

Class 17

Visual Recognition in Primates and Machines

Tomaso Poggio (with Thomas Serre)

Plan for class 16 and 21

□ Class 17: “coarse” description of

--a model that accounts for properties of neurons in the visual cortex

--a model that accounts for human recognition of complex images

□ Class 18: “finer” level of description and questions

□ Class 20: Mathematical framework: towards a theory of learning in cortex

Motivation for studying vision:

trying to understand how the brain works

- Old dream of all philosophers and more recently of AI:
 - understand how the brain works
 - make intelligent machines



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.

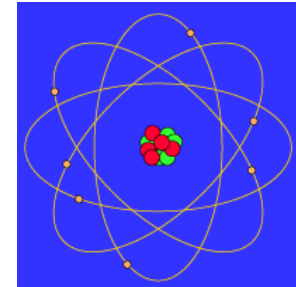
□ **Why hierarchies?** There may be reasons of *efficiency* – computational speed and use of computational resources. For instance, the lowest levels of the hierarchy may represent a dictionary of features that can be shared across multiple classification tasks.

□ 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.

This tutorial:

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

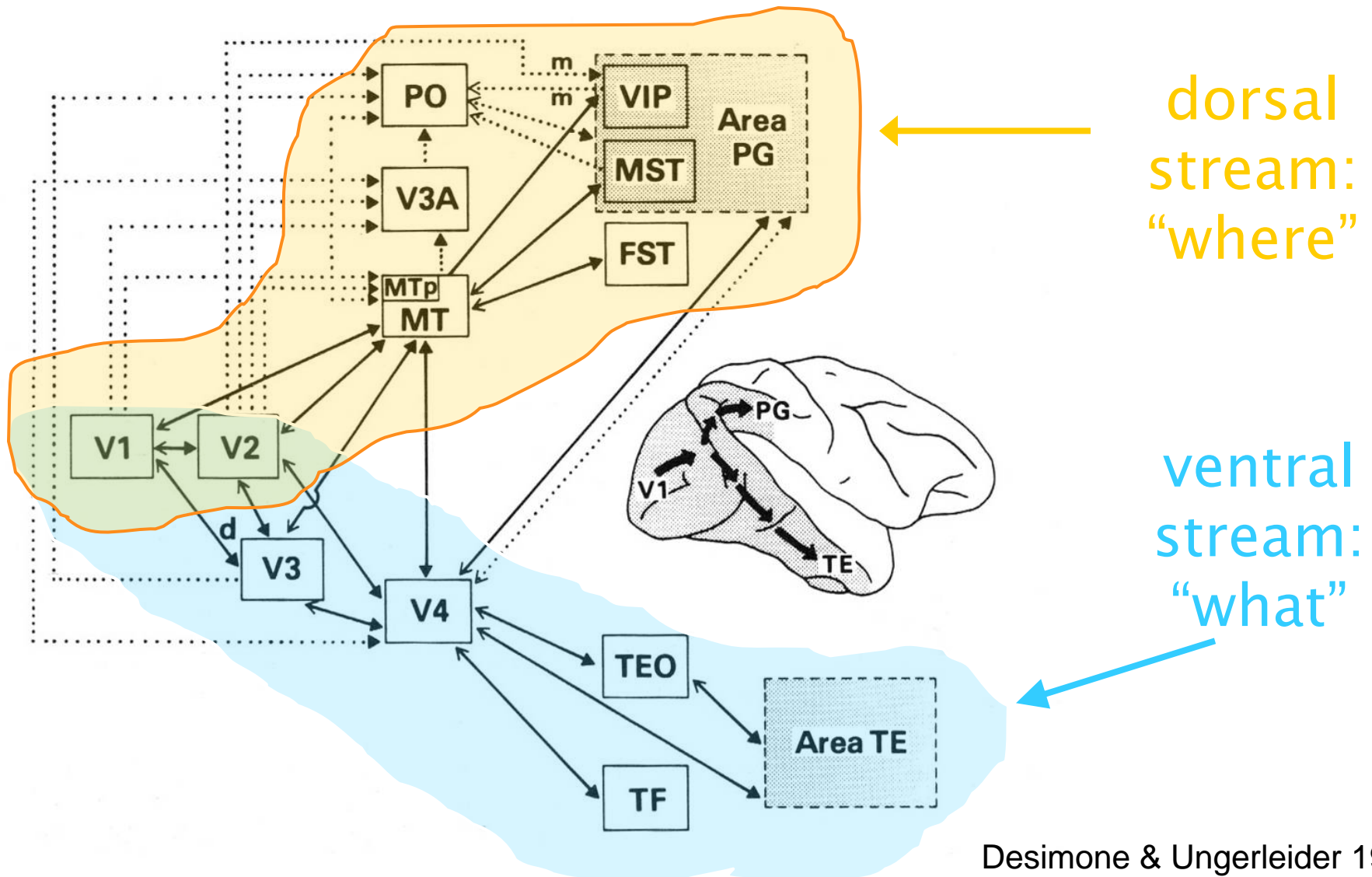
- Models are cartoons of reality, eg Bohr's model of the hydrogen atom



- 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. Data and feedforward hierarchical models
5. What is next?

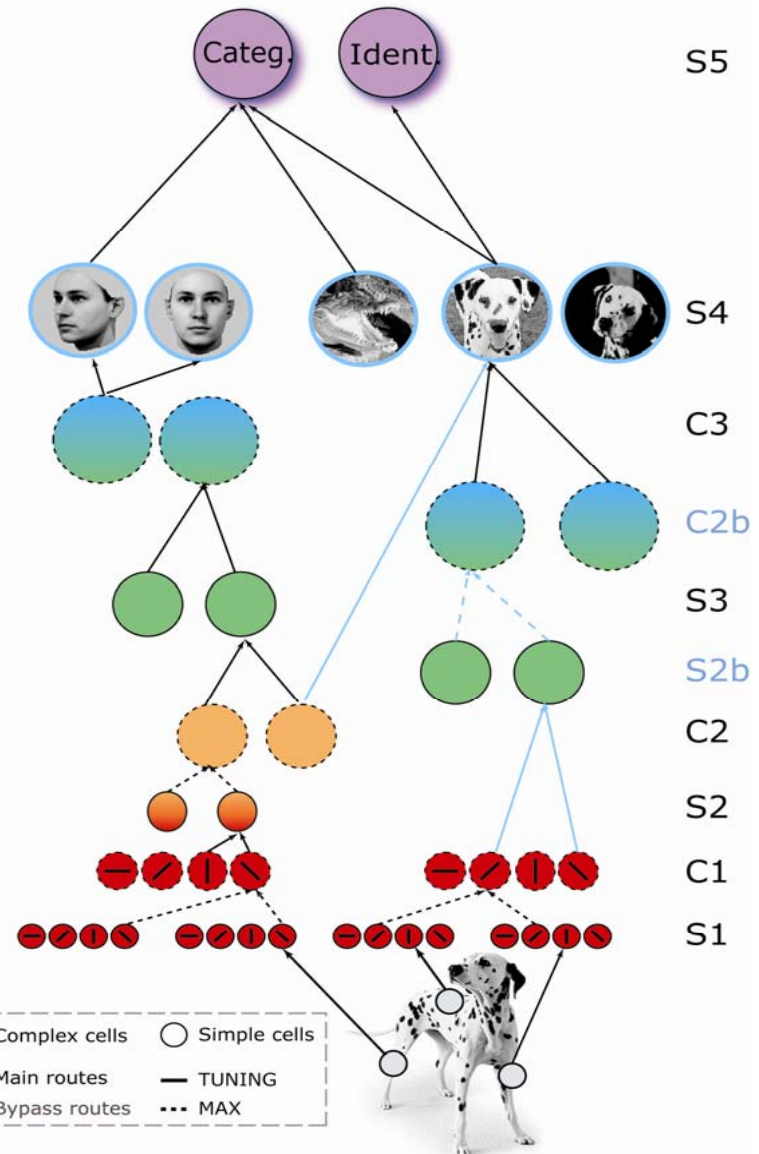
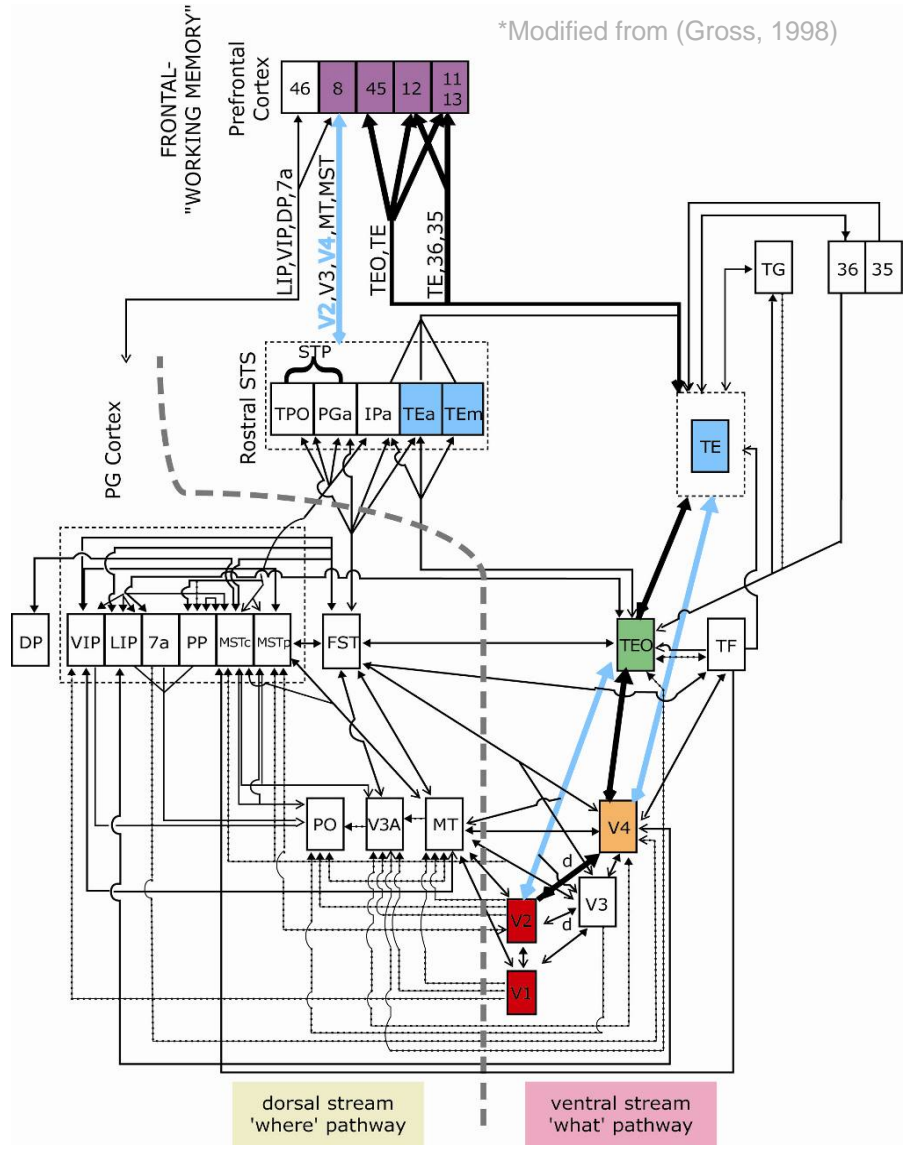
How does visual cortex solve this problem?
How can computers solve this problem?



A “feedforward” version of the problem: rapid categorization (RVSP)



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



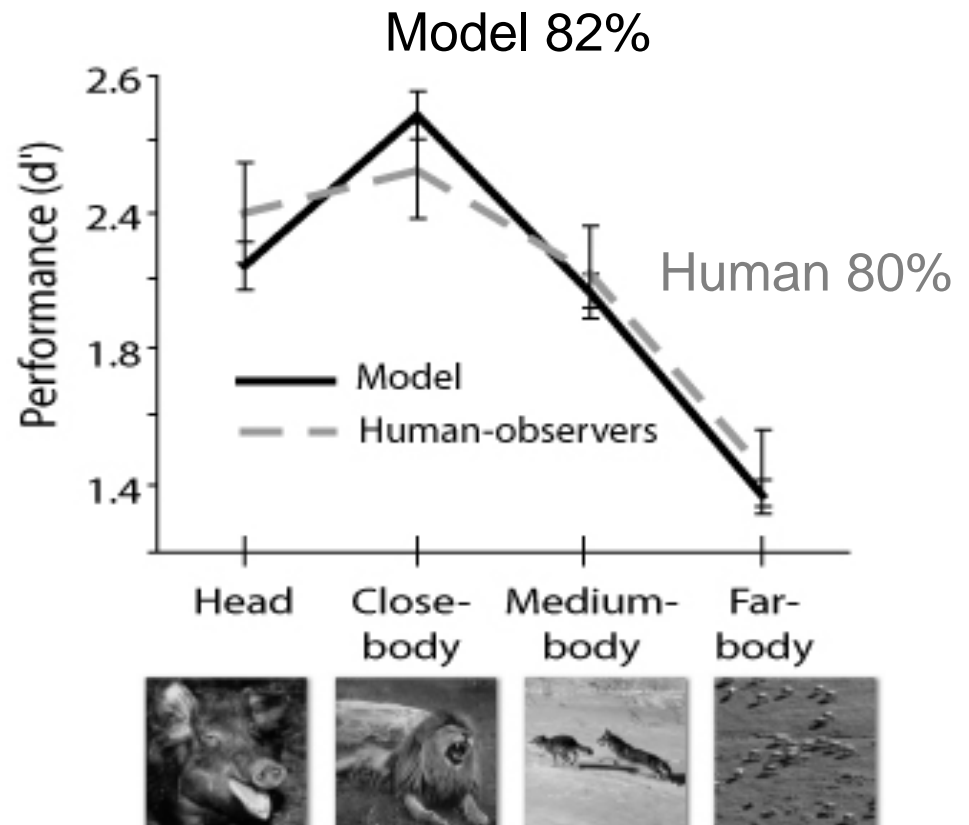
Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich
 Kreiman & Poggio 2005; Serre Oliva Poggio 2007

[software available online]

...”solves” the problem

(if the mask forces feedforward processing)...

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



1. Problem of visual recognition, visual cortex
2. **Historical background**
3. Neurons and areas in the visual system
4. Data and feedforward hierarchical models
5. What is next?

