Statistical Physics and Machine Learning: A 30 years perspective

Physics Next, APS Workshop, Riverhead, October 2018

Naftali Tishby

School of Engineering and Computer Science

The Edmond & Lily Safra Center for Brain Sciences

Hebrew University, Jerusalem, Israel

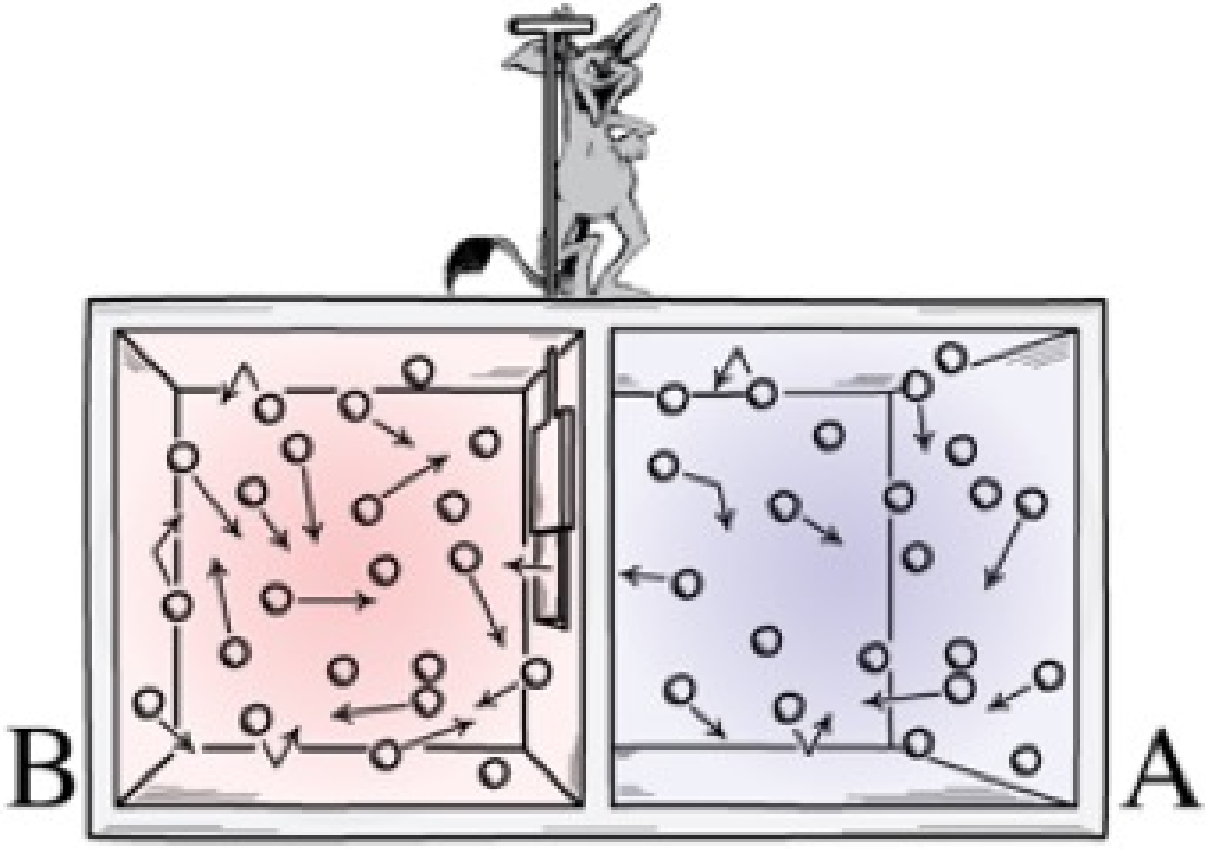
The physics behind Information Science

- 1871 – The origins of AI are in physics – Maxwell, Szilard,…
- 1982 – Hopfield's model & Valient's theory of the learnable
- 1985 – Spin Glass theory of Hopfield's model (AGS) & the PDP book
- Attractors and MLP as competing Neural Network paradigms
- 1987 – Gardner's approach, typical constrained volume calculation
- 1989 – Gardner's approach applied to supervised learning
- 1990 – Learning curves and phase transitions in learning
- 1991 – SM (TAP & Replica) applied to more general inference problems
- 1993 – Rigorous bounds from SM - the power of the Annealed Approximation
- 2000 – AMP and other Graphical models algorithms analyzed with SM
- 2010… - Information Theorists begin to adopt SM methods

The physics behind Information Science

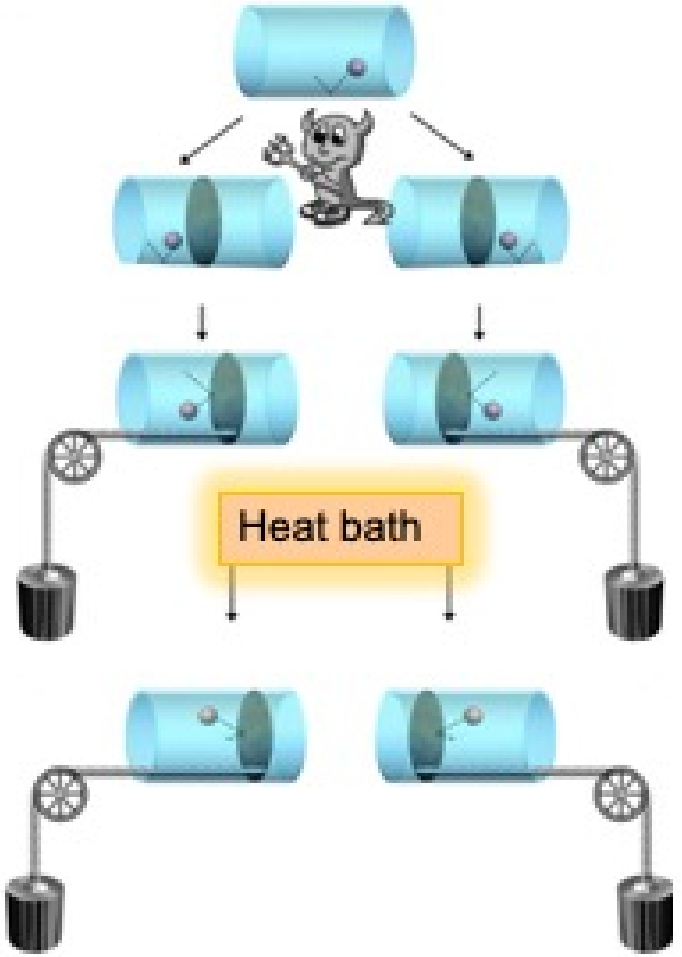
- 1900-95 - The SM of Learning was marginal with tiny influence on ML …
- from ~ 2000 the scale of ML became too large for effective worse case bounds
- Typical behavior theories become essential
- Classical parametric statistics becomes irrelevant and SM & IT take over
- 2008… – Deep Learning and DNNs strike back … no good theory (yet…)
- 2012 – DL is related to RG and emergence of relevant representation theory
- 2015 – suggested connection between DL and representation compression
- 2017 – understanding the role of Stochastic Dynamics (SGD) for DL (?)
- 2010-18 – Replica calculations of Mutual Information in DL and related models
- 2019…. The return of SM as the dominant theoretical framework in ML ???

Maxwell's demon and the simplest life

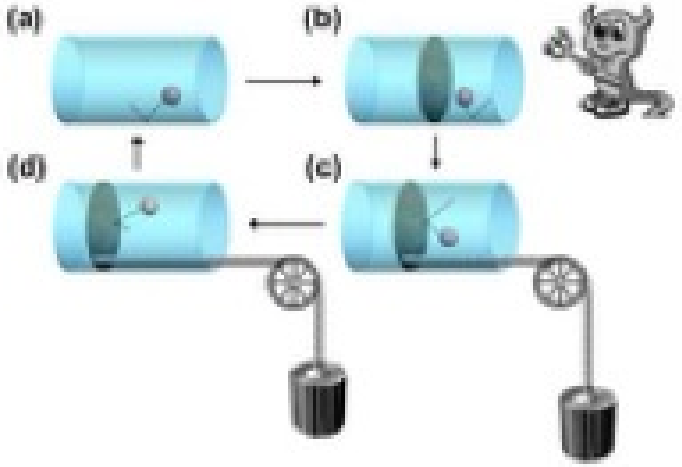


Physics of information 1

Maxwell's Demon & Szilard's Engine



(a) Preparation (reset memory, barrier)
(b) Measurement (noisy) on (R,L)
(c) Action: hang mass on (R,L) side
(d) Gaining work (heat bath, value)



Szilard ["sensing-acting"] cycle

Its all about Information & Computation

Alan Turing
(June 23, 1912 – June 7, 1954)

Claude Elwood Shannon
(April 30, 1916 – February 24, 2001)


The Dream

The Deep Learning revolution

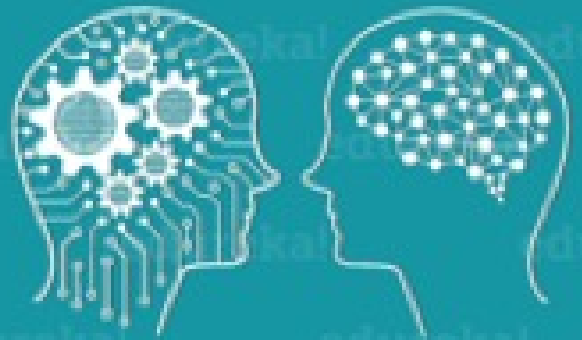
edureka!

