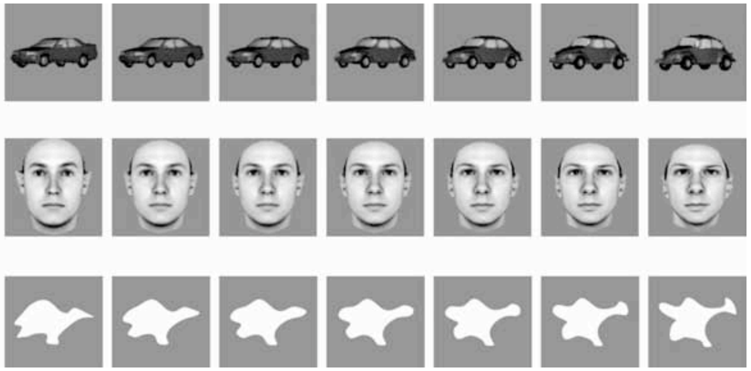


Stimulus conditions

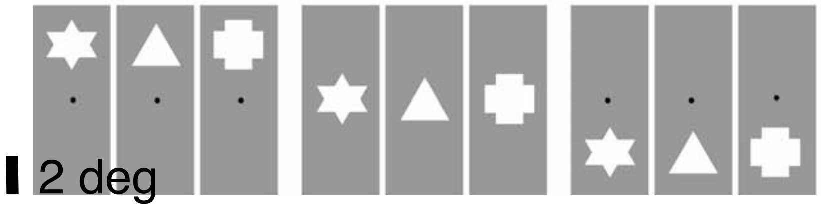
EXPERIMENT 1

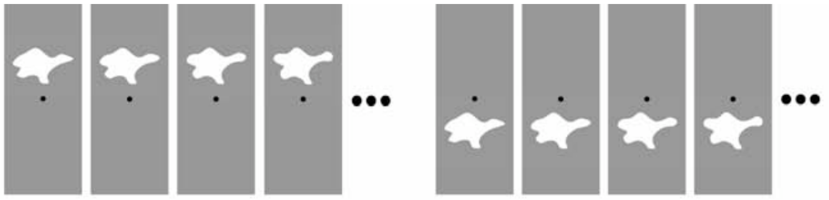
EXPERIMENT 2