Object recognition for computer vision: (personal) historical perspective

Face detection

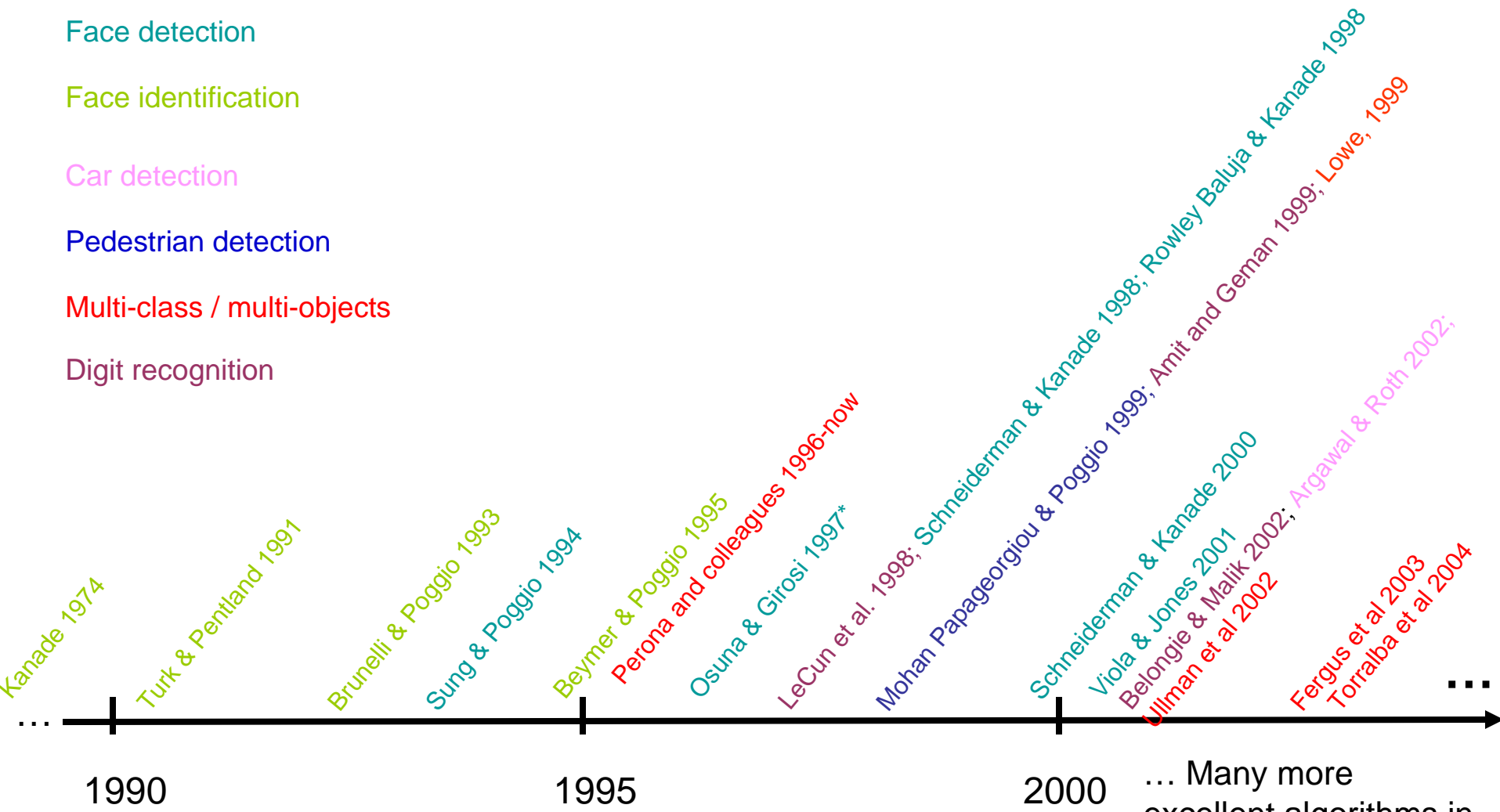
Face identification

Car detection

Pedestrian detection

Multi-class / multi-objects

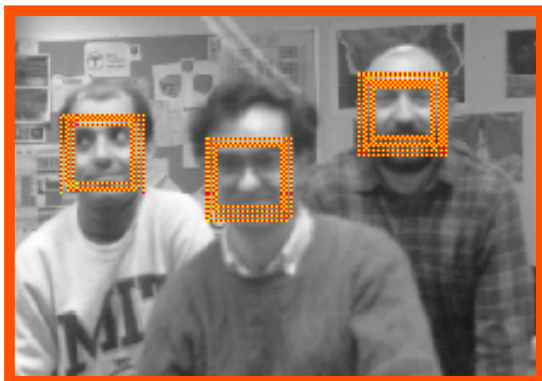
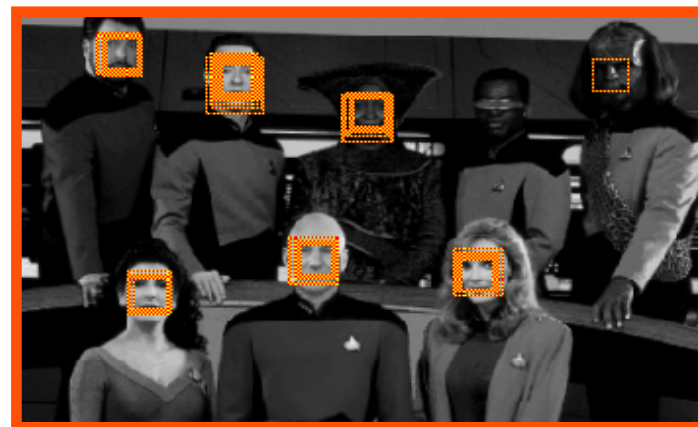
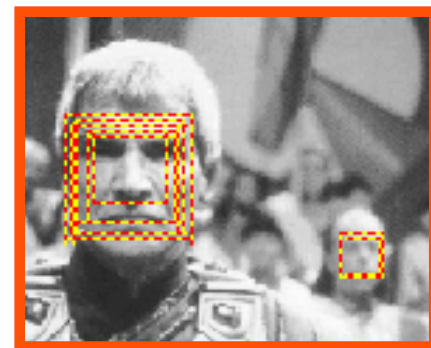
Digit recognition



*Best CVPR'07 paper 10 yrs ago

Examples: Learning Object Detection: Finding Frontal Faces

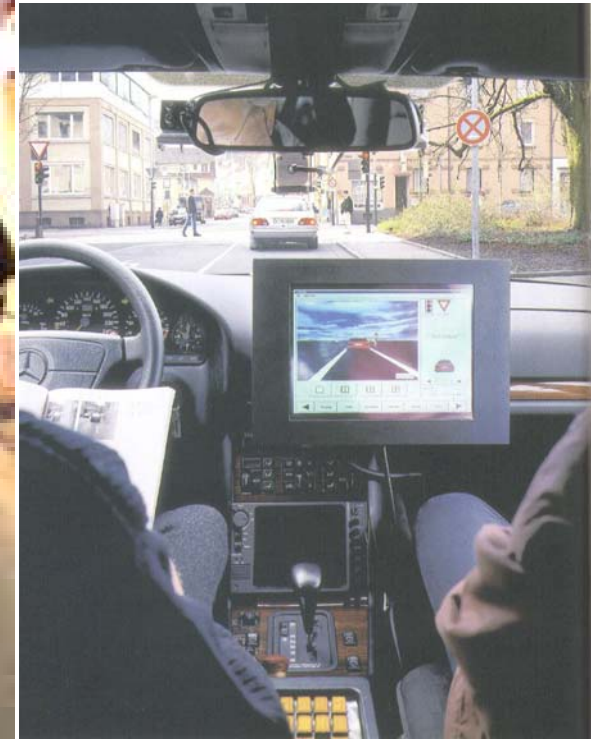
- Training Database
- 1000+ Real, 3000+ VIRTUAL
- 50,000+ Non-Face Pattern



~10 year old CBCL computer vision work:

SVM-based pedestrian detection system in Mercedes
test car...

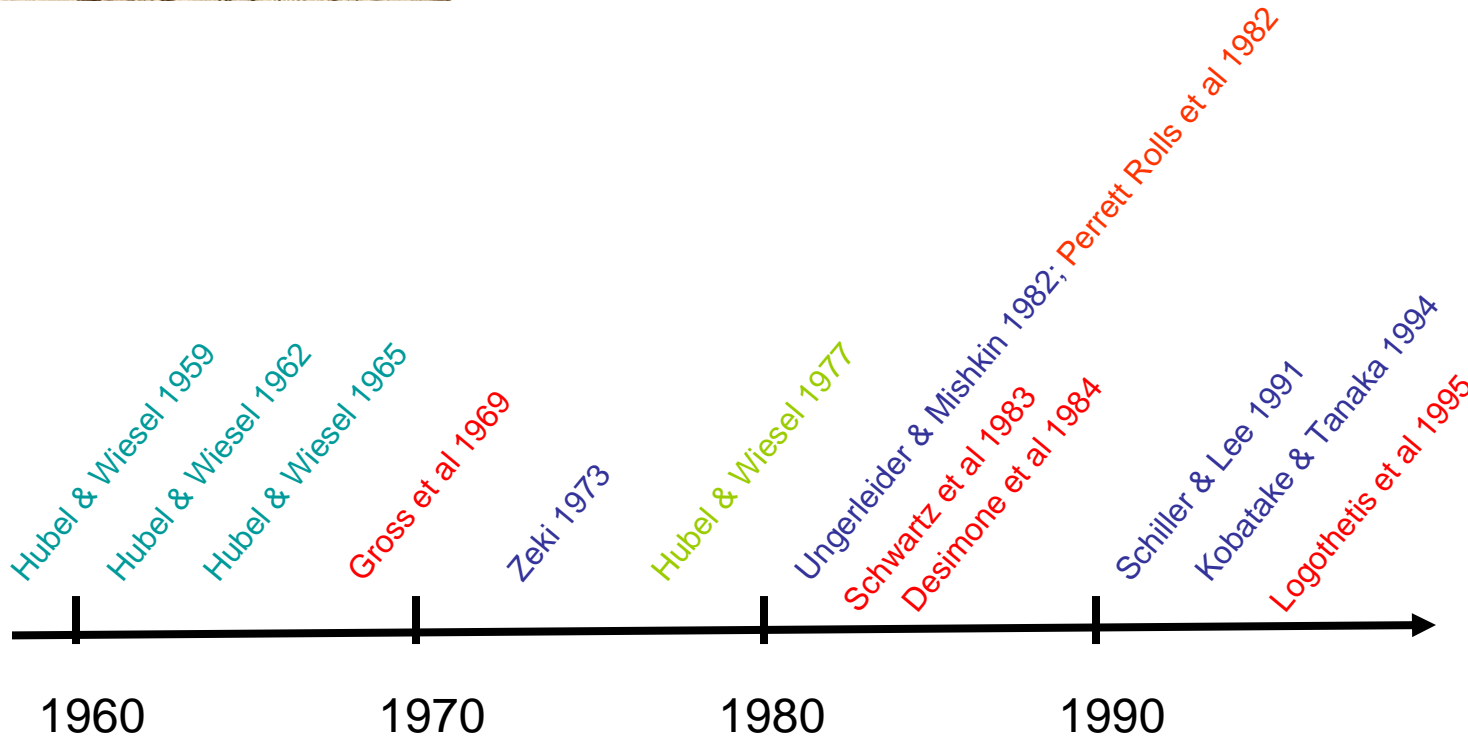
now becoming a product (MobilEye)



Object recognition in cortex: Historical perspective



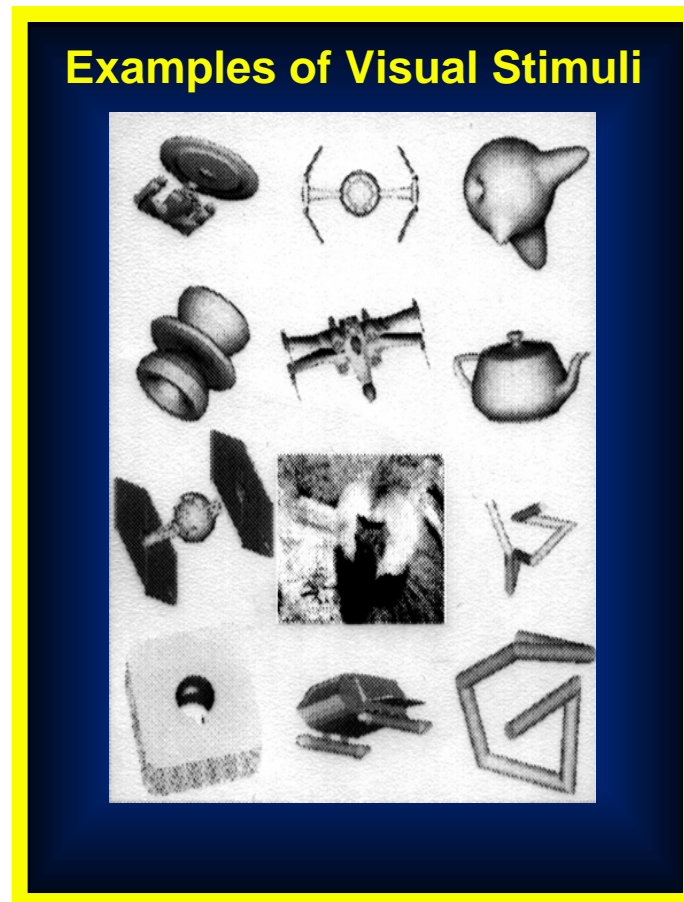
V1 cat
V1 monkey
Extrastriate cortex
IT-STS



... Much
progress in the
past 10 yrs

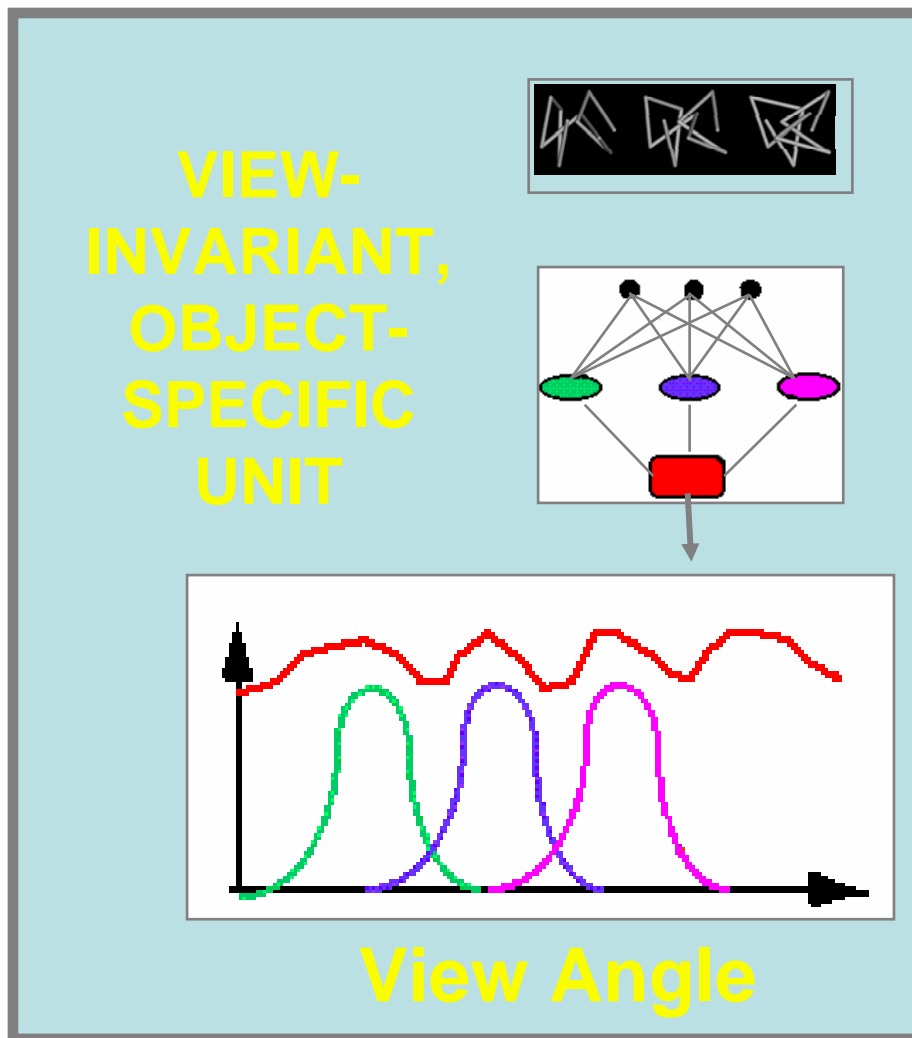
Some personal history:

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



An idea for a module for view-invariant identification

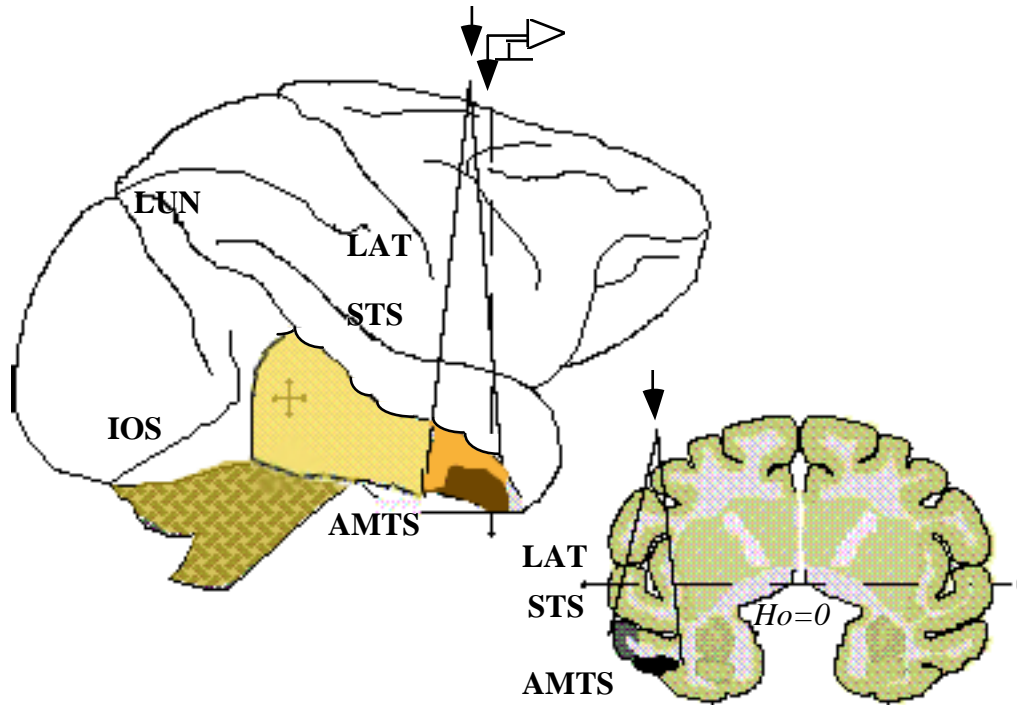
Architecture that accounts for invariances to 3D effects (>1 view needed to learn!)



Prediction:
neurons become
view-tuned
through learning

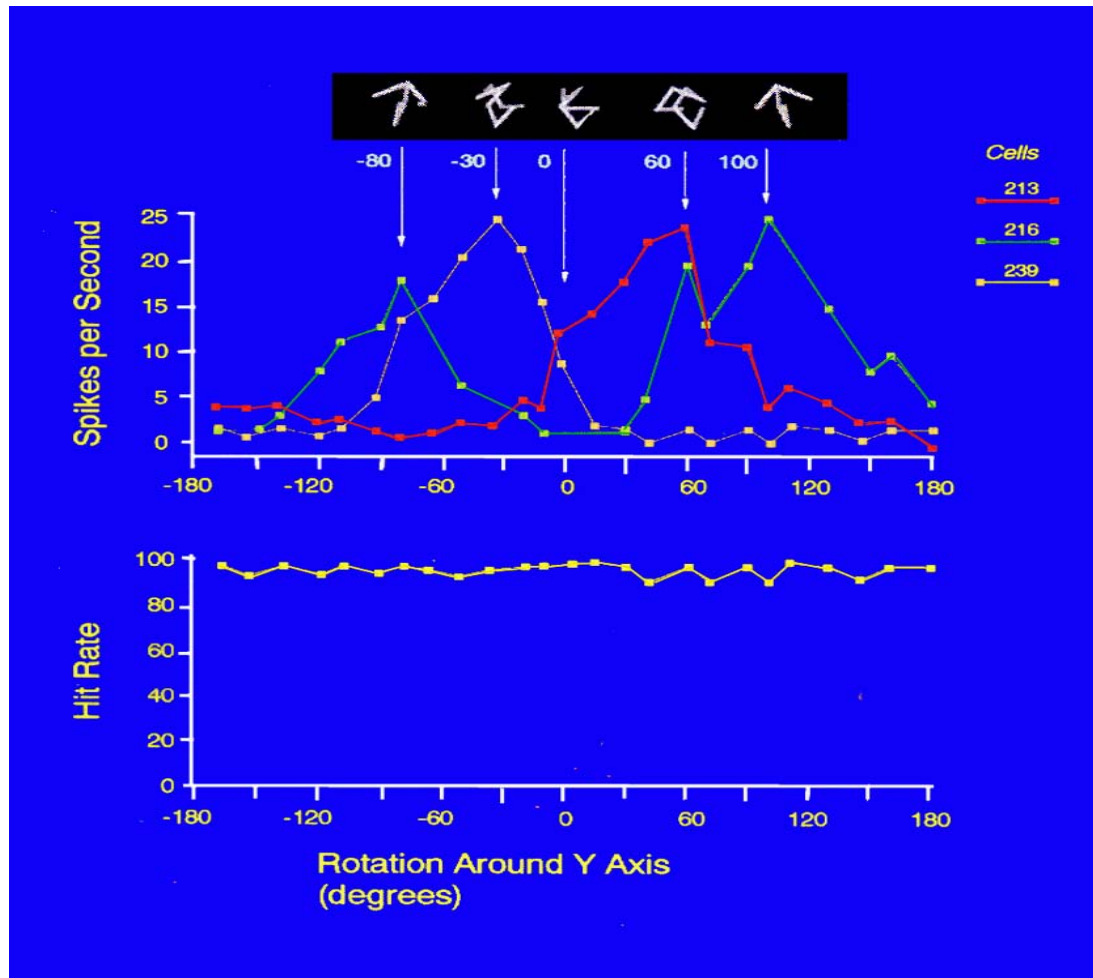
Regularization
Network (GRBF)
with Gaussian kernels

Recording Sites in Anterior IT



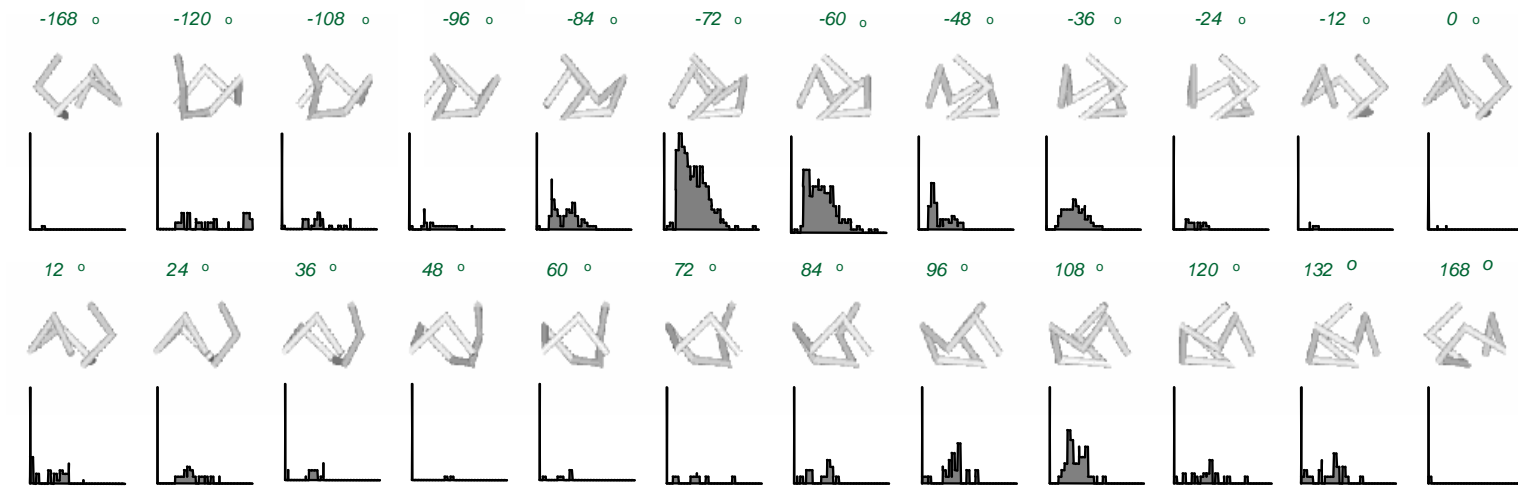
...neurons tuned to faces are intermingled nearby....

Neurons tuned to object views, as predicted by model!

