Pattern Recognition
Introduction

Francesco Tortorella

University of Cassino and
Southern Latium
Cassino, Italy

The Pattern Recognition and Image Analysis Group @ UniCas

Biomedical Image Analysis

- Retinopathy
- Breast Ultra Sound
- Breast Thomosynthesis
- Digital mammography

Biomedical Image Analysis

Pattern Recognition

Image Processing

C++

OpenCV

Learning-based Decision Systems

- Deep Learning
- Designing new features for Pattern Recognition
- Object detection with unbalanced datasets
- AUC/pAUC-based Classifiers
- Cascade-based Classifiers
- ECOC Decision Systems
- ROC-based Classifiers

Organization of the course

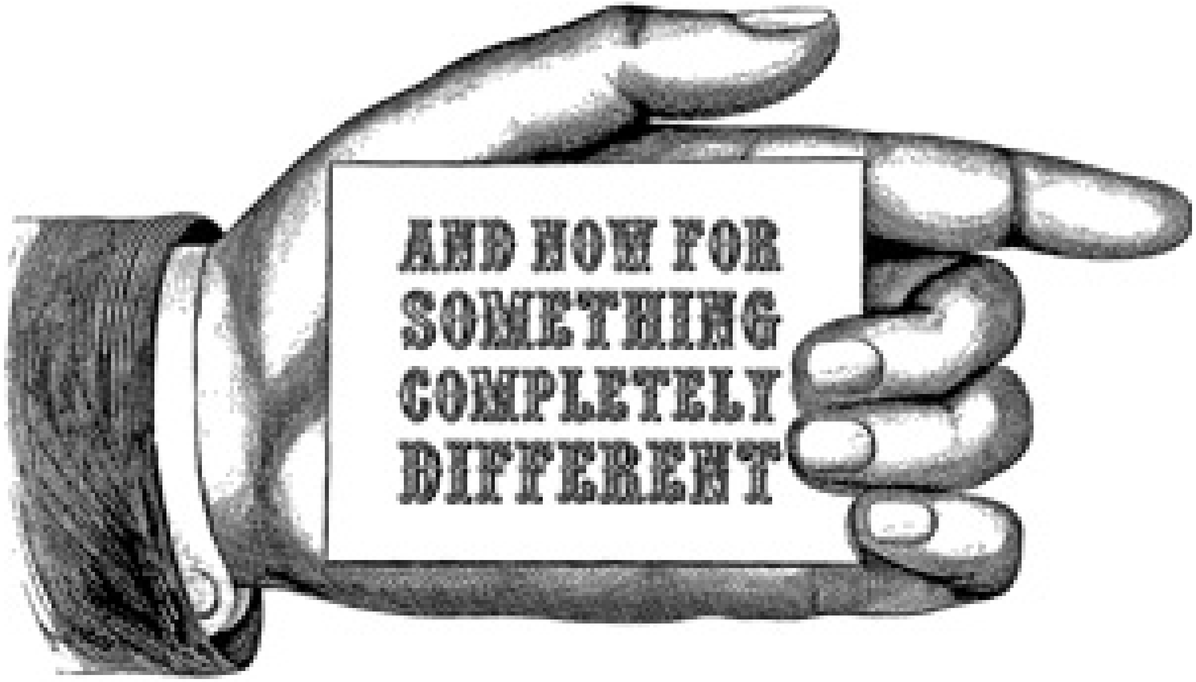
- Topics:
  - Bayes decision theory.
  - Parametric and nonparametric classification.
  - Feature selection and extraction.
  - Kernel-based classifiers
  - Deep learning paradigm.
  - Error assessment.

Texts

- A. Webb, *Statistical Pattern Recognition*, 3rd edition, John Wiley & Sons, Inc., 2011.
- C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2000.
- I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016.

Organization of the course

- Grading:
  - 3 homeworks (40%)
  - Term project (public presentation) (40%)
  - Final exam (20%)


AND NOW FOR
SOMETHING
COMPLETELY
DIFFERENT

Recognizing objects

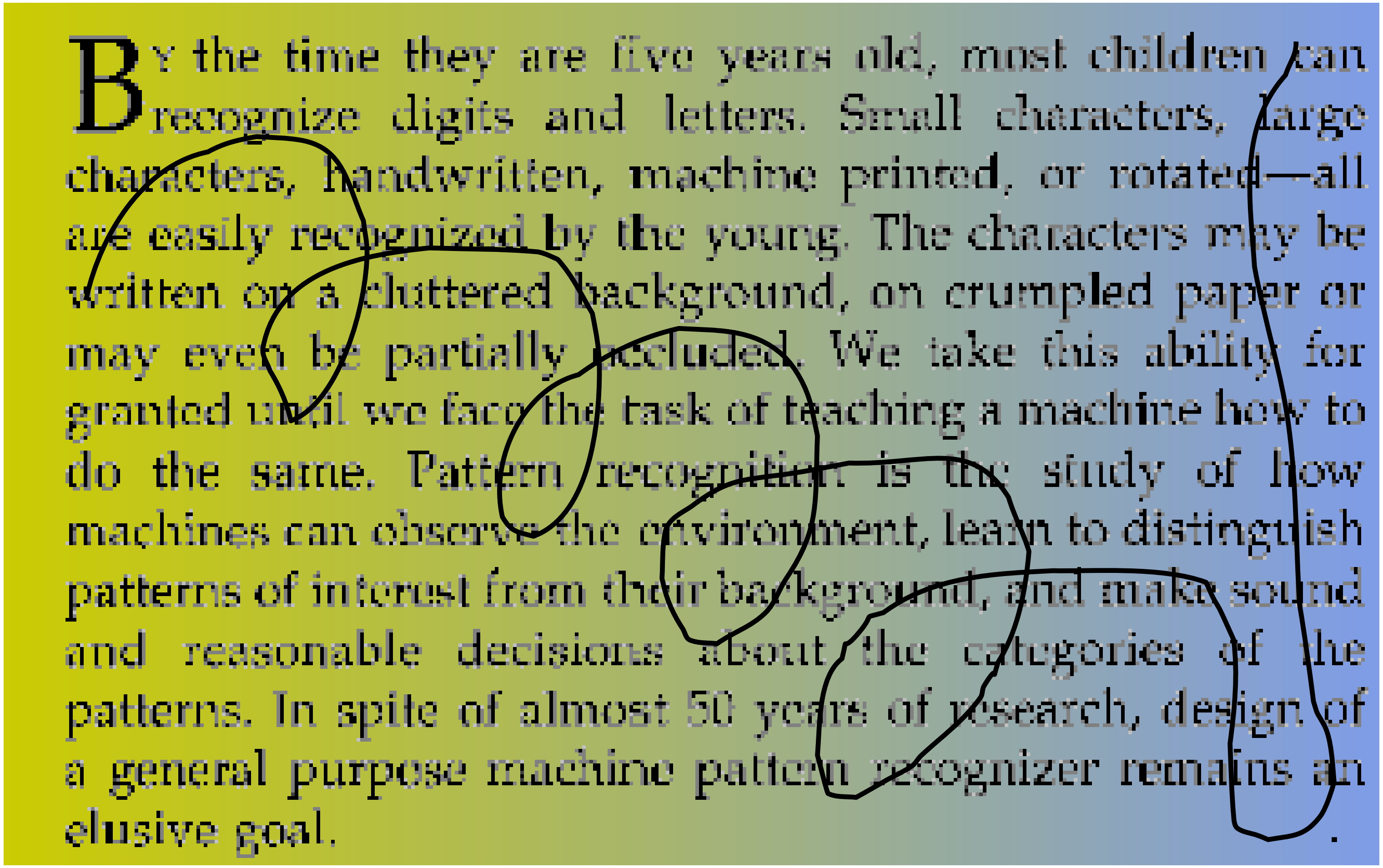
- How to build a system that automatically recognize a certain kind of objects in a scene?
- Human beings are extremely able in this task (thanks, Mr. Darwin !)
- But it is extremely difficult to transfer this ability to an artificial system.

Recognizing objects

- How to build a system that automatically recognize a certain kind of objects in a scene?
- Human beings are extremely able in this task (thanks, Mr. Darwin !)

Recognizing objects

- How to build a system that automatically recognize a certain kind of objects in a scene?
- Human beings are extremely able in this task (thanks, Mr. Darwin !)
- But it is extremely difficult to transfer this ability to an artificial system.

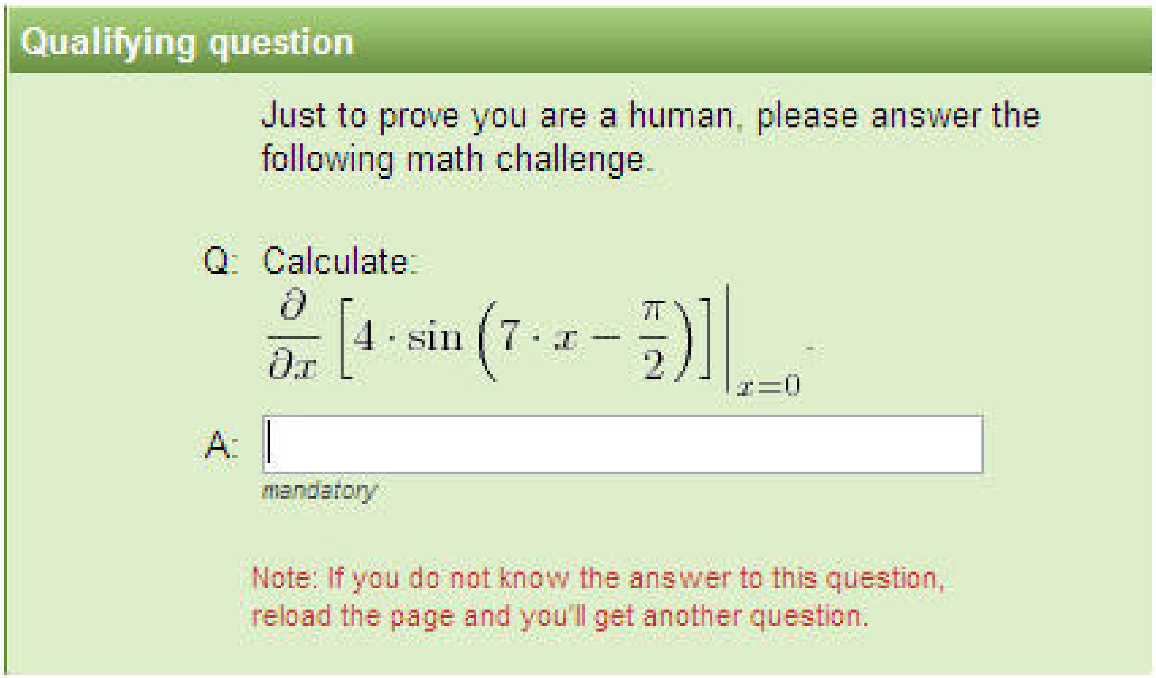


By the time they are five years old, most children can recognize digits and letters. Small characters, large characters, handwritten, machine printed, or rotated—all are easily recognized by the young. The characters may be written on a cluttered background, on crumpled paper or may even be partially occluded. We take this ability for granted until we face the task of teaching a machine how to do the same. Pattern recognition is the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns. In spite of almost 50 years of research, design of a general purpose machine pattern recognizer remains an elusive goal.

A.K. Jain, R.P. Duin, J. Mao
"Statistical Pattern Recognition: a Review"

Used in CAPTCHA to distinguish between human and machine

Completely
Automated
Public
Turing test to tell
Computers and
Humans
Apart


Qualifying question
Just to prove you are a human, please answer the following math challenge.
Q: Calculate:
$\frac{\partial}{\partial x}\left[4 \cdot \sin\left(7 \cdot x - \frac{\pi}{2}\right)\right]\Big|_{x=0}$.
A:
mandatory
Note: If you do not know the answer to this question, reload the page and you'll get another question.

