Computational Linguistics

Clustering

Clayton Greenberg & Stefan Thater

FR 4.7 Allgemeine Linguistik (Computerlinguistik)
Universität des Saarlandes
Summer 2015

Cluster Analysis

Goal:

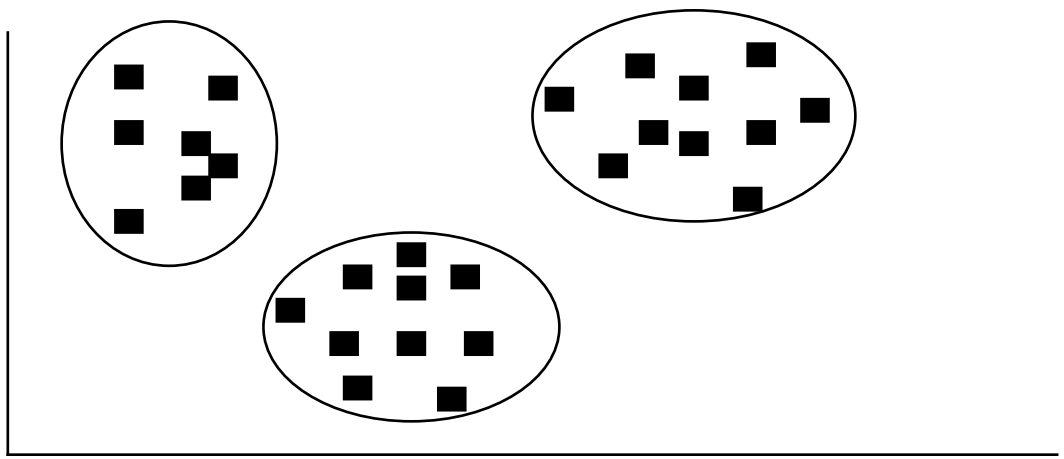
group similar items together in a group

Steps:

define distance between points in the sample

define a loss function

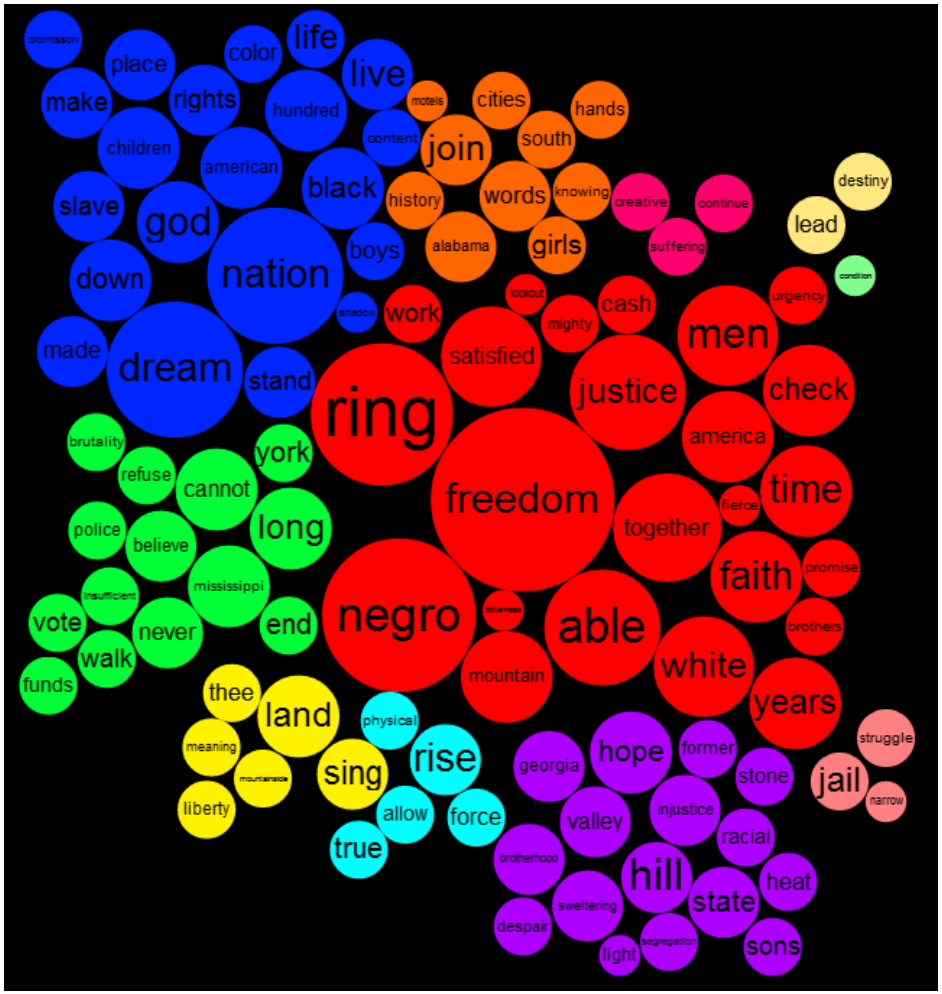
find an algorithm that minimizes this loss function



Examples

Cluster Word

(speech “I have a dream”)



From http://neoformix.com/2011/wcd_KingIHaveADream.png

Cluster Text (e.g. search results)

EisenLab
Commercial use of the ScanAlyze, **Cluster** and/or TreeView executable and/or ... **Cluster** and TreeView are an integrated pair of programs for analyzing and ...
rana.lbl.gov/EisenSoftware.htm - 11k - Cached - Similar pages

Book results for **cluster**

The Linux Enterprise Cluster : build a highly ... - by Karl Kopper - 466 pages
Messier's Nebulae and Star Clusters - by Kenneth Glyn Jones - 456 pages

Searches related to: **cluster**

cluster **headache**	cluster **analysis**	**server** cluster	cluster **sampling**
cluster **windows 2003**	**sql** cluster	**oracle** cluster	**clusty**

Goooooooooogle
1 2 3 4 5 6 7 8 9 10 **Next**

cluster Search

Search within results | Language Tools | Search Tips | Dissatisfied? Help us improve |

Cluster Image Regions:

Image Segmentation

http://people.cs.uchicago.edu/~pff/segment/

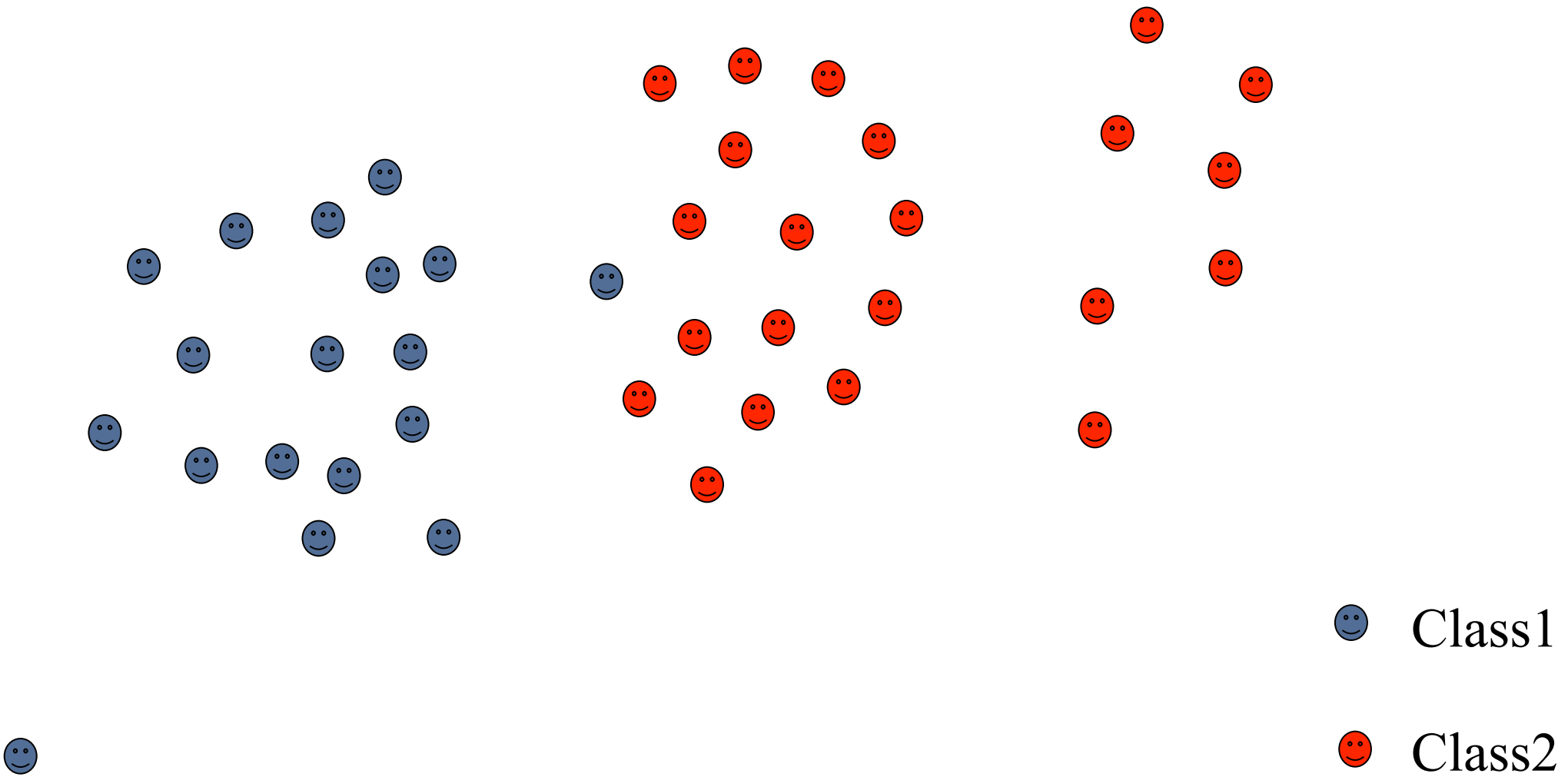
Cluster Image Regions

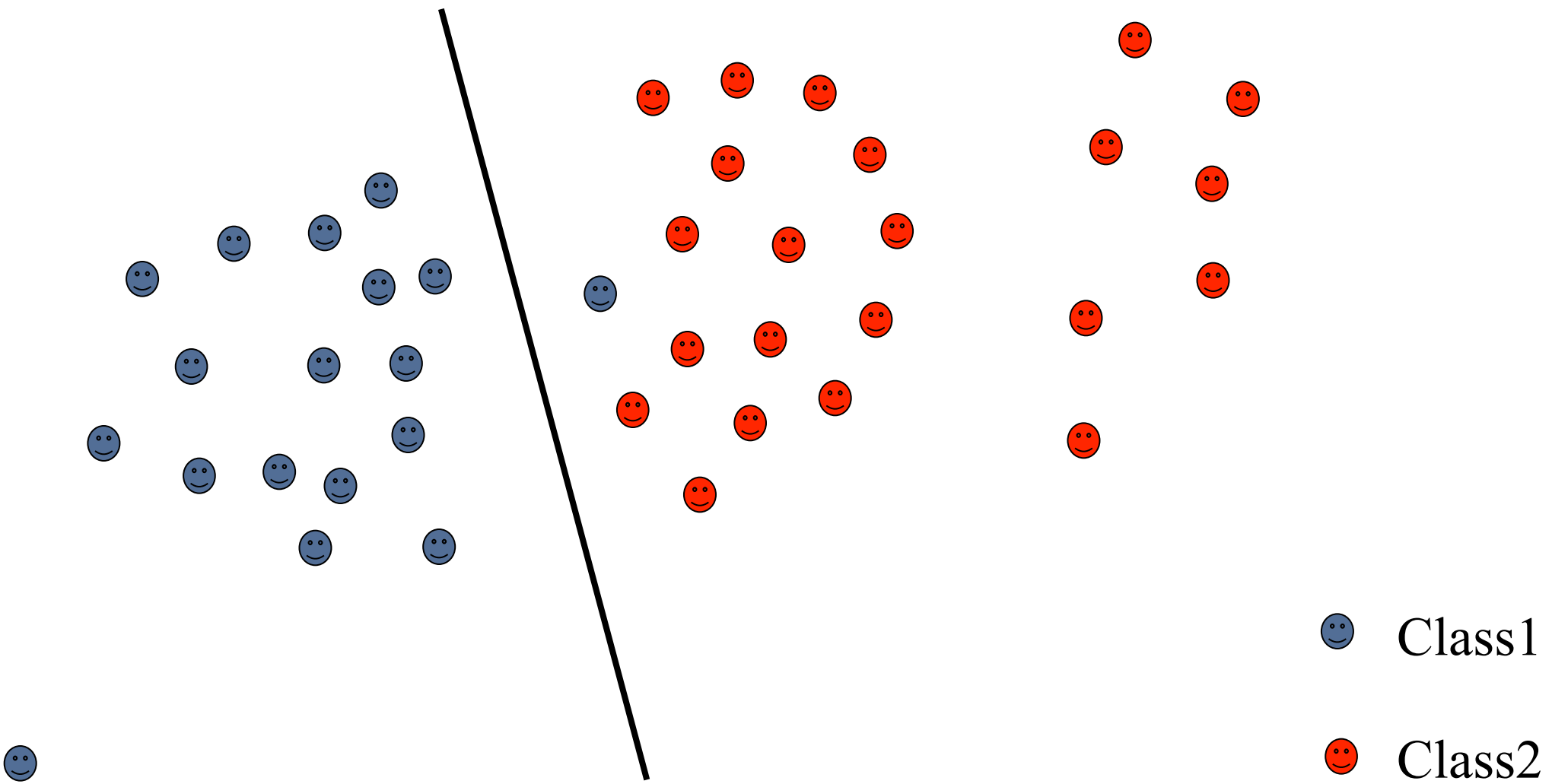
Bishop, PRML

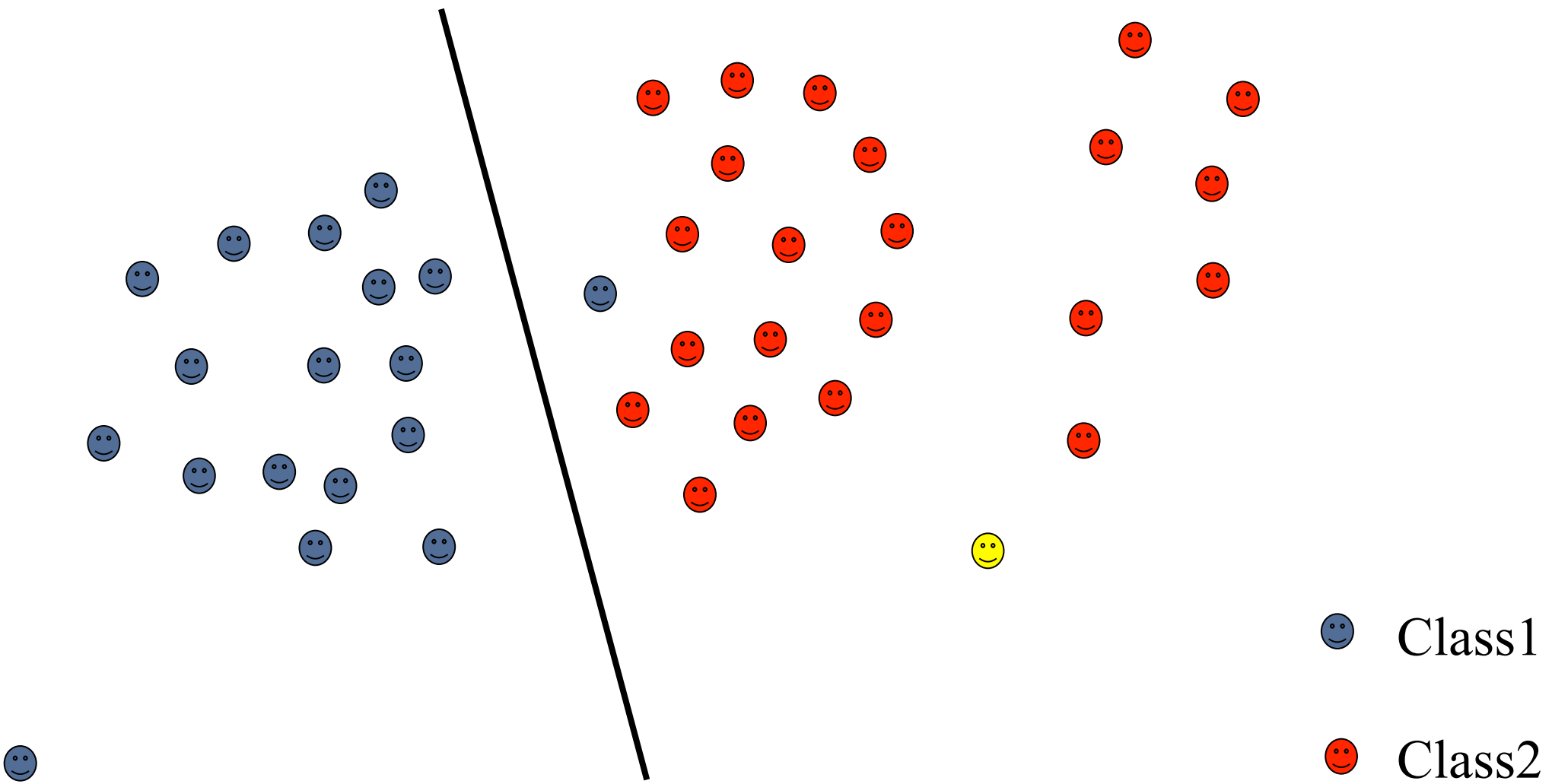
Unsupervised learning

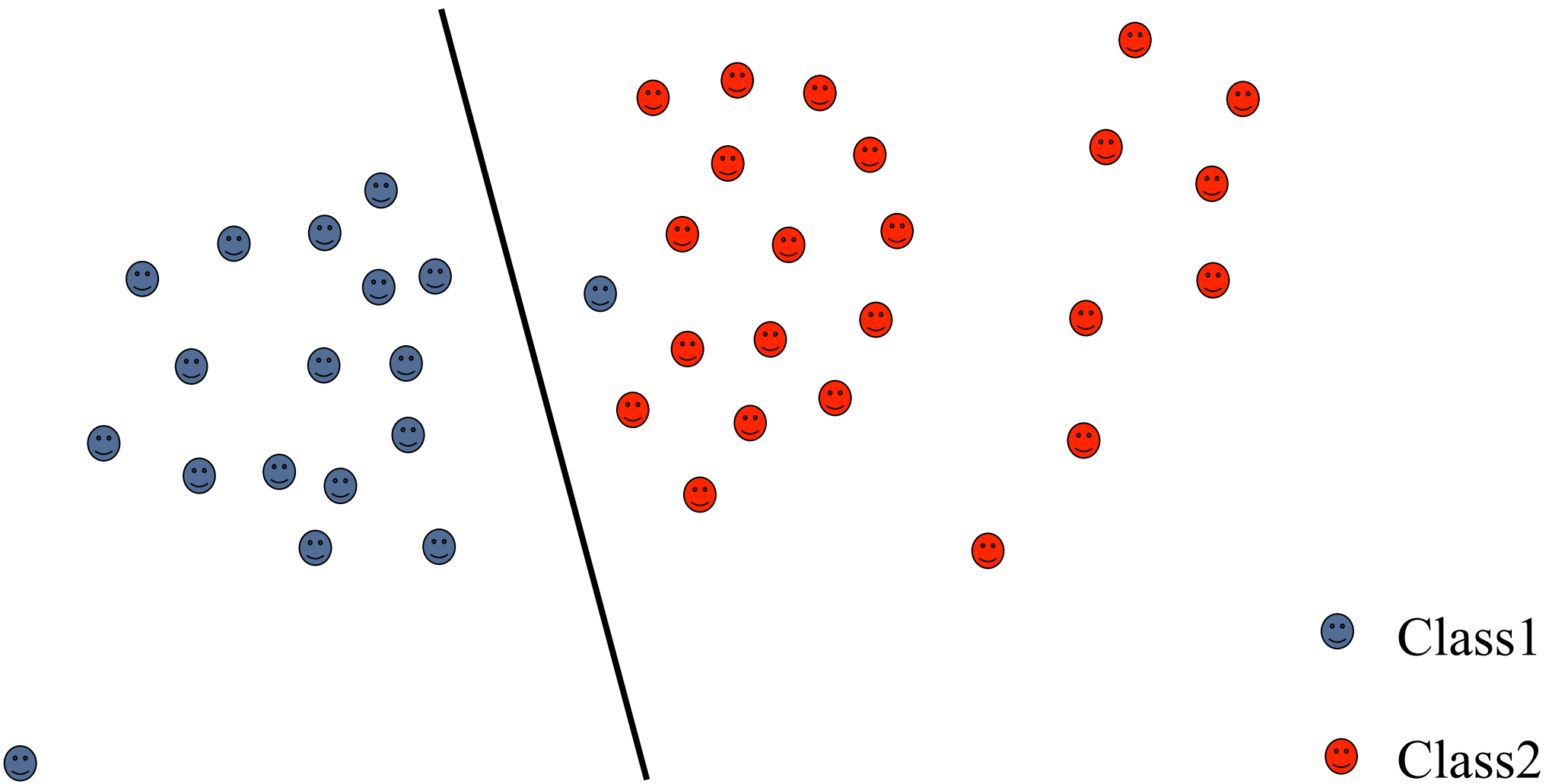
Supervised Classification:

Labels known



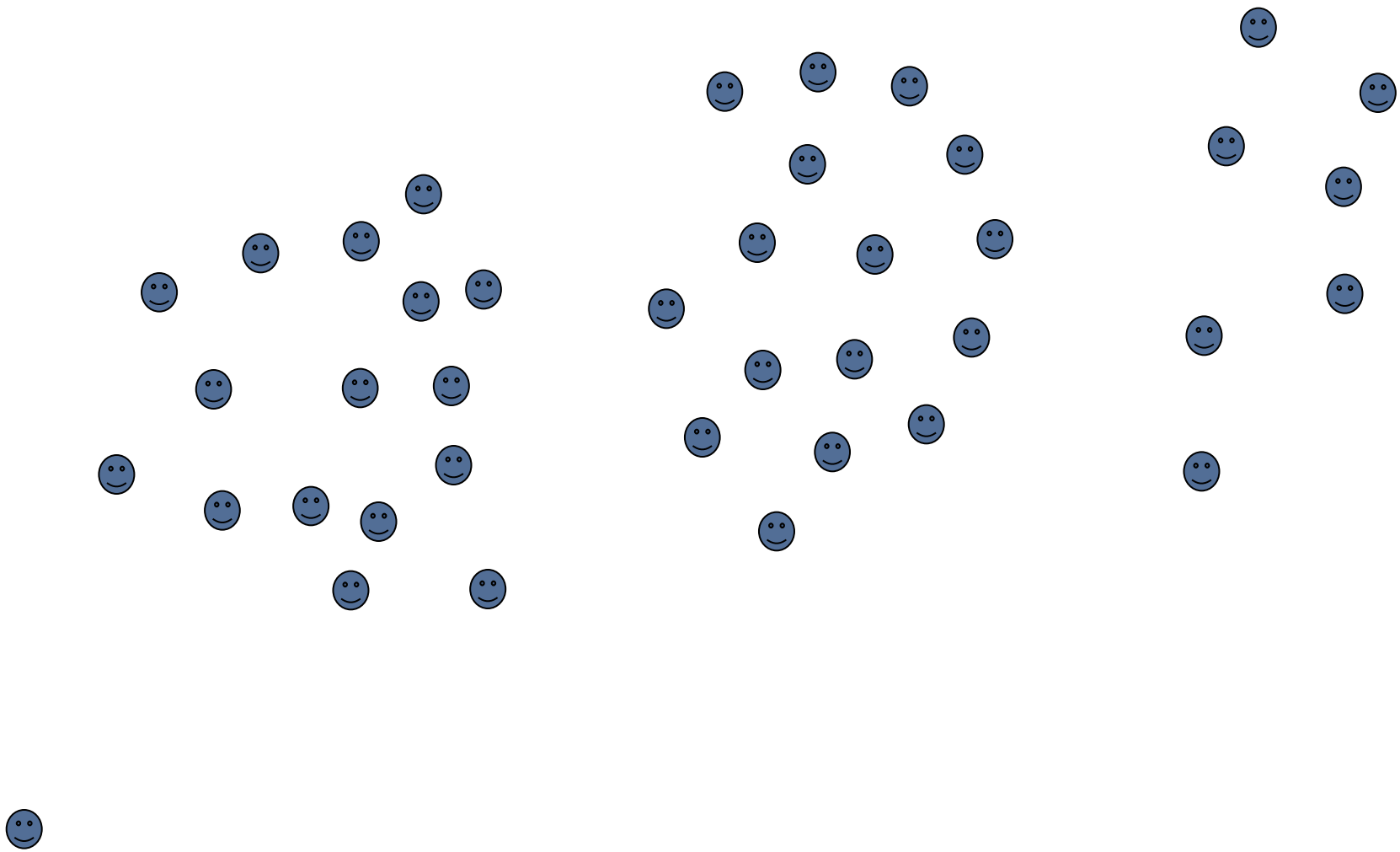

Class1
Class2


Class1
Class2


Class1
Class2

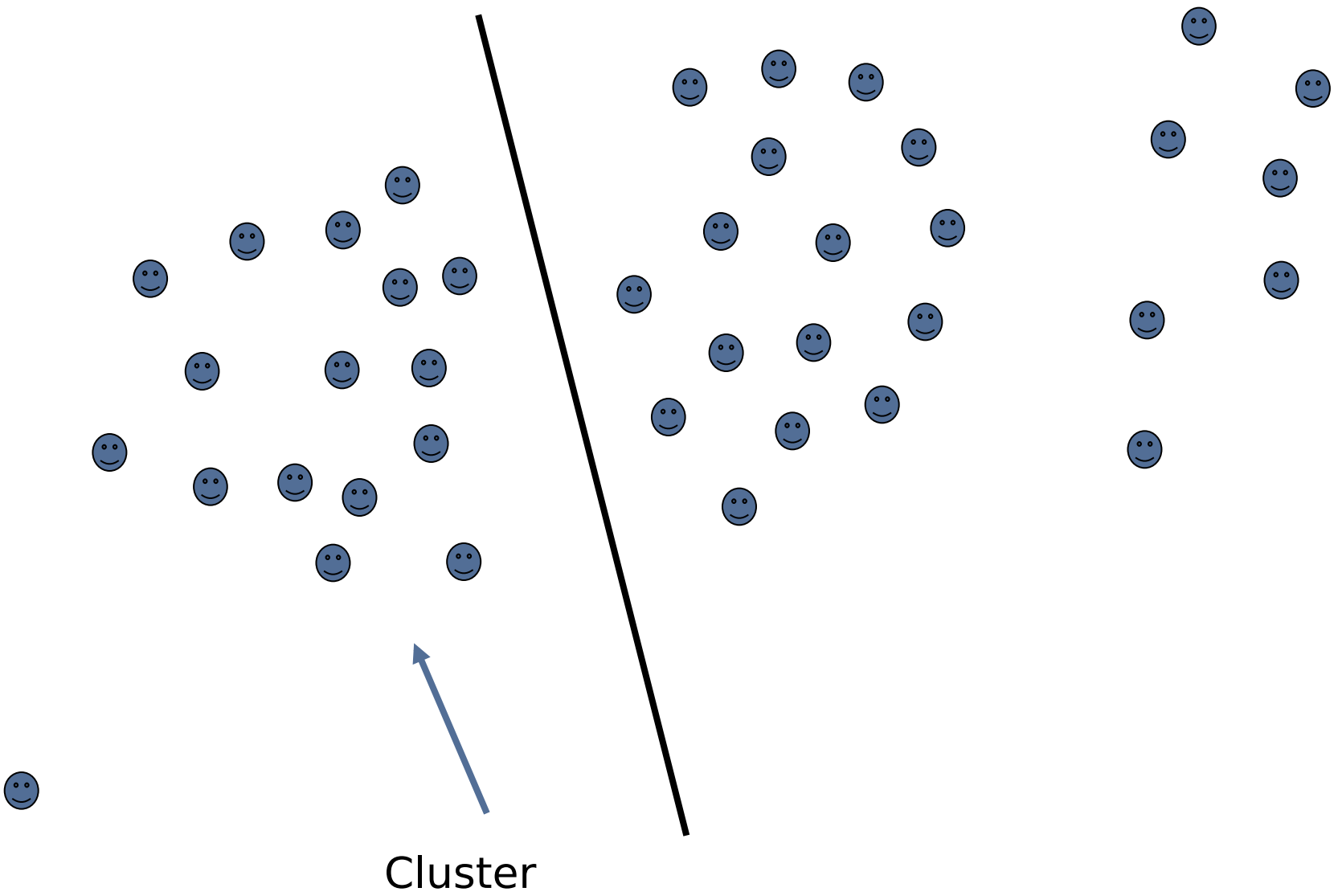
Clustering:

No labels!



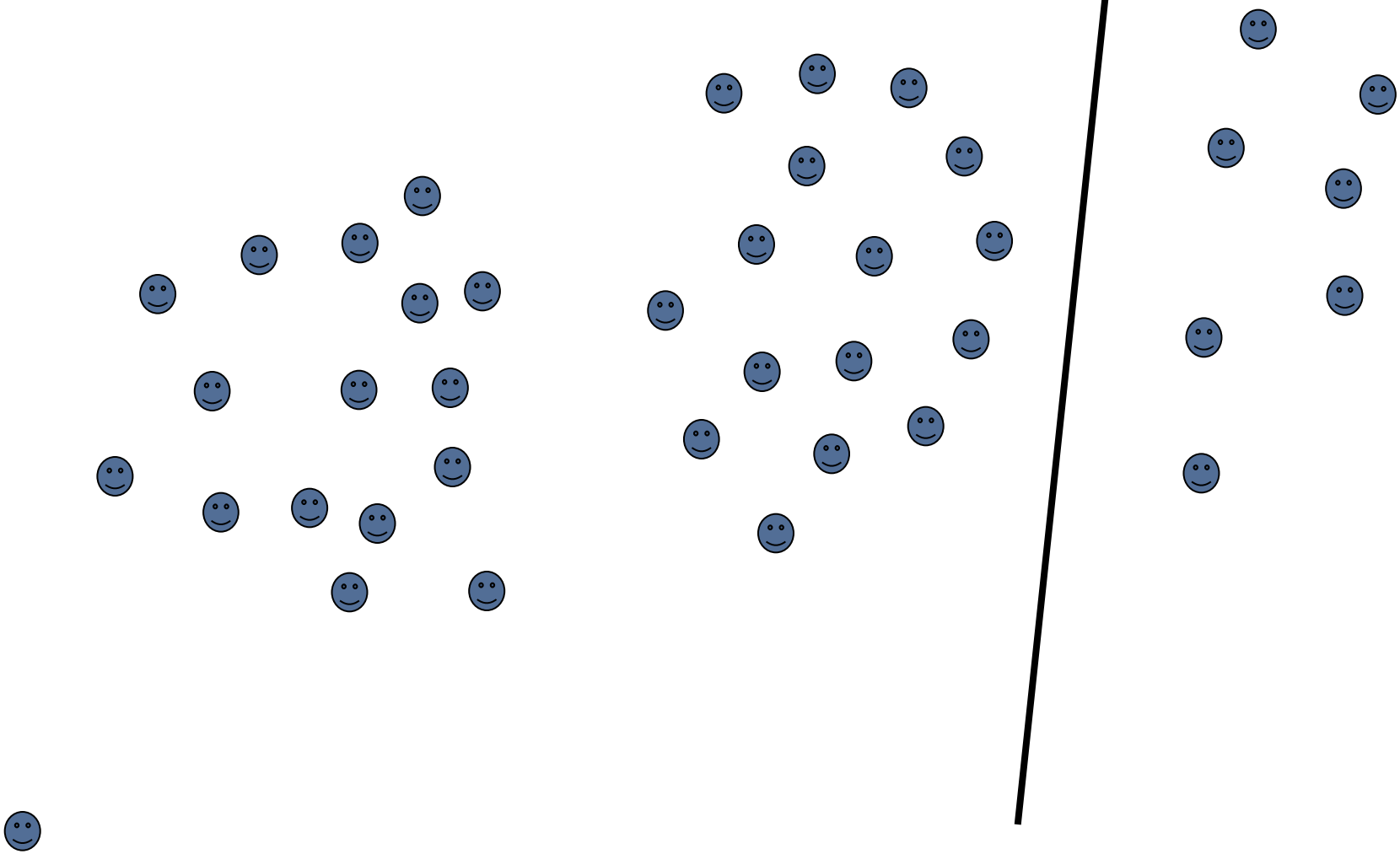
Clustering:

No labels!



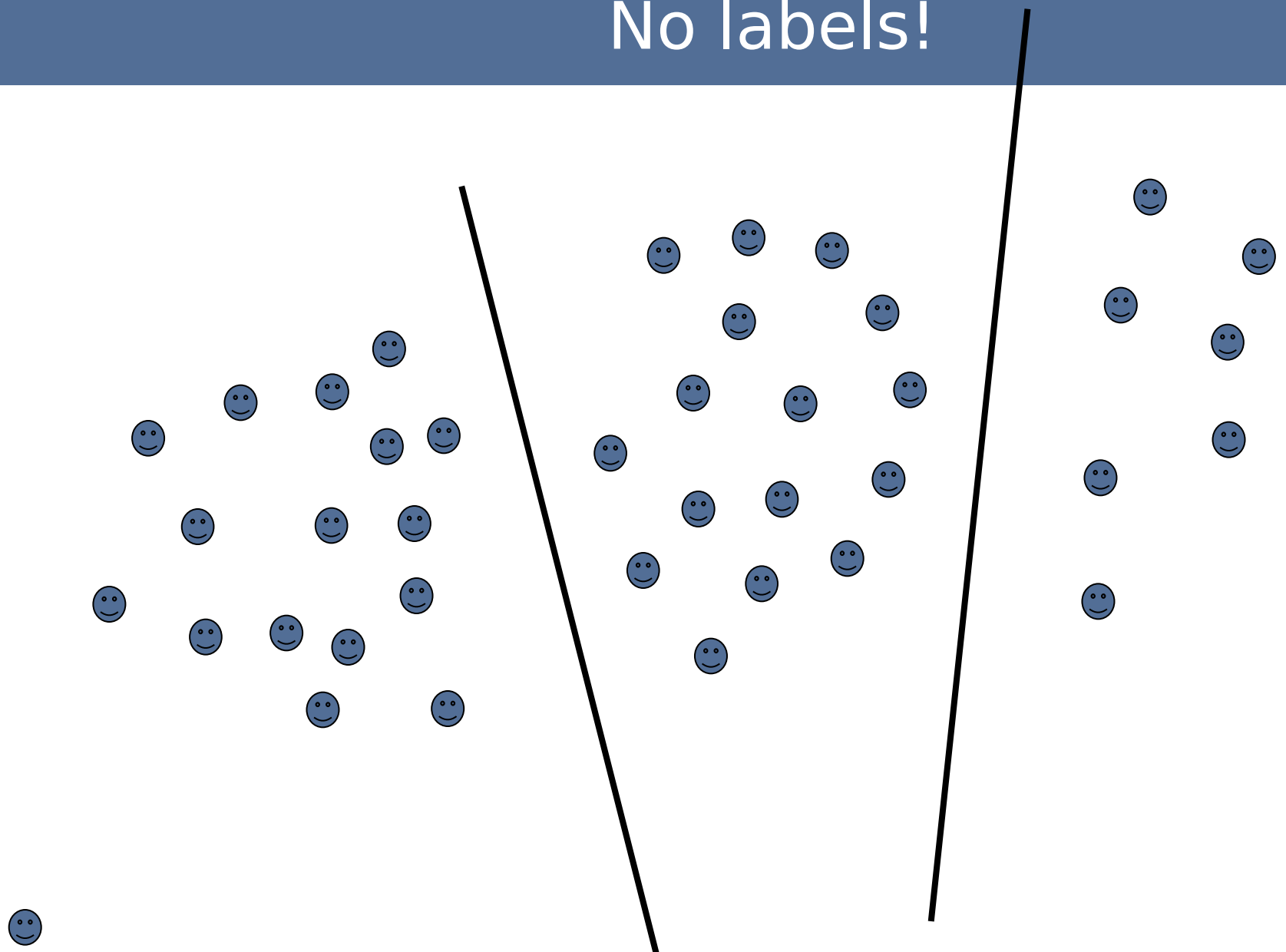
Clustering:

No labels!



