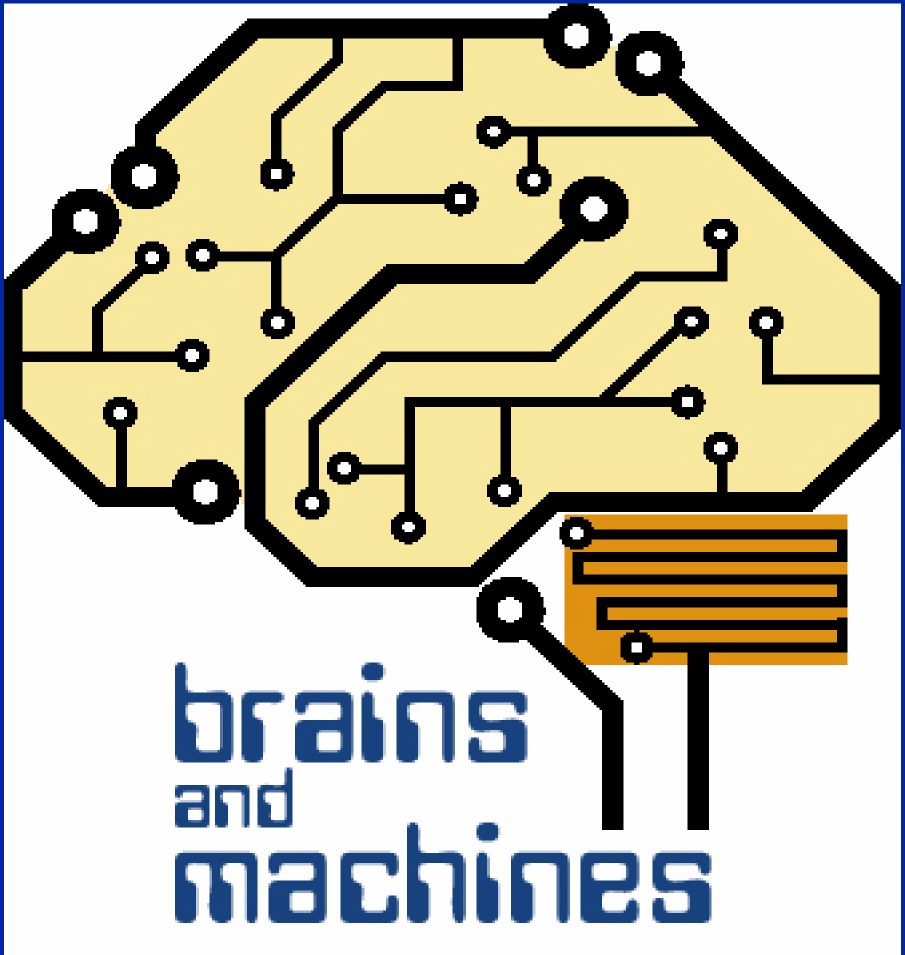
9.520

Statistical Learning Theory and Applications

Sasha Rakhlin and Andrea Caponnetto and Ryan Rifkin + tomaso poggio

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

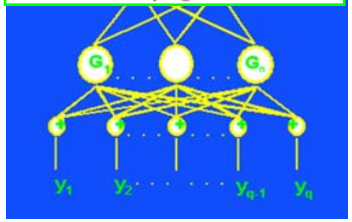
- Plasticity and learning have a central stage in the neurosciences

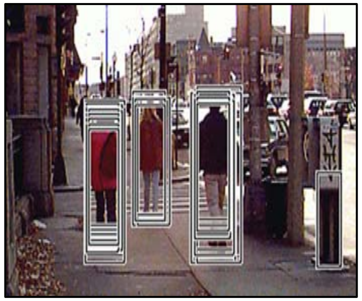
- Until now math and engineering of learning has developed independently of neuroscience…but it may begin to change: we will see the example of learning+computer vision…

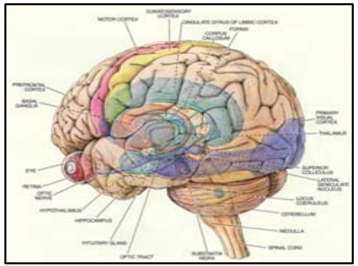
Learning: math, engineering, neuroscience

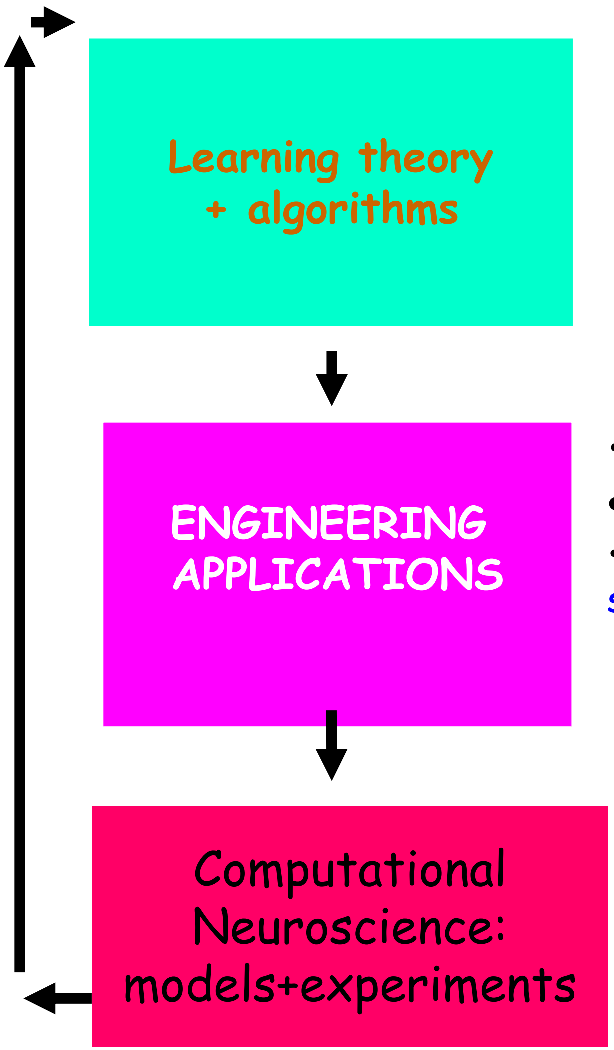
$$\min_{f\in H}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} V(y_i, f(x_i)) + \mu\|f\|_K^2\right]$$

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









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!
mathcamps? Monday late afternoon?

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

Slides on the Web site
Staff mailing list is 9.520@mit.edu
Student list is 9.520-students@mit.edu
Please fill form!

9.520 Statistical Learning Theory and Applications

Class 24: Project presentations

2:30—2:45 "Adaboosting SVMs to recover motor behavior from motor data", Neville Sanjana

2:45-3:00 "Review of Hierarchical Learning", Yann LeTallec

3:00—3:15 "An analytic comparison between SVMs and Bayes Point Machines", Ashis Kapoor

3:15-3:30 "Semi-supervised learning for tree-structured data", Charles Kemp

3:30—3:45 “Unsupervised Clustering with Regularized Least Square classifiers" - Ben Recht

3:40—3:50 "Multi-modal Human Identification." Brian Kim

3:50—4:00 "Regret Bounds, Sequential Decision-Making and Online Learning", Sanmay Das

9.520 Statistical Learning Theory and Applications

Class 25: Project presentations

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

4. (suggested by steve smale) Approximate indicator functions with kernels from a RKHS with very little smoothness. Calculate approx and sample error using bounds such as Cucker Smale etc.. Verify with computer simulations.

5. (also suggested by steve smale) Do careful proof – mimicking theorem 4 in CS p. 37 – that the RKHS defined for unbounded domains through the Mercer-like Fourier representation (Girosi) is the same as the RKHS define through the r.k. without Fourier.

6. (suggested by M. Bertero) Use L_2 compactness of monotonic functions for regularizing density estimation ?

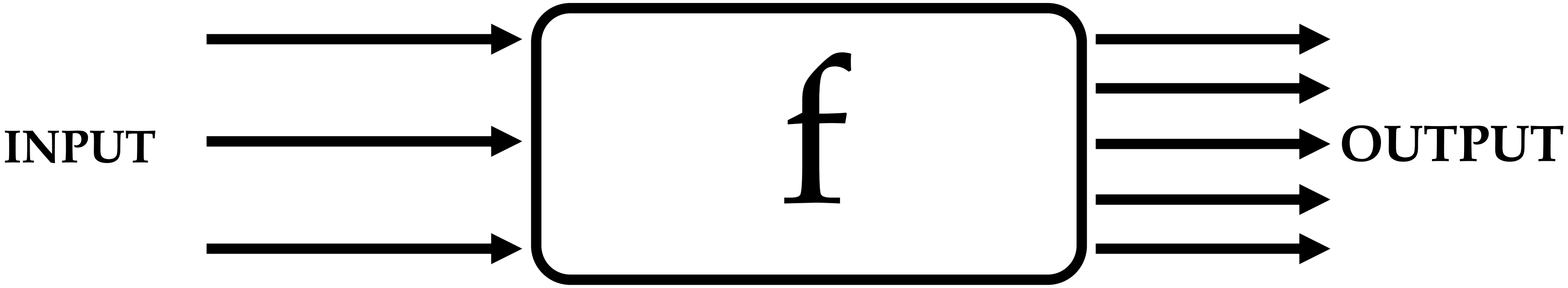
Overview of overview

- The problem of supervised learning: “real” math behind it

- Examples of engineering applications (from our group)

- Learning and the brain (example of object recognition)

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



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

Question: find function *f* such that

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

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

Reason for you to know theory

We will speak today and later about applications…

they are not simply using a black box. The best ones are about the right formulation of the problem (choice of representation (inputs, outputs), choice of examples, validate predictivity, do not datamine)

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

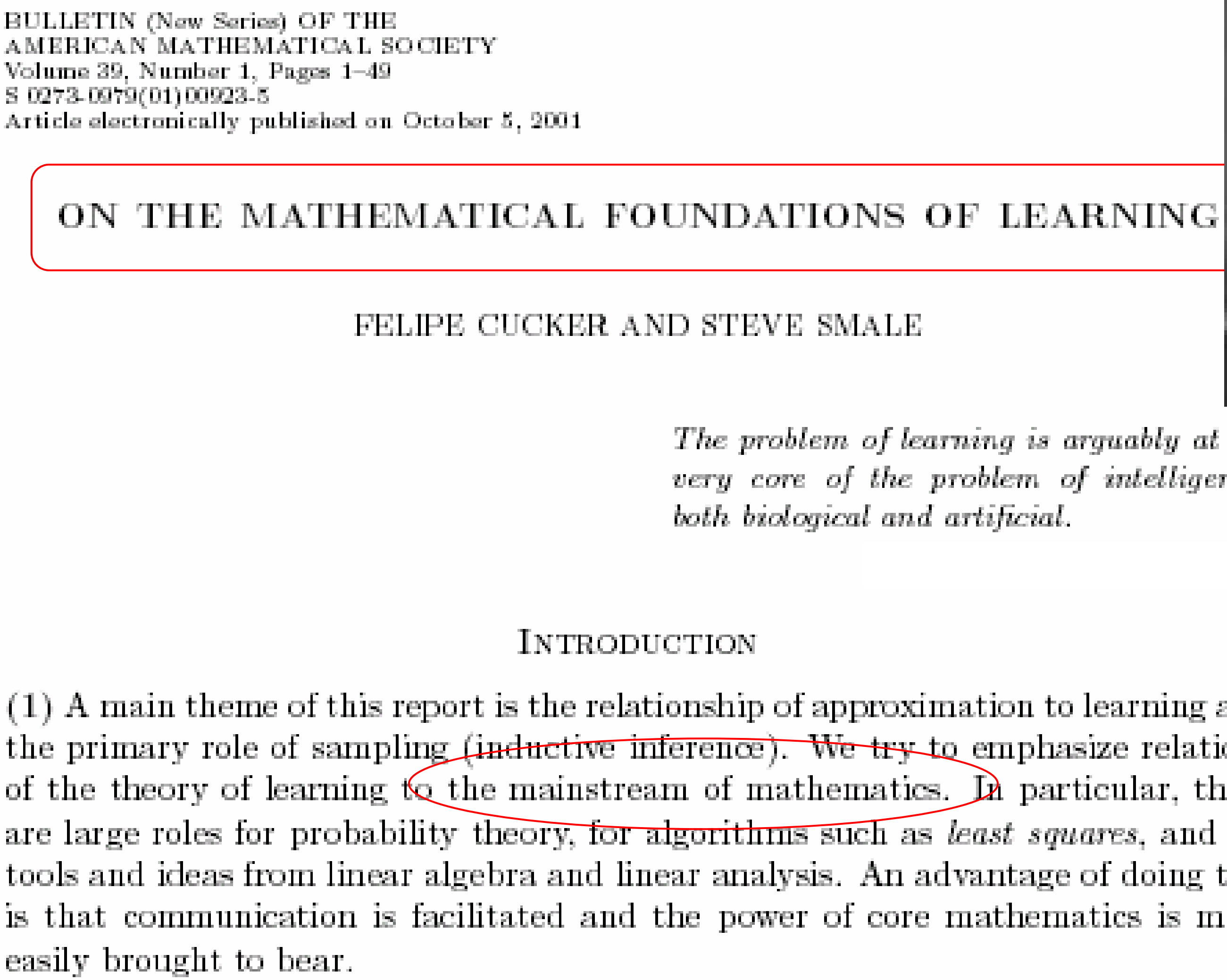
Notes

Two strands in learning theory:

❑ Bayes, graphical models…

❑ Statistical learning theory, regularization (closer to classical math, functional analysis+probability theory+empirical process theory…)

Interesting development: the theoretical foundations of learning are becoming 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

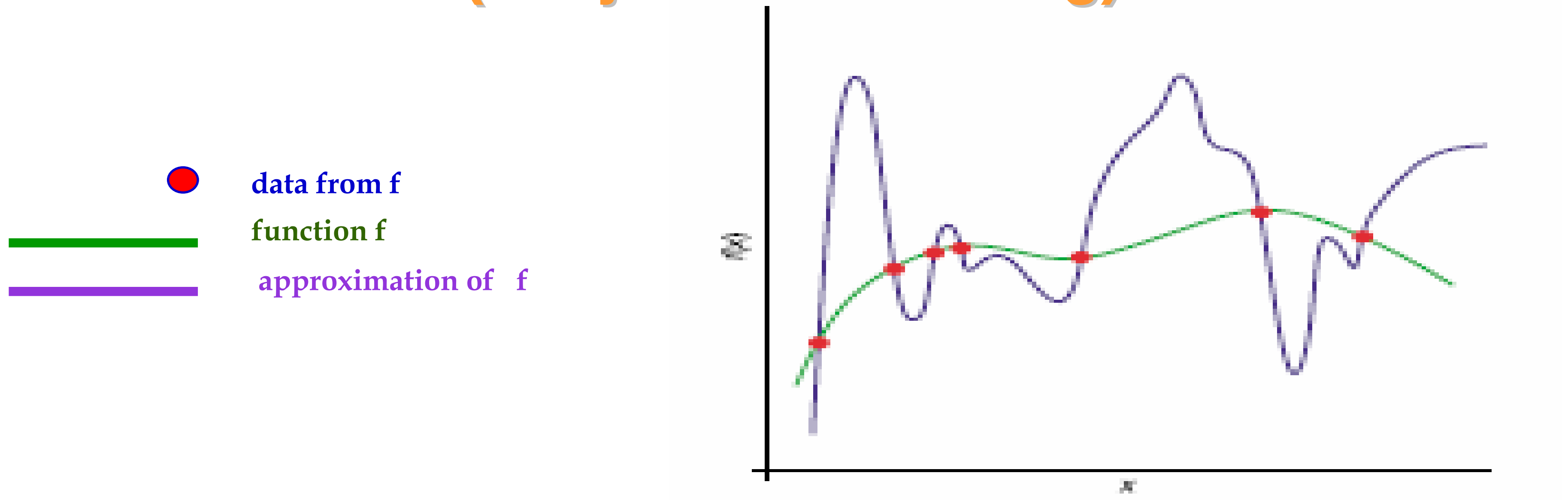
FELIPE CUCKER AND STEVE SMALE

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 learning theory) to solve the problem of learning from examples: *generalization* (and possibly even *consistency*).

