# Machine learning 2
# Evaluation of performances & Unsupervised classification

**Pierre Chainais**

centralelille
ÉCOLE CENTRALE DE LILLE

# Trajectory



1. Evaluation of performances in pattern recognition (PR)
   - Test set
   - Cross-Validation (CV)
   - Confusion matrix
   - ROC curve
2. Unsupervised classification
   - Clustering K-means
   - Examples of applications of K-means
     - Geyser Old Faithful
     - Génétique et tumeurs cancéreuses
     - Vectorial quantization of images
   - Clustering K-medoids
   - Examples of applications of K-medoids
     - Differences of perception between countries
   - Other approaches

1. Evaluation of performances in pattern recognition (PR)
   - Test set
   - Cross-Validation (CV)
   - Confusion matrix
   - ROC curve

2. Unsupervised classification
   - Clustering K-means
   - Examples of applications of K-means
     - Geyser Old Faithful
     - Génétique et tumeurs cancéreuses
     - Vectorial quantization of images
   - Clustering K-medoids
   - Examples of applications of K-medoids
     - Differences of perception between countries
   - Other approaches

Using some training samples, an estimate of the classification error estimated from this same training sample (a.k.a. *substitution method*) may suffer from some optimistic bias (underestimate).

The classifier has been defined to minimize the error rate on the training set: it may be more performant on this data set than on others: **how to accurately estimate the generalization error ?**

Different methods of evaluation have been defined:
**test set & cross-validation**.

# Methode of the test set (*holdout*)



## Trivial solution

Use another data set (not used for training): **the test set**
If the test set is of size $m$ and that $r$ classification errors occur, the test error estimate will be:

$$\hat{E} = r/m$$

This approach has 3 main drawbacks.

# Method of the test set (*holdout*): drawbacks



### Necessitates a large test set

The number of classification errors can be modelled by a binomial random variable of parameter $m$ and $p = P(\text{erreur})$.

Example: to get 95% of chances to know an error rate of $p = 10\%$ with 1% accuracy calls for $m$=3460 observations.

### The size of the test set should remain limited

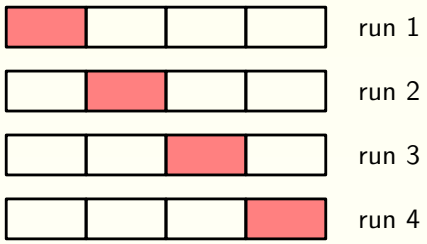
A pity to use too many data for the test rather than for training.

### Pessimistic bias

Generally over-estimates the error rate.

# Cross-validation (*leave one out*, *Jackknife...*)






## Principe

- Division of the data set in $v$ subsets
- Training using the union of $v - 1$ subsets
- Test with the remaining part (1/v fraction of the data set)
- Repeat $v$ times this procedure

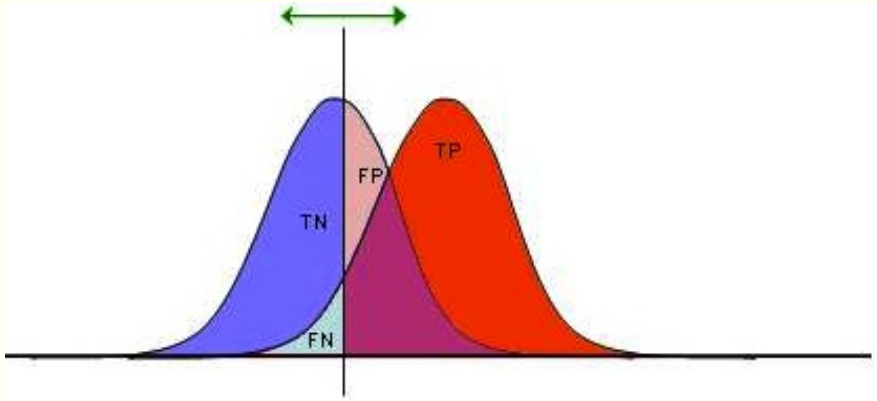
Final error rate estimate finale = average the $v$ estimates

## Leave one out

Extreme case: leave-one-out $\Rightarrow$ $v$ = size of the data set. (test 1 sample only) $\Longrightarrow$ yields good results.

# Confusion matrix






| TP | FP |
|---|---|
| FN | TN |
| 1 | 1 |

## Principle

Evaluation by estimating conditional probabilities such as:

$$P\,(\text{classify } \mathbf{x} \text{ in class } k \mid \mathbf{x} \text{ is in class } \ell)$$

## Computation

These probabilities are gathered in the **confusion matrix** by considering couples of of estimated and true classes.

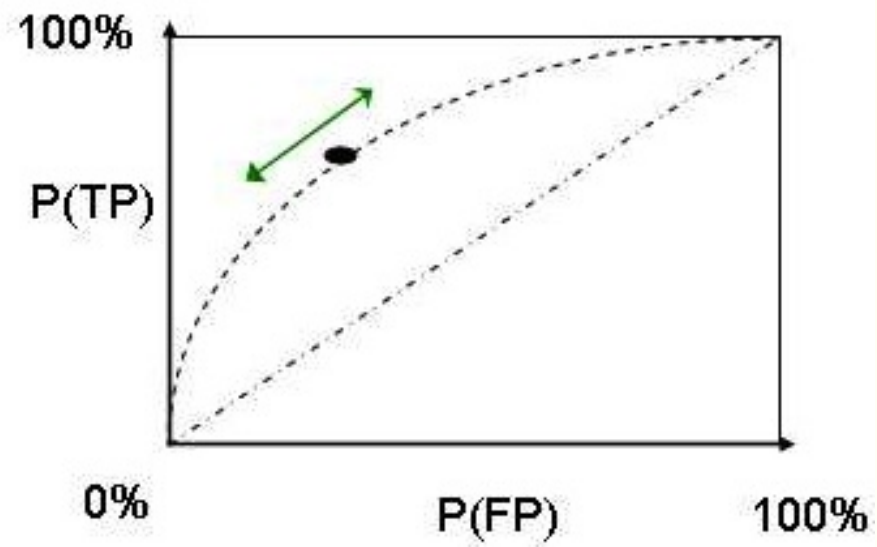
$m$ observations in class $\ell$ (of size $s$) have been affected to class $k$ by the decision function: $\Rightarrow$ the estimated probability is $m/s$

# ROC curve (Receiver Operating Characteristics)



When the decision function depends on some parameter (typically a threshold). A ROC curve represents the compromise between benefits (True Positives) and losses (False Positives).

