Class 14-15

Vision and Visual Neuroscience

Tomaso Poggio
Jim Mutch + Hueihan Jhuang

Plan for class 14-15-16-17

- Class 14: HLM in the ventral stream of visual cortex
- Class 15 Models of the ventral an dorsal stream
- Class 16: Derived Kernels: a mathematical framework for hierarchical learning machines
- Class 17: Attention: a Bayesian extension of the model

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.

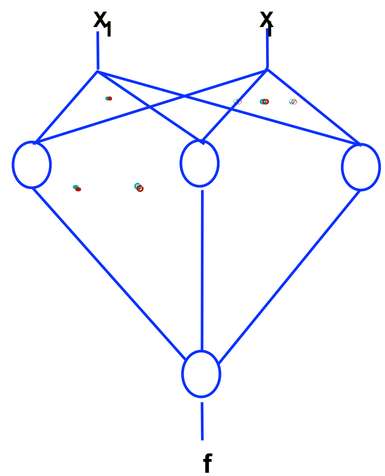
- There may also be the more fundamental issue of *sample complexity*. Learning theory shows that the difficulty of a learning task depends on the size of the required hypothesis space. This complexity determines in turn how many training examples are needed to achieve a given level of generalization 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)

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

implies

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



Remark:

Kernel machines correspond to *shallow* networks

Winning against the curse of dimensionality: new research directions in learning

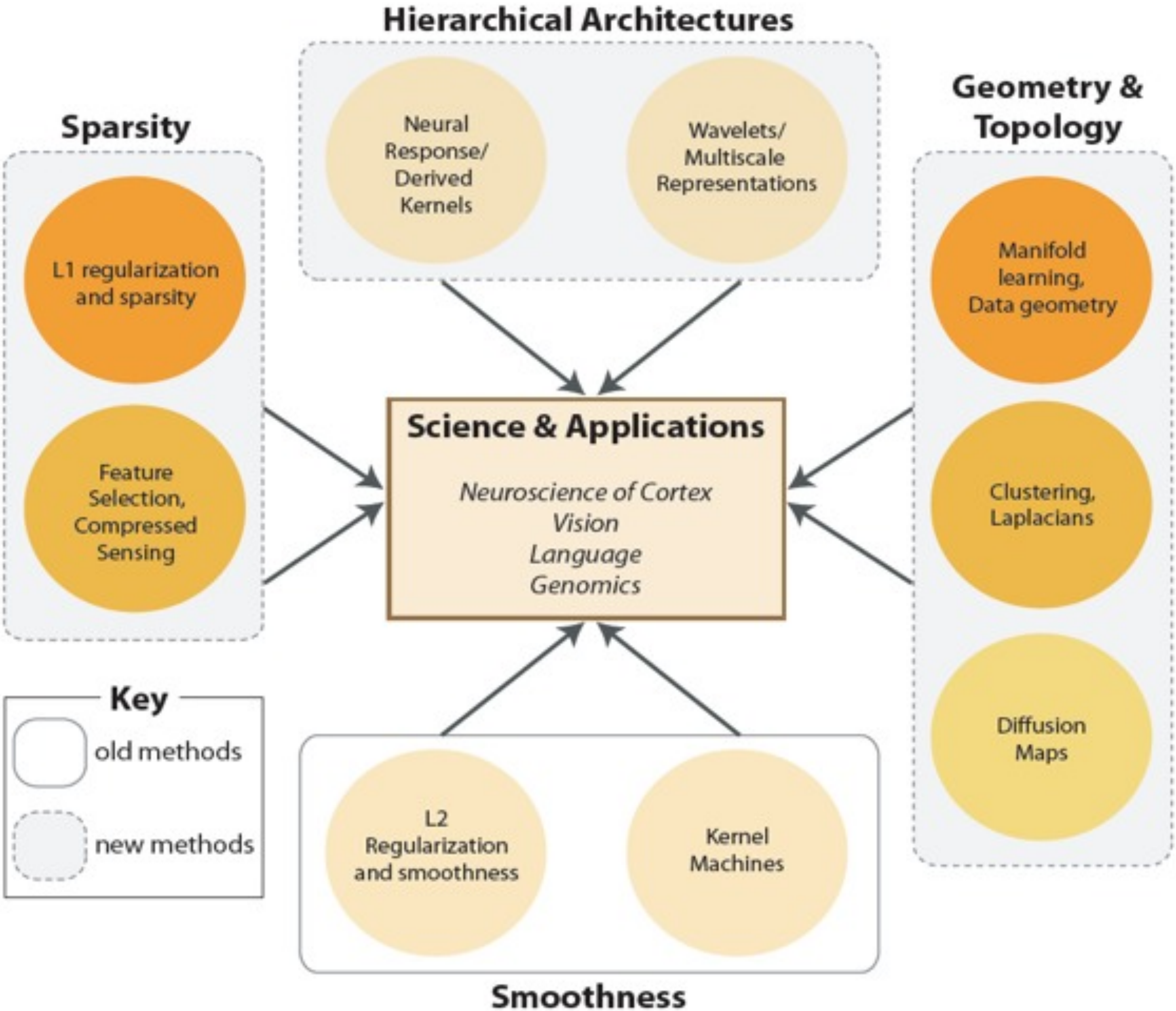
Many processes - physical processes as well as human activities – generate high-dimensional data: *curse of dimensionality or poverty of stimulus*.

There are, however, basic properties of the data generating process that may allow to circumvent the problem of high dimensionality and make the analysis possible:

- smoothness - exploited by L2 regularization techniques
- sparsity - exploited by L1 regularization techniques
- data geometry - exploited by manifold learning techniques
- hierarchical organization – suggested by the architecture of sensory cortex

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

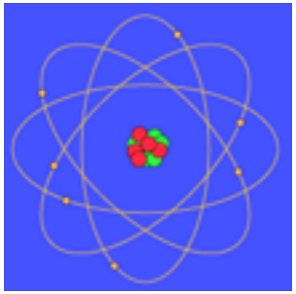
New Research Directions



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



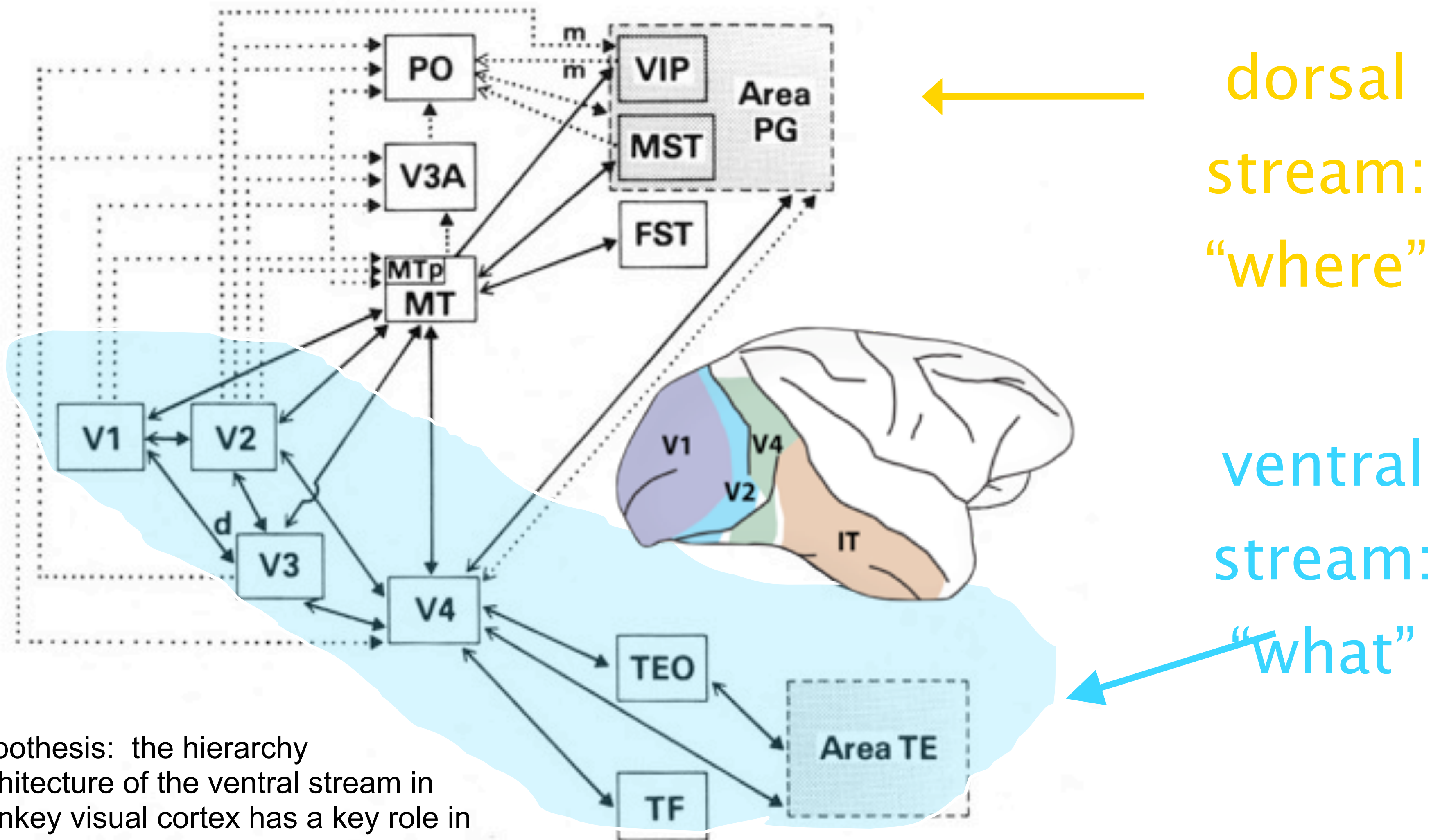
- All models are "wrong"
- Some models can be useful summaries of data and some can be a good starting point for more complete theories

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models
   - Ventral stream model in more details (Jim Mutch)
   - Dorsal stream model (Hueihan Jhuang)

The Ventral Stream

unconstrained visual recognition is a difficult learning problem
(e.g., “is there an animal in the image?”)

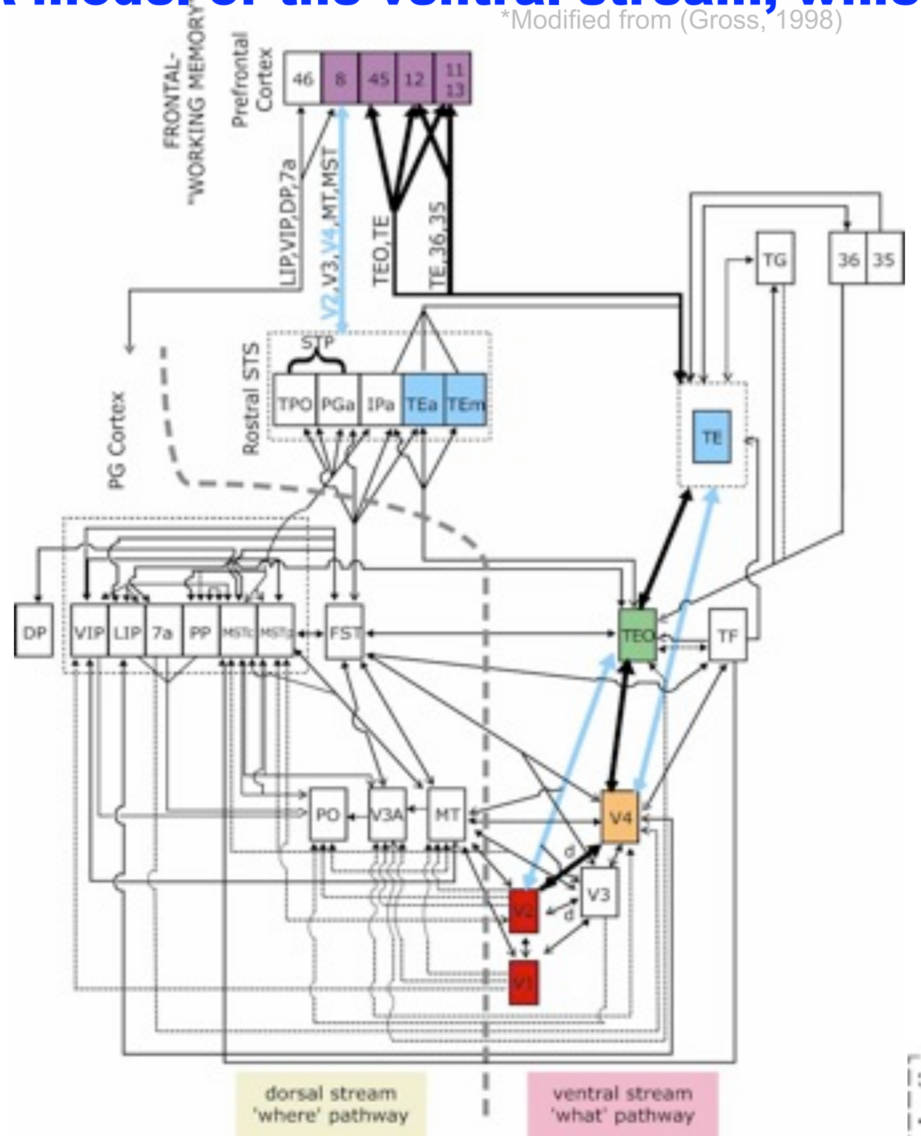
Object Recognition and the Ventral Stream

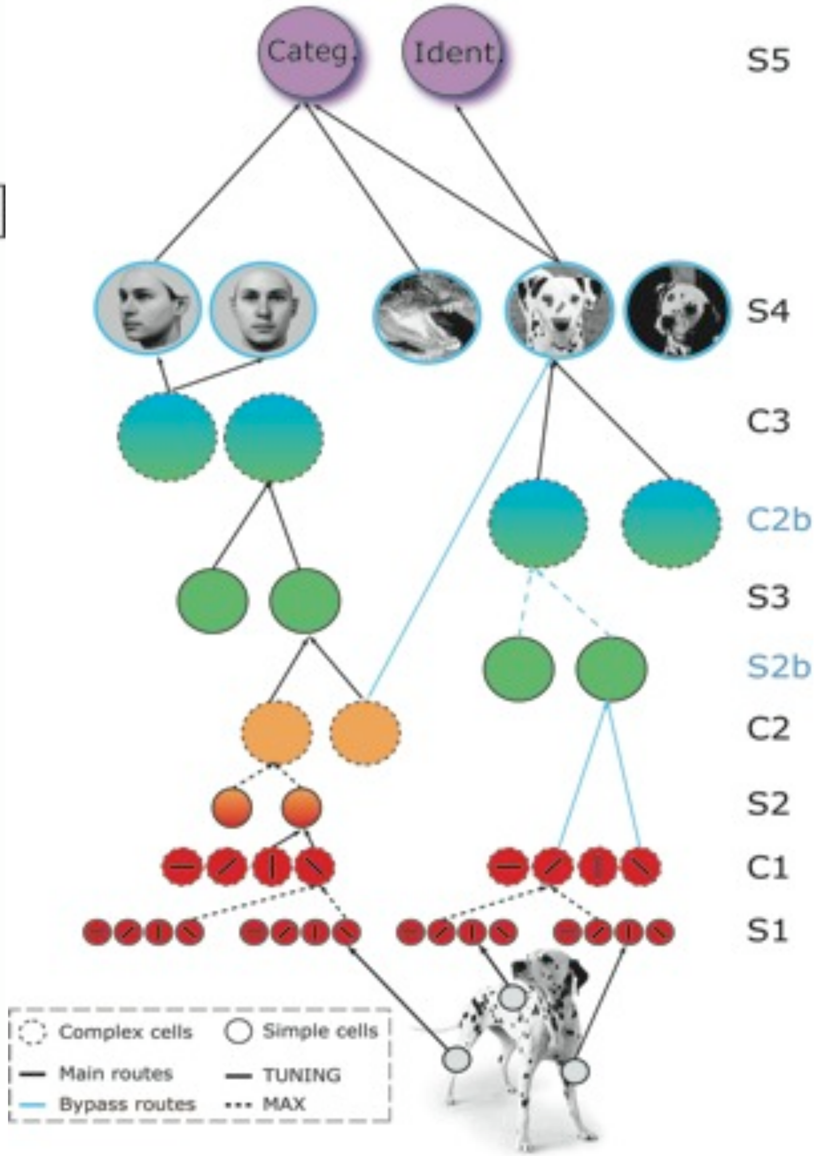


Hypothesis: the hierarchy architecture of the ventral stream in monkey visual cortex has a key role in object recognition…of course subcortical pathways may also be important (thalamus, in particular pulvinar…).

Desimone & Ungerleider 1989

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]

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models
    - Ventral stream model in more details (Jim Mutch)
    - Dorsal stream model (Hueihan Jhuang)

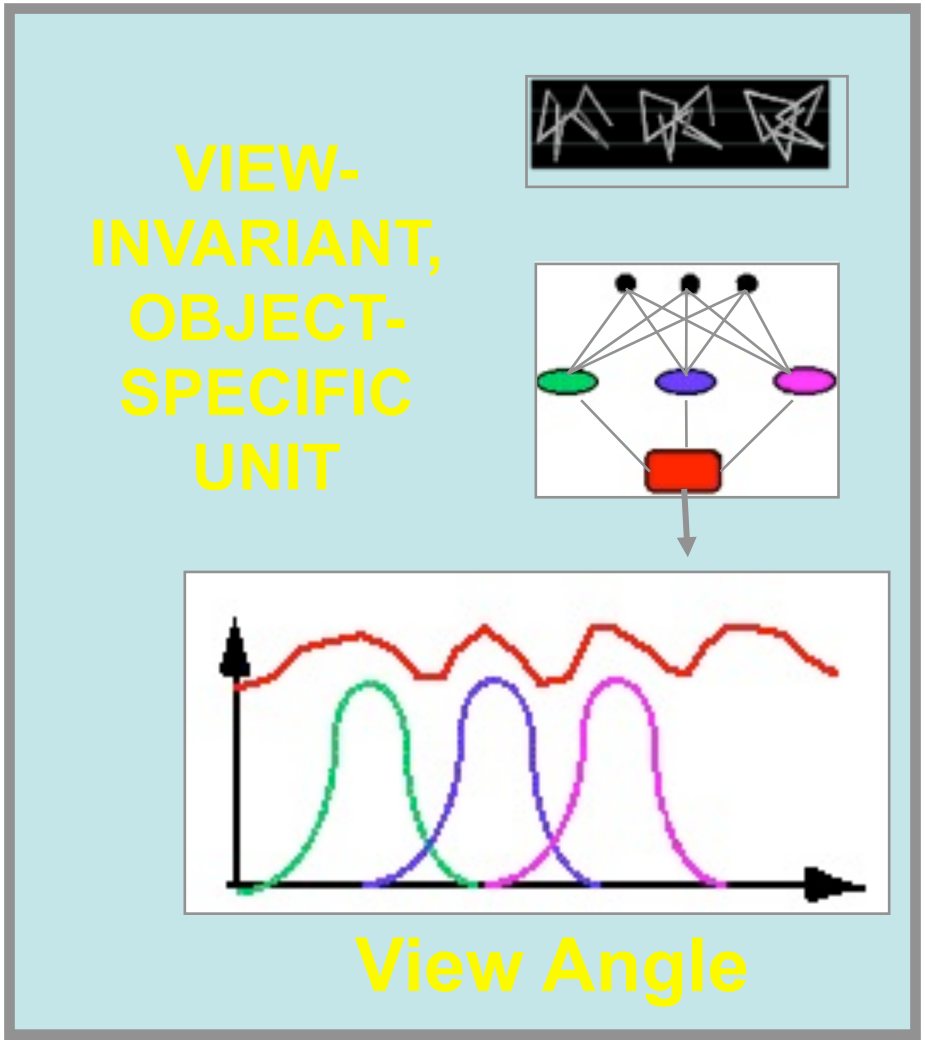
Some personal history:

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!)



Regularization Network (GRBF) with Gaussian kernels

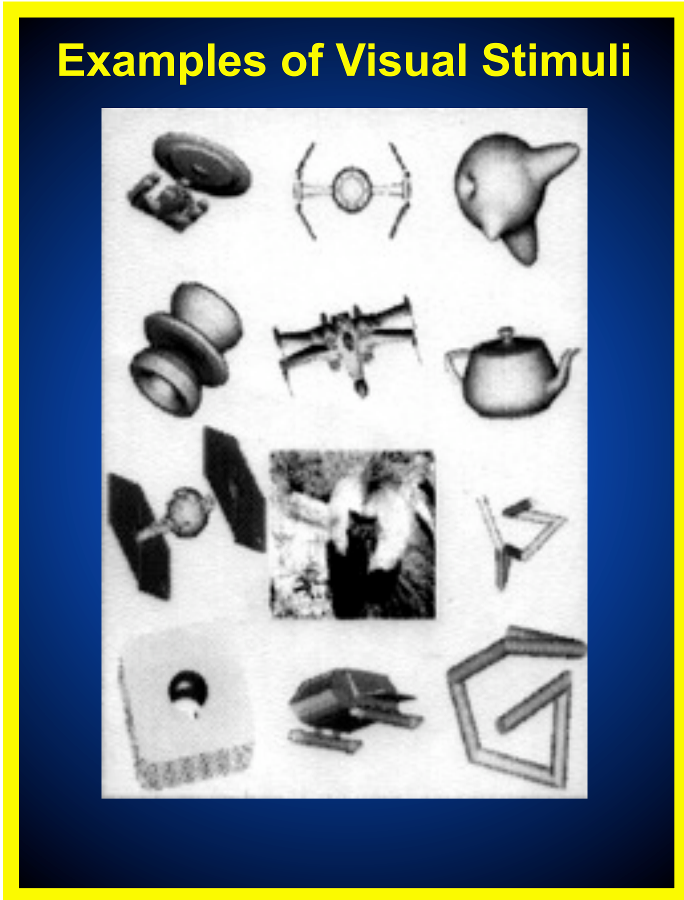
Prediction: neurons become view-tuned through learning

Poggio & Edelman 1990

Learning to Recognize 3D Objects in IT Cortex

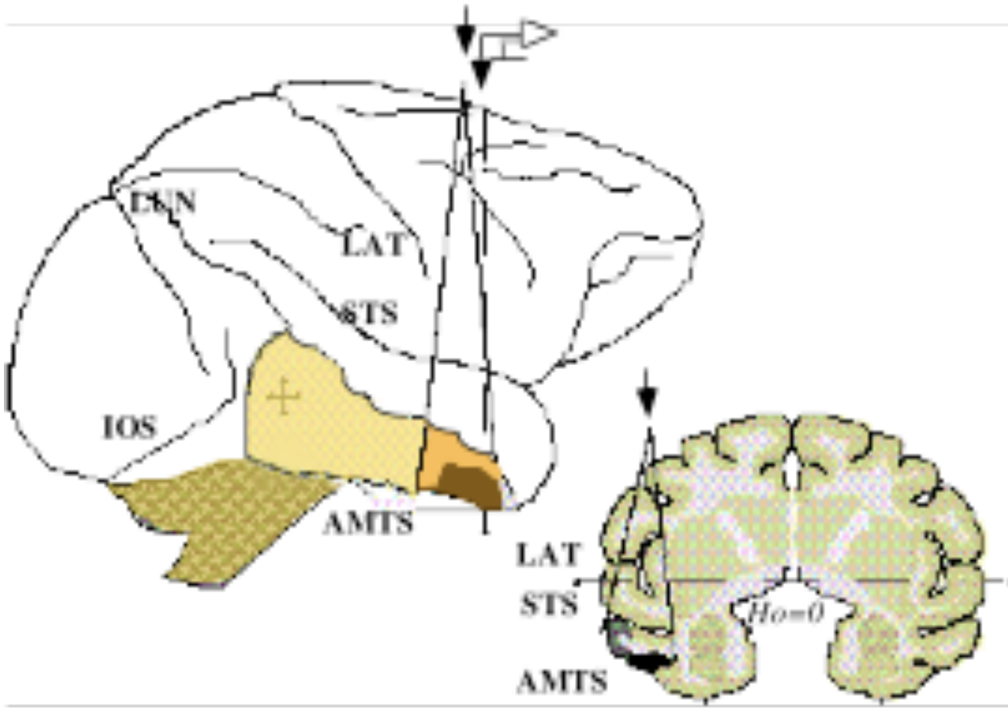
After human psychophysics (Buelthoff, Edelman, Tarr, Sinha, *to be added next year*…), which supports models based on view-tuned units...

… physiology!



Logothetis Pauls & Poggio 1995

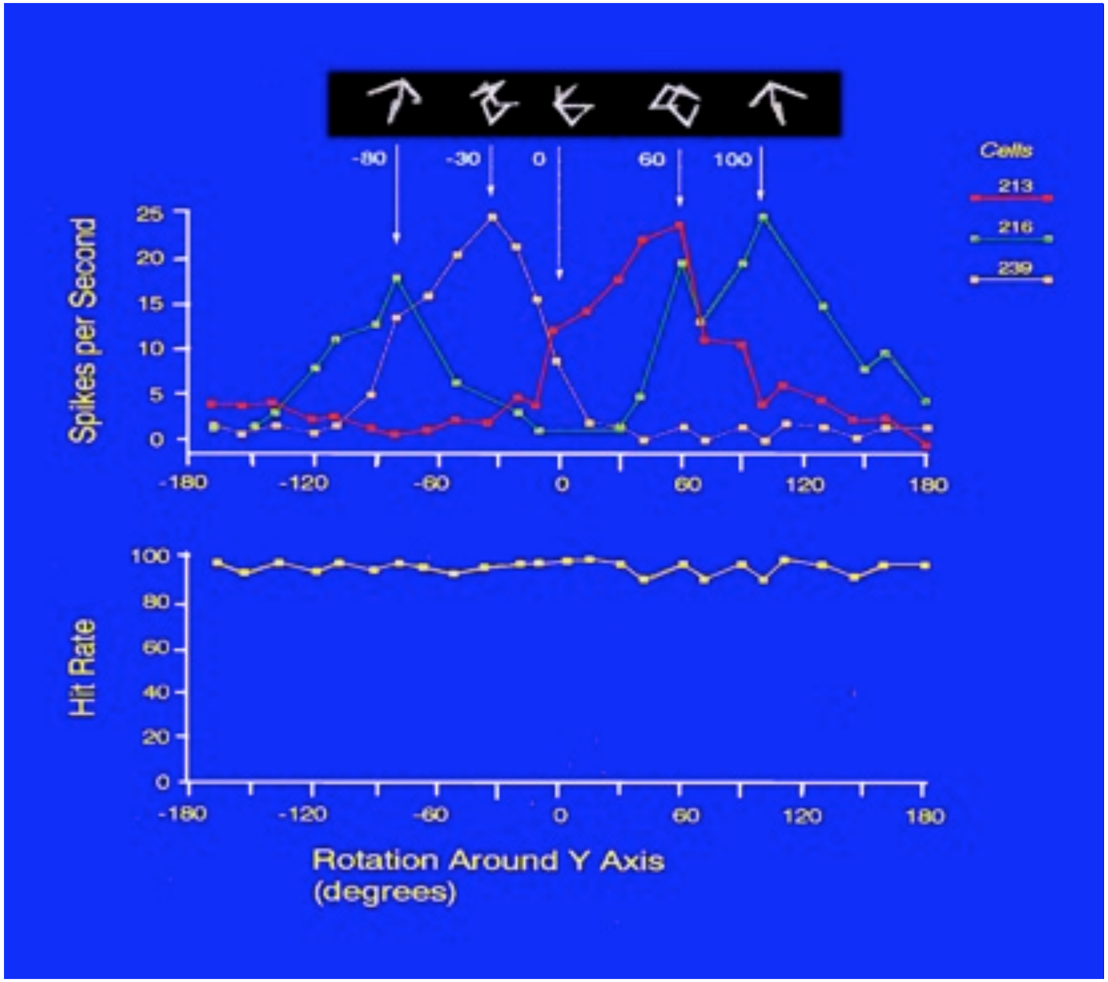
Recording Sites in Anterior IT



…neurons tuned to faces are intermingled nearby….

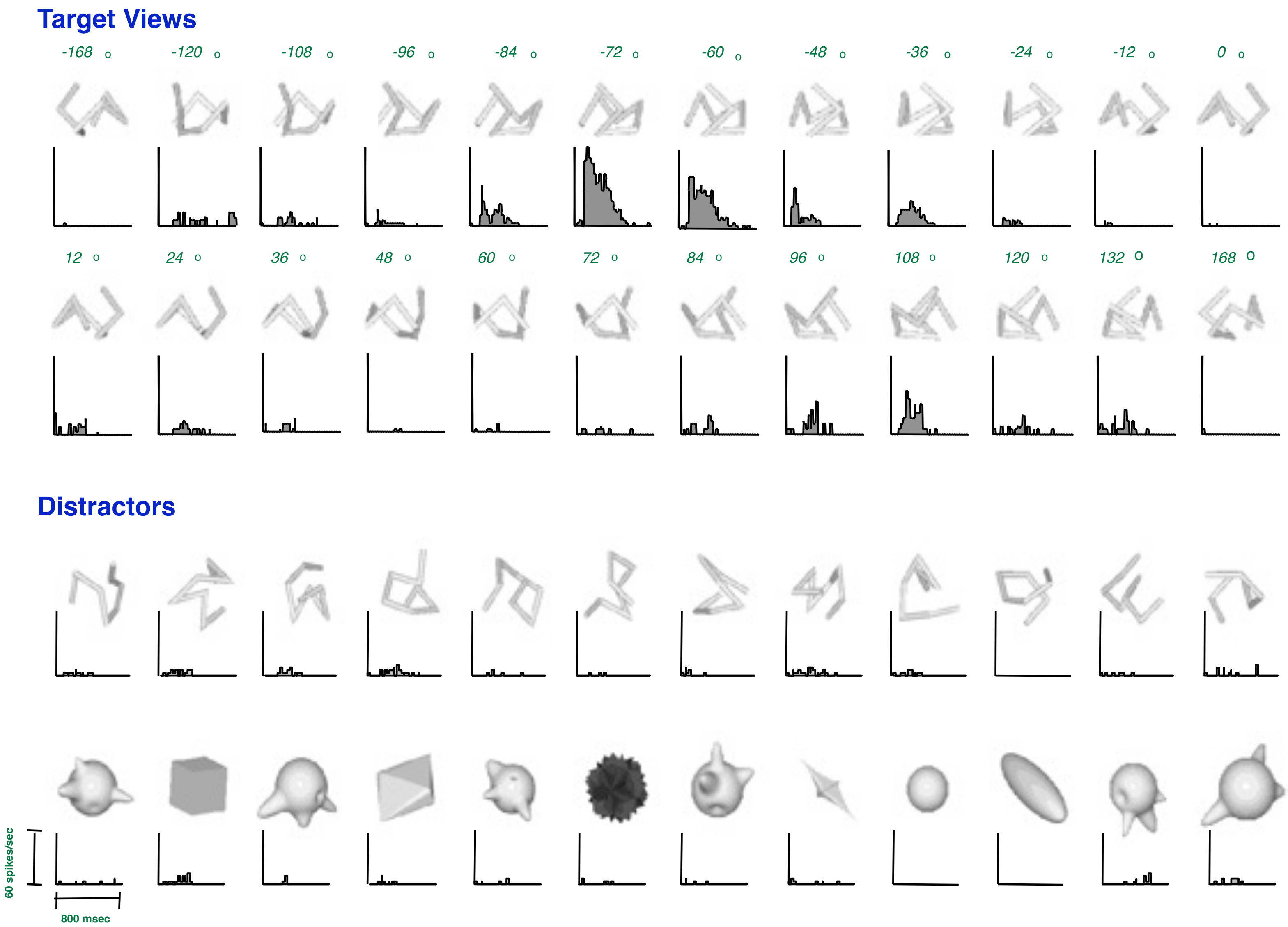
Logothetis, Pauls & Poggio 1995

Neurons tuned to object views, as predicted by model!



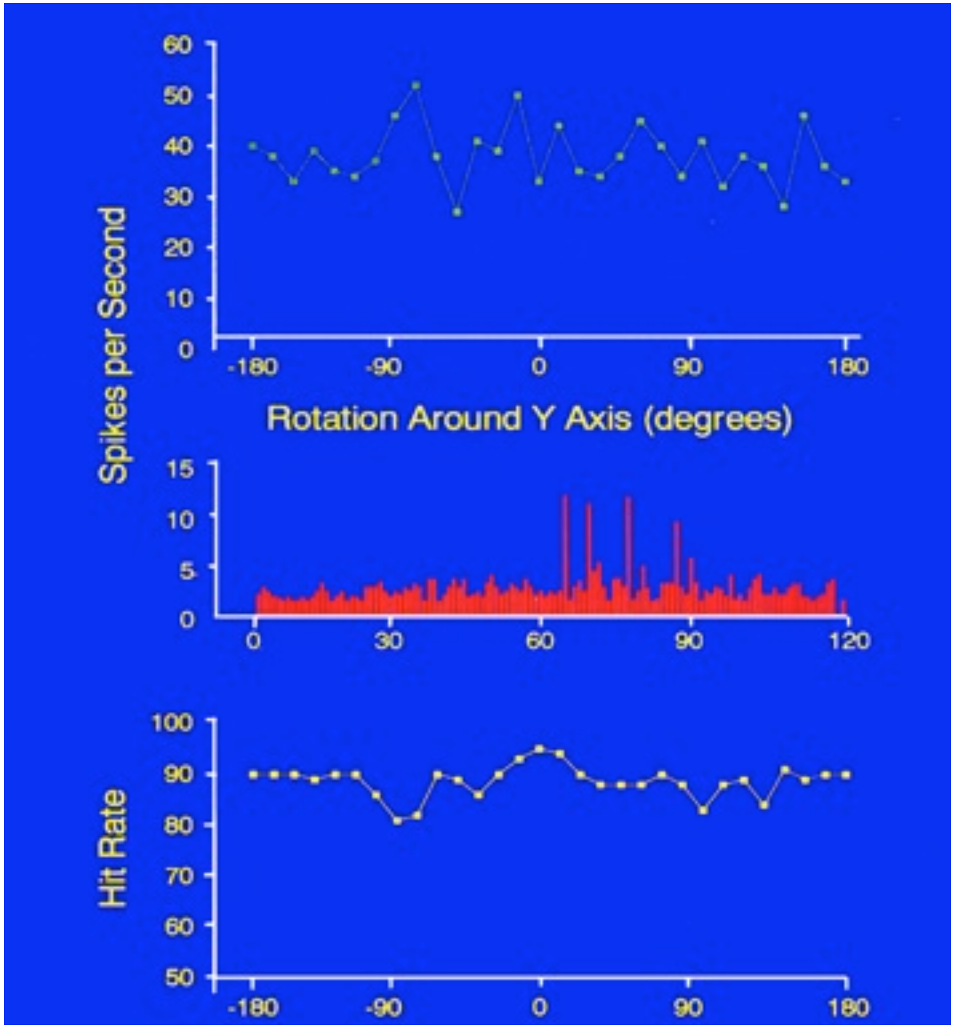
Logothetis Pauls & Poggio 1995

A “View-Tuned” IT Cell



Logothetis Pauls & Poggio 1995

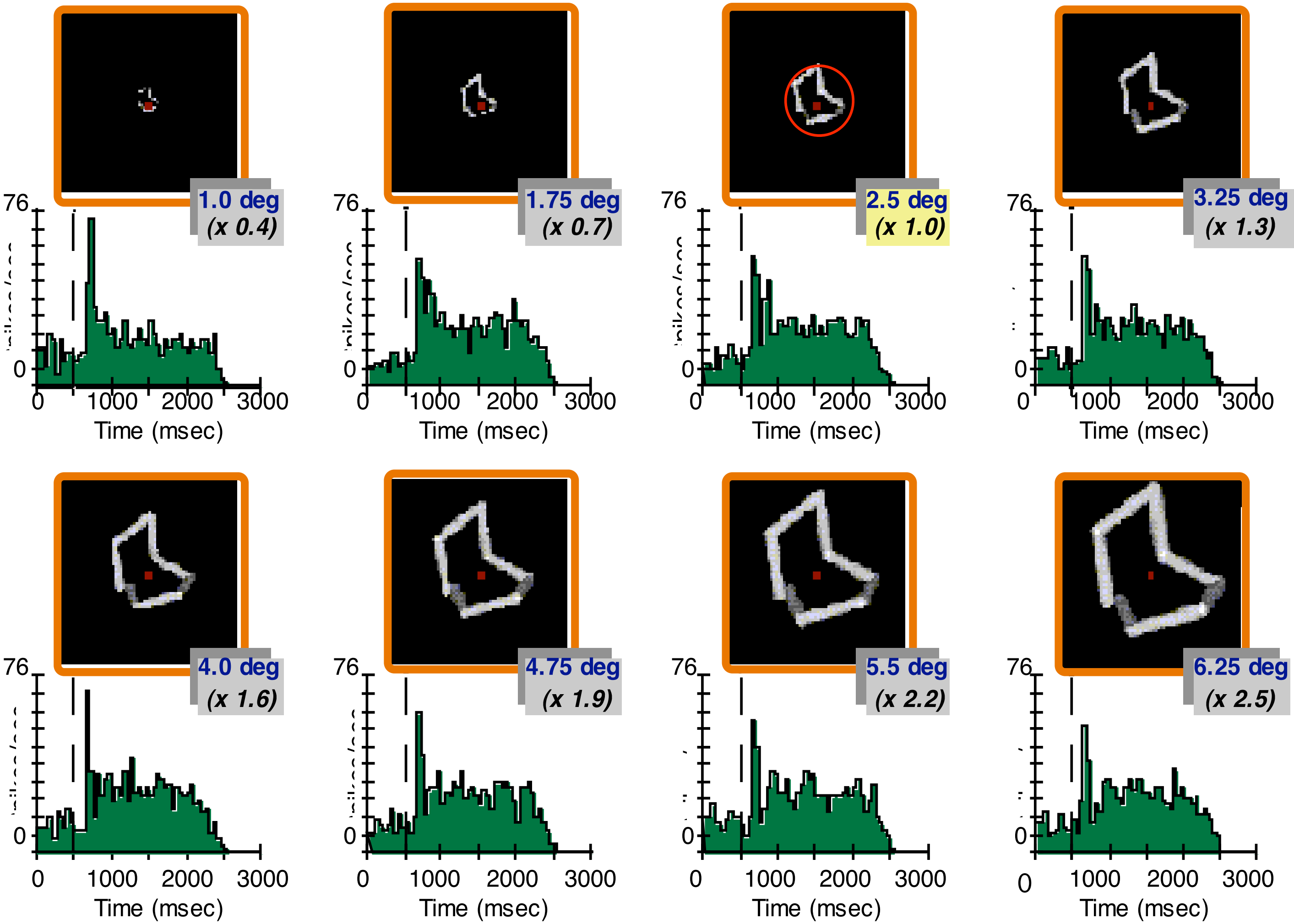
But also view-invariant object-specific neurons (5 of them over 1000 recordings)