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).

Learning from examples: another goal (from inverse problems) is to ensure that 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*

Consider the standard learning algorithm,

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

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

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

More later…..

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*

Classical framework but with more general loss function

The algorithm uses a quite general space of functions or "hypotheses" : RKHSs.

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

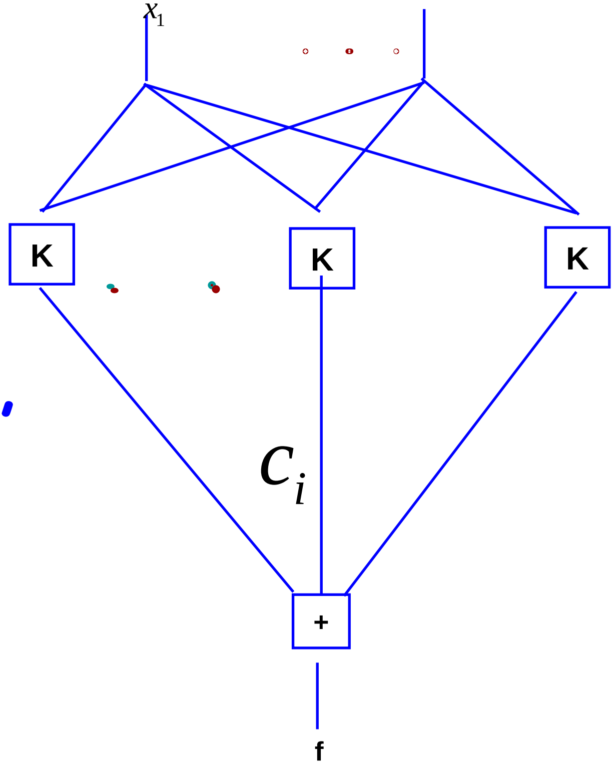
Girosi, Caprile, Poggio, 1990

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"



Theory summary

In the course we will introduce

- Generalization (predictivity of the solution)
- Stability (well-posedness)
- RKHSs hypotheses spaces
- Regularization techniques leading to RN and SVMs
- Manifold Regularization (semisupervised learning)
- Unsupervised learning
- Generalization bounds based on stability
- Alternative classical bounds (VC and Vgamma dimensions)

- Related topics

- Applications

S
y

Syllabus

Overview of overview

- Supervised learning: real math
- Examples of recent and ongoing in-house engineering on applications
- Learning and the brain

Learning from Examples: engineering applications

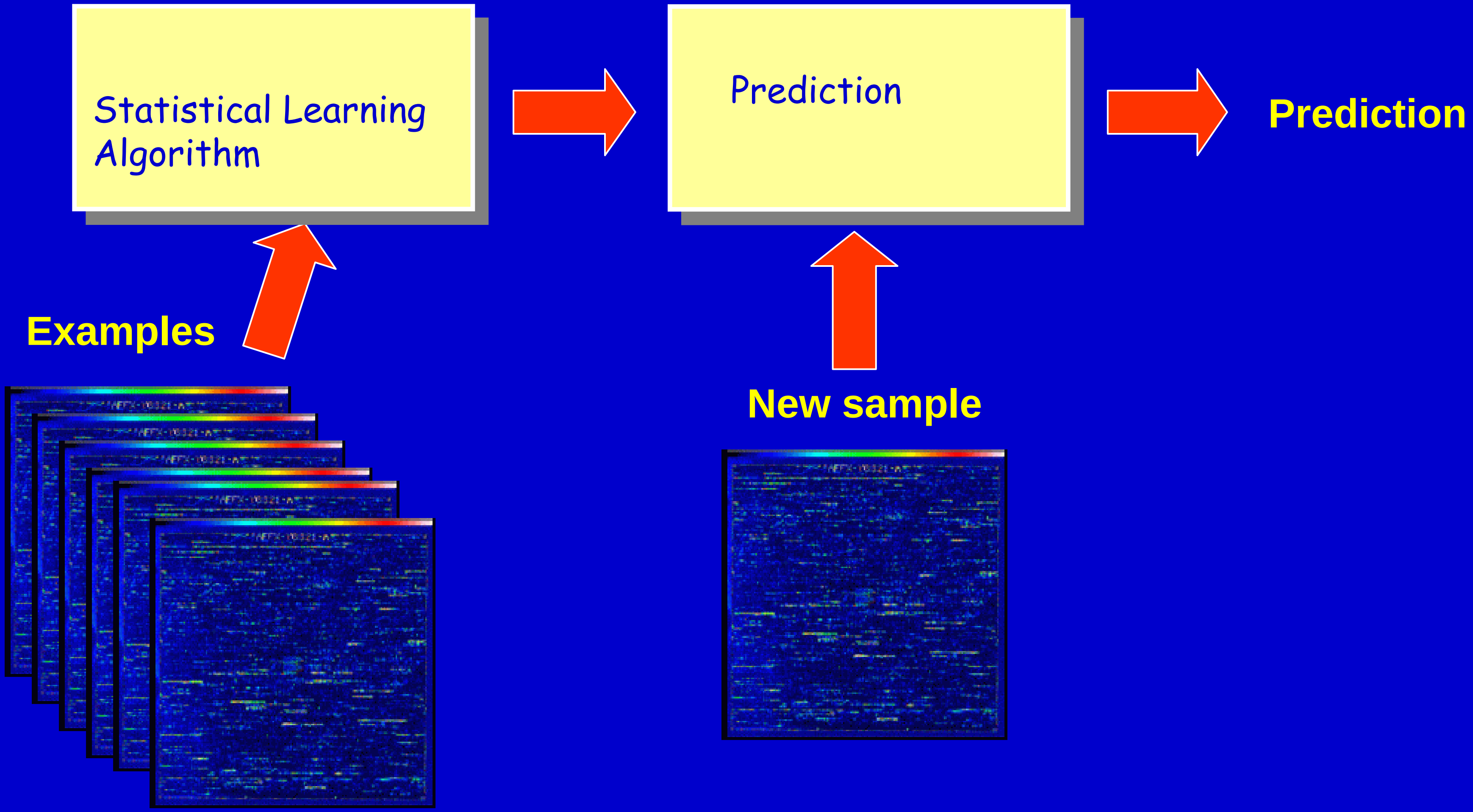
INPUT
OUTPUT

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



Statistical Learning Algorithm

Prediction

Prediction

Examples

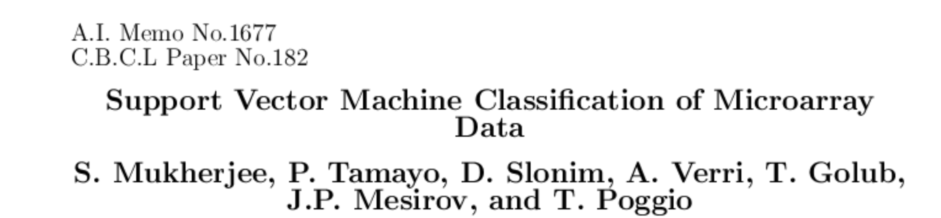
New sample

Bioinformatics application: predicting type of cancer from DNA chips

New feature selection SVM:

Only 38 training examples, 7100 features

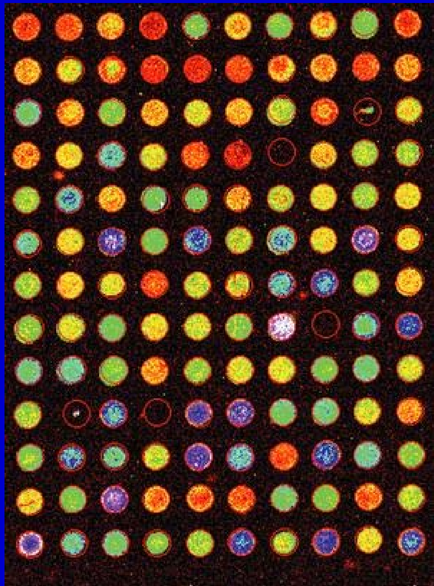
AML vs ALL: 40 genes 34/34 correct, 0 rejects.
5 genes 31/31 correct, 3 rejects of which 1 is an error.



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.

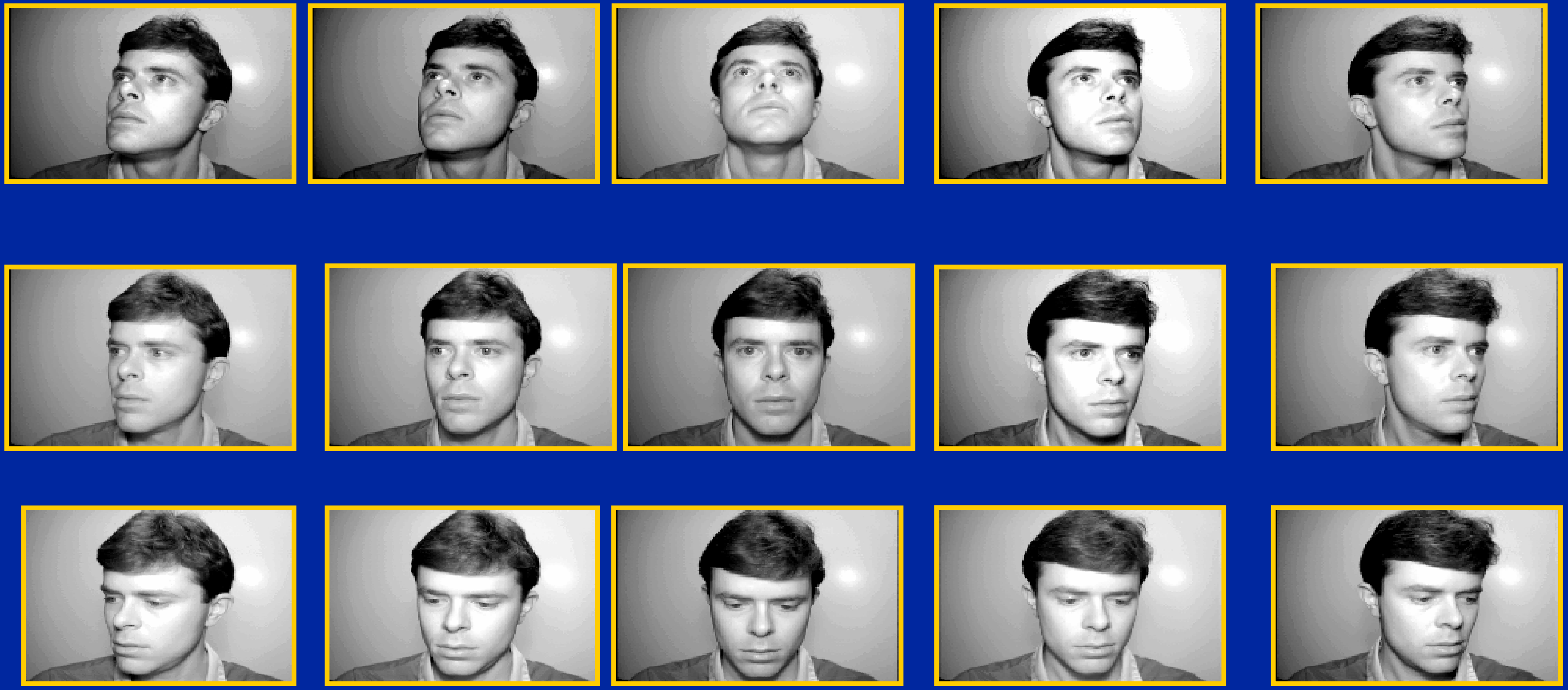
Learning from Examples: engineering applications

INPUT
OUTPUT

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

Face identification: example

An old view-based system: 15 views



Performance: 98% on 68 person database

Beymer, 1995

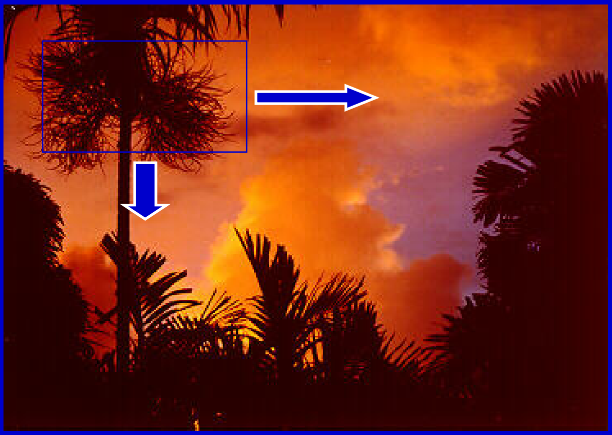
Learning from Examples: engineering applications

INPUT
OUTPUT

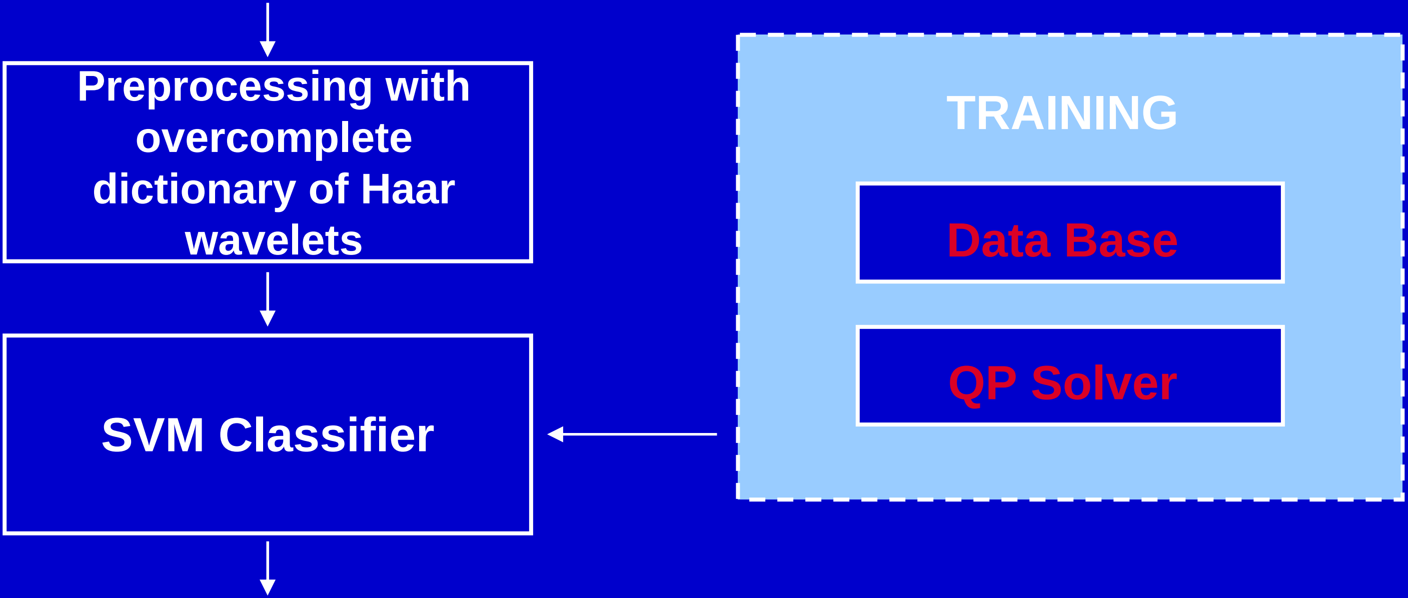
Bioinformatics
Artificial Markets
Object categorization
Object identification
Image analysis
Graphics
Text Classification

.....

