Learning:

neuroscience
and
engineering applications

or…

a motivation for Steve’s talk

Tomaso Poggio

CBCL
Center for Biological
& Computational Learning

McGOVERN INSTITUTE
FOR BRAIN RESEARCH AT MIT

Computational neuroscience <u>may</u> just begin to understand what is going on in visual cortex from a *computational* point of view

- ❑ We have an algorithm that mimics the ability of people to recognize complex images but…

- ❑ …we have a model, **not** a theory: one <u>motivation</u> for Steve's talk!

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

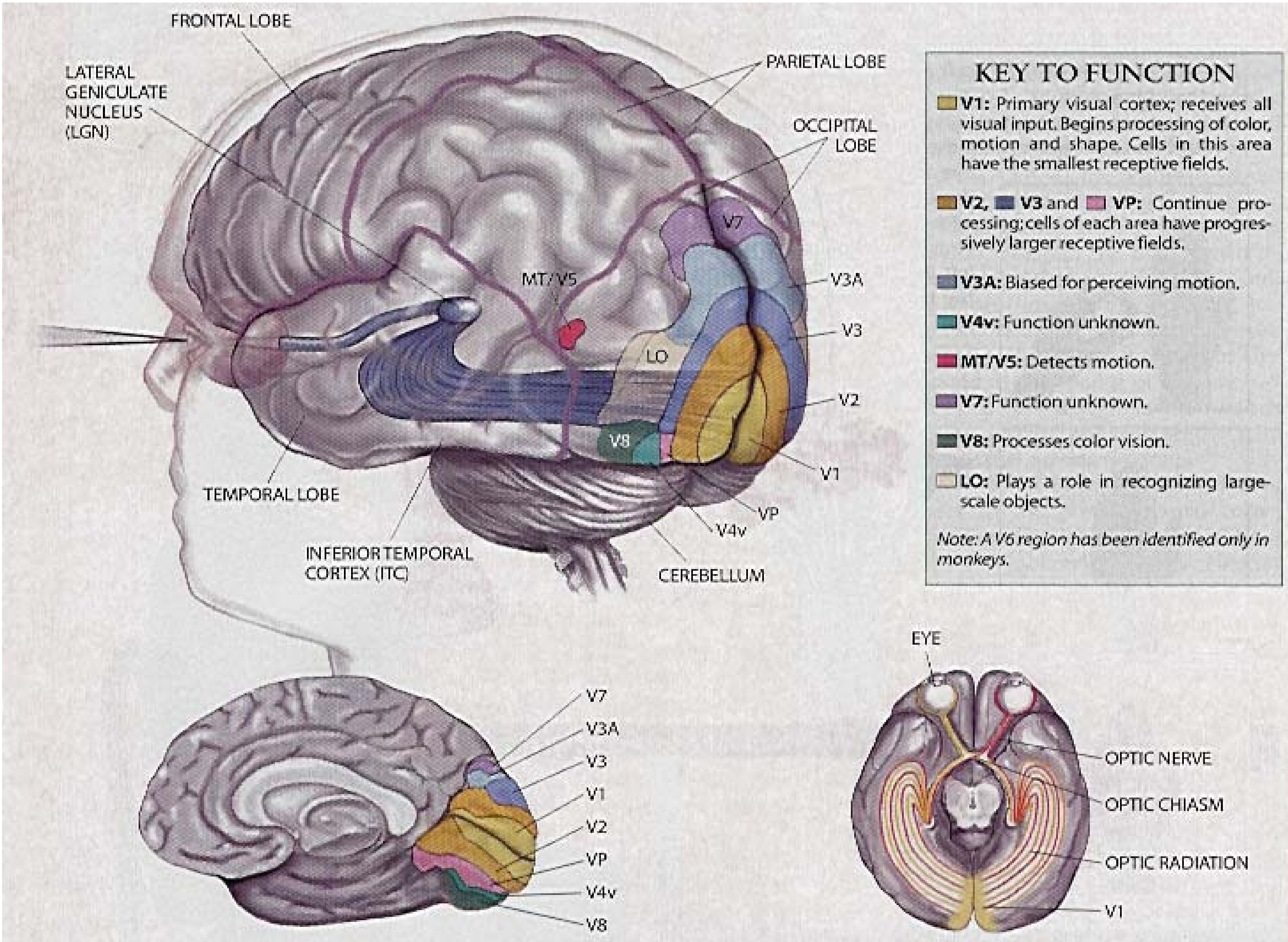
❑ How then do the learning machines described by regularization compare with brains? **One of the most obvious differences is the ability of people and animals to learn from very few examples.**

❑ 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?**

❑ There may also be the more fundamental issue of ***sample complexity***. ... Thus our ability of learning from just a few examples, and its limitations, may be related to the **hierarchical architecture of cortex.**

❑ **Why hierarchies?**

Primer on the Brain



A complex electro-chemical computing machinery

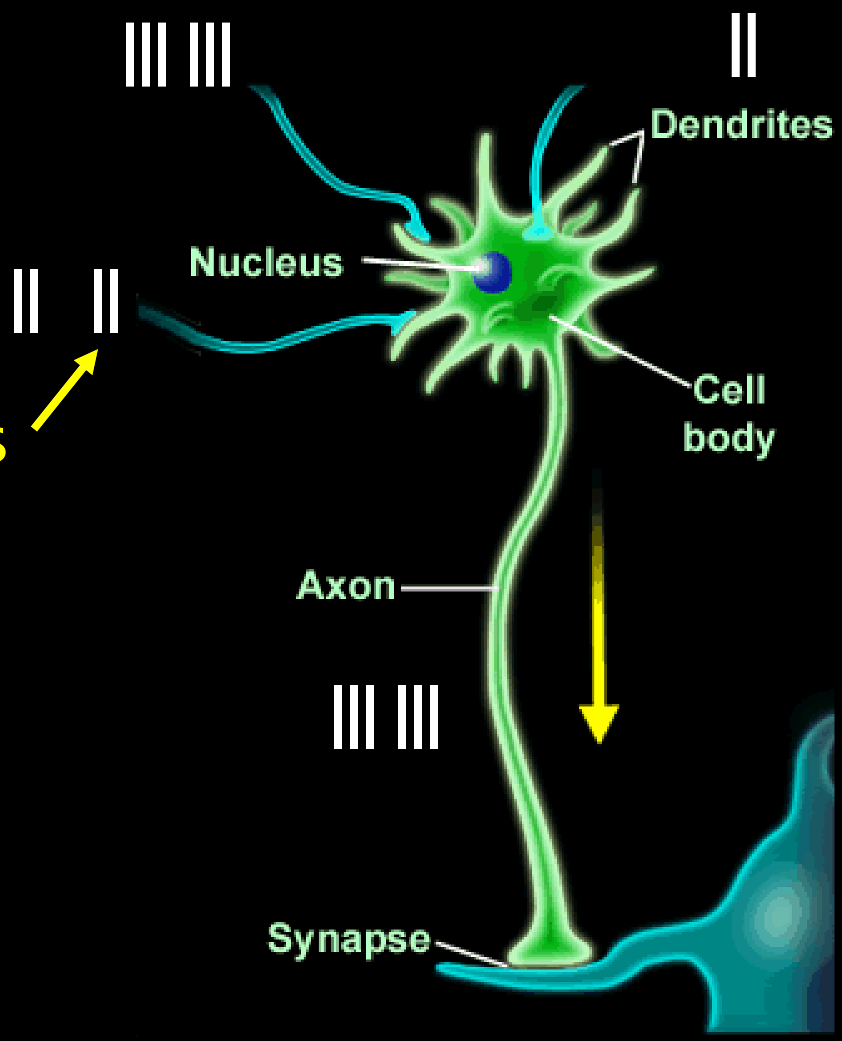
Human Brain

- ~ 10^{11} neurons (1 million flies ☺)
- ~ 10^{14}– 10^{15} synapses
- ~ 5 billion neurons in human visual cortex

Neuron

Fundamental space dimension: fine dendrites : 0.1 μ
Fundamental time length : 1 msec