ARTIFICIAL INTELLIGENCE

Engineering of making Intelligent Machines and Programs

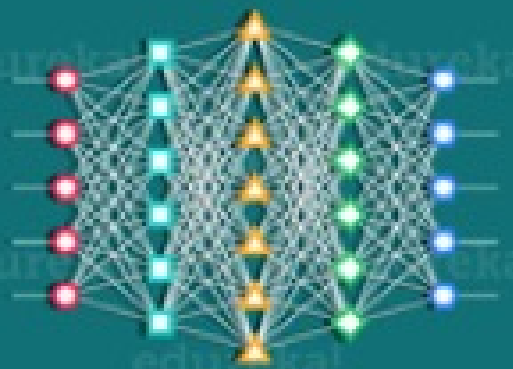
MACHINE LEARNING

Ability to learn without being explicitly programmed



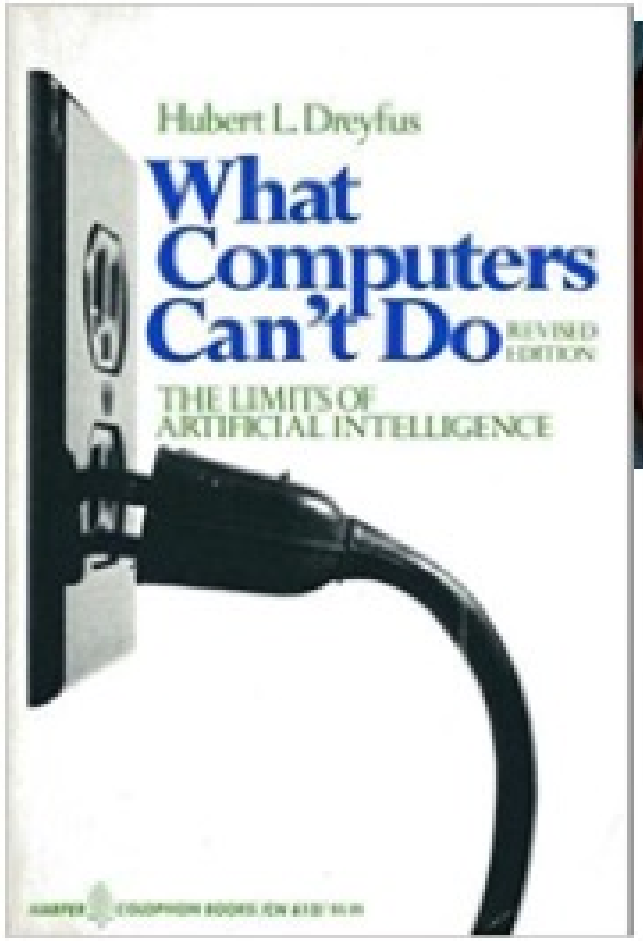
DEEP LEARNING

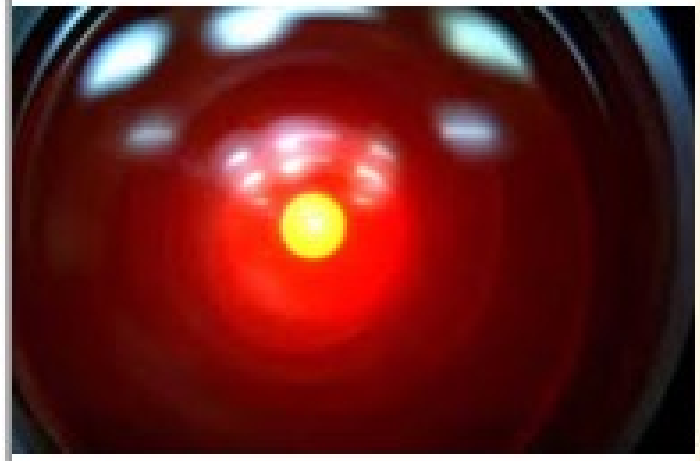
Learning based on Deep Neural Network

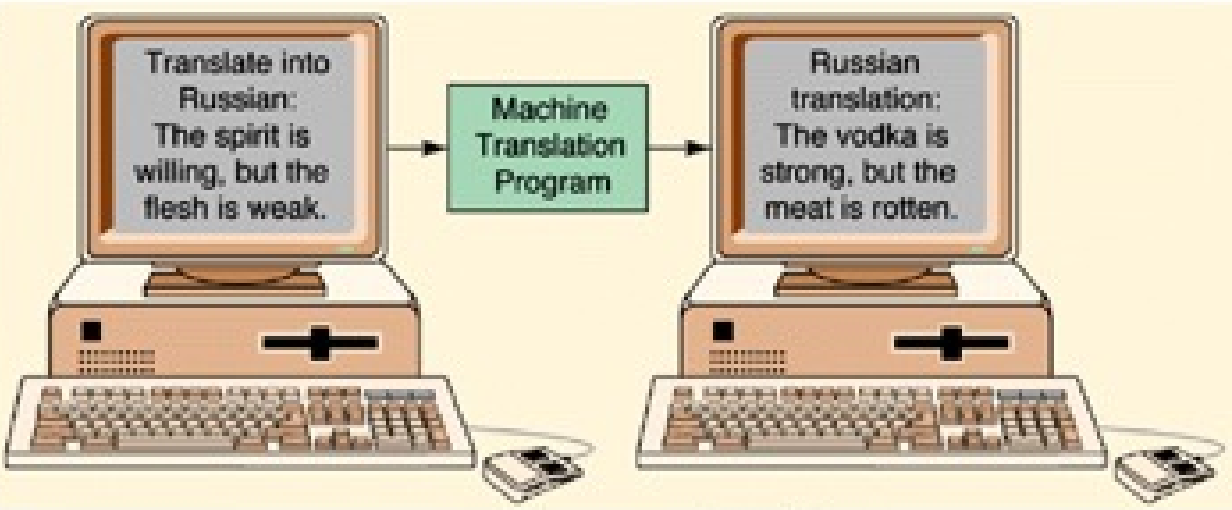


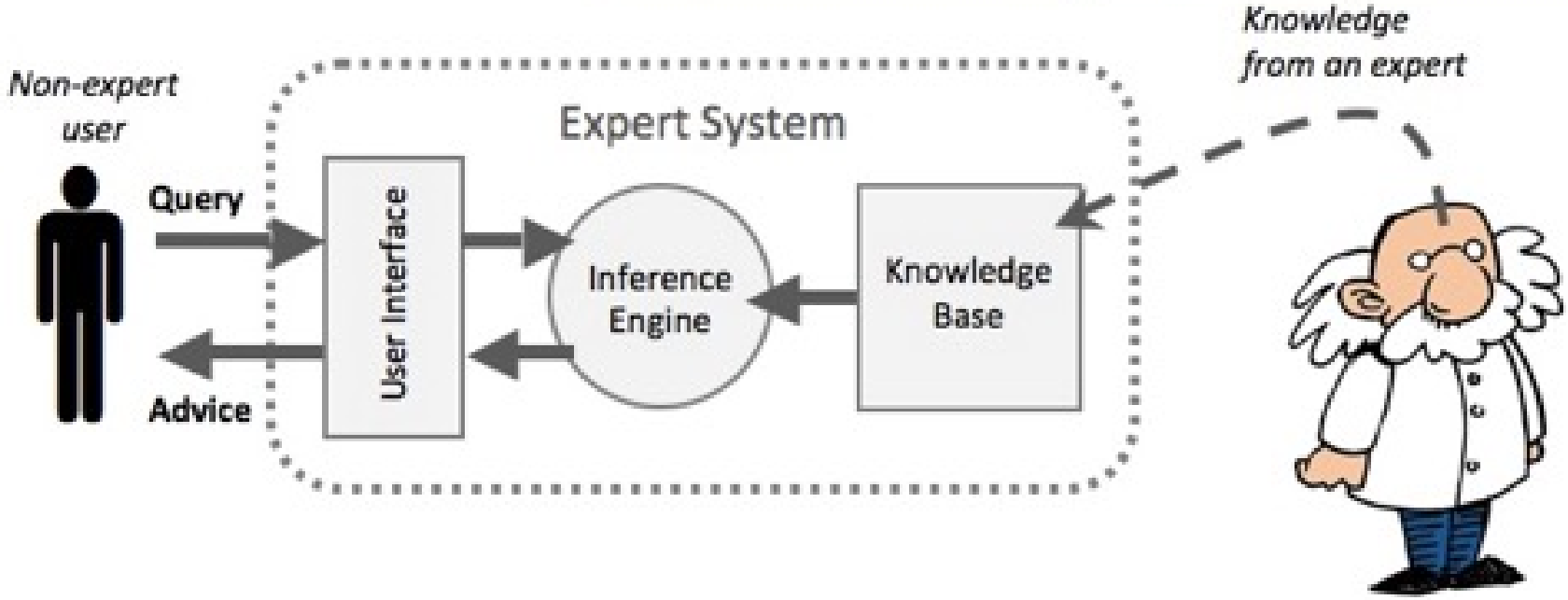
1950's | 1960's | 1970's | 1980's | 1990's | 2000's | 2006's | 2010's | 2012's | 2017's

Failures of the old “AI”… generalization!

