Statistical Learning Theory and Applications

9.520/6.860 in Fall 2018

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/

Contact: 9.520@mit.edu

Tomaso Poggio (TP), Lorenzo Rosasco (LR), Sasha Rakhlin (SR)

TAs:
Andrzej Banburski, David Zhou, Nhat Le, Michael Lee

Today's overview

- Course description/logistic
- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM, the MIT Quest: ***Intelligence, the Grand Vision***
- A bit of history: Statistical Learning Theory, Neuroscience
- A bit of ML history: applications
- Deep Learning

9.520: Statistical Learning Theory and Applications

- Course focuses on algorithms and theory for supervised learning.
- Regularization techniques, Kernel machines, batch and online supervised learning, sparsity.
- Deep learning and theory of it, based on first part of the class

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

9.520/6.860: Statistical Learning Theory and Applications

Class: Mon., Wed. 1:00 - 2:30 pm, 46-3310 (PILM Seminar Room)**
Office Hours: Friday 1:00 pm - 2:00 pm, 46-5156 (Poggio lab lounge) and/or 46-5165 (MIBR Reading Room)

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

Contact: 9.520@mit.edu

Mailing list: 9.520students@mit.edu

- 9.520/6.860 will use Stellar
- Mailing list and web (announcements) for updates

** On **10/22 and 10/24class will be in Building 34 Room 101.**

Material

Slides— will be posted (for most lectures)

Videos— check CBMM

Notes—

L. Rosasco and T. Poggio, **Machine Learning: a Regularization Approach, MIT-9.520 Lectures Notes**, Manuscript, Dec. 2016 **(will be provided)**
For feedback on book (typos, errors, ...) https://goo.gl/forms/pQcewnsAV3lCNoyr1

Syllabus at a glance

Class	Date	Title	Instructor(s)
Class 01	Wed Sep 05	The Course at a Glance	TP
Class 02	Mon Sep 10	Statistical Learning Setting	LR
Class 03	Wed Sep 12	Regularized Least Squaresces	LR
Class 04	Mon Sep 17	Feature Maps and Kernels	LR
Class 05	Wed Sep 19	Logistic Regression and Support Vector Machines	LR
Class 06	Mon Sep 24	Learning with Stochastic Gradients	AR
Class 07	Wed Sep 26	Implicit Regularization	LR
Class 08	Mon Oct 01	Large Scale Learning by Sketching	LR
Class 09	Wed Oct 03	Sparsity Based Regularization	LR
Mon Oct 08 - Columbus Day			
Class 10	Wed Oct 10	Neural networks: Introduction, backpropagation	LR or AB
Class 11	Mon Oct 15	Neural Networks: tips, tricks and SW	QL AB
Class 12	Wed Oct 17	Generative Adversarial Networks	TBA
Class 13	Mon Oct 22	Statistical Learning (from SGD/GD to Stat objective)	AR
Class 14	Wed Oct 24	Uniform Convergence, ERM	AR
Class 15	Mon Oct 29	Sample Complexity via Rademacher Averages I	AR
Class 16	Wed Oct 31	Sample Complexity via Rademacher Averages II	AR
Class 17	Mon Nov 05	Margin Analysis for Classification	AR
Class 18	Wed Nov 07	Local Fitting: Interpolation, Generalization, Bias-Variance	AR
Mon Nov 12 - Veterans Day			
Class 19	Wed Nov 14	Algorithmic Stability and Generalization	AR
Class 20	Mon Nov 19	Privacy and Information-Theoretic Stability	AR
Class 21	Wed Nov 21	Deep Learning Theory: Approximation	TP
Class 22	Mon Nov 26	Sample complexity of Neural Networks I	AR
Class 23	Wed Nov 28	Sample complexity of Neural Networks II	AR
Class 24	Mon Dec 03	Deep Learning Theory: Optimization	TP
Class 25	Wed Dec 05	Deep Learning Theory: Generalization	TP
Class 26	Mon Dec 10	Machine Learning, the Brain and the Next Breakthrough in AI	TP
Wed Dec 12 - 2 poster sessions on Dec. 12			

Grading policies

- **Problem sets (0.6)**
  - 6 problem sets (0.10 each)
    - 2 - 3 questions (exercises and/or MATLAB)
    - 1 week due
  - Late policy on next slide
  - typeset in LaTeX (template will be provided)
  - Online submission by due date; printed submission in next class
- **Project (0.3)**
  - See later
- **Participation (0.1)**
  - *Attending class lectures is required!*
  - Sign-in sheet will be circulated 5 (random) times

Problem sets

- Problem sets (0.6)
  - 6 problem sets (0.10 each)
    - 2 - 3 questions (demonstrations/exercises + short MATLAB)
    - 7 days due!
  - typeset in LaTeX (template provided)
  - *online submission by due date; printed submission in next class*
- Late policy
  - All students have 4 free late days (to be used on psets and project proposal)
  - You may use up to 2 late days per assignment with no penalty
  - Beyond this, we will deduct a late penalty of 50% of the grade per additional late day

Dates (due times are 11:59 pm). Submission online (dbox link).

[pset 1] Wed. Sep. 19, due: Tue., Sep. 25
[pset 2] Wed. Oct. 3, due: Tue., Oct. 09
[pset 3] Wed. Oct. 17, due: Tue., Oct. 23
[pset 4] Wed. Oct. 31, due: Tue., Nov. 06
[pset 5] Wed. Nov. 19, due: Tue., Nov. 25
[pset 6] Wed. Dec. 3, due: Tue., Dec. 11

Collaboration policy: You may discuss with others but need to work out your own solution.

Projects

A) Theory
B) Algorithms
C) Application
   - This is not a data science course, so we will not consider data preparation as contributing to the grade.

D) Coding
E) Wikipedia

- report (NIPS format): 4 pages (+ Appendix), 6 pages max
  OR
- poster session (last week of classes)

Dates

- Abstract and title: Oct. 31
- Feedback and approval: Nov. 7
- Poster and revised abstract submission: Dec. 10
- Poster presentations: Dec. 12
- Report submission: Dec. 12

Today's overview

- Course description/logistic

- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM, the MIT Quest: ***Intelligence, the Grand Vision***

- A bit of history: Statistical Learning Theory, Neuroscience

- A bit of ML history: applications

- Deep Learning

Grand Vision of CBMM, Quest, this course

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

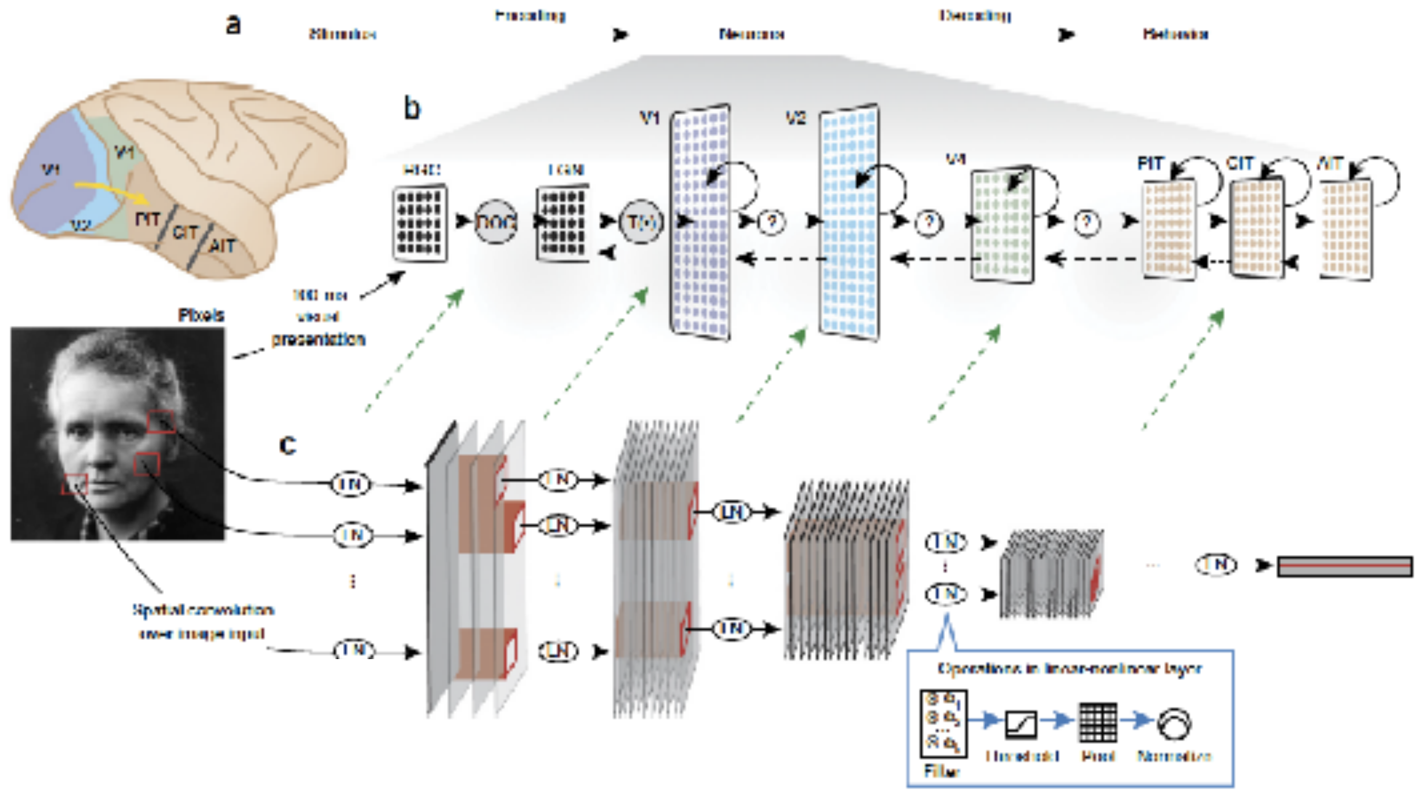
The problem of (human) intelligence is one of the great problems in science, probably the greatest.

Research on intelligence:

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

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.

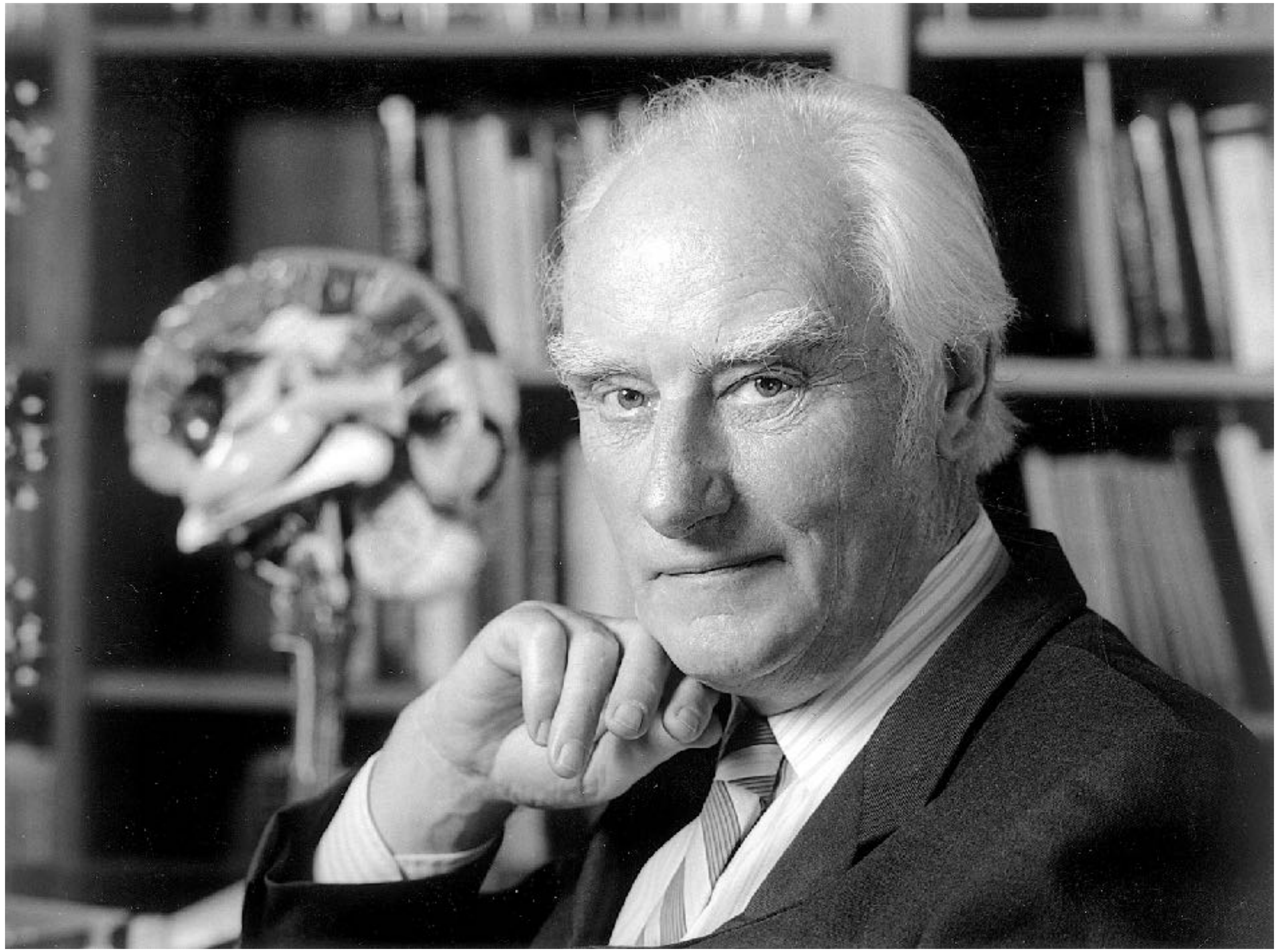


Key recent advances in the engineering of intelligence have their roots in basic research on the brain

CENTER FOR Brains Minds+ Machines

Why (Natural) Science and Engineering?

Just a definition: science is natural science (Francis Crick, 1916-2004)