Neuron basics



spikes

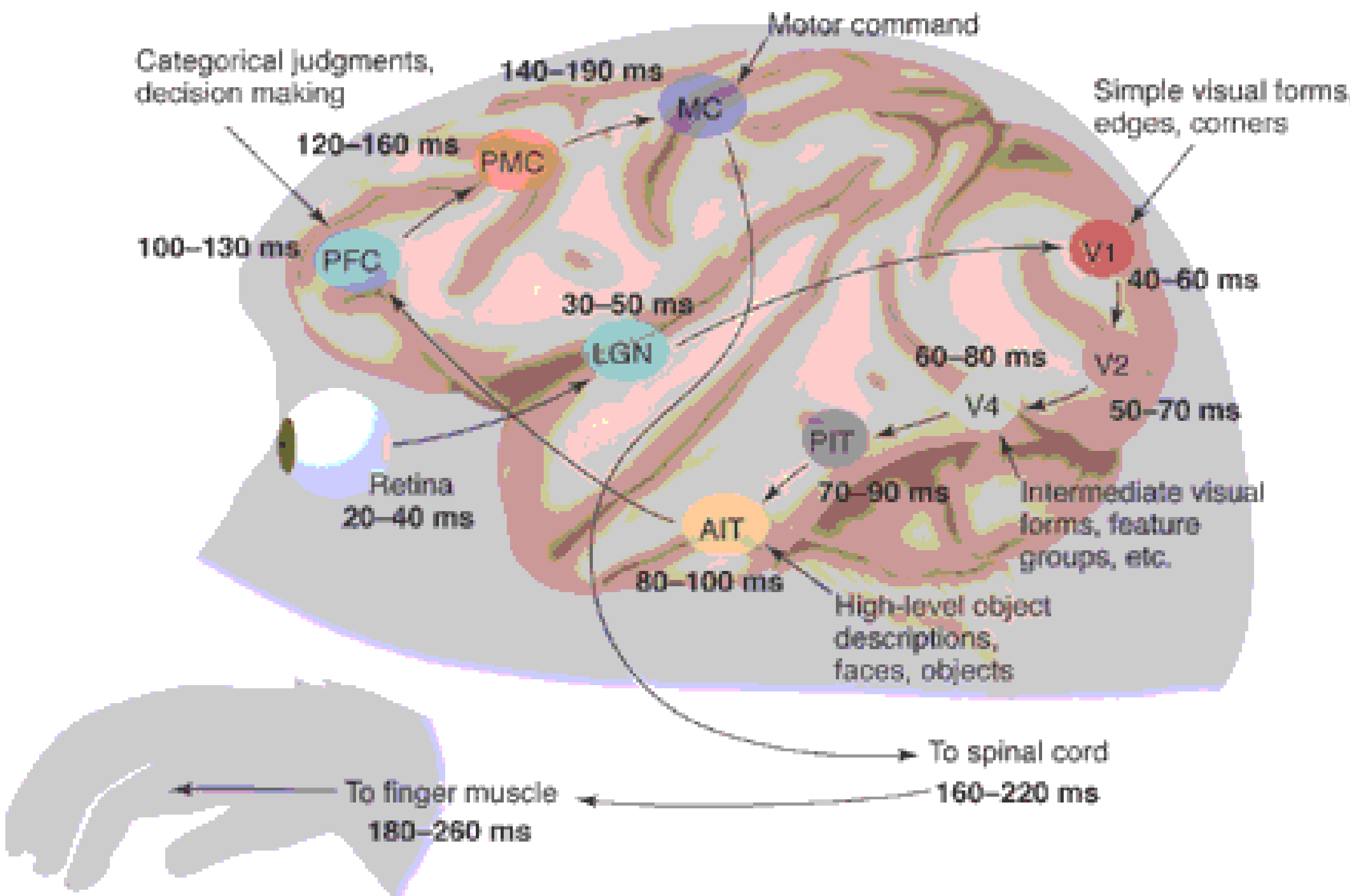
INPUT
= pulses

COMPUTATION

OUTPUT
= pulses

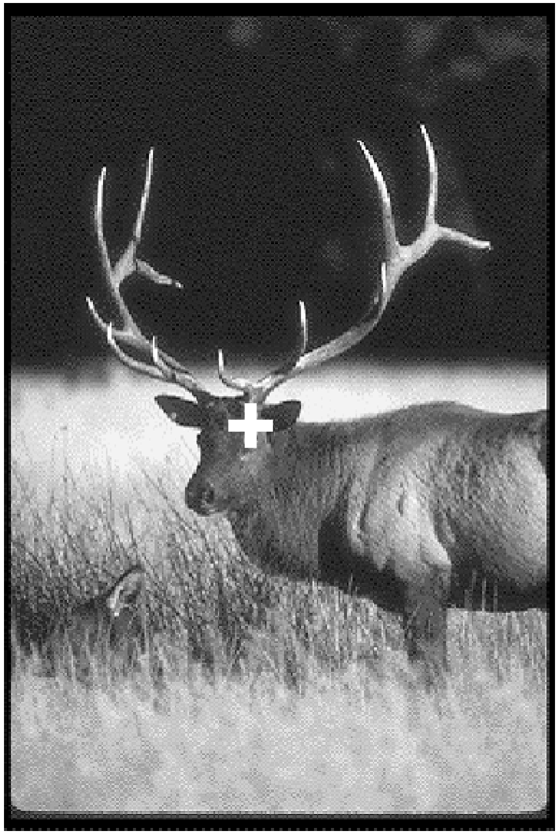
How does the brain recognize objects? Consider how well we do in the first 100 msec (no eye movements)

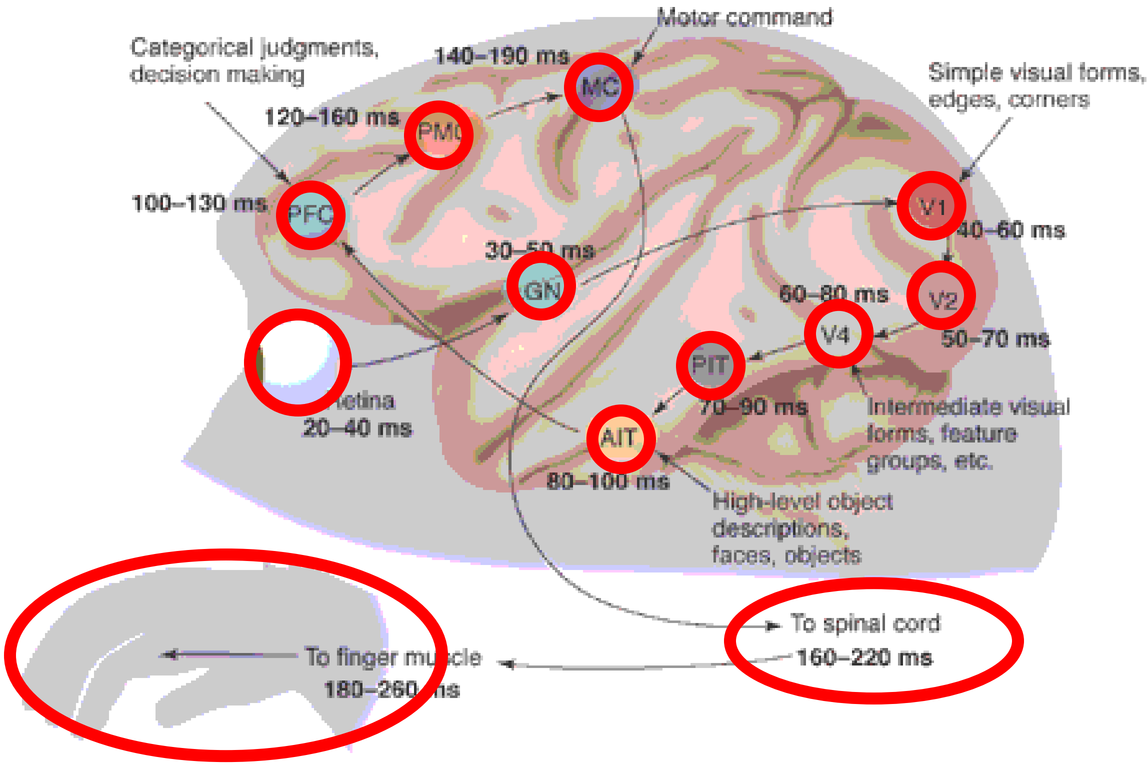
What do we know about cortex and recognition?