Logothetis Pauls & Poggio 1995

View-tuned cells:

scale invariance (one training view only) motivates present model



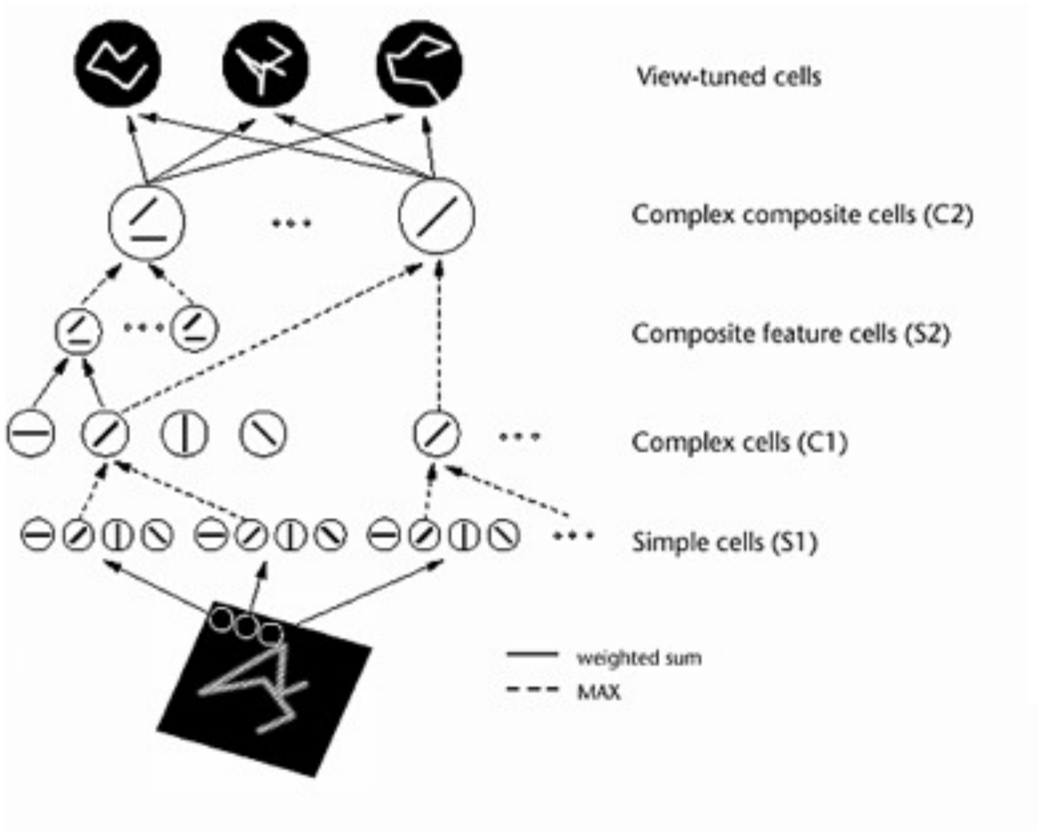
Logothetis Pauls & Poggio 1995

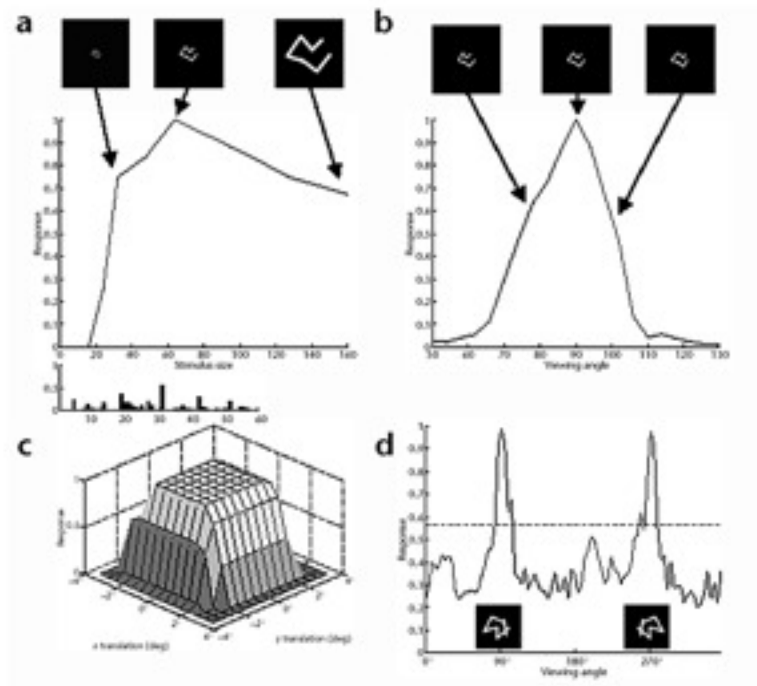
Hierarchy

- 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

Answer: 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
   - Ventral stream model in more details (Jim Mutch)
   - Dorsal stream model (Hueihan Jhuang)

Different shapes and sizes but common structure

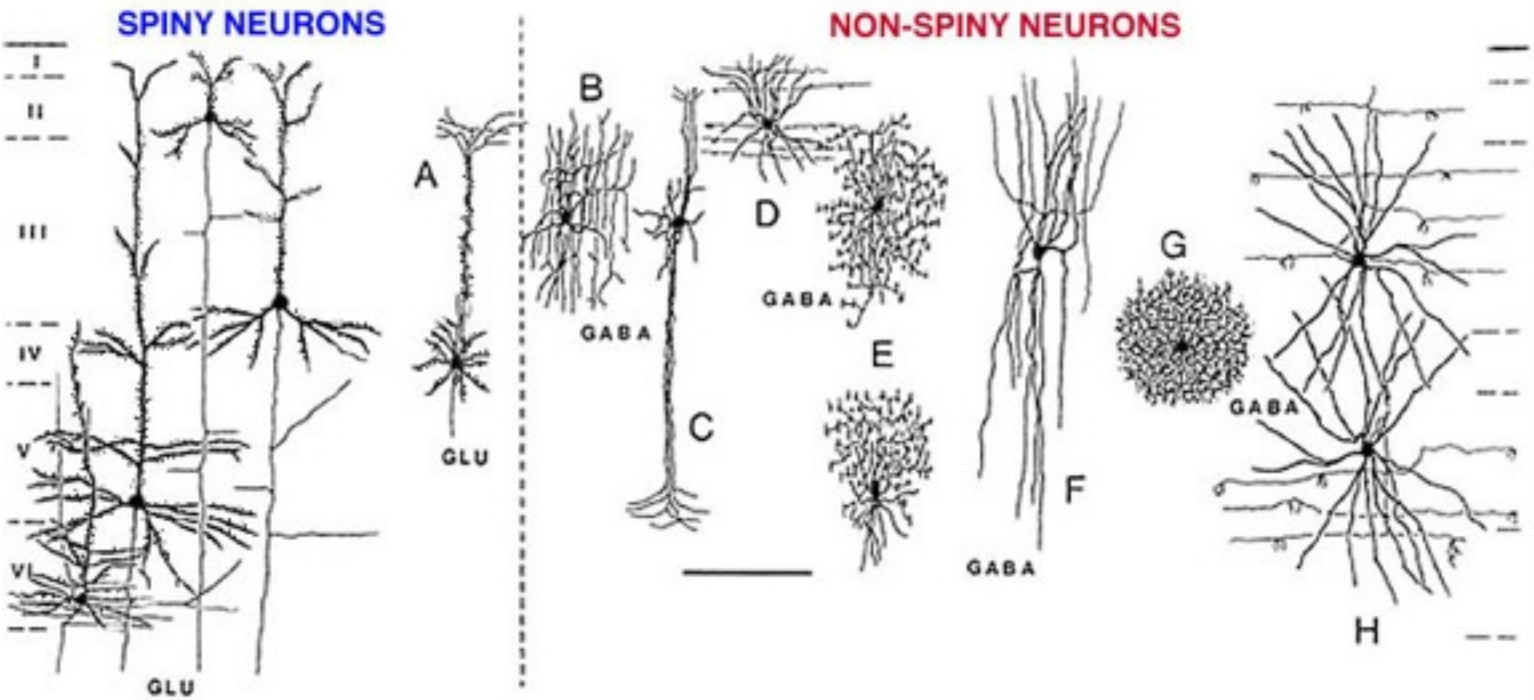
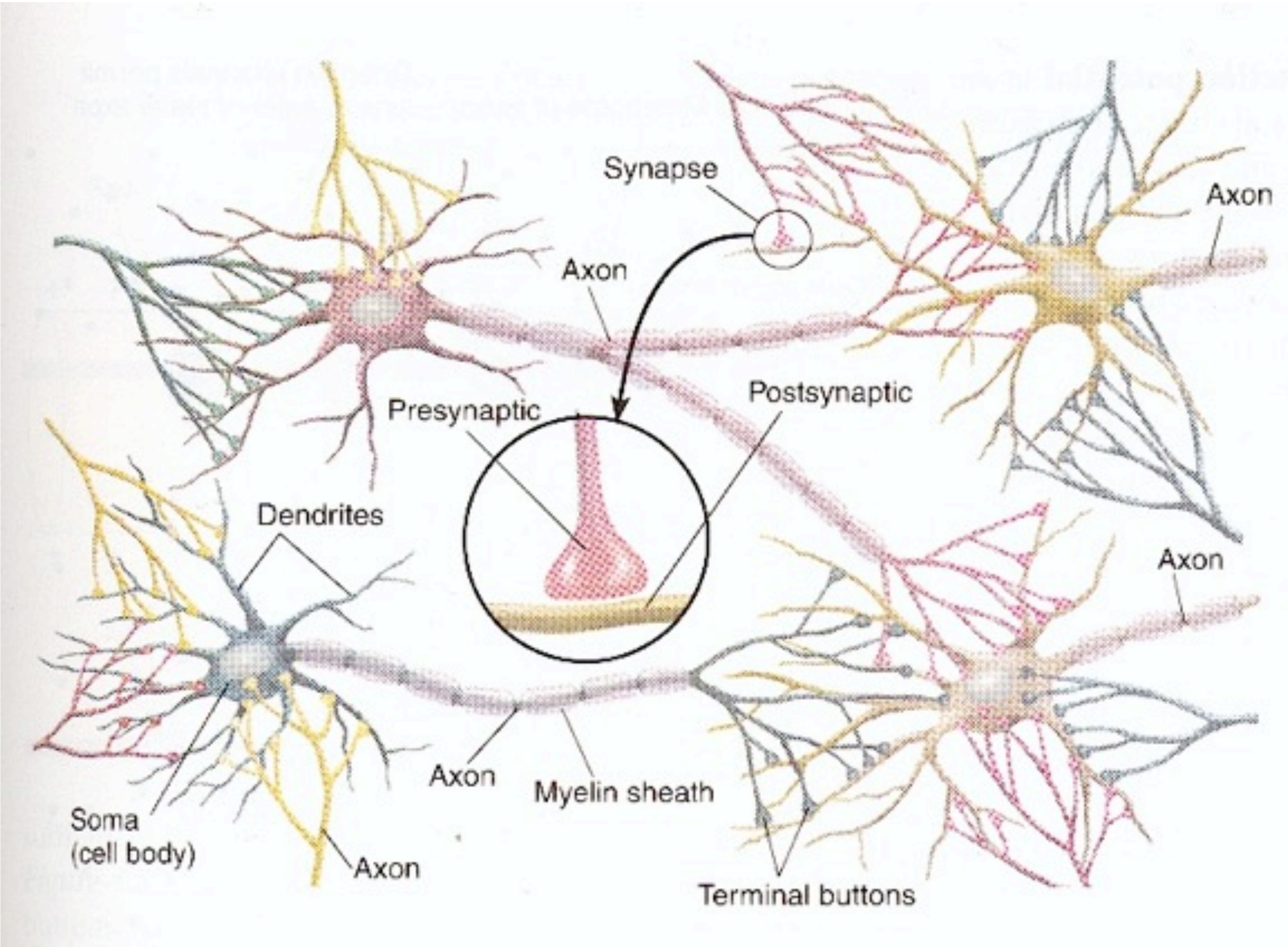


Figure 12. Basic cell types in the monkey cerebral cortex. Left: spiny neurons that include pyramidal cells and stellate cells (A). Spiny neurons utilize the neurotransmitter glutamate (Glu). Right: smooth cells that use the neurotransmitter GABA. B, cell with local axon arcades; C, double bouquet cell; D, H, basket cells; E, chandelier cells; F, bitufted, usually peptide-containing cell; G, neurogliaform cell.

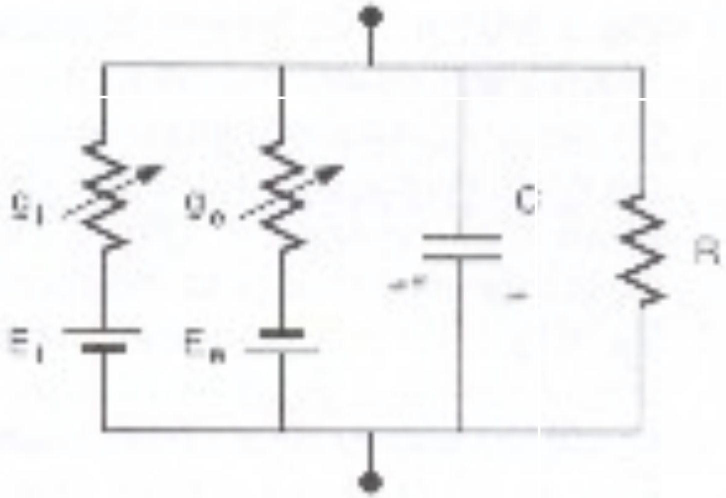
Source: http://webvision.med.utah.edu/

Neural Circuits



Source: Modified from Jody Culham's web slides

Membrane with excitatory and inhibitory synapses

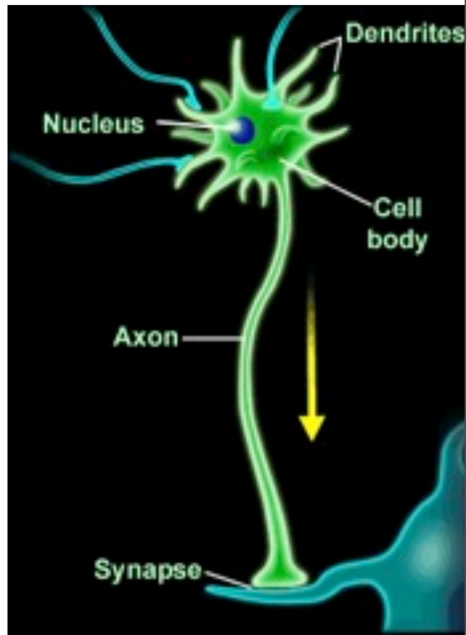


$$C\frac{dV}{dt}+g_i(V-E_i)+g_e(V-E_e)+g_0(V-V_{rest})=0$$

and with $\frac{dV}{dt}\approx 0$, $E_i\approx 0$, $V_{rest}\approx 0$, $\tilde{g}_e=\frac{g_e}{g_0}$ and $\tilde{g}_i=\frac{g_i}{g_0}$ we obtain

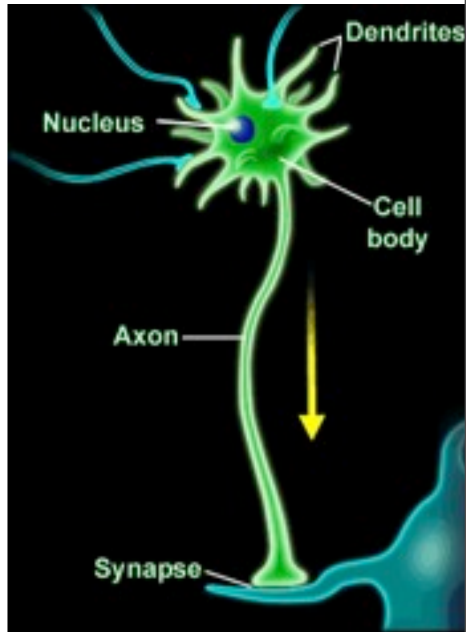
$$V\approx E_e\frac{\tilde{g}_e}{1+\tilde{g}_e+\tilde{g}_i}$$

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

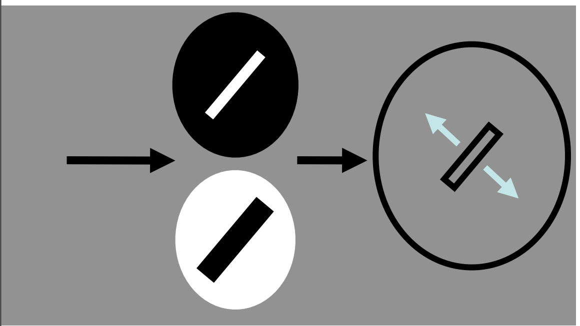
The Ventral Stream

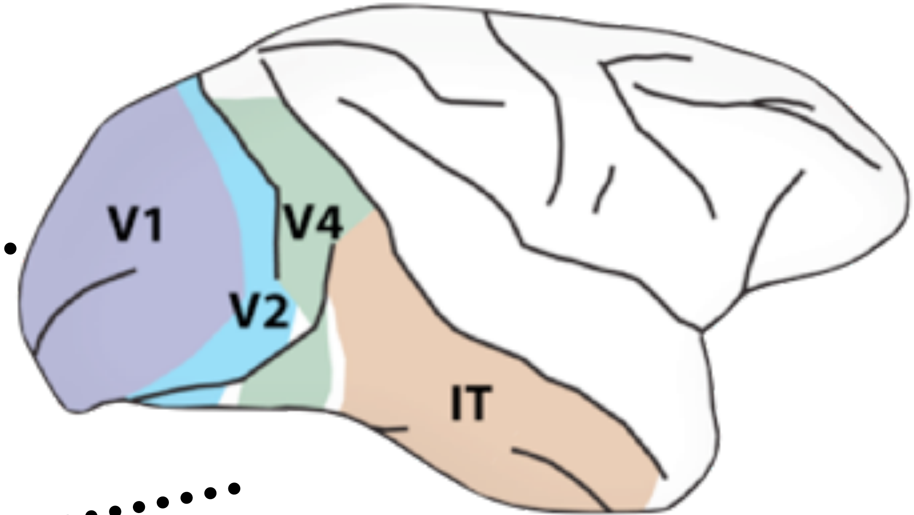


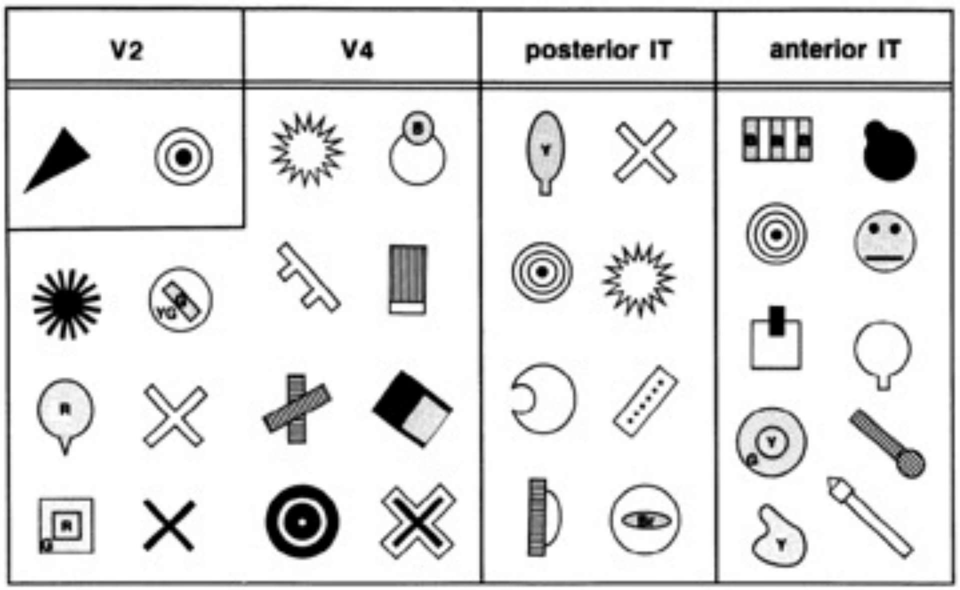
- Human Brain
  - 10^{10}-10^{11} neurons (~1 million flies ☺)
  - 10^{14}- 10^{15} synapses

- Ventral stream in rhesus monkey
  - ~10^{9} neurons in the ventral stream (350 10^{6} in each emisphere)
  - ~15 10^{6} neurons in AIT (Anterior InferoTemporal) cortex

The Ventral Stream







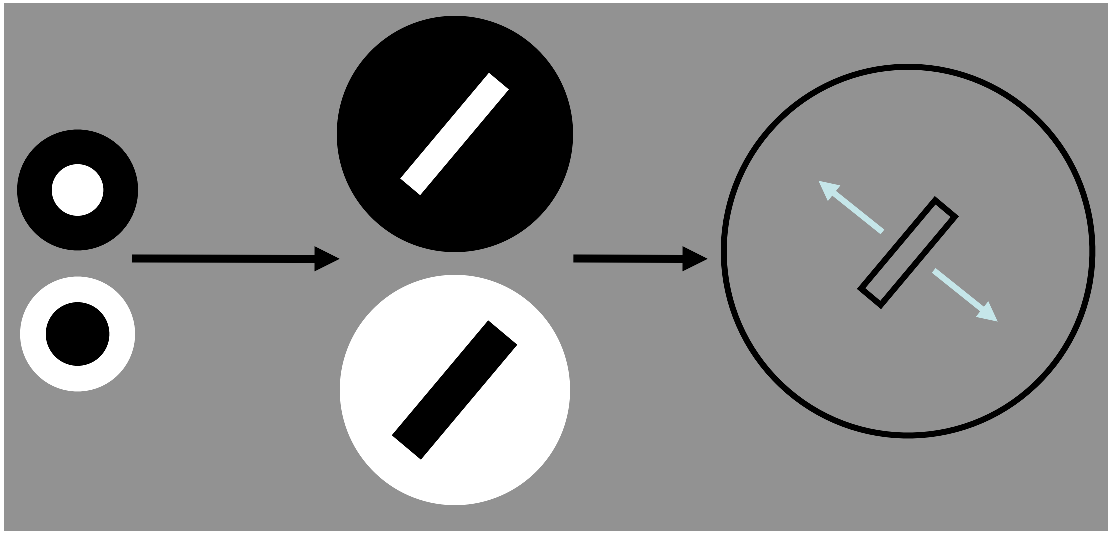
The ventral stream hierarchy: V1, V2, V4, IT

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

V1: hierarchy of simple and complex cells

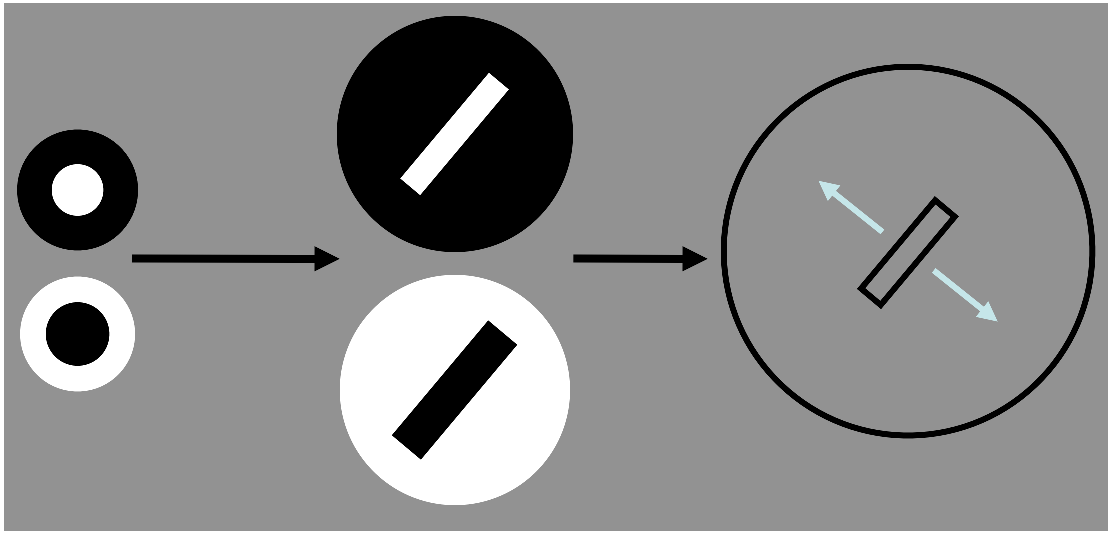
LGN-type cells | Simple cells | Complex cells

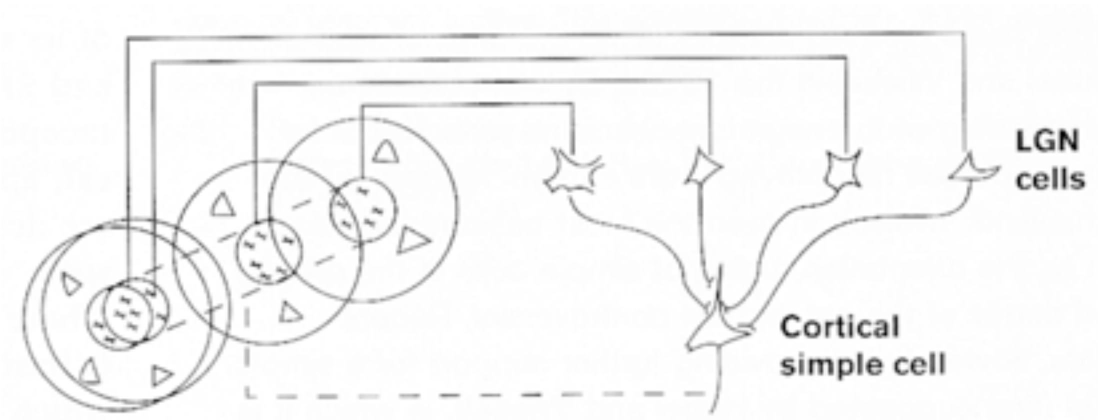


(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

LGN-type cells | Simple cells | Complex cells

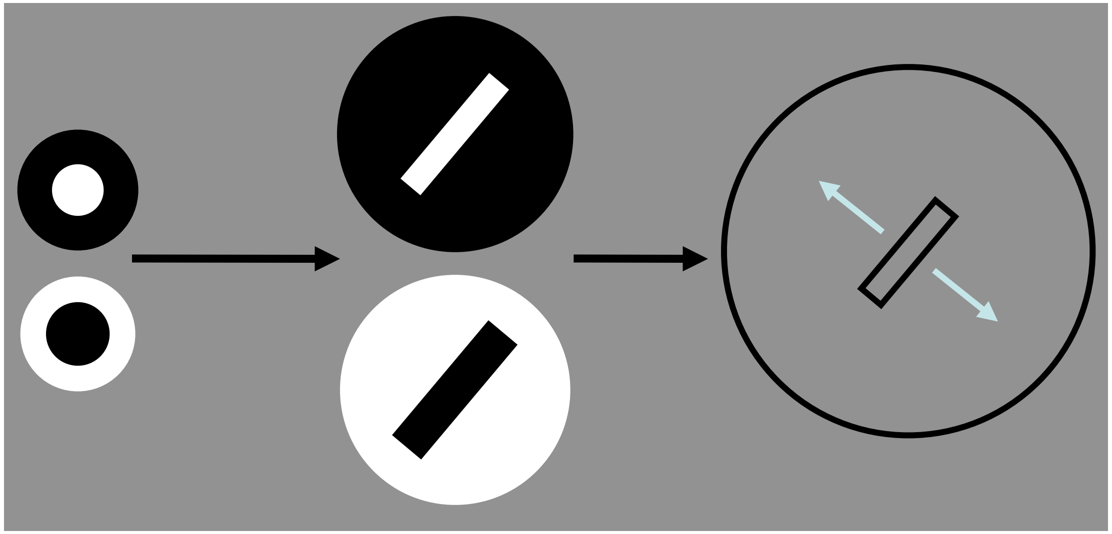


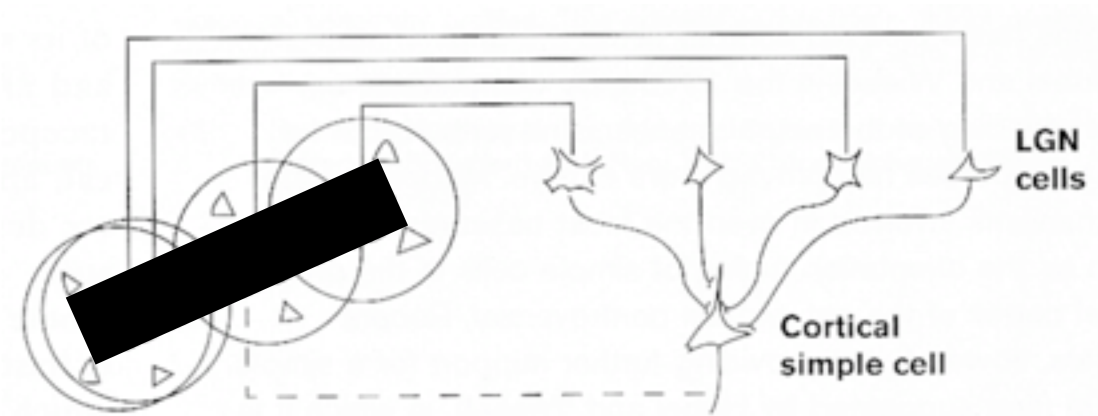


(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

LGN-type cells | Simple cells | Complex cells

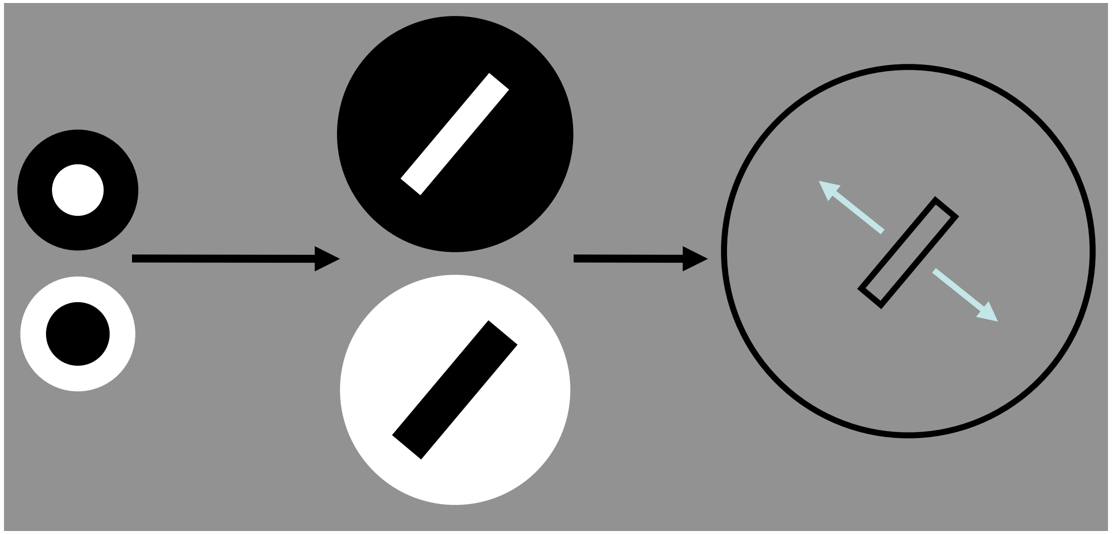




(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

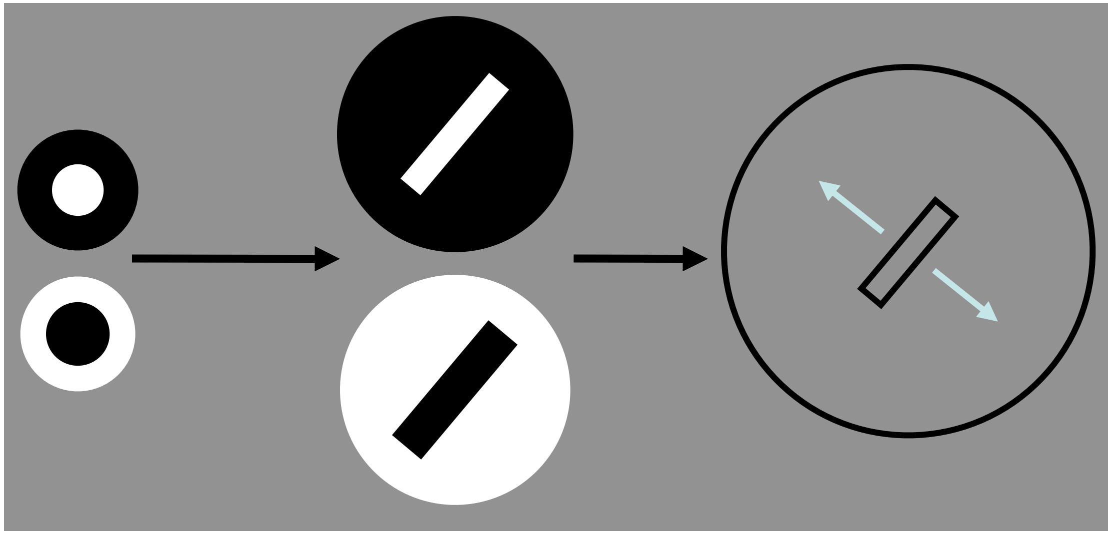
LGN-type cells | Simple cells | Complex cells

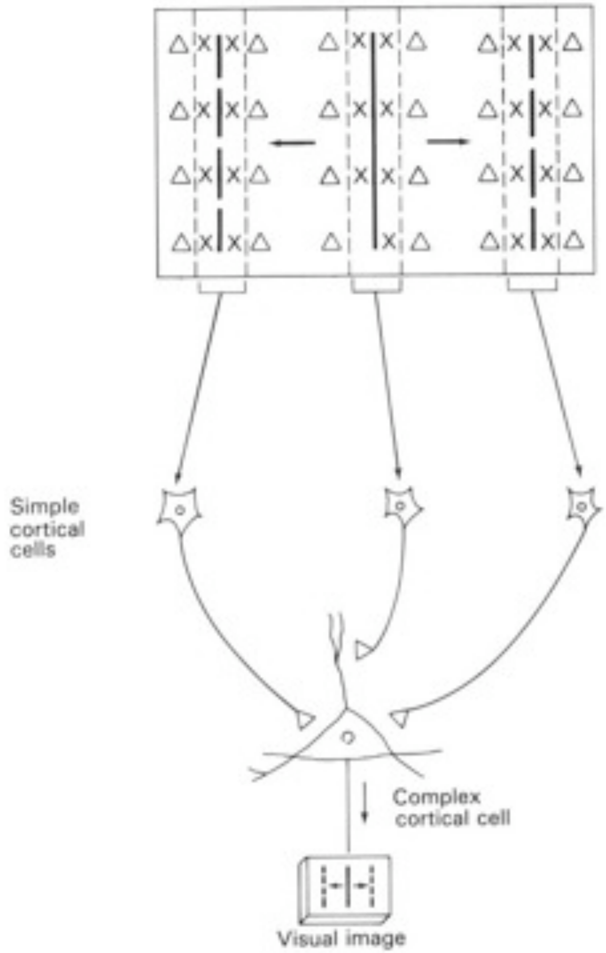


(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

LGN-type cells   Simple cells   Complex cells

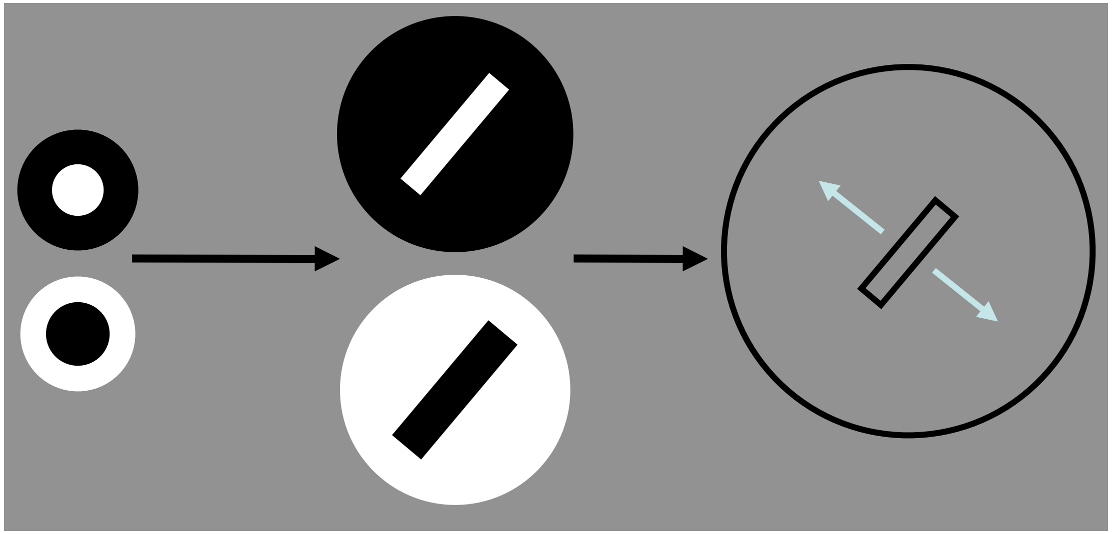


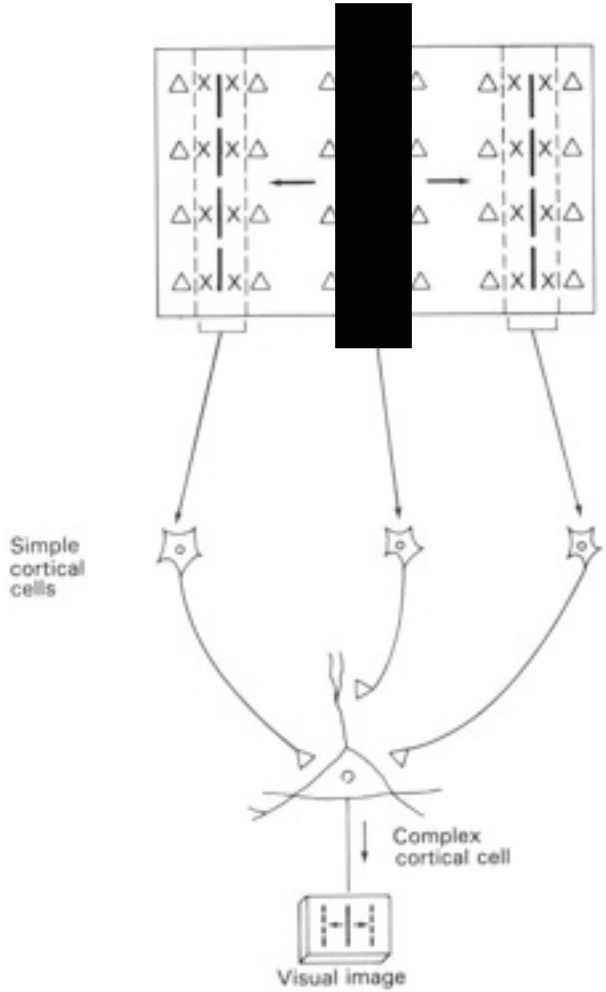


(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

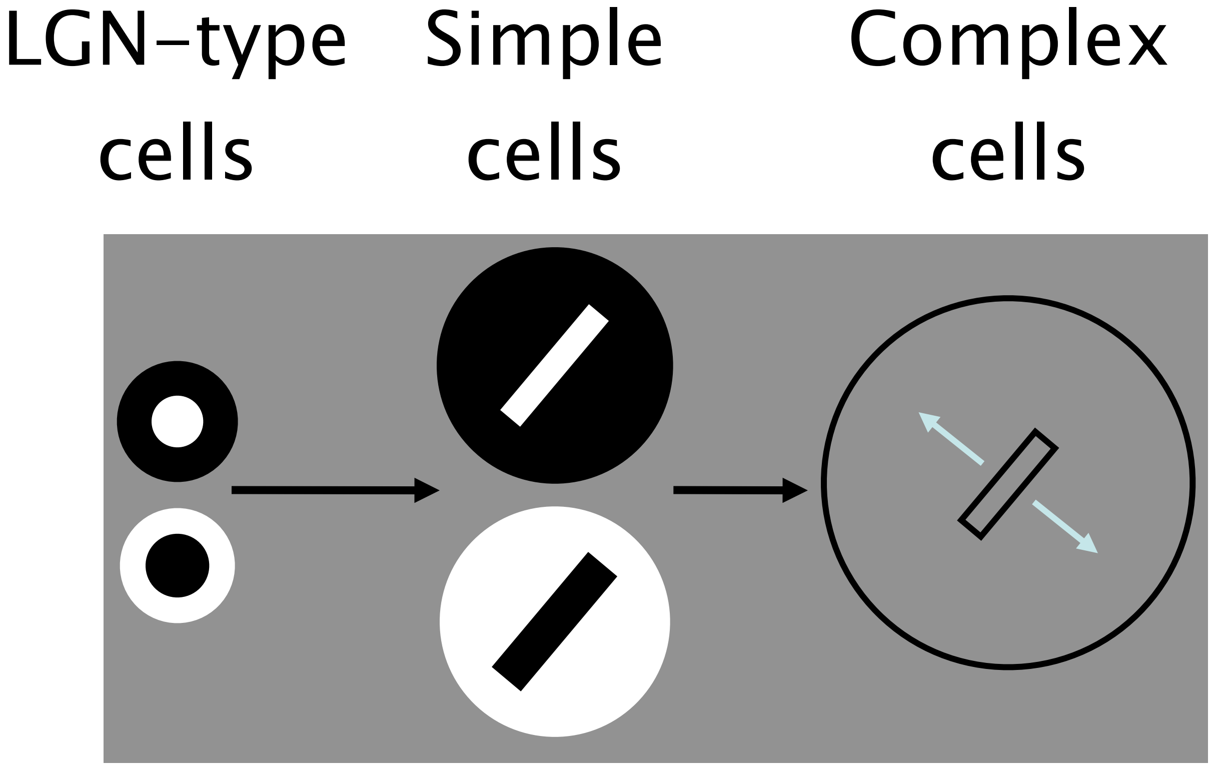
LGN-type cells | Simple cells | Complex cells

