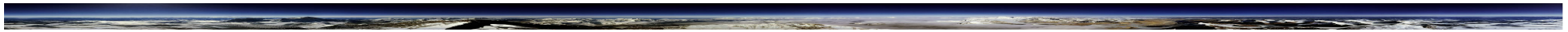
An Introduction to Collaborative (and incremental) Clustering

(1ère partie)



A. Cornuéjols

AgroParisTech – INRA MIA 518

Outline

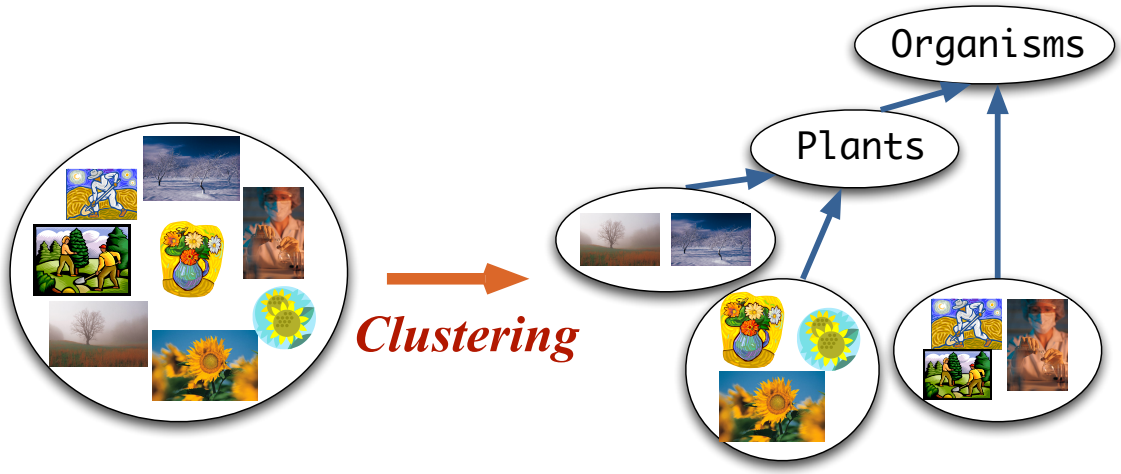
1. Clustering (a reminder)
2. New problems / new approaches: multi-clustering
3. Collaborative clustering
4. Simple examples to stimulate questions
5. How to choose the best collaborators
6. (Partial) conclusions

Clustering

A reminder

Motivation

- **Unsupervised** data
- Goal
  - Organize a set of objects into **contrasted groups contrastés**
  - **Compress** information by **discovering structures** in the data



Issues

- Which optimization criterion?
- How to chose:
  - The description space
  - Distances
  - The number of clusters
- Which optimization algorithm ?
- Validation ?

Questions: what do we want to optimize

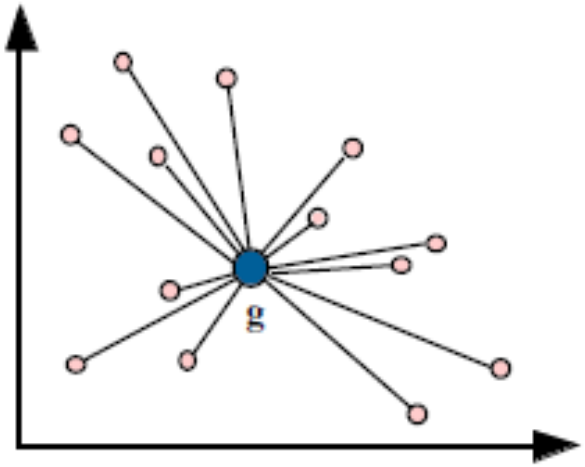
- Group objects into
  - Objects which are **similar** within groups : **intra**-similarity
  - Objects which are **dissimilar** in different groups : **inter**-dissimilarity

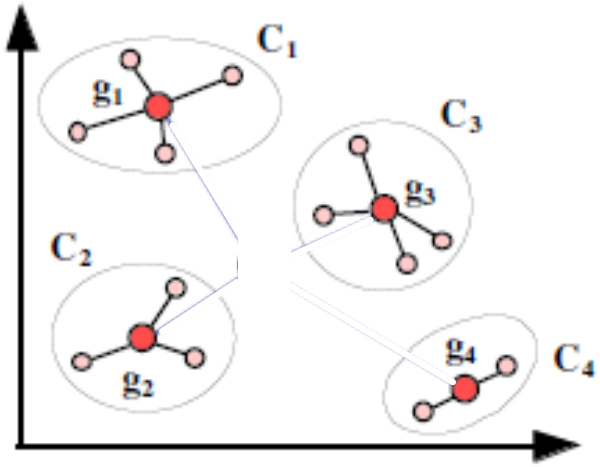
An optimization problem

Optimization criterion

- Formulation :

$$J = \sum_{g=1}^{K} \sum_{i \in C_g} d^2(\mathbf{x}_i, \mu_g)$$

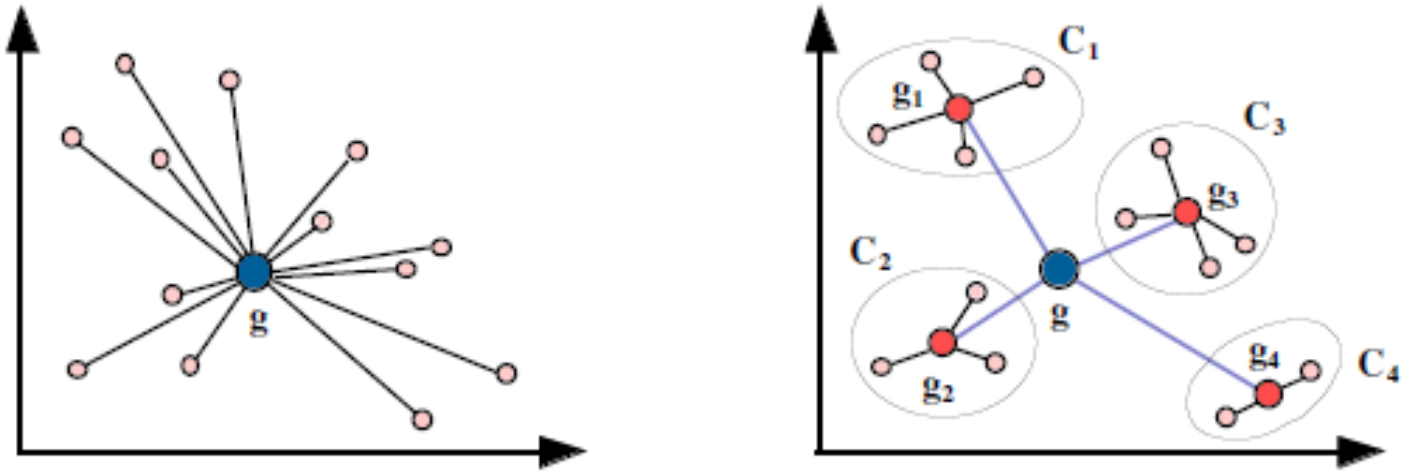




Optimization criterion

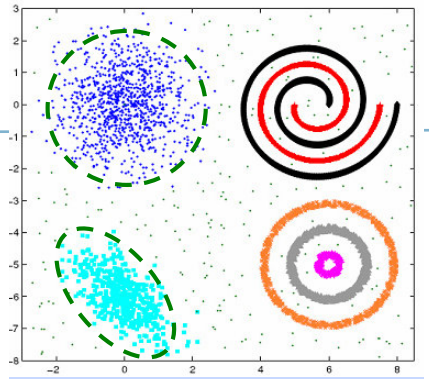
- Another formulation :

$$J = \underbrace{\sum_{g=1}^{K} \sum_{i \in C_g} d^2(\mathbf{x}_i, \mu_g)}_{intra-inertia} + \underbrace{\sum_{g=1}^{K} N_g \; d^2(\mu_g, \mu)}_{extra-inertia}$$

