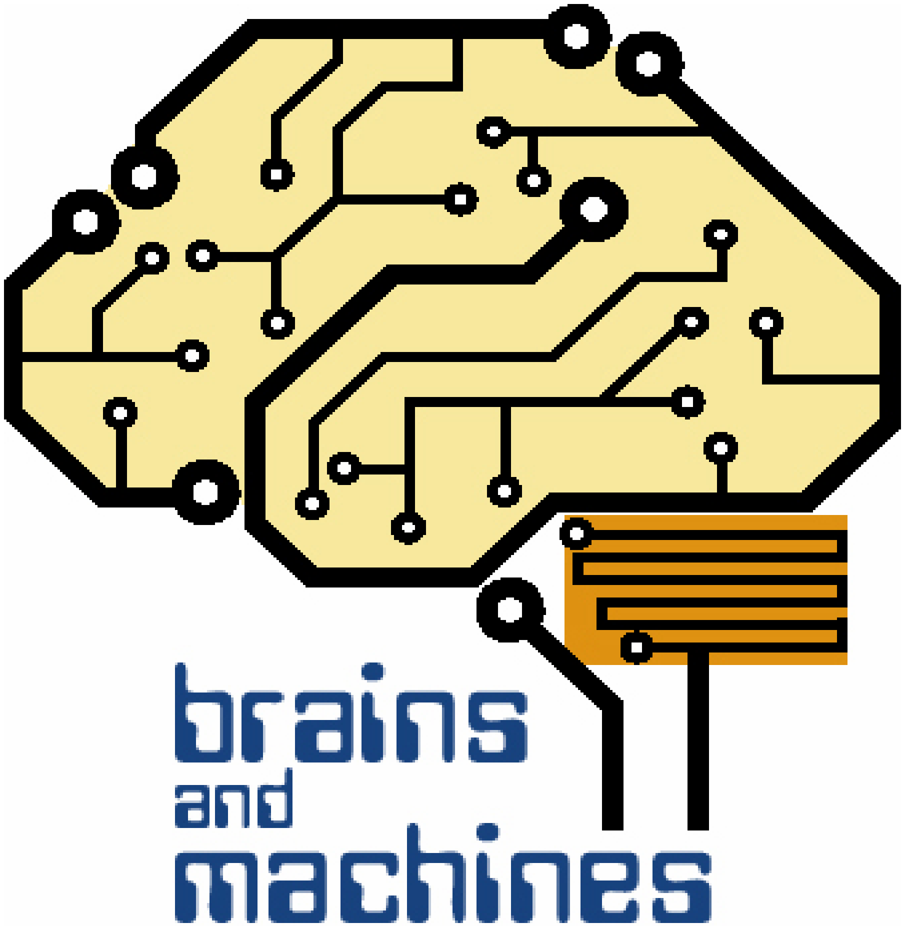
9.250

Statistical Learning Theory and Applications

Lorenzo Rosasco + Jake Bouvrie +
Ryan Rifkin + Tomaso Poggio
McGovern Institute for Brain Research
Center for Biological and Computational Learning
Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Learning: Brains and Machines



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

Problem of learning:
a focus for

- modern math
- computer algorithms
- neuroscience

Learning: much more than memory

- Role of **learning (**theory and applications in many different domains) has grown substantially in CS

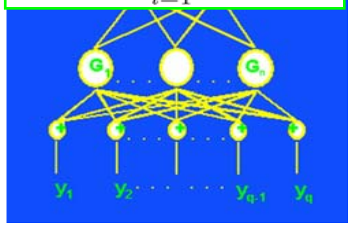
- Plasticity and learning have a central stage in the neurosciences

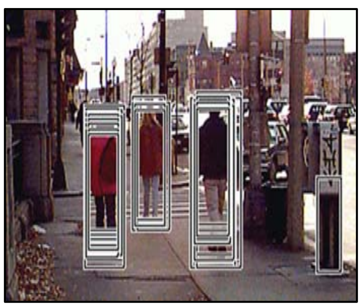
- Until now math and engineering of learning has developed independently of neuroscience…but it may begin to change: we will see in the class the situation in vision…

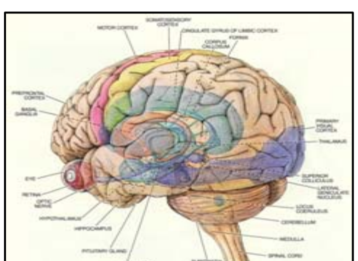
Learning:
math, engineering, neuroscience

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

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







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 – and how it may suggest better computer vision systems

Class

Rules of the game: problem sets (2)
final project (min = review; max = j. paper)
grading
participation!

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

Slides on the Web site
Staff mailing list is 9.520@mit.edu
Student list will be 9.520students@mit.edu
<u>Please fill form!</u>

9.520 Statistical Learning Theory and Applications (2007)

Class 26: Project presentations (past examples)

10:30 - Simon Laflamme "Online Learning Algorithm for Structural Control using Magnetorheological Actuators"

- Emily Shen "Time series prediction"
- Zak Stone "Facebook project"
- Jeff Miller "Clustering features in the standard model of cortex"
- Manuel Rivas "Learning Age from Gene Expression Data"
- Demba Ba "Sparse Approximation of the Spectrogram via Matching Pursuits: Applications to Speech Analysis"
- Nikon Rasumov "Data mining in controlled environment and real data"

9.520 Statistical Learning Theory and Applications (2003)

Class 26: Project presentations (past examples)

2:35-2:50 "Learning card playing strategies with SVMs", David Craft and Timothy Chan

2:50-3:00 "Artificial Markets: Learning to trade using Support Vector Machines", Adlar Kim

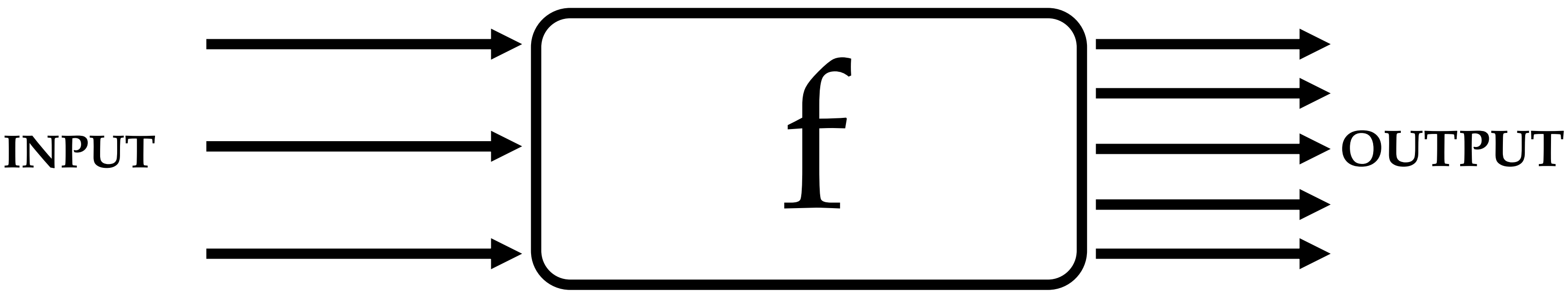
3:00-3:10 "Feature selection: literature review and new development'', Wei Wu

3:10—3:25 "Man vs machines: A computational study on face detection" Thomas Serre

Overview of overview

- The problem of supervised learning: “real” math behind it
- Examples of engineering applications (from our group)
- Learning and the brain

Learning from examples: goal is not to memorize but to **generalize**, eg *predict*.



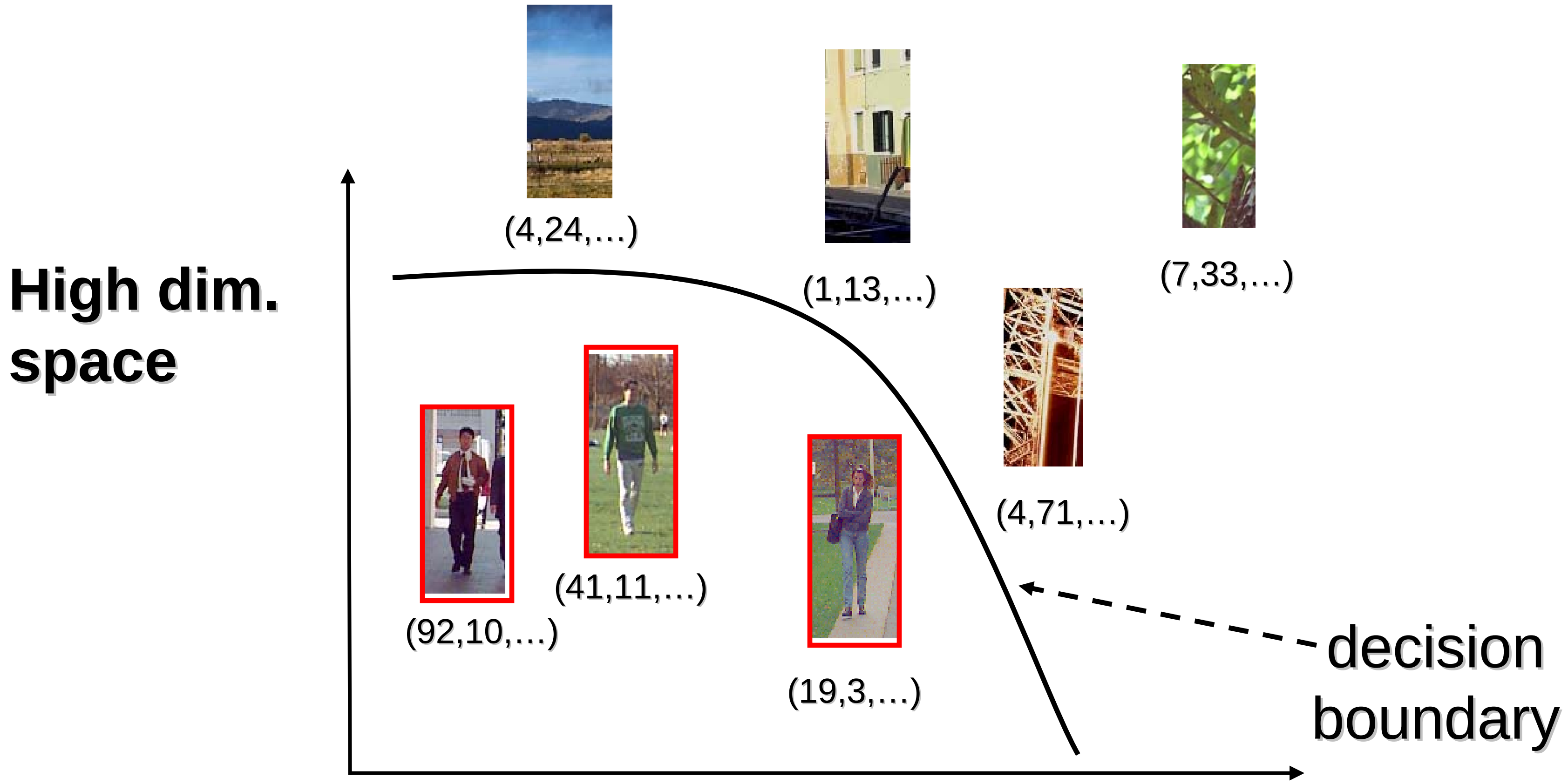
Given a set of ℓ examples (data) $\{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\}$

Question: find function *f* such that

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

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

Binary classification case



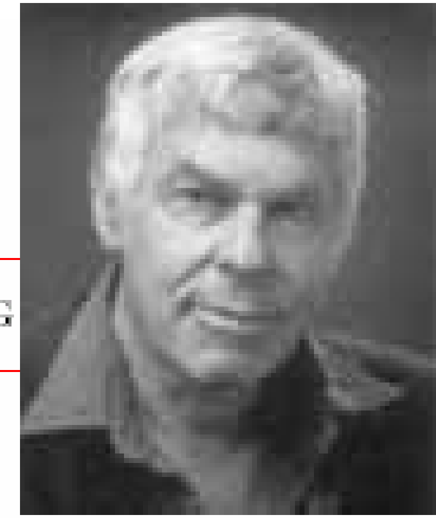
Reason to learn some learning theory

Applications cannot be carried out by simply using a black box.

What is needed: the right formulation of the problem (which is helped by knowledge of theory): choice of representation (inputs, outputs), choice of examples, validate predictivity, do not datamine

… $f(\mathbf{x}) = \mathbf{w}\mathbf{x} + b$

Interesting development: in the last few years he theoretical foundations of learning have become part of mainstream mathematics



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

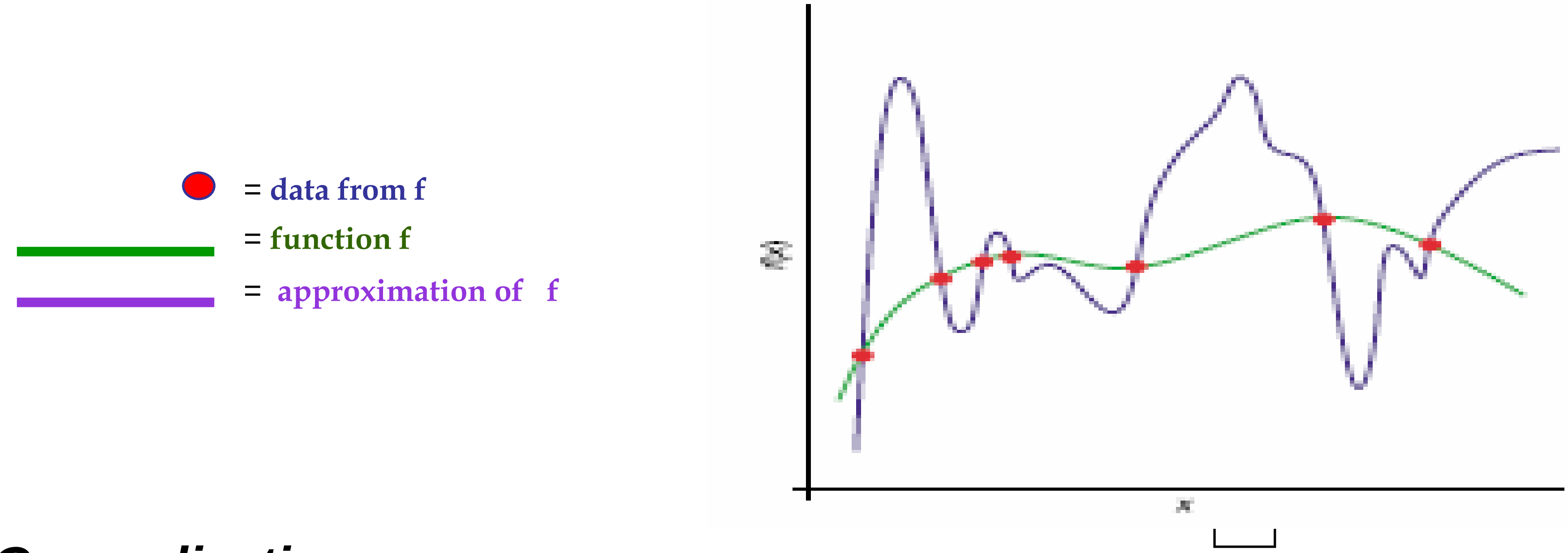
FELIPE CUCKER AND STEVE SMALE

The problem of learning is arguably at the very core of the problem of intelligence, both biological and artificial.

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.

Learning from examples: predictive, multivariate function estimation from sparse data (not just curve fitting)



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

***Regression*:** **function is real valued**

***Classification*:** **function is binary**

The learning problem

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 closed 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.*

Thus....the key requirement (main focus of classical learning theory) to solve the problem of learning from examples: *generalization*

Example:

A standard way to learn from examples is ERM (empirical risk minimization)

$$\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(x_i), y_i)$$

The problem does not have a *predictive* solution in general (just fitting the data does not work). Choosing an appropriate hypothesis space *H* (for instance a compact set of continuous functions) can guarantee generalization (how good depends on the problem and other parameters).

