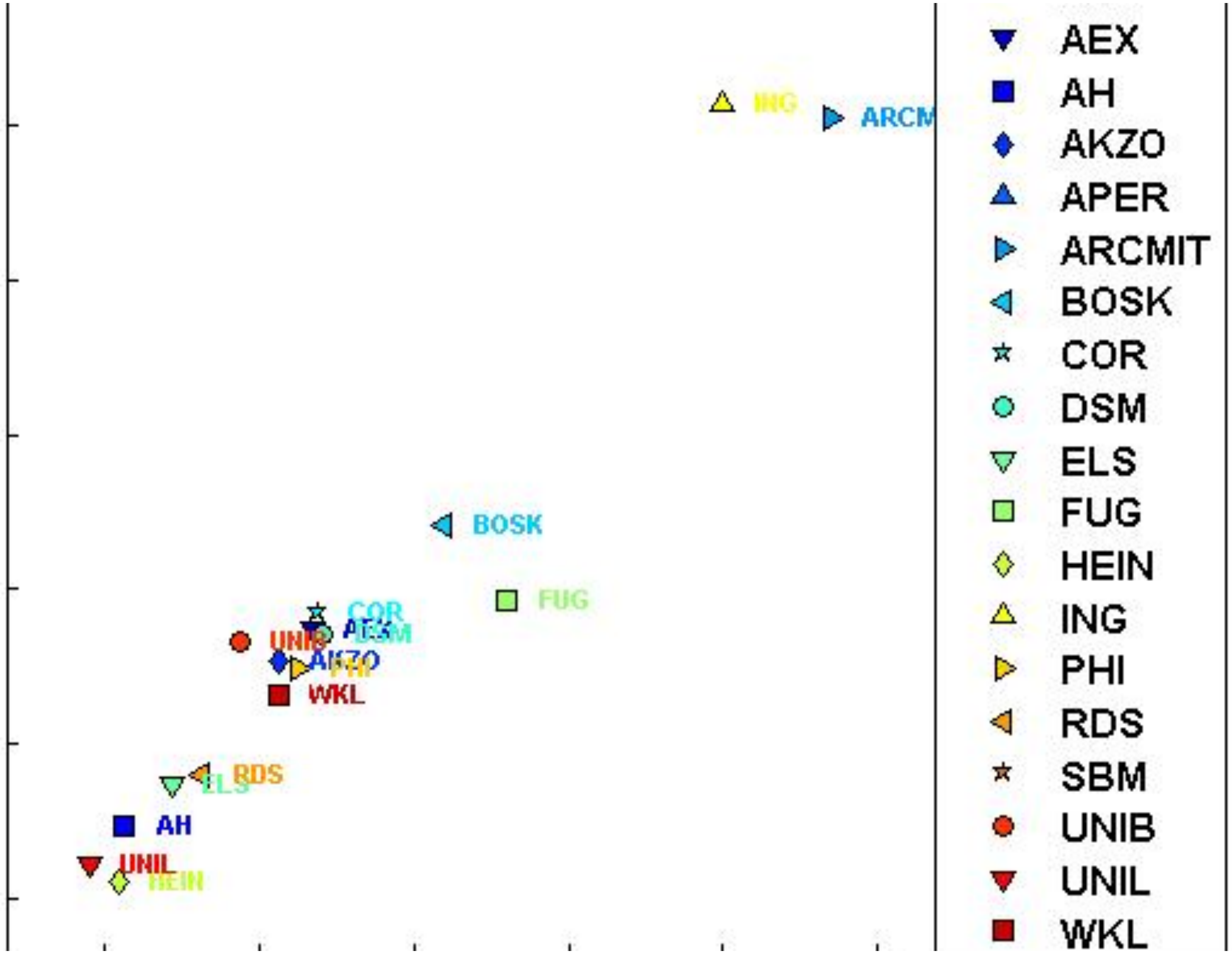
Clustering
by proximity to prototypes

k-means clustering

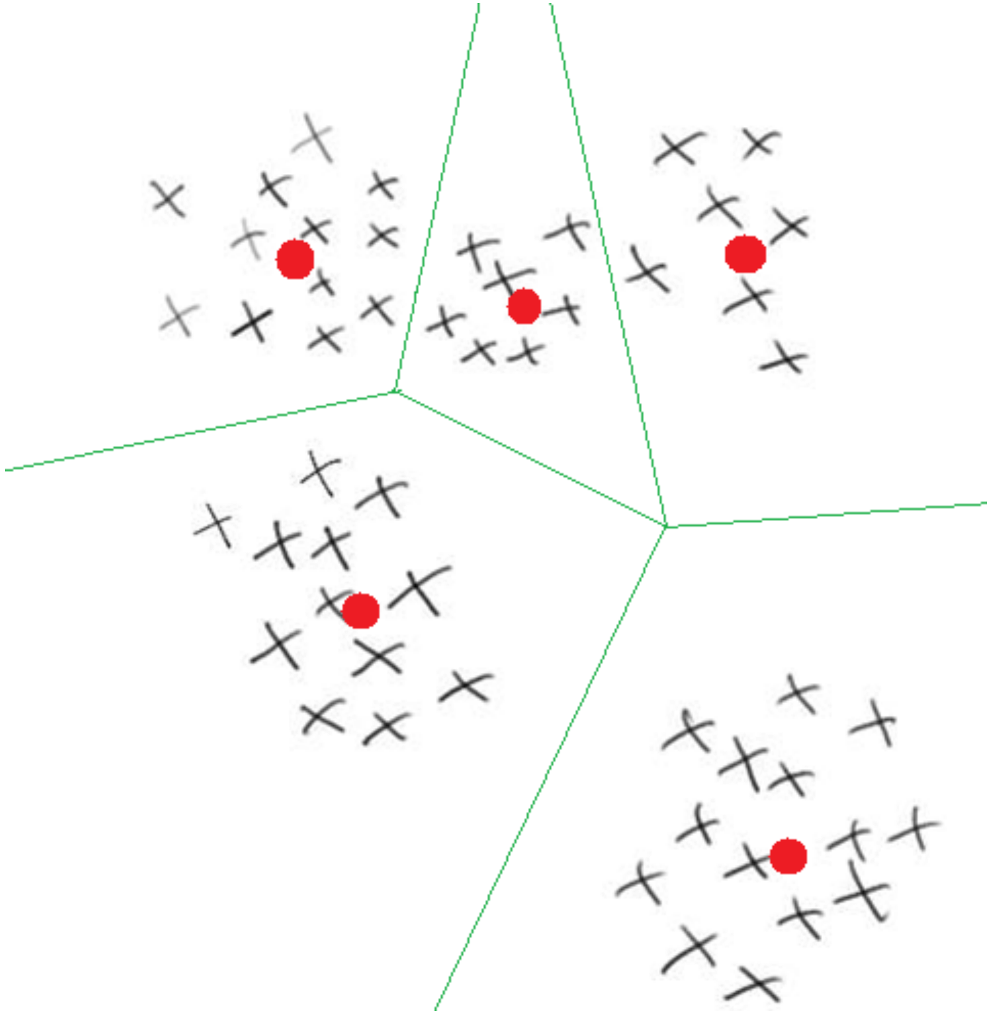
What is the goal of clustering?

Division of a data set X into k disjoint subsets $C_1 \ldots C_k$ such that objects within each subset are similar and objects in different subsets are dissimilar

Example
# of data to be clustered



k-means clustering



Euclidean-distance, prototype-based clustering: assign a data point to the nearest prototype

k-means clustering

- given: elements x^j in R^n, number of clusters k
- Goal: find k prototypes μ^i

  that minimize the quantization error

$$J_e = \frac{1}{2}\sum_{\vec{\mu}^i} \sum_{\vec{x}^j \in C(\vec{\mu}^i)} \left\| \vec{x}^j - \vec{\mu}^i \right\|^2$$

$C(\mu^i)$ – cluster (subset of X) associated with μ^i

(also called receptive field of μ^i)

Lloyd’ s algorithm for k-means clustering

```
begin initialize $\mu^1, \mu^2, ..., \mu^k$ (e.g. take randomly k samples
        from the data set)
    do assign data points to nearest $\mu^i$ (compute $C^i$)
        re-compute $\mu^i$ as the mean of points in $C^i$
    until no change in $\mu^1, \mu^2, ..., \mu^k$
    return $C^1, C^2, ..., C^k$ and $\mu^1, \mu^2, ..., \mu^k$
end
```

Lloyd' s algorithm for k-means clustering

```
1. begin initialize $\mu^1$ , $\mu^2$ ,..., $\mu^k$ (e.g. take randomly k samples
      from the data set)
2.    do  assign  data points to nearest  $\mu^i$  (compute $C^i$)
3.          re-compute  $\mu^i$  as the mean of points in $C^i$
4.    until no change in  $\mu^1, \mu^2,..., \mu^k$
5.    return $C^1$, $C^2$, ... , $C^k$ and  $\mu^1, \mu^2,..., \mu^k$
6. End
```

COMMENT ON HOW TO IMPLEMENT STEP 2 ASSIGNMENT TO CLUSTERS: USE AN INTEGER ATRRAY C such that C(i)=number of cluster to which point i is assigned

Does Lloyd’ s algorithm converge?