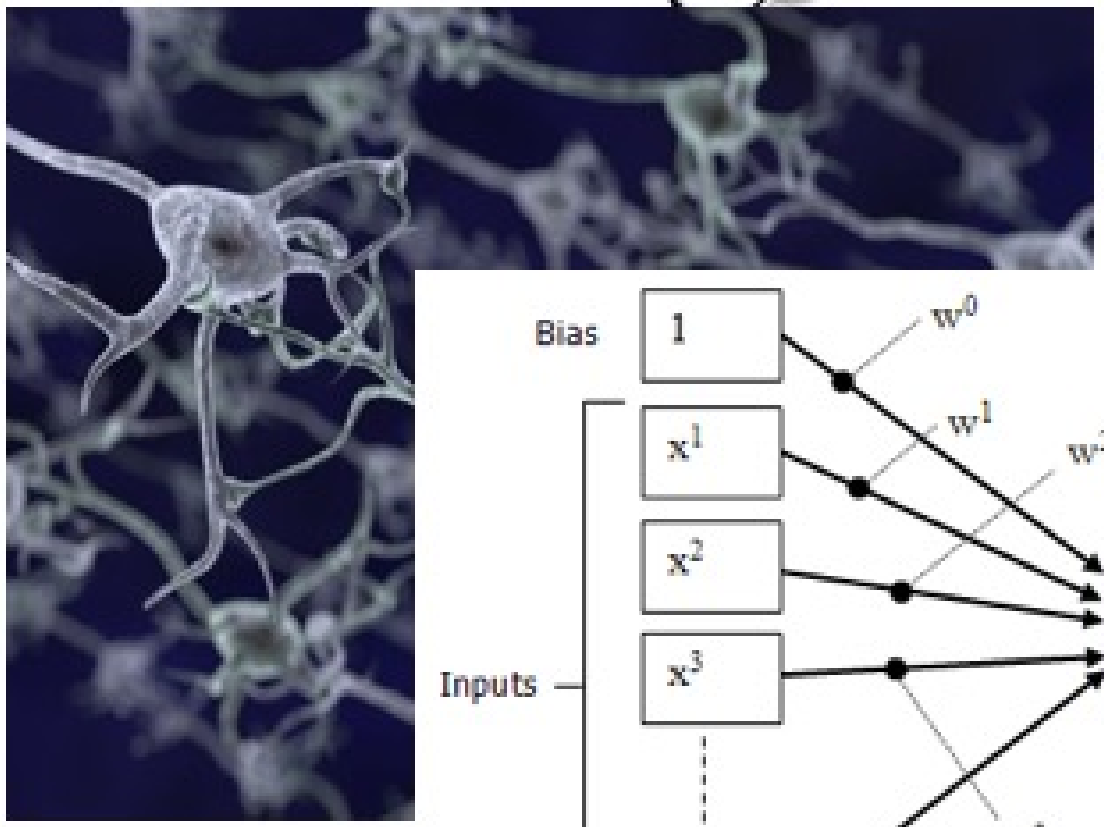


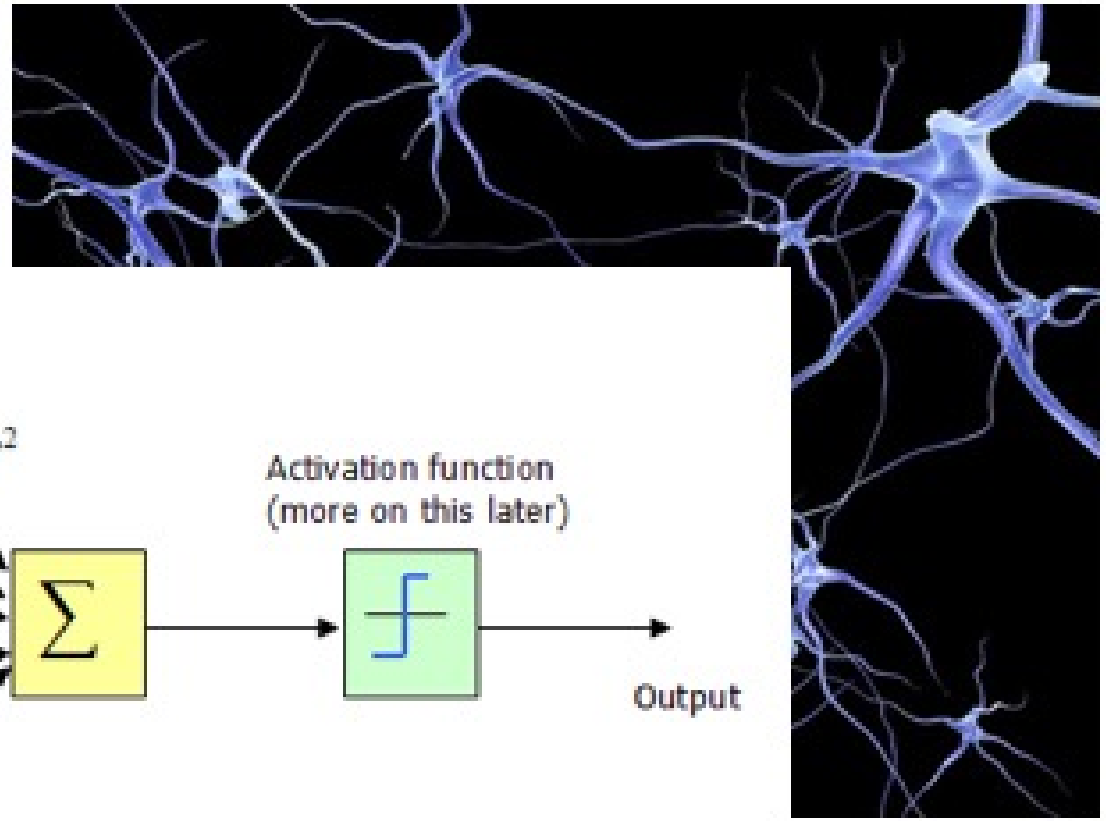


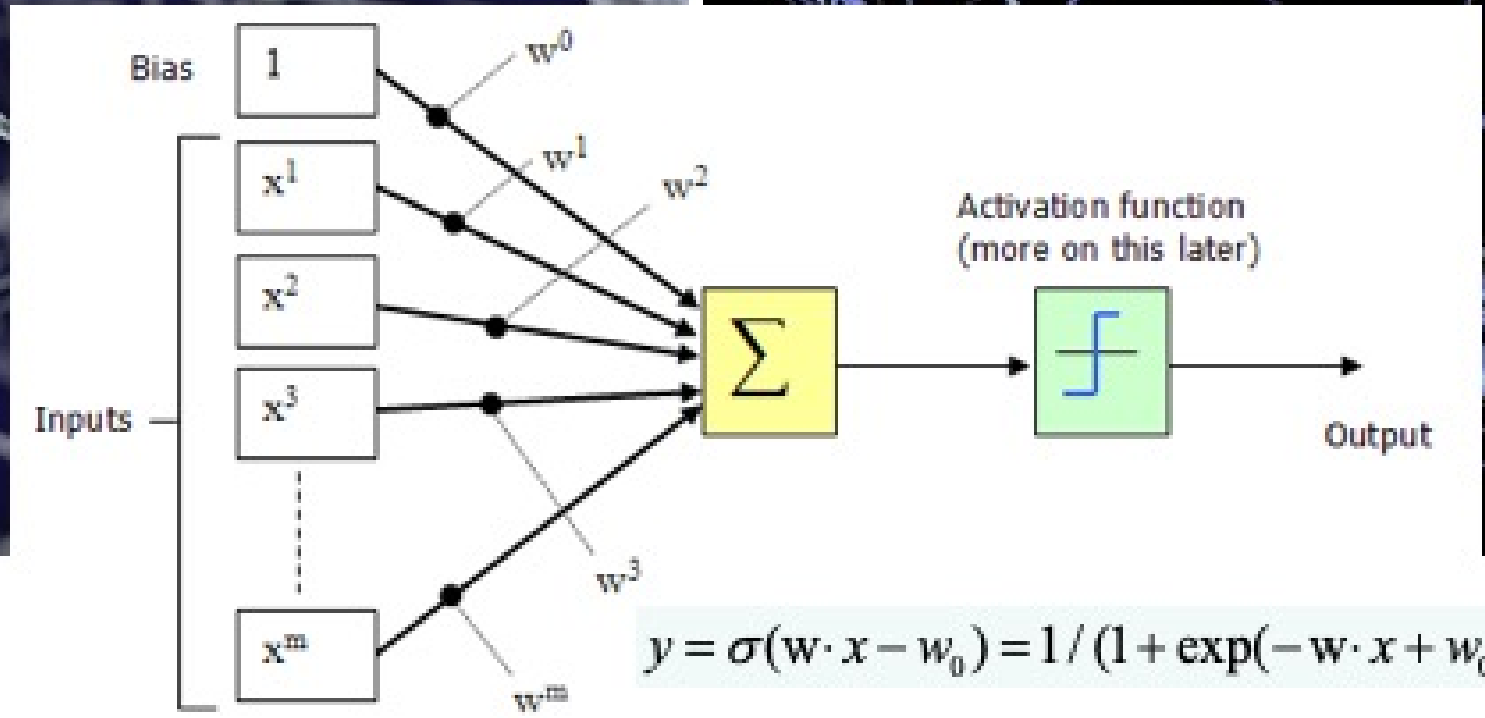




Mimic Biology - Brain like Neurons

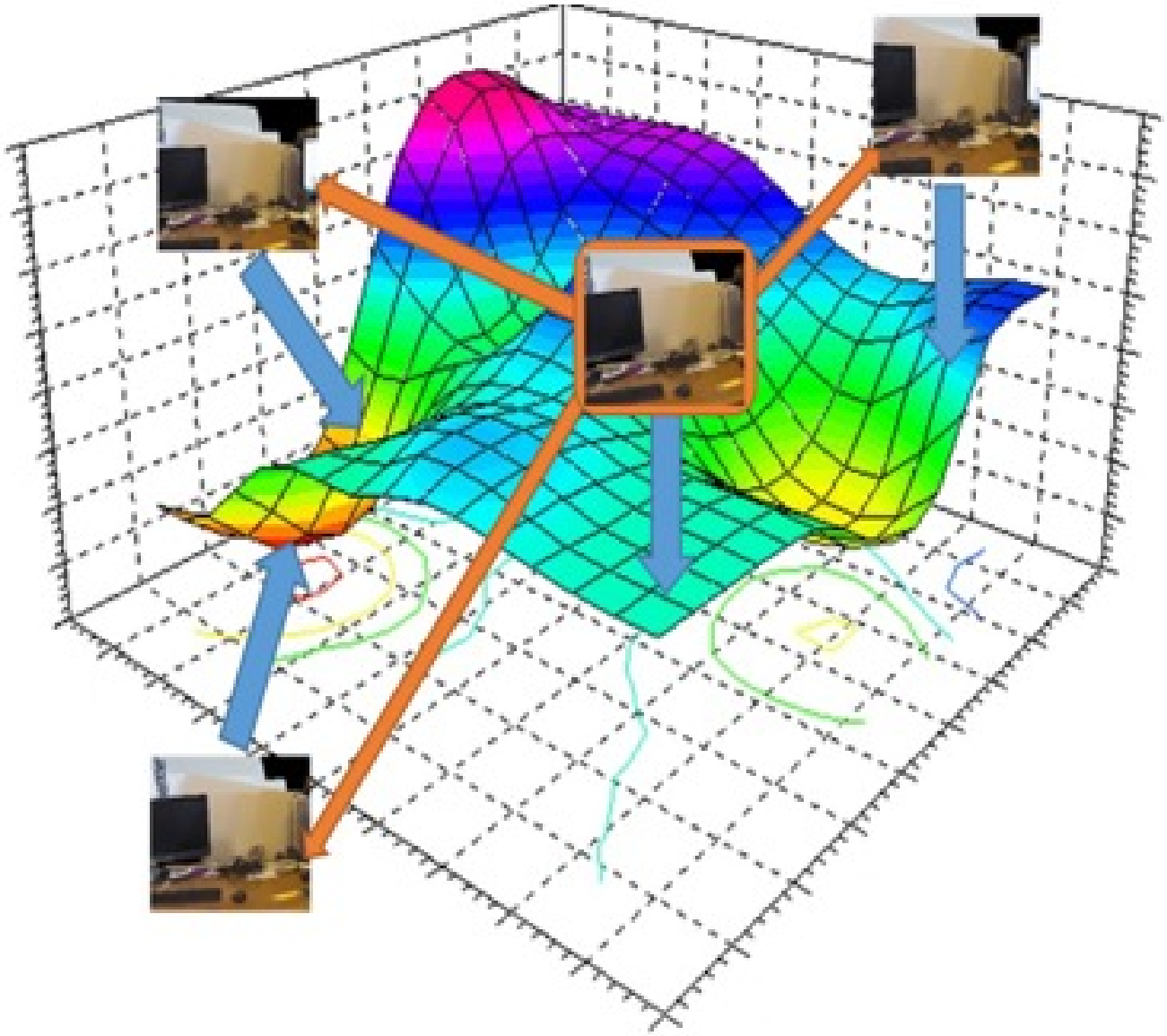


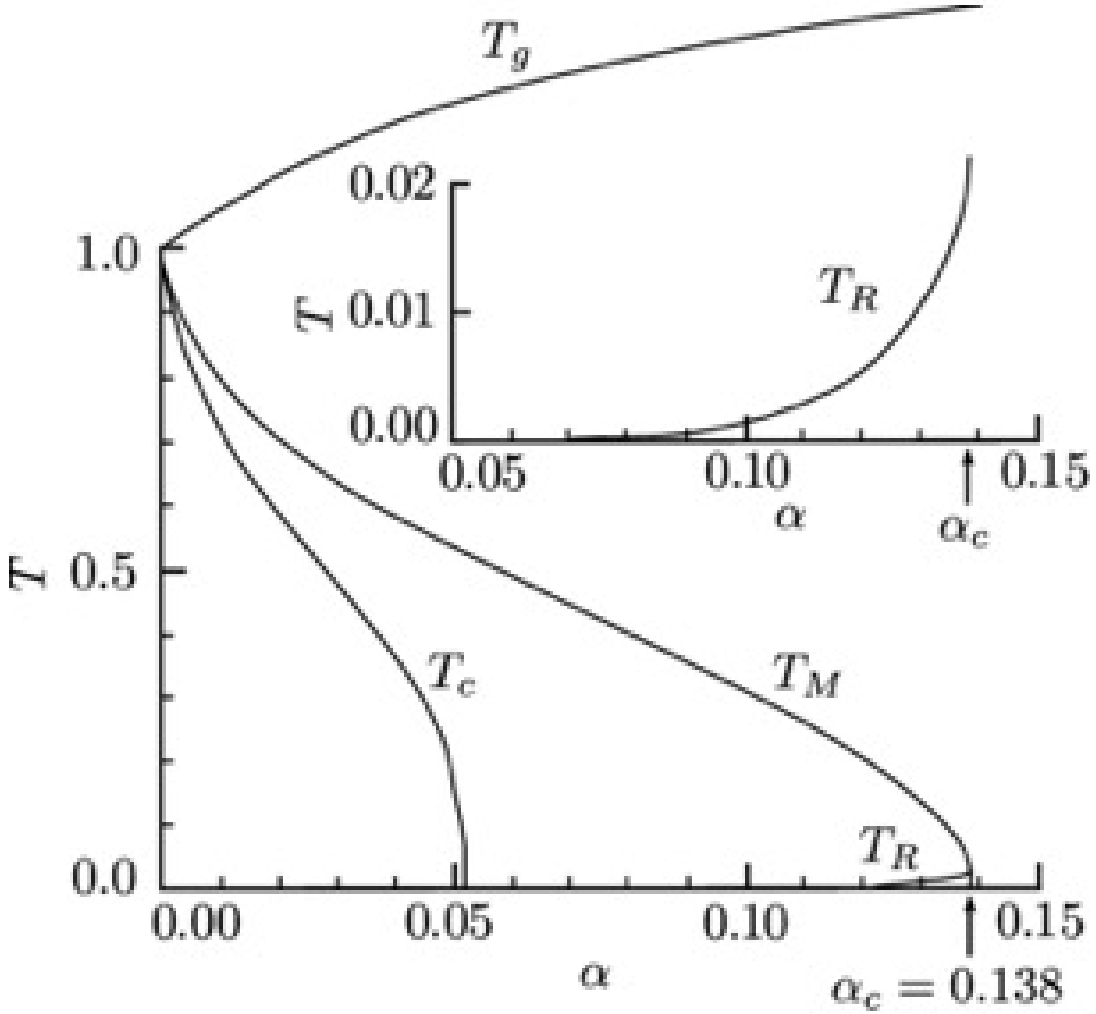




$$y = \sigma(w \cdot x - w_0) = 1/(1+\exp(-w \cdot x + w_0))$$

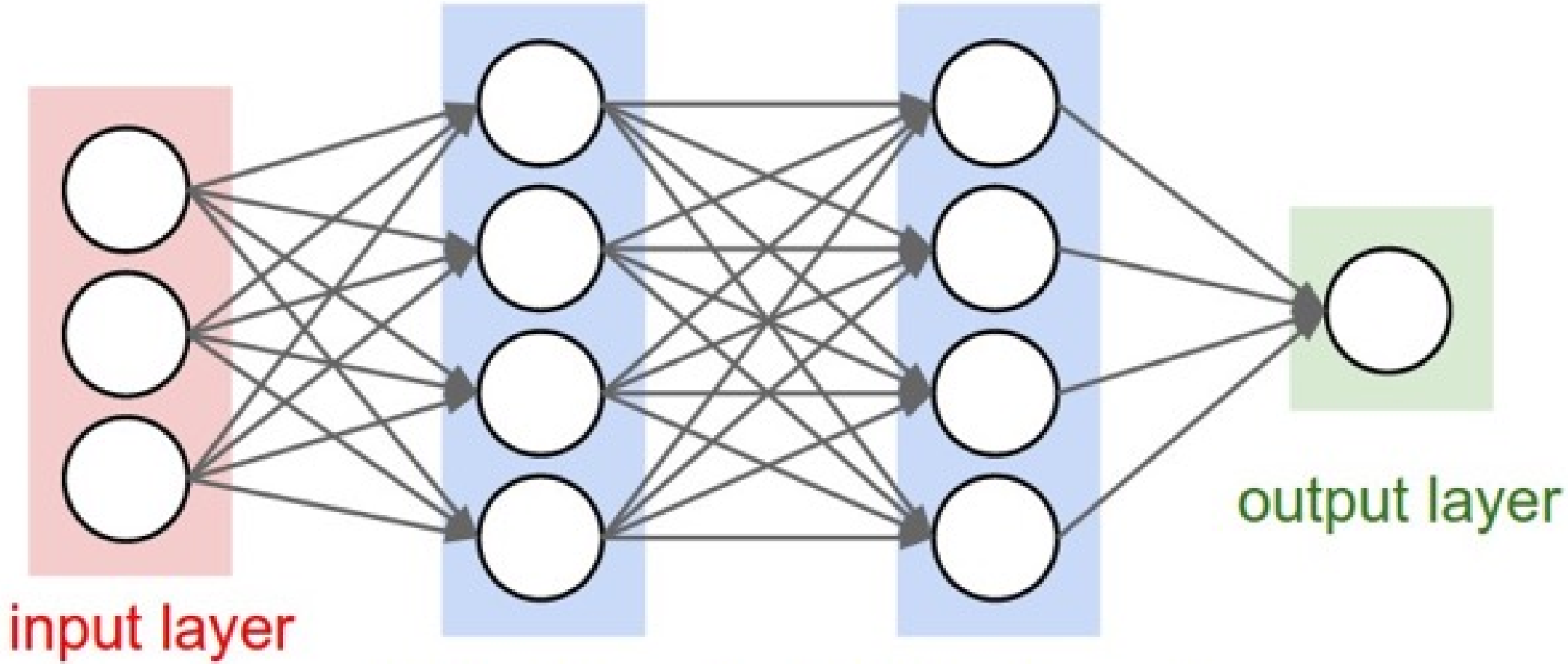
The Hopfield memory model





Amit et.al., 1986

Simple feed forward Neural Networks



Gardner's method

Elizabeth Jane Gardner(1957-1988)

10/26/18

5:07 Mon 8 Oct

journals.aps.org

The space of interactions in neural network models

E Gardner

Department of Physics, Edinburgh University, Mayfield Road, Edinburgh EH9 3JK, UK

Received 13 May 1987, in final form 27 July 1987

Abstract. The typical fraction of the space of interactions between each pair of N Ising spins which solve the problem of storing a given set of p random patterns as N-bit spin configurations is considered. The volume is calculated explicitly as a function of the storage ratio, $\alpha = p/N$, of the value $\kappa(>0)$ of the product of the spin and the magnetic field at each site and of the magnetisation, m. Here m may vary between 0 (no correlation) and 1 (completely correlated). The capacity increases with the correlation between patterns from $\alpha = 2$ for correlated patterns with $\kappa = 0$ and tends to infinity as m tends to 1. The calculations use a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is shown to be locally stable. A local iterative learning algorithm for updating the interactions is given which will converge to a solution of given κ provided such solutions exist.

1. Introduction

There has been a lot of recent interest in McCulloch–Pitts (1943) neural networks (Hebb 1949, Little 1974, Hopfield 1982). Analytic results (Amit *et al* 1985a, b, 1987a, b, Kanter and Sompolinsky 1987, Mézard *et al* 1986, Bruce *et al* 1987, Gardner 1986) have been obtained for thermodynamic and dynamical quantities using particular

Phase Transitions in Learning Curves

Rigorous Learning Curve Bounds from Statistical Mechanics

David Haussler
U.C. Santa Cruz
Santa Cruz, California

Michael Kearns
AT&T Bell Laboratories
Murray Hill, New Jersey

H. Sebastian Seung
AT&T Bell Laboratories
Murray Hill, New Jersey

Naftali Tishby
Hebrew University
Jerusalem, Israel

Abstract

In this paper we introduce and investigate a mathematically rigorous theory of learning curves that is based on ideas from statistical mechanics. The advantage of our theory over the well-established Vapnik-Chervonenkis theory is that our bounds can be considerably tighter in many cases, and are also more reflective of the true behavior (functional form) of learning curves. This behavior can often exhibit dramatic properties such as phase transitions, as well as power law asymptotics not explained by the VC theory. The disadvantages of our theory are that its application requires knowledge of the input distribution, and it is limited so far to finite cardinality function classes.

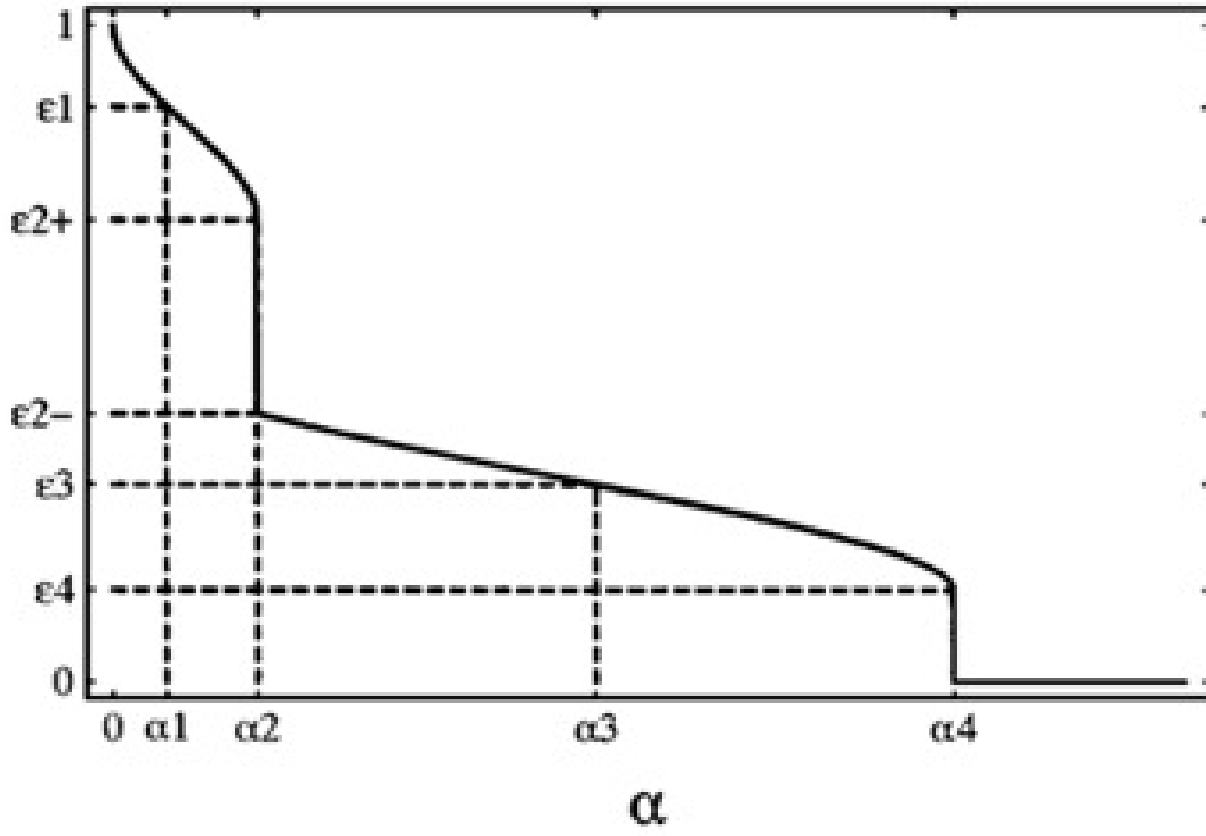
We illustrate our results with many concrete examples of learning curve bounds derived from our theory.

1 Introduction

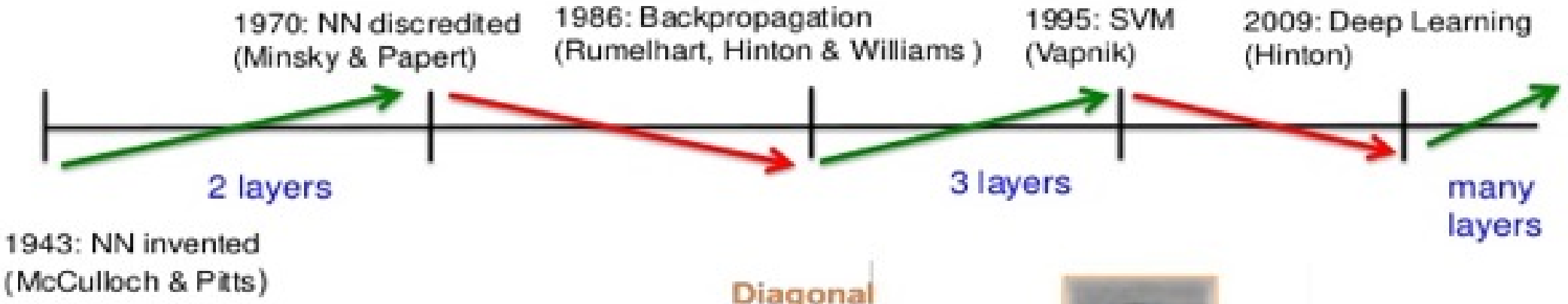
According to the Vapnik-Chervonenkis (VC) theory of learning curves [31, 30], minimizing empirical error within a function class F on a random sample of m examples leads to

However, it is becoming clear that learning curves exhibit a diversity of behaviors. For instance, some researchers have attempted to fit learning curves from backpropagation experiments with a variety of functional forms, including exponentials [4]. Backpropagation experiments with handwritten digits and characters indicate that good generalization error is sometimes obtained for sample sizes considerably smaller than the number of weights (presumed to be roughly the same as the VC dimension) [20], though the VC bounds are vacuous for m smaller than d. Discrepancies between the VC bounds and actual learning curve behavior have also been pointed out and analyzed in other machine learning work [33, 21].