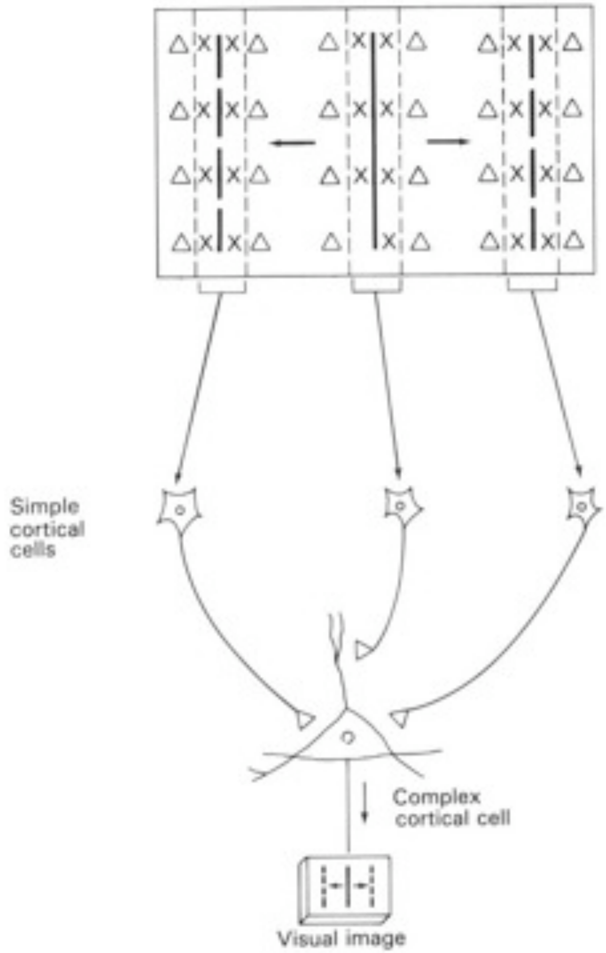




(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

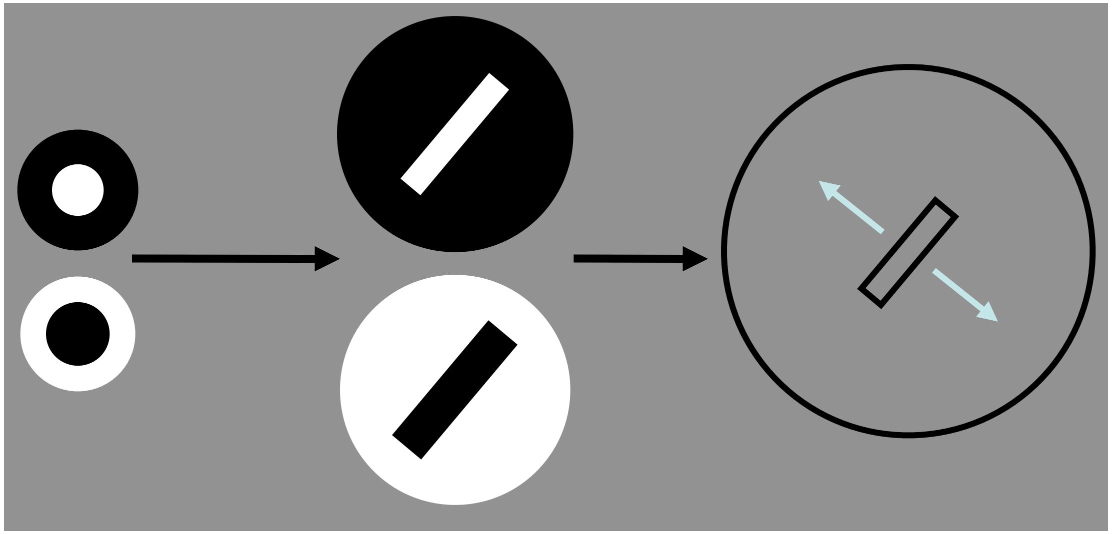


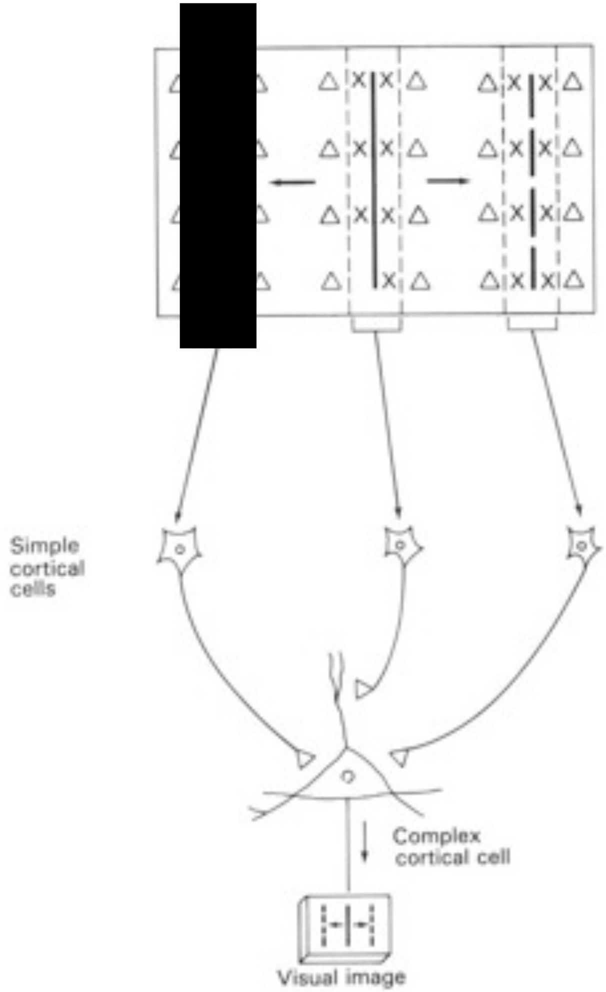


(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

LGN-type cells | Simple cells | Complex cells

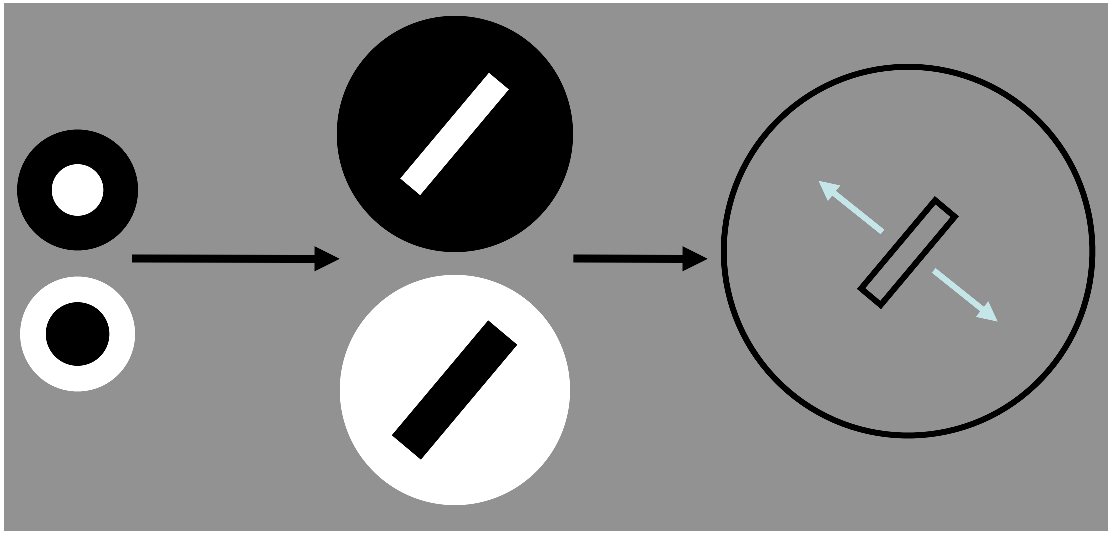


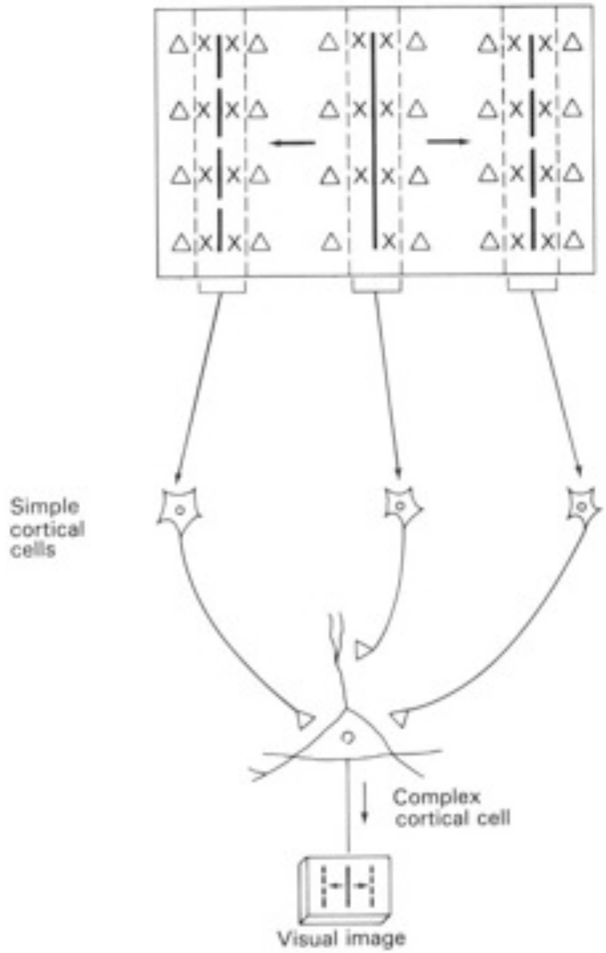


(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

LGN-type cells　Simple cells　Complex cells

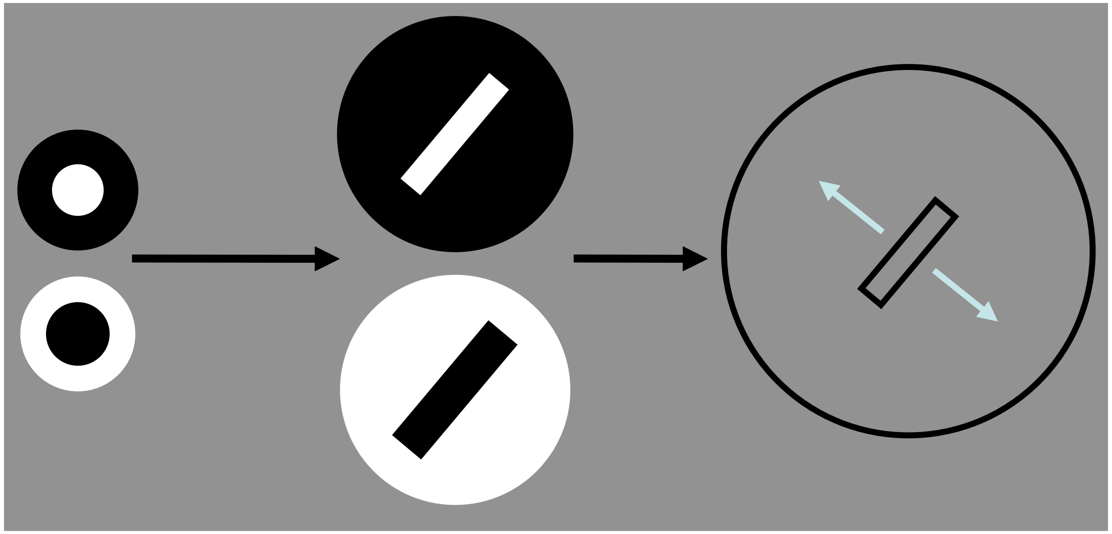


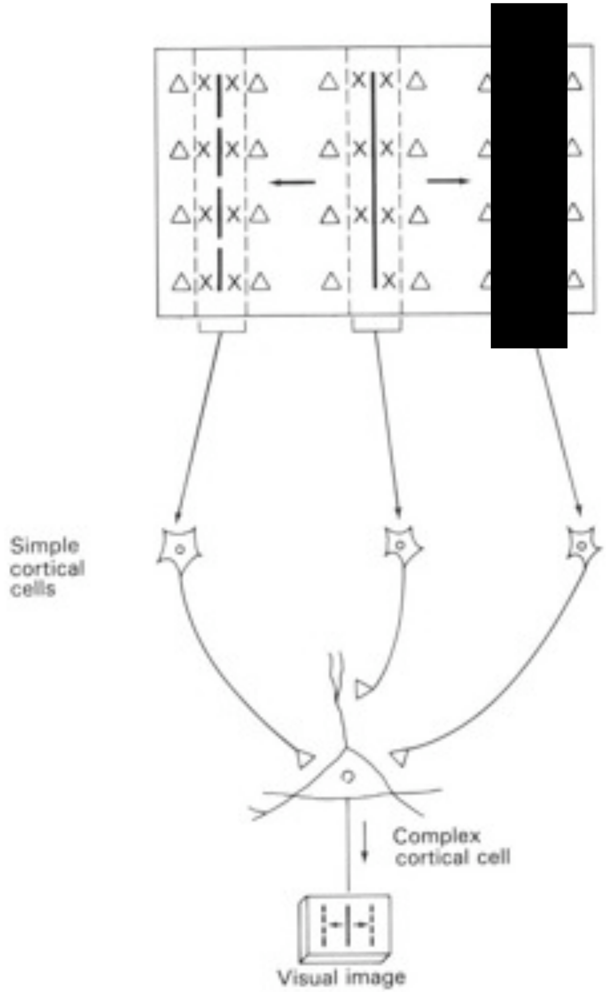


(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

LGN-type cells | Simple cells | Complex cells

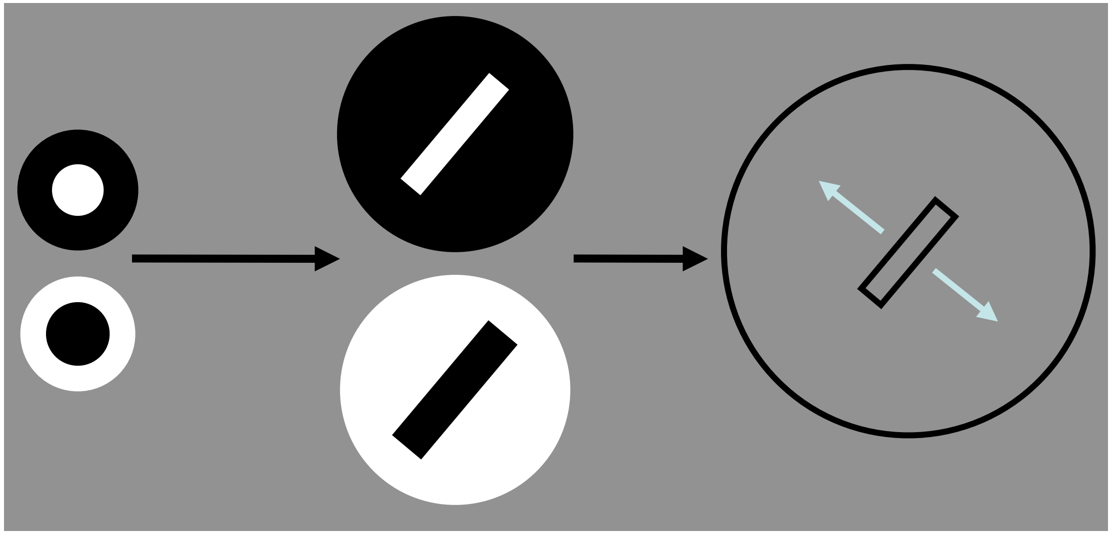


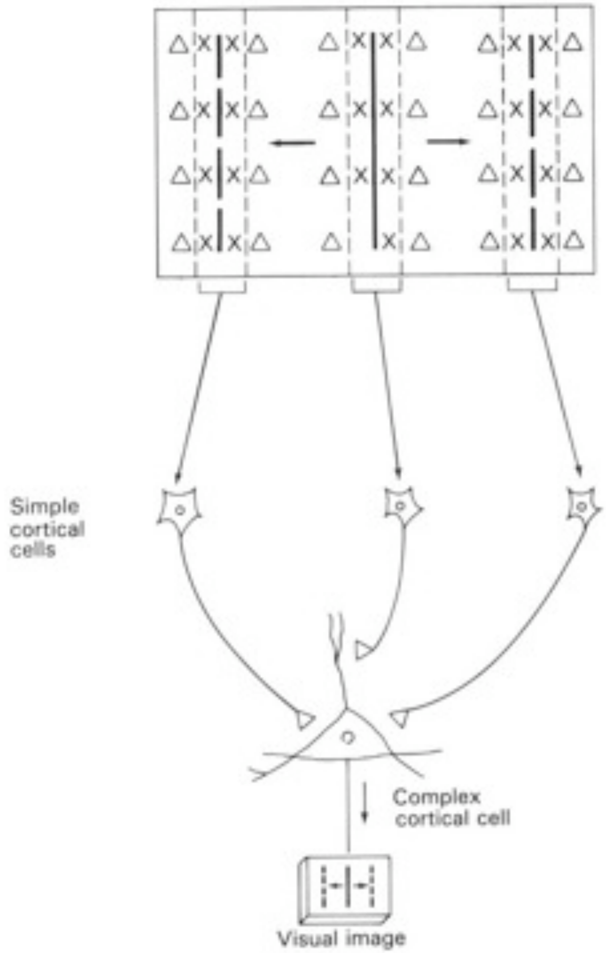


(Hubel & Wiesel 1959)

V1: hierarchy of simple and complex cells

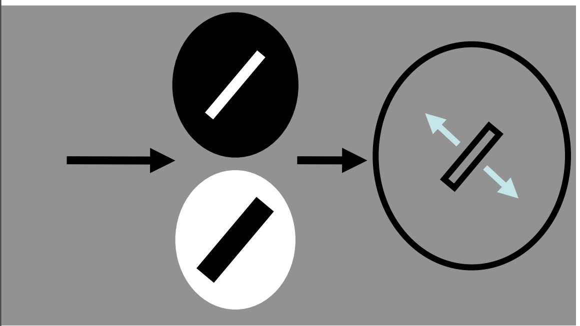
LGN-type cells | Simple cells | Complex cells

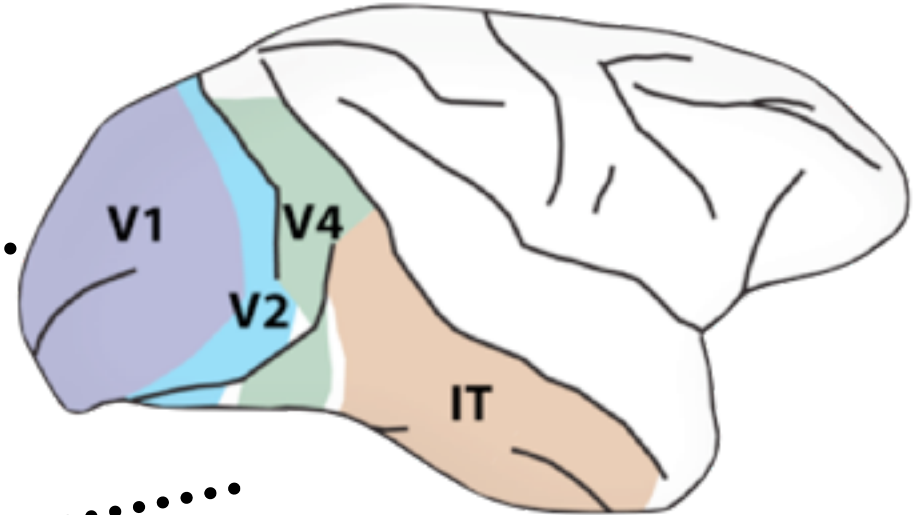


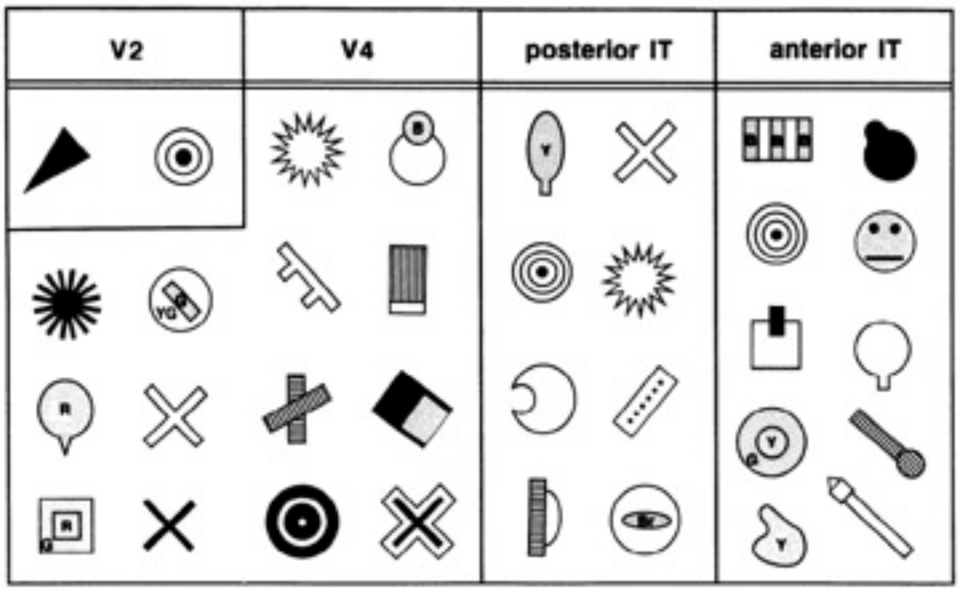


(Hubel & Wiesel 1959)

The Ventral Stream





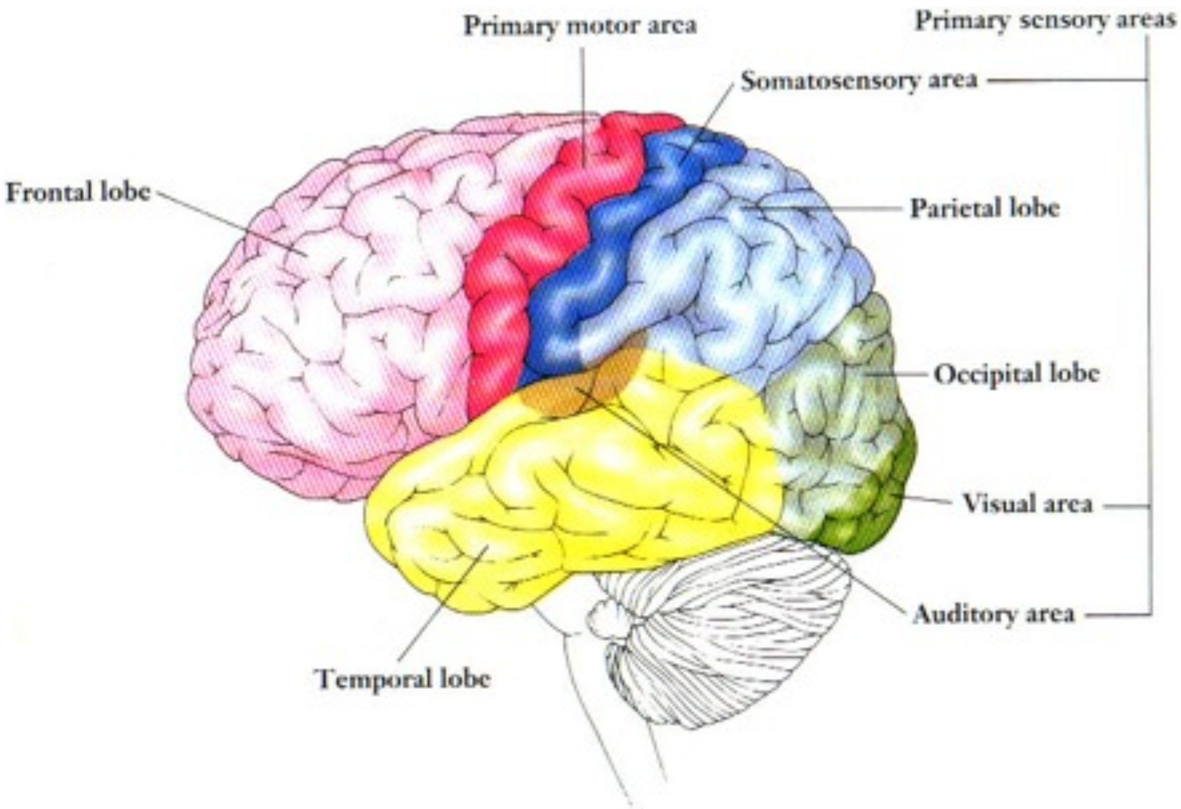


The ventral stream hierarchy: V1, V2, V4, IT

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

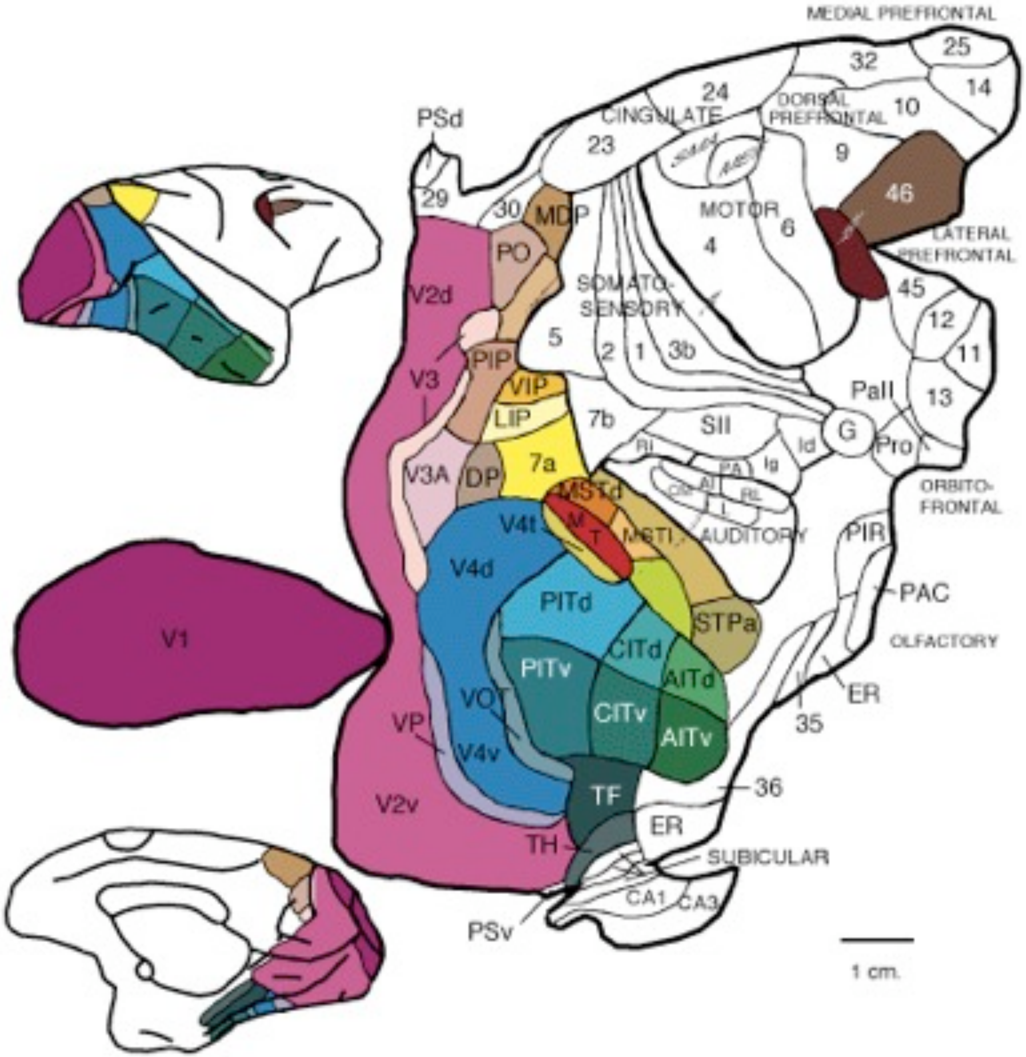
Kobatake & Tanaka, 1994

Gross Brain Anatomy



A large percentage of the cortex devoted to vision

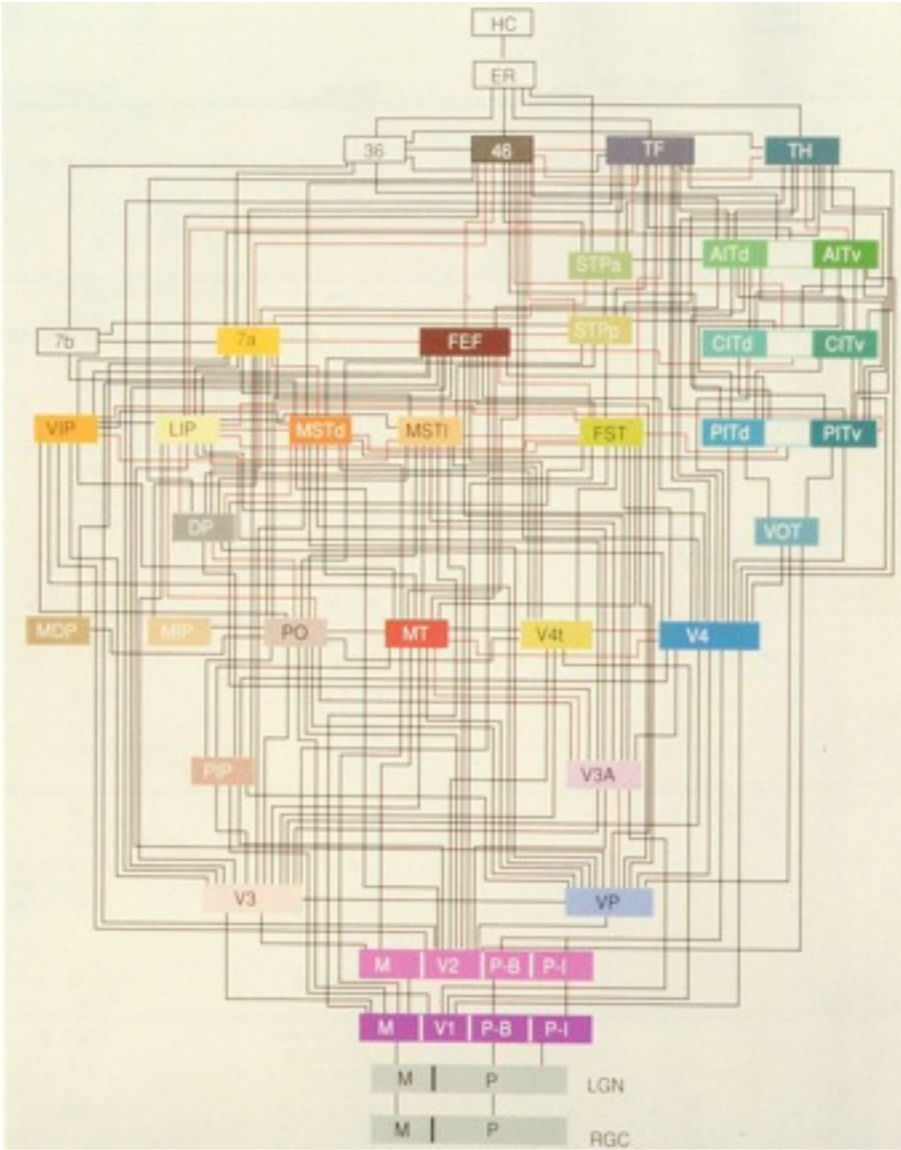
The Visual System



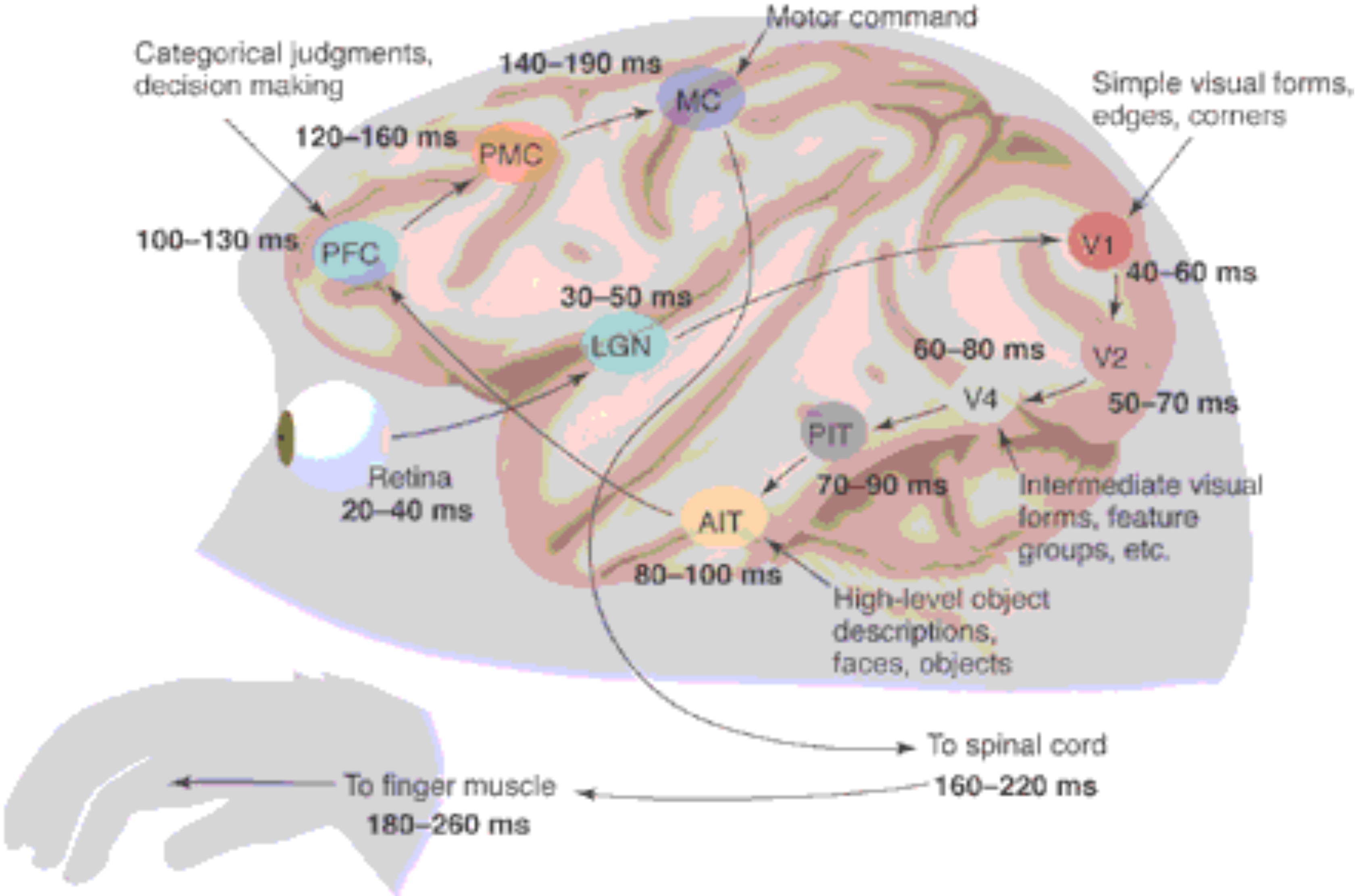
[Van Essen & Anderson, 1990]

The visual system

- Over 30 visual areas
- Over 300 cortico-cortical pathways

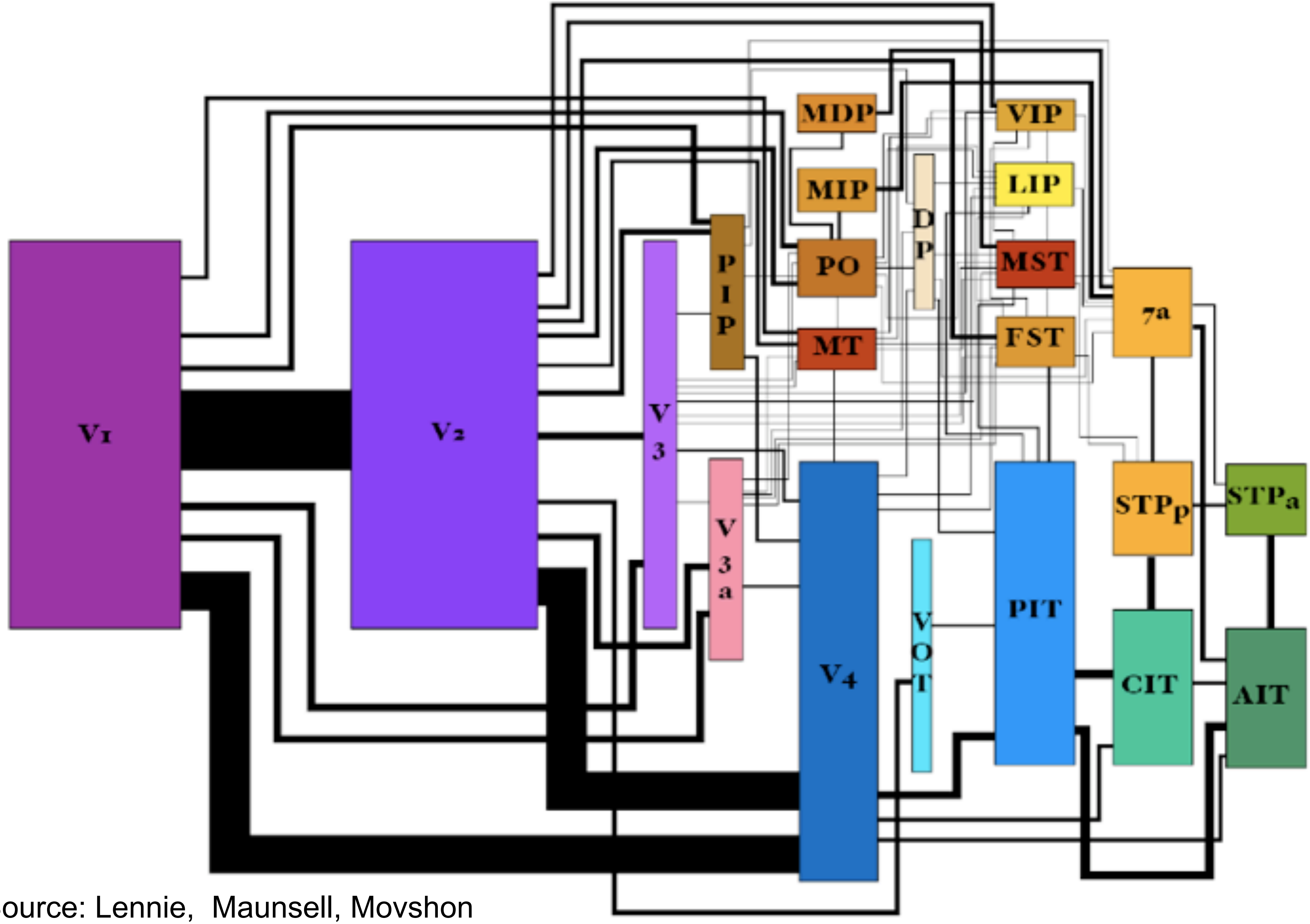


(Felleman & VanEssen 1991)



(Thorpe and Fabre-Thorpe, 2001)

The ventral stream



Source: Lennie, Maunsell, Movshon

1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Feedforward hierarchical models
    - Ventral stream model in more details (Jim Mutch)
    - Dorsal stream model (Hueihan Jhuang)

From HMAX to the present model

How the new version of the model evolved from the original one

1. **The two key operations:** Operations for selectivity and invariance, originally computed in a simplified and idealized form (i.e., a multivariate Gaussian and an exact max, see Section 2) have been replaced by more plausible operations, normalized dot-product and softmax

2. **S1 and C1 layers:** In [Serre and Riesenhuber, 2004] we found that the S1 and C1 units in the original model were too broadly tuned to orientation and spatial frequency and revised these units accordingly. In particular at the S1 level, we replaced Gaussian derivatives with Gabor filters to better fit parafoveal simple cells' tuning properties. We also modified both S1 and C1 receptive field sizes.

