Clustering: evolution of methods to meet new challenges

C. Biernacki

Journée "Clustering", Orange Labs, October 20th 2015

Take home message

cluster

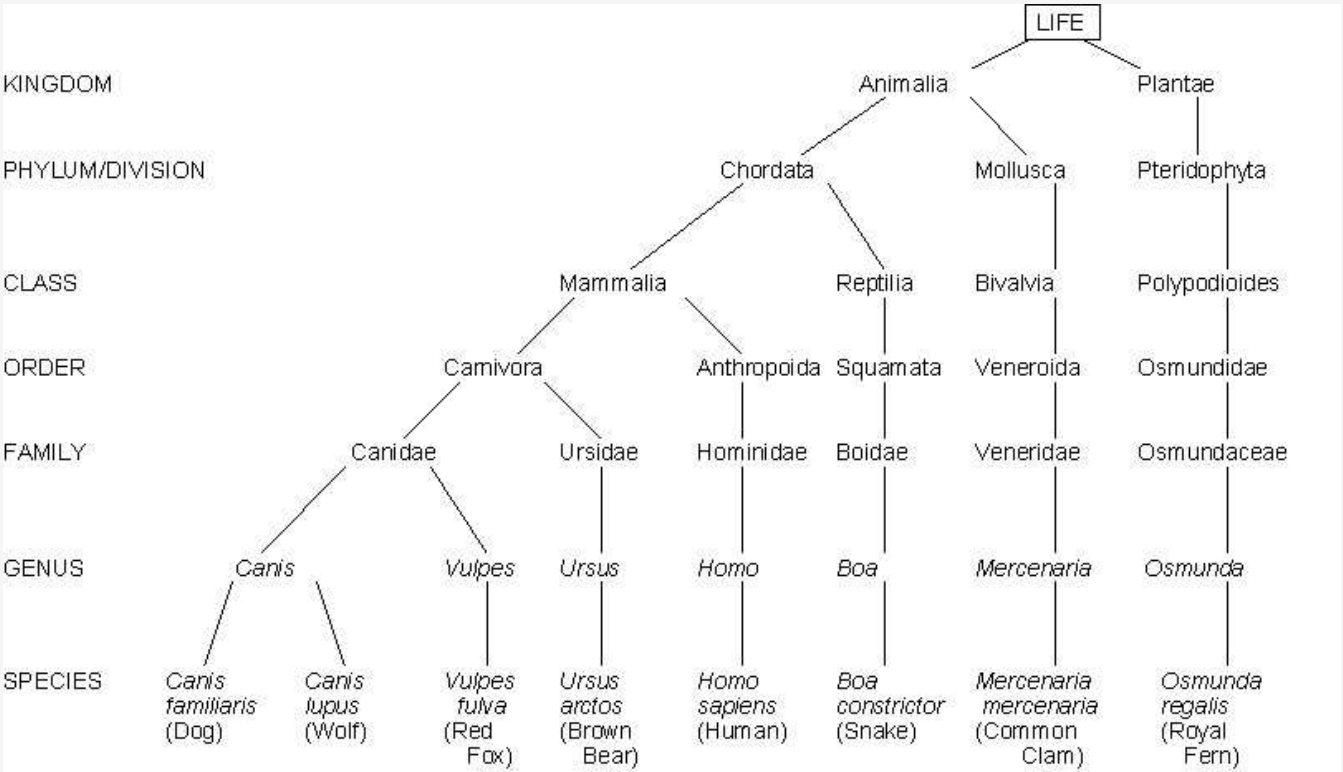
clustering

define both!

Outline

A first systematic attempt

- Carl von Linné (1707–1778), Swedish botanist, physician, and zoologist
- Father of modern taxonomy based on the most visible similarities between species
- *Linnaeus's Systema Naturae* (1st ed. in 1735) lists about 10,000 species of organisms (6,000 plants, 4,236 animals)



Interdisciplinary endeavor

- Medicine[1]: diseases may be classified by etiology (cause), pathogenesis (mechanism by which the disease is caused), or by symptom(s). Alternatively, diseases may be classified according to the organ system involved, though this is often complicated since many diseases affect more than one organ.
- And so on...

Typology

From Wikipedia, the free encyclopedia

Typology is the study of types. **Typology** may refer to:

- Typology (anthropology), division of culture by races
- Typology (archaeology), classification of artifacts according to their characteristics
- Typology (linguistics), study and classification of languages according to their structural features
- Typology (psychology), a model of personality types
- Typology (theology), in Christian theology, the interpretation of some figures and events in the Old Testament as foreshadowing the New Testament
- Typology (urban planning and architecture), the classification of characteristics common to buildings or urban spaces
- Typology (statistics), a concept in statistics, research design and social sciences

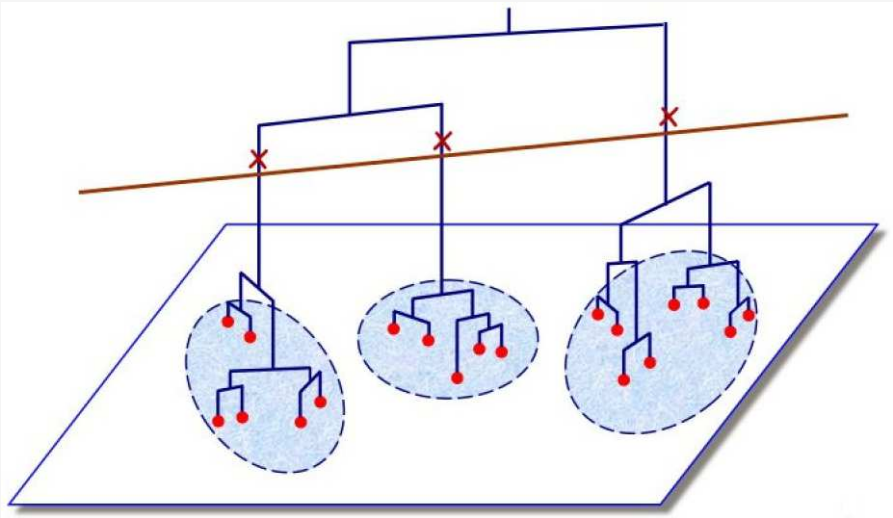
Look up ***typology***, ***typologist***, ***typological***, or ***typologically*** in Wiktionary, the free dictionary.

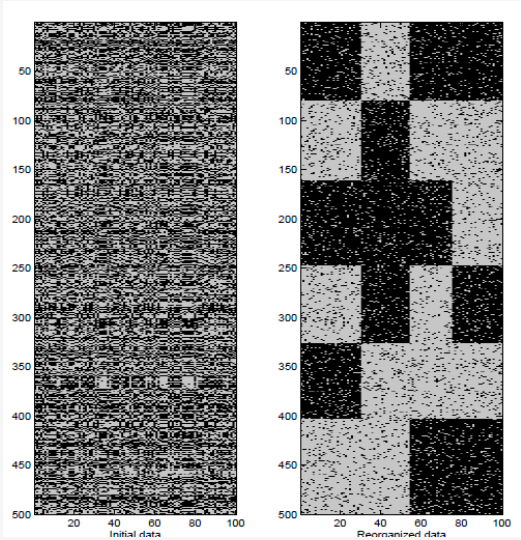
[1] Nosologie méthodique, dans laquelle les maladies sont rangées par classes, suivant le système de Sydenham, & l'ordre des botanistes, par François Boissier de Sauvages de Lacroix. Paris, Hérissant le fils, 1771

Three main clustering structures

Data set of n individuals $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i$ described by d variables

- Partition in K clusters denoted by $\mathbf{z} = (z_1, \ldots, z_n)$, with $z_i \in \{1, \ldots, K\}$
- Hierarchy Nested partitions
- Block partition Crossing simultaneously partitions in individuals and columns





Clustering is the cluster building process

Cluster analysis

From Wikipedia, the free encyclopedia

For the supervised learning approach, see Statistical classification.