A superficially different requirement for learning to be possible is that the problem is *well-posed* (solution exists, *stable*)

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)

Thus....two key requirements to solve the problem of learning from examples: *well-posedness* and *generalization. How are they related?*

Intuition: Consider the standard learning

$$\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(x_i), y_i)$$

The main focus of learning theory is *predictivity* of the solution eg *generalization*. The problem is in addition *ill-posed*. It was known that by choosing an appropriate hypothesis space *H* predictivity is ensured. It was also known that appropriate *H* provide well-posedness.

A couple of years ago it was shown that under quite general assumptions

generalization and well-posedness are *equivalent*, eg one implies the other.

Thus a stable solution is predictive and (for ERM) also viceversa.

Learning theory and natural sciences

Conditions for **generalization** in learning theory

have deep, almost philosophical, implications:

they may be regarded as conditions that guarantee a theory to be *predictive* (that is *scientific*)

We have used a simple algorithm
-- that ensures generalization --
in most of our applications…

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

implies

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

Equation includes Regularization Networks (special cases are splines, Radial Basis Functions and Support Vector Machines). Function is nonlinear and general approximator…

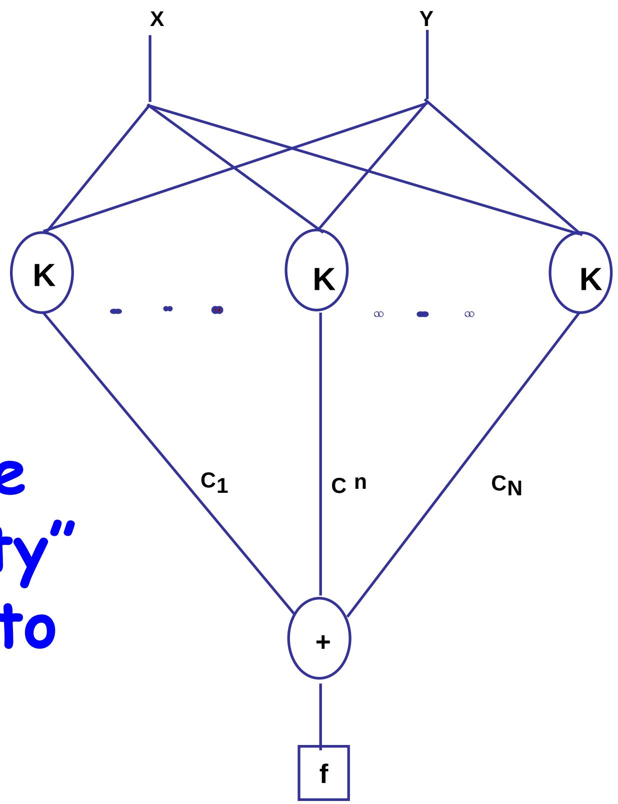
For a review, see Poggio and Smale, ***The Mathematics of Learning****,*
Notices of the AMS, 2003

Another remark: equivalence to networks

Many different V lead to the same solution…

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

…and can be "written" as the same type of network…where the value of K corresponds to the "activity" of the "unit" and the c_i correspond to (synaptic) "weights"



Winning against the curse of dimensionality: new research directions in learning

Many processes - physical processes as well as human activities – generate high-dimensional data. Because of the high dimensionality these data are in general difficult to analyze: their sample complexity is too high (eg *curse of dimensionality or poverty of stimulus*). There are, however, basic properties of the data generating process that may allow to circumvent the problem of high dimensionality and make the analysis possible.

A classical example is smoothness - exploited by L2 regularization techniques: the underlying principle is smoothness of the underlying function space.

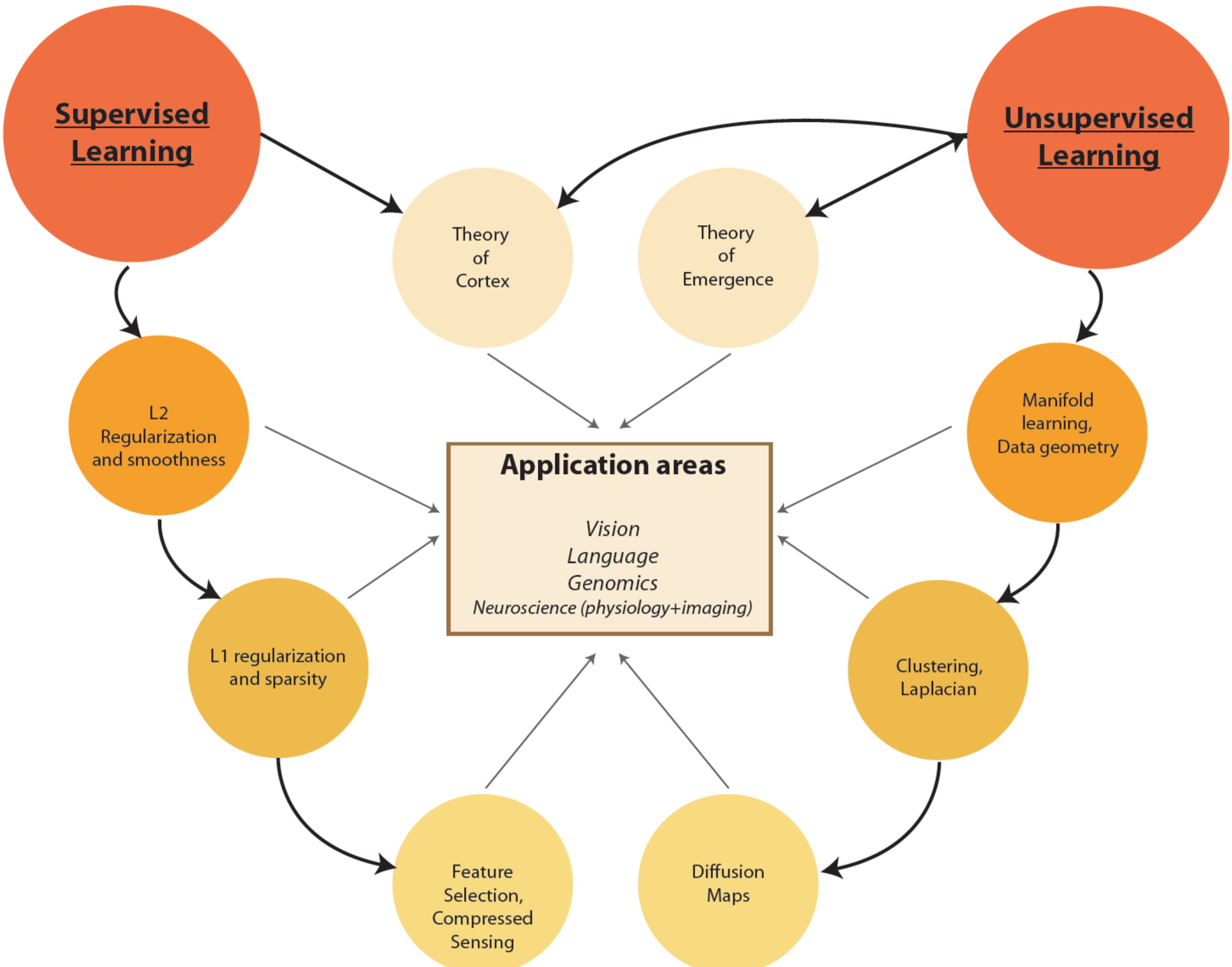
Very recently, mathematicians and computer scientists have been uncovering novel principles that apply to other broad classes of phenomena and allow circumventing the problems posed by the high dimensionality of the data.

Panning for Gold: The Science and Applications of Learning from Data

The Team

Stanley Osher (UCLA), Terence Tao (UCLA),
Joseph Teran (UCLA), Partha Niyogi (U. Chicago),
Stephen Smale (TTI-C, U. Chicago), Ingrid Daubechies (Princeton), Olga Troyanskaya (Princeton),
Yann LeCun (NYU), Tomaso Poggio (MIT)

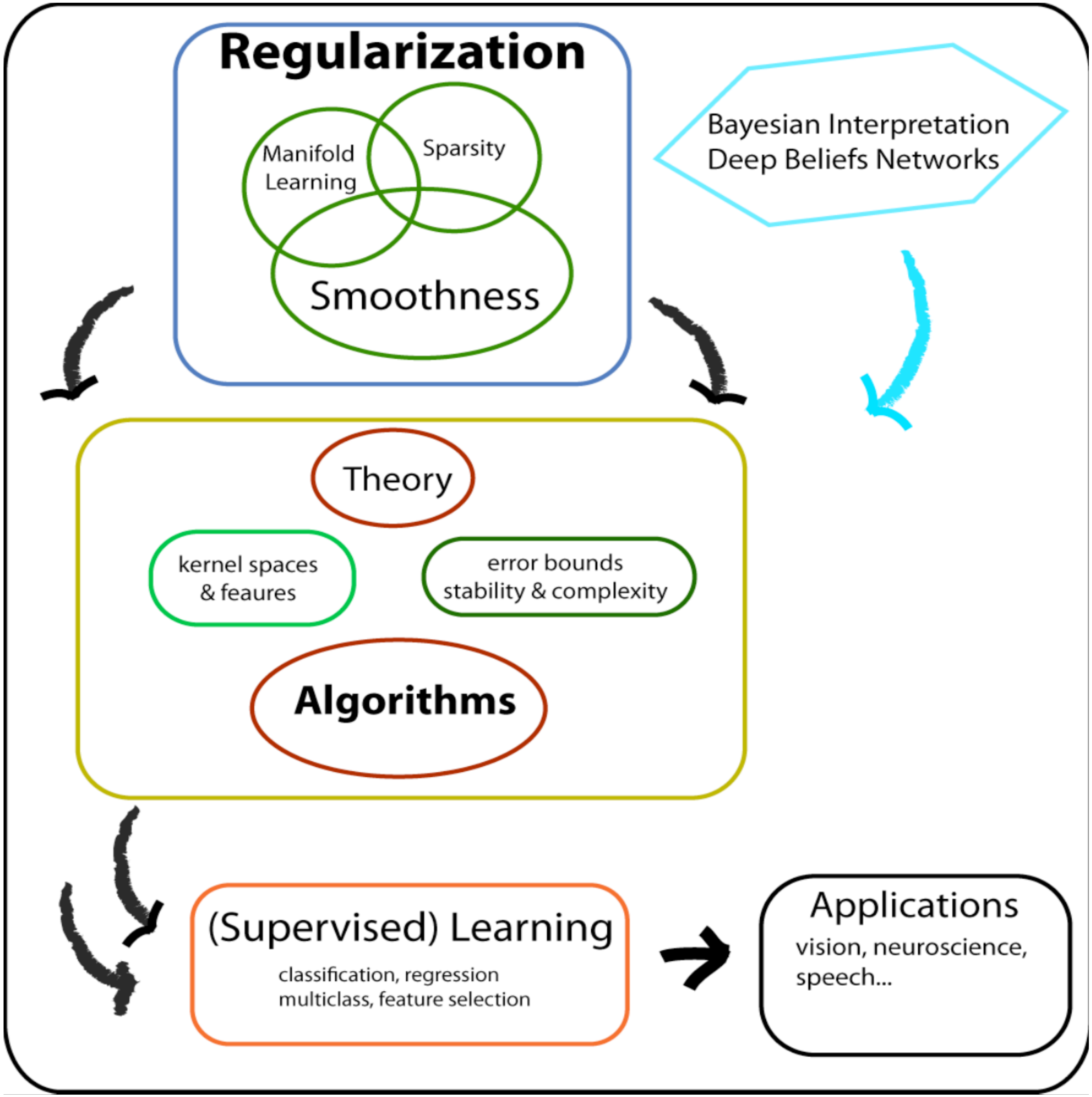
New Research Directions



What are the principles of learning from few data in high dimensional spaces?

How might it be possible to make reliable inferences about the underlying phenomena without running into the curse of dimensionality. There are at least three different points of view from which to approach this question: *smoothness, sparsity, and low dimensional geometry*.

- It has long been known that if f belongs to a Sobolev space of order s, then the rate of convergence for nonparametric learning depends on the ratio of smoothness and dimensionality, eg functions in a Sobolev space of high order (i.e., smoother functions) are learned more easily. A more recent development is that the framework of Mercer kernels and Reproducing Kernel Hilbert Spaces (RKHS) allows one to implicitly capture smoothness classes while allowing for efficient algorithms based on regularization.
- A second point of view is that the function of interest may not be smooth in a classical sense but may be sparse in some suitable basis. This includes the application of wavelet based methods for learning and function approximation as well as recent developments in compressed sensing (*L1 sparsity*).
- A third and more recent point of view is built around the hypothesis that although natural data lives in very high dimensional spaces, they concentrate around lower dimensional geometrically structured objects. The most prominent of these methods assume this lower dimensional object to be a submanifold and show how to build suitable classes of functions on this submanifold from randomly sampled data. The topology and geometry of this submanifold may be revealed through the empirical Laplace operator and the heat kernel on data derived graphs and simplicial complexes (*diffusion maps*).



Regularization
Manifold Learning
Sparsity
Smoothness
Bayesian Interpretation
Deep Beliefs Networks
Theory
kernel spaces & feaures
error bounds stability & complexity
Algorithms
(Supervised) Learning
classification, regression multiclass, feature selection
Applications
vision, neuroscience, speech...

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

Overview

- Supervised learning: real math
- Examples of recent and ongoing in-house engineering applications

Overview of overview

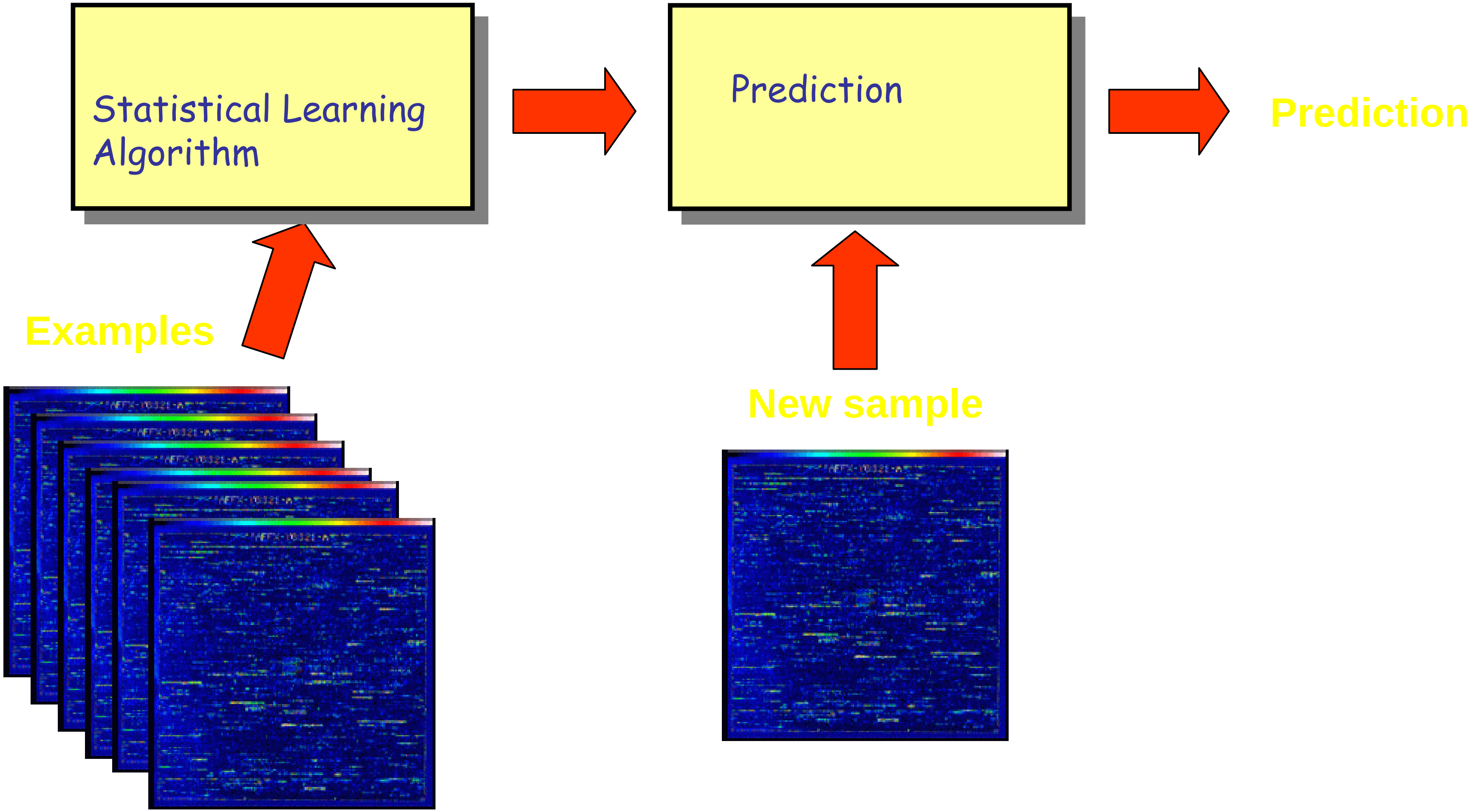
- The problem of supervised learning: "real" math behind it
- Examples of engineering applications (from our group)
- Learning and the brain

Learning from Examples: engineering applications

Bioinformatics
Artificial Markets
Object categorization
Object identification
Image analysis
Graphics
Text Classification
.....