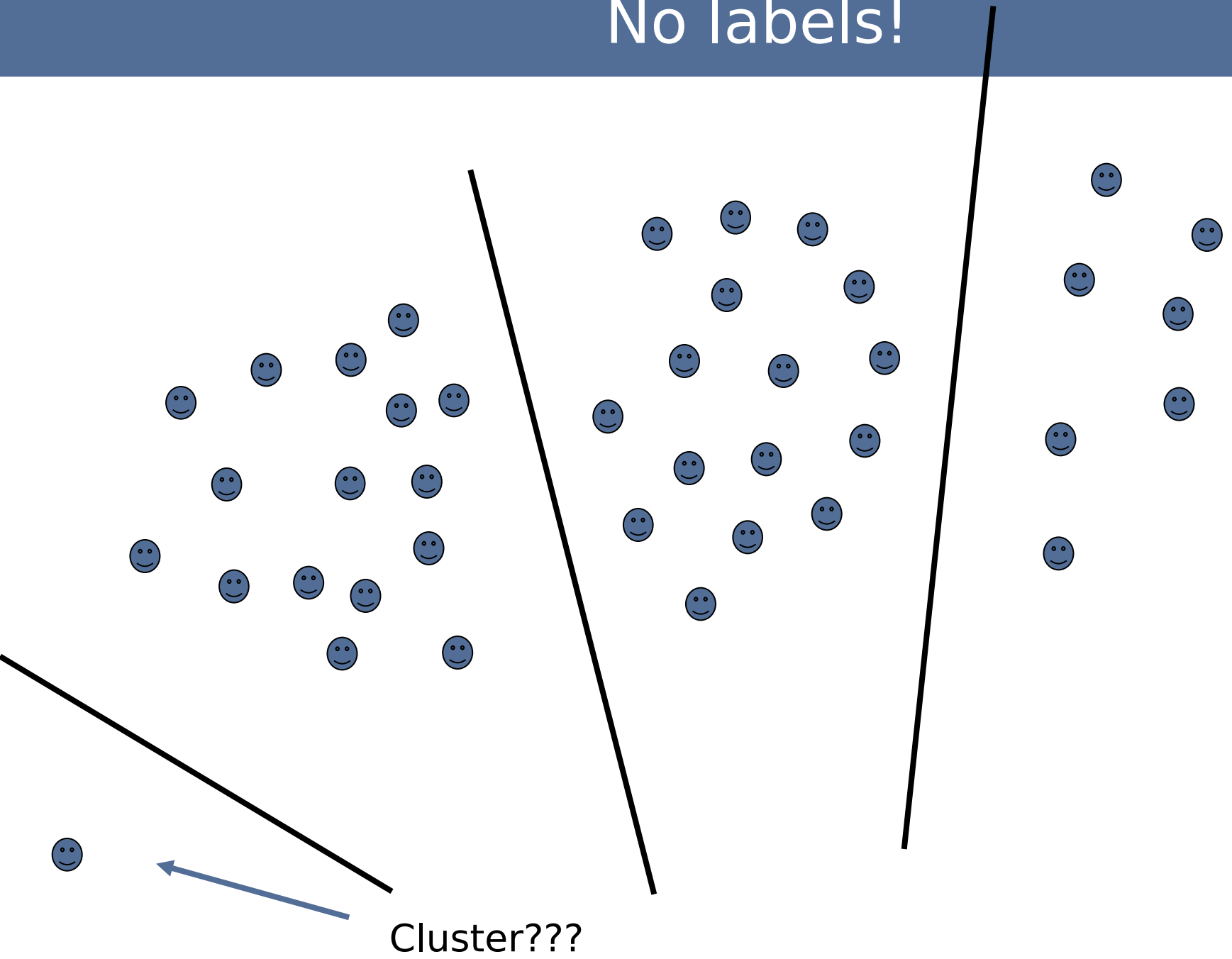
Clustering:

No labels!



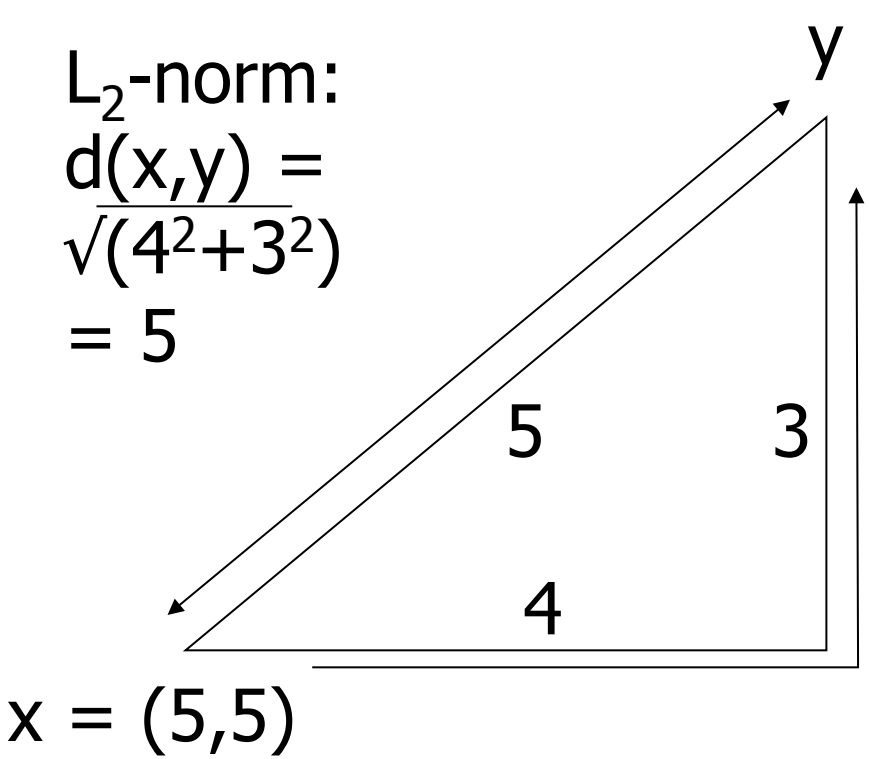
Clustering:

No labels!



Distance Measures

Euclidean Distances



Axioms of a Distance Measure

d is a *distance measure* if it is a function from pairs of points to the real numbers such that:

d(x,y) $\geq$ 0.

d(x,y) = 0 iff x = y.

d(x,y) = d(y,x).

d(x,y) $\leq$ d(x,z) + d(z,y) (*triangle inequality*).

Distances measures

L_1 distance (Manhattan distance)

$$d_1(\vec{x},\vec{y}) = \sum_{k=1}^{K} | x_k - y_k |$$

L_2 distance (Euclidian distance)

$$d_2(\vec{x},\vec{y}) = \sqrt{\sum_{k=1}^{K} | x_k - y_k |^2}$$

r^2 distance (Euclidian squared distance)

$$r^2(\vec{x},\vec{y}) = \sum_{k=1}^{K} | x_k - y_k |^2$$

L_∞ distance (maximum distance)

$$d_\infty(\vec{x},\vec{y}) = \max_k(| x_k - y_k |)$$

Example

- Calculate the distance of

$$\vec{x} = \begin{pmatrix} 3 \\ -1 \\ 0 \\ 3 \end{pmatrix} \qquad \vec{y} = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \end{pmatrix}$$

- Use all four distance measures introduced on the previous slide

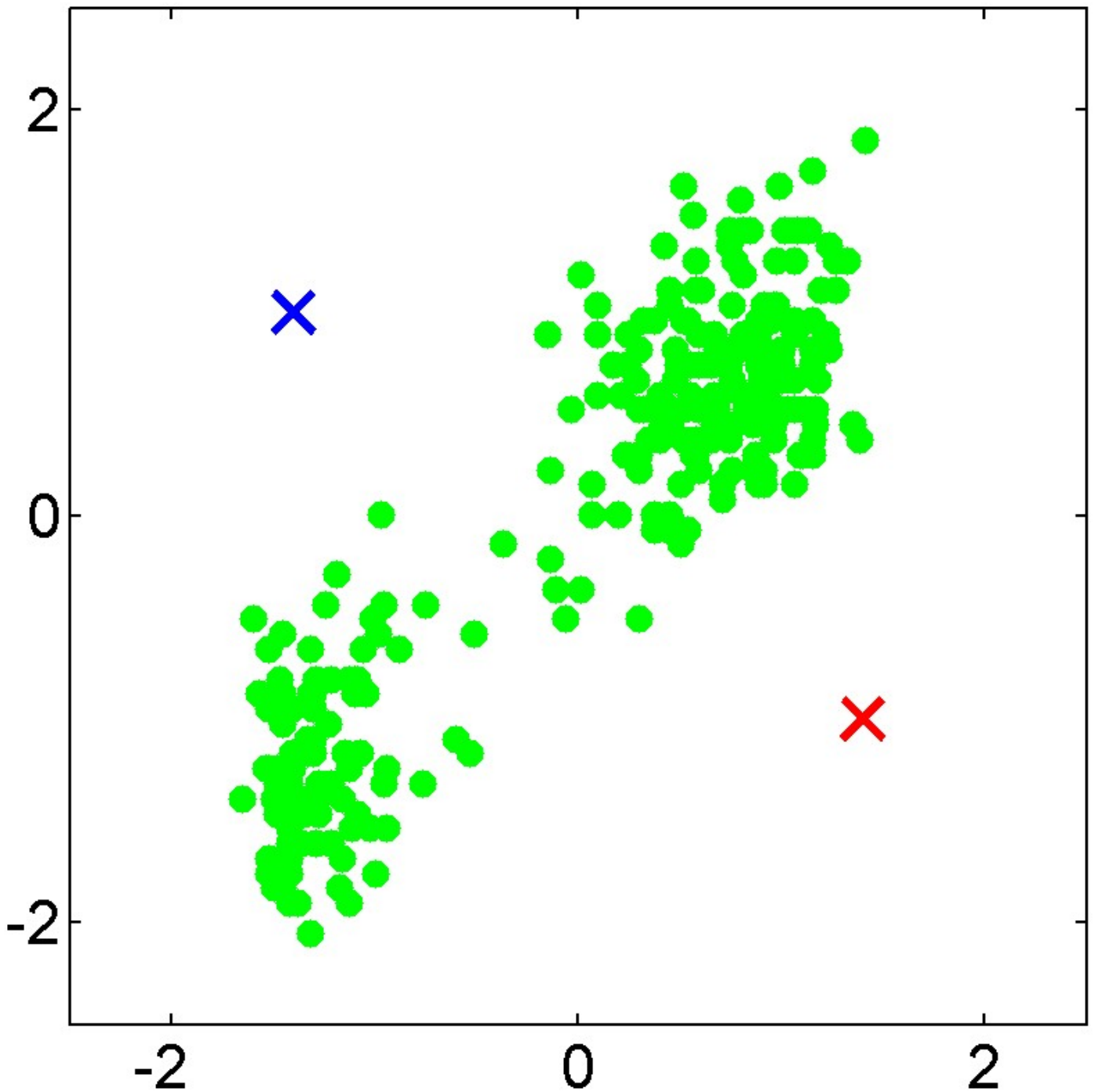
Other distance measures

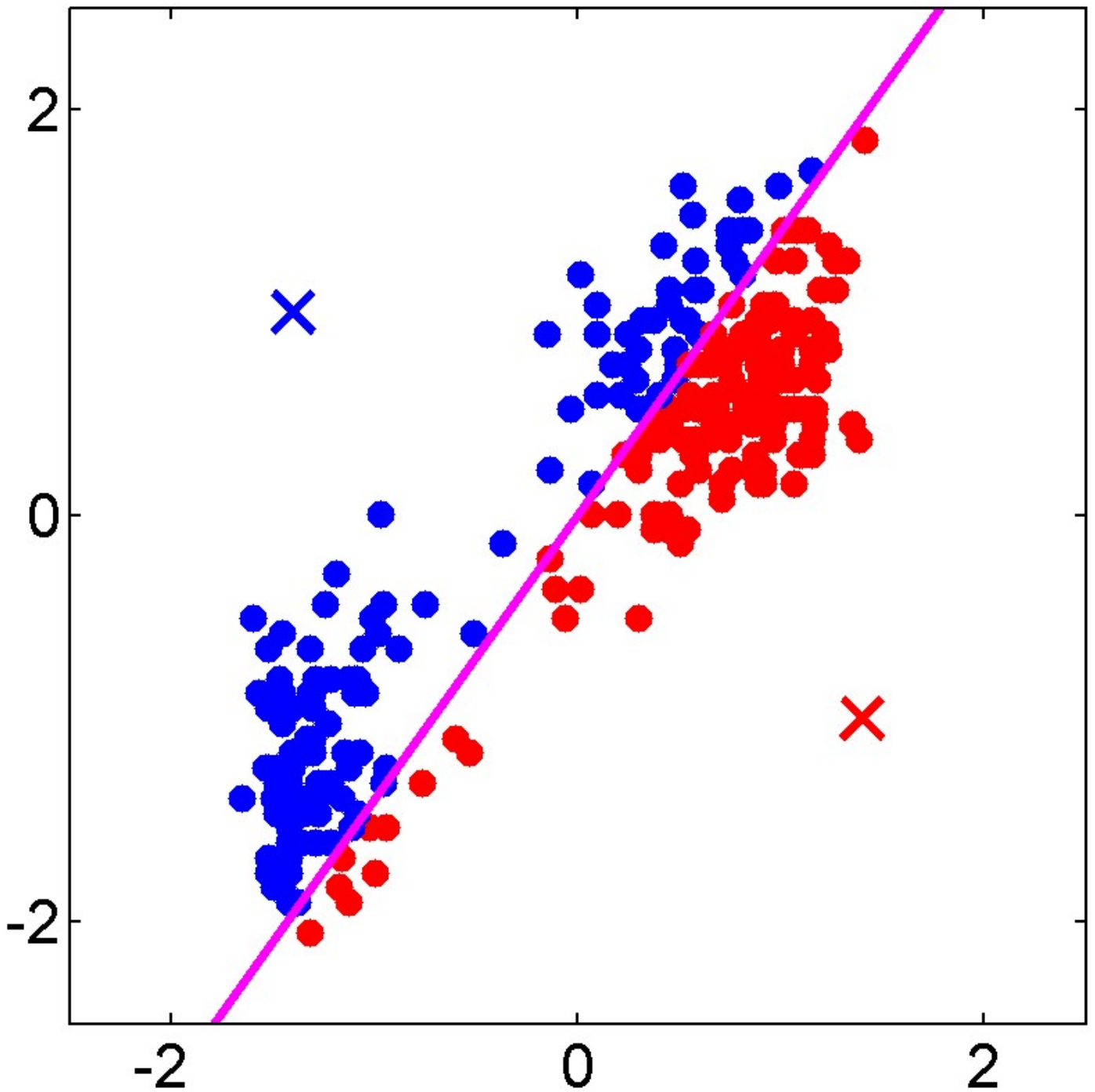
- Cosine*
- Edit distance
- Jaccard
- Kernels