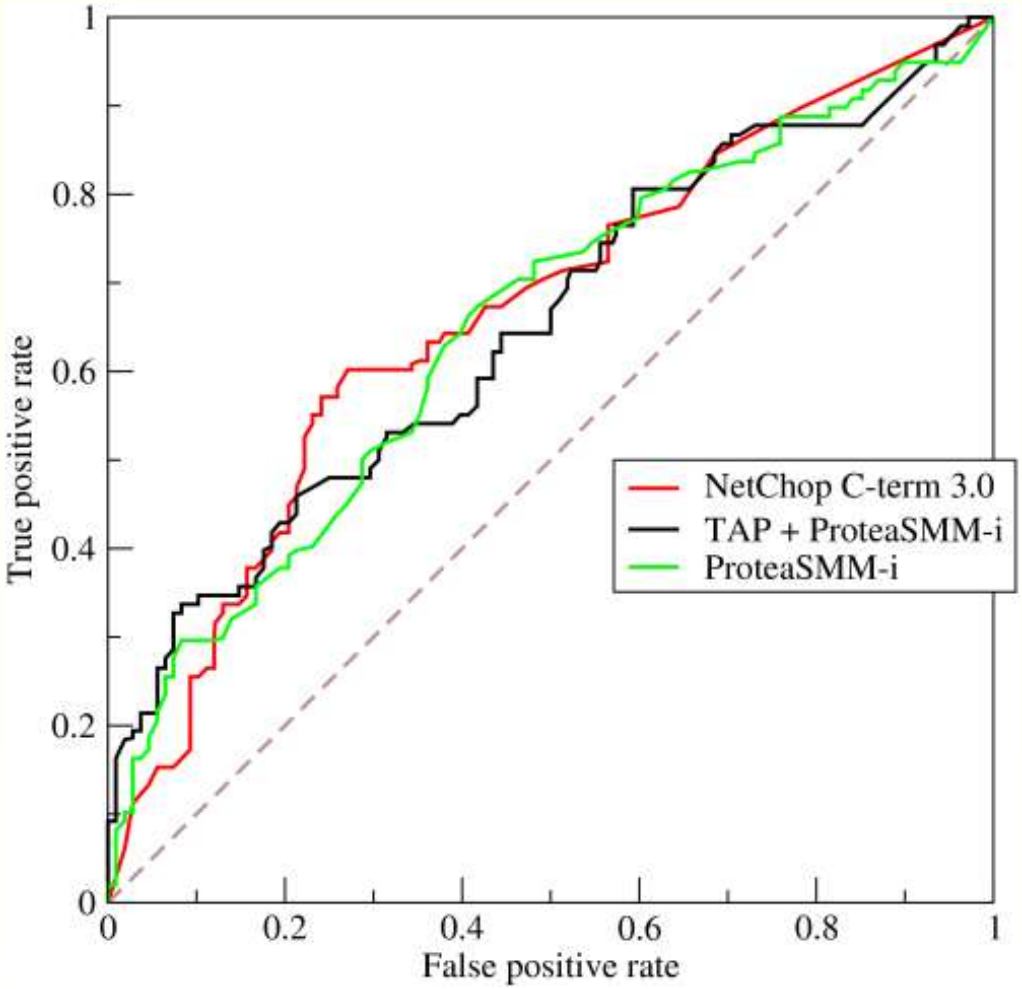
Several important points:




- (0, 0) always predicts 'negative': no FP, no TP;
- (1, 1) always predicts 'positive' : all samples are positive, either FP or TP;
- (0, 1) : perfect classification;
- the diagonal : $(p, p)$ randomly predicts 'positive' with proba. $p$.
- below the diagonal is worse than random decision: it should remain empty in the ROC curve.

# ROC curve (Receiver Operating Characteristics)





# Trajectory



1. Evaluation of performances in pattern recognition (PR)
   - Test set
   - Cross-Validation (CV)
   - Confusion matrix
   - ROC curve

2. Unsupervised classification
   - Clustering K-means
   - Examples of applications of K-means
     - Geyser Old Faithful
     - Génétique et tumeurs cancéreuses
     - Vectorial quantization of images
   - Clustering K-medoids
   - Examples of applications of K-medoids
     - Differences of perception between countries
   - Other approaches

**Objective:** identify the existence of structure in the data (groups/aggregates/clusters)
**ignoring labels.** *(classes are unknown)*

**Principle :**

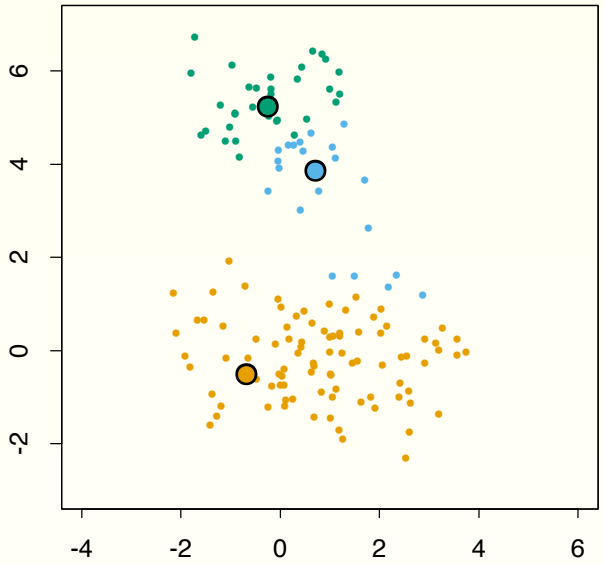
- Choose the number $K$ of expected clusters,
- Start from K initial centers $\mu_k$ :
  - for each $k$, $1 \leq k \leq K$, identify the set of points $\mathcal{V}_k$ closer from $\mu_k$ than any other center,
  - replace $\mu_k$ by the barycentre of $\mathcal{V}_k$
- Iterate until convergence (mathematically proven).

# K-means Clustering

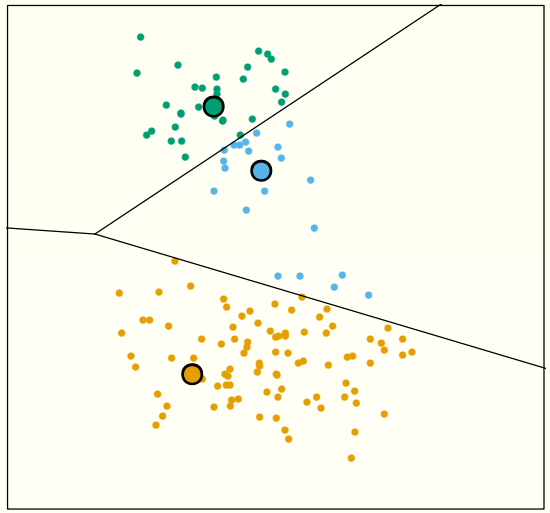
## Illustration



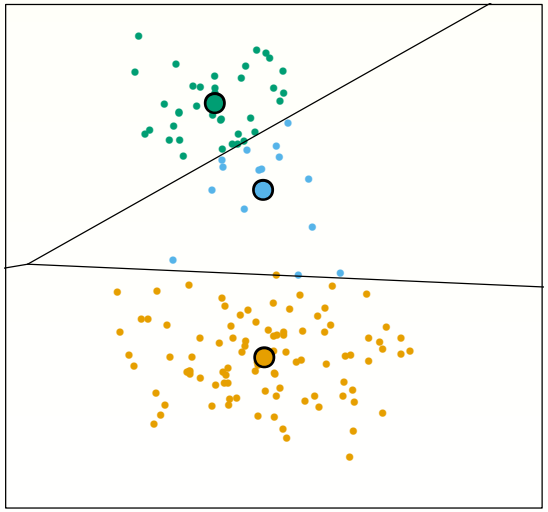
Initial Centroids



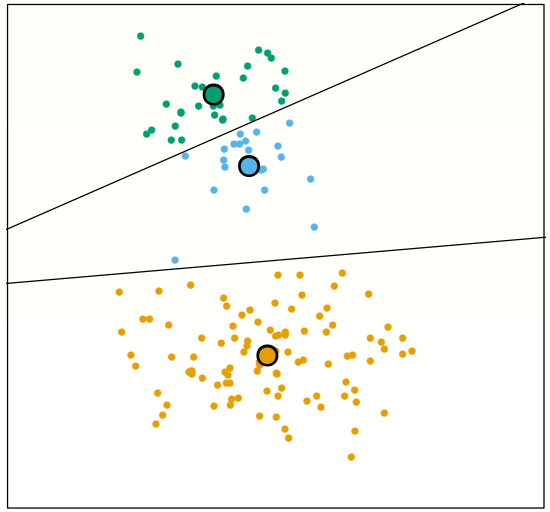

Initial Partition



Iteration Number 2



Iteration Number 20



- **Initial centers** : $K$ random entries...

