Class 21-22

Vision and Visual Neuroscience

Tomaso Poggio and Thomas Serre

Plan for class 21-22-23

- Class 21: Learning in the ventral stream of visual cortex: hierarchical models

Class 22: More on hierarchical models of recognition

- Class 23: Mathematical framework for hierarchical kernel machines: towards 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. The algorithms we have described can learn an object recognition task from a few thousand labeled images but a child, or even a monkey, can learn the same task from just a few examples. Thus an important area for future theoretical and experimental work is learning from partially labeled 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? It seems that the learning theory of the type we have outlined does not offer any general argument in favor of hierarchical learning machines for regression or classification.

❑ Why hierarchies? There may be reasons of *efficiency* – computational speed and use of computational resources. For instance, the lowest levels of the hierarchy may represent a dictionary of features that can be shared across multiple classification tasks.

❑ There may also be the more fundamental issue of *sample complexity*. Learning theory shows that the difficulty of a learning task depends on the size of the required hypothesis space. This complexity determines in turn how many training examples are needed to achieve a given level of generalization error. Thus our ability of learning from just a few examples, and its limitations, may be related to the hierarchical architecture of cortex.

Classical Learning Theory and Kernel Machines (Regularization in RKHS)

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

implies

$$f(\mathbf{x}) = \sum_i^l \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

Equation includes splines, Radial Basis Functions and SVMs (depending on choice of V).

*For a review, see Poggio and Smale, **The Mathematics of Learning**, Notices of the AMS, 2003; see also Schoelkopf and Smola, 2002; Bousquet, O., S. Boucheron and G. Lugosi.*

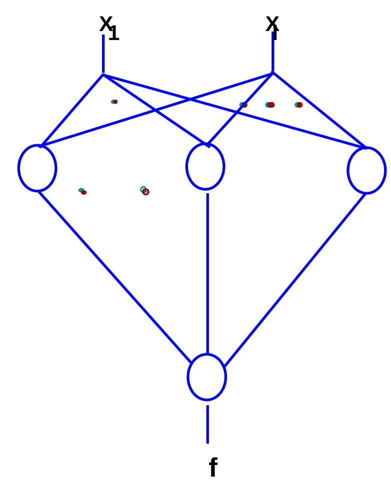
Classical Learning Theory and Kernel Machines (Regularization in RKHS)

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

implies

$$f(\mathbf{x}) = \sum_i^l \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

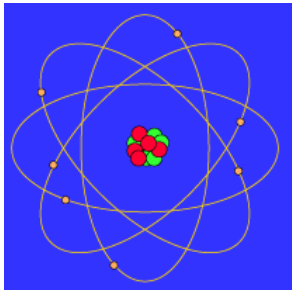
Kernel machines correspond to
shallow networks unlike cortex…



This class:

using a class of models to summarize/interpret experimental results…with *caveats*:

- Models are cartoons of reality, eg Bohr's model of the hydrogen atom

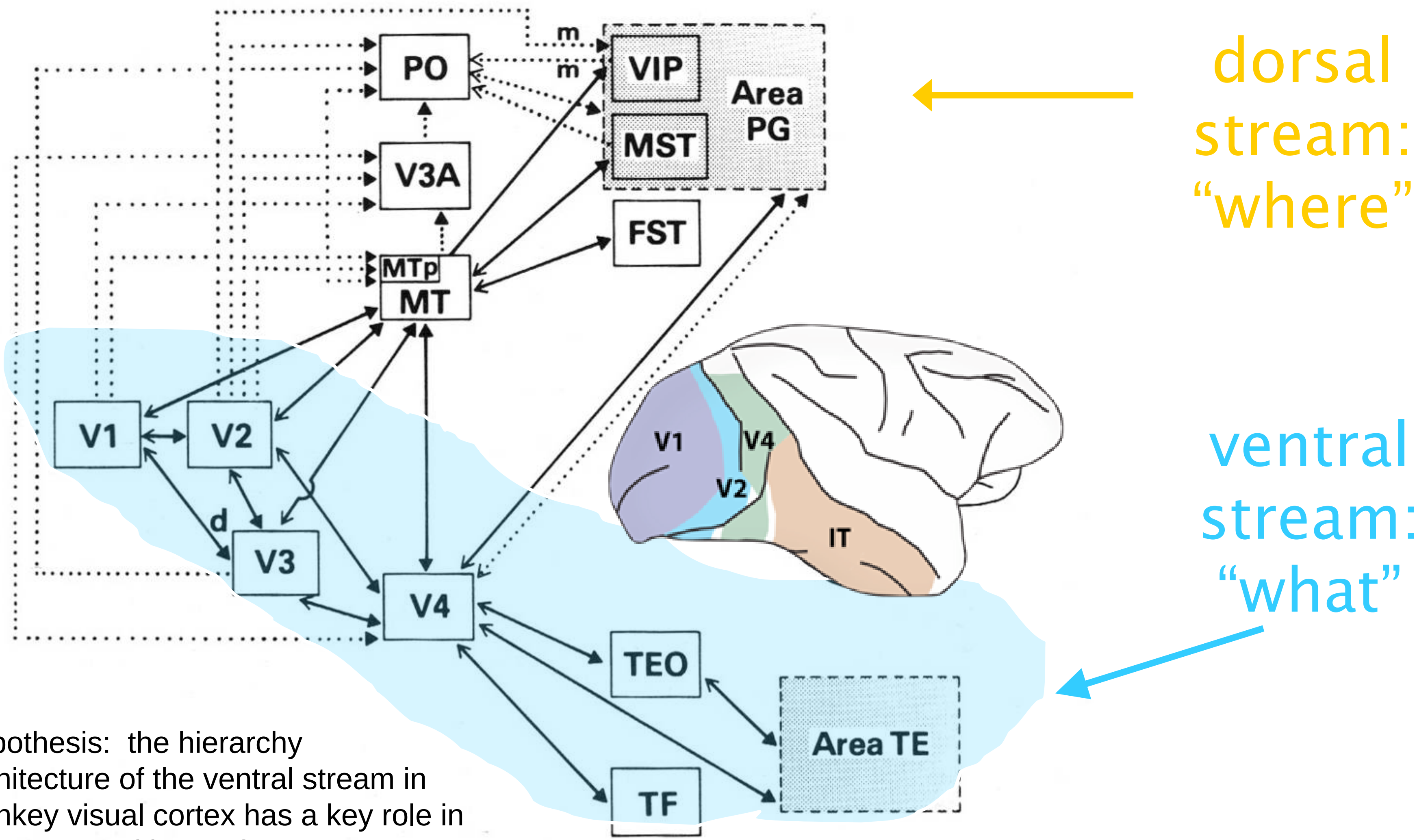


- All models are "wrong"
- Some models can be useful summaries of data and some can be a good starting point for more complete theories

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models

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

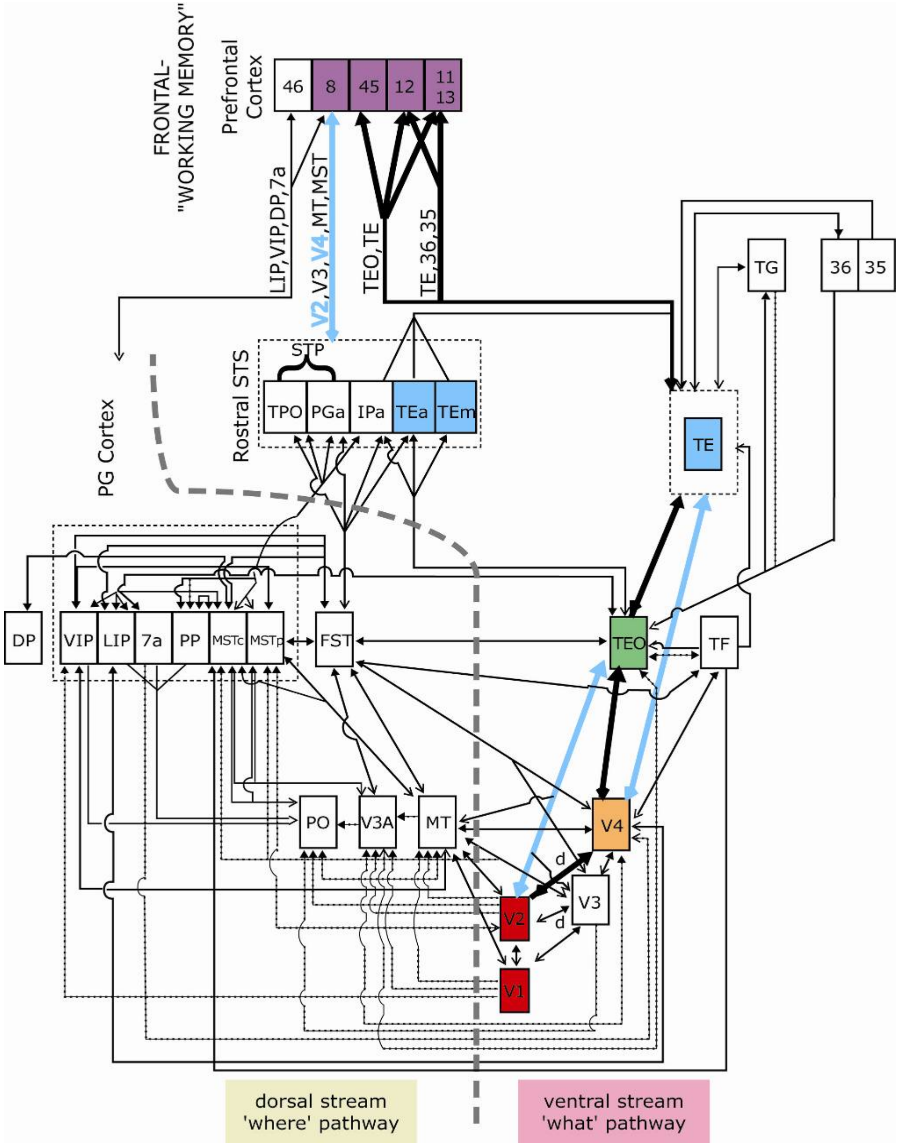
Object Recognition and the Ventral Stream



Hypothesis: the hierarchy architecture of the ventral stream in monkey visual cortex has a key role in object recognition…of course subcortical pathways may also be important (thalamus, in particular pulvinar…).

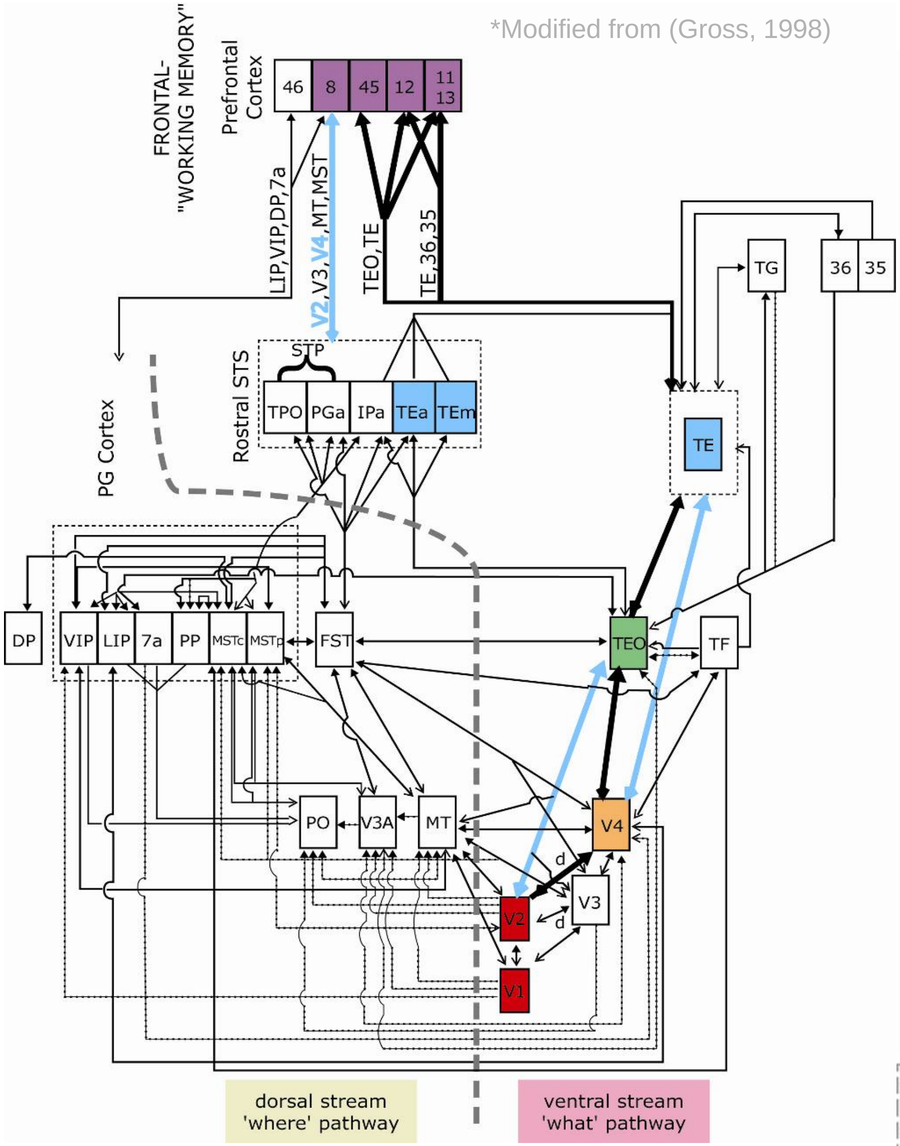
Desimone & Ungerleider 1989

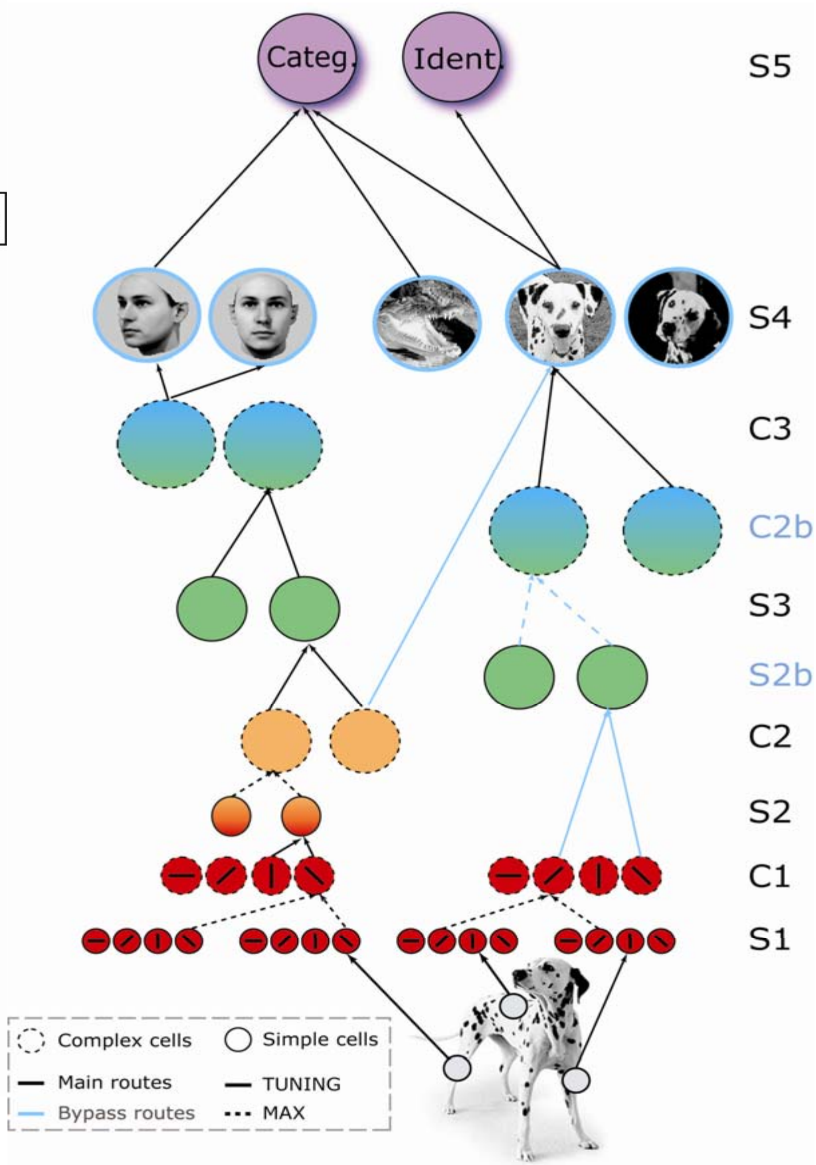
The ventral stream



Feedforward connections only?