Recognition: activity of the brain

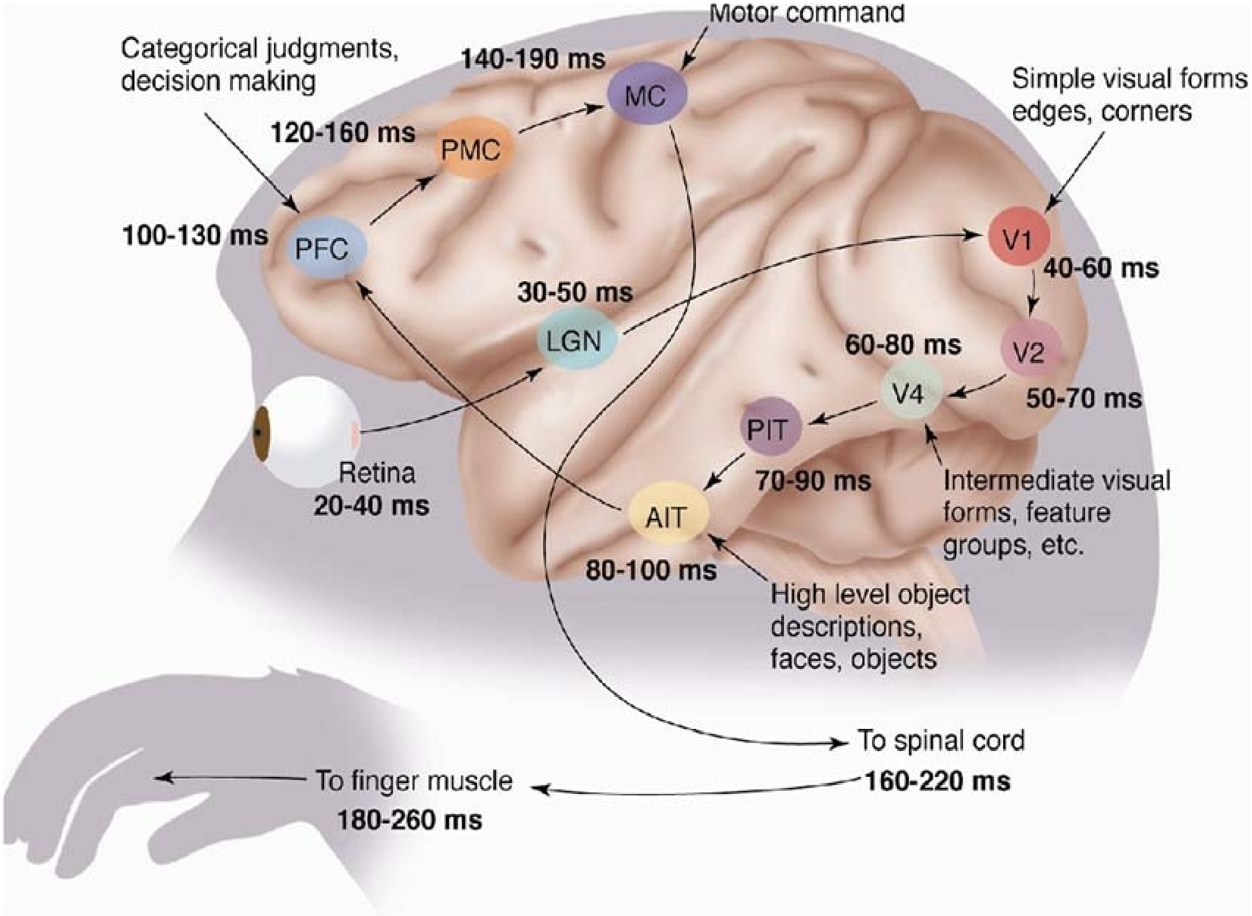
Every attempt of replying the human ability for recognition must take into account the complexity of the human brain.

Recognition: activity of the brain

Every attempt of replying the human ability for recognition must take into account the complexity of the human brain.

The brain –
that's my second most favourite organ!
Woody Allen

What happens under the hood?



Recognition: activity of the brain

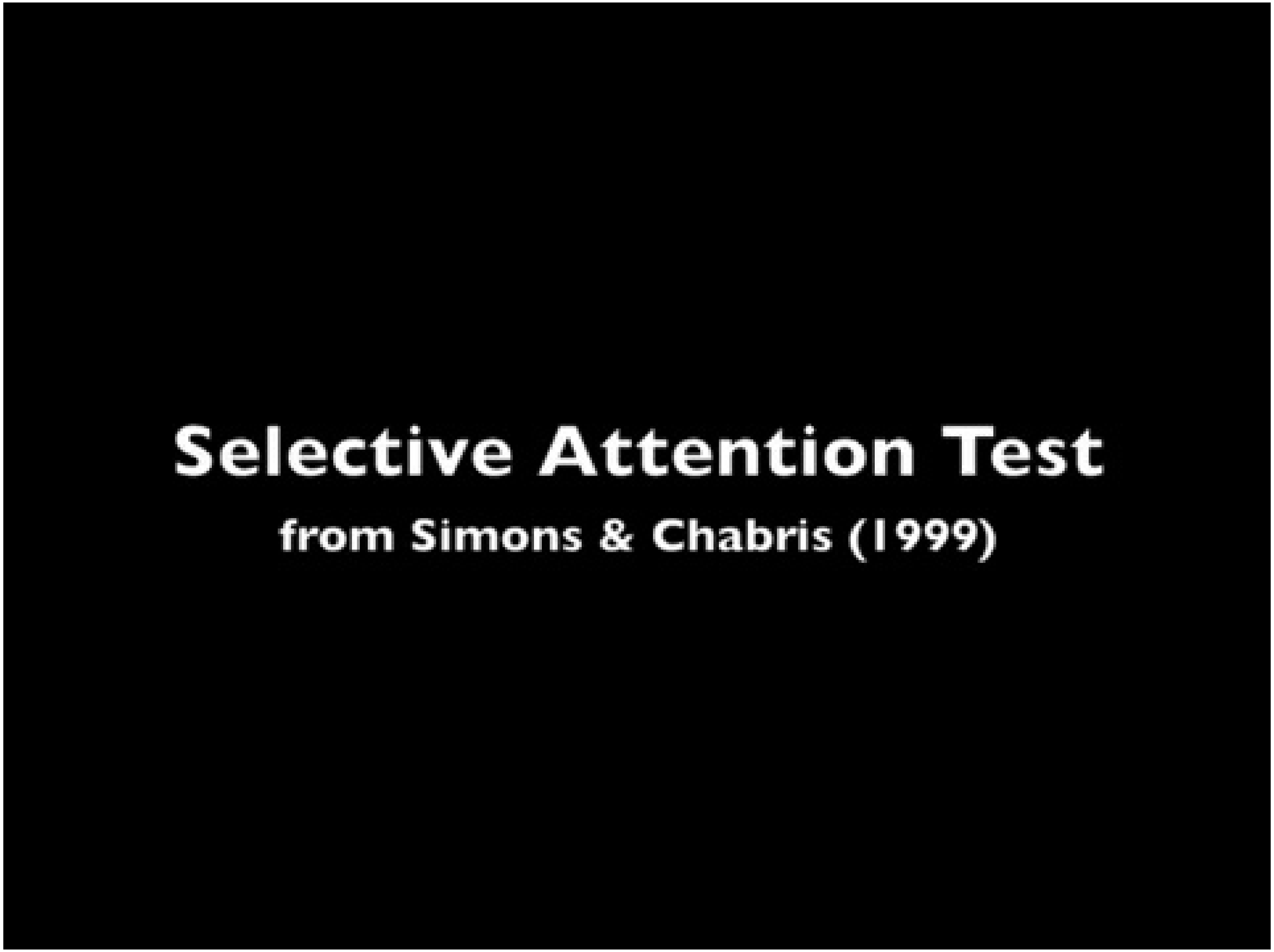
Every attempt of replying the human ability for recognition must take into account the complexity of the human brain.

But do we actually know how it works?

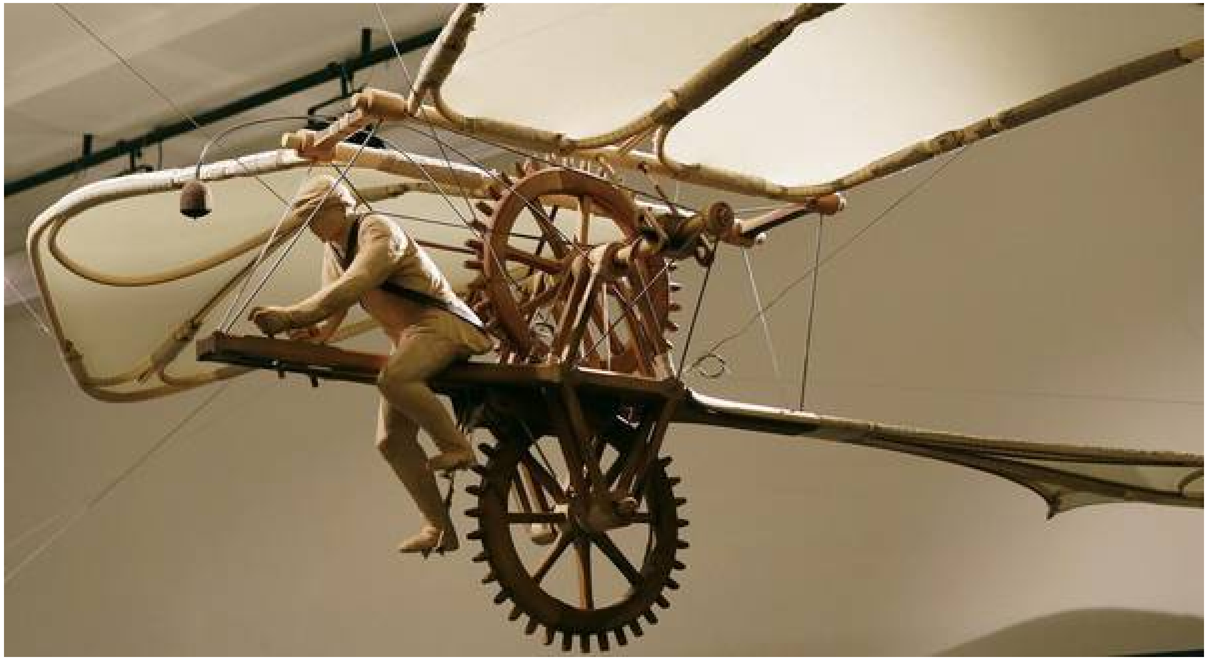
Igor, Whose Brain I did put in?

Mel Brooks

Video


Selective Attention Test
from Simons & Chabris (1999)

Nice idea to imitate the Nature, but probably we could miss some details ...



What is Pattern Recognition?

- The word *pattern* is derived from the same root as the word *patron* and, in its original use, means something which is set up as a perfect example to be imitated.
- Thus Pattern Recognition means the identification of the ideal which a given object was made after.
  (Theo Pavlidis)

What is Pattern Recognition?

- The field of pattern recognition is concerned with *the automatic discovery of regularities in data through the use of computer algorithms* and with *the use of these regularities to take actions such as classifying the data into different categories*
  Chris Bishop

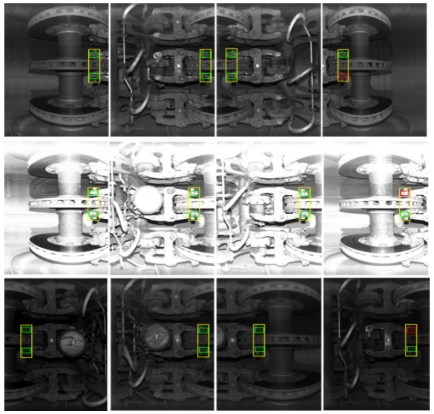
What is Pattern Recognition?

- "The *assignment* of a physical object or event to one of *several pre-specified categories*" (Duda and Hart)
- "A problem of *estimating density functions* in a *high-dimensional space* and *dividing the space into the regions of categories or classes*" (Fukunaga)
- "*Given some examples* of complex signals and the correct decisions for them, *make decisions automatically* for a stream of future examples" (Ripley)
- "The science that concerns the *description or classification (recognition)* of measurements" (Schalkoff)
- "The process of *giving names* ω to *observations* x" (Schürmann)
- Pattern Recognition is concerned with answering the question "**What is this**?" (Morse)

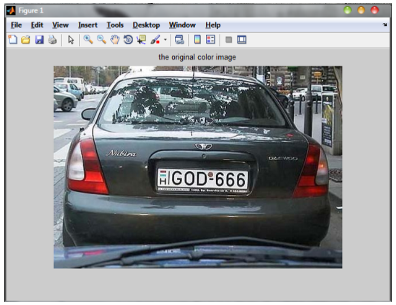
What is Pattern Recognition?