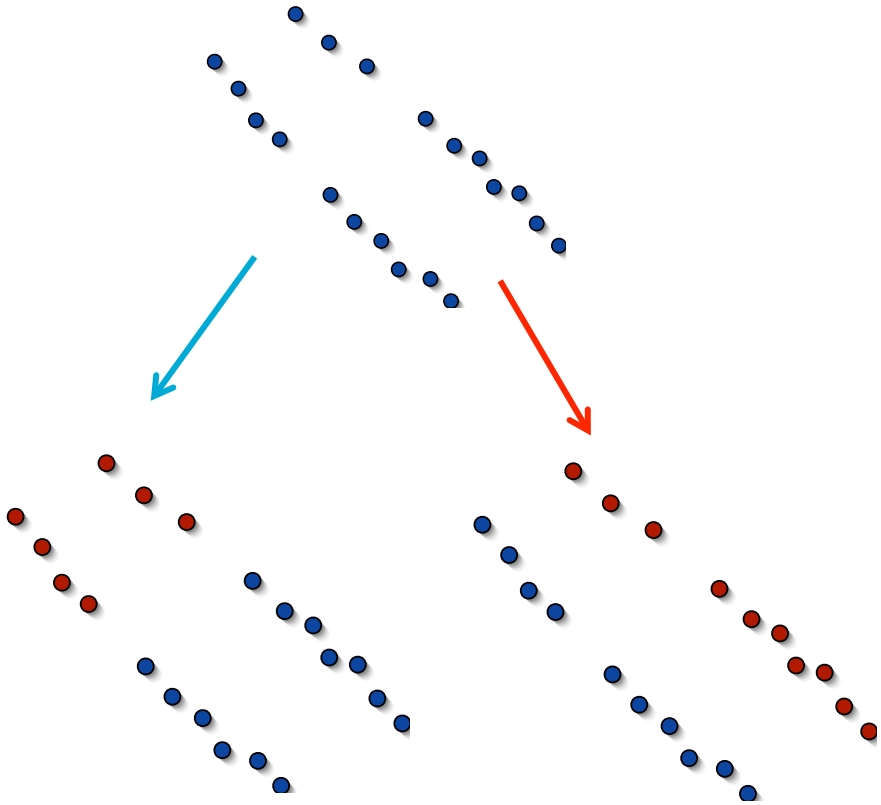


Inertie totale des points = Inertie Intra + Inter

Choice of the distances

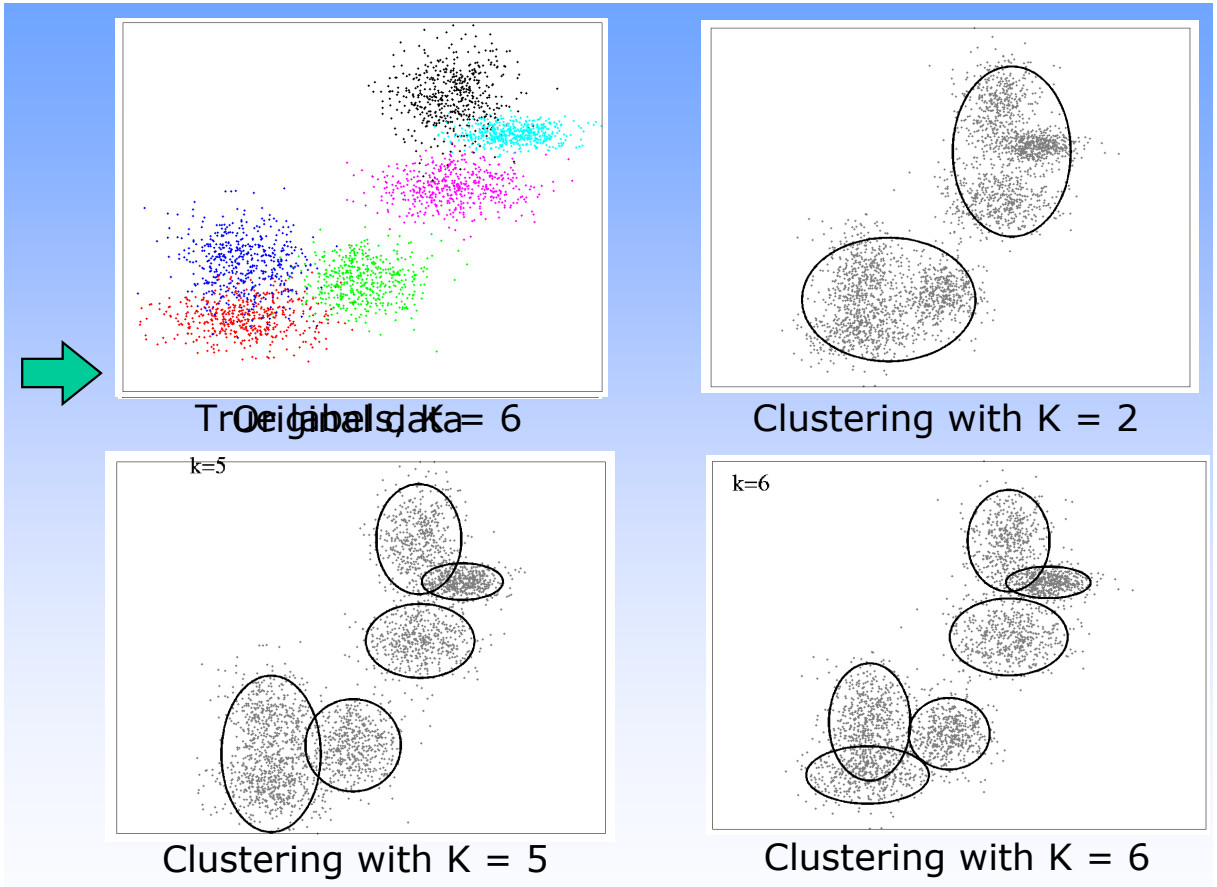


- Examples
  - Intra-distance
    << Inter-distance
  - Intra-connectivity
    << Inter-connectivity



Number of clusters

- How to choose the number?



Description space

- Normalizing dimensions ?

- High dimensionality: which consequences?

Algorithms

- How to find a clustering optimizing the criterion

→ Exploration of the space of configurations

Number of possible clustering

- Number of partitions of *N* objects in *K* groups

$$S_{N,K} = \frac{1}{K!}\sum_{k=0}^{K}(-1)^k\,(K-k)^N\binom{K}{k} \simeq \frac{K^N}{K!} \text{ as } N \to \infty$$

  - Number of partitions of *100* objects in *3* groups = 10^{47}
  - Number of partitions of *100* objects in *5* groups = 10^{68}

- Total number of possible partitions of *N* objects

$$B_N = \sum_{k=1}^{N} S_{N,k}$$

Number of partitions of *25* objects = $4{,}6 \cdot 10^{19}$

Algorithms

- Necessity of **heuristical approaches**

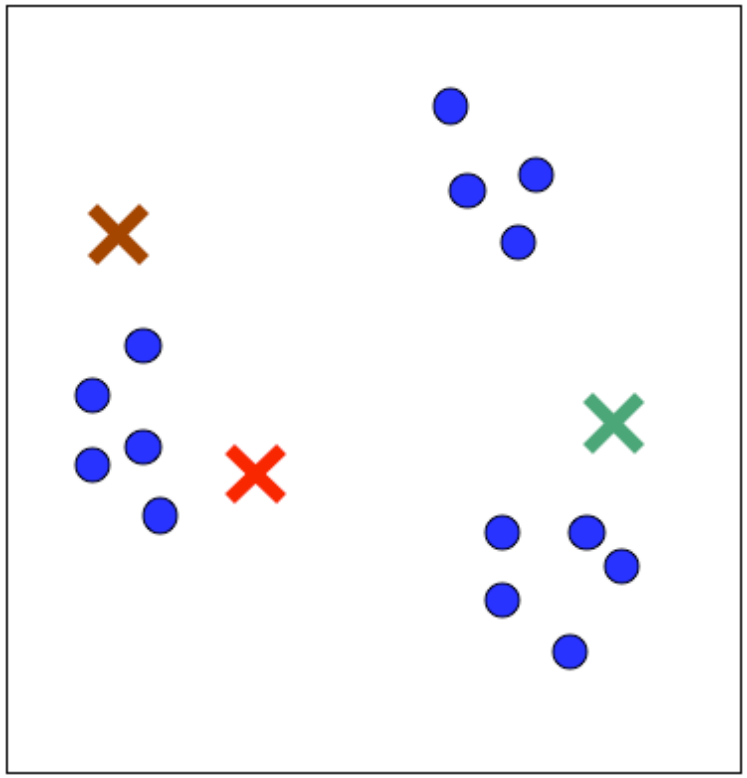
- Families of algorithms
  - Partitioning Methods (K-means, …)
  - Hierarchical Methods (AHC, …)
  - Density-based Methods (Dbscan, SOM, …)
  - Spectral Methods (Graph *Laplacian*)

The K-means algorithm

- Goal criterion: $$J = \sum_{g=1}^{K} \sum_{i \in C_g} d^2(\mathbf{x}_i, \mu_g)$$

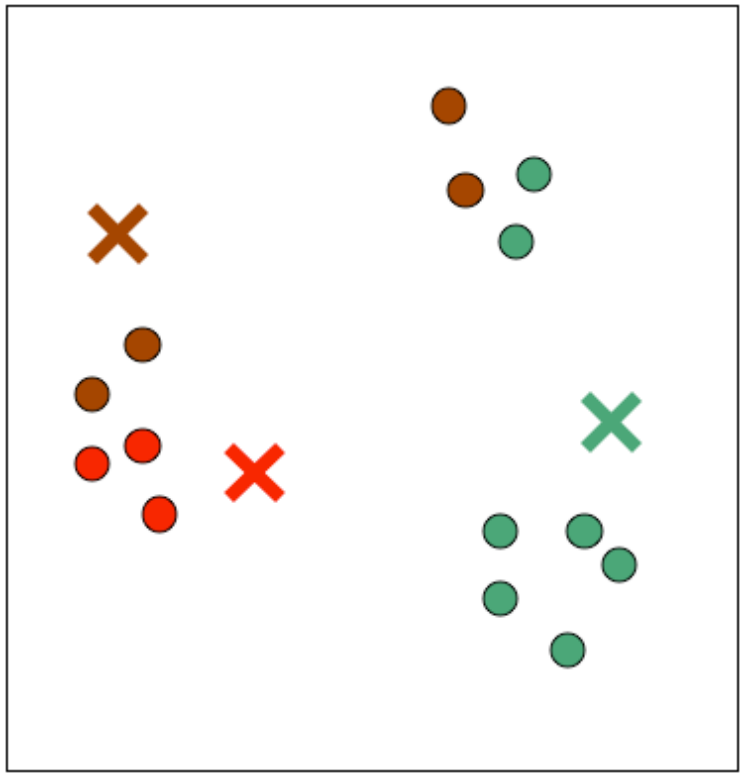
The K-means algorithm

1. Initialization: Choose
   K objects randomly



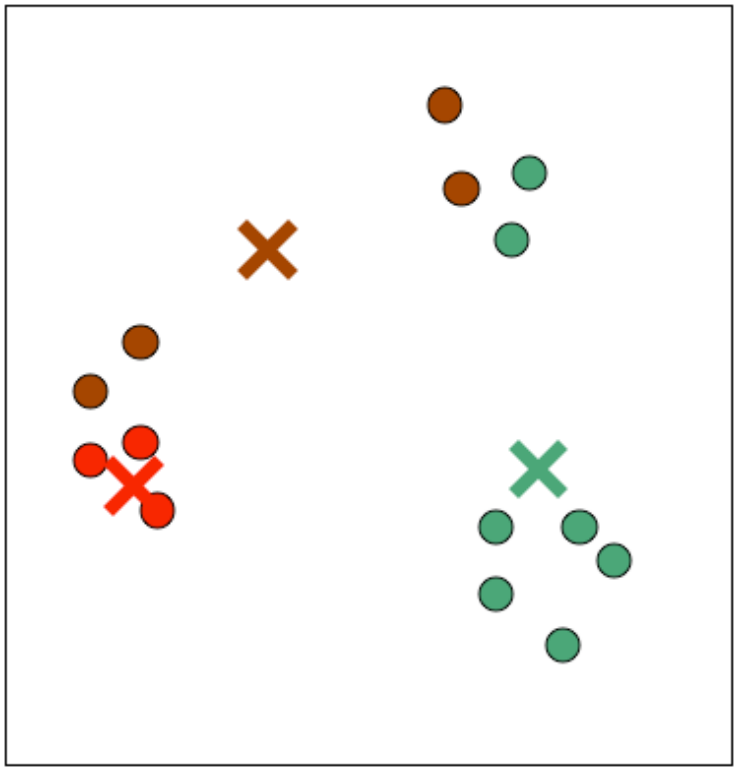
The K-means algorithm

1. **Initialization**: choose *K* objects randomly
2. Affect each object to the nearest centroïd (according to the chosen distance)



The K-means algorithm

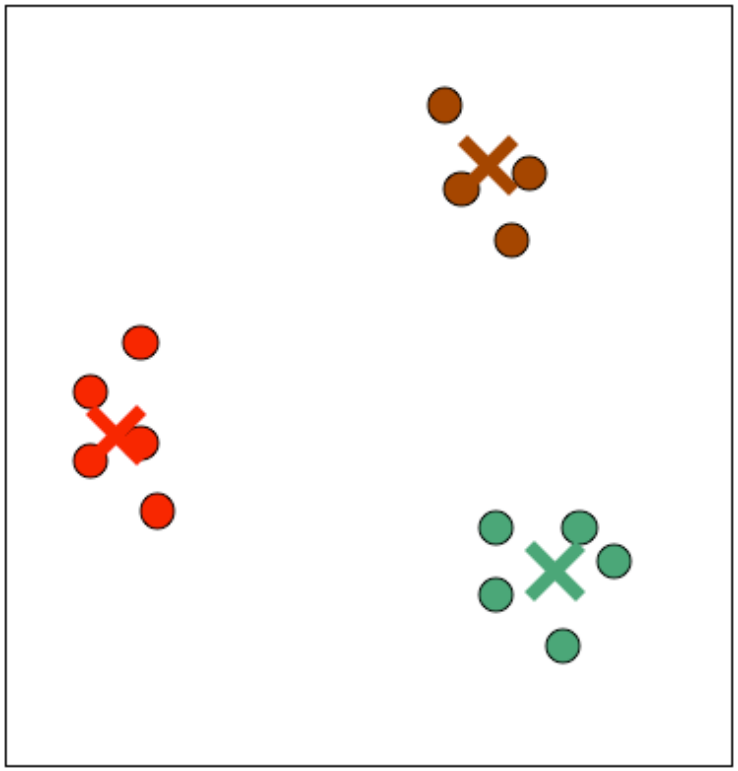
1. **Initialization**: choose *K* objects randomly
2. Affect each object to the nearest centroïd (according to the chosen distance)
3. Move the centroïds to the barycenter of the objects that have been affected to it