A model of the ventral stream, which is also a hierarchical algorithm…





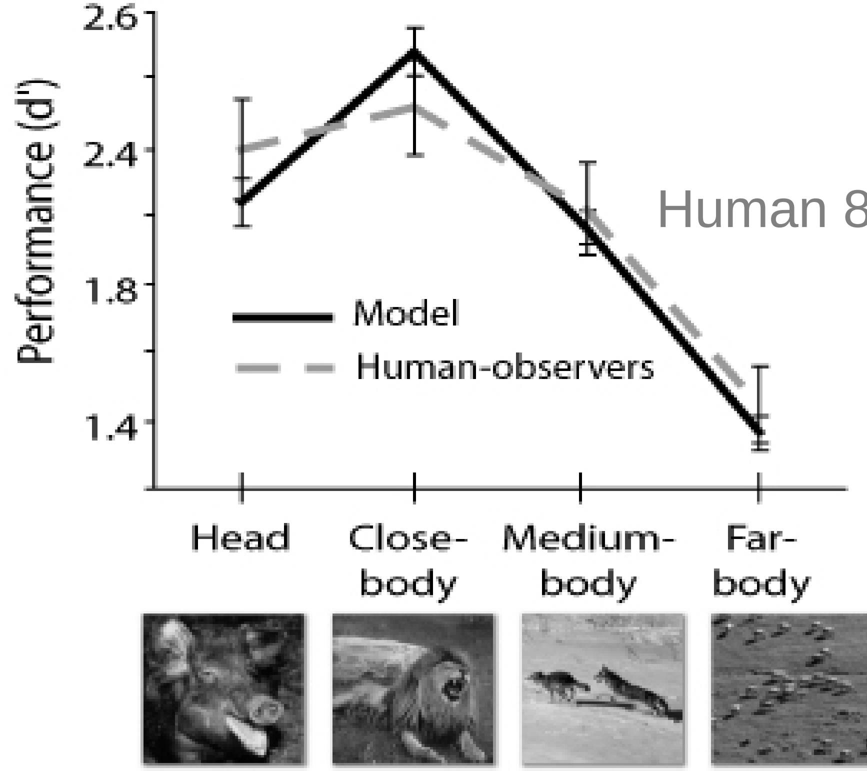
Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

[software available online]

..."solves" the problem
(if the mask forces feedforward processing)...

- d'~ standardized error rate
- the higher the d', the better the performance

Model 82%



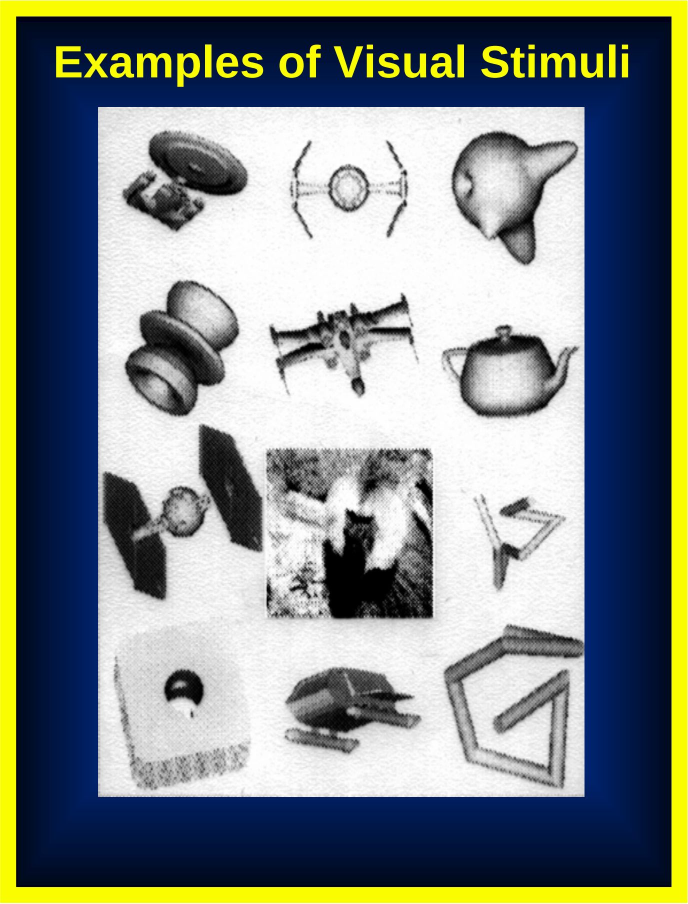
Human 80%

Serre Oliva & Poggio 2007

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models

Some personal history:

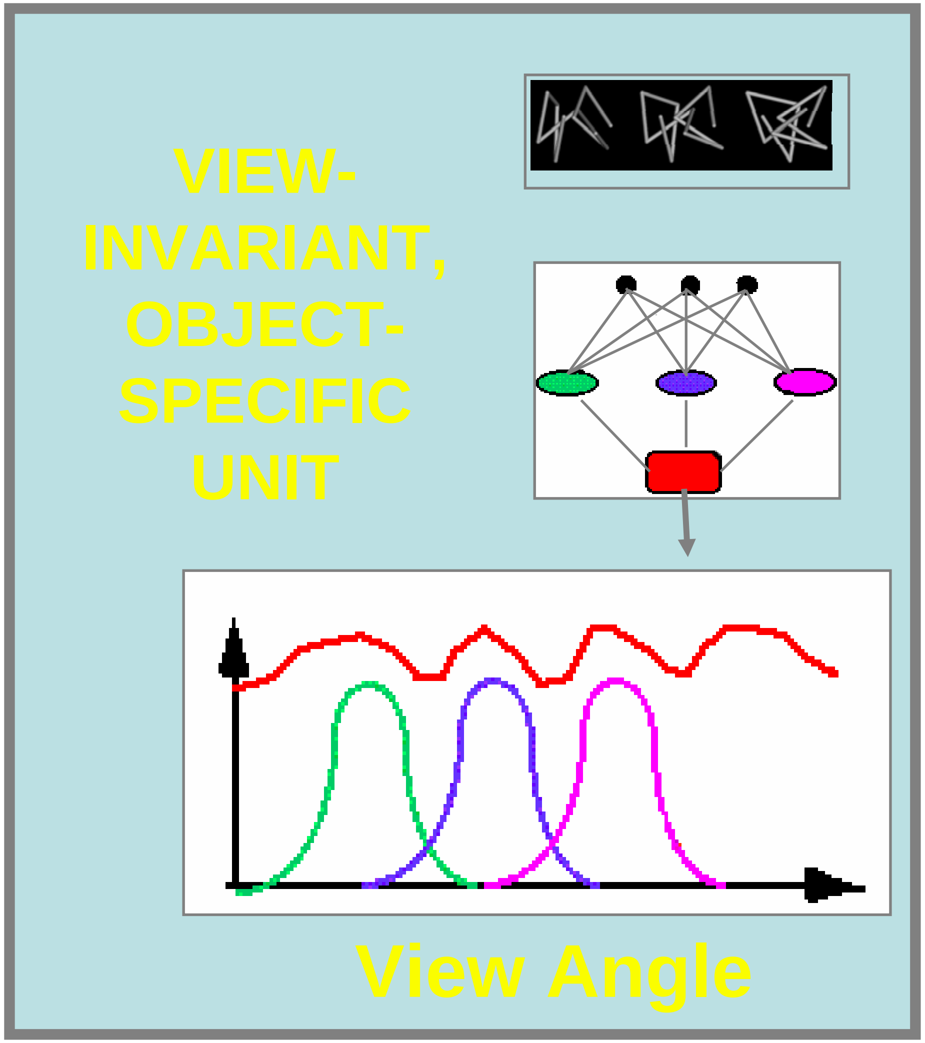
First step in developing a model: learning to recognize 3D objects in IT cortex



Poggio & Edelman 1990

An idea for a module for view-invariant identification

Architecture that accounts for invariances to 3D effects (>1 view needed to learn!)

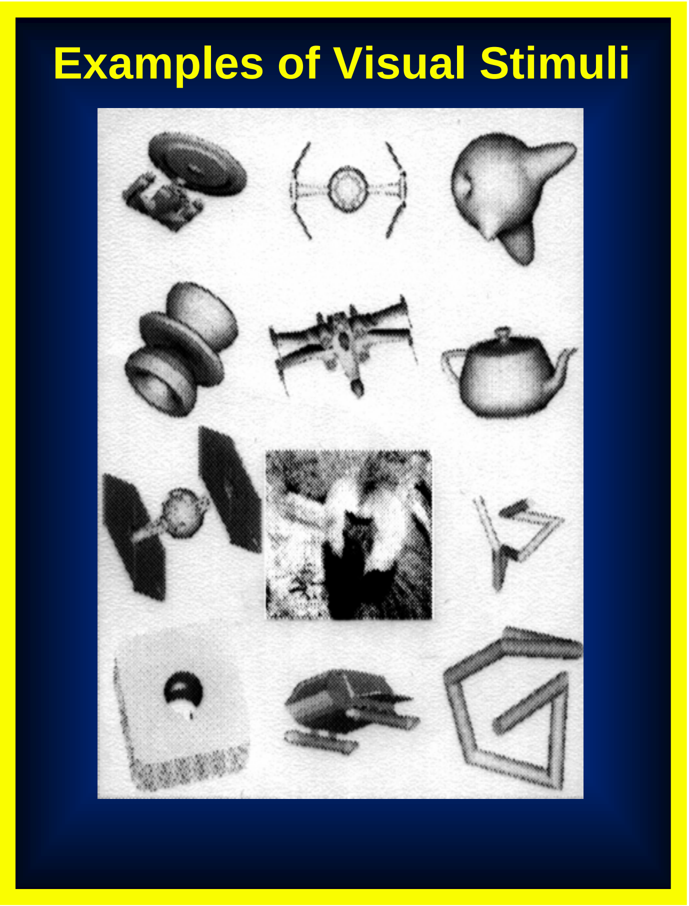


Prediction: neurons become view-tuned through learning

Regularization Network (GRBF) with Gaussian kernels

Poggio & Edelman 1990

Learning to Recognize 3D Objects in IT Cortex

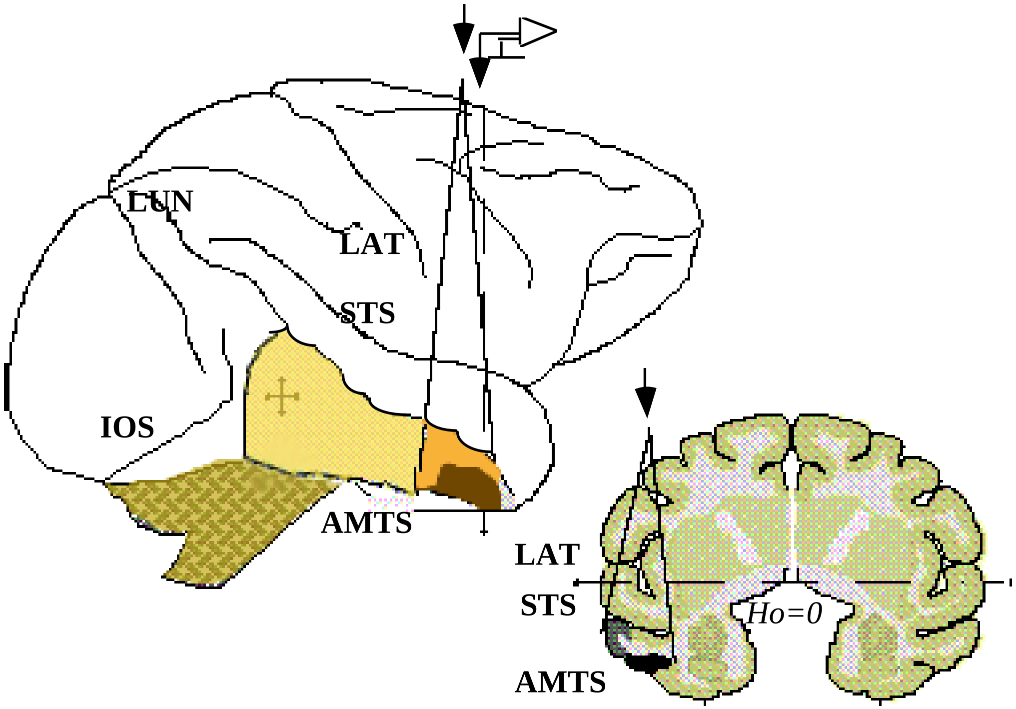


After human psychophysics (Buelthoff, Edelman, Tarr, Sinha, …), which supports models based on view-tuned units...

… physiology!

Logothetis Pauls & Poggio 1995

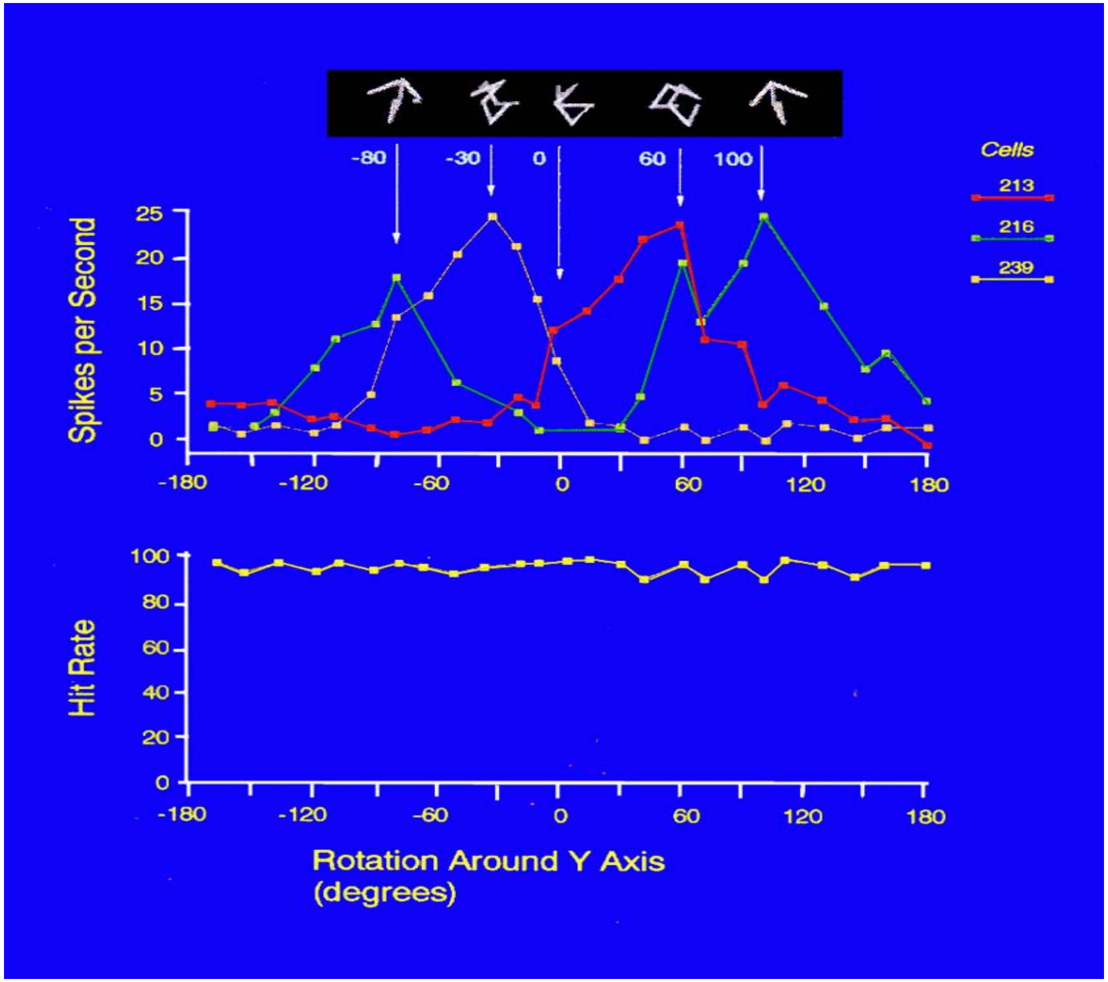
Recording Sites in Anterior IT



…neurons tuned to faces are intermingled nearby….

