i-theory:
visual cortex and deep networks

The Center for Brains, Minds and Machines

tomaso poggio,
CBMM, BCS, CSAIL, McGovern
MIT

Center for Brains,
Minds & Machines

Theoretical/conceptual framework for vision

- The first 100ms of vision: feedforward and invariant: what, who, where
- Top-down needed for verification step and more complex questions: generative models, probabilistic inference, top-down visual routines.

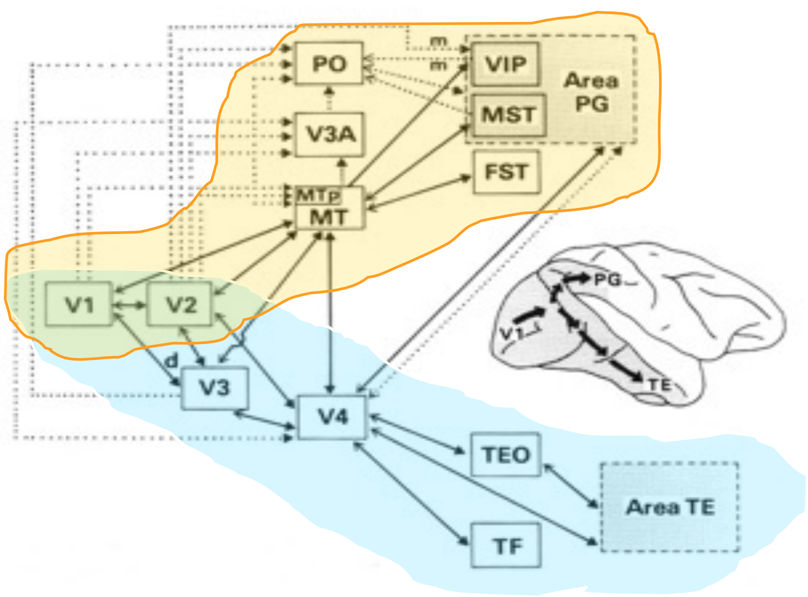
Following this conceptual framework we are working on:

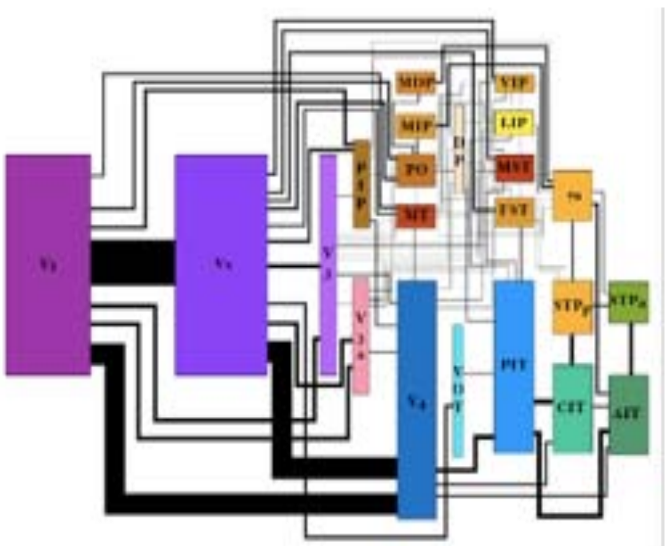
1. *theory of invariance* in feedforward networks (visual cortex)
2. a *generative approach*, probabilistic in nature
3. *visual routines*, and of how they may be learned.

Computational Vision

Object recognition

Figure removed due to copyright restrictions. Please see the video.





Source: Wallisch, Pascal, and J. Anthony Movshon. "Structure and function come unglued in the visual cortex." Neuron 60, no. 2 (2008): 195-197.

Vision: what is where

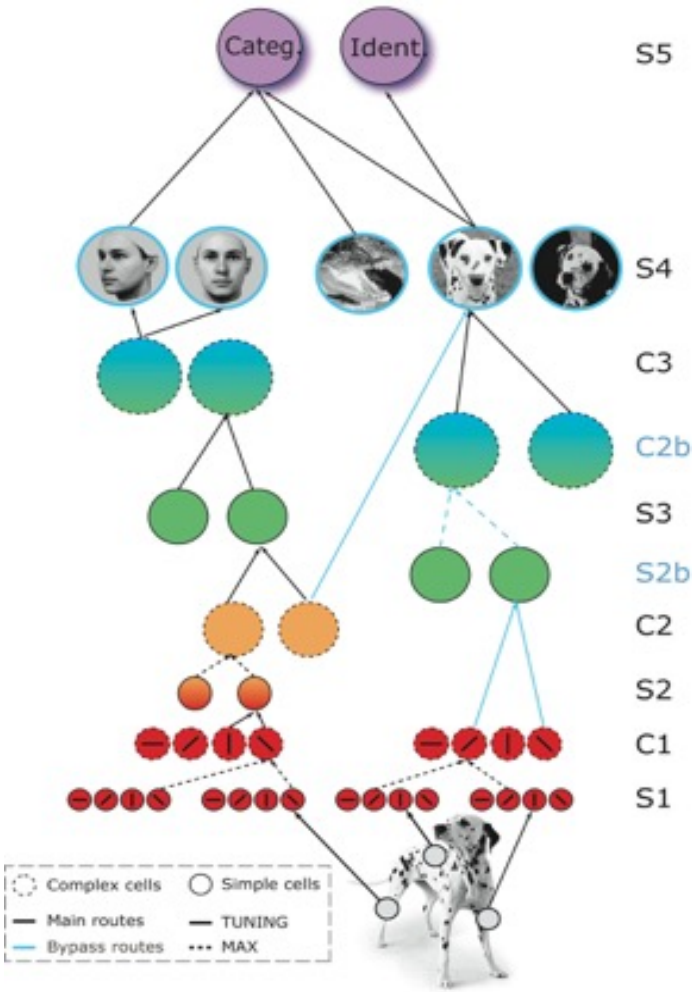
- Human Brain
  - 10^{10}-10^{11} neurons (~1 million flies)
  - 10^{14}- 10^{15} synapses