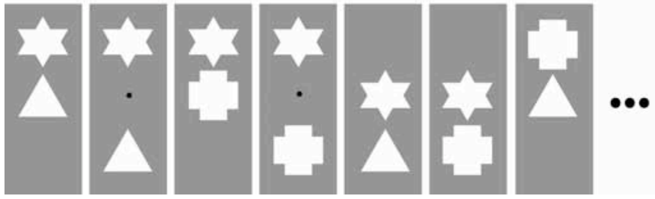


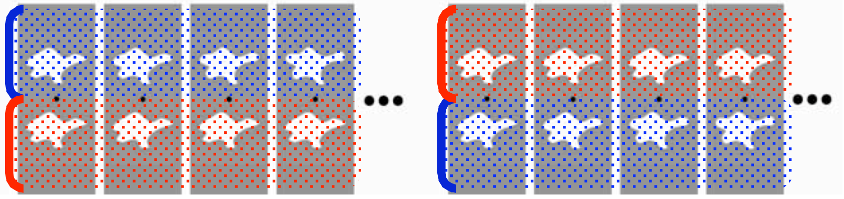
Single objects





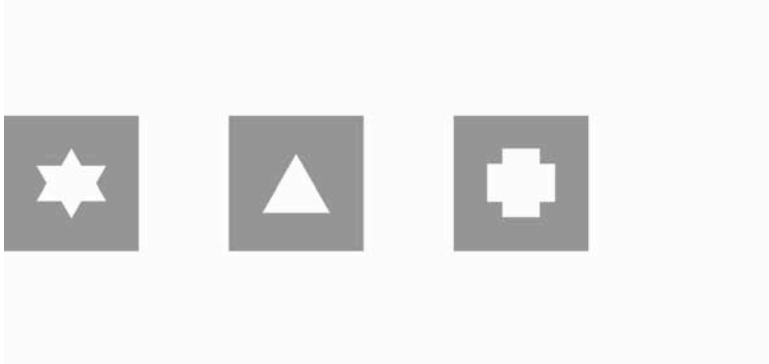
Object pairs



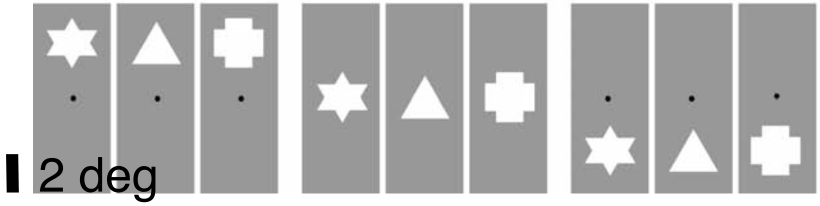


