9.520 in 2012

Statistical Learning Theory and Applications

Class Times:
Monday and Wednesday 10:30-12:00
Units: 3-0-9 H,G
Location:
46-5193
Instructors:
T. Poggio, L. Rosasco, C. Ciliberto, G. Evangelopoulos and C. Frogner

Office Hours:
Friday 1-2 pm in 46-5156, CBCL lounge (by appointment)

Email Contact :
9.520@mit.edu

Class

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

Mathcamps (optional):

- Functional analysis (~45mins)
- Probability (~45mins)

Sept 9th
7pm-9pm???

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 12th
- scribing
- participation
- Grading is based on Psets (25%+25%) + Final Project (30%) + Scribing (10%) + Participation (10%)

Slides on the Web site

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

Problem Set 1: Wed 16 Oct (Class 12)
Problem Set 2: Tues 12 Nov (Class 20)
Final Project : Wed 14 Dec (Class 27 - Currently TBA)

Final Project (this year is different)

The final project can be either a Wikipedia entry or a research project (we highly recommend a Wikipedia entry).

We envision 2 kinds of research project: 1) applications-- evaluate an algorithm on some interesting problem of your choice; and 2) theory and algorithms -- study theoretically or empirically some new machine learning algorithm/problem.

For the Wikipedia article we suggest to post 1-2 pages (short) using Wikipedia standard format (of course); for the research project you should use the template on the Web site.

Project: posting/editing article on Wikipedia

- Computational learning theory: to be redone or new entry in Generalization Bounds
- RKHS is ok but could be improved on the learning side
- Stability in Learning Theory (batch and online) is missing
- Radial basis function network should be rewritten or edited
- VC theory exists in a minimalistic form should be improved
- Regularization networks/theory IS TERRIBLE...EASY TO IMPROVE
- Statistical learning theory is a mess should be edited
- http://en.wikipedia.org/wiki/Deep_learning should be edited and corrected

Research Projects

Examples: Projects 2012

- Project Ideas Contact: Instructors

- Evaluating which Classifiers Work Best for Decoding Neural Data Contact: Ethan Meyers
- Does learning from segmented images aid categorization? Contact: Cheston Tan

-- Learning to rank papers/grants: replacing review panels

-- Oscillations and iterations in optimization

-- Class-specific computations and architecture of recognition

-- Sparseness and recall from visual associative memory

-- The surprising usefulness of sloppy arithmetic: study of bits and their tradeoff in hierarchical architectures

Research Projects

Projects 2013

- Project Ideas Contact: Instructors

-- Learning to rank papers/grants: replacing review panels

-- Simulations of associative memories for object recognition: bounds on # items stored, noiseless coding, sparseness

-- The surprising usefulness of sloppy arithmetic: study of bits and their tradeoff in hierarchical architectures

Overview of overview

- Context for this course: a golden age for new AI and the key role of Machine Learning
- Success stories from past research in Machine Learning: examples of engineering applications
- Statistical Learning Theory
- A new kind of basic research on learning: computer science and neuroscience, learning and the brain: A Center for Brains, Minds and Machines
- A new phase in Machine Learning?

Overview of overview

- Context for this course: a golden age for new AI and the key role of Machine Learning
- Success stories from past research in Machine Learning: examples of engineering applications
- Statistical Learning Theory
- A new cycle of basic research on learning: computer science and neuroscience, learning and the brain
- A Center for Brains, Minds and Machines

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.

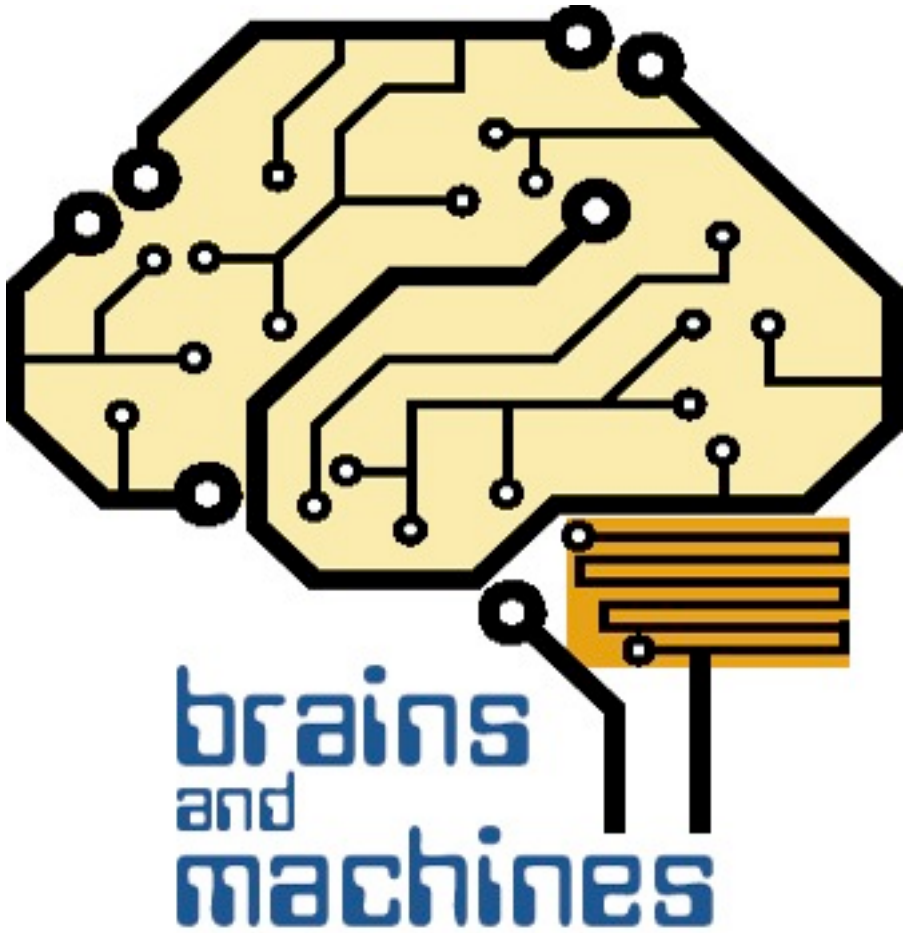
Research on intelligence:

- a great intellectual mission
- will help cure mental diseases and develop more intelligent artifacts
- will improve the mechanisms for collective decisions

These advances will be critical to of our society's

- future prosperity
- education, health, security

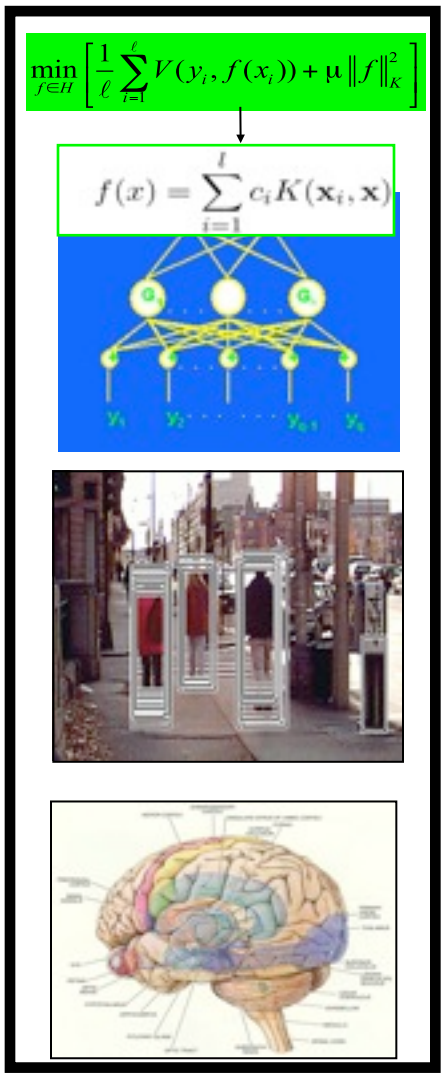
At the core of the problem of Intelligence is the problem of Learning



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

Problem of learning:
a focus for
- math
- computer algorithms
- neuroscience

Machine Learning + Vision @CBCL



LEARNING THEORY + ALGORITHMS

ENGINEERING APPLICATIONS

COMPUTATIONAL NEUROSCIENCE: models+experiments

Mathematics

Engineering

Science

Poggio, T. and F. Girosi. Networks for Approximation and Learning, *Proceedings of the IEEE* 1990) also *Science,* 1990

Poggio, T. and S. Smale. The Mathematics of Learning: Dealing with Data, *Notices American Mathematical Society (AMS)*, 2003

Poggio, T., R. Rifkin, S. Mukherjee and P. Niyogi. General Conditions for Predictivity in Learning Theory, *Nature*, 2004

Beymer, D. and T. Poggio. Image Representation for Visual Learning, *Science*, 272., 1905-1909, 1996

Brunelli, R. and T. Poggio. Face Recognition: Features Versus Templates, *IEEE PAMI*, 1993

Sung, K.K and T. Poggio. Example-Based Learning for View-Based Human Face Detection, IEEE PAMI, 1998 (1995)

Ezzat, T., G. Geiger and T. Poggio. "Trainable Videorealistic Speech Animation," *ACM SIGGRAPH 2002*

Freedman, D.J., M. Riesenhuber, T. Poggio and E.K. Miller. Categorical Representation of Visual Stimuli in Prefrontal Cortex, *Science*, 291, 312-316, 2001.

Riesenhuber, M. and T. Poggio. Hierarchical Models of Object Recognition in Cortex, *Nature Neuroscience*, 2, 1019-1025, 1999.

Serre, T., A. Oliva and T. Poggio. A Feedforward Architecture Accounts for Rapid Categorization, *(PNAS)*, Vol. 104, No. 15, 6424-6429, 2007.

Poggio, T. and E. Bizzi. Generalization in Vision and Motor Control, *Nature*, Vol. 431, 768-774, 2004.

Theory of Learning

- Learning is becoming the *lingua franca* of Computer Science
- Learning is at the center of recent successes in AI over the last 15 years
- The next 10 year will be a golden age for technology based on learning: Google, MobilEye, Siri 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.

Machine Learning is where the action is

Peter Norvig - Jan 26, 2012 - Public

Robert Tibshirani, co-author of one of the best-ever books on statistics / machine learning, describes what it is like to be transformed into **a rockstar, as the field of statistics gains popularity.**

What Are the Odds That Stats Would Be This Popular?

By QUENTIN HARDY | January 26, 2012, 10:30 AM1

"Most of my life I went to parties and heard a little groan when people heard what I did," says Robert Tibshirani, a statistics professor at Stanford University. "Now they're all excited to meet me."

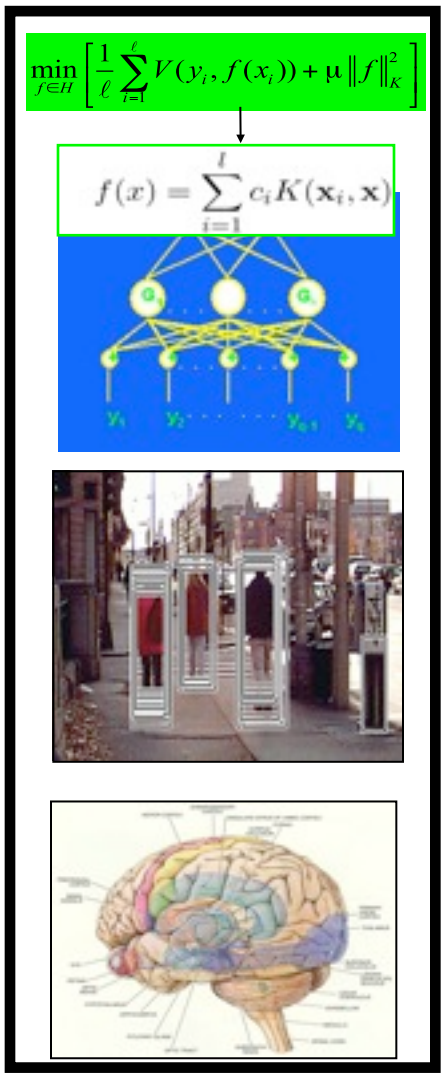
It's not because of a new after-shave. Arcane statistical analysis, the business of making sense of our growing data mountains, has become high tech's hottest calling.

Stanford's Department of Statistics, both renowned and near so many Internet and bioscience companies, is at the center of the boom. It received 800 résumés for next year's 60 graduate positions, twice the number of applications it had three years ago. Graduates head to business school at a starting salary of $150,000 or more, or to Facebook for about $130,000.

Overview of overview

- Context for this course: a golden age for new AI and the key role of Machine Learning
- Success stories from past research in Machine Learning: examples of engineering applications
- Statistical Learning Theory (supervised)
- A new cycle of basic research on learning: computer science and neuroscience, learning and the brain
- A Center for Brains, Minds and Machines

Learning:
Math, Engineering, Neuroscience



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

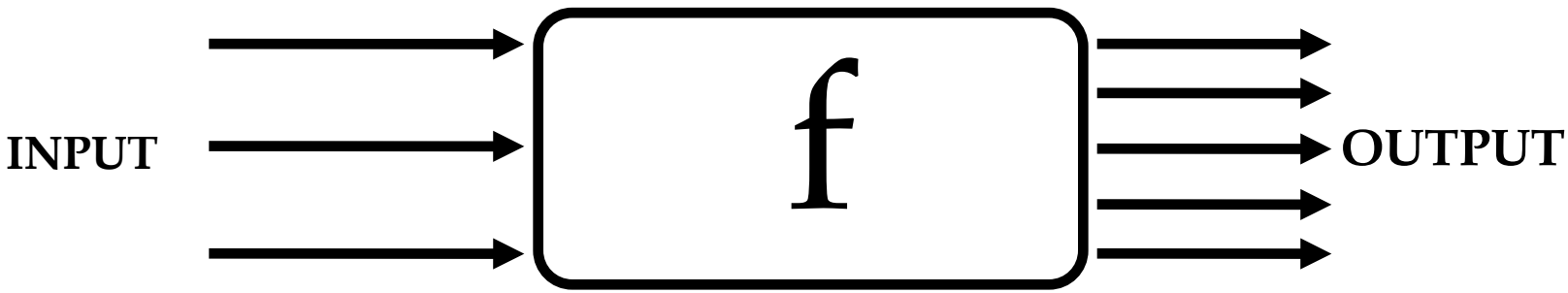
ENGINEERING APPLICATIONS

- 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)

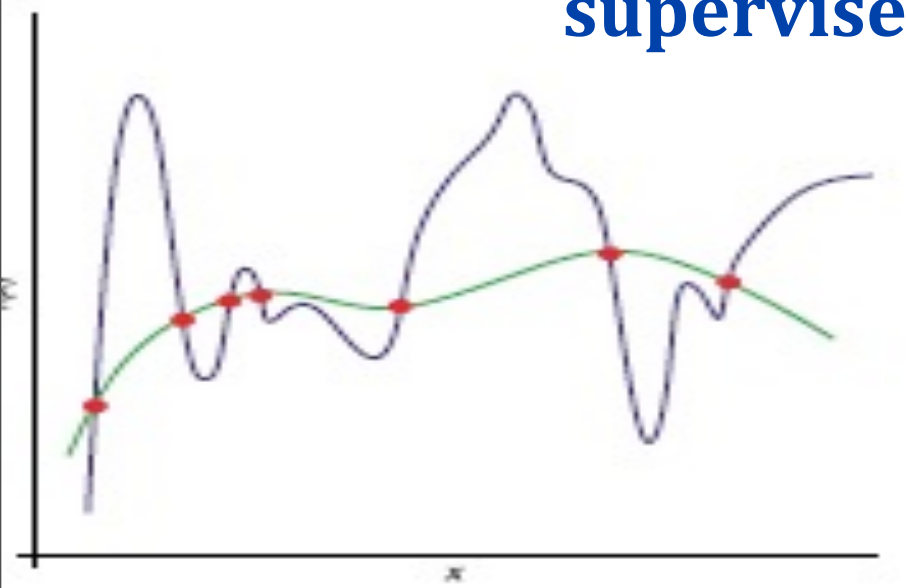
$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\}$$

Question: find function f such that

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

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

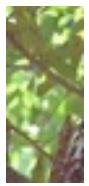
Statistical Learning Theory: supervised learning



Regression

(4,24,…)

(1,13,…)



(7,33,…)

Classification

(92,10,…)

(41,11,…)

(19,3,…)

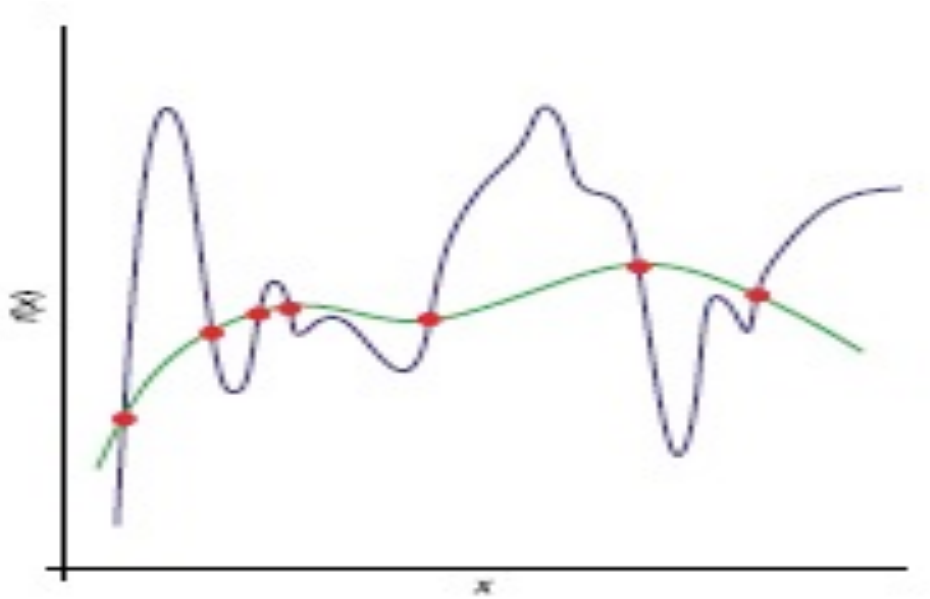
(4,71,…)

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: 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

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

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

which is the **expected loss** on a new example drawn at random from μ.

The **empirical error** of f is:

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

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

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

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

Statistical Learning Theory: supervised learning

Consider a prototypical learning algorithm: ERM (empirical risk minimization)

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

What are the conditions ensuring generalization?

It turns out that choosing an appropriately *simple* hypothesis space *H* (for instance a compact set of continuous functions) can guarantee generalization

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

An algorithm is stable if the removal of any one training sample from any large set of samples results *almost* always in a small change in the learned function.

For ERM the following theorem holds for classification and regression

ERM on H generalizes if and only if the hypothesis space H is uGC and if and only if ERM on H is CV_{loo} *stable*

Mukherjee, Niyogi, Poggio, Rifkin, *Nature*, March 2004

This is an example of foundational results
in learning theory...

Statistical Learning Theory: theorems extending foundations of learning theory

Conditions for **generalization** 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

Statistical Learning Theory: classical 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

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

Equation includes splines, Radial Basis Functions and SVMs (depending on choice of 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

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

Statistical Learning Theory: Regularization

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

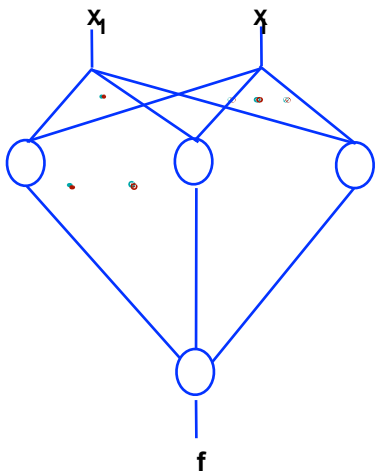
$$\min_{f \in H} \left[\frac{1}{n} \sum_{i=1}^{n} V(f(x_i) - y_i) + \lambda \|f\|_K^2 \right]$$

implies

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

Remark (for later use):

Kernel machines correspond to *shallow* networks



Statistical Learning Theory: note

Two connected and overlapping strands in learning theory both based on *constrained optimization*:

- Bayes, hierarchical models, graphical models…
  - computational intractability?

