BMT 823 Neural & Cognitive Systems
Slides Series 5

Pattern Recognition & Machine Learning

Prof. Dr. rer. nat. Dr. rer. med. habil. Daniel J. Strauss

What is pattern recognition?

"The assignment of a physical object or event to one of several prespecified categeries" -- Duda & Hart

- A **pattern** is an object, process or event that can be given a name.
- A **pattern class** (or category) is a set of patterns sharing common attributes and usually originating from the same source.
- During **recognition** (or **classification**) given objects are assigned to prescribed classes.
- A **classifier** is a machine which performs classification.

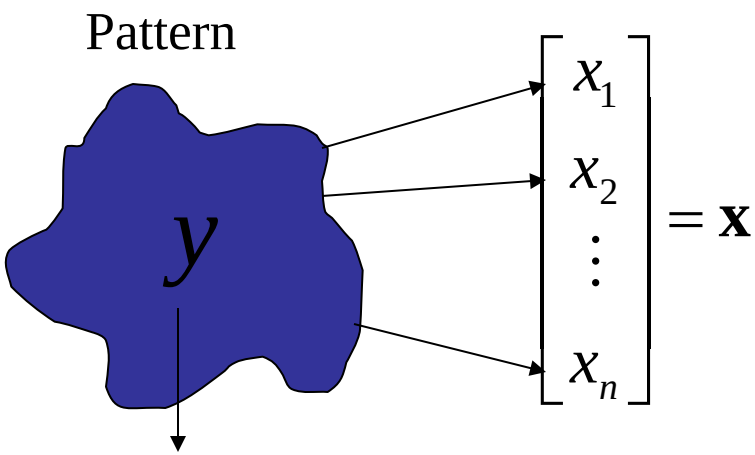
Examples of applications

- **Optical Character Recognition (OCR)**
  - Handwritten: sorting letters by postal code, input device for PDA's.
  - Printed texts: reading machines for blind people, digitalization of text documents.
- **Biometrics**
  - Face recognition, verification, retrieval.
  - Finger prints recognition.
  - Speech recognition.
- **Diagnostic systems**
  - Medical diagnosis: X-Ray, EKG analysis.
  - Machine diagnostics.
- **Military applications**
  - Automated Target Recognition (ATR).
  - Image segmentation and analysis

Approaches

- **Statistical PR:** based on underlying statistical model of patterns and pattern classes.
- **Structural (or syntactic) PR:** pattern classes represented by means of formal structures as grammars, strings, etc.
- **Neural networks:** classifier is represented as a network of cells modeling neurons of the human brain.

Basic concepts



Feature vector $\mathbf{x} \in X$

- A vector of observations (measurements).
- $\mathbf{x}$ is a point in feature space X.

Hidden state $y \in Y$

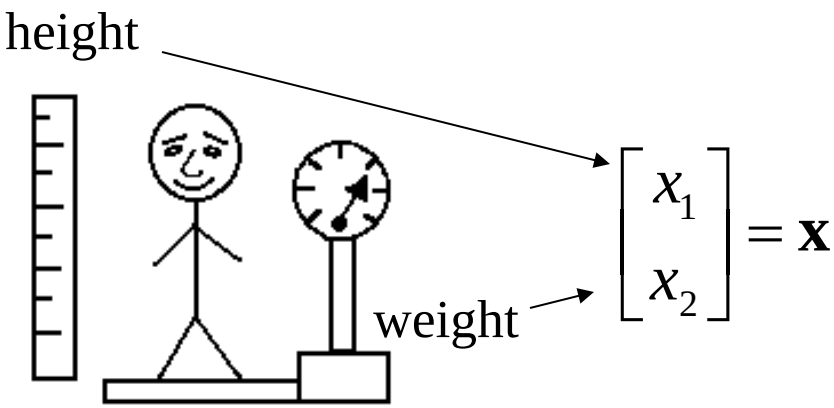
- Cannot be directly measured.
- Patterns with equal hidden state belong to the same class.

Task

- To design a classifier (decision rule) $\mathrm{q} : X \to Y$

which decides about a hidden state based on an observation.

Example



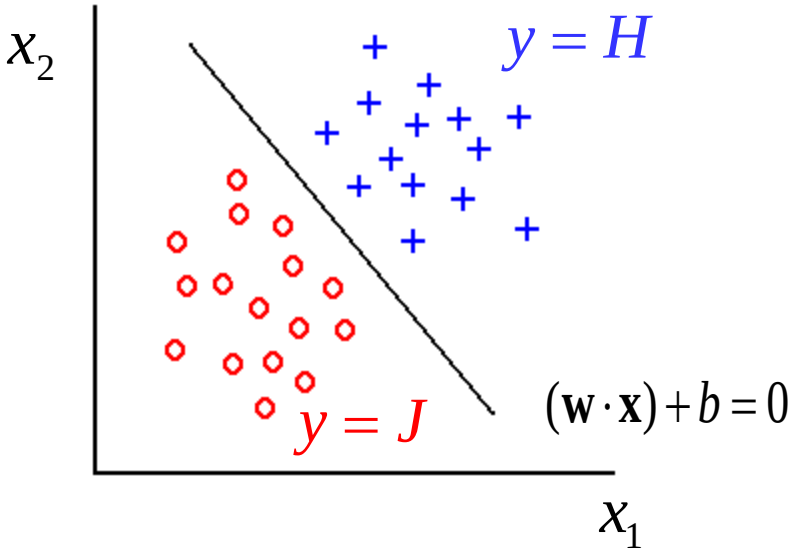
The set of hidden state is $Y = \{H, J\}$

The feature space is $X = \Re^2$

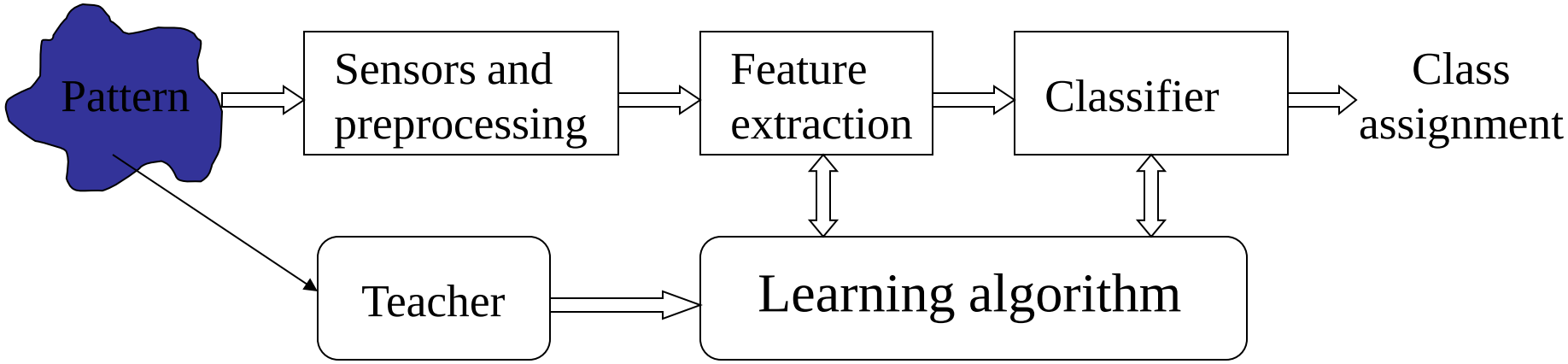
Training examples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$

Linear classifier:

$$q(\mathbf{x}) = \begin{cases} H & if \quad (\mathbf{w} \cdot \mathbf{x}) + b \geq 0 \\ J & if \quad (\mathbf{w} \cdot \mathbf{x}) + b < 0 \end{cases}$$



Components of PR system

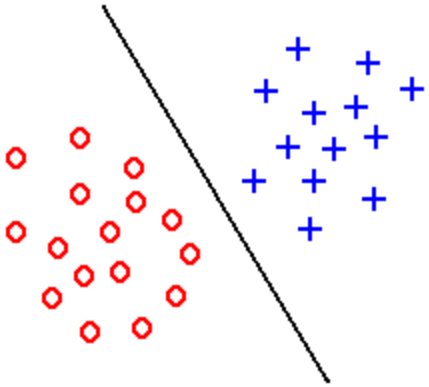


- **Sensors and preprocessing**.
- **A feature extraction** aims to create discriminative features good for classification.
- **A classifier**.
- **A teacher** provides information about hidden state -- supervised learning.
- **A learning algorithm** sets PR from training examples.

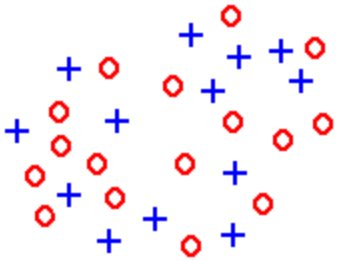
Feature extraction

Task: to extract features which are good for classification.

Good features:
- Objects from the same class have similar feature values.
- Objects from different classes have different values.

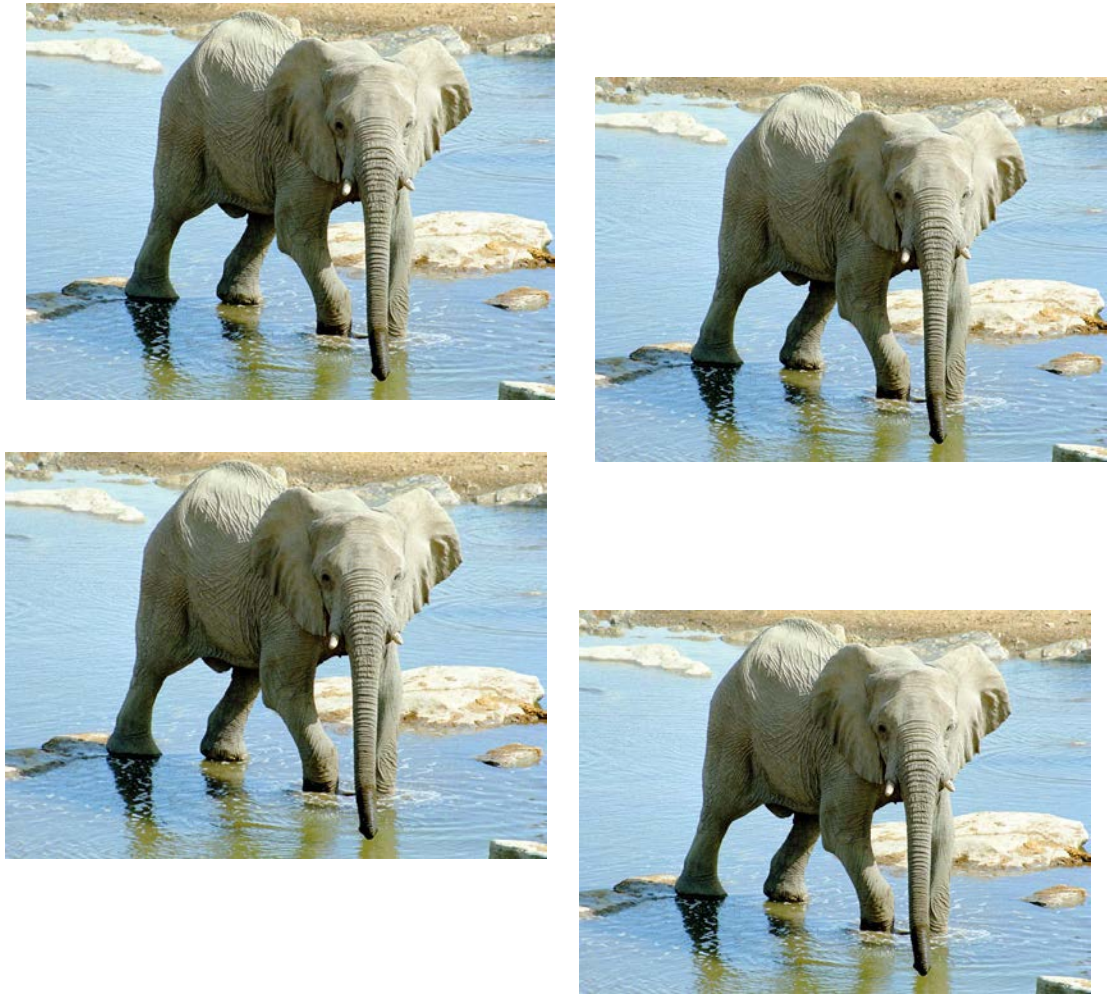


"Good" features



"Bad" features

good and bad features



good and bad features

lousy feature

color:

gray vs. gray

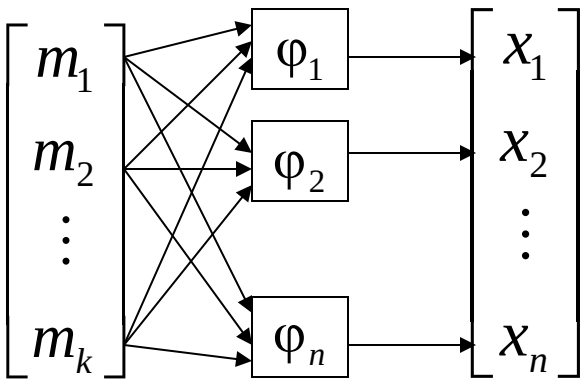
fine feature

weight:

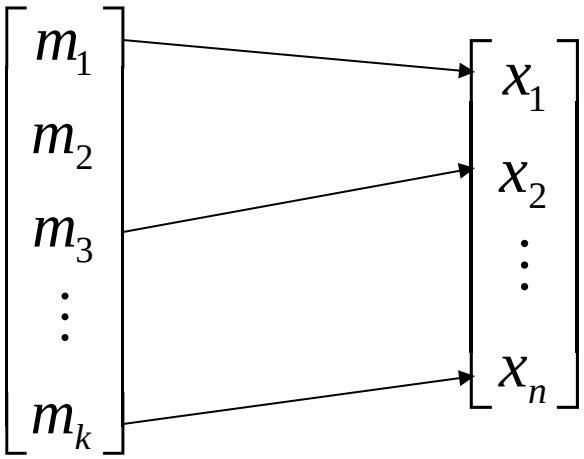
0.020kg vs. 7000.000kg

Feature extraction methods

Feature extraction



Feature selection



Problem can be expressed as optimization of parameters of featrure extractor $\varphi(\theta)$.

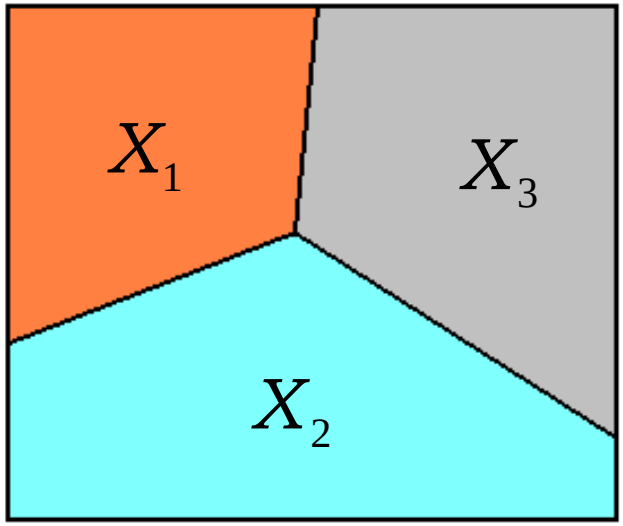
Supervised methods: objective function is a criterion of separability (discriminability) of labeled examples, e.g., linear discriminat analysis (LDA).

