Statistical Learning Theory and Applications

9.520/6.860 in Fall 2017

Class Times:
Monday and Wednesday 1pm-2:30pm in 46-3310 Units: 3-0-9 H,G
Web site: http://www.mit.edu/~9.520/

Email Contact :
9.520@mit.edu

9.520: Statistical Learning Theory and Applications

- Course focuses on regularization techniques for supervised learning.
- Support Vector Machines, manifold learning, sparsity, batch and online supervised learning, feature selection, structured prediction, multitask learning.
- Optimization theory critical for machine learning (first order methods, proximal/ splitting techniques).
- *Focus on deep learning and theory of it, based on first part of the class*

The goal of this class is to provide the theoretical knowledge and the basic intuitions underlying it, which are needed to effectively use and develop machine learning solutions to a variety of problems.

Class

http://www.mit.edu/~9.520/

Mathcamps

- Functional analysis (~45mins)

Linear Algebra

Basic notion and definitions: matrix and vectors norms, positive, symmetric, invertible matrices, linear systems, condition number.

Functional Analysis:

Linear and Euclidean spaces
scalar product, orthogonality
orthonormal bases, norms and semi-norms,
Cauchy sequence and complete spaces
Hilbert spaces, function spaces
and linear functional, Riesz representation
theorem, convex functions, functional calculus.

- Probability (~45mins)

Probability Theory:

Random Variables (and related concepts), Law of Large Numbers, Probabilistic Convergence, Concentration Inequalities.

Class http://www.mit.edu/~9.520/: big picture

Shallow Networks

- Classes 2-9 are the core: foundations + regularization
- Classes 10-20 are state-of-the-art topics for research in — and applications of — ML

Deep Networks

- Classes 21-25 review very recent developments in the theory of multilayer networks (DCLNs)

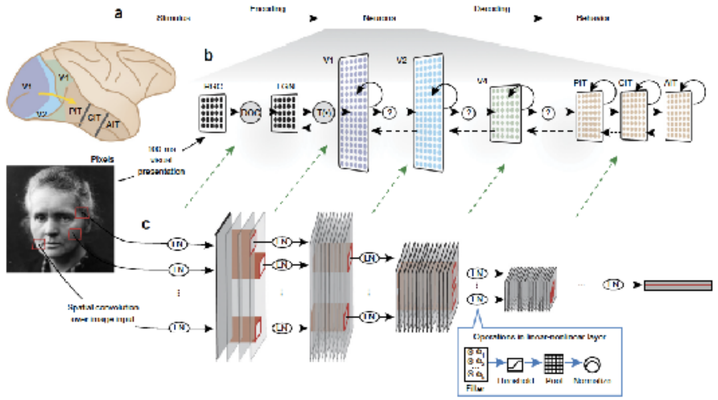
Today's hand wavy overview

- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM
- A bit of history: Statistical Learning Theory, Neuroscience
- A bit of ML history: applications
- Deep Learning

CBMM

CBMM's focus 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.



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

CENTER FOR Brains Minds + Machines

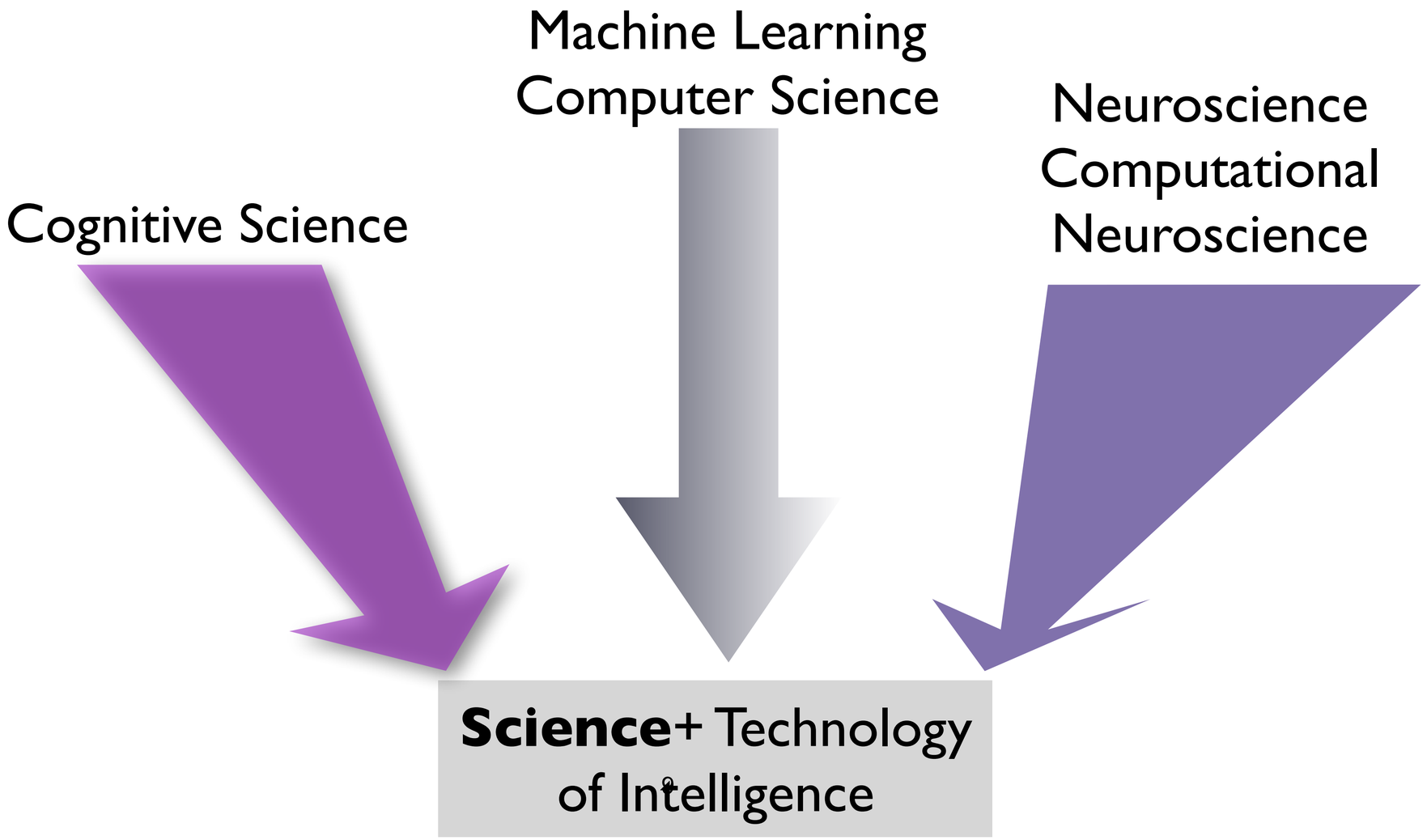
The problem of intelligence: how it arises in the brain and how to replicate it in machines

The problem of (human) intelligence is one of the great problems in science, probably the greatest.

Research on intelligence:

- a great intellectual mission: understand the brain, reproduce it in machines
- will help develop intelligent machines

Interdisciplinary



Research, Education & Diversity Partners

MIT
Boyden, Desimone, DiCarlo, Kanwisher, Katz, McDermott, Poggio, Rosasco, Sassanfar, Saxe, Schulz, Tegmark, Tenenbaum, Ullman, Wilson, Winston

Harvard
Blum, Gershman, Kreiman, Livingstone, Nakayama, Sompolinsky, Spelke

Allen Institute
Koch

Howard U.
Chouika, Manaye, Rwebangira, Salmani

Hunter College
Chodorow, Epstein, Sakas, Zeigler

Johns Hopkins U.
Yuille

Queens College
Brumberg

Rockefeller U.
Freiwald

Stanford U.
Goodman

Universidad Central del Caribe (UCC)
Jorquera

University of Central Florida
McNair Program

UMass Boston
Blaser, Ciaramitaro, Pomplun, Shukla

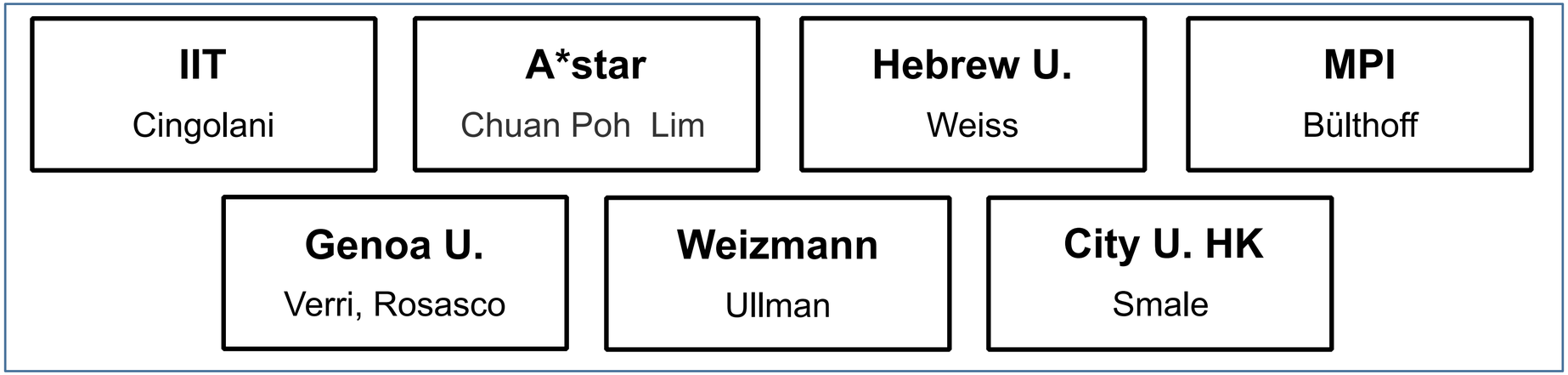
UPR – Mayagüez
Santiago, Vega-Riveros

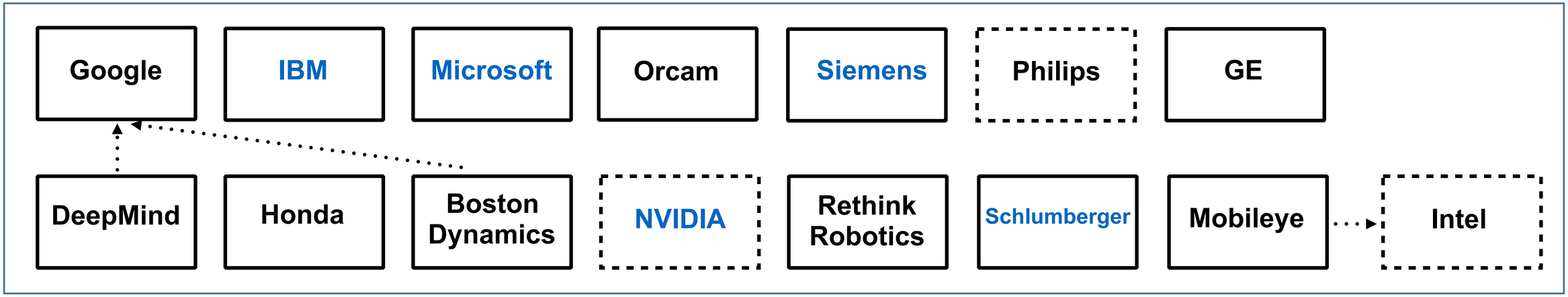
UPR– Río Piedras
Garcia-Arraras, Maldonado-Vlaar, Megret, Ordóñez, Ortiz-Zuazaga

Wellesley College
Hildreth, Wiest, Wilmer

CENTER FOR Brains Minds + Machines

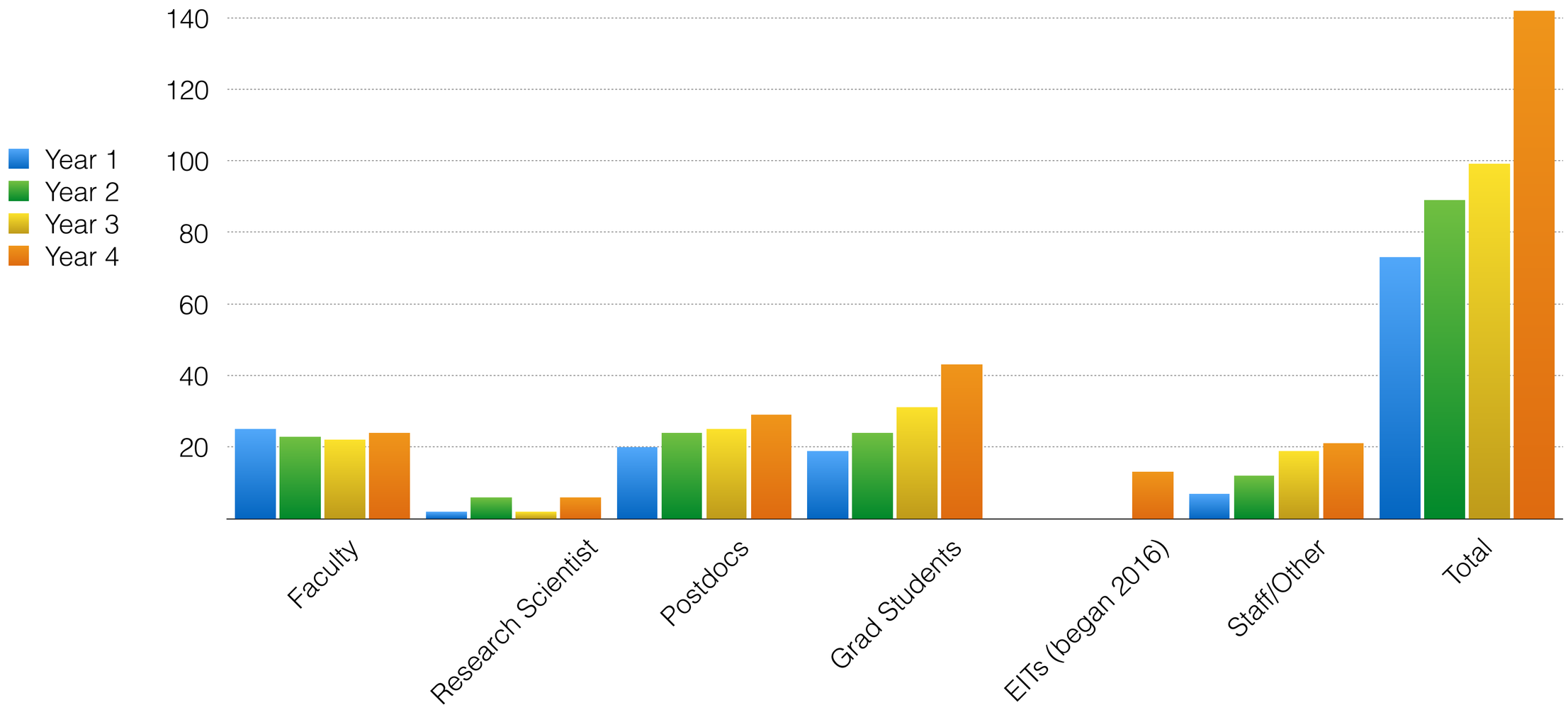
Academic and Corporate Partners





CENTER FOR
Brains
Minds+
Machines

CBMM Participants



CENTER FOR
Brains
Minds+
Machines

Collaboration

- Of all the things that your STC does, what works best to foster inter-institutional collaboration?

CENTER FOR
Brains
Minds+
Machines

Education

CENTER FOR
Brains
Minds +
Machines

CBMM Summer Course at Woods Hole: Our flagship initiative

Brains, Minds & Machines Summer Course

An intensive three-week course gives advanced students a “deep” introduction to the problem of intelligence

A community of scholars between computer science and neuroscience is being formed:

First reunion of alumni of summer school Aug. 26-27 in Woodshole, MA

CENTER FOR
Brains
Minds+
Machines

Recent Achievements in AI

Intelligence in games: the beginning


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

Recent progress in AI

CENTER FOR
Brains
Minds+
Machines

The 2 best examples of the success of new ML

- AlphaGo
- Mobileye

FINANCIAL TIMES

ft.com/comment

You are signed in | Search for...

Subscribe now - Save up to 60% ▸

Home | World | Companies | Markets | Global Economy | Lex | Comment | Management | Life & Arts

Columnists | The Big Read | Opinion | FT View | Instant Insight | EM Squared | The Exchange | Blogs | Letters | Corrections | Obits | Tools

PERSON IN THE NEWS

March 11, 2016 3:14 pm

Demis Hassabis, master of the new machine age

Murad Ahmed

Share | Author alerts | Print | Clip | Comments

The creator of the AI game-playing program makes all the right moves, writes Murad Ahmed

The victories have a human mastermind in Demis Hassabis, co-founder and chief executive of DeepMind. He describes Mr Lee as the "Roger Federer of Go", and for some the computer program's achievement is akin to a robot taking to the lawns of Wimbledon and beating the legendary tennis champion.

"I think it is pretty huge but, ultimately, it will be for

More

PERSON IN THE NEWS

James Comey

Ali al-Naimi

Kyle Bass

THE BIG READ

EDF

TUNISIA

CENTER FOR Brains Minds + Machines


KBA
Google DeepMind
Challenge Match
AlphaGo
KBA
Google

Real Engineering: Mobileye

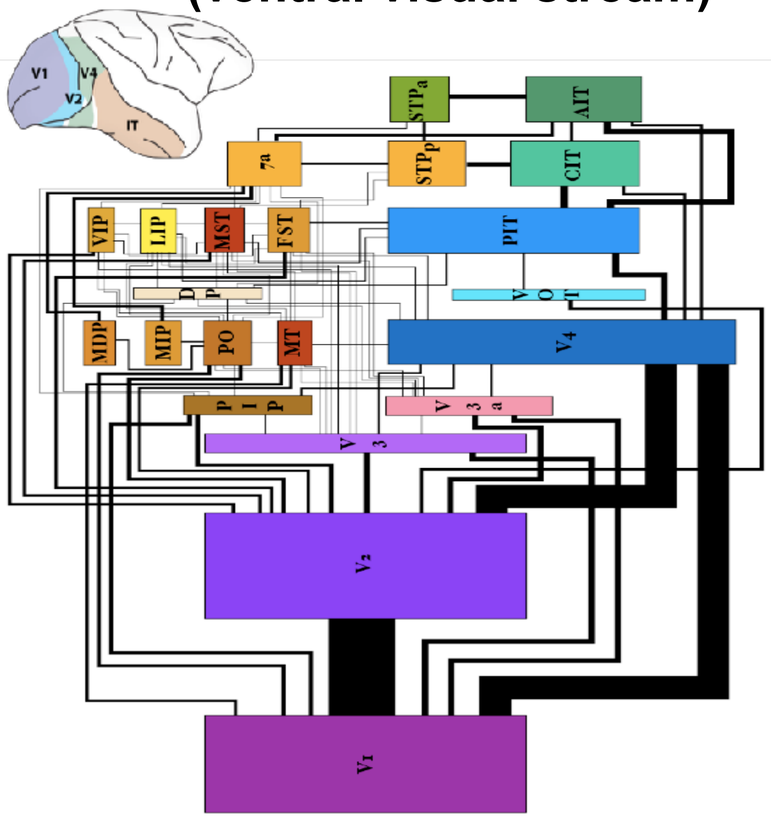
History

Inspiration from Neuroscience

Background: State-of-the-art Machines ("Deep Learning") Have Emerged From the Brain's Visual Processing Architecture

Brains / Minds

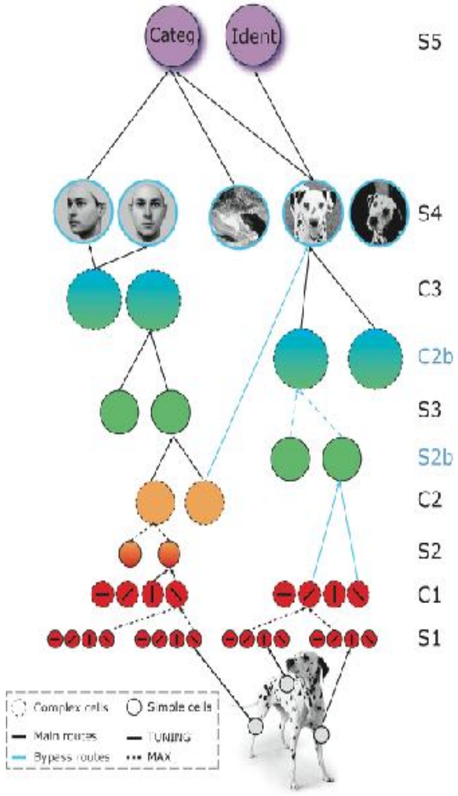
(ventral visual stream)

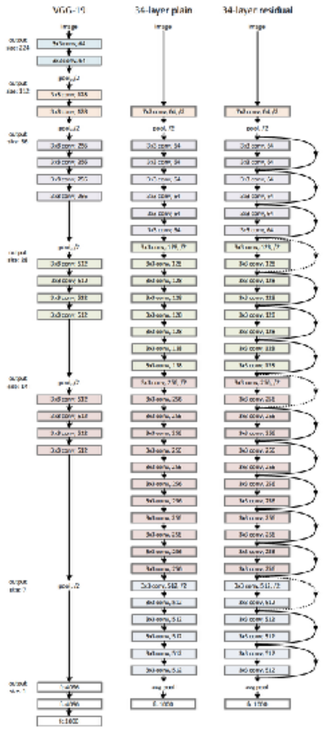


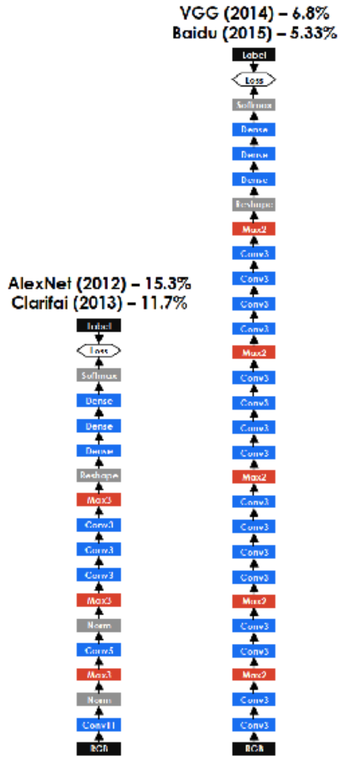
Desimone & Ungerleider 1989; vanEssen+Movshon

What's the engineering of the future?

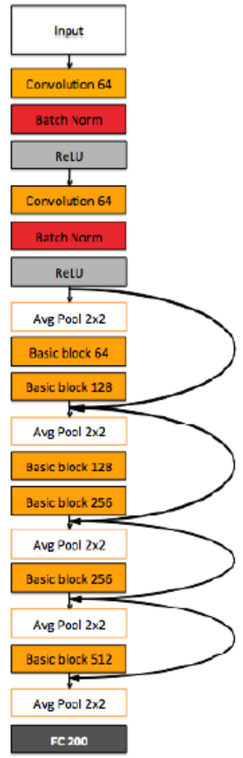
Machines → …







State of the Art ResNets



CENTER FOR Brains Minds + Machines

The Problem of Intelligence is NOT solved as yet….

The Problems of Intelligence and CBMM

Intelligence is not solved,
not as as a scientific problem, not as an engineering problem.

Research is needed:

- for the sake of basic science
- for the engineering of tomorrow

Building Jarvis

MARK ZUCKERBERG · MONDAY, DE

My personal challenge for 2016 was to build a simple AI to
Iron Man. Within 5-10 years we'll have AI systems that are
each of our senses -- vision, hearing, touch, etc, as well as t
impressive how powerful the state of the art for these tools

At the same time, we are still far off from understanding
Everything I did this year – natural language, face recogniti
–are all variants of the same fundamental pattern recognitio
hours building Jarvis this year, **but even if I spent 1,000 m**
be able to build a system that could learn completely ne
made some fundamental breakthrough in the state of AI

CENTER FOR Brains Minds + Machines

For the solution I bet we will need Neuroscience

(suggestion: attend 6.861/9.523)

Neuron
Review

Neuroscience-Inspired Artificial Intelligence

Demis Hassabis,[1,2,*] Dharshan Kumaran,[1,3] Christopher Summerfield,[1,4] and Matthew Botvinick[1,2]
[1]DeepMind, 5 New Street Square, London, UK
[2]Gatsby Computational Neuroscience Unit, 25 Howland Street, London, UK
[3]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, UK
[4]Department of Experimental Psychology, University of Oxford, Oxford, UK
*Correspondence: dhcontact@google.com
http://dx.doi.org/10.1016/j.neuron.2017.06.011

The fields of neuroscience and artificial intelligence (AI) have a long and intertwined history. In more recent times, however, communication and collaboration between the two fields has become less commonplace. In this article, we argue that better understanding biological brains could play a vital role in building intelligent machines. We survey historical interactions between the AI and neuroscience fields and emphasize current advances in AI that have been inspired by the study of neural computation in humans and other animals. We conclude by highlighting shared themes that may be key for advancing future research in both fields.

The successful transfer of insights gained from neuroscience to the development of AI algorithms is critically dependent on the interaction between researchers working in both these fields, with insights often developing through a continual handing back and forth of ideas between fields. In the future, we

INTERVIEW SCIENCE TECH

DeepMind's founder says to build better computer brains, we need to look at our own

What AI can learn from neuroscience, and neuroscience from AI

by James Vincent | @jjvincent | Jul 19, 2017, 12:00pm EDT

Illustration by James Bareham / The Verge

They point out that contemporary AI programs are extremely narrow in their abilities; that they're easily tricked, and simply don't possess those hard-to-define — but easy-to-spot skills we usually sum up as "common sense." They are, in short, not that intelligent.

The question is: how do we get to the next level? For Demis Hassabis, founder of Google's AI powerhouse DeepMind, the answer lies within us. Literally. In a review

The Science of Intelligence

The science of intelligence was at the roots of today's engineering success

We need to make another basic effort leveraging
the old and new
science of intelligence:
neuroscience, cognitive science, learning theory

Today's overview

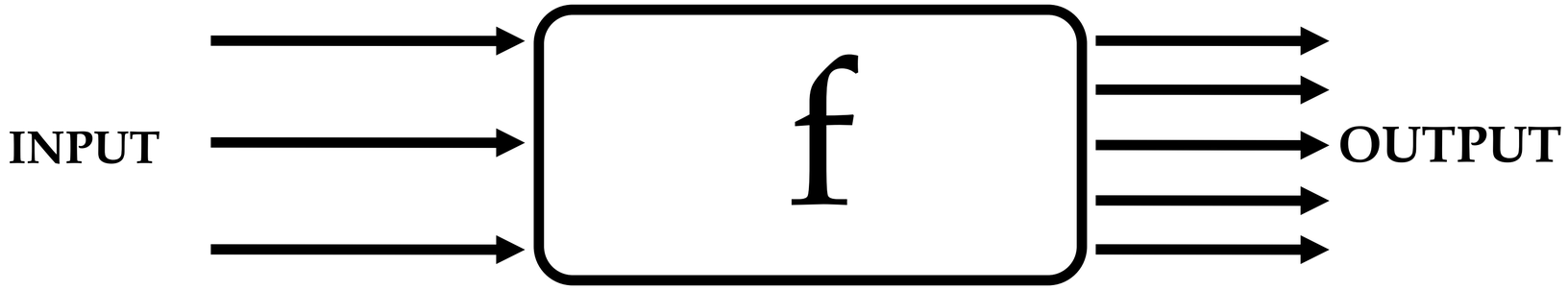
- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM

Summary: I told you about the present great success of ML, its connections with neuroscience, its limitations for full AI. I then told you that we need to connect to neuroscience if we want to realize real AI, in addition to understanding our brain. BTW, even without this extension, the next few years will be a golden age for ML applications. The connection to neuroscience is what we do at CBMM and in the CBMM Summer School: this is an advertisement.