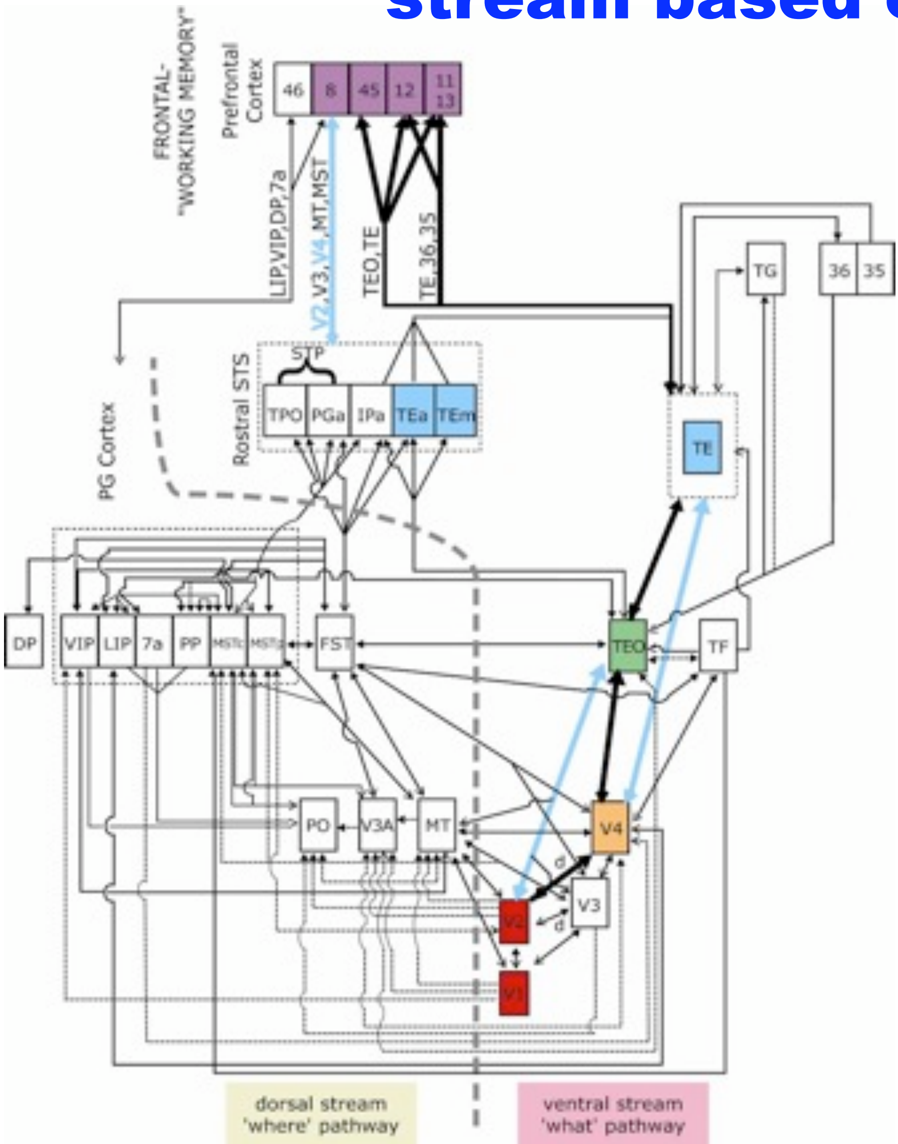
3. **S2 layers:** They are now learned from natural images. S2 units are more complex than the old ones (simple 2 °— 2 combinations of orientations). The introduction of learning, we believe, has b een the key factor for the model to achieve a high-level of performance on natural images, see [Serre et al., 2002].

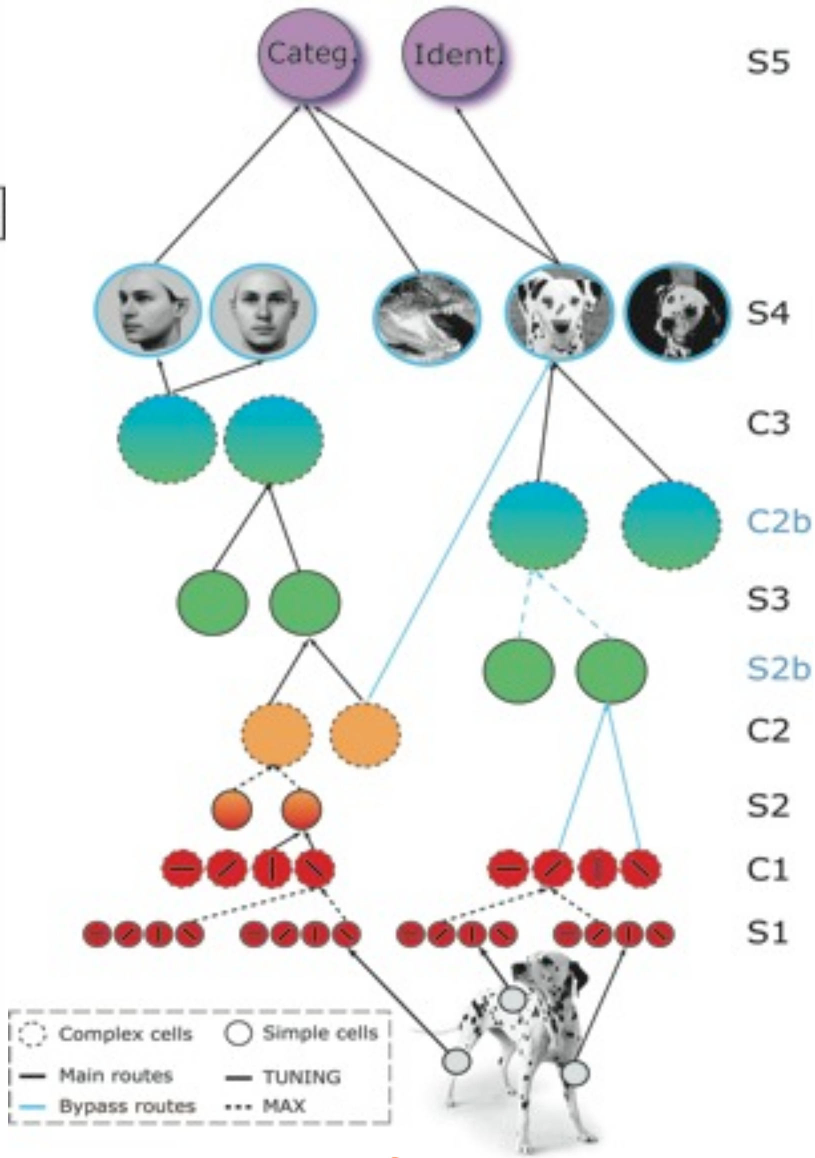
4. **C2 layers:** Their receptive field sizes, as well as range of invariances to scale and position have been decreased so that C2 units now better fit V4 data.

5. **S3 and C3 layers:** They were recently added and constitute the top-most layers of the model along with the S2b and C2b units (see Section 2 and above). The tuning of the S3 units is also learned from natural images.

6. **S2b and C2b layers:** We added those two layers to account for the bypass route (that projects directly from V1/V2 to PIT, thus bypassing V4 [see Nakamura et al., 1993]).

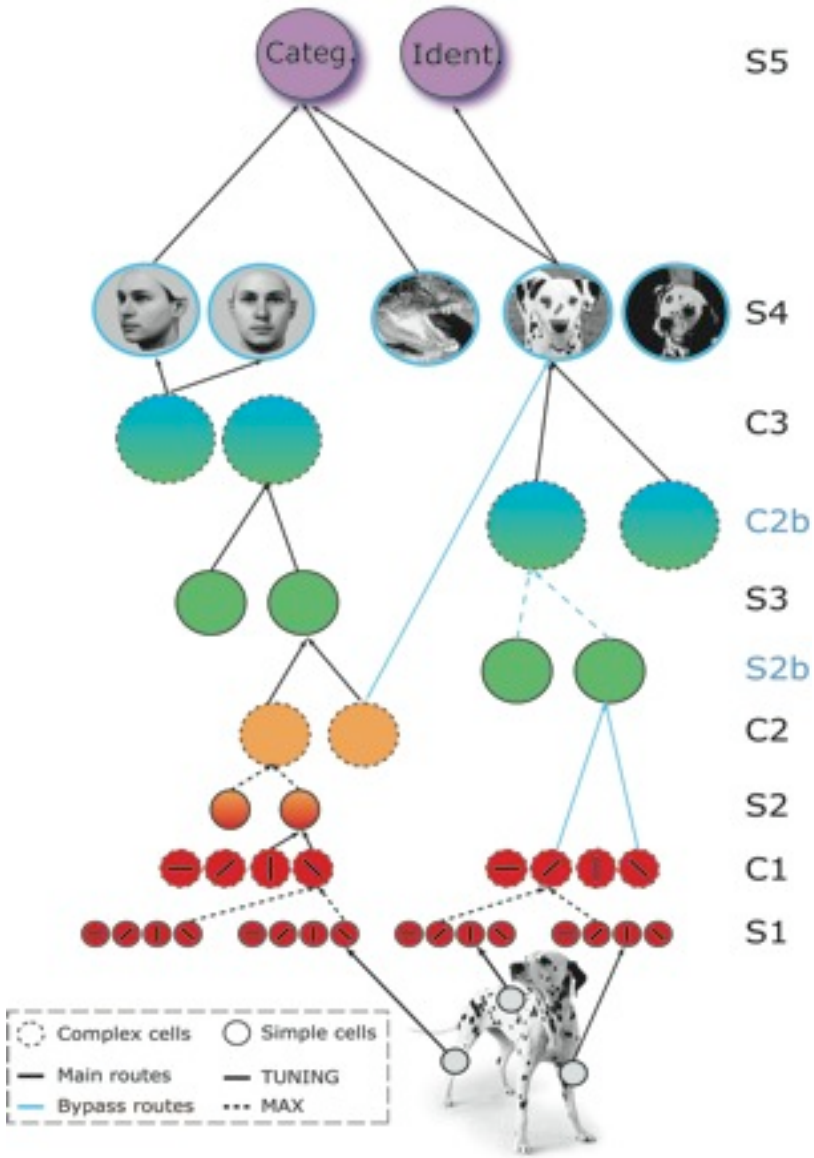
A hierarchical feedforward model of the ventral stream based on neural data





[software available online]

Model of Visual Recognition (millions of units) based on neuroscience of cortex



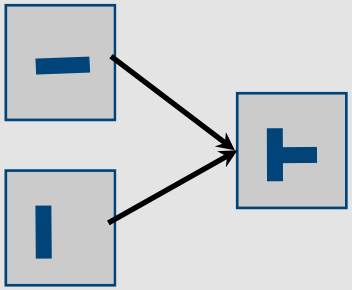
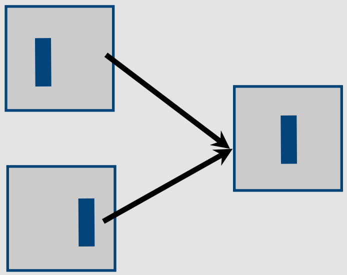
- It is in the family of "Hubel-Wiesel" models (Hubel & Wiesel, 1959; Fukushima, 1980; Oram & Perrett, 1993, Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe, 2002; Ullman et al., 2002; Mel, 1997; Wersing and Koerner, 2003; LeCun et al 1998; Amit & Mascaro 2003; Deco & Rolls 2006…)

- As a biological model of object recognition in the ventral stream – from V1 to PFC -- it is *perhaps* the most quantitative and faithful to known neuroscience

- A model which "copies" the neuroscience. Millions of (model) neurons.

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

Two key computations, suggested by physiology

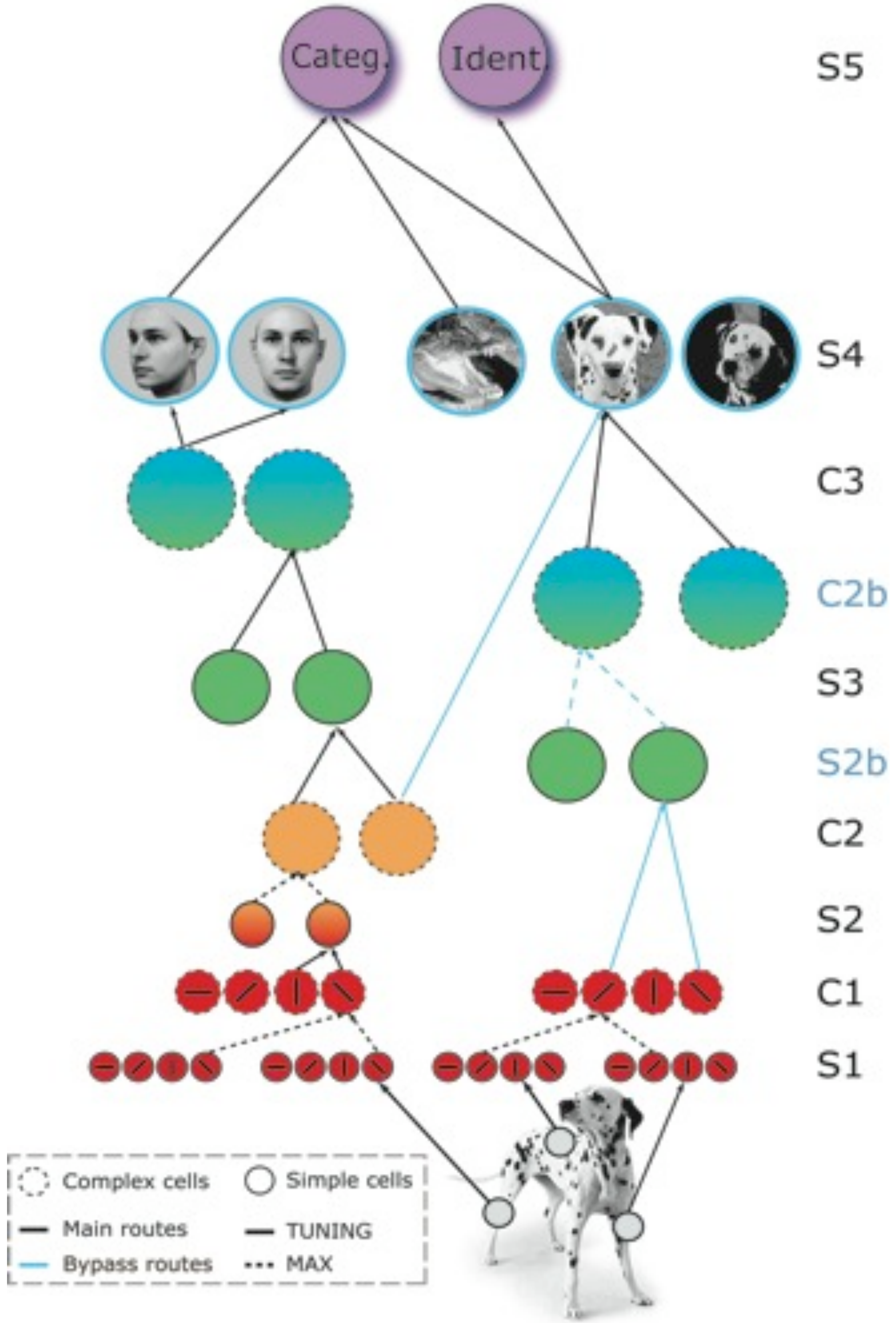
Unit	Pooling	Computation	Operation
Simple		Selectivity / template matching	Gaussian-tuning / AND-like
Complex		Invariance	Soft-max / or-like

➢Gaussian-like tuning operation (and-like)

➢Simple units

➢Max-like operation (or-like)

➢Complex units

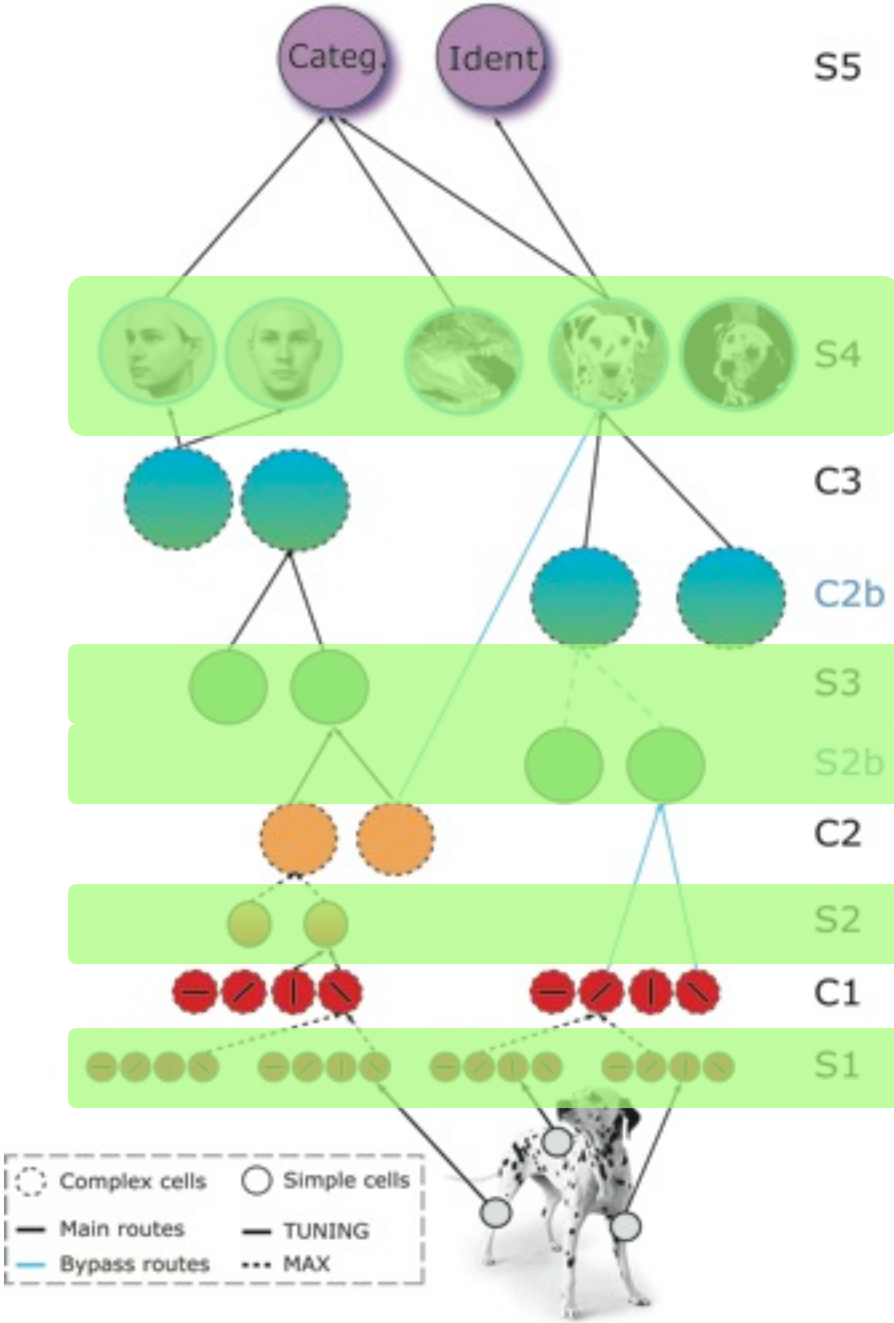


➢Gaussian-like tuning operation (and-like)

➢Simple units

➢Max-like operation (or-like)

➢Complex units

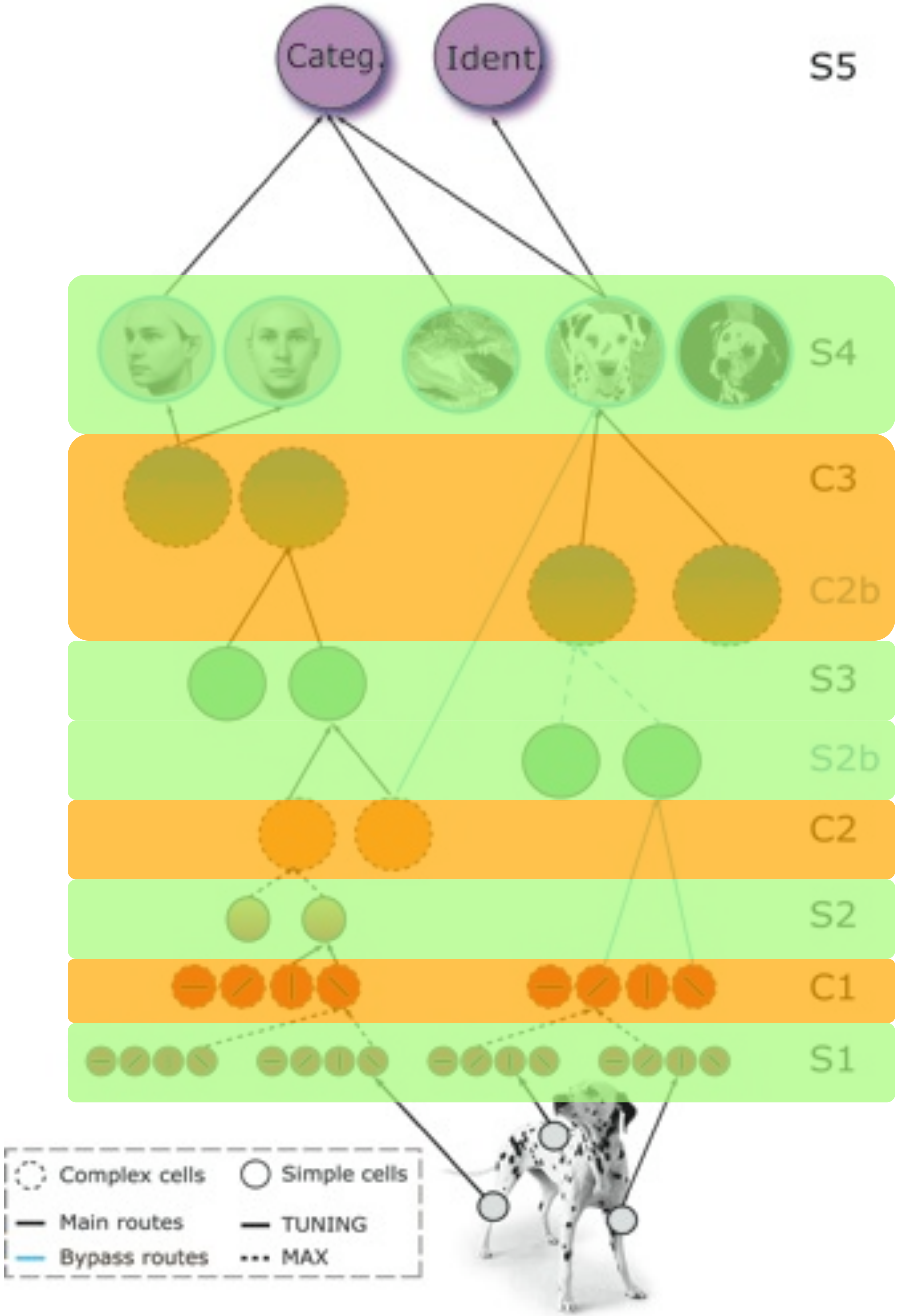


➢Gaussian-like tuning operation (and-like)

➢Simple units

➢Max-like operation (or-like)

➢Complex units



Two operations (~OR, ~AND): disjunctions of conjunctions

- Tuning operation (Gaussian-like, AND-like) $y = e^{-|x-w|^2}$

  or

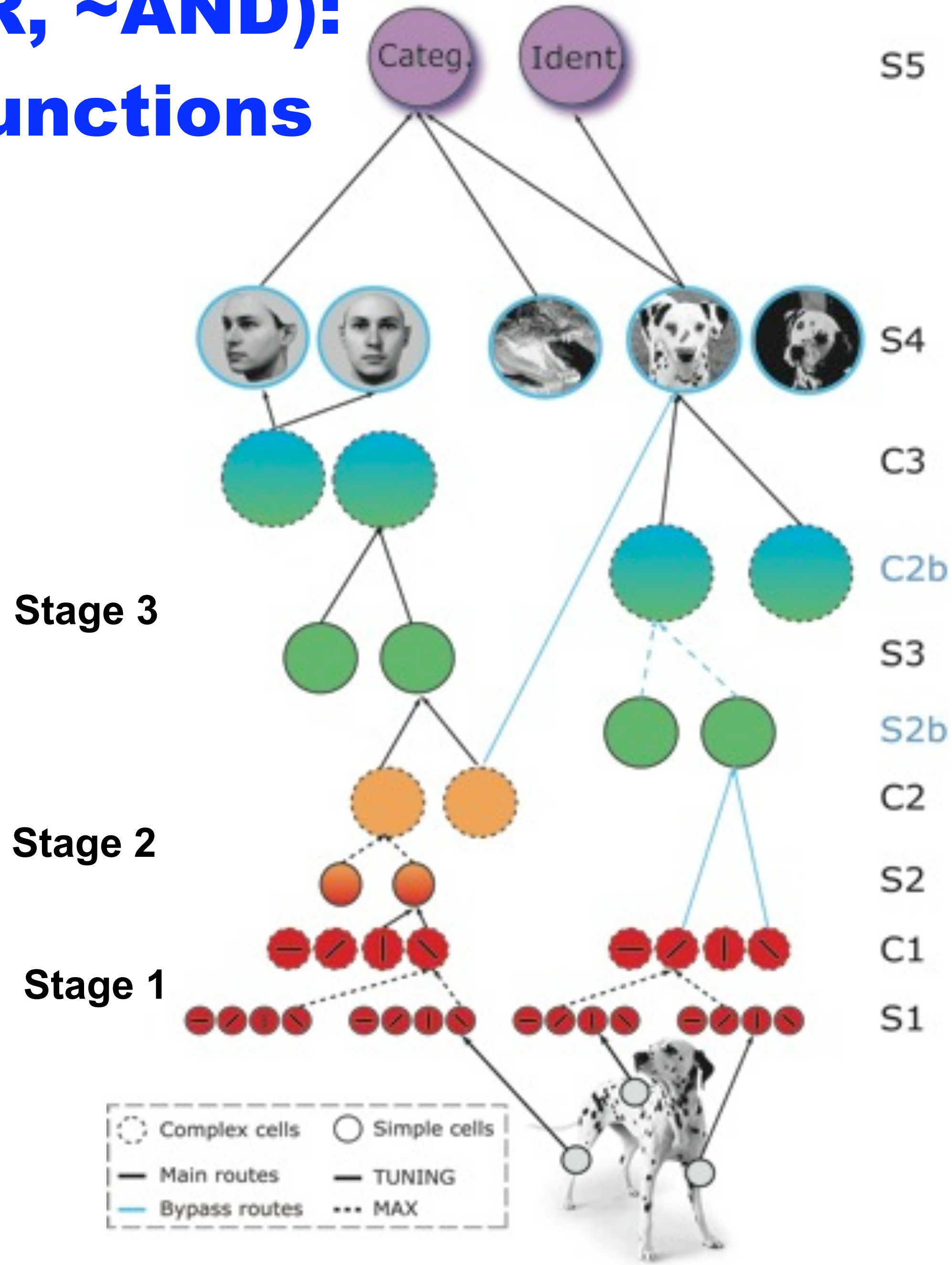
  $$y \sim \frac{x \bullet w}{|x|}$$

- Simple units

- Max-like operation (OR-like)

  $$y = \max\{x1, x2, \ldots\}$$

- Complex units

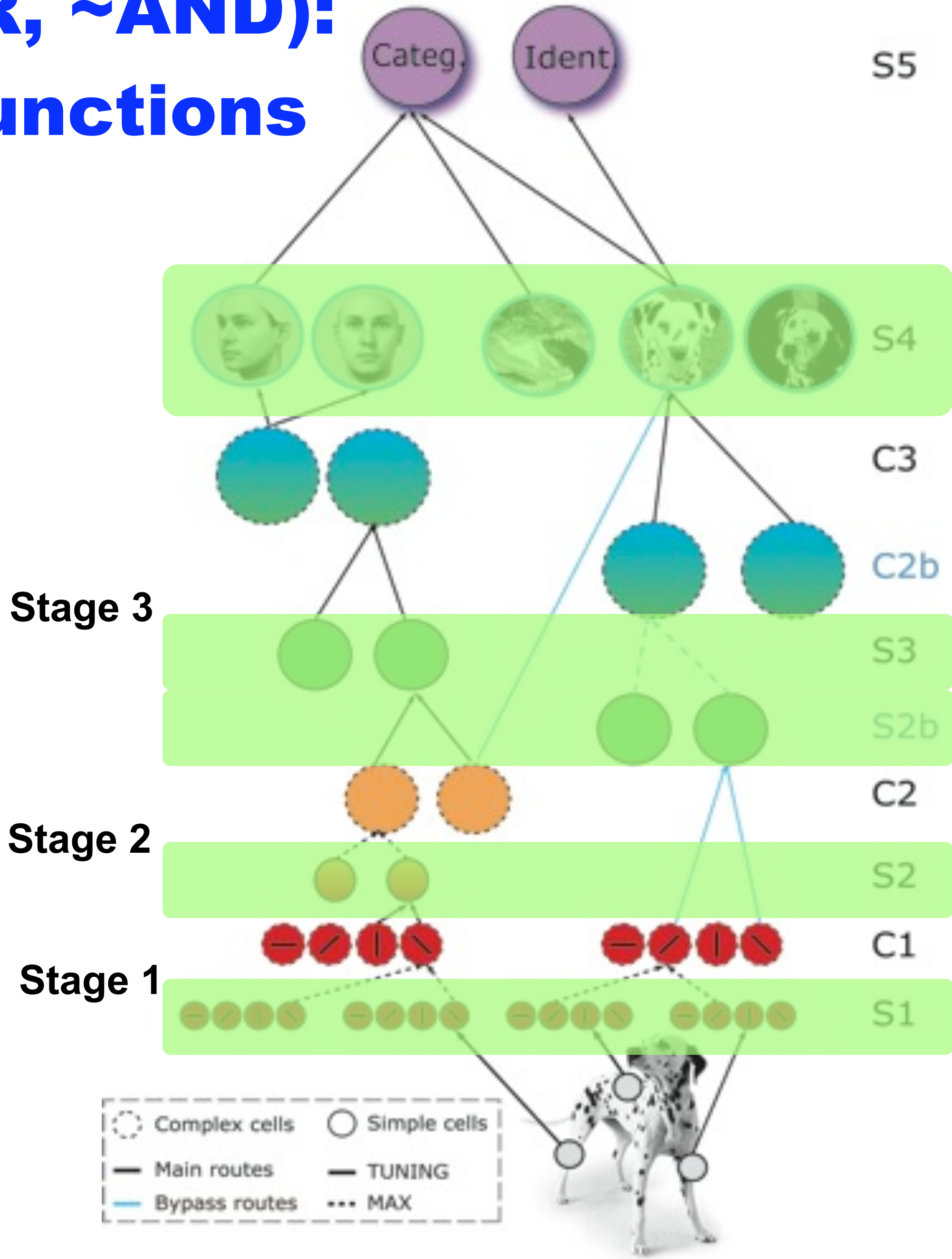


Each operation ~microcircuits of ~100 neurons

Two operations (~OR, ~AND): disjunctions of conjunctions

- Tuning operation (Gaussian-like, AND-like) $$y = e^{-|x-w|^2}$$ *or* $$y \sim \frac{x \bullet w}{|x|}$$
- Simple units

- Max-like operation (OR-like) $$y = \max\{x1, x2, \ldots\}$$
- Complex units



Each operation ~microcircuits of ~100 neurons

Two operations (~OR, ~AND): disjunctions of conjunctions

- Tuning operation (Gaussian-like, AND-like) $y = e^{-|x-w|^2}$

  or

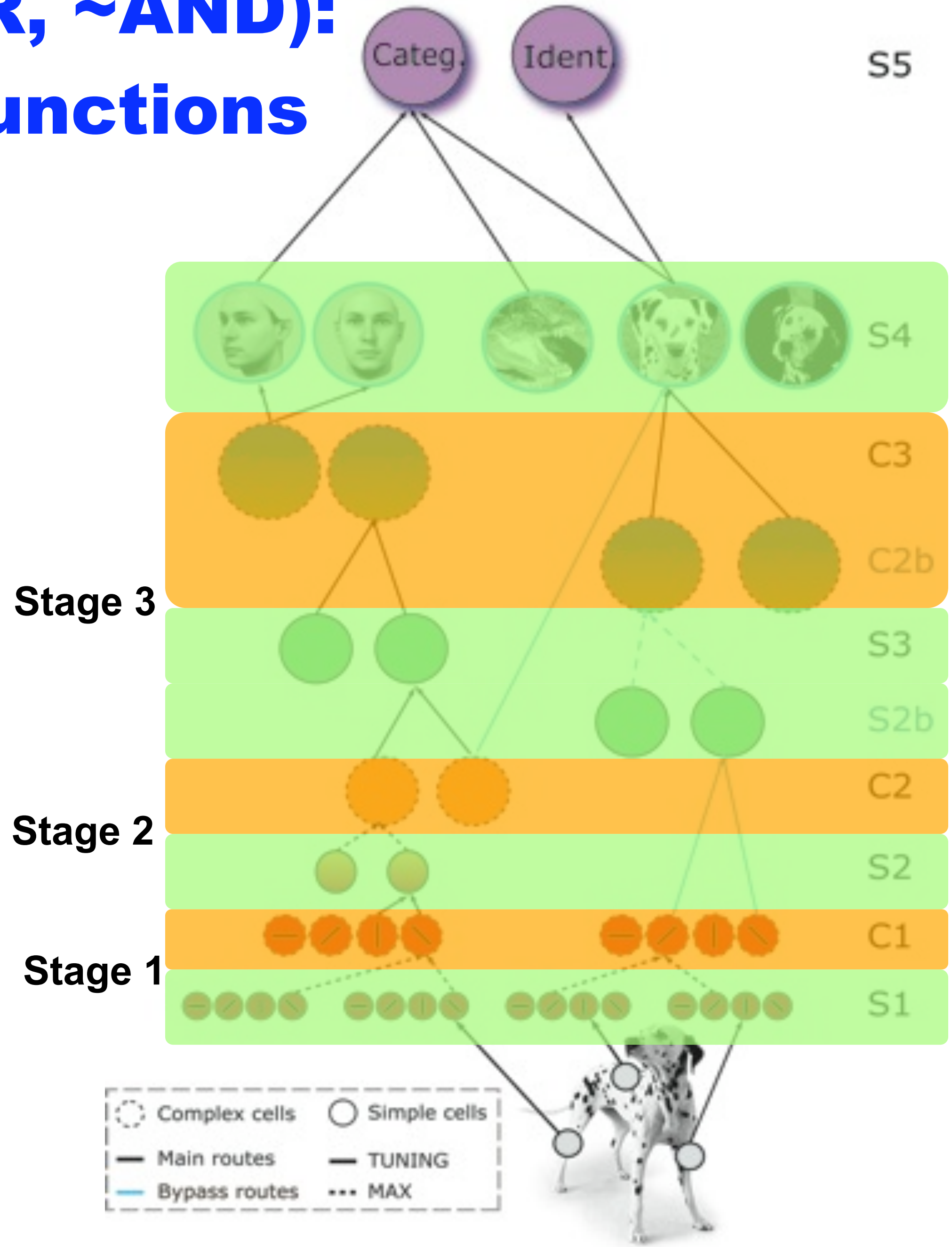
  $$y \sim \frac{x \bullet w}{|x|}$$

- Simple units

- Max-like operation (OR-like)

  $$y = \max\{x1, x2, \ldots\}$$

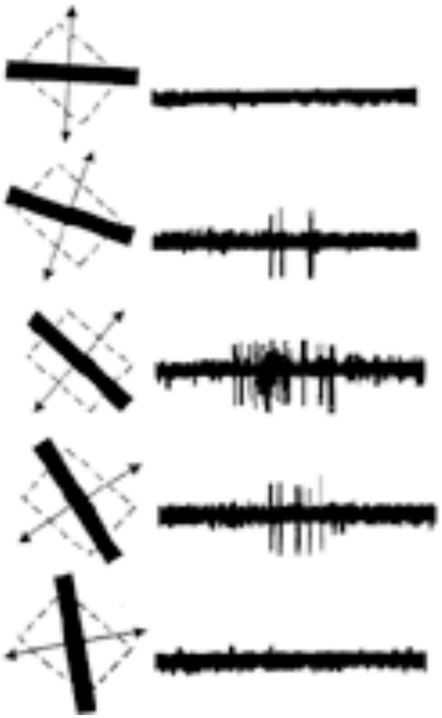
- Complex units



Each operation ~microcircuits of ~100 neurons

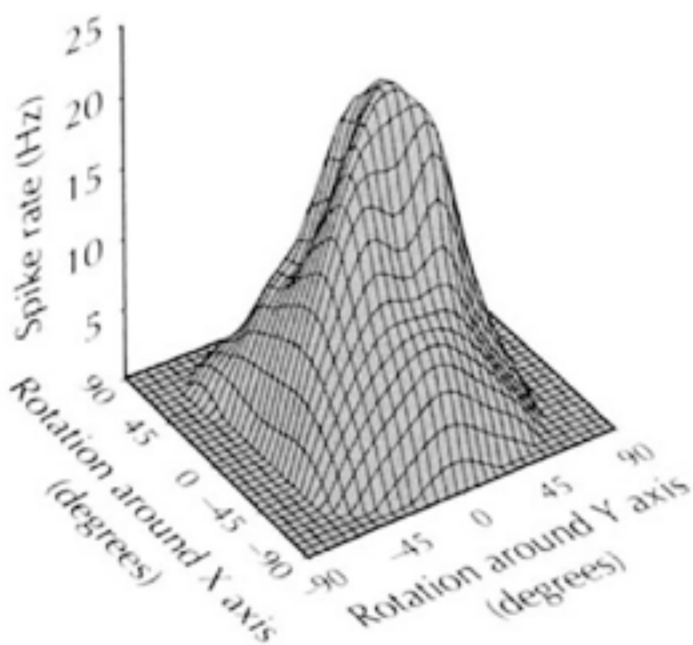
Gaussian tuning

Gaussian tuning in V1 for orientation



Hubel & Wiesel 1958

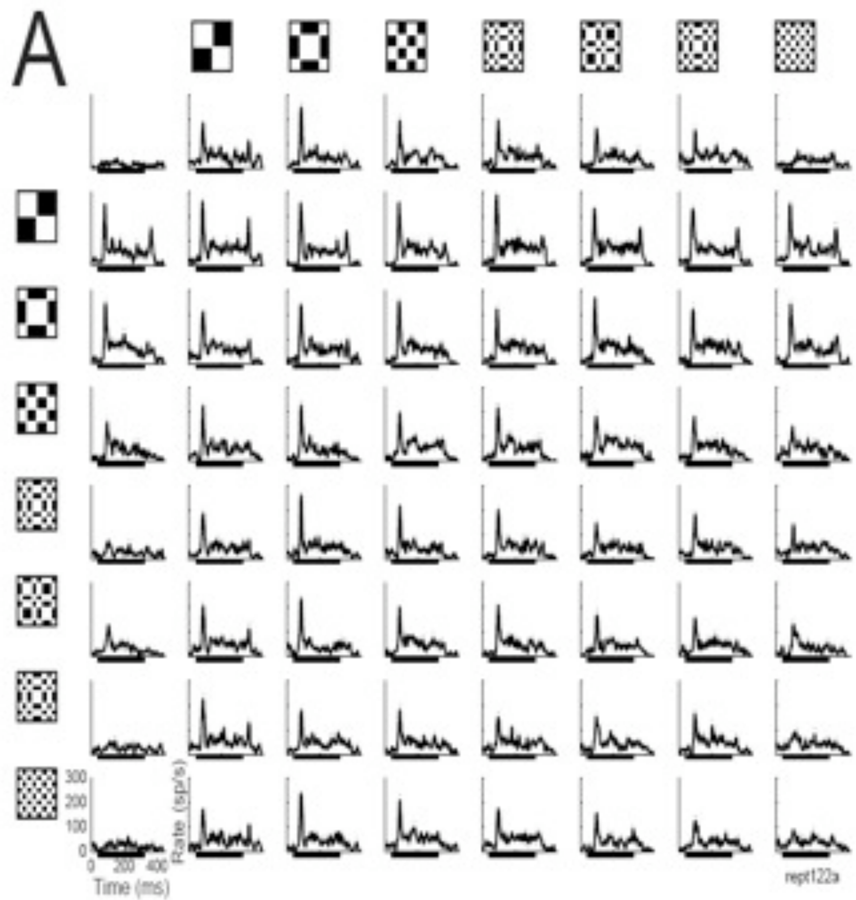
Gaussian tuning in IT around 3D views



Logothetis Pauls & Poggio 1995

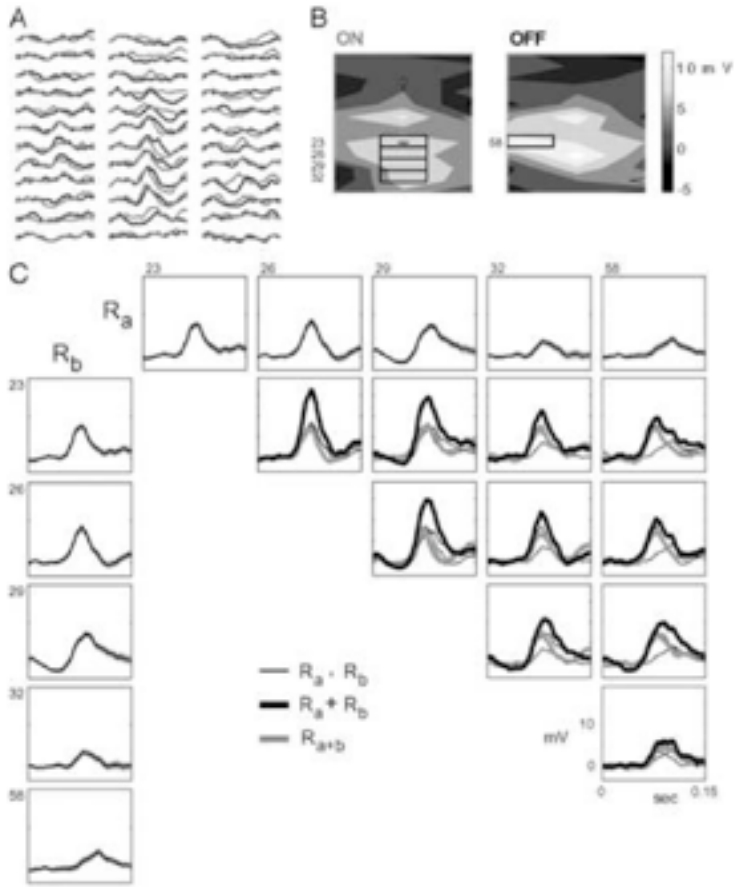
Max-like operation

Max-like behavior in V4



Gawne & Martin 2002

Max-like behavior in V1

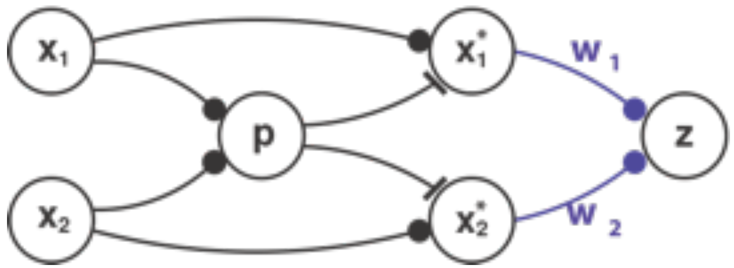


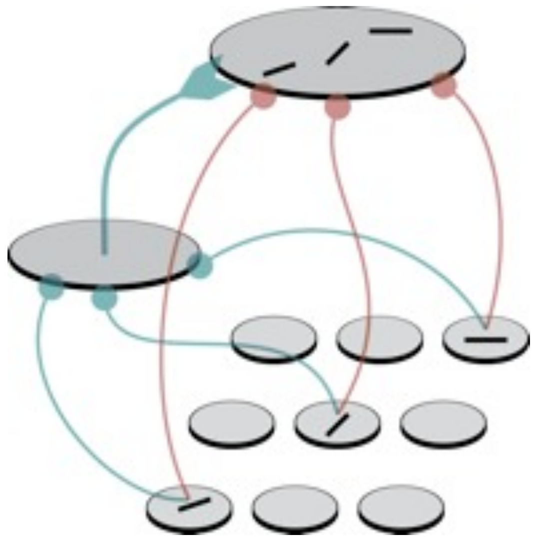
Lampl Ferster Poggio & Riesenhuber 2004
see also Finn Prieber & Ferster 2007

