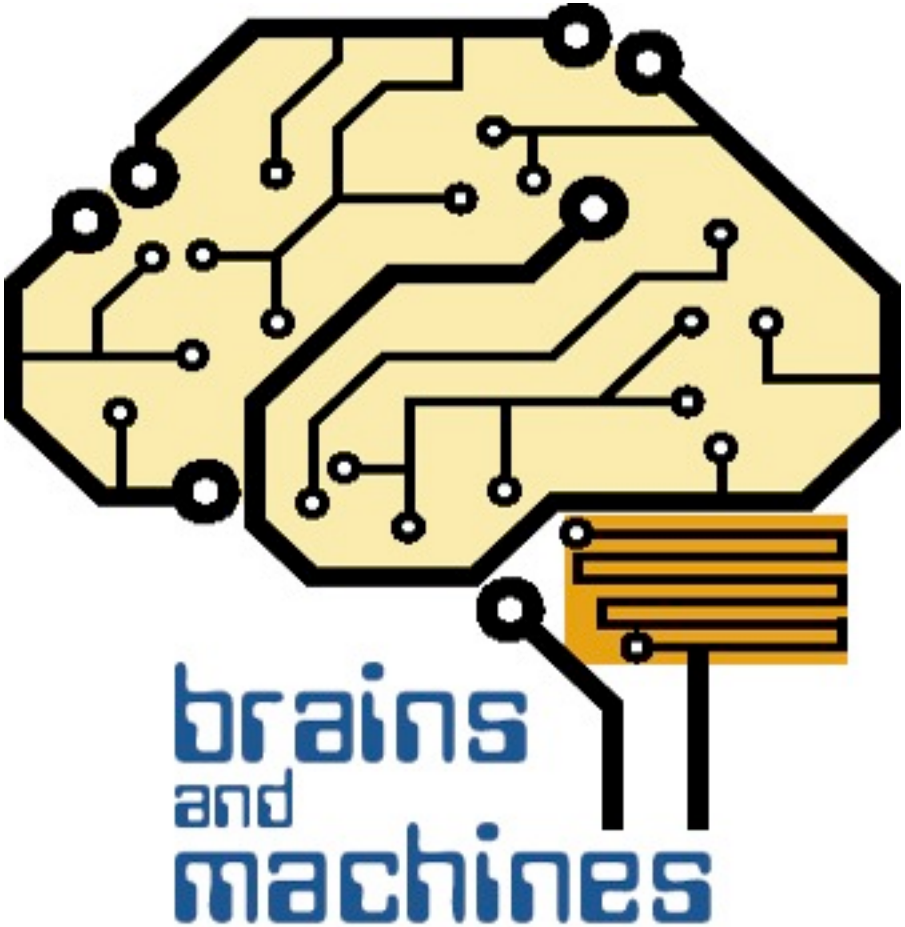
M-theory:

Learning representations for learning like humans learn

brains
and
machines

tomaso poggio
McGovern Institute
I2, CBCL, BCS,
LCSL, CSAIL
MIT



tomaso poggio
McGovern Institute
I2, CBCL, BCS,
LCSL, CSAIL
MIT

The Center for Brains, Minds and Machines

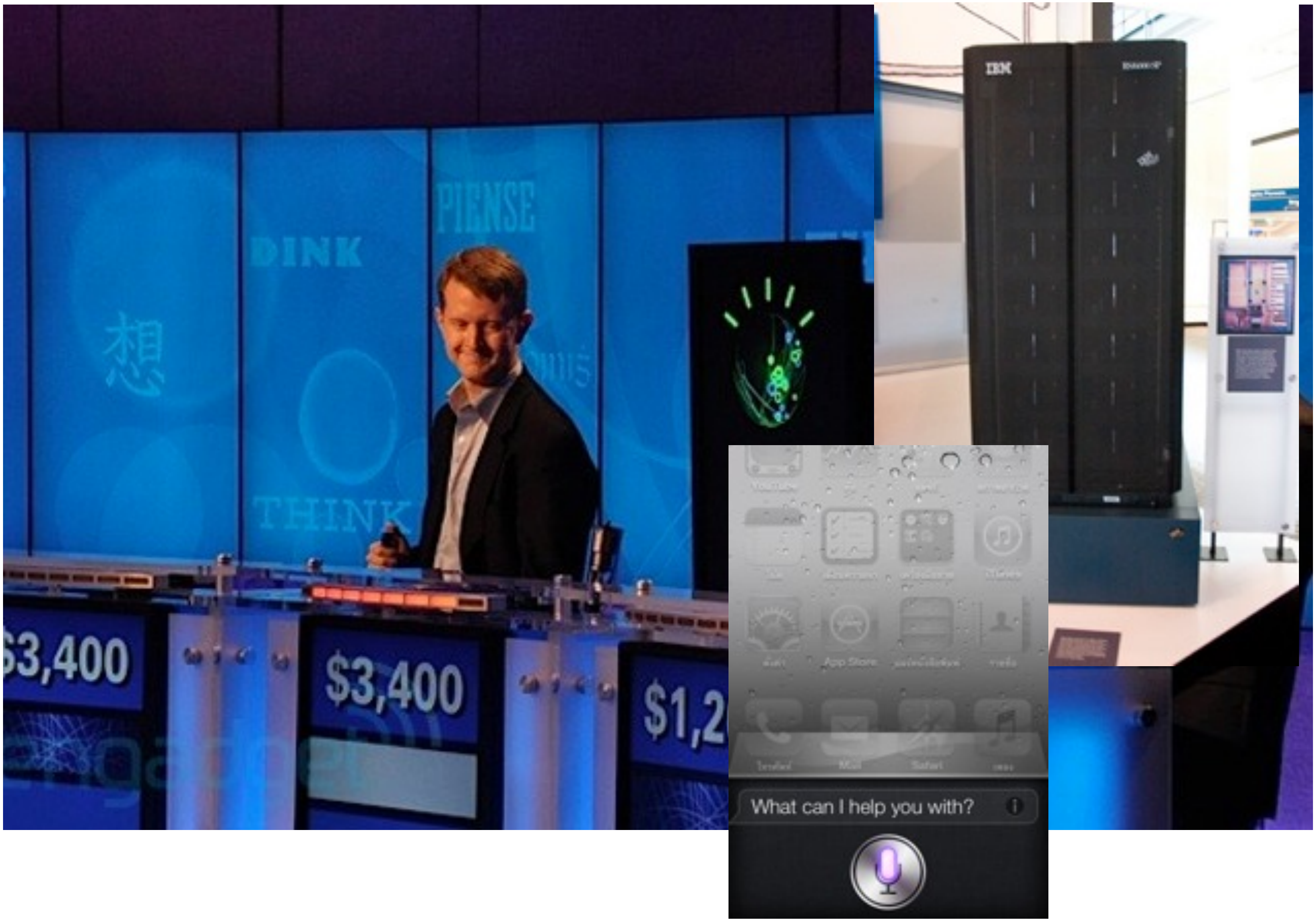
Vision for CBMM

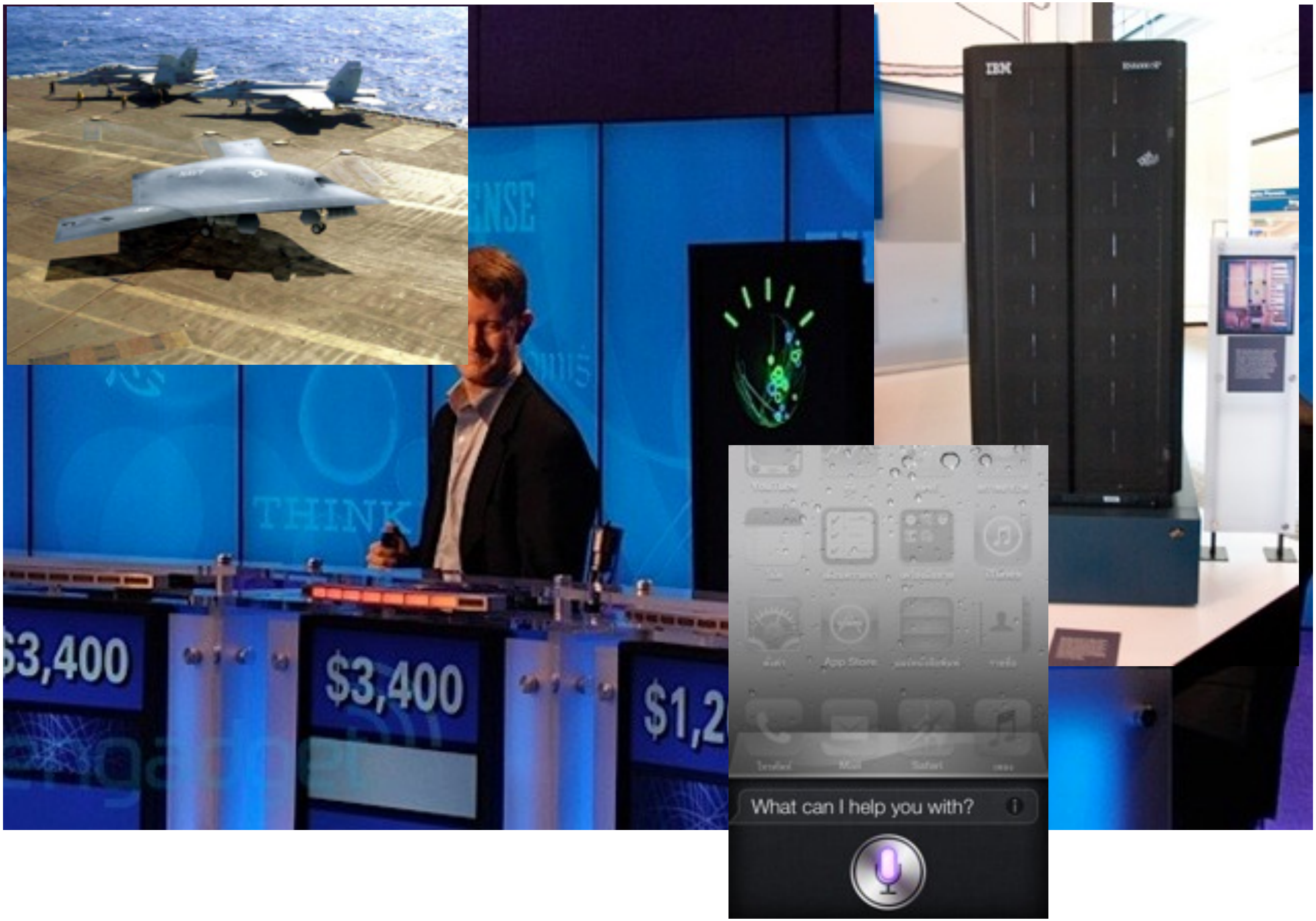
Vision for CBMM

- The problem of intelligence is one of the great problems in science.
- Work so far has led to many systems with impressive but narrow intelligence
- Now it is time to develop a deep computational understanding of human intelligence for its own sake and so that we can take intelligent applications to another level.

DINK
PIENSE
THINK
想
THINK
$3,400
$3,400
$1,200
engadget

PIENSE
DINK
想
THINK
$3,400
$3,400
$1,200
IBM


想
DINK
PIENSE
THINK
$3,400
$3,400
$1,2
IBM
App Store
What can I help you with?



What can I help you with?

Pedestrian accidents occur every day
in our increasingly intensive traffic environment.

In Europe, 14% of all traffic fatalities are pedestrians.

What can I help you with?

MIT

Boyden, Desimone ,Kaelbling , Kanwisher, Katz, Poggio, Sassanfar, Saxe, Schulz, Tenenbaum, Ullman, Wilson, Rosasco, Winston

Harvard

Blum, Kreiman, Mahadevan, Nakayama, Sompolinsky, Spelke, Valiant

Cornell

Hirsh

Allen Institute

Koch

Rockfeller

Freiwald

UCLA

Yuille

Stanford

Goodman

Hunter

Epstein,...

Wellesley

Hildreth, Conway...

Puerto Rico

Bykhovaskaia, Vega...

Howard

Manaye,...

City U. HK
Smale

Hebrew U.
Shashua

IIT
Metta, Rosasco, Sandini

MPI
Buelthoff

NCBS
Raghavan

Genoa U.
Verri

Weizmann
Ullman

City University of Hong Kong

iit
ISTITUTO ITALIANO DI TECNOLOGIA

ncbs

UNIVERSITÀ DEGLI STUDI DI GENOVA

WEIZMANN INSTITUTE OF SCIENCE

Google
Norvig

IBM
Ferrucci

Microsoft UK
Blake

Orcam
Shashua

MobilEye
Shashua

DeepMind
Hassabis

Boston Dynamics
Raibert

Rethink Robotics
Brooks

Willow Garage
Cousins

Google IBM Microsoft MOBILEYE BostonDynamics rethink robotics Willow Garage

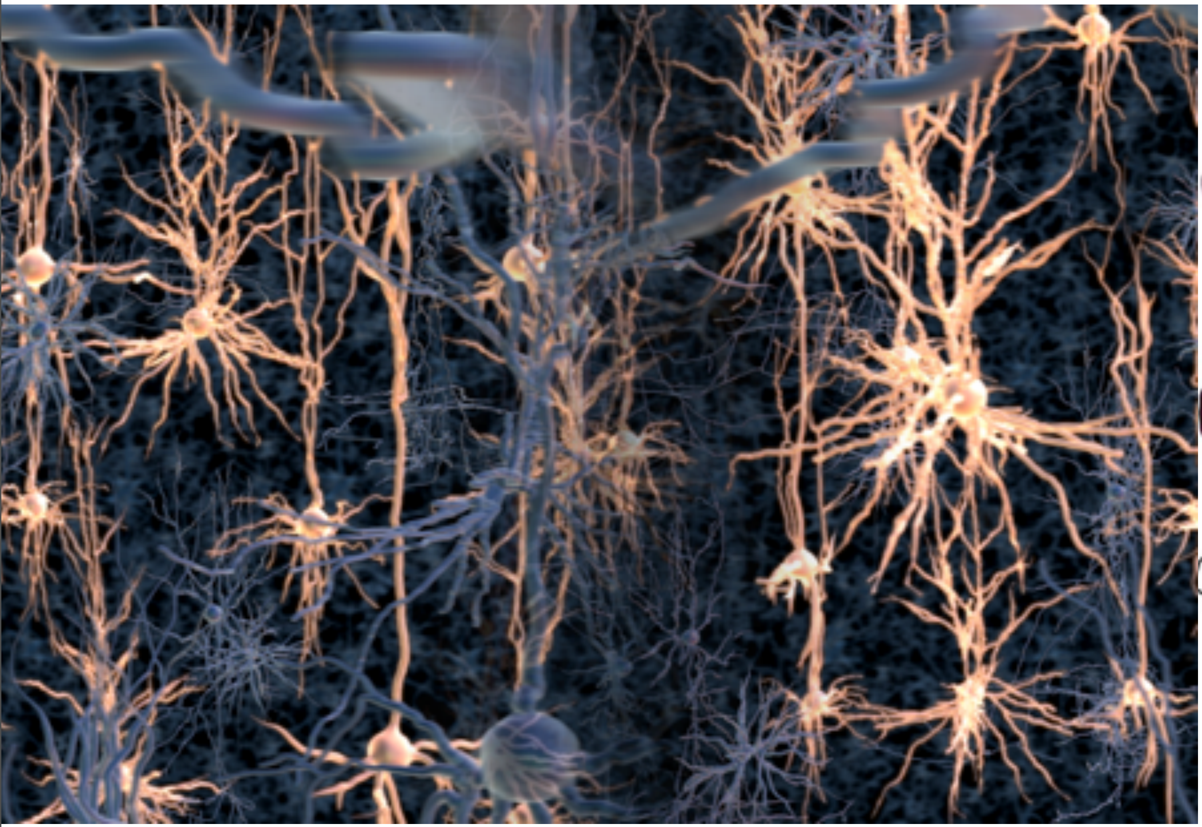
Example : a second phase in machine learning

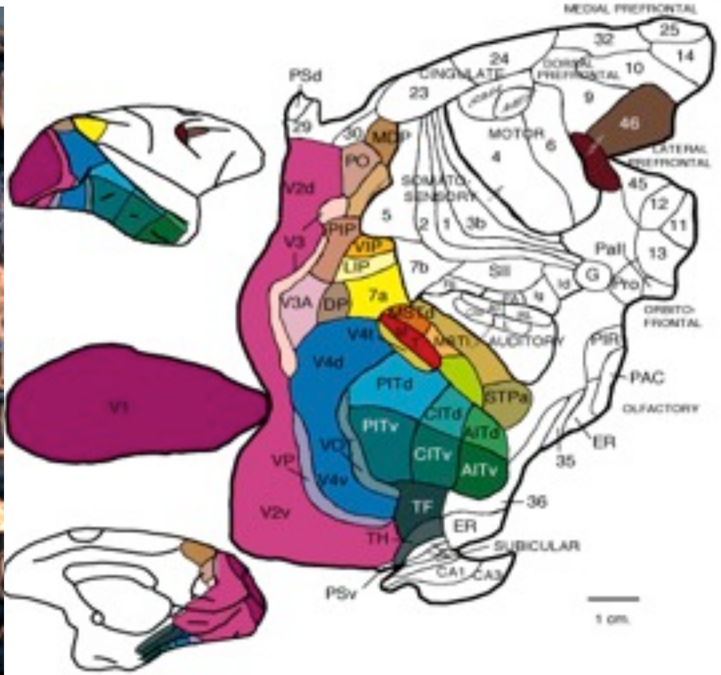
The first phase (and successes) of ML: supervised learning “n--> ∞”

Siri
Use your voice to send messages, set reminders, search for information, and more.
Autonomous Driving

The next phase of ML: unsupervised learning of invariant representations for learning “n--> 0”

Vision in the Brain





Van Essen & Anderson, 1990

- Human Brain
  - **10^{10}-10^{11} neurons (~1 million flies)**
  - **10^{14}- 10^{15} synapses**
  - ~ 30% cortex is vision (more than for
  - language and any other modality)

Recognition in Visual Cortex: hierarchical model

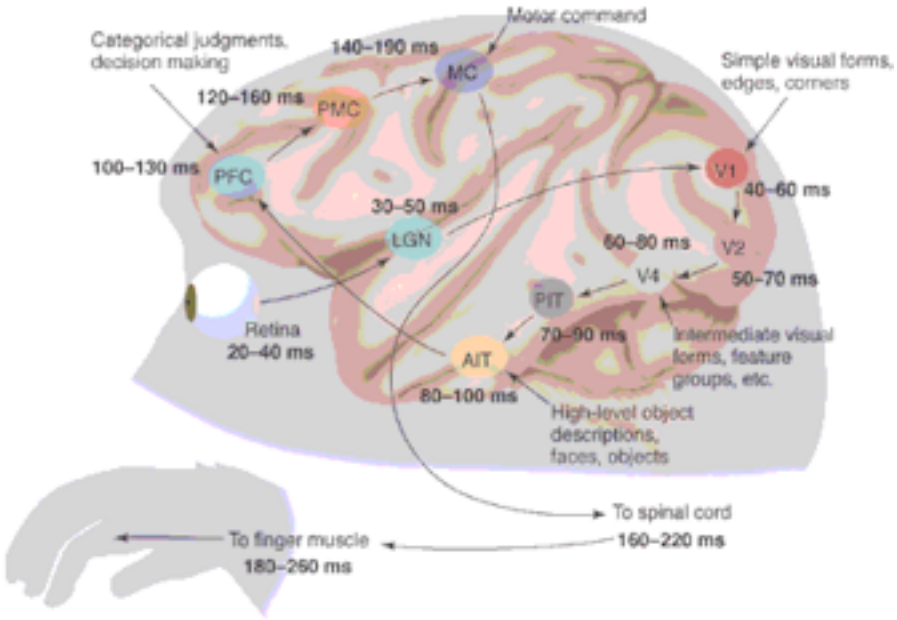
*Modified from (Gross, 1998)

[software available online
with CNS (for GPUs)]

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

Recognition in Visual Cortex: hierarchical model

*Modified from (Gross, 1998)



[software available online with CNS (for GPUs)]

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

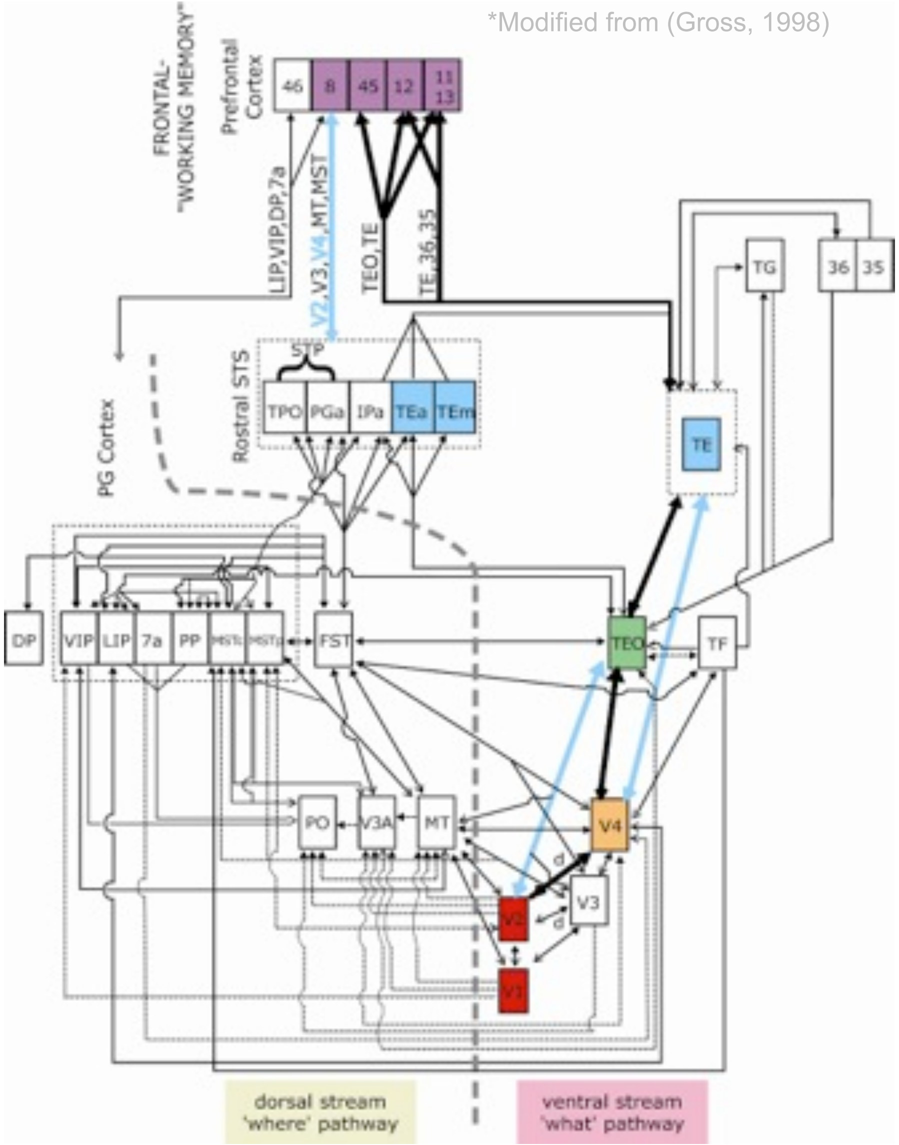
Recognition in Visual Cortex: hierarchical model

*Modified from (Gross, 1998)

[software available online
with CNS (for GPUs)]

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

Recognition in Visual Cortex: hierarchical model

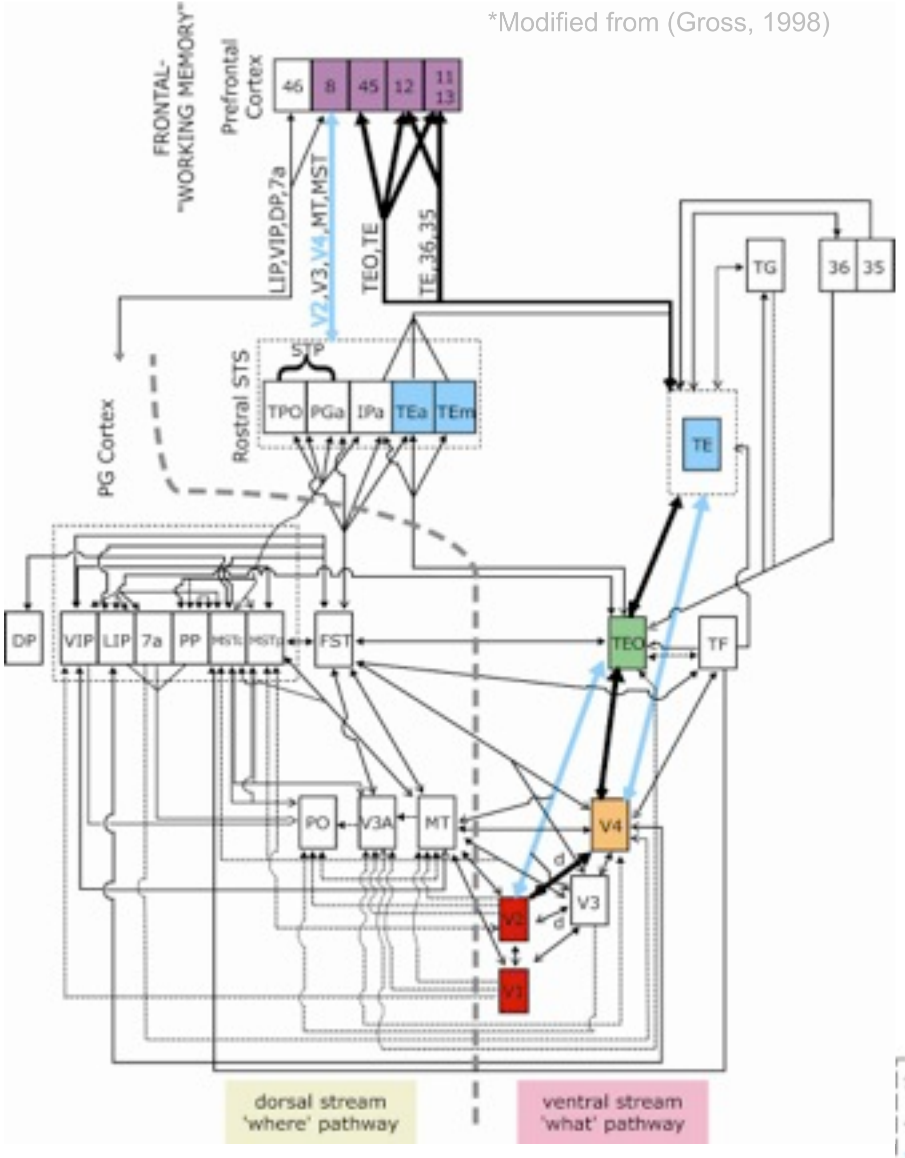


*Modified from (Gross, 1998)

