The Center for
Brains, Minds and Machines

CENTER FOR
Brains
Minds+
Machines

Center for Brains, Minds, and Machines

Panel Discussion

on the mathematics and the neuroscience of Deep Learning

CENTER FOR
Brains
Minds+
Machines

AAAI Symposium , Palo Alto March 2017

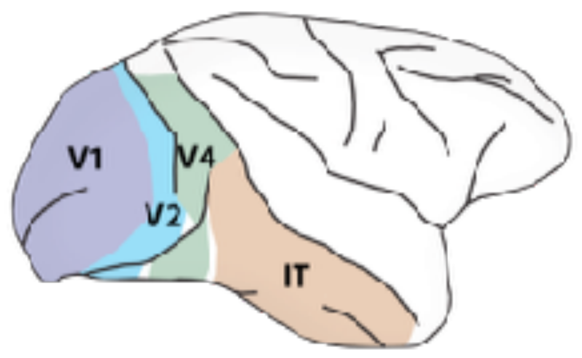
CBMM's <u>main</u> goal is the Science and the Engineering of Intelligence

We aim to make progress in understanding intelligence — that is in understanding how the brain makes the mind, how the brain works and how to build intelligent machines. We believe that the *science of intelligence* will enable better engineering of intelligence.

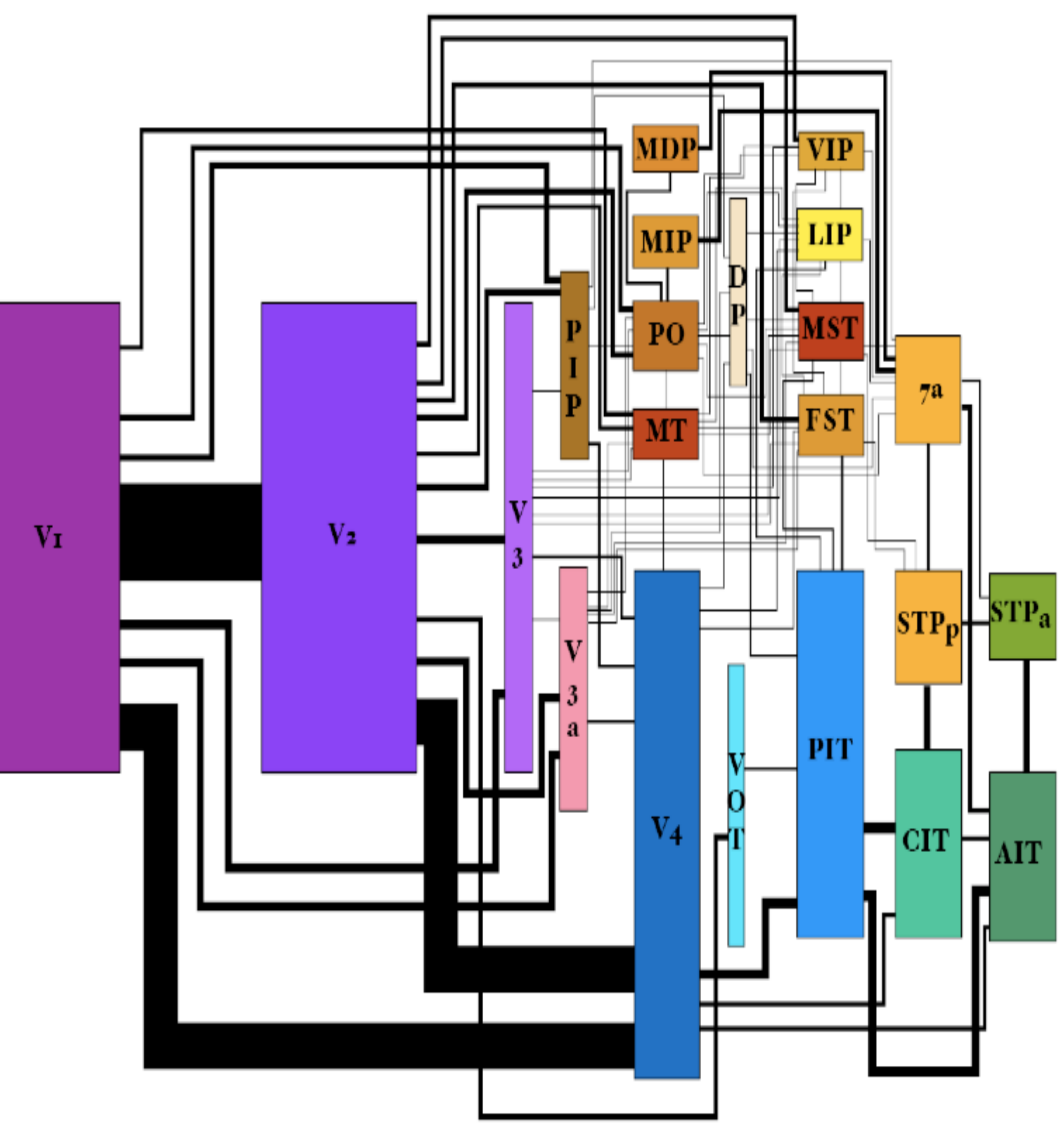
CBMM: the science of intelligence

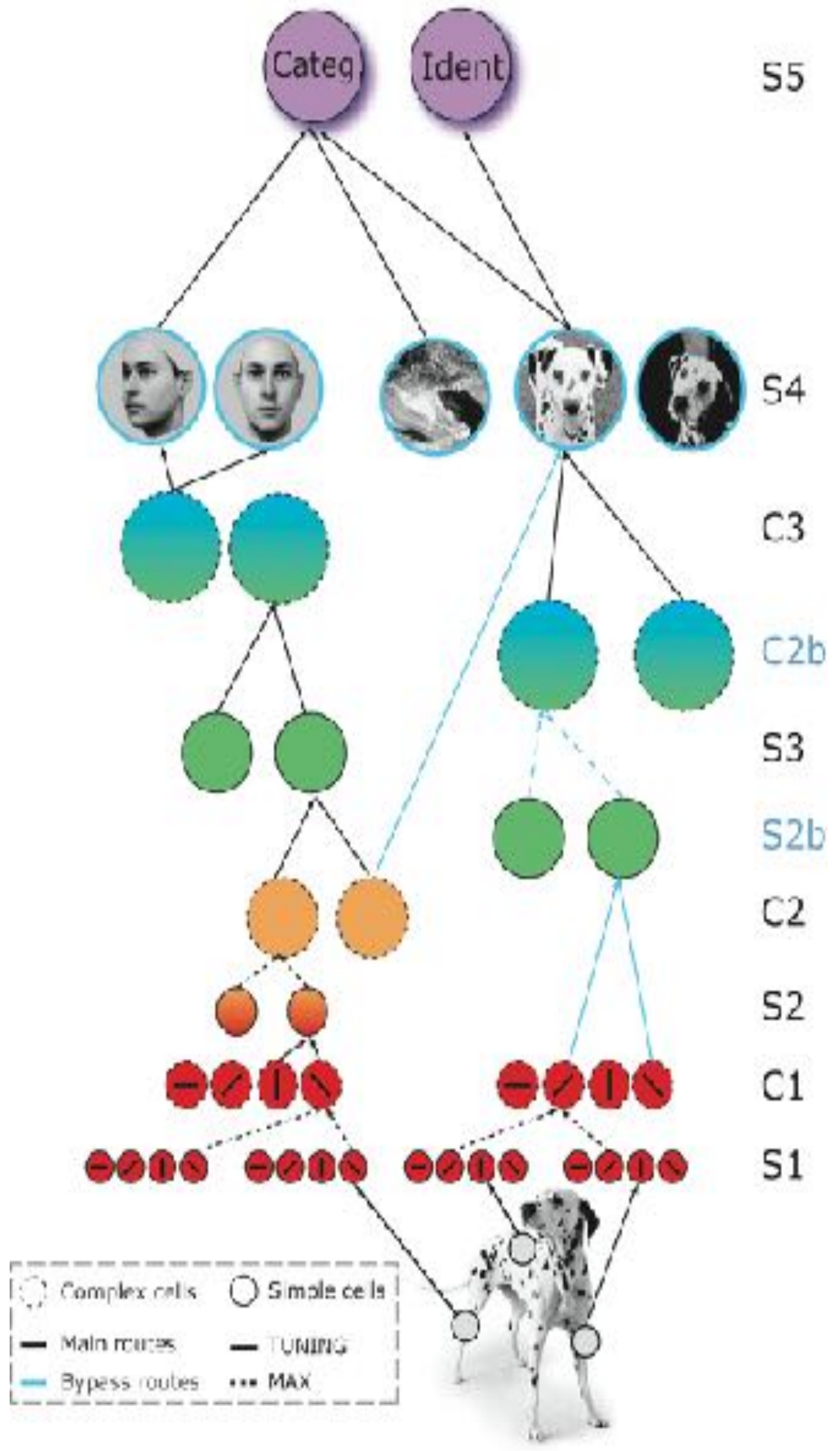
Key recent advances
in the engineering of intelligence
have their roots
in basic research on the brain