Summary of today's overview

- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM
- A bit of history: Statistical Learning Theory
- A bit of history: applications
- Deep Learning

Statistical Learning Theory: **supervised** learning (~1980-2010)



Given a set of l examples (data)

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\}$$

Question: find function f such that

$$f(x) = \hat{y}$$

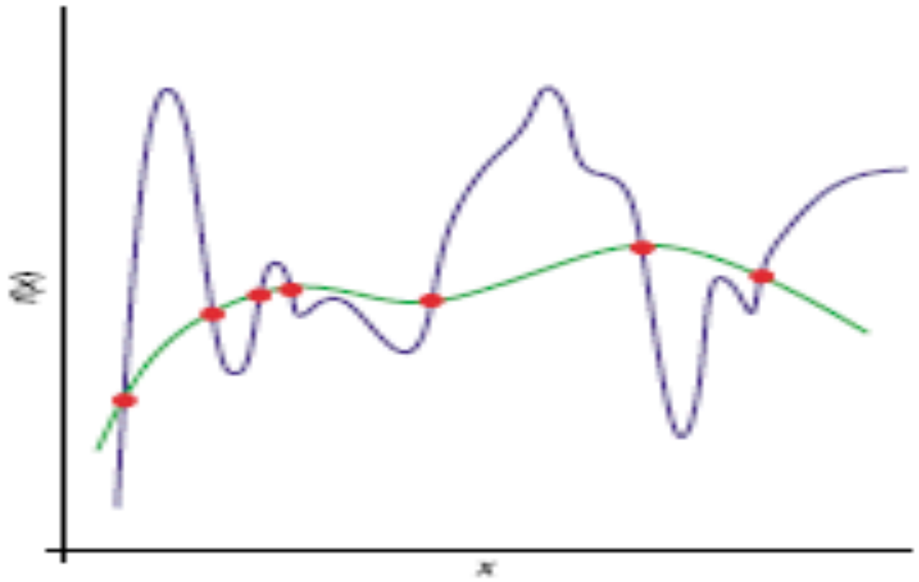
is a **good predictor** of y for a **future** input x (fitting the data is **not** enough!)

Statistical Learning Theory: prediction, not description

= data from f

= function f

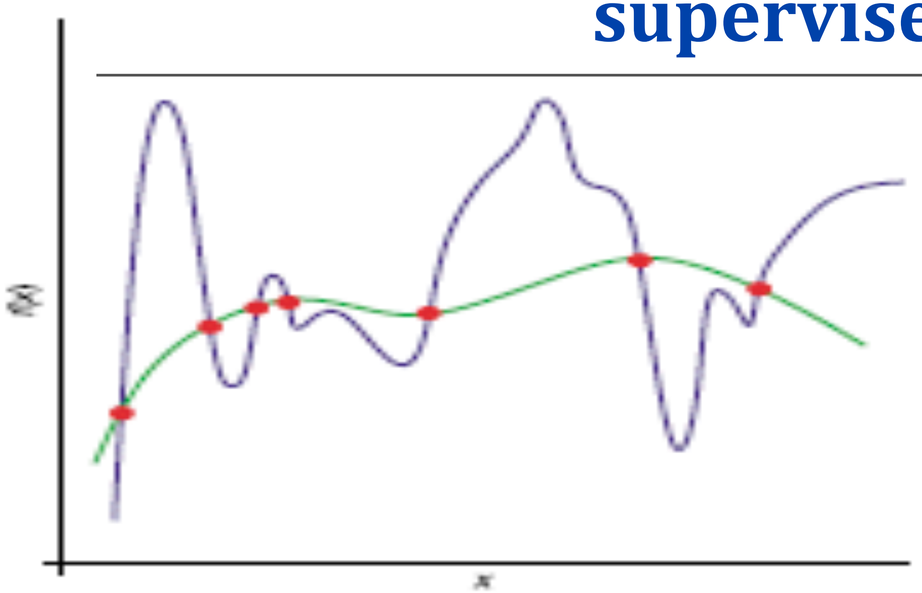
= approximation of f



Generalization:

estimating value of function where there are no data (good generalization means predicting the function well; important is for empirical or validation error to be a good proxy of the prediction error)

Statistical Learning Theory: supervised learning



Regression

(4,24,…)

(1,13,…)

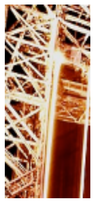
(7,33,…)

Classification

(92,10,…)

(41,11,…)

(19,3,…)



(4,71,…)

Statistical Learning Theory: supervised learning

There is an unknown **probability distribution** on the product space $Z = X \times Y$, written $\mu(z) = \mu(x, y)$. We assume that X is a compact domain in Euclidean space and Y a bounded subset of $\mathbb{R}$. The **training set** $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} = \{z_1, ...z_n\}$

consists of n samples drawn i.i.d. from μ.

$\mathcal{H}$ is the **hypothesis space**, a space of functions $f : X \to Y$.

A **learning algorithm** is a map $L : Z^n \to \mathcal{H}$ that looks at S and selects from $\mathcal{H}$ a function $f_S : \mathbf{x} \to y$ such that $f_S(\mathbf{x}) \approx y$ *in a predictive way*.

Statistical Learning Theory

Given a function f, a loss function V, and a probability distribution μ over Z, the **expected or true error** of f is:

$$I[f] = \mathbb{E}_Z V[f, z] = \int_Z V(f, z) d_\mu(z) \tag{1}$$

which is the **expected loss** on a new example drawn at random from μ.
The **empirical error** of f is:

$$I_S[f] = \frac{1}{n} \sum V(f, z_i) \tag{2}$$

A very natural requirement for f_S is distribution independent **generalization**

$$\forall \mu, \lim_{n \to \infty} |I_S[f_S] - I[f_S]| = 0 \ \textit{in probability} \tag{3}$$

In other words, the training error for the solution must converge to the expected error and thus be a "proxy" for it. Otherwise the solution would not be "predictive".

Statistical Learning Theory: foundational theorems

Conditions for generalization and well-posedness in learning theory have deep, almost philosophical, implications:

they can be regarded as equivalent conditions that guarantee a theory to be predictive and scientific

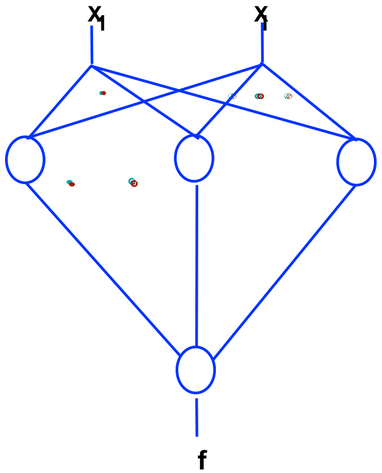
- theory must be chosen from a small hypothesis set (~ Occam razor, VC dimension,…)
- theory should not change much with new data...most of the time (stability)

Classical algorithm:
Regularization in RKHS (eg. kernel machines)

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

implies

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



Remark (for later use):

Classical kernel machines — such as SVMs — correspond to shallow networks

Summary of today's overview

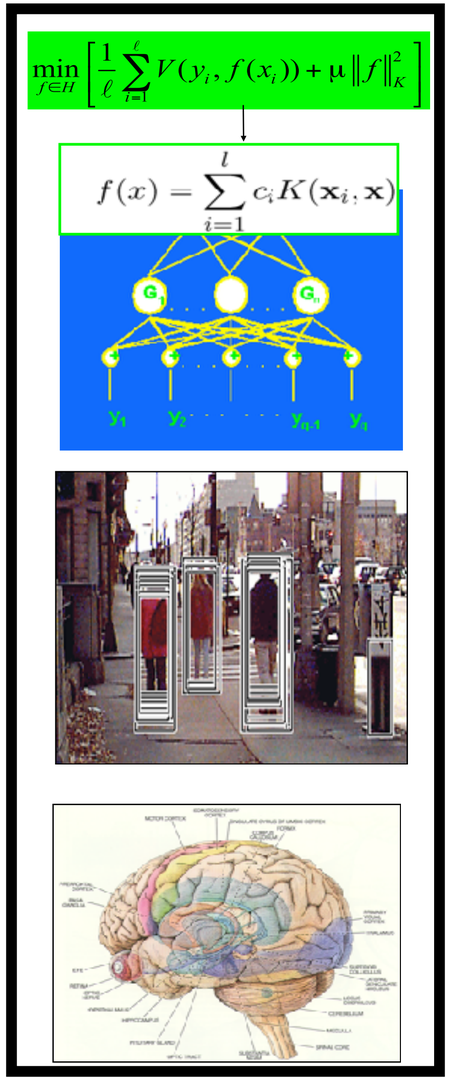
- A bit of history: Statistical Learning Theory

Summary: I told you about learning theory and the concern about productivity and no overfitting. I told you about kernel machines and shallow networks. We will learn a lot about RKHS. Much of this is needed for an eventual theory for deep learning.

Summary of today's overview

- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM
- A bit of history: Statistical Learning Theory, Neuroscience
- A bit of history: old applications
- Deep Learning

Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

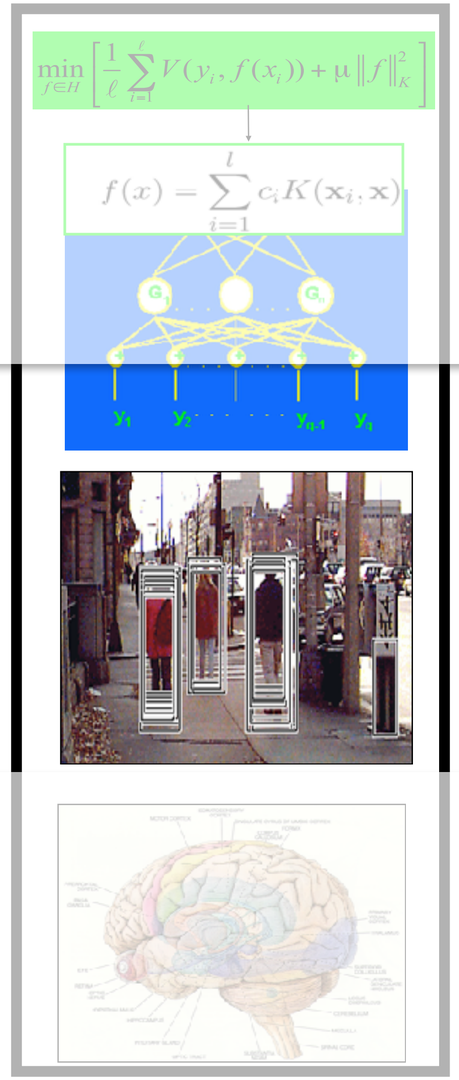
Predictive algorithms

Sung & Poggio 1995, also Kanade& Baluja....

COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

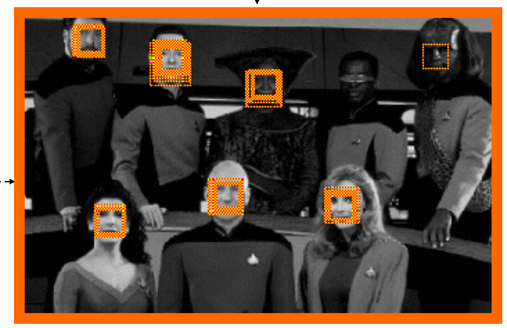
Engineering of Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms



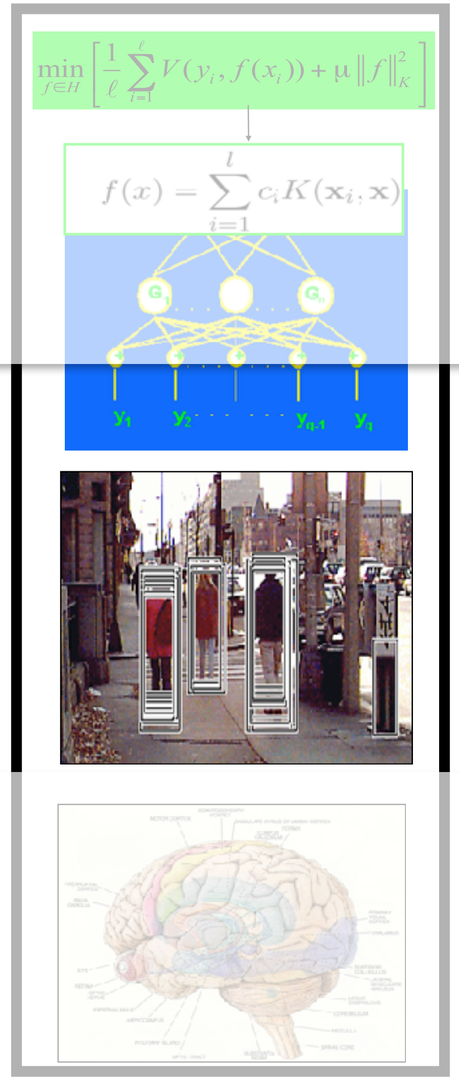
Sung & Poggio 1995

COMPUTATIONAL
NEUROSCIENCE:
models+experiments

How visual cortex works


Image
Output

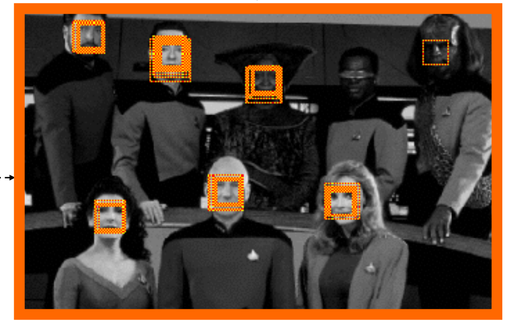
Engineering of Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

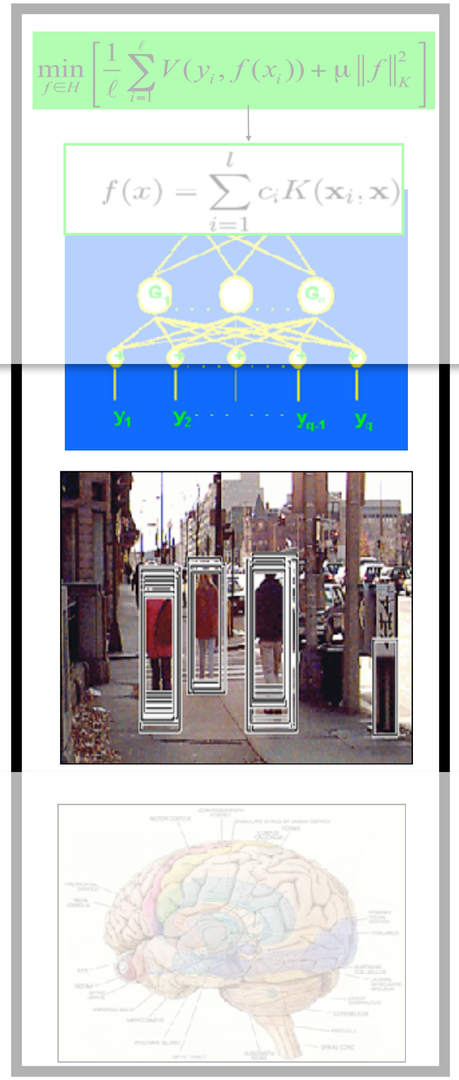


Face detection has been available in digital cameras for a few years now

COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

Engineering of Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

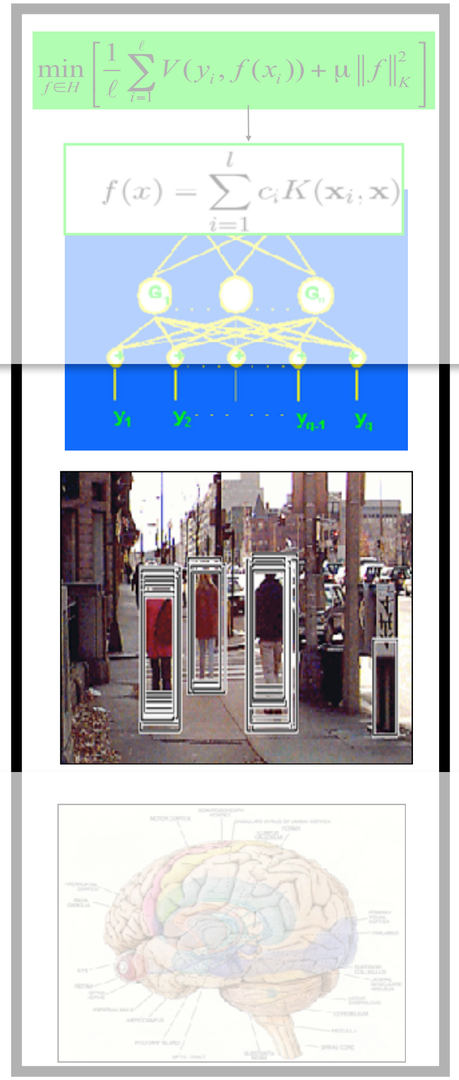
People detection

Papageorgiou&Poggio, 1997, 2000
also Kanade&Scheiderman

COMPUTATIONAL
NEUROSCIENCE:
models+experiments

How visual cortex works

Engineering of Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

Pedestrian detection

Papageorgiou&Poggio, 1997, 2000
also Kanade&Scheiderman

COMPUTATIONAL
NEUROSCIENCE:
models+experiments

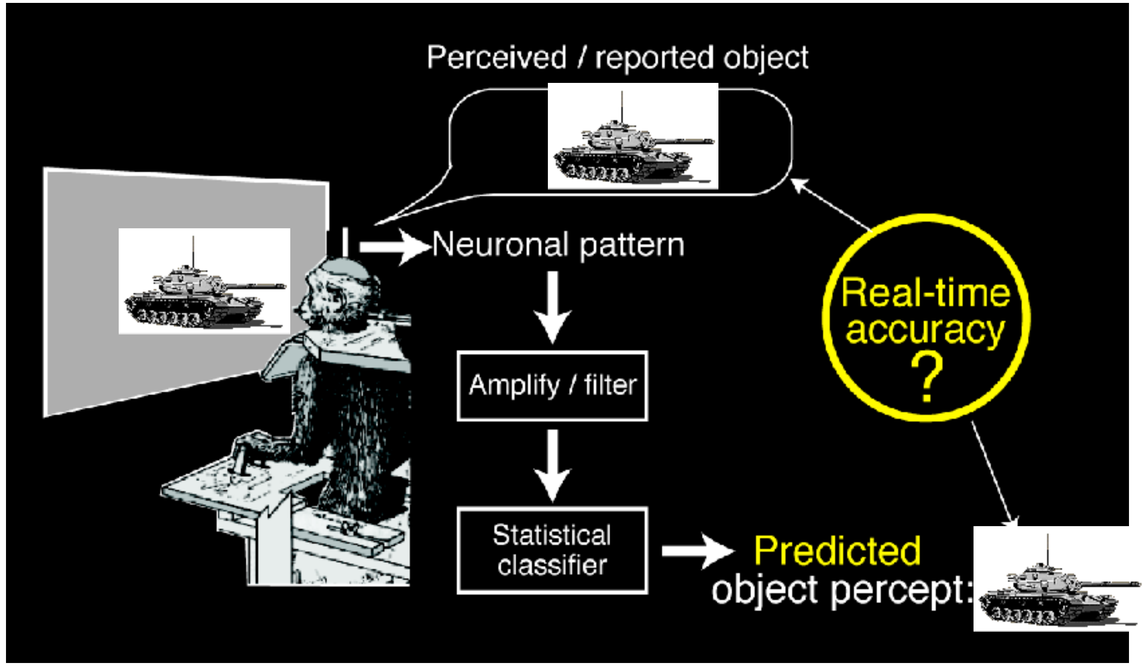
How visual cortex works

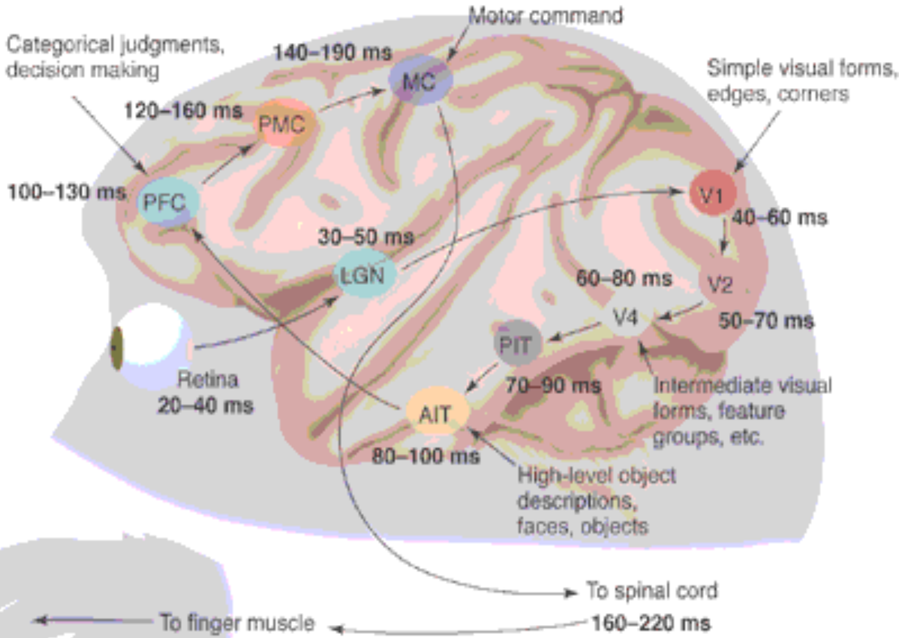
Some other examples of past ML applications from my lab

Computer Vision
- Face detection
- Pedestrian detection
- Scene understanding
- Video categorization
- Video compression
- Pose estimation

Graphics
Speech recognition
Speech synthesis
Decoding the Neural Code
Bioinformatics
Text Classification
Artificial Markets
Stock option pricing
….

Decoding the neural code: Matrix-like read-out from the brain





Hung, Kreiman, Poggio, DiCarlo. *Science 2005*

Learning: bioinformatics

New feature selection SVM:

Only 38 training examples, 7100 features

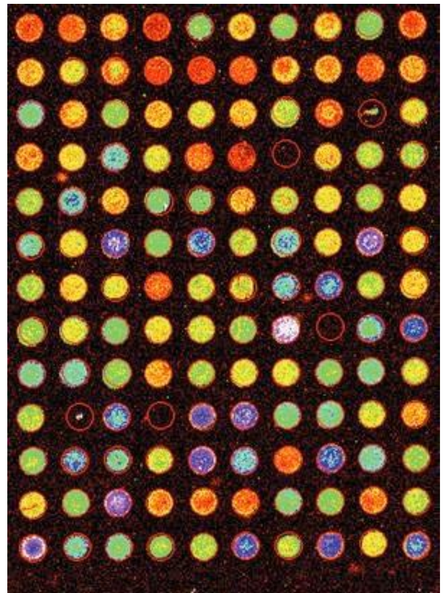
AML vs ALL: 40 genes 34/34 correct, 0 rejects.
5 genes 31/31 correct, 3 rejects of which 1 is an error.

A.I. Memo No.1677
C.B.C.L Paper No.182

Support Vector Machine Classification of Microarray Data

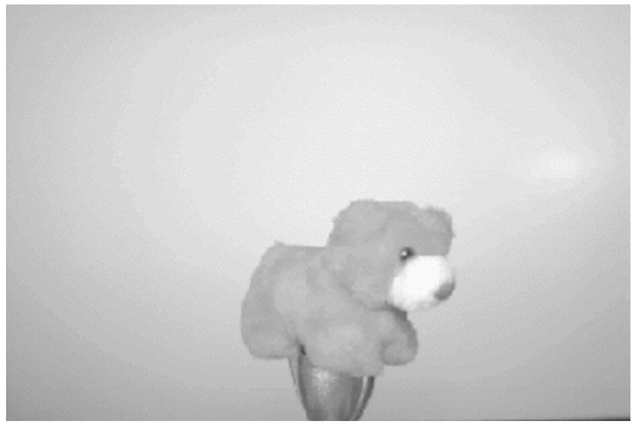
S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J.P. Mesirov, and T. Poggio

Pomeroy, S.L., P. Tamayo, M. Gaasenbeek, L.M. Sturia, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, M.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander and T.R. Golub. Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression, *Nature*, 2002.