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)

Basic circuit is closely related to other models

Operation	(Steady-State) Output	
Canonical	$$y = \frac{\sum_{i=1}^{n} w_i\, x_i^{\,p}}{k + \left(\sum_{i=1}^{n} x_i^{\,q}\right)^{r}}$$	(1)
Energy Model	$$y = \sum_{i=1}^{2} x_i^{\,2}$$	(2)
Gaussian-like	$$y = \frac{\sum_{i=1}^{n} w_i\, x_i}{k + \sum_{i=1}^{n} x_i^{\,2}}$$	(4)
Max-like	$$y = \frac{\sum_{i=1}^{n} x_i^{\,3}}{k + \sum_{i=1}^{n} x_i^{\,2}}$$	(5)

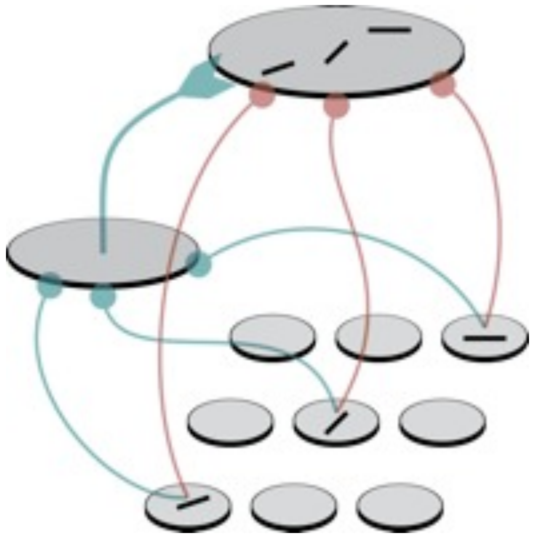
Can be implemented by shunting inhibition (Grossberg 1973, Reichardt et al. 1983, Carandini and Heeger, 1994) and spike threshold variability (Anderson et al. 2000, Miller and Troyer, 2002)

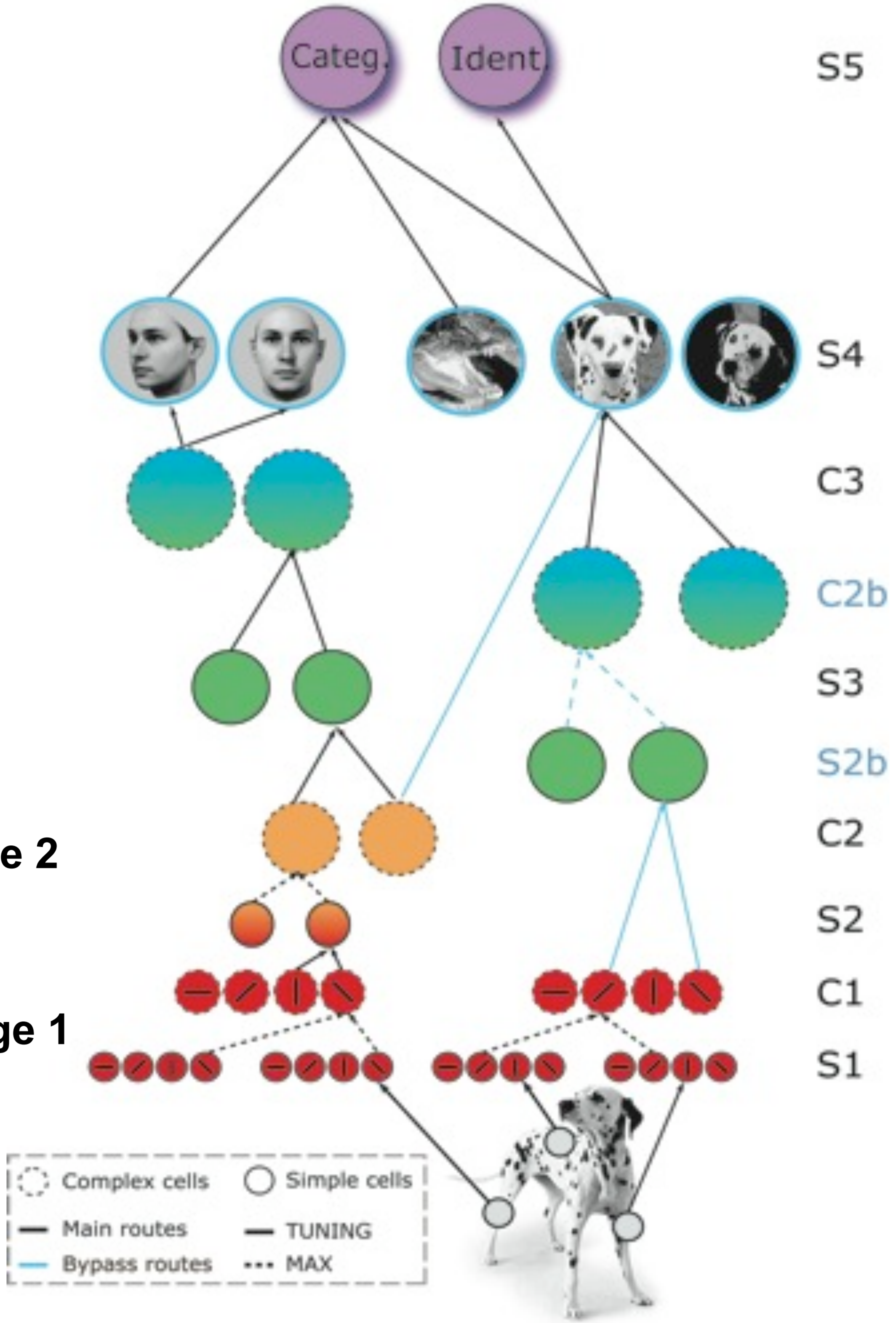
Adelson and Bergen (see also Hassenstein and Reichardt, 1956)

Of the same form as model of MT (Rust et al., Nature Neuroscience, 2007

Biophysics: one circuit

A canonical microcircuit of spiking neurons?

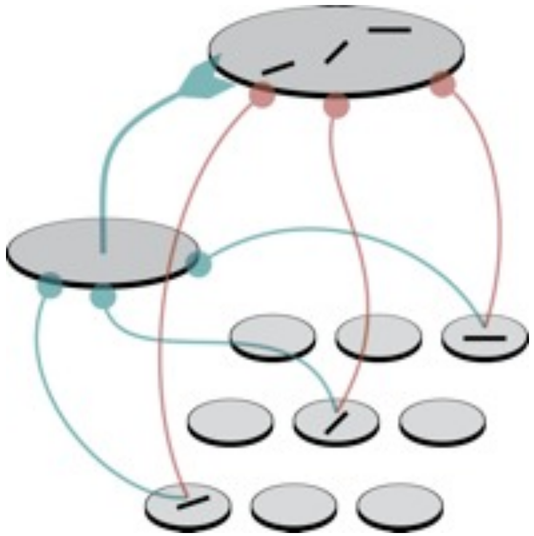


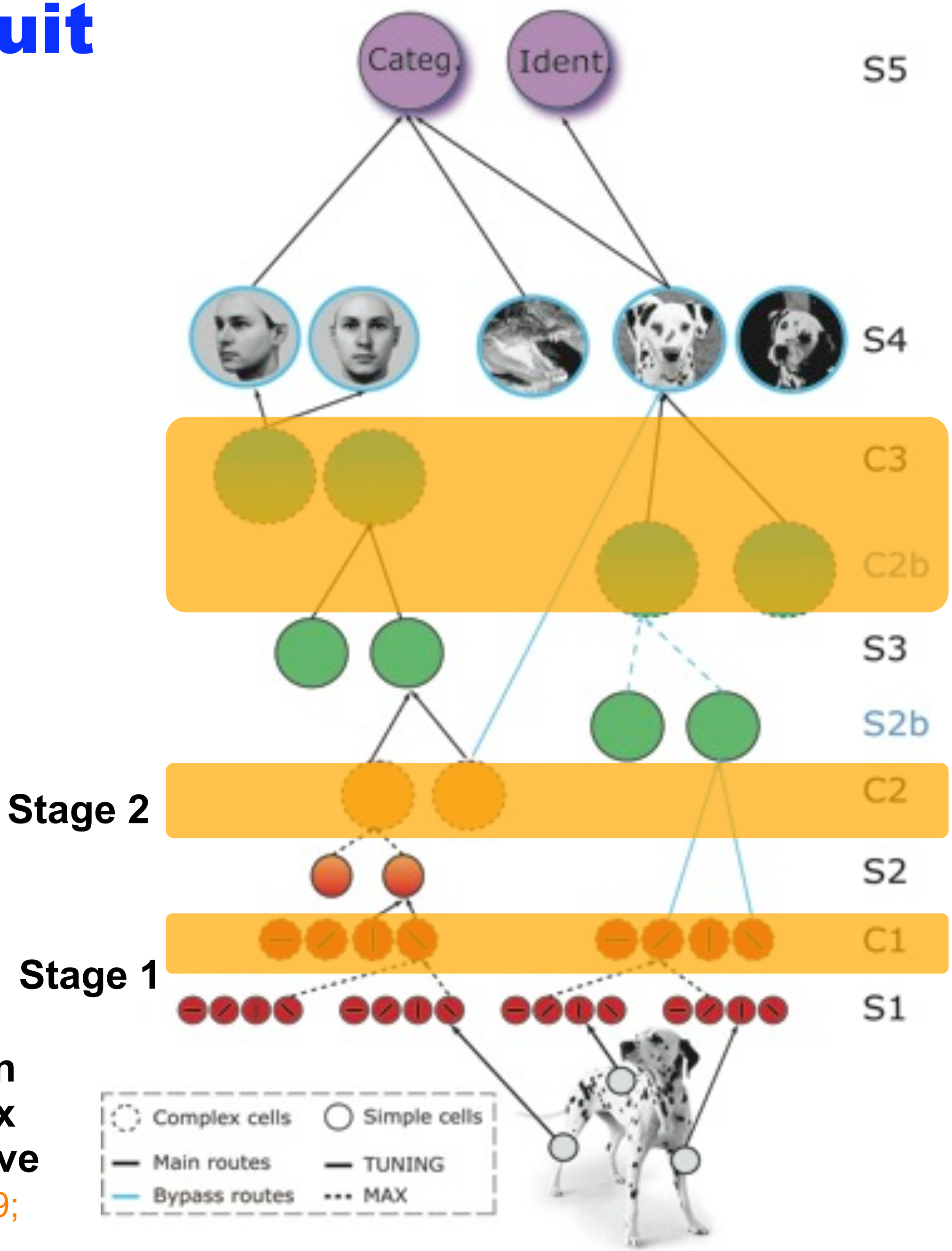


A plausible biophysical implementation for *both* Gaussian tuning (~AND) + max (~OR): normalization circuits with divisive inhibition (Kouh, Poggio, 2008; also RP, 1999; Heeger, Carandini, Simoncelli,…)

Biophysics: one circuit

A canonical microcircuit of spiking neurons?

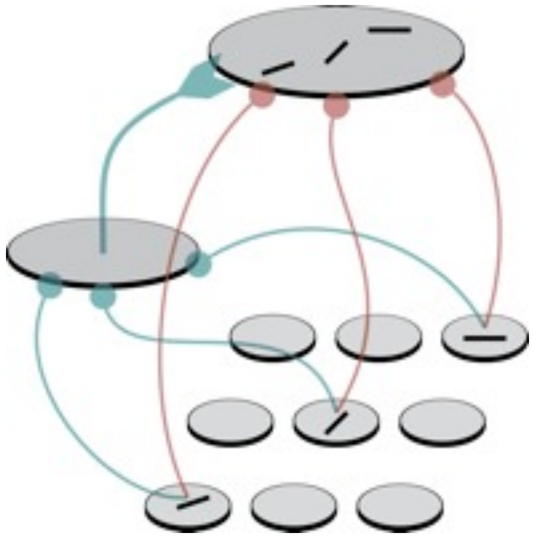


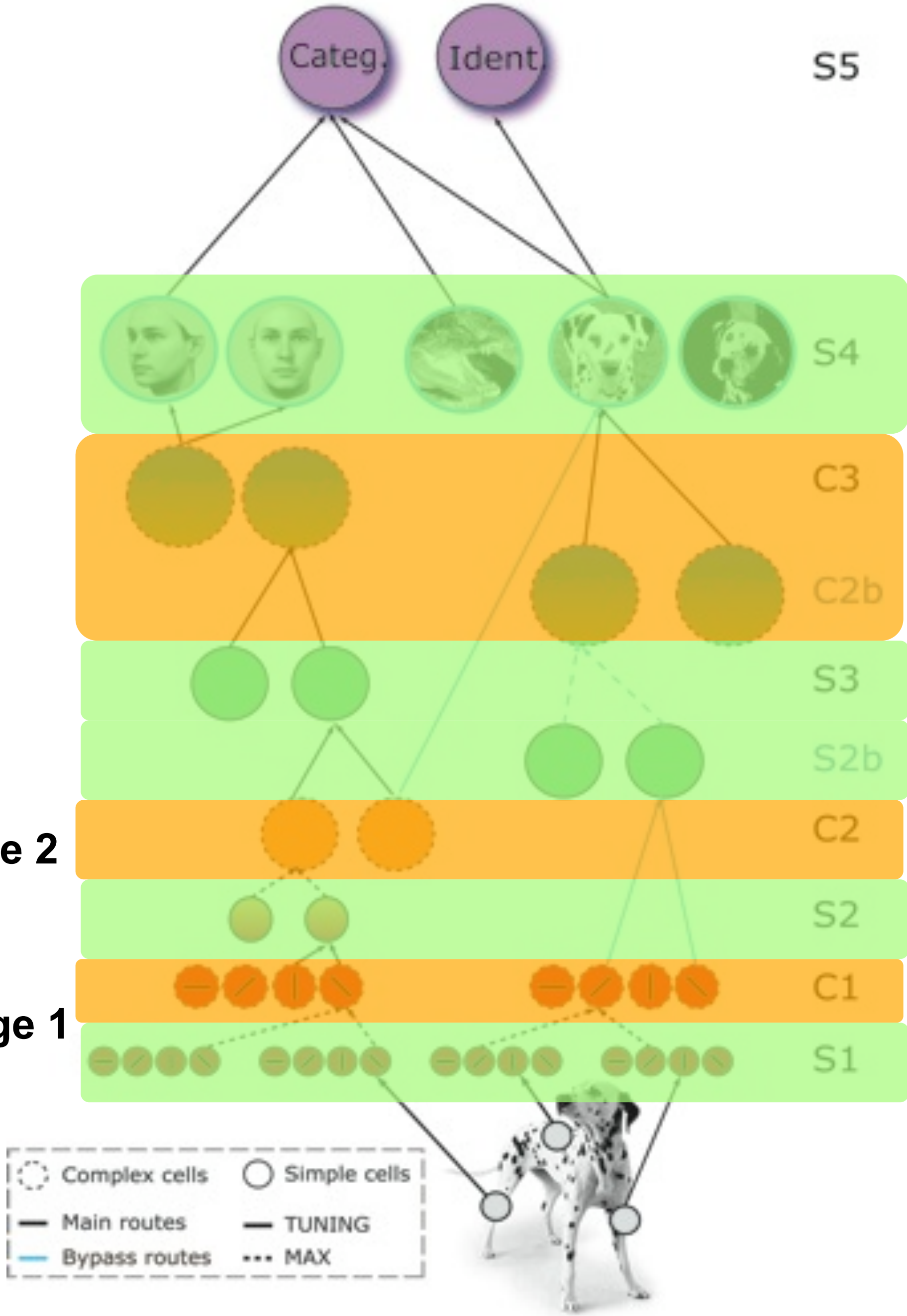


A plausible biophysical implementation for *both* Gaussian tuning (~AND) + max (~OR): normalization circuits with divisive inhibition (Kouh, Poggio, 2008; also RP, 1999; Heeger, Carandini, Simoncelli,…)

Biophysics: one circuit

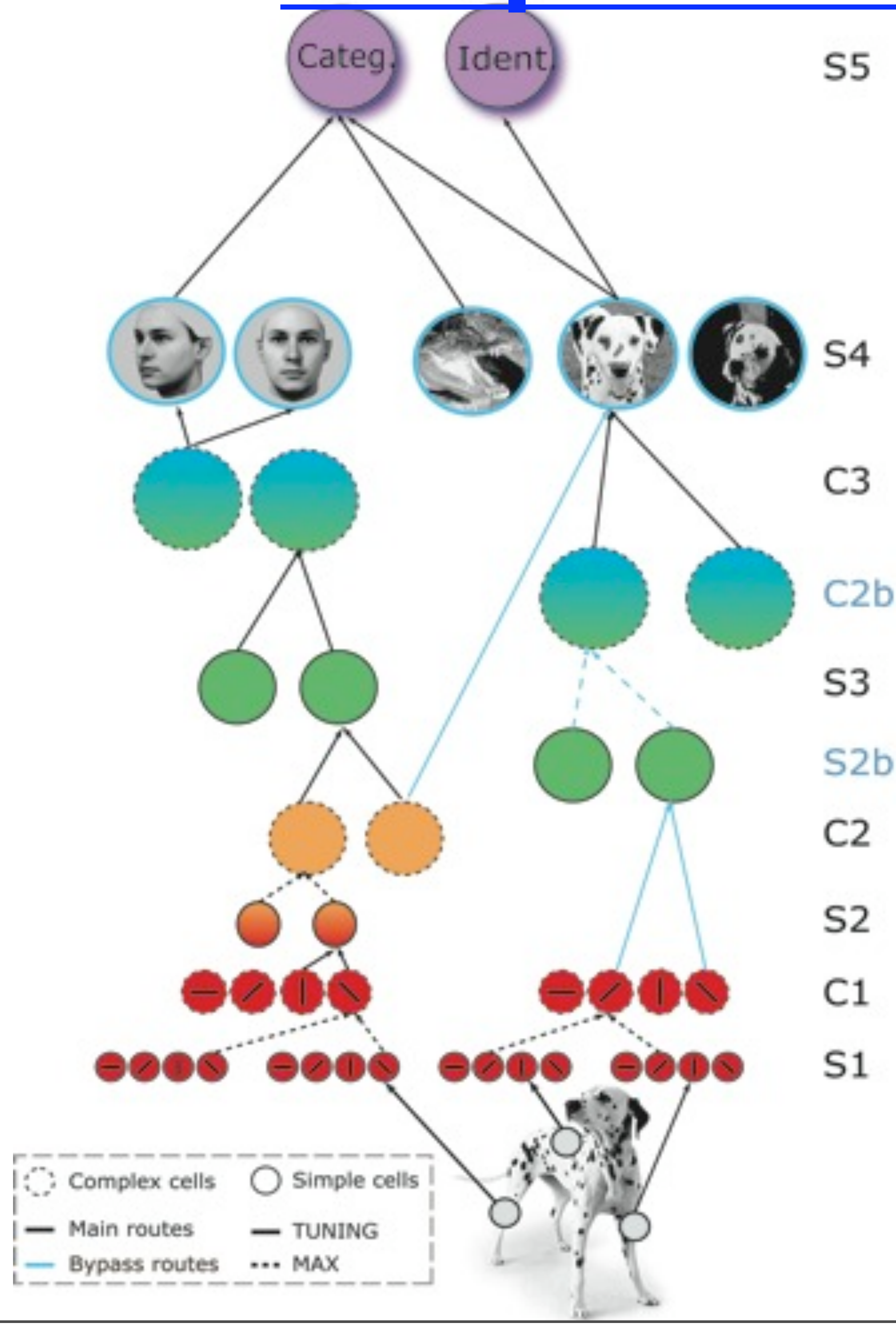
A canonical microcircuit of spiking neurons?



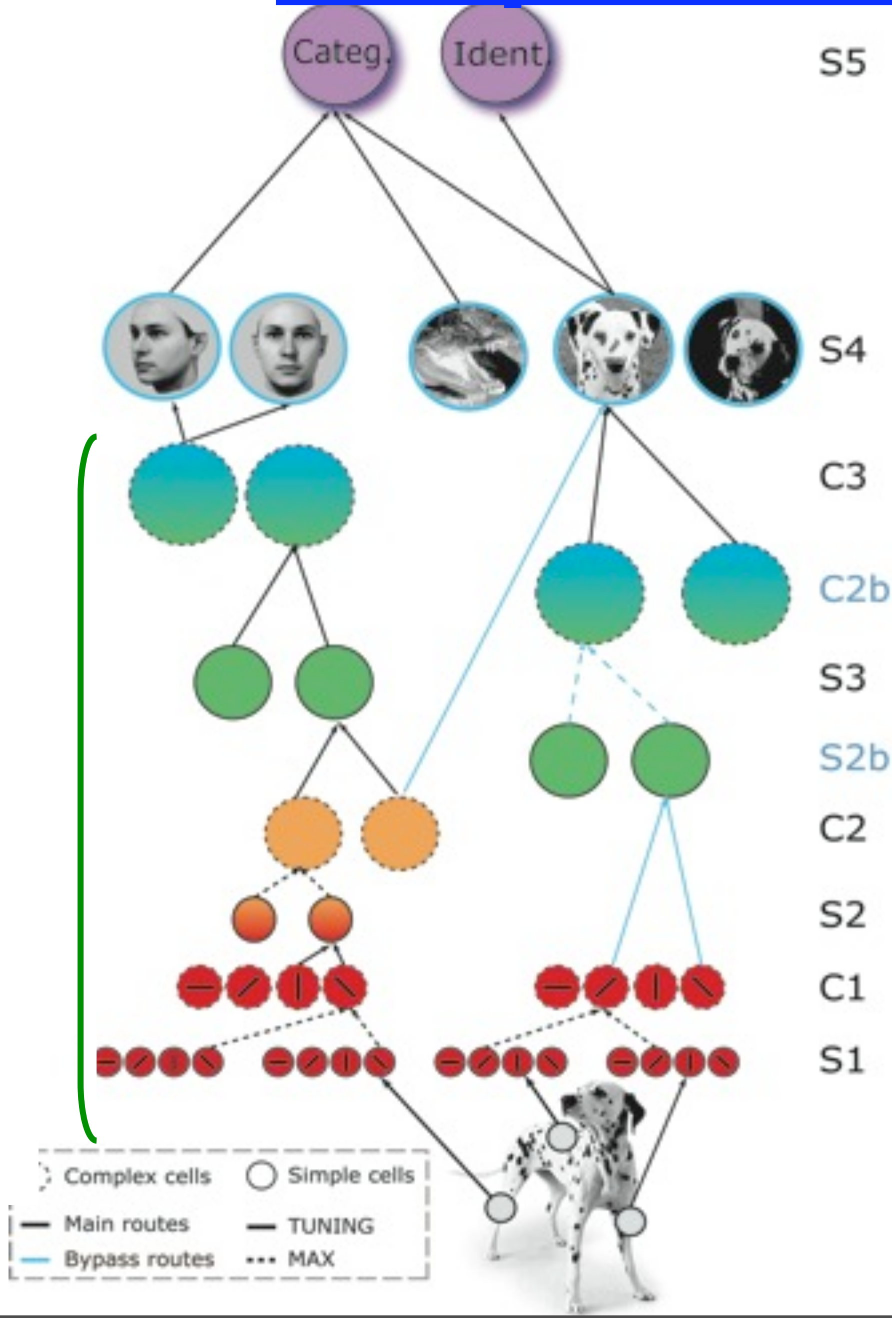


A plausible biophysical implementation for *both* Gaussian tuning (~AND) + max (~OR): normalization circuits with divisive inhibition (Kouh, Poggio, 2008; also RP, 1999; Heeger, Carandini, Simoncelli,…)

Learning: supervised and unsupervised



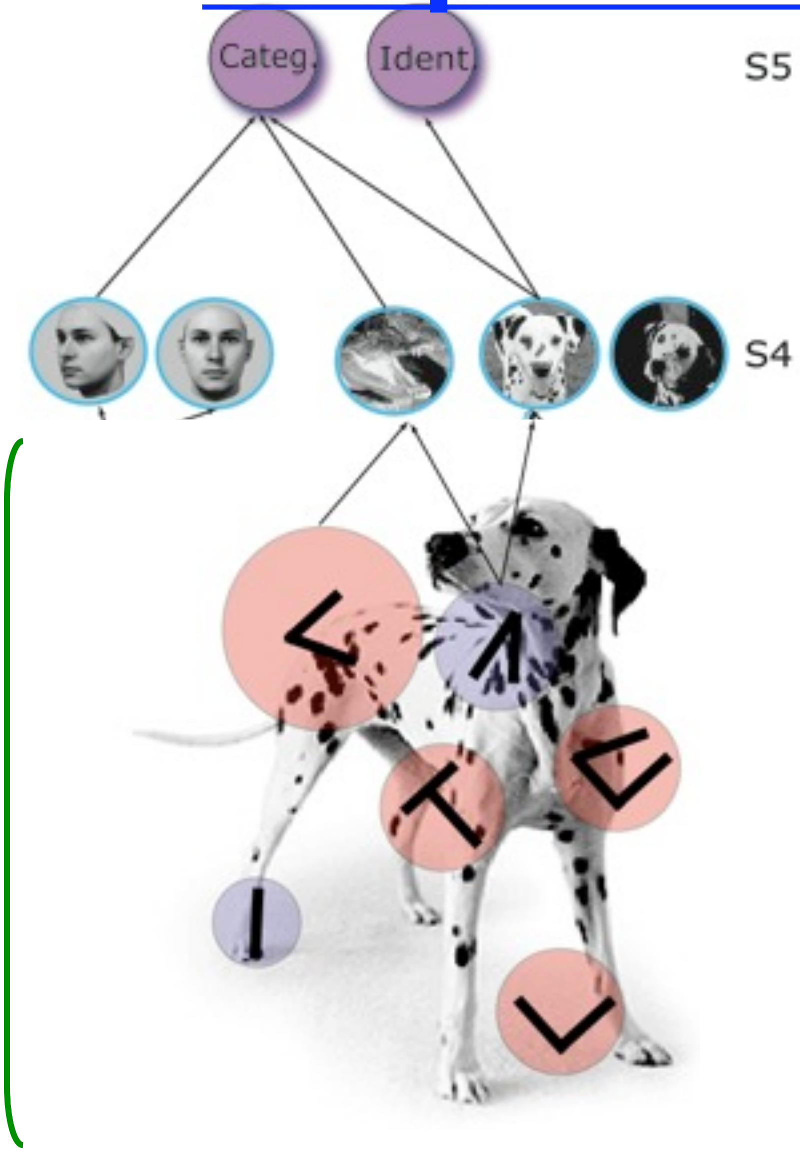
Learning: supervised and <u>unsupervised</u>



- Generic, overcomplete dictionary of “templates” or image components (from V1 to IT) represented by tuning of cells generated during **<u>unsupervised</u>** 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 <u>unsupervised</u>

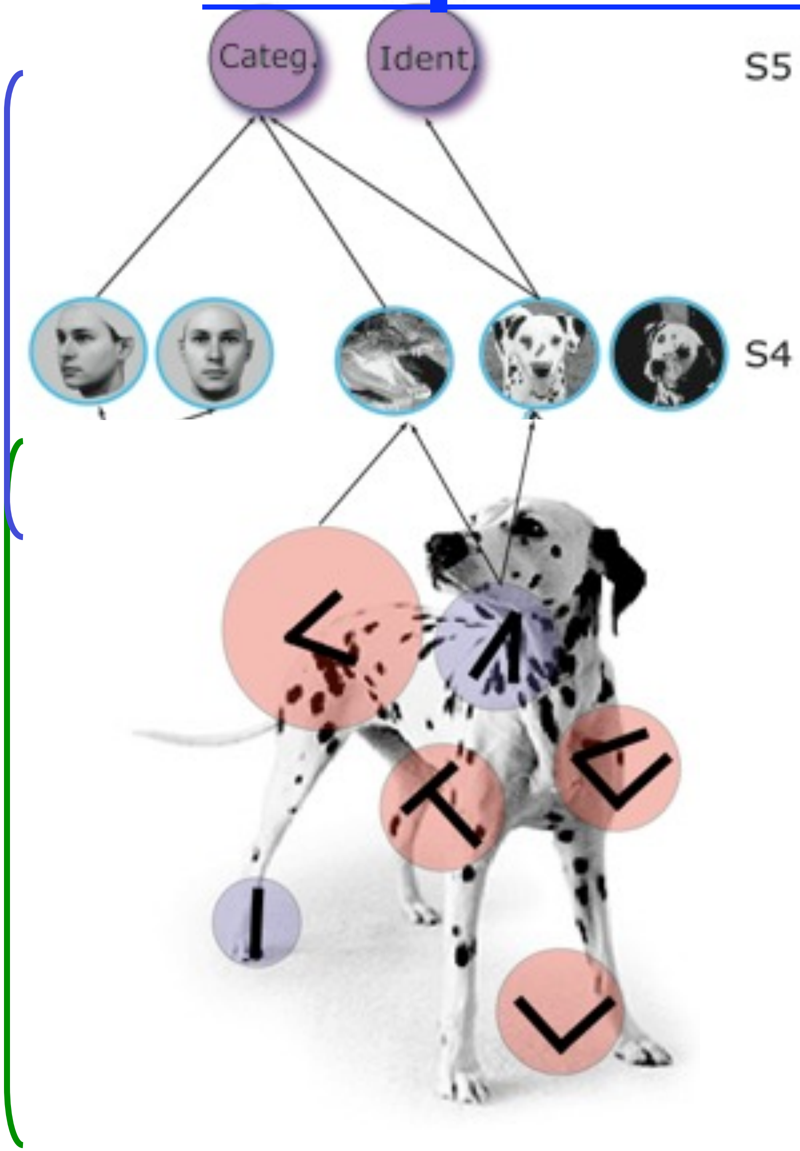


- Generic, overcomplete dictionary of "templates" or image components (from V1 to IT) represented by tuning of cells generated during **<u>unsupervised</u>** learning (from ~10,000 natural images) during a developmental-like stage

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

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)

More on feedforward (CBCL) models

S and C layers and parameters

unsupervised, developmental learning

software, GPUs and optimization

Jim Mutch

Max operation in cortex	The model predicted the existence of complex cells in V1 [Lampl et al., 2004] and V4 [Gawne and Martin, 2002] performing a soft-max pooling operation
Tolerance to eye movements	From the softmax operation – originally introduced to explain invariance to translation in IT – the model predicts stability of complex cells responses relative to small eye motions
Tuning properties of view-tuned units in IT	The model has been able to duplicate quantitatively the generalization properties of IT neurons that remain highly selective for particular objects, while being invariant to some transformations [Logothetis et al., 1995; Riesenhuber and Poggio, 1999b] their tuning for pseudo-mirror views and generalization over contrast reversal. Also, the model qualitatively accounts for IT neurons responses to altered stimuli [Riesenhuber and Poggio, 1999b], *i.e.*, scrambling [Vogels, 1999], presence of distractors within units receptive fields [Sato, 1989] and clutter [Missal et al., 1997]
Role of IT and PFC in categorization tasks	After training monkeys to categorize between "cats" and "dogs", we found that the ITC seems more involved in the analysis of currently viewed shapes, whereas the PFC showed stronger category signals, memory effects, and a greater tendency to encode information in terms of its behavioral meaning [Freedman et al., 2002] (see also subsection 4.4)
Learned model C2 units compatible with V4 data	We have recently shown (see Subsection 4.2) that C2 units that were passively learned from natural images seem consistent with V4 data, including tuning for boundary conformations [Pasupathy and Connor, 2001], two-spot interactions[Freiwald et al., 2005], gratings [Gallant et al., 1996], as well as the biased-competition model [Reynolds et al., 1999]
"Face inversion" effect	The model has helped [Riesenhuber et al., 2004] guide control conditions in psychophysical experiments to show that an effect that appeared to be incompatible with the model turned out to be an artifact

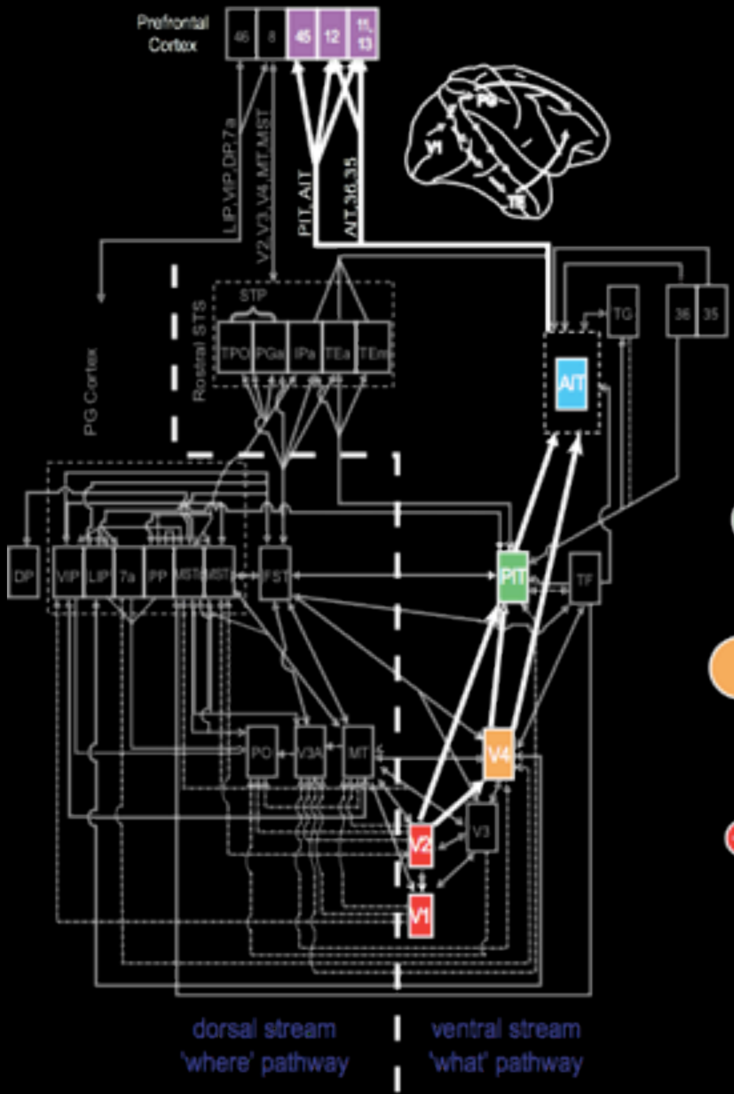
Table 3: Some of the correct predictions by the model

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)