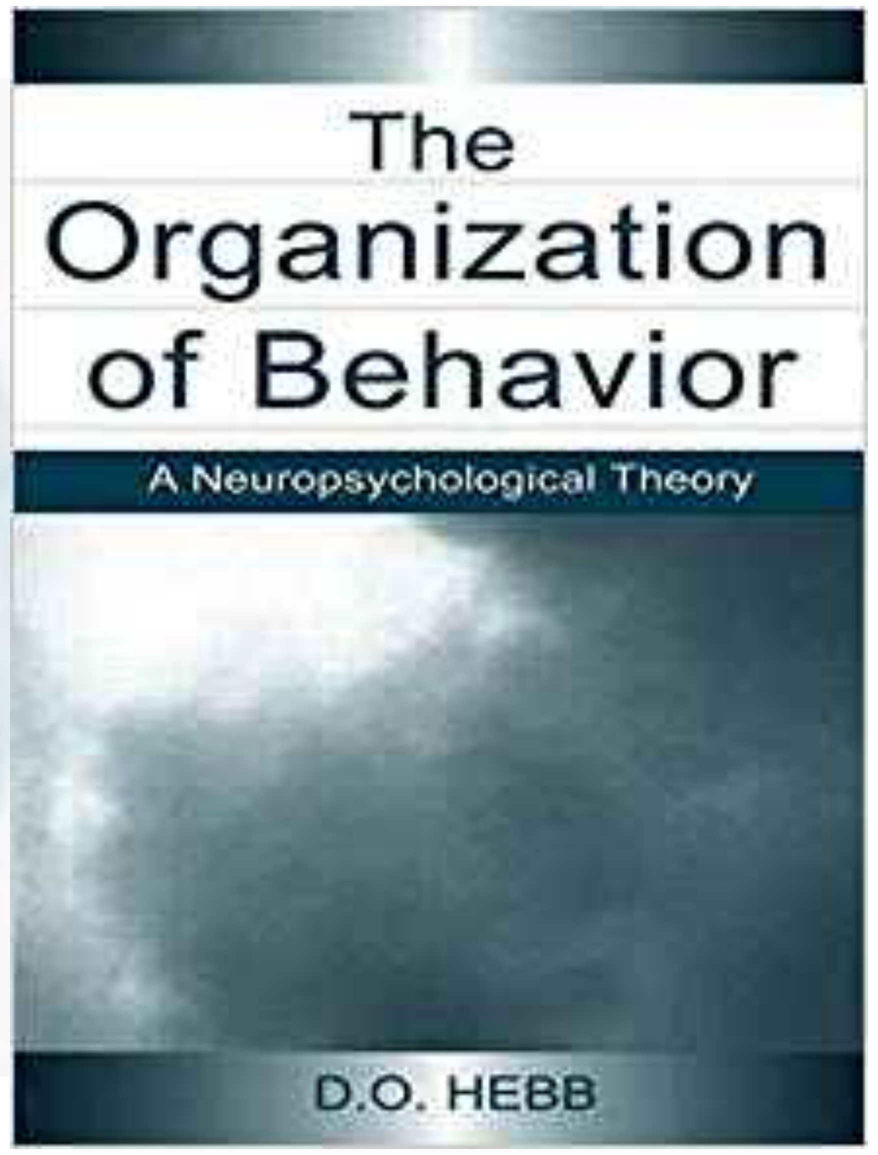


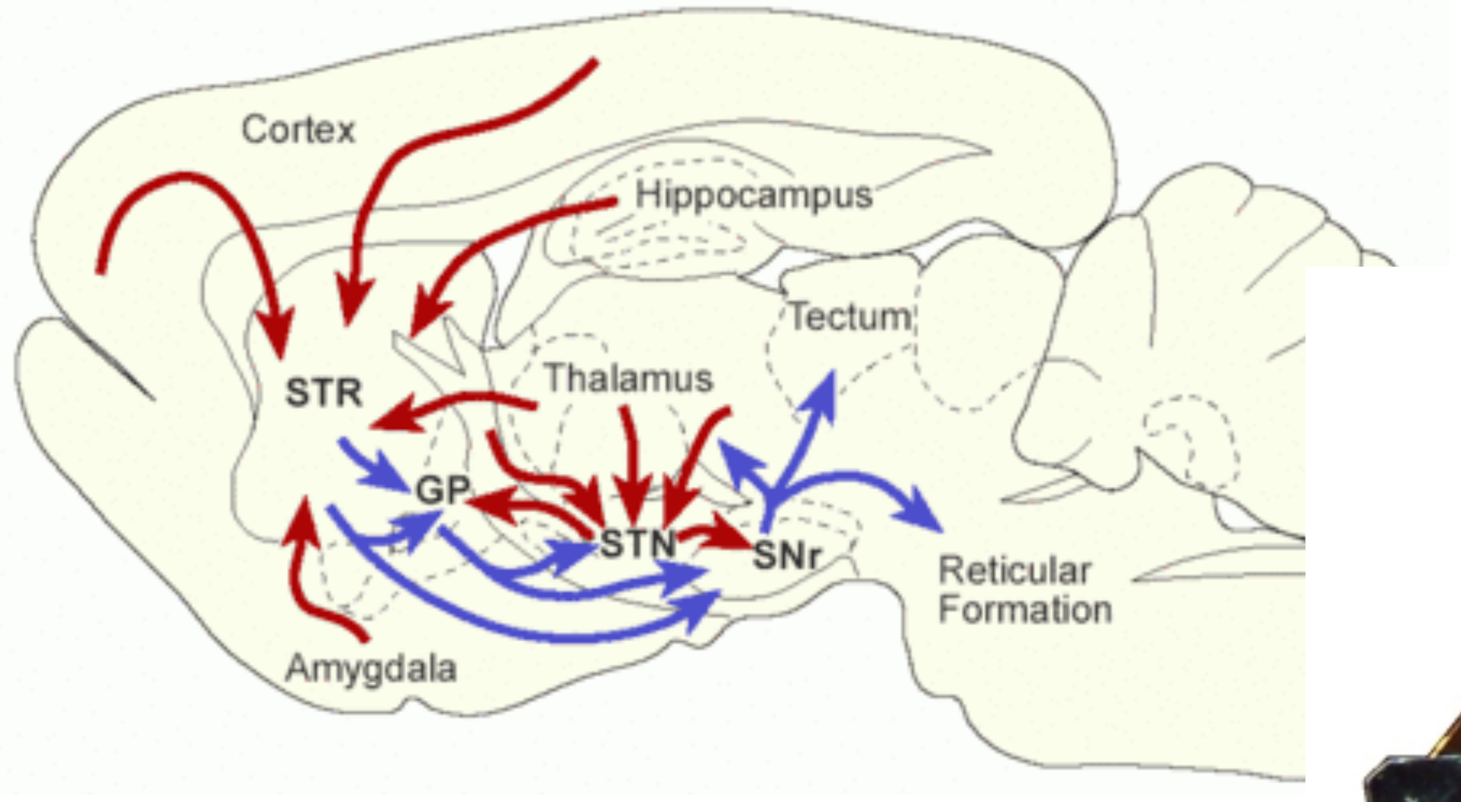
Two Main Recent Success Stories in AI

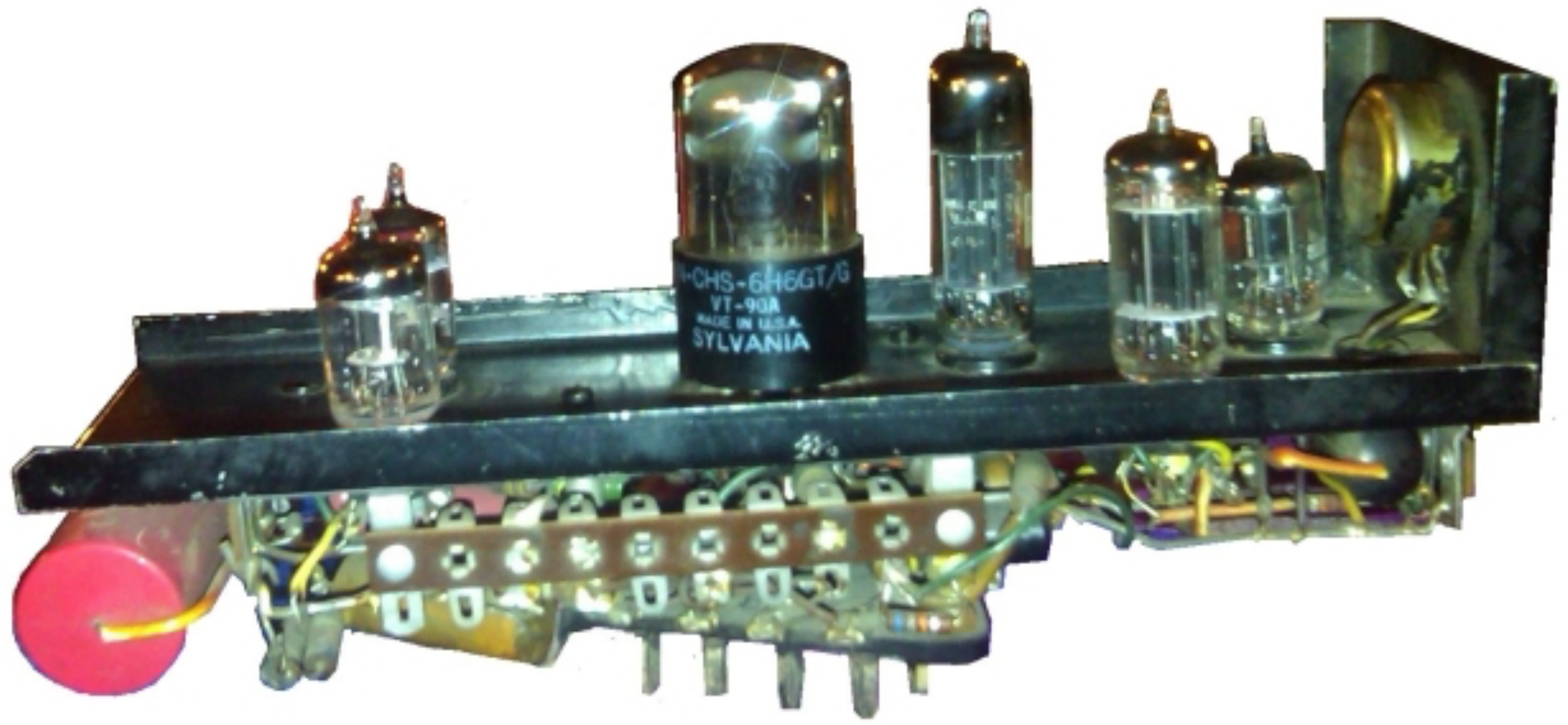
DL and RL come from neuroscience

R

L







Minsky's SNARC

D

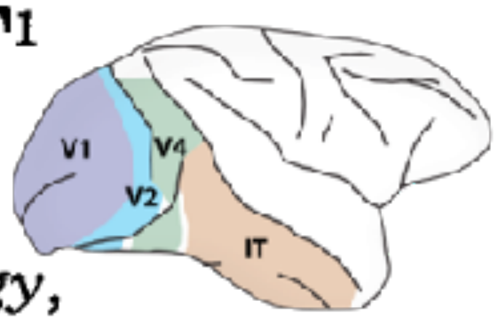
L

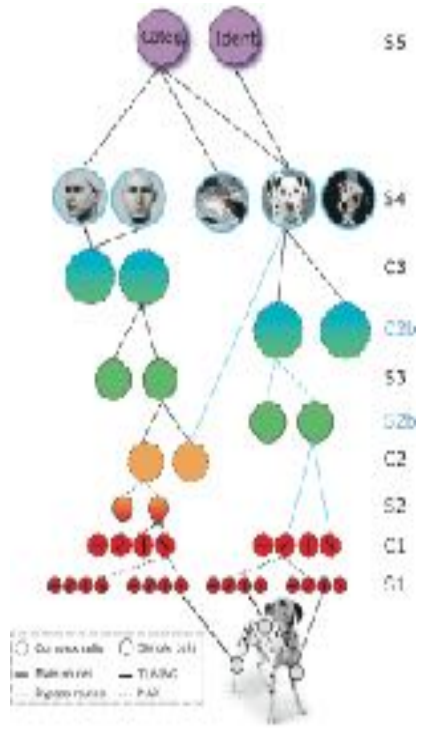
RECEPTIVE FIELDS AND FUNCTIONAL ARCHITECTURE IN TWO NONSTRIATE VISUAL AREAS (18 AND 19) OF THE CAT[1]

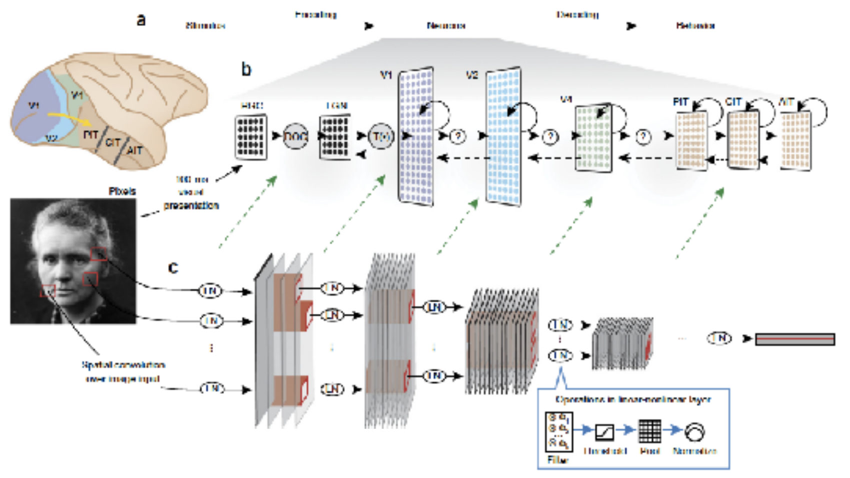
DAVID H. HUBEL AND TORSTEN N. WIESEL

Neurophysiology Laboratory, Department of Pharmacology, Harvard Medical School, Boston, Massachusetts

(Received for publication August 24, 1964)







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
combining them with learning theory

(suggestion: attend 6.861/9.523)

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

CENTER FOR Brains Minds+ Machines

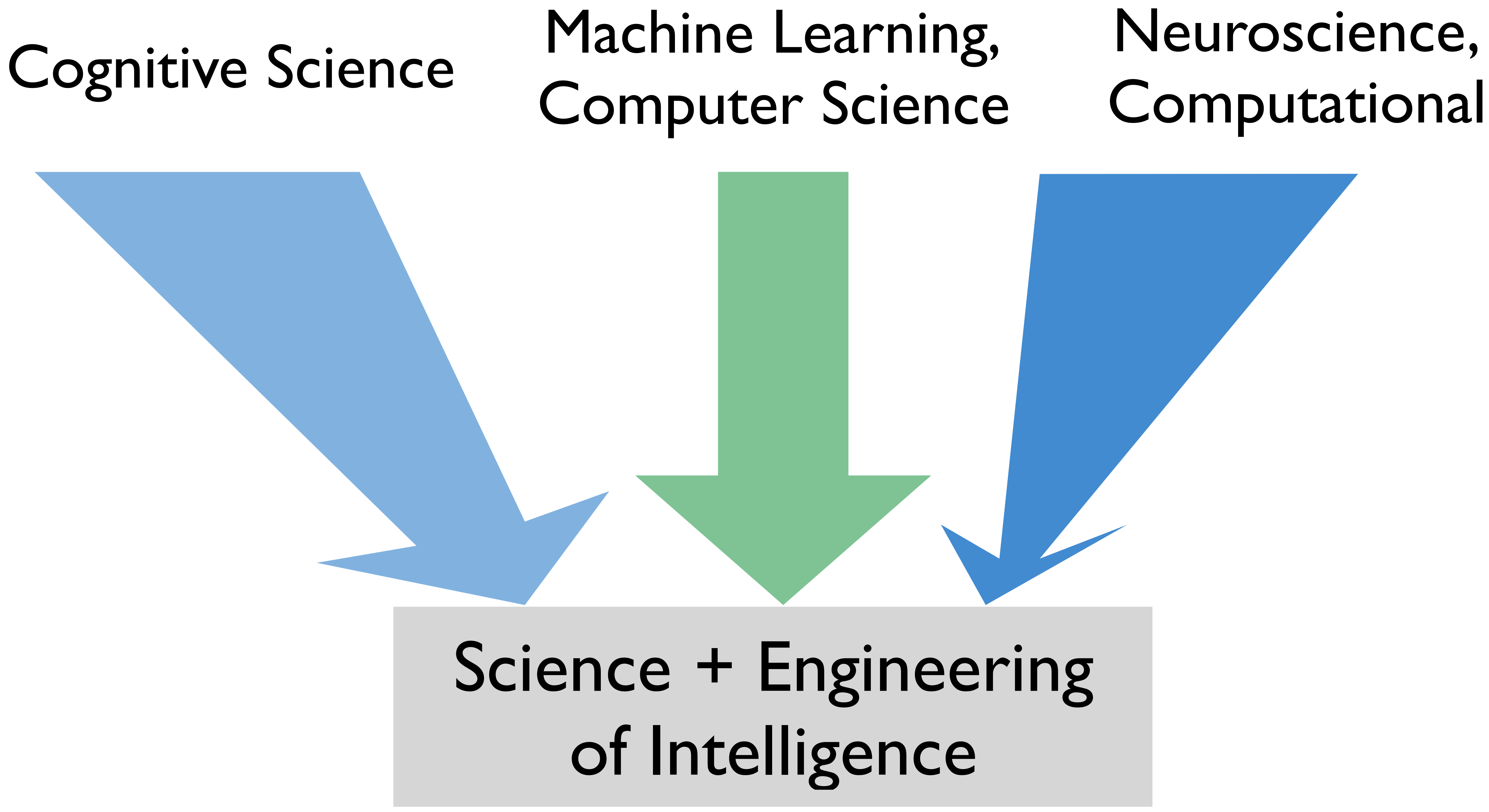
CBMM and the MIT Quest

CBMM Overview

CENTER FOR
Brains
Minds+
Machines

The Center for Brains, Minds and Machines (CBMM) is a multi-institutional NSF Science and Technology Center dedicated to the study of intelligence - how the brain produces intelligent behavior and how we may be able to replicate intelligence in machines. We believe in the synergy between the science and the engineering of intelligence.

Funding 2013-2023	**~$50M**
Research Institutions	**~4**
Educational Institutions	**12**
Faculty (CS+BCS+…)	**~23**
Researchers	**223**
Publications	**397**



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

Boston Children's Hospital
Kreiman

Florida International U.
Diaz, Finlayson

Harvard Medical School
Kreiman, Livingstone

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

Academic and Corporate Partners

IIT
Cingolani

A*star
Chuan Poh Lim

Hebrew U.
Weiss

MPI
Bülthoff

Genoa U.
Verri, Rosasco

Weizmann
Ullman

City U. HK
Smale

Google

IBM

Microsoft

Orcam

Siemens

Philips

GE

Neuroelectrics

DeepMind

Honda

Boston Dynamics

NVIDIA

Rethink Robotics

Schlumberger

Mobileye

Intel

CENTER FOR
Brains
Minds+
Machines

Summer Course at Woods Hole: Our flagship initiative

Brains, Minds & Machines Summer Course

Gabriel Kreiman + Boris Katz

A community of scholars is being formed:

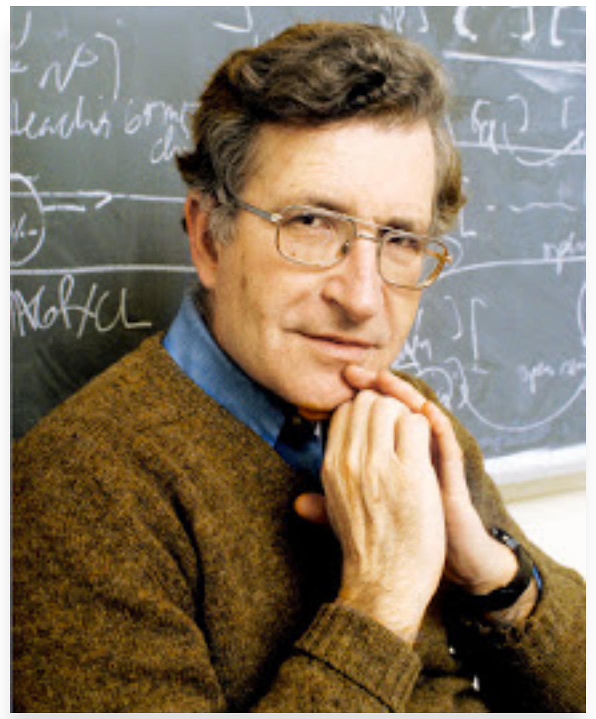
Intelligence, the MIT Quest

The MIT Intelligence Quest will advance the science and engineering of both human and machine intelligence. Launched on February 1, 2018, this effort seeks to discover the foundations of human intelligence and drive the development of technological tools that can positively influence virtually every aspect of society.

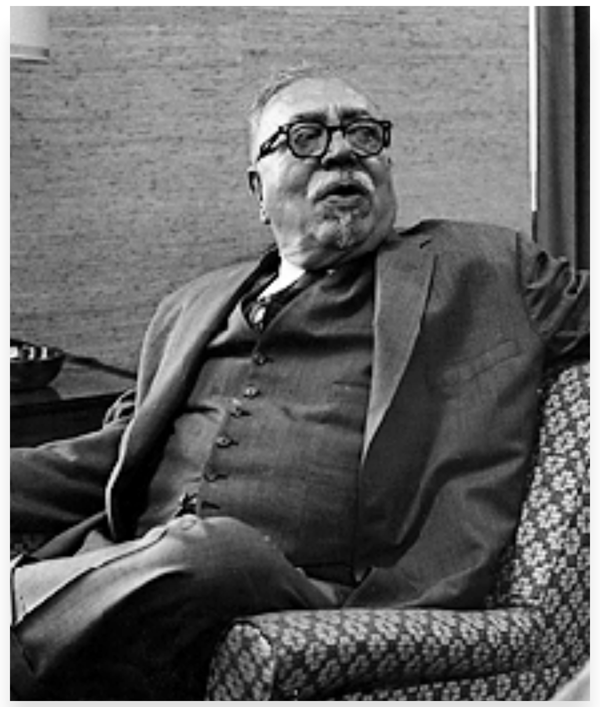
The Institute's culture of collaboration will encourage life scientists, computer scientists, social scientists, and engineers to join forces to investigate the societal implications of their work as they pursue hard problems lying beyond the current horizon of intelligence research. By uniting diverse fields and capitalizing on what they can teach each other, we seek to answer the deepest questions about intelligence.

Historical timeline…

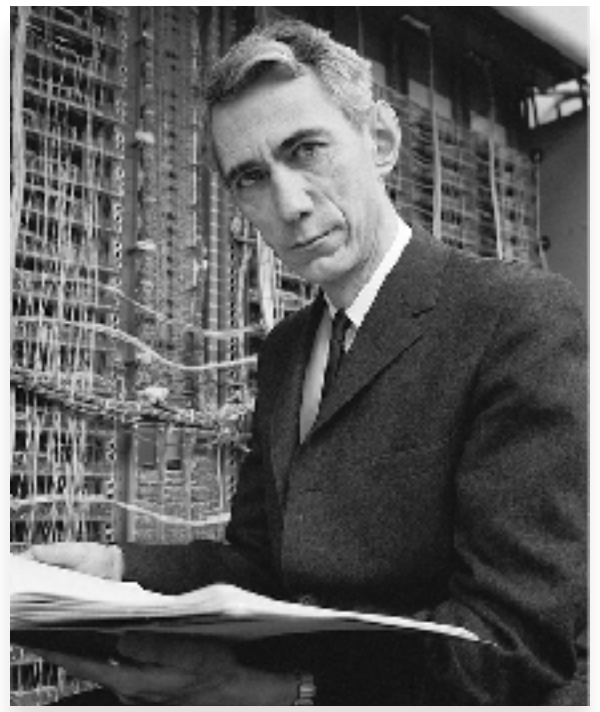
Marvin Minsky



Noam Chomsky



Norbert Wiener



Claude Shannon

"The Golden Age" 1950 - 1970

MIT Intelligence Initiative

CENTER FOR
Brains
Minds+
Machines

Intelligence:
The MIT Quest

2008

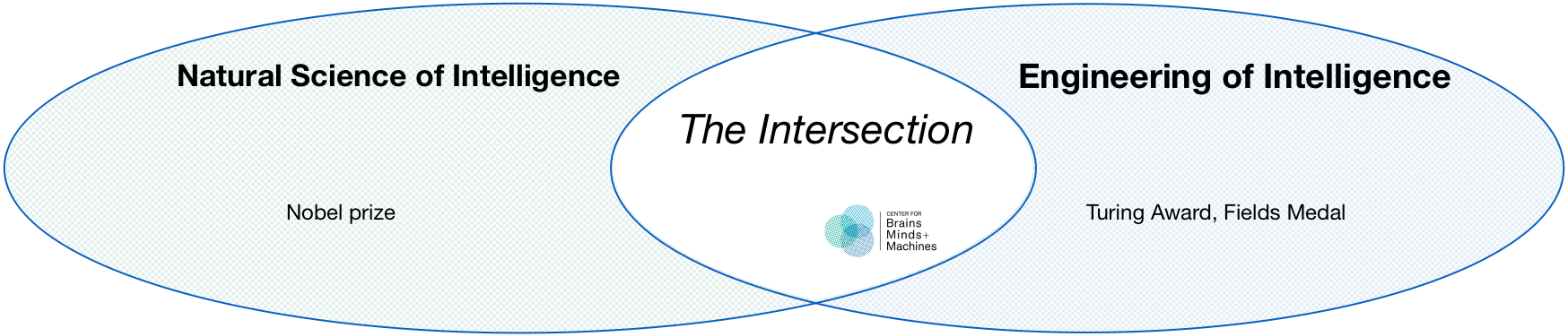
2012 - 2013

2018

Intelligence: The MIT Quest

BRIDGE

CORE: Cutting-Edge Research on the Science + Engineering of Intelligence



Summary

- 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 <u>real</u> AI, in addition to understanding our brain. BTW, even without this extension, the next few years will be a golden age for ML applications.

Today's overview

- Course description/logistic

- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM, the MIT Quest: ***Intelligence, the Grand Vision***

- A bit of history: Statistical Learning Theory and Applications

- Deep Learning

Why theory of Learning

- Learning is now the lingua franca of Computer Science
- Learning is at the center of recent successes in AI over the last 15 years
- Now and the next 10 year will be a golden age for technology based on learning: Google, Siri, Mobileye, Deep Mind etc.
- The next 50 years will be a golden age for the science and engineering of intelligence. Theories of learning and their tools will be a key part of this.