Learning: image analysis

⇒ **Bear (0° view)**



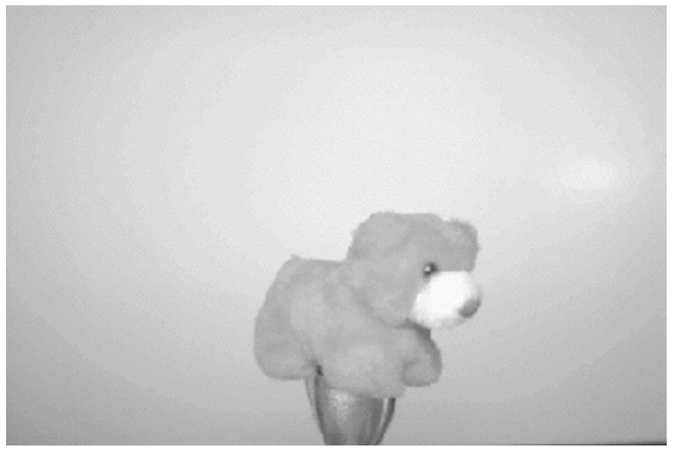
⇒ **Bear (45° view)**

Learning: image synthesis

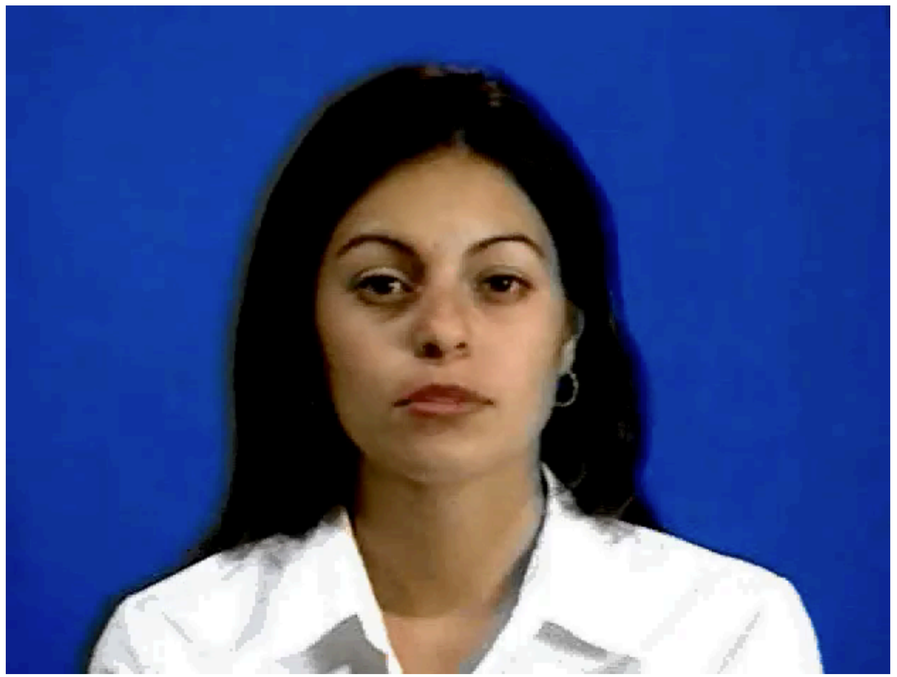
UNCONVENTIONAL GRAPHICS

Θ = 0° view $\Rightarrow$

Θ = 45° view $\Rightarrow$



Extending the same basic learning techniques (in 2D): Trainable Videorealistic Face Animation



Mary101

A- more in a moment

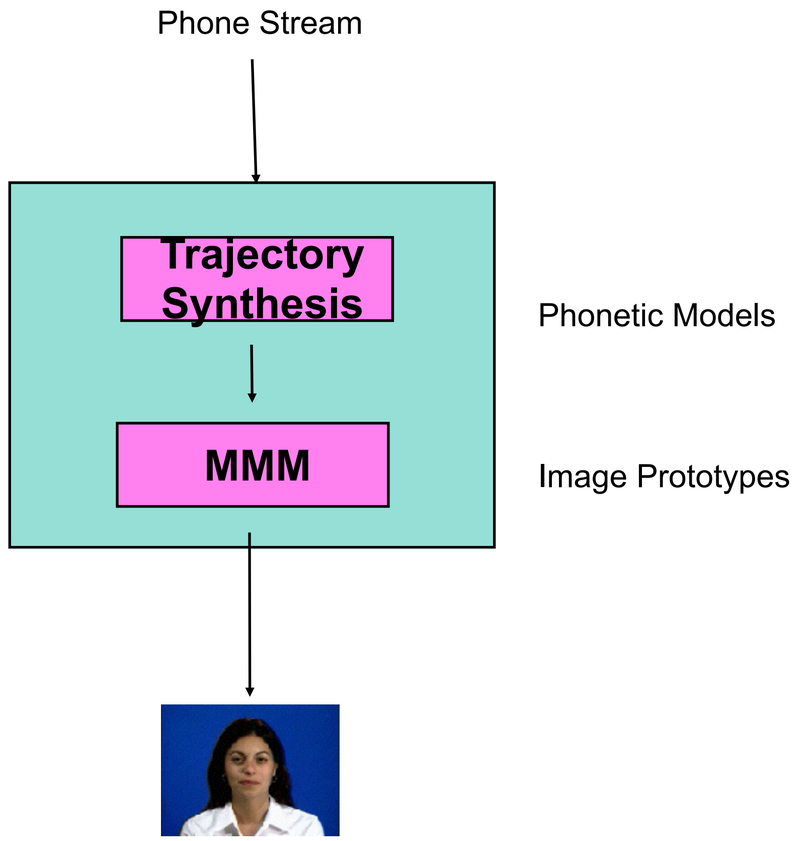
Tony Ezzat,Geiger, Poggio, **SigGraph 2002**

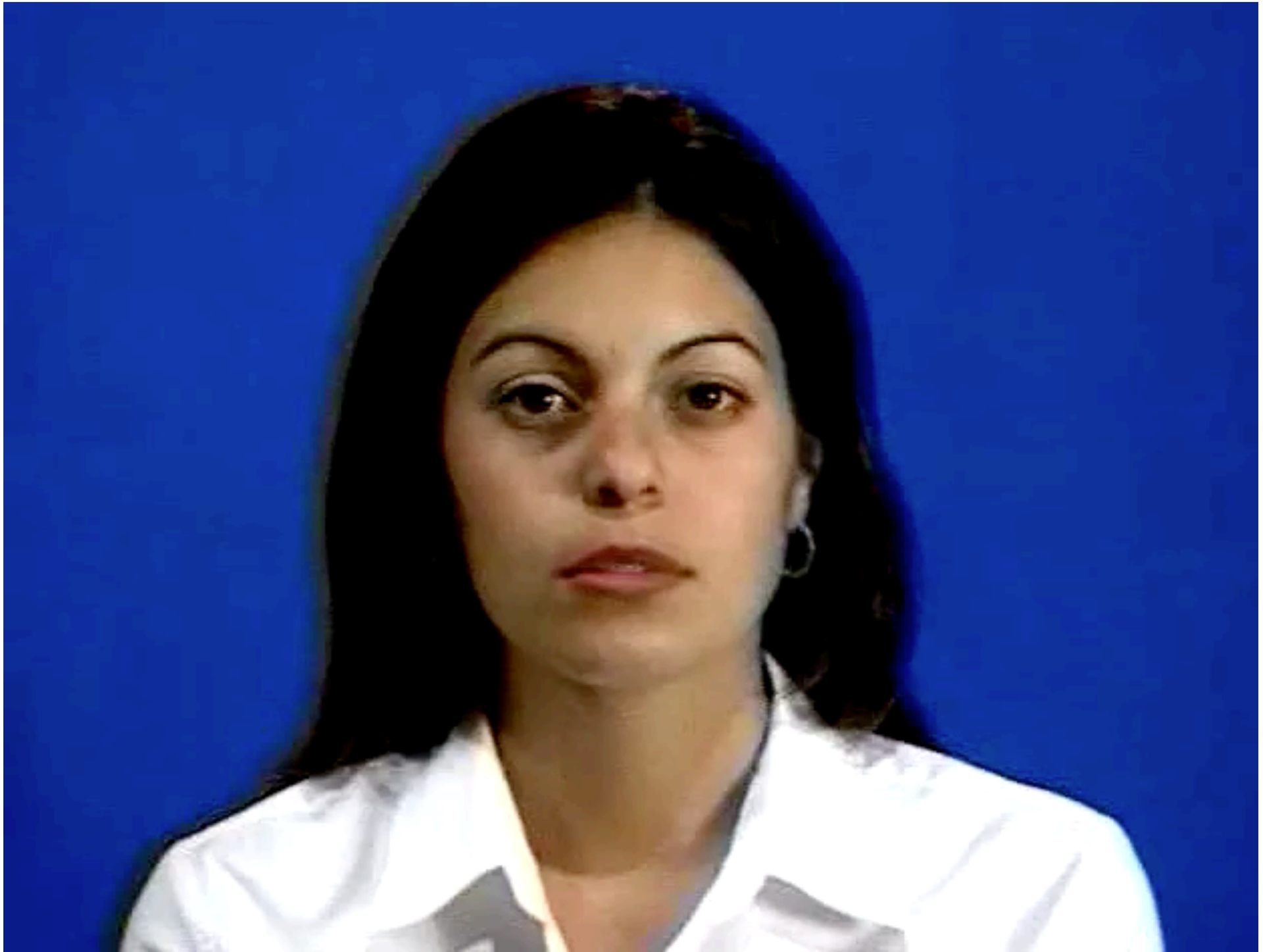
1. Learning

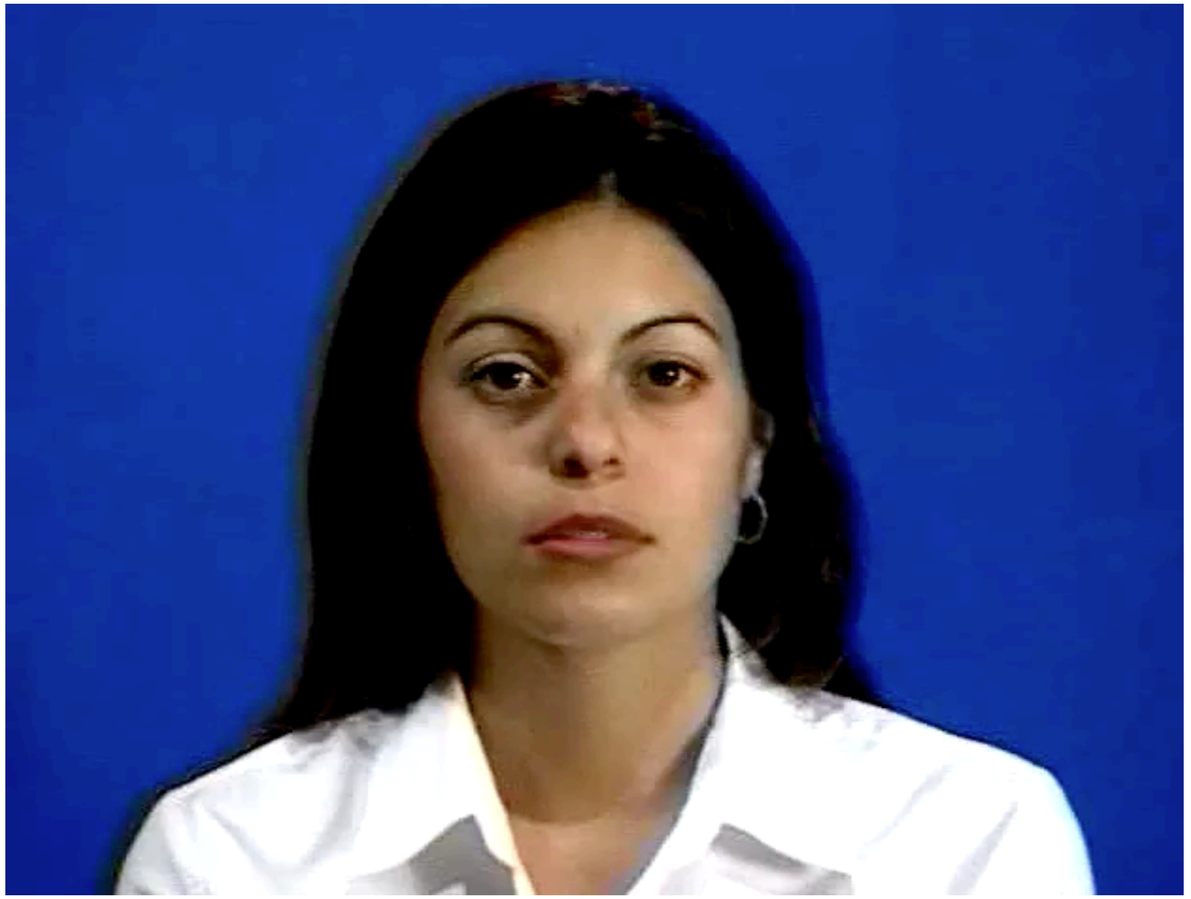
System learns from 4 mins of video face appearance (Morphable Model) and speech dynamics of the person

2. Run Time

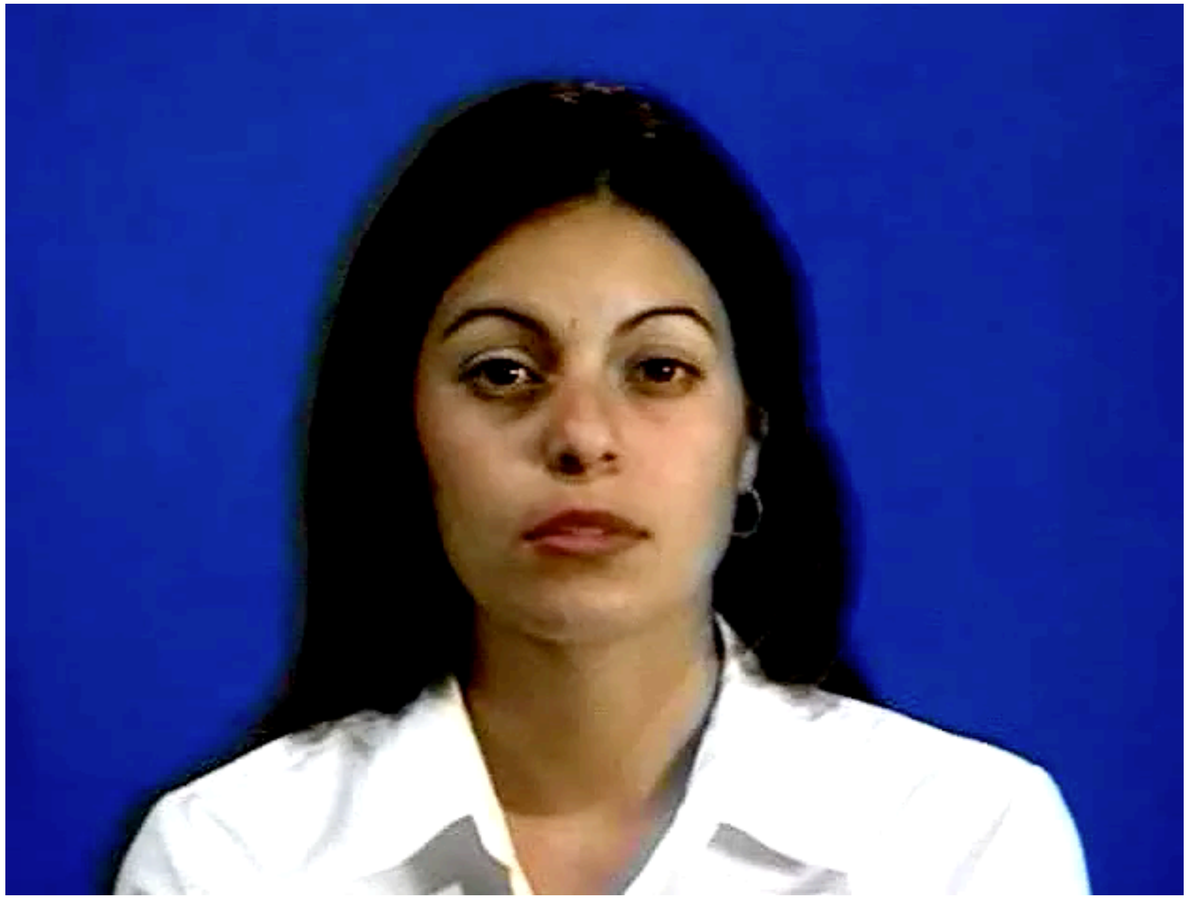
For any speech input the system provides as output a synthetic video stream



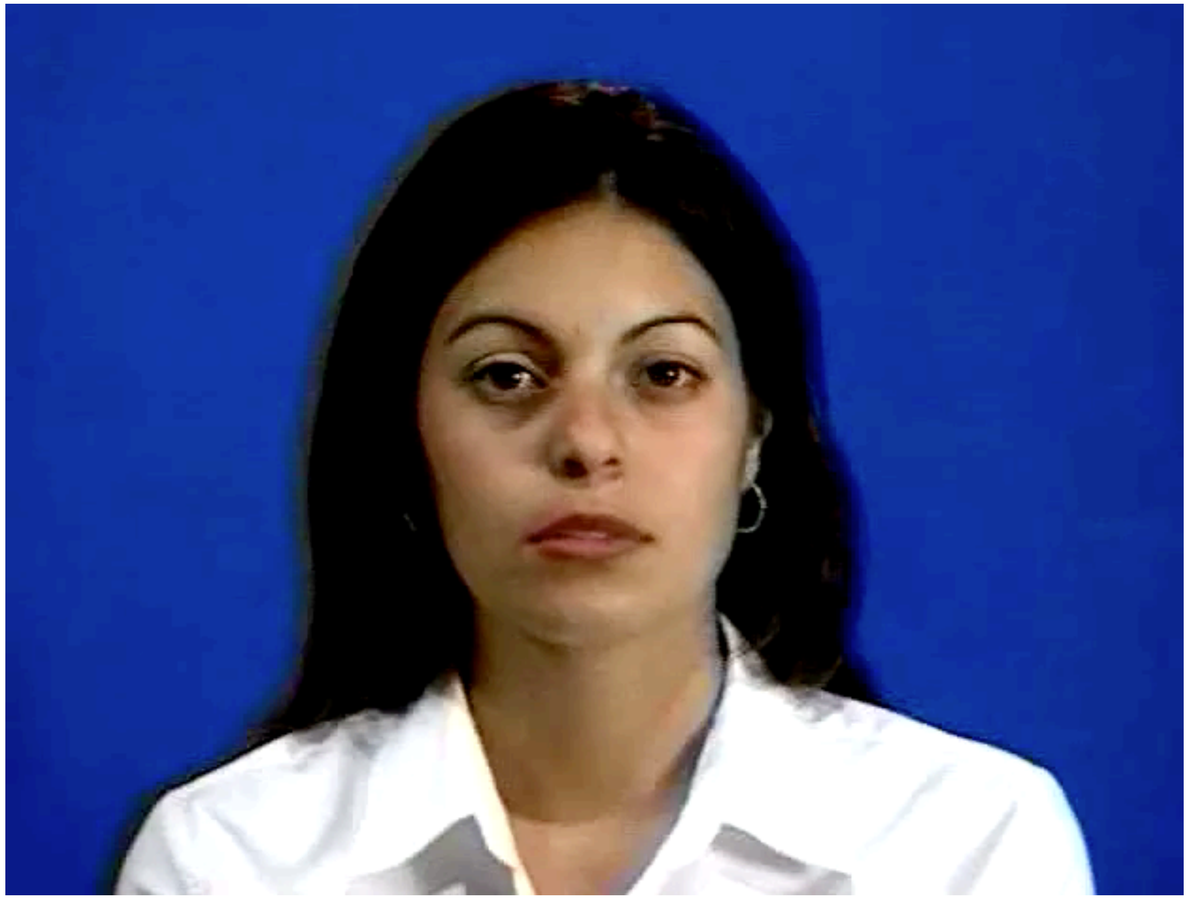




B-Dido



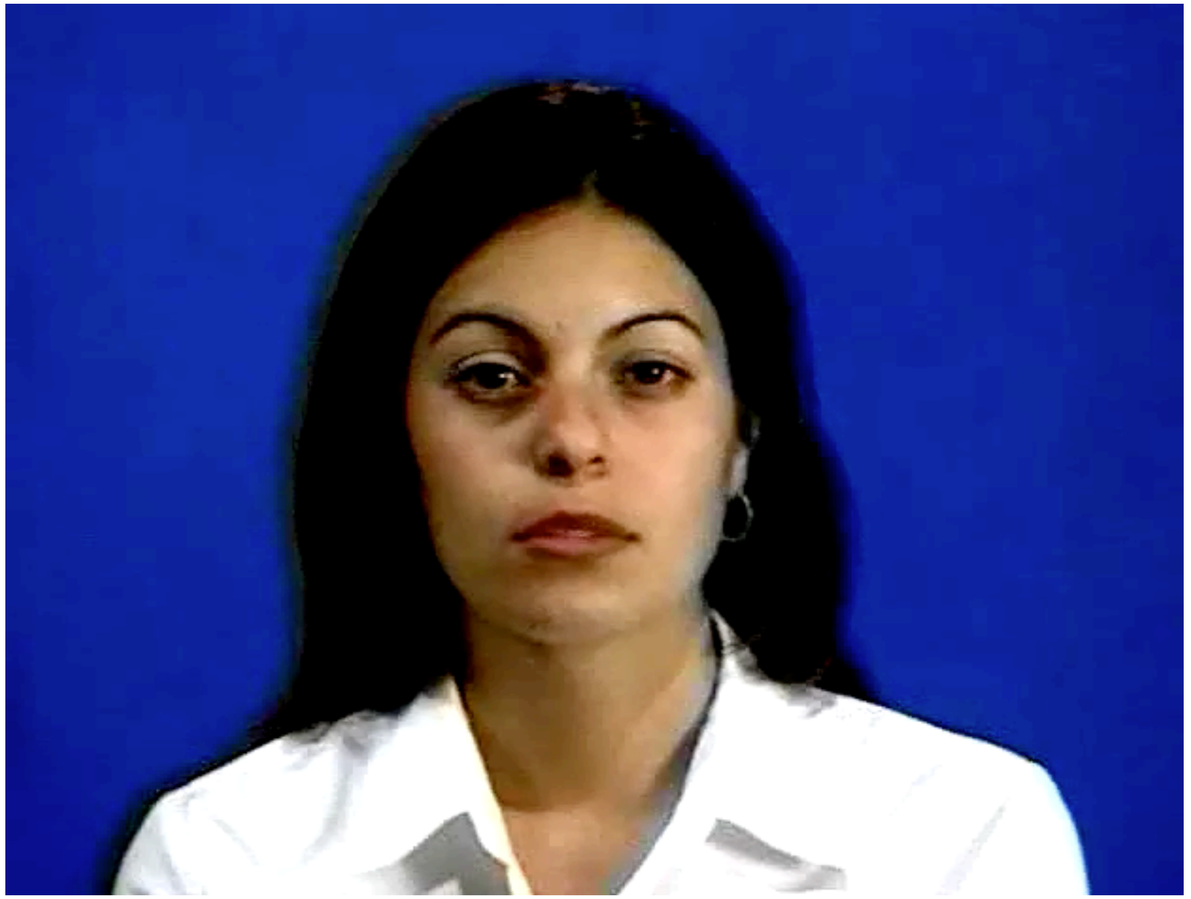
C-Hikaru



D-Denglijun

E-Marylin

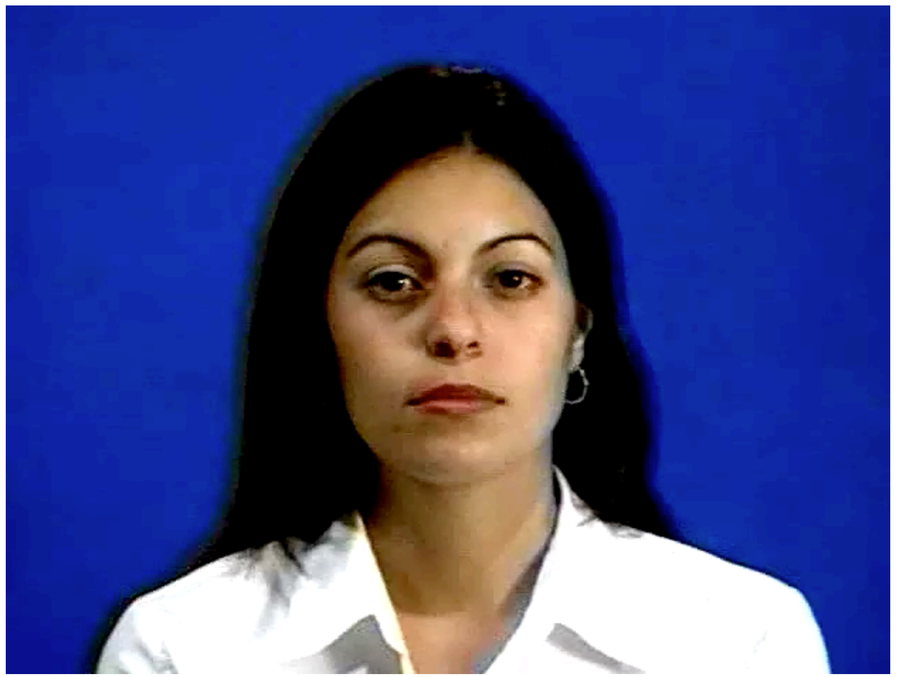
G-Katie



H-Rehema

I-Rehemax

A Turing test: what is real and what is synthetic?



L-real-synth

A Turing test: what is real and what is synthetic?

Experiment	# subjects	% correct	t	p<
Single pres.	22	54.3%	1.243	0.3
Fast single pres.	21	52.1%	0.619	0.5
Double pres.	22	46.6%	-0.75	0.5

Table 1: Levels of correct identification of real and synthetic sequences. t represents the value from a standard t-test with significance level of p<.

Tony Ezzat,Geiger, Poggio, **SigGraph 2002**

Opportunity for a good project!

Summary of today’s overview

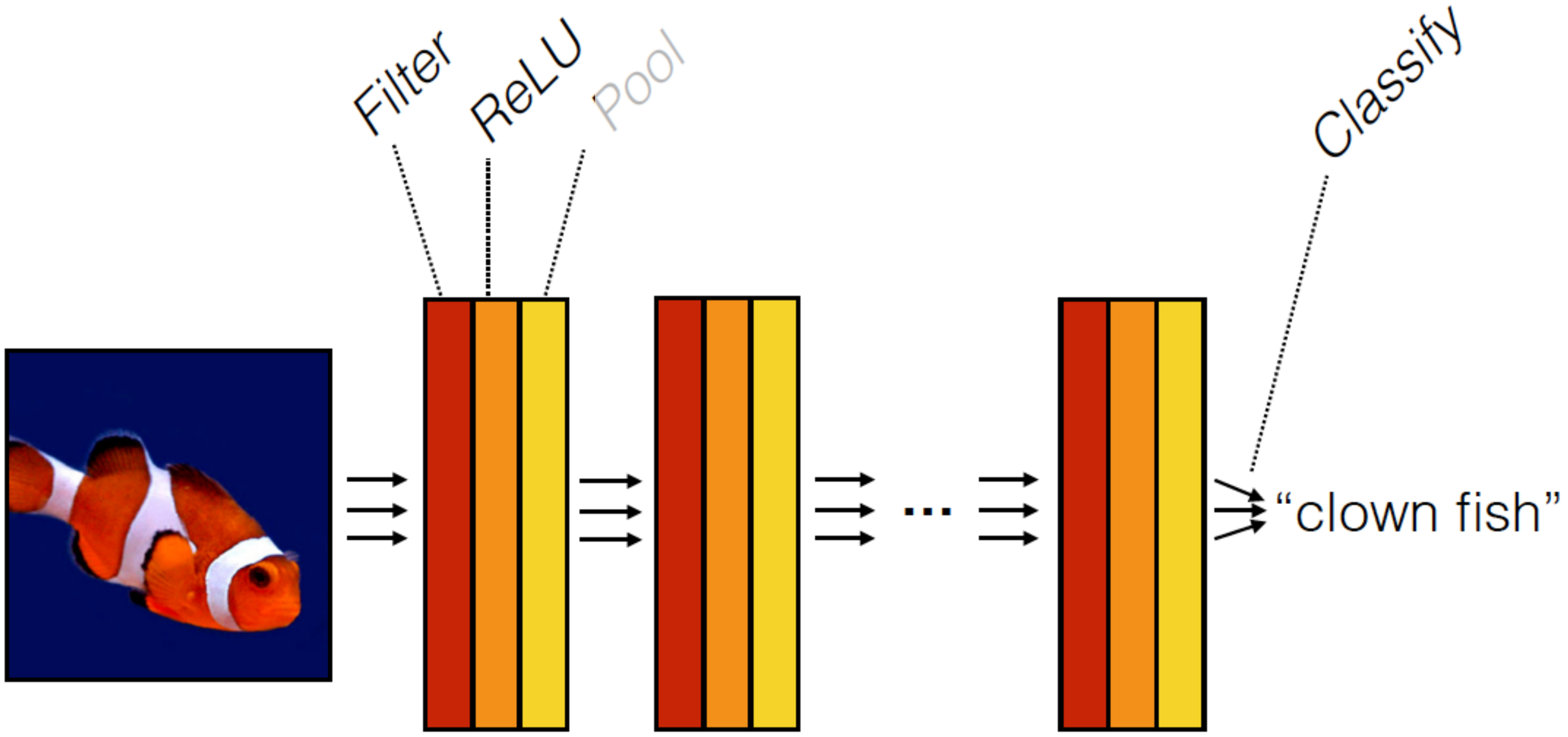
- A bit of history: old applications

Summary: I told you about old applications of ML, mainly kernel machines. I wanted to give you a feeling for how broadly powerful is the supervised learning approach: you can apply it to visual recognition, to decode neural data, to medical diagnosis, to finance, even to graphics. I also wanted to make you aware that ML does not start with deep learning and certainly does not finish with it.