  $\Rightarrow$ run the algorithm several times and keep the best result,

- **Best result** : minimizes $\sum_{k=1}^{K} \sum_{C(n)=k} \|\mathbf{x}_n - \mu_k\|^2$

- **Limitations** :
  - choice of $K$? non hierarchy when decreasing K...
  - no control of the relative importance of clusters...
- **soft generative version**: GMM & EM algorithm EM (Expectation-Minimization $\Rightarrow$ see Statistics 2 or Hastie p. 272, Bishop p. 430)

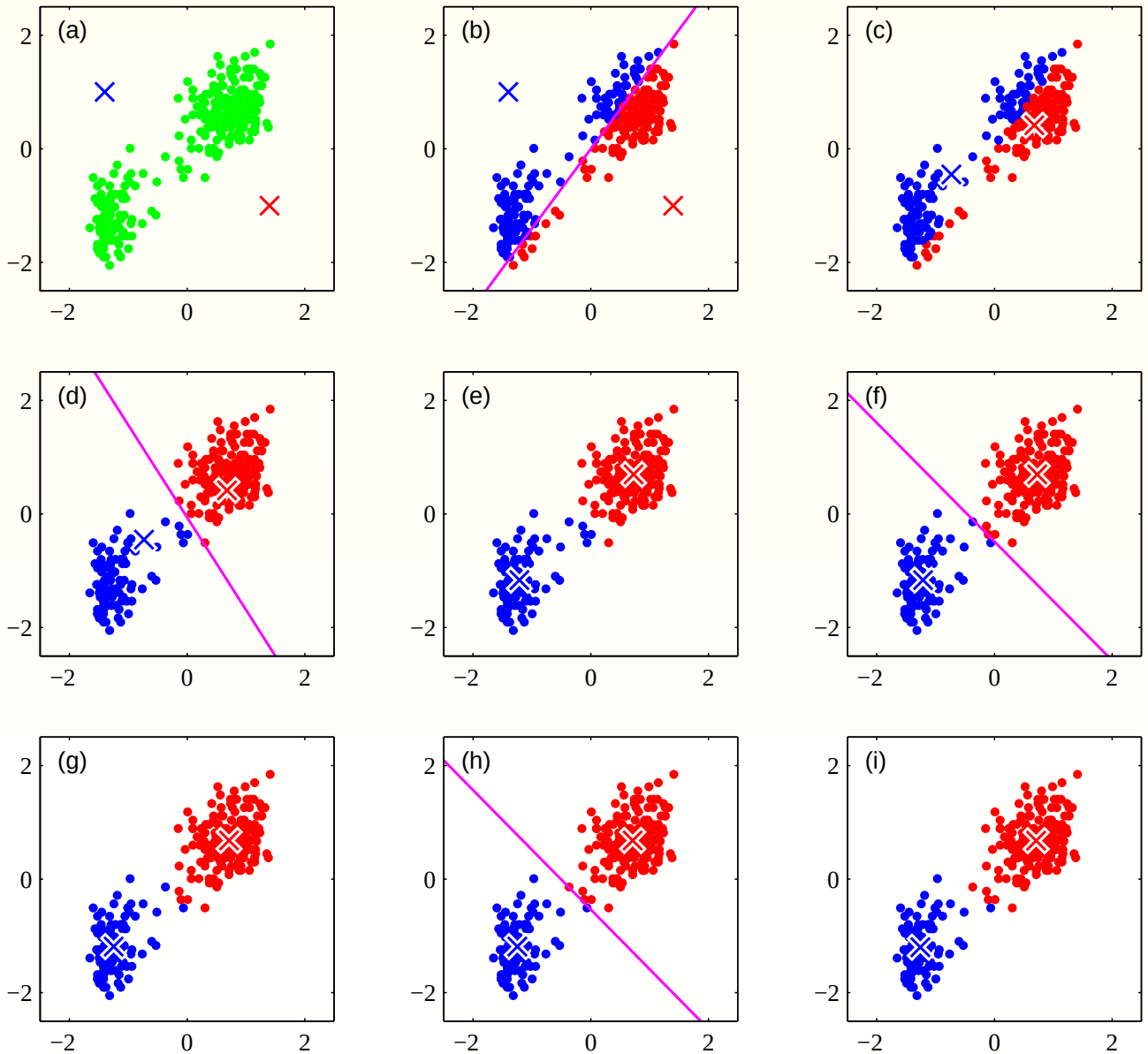
**Données :**

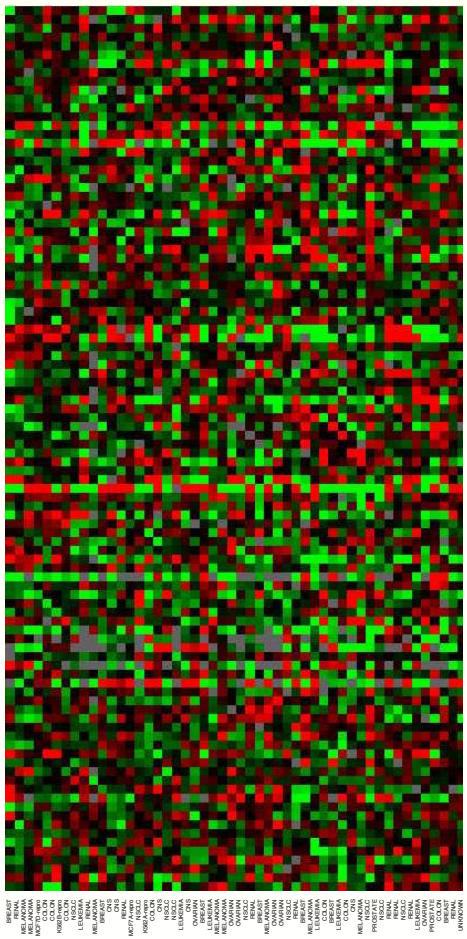
- 272 measures, eruptions of the Old Faithful geyser in Yellowstone
  - $x_1$ : durations of eruptions in minutes,
  - $x_2$ : waiting time before the next eruption
- identification of 2 types of eruptions

# Examples of applications of K-means

## Old Faithful geyser







## Données :

- Matrix 6830x64 real numbers measuring:
  - the expression of some particular gene = lines
  - for some patient = columns
- 64 input vectors $\mathbf{x}_n$ with $D = 6830$ values.
- labels/targets : $t_n$ = breast (cancer), melanoma (skin)... :

$\Rightarrow$ used a posteriori to check the relevance of the proposed clustering.