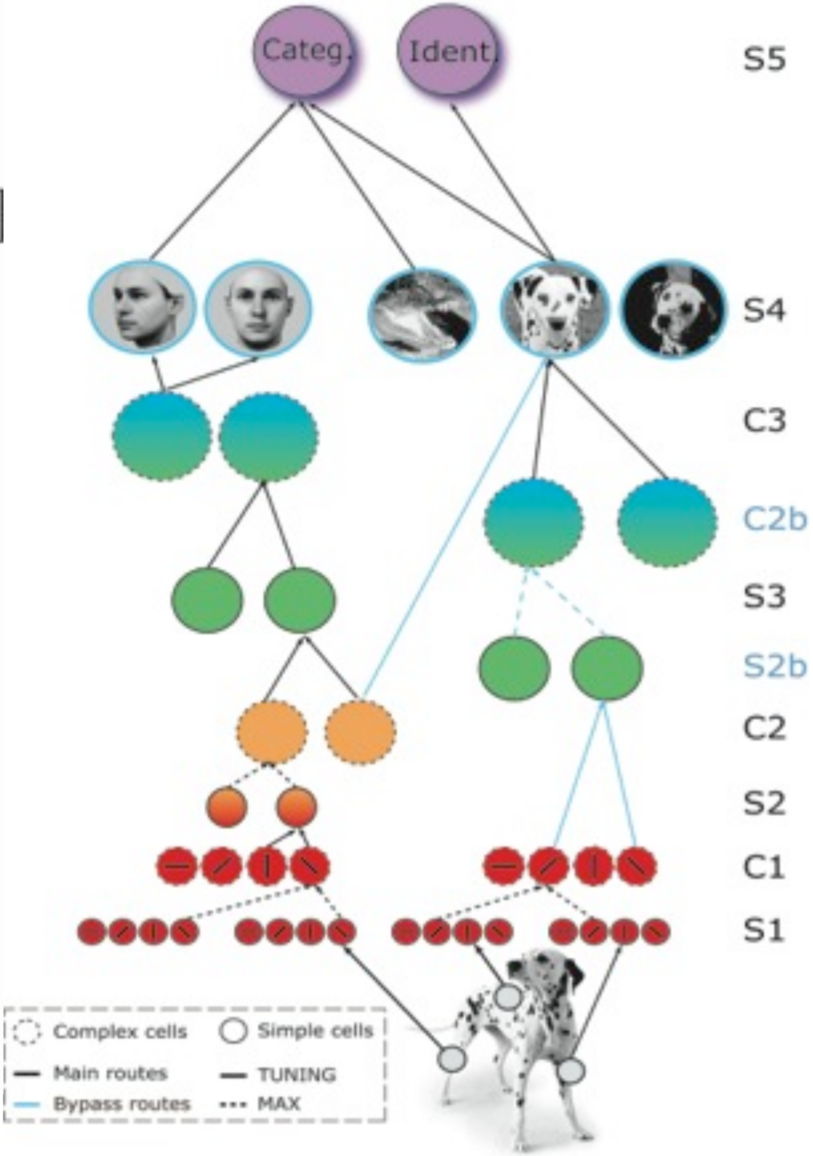
[software available online
with CNS (for GPUs)]

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

Recognition in Visual Cortex: hierarchical model



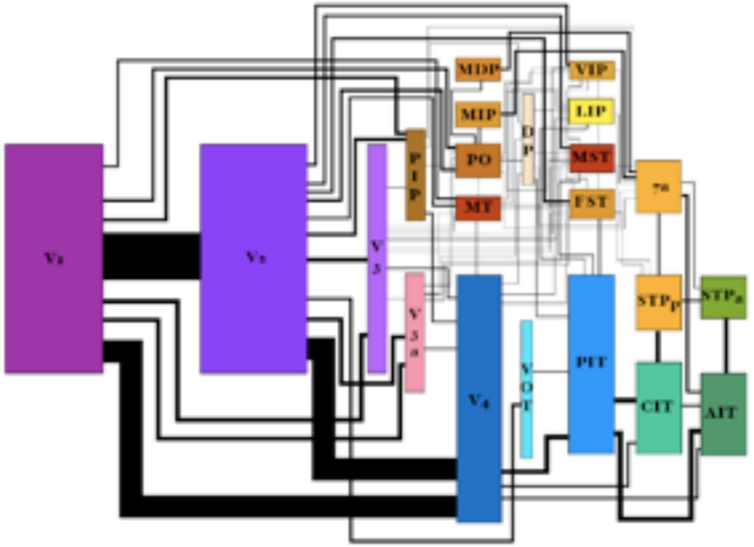
*Modified from (Gross, 1998)

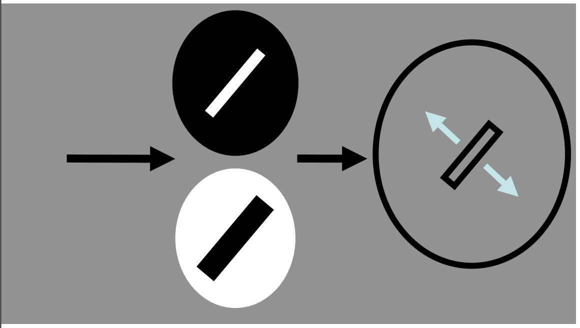


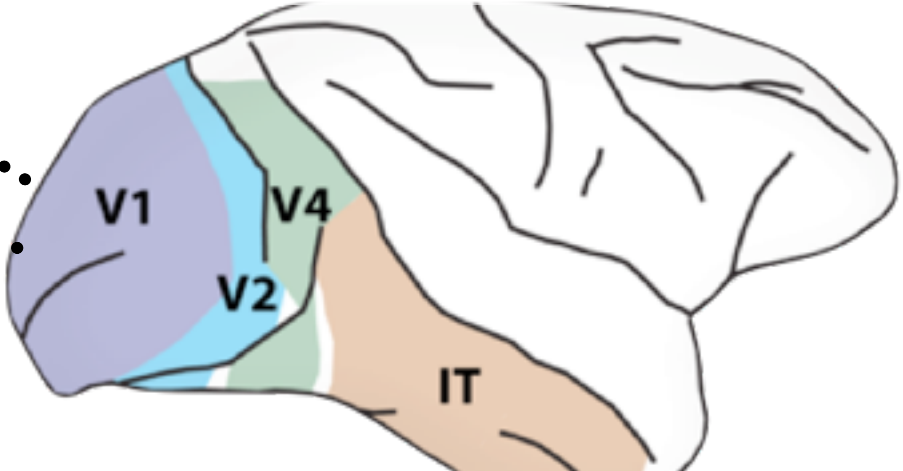
[software available online with CNS (for GPUs)]

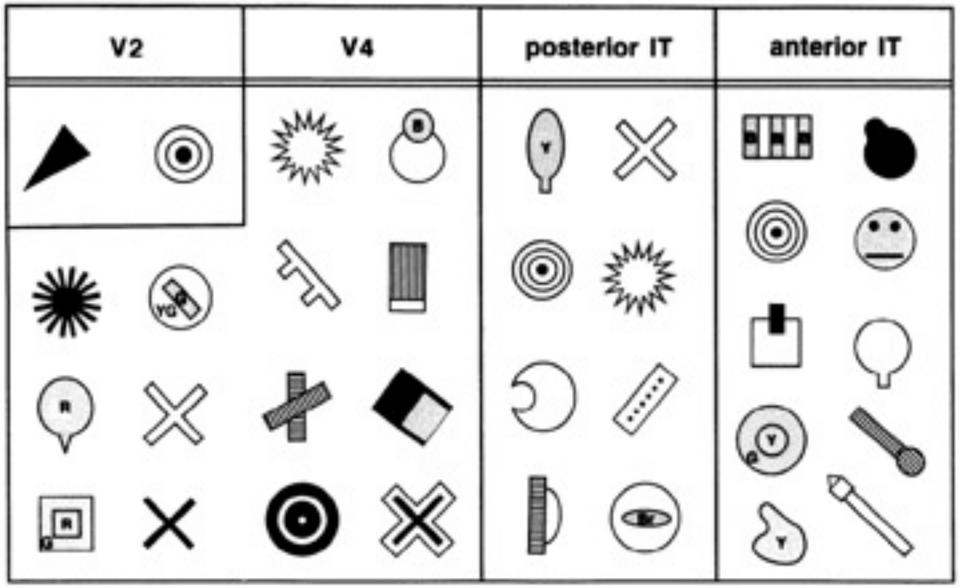
Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Visual Object Recognition: the ventral stream (macaque)









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

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

Kobatake & Tanaka, 1994

Recognition in Visual Cortex: hierarchical model

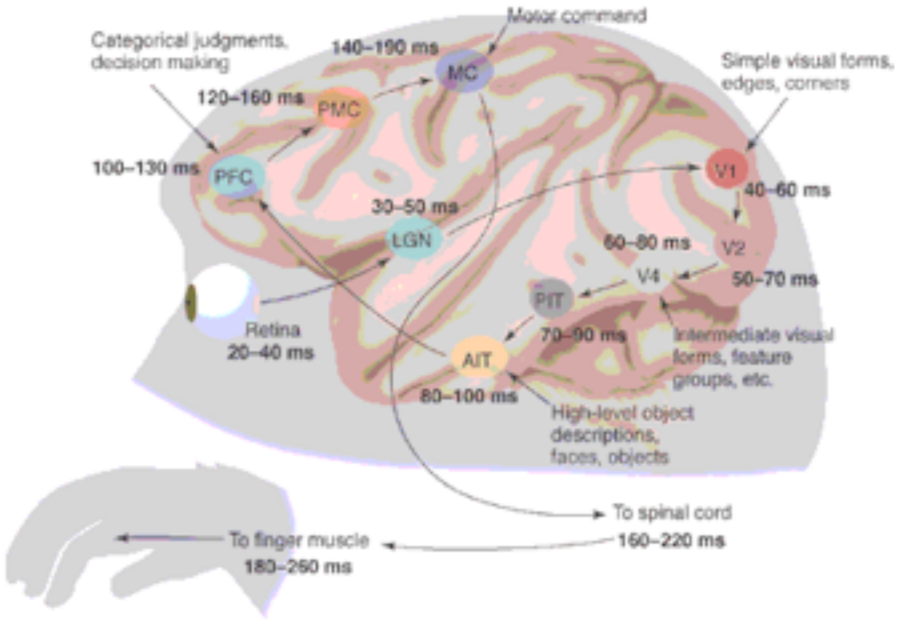
*Modified from (Gross, 1998)

[software available online
with CNS (for GPUs)]

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

Recognition in Visual Cortex: hierarchical model

*Modified from (Gross, 1998)



[software available online
with CNS (for GPUs)]

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

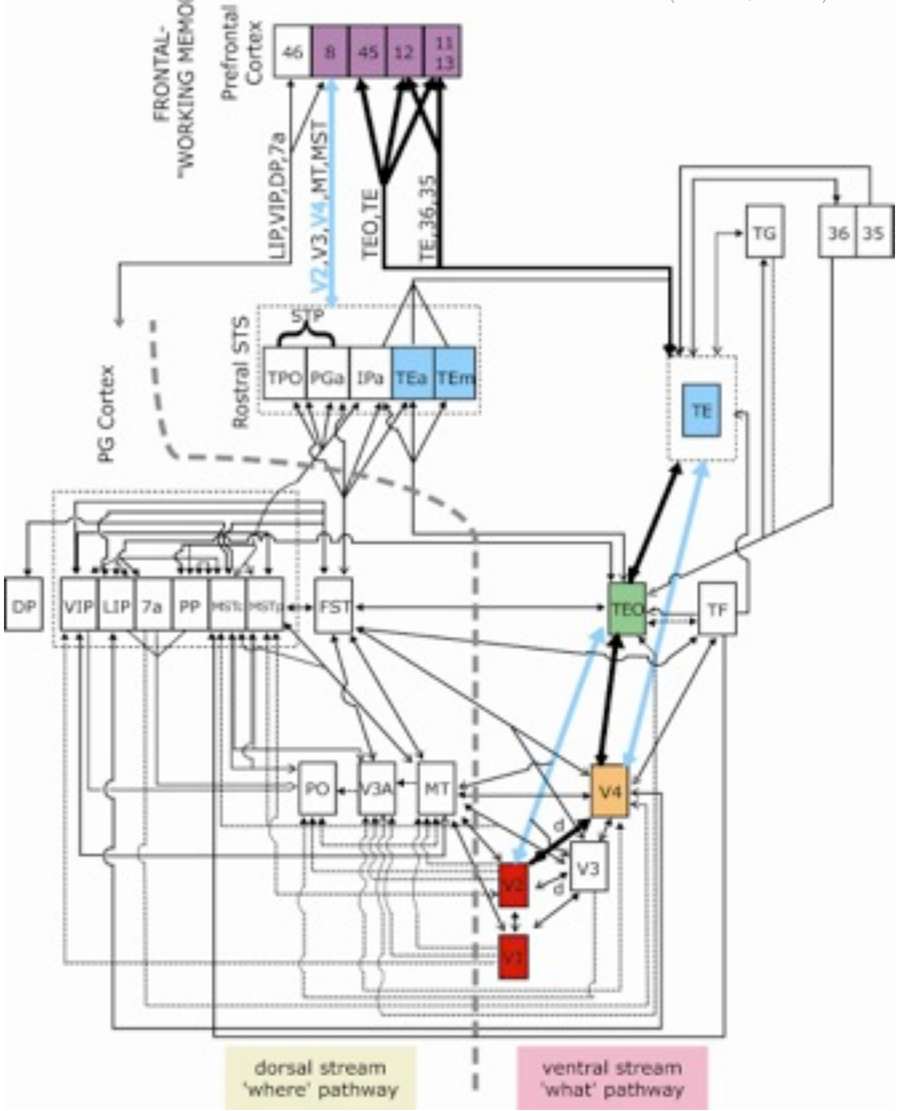
Recognition in Visual Cortex: hierarchical model

*Modified from (Gross, 1998)

[software available online
with CNS (for GPUs)]

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

Recognition in Visual Cortex: hierarchical model

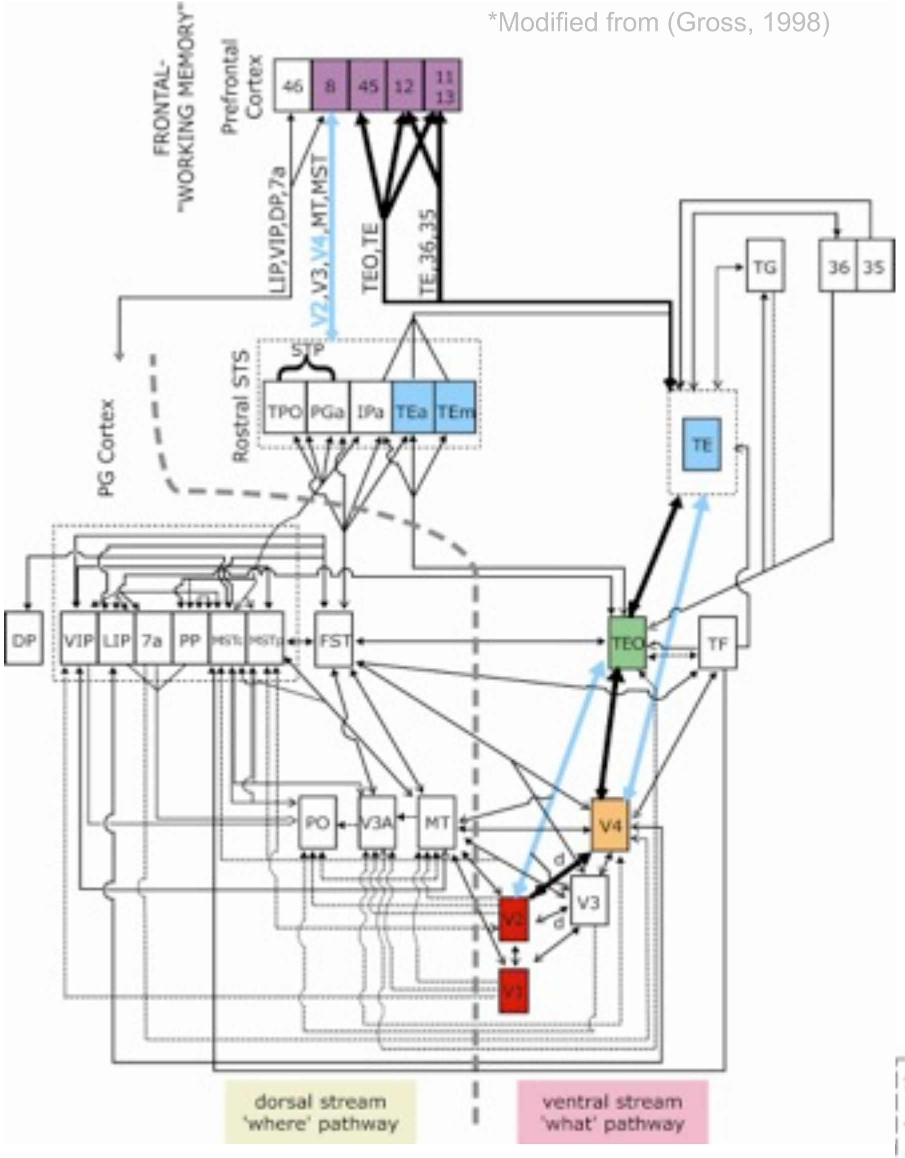


*Modified from (Gross, 1998)

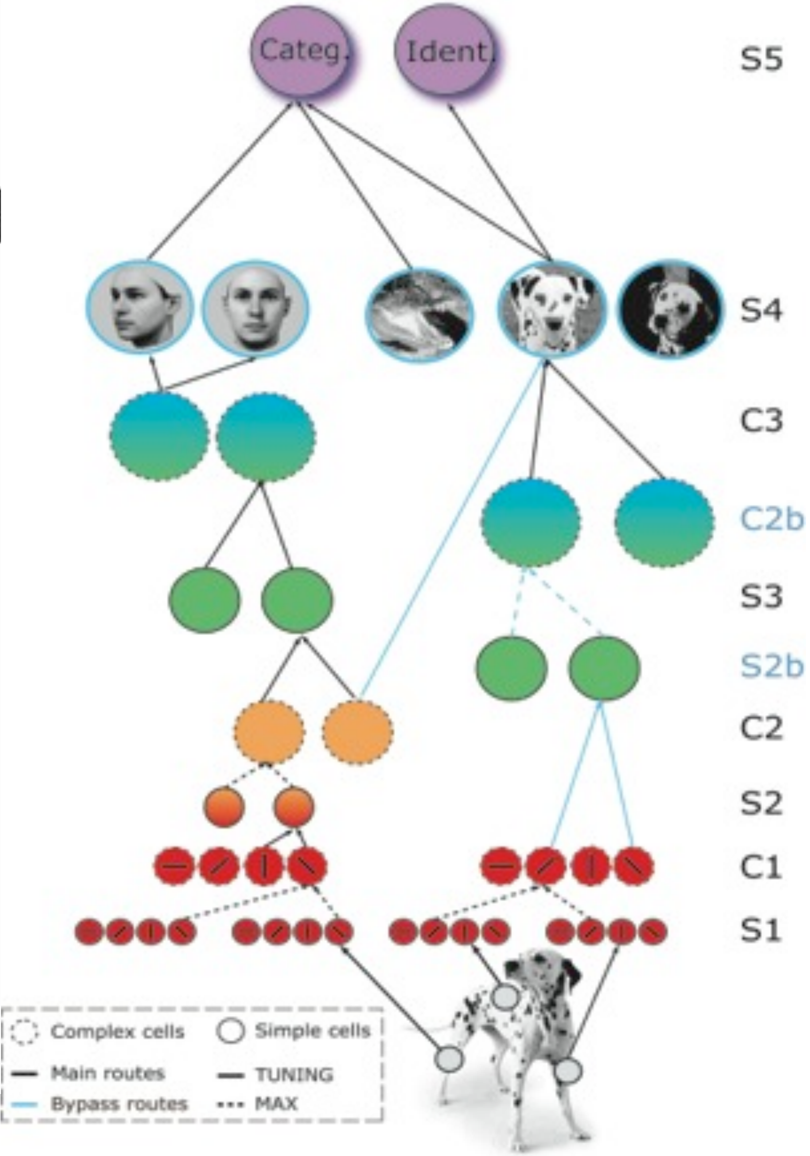
[software available online
with CNS (for GPUs)]

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

Recognition in Visual Cortex: hierarchical model



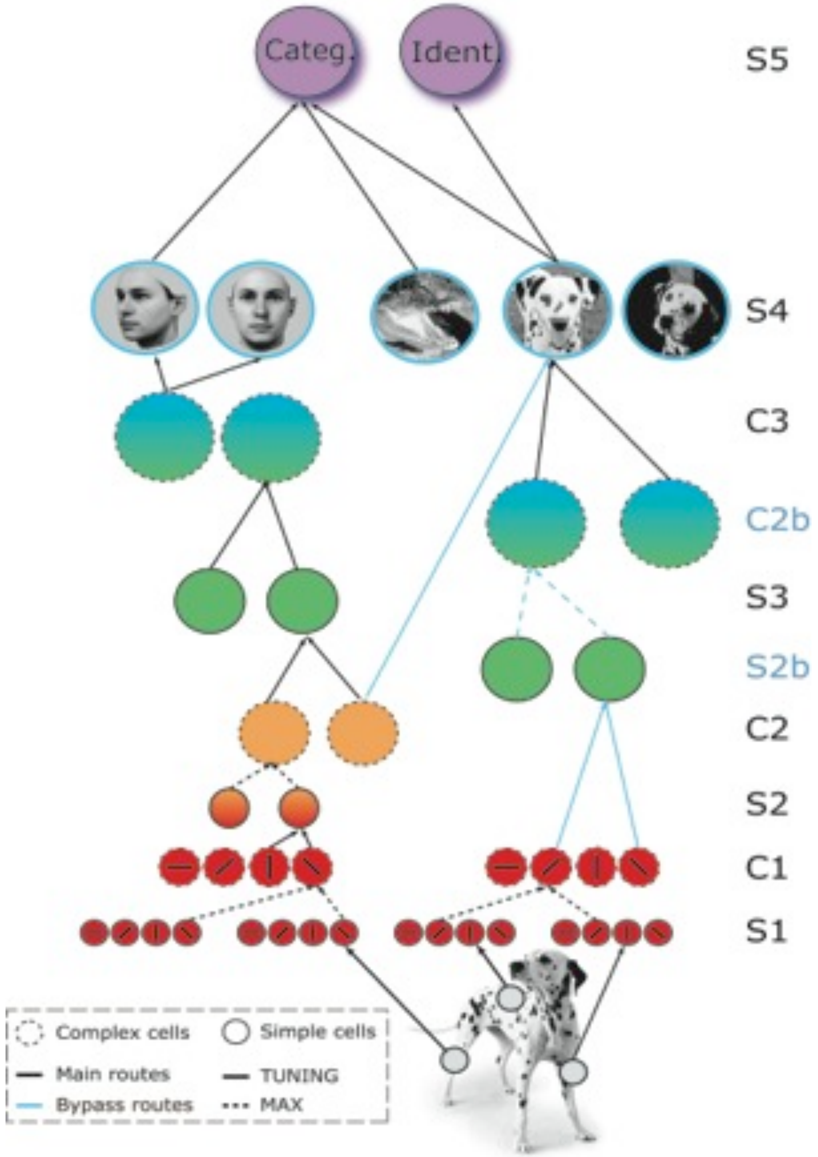
*Modified from (Gross, 1998)



[software available online
with CNS (for GPUs)]

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

Recognition in Visual Cortex: "classical model", selective and invariant



- It is in the family of "Hubel-Wiesel" models (Hubel & Wiesel, 1959: *qual.* **Fukushima**, 1980: *quant*; Oram & Perrett, 1993: *qual*; Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe, 2002; Ullman et al., 2002; Mel, 1997; Wersing and Koerner, 2003; LeCun et al 1998: *not-bio*; Amit & Mascaro, 2003: *not-bio*; Hinton, LeCun, Bengio *not-bio;* Deco & Rolls 2006…)

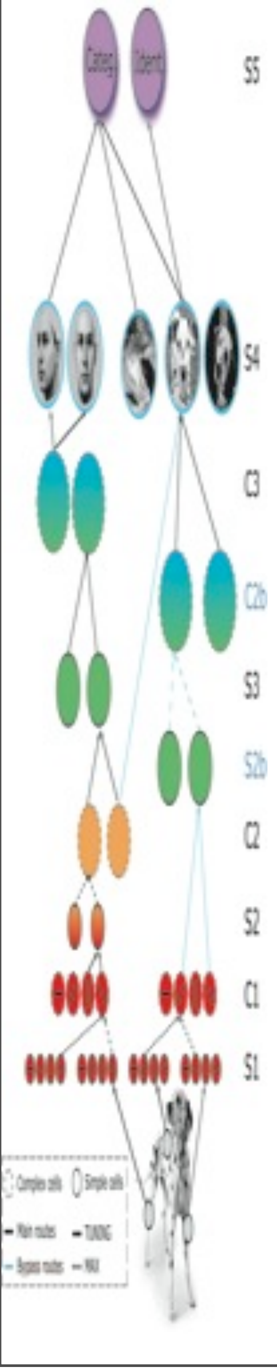
- As a biological model of object recognition in the ventral stream – from V1 to PFC -- it is *perhaps* the most quantitatively faithful to known neuroscience data

[software available online]

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

Model “works”:
it accounts for physiology

Hierarchical Feedforward Models:
is consistent with or predict 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)

V2:

Subunits and their tuning (Anzai, Peng, Van Essen 2007)

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)

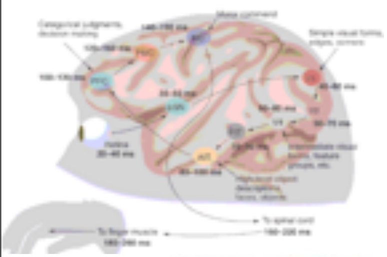
Pseudo-average effect in IT (Zoccolan Cox & DiCarlo 2005; Zoccolan Kouh Poggio & DiCarlo 2007)