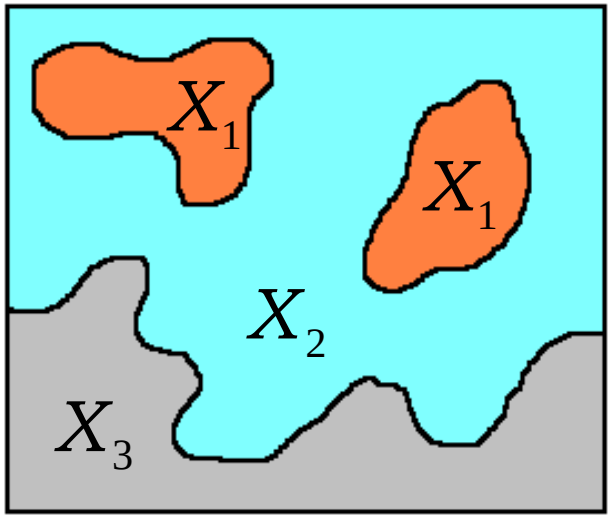
Unsupervised methods: lower dimesional representation which preserves important characteristics of input data is sought for, e.g., principal component analysis (PCA).

Classifier

A classifier partitions feature space X into **class-labeled regions** such that

$$X = X_1 \cup X_2 \cup \ldots \cup X_{|Y|} \quad \text{and} \quad X_1 \cap X_2 \cap \ldots \cap X_{|Y|} = \{0\}$$





The classification consists of determining to which region a feature vector **x** belongs to.

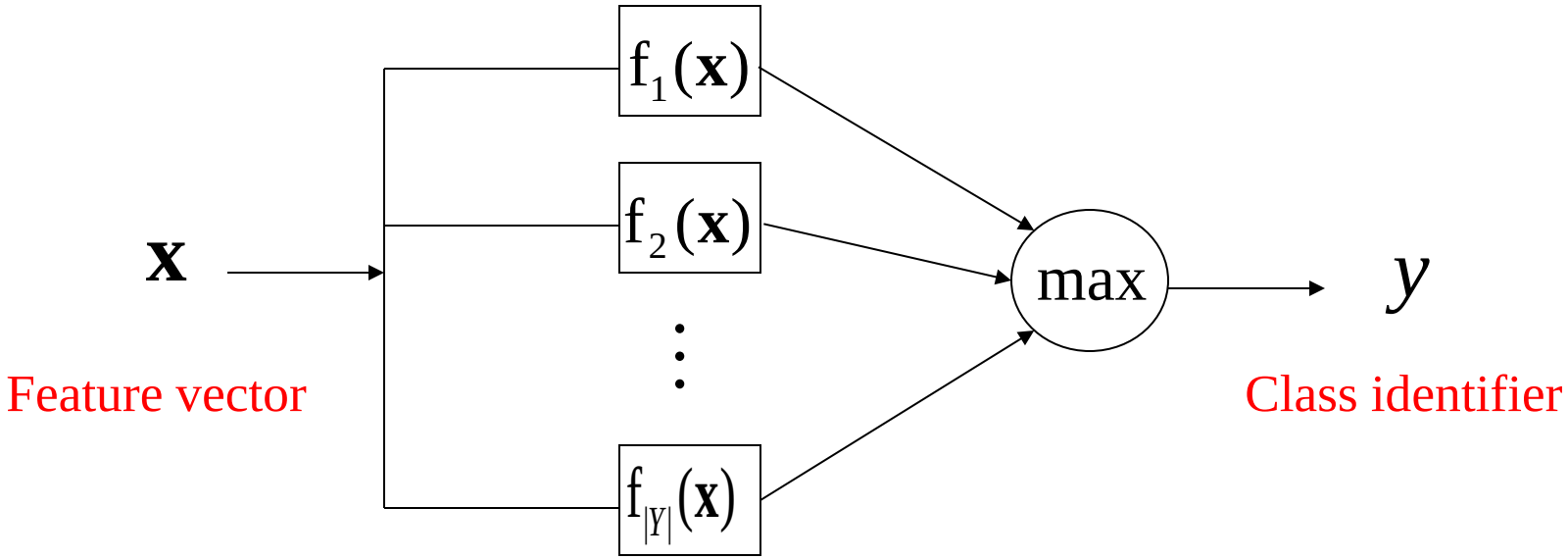
Borders between **decision boundaries** are called decision regions.

Representation of classifier

A classifier is typically represented as a set of discriminant functions

$$\mathrm{f}_i(\mathbf{x}) : X \rightarrow \Re, i = 1, \ldots, |Y|$$

The classifier assigns a feature vector **x** to the *i*-the class if $\mathrm{f}_i(\mathbf{x}) > \mathrm{f}_j(\mathbf{x}) \quad \forall j \neq i$



Discriminant function

Bayesian decision making

- The Bayesian decision making is a fundamental statistical approach which allows to design the optimal classifier if complete statistical model is known.

Definition:

Obsevations	X	A loss function	$W : Y \times D \to R$
Hidden states	Y	A decision rule	$\mathrm{q} : X \to D$
Decisions	D	A joint probability	$\mathrm{p}(\mathbf{x}, y)$

Task: to design decision rule q which minimizes Bayesian risk

$$\mathrm{R}(\mathrm{q}) = \sum_{y \in Y} \sum_{x \in X} \mathrm{p}(\mathbf{x}, y)\, \mathrm{W}(\mathrm{q}(\mathbf{x}), y)$$

Example of Bayesian task

Task: minimization of classification error.

A set of decisions D is the same as set of hidden states Y.

0/1 - loss function used $\mathrm{W}(\mathrm{q}(\mathbf{x}), y) = \begin{cases} 0 & if \quad \mathrm{q}(\mathbf{x}) = y \\ 1 & if \quad \mathrm{q}(\mathbf{x}) \neq y \end{cases}$

The Bayesian risk R(q) corresponds to probability of misclassification.

The solution of Bayesian task is

$$\mathrm{q}^* = \arg\min_{\mathrm{q}} \mathrm{R}(\mathrm{q}) \Rightarrow y^* = \arg\max_{y} p(y \mid \mathbf{x}) = \arg\max_{y} \frac{\mathrm{p}(\mathbf{x} \mid y)\,\mathrm{p}(y)}{\mathrm{p}(\mathbf{x})}$$

Limitations of Bayesian approach

- The statistical model $p(\mathbf{x},y)$ is mostly not known therefore learning must be employed to estimate $\overline{p}(\mathbf{x},y)$ from training examples $\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_\ell,y_\ell)\}$ -- plug-in Bayes.
- Non-Bayesian methods offers further task formulations:
  - A partial statistical model is avaliable only:
    - $p(y)$ is not known or does not exist.
    - $p(\mathbf{x}|y,\theta)$ is influenced by a non-random intervetion θ.
  - The loss function is not defined.

Discriminative approaches

Given a class of classification rules $q(\mathbf{x};\boldsymbol{\theta})$ parametrized by $\boldsymbol{\theta}\in\Xi$ the task is to find the "best" parameter $\boldsymbol{\theta}^*$ based on a set of training examples $\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_\ell,y_\ell)\}$ -- supervised learning.

The task of learning: recognition which classification rule is to be used.

The way how to perform the learning is determined by a selected inductive principle.

Empirical risk minimization principle

The true expected risk R(q) is approximated by empirical risk

$$\mathrm{R}_{\mathrm{emp}}(\mathrm{q}(x;\boldsymbol{\theta})) = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathrm{W}(\mathrm{q}(\mathbf{x}_i;\boldsymbol{\theta}), y_i)$$

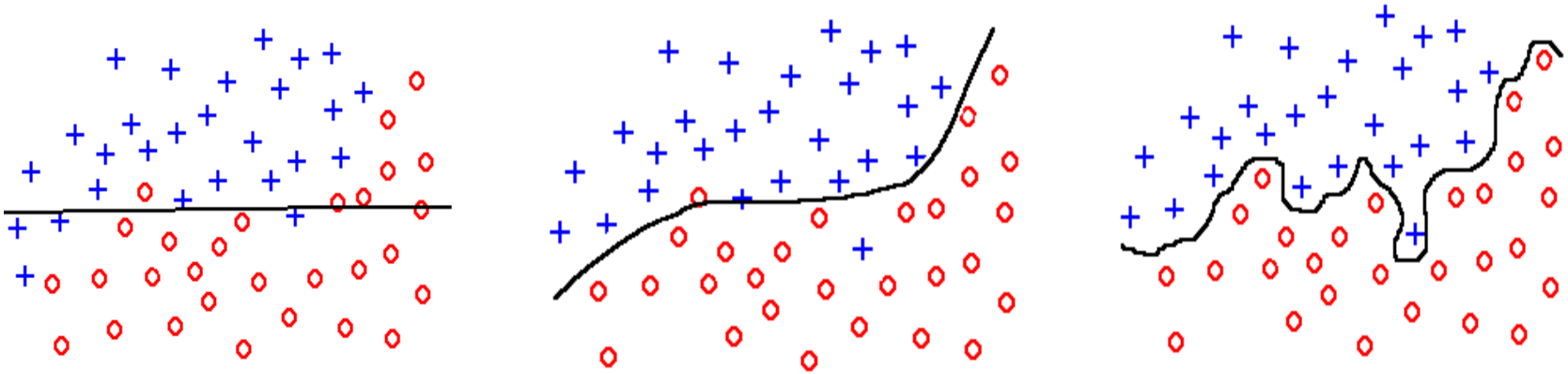
The learning based on the empirical minimization principle is defined as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathrm{R}_{\mathrm{emp}}(\mathrm{q}(\mathbf{x};\boldsymbol{\theta}))$$

Examples of algorithms: Perceptron, Back-propagation, etc.

Overfitting and underfitting

Problem: how rich class of classifications $q(\mathbf{x};\boldsymbol{\theta})$ to use.



underfitting good fit overfitting

Problem of generalization: a small emprical risk R_{emp} does not imply small true expected risk R.

Structural risk minimization principle

Statistical learning theory -- Vapnik & Chervonenkis.

An upper bound on the expected risk of a classification rule $q \in Q$

$$R(q) \le R_{emp}(q) + R_{str}\left(\frac{1}{\ell}, h, \log\frac{1}{\sigma}\right)$$

where ℓ is number of training examples, h is VC-dimension of class of functions Q and $1-\sigma$ is confidence of the upper bound.

SRM principle: from a given nested function classes $Q_1, Q_2, \ldots, Q_m$, such that

$$h_1 \le h_2 \le \ldots \le h_m$$

select a rule q^* which minimizes the upper bound on the expected risk.

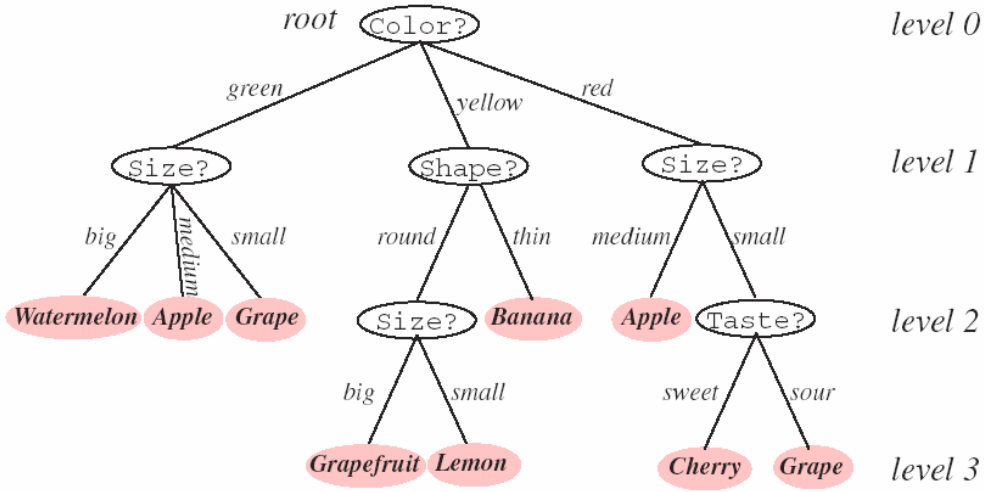


FIGURE 8.1. Classification in a basic decision tree proceeds from top to bottom. The questions asked at each node concern a particular property of the pattern, and the downward links correspond to the possible values. Successive nodes are visited until a terminal or leaf node is reached, where the category label is read. Note that the same question, `Size?`, appears in different places in the tree and that different questions can have different numbers of branches. Moreover, different leaf nodes, shown in pink, can be labeled by the same category (e.g., **Apple**). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

an example for machine perception

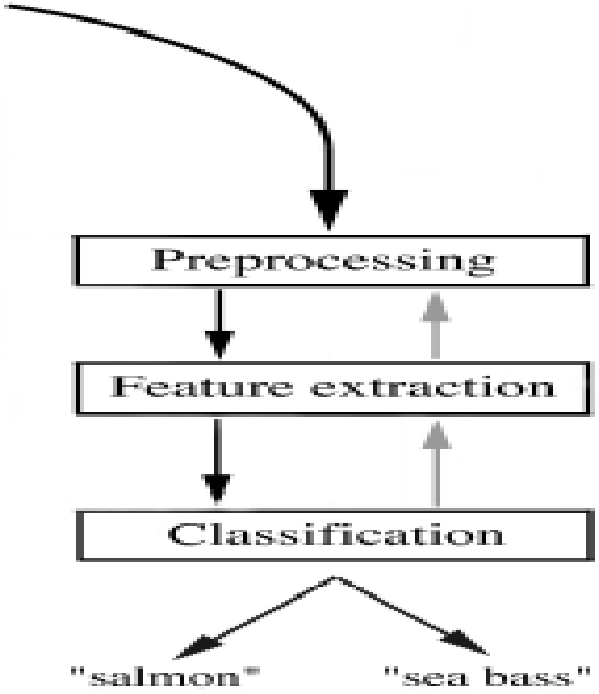
An Example

- "Sorting incoming Fish on a conveyor according to species using optical sensing"