Figure removed due to copyright restrictions. Please see the video.
Source: Figure 2 from Felleman, Daniel J., and David C. Van Essen. "Distributed hierarchical processing in the primate cerebral cortex." Cerebral cortex 1, no. 1 (1991): 1-47.

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

- ~200M in V1, ~200M in V2, 50M in V4

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



Source: Serre, Thomas, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. No. AI MEMO-2005-036. Massachusetts Institute of Technology Center for Biological and Computational Learning, 2005.

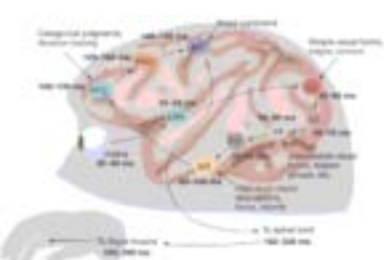
[software available online]

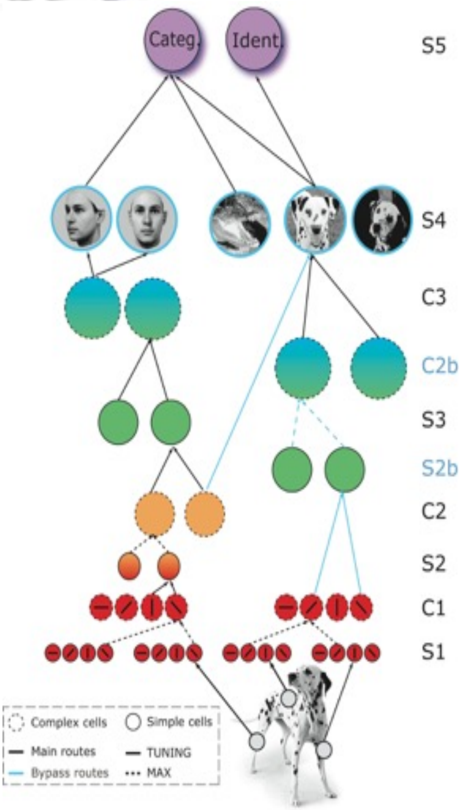
- It is in the family of "Hubel-Wiesel" models (Hubel & Wiesel, 1959: *qual.* Fukushima, 1980: *quant*; Oram & Perrett, 1993: *qual*; Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe, 2002; Ullman et al., 2002; Mel, 1997; Wersing and Koerner, 2003; LeCun et al 1998: *not-bio*; Amit & Mascaro,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

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

Hierarchical feedforward models of the ventral stream





Source: Serre, Thomas, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. No. AI MEMO-2005-036. Massachusetts Institute of Technology Center for Biological and Computational Learning, 2005.

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

Figure removed due to copyright restrictions. Please see the video.

Why do these networks including DLCNs work so well?

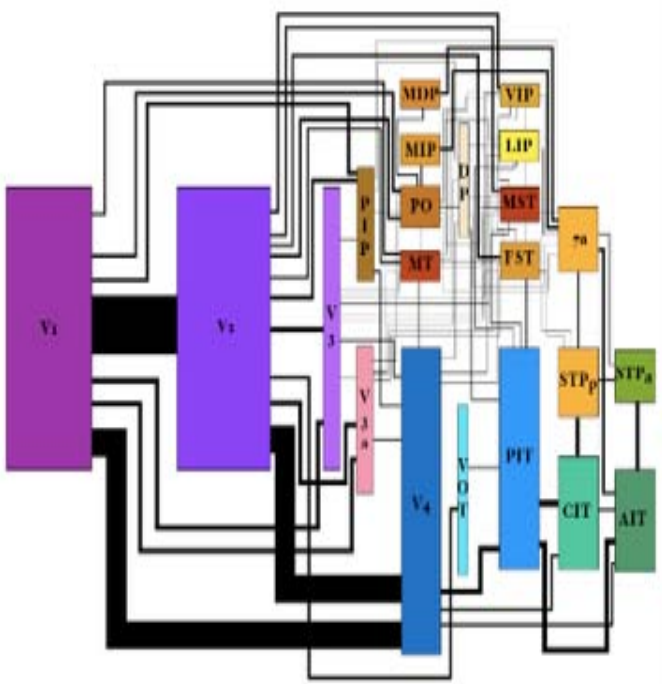
Models are not enough…
we need a theory!