Simon Thorpe

Consider feedforward processing in the ventral stream (we neglect feedback for now)





[Thorpe and Fabre-Thorpe, 2001]

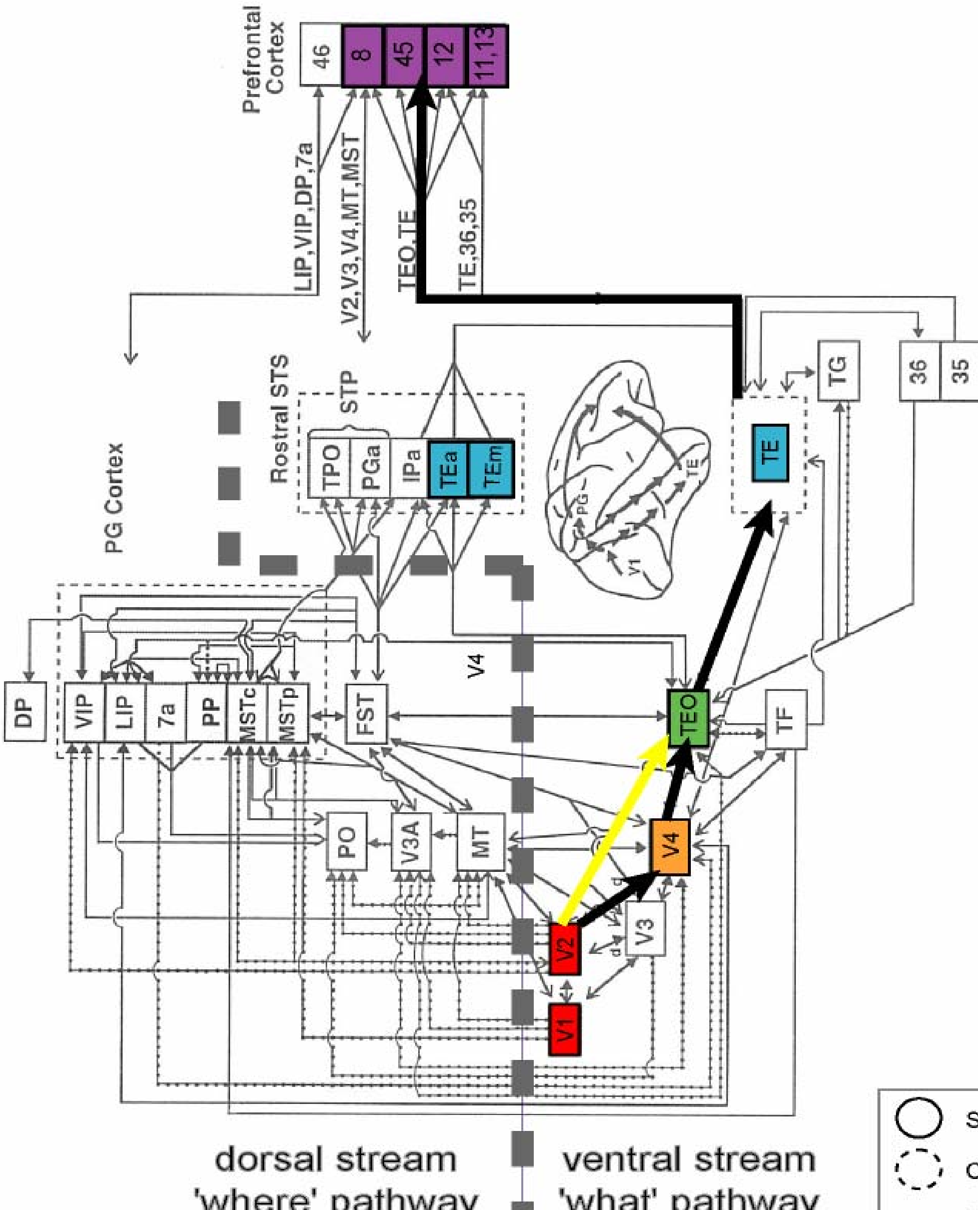
A computational model

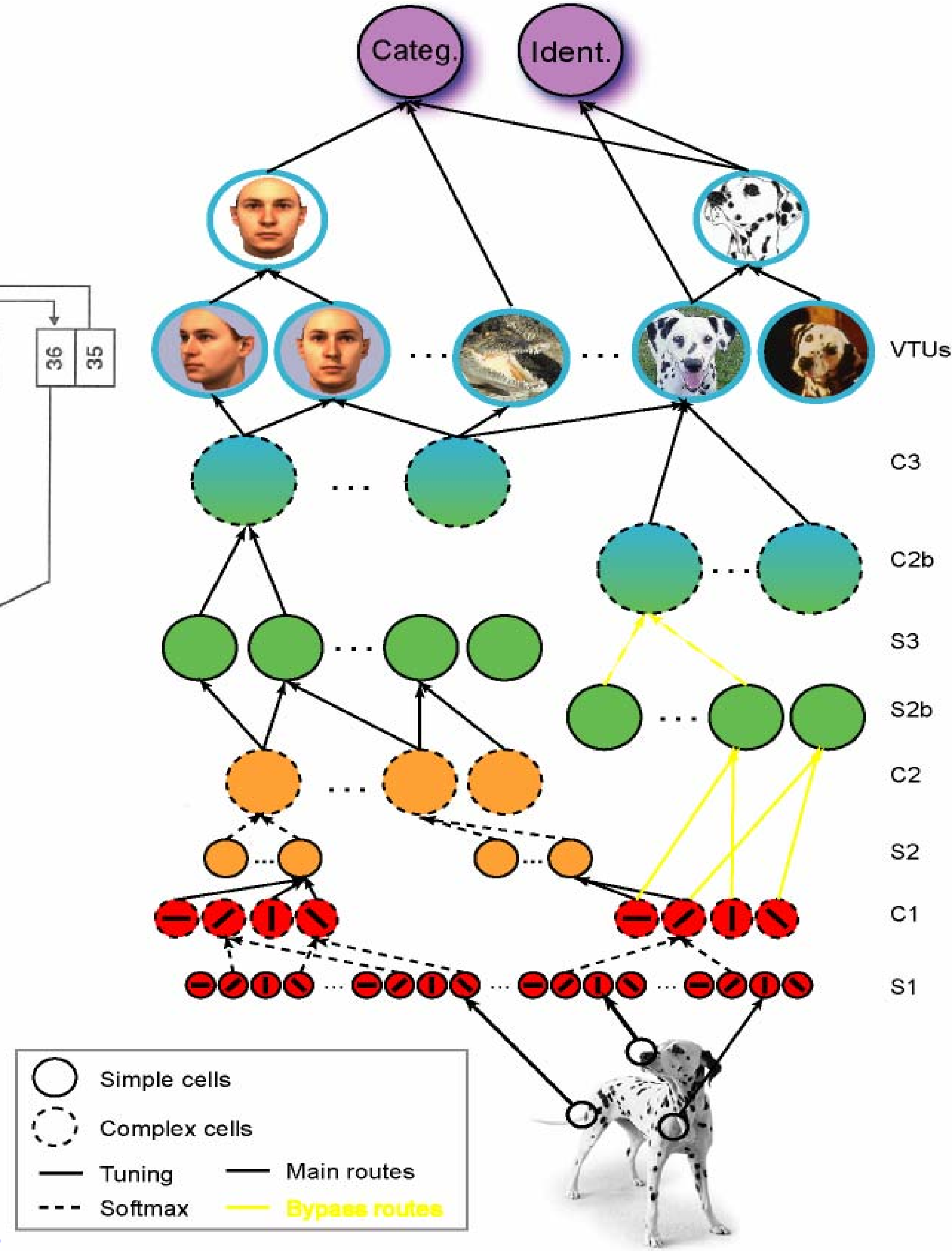
(which can be simulated on a computer)

A model: feedforward → accounting only for immediate perception

- It is in the family of "Hubel-Wiesel" models (Hubel & Wiesel, 1959; Fukushima, 1980; Oram & Perrett, 1993, Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe; Mel; Koerner...)
- It is the most quantitative and faithful to known biology (though many details/facts are unknown or still to be incorporated)

Quant model derived from anatomy+physiology data in V1, V4, IT, PFC
(millions of artificial neurons simulated on a computer)