Iteration = 1

The K-means algorithm

1. **Initialization**: choose *K* objects randomly
2. **Affect** each object to the nearest centroïd (according to the chosen distance)
3. **Move** the centroïds to the barycenter of the objects that have been affected to it
4. **Repeat** until the changes fall below a given threshold



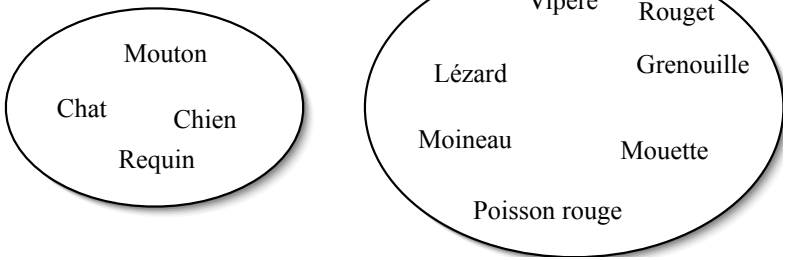
Iteration = 3

The K-means algorithm

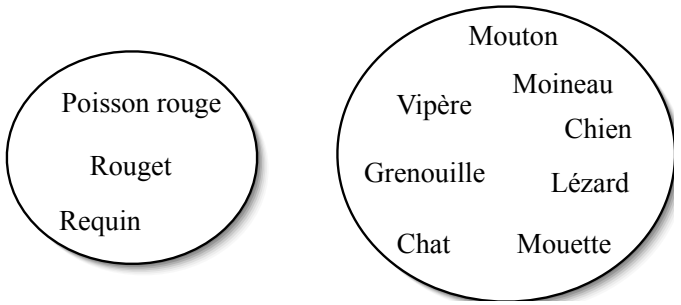
- Properties :
  - Converges (towards a local minimum)
  - Simple and efficient in ($O(K\,N\,I)$) (I itérations) and in space
    - Often <= 10 itérations
    - Applicable to large N
  - BUT:
    - Sensitive to the initialization
    - K must be specified
    - Sensitive to outliers
    - Not appropriate for discovering non convex clusters
    - Not directly applicable to categorial data

The question to the validation

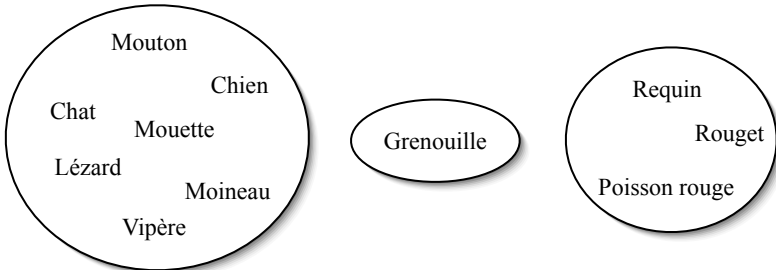
- There is no intrinsically perfectclustering



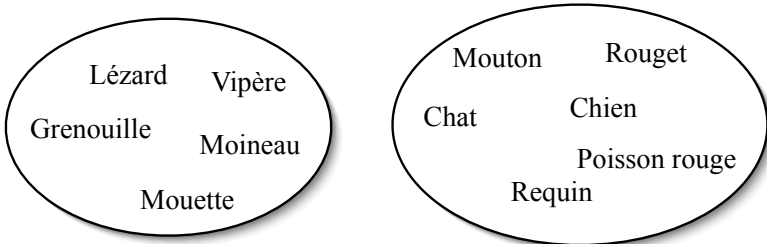
Types of bearing



Types of lungs



Types of environnement



Types of bearing and environnement

- Every criterion is **arbitrary**

The question to the validation

- **Impossible to « test »** the hypothetical structure over a test sample (unlike in supervised learning)

- Unfortunately:
  - Large impact of **numerous choices** and **parameters**
    - *Distance*
    - *Number of clusters*
    - *Description space*
    - *Algorithm*
    - *Initialization*

[Handl et al., 2005]

→ ***How to proceed ?***

The question to the validation: very directly

- Do the uncovered strutures
  - Correspond to « **the reality** »?
  - Or only **reflect** the chosen **clustering method**?

Validation: approaches

- Extrinsic: *information external to the data*
  - Expert opinions
  - True class of some data points
  - ➢ F-measure
  - ➢ Consistency measure (or purity measure)
    - *Rand Index, Adjusted Rand Index (ARI), …*

- Intrinsic: *measure on the discovered clusters*
  - Compacity
  - Connexity
  - Separation

Linear combination: *Validity index SD*

Non linear combination : *Davies-Bouldin index*, *Dunn-like index*, *Silhouette index*

Validation and stability criterion

- Assumption:
  - **If** stable results in spite of variations on the choice
  - **Then** true regularities

- Temptation to base clustering on the robustness XXX
  - Relative index: comparison of clusterings

- One of the foundations of the multi-clustering approaches

Multi-clustering

New approaches

- **Consensus Clustering**, Clustering Aggregation
  - [Gionis et al., 2007]

- Multiview Clustering, Non-redundant Clustering, **Alternative Clustering**
  - [Gondek et al., 2007 ; Cui et al. 2007 ; Dang et al. 2015]

- **Collaborative clustering**
  - [Bennani et al., in preparation]

Consensus clustering

- Motivation
  - Difficult to know which clustering method to use (they are all biased)
- Principle
  - Rely on various clustering methods
  - In the hope of escaping bad biases
  - But *what could be a sound justification?*
- **Orthogonality** measure
  - ➢ How to choose appropriate methods?
  - Complementary methods

(dis)-Agreement measure

(dis)-Agreement measure

- A **disagreement** = a pair of objects
  - That are assigned to the **same** cluster by the clustering C
  - And **not** by the other clustering C'

$$d(C, C') = \text{nb of disagreements between } C \text{ and } C'$$

[A. Gionis, H. Mannila & P. Tsaparas (2007): Clustering Aggregation.
ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 1, N° 1, 2007.]

(dis)-Agreement measure

	C_1	C_2	C_3	
x_1	1	a	α	
x_2	1	b	β	
x_3	2	a	α	
x_4	2	b	β	
x_5	3	c	χ	
x_6	3	d	χ	

pairs	C_1C_2	C_1C_3	C_2C_3
1,2			
1,3			
1,4			
1,5			
1,6			
2,3			
2,4			
2,5			
2,6			
3,4			
3,5			
3,6			
4,5			
4,6			
5,6			

(dis)-Agreement measure

	C_1	C_2	C_3	
x_1	1	a	α	
x_2	1	b	β	
x_3	2	a	α	
x_4	2	b	β	
x_5	3	c	χ	
x_6	3	d	χ	

pairs	C_1C_2	C_1C_3	C_2C_3
1,2	1	1	-
1,3	1	1	-
1,4	-	-	-
1,5	-	-	-
1,6	-	-	-
2,3	-	-	-
2,4	-	-	-
2,5	-	-	-
2,6	-	-	-
3,4	1	1	-
3,5	-	-	-
3,6	-	-	-
4,5	-	-	-
4,6	-	-	-
5,6	1	-	1

(dis)-Agreement measure

	C_1	C_2	C_3	Consensus
x_1	1	a	α	i
x_2	1	b	β	j
x_3	2	a	α	i
x_4	2	b	β	j
x_5	3	c	χ	k
x_6	3	d	χ	k

Consensus clustering minimizes the

total number of disagreements

with the base clusterings

pairs	C_1C_2	C_1C_3	C_2C_3
1,2	1	1	-
1,3	1	1	-
1,4	-	-	-
1,5	-	-	-
1,6	-	-	-
2,3	-	-	-
2,4	-	-	-
2,5	-	-	-
2,6	-	-	-
3,4	1	1	-
3,5	-	-	-
3,6	-	-	-
4,5	-	-	-
4,6	-	-	-
5,6	1	-	1

Alternative clustering

- Motivation
  - Specially in high dimension
  - Data can be grouped into **different yet meaningful** ways
- Principle
  - Favor **different clusterings** of the data set

➢ **Orthogonality measure** between clusterings

Measuring uncorrelation between clusterings

- Dot-product between pairwise centroïds of the clusterings [Jain et al. 2008]
  - Characterize the existing clustering by the space spanned by the existing centroïds
  - Search a **clustering in the orthogonal space** [Cui et al. 2007]

- Mutual information between the clusterings [Dang & Bailey, 2010]
  - Find a new clustering that **maximizes the likelihood of the data** while its **mutual information with existing clusterings is minimized** [Dang & Bailey, 2015]

$$\hat{\Theta} = \underset{\Theta}{\text{Argmax}}\, L(\mathcal{X}|\Theta)$$

$$= \underset{\Theta}{\text{Argmax}} \left\{ \sum_{n=1}^{m} \log \sum_{i=1}^{K} \alpha_i\, p(\mathbf{x}_n|\theta_i) - \gamma \sum_{s} I(C; C^{(s)}) \right\}$$

- ...

Collaborative Clustering

Motivations

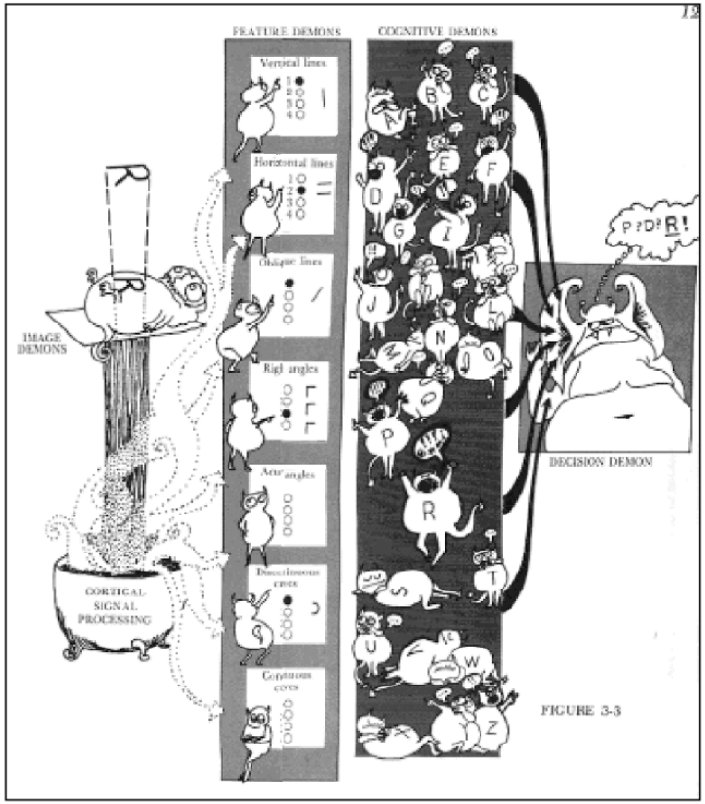
- Distributed computation, **local informations**: **multi-sources**
  - « **Vertical** » clustering
    - Examples **=** ; features **≠**
      - *No center has all information about the clients*
  - « **Horizontal** » clustering
    - Examples **≠** ; features **=**
      - *Branches of a bank: not the same clients*
  - « **Mixte = vertical + horizontal** » clustering
    - Examples **≠** ; features **≠**
      - *Different patients in various medical centers*