Species

Sea bass

Salmon


Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

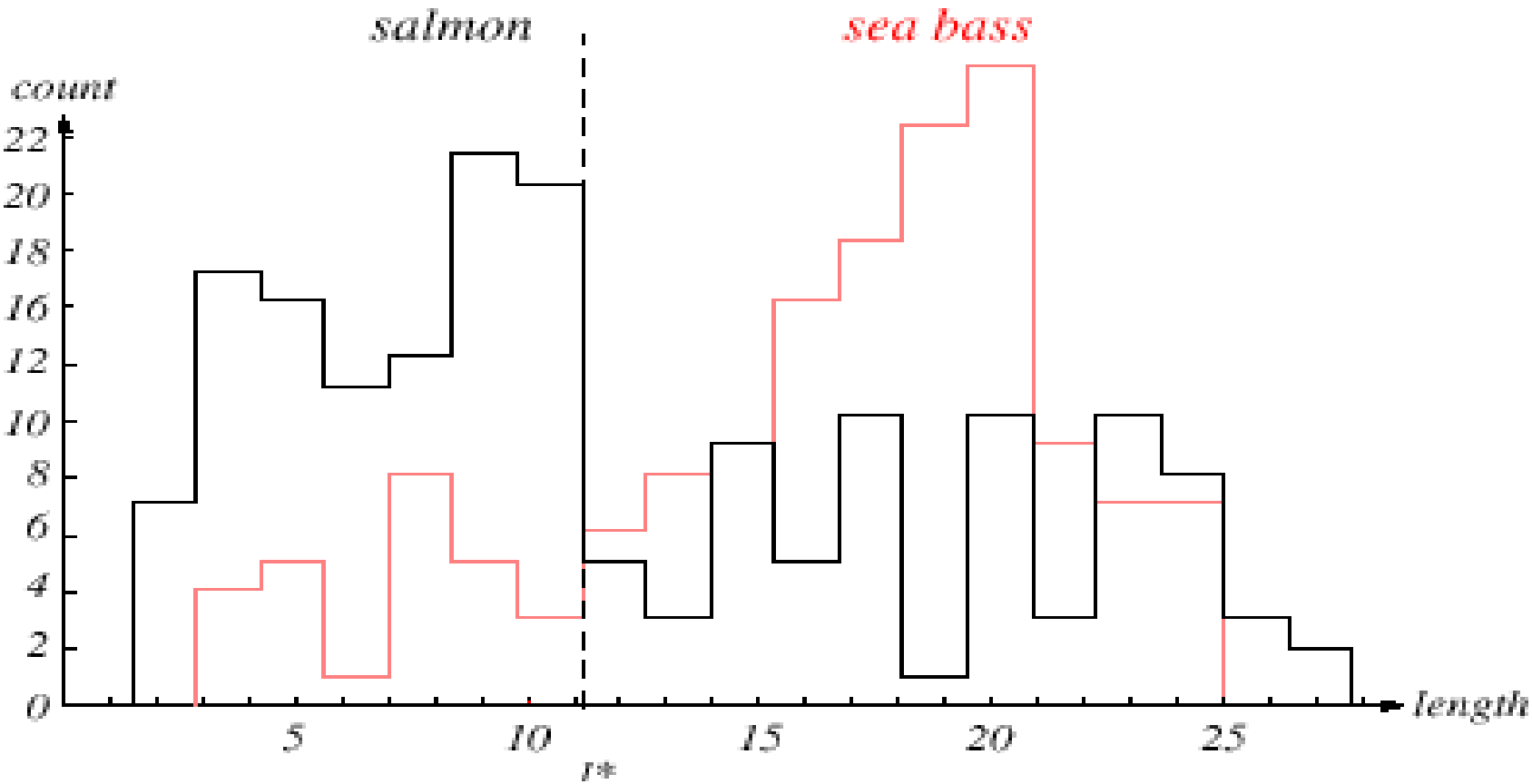
- Problem Analysis
  - Set up a camera and take some sample images to extract features

    - Length
    - Lightness
    - Width
    - Number and shape of fins
    - Position of the mouth, etc…

    - This is the set of all suggested features to explore for use in our classifier!

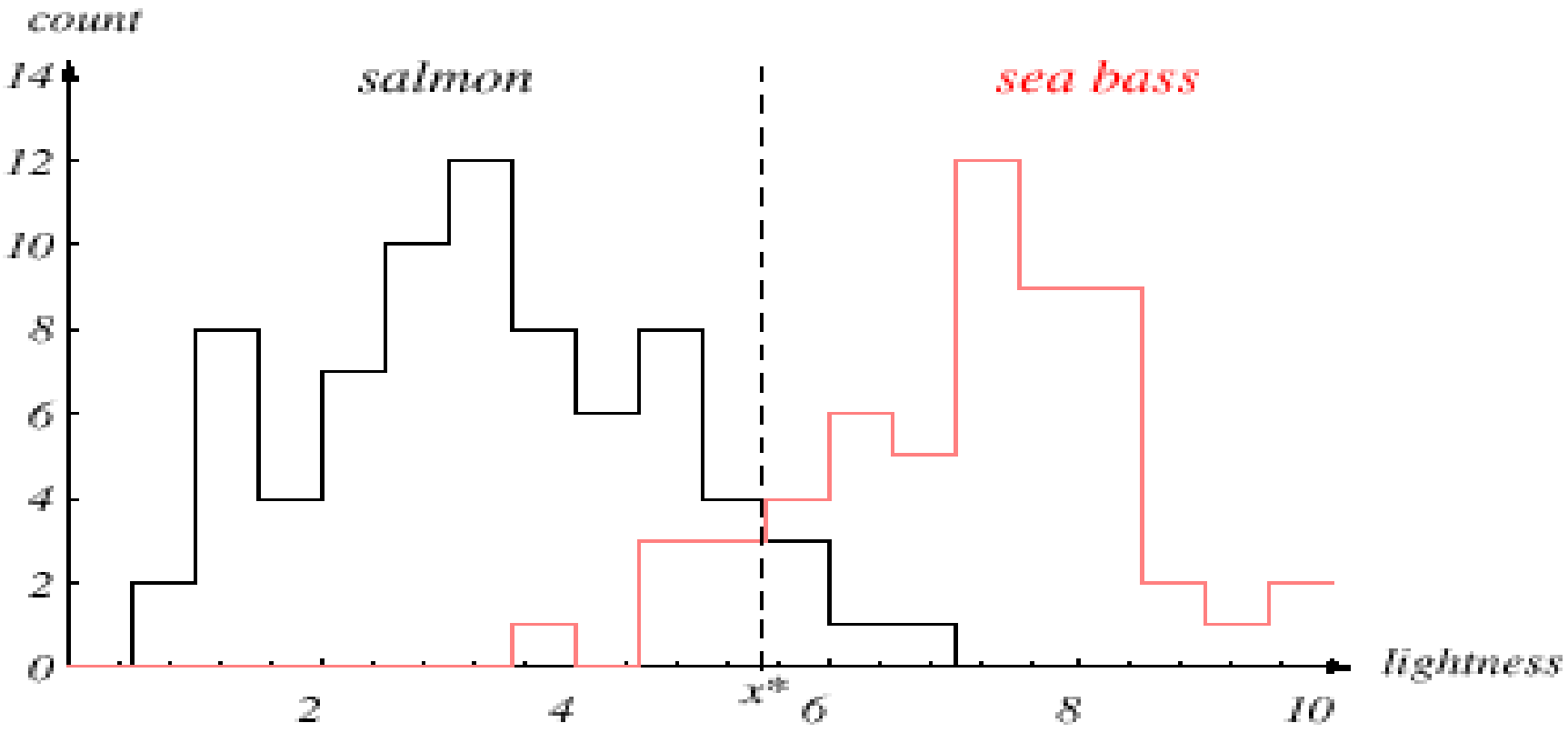
- Preprocessing
  - Use a segmentation operation to isolate fishes from one another and from the background
- Information from a single fish is sent to a feature extractor whose purpose is to reduce the data by measuring certain features
- The features are passed to a classifier

- Classification
  - Select the length of the fish as a possible feature for discrimination


salmon
sea bass
count
22
20
18
16
12
10
8
6
4
2
0
5
10
15
20
25
length
l*

The length is a poor feature alone!

Select the lightness as a possible feature.


count
salmon
sea bass
14
12
10
8
6
4
2
0
2
4
x*
6
8
10
lightness

- Threshold decision boundary and cost relationship
  - Move our decision boundary toward smaller values of lightness in order to minimize the cost (reduce the number of sea bass that are classified salmon!)

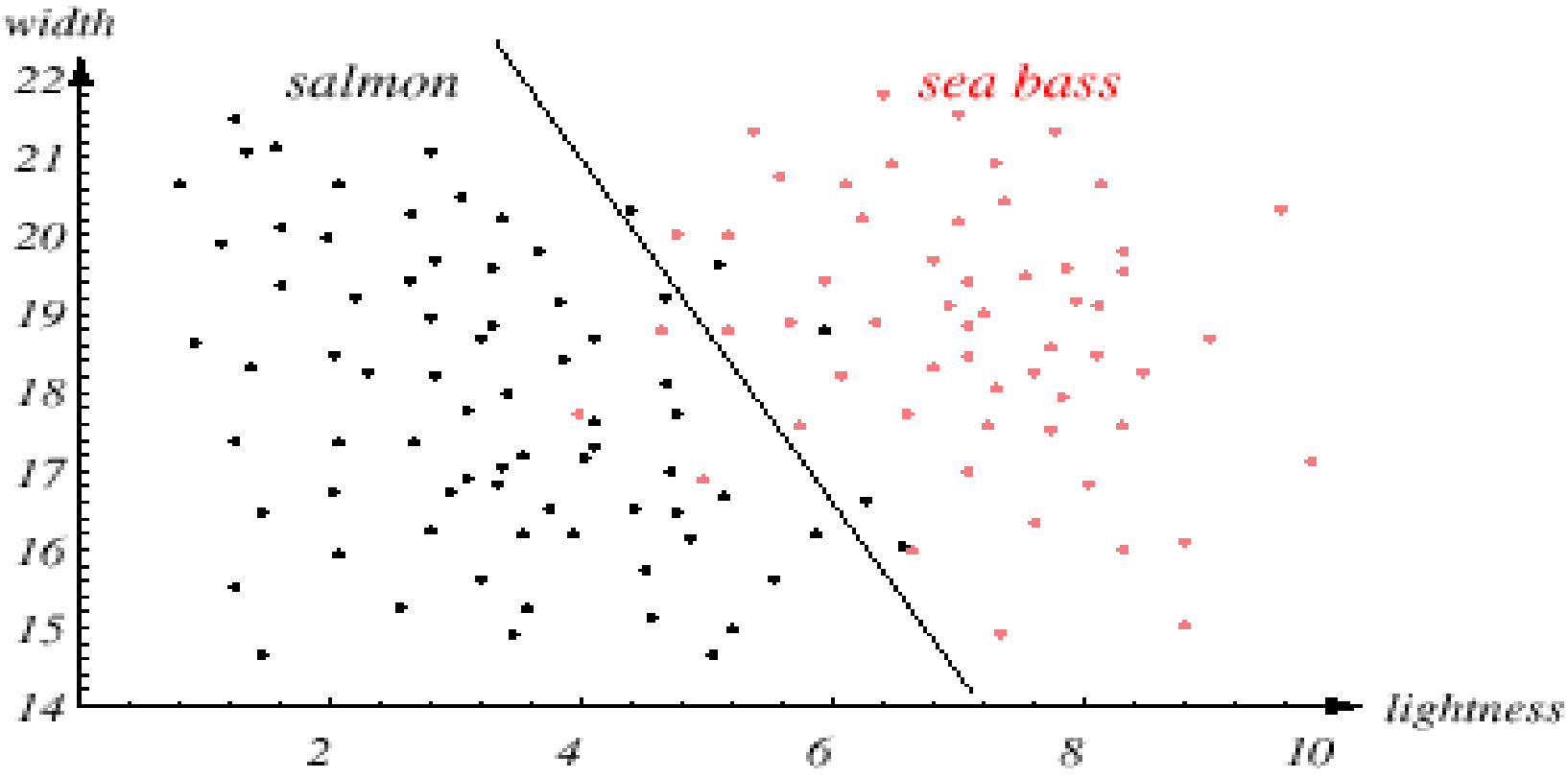
Task of decision theory

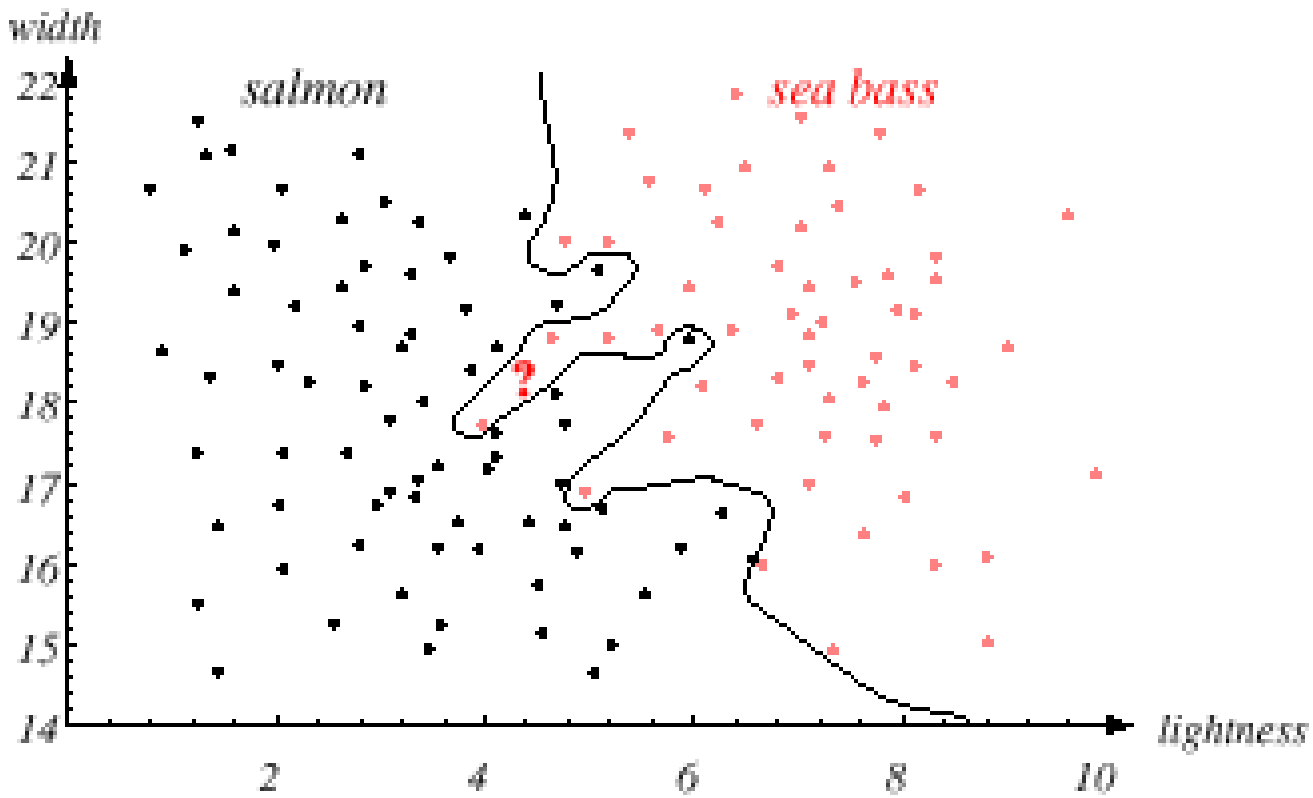
- Adopt the lightness and add the width of the fish