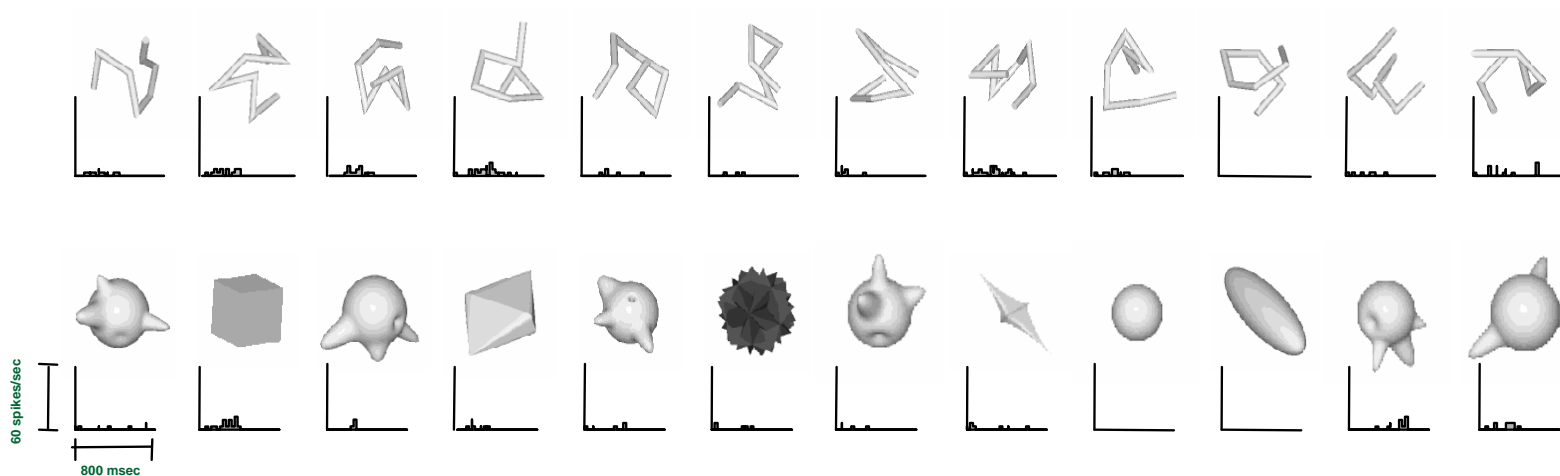


A "View-Tuned" IT Cell

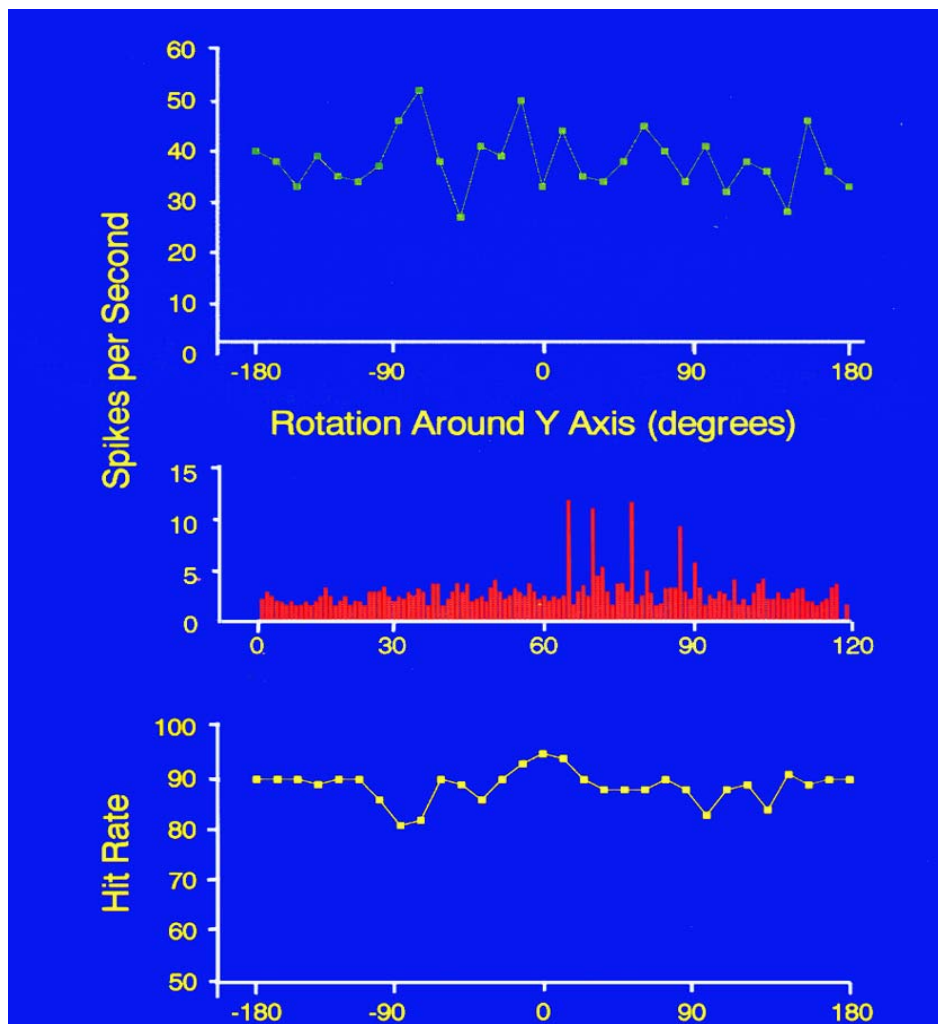
Target Views



Distractors

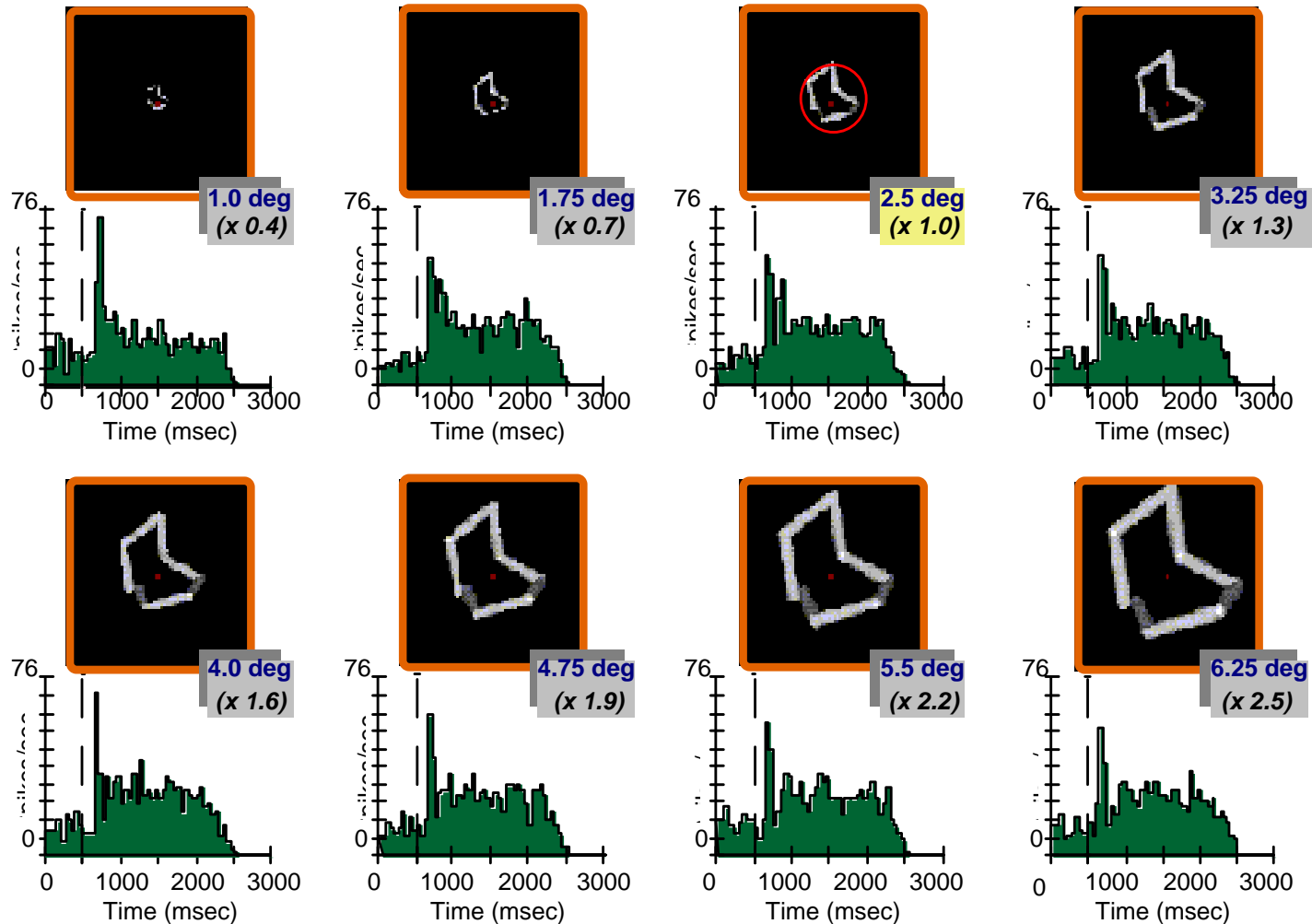


But also view-invariant object-specific neurons
(5 of them over 1000 recordings)

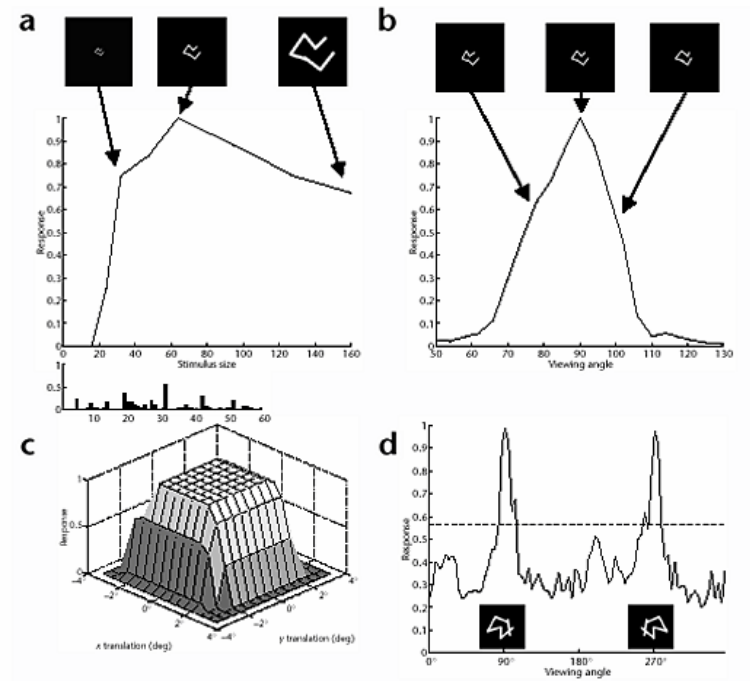
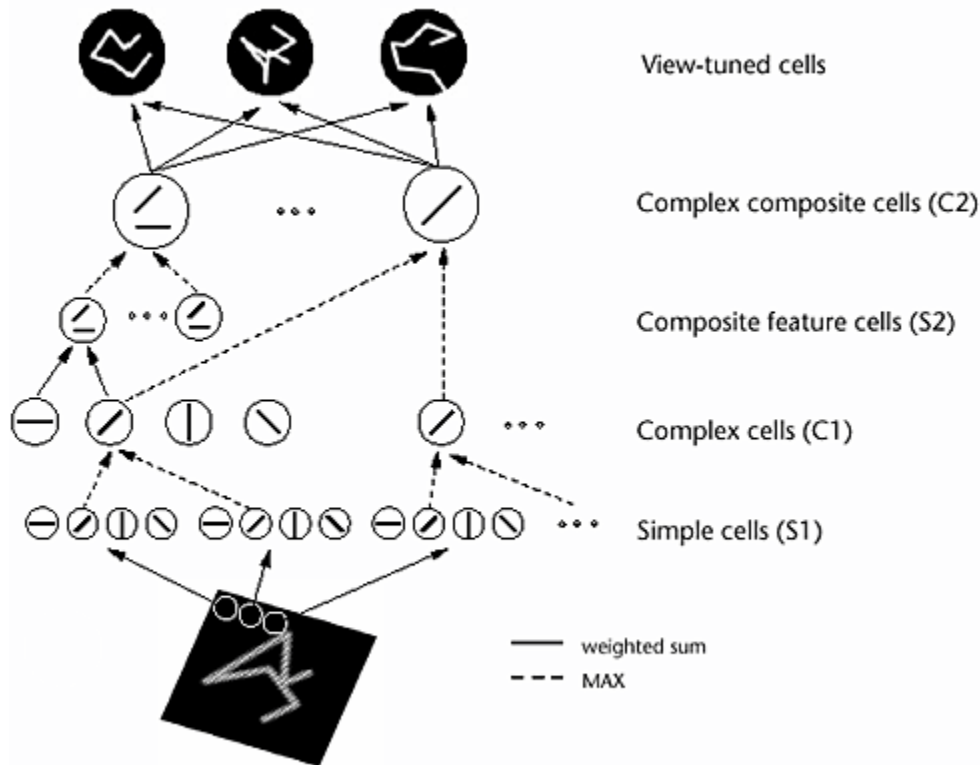


View-tuned cells:

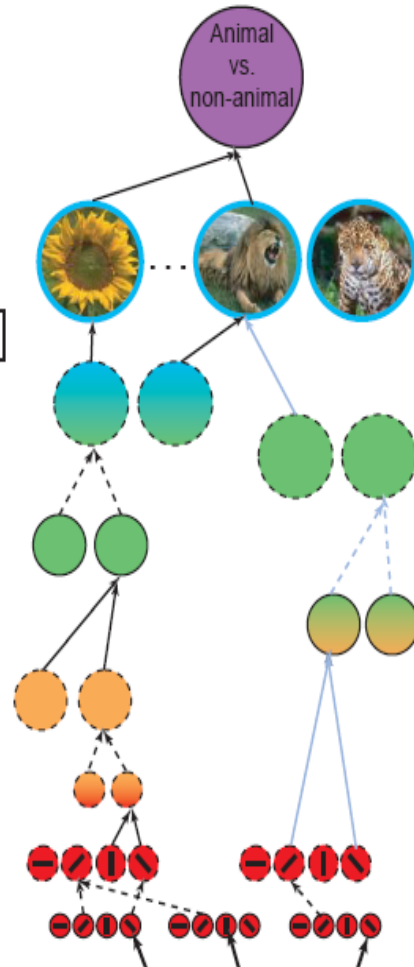
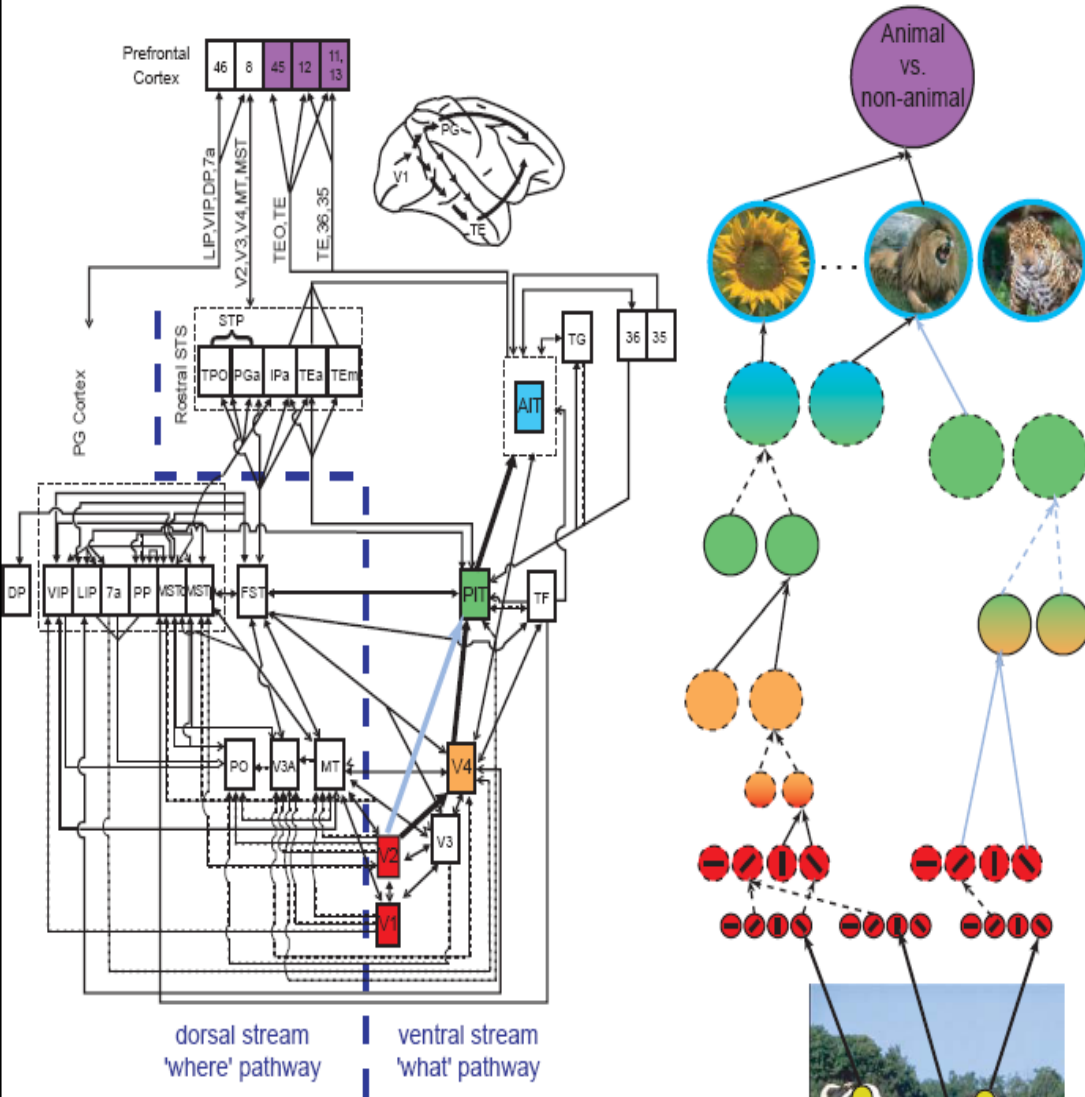
scale invariance (one training view only) motivates present model



The "HMAX" model

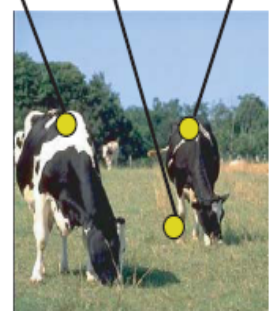


How to obtain selectivity and invariance: from "HMAX" to the model now...



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

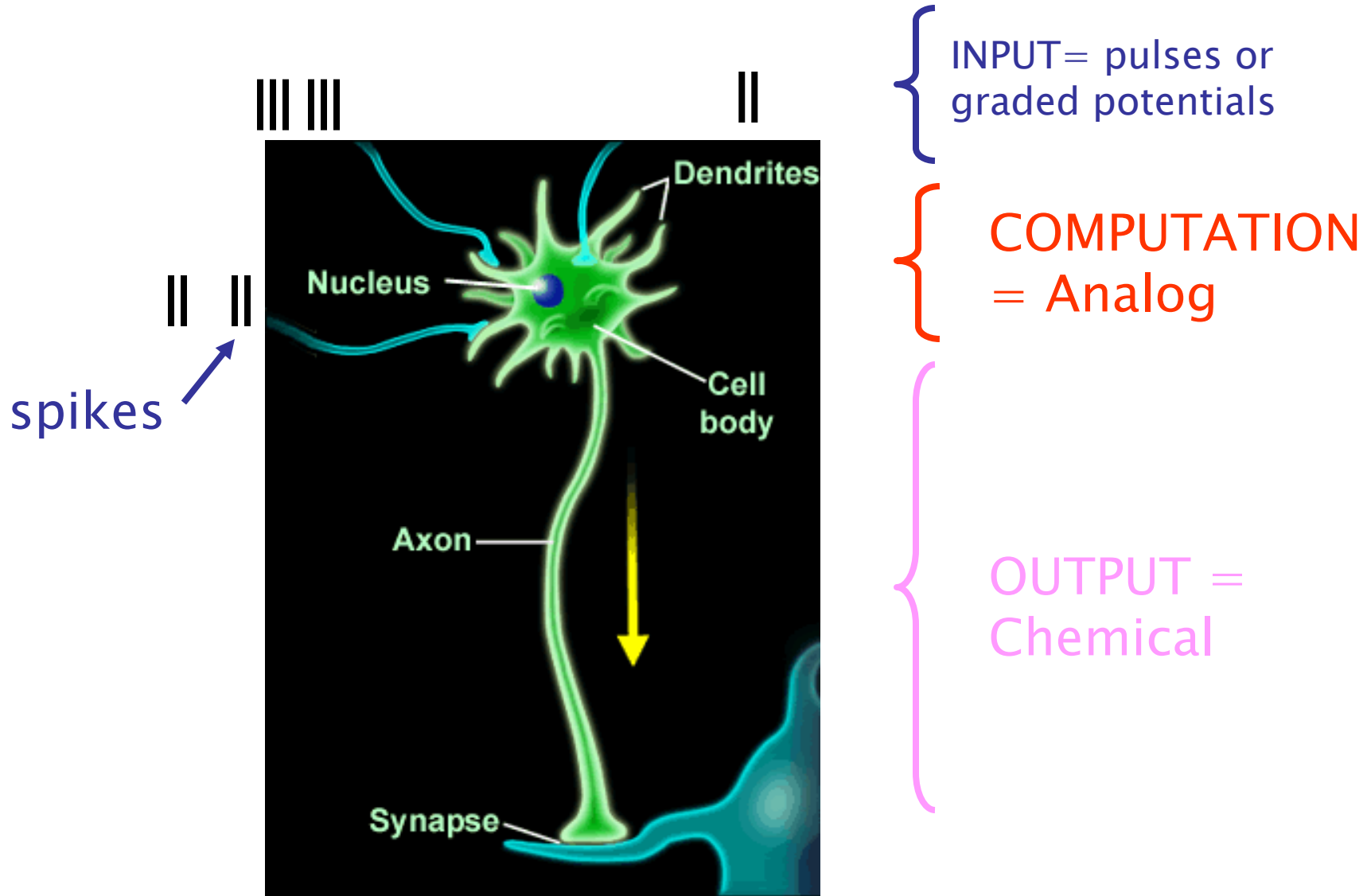
↑ **Supervised task-dependent learning**
 ↓ **Unsupervised task-independent learning**
 ↑ increase in complexity (number of subunits), RF size and invariance



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

1. Problem of visual recognition, visual cortex
2. Historical background
3. **Neurons and areas in the visual system**
4. Data and feedforward hierarchical models
5. What is next?