Motivations

- **Multi-expertises**
  - Image analysis
    - At different scales
    - With various sources of expertise
      - Analysis at different wavelength
      - Analysis of the shadows
      - Knowledge about the territoires
  - Speech recognition
    - The Hearsay II system

Pandemonium

- First **Pandemonium** (1958)
  - Oliver Selfridge « Pandemonium: A Paradigm for Learning »
  - A **hierarchical architecture** of « demons » to solve problems + a suggestion for a **learning mechanism**
    - « **Data** demons : specialized in some types of input data (horizontal line, circle subparts, …)
    - « **Cognitive** demons »: integrate information coming from sub-levels demons
    - « **Decision** demons » : make decisions about the interpretation
    - Demons shout with a strength in proportion to their certainty in their claim
    - This « strength » is set through **learning**



Hearsay II (1975)

- Speech recognition
  - The DARPA Speech Understanding Research (SUR) program

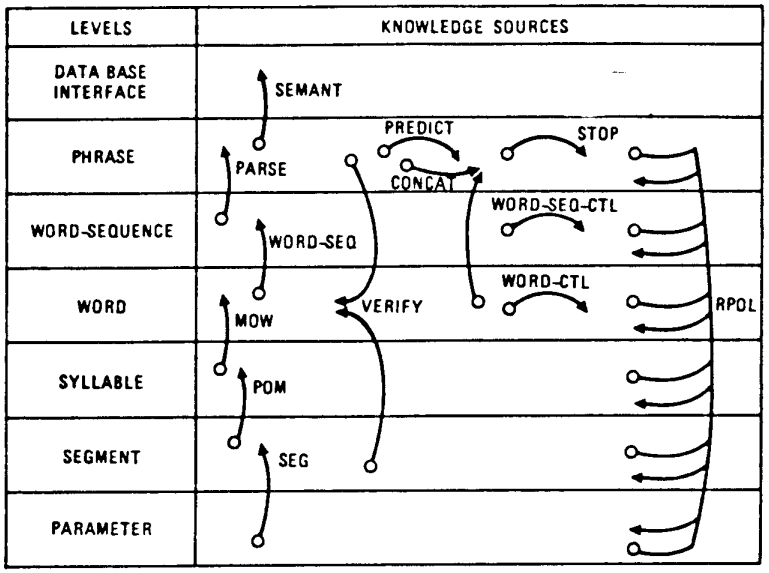


FIGURE 2. The levels and knowledge sources of September 1976. KSs are indicated by vertical arcs with the circled ends indicating the input level and the pointed ends indicating output level.

Signal Acquisition, Parameter Extraction, Segmentation, and Labeling:
- SEG: Digitizes the signal, measures parameters, and produces a labeled segmentation.

Word Spotting:
- POM: Creates syllable-class hypotheses from segments.
- MOW: Creates word hypotheses from syllable classes.
- WORD-CTL: Controls the number of word hypotheses that MOW creates.

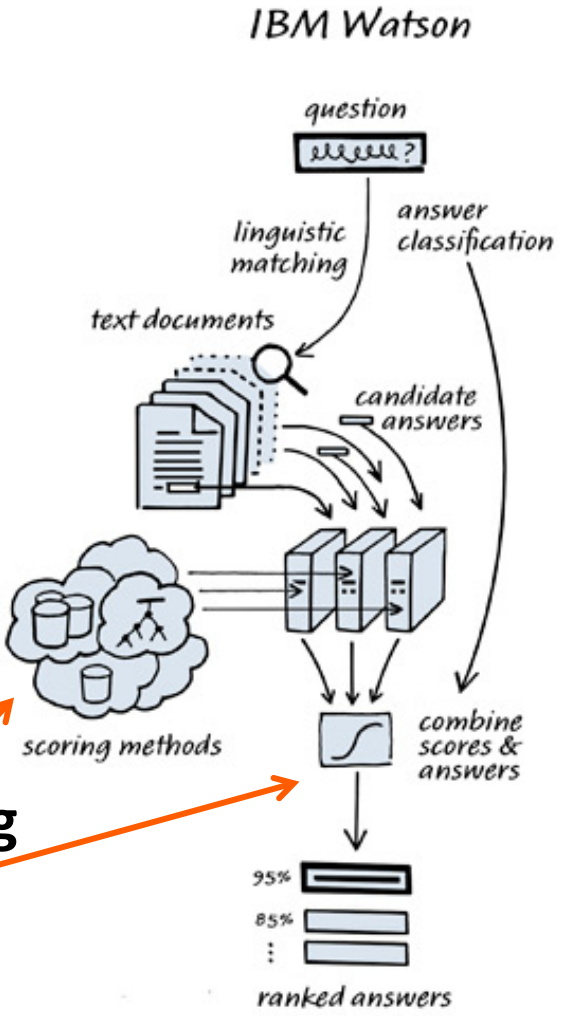
Phrase-Island Generation:
- WORD-SEQ: Creates word-sequence hypotheses that represent potential phrases from word hypotheses and weak grammatical knowledge.
- WORD-SEQ-CTL: Controls the number of hypotheses that WORD-SEQ creates.
- PARSE: Attempts to parse a word sequence and, if successful, creates a phrase hypothesis from it.

Phrase Extending:
- PREDICT: Predicts all possible words that might syntactically precede or follow a given phrase.
- VERIFY: Rates the consistency between segment hypotheses and a contiguous word-phrase pair.
- CONCAT: Creates a phrase hypothesis from a verified contiguous word-phrase pair.

Rating, Halting, and Interpretation:
- RPOL: Rates the credibility of each new or modified hypothesis, using information placed on the hypothesis by other KSs.
- STOP: Decides to halt processing (detects a complete sentence with a sufficiently high rating, or notes the system has exhausted its available resources) and selects the best phrase hypothesis or set of complementary phrase hypotheses as the output.
- SEMANT: Generates an unambiguous interpretation for the information-retrieval system which the user has queried.

More recently: WATSON

- Search the best answer to open questions
- Require a the exploration of a huge serach space
  - Documents
  - Internet



But those are **supervised learning** systems

Collaborative clustering

Algorithm 3: Algorithm for *collaborative clustering*

Data: N subsets $\{\mathcal{S}_{i\ (1\leq i\leq N)}\}$ of $\mathcal{S}$ not necessarily disjoint
N algorithms $\{A_{i\ (1\leq i\leq N)}\}$

Result: N clusterings $\mathcal{C}_i$, one for each data set

repeat

Run in parallel the N algorithms A_i each on its own data set $\mathcal{S}_i$;

Compute the resulting N clusterings $\mathcal{C}_i$;

Exchange information between the/some algorithms ;

until *stabilization of the local clusterings* ;

Collaborative learning: issues

- Information **exchanges**
  - What types of information?
  - Which weight?

- **How many** information exchanges: the control
  - Stopping criterion?

- **With whom** to exchange?
  - Which collaborations are the most favorable?

Illustrations

Algorithms

- Supposedly all ***K*-means**
  - Values for *K* possibly different
  - Choice of distances possibly different
  - Choices of initializations possibly different

- Can **exchange?**
  - …
  - …
  - …
  - …

Algorithms

- Supposedly all ***K*-means**
  - Values for *K* possibly different
  - Choice of distances possibly different
  - Choices of initializations possibly different

- Can **exchange**
  - The number of clusters they envision
  - The coordinates of the centroïds of the clusters
  - The proportion of examples assigned to each cluster
  - The identifiers of the examples assigned to each cluster

Scenarios

1. The various algorithms have access to the **same data**

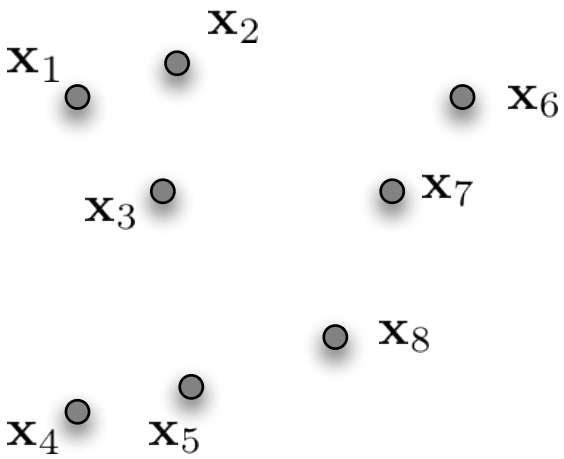
2. **Same examples**, but **different descriptors** (vertical)

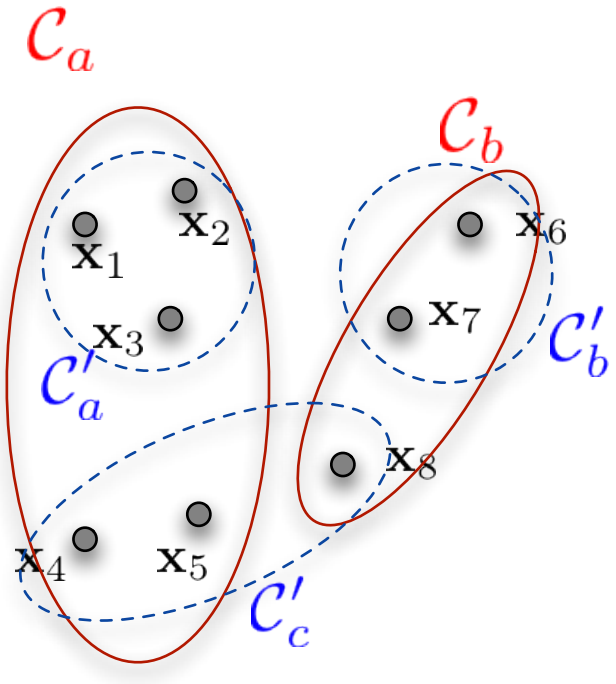
3. **Different examples**, but **same descriptors** (horizontal)

4. **Different examples** and **different descriptors**

Scenario 1: The algorithms have access to the **same data**

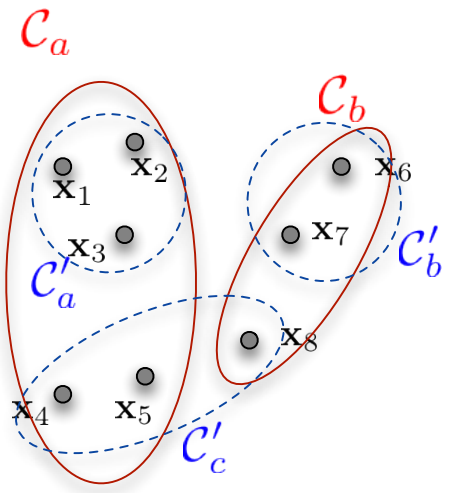
- Illustration





Scenario 1: The algorithms have access to the **same data**

- Can exchange:
  - Assignements of the examples to the clusters
    - Thanks to the identifiers
  - The centroïdes

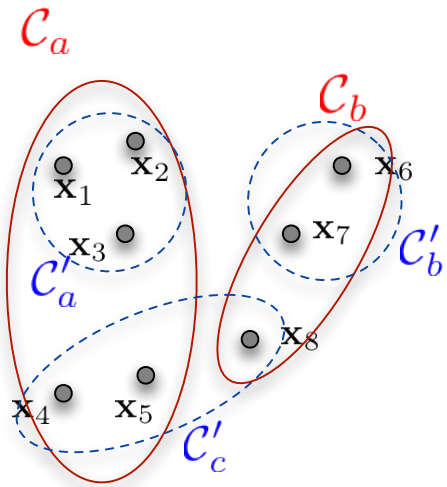


- Conflict detection:
  - How?