Plan

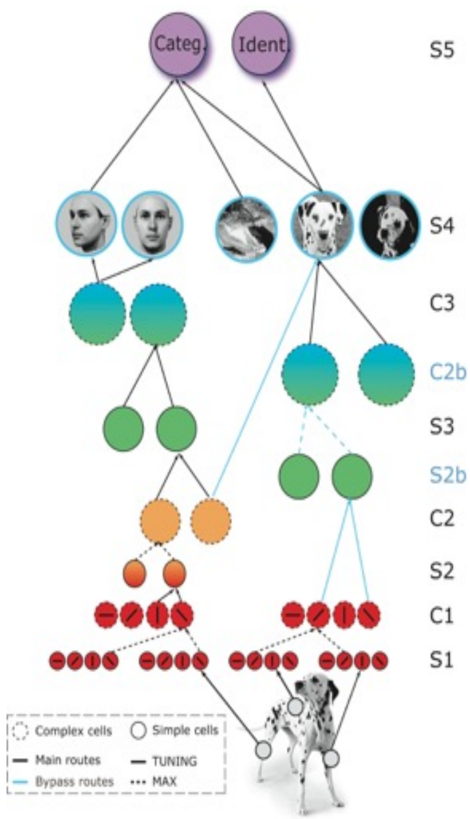
- i-theory (main results)
- equivalence to DCLNs, theory notes on DCLNs
- Some predictions + perspectives in i-theory
- Details and ML remarks

i-theory

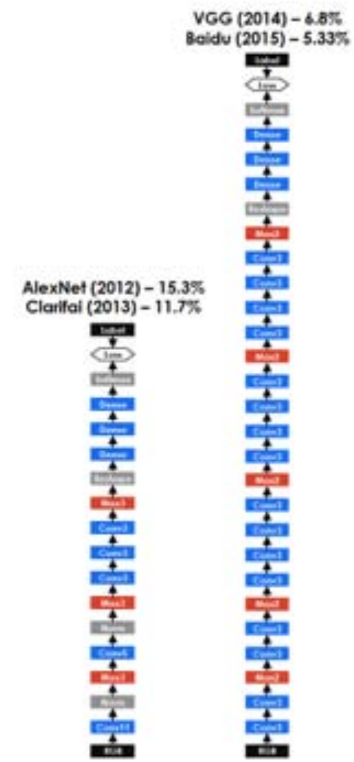
Learning of *invariant&selective* Representations in Sensory Cortex



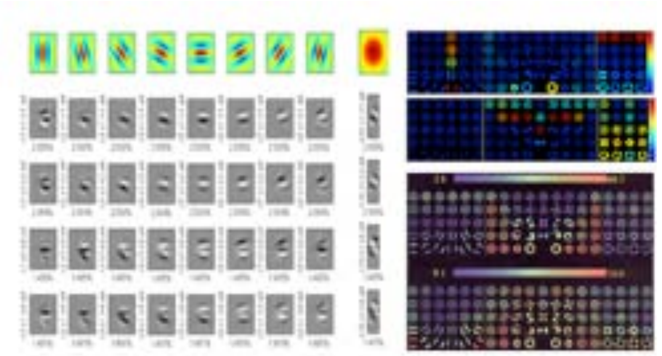
Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Wallisch, Pascal, and J. Anthony Movshon. "Structure and function come unglued in the visual cortex." Neuron 60, no. 2 (2008): 195-197.



Source: Serre, Thomas, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. No. AI MEMO-2005-036. Massachusetts Institute of Technology Center for Biological and Computational Learning, 2005.



What i-theory can answer for you



Courtesy of Tomaso Poggio, Jim Mutch, Fabio Anselmi, Andrea Tacchetti, Lorenzo Rosasco and Joel Leibo. Used with permission.
Source: Poggio, Tomaso, Jim Mutch, Fabio Anselmi, Andrea Tacchetti, Lorenzo Rosasco, and Joel Z. Leibo. "Does invariant recognition predict tuning of neurons in sensory cortex?" (2013).

- why some hierarchical nets work well
- what is visual cortex computing?
- function and circuits of simple-complex cells
- why Gabor-like tuning in simple cells?
- why generic, Gabor-like tuning in early areas and specific selective tuning higher up?
- what is the computational reason for the eccentricity-dependent size of RFs in V1, V2, V4?
- what are the roles of back projections?

