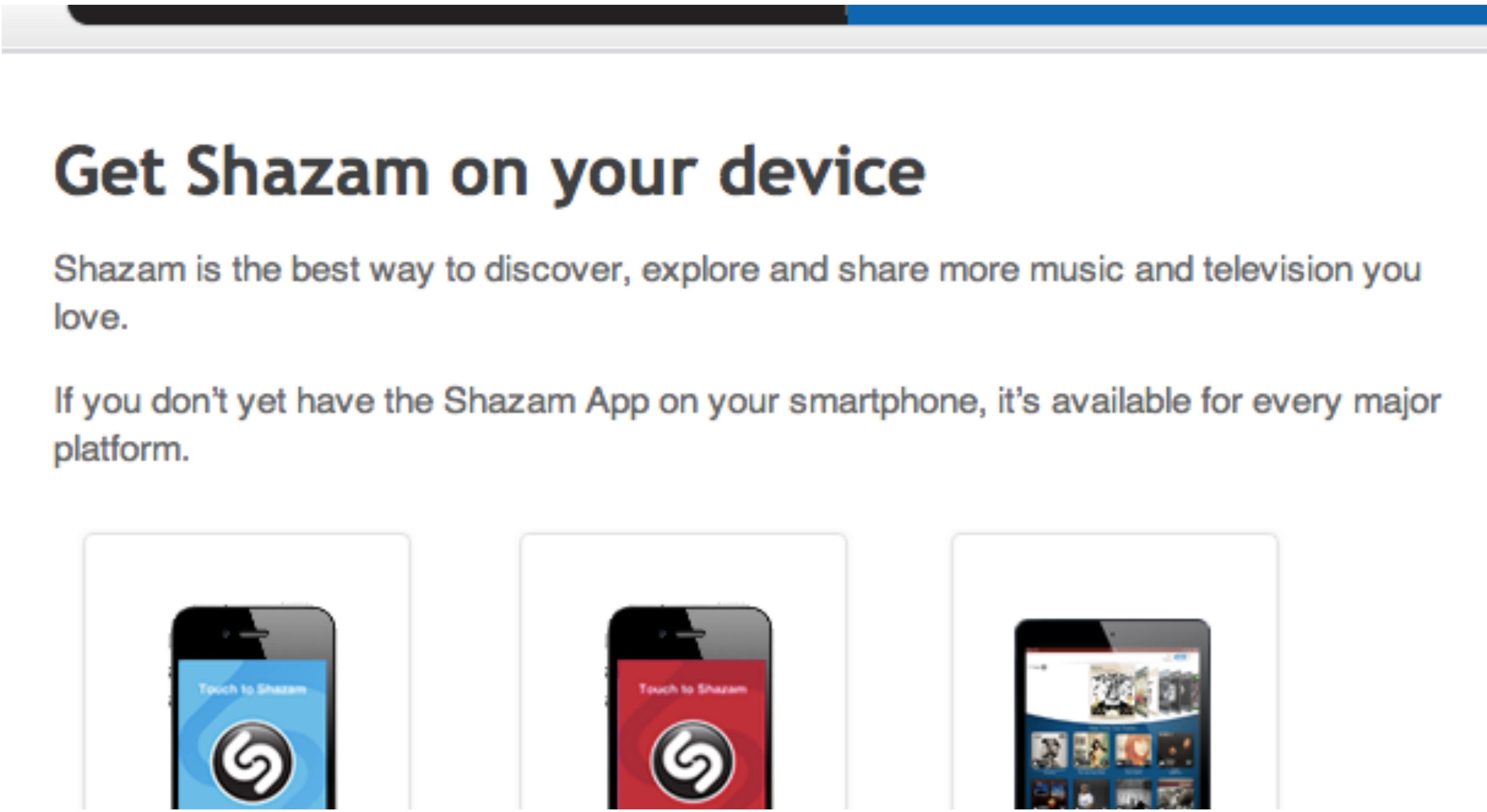
Patterns
and
Pattern Recognition

Examples

Music recognition



MelApp

OTCBB: HDVY.OB

Smart Health Applications from Health Discovery Corporation

HOME | ABOUT HDC | PHYSICIANS | CONTACT

How it Works

health discovery corporation

Melanoma is the fastest growing cancer worldwide and the most deadly of all skin cancers, if not caught early. Performing regular self-exams could save your life or that of a loved one.

Checking a mole or freckle is quick and easy:

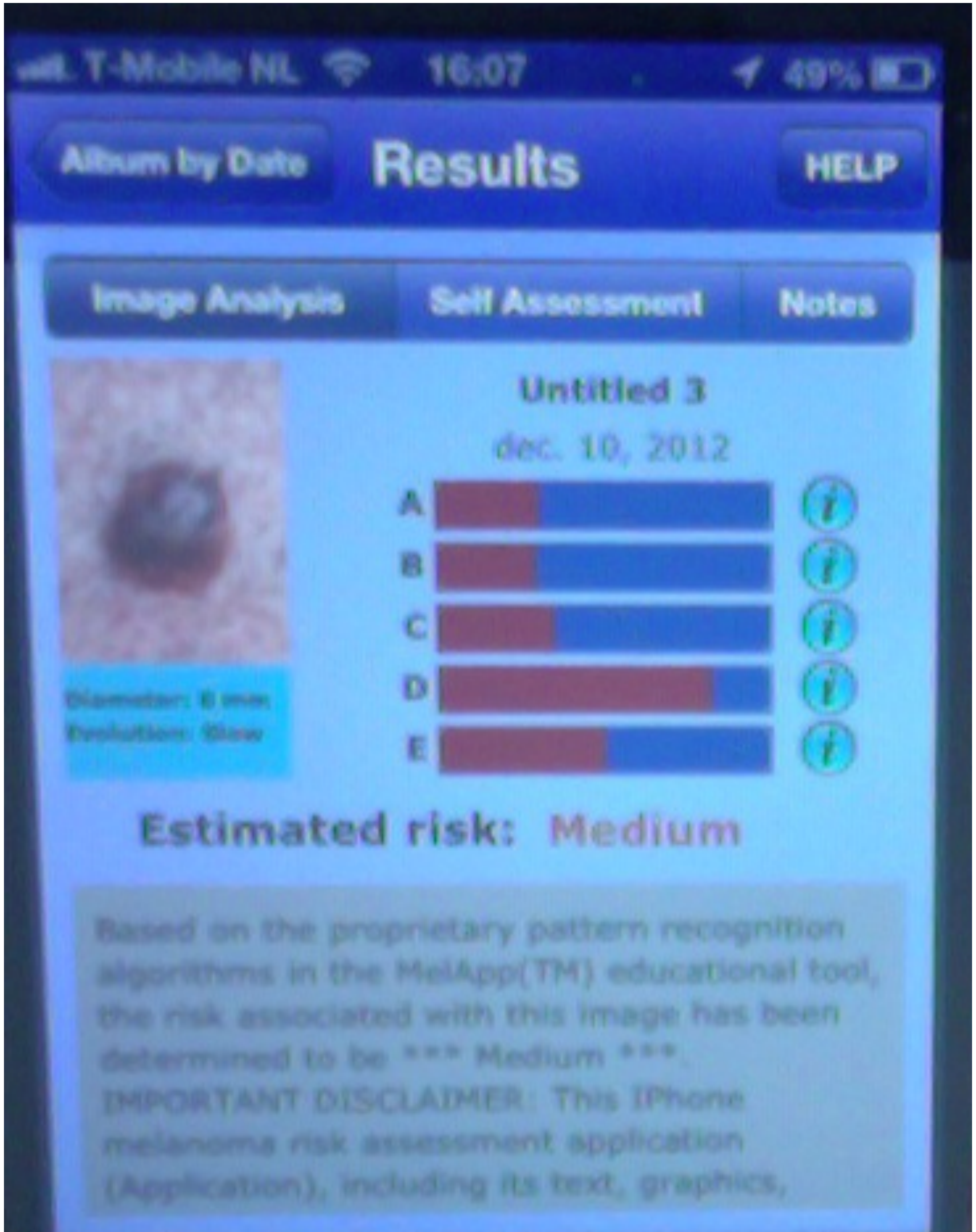
1. Use MelApp to take a picture of the skin lesions of concern with a smartphone's camera, enlarging it with the zoom feature to fit into the green box, then

2. Pin point the mole size and its evolution by sliding the corresponding indicator bar and tap on "Check Risk." Within seconds MelApp will provide a risk analysis of the uploaded picture being a melanoma.

MelApp uses highly sophisticated patent protected state-of-the-art mathematical algorithms and image based pattern recognition technology to analyze the uploaded image. The app was validated using DermAtlas, an open access, physician-edited database of over 10,000 high quality histological and clinical images of skin conditions.

These pictures also can be stored on MelApp and saved according to date, label or risk. Archiving your pictures lets you review them for changes in the skin lesions occurring over time.

The concept of feature vector



Five numbers are extracted or provided that characterize the mole:

A – asymmetry

B – fractalness of the boundary

C – color (black is worse)

D – diameter

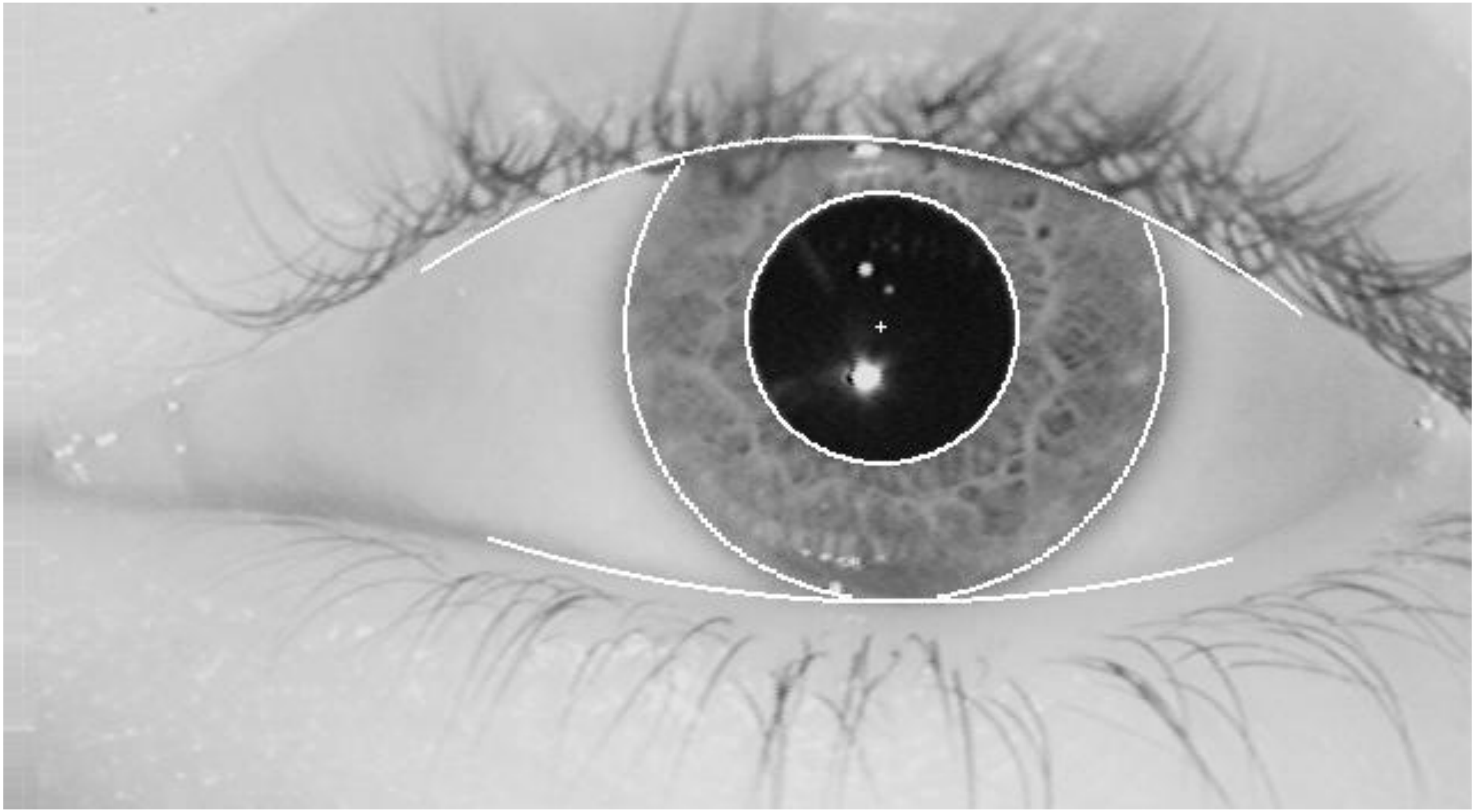
E – evolution (speed of change)