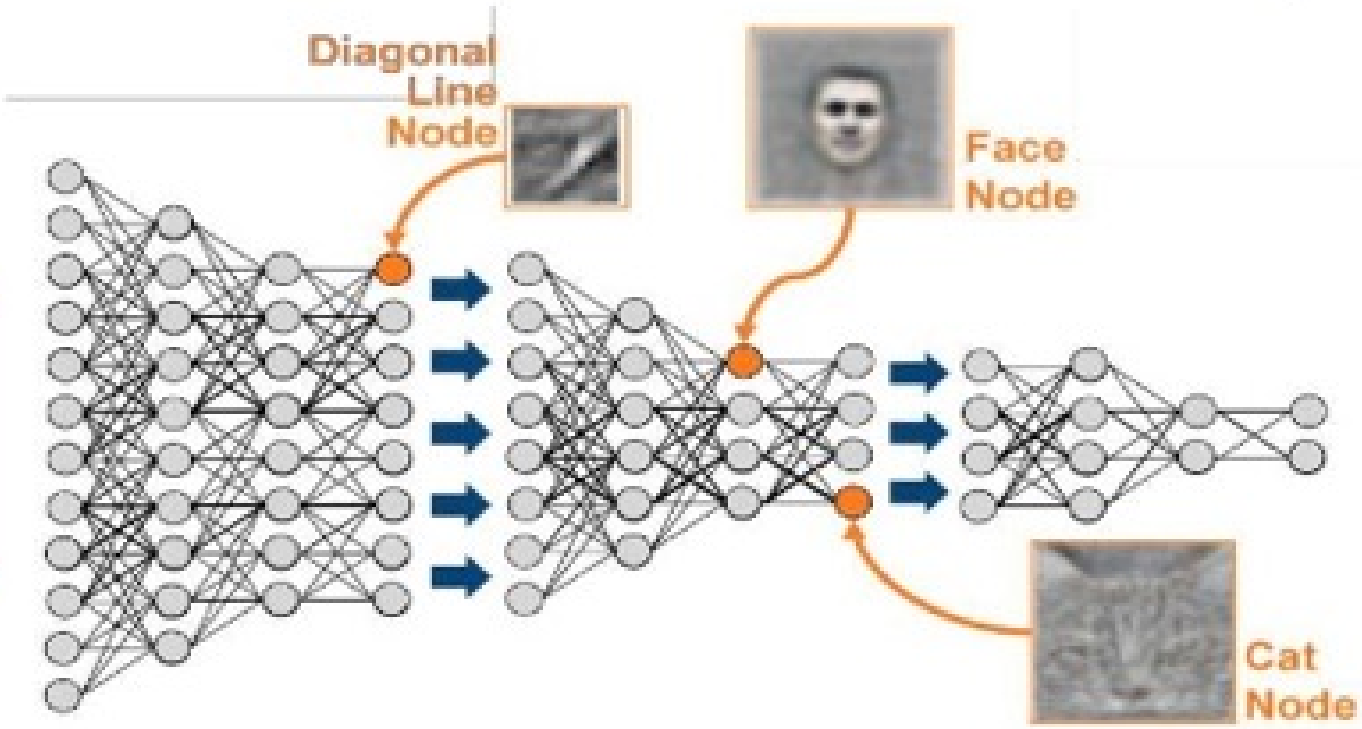
Of course, the VC bounds might simply be inapplicable to these experiments, because backpropagation is not equivalent to empirical error minimization. Vapnik has conjectured that backpropagation can access only a limited portion of the function space, so that the "effective dimension" is much smaller than the VC dimension. According to this type of reasoning, learning curves are heavily affected by the specifics of the algorithm. Another possibility is that the VC bounds are applicable, but sometimes fail to capture the true behavior of particular learning curves because



Deep Learning: Neural-Nets strike back

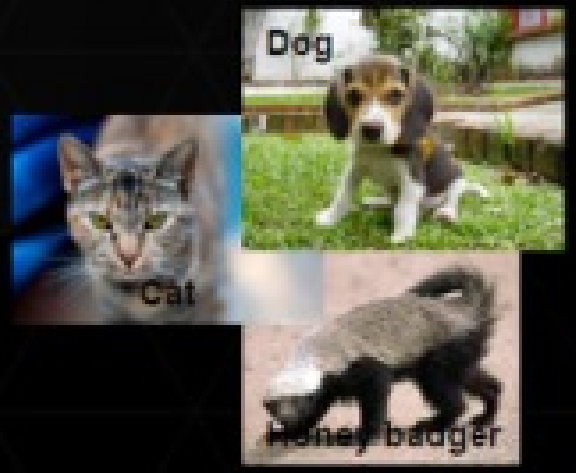


-Model Size: 10B parameters
-Used by: Yahoo!, Google, Microsoft, Baidu, IBM, Scyfer ☺



DEEP LEARNING APPROACH

Train:



Errors

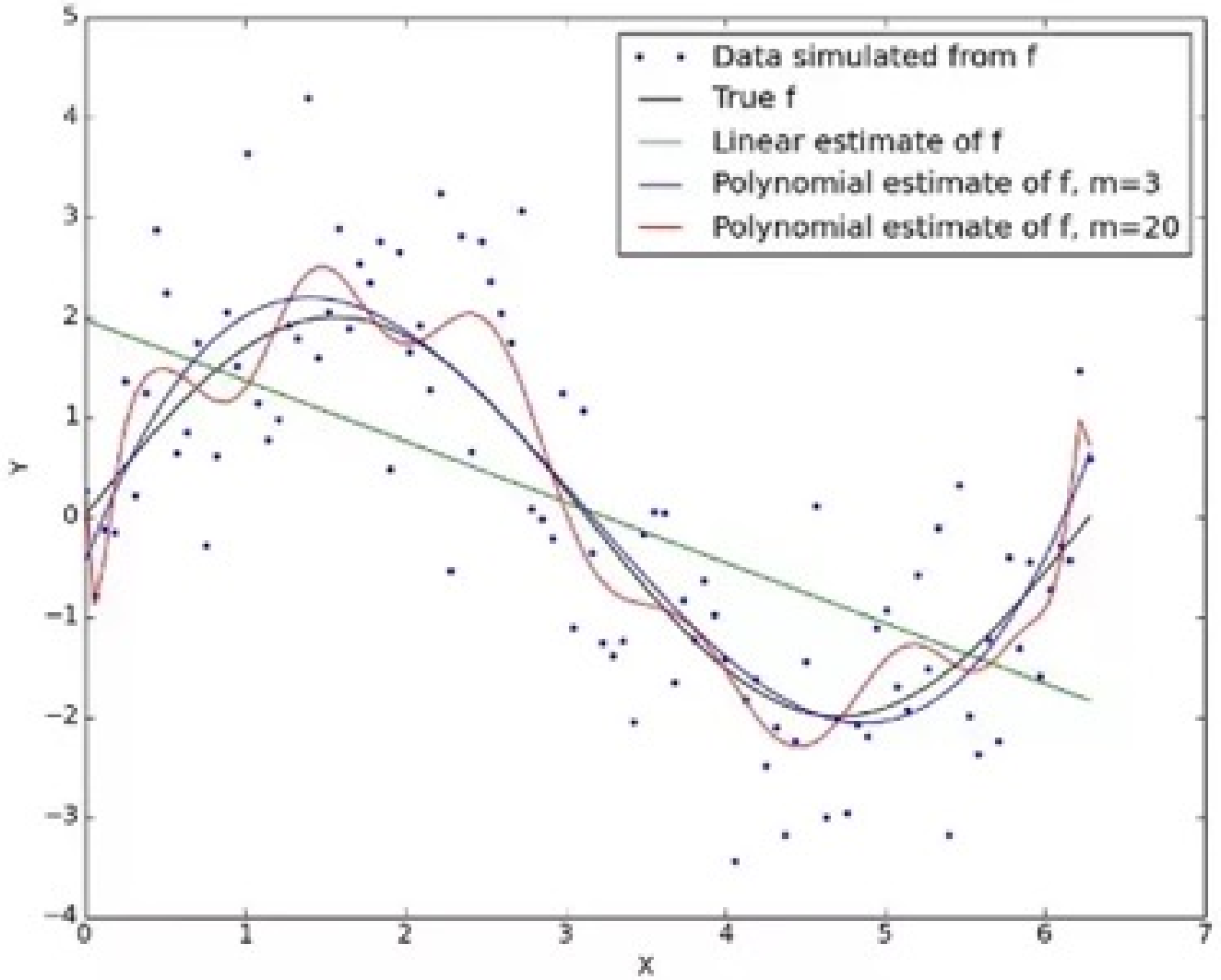
Dog
Cat
Raccoon

Surprise!

Deploy:

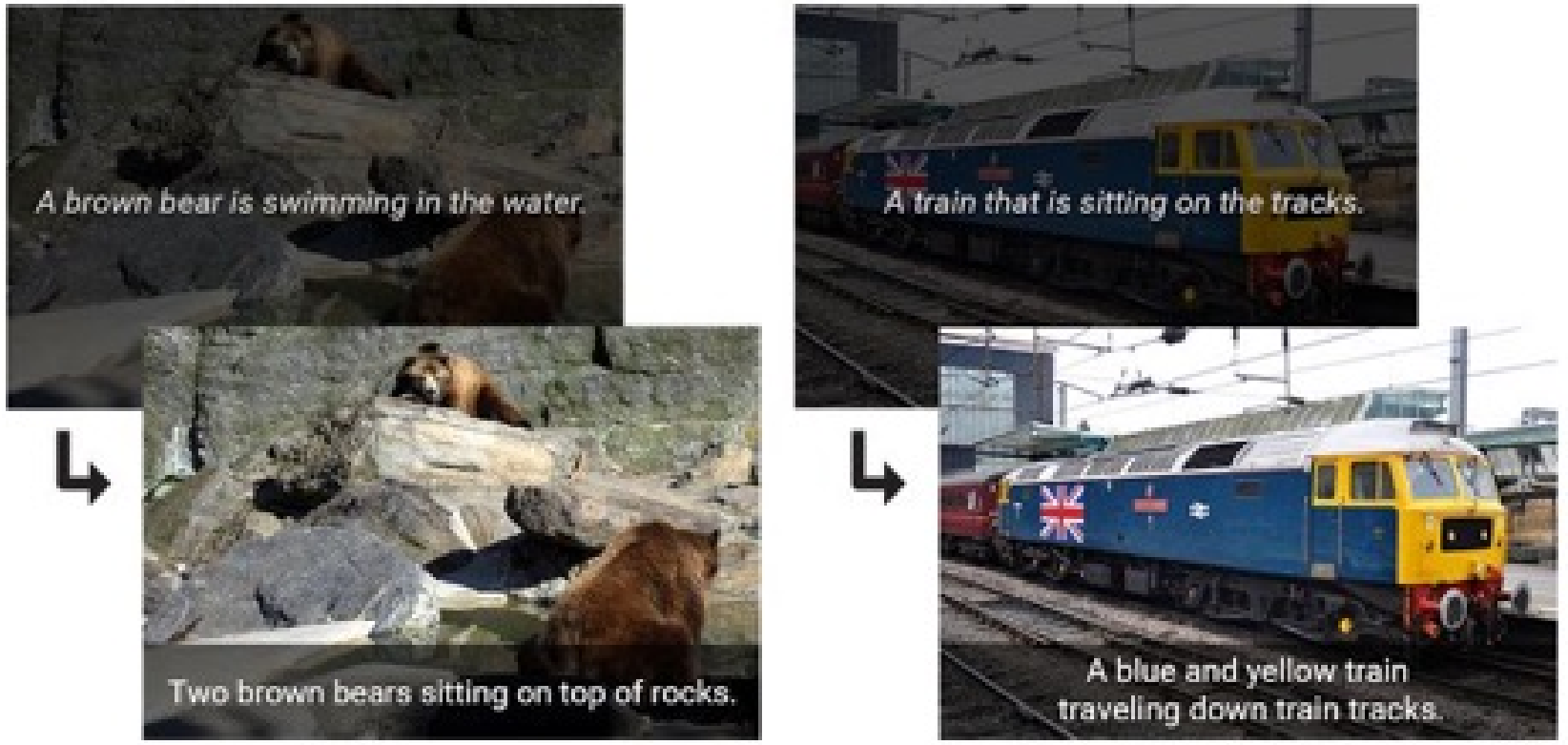
Dog

Is it more than high-dimensional curve fitting?



- Classical [least-squares] regression
- Main issue: **generalization**
- How to avoid **over-fitting?**
- # of data points ~ # of parameters
- Choose the right hypotheses class!

Scene understanding



Rethinking Learning Theory

"Old" Generalization bounds:

$$\varepsilon^2 < \frac{\log|H_\varepsilon| + \log 1/\delta}{2m}$$

ε - generalization error

δ - confidence

m - number of training examples

H_ε - ε-cover of the Hypothesis class

typically we assume: $|H_\varepsilon| \sim \left(\frac{1}{\varepsilon}\right)^d$

d - the class (VC,...) dimension

... Don't work for Deep Learning!

Higher expressivity - worse bound!

