Project in computational neuroscience:

Detection and Recognition of objects in visual cortex

NIH Conte Center
with
J. DiCarlo, E. Miller, D. Ferster, C. Koch, M. Riesenhuber, T. Poggio
(MIT, CalTech, Northwestern, Georgetown)

A theory of visual recognition is used as a tool to integrate and drive multidisciplinary research in different experimental neuroscience labs.

Object Recognition (for biology and for machines) is difficult: trade-off between **selectivity** and **invariance**

Many different images can correspond to the same type of object…

…while similar activation patterns can correspond to different objects

Thanks to Ezra Rosen for images

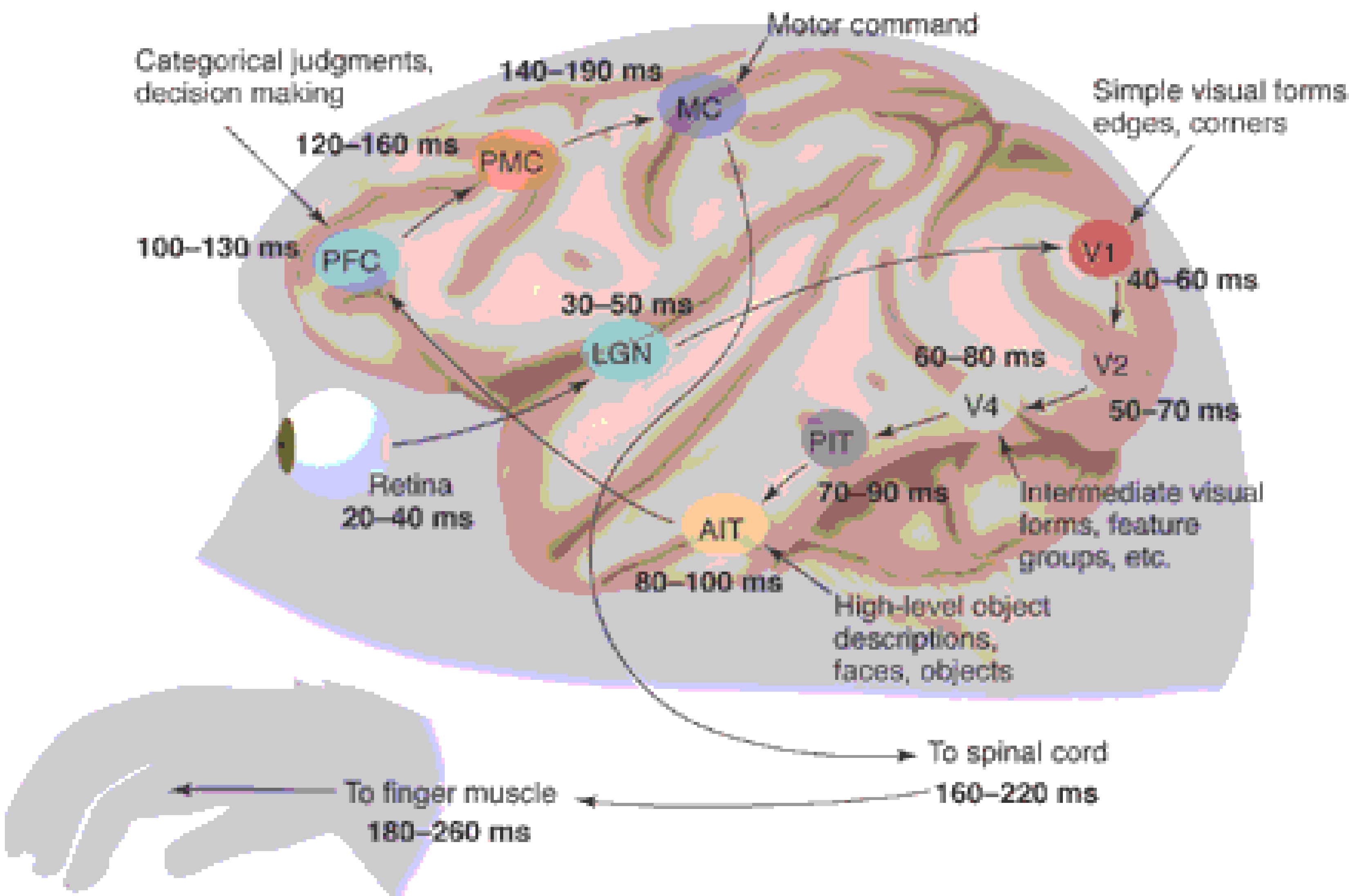
The first 100 msec of visual recognition…

…these are the kind of visual tasks we would like to explain with a feedforward model, *extending* Hubel and Wiesel

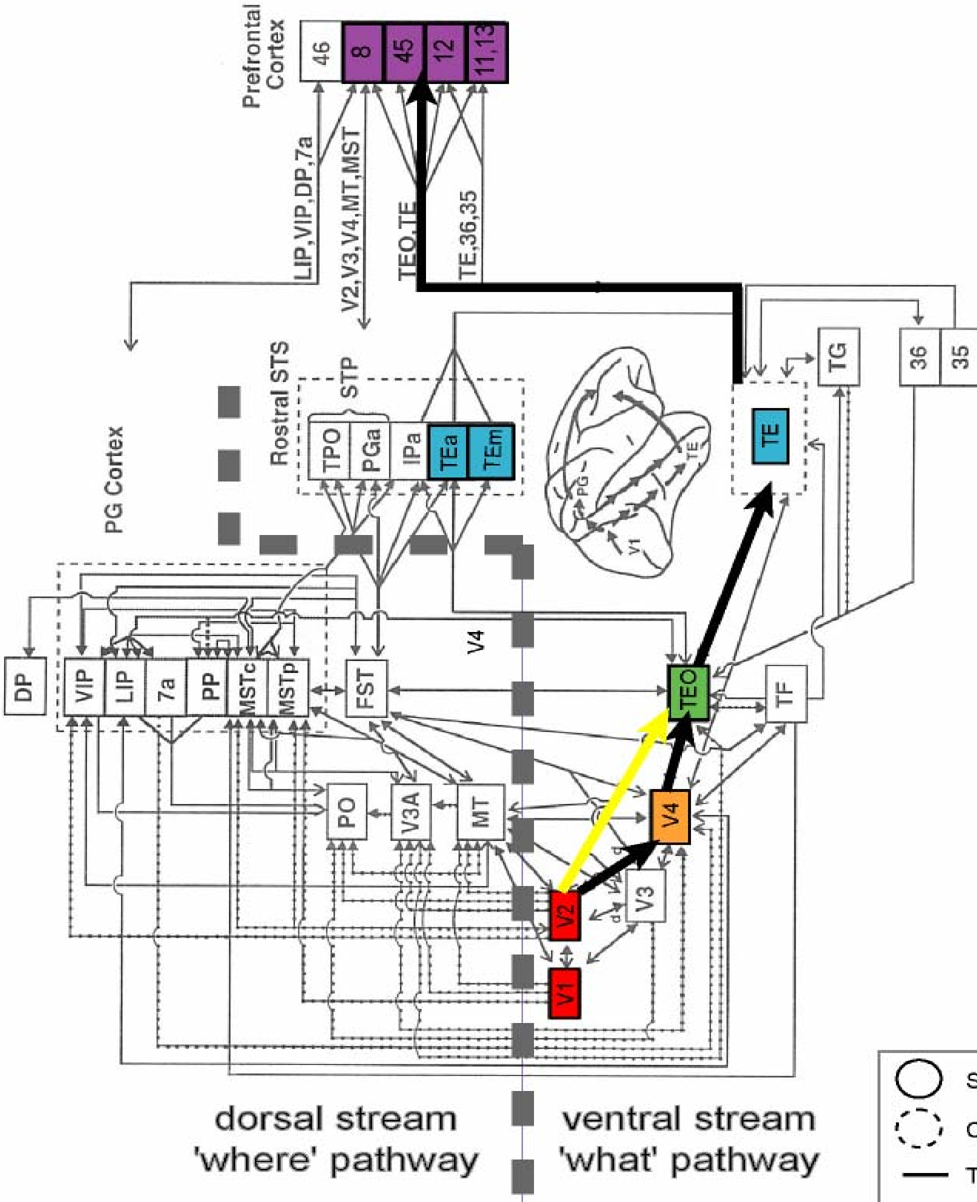
from V1 to PFC

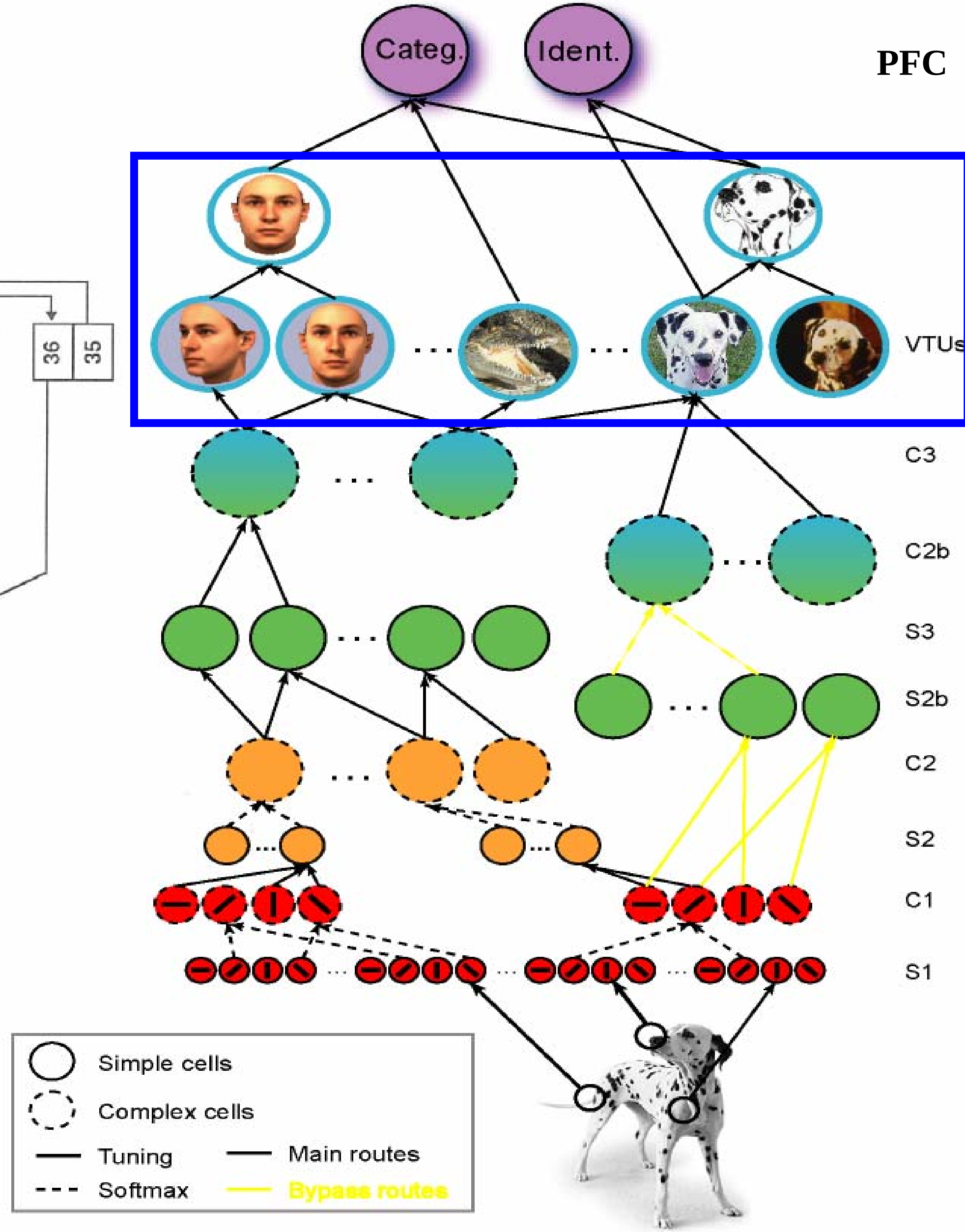
Movie courtesy of Jim DiCarlo

Ventral stream in visual cortex



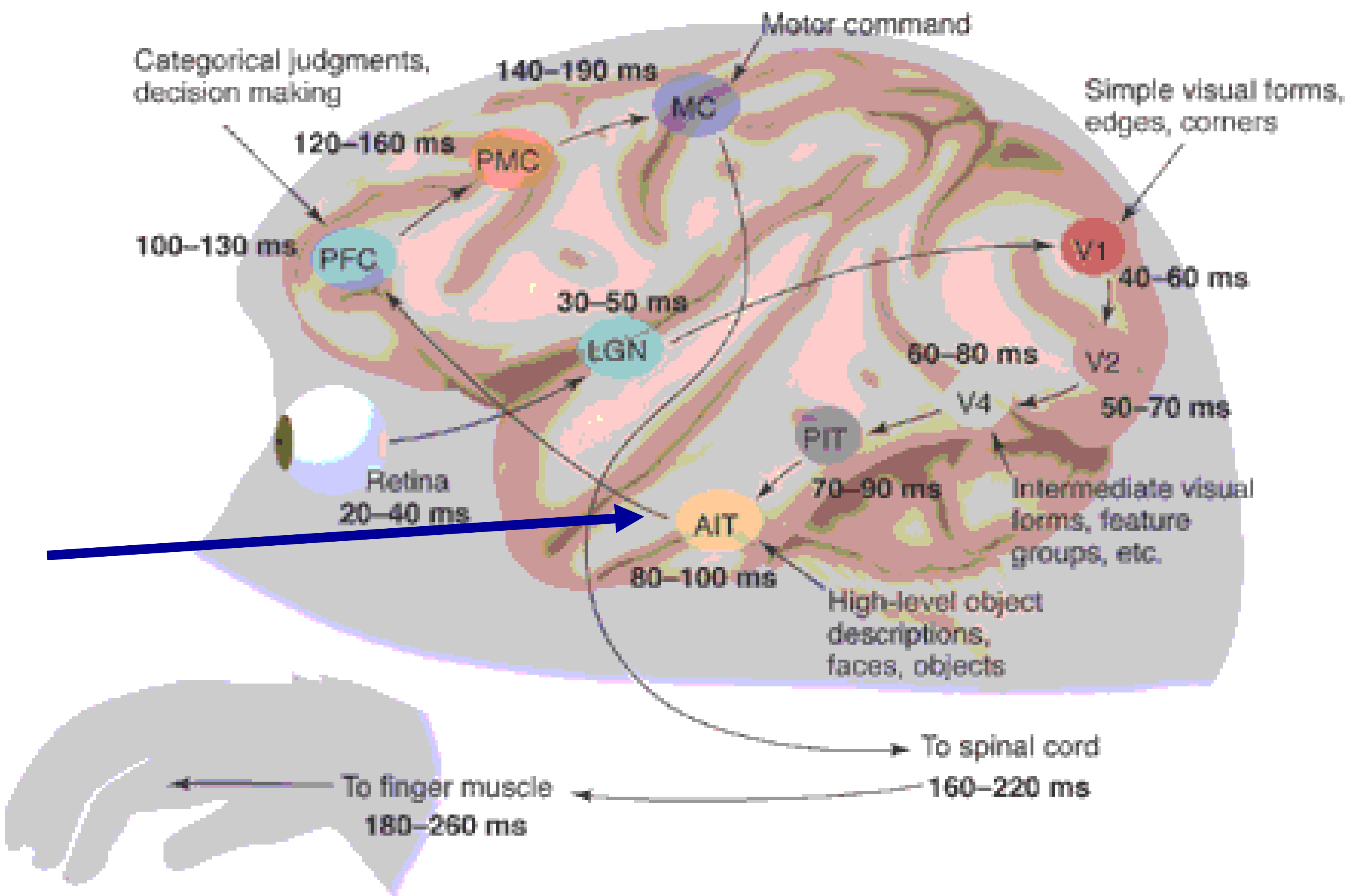
Mapping the ventral stream into a theory





Serre, Kouh, Cadieu, Knoblich, Poggio, 2005

Ventral stream in visual cortex



IT is the final visual stage in the theory…

thus let us give a (new) look at the representation in IT:

classifiers (eg learning algorithms) for *read-out* from IT

Chou Hung, Gabriel Kreiman, Tomaso Poggio, James DiCarlo
(with help from Rodrigo Quiroga and and Alexander Kraskov and from DARPA)

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

Goal 1

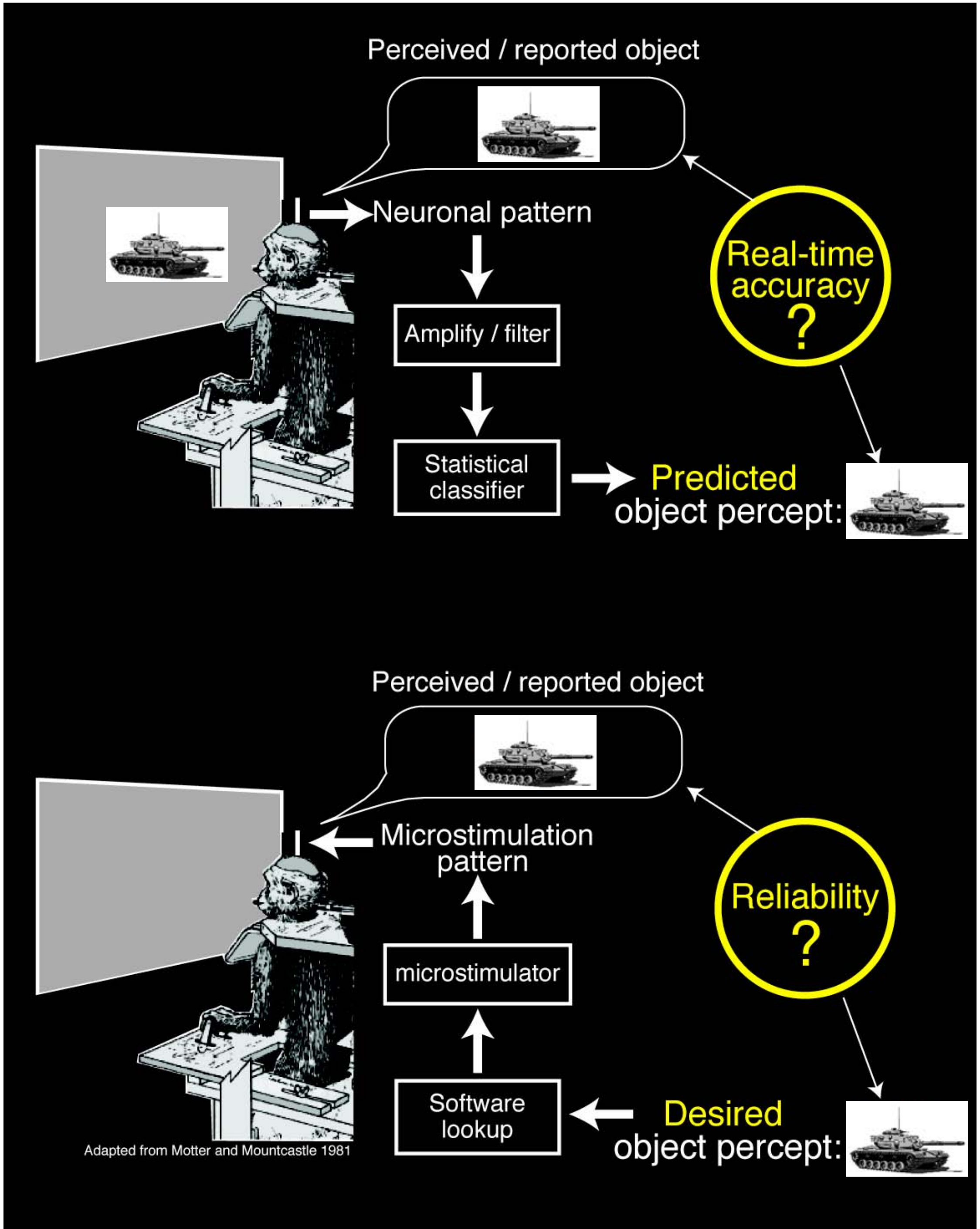
(Read-out eg analysis):

