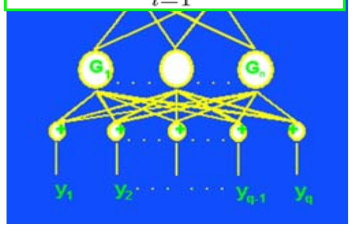
Visual Recognition in primate cortex: from Neuroscience to a new AI?

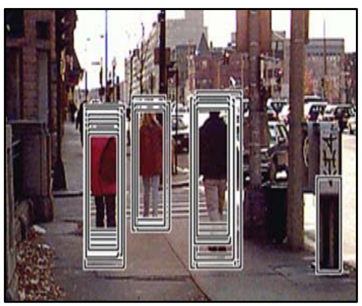
Tomaso Poggio
McGovern Institute for Brain Research
Center for Biological and Computational Learning
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

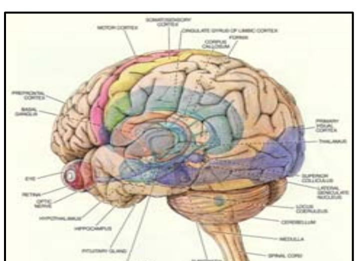
Learning: math, engineering, neuroscience *(until recently)*

$$\min_{f \in H} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(x_i)) + \mu \|f\|_K^2 \right]$$

$$f(x) = \sum_{i=1}^{l} c_i K(\mathbf{x}_i, \mathbf{x})$$







Learning theory + algorithms

Theorems on foundations of learning

Predictive algorithms

ENGINEERING APPLICATIONS

- Bioinformatics
- Computer vision
- Computer graphics, speech synthesis, creating a virtual actor

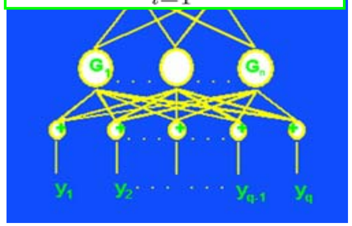
Computational Neuroscience: models+experiments

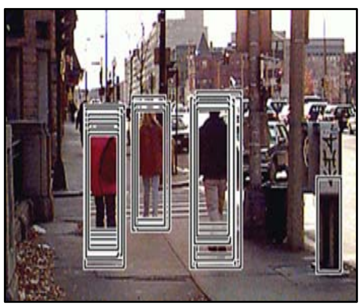
How visual cortex works

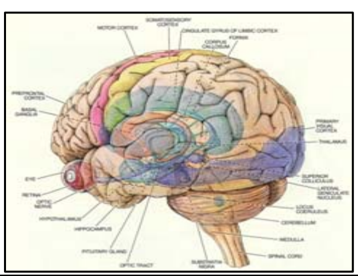
Learning:
math, engineering, neuroscience *(now)*

$$\min_{f \in H} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(x_i)) + \mu \| f \|_K^2 \right]$$

$$f(x) = \sum_{i=1}^{l} c_i K(\mathbf{x}_i, \mathbf{x})$$







Learning theory + algorithms

Theorems on foundations of learning

Predictive algorithms

ENGINEERING APPLICATIONS

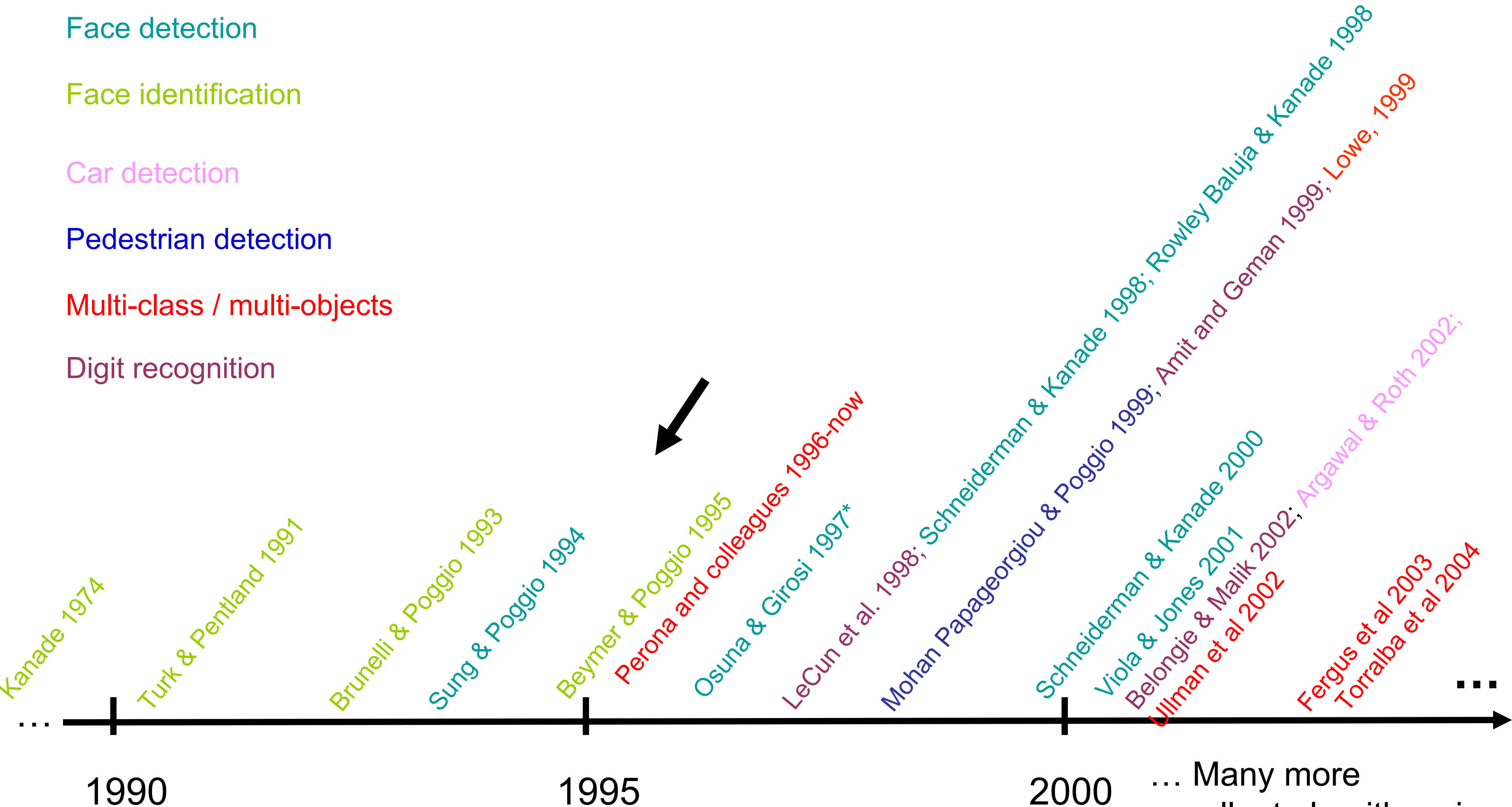
- Bioinformatics
- Computer vision
- Computer graphics, speech synthesis, creating a virtual actor

Computational Neuroscience: models+experiments

How visual cortex works

1. “Old” computer vision and learning work
2. Recent work in neuroscience of recognition can account for cell properties, human performance and provide good computer vision algorithms
3. Future: recognition in videos, a new learning theory inspired by cortex and extending approach to image inference tasks

Object recognition for computer vision: (personal) historical perspective

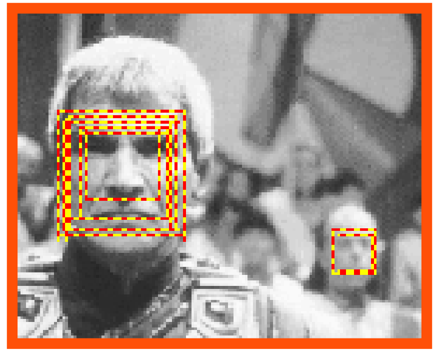


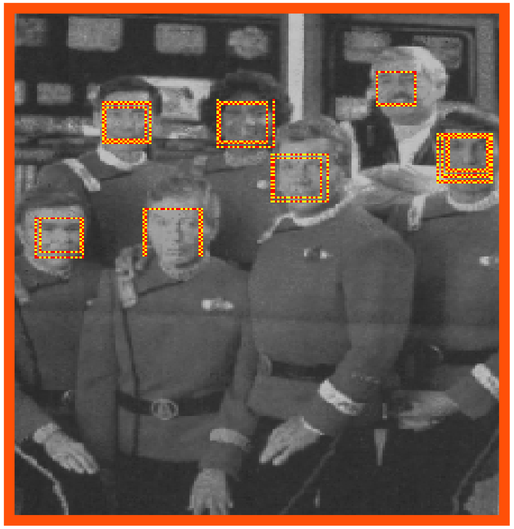
… Many more excellent algorithms in the past few years…

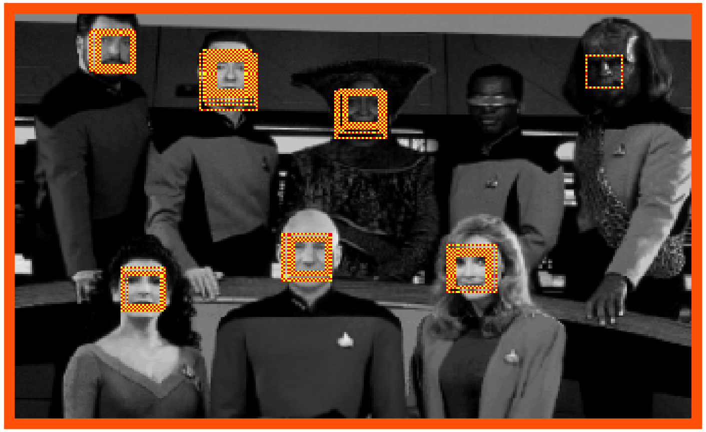
*Best CVPR'07 paper 10 yrs ago

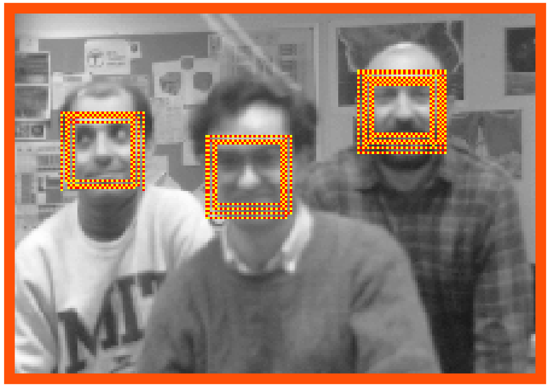
Examples: Learning Object Detection: Finding Frontal Faces

- Training Database
- 1000+ Real, 3000+ VIRTUAL
- 50,0000+ Non-Face Pattern









Sung & Poggio 1995

~10 year old CBCL computer vision work: SVM-based pedestrian detection system in Mercedes test car...
now becoming a product (MobilEye, Israeli company)

Parallel development of (classical) learning theory and learning algorithms from perceptrons to learning theory to Vapnik and to Smale (and many others…)

In the last few years the theoretical foundations of learning have become part of mainstream mathematics (many papers/results on the mathematical foundations and on algorithms)

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 39, Number 1, Pages 1–49
S 0273-0979(01)00923-5
Article electronically published on October 5, 2001

ON THE MATHEMATICAL FOUNDATIONS OF LEARNING

FELIPE CUCKER AND STEVE SMALE

The problem of learning is arguably at the very core of the problem of intelligence, both biological and artificial.

T. Poggio and C.R. Shelton

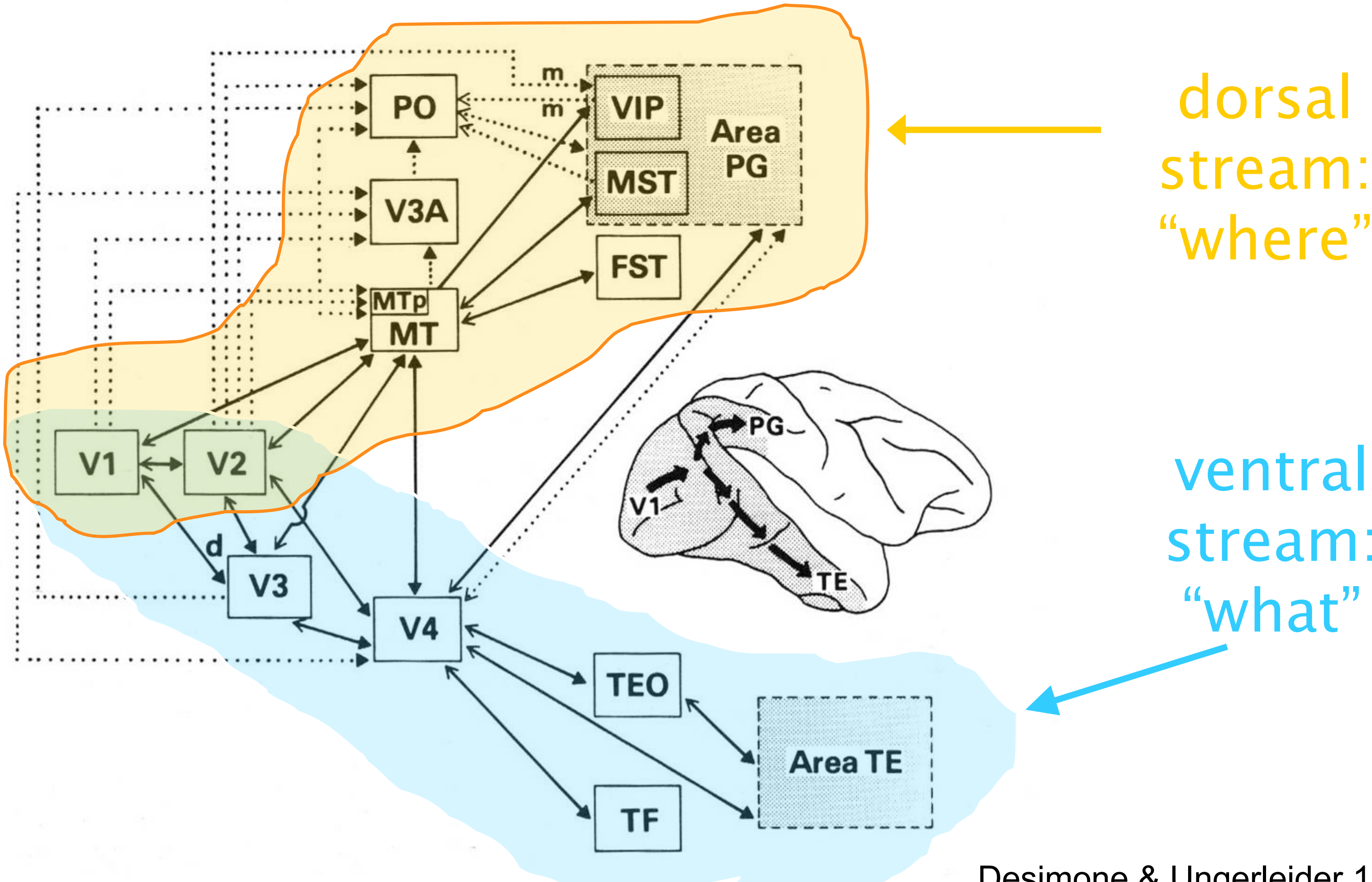
Introduction

(1) A main theme of this report is the relationship of approximation to learning and the primary role of sampling (inductive inference). We try to emphasize relations of the theory of learning to the mainstream of mathematics. In particular, there are large roles for probability theory, for algorithms such as *least squares*, and for tools and ideas from linear algebra and linear analysis. An advantage of doing this is that communication is facilitated and the power of core mathematics is more easily brought to bear.