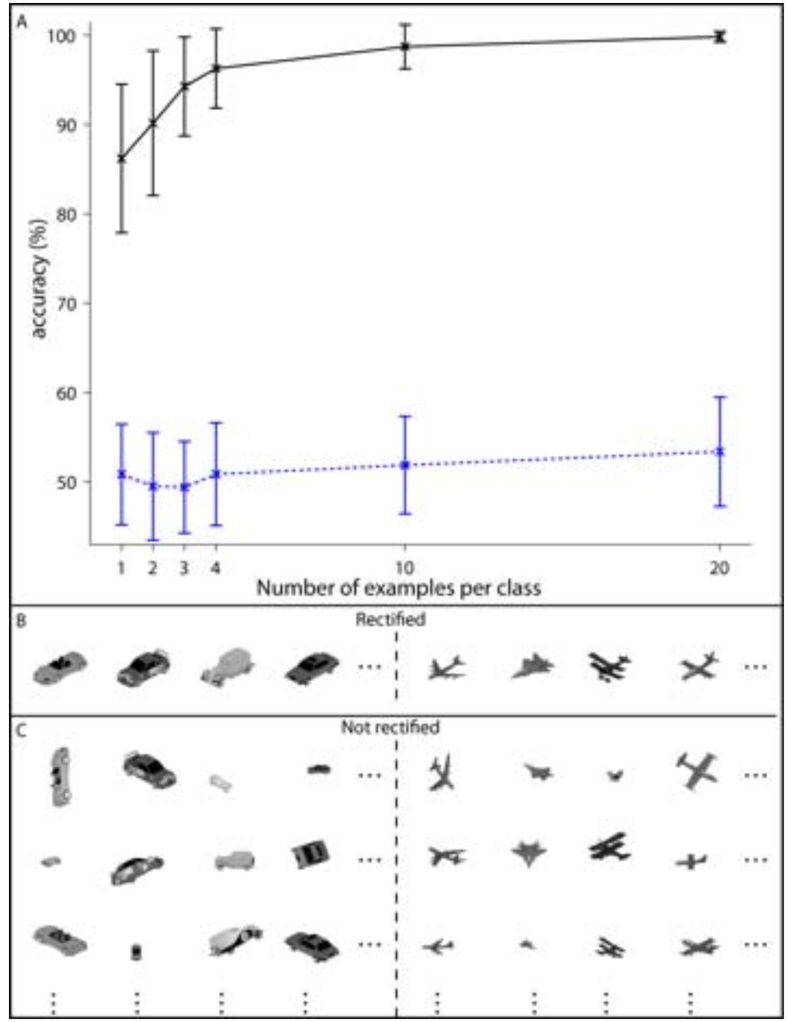
i-theory: exploring a new hypothesis

A main computational goal of the *feedforward* ventral stream hierarchy — and of vision — is to compute a representation for each incoming image which is invariant to transformations previously experienced in the visual environment.

Empirical demonstration: invariant representation leads to lower sample complexity for a supervised classifier

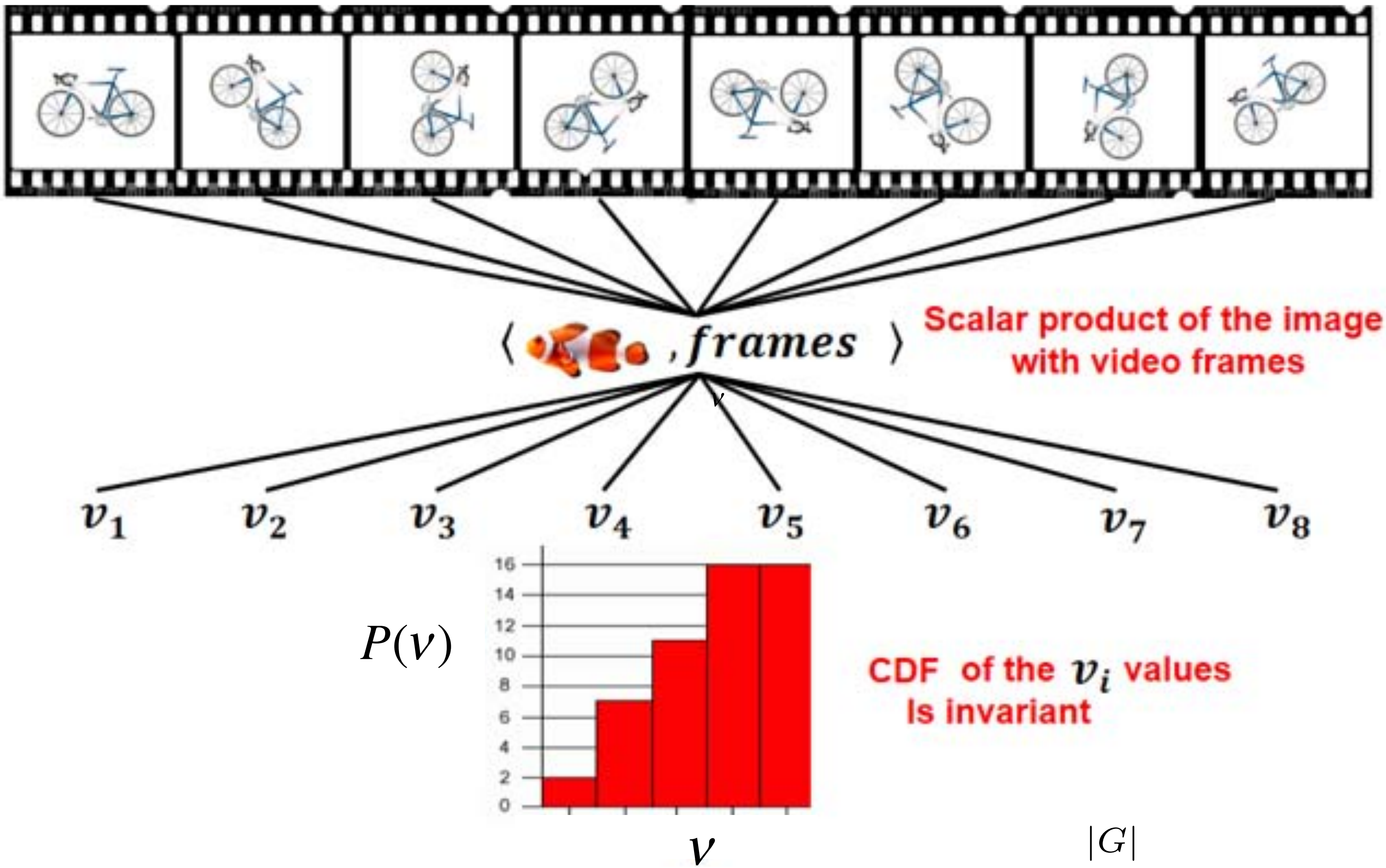
Theorem *(translation case)* Consider a space of images of dimensions $d \times d$ pixels which may appear in any position within a window of size $rd \times rd$ pixels. The usual image representation yields a sample complexity (of a linear classifier) of order $m = O(r^2 d^2)$;the oracle representation (invariant) yields (because of much smaller covering numbers) a sample complexity of order