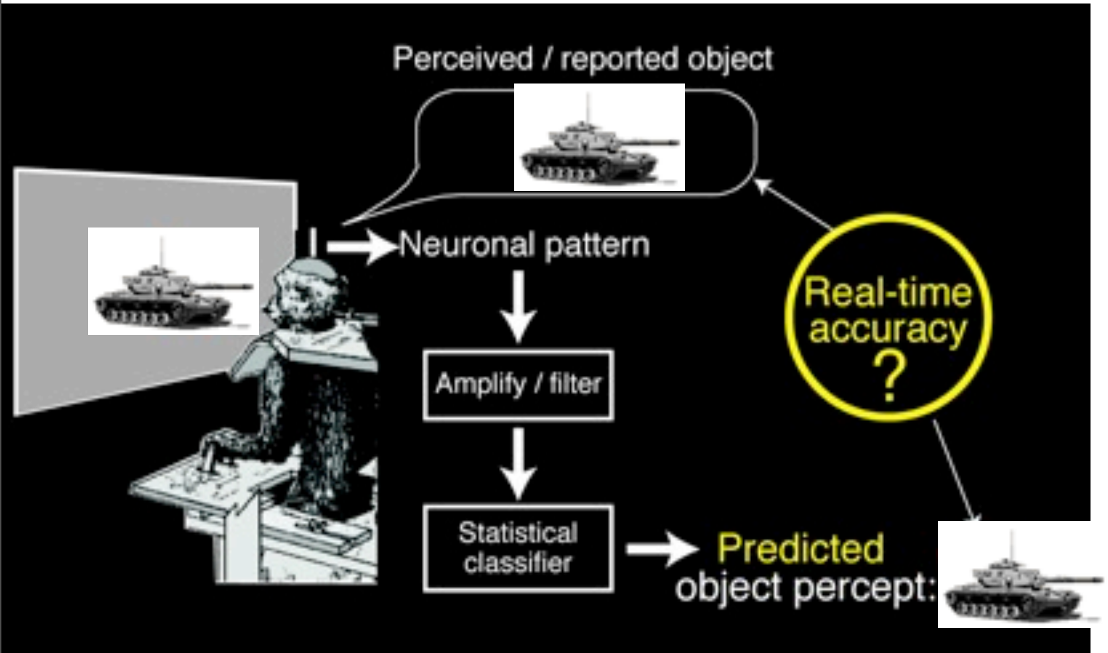
Comparison w/ neural data

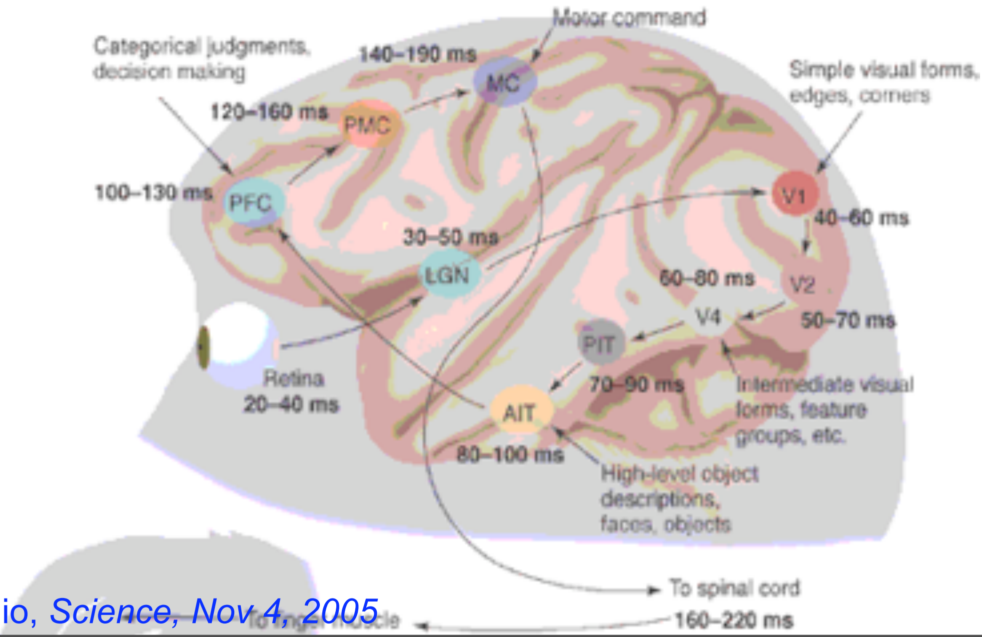


- Simple and complex cells tuning properties (Schiller et al 1976; Hubel & Wiesel 1965; Devalois et al 1982)
- MAX operation in subset of complex cells (Lampl et al 2004)
- Tuning for two-bar stimuli (Reynolds Chelazzi & Desimone 1999)
- MAX operation (Gawne et al 2002)
- Two-spot interaction (Freiwald et al 2005)
- Tuning for boundary conformation (Pasupathy & Connor 2001)
- Tuning for Cartesian and non-Cartesian gratings (Gallant et al 1996)
- Tuning and invariance properties (Logothetis et al 1995)
- Differential role of IT and PFC in categorization (Freedman et al 2001 2002 2003)
- Read out data (Hung Kreiman Poggio & DiCarlo 2005)
- Average effect in IT (Zoccolan Cox & DiCarlo 2005; Zoccolan Kouh Poggio & DiCarlo in press)
- Rapid animal categorization (Serre Oliva Poggio 2007)

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

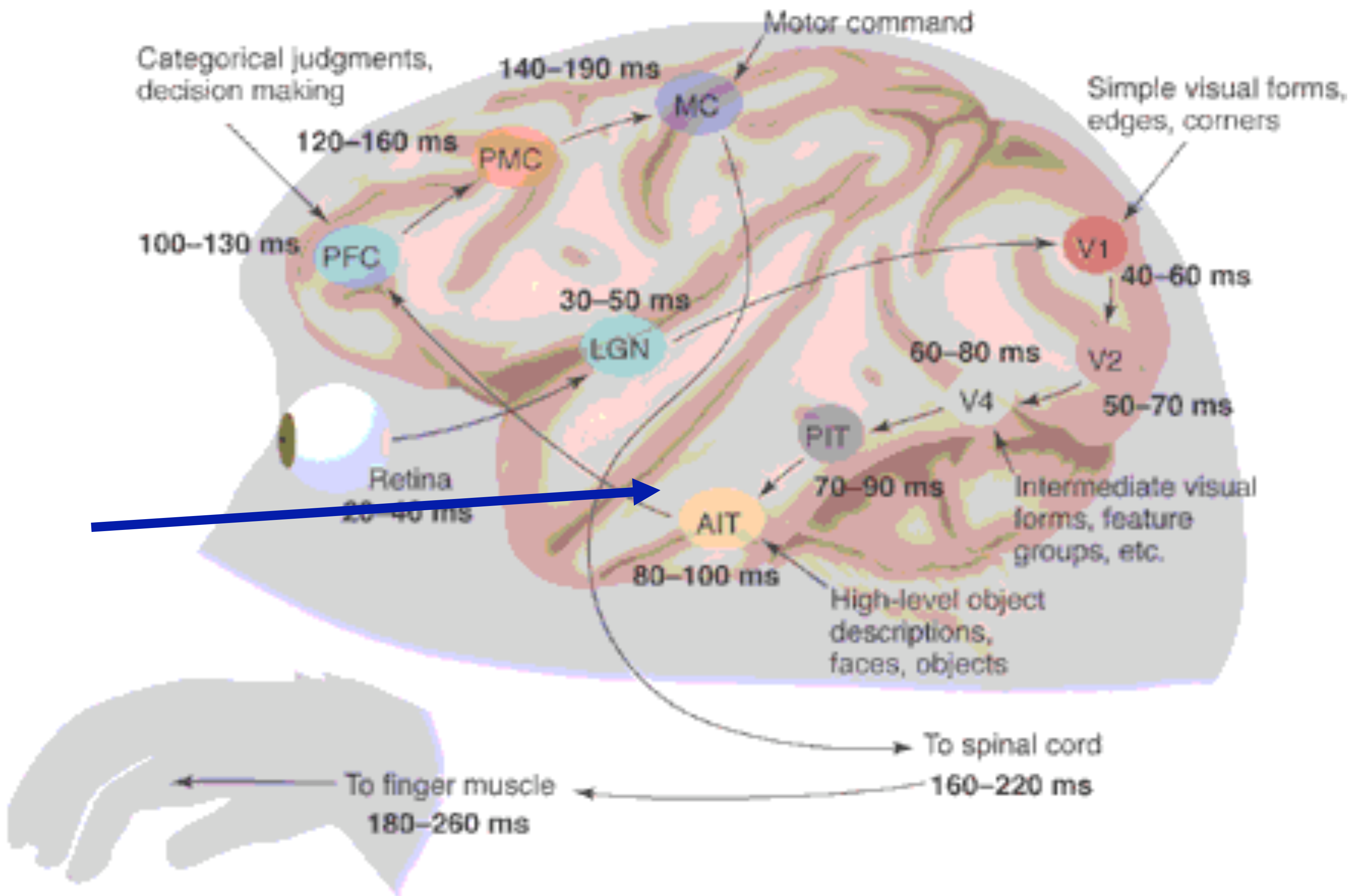
Agreement of model w| IT Readout data





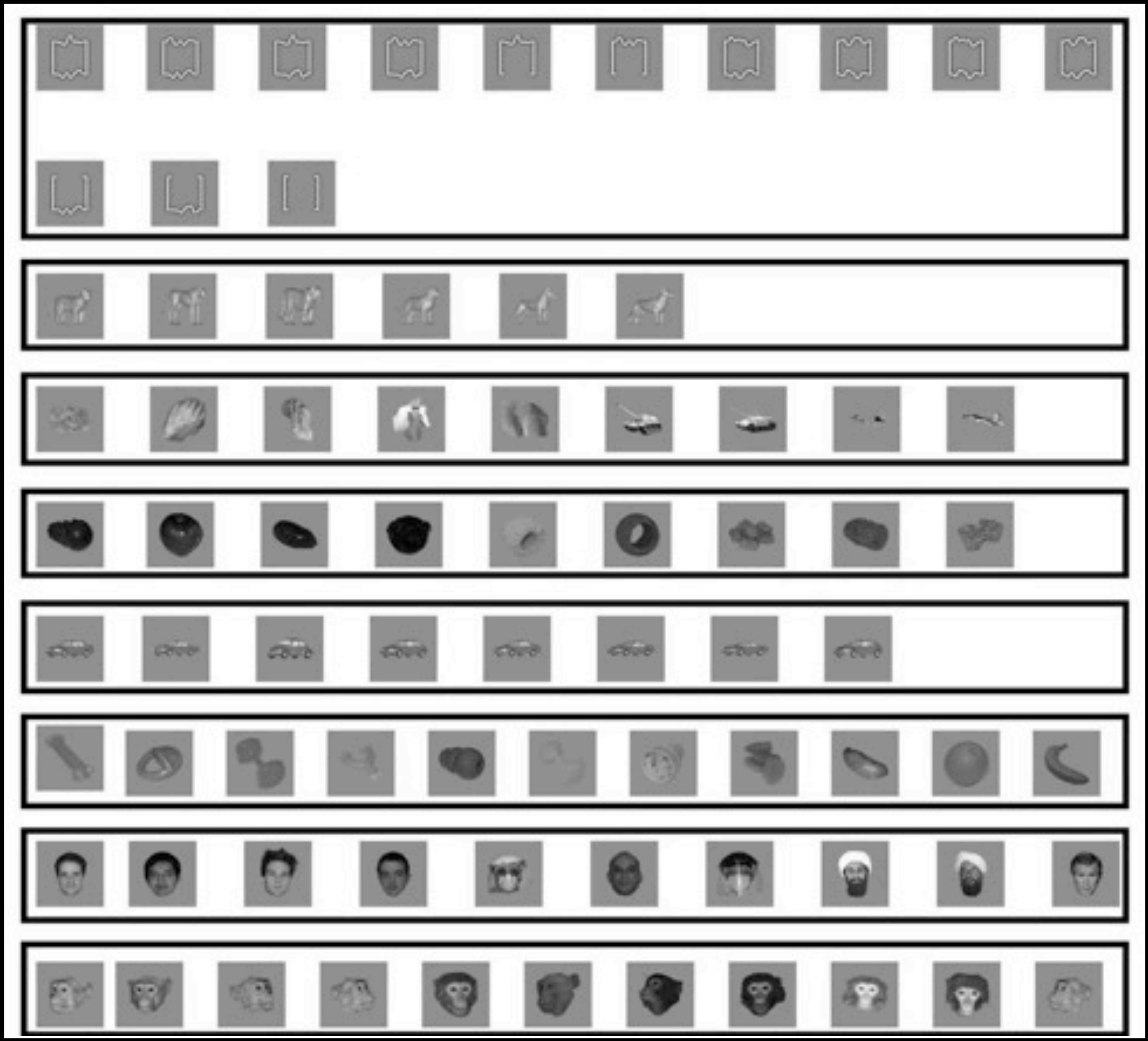
Chou Hung, Gabriel Kreiman, James DiCarlo, Tomaso Poggio, *Science, Nov 4, 2005*

The end station of the ventral stream in visual cortex is IT



IT Readout

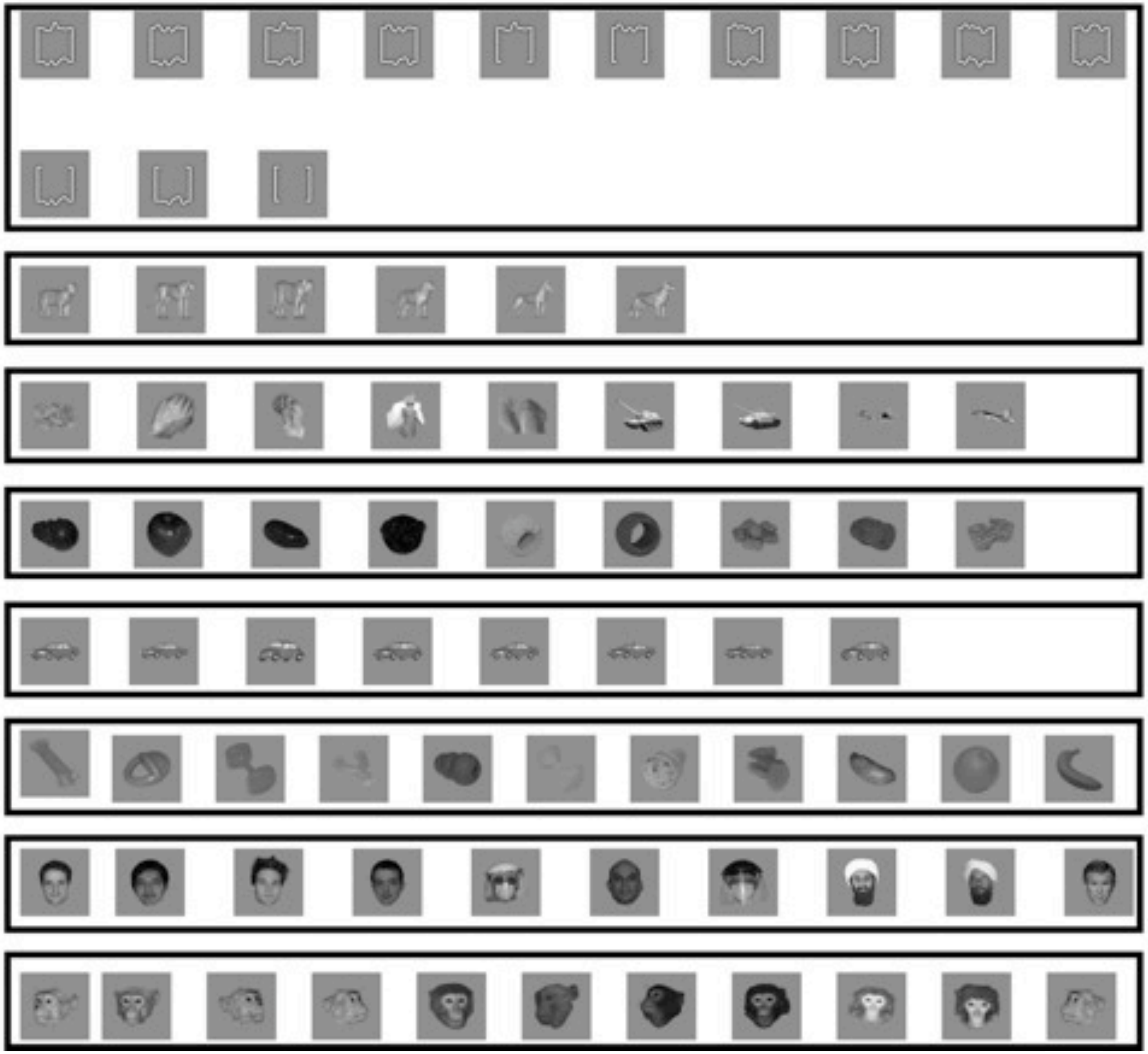
77 objects,
8 classes



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

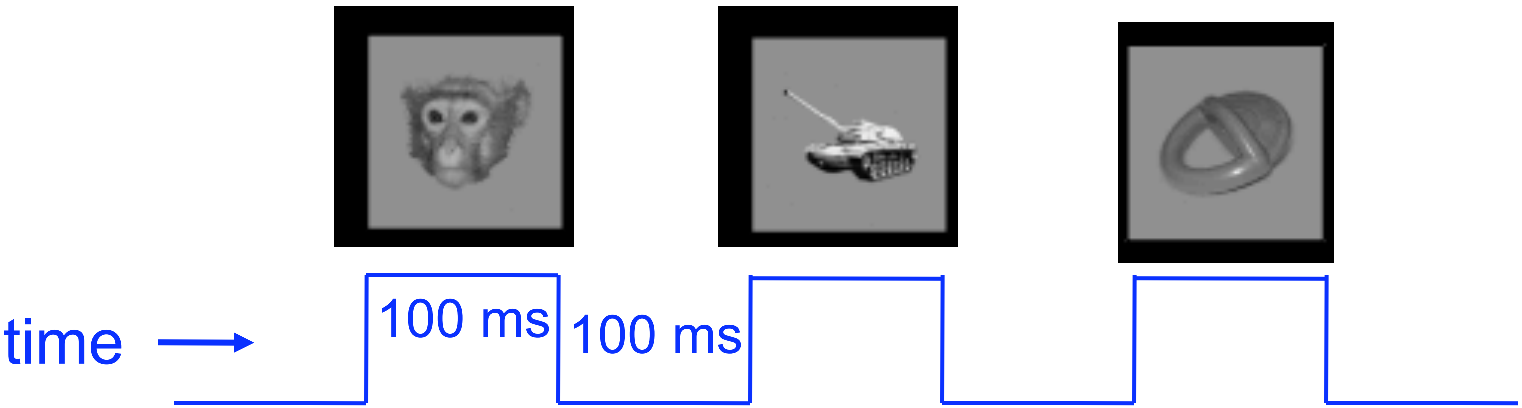
Reading-out the neural code in AIT

77 objects,
8 classes



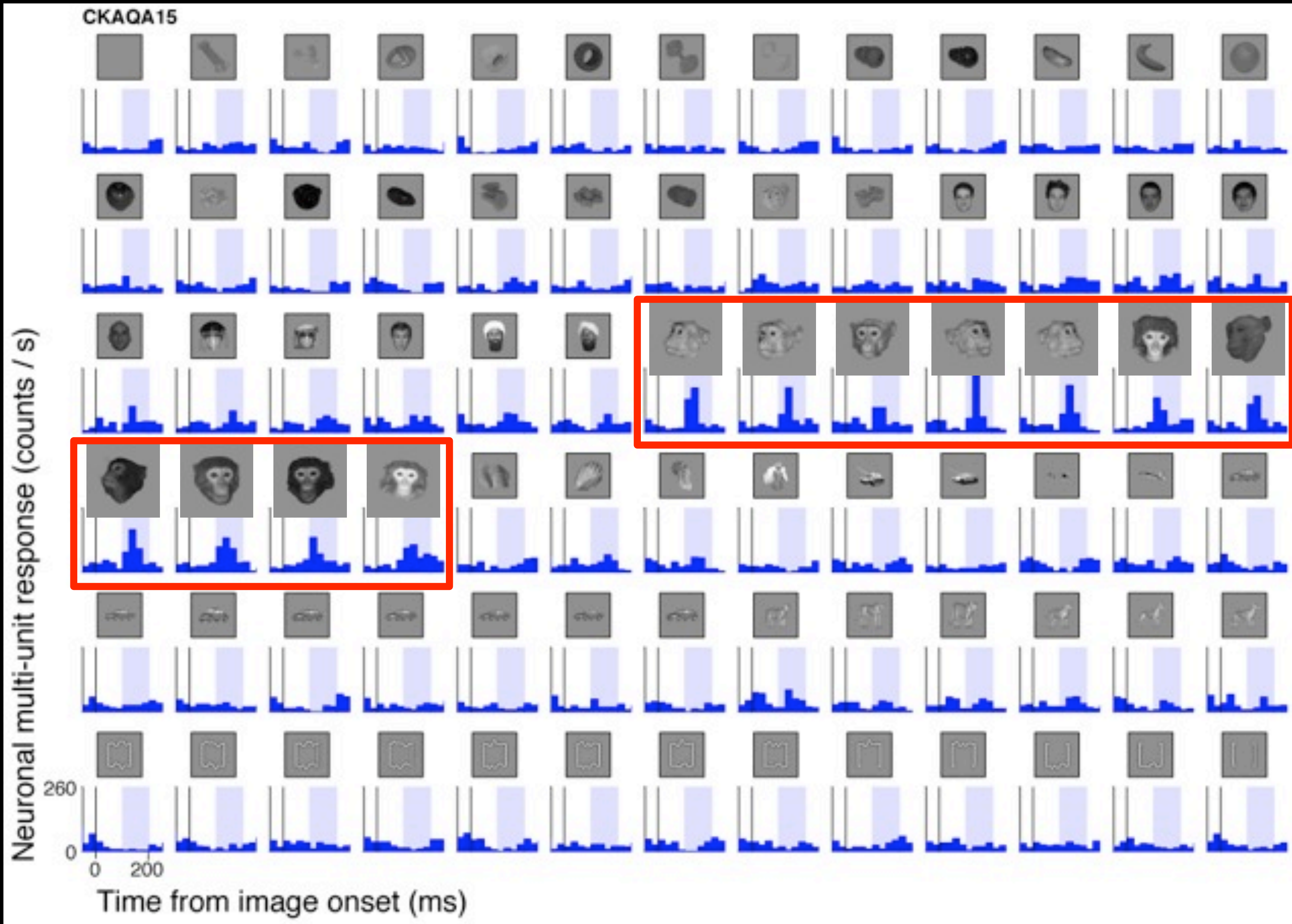
Chou Hung, Gabriel Kreiman, James DiCarlo, Tomaso Poggio, *Science, Nov 4, 2005*

Recording at each recording site during passive viewing

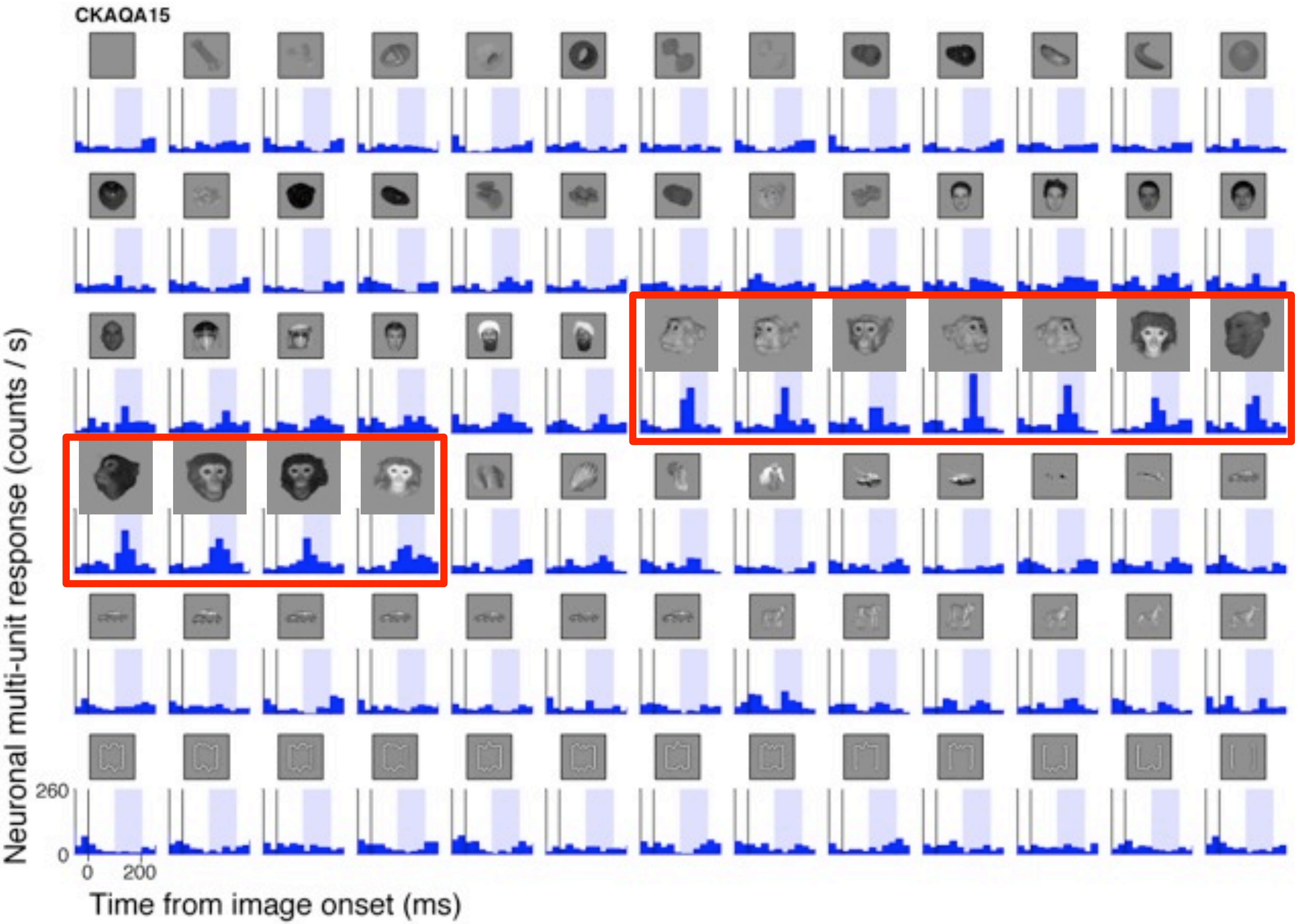


- 77 visual objects
- 10 presentation repetitions per object
- presentation order randomized and counter-balanced

Example of One IT Cell

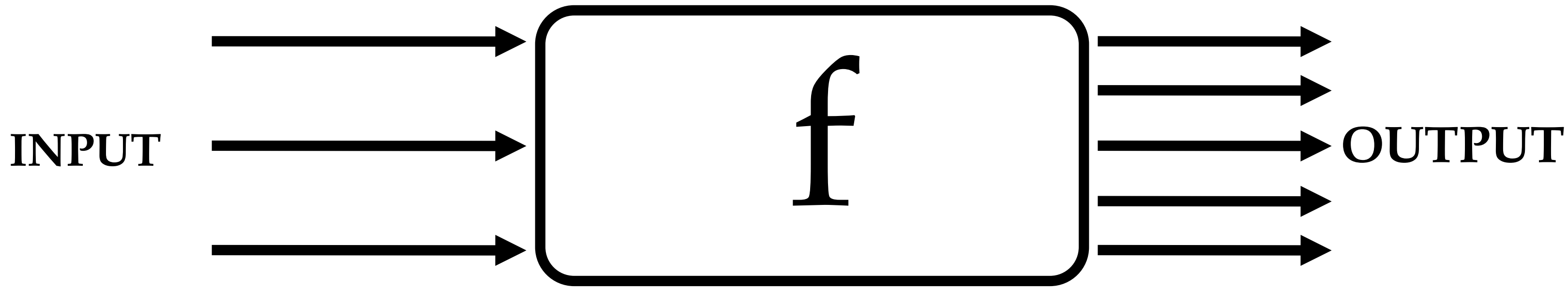


Agreement of model w| IT Readout data



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

Training a classifier on neuronal activity.



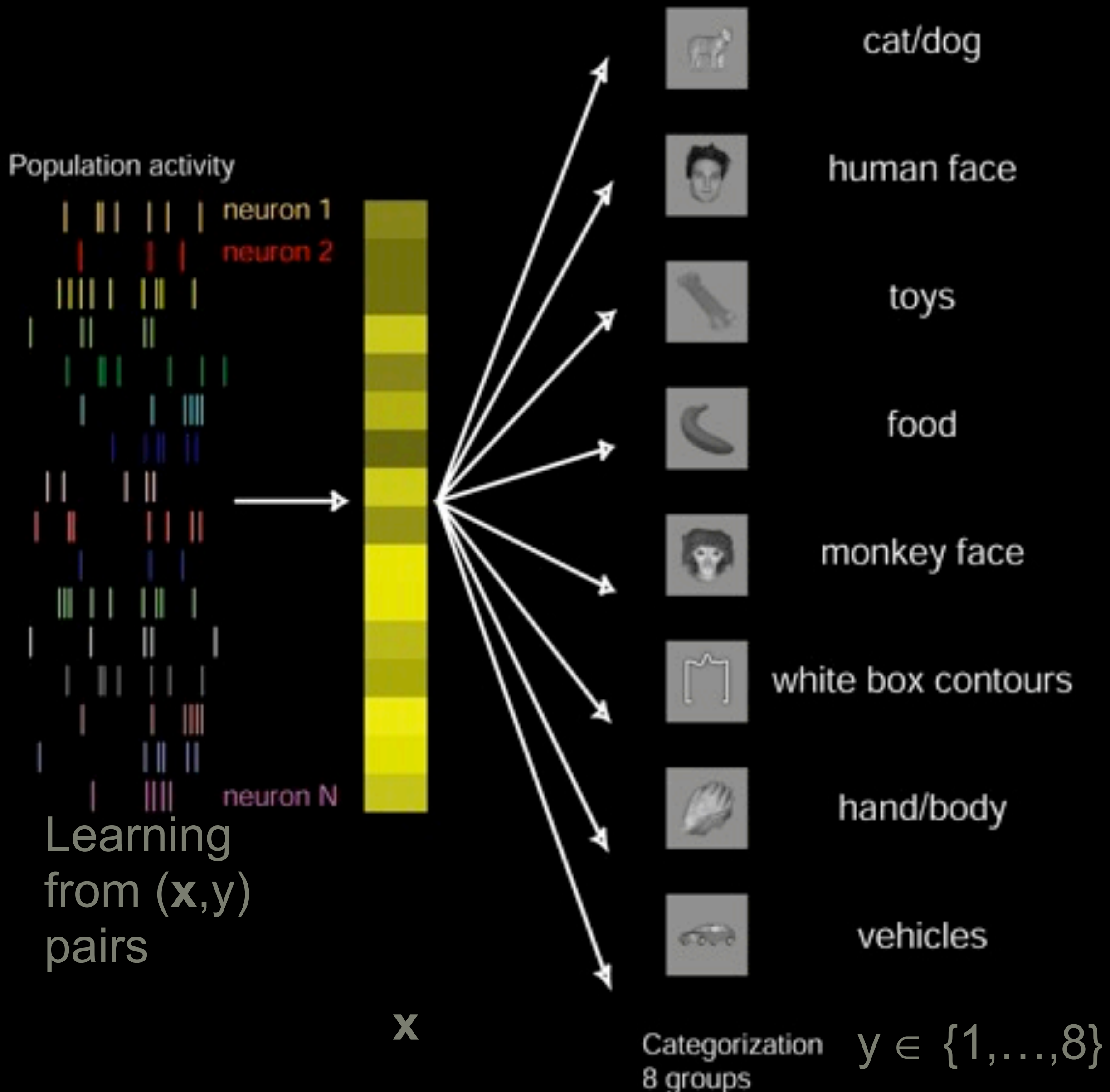
From a set of data (vectors of activity of n neurons (x) and object label (y)

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

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

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

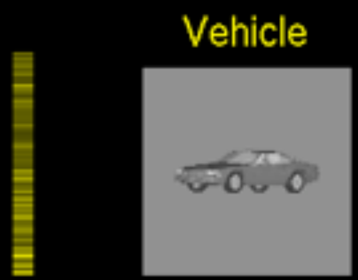
Decoding the Neural Code ...
population response (using a classifier)



Hung*, Kreiman, Poggio, DiCarlo. *Science 2005*

From neuronal population activity…

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



Categorization

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

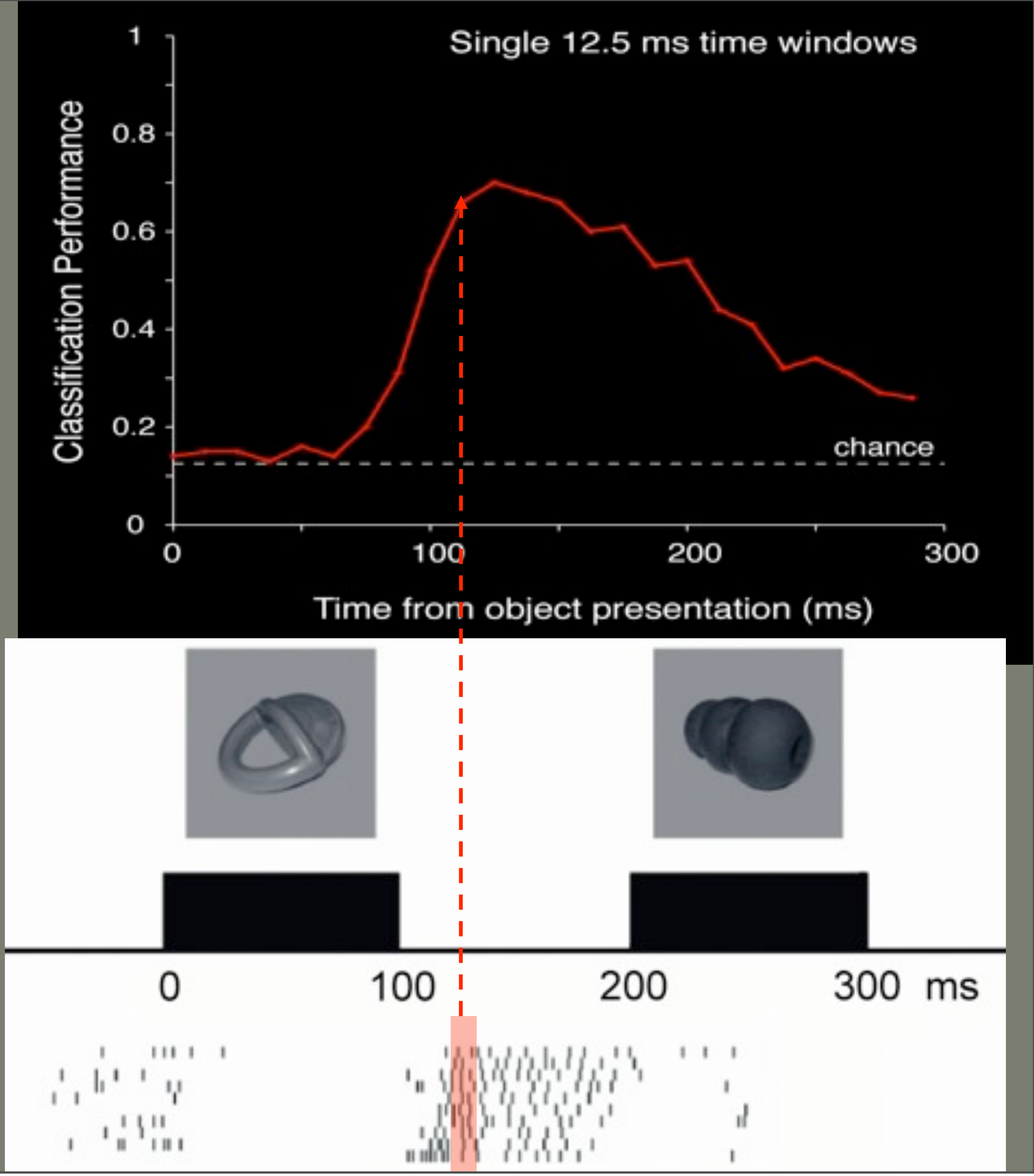
Video speed: 1 frame/sec
Actual presentation rate: 5 objects/sec

80% accuracy in read-out from ~200 neurons

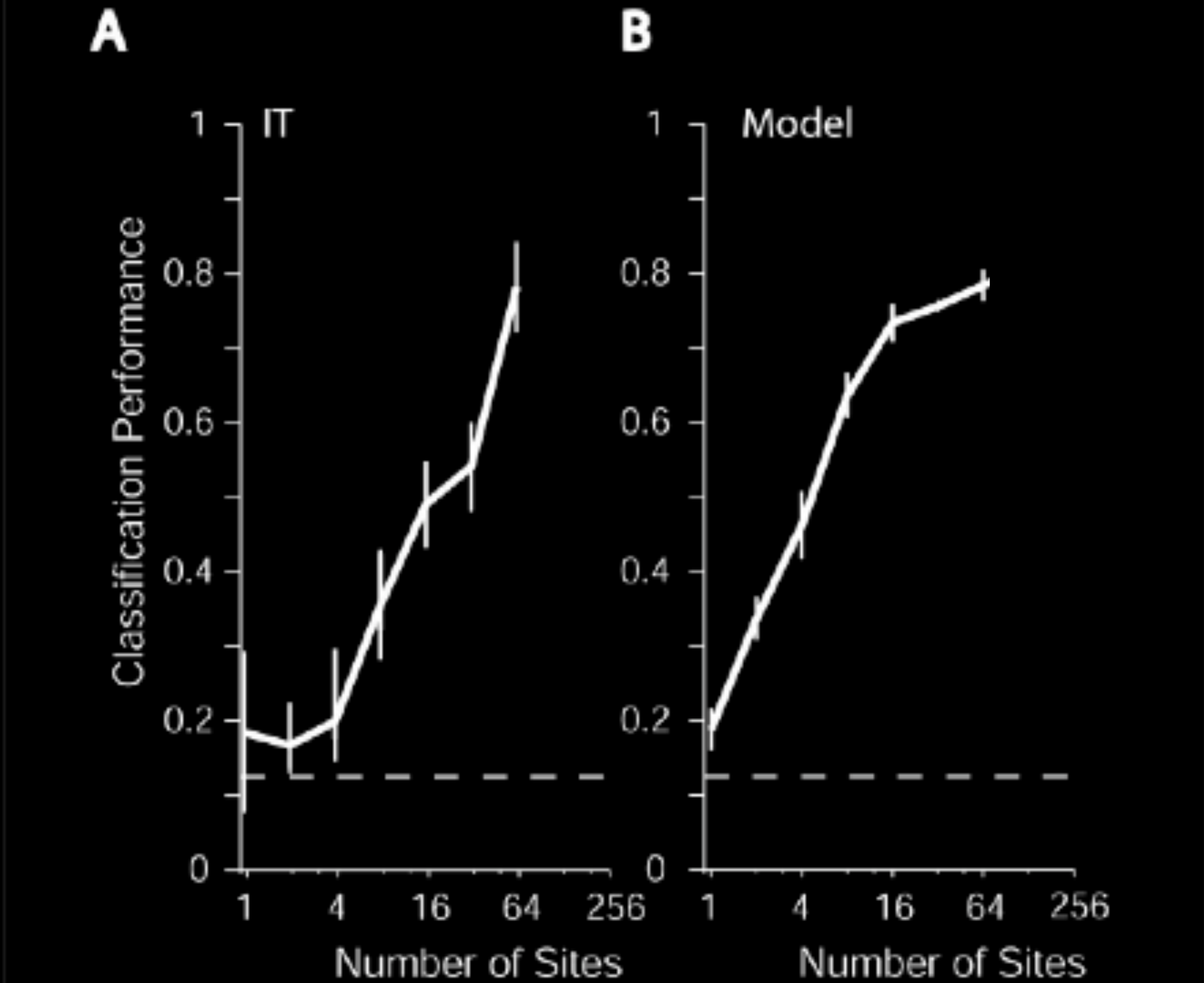
A result (C. Hung, et al., 2005): very rapid read-out of object information rapid (80-100 ms from onset of stimulus)

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



It turns out that the model agrees with IT data: we can decode from model units as well as from IT



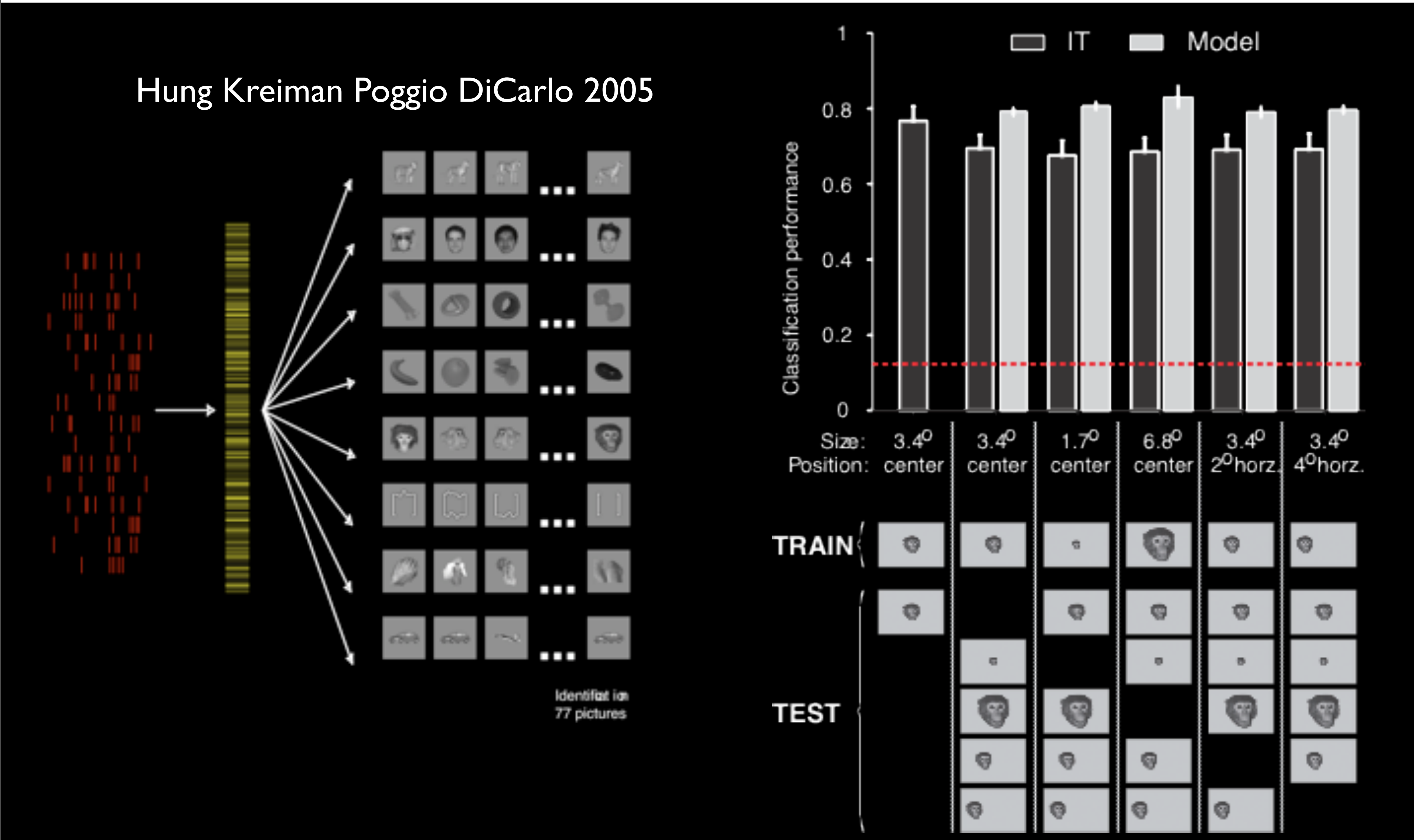
Serre, Kouh, Cadieu, Knoblich, Kreiman, Poggio. *MIT AI Memo* 2005

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 !!!