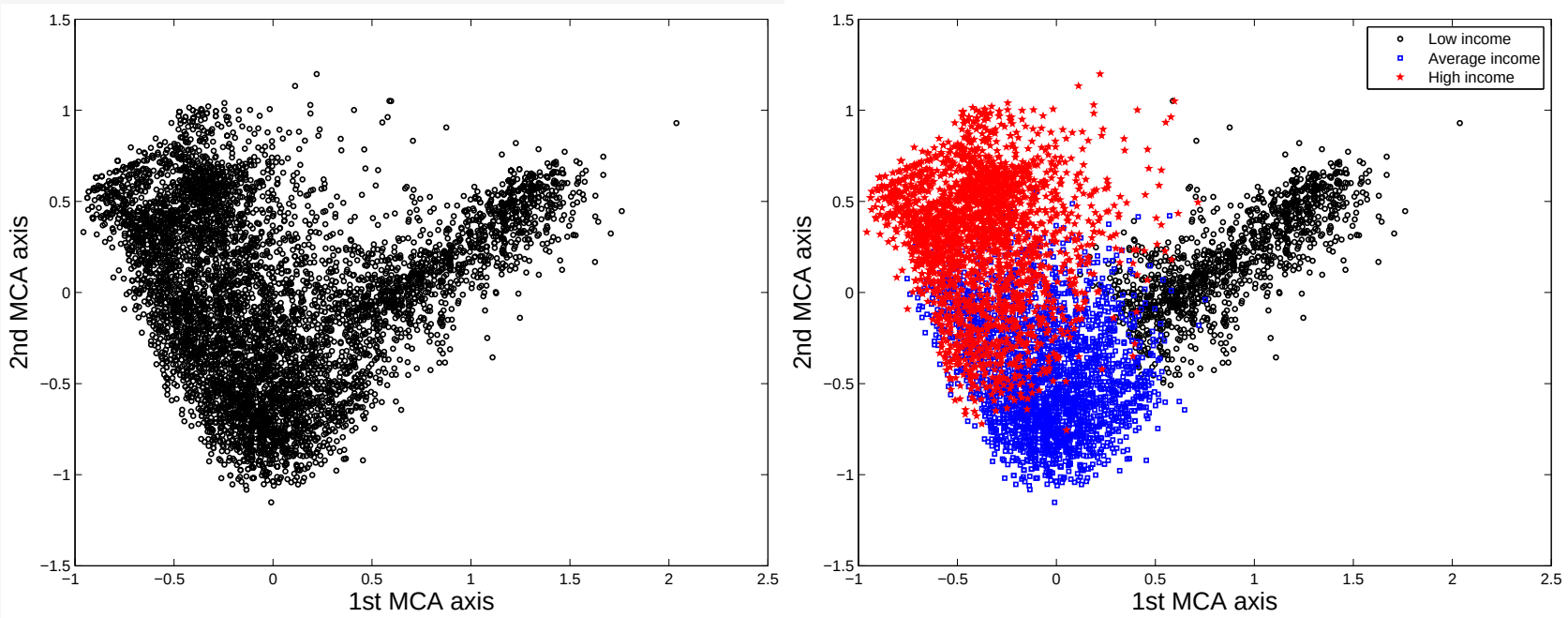
Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them.

- According to JSTOR, data clustering first appeared in the title of a 1954 article dealing with anthropological data
- Need to be automatic (algorithms) for complex data: mixed features, large data sets, high-dimensional data...

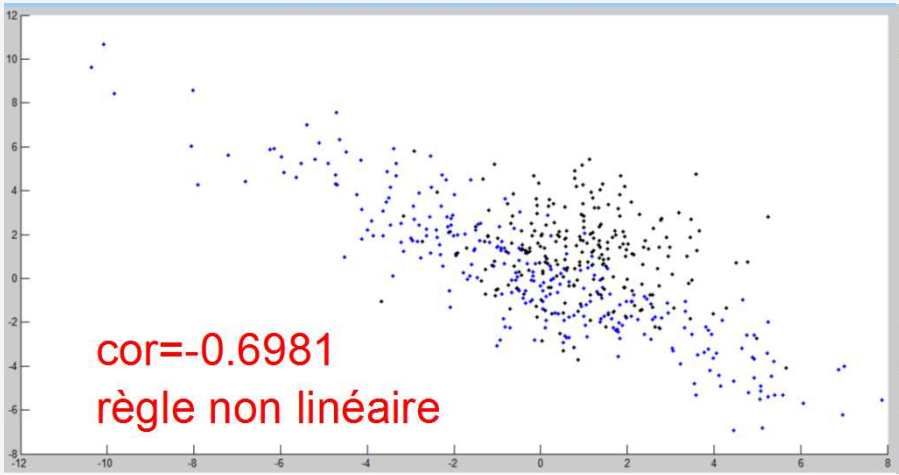
A 1st aim: explanatory task

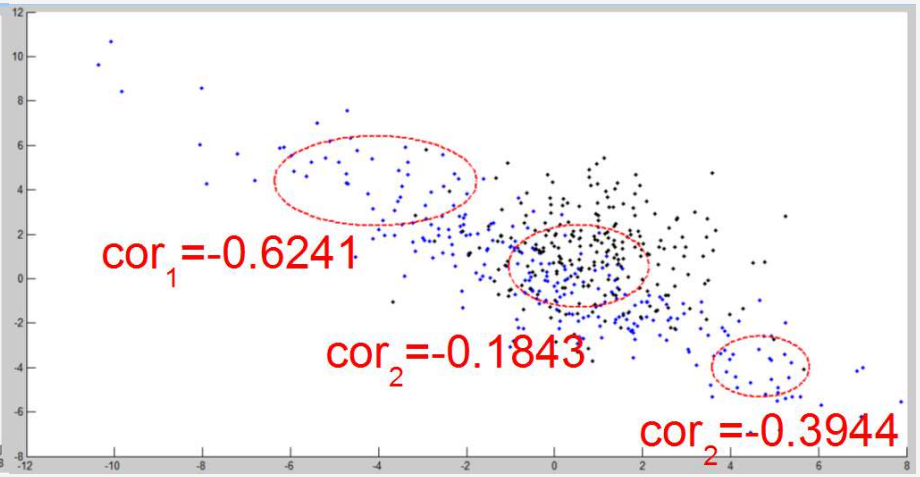
- A clustering for a marketing study
- Data: $d = 13$ demographic attributes (nominal and ordinal variables) of $n = 6\,876$ shopping mall customers in the San Francisco Bay (SEX (1. Male, 2. Female), MARITAL STATUS (1. Married, 2. Living together, not married, 3. Divorced or separated, 4. Widowed, 5. Single, never married), AGE (1. 14 thru 17, 2. 18 thru 24, 3. 25 thru 34, 4. 35 thru 44, 5. 45 thru 54, 6. 55 thru 64, 7. 65 and Over), *etc.*)
- Partition: retrieve less that 19 999$ (group of "low income"), between 20 000$ and 39 999$ group of "average income"), more than 40 000$ (group of "high income")