We call these characteristics 'features' and put their values in a 'feature vector' (A,B,C,D,E)

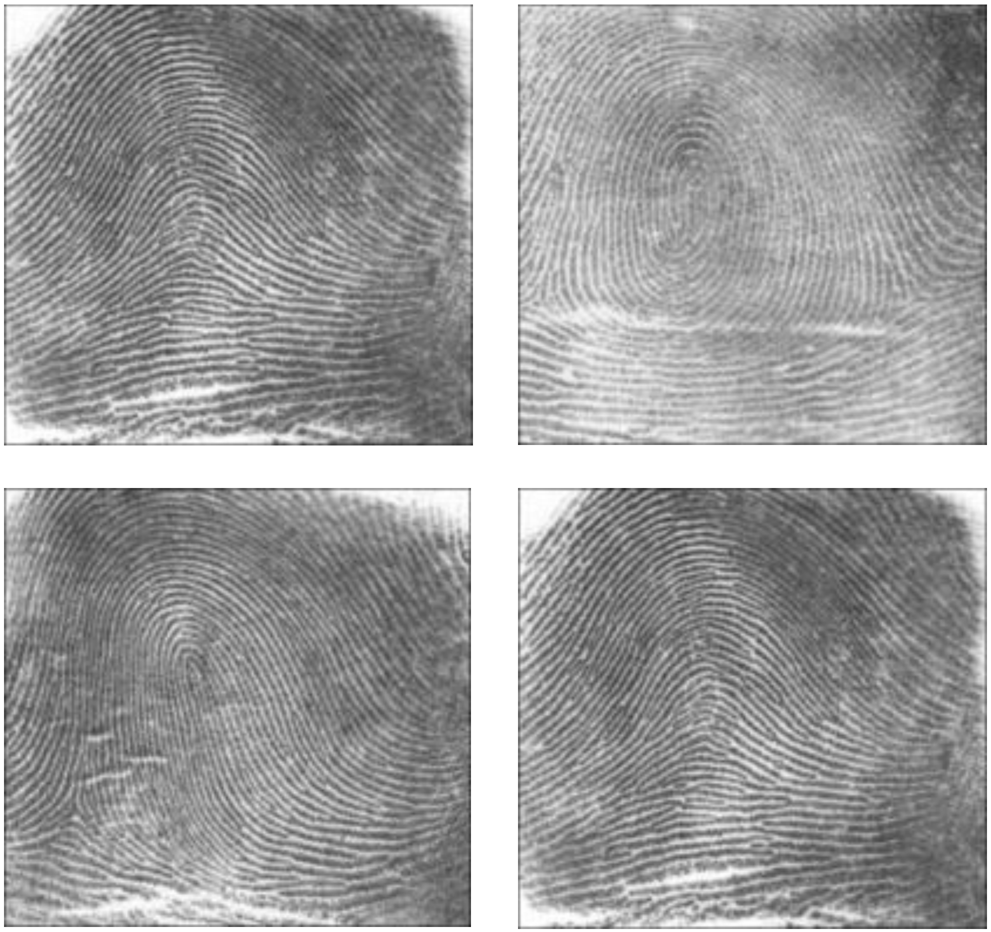
The pattern to be classified is thus represented by a feature vector.

Product recognition / identification

Iris authentication

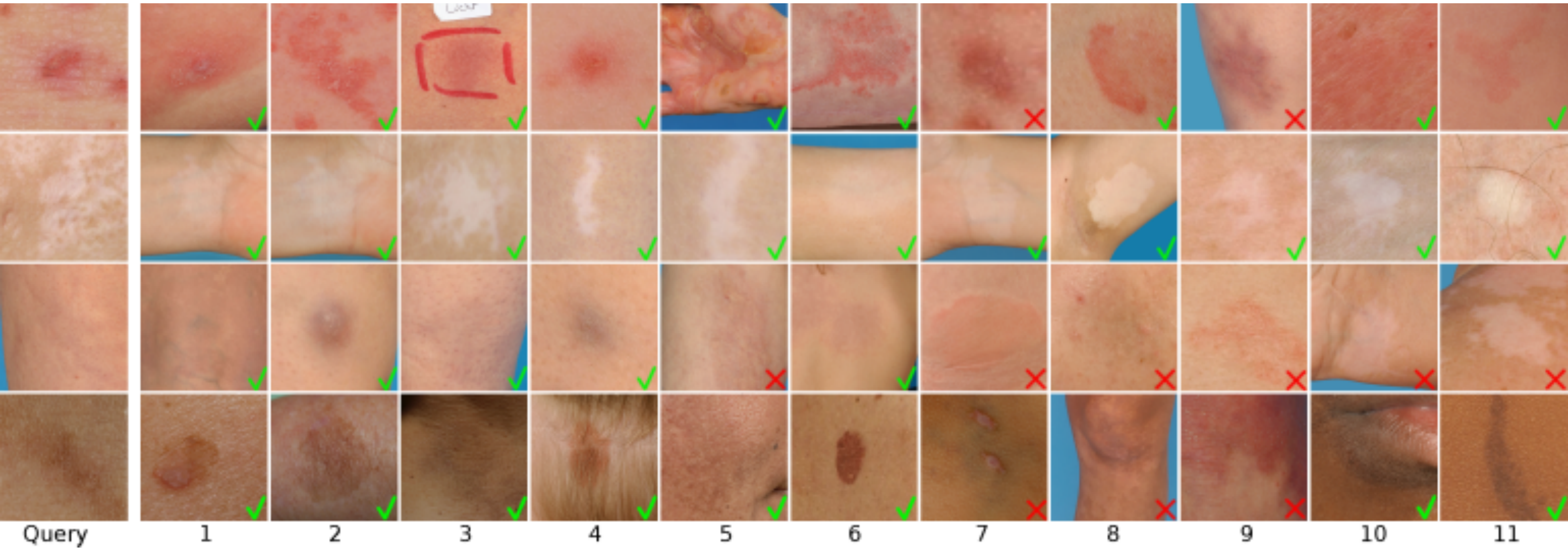


Finger print identification

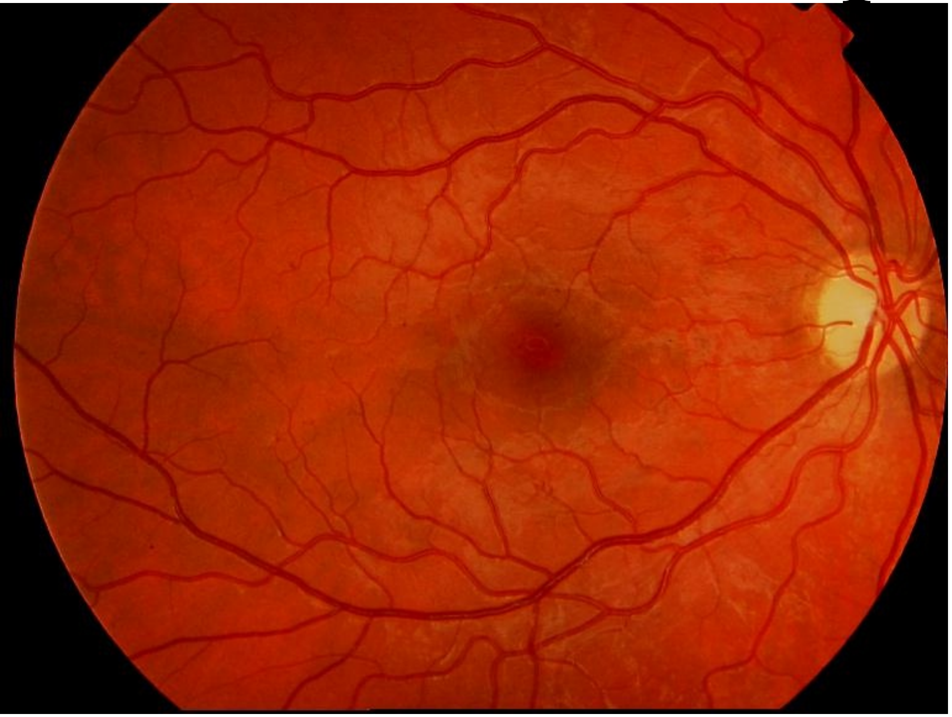


Hand/palm authentication

Medicine – computer aided diagnosis



Medicine – computer aided diagnosis



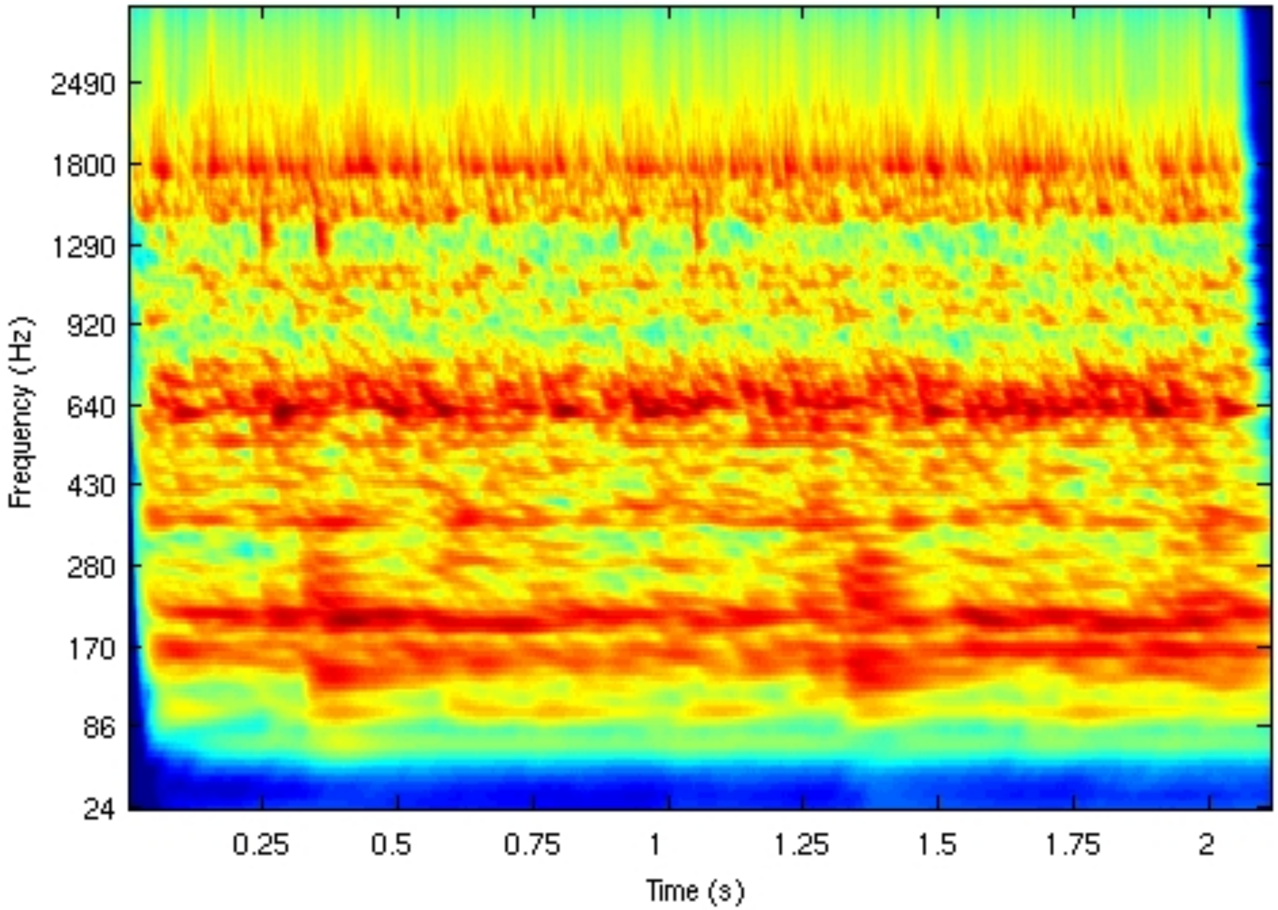
Normal retina (2002)