Algorithm	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
A	C_a	C_a	C_a	C_a	C_a	C_b	C_b	C_b
B	C'_a	C'_a	C'_a	C'_c	C'_c	C'_b	C'_b	C'_c

Scenario 1: The algorithms have access to the **same data**

Algorithm	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
A	C_a	C_a	C_a	C_a	C_a	C_b	C_b	C_b
B	C'_a	C'_a	C'_a	C'_c	C'_c	C'_b	C'_b	C'_c



- Possible solution:

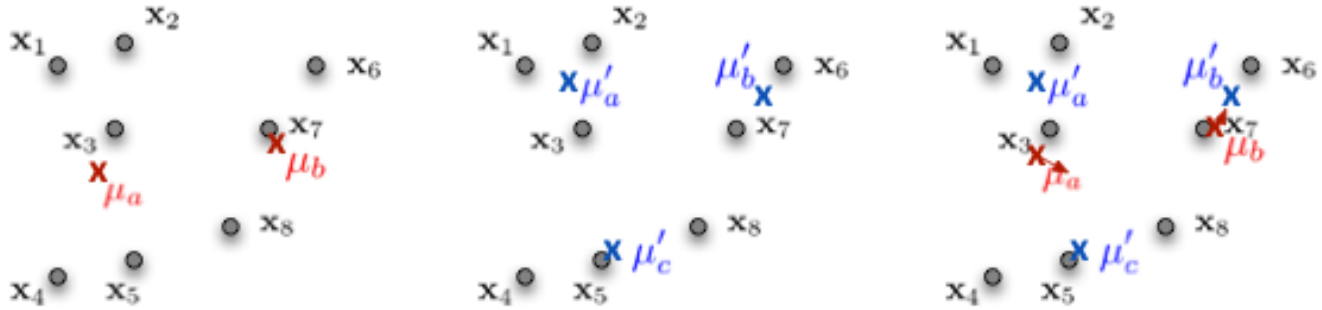


Figure 2: (Left) The prototypes computed by algorithm A at time step t. (Center) The prototypes computed by algorithm B. (Right) The updates of the prototypes of algorithm A when taking into account the prototypes communicated by algorithm B.

Scenario 1: The algorithms have access to the **same data**

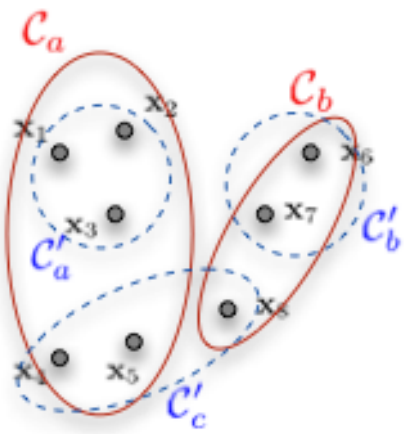
- Which « bandwith »: cost of communication?
- Optimal exchanges?
- If more than 2 agents?
- Convergence?

→ Which algorithm?

Scenario 2: **Same examples**, but with **different descriptors**

Scenario 2: **Same examples**, but with **different descriptors**

- Can exchange:
  - The **identifiers** of the examples
  - **Not** the coordinates of the centroïds



	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1		1	1	½	½	0	0	0
X_2	1	-	1	½	½	0	0	0
X_3	1	1	-	½	½	0	0	0
X_4	½	½	½	-	.	0	0	½
X_5	½	½	½	.	-	0	0	½
X_6	0	0	0	0	0	-	'	½
X_7	0	0	0	0	0	'	-	½
X_8	0	0	0	½	½	½	½	

Figure 3: (Left) The data set and the clusterings by algorithms A and B. (Right) The corresponding consensus matrix. The shaded parts denotes the objects for which there is an agreement between the clusterings.

Scenario 3 : **Different examples**, but **same descriptors**

Scenario 3 : **Different examples**, but **same descriptors**

- One must assumes that *the local data sets have been generated by the same underlying distributions*

- One can again compute a consensus matrix
- It is possible to exchange centroïds
- **Same** type of **algorithm than in scenario 1**

Scenario 4: **Different examples** and **different descriptors**

- Exchange of the centroïds is still possible if there are some common descriptors
- And else … ?

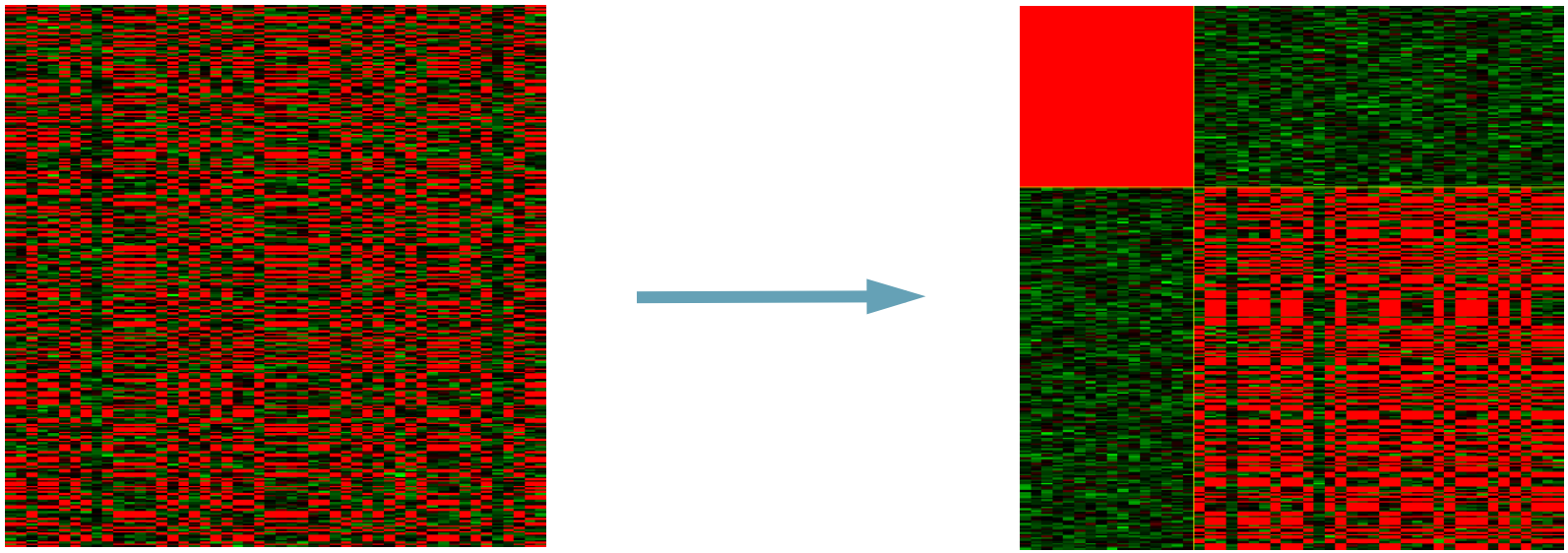
Bi-clustering

An elementary case of collaborative clustering

Illustration

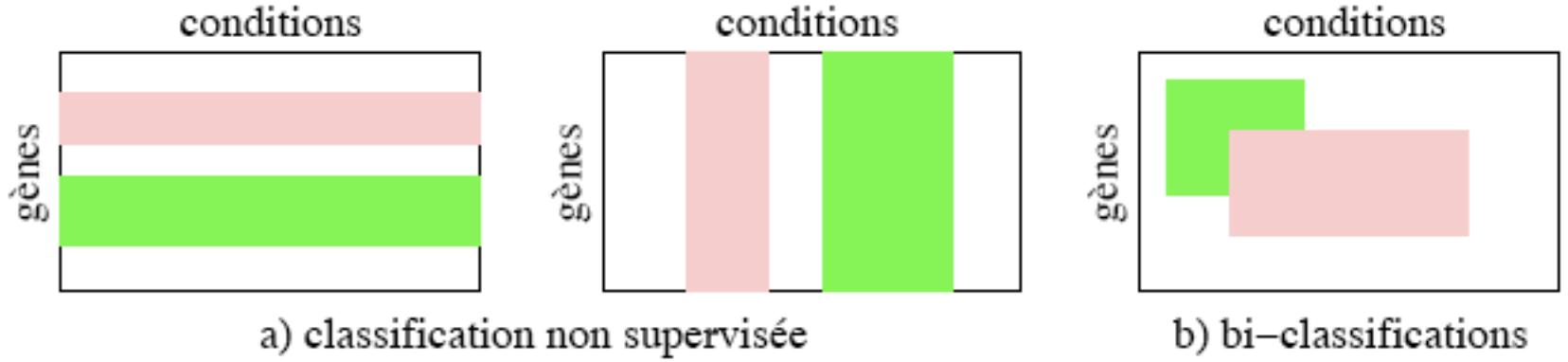
- Matrices
  - X = {Gènes}
  - Y = {Conditions}

Some genes can be co-expressed under some conditions
and behave almost independently in other conditions



How?

- Clustering on each dimension independently



How?