Bioinformatics application: predicting type of cancer from DNA chips signals

Learning from examples paradigm



Bioinformatics application: predicting type of cancer from DNA chips

New feature selection SVM:

Only 38 training examples, 7100 features

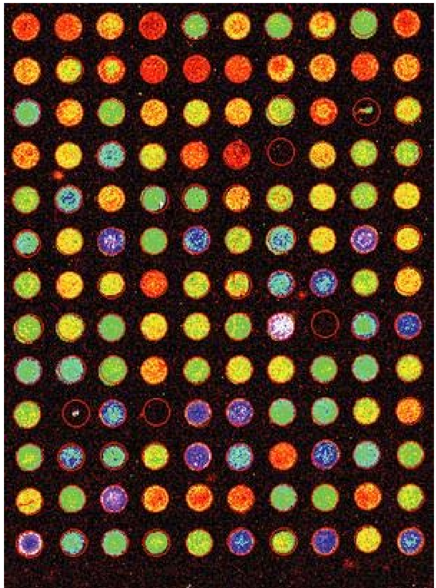
AML vs ALL: 40 genes 34/34 correct, 0 rejects.
5 genes 31/31 correct, 3 rejects of which 1 is an error.

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 from Examples: engineering applications

Bioinformatics
Artificial Markets
Object categorization
Object identification
Image analysis
Graphics
Text Classification
.....

Object recognition for computer vision: (personal) historical perspective

Face detection

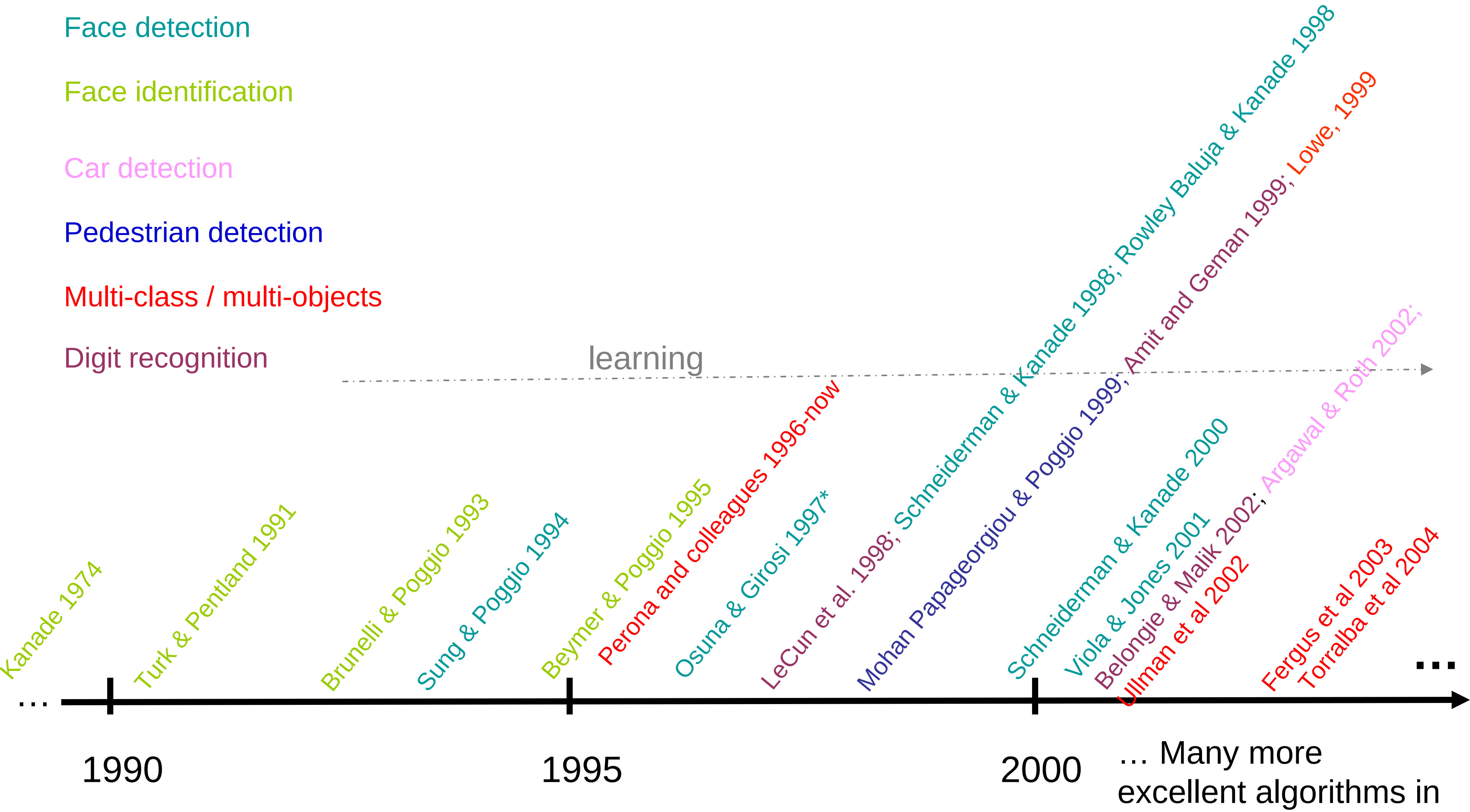
Face identification

Car detection

Pedestrian detection

Multi-class / multi-objects

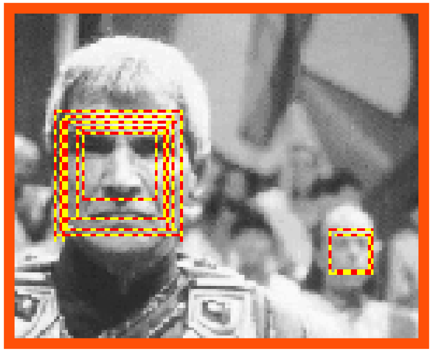
Digit recognition

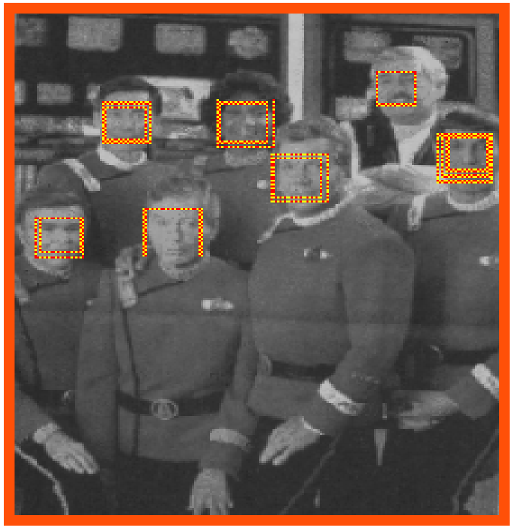


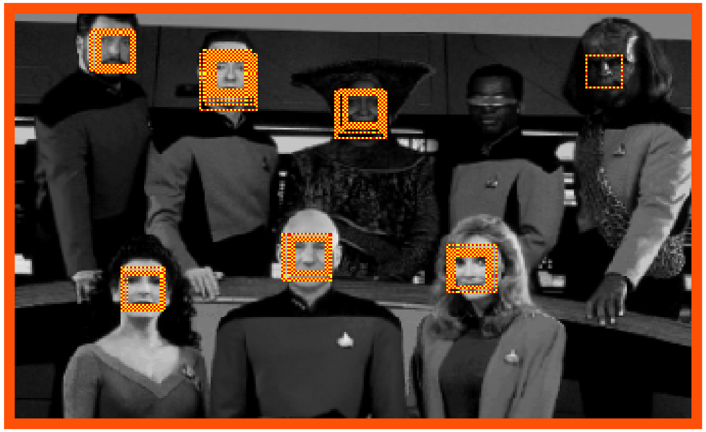
*Best CVPR'07 paper 10 yrs ago

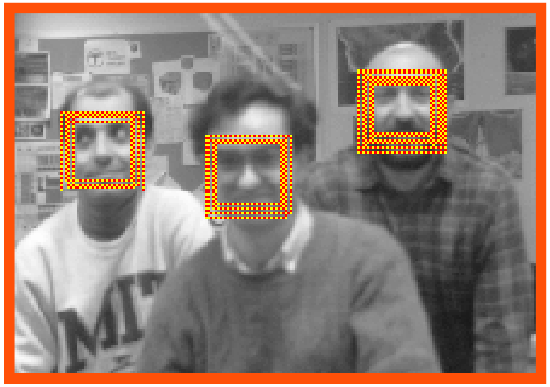
Examples: Learning Object Detection: Finding Frontal Faces

- Training Database
- 1000+ Real, 3000+ VIRTUAL
- 50,0000+ Non-Face Pattern



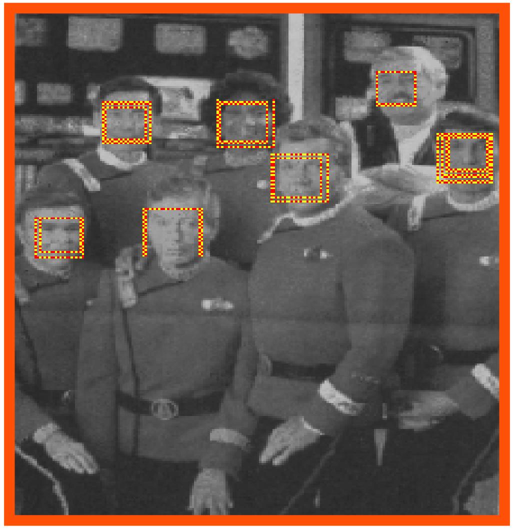


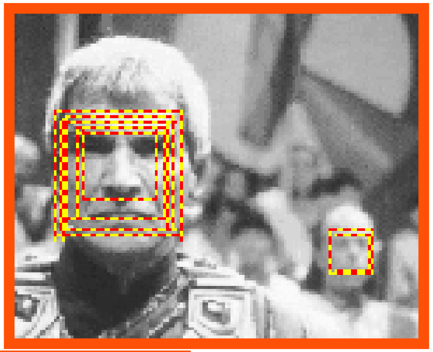


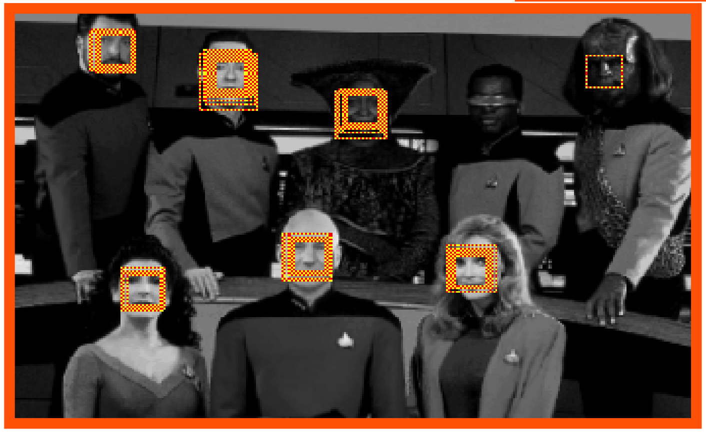


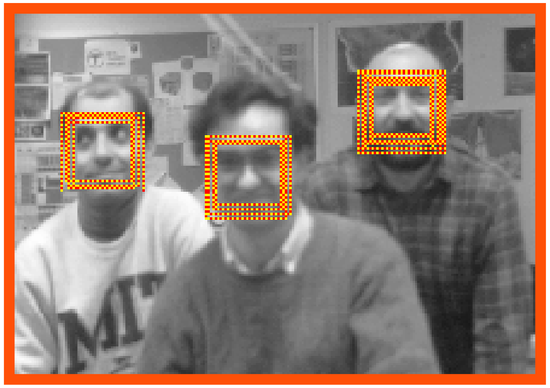
Sung & Poggio 1995

Learning Object Detection: Finding Frontal Faces ...