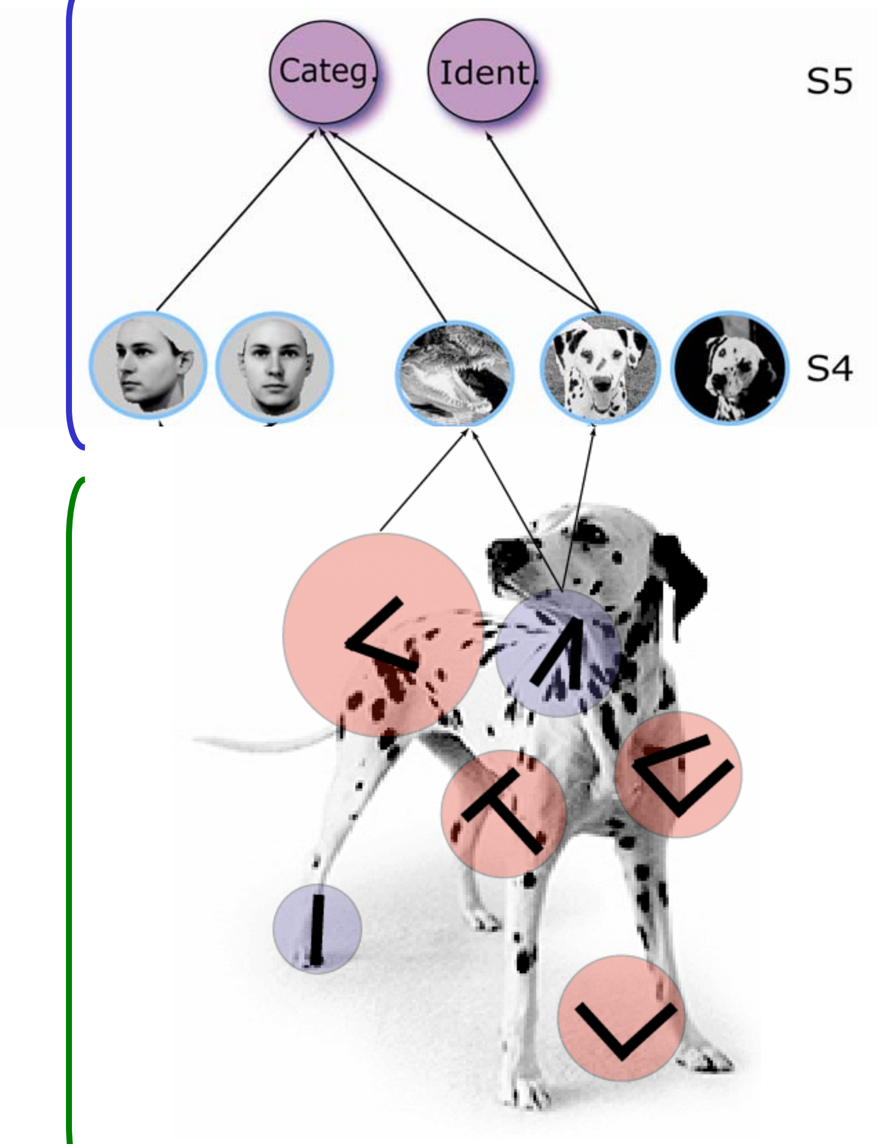


Serre, Kouh, Cadieu, Knoblich, Poggio, 2005;
Serre, Oliva, Poggio, PNAS, 2007

Layers of cortical processing units

- Task-specific circuits (from IT to PFC)
  - Supervised learning

- Generic dictionary of shape components (from V1 to IT)
  - Unsupervised learning during a developmental-like stage learning dictionaries of "templates" at different S levels



SCP model

- $$f \circ h : v \to [0,1] \in Im(v)$$

  if $f \in Im(v')$ and $h \in H(v)$,

- $$f \circ h' : v' \to [0,1] \in Im(v')$$

  if $f \in Im(R)$ and $h' \in H'(v')$,

1. The process starts with a distance on v provided by

$$d'_0(f,g) = d(f,g) = ||f-g||_p$$

   Then we define a *Neural Similarity* following Steve:

$$N_t^{1,C}(f) = \min_{h \in H} d(f \circ h, t)$$

   where $N_t^{1,C}(f)$ corresponds to the response of a C1 cell with template t and with receptive field – the region over which the min is taken – corresponding to v'.

2. We now can repeat the process by defining the derived distance of Steve on $Im(v')$ as

$$d'_1(f,g) = ||N^{1,C}(f) - N^{1,C}(g)||_p$$

   and the second stage *Neural Similarity* with

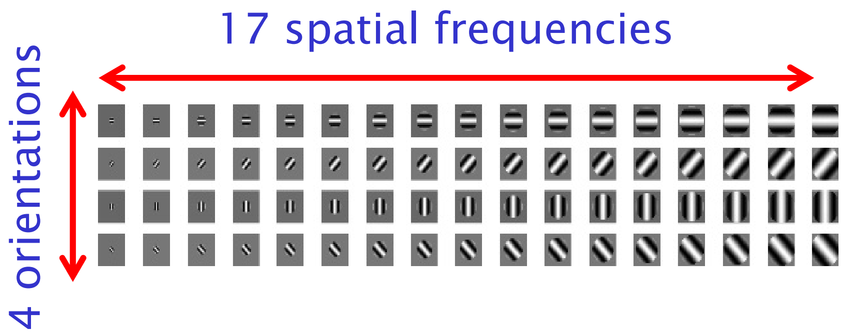
$$N_{t'}^{2,C}(f) = \min_{h' \in H'} d'_1(f \circ h', t')$$

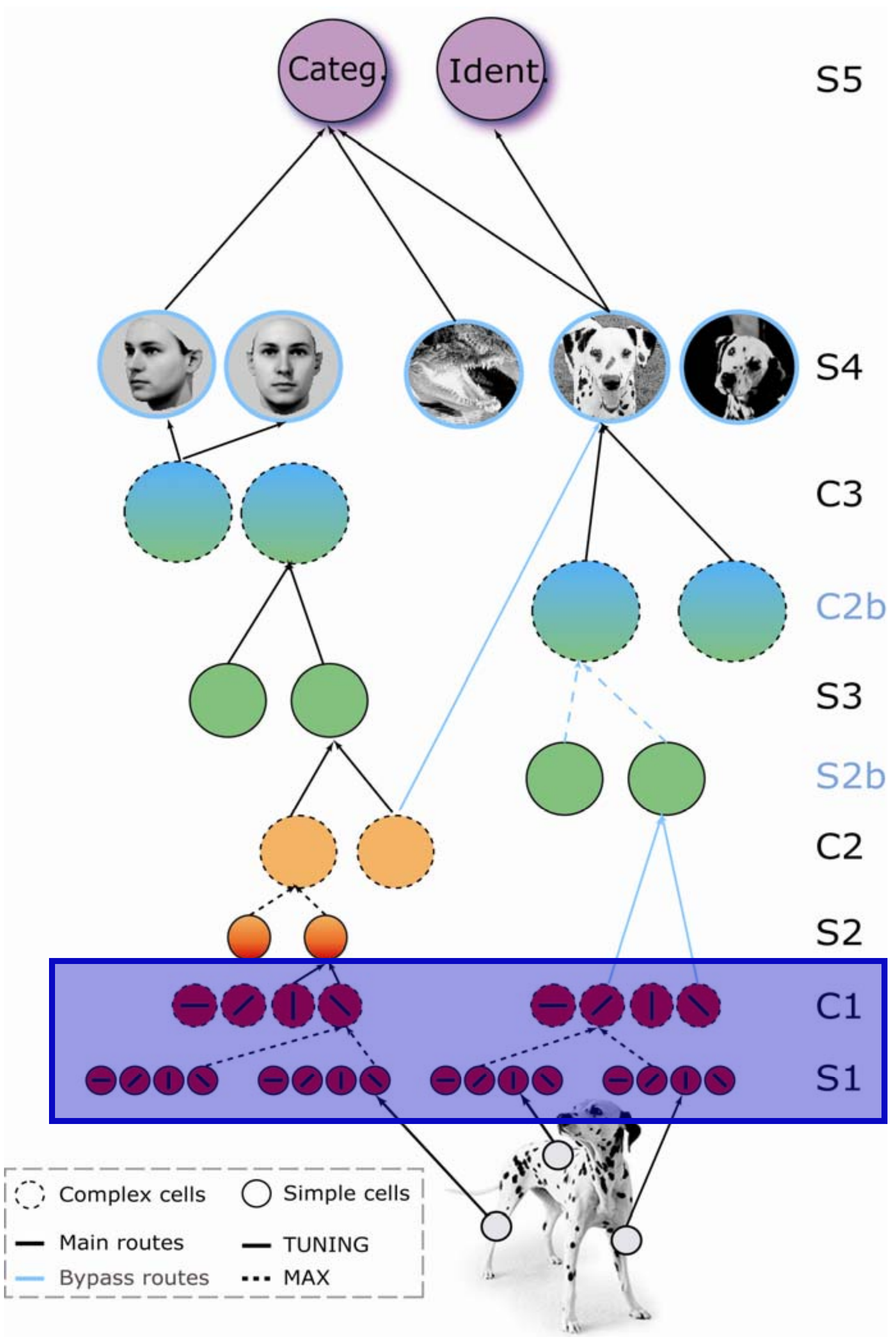
   which is the response of a C2 cell with receptive field R.

