Computer Science Laboratory of
Paris 13 University

Paris 13 University - Institut Galilée - LIPN, UMR 7030 du CNRS
99 Avenue J-B. Clément - 93430 Villetaneuse - France

# TUTORIAL

## Unsupervised Collaborative Learning

Nistor Grozavu

EPAT

EPAT'14 : École de Printemps sur l'Apprentissage arTificiel 2014

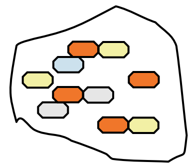
7-12 juin 2014 Carry-le-Rouet (France)

# Plan

- Introduction
- The problem
  - **Consenus Clustering & Collaborative Clustering**
- Collaborative clustering
  - **Horizontal collaboration**
  - **Vertical collaboration**
- Topological Collaborative Clustering
- Diversity Analysis
  - **The problem**
  - **Proposed solutions**
- Real applications
- Conclusions

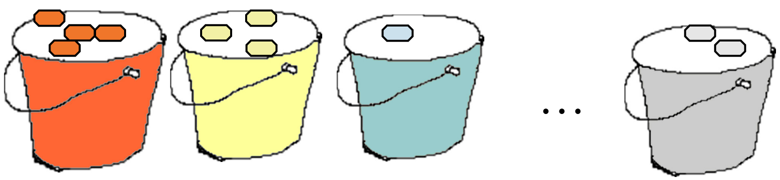
# Introduction

**Classification**



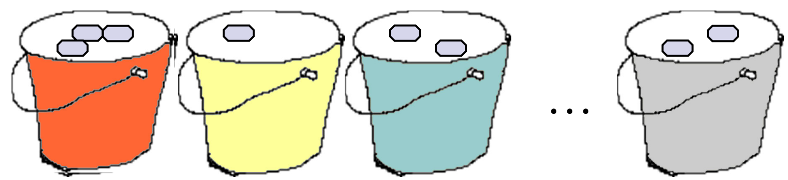
**Clustering**

The labels and the number of classes are known



1 2 3 … n

The labels and the number of classes are unknown



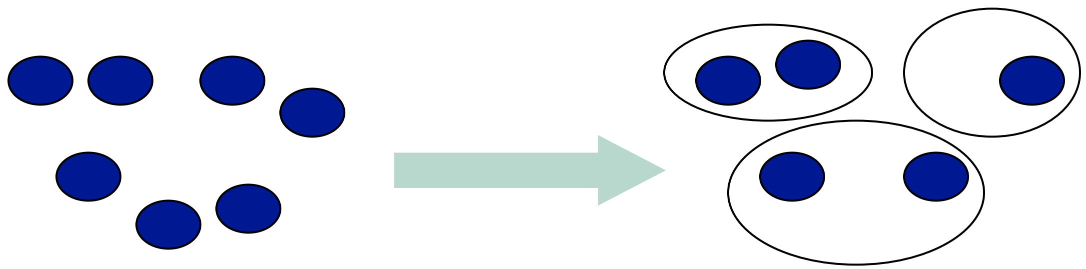
? ? ? … ?

**Difficulties**

- *The objects are not labeled, …*
- *We need to use a similarity measure (for which variables?)*
- *Do we need to know a priori the number of classes?*
- *How to characterize clusters?*

# Introduction : Clustering

- Grouping together of “similar” objects
  - Hard Clustering -- Each object belongs to a single cluster
  - Soft Clustering -- Each object is probabilistically assigned to clusters

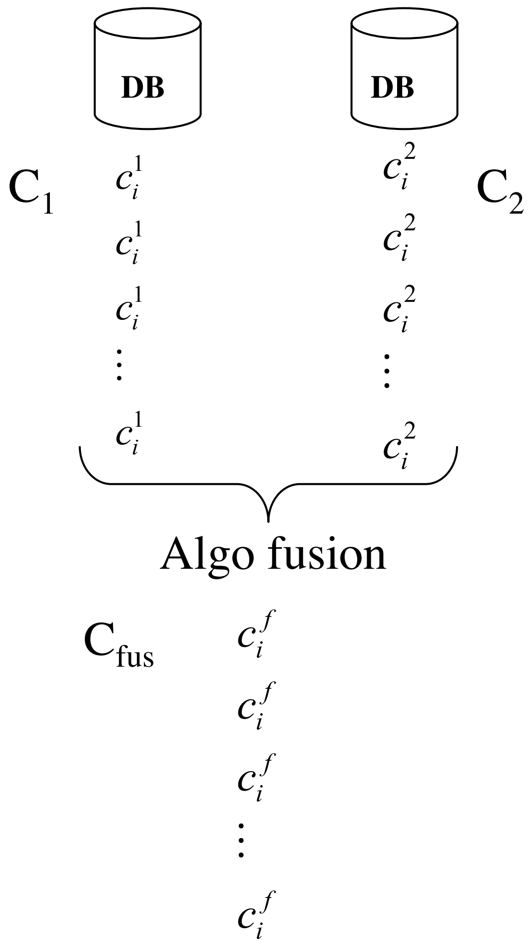


In general, the formalization of the Clustering problem is determined by the following components:

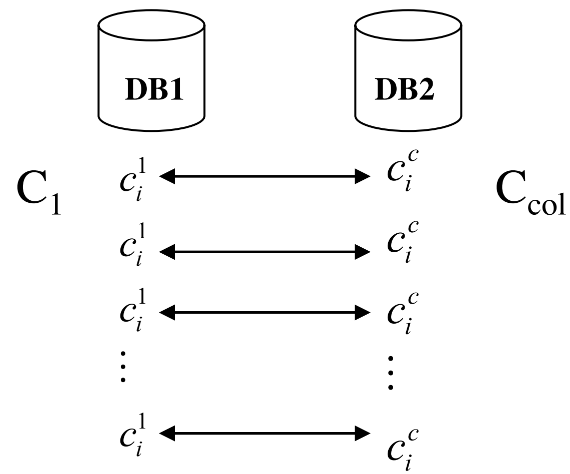
Data representation (categorical, binary, graph…)
The affinity measures (similarity, distance,..)
The objective function
The optimization procedure
Data distribution …

# Introduction - Fusion vs Collaboration

**The principle of the Fusion**

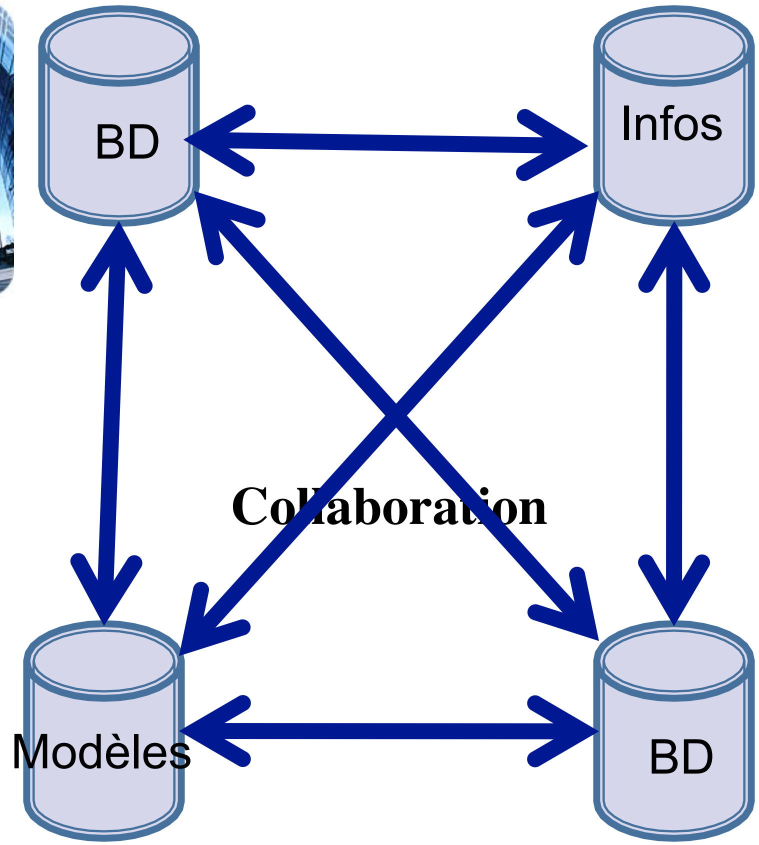



**The principle of the Collaboration**




- Collaborate the datasets of different size;
- Use the same clustering method + a collaboration step;
- Use this schema for different datasets or for the multi-views datasets;

# Collaboration : principle



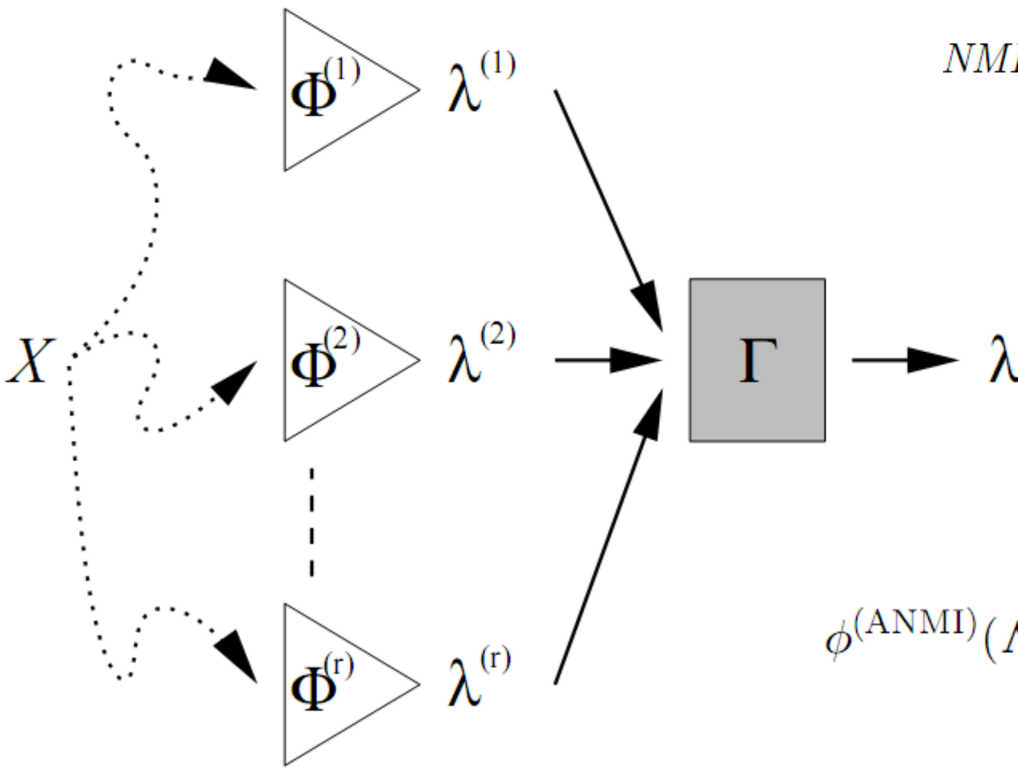

# The problem

- The collaborative clustering is an emerging problem
- Some works (fusion & collaboration) :
  - Pedrycz & Rai  2008 (Collaboration);
  - Costa da Silva & Klusch, 2006 (Collaboration);
  - Wemmert & al., 2007 (Collaborative and Fusion);
  - Cleuziou et al., 2009 (Horizontal Collaboration);
  - Forestier et al., 2009 (Fusion/Collaboration);
  - Grozavu et al., 2009 (Fusion, Collaboration);
  - Strehl & Ghosh, 2002 (Fusion).

- Collaborative Topological Learning uses the principle of the Collaborative Fuzzy c-means (Pedrycz  & Rai, 2002)

# Strehl & Ghosh, 2002 (Fusion)

- Compute the normalized mutual information (NMI) for each dataset ;
- Compute the mean of the NMI for a set of *r* classes (labels) - ANMI;




$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}}$$

$$\phi^{(\mathrm{ANMI})}(\mathbf{\Lambda}, \hat{\lambda}) = \frac{1}{r}\sum_{q=1}^{r} \phi^{(\mathrm{NMI})}(\hat{\lambda}, \lambda^{(q)}).$$

# Costa da Silva & Klusch, 2006 (Collaboration)

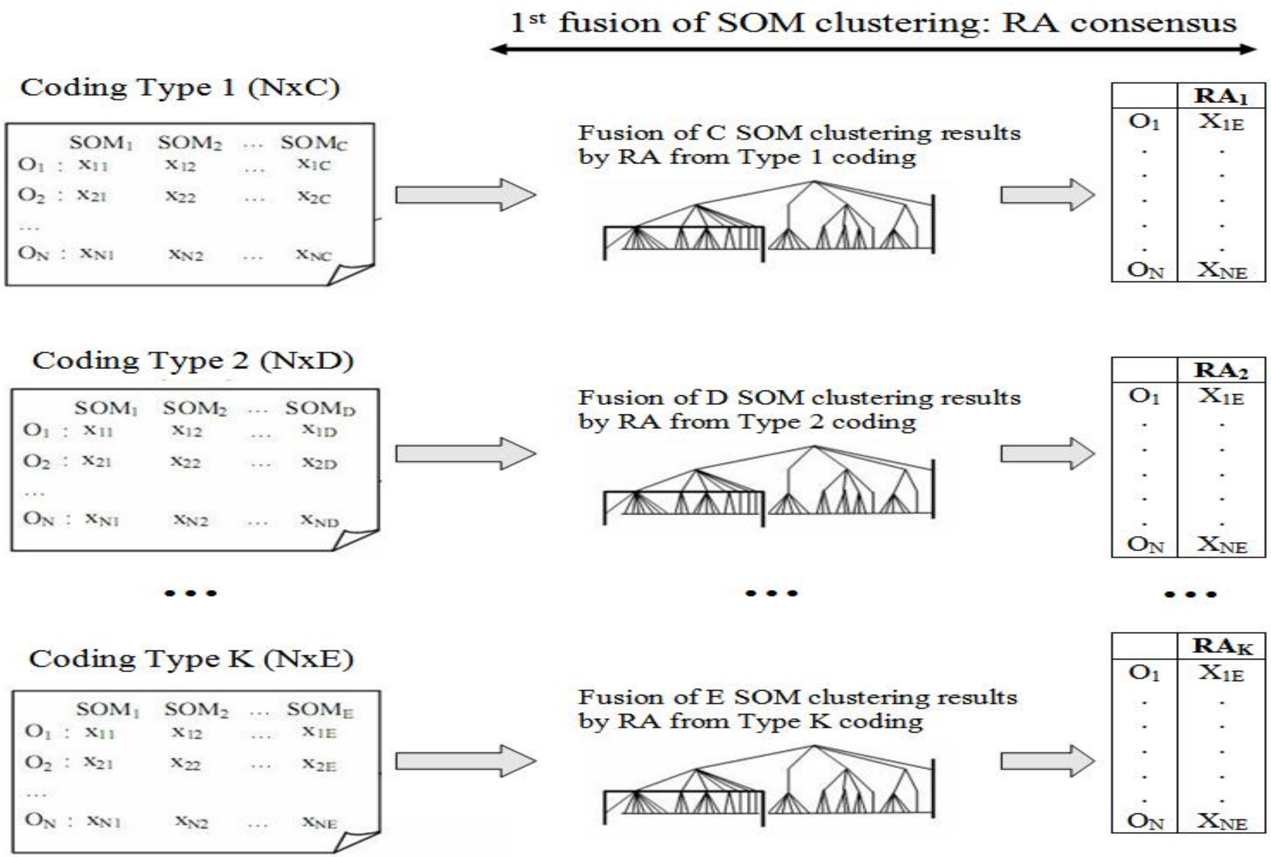
- **Distributed Data Clustering (DDC) :**
  - **KDEC – Density estimation based Distributed Clustering;**
  - **Compute the densities for each local DB:**

$$\hat{\varphi}_{K,h}[S](\vec{x}) = \sum_{i=1}^{N} K\left(\frac{d(\vec{x},\vec{x}_i)}{h}\right)$$

  - **Send these densities to a « *helper site* » which will build the global clustering and send these information to other local sites.**

# Bennani et al., 2009 (Fusion)

- Fusion of several classifications using Relational Analysis approach (AR)