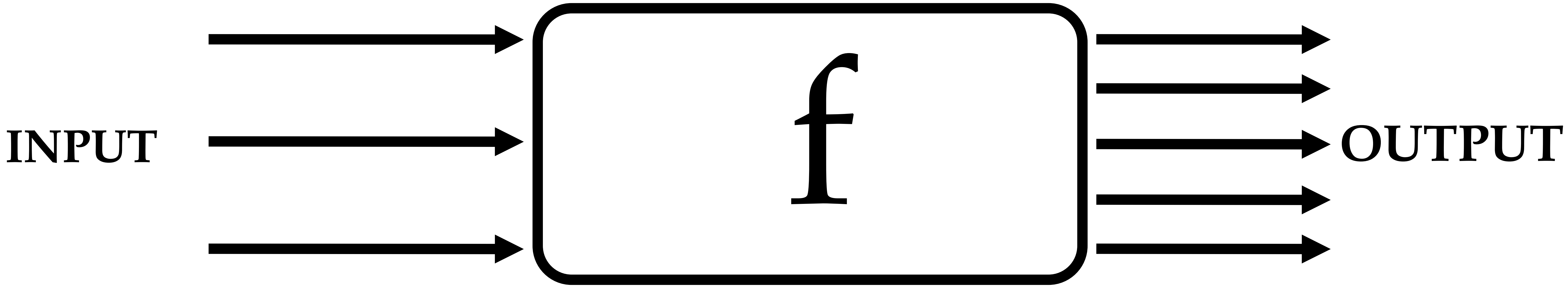
2015

CENTER FOR
Brains
Minds+
Machines

~1995

Statistical Learning Theory

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



Given a set of l examples (data)

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

Question: find function f such that

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

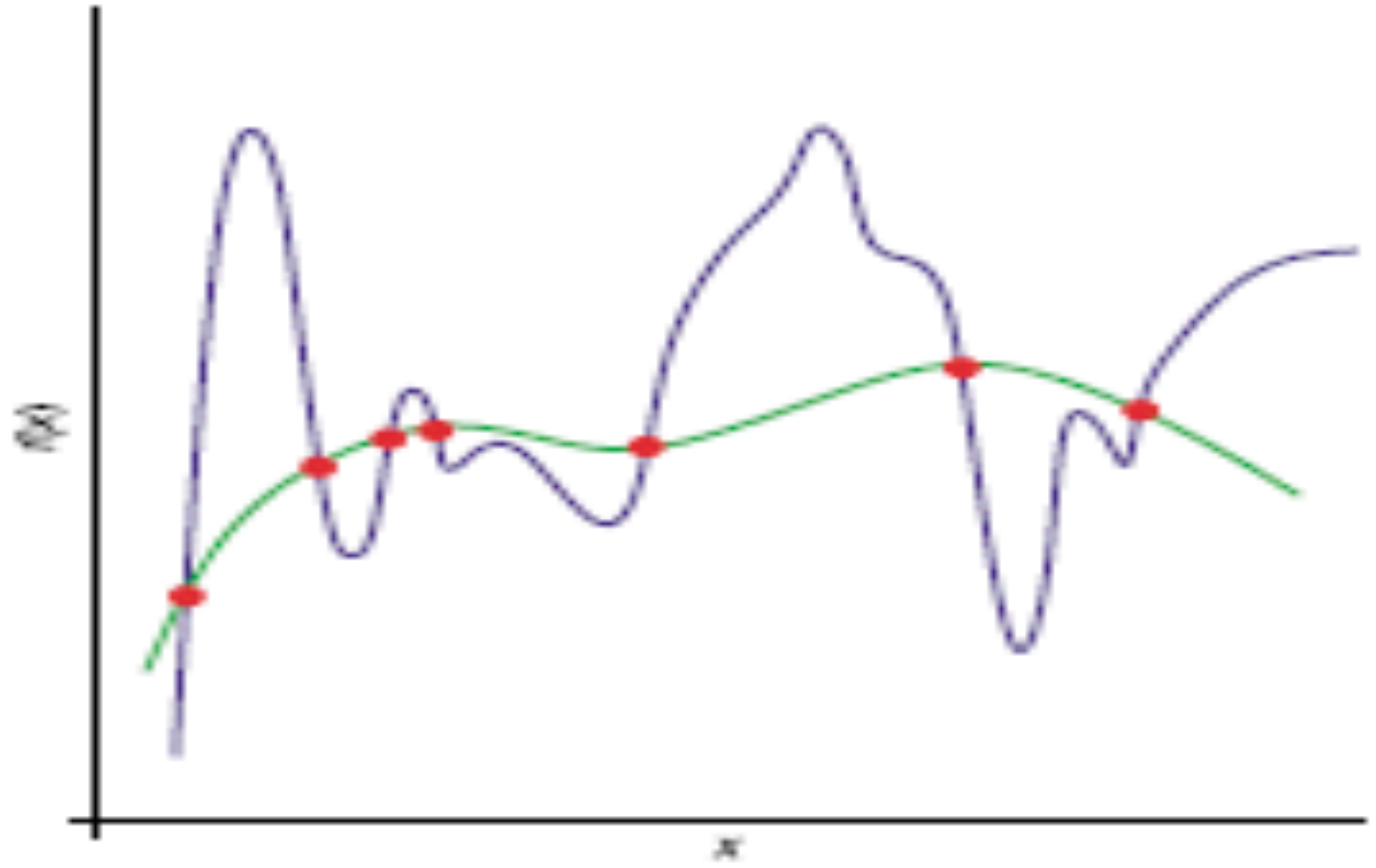
is a **good predictor** of y for a **future** input x (fitting the data is **not** enough!)

Statistical Learning Theory: prediction, not description

= data from f

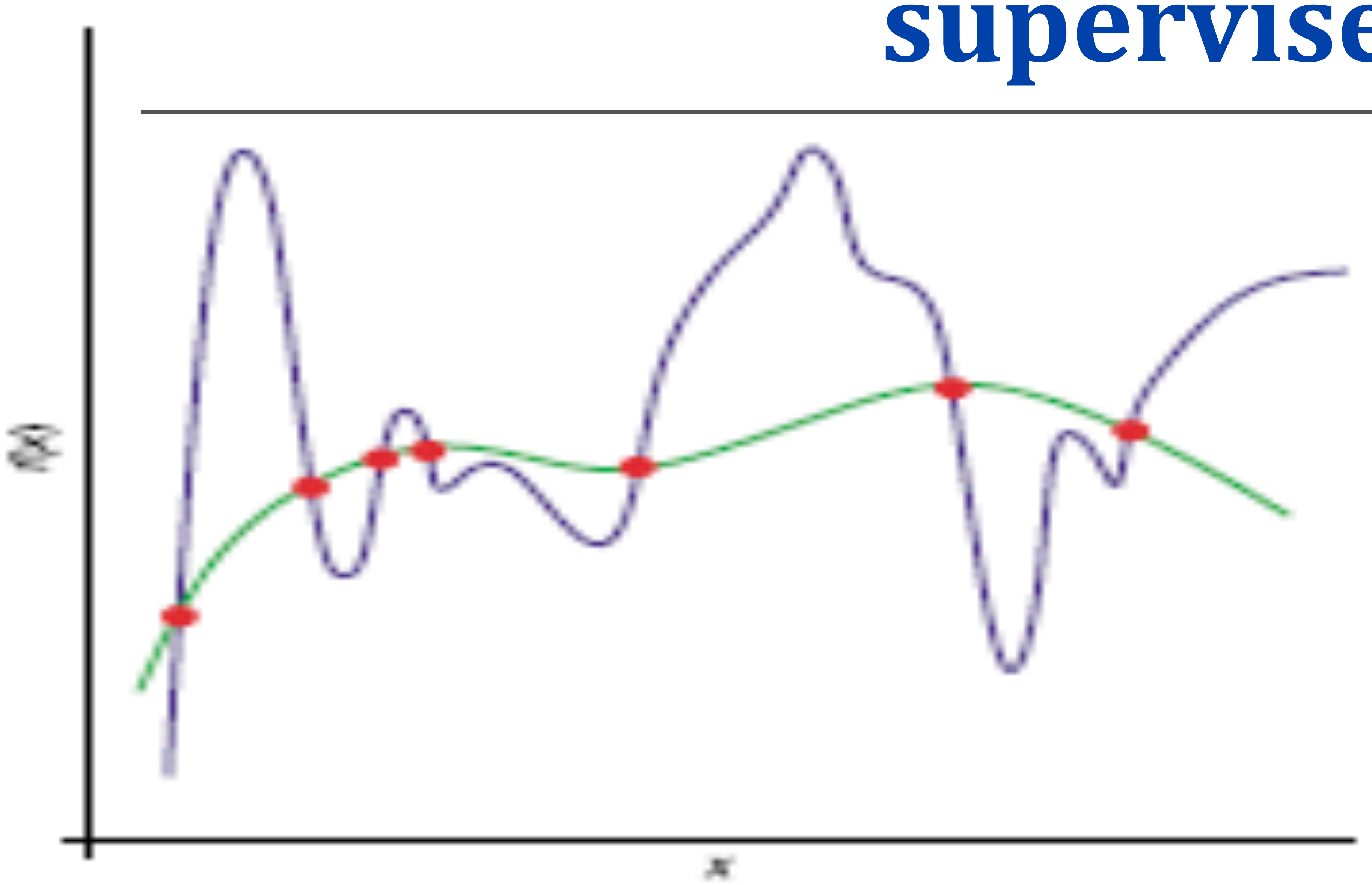
= function f

= approximation of f



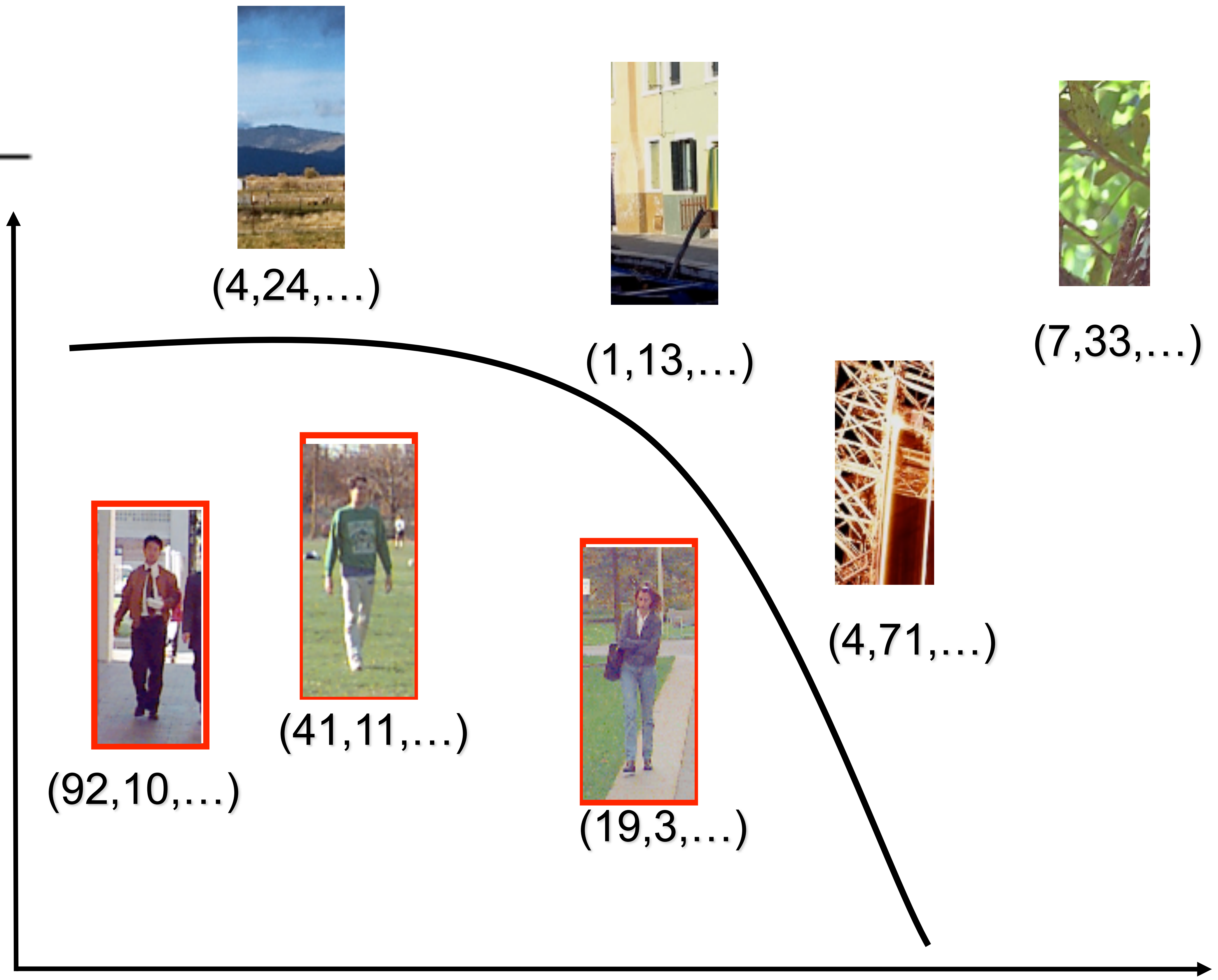
Intuition: Learning from data to predict well the value of the function where there are no data

Statistical Learning Theory: supervised learning



Regression

Classification



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.

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

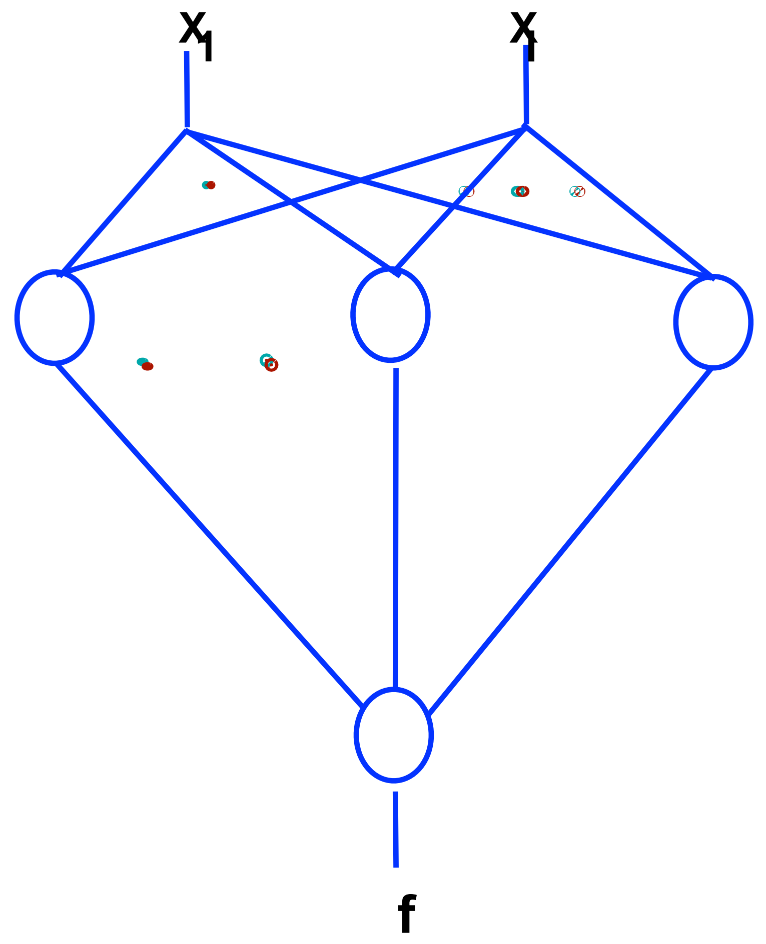
- theory must be chosen from a small hypothesis set (~ Occam razor, VC dimension,…)
- theory should not change much with new data...most of the time (stability)

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

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



Classical kernel machines — such as SVMs — correspond to <u>shallow</u> networks

Summary

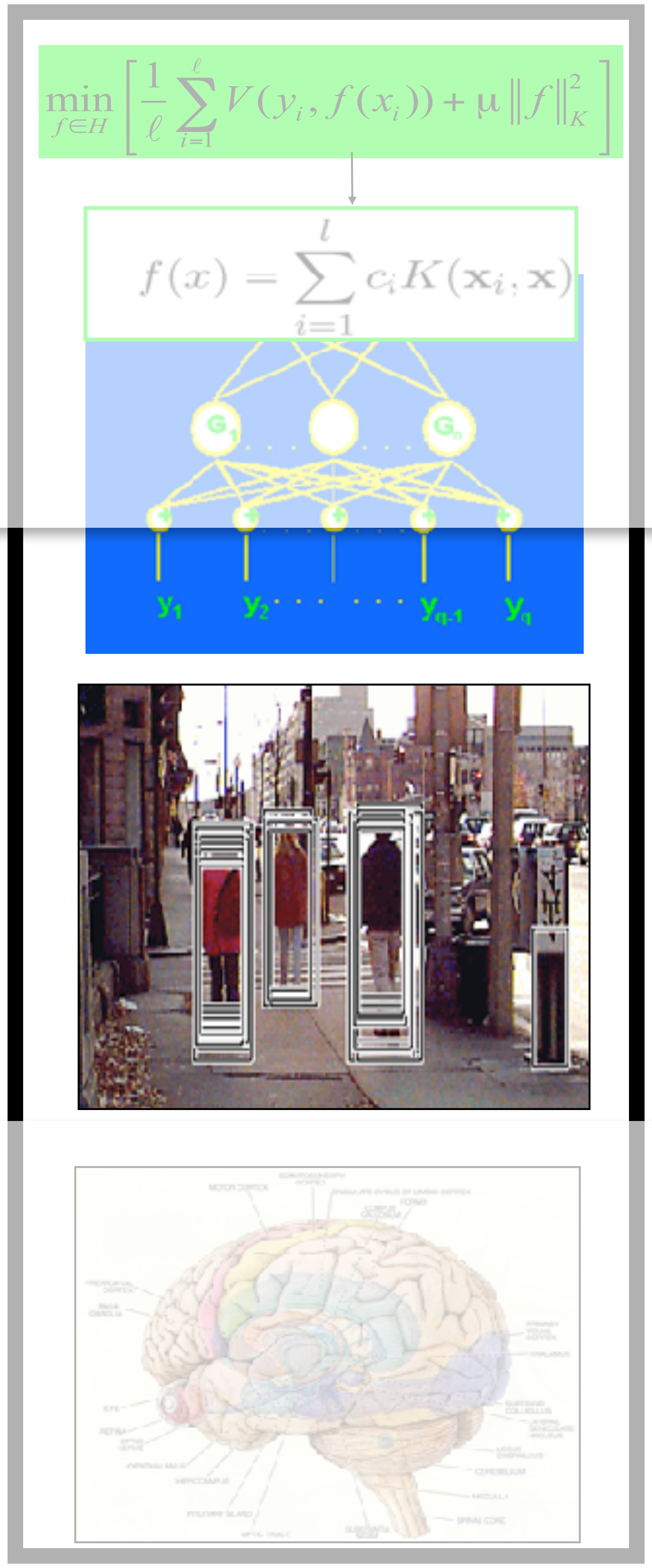
- A bit of history: Statistical Learning Theory

Summary: I told you about learning theory and predictivity. I told you about kernel machines and shallow networks.

Historical perspective: Examples of old Applications


Image
Output

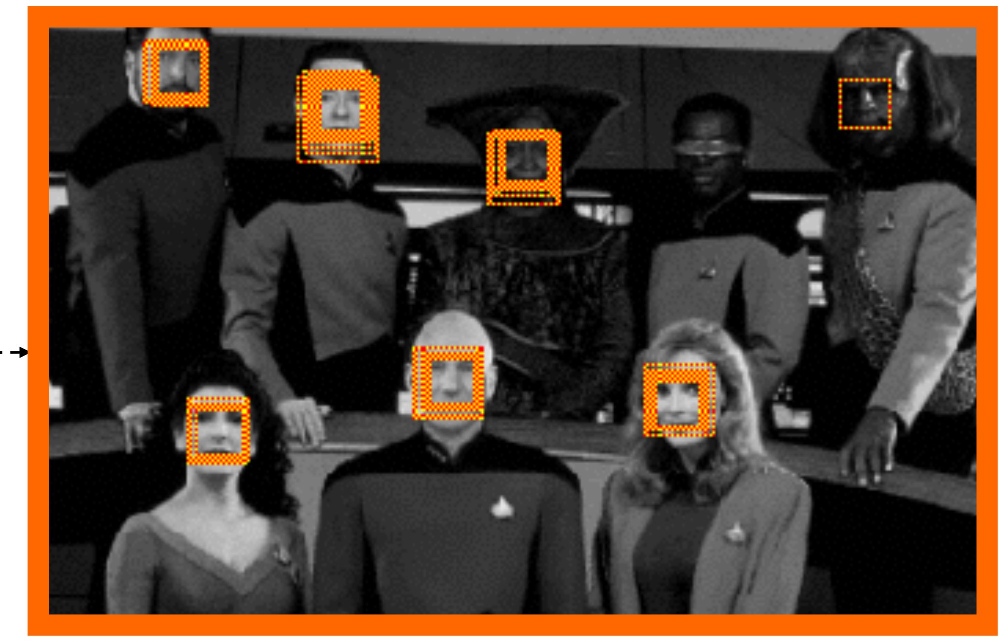
Engineering of Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms



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

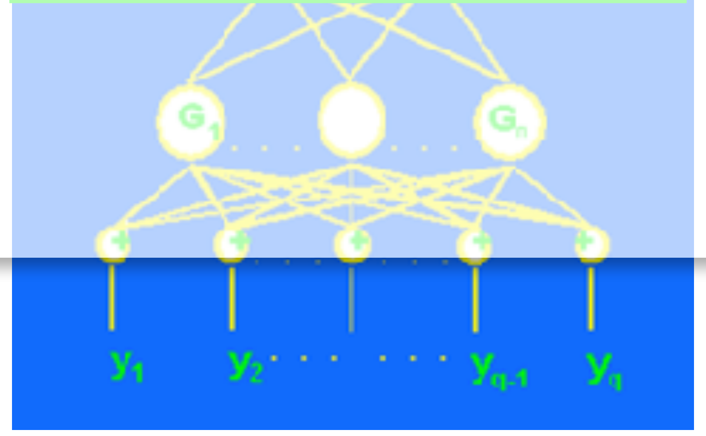
COMPUTATIONAL NEUROSCIENCE: models+experiments

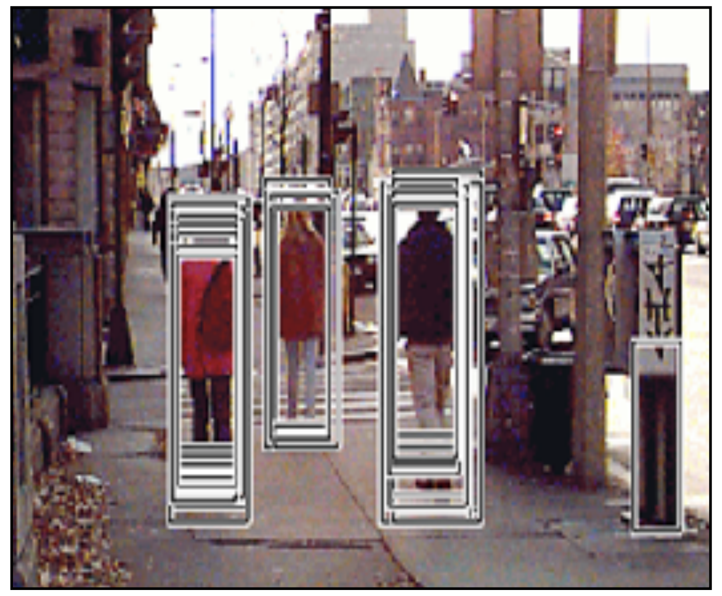
How visual cortex works

Engineering of Learning

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

$$f(x) = \sum_{i=1}^{l} c_i K(\mathbf{x}_i, \mathbf{x})$$

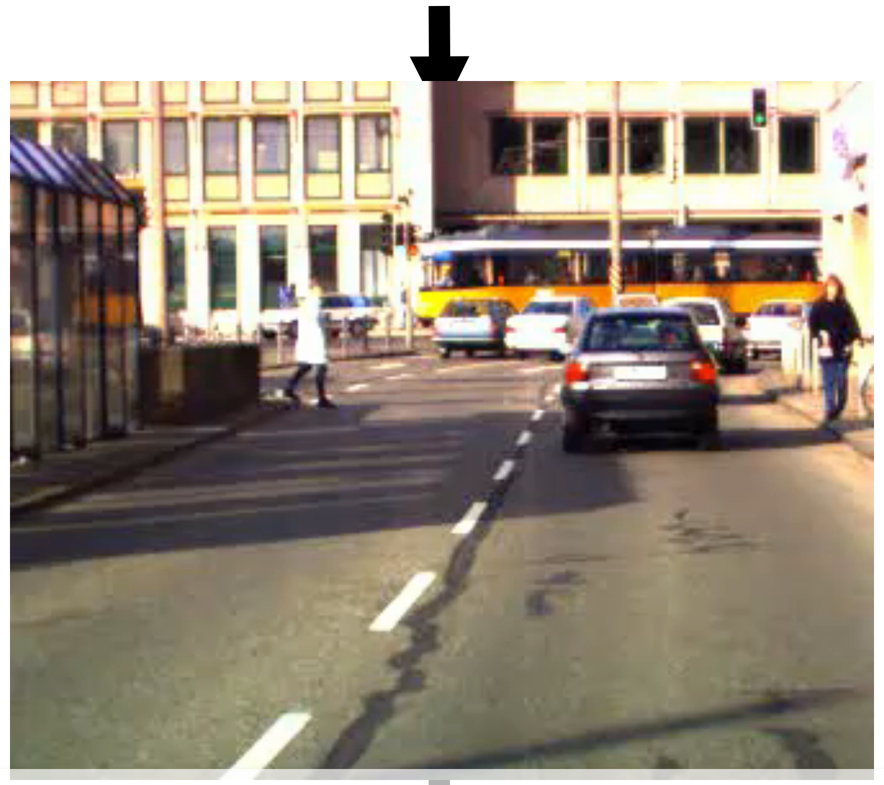




LEARNING THEORY
+
ALGORITHMS

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

….