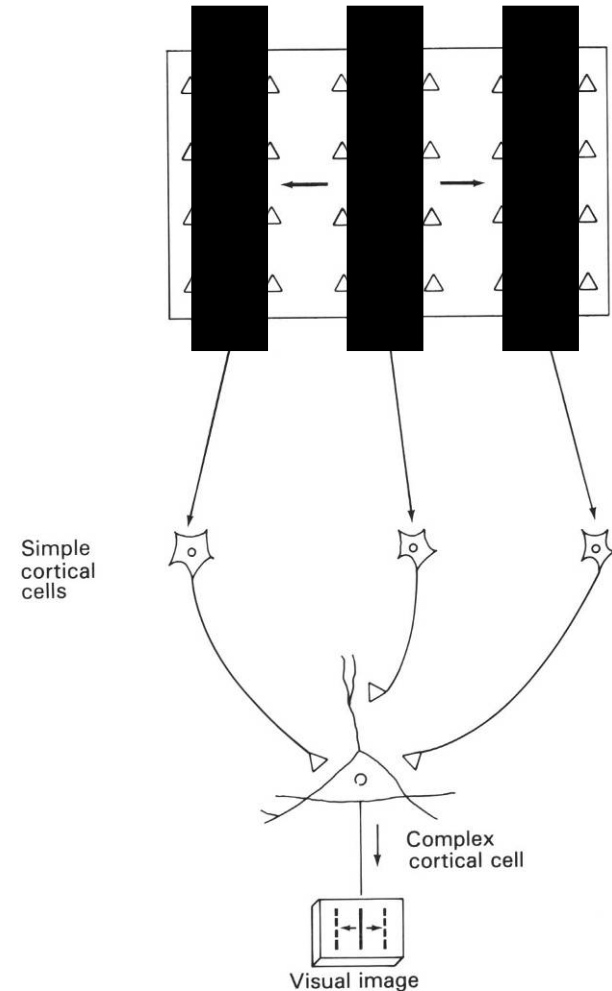
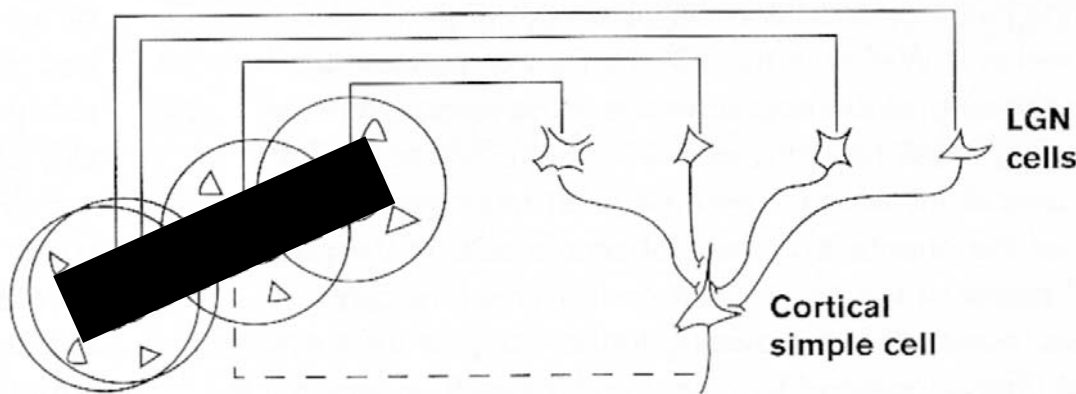
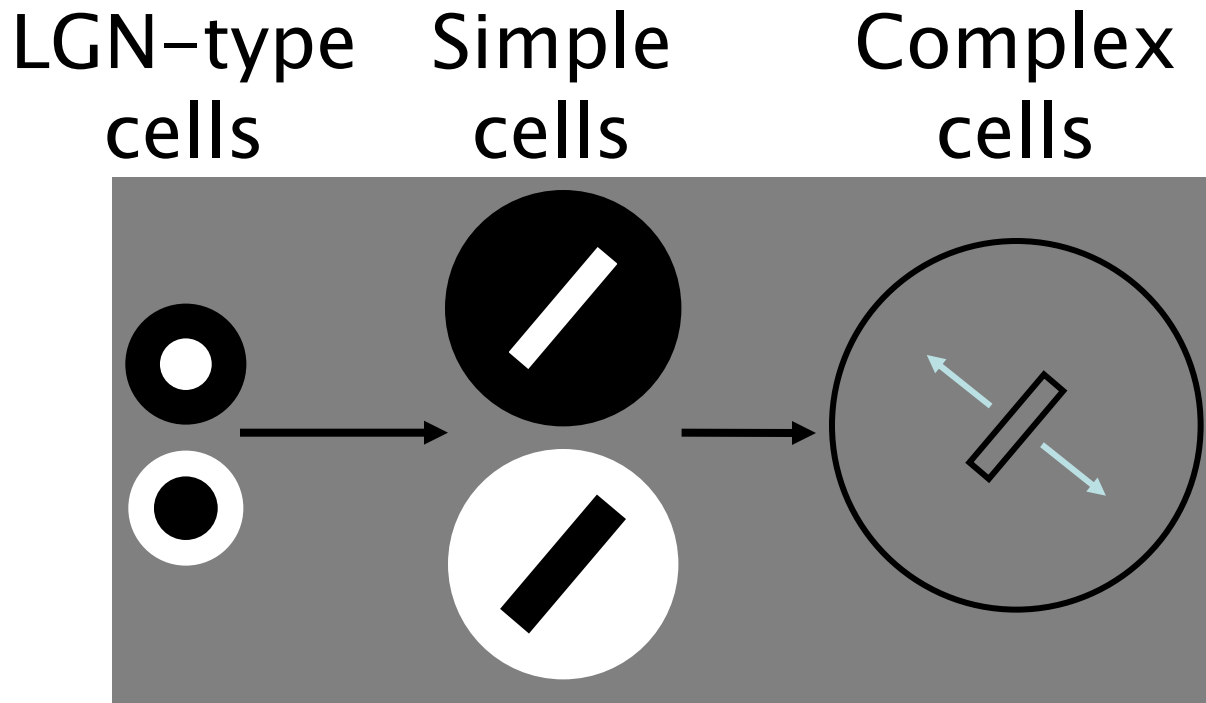
Neuron basics



Some numbers

- Human Brain
 - 10^{11} - 10^{12} neurons (1 million flies 😊)
 - 10^{14} - 10^{15} synapses
- Neuron
 - Fundamental space dimensions:
 - fine dendrites : 0.1 μ diameter; lipid bilayer membrane : 5 nm thick; specific proteins : pumps, channels, receptors, enzymes
 - Fundamental time length : 1 msec

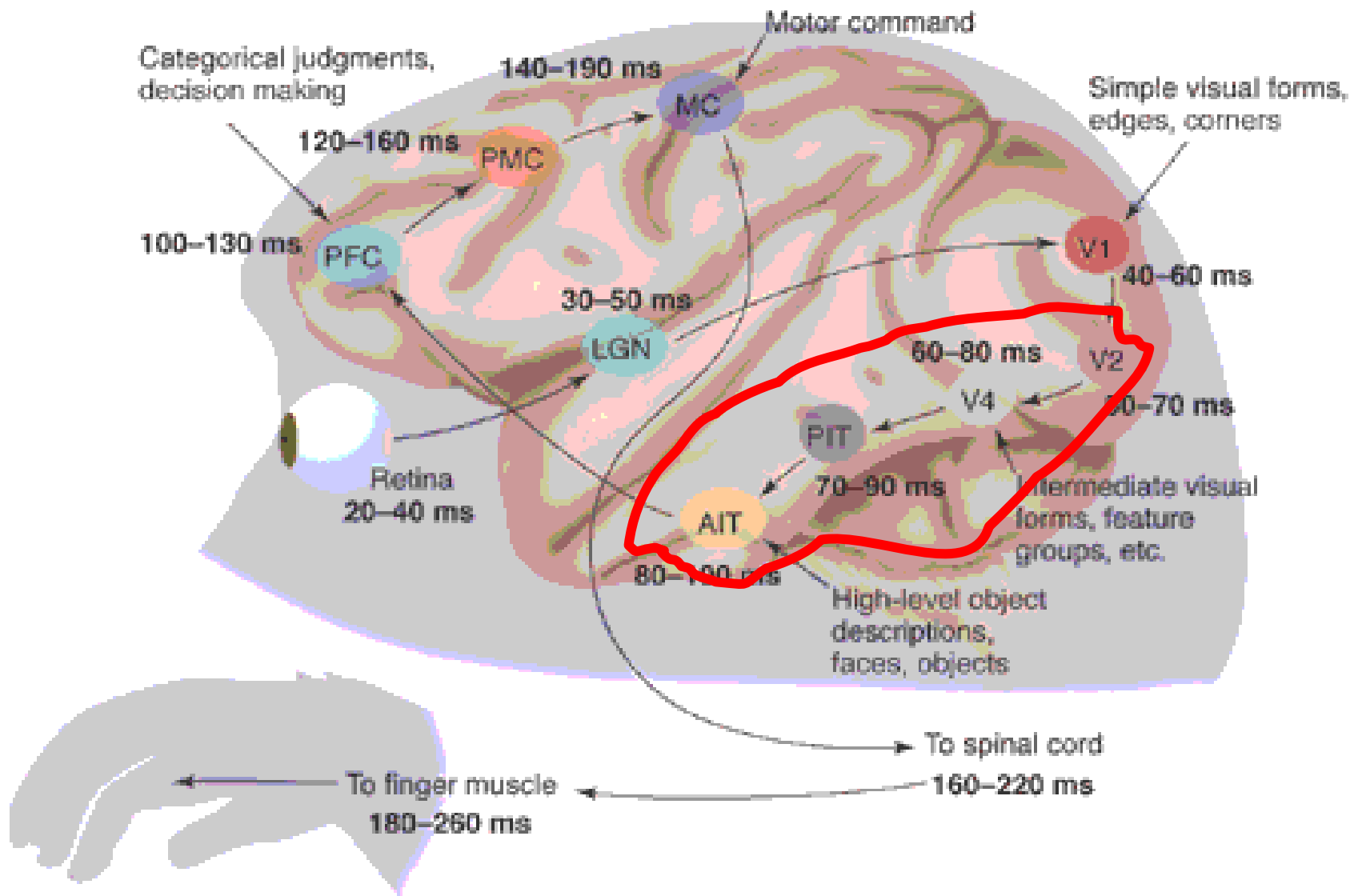
V1: hierarchy of simple and complex cells



(Hubel & Wiesel 1959)









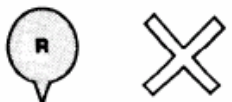


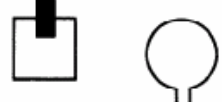
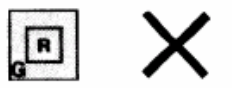



V1: Orientation selectivity

**Hubel & Wiesel
movie (later)**

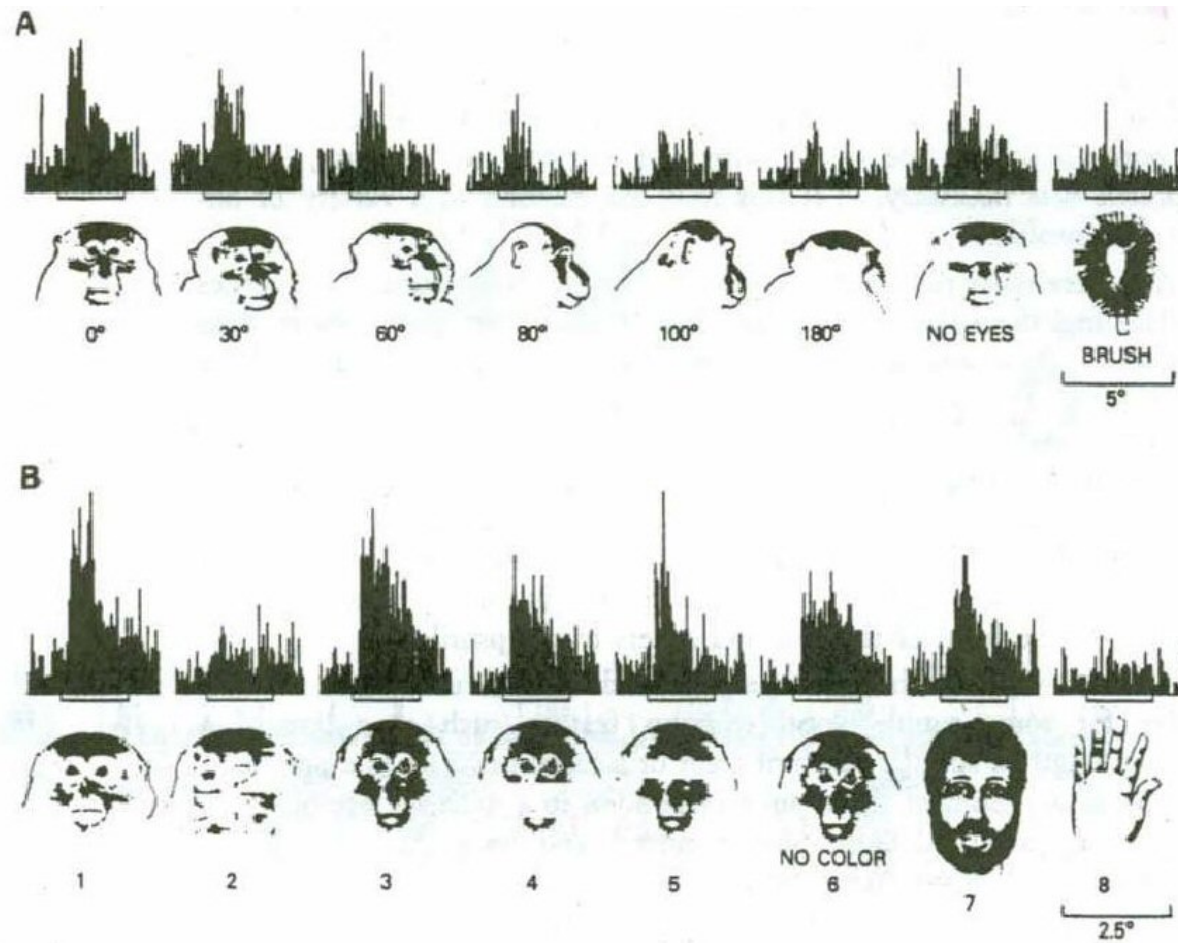


(Thorpe and Fabre-Thorpe, 2001)