# Costa da Silva & Klusch, 2006 (Collaboration)

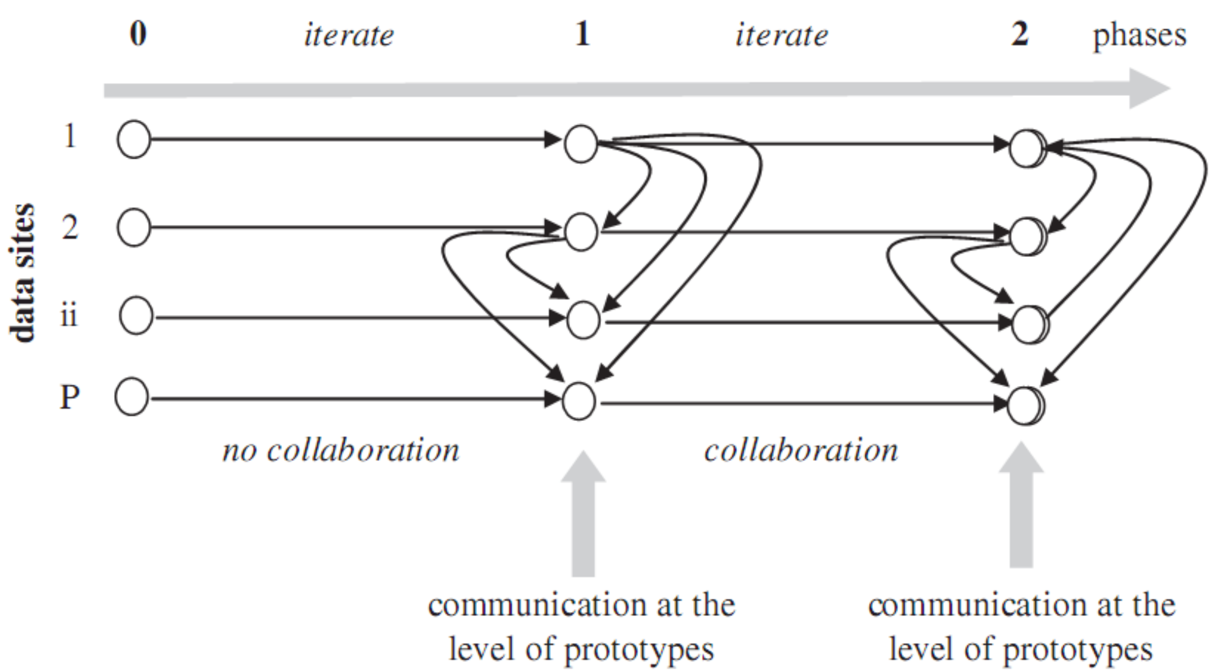
- **Distributed Data Clustering (DDC) :**
  - **KDEC – Density estimation based Distributed Clustering;**
  - **Compute the densities for each local DB:**

$$\hat{\varphi}_{K,h}[S](\vec{x}) = \sum_{i=1}^{N} K\left(\frac{d(\vec{x},\vec{x}_i)}{h}\right)$$

  - **Send these densities to a « *helper site* » which will build the global clustering and send these information to other local sites.**

## Pedrycz & Rai  2008 (Collaboration)

- Fuzzy C-Means Clustering (FCM) :
  - For each dataset, build granular prototypes using the partitions matrix;

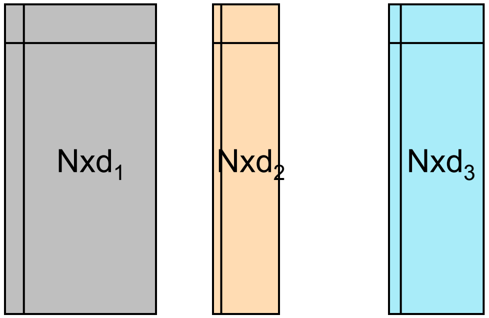


# Collaborative Clustering

## Three main types of collaboration :

### 1. Horizontal

All datasets are described by the same observations but in different spaces Of description (different variables).

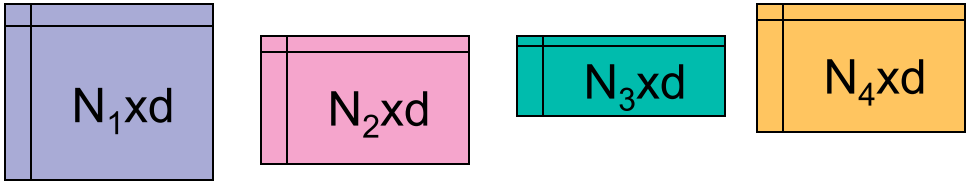



### 2. Vertical

All the datasets have the same variables (same description space), but have different observations.

### 3. Hybrid

Combination between 1 & 2.



# Horizontal collaboration

| ID | Att1 | Att2 | Att3 |
|---|---|---|---|
| id1 | | | |
| id2 | | | |
| id3 | | | |
| id4 | | | |
| id5 | | | |
| id6 | | | |
| id7 | | | |
| id8 | | | |
| id9 | | | |

Dataset [1]

| ID | Att4 | Att5 |
|---|---|---|
| id1 | | |
| id2 | | |
| id3 | | |
| id4 | | |
| id5 | | |
| id6 | | |
| id7 | | |
| id8 | | |
| id9 | | |

Dataset [2]

Same samples

...

| ID | Att6 | Att7 |
|---|---|---|
| id1 | | |
| id2 | | |
| id3 | | |
| id4 | | |
| id5 | | |
| id6 | | |
| id7 | | |
| id8 | | |
| id9 | | |

Dataset [P]

# Vertical Collaboration

Dataset [1]

| ID | Att 1 | Att 2 | Att 3 | Att 4 |
|---|---|---|---|---|
| Id1 | | | | |
| Id2 | | | | |
| Id3 | | | | |
| Id4 | | | | |

Same variables

Dataset [2]

| ID | Att 1 | Att 2 | Att 3 | Att 4 |
|---|---|---|---|---|
| Id5 | | | | |
| Id6 | | | | |
| Id7 | | | | |

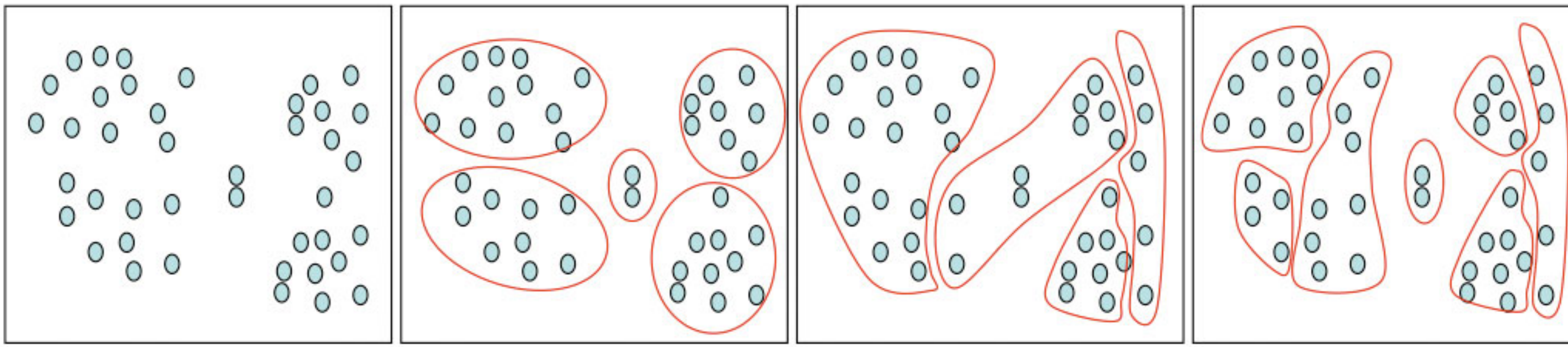
⋮

Dataset [P]

| ID | Att 1 | Att 2 | Att 3 | Att 4 |
|---|---|---|---|---|
| Id8 | | | | |
| Id9 | | | | |

# The problem

## Horizontal collaboration vs Vertical collaboration

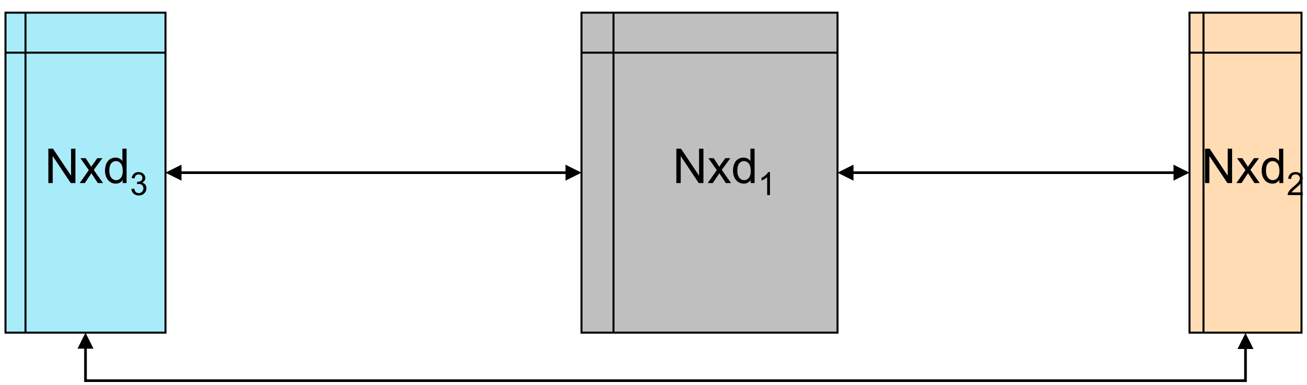


Group of houses

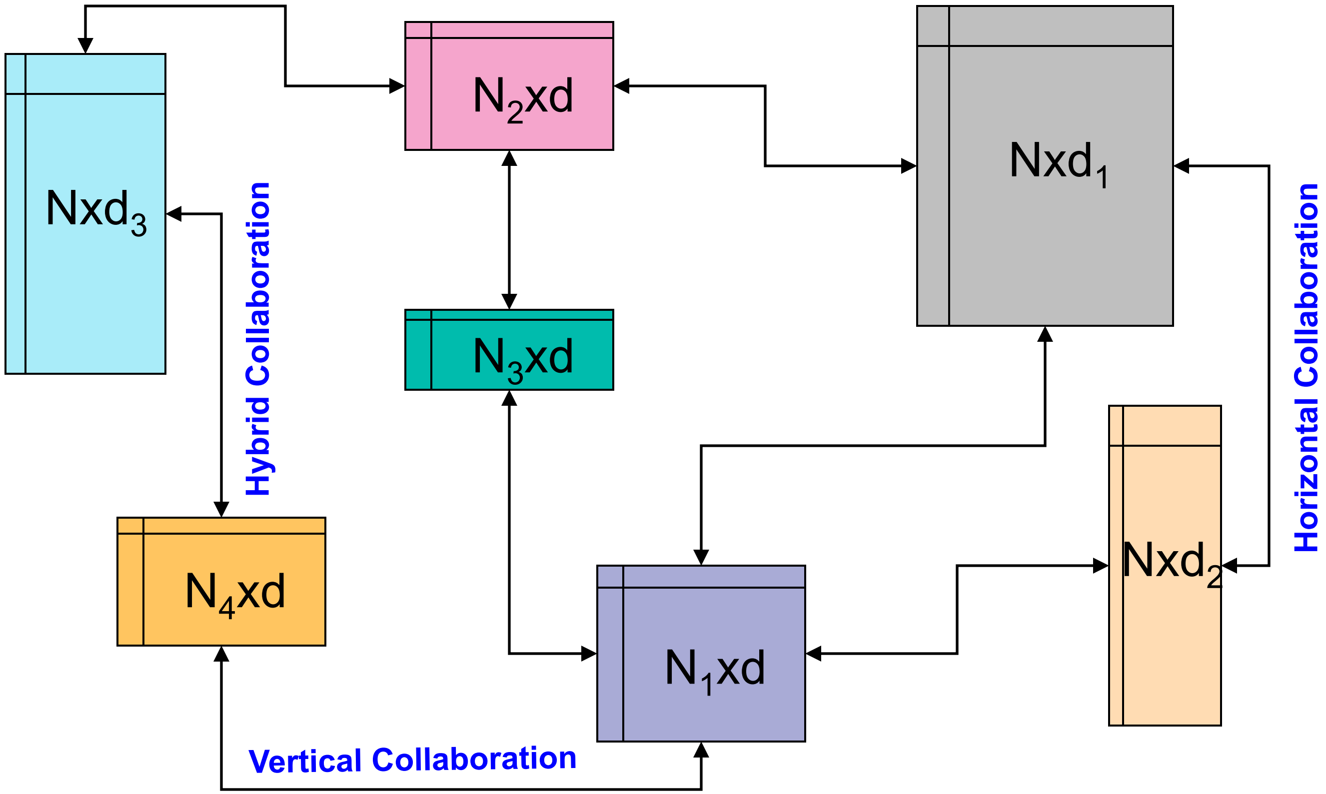
Clustered based on geographic distance

Clustered based on value

Clustered based on size and value



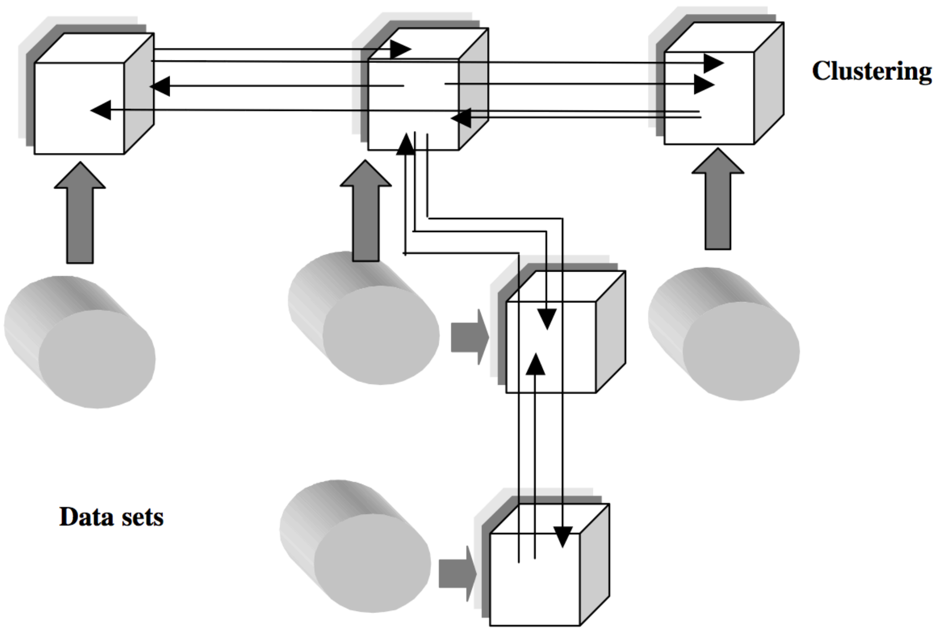
# The problem




- **How to improve the local clustering derived out of a set of distant clustering results without sharing the initial data ?**

# Collaborative FCM (Pedrycz, 2002)

# Collaborative FCM (Pedrycz, 2002)

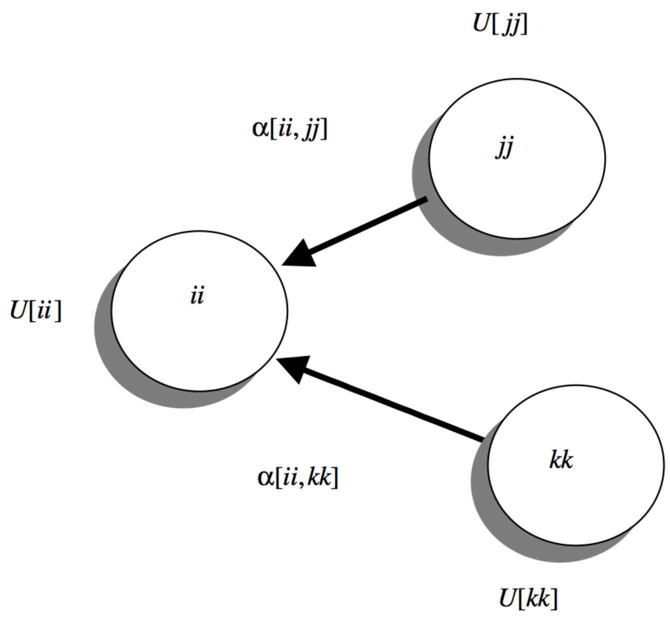



$$\sum_{k=1}^{N} \sum_{i=1}^{c} u_{ik}^2[ii] d_{ik}^2[ii]$$

The distance function between the **i**th prototype and the **k**th pattern in the same subset is denoted by $d_{ik}{}^2[ii]$, i = 1,2,...,c, k = 1,2,...,N and ii = 1, 2, . . . , P

# Collaborative FCM (Pedrycz, 2002)

Each entry of the collaborative matrix describes the intensity of the interaction. In general, α[ii,kk] assumes nonnegative values.




Collaboration in the clustering scheme represented by the matrix of collaboration levels between the subsets; the partition matrices generated for each data set are shown.

# General collaborative clustering scheme (Pedrycz, 2002)

*Given*: subsets of patterns $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_P$

*Select*: distance function, number of clusters ($c$), termination criterion, and collaboration matrix $\alpha[ii, jj]$.