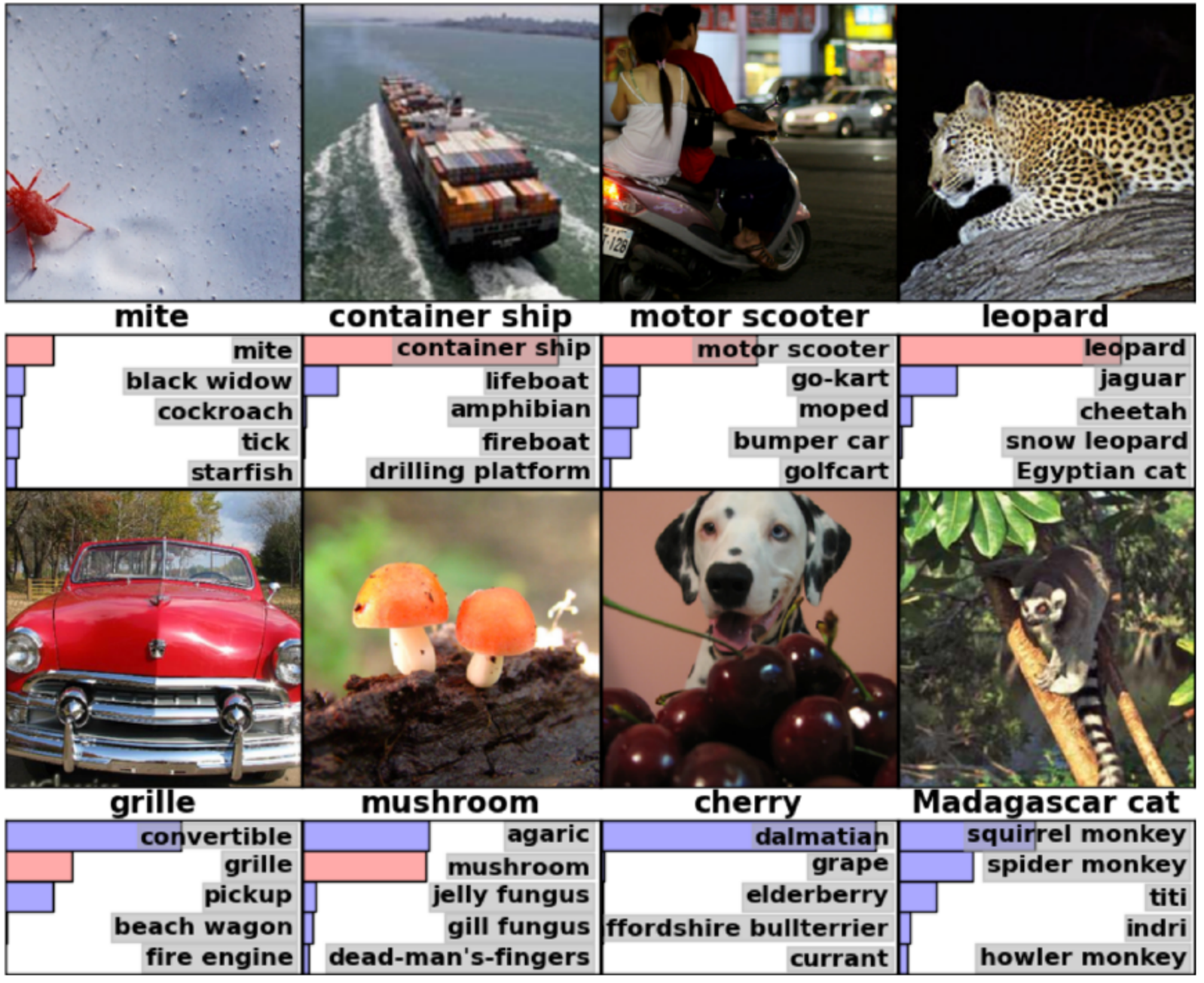
Today's overview

- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM
- A bit of history: Statistical Learning Theory, Neuroscience
- A bit of history: old applications
- Deep Learning, theory questions:
  - why depth works
  - why deep networks do not overfit
  - the challenge of sampling complexity

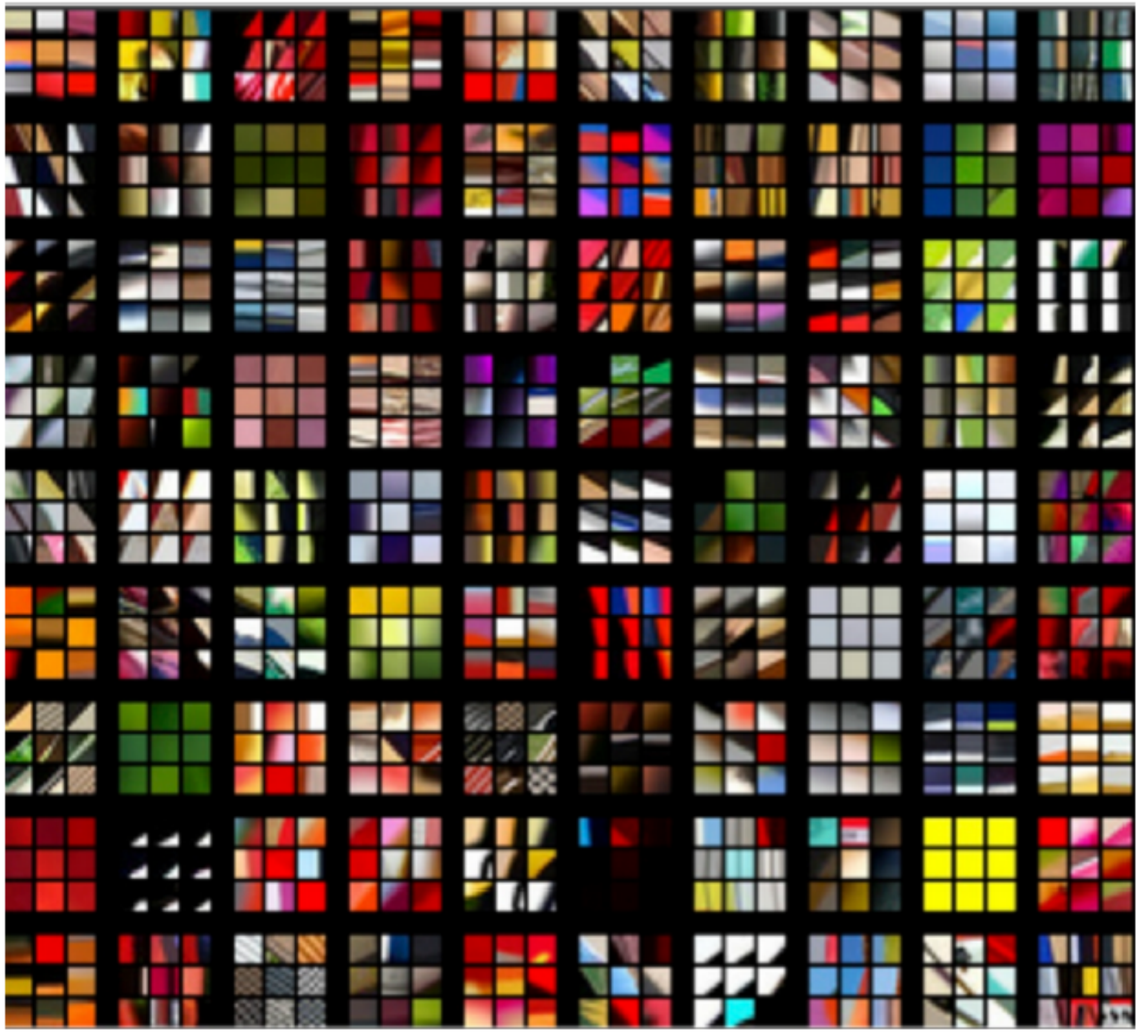
Computation in a neural net



$$f(\mathbf{x}) = f_L(\ldots f_2(f_1(\mathbf{x})))$$



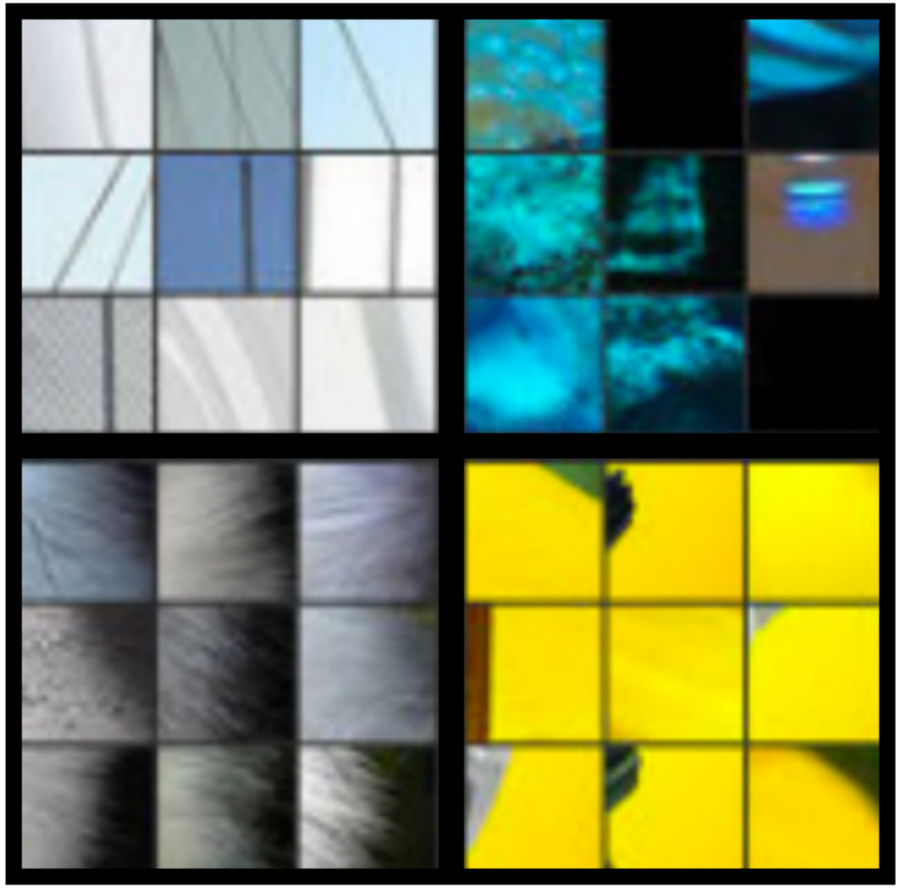
Krizhevsky et al. NIPS 2012



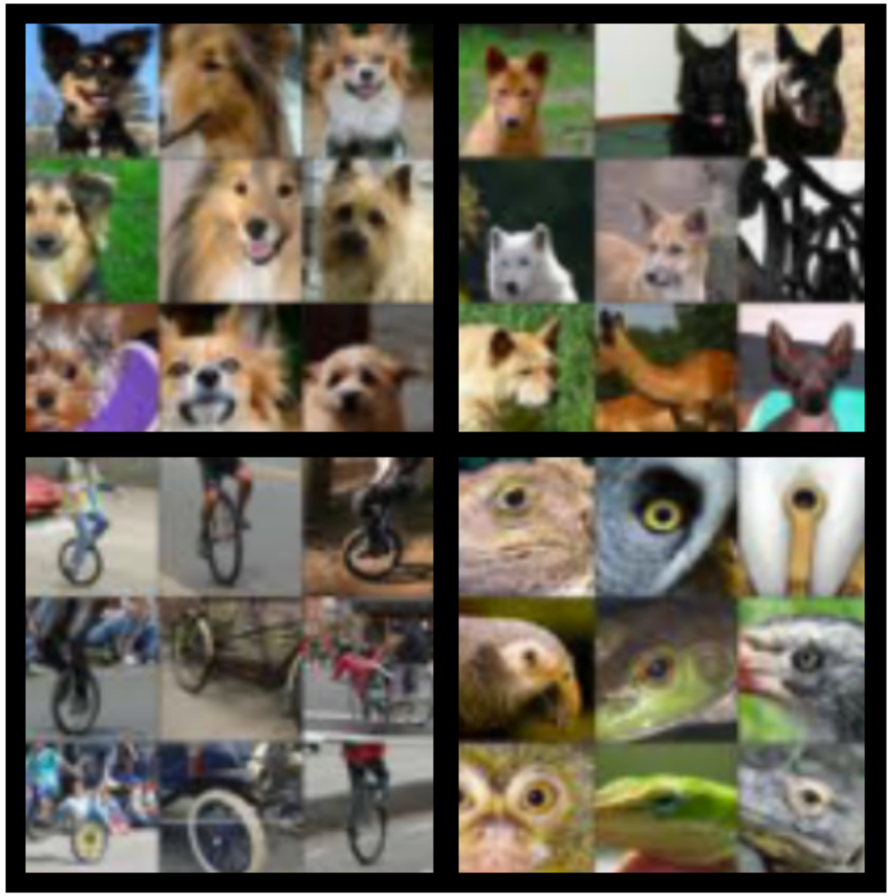
patches that strongly activate first layer filters

Zeiler & Fergus, ECCV 2014

Layer 2



Layer 5

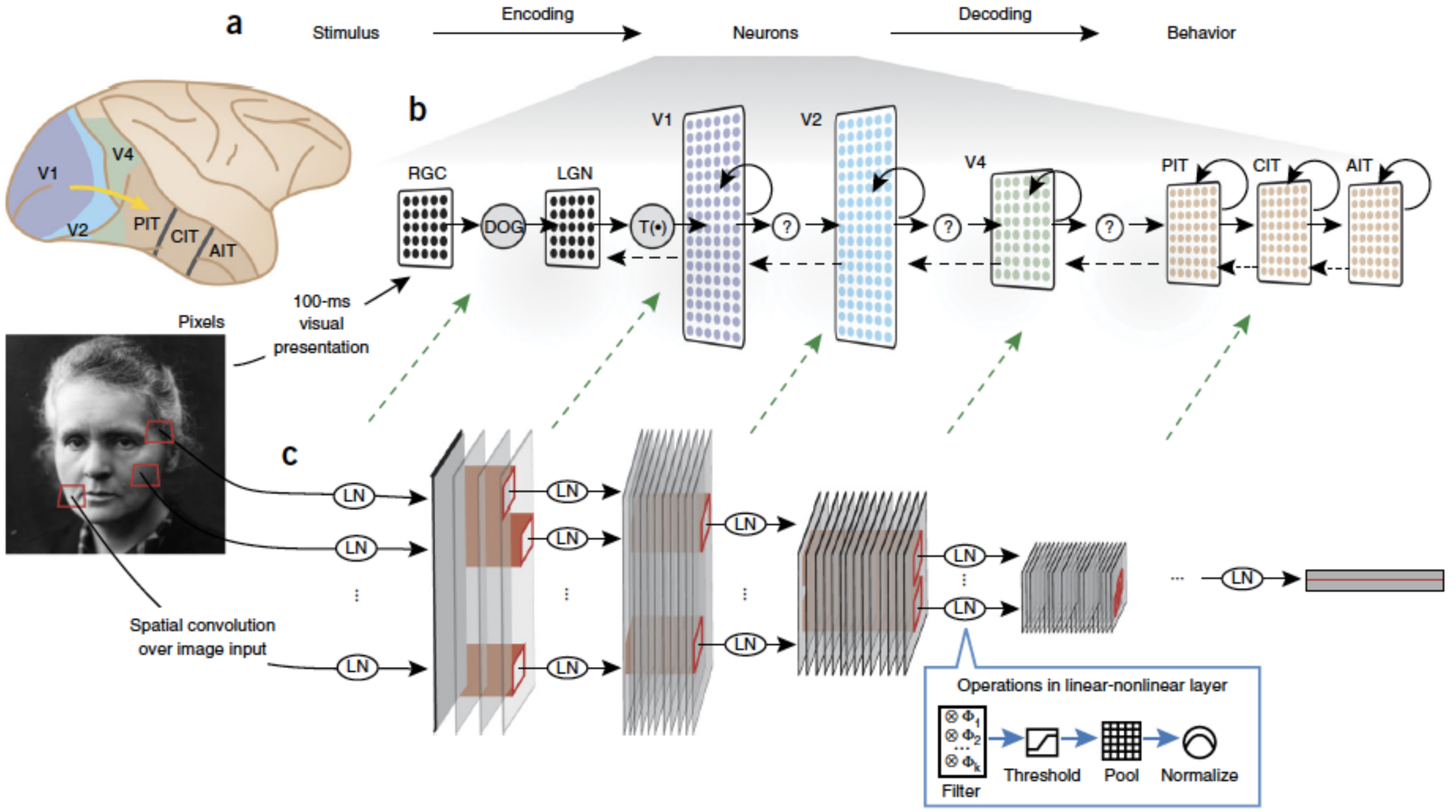


patches that strongly activate neurons on specified layer

Zeiler & Fergus, ECCV 2014

Using goal-driven deep learning models to understand sensory cortex

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



Deep nets : a theory is needed

Theories of Deep Learning (STATS 385)

Stanford University, Fall 2017

The spectacular recent successes of deep learning are purely empirical. Nevertheless intellectuals always try to explain important developments theoretically. In this literature course we will review recent work of Bruna and Mallat, Mhaskar and Poggio, Papyan and Elad, Bolcskei and co-authors, Baraniuk and co-authors, and others, seeking to build theoretical frameworks deriving deep networks as consequences. After initial background lectures, we will have some of the authors presenting lectures on specific papers. This course meets once weekly.

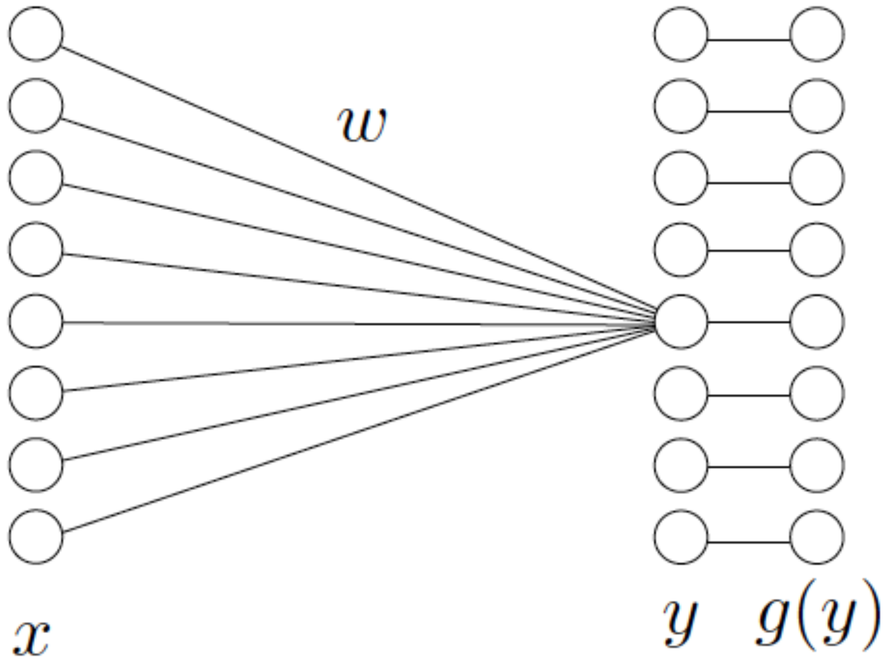
Instructors:


David Donoho

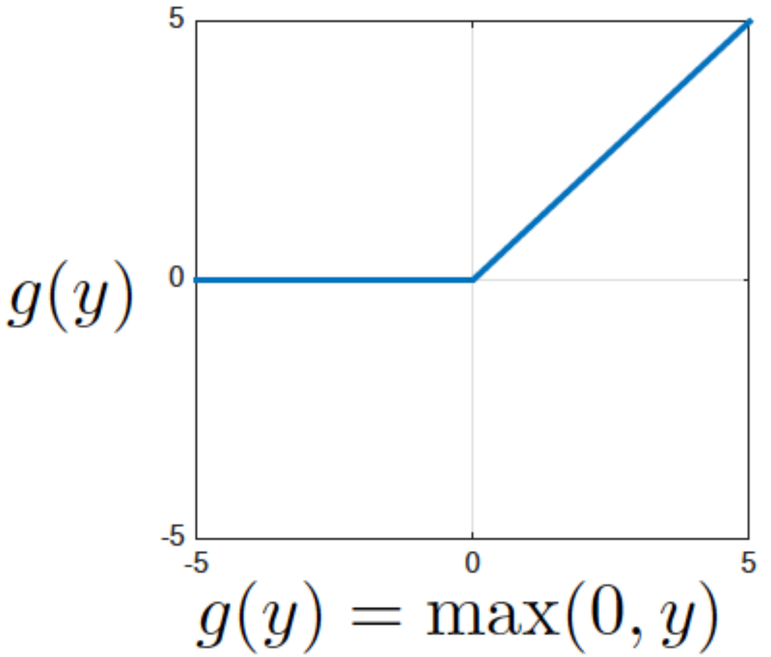

Hatef Monajemi


Vardan Papyan

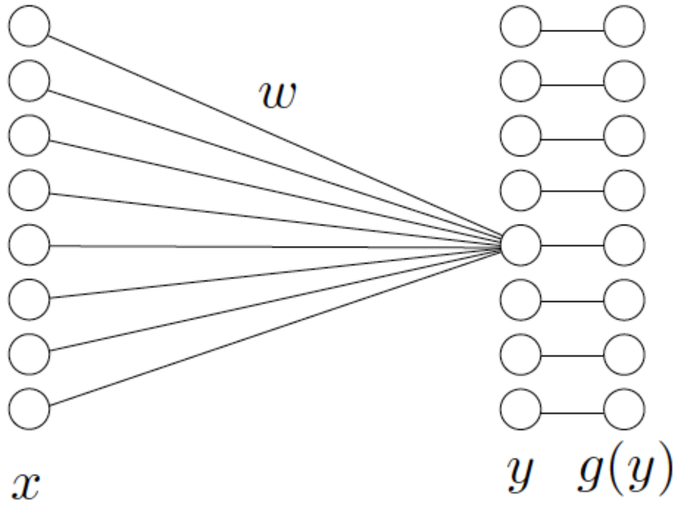
Computation in a neural net



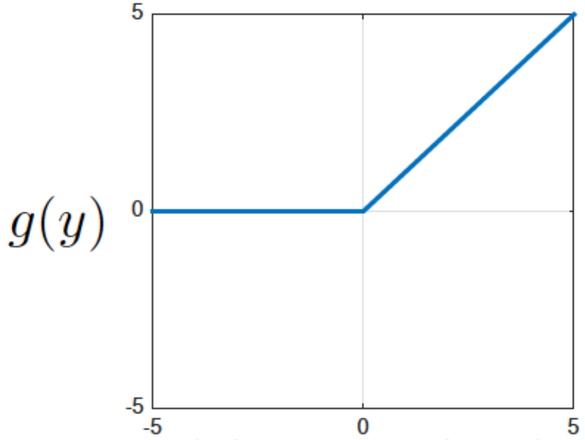
Rectified linear unit (ReLU)

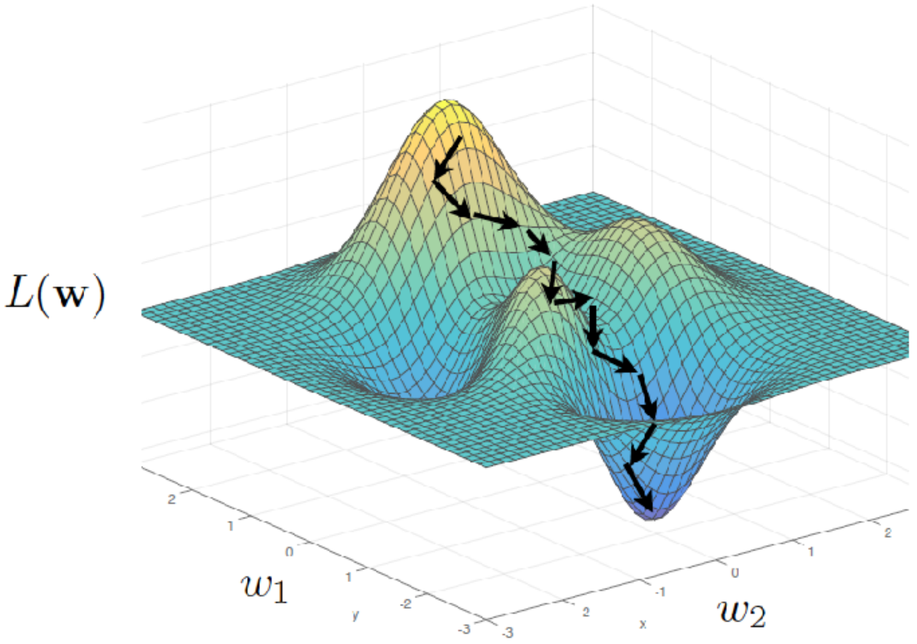


Deep nets architecture and SGD training



Rectified linear unit (ReLU)





CENTER FOR
Brains
Minds+
Machines

Gradient descent

$$\underset{\mathbf{w}}{\operatorname{argmin}} \quad \sum_i \ell(\mathbf{z}_i, f(\mathbf{x}_i; \mathbf{w})) = L(\mathbf{w})$$

One iteration of gradient descent:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t \frac{\partial L(\mathbf{w}^\mathbf{t})}{\partial \mathbf{w}}$$

learning rate

Summary of today's overview

- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM
- A bit of history: Statistical Learning Theory, Neuroscience
- A bit of history: old applications
- Deep Learning, theory questions
  - why depth works
  - why deep networks do not overfit
  - the challenge of sampling complexity

DLNNs: three main scientific questions

Approximation theory: when and why are deep networks better - no curse of dimensionality — than shallow networks?

Optimization: what is the landscape of the empirical risk?

Generalization by SGD: how can overparametrized networks generalize?

Work with Hrushikeshl Mhaskar+Lorenzo Rosasco+Fabio Anselmi+Chiyuan Zhang+Qianli Liao +Sasha Rakhlin + Noah G + Xavier B

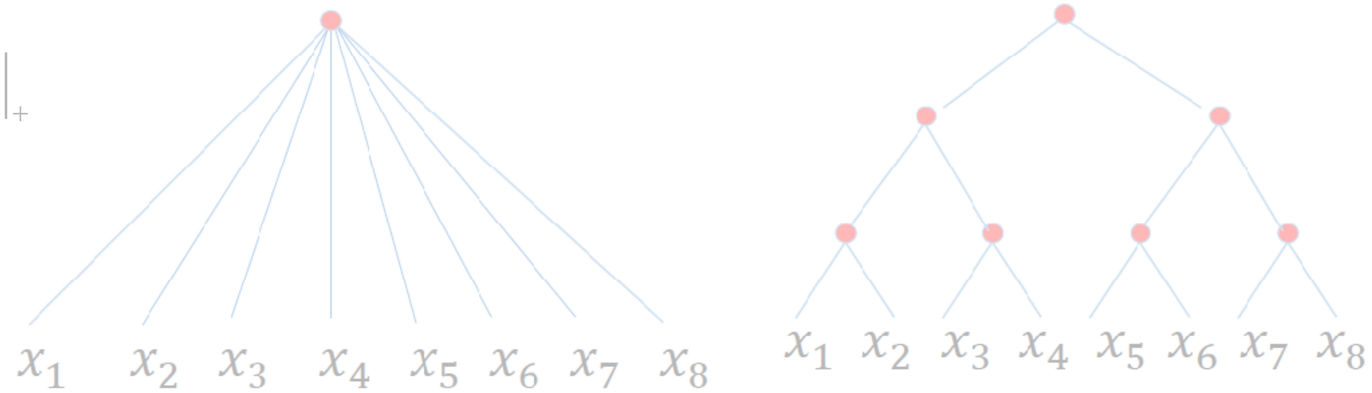
Minds + Machines

Opportunity for theory projects!

