9.520 in 2015

Statistical Learning Theory and Applications

Class Times:
Monday and Wednesday 1pm-2:30pm
Units: 3-0-9 H,G
Location:
46-5193
Instructors: Carlo Ciliberto, Georgios Evangelopoulos, Maximilian Nickel, Ben Deen, Hongyi Zhang, Steve Voinea, Owen Lewis,
T. Poggio, L. Rosasco

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

Office Hours:
Friday 2-3 pm in 46-5156, CBCL lounge (by appointment)

Email Contact :
9.520@mit.edu

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

Class 3 (Wed, Sept 16): Mathcamps

- Functional analysis (~45mins)

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

& Multivariate Calculus:
Extremal problems, differential, gradient.

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.

9.520: Statistical Learning Theory and Applications, Fall 2015

- Course focuses on regularization techniques, that provide a theoretical foundation to high- dimensional supervised learning.
- Support Vector Machines, manifold learning, sparsity, batch and online supervised learning, feature selection, structured prediction and multitask learning.
- Optimization theory critical for machine learning (first order methods, proximal/splitting techniques).
- In the final part focus on deep theory: deep learning networks, theory of invariance, extension of convolutional layers, learning invariance, connection of DCLNs with hierarchical splines, possibility of theory.

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

Class

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

Rules of the game:

• problem sets (2)
• final project: you have to give us title + abstract before November 25th
• participation
• Grading is based on Psets (27.5%+27.5%) + Final Project (32.5%) + Participation (12.5%)

Slides on the Web site (most classes on blackboard)
Staff mailing list is 9.520@mit.edu
Student list will be 9.520students@mit.edu
Please fill form!

send email to us if you want to be added to mailing list

Friday 2-3 pm in 46-5156,
CBCL lounge (by appointment)
Problem Set 1: 05 Oct (Class 8)
Problem Set 2: 09 Nov (Class 18)
Final Project Decision: 25 Nov (Class 22)

Final Project

The final project can be

- a Wikipedia entry or
- problems for chapters of the textbook of the class or
- contributions to GURLs (GURLS: a Toolbox for Regularized Least Squares Learning) or
- a research project.

For the Wikipedia article we suggest to post 1-2 pages (short) using Wikipedia standard format (of course).

For the research project (either Application or Theory) you should use the template on the Web site.

Project: posting/editing article on Wikipedia (past examples below)

- Kernel methods for vector output : http://en.wikipedia.org/wiki/Kernel_methods_for_vector_output
- Principal component regression : http://en.wikipedia.org/wiki/Principal_component_regression
- Reproducing kernel Hilbert space : http://en.wikipedia.org/wiki/Reproducing_kernel_Hilbert_space
- Proximal gradient methods for learning : http://en.wikipedia.org/wiki/Proximal_gradient_methods_for_learning
- Regularization by spectral filtering : https://en.wikipedia.org/wiki/Regularization_by_spectral_filtering
- Online learning and stochastic gradient descent : http://en.wikipedia.org/wiki/Online_machine_learning
- Kernel embedding of distributions : http://en.wikipedia.org/wiki/Kernel_embedding_of_distributions
- Vapnik–Chervonenkis theory : https://en.wikipedia.org/wiki/VC_theory
- Deep learning : http://en.wikipedia.org/wiki/Deep_learning
- Early stopping and regularization : http://en.wikipedia.org/wiki/Early_stopping
- Statistical learning theory : http://en.wikipedia.org/wiki/Statistical_learning_theory
- Representer theorem : http://en.wikipedia.org/wiki/Representer_theorem
- Regularization perspectives on support vector machines : http://en.wikipedia.org/wiki/Regularization_perspectives_on_support_vector_machines
- Semisupervised learning : http://en.wikipedia.org/wiki/Semi_supervised_learning
- Bayesian interpretation of regularization :

- Statistical learning theory : http://en.wikipedia.org/wiki/Statistical_learning_theory
- Representer theorem : http://en.wikipedia.org/wiki/Representer_theorem
- Regularization perspectives on support vector machines : http://en.wikipedia.org/wiki/Regularization_perspectives_on_support_vector_machines
- Semisupervised learning : http://en.wikipedia.org/wiki/Semi_supervised_learning
- Bayesian interpretation of regularization : http://en.wikipedia.org/wiki/Bayesian_interpretation_of_regularization
- Regularized least squares (RLS) : http://en.wikipedia.org/wiki/User:Bdeen/sandbox
- Occam Learning (PAC Learning) : https://en.wikipedia.org/wiki/Occam_learning
- Multiple Kernel Learning: https://en.wikipedia.org/wiki/Multiple_kernel_learning
- Loss Function for Classification : https://en.wikipedia.org/wiki/Loss_functions_for_classification
- Online Machine Learning : https://en.wikipedia.org/wiki/Online_machine_learning
- Sparse PCA : https://en.wikipedia.org/wiki/Sparse_PCA
- Distribution Learning Theory : https://en.wikipedia.org/wiki/Distribution_learning_theory
- Sample Complexity : https://en.wikipedia.org/wiki/Sample_complexity
- Hyper Basis Function Network : https://en.wikipedia.org/wiki/Hyper_basis_function_network
- Diffusion Map : https://en.wikipedia.org/wiki/Diffusion_map
- Matrix Regularization: https://en.wikipedia.org/wiki/Matrix_regularization
- Mtheory (Learning Framework) : https://en.wikipedia.org/wiki/MTheory_(learning_framework)
- Feature Learning : https://en.wikipedia.org/wiki/Feature_learning

Done but not submitted in (public) Wikipedia

====================================

- Lasso Regression : https://en.wikipedia.org/wiki/User:Rezamohammadighazi/sandbox
- Unsupervised Learning: Dim. Red. : https://en.wikipedia.org/wiki/User:Iloverobotics/sandbox
- Regularized Least Squares : https://en.wikipedia.org/wiki/User:Yakirrr
- Error Tolerance (PAC Learning): https://en.wikipedia.org/wiki/User:Alex_e_e_alex/sandbox

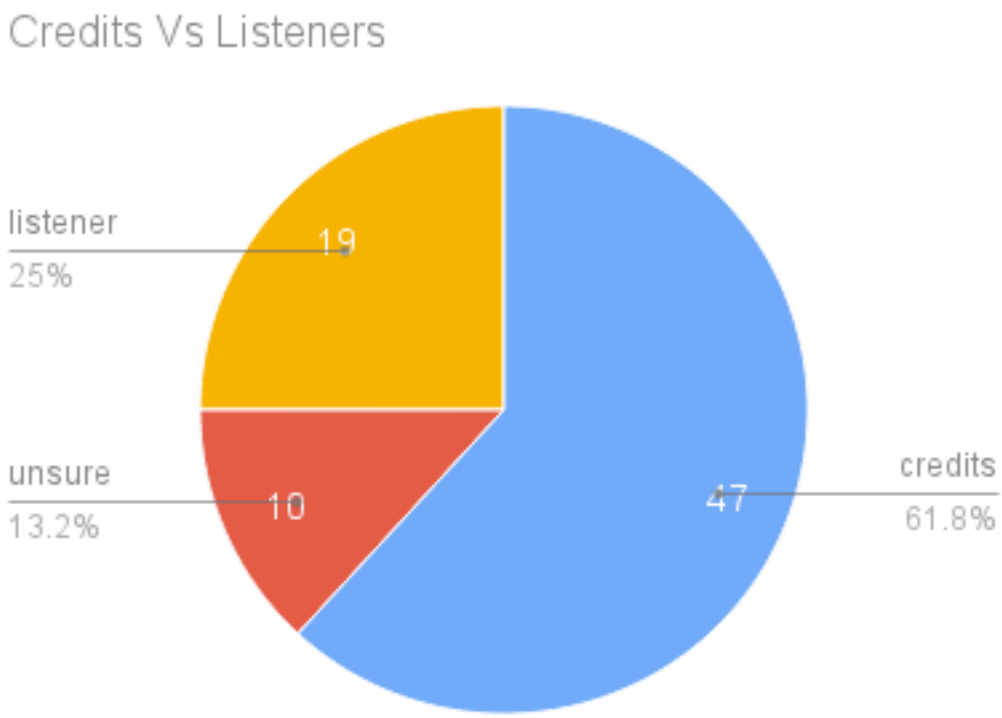
Done but not submitted in (public) Wikipedia
====================================
- Lasso Regression : https://en.wikipedia.org/wiki/User:Rezamohammadighazi/sandbox
- Unsupervised Learning: Dim. Red. : https://en.wikipedia.org/wiki/User:Iloverobotics/sandbox
- Regularized Least Squares : https://en.wikipedia.org/wiki/User:Yakirrr
- Error Tolerance (PAC Learning): https://en.wikipedia.org/wiki/User:Alex_e_e_alex/sandbox
- Desnity Estimation : https://en.wikipedia.org/wiki/User:Linjing1119/sandbox
- Matrix Completion : https://en.wikipedia.org/wiki/User:Milanambiar/sandbox
- Multiple Instance Learning : we have Wiki markup
- Uniform Stability and Generalization in Learning Theory : https://en.wikipedia.org/wiki/Draft:Uniform_Stability_and_Generalization_in_learning_theory
- Generalization Error: https://en.wikipedia.org/wiki/User:Agkonings/sandbox
- Tensor Completion : https://en.wikipedia.org/wiki/User:Aali9520/Tensor_Completion
- Structured Sparsity Regularization : https://en.wikipedia.org/wiki/User:A.n.campero/sandbox
- Proximal Operator for Matrix Function : https://en.wikipedia.org/wiki/User:Lovebeloved/sandbox
- Sparse Dictionary Learning : we have pdf
- PAC Learning : https://en.wikipedia.org/wiki/User:Scott.linderman/sandbox
- Convolutional Neural Networks : https://en.wikipedia.org/wiki/User:Wfwhitney/sandbox
- Frames/Basis Functions: https://en.wikipedia.org/wiki/Frame_(linear_algebra)

Class

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

- The pace is fast on purpose...
- Big picture will be provided today and repeated at the end of the course...
- Be ready for a lot of material: this is MIT.
- If you need a refreshment in Fourier analysis you should not be in this class.
- We do not compare the approach in this class to others -- such as Bayesian one -- because we do not like to complain too much about others.

9.520 in 2015


Credits Vs Listeners
listener
19
25%
unsure
10
13.2%
credits
47
61.8%

Summary of today's overview

- Motivations for this course: a golden age for new AI (and the key role of Machine Learning)
- Statistical Learning Theory
- Success stories from past research in Machine Learning: examples of engineering applications
- In this machine learning class: computer science and neuroscience, developing a theory for deep learning.

Summary of today's overview

- Motivations for this course: a golden age for new AI (and the key role of Machine Learning)

- Statistical Learning Theory

- Success stories from past research in Machine Learning: examples of engineering applications

- A new phase in machine learning: computer science and neuroscience, learning and the brain, CBMM:

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

The problem of 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

These advances will be critical to of our society's

- future prosperity
- education, health, security

The Center for Brains, Minds and Machines

MIT

Boyden, Desimone ,Kaelbling , Kanwisher, Katz, Poggio, Sassanfar, Saxe, Schulz, Tenenbaum, Ullman, Wilson, Rosasco, Winston

Harvard

Blum, Kreiman, Mahadevan, Nakayama, Sompolinsky, Spelke, Valiant

Rockefeller
Freiwald

Allen Institute
Koch

UCLA
Yuille

Stanford
Goodman

Cornell
Hirsh

Hunter
Epstein,Sakas, Chodorow

Wellesley
Hildreth, Conway, Wiest

Puerto Rico
Bykhovaskaia, Ordonez, Arce Nazario

Howard
Manaye, Chouikha, Rwebargira

Industrial partners

IIT
Metta, Rosasco, Sandini

A*star
Tan

Hebrew U.
Shashua

MPI
Bueltheff

Genoa U.
Verri

Weizmann
Ullman

City U. HK
Smale

NCBS
Raghavan

香港城市大學 City University of Hong Kong

iit ISTITUTO ITALIANO DI TECNOLOGIA

ncbs

UNIVERSITÀ DEGLI STUDI DI GENOVA

מכון ויצמן למדע WEIZMANN INSTITUTE OF SCIENCE

Google
Norvig

IBM
Lemnios

Microsoft
Blake

Siemens

Schlumberger

GE

DeepMind
Hassabis

Boston Dynamics
Raibert

Orcam
Shashua

Rethink Robotics
Brooks

MobilEye
Shashua

At the core of the problem of Intelligence
is the problem of Learning

Learning is the gateway to understanding the brain and to making intelligent machines.

Problem of learning:
a focus for

- math
- computer algorithms
- neuroscience

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.

Class

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

- The pace is fast on purpose, otherwise we get too bored.
- Big picture will be provided today and repeated at the end of the course. Listen carefully.
- Be ready for a lot of material: this is MIT.
- If you think that the course is disorganized, it means you have not really understood it.a
- I am passionate about ML and I will show it today. If you think Lorenzo is not, complain to him, not to me!
- Notation is kept inconsistent throughout the course on purpose to train you to read and understand different papers with different notations.
- If you need a refreshment in Fourier analysis you should not be in this class.
- We do not compare the approach in this class to others -- such as Bayesian one -- because we do not like to complain too much about others.

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

- Classes 2-9 are the core: foundations + regularization
- Classes 10-20 are state-of-the-art topics for research in — and applications of — ML
- Classes 21-26 are mostly new, about multilayer networks (DCLNs)

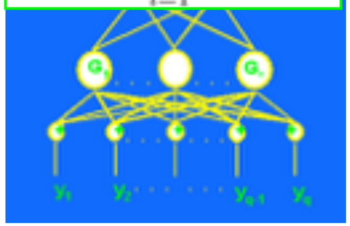
Summary of today's overview

- Motivations for this course: a golden age for new AI and the key role of Machine Learning
- Statistical Learning Theory
- Success stories from past research in Machine Learning: examples of engineering applications
- A new phase in machine learning: computer science and neuroscience, learning and the brain, CBMM:

Learning:
Math, Engineering, Neuroscience

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

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



LEARNING THEORY + ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

ENGINEERING APPLICATIONS

- Bioinformatics
- Computer vision
- Computer graphics, speech synthesis, creating a virtual actor

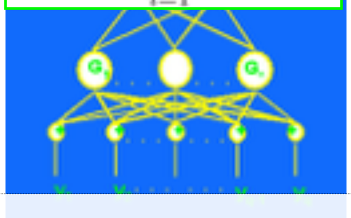
COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

Statistical Learning Theory

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

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



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

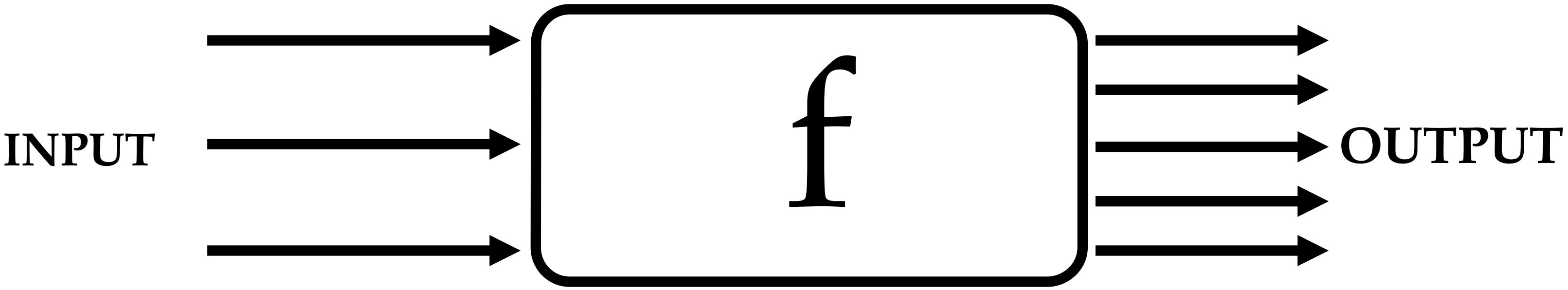
ENGINEERING APPLICATIONS

- Bioinformatics
- Computer vision
- Computer graphics, speech synthesis, creating a virtual actor

COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

Statistical Learning Theory: **supervised** learning



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

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

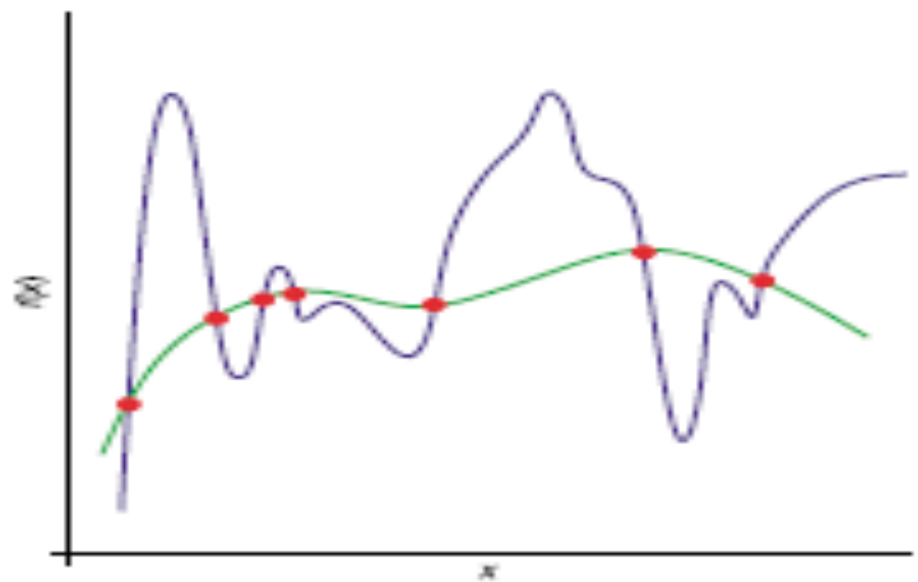
is a **good predictor** of y for a **future** input x (fitting the data is **not** enough!)

Statistical Learning Theory: prediction, not curve fitting

= data from f

= function f

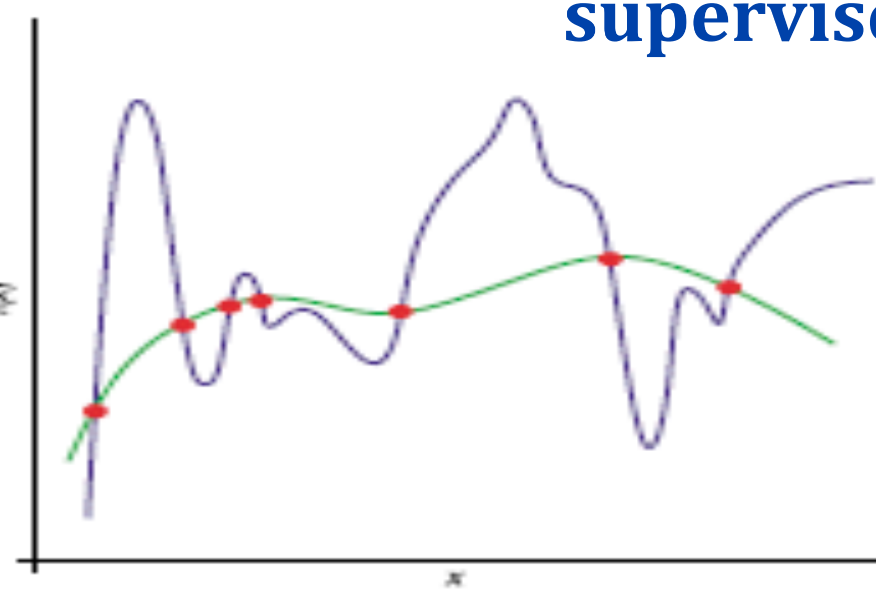
= approximation of f



Generalization:

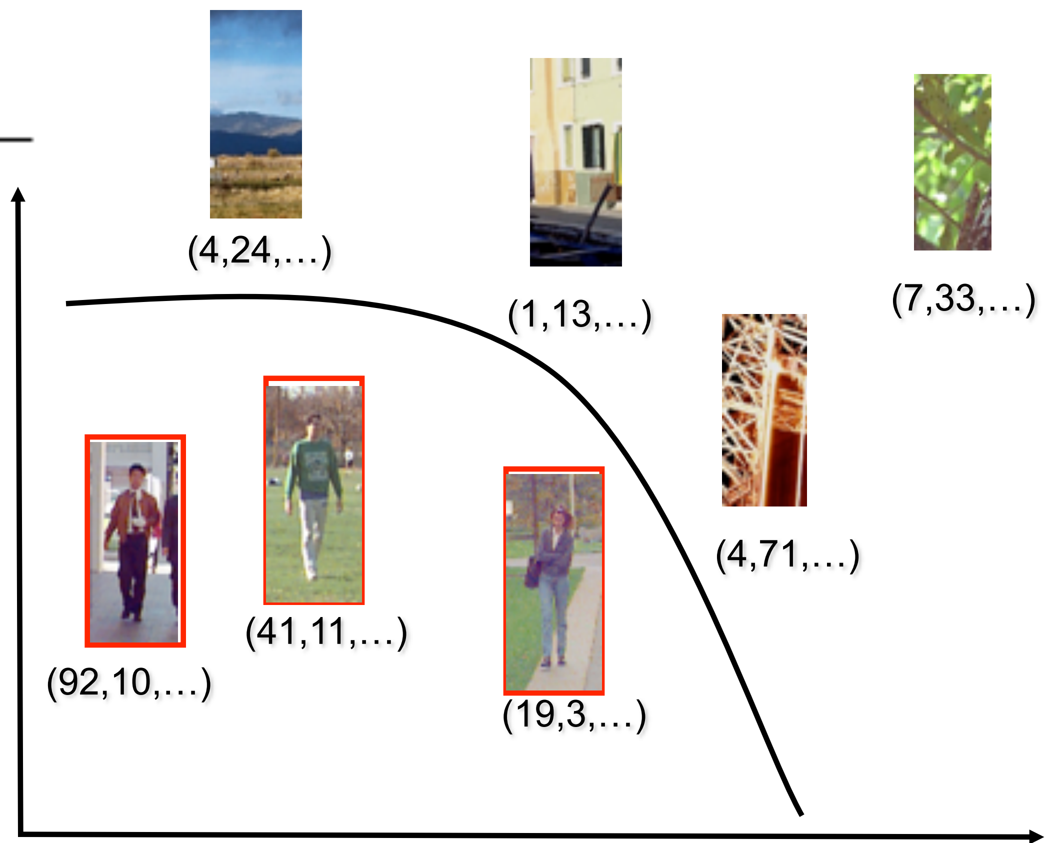
estimating value of function where there are no data (good generalization means predicting the function well; important is for empirical or validation error to be a good proxy of the prediction error)

Statistical Learning Theory: supervised learning



Regression

Classification



Statistical Learning Theory: part of mainstream math not just statistics (Valiant, Vapnik, Smale, Devore...)

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 39, Number 1, Pages 1–49
S 0273-0979(01)00923-5
Article electronically published on October 5, 2001

ON THE MATHEMATICAL FOUNDATIONS OF LEARNING

FELIPE CUCKER AND STEVE SMALE

The problem of learning is arguably at the very core of the problem of intelligence, both bi

T. Poggio and C.R. Shelton

Introduction

(1) A main theme of this report is the relationship of approximation to learning and the primary role of sampling (inductive inference). We try to emphasize relations of the theory of learning to the mainstream of mathematics. In particular, there are large roles for probability theory, for algorithms such as *least squares*, and for tools and ideas from linear algebra and linear analysis. An advantage of doing this is that communication is facilitated and the power of core mathematics is more easily brought to bear.

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: the learning problem should be well-posed

J. S. Hadamard, 1865-1963

A problem is well-posed if its solution

exists, unique and

is stable, eg depends continuously on the data (here examples)

Statistical Learning Theory: theorems extending foundations of learning theory

Conditions for <u>generalization</u> in learning theory

have deep, almost philosophical, implications:

they can be regarded as equivalent conditions that guarantee a theory to be predictive (that is scientific)

- theory must be chosen from a small set
- theory should not change much with new data...most of the time

A classical algorithm in Statistical Learning Theory:

Kernel Machines eg Regularization in RKHS

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

implies

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

Equation includes splines, Radial Basis Functions and SVMs (depending on choice of K and V).

For a review, see Poggio and Smale, 2003; see also Schoelkopf and Smola, 2002; Bousquet, O., S. Boucheron and G. Lugosi; Cucker and Smale; Zhou and Smale...

Statistical Learning Theory: classical algorithms: Regularization

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

has a Bayesian interpretation:

data term is a model of the noise and the stabilizer is a prior on the hypothesis space of functions f. That is, Bayes rule

$$\mathcal{P}[f|D_\ell] = \frac{\mathcal{P}[D_\ell|f] \; \mathcal{P}[f]}{P(D_\ell)}$$

leads to

$$\mathcal{P}[f|D_\ell] = \frac{1}{Z_D Z_L Z_r} e^{-\left(\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}(y_i - f(\mathrm{x}_i))^2 + \|f\|_K^2\right)}$$

Statistical Learning Theory: classical algorithms: Regularization

Classical learning algorithms: Kernel Machines (eg Regularization in RKHS)

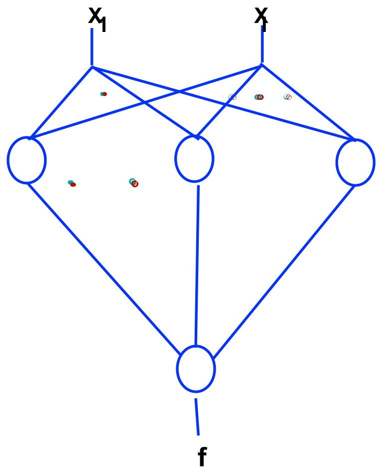
$$\min_{f \in H} \left[\frac{1}{n} \sum_{i=1}^{n} V(f(x_i) - y_i) + \lambda \; \|f\|_K^2 \right]$$

implies

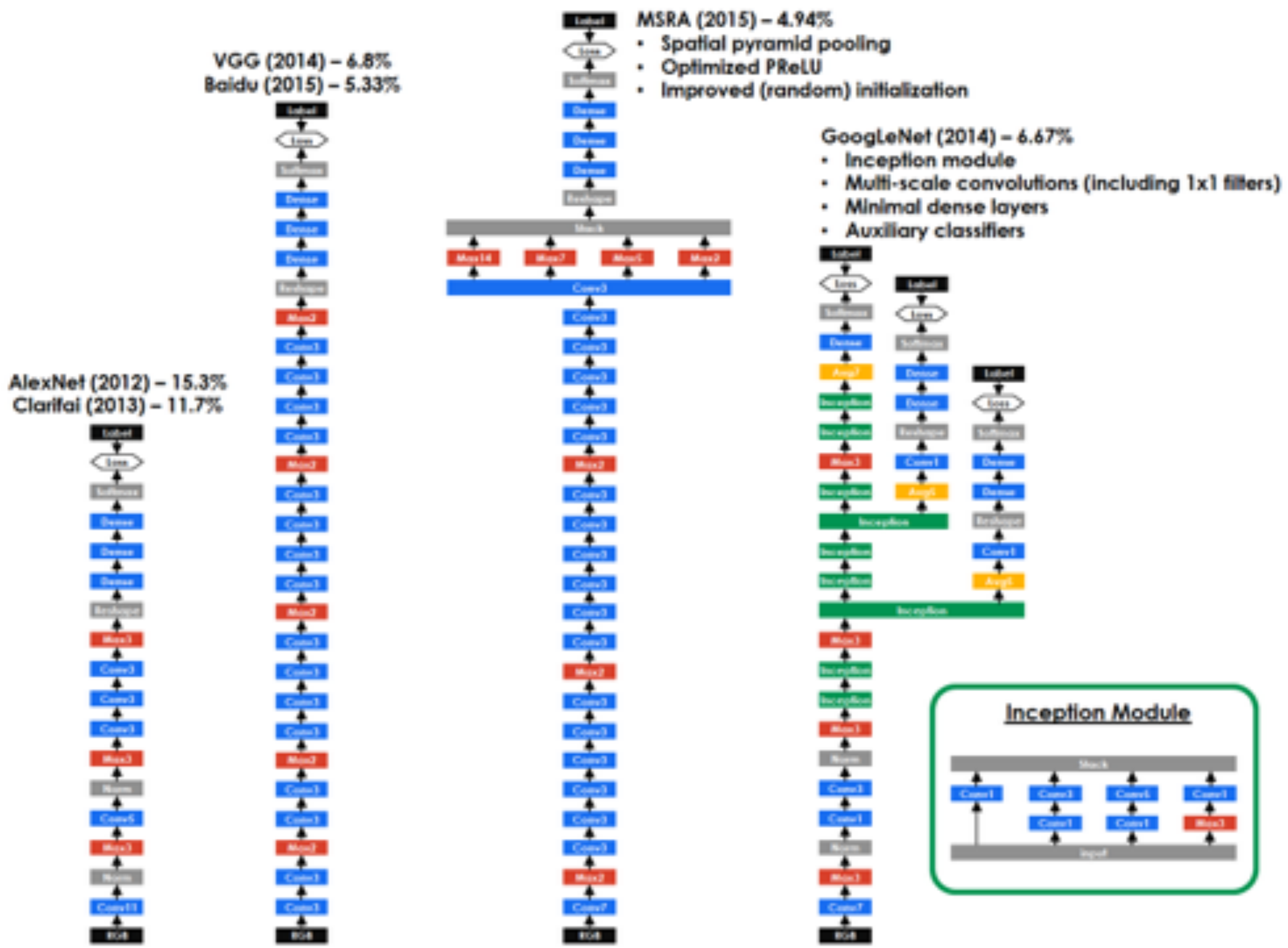
$$f(\mathbf{x}) = \sum_i^n \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

Remark (for later use):

Classical kernel machines correspond to shallow networks



A present challenge: a theory for Deep Learning



Statistical Learning Theory: note

Two connected and overlapping strands in learning theory:

- ❑ Bayes, hierarchical models, graphical models…
- ❑ Statistical learning theory, regularization

Summary of today's overview

- Motivations for this course: a golden age for new AI and the key role of Machine Learning
- Statistical Learning Theory
- Success stories from past research in Machine Learning: examples of engineering applications
- A new phase in machine learning: computer science and neuroscience, learning and the brain, CBMM:

Supervised learning

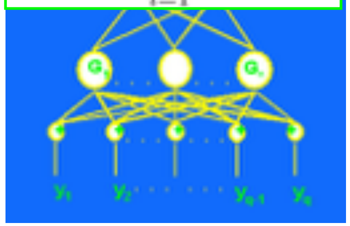
Since the introduction of supervised learning techniques 20 years ago, AI has made significant (and not well known) advances in a few domains:

- *Vision*
- *Graphics and morphing*
- *Natural Language/Knowledge retrieval (Watson and Jeopardy)*
- *Speech recognition (Nuance, Microsoft, Google)*
- *Games (Go, chess, Atari games…)*
- *Semiautonomous driving*