System Architecture



Scanning in x,y and scale



Preprocessing with overcomplete dictionary of Haar wavelets

SVM Classifier

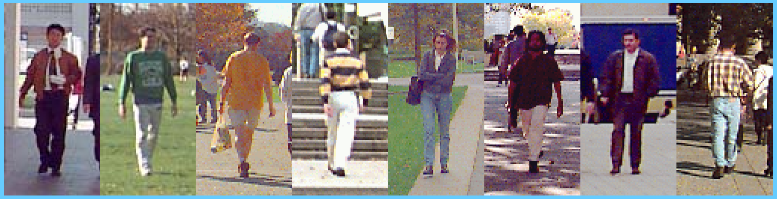
TRAINING

Data Base

QP Solver

Sung, Poggio 1994; Papageorgiou and Poggio, 1998

People classification/detection: training the system



. . .

1848 patterns

. . .

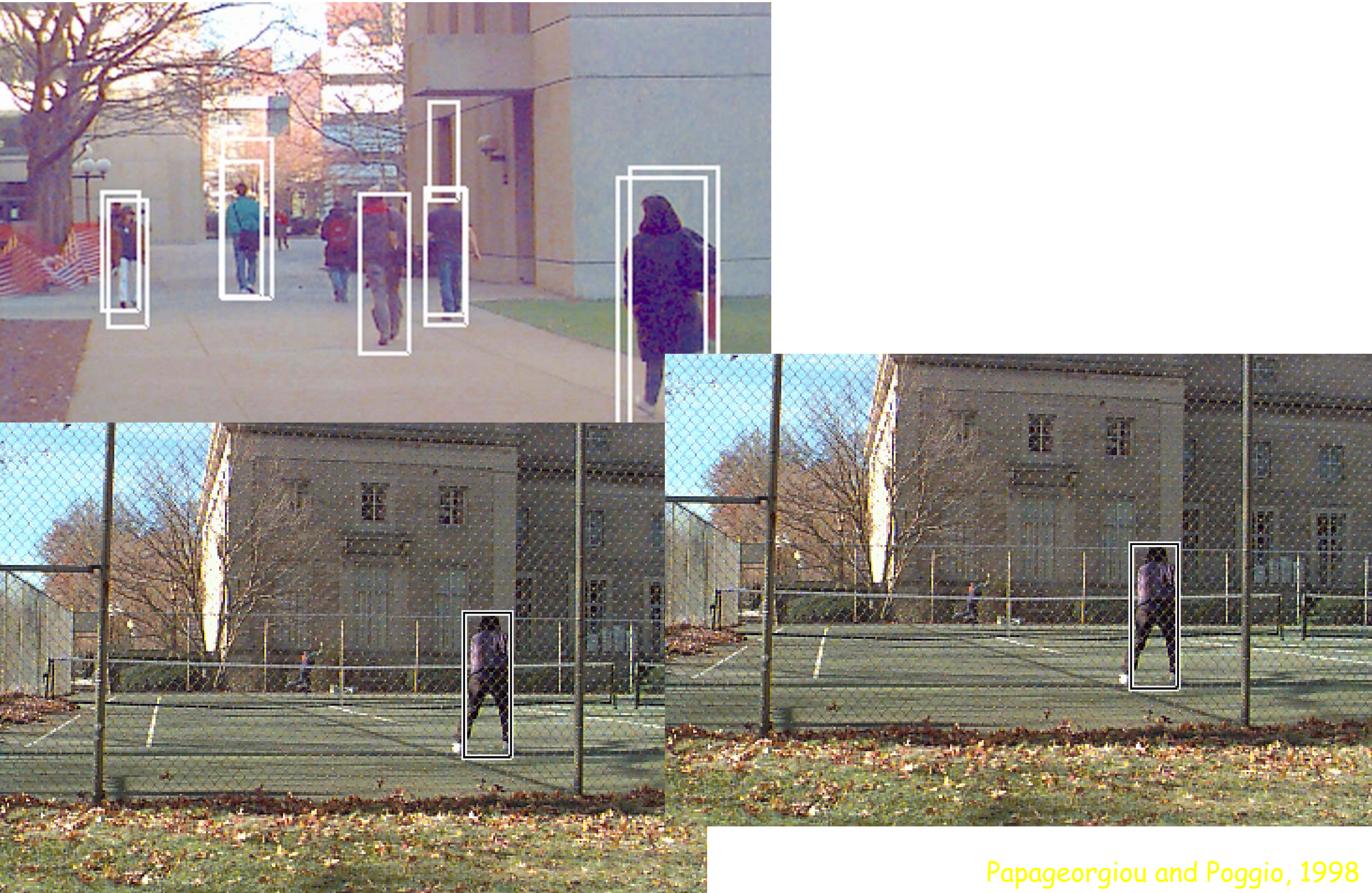
7189 patterns

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

Core learning algorithm: Support Vector Machine classifier

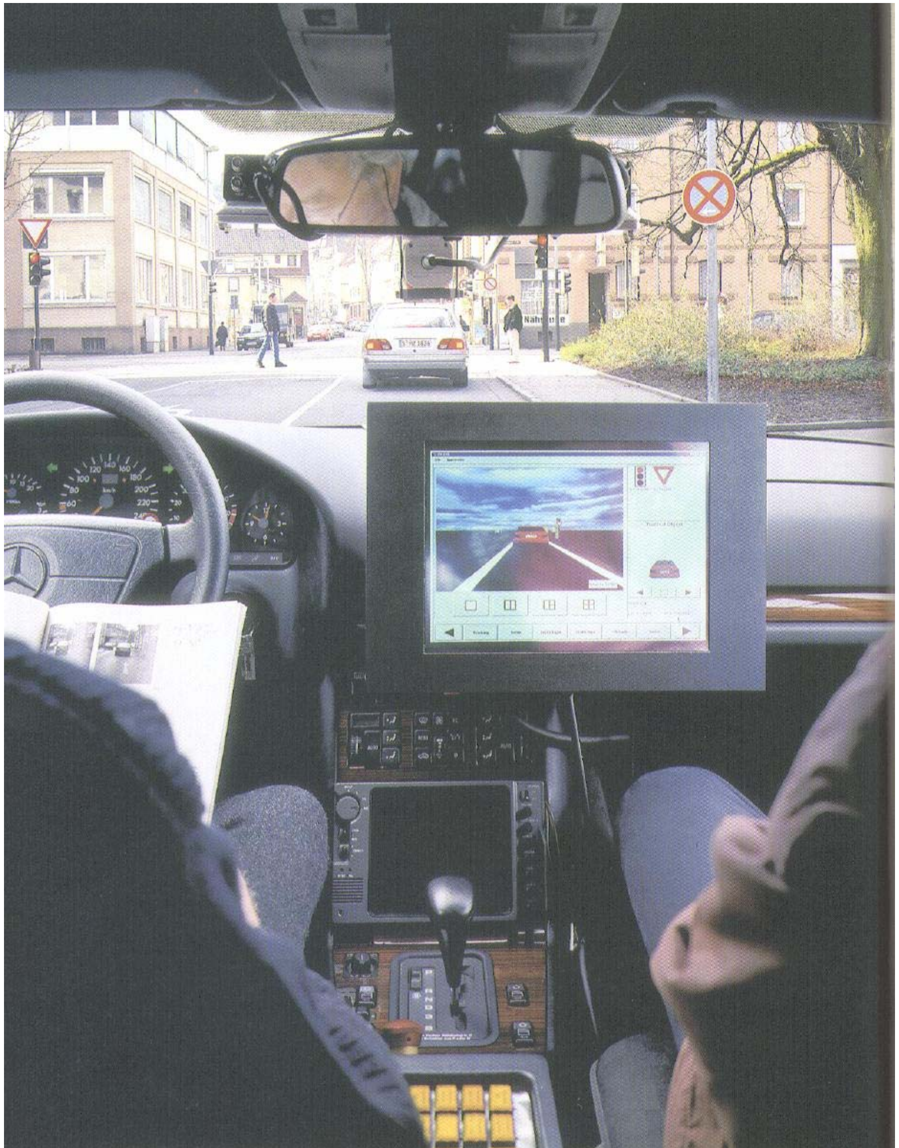
pedestrian detection system

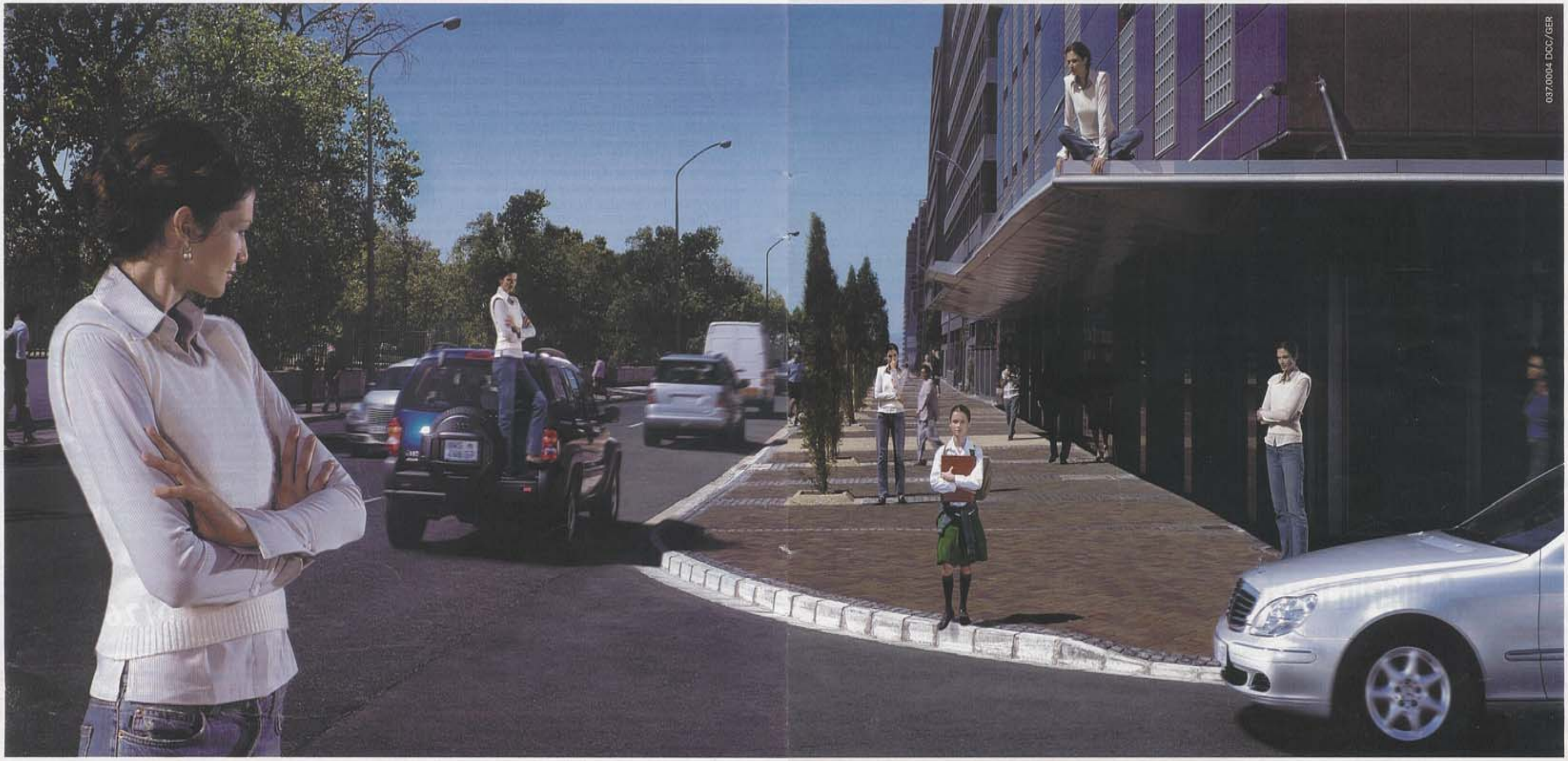
Trainable System for Object Detection: Pedestrian detection - Results



Papageorgiou and Poggio, 1998

The system was tested in a test car (Mercedes)