3. The new derived distance is now on $Im(R)$

$$d'_2(f,g) = ||N^{2,C}(f) - N^{2,C}(g)||_p.$$

S1 and C1 units





Partial list of detailed 'predictions" of physiology and psychophysics data

- MAX in V1 (Lampl et al, 2004) and V4 (Gawne et al, 2002)
- Tow-spot reverse correlation in V1 (Livingstone and Conway, 2003; Serre et al, 2005)
- Tuning for boundary conformation (Cadieu et al., 2007, Pasupathy & Connor, 2001) in V4
- Tuning for gratings in V4 (Gallant et al, 1996; Serre et al, 2005)
- Tuning for two-bar stimuli in V4 (Reynolds et al, 1999; Serre et al, 2005)
- Tuning to Cartesian and non-Cartesian gratings in V4 (Serre et al, 2005)
- Two-spot interaction in V4 (Freiwald et al, 2005; Cadieu et al. 2007)
- Tuning and invariance properties in AIT (Logothetis et al, 1995)
- Average "average effect" in IT (Zoccolan, Cox & DiCarlo, 2005)
- IT *read out* data (Hung et al, 2005)
- Differential role of IT and PFC in categ. (Freedman et al, 2001,2002,2003)
- Trade-off of selectivity and invariance in IT (Zoccolan, Kouh, DiCarlo, 2007)
- Face inversion effect (Riesenhuber, Sinha et al, 2004)
- Rapid categorization (Serre et al., 2005, 2007)

Serre, Kouh, Cadieu, Knoblich, Kreiman, Poggio. *MIT AI Memo* 2005

So…the model fits many physiological data (V1, V4, IT, PFC…), predicts several new ones…

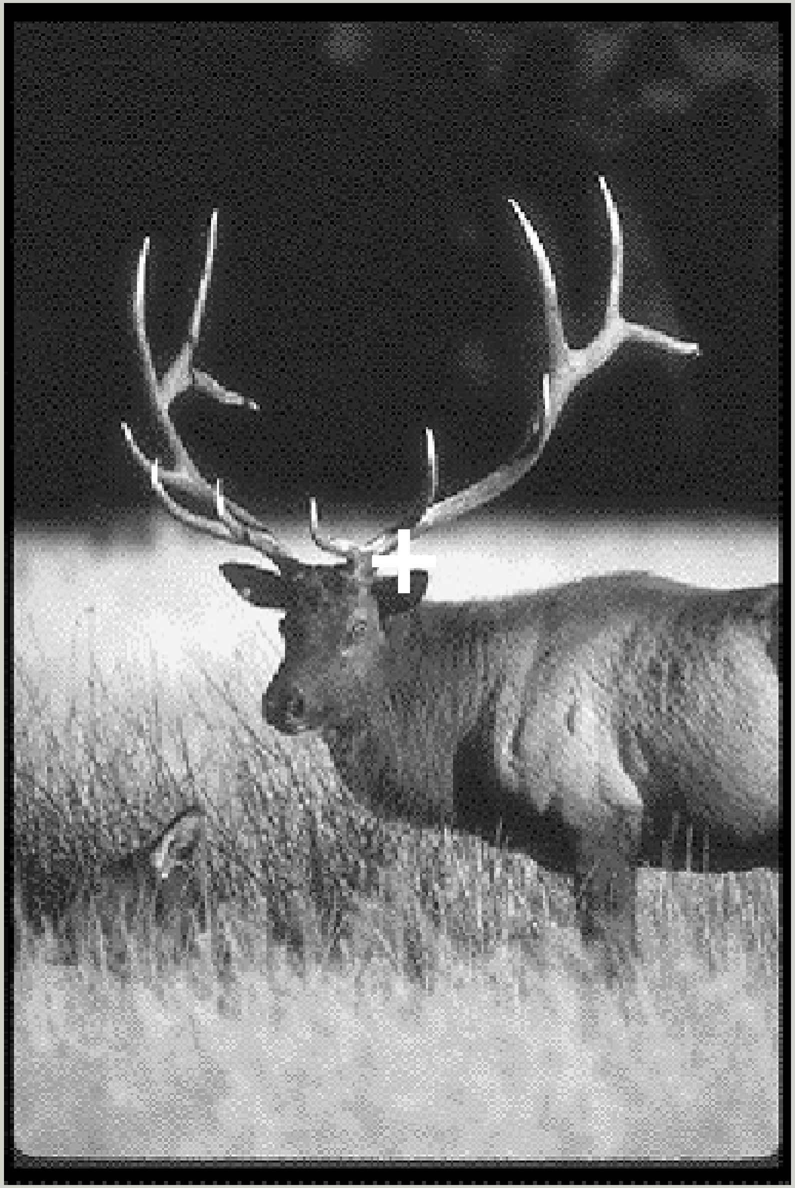
recently it provided a surprise (for me)…

…when we compared its performance with human vision on rapid categorization of complex natural images