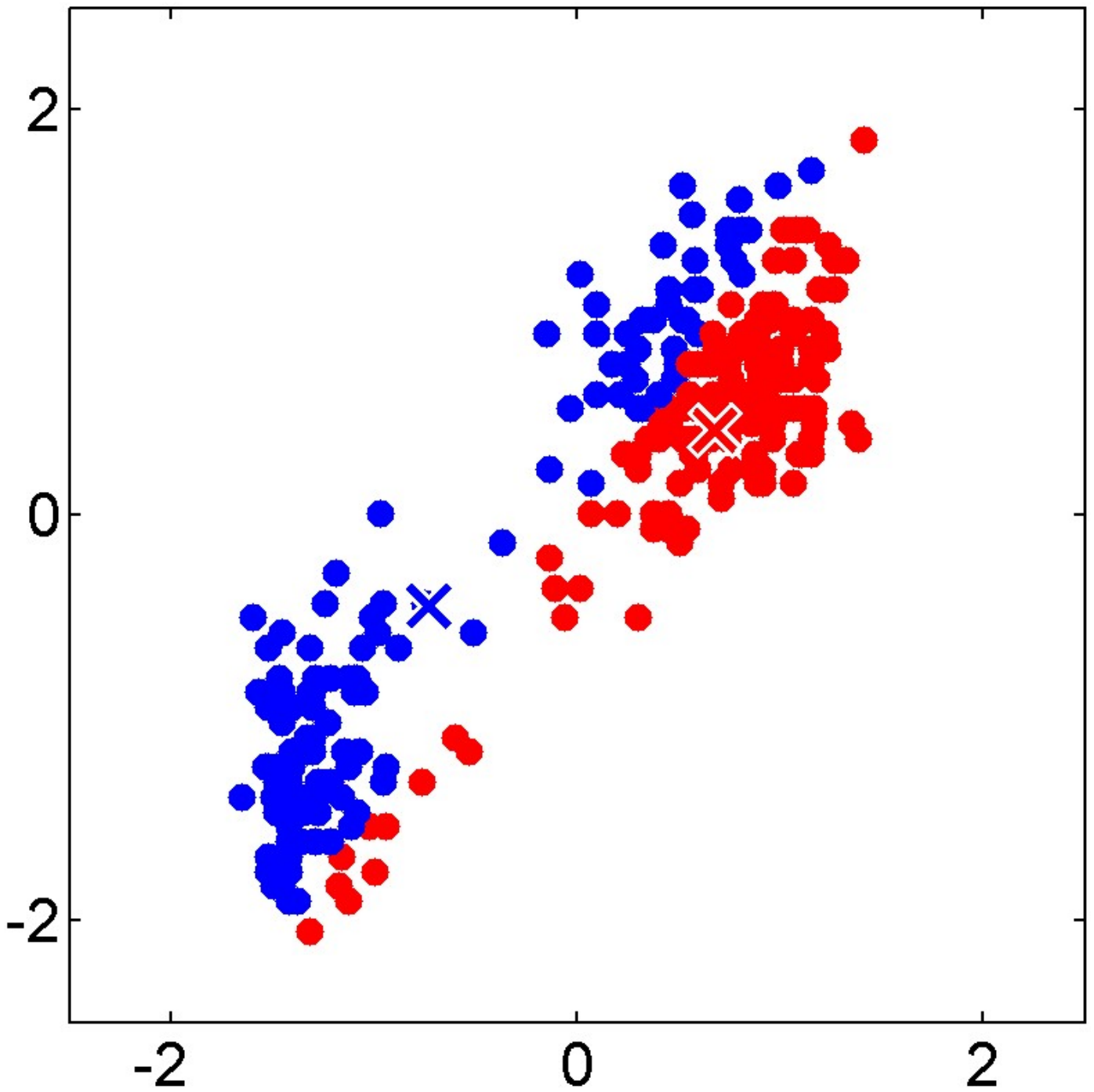
K-Means Clustering

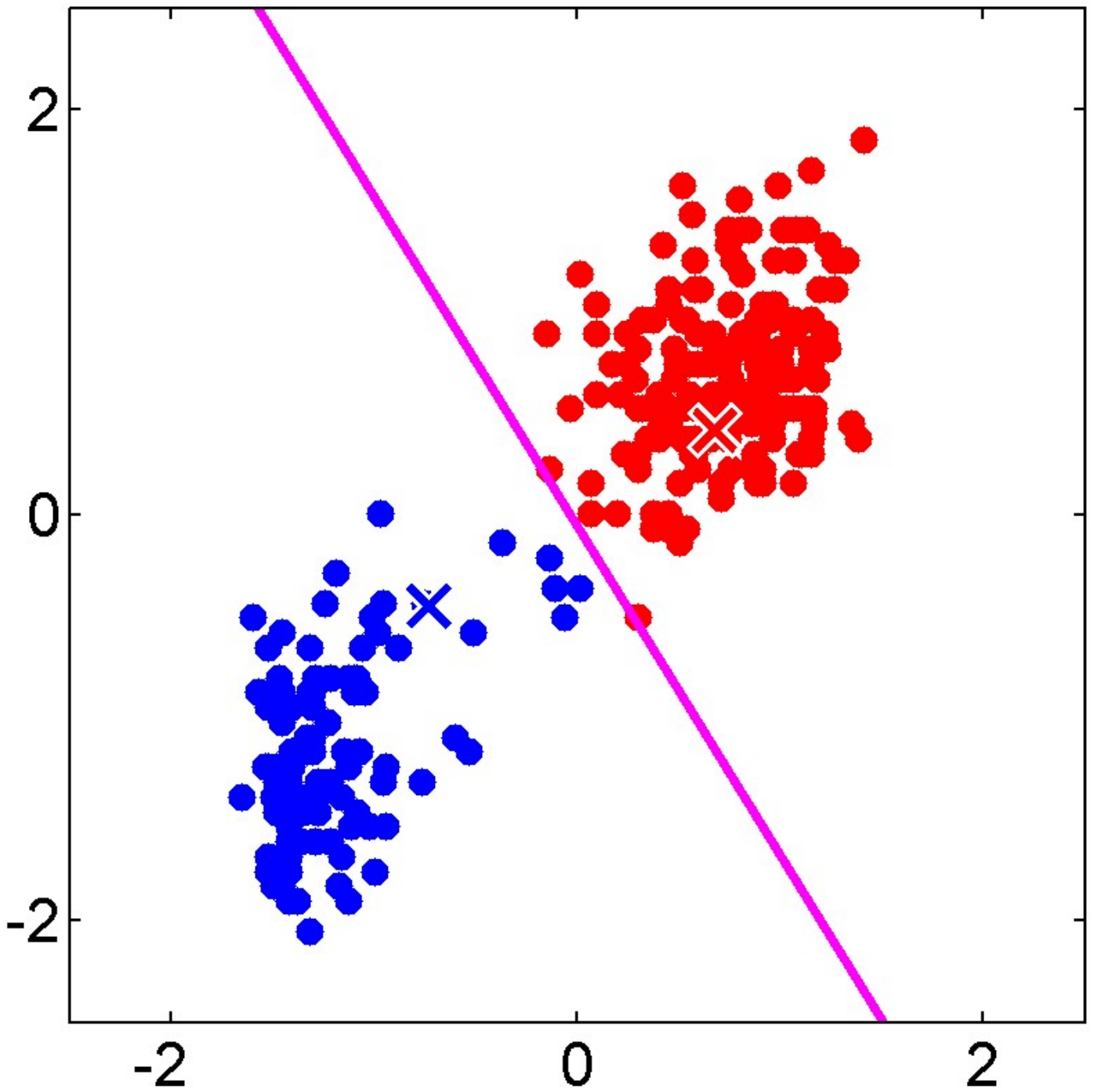
The K-means algorithm

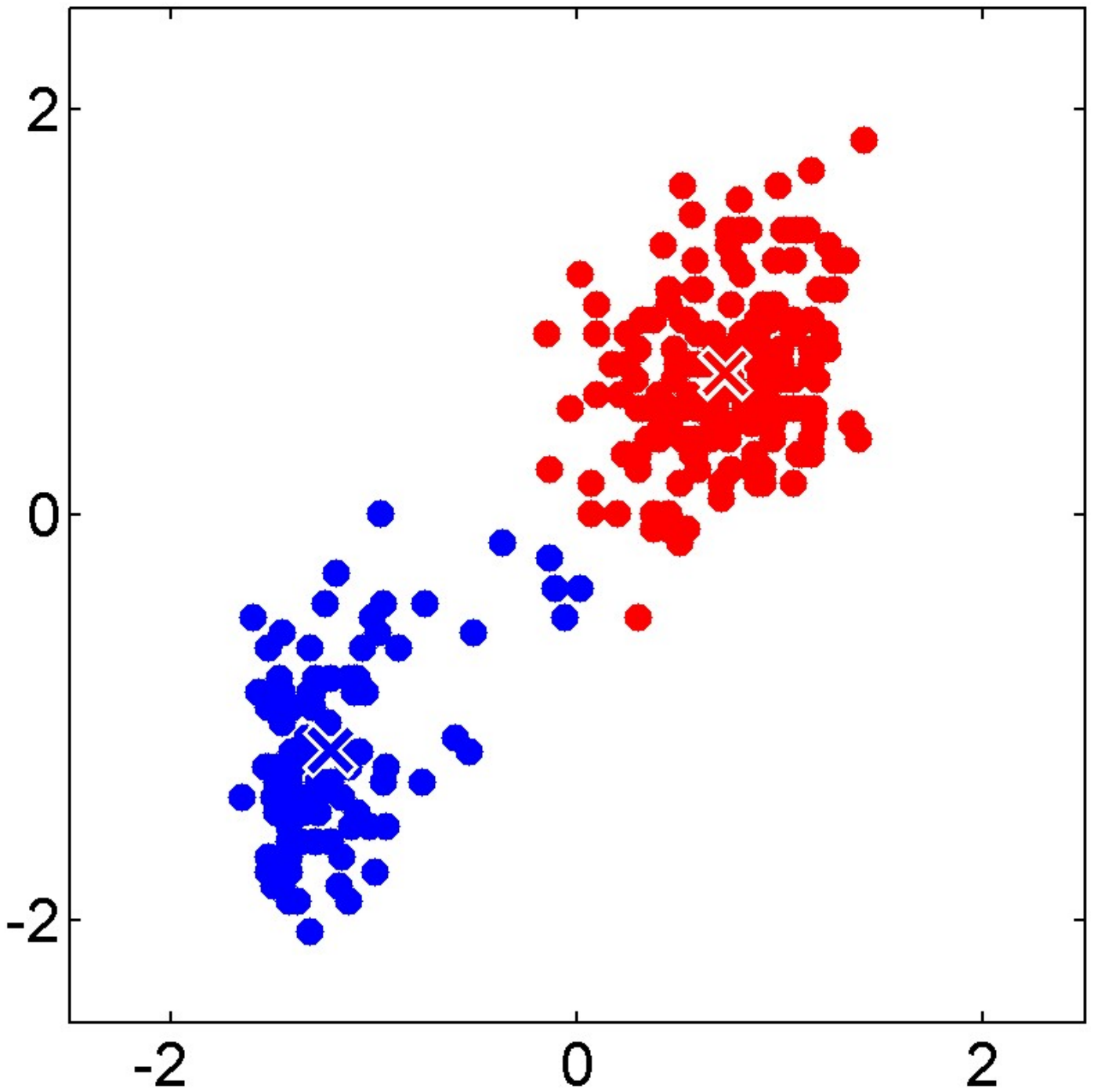
1. For each cluster, decide on a mean
2. Assign each data point to the nearest mean
3. Recalculate means according to assignments
4. If mean changed go back to step 1