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

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

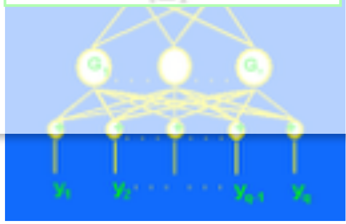
COMPUTATIONAL NEUROSCIENCE: models+experiments

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

Sung & Poggio 1995

COMPUTATIONAL NEUROSCIENCE: models+experiments

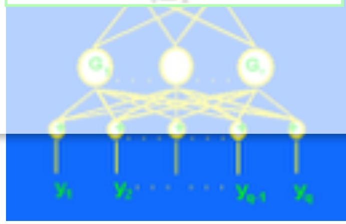
How visual cortex works


Image
Output
13

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

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

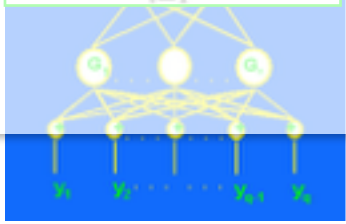
COMPUTATIONAL NEUROSCIENCE:
models+experiments

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

People detection

Papageorgiou&Poggio, 1997, 2000
also Kanade&Scheiderman

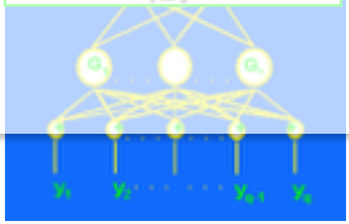
COMPUTATIONAL NEUROSCIENCE:
models+experiments

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

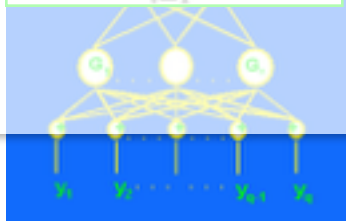
COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

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 and car detection are also "solved" (commercial systems, *MobilEye*)

COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

Recent progress in AI
and
machine learning

Why now: recent progress in AI


PIENSE
THINK
THINK
$3,400
$1,200


NAVY
505

Why now: very recent progress in AI

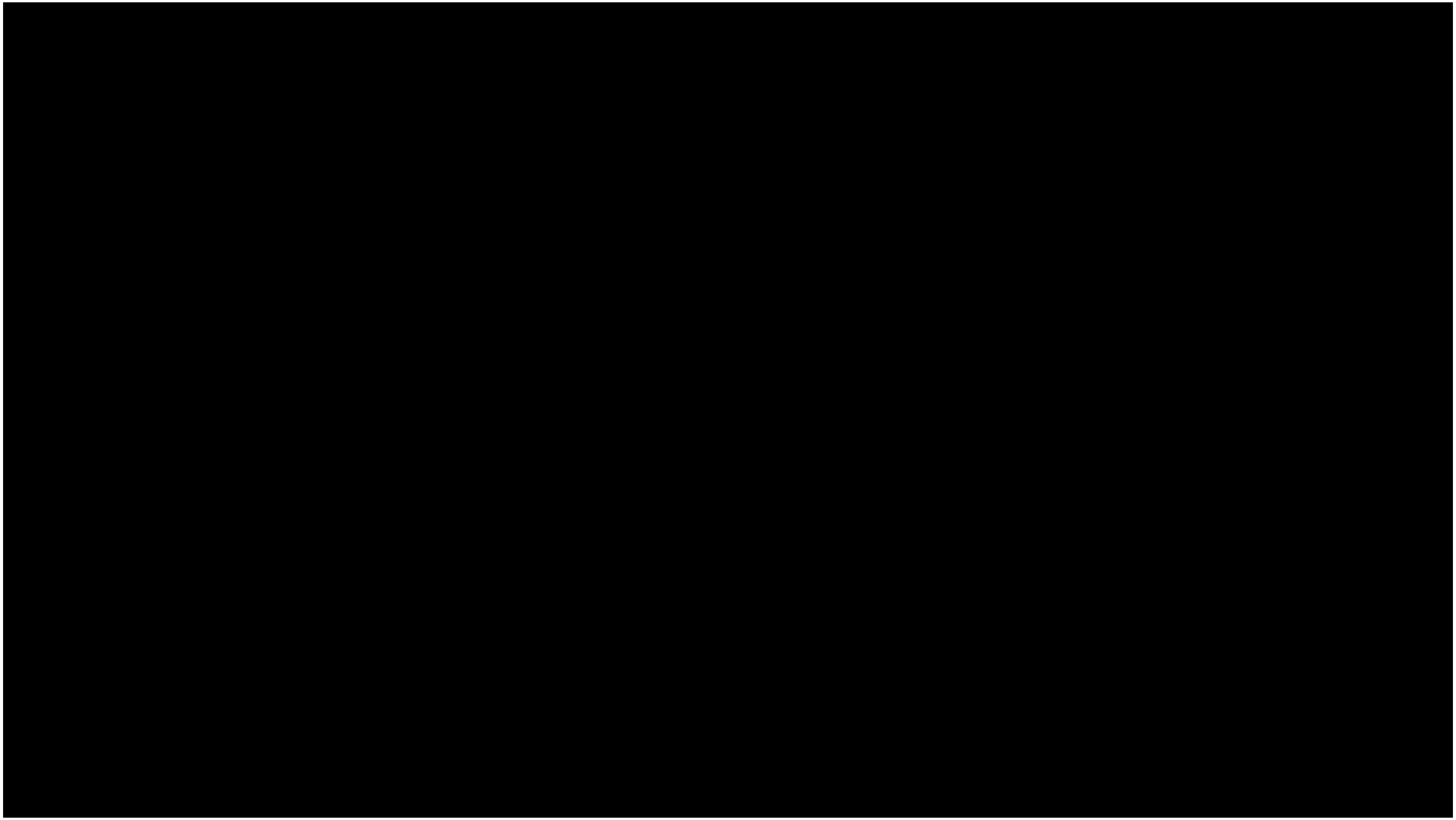

nature INSIGHT
MACHINE
INTELLIGENCE

Pedestrian accidents occur every day
in our increasingly intensive traffic environment.

DIVX VIDEO

Center for Brains,
Minds & Machines

Why now: very recent progress in AI



Center for Brains, Minds & Machines

CBMM Summer School Schedule

August 13th through September 2nd
Organized by the Center for Brains, Minds, and Machines
At the Marine Biology Lab at Woods Hole

		Morning 9-12	Afternoon 1:30-5:30	Evening 8-9	
Th	13		Reception - 5PM - Swope		
F	14	Introduction	Student introductions	Project introductions	Social
Sa	15	Linear algebra, probability	Neuroscience, programming	Project discussion	
Su	16	iCub, Google Glass		Dinner - 6:30 - Swope	
M	17	Computational neuroscience/ Propagation of sensory representations in cortex-like deep architectures Gabriel Kreiman/Haim Sompolinsky	Biological and computer vision Jim DiCarlo	Larry Abbott	Reception
Tu	18	Cognitive Neuroscience and Face Recognition Winrich Freiwald, Nancy Kanwisher	Computer vision, deep learning Andrei Barbu	Tom Mitchell	Reception
W	19	Machine learning Lorenzo Rosasco	Machine learning Lorenzo Rosasco	Demis Hassabis	Reception
Th*	20	Surya Ganguli	Robotics Afternoon		
F	21	Computational cogsci Josh Tenenbaum, Tomer Ullman	Church Tomer Ullman		
Sa	22				
Su	23	Martha's Vineyard trip			
M	24	Memory Matt Wilson, Aude Oliva	AI / Vision Shimon Ullman	Eero Simoncelli	Reception
Tu	25	Development I Liz Spelke, Alia Martin	Psychophysics and mTurk Leyla Isik, Tomer Ullman	Dorin Comaniciu	Reception
W*	26	Social perception Rebecca Saxe, Ken Nakayama	Neural data analysis Ethan Meyers		
Th	27	Development II Laura Schulz, Tomer Ullman	Invariance, inverse problems Tomaso Poggio, Mahadevan	Jessica Sommerville	Reception
F*	28	AI / Language Patrick Winston, Boris Katz	AI / Vision Andrei Barbu		
Sa	29	Audition and speech Josh McDermott, Hynek Hermansky	Audition/vision panel J. McDermott, H. Hermansky, D. Yamins		
Su*	30			Dinner - 6:30 - Swope	
M	31			Amnon Shashua	Reception
Tu	1				
W	2	Student presentations		Closing reception - 7PM	

talk	social	panel
project	tutorial	

*Starred days will feature a journal club.

Talks (red) are in Lillie Audtiorium.
Tutorials and Projects (orange, green) are in Loeb 306.

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

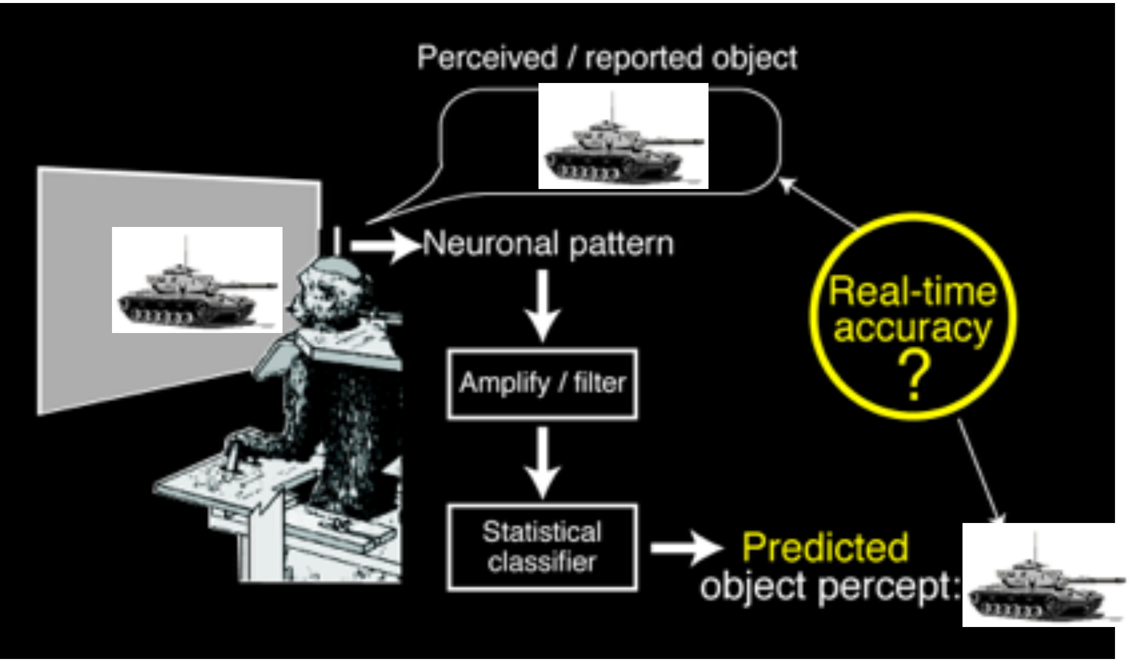
Text Classification

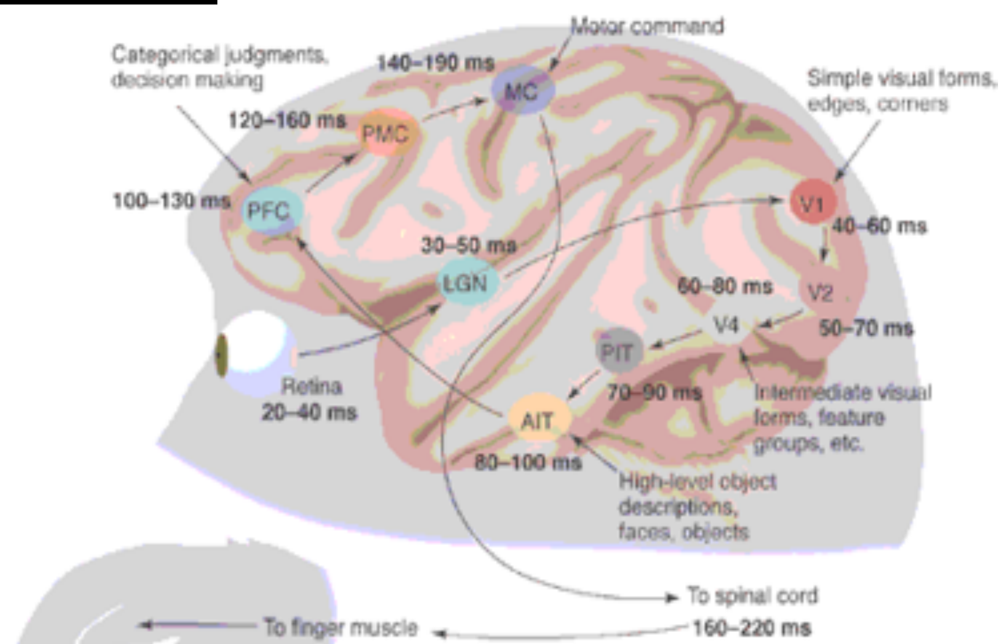
Artificial Markets

Stock option pricing

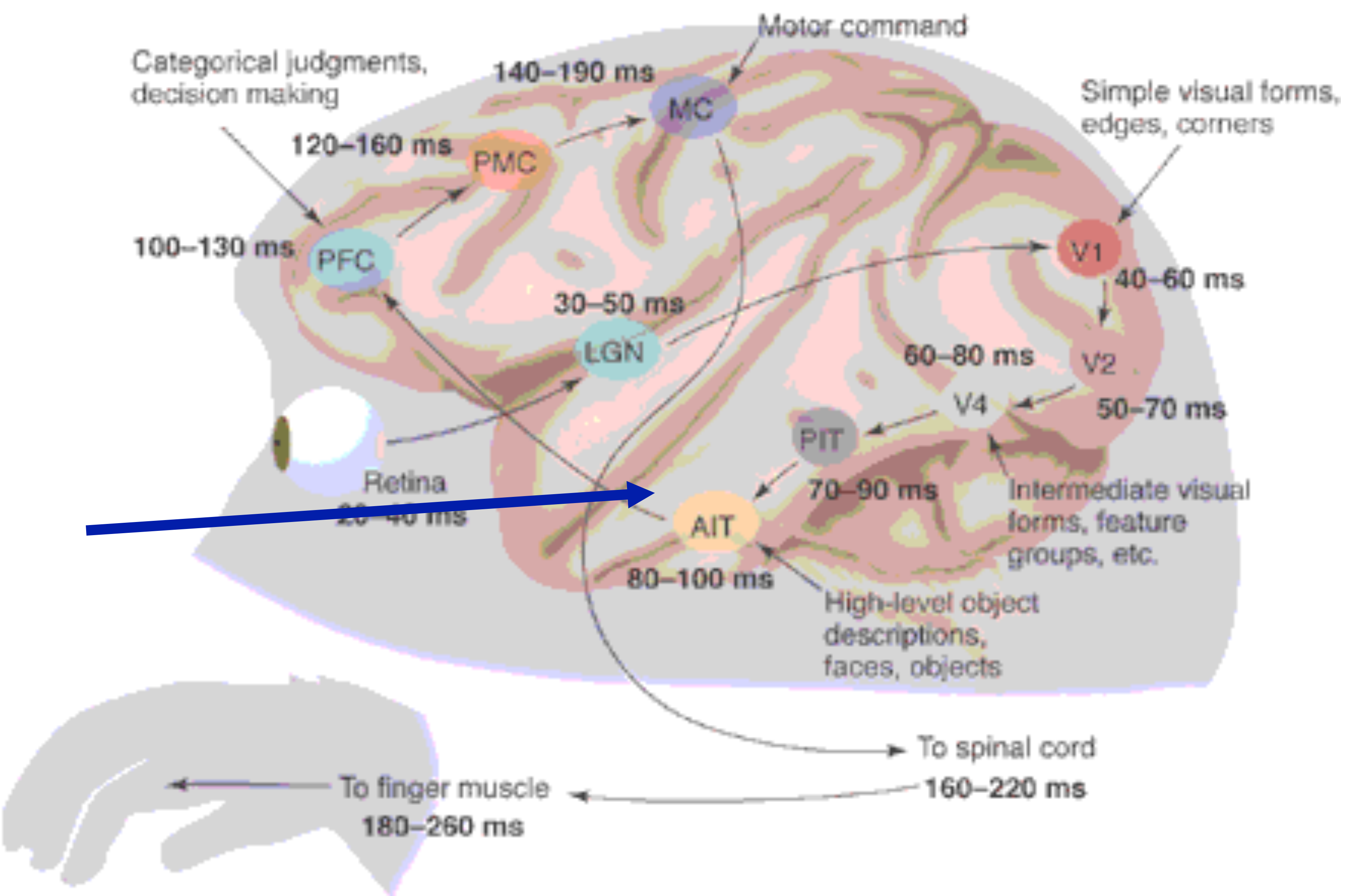
….