**Results for K=3 :**

| Cluster | Breast | CNS | Colon | K562 | Leukemia | MCF7 |
|---|---|---|---|---|---|---|
| 1 | 3 | 5 | 0 | 0 | 0 | 0 |
| 2 | 2 | 0 | 0 | 2 | 6 | 2 |
| 3 | 2 | 0 | 7 | 0 | 0 | 0 |
| Cluster | Melanoma | NSCLC | Ovarian | Prostate | Renal | Unknown |
| 1 | 1 | 7 | 6 | 2 | 9 | 1 |
| 2 | 7 | 2 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |

# Vectorial quantization of images

## Compression / dimension reduction

1 pixel $= \mathbf{x}_n = (x_R, x_V, x_B) \Longrightarrow$ clustering 3d

using $K$ colours only

$K = 2$

$K = 3$

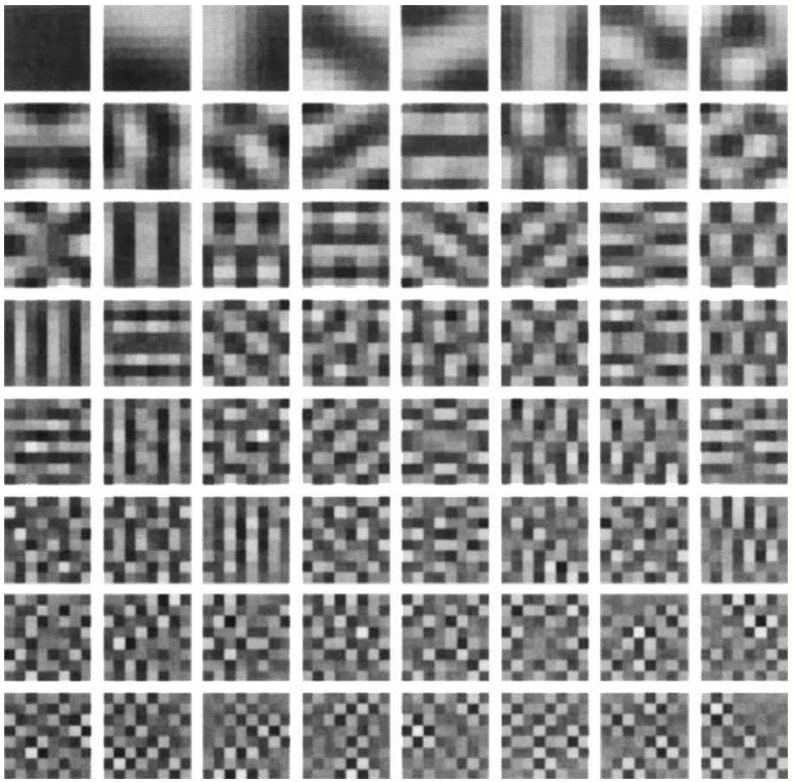
$K = 10$

Original image

# Vectorial quantization of images

## Learning a dictionary of patterns

Images of reference $\Longrightarrow$ $\mathbf{x}_n$ = 8x8 blocks $\Rightarrow$ clustering in dim. 64
using $K$ patterns (atoms) to represent/denoise... images



Think of block cosines in JPEG...

**Objective:** identify the existence of clusters within unlabelled data, *from distances or similarities between points only*

**Principle :**

- Choose the number $K$ of clusters,
- Given an affectation of points $\mathbf{x}_n$ to clusters $\mathcal{C}_k \Leftrightarrow \mu_k$ :
  - find the observation that minimizes the distance to all other points inside its cluster :

$$1 \leq k \leq K, \quad n_k^* = \text{argmin}_{n \in \mathcal{C}_k} \sum_{q \in \mathcal{C}_k} d(\mathbf{x}_n, \mathbf{x}_q)$$

  - the centres become $\mu_k = \mathbf{x}_{n_k^*}$
  - associate each input sample $\mathbf{x}_n$ to the class of the closest center
- Iterate until convergence (guaranteed).

# Examples of applications of K-medoids

## Differences of perception between countries

**Data set:** in 1990, students from political sciences have rated (over 10) "similarities" between pairs of countries among 12 countries (Belgium, Brazil, Chile...)

| | BEL | BRA | CHI | CUB | EGY | FRA | IND | ISR | USA | USS | YUG |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BRA | 5.58 | | | | | | | | | | |
| CHI | 7.00 | 6.50 | | | | | | | | | |
| CUB | 7.08 | 7.00 | 3.83 | | | | | | | | |
| EGY | 4.83 | 5.08 | 8.17 | 5.83 | | | | | | | |
| FRA | 2.17 | 5.75 | 6.67 | 6.92 | 4.92 | | | | | | |
| IND | 6.42 | 5.00 | 5.58 | 6.00 | 4.67 | 6.42 | | | | | |
| ISR | 3.42 | 5.50 | 6.42 | 6.42 | 5.00 | 3.92 | 6.17 | | | | |
| USA | 2.50 | 4.92 | 6.25 | 7.33 | 4.50 | 2.25 | 6.33 | 2.75 | | | |
| USS | 6.08 | 6.67 | 4.25 | 2.67 | 6.00 | 6.17 | 6.17 | 6.92 | 6.17 | | |
| YUG | 5.25 | 6.83 | 4.50 | 3.75 | 5.75 | 5.42 | 6.08 | 5.83 | 6.67 | 3.67 | |
| ZAI | 4.75 | 3.00 | 6.08 | 6.67 | 5.00 | 5.58 | 4.83 | 6.17 | 5.67 | 6.50 | 6.92 |