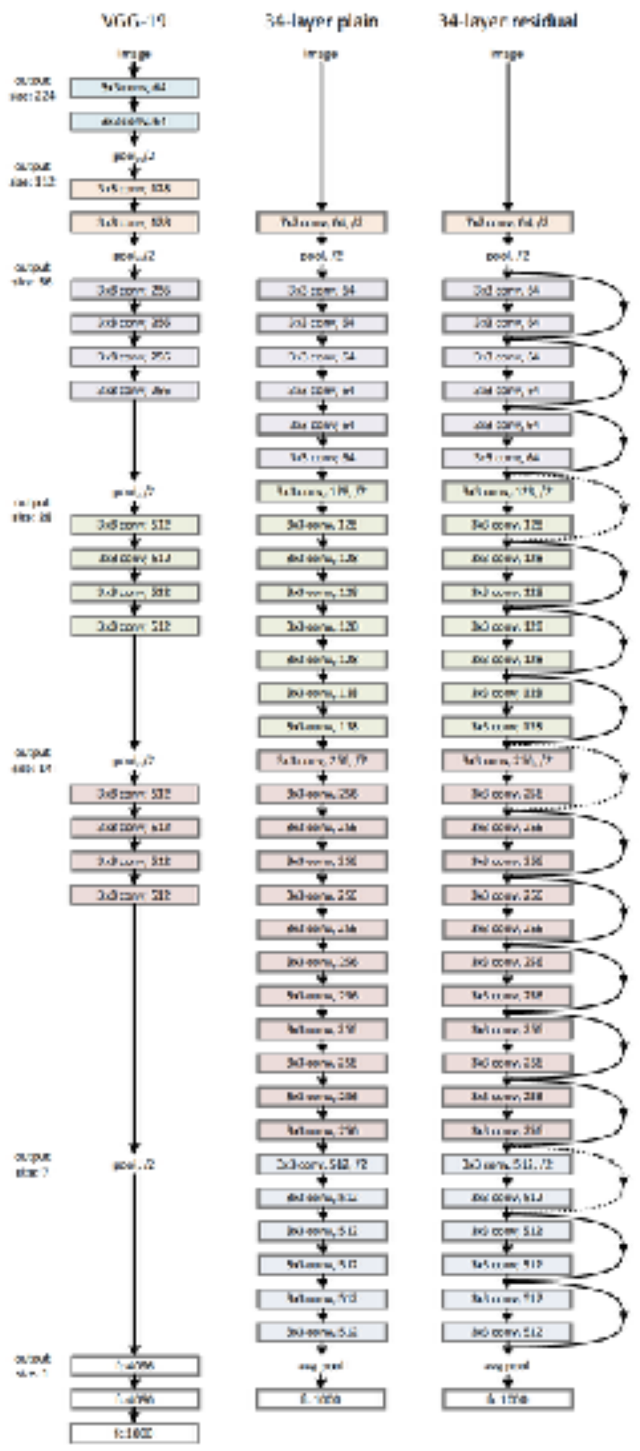
CENTER FOR
Brains
Minds+
Machines

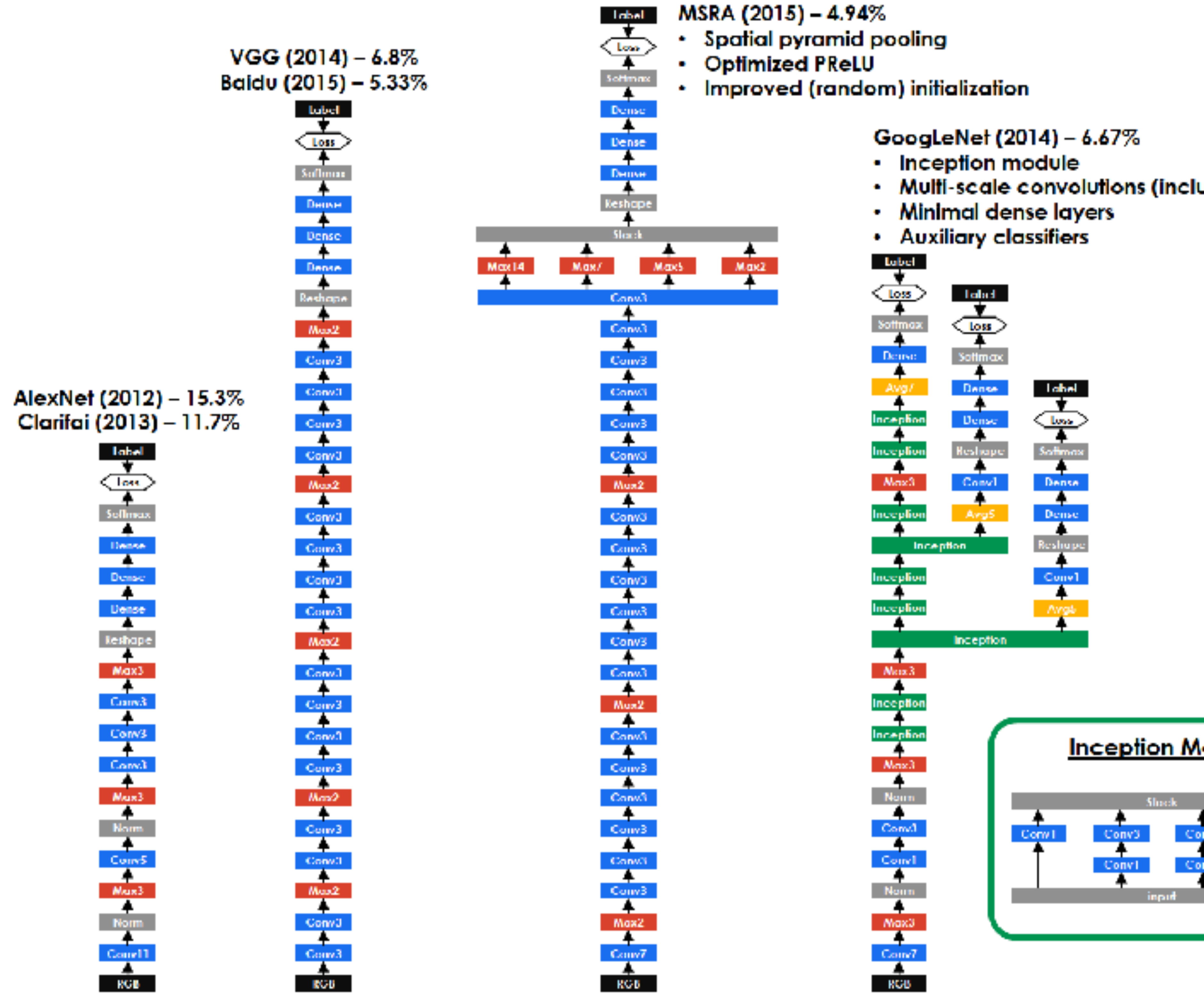


The same hierarchical architectures in the cortex, in models of vision and in Deep Learning networks