New: Typical Problem dependent:

$$|H_\varepsilon| \sim 2^{|X|} \to 2^{|T_\varepsilon|}$$

T_ε - ε-partition of the input variable X

with respect to the distortion:

$$d_{IB}(x,t) = D[p(y|x) \| p(y|t)]$$

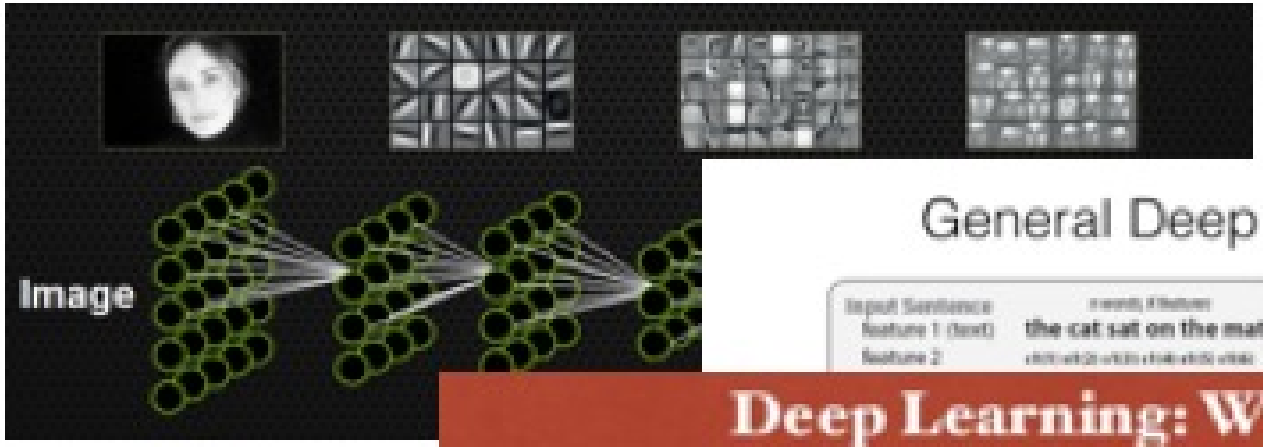
$$\geq 1/2\ln 2 \, \| p(y|x) - p(y|t) \|_1^2$$

when $p(y|t) = \sum_x p(y|x)p(x|t)$

$$\langle d_{IB} \rangle = I(X;Y) - I(T;Y)$$

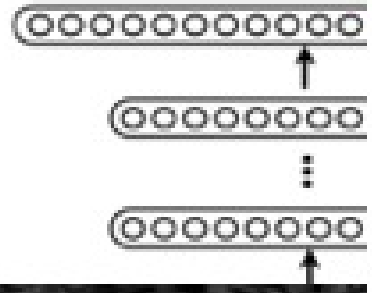
small IB distortion , or high $I(T;Y)$,

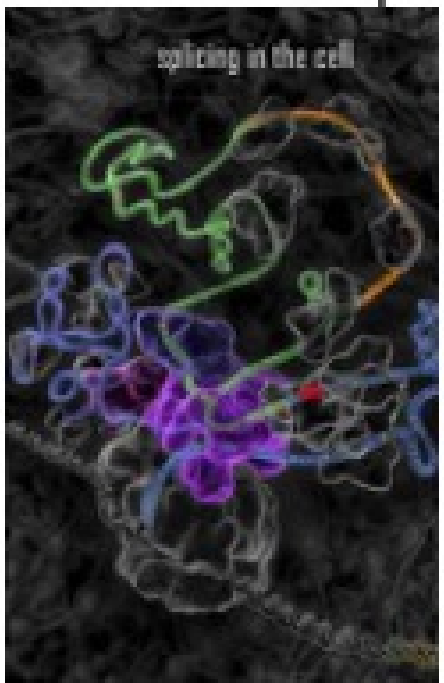
$\Rightarrow$ small [typical] generalization error



General Deep Architecture for NLP

Deep Learning: Why for NLP ?



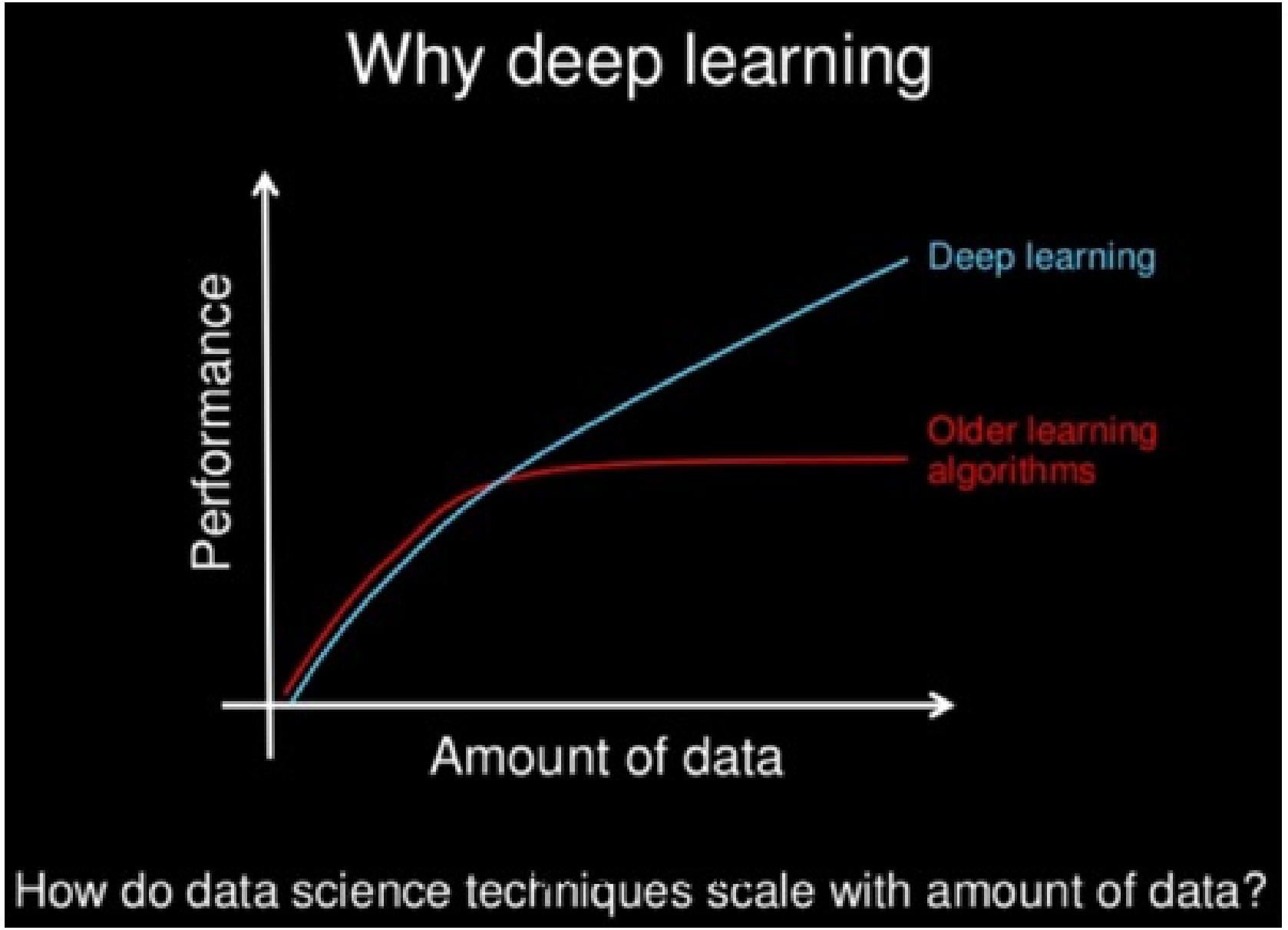


One Model rules them all ?

DL approaches have been successfully applied to:

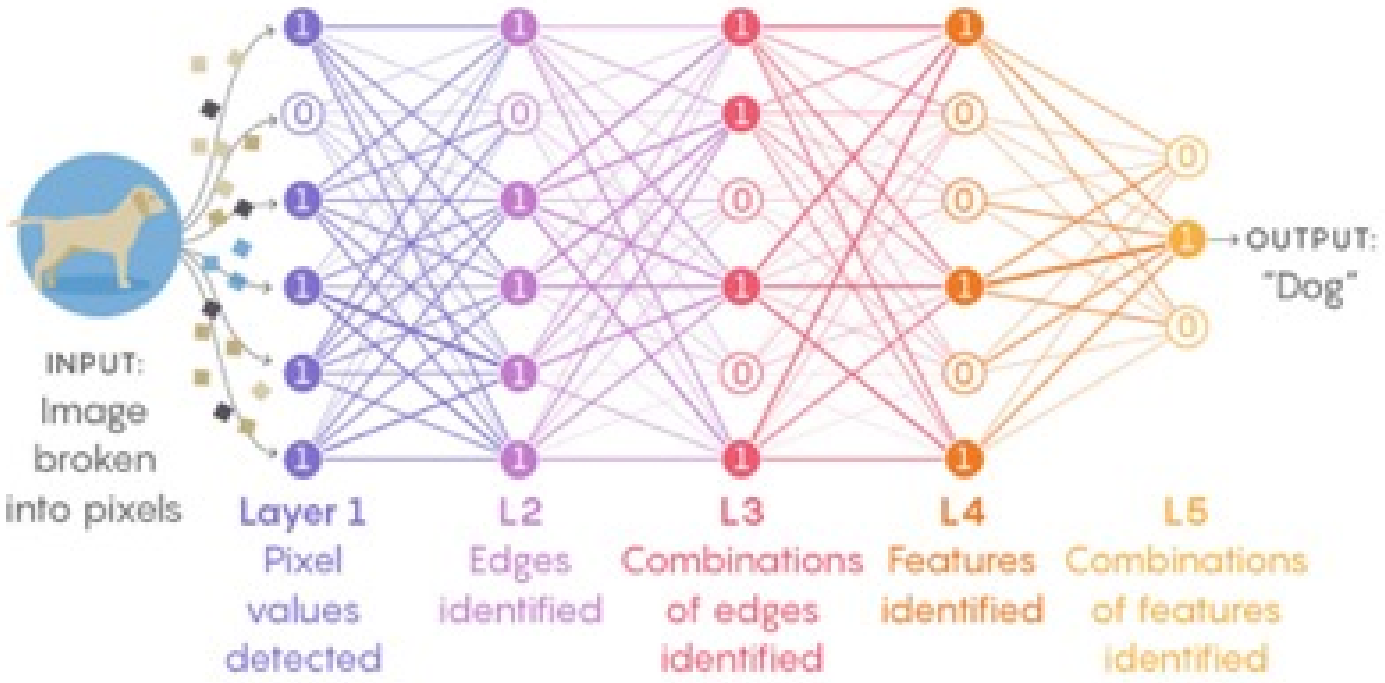
- Automatic summarization
- Coreference resolution
- Discourse analysis
- Machine translation
- Morphological segmentation
- Named entity recognition (NER)
- Natural language generation
- Word sense disambiguation
- Relationship extraction
- Speech processing
- Part-of-speech tagging
- sentence boundary disambiguation
- Sentiment analysis
- Optical character recognition (OCR)
- Question answering
- Parsing
- Word segmentation
- Natural language understanding
- Information retrieval (IR)
- Speech recognition
- Topic segmentation and recognition
- Speech segmentation
- Information extraction (IE)

Deep learning & “big data”



Learning From Experience

Deep neural networks learn by adjusting the strengths of their connections to better convey input signals through multiple layers to neurons associated with the right general concepts.



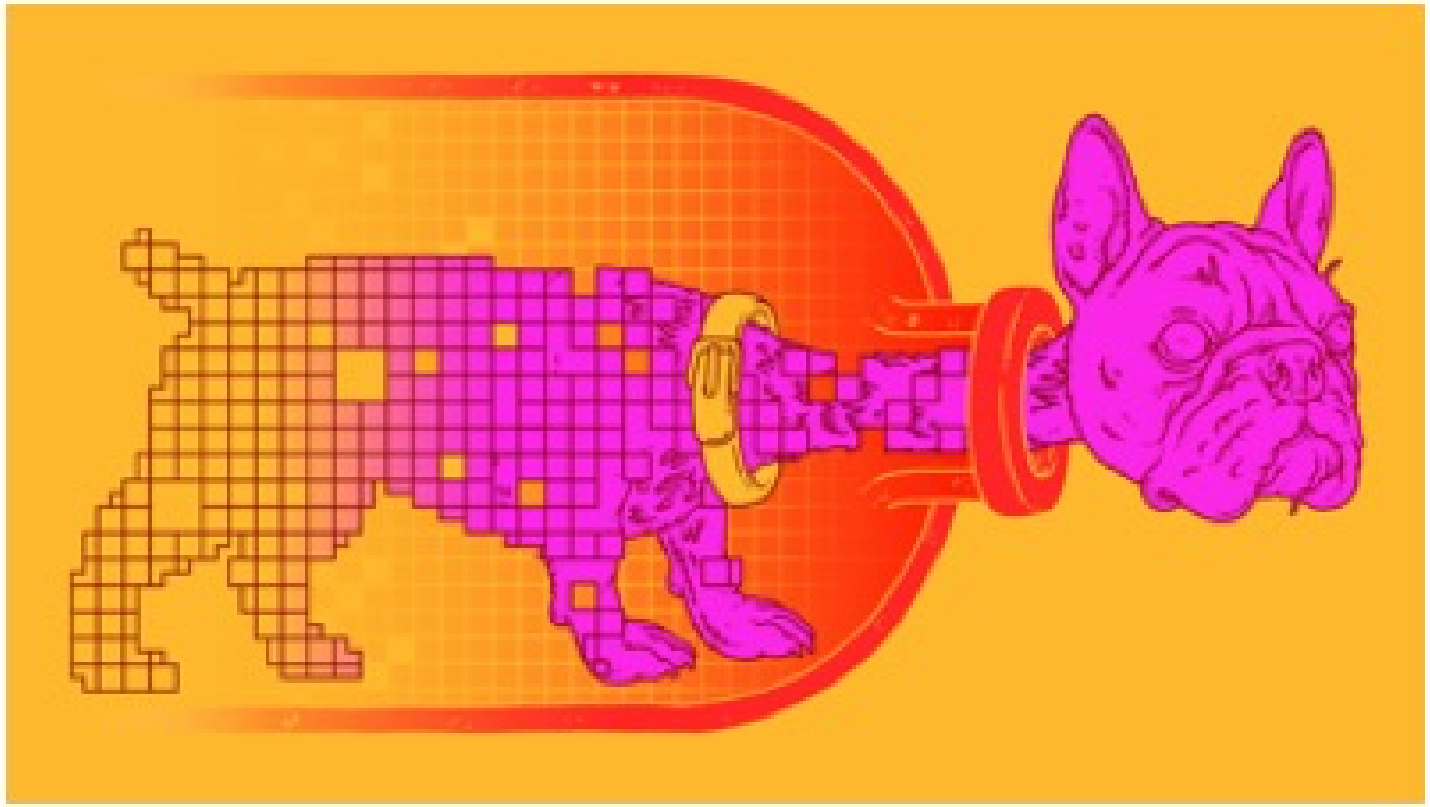
When data is fed into a network, each artificial neuron that fires (labeled "1") transmits signals to certain neurons in the next layer, which are likely to fire if multiple signals are received. The process filters out noise and retains only the most relevant features.