037.0004 DCC/GER

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

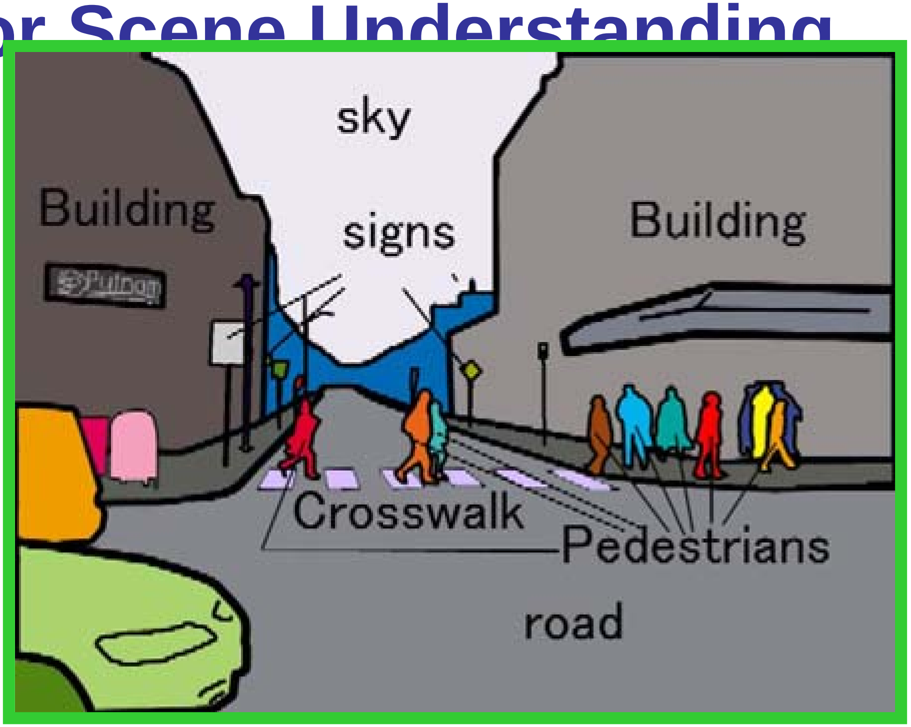
Face identification: training the system

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

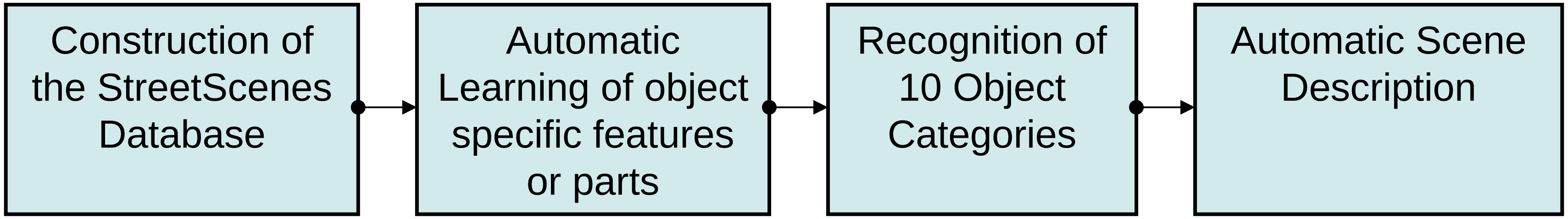
face identification

Computer vision: new StreetScenes Project

Learning Algorithms for Scene Understanding



Project Timeline



Construction of the StreetScenes Database

Automatic Learning of object specific features or parts

Recognition of 10 Object Categories

Automatic Scene Description

Lior Wolf, Stan Bileschi, …

Learning from Examples: Applications

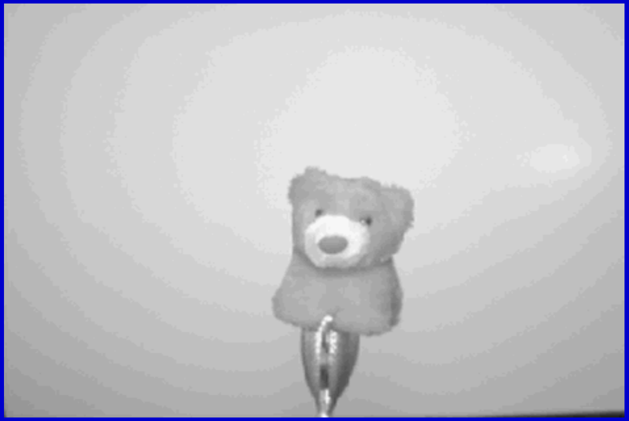
INPUT
OUTPUT

Object identification
Object categorization
Image analysis
Graphics
Finance
Bioinformatics

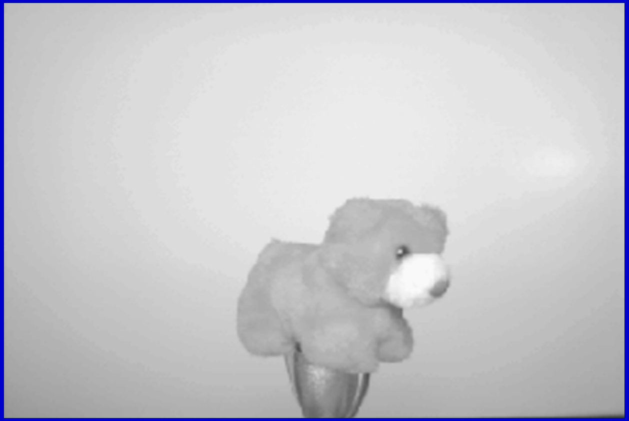
…

Image Analysis

IMAGE ANALYSIS: OBJECT RECOGNITION AND POSE ESTIMATION

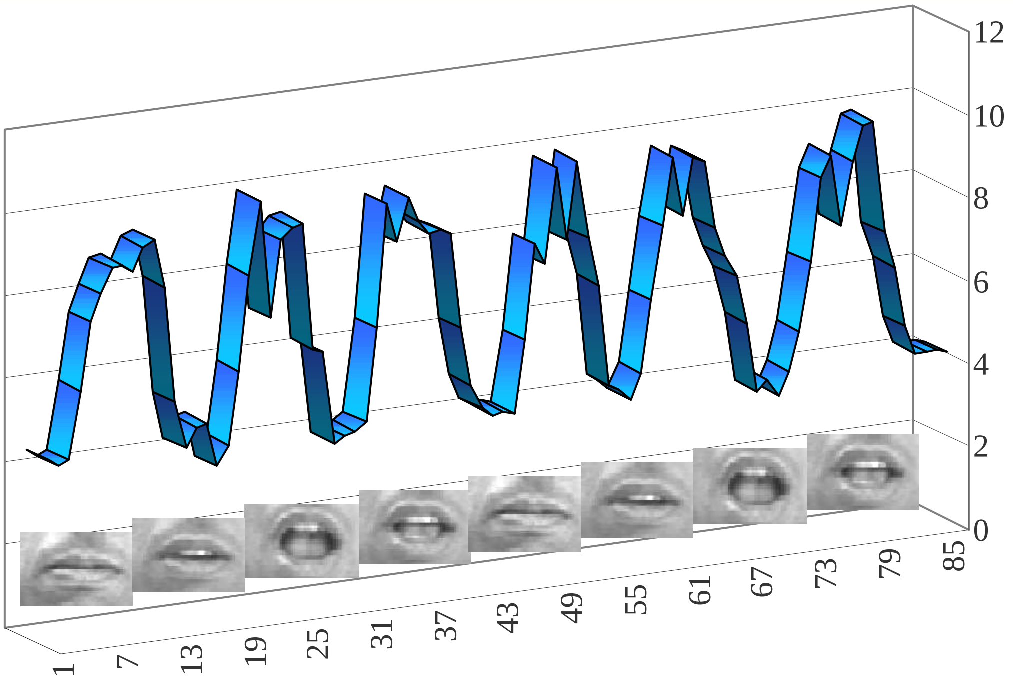


⇒ Bear (0° view)



⇒ Bear (45° view)

Computer vision: analysis of facial expressions



The main goal is to estimate basic facial parameters, e.g. degree of mouth openness, through learning. One of the main applications is video-speech fusion to improve speech recognition systems.

Learning from Examples: engineering applications

INPUT
OUTPUT

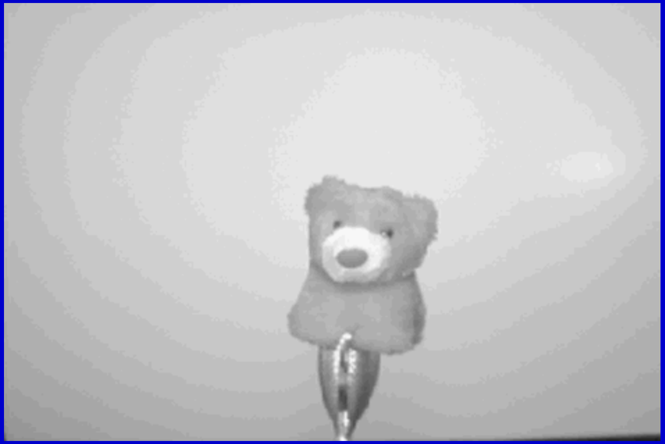
Bioinformatics
Artificial Markets
Object categorization
Object identification
Image analysis
Image synthesis, eg Graphics
Text Classification

…..