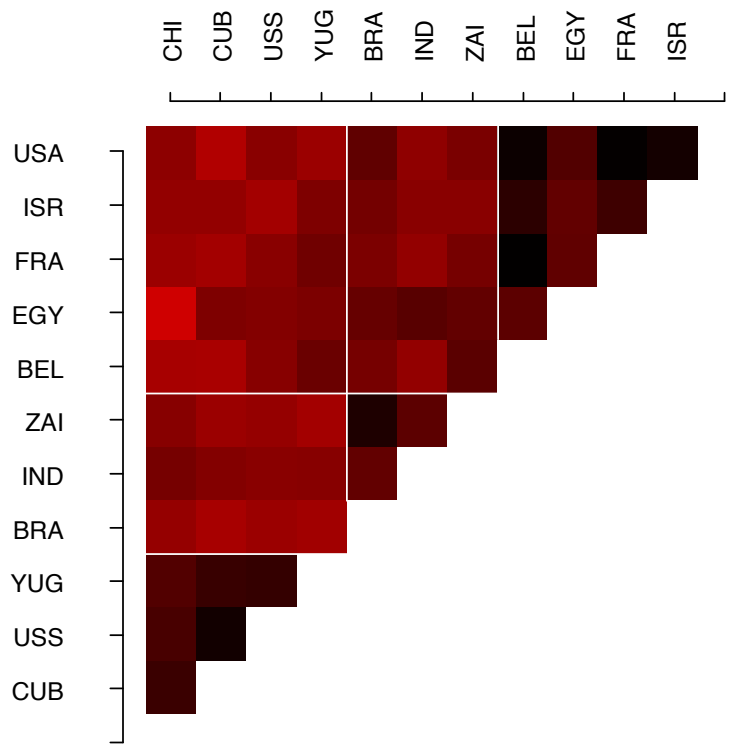
# Examples of applications of K-medoids



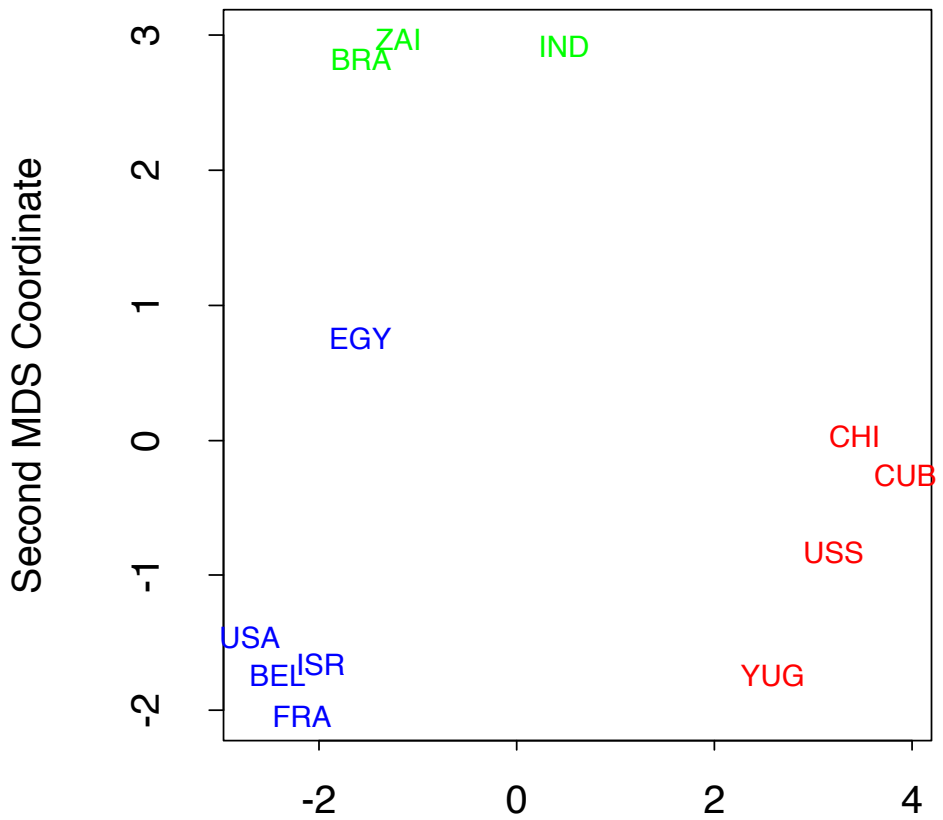
## Differences of perception between countries

**Data set:** in 1990, students from political sciences have rated (over 10) "similarities" between pairs of countries among 12 countries (Belgium, Brazil, Chile...)

Clustering en K=3 groupes + représentation simplifiée :




Reordered Dissimilarity Matrix



# Other approaches

## Spectral clustering, MDS, SOM...

- **Spectral clustering** : similarities $w_{ij}$ only
  notion de Laplacien d'un graphe... voir Hastie p. 544
  *find* $\mathbf{f}$ *that minimizes* $\sum_i \sum_j w_{ij}(f_j - f_i)^2$
- **Multi-Dimensional Scaling (MDS)** : visualization
  *find* $(\mathbf{z}_1, ..., \mathbf{z}_N) \in \mathbb{R}^K$, $K < D$, *minimizing the strain function*

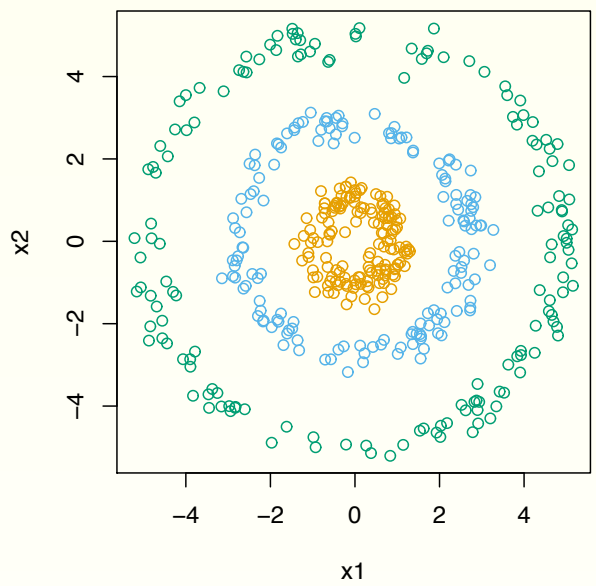
$$S_{MDS}(\mathbf{z}_1, ..., \mathbf{z}_N) = \sum_{i \neq j} (d_{ij} - \|\mathbf{z}_j - \mathbf{z}_i\|)^2$$

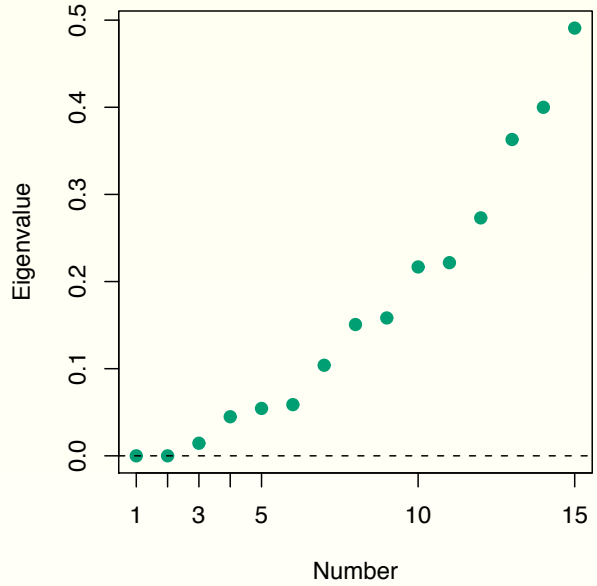
- **Self-Organized Maps de Kohonen (SOM)** (1982)
  grid of neuron with responses $\|$ neighborhoods in features space
- **Local Linear Embedding** : interpolation between neighbours... (Hastie p. 573)