Stimulus conditions

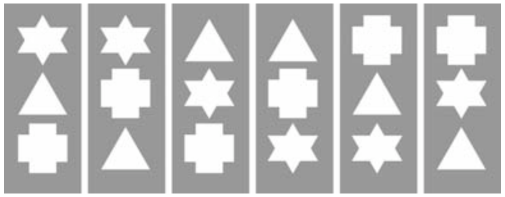
EXPERIMENT 1



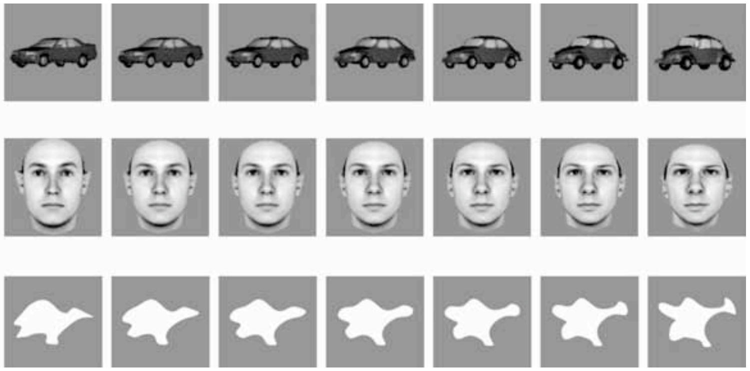
Single objects

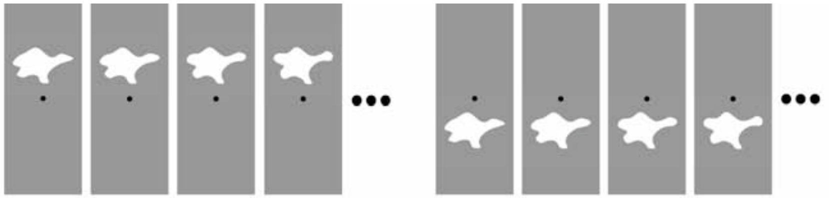