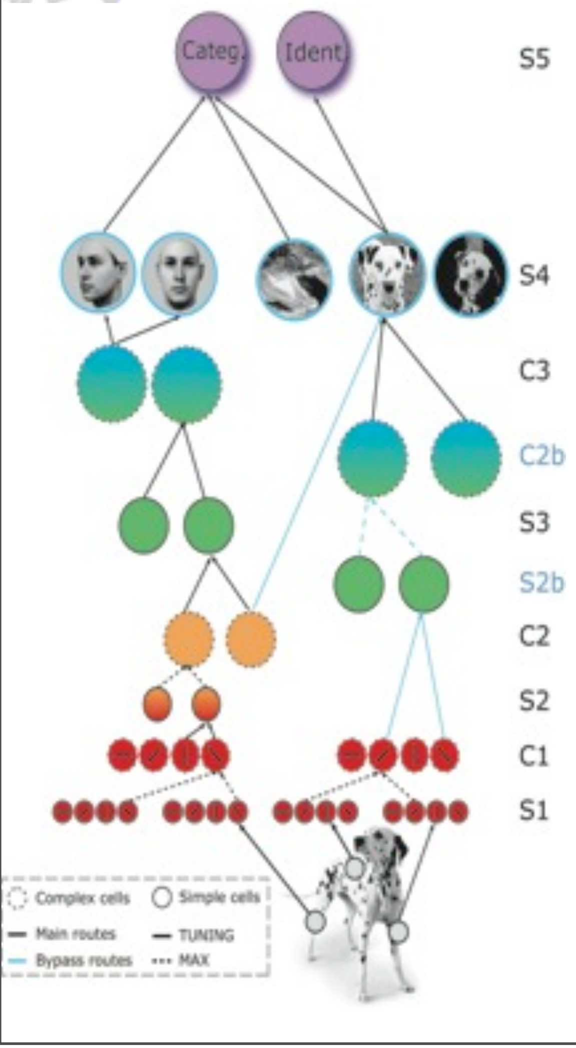
Human:

Rapid categorization (Serre Oliva Poggio 2007)

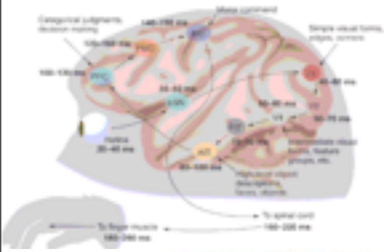
Face processing (fMRI + psychophysics) (Riesenhuber et al 2004; Jiang et al 2006)

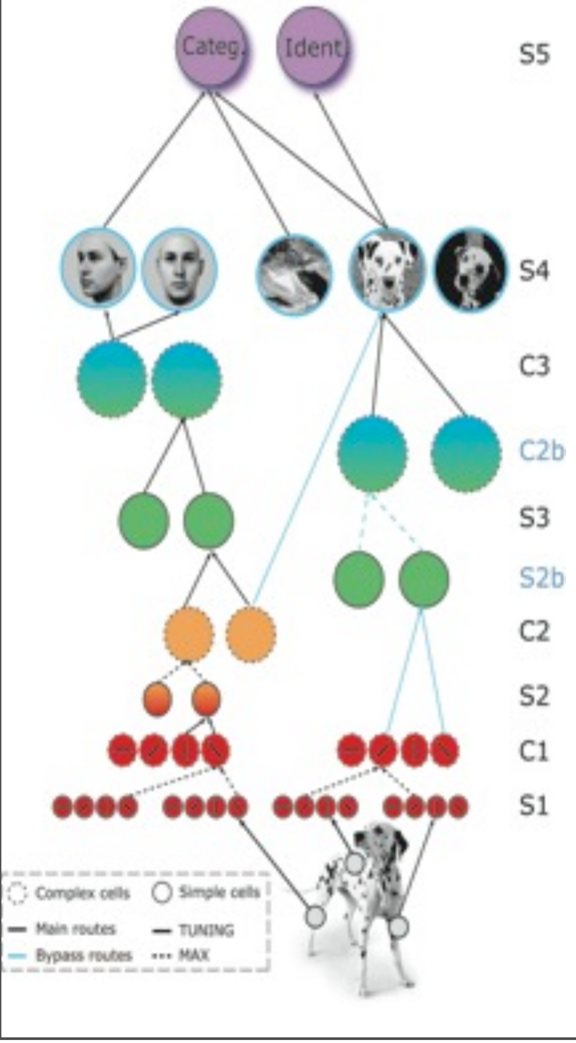
Model "works": it accounts for psychophysics


Categorical judgments, decision making
Motor command
Simple visual forms, edges, corners
PFC
PMC
MC
V1
V2
V4
PIT
AIT
LGN
Retina
20–40 ms
Intermediate visual forms, feature groups, etc.
High-level object descriptions, faces, objects
To spinal cord
160–220 ms
To finger muscle
180–260 ms

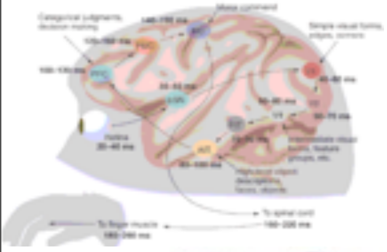


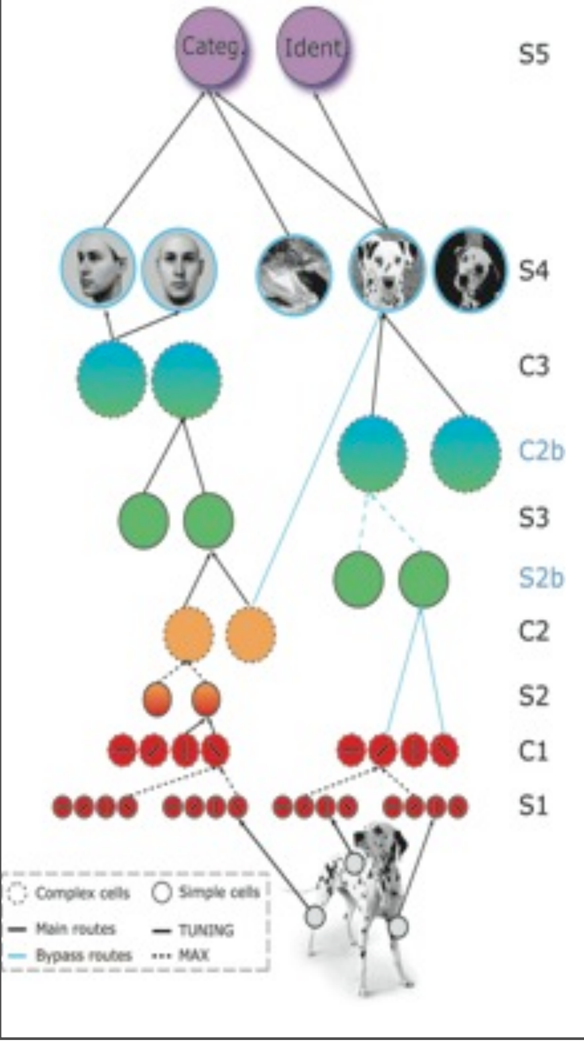
Model "works": it accounts for psychophysics





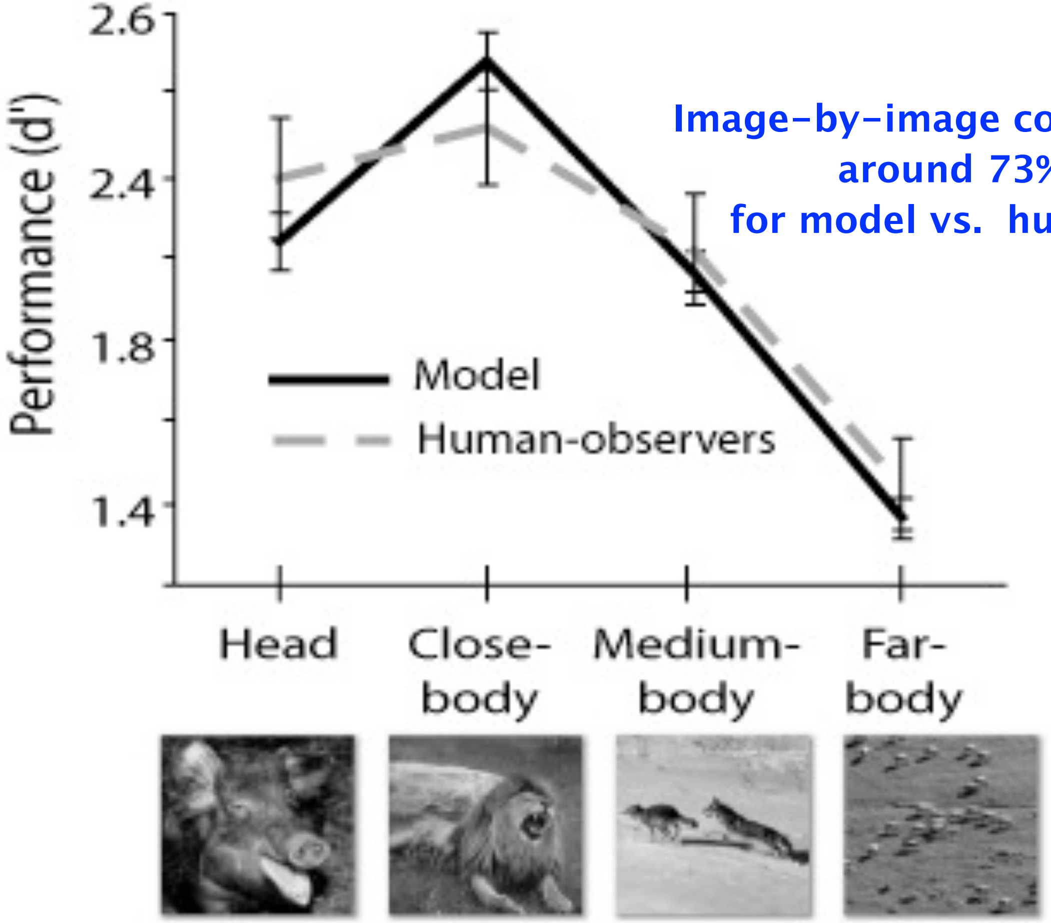
Model "works": it accounts for psychophysics





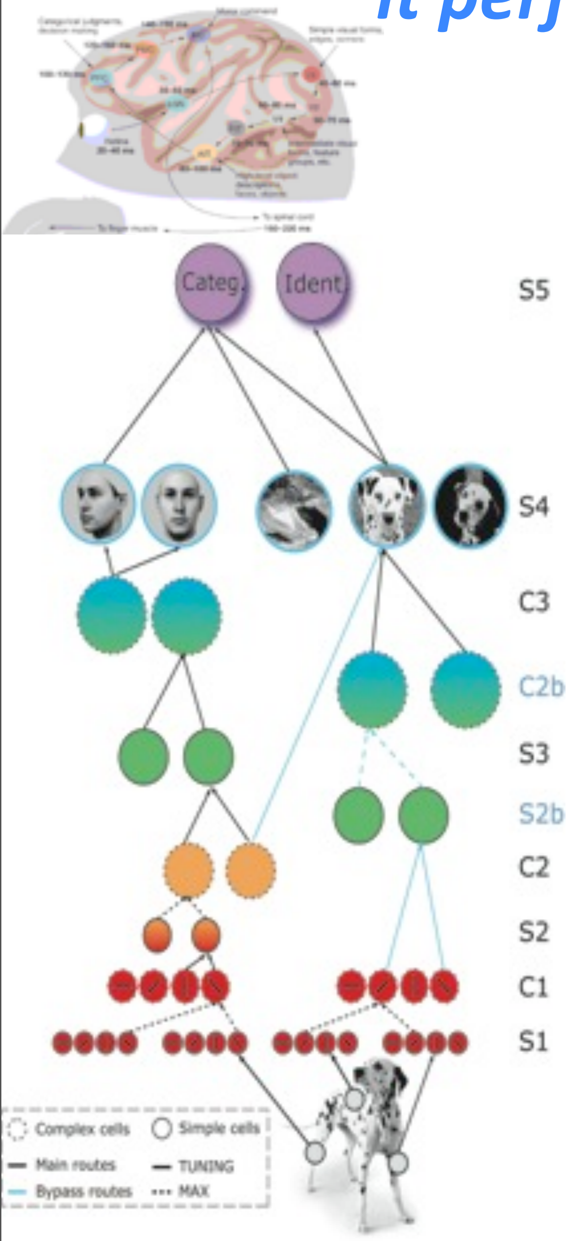
Feedforward Models: "predict" rapid categorization (82% model vs. 80% humans)

Image-by-image correlation: around 73% for model vs. humans)



Model “works”:

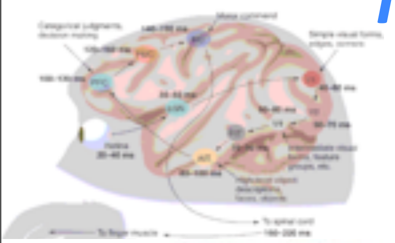
it performs well at computational level

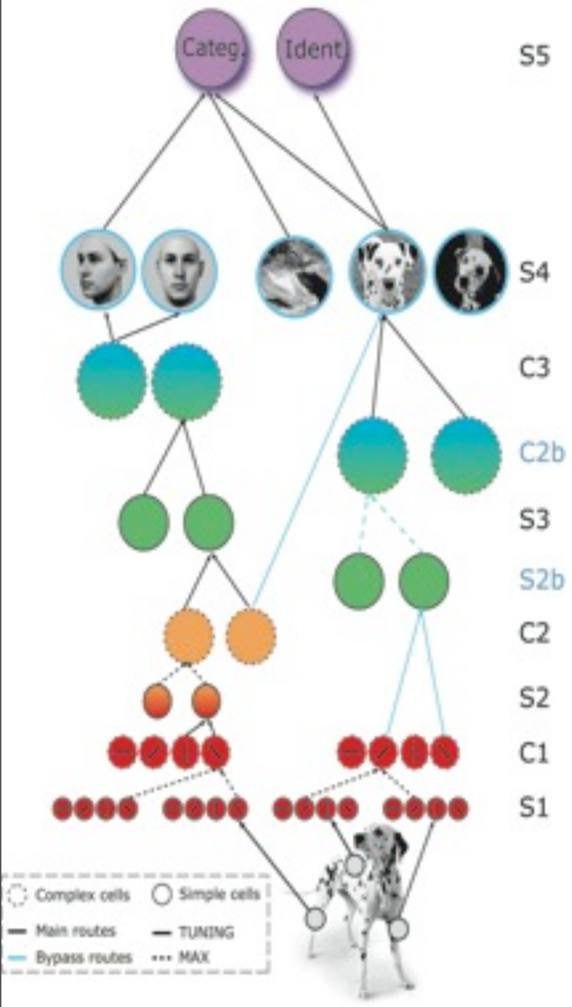


Models of the *ventral stream* in cortex
perform well compared to
engineered computer vision systems (in 2006)
on several databases

Bileschi, Wolf, Serre, Poggio, 2007

Model "works":
it performs well at computational level

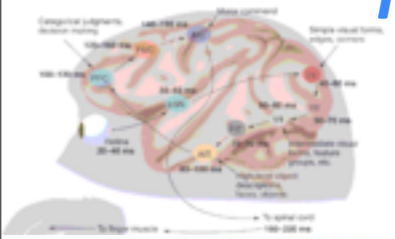


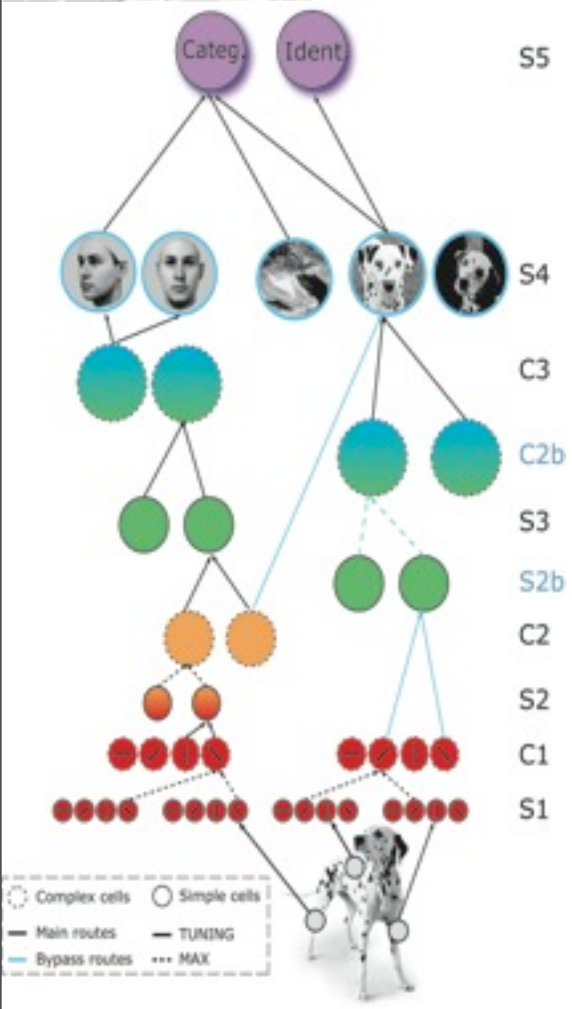


Models of the *ventral stream* in cortex perform well compared to engineered computer vision systems (in 2006) on several databases

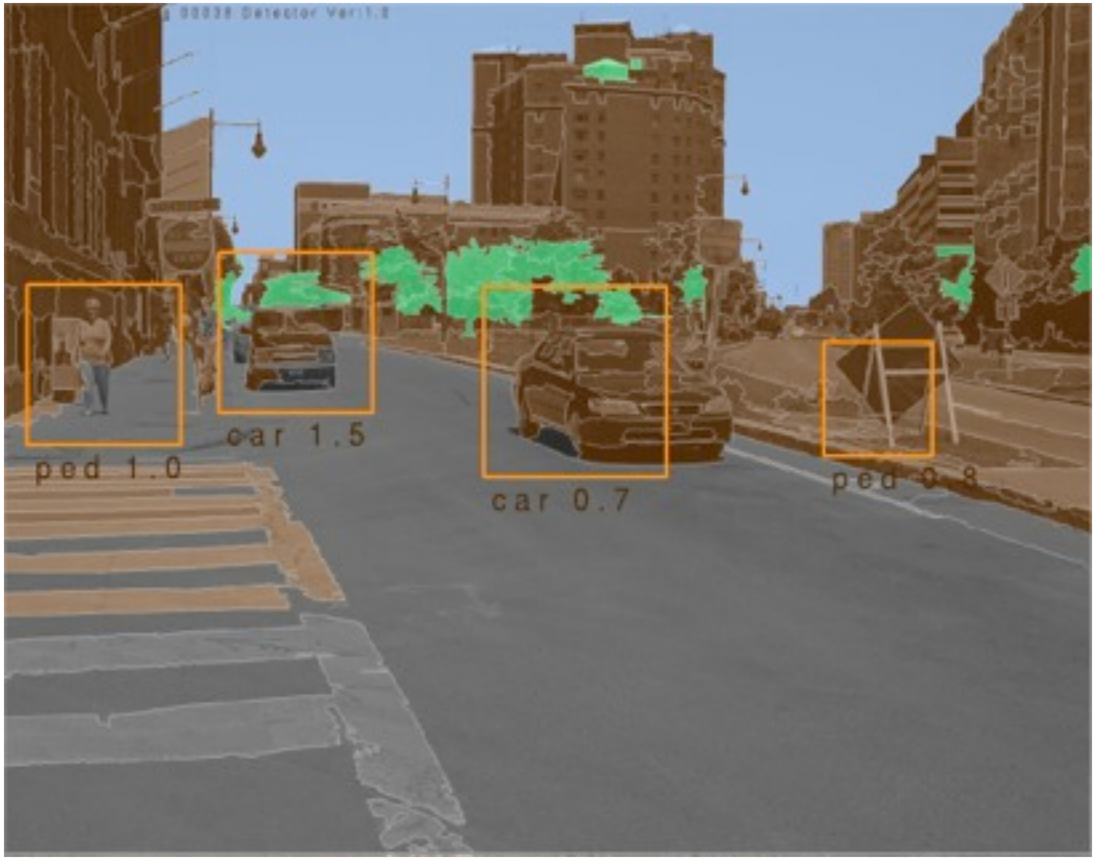
Bileschi, Wolf, Serre, Poggio, 2007

Model "works": it performs well at computational level





Models of the <u>*ventral stream*</u> in cortex perform well compared to engineered computer vision systems (in 2006) on several databases



Bileschi, Wolf, Serre, Poggio, 2007

Model "works":
it performs well at computational level

Performance

human agreement	72%
proposed system	77%
commercial system	61%
chance	12%

Models of cortex lead to better systems for action recognition in videos: automatic phenotyping of mice

Jhuang , Garrote, Yu, Khilnani, Poggio, Mutch Steele, Serre, Nature Communicatons, 2010

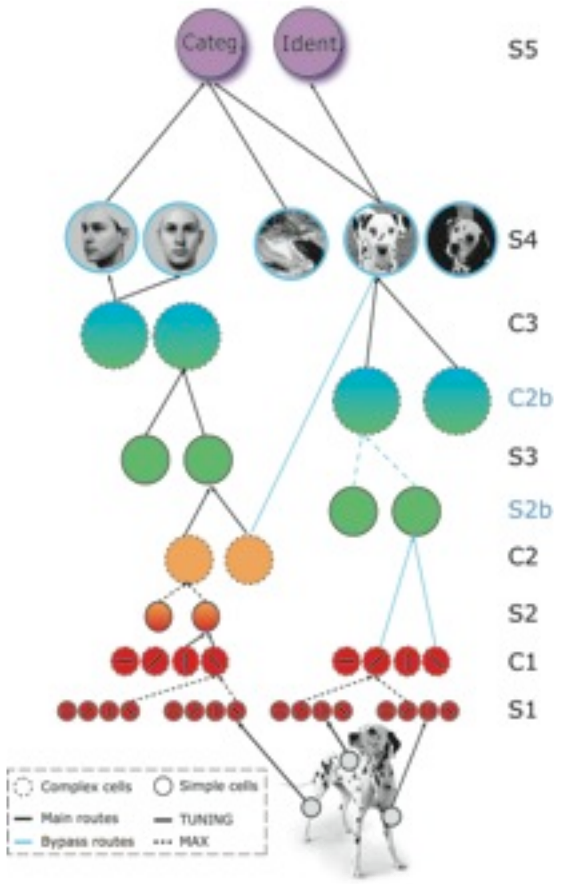
Model "works":
it performs well at computational level

Performance

human agreement	72%
proposed system	77%
commercial system	61%
chance	12%

Models of cortex lead to better systems for action recognition in videos: automatic phenotyping of mice

Jhuang , Garrote, Yu, Khilnani, Poggio, Mutch Steele, Serre, Nature Communicatons, 2010



Found Comput Math (2010) 10: 67–91
DOI 10.1007/s10208-009-9049-1

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Mathematics of the Neural Response

S. Smale · L. Rosasco · J. Bouvrie · A. Caponnetto · T. Poggio

A puzzle

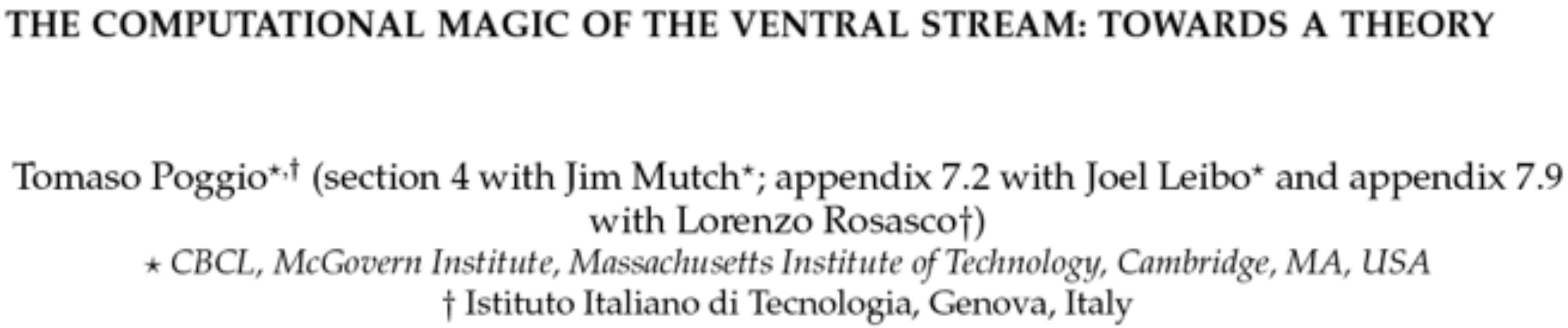
Hierarchical, HMAX-type models of visual cortex very well as
computer vision systems
but...why?

Very similar *convolutional networks*
now called deep learning networks
(LeCun, Hinton,...) are
unreasonably successful
in vision and speech (ImageNet+Timit)...

why?

We need theories!