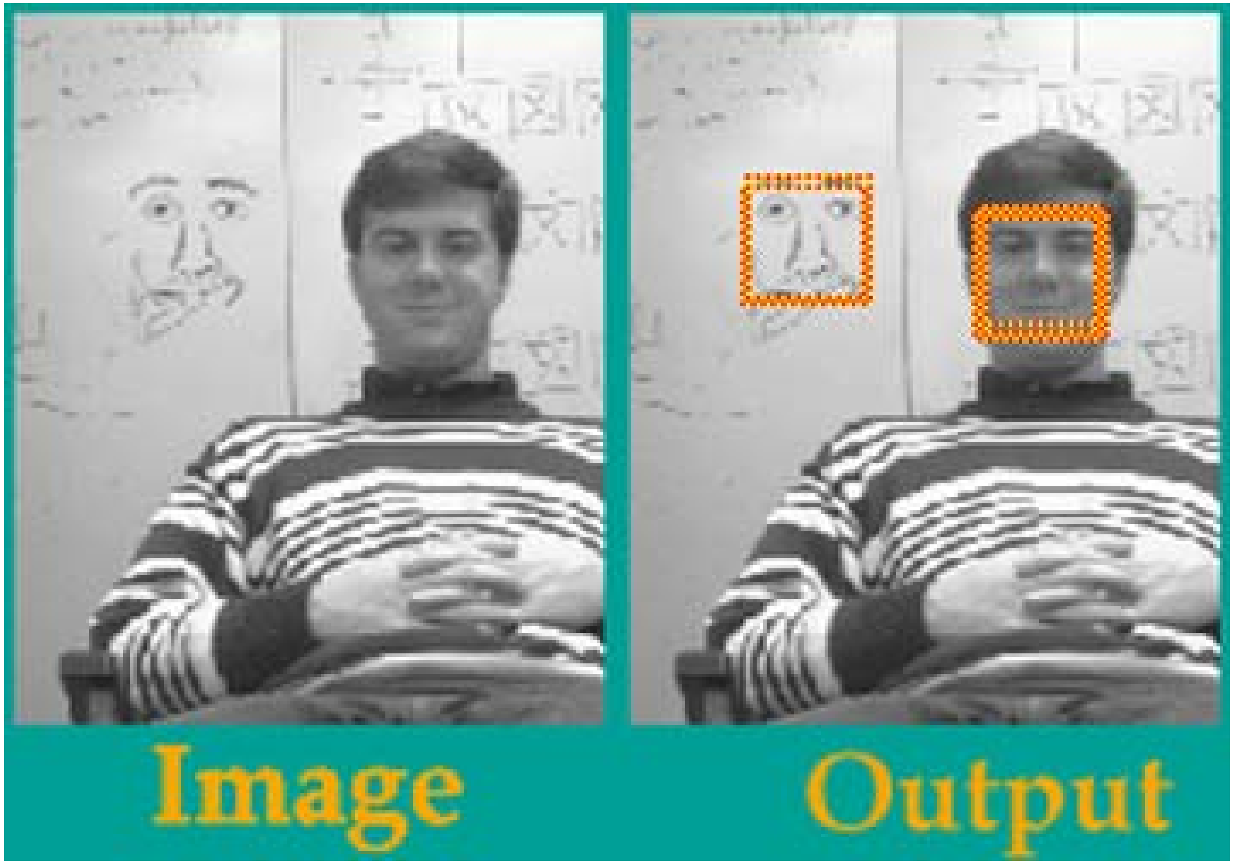




Training Database
1000+ Real, 3000+ *VIRTUAL*
50,0000+ Non-Face Pattern

Sung, Poggio 1995

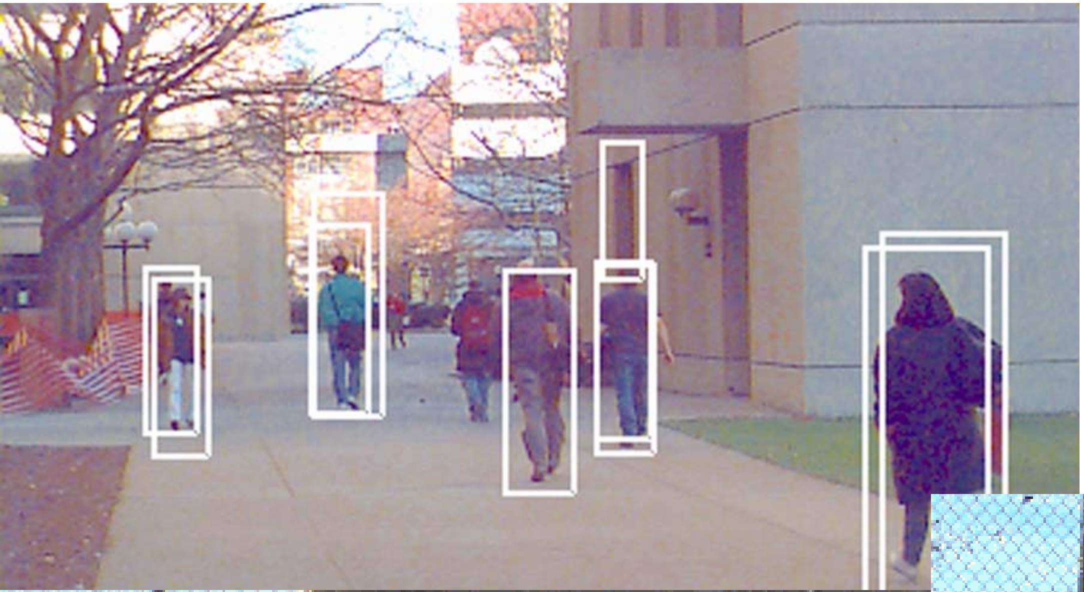
Learning Face Detection

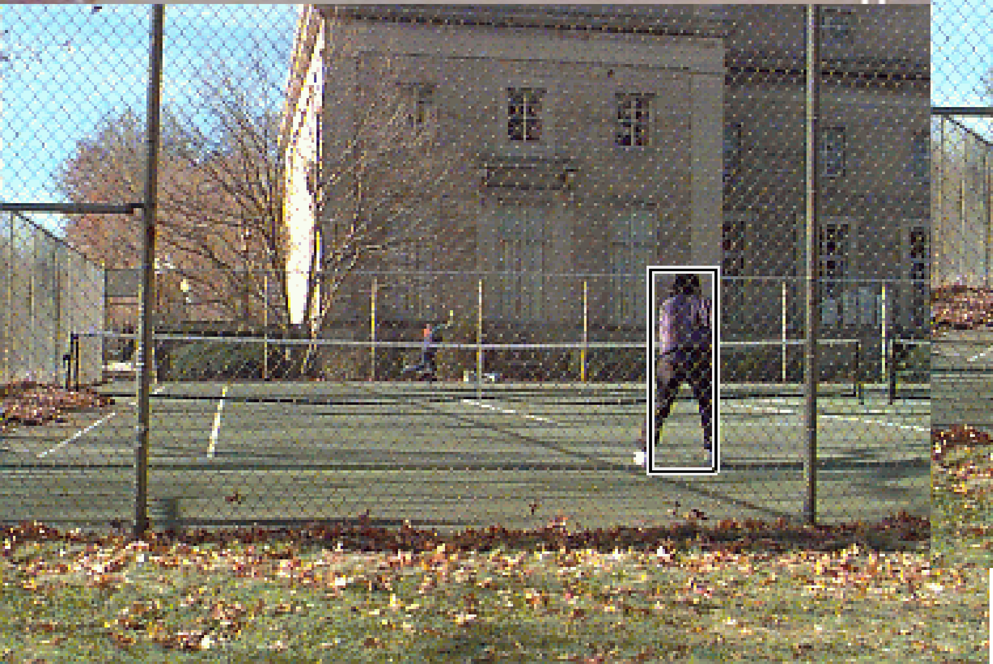


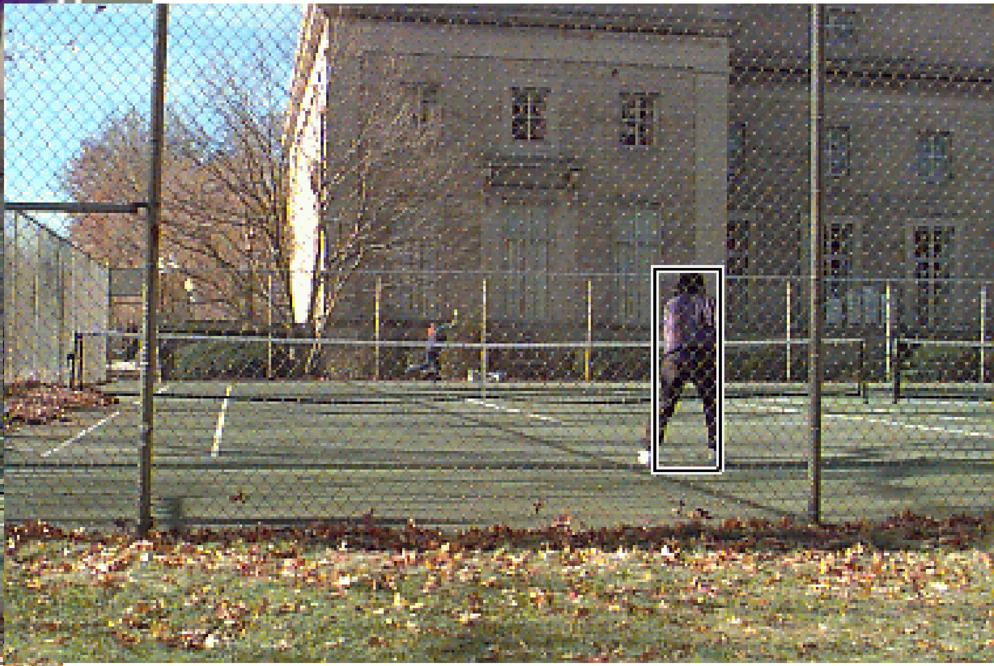
Sung, Poggio
1994

Face detection:...

Trainable System for Object Detection: Pedestrian detection - Results

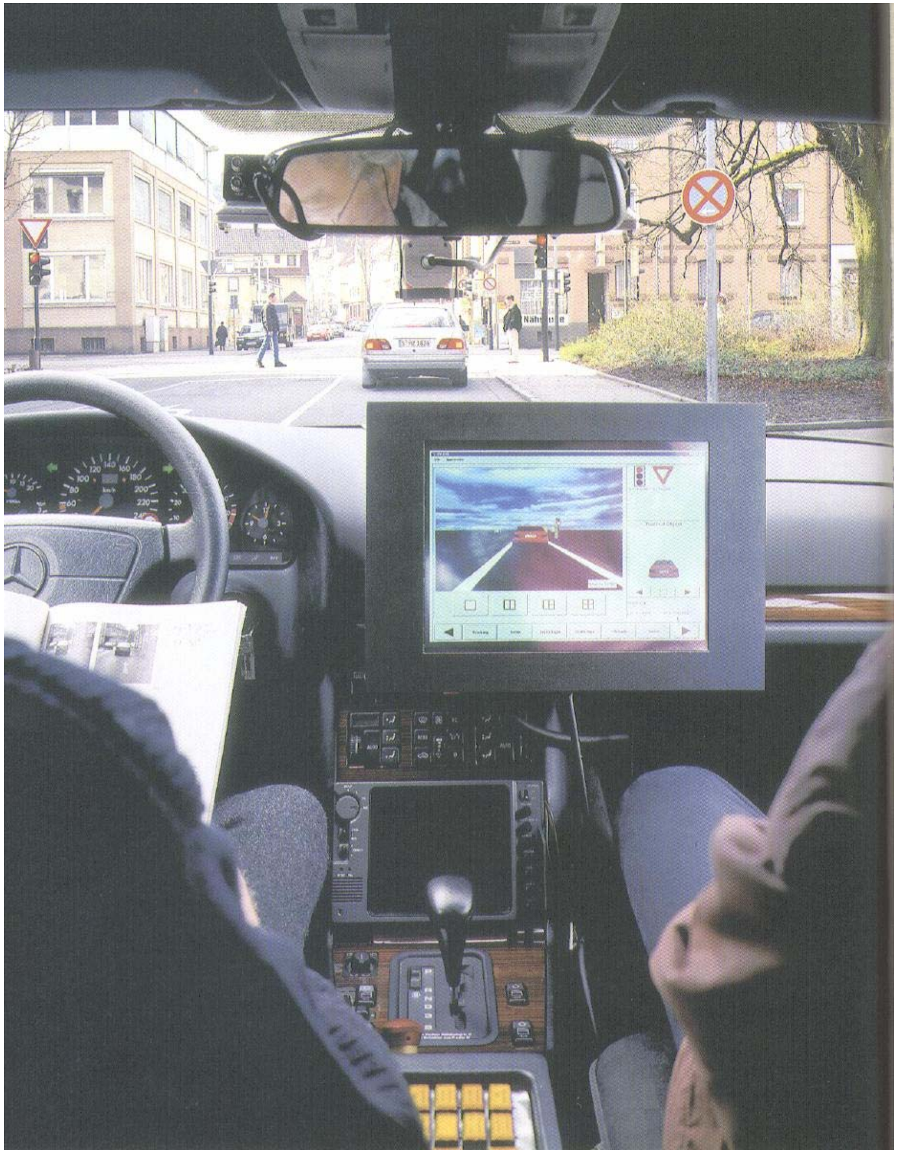




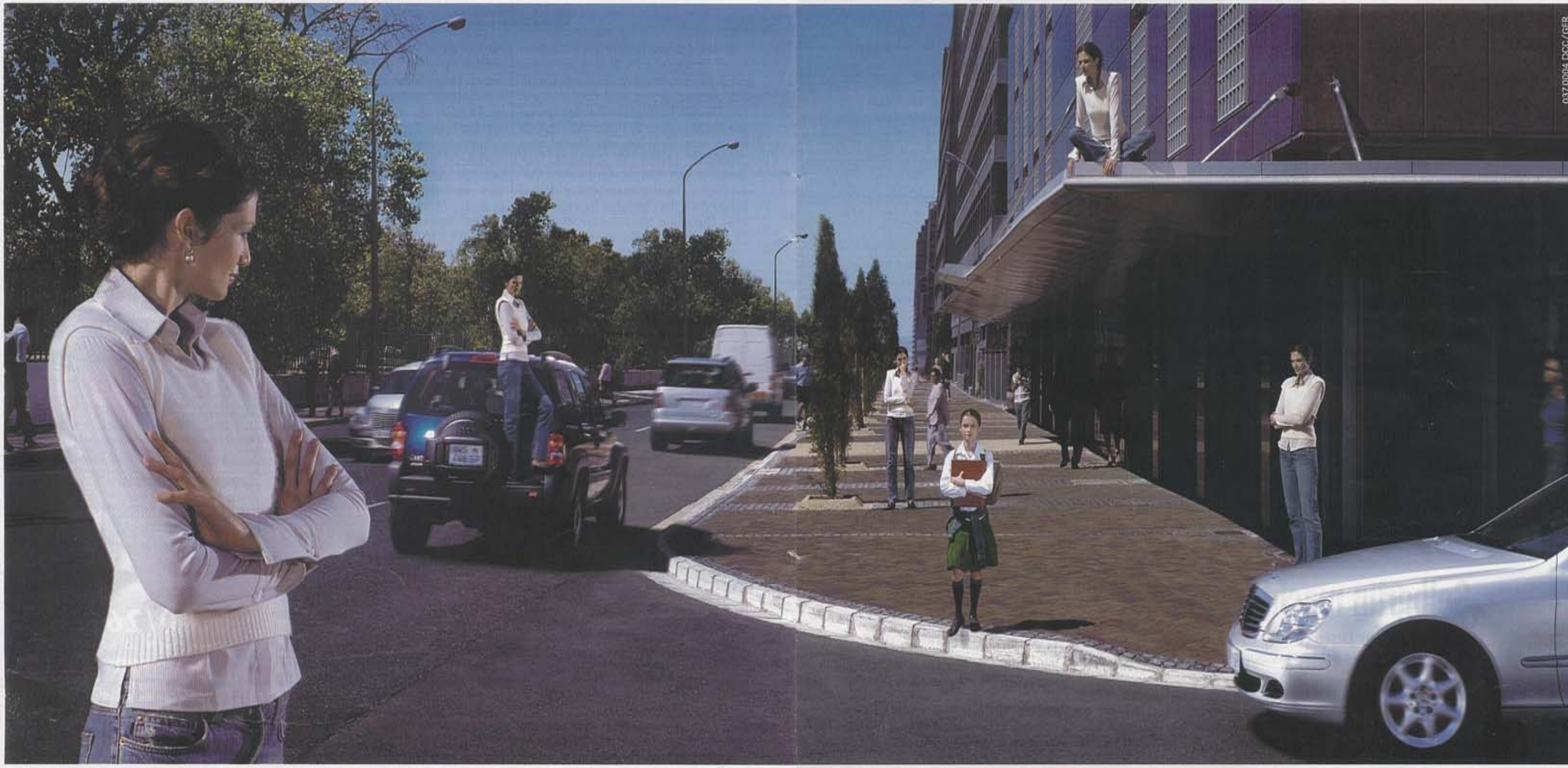


Papageorgiou and Poggio, 1998

The system was tested in a test car (Mercedes)



~10 year old CBCL computer vision work:
SVM-based pedestrian detection system in Mercedes test car...
now becoming a product (MobilEye)



Wir bringen unseren Autos das Sehen bei, weil eine Mutter nicht überall sein kann.