Object triplets

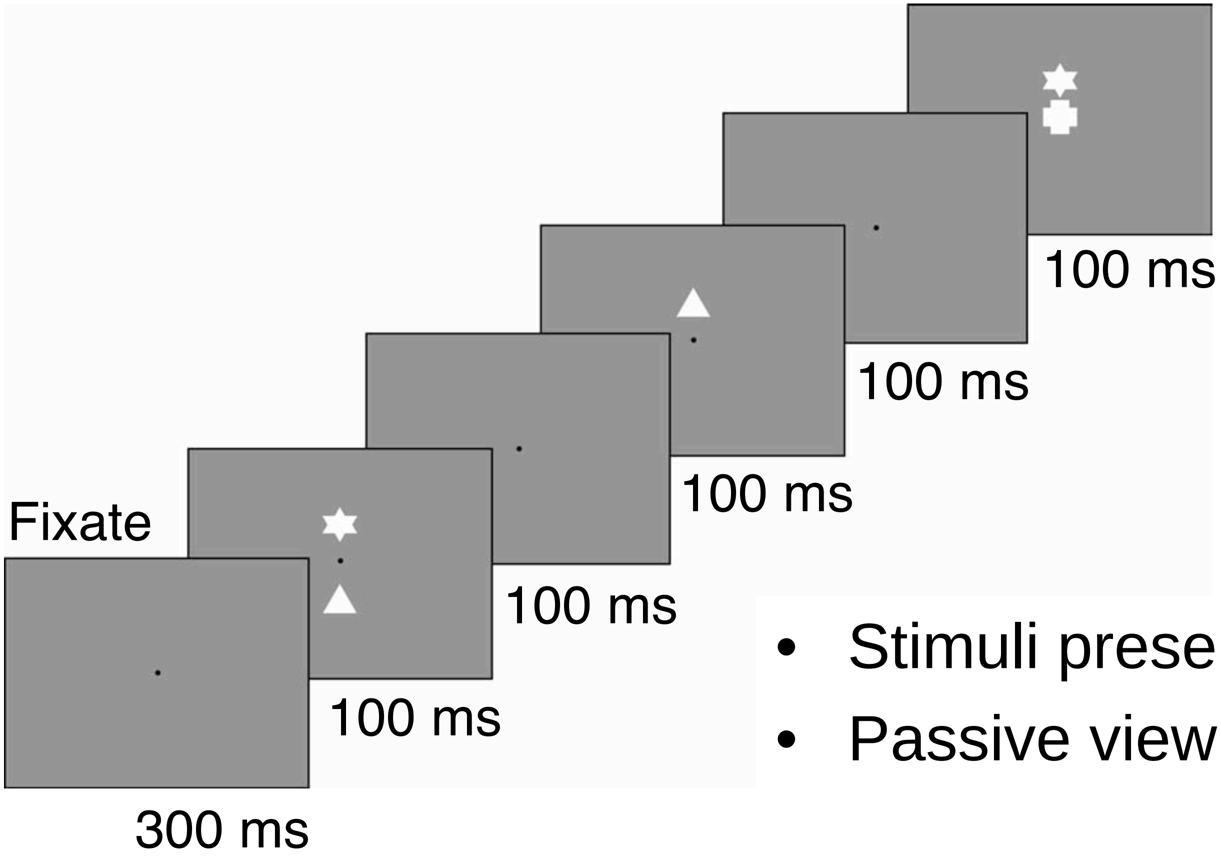


EXPERIMENT 2





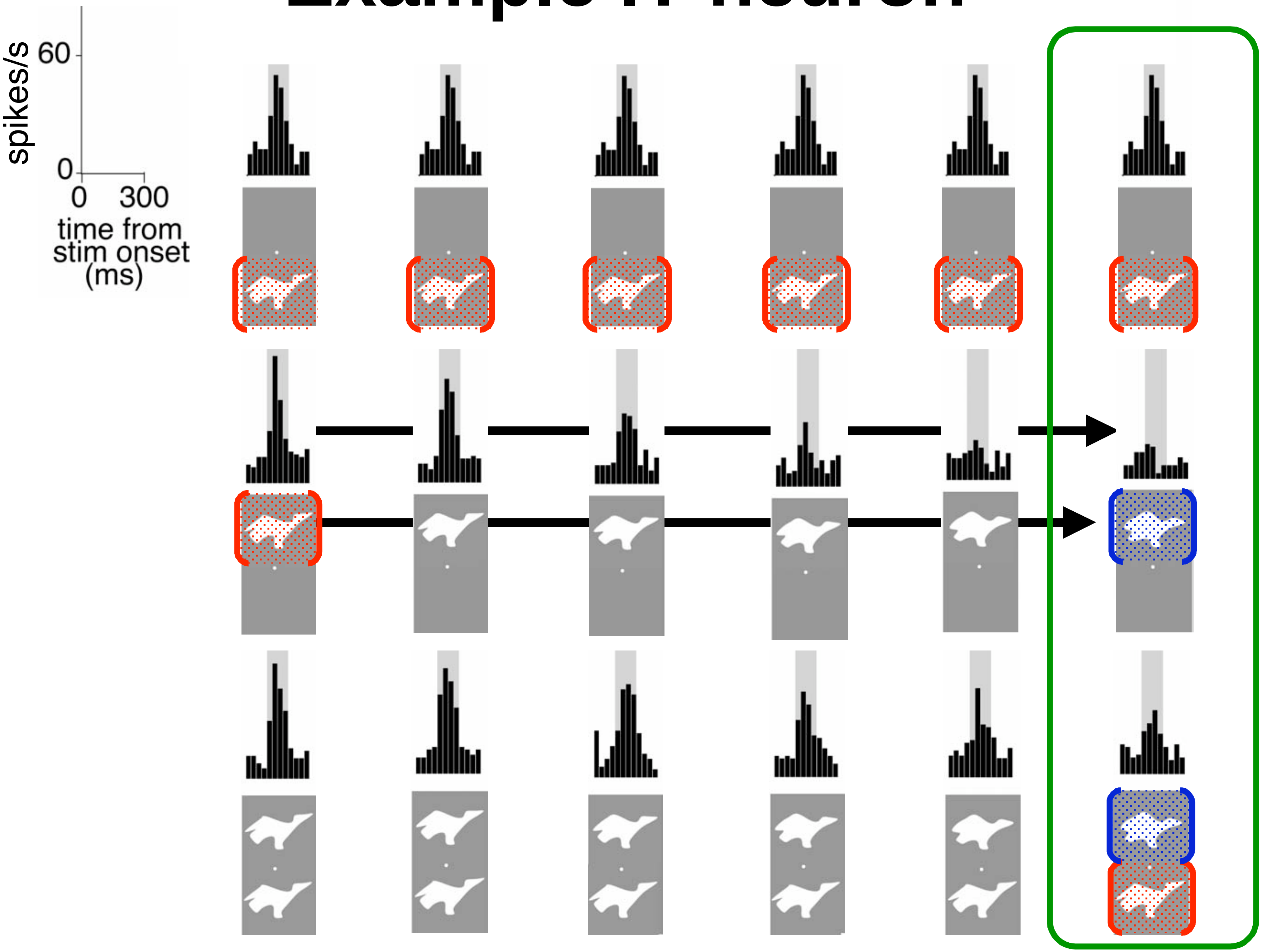
Core response: Rapid visual presentation