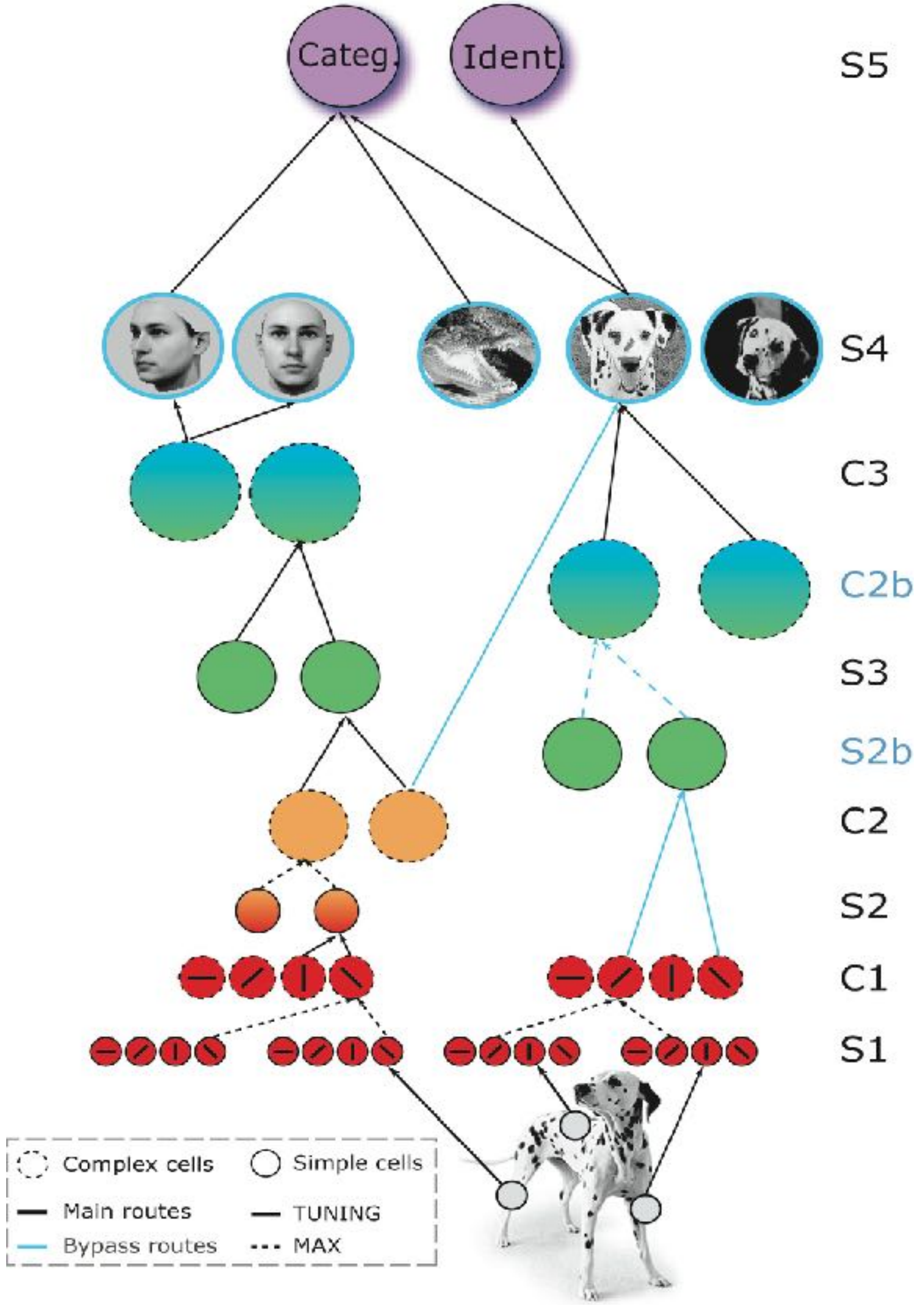




Desimone & Ungerleider 1989; vanEssen+Movshon

Desimone & Ungerleider 1989; vanEssen+Movshon

Convolutional networks



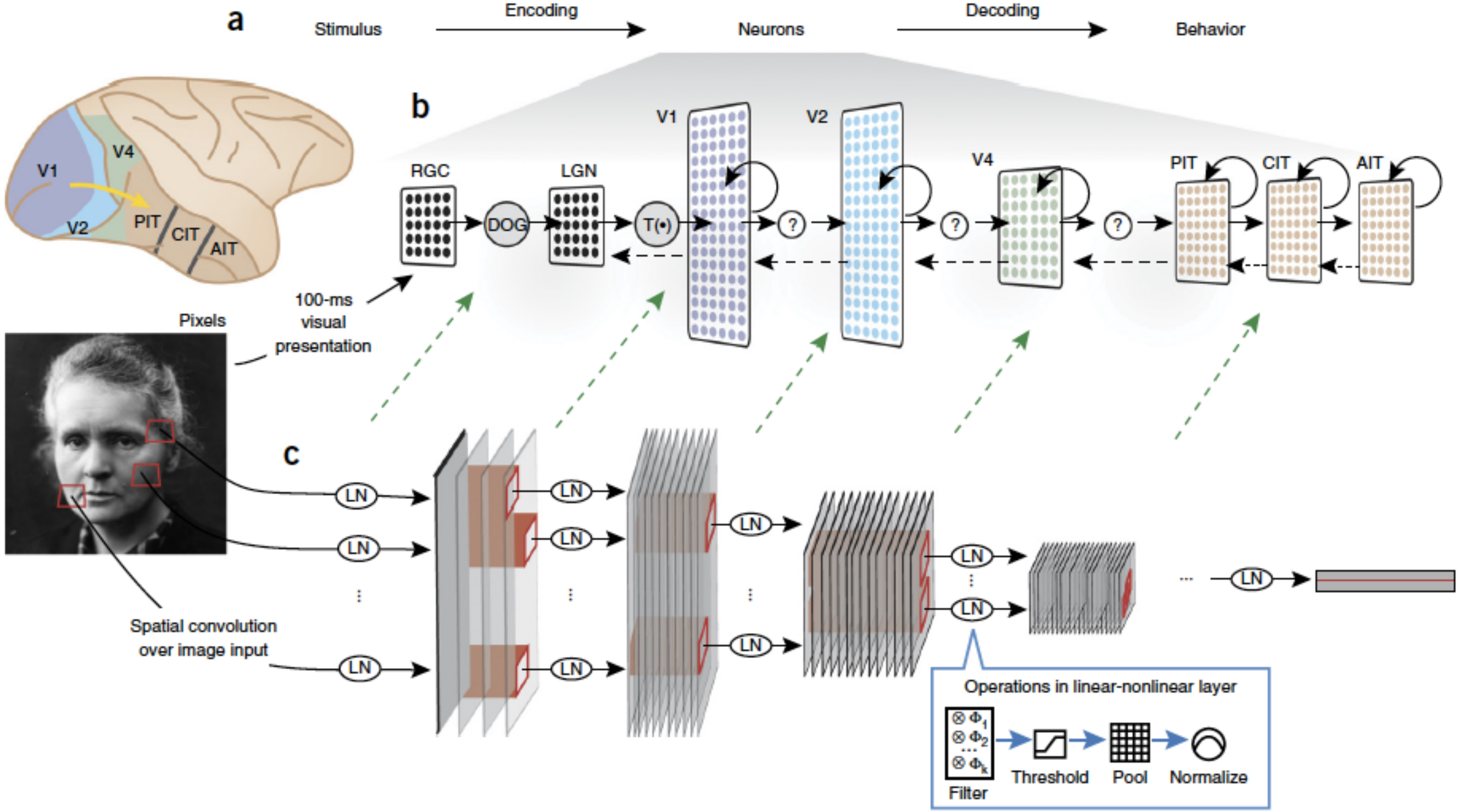
"Hubel-Wiesel" models include

Hubel & Wiesel, 1959: Fukushima, 1980, Wallis & Rolls, 1997; Mel, 1997; LeCun et al 1998; Riesenhuber & Poggio, 1999; Thorpe, 2002; Ullman et al., 2002; Wersing and Koerner, 2003; Serre et al., 2007; Freeman and Simoncelli, 2011….

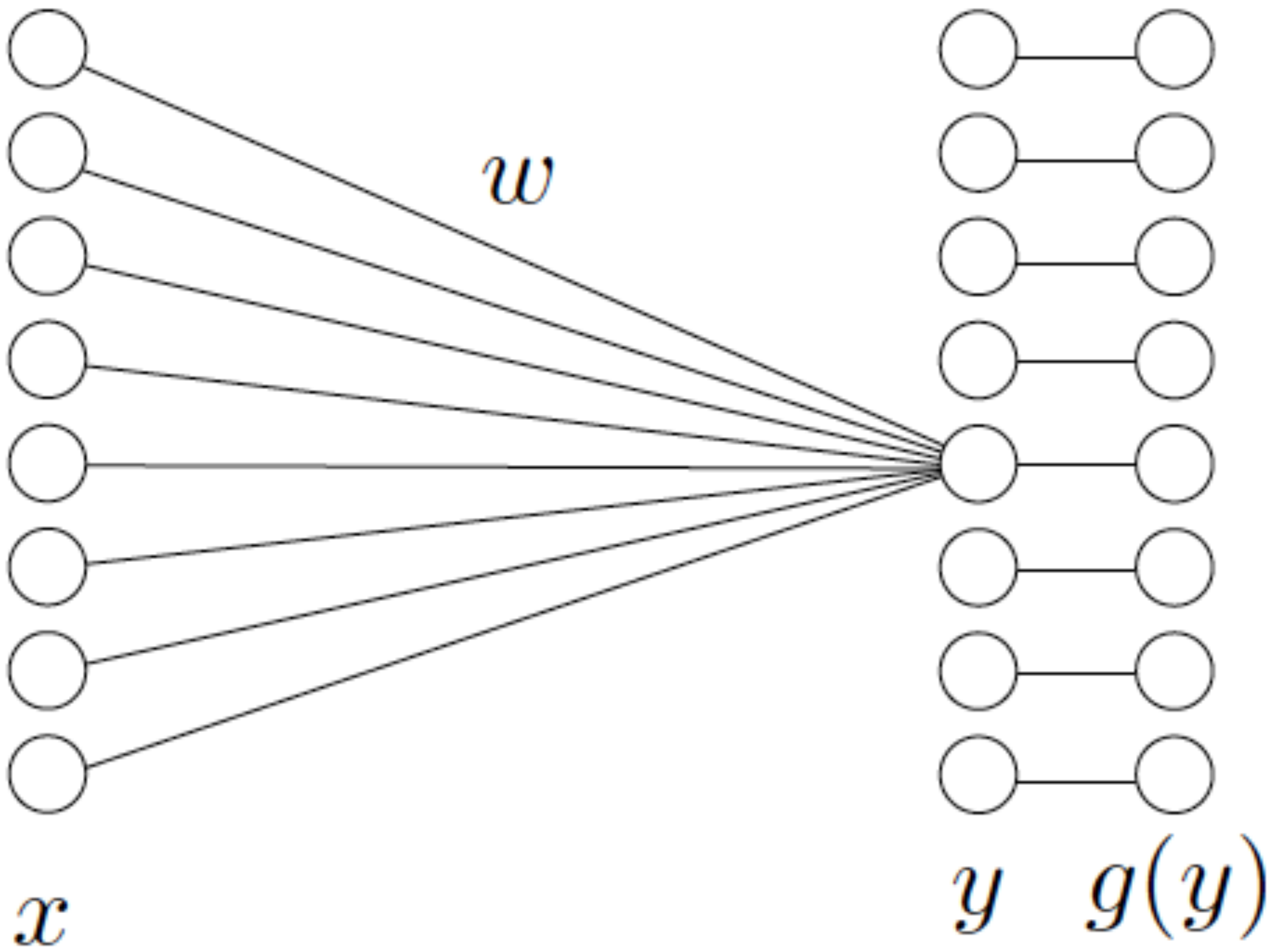
Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Using goal-driven deep learning models to understand sensory cortex

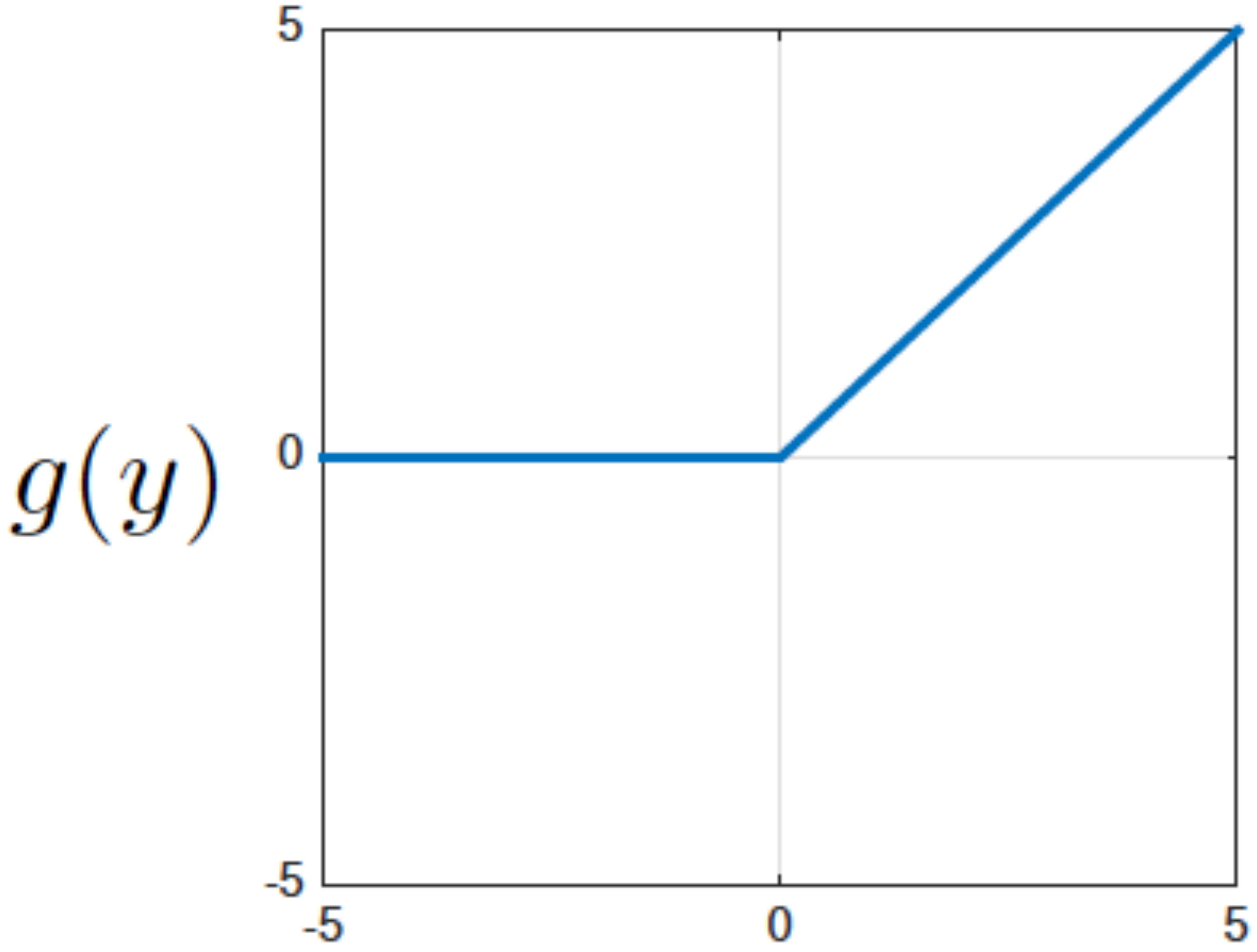
Daniel L K Yamins[1,2] & James J DiCarlo[1,2]