- A summary view:
  Given the *description* of an object that can belong to one out of N possible classes, the system should assign the object to a class on the basis of a *knowledge base previously built*

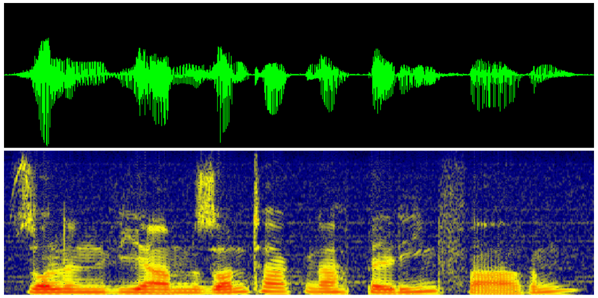
Applications



- Machine vision
  - Visual inspection
- Character recognition
  - Automated mail sorting, processing bank checks
  - Written text is scanned and the image is converted into constituent characters



- Speech recognition
  - Human Computer Interaction
  - Microphone records acoustic signal
  - Speech signal is classified into phonemes and/or words



Applications

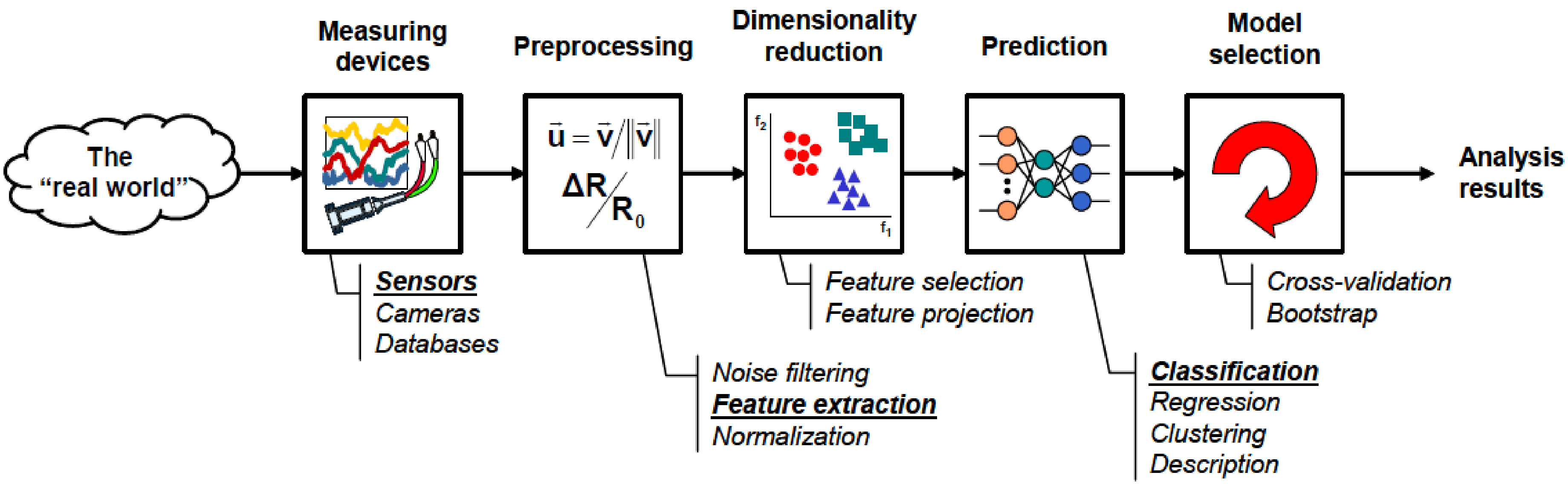
- Computer aided diagnosis
  - Medical imaging, EEG, ECG signal analysis
  - Designed to assist (not replace) physicians
  - Example: X-ray mammography

Components of a pattern recognition system

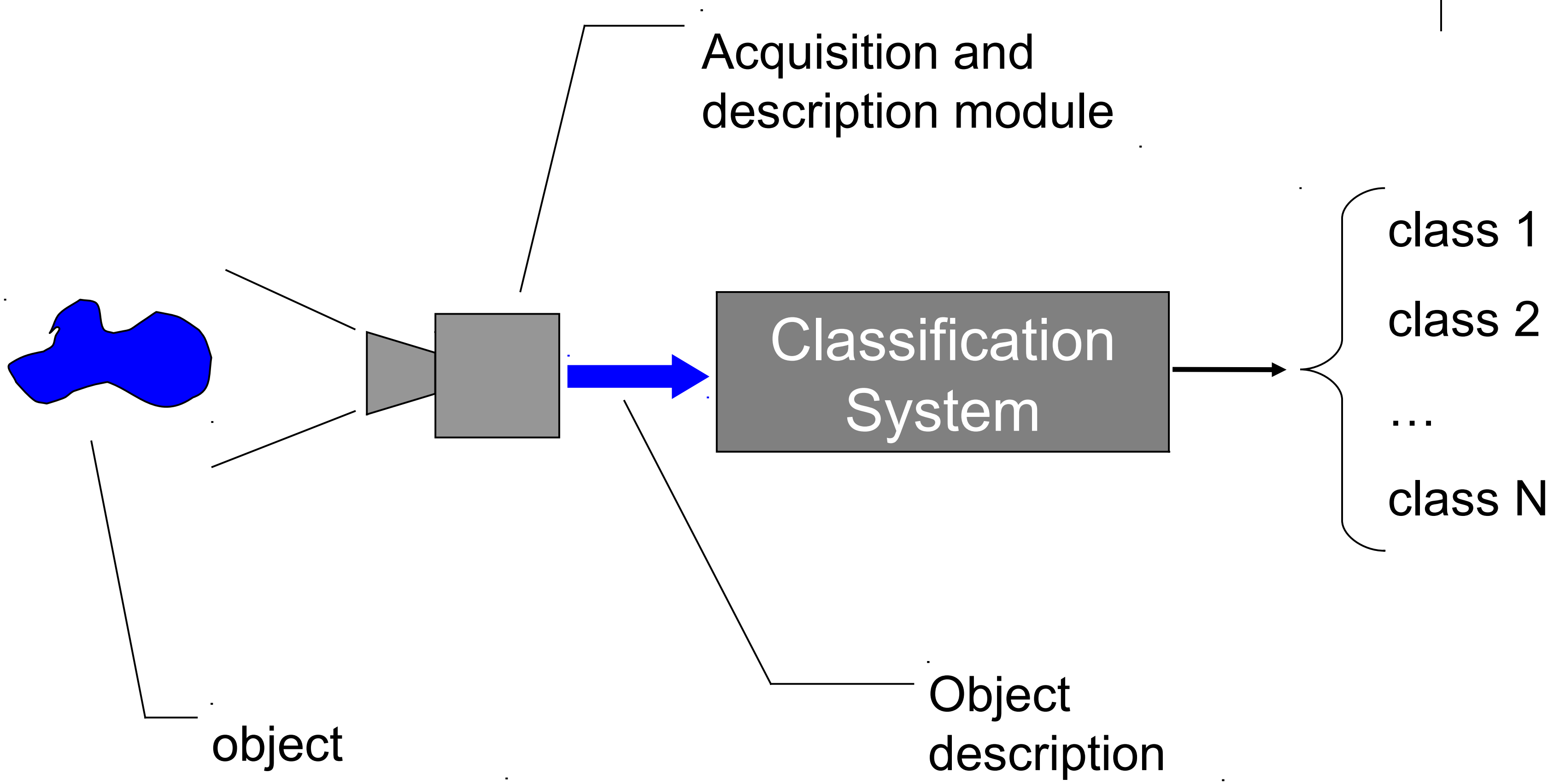
An ideal model of a recognition system contains:

- a sensor
- a preprocessing mechanism
- a feature extraction mechanism (manual or automated)
- a classification algorithm

A complete recognition system

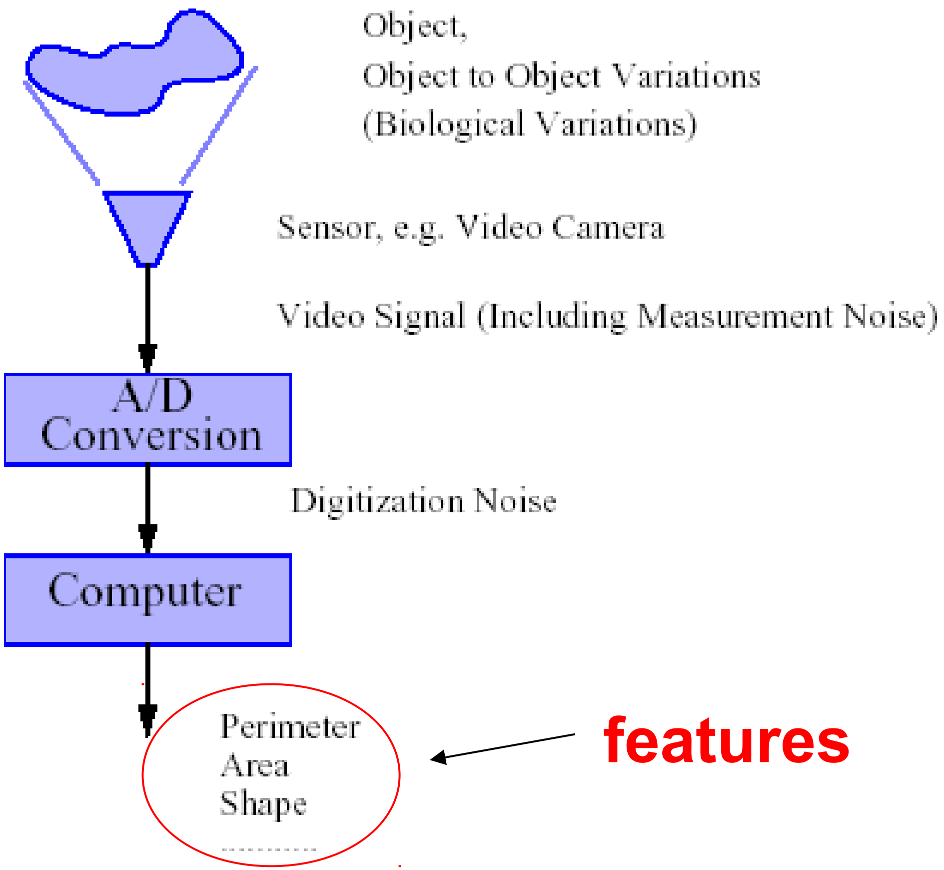


The classification system

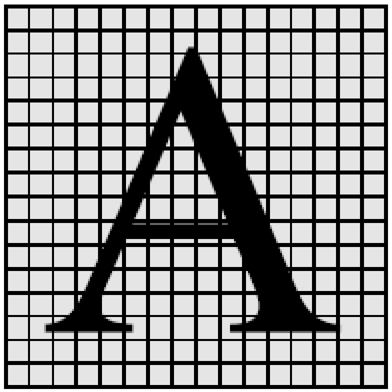


Description

- The classification system receives as input a **description**, i.e. a set of measures characterizing the object (**features**).
- The features are chosen on the basis of the specific requirements

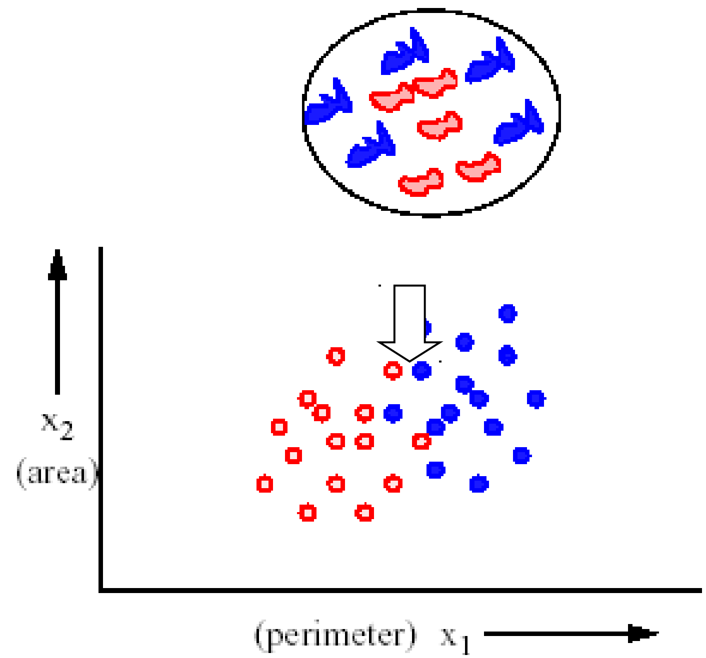


Which features?