1. “Old” computer vision and learning work
2. Now: recent work in neuroscience of recognition can account for cell properties, human performance and provide good computer vision –and perhaps learning -- algorithms
3. Future: recognition in videos, a new learning theory inspired by cortex and extending approach to image inference tasks

The problem:
recognition in natural images
(e.g., “is there an animal in the image?”)

The hypothesis is that visual cortex has a key role in solving this problem: how?



Desimone & Ungerleider 1989

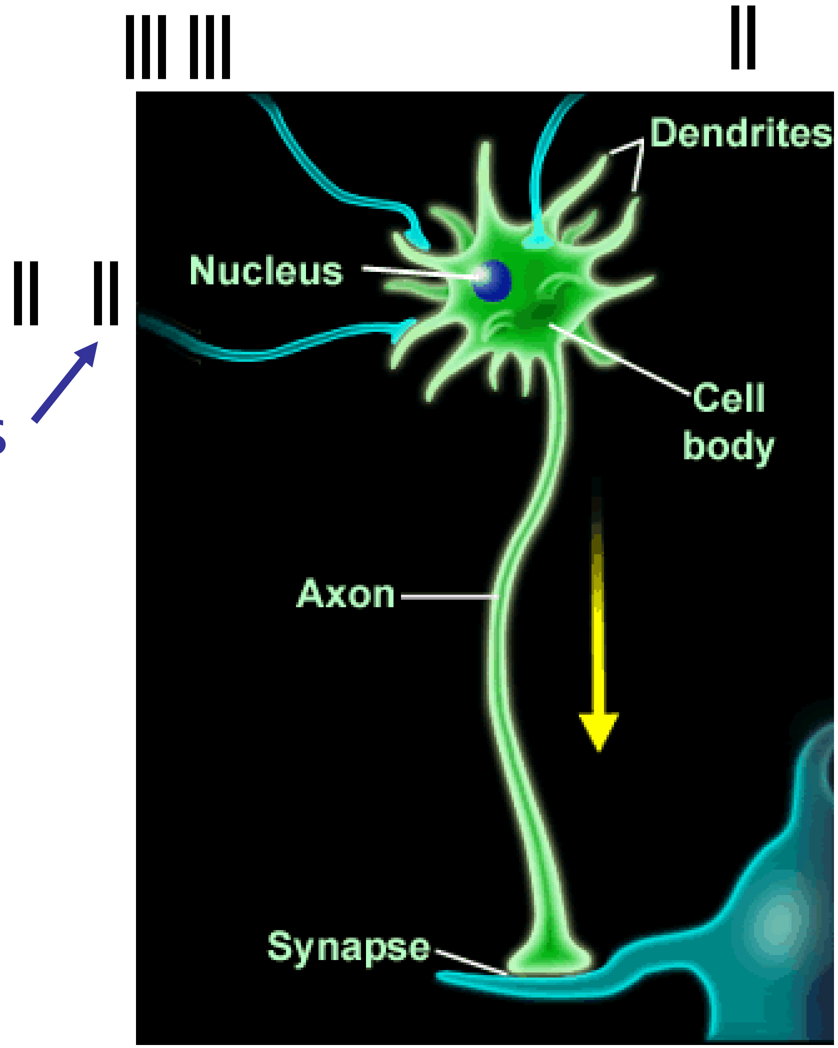
Neuron basics

INPUT= pulses or graded potentials

COMPUTATION = Analog

OUTPUT = Chemical

spikes

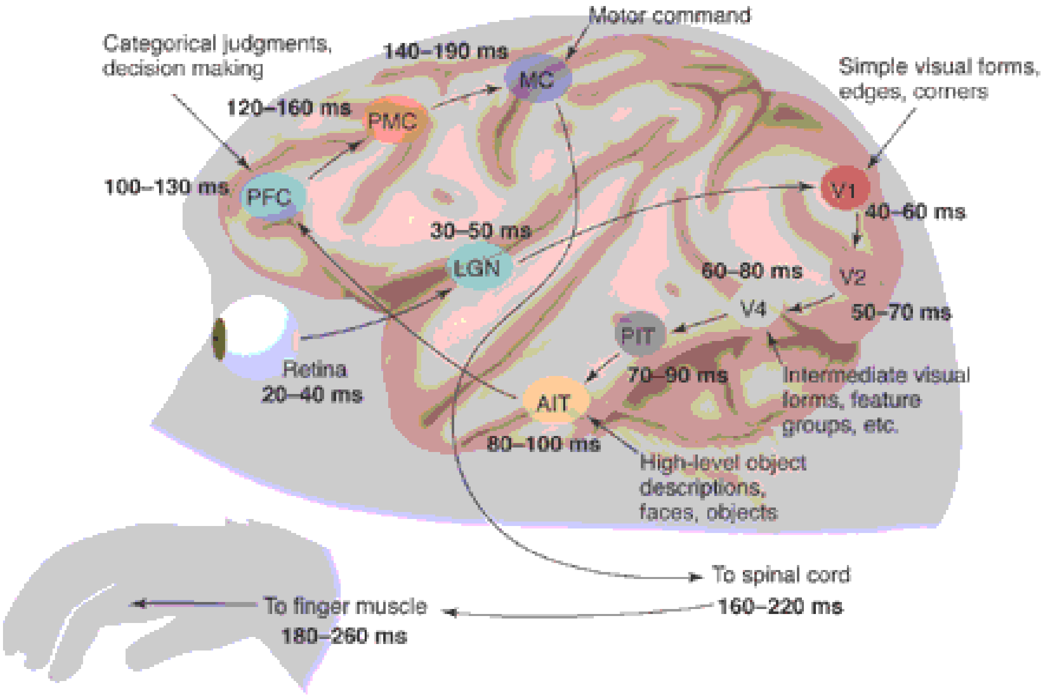


Some numbers

- Human Brain
  - 10^{11}-10^{12} neurons (1 million flies ☺)
  - 10^{14}- 10^{15} synapses

- Neuron
  - Fundamental space dimensions:
    - fine dendrites : 0.1 μ diameter; lipid bilayer membrane : 5 nm thick; specific proteins : pumps, channels, receptors, enzymes
  - Fundamental time length : 1 msec

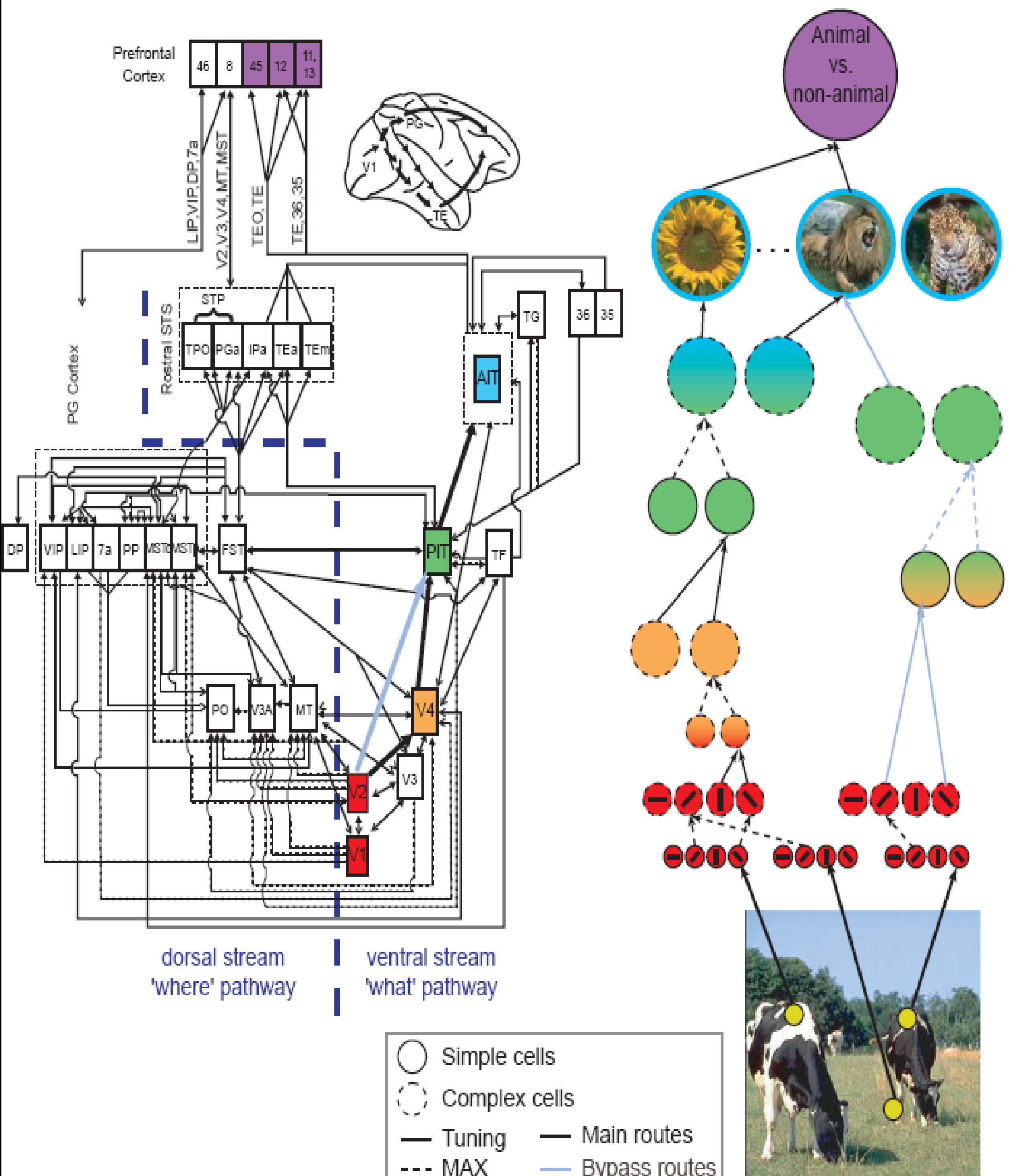
It turns out the brain may teach us something about computer vision and learning: a model of the ventral stream of visual cortex



Theory of Object Recognition: Computations and Circuits in the Feedforward path of the Ventral Stream in Primate Visual Cortex

Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich and Tomaso Poggio, December 2005

Models of Visual Recognition with learning (unsupervised and supervised)



Model layers	Corresponding brain area (tentative)	RF sizes	Number units	
classifier	PFC		$1.0\ 10^{0}$	
S4	AIT	>4.4°	$1.5\ 10^{2}$	~ 5,000 subunits
C3	PIT - AIT	>4.4°	$2.5\ 10^{3}$	
C2b	PIT	>4.4°	$2.5\ 10^{3}$	
S3	PIT	1.2° - 3.2°	$7.4\ 10^{4}$	~ 100 subunits
S2b	V4 - PIT	0.9° - 4.4°	$1.0\ 10^{7}$	~ 100 subunits
C2	V4	1.1° - 3.0°	$2.8\ 10^{5}$	
S2	V2 - V4	0.6° - 2.4°	$1.0\ 10^{7}$	~ 10 subunits
C1	V1 - V2	0.4° - 1.6°	$1.2\ 10^{4}$	
S1	V1 - V2	0.2° - 1.1°	$1.6\ 10^{6}$	

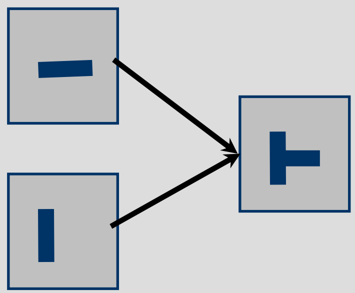
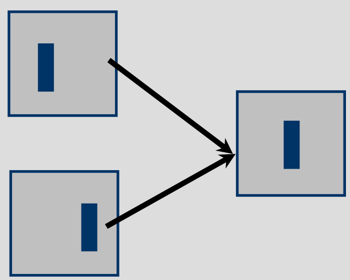
Supervised task-dependent learning

Unsupervised task-independent learning

increase in complexity (number of subunits), RF size and invariance

Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Two key computations, suggested by physiology

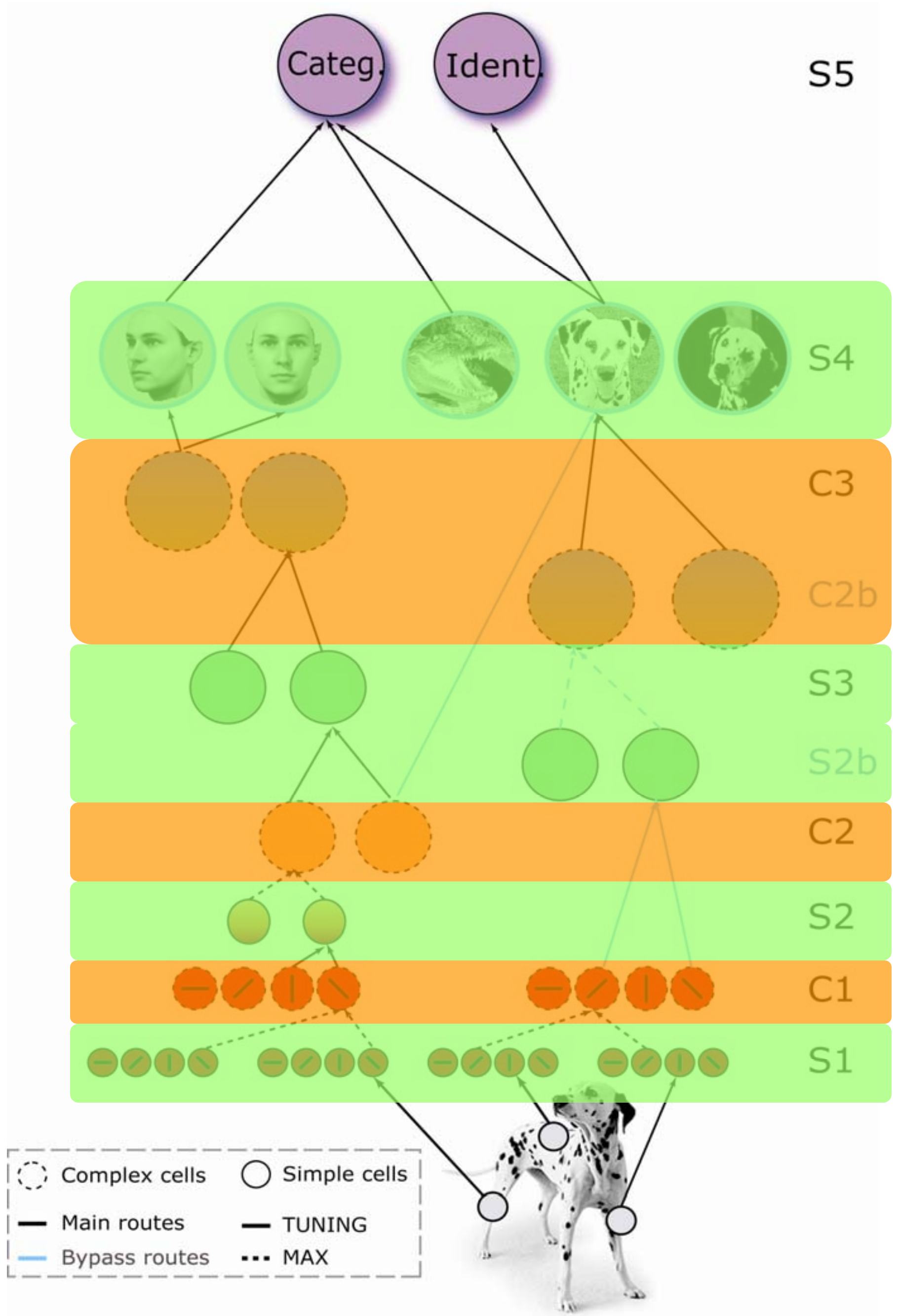
Unit types	Pooling	Computation	Operation
Simple		Selectivity / template matching	Gaussian-tuning / AND-like
Complex		Invariance	Soft-max / OR-like