*Compute*: initiate randomly all partition matrices $U[1], U[2], \ldots, U[P]$

***Phase I***

For each data

*repeat*

compute prototypes $\{\mathbf{v}_i[ii]\}, i = 1, 2, \ldots, c$ and partition matrices $U[ii]$ for all subsets of patterns

*until* a termination criterion has been satisfied

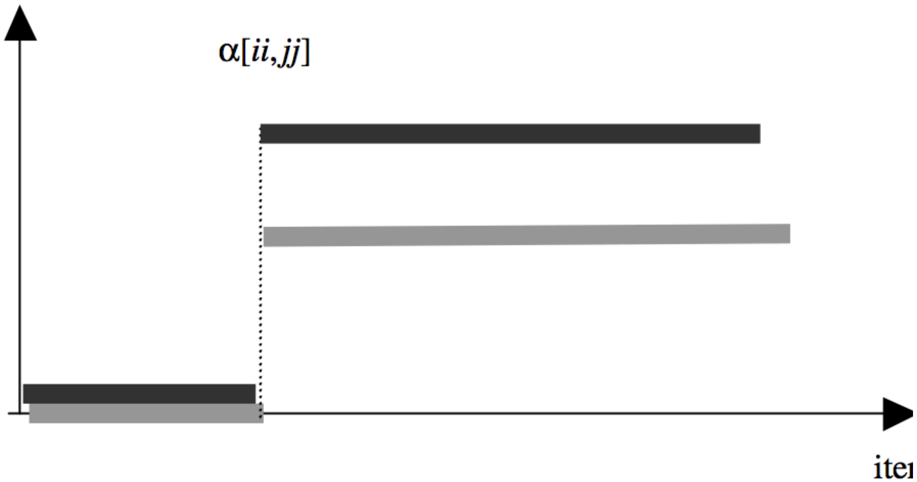
***Phase II***

*repeat*

For the given matrix of collaborative links $\alpha[ii, jj]$, compute prototypes and partition matrices $U[ii]$ using (7.4) and (7.7)

*until* a termination criterion has been satisfied

# Quantification of the Collaborative Phenomenon of Clustering (Pedrycz, 2002)




**The intensity of collaboration** :

$$\delta = \|U[ii] - U_{\text{ref}}[ii]\|$$

$U_{\text{ref}}$[ii] to denote the partition matrix produced independently of other sources

**Consistency measure :**

$$\phi[ii, jj] = \|U[ii] - U[jj]\|$$

indicates the structural differences between the partition matrices defined over two data sets (ii and jj, respectively)

# Topological Collaborative Clustering

# Prototype based Clustering (SOM)

$$K_{\delta(i,j)} = \exp\left(-\frac{\delta^2(i,j)}{T^2}\right)$$

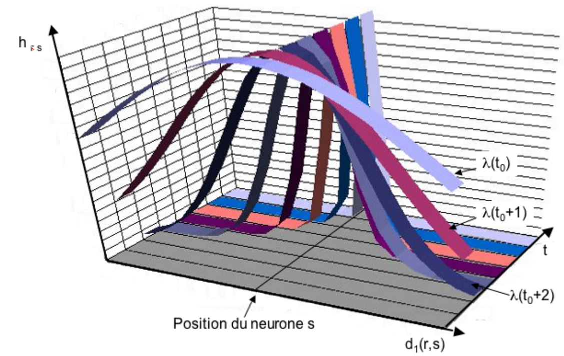
Neighborhood function

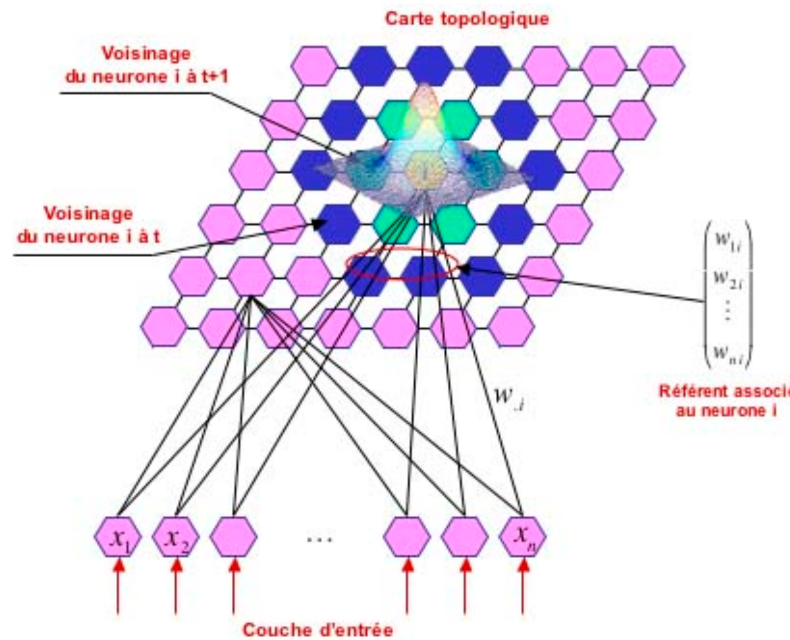
$$w^* = \arg\min_{w}\left\{\sum_{i=1}^{N}\sum_{j=1}^{|w|} K_{\delta(j,\chi(x_i))}\left\|x_i - w_j\right\|^2\right\}$$

Assignment function

$$\chi(x_i) = \arg\min_{j}\left(\left\|x_i - w_j\right\|^2\right)$$

N : number of observations x, |W| : number of prototypes w



# Horizontal Collaboration

$$w^* = \arg\min_{w} \left\{ R_{SOM}^{[ii]}(\chi, w) + \boxed{R_{Col\_H}^{[ii]}(\chi, w)} \right\}$$

**Collaboration coefficient**

Collaboration term

$$R_{Col\_H}^{[ii]}(\chi, w) = \sum_{jj=1,\, jj \neq ii}^{P} \alpha_{[ii]}^{[jj]} \sum_{i=1}^{N} \sum_{j=1}^{|w|} \left( K_{\delta(j,\chi(x_i))}^{[ii]} - K_{\delta(j,\chi(x_i))}^{[jj]} \right)^2 \left\| x_i^{[ii]} - w_j^{[ii]} \right\|^2$$

$$R_{SOM}^{[ii]}(\chi, w) = \sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\delta(j,\chi(x_i))}^{[ii]} \left\| x_i^{[ii]} - w_j^{[ii]} \right\|^2$$

$N^{[ii]}$ : the number of observations on the datasets [ii], P : the number of datasets

# Learning Algorithm

**Algorithm 1:** The horizontal collaboration algorithm

Fix the collaboration matrix $\alpha_{[ii]}^{[jj]}$

**1. Local step:**

For each dataset $BD[ii]$, $ii = 1$ to $P$ :

Find the prototypes minimizing the classical SOM objective function:

$$w^* = \arg\min_{w} \left[ R_{SOM}^{[ii]}(\chi, w) \right]$$

**2. Collaboration step:**

For the horizontal collaboration of the $[ii]$-th map with the $[jj]$-th map:

Update the prototypes of the $[ii]$-th map minimizing the objective function of the horizontal collaboration:

$$w_{jk}^{*[ii]} = \frac{\sum_{i=1}^{N} K_{\sigma(j,\chi(x_i))}^{[ii]} x_{ik}^{[ii]} + \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]} \right)^2 x_{ik}^{[ii]}}{\sum_{i=1}^{N} K_{\sigma(j,\chi(x_i))}^{[ii]} + \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]} \right)^2}$$

# Vertical Collaboration

$$w^{*} = \arg\min_{w}\left\{R_{SOM}^{[ii]}(\chi,w) + \boxed{R_{Col\_V}^{[ii]}(\chi,w)}\right\}$$

**Collaboration coefficient**

Collaboration term

$$R_{Col\_V}^{[ii]}(\chi,w) = \sum_{jj=1,\, jj\neq ii}^{P} \alpha_{[ii]}^{[jj]} \sum_{i=1}^{N^{[ii]}} \sum_{j=1}^{|w|} \left( K_{\delta(j,\chi(x_i))}^{[ii]} - K_{\delta(j,\chi(x_i))}^{[jj]} \right)^2 \left\| w_j^{[ii]} - w_j^{[ii]} \right\|^2$$

$$R_{SOM}^{[ii]}(\chi,w) = \sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\delta(j,\chi(x_i))}^{[ii]} \left\| x_i^{[ii]} - w_j^{[ii]} \right\|^2$$

$N^{[ii]}$ : the number of observations on dataset [ii], P : the number of datasets sites

# Learning algorithm

**Algorithm 2:** Vertical Collaboration algorithm

Fix the collaboration parameter $\alpha_{[ii]}^{[jj]}$

**1. Local step:**

For each dataset $BD[ii]$, $ii = 1$ to $P$ :

Find the prototypes minimizing the classical SOM objective function:

$$w^* = \arg\min_w \left[ R_{SOM}^{[ii]}(\chi, w) \right]$$

**2. Collaboration step:**

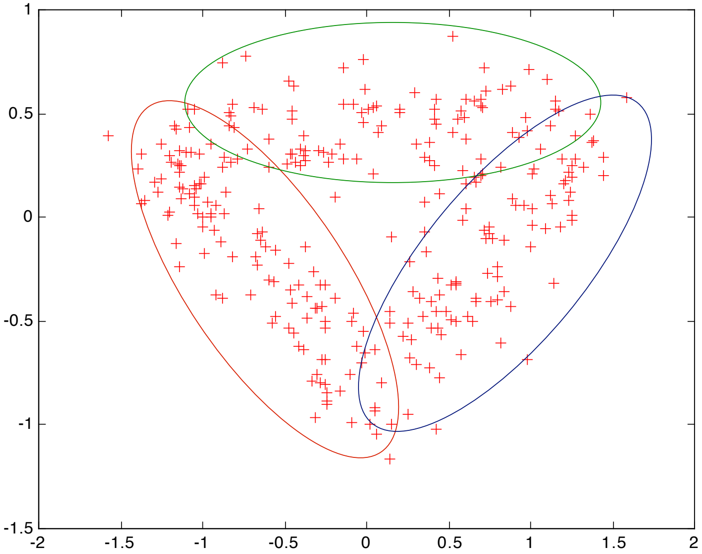
For the vertical collaboration of the $[ii]$-th map with the map $[jj]$:

Update the prototypes of the $[ii]$-th map minimizing the objective function of the vertical collaboration:

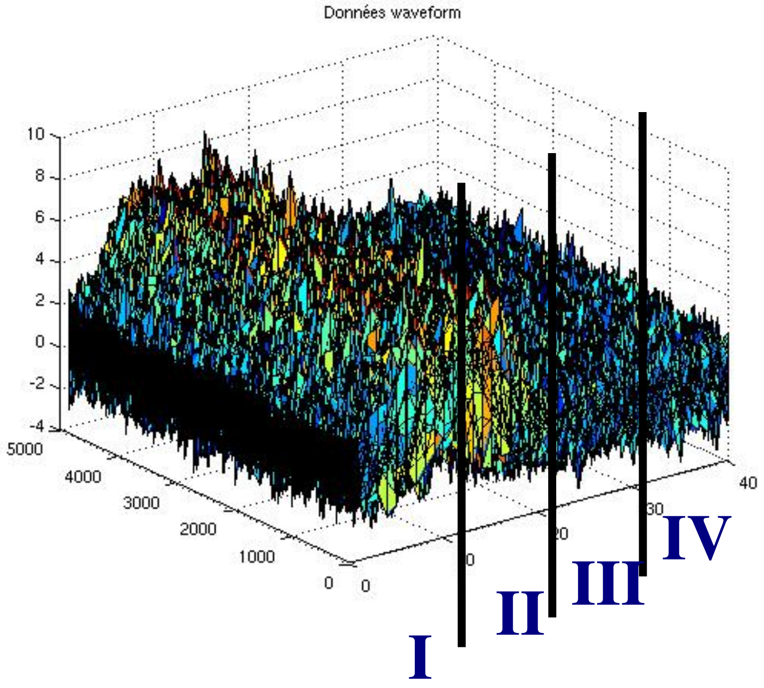
$$w_{jk}^{*[ii]} = \frac{\sum_{i=1}^{N^{[ii]}} K_{\sigma(j,\chi(x_i))}^{[ii]} x_{ik}^{[ii]} + \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N^{[ii]}} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]} \right)^2 w_{ik}^{[jj]}}{\sum_{i=1}^{N} K_{\sigma(j,\chi(x_i))}^{[ii]} + \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j,\chi(x_i))}^{[ii]} - K_{\sigma(j,\chi(x_i))}^{[jj]} \right)^2}$$

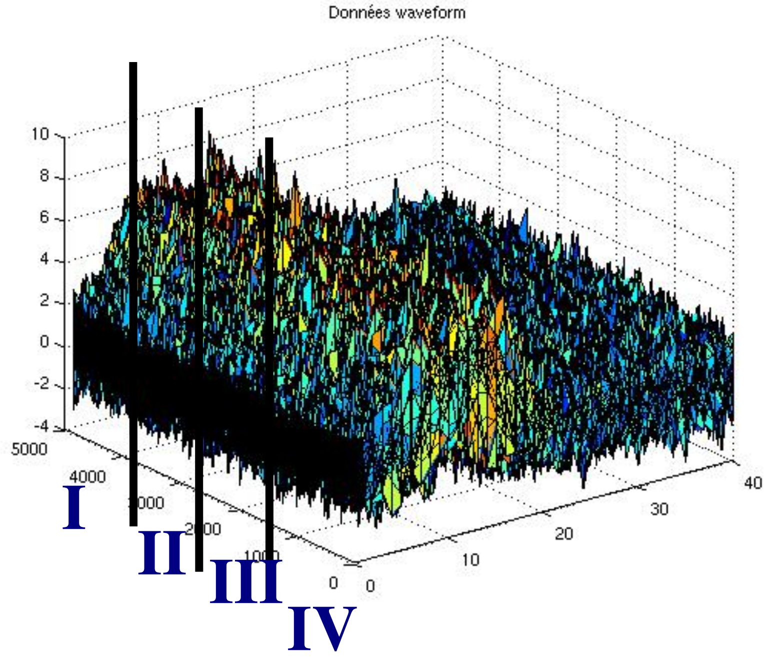
# Ilustrative example

- Waveform dataset
  - 5000 samples
  - 40 variables where 19 variables are Gaussian noisy
  - 3 classes

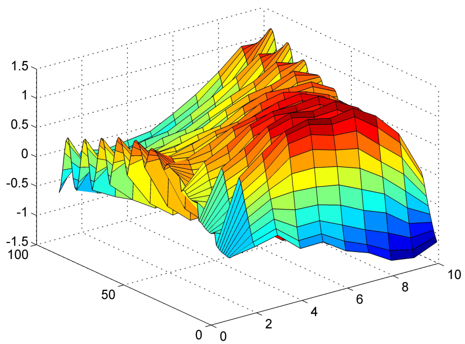


# Horizontal Collaboration (waveform)



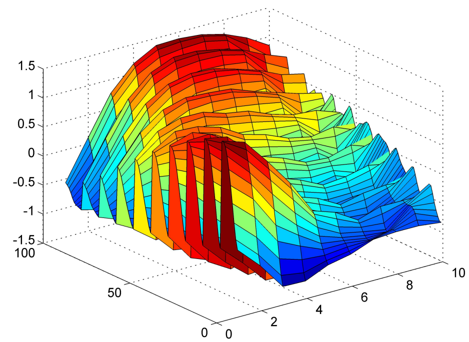



# Horizontal Collaboration (waveform)



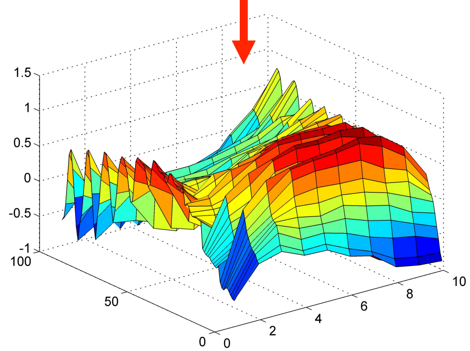

The prototypes of the 1st dataset before the collaboration : SOM1

**75.71%**




The prototypes of the 2nd dataset before the collaboration : SOM2

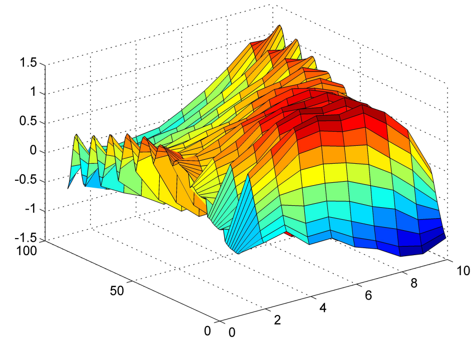
**79.61%**




The prototypes of the 1st dataset after the collaboration with the SOM2 map : SOM12

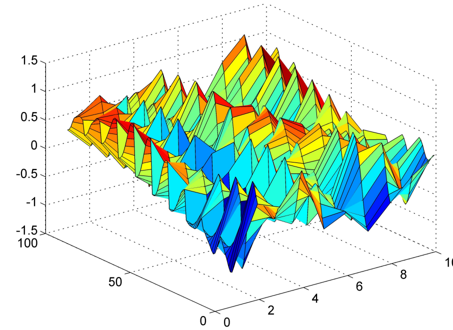
**76.21%**

# Horizontal Collaboration (waveform)



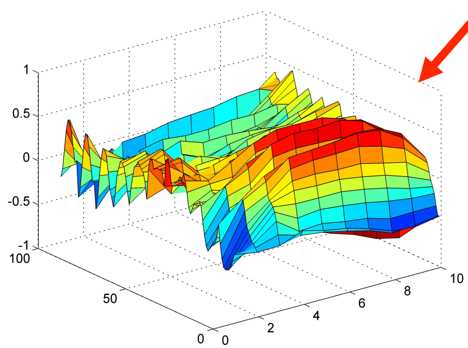

The prototypes of the 1st map obtained from the
1st dataset before the collaboration : SOM1