…

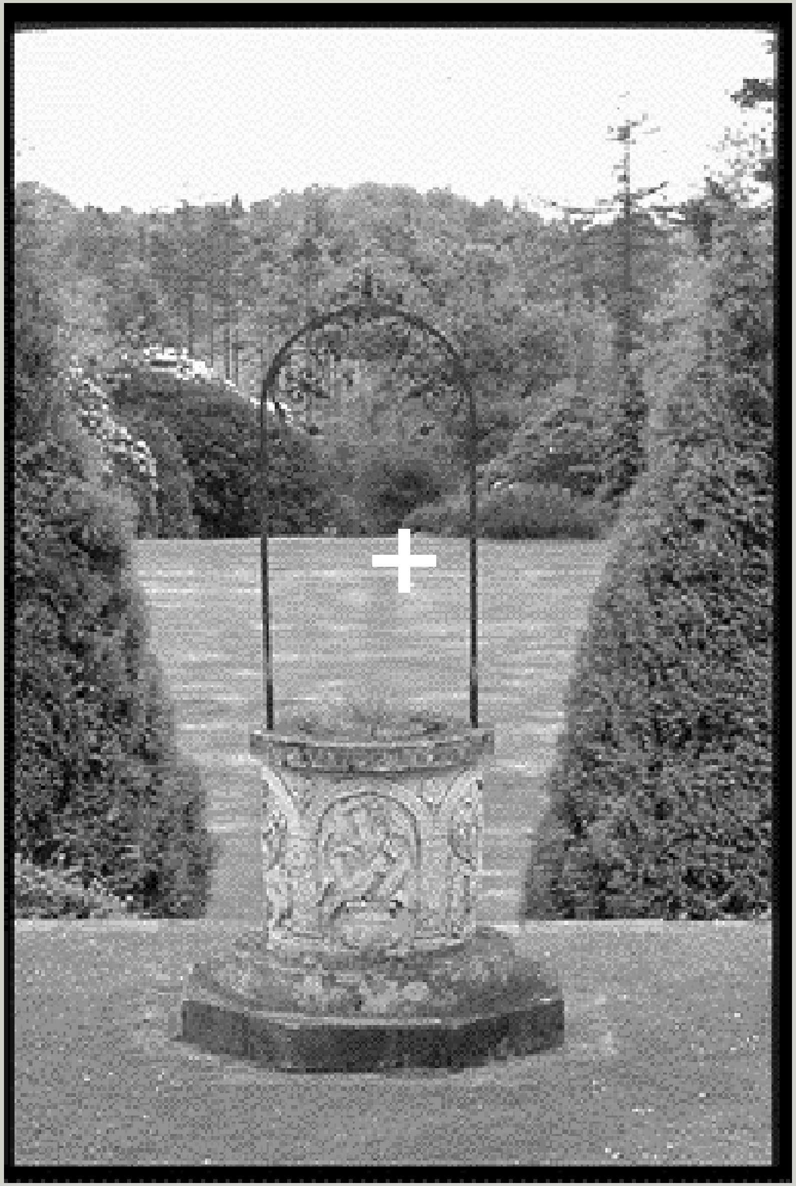
Serre, Oliva, Poggio, 2007

Rapid categorization

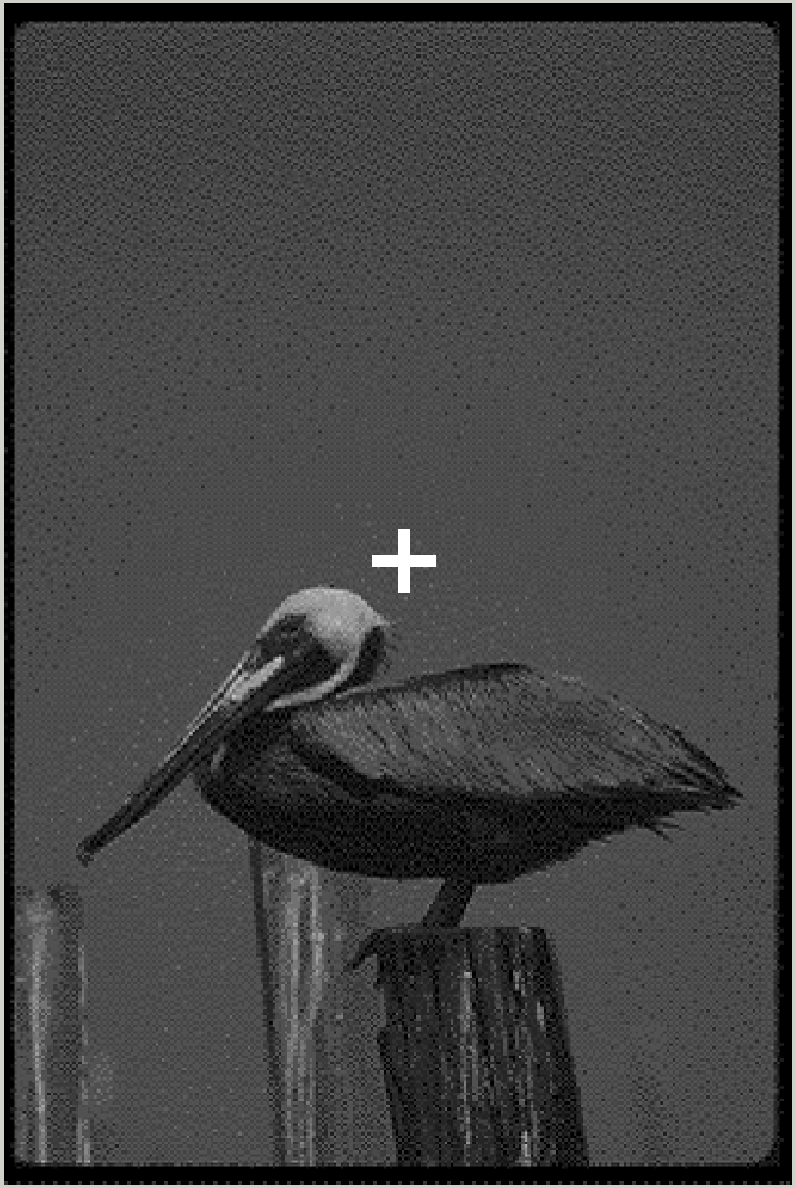


Source: Modified from Tim Masquellier, ECVP'05

Source: Modified from Tim Masquellier, ECVP'05



Source: Modified from Tim Masquellier, ECVP’05

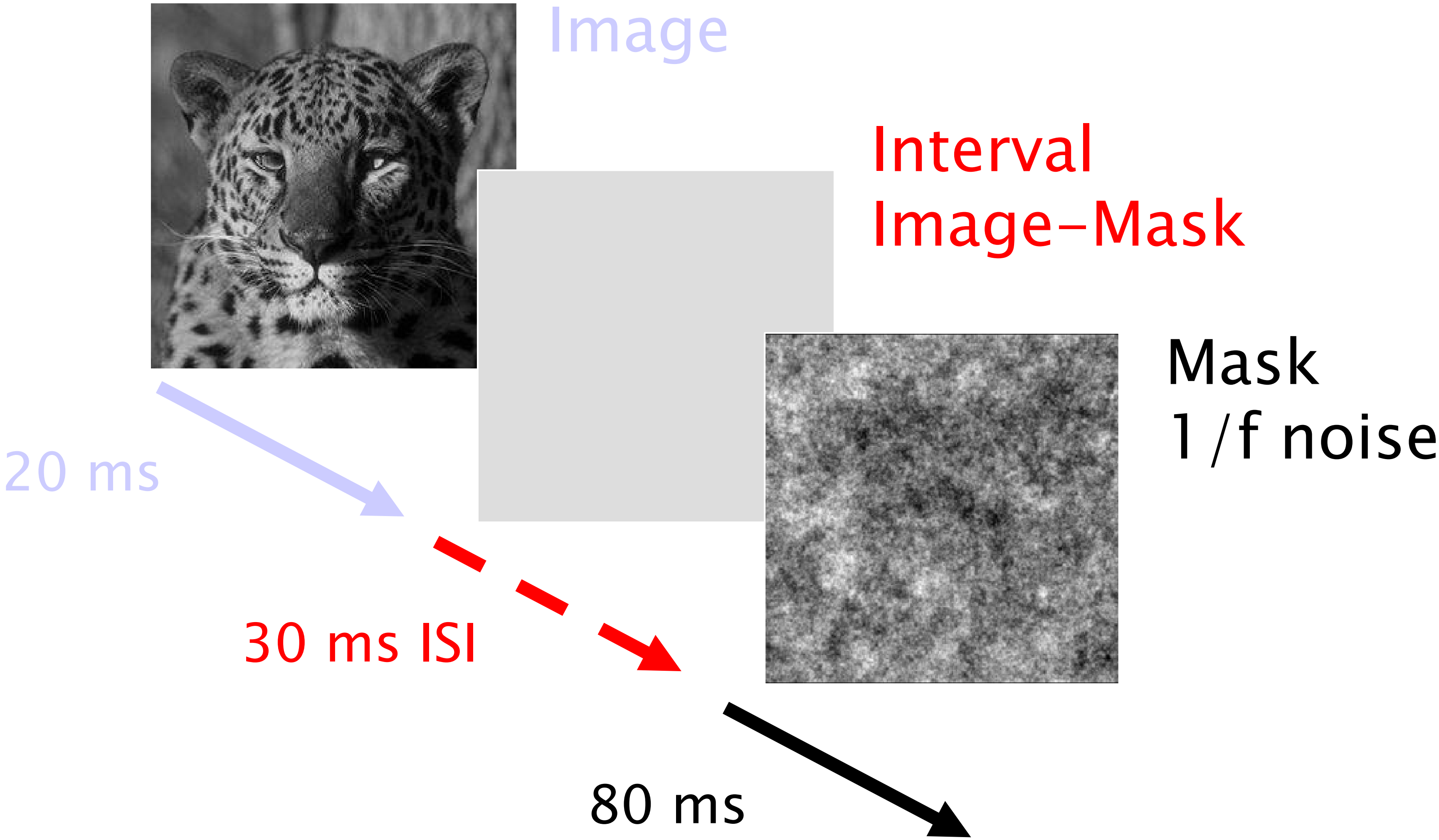


Source: Modified from Tim Masquellier, ECVP’05