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

Overview of overview

- Context for this course: a golden age for new AI and the key role of Machine Learning
- Success stories from past research in Machine Learning: examples of engineering applications
- Statistical Learning Theory
- A new cycle of basic research on learning: computer science and neuroscience, learning and the brain
- A Center for Brains, Minds and Machines

Supervised learning

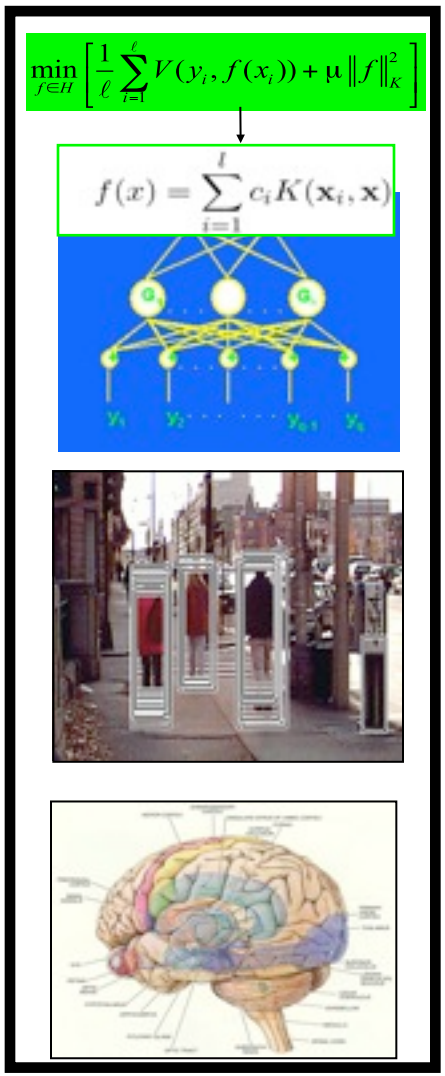
Since the introduction of *supervised learning* techniques 20 years ago, AI has made significant (and not well known) advances in a few domains:

- *Vision*
- *Graphics and morphing*
- *Natural Language/Knowledge retrieval* (Watson and Jeopardy)
- *Speech recognition* (Nuance)
- *Games* (Go, chess,...)


IBM


NAVY
505

Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

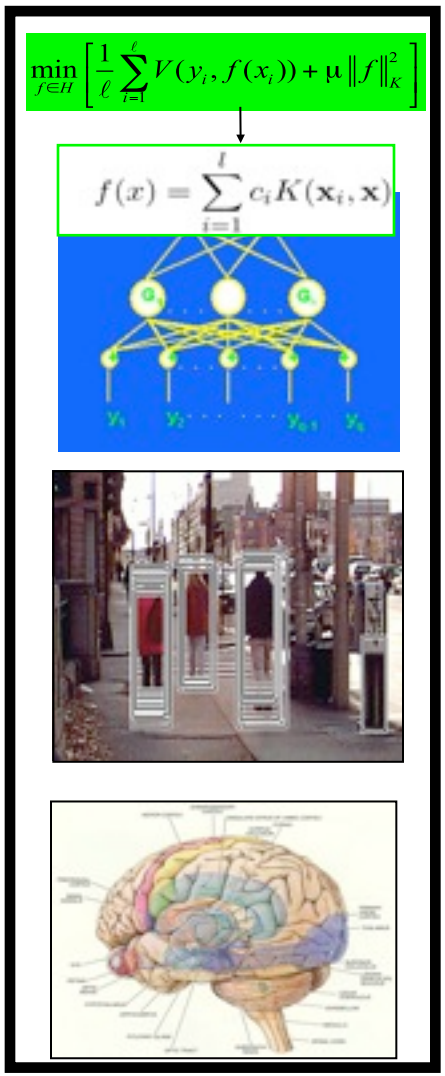
Sung & Poggio 1995

COMPUTATIONAL
NEUROSCIENCE:
models+experiments

How visual cortex works


Image
Output

Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

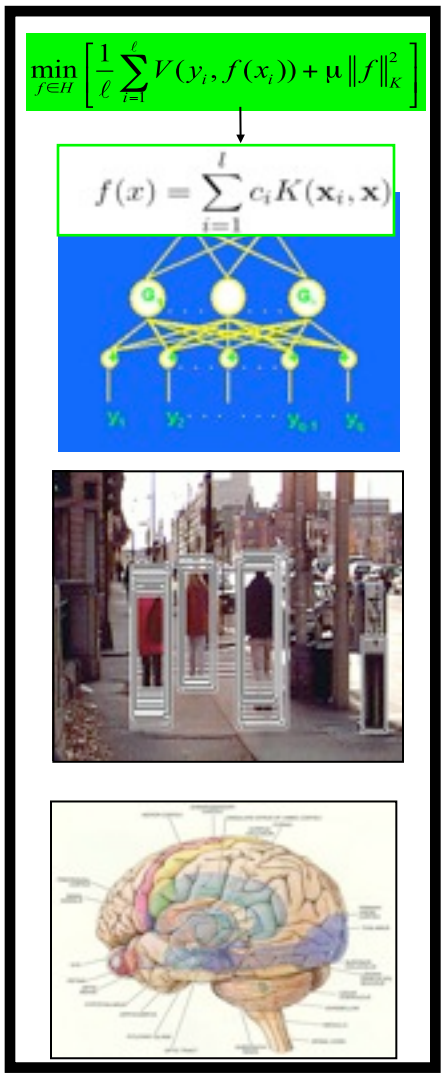
Predictive algorithms

Face detection is now available in digital cameras (commercial systems)

COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

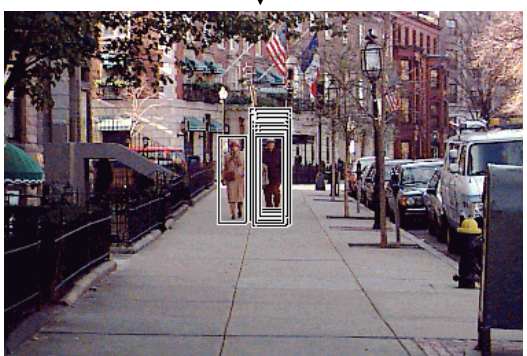
Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

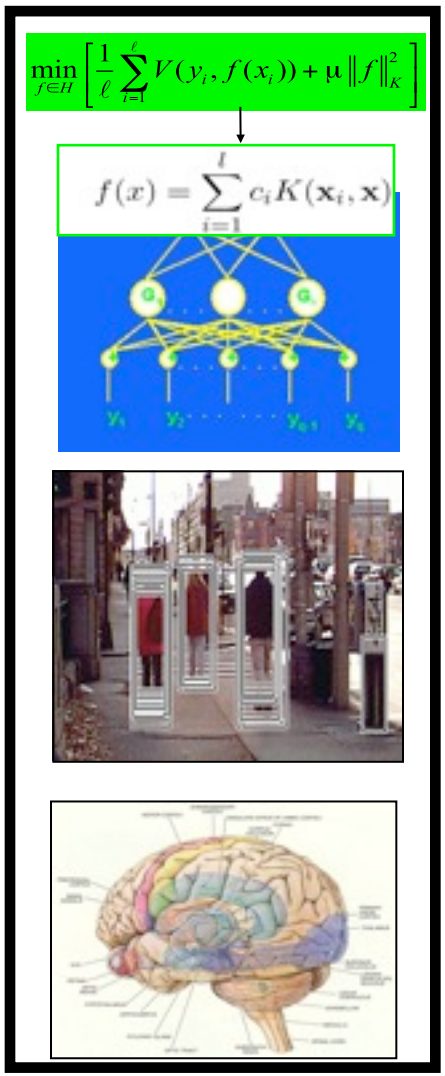


Papageorgiou&Poggio, 1997, 2000
also Kanade&Scheiderman

COMPUTATIONAL
NEUROSCIENCE:
models+experiments

How visual cortex works

Learning



LEARNING THEORY + ALGORITHMS

Theorems on foundations of learning

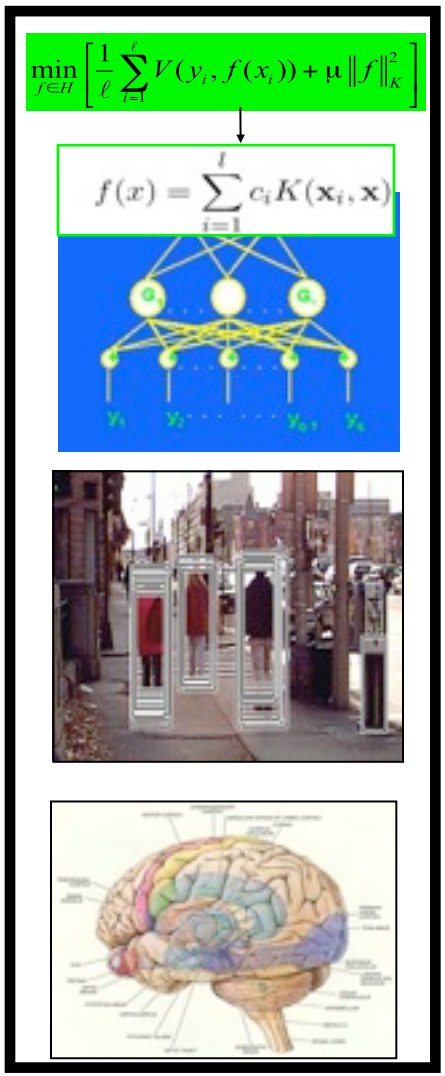
Predictive algorithms

Papageorgiou&Poggio, 1997, 2000
also Kanade&Scheiderman

COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

Learning



LEARNING THEORY
+
ALGORITHMS

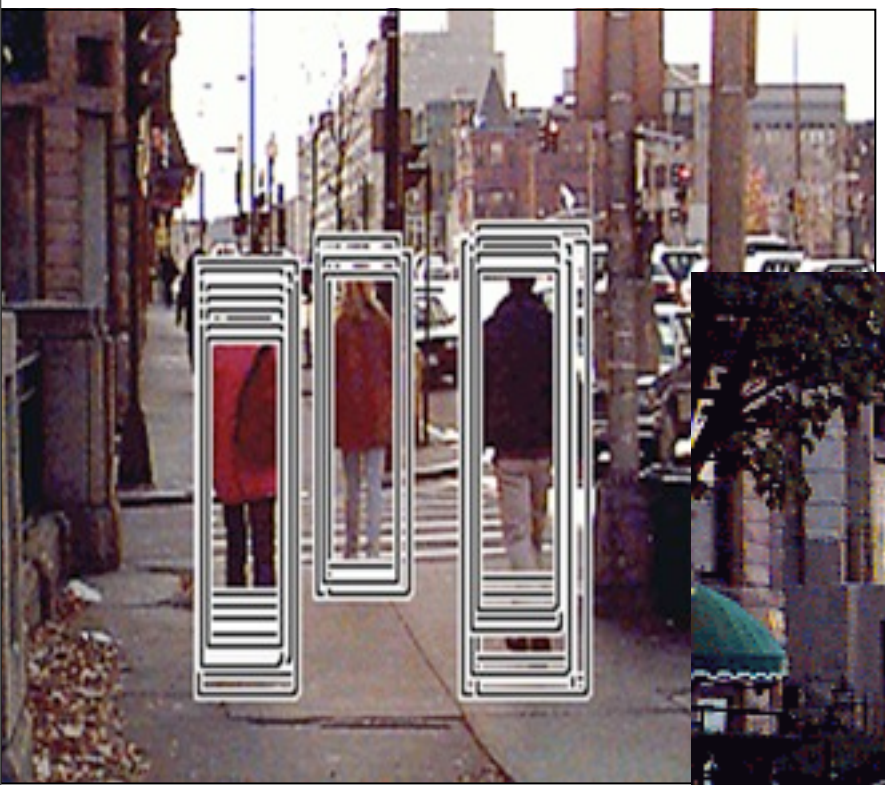
Theorems on foundations of learning

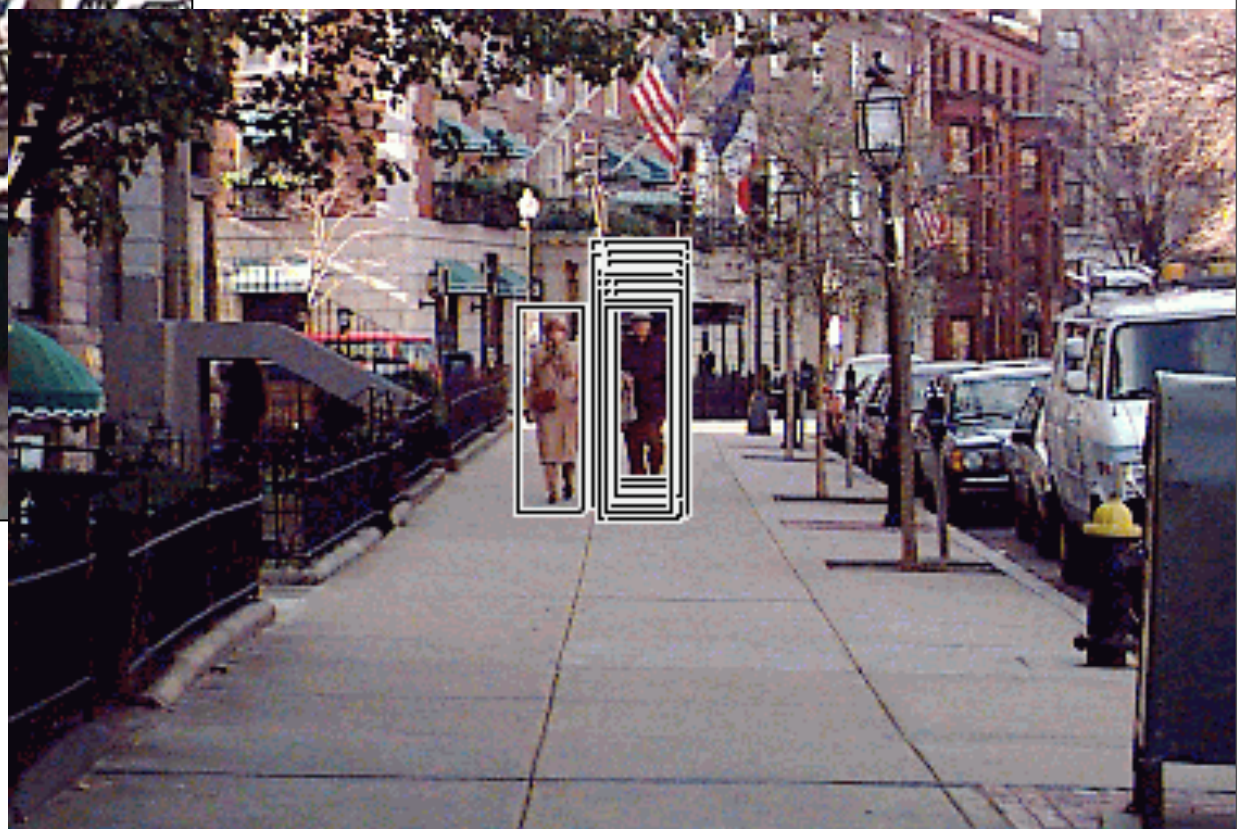
Predictive algorithms

Papageorgiou&Poggio, 1997, 2000
also Kanade&Scheiderman

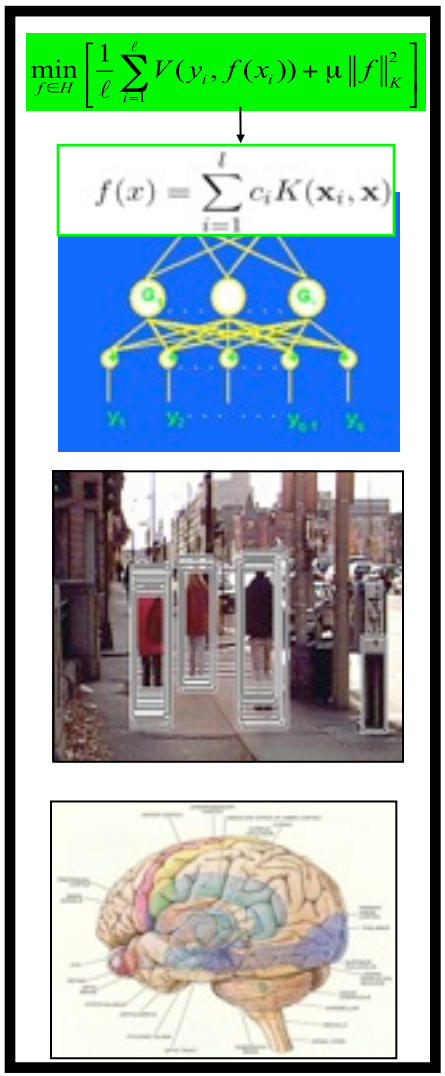
COMPUTATIONAL
NEUROSCIENCE:
models+experiments

How visual cortex works





Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

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

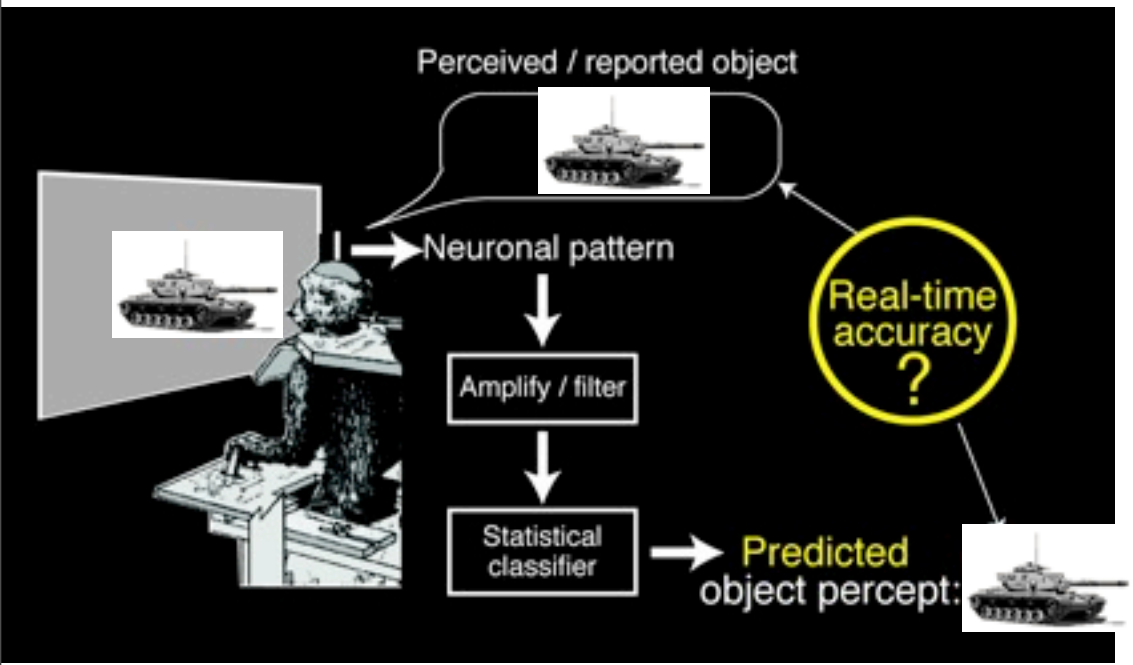
COMPUTATIONAL NEUROSCIENCE: models+experiments

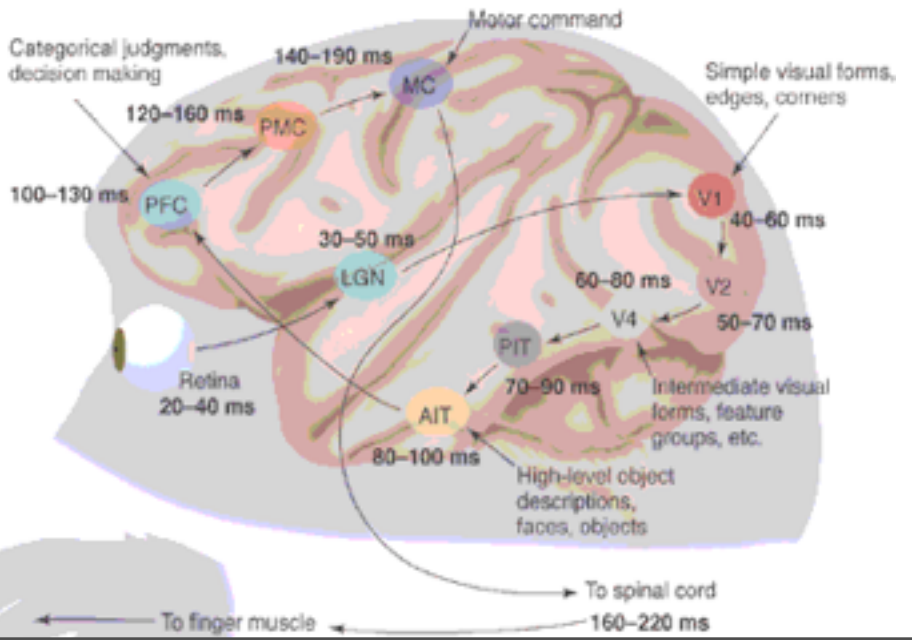
How visual cortex works

http://www.volvocars.com/us/all-cars/volvo-s60/pages/5-things.aspx?p=5

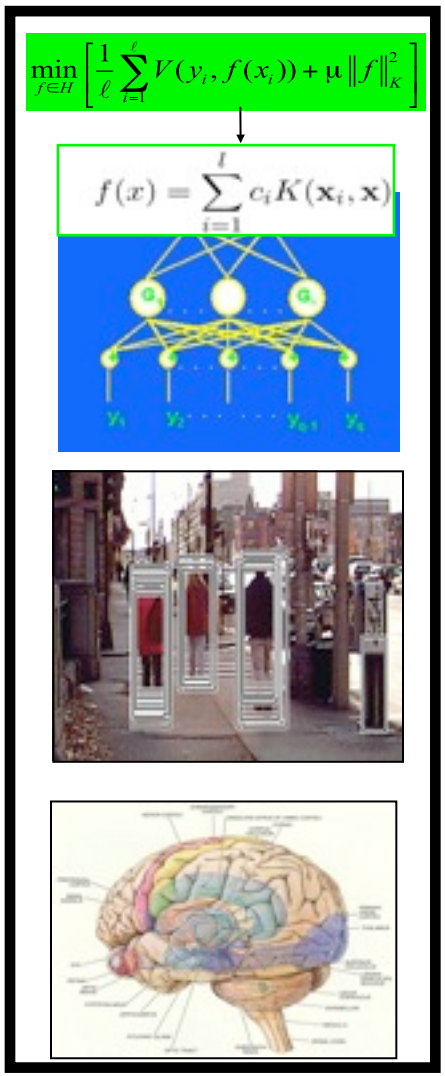
http://www.volvocars.com/us/all-cars/volvo-s60/pages/5-things.aspx?p=5