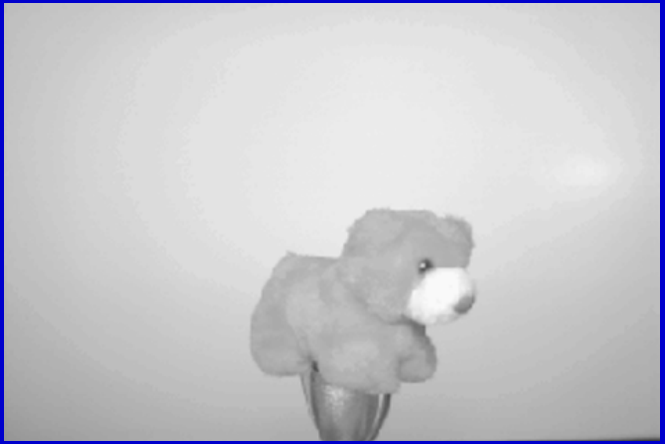
Image Synthesis

Metaphor for UNCONVENTIONAL GRAPHICS

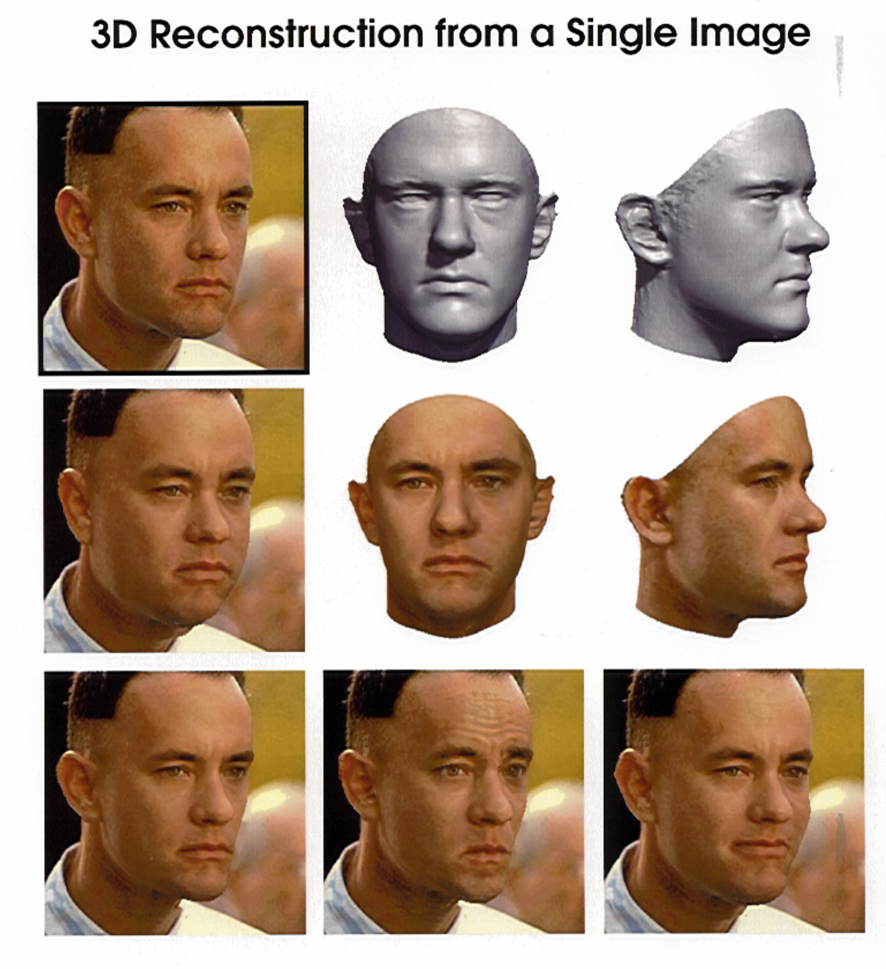
Θ = 0° view ⇒



Θ = 45° view ⇒

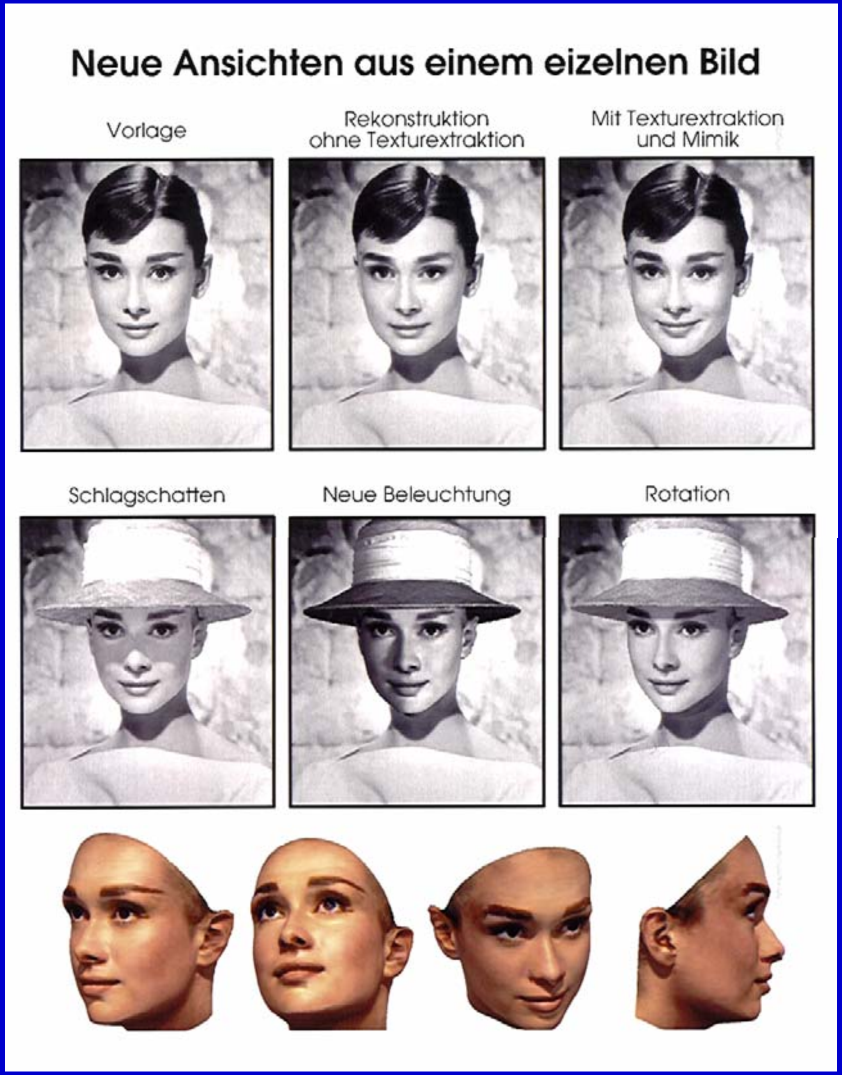


Reconstructed 3D Face Models from 1 image



Blanz and Vetter,
MPI
SigGraph '99

Reconstructed 3D Face Models from 1 image



Neue Ansichten aus einem eizelnen Bild

Vorlage

Rekonstruktion ohne Texturextraktion

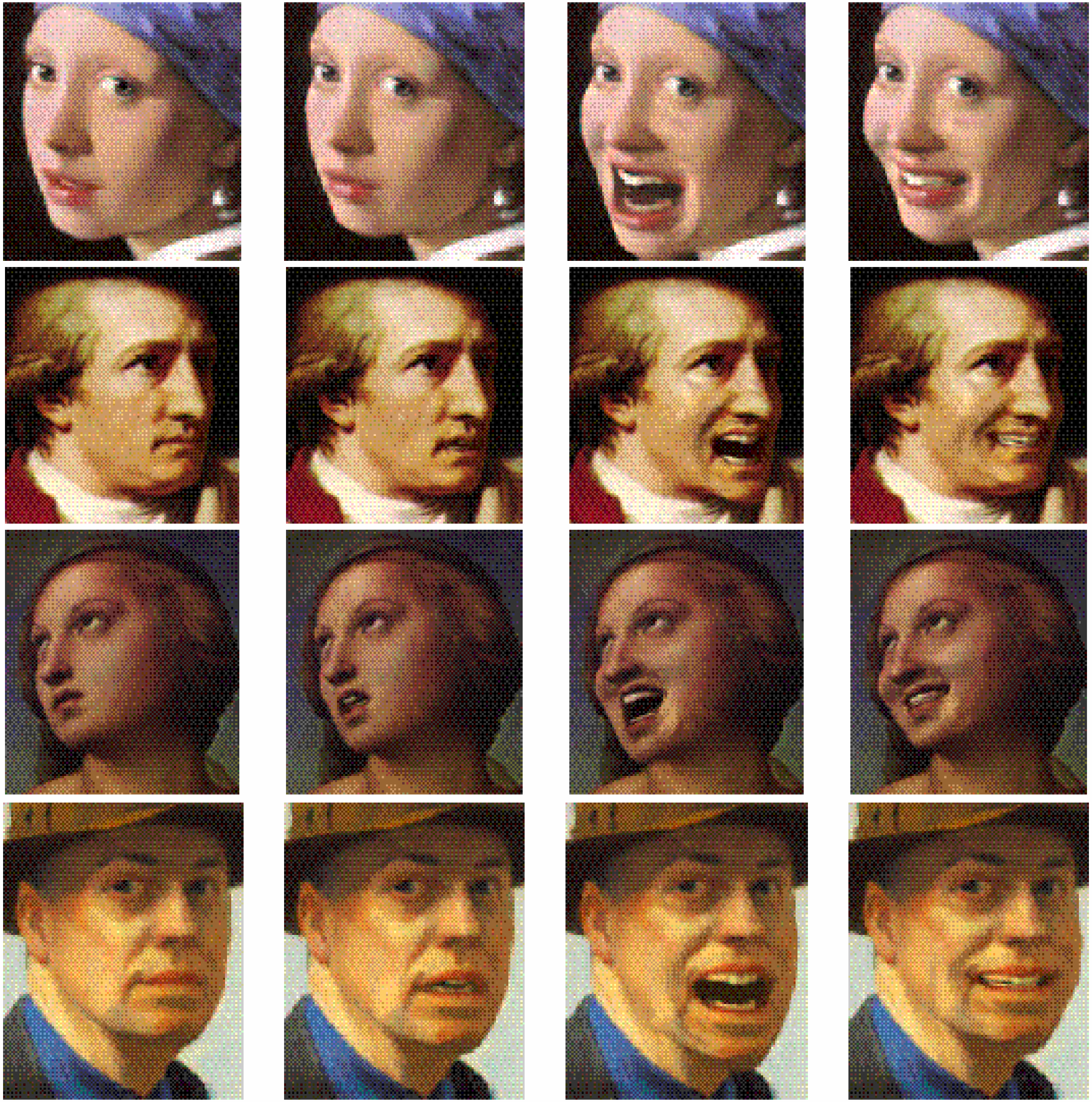
Mit Texturextraktion und Mimik

Schlagschatten

Neue Beleuchtung

Rotation

Blanz and Vetter,
MPI
SigGraph '99



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

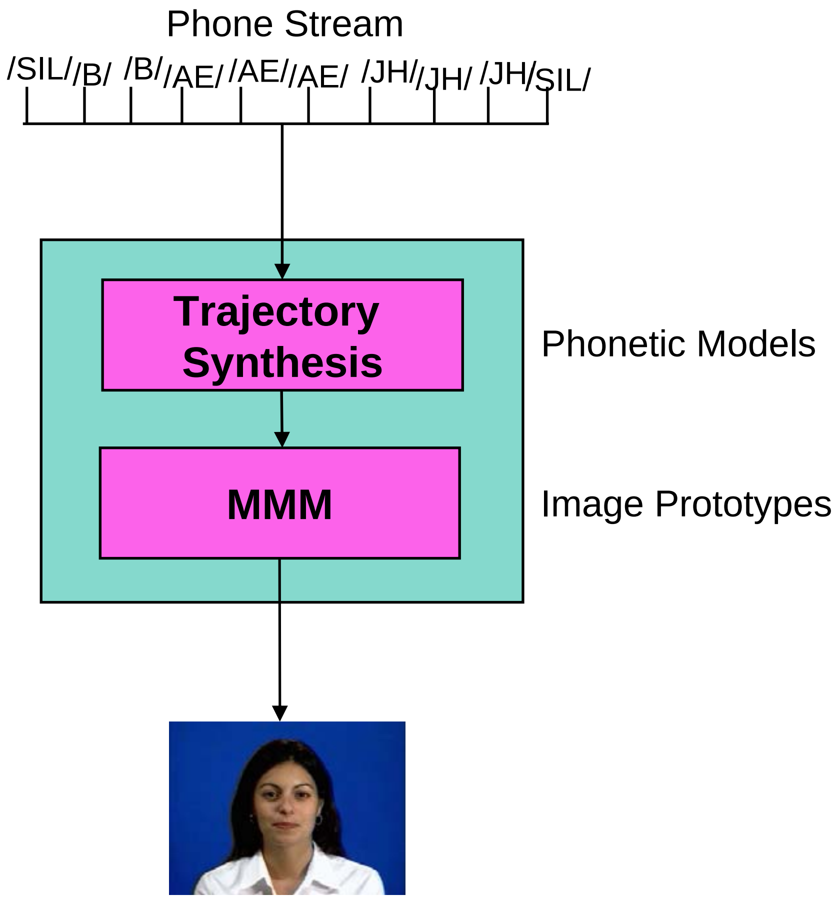
1. Learning

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

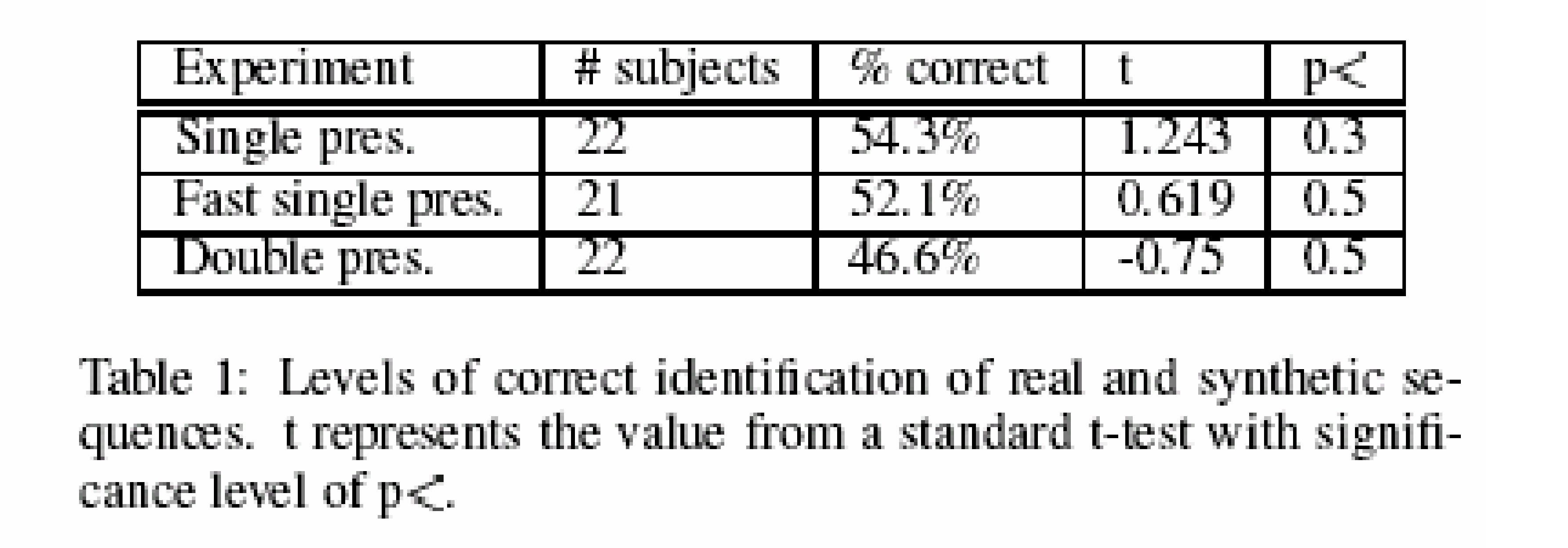
For any speech input the system provides as output a synthetic video stream



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.



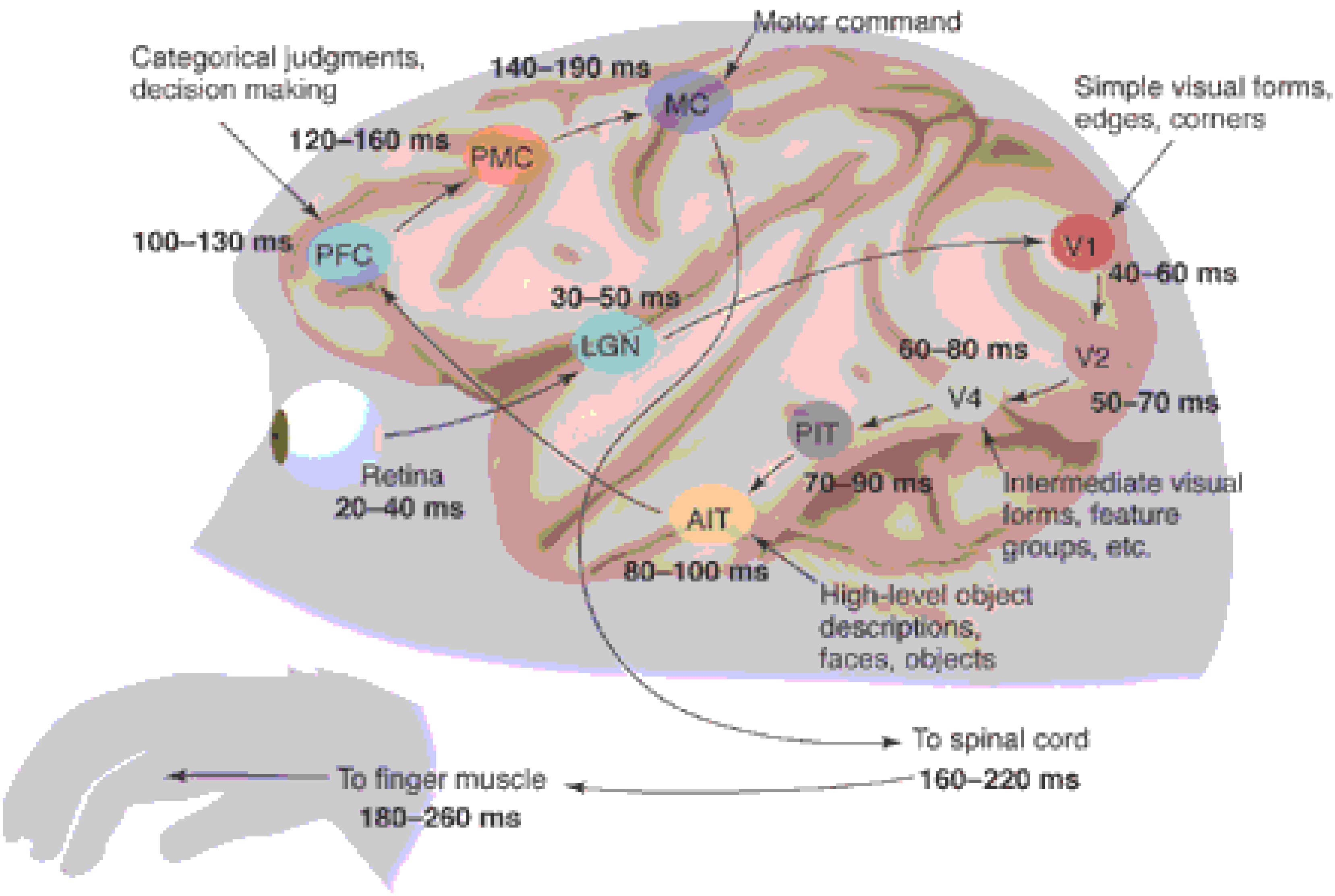
Experiment	# subjects	% correct	t	p<
Single pres.	22	54.3%	1.243	0.3
Fast single pres.	21	52.1%	0.619	0.5
Double pres.	22	46.6%	-0.75	0.5

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

- Supervised learning: the problem and how to frame it within classical math
- Examples of in-house applications
- Learning and the brain

Learning to recognize objects and the ventral stream in visual cortex



Some numbers

Human Brain

10^{11}… 10^{12} neurons

10^{14} + synapses

Neuron

Fine dendrites : 0.1 μ diameter

Lipid bylayer membrane : 5 nm thick

Specific proteins : pumps, channels, receptors, enzymes

Synaptic packet of transmitter opens 2 x 10^3 channels (with 10^4 AcH molecules)