- Yes, in a finite number of steps, because a non-negative cost function (the quantization error) decreases (or remains constant) with each step:

$$J_e = \frac{1}{2}\sum_{\vec{\mu}^i}\sum_{\vec{x}^j \epsilon C(\vec{\mu}^i)} ||\vec{x}^j - \vec{\mu}^i||^2$$

$$\vec{\mu}^i = \frac{1}{n}\sum_{\vec{x}^j \epsilon C(\vec{\mu}^i)} \vec{x}$$

- However, there is no guarantee that a global minimum is reached

Does Lloyd’s algorithm converge?

- THE QUANTIZATION ERROR AS A FUNCTION OF THE INTERATION NUMBER MUST DECREASE MONOTONOUSLY
- IF THE QUANTIZATION ERROR SHOWS OSCILLATIONS (goes up and down) THERE MUST BE A BUG IN THE CODE

Intitialisation of k means

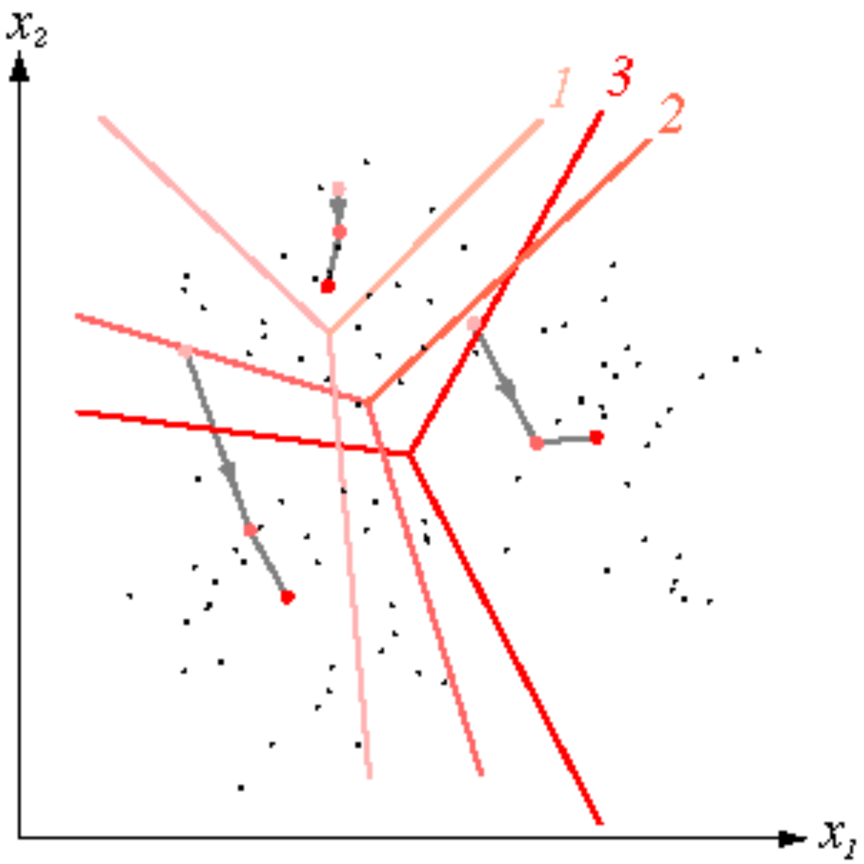
- MULTIPLE INITIALIZATIONS, e.g. take data points randomly
- RUN THE K-NN ALGORITHM FOR DIFFERENT INITIALISATIONS AND TAKE THE RESULT FOR WHICH THE QUANITIZATION ERROR IS MINIMUM

HOW TO CHOOSE K?