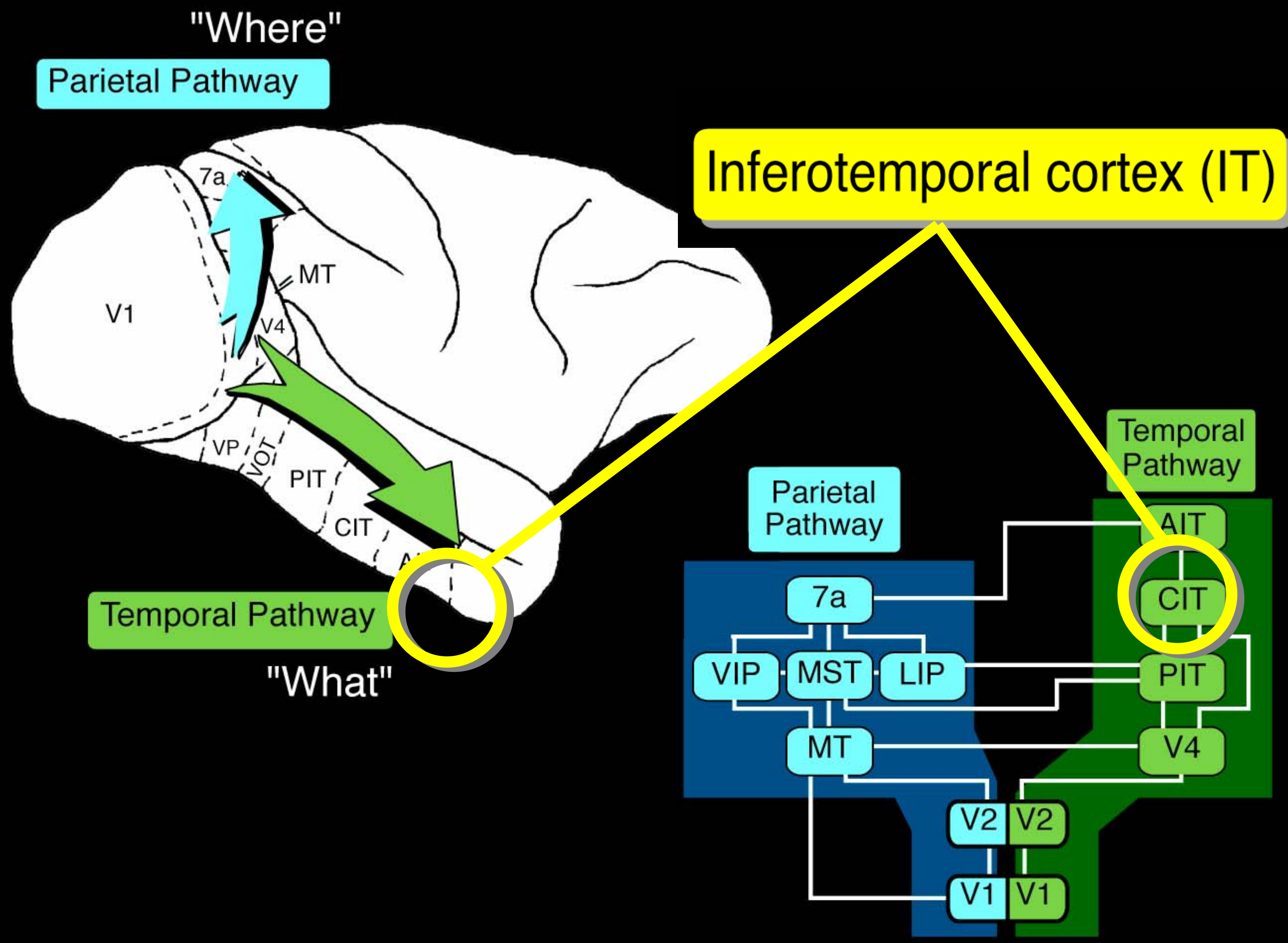
Can we “read-out” the subject’s object percept?

Goal 2

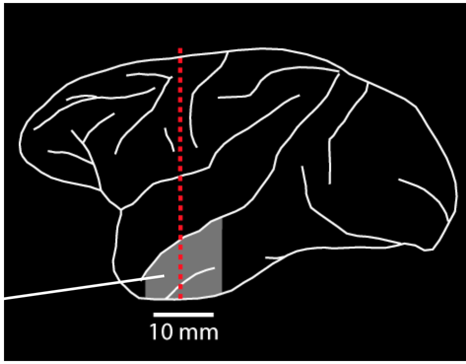
(Write-in eg synthesis):

Can we “write-in” (induce) an object percept?



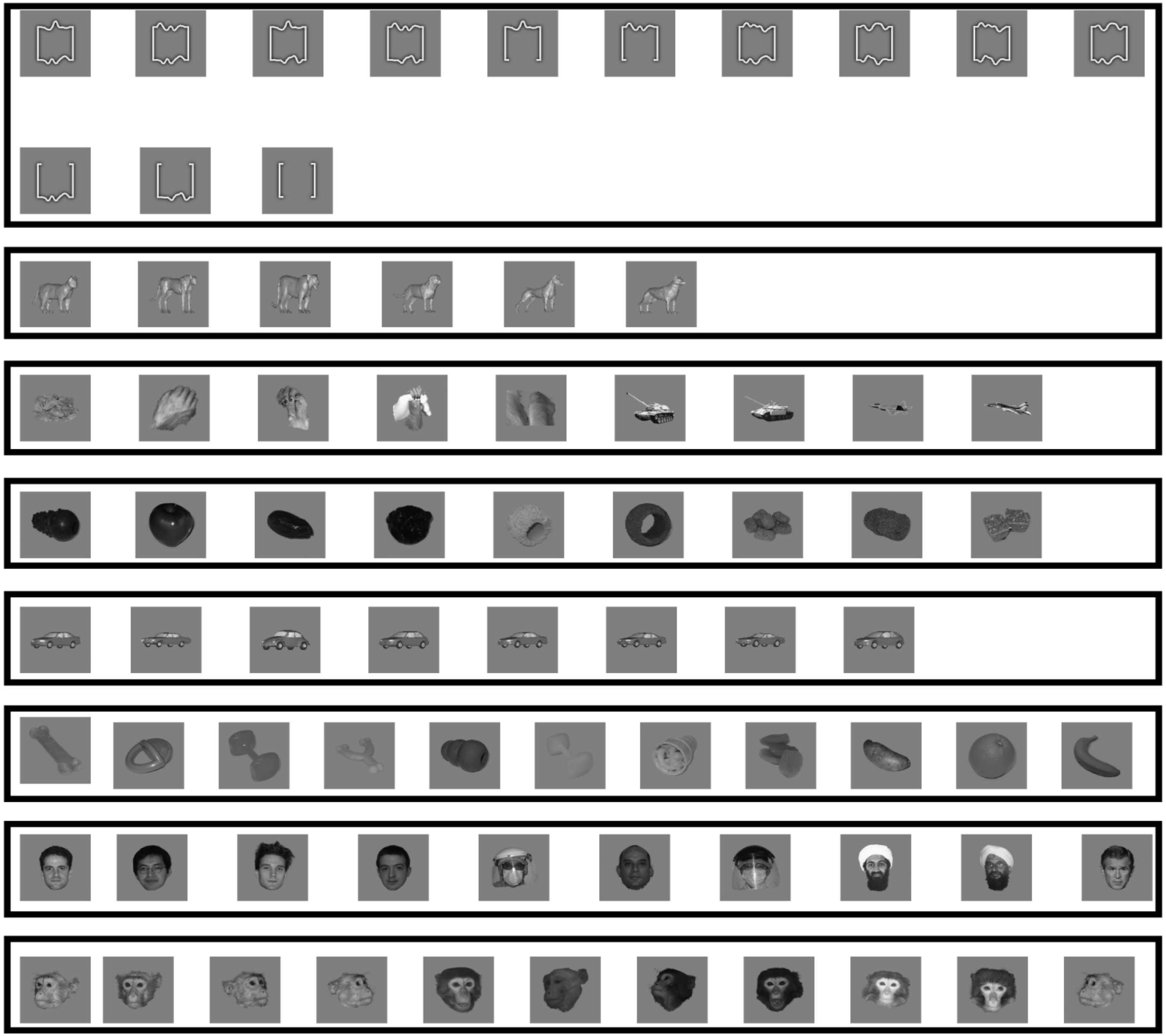

"Where"
Parietal Pathway
Inferotemporal cortex (IT)
7a
MT
V1
V4
VP
VOT
PIT
CIT
Temporal Pathway
"What"
Temporal Pathway
Parietal Pathway
AIT
7a
CIT
VIP
MST
LIP
PIT
MT
V4
V2
V2
V1
V1

Can we “read-out” the subject’s object percept from IT?

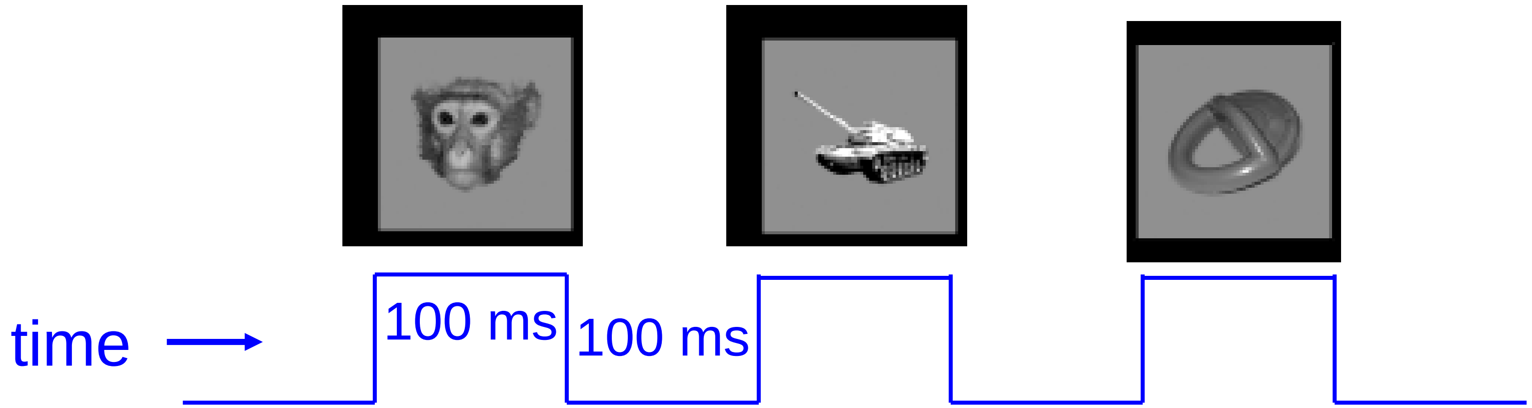


- **number of sites for reliable, real-time performance**
- **temporal properties (onset + integration scale) of object information**
- **neural code for different tasks**
- **invariance to object position, size, pose, illumination, clutter**
- **recognition: ‘classification’ vs. ‘identification’ ?**
- spatial scale of object information (single unit, multi-unit, LFP)
- stability of these neuronal codes?
- improvement with experience?
- …

77 objects, 8 classes

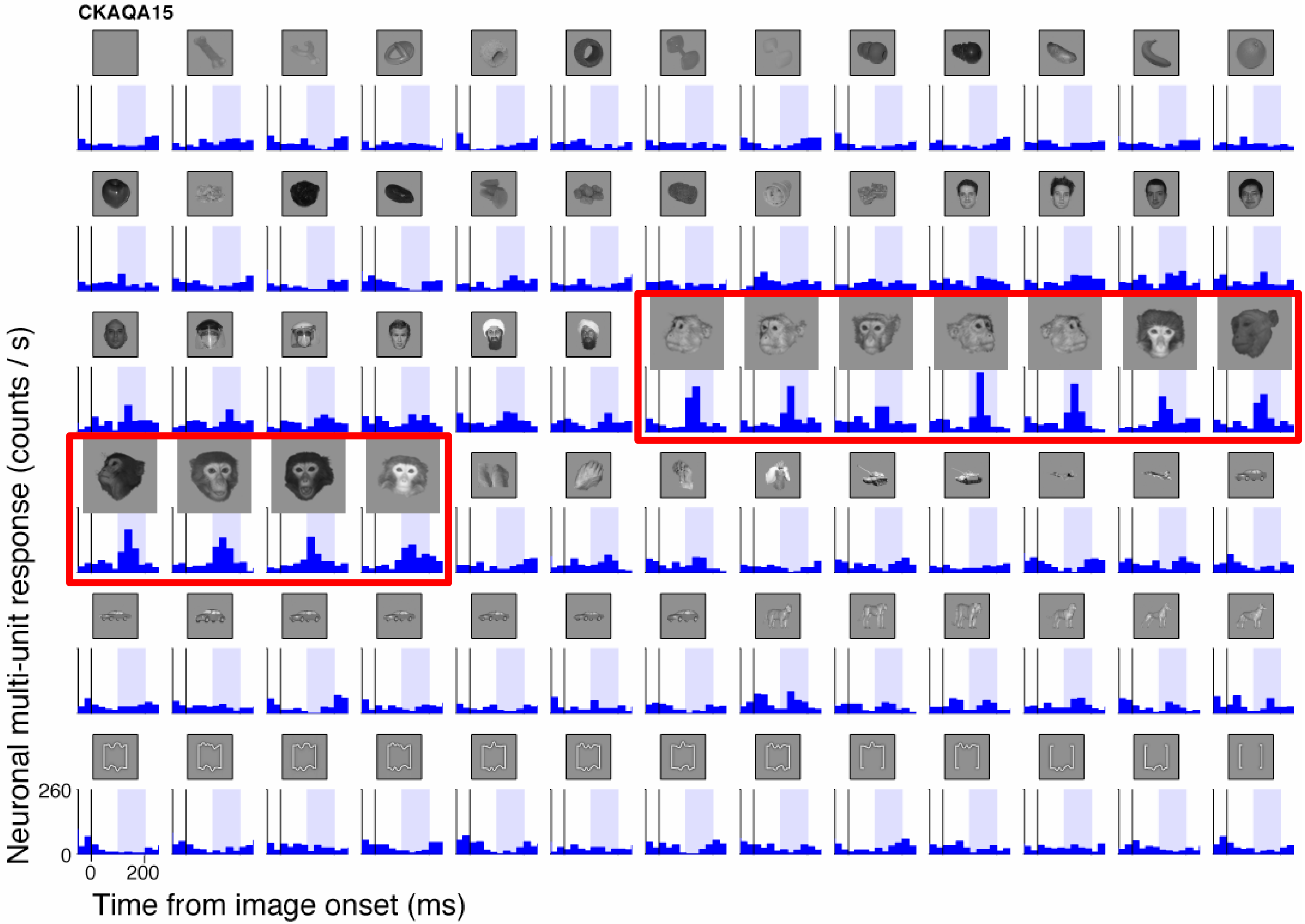


Rapid assessment of stimulus selectivity at each recording site during passive viewing

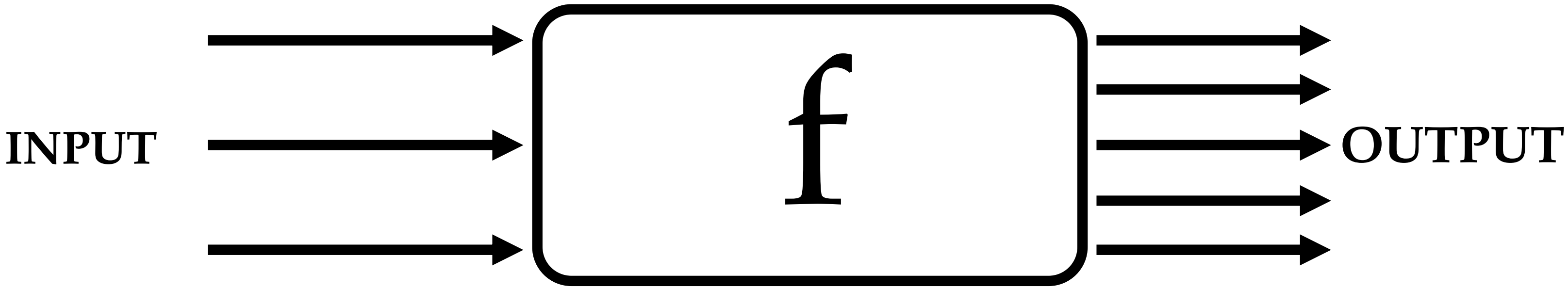


- 77 visual objects
- 10 presentation repetitions per object
- presentation order randomized and counter-balanced

Example AIT recording site



Training a classifier on neuronal activity.

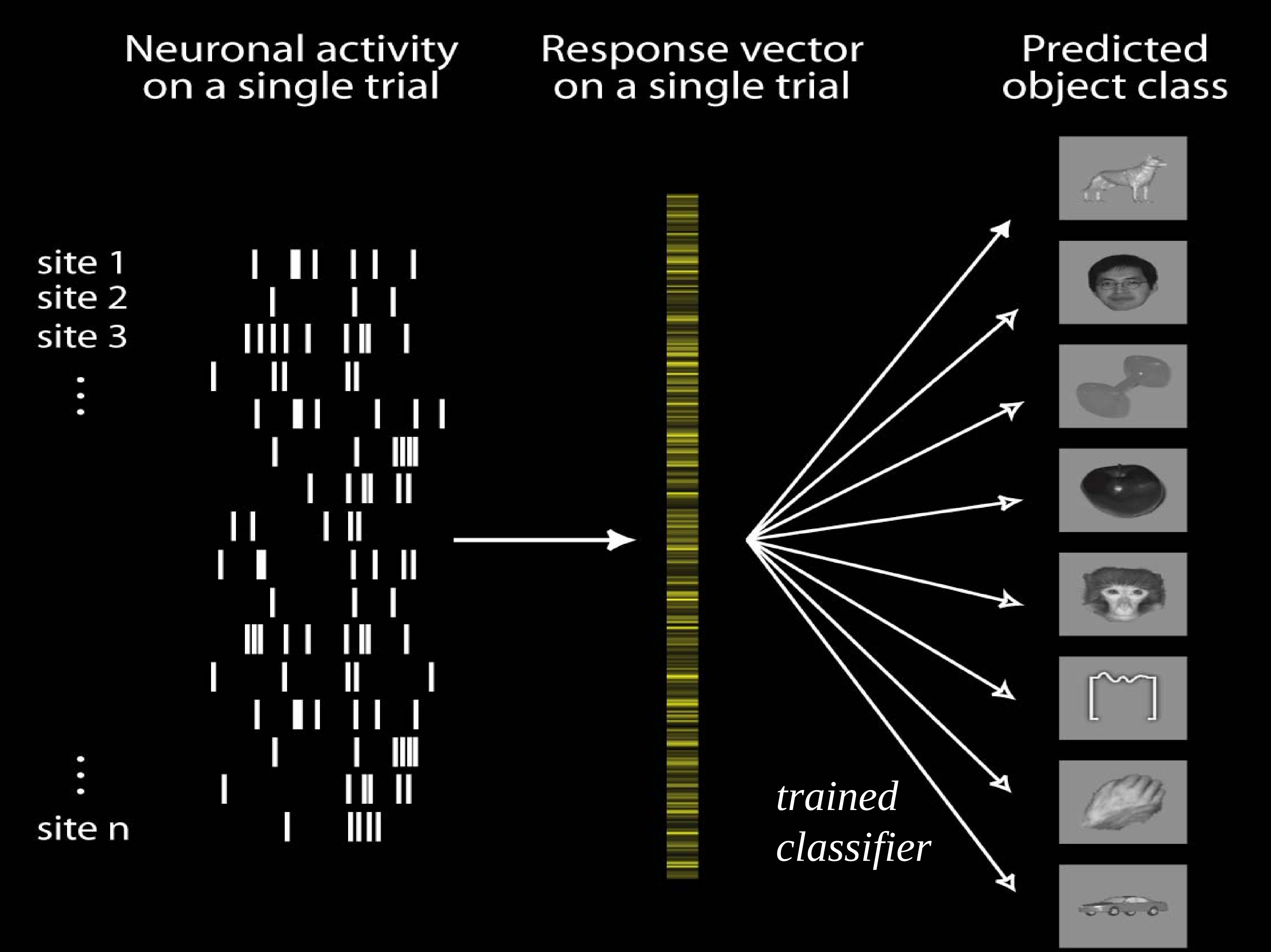


From a set of data (vectors of activity of n neurons (x) and object label (y)