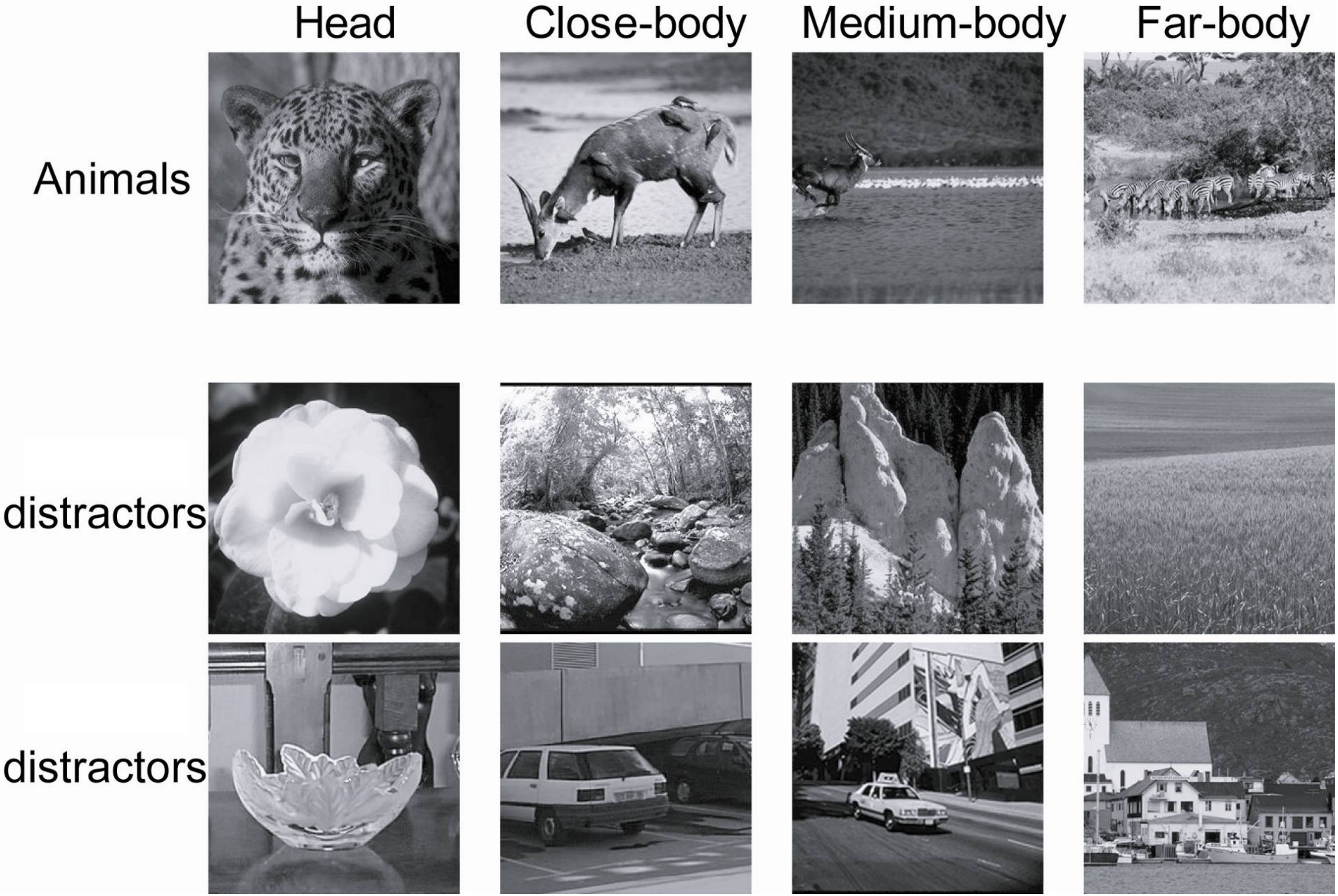
Source: Modified from Tim Masquellier, ECVP'05

Source: Modified from Tim Masquellier, ECVP'05

Rapid categorization task



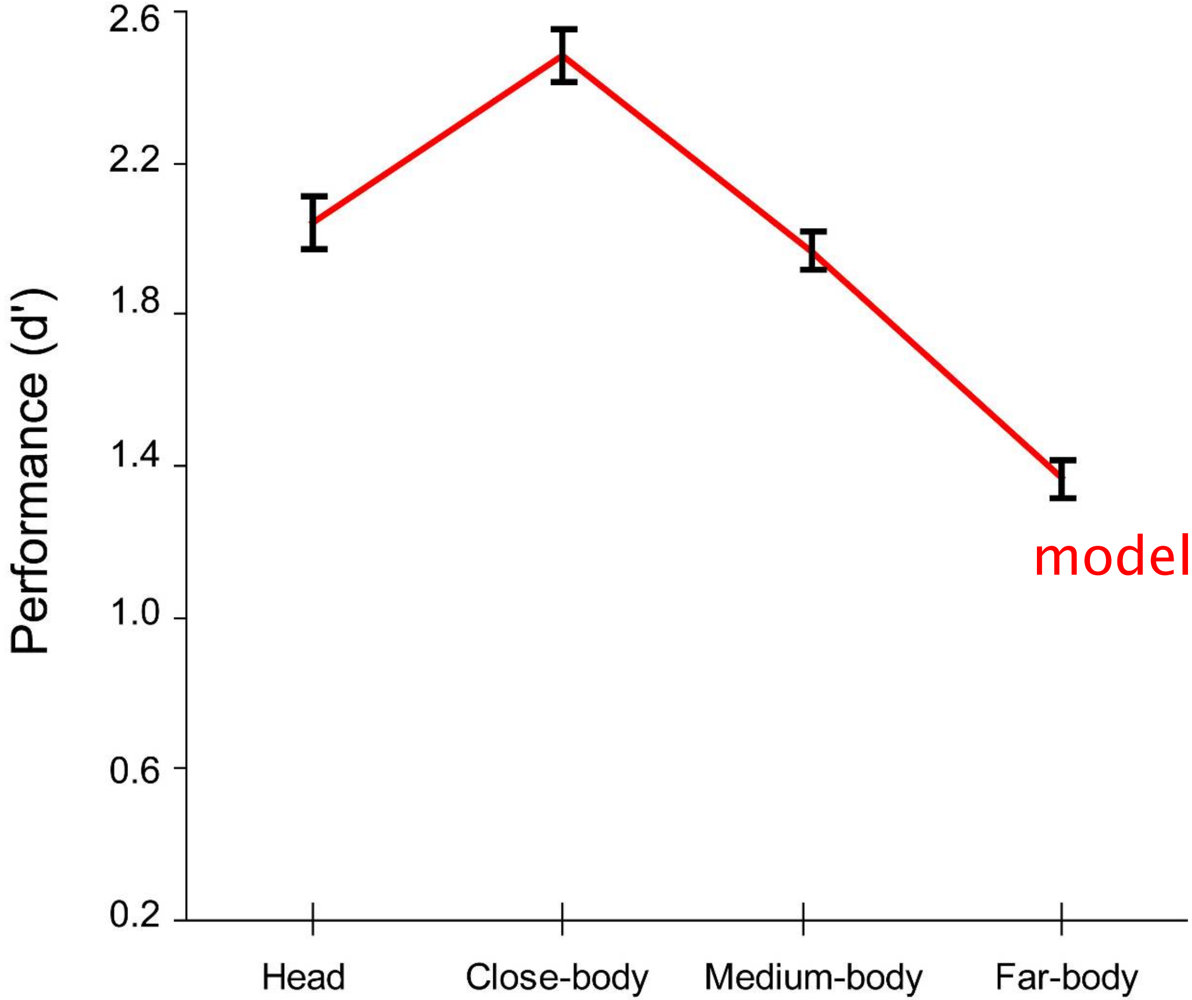
(Thorpe et al, 1996; Van Rullen & Koch, 2003;
Bacon-Mace et al, 2005; Oliva & Torralba, in press)



(Torralba & Oliva, 2003; Oliva & Torralba, in press)

The model predicts human perf.