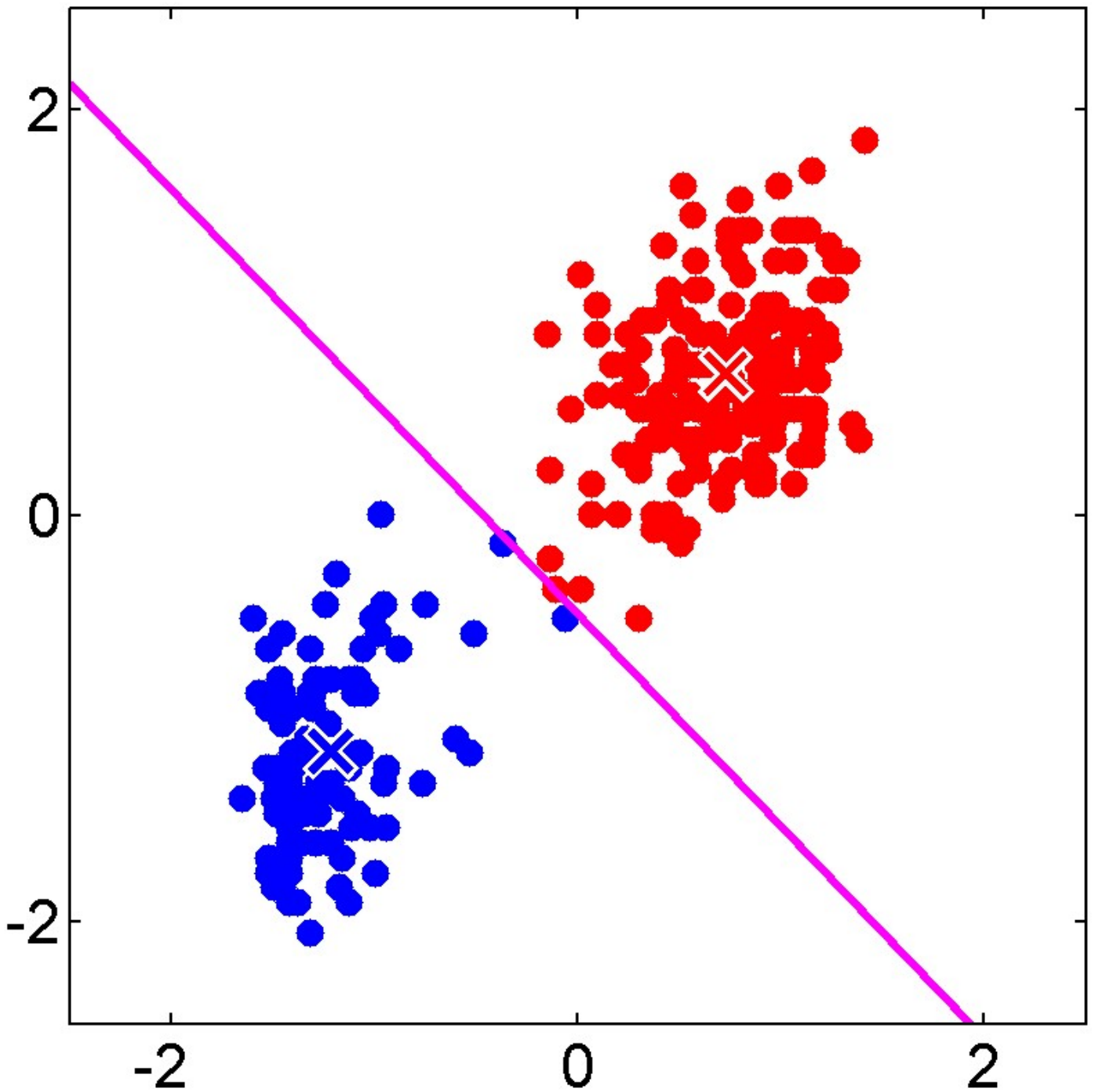

2
0
-2
-2
0
2

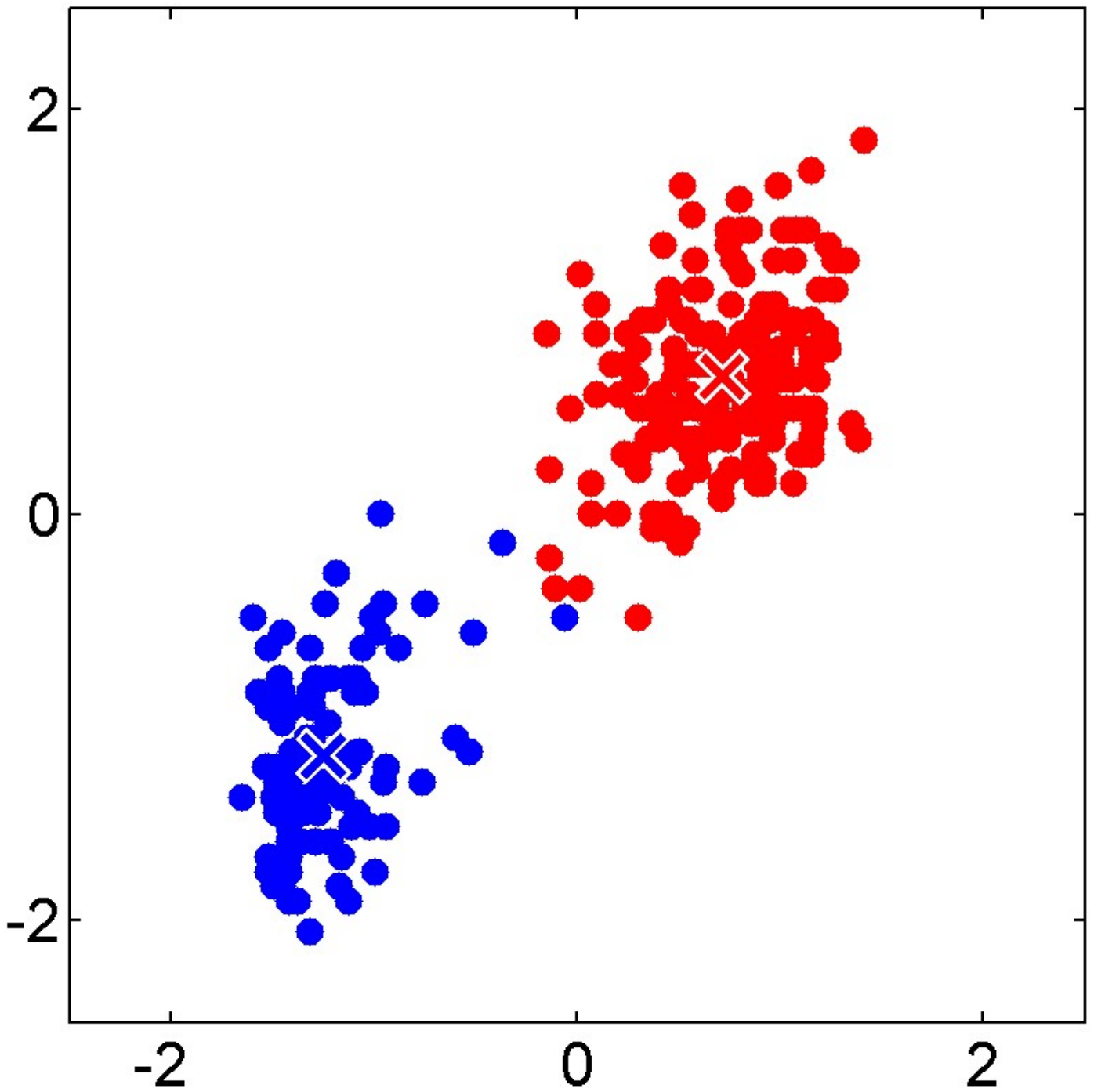

2
0
-2
-2
0
2

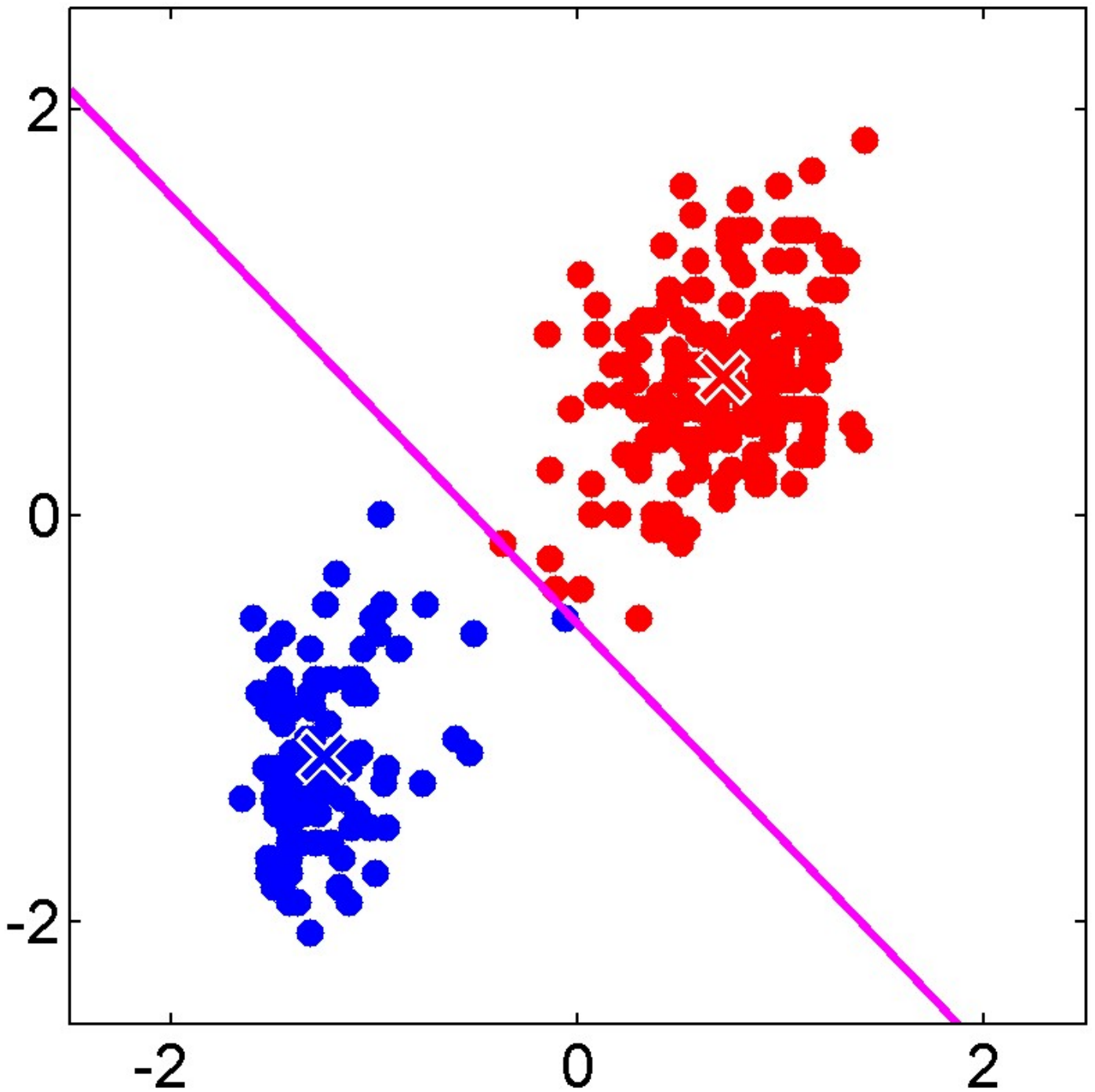

2
0
-2
-2
0
2

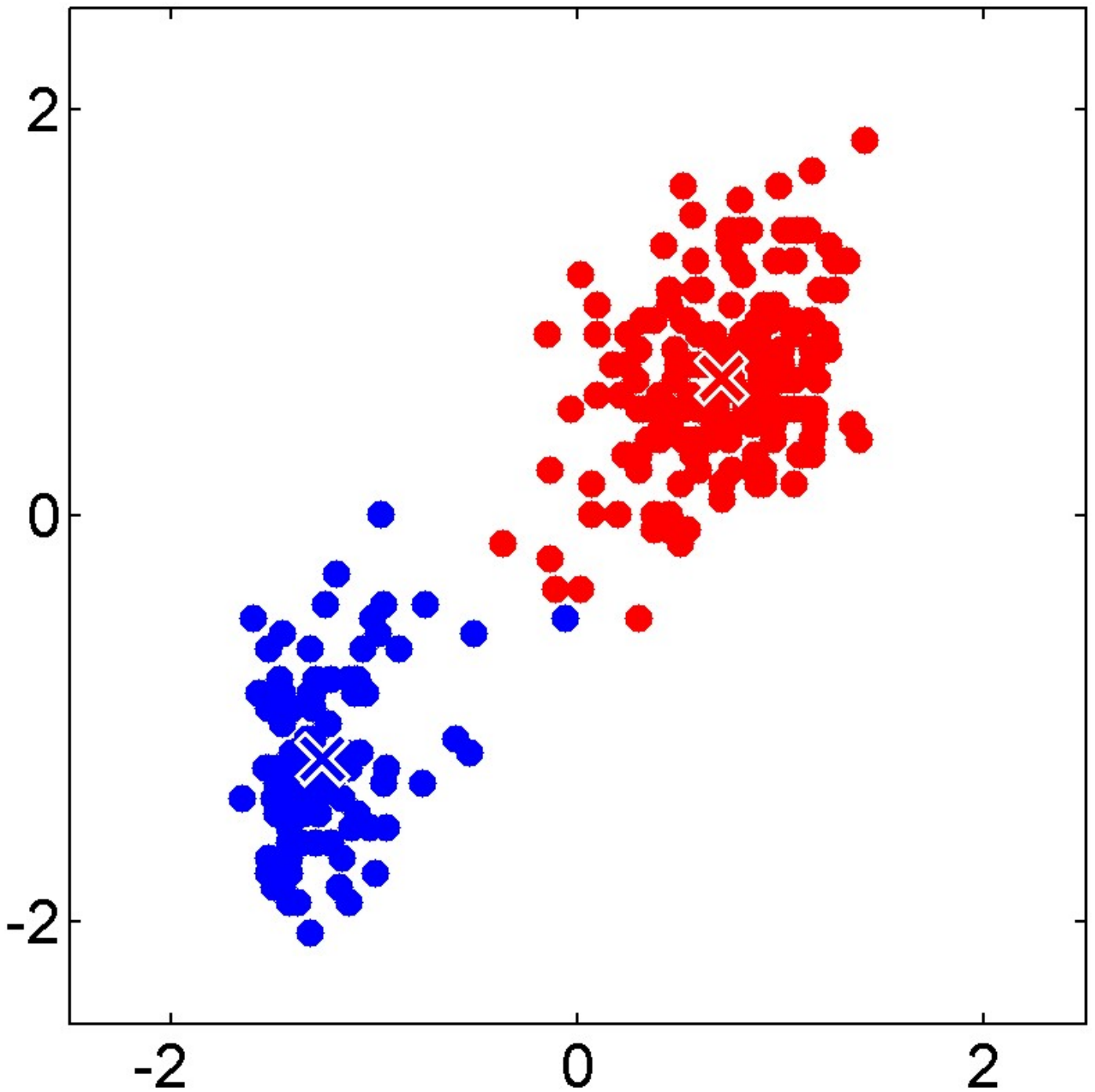

2
0
-2
-2
0
2


2
0
-2
-2
0
2


2
0
-2
-2
0
2


2
0
-2
-2
0
2


2
0
-2
-2
0
2


2
0
-2
-2
0
2

Assigning points to clusters

$$r_{n,k} = \begin{cases} 1 \text{ if } k = \arg\min_j d\left(\vec{x}_n, \vec{\mu}_j\right) \\ 0 \text{ otherwise} \end{cases}$$

$\vec{x}_n$: n-th training sample (vector)

$\vec{\mu}_j$: mean of the j-th cluster

$d\left(\vec{x}_n, \vec{\mu}_j\right)$: distance (your choice, e.g. L_2)

Example

$$r_{n,k} = \begin{cases} 1 \text{ if } k = \arg\min_j d\left(\vec{x}_n, \vec{\mu}_j\right) \\ 0 \text{ otherwise} \end{cases}$$

See white board