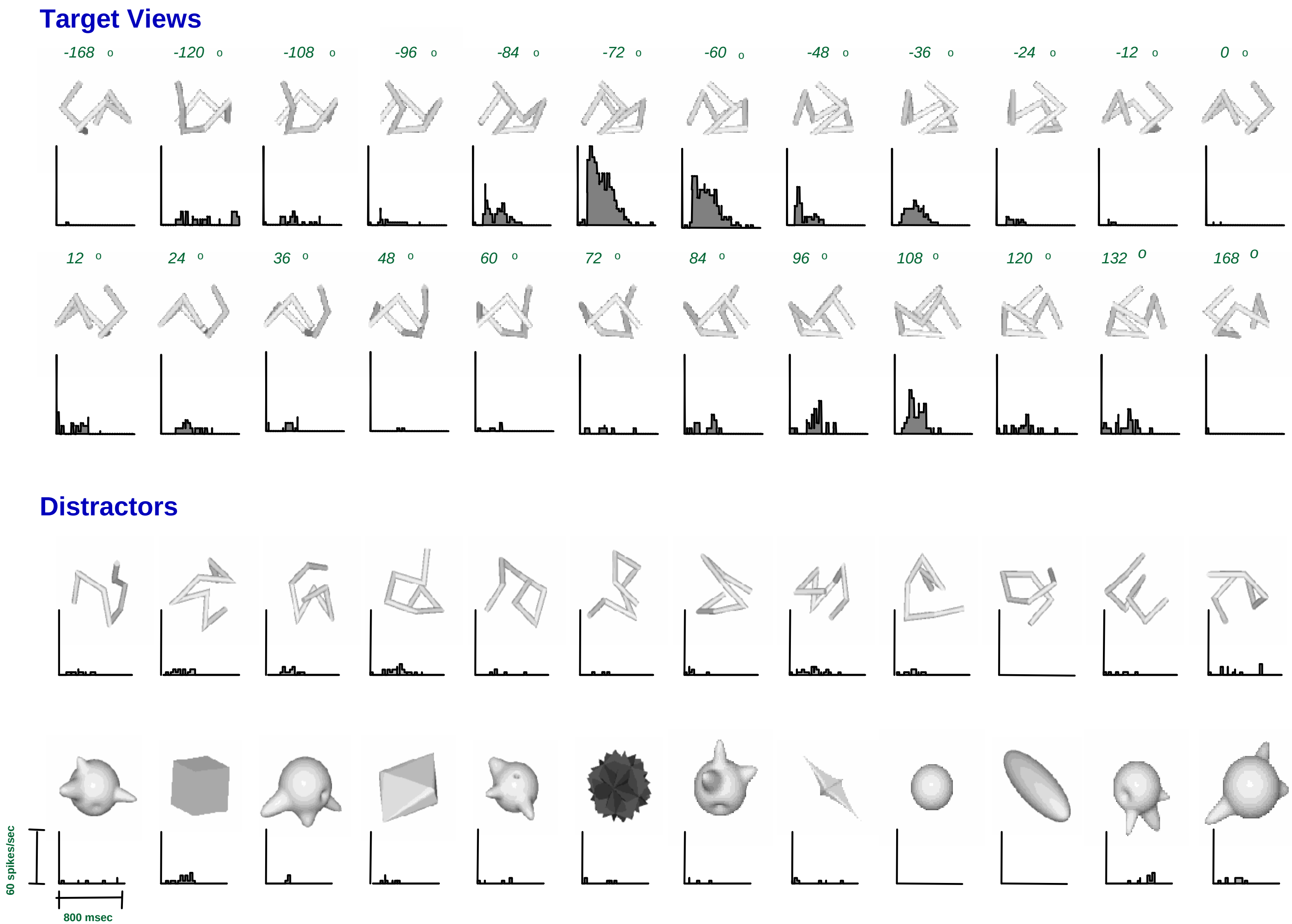
Logothetis, Pauls & Poggio 1995

Neurons tuned to object views, as predicted by model!



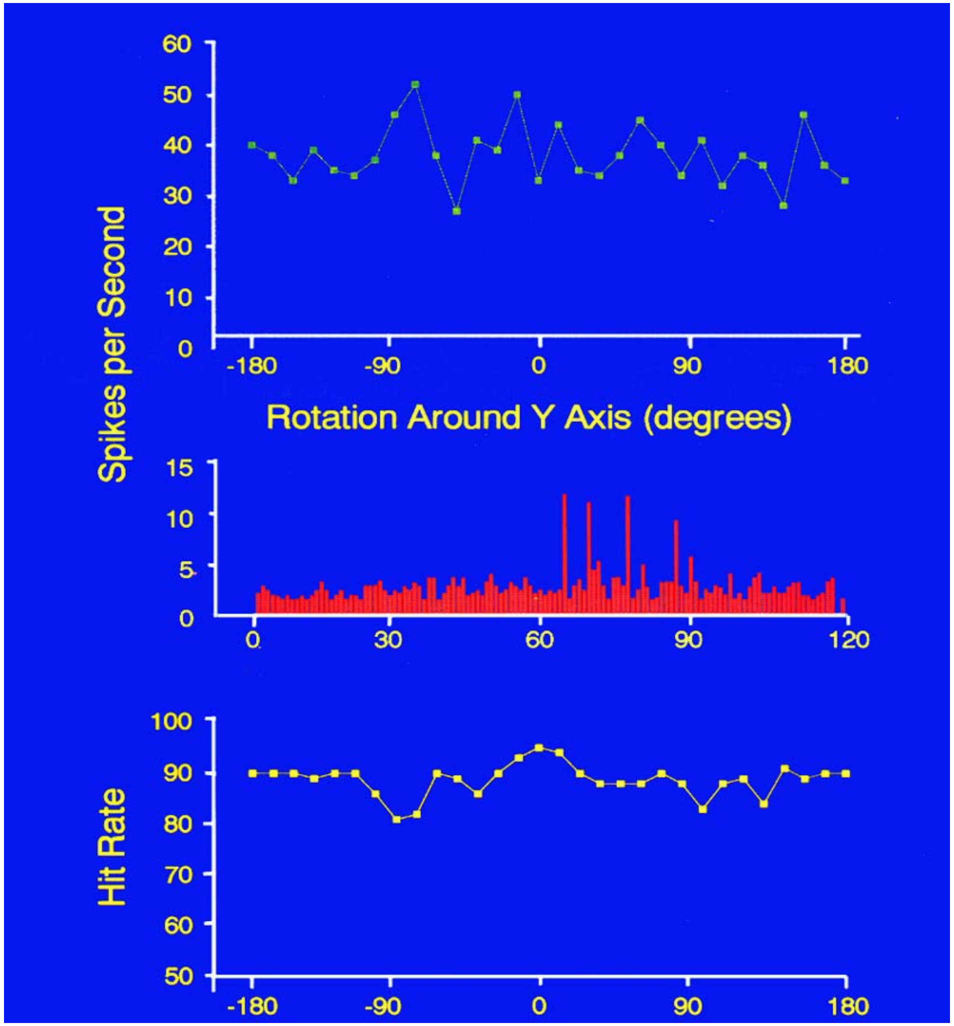
Logothetis Pauls & Poggio 1995

A “View-Tuned” IT Cell



Logothetis Pauls & Poggio 1995

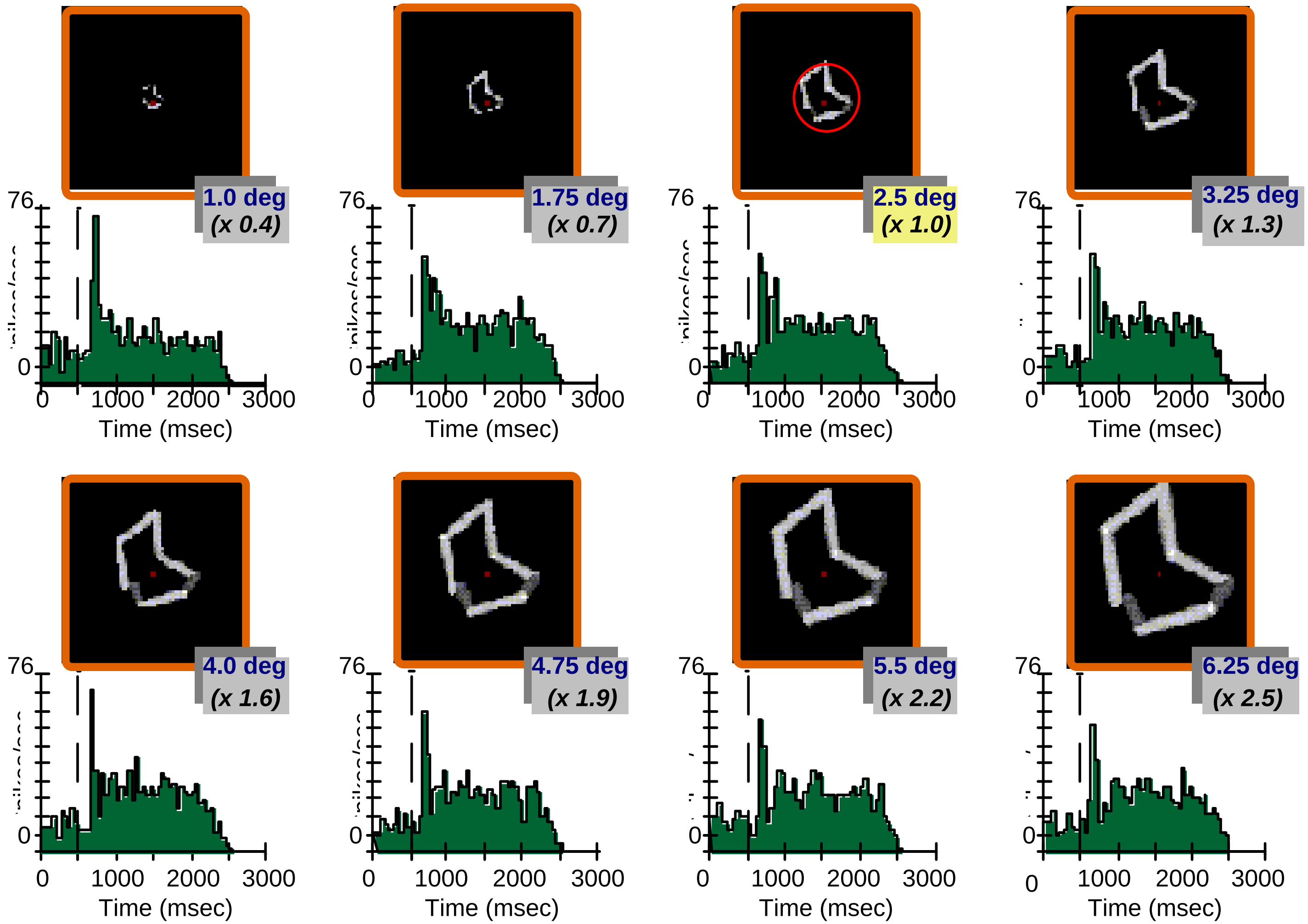
But also view-invariant object-specific neurons (5 of them over 1000 recordings)



Logothetis Pauls & Poggio 1995

View-tuned cells:

scale invariance (one training view only) motivates present model



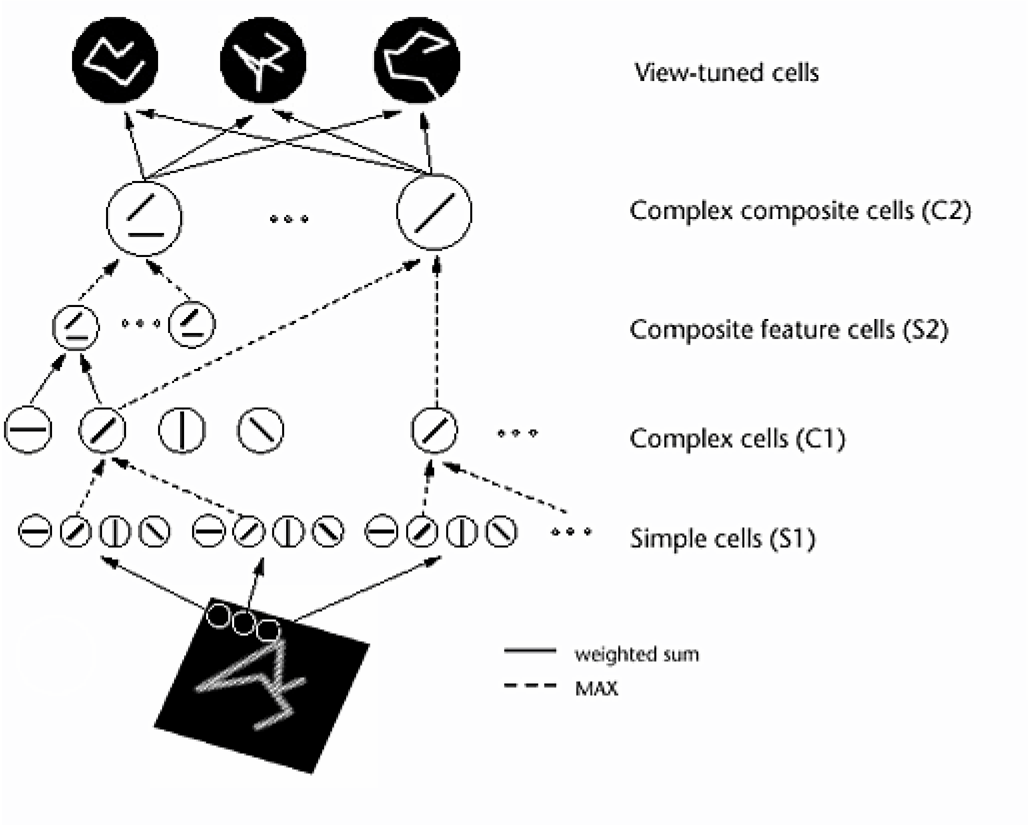
Logothetis Pauls & Poggio 1995

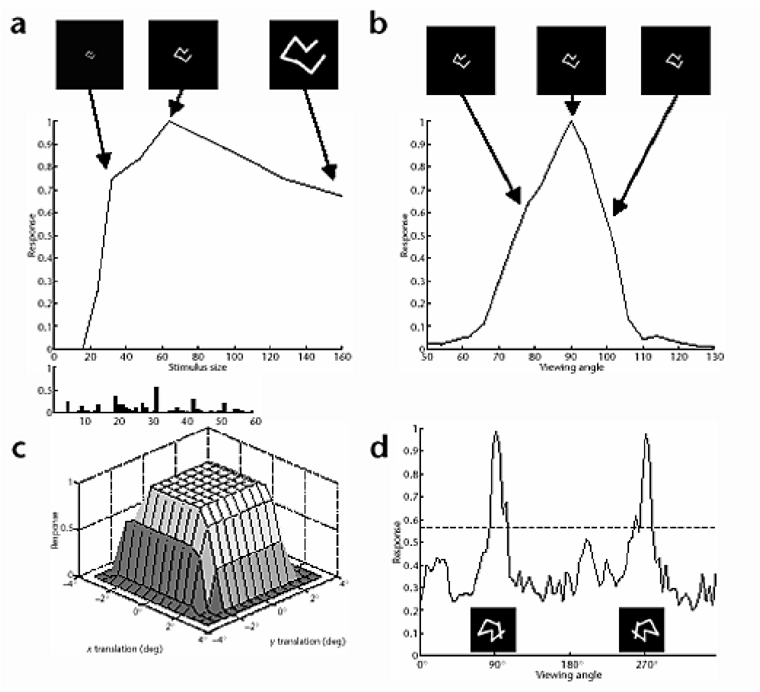
Hierarchy

- Gaussian centers (Gaussian Kernels) tuned to complex multidimensional features as composition of lower dimensional Gaussian
- What about tolerance to position and scale?