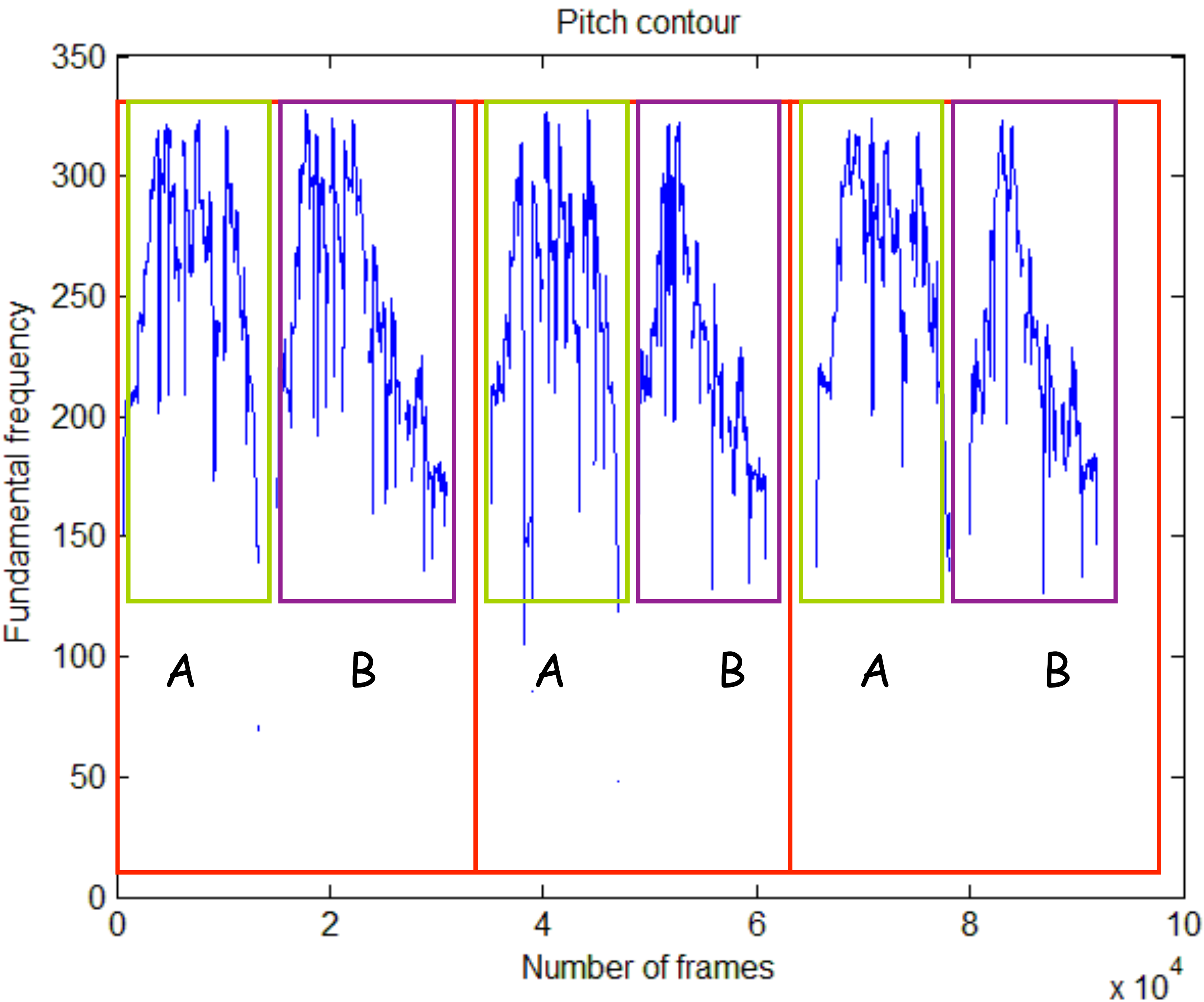
Diabetic retinopathy (2009)

License plate recognition

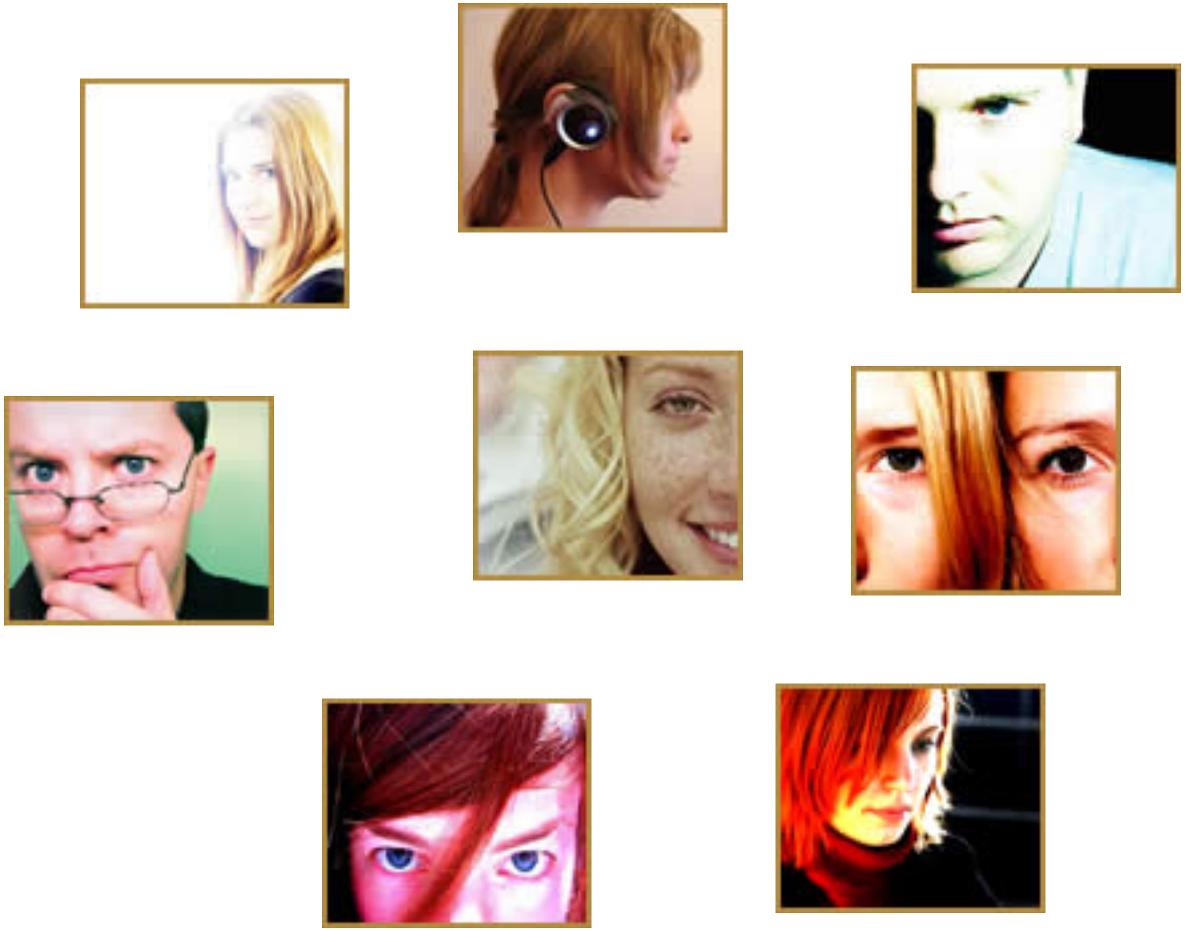
http://www.dacolian.com/
http://www.q-free.com/

Sound/speech recognition




Pitch contour
Fundamental frequency
Number of frames
x 10^4
0
50
100
150
200
250
300
350
0
2
4
6
8
10
A
B
A
B
A
B

Face identification/detection



http://www.face-rec.org

Spam detection

Dear Sir or Madam:

Please reply to
Receiver: China Enterprise Management Co., Ltd. (CMC)
E-mail: unido@chinatop.net

As one technical organization supported by China Investment and Technical Promotion Office of United Nation Industry Development Organization (UNIDO), we cooperate closely with the relevant Chinese Quality Supervision and Standardization Information Organization. We provide the most valuable consulting services to help you to open Chinese market within the shortest time:

1. Consulting Service on Mandatory National Standards of The People's Republic of China.

2. Consulting Service on Inspection and Quarantine Standards of The People's Republic of China.

3. Consulting Service for Permission to Enter Chinese Market

We are very sorry to disturb you!

More information, please check our World Wide Web: http://www.chinatop.net

Sincerely yours

Phishing detection

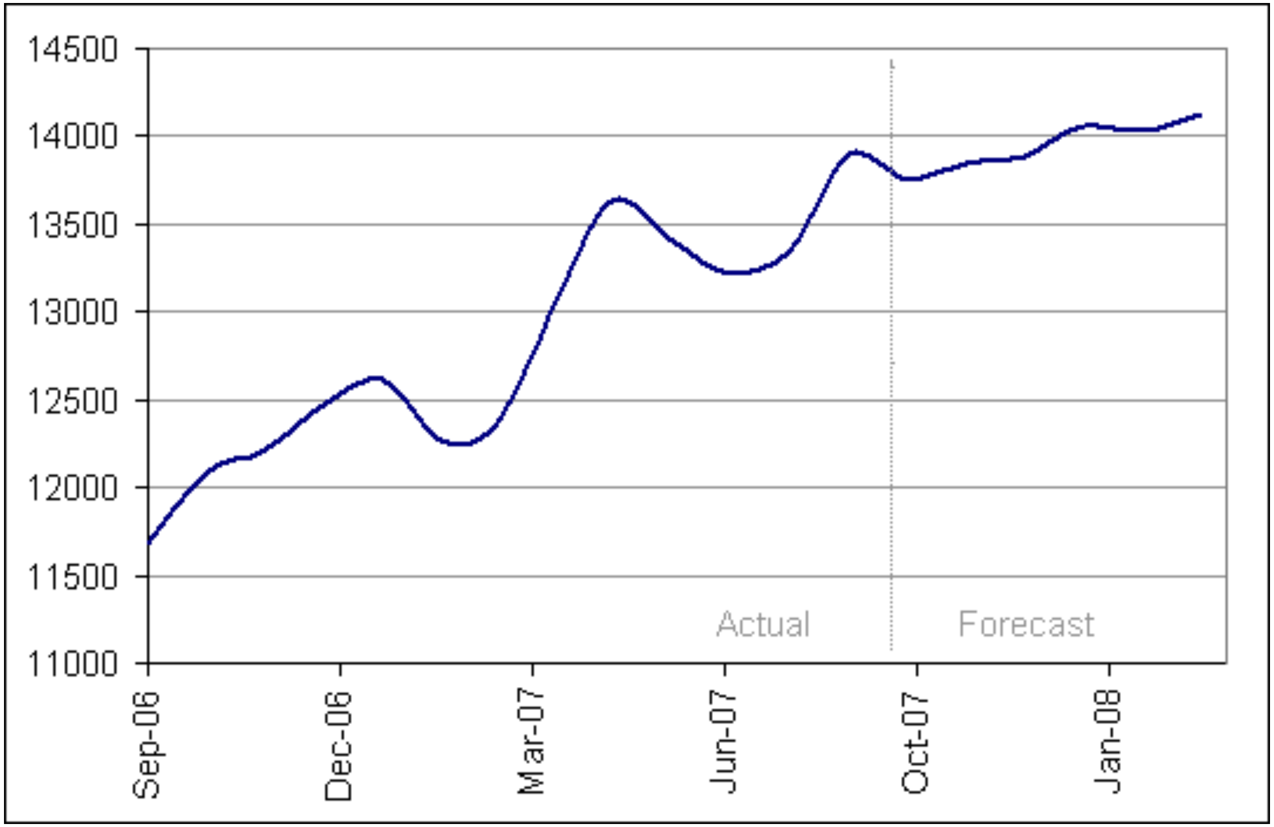
Your mailbox has exceeded the storage limit set by your administrator. You may not be able to send or receive new mail until your mailbox size is increased by your system administrator. You are required to contact your system administrator through e-mail with your Username: { } and Password:{ } to increase your storage limit.