A theory (unpublished) of visual cortex and of so-called deep learning architectures



THE COMPUTATIONAL MAGIC OF THE VENTRAL STREAM: TOWARDS A THEORY

Tomaso Poggio[*,†] (section 4 with Jim Mutch[*]; appendix 7.2 with Joel Leibo[*] and appendix 7.9 with Lorenzo Rosasco†)

⋆ *CBCL, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA*

† Istituto Italiano di Tecnologia, Genova, Italy

Theory of visual cortex:
from invariance it predicts tuning of cells
and architecture and function of cortex

Theory of visual cortex:
from invariance it predicts tuning of cells
and architecture and function of cortex

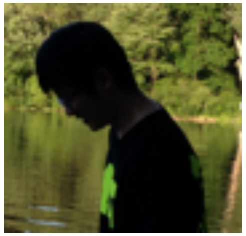
It may explain why deep convolutional
architecture do
so well in object recognition (ImageNet) and
speech recognition
and
how to make them better

Theory of visual cortex:
from invariance it predicts tuning of cells
and architecture and function of cortex

It may explain why deep convolutional
architecture do
so well in object recognition (ImageNet) and
speech recognition
and
how to make them better

It is a theory of
unsupervised learning
of representations
for supervised learning

Collaborators (MIT-IIT, LCSL) in recent work



F. Anselmi, J. Mutch , J. Leibo, L. Rosasco, A. Tacchetti, Q. Liao

+ +

L. Isik, S. Ullman, S. Smale, C. Tan

Also: M. Riesenhuber, T. Serre, G. Kreiman, S. Chikkerur, A. Wibisono, J. Bouvrie, M. Kouh, J. DiCarlo, E. Miller, C. Cadieu, A. Oliva, C. Koch, A. Caponnetto ,D. Walther, U. Knoblich, T. Masquelier, S. Bileschi, L. Wolf, E. Connor, D. Ferster, I. Lampl, S. Chikkerur, G., N. Logothetis, H. Buelthoff

Motivation

Cardinality of the universe of possible images generated by an object:

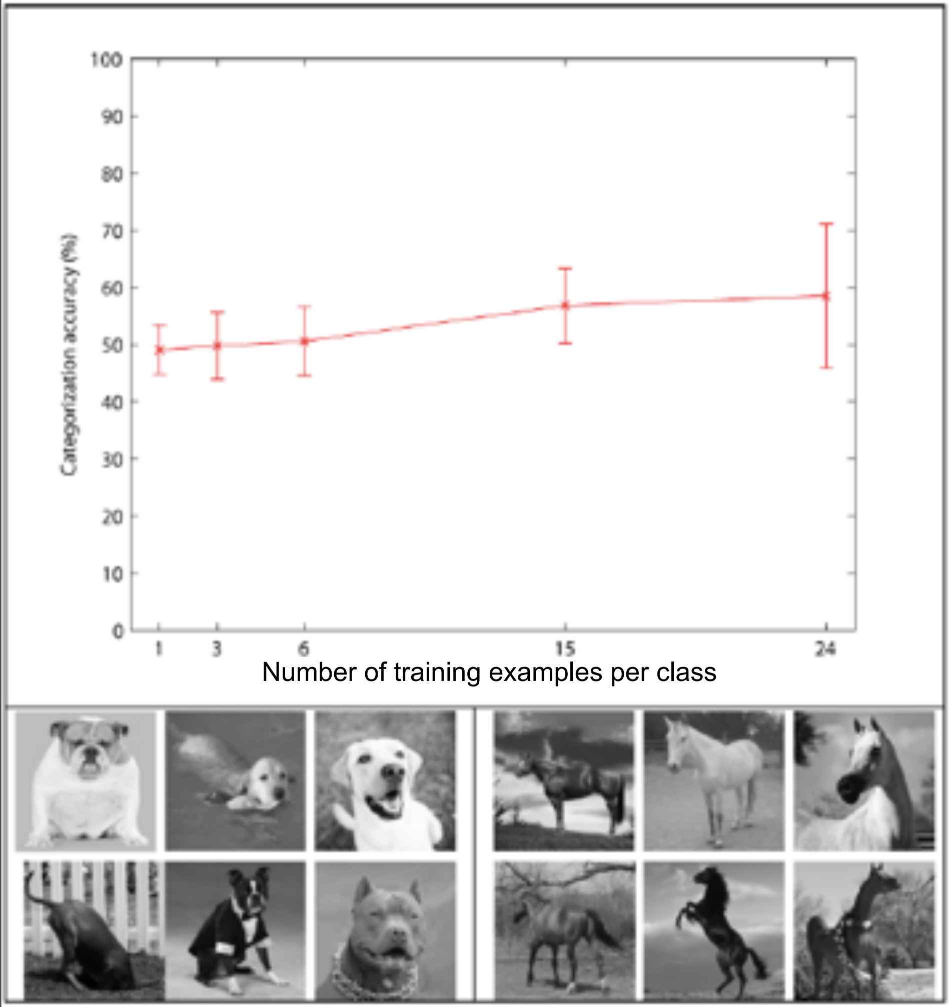
- Assuming: a granularity of a few minutes of arc + a visual field of say 10 degrees
- then
  - $10^3 - 10^5$ different images of the same object from x, y translations
  - $10^3 - 10^5$ from rotations in depth
  - a factor of $10 - 10^2$ from rotations in the image plane
  - another factor of $10 - 10^2$ from scaling.

  for a total $10^8 - 10^{14}$ distinguishable images for a single object.

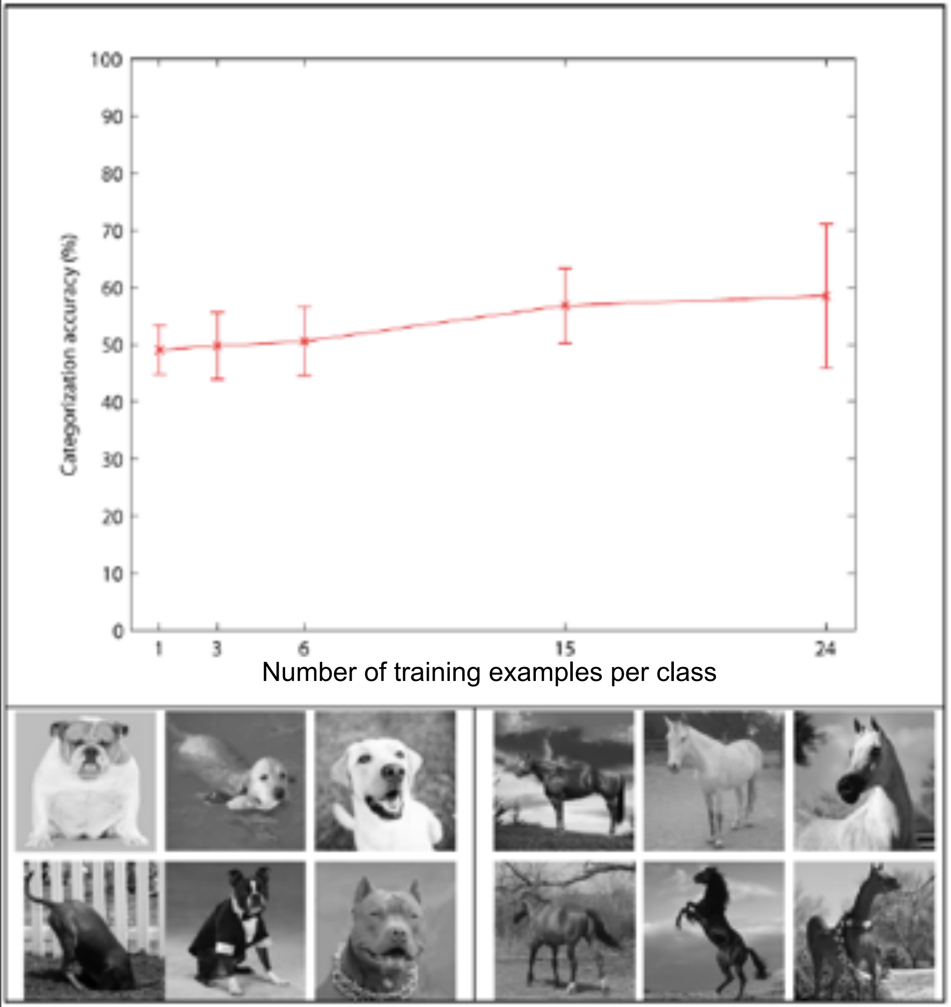
How many different types of dogs exist within the "dog" category? No more than, say, $10^2 - 10^3$. Thus it is greater win to be able to factor out the geometric transformations than the intracategory differences.

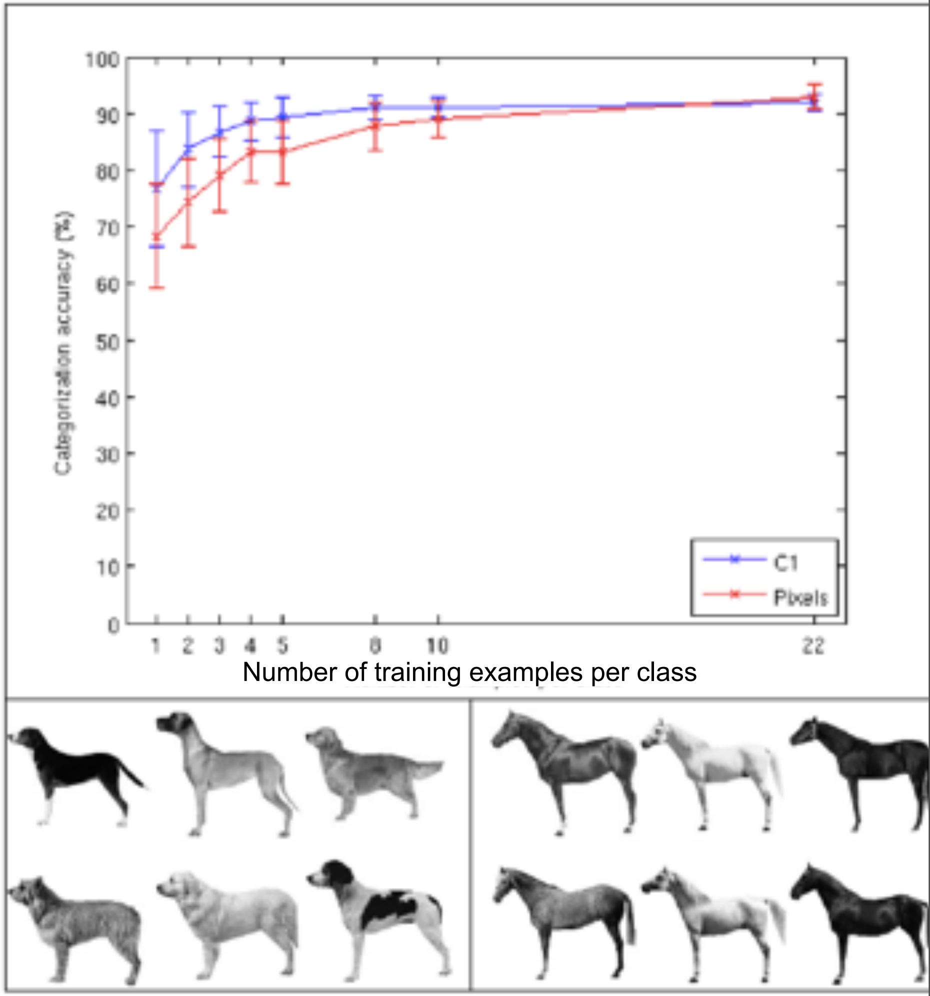
Computing invariant representations for perception:
is this the computational goal of the ventral stream?
the magic of sensory cortex?

Computing invariant representations for perception: is this the computational goal of the ventral stream? the magic of sensory cortex?

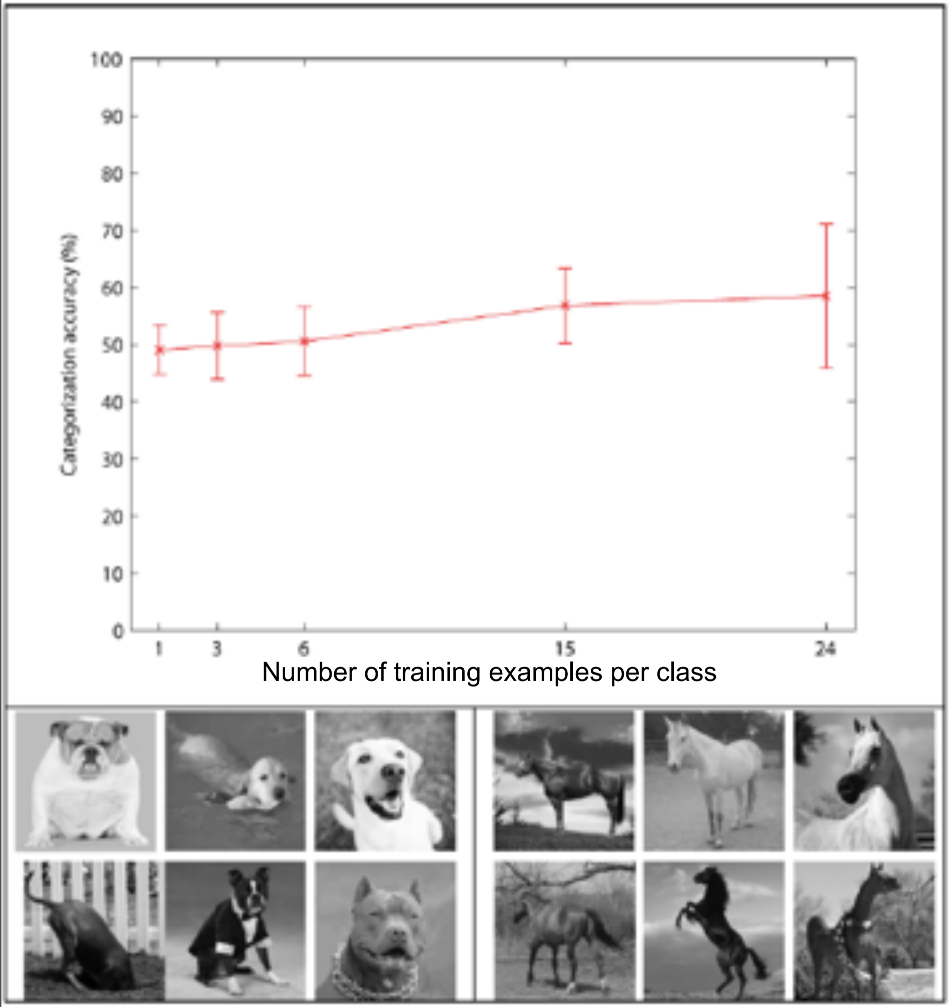


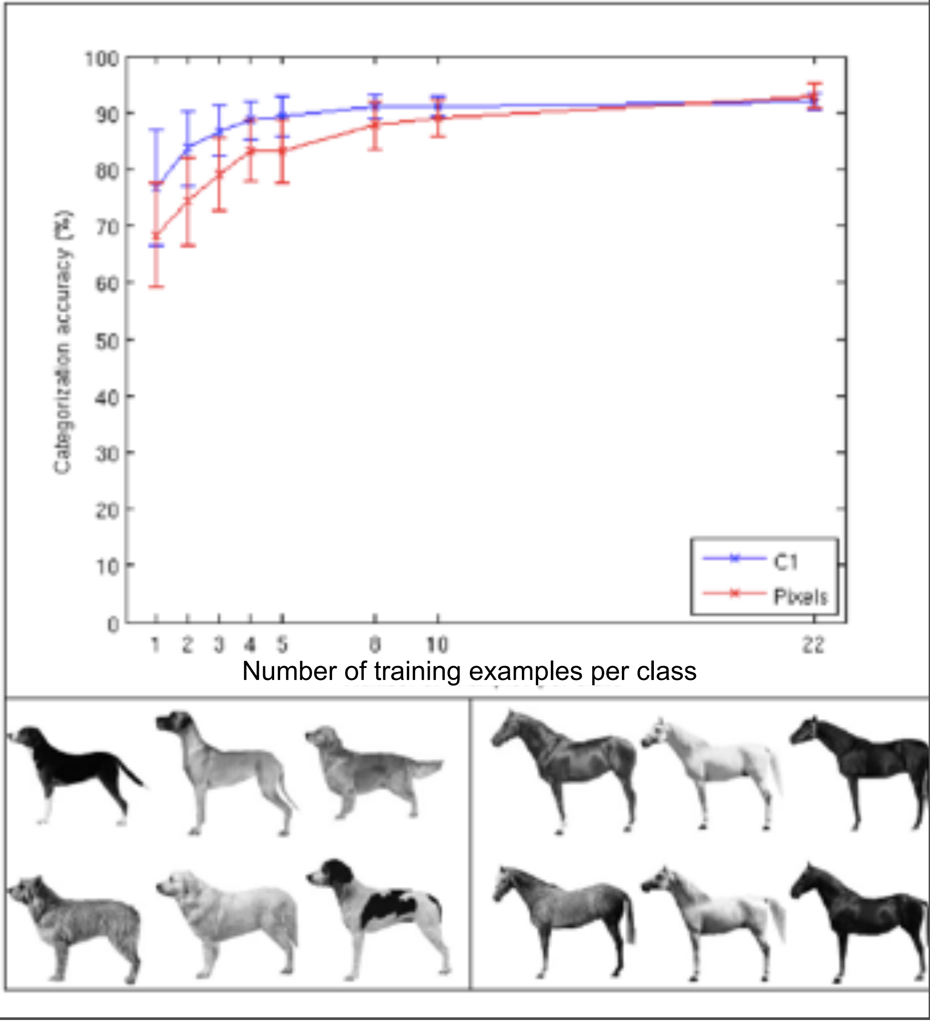
Computing invariant representations for perception: is this the computational goal of the ventral stream? the magic of sensory cortex?

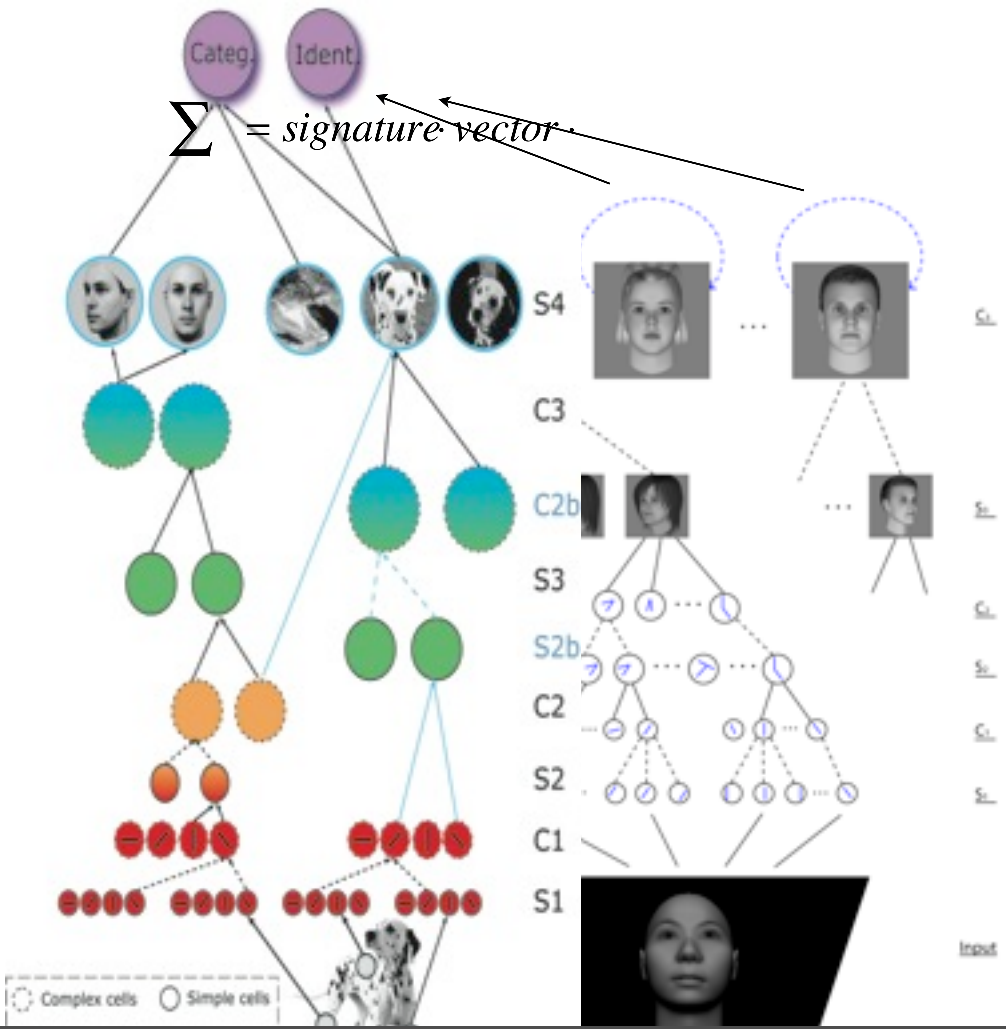




Computing invariant representations for perception: is this the computational goal of the ventral stream? the magic of sensory cortex?

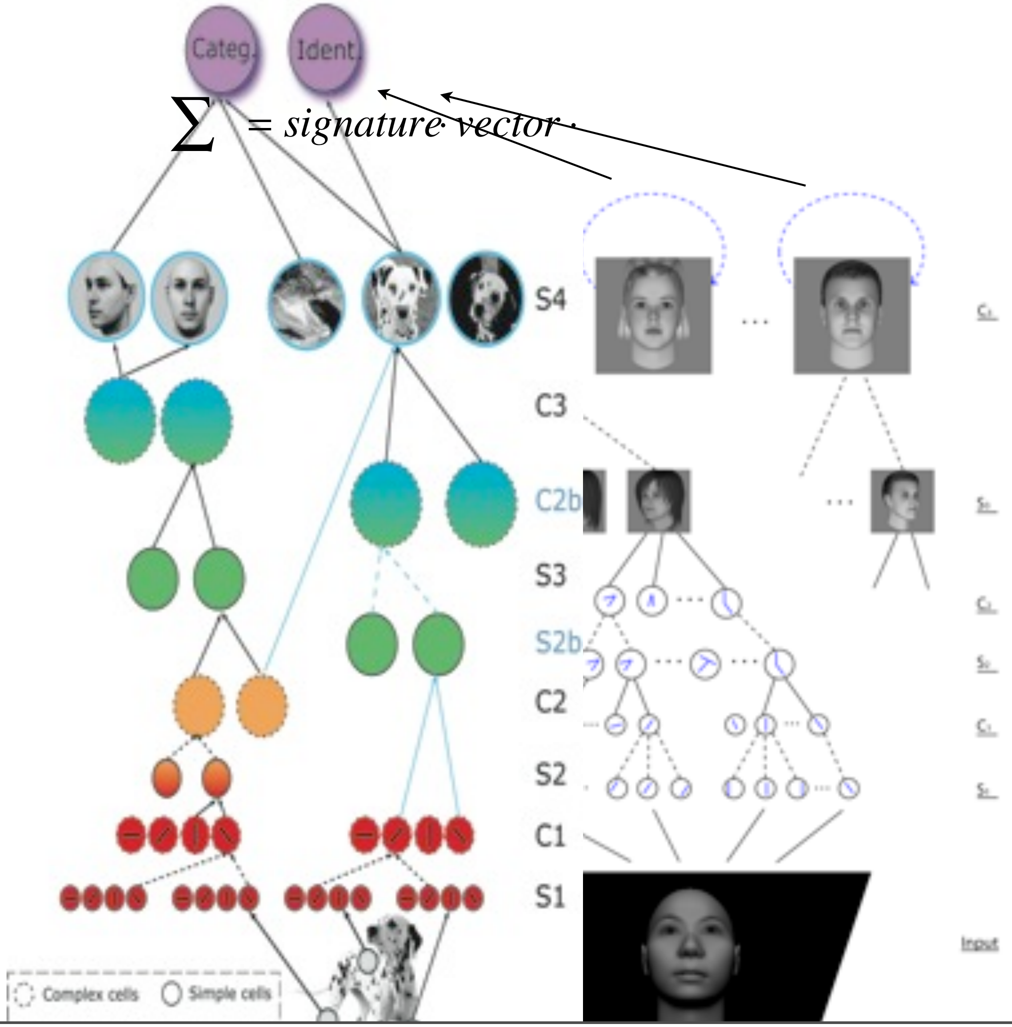




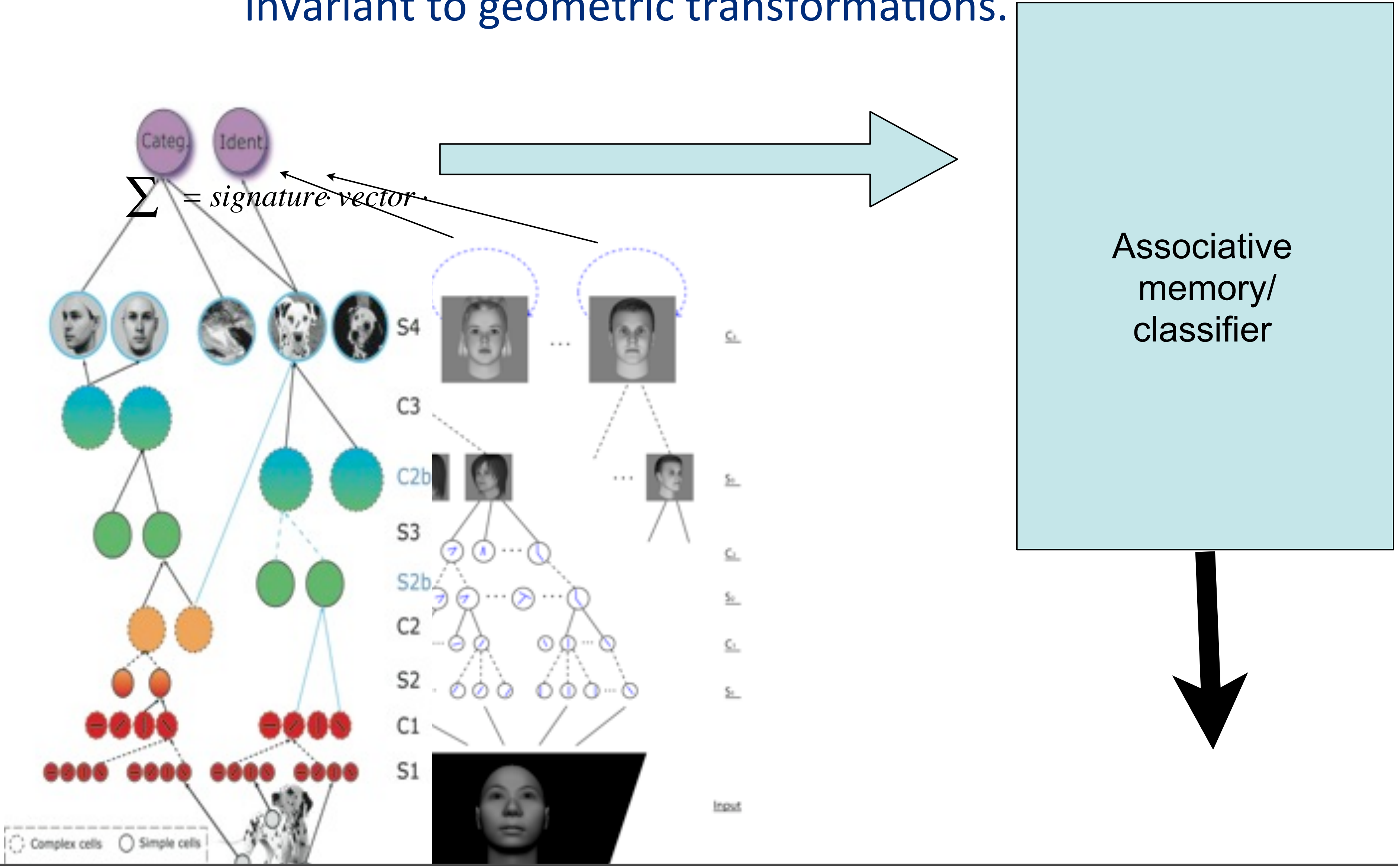


$\sum$ = *signature vector* ·

Conjecture: the key computational problem “solved” by the ventral stream is object recognition from a single training image, invariant to geometric transformations.



Conjecture: the key computational problem "solved" by the ventral stream is object recognition from a single training image, invariant to geometric transformations.