Fish → $x^T = [x_1, x_2]$

x_1 → **Lightness**

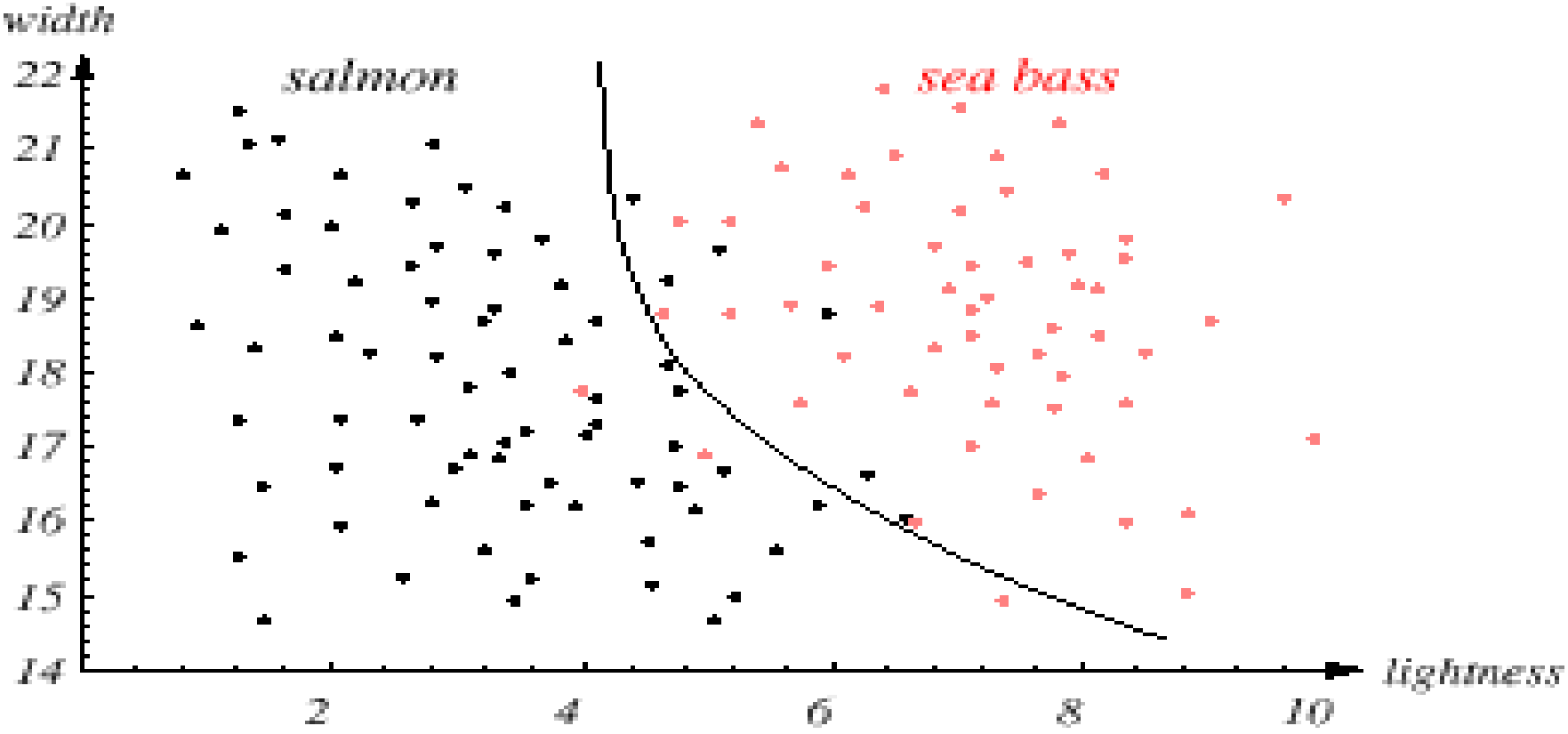
x_2 → **Width**


width
22
21
20
19
18
17
16
15
14
salmon
sea bass
2
4
6
8
10
lightness


width
salmon
sea bass
?
22
21
20
19
18
17
16
15
14
2
4
6
8
10
lightness

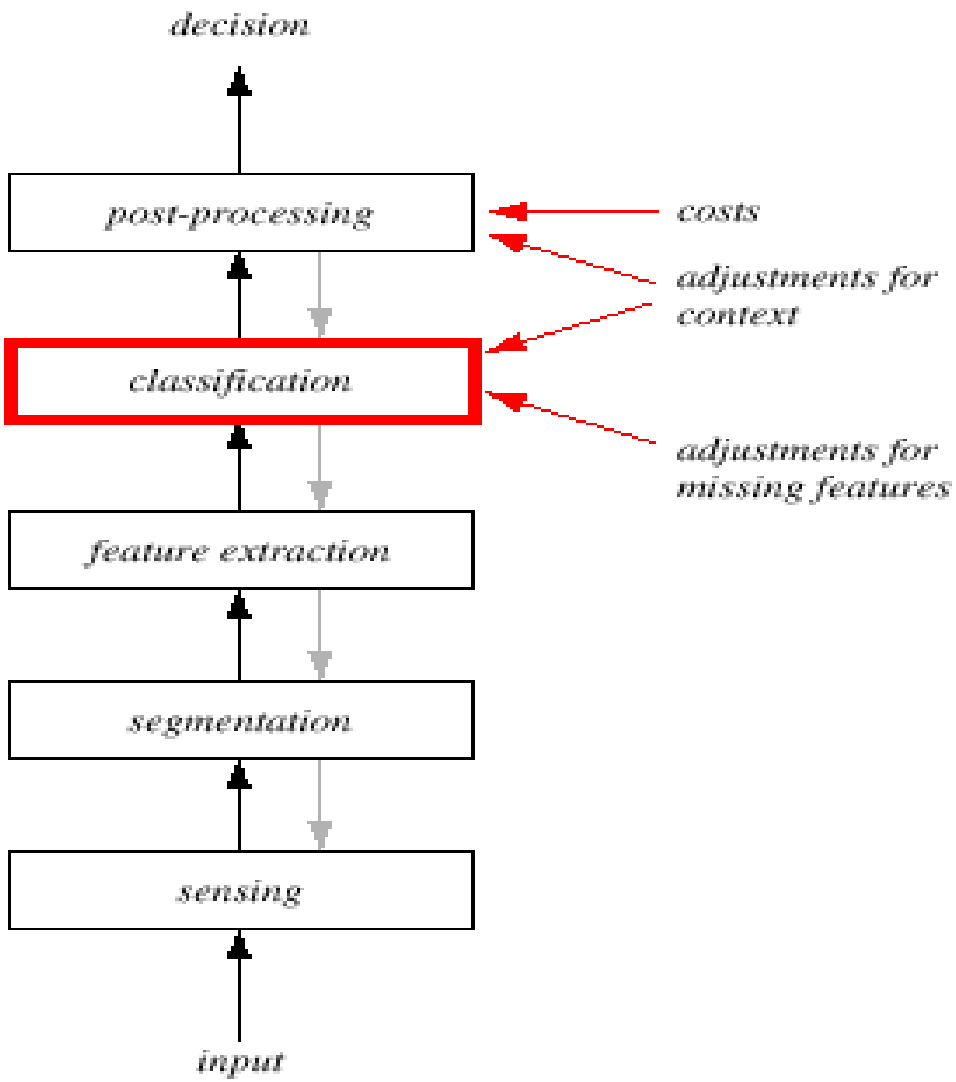
- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input

Issue of generalization!

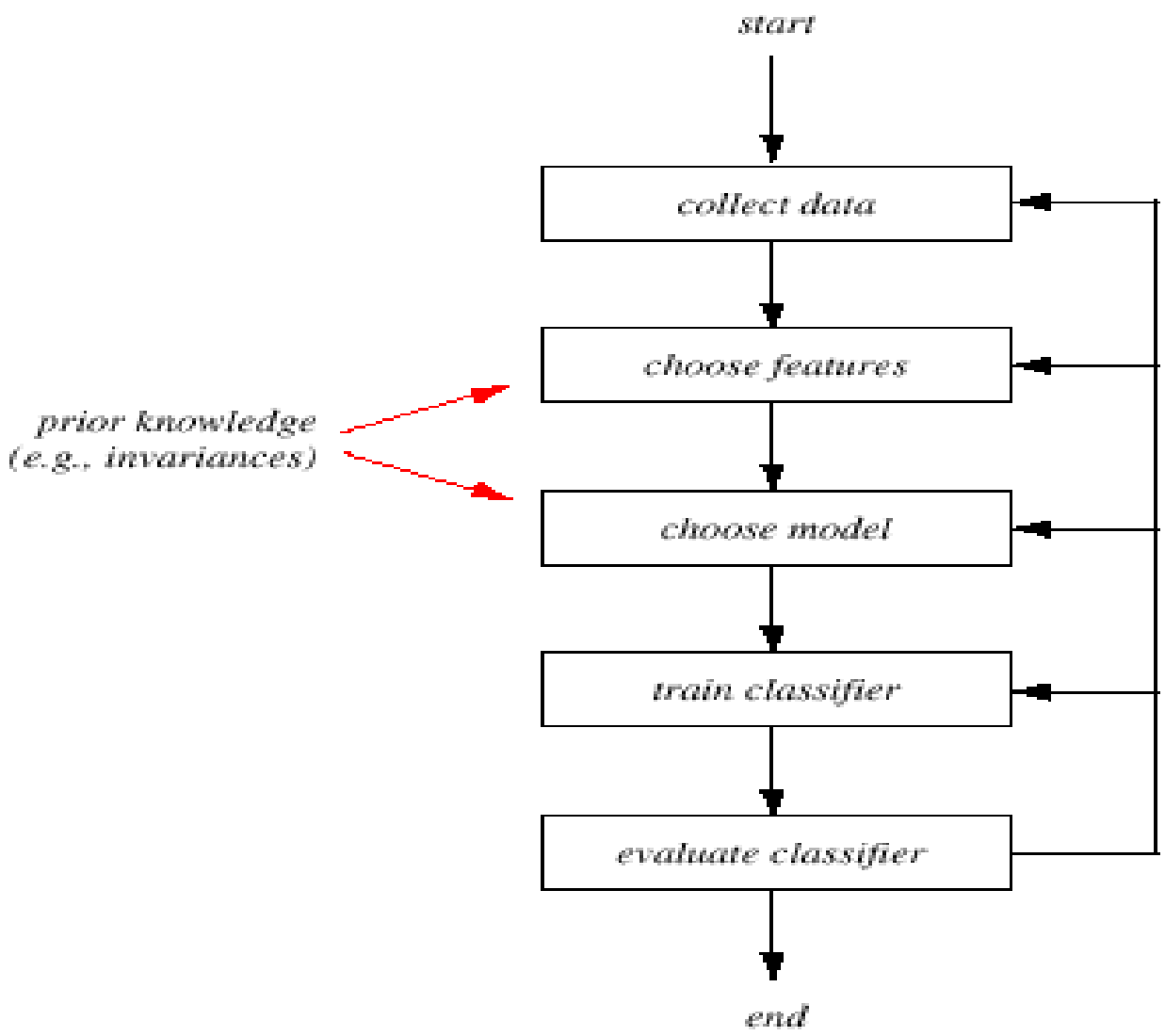

width
salmon
sea bass
22
21
20
19
18
17
16
15
14
2
4
6
8
10
lightness

Pattern Recognition Systems

- Sensing
  - Use of a transducer (camera or microphone)
  - PR system depends of the bandwidth, the resolution sensitivity distortion of the transducer
- Segmentation and grouping
  - Patterns should be well separated and should not overlap


decision
post-processing
costs
adjustments for
context
classification
adjustments for
missing features
feature extraction
segmentation
sensing
input

- Feature extraction
  - Discriminative features
  - Invariant features with respect to translation, rotation and scale.
- Classification
  - Use a feature vector provided by a feature extractor to assign the object to a category
- Post Processing Exploit context input dependent information
  - other than from the target pattern itself to improve performance


start
collect data
choose features
prior knowledge
(e.g., invariances)
choose model
train classifier
evaluate classifier
end

- Data Collection
  - How do we know when we have collected an adequately large and representative set of examples for training and testing the system?

- Feature Choice

  – Depends on the characteristics of the problem domain. Simple to extract, invariant to irrelevant transformation insensitive to noise.

- Model Choice
  - Unsatisfied with the performance of our fish classifier and want to jump to another class of model

- Training
  - Use data to determine the classifier. Many different procedures for training classifiers and choosing models

- Evaluation
  - Measure the error rate (or performance and switch from one set of features to another one

- Computational Complexity
  - What is the trade-off between computational ease and performance?
  - (How an algorithm scales as a function of the number of features, patterns or categories?)

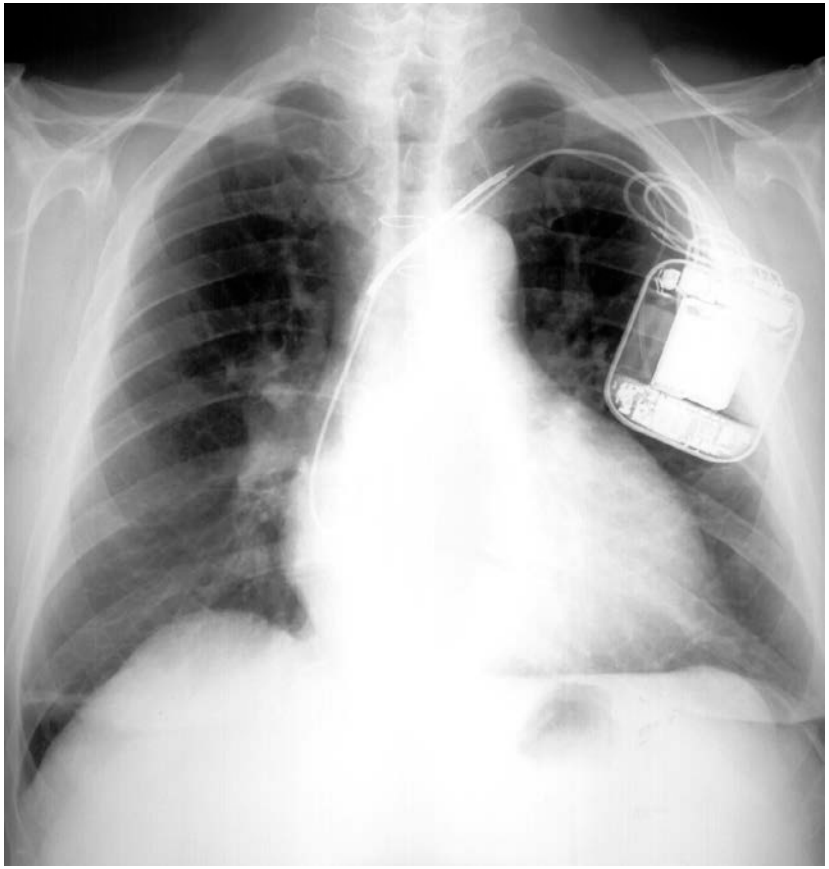
Learning and Adaptation

- Supervised learning
  - A teacher provides a category label or cost for each pattern in the training set
- Unsupervised learning
  - The system forms clusters or “natural groupings” of the input patterns