Deep nets architecture

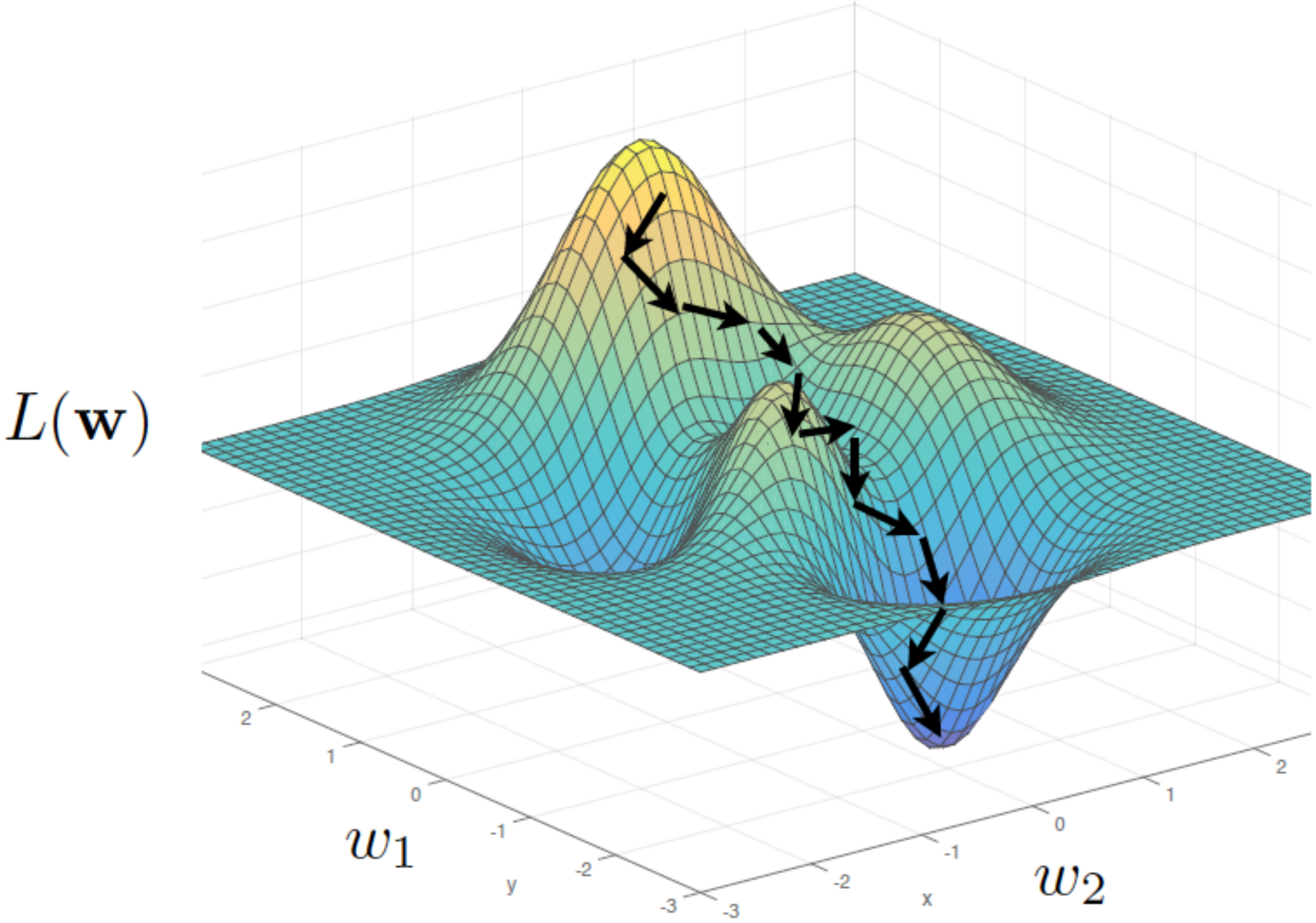


Rectified linear unit (ReLU)



$$g(y) = \max(0, y)$$

Deep nets training: stochastic gradient descent



CENTER FOR
Brains
Minds+
Machines

A theory of Deep Learning

- When and why are deep networks better than shallow networks?

- What is the landscape of the empirical risk?

- How can deep learning generalize so well?

Masker, Poggio et al, 2017

DLNNs: three main scientific questions

Approximation theory: when and why are deep networks better than shallow networks?

Optimization: what is the landscape of the empirical risk?

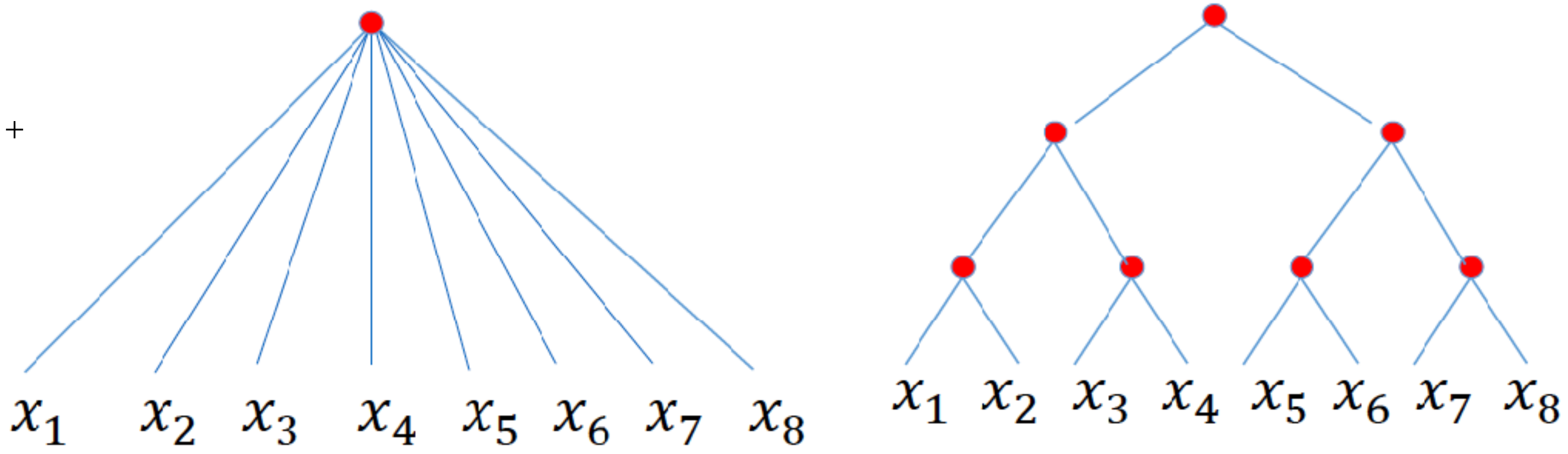
Generalization by SGD: how can overparametrized networks generalize?

Work with Hrushikeshl Mhaskar;
initial parts with L. Rosasco and F. Anselmi

Theory I:
Why and when are deep networks better than shallow networks?

$$f(x_1,x_2,\ldots,x_8) = g_3(g_{21}(g_{11}(x_1,x_2),g_{12}(x_3,x_4))g_{22}(g_{11}(x_5,x_6),g_{12}(x_7,x_8)))$$

$$g(x) = \sum_{i=1}^{r} c_i \left| < w_i, x > + b_i \right|_+$$



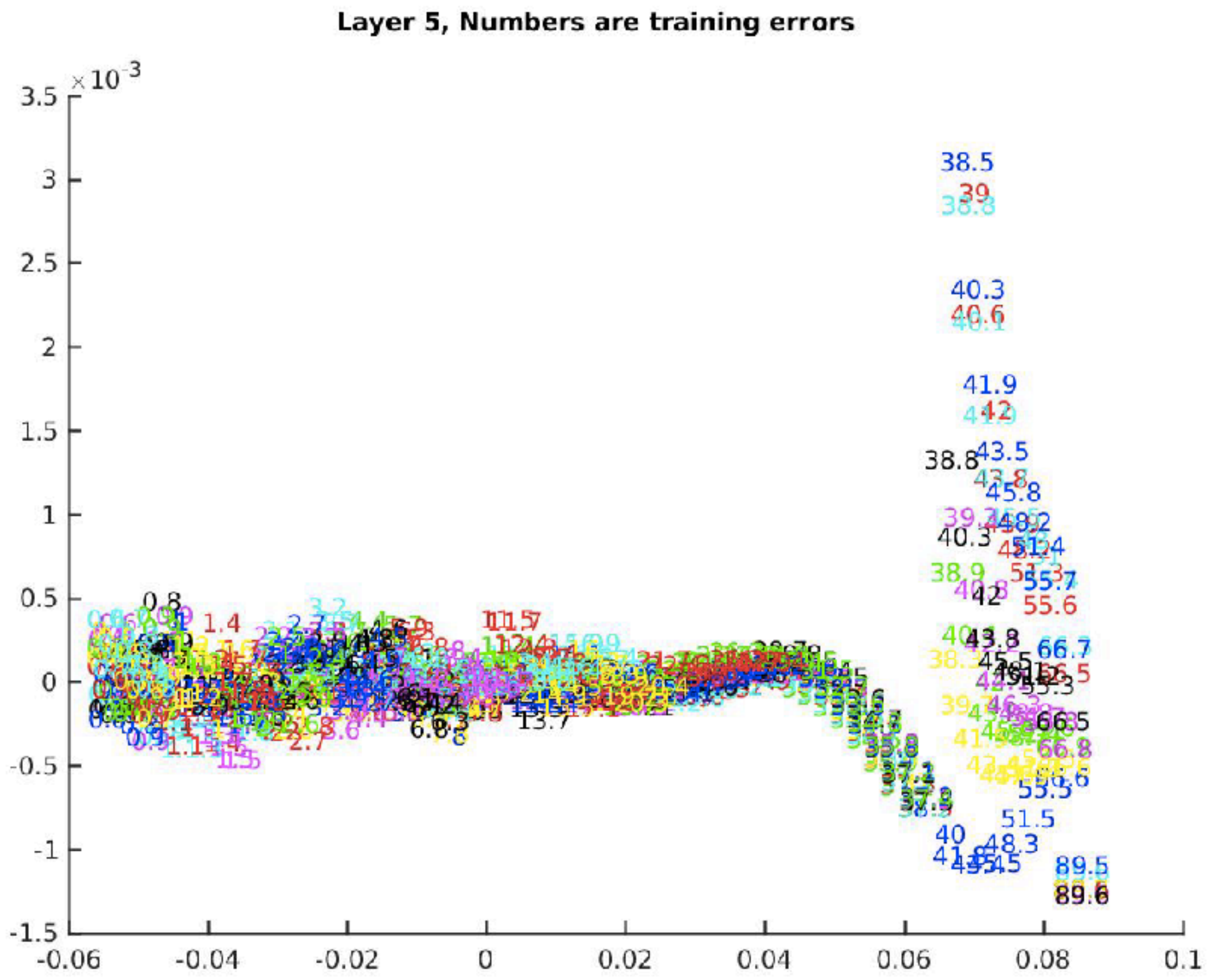
Theorem (informal statement)

Suppose that a function of d variables is compositional . Both shallow and deep network can approximate f equally well. The number of parameters of the shallow network depends exponentially on d as $O(\varepsilon^{-d})$with the dimension whereas for the deep network dance is dimension independent, i.e. $O(\varepsilon^{-2})$

Theory II:
What is the Landscape of the empirical risk?