Associative memory/ classifier

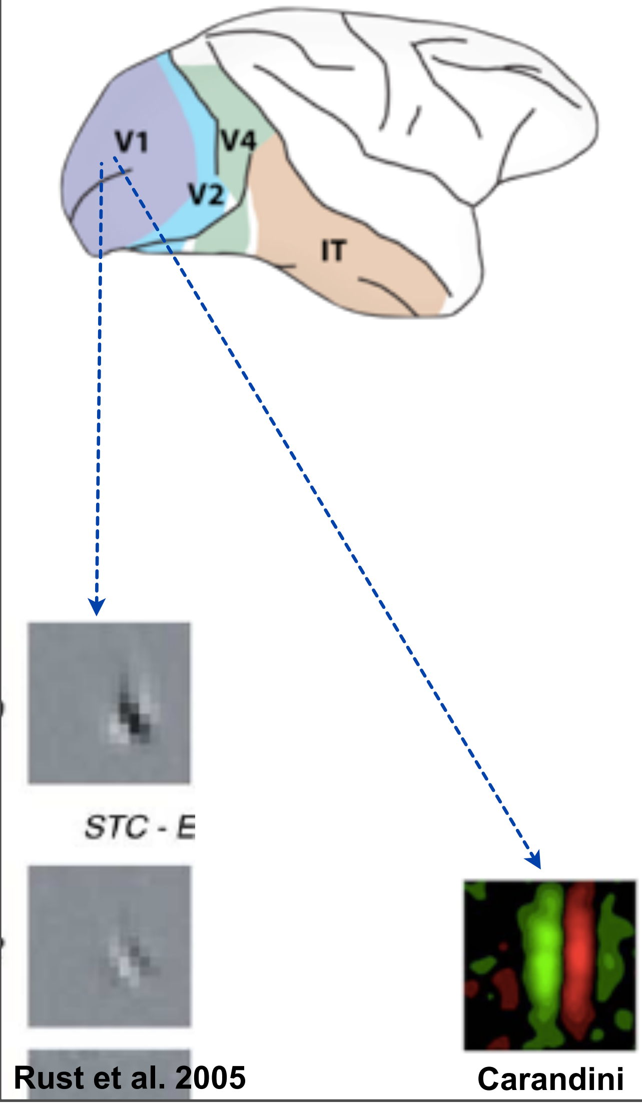
Conjecture: the key computational problem
“solved” by the ventral stream is
object recognition from a single training image,
invariant to geometric transformations.

The goal of the ventral stream would be preprocessing of image into a representation which is invariant: this would reduce significantly the sampling complexity of the learning problem for the classifier -->
learning from ~ one example.

Some of the questions answered by the theory

- What is the main computational task of the ventral stream?
- Why do simple cells in V1 have Gabor tuning curves?
- What are V2, V4, IT computing?
- Why do cells in the AL *face* patch show mirror symmetric tuning curves?

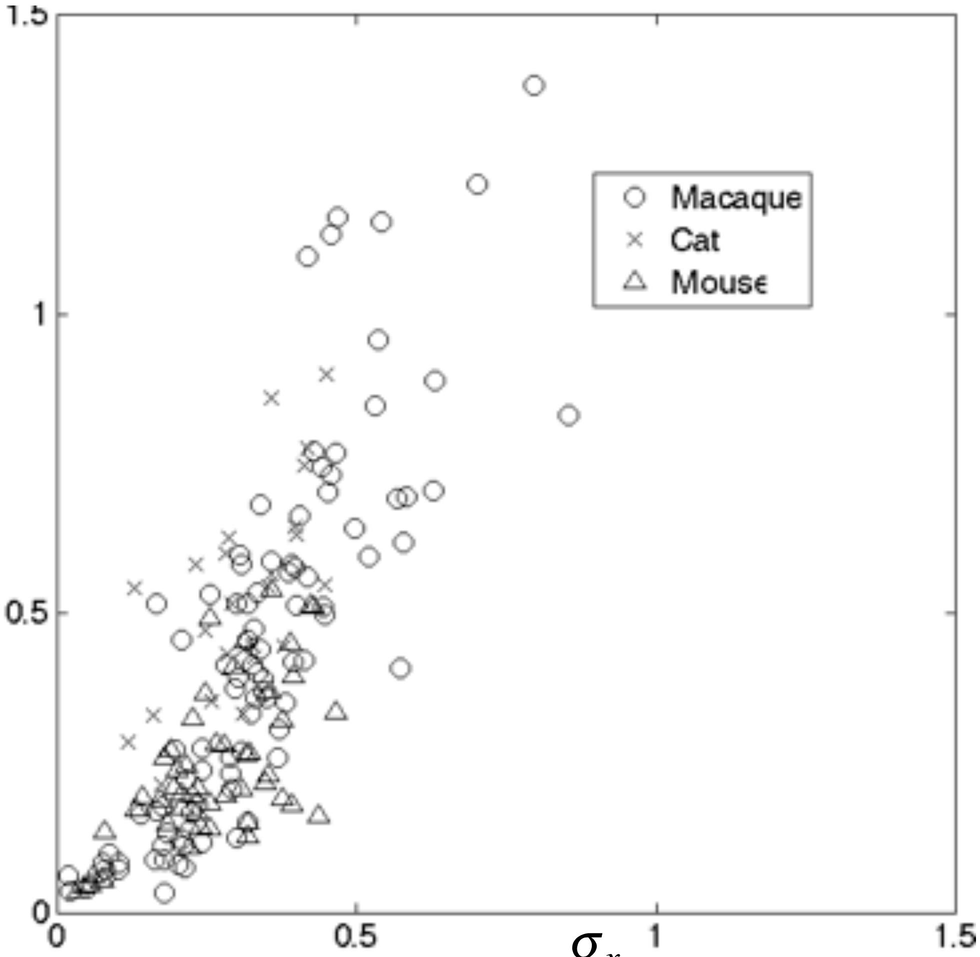
Gabor-like tuning with “universal constants” in simple cells (Jones and Palmer, 1987; Ringach, 2002; Niell and Stryker, 2008): why?



Rust et al. 2005

Carandini

$$n_y == \frac{\sigma_y}{\lambda}$$



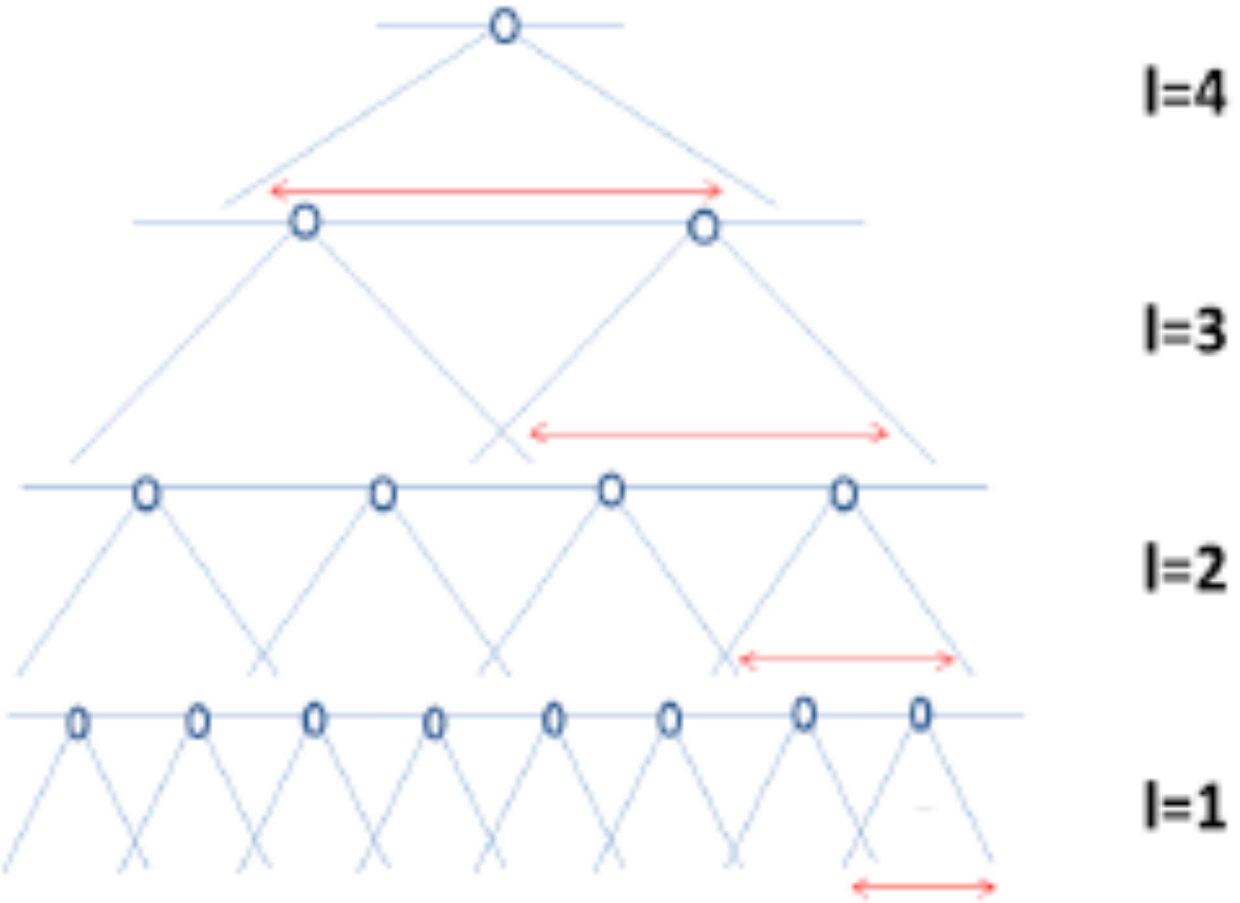
$$n_x = \frac{\sigma_x}{\lambda}$$

2 Different stages in the theory

1. development: learning of transformations (and acquiring invariance) via motion sequences

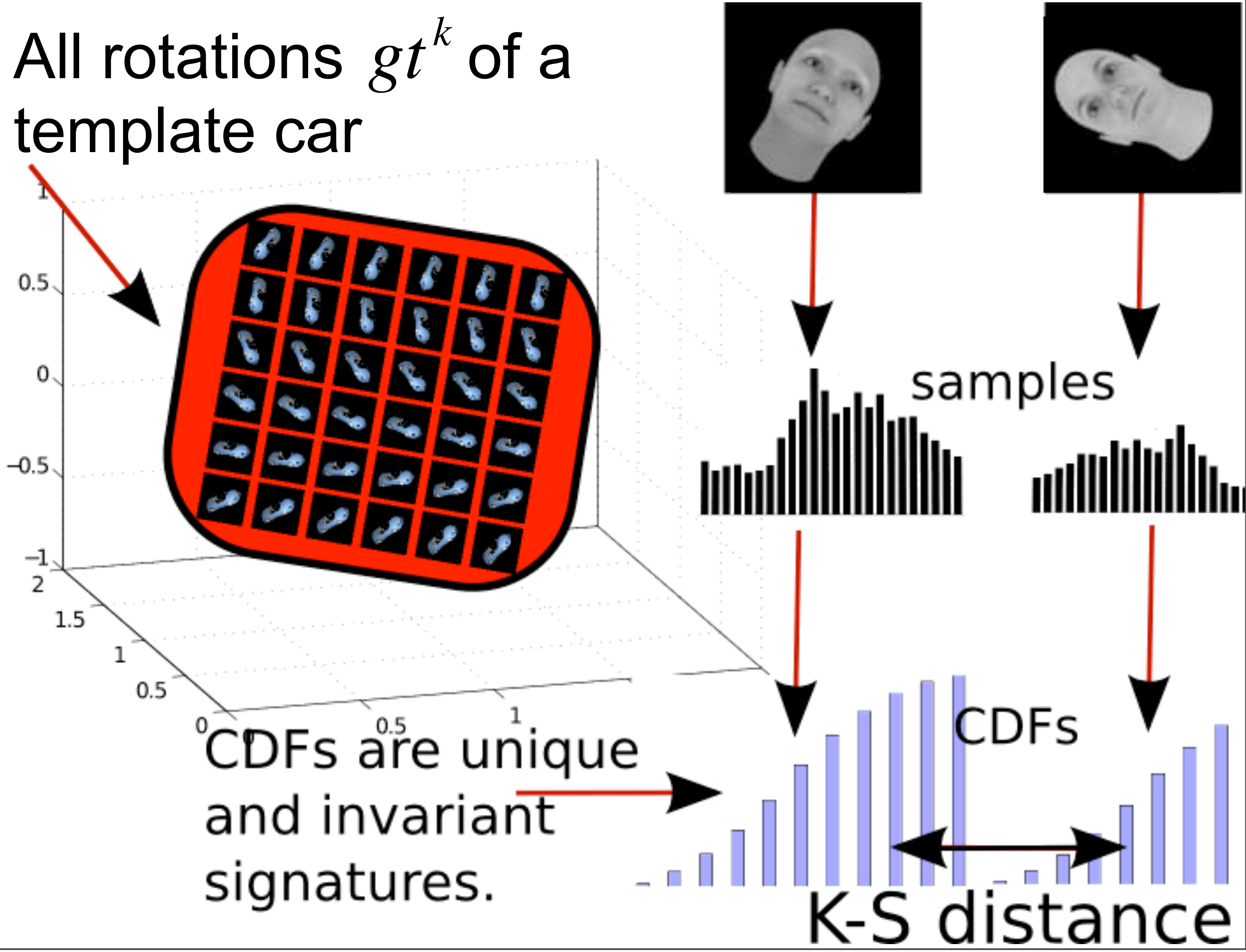
2. mature stage: acquire an object (single image) and (later) recognize it (from single image)

Multilayer architectures



Basic Idea

All rotations gt^k of a template car



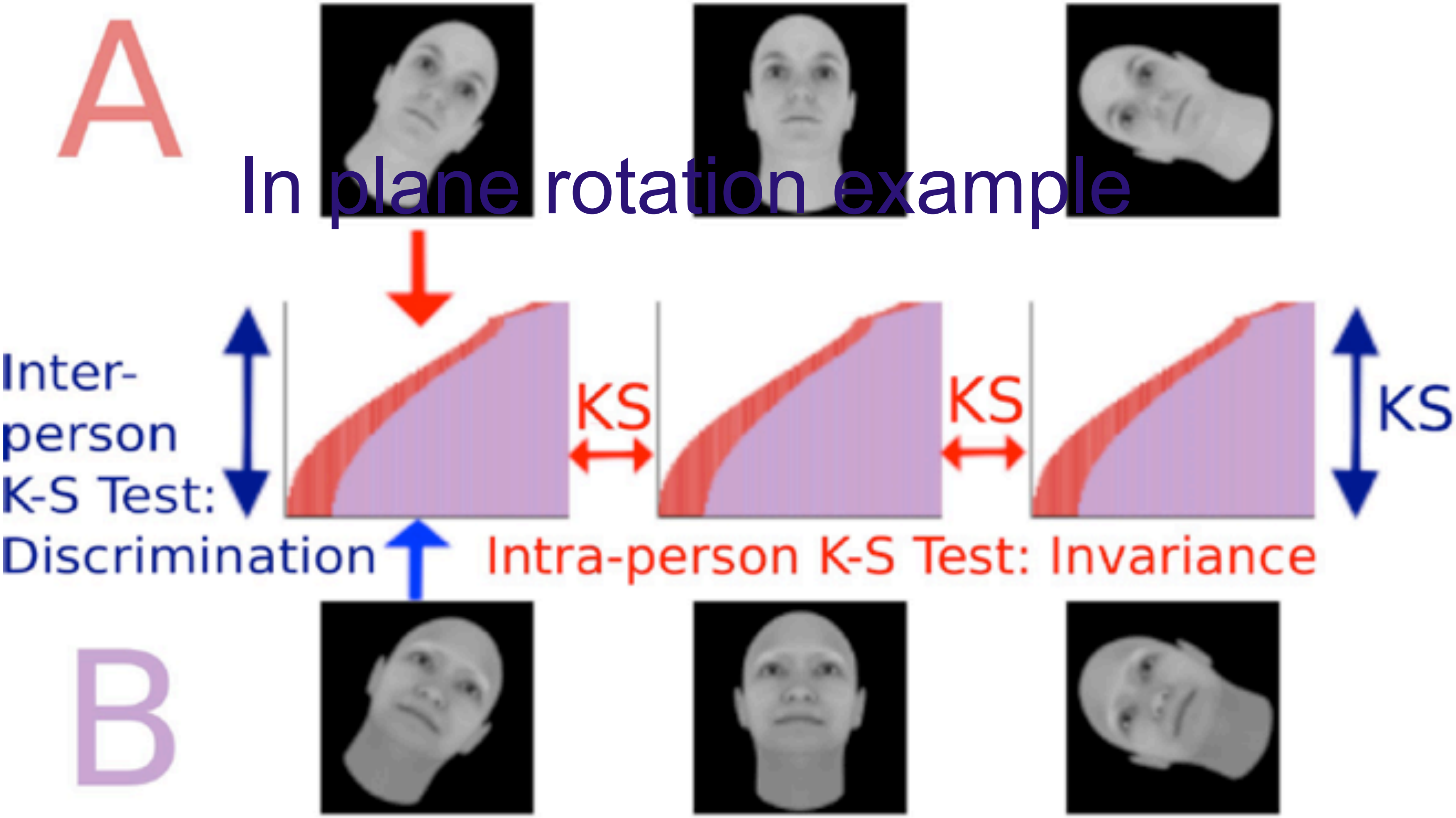
samples

CDFs

CDFs are unique and invariant signatures.

K-S distance

In plane rotation example


A
Inter-person K-S Test: Discrimination
KS
KS
KS
Intra-person K-S Test: Invariance
B

In plane rotation example

Focus of theory:

in multilayer architectures covariance of hierarchy

means

study the basic module!

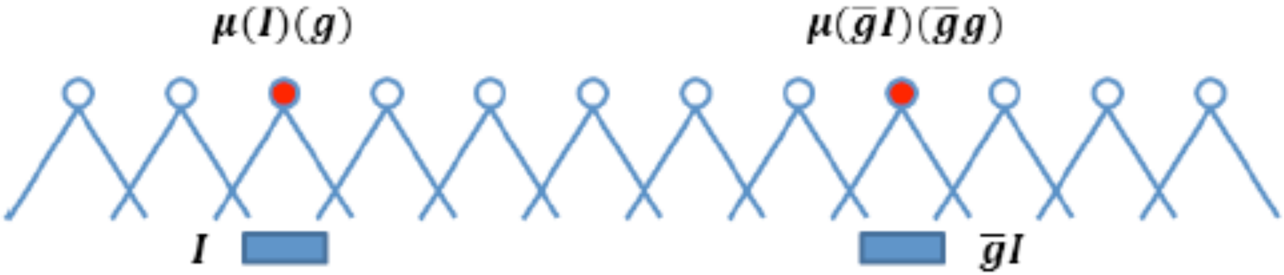


Fig. 3: *Covariance: the response for an image I at position g is equal to the response of the group shifted image at the shifted position.*

Hierarchy and covariance

- Each module provides a feature vector, that we call a signature, invariant to affine transformations of the images within its receptive field (or better pooling range).

- The hierarchical architecture, since it computes a set of signatures for different parts of the image, is invariant to the rather general family of locally affine transformations (which includes globally affine transformations of the whole image).

- This property of hierarchical architectures (see Fig.1) follows from *covariance* of the architecture for image transformations and from the uniqueness and invariance of the individual module signatures.

Affine transformations

We define as geometric transformations of the image I transformations $T \circ I$ such that:

$$T \circ I(x, y) = I(x', y')$$

An example of T is the affine case, eg

$$\mathbf{x}' = A\mathbf{x} + \mathbf{t}_{\mathbf{x}}$$

Image representation in the ventral stream

Image representation in the ventral stream

- Images can be represented by a set of functionals on the image, eg a set of measurements

Image representation in the ventral stream

Image representation in the ventral stream

- Neuroscience suggests that natural functionals for a neuron to compute is a high-dimensional dot product between an "image patch" and another image patch (called *template)* which is stored in terms of synaptic weights (synapses per neuron $\sim 10^2 - 10^5$)

Image representation in the ventral stream

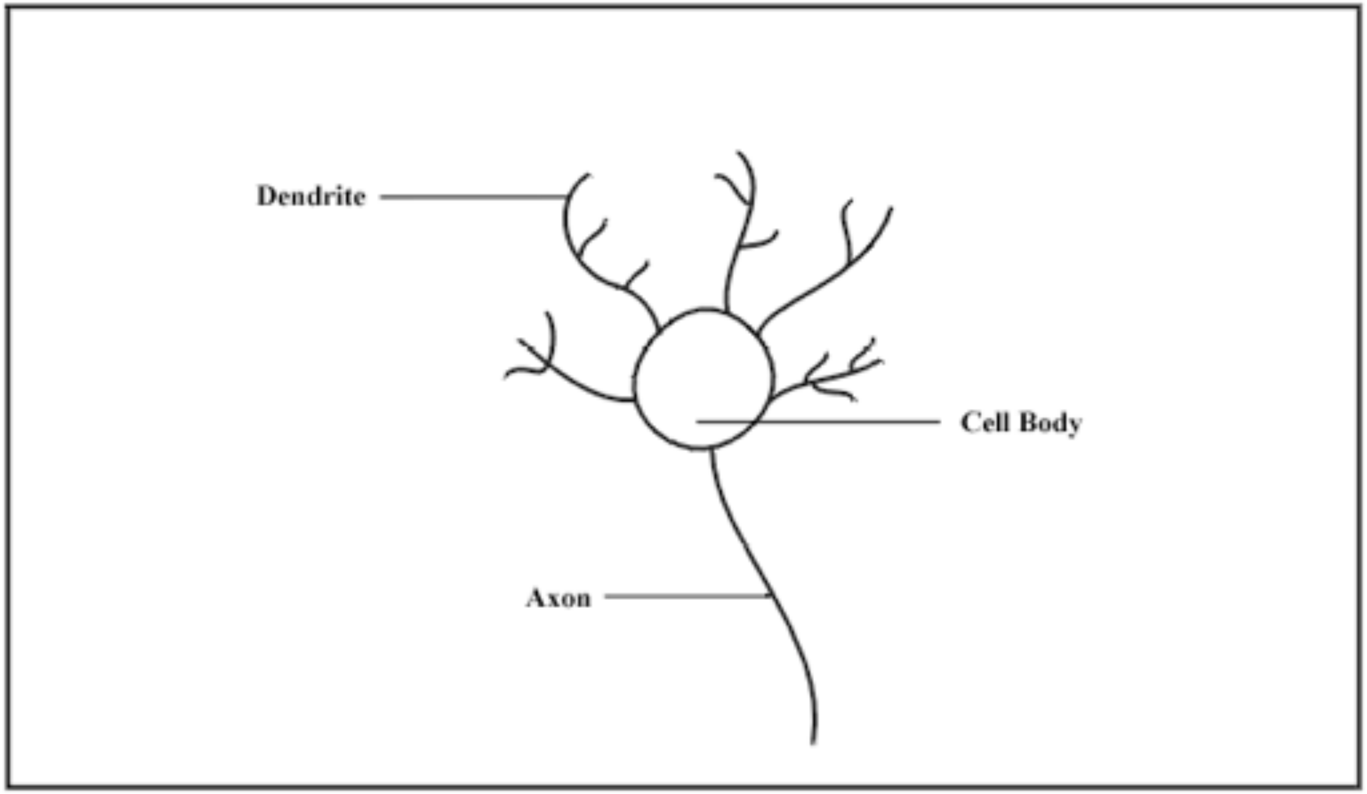
- Neuroscience suggests that natural functionals for a neuron to compute is a high-dimensional dot product between an "image patch" and another image patch (called *template)* which is stored in terms of synaptic weights (synapses per neuron $\sim 10^2 - 10^5$)

$$x \bullet t$$

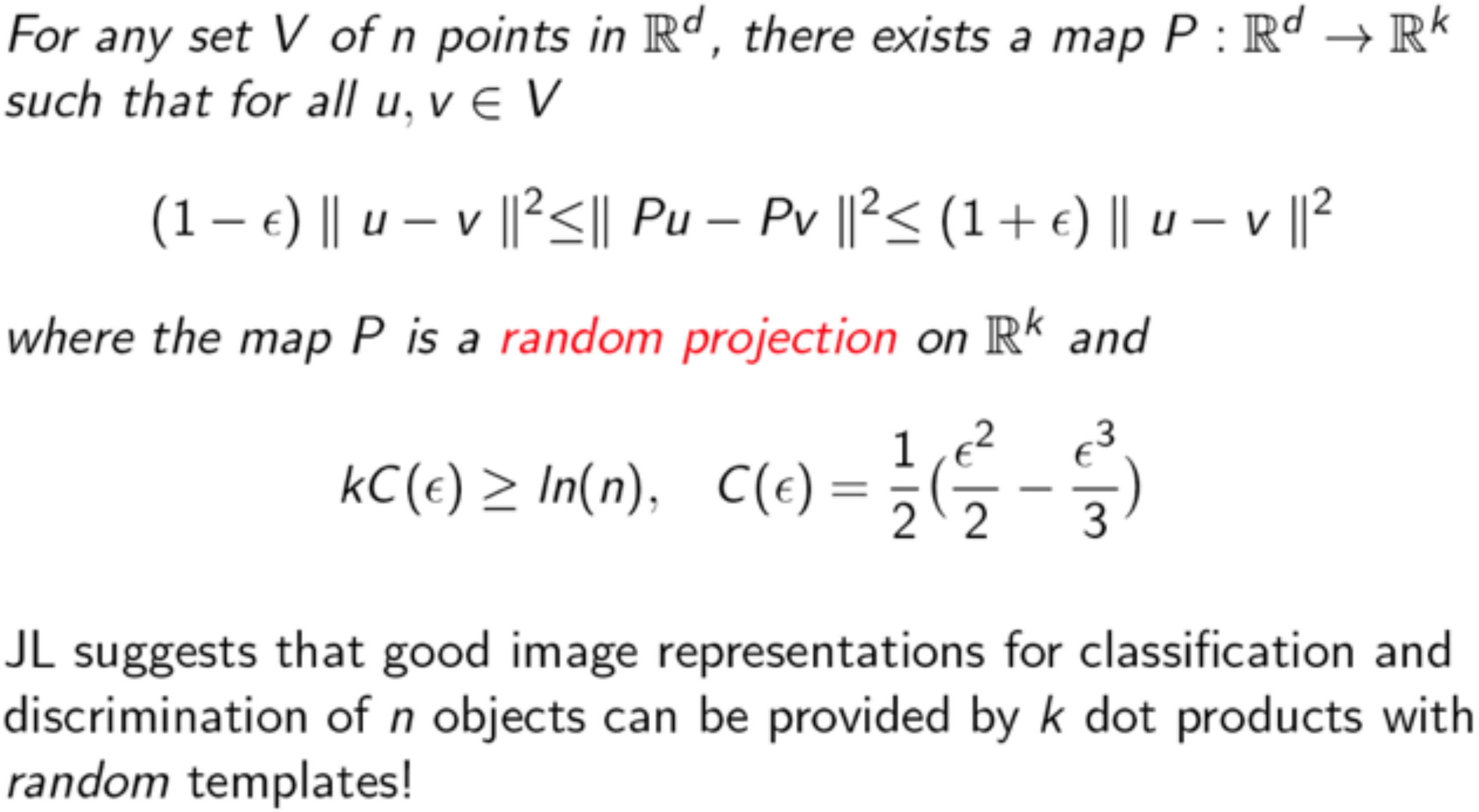
Image representation in the ventral stream

- Neuroscience suggests that natural functionals for a neuron to compute is a high-dimensional dot product between an "image patch" and another image patch (called *template)* which is stored in terms of synaptic weights (synapses per neuron $\sim 10^2 - 10^5$)

$$x \bullet t$$



A motivation for signatures: the Johnson-Lindenstrauss theorem (features do not matter much!)