- 16x16=256 pixels → 256 features/sample
- Area, perimeter → 2 features/sample

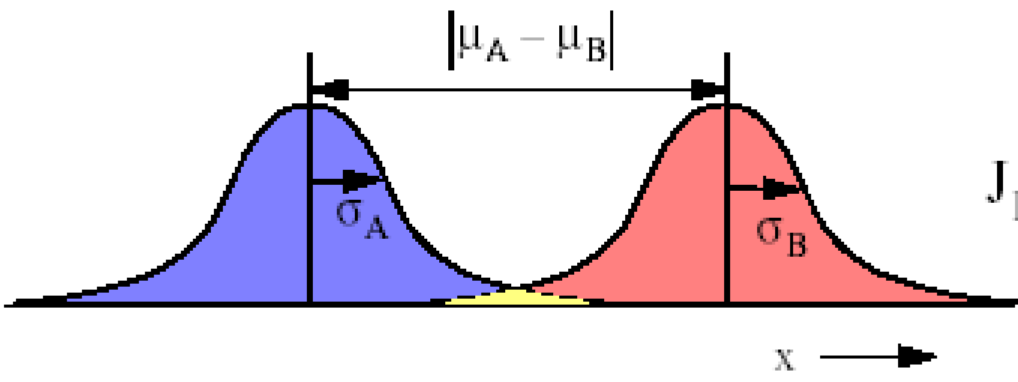
Each object corresponds to a *feature vector* which can be seen as a point in the *feature space*



Which features?

- Features must be discriminant
- Which criteria to be used?
  - heuristic
  - statistic

e.g. Fisher Criterion:



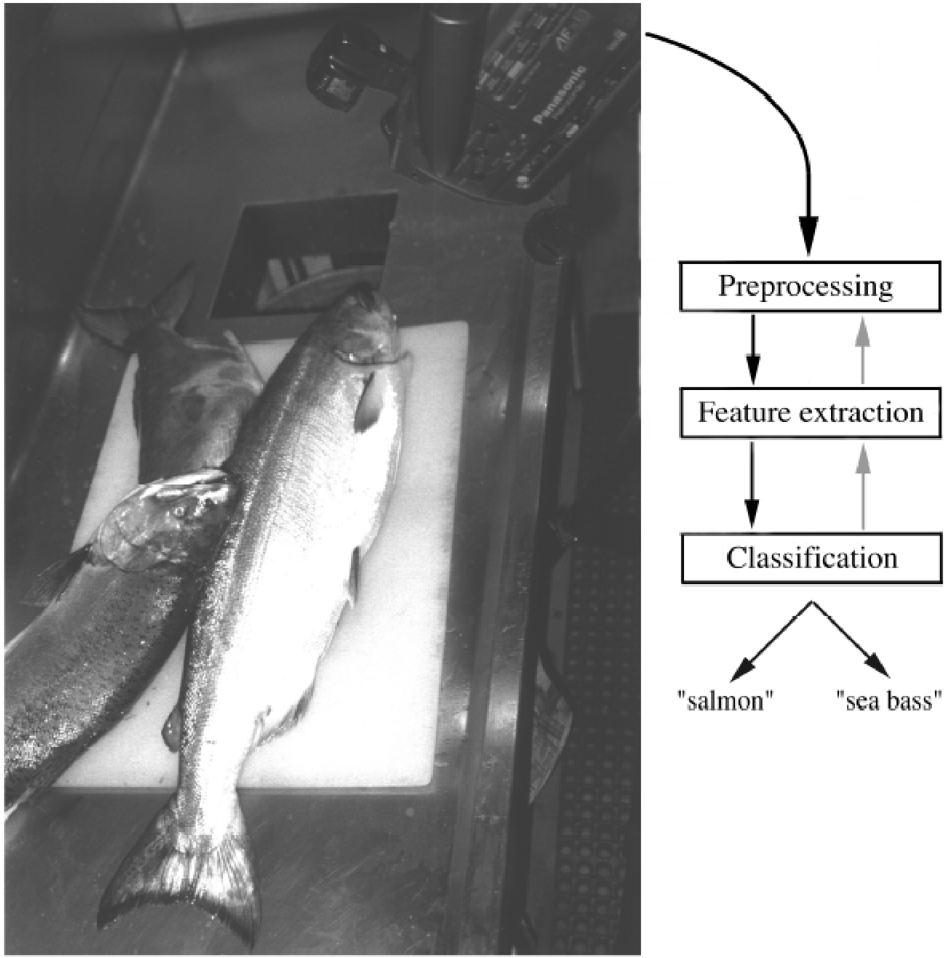
$$J_F(x) = \frac{|\mu_A - \mu_B|^2}{\sigma_A^2 + \sigma_B^2}$$

A fishing example
(Duda, Hart, Stork)

Problem:

for fishes going on a conveyor belt

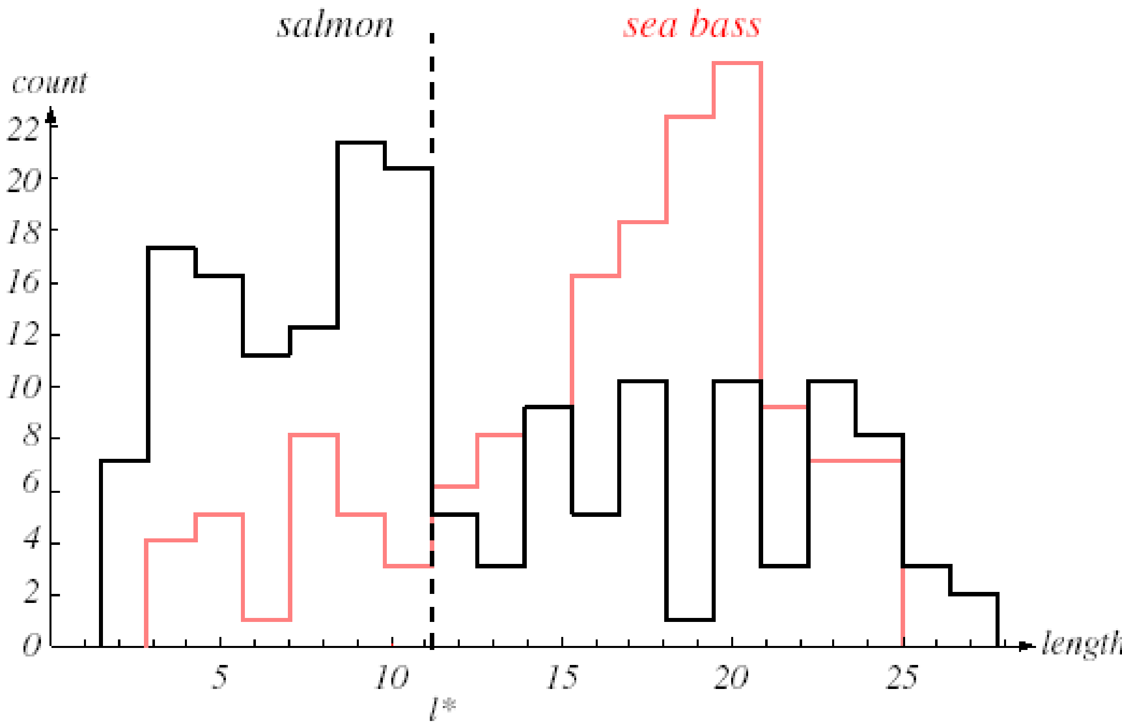
separate sea bass from salmon using optical sensing



Choosing the features

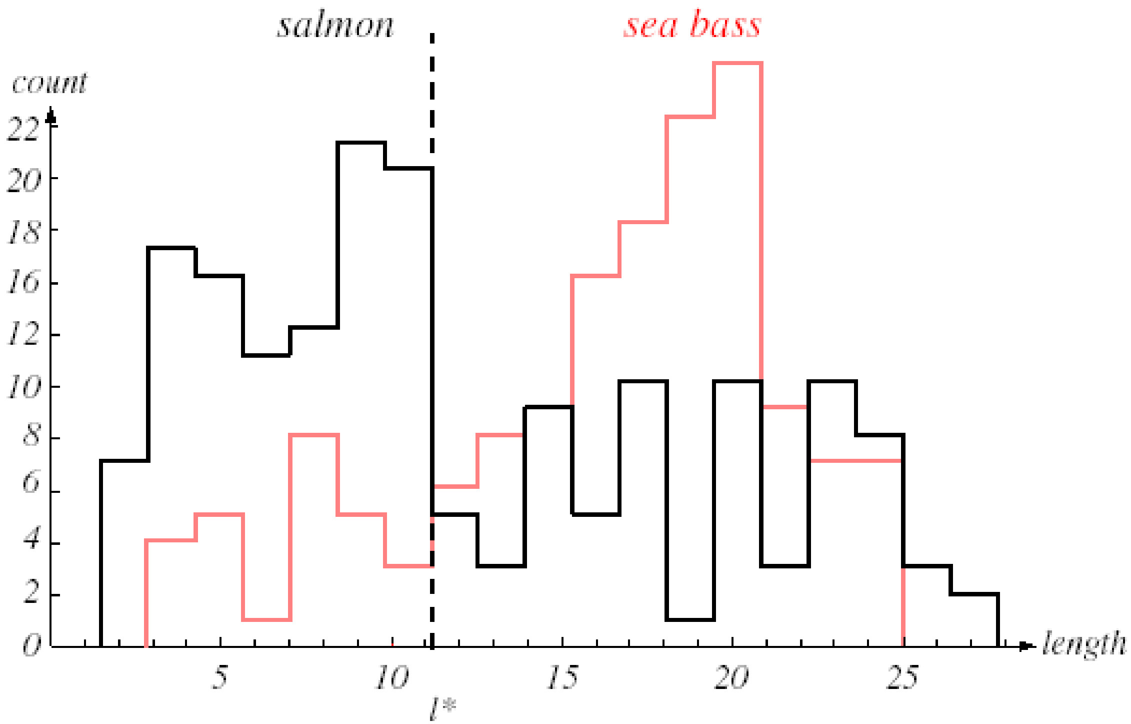
- We could initially choose the *length* of the fish.
- **Problem**: which *a priori* knowledge have we about the distribution of this feature in the two classes?
- We could measure such distribution on a set of samples of the two classes.

Choosing the features



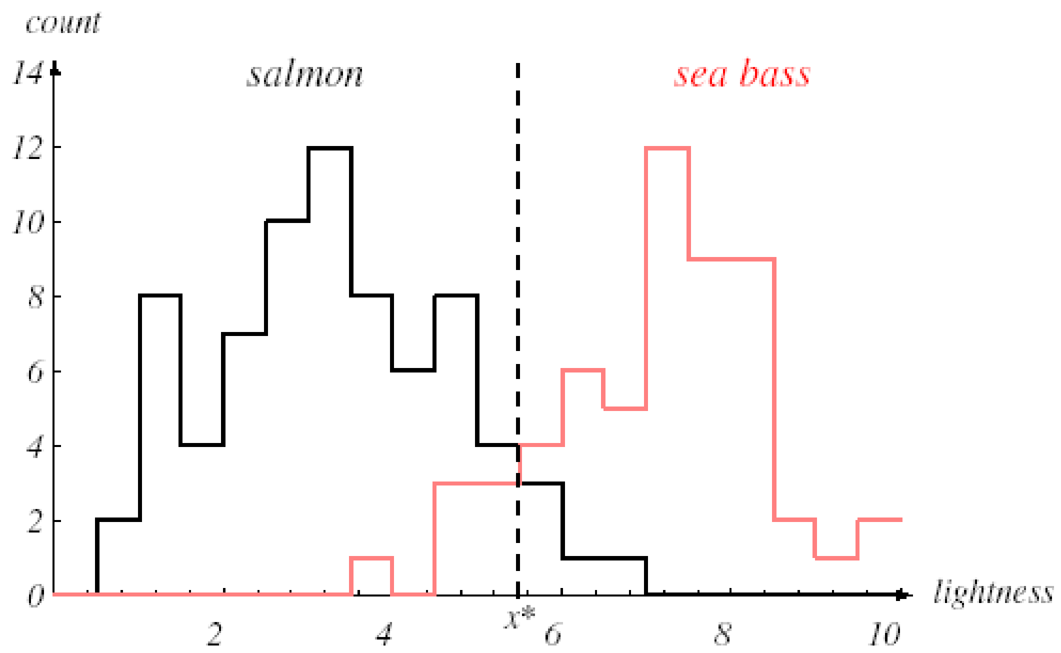
Not very discriminant feature. Not possible to perfectly separate the two categories.

Choosing the features



Not very discriminant feature. Not possible to perfectly separate the two categories.