Learning: read-out of thoughts





Learning



LEARNING THEORY
+
ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

ENGINEERING
APPLICATIONS

- **Bioinformatics**
- Computer vision
- Computer graphics, speech synthesis, creating a virtual actor
- Neuroinformatics, read-out

COMPUTATIONAL
NEUROSCIENCE:
models+experiments

How visual cortex works

Learning: bioinformatics

New feature selection SVM:

Only 38 training examples, 7100 features

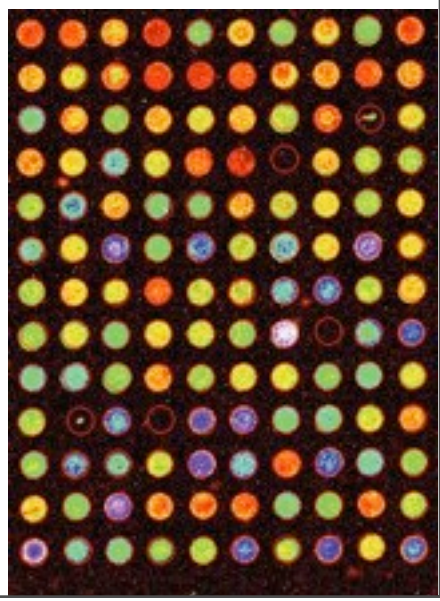
AML vs ALL: 40 genes 34/34 correct, 0 rejects.
5 genes 31/31 correct, 3 rejects of which 1 is an error.

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

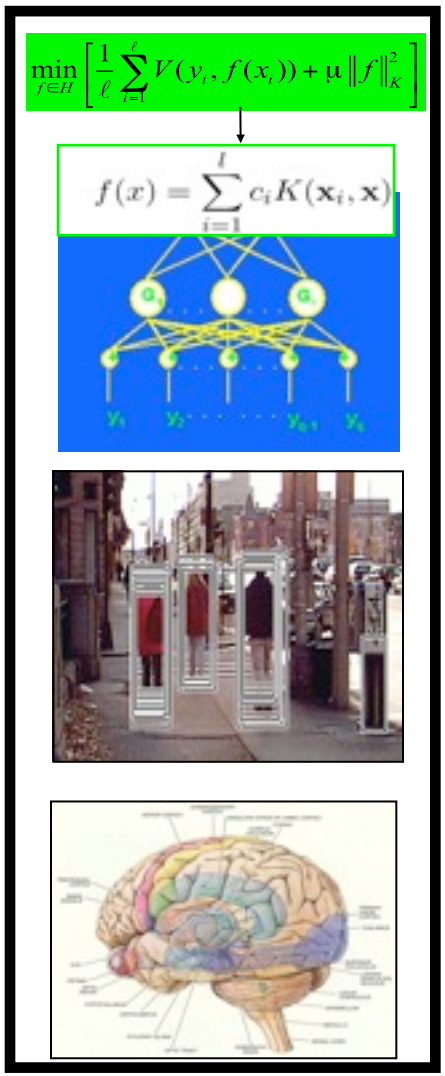
Support Vector Machine Classification of Microarray Data

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

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



Learning



LEARNING THEORY + ALGORITHMS

Theorems on foundations of learning

Predictive algorithms

ENGINEERING APPLICATIONS

- Bioinformatics
- Computer vision
- **Computer graphics, speech synthesis**
- Neuroinformatics, read-out

COMPUTATIONAL NEUROSCIENCE: models+experiments

How visual cortex works

Learning: image analysis

⇒ **Bear (0° view)**

⇒ **Bear (45° view)**

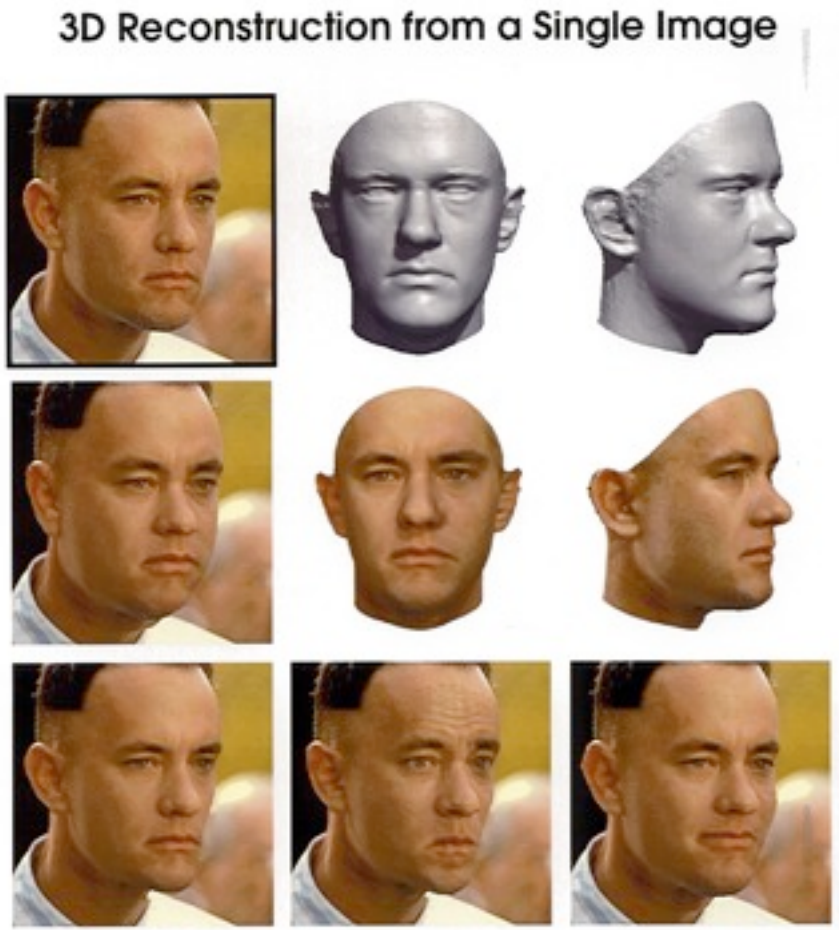
Learning: image synthesis

UNCONVENTIONAL GRAPHICS

Θ = 0° view $\Rightarrow$

Θ = 45° view $\Rightarrow$

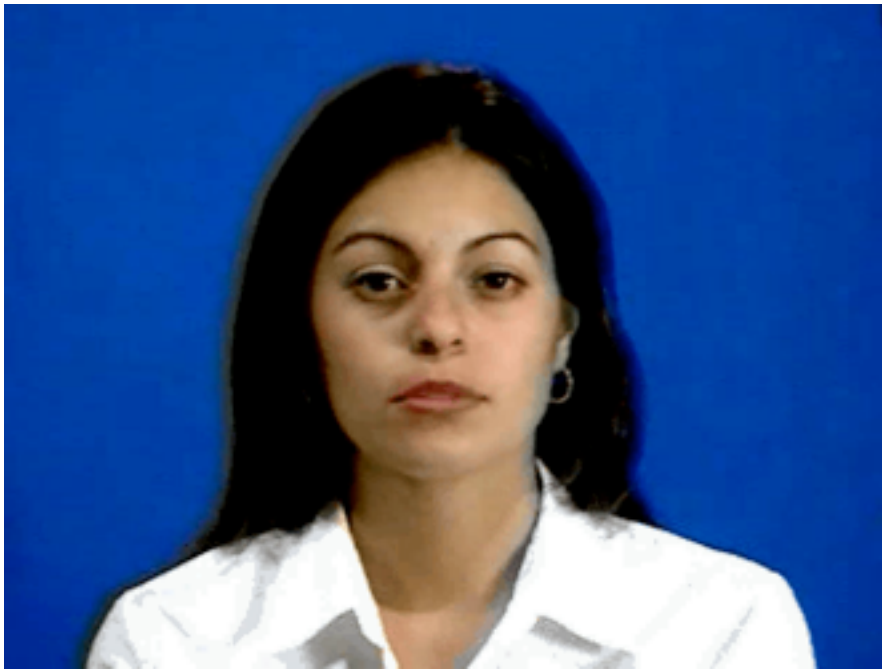
Learning: image synthesis



Blanz and Vetter,
MPI
SigGraph '99

Learning: image synthesis

Blanz and Vetter, MPI
SigGraph '99



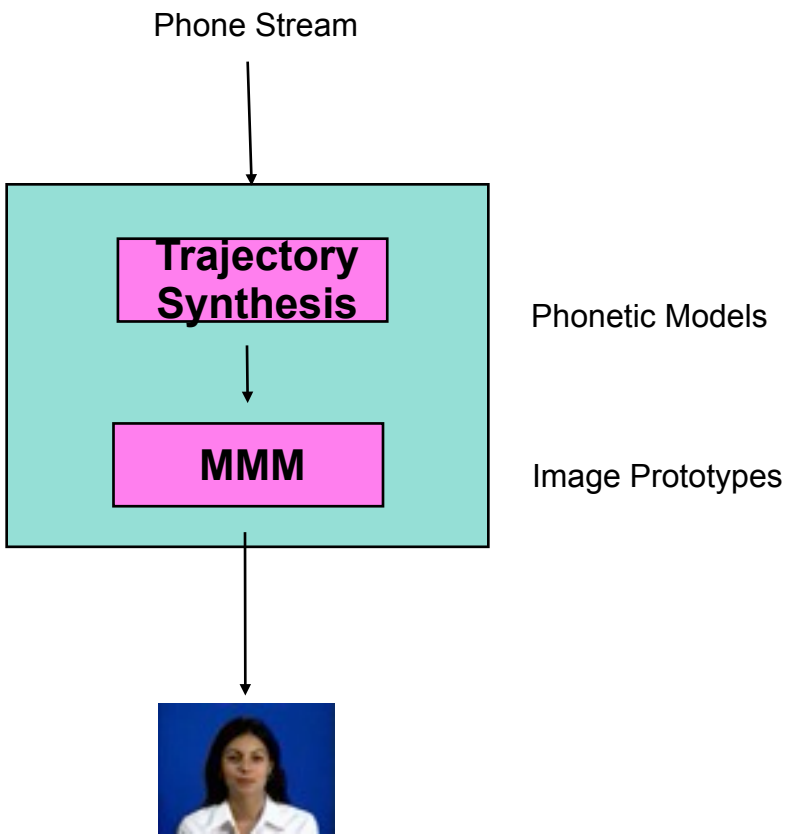
A- more in a moment

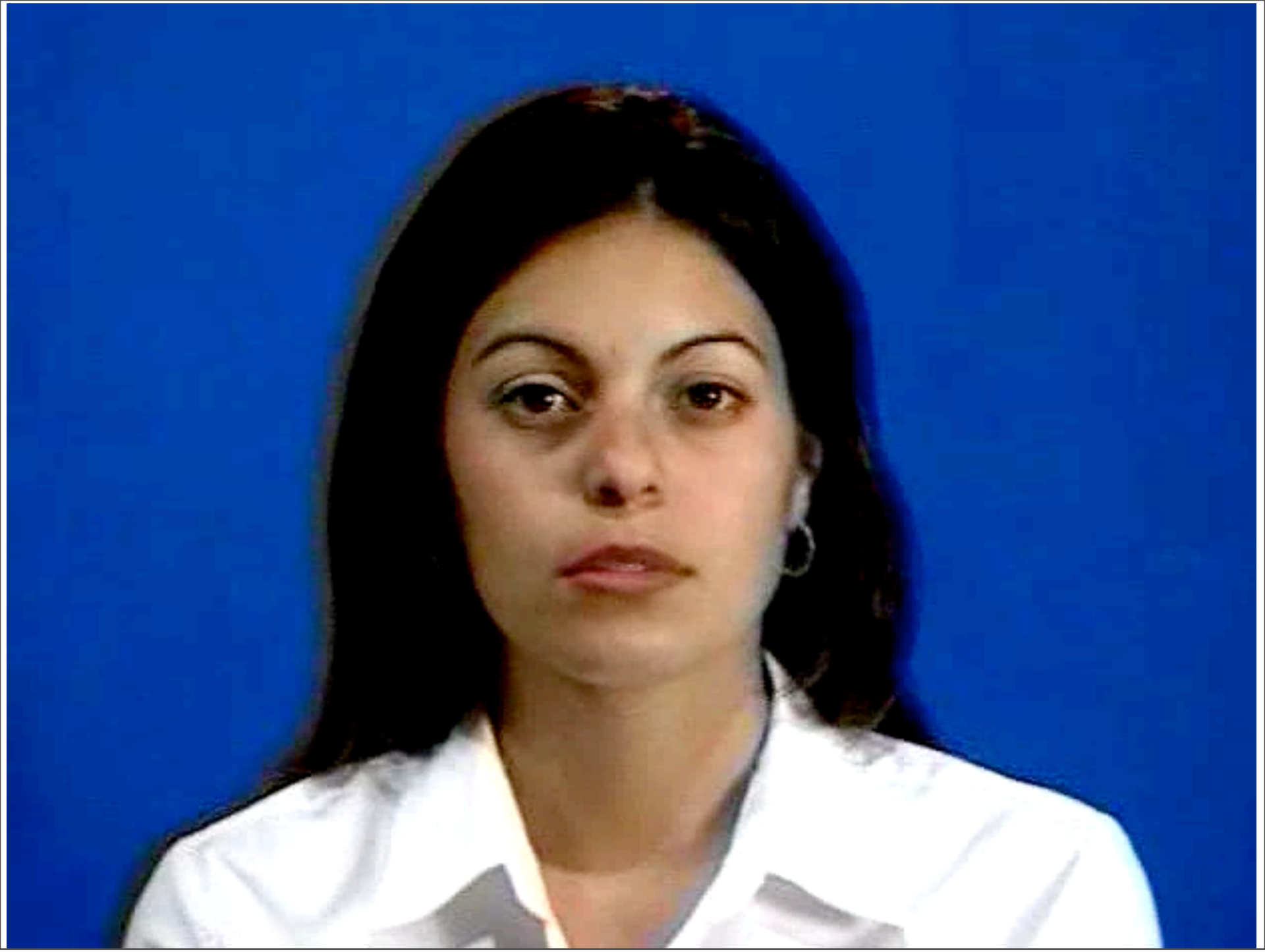
1. Learning

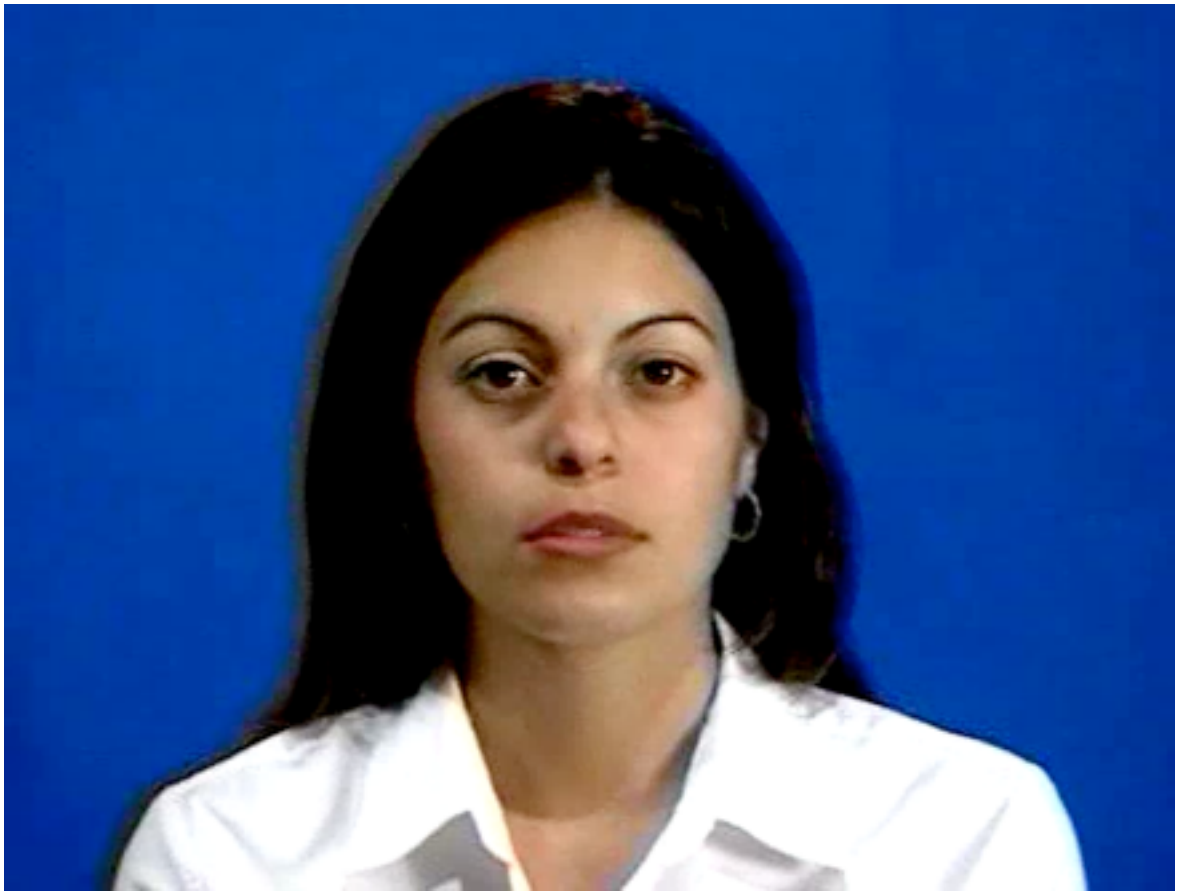
System learns from 4 mins of video face appearance (Morphable Model) and speech dynamics of the person

2. Run Time

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

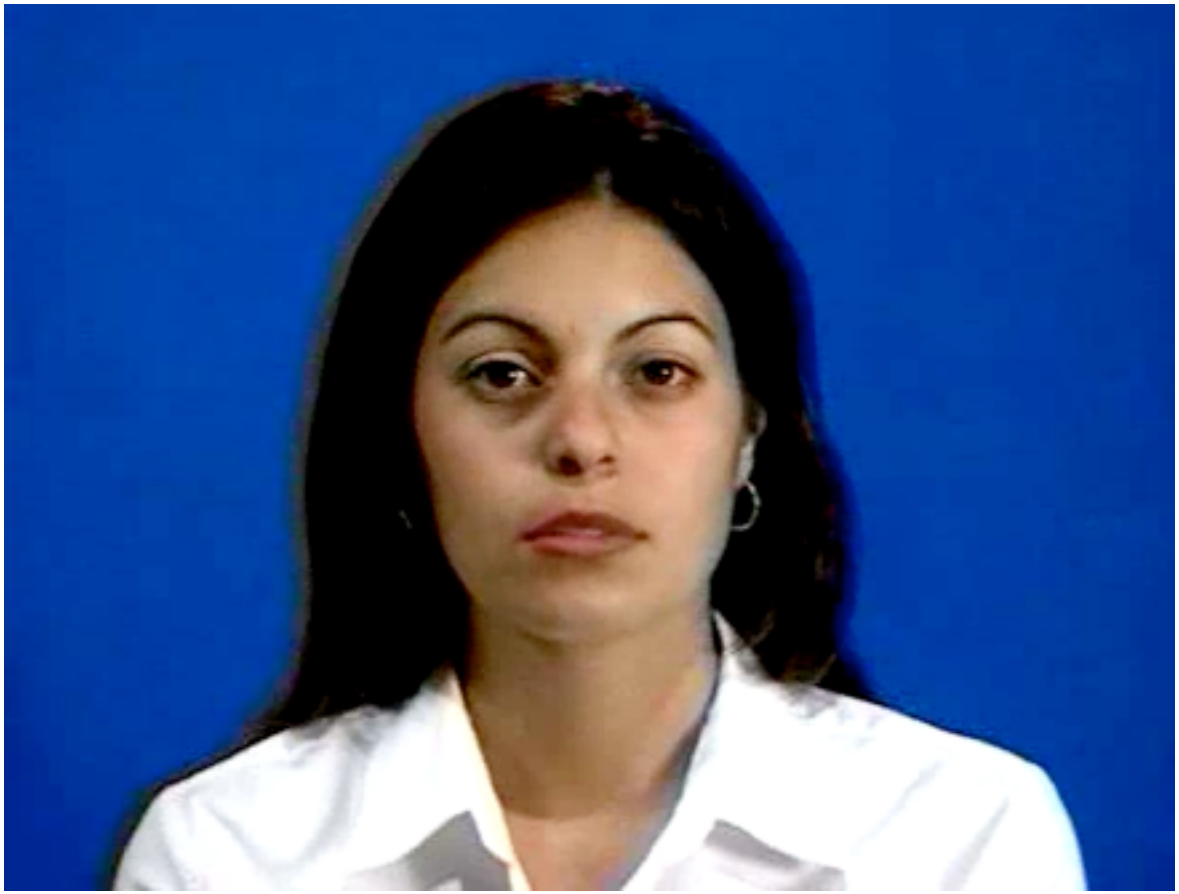






B-Dido

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?