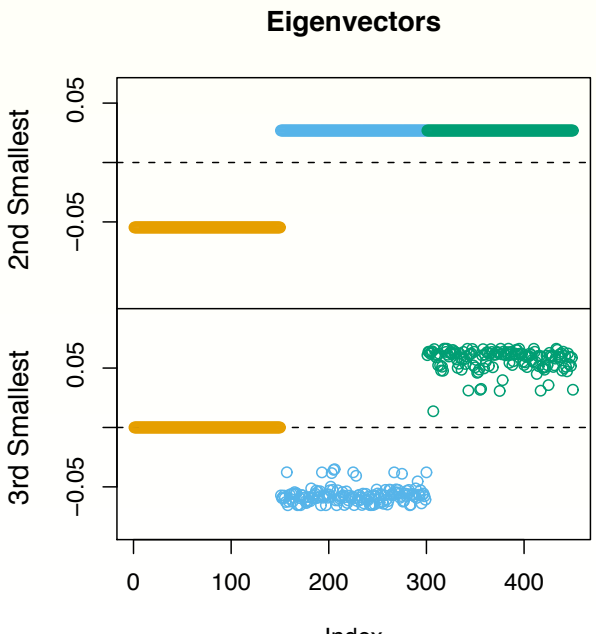
# Other approaches

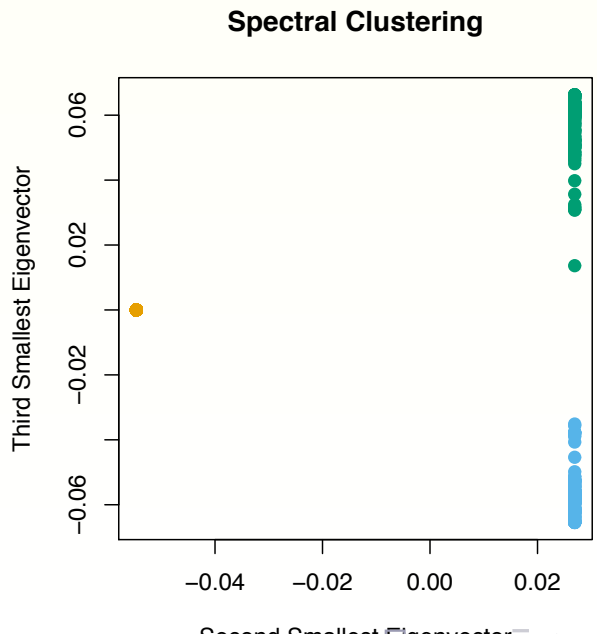
## Spectral clustering, MDS, SOM...





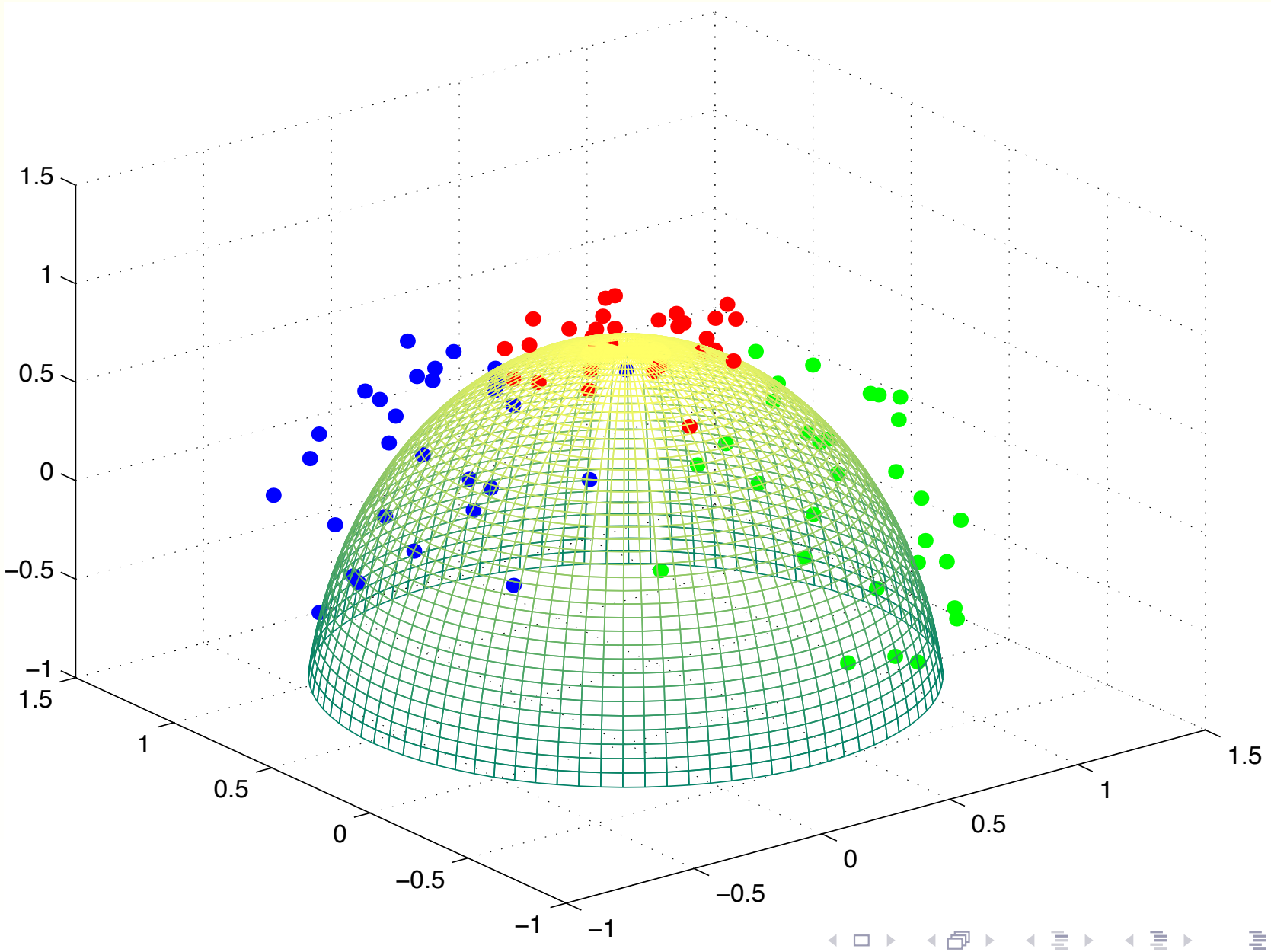



# Other approaches

## Spectral clustering, MDS, SOM...



Synthetic dataset close to the half-sphere :