Learning: bioinformatics

New feature selection SVM:

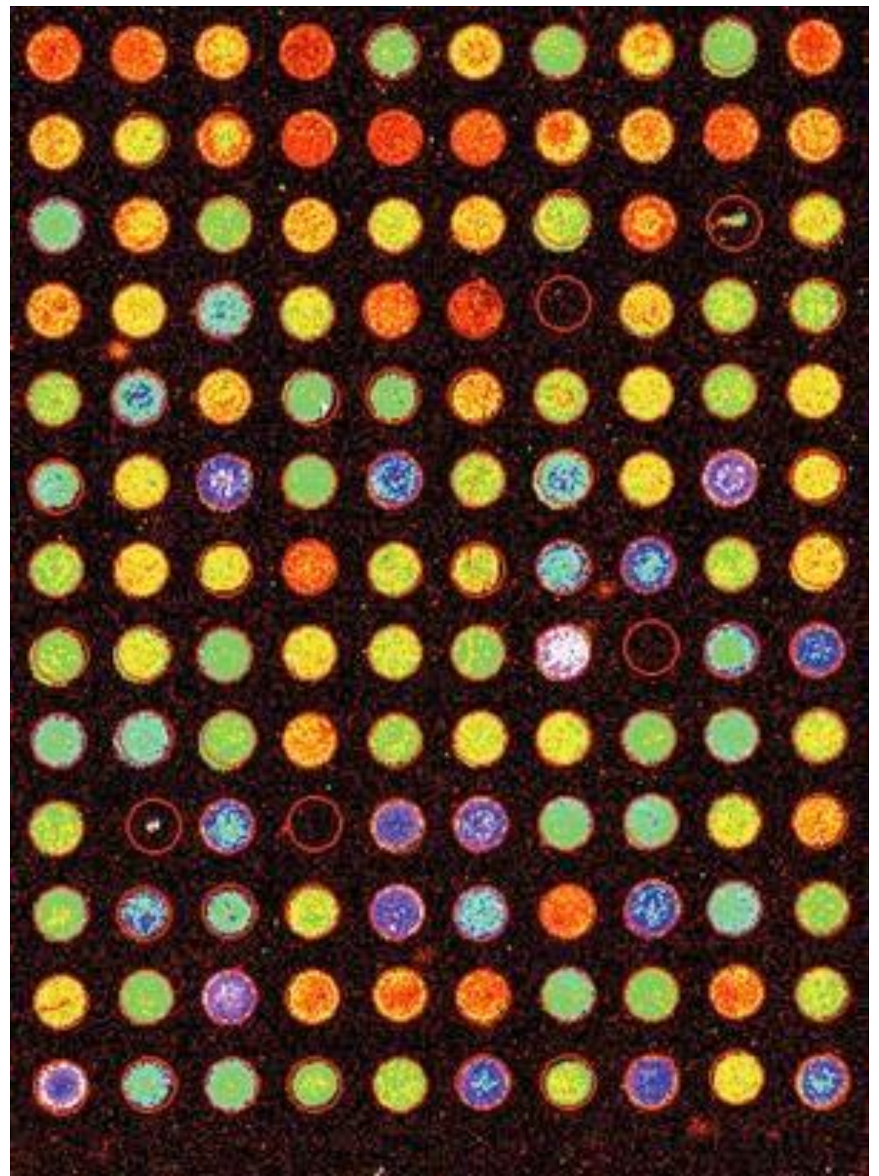
Only 38 training examples, 7100 features

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

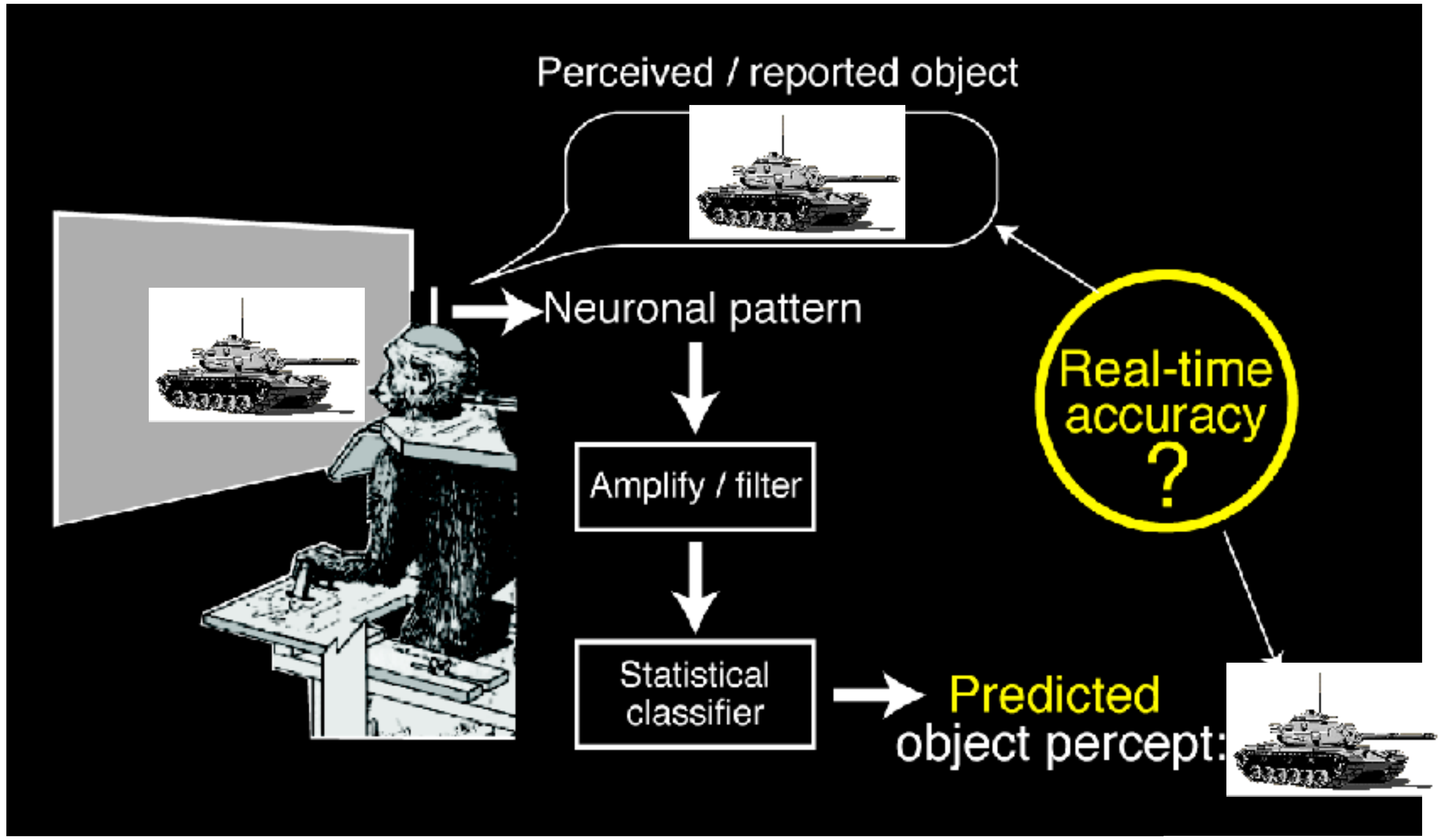
Support Vector Machine Classification of Microarray Data

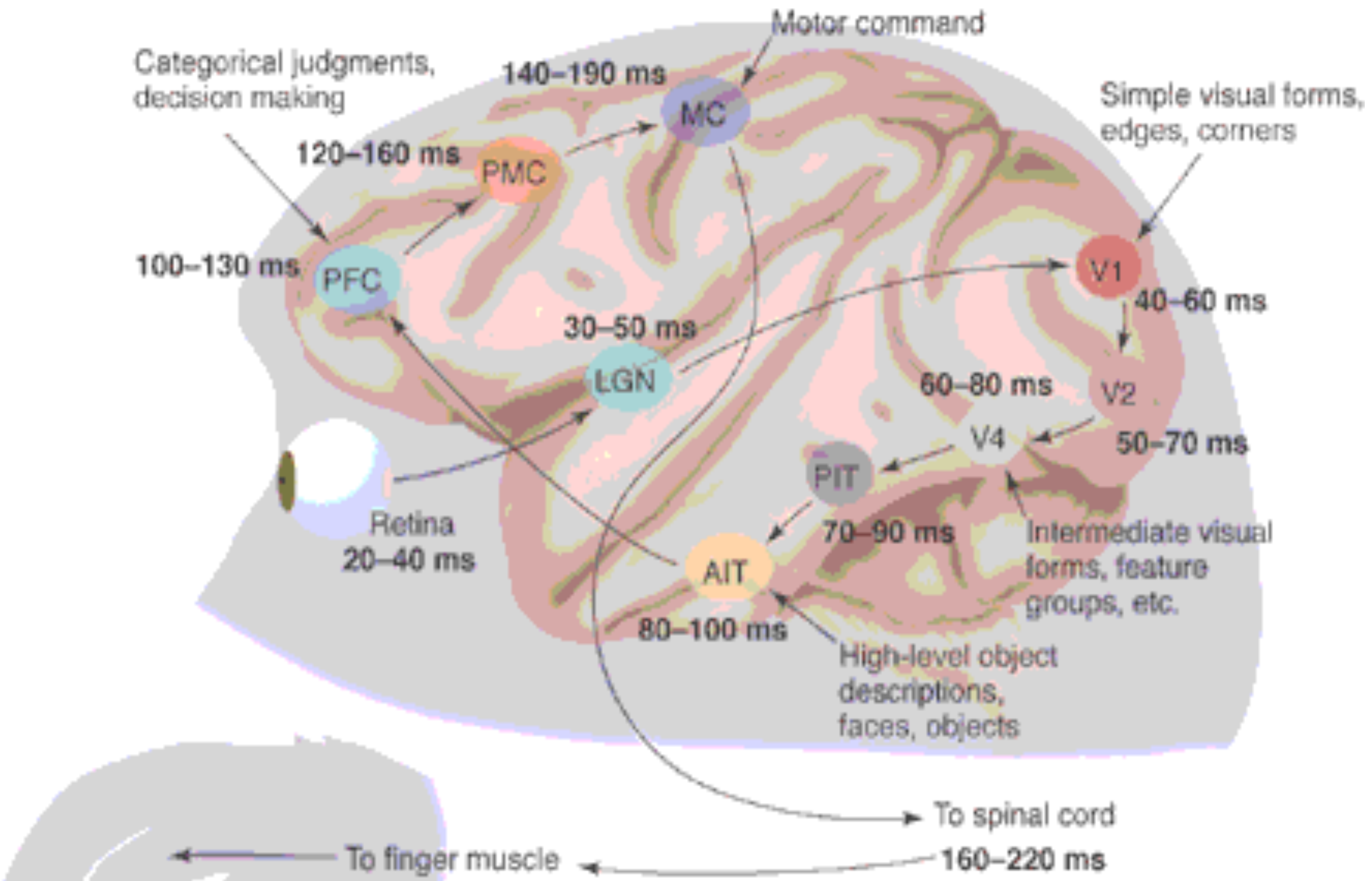
S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J.P. Mesirov, and T. Poggio



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.

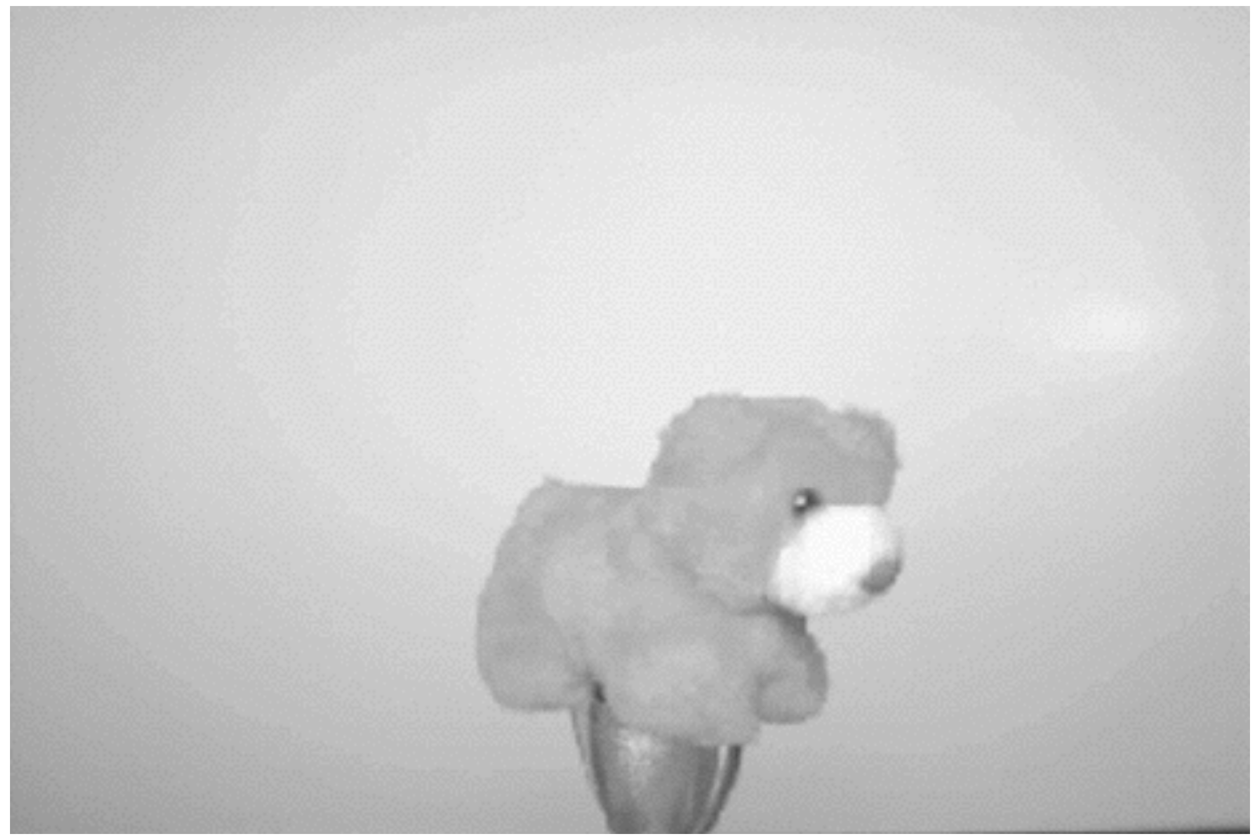
Decoding the neural code: Matrix-like read-out from the brain