Each channel: conductance g = 10^{-11} mho

Fundamental time length : 1 msec

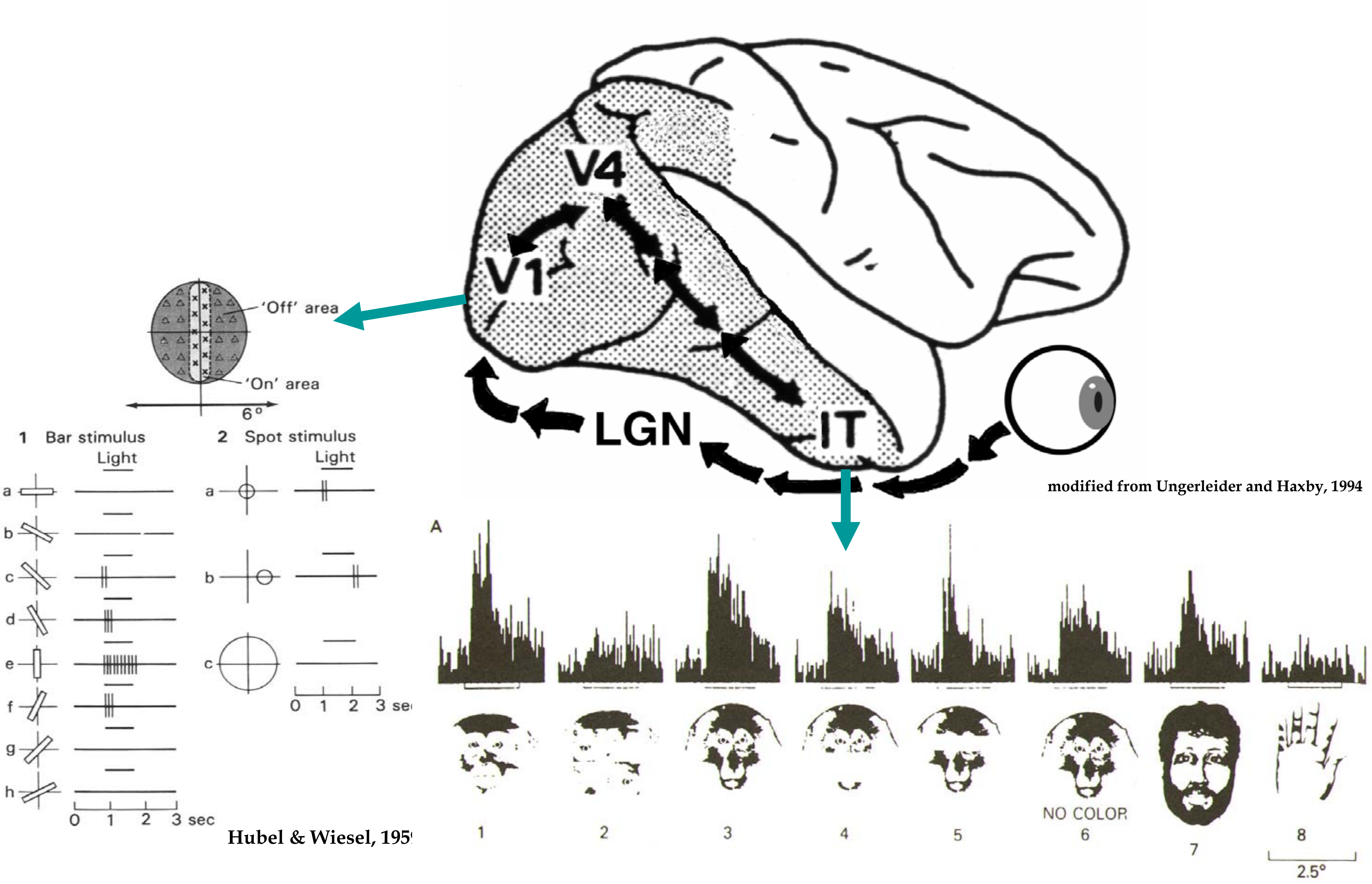
A theory
# of the ventral stream of visual cortex

Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich
and Tomaso Poggio

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

department of brain and
cognitive sciences at MIT

The Ventral Visual Stream: From V1 to IT



modified from Ungerleider and Haxby, 1994

Hubel & Wiesel, 195

Summary of “basic facts”

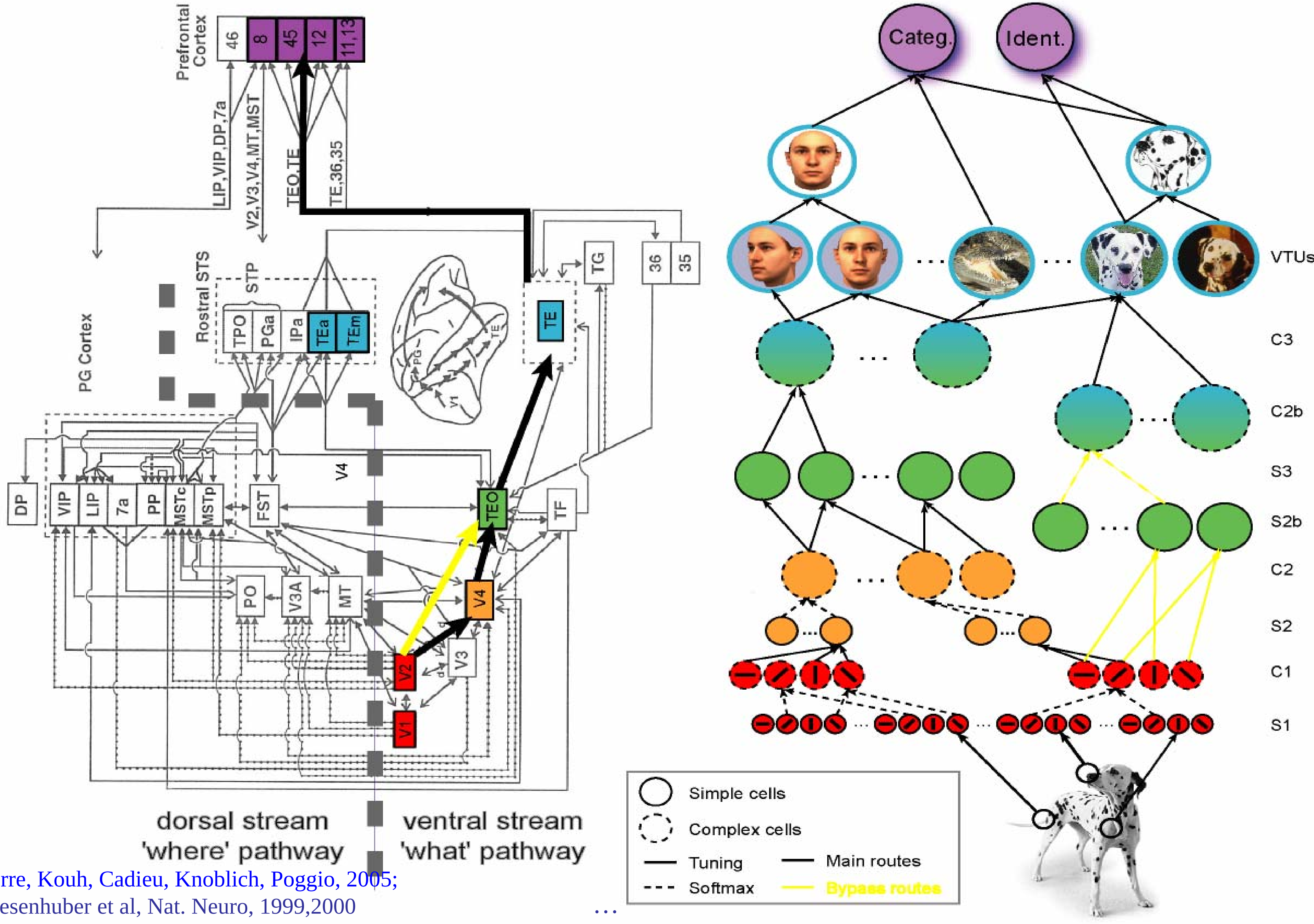
Accumulated evidence points to three (mostly accepted) properties of the ventral visual stream architecture:

- Hierarchical build-up of invariances (first to translation and scale, then to viewpoint etc.) , size of the receptive fields and complexity of preferred stimuli

- Basic feed-forward processing of information (for “immediate” recognition tasks)

- Learning of an individual object generalizes to scale and position

Mapping the ventral stream into a model



Serre, Kouh, Cadieu, Knoblich, Poggio, 2005;
Riesenhuber et al, Nat. Neuro, 1999,2000

…

The model

Claims to interpret or predict several existing data in microcircuits and system physiology, and also in cognitive science:

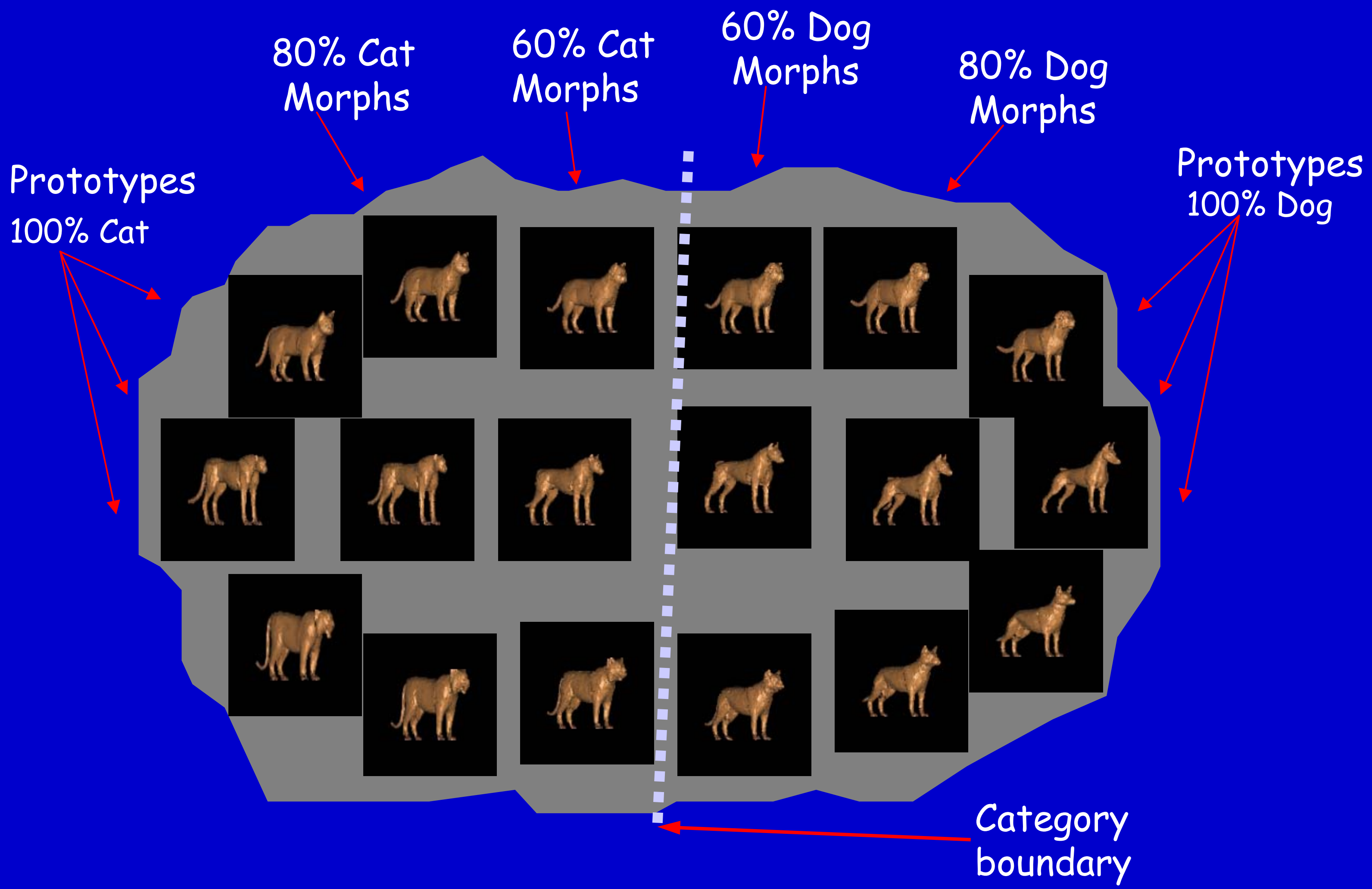
- What some complex cells in V1 and V4 do and why: MAX…

- View-tuning of IT cells (Logothetis)
- Response to pseudomirror views
- Effect of scrambling
- Multiple objects
- Robustness/sensitivity to clutter
- K. Tanaka's simplification procedure
- Categorization tasks (cats vs dogs)
- Invariance to translation, scale etc…
- Read-out data…

- Gender classification
- Face inversion effect : experience, viewpoint, other-race, configural vs. featural representation
- Binding problem, no need for oscillations…

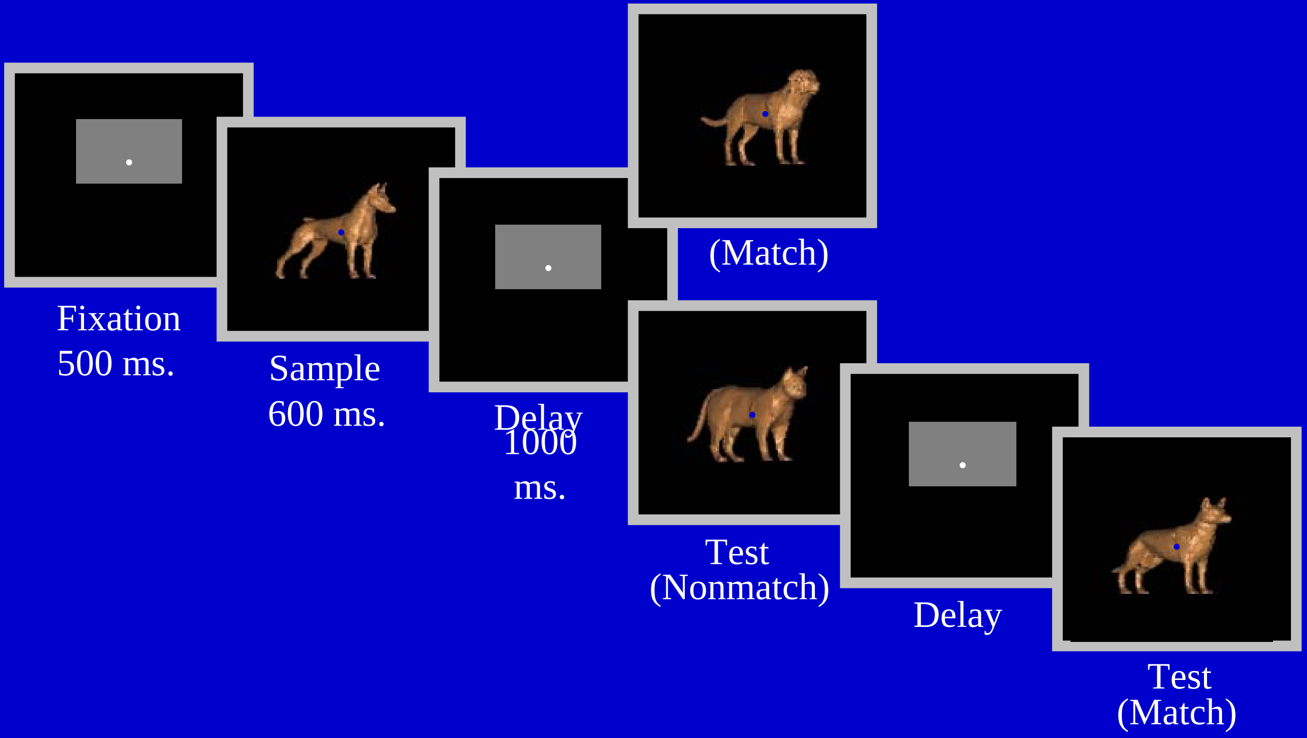
Neural Correlate of Categorization (NCC)