- Answer: hierarchy of invariance and tuning operations

The "HMAX" model

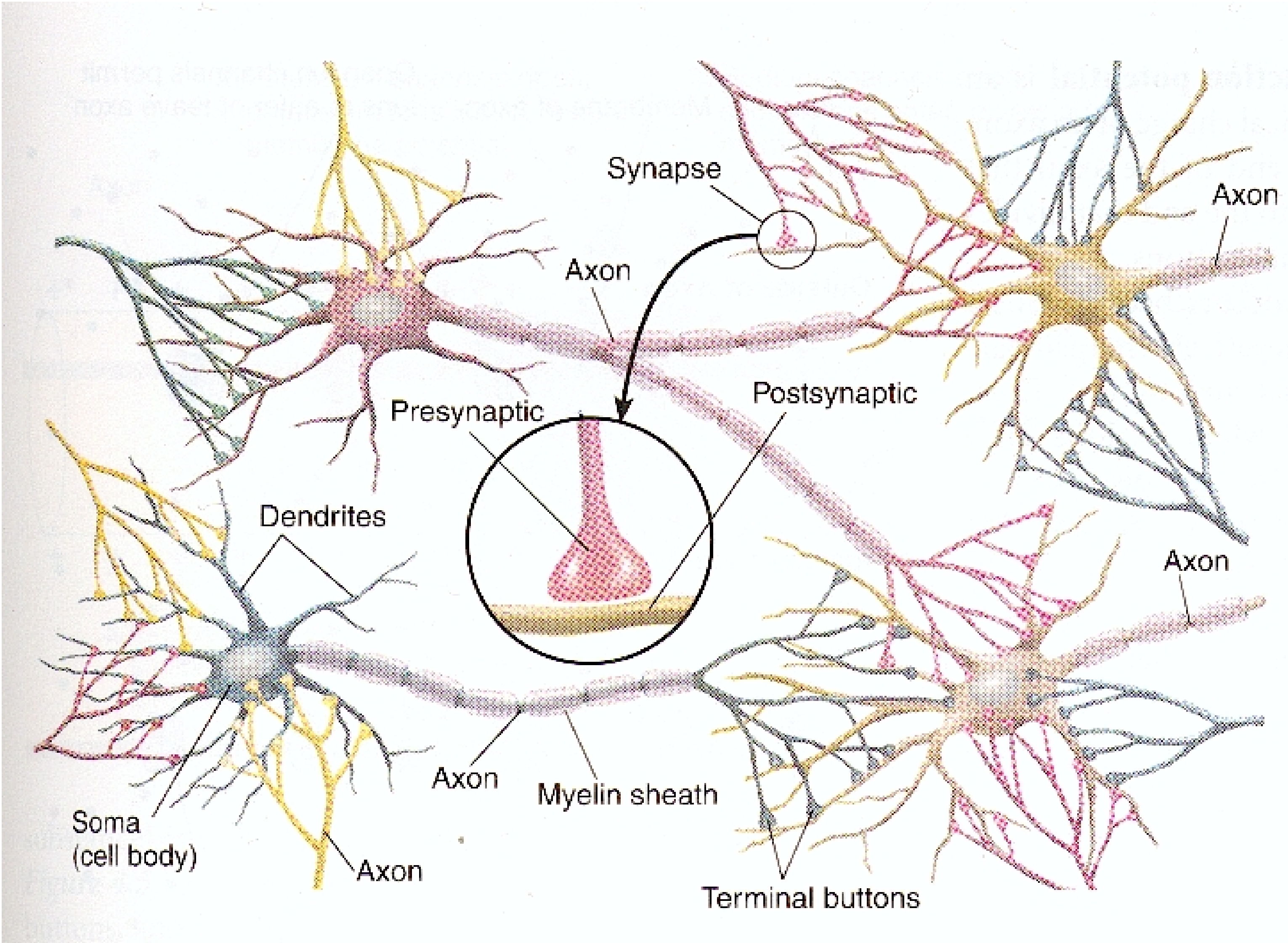




Riesenhuber & Poggio 1999, 2000

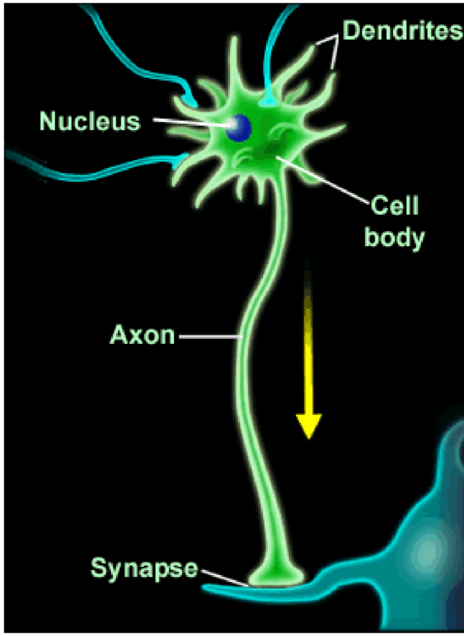
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models

Neural Circuits



Source: Modified from Jody Culham's web slides

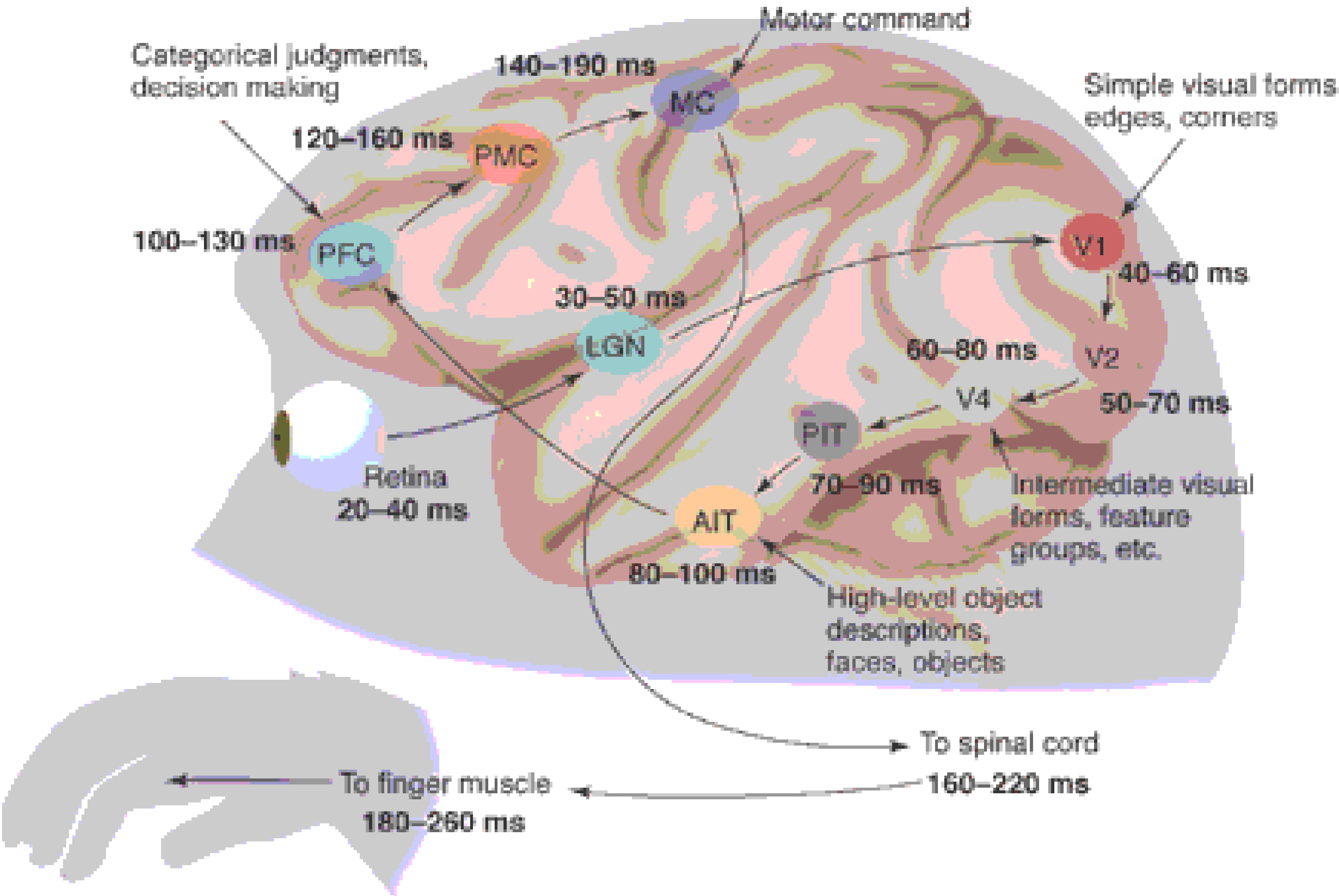
Object Recognition and the Ventral Stream



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

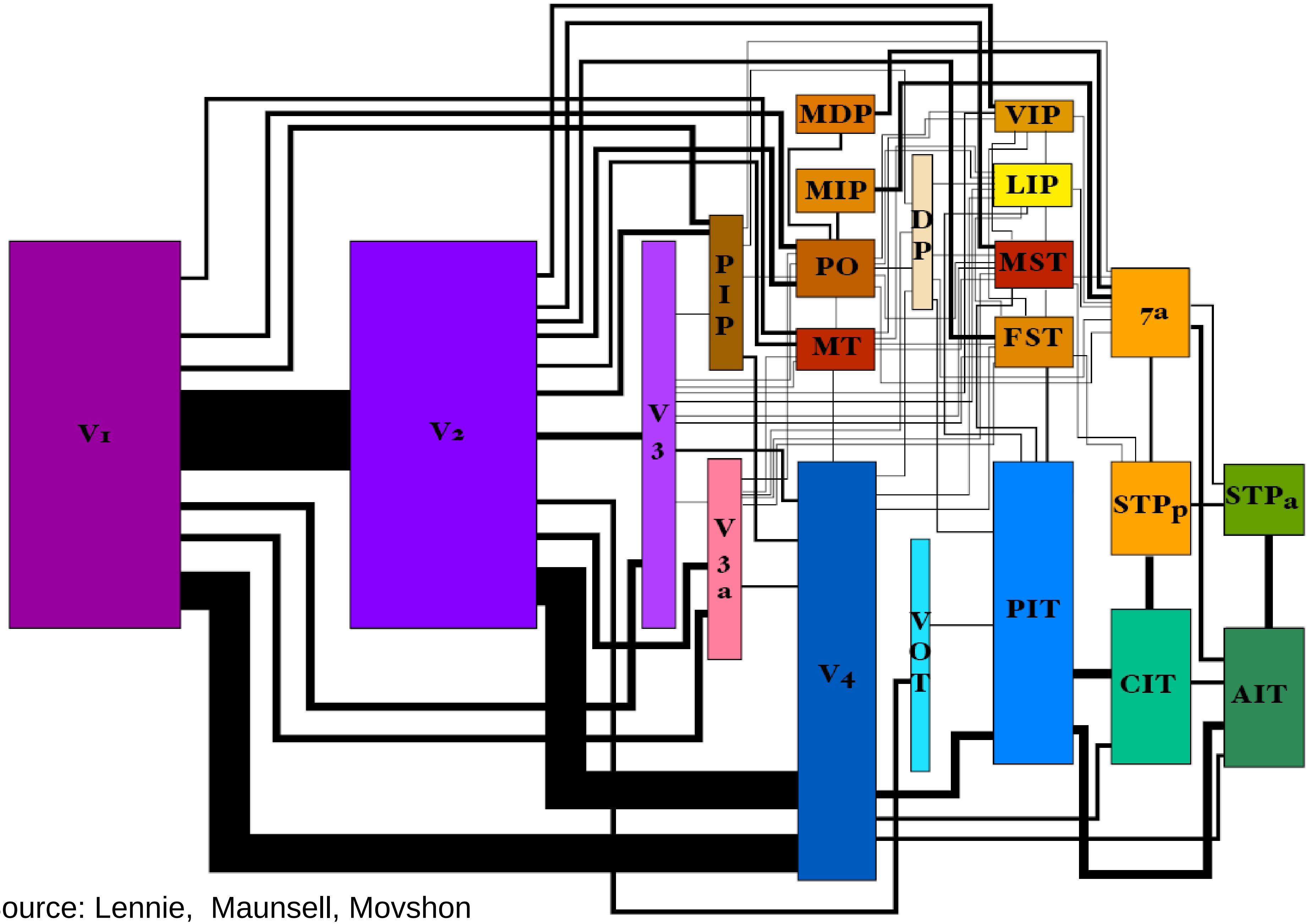
- Ventral stream in rhesus monkey
  - 10^{9} neurons
  - 5 10^{6} neurons in AIT