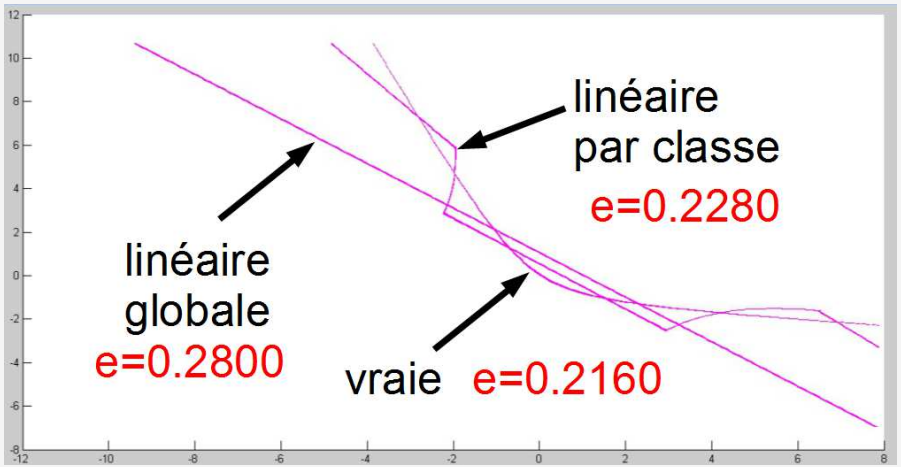


A 2nd aim: preprocessing step

- Logit model:
  - Not very flexible since linear borderline
  - Unbiased ML estimate by asymptotic variance $\sim n(\mathbf{x}'\mathbf{w}\mathbf{x})^{-1}$ is influenced by correlations
- A clustering may improve logistic regression prediction
  - More flexible borderline: piecewise linear
  - Decrease correlation so decrease variance