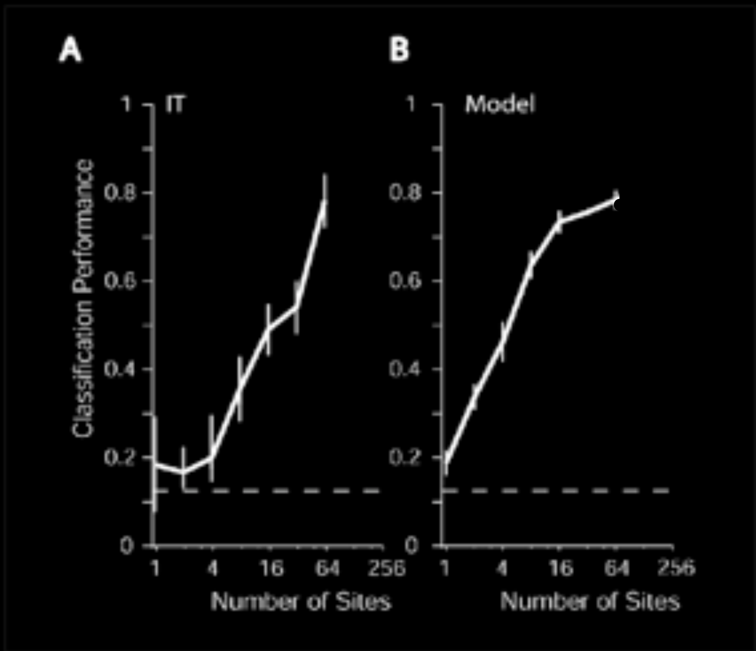
Agreement of model w| IT Readout data

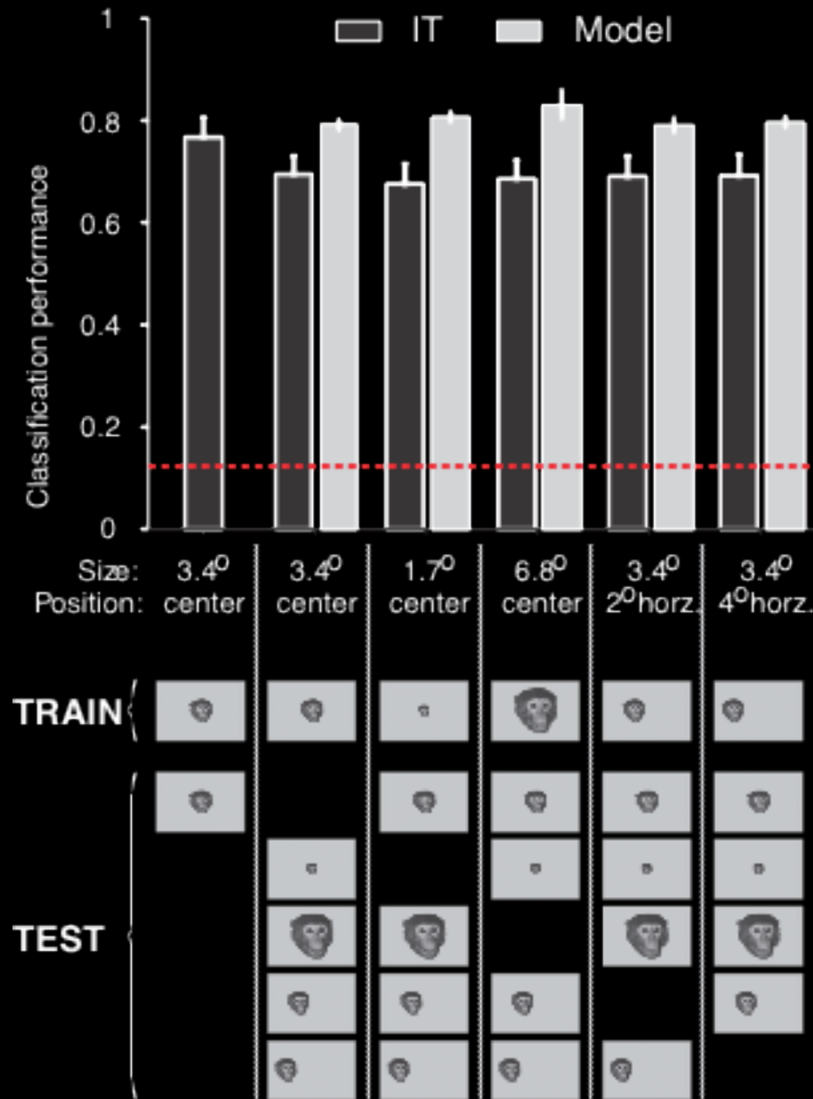
Reading out category and identity invariant to position and scale



Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005

Agreement of Model w| IT Readout data

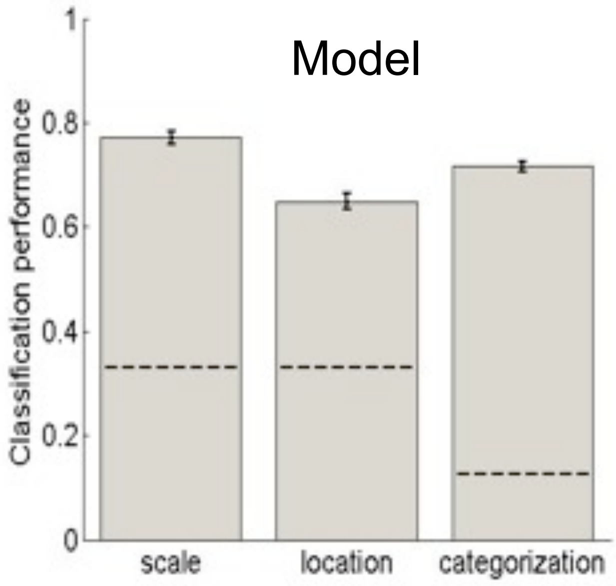


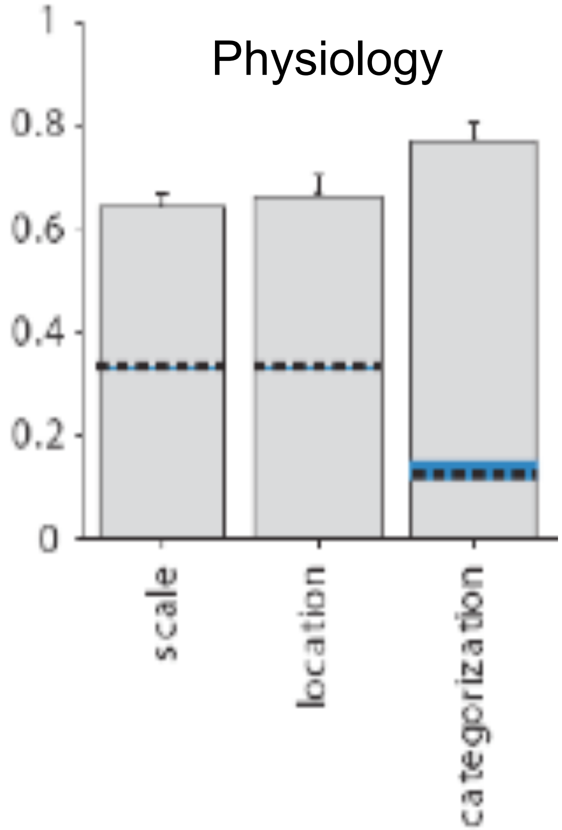


Reading out category and identity "*invariant*" to position and scale

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

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

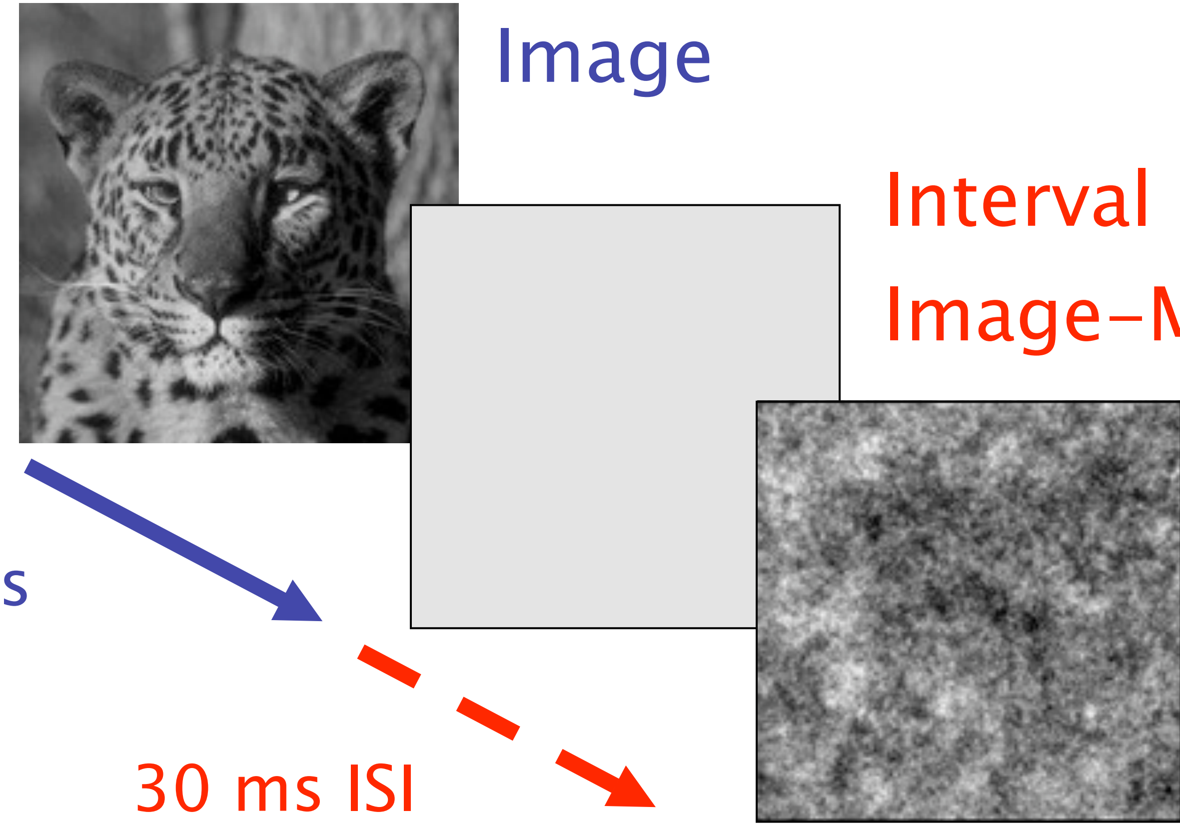




- **70/30 train/test (20 splits)**
- **64 randomly selected C3/C2b features**
  - to match 64 recording sites
- **Scale:** 77.2 ± 1.25% vs. ~63% (physiology)
- **Location:** 64.9 ± 1.44% vs. ~65% (physiology)
- **Categorization:** 71.6 ± 0.91% vs. ~77% (physiology)

Tan, Serre, Poggio, 2008

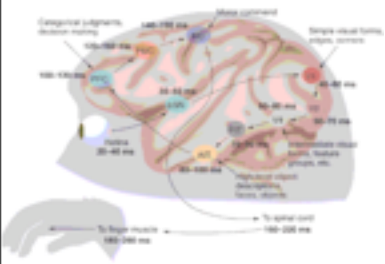
Hierarchical feedforward models of the ventral stream



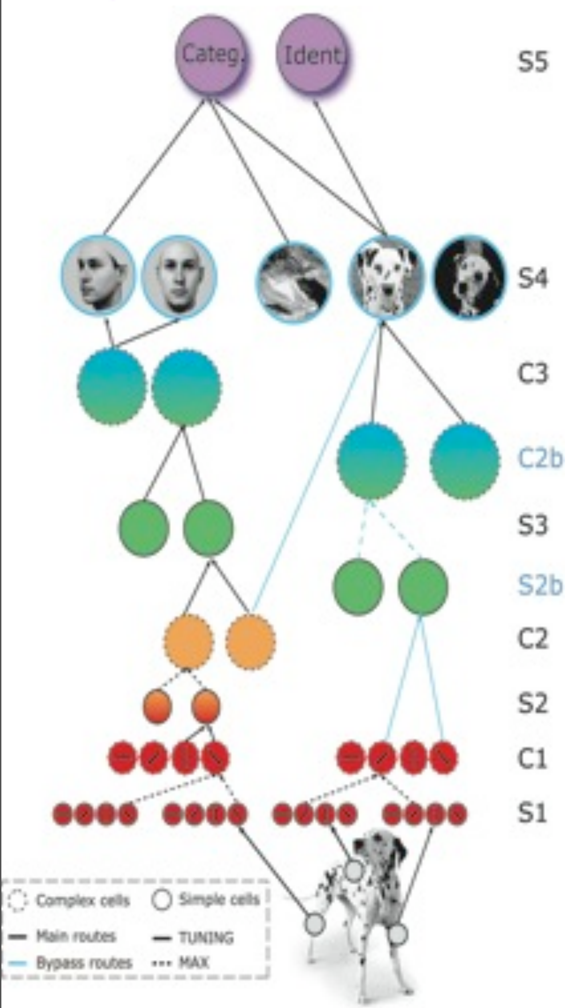
Rapid Categorization:
mask should force visual cortex to operate in feedforward mode

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

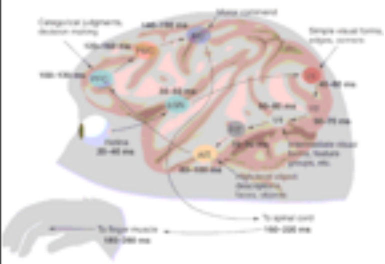
Hierarchical feedforward models of the ventral stream

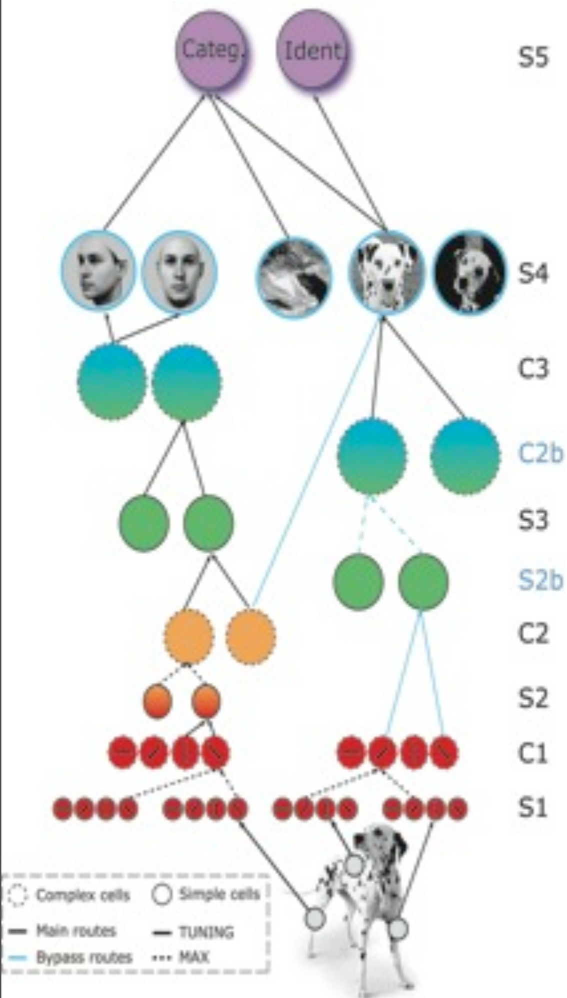


Rapid Categorization



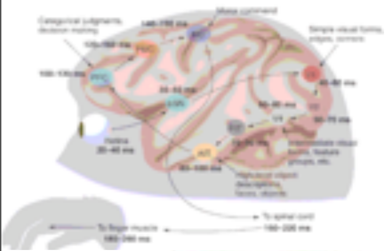
Hierarchical feedforward models of the ventral stream

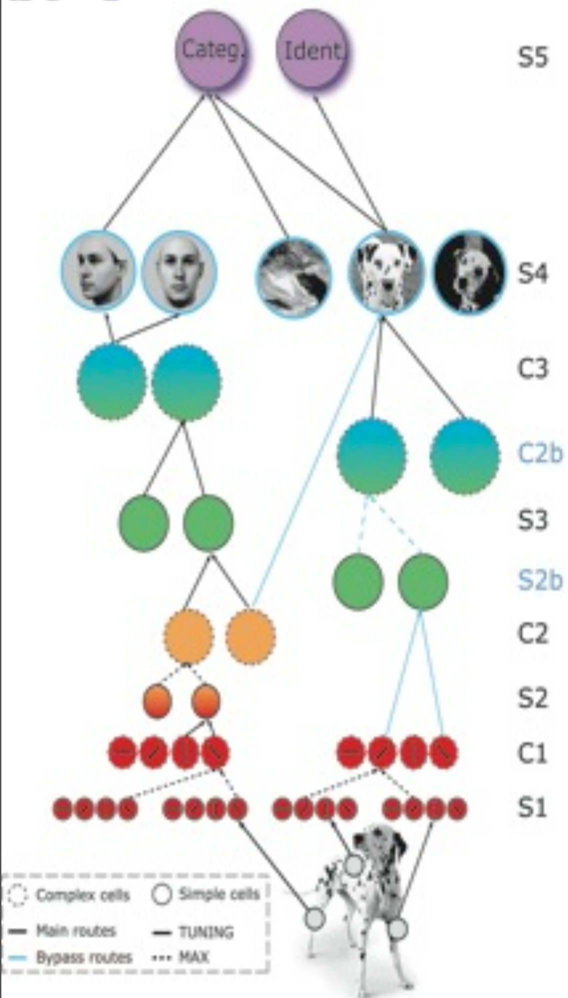




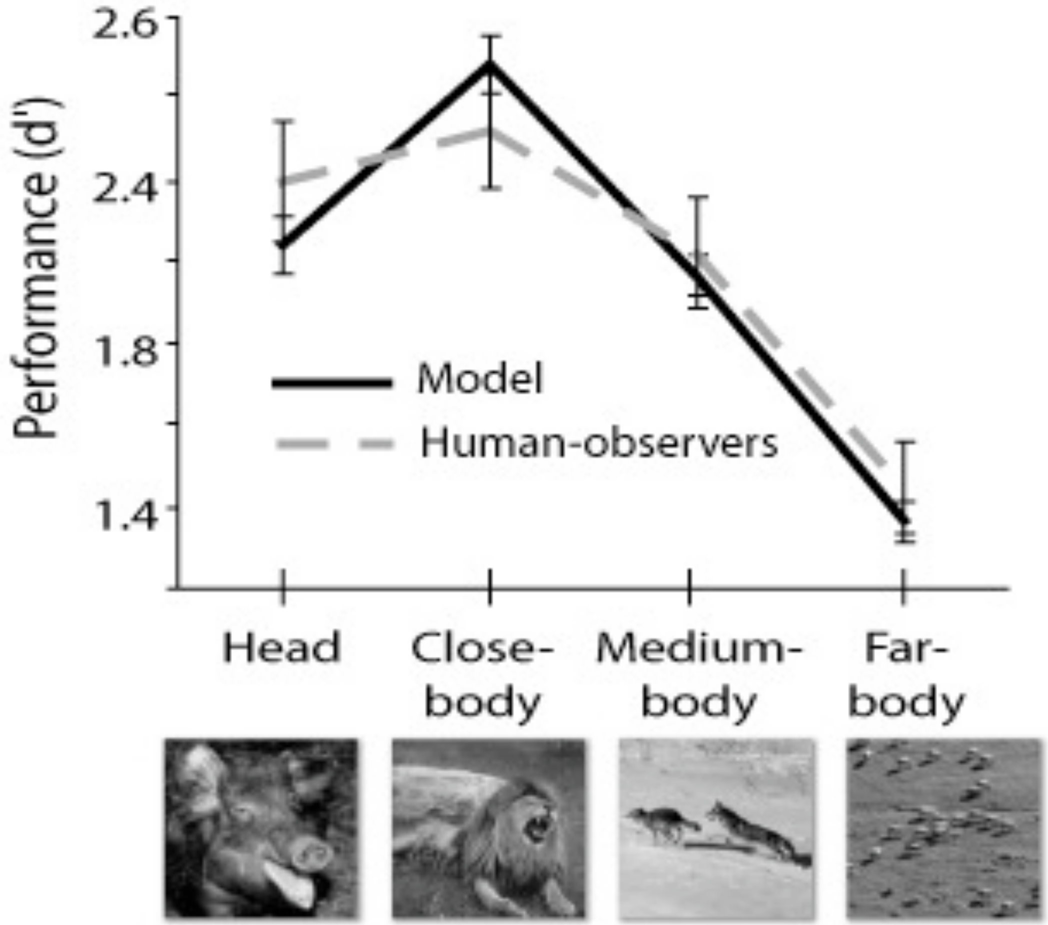
Rapid Categorization

Hierarchical feedforward models of the ventral stream

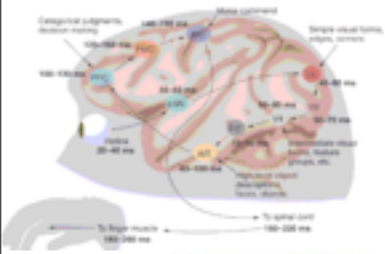


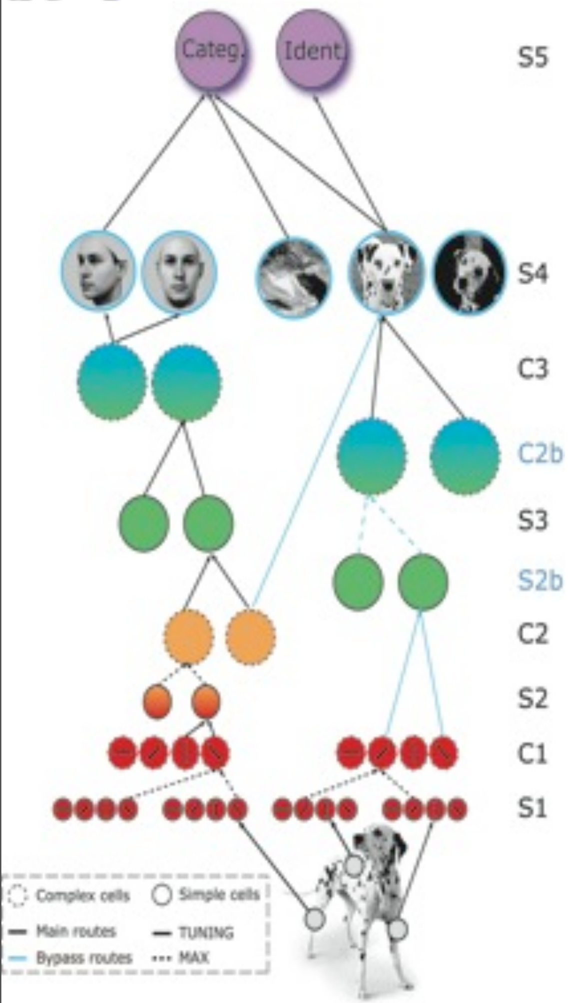


Feedforward Models:
“predict” rapid categorization
(82% model vs. 80% humans)

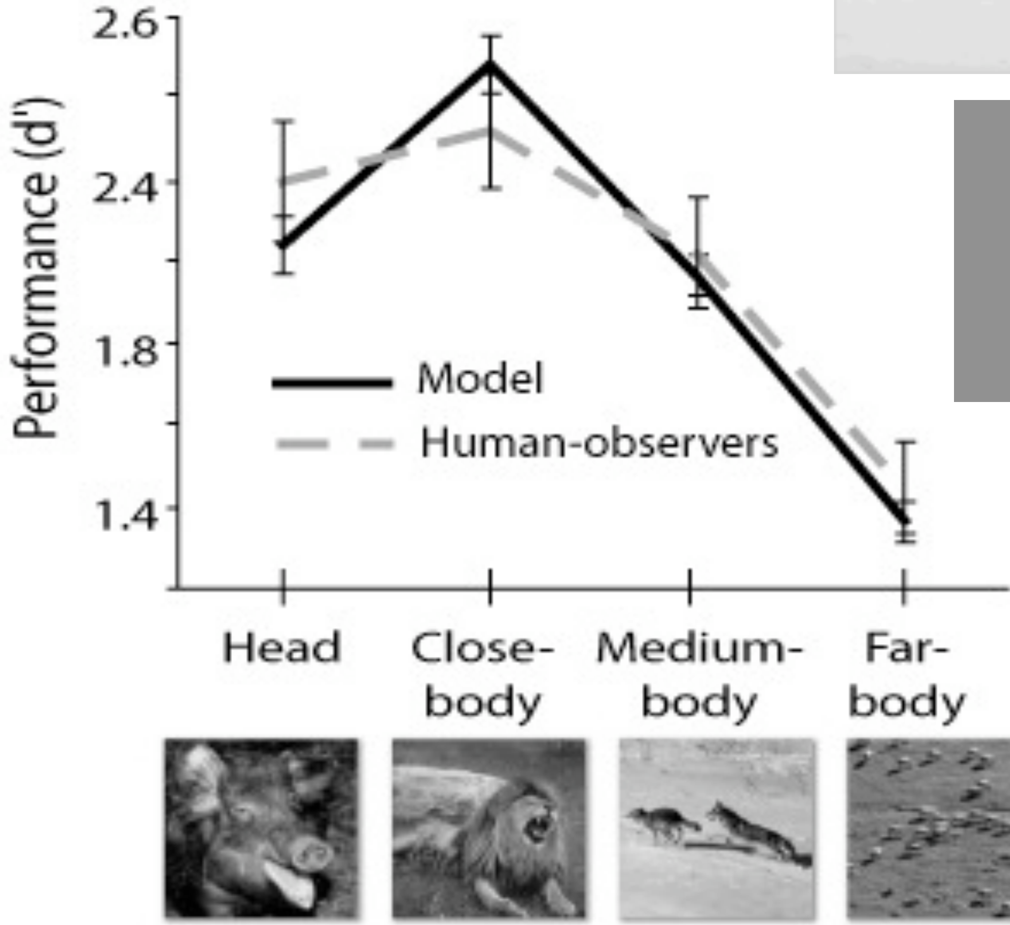


Hierarchical feedforward models of the ventral stream

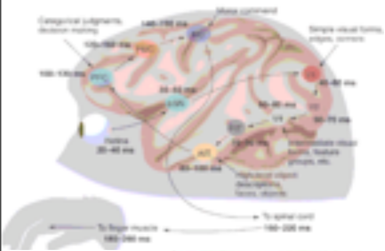


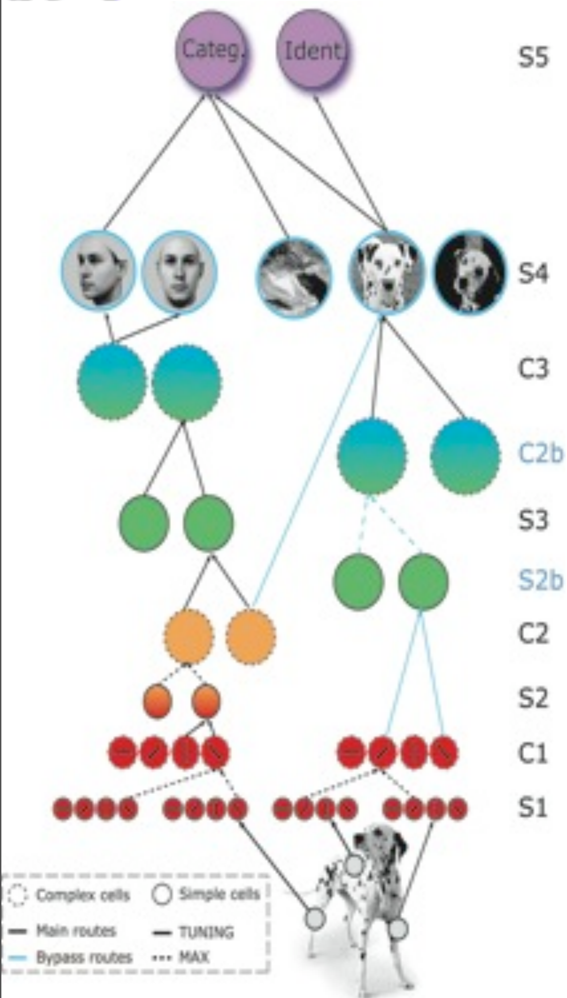


Feedforward Models:
“predict” rapid categorization
(82% model vs. 80% humans)

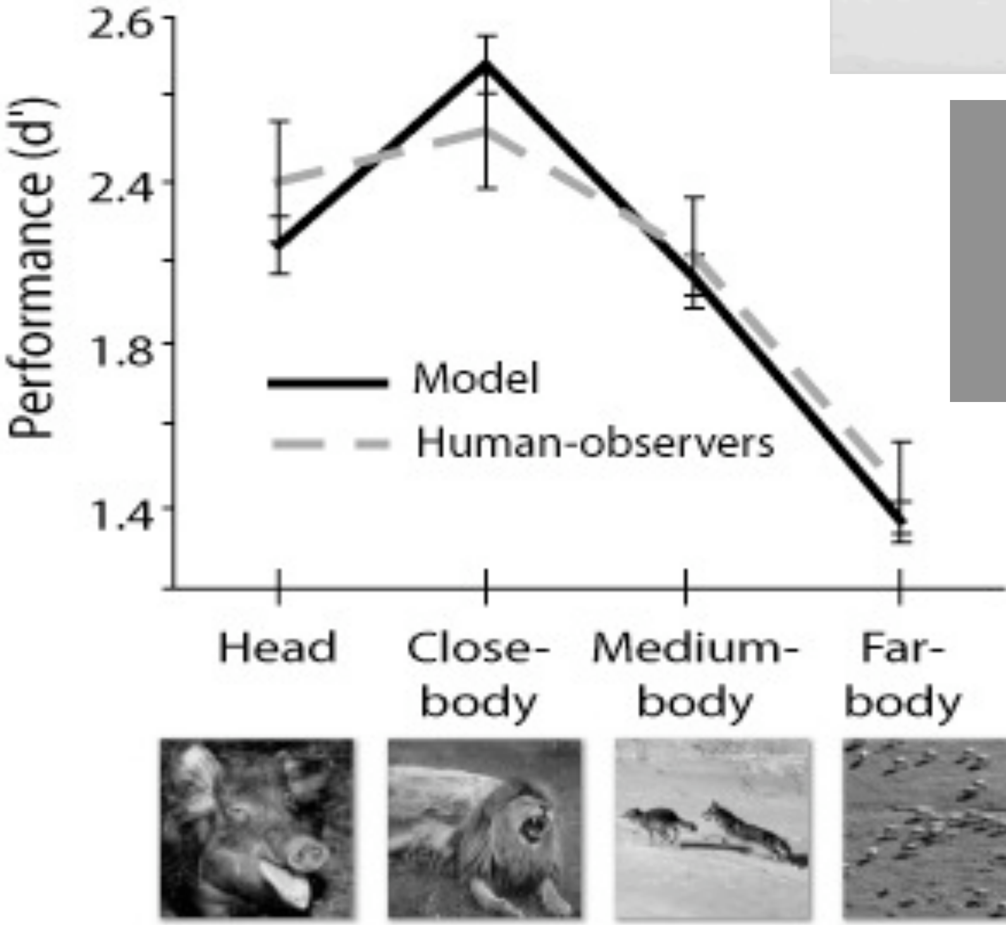


Hierarchical feedforward models of the ventral stream

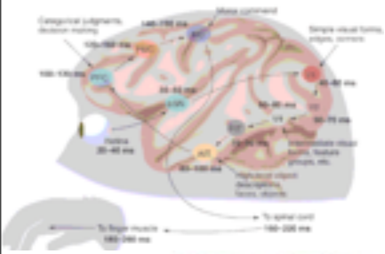


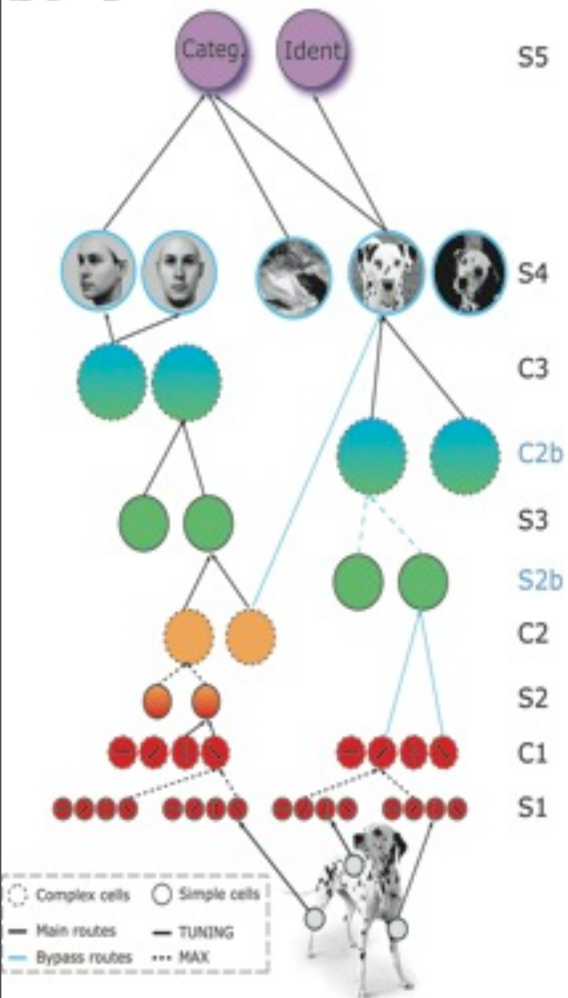


Feedforward Models:
"predict" rapid categorization
(82% model vs. 80% humans)

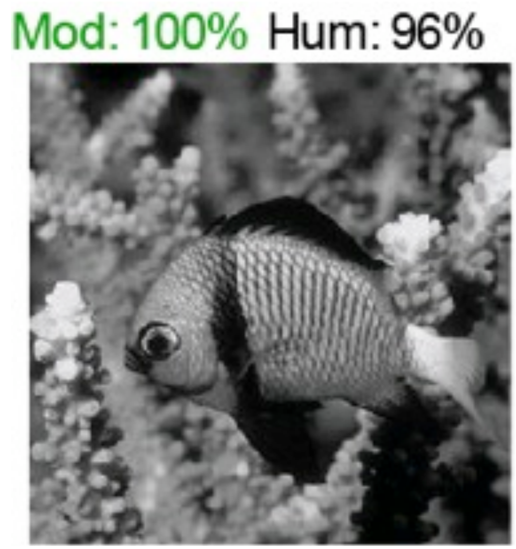


Hierarchical feedforward models of the ventral stream

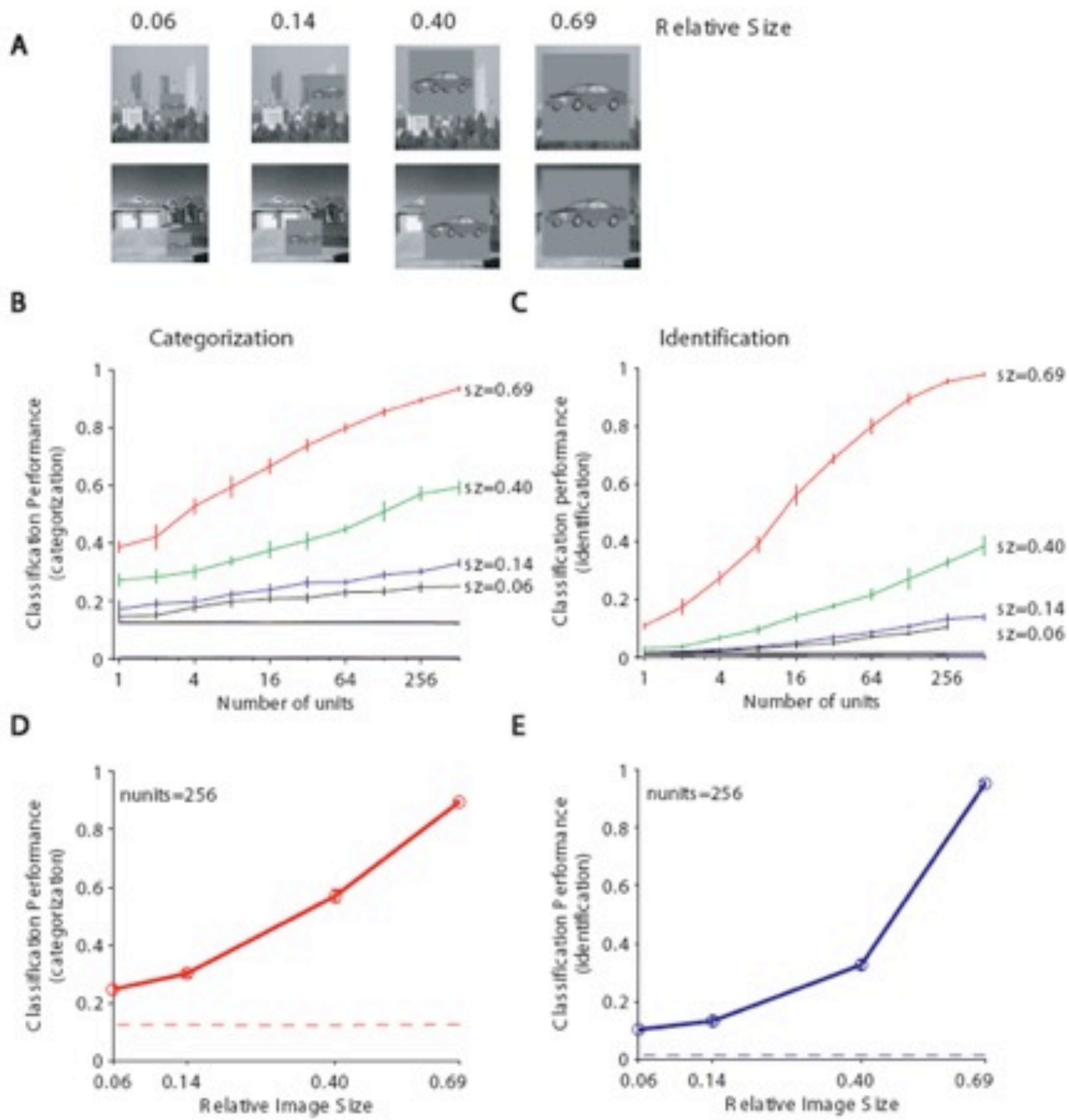




- Image-by-image correlation:
  - Heads: ρ=0.71
  - Close-body: ρ=0.84
  - Medium-body: ρ=0.71
  - Far-body: ρ=0.60



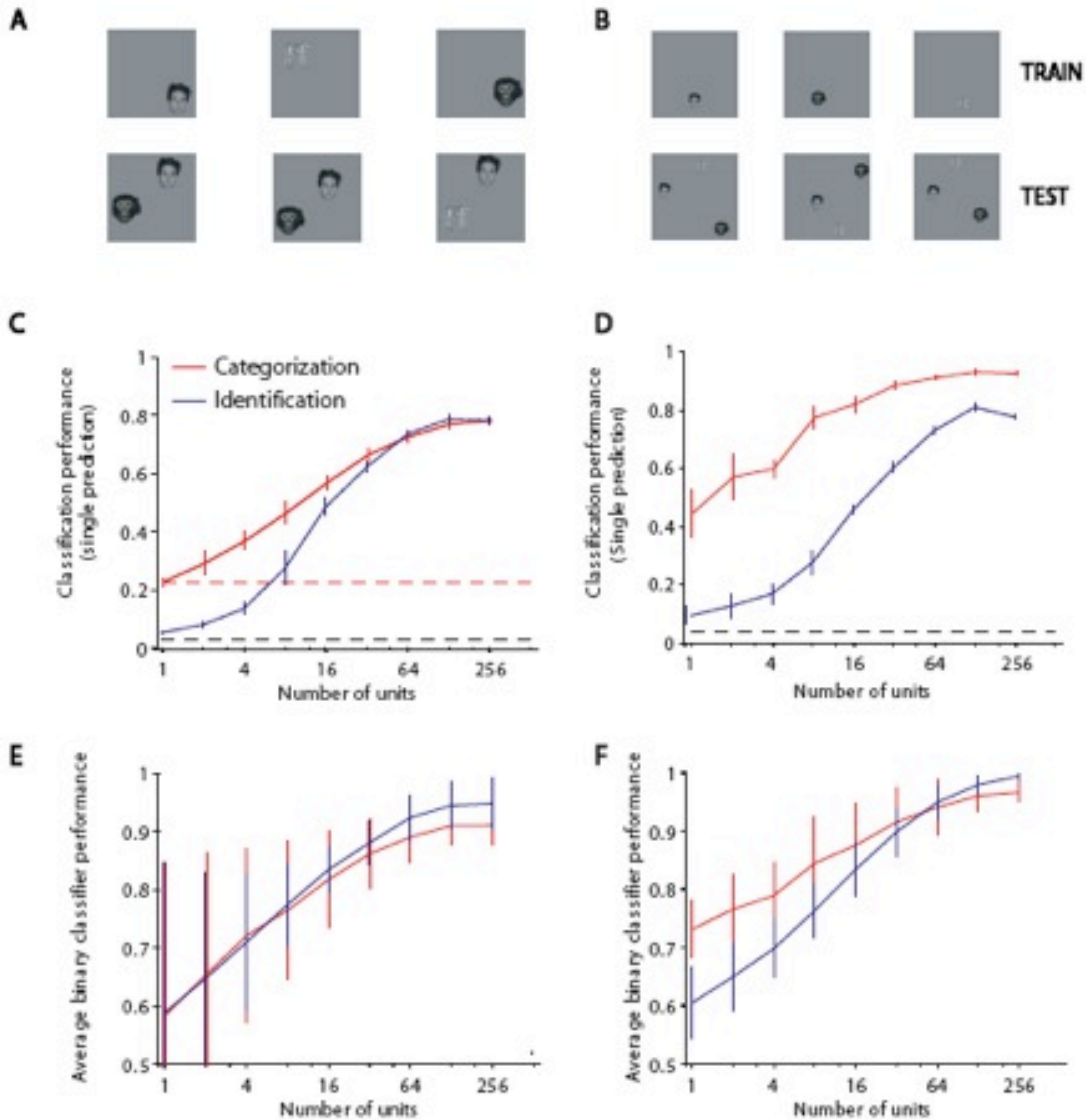
Read-out of object category in clutter



Read-out of object category in clutter

A. Sample of the objects pasted in complex backgrounds. Here we show a single object (a car) out of the 77 objects that were used in this experiment. Here we show the object overlayed onto two different complex background scenes (city landscape, top and house exterior, bottom) out of the 98 different background scenes that we used in this experiment. We did not attempt to generate a ”meaningful” image, objects (including their surrounding gray background) were merely overlayed onto the background scenes. We used four different relative sizes of the object and background images. The center of each object was randomly positioned in the image. **B, C**. Classification performance (**B**. categorization, **C**. identification) as a function of the number of C2 units used to train the classifier. The classifier was trained using 20 % of the 98 backgrounds and the performance was tested with the same objects presented under different backgrounds. Object position within the image was randomized (both for the training and testing images). The different colors correspond to different relative sizes for the object with respect to the background. **D, E**. Classification performance (**D**. categorization, **E**. identification) using 256 units as a function of the relative size of object to background. The horizontal dashed lines indicate chance performance obtained by randomly shuffling the object labels during training.