Define categories in morph space



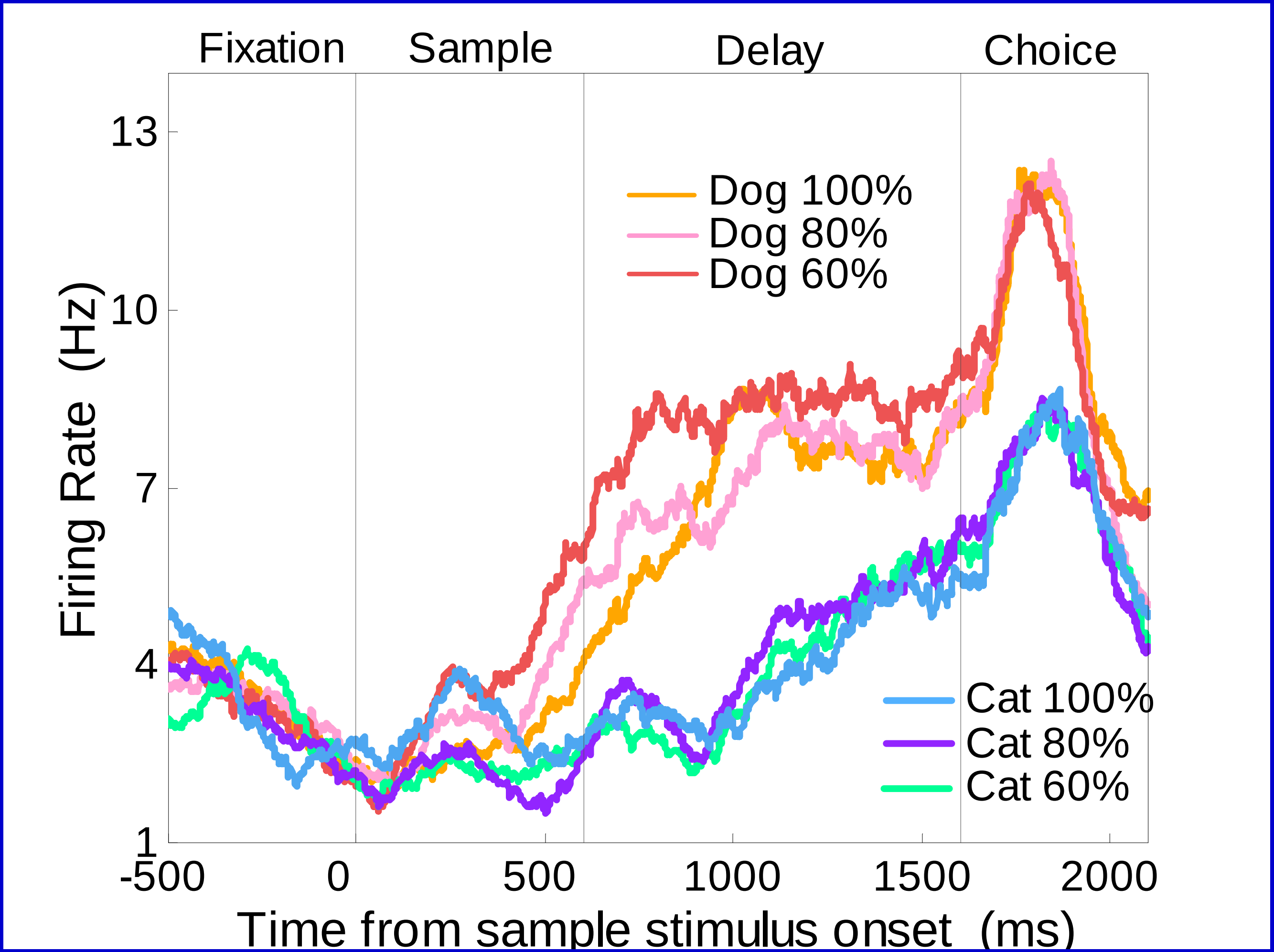
Categorization task

Train monkey on categorization task

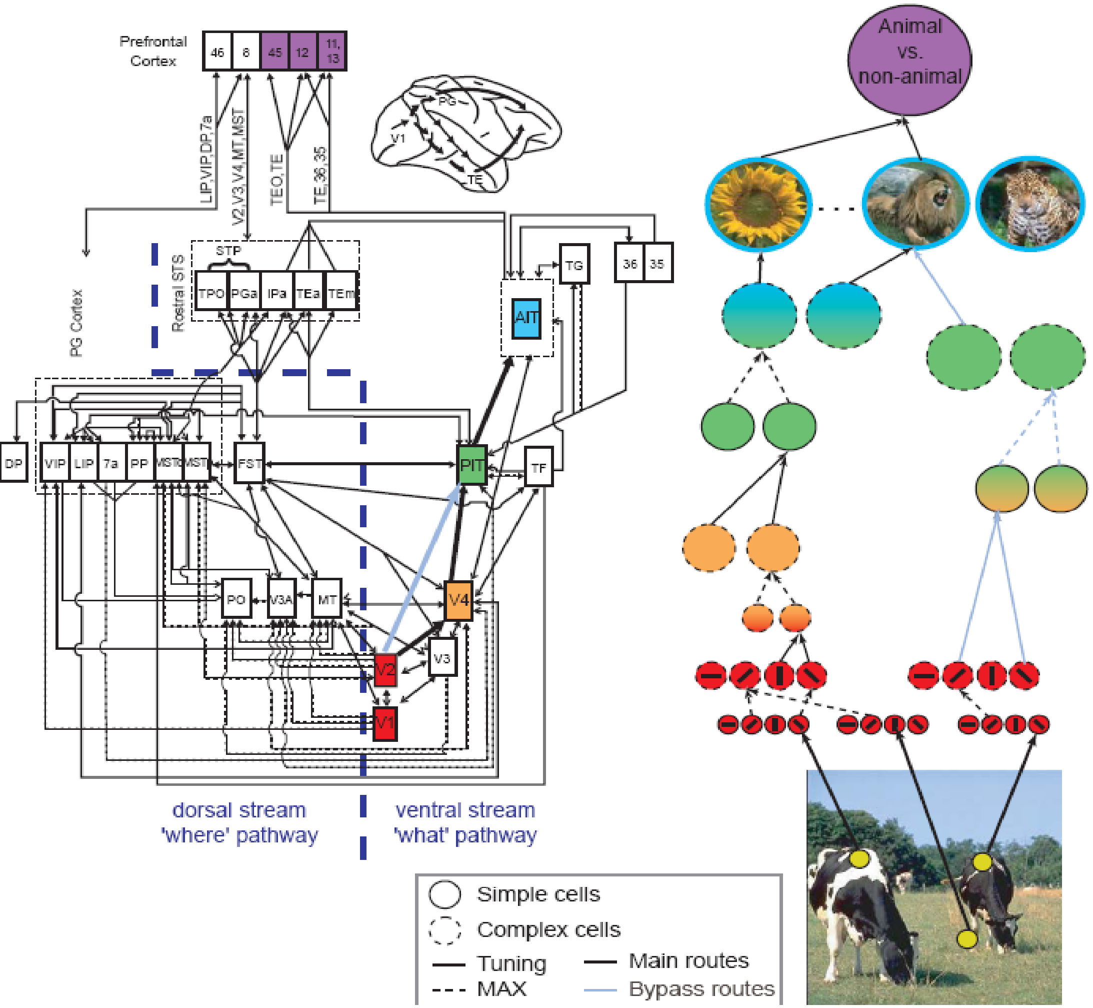


After training, record from neurons in IT & PFC

Single cell example: a "categorical" PFC neuron that responds more strongly to DOGS than CATS



D. Freedman + E. Miller + M. Riesenhuber+T. Poggio (Science, 2001)



Model layers	Corresponding brain area (tentative)	RF sizes	Number units	
classifier	PFC		1.0 10^0	
S4	AIT	>4.4°	1.5 10^2	~ 5,000 subunits
C3	PIT - AIT	>4.4°	2.5 10^3	
C2b	PIT	>4.4°	2.5 10^3	
S3	PIT	1.2°- 3.2°	7.4 10^4	~ 100 subunits
S2b	V4 - PIT	0.9°- 4.4°	1.0 10^7	~ 100 subunits
C2	V4	1.1°- 3.0°	2.8 10^5	
S2	V2 - V4	0.6°- 2.4°	1.0 10^7	~ 10 subunits
C1	V1 - V2	0.4°- 1.6°	1.2 10^4	
S1	V1 - V2	0.2°- 1.1°	1.6 10^6	