Learning: image analysis

⇒ **Bear (0° view)**



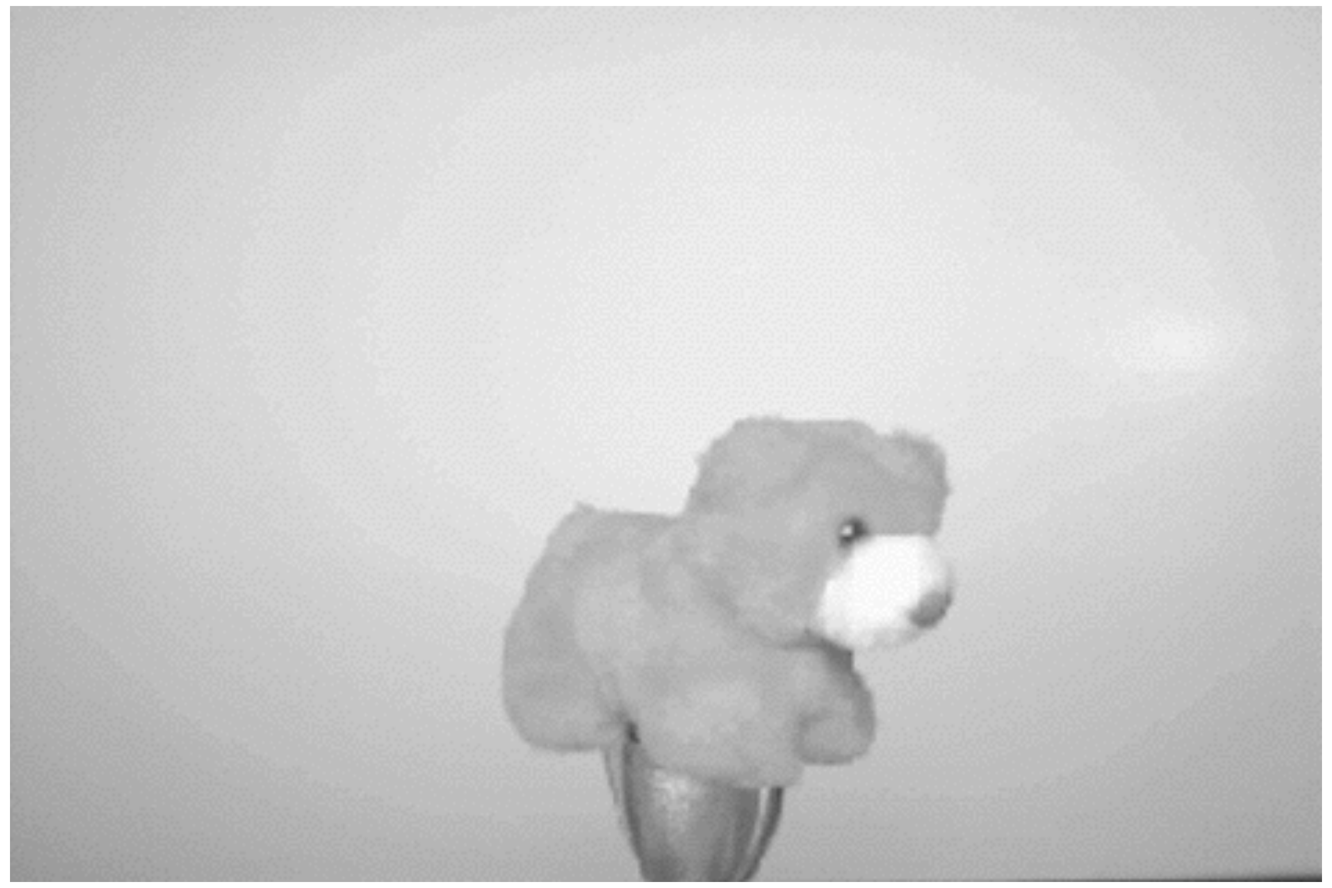
⇒ **Bear (45° view)**

Learning: image synthesis

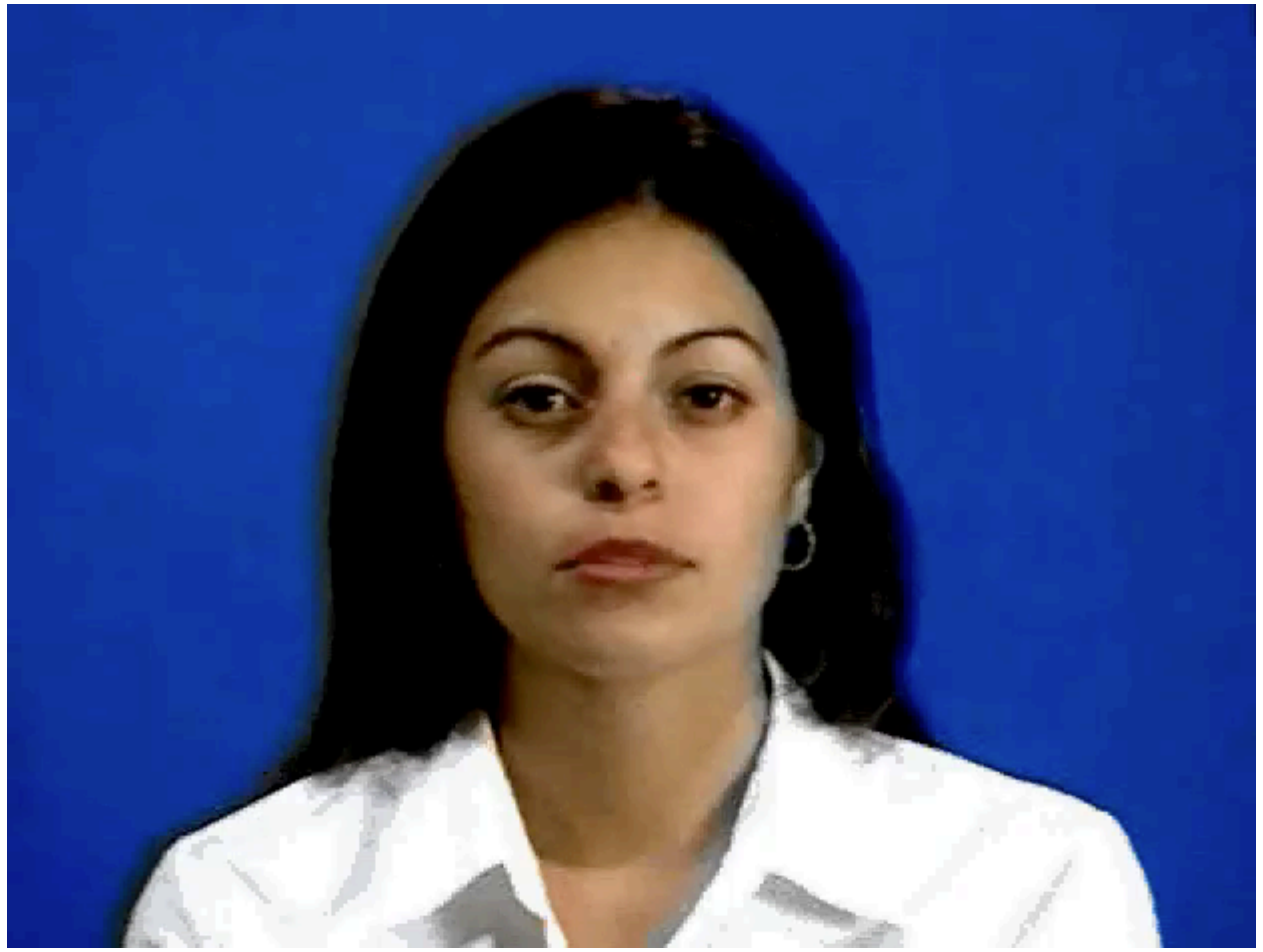
UNCONVENTIONAL GRAPHICS

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

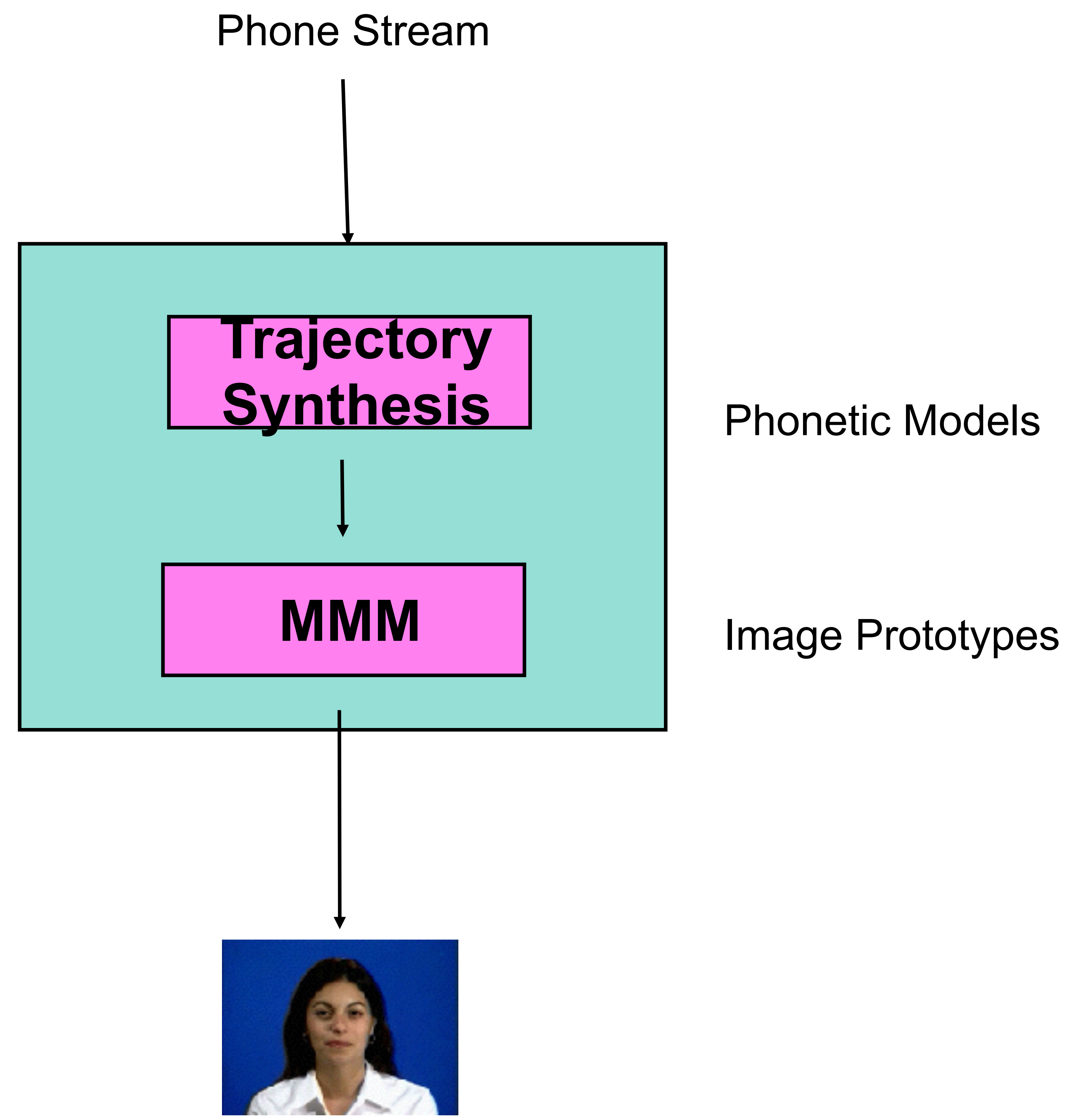
Tony Ezzat,Geiger, Poggio, **SigGraph 2002**

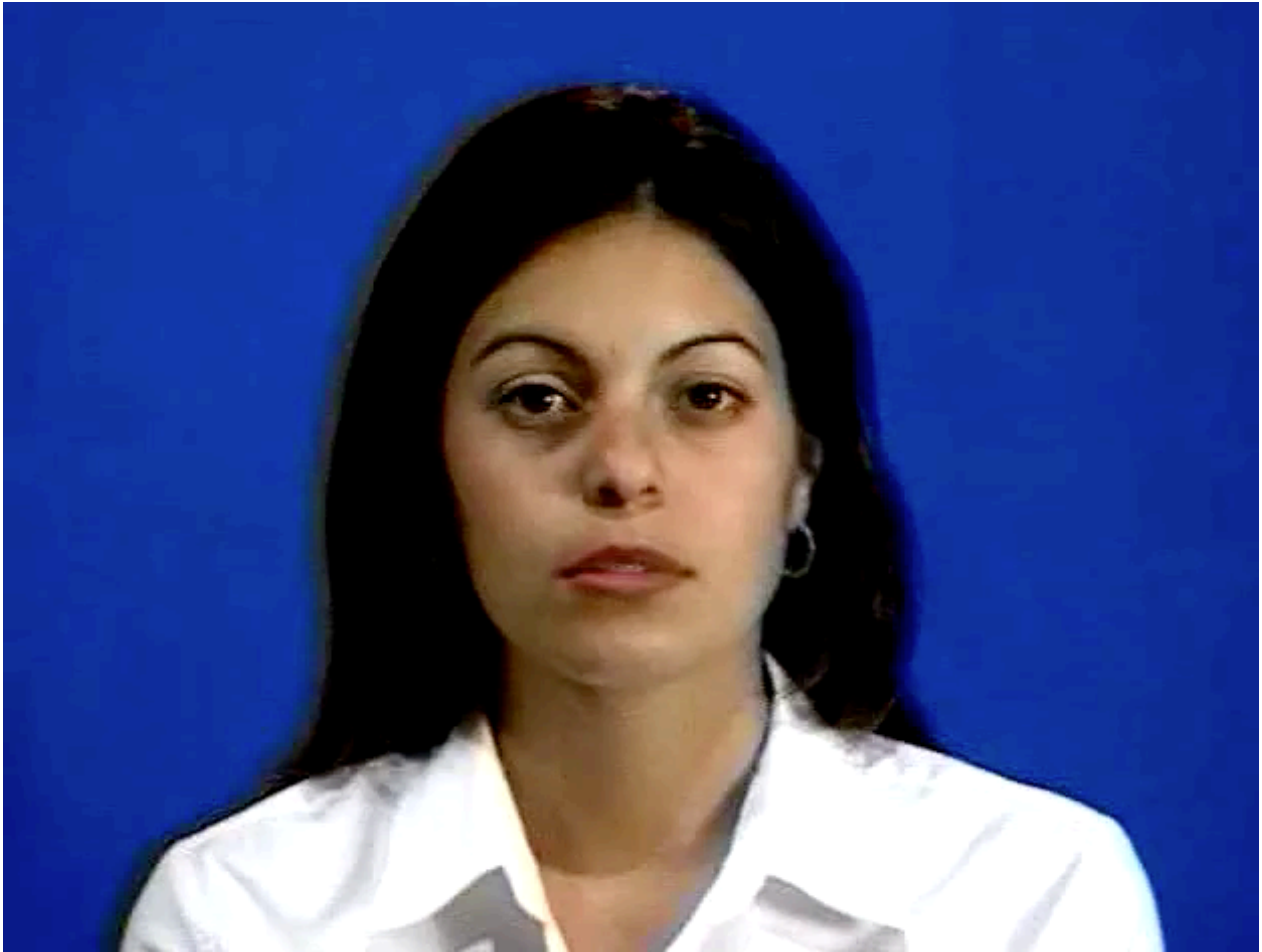
1. Learning

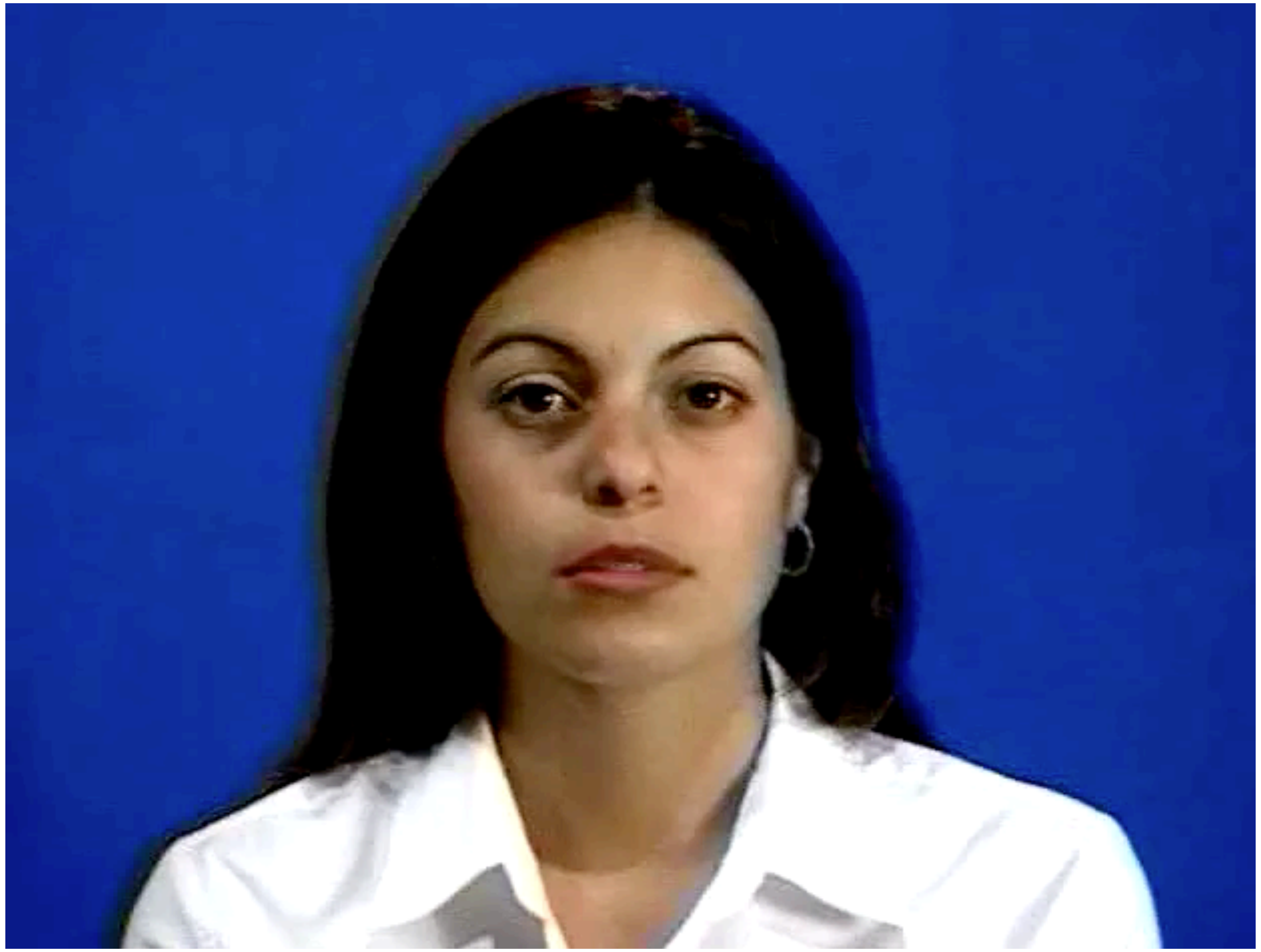
System learns from 4 mins of video face appearance (Morphable Model) and speech dynamics of the person

2. Run Time

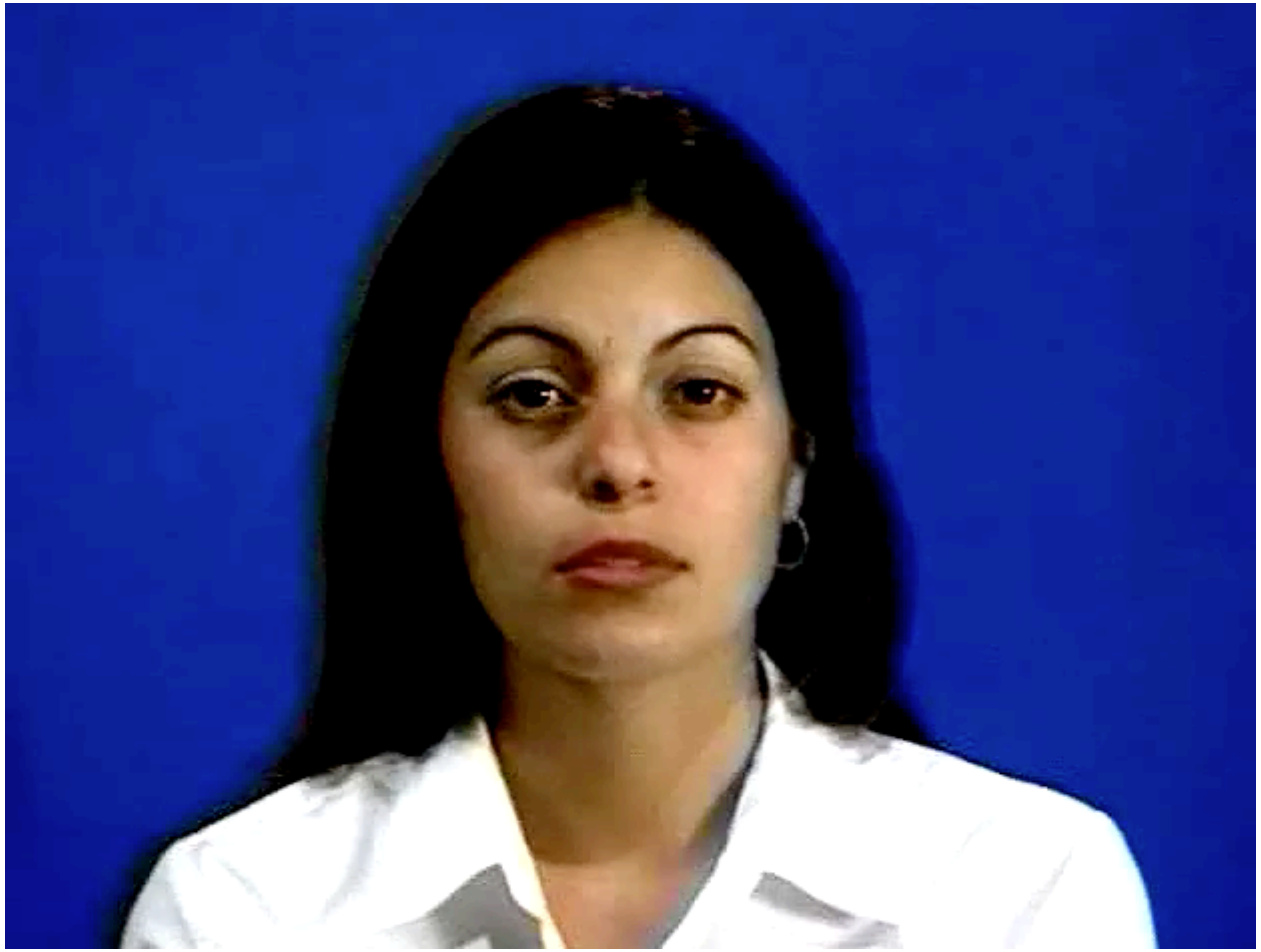
For any speech input the system provides as output a synthetic video stream