- 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



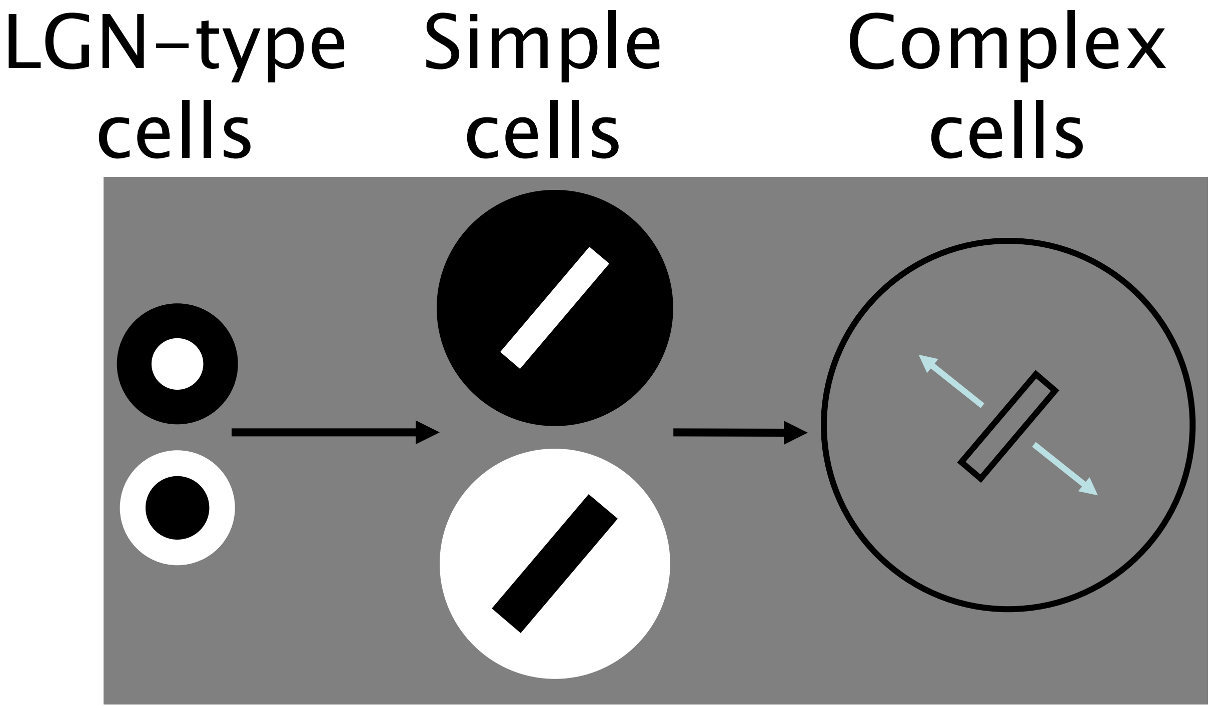
(Thorpe and Fabre-Thorpe, 2001)

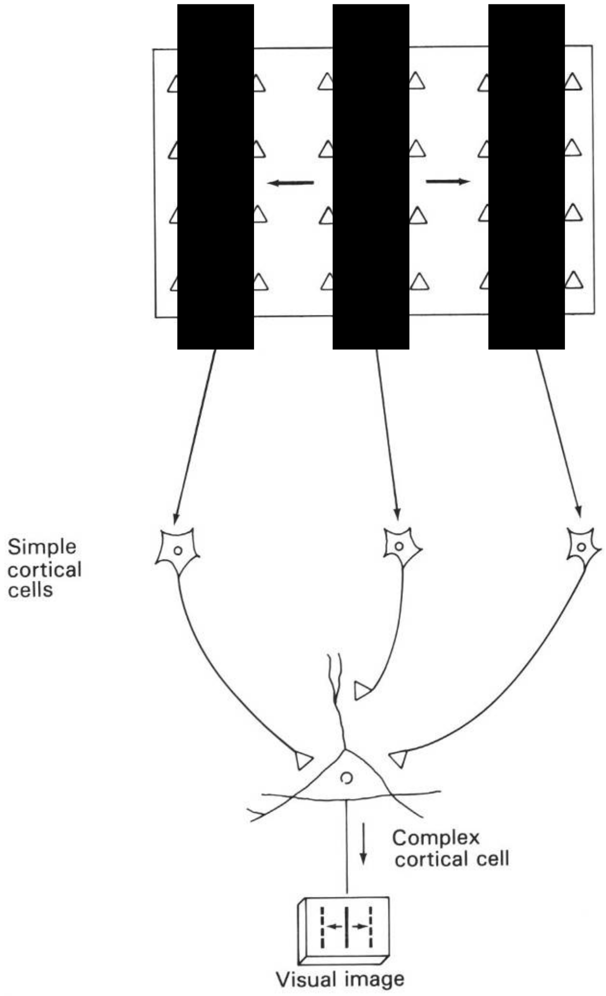
The ventral stream

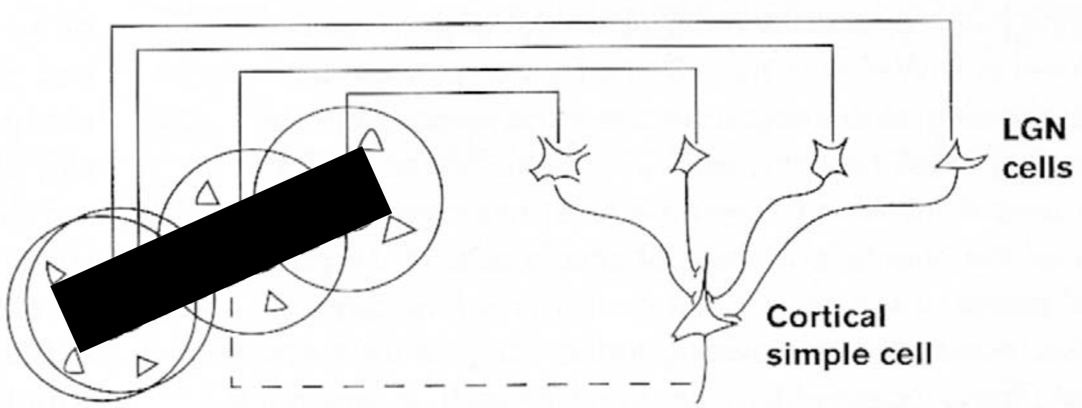


Source: Lennie, Maunsell, Movshon

V1: hierarchy of simple and complex cells

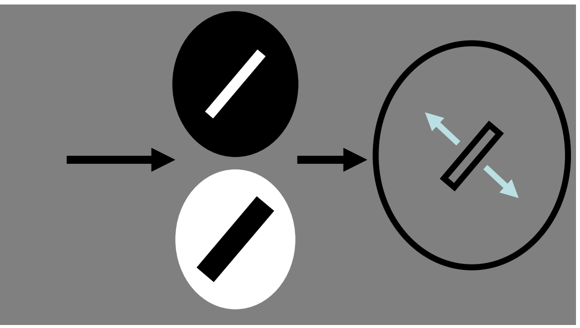


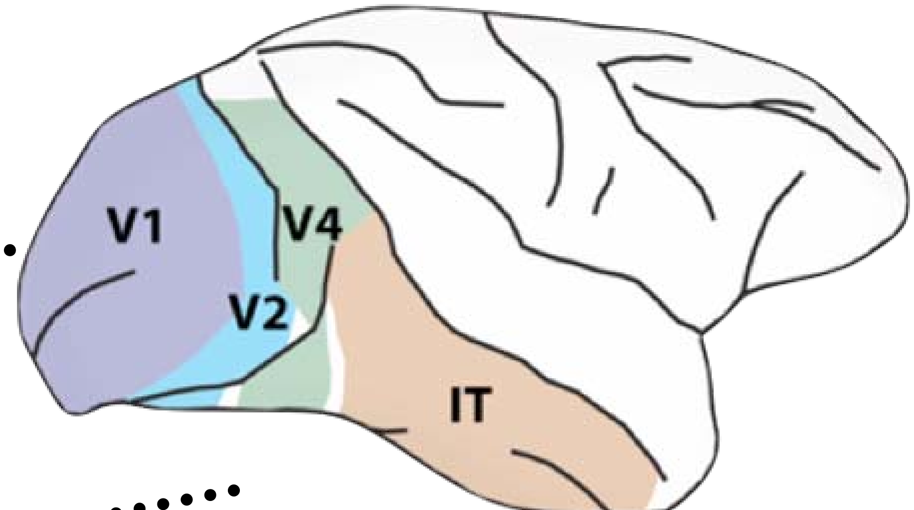


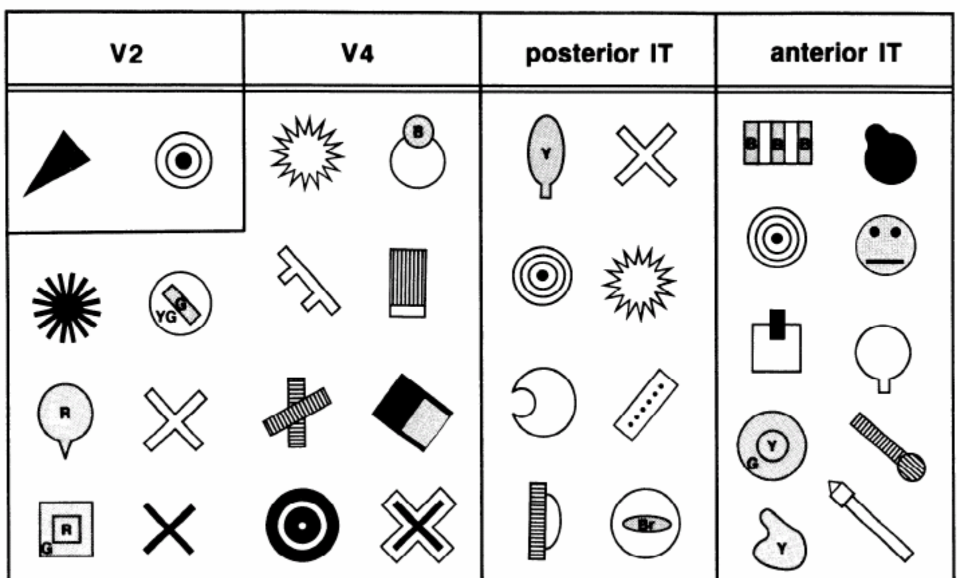


(Hubel & Wiesel 1959)

The Ventral Stream Hierarchy: V1, V2, V4, IT





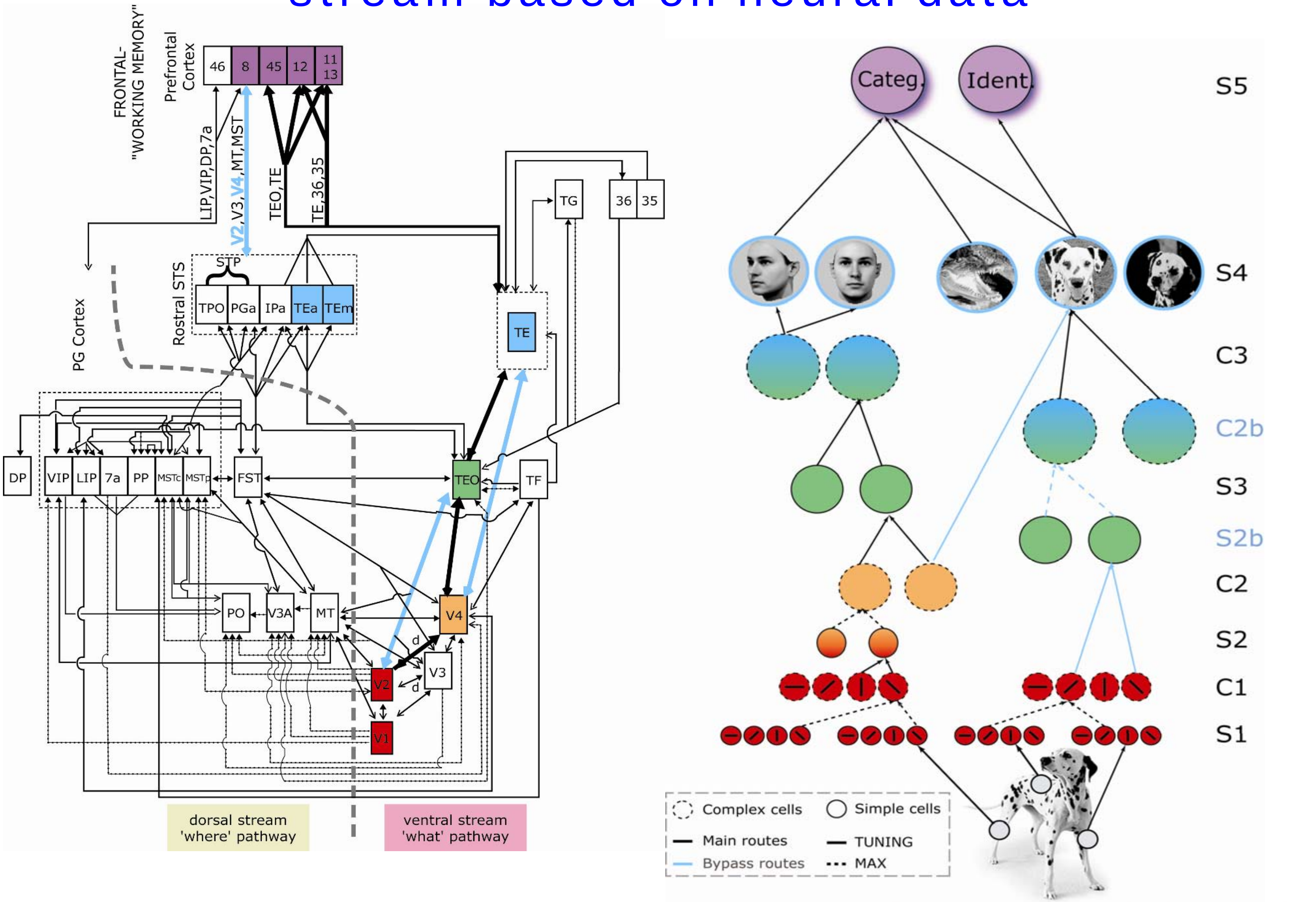


A gradual increase in the receptive field size, in the complexity of the preferred stimulus, in tolerance to position and scale changes

Kobatake & Tanaka, 1994

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models

A hierarchical feedforward model of the ventral stream based on neural data

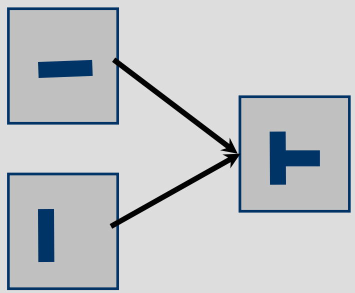
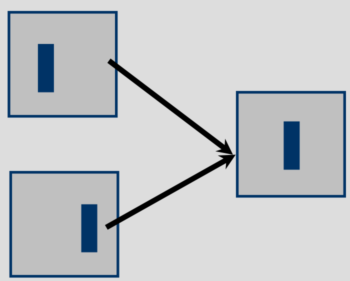


[software available online]

Our present model of the ventral stream: feedforward, accounting only for "immediate recognition"

- It is in the family of "Hubel-Wiesel" models (Hubel & Wiesel, 1959; Fukushima, 1980; Oram & Perrett, 1993, Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe, 2002; Ullman et al., 2002; Mel, 1997; Wersing and Koerner, 2003; LeCun et al 1998; Amit & Mascaro 2003; Deco & Rolls 2006…)
- As a biological model of object recognition in the ventral stream it is *perhaps* the most quantitative and faithful to known neuroscience (though many details/facts are unknown or still to be incorporated)