**75.71%**



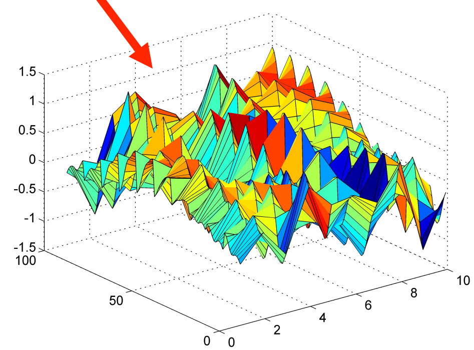

The prototypes of the map from the 3rd
dataset before the collaboration : SOM3

**47.19%**




The prototypes of the map obtained from the 1st
dataset after the collaboration with SOM3 : SOM13

**62.47%**




The prototypes of the map obtained from the 3rd
dataset after the collaboration with SOM1 : SOM31

**54.63%**

# Validation de l' approche collaboratif sur différents bases de données

## Collaboration horizontale

| Dataset | DB.. | Purity | QE |
|---|---|---|---|
| wdbc | SOM1 | 94.9550 | 1.9993 |
| | SOM2 | 97.2777 | 2.0749 |
| | SOM12 | 👍 | 👍 |
| | SOM21 | 👍 | 👍 |
| isolet 5x5 | SOM1 | 81.2081 | 12.6149 |
| | SOM2 | 95.1220 | 14.4591 |
| | SOM12 | 👍 | 👍 |
| | SOM21 | 👍 | 👍 |
| madelon | SOM1 | 60.8879 | 15.5896 |
| | SOM2 | 62.6402 | 15.5065 |
| | SOM12 | 👍 | 👍 |
| | SOM21 | 👍 | 👍 |
| spam | SOM1 | 83.8603 | 3.4582 |
| | SOM2 | 85.7205 | 2.5580 |
| | SOM12 | 👎 | 👍 |
| | SOM21 | 👎 | 👍 |

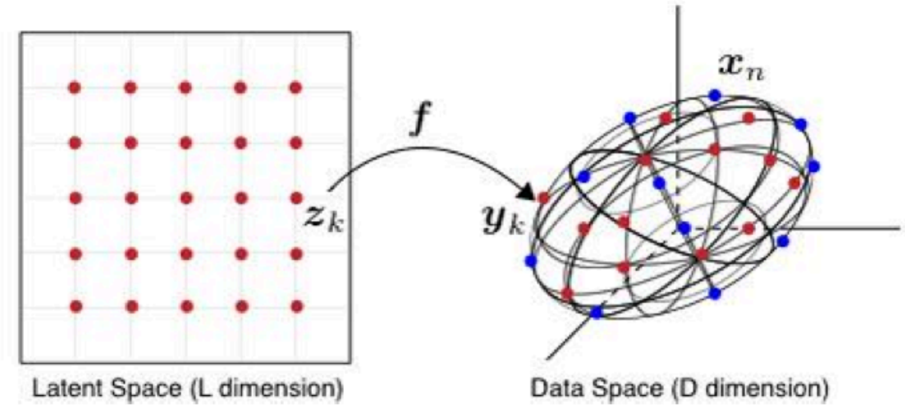
## Collaboration verticale

| Dataset | DB.. | Purity | QE |
|---|---|---|---|
| wdbc | SOM1 | 96.7153 | 90.5413 |
| | SOM2 | 97.8723 | 67.6035 |
| | SOM12 | 👍 | 👍 |
| | SOM21 | 👎 | 👍 |
| isolet 5x5 | SOM1 | 98.8506 | 8.1904 |
| | SOM2 | 98.4615 | 8.7671 |
| | SOM12 | 👎 | 👎 |
| | SOM21 | 👎 | 👎 |
| madelon | SOM1 | 69.7198 | 612.3251 |
| | SOM2 | 69.8718 | 611.5365 |
| | SOM12 | 👍 | 👍 |
| | SOM21 | 👍 | 👍 |
| spam | SOM1 | 76.2624 | 61.8324 |
| | SOM2 | 70.4306 | 48.2763 |
| | SOM12 | 👎 | 👍 |
| | SOM21 | 👎 | 👍 |

# Probabilistic Collaborative Clustering

# Probabilistic Clustering

**Generative Topographic Mapping [Bishop 95]**




$$\boxed{y = y(z, W) = W\Phi(z)}$$

$$\boxed{p(x_n|z, W, \beta) = \mathcal{N}(y(z, W), \beta)}$$

$$\mathcal{L}(W, \beta) = \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{i=1}^{K} p(x_n|z_i, W, \beta) \right\} \Longrightarrow \boxed{\textbf{EM}\ \text{Algorithm}}$$

# E & M steps

## E step - Computing posterior probabilites

$$
\begin{aligned}
r_{in} &= p(z_i | x_n, W_{old}, \beta_{old}) \\
&= \frac{p(x_n | z_i, W_{old}, \beta_{old})}{\sum_{i'=1}^{K} p(x_n | z_i', W_{old}, \beta_{old})}
\end{aligned}
$$

## M step - Updating parameters

$$\mathbb{E}[\mathcal{L}_{comp}(W, \beta)] = \sum_{n=1}^{N} \sum_{i=1}^{K} r_{in} \ln\{p(x_n | z_i, W, \beta)\}$$

$$\Phi^T G \Phi W_{new}^T = \Phi^T R X$$

$$\frac{1}{\beta_{new}} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{i=1}^{K} r_{in} \|x_n - W^{new} \phi(z_i)\|^2$$

# Topological Collaborative Clustering

Collaborative Clustering : **local step + collaboration step**

$$R_H^{[ii]}(W) = R_{Quantiz}(W) + R_{Collab}(W)$$

- Prototype based Clustering

$$R_{Quantiz}(W) = \sum_{jj=1, jj \neq ii}^{P} \alpha_{[ii]}^{[jj]} \sum_{i=1}^{N} \sum_{j=1}^{|w|} \mathcal{K}_{\sigma(j,\chi(x_i))}^{[ii]} \|x_i^{[ii]} - w_j^{[ii]}\|^2$$

$$R_{Collab}(W) = \sum_{jj=1, jj \neq ii}^{P} \beta_{[ii]}^{[jj]} \sum_{i=1}^{N} \sum_{j=1}^{|w|} \left(\mathcal{K}_{\sigma(j,\chi(x_i))}^{[ii]} - \mathcal{K}_{\sigma(j,\chi(x_i))}^{[jj]}\right)^2 * \|x_i^{[ii]} - w_j^{[ii]}\|^2$$

- Probabilistic Clustering

$$\mathcal{L}^{hor}[ii] = \mathbb{E}[\mathcal{L}_{comp}(W^{[ii]}, \beta^{[ii]})] - \sum_{[jj]=1, [jj] \neq [ii]}^{P} \alpha_{[ii]}^{[jj]} \sum_{n=1}^{N} \sum_{i=1}^{K} \frac{\beta^{[ii]}}{2} (r_{in}^{[ii]} - r_{in}^{[jj]})^2 \|x_n - W^{[ii]} \phi^{[ii]}(z_i)\|^2$$

# Collaborative Generative Topographic Mapping

## Horizontal approach

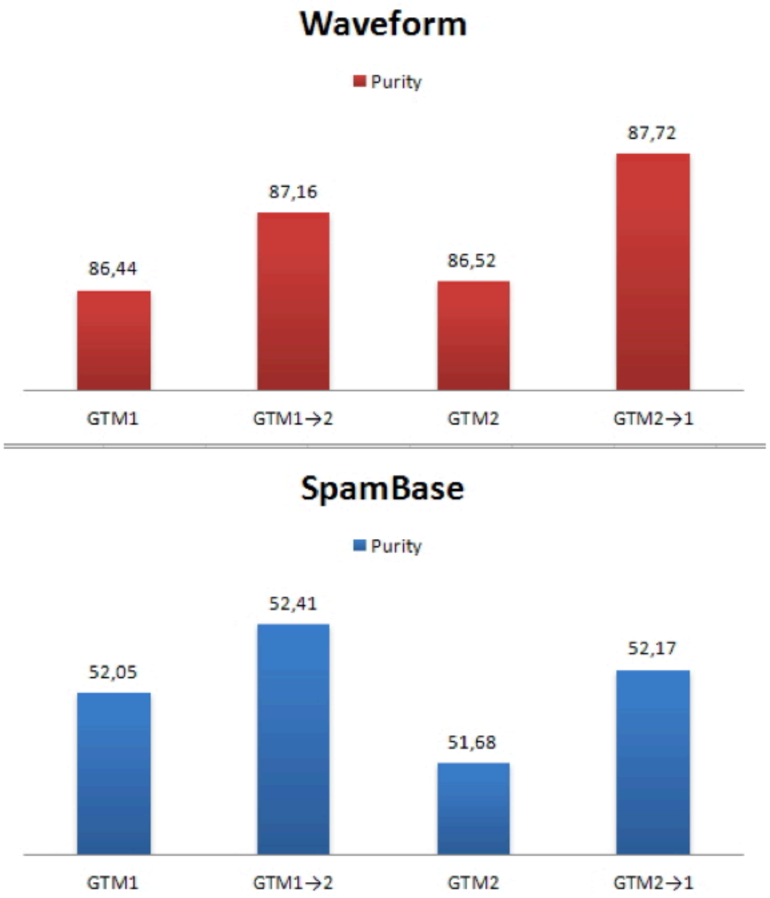
$$\mathcal{L}^{hor}[ii] = \mathbb{E}[\mathcal{L}_{comp}(W^{[ii]}, \beta^{[ii]})] - \sum_{[jj]=1,[jj]\neq[ii]}^{P} \alpha_{[ii]}^{[jj]} \sum_{n=1}^{N} \sum_{i=1}^{K} \frac{\beta^{[ii]}}{2} (r_{in}^{[ii]} - r_{in}^{[jj]})^2 \| x_n - W^{[ii]} \phi^{[ii]}(z_i) \|^2$$

## Vertical approach

$$\mathcal{L}^{ver}[ii] = \mathbb{E}[\mathcal{L}_{comp}(W^{[ii]}, \beta^{[ii]})] - \sum_{[jj]=1,[jj]\neq[ii]}^{P} \alpha_{[ii]}^{[jj]} \sum_{n=1}^{N[ii]} \sum_{i=1}^{K} r_{in} \frac{\beta^{[ii]}}{2} \| W^{[ii]} \phi^{[ii]}(z_i) - W^{[jj]} \phi^{[jj]}(z_i) \|^2$$

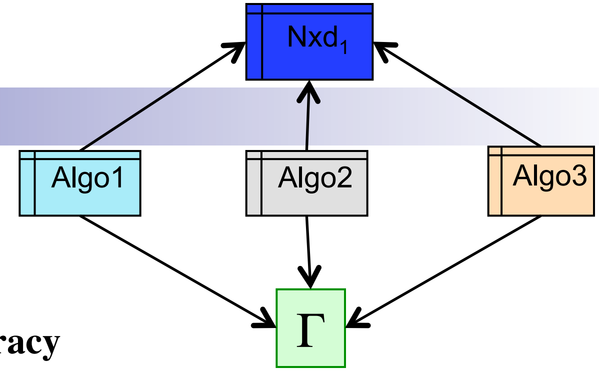
# Some experimental results

| Dataset | Map | Purity |
|---|---|---|
| Waveform | $GTM_1$ | 86.44 |
| | $GTM_2$ | 86.52 |
| | $GTM_{1\to 2}$ | 87.16 |
| | $GTM_{2\to 1}$ | 87.72 |
| Wdbc | $GTM_1$ | 96 |
| | $GTM_2$ | 96.34 |
| | $GTM_{1\to 2}$ | 96.08 |
| | $GTM_{2\to 1}$ | 96.15 |
| Isolet | $GTM_1$ | 87.17 |
| | $GTM_2$ | 86.83 |
| | $GTM_{1\to 2}$ | 87.29 |
| | $GTM_{2\to 1}$ | 85.87 |
| SpamBase | $GTM_1$ | 52.05 |
| | $GTM_2$ | 51.68 |
| | $GTM_{1\to 2}$ | 52.41 |
| | $GTM_{2\to 1}$ | 52.17 |



# Collaborative Clustering
# Diversity analysis

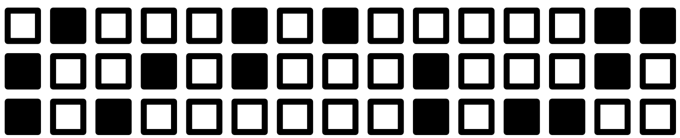
# Diversity : why?




**Studied in Consensus clustering**

Dataset X containing 15 samples

**accuracy**



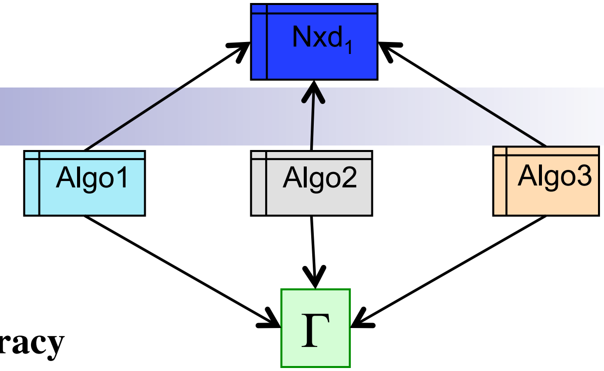
**Algo1** 10/15 = **0.667**

**Algo2** 10/15 = **0.667**

**Algo3** 10/15 = **0.667**

□ Correct ■ Wrong

# Diversity : why?




**Studied in Consensus clustering**

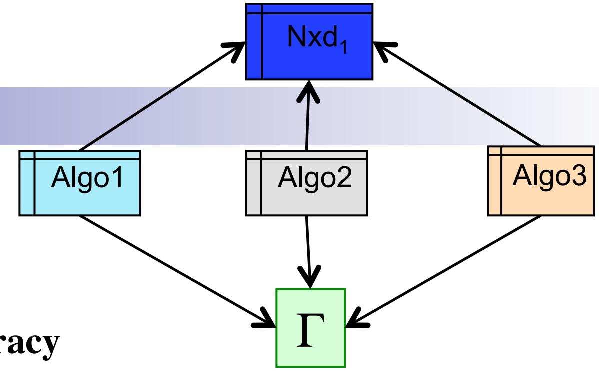
Dataset X containing 15 samples

| | | | | | | | | | | | | | | | | **accuracy** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Algo1** | □ | ■ | □ | □ | □ | ■ | □ | ■ | □ | □ | □ | □ | □ | ■ | ■ | 10/15 = **0.667** |
| **Algo2** | ■ | □ | □ | ■ | □ | ■ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | 10/15 = **0.667** |
| **Algo3** | ■ | □ | ■ | □ | □ | □ | □ | □ | □ | ■ | □ | ■ | ■ | □ | □ | 10/15 = **0.667** |
| **Fusion** | ■ | □ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | 11/15 = **0.773** |

□ Correct ■ Wrong

Majority vote rule

# Diversity : why?




## Studied in Consensus clustering

Dataset X containing 15 samples

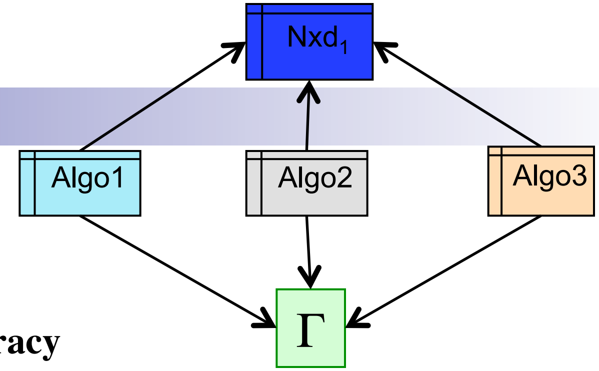
| | | | | | | | | | | | | | | | | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Algo1** | □ | ■ | □ | □ | □ | ■ | □ | ■ | □ | □ | □ | □ | □ | ■ | ■ | 10/15 = **0.667** |
| **Algo2** | ■ | □ | □ | ■ | □ | ■ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | 10/15 = **0.667** |
| **Algo3** | ■ | □ | ■ | □ | □ | □ | □ | □ | □ | ■ | □ | ■ | ■ | □ | □ | 10/15 = **0.667** |
| **Fusion** | ■ | □ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | 11/15 = **0.773** |

□ Correct ■ Wrong

Majority vote rule

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Algo1** | ■ | □ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | □ | □ | □ | 10/15 = **0.667** |
| **Algo2** | ■ | □ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | □ | □ | □ | 10/15 = **0.667** |
| **Algo3** | ■ | □ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | □ | □ | □ | 10/15 = **0.667** |