System Administrator
E-mail: …

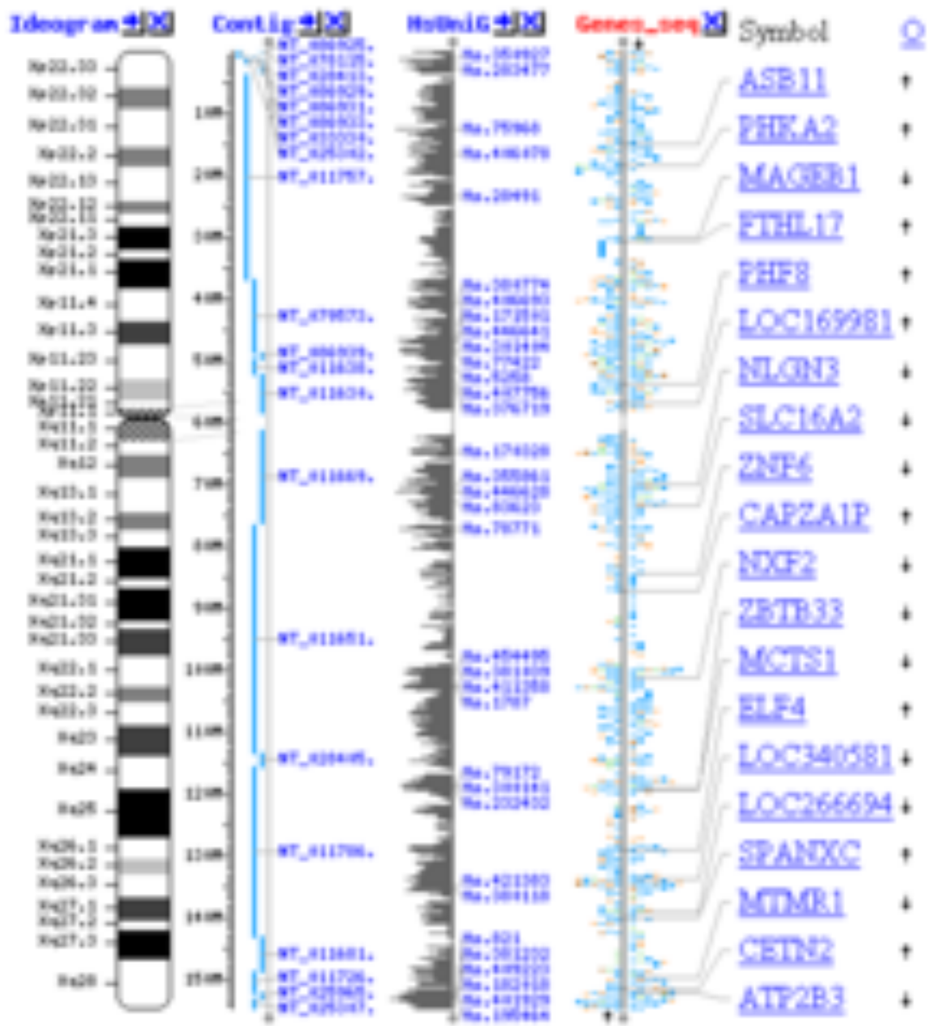
You will continue to receive this warning message periodically if your inbox size continues to exceed its size limit.

This email is intended only for the use of the individual or entity to which it is addressed and contains information that is privileged and confidential.

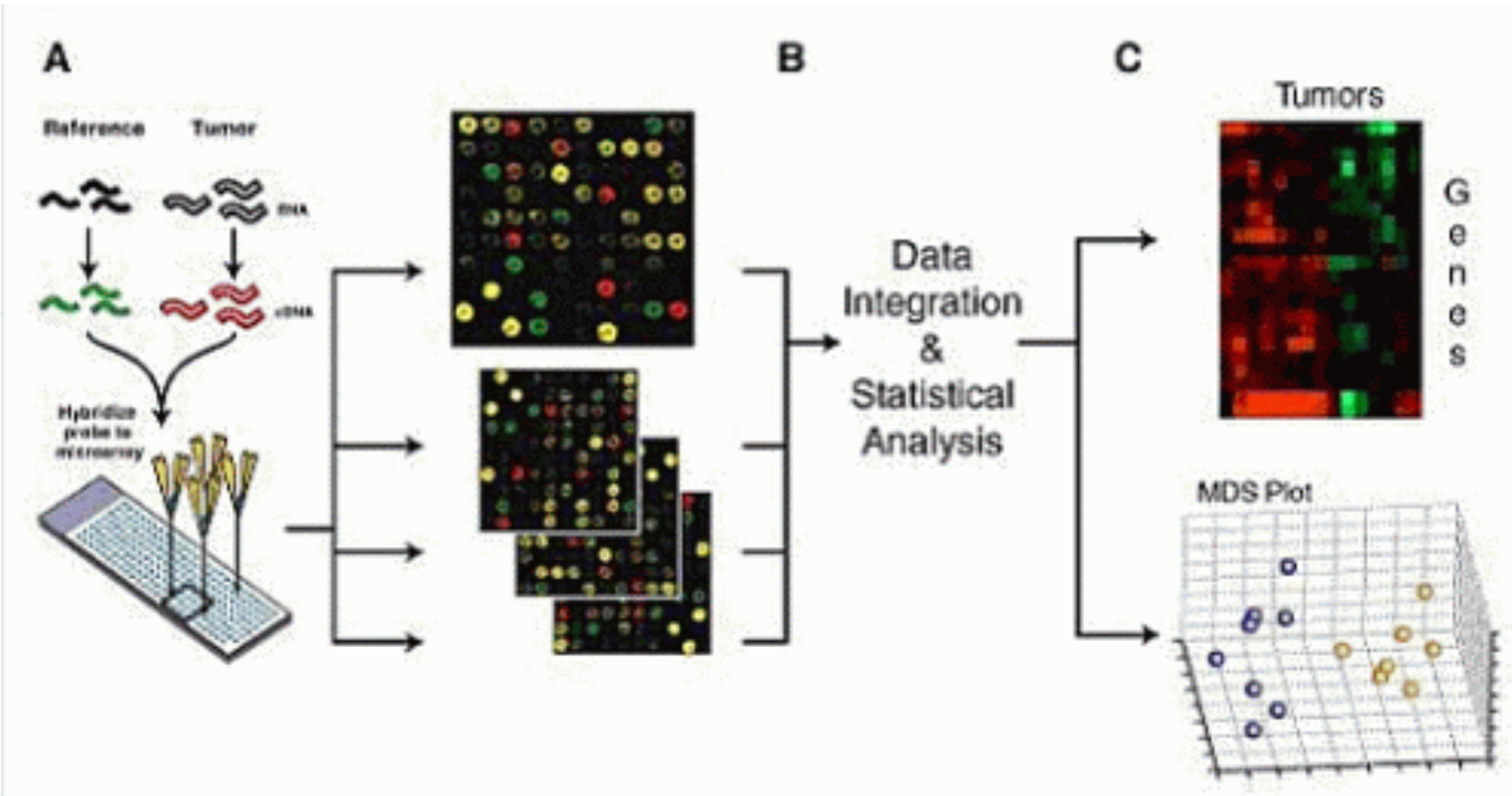
Financial analysis/forecast



Gene finding/recognition



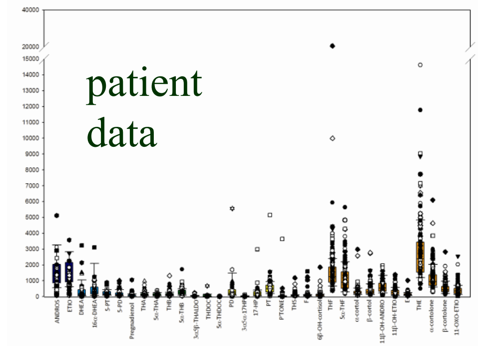
Bioinformatics

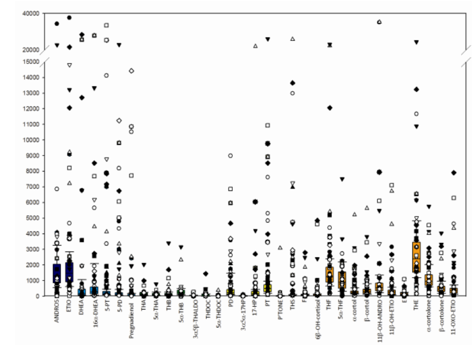


Tumor Classification

adrenocortical adenomas (ACA) carcinomas(ACC)

adrenal
gland





32 medical tests are done with each patient. Each test delivers a number that represents the concentration of a certain steroid in the urine. Thus each patient is represented by a feature vector of 32 numbers.
The above plots are superpositions of the data from many patients.

Definitions

Pattern: (relevant part of) acquired or sensed data

(e.g. a face or an iris in an image, the sound of a gun shot in an audio signal, a specific combination of words in a text)

Pattern recognition:

- data acquisition
- segmentation of relevant data
- extraction of features (descriptors)
- classification / detection / identification / authentication

Bar code – a pattern easily recognized by a computer

Bar codes

Wikipedia

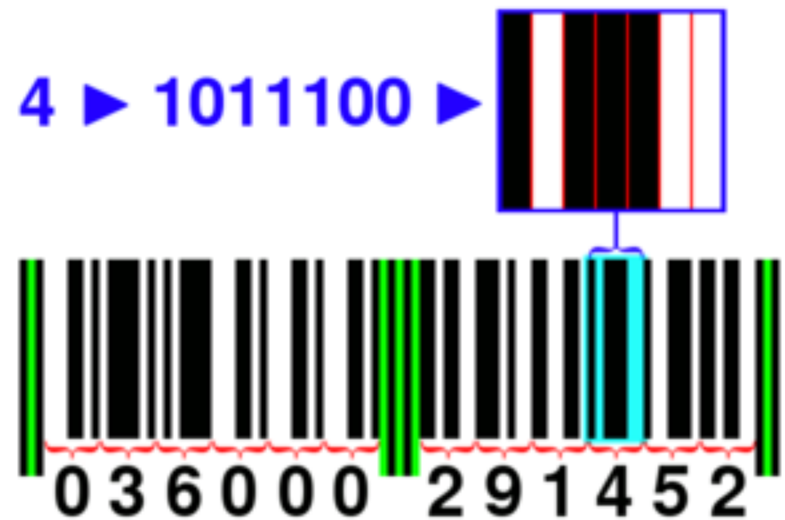
Wikipedia

0 36000 29145 2

Universal product code (UPC)

Universal product code (UPC)

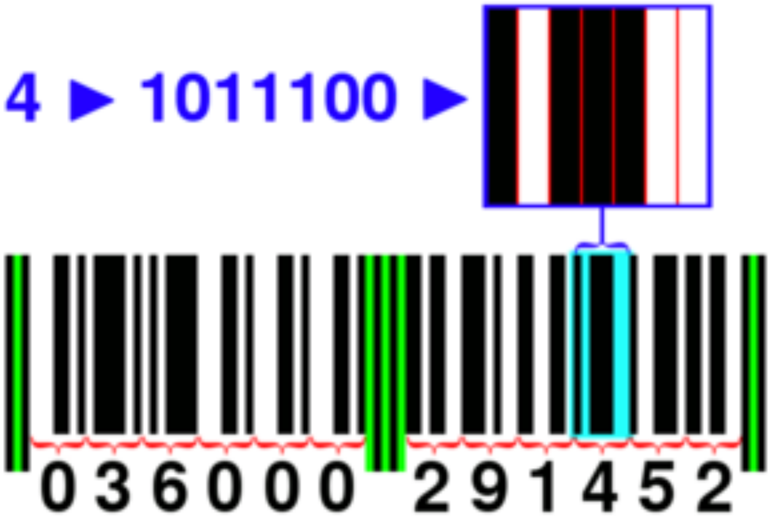
- 30 bars that code for 95 bits
- Starts and ends with 101
- 01010 in the middle
- A group of 7 bits codes a decimal digit, e.g. 1011100 = 4