Supervised task-dependent learning

Unsupervised task-independent learning

increase in complexity (number of subunits), RF size and invariance

The model fits many physiological data,
predicts several new ones…

recently it provided a surprise (for us)…

…when we compared its performance with machine vision…

Sample Results on the CalTech 101-object dataset

crocodile head : 96.90

panda : 94.20

emu : 90.40

metronome : 96.9

gramophone : 92.80

lobster : 90.80

saxophone : 95.50

snoopy : 94.20

brontosaurus : 95.70

camera : 91.20

headphone : 96.70

crocodile : 95.30

mandolin : 91.40

pigeon : 92.00

hedgehog : 91.50

scissors : 97.90

pagoda : 97.10

scissors : 97.90

rooster : 94.60

octopus : 94.80

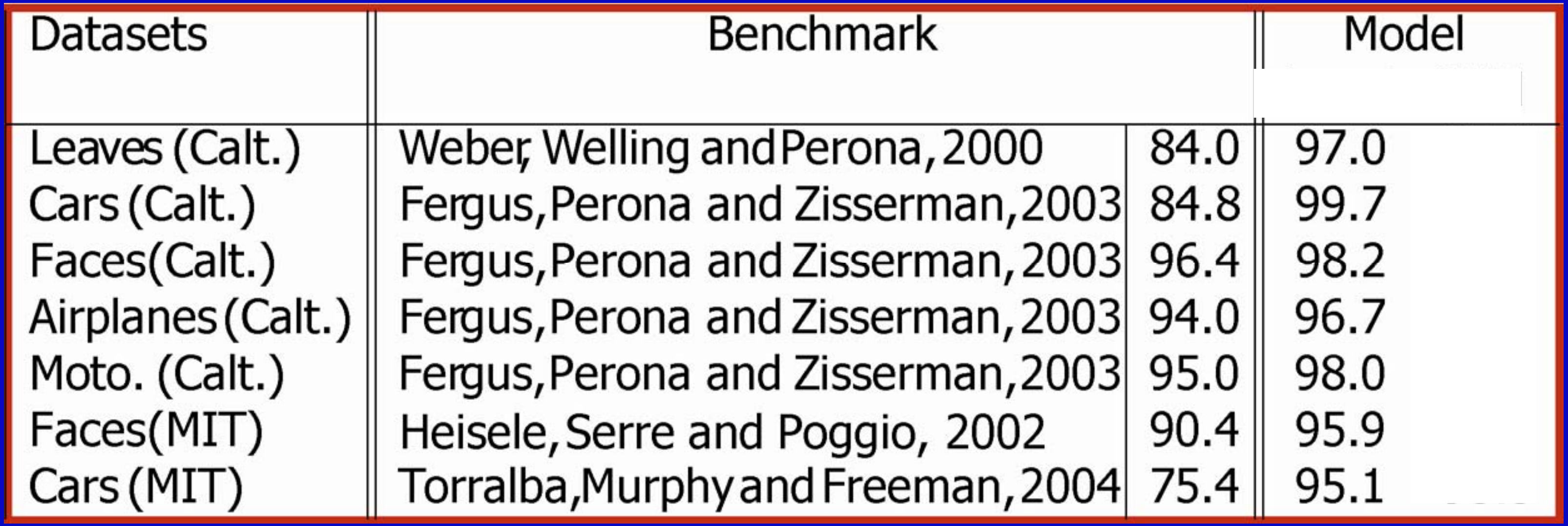
headphone : 96.70

ant : 94.60

platypus : 91.60

gramophone : 92.80

The model performs at the level of the best computer vision systems

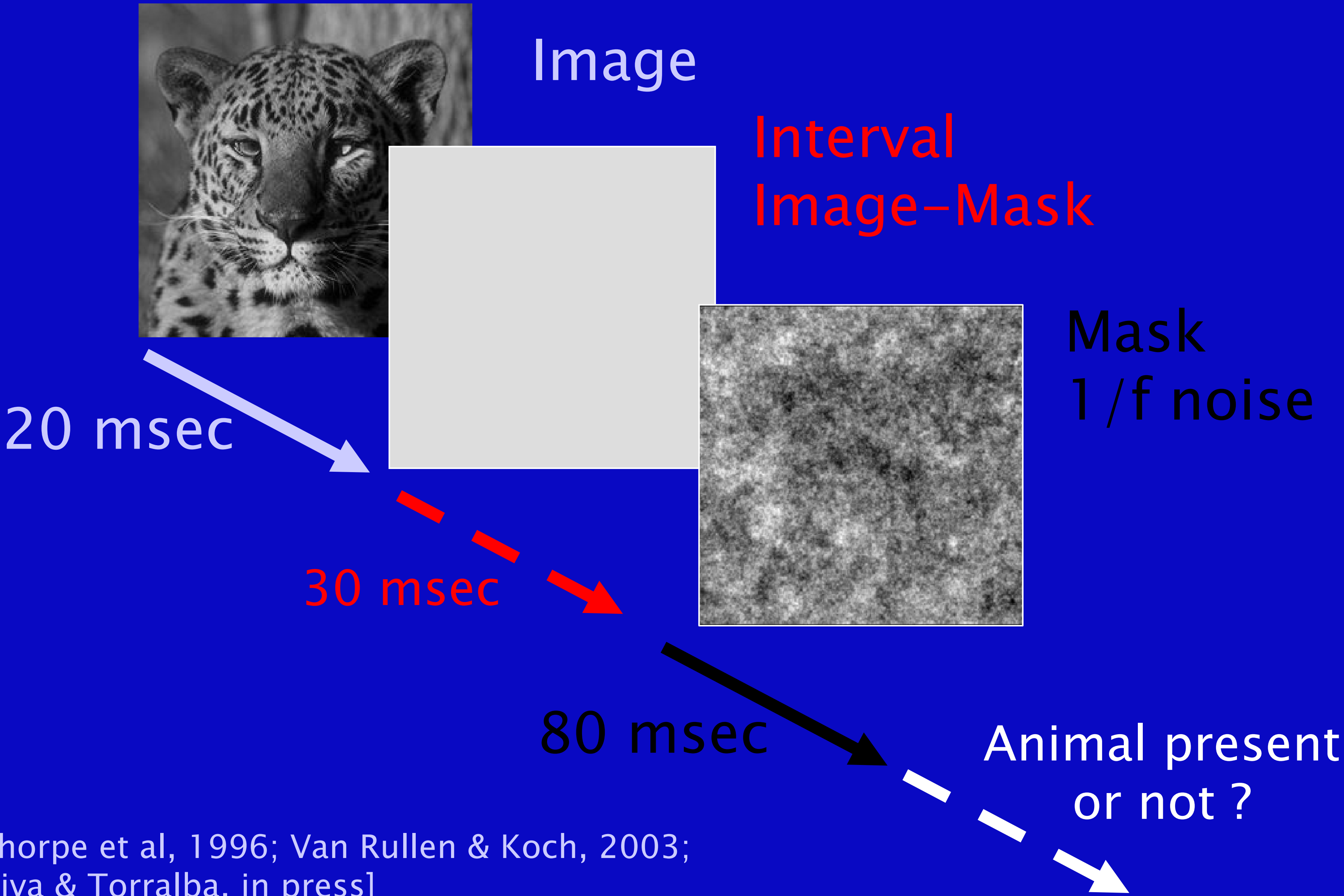


Datasets	Benchmark		Model
Leaves (Calt.)	Weber, Welling and Perona, 2000	84.0	97.0
Cars (Calt.)	Fergus, Perona and Zisserman, 2003	84.8	99.7
Faces(Calt.)	Fergus, Perona and Zisserman, 2003	96.4	98.2
Airplanes (Calt.)	Fergus, Perona and Zisserman, 2003	94.0	96.7
Moto. (Calt.)	Fergus, Perona and Zisserman, 2003	95.0	98.0
Faces(MIT)	Heisele, Serre and Poggio, 2002	90.4	95.9
Cars (MIT)	Torralba, Murphy and Freeman, 2004	75.4	95.1

…and another surprise…

… was the comparison with human performance
(Thomas Serre with Aude Oliva)
on rapid categorization of complex natural images

Experiment: rapid (to avoid backprojections) animal detection in natural images



Image

Interval
Image-Mask

Mask
1/f noise

20 msec

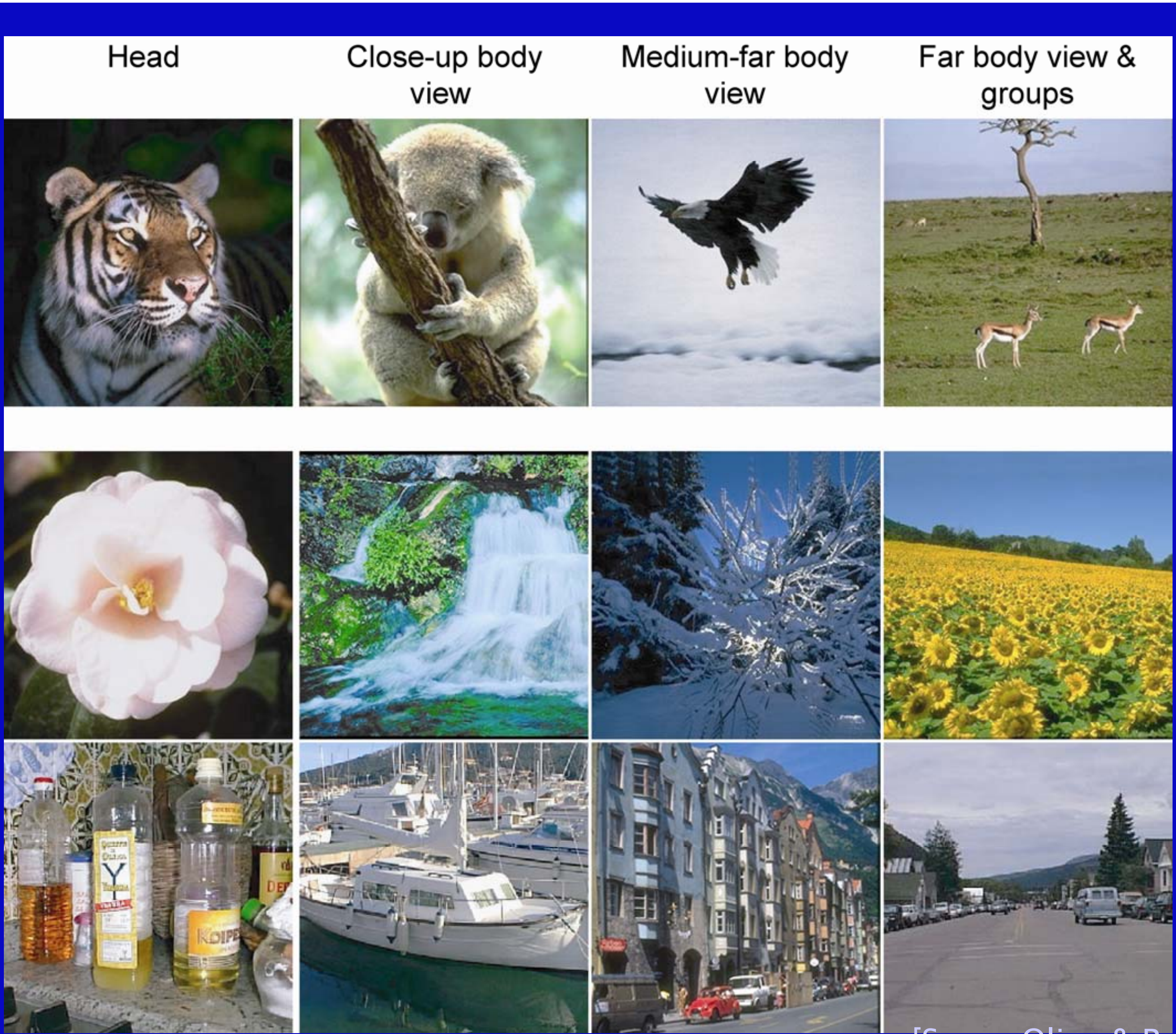
30 msec

80 msec

Animal present
or not ?

[Thorpe et al, 1996; Van Rullen & Koch, 2003;
Oliva & Torralba, in press]

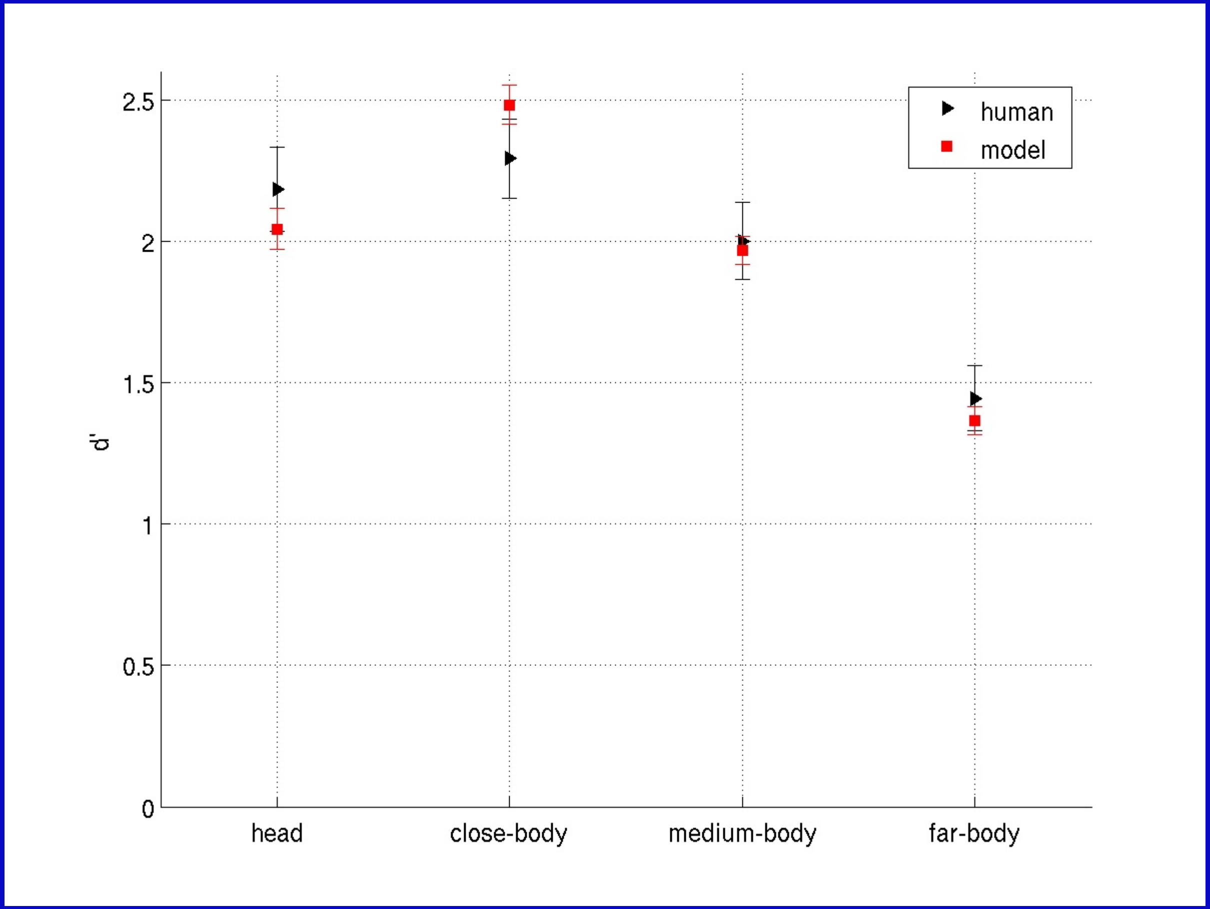
Targets and distractors



[Serre, Oliva & Poggio, in prep]

Humans achieve model–level performance

Model results obtained without tuning a single parameter!



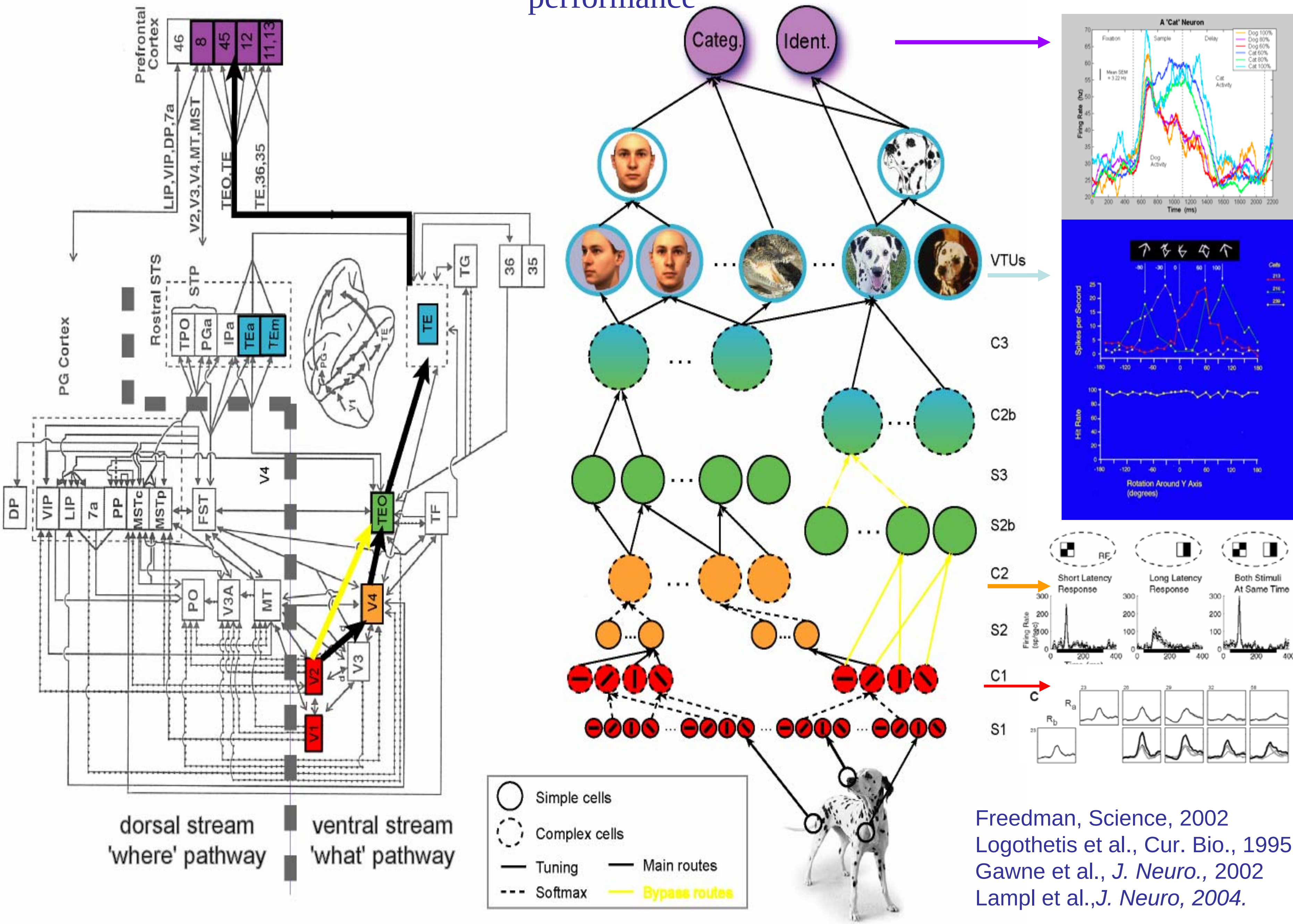
Human: 80% correct

vs.

Model: 82% correct

[Serre, Oliva & Poggio, in prep]

Theory supported by data in V1, V4, IT; works as well as the best computer vision; mimics human performance



Freedman, Science, 2002
Logothetis et al., Cur. Bio., 1995
Gawne et al., *J. Neuro.*, 2002
Lampl et al.,*J. Neuro, 2004.*

A challenge for learning theory:

an unusual, hierarchical architecture
with unsupervised and supervised learning
and learning of invariances…

We will see later why this is unusual and interesting for learning theory!