$$m_{oracle} = O(d^2) = \frac{m_{image}}{r^2}$$



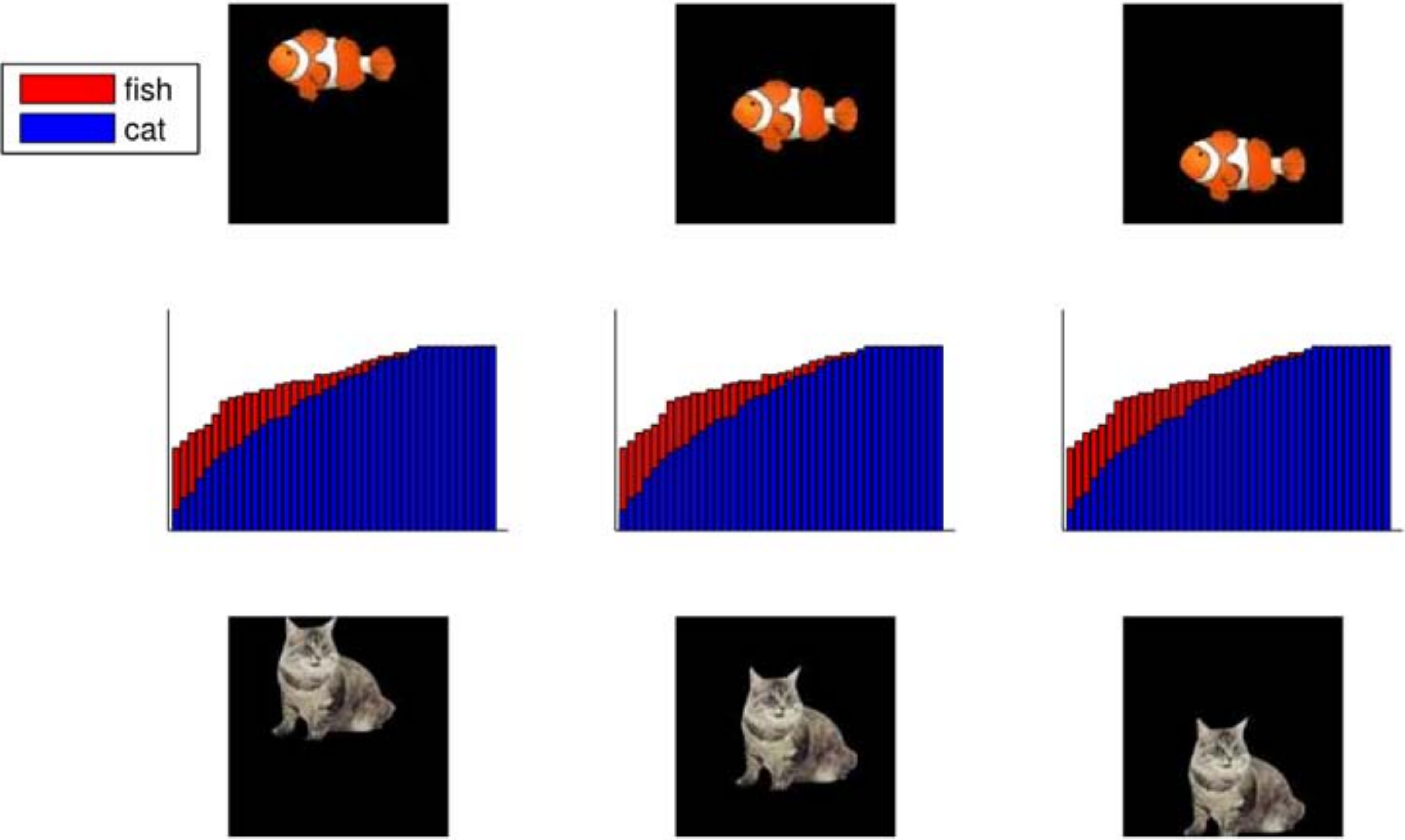
Source: Anselmi, Fabio, Joel Z. Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. "Unsupervised learning of invariant representations. " Theoretical Computer Science 633 (2016): 112-121.