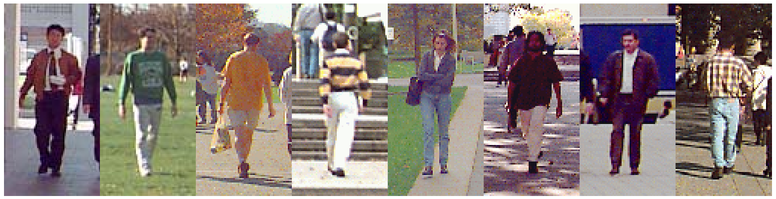
Eine Mutter kann ihre Kinder nicht immer beschützen. Besonders dann nicht, wenn sie alleine im Straßenverkehr unterwegs sind. Deshalb arbeiten wir an Fußgängererkennungs-Systemen für unsere Autos, die dem Fahrer helfen, Menschen auf der Straße schneller zu erkennen. Innerhalb von Bruchteilen einer Sekunde warnt das System den Fahrer, damit er besser reagieren kann. Diese intelligenten Technologien zur Vermeidung von Unfällen entwickelt die DaimlerChrysler Forschung schon heute. Für die Automobile von morgen.

Tiefere Einblicke in die Vision vom ‚Unfallfreien Fahren' erhalten Sie unter: www.daimlerchrysler.com

DaimlerChrysler

Answers for questions to come.

People classification/detection: training the system



1848 patterns

7189 patterns

Representation: overcomplete dictionary of Haar wavelets; high dimensional feature space (>1300 features)

pedestrian detection

Face classification/detection: training the system

. . .

. . .

Representation: grey levels (normalized) or overcomplete dictionary of Haar wavelets

face detection

Face identification: training the system

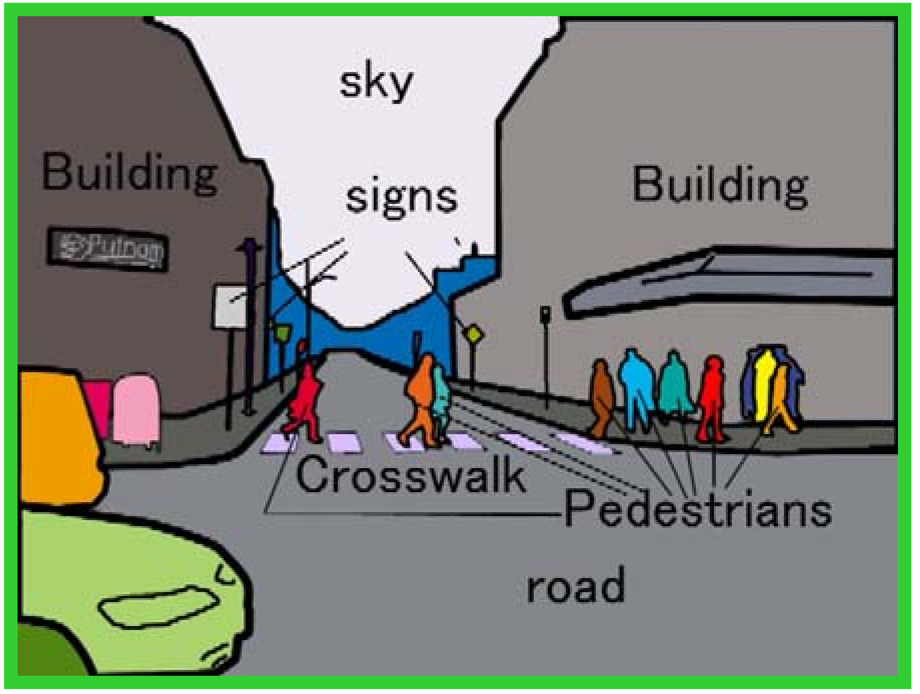
. . .

. . .

Representation: grey levels (normalized) or overcomplete dictionary of Haar wavelets

face identification

What about the model and computer vision?
The street scene project



Source: Bileschi, Wolf & Poggio

This was a project in computer vision
until we found out – as I already mentioned -- that
a separate neuroscience project
was giving us a very good system to solve recognition
problems of this type...more tomorrow in the
neuroscience day!

Learning from Examples: engineering applications

Bioinformatics
Artificial Markets
Object categorization
Object identification
Image analysis
Decoding the Neural Code
Graphics
Text Classification
.....

Another application:
using learning algorithms to *decrypt* the brain code

Chou Hung, Gabriel Kreiman, James DiCarlo, Tomaso Poggio,

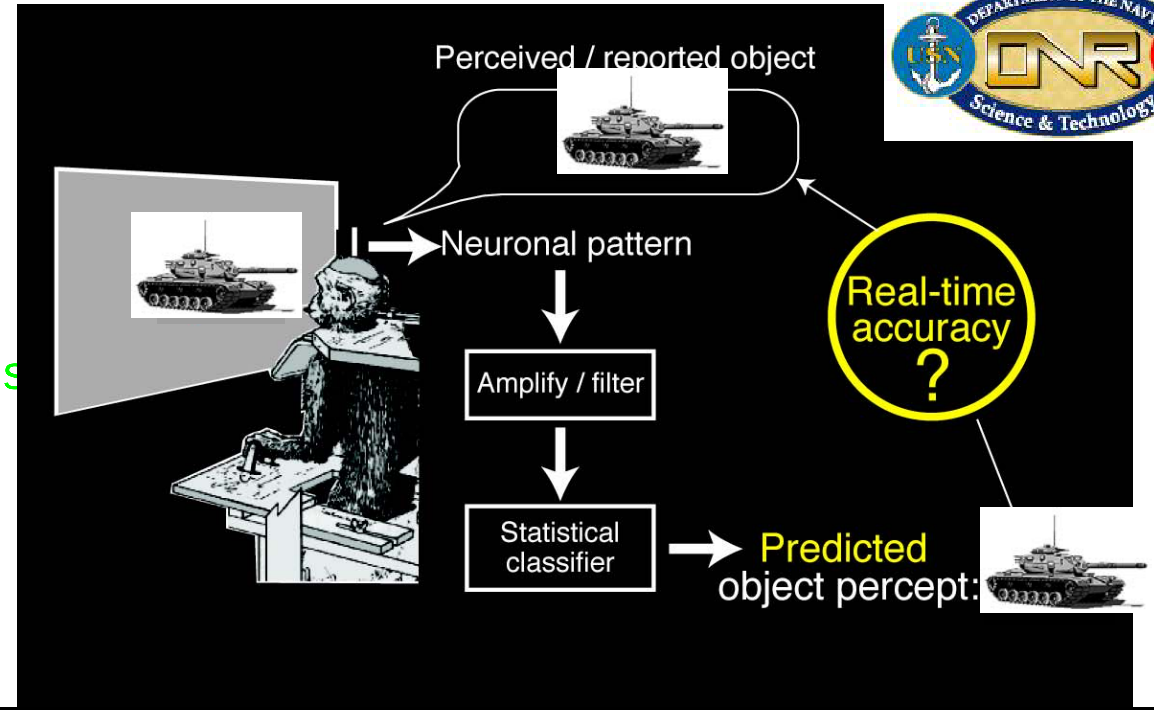
The McGovern Institute for Brain Research, Department of Brain Sciences
Massachusetts Institute of Technology, Cambridge MA

Science, Nov 4, 2005

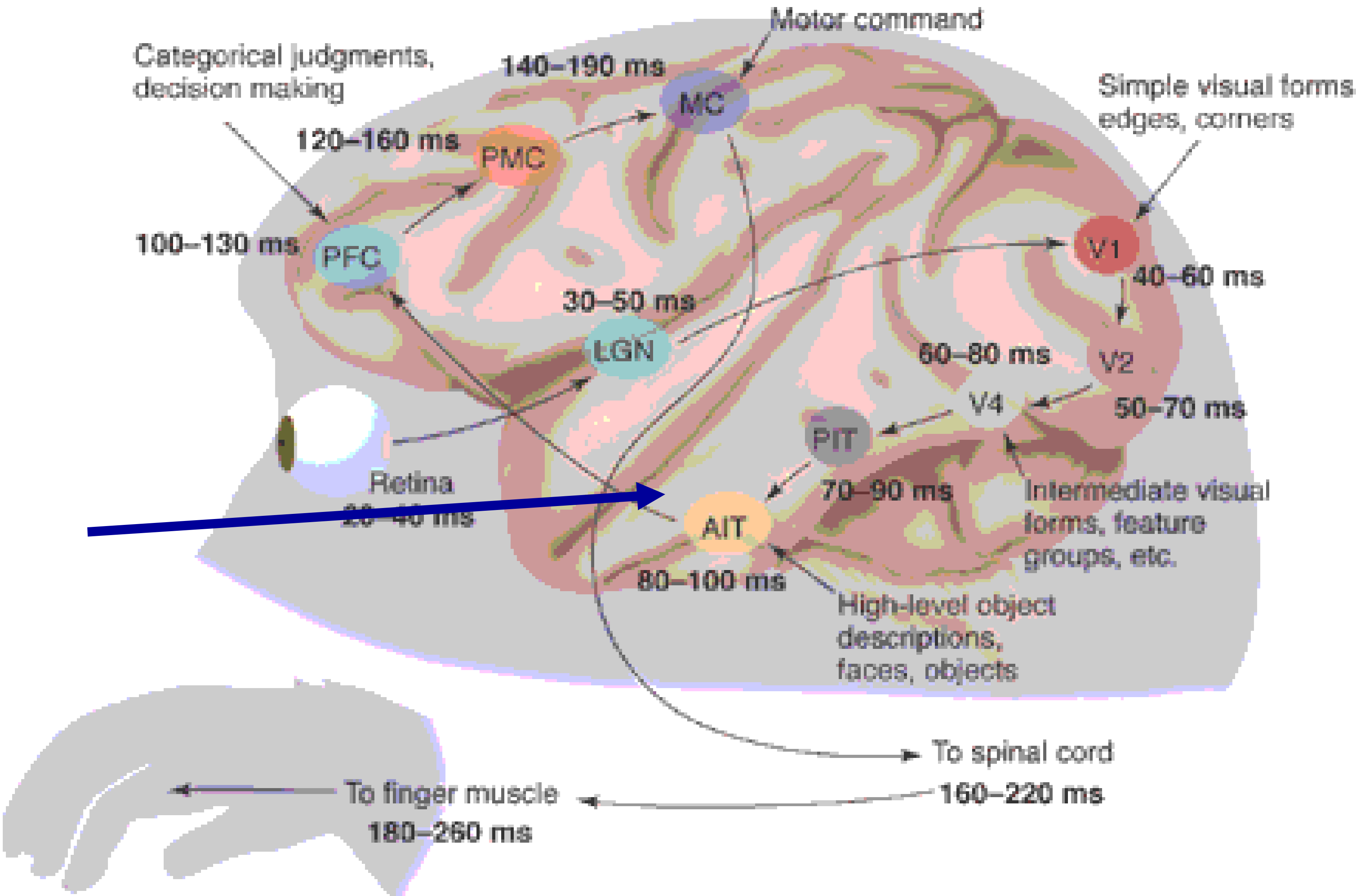
DARPA Neovision

Goal (analysis):

Can we “read-out” the subject’s object percept?