➢Gaussian-like tuning operation (and-like)

➢Simple units

➢Max-like operation (or-like)

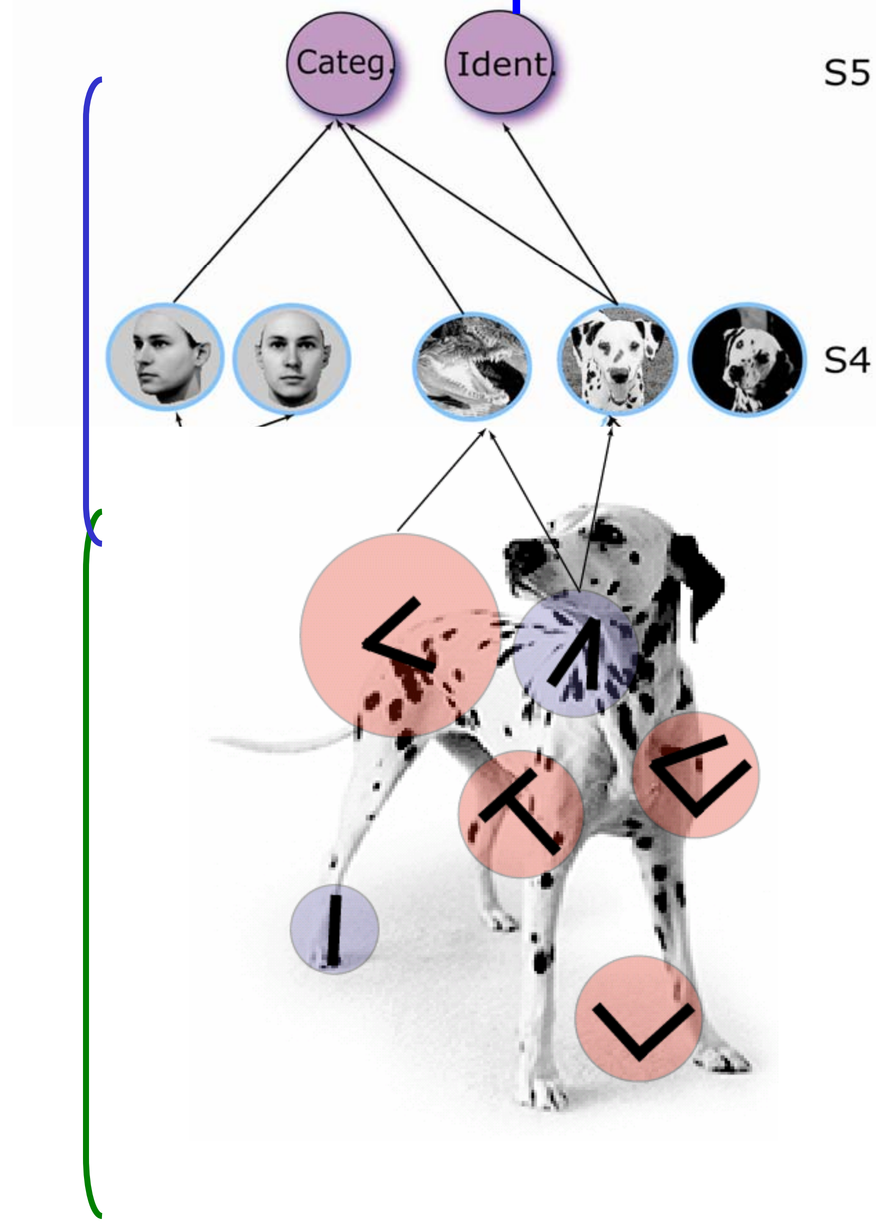
➢Complex units



Learning: supervised and unsupervised

Task-specific circuits (from IT to PFC)

- Supervised learning: ~ Gaussian RBF

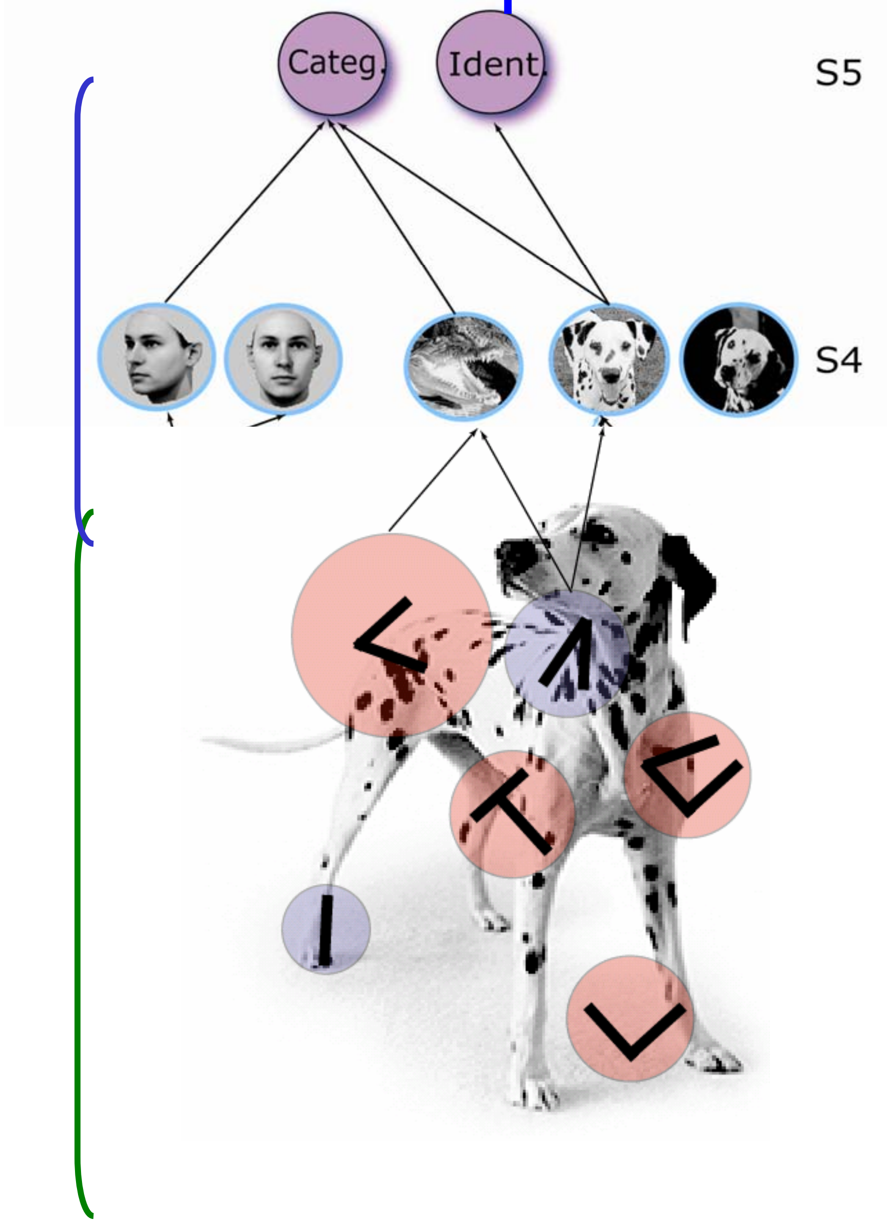


- Generic, overcomplete dictionary of reusable shape components (from V1 to IT) provide unique representation
  – Unsupervised learning (from ~10,000 natural images) during a developmental-like stage

see also (Foldiak 1991; Perrett et al 1984; Wallis & Rolls, 1997; Lewicki and Olshausen, 1999; Einhauser et al 2002; Wiskott & Sejnowski 2002; Spratling 2005)

Learning: supervised and unsupervised

Supervised learning



- Hierarchy – and related unsupervised learning (layer-by-layer – decreases sample complexity for classifier at the top

Can the model explain tuning and invariance properties of *neurons* in the ventral stream?

Feedforward models: comparison w| some neural data

- V1:
  - Simple and complex cells tuning (Schiller et al 1976; Hubel & Wiesel 1965; Devalois et al 1982)
  - MAX-like operation in subset of complex cells (Lampl et al 2004)
- V4:
  - Tuning for two-bar stimuli (Reynolds Chelazzi & Desimone 1999)
  - MAX-like operation (Gawne et al 2002)
  - Two-spot interaction (Freiwald et al 2005)
  - Tuning for boundary conformation (Pasupathy & Connor 2001, Cadieu, Kouh, Connor et al., 2007)
  - Tuning for Cartesian and non-Cartesian gratings (Gallant et al 1996)
- IT:
  - Tuning and invariance properties (Logothetis et al 1995, paperclip objects)
  - Differential role of IT and PFC in categorization (Freedman et al 2001, 2002, 2003)
  - **<u>Read out data</u>** (Hung Kreiman Poggio & DiCarlo 2005)
  - Pseudo-average effect in IT (Zoccolan Cox & DiCarlo 2005; Zoccolan Kouh Poggio & DiCarlo 2007)
- Human:
  - Rapid categorization (Serre Oliva Poggio 2007)
  - Face processing (fMRI + psychophysics) (Riesenhuber et al 2004; Jiang et al 2006)

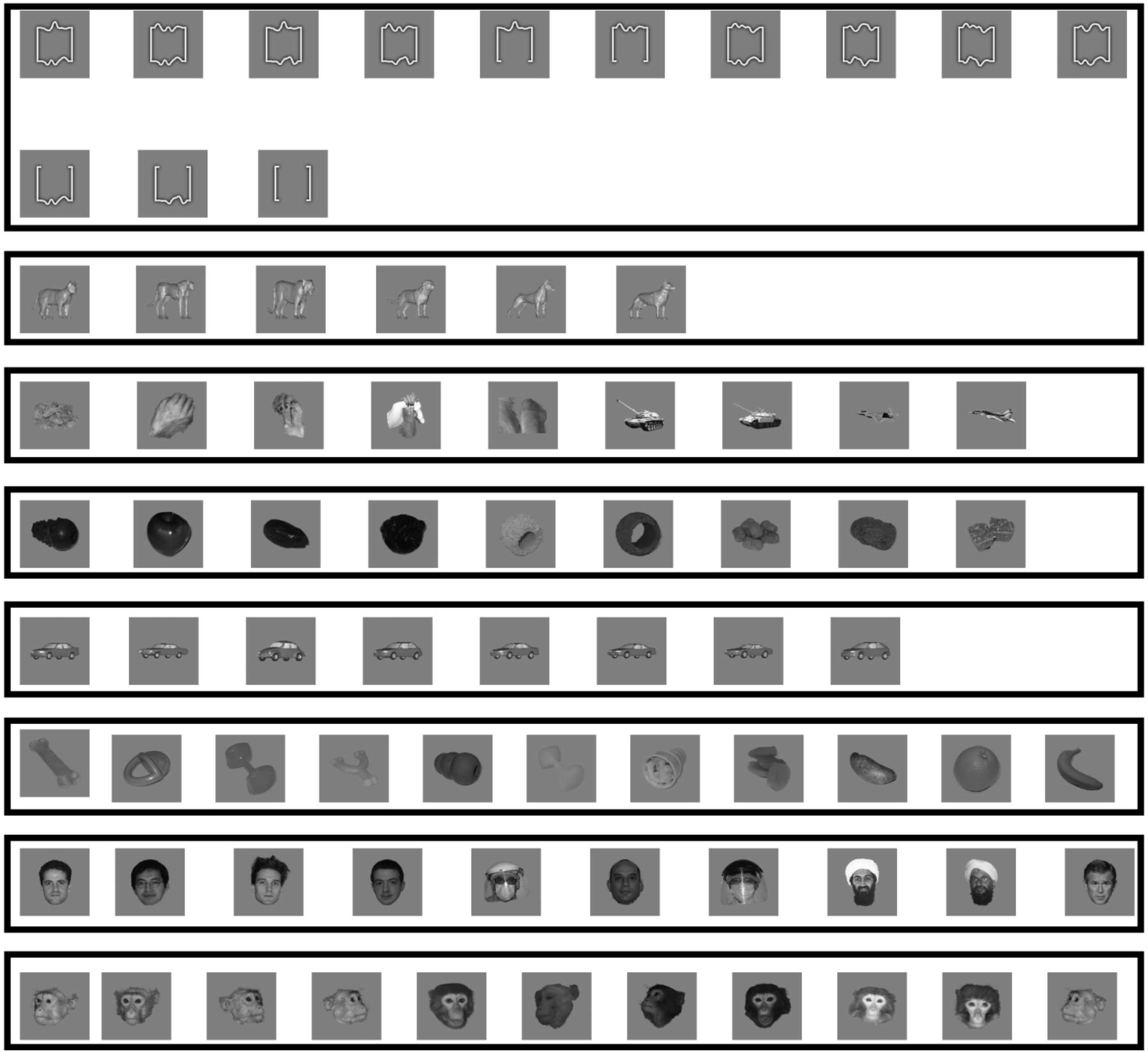
(Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005)

- Just one example…:

Read out data (Hung Kreiman Poggio & DiCarlo 2005)