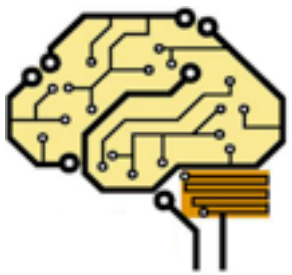
Experiment	# subjects	% correct	t	p<
Single pres.	22	54.3%	1.243	0.3
Fast single pres.	21	52.1%	0.619	0.5
Double pres.	22	46.6%	-0.75	0.5

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**

Overview of overview

- Context for this course: a golden age for new AI and the key role of Machine Learning
- Success stories from past research in Machine Learning: examples of engineering applications
- Statistical Learning Theory
- A new kind of basic research on learning based on computer science and neuroscience, learning and the brain: A Center for Brains, Minds and Machines
- A new phase in Machine Learning?



The Center for Brains, Minds and Machines

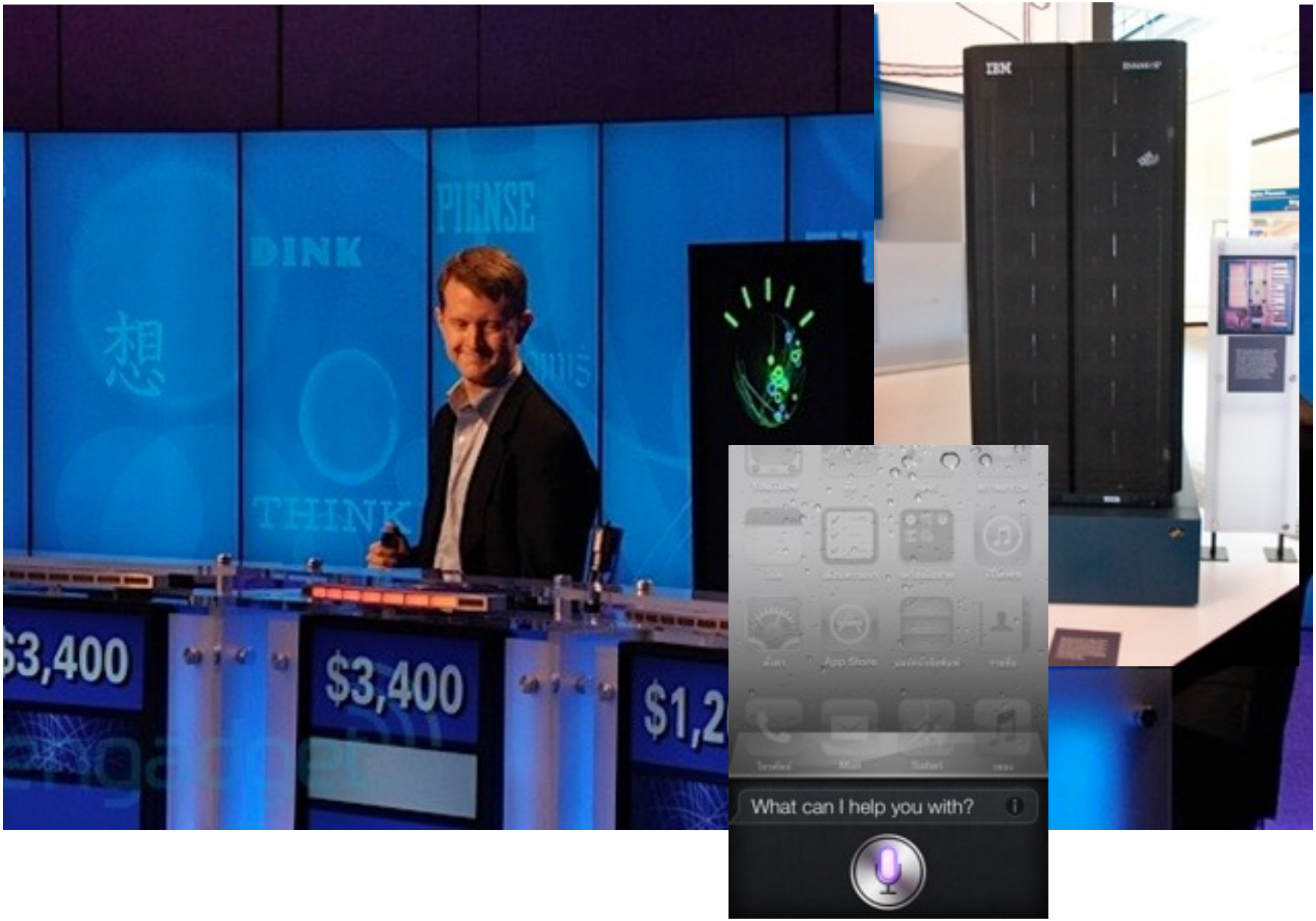
Vision for CBMM

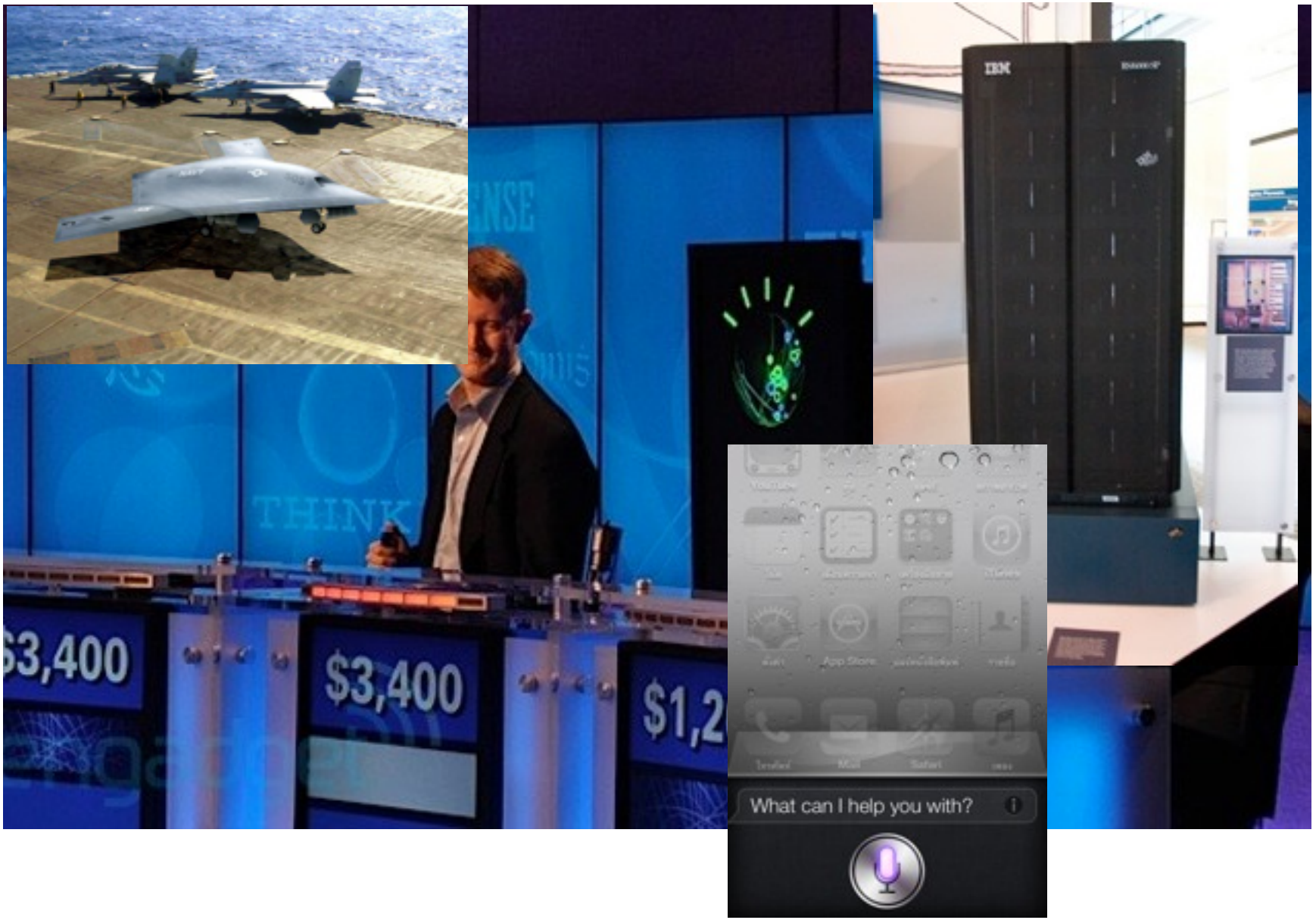
Vision for CBMM

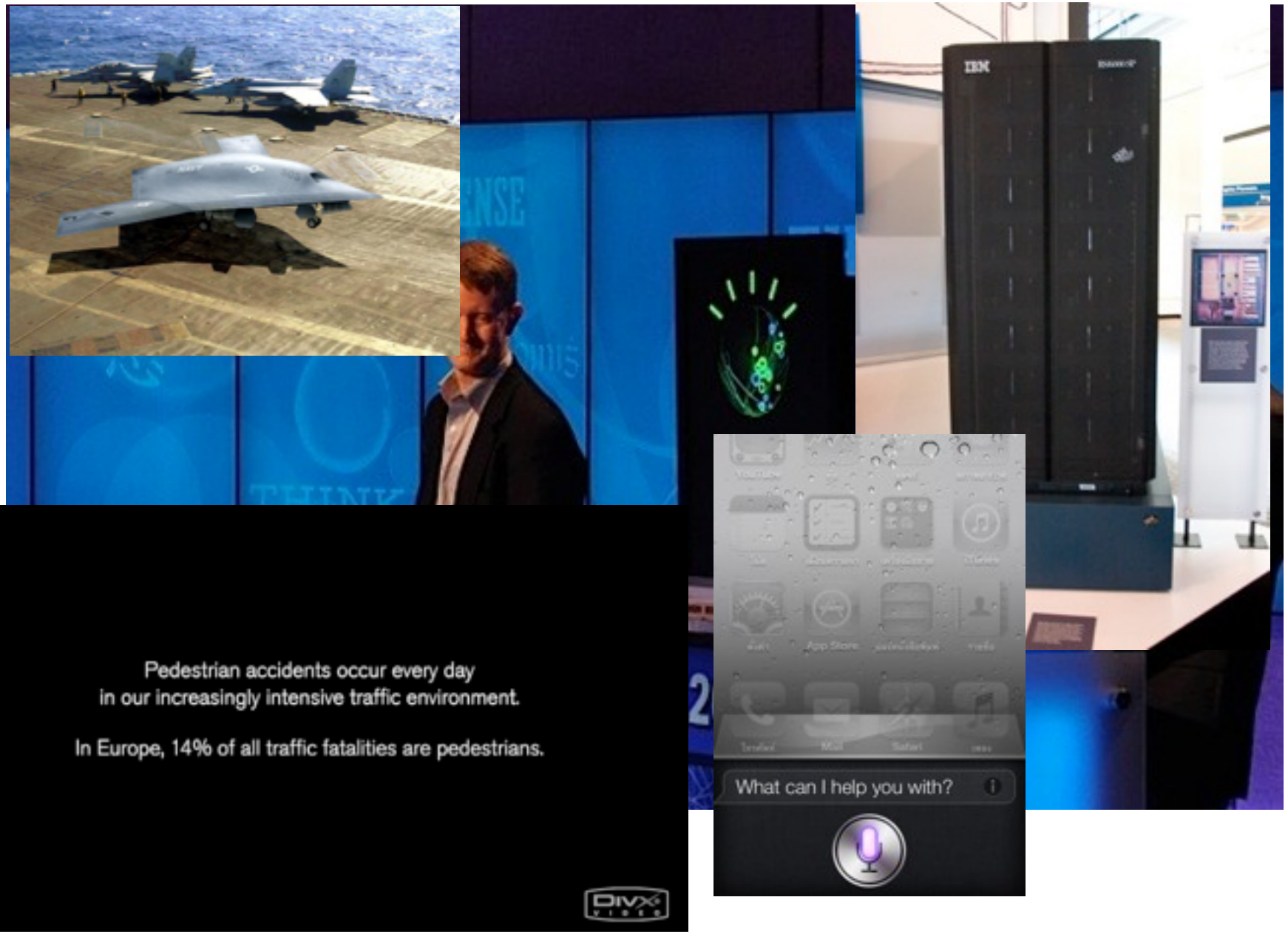
- The problem of intelligence is one of the great problems in science.
- Work so far has led to many systems with impressive but narrow intelligence
- Now it is time to develop a deep computational understanding of human intelligence for its own sake and so that we can take intelligent applications to another level.


PIENSE
THINK
想
$3,400
$3,400
$1,200


PIENSE
THINK
$3,400
$3,400
$1,200
IBM


PIENSE
DINK
想
THINK
$3,400
$3,400
$1,2
IBM
What can I help you with?


IBM
$3,400
$3,400
What can I help you with?


Pedestrian accidents occur every day
in our increasingly intensive traffic environment.
In Europe, 14% of all traffic fatalities are pedestrians.
What can I help you with?

The problem is that none of the systems is able to pass a full Turing test

MIT

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

Harvard

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

Cornell

Hirsh

Allen Institute

Koch

Rockfeller

Freiwald

UCLA

Yuille

Stanford

Goodman

Hunter

Epstein,...

Wellesley

Hildreth, Conway...

Puerto Rico

Bykhovaskaia, Vega...

Howard

Manaye,...