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\}$$

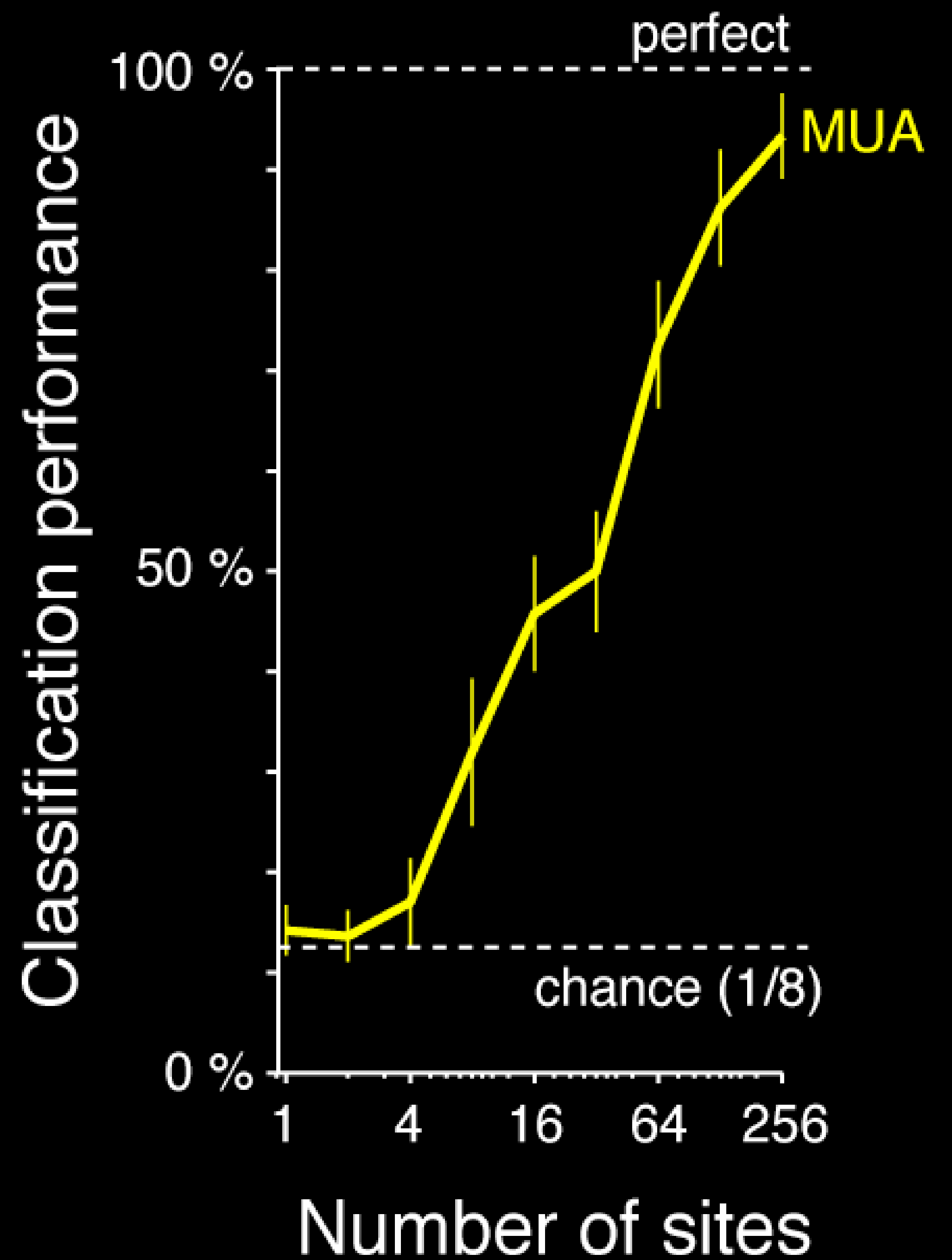
Synthesize (by training) a classifier eg a function *f* such $f(x) = \hat{y}$

is a ***good predictor*** of object label *y* for a ***future*** neuronal activity *x*



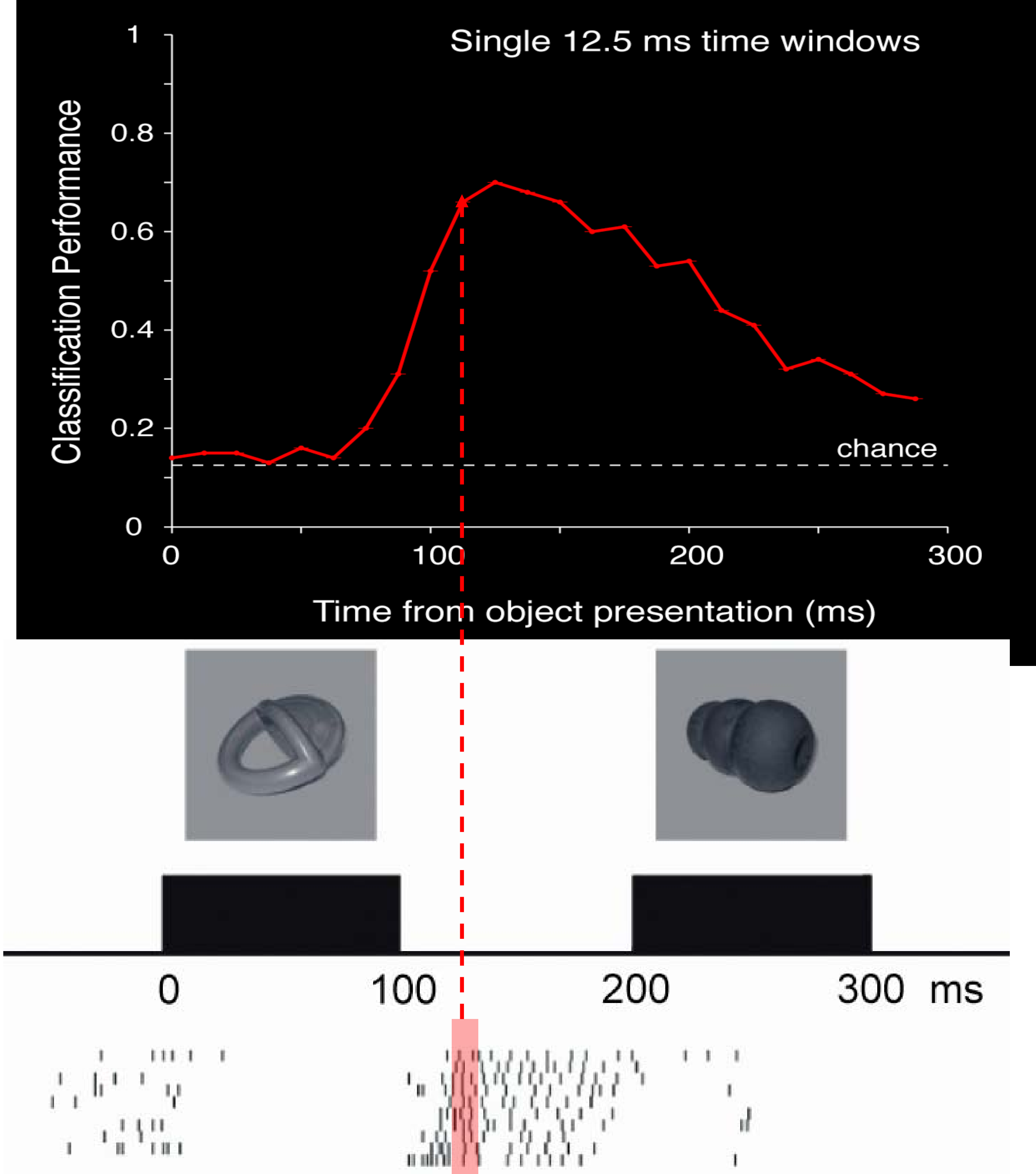
Neuronal activity on a single trial
Response vector on a single trial
Predicted object class
site 1
site 2
site 3
site n
trained classifier

First result: quite reliable object categorization using ~100 <u>arbitrary</u> AIT sites

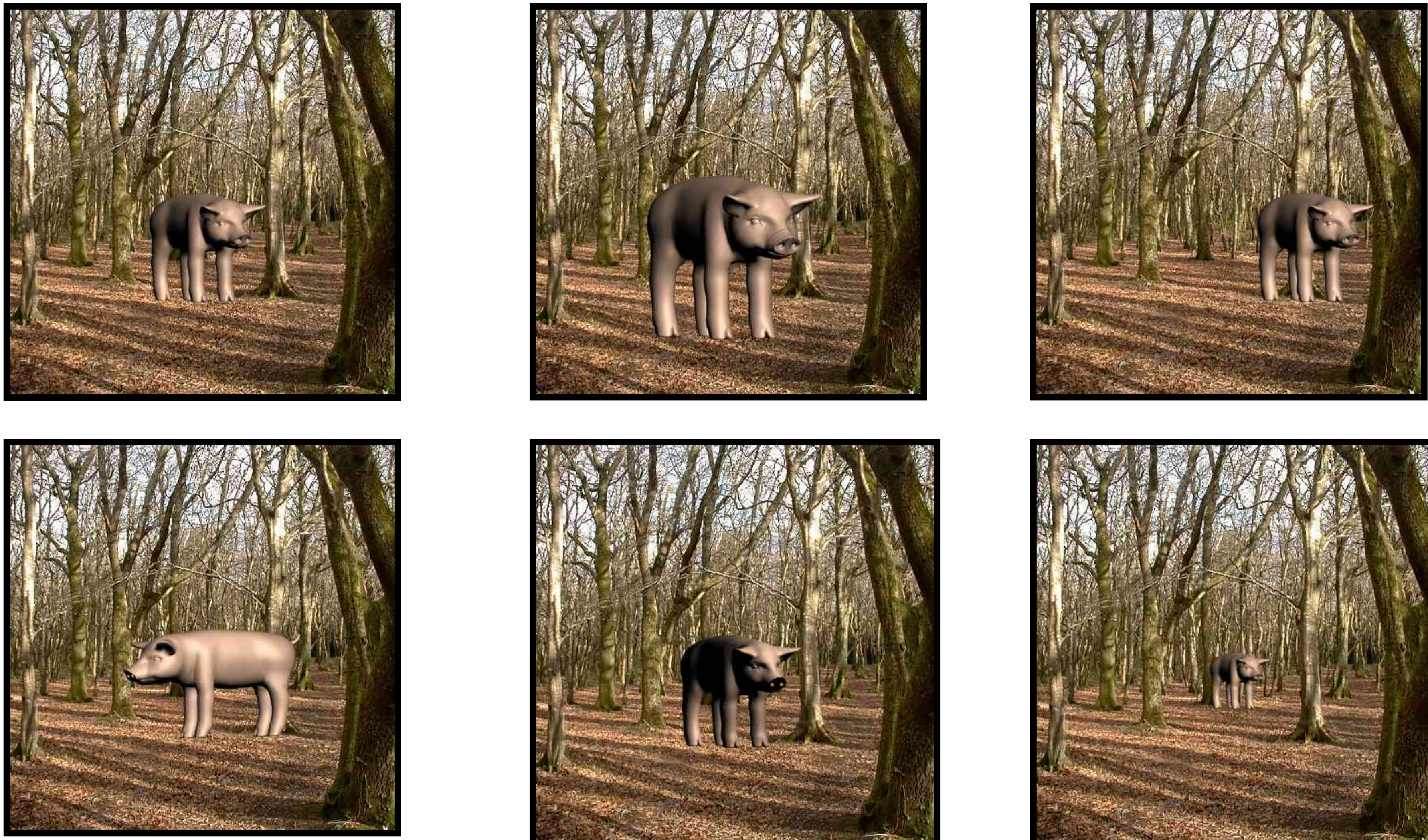


- *[100-300 ms] interval*
- *50 ms bin size*
- *4 bins per site*

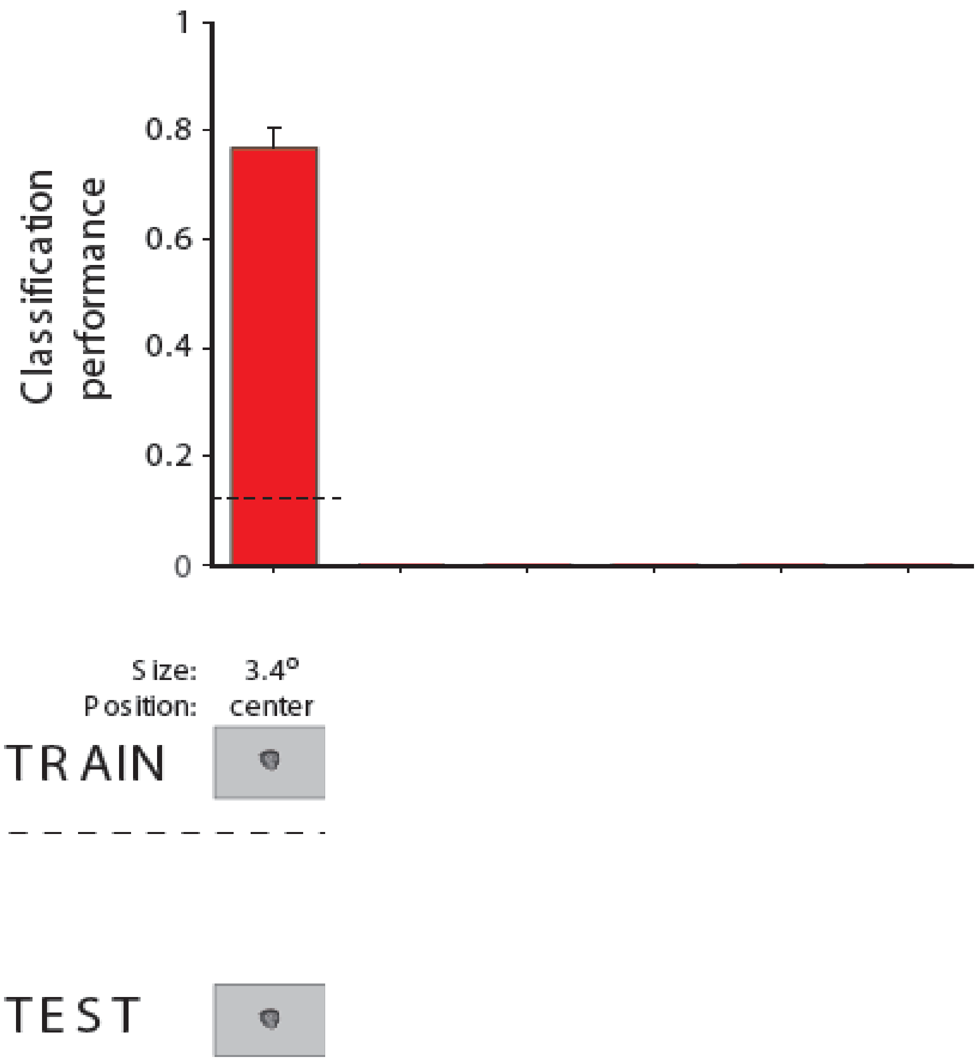
Very rapid read-out of object information



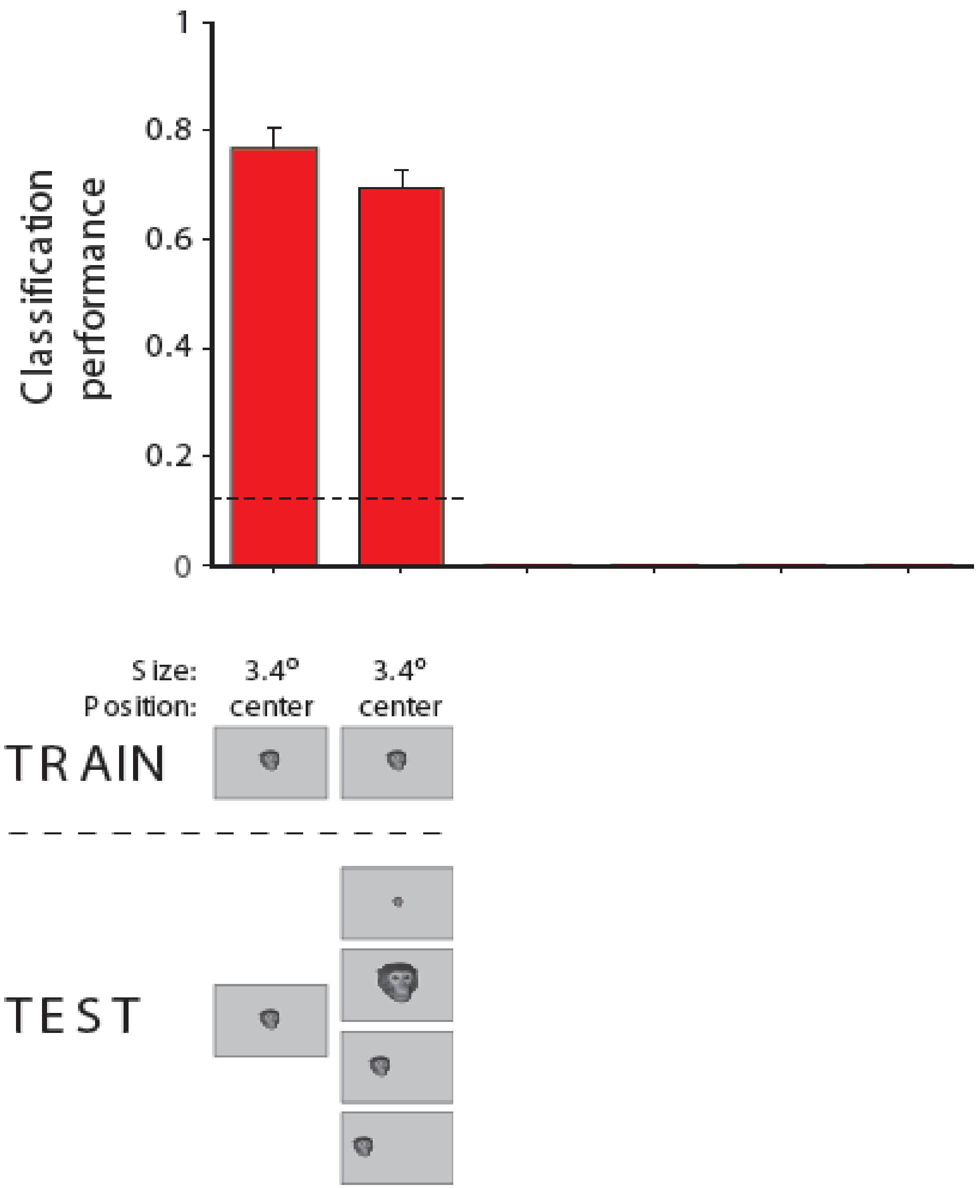
Is the representation in IT *selective* and *invariant* (which is the main goal of ventral stream)?