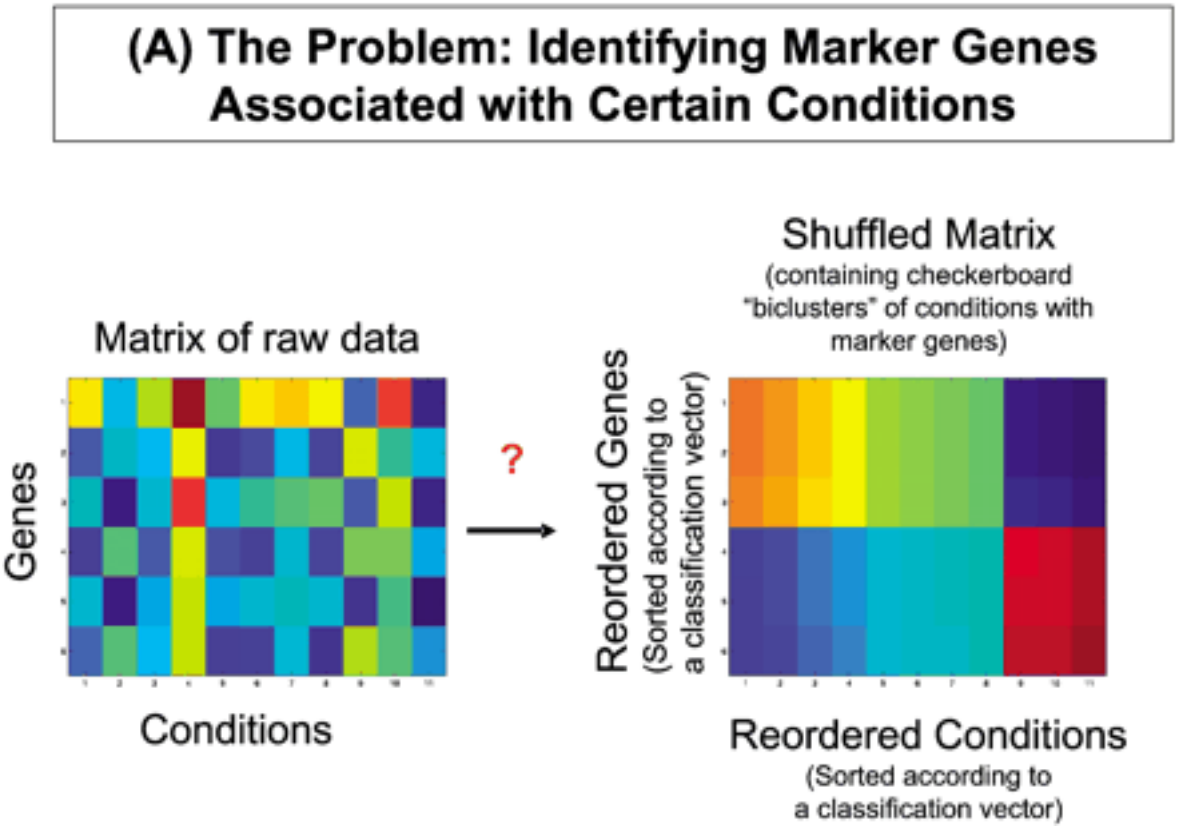
- Naïve approach
  - Through successives permutations on X_1 and on X_2
  - Until?
    - **A criterion quality** is maximized?

Number of permutations

- $|X_1|! \times |X_2|!$
- E.g. : $100! \times 100! \approx 8 \cdot 10^{315}$

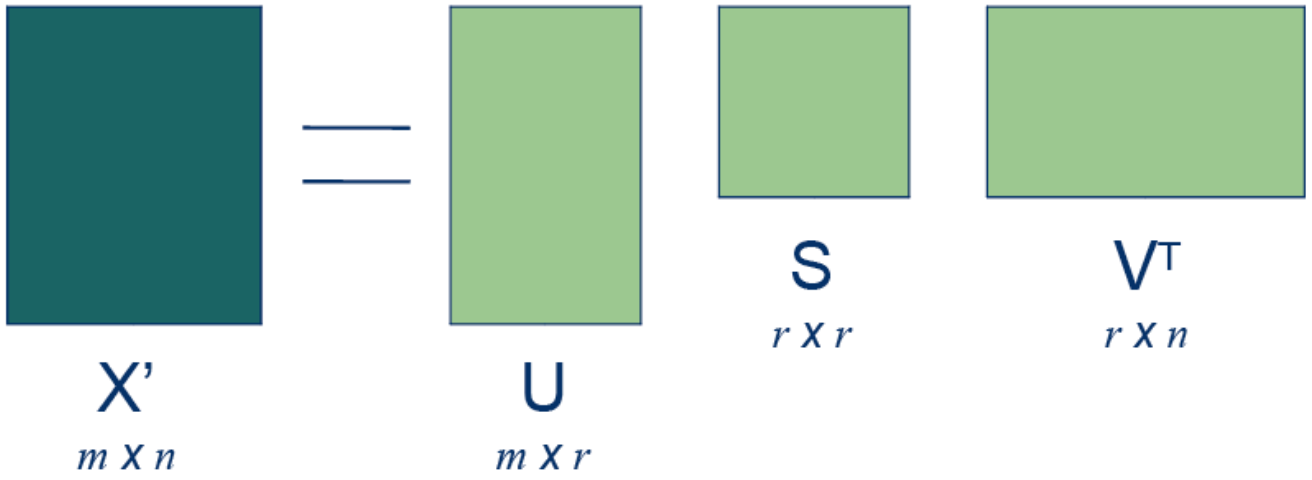
→ Heuristical approach!

How?


(A) The Problem: Identifying Marker Genes Associated with Certain Conditions
Matrix of raw data
Genes
Conditions
?
Shuffled Matrix
(containing checkerboard "biclusters" of conditions with marker genes)
Reordered Genes
(Sorted according to a classification vector)
Reordered Conditions
(Sorted according to a classification vector)

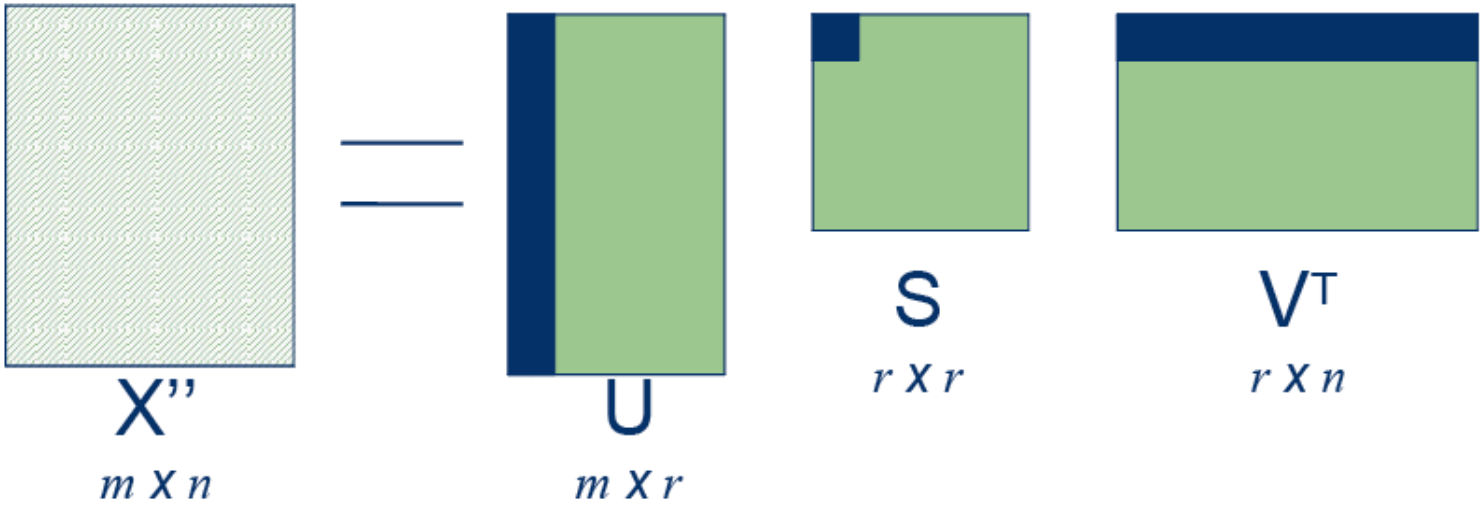
Singular Value Decomposition (SVD)

- Any matrix **X** of rank *r* can be decomposed in $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^\top$
  - Where **U** and **V** are orthogonal matrices $\mathbf{U}\,\mathbf{U}^\top = \mathbf{V}\,\mathbf{V}^\top = \mathbf{I}$
  - And **S** is a diagonal matrix composed of singular vlaues of **X**



LSA = SVD + selection of singular values

- If the lowest singular values of **S** are set to zero
  - One gets an approximation of **X**



Bi-clustering …

- Bi-clustering = LSA + sparsity constraints on **U** and **V**
  - Few elements ≠ 0
  - One wants few clusters along each dimension

Sparse Singular Value Decomposition approach (SSVD)

- The basis of numerous systems
- SSVD: the matrix of the data **X** is supposed to be approximable through a matrix of lower rank **U**

$$\underbrace{\mathbf{X}}_{\mathbb{R}^{n.p}} = \underbrace{\mathbf{U}}_{\mathbb{R}^{n.r}} \underbrace{\mathbf{D}}_{\mathbb{R}^{r.r}} \underbrace{\mathbf{V}^\top}_{\mathbb{R}^{r.p}} = \sum_{k=1}^{r} \mathbf{u}_k \, s_k \, \mathbf{v}_k^\top$$

Only the *K* highest values of **D** are kept

$$\underbrace{\mathbf{X}}_{\mathbb{R}^{n.p}} \approx \mathbf{X}^{(K)} = \underbrace{\mathbf{U}}_{\mathbb{R}^{n.K}} \underbrace{\mathbf{D}}_{\mathbb{R}^{K.K}} \underbrace{\mathbf{V}^\top}_{\mathbb{R}^{K.p}} = \sum_{k=1}^{K} \mathbf{u}_k \, s_k \, \mathbf{v}_k^\top$$

$\mathbf{X}^{(K)}$ minimizes the Froebenius norm:

$$\mathbf{X}^{(K)} = \underset{\mathbf{X}^\star \in \mathcal{A}_K}{\text{Argmin}} \, ||\mathbf{X} - \mathbf{X}^\star||_F^2 = \underset{\mathbf{X}^\star \in \mathcal{A}_K}{\text{Argmin}} \, \text{tr}\left\{(\mathbf{X} - \mathbf{X}^\star)(\mathbf{X} - \mathbf{X}^\star)^\top\right\}$$

Sparse Singular Value Decomposition approach (SSVD)

- We want to optimize (1) :

$$||\mathbf{X} - \mathbf{u}_k\, s_k\, \mathbf{v}_k^\top||_F^2 + \lambda_u\, P_1(\mathbf{s}, \mathbf{u}) + \lambda_v\, P_2(\mathbf{s}, \mathbf{v})$$

- With a fixed **u**, minimize (1) with respect to **v** :

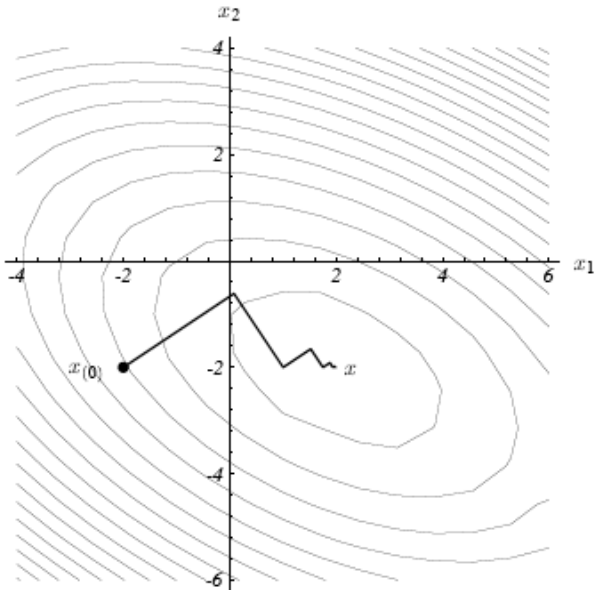
- Idem with a fixed **v**, minimize (1) with respect to **u** :

→ Iterative algorithm using **conjugated gradient**

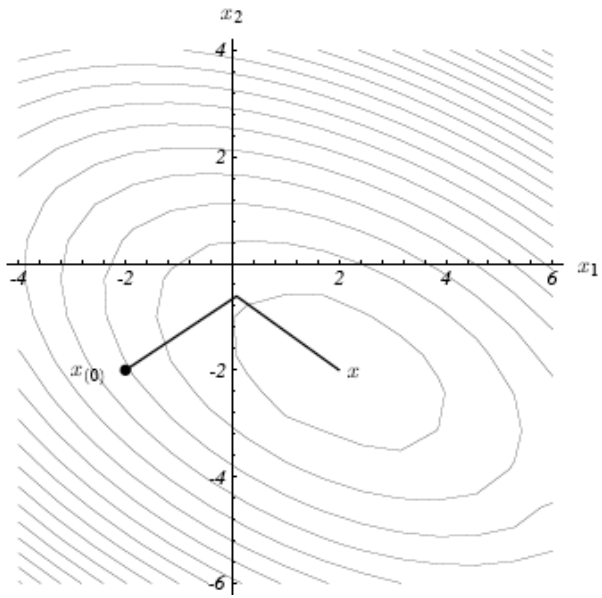
[Milhee Lee, Haiping Shen, et al. « Biclustering via sparse singular value decomposition », Biometrics, 66, 1087-1095, (2010)]

Conjugated gradient

- Follow each direction once with the appropriate step length



Simple gradient
(steepest gradient)



Conjugated gradient

Conjugated gradient: when does it work?