‘ELBOW’ METHOD:

1. Run the k-nn algorithm for multiple values of k and for each value of k record the value of the quantization error upon convergence
2. Plot the reached quantization error as a function of k
3. If the plot shows an ‘elbow’ for a certain k, take that k

Example of k-means clustering



Evolution of the (3) computed means (and Voronoi cells) during 3-means clustering

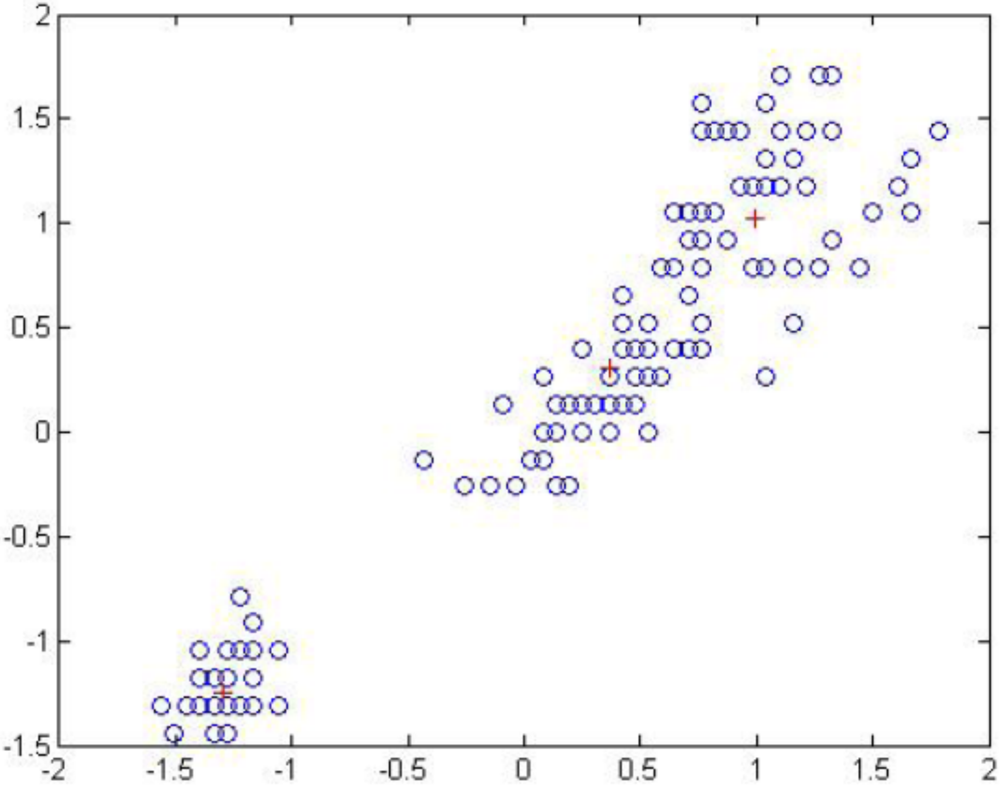
from Duda, Hart, Stork (2001) Pattern classification