Choosing the features

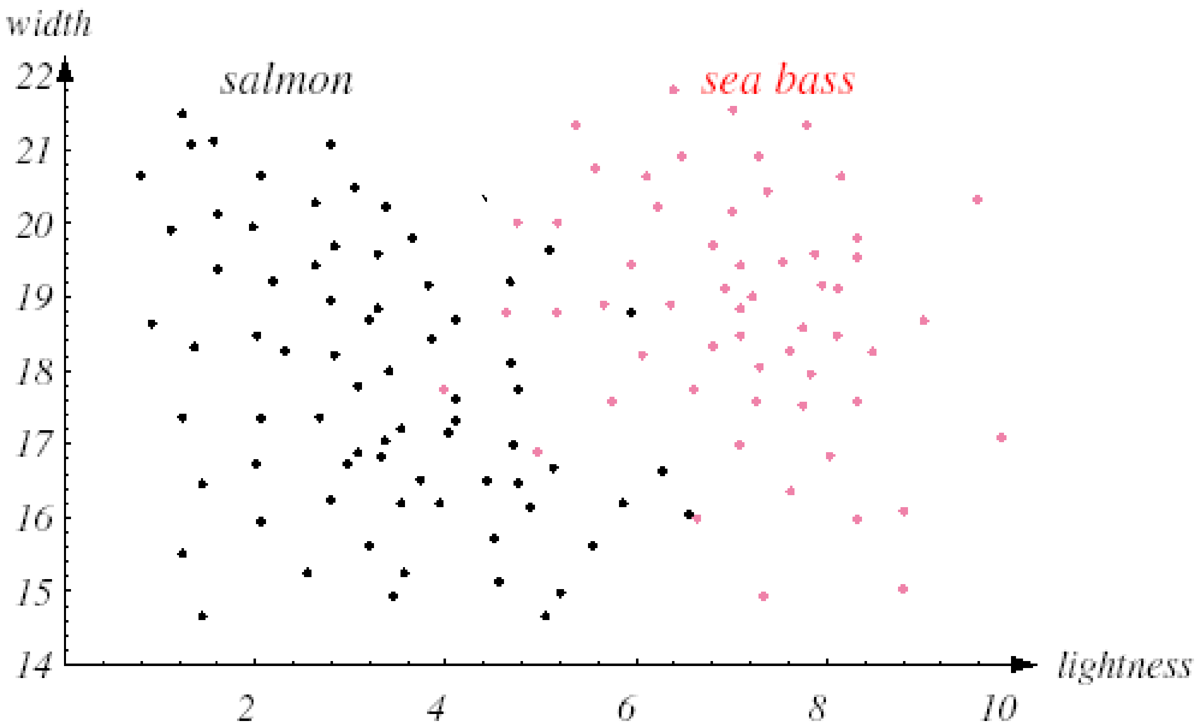


Better feature, but still not perfect. The threshold x* leads to the smallest number of errors.

Choosing the features

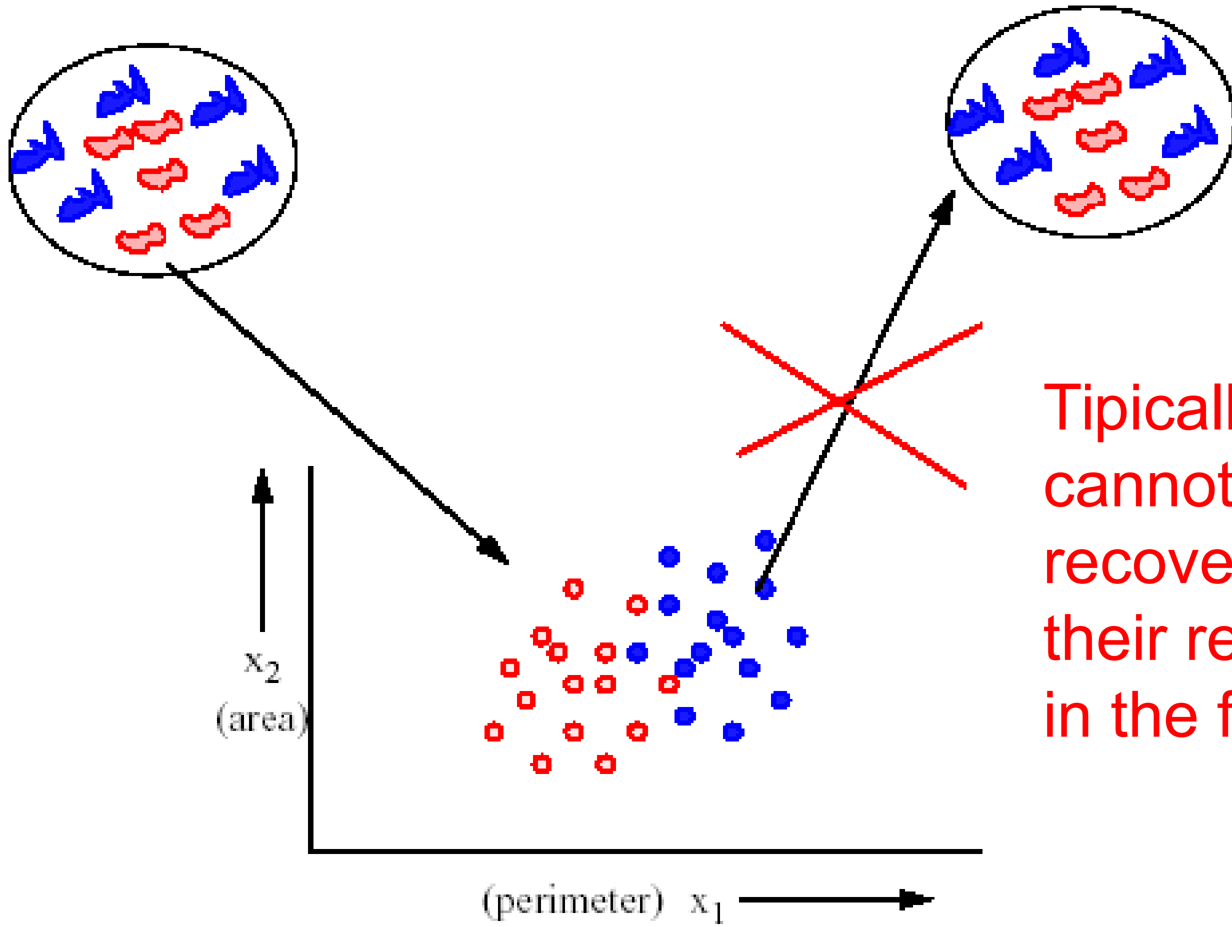
- Neither feature provides a satisfying result.
- What if we use both of them?
- In this case, each sample is described not by only one value, but by a set of values.
- From a formal point of view, each sample is described by a *feature vector* with N components and can be represented by a point in a N-dimensional space called *feature space*.

Feature space



In this case the feature space is a plane and each sample is described by a feature vector $[x_1\ x_2]^T$.

Feature space: characteristics



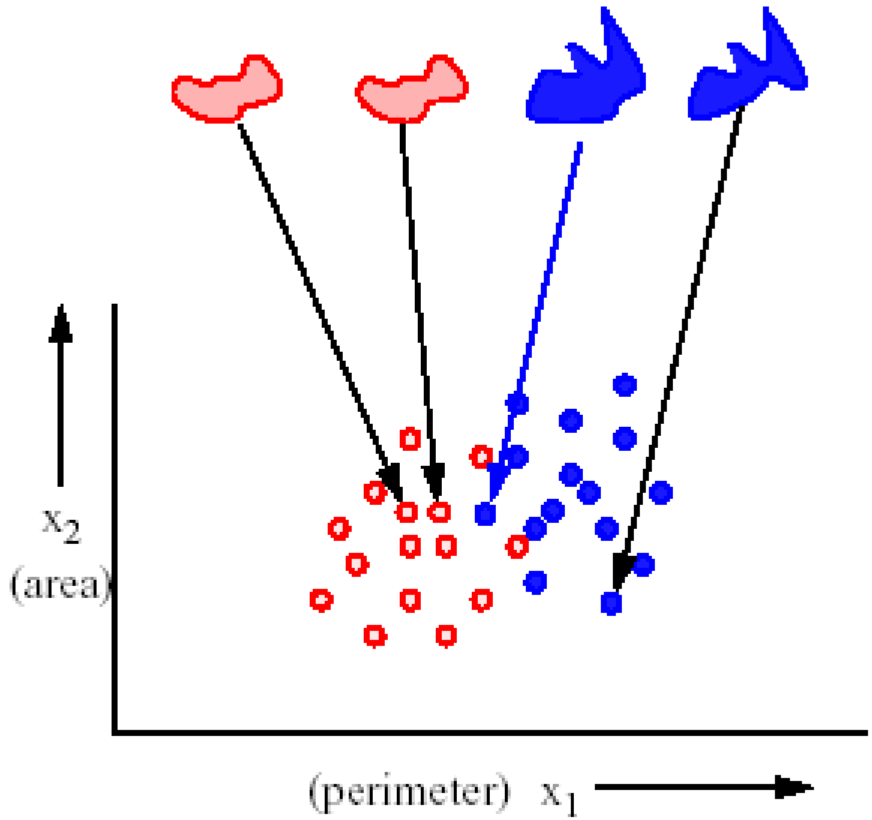
Tipically the objects cannot be recovered from their representation in the feature space

Similarity assumption

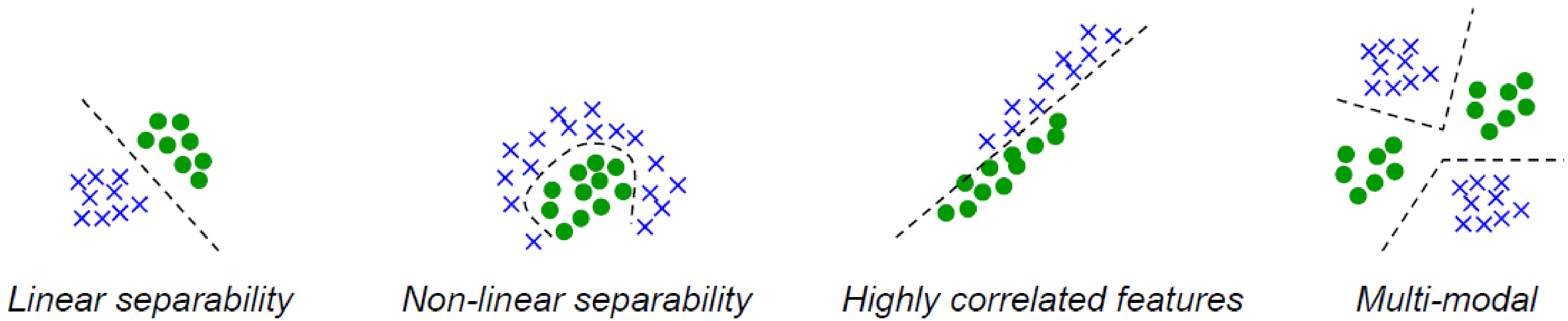
Similar objects are near each other in the feature space