City U. HK
Smale

Hebrew U.
Shashua

IIT
Metta, Rosasco, Sandini

MPI
Buelthoff

NCBS
Raghavan

Genoa U.
Verri

Weizmann
Ullman



Google
Norvig

IBM
Ferrucci

Microsoft
Blake

Orcam
Shashua

MobilEye
Shashua

DeepMind
Hassabis

Boston Dynamics
Raibert

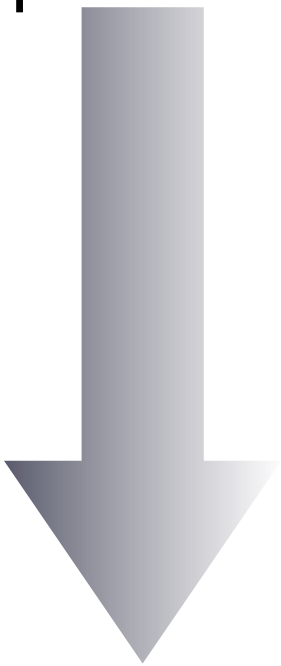
Rethink Robotics
Brooks

Willow Garage
Cousins


Rational for a Center

Convergence of progress: a key opportunity

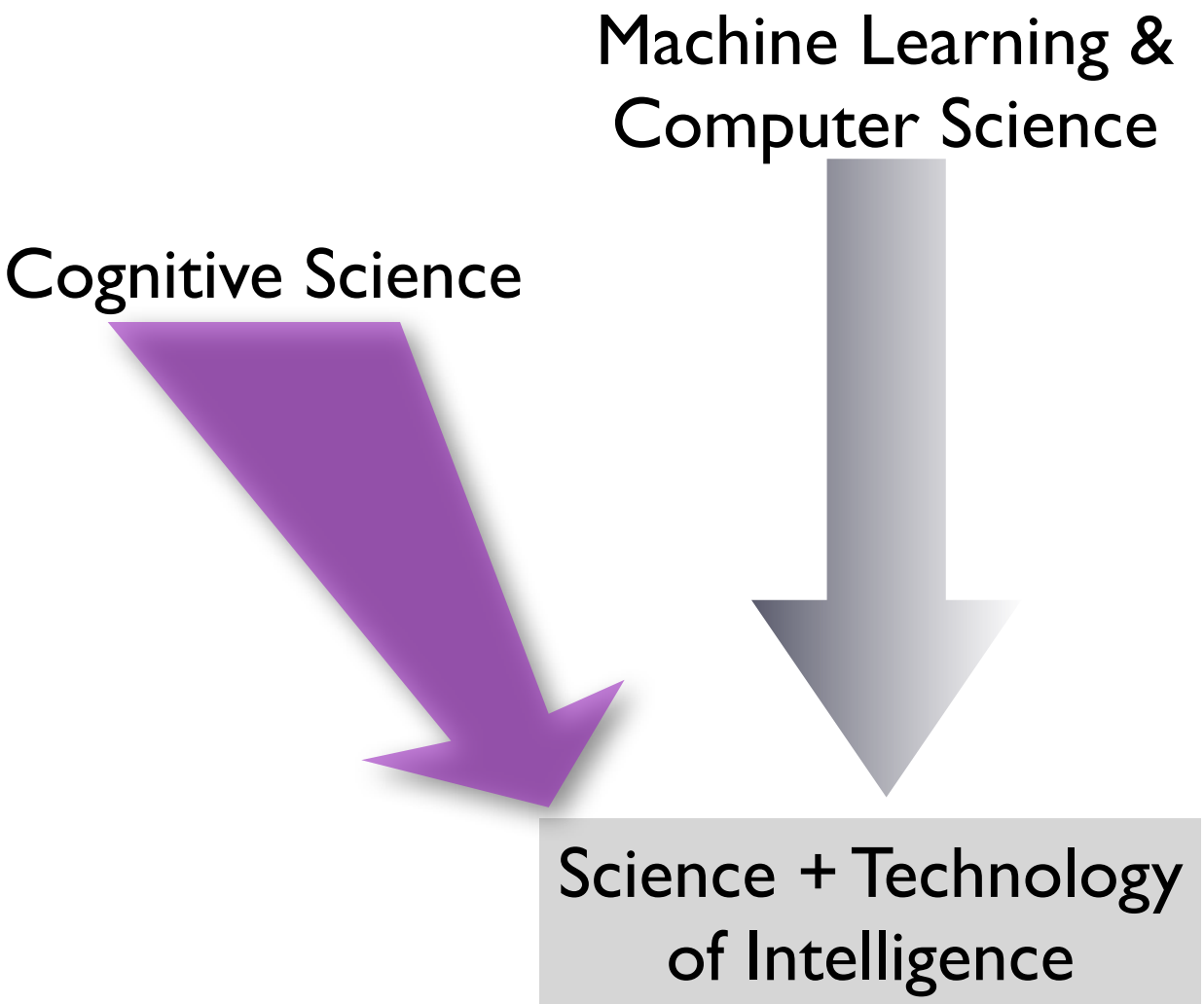
Machine Learning &
Computer Science



Science + Technology
of Intelligence

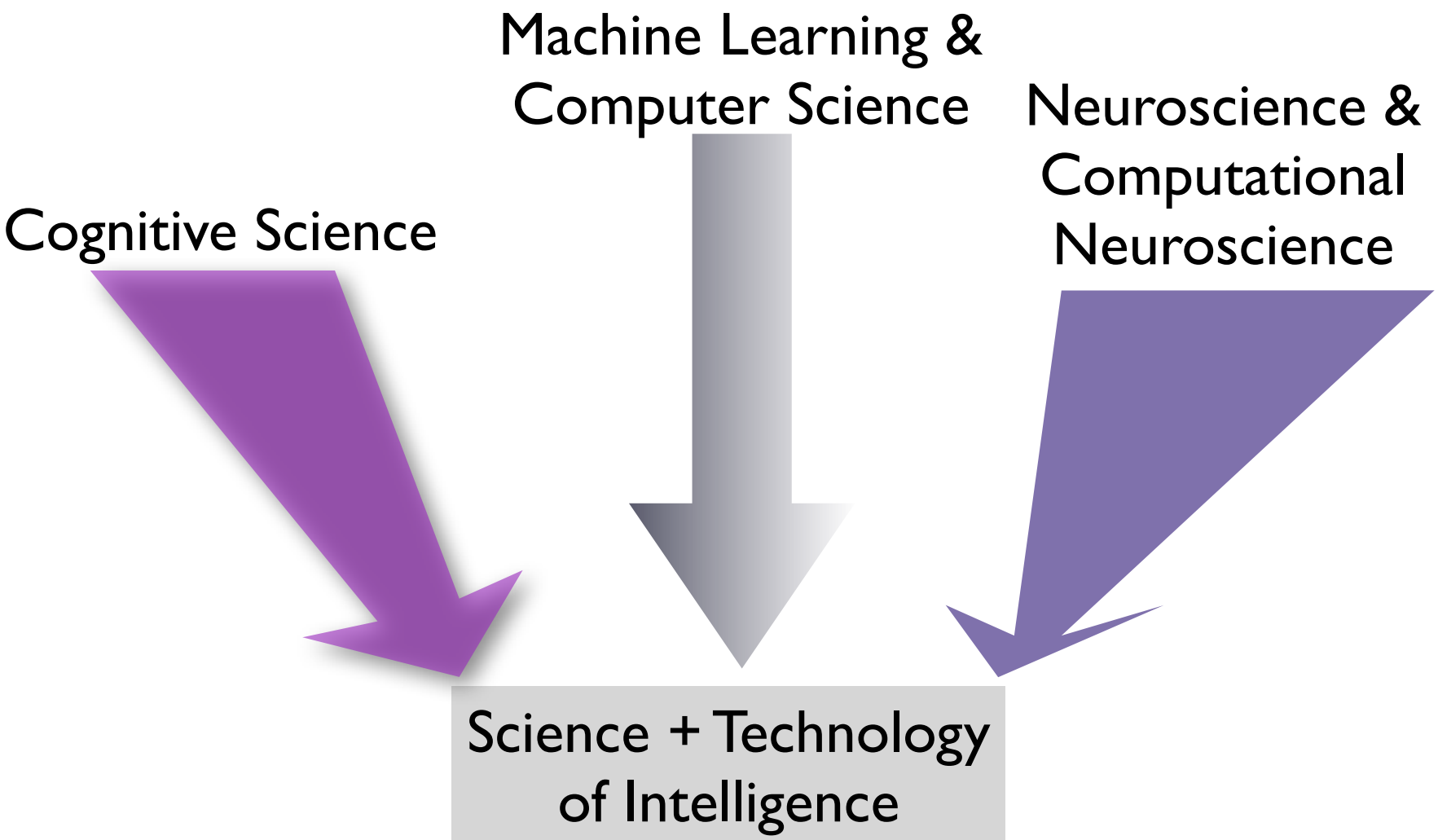
Rational for a Center

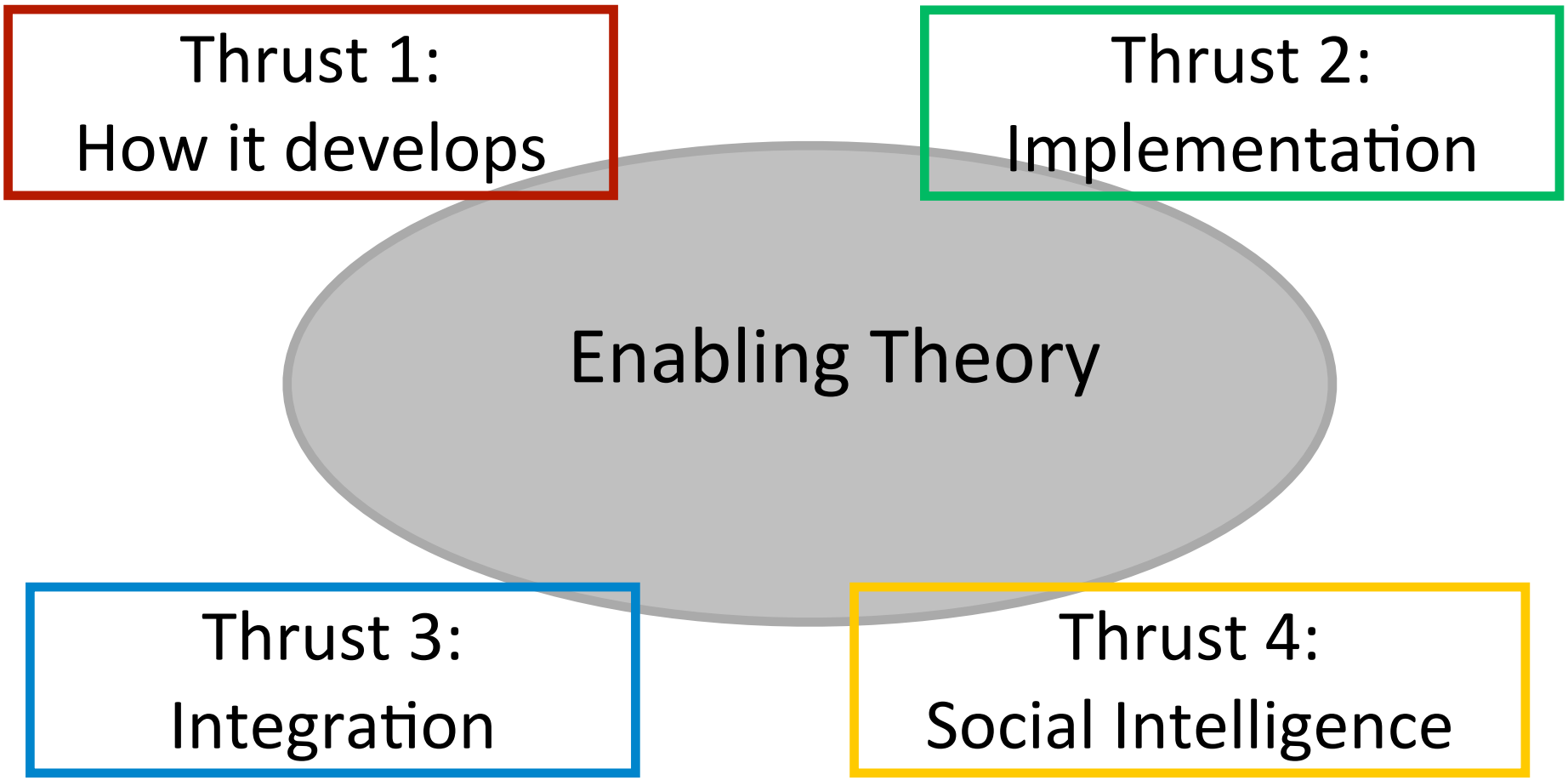
Convergence of progress: a key opportunity



Rational for a Center

Convergence of progress: a key opportunity




Thrust 1:
How it develops
Thrust 2:
Implementation
Enabling Theory
Thrust 3:
Integration
Thrust 4:
Social Intelligence

Example: A “Turing’’ test for vision


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

Example: A "Turing" test for vision

My personal bet: we may need to understand cortex (and the brain!) to achieve, for instance, scene understanding at human level, and thereby develop systems that pass a full Turing test.

Example : a second phase in machine learning

The first phase (and successes) of ML: supervised learning "n-->∞"

The next phase of ML: unsupervised learning of invariant representations for learning "n--> 0"

Contributions to come

- A better understanding of ourselves
- A better understanding of each other
- A substantial contribution to our societies future education, prosperity, health, and security

Overview of overview

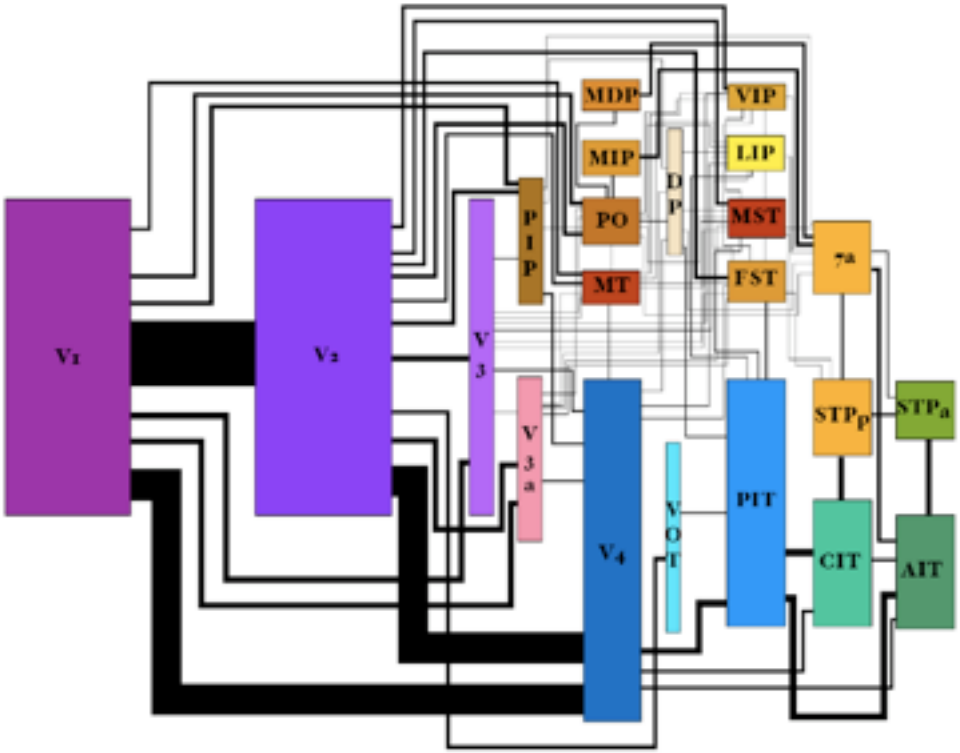
- Context for this course: a golden age for new AI and the key role of Machine Learning
- Success stories from past research in Machine Learning: examples of engineering applications
- Statistical Learning Theory
- A new kind of basic research on learning: computer science and neuroscience, learning and the brain: A Center for Brains, Minds and Machines
- A new phase in Machine Learning?

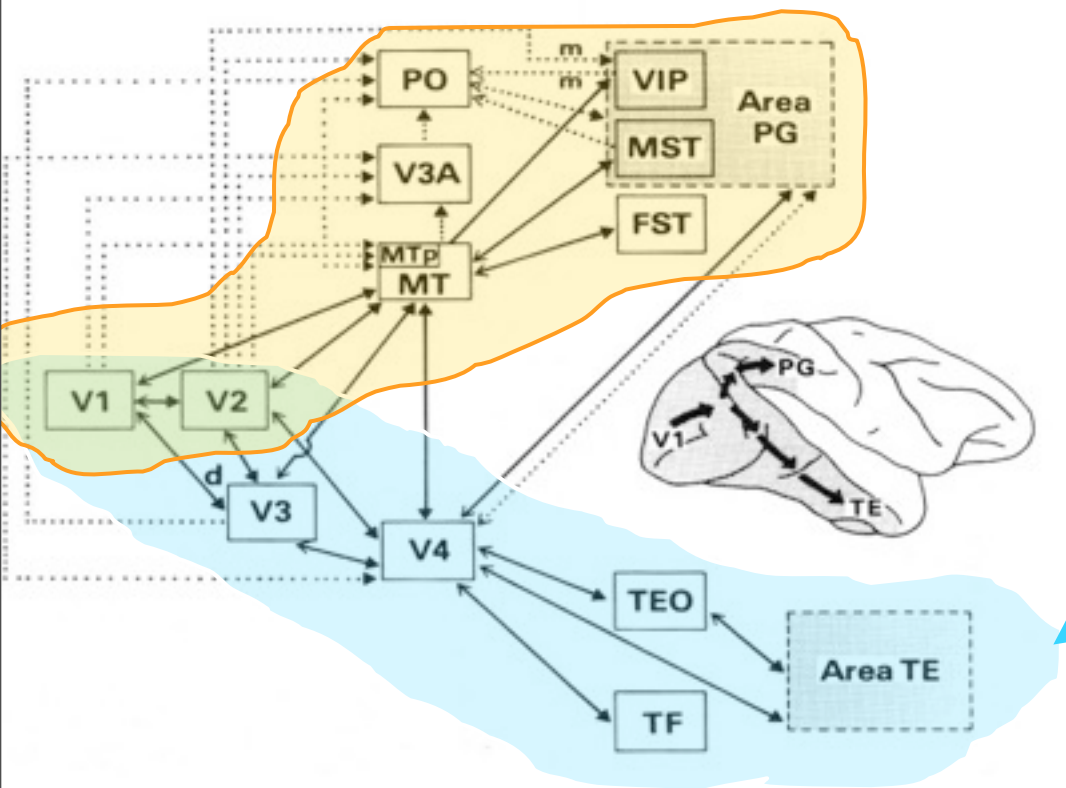
Example : a second phase in machine learning

The first phase (and successes) of ML: supervised learning "n-->∞"

The next phase of ML: unsupervised learning of invariant representations for learning "n--> 0"

- Human Brain
  - 10^{10}-10^{11} neurons (~1 million flies)
  - 10^{14}- 10^{15} synapses

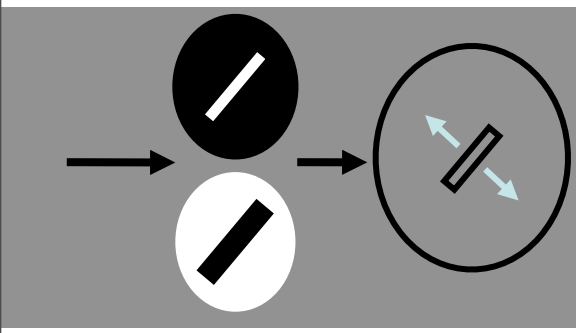


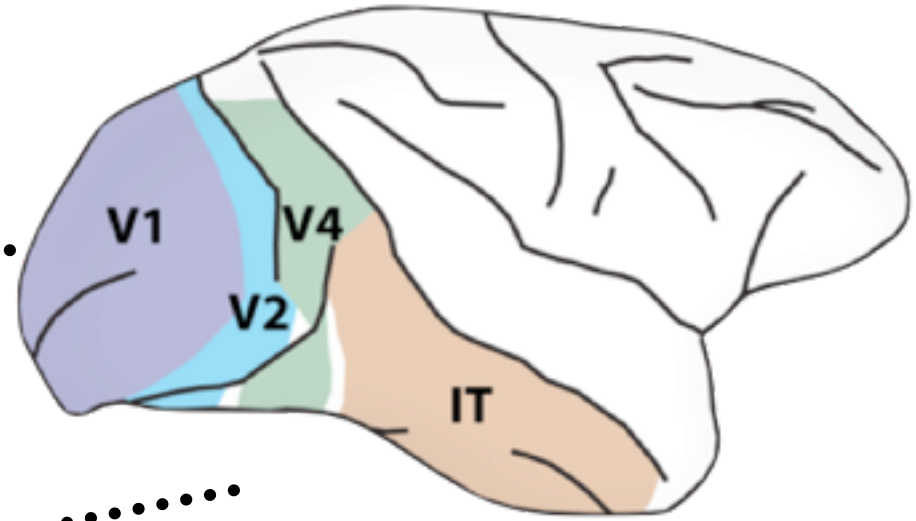


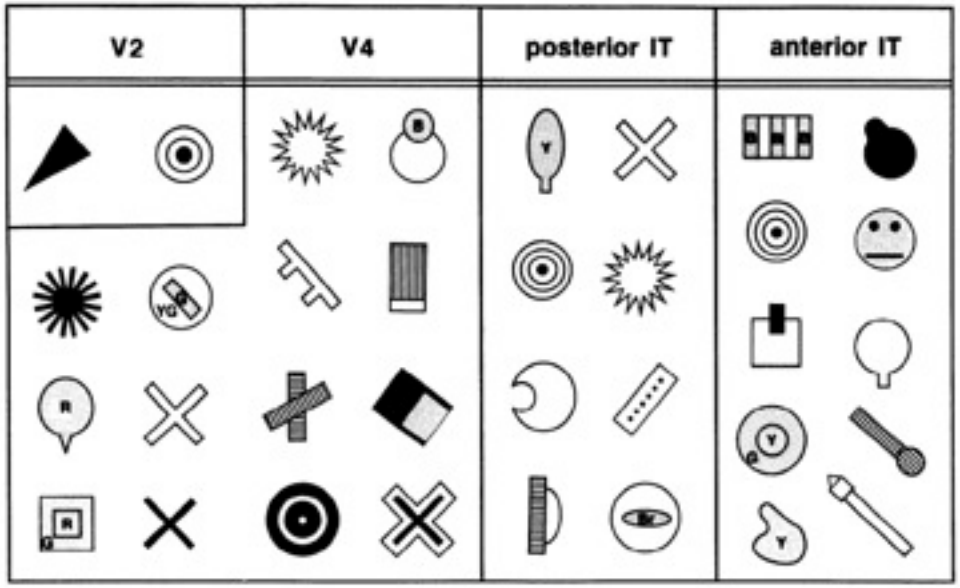
Ventral stream in rhesus monkey

- ~10^9 neurons in the ventral stream ($350\ 10^6$ in each emisphere)
- ~$15\ 10^6$ neurons in AIT