Beyond V1: A gradual increase in the receptive field size and in the complexity of the preferred stimulus

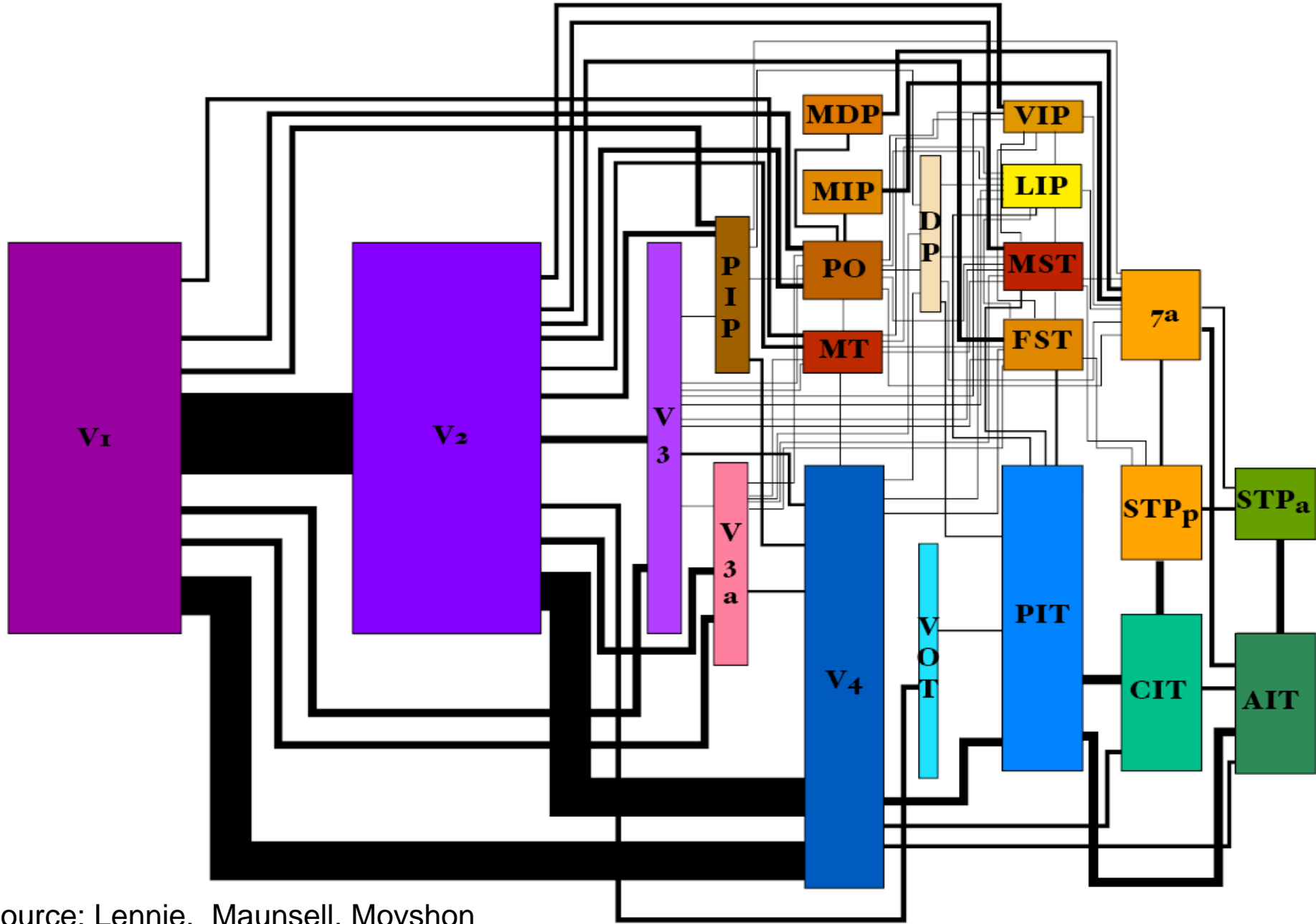
V2	V4	posterior IT	anterior IT
			
			
			
			

AIT: Face cells



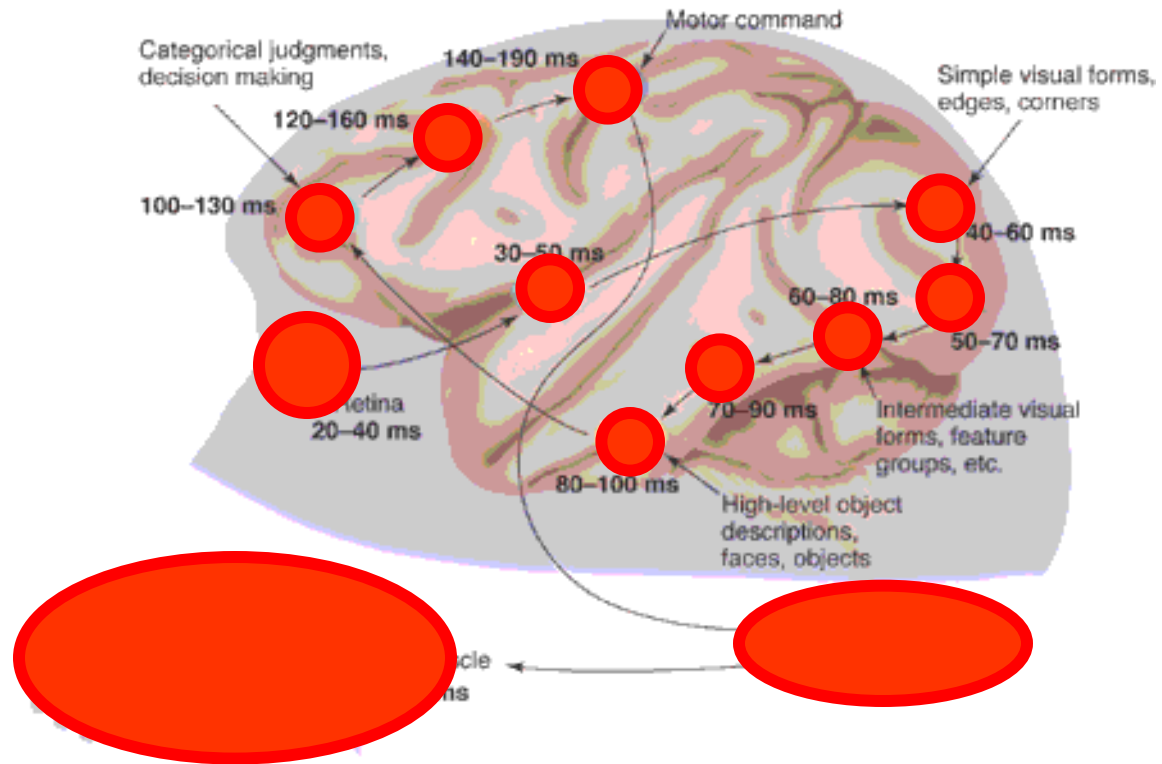
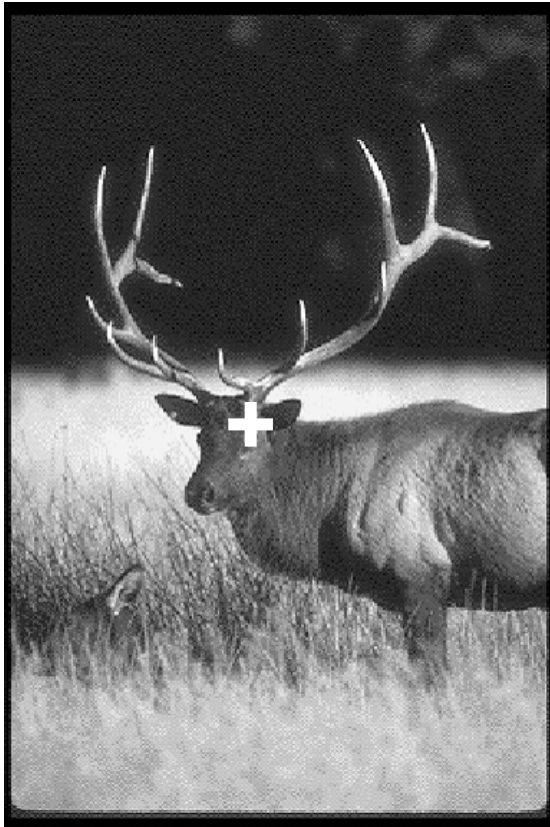
1. Problem of visual recognition, visual cortex
2. Historical background
3. Neurons and areas in the visual system
4. Data and feedforward hierarchical models
5. What is next?

The ventral stream



Source: Lennie, Maunsell, Movshon

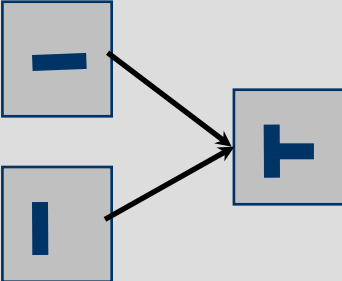
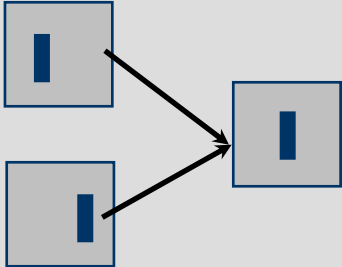
We consider feedforward architecture only



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

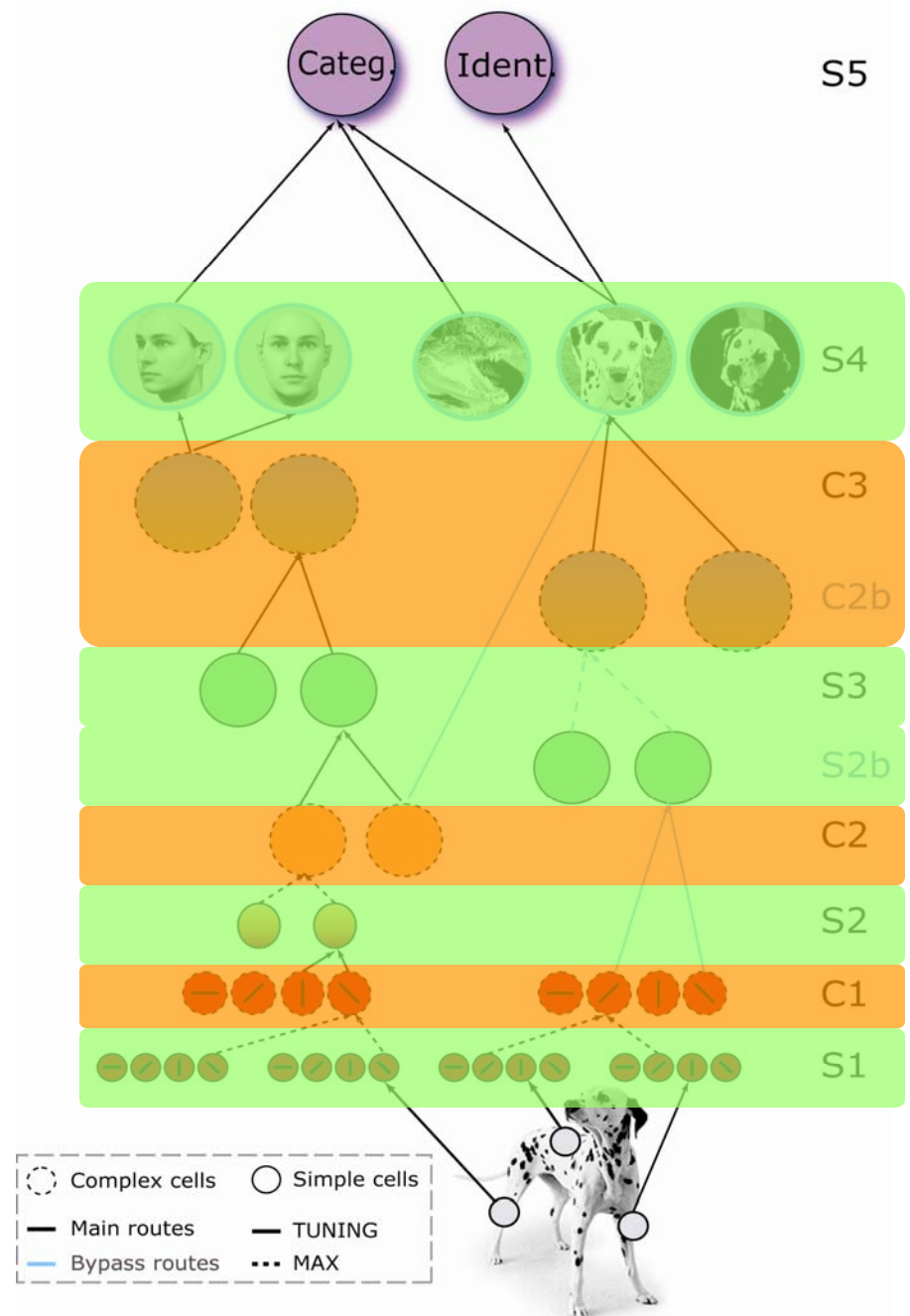
Unit types	Pooling	Computation	Operation
Simple		Selectivity / template matching	Gaussian- tuning / AND-like
Complex		Invariance	Soft-max / or-like

➤ Gaussian-like tuning operation (and-like)

➤ Simple units

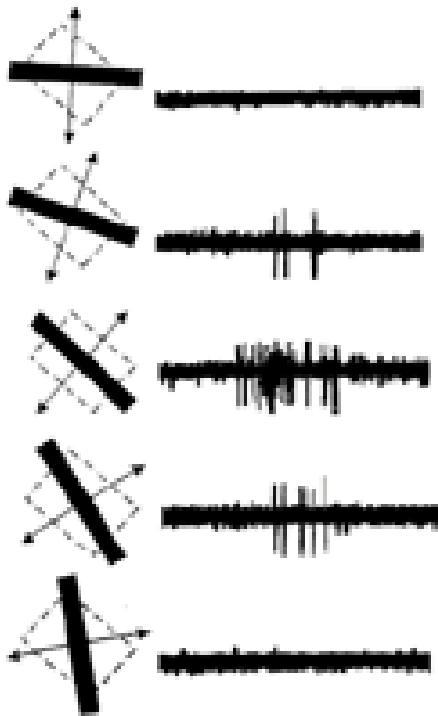
➤ Max-like operation (or-like)