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

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

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



Theorem (informal statement)

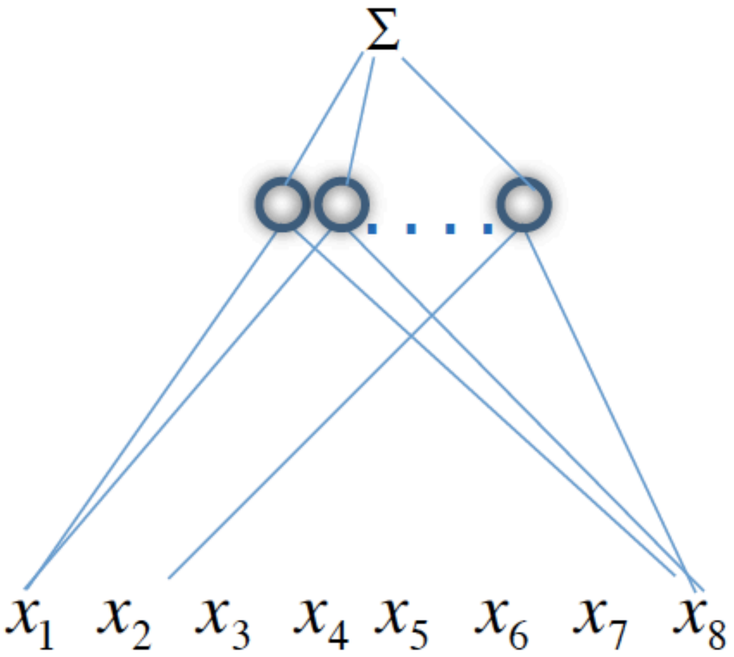
Suppose that a function of d variables is compositional . Both shallow and deep network can approximate f equally well. The number of parameters of the shallow network depends exponentially on d as $O(\varepsilon^{-d})$with the dimension whereas for the deep network dance is dimension independent, i.e.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

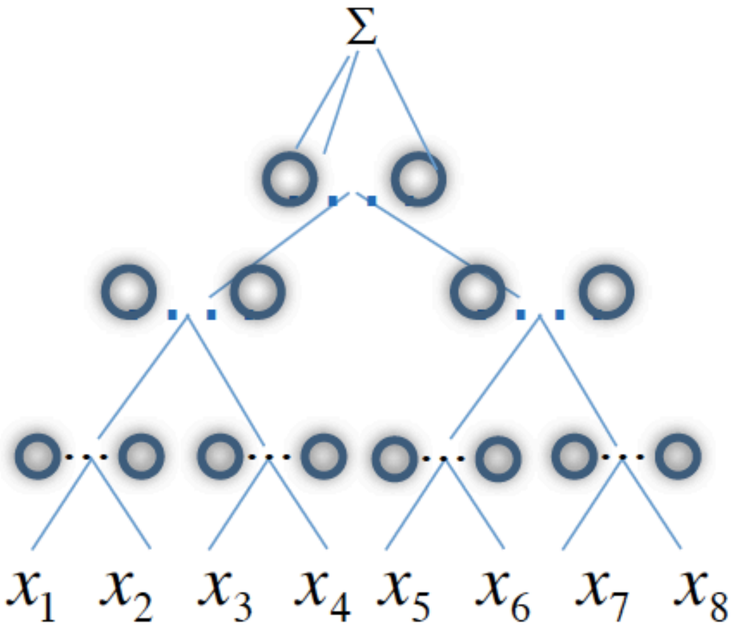
CENTER FOR Brains Minds+ Machines

Mhaskar, Poggio, Liao, 2016

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, ….

Curse of dimensionality

$$y = f(x_1, x_2, \ldots, x_8)$$

Both shallow and deep network can approximate a function of d variables equally well. The number of parameters in both cases depends exponentially on d as $O(\varepsilon^{-d})$.

CENTER FOR Brains Minds + Machines

Mhaskar, Poggio, Liao, 2016

When can the curse of dimensionality be avoided

Generic functions

$$f(x_1, x_2, \ldots, x_8)$$

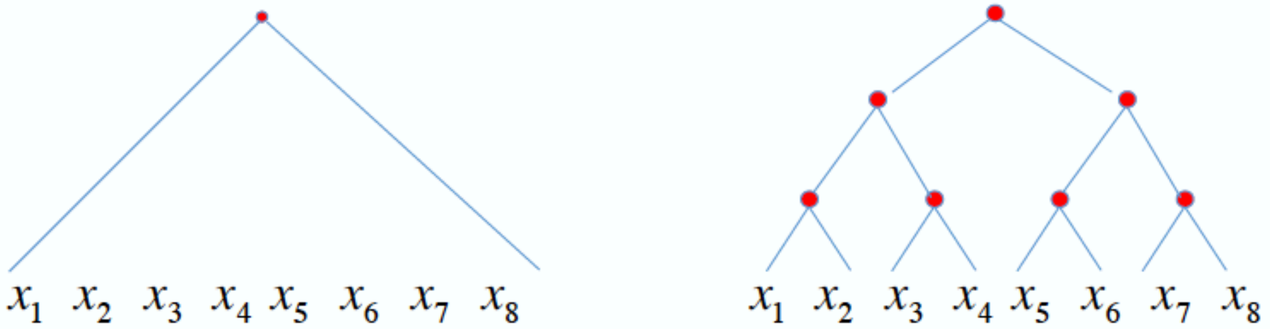
Compositional functions

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

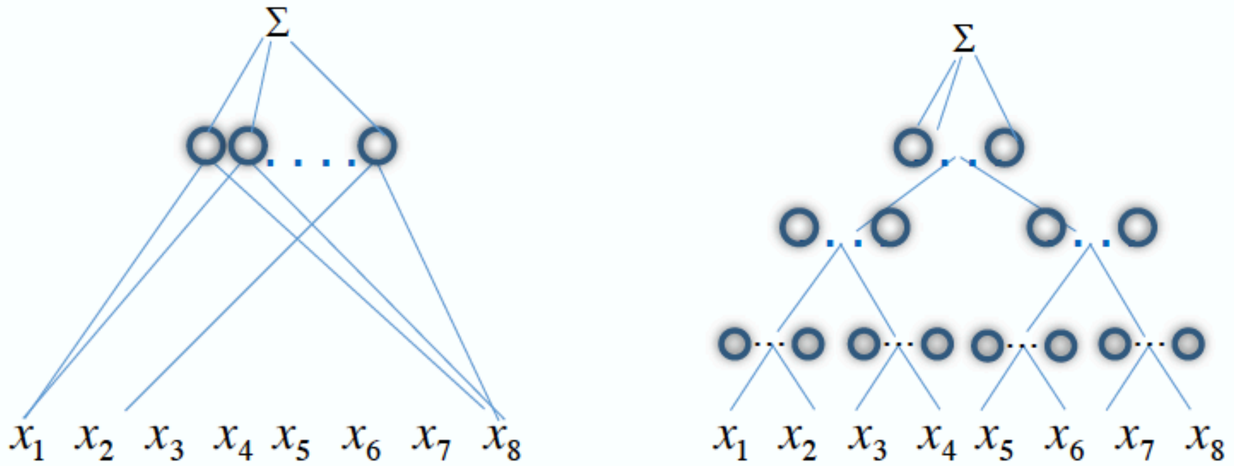
Mhaskar, Poggio, Liao, 2016

Microstructure of compositionality

target function



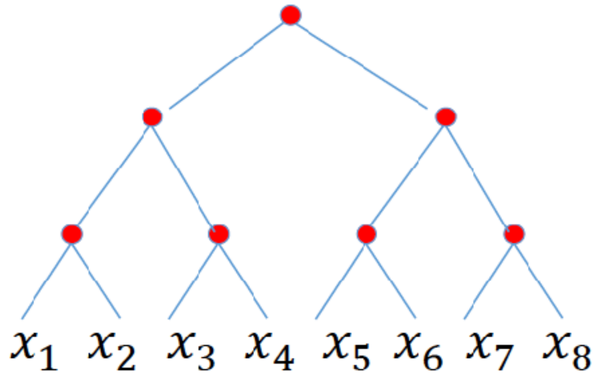
approximating function/network



a *b*

Hierarchically local compositionality

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



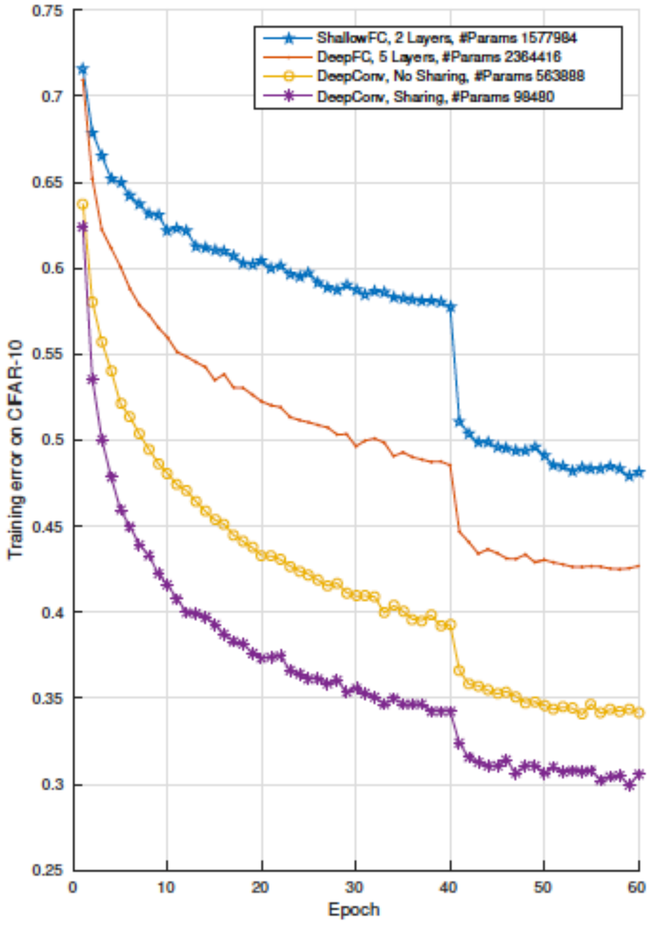
Theorem (informal statement)

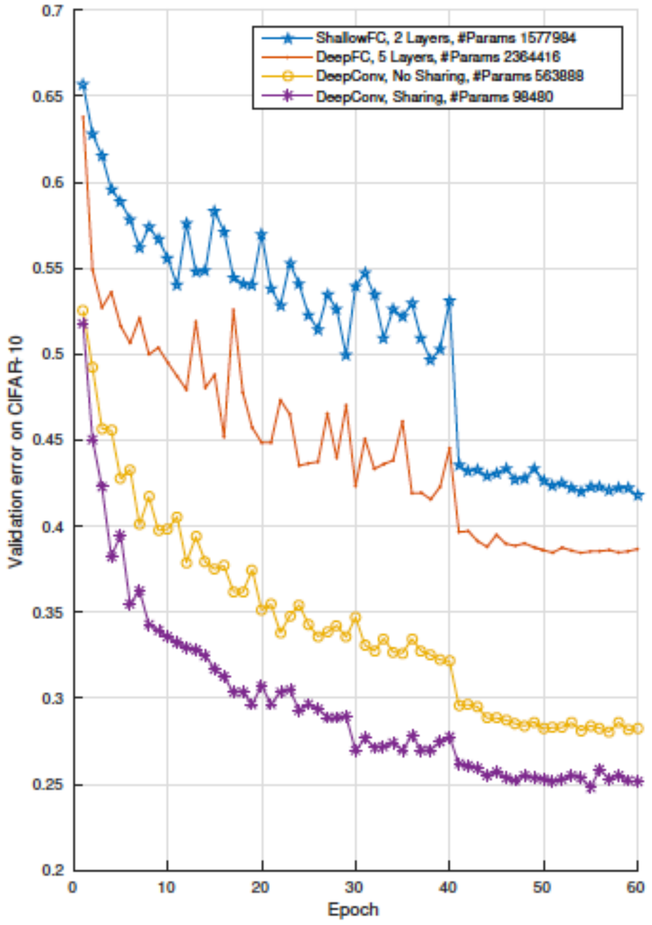
Suppose that a function of d variables is hierarchically, locally, compositional . Both shallow and deep network can approximate f equally well. The number of parameters of the shallow network depends exponentially on d as $O(\varepsilon^{-d})$ with the dimension whereas for the deep network dance is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

CENTER FOR Brains Minds + Machines

Mhaskar, Poggio, Liao, 2016

Locality of constituent functions is key not weight sharing: CIFAR





Open problem: why compositional functions are important for perception?

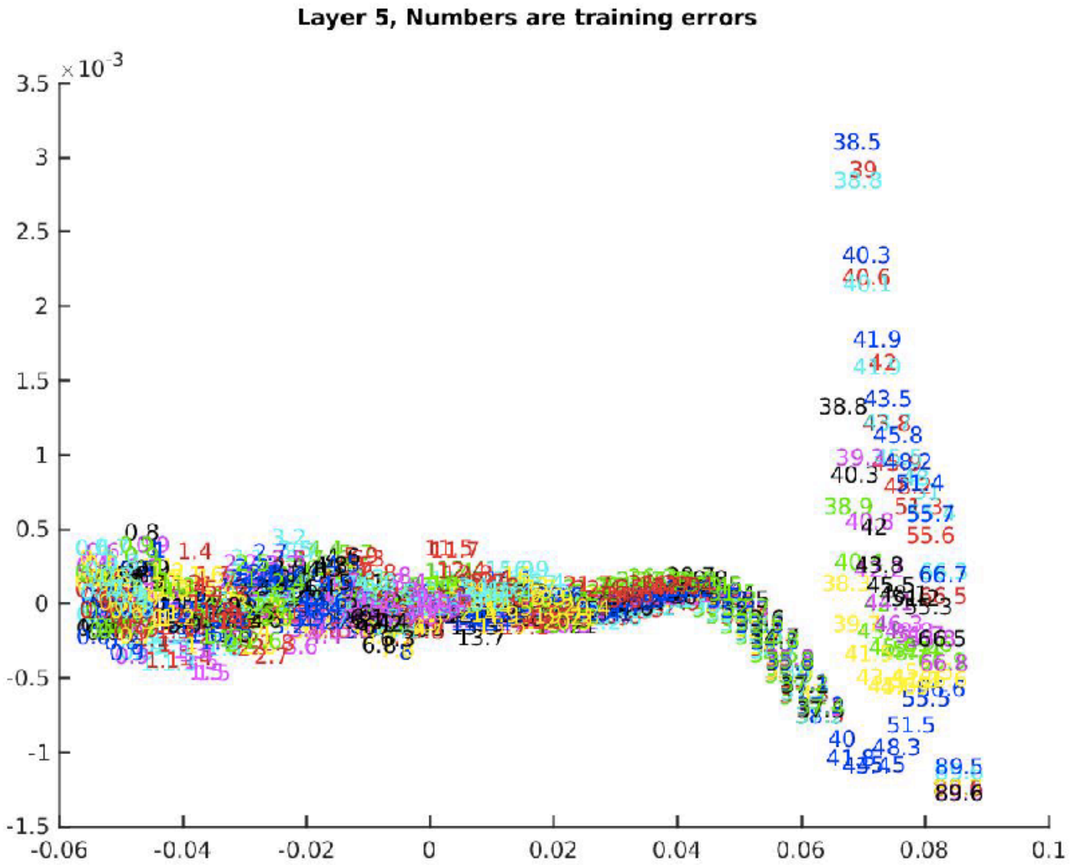
Which one of these reasons:
Physics?
Neuroscience? <===
Evolution?

Opportunity for theory projects!

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

Theorem (informal statement)

Replacing the RELUs with univariate polynomial approximation, Bezout theorem implies that the system of polynomial equations corresponding to zero empirical error has a very large number of degenerate solutions. The global zero-minimizers correspond to flat minima in many dimensions (generically unlike local minima). Thus SGD is biased towards finding global minima of the empirical risk.



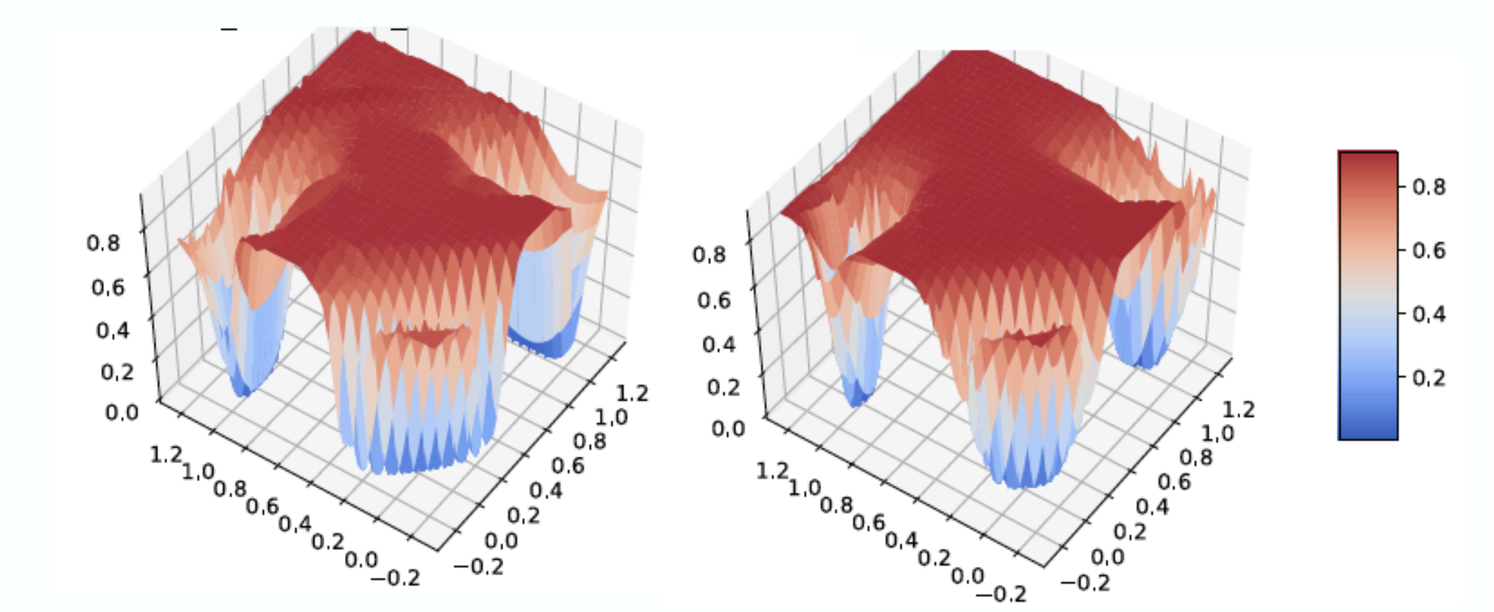
CENTER FOR Brains Minds + Machines

Liao, Poggio, 2017

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

Results

- SGD finds with very high probability large volume, flat zero-minimizers;
- Flat minimizers correspond to degenerate zero-minimizers and thus to global minimizers;
- SGD minimizers select minima that correspond to small norm solutions and “good” expected error;



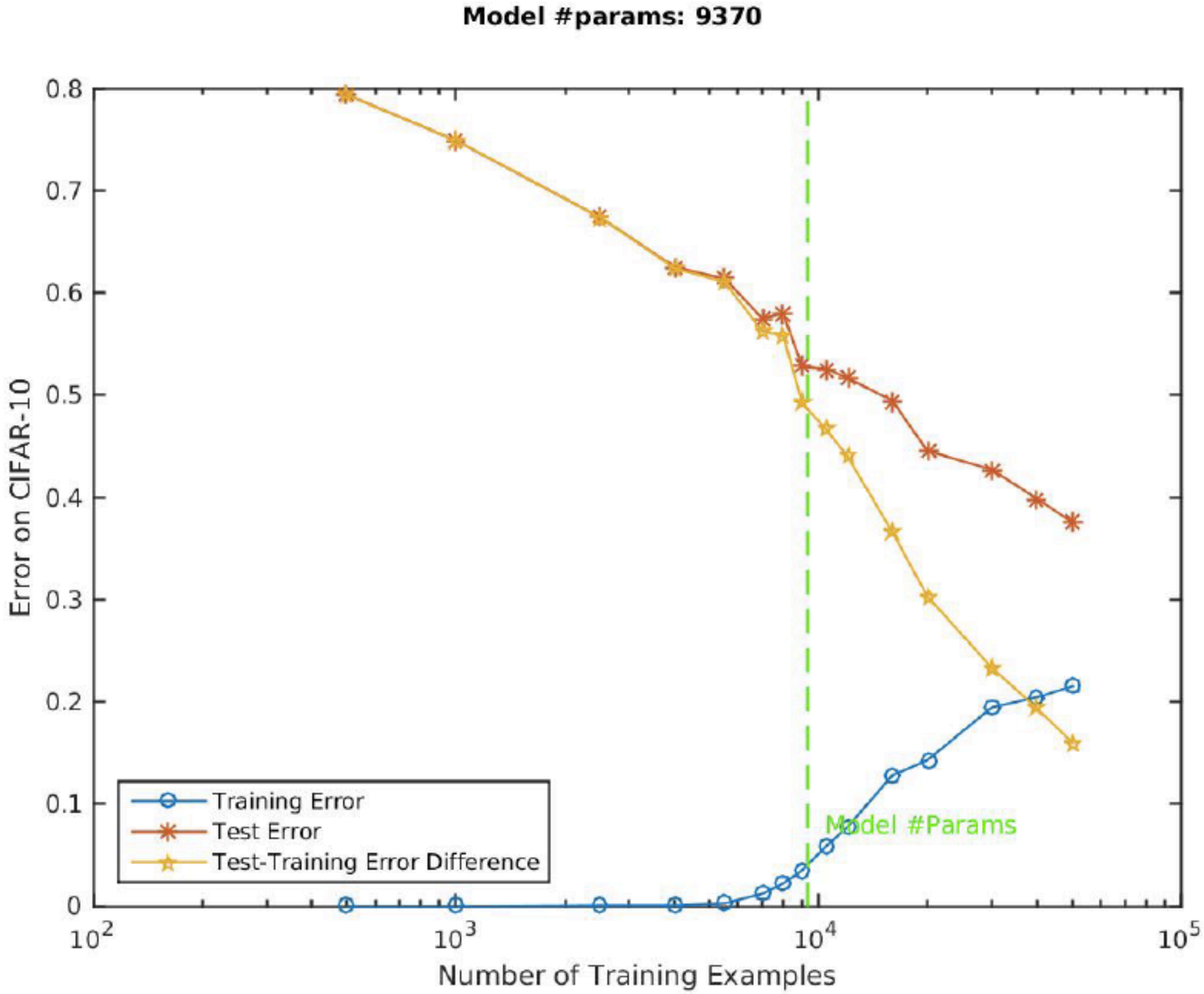
CIFAR-10: Natural Labels

Random Labels

Poggio, Rakhlin, Golovitc, Zhang, Liao, 2017

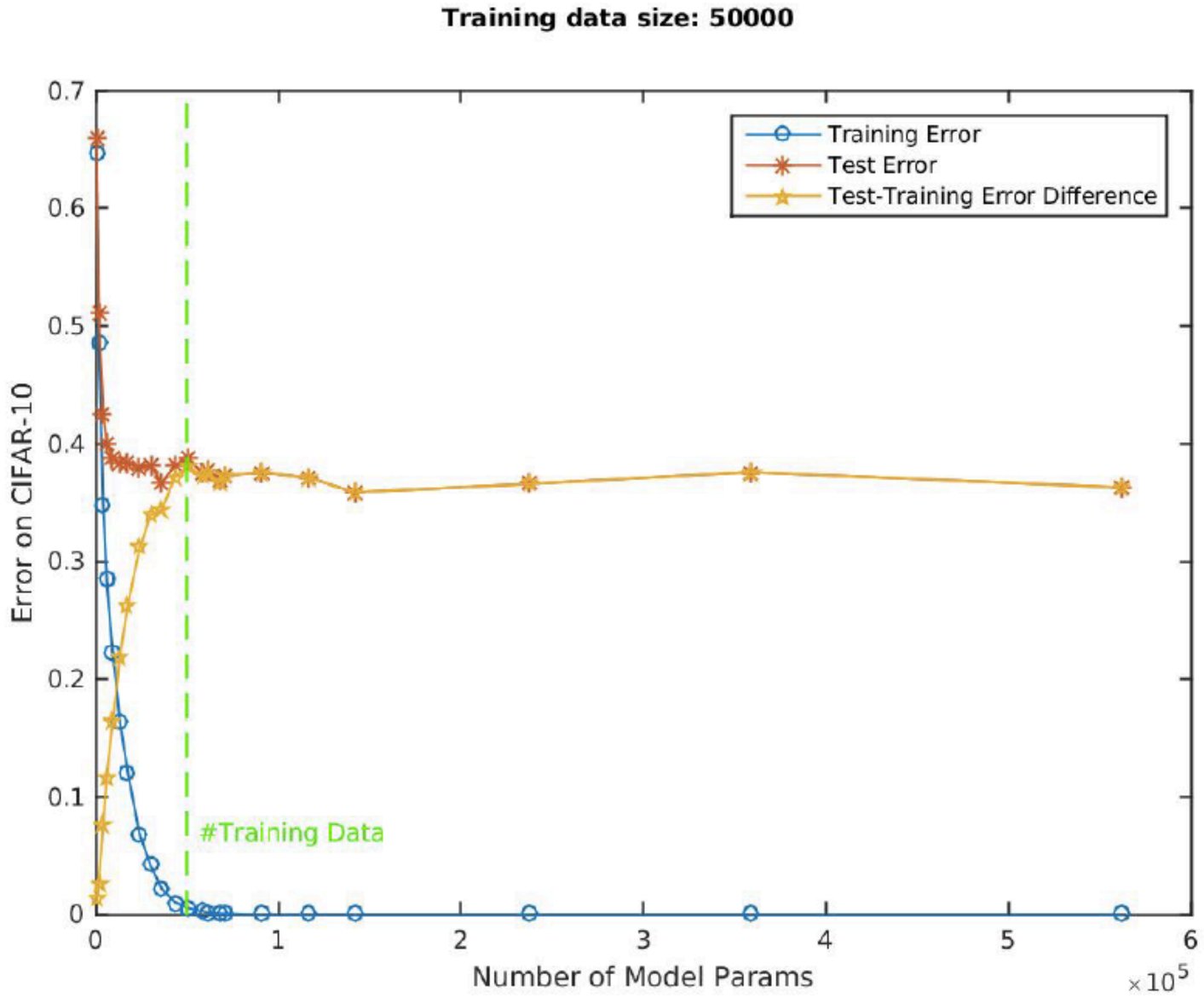
CENTER FOR
Brains
Minds
Machi

Good generalization with less data than # weights



CENTER FOR
Brains
Minds+
Machines

No overfitting



CENTER FOR
Brains
Minds+
Machines

Beyond today's DLNNs:
several scientific questions…

Why do Deep Learning Networks work? ===>

In which cases will they fail?

Is it possible to improve them?

Is it possible to reduce the number of labeled examples?

CENTER FOR
Brains
Minds+
Machines

Opportunity for a good project!

Beyond today's DLNNs: neurocognitive science

- State-of-the-art DLNNs require ~1M labeled examples
- This is not how we learn, how children learn

Today's science, tomorrow's engineering: learn like children learn

The first phase (and successes) of ML:
supervised learning, big data: $n \to \infty$

from programmers…
…to labelers…
…to computers that learn like children…

The next phase of ML: implicitly supervised learning,
learning like children do, small data: $n \to 1$

Summary of today's overview

- Deep Learning, theory questions:
  - why depth works
  - why deep networks do not overfit
  - the challenge of sampling complexity

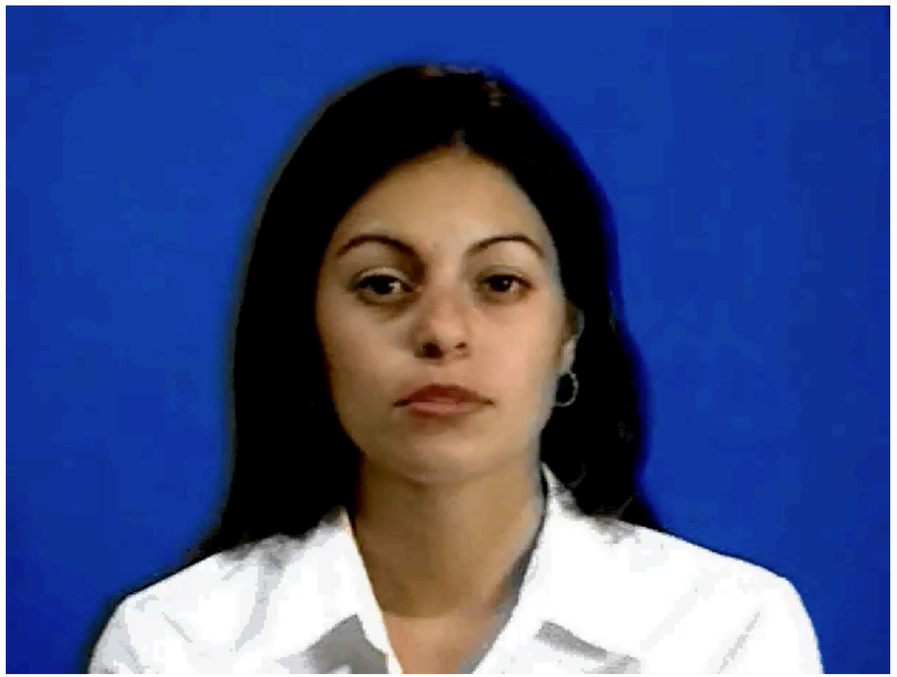
Summary: I told you why and when deep learning can avoid the curse of dimensionality while shallow nets cannot. I told you why SGD finds global minima and why they are likely to exist in overparametrized networks. I told you how the theory you learned in class 2-9 explain the puzzle of non-overfitting and good generalization by deep nets.