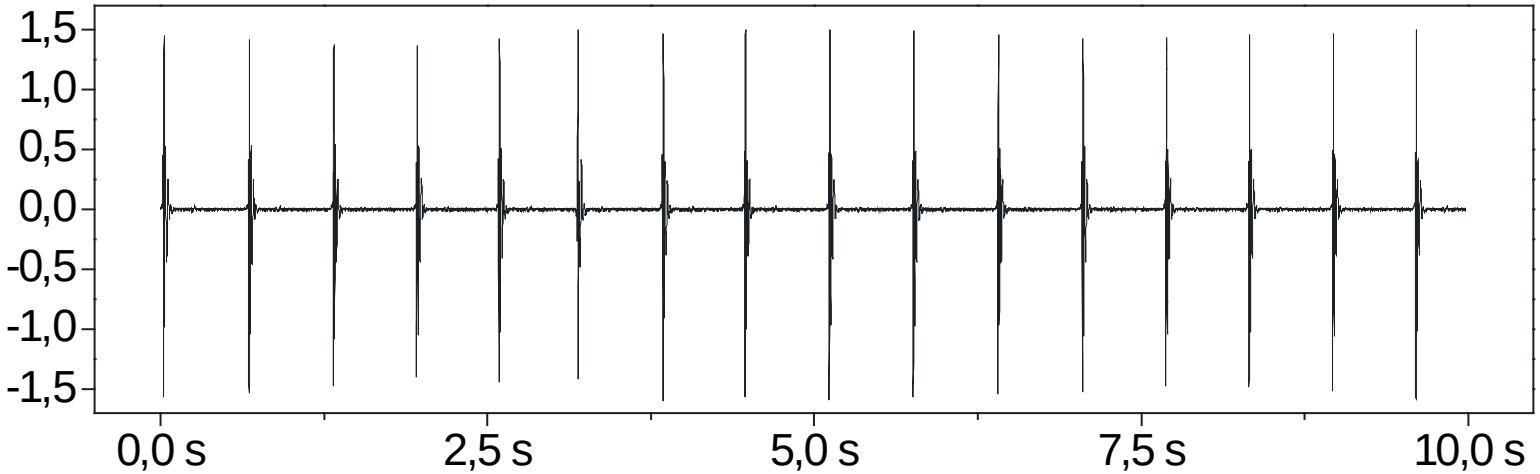
Classification of Atrial Activation Patterns

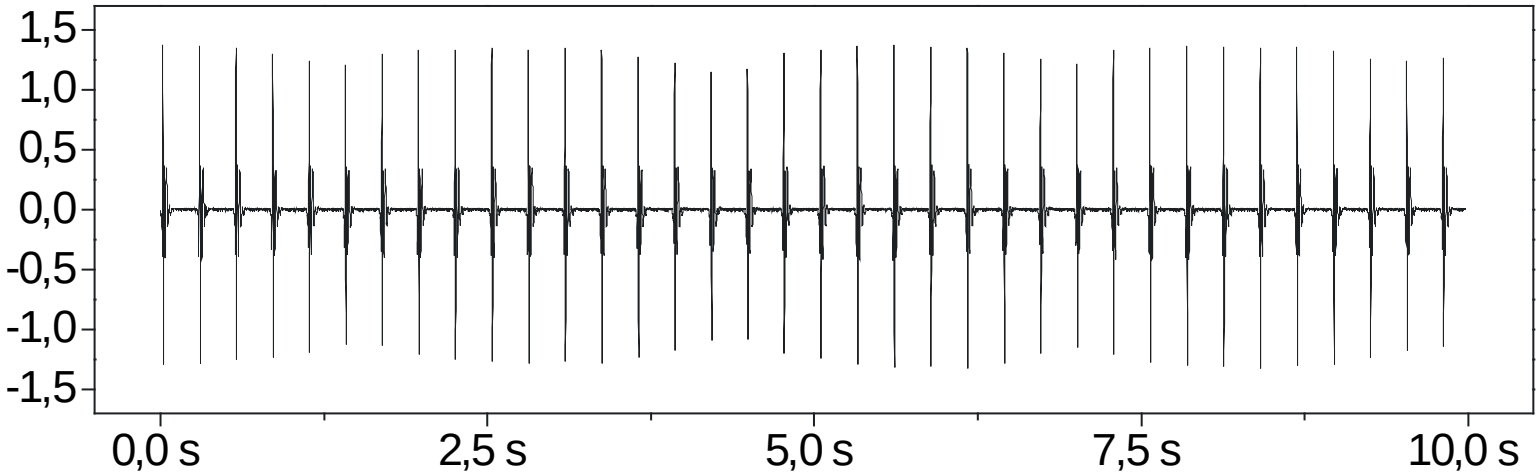
Collaboration with the Internal Medicine III, Saarland University Hospital

Dual Chamber ICD



Endocardial Electrograms





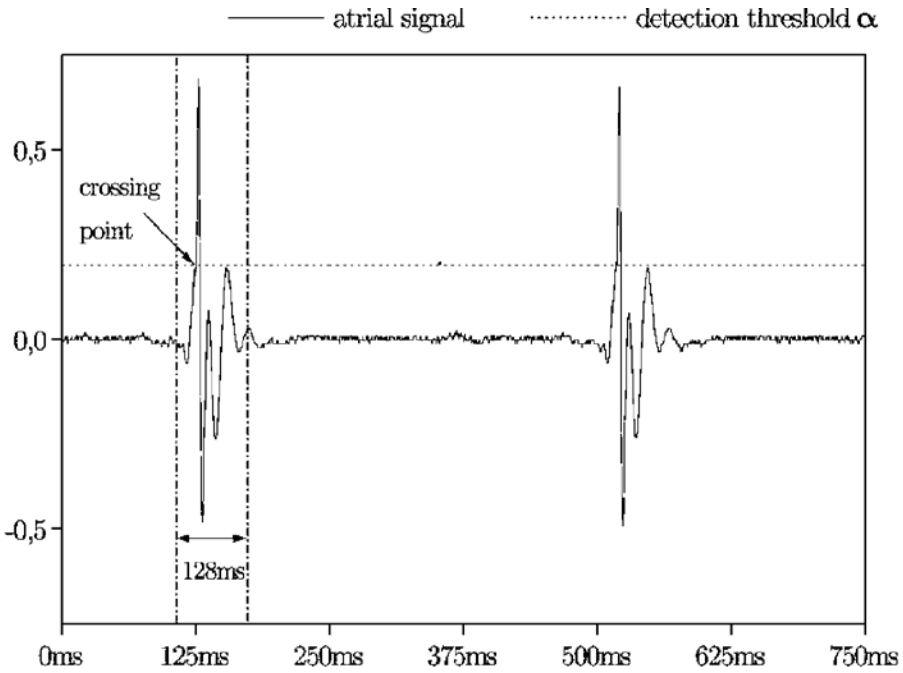
rate based algorithms

VR > RS und VR > AR -> life-threatening ventricular tachyarrhythmia

VR > RS und VR < AR -> non life-threatening atrial tachyarrhythmia

VR > RS und VR = AR -> ? sinus tachycardia or ventricular tachycardia with 1:1 retrograde conduction

Morphologische Analyse



Erkennung von atrialen Aktivierungsmustern

Gut: - Lernen mit Unterweisung ist möglich

Schlecht: - Individuelle Adaption des Verfahrens erforderlich

- Aktivierungsmuster können sich nur geringfügig unterscheiden

- Hohes Maß der Gerneralisierung erforderlich

Bekannte morphologische Detektionsverfahren

Template-Matching im Zeitbereich

Correlation Waveform Analysis (CWA),
Area of Difference (AOD),
Bin Area Method (BAM)

Jenkins-Units

MATIC

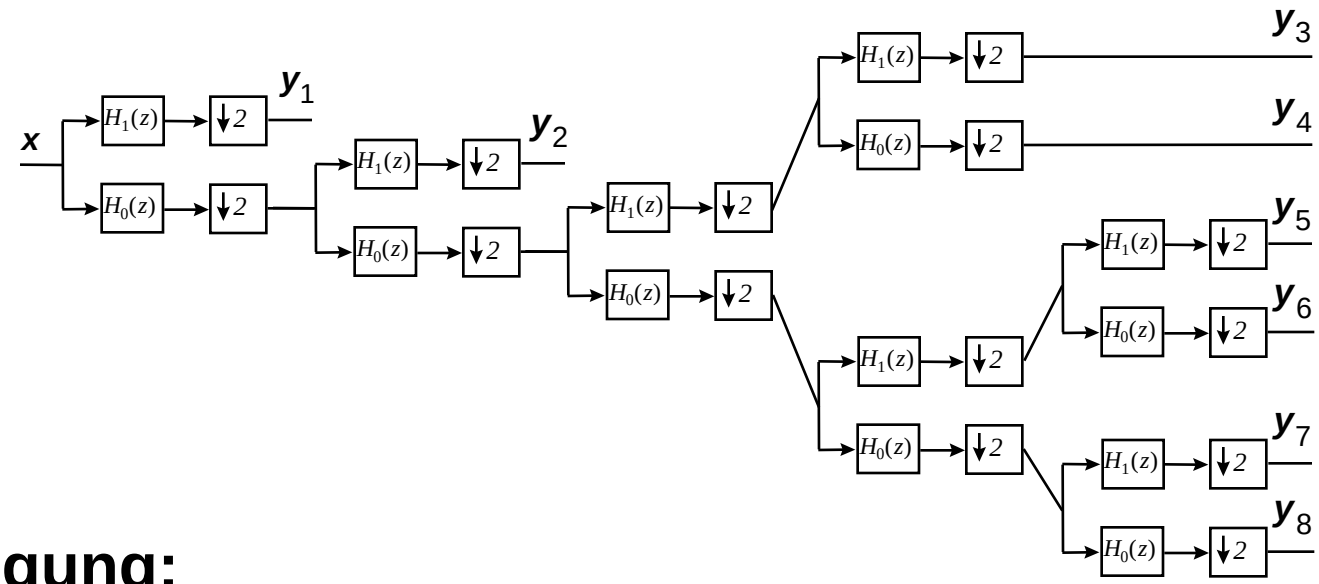
(*Morphology And Timing Intracardiac Classifier*)