- for SOA ~ 50ms



- d' ~ standardized error rate (the higher the d' the better the performance)

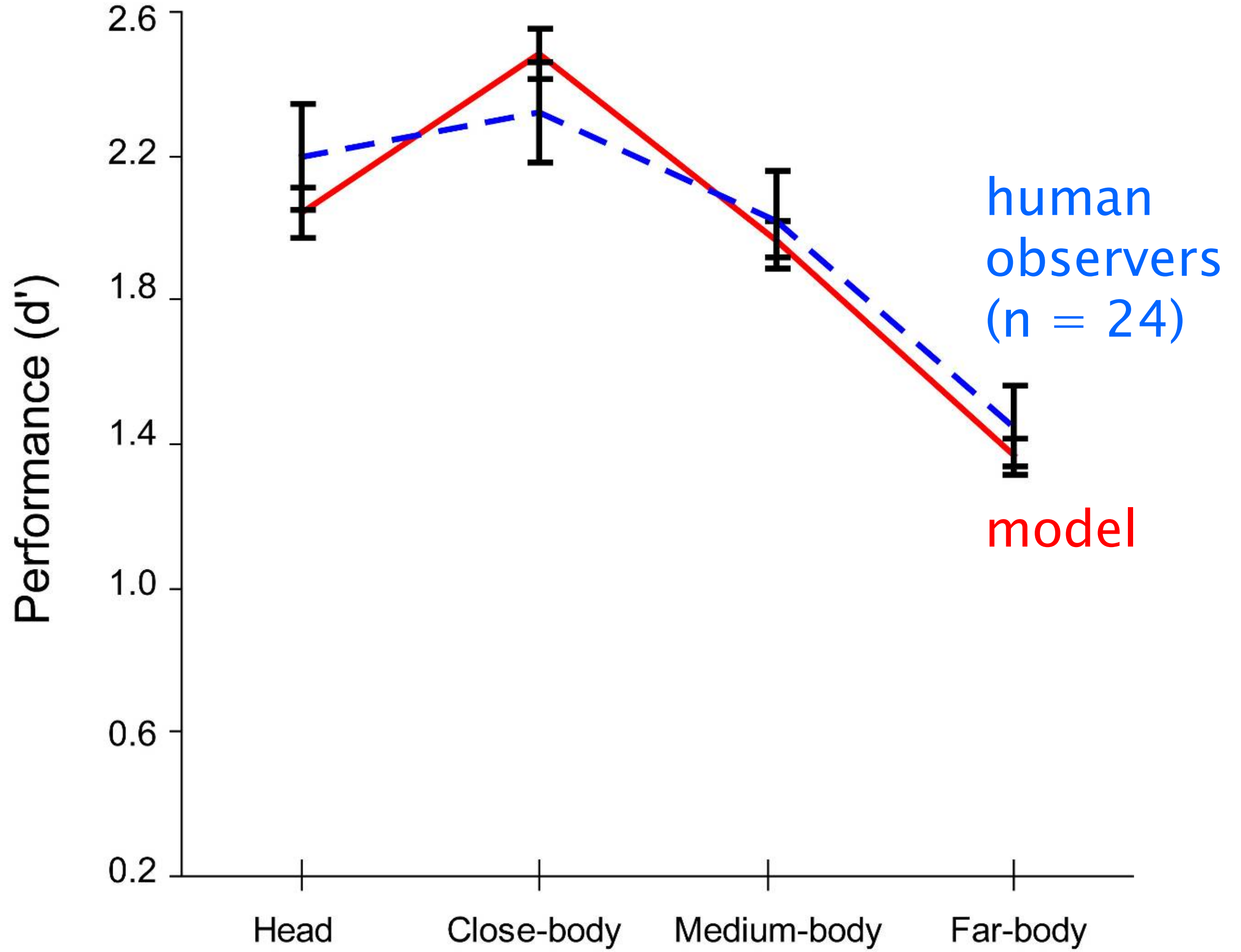
(Serre, Oliva and Poggio, 2007)

The model predicts human perf.

Model 82%

vs.

humans 80%



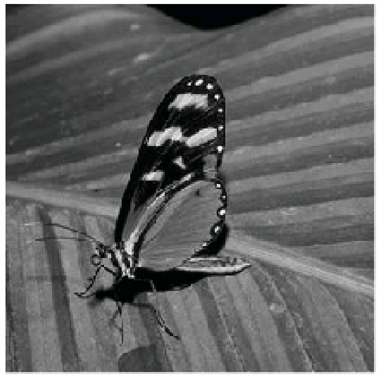
(Serre, Oliva and Poggio, in sub)

Similar correct and similar wrong answers

Some hits

Mod: 100% Hum: 96%

Mod: 91% Hum: 83%

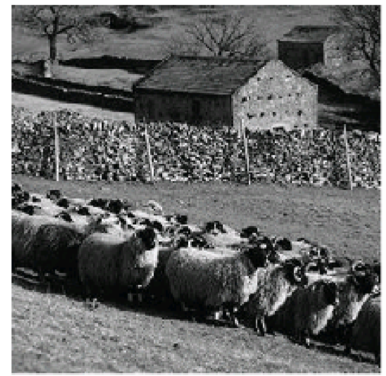


Mod: 100% Hum: 96%

Mod: 100% Hum: 91%

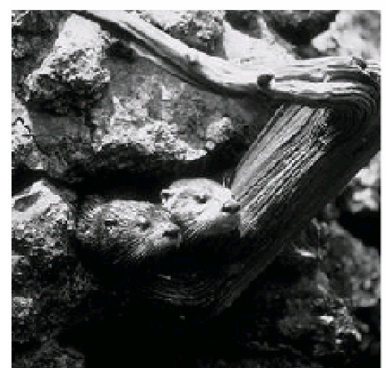
Some misses

Mod: 22% Hum: 21%



Mod: 0% Hum: 21%

Mod: 33% Hum: 21%



Mod: 0% Hum: 29%

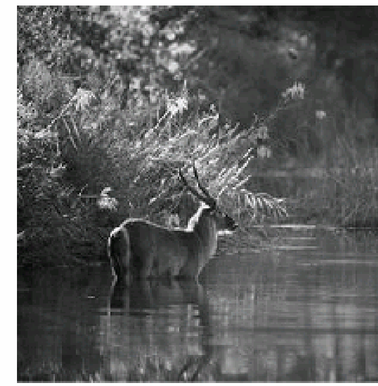
High correlation of correct answers and errors between humans and computer model: ρ~0.71

Mod: 40% Hum: 38%

Mod: 91% Hum: 33%

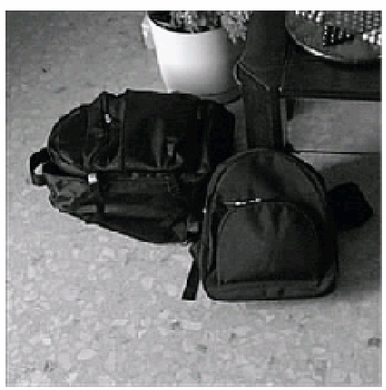
Mod: 20% Hum: 75%

Mod: 100% Hum: 25%



Mod: 44% Hum: 42%

Mod: 82% Hum: 63%



Mod: 10% Hum: 71%

Mod: 100% Hum: 29%

…another surprise…

… was that it works as well as the best machine vision systems…

The model performs at the level of the best computer vision systems on a variety of simple and difficult databases such as the SceneStreets database

Bileschi, Wolf, Serre and Poggio, PAMI, 2006

StreetScenes Database. Subjective Results

Results

Conclusions

We have a model describing how visual cortex computes. The model can:

- predict the neuronal properties in several cortical areas
- mimic human object recognition performance (no eye movements)
- perform better than many computer vision systems

Conclusions

Hierarchical architecture of (visual) cortex: why?

A challenge to (classical) learning theory:
a hierarchical architecture
with unsupervised and supervised learning
and learning of invariances…

and potential advantages for learning from very few supervised examples…

possibly because the organization of the visual world is hierarchical in space and time

High-level Musings

- ❑ We have a model of part of the ventral stream-- ***not*** *a theory* (as yet)!

- ❑ Is it possible that we will be able to replicate the brain without being able to develop a theory of it?

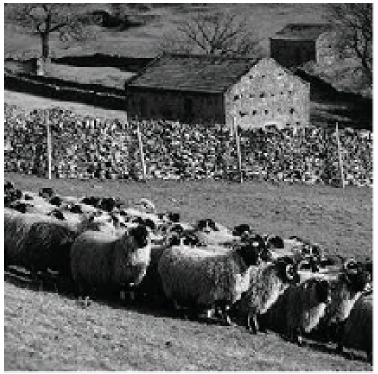
- ❑ Is it possible that understanding the brain will not go beyond simulations of a model of it?

Which ones are we going to get right or wrong?

A simulation can answer…but no simple intuition beforehand…

We need a theory!

Mod: 22% Hum: 21%



Mod: 0% Hum: 21%

Mod: 33% Hum: 21%

Mod: 0% Hum: 29%

Collaborators

- **T. Serre**
- S. Bileschi
- C. Cadieu
- U. Knoblich
- M. Kouh
- G. Kreiman
- M. Riesenhuber
- A. Oliva
- D. Ferster
- E. Connors
- J. Dicarlo
- S. Smale
- A. Caponnetto
- L. Rosasco

The end…