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



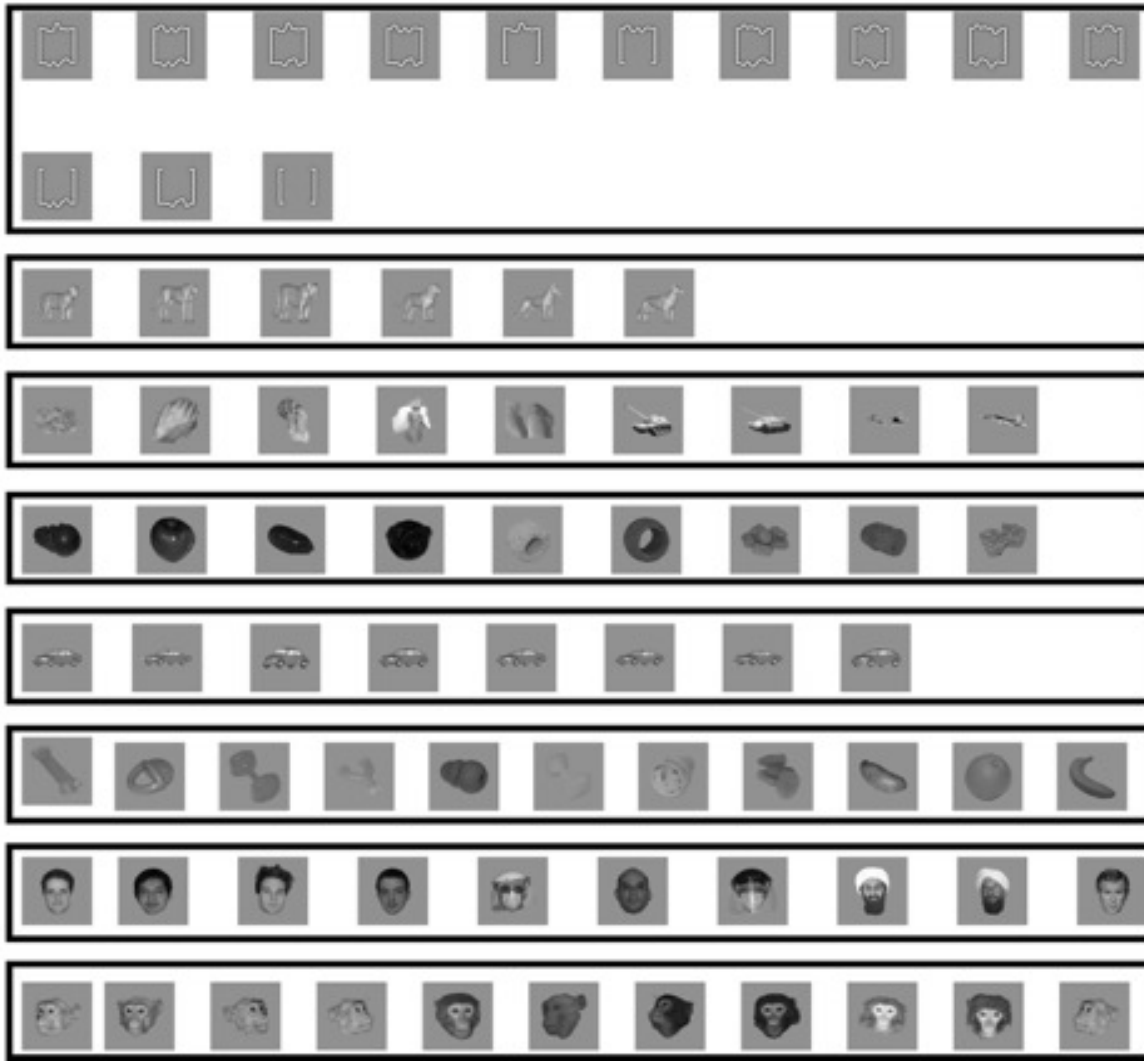


The end station of the ventral stream in visual cortex is IT



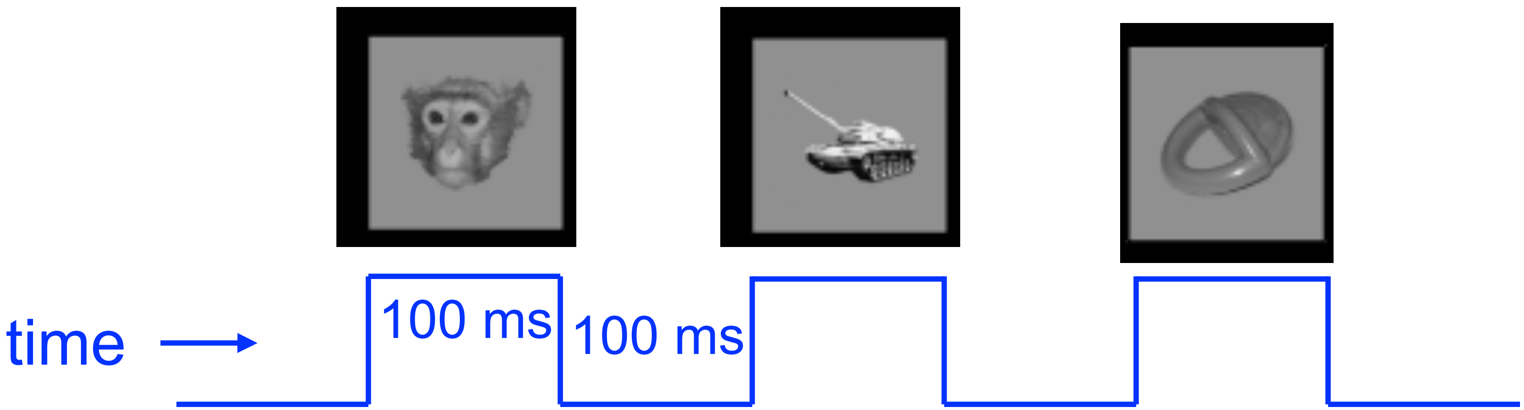
Reading-out the neural code in AIT

77 objects, 8 classes



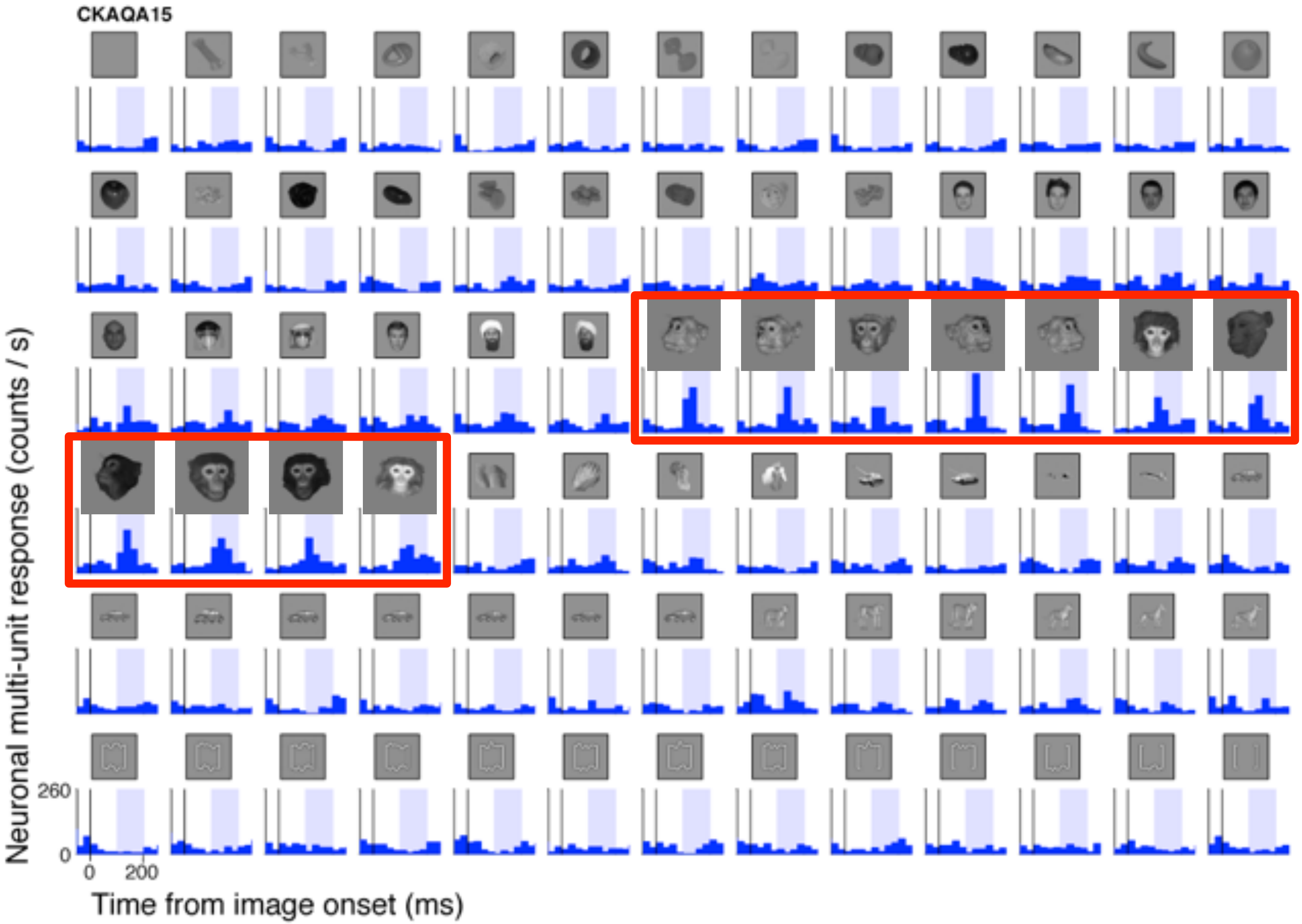
Chou Hung, Gabriel Kreiman, James DiCarlo, Tomaso Poggio, Science, Nov 4, 2005

Recording at each recording site during passive viewing

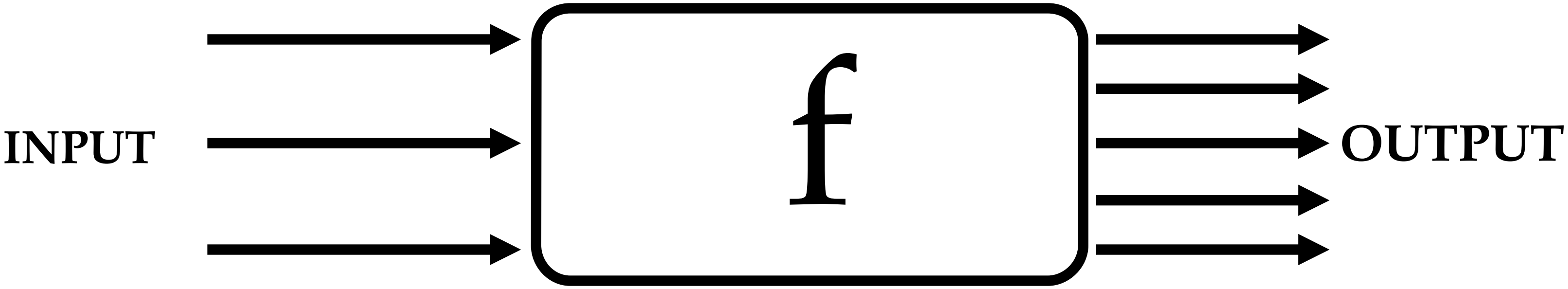


- 77 visual objects
- 10 presentation repetitions per object
- presentation order randomized and counter-balanced

Example of one AIT cell



Learning: read-out from the brain



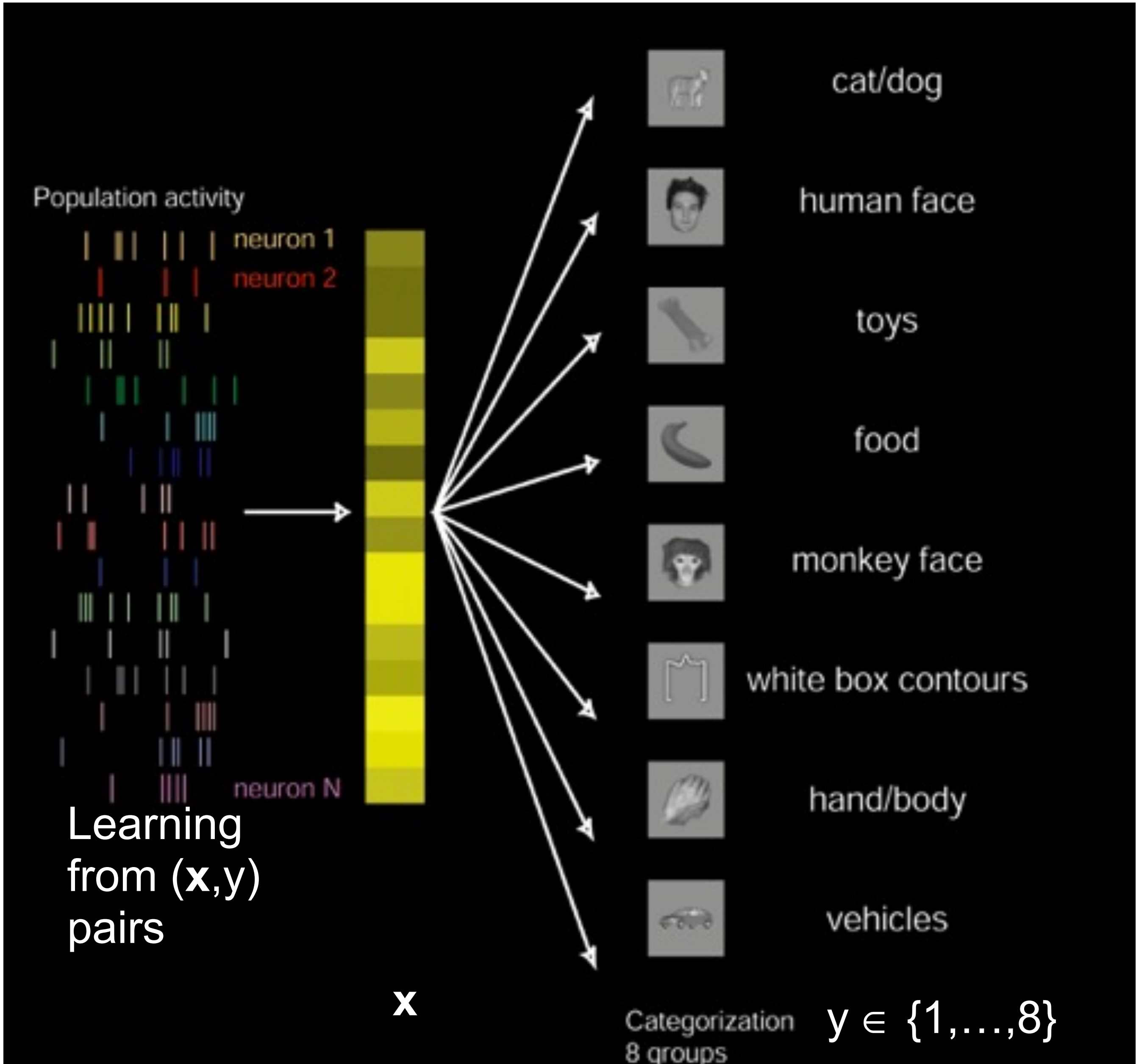
From a set of data (vectors of activity of n neurons (x) and object label (y)