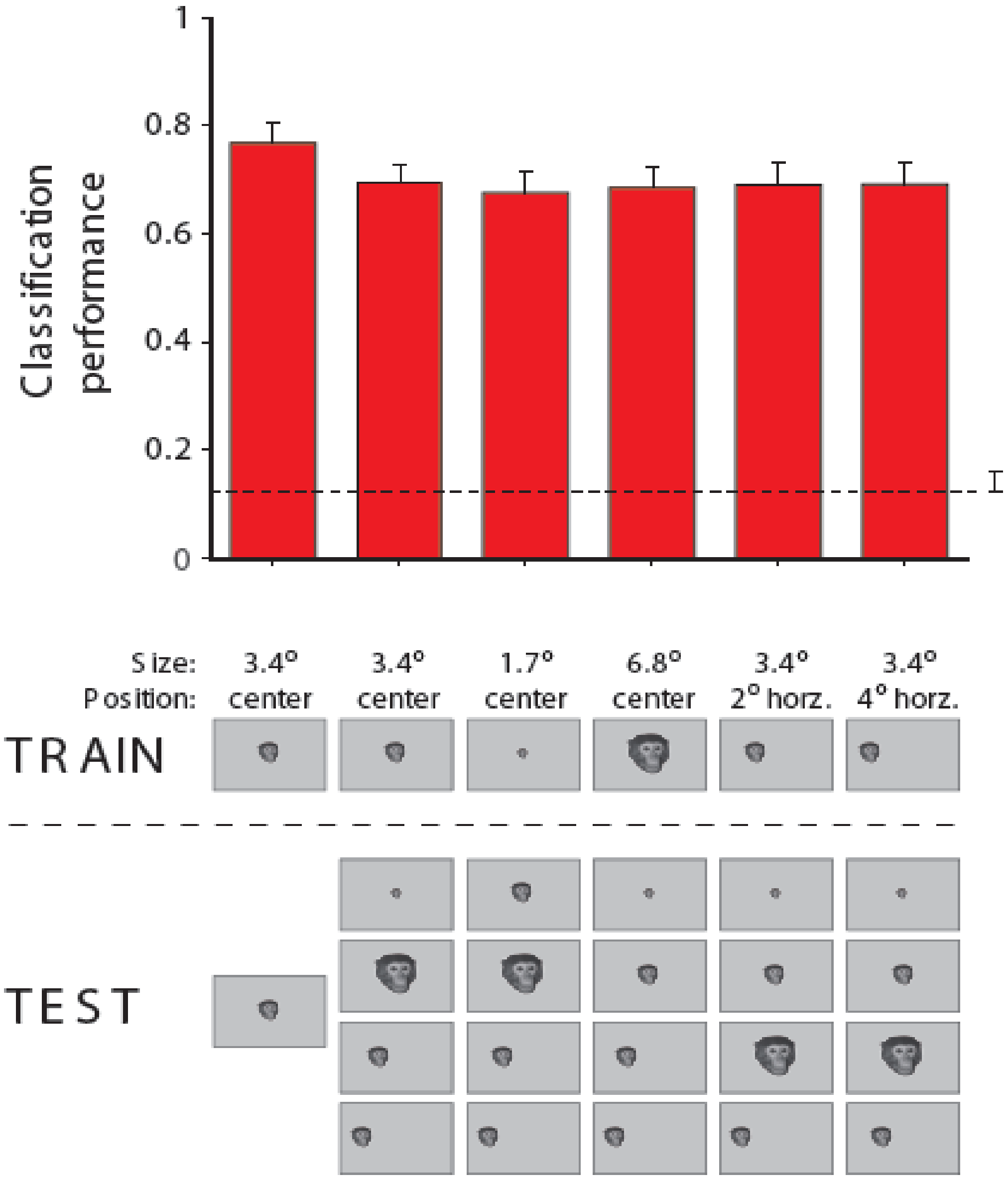
IT representation is invariant to changes in position and size



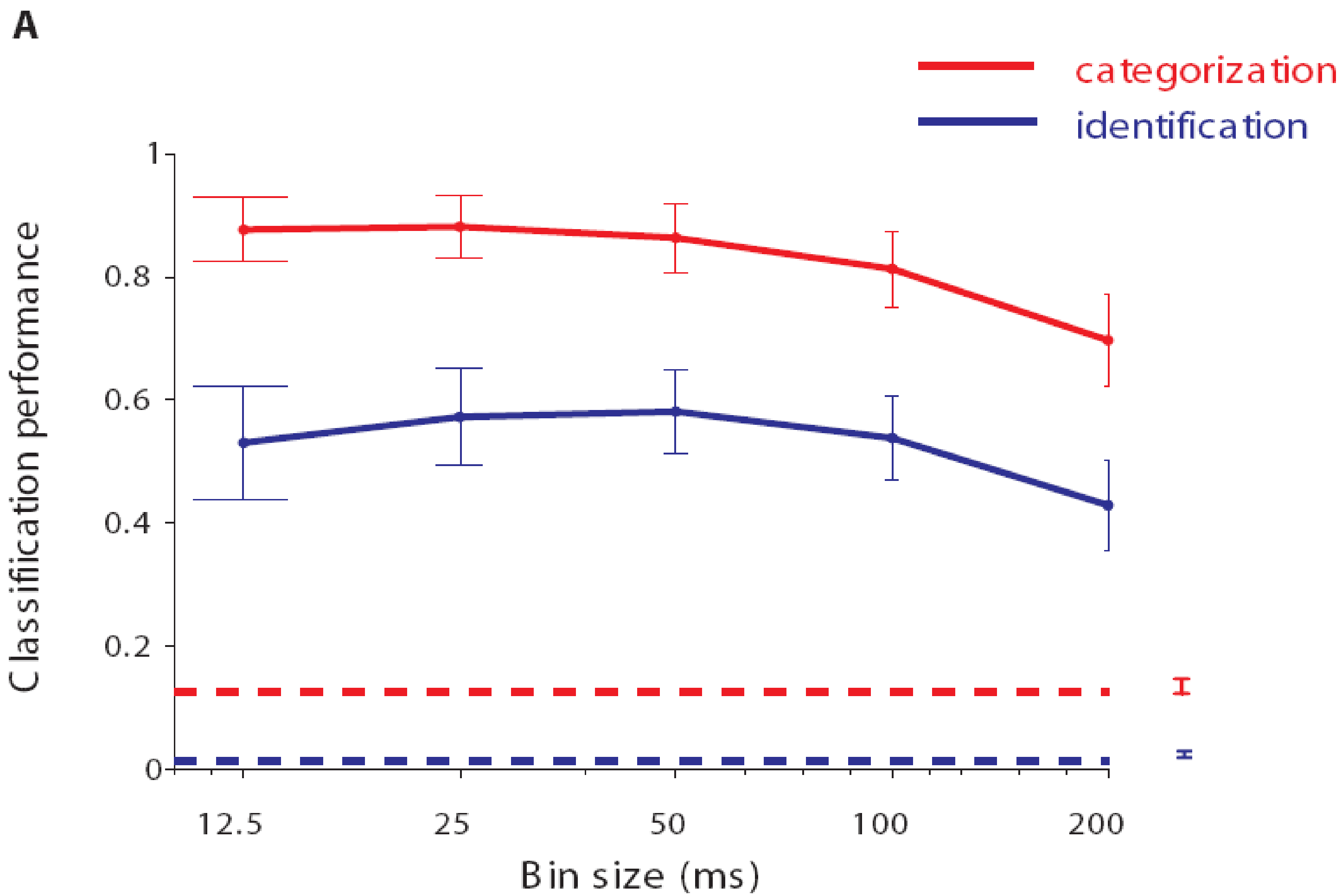
IT representation is invariant to changes in position and size



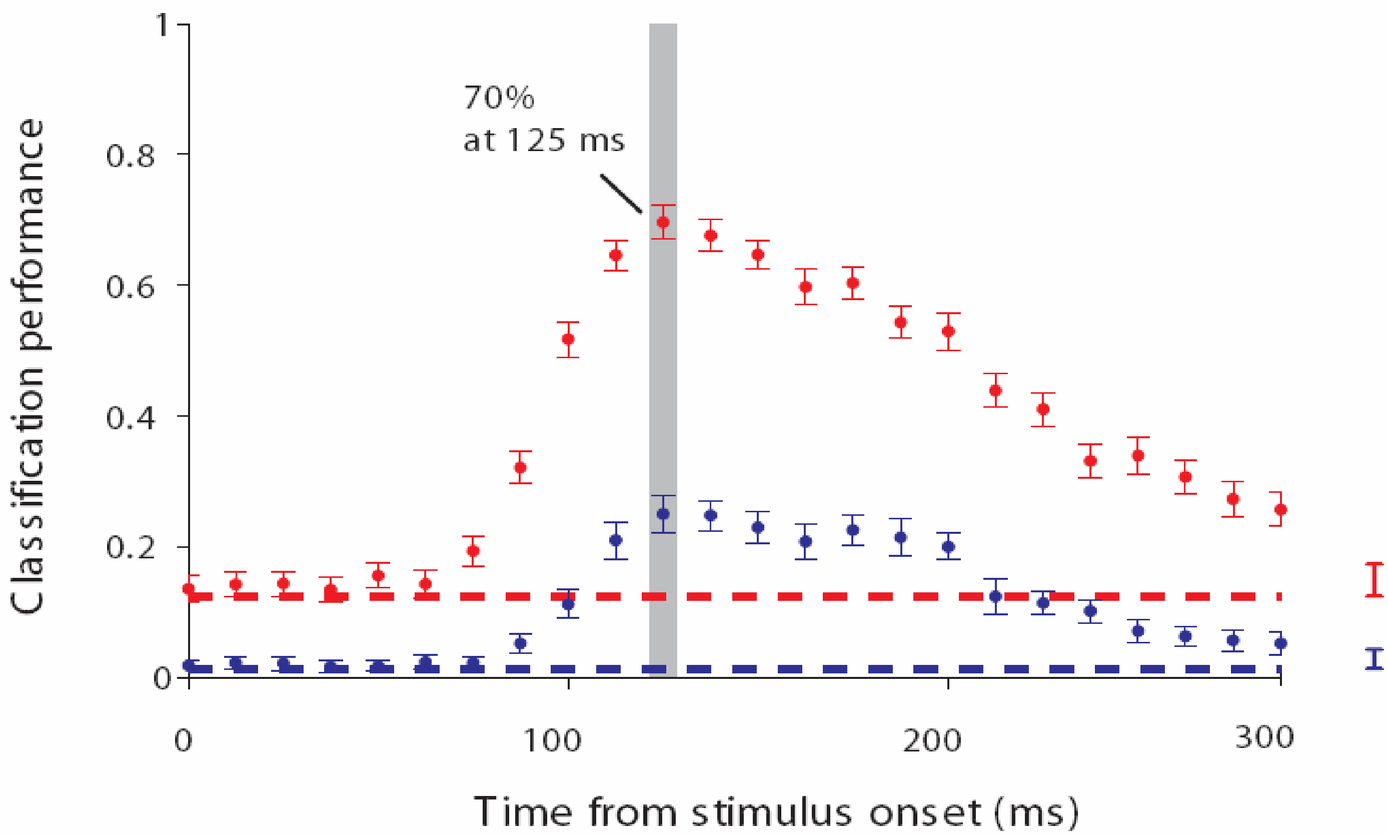
IT representation is invariant to changes in position and size