The main part of Learning is… forgetting

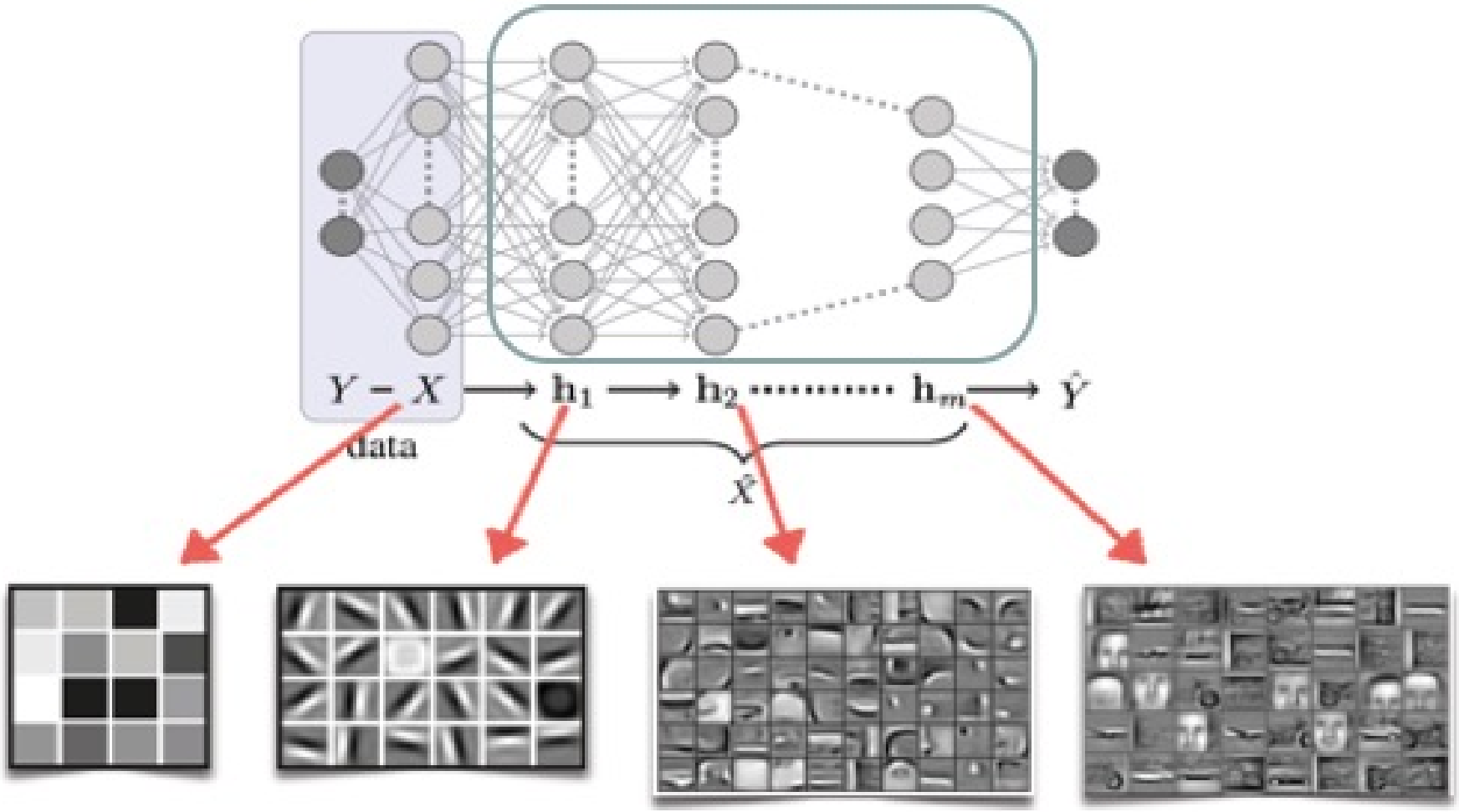
Quanta Magazine: Sep. 2017

New Theory Cracks Open the Black Box of Deep Learning

A new idea called the "information bottleneck" is helping to explain the puzzling success of today's artificial-intelligence algorithms — and might also explain how human brains learn.



Deep Neural Nets and Information



Concentration of Mutual Information

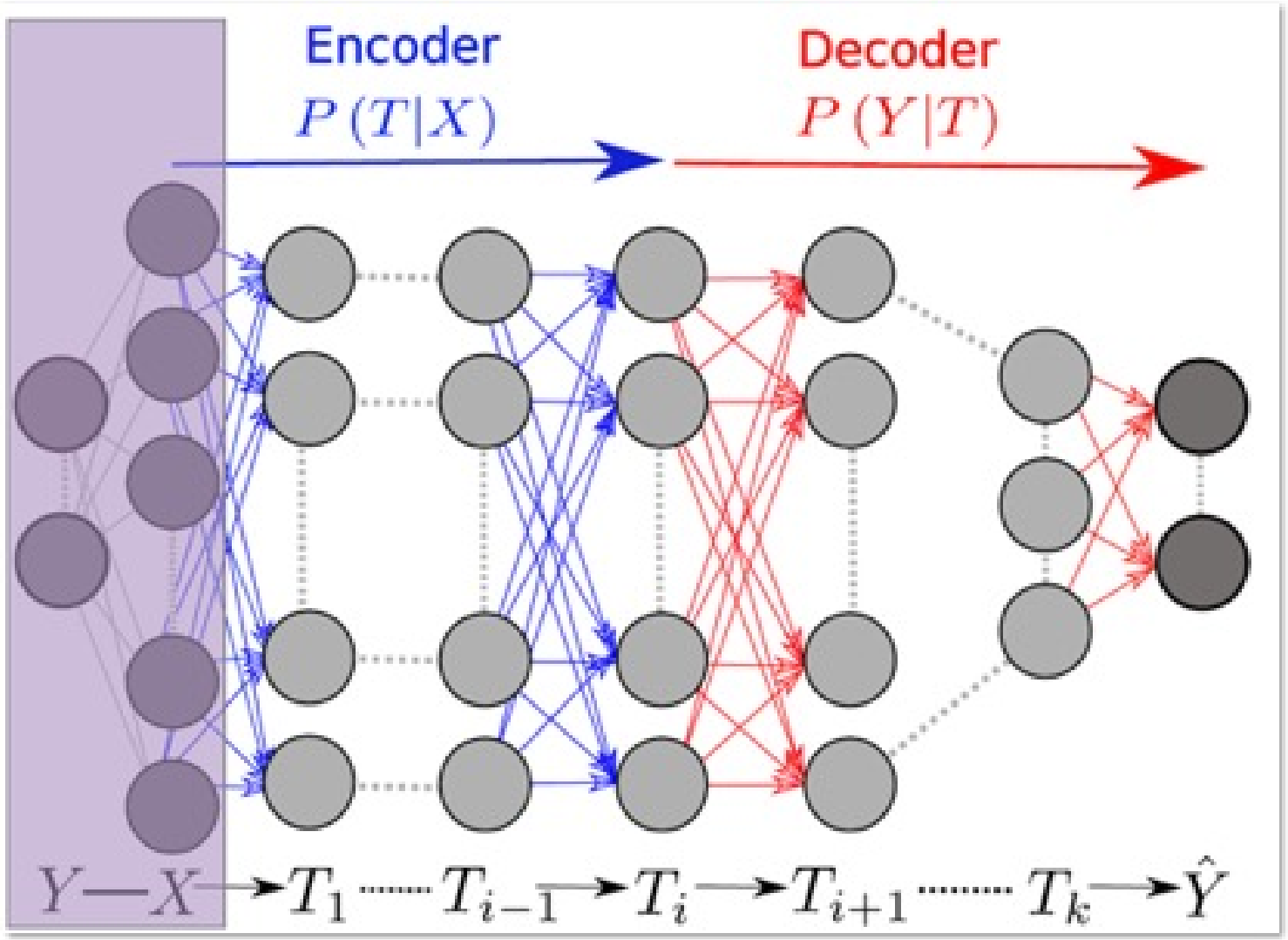
$$I(X;T)=\left\langle \log \frac{p(x|t)}{p(x)} \right\rangle_{X,T} = \left\langle \log \prod_i \frac{p(x_i|Pa(x_i),t)}{p(x_i|Pa(x_i))} \right\rangle_{X,T} = \left\langle \sum_i \log \frac{p(x_i|Pa(x_i),t)}{p(x_i|Pa(x_i))} \right\rangle_{X,T}$$

$$I(T;Y)=\left\langle \log \sum_x p(y|x)p(x|t) - \log p(y) \right\rangle_{Y,T} = \left\langle \log \left[\sum_x p(y|x) \prod_i p(x_i|Pa(x_i),t) \right] - \log p(y) \right\rangle_{Y,T}$$

Proposition:

1. Both I(T;X) and I(T;Y), as defined, concentrate, uniformly, under the partition typicality assumption.
2. Both can be estimated uniformly well (over the partitions) from a sample of p(X,Y).

Each layer is characterized by its Encoder & Decoder Information

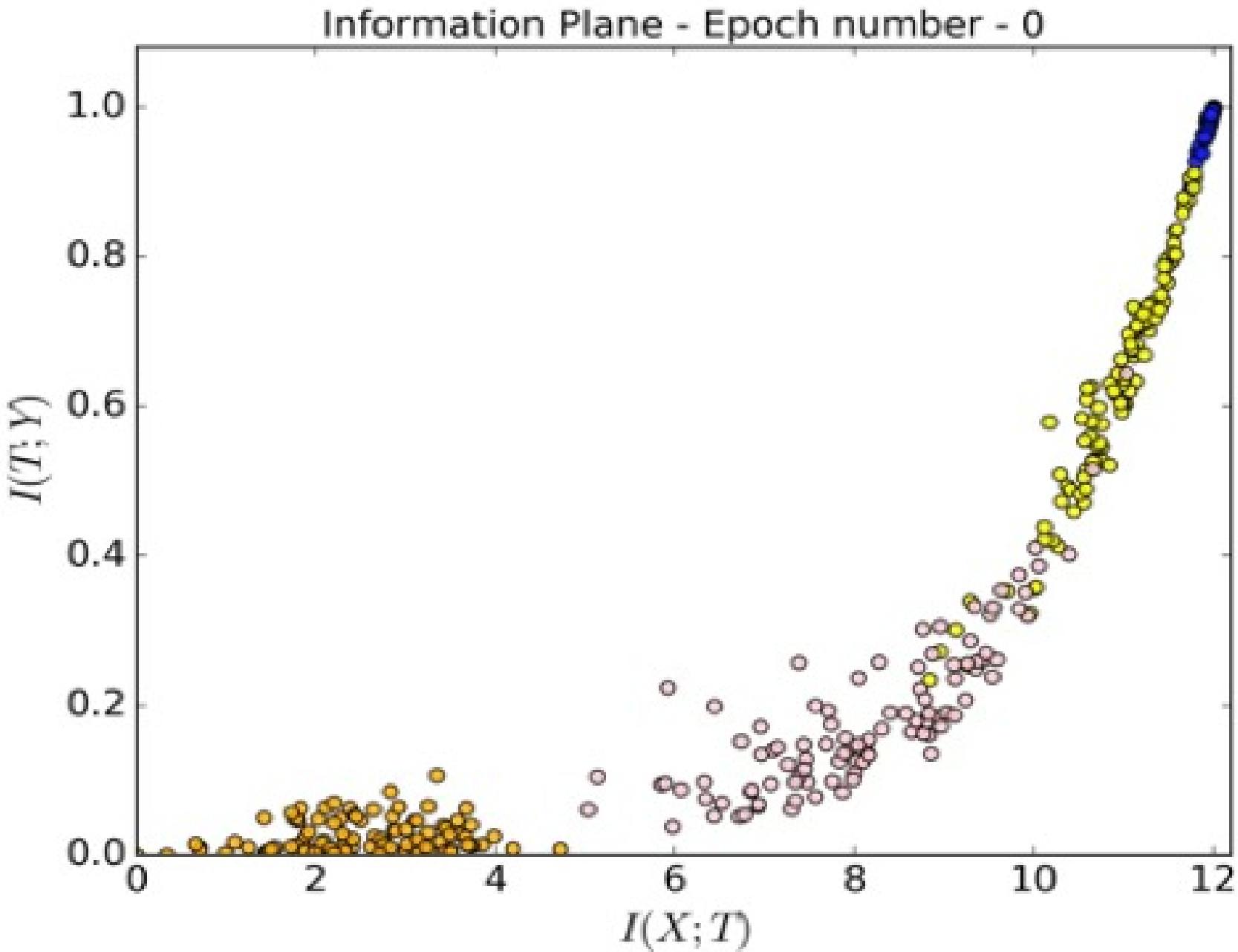


Theorem (Information Plane):

*For large typical **X**, the sample complexity of a DNN is completely determined by the encoder mutual information, **I(X;T)**, of the last hidden layer; the accuracy (generalization error) is determined by the decoder information, **I(T;Y)**, of the last hidden layer.*

The complexity of the problem shifts from the decoder to the encoder, across the layers…

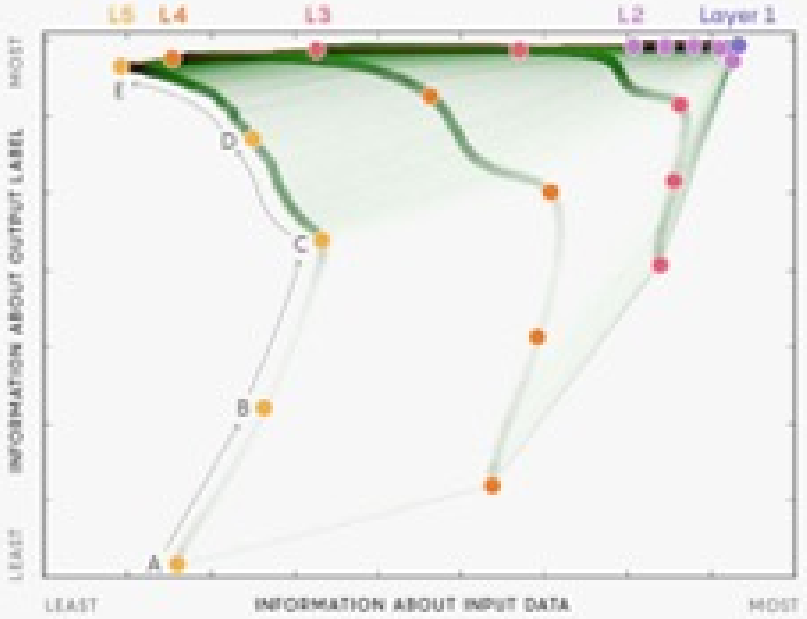
100 DNN Layers in Info-Plane without averaging



How does it work?

Inside Deep Learning

New experiments reveal how deep neural networks evolve as they learn.



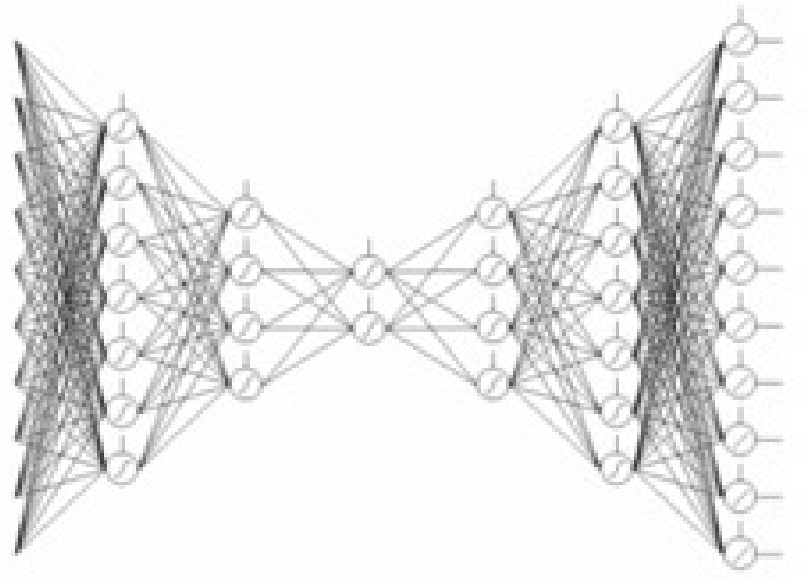
A INITIAL STATE: Neurons in Layer 1 encode everything about the input data, including all information about its label. Neurons in the highest layers are in a nearly random state bearing little to no relationship to the data or its label.