Leong and Jabri

Problems: - Dimension of the problem is large
- Selection window

$$S \subseteq R^n \qquad\qquad C = \{0,1\}$$

$$f : S \to F \qquad F \subseteq R^m, \ (m \leq n) \qquad c : F \to C$$

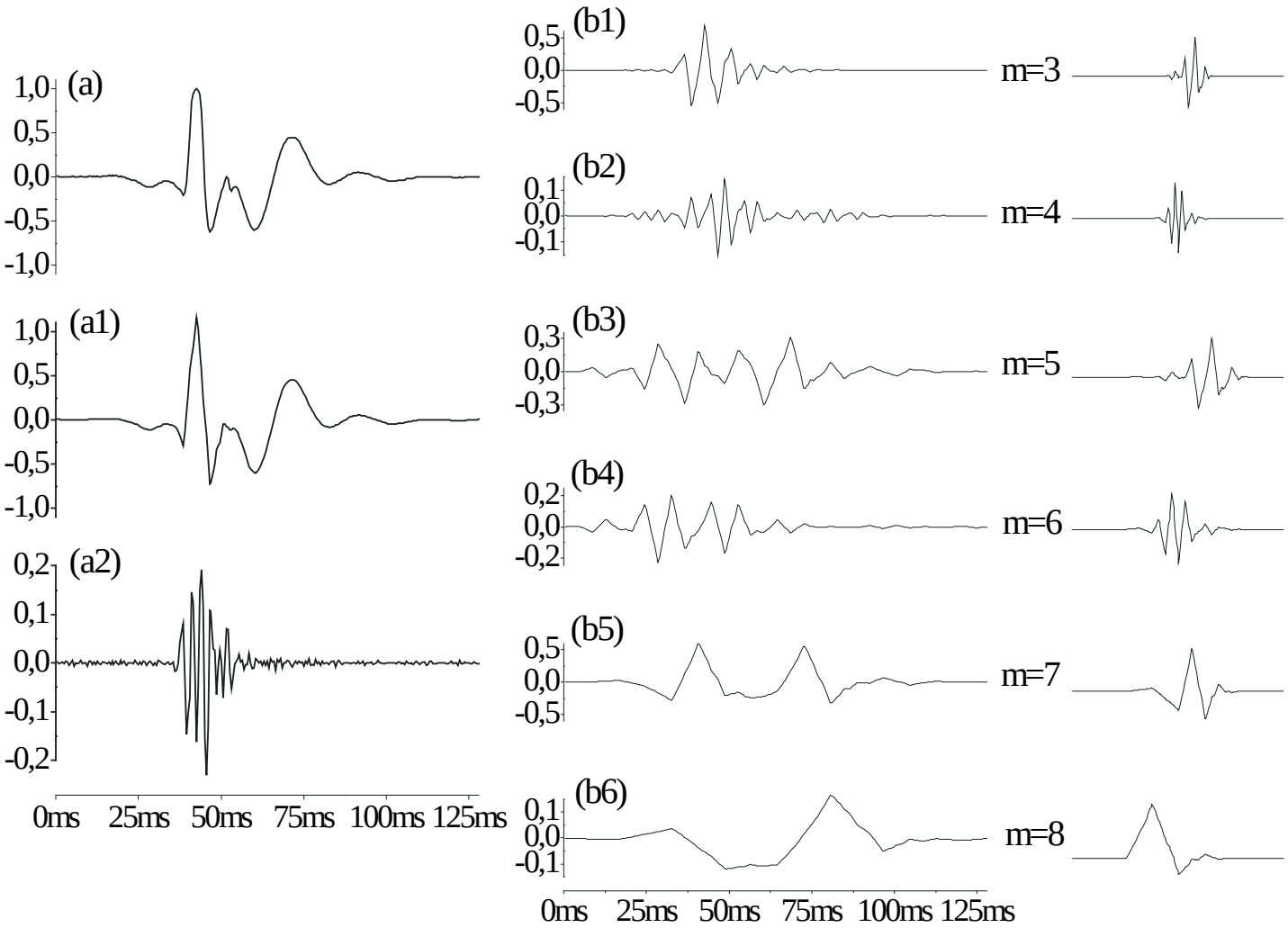


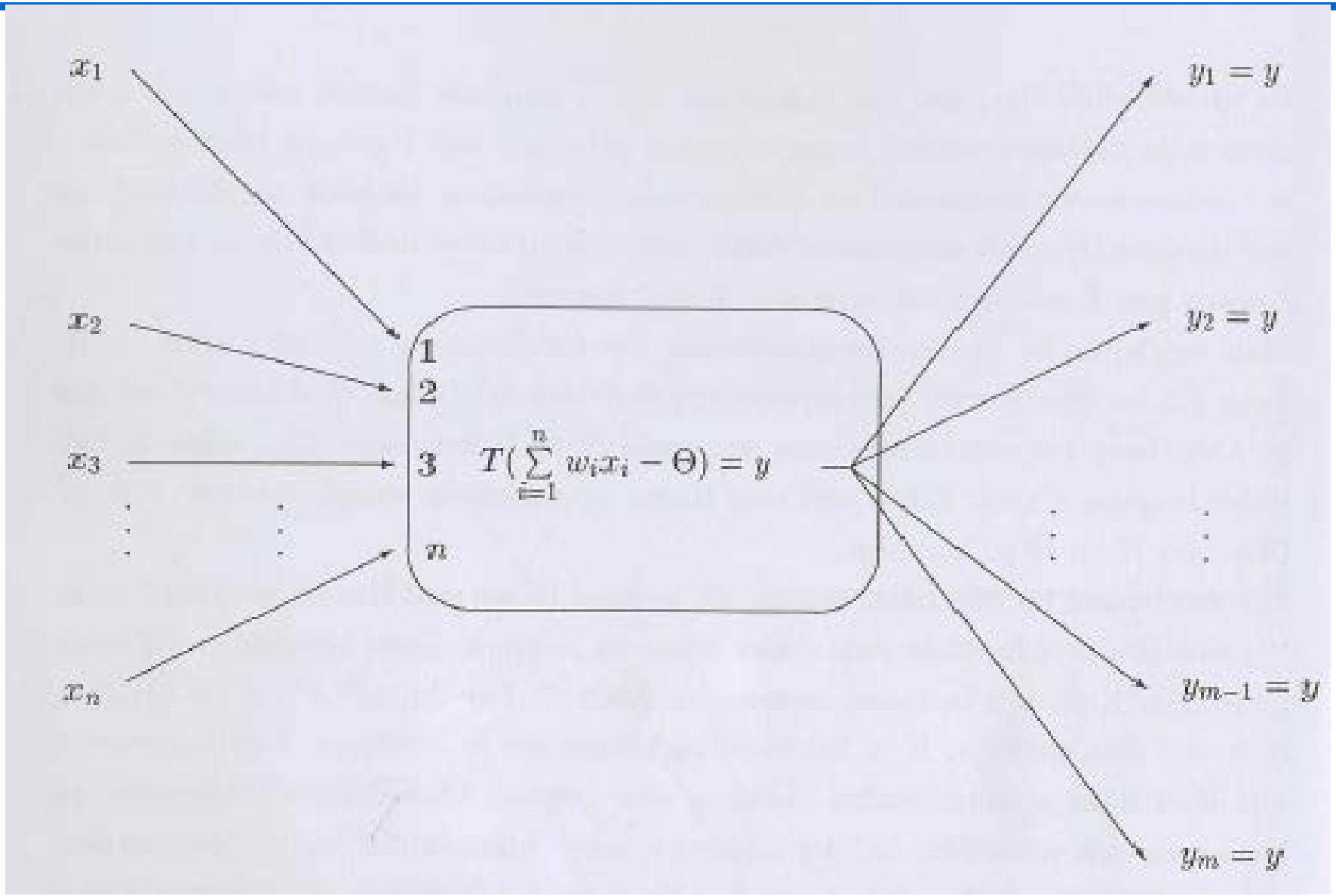
Zerlegung:

$$(\mathbf{y}_1,\ldots,\mathbf{y}_8) = (Q\mathbf{x}, QP\mathbf{x}, Q^2P^2\mathbf{x}, PQP^2\mathbf{x}, Q^2P^3\mathbf{x}, PQP^3\mathbf{x}, QP^4\mathbf{x}, P^5\mathbf{x})$$

Merkmalsvektor:

$$\square = (x_1,\ldots,x_6) := (\|\mathbf{y}_3\|_{\ell^1},\ldots,\|\mathbf{y}_8\|_{\ell^1}) \qquad \|\square\|_{\ell^1} = 1$$


(a)
1,0
0,5
0,0
-0,5
-1,0
(a1)
1,0
0,5
0,0
-0,5
-1,0
(a2)
0,2
0,1
0,0
-0,1
-0,2
0ms 25ms 50ms 75ms 100ms 125ms
(b1)
0,5
0,0
-0,5
m=3
(b2)
0,1
0,0
-0,1
m=4
(b3)
0,3
0,0
-0,3
m=5
(b4)
0,2
0,0
-0,2
m=6
(b5)
0,5
0,0
-0,5
m=7
(b6)
0,1
0,0
-0,1
m=8
0ms 25ms 50ms 75ms 100ms 125ms


x_1
x_2
x_3
x_n
1
2
3
n
$T(\sum_{i=1}^{n} w_i x_i - \Theta) = y$
$y_1 = y$
$y_2 = y$
$y_{m-1} = y$
$y_m = y$

- Die *Identität* T_I,

$$T(x) := x =: T_I(x).$$

- Die *Einschalt-, Sprung-* oder *1-Funktion* $T_{\mathbf{1}}$,

$$T(x) := \left\{ \begin{array}{lll} 0 & \text{für} & x < 0 \\ 1 & \text{für} & x \geq 0 \end{array} \right\} =: T_{\mathbf{1}}(x).$$

- Die *Fermi-Funktion* T_F,

$$T(x) := \frac{1}{1 + e^{-x}} =: T_F(x).$$

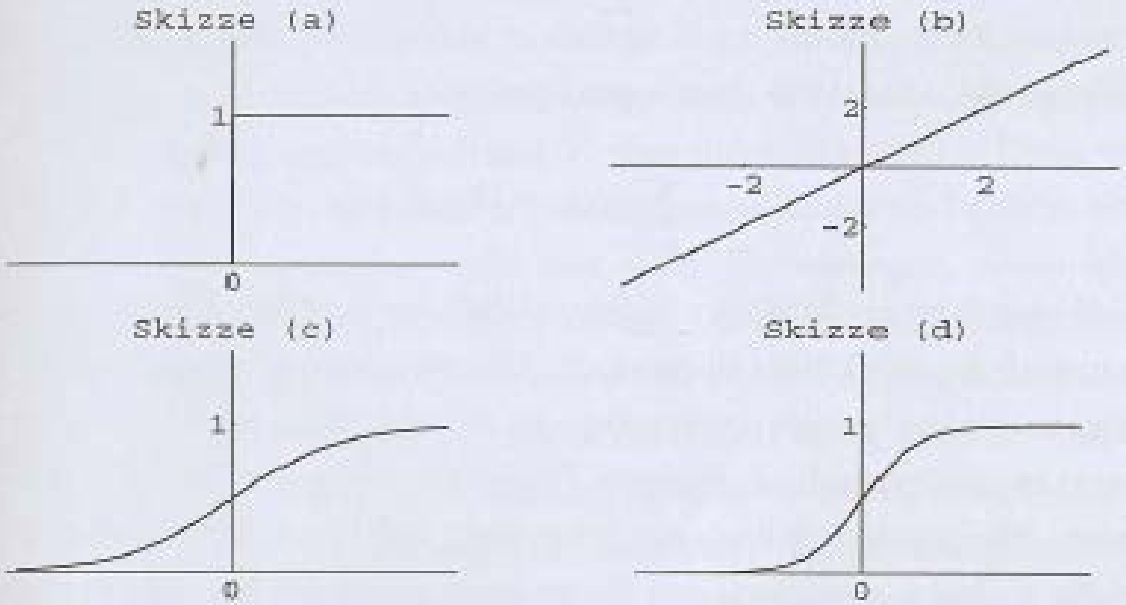
- Die *Tangenshyperbolicus-Funktion* T_H,

$$T(x) := \frac{1 + \tanh(x)}{2} =: T_H(x).$$

- Alle beschränkten Funktionen $T : \mathbb{R} \to \mathbb{R}$ mit

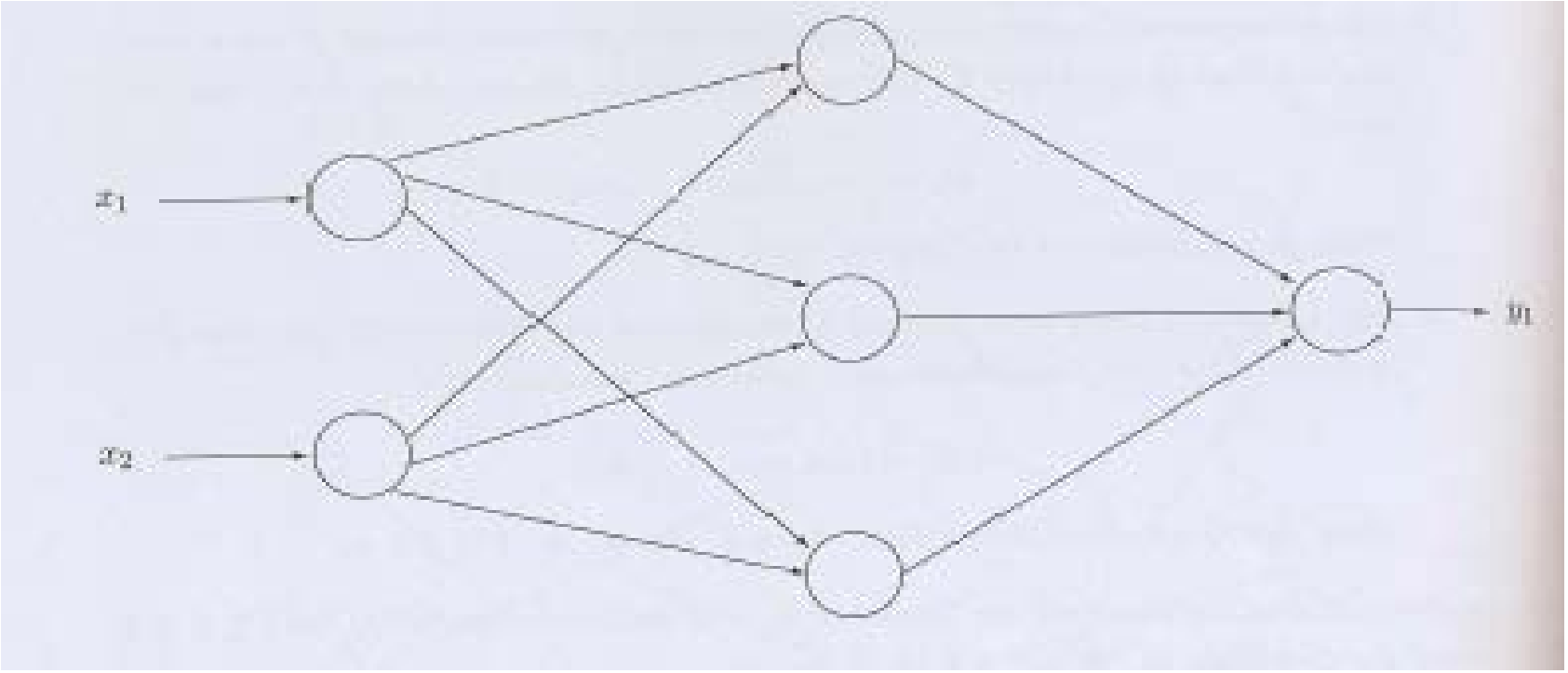
$$\lim_{x \to -\infty} T(x) = 0 \quad \text{und} \quad \lim_{x \to \infty} T(x) = 1,$$

die sogenannten *S-förmigen* oder *sigmoidalen Funktionen*.



Skizze (a): $T_{\perp}$ Skizze (b): T_I

Skizze (c): T_F Skizze (d): T_H


x_1
x_2
y_1

Classifier

Feed-Forward Neural Network

Input-Layer: $N^{(I)} = \dim F$

Output-Layer: $N^{(O)} = 1$

Hidden-Layer: $N^{(H)} = 4$

$$r = \sum_{p=1}^{N^{(H)}} w_p^{(O)} T\left(\sum_{i=1}^{N^{(I)}} w_{ip}^{(H)} x_i (\square) - t_p \right)$$

FFNN Learning

Minimization of

$$\sum_{x^{ant}_{ret} \in A^{ant}_{ret}} \left(r^{ant}_{ret} - \sum_{p=1}^{N^{(H)}} w_p^{(O)} T\left(\sum_{i=1}^{N^{(I)}} w_{ip}^{(H)} x_i^{ant}(\square) - t_p \right) \right)^2$$

Backpropagation-Algorithm

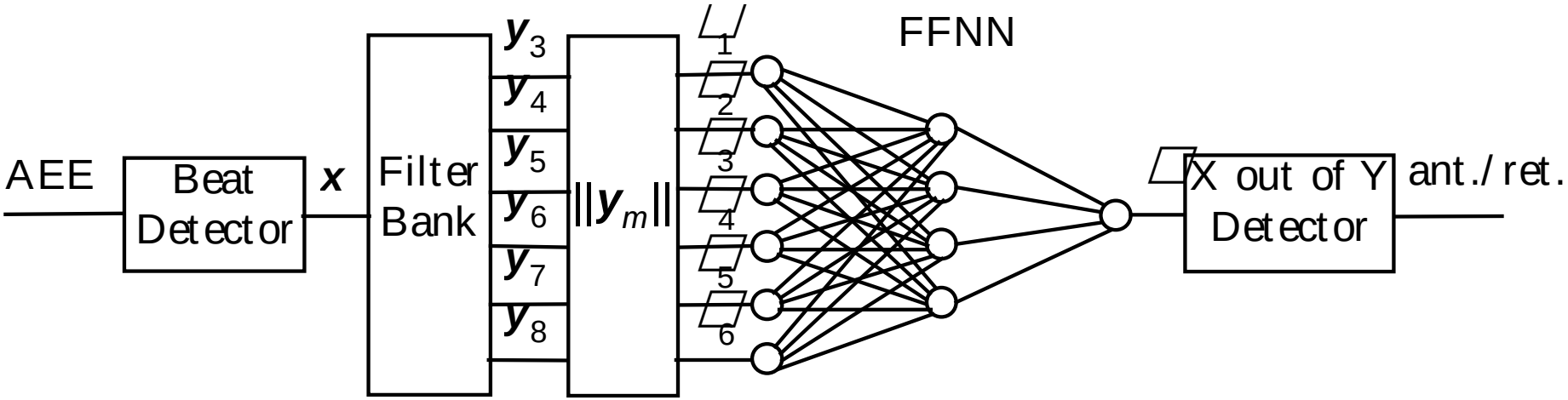
$$r^{ant} = 0 \quad \text{und} \quad r^{ret} = 1$$