➤ Complex units



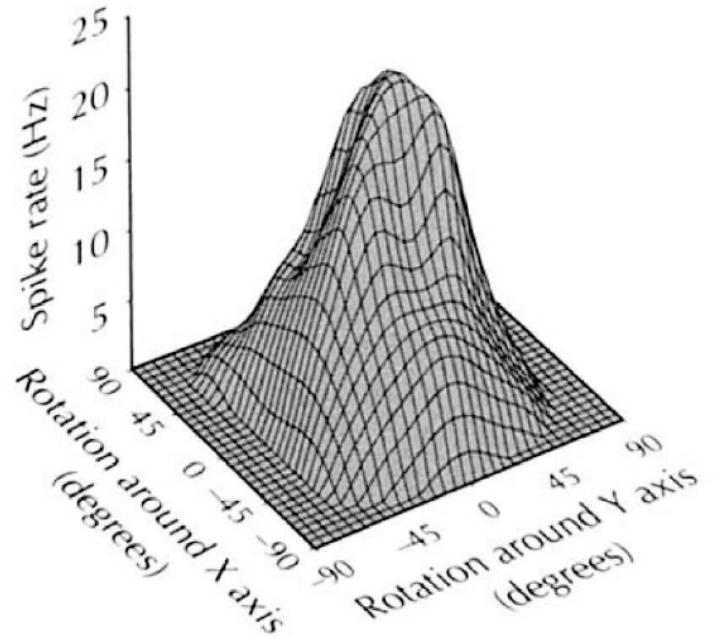
Gaussian tuning

Gaussian tuning in V1 for orientation



Hubel & Wiesel 1958

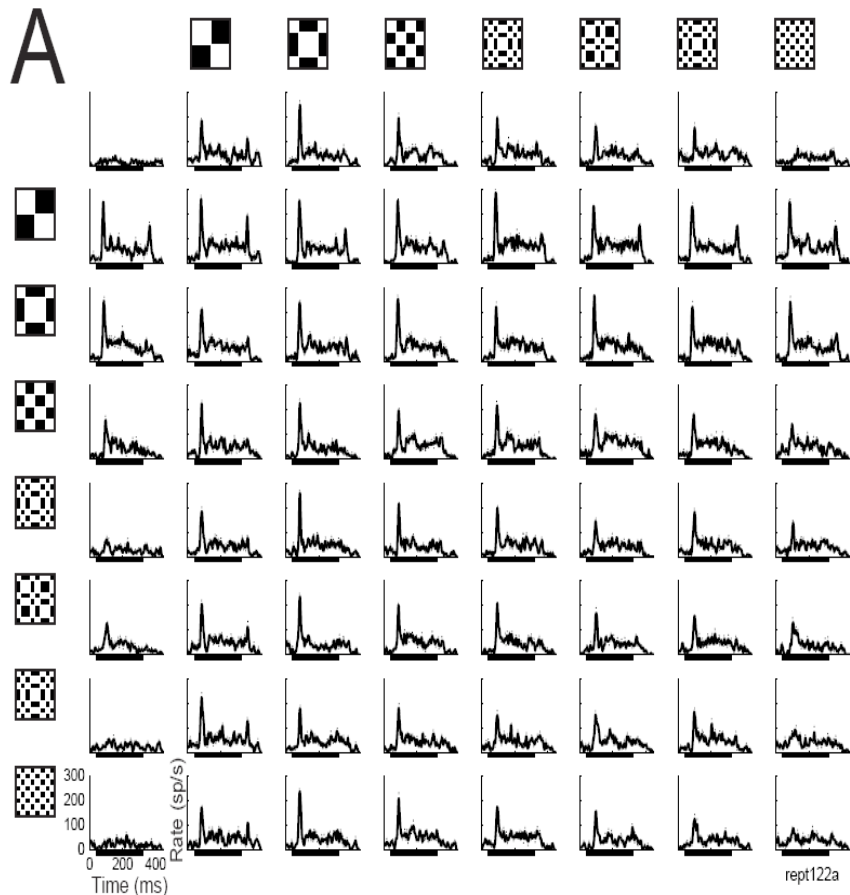
Gaussian tuning in IT around 3D views



Logothetis Pauls & Poggio 1995

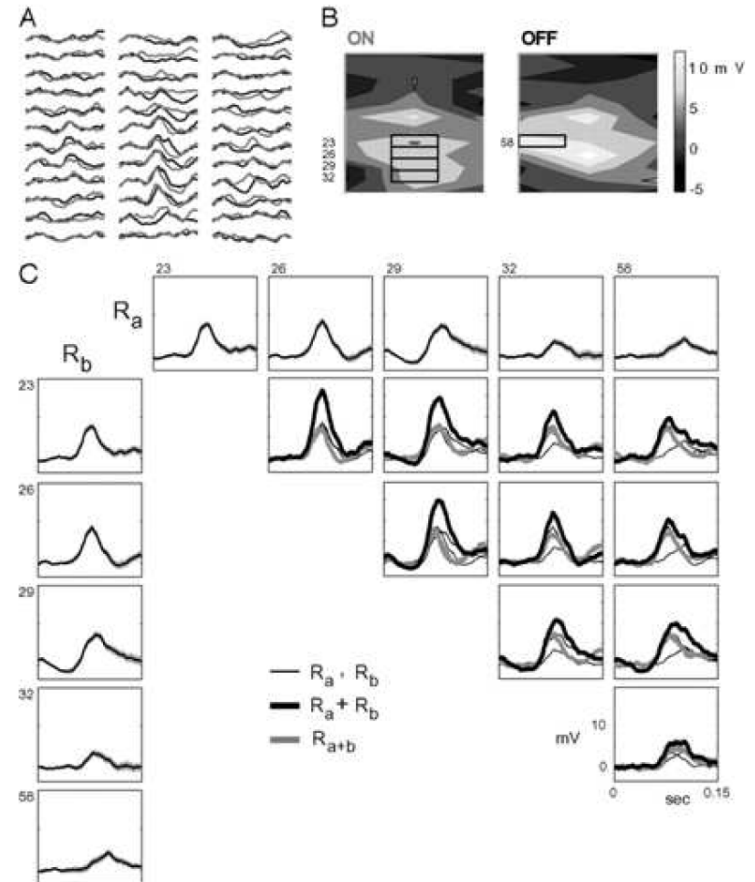
Max-like operation

Max-like behavior in V4



Gawne & Martin 2002

Max-like behavior in V1

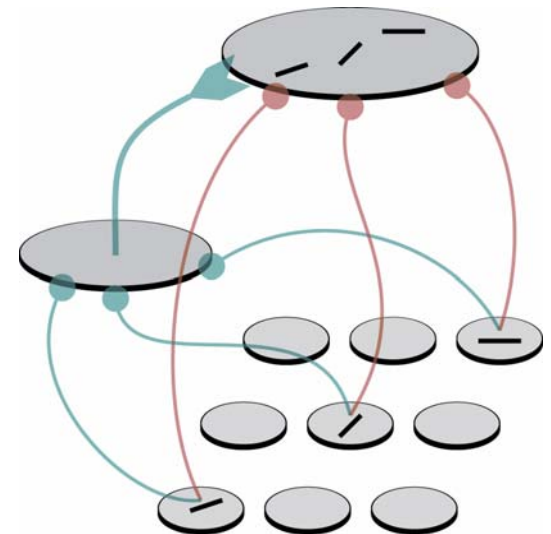
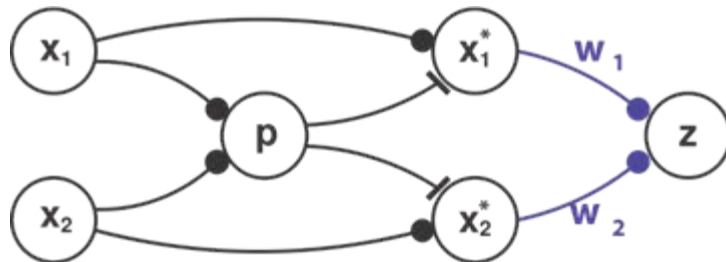


Lampl Ferster Poggio & Riesenhuber 2004
see also Finn Prieber & Ferster 2007

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.

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



Basic circuit is closely related to other models

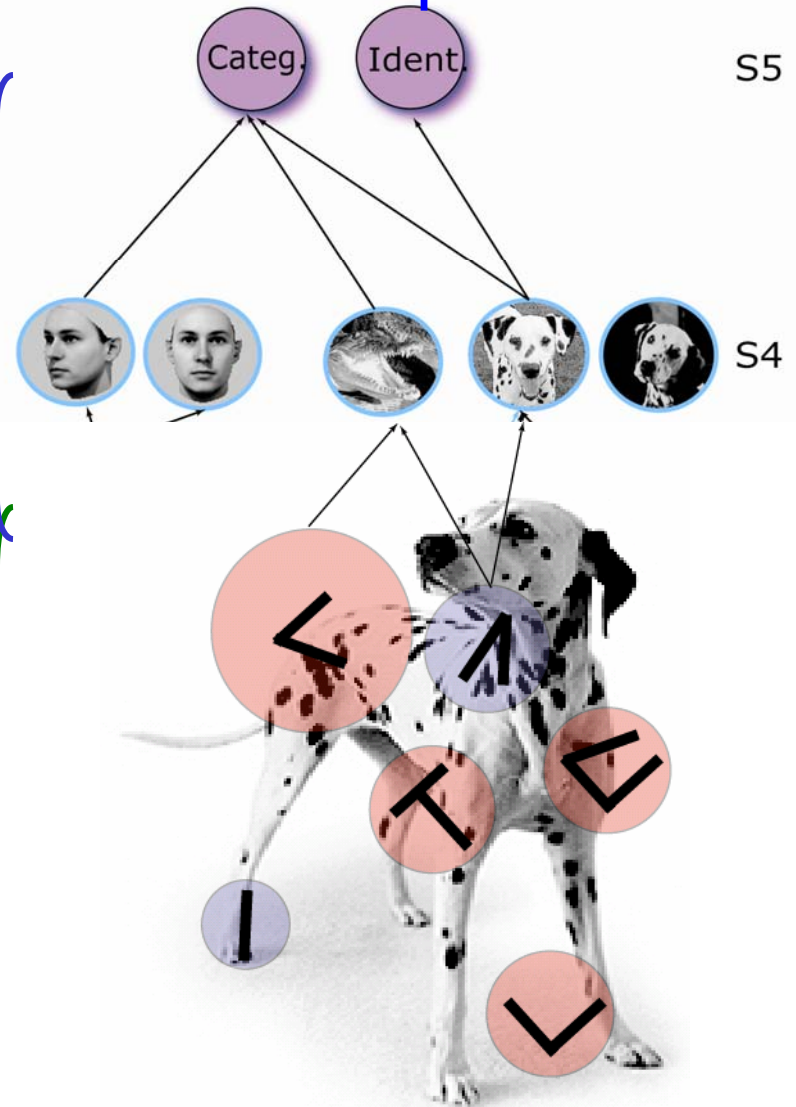
Operation	(Steady-State) Output	
Canonical	$y = \frac{\sum_{i=1}^n w_i x_i^p}{k + \left(\sum_{i=1}^n x_i^q \right)^r} \quad (1)$	Can be implemented by shunting inhibition (Grossberg 1973, Reichardt et al. 1983, Carandini and Heeger, 1994) and spike threshold variability (Anderson et al. 2000, Miller and Troyer, 2002)
Energy Model	$y = \sum_{i=1}^2 x_i^2 \quad (2)$	Adelson and Bergen (see also Hassenstein and Reichardt, 1956)
Gaussian-like	$y = \frac{\sum_{i=1}^n w_i x_i}{k + \sum_{i=1}^n x_i^2} \quad (4)$	Of the same form as model of MT (Rust et al., Nature Neuroscience, 2007)
Max-like	$y = \frac{\sum_{i=1}^n x_i^3}{k + \sum_{i=1}^n x_i^2} \quad (5)$	

Learning: supervised and unsupervised

Task-specific circuits (from IT to PFC)

- Supervised learning: ~ Gaussian RBF

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

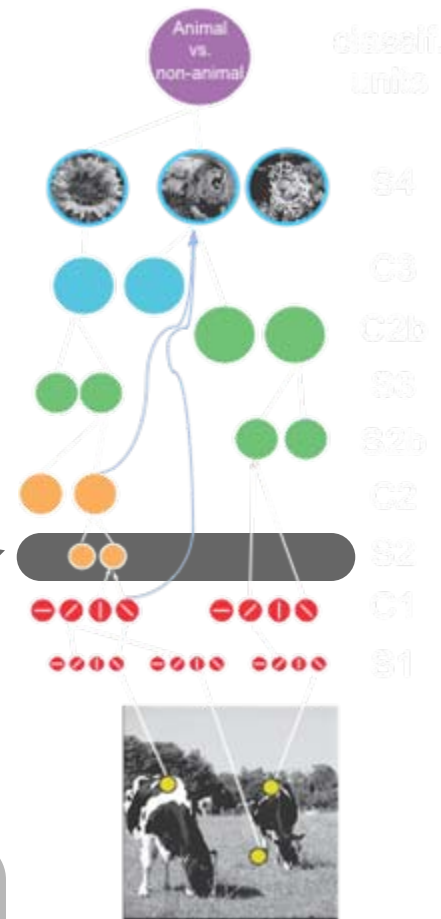
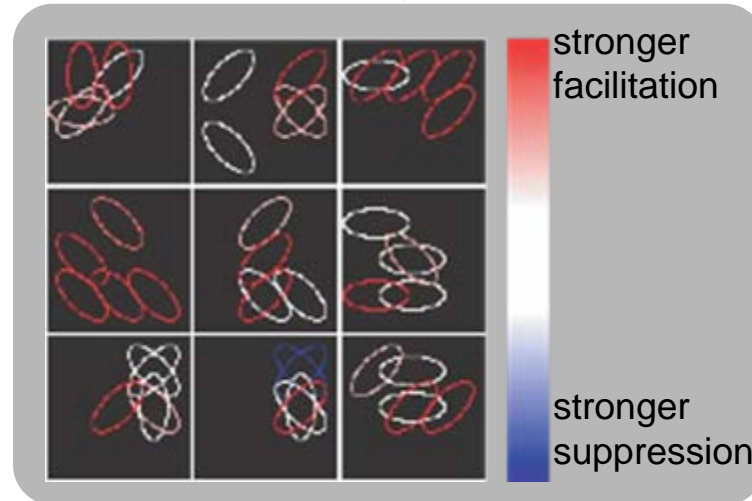


see also (Foldiak 1991; Perrett et al 1984; Wallis & Rolls, 1997; Lewicki and Olshausen, 1999; Einhauser et al 2002; Wiskott & Sejnowski 2002; Spratling 2005)