Mixed features

categorical: Marital status **maried**

integer: Children **3**

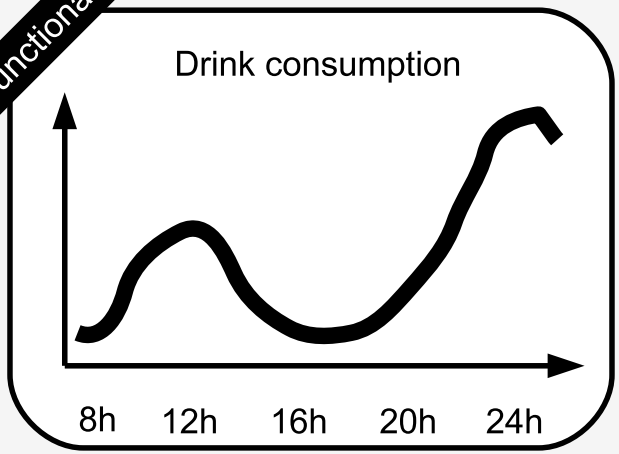
missing: Size (m) **?**

rank: Drink preference **beer > soda > water**

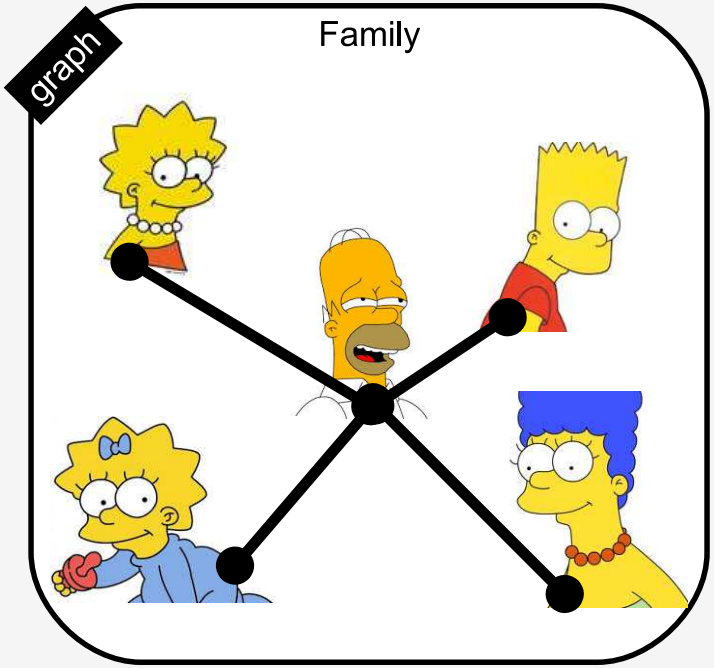
ordinal: Intelligence **low**

continuous: Weight (kg) **119.5**

functional

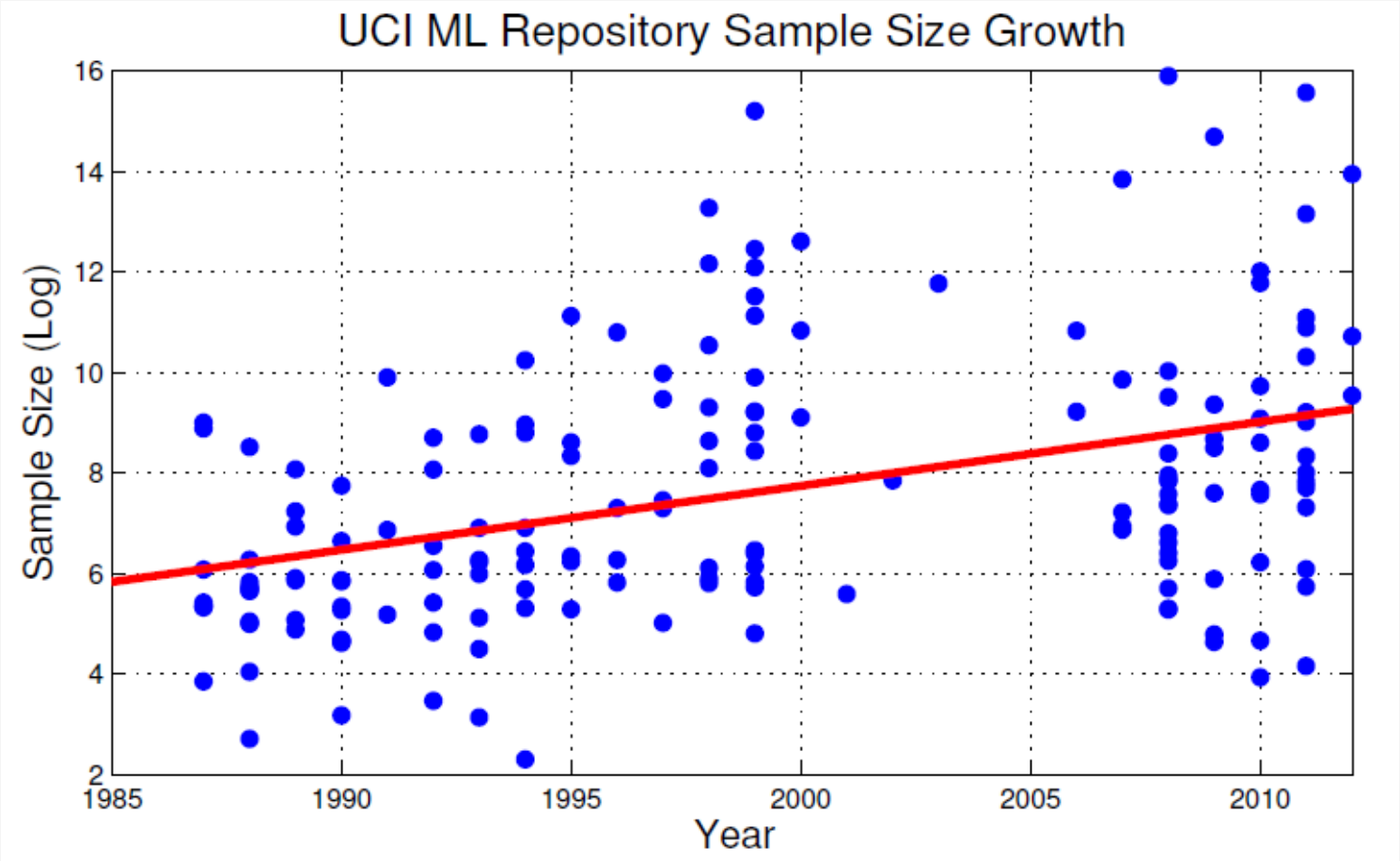


graph



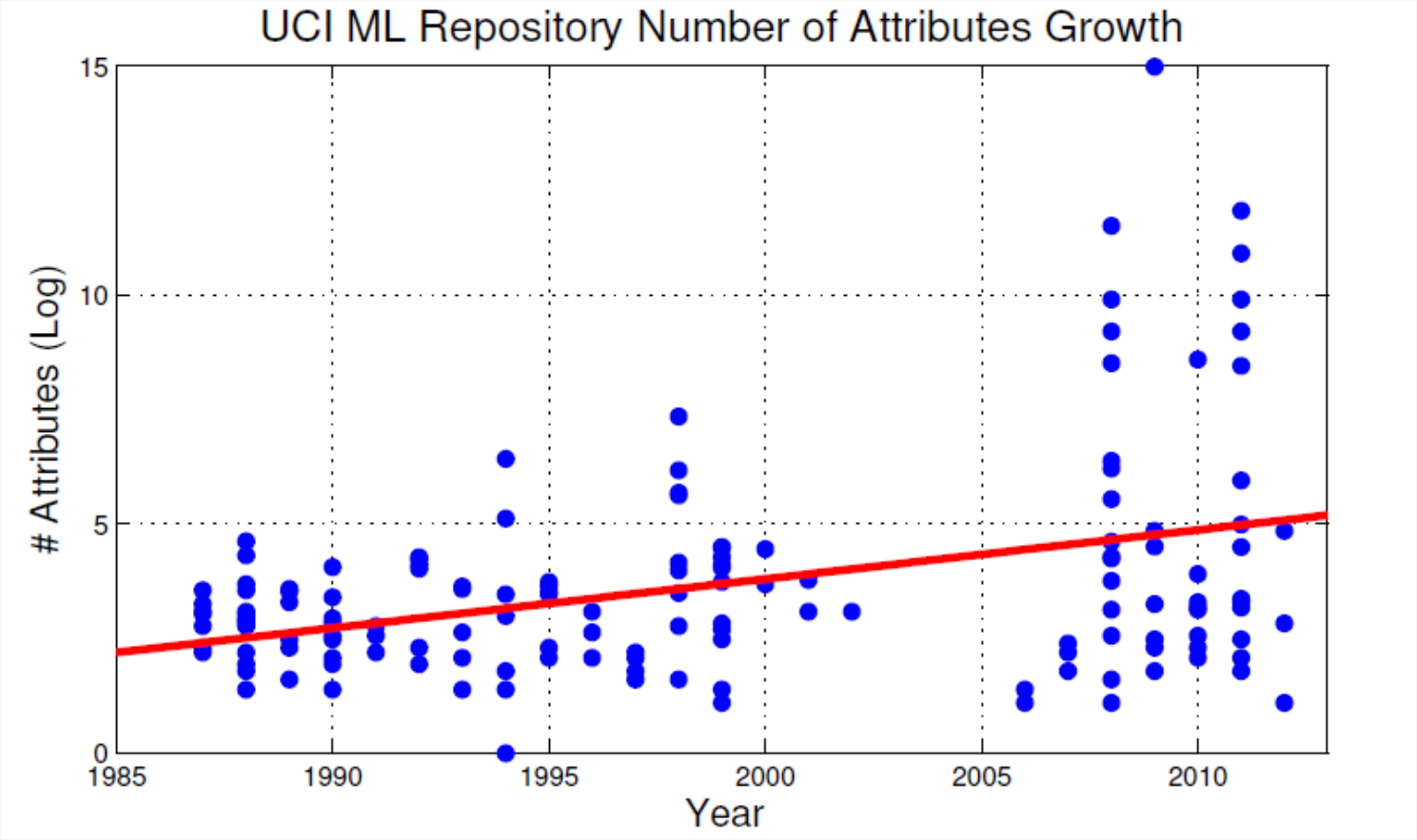
And so on...

Large data sets[2]



[2]S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, **29**

High-dimensional data[3]



[3] S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, **29**

Genesis of "Big Data"

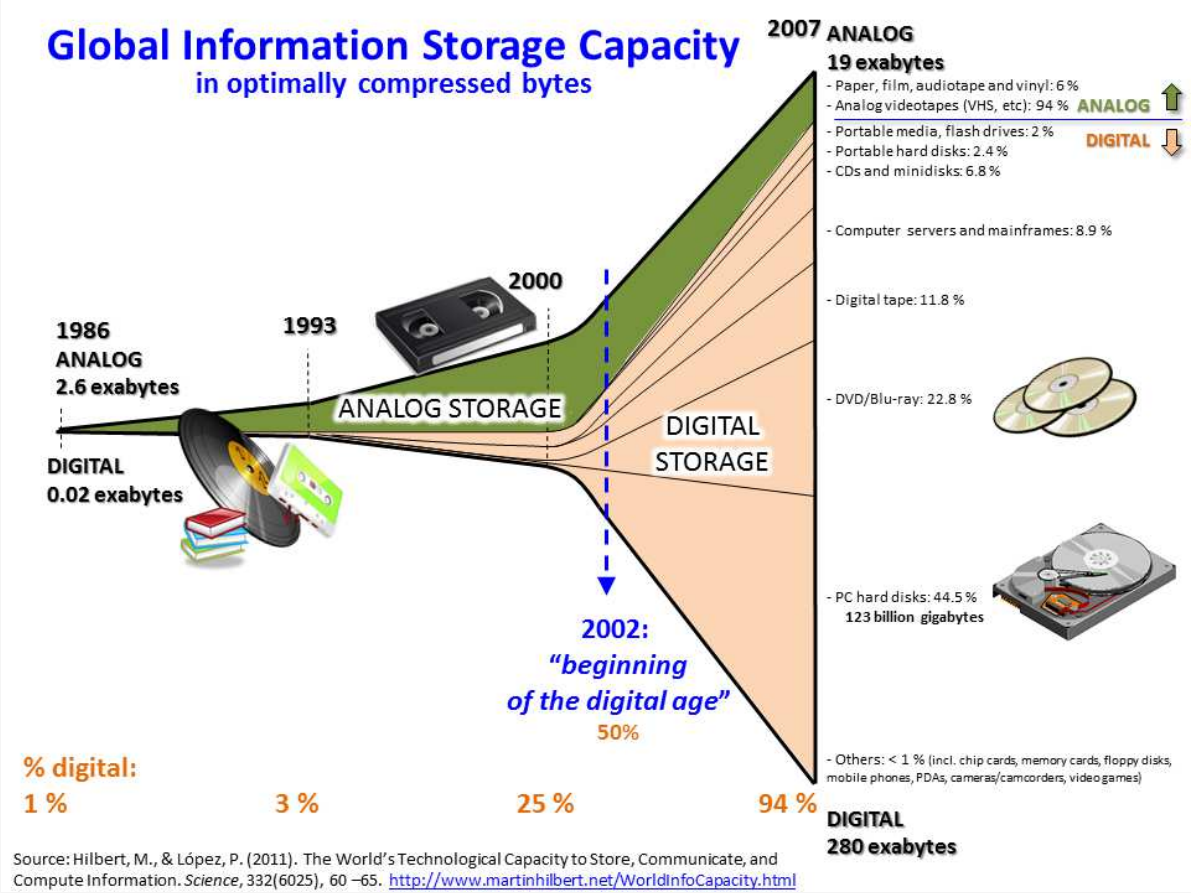
The Big Data phenomenon mainly originates in the increase of computer and digital resources at an ever lower cost