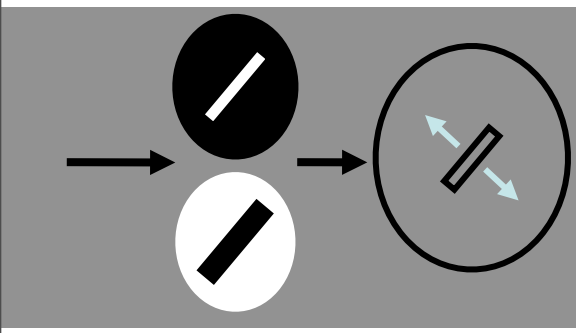
Vision: ventral stream

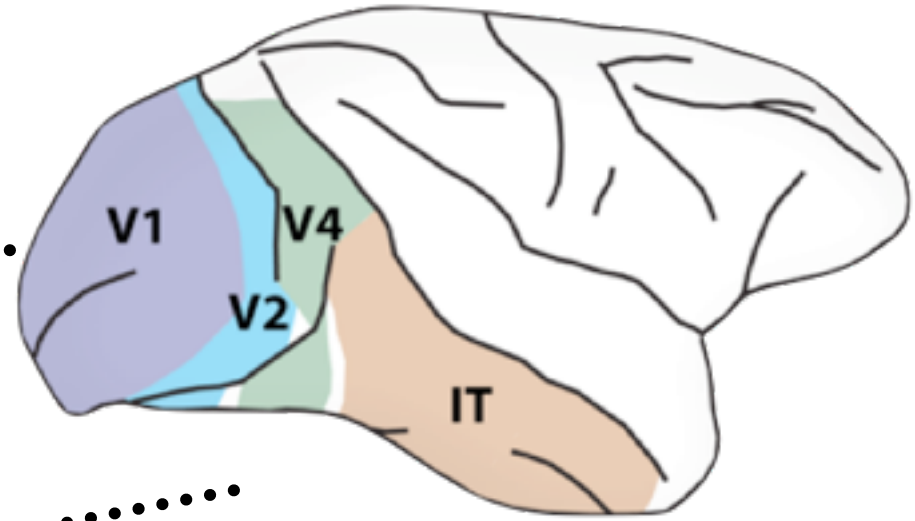


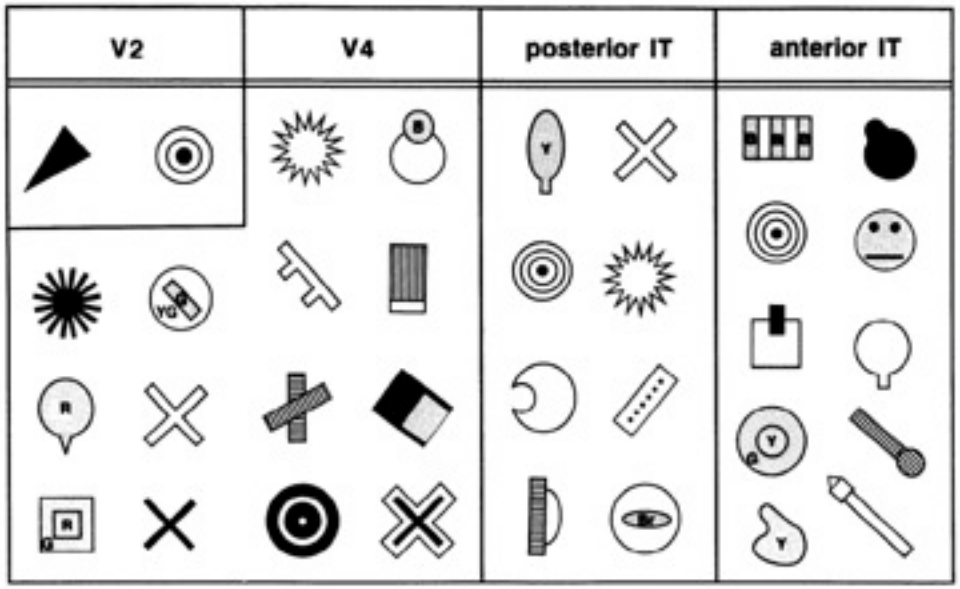




Vision: ventral stream







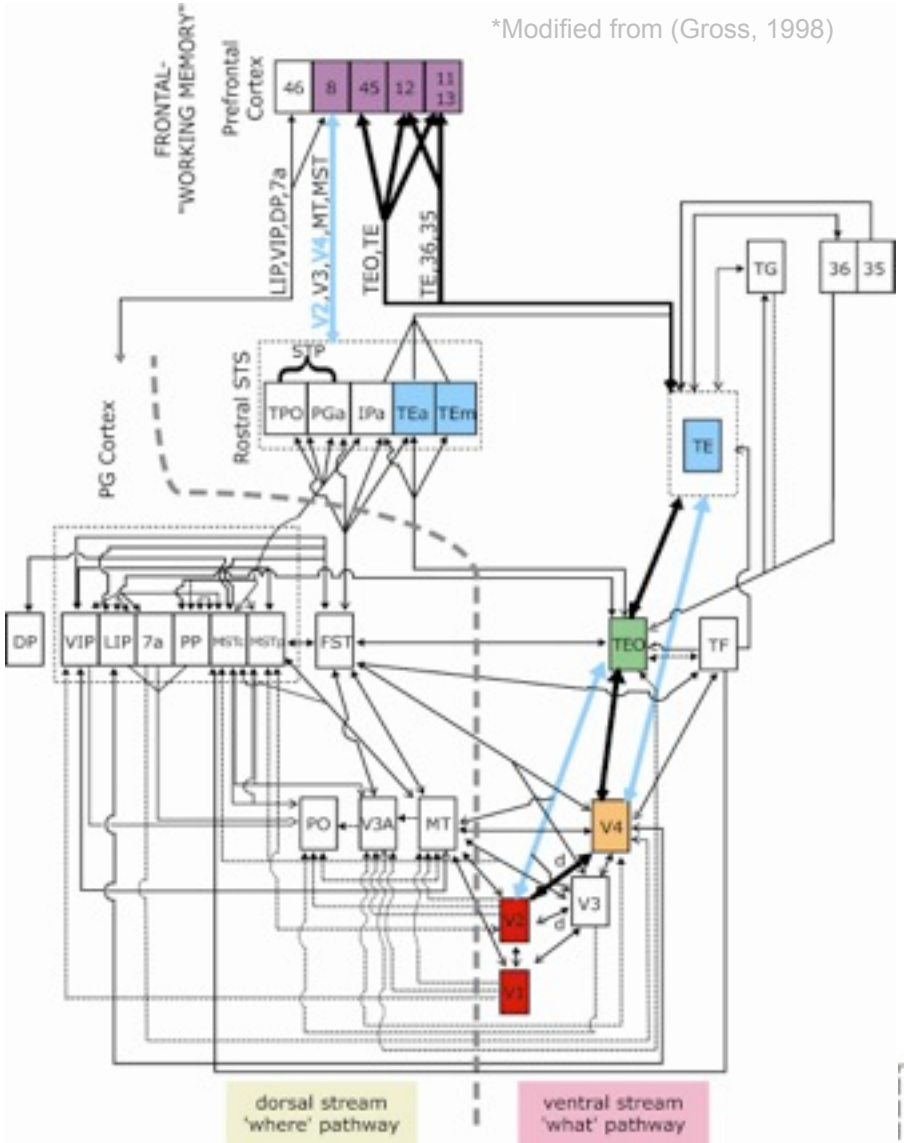
The ventral stream hierarchy: V1, V2, V4, IT

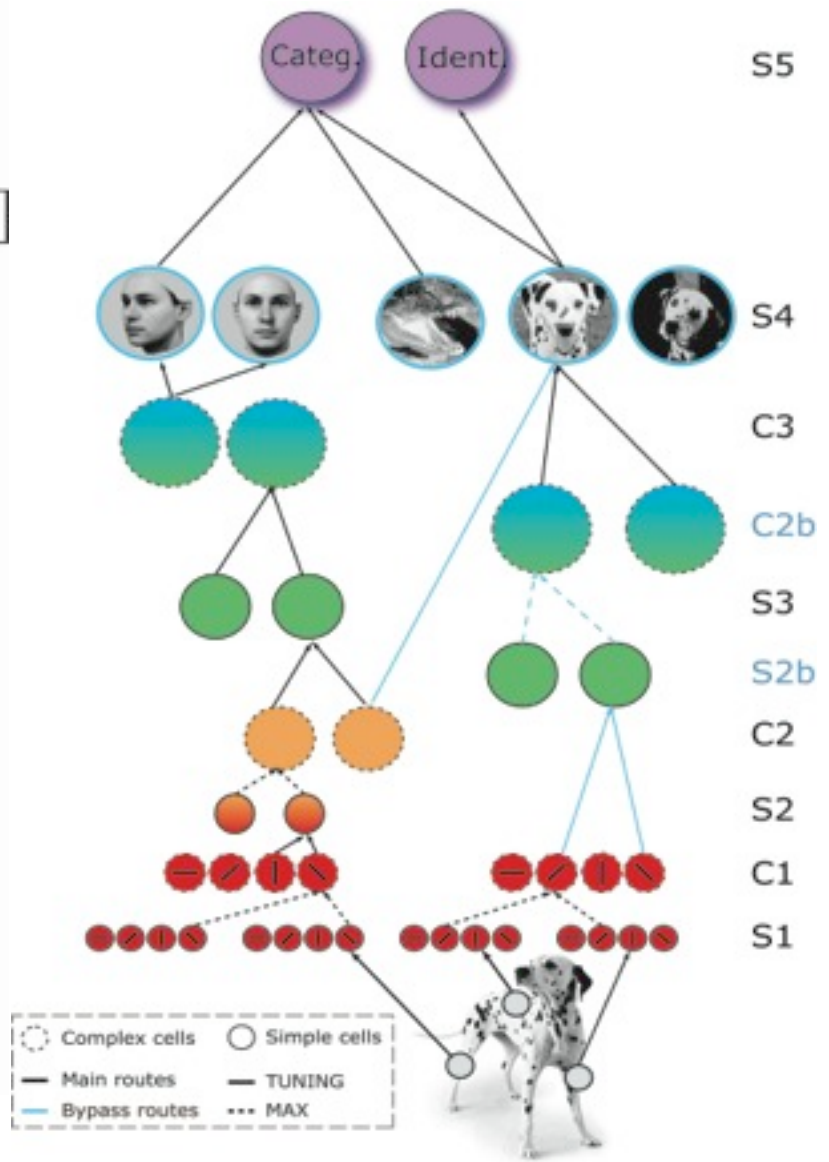
A gradual increase in the receptive field size, in the "**complexity**" of the preferred stimulus, in **"invariance"** to position and scale changes

Kobatake & Tanaka, 1994

Learning in visual cortex

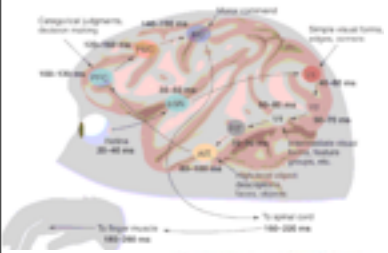
*Modified from (Gross, 1998)

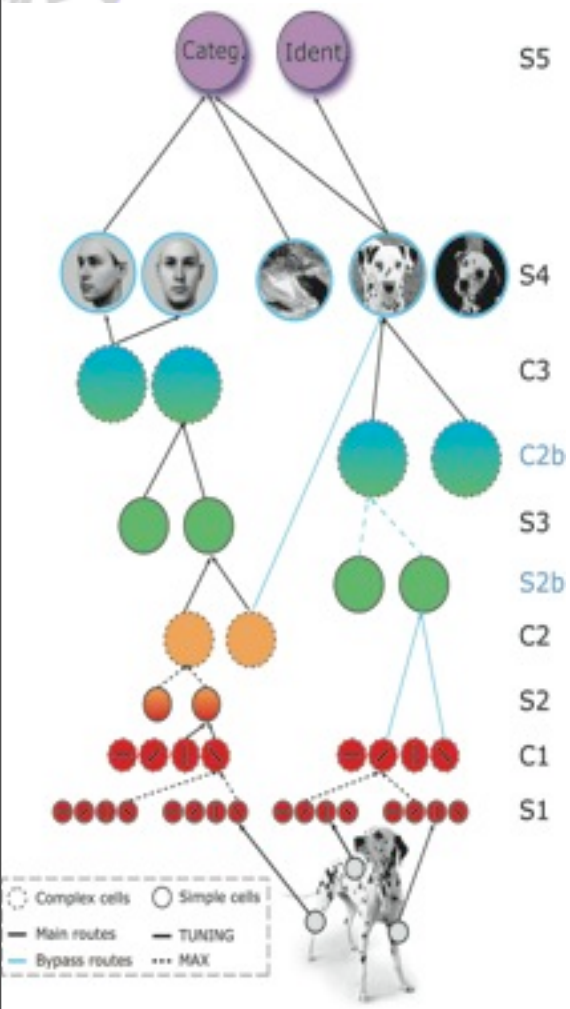


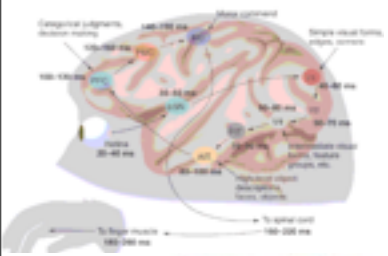


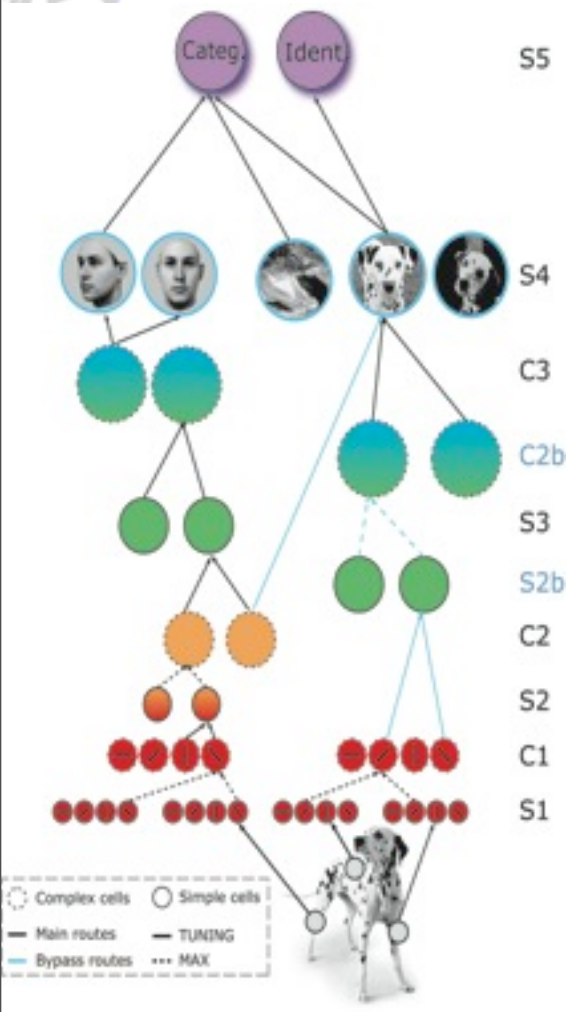
[software available online with CNS (for GPUs)]

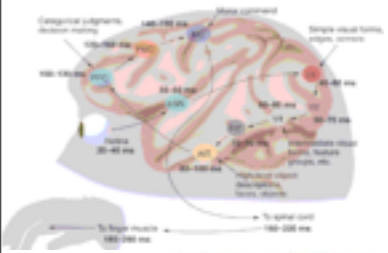
Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

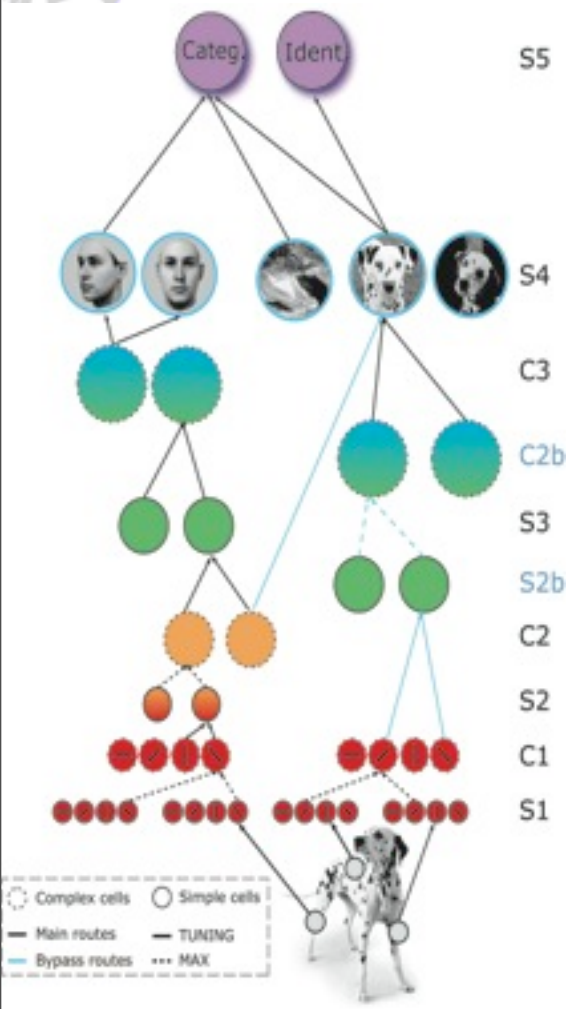


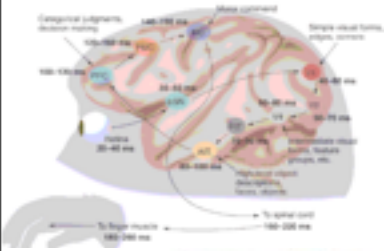

Categ.
Ident.
S5
S4
C3
C2b
S3
S2b
C2
S2
C1
S1
Complex cells
Simple cells
Main routes
TUNING
Bypass routes
MAX

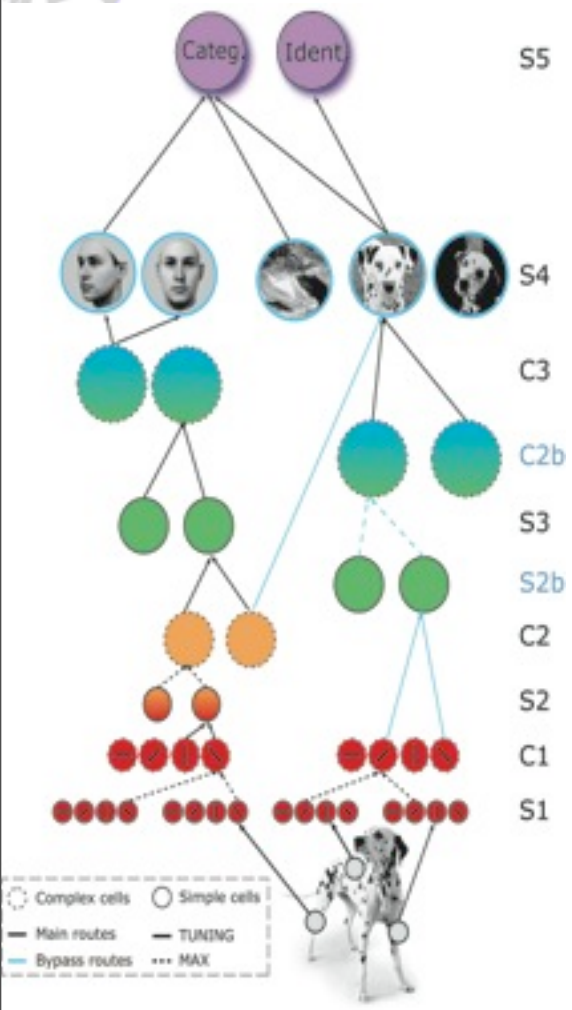


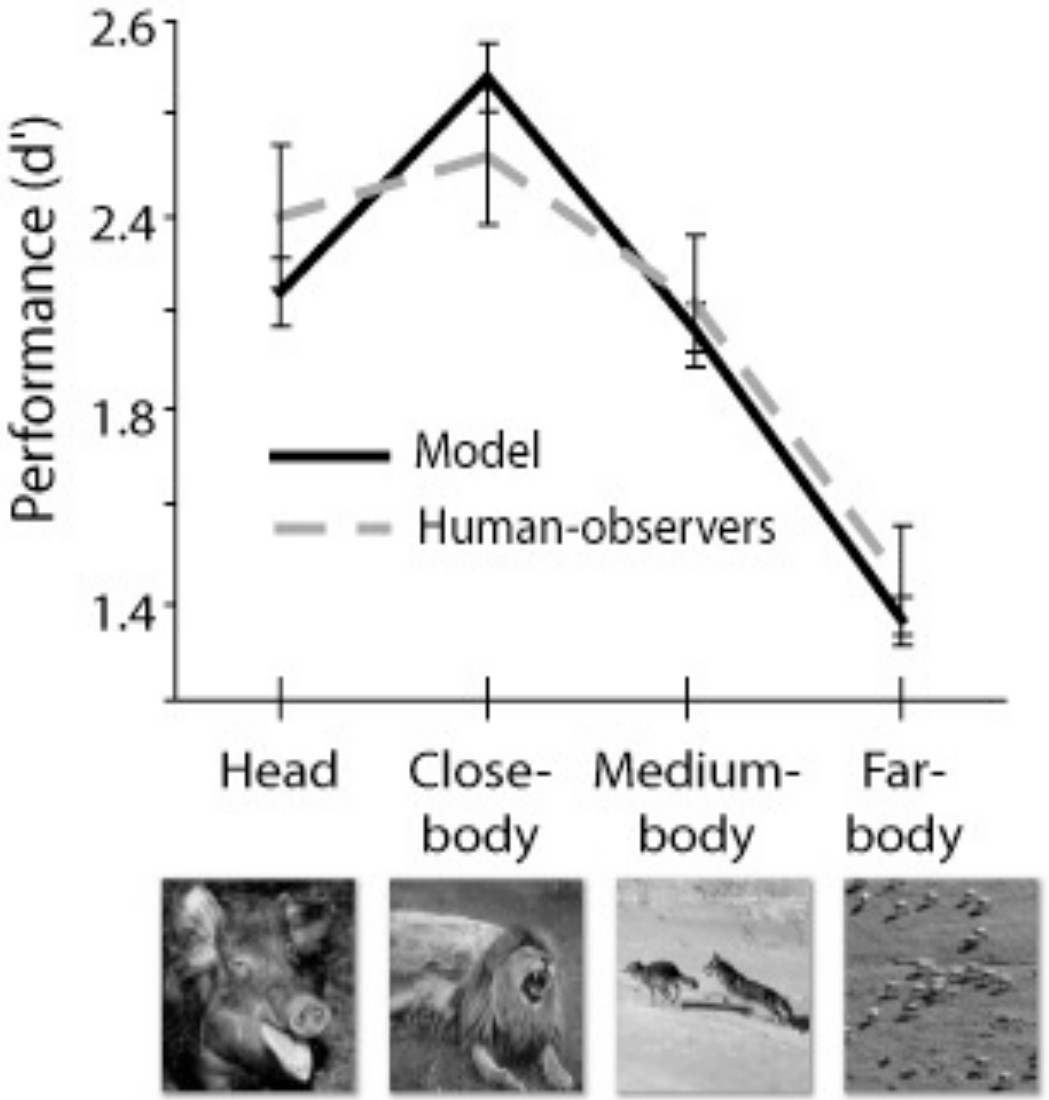

Categ.
Ident.
S5
S4
C3
C2b
S3
S2b
C2
S2
C1
S1
Complex cells
Simple cells
Main routes
TUNING
Bypass routes
MAX