Automatic UPC recognition

Feature extraction

- segment UPC in an image
- divide UPC in 95 segments of equal width
- assign 1 or 0 to dark or light segments
- result: 95 bits (binary features);
- for a given product ALWAYS THE SAME, NO SPREAD IN VALUES

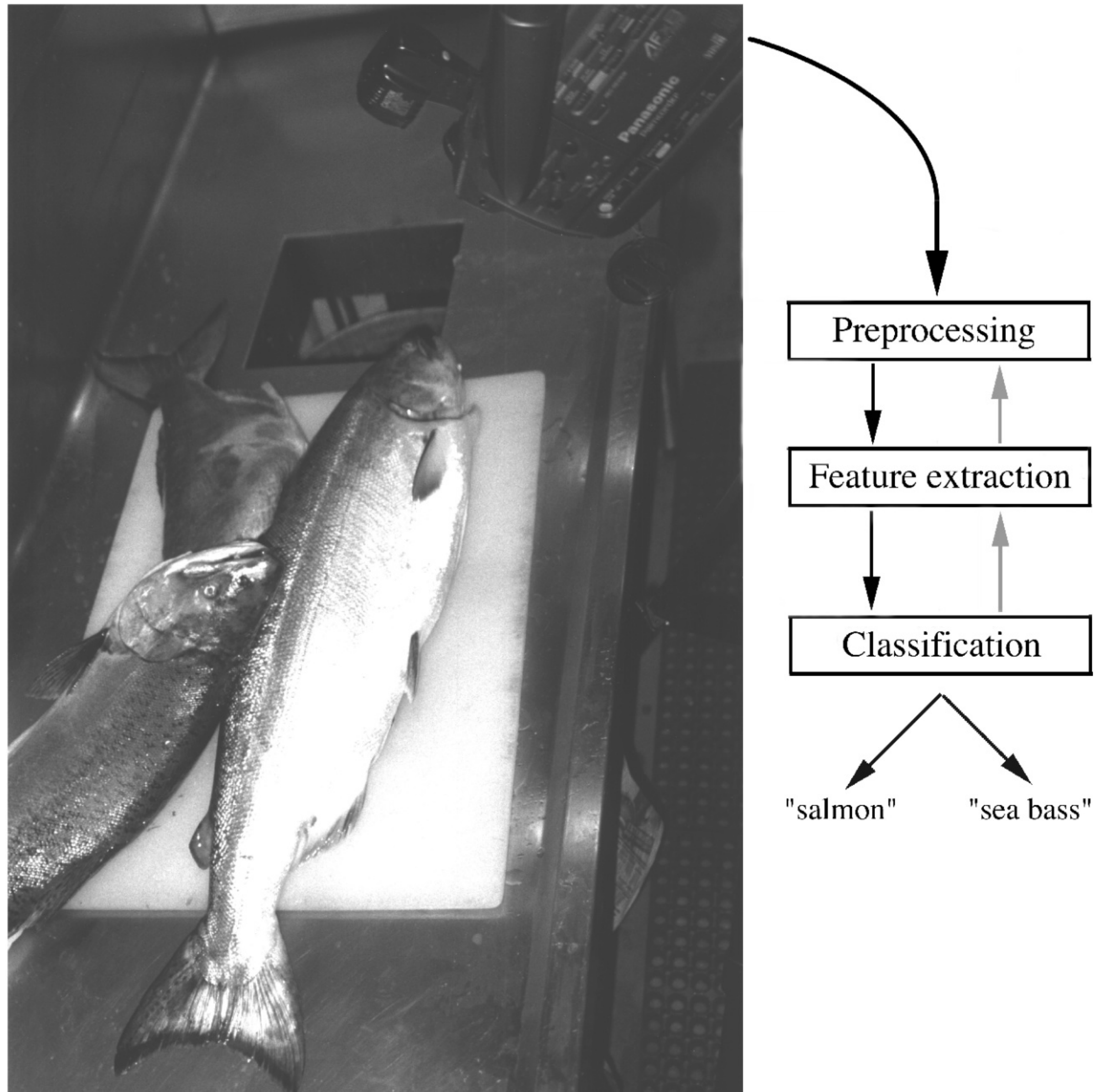


Recognizing natural patterns

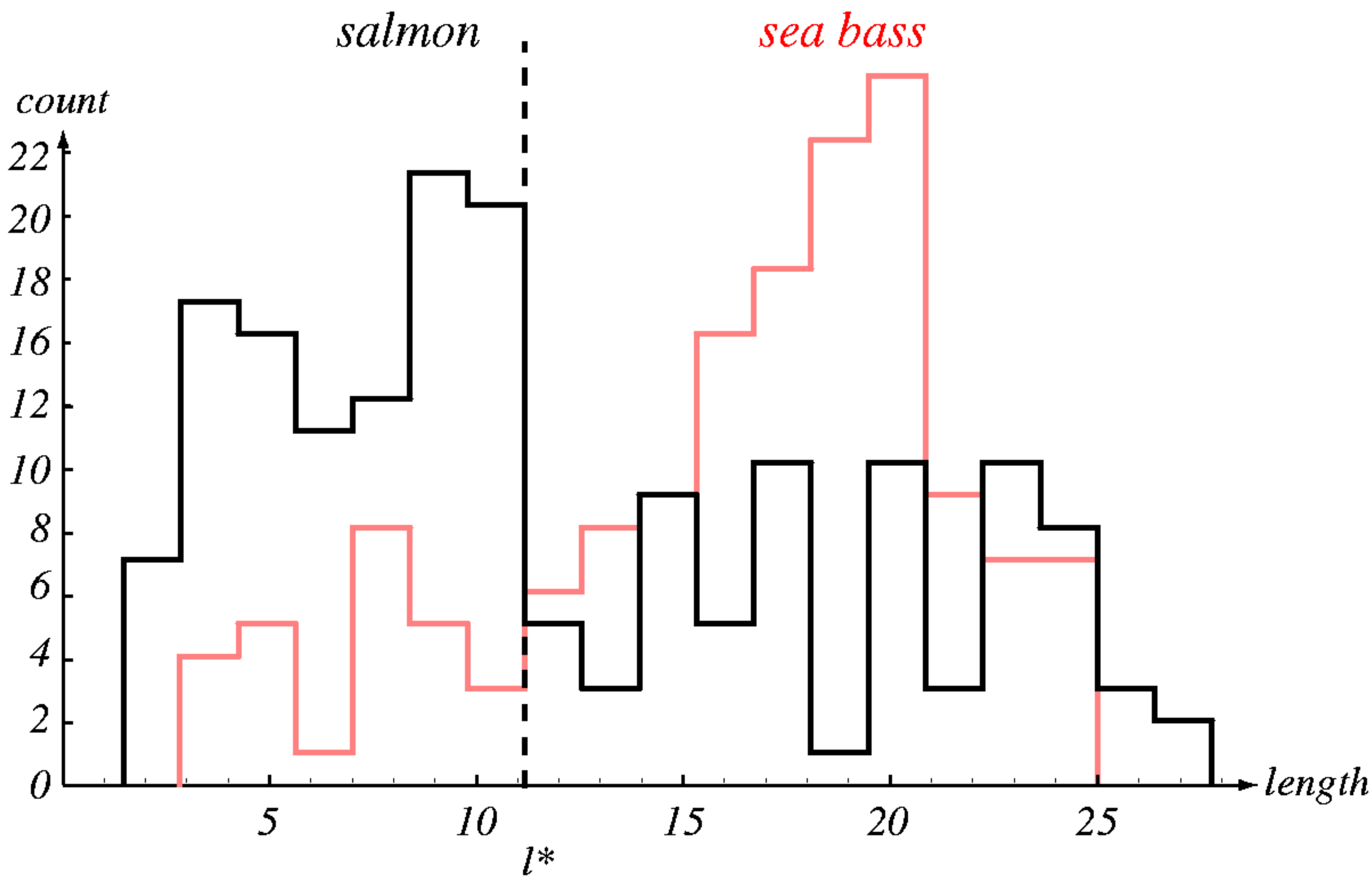
Recognizing natural objects

Example:

Automatic sorting of fish on a conveyor belt

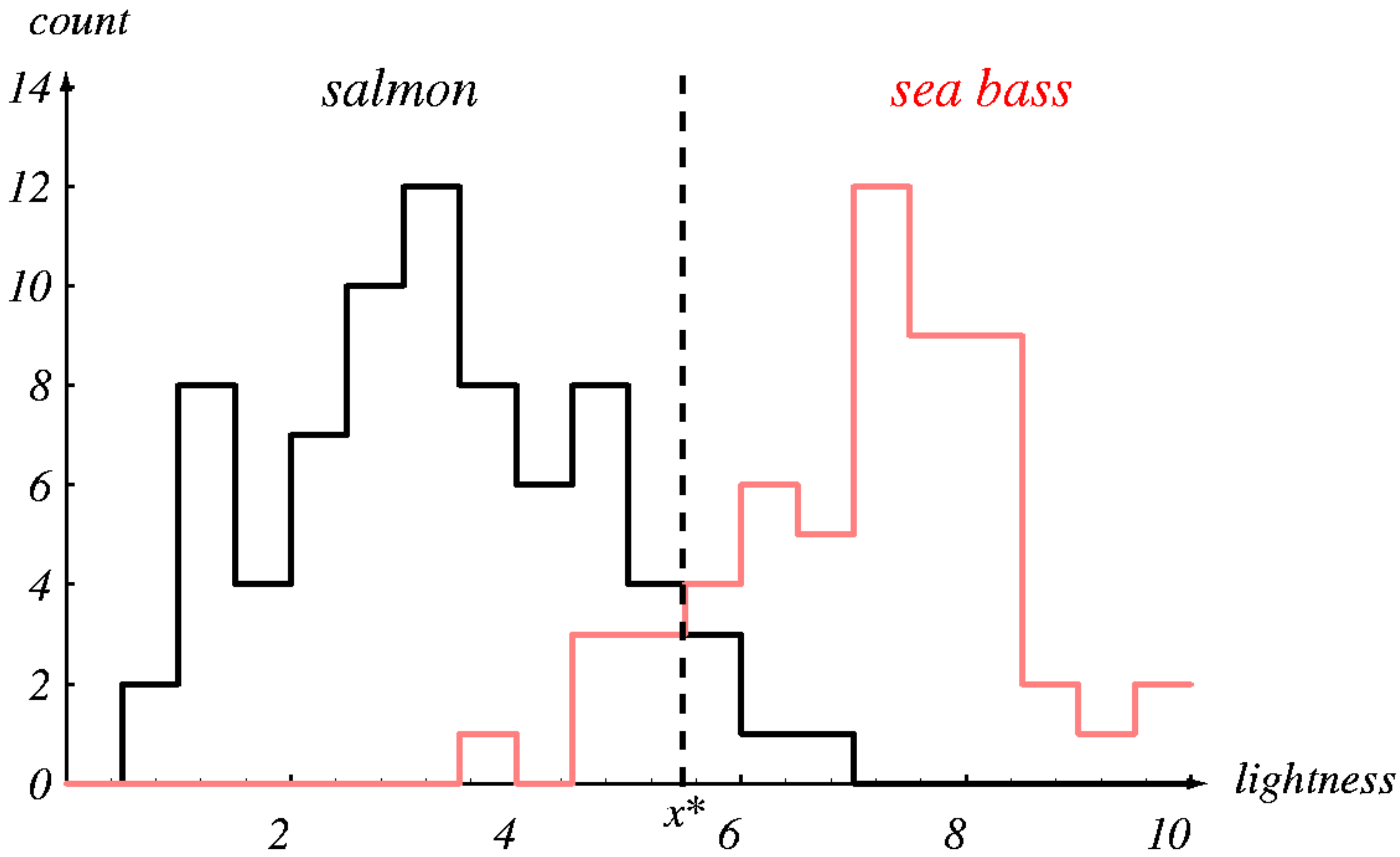


Feature values extracted from natural objects



Histograms of the "length" feature for the two categories
(from Duda, Hart, Stork (2001) Pattern classification)