For any set V *of* n *points in* $\mathbb{R}^d$, *there exists a map* $P : \mathbb{R}^d \to \mathbb{R}^k$ *such that for all* $u, v \in V$

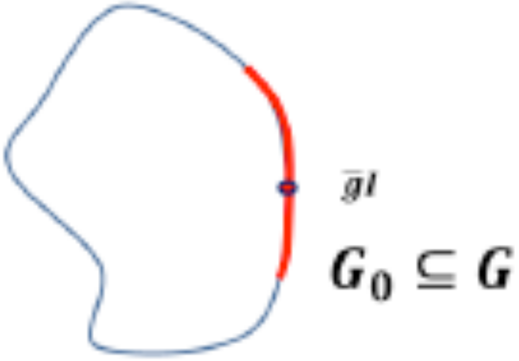
$$(1-\epsilon) \parallel u - v \parallel^2 \leq \parallel Pu - Pv \parallel^2 \leq (1+\epsilon) \parallel u - v \parallel^2$$

where the map P *is a random projection on* $\mathbb{R}^k$ *and*

$$kC(\epsilon) \geq ln(n), \quad C(\epsilon) = \frac{1}{2}(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3})$$

JL suggests that good image representations for classification and discrimination of n objects can be provided by k dot products with *random* templates!

Images, groups and orbit



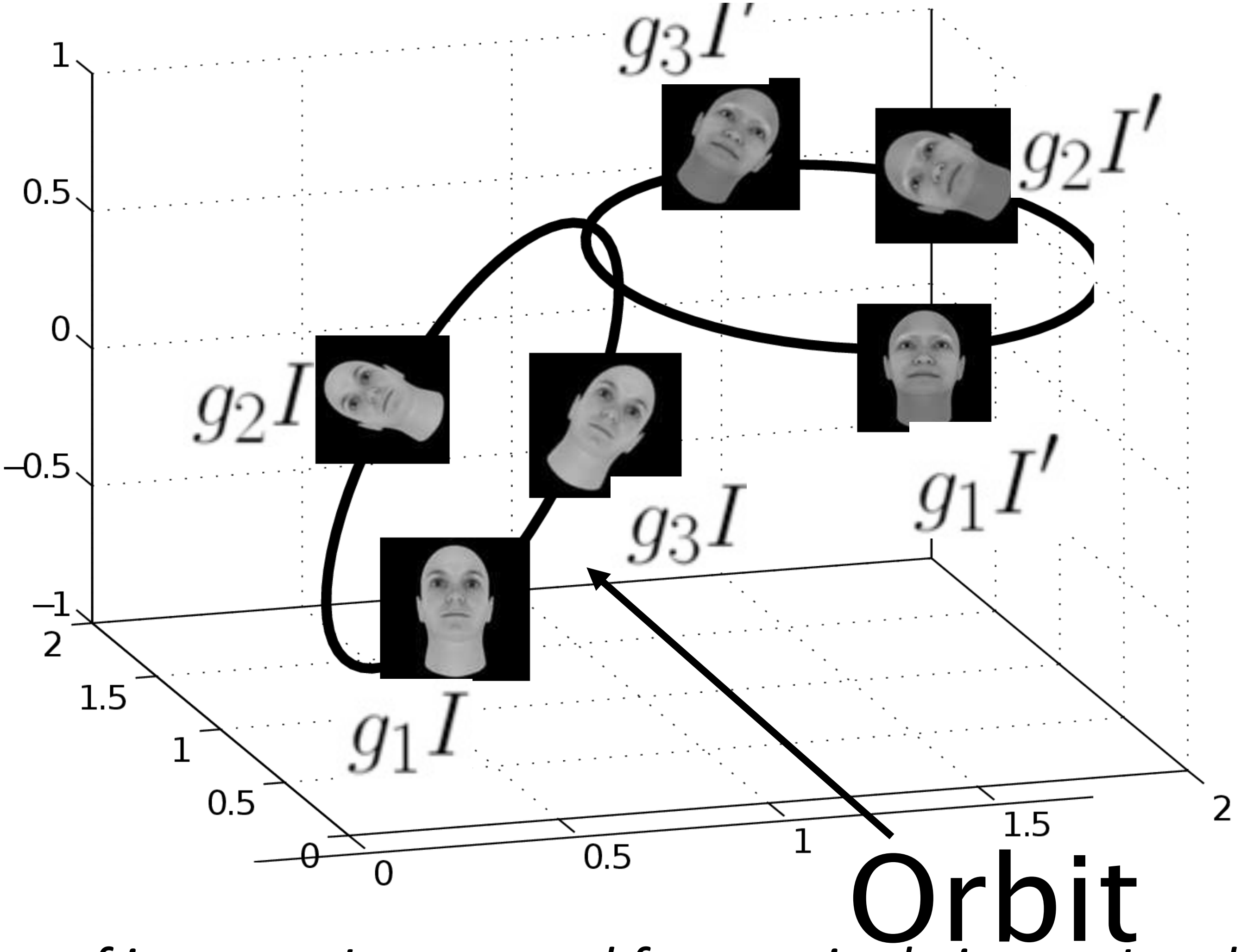
Orbit O_I

$I \sim I' if \quad \exists g \in G$ s.t. $I' = gI$

Orbit O_I can be proved to be *invariant* and *unique*

Orbit is unique and invariant

$$I \sim I' \iff O_I = O_{I'}$$



Orbit: set of images gI generated from a single image I under the action of the group

Preview: group invariance theorems

- An orbit is fully characterized by the probability density$P_G(gI)$

- An application of Cramer-Wold theorems suggests that that a proxy for $P_G(gI)$ is a set of K one-dimensional $P_G(< gI, t^k >)$

- Since $P_G(< gI, t^k >) = P_G(< I, g^{-1} t^k >)$ it is possible to get an invariant representation from a single image I if all transformations of t^k are stored.

Projections of Probabilities: Cramer-Wold

As argued later, simple operations for neurons are (high-dimensional) dot products between inputs and stored "templates" which are images. It turns out that classical results (such as the Cramer-Wold theorem) ensure that lower dimensional projections of a probability distribution on the unit ball uniquely characterize it.

Theorem *Let P and Q two probability distributions on $\mathbb{R}^d$. Let $\Gamma = (t \in \mathbb{S}(\mathbb{R}^d),\ s.t.\ P_t = \langle P, t \rangle = \langle Q, t \rangle = Q_t)$, where $\mathbb{S}(\mathbb{R}^d)$ is the unit ball in $\mathbb{R}^d$. Let $\lambda(\Gamma)$ its normalized measure. We have that if $\lambda(\Gamma) > 0$ then $P = Q$. This implies that the probability of choosing t such that $P_t = Q_t$ is equal to 1 if and only if $P = Q$ and the probability of choosing t such that $P_t = Q_t$ is equal to 0 if and only if $P \neq Q$.*

Invariant projections theorem

Consider

$$d(P_I, P_{I'}) = \int d_0(P_{\langle I,t\rangle}, P_{\langle I',t\rangle})d\lambda(t), \quad \forall I, I' \in \mathcal{X},$$

$$d(P_I, P_{I'}) \approx \int d_\mu(\mu^t(I), \mu^t(I'))d\lambda(t), \quad \forall I, I' \in \mathcal{X},$$

where d_μ is a metric on histograms induced by d_0.

$$d_\mu(\mu^k(I), \mu^k(I')) = \left\|\mu^k(I) - \mu^k(I)\right\|_{\mathbb{R}^N}$$

where $\|\cdot\|_{\mathbb{R}^N}$ is the Euclidean norm in $\mathbb{R}^N$

Theorem *Consider n images $\mathcal{X}_n$ in $\mathcal{X}$. Let $K \geq \frac{c}{\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant. Then*

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| \leq \epsilon,$$

with probability $1 - \delta^2$, for all $I, I' \in \mathcal{X}_n$.

$$I \sim I' \iff O_I = O_{I'} \iff P_I = P_{I'}.$$

The orbit is invariant and unique

$$\begin{tikzcd}
P_I \arrow[r, leftrightarrow, "\text{Cramer-Wold Theorem}"'] \arrow[d, leftrightarrow] & P_{\langle I, t^k \rangle} \arrow[d, leftrightarrow] \\
gI & \langle gI, t^k \rangle = \langle I, \underbrace{g^{-1} t^k}_{\downarrow} \rangle
\end{tikzcd}$$

Invariant and unique 1D distribution from a SINGLE image

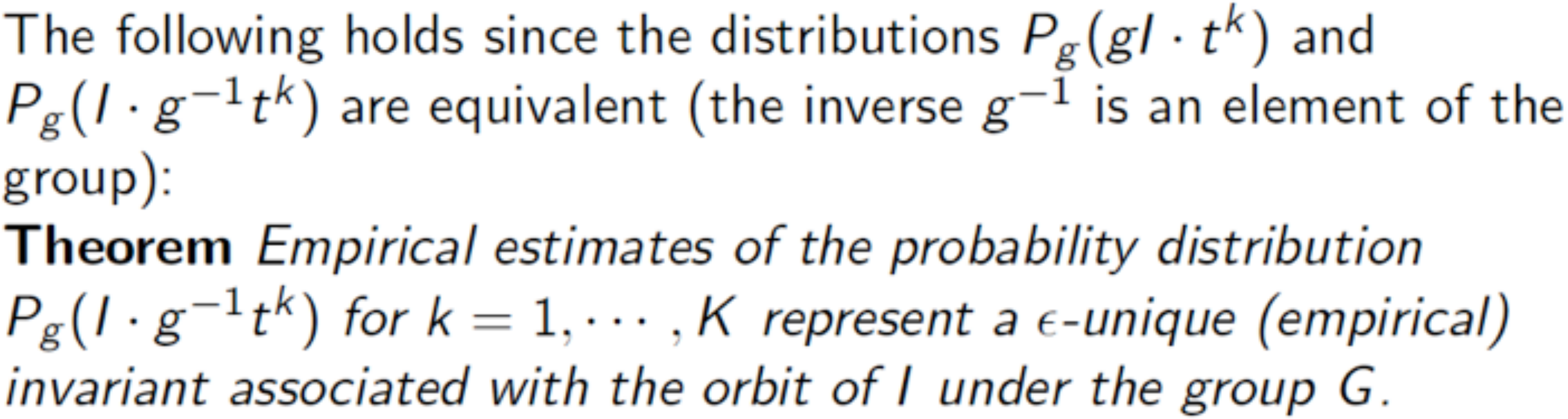
$t^k, \ k = 1, ..., K$ are a set of templates.

This “movie”
is stored during development: unsupervised learning

Group Invariance

The estimation of $P(gI \cdot t^k)$ seems to require the observation of the image *and* "all" its transforms. Ideally we would like to compute an invariant signature for a new object seen only once (we can recognize a face at a different distances after just one observation). The key here is the simple observation that $gI \cdot t^k = I \cdot g^{-1}t^k$. Thus it is possible for the system to store for each template t^k all its transformations gt^k and thus later obtain an invariant signature for new images.

Group Invariance



The following holds since the distributions $P_g(gI \cdot t^k)$ and $P_g(I \cdot g^{-1}t^k)$ are equivalent (the inverse g^{-1} is an element of the group):

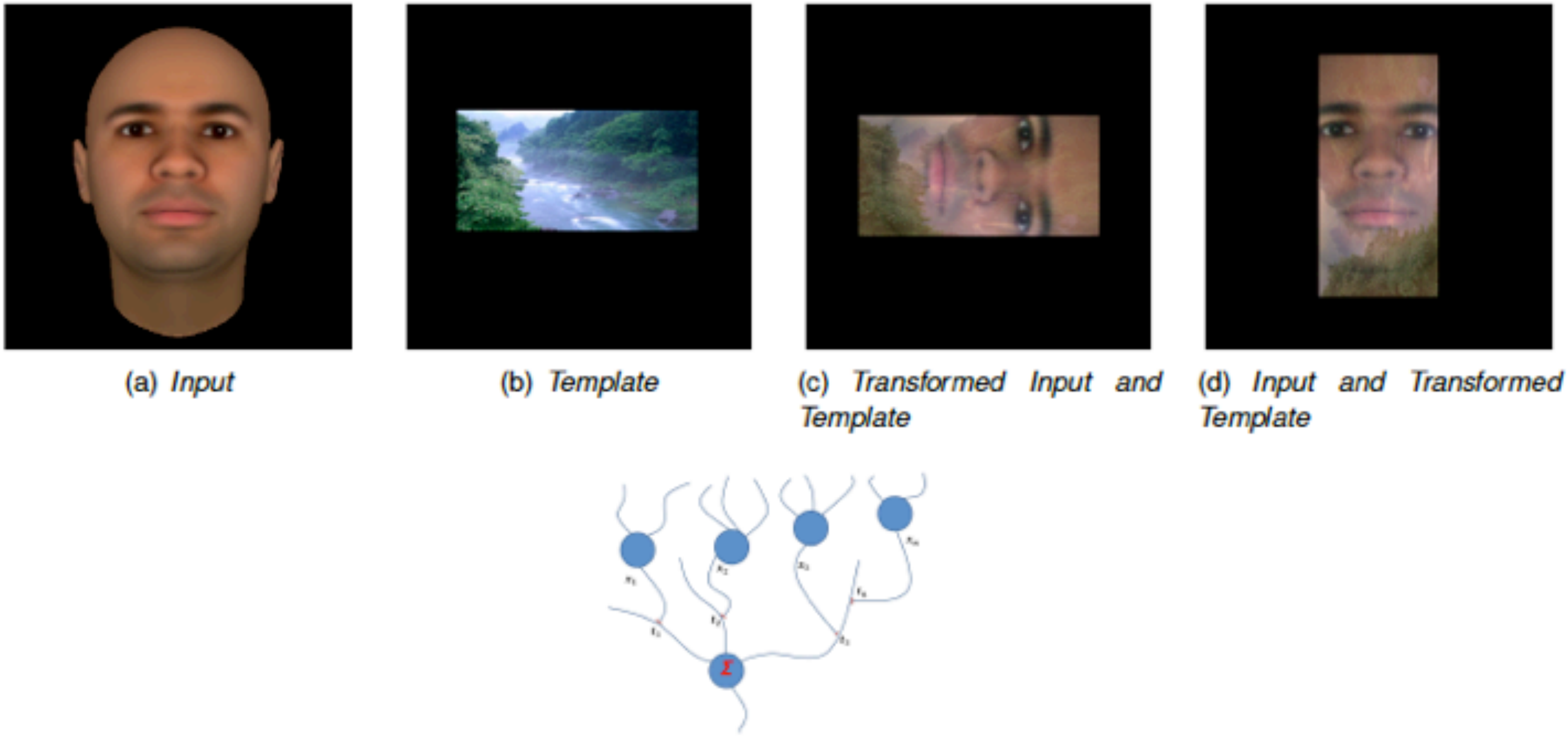
Theorem *Empirical estimates of the probability distribution $P_g(I \cdot g^{-1}t^k)$ for $k = 1, \cdots, K$ represent a ϵ-unique (empirical) invariant associated with the orbit of I under the group G.*

Group Invariance: summary

- The full $P(gI)$ is a probability density induced by "all" $g \in G$; not surprisingly it is a full and invariant characterization of I and all its transforms.
- The Cramer Wold-like theorems say that a proxy for $P(gI)$ is a set of K one dimensional $P(gI \cdot t^k)$. This still requires observation of all the transformations of I induced by the group.
- Since $gI \cdot t^k = I \cdot g^{-1} t^k$ it is however possible possible to obtain an invariant signature from a single image I by storing for each template t^k all its transformations gt^k.

Neuron’s ways to compute invariance

During development of the visual system a group of $|G|$ (simple) cells store in their synapses an image patch t^k and its transformations $g_1 t^k, ..., g_{|G|} t^k$. This is done for several image patches (templates). Later when an image is presented, the simple cells compute $I \cdot g_i t^k$ for $i = 1, ..., |G|$. Complex cells pool the outputs of the simple cells and compute $\mu_n^k = \Sigma_{i=1}^{|G|} \sigma(I \cdot g_i t^k + n\Delta)$ where σ is a smooth step function ($\sigma(x) = 0$ for $x \leq 0$, $\sigma(x) = 1$ for $x > 0$) and $n = 1, ..., N$.



(a) *Input*

(b) *Template*

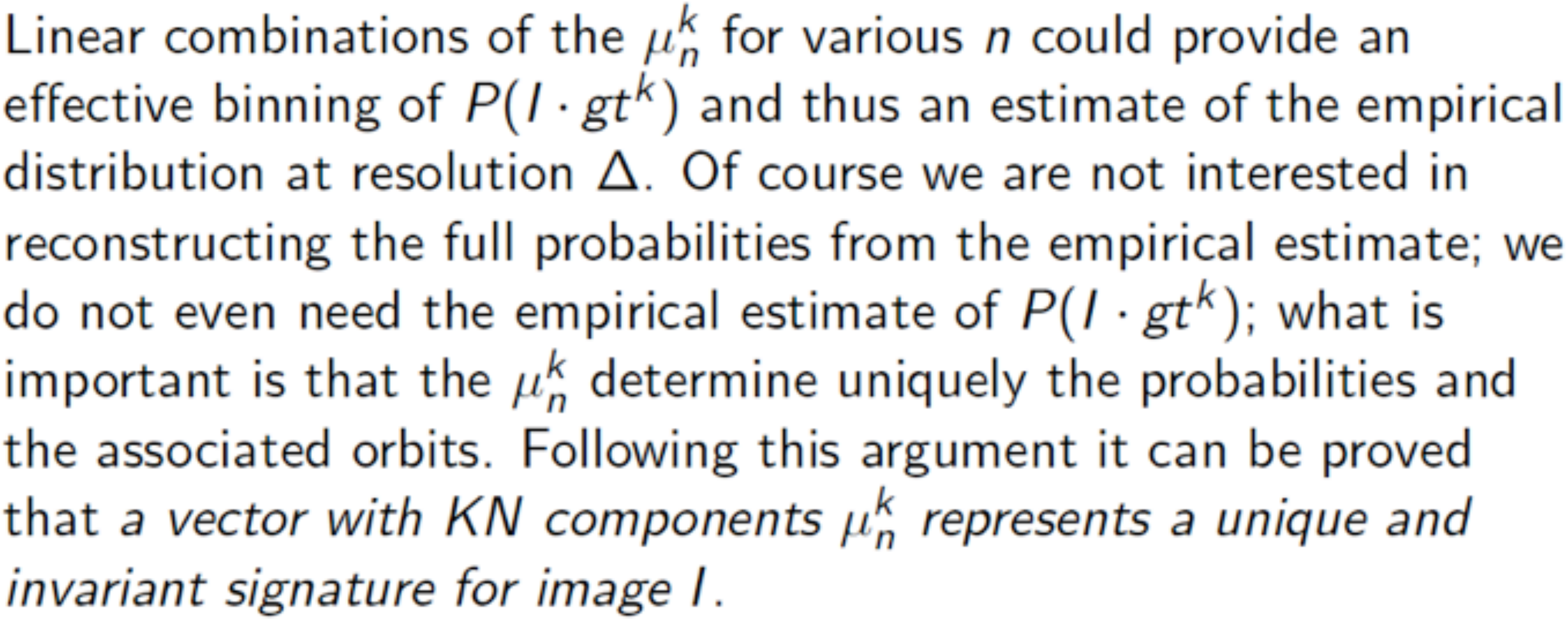
(c) *Transformed Input and Template*

(d) *Input and Transformed Template*

(e) *A neuron's dendritic tree with inputs at its synapses*

Figure 2: *The dot product between a transformed image and a template (c) is equivalent to the dot product between the image with the inversely transformed template (d). Neurons can easily perform high-dimensional dot products between inputs on their dendritic tree and stored synapses weights (indicated in (d)).*

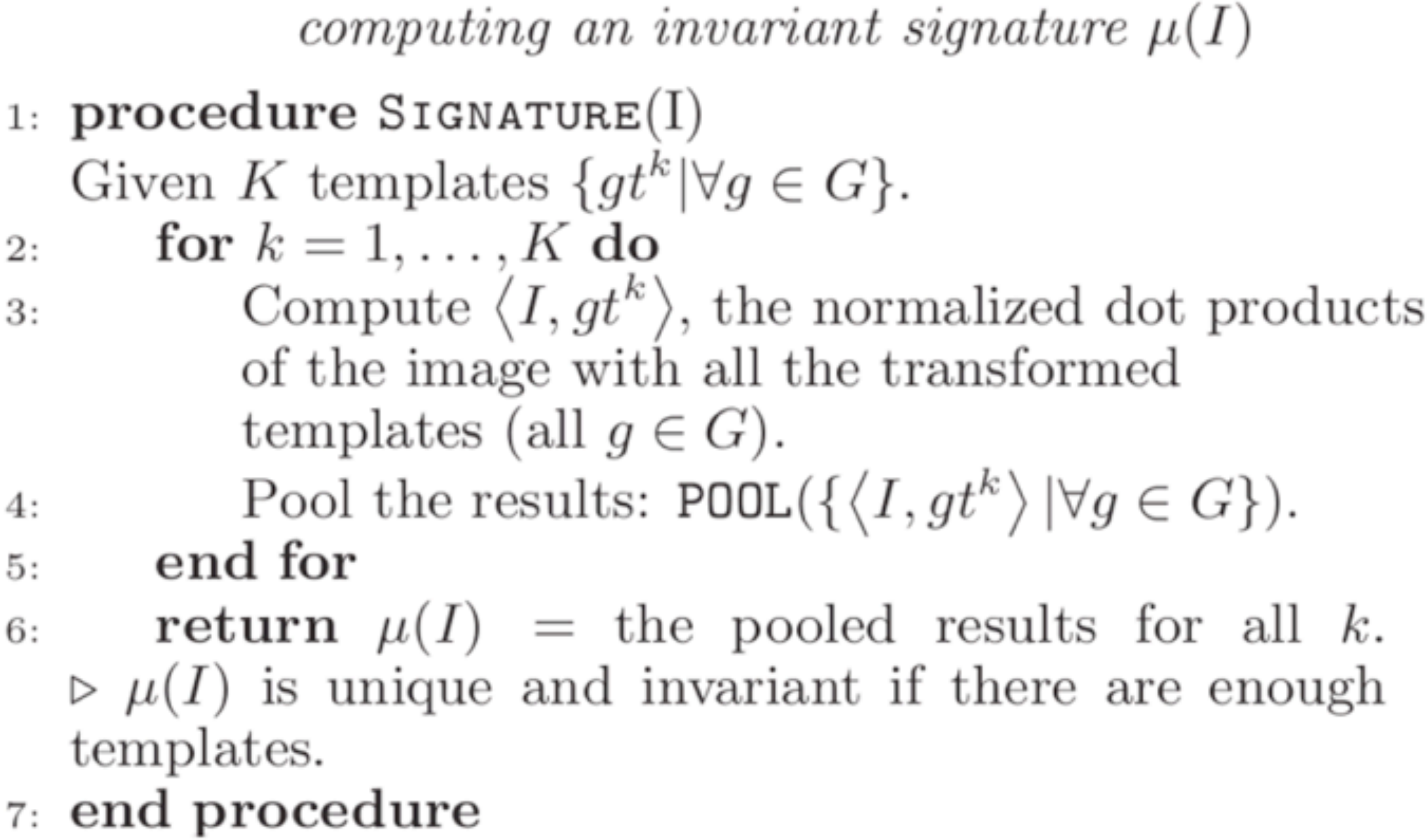
Neural signature: invariance and *uniqueness*



Linear combinations of the μ_n^k for various n could provide an effective binning of $P(I \cdot gt^k)$ and thus an estimate of the empirical distribution at resolution Δ. Of course we are not interested in reconstructing the full probabilities from the empirical estimate; we do not even need the empirical estimate of $P(I \cdot gt^k)$; what is important is that the μ_n^k determine uniquely the probabilities and the associated orbits. Following this argument it can be proved that *a vector with KN components μ_n^k represents a unique and invariant signature for image I.*

Neural signature: energy model

An invariant signature can be computed in other, equivalent ways at the level of complex cells. Instead of the μ_n^k components, the moments $m_n^k = \int_G (I \cdot g_i t^k)^n dg$ can be computed (they characterize the projections of the probability distributions and can be regarded as group averages. Under some rather weak conditions, they characterize uniquely the distribution $P(I \cdot t)$. For $n = 2$ this corresponds to an energy model of complex cells; for very large n it corresponds to a *max* operation by complex cells. Other nonlinearities are also possible. The available evidence suggests that simple/complex cells in V1 and cells in AL may be described better in terms of energy models than in terms of the sigmoidal nonlinearity.



computing an invariant signature $\mu(I)$

```
procedure SIGNATURE(I)
Given $K$ templates $\{gt^k | \forall g \in G\}$.
    for $k = 1, \dots, K$ do
        Compute $\langle I, gt^k \rangle$, the normalized dot products
        of the image with all the transformed
        templates (all $g \in G$).
        Pool the results: POOL($\{\langle I, gt^k \rangle | \forall g \in G\}$).
    end for
    return $\mu(I)$ = the pooled results for all $k$.
▷ $\mu(I)$ is unique and invariant if there are enough
templates.
end procedure
```

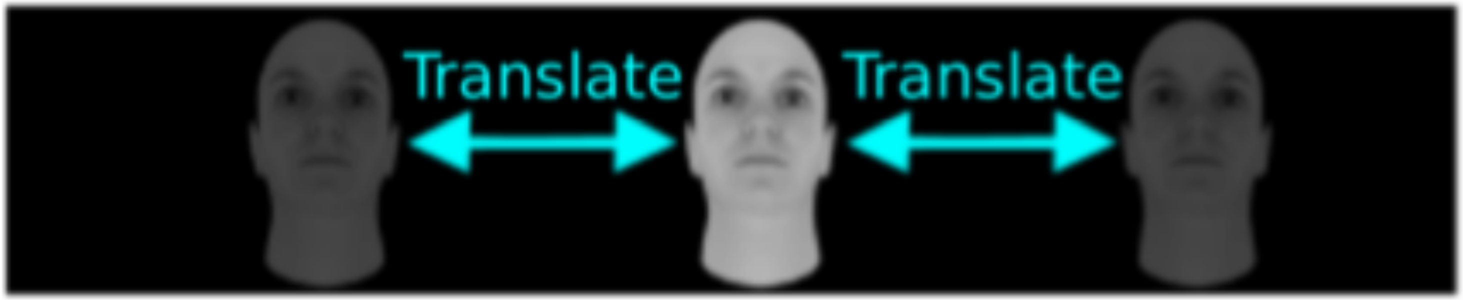
The basic magic module

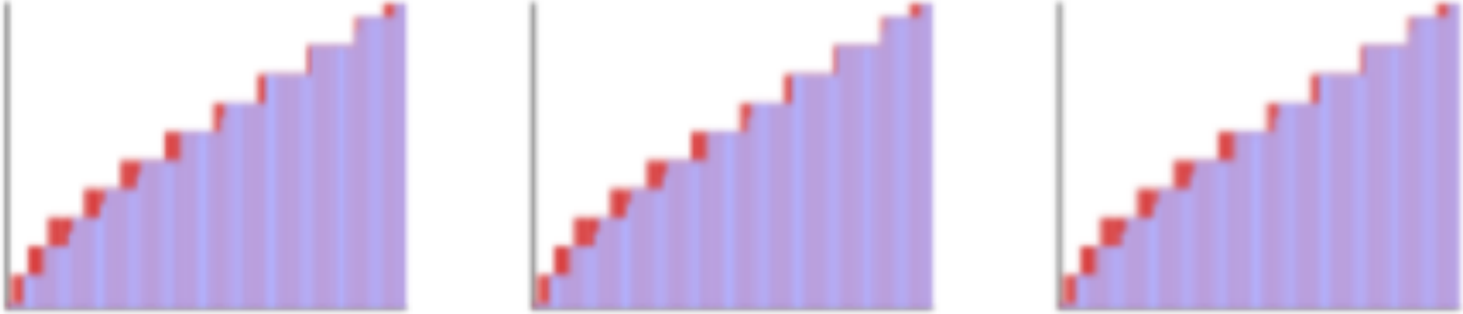
So far (for simplicity): compact groups in R^2