but

different objects are not necessarily far from each other

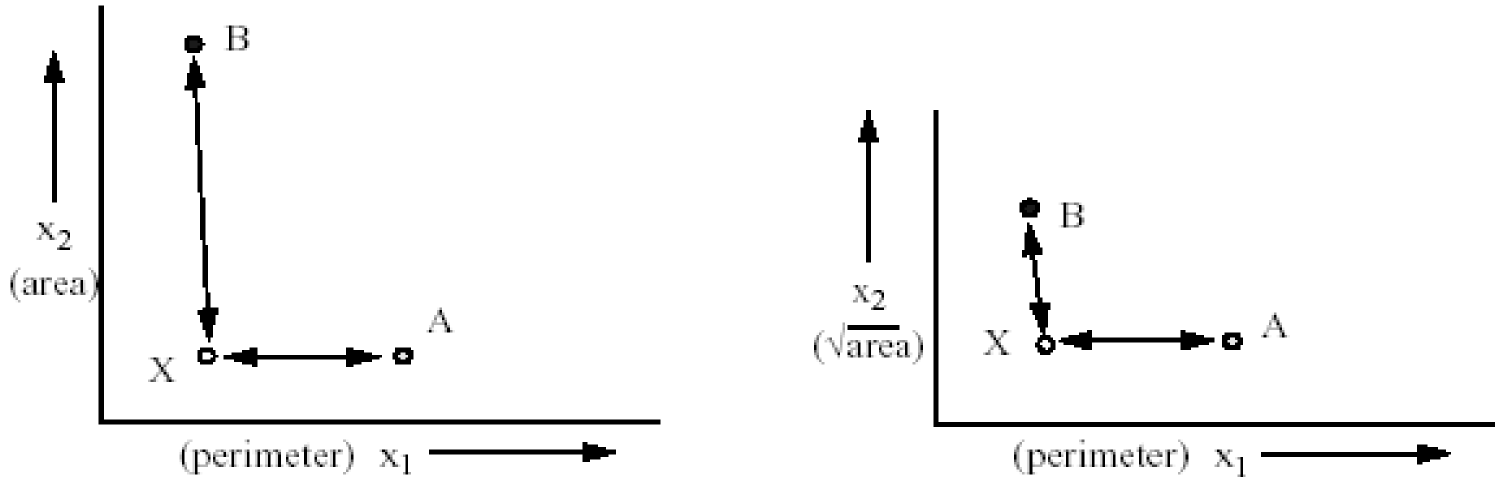


More feature properties



Not homogeneous features

A dimensional normalization could be needed



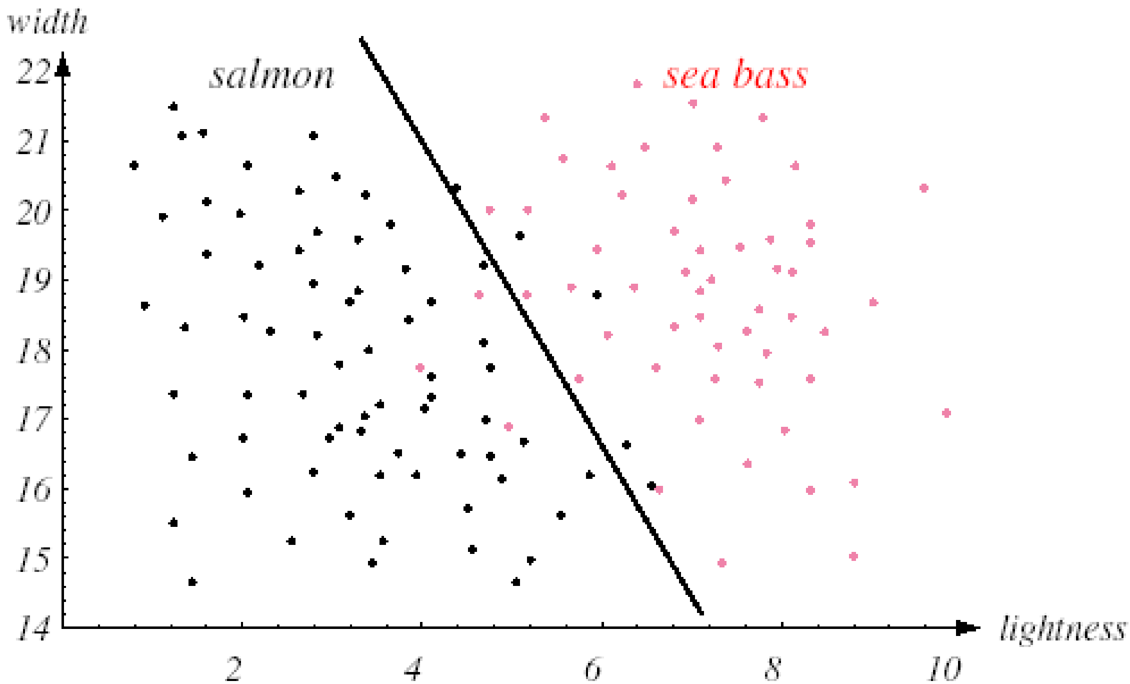
Before the normalization:
d(X,B) > d(X,A)

After the normalization:
d(X,B) < d(X,A)

Decision regions

- Defining a classification system implies dividing the feature space in regions, each of which belongs to one of the given classes.
- *Decision regions* separated by a *decision boundary.*
- In this way it is possible to classify a new sample on the basis of its position in the feature space.

Decision regions



The simplest choice: a linear boundary.

Better solutions?

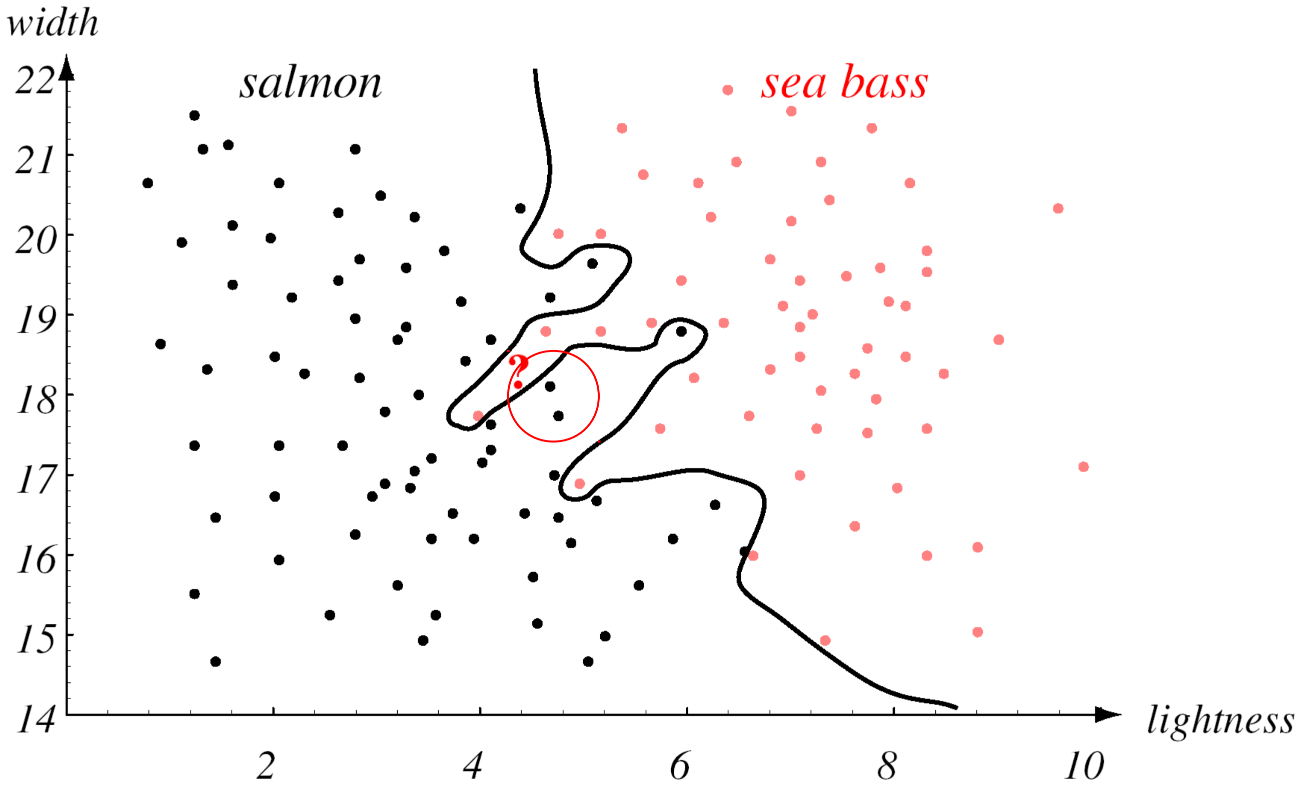
- Possible to eliminate all the errors by means of a more complex boundary?
- The decision boundary is defined by the classification system, thus

more complex boundary

$\Updownarrow$

more complex classifier

Better solutions?



A very complex decision boundary ***with no errors***.
Which is the decision for the new sample?

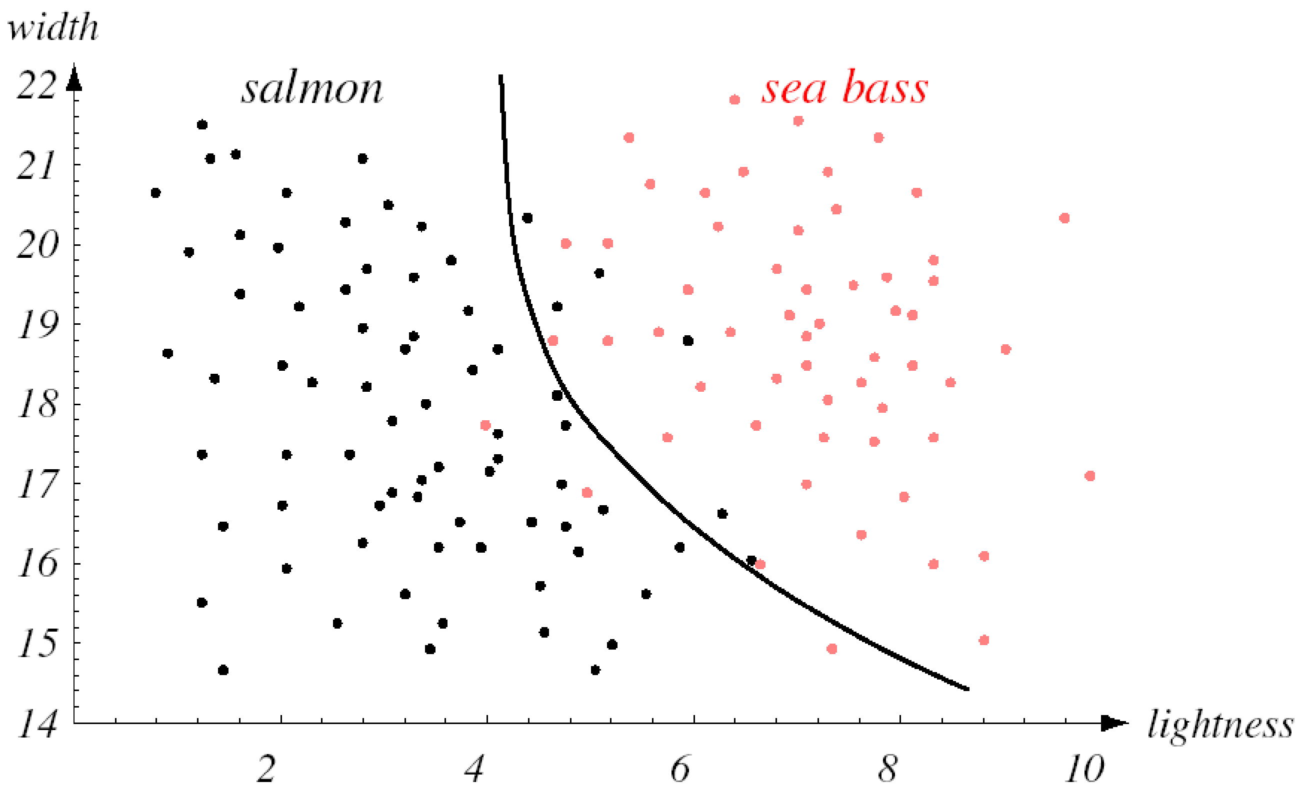
The generalization problem

- We can achieve no error **on the set of data** we use for building the classifier by means of a complex decision boundary.
- **BUT** a good recognition quality is not assured on the new samples.
- The generalization problem is **fundamental** when designing the recognition system

The generalization problem

- We must consider a trade-off between
  - Recognition quality on the data set
  - Recognition capability on the new samples (generalization capability)
- It would be better to tolerate some errors on the data set if this leads to a classifier with better generalization capability

The generalization problem



A decision boundary slightly more complex than the linear one.
A good trade-off?

How to build a classifier?

- Once we know the characteristics a good classifier should have, the questions is: how to build it?
- **Algorithmic solution?**
  Is it possible to write an algorithm able to recognize characters, speech, images?
- After 50 years of researchs, the answer is definitely NO.
- Why?
- Other possible choices?

Learning

- *Learning*: a method that combines empirical information coming from the environment with a priori knowledge about the problem in order to build a classifier
- The empirical information is usually provided by a set of samples of the objects.
- The a priori knowledge is in the form of invariants, correlations among data, ...

Machine Learning Algorithm

- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E” (Mitchell, 1997)

The Tasks

- Tasks that are too difficult to solve with fixed programs written and designed by human beings
- The process of learning itself is not the task. Learning is the means of attaining the ability to perform the task.

The Tasks

- Tasks are usually described in terms of how the machine learning system should process an *example*.
- An *example* is a collection of features that have been quantitatively measured from some object or event that we want the machine learning system to process.

Types of tasks

- **Classification**: the computer program is asked to specify which of k categories some input belongs to. Special case k=2.
- To solve this task, the learning algorithm is usually asked to produce a function $f : R^n \to \{1, 2, \ldots, k\}$. When $y = f(\mathbf{x})$, the model assigns an input described by vector $\mathbf{x}$ to a category identified by numeric code y.

Types of tasks

- **Detection**: the computer program is asked to decide if the input belongs to an *object-class*.
- A sort of classification task in which one must discriminate between the object class and the *Rest-of-theWorld* class.
- The learning algorithm is usually asked to produce a function $f : R^n \to \{0, 1\}$, where $1 \longrightarrow$ object, $0 \longrightarrow$ non-object.