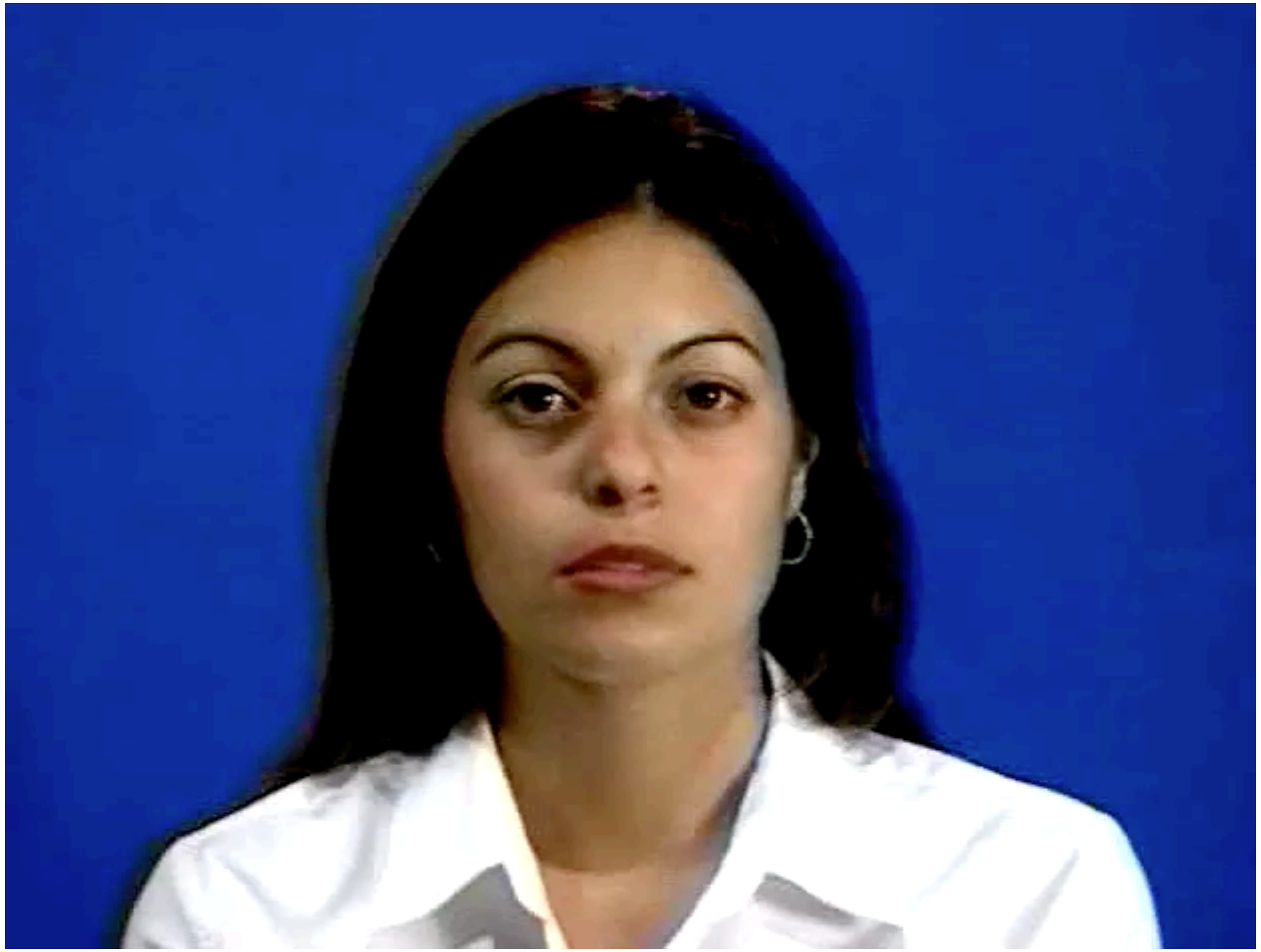




B-Dido

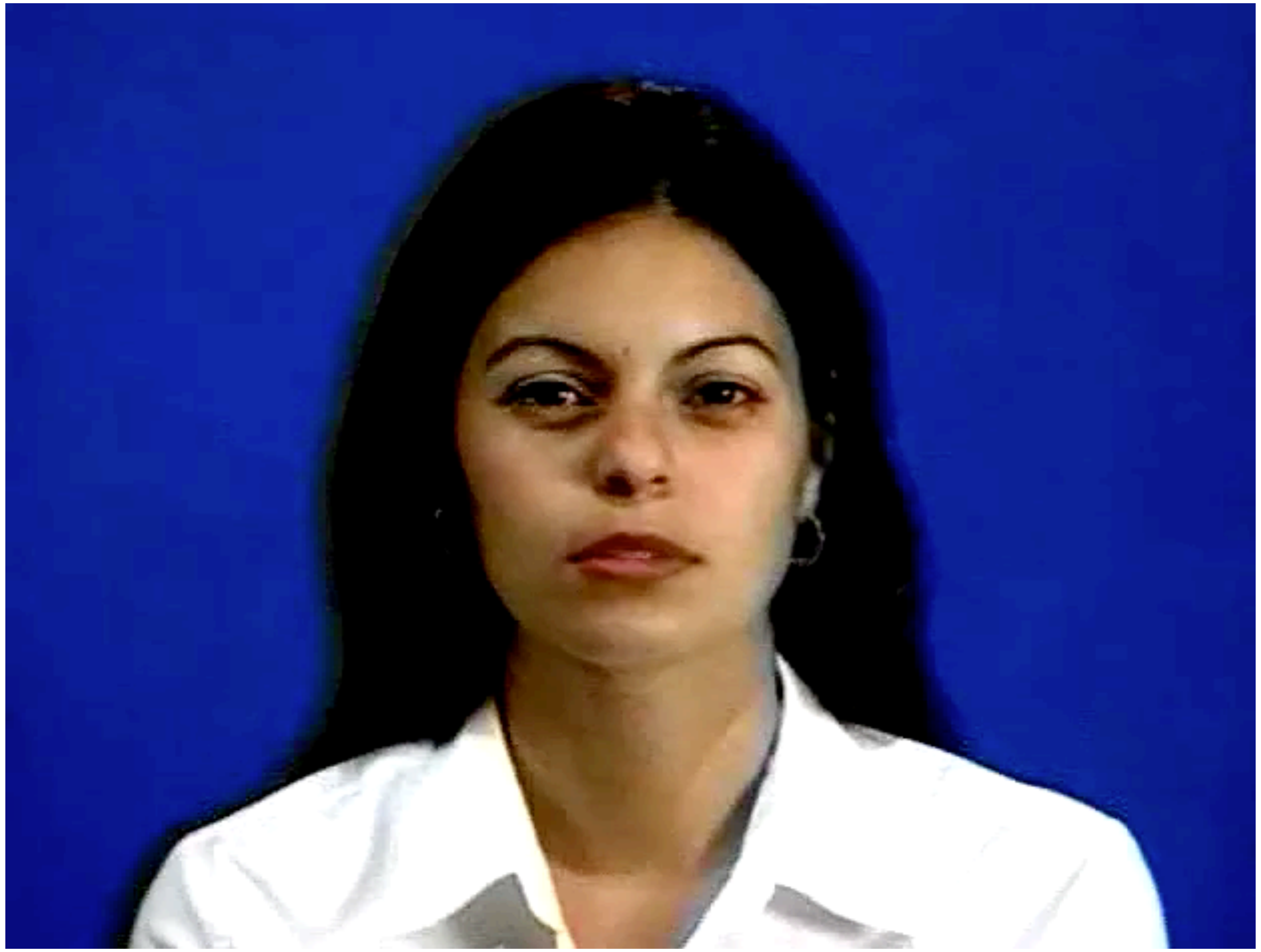


C-Hikaru

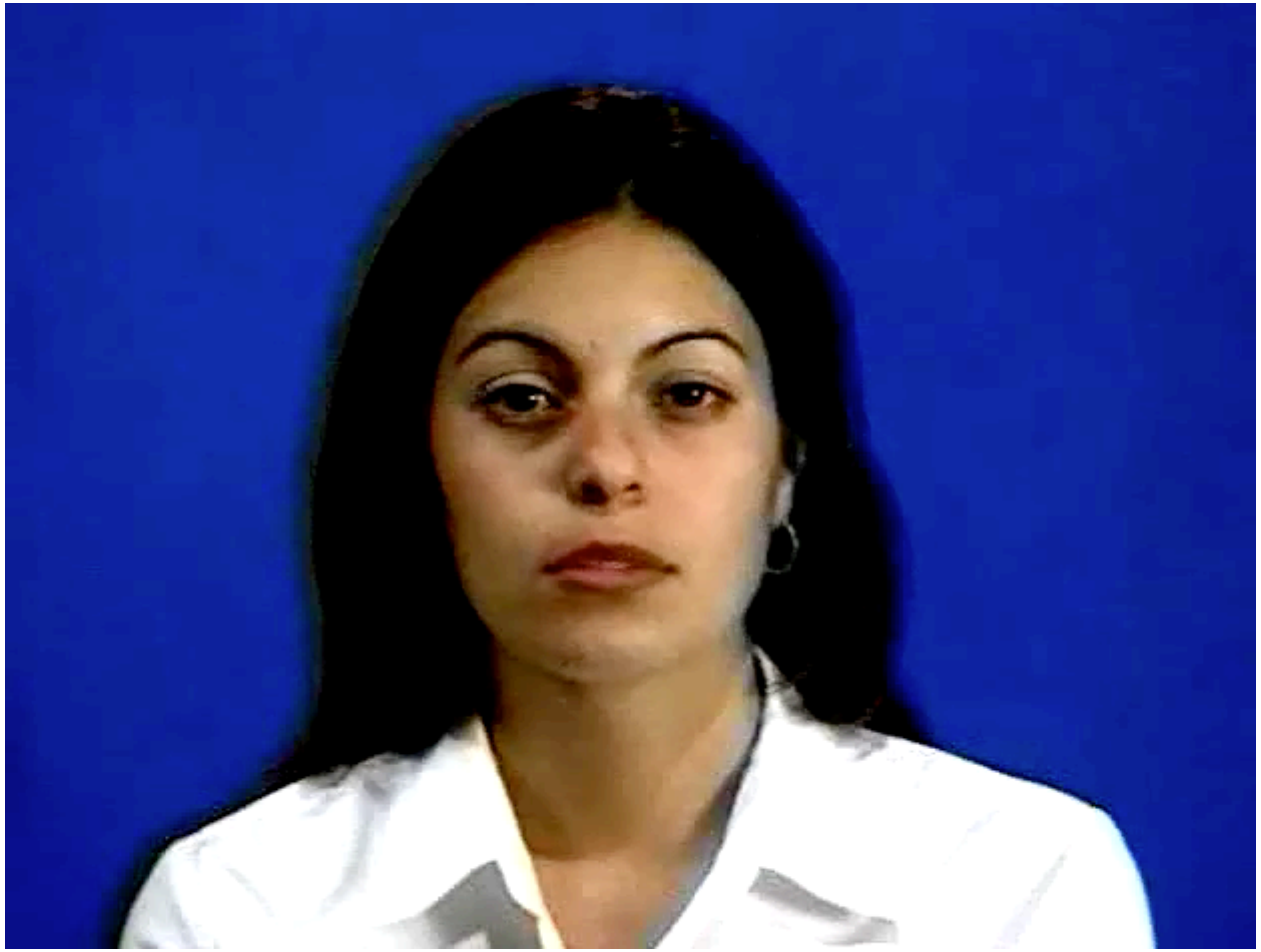


D-Denglijun

E-Marylin



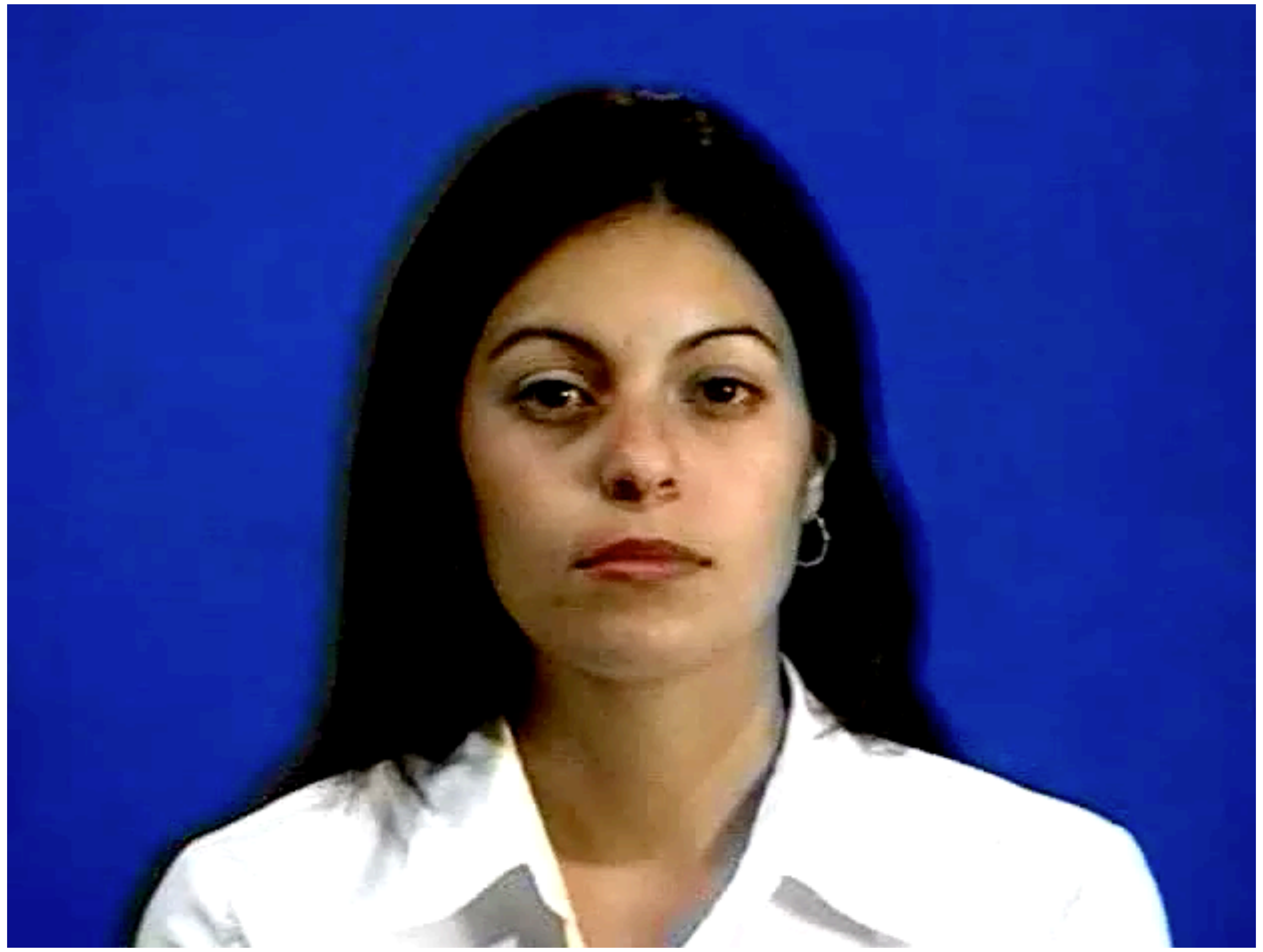
G-Katie



H-Rehema

I-Rehemax

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



L-real-synth

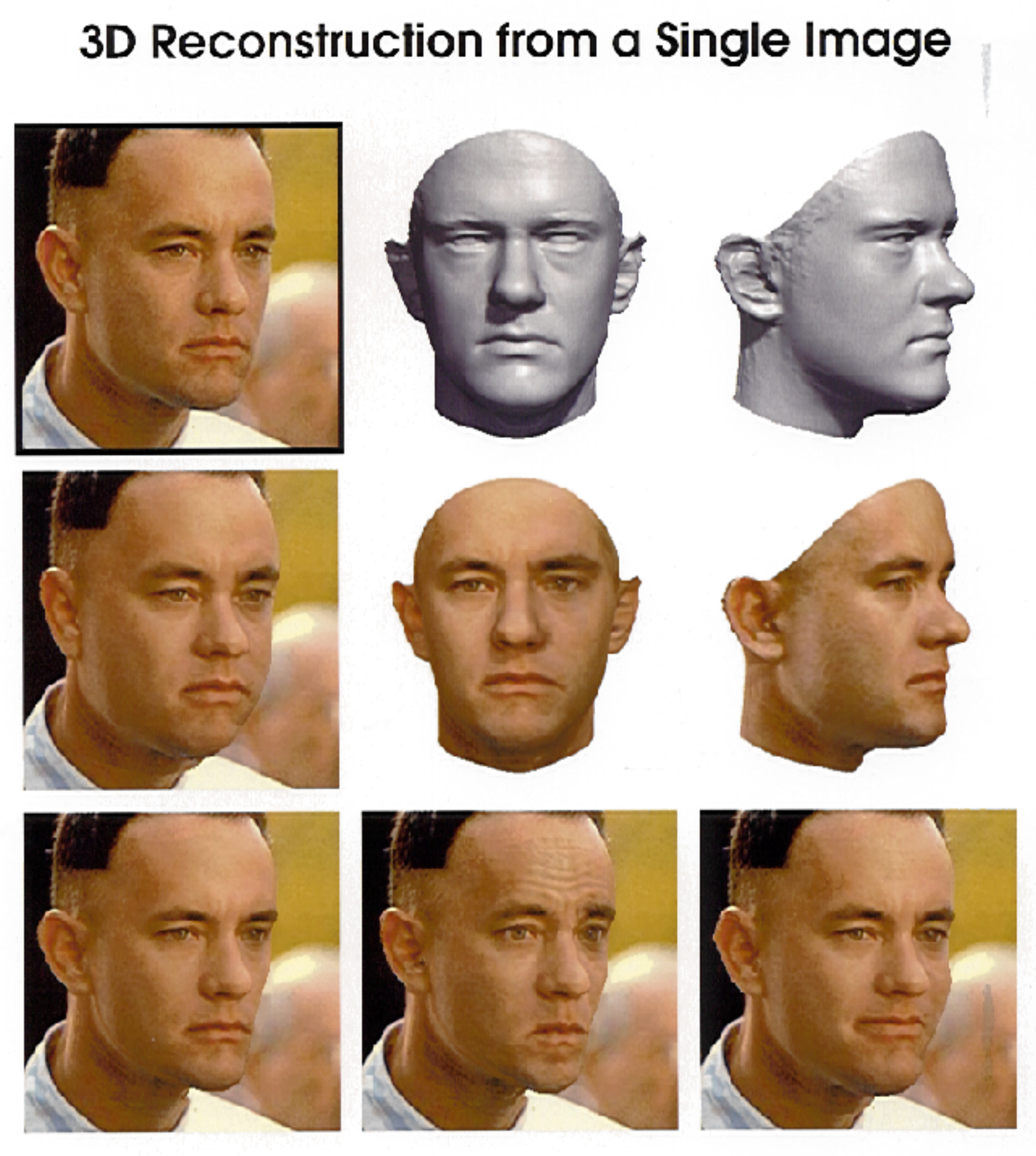
A Turing test: what is real and what is synthetic?

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**

Learning: image synthesis



Blanz and Vetter,
MPI
SigGraph ‘99

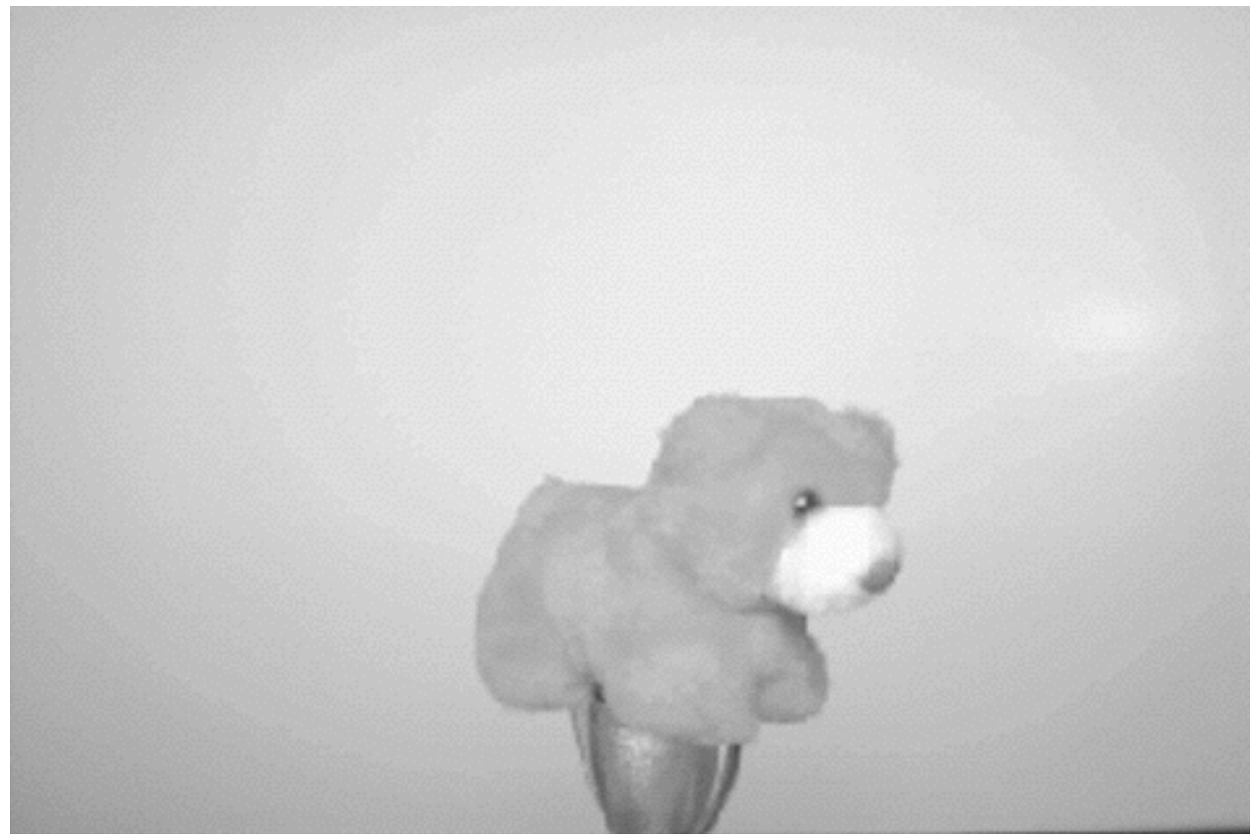
Learning: image synthesis

Blanz and Vetter,
MPI
SigGraph '99

Similar to today's GANs

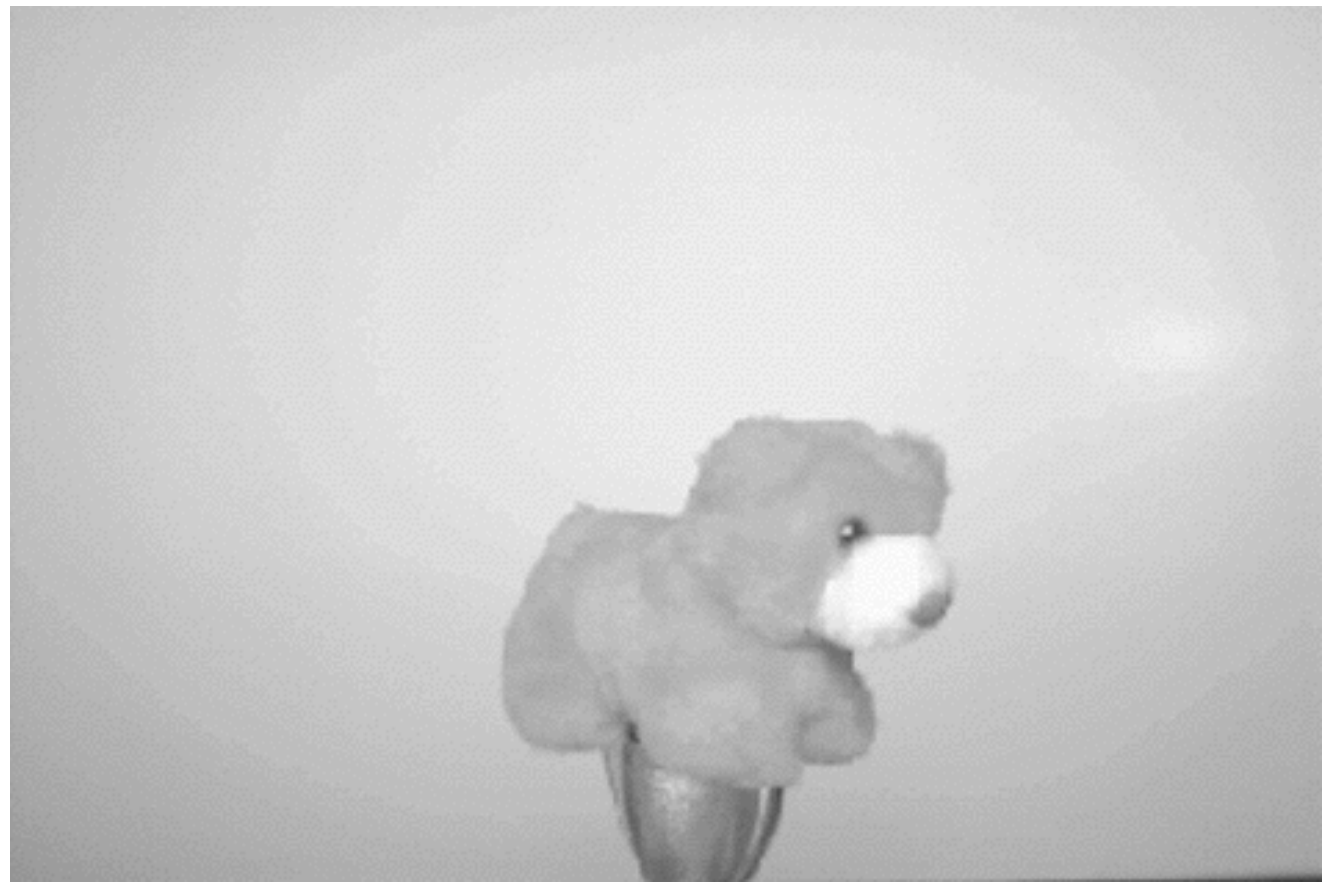
============>

Θ **= 0° view** $\Rightarrow$



===========>

Θ **= 45° view** $\Rightarrow$



Very Low-Band Video
E-Mail Demonstration

Summary

- A bit of history: old applications

Summary: I told you about old applications of ML, mainly kernel machines to give 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