Threshold: $k = 0.5$

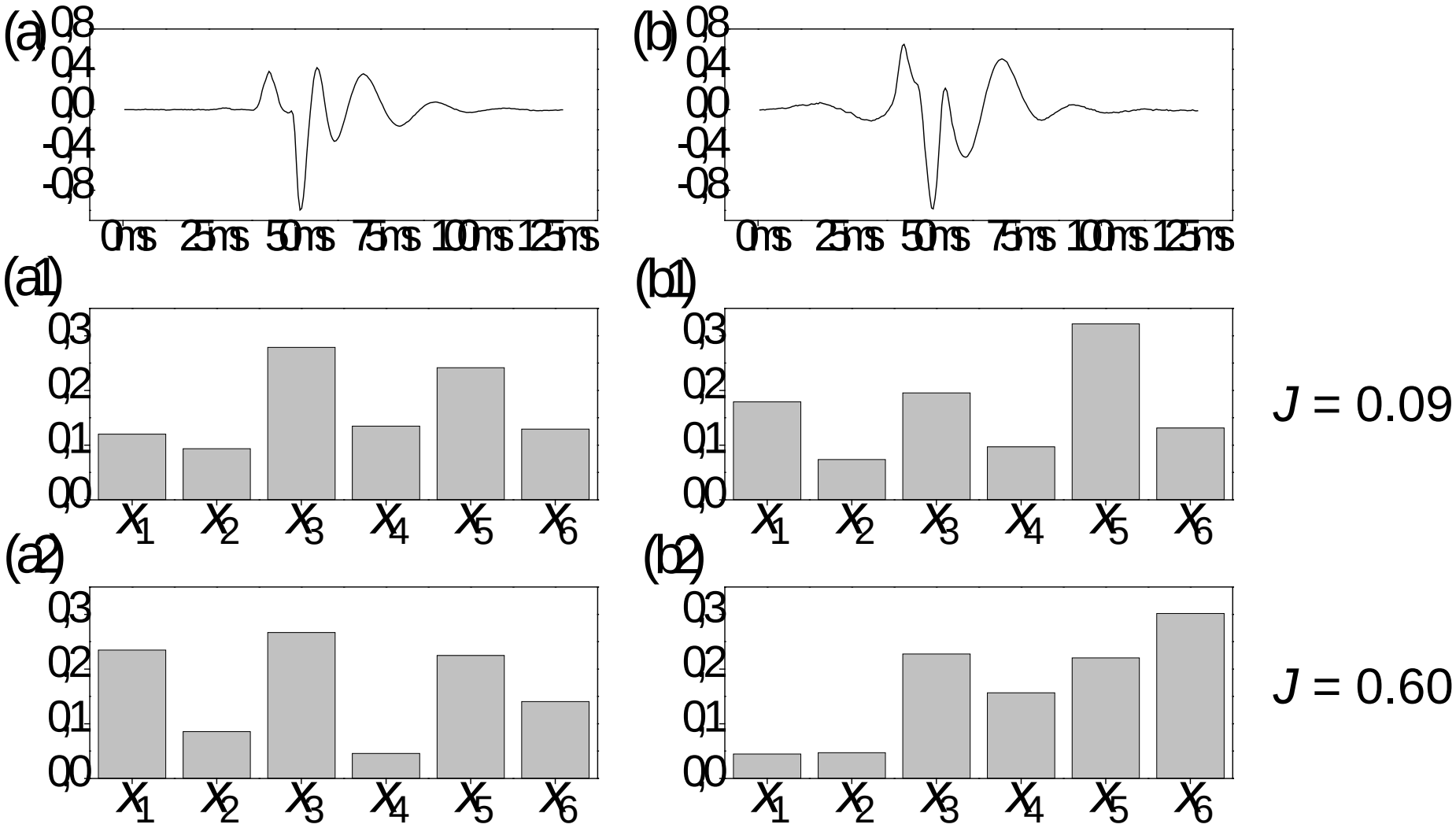
$k < 0.5$ = antegrade activation -> physiological

$k \geq 0.5$ = retrograde activation -> pathological

Hybrid System:



Eigenschaftsextraktion



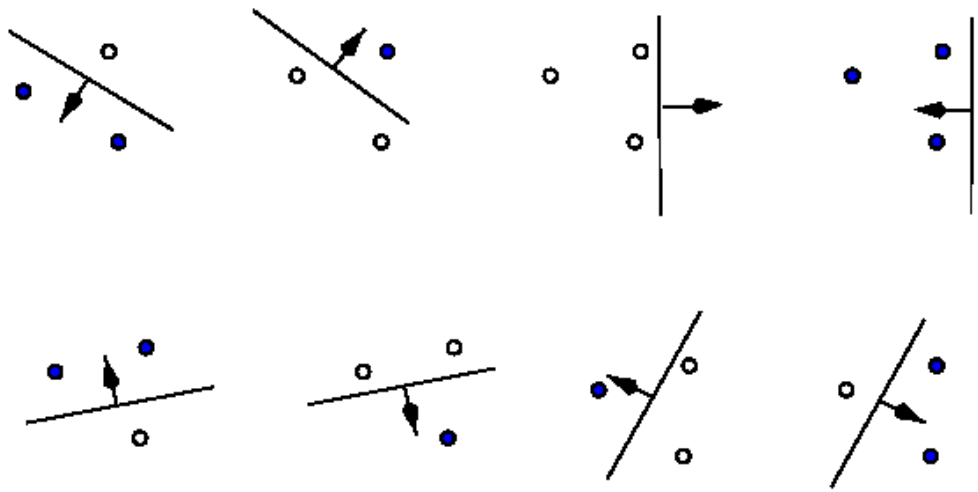
30 Patienten

	ρ_{min}	ρ_{max}	ρ_{mean}	SD	Activations
Antegrade:	−0.45	0.48	0.03	0.09	726
Retrograde:	0.51	1.54	0.98	0.13	678

TABLE 2. $\rho_{mean} \pm$ SD for sinus rhythm at BL and after AS.

	P1	P6	P9	P30
CLR	40%	51%	31%	36%
BL	0.02±0.06	0.01±0.03	0.03±0.05	0.03±0.05
AS	−0.01±0.06	0.04±0.03	0.05±0.09	0.05±0.07

CLR: cycle length reduction
AS : adrenergic stimulation (**adrenerge Stimulation**)
BL : base line



2.1. The VC Dimension

The VC dimension is a property of a set of functions $\{f(\alpha)\}$ (again, we use α as a generic set of parameters: a choice of α specifies a particular function), and can be defined for various classes of function f. Here we will only consider functions that correspond to the two-class pattern recognition case, so that $f(\mathbf{x}, \alpha) \in \{-1, 1\}\ \forall \mathbf{x}, \alpha$. Now if a given set of l points can be labeled in all possible 2^l ways, and for each labeling, a member of the set $\{f(\alpha)\}$ can be found which correctly assigns those labels, we say that that set of points is *shattered* by that set of functions. The VC dimension for the set of functions $\{f(\alpha)\}$ is defined as the maximum number of training points that can be shattered by $\{f(\alpha)\}$. Note that, if the VC dimension is h, then there exists at least one set of h points that can be shattered, but it in general it will not be true that *every* set of h points can be shattered.

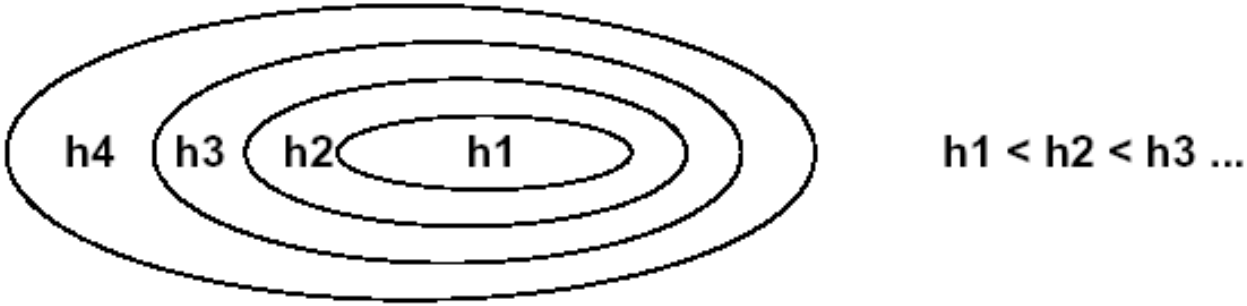
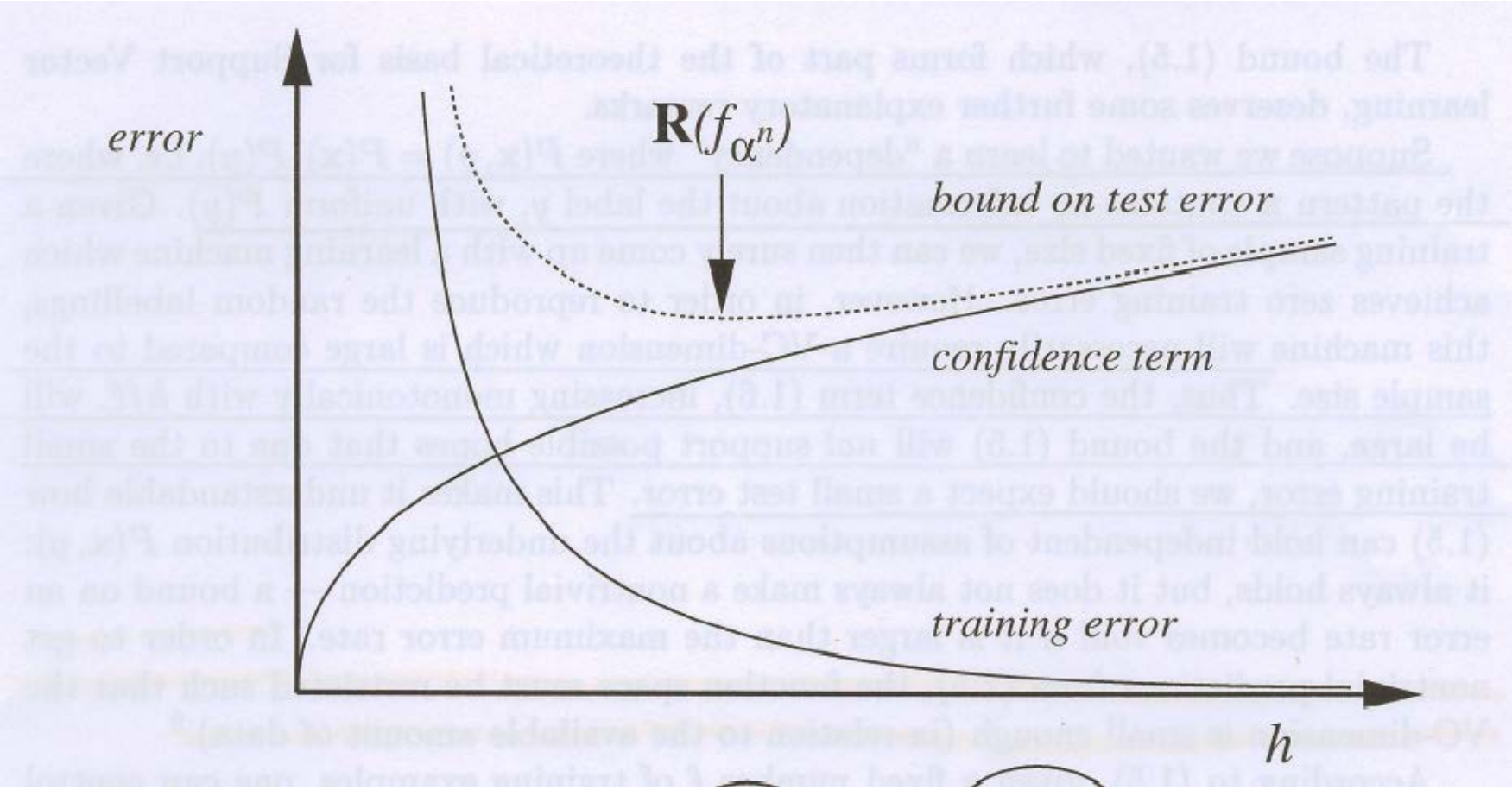


Figure 4. Nested subsets of functions, ordered by VC dimension.

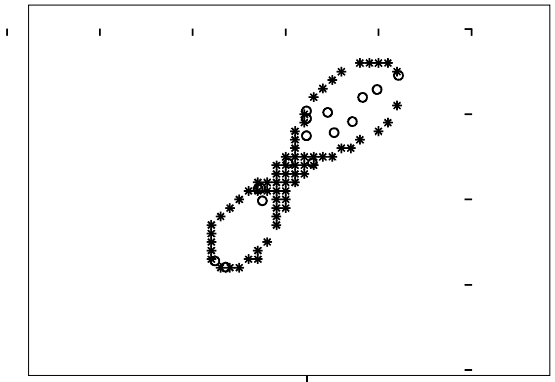

error
$\mathbf{R}(f_{\alpha^n})$
bound on test error
confidence term
training error
h

Kernel Machines

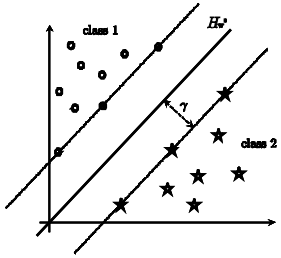
We are given a set of associations

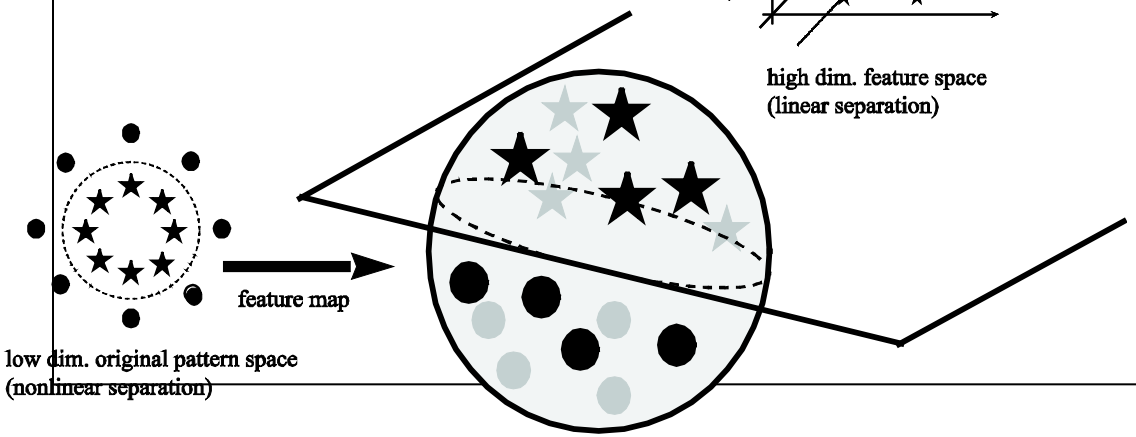
$$A = \{(\mathbf{x}_i, y_i) \in IR^d \times \{-1,1\} : i = 1, \ldots, M\}$$

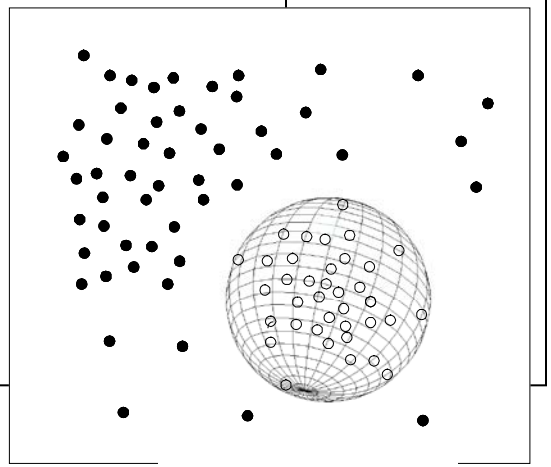
Novelty Detection



Binary Classification







Feature Map & Space

Let us introduce the **feature map** $\Phi : X \subset IR^n \to \ell^2$ by

$$\Phi(\cdot) = \left(\sqrt{\eta_i}\,\alpha(\cdot)\right)_{i \in IN}$$

which is induced by a kernel *K* of a reproducing kernel Hilbert space

$$K(\mathbf{x},\mathbf{y}) = \sum_{i \in IN} \eta_i \alpha(\mathbf{x}) \alpha(\mathbf{y}) \qquad \| \Phi(\mathbf{x}) \|_{\ell^2}^2 = \sum_{i \in IN} \eta_i \alpha_i(\mathbf{x}) = K(\mathbf{x},\mathbf{x})$$