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

Find (by training) a classifier eg a function f such that $f(x) = \hat{y}$

is a **good predictor** of object label y for a **future** neuronal activity x

Decoding the neural code … using a classifier

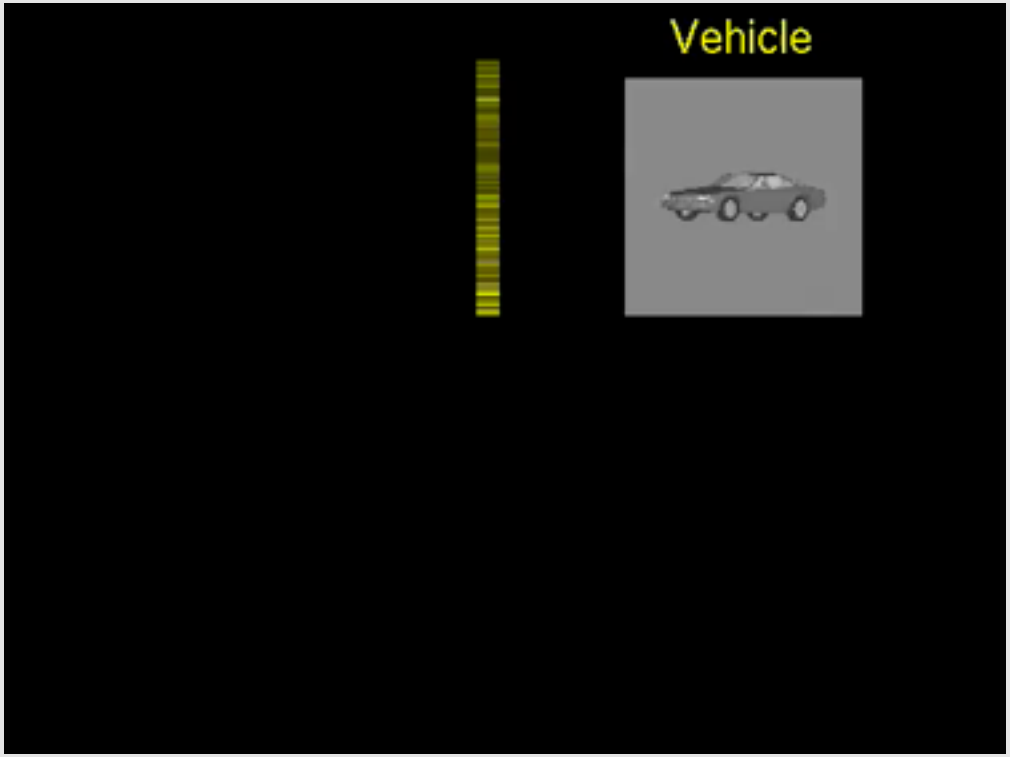


We can decode the brain's code and read-out from neuronal populations:

reliable object categorization (>90% correct) using ~200 <u>arbitrary</u> AIT "neurons"

Video speed: 1 frame/sec

Actual presentation rate: 5 objects/sec



<u>Categorization</u>

Toy

Body

Human Face

Monkey Face

Vehicle

Food

Box

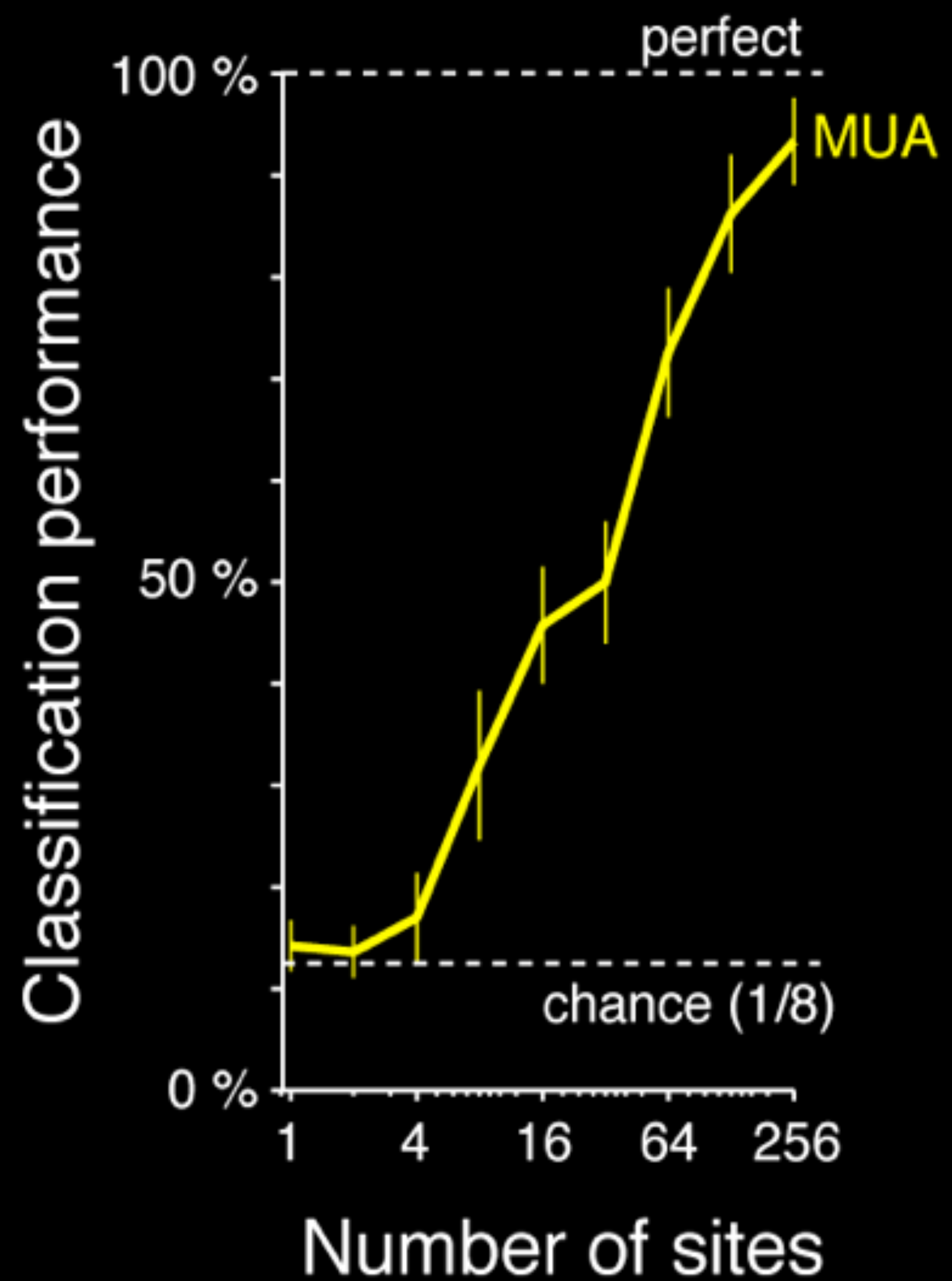
Cat/Dog

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

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

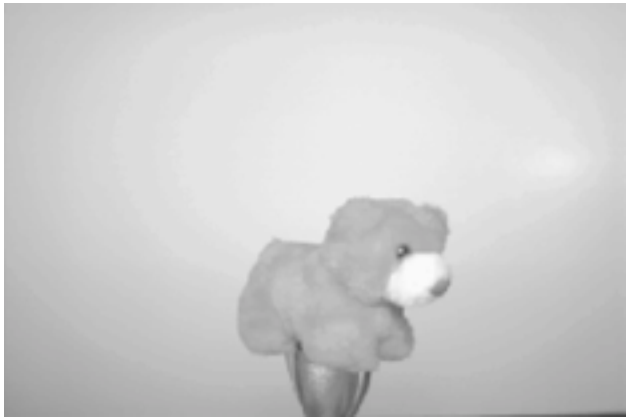
reliable object categorization using ~100 arbitrary AIT sites

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



Learning: image analysis

⇒ **Bear (0° view)**



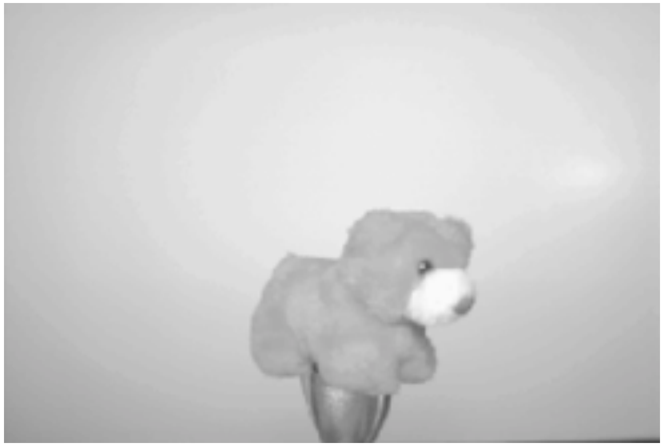
⇒ **Bear (45° view)**

Learning: image synthesis

UNCONVENTIONAL GRAPHICS

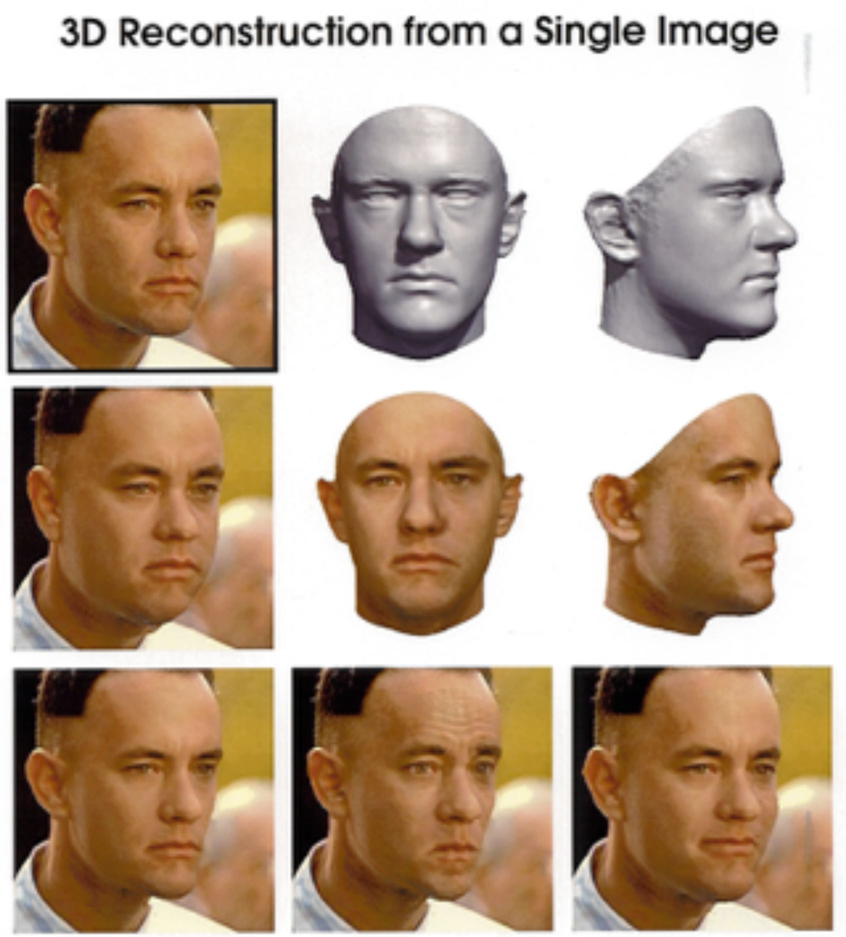
Θ = 0° view $\Rightarrow$

Θ = 45° view $\Rightarrow$



[redacted]

Learning: image synthesis

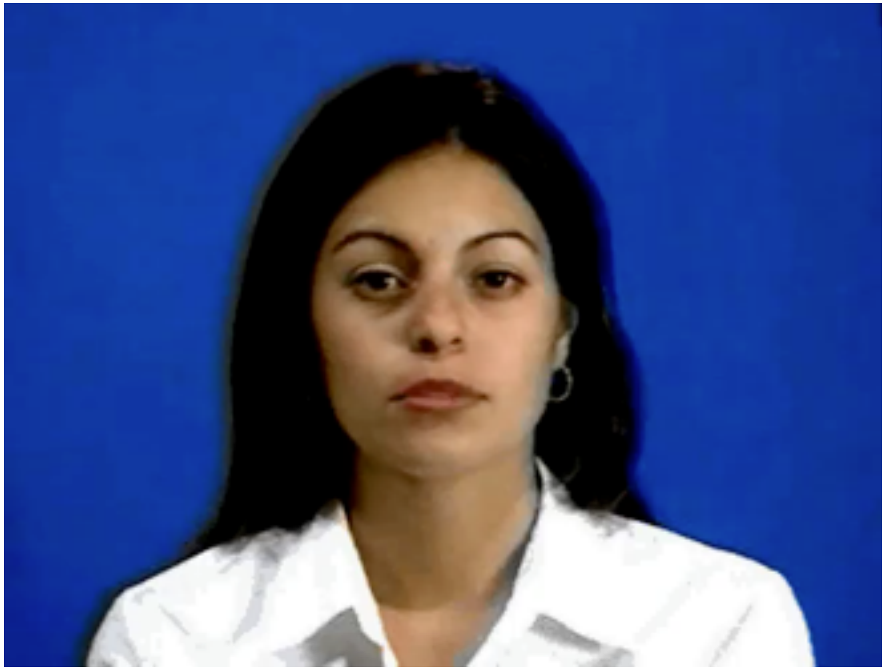


Blanz and Vetter,
MPI
SigGraph '99

Learning: image synthesis

Blanz and Vetter,
MPI
SigGraph '99

Mary101



A- more in a moment

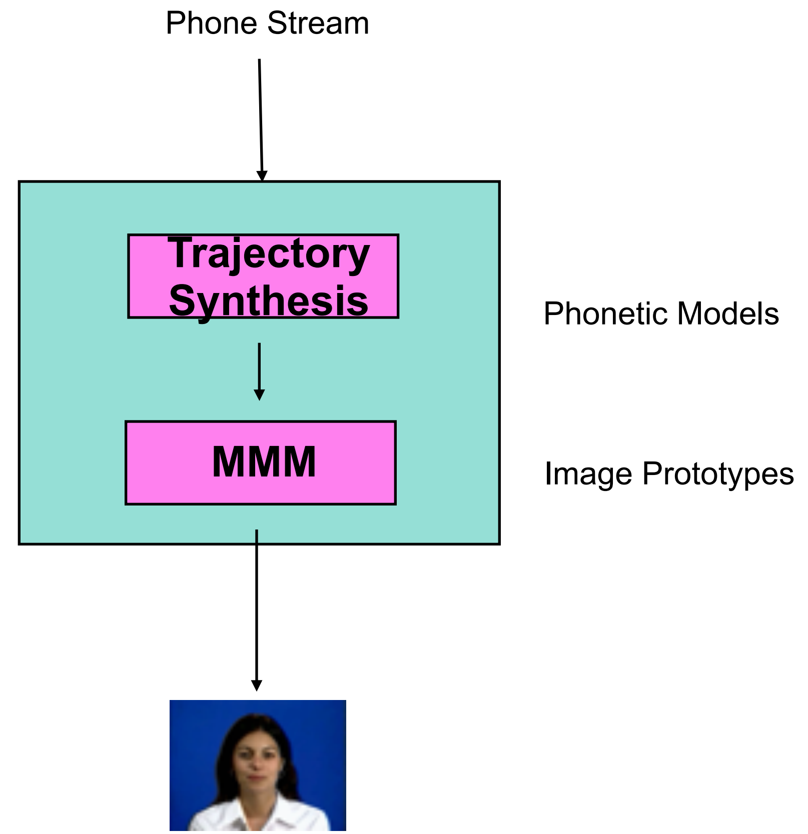
Tony Ezzat,Geiger, Poggio, **SigGraph 2002**