Two key computations, suggested by physiology

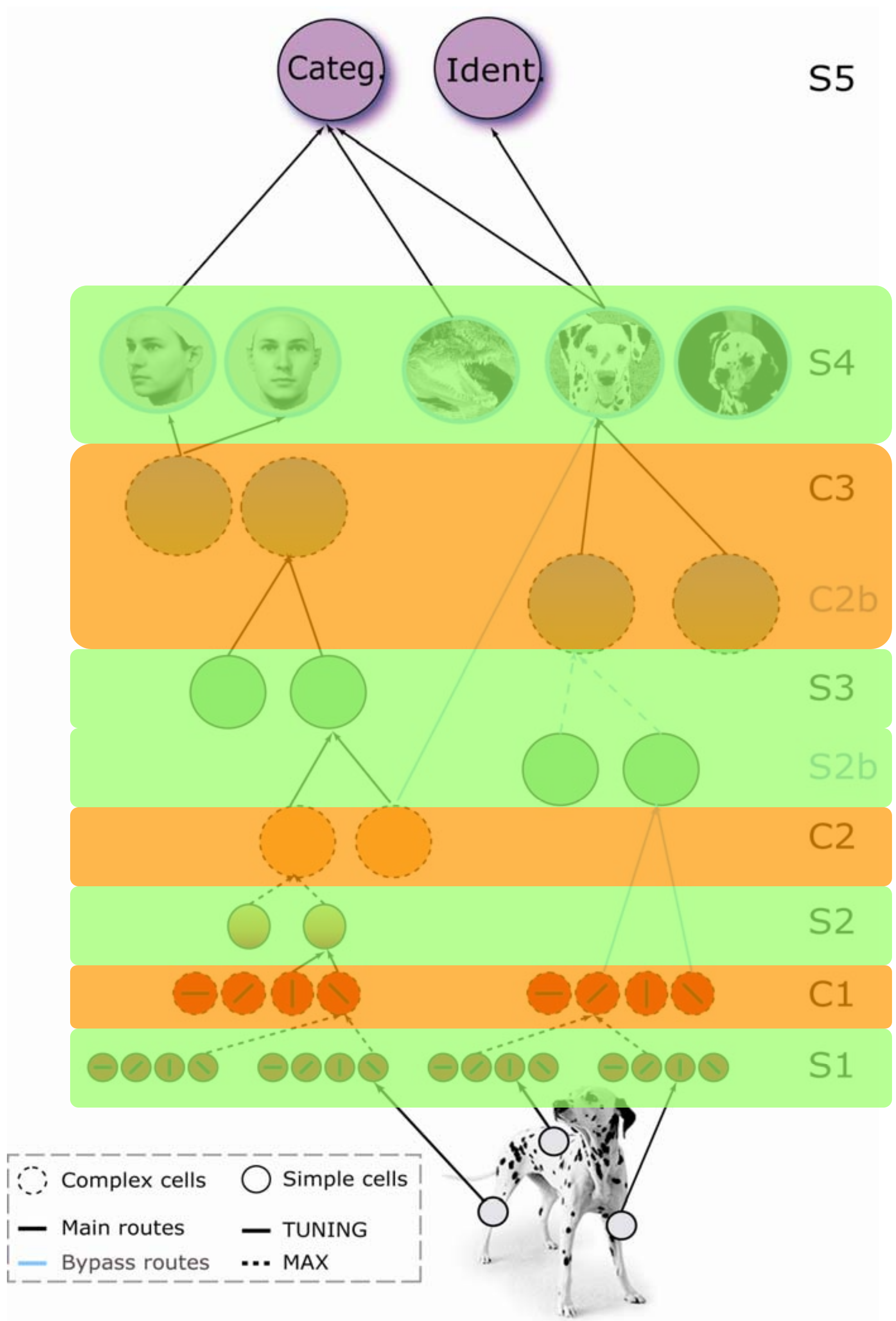
Unit types	Pooling	Computation	Operation
Simple		Selectivity / template matching	Gaussian-tuning / AND-like
Complex		Invariance	Soft-max / or-like

➢Gaussian-like tuning operation (and-like)

➢Simple units

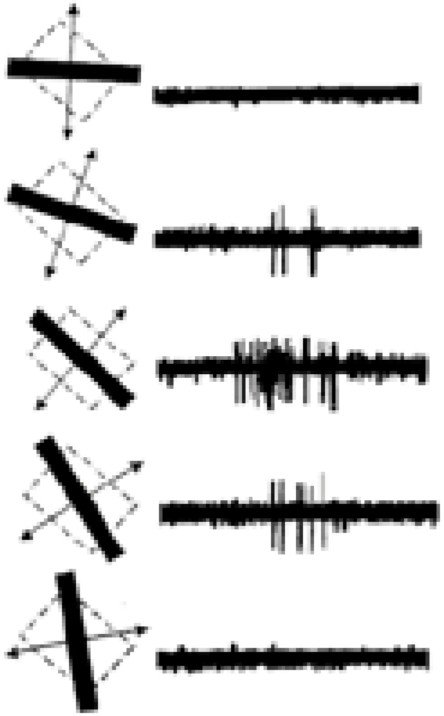
➢Max-like operation (or-like)

➢Complex units



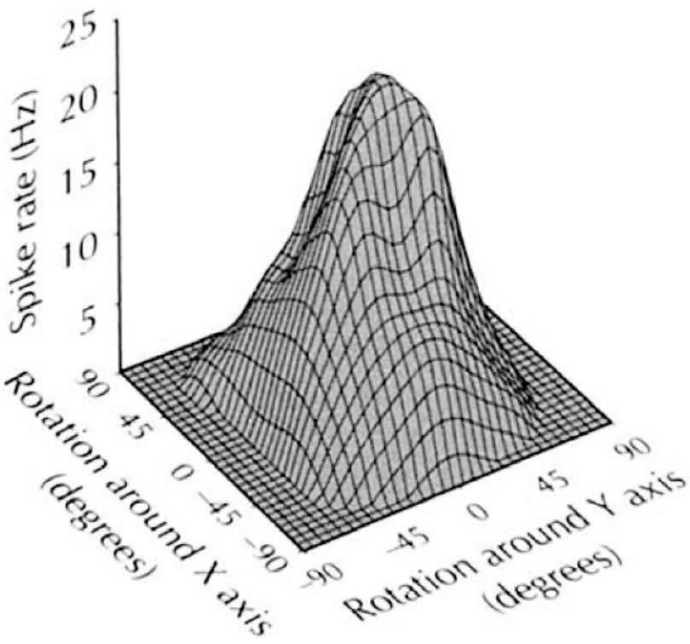
Gaussian tuning

Gaussian tuning in V1 for orientation



Hubel & Wiesel 1958

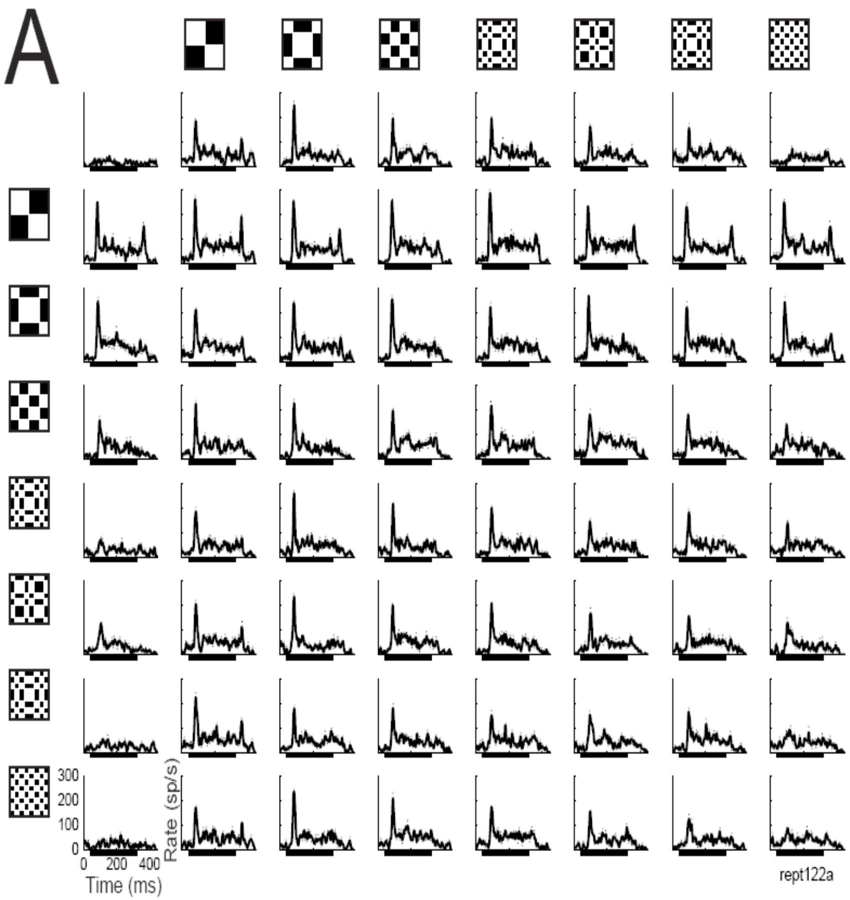
Gaussian tuning in IT around 3D views



Logothetis Pauls & Poggio 1995

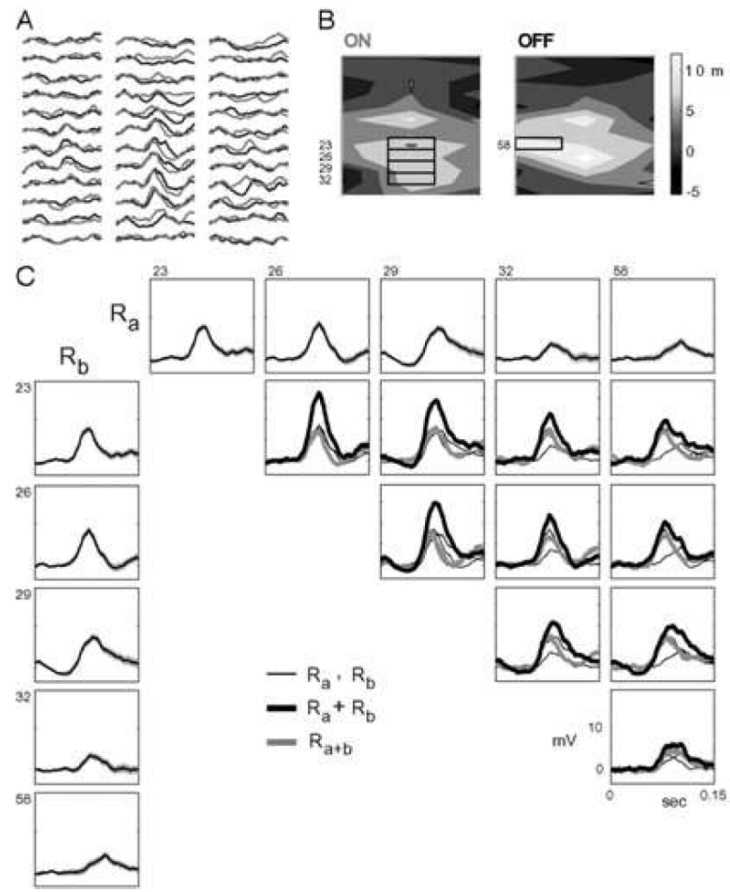
Max-like operation

Max-like behavior in V4



Gawne & Martin 2002

Max-like behavior in V1

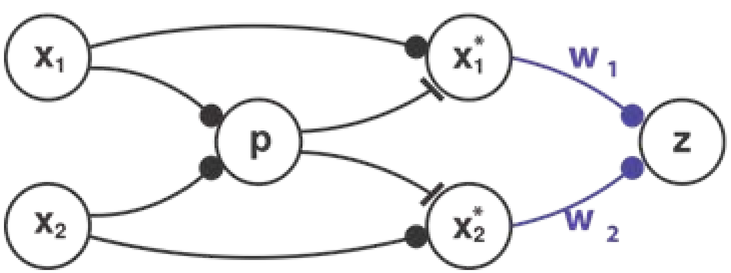


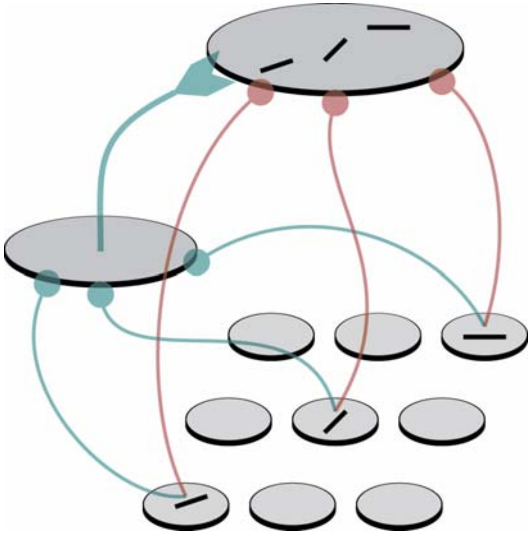
Lampl Ferster Poggio & Riesenhuber 2004
see also Finn Prieber & Ferster 2007

Plausible biophysical implementations

- Max and Gaussian-like tuning can be approximated with same canonical circuit using shunting inhibition. Tuning (eg “center” of the Gaussian) corresponds to synaptic weights.

$$y = \frac{\sum_{j=1}^{n} w_j^* \, x_j^p}{k + \left(\sum_{j=1}^{n} x_j^q \right)^r},$$





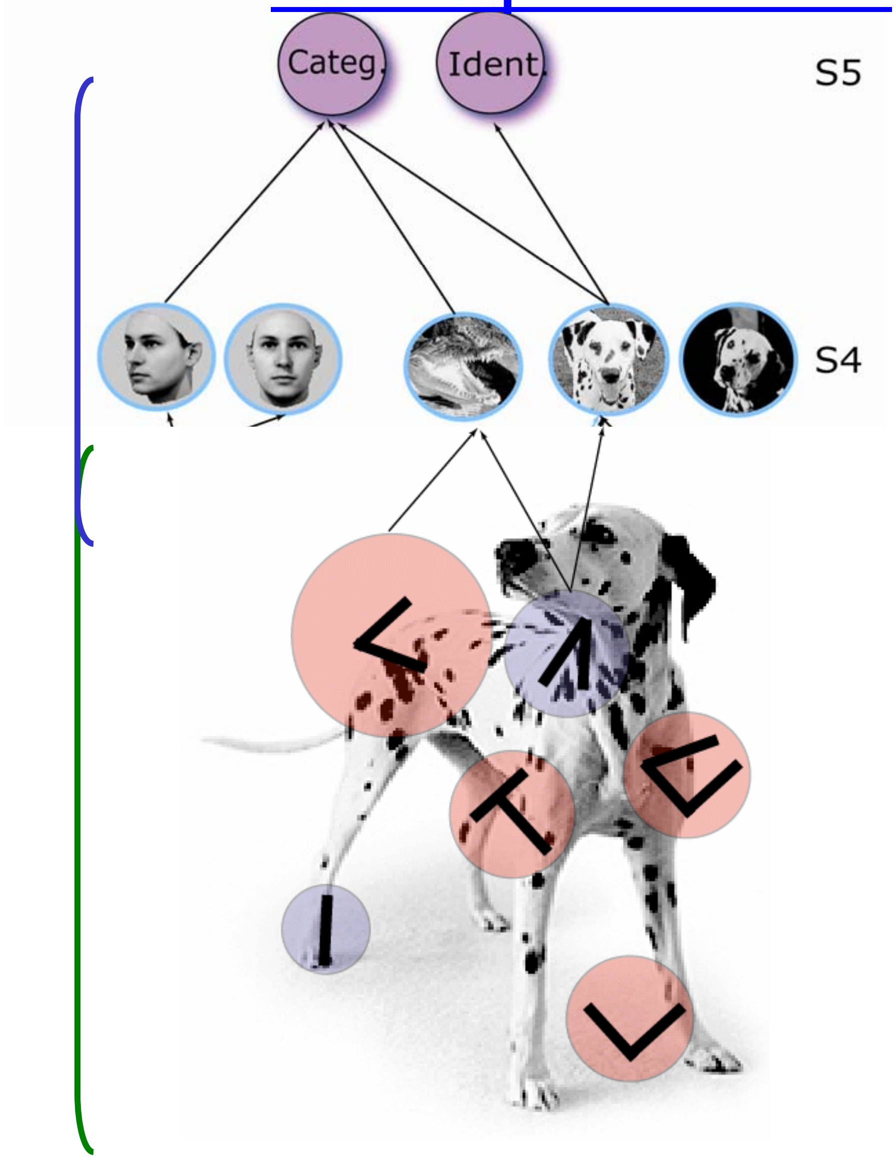
(Knoblich Koch Poggio in prep; Kouh & Poggio 2007; Knoblich Bouvrie Poggio 2007)