Types of tasks

- **Regression**: the computer program is asked to predict a numerical value given some input. To solve this task, the learning algorithm is asked to output a function $f : R^n \to R$.
- Task similar to classification, except that the format of output is different.

The Experience

- The experience is tipically provided to the learning algorithm in terms of a *dataset.*
- A dataset is a collection of many examples (aka *data points*).
- Two possible paradigms:
  - *Supervised Learning*:

    the class of each example is known (*label* or *target*). The goal is to minimize the errors.
  - *Unsupervised Learning (Clustering)*:

    the class of each example is not given. The goal is to extract some features of the data: e.g. probability distributions. Another possible goal is to *cluster* the samples tipically on the basis of a distance. The number of clusters to be produced could be known in advance.

The Performance Measure

- A quantitative measure of the performance of the learning algorithm specific to the task T being carried out by the system.
- Classification → **Accuracy**: the proportion of examples for which the model produces the correct output.
- Equivalent information given by measuring the error rate, the proportion of examples for which the model produces an incorrect output.
- Questions:
  - Which is the relative size of each class?
  - Do differente errors have the same consequences, and thus the same costs?

Building a machine learning system

- Two stages
  - Training
  - Performance assessment (test)

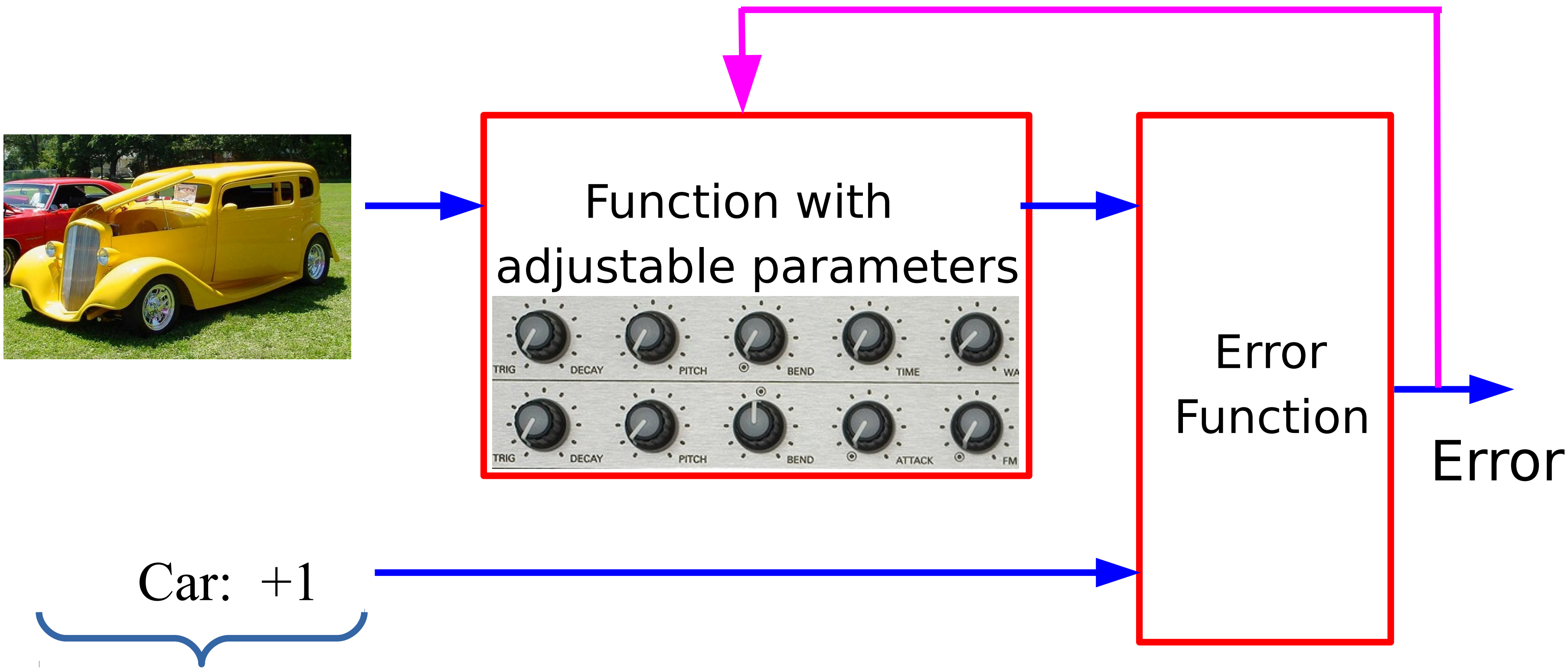
Training

- The experience is used to set the parameters of the computer program so as to maximize a performance measure.
- The esperience is given in terms of a set of examples: the *training set*.
- A training set is tipically given as a set of pairs $(\mathbf{x}_i, y_i)$with $i = 1, \cdots, n_{train}$ where, for each example $\mathbf{x}_i$ the corresponding label y_i is provided.

Performance assessment

- Question: how well the machine learning algorithm performs on data that it has not seen before?
- This determines how well it will work when deployed in the real world (that is our real goal!)
- We therefore evaluate these performance measures using a *test set* of data <u>that is separate from the data used for training the machine learning system</u>.
- The test set also is given as a set of pairs $(\mathbf{x}_i, y_i)$ with $i = 1, \cdots, n_{test}$

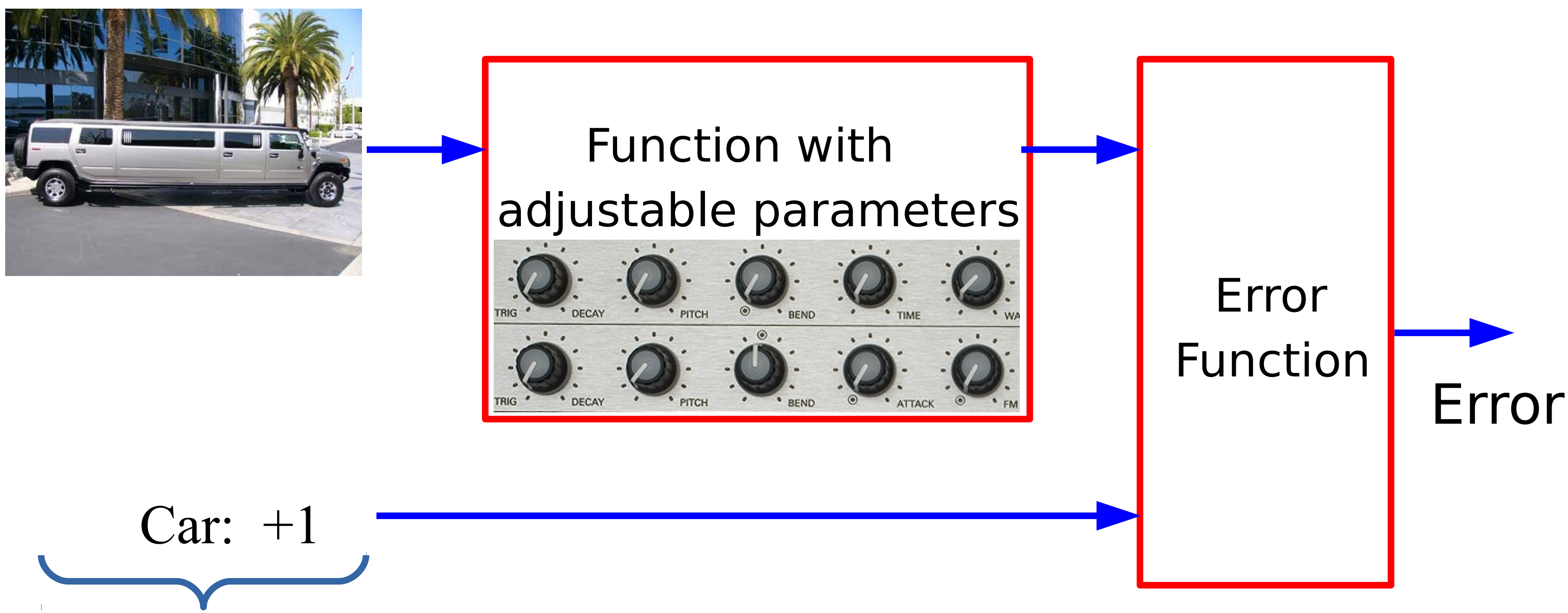
Training



Training Set

Figure: Yann Le Cun
Facebook AI Research,

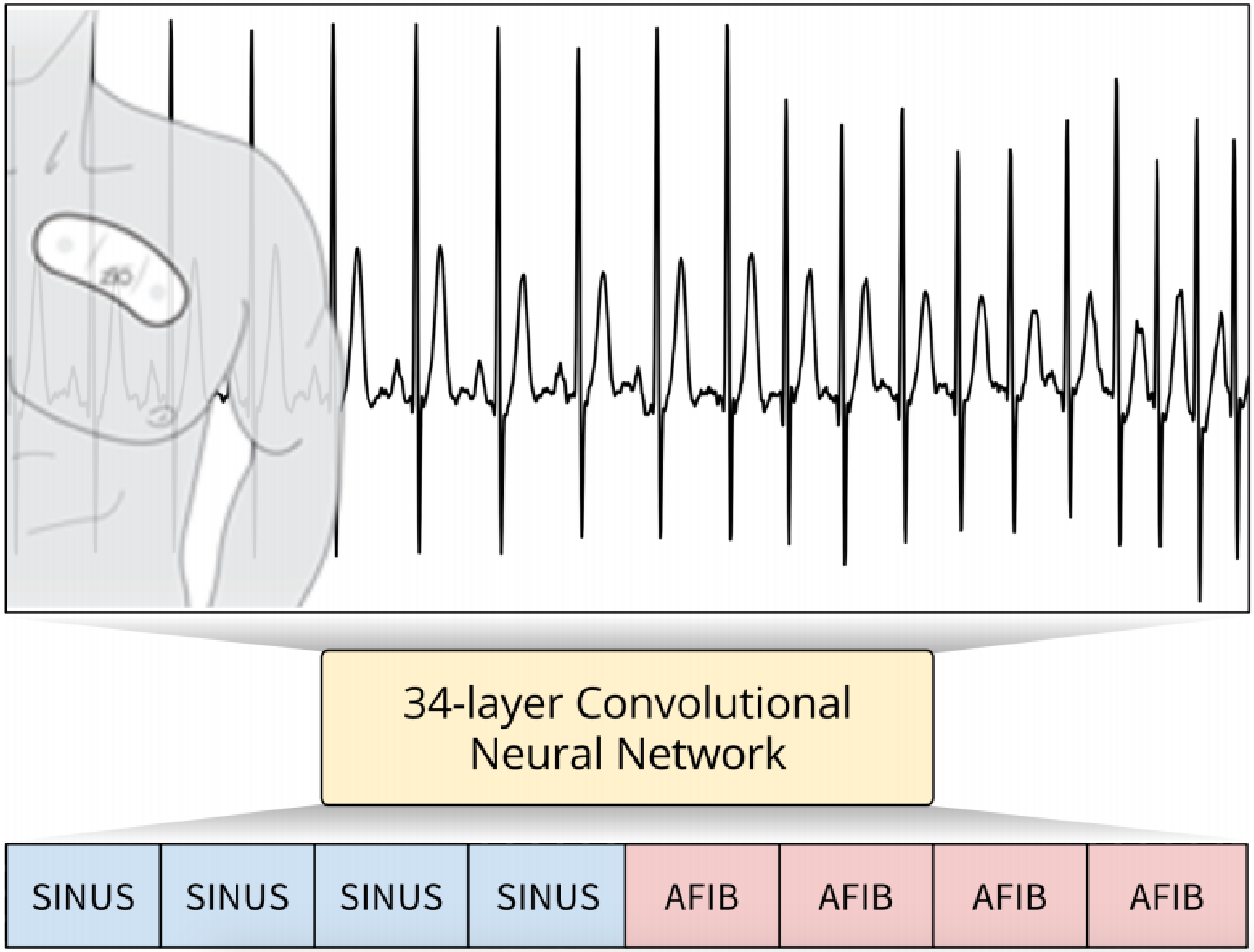
Test



Test Set

Figure: Yann Le Cun
Facebook AI Research,

Classification task



sinus rhythm	Atrial Fibrillation

Mapping a sequence of ECG samples to a sequence of rhythm classes

"Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks"