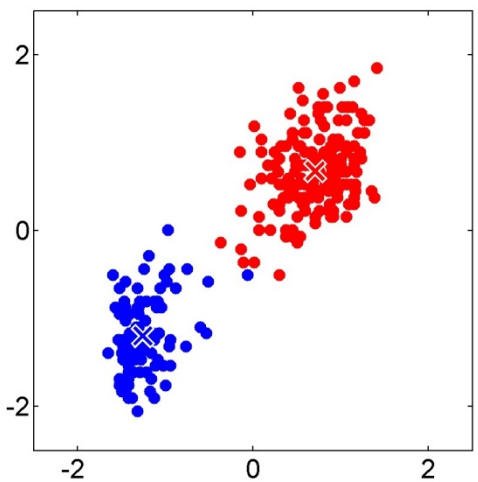
Update mean

$$\vec{\mu}_k = \frac{\sum_{n=1}^{N} r_{n,k} \vec{x}_n}{\sum_{n=1}^{N} r_{n,k}}$$

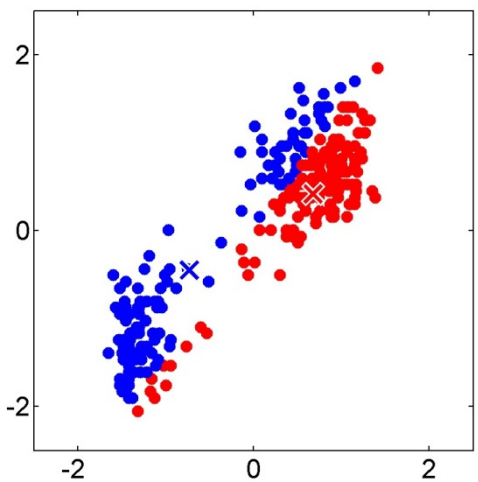
Interpret the denominator

Loss Function: Distortion Measure

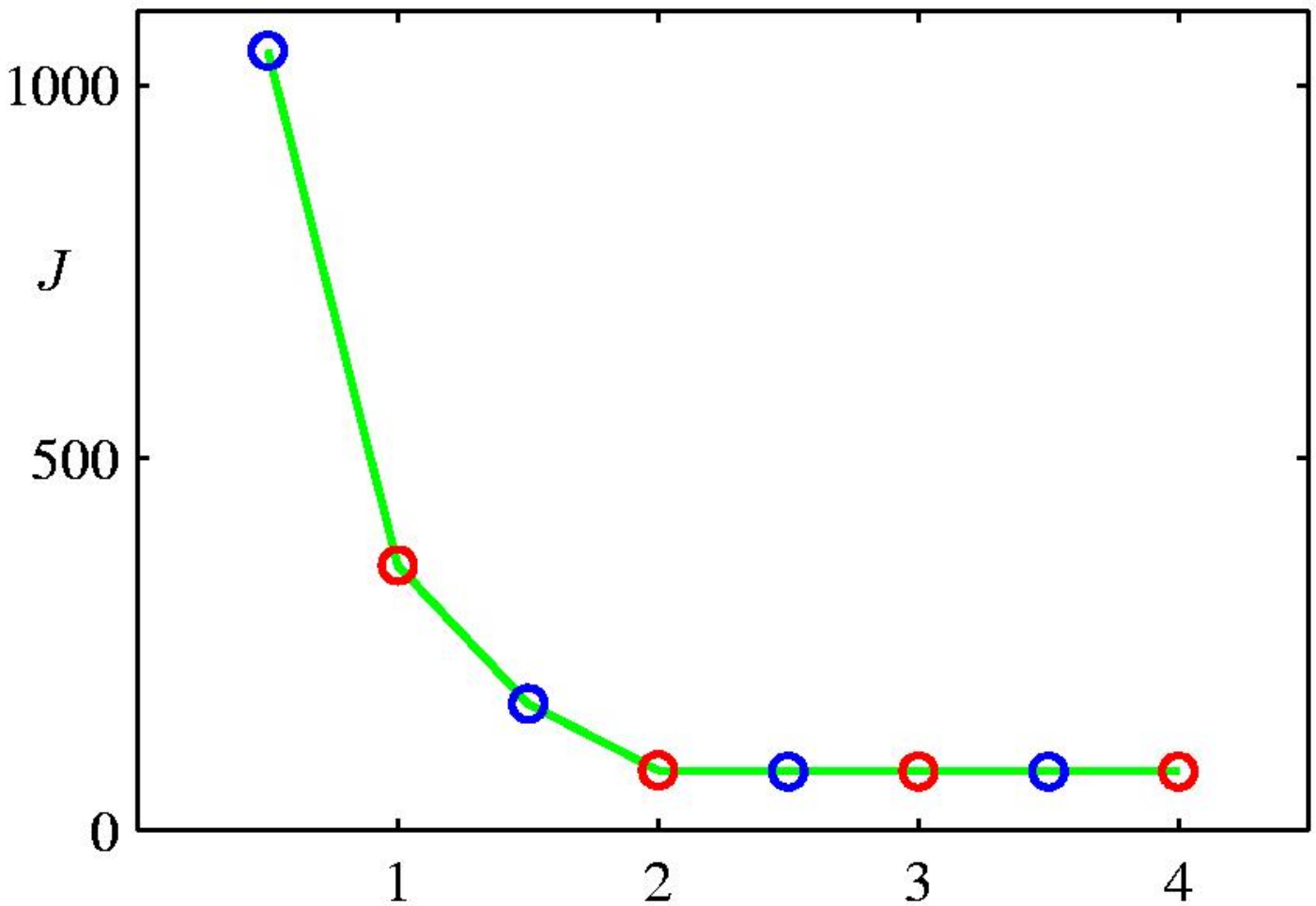
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} d(x_n, \mu_k)$$



Which of these has the smaller J?



Distortion Function after each iteration



How to initialize K-Means

- Converges to local optimum
- Outcome of clustering depends on initialization
- Heuristic:
  pick K vectors from training data
  (being furthest apart)

One way to choose K:

The Variance Ratio Criterion

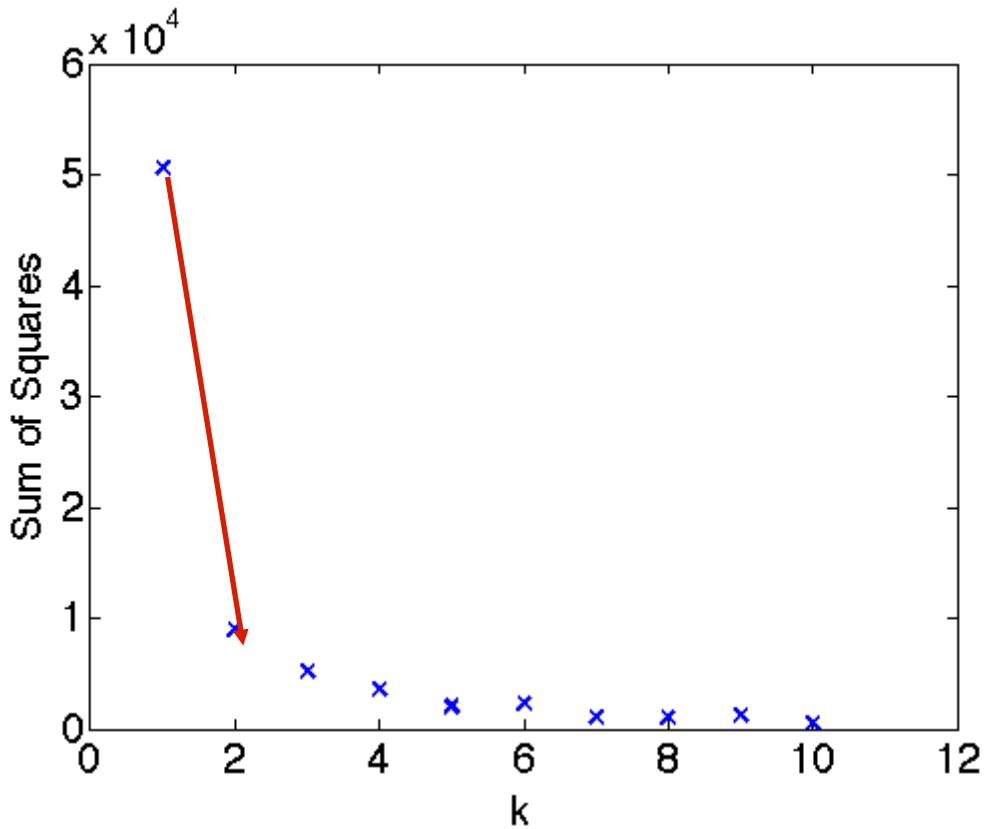
How to determine K

What about picking K such that J becomes as small as possible?