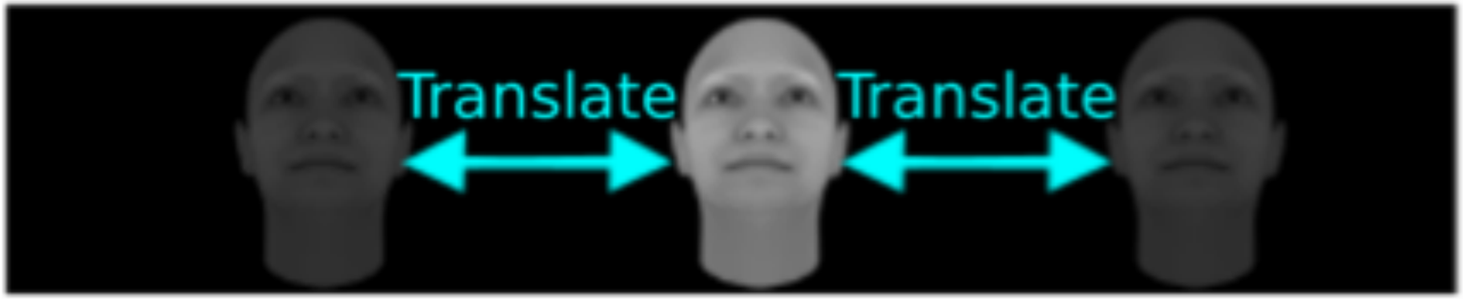
M-theory extend result to

- **non compact groups**
- hierarchies of magic modules (multilayer)
- non-group transformations

Translation







J. Leibo, Q. Liao

Non compact groups

We assume that the dot products is “normalized": the signals x and t are zero-mean and norm = 1. Thus starting with x”, t”

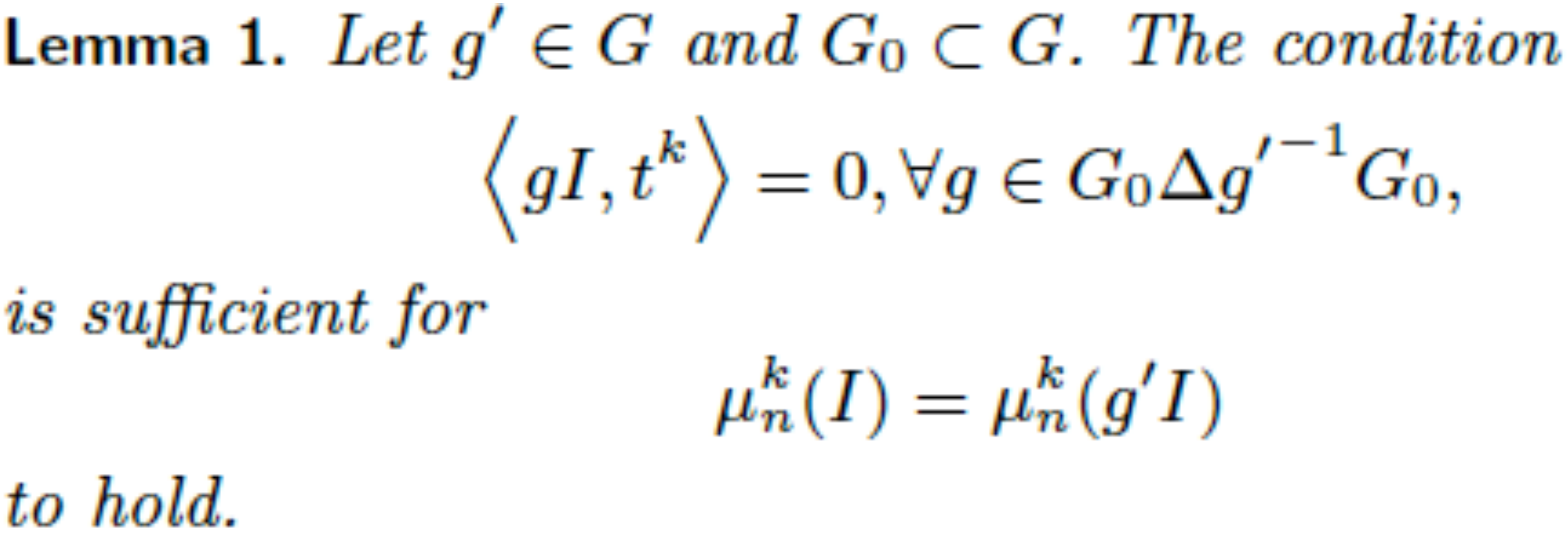
$$x' = x'' - E(x''), \qquad x = \frac{x'}{|x'|};$$

$$\mathrm{t}' = t'' - E(t''), \qquad \mathrm{t} = \frac{t'}{|t'|}$$

We assume that the empty surround of an isolated image patch has value 0, being equal to the average value over the ensemble of images. In particular the dot product of a template and the region outside an isolated image patch is 0.

Partially Observable Groups

For a transformation observed via a “receptive field” there is only “partial invariance”



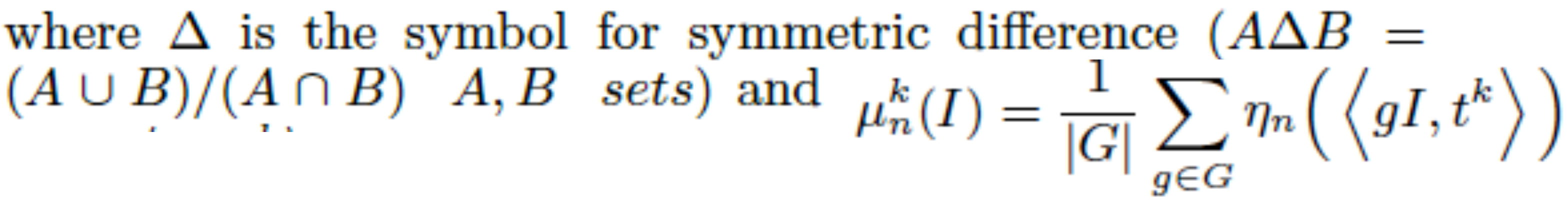
Lemma 1. *Let $g' \in G$ and $G_0 \subset G$. The condition*

$$\left\langle gI, t^k \right\rangle = 0, \forall g \in G_0 \Delta {g'}^{-1} G_0,$$

is sufficient for

$$\mu_n^k(I) = \mu_n^k(g'I)$$

to hold.



where Δ is the symbol for symmetric difference ($A \Delta B = (A \cup B)/(A \cap B)$ A, B *sets*) and $\mu_n^k(I) = \frac{1}{|G|} \sum_{g \in G} \eta_n \Big(\left\langle gI, t^k \right\rangle \Big)$

Partially Observable Groups

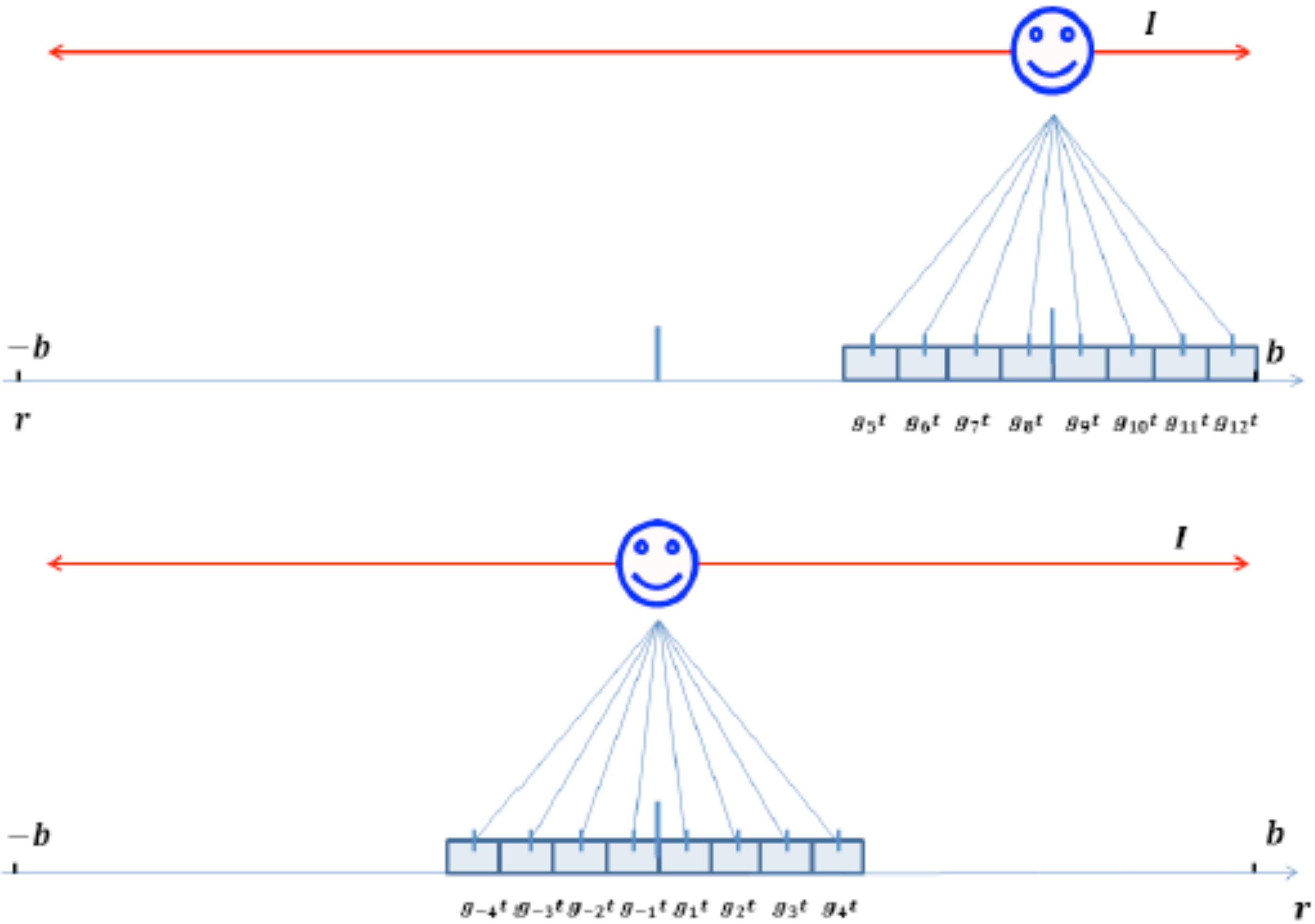
Invariance for POGs implies a property that can be called *localization* or

- *sparsity* of I wrt the dictionary of the t

Example: consider the case of a 1D parameter translation group: invariance within the pooling region $[-b,b]$ is ensured for

$$\boxed{\begin{array}{l} -b+a \leq r \leq b-a \\ \text{if } < I, g_r t >= 0 \quad \text{for } r > a \end{array}}$$

Partially Observable Groups


I
$-b$
b
r
$g_5{}^t$ $g_6{}^t$ $g_7{}^t$ $g_8{}^t$ $g_9{}^t$ $g_{10}{}^t$ $g_{11}{}^t$ $g_{12}{}^t$
I
$-b$
b
$g_{-4}{}^t$ $g_{-3}{}^t$ $g_{-2}{}^t$ $g_{-1}{}^t$ $g_1{}^t$ $g_2{}^t$ $g_3{}^t$ $g_4{}^t$
r

Invariance, localization, wavelets

Localization/sparsity implies, and is implied by, invariance.
Localization can be satisfied in two different regimes:

- *exact* localization for *generic* images holds for affine group: expected for the first layers, yields Gabor wavelets

- *approximate* sparsity of a subclass of I wrt dictionary of templates t^k holds locally for any smooth transformation: expected for highest layers, yields very specific decoherent tunings

Theorem:
optimal x,s invariance implies Gabor wavelets
(in the generic regime)

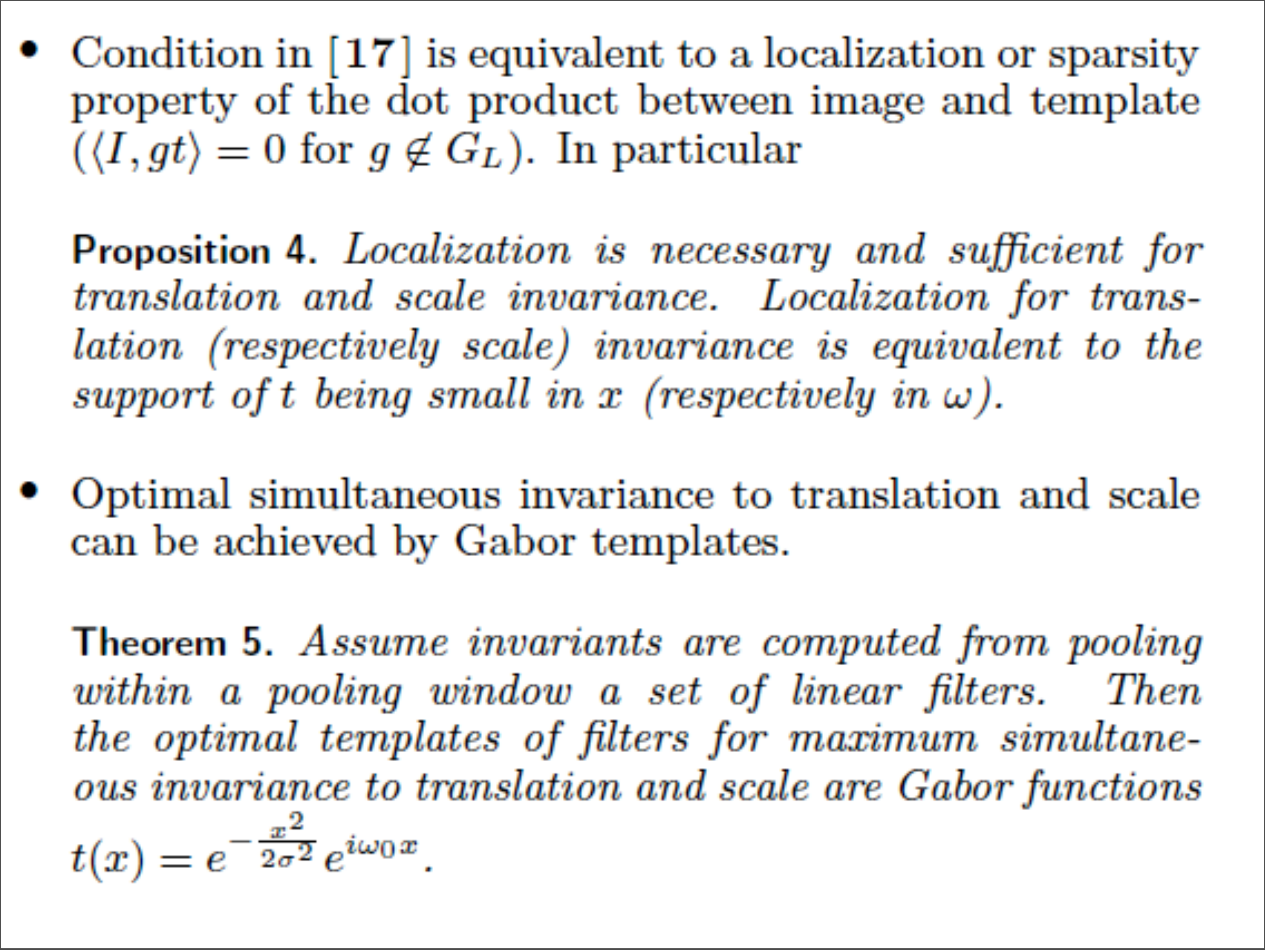
“

Invariance (for scale+translation) ⇔

⇔ localization in x and ω ⇔ wavelets

Maximum joint (x, s) invariance ⇔ Gabor-like wavelets

Full information ≈ frame of Gabor-like wavelets



- Condition in [17] is equivalent to a localization or sparsity property of the dot product between image and template ($\langle I, gt\rangle = 0$ for $g \notin G_L$). In particular

  Proposition 4. *Localization is necessary and sufficient for translation and scale invariance. Localization for translation (respectively scale) invariance is equivalent to the support of t being small in x (respectively in ω).*

- Optimal simultaneous invariance to translation and scale can be achieved by Gabor templates.

  Theorem 5. *Assume invariants are computed from pooling within a pooling window a set of linear filters. Then the optimal templates of filters for maximum simultaneous invariance to translation and scale are Gabor functions $t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}$.*

Class-specific regime:

sparsity of subclass of images I_C wrt to templates t_k under the group G

$$\text{Invariance} \Leftrightarrow \text{sparsity} \Leftrightarrow \text{complex, templates with delta-like autocorr}$$

non-group transformations

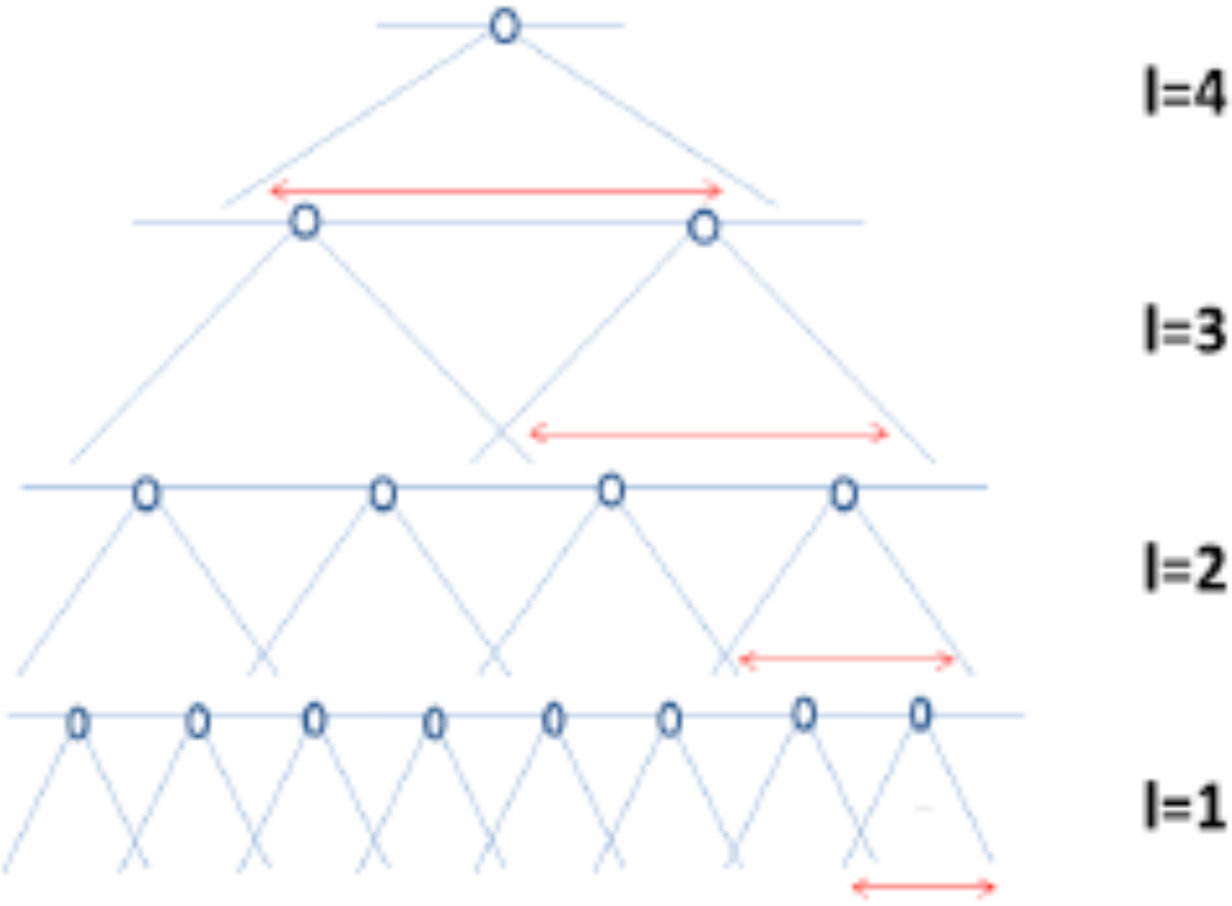
The basic magic module

So far (for simplicity): compact groups in R^2

M-theory extend result to

- non compact groups
- **hierarchies of magic modules (multilayer)**
- non-group transformations

Multilayer architectures: key property: covariance



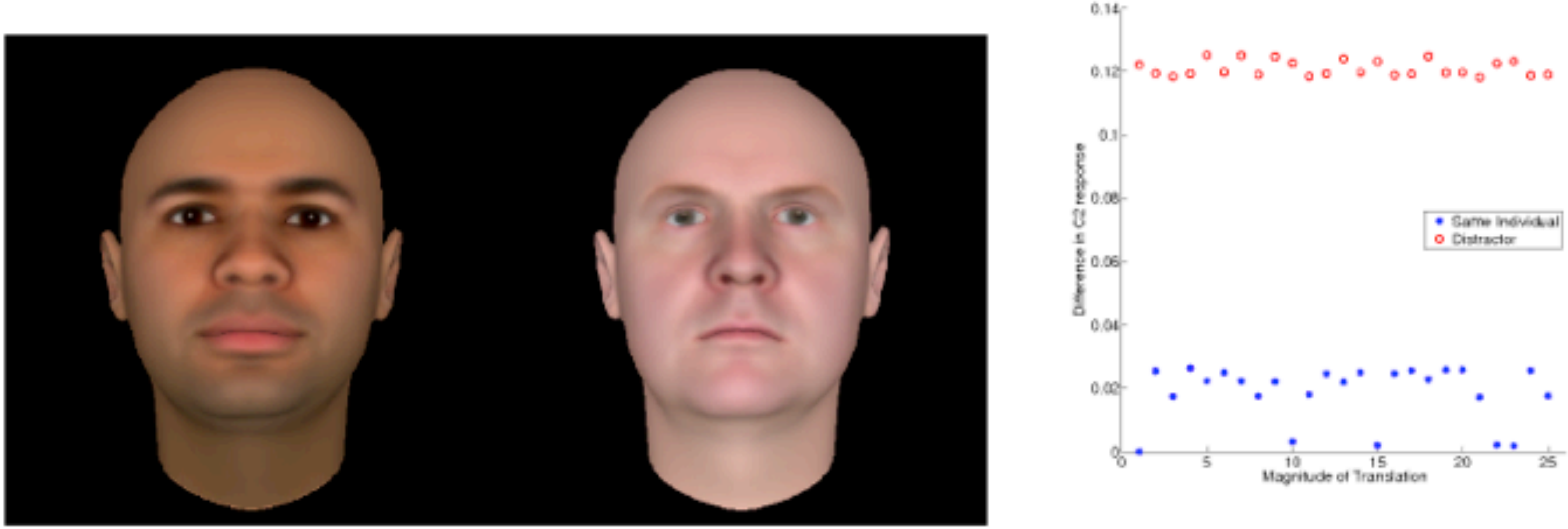
Why multilayer architectures

- **Compositionality**

- **Factorization of invariance ranges**

- **Memory access minimizing clutter effects**

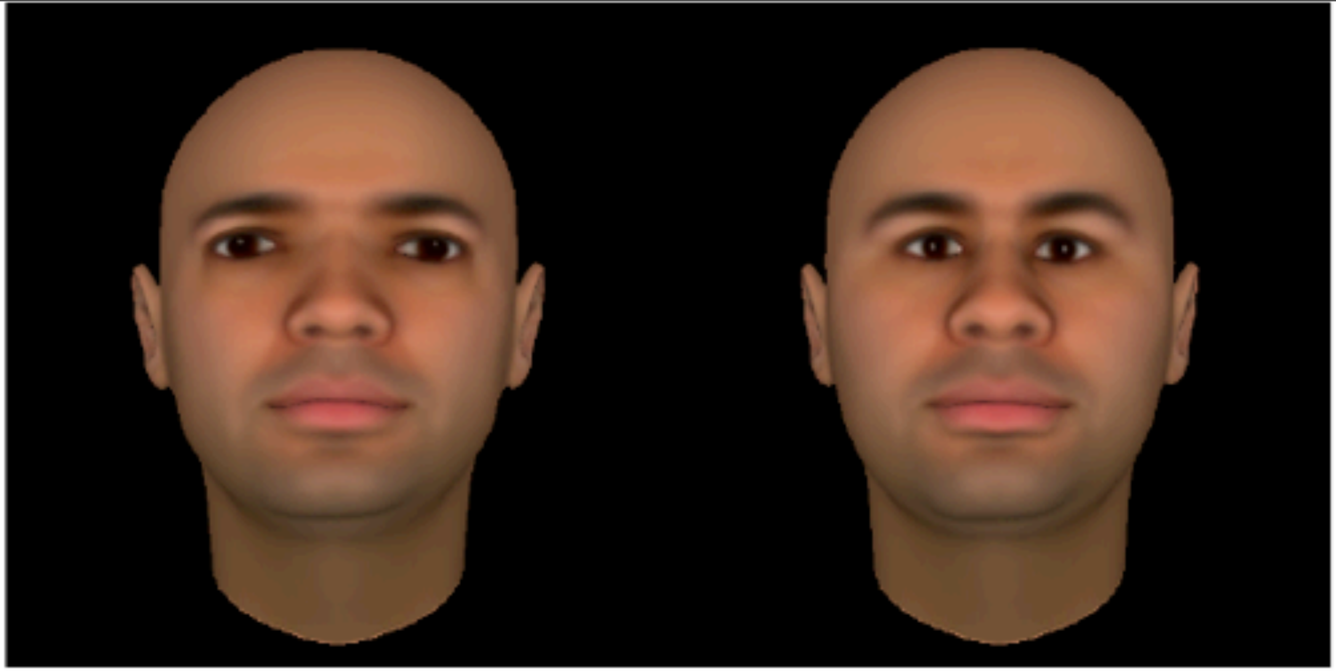
- **Optimization of local connections**



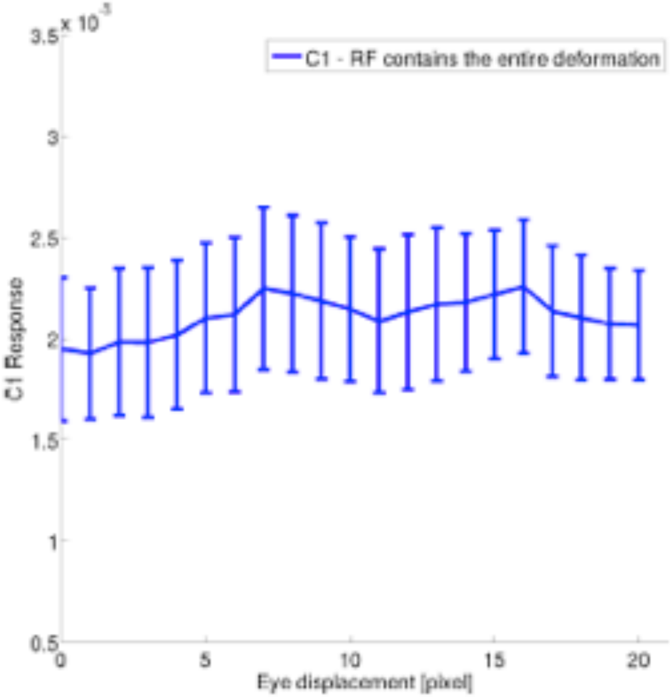
(a) *Reference input and distractor.*

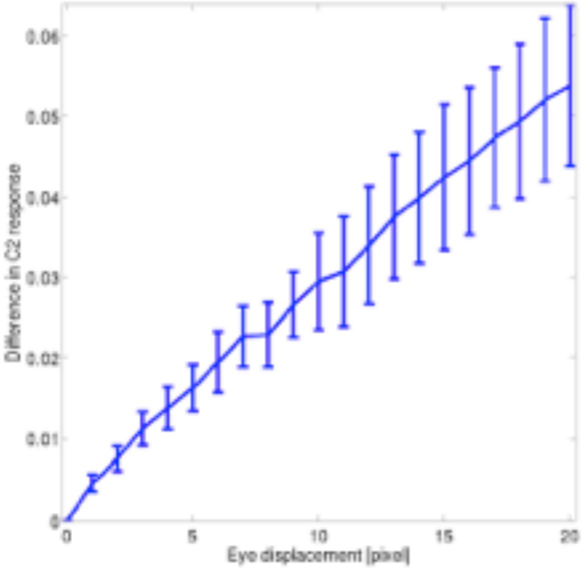
(b)