# Diversity : why?




## Studied in Consensus clustering

Dataset X containing 15 samples

□ Correct ■ Wrong

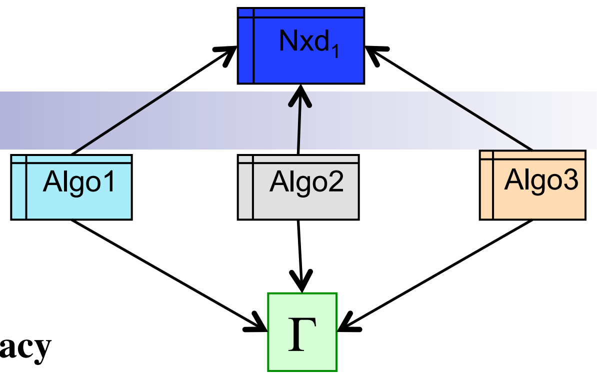
| | | accuracy |
|---|---|---|
| **Algo1** | □■□□□■□■□□□□□■■ | 10/15 = **0.667** |
| **Algo2** | ■□□■□■□□□■□□□■□ | 10/15 = **0.667** |
| **Algo3** | ■□■□□□□□□■□■■□□ | 10/15 = **0.667** |
| **Fusion** | ■□□□□■□□□■□□□■□ | 11/15 = **0.773** |

Majority vote rule

| | | |
|---|---|---|
| **Algo1** | ■□■■■■□□□□□□□□□ | 10/15 = **0.667** |
| **Algo2** | ■□■■■■□□□□□□□□□ | 10/15 = **0.667** |
| **Algo3** | ■□■■■■□□□□□□□□□ | 10/15 = **0.667** |
| **Fusion** | ■□■■■■□□□□□□□□□ | 10/15 = **0.667** |

Majority vote rule

# Diversity : why?




## Studied in Consensus clustering

Dataset X containing 15 samples

□ Correct ■ Wrong

| | | | | | | | | | | | | | | | | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Algo1** | □ | ■ | □ | □ | □ | ■ | □ | ■ | □ | □ | □ | □ | □ | ■ | ■ | 10/15 = **0.667** |
| **Algo2** | ■ | □ | □ | ■ | □ | ■ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | 10/15 = **0.667** |
| **Algo3** | ■ | □ | ■ | □ | □ | □ | □ | □ | □ | ■ | □ | ■ | ■ | □ | □ | 10/15 = **0.667** |
| **Fusion** | ■ | □ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | □ | □ | ■ | □ | 11/15 = **0.773** |

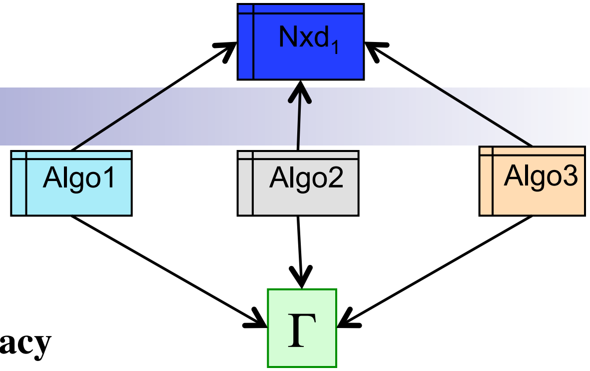
Majority vote rule

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Algo1** | ■ | □ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | □ | □ | □ | 10/15 = **0.667** |
| **Algo2** | ■ | □ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | □ | □ | □ | 10/15 = **0.667** |
| **Algo3** | ■ | □ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | □ | □ | □ | 10/15 = **0.667** |
| **Fusion** | ■ | □ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | □ | □ | □ | 10/15 = **0.667** |

Majority vote rule

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Algo1** | ■ | □ | ■ | ■ | □ | □ | □ | ■ | □ | ■ | □ | □ | □ | □ | □ | 10/15 = **0.667** |
| **Algo2** | ■ | □ | □ | ■ | □ | ■ | □ | □ | □ | □ | ■ | □ | □ | ■ | □ | 10/15 = **0.667** |
| **Algo3** | □ | □ | ■ | □ | □ | ■ | □ | □ | □ | ■ | ■ | □ | □ | ■ | □ | 10/15 = **0.667** |

# Diversity : why?




## Studied in Consensus clustering

Dataset X containing 15 samples

□ Correct ■ Wrong

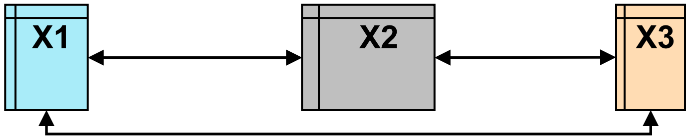
| | Samples | accuracy | |
|---|---|---|---|
| **Algo1** | □■□□□■□■□□□□□■■ | 10/15 = **0.667** | |
| **Algo2** | ■□□■□■□□□■□□□■□ | 10/15 = **0.667** | |
| **Algo3** | ■□■□□□□□□■□■■□□ | 10/15 = **0.667** | |
| **Fusion** | ■□□□□■□□□■□□□■□ | 11/15 = **0.773** | Majority vote rule |

| | Samples | accuracy | |
|---|---|---|---|
| **Algo1** | ■□■■■■□□□□□□□□□ | 10/15 = **0.667** | |
| **Algo2** | ■□■■■■□□□□□□□□□ | 10/15 = **0.667** | |
| **Algo3** | ■□■■■■□□□□□□□□□ | 10/15 = **0.667** | |
| **Fusion** | ■□■■■■□□□□□□□□□ | 10/15 = **0.667** | Majority vote rule |

| | Samples | accuracy | |
|---|---|---|---|
| **Algo1** | ■□■■□□□■□■□□□□□ | 10/15 = **0.667** | |
| **Algo2** | ■□□■□■□□□□■□□■□ | 10/15 = **0.667** | |
| **Algo3** | □□■□□■□□□■■□□■□ | 10/15 = **0.667** | |
| **Fusion** | ■□■■□■□□□■■□□■□ | 8/15 = **0.533** | Majority vote rule |

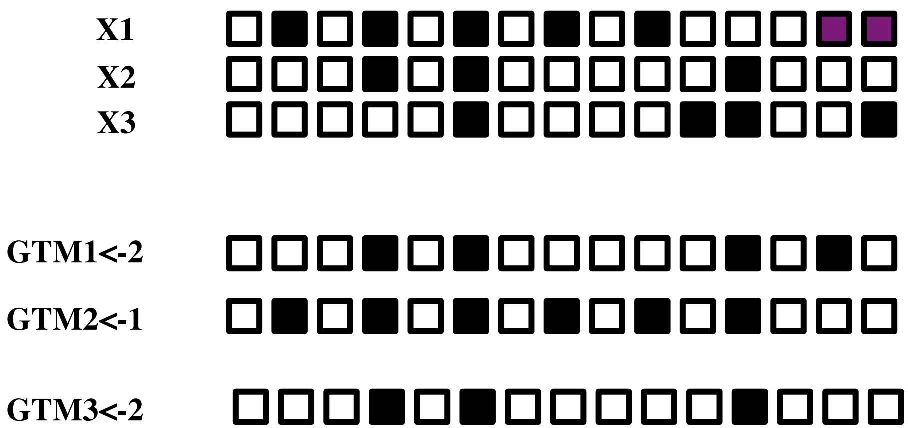
# Diversity (2)

**Collaborative clustering**




Dataset X1 containing 15 samples
Dataset X2 containing 15 samples
Dataset X3 containing 15 samples

□ Correct ■ Wrong




accuracy

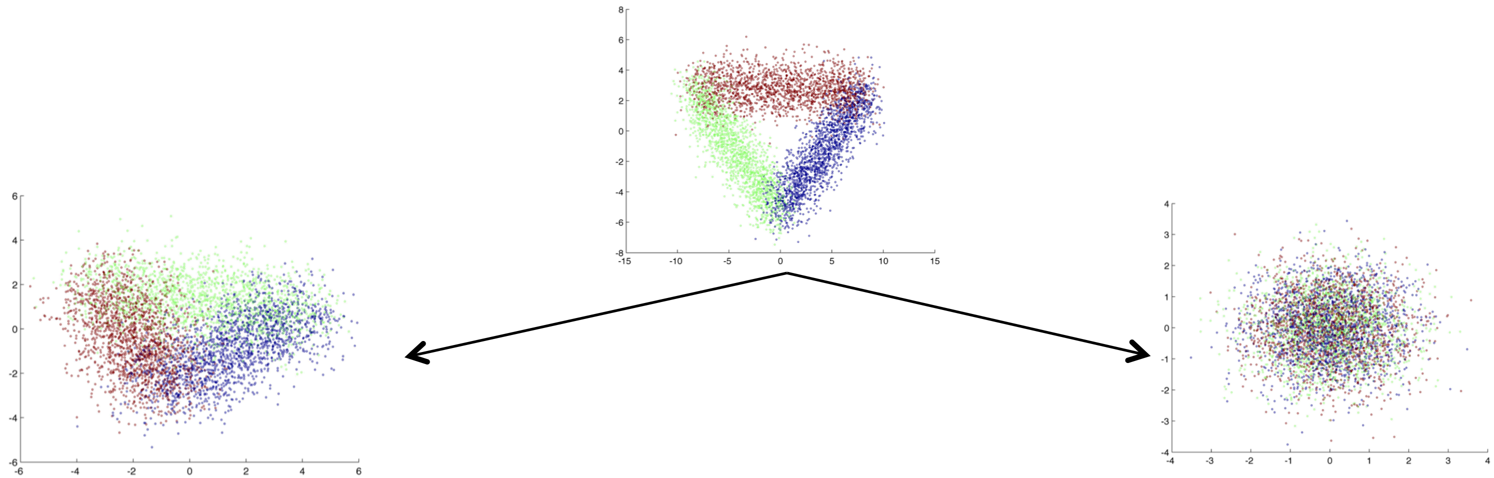
8/15 = **0.533**
12/15 = **0.8**
11/15 = **0.733**

11/15 = **0.733**
10/15 = **0.6**
12/15 = **0.8**

# Diversity measures

| index | formula |
|---|---|
| Rand index | $Rand = \frac{a_{00} + a_{11}}{a_{00} + a_{01} + a_{10} + a_{11}}$ |
| Adjusted Rand index | $AdjustedRand = \frac{a_{00} + a_{11} - n_c}{a_{00} + a_{01} + a_{10} + a_{11} - n_c}$ |
| Jaccard index | $Jaccard = \frac{a_{11}}{a_{01} + a_{10} + a_{11}}$ |
| Wallace's coefficient | $W_{P1 \to P2} = \frac{a_{11}}{a_{11} + a_{10}}$ and $W_{P2 \to P1} = \frac{a_{11}}{a_{11} + a_{01}}$ |
| Adjusted Wallace index | $AW_{P1 \to P2} = \frac{W_{P1 \to P2} - Wi_{P1 \to P2}}{1 - Wi_{P1 \to P2}}$ |
| Normalized Mutual Information | $NMI = \frac{-2 \sum_{ij} n_{ij} log \frac{n_{ij} N}{n_i n_j}}{\sum_i n_i log \frac{n_i}{N} + \sum_j n_j log \frac{n_j}{N}}$ |
| Variation of Information | $VI = -2 \sum_{ij} \frac{n_{ij}}{N} log \frac{n_{ij} N}{n_i n_j} - \sum_i \frac{n_i}{N} log \frac{n_i}{N} - \sum_j \frac{n_j}{N} log \frac{n_j}{N}$ |

# Diversity measures on waveform datasets



(a) dataset db1

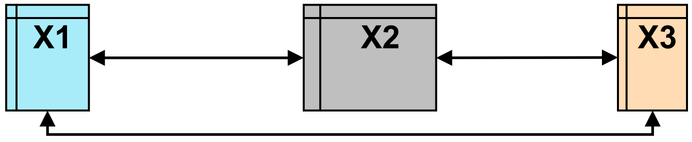
(b) dataset db6

**Table 1: Diversity measure on the waveform subsets**

| Subset | Relevant datasets | | Relevant vs Noisy datasets | | Noisy datasets | |
|---|---|---|---|---|---|---|
| Diversity index | db2/db3 | db3/db4 | db2/db8 | db4/db9 | db7/db8 | db9/db10 |
| Rand | 0.6707 | 0.7042 | 0.5539 | 0.555 | 0.543 | 0.5553 |
| Adjusted Rand | 0.2625 | 0.3356 | 0.00008 | 0.0002 | 0.00002 | 0.00004 |
| Jaccard | 0.3429 | 0.3869 | 0.2017 | 0.2008 | 0.2 | 0.2003 |
| Wallace's coefficient | 0.5079 | 0.5578 | 0.3332 | 0.3342 | 0.33 | 0.3334 |
| Adjusted Wallace | 0.5135 | 0.5581 | 0.3383 | 0.3347 | 0.35 | 0.3411 |
| Normal Mutual Information | 0.262 | 0.3072 | 0.0002 | 0.0006 | 0.0003 | 0.0004 |
| Variation of Information | 2.334 | 2.1918 | 3.1577 | 3.1631 | 3.168 | 3.1664 |

# Diversity (2)

**Collaborative clustering**




Dataset X1 containing 15 samples
Dataset X2 containing 15 samples
Dataset X3 containing 15 samples

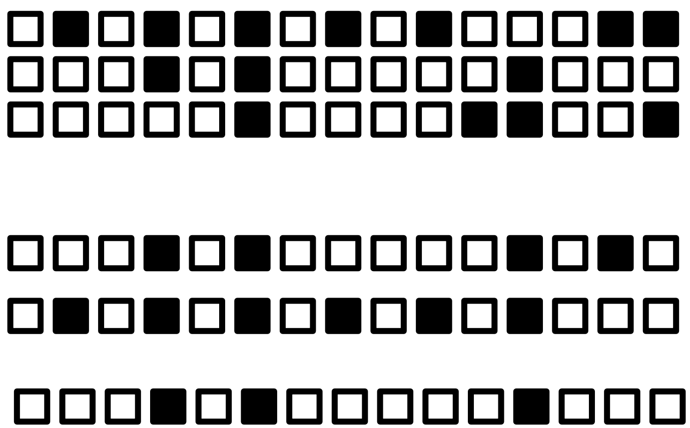
Correct Wrong

accuracy

diversity

X1-X2 =**0.956**

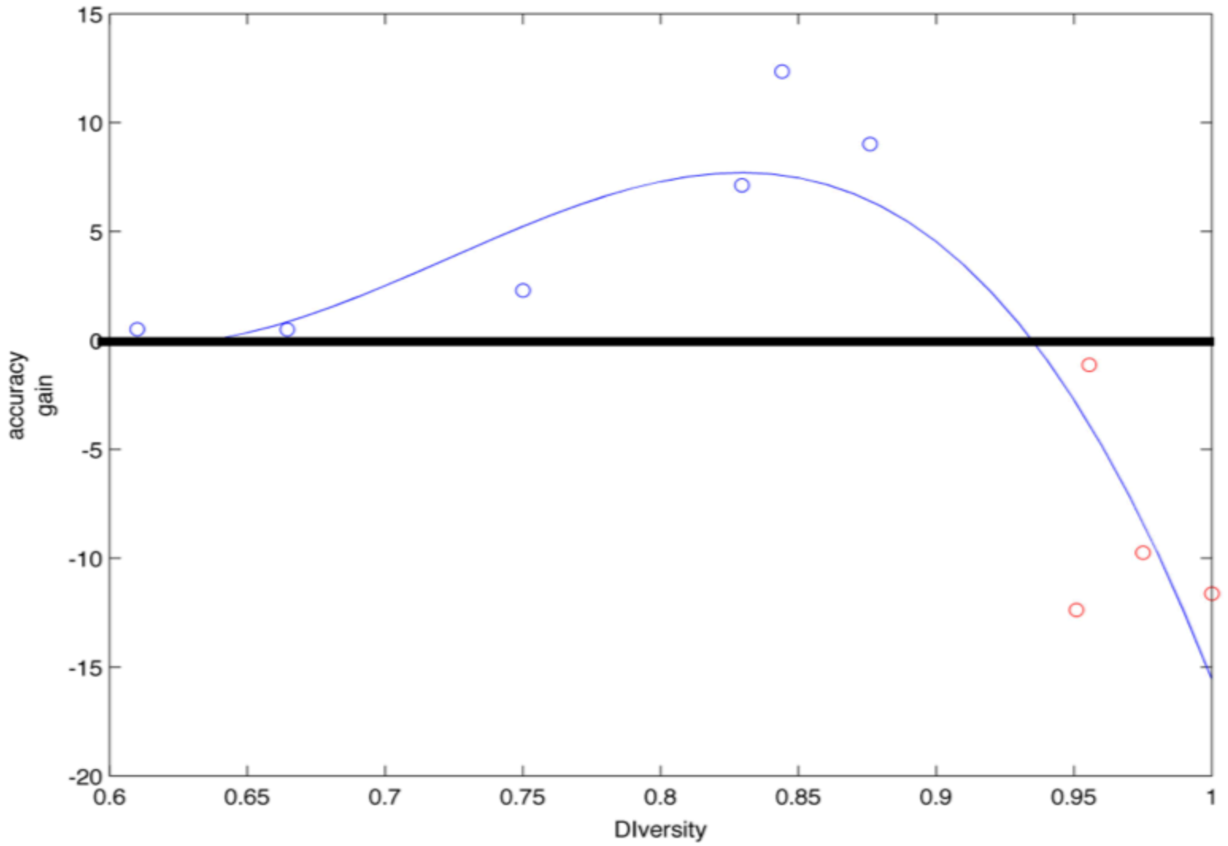
X2-X3 = **0.678**



| | accuracy |
|---|---|
| **X1** | 8/15 = **0.533** |
| **X2** | 12/15 = **0.8** |
| **X3** | 11/15 = **0.733** |
| **GTM1<-2** | 11/15 = **0.733** |
| **GTM2<-1** | 10/15 = **0.6** |
| **GTM3<-2** | 12/15 = **0.8** |