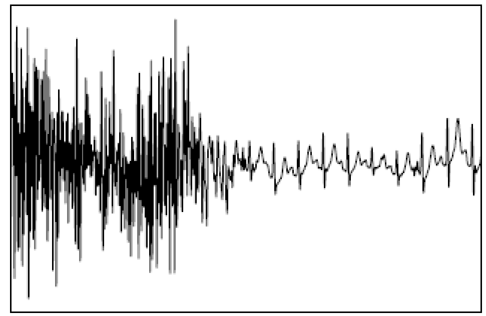
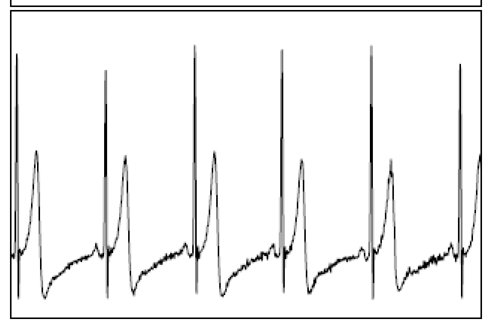
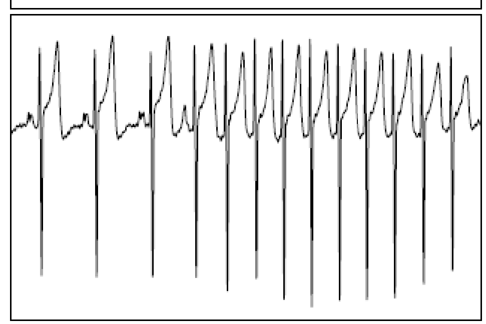
P. Rajpurkar et al.

arXiv:1707.01836v1 [cs.CV]

Classification task

- Example: a 30 seconds single-lead ECG signal sampled at 200Hz
- Classes: 14 classes
  - 12 heart arrhythmias
  - sinus rhythm
  - Noise
- Training set: 64,121 ECG records from 29,163 patients
- Test set: 336 records from 328 unique patients

Classification task

NOISE	Noise	
SINUS	Sinus Rhythm	
SVT	Supraventricular Tachycardia	

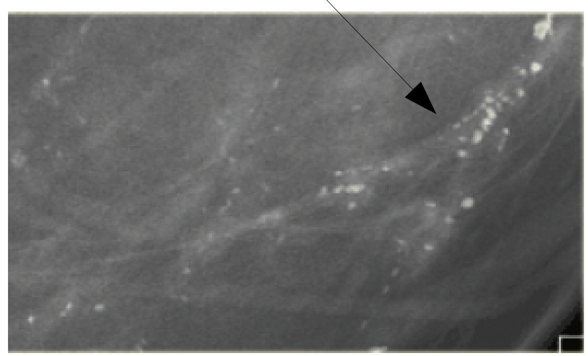
Classification task

	Seq		Set	
	Model	Cardiol.	Model	Cardiol.
Class-level F1 Score				
AFIB	**0.604**	0.515	**0.667**	0.544
AFL	**0.687**	0.635	**0.679**	0.646
AVB_TYPE2	**0.689**	0.535	**0.656**	0.529
BIGEMINY	**0.897**	0.837	**0.870**	0.849
CHB	**0.843**	0.701	**0.852**	0.685
EAR	**0.519**	0.476	**0.571**	0.529
IVR	**0.761**	0.632	**0.774**	0.720
JUNCTIONAL	0.670	**0.684**	**0.783**	0.674
NOISE	**0.823**	0.768	**0.704**	0.689
SINUS	**0.879**	0.847	**0.939**	0.907
SVT	**0.477**	0.449	**0.658**	0.556
TRIGEMINY	**0.908**	0.843	**0.870**	0.816
VT	0.506	**0.566**	0.694	**0.769**
WENCKEBACH	**0.709**	0.593	**0.806**	0.736
Aggregate Results				
Precision (PPV)	**0.800**	0.723	**0.809**	0.763
Recall (Sensitivity)	**0.784**	0.724	**0.827**	0.744
F1	**0.776**	0.719	**0.809**	0.751

Table 1. The top part of the table gives a class-level comparison of the expert to the model F1 score for both the Sequence and the Set metrics. The bottom part of the table shows aggregate results over the full test set for precision, recall and F1 for both the Sequence and Set metrics.

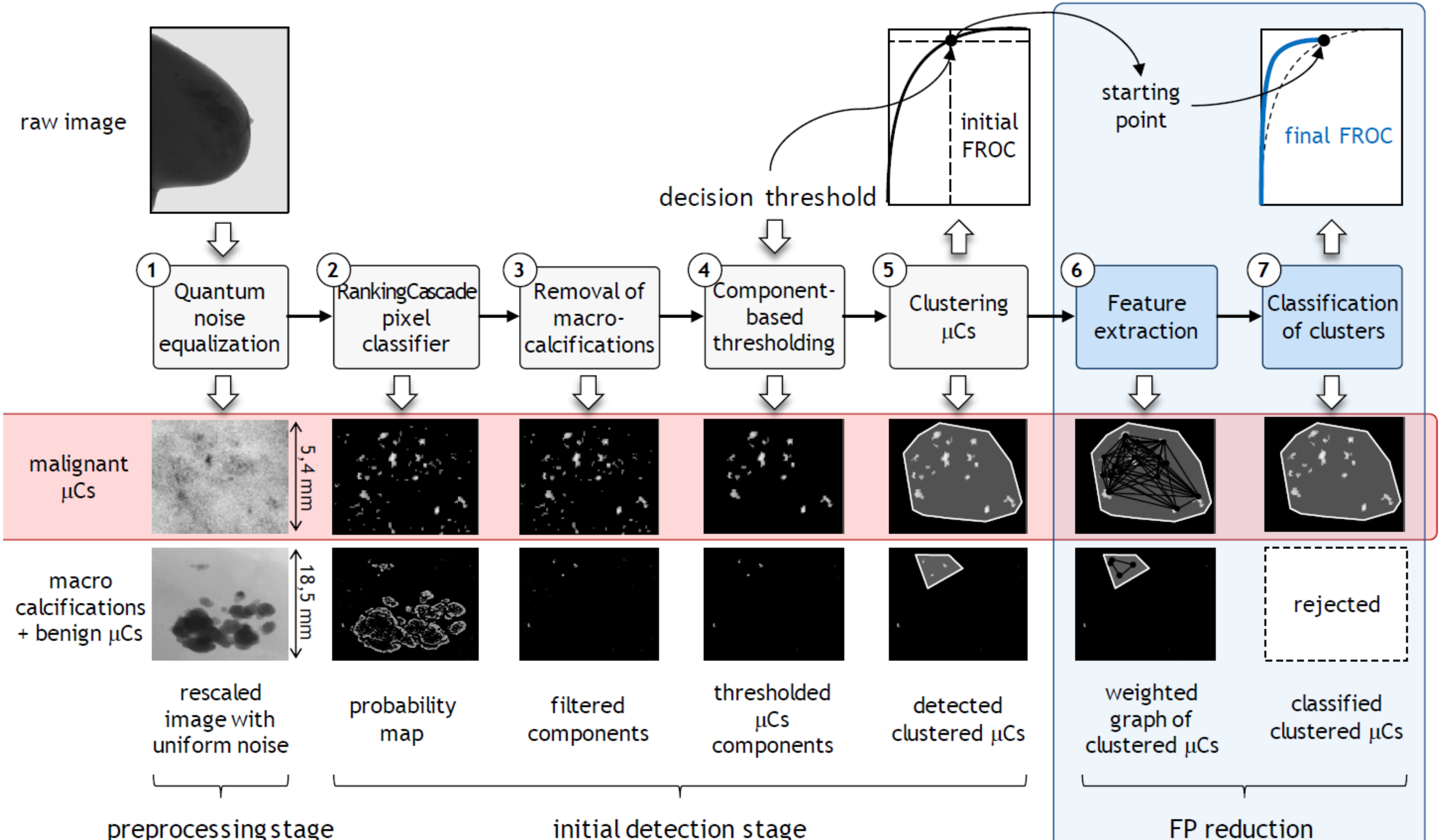
Detection task

calcifications form within the duct



"Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications"
A. Bria, N. Karssemeijer, F. Tortorella
Medical Image Analysis, 2014

Detection task



Detection task

- Example: window of 12x12 pixels
- Objects to be detected:
  - Microcalcifications
  - Clusters of microcalcifications
- Training Set: 252 images from 129 abnormal cases (70 benign and 59 malignant)
- Test Set: 447 images from 242 abnormal cases (186 benign and 56 malignant)

Detection task

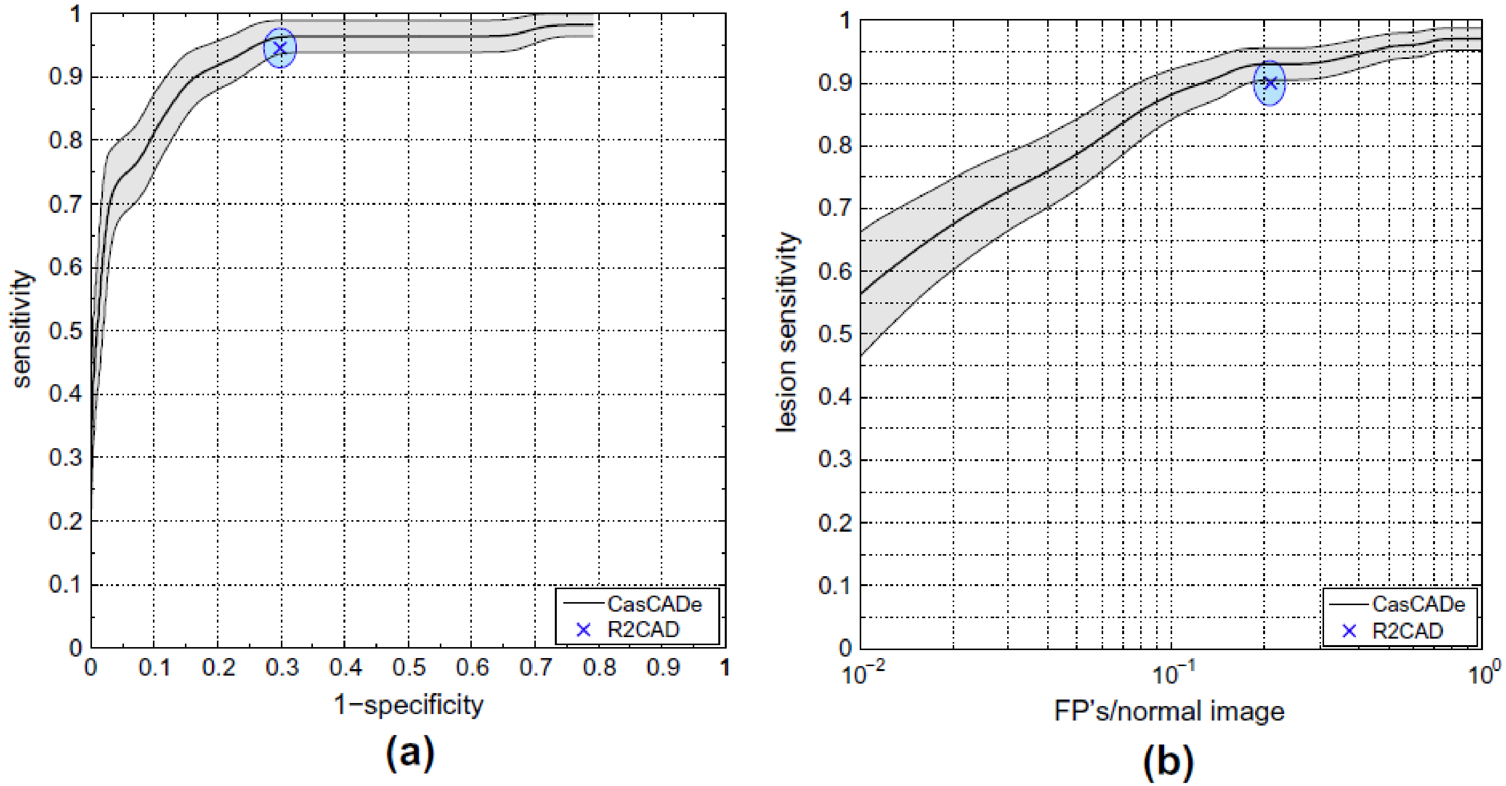


Fig. 11. Average case-based ROC curves (a) and lesion-based FROC curves and (b) of CasCADe and R2CAD obtained from 5000 bootstrap samples. For CasCADe, confidence bands indicate the standard deviation along the sensitivity axis. For R2CAD, the marks denote the average operating points and the ellipses indicate the standard deviation along both the axes.