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} d\left(x_n, \mu_k\right)$$

How to determine K

- For K=N the distortion J=0
- Solution: find large jump



The Variance Ratio Criterion (VRC)

If we use squared distance (r^2):
$SS_{Total} = SS_{Between} + SS_{Within}$

SS_{Within} is also known as SS_{Error}

1. Compute VRC (F score) for each k:

$$VRC(k) = \frac{SS_B}{k-1} / \frac{SS_W}{n-k}$$

2. Choose K such that $(VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1})$ is minimized.

Slide from the Statistics with R course

$$SS_{Total} = SS_{Treatment} + SS_{Error}$$

$$F = \frac{\frac{\sum n_{in\ a\ group} * (\bar{x}_{group} - \bar{x}_G)^2}{k-1}}{\underbrace{\frac{\sum (x_i - \bar{x}_{group})^2}{n_{total} - k}}_{MS_{error}}}$$

$SS_{Treatment\ /\ Numerator}$

$df_{Treatment\ /\ Numerator}$

$SS_{Error\ /\ Residual\ /\ Denominator}$

$df_{Error\ /\ Residual\ /\ Denominator}$

k = number of categories / groups

n_{total} = total number of scores

$\bar{x}_G$ = grand mean

$n_{in\ a\ group}$ = number of scores in each group

$\bar{x}_{group}$ = mean of the group the score comes from

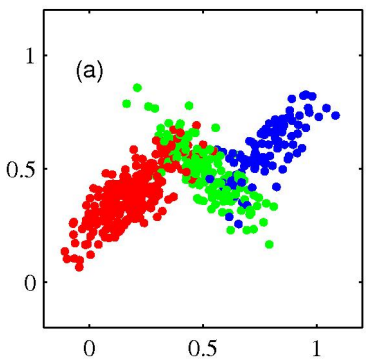
x_i = individual score

Other Clustering Algorithms

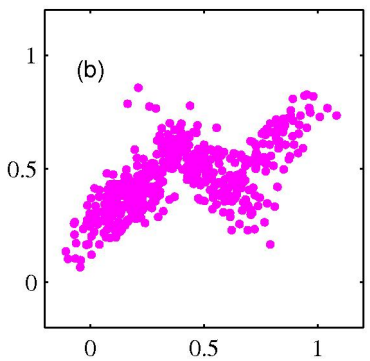
Soft clustering (e.g. Expectation-Maximization)

No strict assignment to a cluster

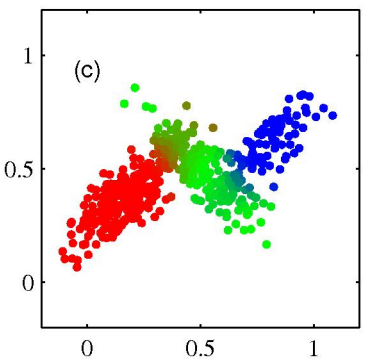
Just probabilities



Original data
Overlapping class regions



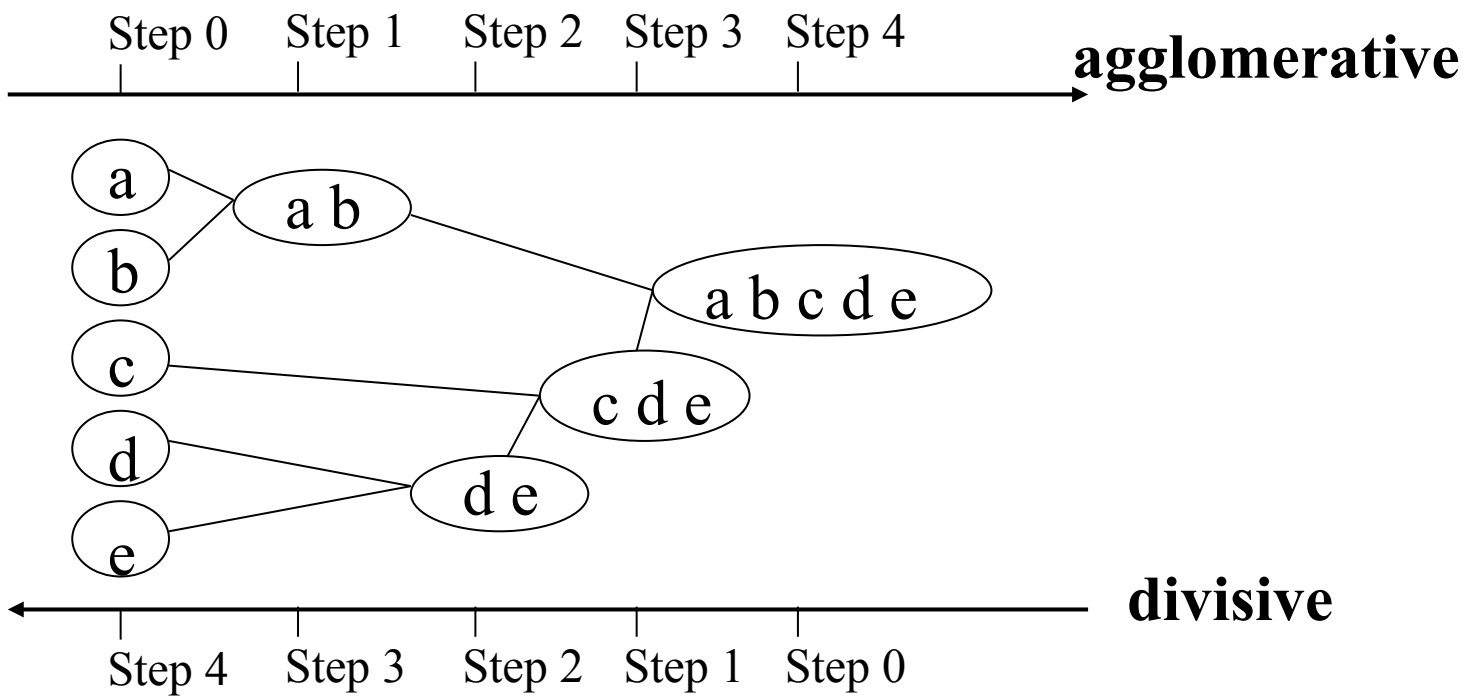
No class
information



Soft clustering

Hierarchical Clustering (Brown Algorithm)

Organize cluster in a hierarchy



The Exchange Algorithm

start with some initial mapping $w \rightarrow g_w$
  for each word w of the vocabulary do
    for each class k do
      tentatively exchange word w from class g_w to class k and update counts
      compute perplexity for this tentative exchange
    exchange word w from class g_w to class k with minimum perplexity
do until stopping criterion is met

g_w : calls of word w

Application to Named Entity Tagging

Possible features of words

- Frequency
- TF-IDF
- Stop wording?
- Stemming?

Idea

Cluster words together that have similar neighbours

Minimize perplexity on training test

Example clustering

Cluster	Example members
1	Groß, Rau, Müller, Zimmermann, Frei, Becker, Möllemann, Schmidt
2	Düsseldorf, Berlin, München, Köln, Stuttgart, Hannover, Hamburg
3	nahmen, macht, zeigt, gleichen, bringt, biete, machte, sorgt, enthält

Class labels as features (1/2)

Training

Word	Class label	Tag
Düsseldorf	C2	City
is	X	O
the	X	O
capital	X	O
of	X	O
NRW	X	O

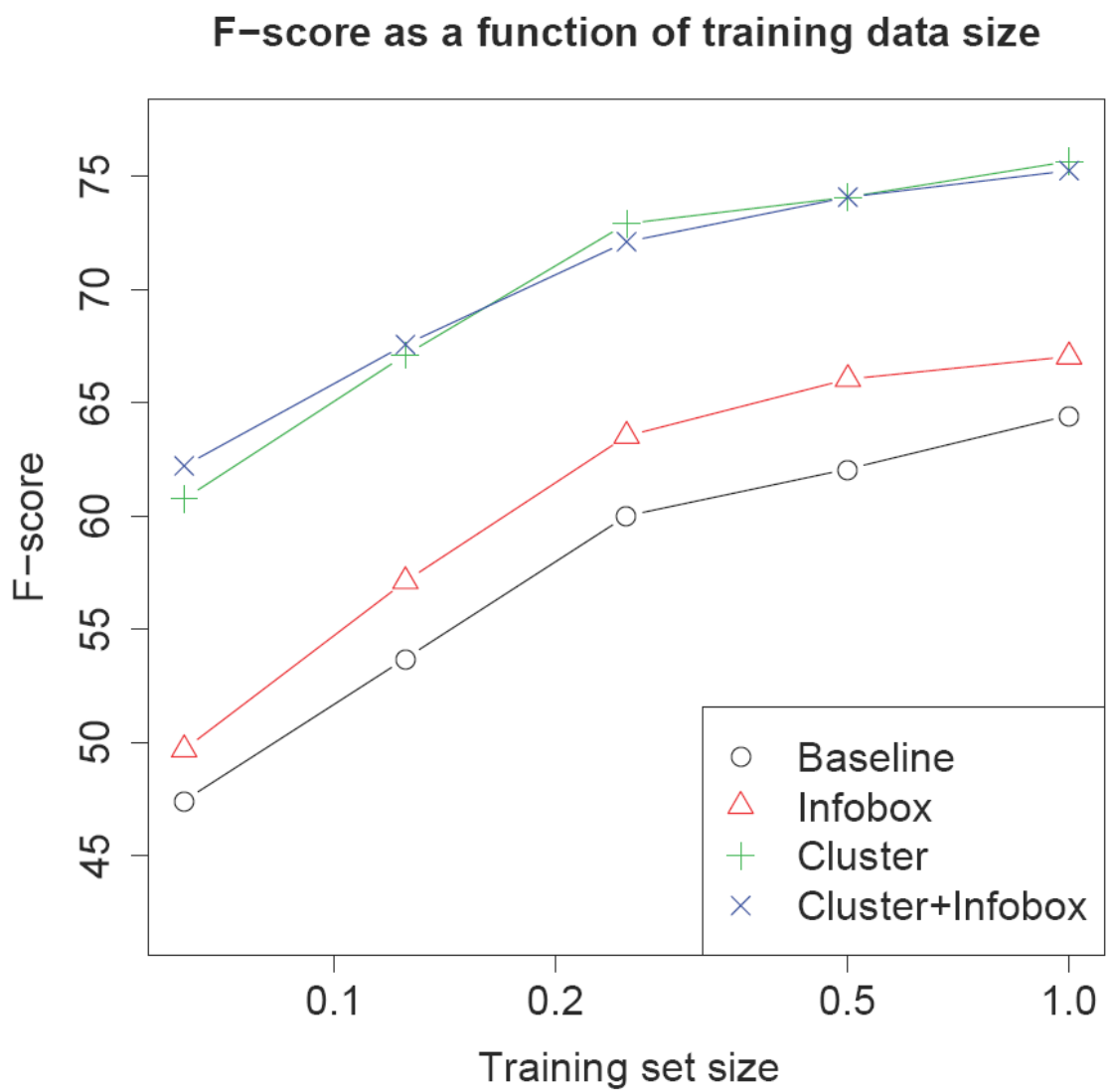
Class labels as features (2/2)

Testing

Word	Class label	Tag
The	X	O
Hofbräuhaus	X	O
is	X	O
in	X	O
Munich	C2	???

How to tag if Munich is not in the training data?

Results


F−score as a function of training data size
F−score
45
50
55
60
65
70
75
0.1
0.2
0.5
1.0
Training set size
Baseline
Infobox
Cluster
Cluster+Infobox

Summary of topics

- Clustering: finding similar items
- Distance metrics
- The K-means algorithm
- The Variance Ratio Criterion
- Other clustering algorithms
- Application to named entity tagging