- Storage cost per MB: 700$ in 1981, 1$ in 1994, 0.01$ in 2013
  $\rightarrow$ price divided by 70,000 in thirty years
- Storage capacity of HDDs: $\approx$1.02 Go in 1982, $\approx$8 To today
  $\rightarrow$ capacity multiplied by 8,000 over the same period
- Computeur processing speed: 1 gigaFLOPS[4] in 1985, 33 petaFLOPS in 2013
  $\rightarrow$ speed multiplied by 33 million

[4] FLOP = FLoating-point Operations Per Second

Digital flow

- Digital in 1986: 1% of the stored information, 0.02 Eo[5]
- Digital in 2007: 94% of the stored information, 280 Eo (multiplied by 14,000)



[5] Exabyte

Societal phenomenon

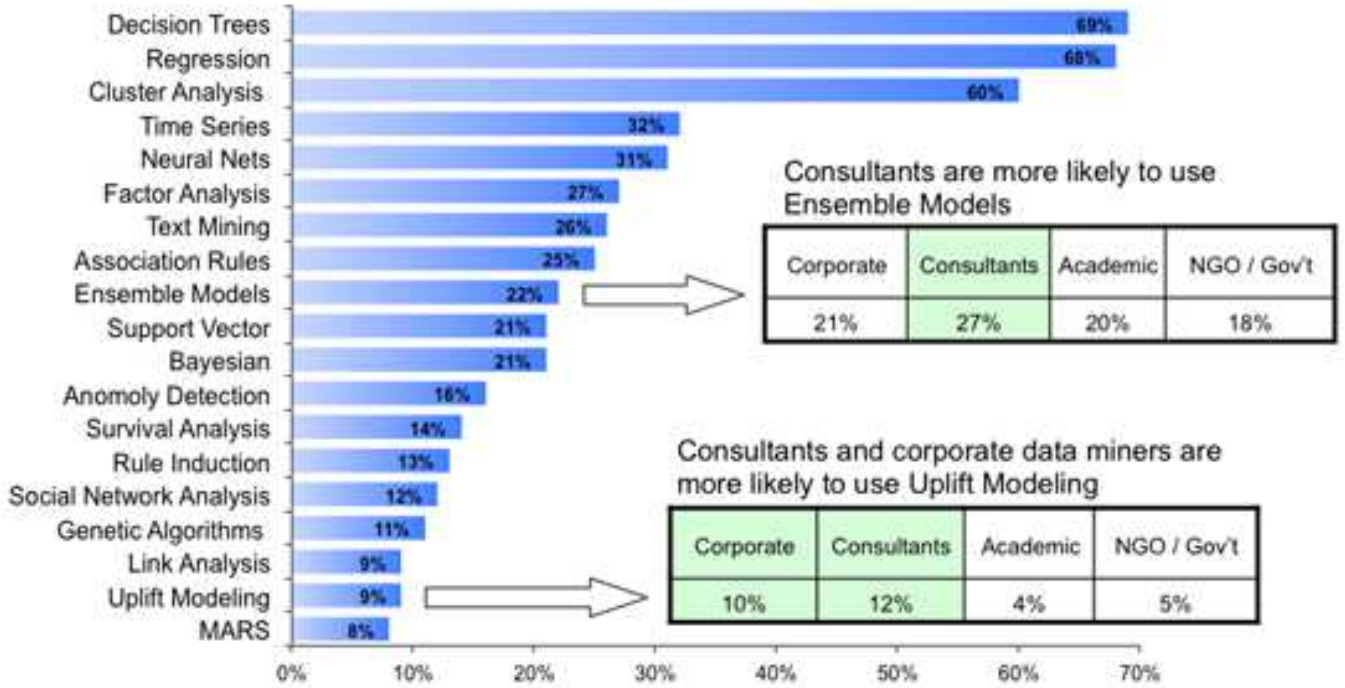
All human activities are impacted by data accumulation

- Trade and business: corporate reporting system , banks, commercial transactions, reservation systems. . .
- Governments and organizations: laws, regulations, standardizations , infrastructure. . .
- Entertainment: music, video, games, social networks. . .
- Sciences: astronomy, physics and energy, genome,. . .
- Health: medical record databases in the social security system. . .
- Environment: climate, sustainable development , pollution, power. . .
- Humanities and Social Sciences: digitization of knowledge , literature, history , art, architecture, archaeological data. . .

New data. . . but classical answers[6]

Data Mining Algorithms

- Decision trees, regression, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been very consistent over time.
- However, a wide variety of algorithms are being used.



[6] Rexer Analytics's Annual Data Miner Survey is the largest survey of data mining, data science, and analytics professionals in the industry (survey of 2011)

Outline

Clustering of clustering algorithms[7]

- Jain *et al.* (2004) hierarchical clustered 35 different clustering algorithms into 5 groups based on their partitions on 12 different datasets.
- It is not surprising to see that the related algorithms are clustered together.
- For a visualization of the similarity between the algorithms, the 35 algorithms are also embedded in a two-dimensional space obtained from the 35x35 similarity matrix.