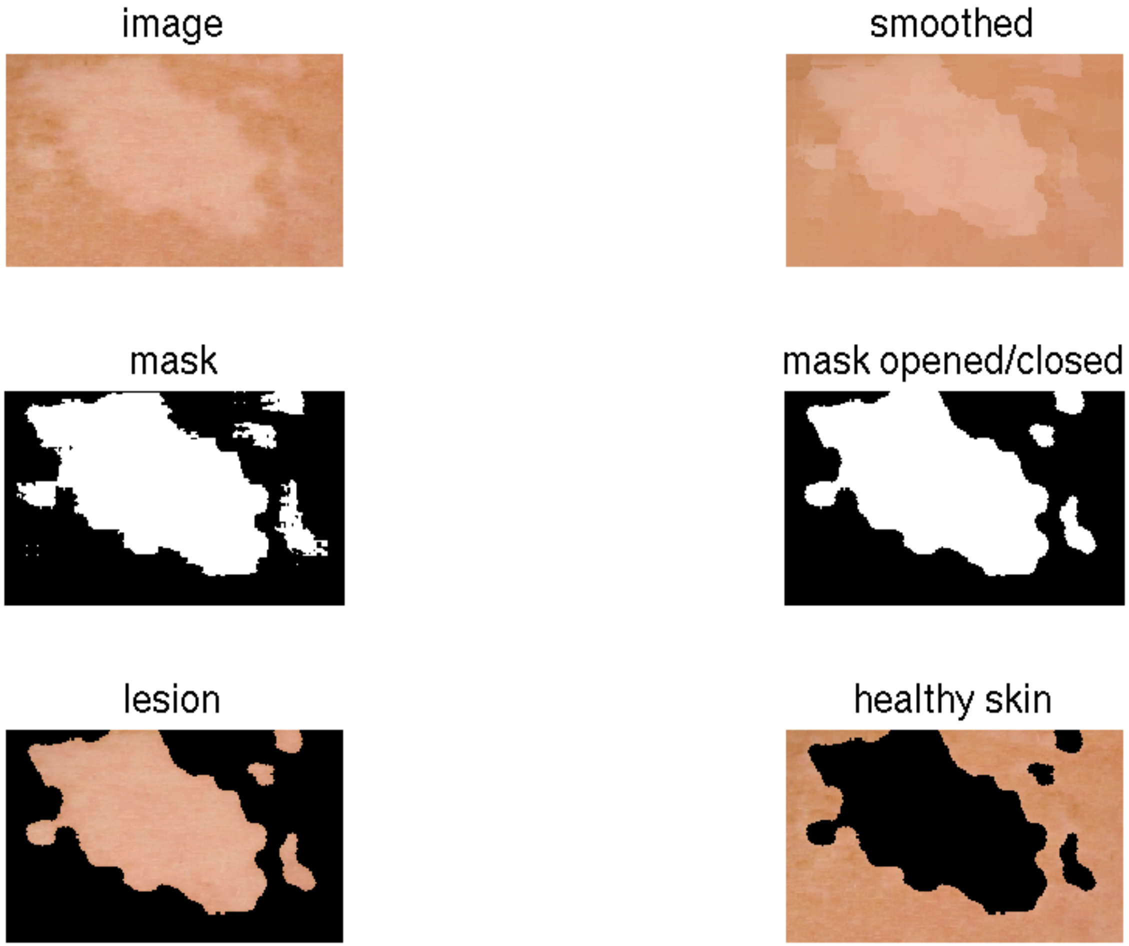
k-means applet

- http://www.cs.rug.nl/~petkov/teaching/PatternRecognition/supplements/k-means/

Iris data

- see UCI repository: http://archive.ics.uci.edu/ml/
  - – 150 points
  - – 4 dimensional
  - – 3 classes