Old applications

Old applications

Old applications

Mary101



A- more in a moment

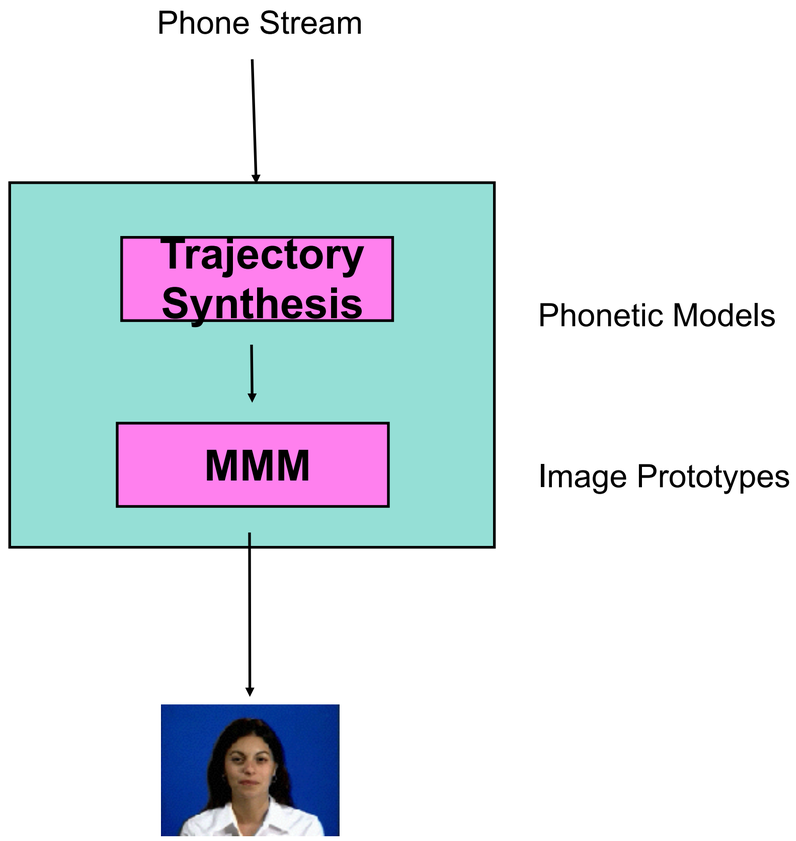
Tony Ezzat,Geiger, Poggio, **SigGraph 2002**

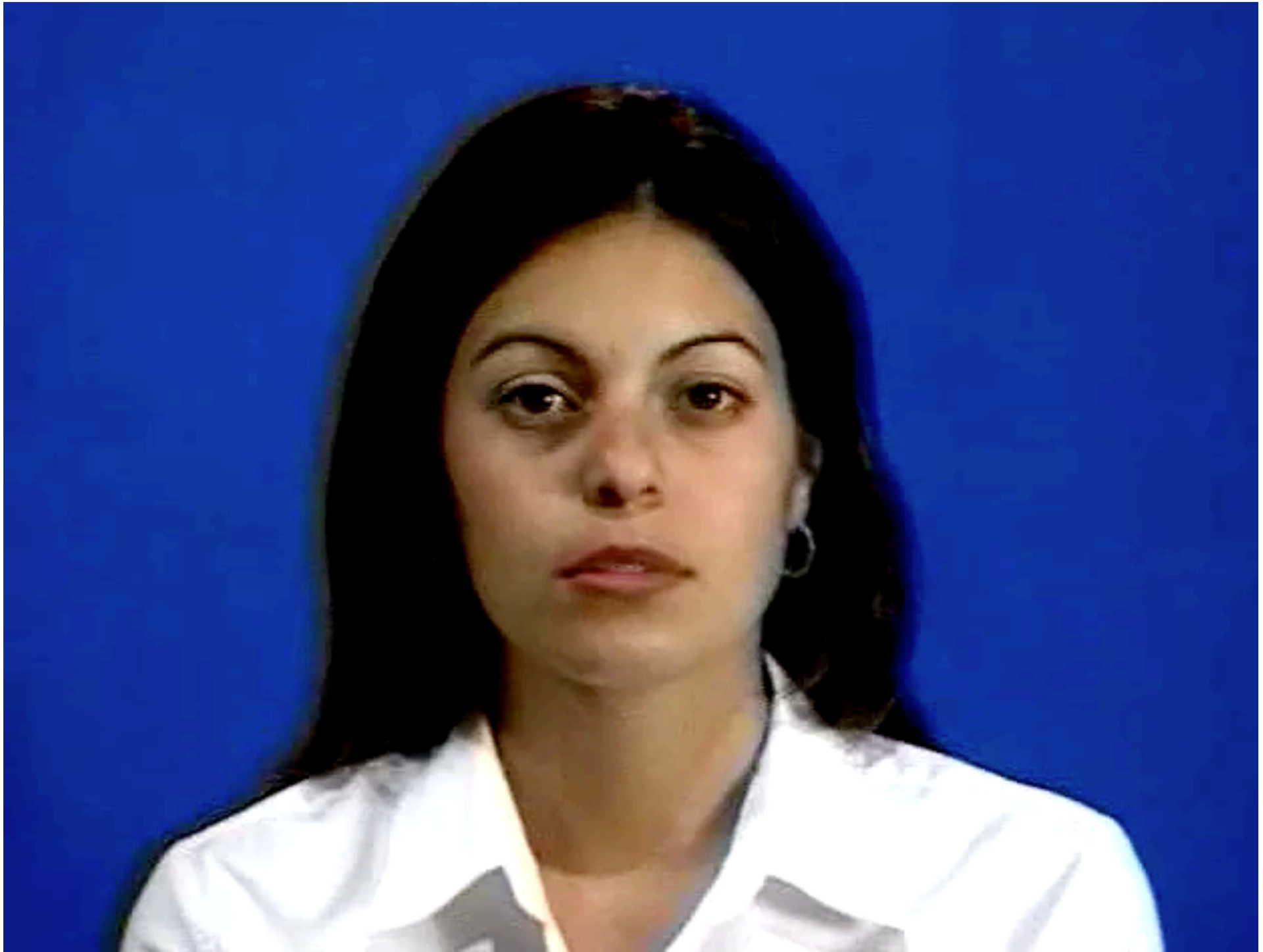
1. Learning

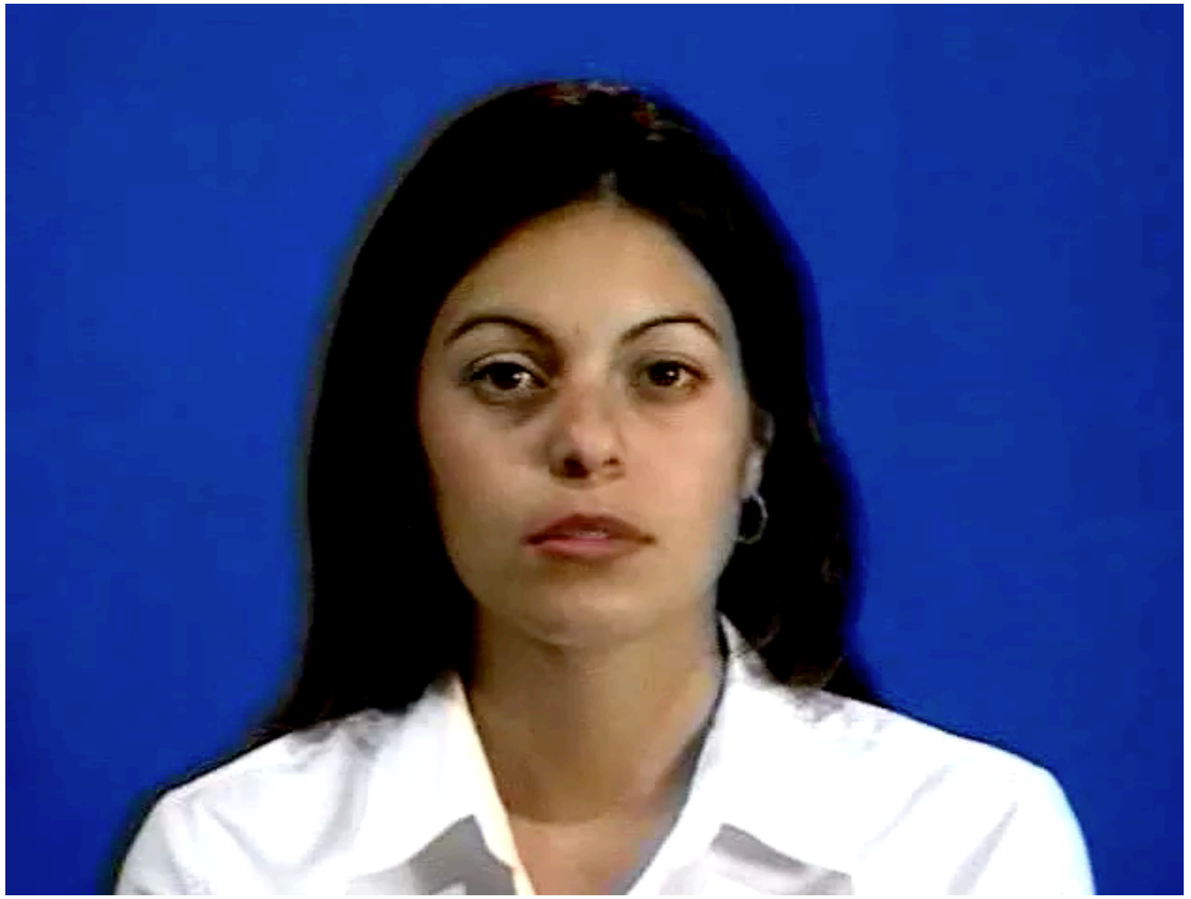
System learns from 4 mins of video face appearance (Morphable Model) and speech dynamics of the person

2. Run Time

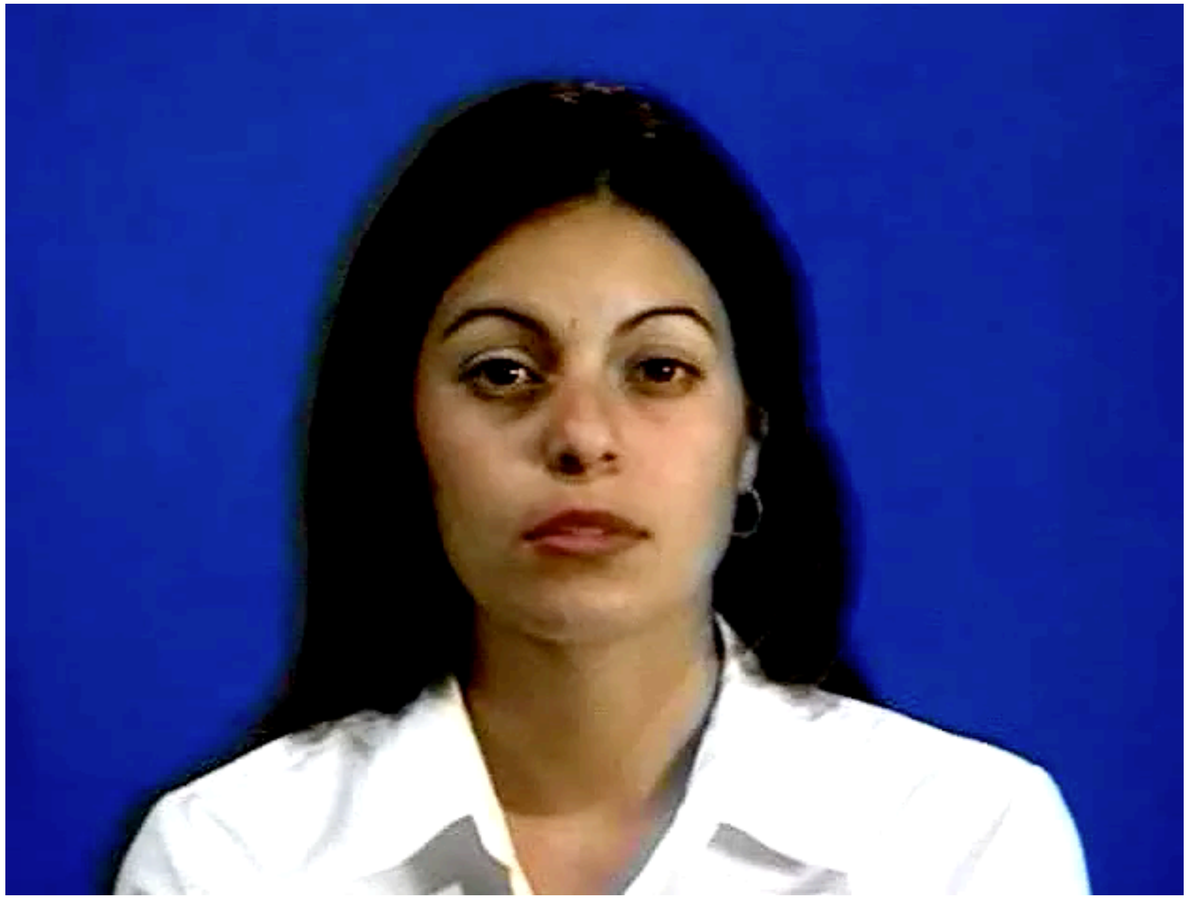
For any speech input the system provides as output a synthetic video stream



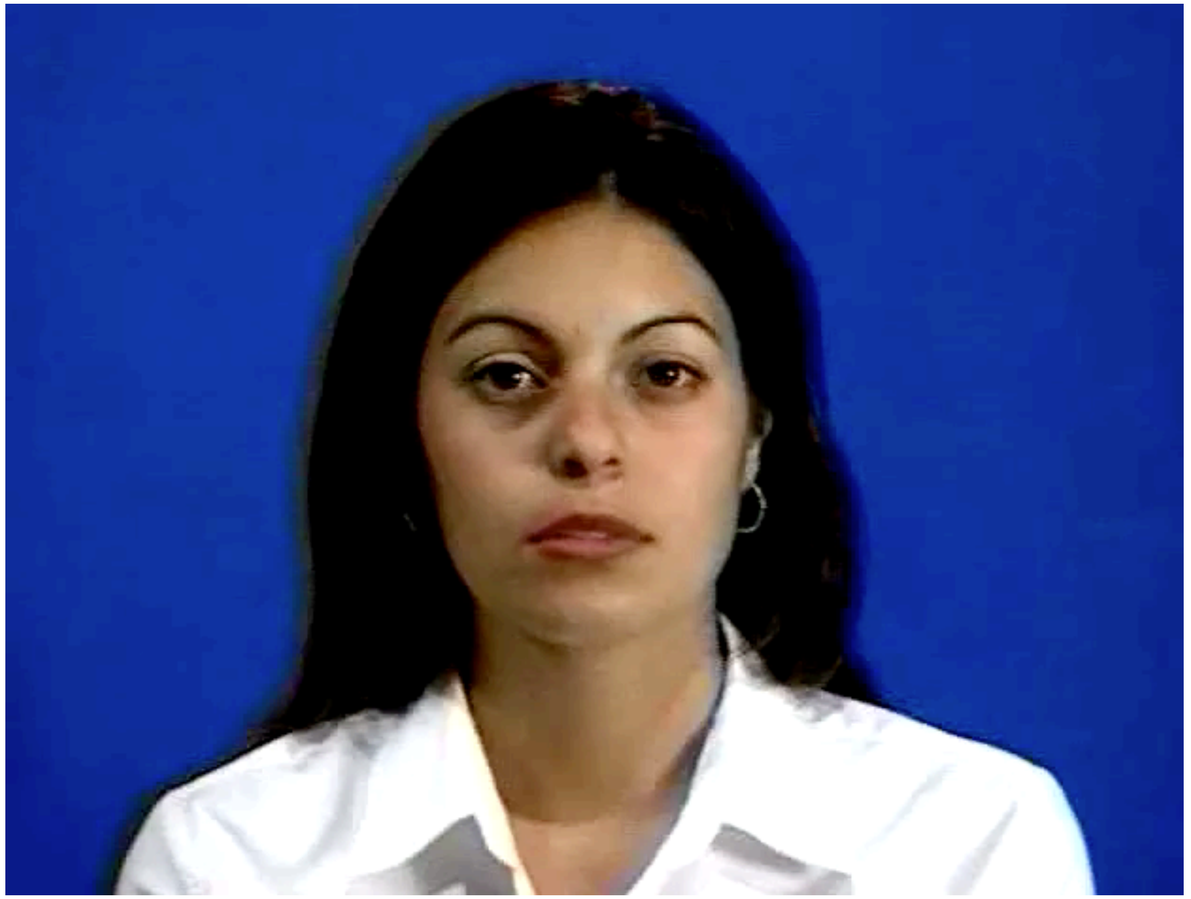




B-Dido



C-Hikaru

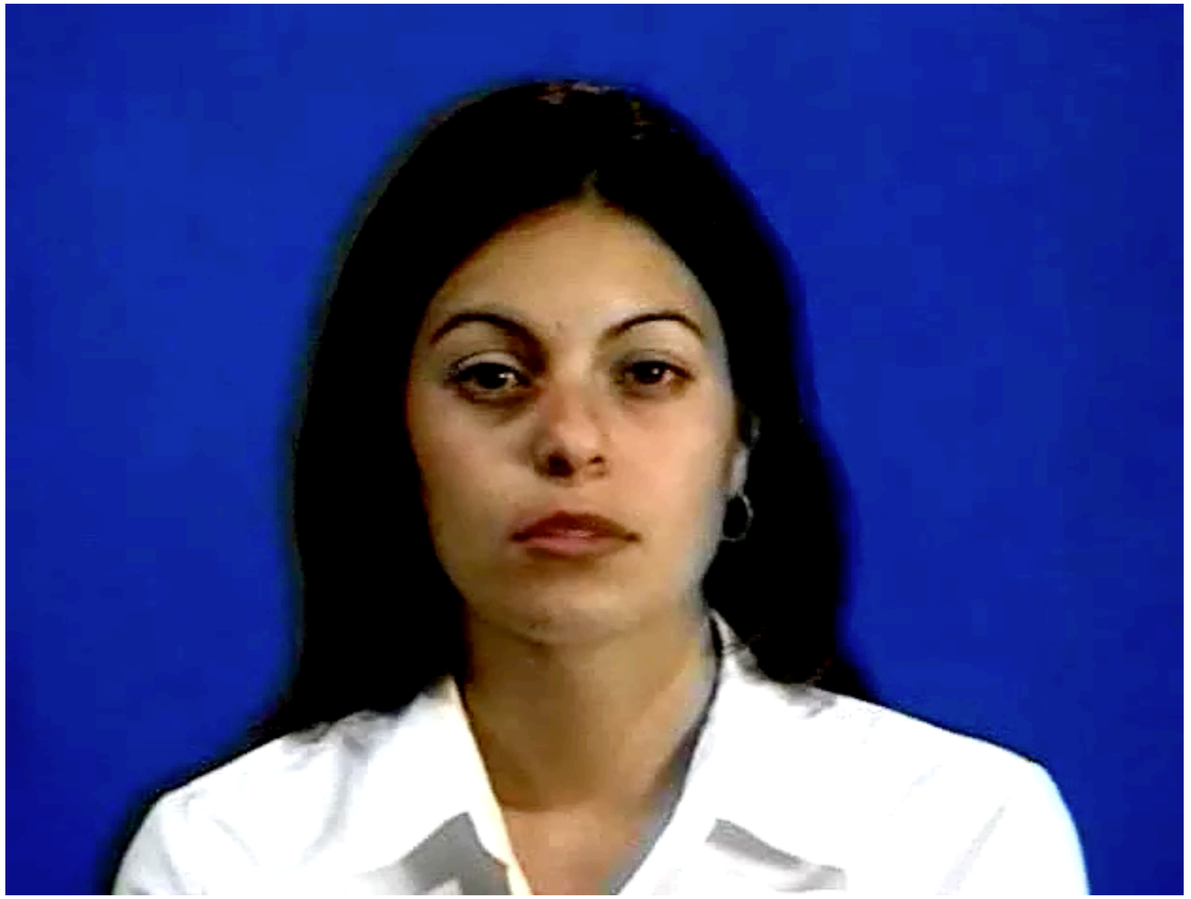


D-Denglijun

E-Marylin

F-Katie Couric

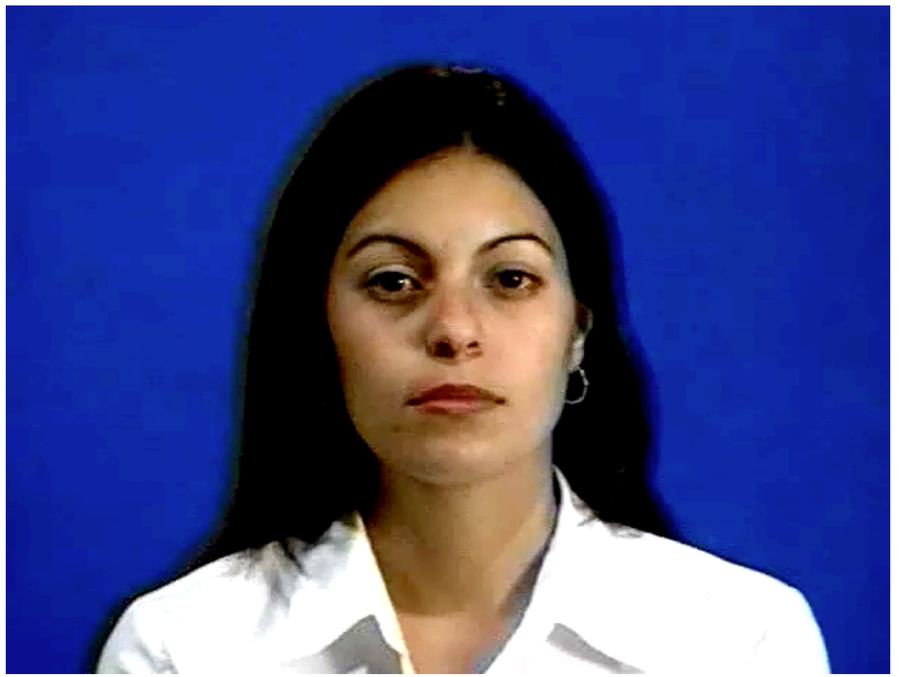
G-Katie



H-Rehema

I-Rehemax

A Turing test: what is real and what is synthetic?



L-real-synth

A Turing test: what is real and what is synthetic?

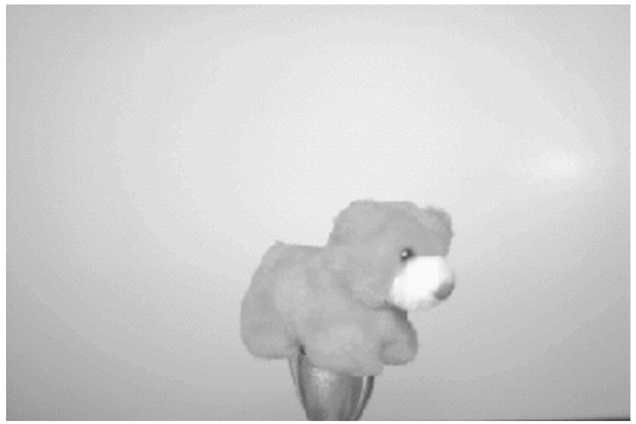
Experiment	# subjects	% correct	t	p<
Single pres.	22	54.3%	1.243	0.3
Fast single pres.	21	52.1%	0.619	0.5
Double pres.	22	46.6%	-0.75	0.5

Table 1: Levels of correct identification of real and synthetic sequences. t represents the value from a standard t-test with significance level of p<.