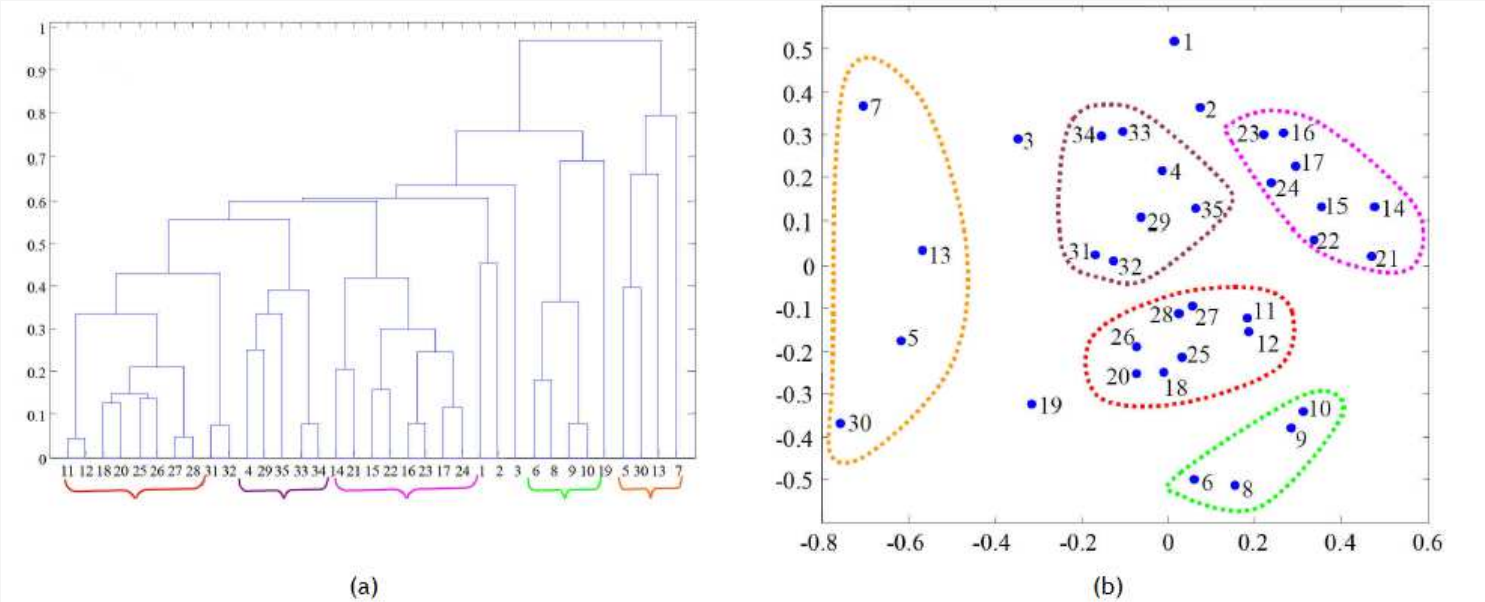
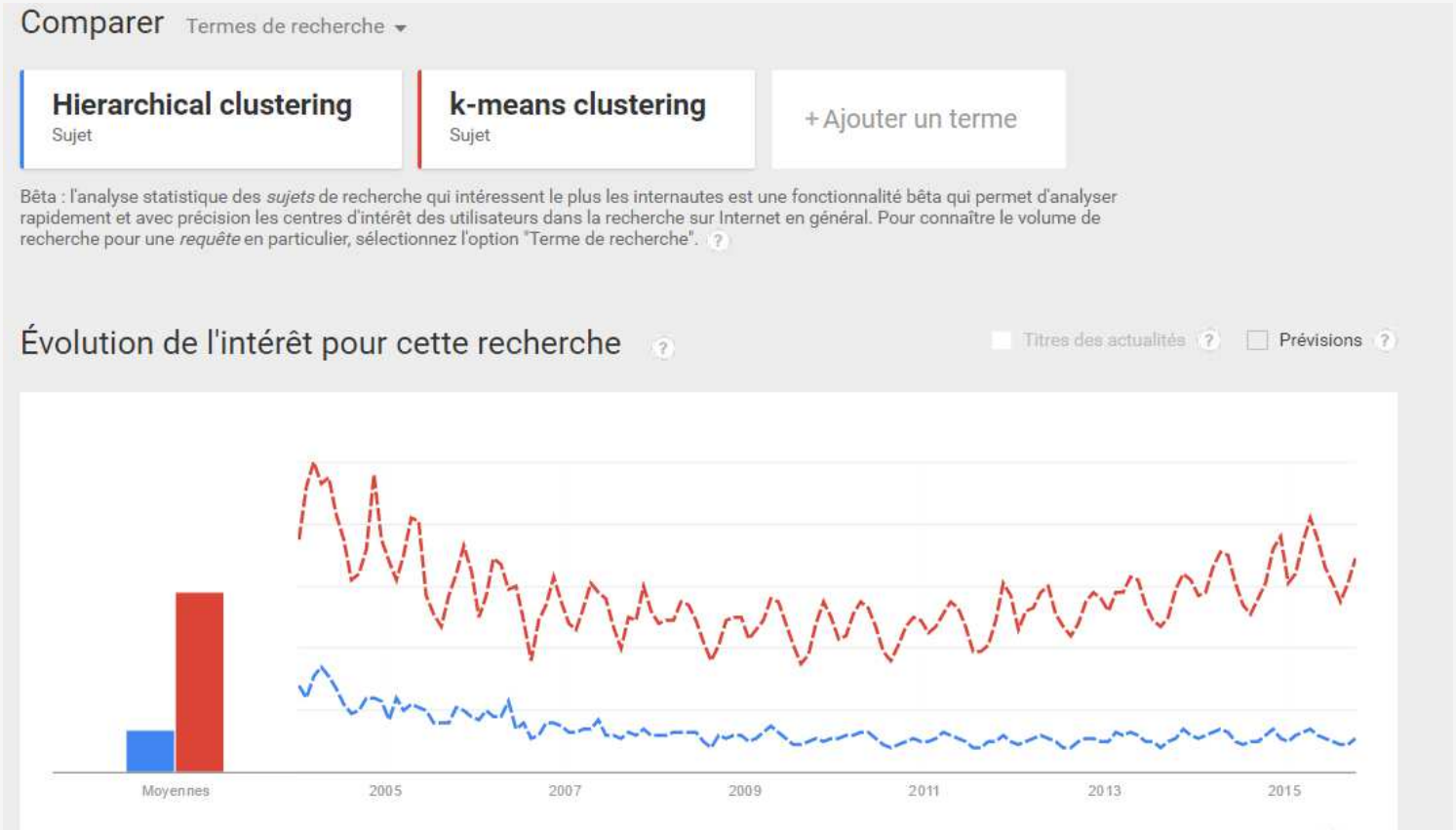


Figure 10 Clustering of clustering algorithms. (a) Hierarchical clustering of 35 different algorithms; (b) Sammon's mapping of the 35 algorithms into a two-dimensional space, with the clusters highlighted for visualization. The algorithms in the group (4, 29, 31-35) correspond to K-means, spectral clustering, Gaussian mixture models, and Ward's linkage. The algorithms in group (6, 8-10) correspond to CHAMELEON algorithm with different objective functions.

[7] A.K. Jain (2008). Data Clustering: 50 Years Beyond K-Means.

Popularity of K-means and hierarchical clustering

Even K-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering for several reasons: ease of implementation, simplicity, efficiency, empirical success. . . and model-based interpretation (see later)



Within-cluster inertia criterion

Select the partition $\mathbf{z}$ minimizing the criterion

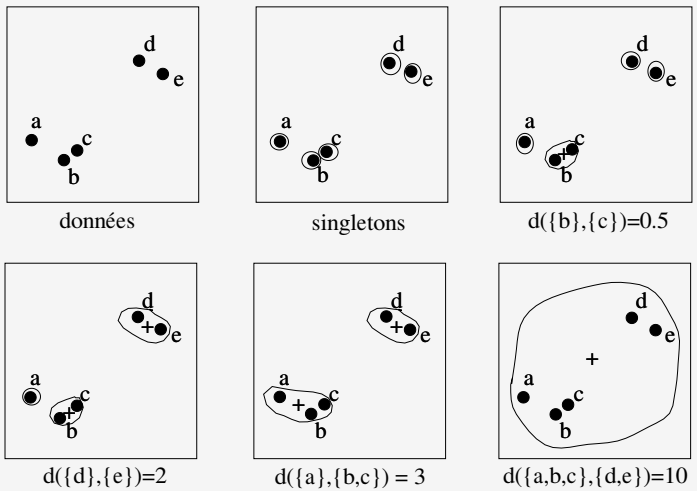
$$W_{\mathbf{M}}(\mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|_{\mathbf{M}}^2$$