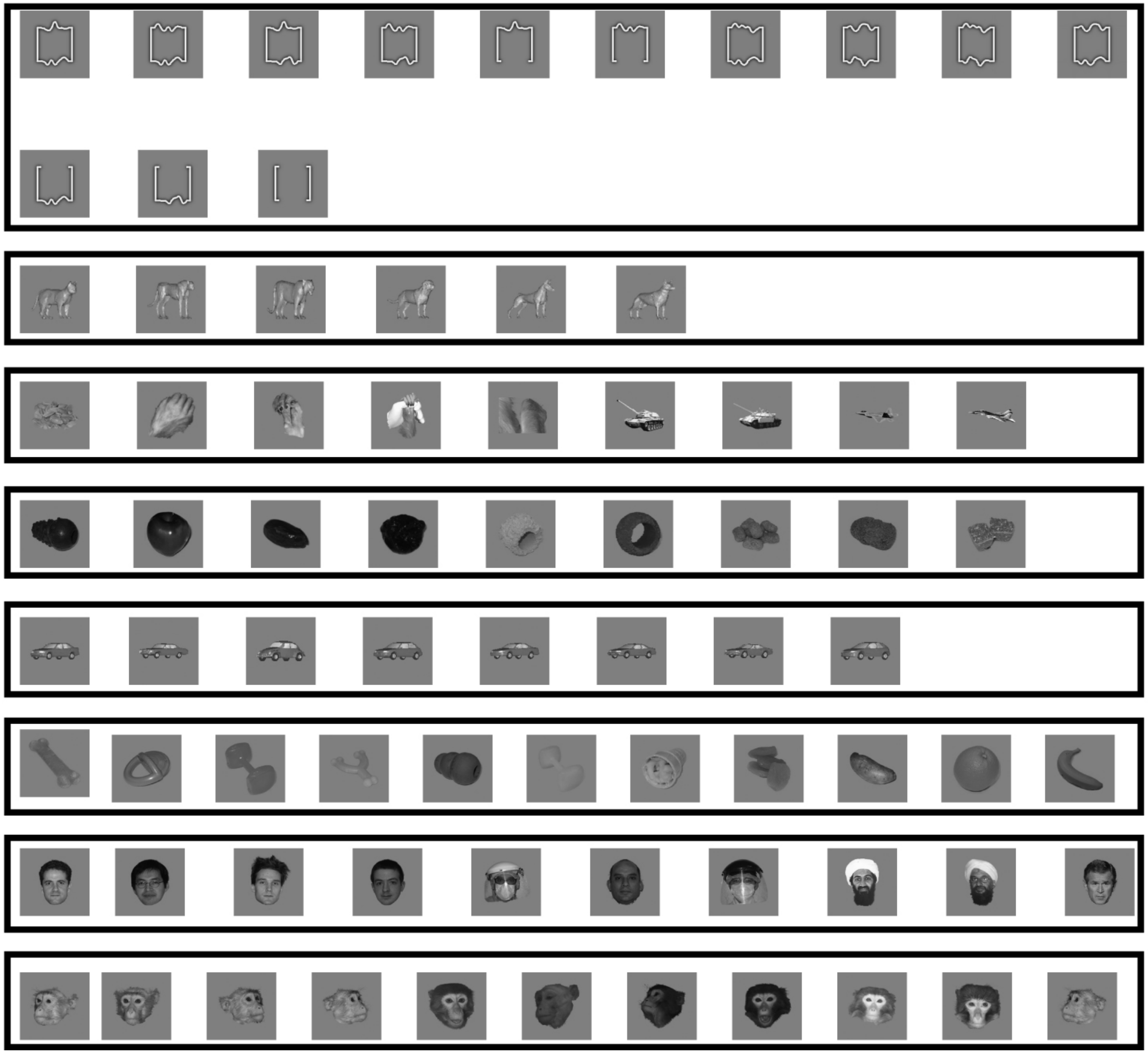


The end station of the ventral stream in visual cortex is IT



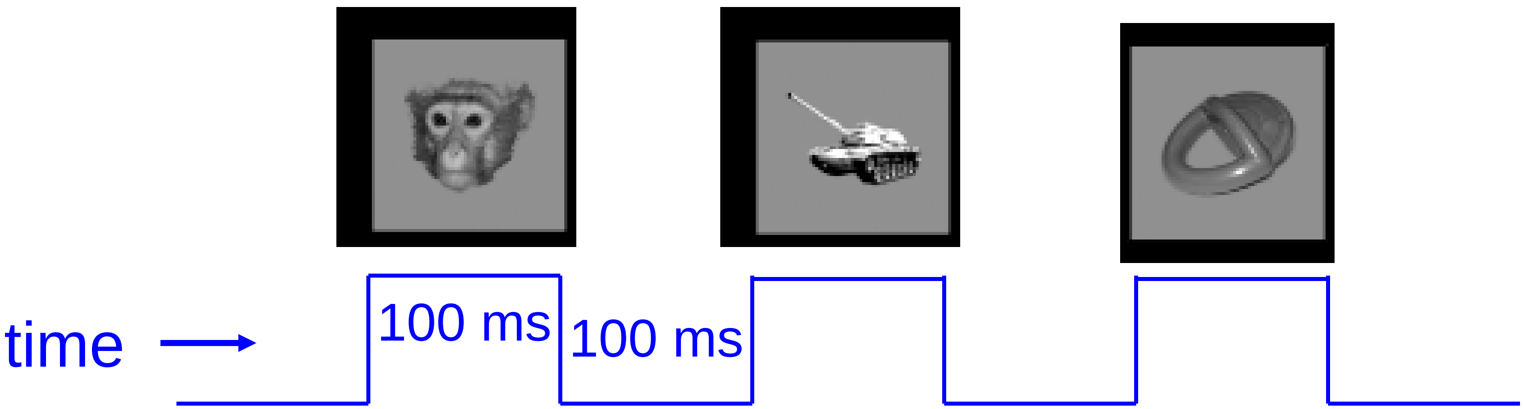
Reading-out the neural code in AIT

77 objects,
8 classes



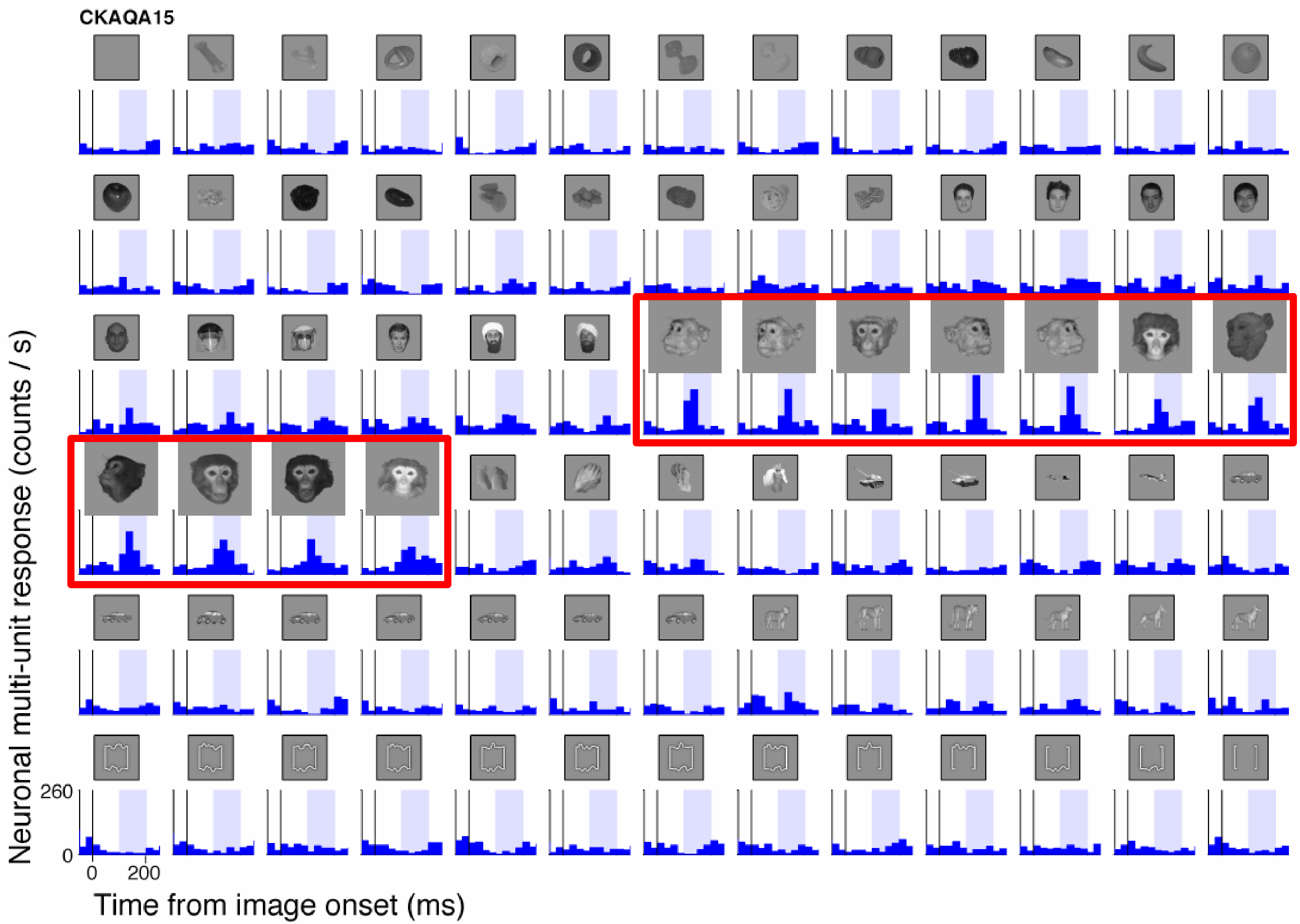
Chou Hung, Gabriel Kreiman, James DiCarlo, Tomaso Poggio, *Science, Nov 4, 2005*

Recording at each recording site during passive viewing

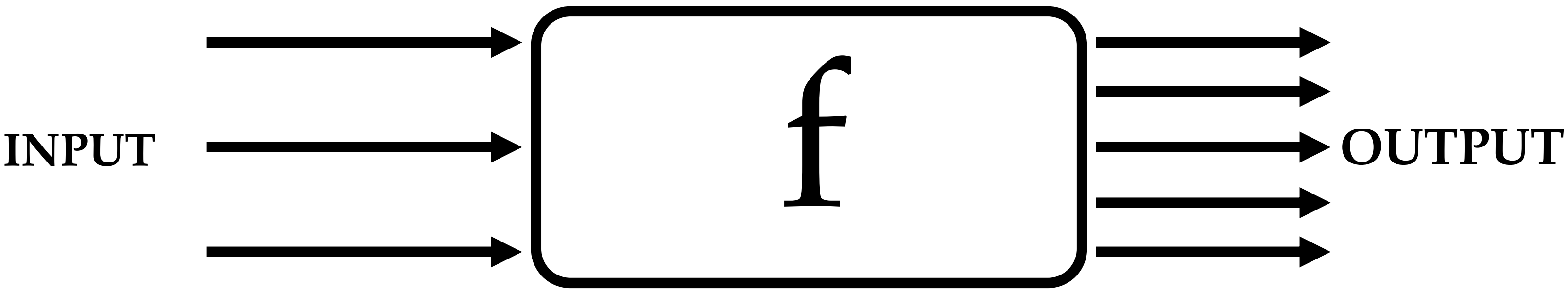


- 77 visual objects
- 10 presentation repetitions per object
- presentation order randomized and counter-balanced

Example of one AIT cell



Training a classifier on neuronal activity.



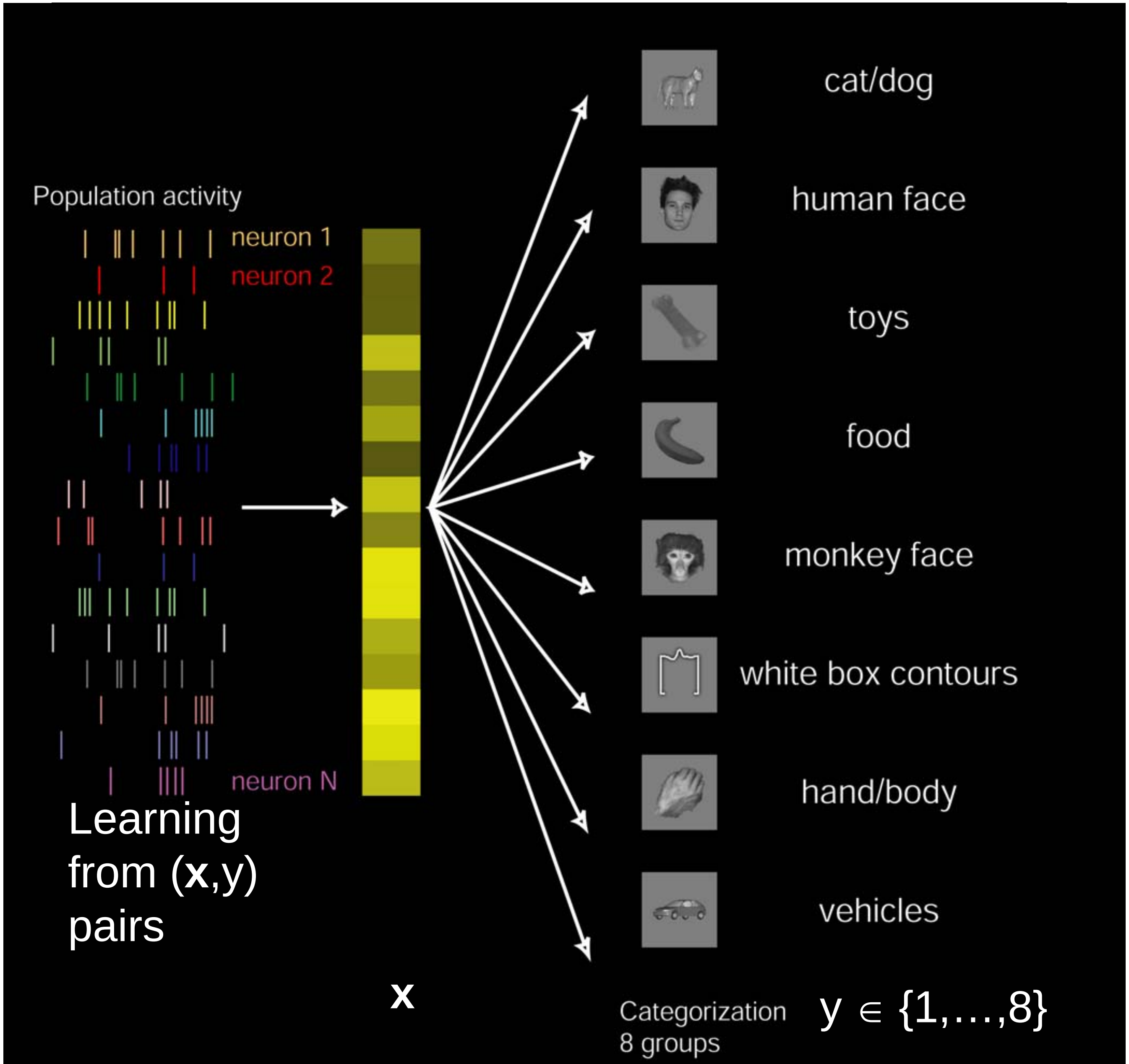
From a set of data (vectors of activity of n neurons (x) and object label (y)

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

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*

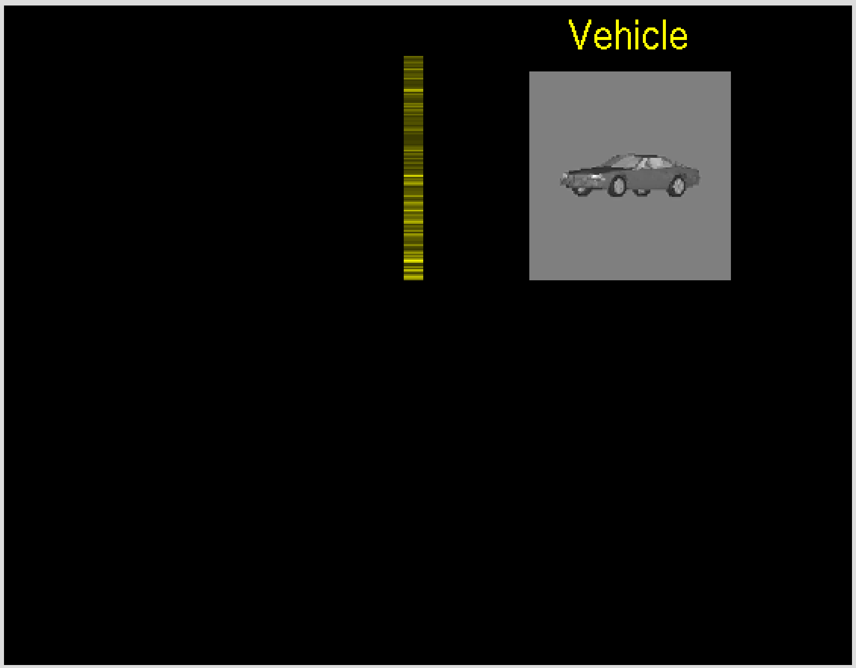
Decoding the neural code ...
population response (using a classifier)



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

Neuronal population activity

Classifier prediction



Categorization

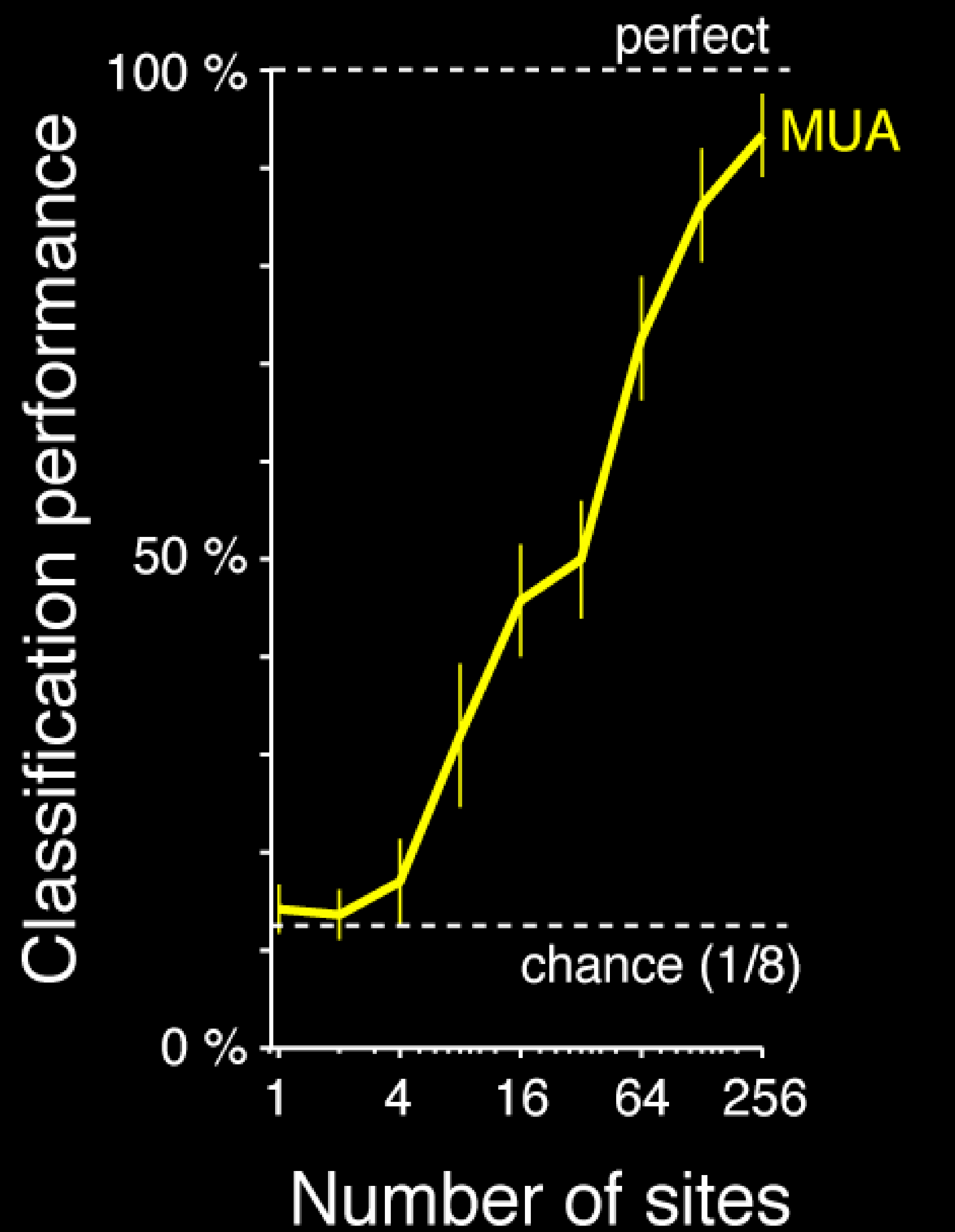
- Toy
- Body
- Human Face
- Monkey Face
- Vehicle
- Food
- Box
- Cat/Dog

Video speed: 1 frame/sec
Actual presentation rate: 5 objects/sec

We can decode the brain's code and read-out
from the cortex
(as from the model, see later)

Results:

reliable object categorization using ~100 <u>arbitrary</u> AIT sites



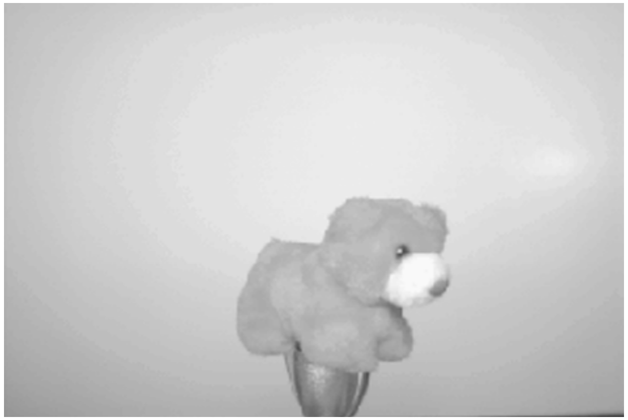
- *[100-300 ms] interval*
- *50 ms bin size*

Learning from Examples: engineering applications

Bioinformatics
Artificial Markets
Object categorization
Object identification
Image analysis
Image synthesis, eg Graphics
Text Classification
.....