Feature values extracted from natural objects



Histograms of the "lightness" feature for the two categories

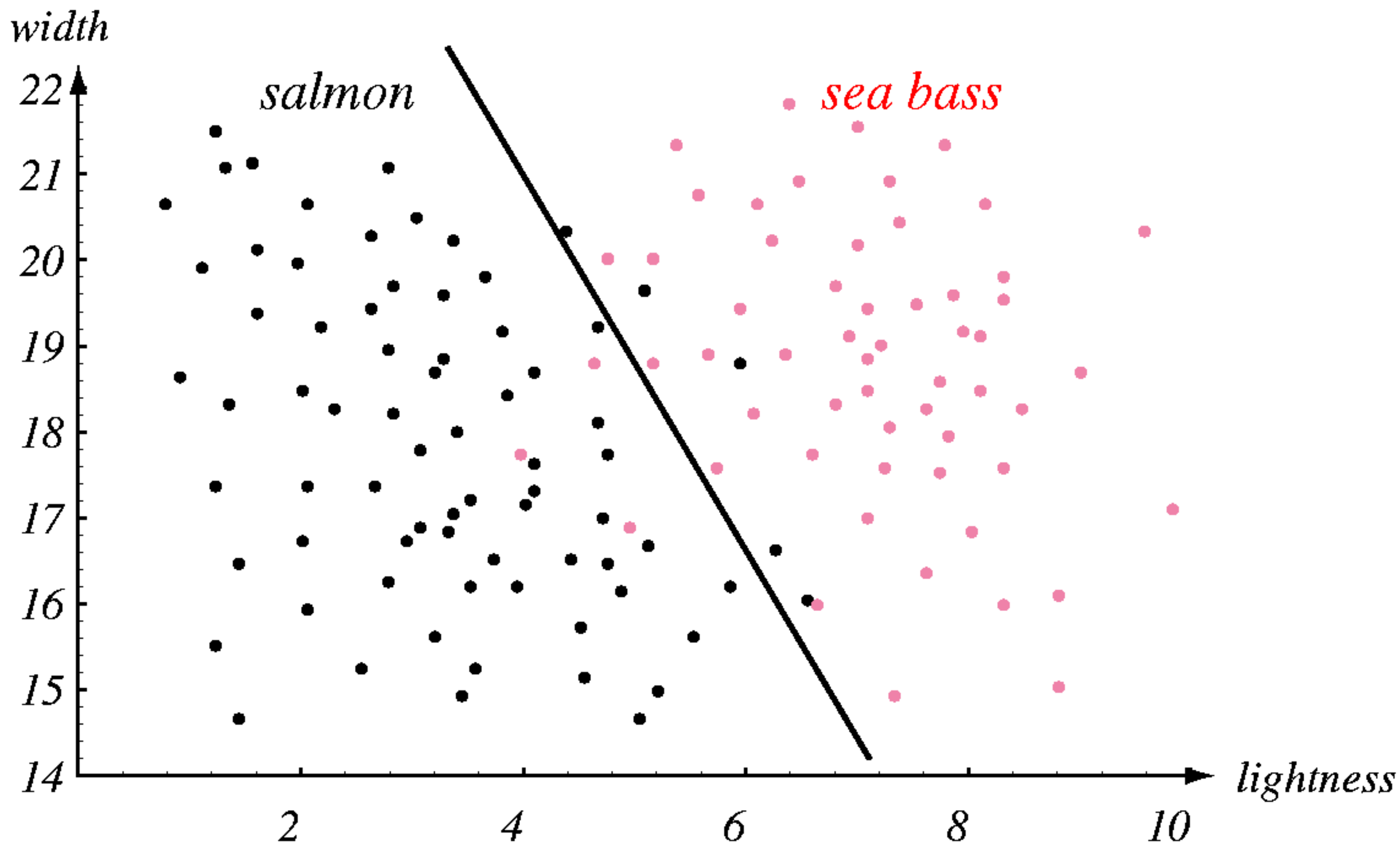
(from Duda, Hart, Stork (2001) Pattern classification)

Classification using a decision boundary

In the above examples, classification is done by taking a certain value, called *decision criterion*, and classifying objects to class A or B, depending on whether their feature values are larger or smaller than the decision criterion.

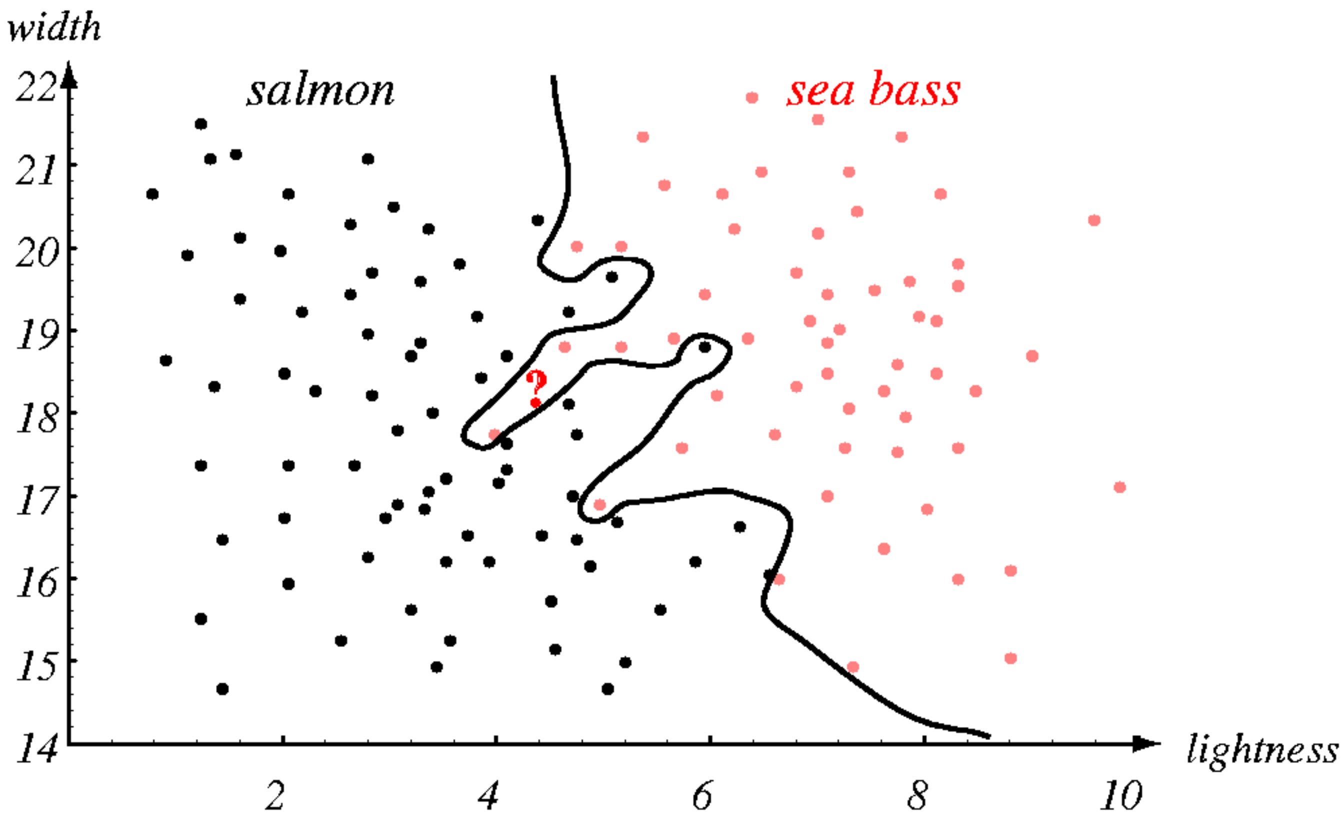
One single feature (length or lightness) might not be enough to achieve good classification. We can pack them in a feature vector x = (length, lightness). In that case, we use a *decision boundary*.

One possible decision boundary



(from Duda, Hart, Stork (2001) Pattern classification)

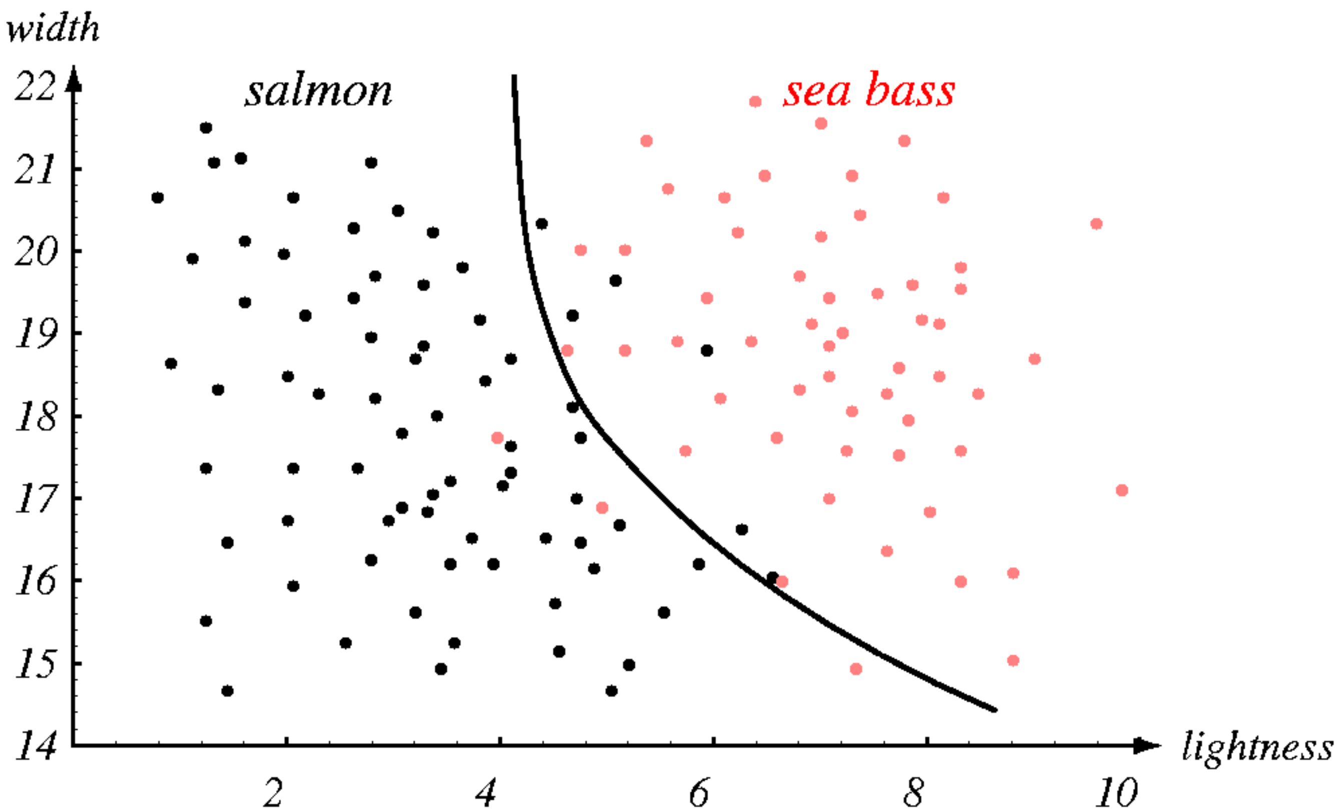
Another possible decision boundary



A complex decision boundary

(from Duda, Hart, Stork (2001) Pattern classification)