An algorithm that learns in an unsupervised way to compute invariant representations



$$\mu_n^k(I) = 1/|G| \sum_{i=1}^{|G|} \sigma(I \cdot g_i t^k + n\Delta)$$

Invariant signature from a single image of a new object



We need only a finite number of projections, K, to distinguish among n images. Similar in spirit to Johnson-Lindestrauss

$d(I, I')$ distance using all templates

$\hat{d}_K(I, I')$ distance using K templates

Suppose we have n images

$\|d(I, I') - \hat{d}_K(I, I')\| \leq \varepsilon$ with probability $1 - \delta^2$ if

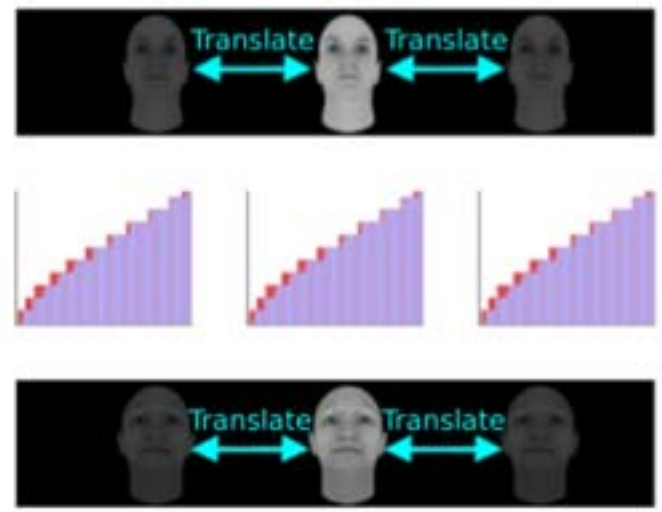
$$K \geq \frac{2}{c\varepsilon^2} \log(\frac{n}{\delta})$$

I-Theory

So far: compact groups in R^2

I-theory extend proves invariance+uniqueness theorems for

- partially observable groups
- non-group transformations
- hierarchies of magic HW modules (multilayer)



Source: Liao, Qianli, Joel Z. Leibo, and Tomaso Poggio. "Learning invariant representations and applications to face verification." In Advances in Neural Information Processing Systems, pp. 3057-3065. 2013.

Invariance, sparsity, wavelets

Theorem: Sparsity is *necessary and sufficient* condition for translation and scale invariance. Sparsity for translation (respectively scale) invariance is equivalent to the support of the template being small in space (respectively frequency).

Theorem: Maximum simultaneous invariance to translation and scale is achieved by Gabor templates:

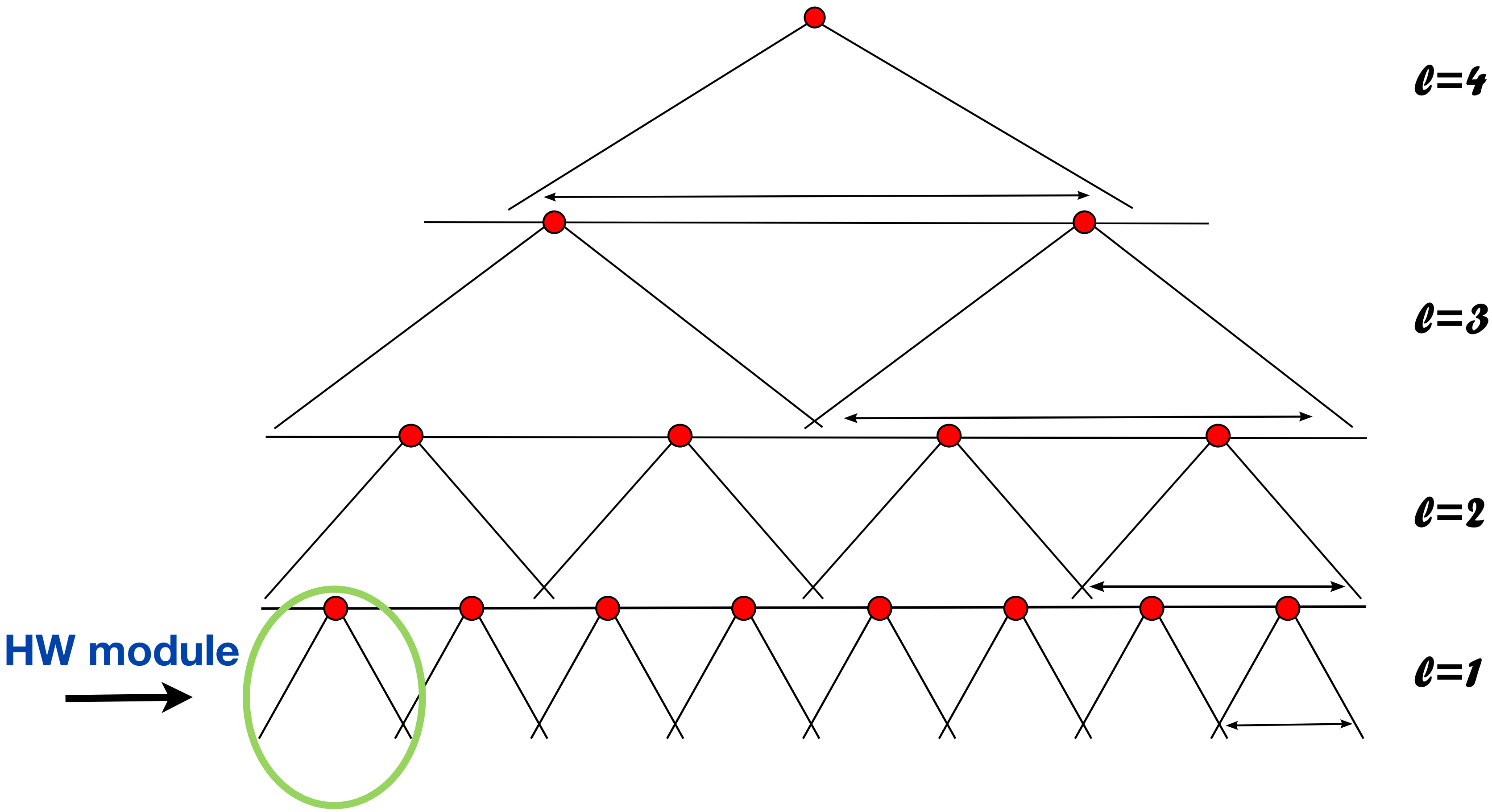
$$t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}$$