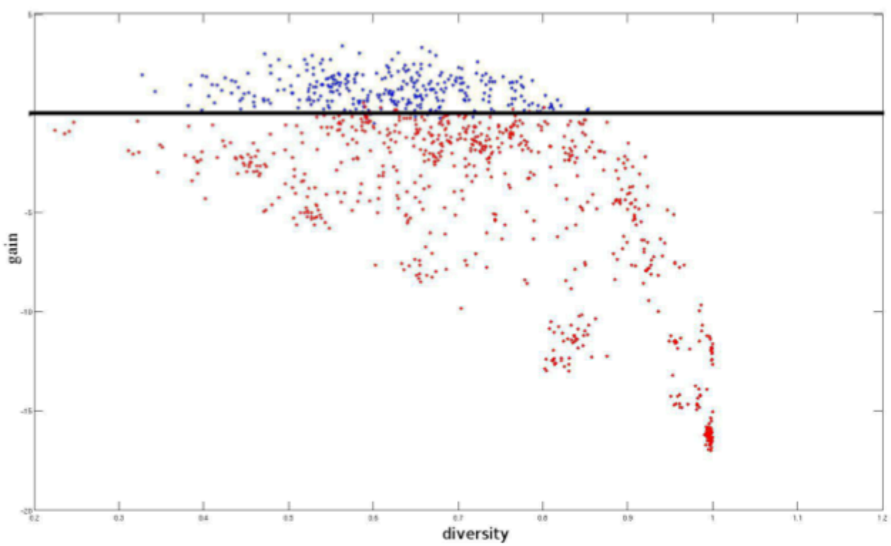
**Need to study the local quality.**

# Results : 10 waveform sub-sets

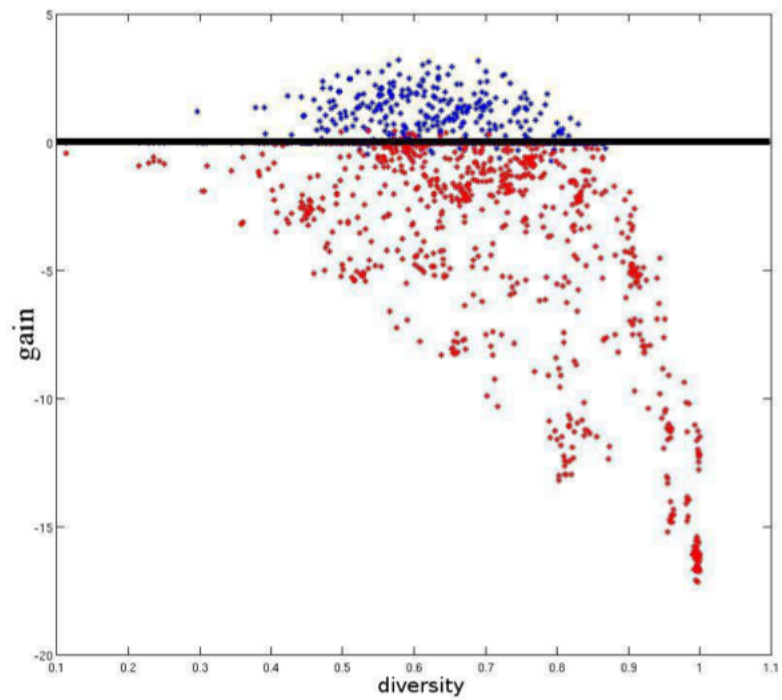



The plot of diversity and the accuracy difference after collaboration

# Results : 1-1.000 waveform sub-sets




(a) waveform subset 1

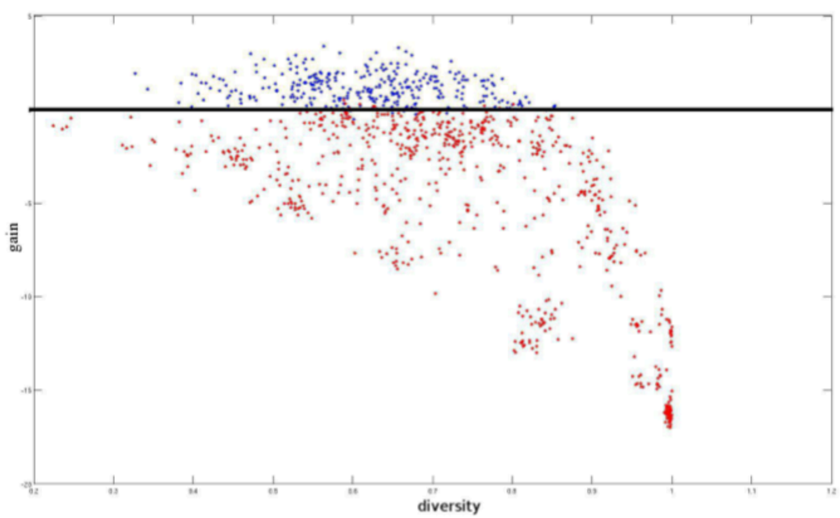



(b) waveform subset 2

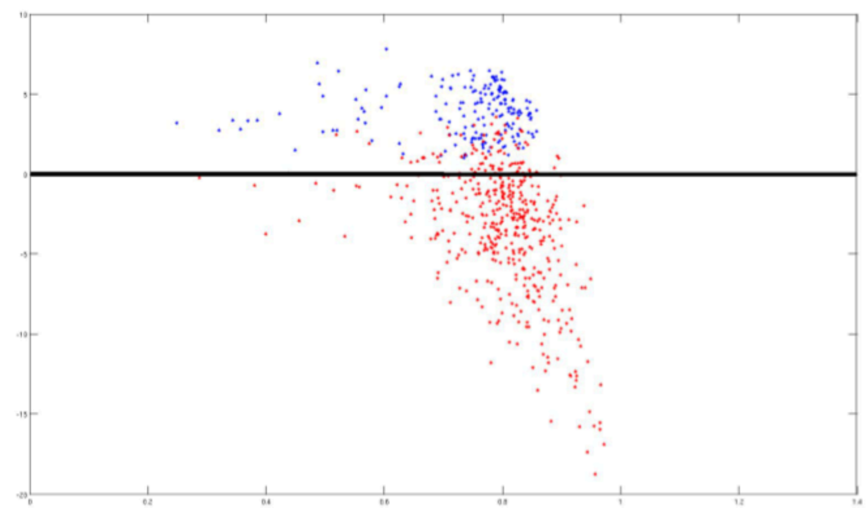
Waveform datasets: Collaboration results between a fixed subset and 1000 randomly subsets (axe X represents the Diversity and axe Y - the Accuracy gain)

# Collaboration results (1)

Collaboration results between a fixed subset and 1000 randomly subsets




(a) Waveform dataset

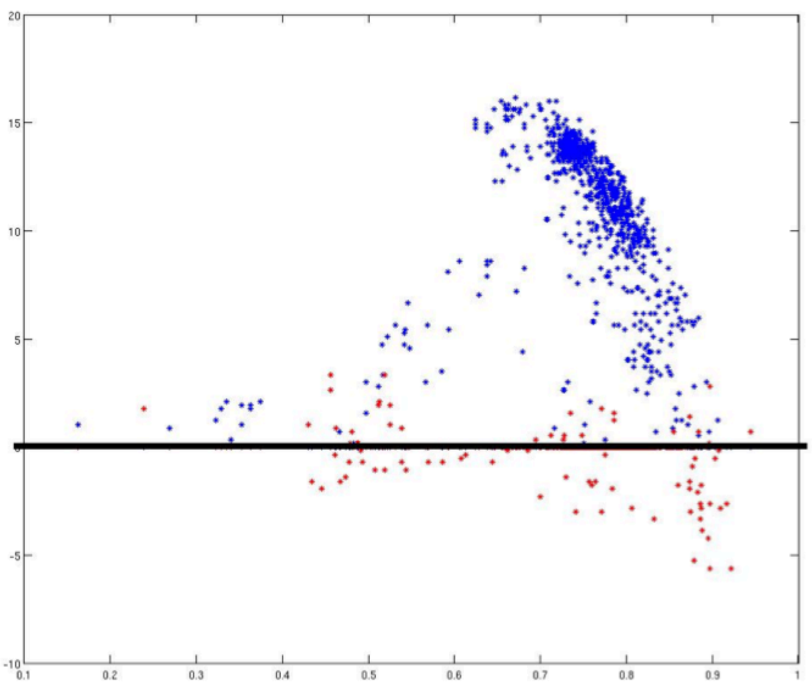


(d) Wdbc dataset

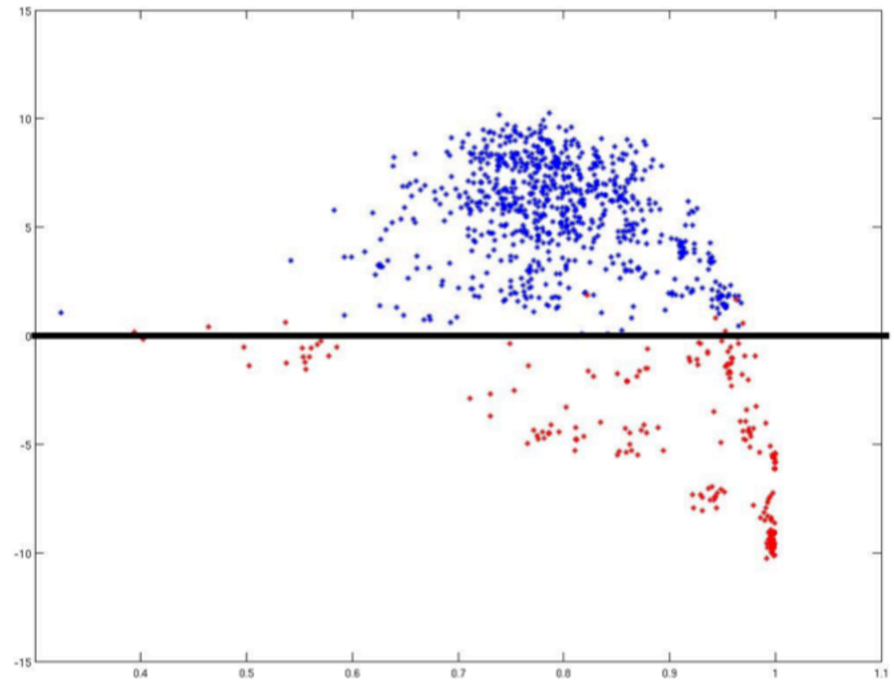
axe X represents the Diversity and axe Y - the Accuracy gain

# Collaboration results (2)

Collaboration results between a fixed subset and 1000 randomly subsets




(b) SpamBase dataset




(c) Isolet dataset

axe X represents the Diversity and axe Y - the Accuracy gain

Collaborative Clustering and Consensus Clustering

real applications

# Images : Strasbourg satellite image (1)

**Projet COCLICO**

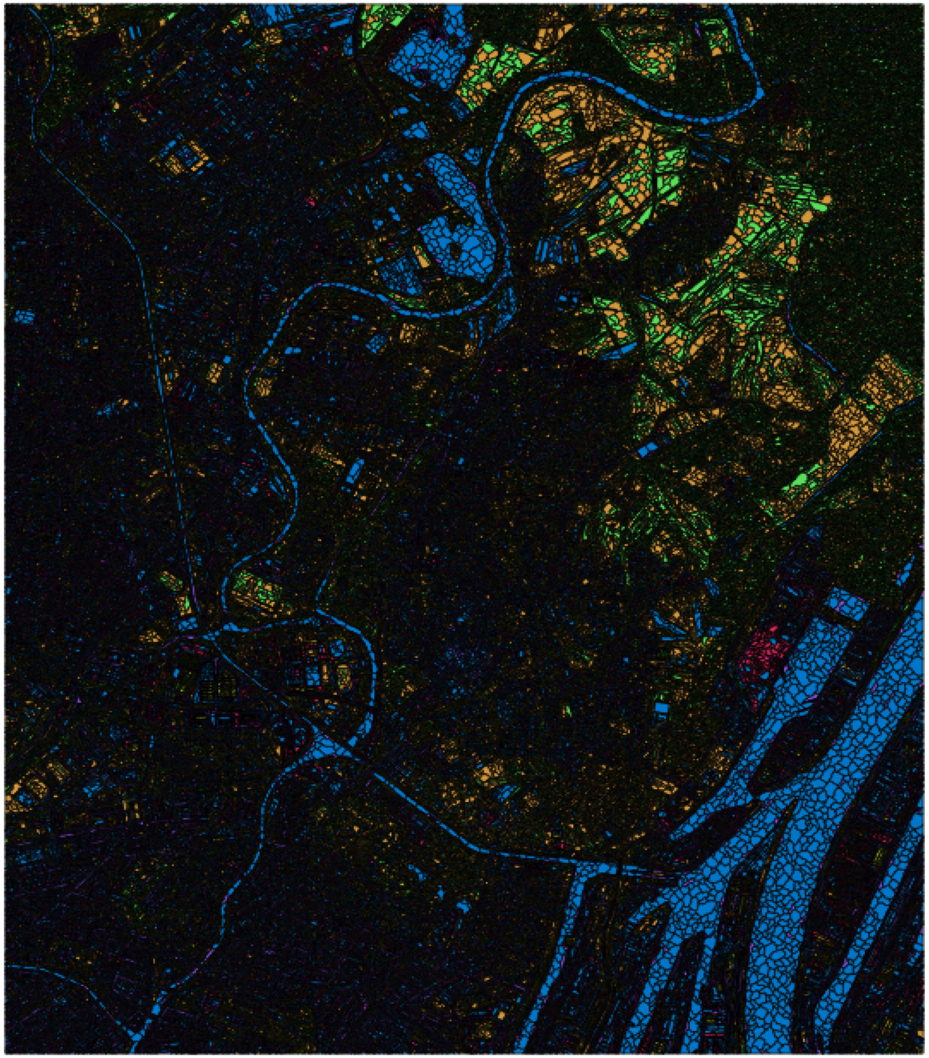
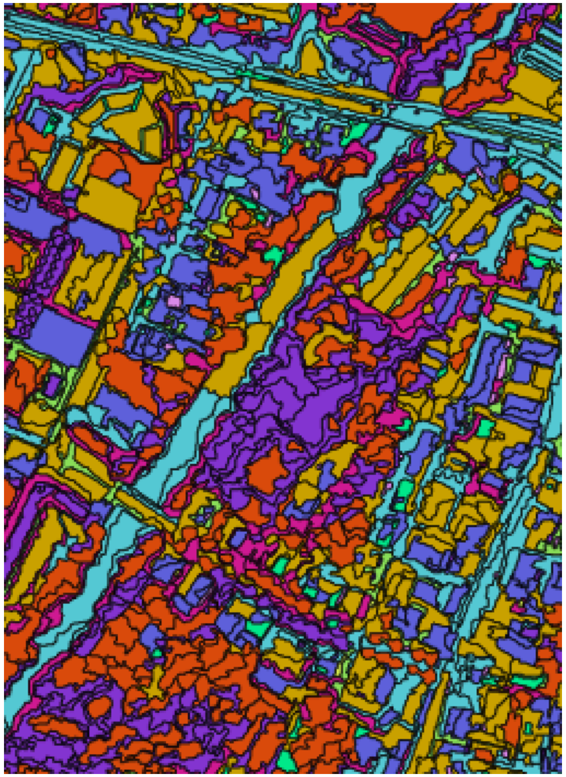


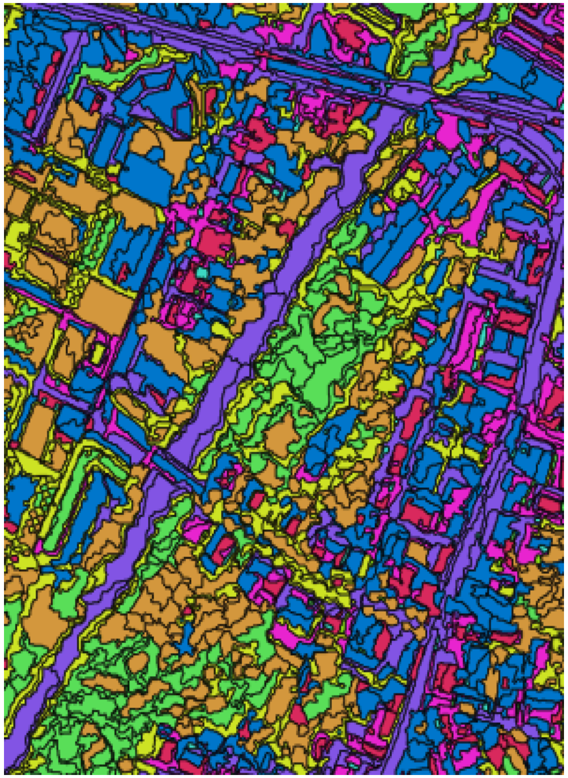
Image Ville Environnement

The authors would like to thank CESBIO (Danielle Ducrot, ClaireMarais-Sicre, Olivier Hagolle, Mireille Huc and Jordi Inglada) for providing the land-cover maps and the geometrically and radiometrically corrected Formosat-2 images.