Yet another possible decision boundary



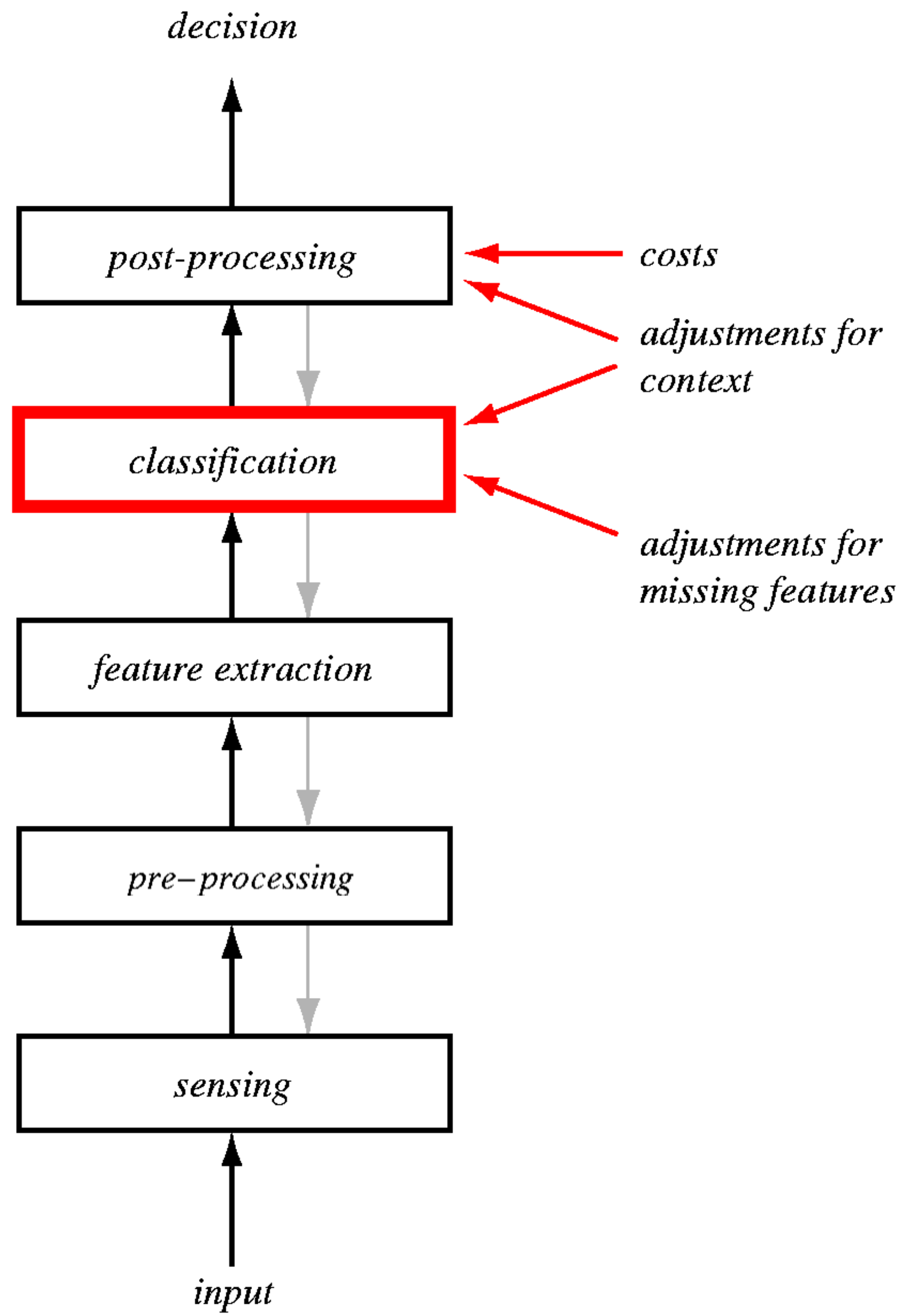
Optimal (?) decision boundary

(from Duda, Hart, Stork (2001) Pattern classification)

Statistical Pattern Recognition

- The values of the features extracted from natural objects exhibit a certain spread
- This spread is dealt with by statistical methods
- The inferences made are stated in terms of probabilities
- The quality of classification is measured in terms of error rates

Components of a pattern recognition System



Stages of PR

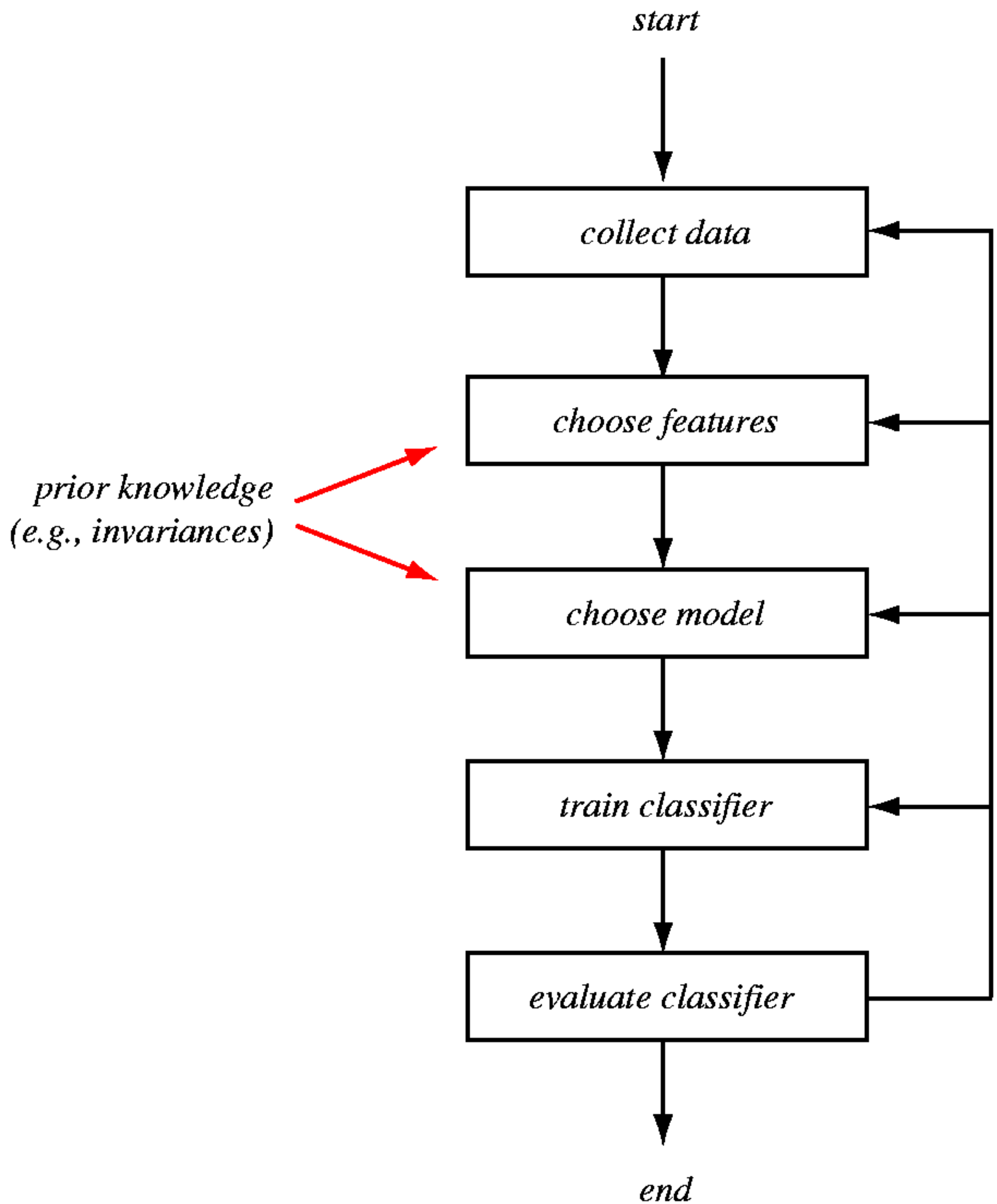
Pre-processing/segmentation – select relevant information from input data

Feature extraction - extract numerical or symbolic values

Classification – assign a category to **an object**

Post-processing – e.g. take some action taking into account error/risk/cost

Design of a pattern recognition system



Learning

A pattern (object) that is represented by a feature vector needs to be classified. This can be achieved by comparing this feature vector with previously stored feature vectors for which it is known what types of object they represent.

The process of providing an automatic pattern recognition system with such feature vectors can be considered as ‘learning’. More generally, ‘*learning*’ means using training data to automatically adjust some inner parameters of a classifier system. The commonly used term is ‘*machine learning*’.

The type of learning in which a feature vector is provided with a label that specifies the class of the object is called ‘*supervised learning’*.

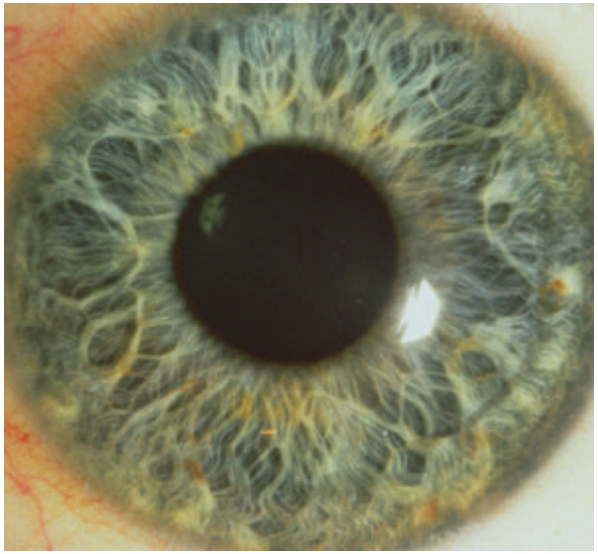
Summary of concepts and facts

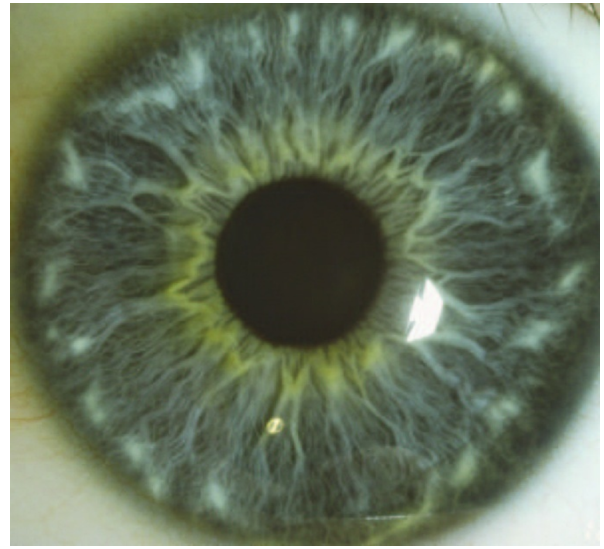
- Pattern, pattern recognition, examples
- Bar code, UPC
- Features, mapping of objects into a multi-dimensional feature space
- Decision criterion/boundary
- Difference between bar code and feature vectors of natural objects
- Aspects of statistical pattern recognition
- Components of a pattern recognition system
- Learning

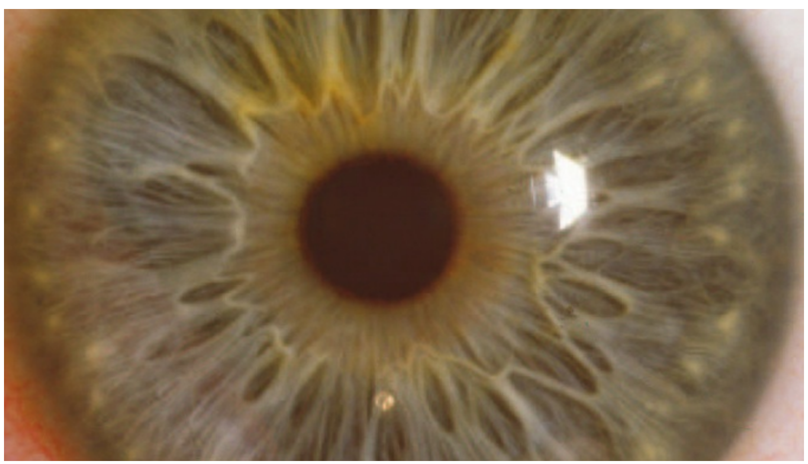
Recognition based on (dis)similarity to a prototype
and
Statistical decision theory (Hypothesis testing)