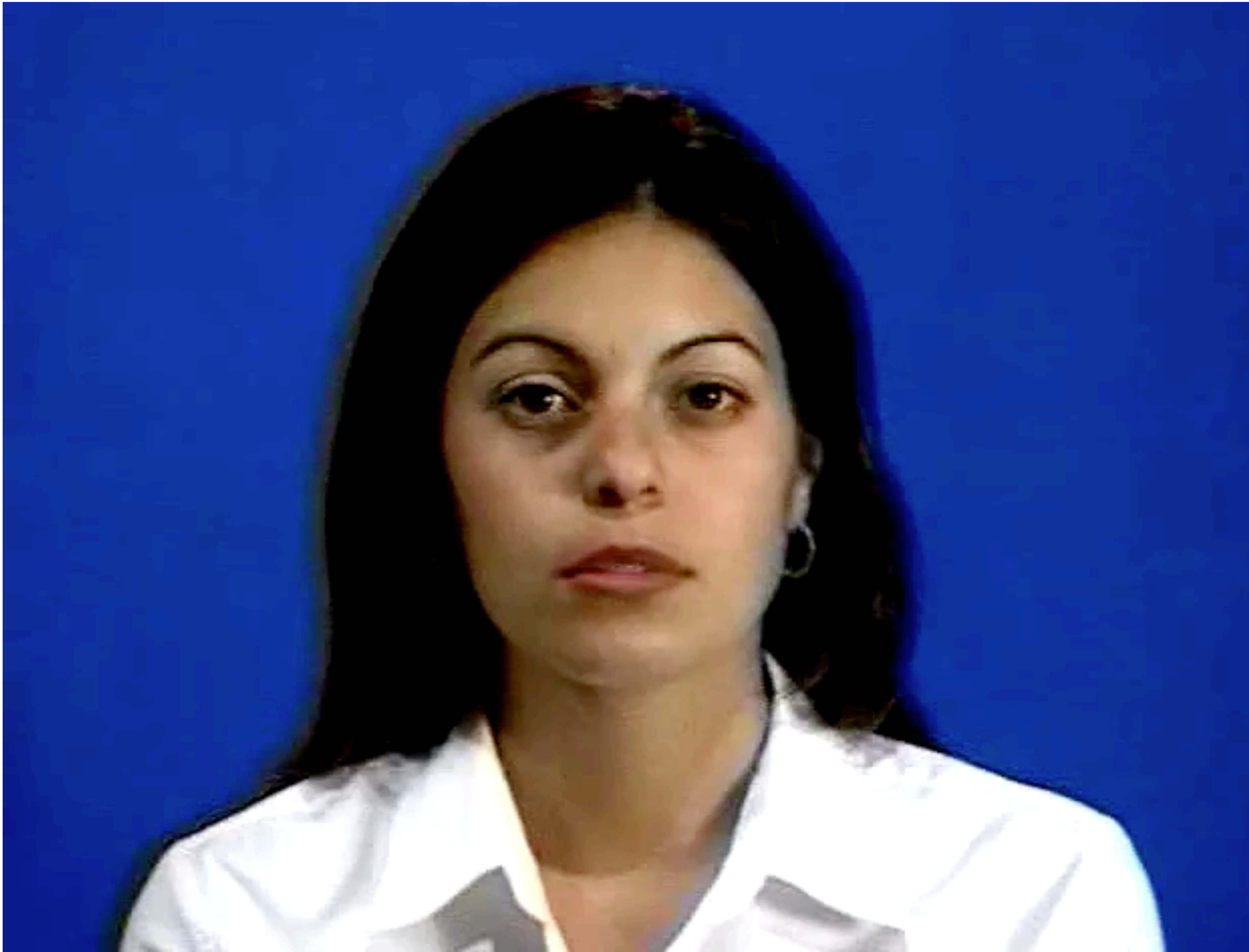
1. Learning

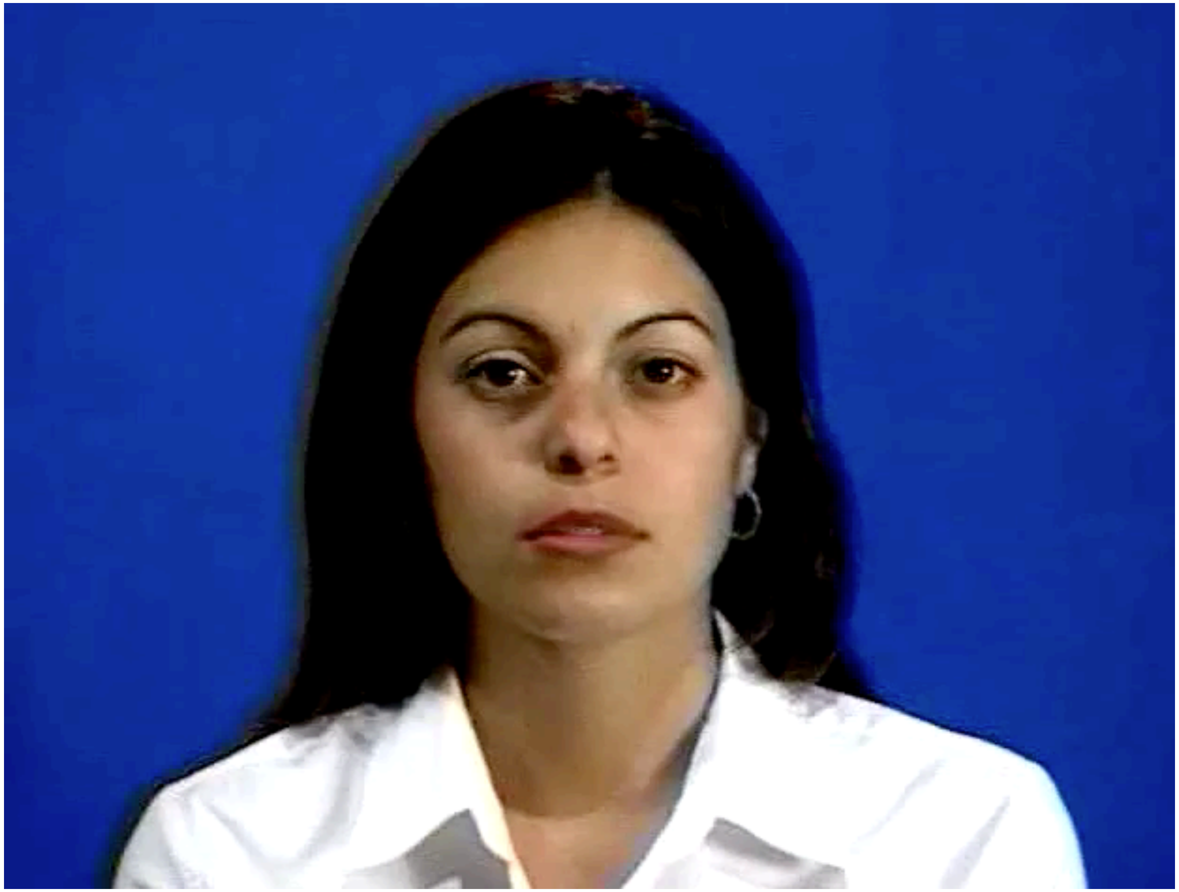
System learns from 4 mins of video face appearance (Morphable Model) and speech dynamics of the person

2. Run Time

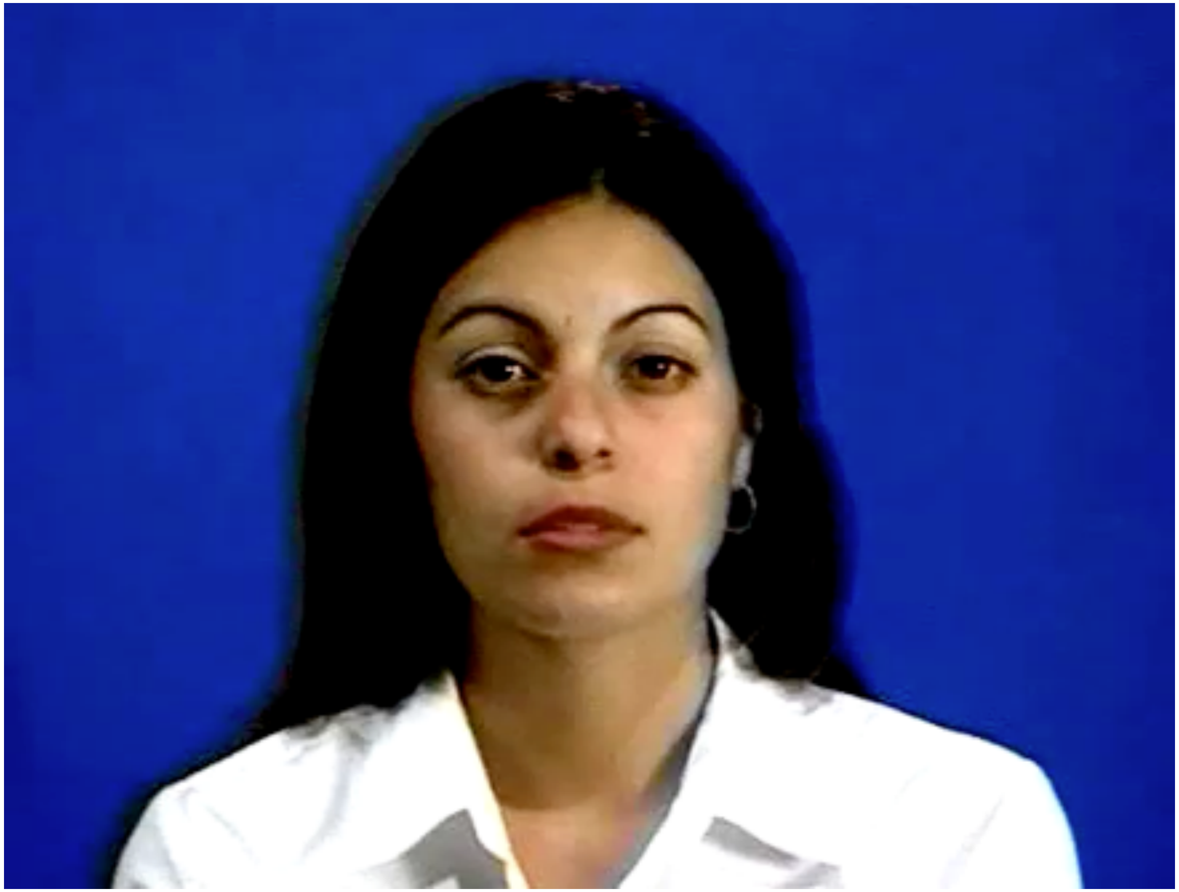
For any speech input the system provides as output a synthetic video stream