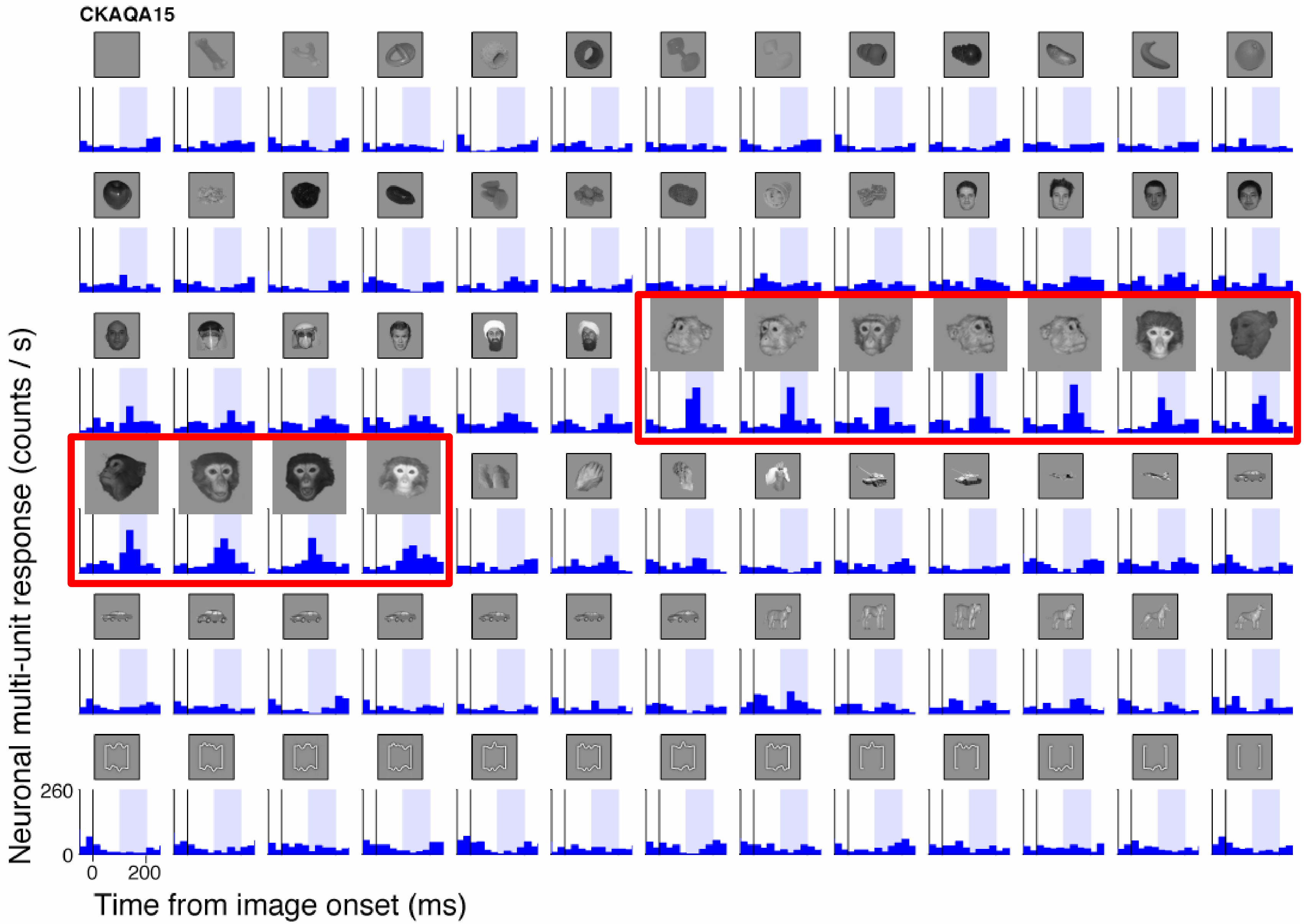
IT Readout data

77 objects,
8 classes

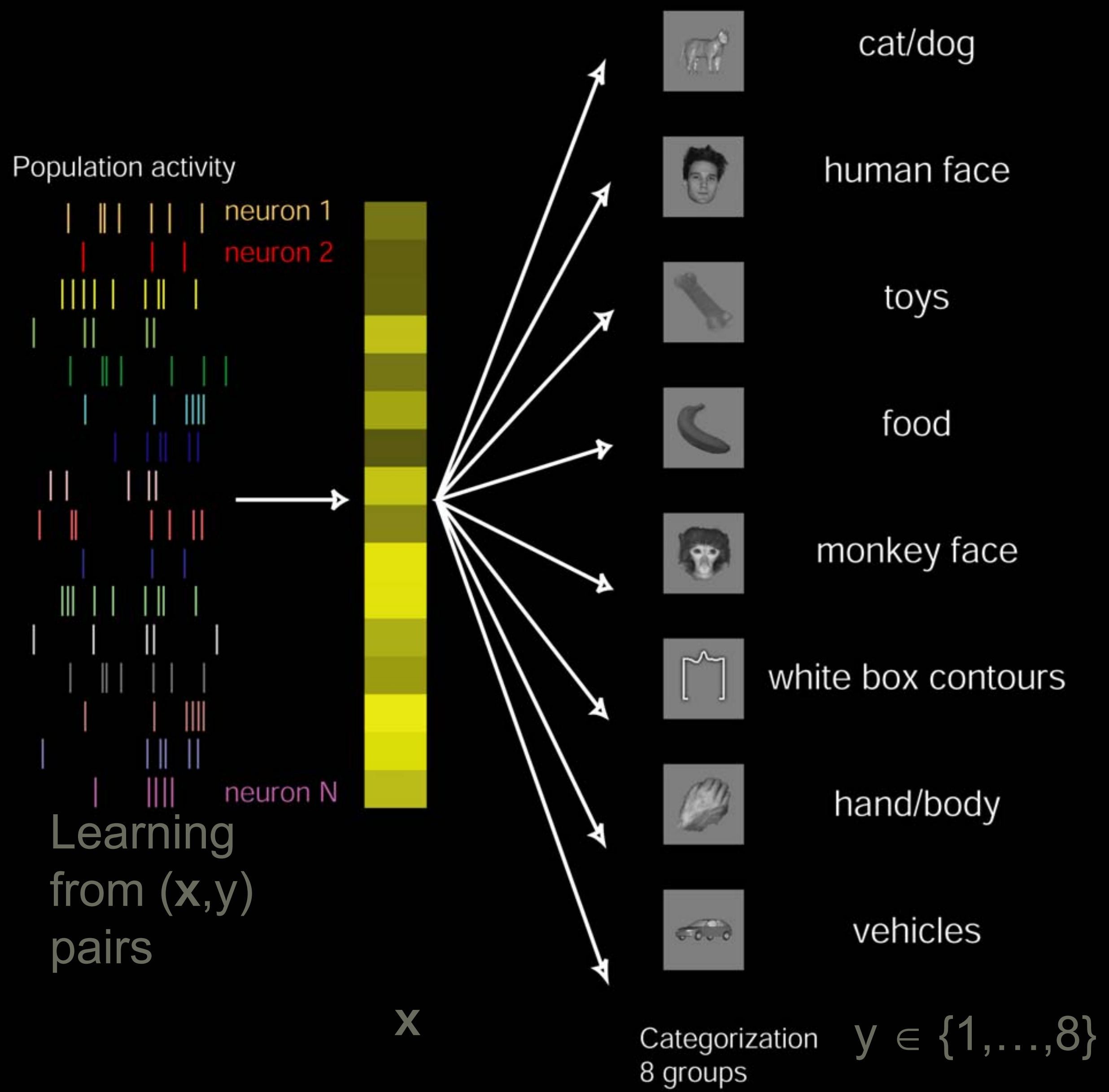


Chou Hung, Gabriel Kreiman, James DiCarlo, Tomaso Poggio, *Science, Nov 4, 2005*

Example of one AIT cell



Decoding the neural code …
population response (using a classifier)



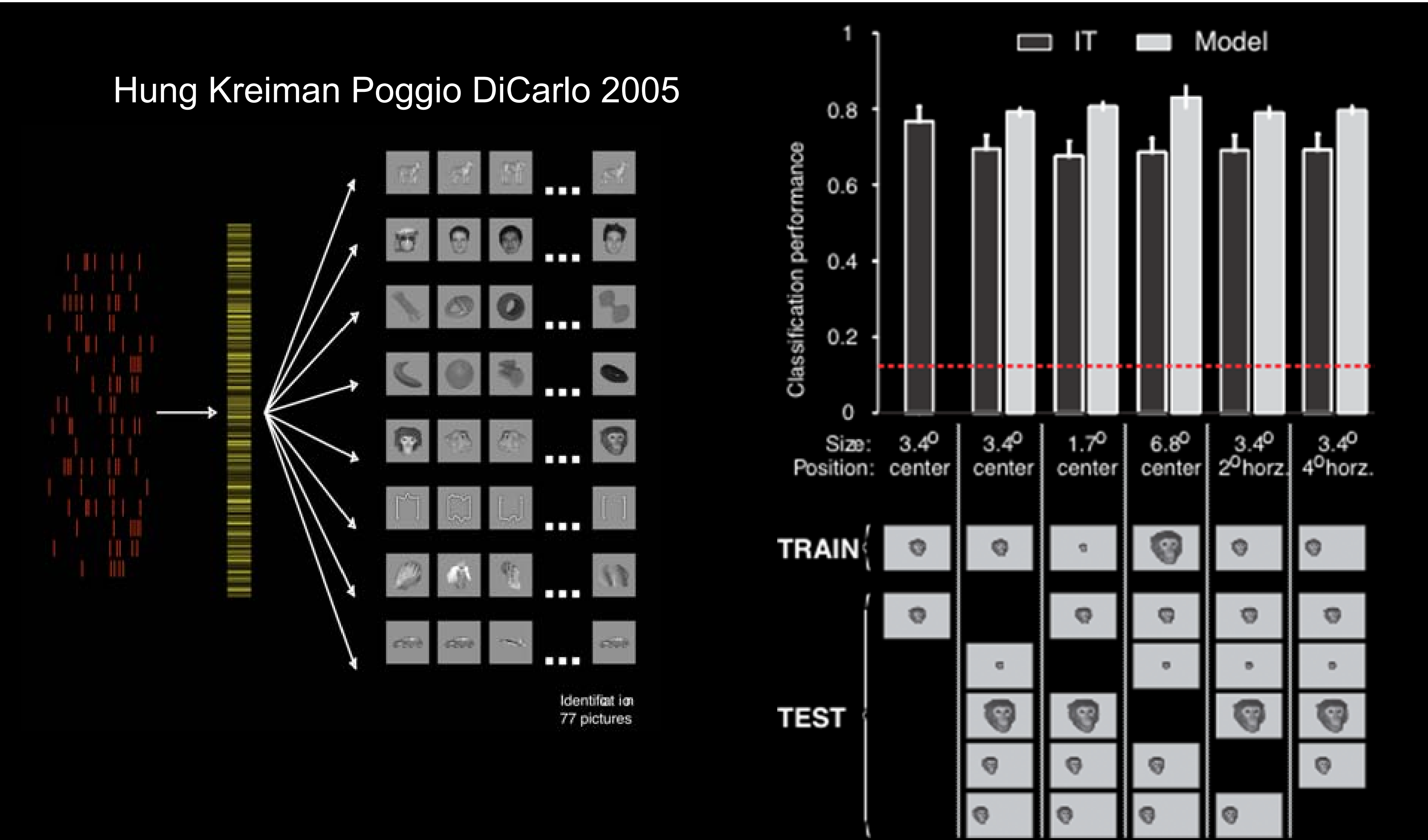
So…we can decode the brain’s code and read-out from neural activity what the monley is seeing

We can also read-out with similar results from the model !!!

We can decode from model units as well as from IT

Agreement of model w| IT Readout data

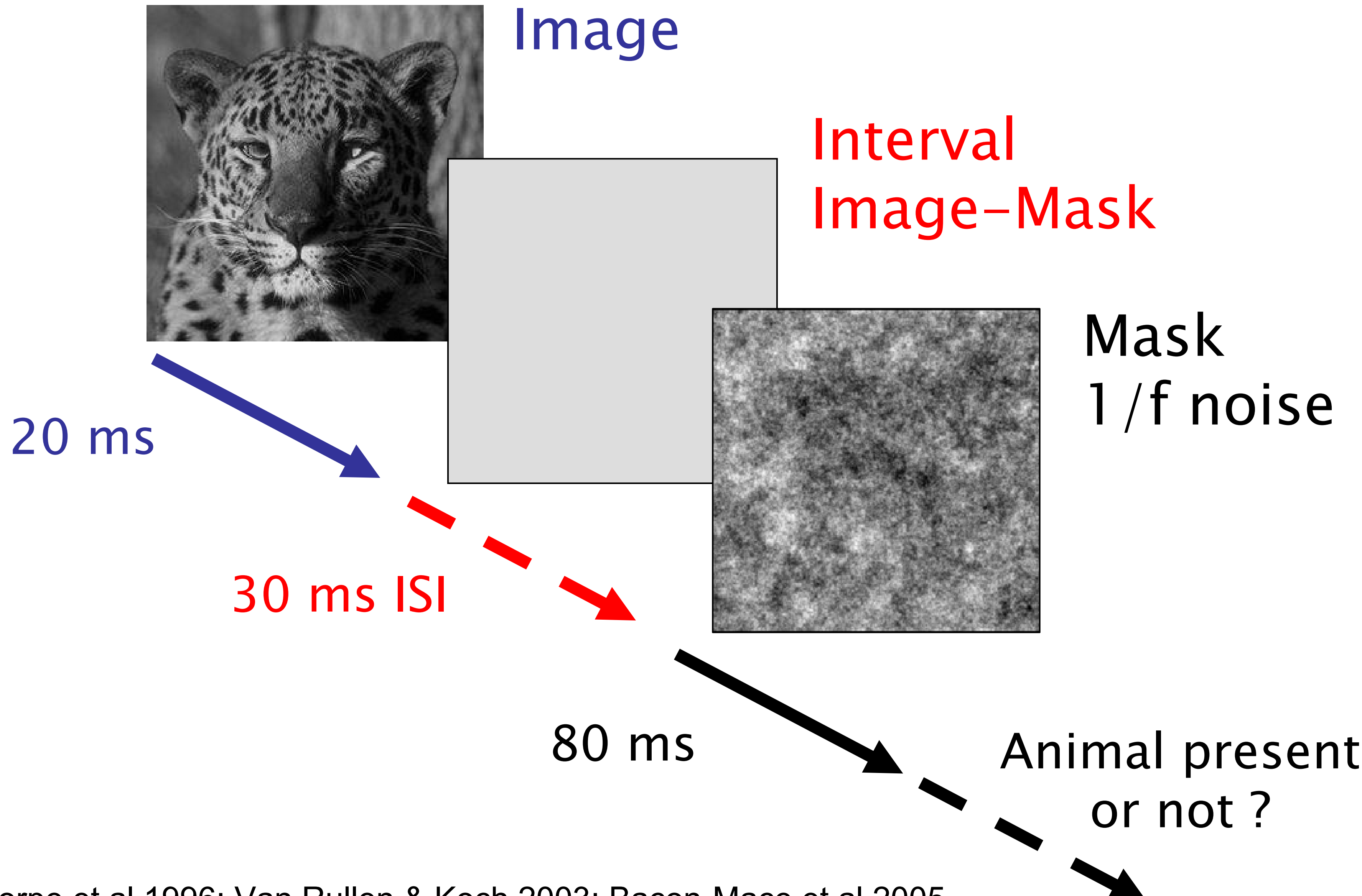
Reading out category and identity invariant to position and scale



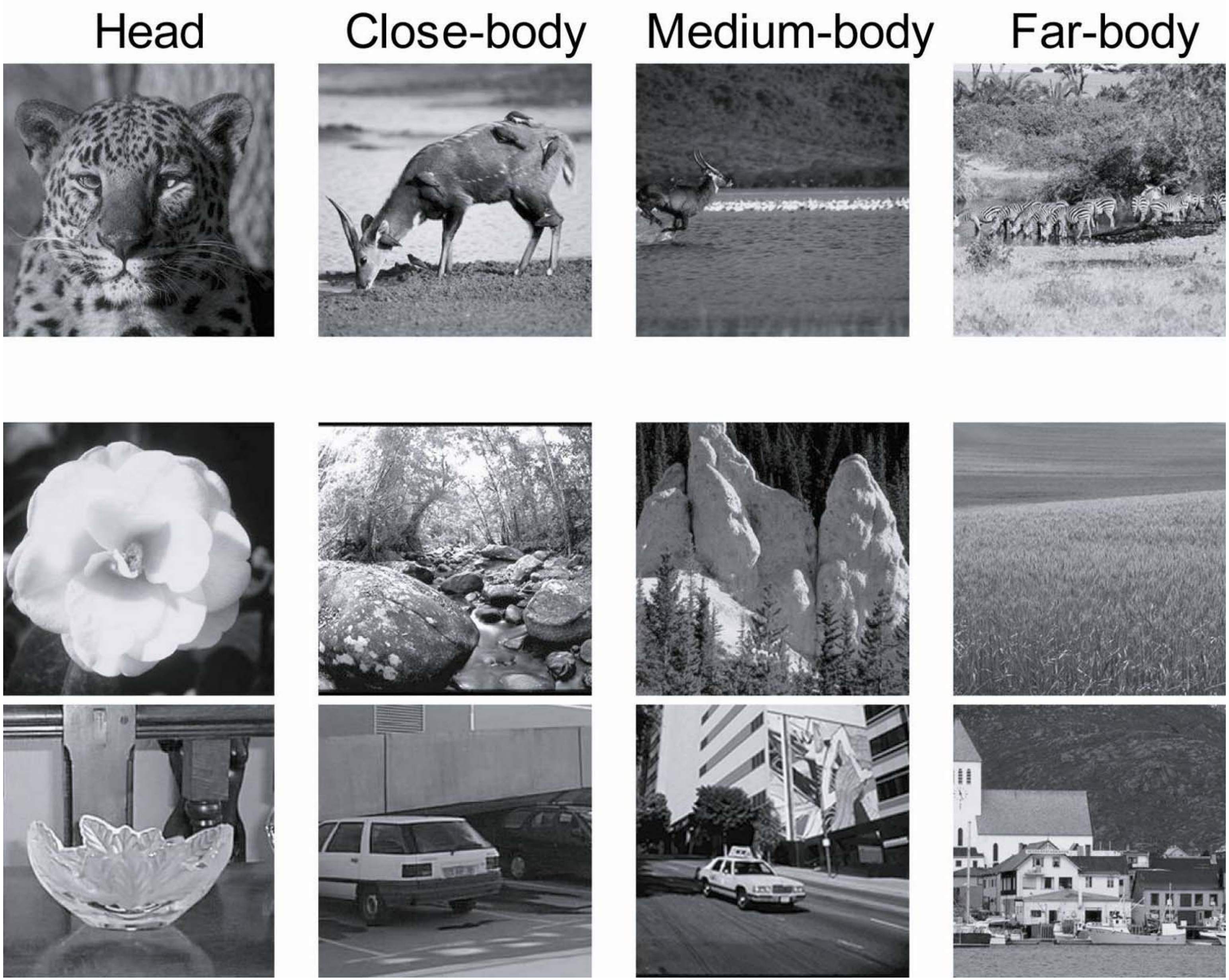
Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005

Can the (feedforward) model then account for rapid categorization by human subjects?