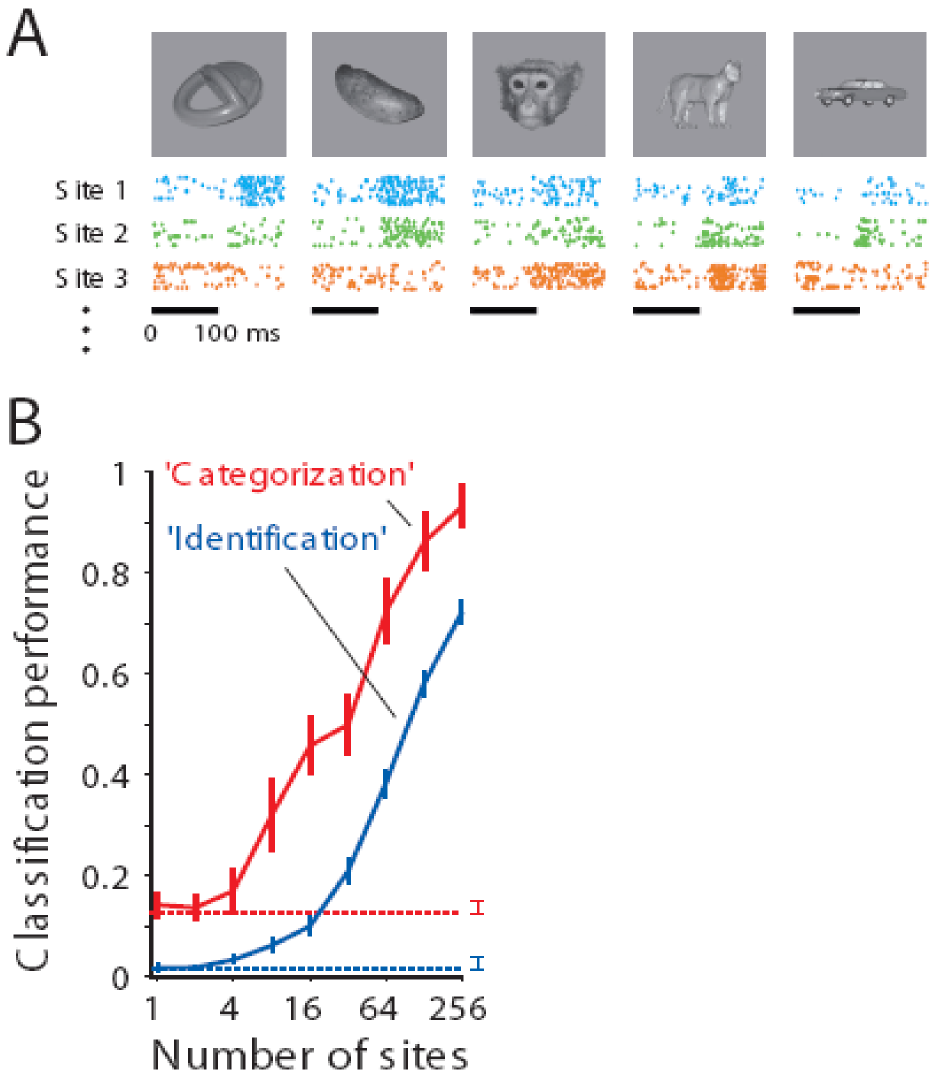
Neural code in IT: time resolution



Neural code in IT: latency and integration time

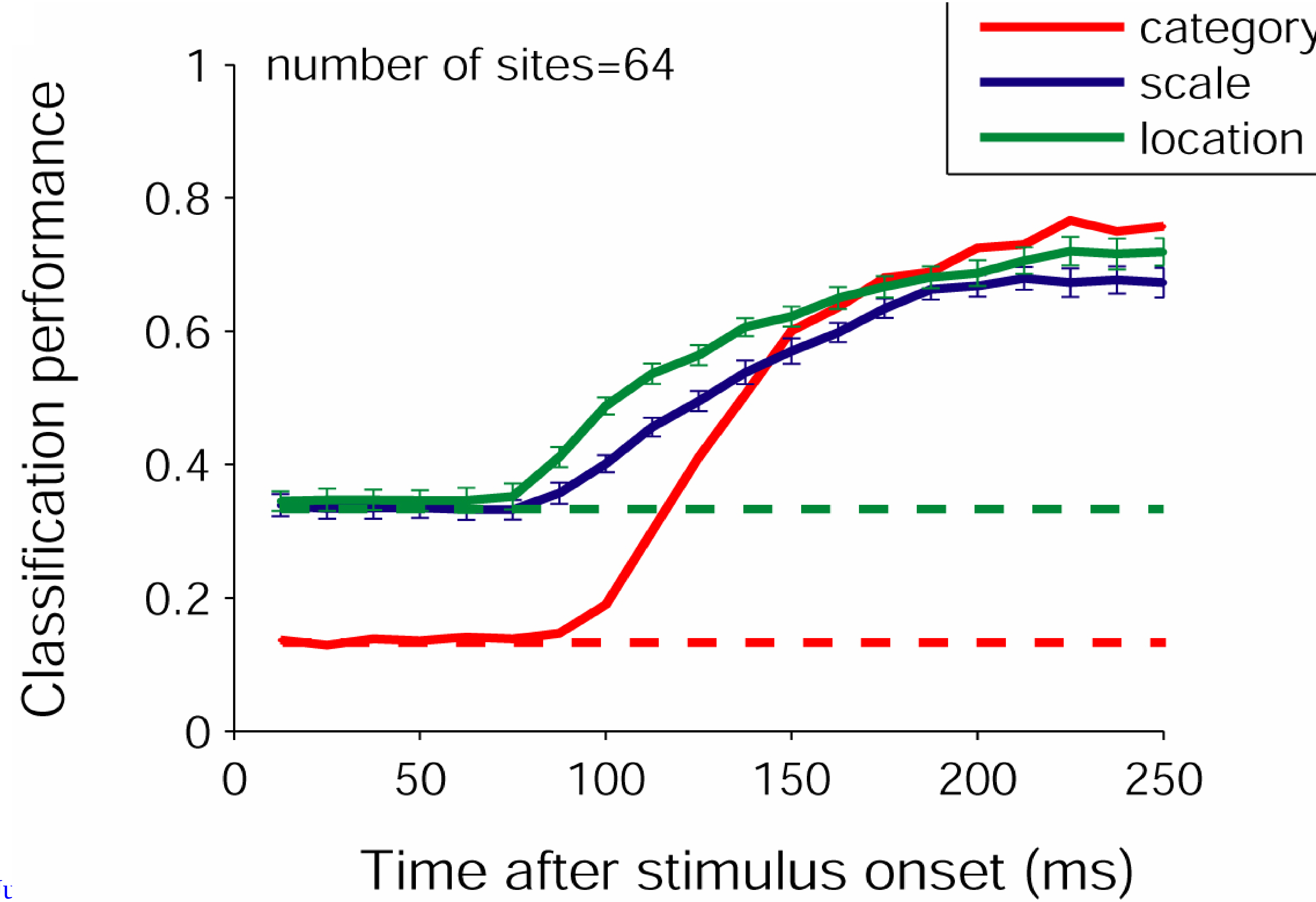


Categorization and identification

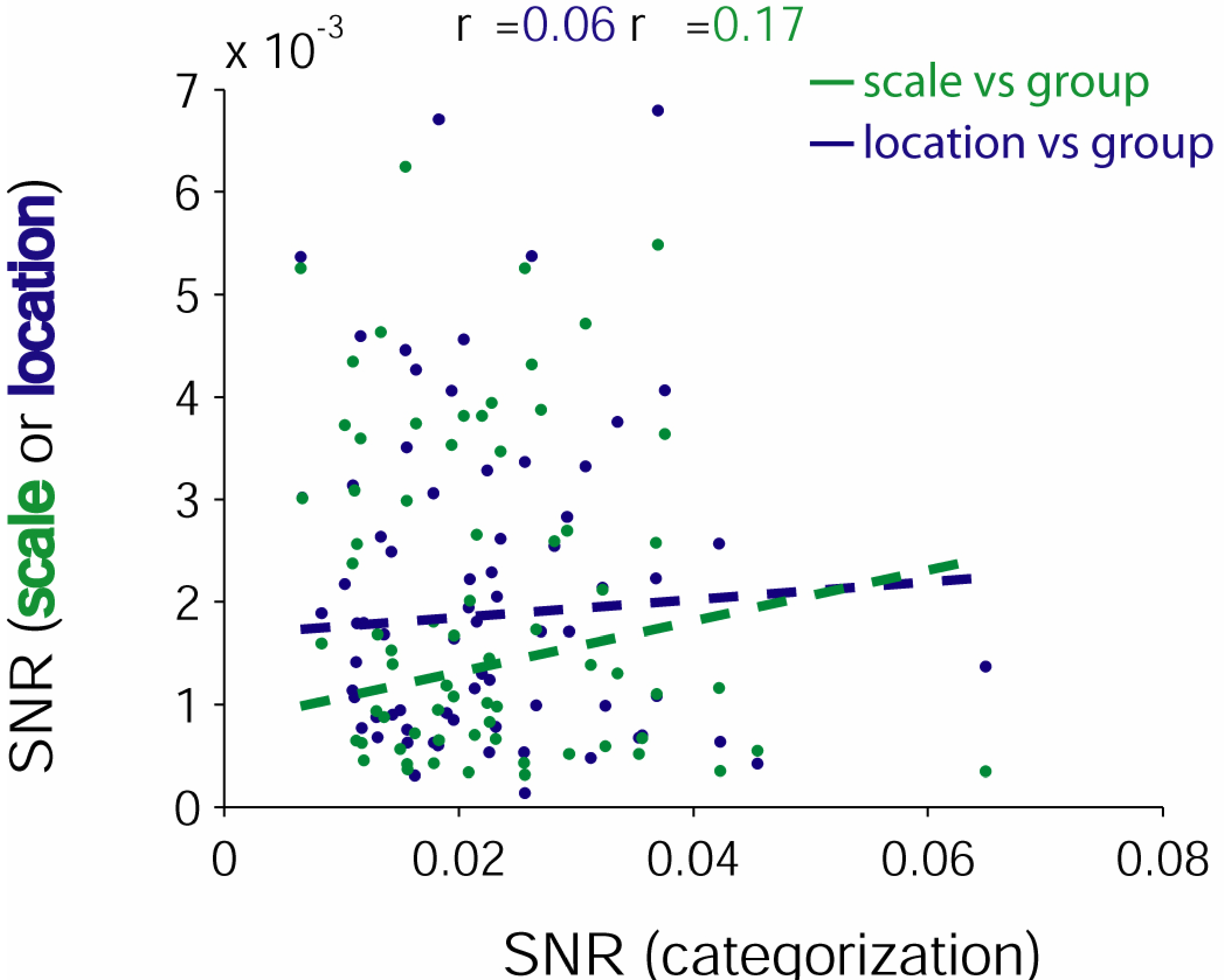


Some more details…

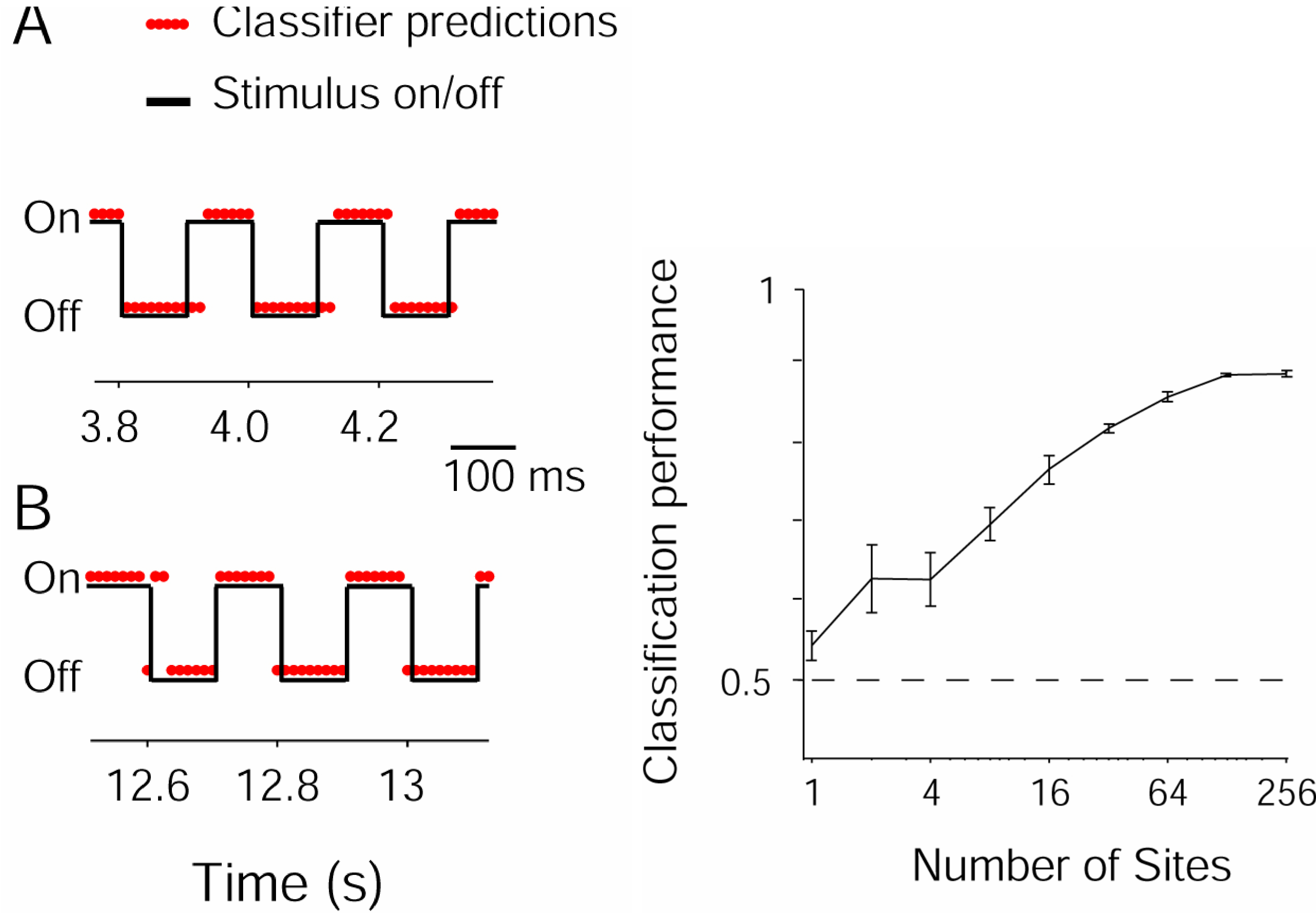
Reading out another type of object info: scale and location



How are different kinds of information coded?



Reading out another type of object info: stimulus onset



Thus IT contains a representation
which is invariant and selective enough
to allow very good,fast performance
by a linear classifier:

at the level of IT the recognition problem -
selectivity and invariance -- is “solved”.

How does the ventral stream do it?

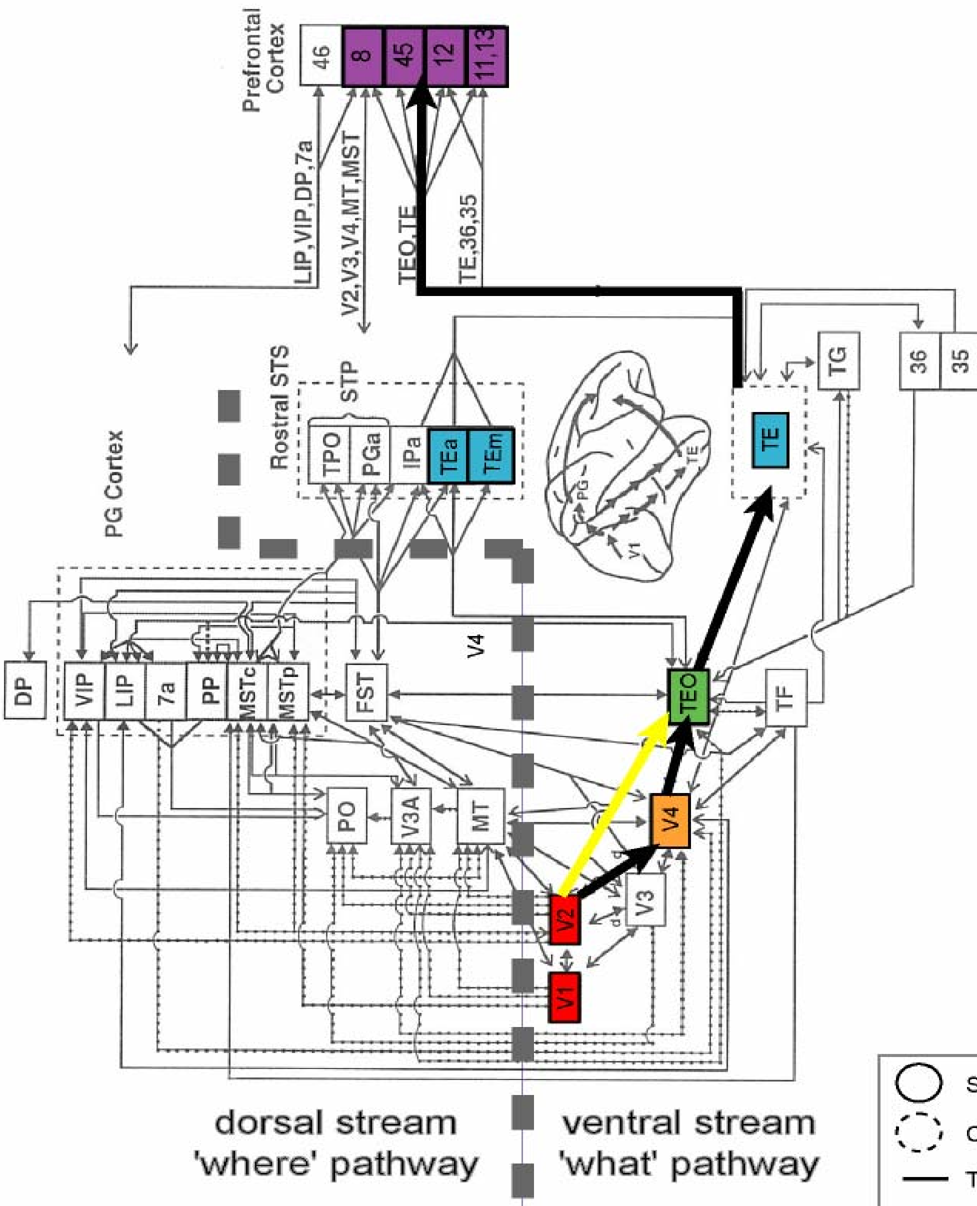
Now…back to *the* theory of the ventral stream of visual cortex

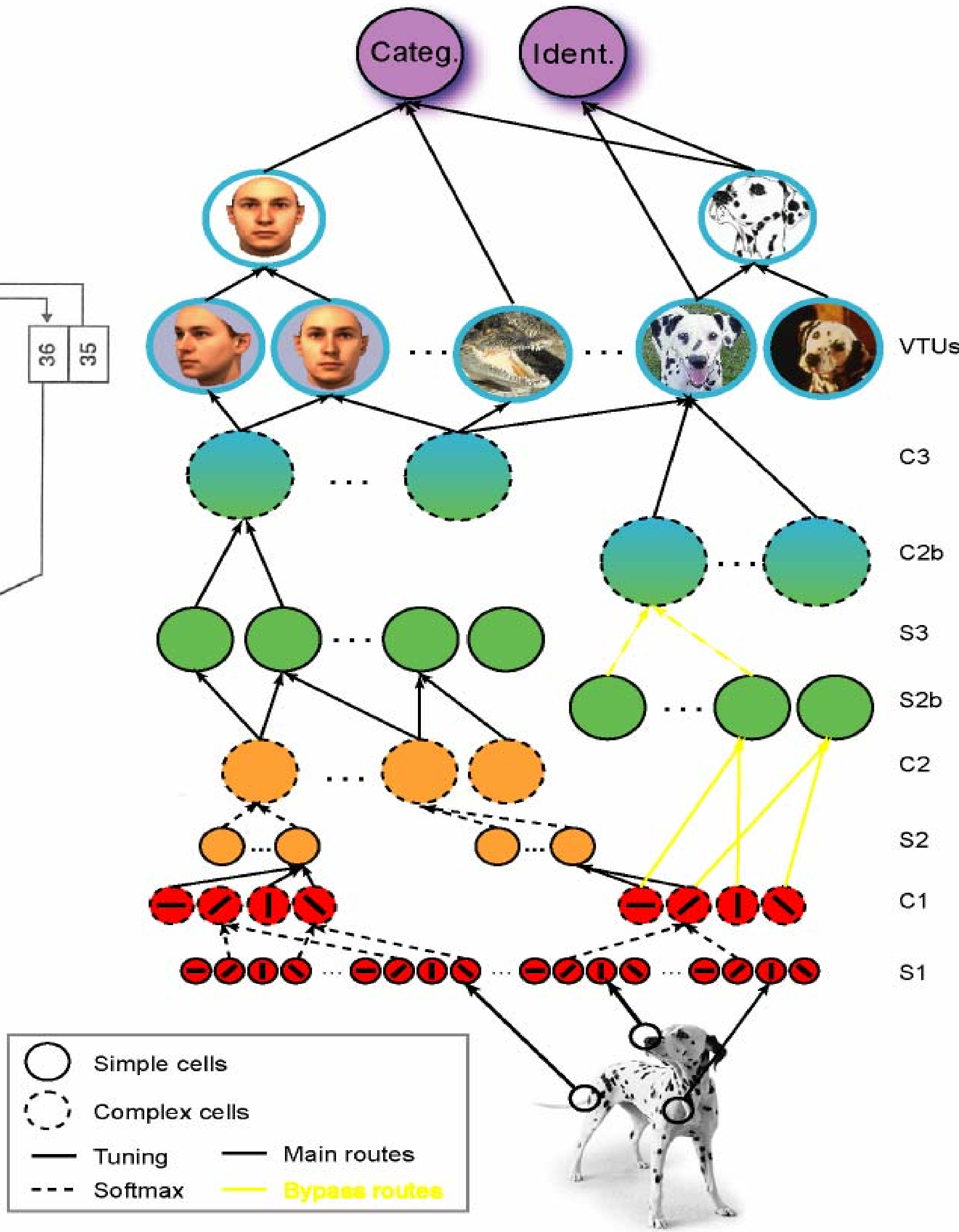
Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich
and Tomaso Poggio

The McGovern Institute for Brain Research,
Department of Brain Sciences
Massachusetts Institute of Technology

CalTech, July 2005

Mapping the ventral stream into a model





Serre, Kouh, Cadieu, Knoblich, Poggio, 2005

Main assumptions of theory

- **Feedforward architecture**

- **Two basic operations**
  -- *tuning* and *softmax* --
  repeated at simple and complex stages
  from V1 to V2 to V4 and IT
  underlie selectivity and invariance of recognition

- **Learning (passive, task independent) at S levels and supervised, task dependent at the level IT→ PFC**

Poggio, T. and E. Bizzi. Generalization in Vision and Motor Control, *Nature*, Vol. 431, 768-774, 2004

Two basic operations

Tuning in simple cells for selectivity:

$$S = \frac{\sum_{j=1}^{n} w_j \; x_j^p}{c + \left(\sum_{j=1}^{n} x_j^q\right)^r}$$

Extra sigmoid transfer function can control the sharpness of tuning to approximate full RBF tuning

Soft-max in complex cells for invariance:

$$C = \frac{\sum_{j=1}^{n} x_j^{q+1}}{c + \sum_{j=1}^{n} x_j^q}$$

Two basic operations

Tuning in simple cells for selectivity:

$$S = \frac{\sum_{j=1}^{n} w_j \; x_j^p}{c + \left(\sum_{j=1}^{n} x_j^q \right)^r}$$

Combine units with different preferred stimulus

Extra sigmoid transfer function can control the sharpness of tuning, approximate RBF tuning

Soft-max in complex cells for invariance:

Combine units with same preferred stimulus but at slightly different scale and position

$$C = \frac{\sum_{j=1}^{n} x_j^{q+1}}{c + \sum_{j=1}^{n} x_j^q}$$

Two basic operations

How could those two types of receptive field be learned from visual experience?

Tuning in simple cells for selectivity:

$$S = \frac{\sum_{j=1}^{n} w_j \; x_j^p}{c + \left(\sum_{j=1}^{n} x_j^q \right)^r}$$

Combine units with different preferred stimulus

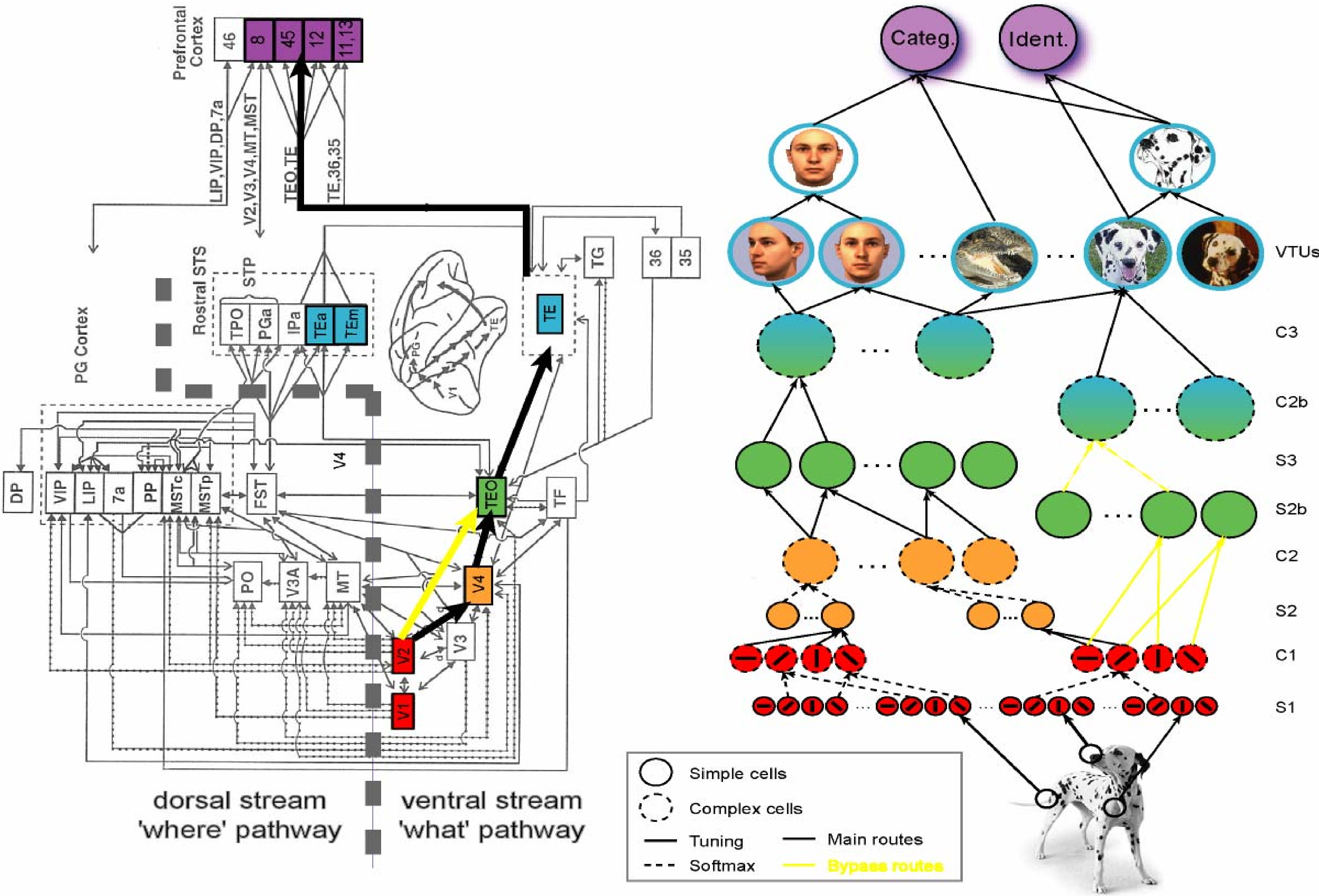
Extra sigmoid transfer function can control the sharpness of tuning, approximate RBF tuning

Soft-max in complex cells for invariance:

Combine units with same preferred stimulus but at slightly different scale and position

$$C = \frac{\sum_{j=1}^{n} x_j^{q+1}}{c + \sum_{j=1}^{n} x_j^q}$$

Mapping the ventral stream into a model



Serre, Kouh, Cadieu, Knoblich, Poggio, 2005

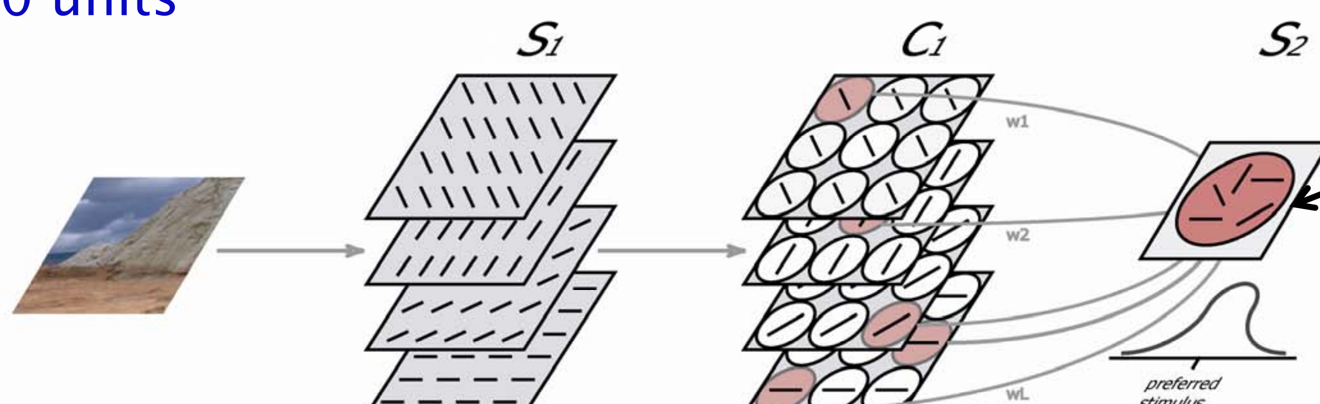
Learning a large universal and overcomplete dictionary of visual shape-components (a version of trace rule)

$$S = \frac{\sum_{j=1}^{n} w_j \; x_j^p}{c + \left(\sum_{j=1}^{n} x_j^q \right)^r}$$