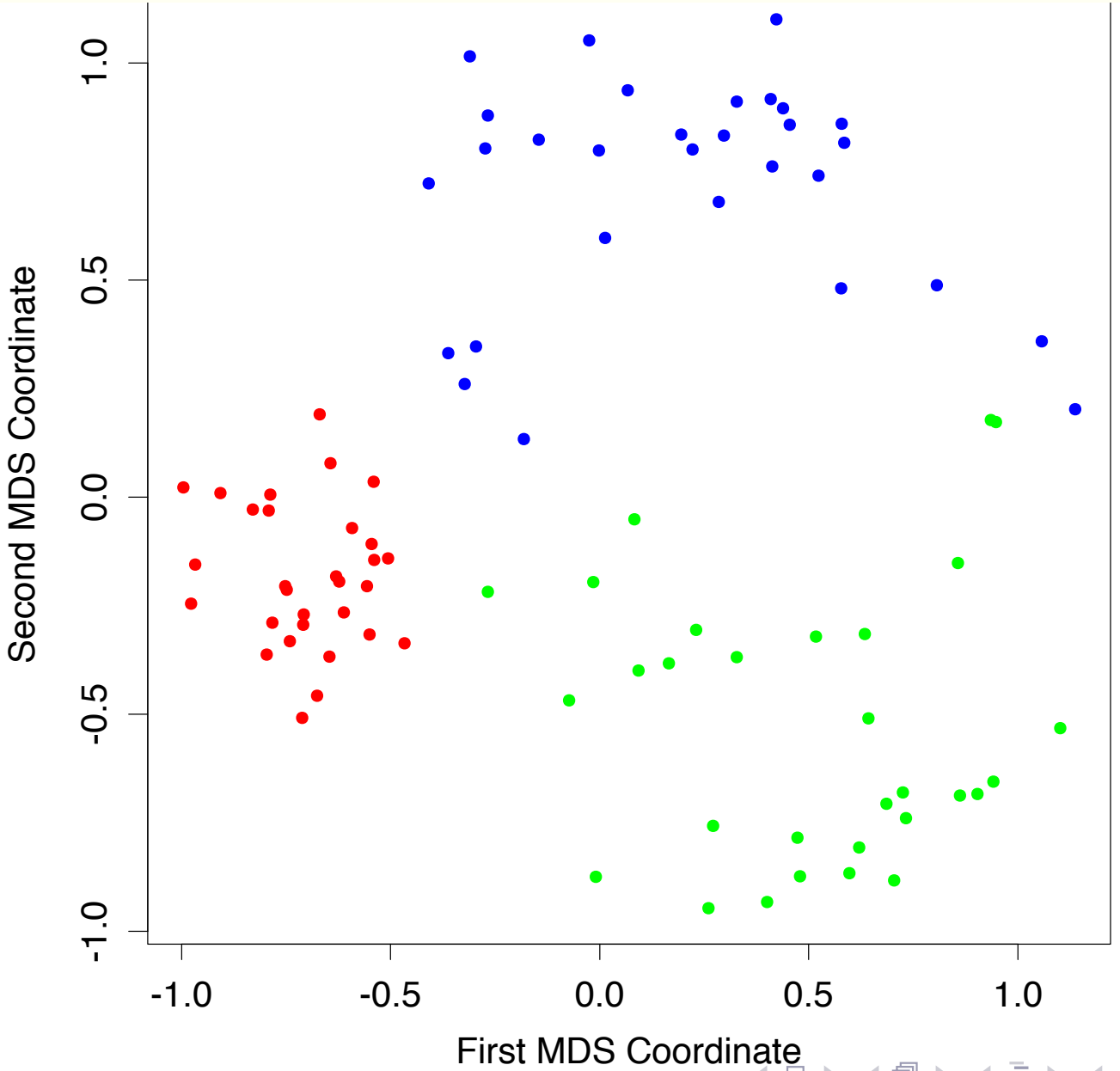


# Other approaches

## Spectral clustering, MDS, SOM...



Result of MDS for the half-sphere dataset :