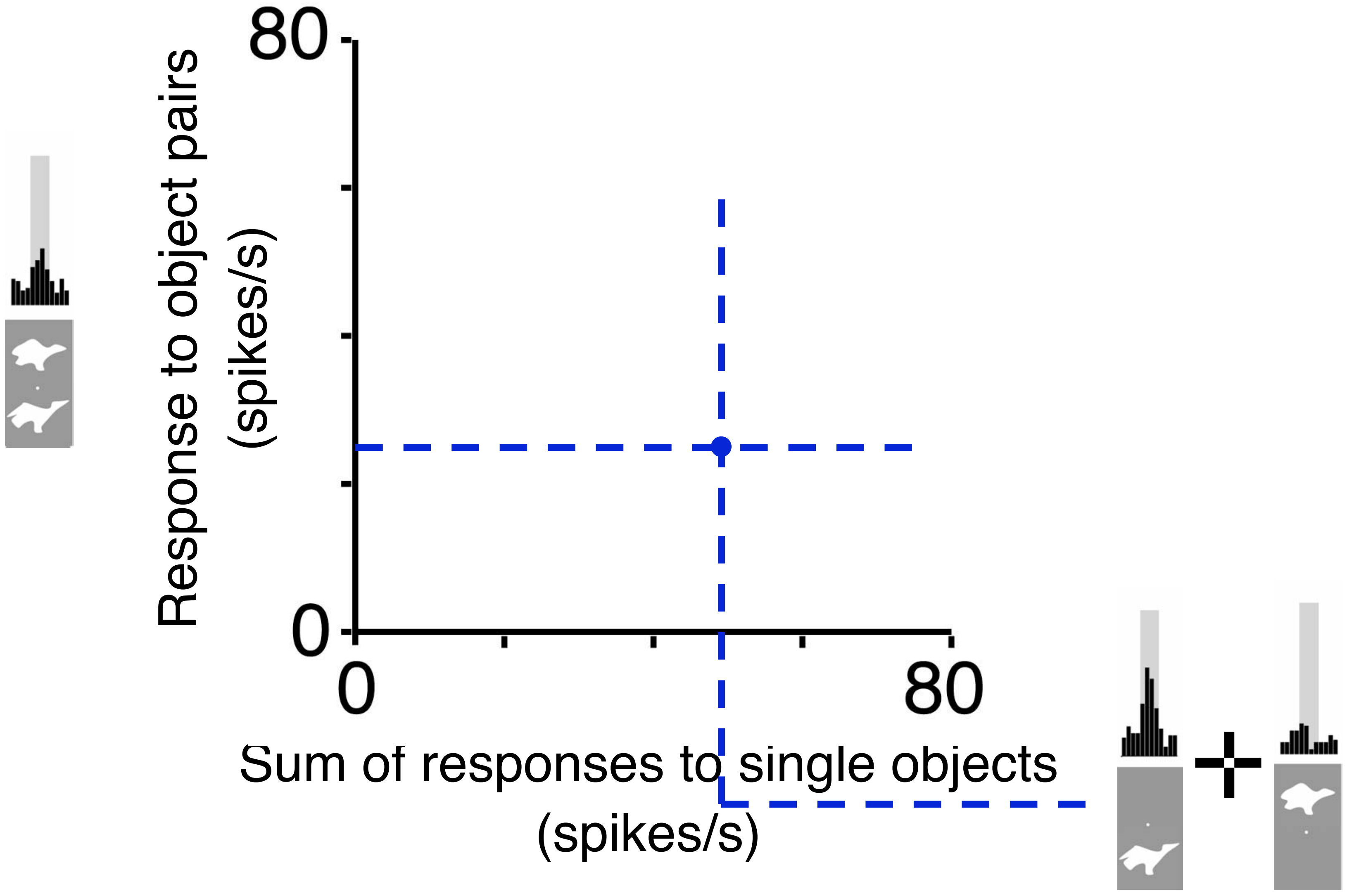
- Stimuli presented at **5 per sec**
- Passive viewing