Image Analysis

⇒ **Bear (0° view)**



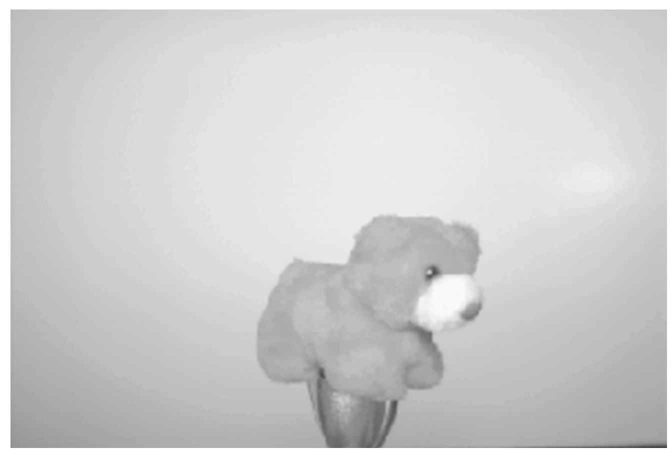
⇒ **Bear (45° view)**

Image Synthesis

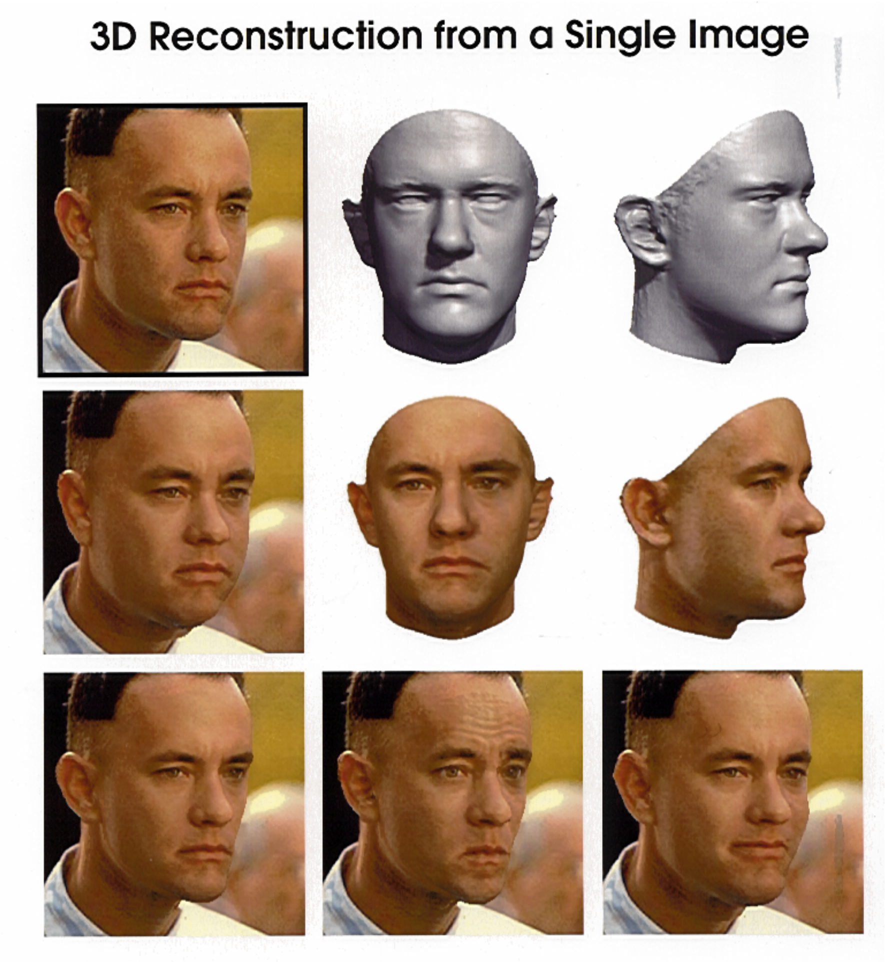
UNCONVENTIONAL GRAPHICS

Θ = 0° view $\Rightarrow$

Θ = 45° view $\Rightarrow$



Reconstructed 3D Face Models from 1 image



Blanz and Vetter,
MPI
SigGraph '99

Reconstructed 3D Face Models from 1 image

Blanz and Vetter,
MPI
SigGraph '99

Vermeer,
Tischbein,
raffaello,
Hopper

V. Blanz, C. Basso,
T. Poggio
and
T. Vetter, 2003

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

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

Trainable Videorealistic Face Animation

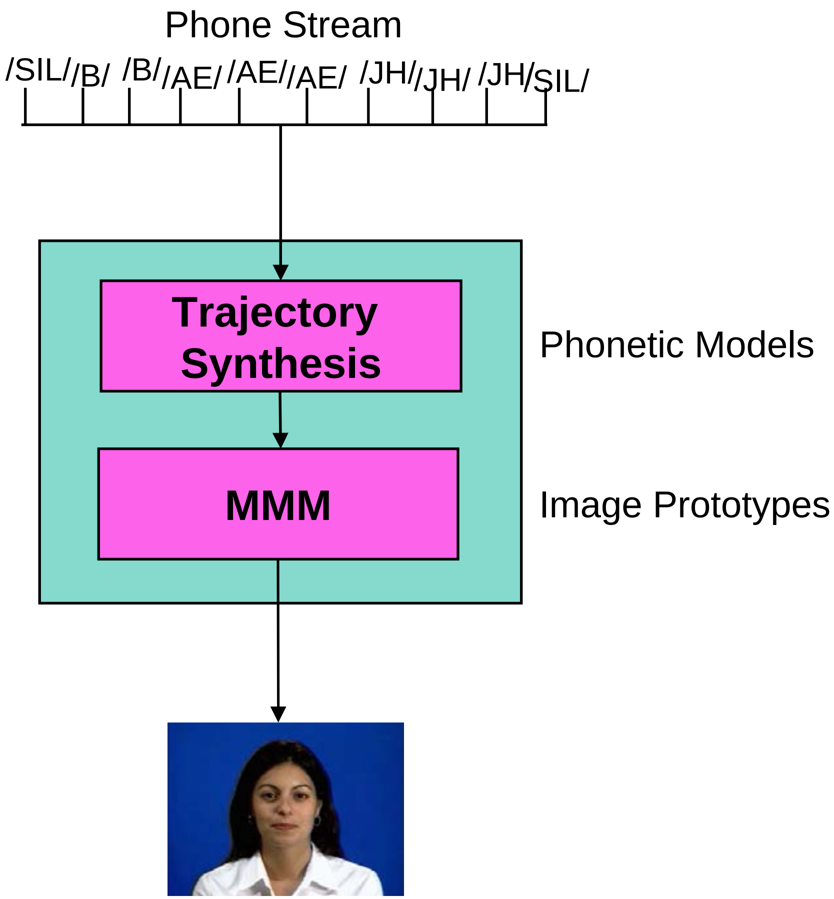
1. Learning

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

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

2. Run Time

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



Movies

Marylin,
Rehema

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

We assessed the realism of the talking face with psychophysical experiments.

Data suggest that the system passes a visual version of the Turing test.

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<.

Overview of overview

- The problem of supervised learning: “real” math behind it
- Examples of engineering applications (from our group)
- Learning and the brain

Learning how the brain works

This is the old dream of all philosophers
and more recently of AI:

understand how the brain works,
make intelligent machines

Hopes

- ❑ Neuroscience may be beginning to understand how a part of cortex works, in terms of its information processing

- ❑ As a consequence, we begin to develop software programs that mimic the ability of people to recognize complex images and understand sounds

- ❑ Will neuroscience determine future development of a new AI?

Some numbers

Human Brain

$10^{11} \ldots 10^{12}$ neurons (1 million flies ☺)

10^{14}- 10^{15} synapses

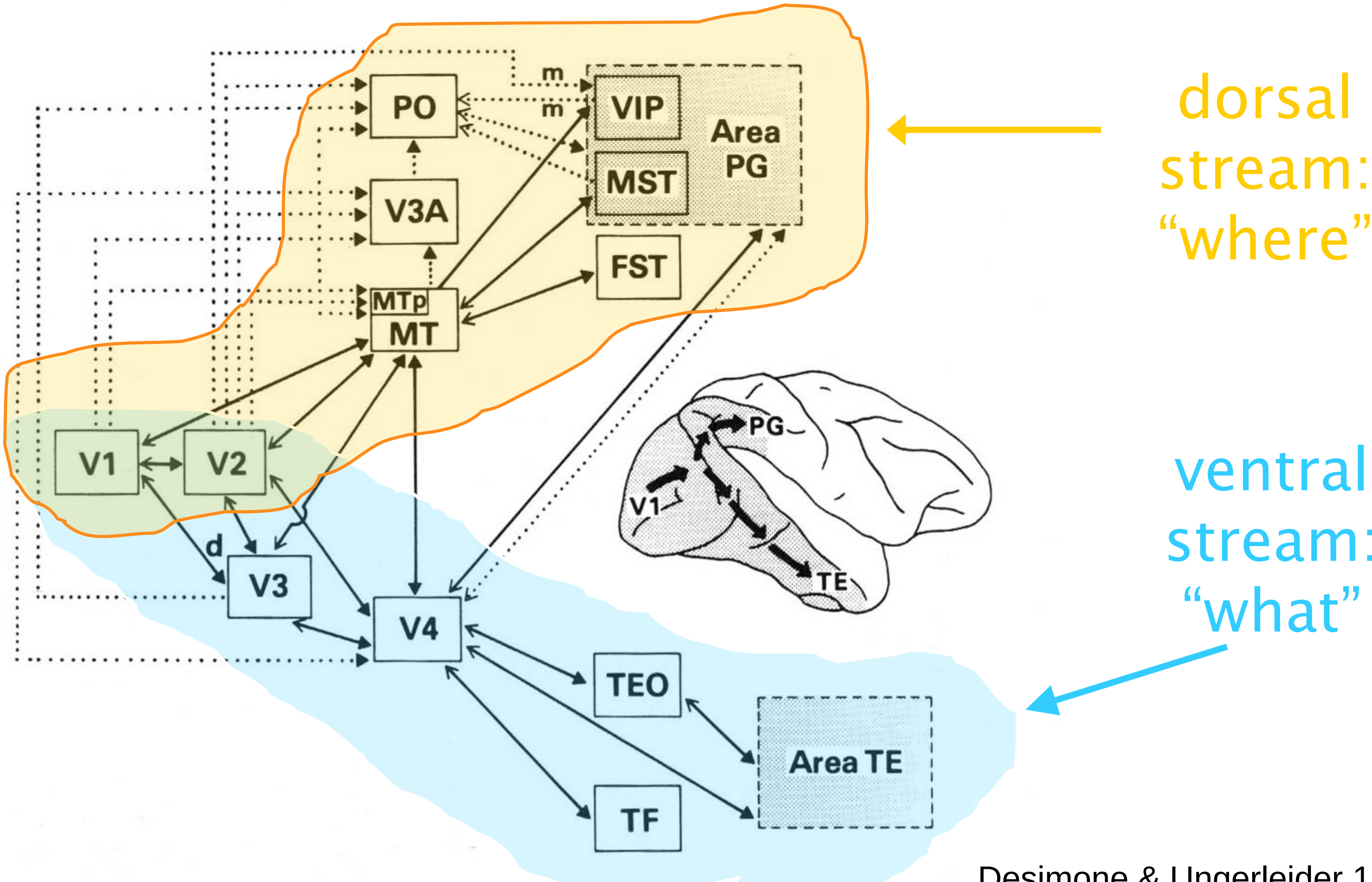
Neuron

Fundamental space dimension: fine dendrites : 0.1 μ diameter; lipid bylayer membrane : 5 nm thick; specific proteins : pumps, channels, receptors, enzymes

Fundamental time length : 1 msec

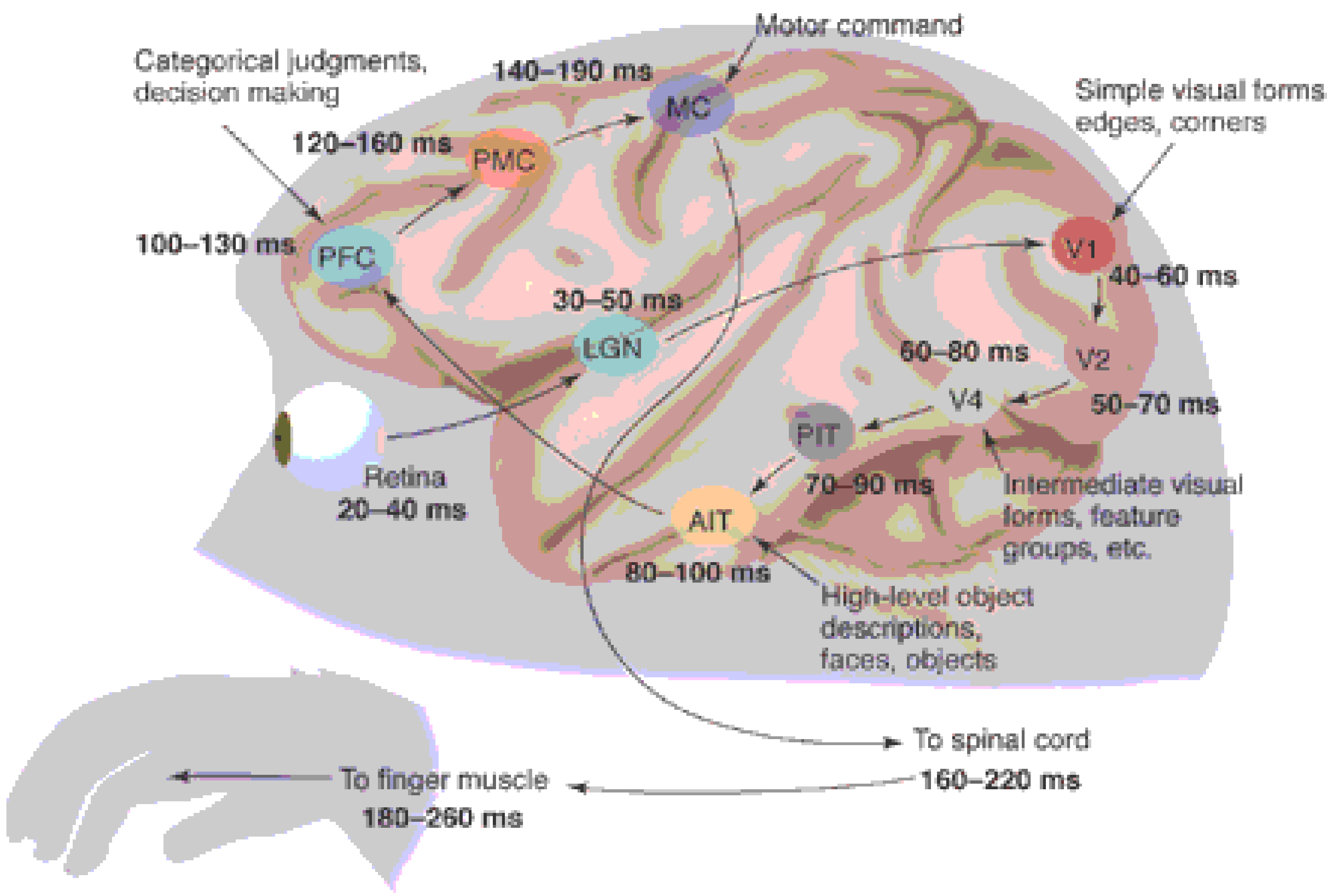
The problem: recognition in natural images (e.g., “is there an animal in the image?”)