Non-group transformations: approximate invariance in class-specific regime

$\mu_n^k(I)$ is locally invariant if:

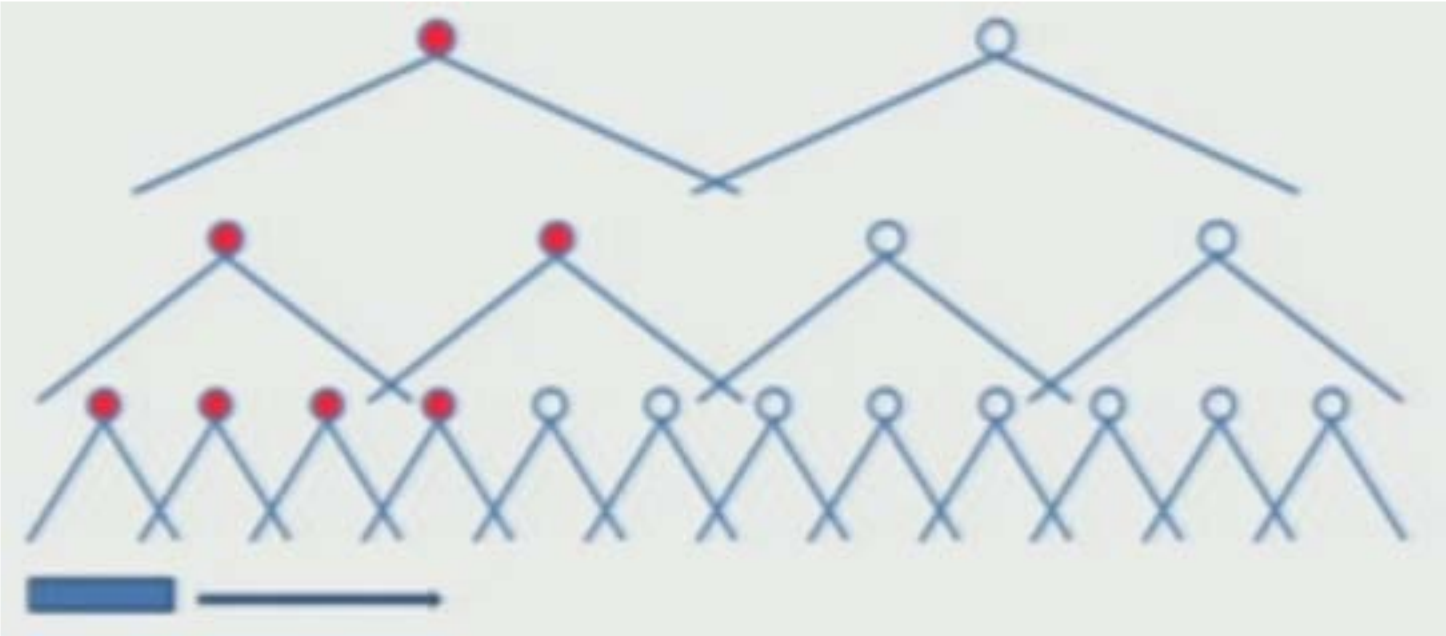
- I is sparse in the dictionary of t^k
- I transforms in the same way (belong to the same class) as t^k
- the transformation is sufficiently smooth

Hierarchies of magic HW modules: key property is covariance



Courtesy of The Center for Brains, Minds and Machines, MIT.