Rapid categorization task (with mask to test feedforward model)



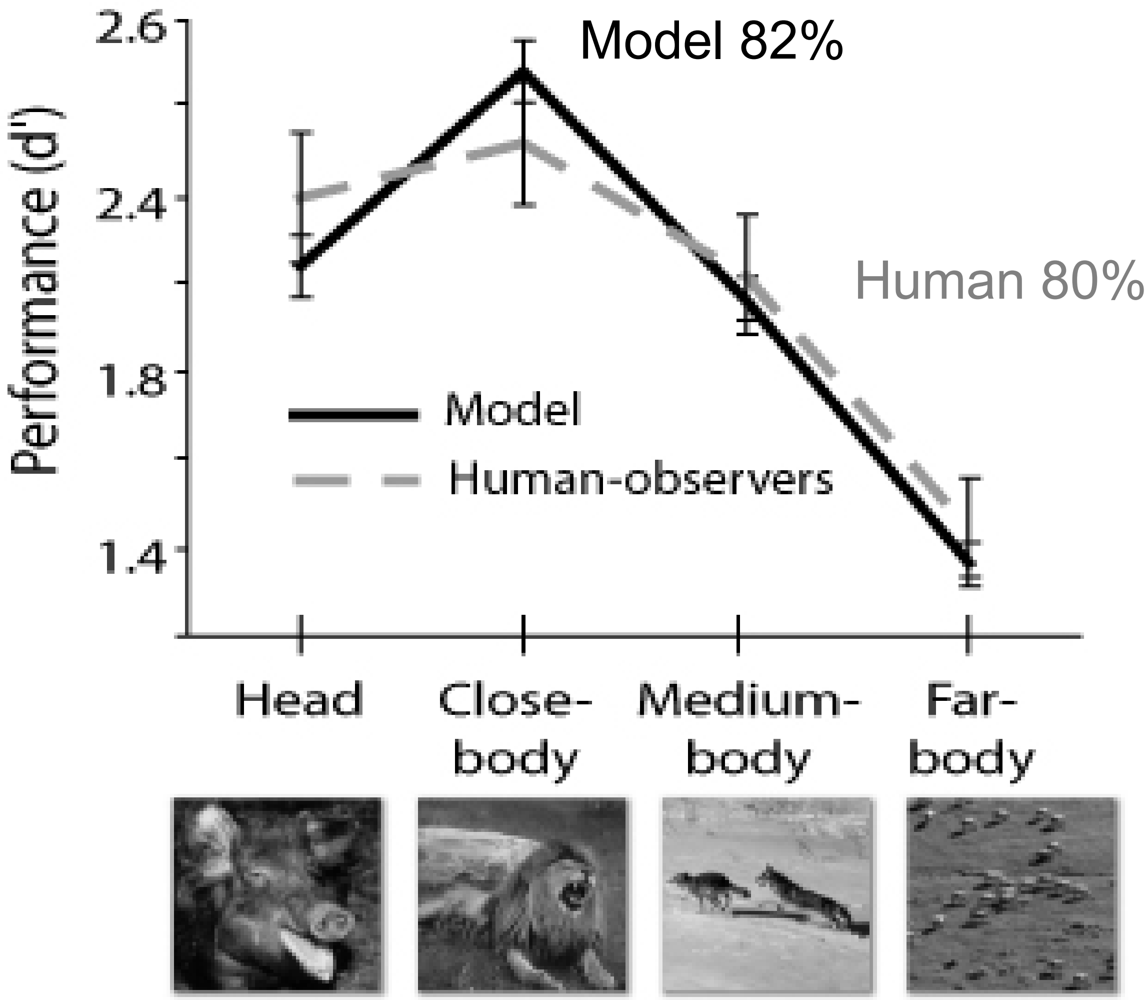
Thorpe et al 1996; Van Rullen & Koch 2003; Bacon-Mace et al 2005



(Torralba & Oliva, 2003)

Model "predicts" human "feedforward" performance

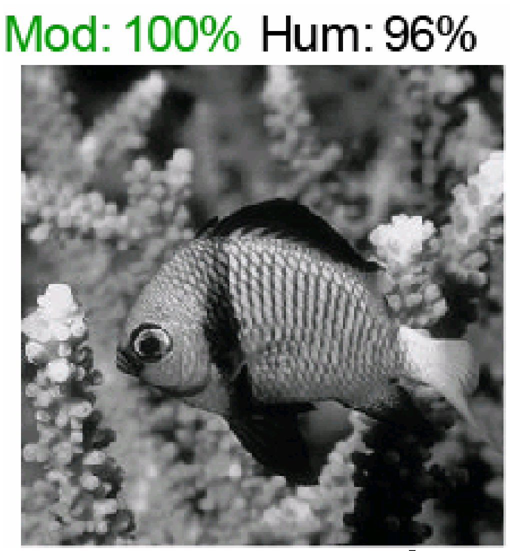
- d'~ standardized error rate
- the higher the d', the better the perf.



Serre Oliva & Poggio 2007

Further comparisons

- Image-by-image correlation:
  - Heads: ρ=0.71
  - Close-body: ρ=0.84
  - Medium-body: ρ=0.71
  - Far-body: ρ=0.60

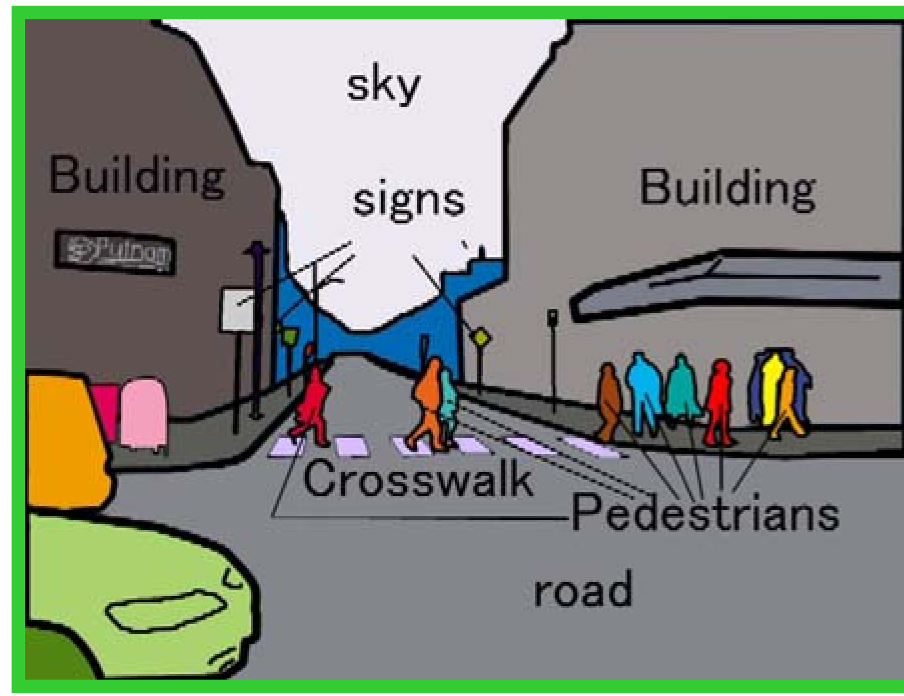


- Model predicts level of performance on rotated images (90 deg and inversion)

Serre Oliva & Poggio PNAS 2007

…a surprise for me was that the neuroscience model worked well compared with several good machine vision systems (in 2005) on a variety of databases (Caltech 101, faces, Weizman) including our own Scene Street database…

The street scene database

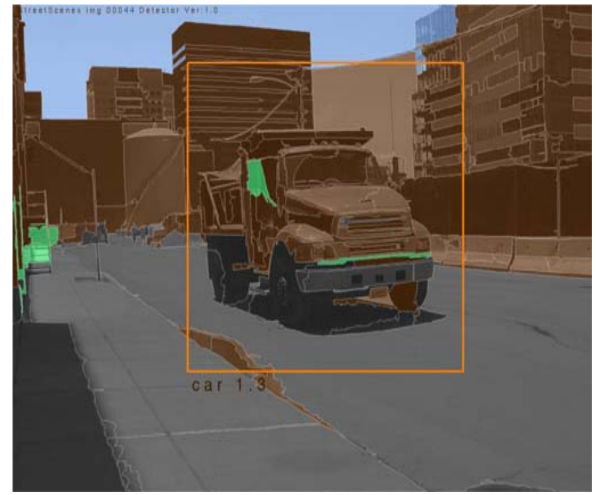


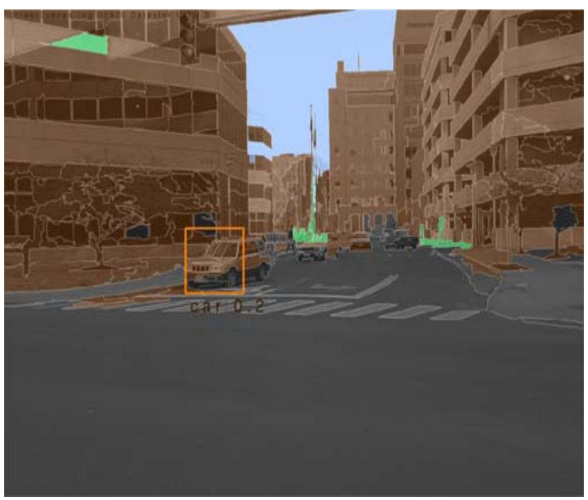
Source: Bileschi, Wolf & Poggio

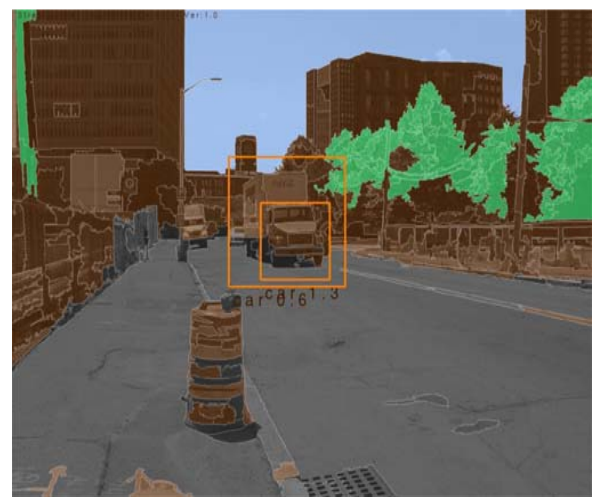
StreetScenes Database. Subjective Results