- Course description/logistic
- Motivations for this course: a golden age for new AI, the key role of Machine Learning, CBMM, the MIT Quest: ***Intelligence, the Grand Vision***
- A bit of history: Statistical Learning Theory and Applications
- Deep Learning

Deep Learning

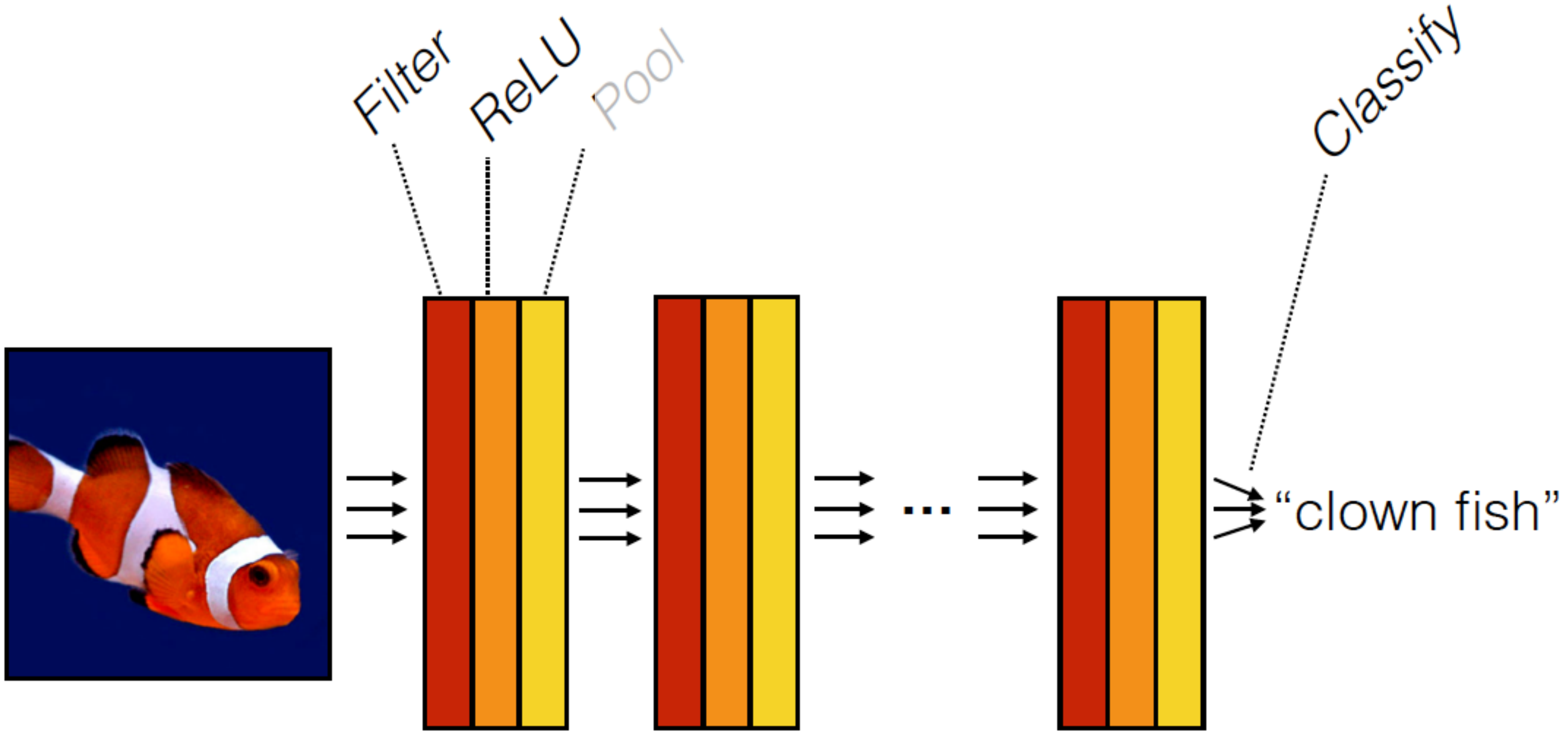
9.520/6.860

- Course focuses on algorithms and theory for supervised learning.

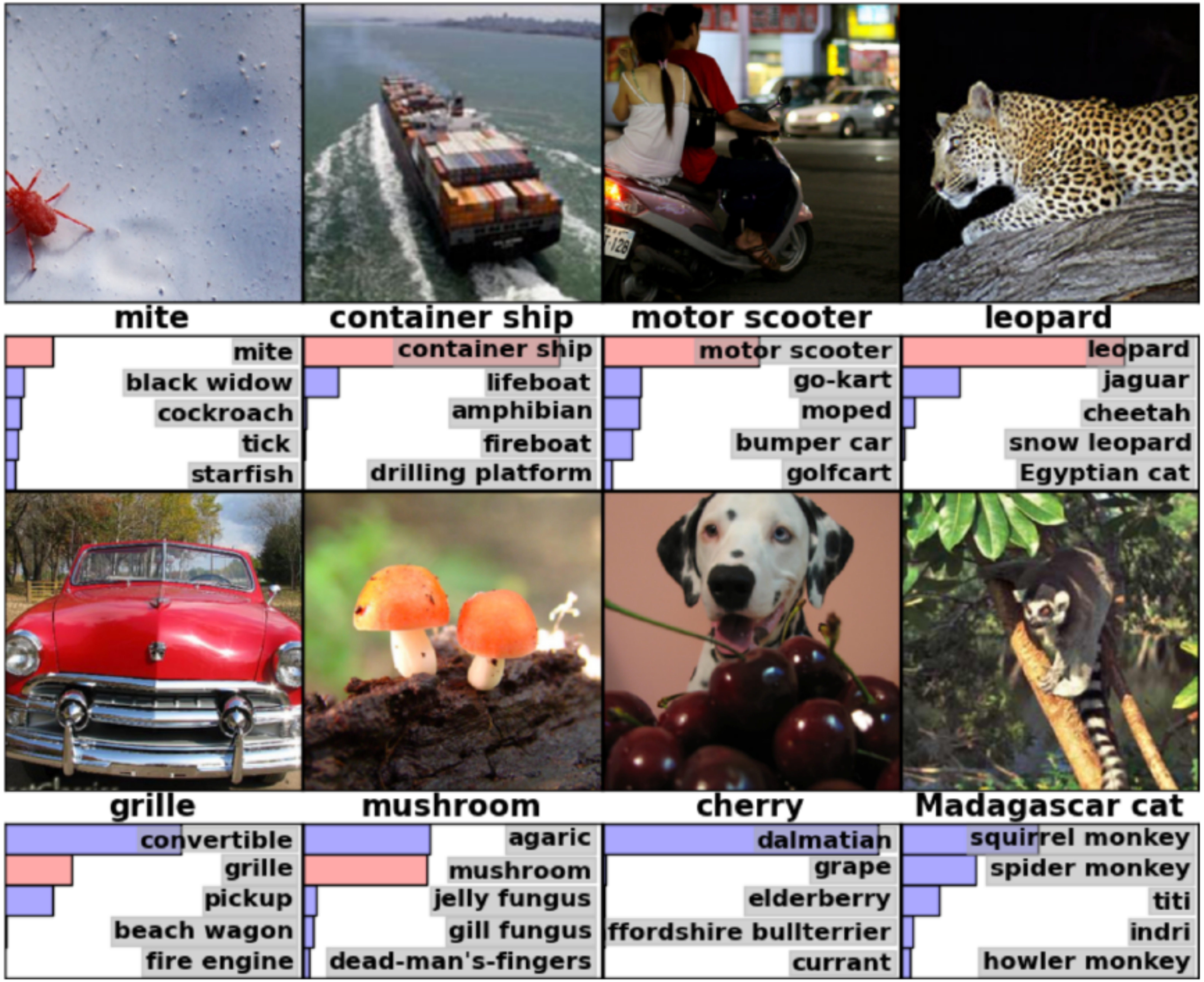
- Regularization techniques, Kernel machines, batch and online supervised learning, sparsity.

- Deep learning and theory of it, based on first part of the class

Computation in a neural net



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



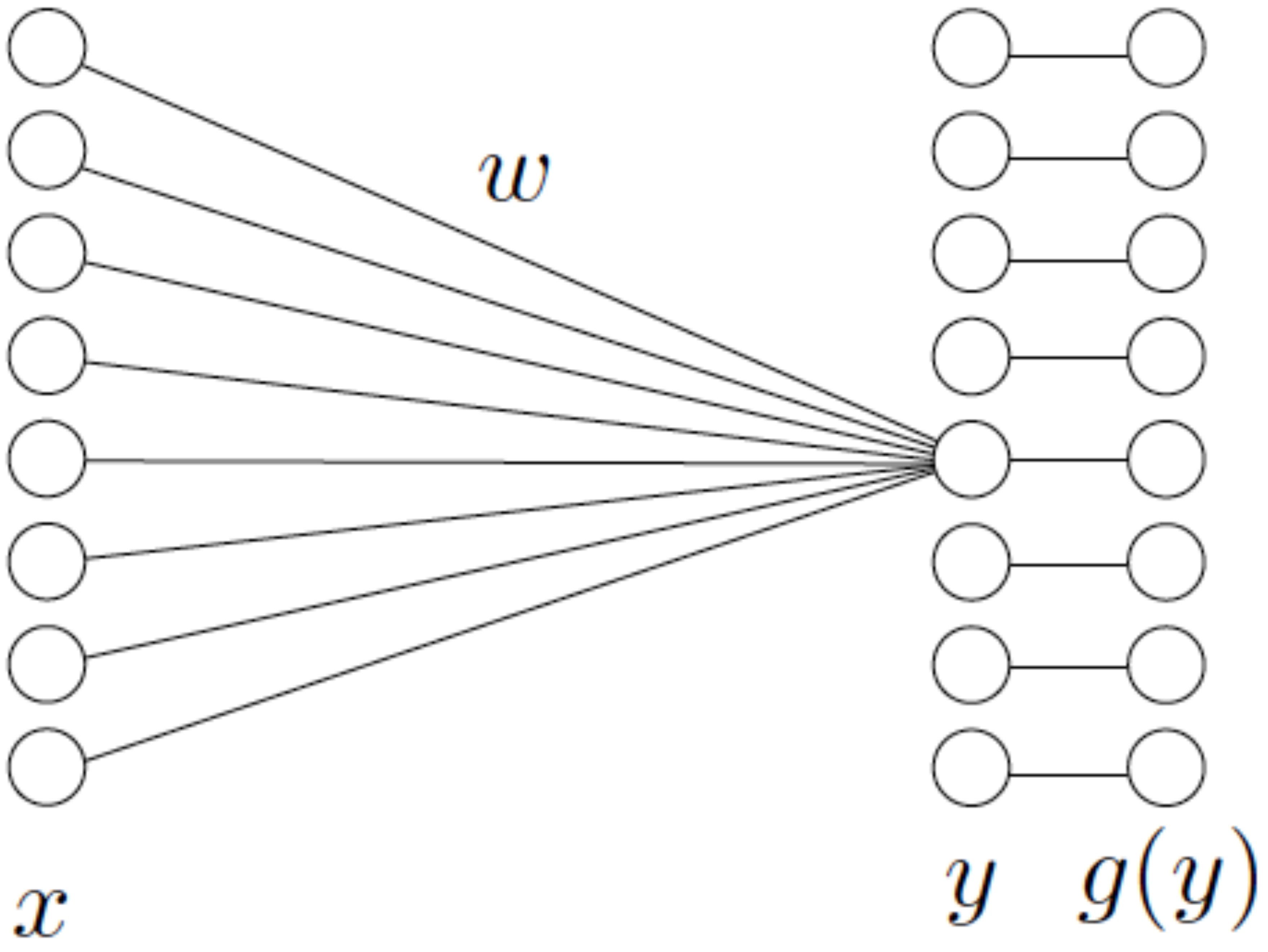
Krizhevsky et al. NIPS 2012

Is the lack of a theory a problem for DCLNs?

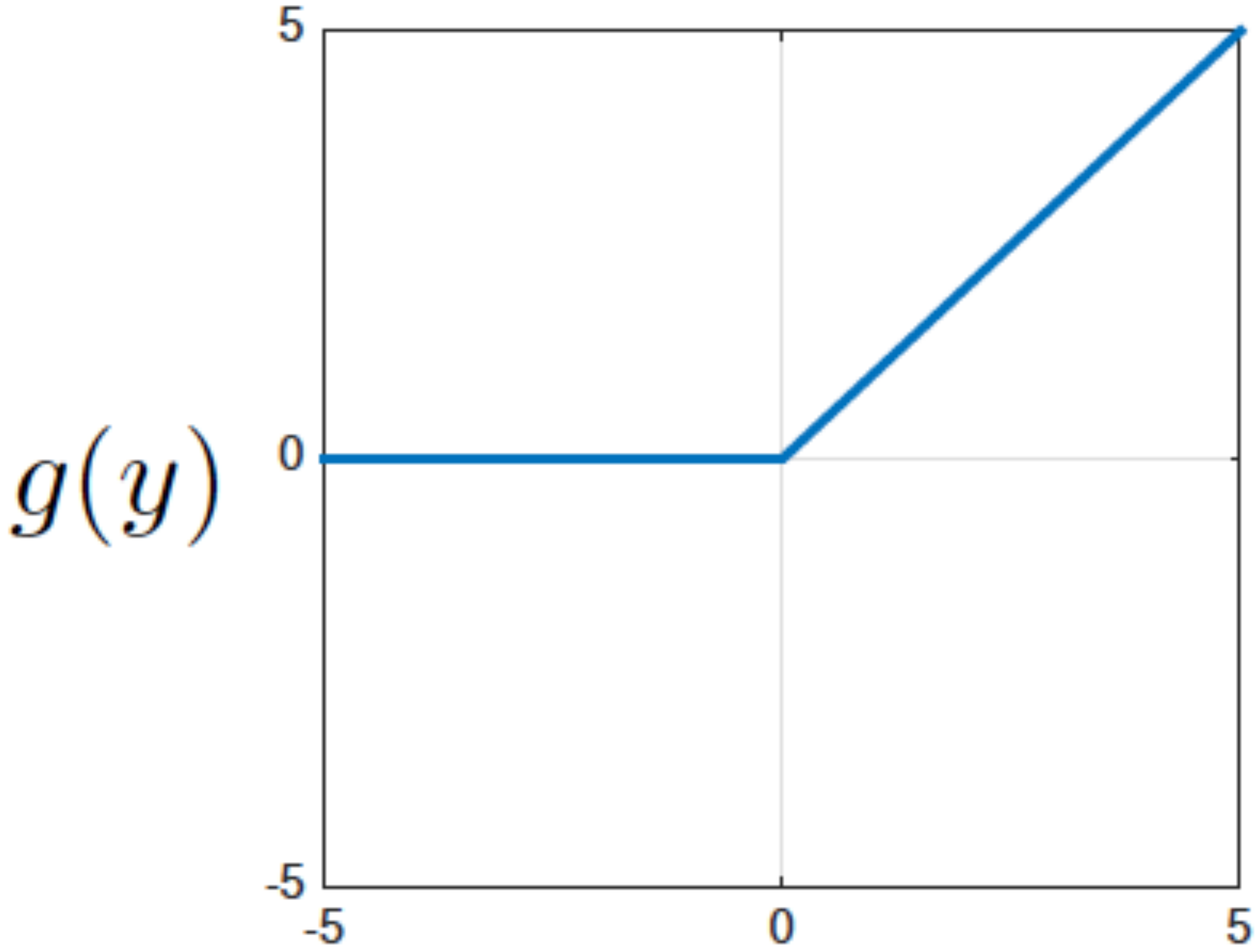
In Poggio and Smale (2003) we wrote "*A comparison with real brains offers another, and probably related, challenge to learning theory. The ``learning algorithms'' we have described in this paper correspond to one-layer architectures. Are hierarchical architectures with more layers justifiable in terms of learning theory?* Fifteen years later, a most interesting theoretical question, both for machine learning and neuroscience, is indeed *why hierarchies.*

Deep nets : a theory is needed
(after alchemy, chemistry)

Computation in a neural net

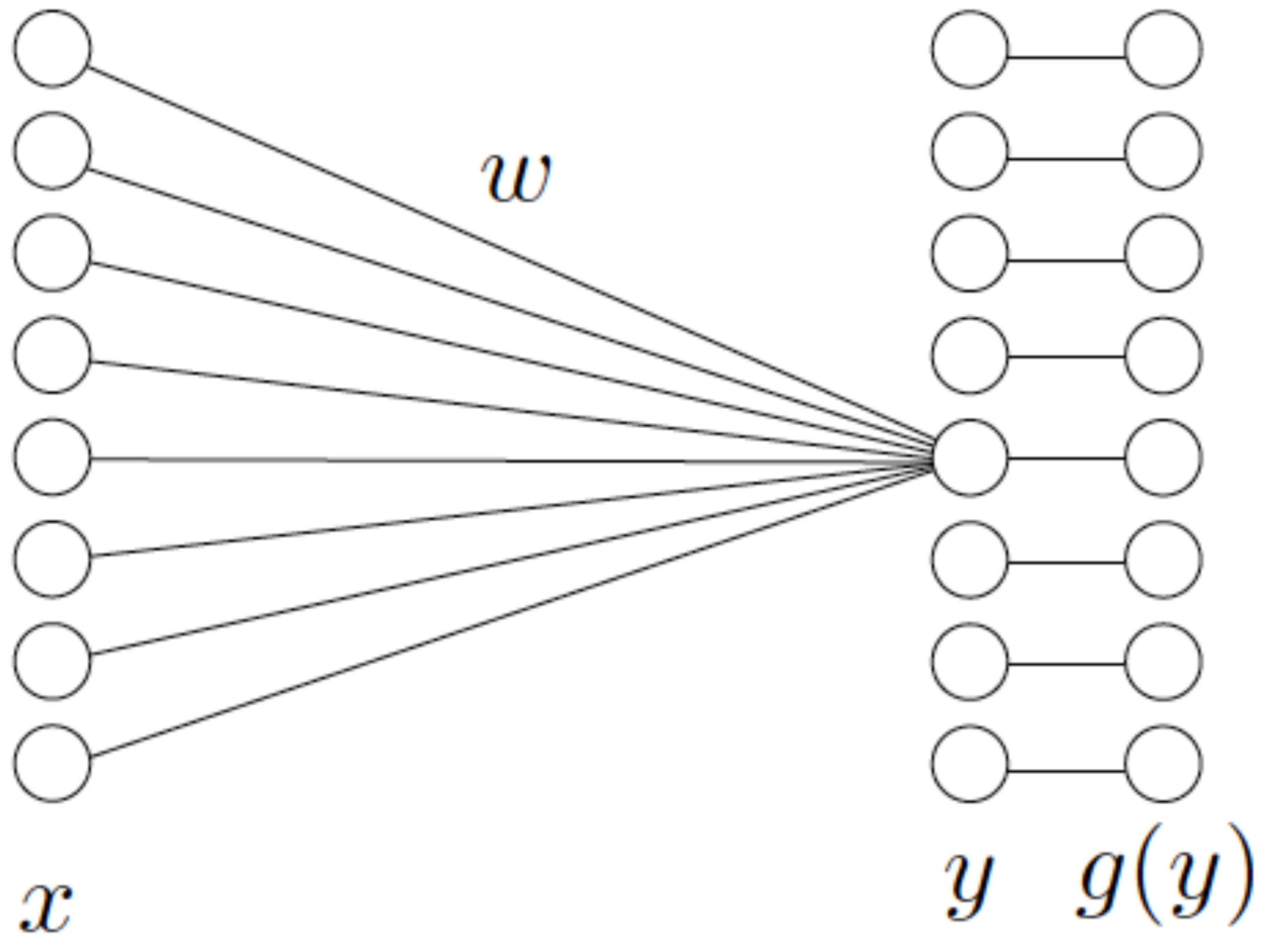


Rectified linear unit (ReLU)