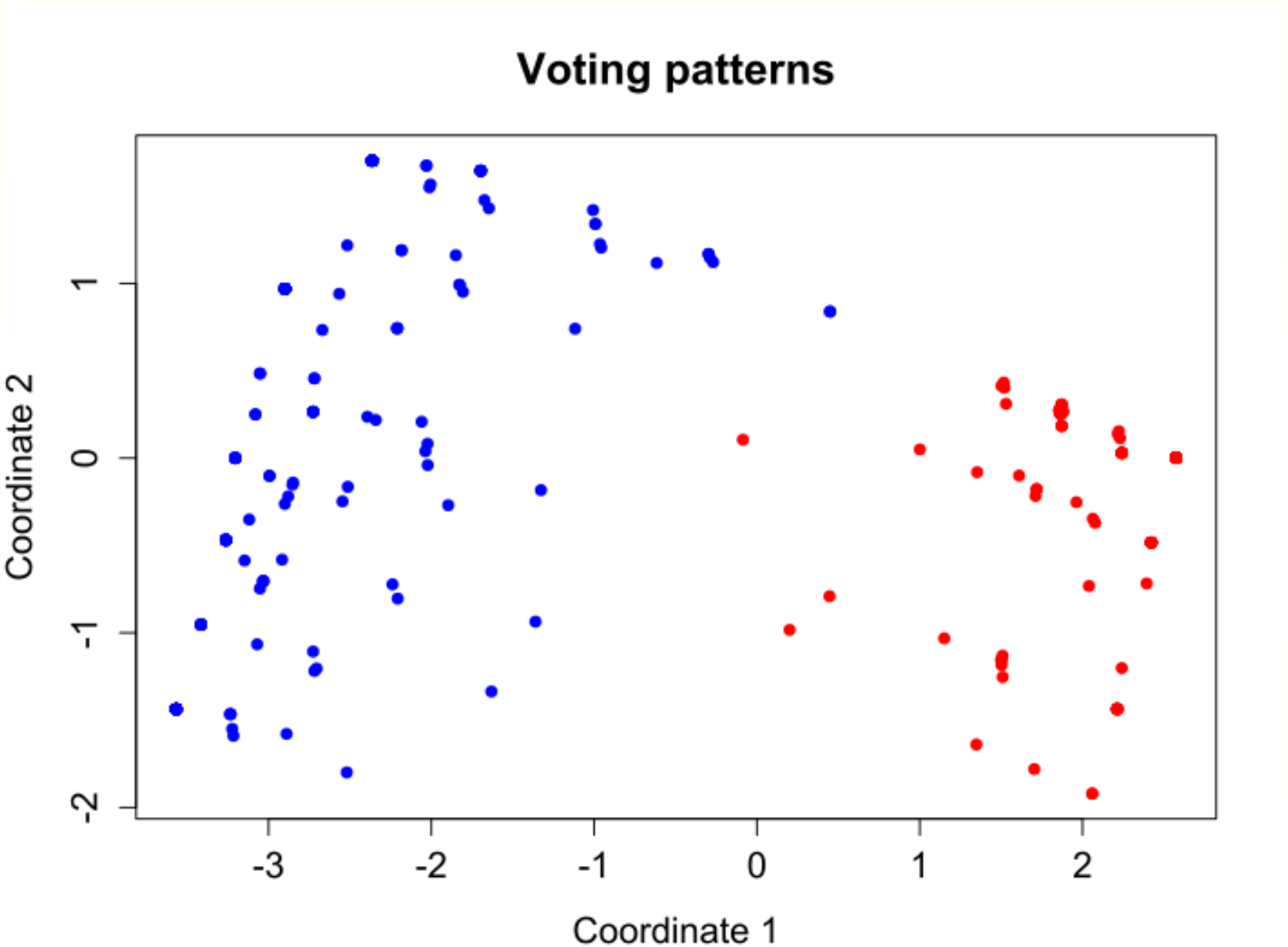


# Other approaches

## Spectral clustering, MDS, SOM...



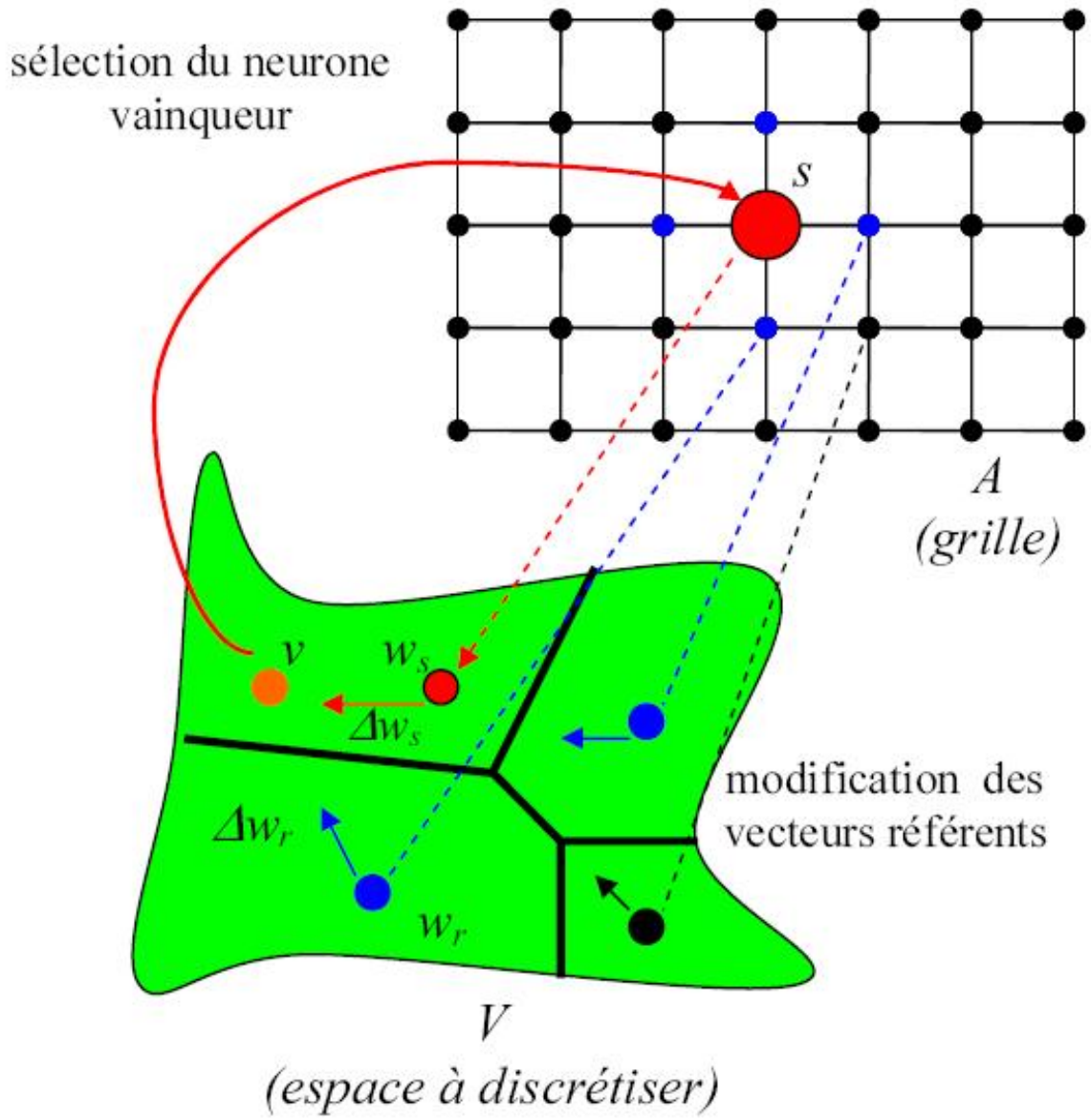
Voting patterns :



# Other approaches

## Spectral clustering, MDS, SOM...

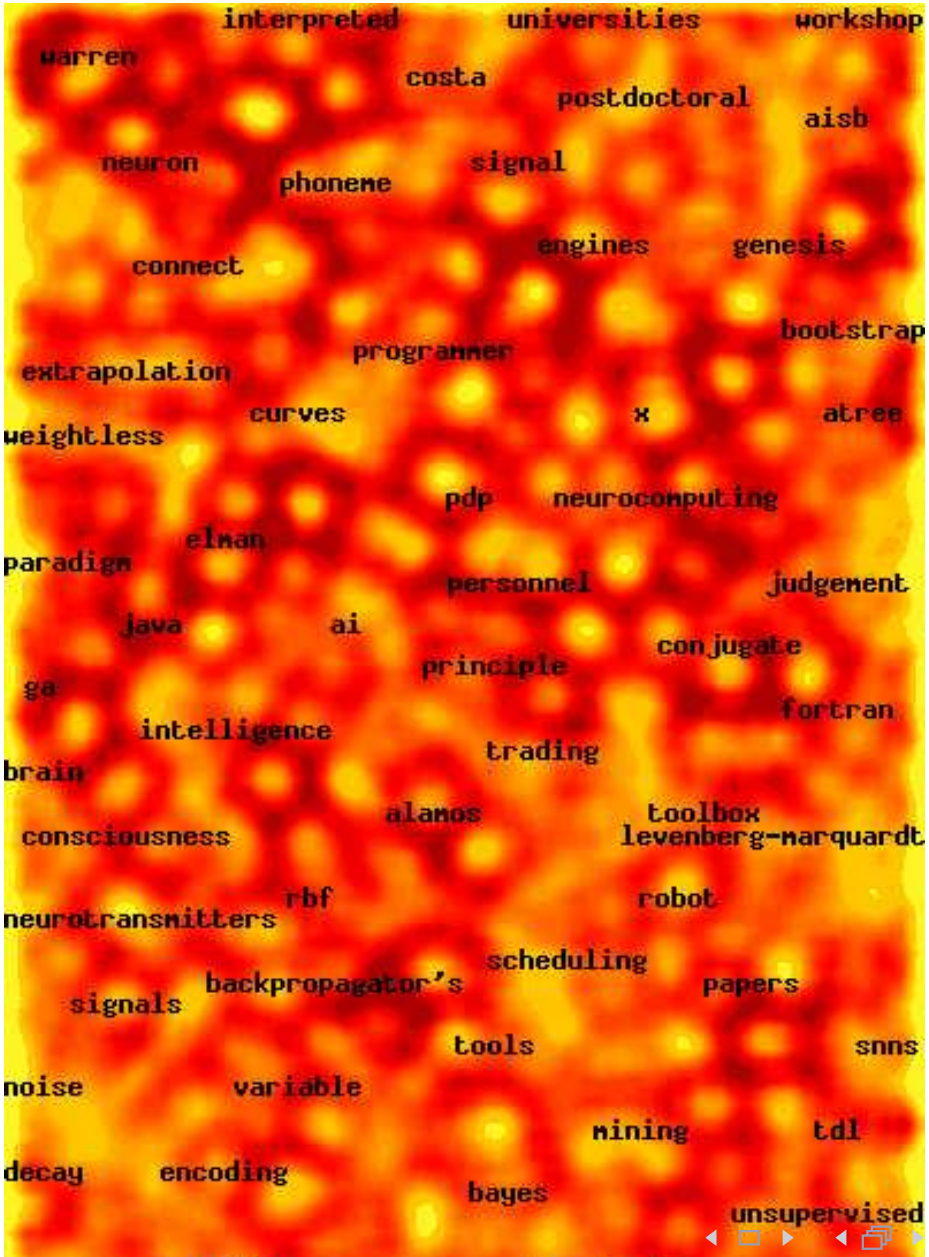
SOM algorithm :



# Other approaches

## Spectral clustering, MDS, SOM...

SOM algorithm for news groups :



# Other approaches

## Spectral clustering, MDS, SOM...



Local Linear Embedding for video-sequences :