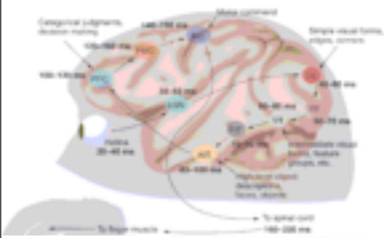
Read-out of object category and identity in images containing multiple objects

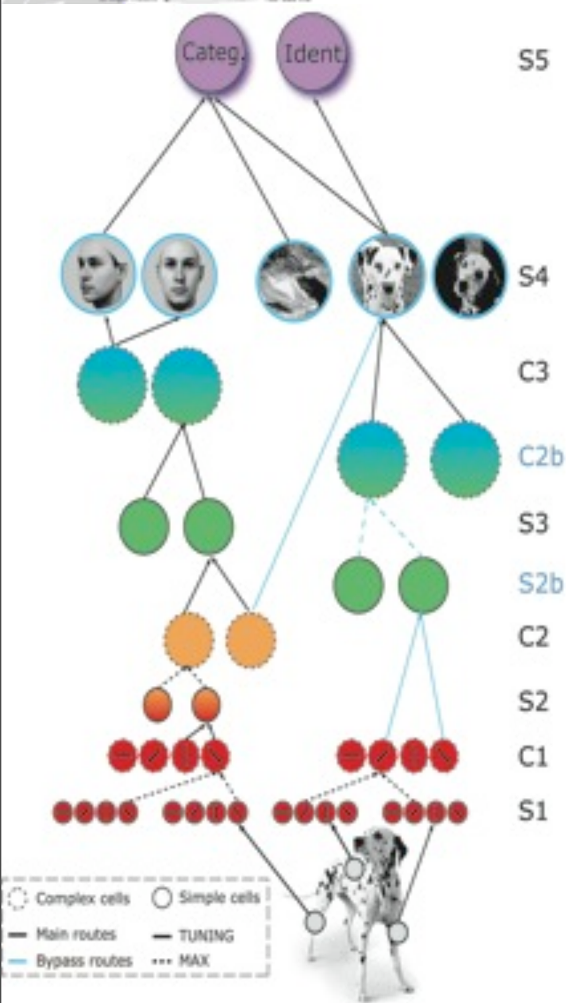


Read-out of object category and identity in images containing multiple objects

Classification performance for reading out object category (red)or object identity (blue) in the presence of two objects (**A, C, E**) or three objects (**B, D, F**). **A, B** Examples of the images used in training (top) and testing (bottom).
Here, we show images containing single objects to train the classifier (top). However, performance was not significantly different when we used images containing multiple objects to train the classifier (see text and Appendix A.9 for details).
C, D Classification performance as a function of the number of C2 units used to train the classifier. Here we used a multi-class classifier approach; the output of the classifier for each test point was a single possible category (or object identity) and a we considered the prediction to be a hit if this prediction matched any of the objects present in the image. The dashed lines show chance performance levels and the error bars correspond to one standard deviation from 20 random choices of which units were used to train the classifier. We exhaustively evaluated every possible object pair or triplet. **E, F** Average performance for each of the binary classifiers as a function of the number of C2 units used for training. The number of binary classifiers was 8 for categorization (red) and 77 for identification (blue). The error bars show one standard deviation over 20 random choices of C2 units.

Hierarchical feedforward models of the ventral stream

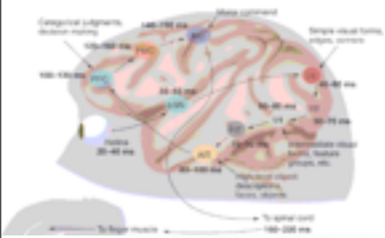


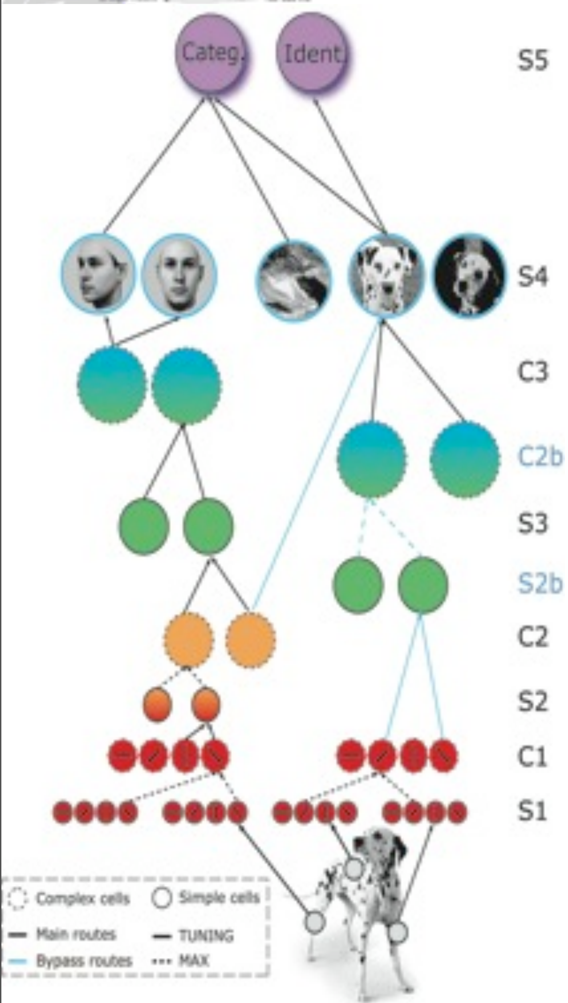


Feedforward Models:
perform well compared to
engineered computer vision systems (in 2006)

Bileschi, Wolf, Serre, Poggio, 2007

Hierarchical feedforward models of the ventral stream

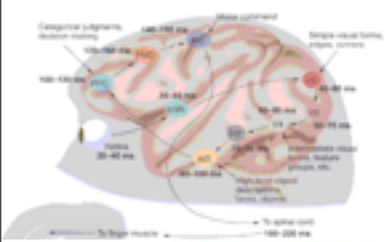


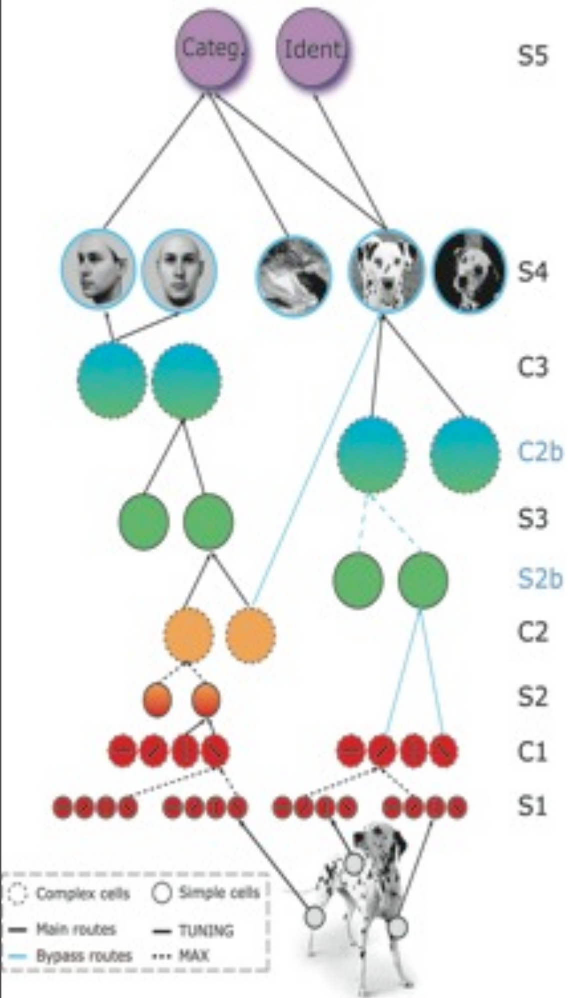


Feedforward Models:
perform well compared to
engineered computer vision systems (in 2006)

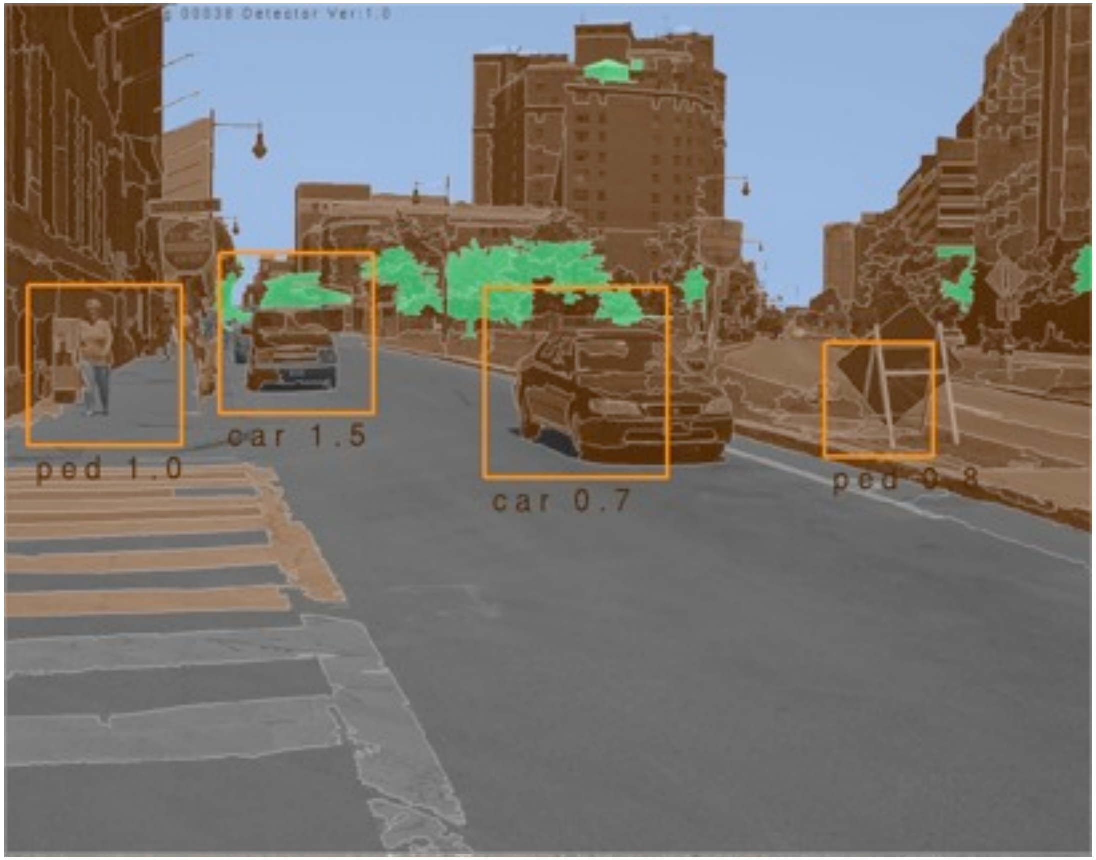
Bileschi, Wolf, Serre, Poggio, 2007

Hierarchical feedforward models of the ventral stream





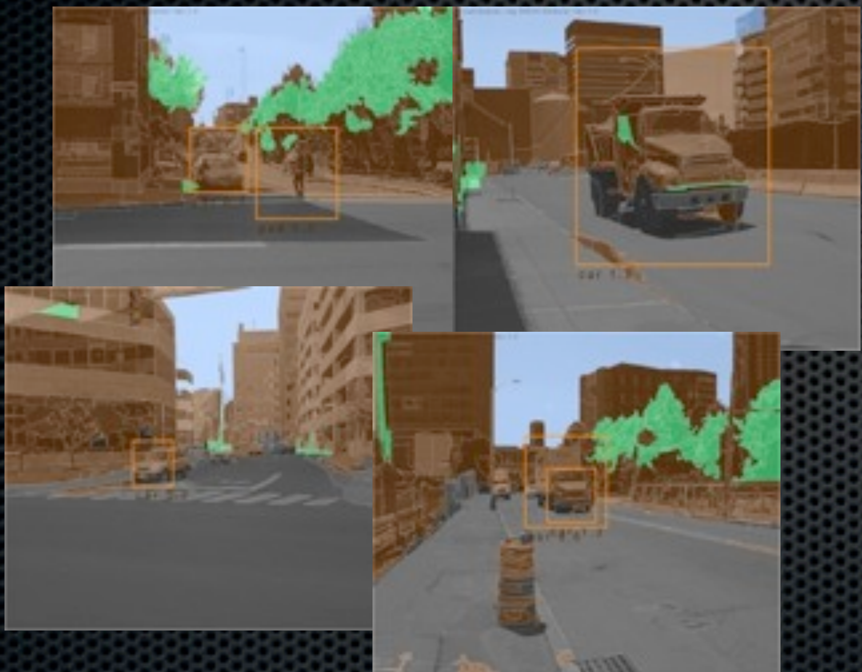
Feedforward Models:
perform well compared to
engineered computer vision systems (in 2006)



Bileschi, Wolf, Serre, Poggio, 2007

Bio-motivated computer vision

Scene parsing and object recognition



Serre Wolf & Poggio 2005; Wolf & Bileschi 2006; Serre et al 2007

Speed improvement since 2006

image size	multi-thread	GPU (cuda)
64x64	4.5x	14x
128x128	3.5x	14x
256x256	1.5x	17x
512x512	2.5x	25x

From ~1 min down to ~1 sec !!

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

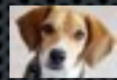
A detailed description of much of the work is in the "supermemo" at

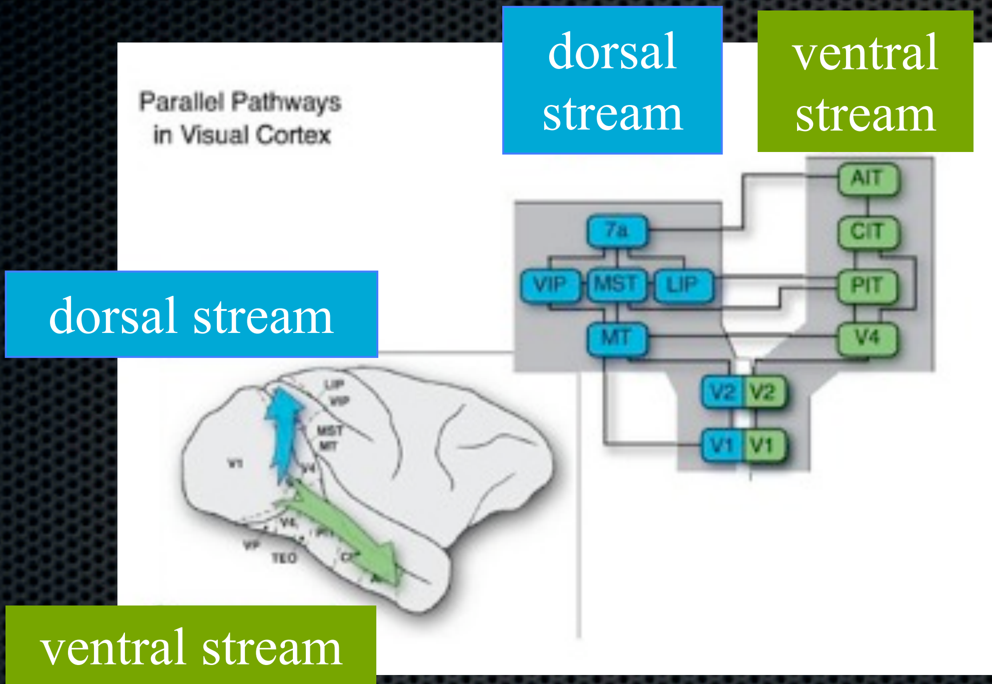
http://cbcl.mit.edu/projects/cbcl/publications/ai-publications/2005/AIM-2005-036.pdf

Other recent publications and references can be found at

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

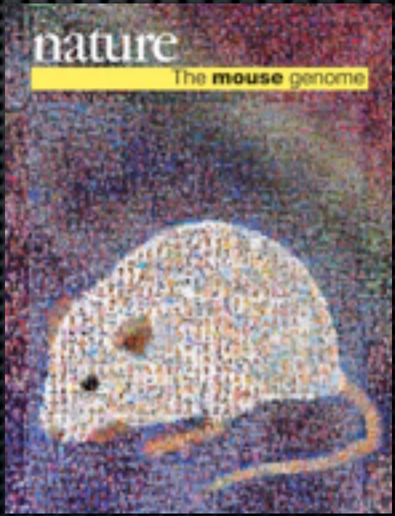
Model extension to the dorsal stream: Recognition of actions





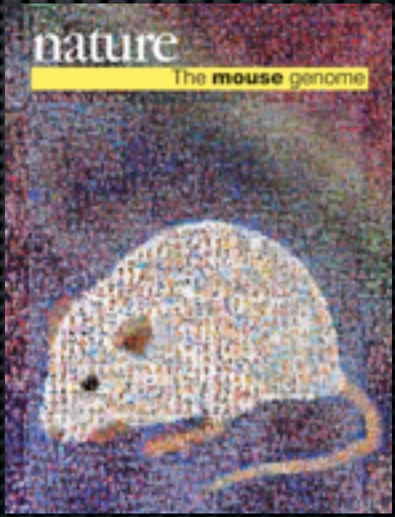
Thomas Serre, Hueihan Jhuang & Tomaso Poggio collaboration with David Sheinberg at Brown University

Quantitative automatic phenotyping



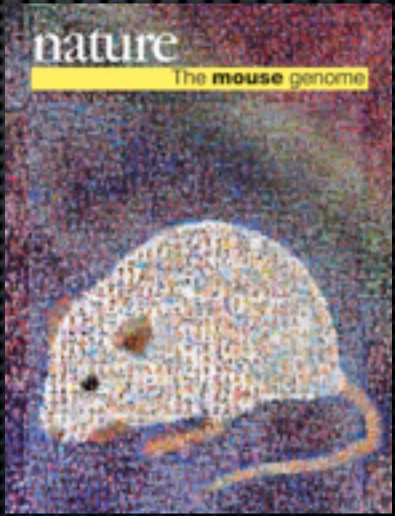
Quantitative automatic phenotyping

- Behavioral analyses of mouse behavior needed to:



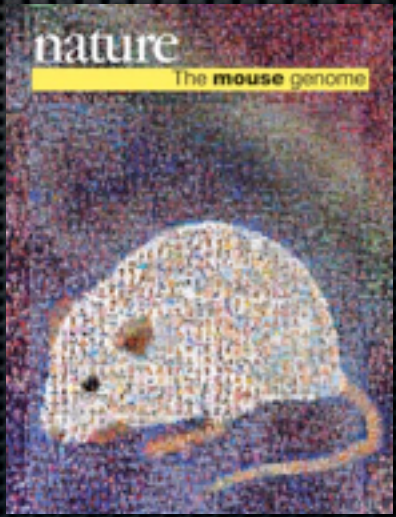
Quantitative automatic phenotyping

- Behavioral analyses of mouse behavior needed to:
  - Assess functional roles of genes



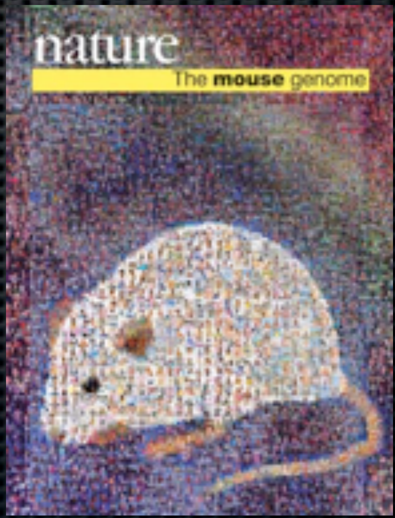
Quantitative automatic phenotyping

- Behavioral analyses of mouse behavior needed to:
  - Assess functional roles of genes
  - Validate models of mental diseases



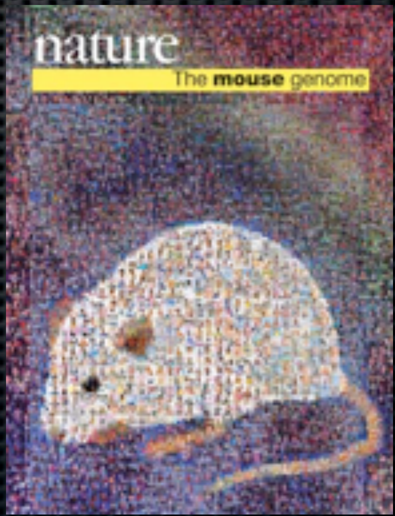
Quantitative automatic phenotyping

- Behavioral analyses of mouse behavior needed to:
  - Assess functional roles of genes
  - Validate models of mental diseases
  - Help assess efficacy of drugs



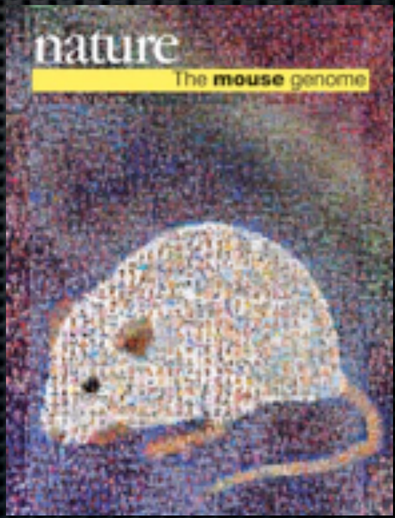
Quantitative automatic phenotyping

- Behavioral analyses of mouse behavior needed to:
  - Assess functional roles of genes
  - Validate models of mental diseases
  - Help assess efficacy of drugs
- Automated quant system to help:



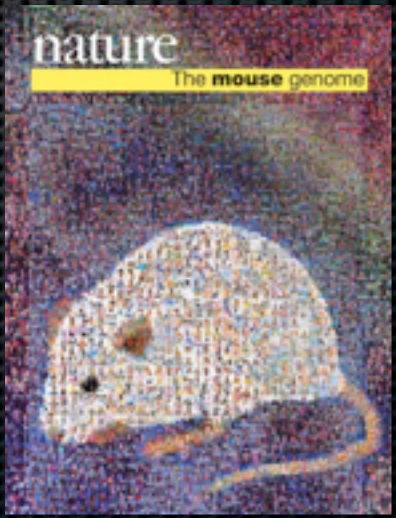
Quantitative automatic phenotyping

- Behavioral analyses of mouse behavior needed to:
  - Assess functional roles of genes
  - Validate models of mental diseases
  - Help assess efficacy of drugs
- Automated quant system to help:
  - Limit subjectivity of human intervention



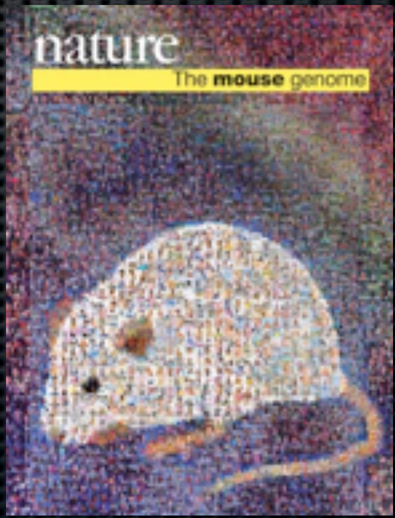
Quantitative automatic phenotyping

- Behavioral analyses of mouse behavior needed to:
  - Assess functional roles of genes
  - Validate models of mental diseases
  - Help assess efficacy of drugs
- Automated quant system to help:
  - Limit subjectivity of human intervention
  - 24/7 home-cage analysis of behavior



Quantitative automatic phenotyping

- Behavioral analyses of mouse behavior needed to:
  - Assess functional roles of genes
  - Validate models of mental diseases
  - Help assess efficacy of drugs
- Automated quant system to help:
  - Limit subjectivity of human intervention
  - 24/7 home-cage analysis of behavior
  - 24/7 monitoring of animal well-being



More on models of the dorsal stream: action recognition and applications

Hueihan Jhuang

Hierarchical feedforward models of visual cortex may be wrong …but present a challenge for "classical" learning theory:

an unusual, <u>hierarchical</u> architecture with unsupervised and supervised learning working well.
But...ironically, we do not understand why these models work well
(see LeCun, Poggio, Hinton...)

…so,
we need theories -- not just models!

Theory of Hierarchical Learning Machines (HLM)

GOAL:
Hierarchical architectures to preprocess images/signals in order to reduce the sampling complexity of a classifier trained with labeled examples.
The hierarchical architecture is synthesized from a large number of unsupervised examples.

Joint work with Steve Smale, Jake Bouvrie, Andrea Caponnetto, Lorenzo Rosasco

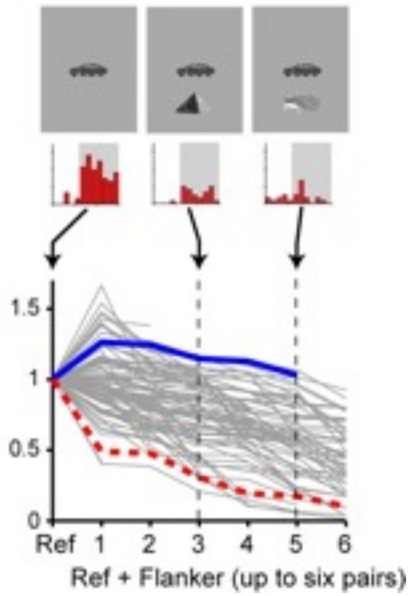
Mathematics of the Neural Response, J. Foundations of Comp. Mathematics, 2009

HLMs:

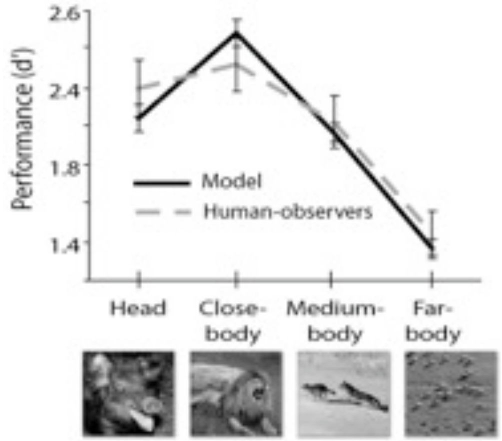
a mathematical framework for hierarchical learning machines

Lorenzo Rosasco + Andre Wibisono: Class 16

Extension to attention: dealing with clutter



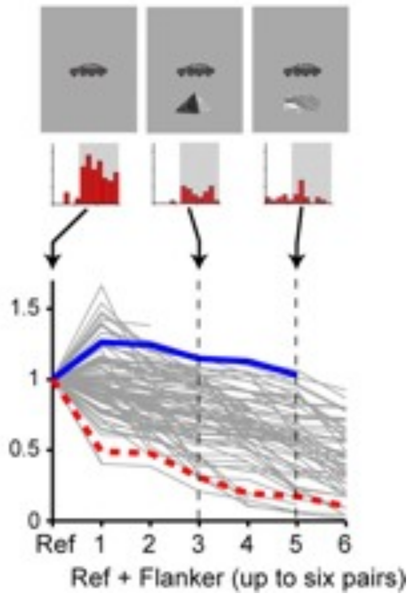
Zoccolan Kouh Poggio DiCarlo 2007



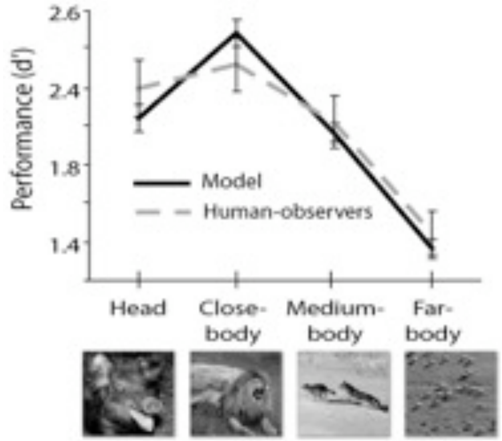
Serre Oliva Poggio 2007

see also Broadbent 1952 1954; Treisman 1960; Treisman & Gelade 1980; Duncan & Desimone 1995; Wolfe, 1997; Tsotsos and many others

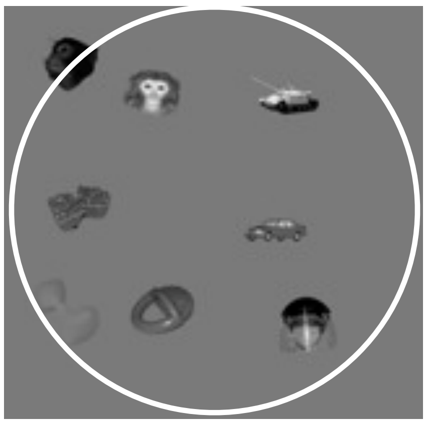
Extension to attention: dealing with clutter

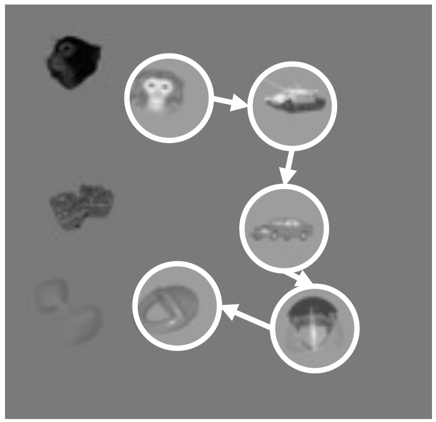


Zoccolan Kouh Poggio DiCarlo 2007



Serre Oliva Poggio 2007

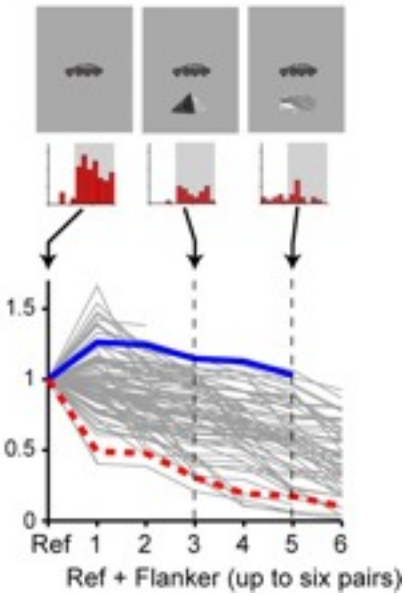




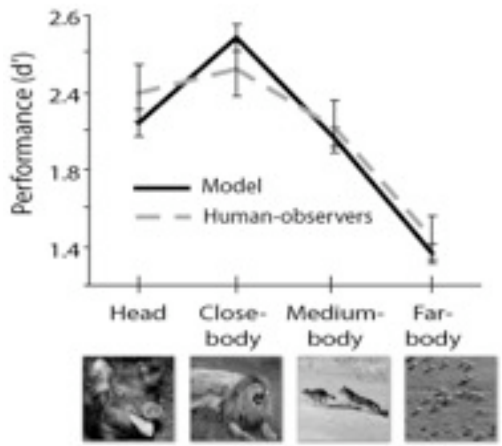
Parallel processing (No attention)

see also Broadbent 1952 1954; Treisman 1960; Treisman & Gelade 1980; Duncan & Desimone 1995; Wolfe, 1997; Tsotsos and many others

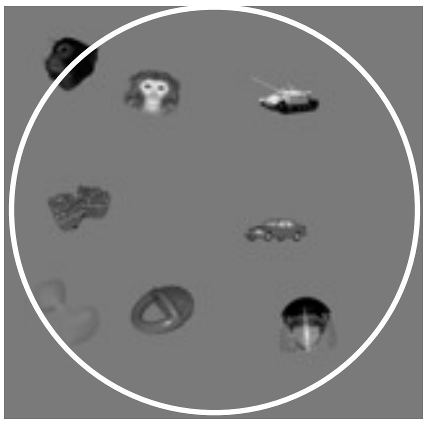
Extension to attention: dealing with clutter



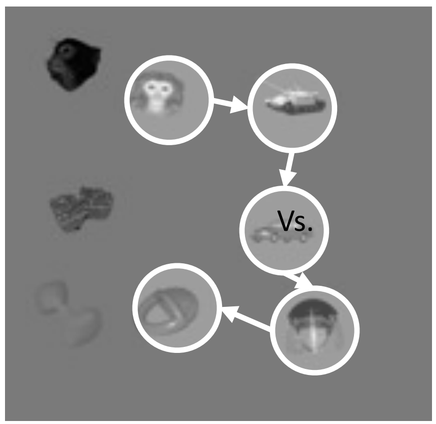
Zoccolan Kouh Poggio DiCarlo 2007



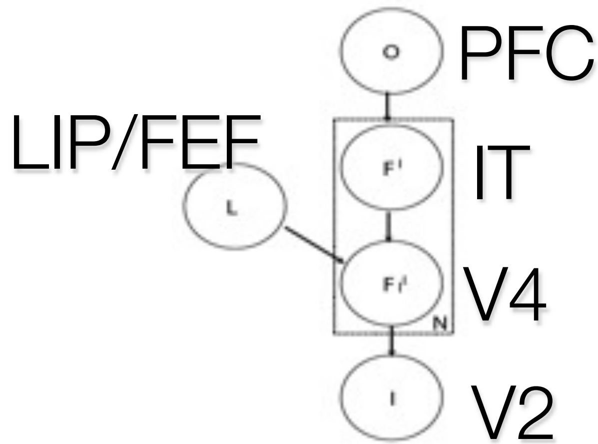
Serre Oliva Poggio 2007



Parallel processing (No attention)



Serial processing (With attention)



see also Broadbent 1952 1954; Treisman 1960; Treisman & Gelade 1980; Duncan & Desimone 1995; Wolfe, 1997; Tsotsos and many others

Extending feedforward models with an additional attention module

Sharat Chikkerur: Class 17

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

A Turing Test for Vision

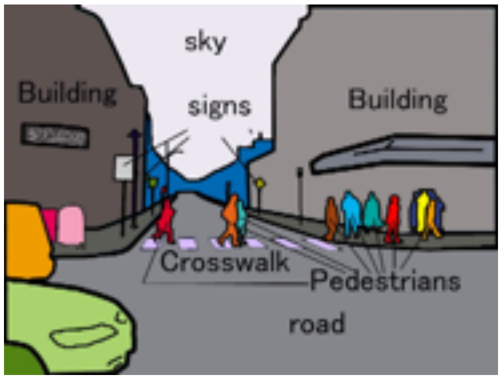
- Vision is <u>more</u> than categorization or identification: it is image understanding/inference/parsing

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

A Turing Test for Vision

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

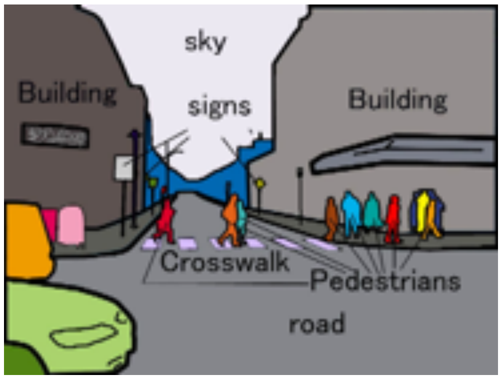
- Our visual system can “answer” almost any kind of question about an image or video (a Turing test for vision…)



A Turing Test for Vision

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

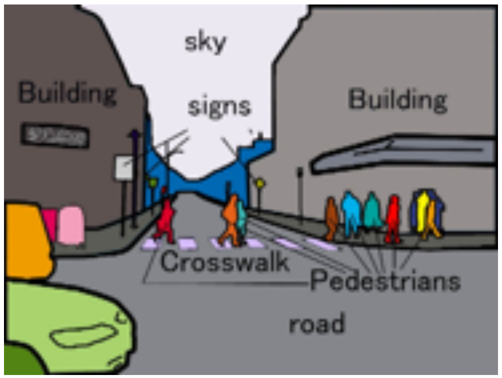
- Our visual system can "answer" almost any kind of question about an image or video (a Turing test for vision…)



A Turing Test for Vision

- 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…)



Collaborators

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