Categ.
Ident.
S5
S4
C3
C2b
S3
S2b
C2
S2
C1
S1
Complex cells
Simple cells
Main routes
TUNING
Bypass routes
MAX




Categ.
Ident.
S5
S4
C3
C2b
S3
S2b
C2
S2
C1
S1
Complex cells
Simple cells
Main routes
TUNING
Bypass routes
MAX

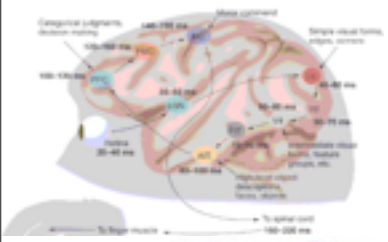

Performance (d')
2.6
2.4
1.8
1.4
Model
Human-observers
Head
Close-body
Medium-body
Far-body

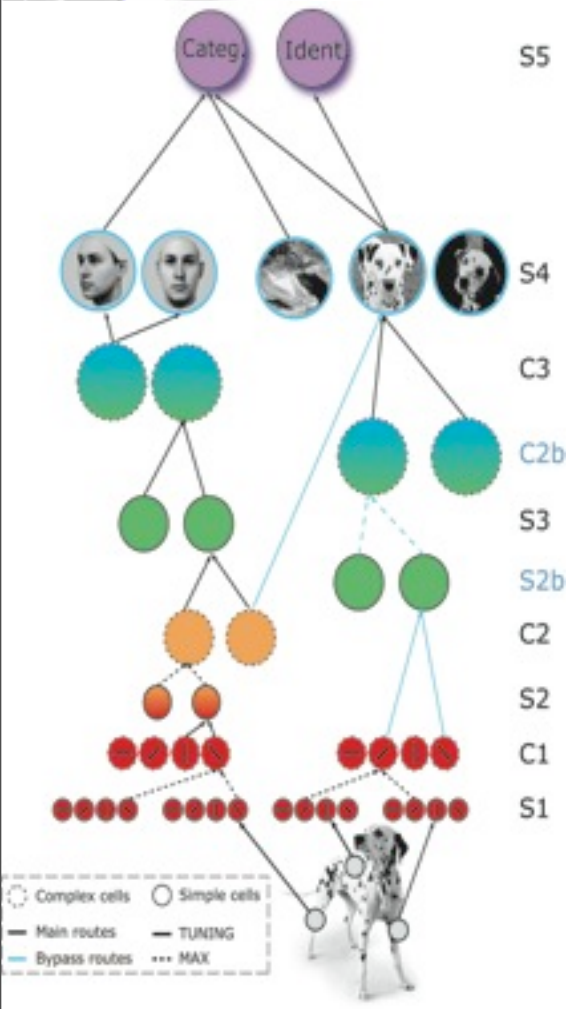
…predicts and is consistent with neural data…

- 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)

Neuroscience to good computer vision

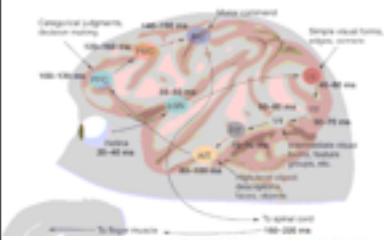


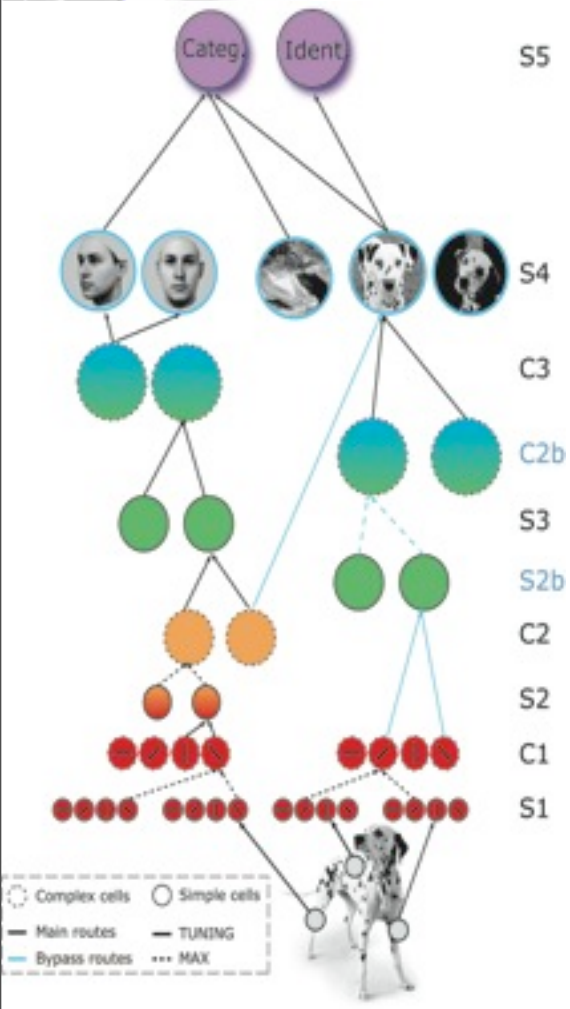


Models of the *ventral stream* in cortex
perform well compared to
engineered computer vision systems (in 2006)
on several databases

Bileschi, Wolf, Serre, Poggio, 2007

Neuroscience to good computer vision

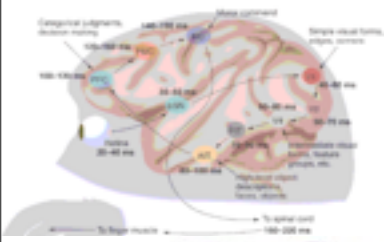


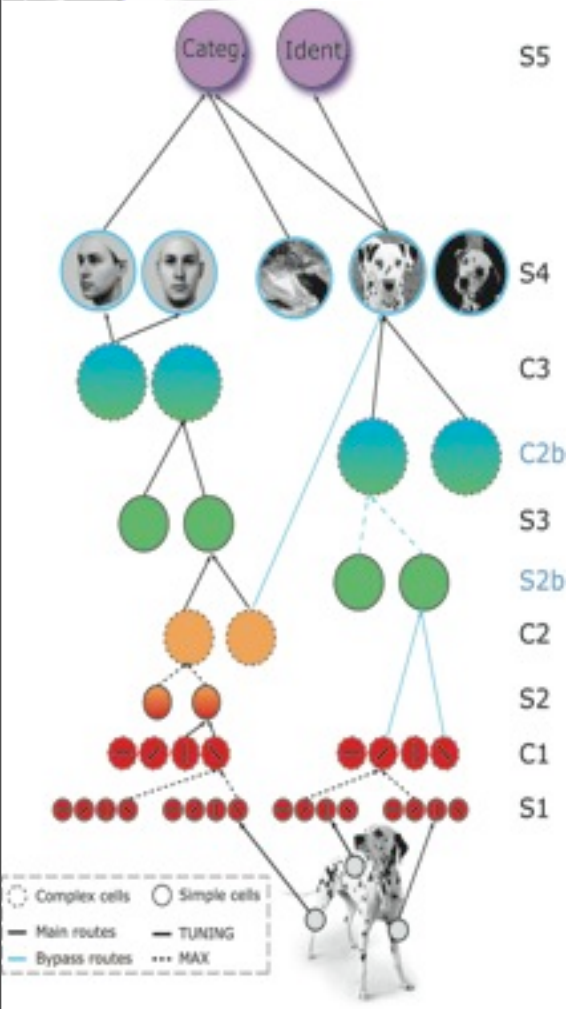


Models of the *ventral stream* in cortex
perform well compared to
engineered computer vision systems (in 2006)
on several databases

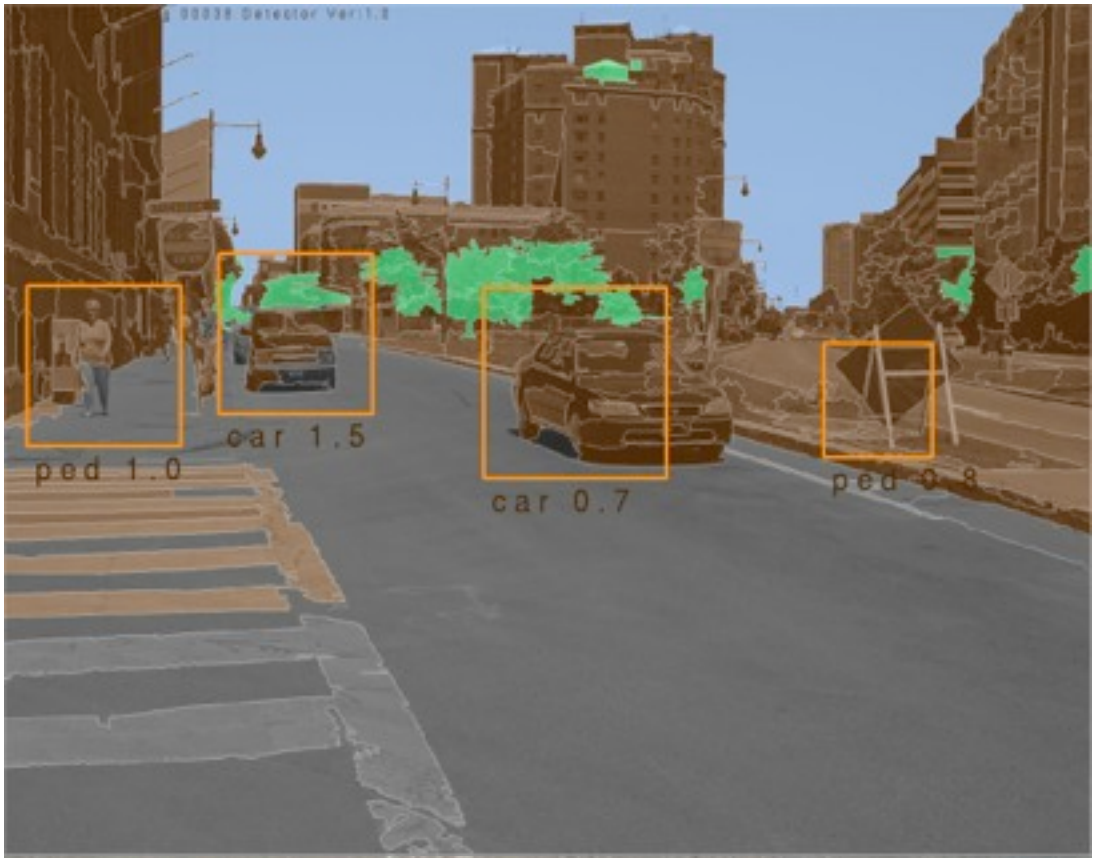
Bileschi, Wolf, Serre, Poggio, 2007

Neuroscience to good computer vision

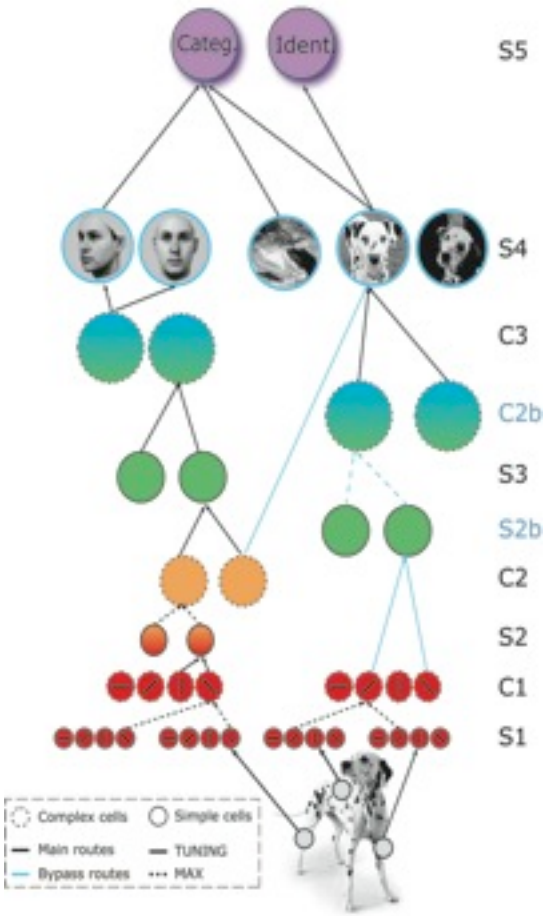




Models of the *ventral stream* in cortex
perform well compared to
engineered computer vision systems (in 2006)
on several databases



Bileschi, Wolf, Serre, Poggio, 2007


Categ.
Ident.
S5
S4
C3
C2b
S3
S2b
C2
S2
C1
S1
Complex cells
Simple cells
Main routes
TUNING
Bypass routes
MAX


Categ.
Ident.
Complex cells
Simple cells
Main routes
TUNING
Bypass routes
MAX

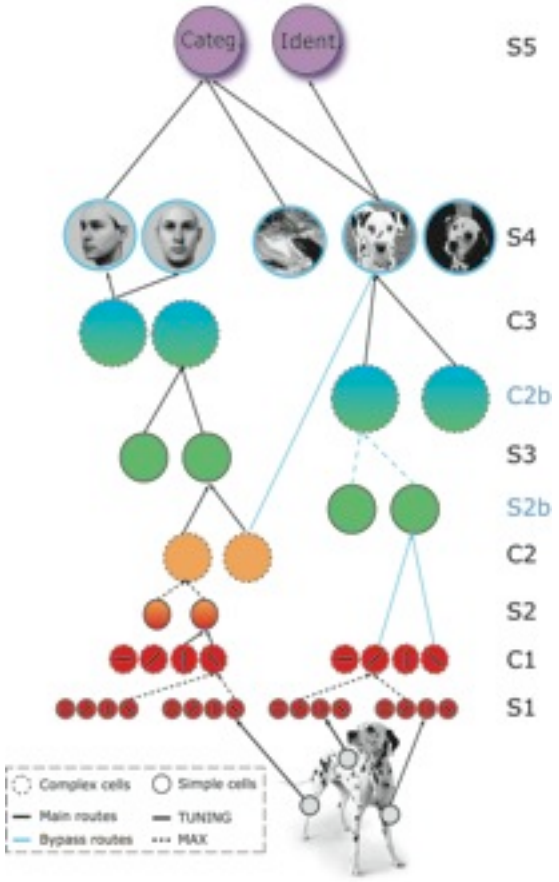

walk

Performance

human agreement	72%
proposed system	77%
commercial system	61%
chance	12%


walk

A puzzle



Hierarchical, HMAX-type models of visual cortex very well as *computer vision systems* but...why?

Very similar *convolutional networks* now called deep learning networks (LeCun, Hinton,...) are *unreasonably successful* in vision and speech (ImageNet+Timit)...

why?

We need theories!

Found Comput Math (2010) 10: 67–91
DOI 10.1007/s10208-009-9049-1

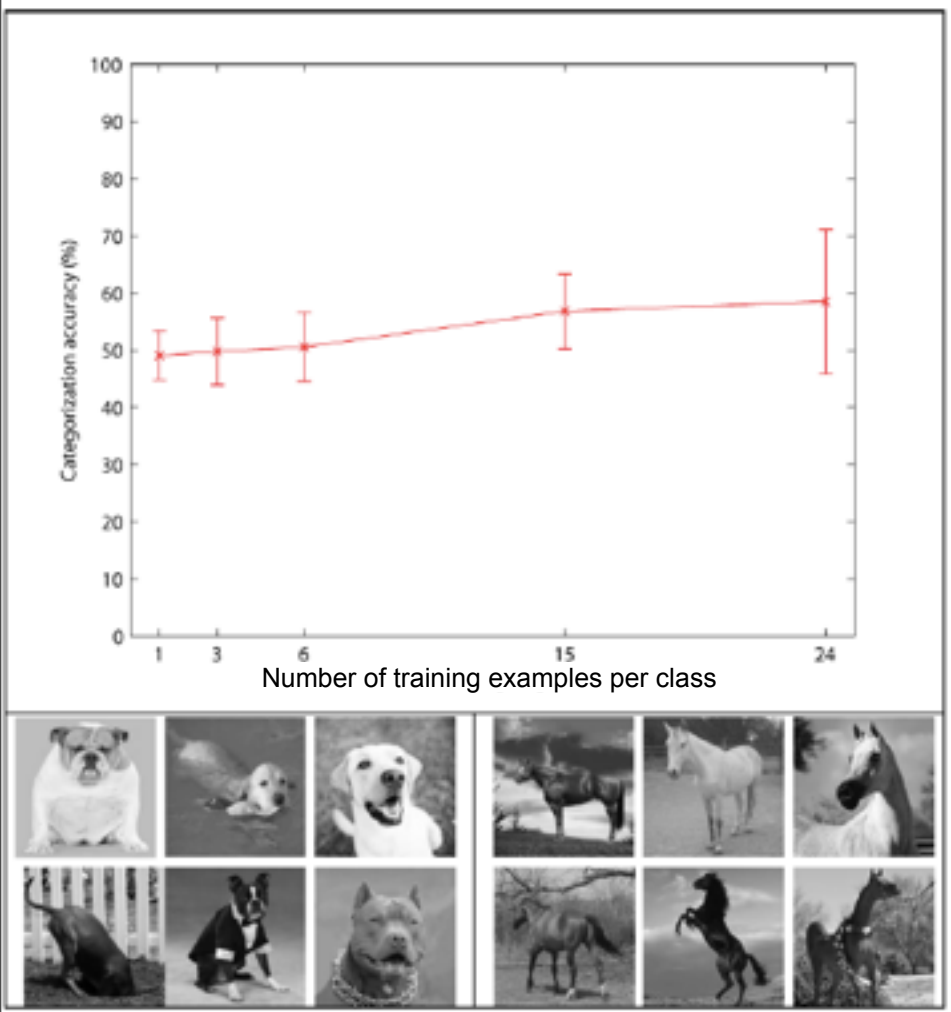
FOUNDATIONS OF COMPUTATIONAL MATHEMATICS

Mathematics of the Neural Response

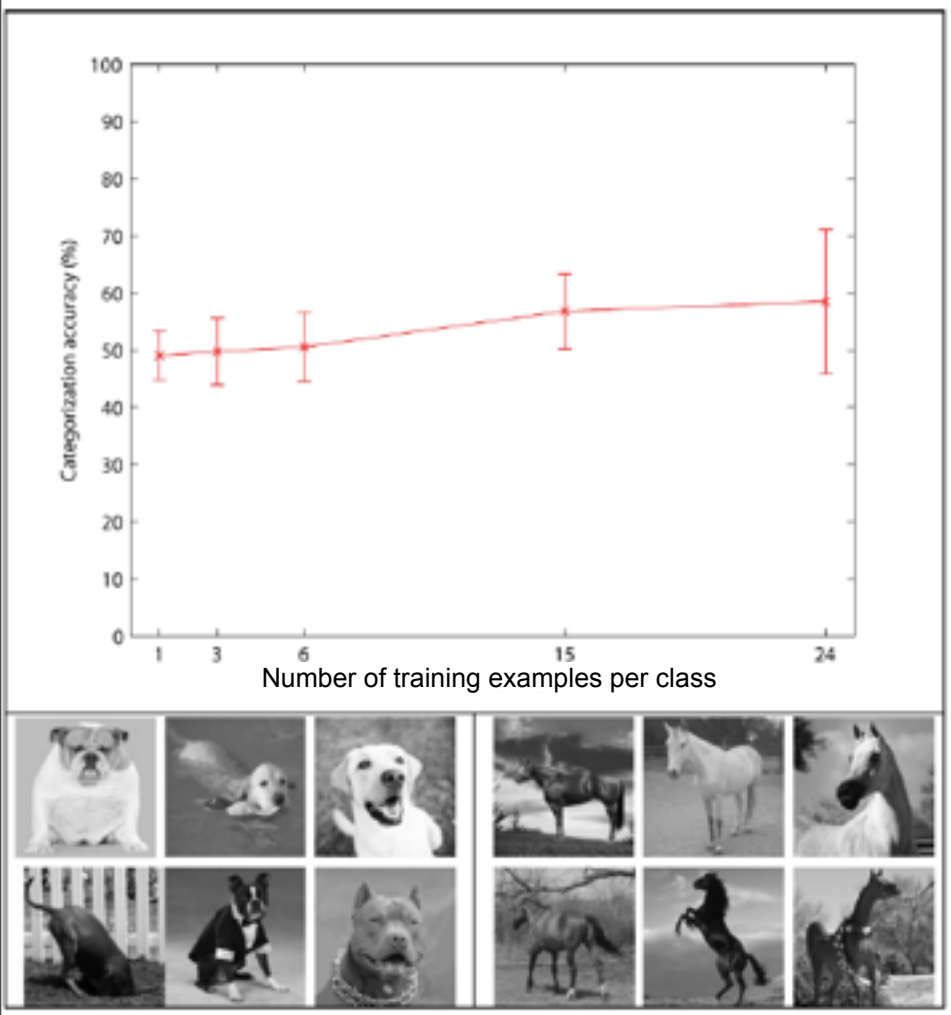
S. Smale · L. Rosasco · J. Bouvrie · A. Caponnetto · T. Poggio