Theorem (informal statement)

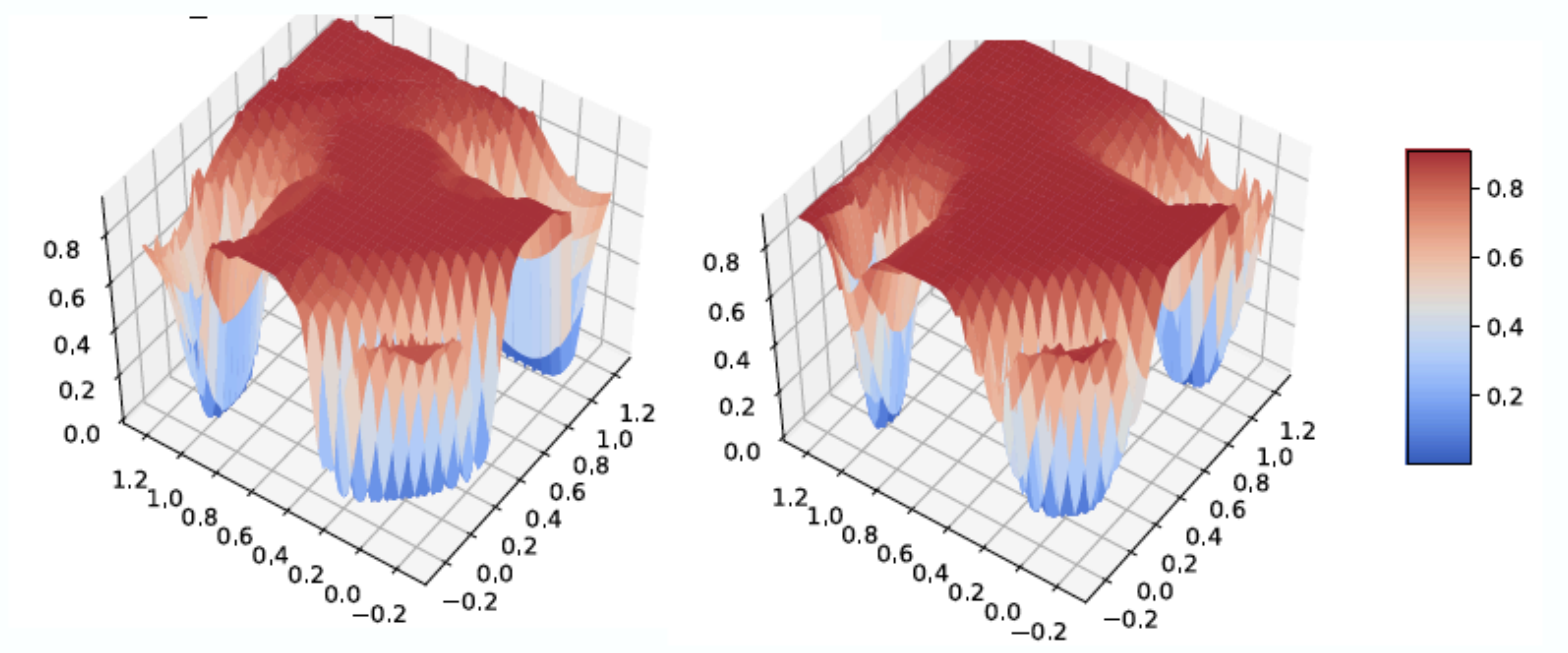
The system of equations for zero empirical error have a very large number of degenerate solutions. Thus there are many zero-minimizers which in the case of classification have a flat non-zero region in all dimensions, that is have a non-zero margin and are degenerate.



Theory III:
How can the underconstrained solutions found by SGD generalize?

Theorem (informal statement)

SGD finds with very high probability degenerate zero-minimizers with large margin. Bounds in terms of margin and Radamacher averages show that generalization is better if the margin is larger.



CIFAR-10: Natural Labels

Random Labels

CENTER FOR
Brains
Minds
Machines

Poggio, Rakhlin, Zhang, Liao, 2017 to appear

Classical learning theory and Kernel Machines (Regularization in RKHS)

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

implies

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

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

RKHS were explicitly introduced in learning theory by Girosi (1997), Vapnik (1998).
Moody and Darken (1989), and Broomhead and Lowe (1988) introduced RBF to learning theory. Poggio and Girosi (1989) introduced Tikhonov regularization in learning theory and worked (implicitly) with RKHS. RKHS were used earlier in approximation theory (eg Parzen, 1952-1970, Wahba, 1990).

Mhaskar, Poggio, Liao, 2016

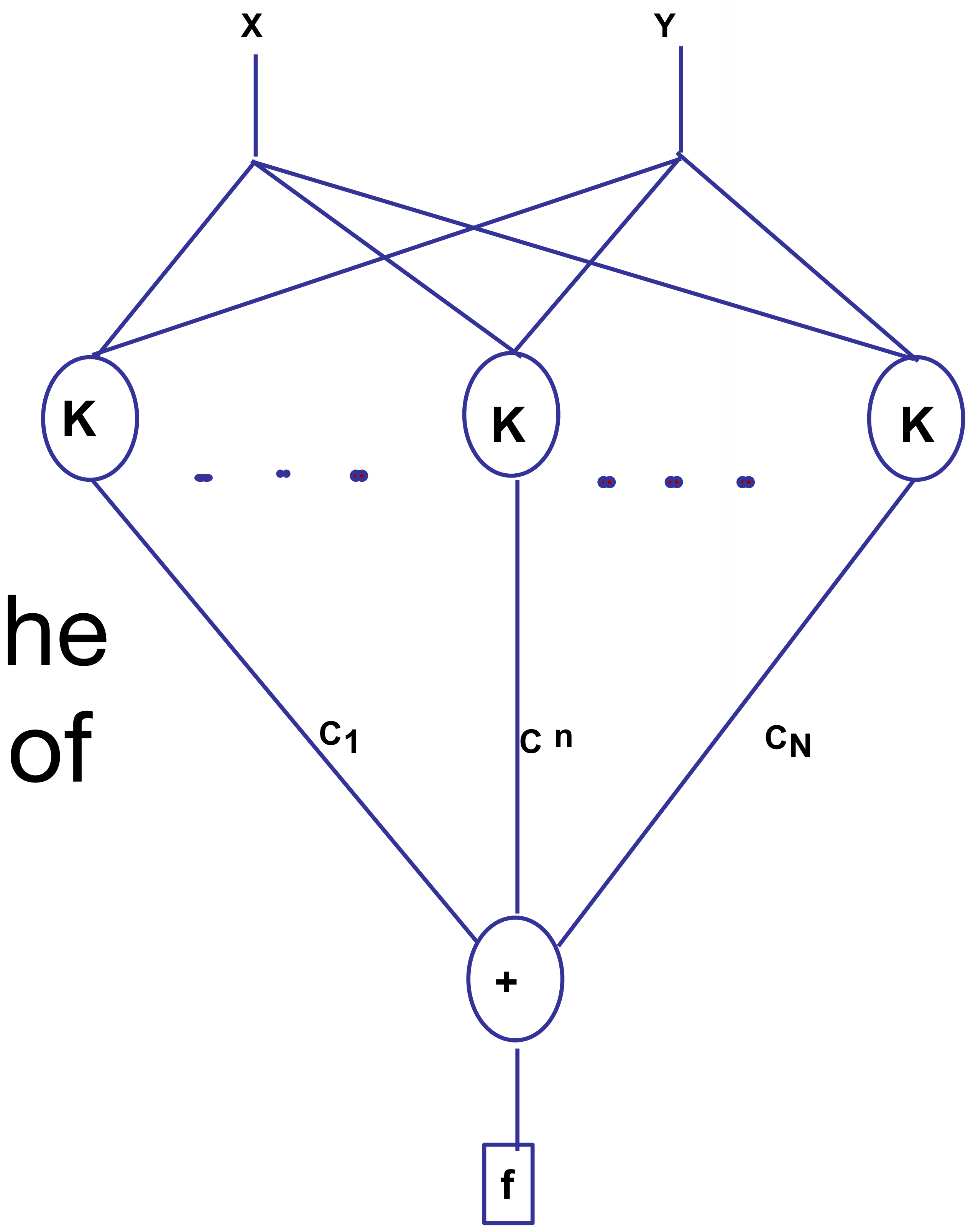
CENTER FOR
Brains
Minds+
Machines

Classical kernel machines are equivalent to shallow networks

Kernel machines…

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

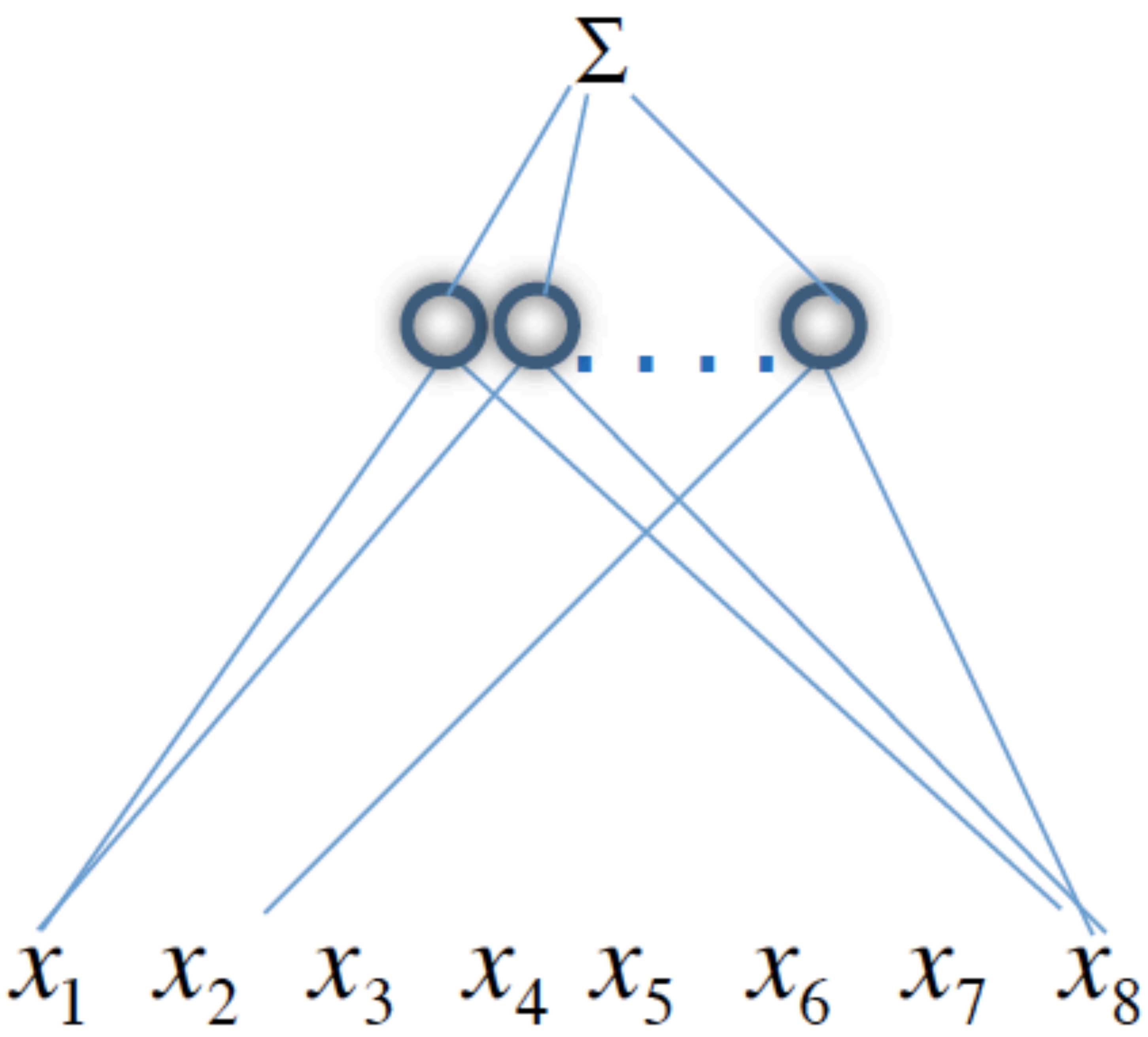
can be “written” as shallow networks: the value of K corresponds to the “activity” of the “unit” for the input and the correspond to “weights”