Tony Ezzat,Geiger, Poggio, **SigGraph 2002**

Learning: image analysis

⇒ **Bear (0° view)**



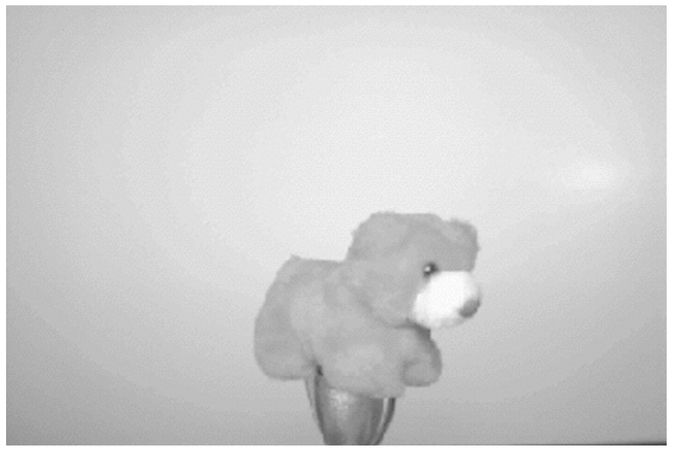
⇒ **Bear (45° view)**

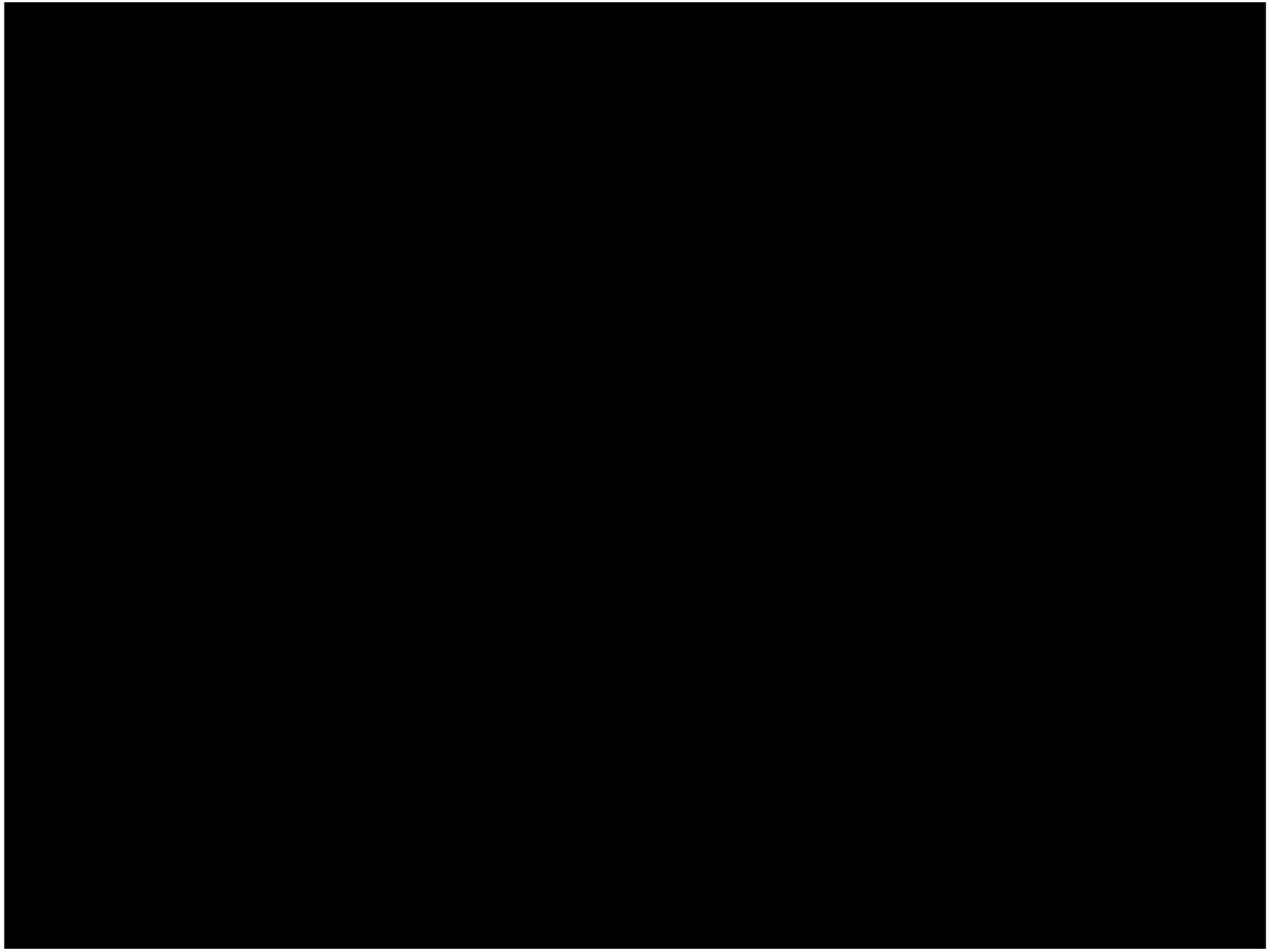
Learning: image synthesis

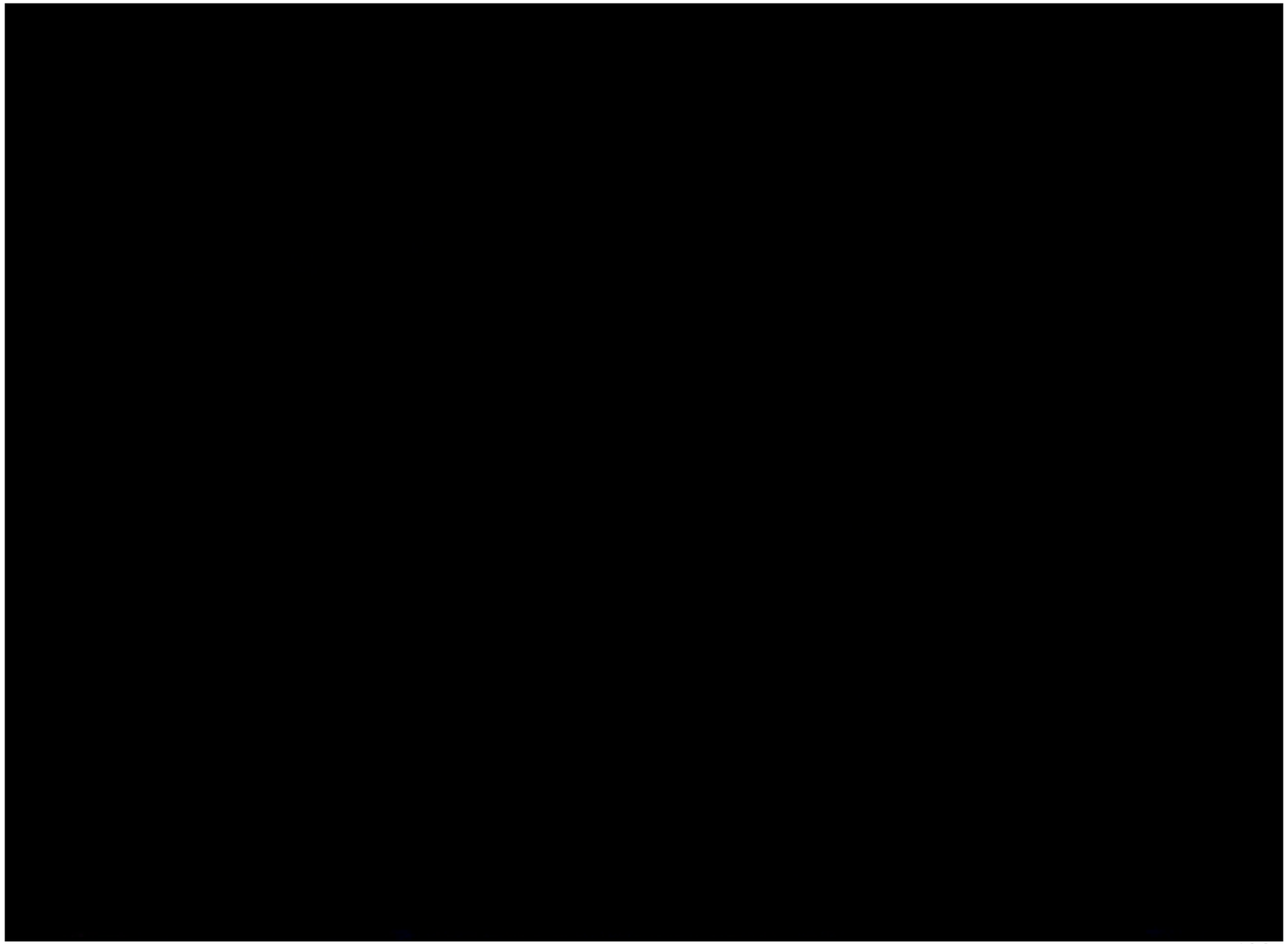
UNCONVENTIONAL GRAPHICS

Θ = 0° view $\Rightarrow$

Θ = 45° view $\Rightarrow$



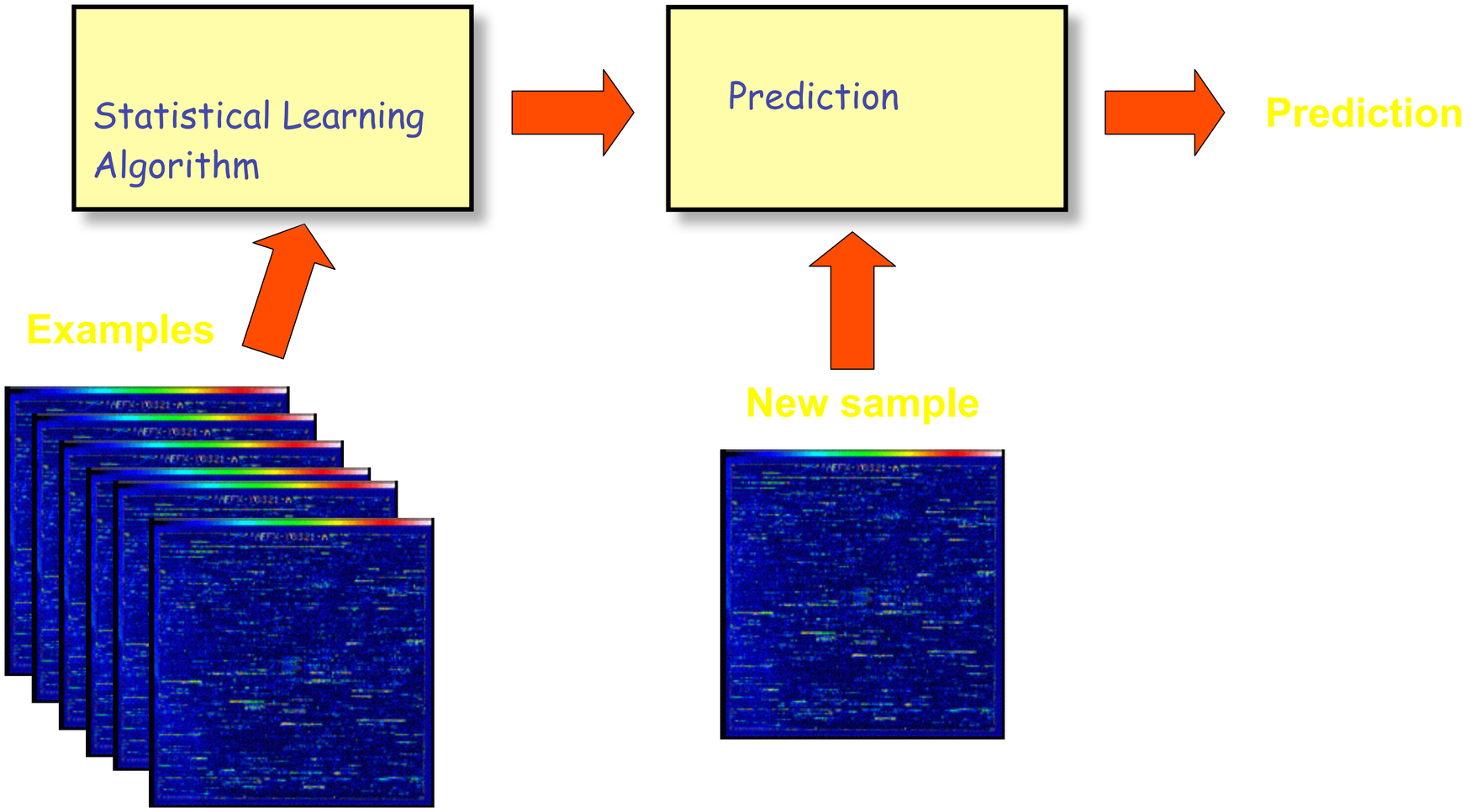




CENTER FOR
Brains
Minds +
Machines

Bioinformatics application: predicting type of cancer from DNA chips signals

Learning from examples paradigm



Learning: bioinformatics

New feature selection SVM:

Only 38 training examples, 7100 features

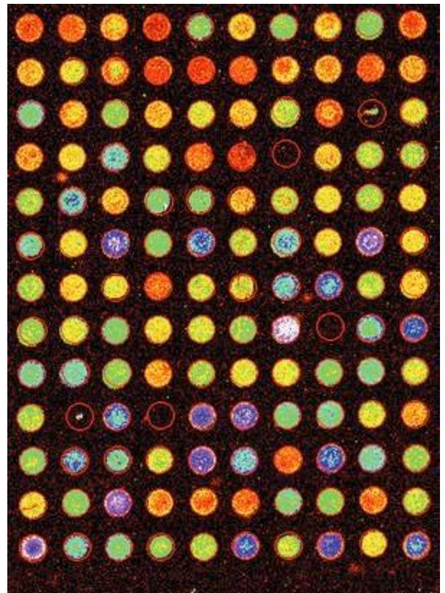
AML vs ALL: 40 genes 34/34 correct, 0 rejects.
5 genes 31/31 correct, 3 rejects of which 1 is an error.

A.I. Memo No.1677
C.B.C.L Paper No.182

Support Vector Machine Classification of Microarray Data

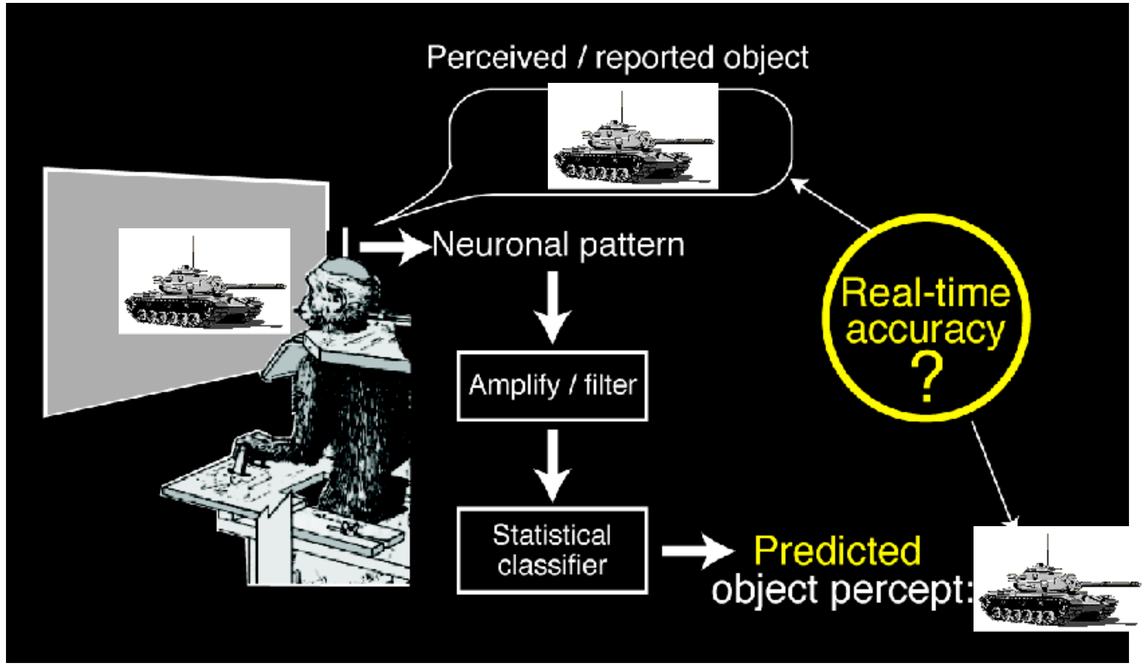
S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J.P. Mesirov, and T. Poggio

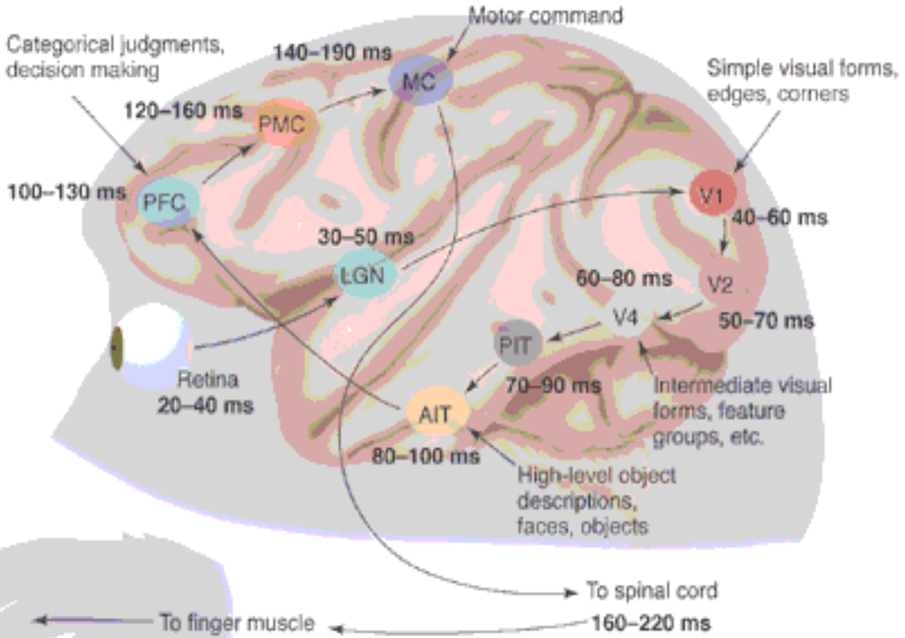
Pomeroy, S.L., P. Tamayo, M. Gaasenbeek, L.M. Sturia, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, M.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander and T.R. Golub. Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression, *Nature*, 2002.



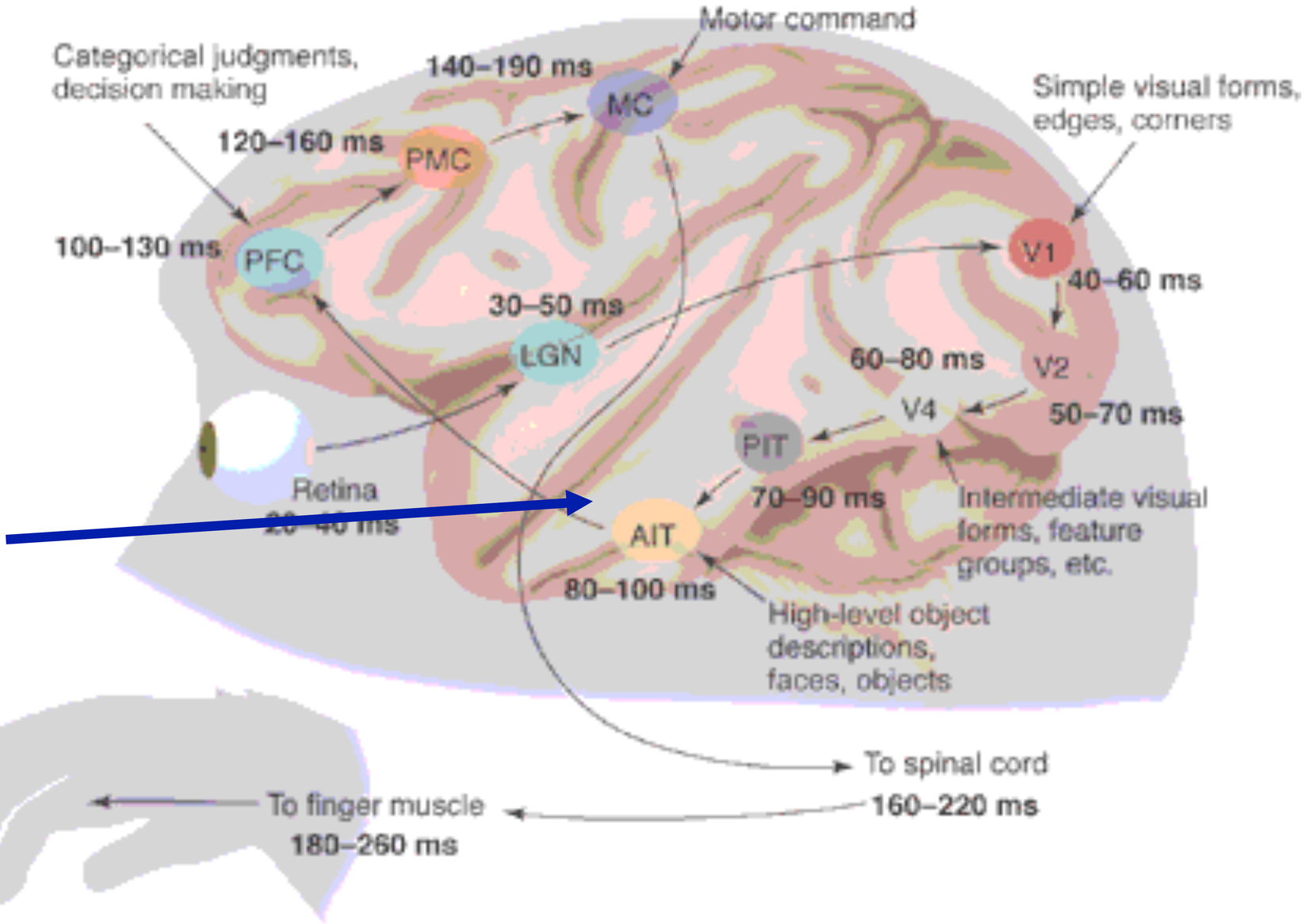
CENTER FOR
Brains
Minds +
Machines

Decoding the neural code: Matrix-like read-out from the brain



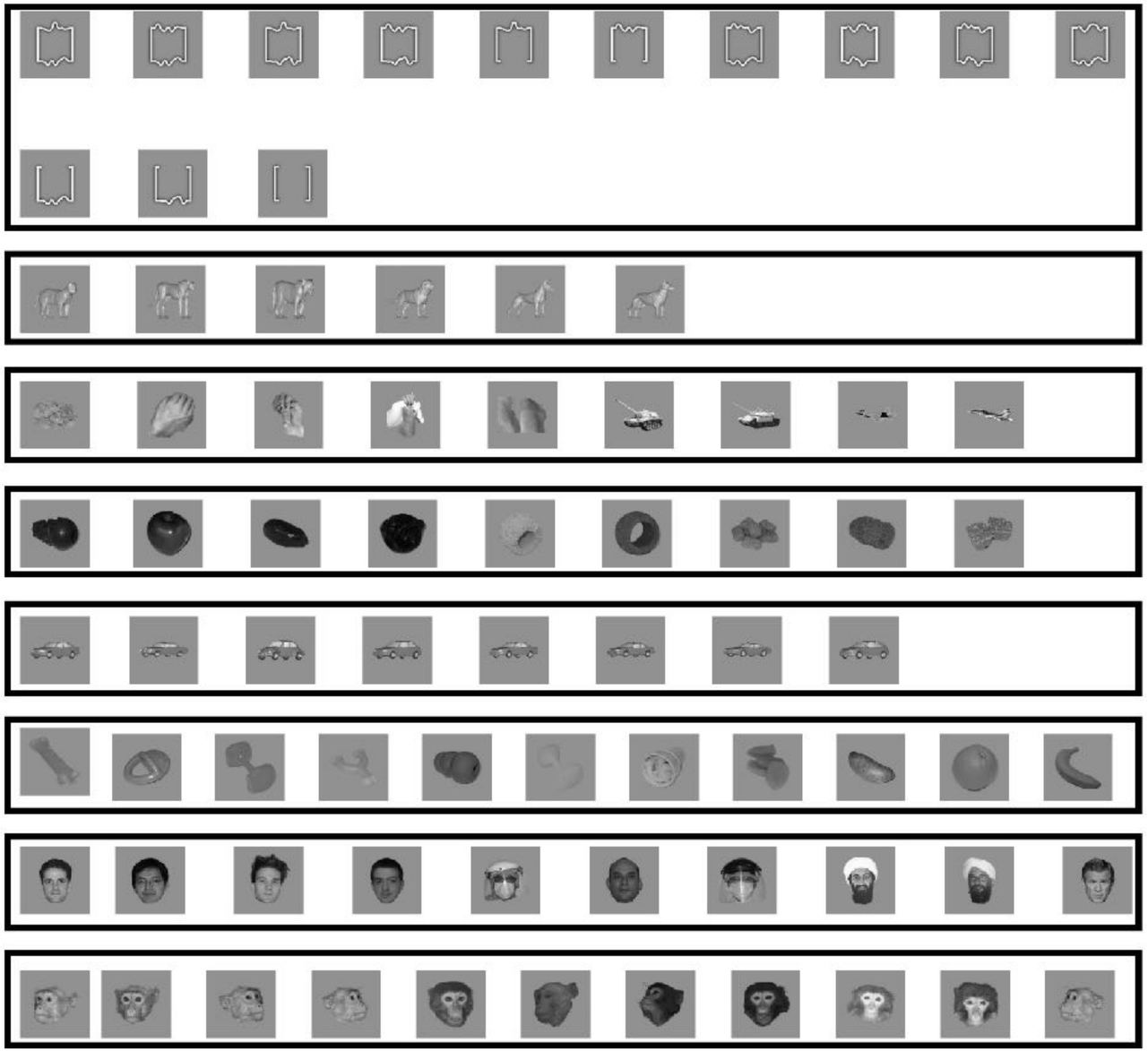


The end station of the ventral stream in visual cortex is IT



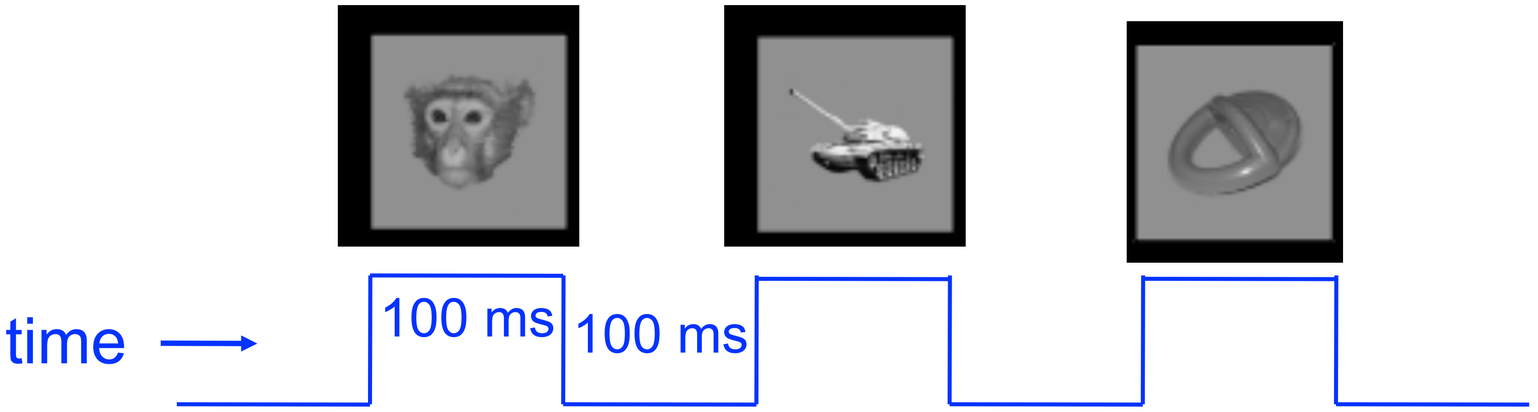
Reading-out the neural code in AIT

77 objects,
8 classes



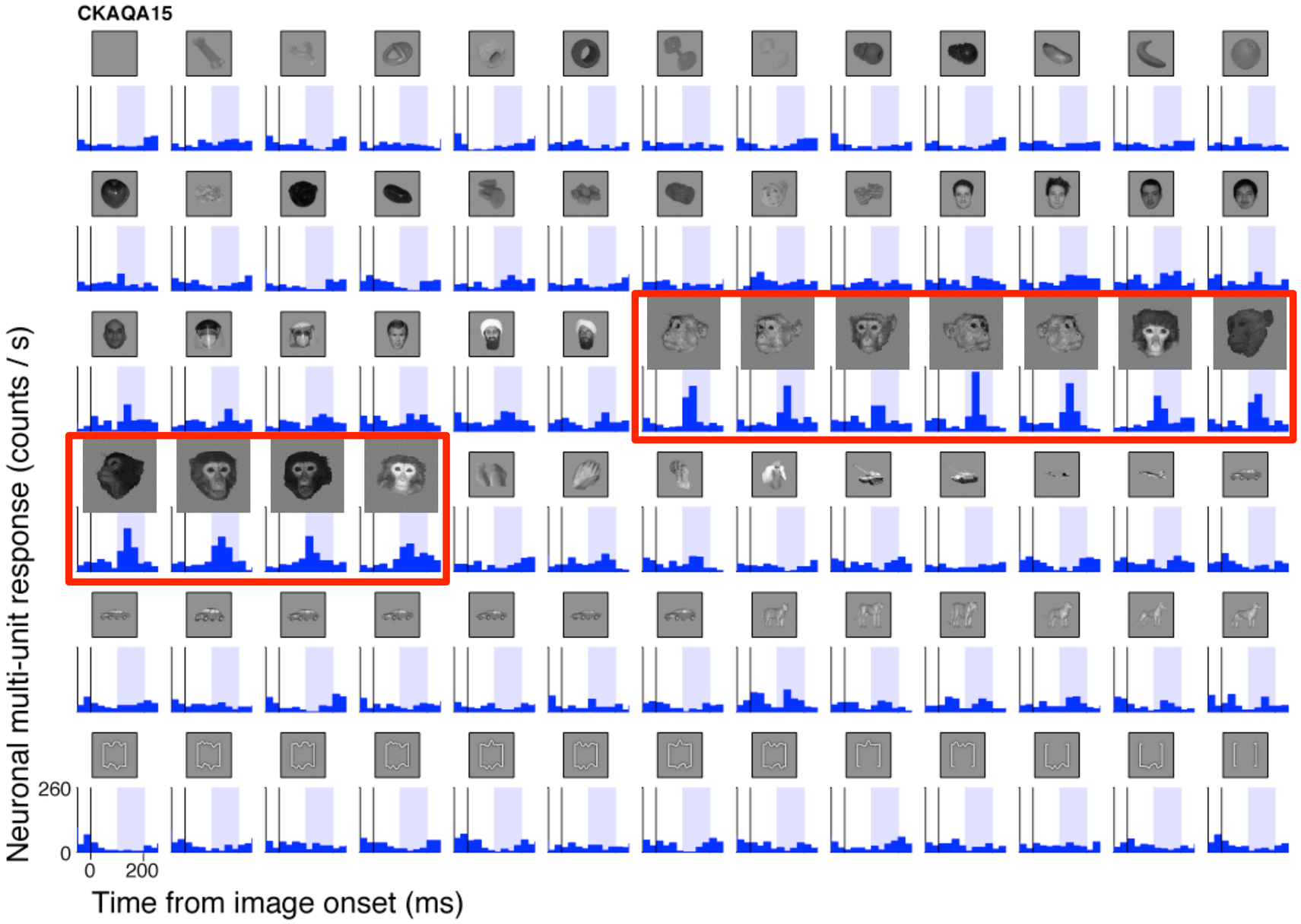
Chou Hung, Gabriel Kreiman, James DiCarlo, Tomaso Poggio, Science, Nov 4, 2005