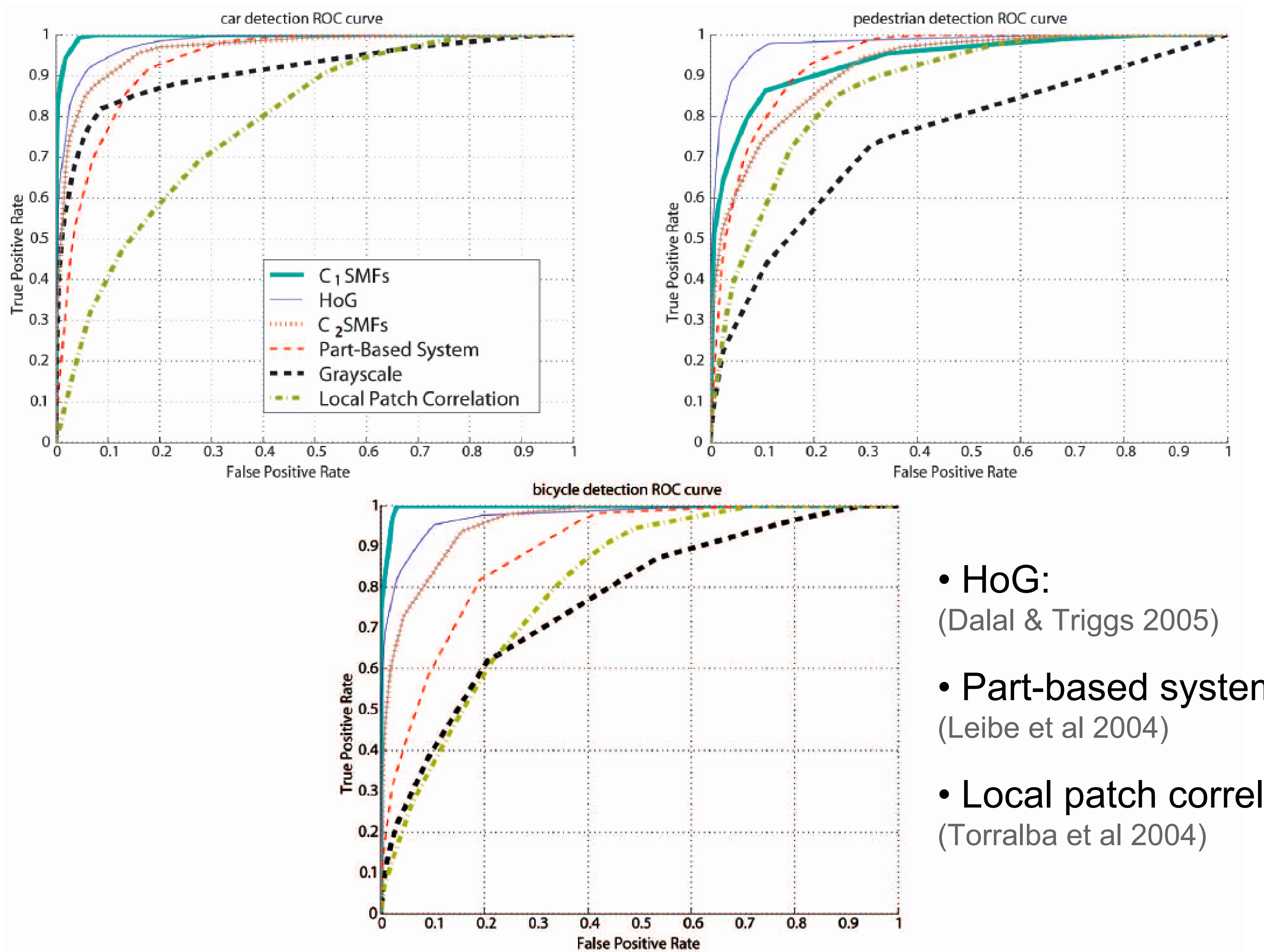
Results

Examples



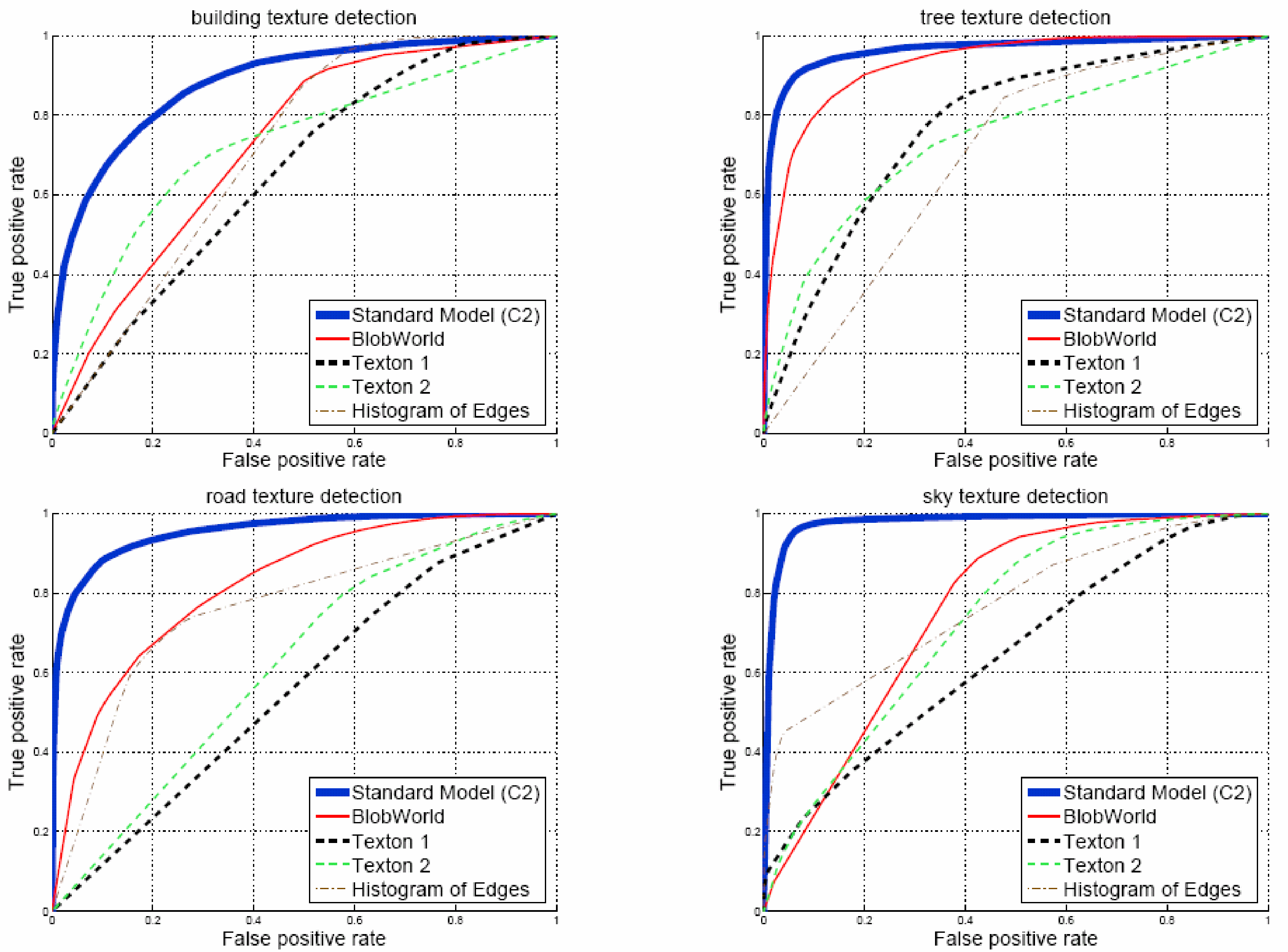






- HoG:
(Dalal & Triggs 2005)
- Part-based system:
(Leibe et al 2004)
- Local patch correlation:
(Torralba et al 2004)

Serre Wolf Bileschi Riesenhuber & Poggio PAMI 2007

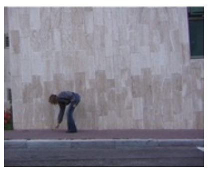


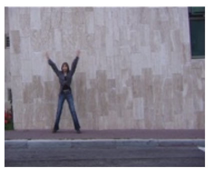
Serre Wolf Bileschi Riesenhuber & Poggio PAMI 2007

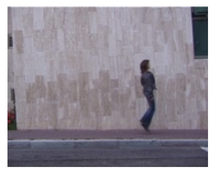
1. “Old” computer vision and learning work
2. Recent work in neuroscience of recognition can account for cell properties, human performance and provide good computer vision algorithms
3. Future: recognition in videos, a new learning theory inspired by cortex and extending approach to image inference tasks

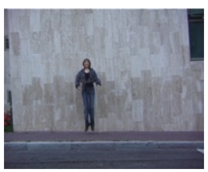
The problem: action recognition

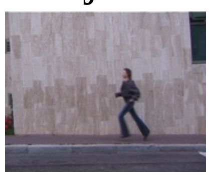
Training Videos


bend

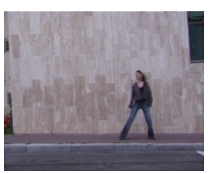

jack

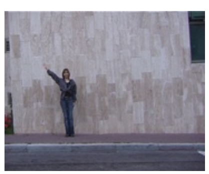

jump

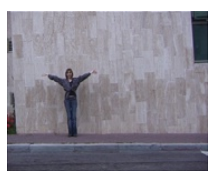

jump


run


walk


side


wave1

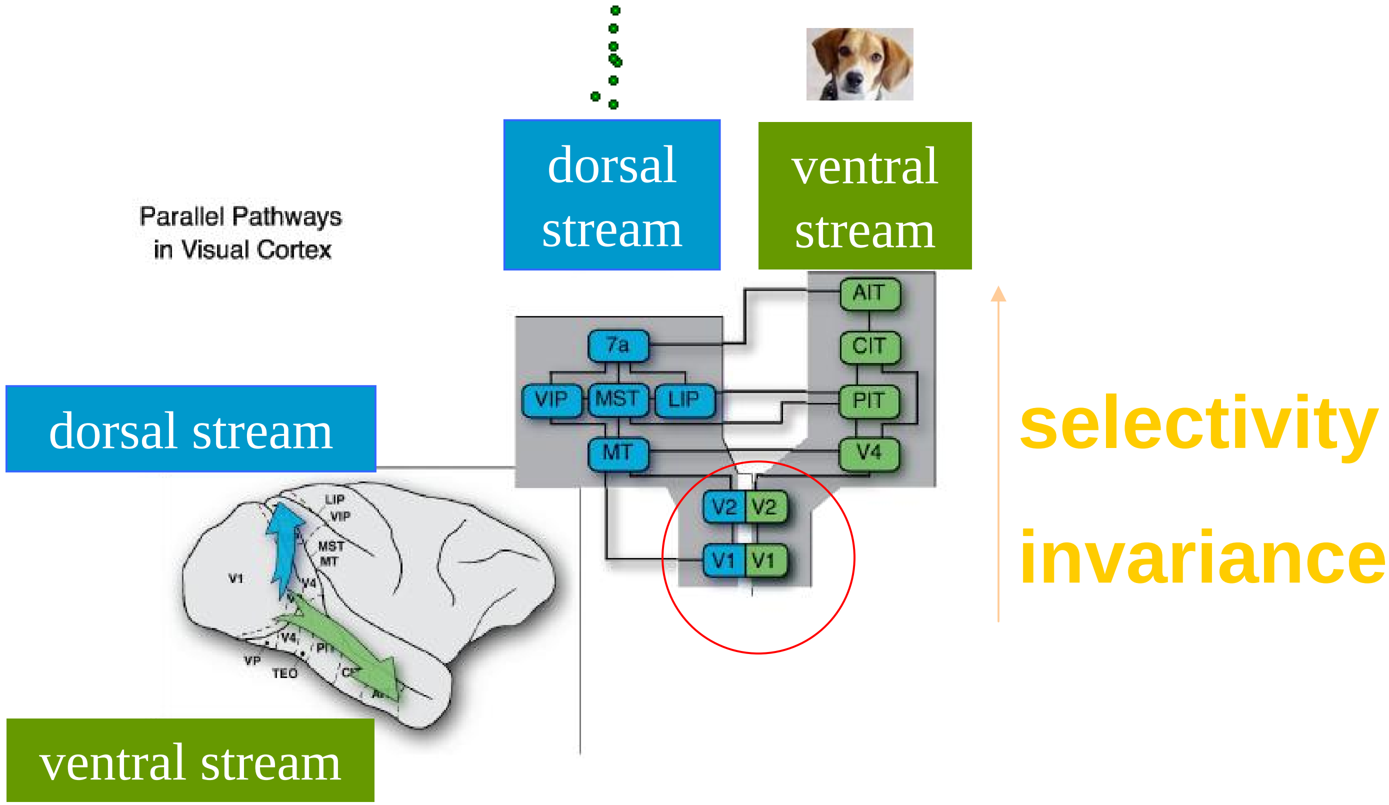

wave2

*each video~4s, 50~100 frames

Testing videos

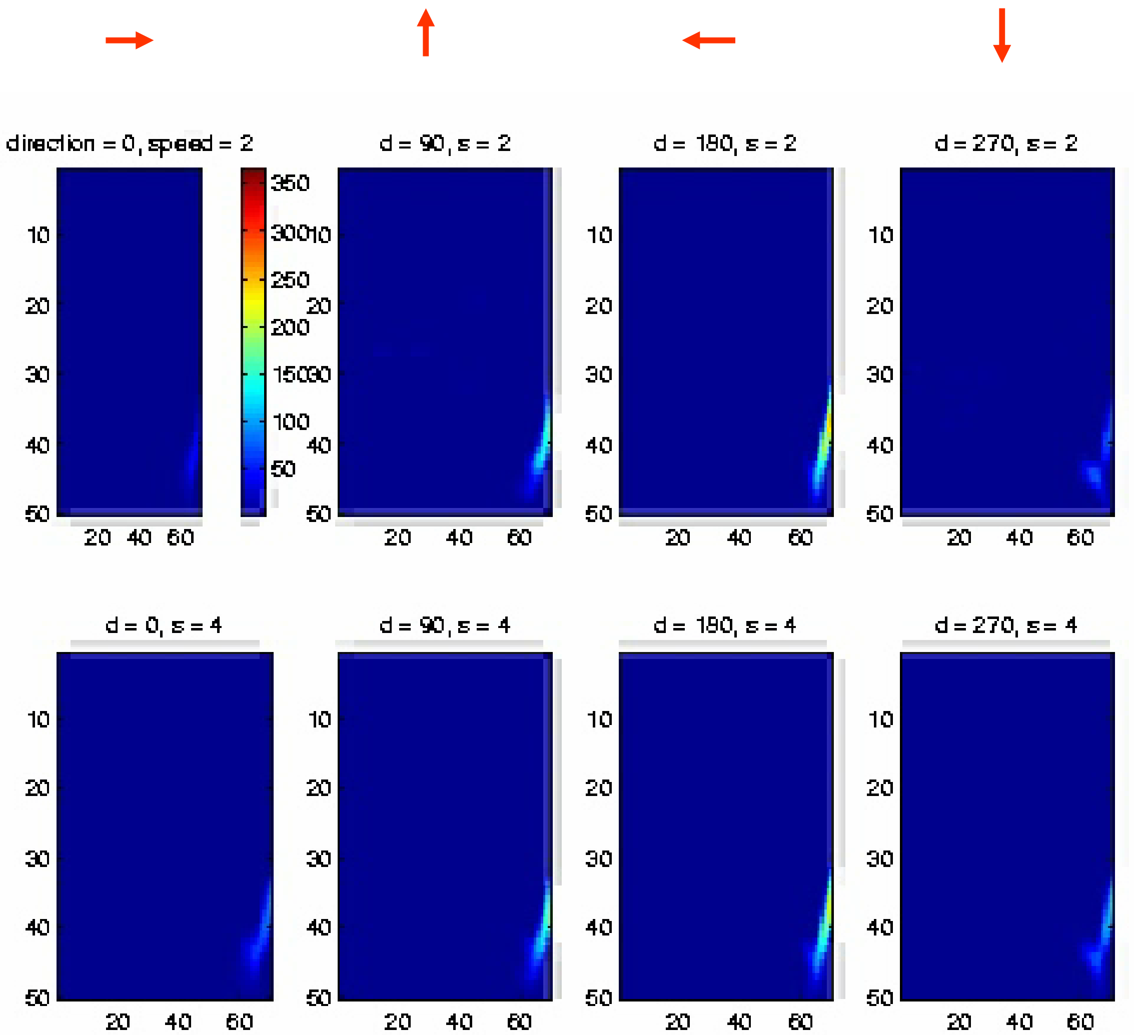
Dataset from (Blank et al, 2005)

A new model of the dorsal stream (motion) following the ventral stream model



Adapted from (Merigan & Maunsell, 1993; Maunsell & Newsome 1987)

Motion features: Spatio-temporal filters (S1 units in "V1")



Unsupervised learning in MT (S2) from natural video sequences

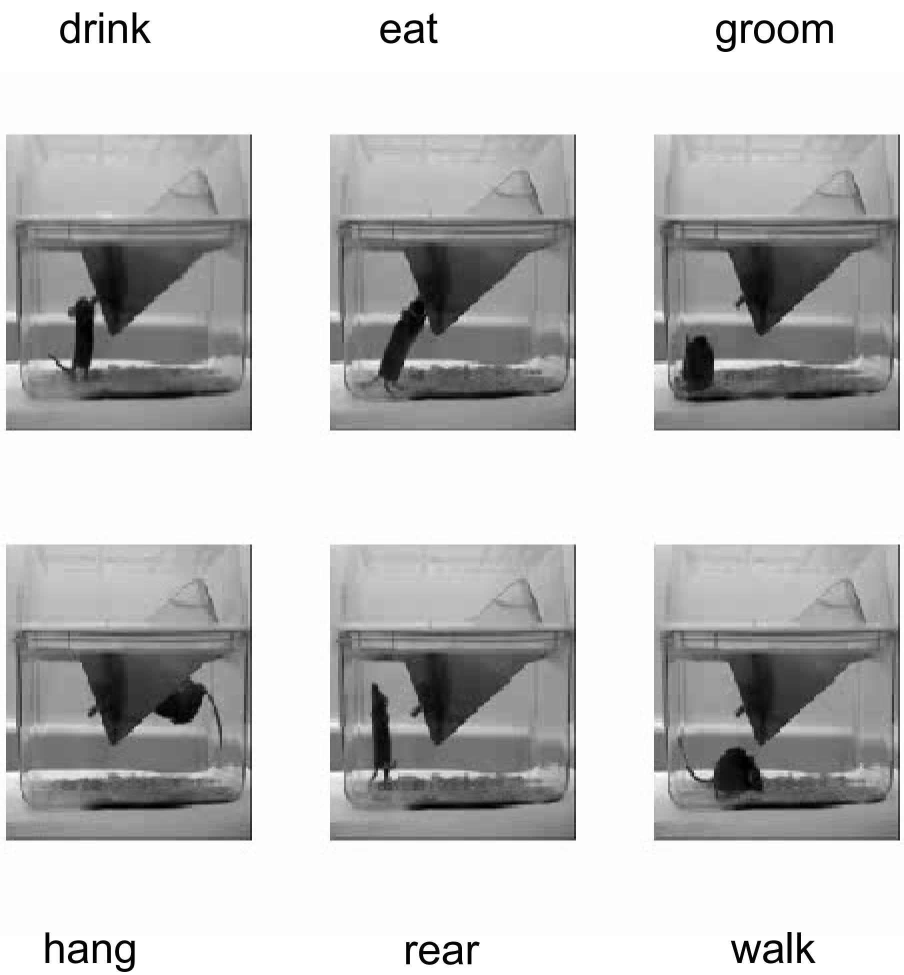
Using a large dictionary of MT-like units for action recognition works well!