Before collaboration



After collaboration



# System for searching visual information

# Multimodal information : prediction

## Application context:

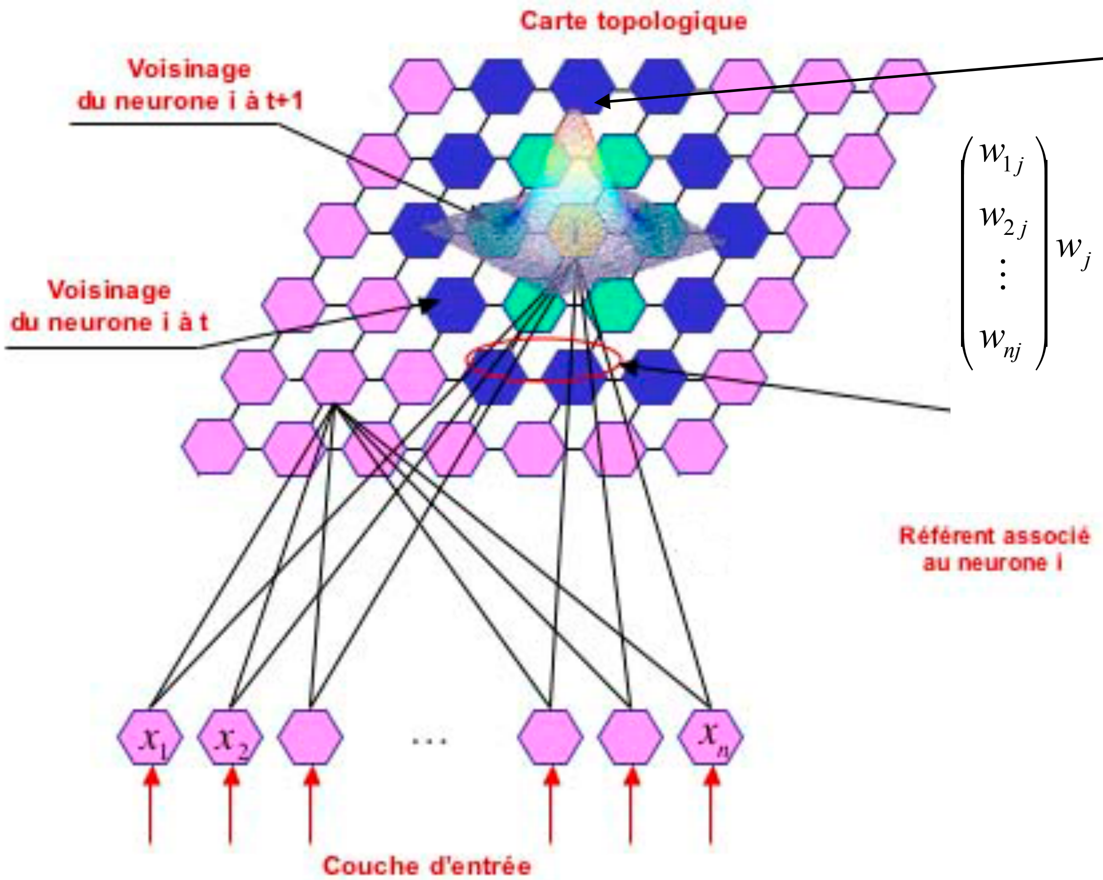
- A Wikipedia dataset containing a set of 20.000 images from wikipedia and ttransformed into numerical values by Xerox research center.

## Resulats:

- Reduce the dimensionality of this dataset of size 20.000 x 12.800 into 20.000 x 10
- **Patent** (THALES, Paris 13 University) N°: 08 06947
  Inventors: BENHADDA H., BENNANI Y., LEBBAH M., GROZAVU N.

# Recall : topological learning

$$\Re_{SOM}(\chi, W) = \sum_{i=1}^{N} \sum_{j=1}^{|W|} \mathrm{K}_{j,\chi(x_i)} \left\| x_i - w_j \right\|^2$$



# Conventional search engine


Example of image search using a traditional search engine *(France flag)*

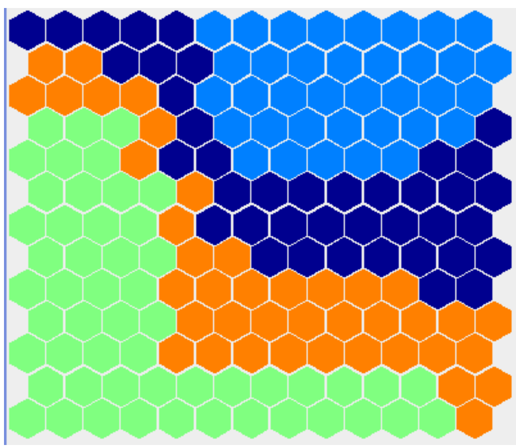
Research time: **0,11s** ;

Browsing the collection of images by user : **15 days**

*( 13.700.000 images / 21 images/pages = 652.380 pages * 2 s = 1.304.760s (362h or 15 days) )*

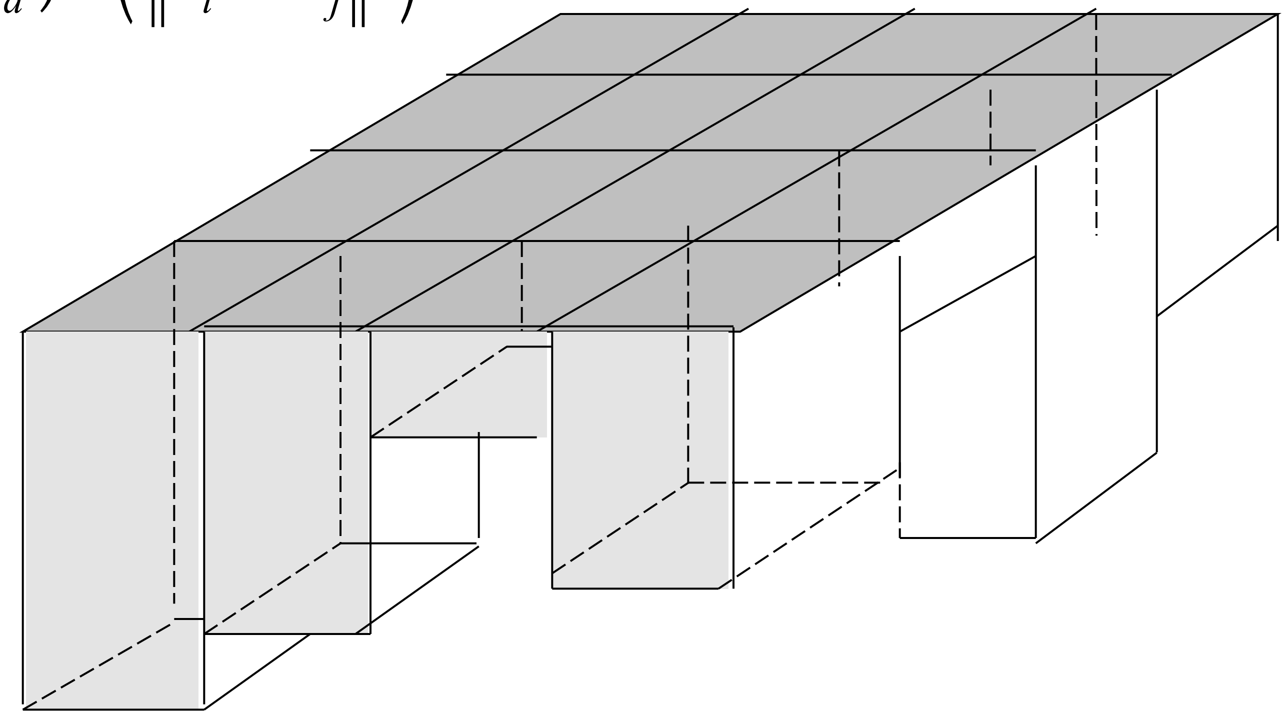
# Map of images

Wikipedia (19.000 x 6400) x 2

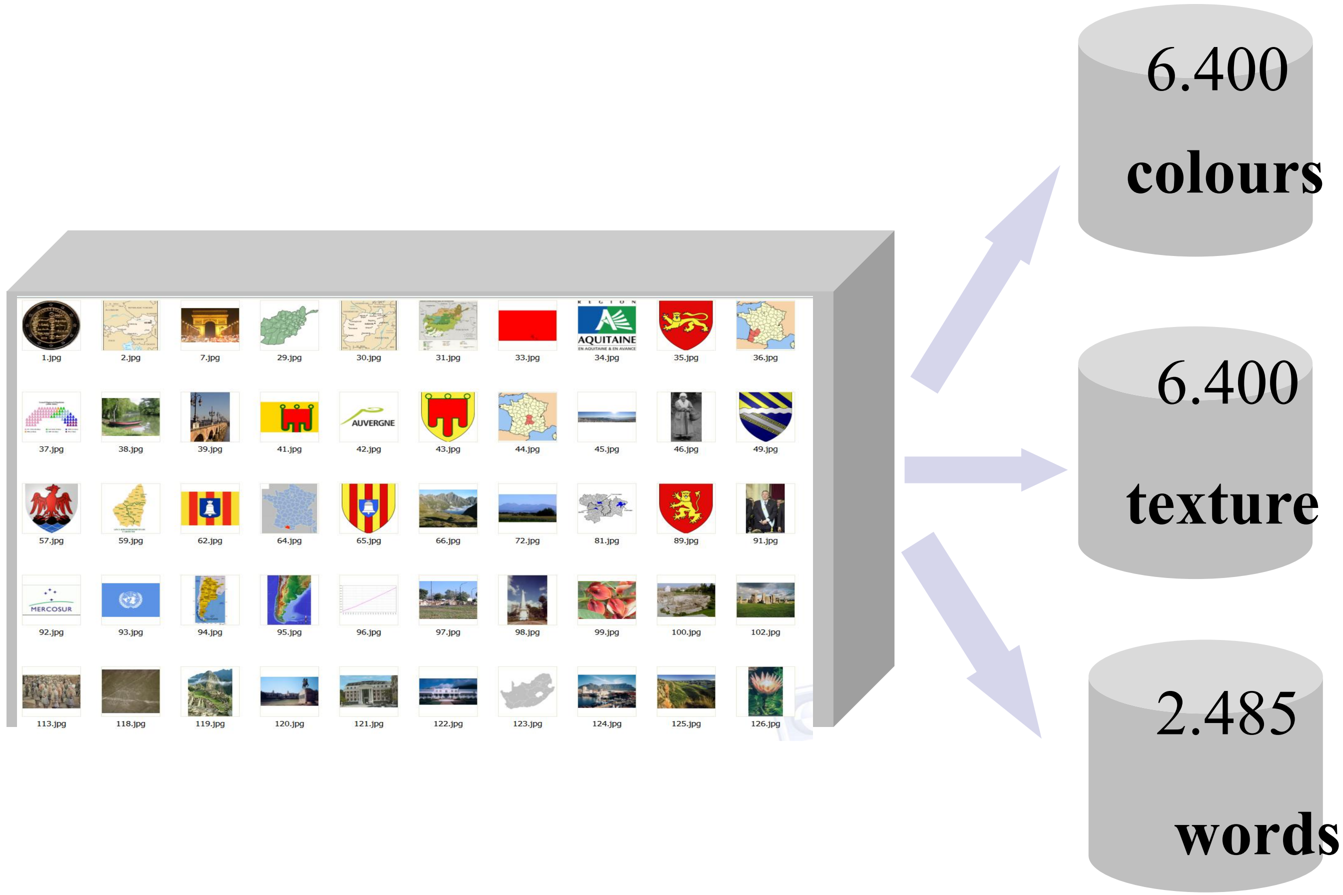


# Hierarchical SOM (3D view)

$$\chi\left(M_d\right) = \left(\left\|x_i - w_j\right\|^2\right)$$



# Intelligent system based on the Topological Learning




**Feature extraction from images**

# Principle

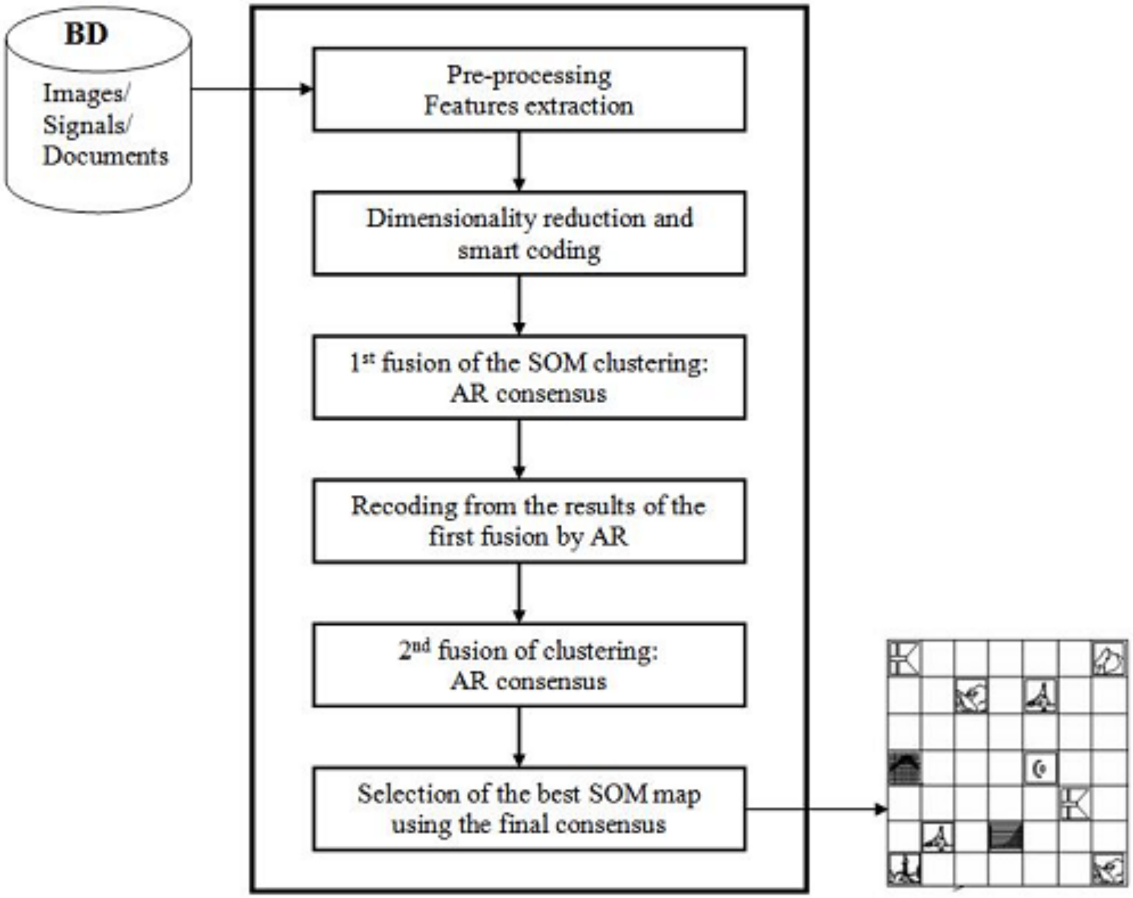
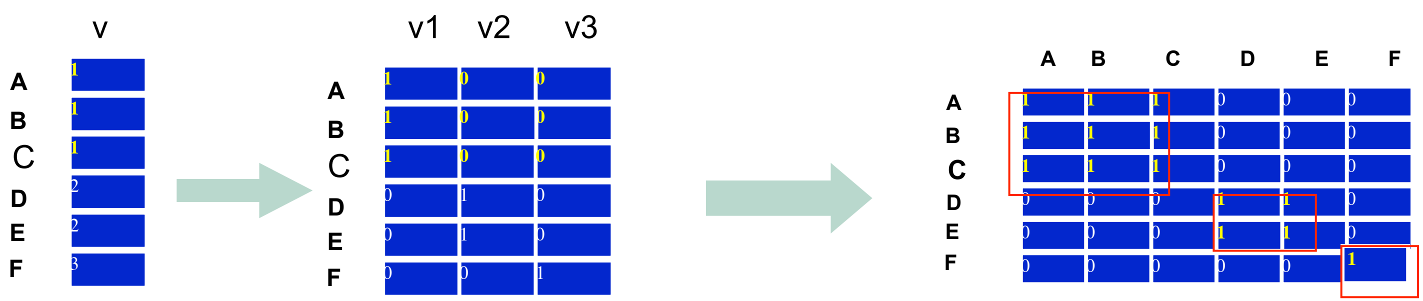