How does visual cortex solve this problem?
How can computers solve this problem?



Desimone & Ungerleider 1989

Learning to recognize objects and the ventral stream in visual cortex

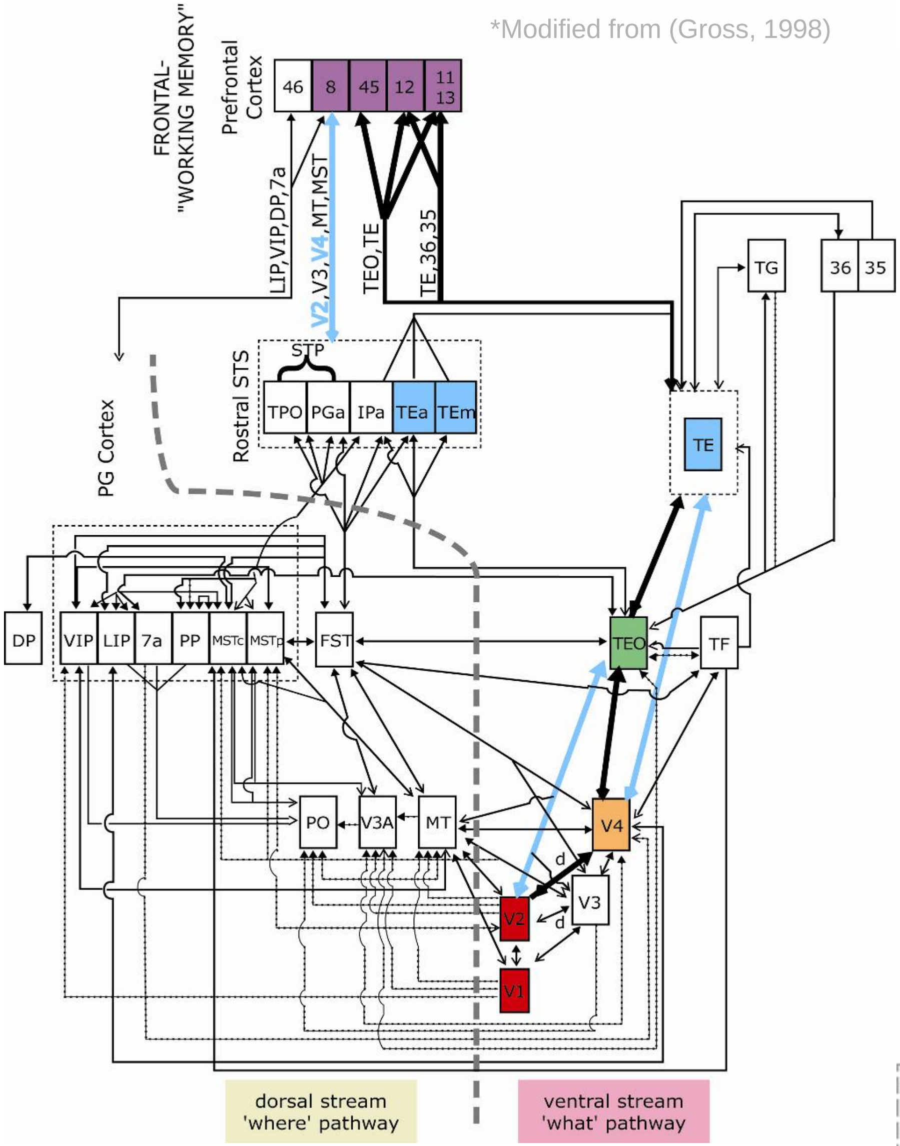


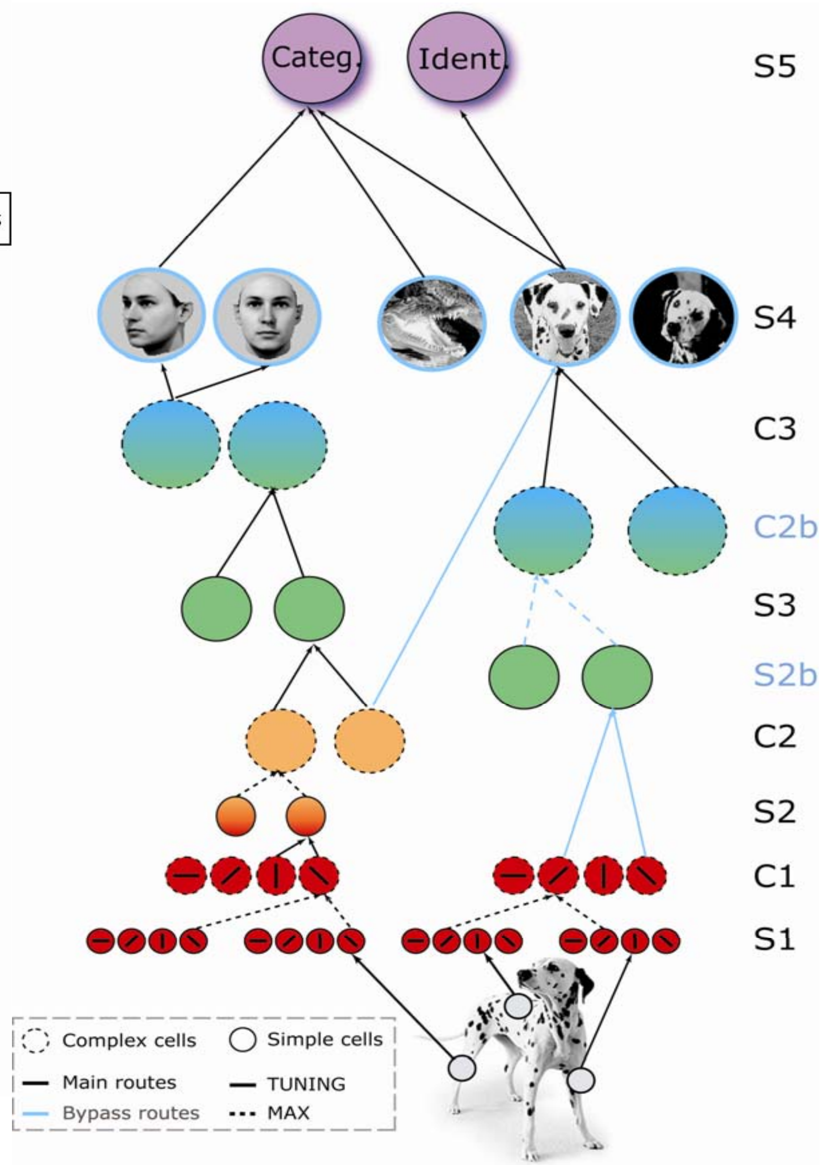
A “feedforward” version of the problem: rapid categorization

SHOW RSVP
MOVIE

Biederman 1972; Potter 1975; Thorpe et al 1996

A model of the ventral stream, which is also an algorithm...





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

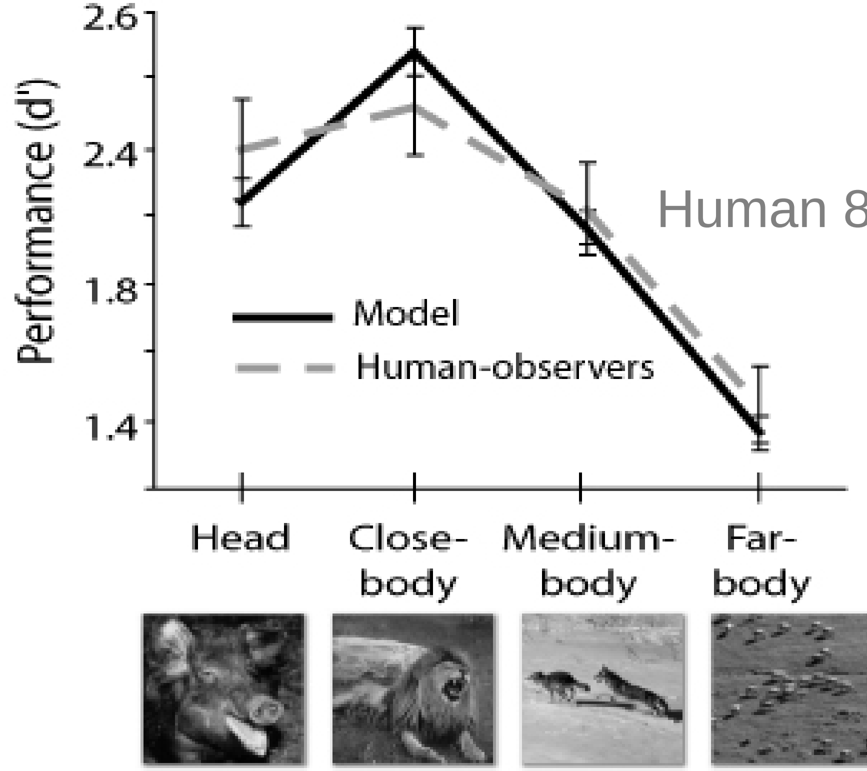
[software available online]

..."solves" the problem

(if the mask forces feedforward processing)...

- d'~ standardized error rate
- the higher the d', the better the performance

Model 82%



Human 80%

Serre Oliva & Poggio 2007

Extensive comparison w| neural data

- V1:
  - Simple and complex cells tuning (Schiller et al 1976; Hubel & Wiesel 1965; Devalois et al 1982)
  - MAX operation in subset of complex cells (Lampl et al 2004)
- V4:
  - Tuning for two-bar stimuli (Reynolds Chelazzi & Desimone 1999)
  - MAX operation (Gawne et al 2002)
  - Two-spot interaction (Freiwald et al 2005)
  - Tuning for boundary conformation (Pasupathy & Connor 2001, Cadieu et al., 2007)
  - Tuning for Cartesian and non-Cartesian gratings (Gallant et al 1996)
- IT:
  - Tuning and invariance properties (Logothetis et al 1995)
  - Differential role of IT and PFC in categorization (Freedman et al 2001, 2002, 2003)
  - Read out data (Hung Kreiman Poggio & DiCarlo 2005)
  - Pseudo-average effect in IT (Zoccolan Cox & DiCarlo 2005; Zoccolan Kouh Poggio & DiCarlo 2007)
- Human:
  - Rapid categorization (Serre Oliva Poggio 2007)
  - Face processing (fMRI + psychophysics) (Riesenhuber et al 2004; Jiang et al 2006)

(Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005)

an unusual, <u>hierarchical</u> architecture
with unsupervised and supervised learning
and learning of invariances…

Notices of the American Mathematical Society (AMS), Vol. 50, No. 5, 537-544, 2003.

The Mathematics of Learning: Dealing with Data

Tomaso Poggio and Steve Smale

How then do the learning machines described in the theory compare with brains?

- One of the most obvious differences is the ability of people and animals to learn from very few examples.

- A comparison with real brains offers another, 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?

- Why hierarchies? For instance, the lowest levels of the hierarchy may represent a dictionary of features that can be shared across multiple classification tasks.

- There may also be the more fundamental issue of *sample complexity*. Thus our ability of learning from just a few examples, and its limitations, may be related to the hierarchical architecture of cortex.

Formalizing the hierarchy: towards a theory

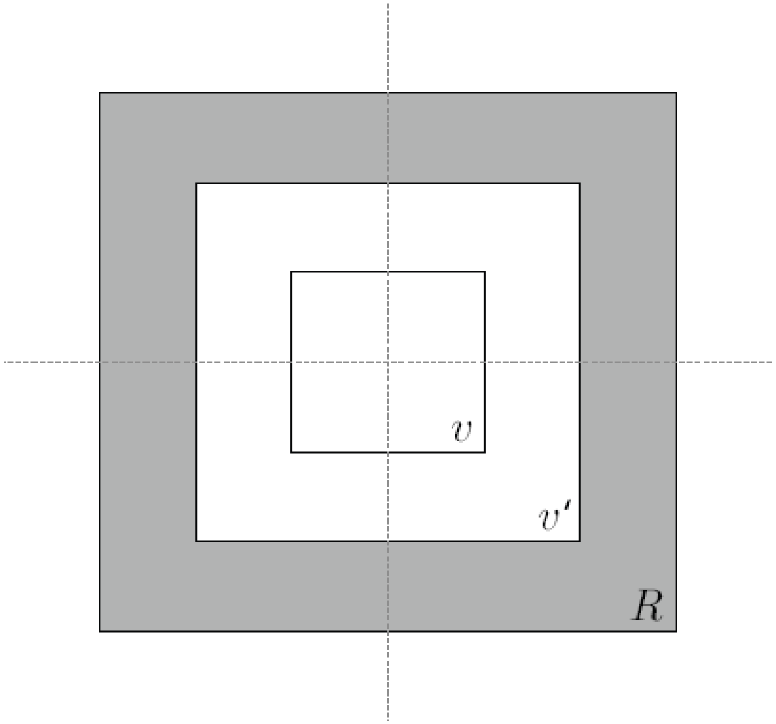


Figure 1: *Nested d*

Smale, S., T. Poggio, A. Caponnetto, and J. Bouvrie. Derived Distance: towards a mathematical theory of visual cortex, *CBCL Paper*, Massachusetts Institute of Technology, Cambridge, MA, November, 2007.

Axiom: $f \circ h : v \to [0,1]$ is in $Im(v)$ if $f \in Im(v')$ and $h \in H$, that is *the restriction of an image is an image* and similarly for H'. Thus

$$f \circ h : v \to [0,1] \in Im(v) \text{ if } f \in Im(v') \text{ and } h \in H,$$
$$f \circ h' : v' \to [0,1] \in Im(v') \text{ if } f \in Im(R) \text{ and } h' \in H'.$$

It is just possible that the brain ....

…will tell us more learning theory!