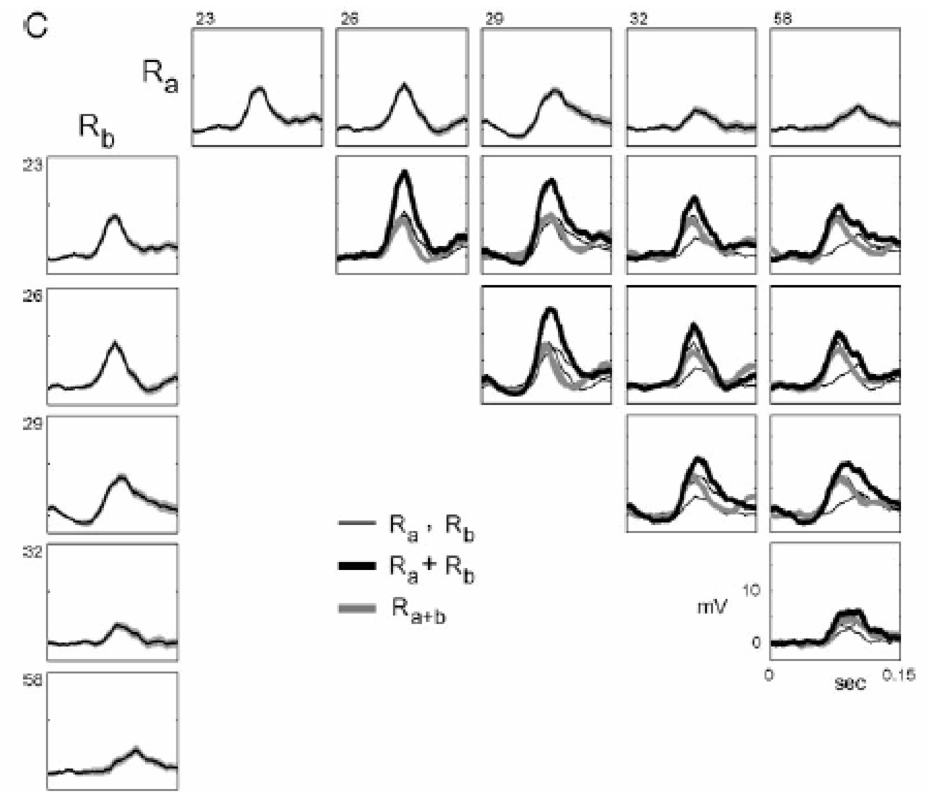
Passive exposure of patches of natural images

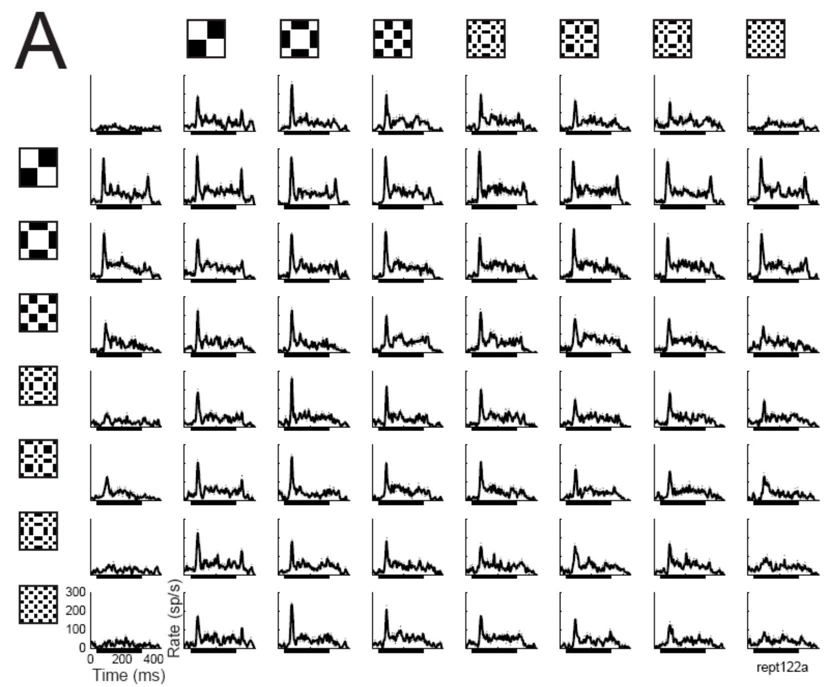
Imprinting of the synaptic weights

~100,000 units



Experimental support for a Max operation in complex cells (cat area 17) and in V4?

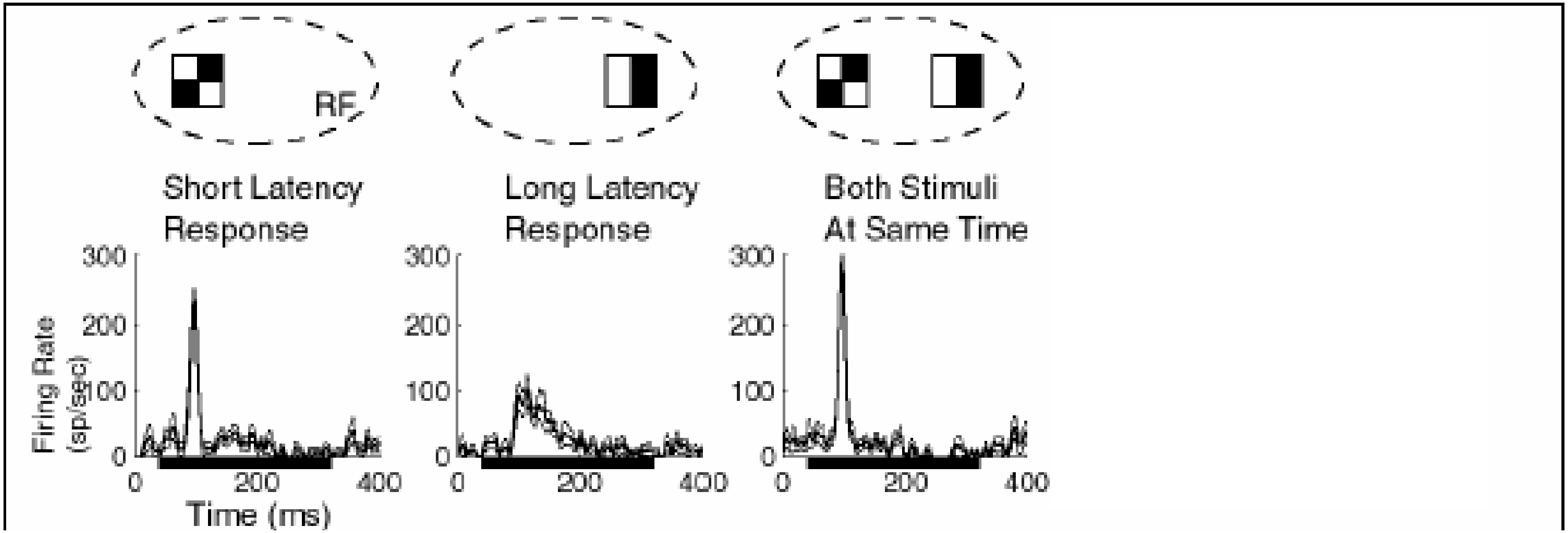




Gawne & Martin, *J. Neurophys.*, 2002

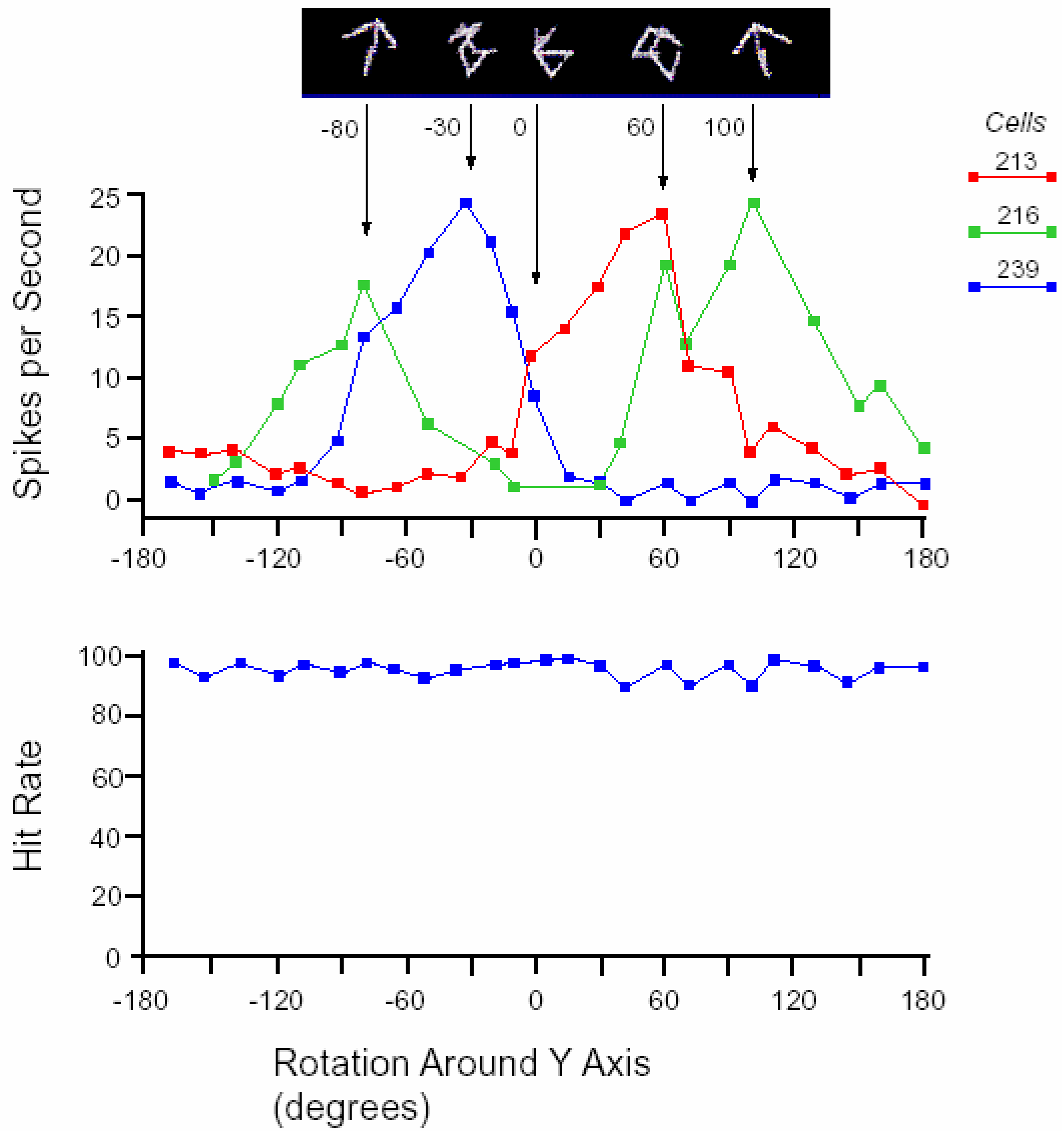
Lampl, Ferster, Poggio, Riesenhuber, *J. Neurophys, 2004.*

Under appropriate conditions…Max operation in V4 cells?



Gawne et al., *J. Neuro.*, 2002

There is also evidence for Gaussian-like tuning in V1, V2 and IT cortex….



Logothetis et al., Cur. Bio., 1995

Summary I: support for the model

- Several complex cell-like neurons (in V1 and V4) seem to perform a softmax operation
- Quantitative generalization properties in IT
- IT response to scrambling , presence of distractors and clutter.
- Learning a categorization task (cats vs. dogs) in IT and PFC units.
- Model learns from natural images and generates a vocabulary of C2 units consistent with V4 data.
- At the cognitive level model predicts several aspects of the face

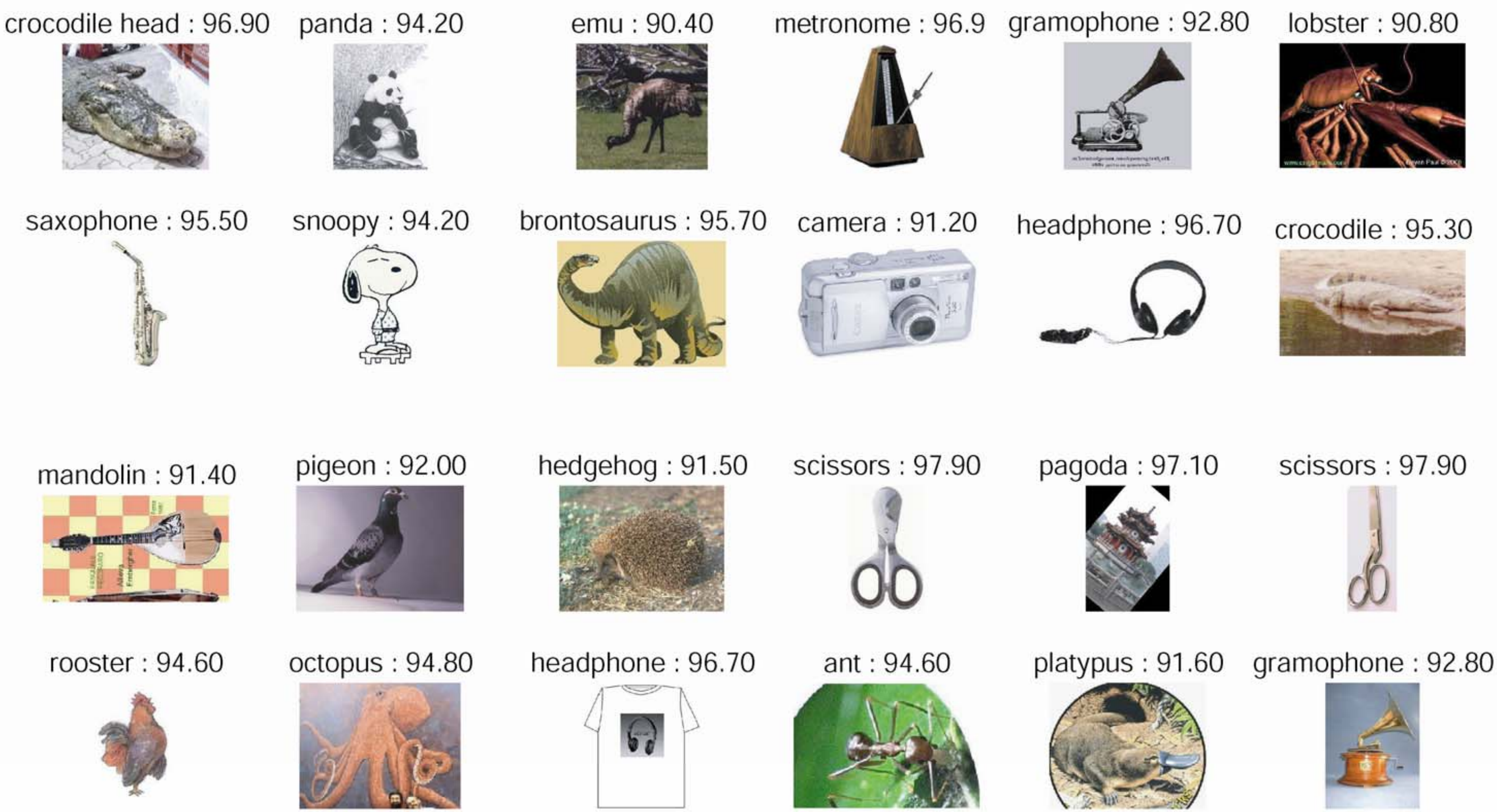
inversion effect.

Now a surprise (for us)…

…comparison of the updated model

with machine vision performance

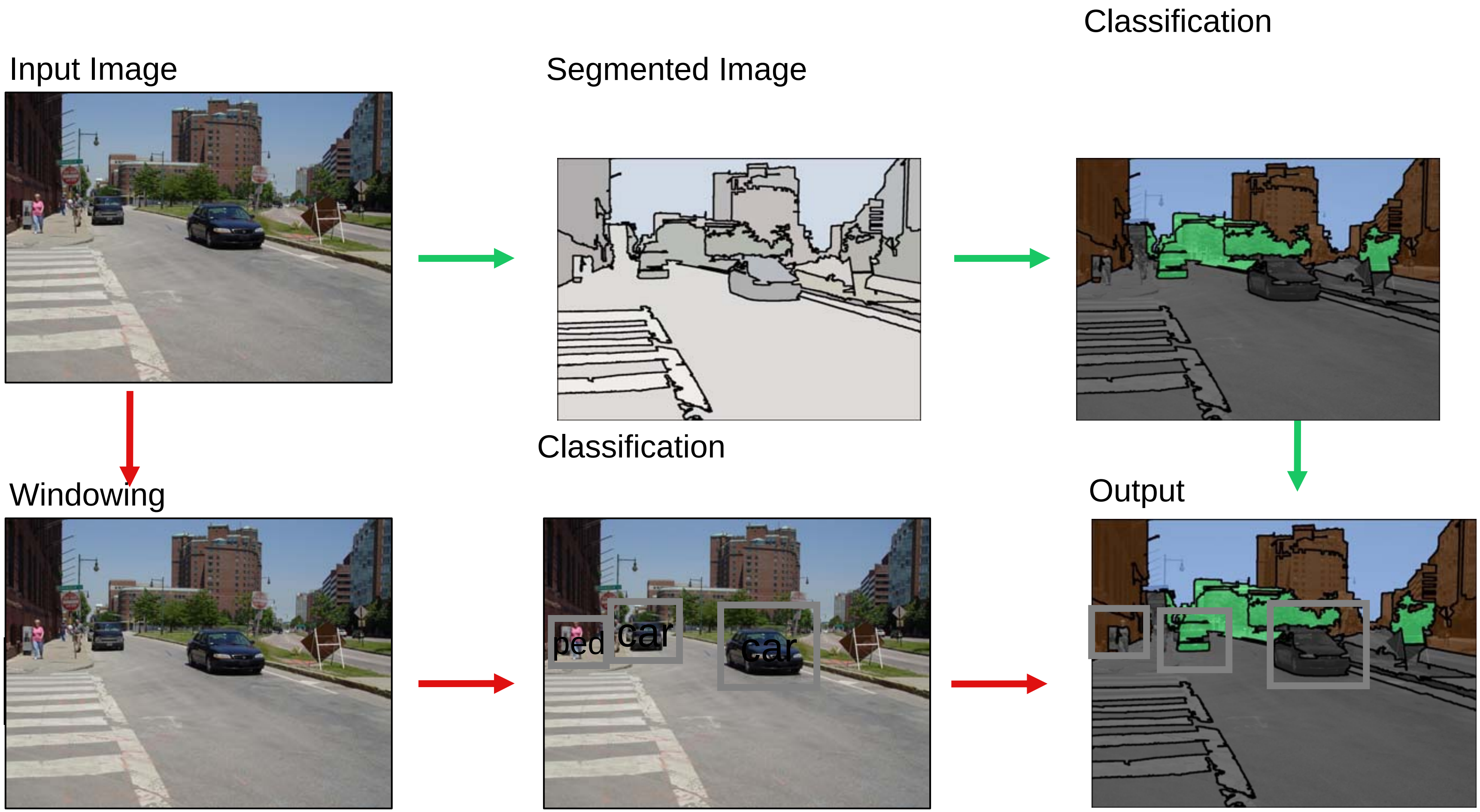
Sample Results on the 101-object dataset



The model performs at the level of the best computer vision systems

Datasets	Benchmark		Model
Leaves (Calt.)	Weber, Welling and Perona, 2000	84.0	97.0
Cars (Calt.)	Fergus, Perona and Zisserman, 2003	84.8	99.7
Faces(Calt.)	Fergus, Perona and Zisserman, 2003	96.4	98.2
Airplanes (Calt.)	Fergus, Perona and Zisserman, 2003	94.0	96.7
Moto. (Calt.)	Fergus, Perona and Zisserman, 2003	95.0	98.0
Faces(MIT)	Heisele, Serre and Poggio, 2002	90.4	95.9
Cars (MIT)	Torralba, Murphy and Freeman, 2004	75.4	95.1