Learning: supervised and unsupervised

Task-specific circuits (from IT to PFC)

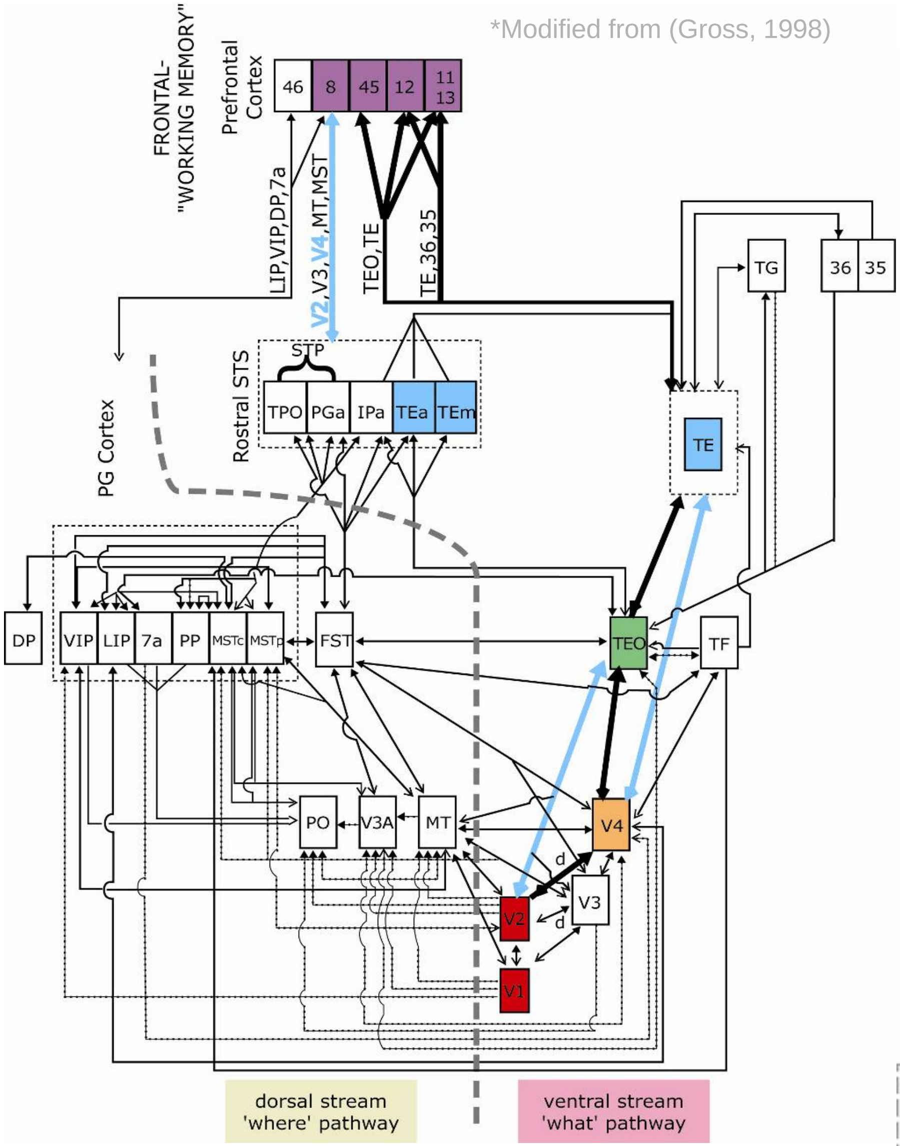
- Supervised learning: ~ classifier

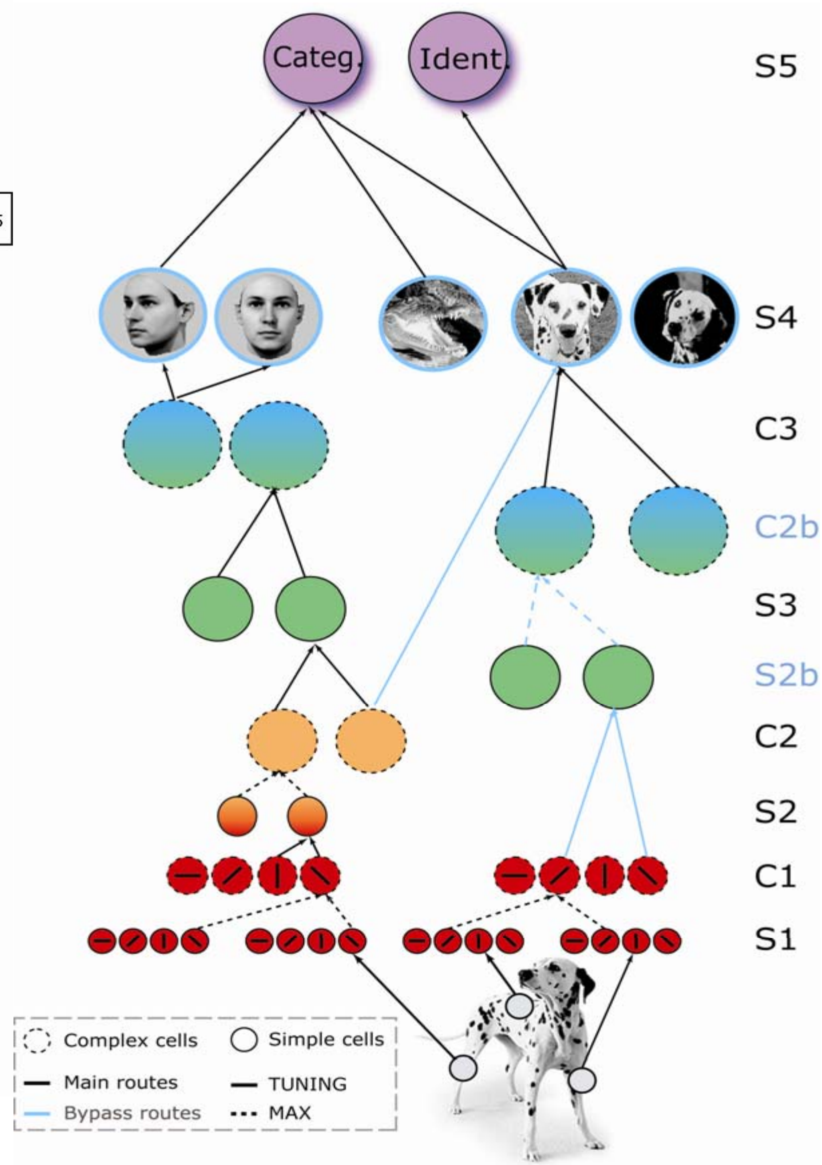
- Generic, overcomplete dictionary of “templates” or image components (from V1 to IT) represented by tuning of cells generated during **unsupervised** learning (from ~10,000 natural images) during a developmental-like stage



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

A hierarchical algorithm...





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

[software available online]

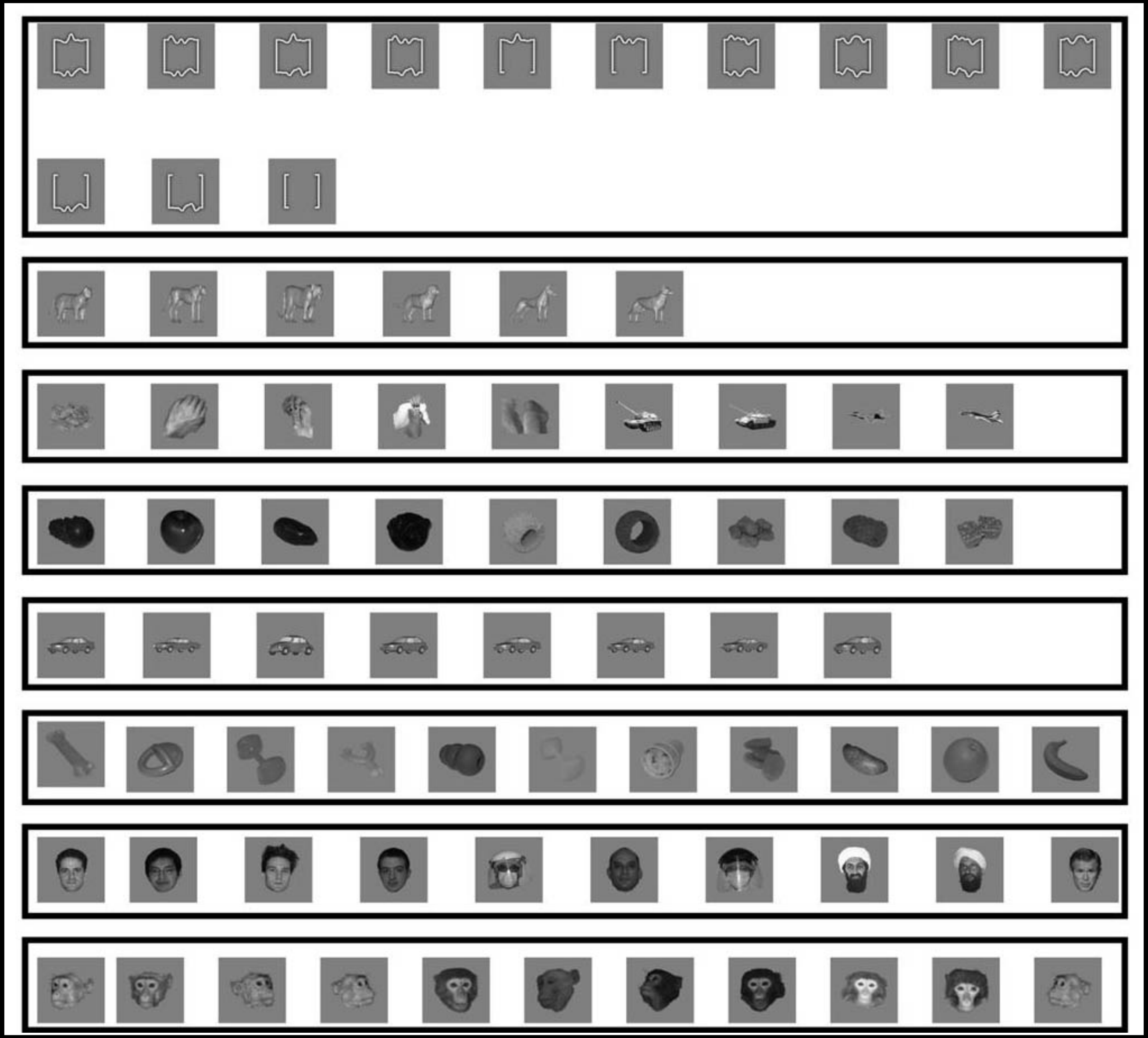
Feedforward Models: comparison w/ neural data

- 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)
  - **Read out results** (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)