Local and global invariance: whole-parts theorem



Source: Serre, Thomas, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. No. AI MEMO-2005-036. Massachusetts Institute of Technology Center for Biological and Computational Learning, 2005.

For any signal (image) there is a layer in the hierarchy such that the response is invariant w.r.t. the signal transformation.

biophysics: prediction on simple-complex cell

Basic machine: a HW module

(dot products and histograms/moments for image seen through RF)

- The cumulative histogram (empirical cdf) can be be computed as

$$\mu_n^k(I) = \frac{1}{|G|}\sum_{i=1}^{|G|}\sigma(\langle I, g_i t^k\rangle + n\Delta)$$

- This maps directly into a set of simple cells with threshold $n\Delta$

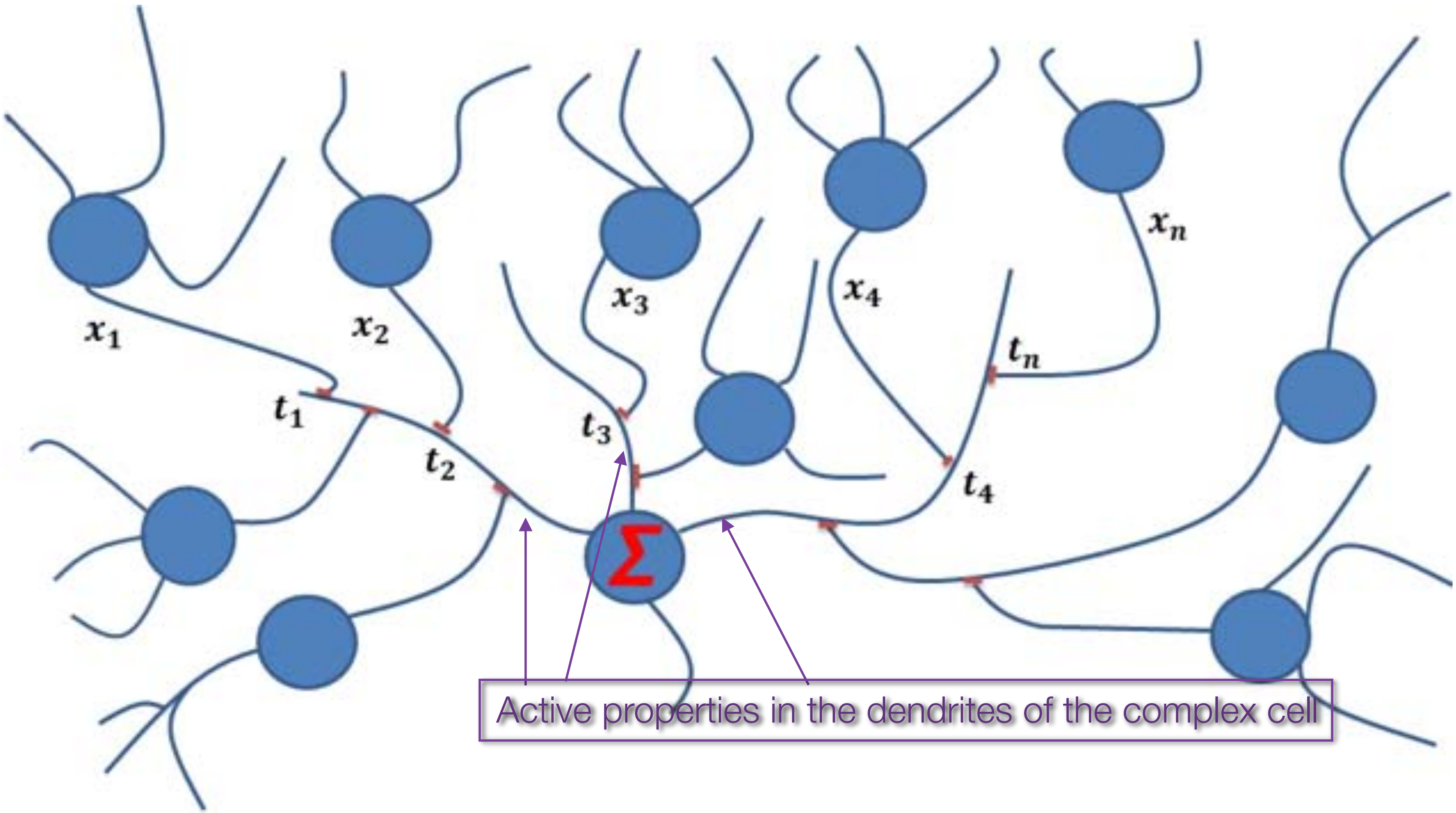
- …and a complex cell indexed by n and k summating the simple cells

The nonlinearity can be rather arbitrary for invariance provided it is stationary in time

Robust and bio plausible

- nonlinearity can be almost anything
- pooling is average but softmax is OK
- low bit precision
- Details and ML remarks

Dendrites of a complex cells *as simple cells*…



Resource: Brains, Minds and Machines Summer Course
Tomaso Poggio and Gabriel Kreiman