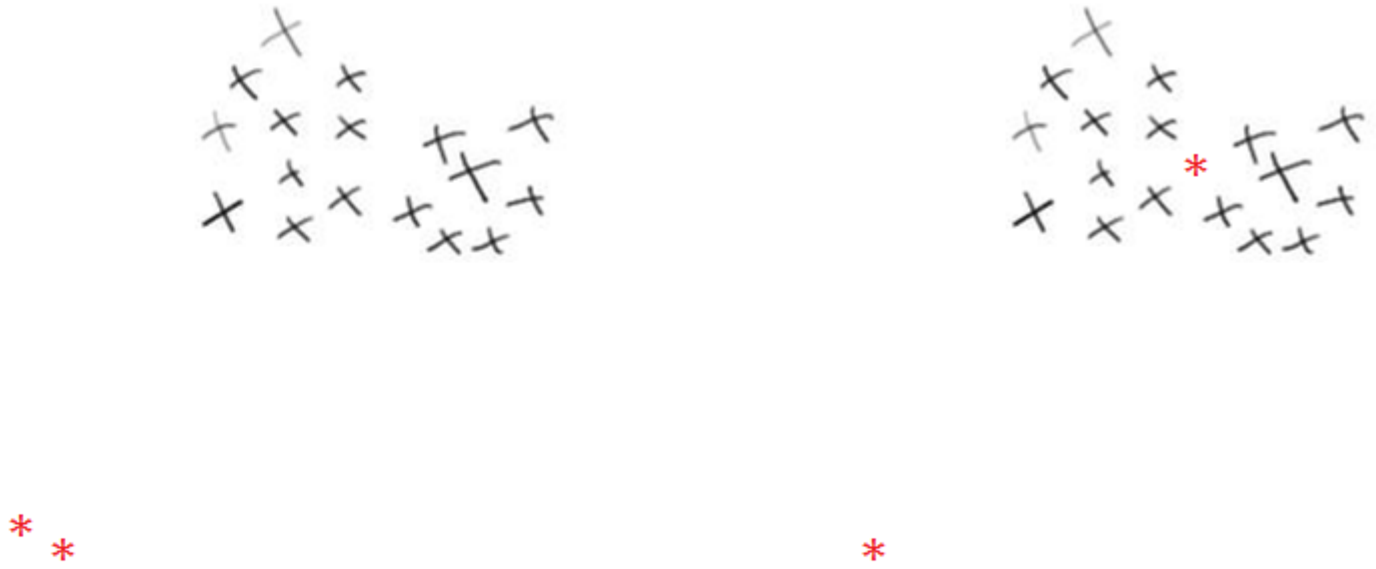




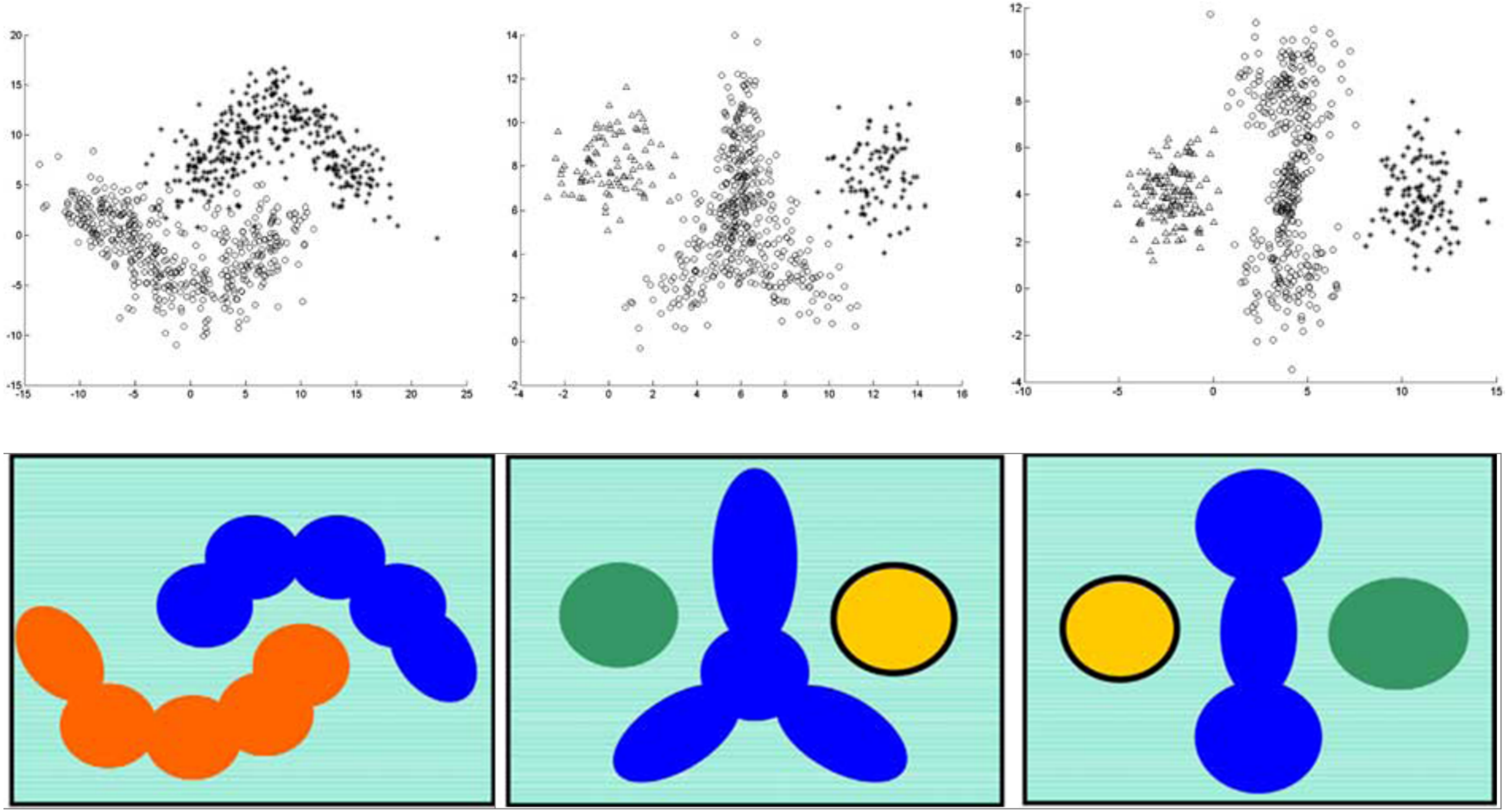
Example of 2-means clustering: a skin image is segmented in two regions of lesion and healthy skin by grouping pixels in two clusters according to their color (result shown in image mask)