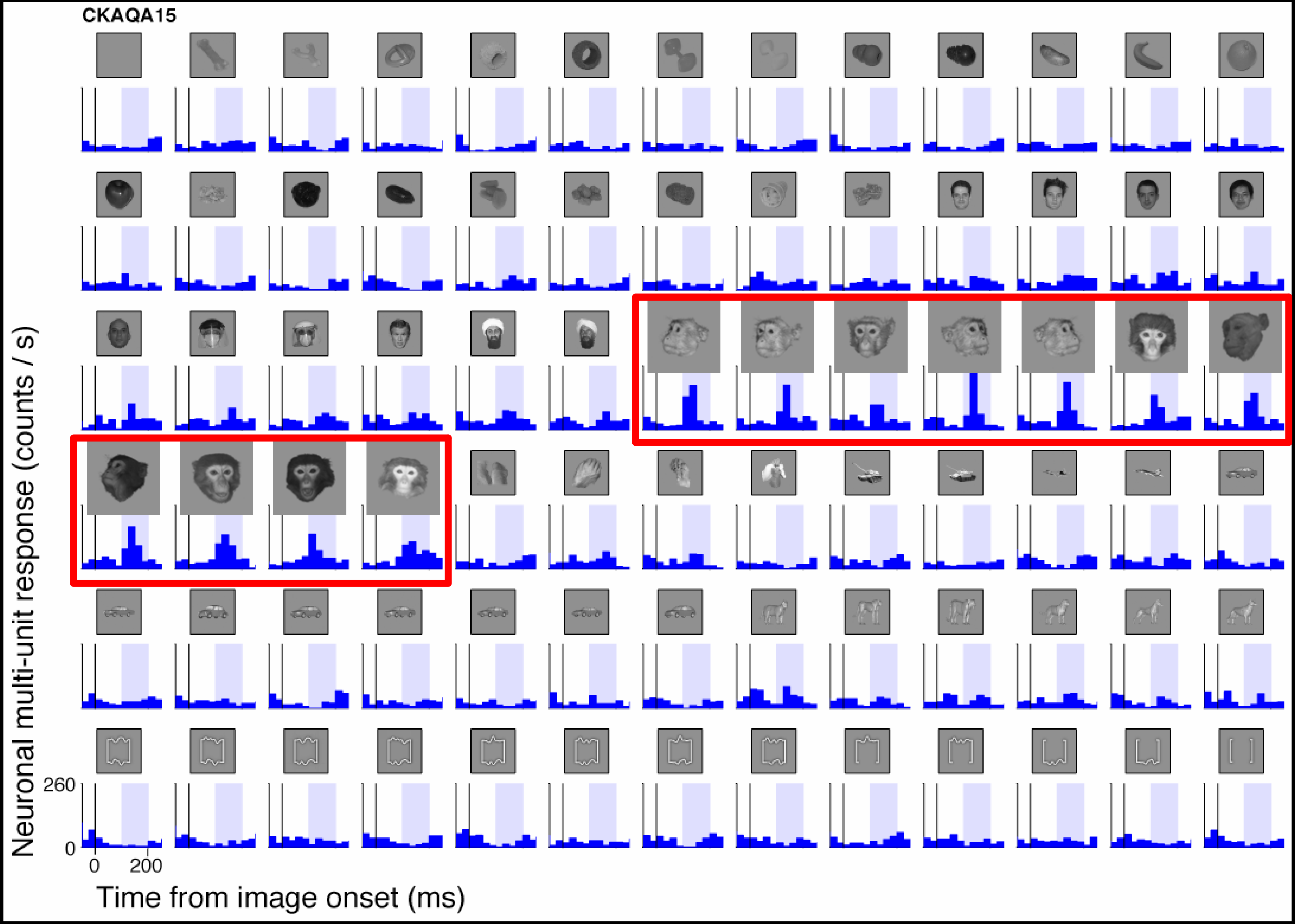
IT Readout

77 objects,
8 classes



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

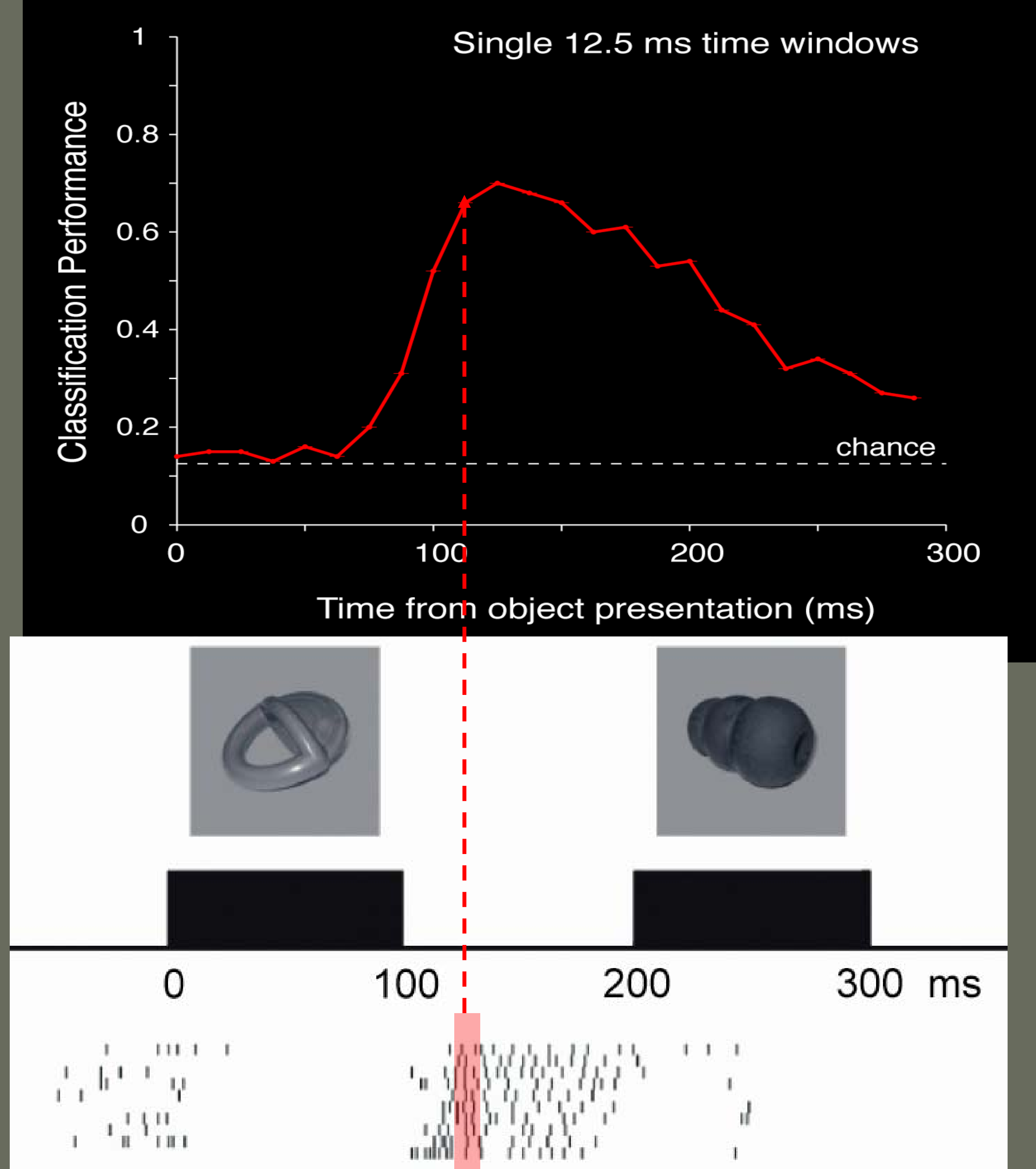
Example of One IT Cell



A result (C. Hung, et al., 2005): very rapid read-out of object information rapid (80-100 ms from onset of stimulus)

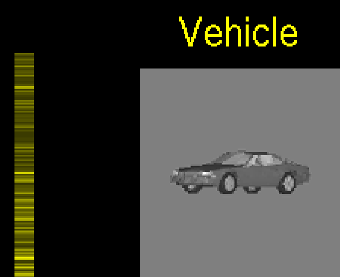
Information represented by population of neurons over very short times (over 12.5ms bin)

Very strong constraint on neural code (not firing rate). Consistent with our IF circuits for max and tuning



From neuronal population activity in IT…

…a classifier trained on examples can decode and guess what the monkey was seeing…



Video speed: 1 frame/sec
Actual presentation rate: 5 objects/sec

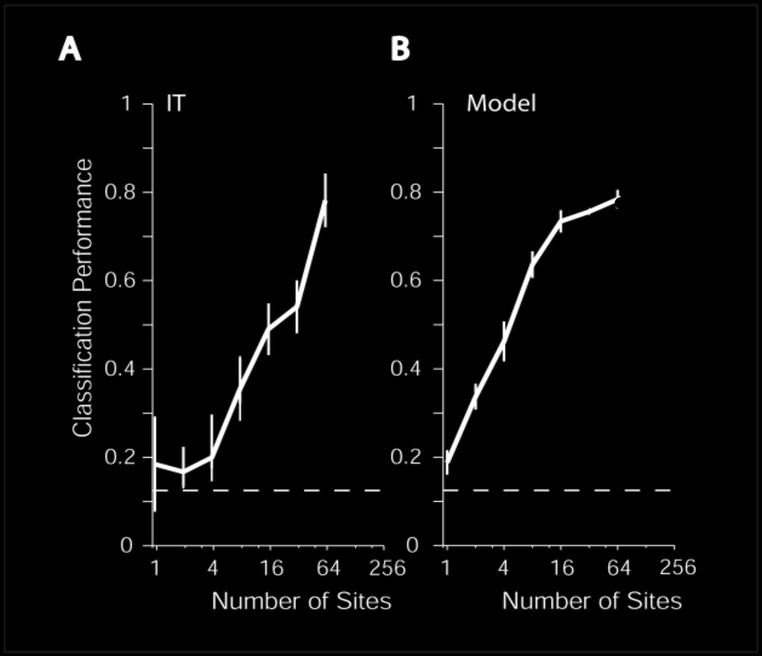
Categorization

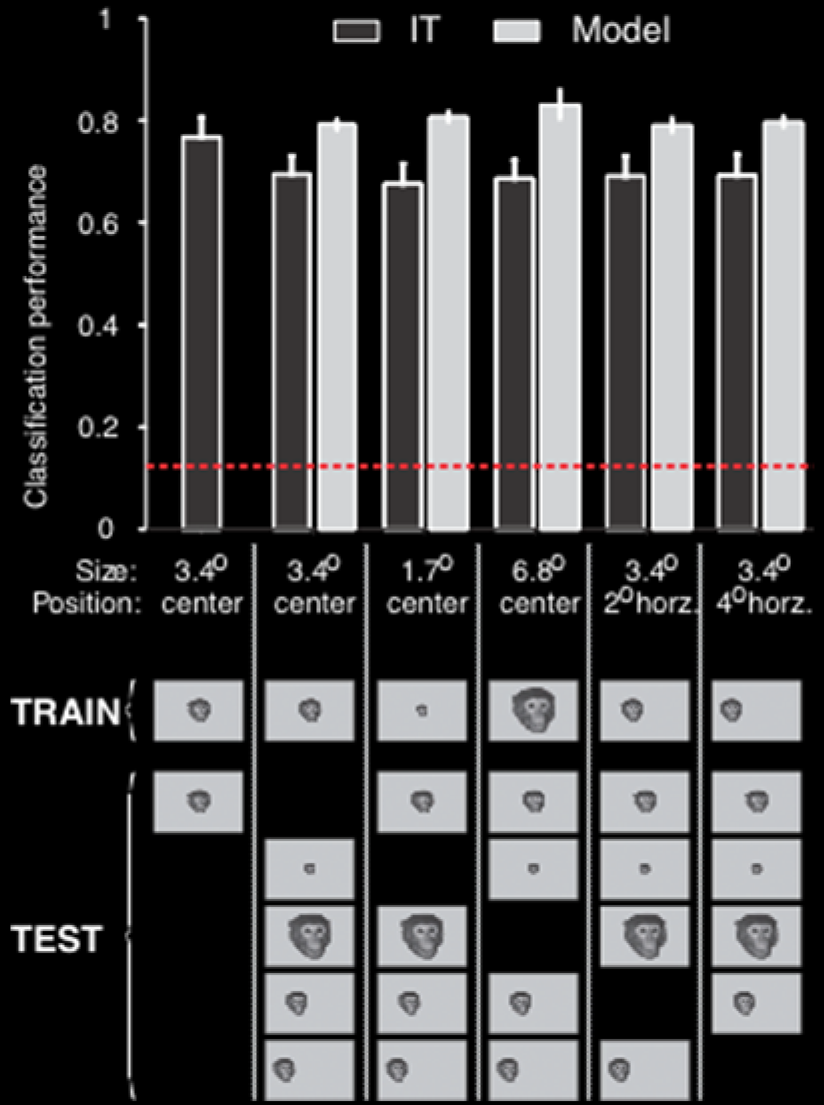
- Toy
- Body
- Human Face
- Monkey Face
- Vehicle
- Food
- Box
- Cat/Dog

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

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

Agreement of Model w| IT Readout data

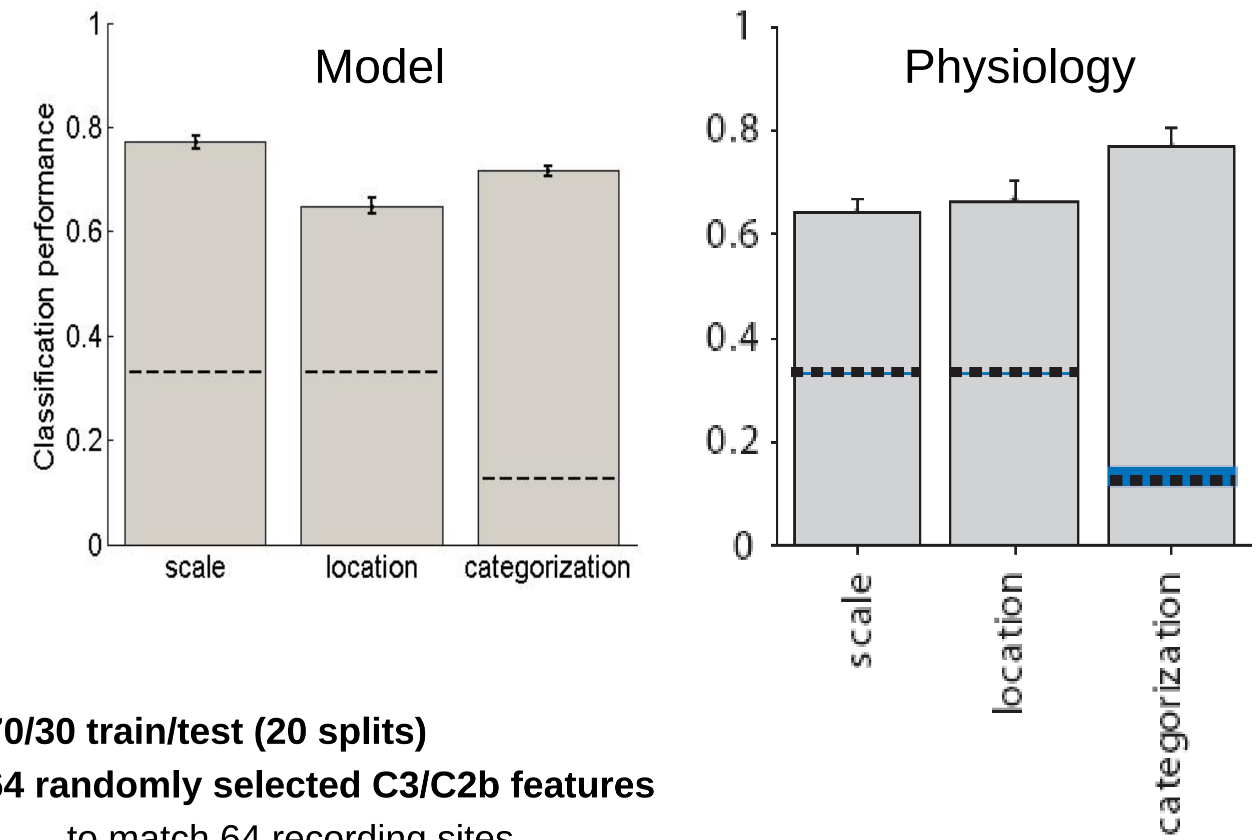




Reading out category and identity "*invariant*" to position and scale

Hung, et al. 2005; Serre et al., 2005

Reading Out Scale and Position Information: comparing the model to Hung et al.



- **70/30 train/test (20 splits)**
- **64 randomly selected C3/C2b features**
  - to match 64 recording sites
- **Scale:** 77.2 ± 1.25% vs. ~63% (physiology)
- **Location:** 64.9 ± 1.44% vs. ~65% (physiology)
- **Categorization:** 71.6 ± 0.91% vs. ~77% (physiology)

Tan, Serre, Poggio, 2008

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 model 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 invariant features

Readings on the work with many relevant references

A detailed description of much of the work is in the "supermemo" at

http://cbcl.mit.edu/projects/cbcl/publications/ai-publications/2005/AIM-2005-036.pdf

Other recent publications and references can be found at

http://cbcl.mit.edu/publications/index-pubs.html

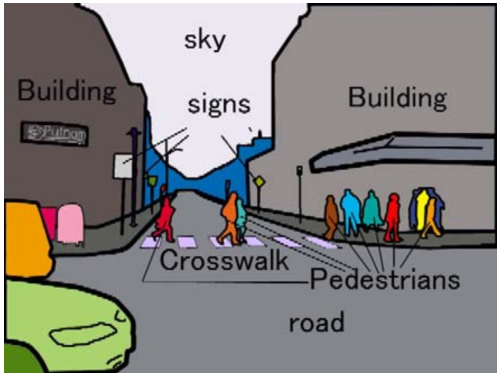
Limitations of present feedforward hierarchical models

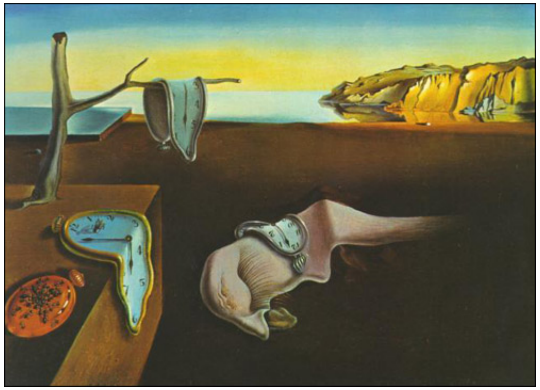
- Most existing models of visual cortex do not account
  - -- for cortical backprojections
  - -- for the emerging detailed connectivity among cortical areas or patches (e.g. “network of face patches….)
  - -- for subcortical pathways and noncortical brain regions e.g. pulvinar…)
- More data from physiology and fMRI are needed

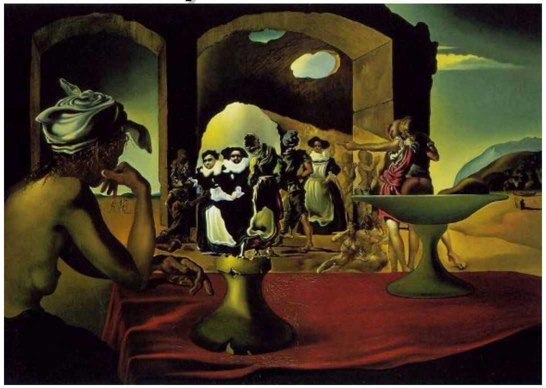
Limitations of present feedforward hierarchical models

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

- Our visual system can "answer" almost any kind of question about an image or video (a Turing test for vision…)







- Two options: 1) top-down (attentional) control of task-dependent routines 2) probabilistic inference in the ventral stream

Collaborators

T. Serre

- Model
  - ✓ A. Oliva
  - ✓ C. Cadieu
  - ✓ U. Knoblich
  - ✓ M. Kouh
  - ✓ G. Kreiman
  - ✓ M. Riesenhuber
- Comparison w| humans
  - ✓ A. Oliva
- Action recognition
  - ✓ H. Jhuang
- Attention
  - ✓ S. Chikkerur
  - ✓ C. Koch
  - ✓ D. Walther
- Computer vision
  - S. Bileschi
  - L. Wolf
- Learning invariances
  - T. Masquelier
  - S. Thorpe