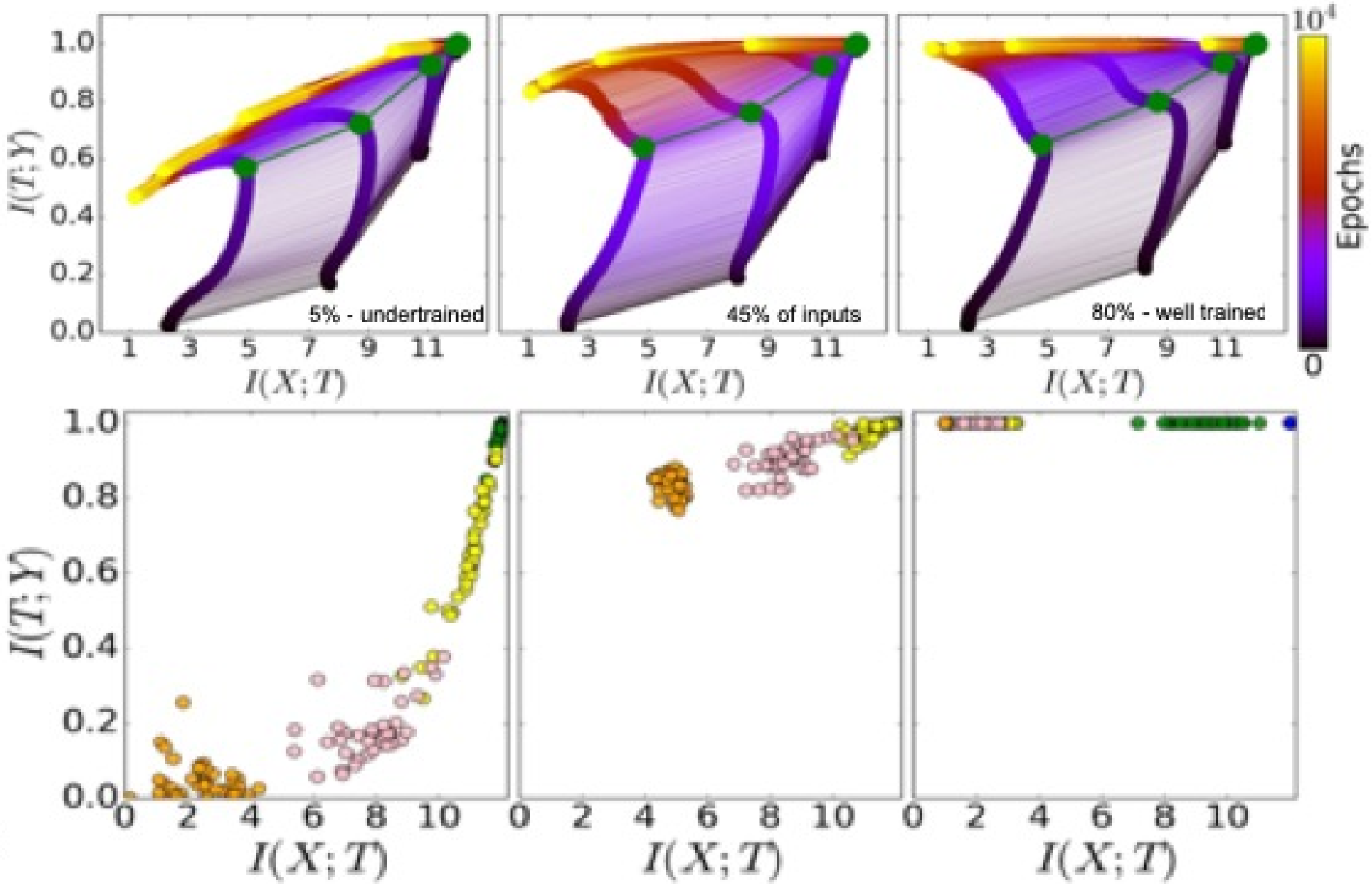
B FITTING PHASE: As deep learning begins, neurons in higher layers gain information about the input and get better at fitting labels to it.

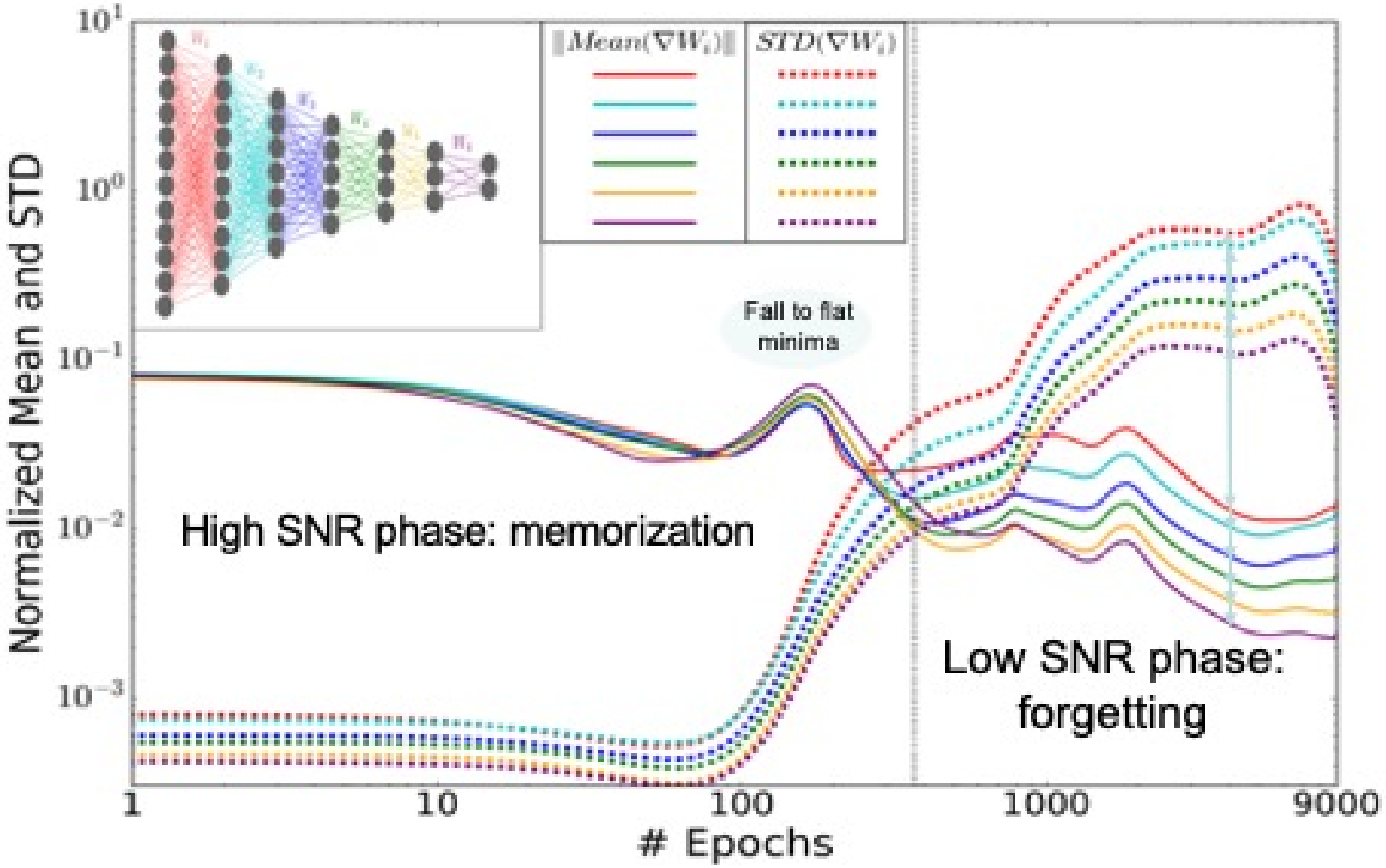
C PHASE CHANGE: The layers suddenly shift gears and start to "forget" information about the input.

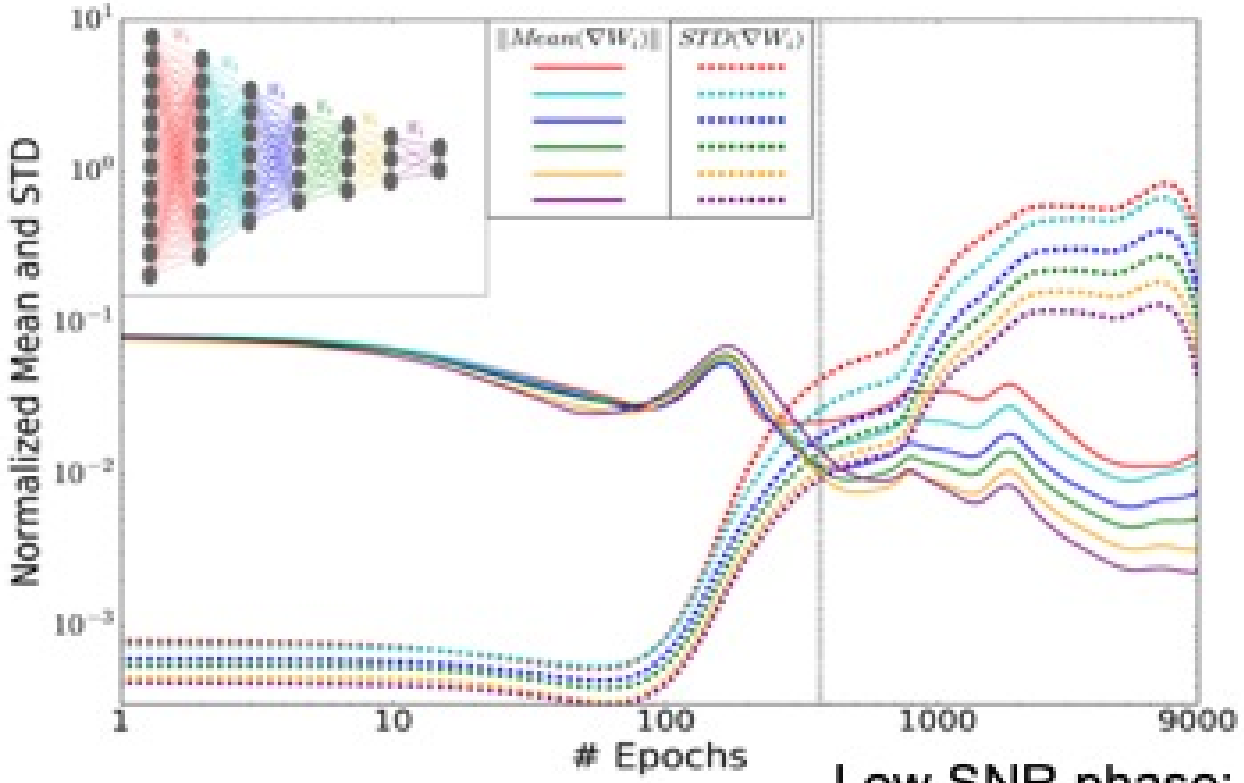
D COMPRESSION PHASE: Higher layers compress their representation of the input data, keeping what is most relevant to the output label. They get better at predicting the label.

E FINAL STATE: The last layer achieves an optimal balance of accuracy and compression, retaining only what is needed to predict the label.



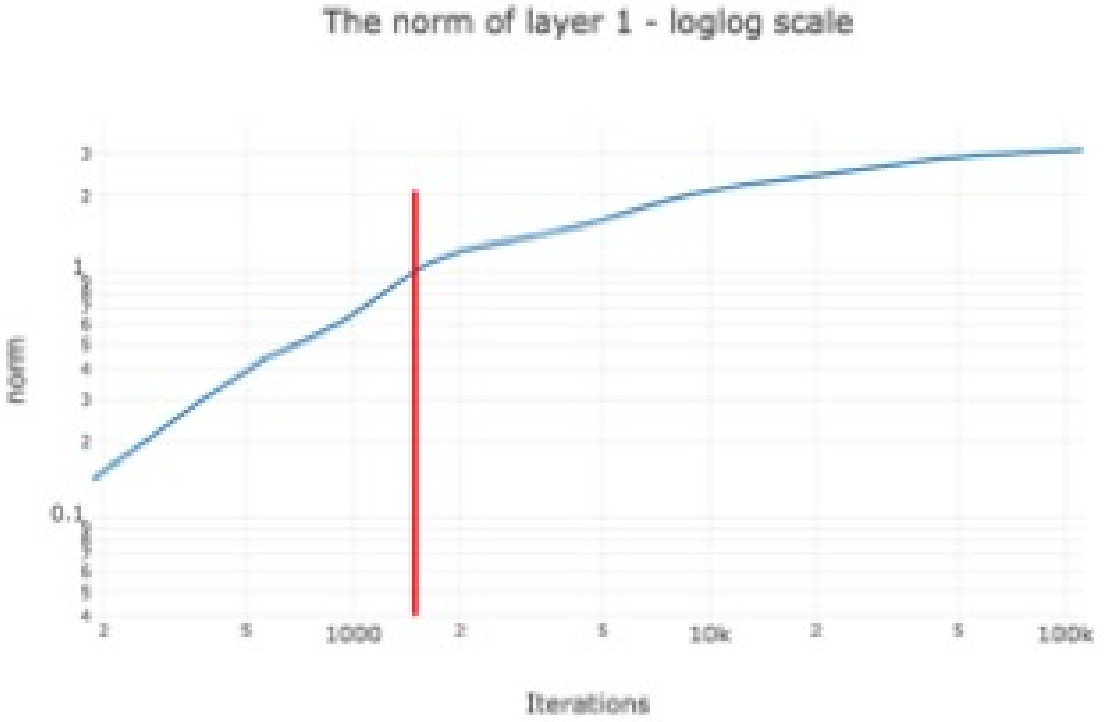

5% - undertrained
45% of inputs
80% - well trained
$I(T;Y)$
$I(X;T)$
Epochs
10^4
0
1.0
0.8
0.6
0.4
0.2
0.0
1 3 5 7 9 11
0 2 4 6 8 10


$\|Mean(\nabla W_i)\|$
$STD(\nabla W_i)$
Fall to flat minima
High SNR phase: memorization
Low SNR phase: forgetting
Normalized Mean and STD
Epochs
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
1
10
100
1000
9000



High SNR phase: memorization

Low SNR phase: forgetting



High SNR phase: Linear drift

Low SNR phase: Diffusion

In the noisy phase the weights diffuse and grow like $O(\sqrt{t})$

The Information Bottleneck optimality bound

(Tishby, Pereira, Bialek, 1999)