S2 units

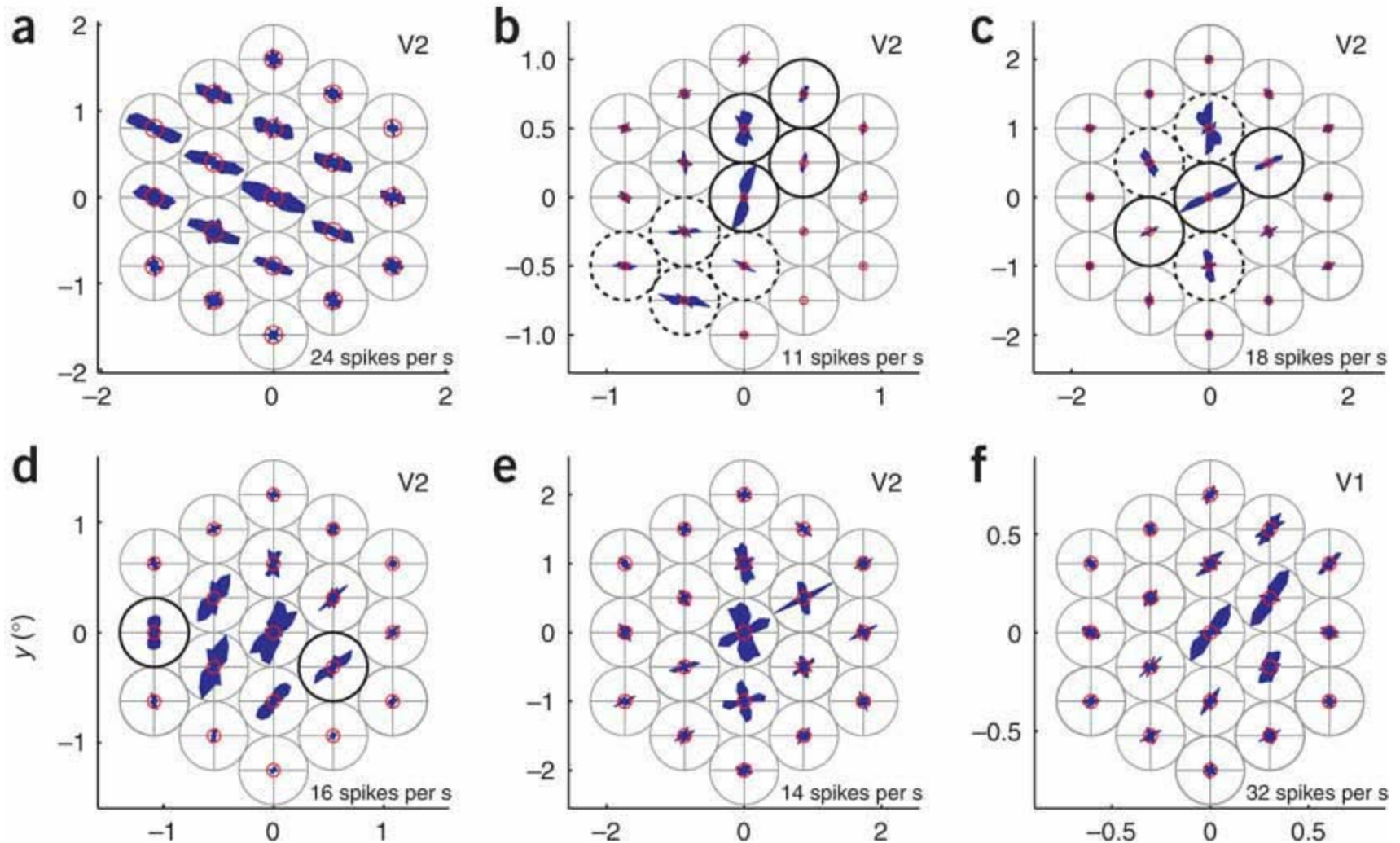
- Features of moderate complexity (n~1,000 types)
- Combination of V1-like complex units at different orientations

- Synaptic weights w learned from natural images
- 5-10 subunits chosen at random from all possible afferents (~100-1,000)



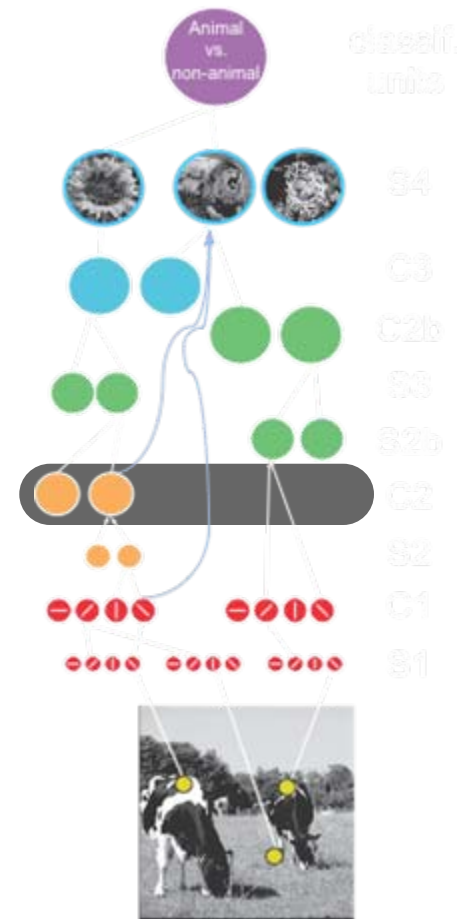
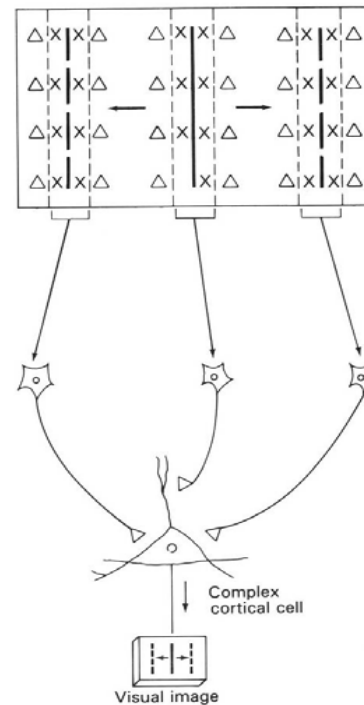
Neurons in monkey visual area V2 encode combinations of orientations

Akiyuki Anzai, Xinmiao Peng & David C Van Essen



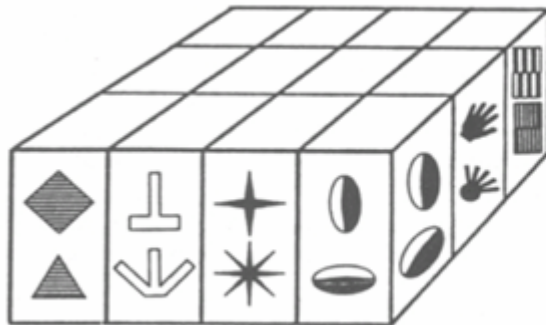
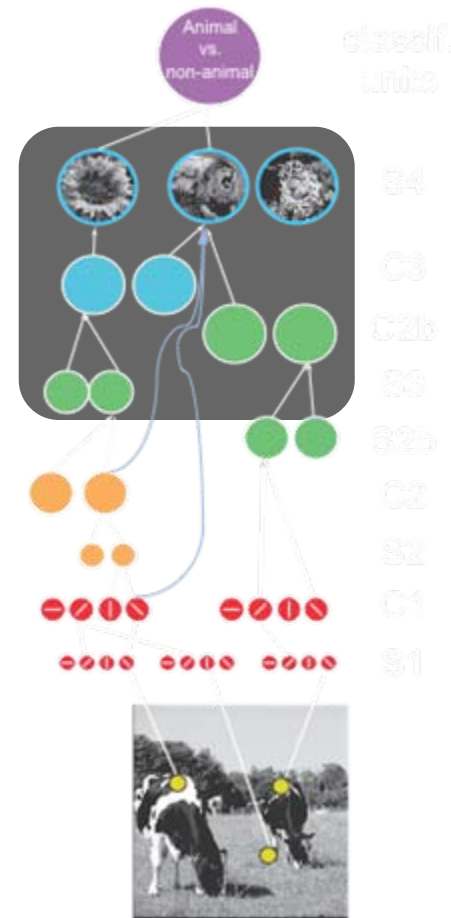
C2 units

- Same selectivity as S2 units but increased tolerance to position and size of preferred stimulus
- Local pooling over S2 units with same selectivity but slightly different positions and scales
- A prediction to be tested: **S2 units in V2** and **C2 units in V4**?

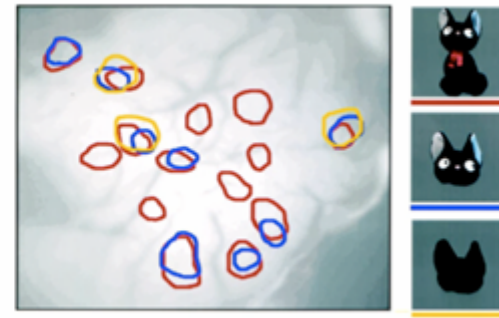


Beyond C2 units

- Units increasingly complex and invariant
- S3/C3 units:
 - Combination of V4-like units with different selectivities
 - Dictionary of ~1,000 features = num. columns in IT (Fujita 1992)



Tanaka et al.

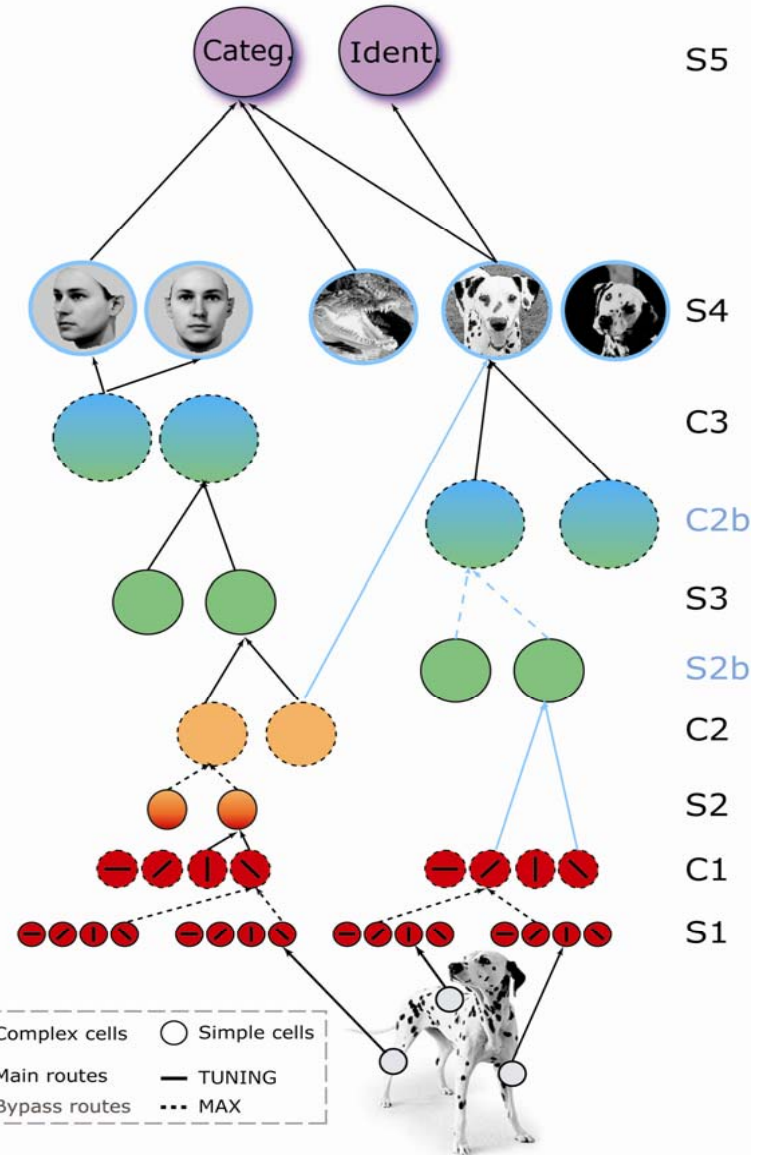
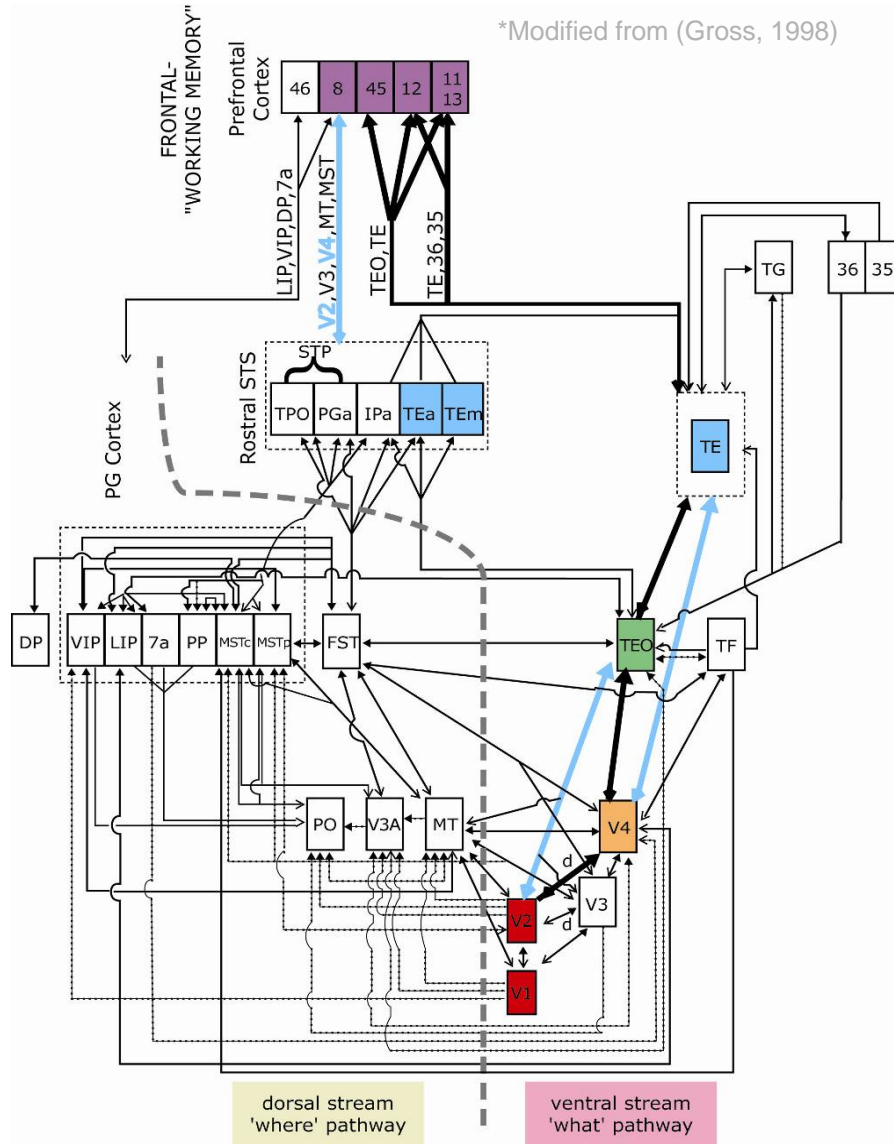


Tsunoda et al.

A loose hierarchy

- Bypass routes along with main routes:
 - From V2 to TEO (bypassing V4) (Morel & Bullier 1990; Baizer et al 1991; Distler et al 1991; Weller & Steele 1992; Nakamura et al 1993; Buffalo et al 2005)
 - From V4 to TE (bypassing TEO) (Desimone et al 1980; Saleem et al 1992)
- “Replication” of simpler selectivities from lower to higher areas
- Rich dictionary of features – across areas -- with various levels of selectivity and invariance

A hierarchical algorithm...



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

[software available online]

- Radial basis function (RBF) networks [w/ Gaussian kernel] can:
 - learn effectively from “small” training sets
 - generalize input-output mapping to new set of data (Poggio, 1990; Poggio and Bizzi, 2004)

- An advantage of the networks with the Gaussian-like tuning units [as opposed to a perceptron-like network with the sigmoidal neural units only] is the speed and ease of learning the parameters in the network (Moody and Darken, 1989; Poggio and Girosi, 1989)

Next: can we falsify feedforward models? comparison w| neural data

- V1:
 - Simple and complex cells tuning (Schiller et al 1976; Hubel & Wiesel 1965; Devalois et al 1982)
 - MAX operation in subset of complex cells (Lampl et al 2004)
- V4:
 - 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, Cadieu et al., 2007)
 - Tuning for Cartesian and non-Cartesian gratings (Gallant et al 1996)
- IT:
 - 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)
 - 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)

Next class and
Hubel & Wiesel
movie