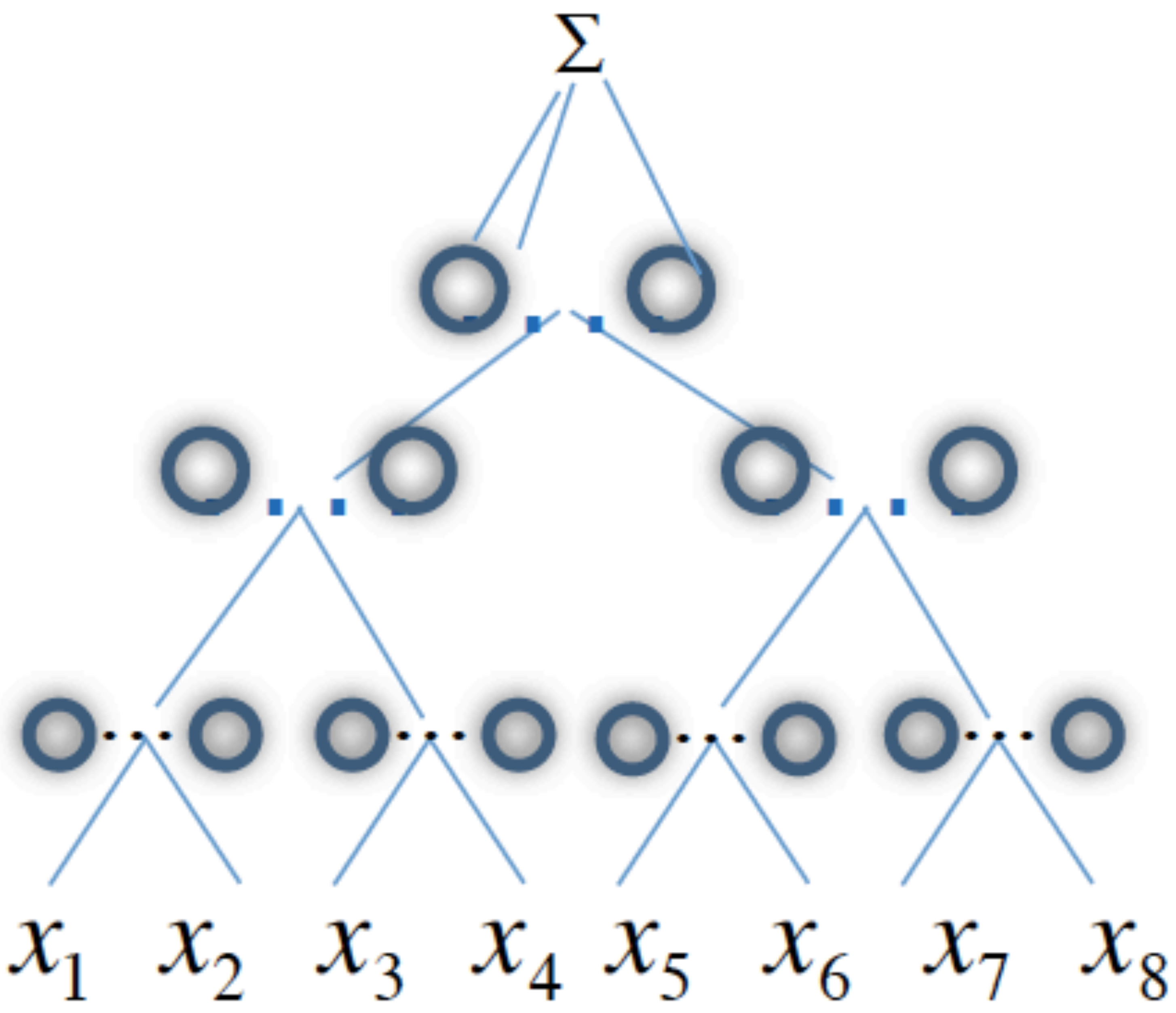


Mhaskar, Poggio, Liao, 2016

Deep and shallow networks: universality

Theorem *Shallow, one-hidden layer networks with a nonlinear $\phi(x)$ which is not a polynomial are universal. Arbitrarily deep networks with a nonlinear $\phi(x)$ (including polynomials) are universal.*





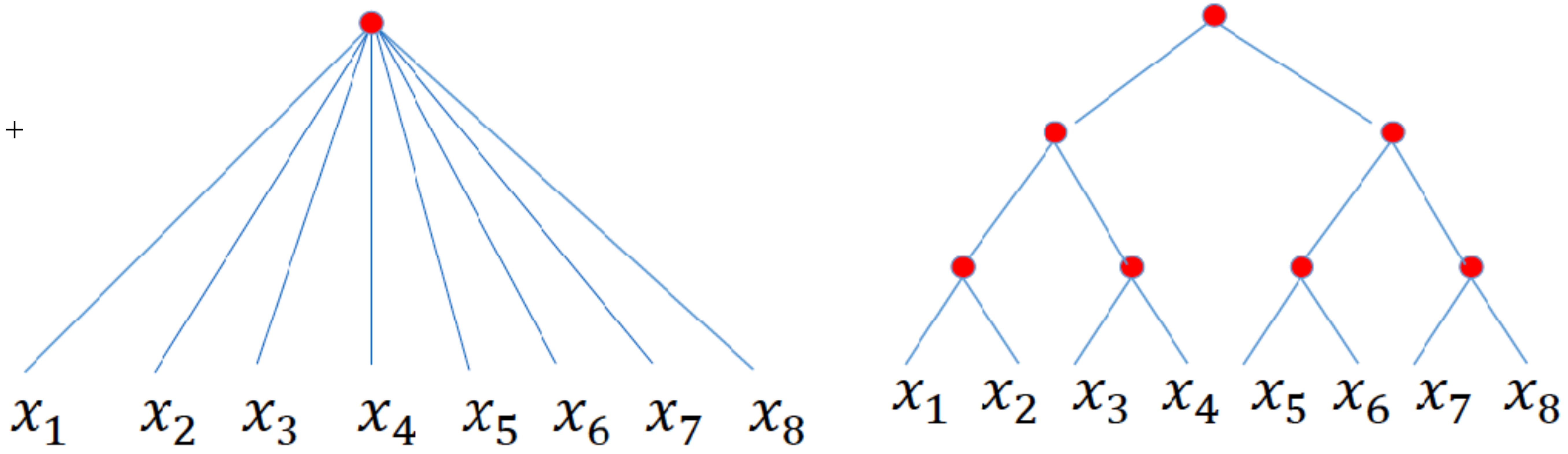
$$\phi(x) = \sum_{i=1}^{r} c_i \left| < w_i, x > + b_i \right|_+$$

Cybenko, Girosi, ….

Theory I:
Why and when are deep networks better than shallow networks?

$$f(x_1,x_2,...,x_8) = g_3(g_{21}(g_{11}(x_1,x_2),g_{12}(x_3,x_4))g_{22}(g_{11}(x_5,x_6),g_{12}(x_7,x_8)))$$

$$g(x) = \sum_{i=1}^{r} c_i \left| < w_i, x > + b_i \right|_+$$



Theorem (informal statement)

Suppose that a function of d variables is compositional . Both shallow and deep network can approximate f equally well. The number of parameters of the shallow network depends exponentially on d as $O(\varepsilon^{-d})$ with the dimension whereas for the deep network dance is dimension independent, i.e. $O(\varepsilon^{-2})$

CENTER FOR
Brains
Minds+
Machines

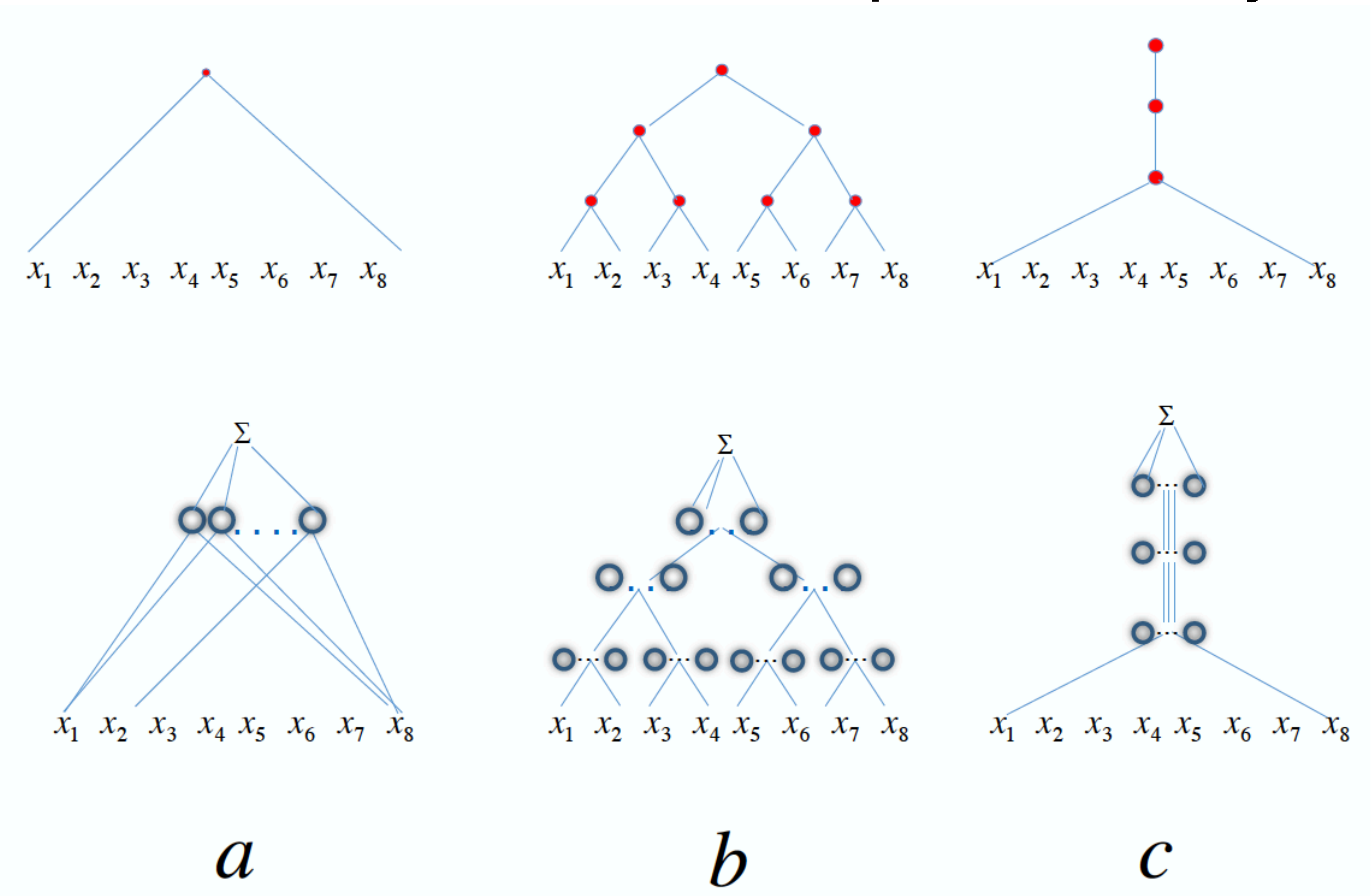
Mhaskar, Poggio, Liao, 2016

Definition Compositionality

$$F(H\ldots(x))$$

Mhaskar, Poggio, Liao, 2016

Microstructure of compositionality



Theorem 3. *Let $\mathcal{G}$ be a DAG, n be the number of source nodes, and for each $v \in V$, let d_v be the number of in-edges of v. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a compositional $\mathcal{G}$-function, where each of the constituent function is in $W^{d_v}_{m_v}$. Consider shallow and deep networks with infinitely smooth activation function as in Theorem 1. Then deep networks – with an associated graph that corresponds to the graph of f – avoid the curse of dimensionality in approximating f for increasing n, whereas shallow networks cannot directly avoid the curse. In particular, the complexity of the best approximating shallow network is exponential in n*