Person authentication by iris pattern

Iris images

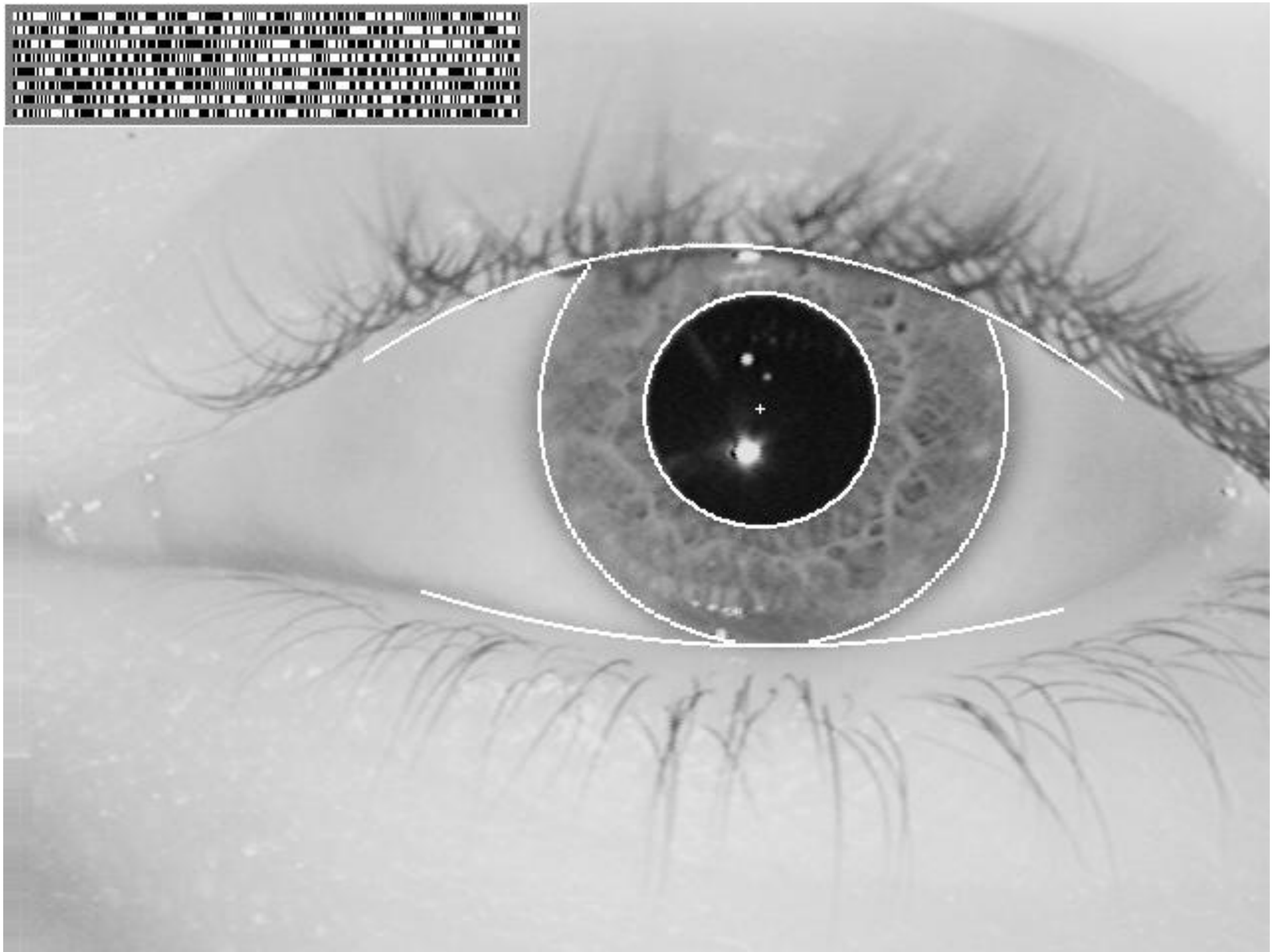




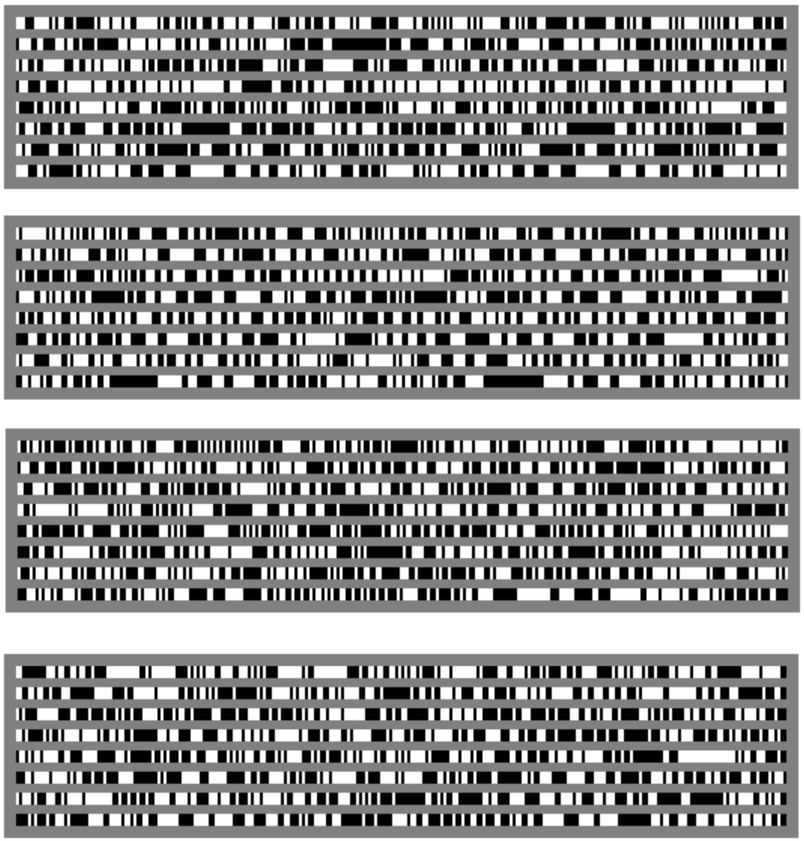


Iris Scan

Extracting binary code from iris



Binary iris codes



Person 1

Person 2

Person3

Person 4

Hamming distance between two iris codes

I_t – test iris code, I_p – prototype iris code

$H(I_t, I_p) = | I_t \text{ XOR } I_p|$ - Hamming distance

$h(I_t, I_p) = H(I_t, I_p)/\dim(I)$ - normalized H. distance

Example:

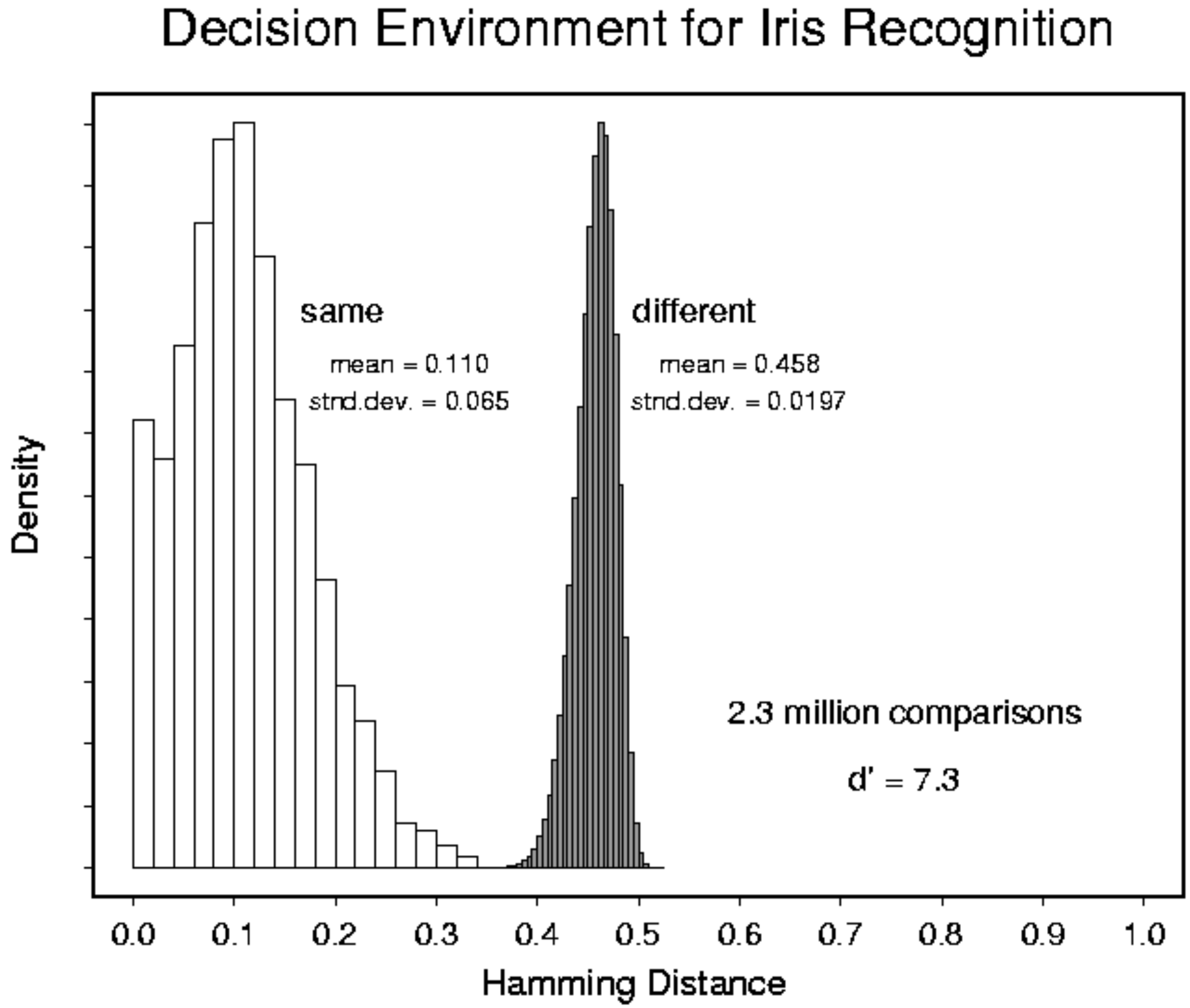
$$I_t = (1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0)$$

$$I_p = (0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0)$$

$$I_t \text{ XOR } I_p = (1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0)$$

$|I_t \text{ XOR } I_p| = 7$, $h(I_t, I_p) = 8/14 = 0.56$

Hamming distance of iris codes



Histograms of the Hamming distances of pairs of iris codes (left – for the comparison of two iris patterns of the same person; right – for different persons)

Model

The distribution of values of the Hamming distance of the iris codes of *different* persons can be modeled very well by a normal distribution, i.e. the envelope of that histogram can be fitted by a Gaussian function

$$N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

Why do we need a mathematical model?

Because we can use it to compute the probability of measuring a certain value of the Hamming distance also in domains in which the histogram contains no data. (The fact that a histogram does not contain any data in a given domain does not mean that no such values will ever be measured.)

Statistical decision theory (Hypothesis testing)

- Given the pdf(HD) that two irises are different
- Formulate hypotheses
  - H_0, null hypothesis: two irises are different
  - H_a, alternative: the two irises are identical
- Select a decision criterion (HD value) C and a rejection region according to the desired significance level (odds of false decision or false rejection of H_0)
- Compare two iris codes and obtain their HD
- Test H_0 hypothesis
  - do not reject H_0 if HD > C (type II error, fn)
  - reject H_0 and accept H_a if HD $\leq$ C (type I error, fp)

Decision error of type I
(False acceptance of an imposter)

HD Criterion	Odds of false decision
0.26	1 in 10^{13}
0.27	1 in 10^{12}
0.28	1 in 10^{11}
0.29	1 in 13 billion
0.30	1 in 1.5 billion
0.31	1 in 185 million
0.40	1 in 170

Applications of Iris Recognition

- National border control
- Computer login
- Data security (control of privileged access)
- Anti-terrorism (e.g. passenger screening)
- Secure e-banking
- Bank ATM cash dispensers
- Anonymous authentication
- Drivers licenses, other personal certificates
- Birth certificates; missing or wanted persons
- Car ignition and unlocking, anti-theft devices
- Any existing use of keys, cards, PINs, passwords

An Afghanistan story

An Afghan girl on the cover page of National Geographic (1984)

Who is the authentic person?

Sharbat Gula (1984) identified (2002) by iris comparison

More on Iris Scan

See: John Daugman

http://www.cl.cam.ac.uk/~jgd1000/

More on hypothesis testing

http://wikipedia.org

Summary of concepts and facts

- Authentication by iris pattern
- Iris code extraction
- Hamming distance of iris codes
- Histogram and probability density function
- Hypothesis testing
- Errors of type I and II