We define the **feature space** $F_K \subset \ell^2$ by

$$F_K = \overline{\operatorname{span}\{\Phi(\mathbf{x}) : \mathbf{x} \in X\}}$$

We do not know the map Φ
we only know the reproducing kernel

Optimization Problems

Classification:

$$\hat{\vartheta} = \arg\max_{\vartheta \in P^K} \left\{ \min_{i \in M_+, j \in M_-} \left\| \Phi(\zeta_i(\vartheta)) - \Phi(\zeta_j(\vartheta)) \right\|_{F_K}^2 \right\}$$

large margin criterion

This can be simplified to (Strauss and Steidl, JCAM, 2002):

$$\hat{\vartheta} = \arg\max_{\vartheta \in P^K} \left\{ \left\| \bar{\zeta}_i(\vartheta) - \bar{\zeta}_j(\vartheta) \right\|_2^2 \right\}$$

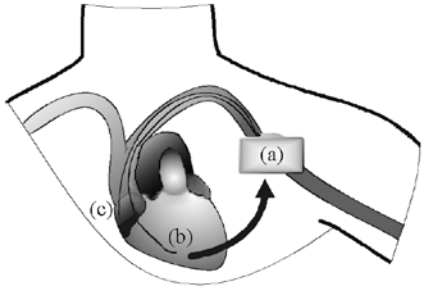
Novelty Detection:

$$\hat{\vartheta} = \arg\min_{\vartheta \in P^K} \left\{ \sum_{i \in M_+} \left\| \Phi(\zeta_i(\vartheta)) - \Phi(\bar{\zeta}(\vartheta)) \right\|_{F_K}^2 \right\}$$

clustering in feature spaces

Applications

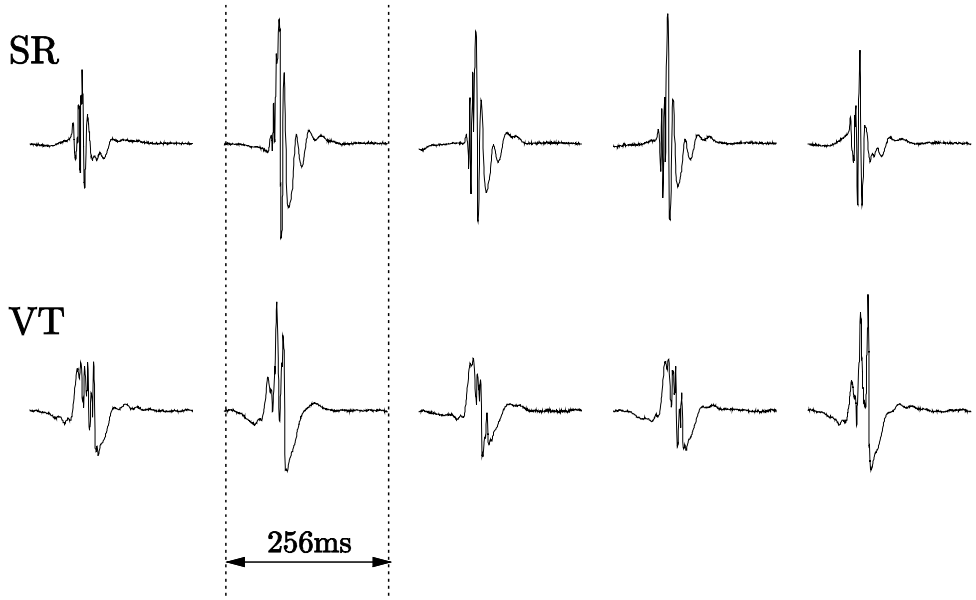
Classification of ventricular endocardial electrograms



Detection of otoacoustic emissions by novelty detection

Endocardial Signal Classification

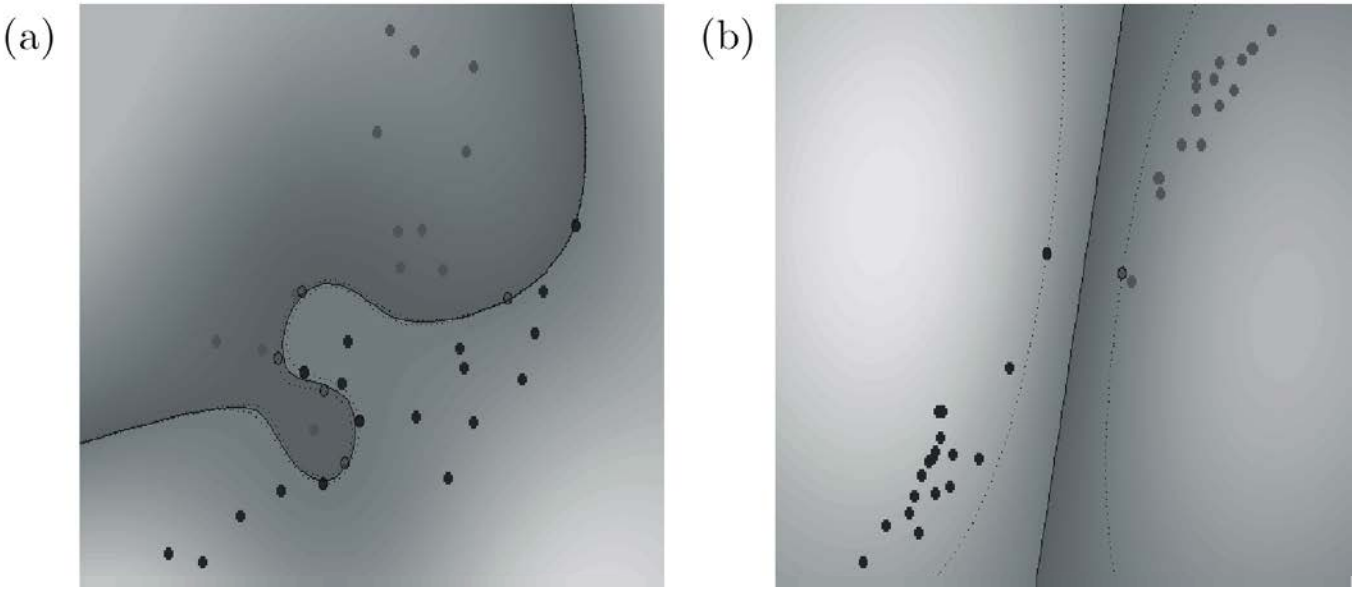
- Currently used rate based algorithms in implantable cardioverter-defibrillators (ICDs) fail in some clinical situation
- A rate-independent morphological recognition scheme may help to overcome this problem



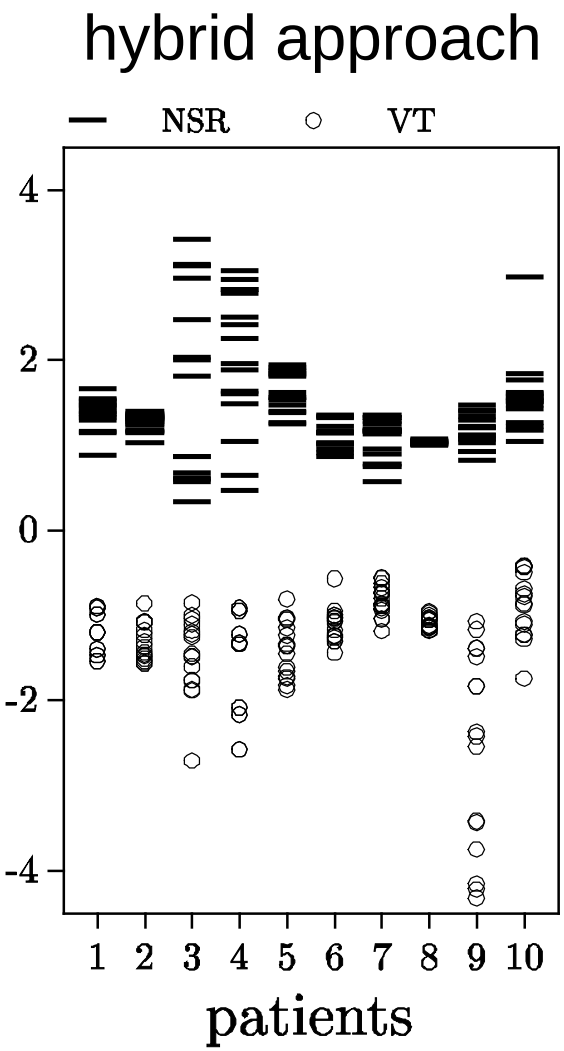
Decision Function

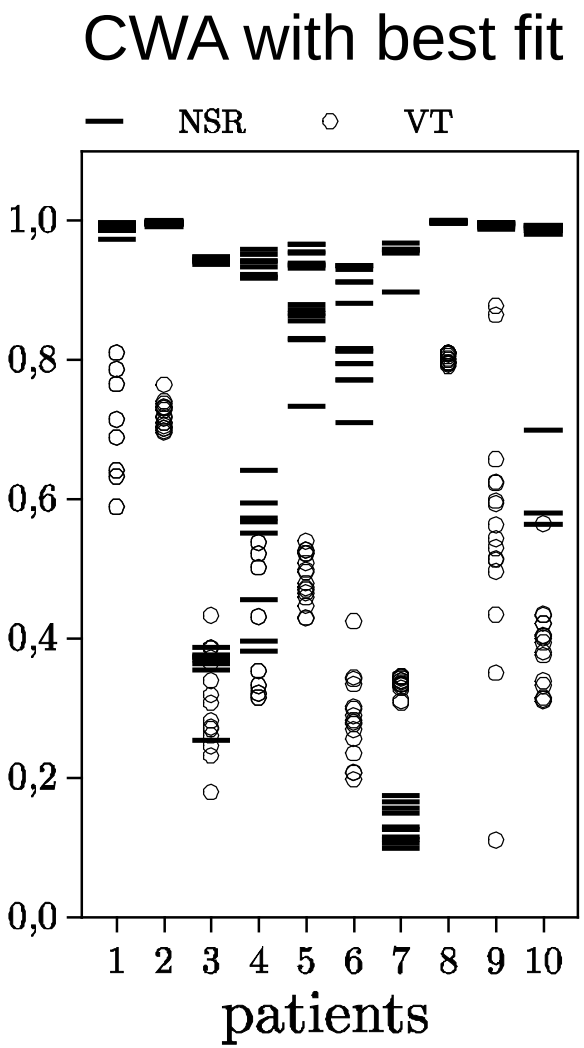
non-adapted

adapted



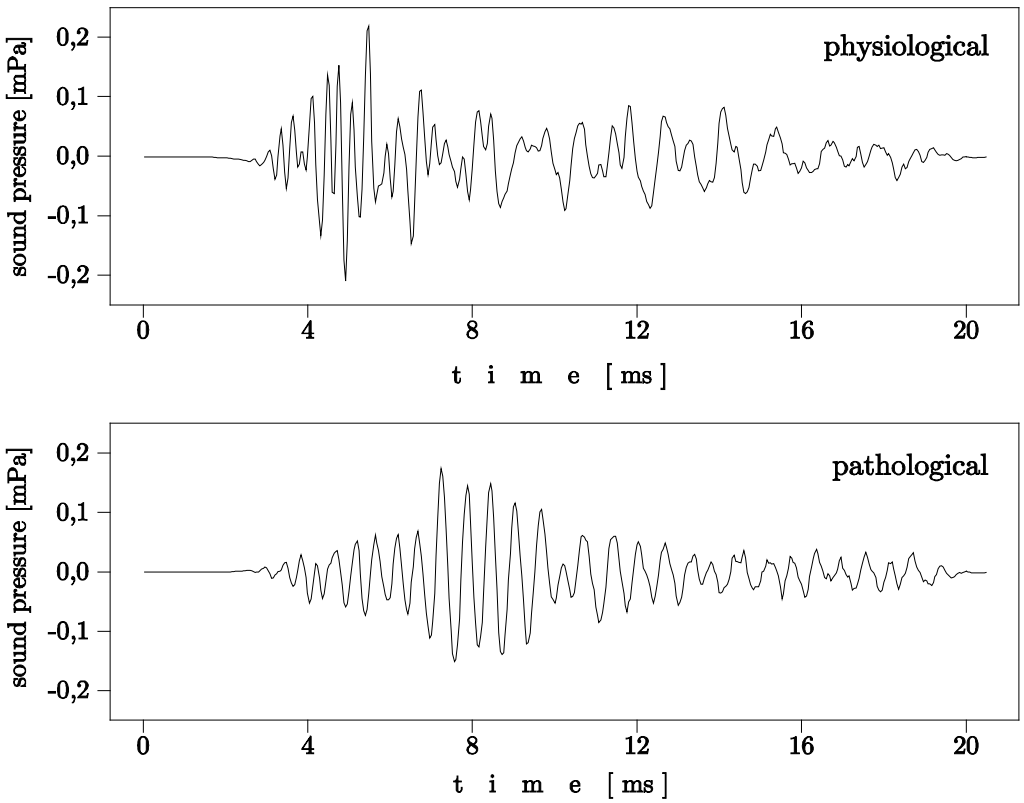
Performance



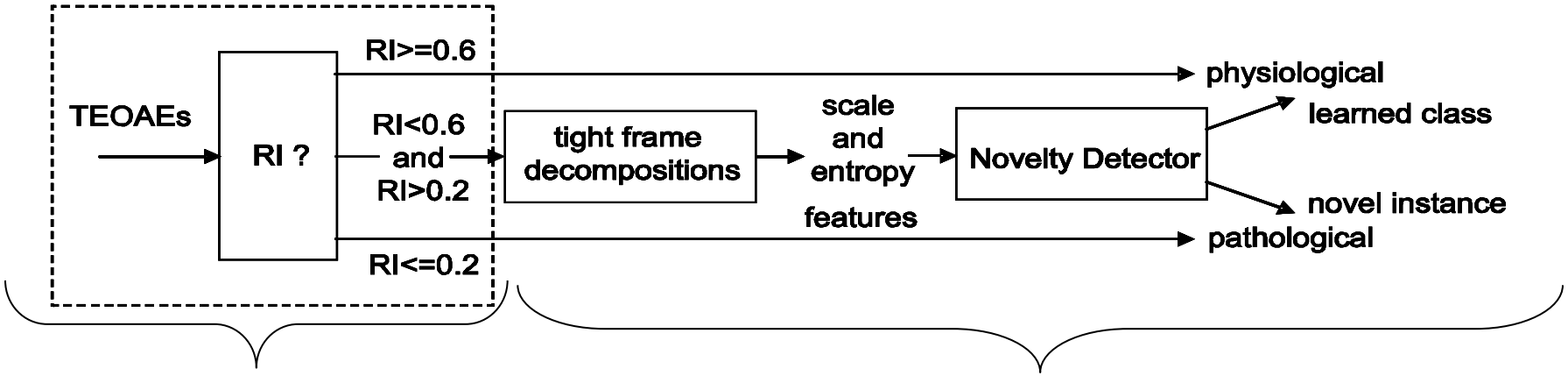


Detection of Otoacoustic Emissions

- The sensitivity of audiological screening procedures for the detection of hearing losses can be considered to be 100%
- A relatively low specificity may require a large number of repetitive measurements and thus cost in universal newborn screening programs



The Detection Scheme



conventional analysis of the reproducibility index (RI)

machine learning stage

Results:

learning associations: 1743 normal measurements (verified TEOAEs)
test set: 170 measurements (140 physiological, 30 pathological)

conventional screening test: 92 % specificity
the hybrid scheme: 96 % specificity