Recording at each recording site during passive viewing

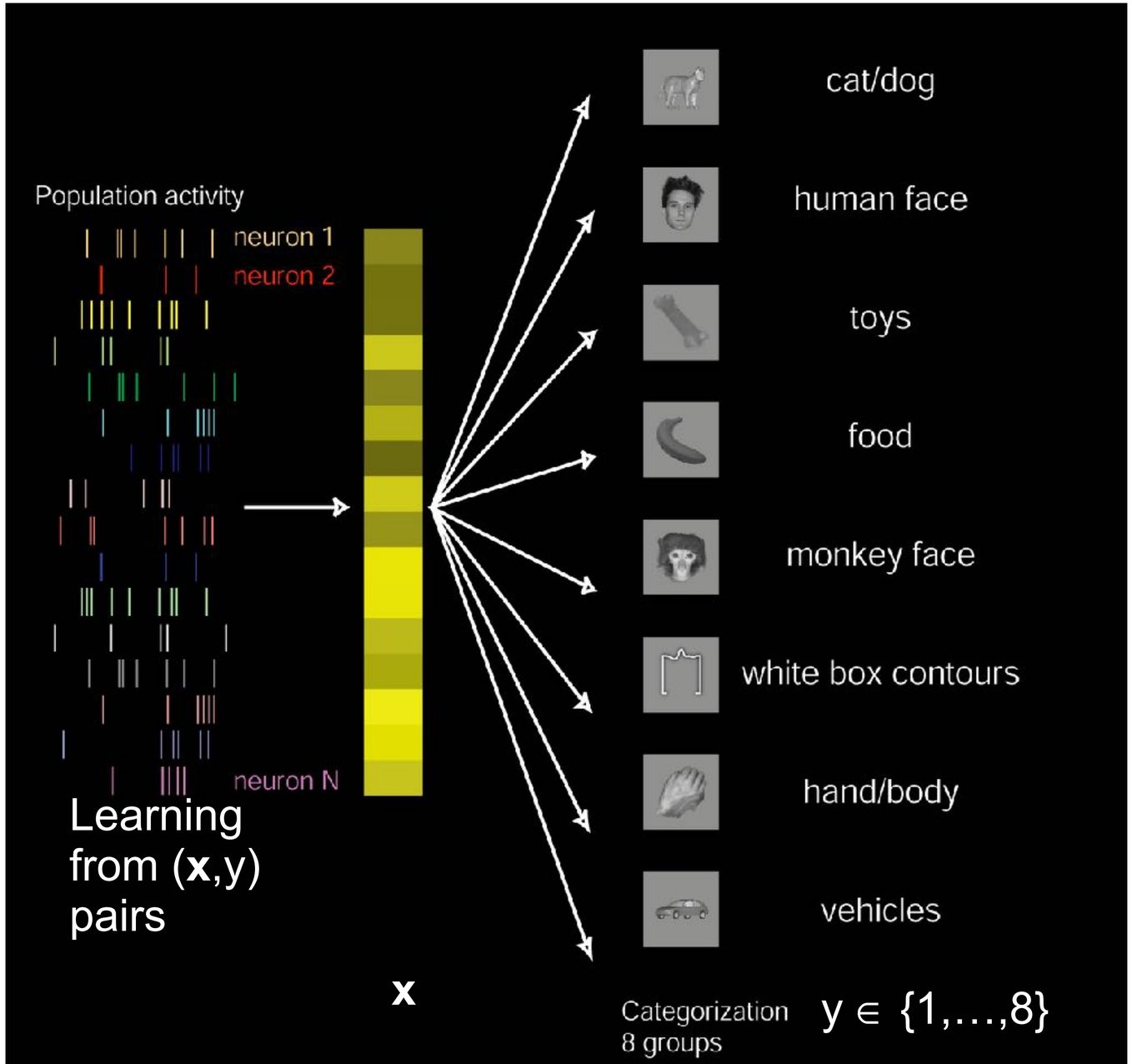


- 77 visual objects
- 10 presentation repetitions per object
- presentation order randomized and counter-balanced

Example of one AIT cell



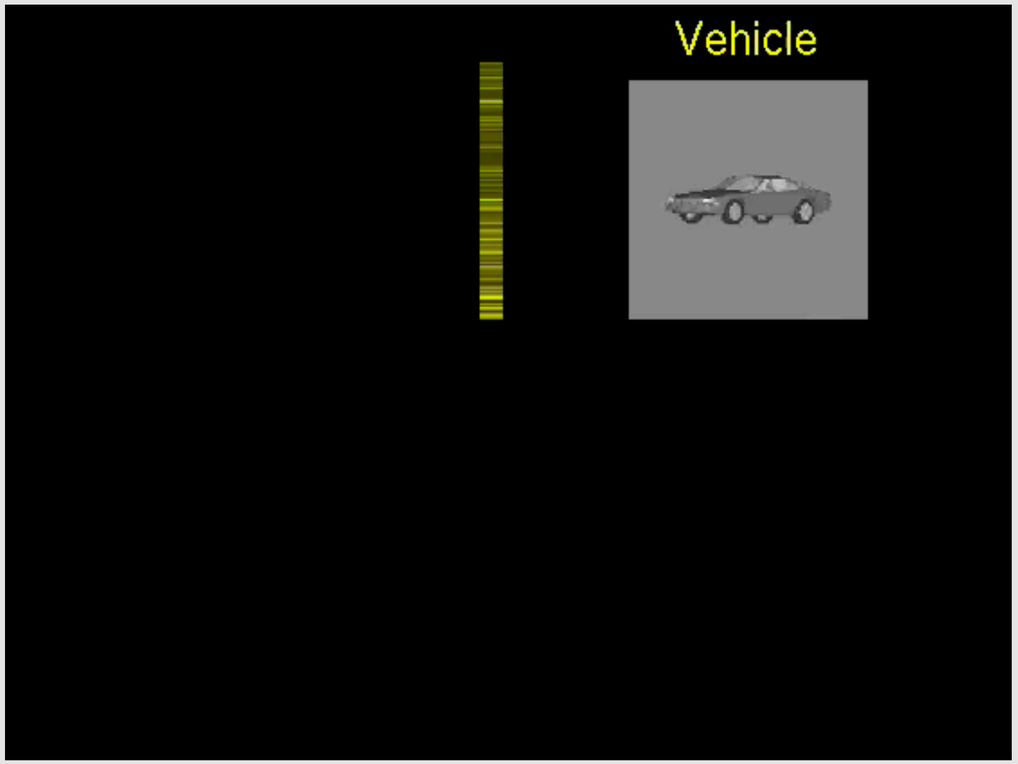
Decoding the neural code … using a classifier



We can decode the brain's code and read-out from neuronal populations:
reliable object categorization (>90% correct) using ~200 arbitrary AIT "neurons"

Video speed: 1 frame/sec

Actual presentation rate: 5 objects/sec



Categorization

Toy

Body

Human Face

Monkey Face

Vehicle

Food

Box

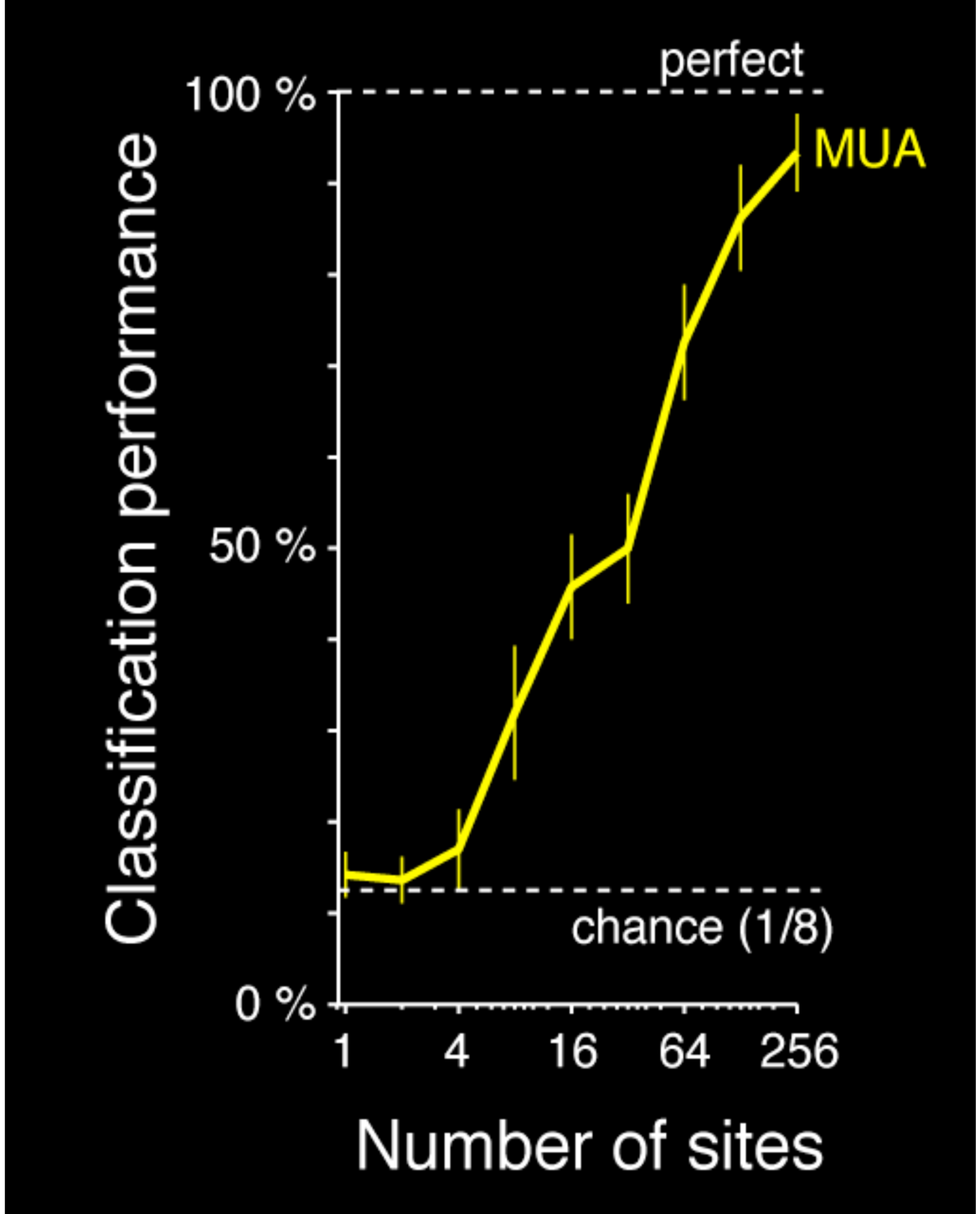
Cat/Dog

Hung, Kreiman, Poggio, DiCarlo. *Science 2005*

We can decode the brain's code and read-out from neuronal populations:

reliable object categorization using ~100 arbitrary AIT sites

- *[100-300 ms] interval*
- *50 ms bin size*

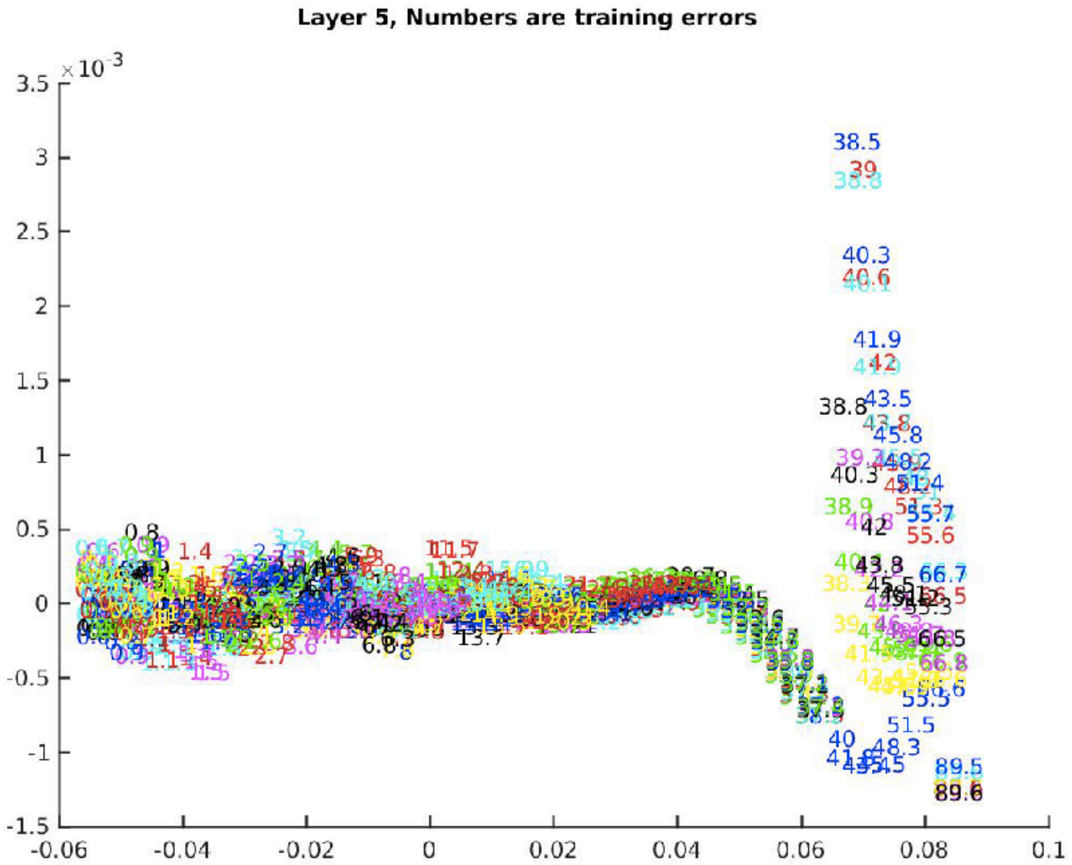


CENTER FOR
Brains
Minds +
Machines

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

Theorem (informal statement)

Replacing the RELUs with univariate polynomial approximation, Bezout theorem implies that the system of polynomial equations corresponding to zero empirical error has a very large number of degenerate solutions. The global zero-minimizers correspond to flat minima in many dimensions (generically unlike local minima). Thus SGD is biased towards finding global minima of the empirical risk.



CENTER FOR Brains Minds + Machines

Liao, Poggio, 2017

Bezout theorem

$$p(x_i) - y_i = 0 \quad \text{for } i = 1,\ldots,n$$

The set of polynomial equations above with *k= degree of p(x)* has a number of distinct zeros (counting points at infinity, using projective space, assigning an appropriate multiplicity to each intersection point, and excluding degenerate cases) equal to

$$Z = k^n$$

the product of the degrees of each of the equations. As in the linear case, when the system of equations is underdetermined – as many equations as data points but more unknowns (the weights) – the theorem says that there are an infinite number of global minima, under the form of Z regions of zero empirical error.

Global and local zeros

$f(x_i) - y_i = 0 \quad \text{for } i = 1,\ldots,n$ n equations in W unknowns with $W >> n$

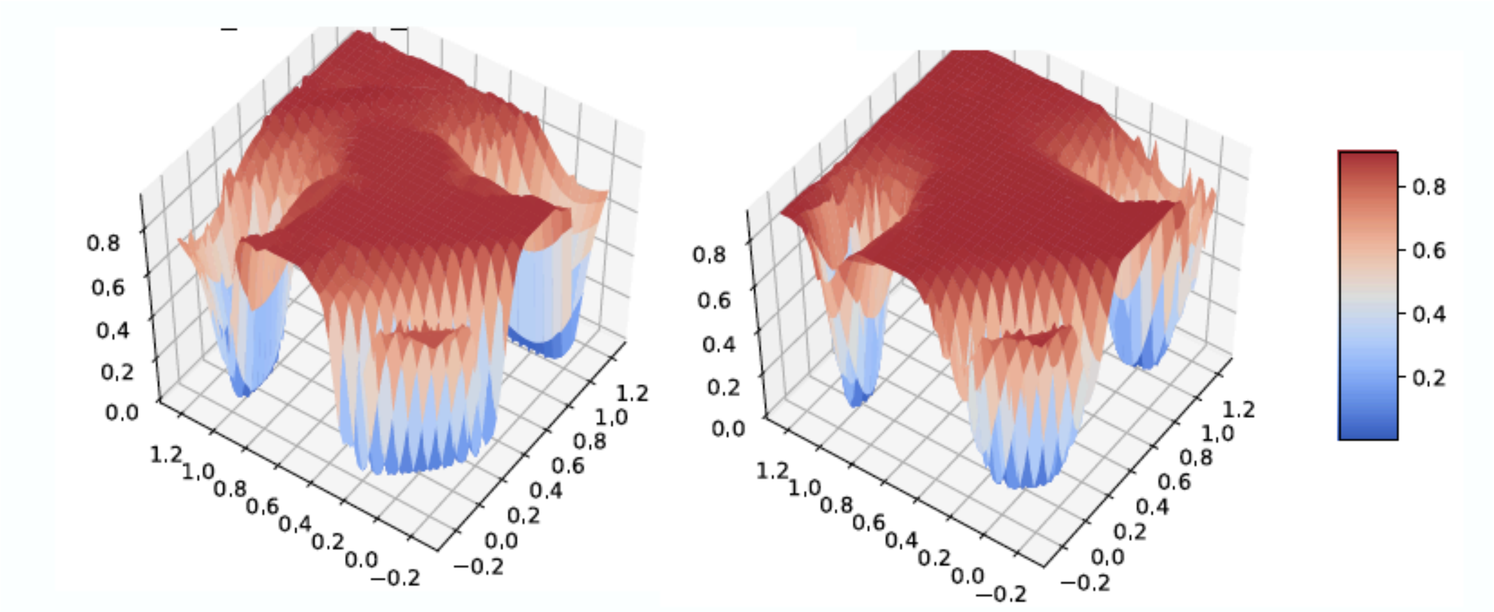
$$\nabla_w \sum_{i=1}^{N} (f(x_i) - y_i)^2) = 0$$ W equations in W unknowns

There are a very large number of zero-error minima which are highly degenerate unlike the local non-zero minima.

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

Results

- SGD finds with very high probability large volume, flat zero-minimizers;
- Flat minimizers correspond to degenerate zero-minimizers and thus to global minimizers;
- SGD minimizers select minima that correspond to small norm solutions and “good” expected error;



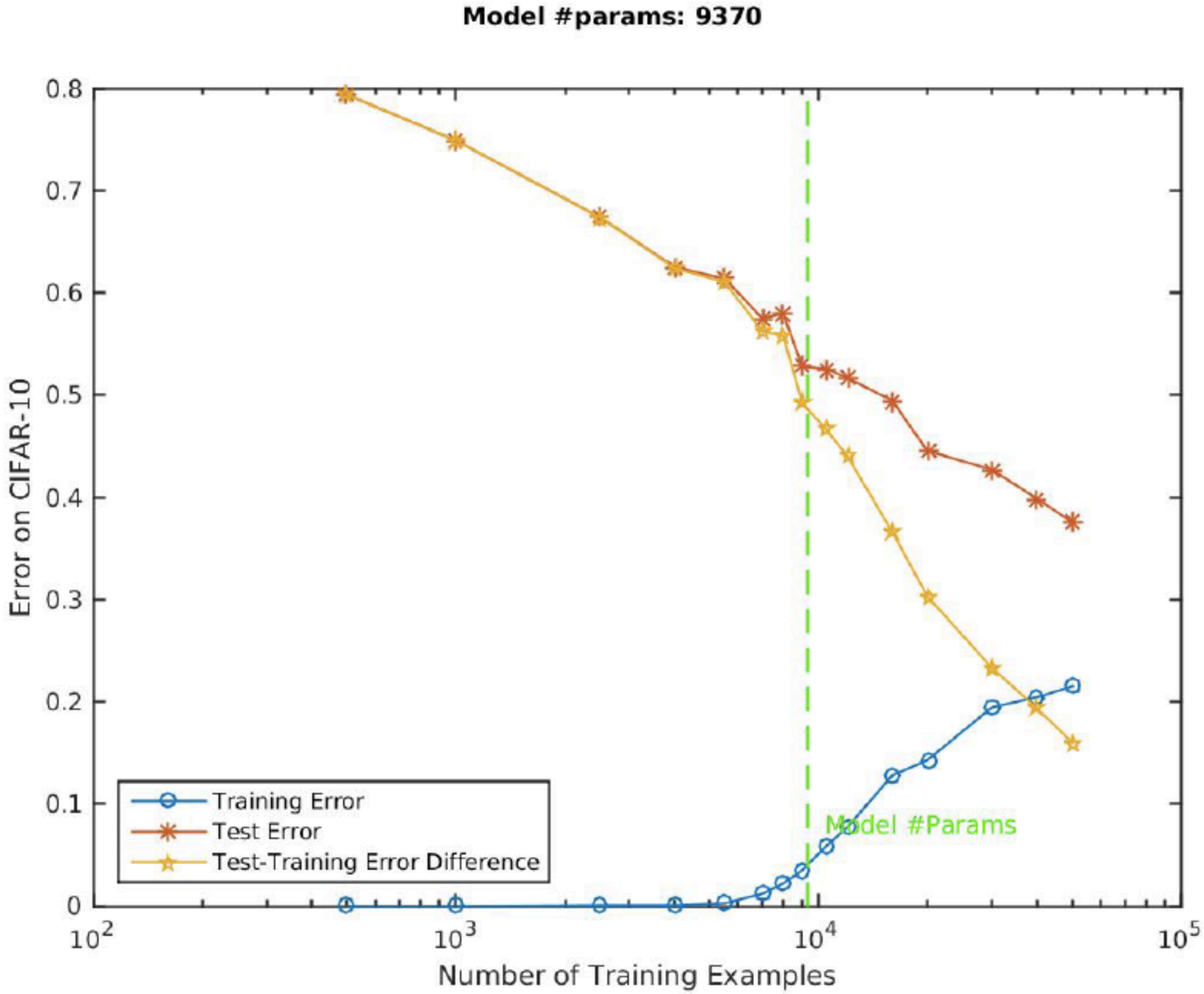
CIFAR-10: Natural Labels

Random Labels

Poggio, Rakhlin, Golovitc, Zhang, Liao, 2017

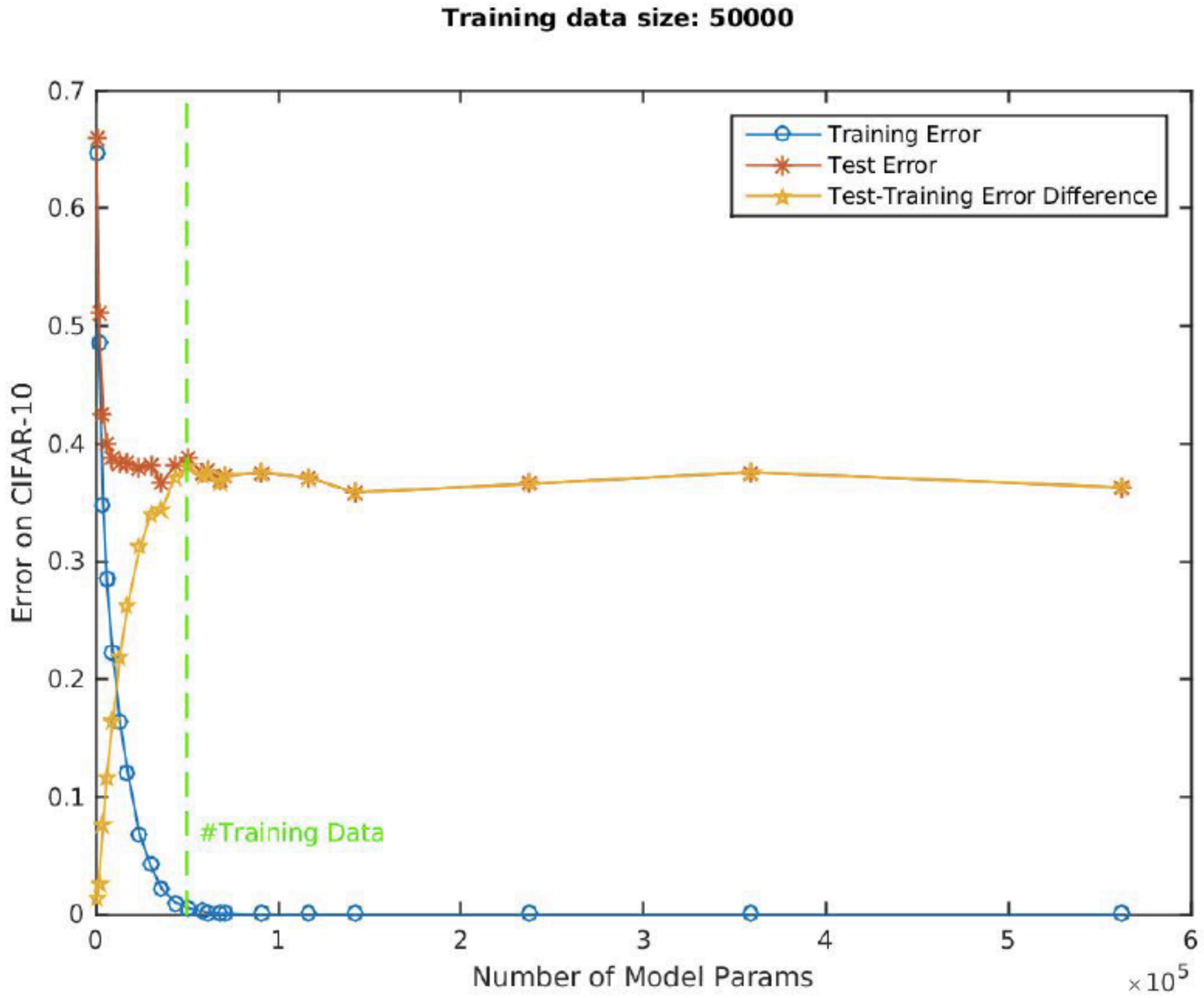
CENTER FOR
Brains
Minds
Machi

Good generalization with less data than # weights



CENTER FOR
Brains
Minds+
Machines

No overfitting



CENTER FOR
Brains
Minds+
Machines

Beyond today's DLNNs:
several scientific questions…

Why do Deep Learning Networks work? ===>

In which cases will they fail?

Is it possible to improve them?

Is it possible to reduce the number of labeled examples?

CENTER FOR
Brains
Minds+
Machines

Opportunity for a good project!

Beyond today's DLNNs: neurocognitive science

- State-of-the-art DLNNs require ~1M labeled examples
- This is not how we learn, how children learn

Today's science, tomorrow's engineering: learn like children learn

The first phase (and successes) of ML:
supervised learning, big data: $n \to \infty$

from programmers…
…to labelers…
…to computers that learn like children…

The next phase of ML: implicitly supervised learning,
learning like children do, small data: $n \to 1$