Problems with k-means clustering: dead units (poor initialization)

if some prototypes are initialized far away from the input data set, no data points are assigned to them and they are never updated

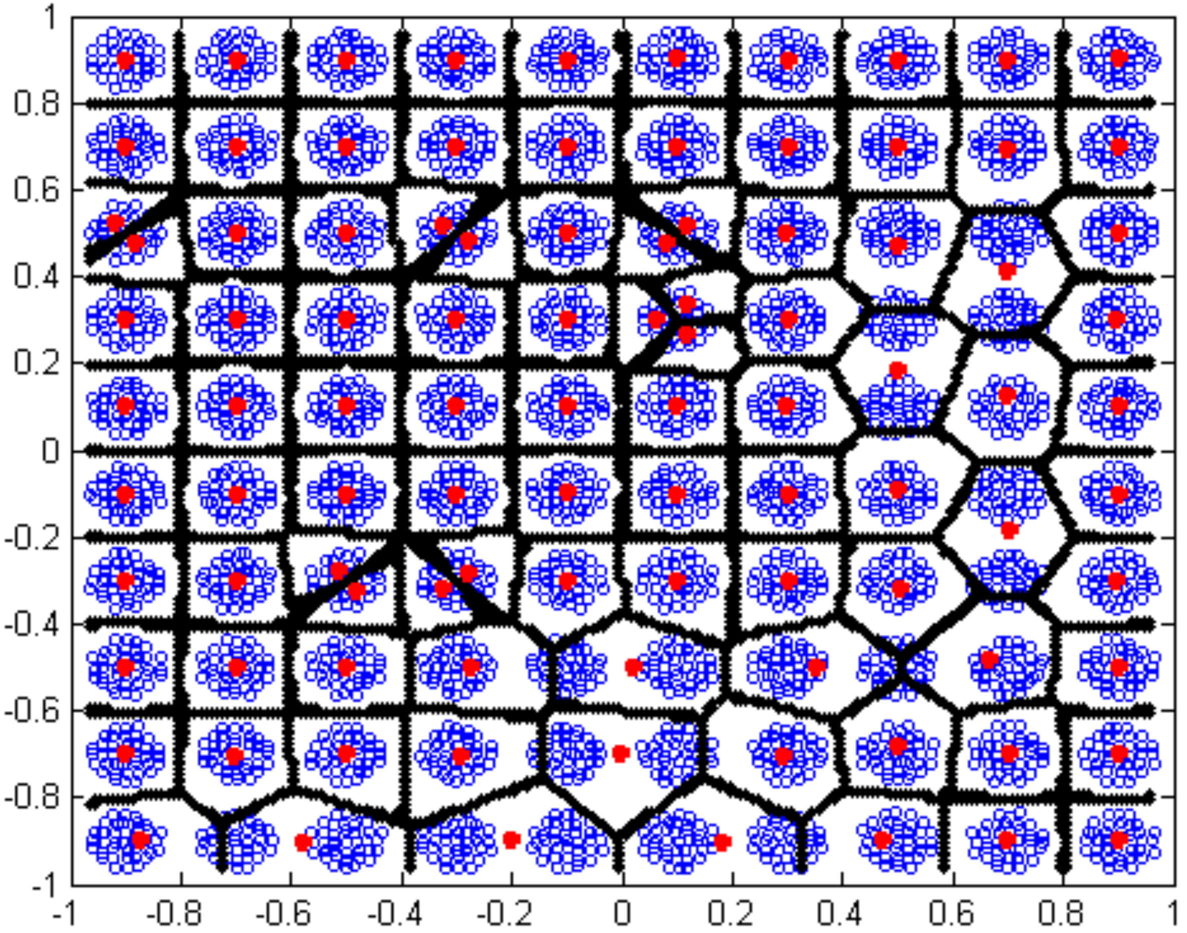


Problems with k-means clustering: non-spherical clusters



Examples of non-spherical clusters: (a) Teaeguk, (b) Triangle, (c) Xours (Cho et al.,2006)

Problems: local optima



Checkerboard data with 100 data clusters and their cluster centers

Summary of concepts

- Clustering by distance to prototypes
- k-means clustering
- Quantization error
- Lloyd' s algorithm
- Problems with Lloyd' s algorithm
- Examples for the application of k-means clustering