104 neurons recorded in three monkeys

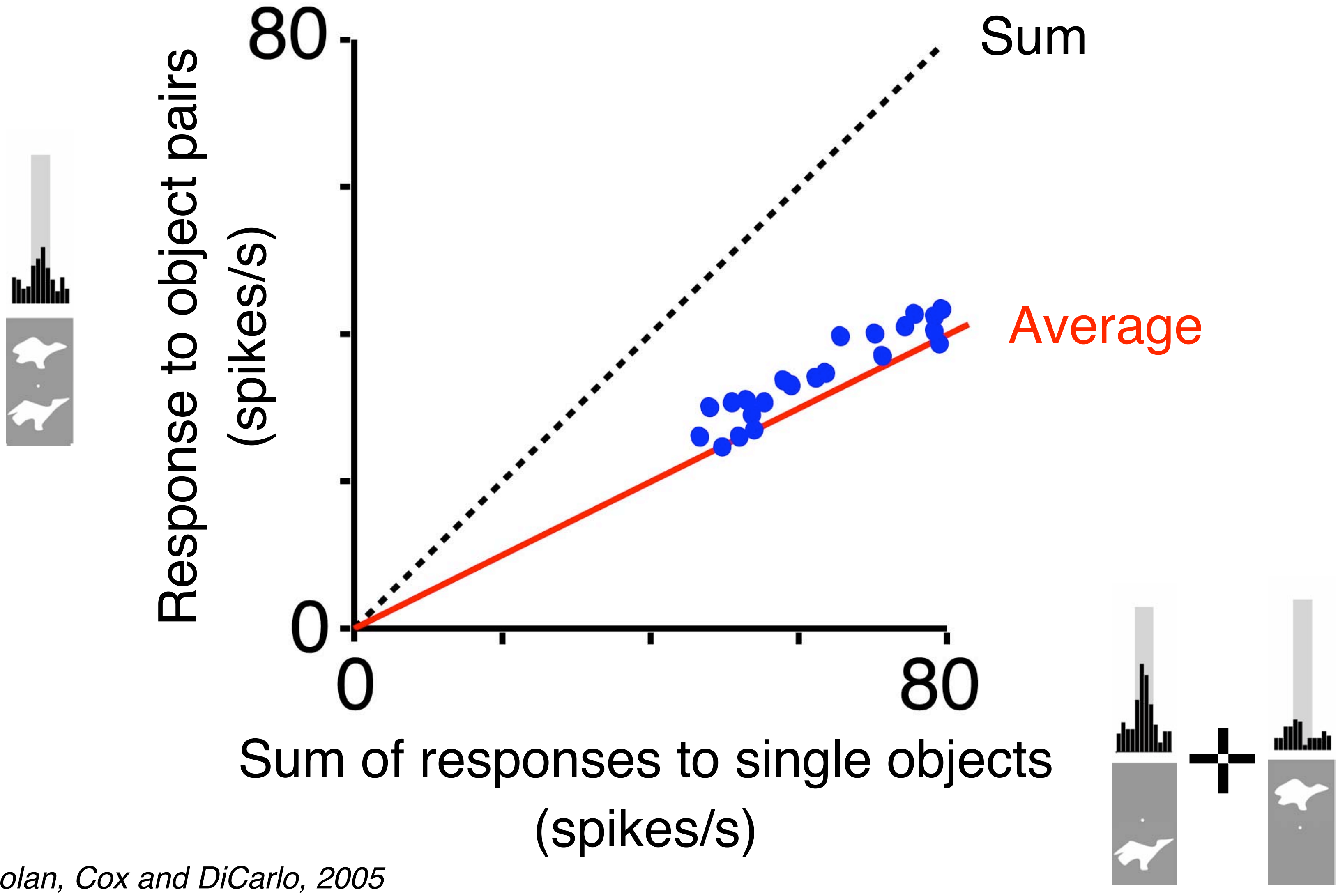
Example IT neuron



Example IT neuron

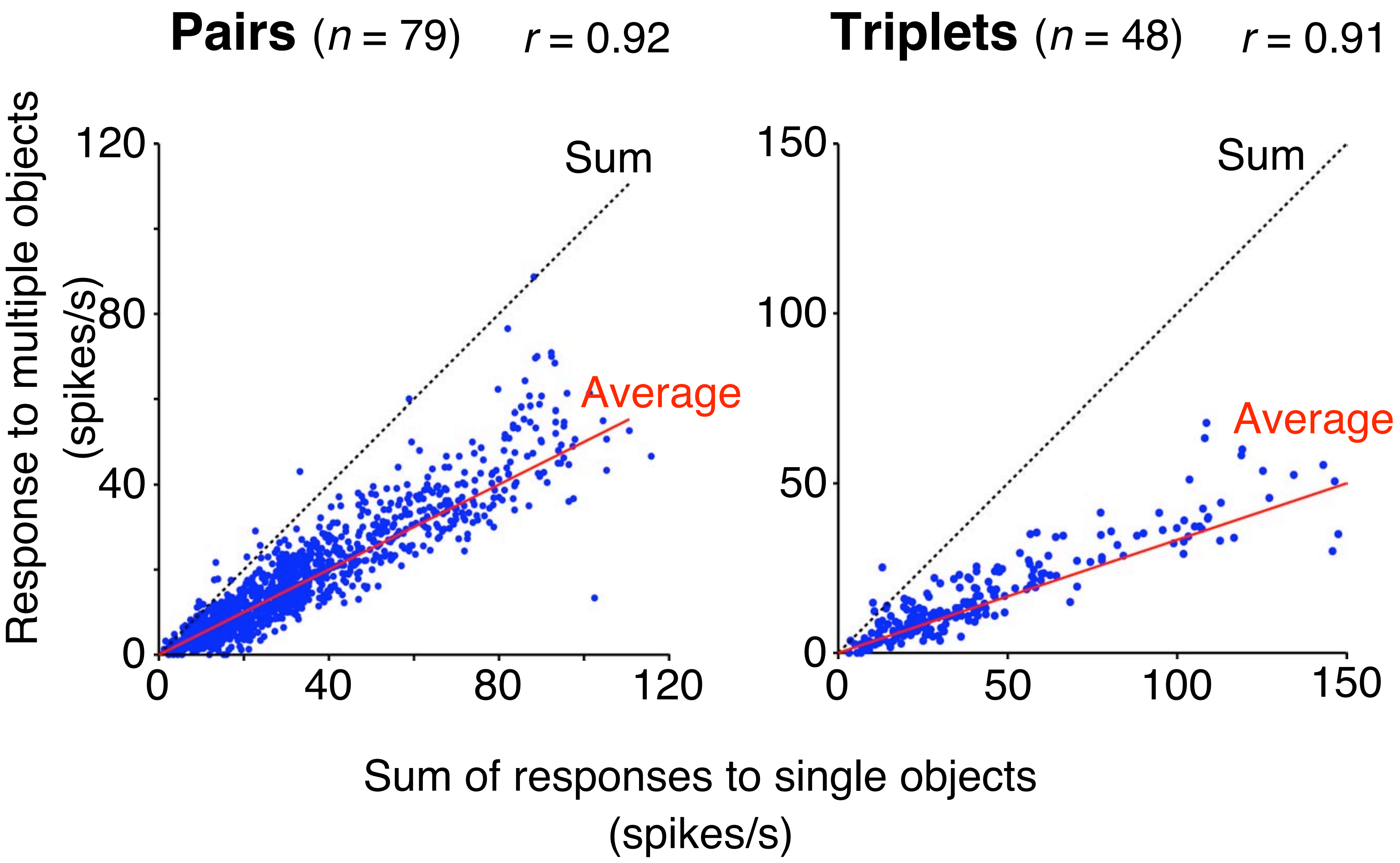


Example IT neuron



Zoccolan, Cox and DiCarlo, 2005

Population analysis



Zoccolan, Cox and DiCarlo, 2005

Summary: The Core and multiple objects

Under the conditions described here**:**

- An “average rule” is a very good predictor of the response of individual IT neurons

  (explains **~63%** of response variance → ***r*** **≈ 0.8**)

- => The response pattern of The Core can be predicted by the response pattern to each constituent object

- => useful for supporting the simultaneous representation of multiple objects

The Core and "natural vision"

How does "natural vision" challenge the basic model of core vision?

1) Eye movements ("free viewing")
2) Clutter / Scene / Context: objects appear among other objects and on backgrounds
3) Goal directed (e.g. feature and spatial attention, motor preparation to act, arousal)

Not much to worry about here.

DiCarlo and Maunsell, 2000

Sheinberg and Logothetis, 2001

Very important challenge.

Beginnings of a systematic understanding.

Zoccolan, Cox and DiCarlo, 2005