	(Dolllar et al. 2005)	model	chance
KTH Human	81.3%	**91.6%**	16.7%
UCSD Mice	75.6%	**79.0%**	20.0%
Weiz. Human	86.7%	**96.3%**	11.1%

- Cross-validation: 2/3 training, 1/3 testing, 10 repeats
- Source code for benchmark graciously provided by Piotr Dollar

(Jhuang Serre Wolf & Poggio ICCV 2007)

A twist: a vision system derived from visual cortex may help biology: Automatic classification of abnormal behavior in mutant vs. wild mice



over 95% correct for 6 class-classification

Serre, Steele, Jhuang, Garrote & Poggio

1. "Old" computer vision and learning work
2. Recent work in neuroscience of recognition can account for cell properties, human performance and provide good computer vision algorithms
3. Future: recognition in videos, a new learning theory inspired by cortex and extending approach to image inference tasks

From a model to a theory

Notices of the American Mathematical Society (AMS), Vol. 50, No. 5, 537-544, 2003.

The Mathematics of Learning: Dealing with Data

Tomaso Poggio and Steve Smale

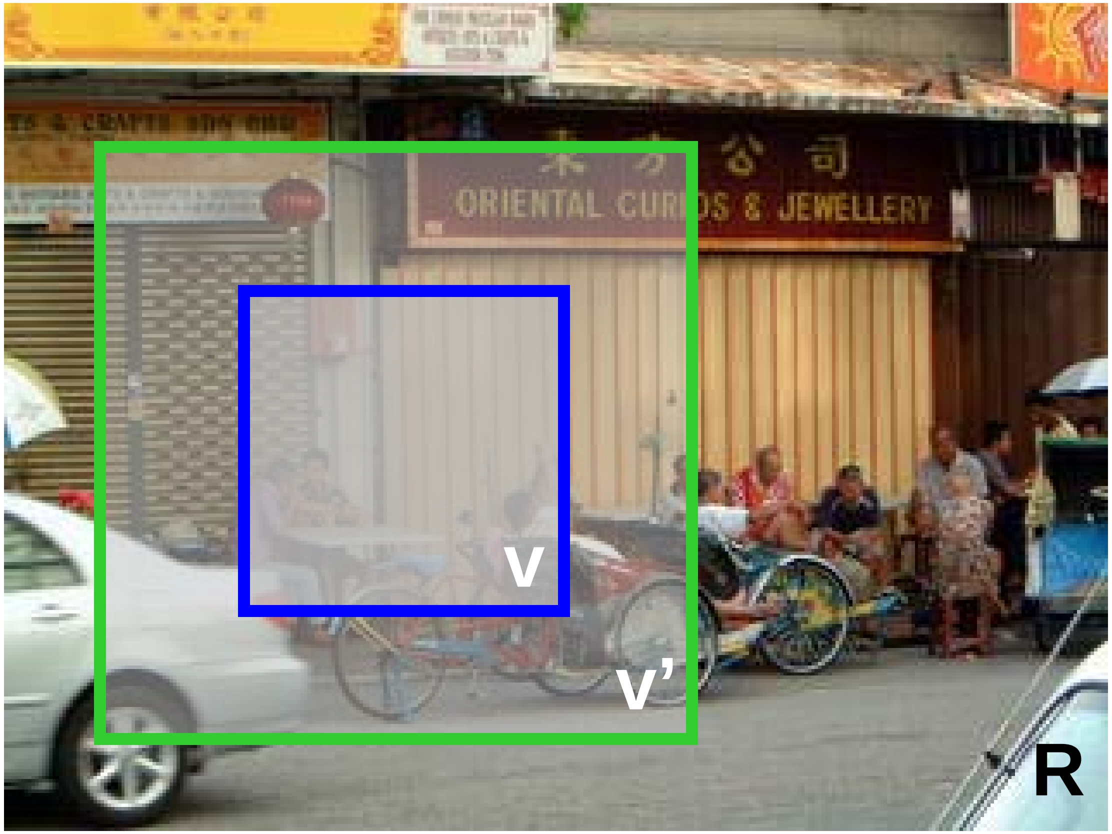
How then do the learning machines described in the theory compare with brains?

- One of the most obvious differences is the ability of people and animals to learn from very few examples.

- A comparison with real brains offers another, related, challenge to learning theory. The "learning algorithms" we have described in this paper correspond to one-layer architectures. Are hierarchical architectures with more layers justifiable in terms of learning theory?

- Why hierarchies?

Formalizing the cortical hierarchy: towards a new class of learning theories?



Derived Distance:

- Iterated analysis with arbitrary transforms and nonlinearities in between layers.
- Template dictionaries at each layer.
- First layer performs simple template matching over the set of allowed transformations.
- At higher layers, we work with representations based on previous layers' templates.

Axiom: $f \circ h : v \to [0,1]$ is in $Im(v)$ if $f \in Im(v')$ and $h \in H$, that is *the restriction of an image is an image* and similarly for H'. Thus

$$f \circ h : v \to [0,1] \in Im(v) \text{ if } f \in Im(v') \text{ and } h \in H,$$
$$f \circ h' : v' \to [0,1] \in Im(v') \text{ if } f \in Im(R) \text{ and } h' \in H'.$$

Smale, S., T. Poggio, A. Caponnetto, and J. Bouvrie. Derived Distance: towards a mathematical theory of visual cortex, *CBCL Paper*, Massachusetts Institute of Technology, Cambridge, MA, November, 2007.

1. “Old” computer vision and learning work
2. Recent work in neuroscience of recognition can account for cell properties, human performance and provide good computer vision algorithms
3. Future: recognition in videos, a new learning theory inspired by cortex and extending approach to image inference tasks

Future directions

- Normal vision is <u>much </u>more than categorization or identification: it is image understanding/inference/parsing

- Our visual system can “answer” almost any kind of question about an image: a Turing test…

Future Directions:
beyond feedforward models

Image inference:

at least two classes of possible models

- o Attentional (with visual routines)

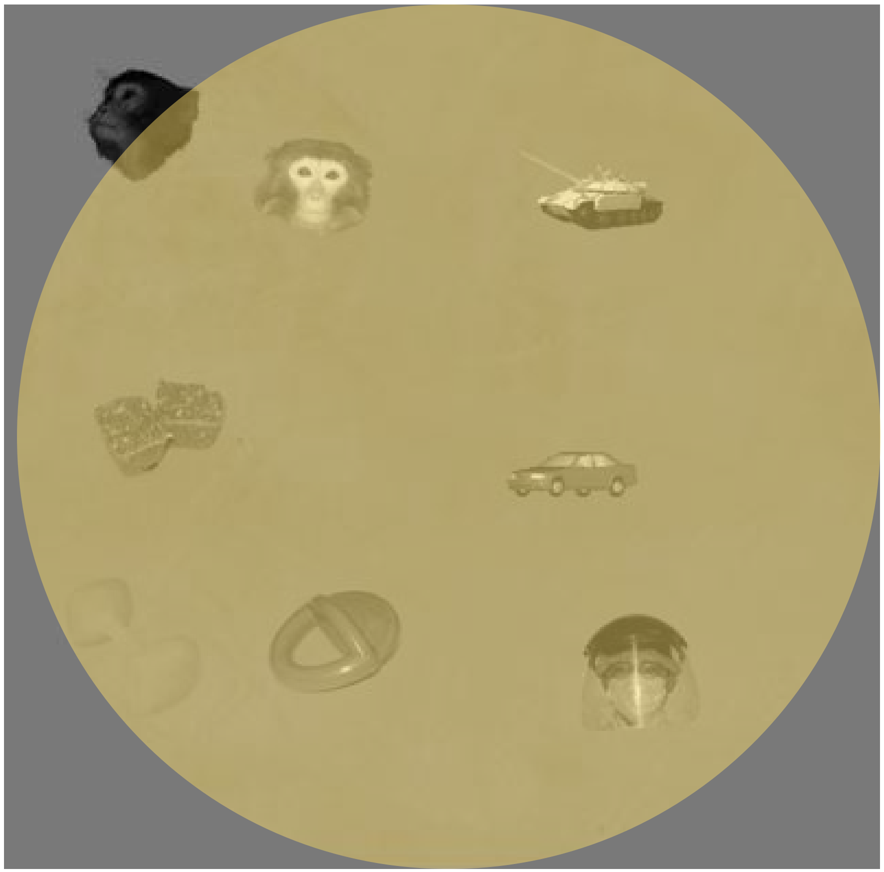
or

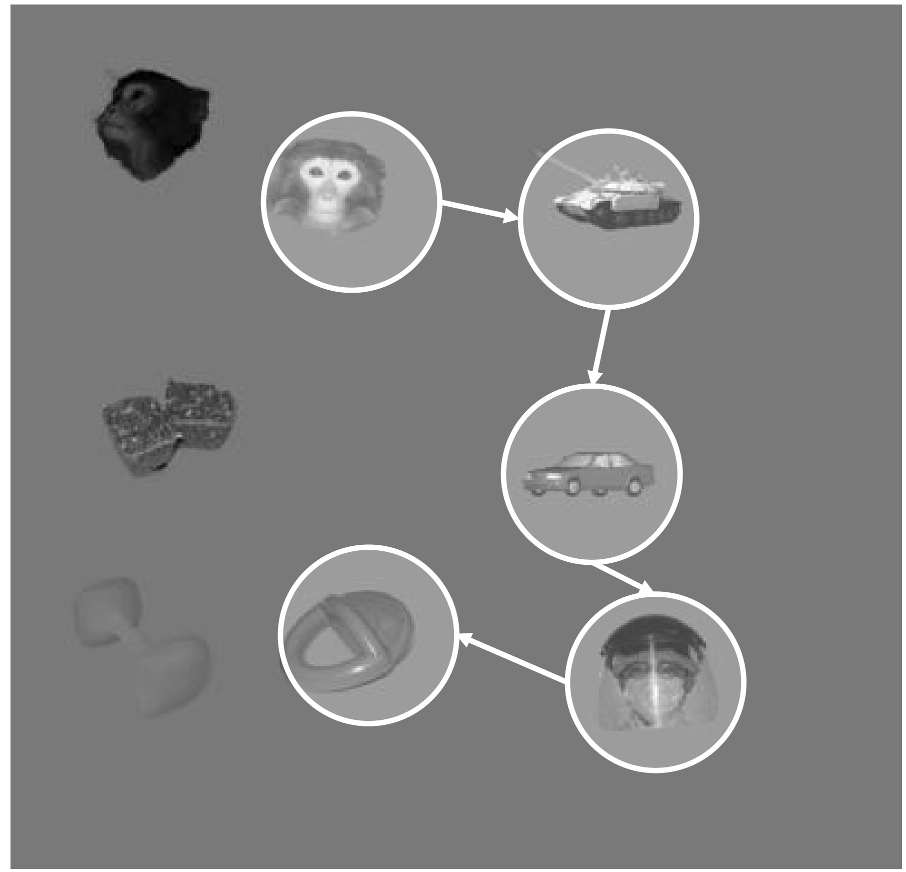
- o Bayesian

?

Lee and Mumford, 2003; Dean,2005; Rao, 2004; Hawkins, 2004; Ullman, 2007, Hinton, 2005;……

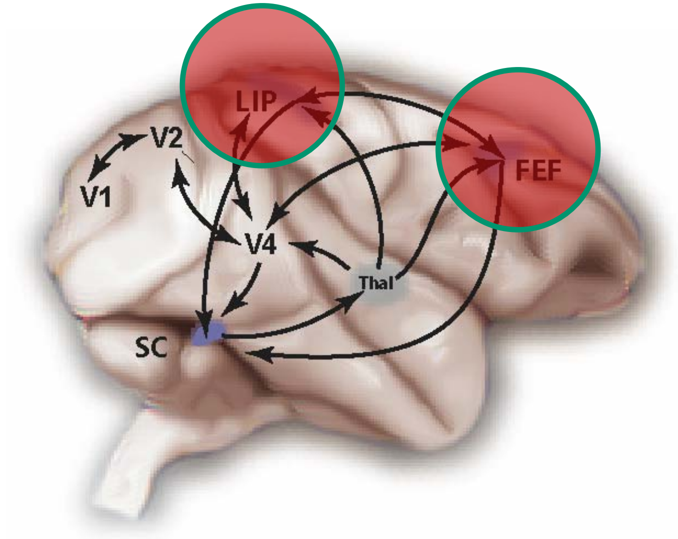
Attention is needed for robust recognition in clutter and for inspecting an image...