- $\|\cdot\|_{\mathbf{M}}$ is the Euclidian distance with metric $\mathbf{M}$ in $\mathbb{R}^d$
- $\bar{\mathbf{x}}_k$ is the mean of the kth cluster

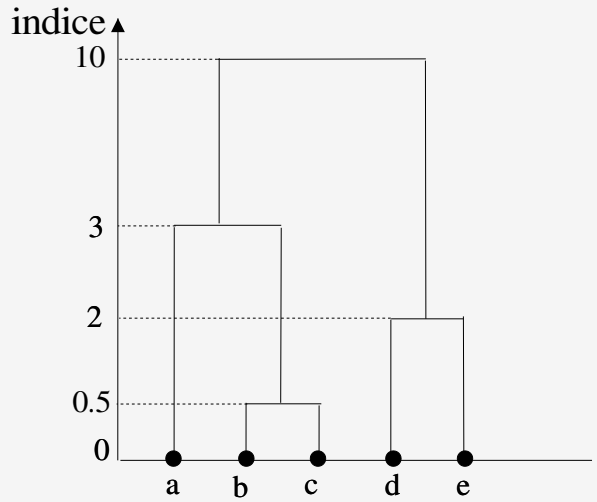
$$\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} \mathbf{x}_i$$

and $n_k = \sum_{k=1}^{n} z_{ik}$ indicates the number of individuals in cluster k

Ward hierarchical clustering



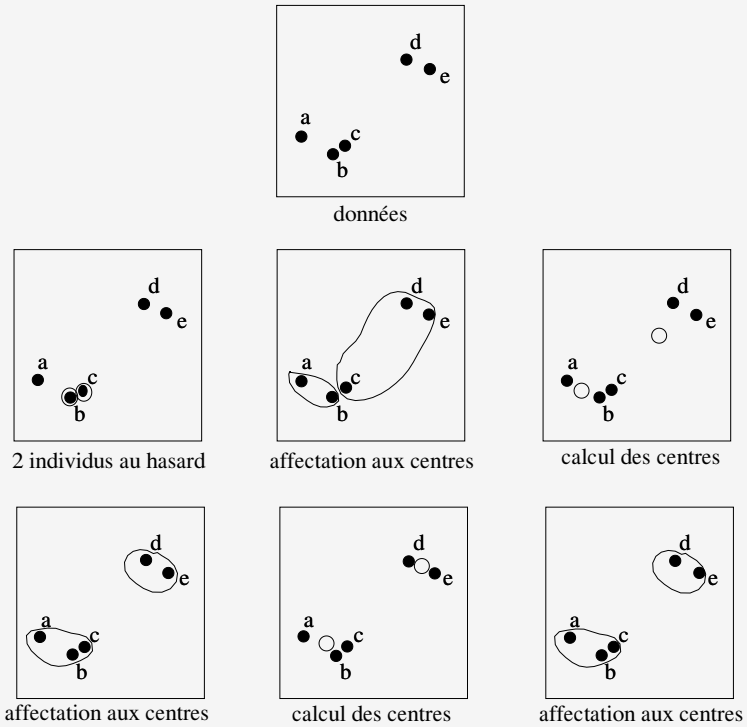
Classification hiérarchique ascendante
(méthode de Ward)



Dendogramme

- Suboptimal optimisation of $W_{\mathbf{M}}(\cdot)$
- A partition is obtained by cuting the dendrogram
- A dissimilarity matrix between pairs of individuals is enough