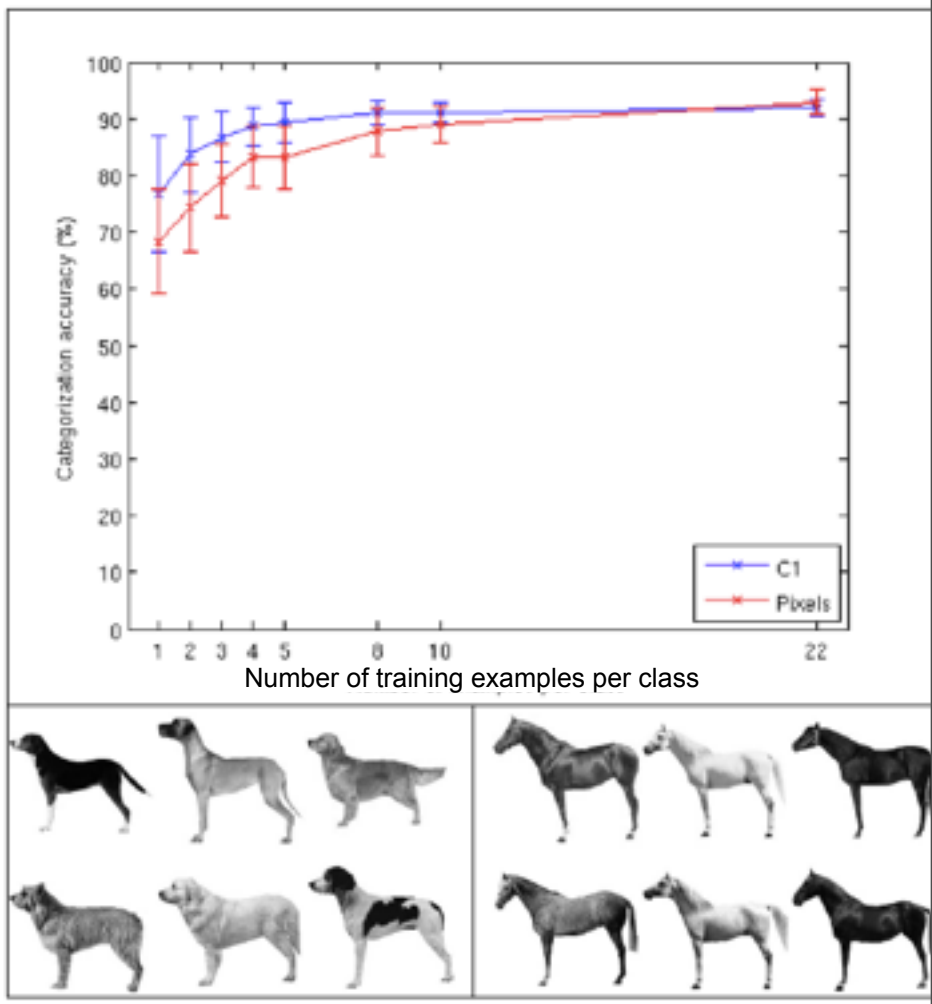
Computing invariant representations for perception:
is this the computational goal of the ventral stream?
the magic of sensory cortex?

Computing invariant representations for perception: is this the computational goal of the ventral stream? the magic of sensory cortex?

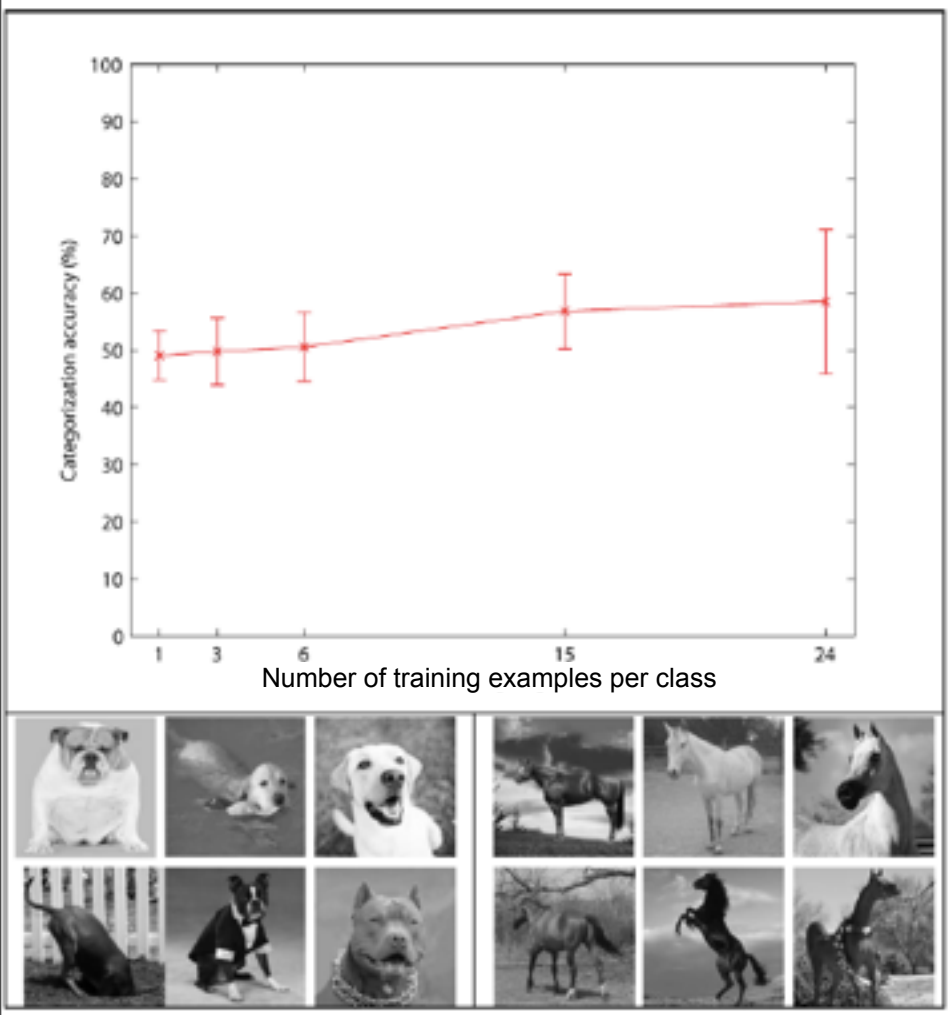


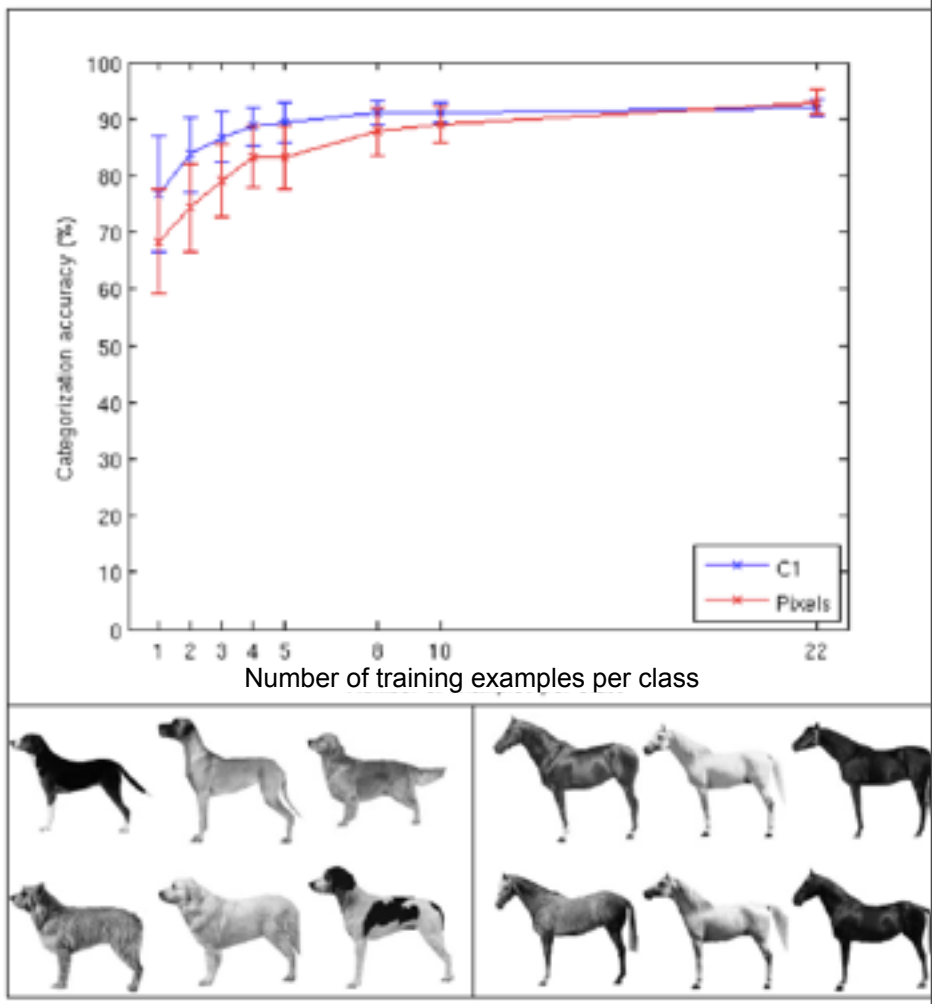
Computing invariant representations for perception: is this the computational goal of the ventral stream? the magic of sensory cortex?



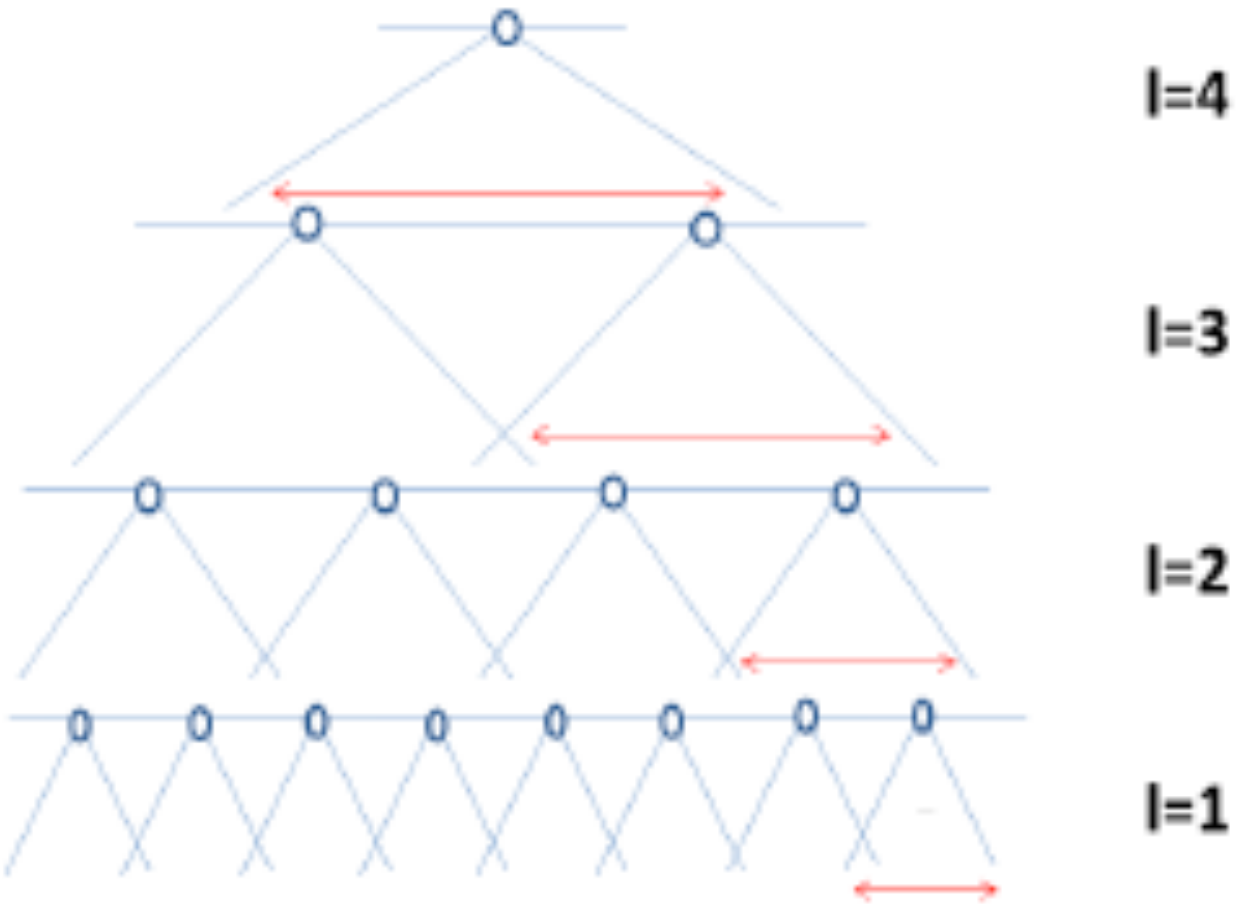


Computing invariant representations for perception: is this the computational goal of the ventral stream? the magic of sensory cortex?



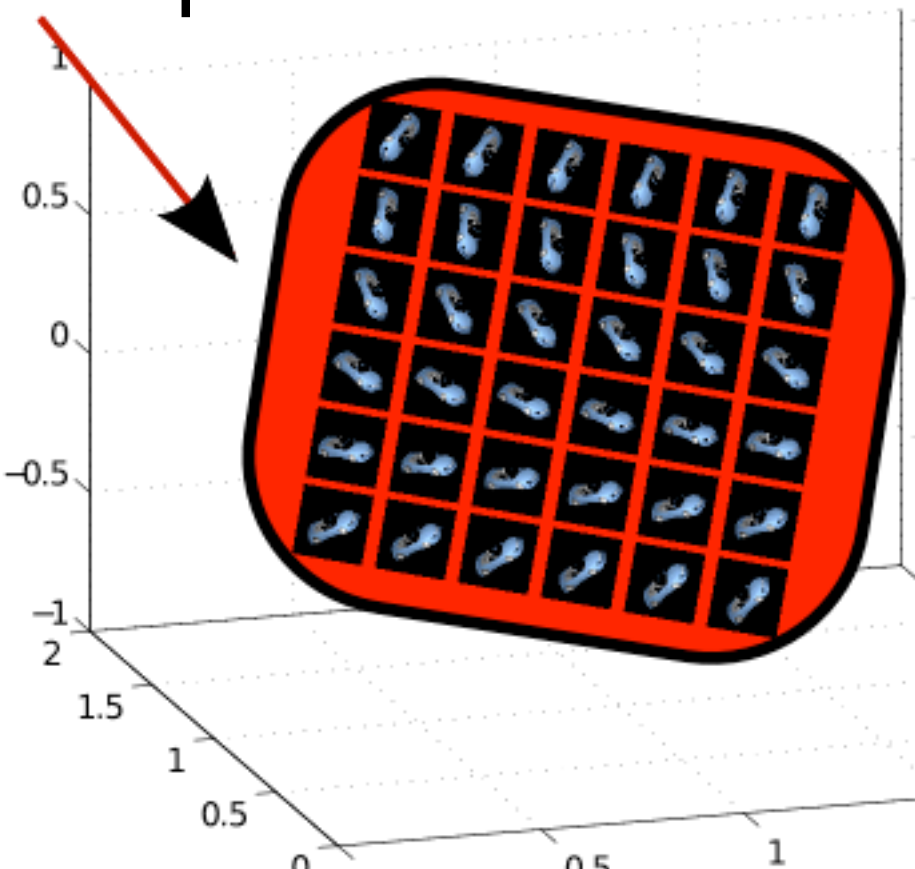


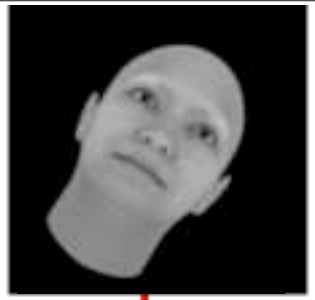
Multilayer architectures



Basic Idea

All rotations gt^k of a template car





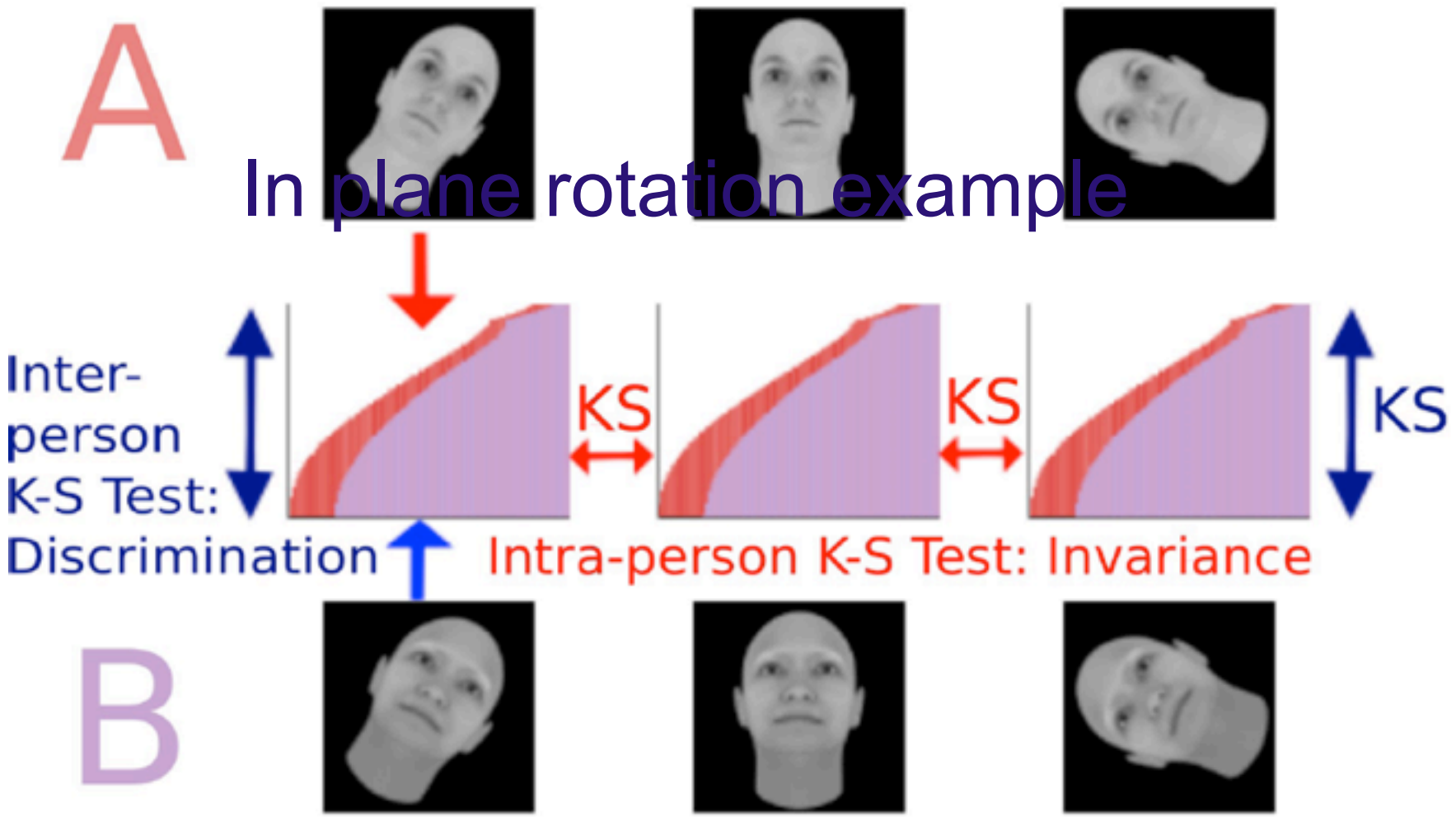
samples

CDFs are unique and invariant signatures.

CDFs

K-S distance

In plane rotation example



A second phase in Machine Learning

- The first phase -- from ~1980s -- led to a rather complete theory of *supervised learning* and to practical systems (MobilEye, Orcam,...) that need *lots of examples for training*

- The second phase may be about *unsupervised learning of (invariant) representations* that make supervised learning possible with *very few examples*