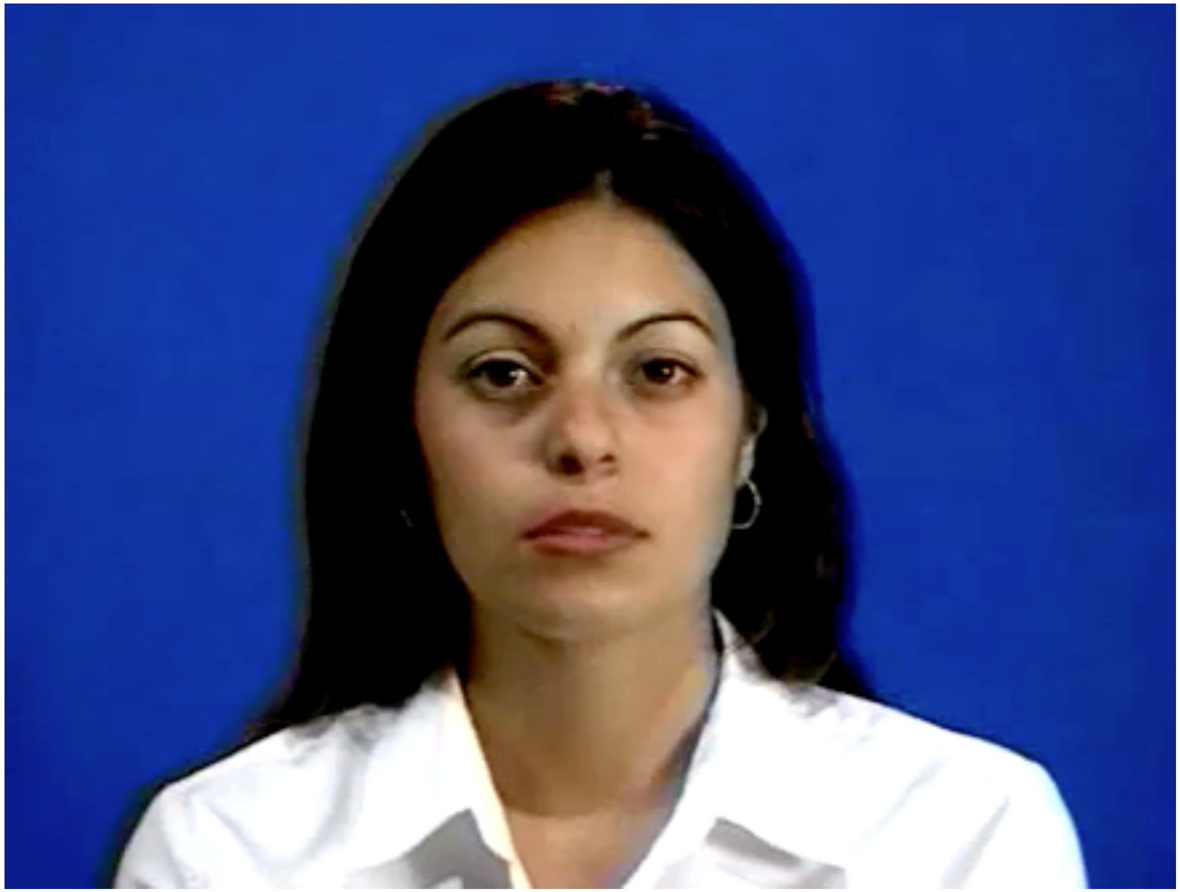




B-Dido



C-Hikaru

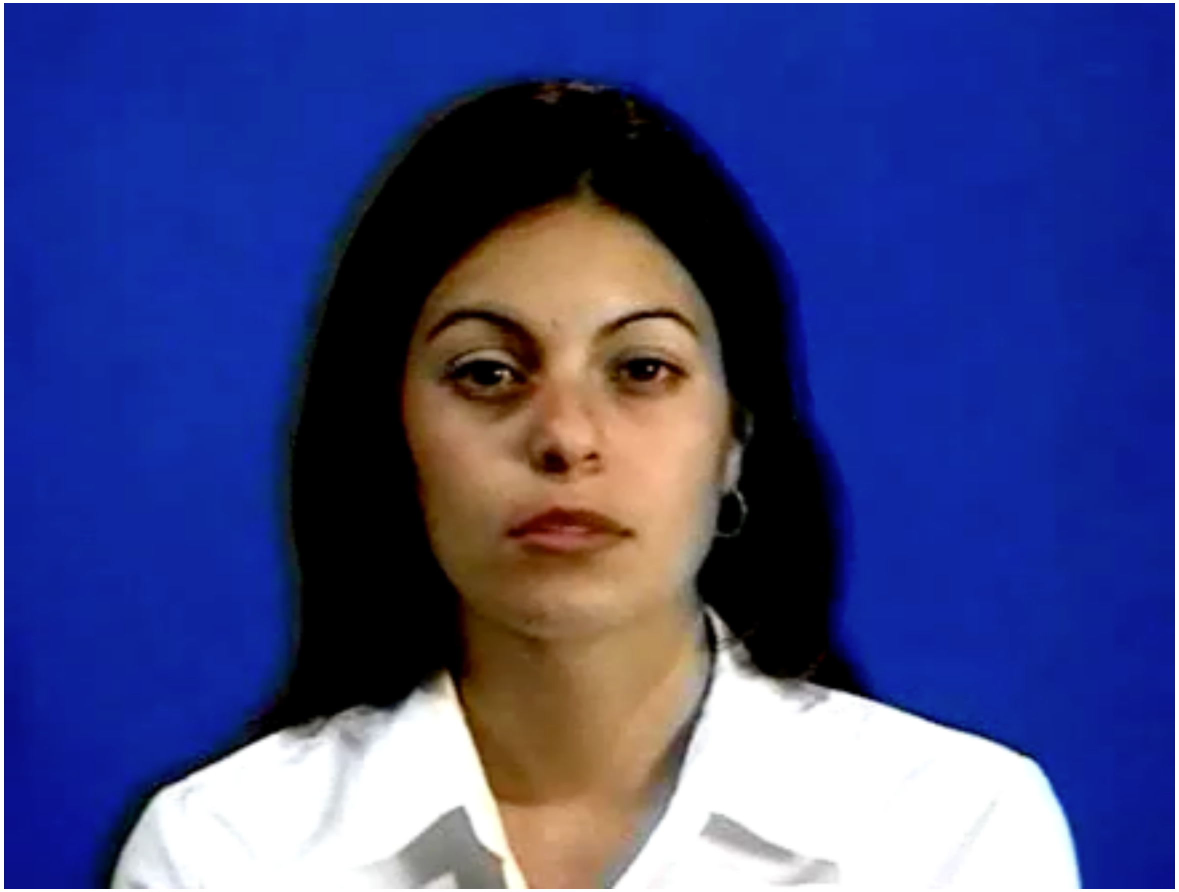


D-Denglijun

E-Marylin

F-Katie Couric

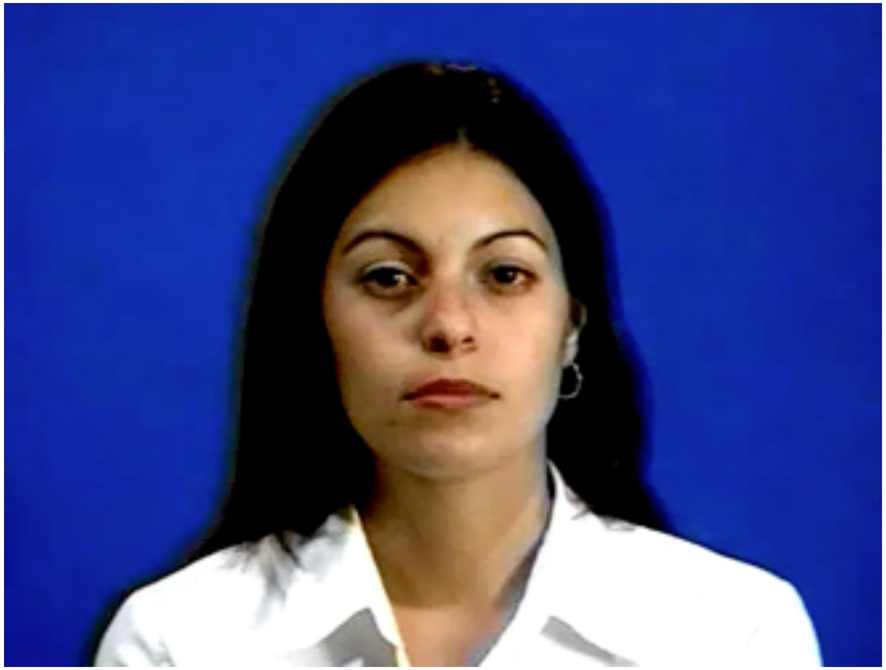
G-Katie



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**

Summary of today's overview

- Motivations for this course: a golden age for new AI and the key role of Machine Learning
- Statistical Learning Theory
- Success stories from past research in Machine Learning: examples of engineering applications
- Our machine learning class: science of intelligence, learning and the brain, CBMM.


What is this?
What is Hueihan doing?
What does Hueihan think about Joel’s thoughts about her?

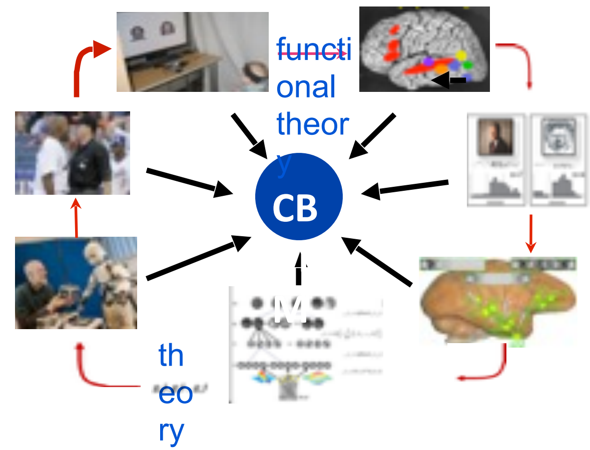
Intelligence and Turing^{++} Questions

- Intelligence —> <u>Human</u> Intelligence

- (Human) Intelligence: one word, many problems

- A CBMM mission: define and “answer” these *Turing^{++} Questions*

Turing^{++} Questions



The challenge is to develop computational models that answer questions about images and videos such as

what is there / who is there / what is the person doing

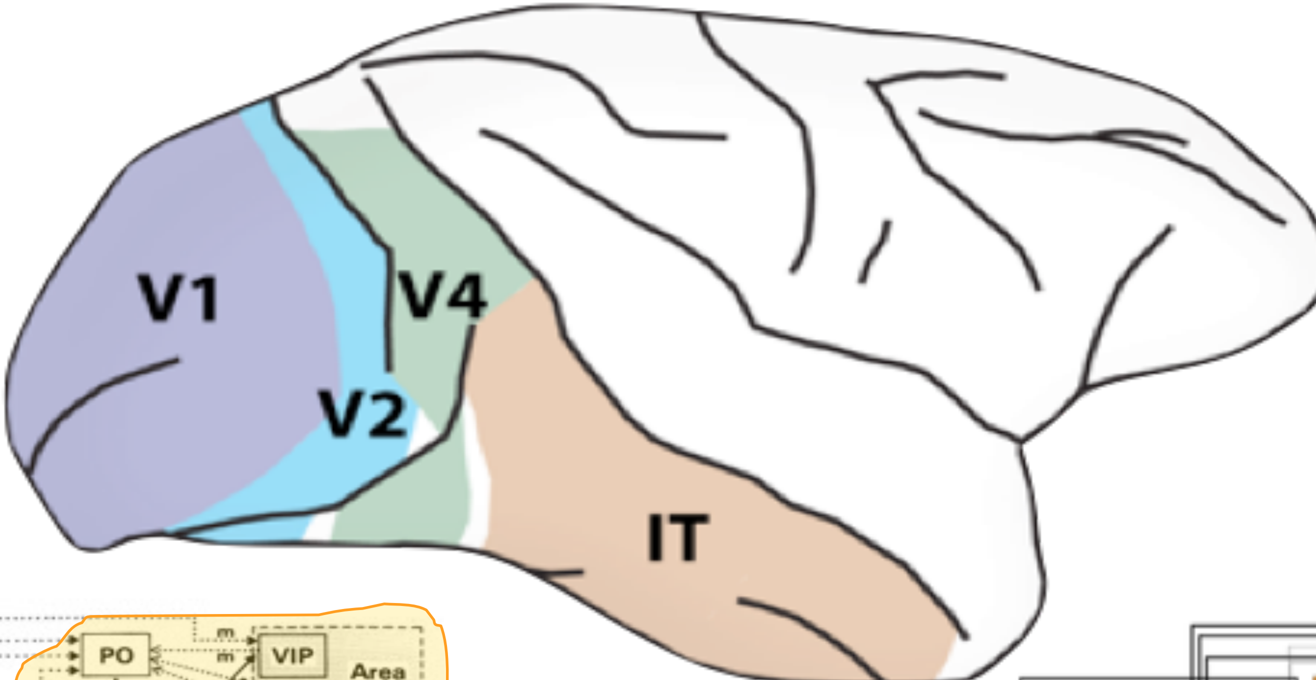
and eventually more difficult questions such as

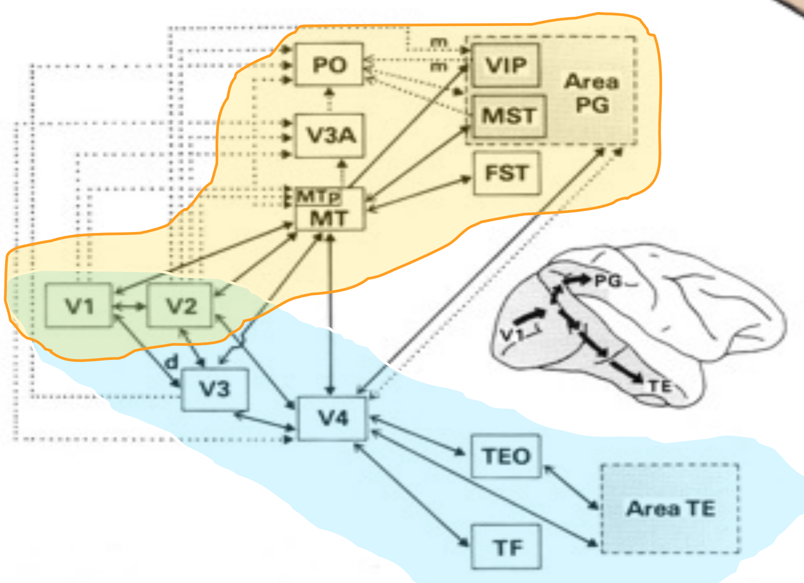
who is doing what to whom?

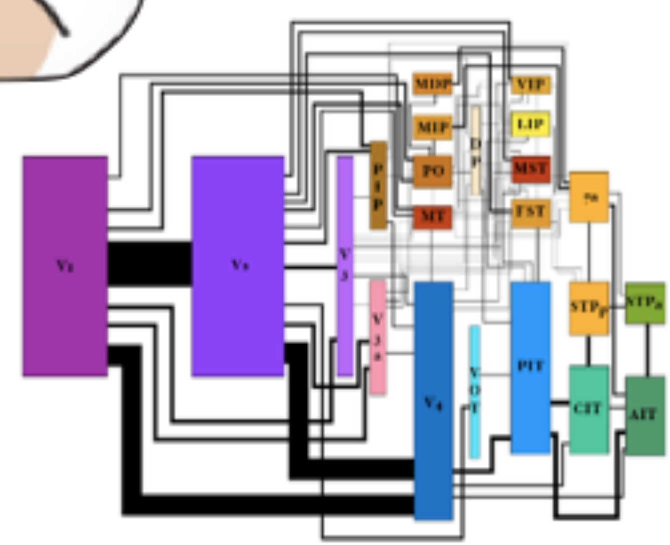
• *what happens next?*

at the <u>computational</u>, <u>**psychophysical**</u> and <u>**neural**</u> levels.

Object recognition



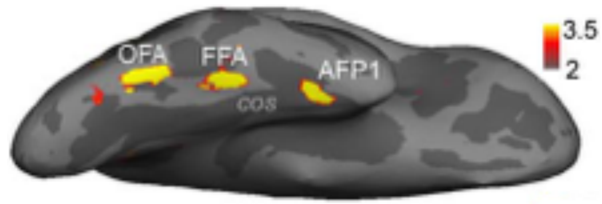


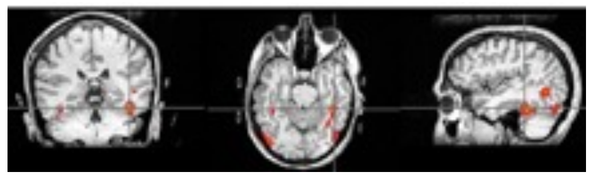


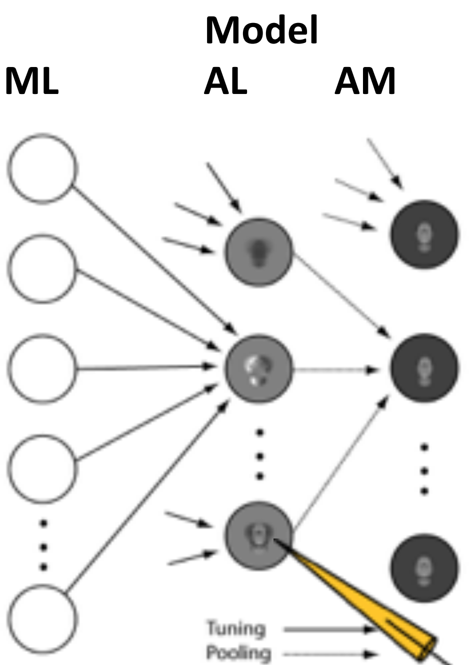
The who question: face recognition

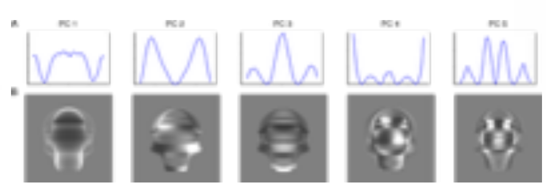
from experiments to theory

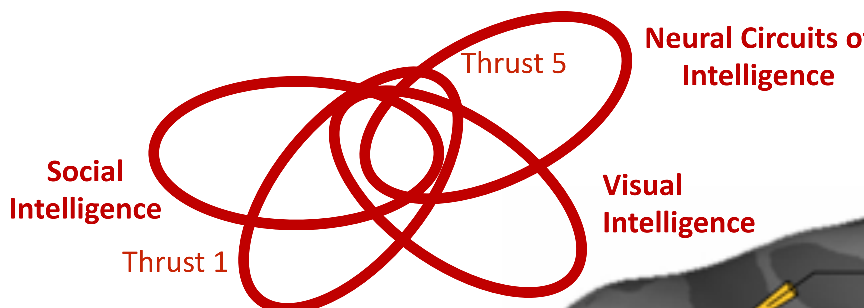
(Workshop, Sept 4-5, 2015)