$$N_s = \mathcal{O}(\epsilon^{-\frac{n}{m}}), \tag{9}$$

where $m = \min_{v \in V} m_v$, while the complexity of the deep network is

$$N_d = \mathcal{O}(\sum_{v \in V} \epsilon^{-d_v/m_v}). \tag{10}$$

Mhaskar, Poggio, Liao, 2016

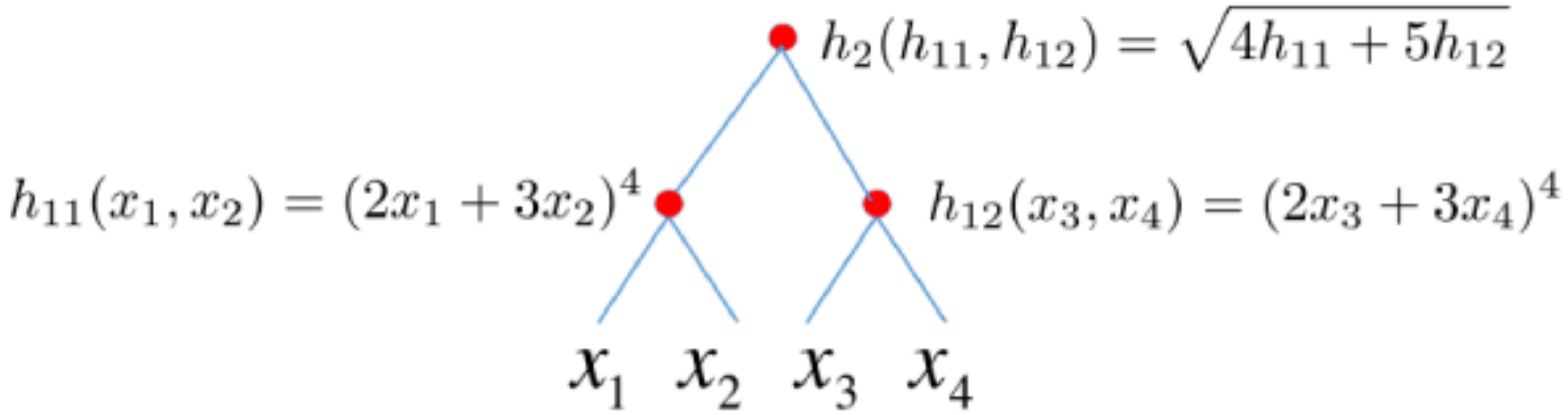
The *curse of dimensionality* and
3 blessings of compositionality