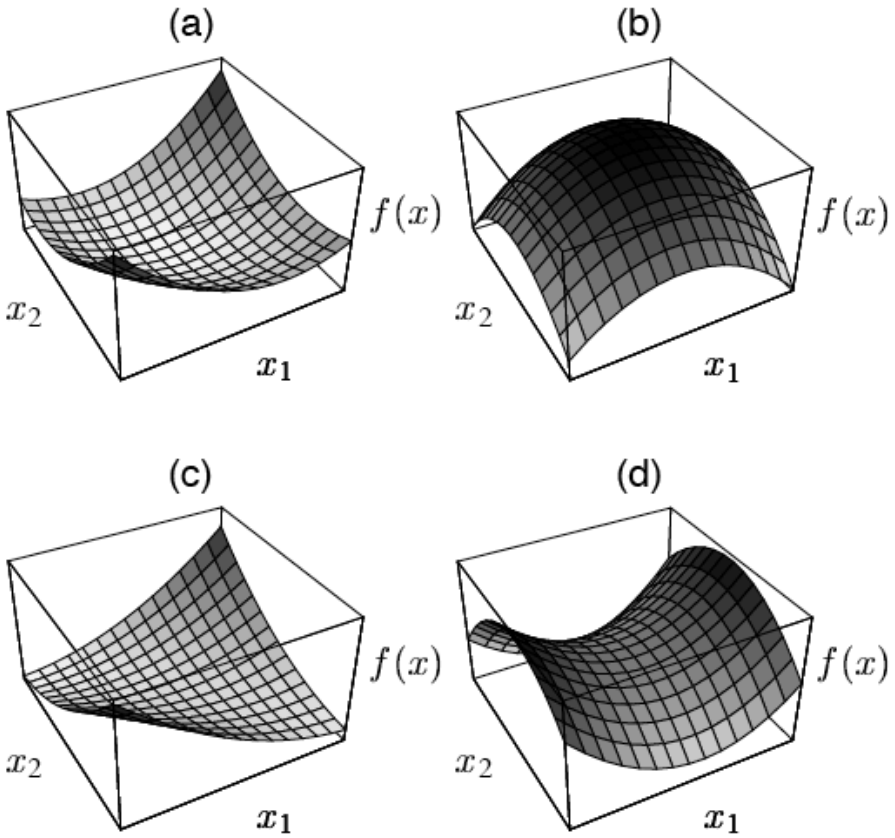


Figure 5: (a) Quadratic form for a positive-definite matrix. (b) For a negative-definite matrix. (c) For a singular (and positive-indefinite) matrix. A line that runs through the bottom of the valley is the set of solutions. (d) For an indefinite matrix. Because the solution is a saddle point, Steepest Descent and CG will not work. In three dimensions or higher, a singular matrix can also have a saddle.

Open question

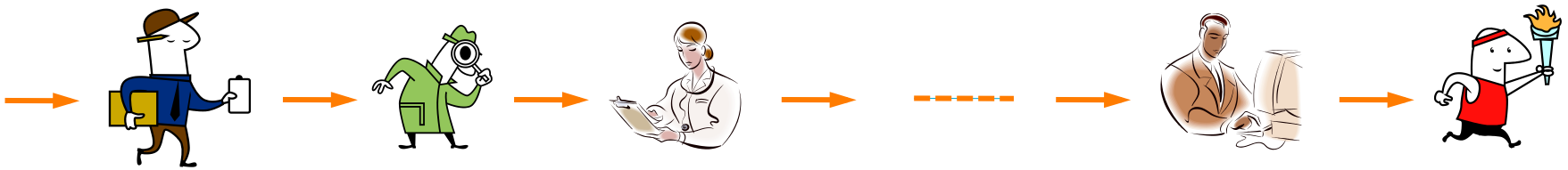
- Can we interpret conjugated gradient like a method for information exchange between dimensions?
  - Optimal transfer rate
  - Orthogonality condition between methods

A general framework?

Extension to the incremental case ?

Incrementality and collaboration

- Incrementality as …
  - a situation of **uni-directional** (vertical) collaboration between agents



- Transmission of information
  - Which information?
  - Which weight?
  - Also function of the changes of the environnement

Choice of the collaborations

An illustrative case

*Unsupervised Identification of **one class of interest***

by a collaborative method

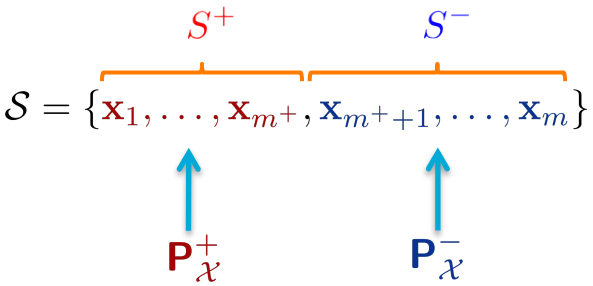
[Cornuéjols & Martin, AAFD-2014, SFC-2014]

Discovery of a class of interest

- Applications
  - Identification of « **anomalous** » behaviors (e.g. *persons guilty of fraud*)
  - **Genes** activated in one environmental condition
  - **Proteins** that interact with a molecule

The problem

- An unsupervised ***data sample*** $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$
- One assumes $\mathcal{S} = \mathcal{S}^+ \cup \mathcal{S}^-$



- One assumes $\mathcal{S}^- \underset{\text{i.i.d.}}{\sim} \mathbf{P}^-_{\mathcal{X}}$ and $\mathcal{S}^+ \underset{\text{i.i.d.}}{\sim} \mathbf{P}^+_{\mathcal{X}}$
- A set of ***score functions*** $\mathcal{F} = \{f_i\}_{1 \leq i \leq N}$

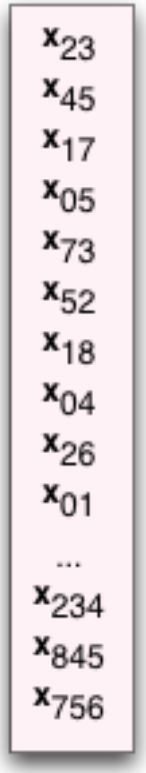
  with $f_i : \mathcal{X} \to \mathbb{R}$
- Goal: **identify** S^+

Score functions

- Function

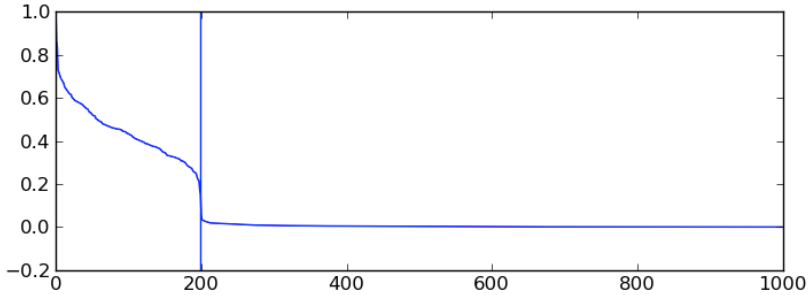
$$f_i : \mathcal{X} \rightarrow \mathbb{R}$$

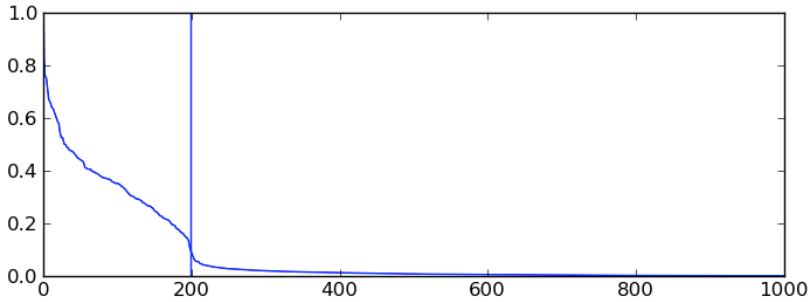
- That can be used to rank « objects »



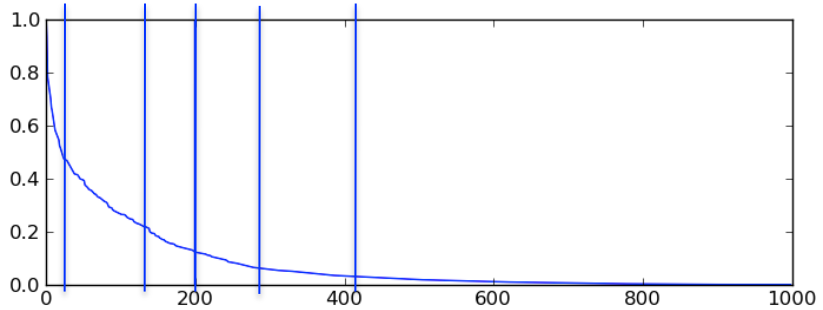
« Filter » methods: which score function is appropriate?

- Ranking of the examples





Threshold? →



? ? ? ? ?