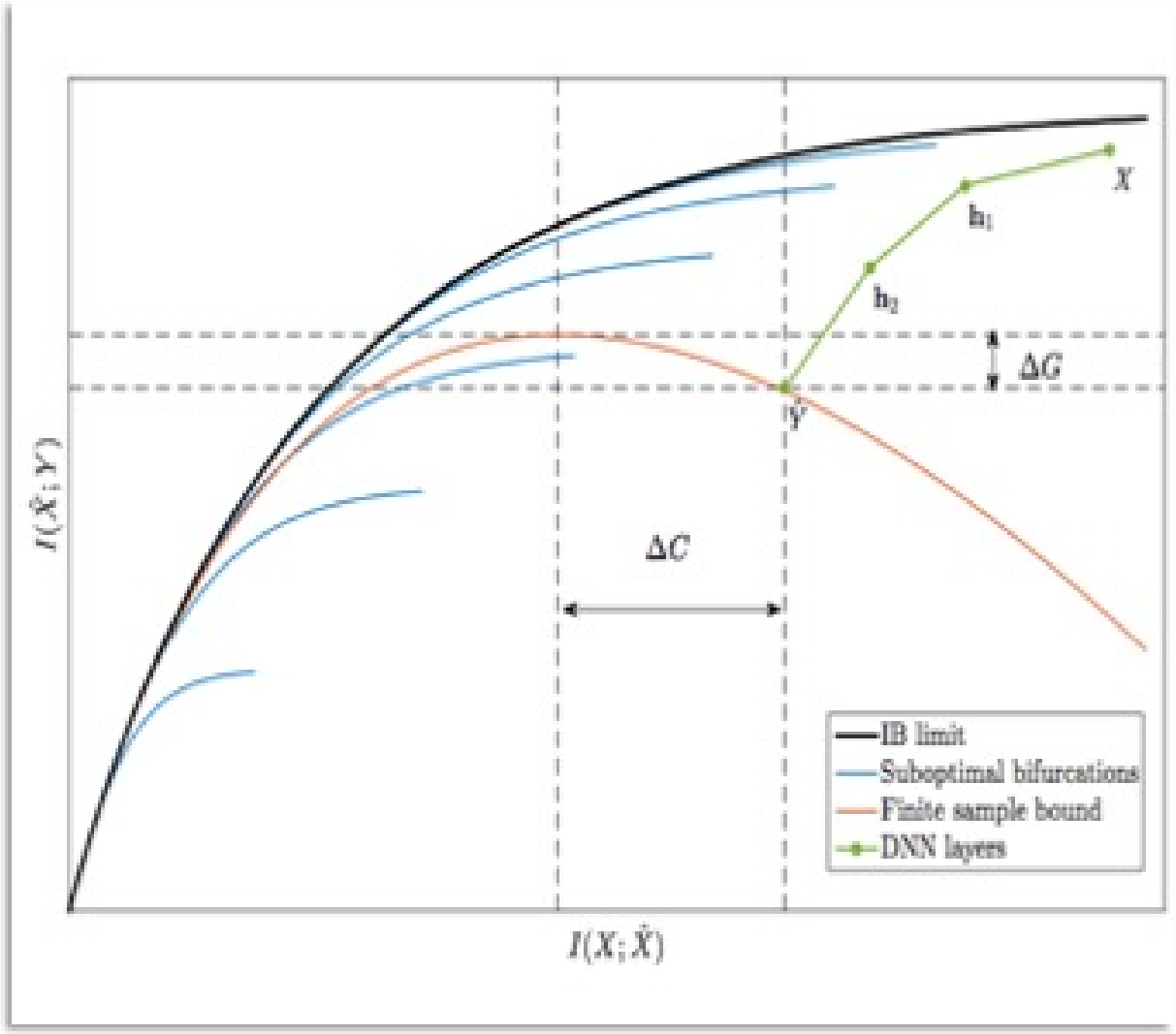
The IB bound optimality equations:

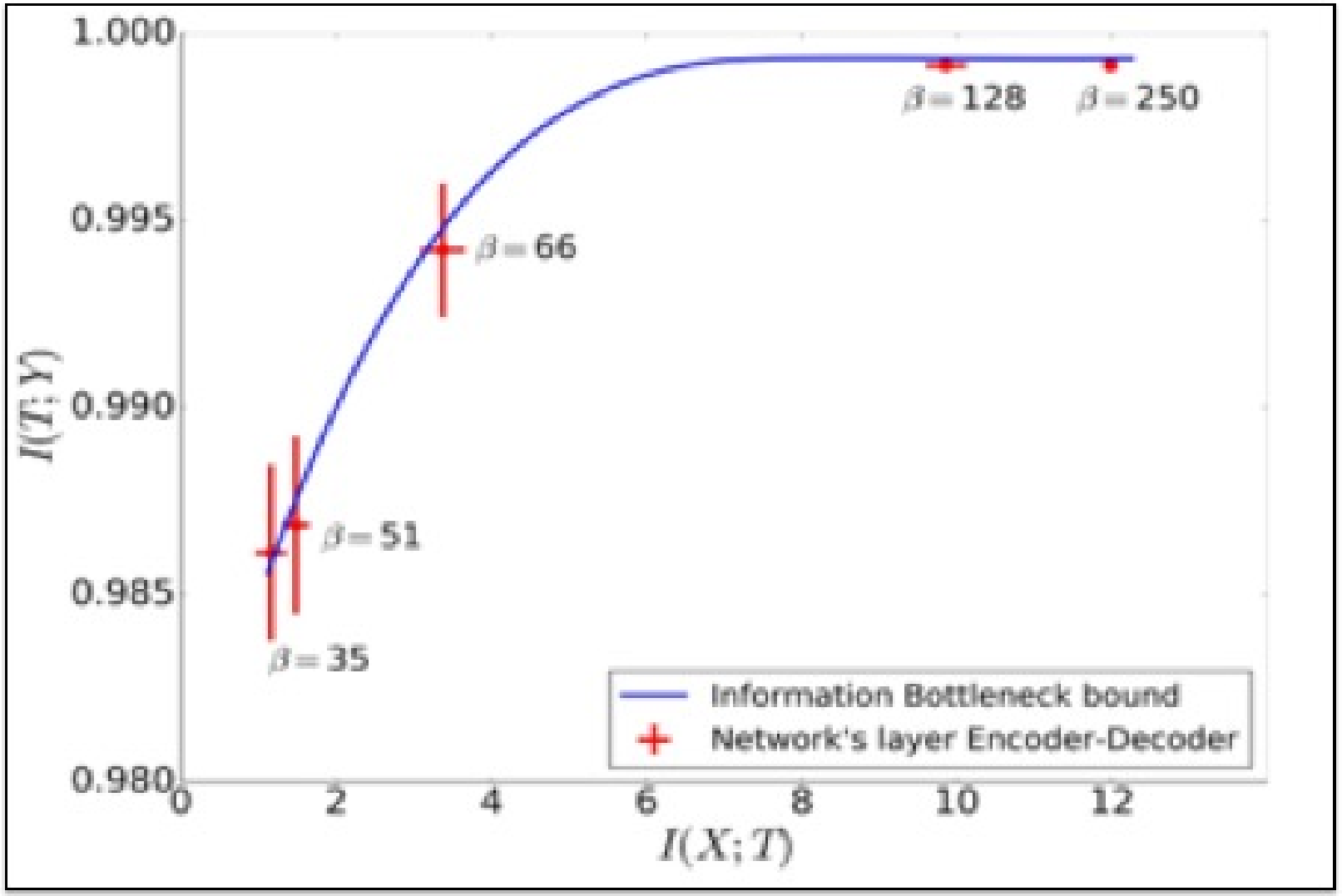
$$\min_{p(\hat{x}|x):\, Y\to X\to\hat{X}} I(\hat{X};X) - \beta I(\hat{X};Y)\ ,\ \ \beta > 0$$

$$\begin{cases} p(x\,|\,\hat{x}) = \dfrac{p(x)}{Z(x,\beta)} \exp(-\beta D[p(y\,|\,x)\,\|\,p(y\,|\,\hat{x}]) \\ Z(x,\beta) = \sum_{\hat{x}} p(\hat{x}) \exp(-\beta D[p(y\,|\,x)\,\|\,p(y\,|\,\hat{x}]) \\ p(\hat{x}) = \sum_{x} p(\hat{x}\,|\,x)\,p(x) \\ p(y\,|\,\hat{x}) = \sum_{x} p(y\,|\,x)\,p(x\,|\,\hat{x}) \end{cases}$$

Solved by Arimoto-Blahut like iterations,

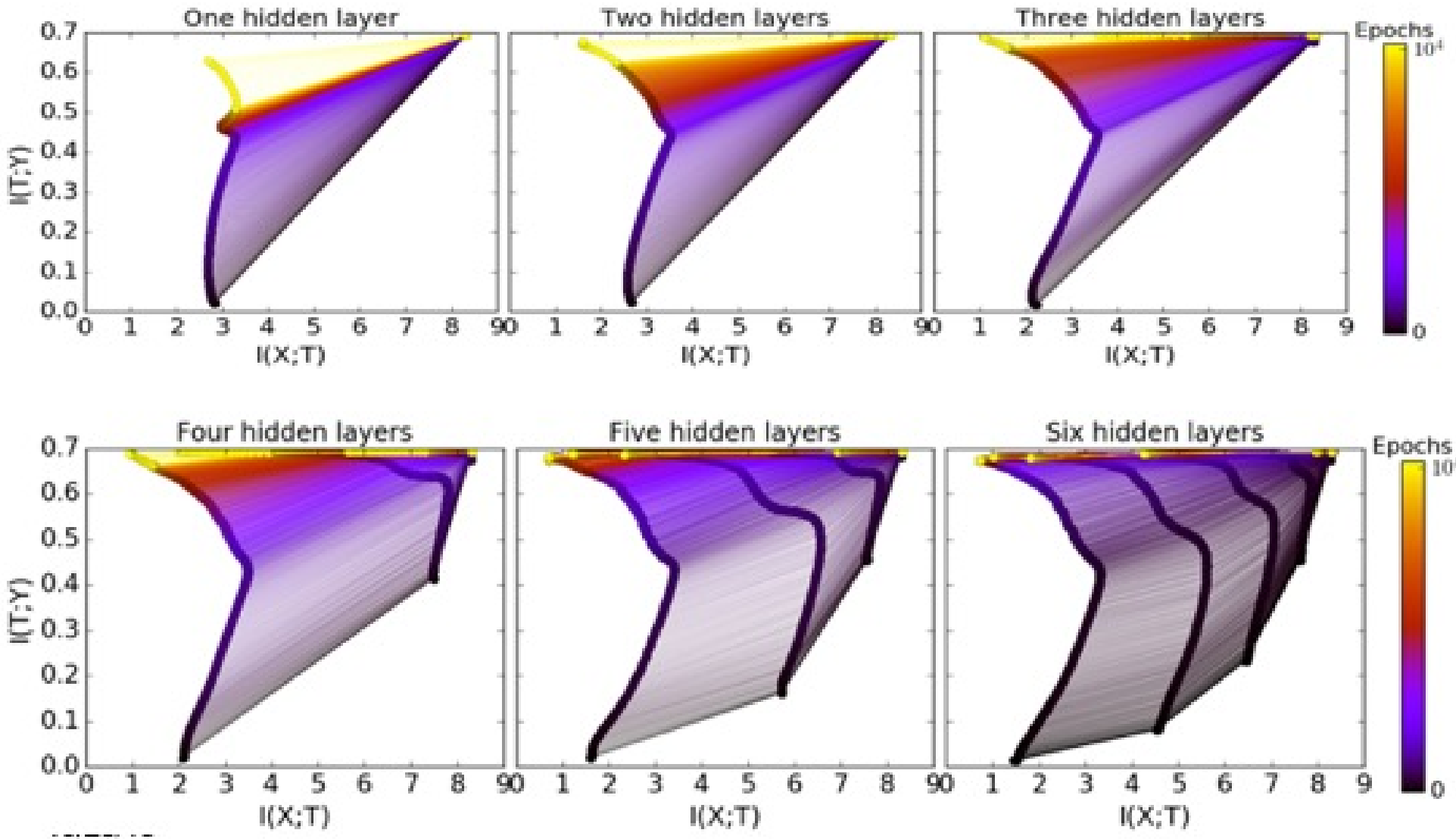
but with possibly sub-optimal solutions, bifurcations (!),





- Layers of optimal DNN converge to [a successively refineable approximation of] the optimal finite-sample IB limit information-curve
- Layers must be in "different topological phases" of the IB solutions
- The DNN encoder & decoder for each layer satisfy the IB self-consistent equations

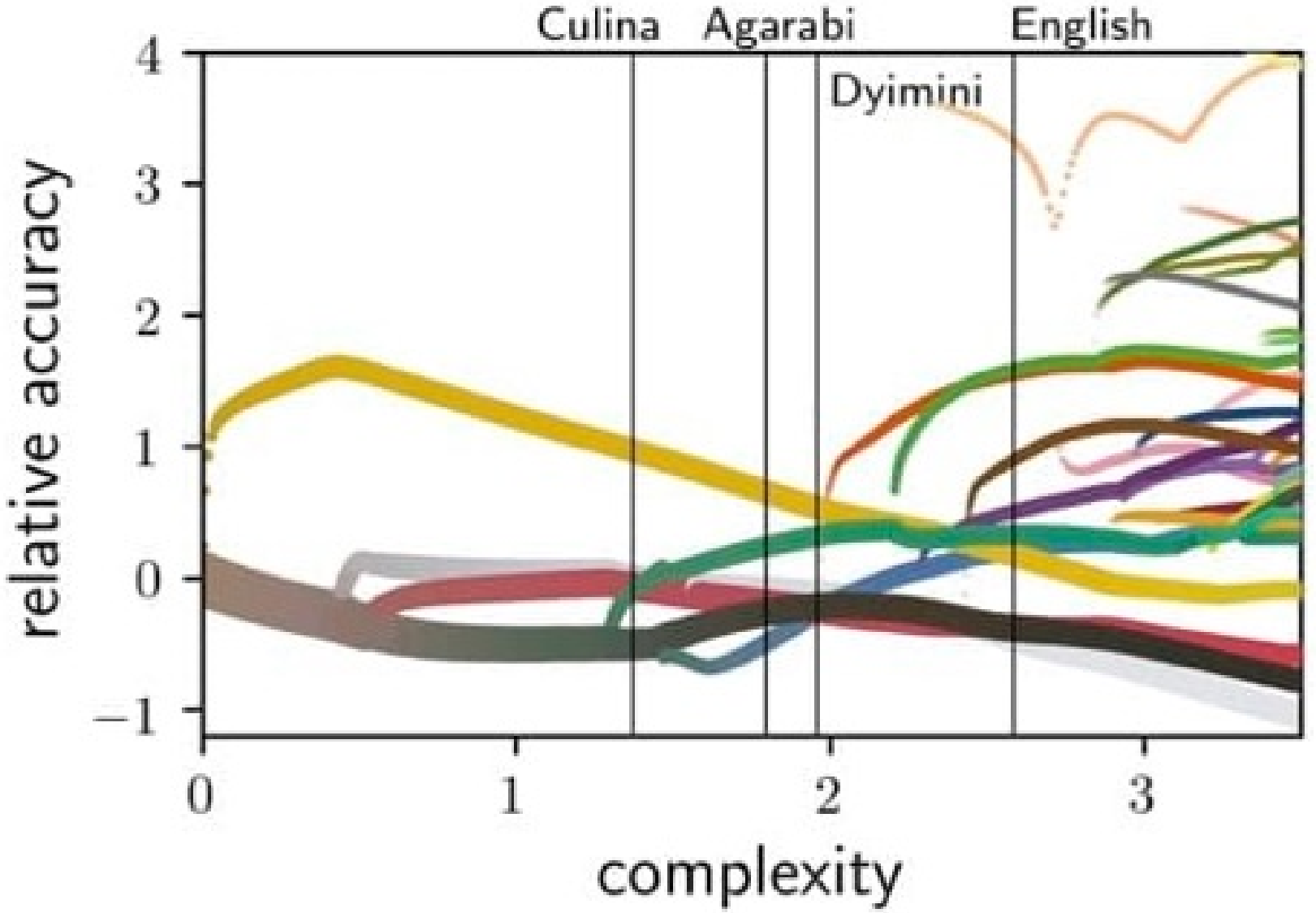
The benefit of the hidden layers is computational!



More layers take much FEWER training epochs for good generalization.

The optimization time depend super-linearly (exponentially?) on the compressed information, delta Ix, for each layer.

Shannon's optimality in color names evolution



Noga Zaslavsky
2018

The Future?