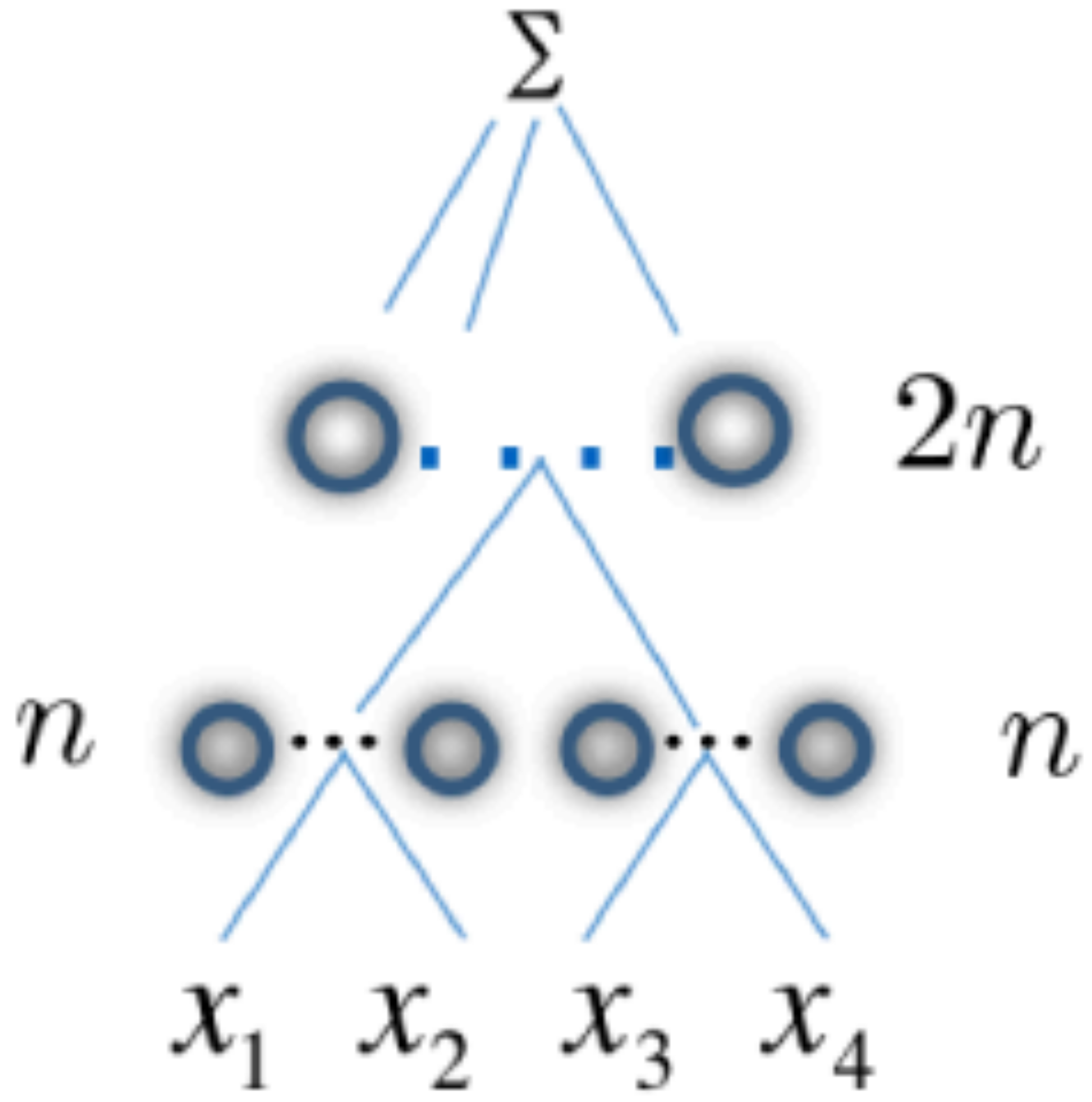
Center for Brains, Minds & Machines

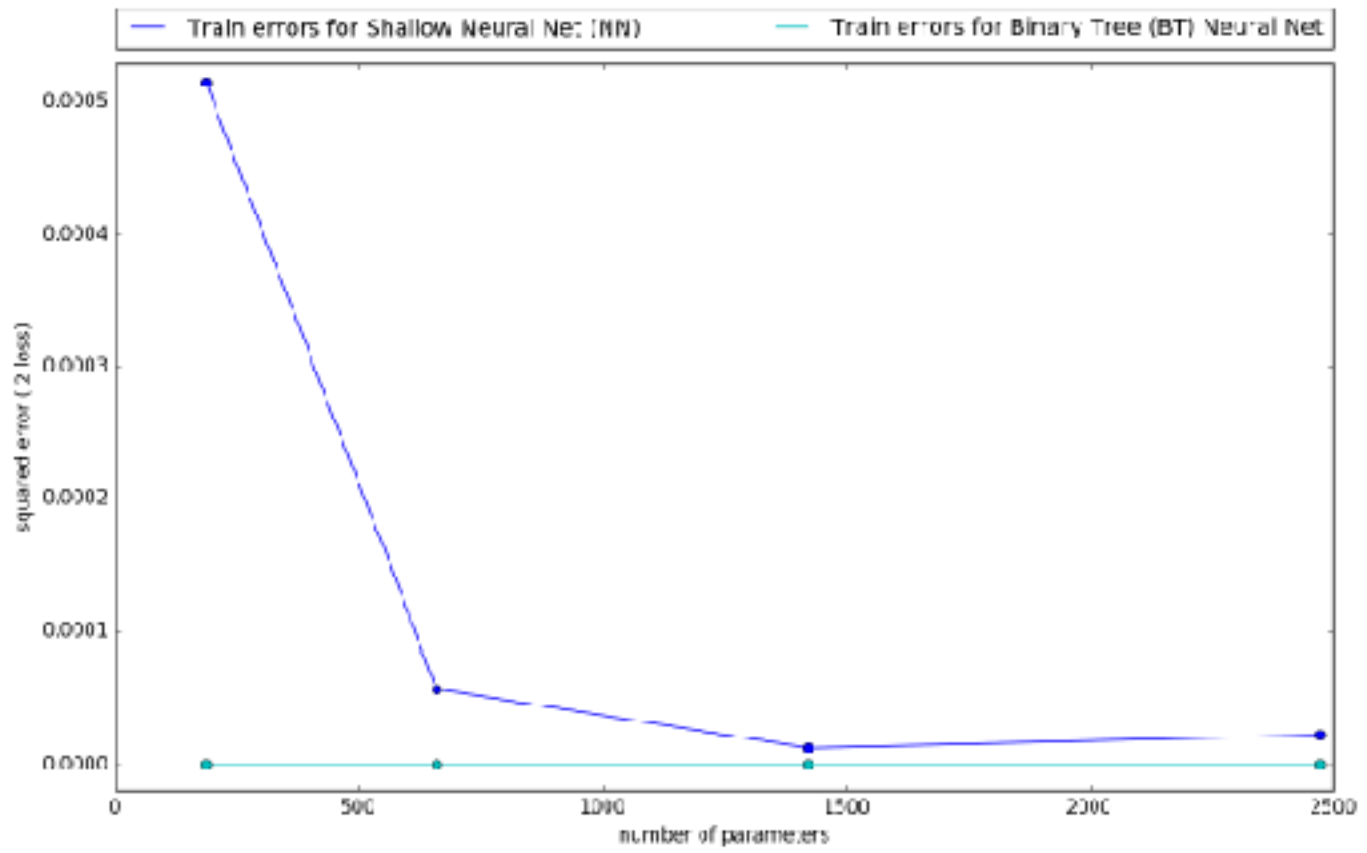
Mhaskar, Poggio, Liao, 2016

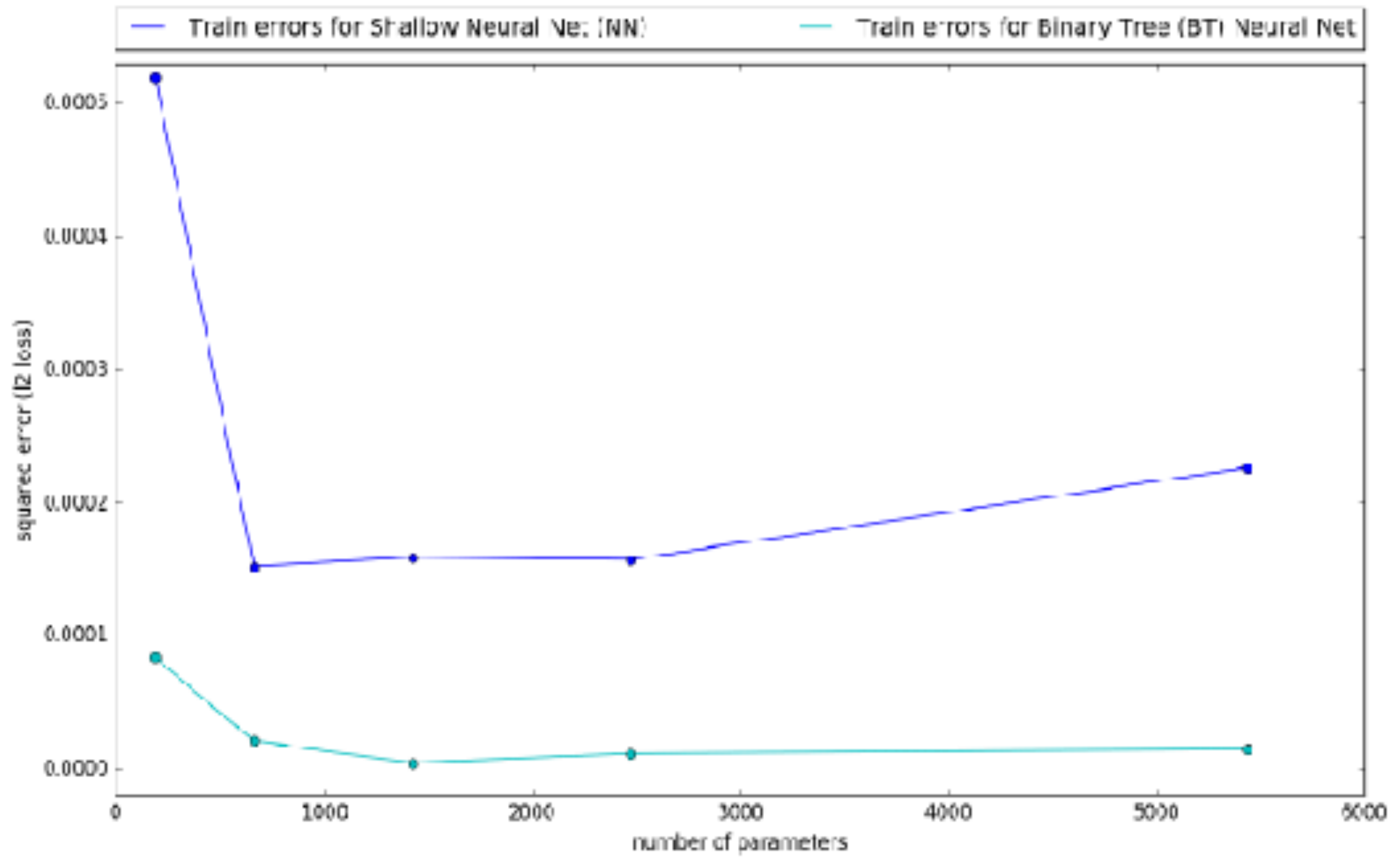
3 blessings of compositionality

- low dim of constituent functions h
- high smoothness of some of the h
- sharing across tasks of K in F= H K for better generalization



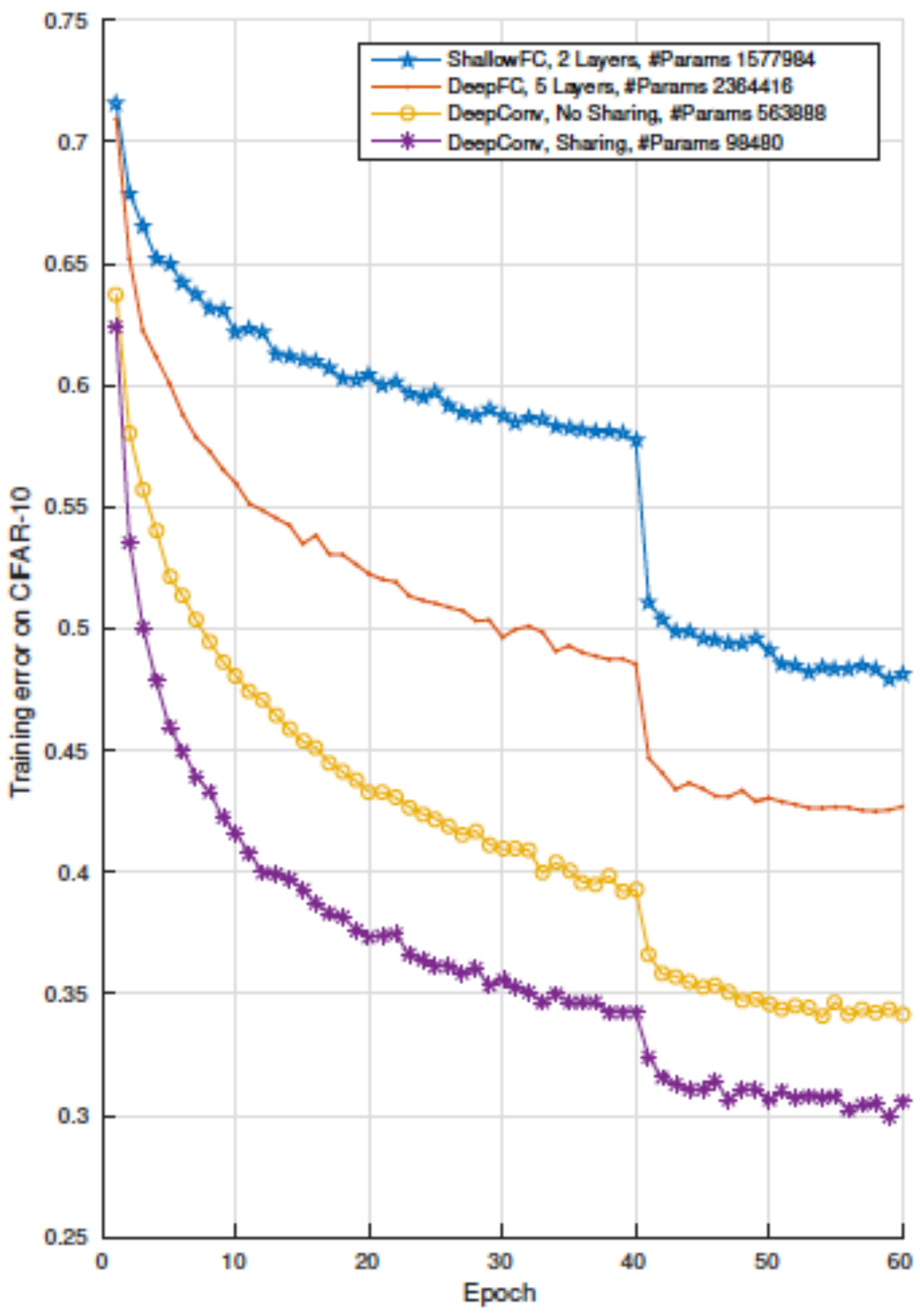


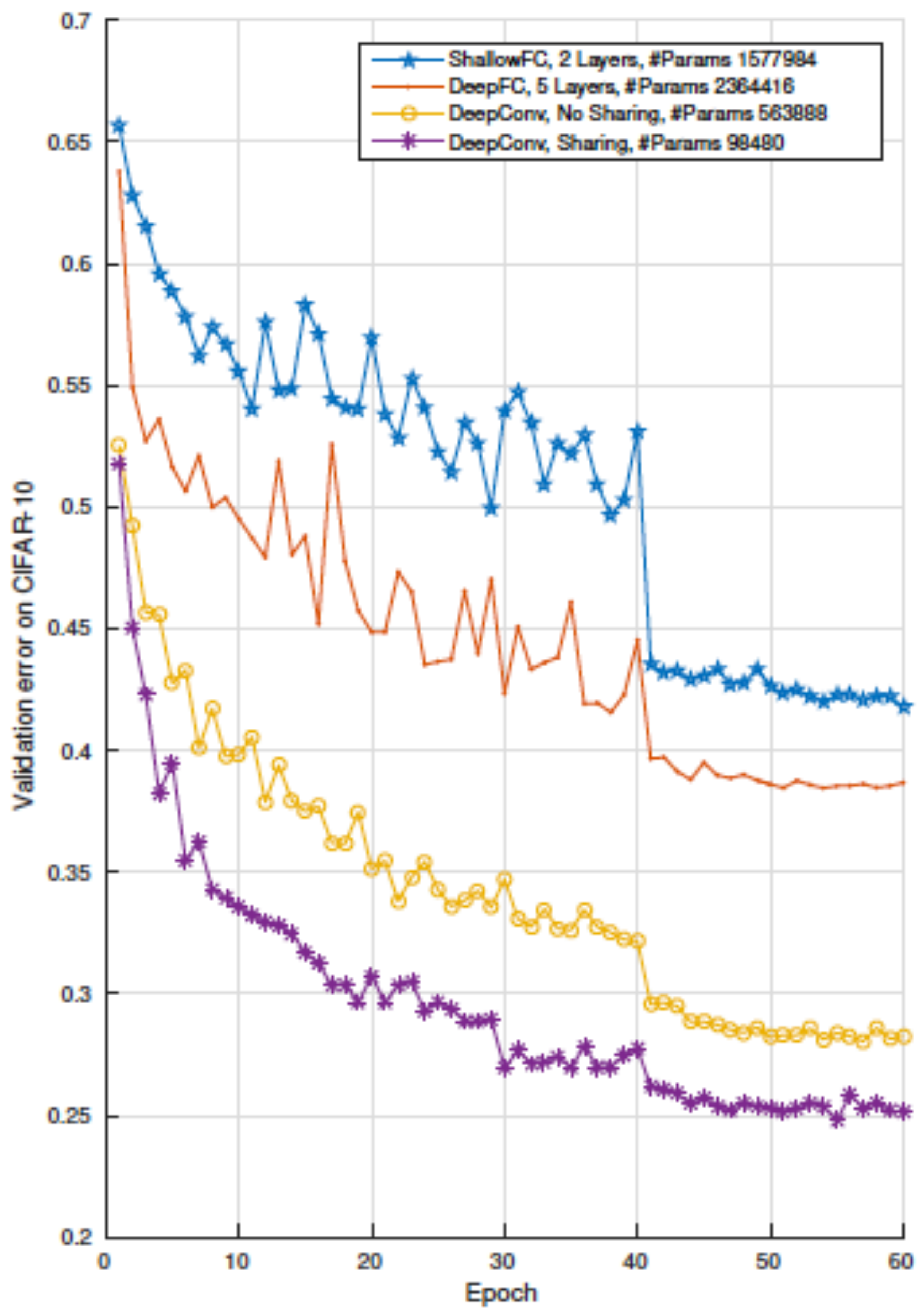




Center for Brains,
Minds & Machines

CIFAR





Center for Brains, Minds & Machines

Panel topics

- mathematical foundations for Deep Learning
- neuroscience plausibility
- epistemology