Sample results on the CBCL StreetScenes database



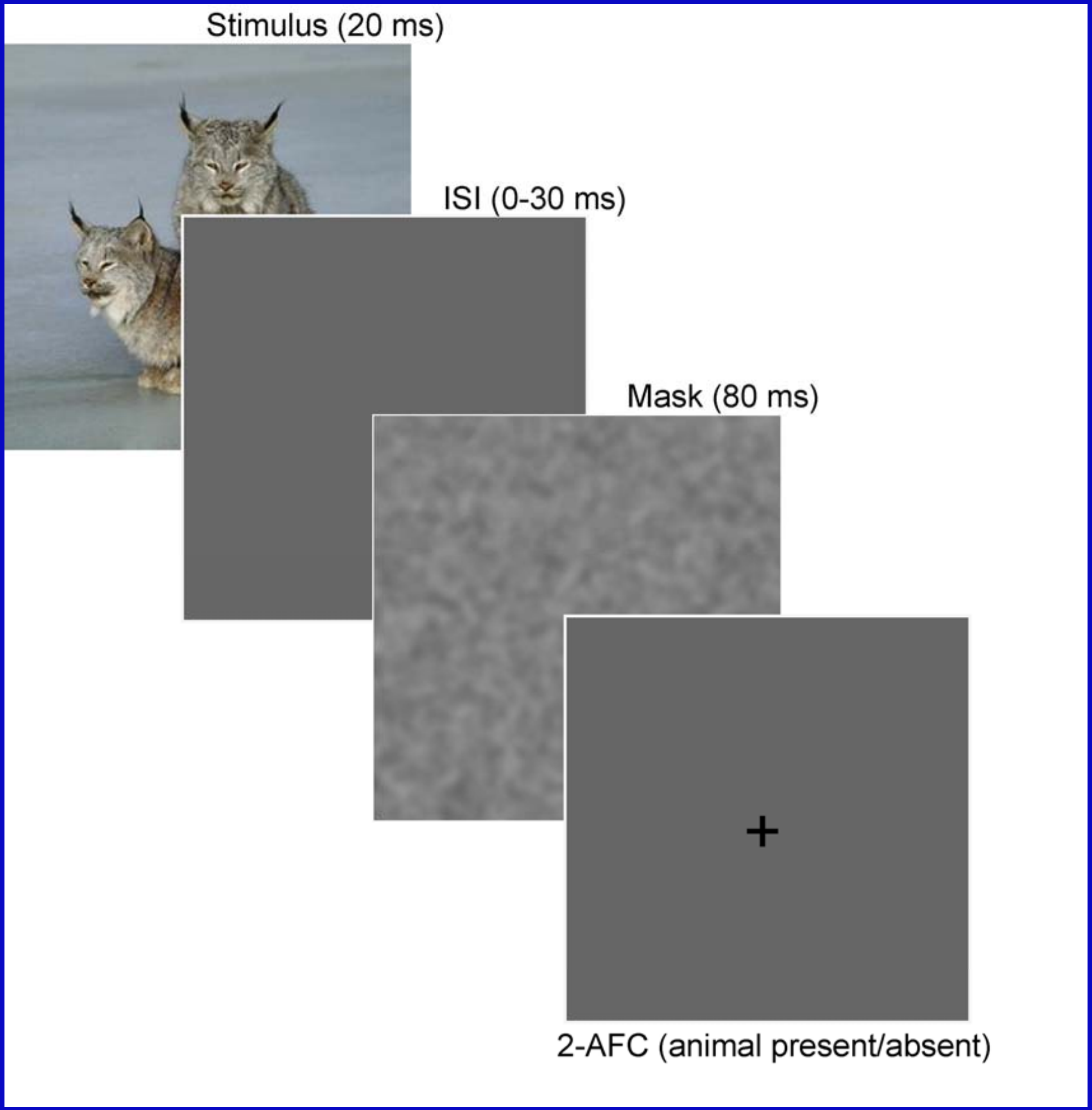
Texture-based objects (e.g., trees, road, sky, buildings)

Shape-objects (e.g., pedestrians, cars)

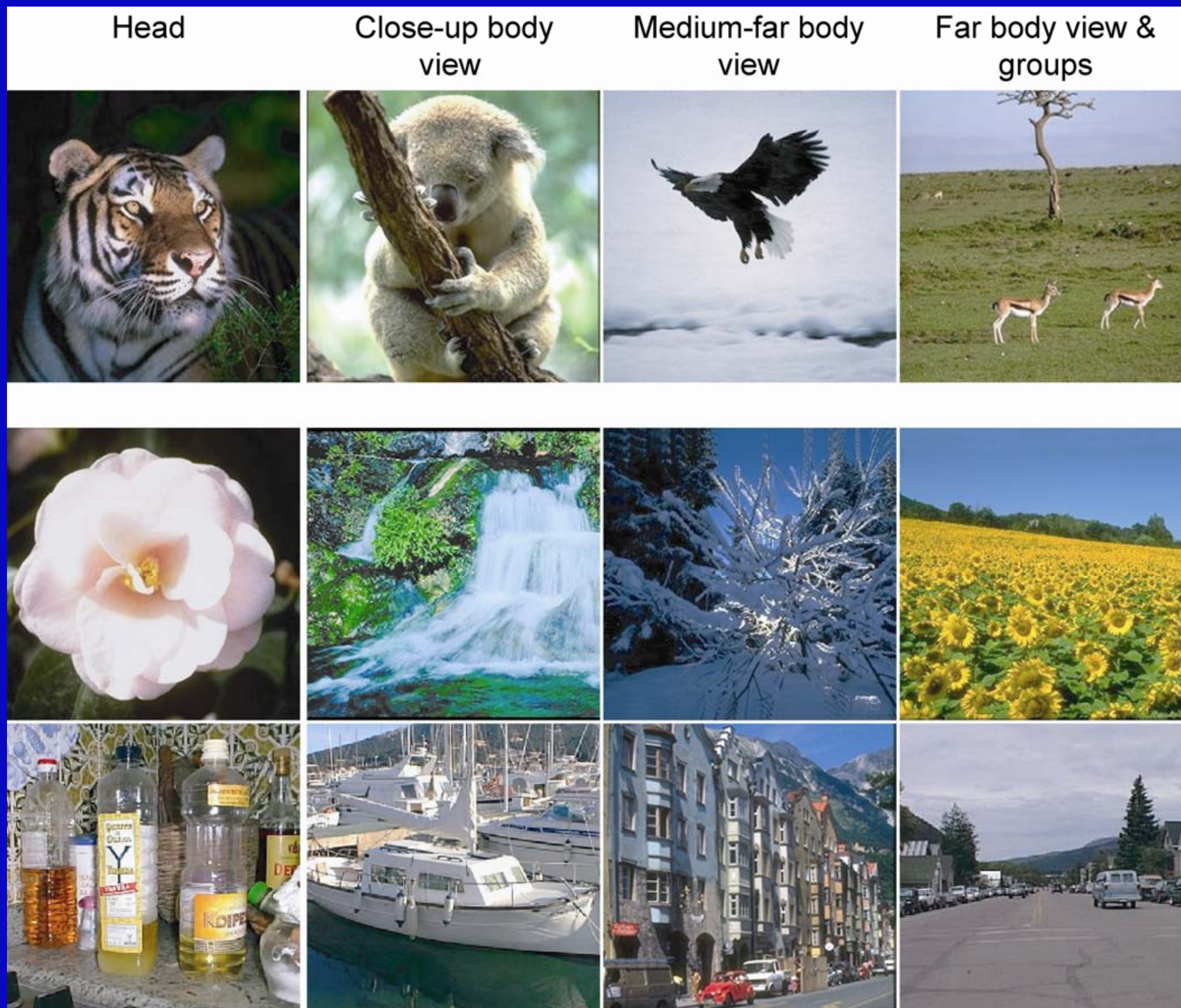
…and another surprise…

… was the comparison with human performance
(Thomas Serre with Aude Oliva)
on rapid categorization of complex natural images

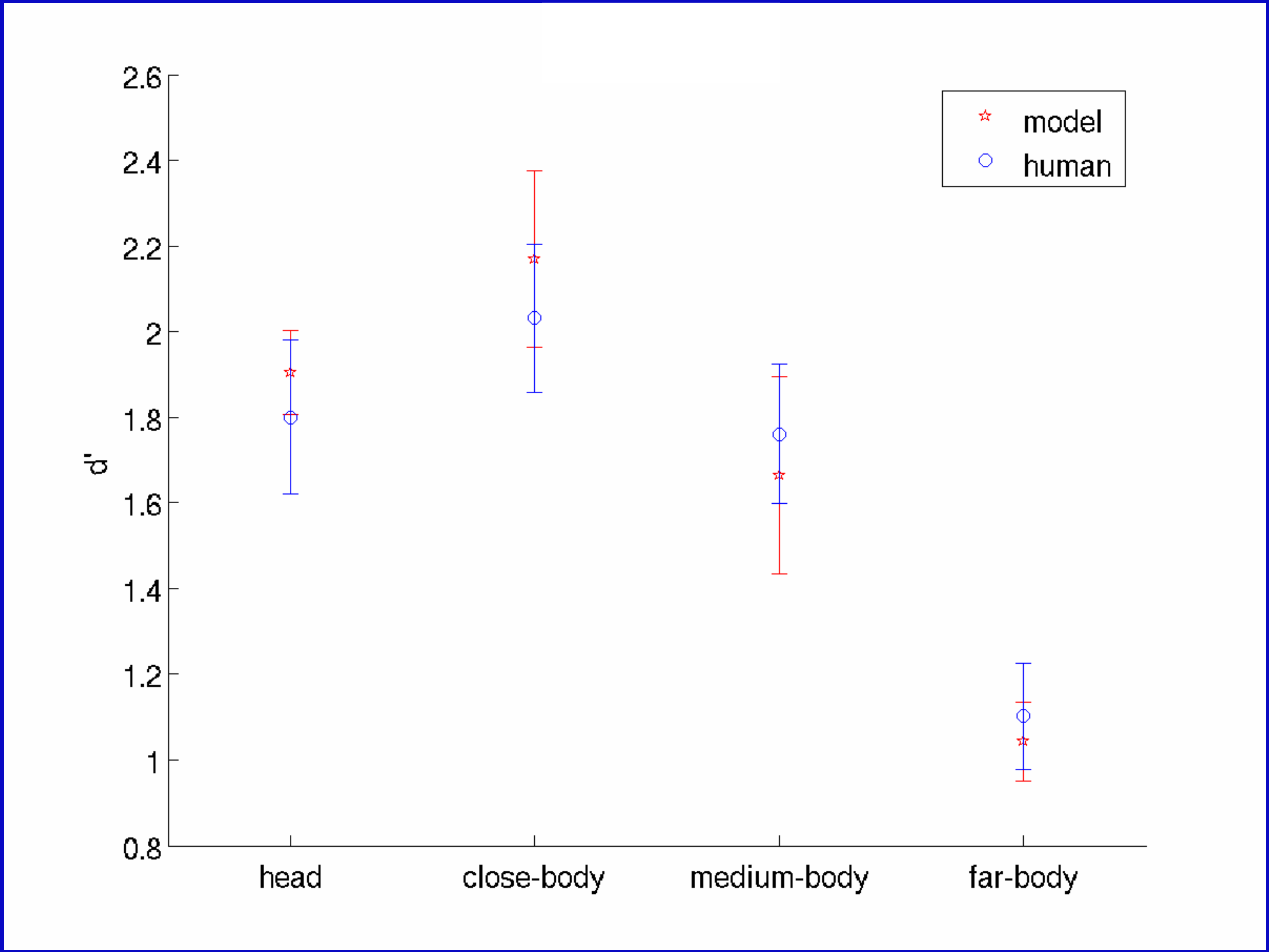
Comparison with Humans



Comparison with Humans



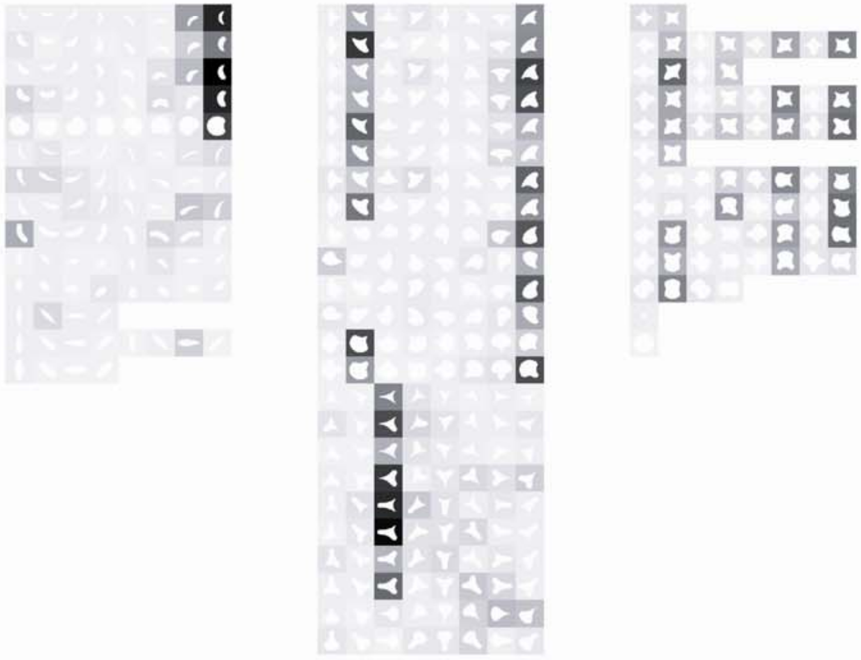
Model vs. Human Subjects



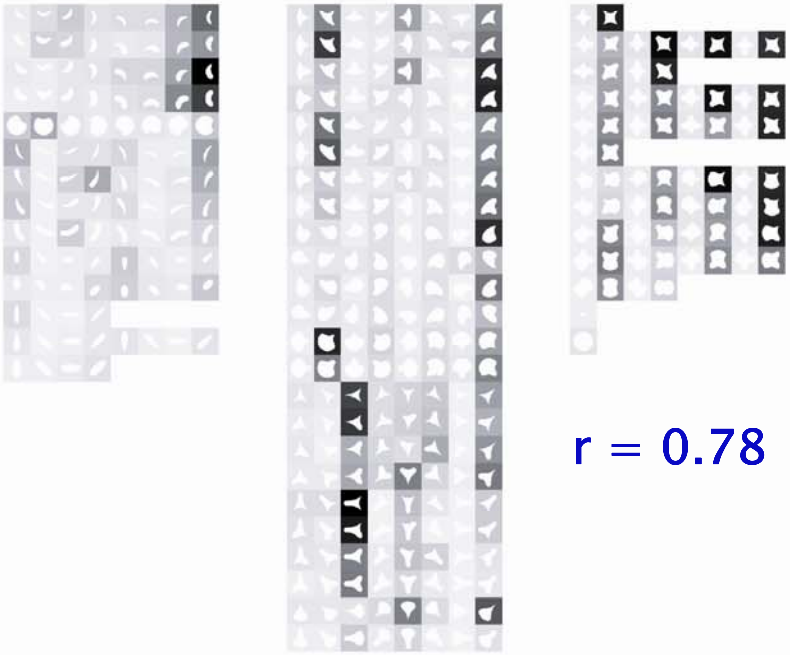
Furthermore…model S2 units are congruent with V4 neural data

Learned Model Units are Congruent with V4 data

Response to a V4 neuron to a parameterized space of shapes



Best model unit from a pool of 109 units learned from natural images



r = 0.78

[Pasupathy & Connors, 2001]
[Cadieu et al., 2005]

Summary II

A simple learning rule generates a large dictionary of visual shape-components

With this learning rule, the model competes with the best computer vision systems on all the categorization datasets we have compared it to (so far)

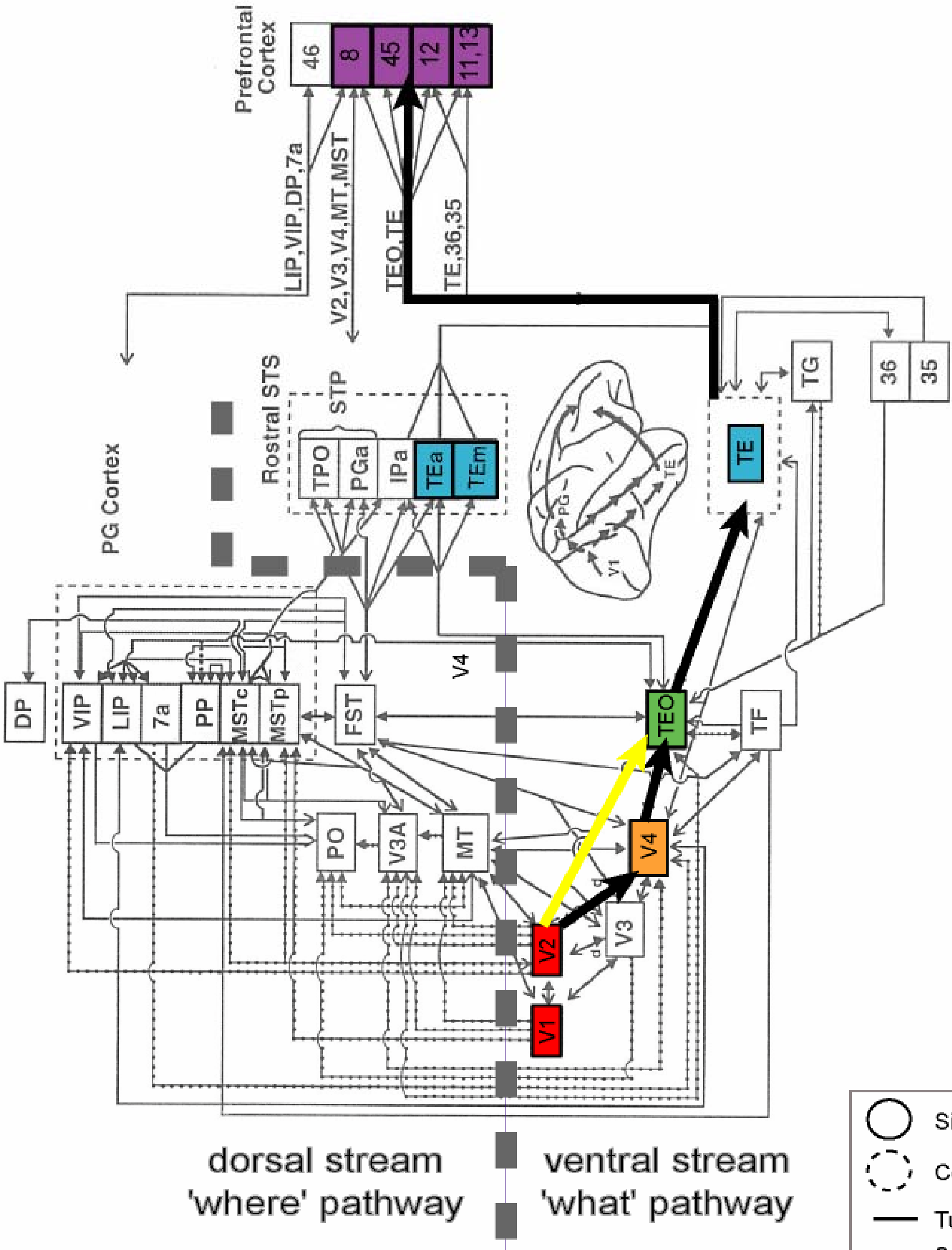
The model performs at the same level of performance as humans on an ultra-rapid animal / non-animal categorization task