Figure 3: *Two distinct stimuli (left) are presented at various location in the visual field. The Euclidean distance between C2 response vectors in HMAX is reported (right). It can be seen how the response are invariant to global translation and discriminative. The C2 units represent the top of a hierarchical, convolutional architecture.*

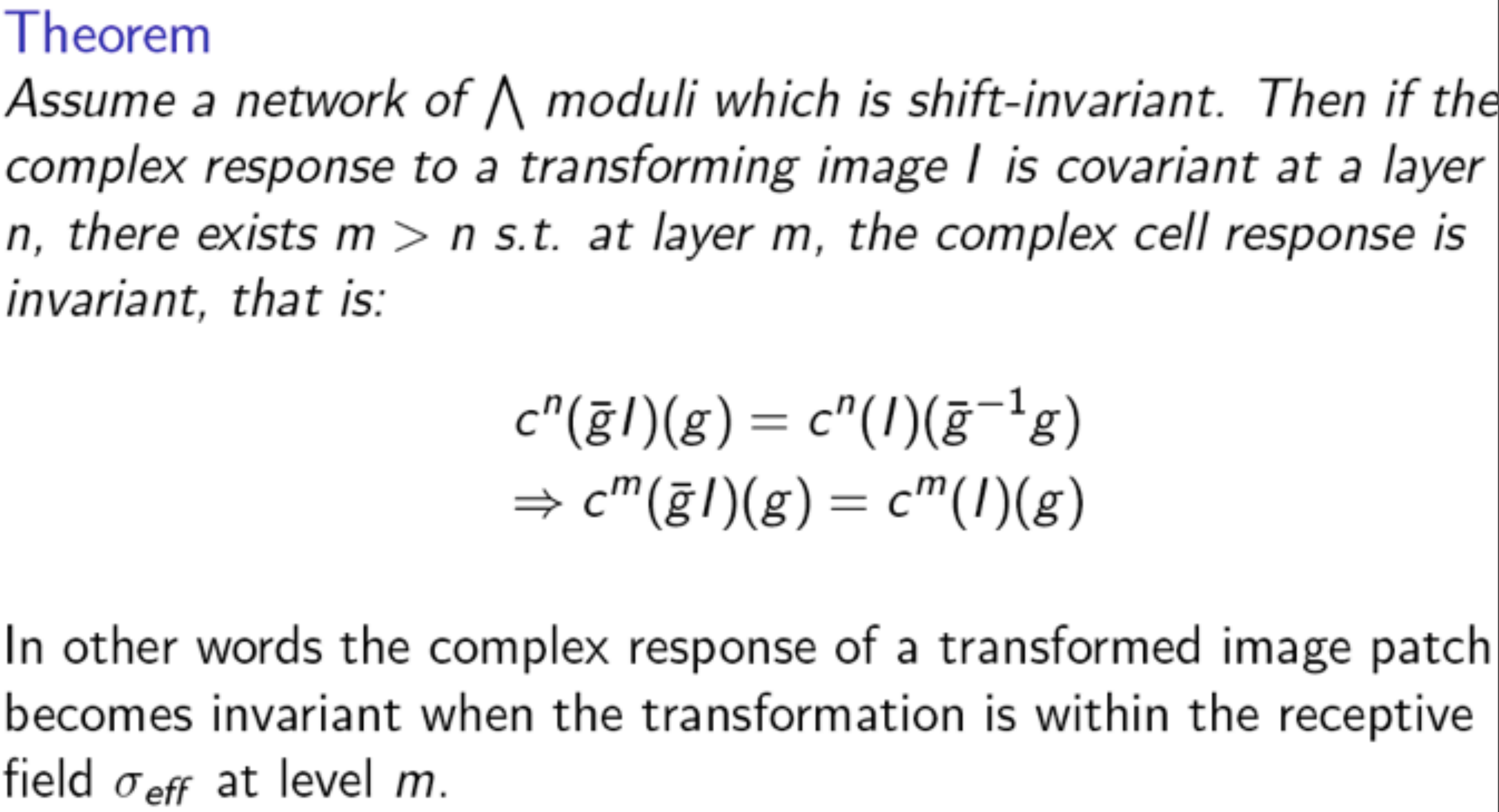


(a)


x 10^-3
3.5
3
2.5
2
1.5
1
0.5
C1 - RF contains the entire deformation
C1 Response
0
5
10
15
20
Eye displacement [pixel]


0.06
0.05
0.04
0.03
0.02
0.01
0
Difference in C2 response
0
5
10
15
20
Eye displacement [pixel]

Local and global invariance: Whole-Parts theorem



Theorem

Assume a network of $\bigwedge$ moduli which is shift-invariant. Then if the complex response to a transforming image I is covariant at a layer n, there exists $m > n$ s.t. at layer m, the complex cell response is invariant, that is:

$$c^n(\bar{g}I)(g) = c^n(I)(\bar{g}^{-1}g)$$
$$\Rightarrow c^m(\bar{g}I)(g) = c^m(I)(g)$$

In other words the complex response of a transformed image patch becomes invariant when the transformation is within the receptive field σ_{eff} at level m.

The basic magic module

For simplicity here: compact groups in R^2

M-theorems extend result to

- non compact groups
- hierarchies of magic modules (multilayer)
- **class-specific, non-group transformations**

- Approximate invariance can be obtained if there is approximate sparsity of the image in the dictionary of templates.

  Proposition 6. *Approximate localization (defined as $\langle t, gt\rangle < \delta$ for $g \notin G_L$, where δ is small in the order of $\approx \frac{1}{\sqrt{d}}$ and $\langle t, gt\rangle \approx 1$ for $g \in G_L$) is satisfied by templates (vectors of dimensionality d) that are similar to images in the set and are sufficiently "large" to be incoherent for "small" transformations.*

- Approximate invariance for smooth (non group) transformations.

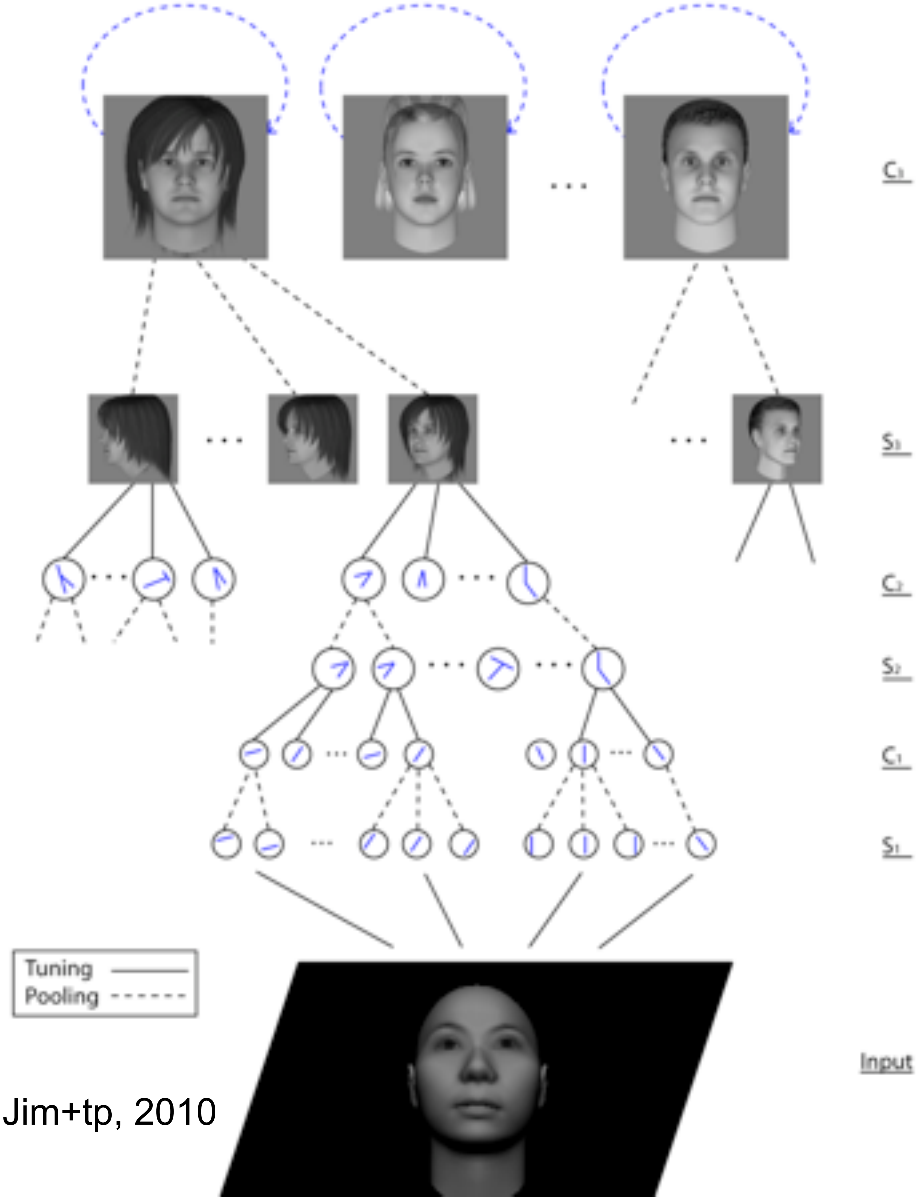
  Proposition 7. $\mu^k(I)$ *is locally invariant* **if**

  – *I is sparse in the dictionary t^k;*

  – *I and t^k transform in the same way (belong to the same class);*

  – *the transformation is sufficiently smooth.*

Pose-invariant face recognition



Joel+Jim+tp, 2010

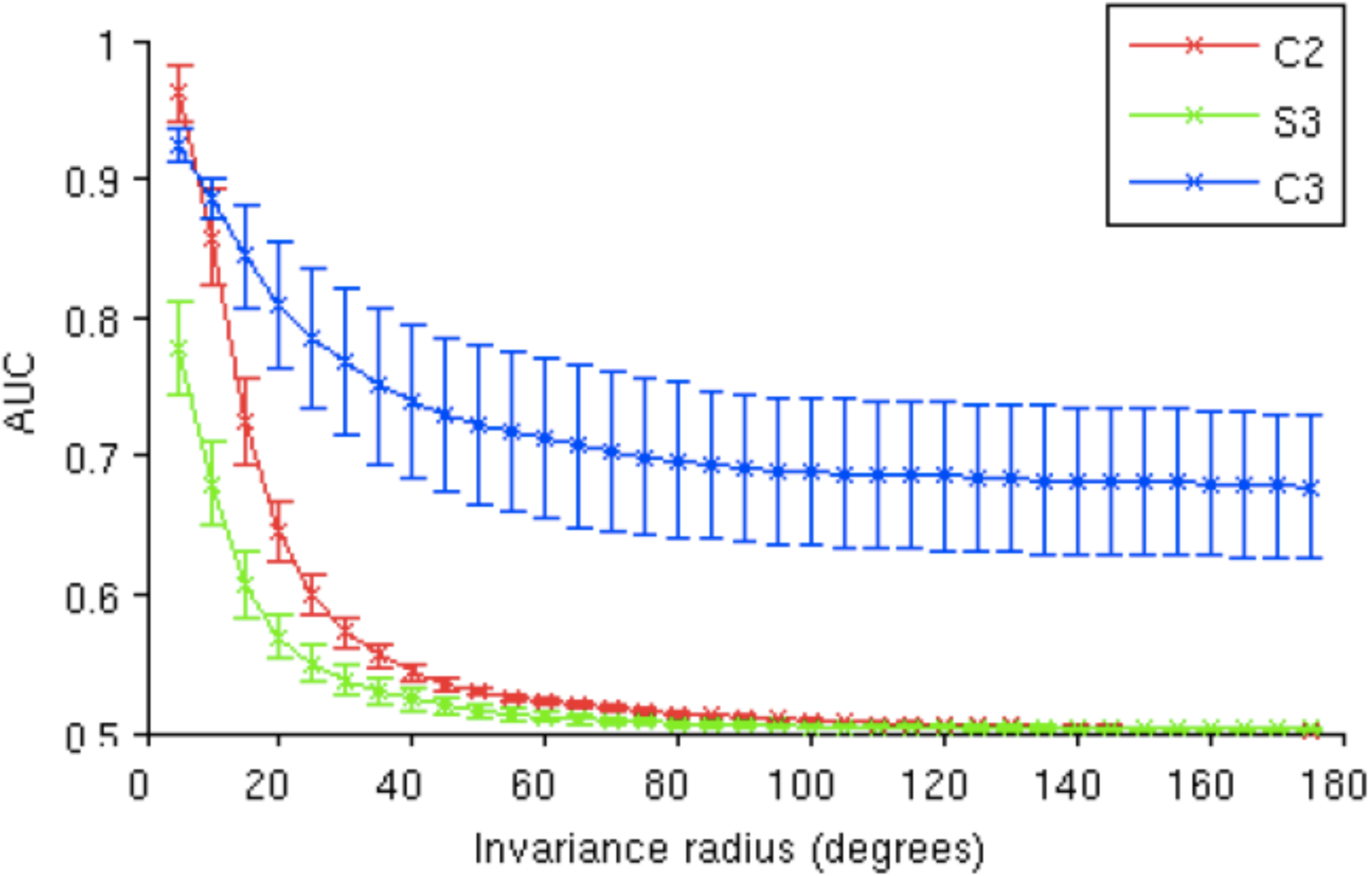
Viewpoint tolerant units
(complex units)

View-tuned units, tuned to full-face
templates for different view angles

HMAX

Tolerance to a transformation may be learned unsupervised

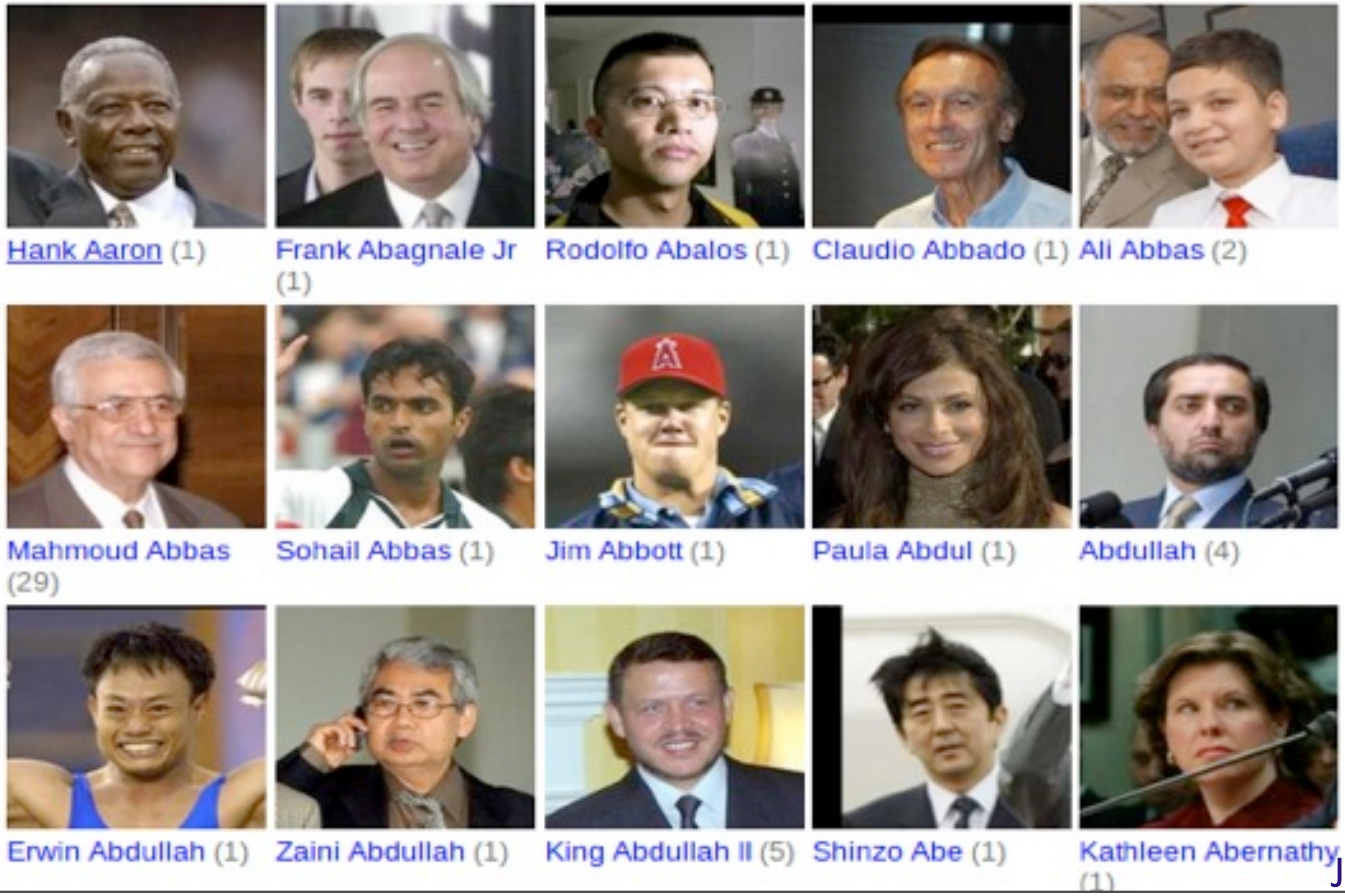
Learning class specific transformations: quasi-invariance to pose for faces



Joel+Jim+tp, NIPS 2012

Labeled Faces in the Wild

Contains 13,233 images of 5,749 people



Hank Aaron (1) Frank Abagnale Jr (1) Rodolfo Abalos (1) Claudio Abbado (1) Ali Abbas (2)

Mahmoud Abbas (29) Sohail Abbas (1) Jim Abbott (1) Paula Abdul (1) Abdullah (4)

Erwin Abdullah (1) Zaini Abdullah (1) King Abdullah II (5) Shinzo Abe (1) Kathleen Abernathy (1)

J. Leibo, Q. Liao

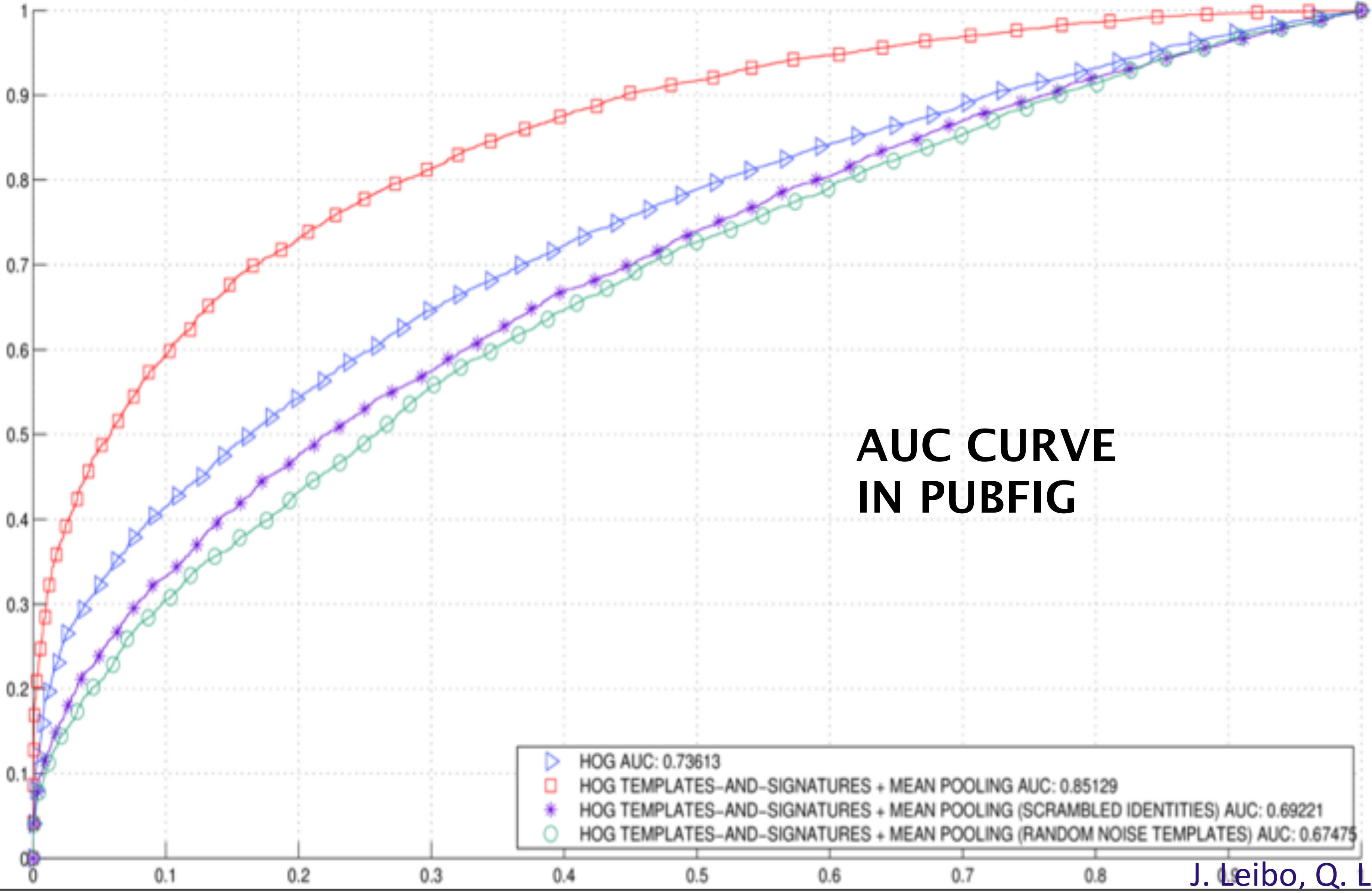
Pubfig

- Originally, 58,797 images of 200 people
- Unfortunately there are only ~21000 left now

J. Leibo, Q. Liao

Does our method work?

Yes.



AUC CURVE IN PUBFIG

J. Leibo, Q. Liao

Performance Summary

- Pubfig State-of-the-art:

78.65% (original training and testing set)

- Our current performance:

HOG based: ~78.3%

LBP based: ~78.5%

LBP + HOG based: ~80.5%

- We did not touch their training data at all.

J. Leibo, Q. Liao

A second phase in Machine Learning

- The first phase -- from ~1980s -- led to a rather complete theory of *supervised learning* and to practical systems (MobilEye, Orcam,...) that need *lots of examples for training*

- The second phase may be about *unsupervised learning of (invariant) representations* that make supervised learning possible with *very few examples*

J. Leibo, Q. Liao

A theory of feedforward vision: will it tell us what cortex computes and properties of its neurons?

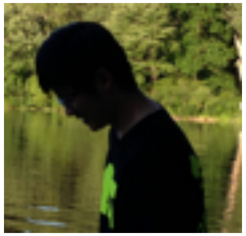
- The basic equation of physics can be derived from a small number of symmetry properties: invariance wrt space+time, conservation of energy, invariance to measurement units....

- Is the architecture and tuning properties of visual (and auditory...) cortex predictable from basic symmetries of geometric transformations of images?

- The brain would be a mirror of the physical world and the tuning of its neurons would reflect symmetry properties of basic physics and geometry.

J. Leibo, Q. Liao

Collaborators in recent work



F. Anselmi, J. Mutch , J. Leibo,   L. Rosasco, A. Tacchetti, Q. Liao

+ +

L. Isik, S. Ullman, S. Smale, C. Tan

Also: M. Riesenhuber, T. Serre, G. Kreiman, S. Chikkerur, A. Wibisono, J. Bouvrie, M. Kouh, J. DiCarlo, E. Miller, C. Cadieu, A. Oliva, C. Koch, A. Caponnetto ,D. Walther,   U. Knoblich, T. Masquelier, S. Bileschi, L. Wolf, E. Connor, D. Ferster, I. Lampl, S. Chikkerur, G., N. Logothetis, H. Buelthoff