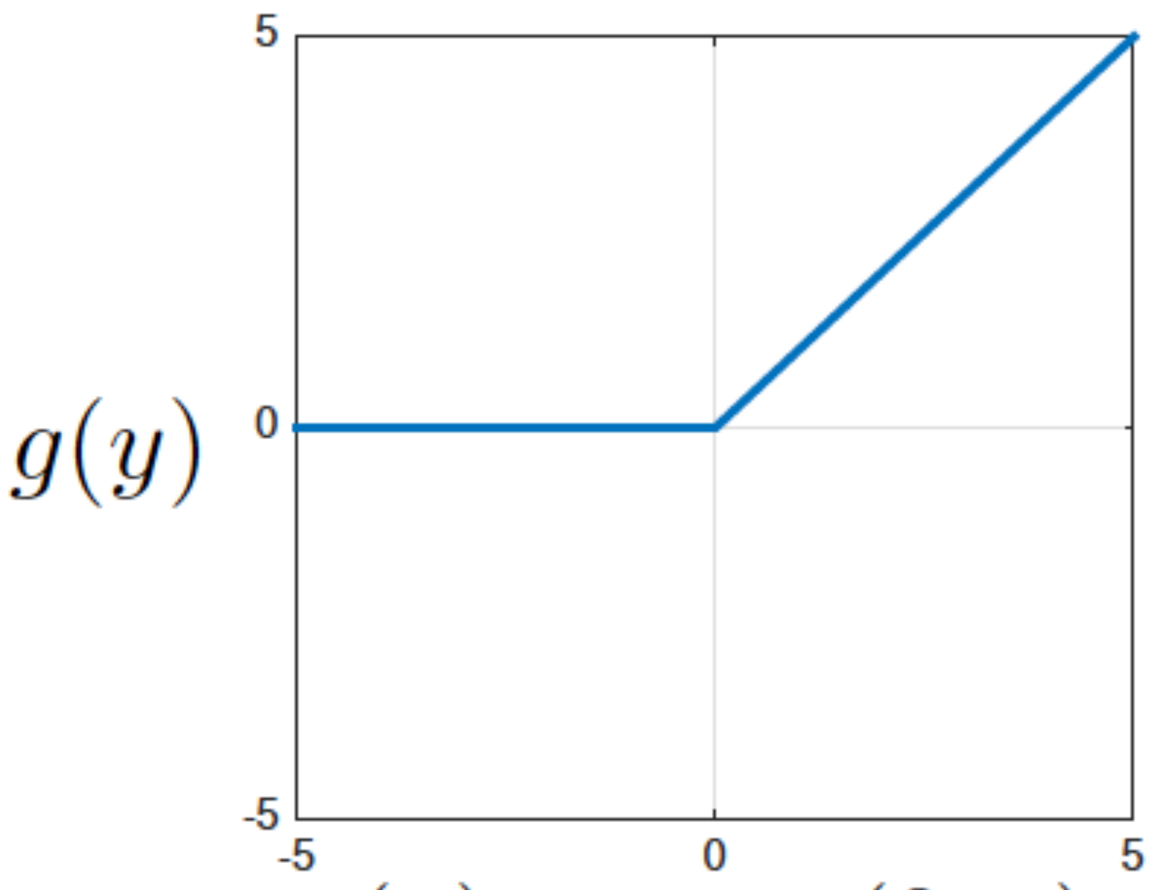


$$g(y) = \max(0, y)$$

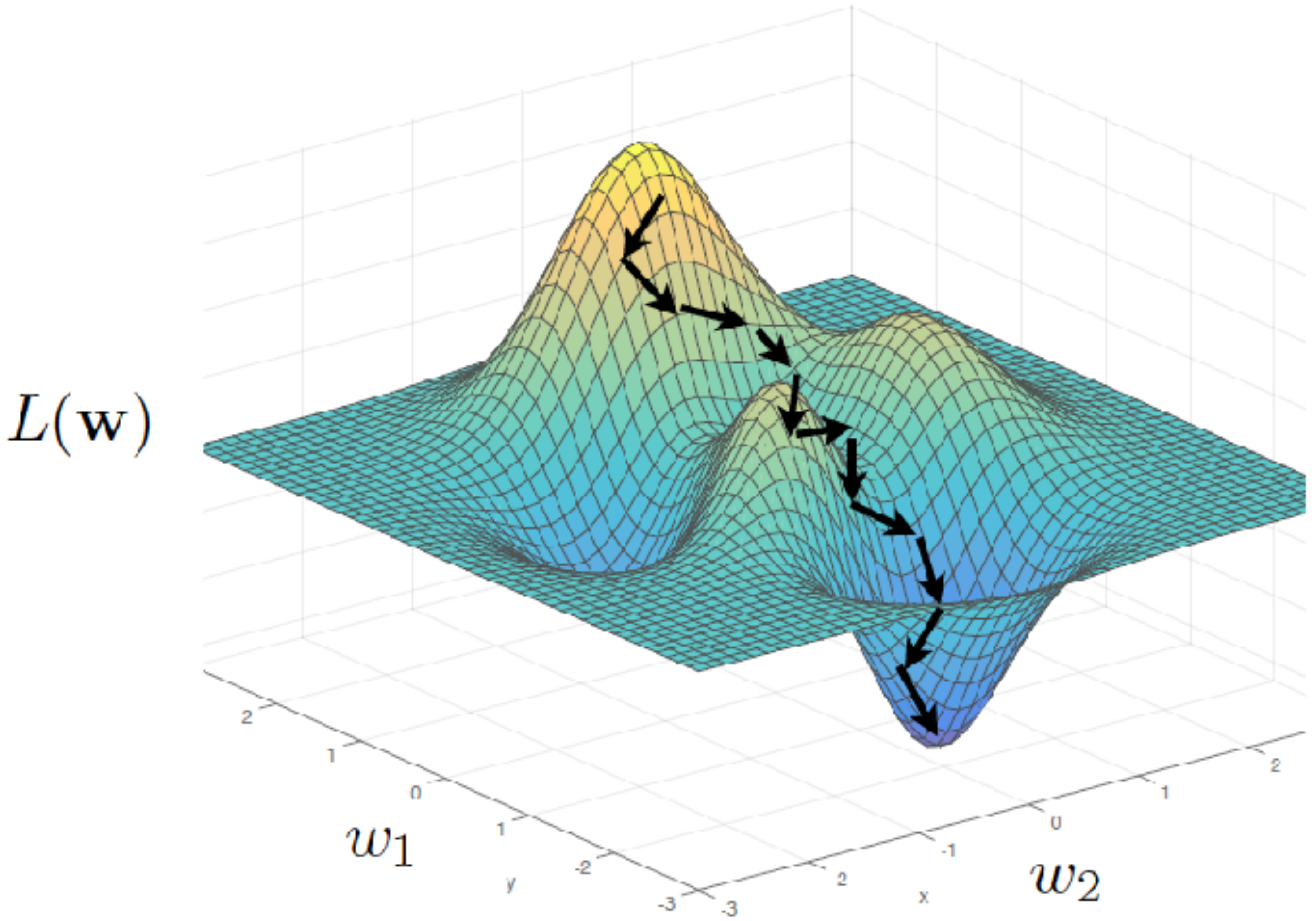
Deep nets architecture and SGD training



Rectified linear unit (ReLU)



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



CENTER FOR
Brains
Minds+
Machines

Gradient descent

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

One iteration of gradient descent:

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

learning rate

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?

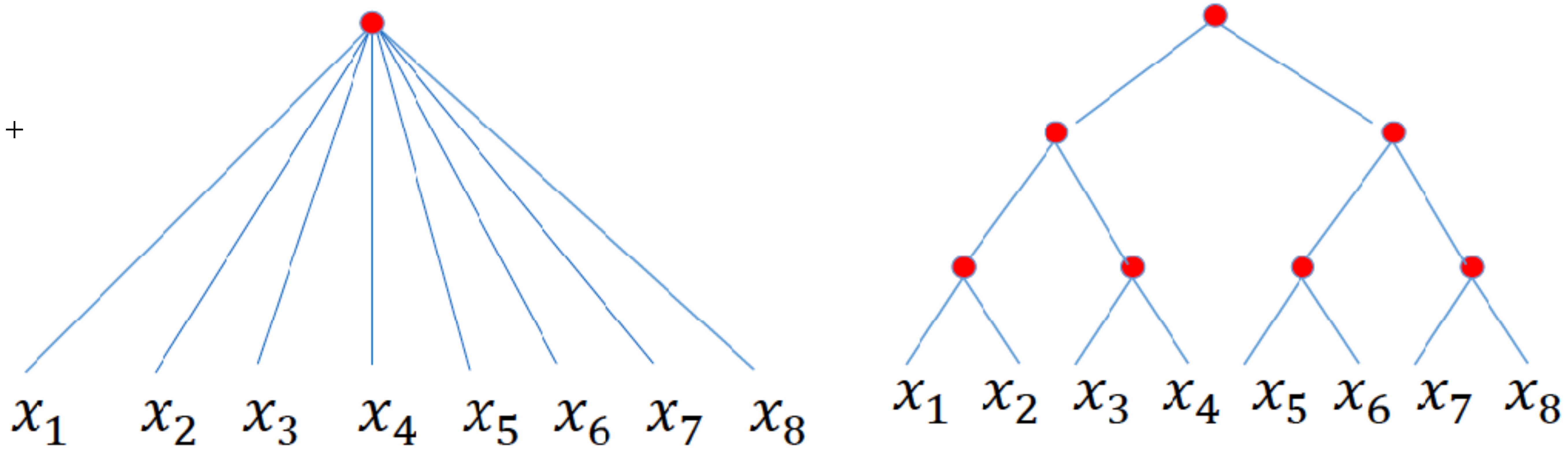
Generalization by SGD: how can overparametrized networks generalize?

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

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. $O(\varepsilon^{-2})$

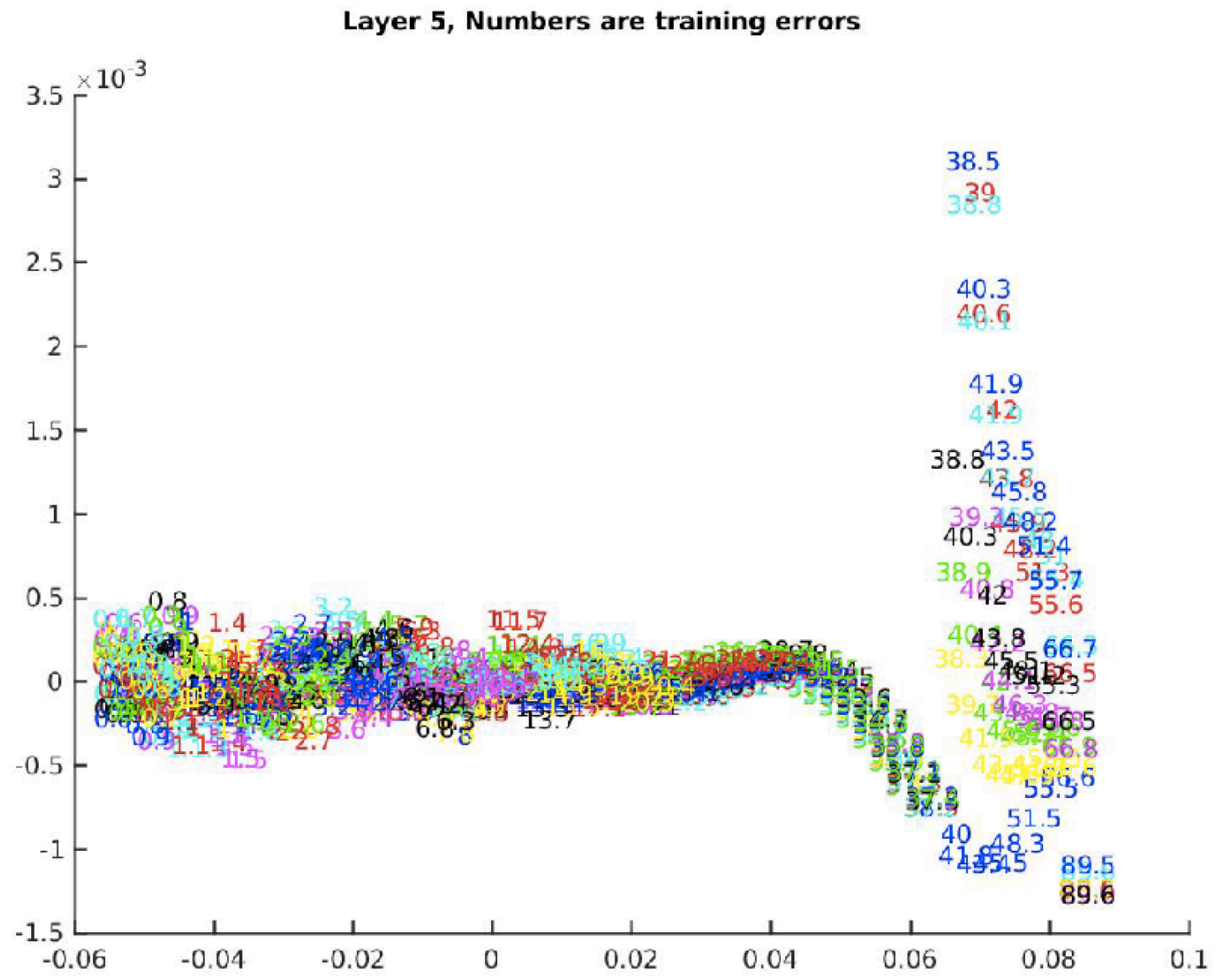
CENTER FOR
Brains
Minds+
Machines

Mhaskar, Poggio, Liao, 2016

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

Theorem (informal statement)

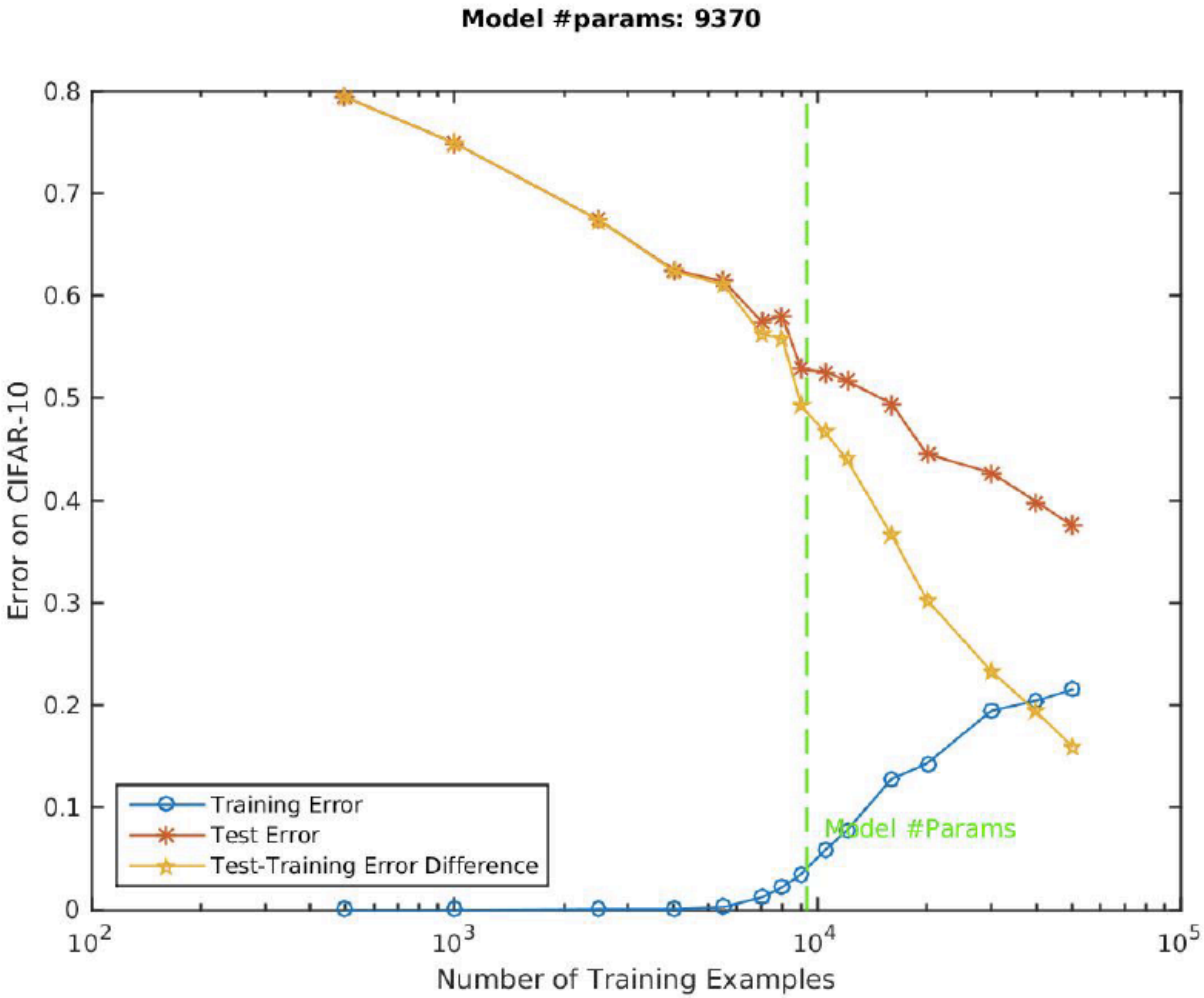
Replacing the RELUs with univariate polynomial approximation, Bezout theorem implies that the system of polynomial equations corresponding to zero empirical error has a very large number of degenerate solutions. The global zero-minimizers correspond to flat minima in many dimensions (generically unlike 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 underconstrained solutions generalize?



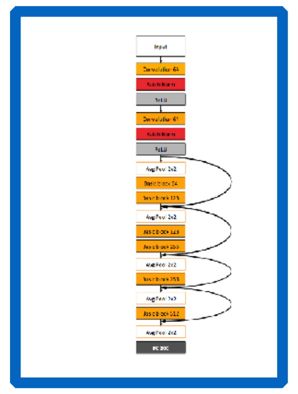
CENTER FOR
Brains
Minds+
Machines

Summary: Deep Learning, theory questions

- why depth works
- why optimization works so nicely
- why deep networks do not overfit and do generalize

Musings on Near Future Breakthroughs

- new architectures/class of applications from basic DCN block (example GAN + RL/DL + …)



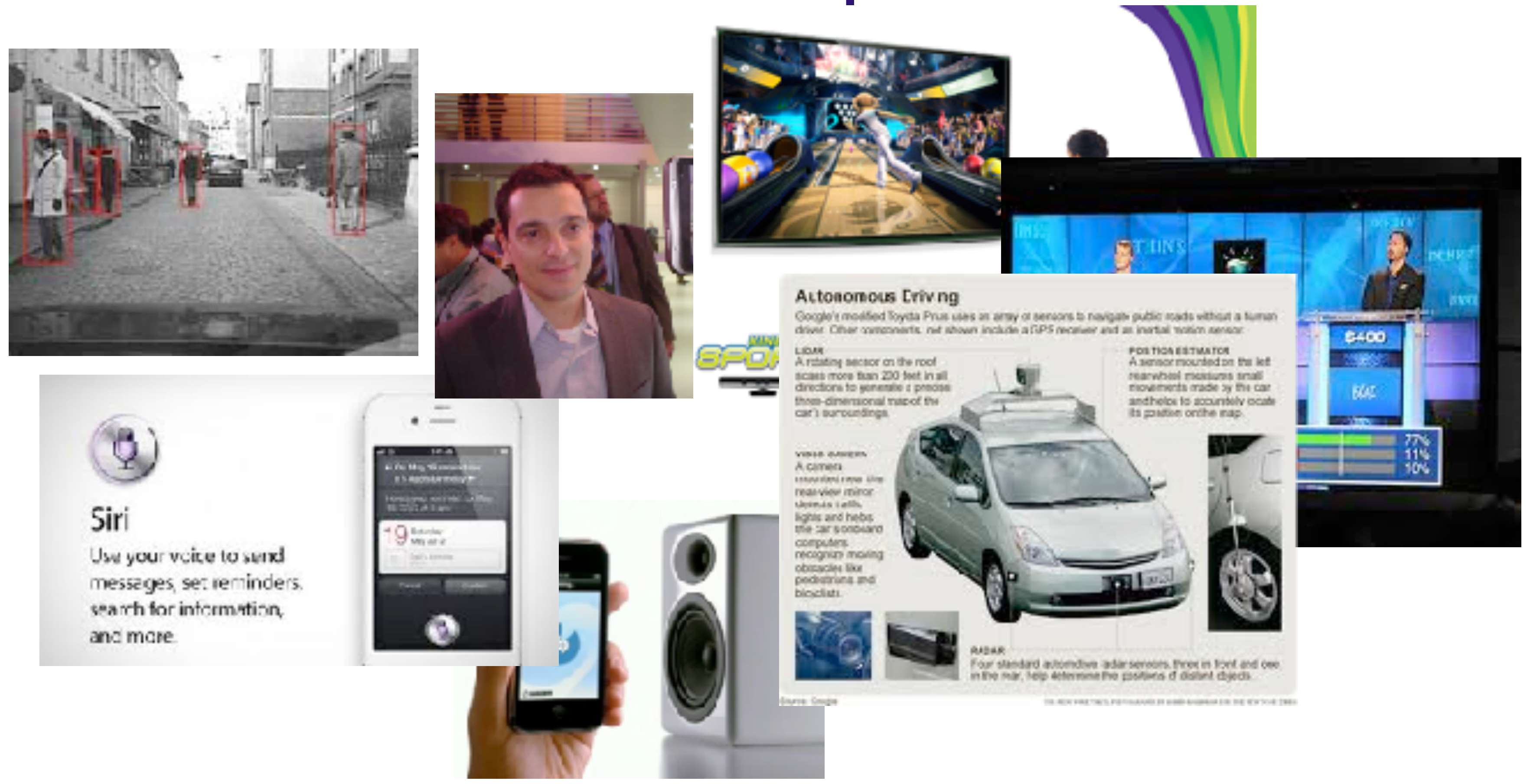
- new semisupervised training framework, avoiding labels: implicit labeling…predicting next “frame”…

- new basic supervised block/circuit $\boxed{?}$

- new learning algorithm (Shim) instead of SGD …

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: $n \to 1$

General musings

The evolution of computer science

- there were programmers
- there are now labelers
- there may be schools for bots…