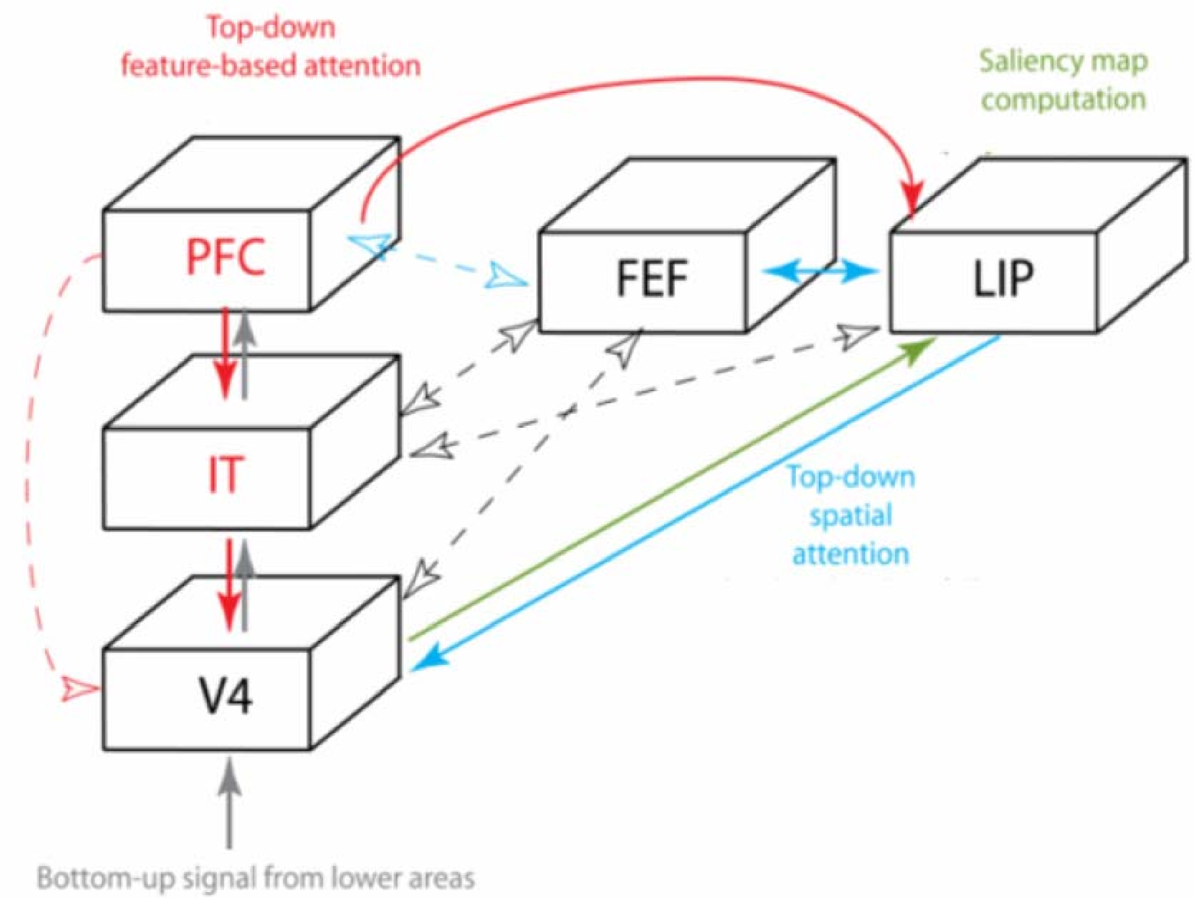




Wolfe, Tsotsos, …

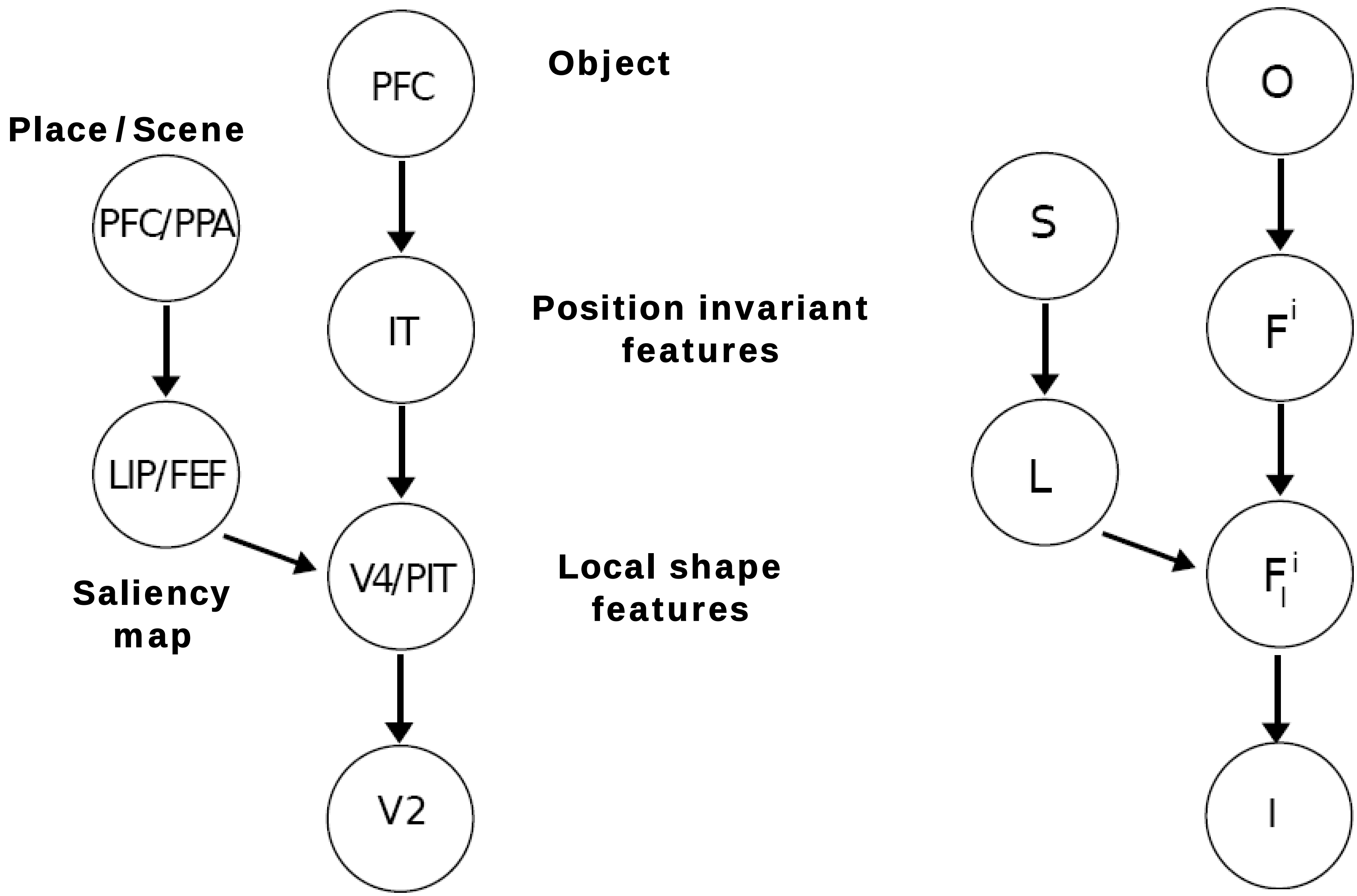
Biology of attention





Computational model: A Bayesian approach

Bayesian Model



Comparing this

top-down attentional model

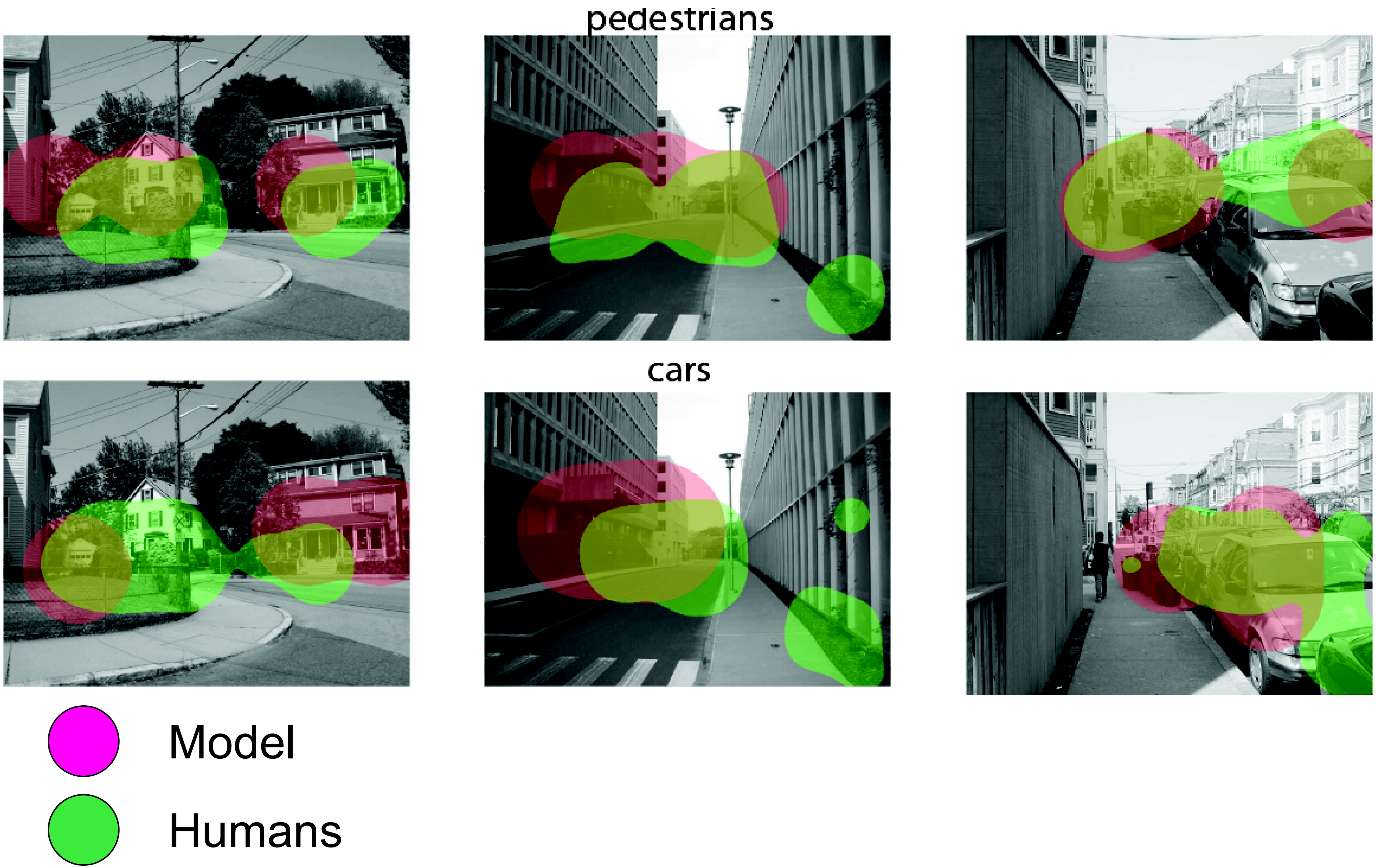
with human eye fixations

in natural scenes
(we get better results than bottom-up
models such as Itti-Koch)

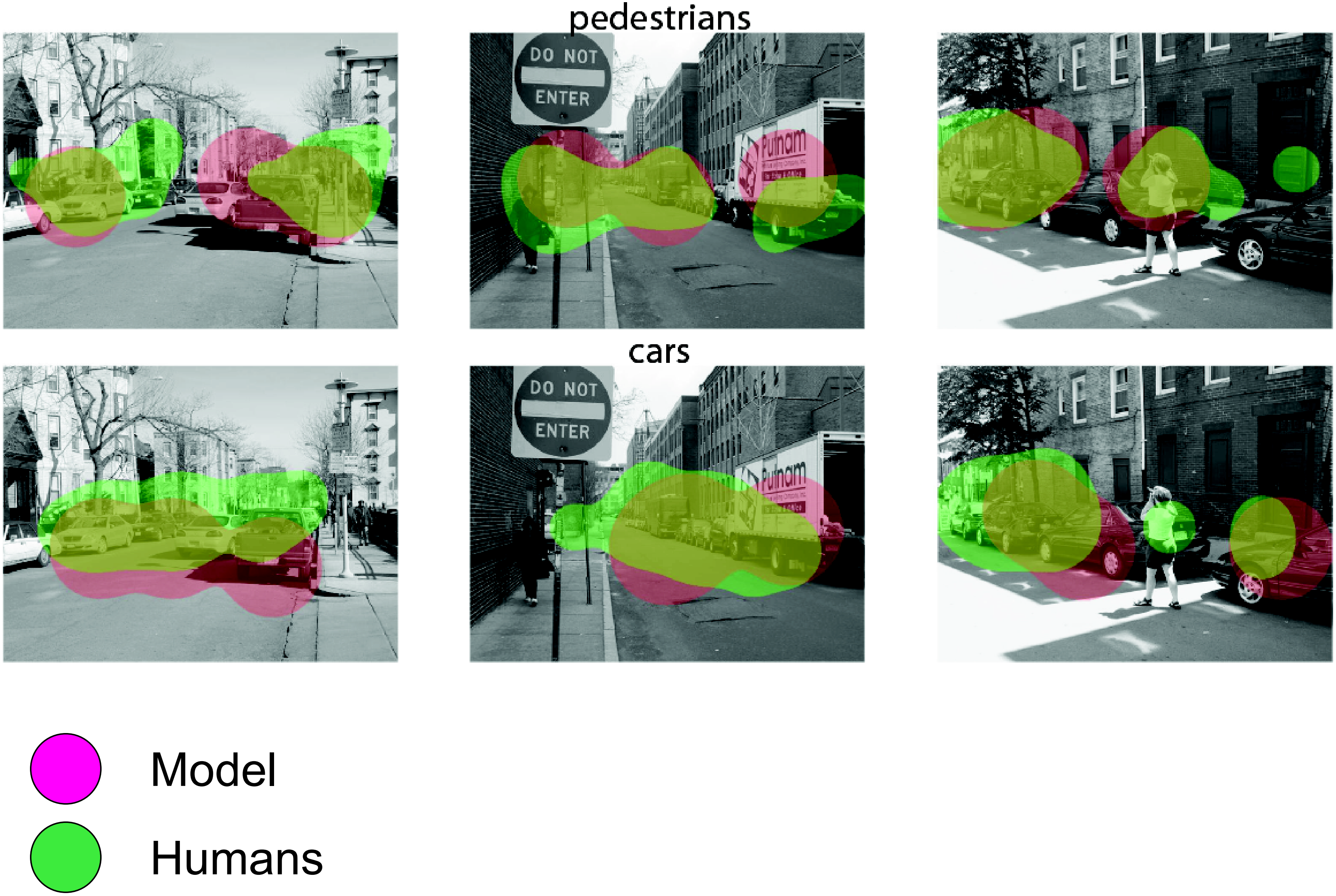
Pyschophysics

- **Dataset**
  - 100 CBCL street-scenes images having cars & pedestrians
  - 20 images with neither objects
- **Experiment**
  - 8 subjects (drawn from the university undergraduate population) where shown these 120 images in random order.
  - The stimuli extends about 12◦ visual angle.
  - Each image in the stimuli-set was presented twice
  - The subjects were asked to count the number of cars/pedestrians
  - For each of these block trials, the subject's eye movements were recorded using an infra-red eye tracker.

Example Stimuli



Example Stimuli



The top-down attentional model
also seems to improve performance in
object recognition in clutter
(very preliminary results)

Future Directions: beyond feedforward models

Image inference

(vision is more than categorization):

at least two classes of possible models

- o Attentional (with visual routines)

or

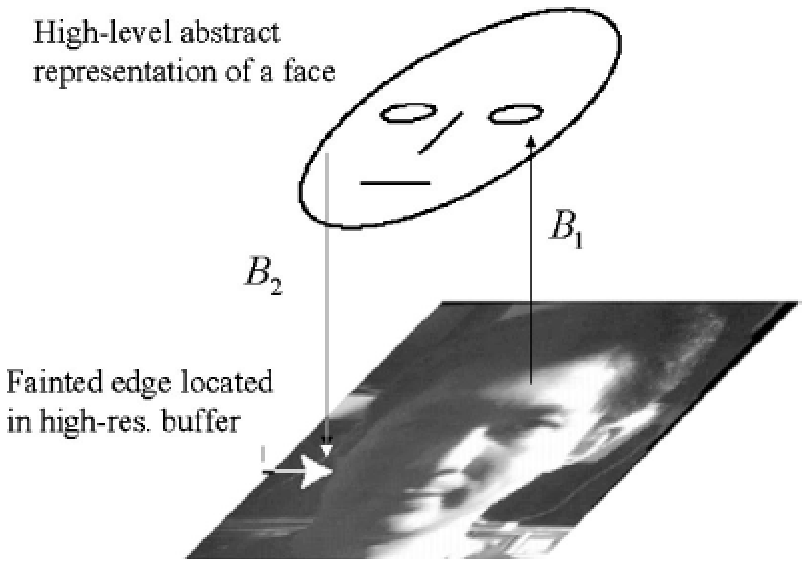
- o Bayesian

?

Lee and Mumford, 2003; Dean,2005; Rao, 2004; Hawkins, 2004; Ullman, 2007, Hinton, 2005;……

2. Bayesian models

Analysis-by-synthesis models, eg probabilistic inference in the ventral stream: neurons represent conditional probabilities of the bottom-up sensory inputs given the top-down hypothesis and converge to globally consistent values



Lee and Mumford, 2003; Dean,2005; Rao, 2004; Hawkins, 2004; Ullman, 2007, Hinton, 2005

Discussion topics

Human vision is much better than feedforward models…

Are attentional models of the type we are exploring – and which *predict well* human eye fixations and seem to *improve recognition* in clutter – likely to fully bridge the gap?

Neurally plausible models may just beginning to provide new insights on how to implement intelligence in machines

Collaborators in recent work

T. Serre

- ❑ Comparison w| humans
  - ✓ A. Oliva
- ❑ Action recognition
  - ✓ H. Jhuang
- ❑ Read-out
  - ✓ E. Meyers
  - ✓ W. Freiwald
- ❑ Attention
  - ✓ S. Chikkerur
  - ✓ C. Tan

Also: C. Koch, D. Walther, C. Cadieu, U. Knoblich, M. Kouh, G. Kreiman. M. Riesenhuber,T. Masquelier, S. Bileschi, L. Wolf, J. Dicarlo, E. Miller, B. Desimone, E. Connor. D. Ferster, I. Lampl, A. Pasupathy