Figure 5.6: General Schema of the Fusion Procedure

# Relational Analysis (Marcotorchino al. 1980)

## 1- Pairwise comparison principle



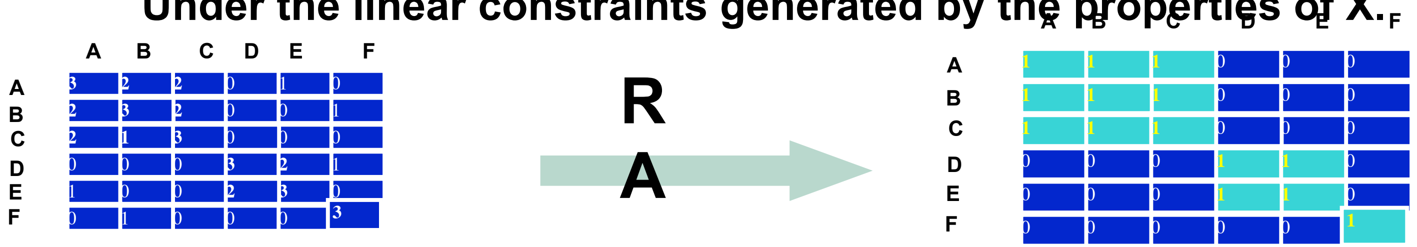

Linear coding Complete Disjonctive Coding Relational Coding

## 2- (0,1) Linear programming modelling

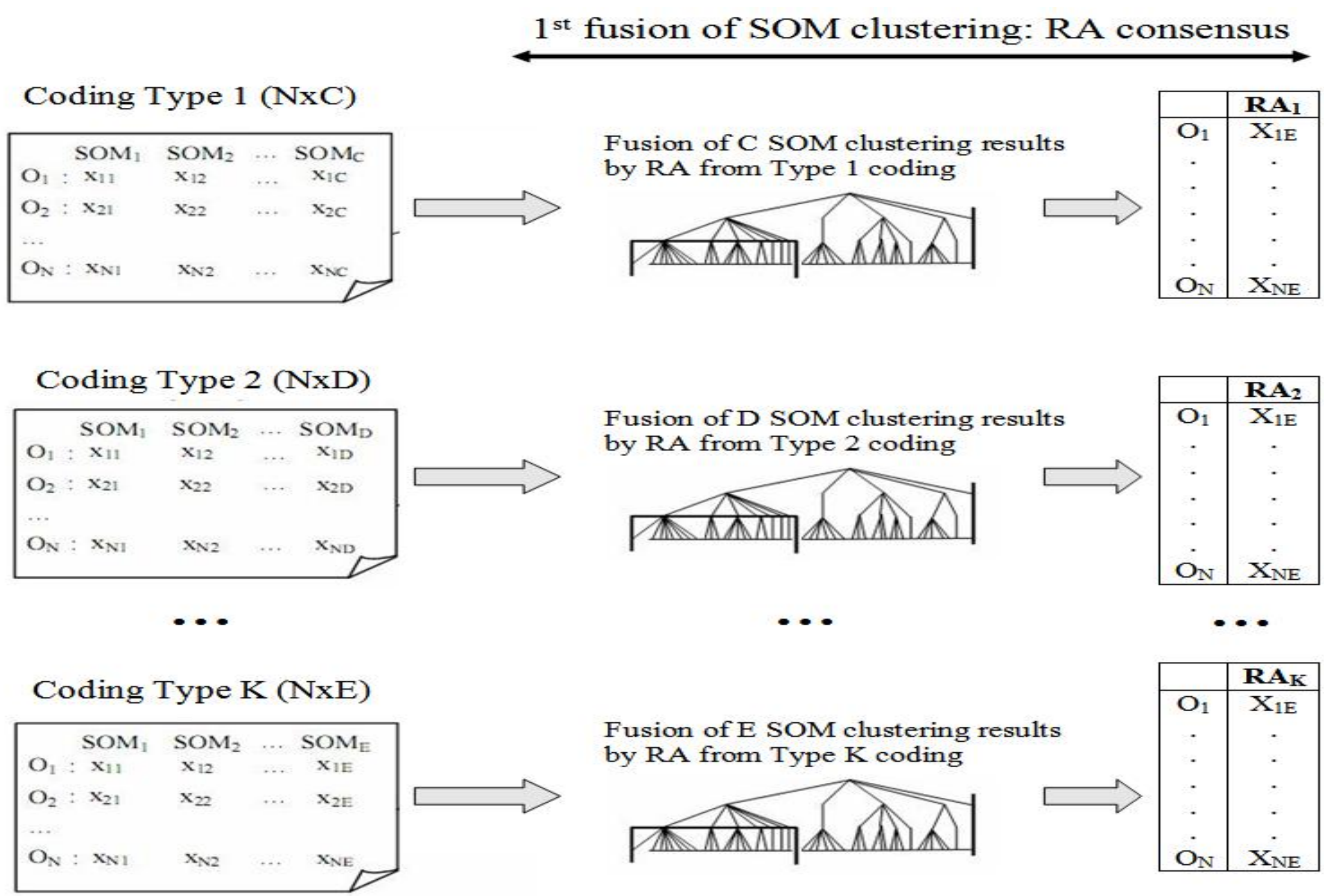
**We denote F(R, X) - the linear criterion measuring the adequacy between the data R and the solution X, the mathematical formulation of the problem is:**

$$\max_{X} F(R, X)$$

**Under the linear constraints generated by the properties of X.**



# Coding et Fusion



# Dimensionality reduction

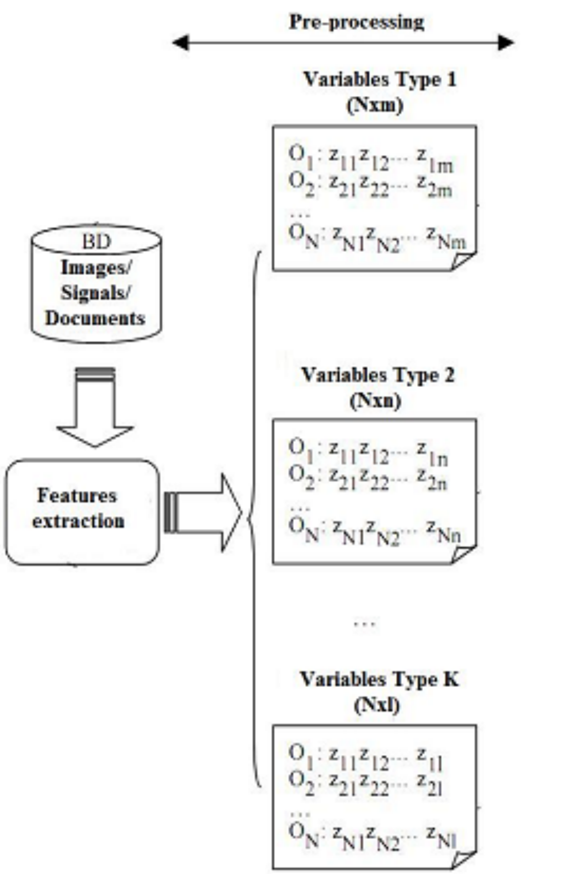



Figure 3.7: Pre-procession of the images dataset

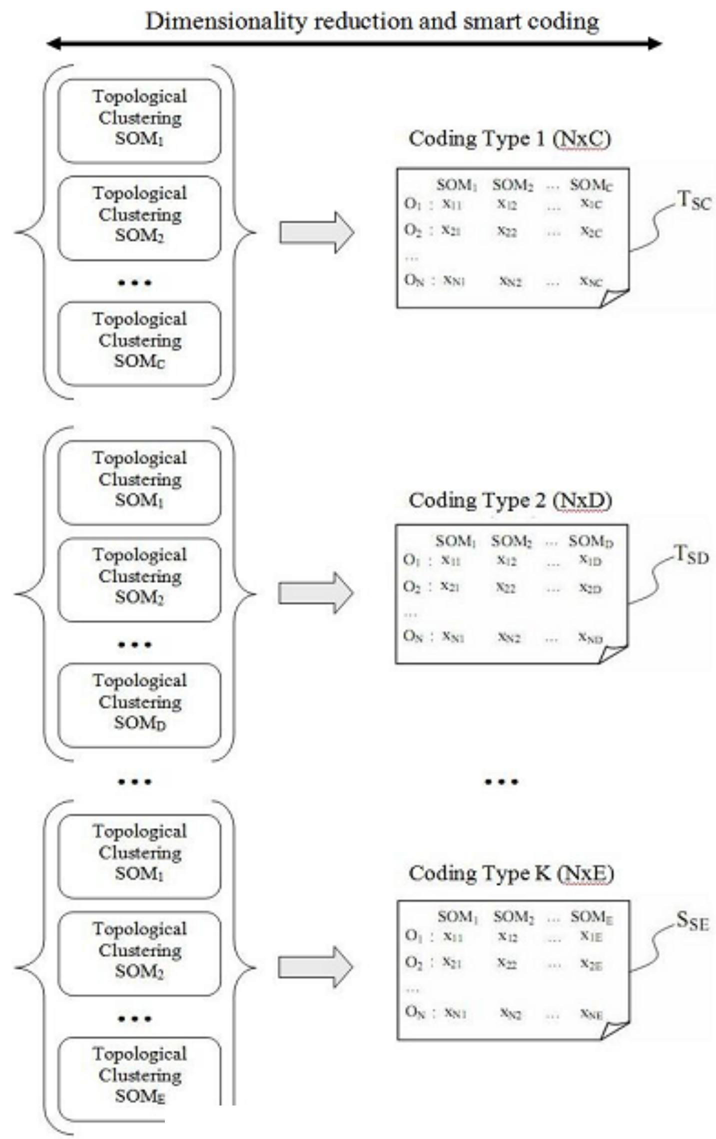
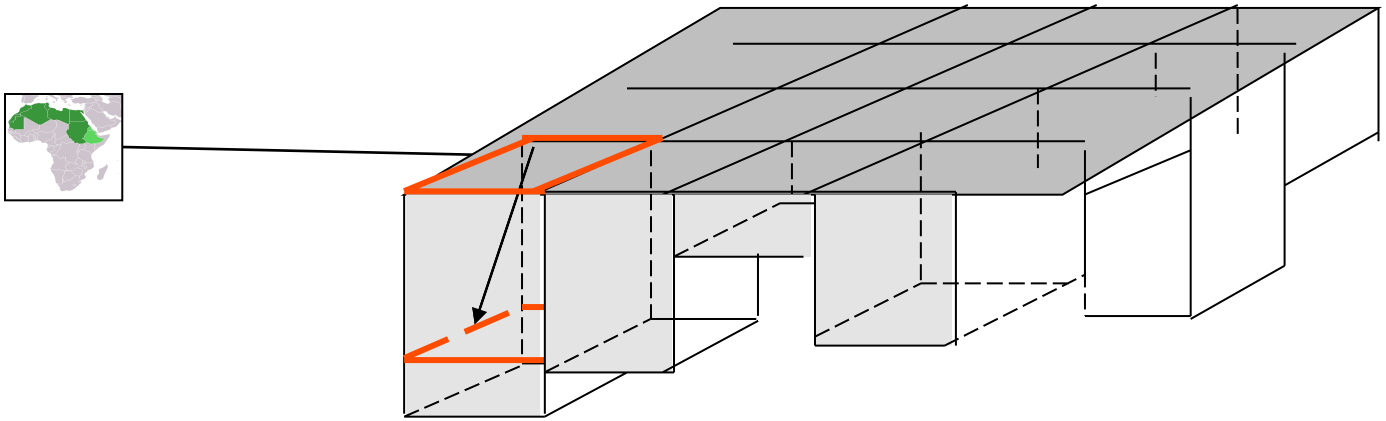



Figure 3.8: Dimensionality Reduction

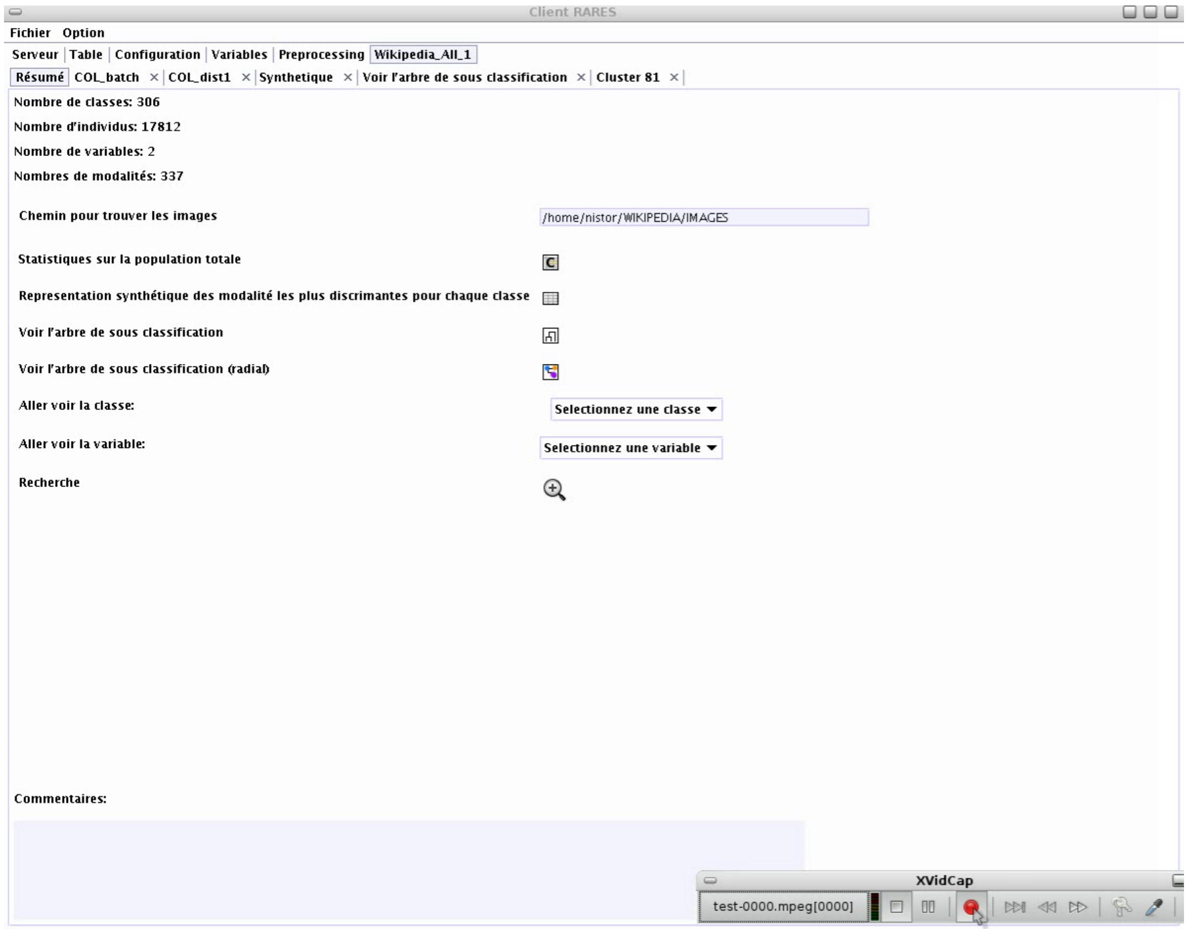
# New image assignment



$$\chi(x_i) = \arg\min_j \left( \left\| x_i - w_j \right\|^2 \right)$$

# Demonstration video

## BREVET (THALES, Université Paris 13 ) N°: 08 06947





BENHADDA H., BENNANI Y., LEBBAH M., GROZAVU N. «SYSTEME DE RECHERCHE D'INFORMATION VISUELLE», **BREVET** 08 06947.

LEBBAH M., BENNANI Y., BENHADDA H., GROZAVU N., (2009), «Relational Analysis for Clustering Consensus», Invited Book Chapter, Machine Learning, ISBN 978-953-7619-X-X, IN-TECH Publisher.

# Conclusions

- The collaborative clustering allows:
  - An interaction between different datasets
  - Reveal underlying structures and patterns within data sets.

- During the collaboration step, where is no need of data, the algorithm requires only the clustering results of other datasets.
  - obtain a new classification that is as close as possible to that which would have obtained if we had centralized datasets and then make a partition.

- The quality of the local clustering algorithm is very important for the collaboration's quality improvement regarding the diversity index
  - Overall, the variability of the collaboration's quality increase with the diversity

- Create a «*helper site*» which will build the global clustering and send these information to other local sites

- Use the diversity for Selective Collaborative Clustering