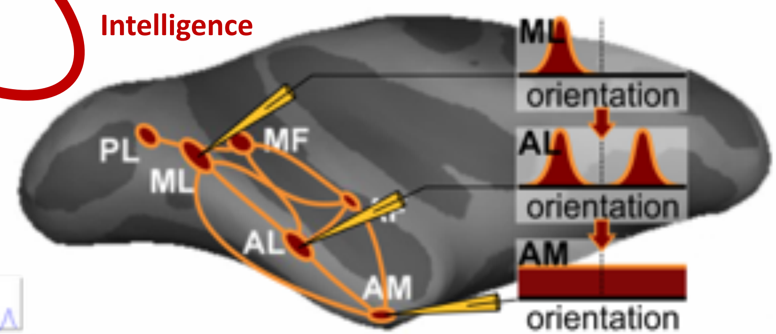






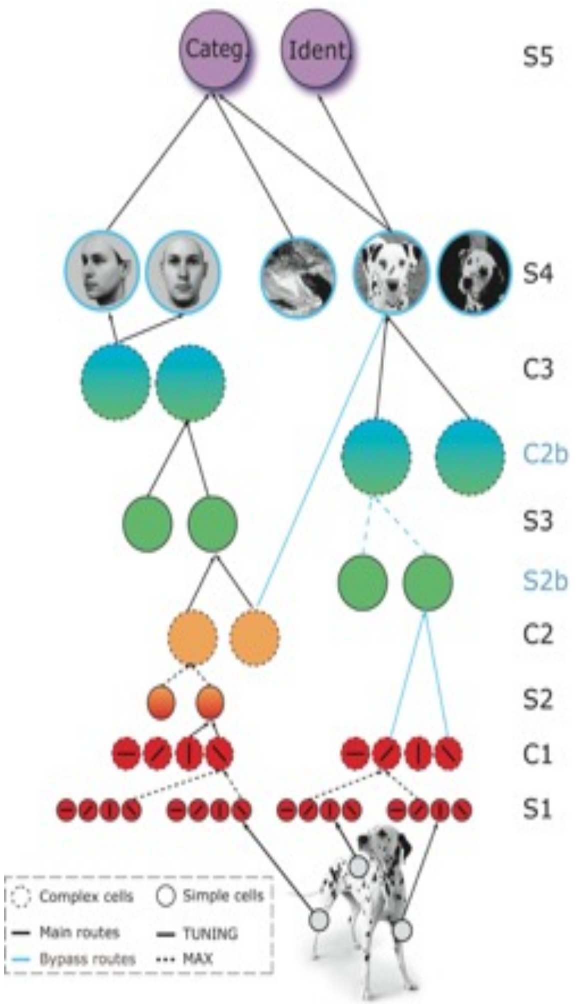


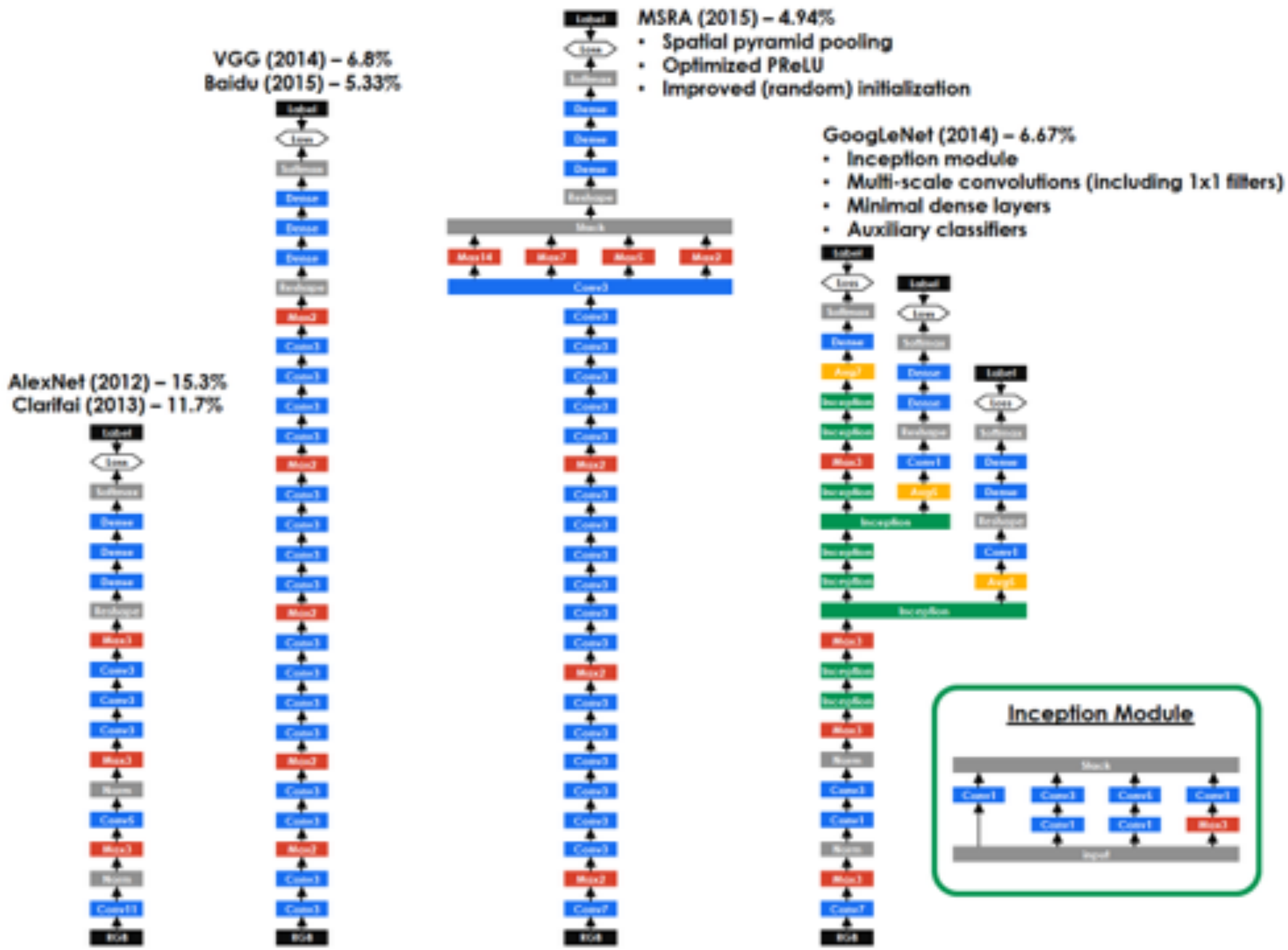




Extended i-theory

Learning of *invariant&selective* Representations

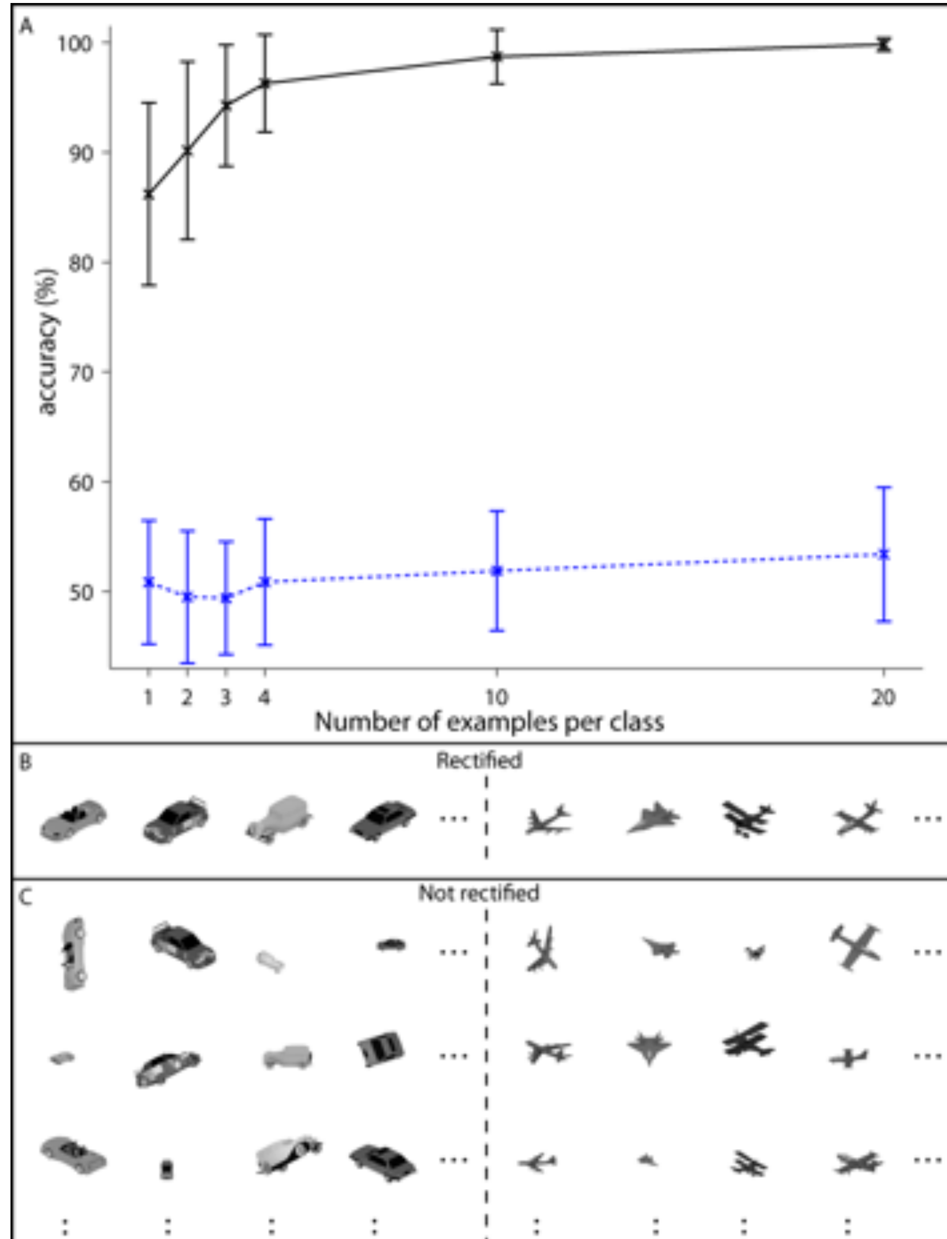




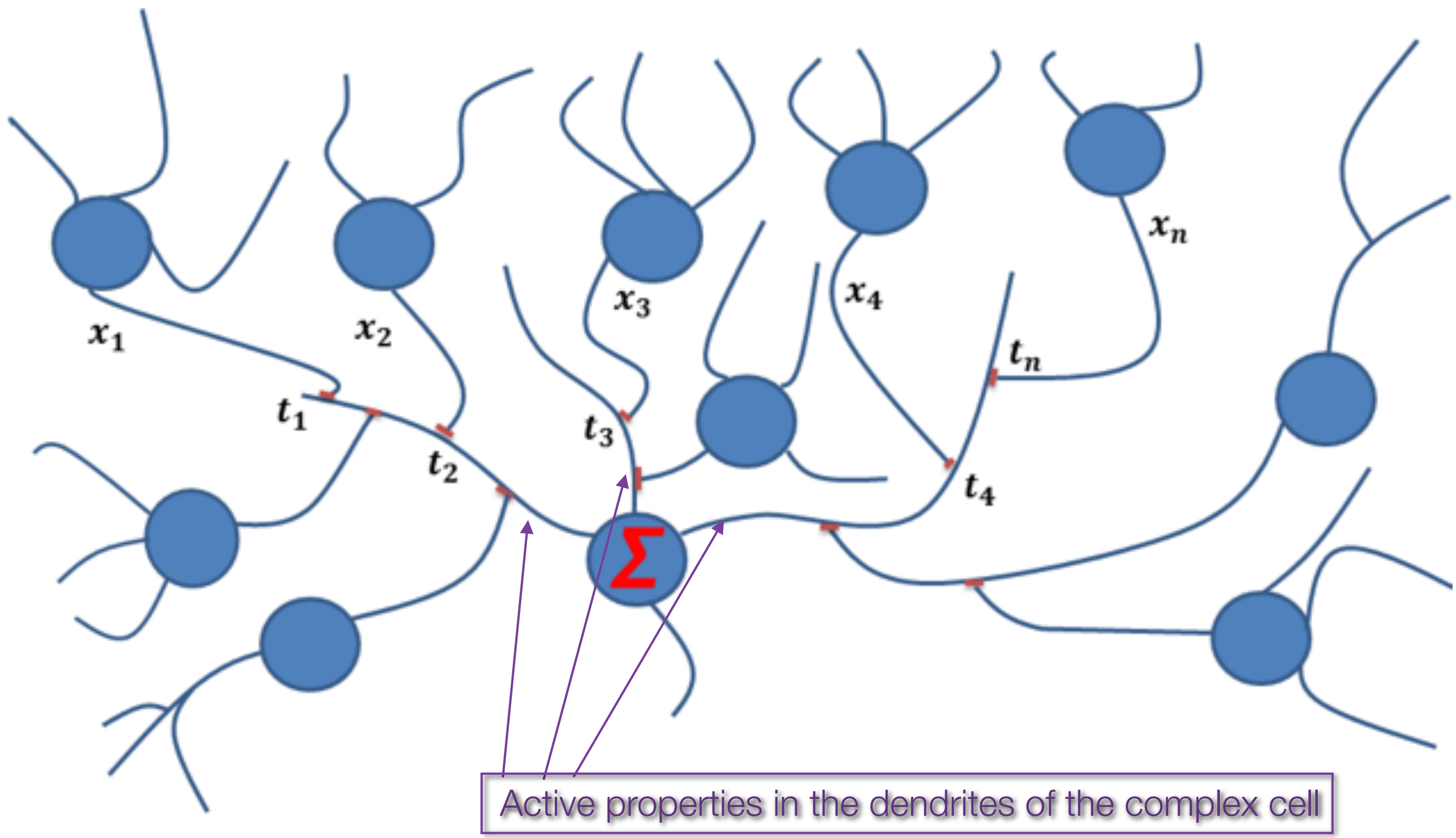
i-theory: invariant representations lead to lower sample complexity for a supervised classifier

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

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



Dendrites of a complex cells *as simple cells*…



I am now more in favor of deep learning as models of parts of the brain

WHY?

The background:

DCLNs (Deep Convolutional Learning Networks)

are doing very well

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?* Twelve years later, a most interesting theoretical question that still remains open, both for machine learning and neuroscience, is indeed *why hierarchies.*

What if DCLNs are the secret of the brain?

Is supervised training with millions of labeled examples biologically plausible?

Implicitly Labeled Examples (ILEs):

interesting research here!

Deep Convolutional Learning Networks like HMAX can be trained effectively with large numbers of labeled examples. This may be biologically plausible if we can show that ILEs could be be used to the same effect. What needs to be done is to train, with a plausible number of ILEs, biologically plausible multilayer architectures. For instance, for visual cortex take into account known parameters, such as receptive field sizes, related range of pooling and especially eccentricity dependence of RF.

Through a new theory for DCLNs to the next frontier in machine learning

<u>The first phase</u> (and successes) of ML:
supervised learning: $n \rightarrow \infty$

<u>The next phase</u> of ML: unsupervised and *implicitely supervised* learning
of invariant representations for learning: $n \rightarrow 1$