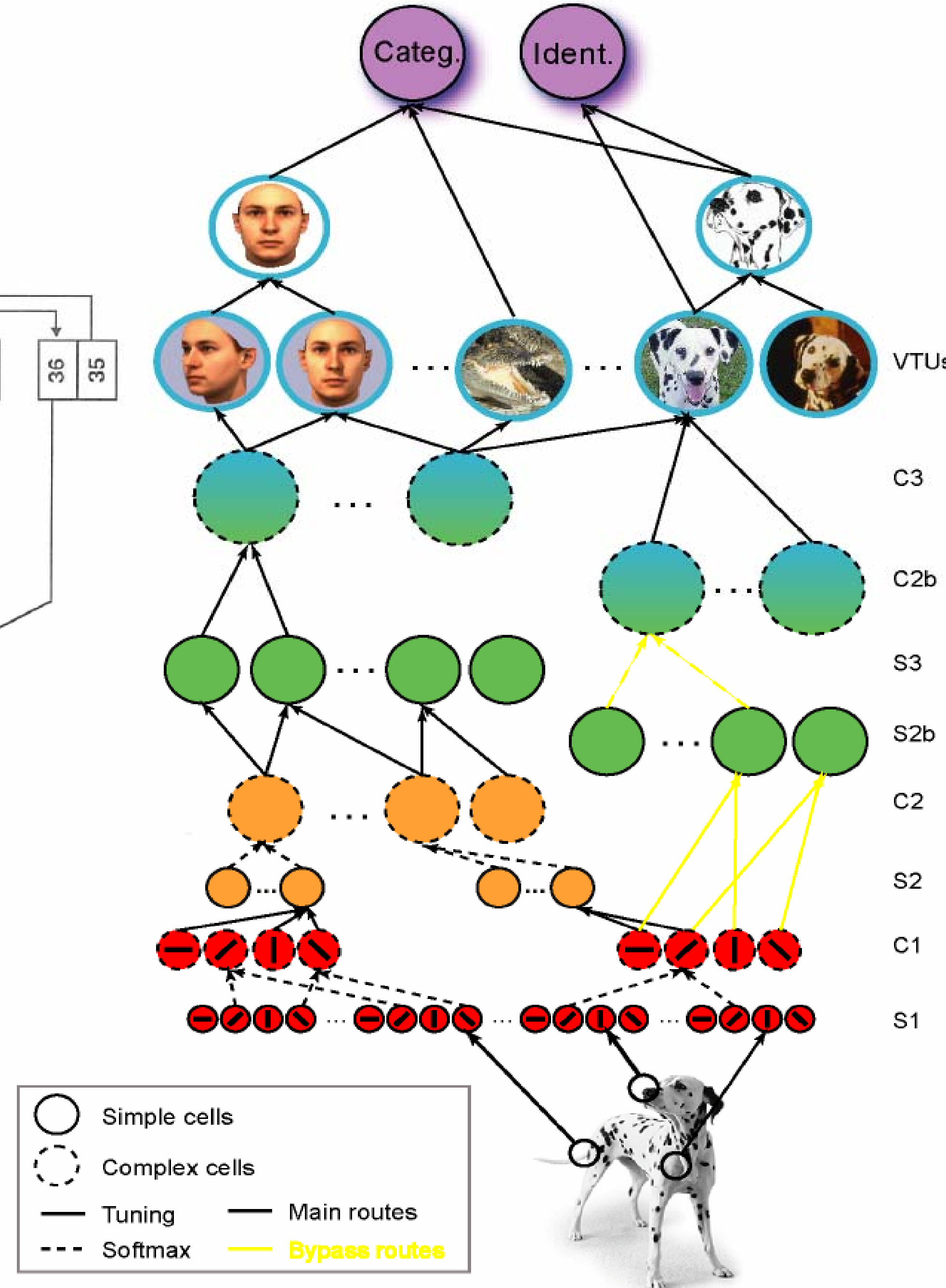
The S2 units learned from natural images are consistent with the tuning properties of V4 neurons

Remarks

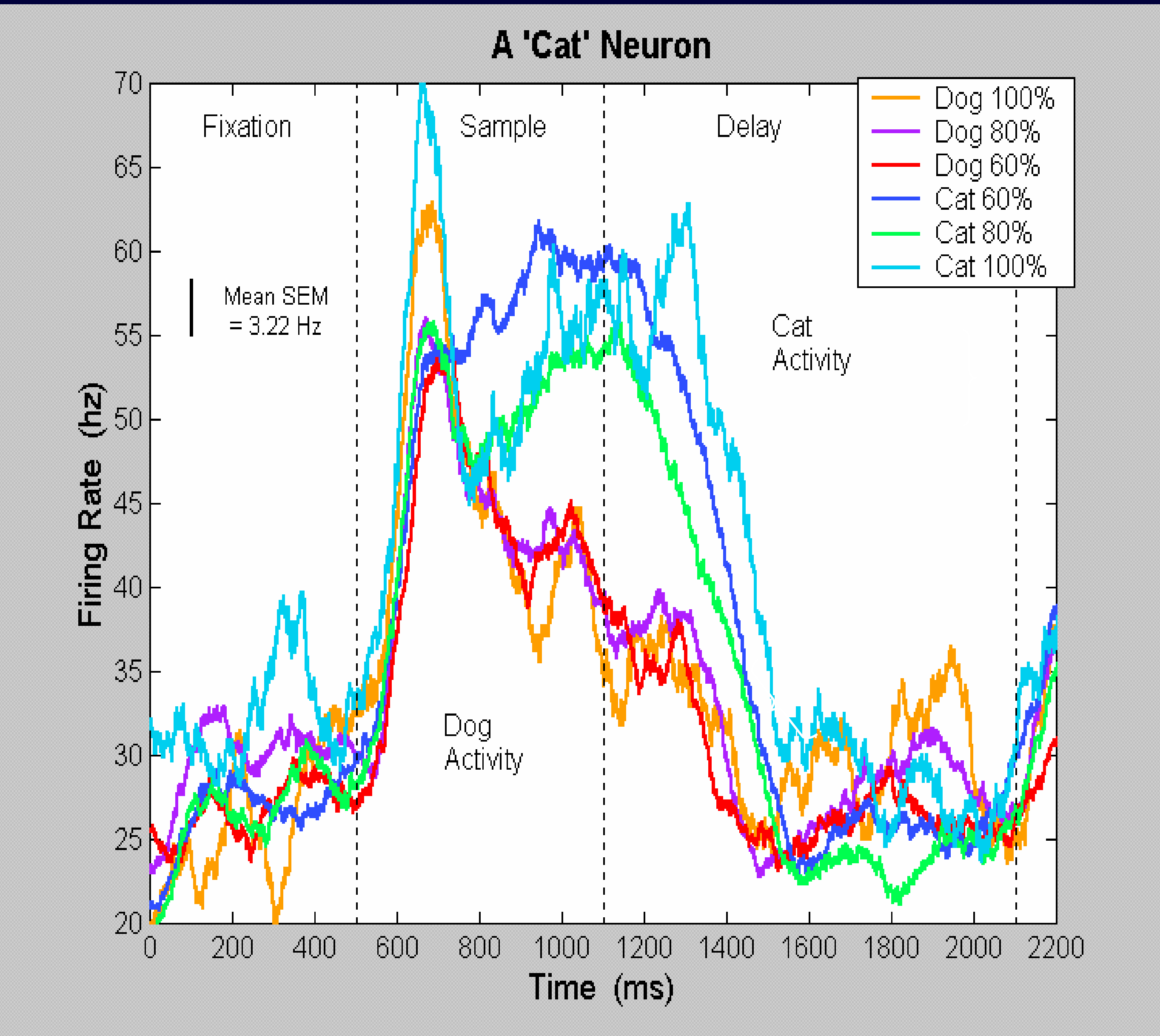
- The stage that includes [V4-PIT]→AIT→PFC represents a learning network of the Gaussian RBF type that is known (from learning theory) to generalize well

- In the theory the stage between IT and "PFC" is a linear classifier – like the one used in the read-out experiments


Prefrontal Cortex
46
8
45
12
11,13
LIP,VIP,DP,7a
V2,V3,V4,MT,MST
TEO,TE
TE,36,35
Rostral STS
STP
TPO
PGa
IPa
TEa
TEm
PG Cortex
TE
TG
36
35
DP
VIP
LIP
7a
PP
MSTc
MSTp
FST
TEO
TF
V4
PO
V3A
MT
V4
V3
V2
V1
dorsal stream
'where' pathway
ventral stream
'what' pathway


Categ.
Ident.
VTUs
C3
C2b
S3
S2b
C2
S2
C1
S1
Simple cells
Complex cells
Tuning
Softmax
Main routes
Bypass routes

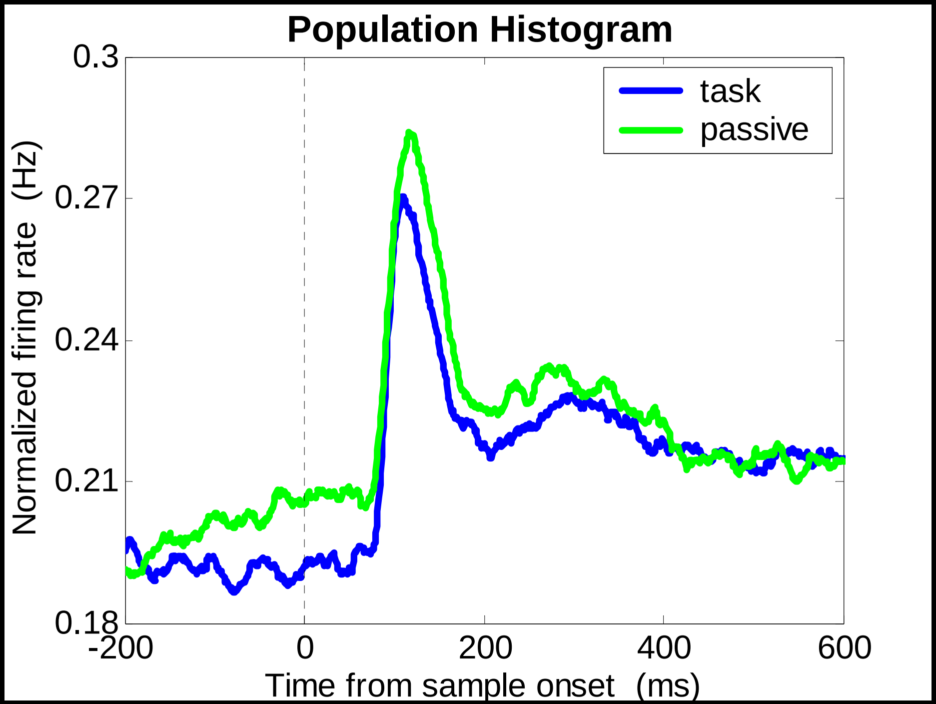
Model performance compares well with recordings from monkey Prefrontal Cortex



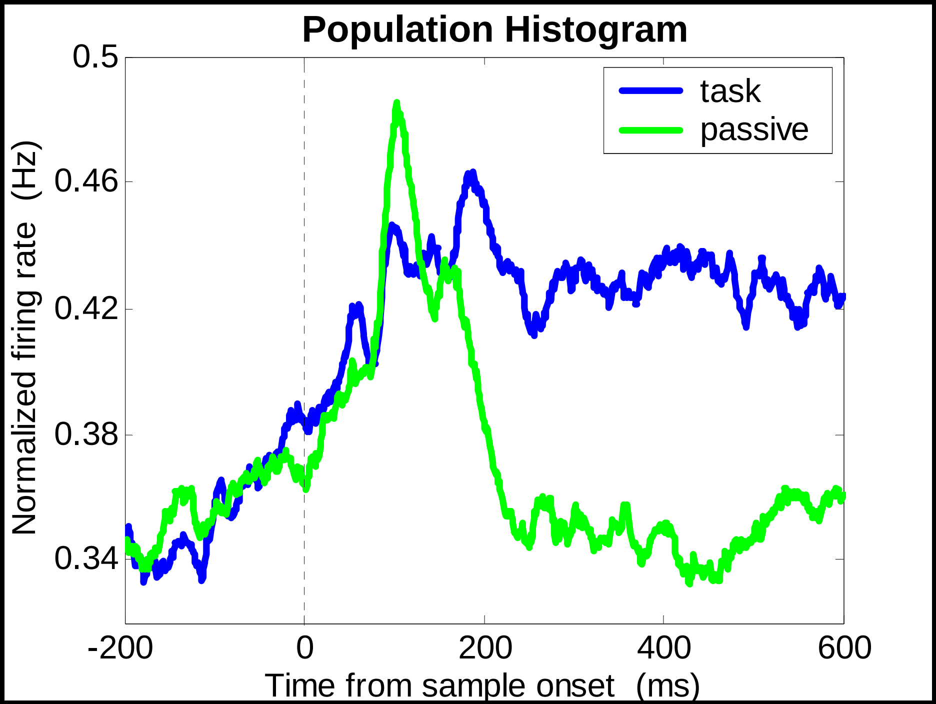
D. Freedman + E. Miller + M. Riesenhuber+T. Poggio (Science, 2001)

Comparison of firing rates to cats/dogs during task and passive viewing.

ITC:



PFC:



ITC activity similar between task and passive viewing.

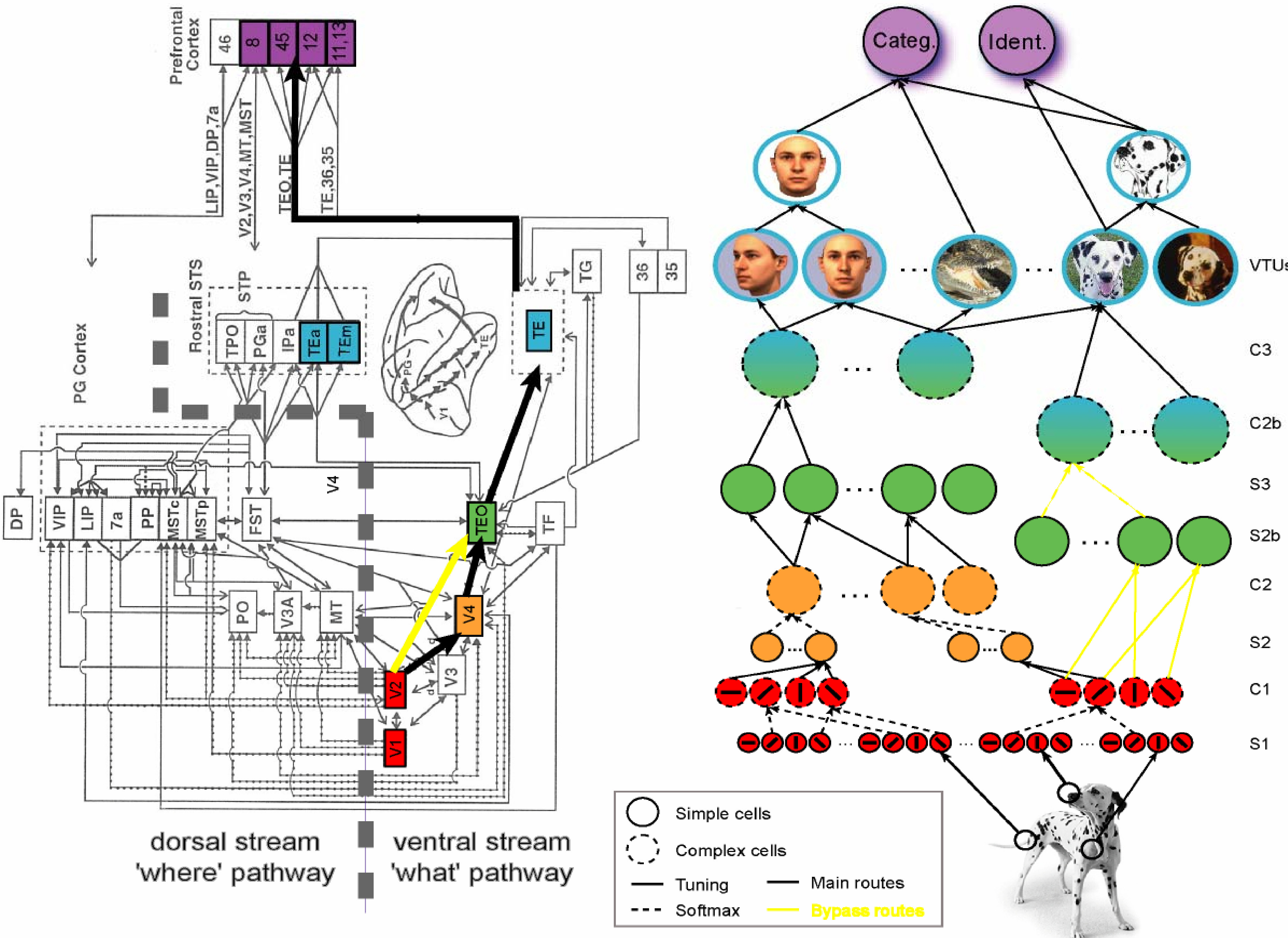
PFC responses were more task-dependent.

How was category selectivity modulated by task demands?

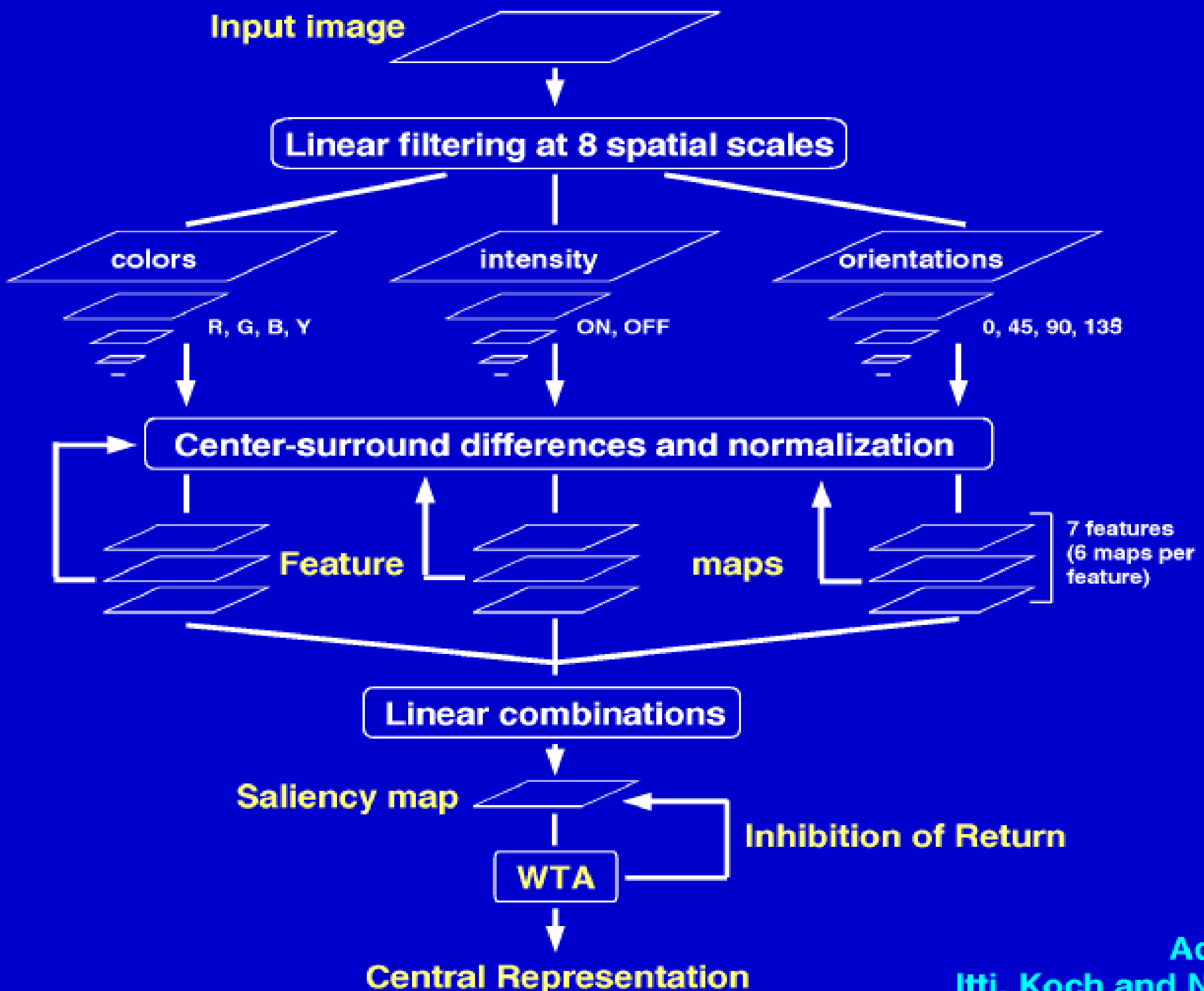
Remarks

- The stage that includes (V4-PIT)-AIT-PFC represents a learning network of the Gaussian RBF type that is known (from learning theory) to generalize well
- In the theory the stage between IT and "PFC" is a linear classifier – like the one used in the read-out experiments
- The inputs to IT are a large dictionary of selective and <u>*invariant*</u> features

FUTURE: extension of the model to include…



…top-down and attention and CalTech (Walther+Koch)



Adapted from
Itti, Koch and Niebur (1998)

…but what if…

it may just be that if the mind were simple enough for us

to understand it

then we may be too simple

to understand it