An **unsupervised** ensemble method

- A **set** $\mathcal{F}$ of score functions

- Measure the **correlation of the ranks**
  - Over S
  - Over random samples

- Keep functions that are
  - Over-correlated over S
    - Agree on S
    - And not (or almost not) in general

Selection algorithm

Algorithme 1: Sélection de fonctions de base pertinentes

Entrées : La base d'exemples $\mathcal{S}$
L'ensemble $\mathcal{F}$ de fonctions d'évaluation de base
Sorties : Un sous-ensemble $\mathcal{F}'' \in \mathcal{F}$ de fonctions de base

Génération de N échantillons "aléatoires" $\mathcal{S}_0$;

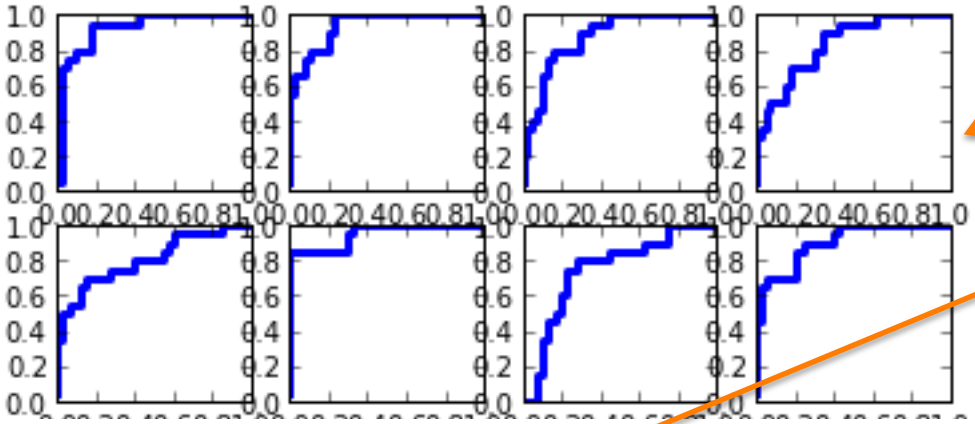
pour tous les *couples de fonctions d'évaluation* $(f_i, f_j)_{(i \neq j)} \in \mathcal{F}$ **faire**
  Calculer la surcorrélation de (f_i, f_j) sur $\mathcal{S}$ par rapport à la corrélation moyenne sur les échantillons $\mathcal{S}_0$
fin pour tous

Sélectionner les fonctions d'évaluation $f_i \in \mathcal{F}$ de surcorrélation $\geq$ seuil_min_surcor : soit $\mathcal{F}'$

Initialisation : $\mathcal{F}'' = \emptyset$

pour tous les $f_i \in \mathcal{F}'$ **faire**
  si $\sum_{j \neq i} surcorr(f_i, f_j) \geq seuil$ **alors**
    Mettre f_i dans $\mathcal{F}''$
fin pour tous

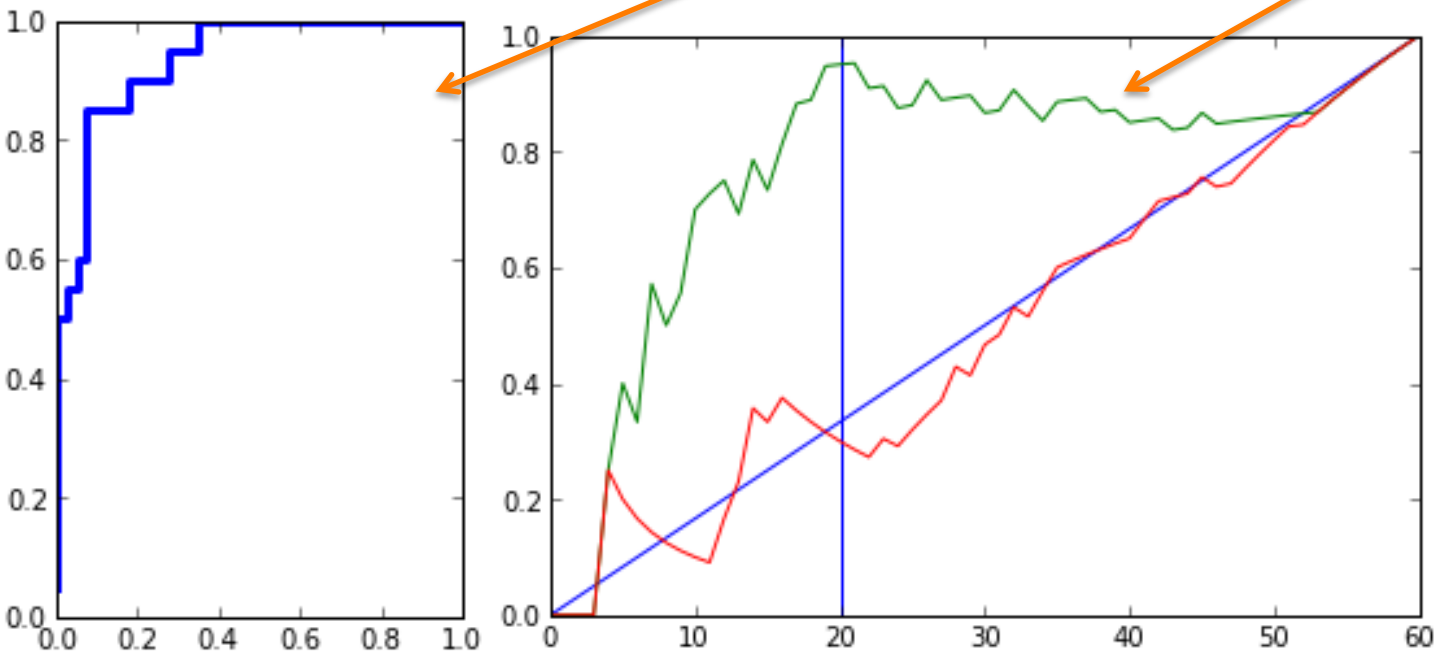
Illustration



ROC curves of the selected functions

ROC curve of the combination

Correlation the two first functions



Experimental protocol

- **320 examples**
  - **40** (soit 1/8) ; **80** (soit 1/4) ; **120** (soit 1/3) d'exemples `+'
  - in $\mathbb{R}^{20}$
- 2 Gaussians : $\mathbf{P}^{+}_{\mathcal{X}}$et $\mathbf{P}^{-}_{\mathcal{X}}$
  - $|\mu_+ - \mu_-|_2 = 3$
  - $\sigma = 1.5$ or 2.5 or 3.5 or 4.5 (noise rate)
- **45 score functions**
  - 22 positively aligned
  - 22 négatively aligned
  - 1 random function

Experimental results

σ	$\frac{m^+}{m}$	*Before selection*		*After selection*			
		auc_m	auc^M	auc_m	auc^M	$\overline{auc}$	AUC comb
1.5	$\frac{40}{320}$	0 ± 0	1 ± 0	$\mathbf{0.92} \pm 0.03$	1 ± 0	0.98 ± 0.01	$\mathbf{1} \pm 0$
	$\frac{80}{320}$	0 ± 0	1 ± 0	$\mathbf{0.87} \pm 0.06$	1 ± 0	0.97 ± 0.01	$\mathbf{1} \pm 0$
	$\frac{120}{320}$	0 ± 0	1 ± 0	$\mathbf{0.84} \pm 0.07$	1 ± 0	0.95 ± 0.01	$\mathbf{1} \pm 0$
2.5	$\frac{40}{320}$	0.02 ± 0.01	0.98 ± 0.01	$\mathbf{0.94} \pm 0.03$	0.98 ± 0.00	0.96 ± 0.02	$\mathbf{0.98} \pm 0.01$
	$\frac{80}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.85} \pm 0.05$	0.98 ± 0.01	0.91 ± 0.02	$\mathbf{0.97} \pm 0.01$
	$\frac{120}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.76} \pm 0.03$	0.98 ± 0.01	0.88 ± 0.02	$\mathbf{0.97} \pm 0.01$
	$\frac{160}{320}$	0.03 ± 0.01	0.98 ± 0.01	$\mathbf{0.73} \pm 0.04$	0.97 ± 0.01	0.85 ± 0.02	$\mathbf{0.95} \pm 0.01$
3.5	$\frac{40}{320}$	0.09 ± 0.02	0.91 ± 0.02	$\mathbf{0.75} \pm 0.06$	0.90 ± 0.03	0.83 ± 0.01	$\mathbf{0.90} \pm 0.03$
	$\frac{80}{320}$	0.09 ± 0.02	0.92 ± 0.02	$\mathbf{0.65} \pm 0.05$	0.92 ± 0.02	0.79 ± 0.02	$\mathbf{0.90} \pm 0.02$
	$\frac{120}{320}$	0.09 ± 0.02	0.91 ± 0.01	$\mathbf{0.64} \pm 0.04$	0.91 ± 0.01	0.77 ± 0.02	$\mathbf{0.89} \pm 0.02$
	$\frac{160}{320}$	0.10 ± 0.01	0.91 ± 0.02	$\mathbf{0.63} \pm 0.03$	0.91 ± 0.02	0.76 ± 0.02	$\mathbf{0.88} \pm 0.02$
4.5	$\frac{40}{320}$	0.13 ± 0.02	0.86 ± 0.02	$\mathbf{0.67} \pm 0.03$	0.86 ± 0.02	0.76 ± 0.02	$\mathbf{0.86} \pm 0.02$
	$\frac{80}{320}$	0.15 ± 0.02	0.85 ± 0.02	$\mathbf{0.65} \pm 0.03$	0.84 ± 0.03	0.75 ± 0.02	$\mathbf{0.84} \pm 0.03$
	$\frac{120}{320}$	0.15 ± 0.02	0.84 ± 0.02	$\mathbf{0.62} \pm 0.06$	0.84 ± 0.02	0.73 ± 0.03	$\mathbf{0.84} \pm 0.02$
	$\frac{160}{320}$	0.15 ± 0.01	0.85 ± 0.01	$\mathbf{0.61} \pm 0.03$	0.85 ± 0.01	0.72 ± 0.02	$\mathbf{0.83} \pm 0.03$

Table 1: Experimental results in function of the noise parameter σ and the proportion of the class '+'.

Conclusions

- **Unsupervised** learning (2 classes)
  - New « clustering » criterion : **sur-corrélation** of the rankings (score functions)
  - 1st **ensemble method** which **does not assume** that the « experts » are good

- **Theoretical** study
  - One can improve precision and recall as much as we want (can)

- **Empirical** study
  - Good results that confirm the soundness of the proposed approach
    - Better results than competition when the noise rate is high

2nd study

Evaluation of the potential contribution of collaborators

*by measuring **quality** and **diversity***

[Sublime et al., en préparation]

Conclusions

Conclusions (1)

- Clustering is useful
- Complex tasks
  - involve various fields of expertise
  - are often naturally distributed

  ➢ **Collaborative clustering**

- Different from
  - Consensus clustering
  - Alternative clustering

Conclusions (2)

- **Collaborative** clustering
  - A new research question
  - Which collaborators?
  - How to control the exchanges?

- Essentially
  - Heuristical approaches
  - Experimental results
  - Lack a theoretical ground

References

- Y. Bennani, A. Cornuéjols, P. Gançarski & C. Wemmert (en préparation): Collaborative Clustering: Why, When, What and How.
- X. H. Dang & J. Balley (2015): A framework to uncover multiple alternative clusterings. *Machine Learning J*. (2015), 98, 7-30.
- A. Gionis, H. Mannila & P. Tsaparas (2007): Clustering Aggregation. *ACM Transactions on Knowledge Discovery from Data* (TKDD), Vol. 1, N° 1, 2007.
- D. Gondek & Th. Hofmann (2007): Non-redundant data clustering. *Knowledge and Information Systems*, Vol. 12, N° 1, 1-24, 2007.
- Handl, Julia & Knowles, Joshua & Kell, Douglas (2005): Computational cluster validation in post-genomic data analysis. *Bioinformatics*, Vol.21, N°15, 3201—3212, 2005.
- C. S. Madeira & L. O. Oliveira (2004): Biclustering Algorithms for Biological Data Analysis : A Survey, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:24-25, 2004.