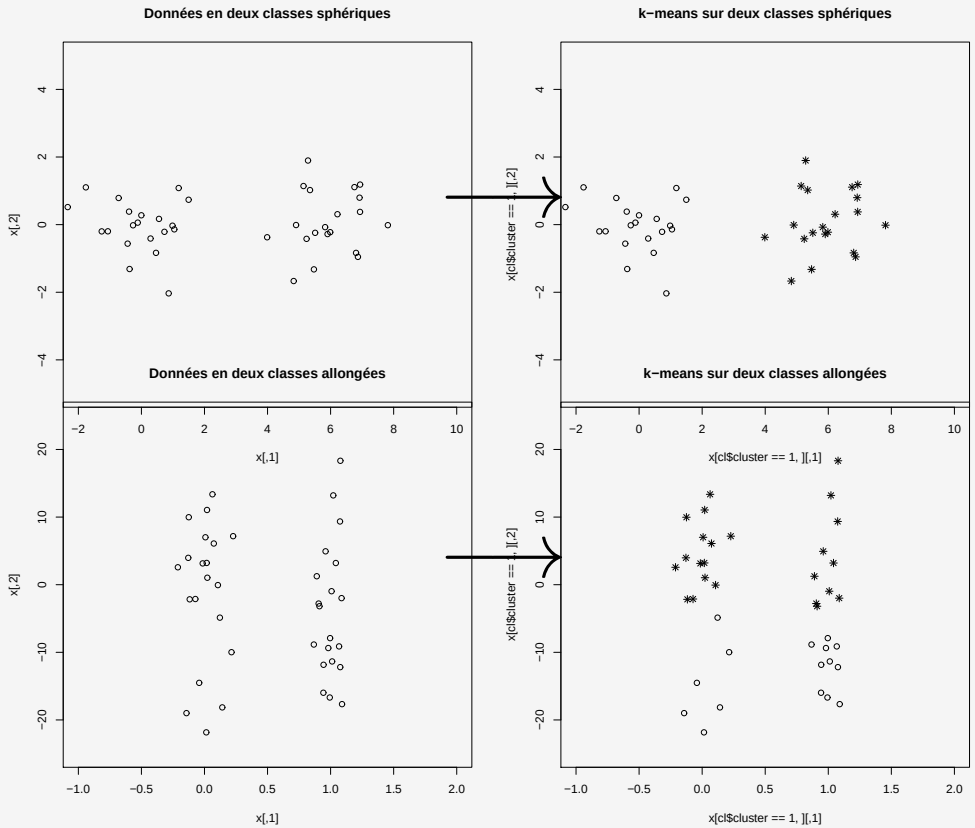
K-means algorithm



Algorithme des centres mobiles

Alternating optimization between the partition and the center of clusters

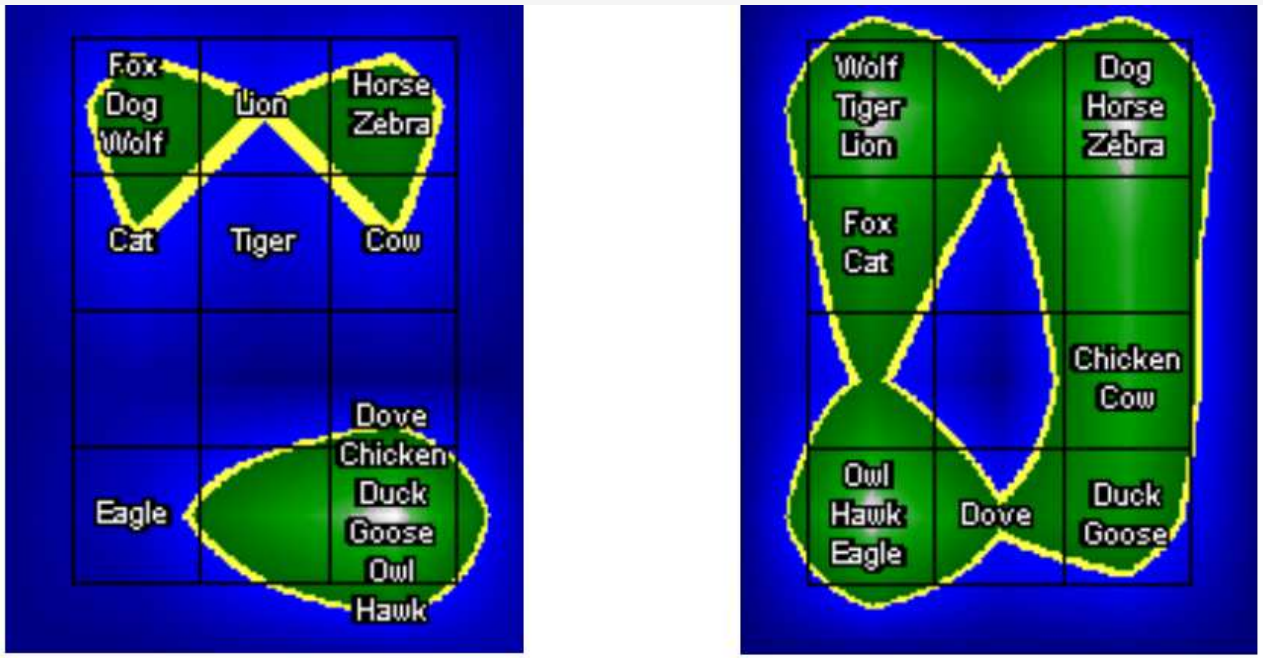
The identity metric $\mathbf{M}$: a classical but hazardous choice



Alternative: estimate $\mathbf{M}_{(k)}$ by minimizing $W_{\mathbf{M}_{(k)}}(\mathbf{z})$ over $(\mathbf{z}, \mathbf{M}_{(k)})$

Effect of the metric **M** through a real example

- Animals represented by 13 Boolean features related to appearance and activity
- Large weight on the appearance features compared to the activity features: the animals were clustered into mammals vs. birds
- Large weight on the activity features: partitioning predators vs. non-predators
- Both partitions are equally valid, and uncover meaningful structures in the data
- The user has to carefully choose his representation to obtain a desired clustering

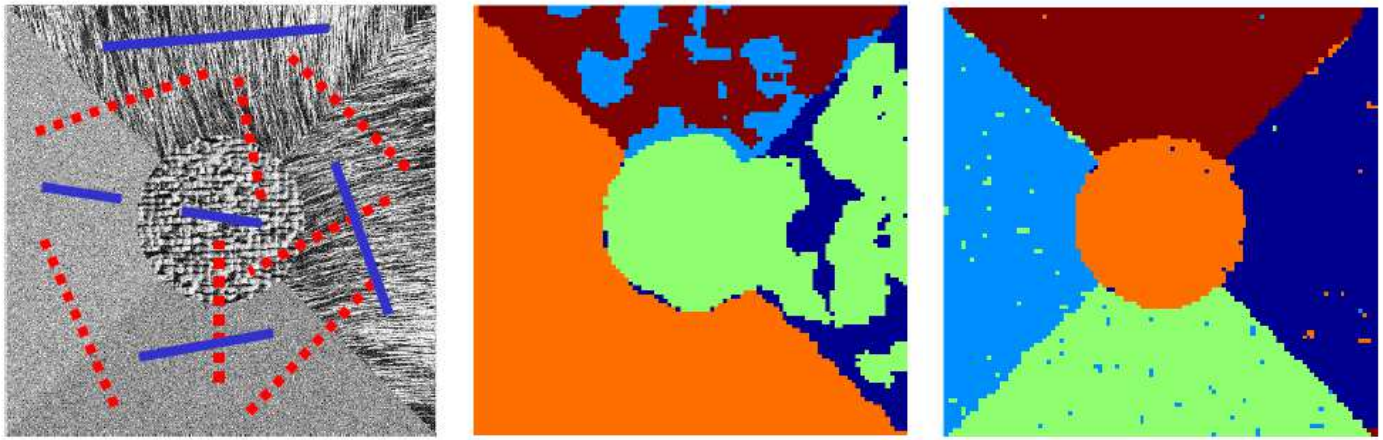


(a) (b)

Figure 6 Different weights on features result in different partitioning of the data. Sixteen animals are represented based on 13 Boolean features related to appearance and activity. (a) partitioning with large weights assigned to appearance based features; (b) a partitioning with large weights assigned to the activity features (figure reproduced from [**Pampalk *et al.* , 2003**]).

Semi-supervised clustering[8]

- The user has to provide any external information he has on the partition
- Pair-wise constraints:
  - A must-link constraint specifies that the point pair connected by the constraint belong to the same cluster
  - A cannot-link constraint specifies that the point pair connected by the constraint do not belong to the same cluster
- Attempts to derive constraints from domain ontology and other external sources into clustering algorithms include the usage of WordNet ontology, gene ontology, Wikipedia, *etc.* to guide clustering solutions



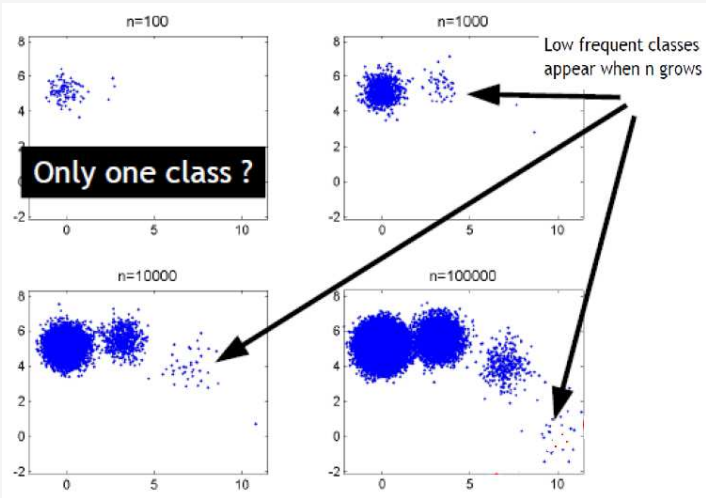
(a) Input image and constraints (b) No constraints (c) 10% pixels in constraints

Figure 12 Semi-supervised learning. (a) Input image with must-link (solid blue lines) and must not link (broken red lines) constraints. (b) Clustering (segmentation) without constraints. (c) Improved clustering with 10% of the data points included in the pair-wise constraints [6].

[8] O. Chapelle *et al.* (2006), A.K. Jain (2008). Data Clustering: 50 Years Beyond K-Means.

Online clustering

- Dynamic data are quite recent: blogs, Web pages, retail chain, credit card transaction streams, network packets received by a router and stock market, *etc.*
- As the data gets modified, clustering must be updated accordingly: ability to detect emerging clusters, *etc.*
- Often all data cannot be stored on a disk
- This imposes additional requirements to traditional clustering algorithms to rapidly process and summarize the massive amount of continuously arriving data
- Data stream clustering a significant challenge since they are expected to involved single-pass algorithms



Data representation challenge[9]

The data unit can be crucial for the data clustering task

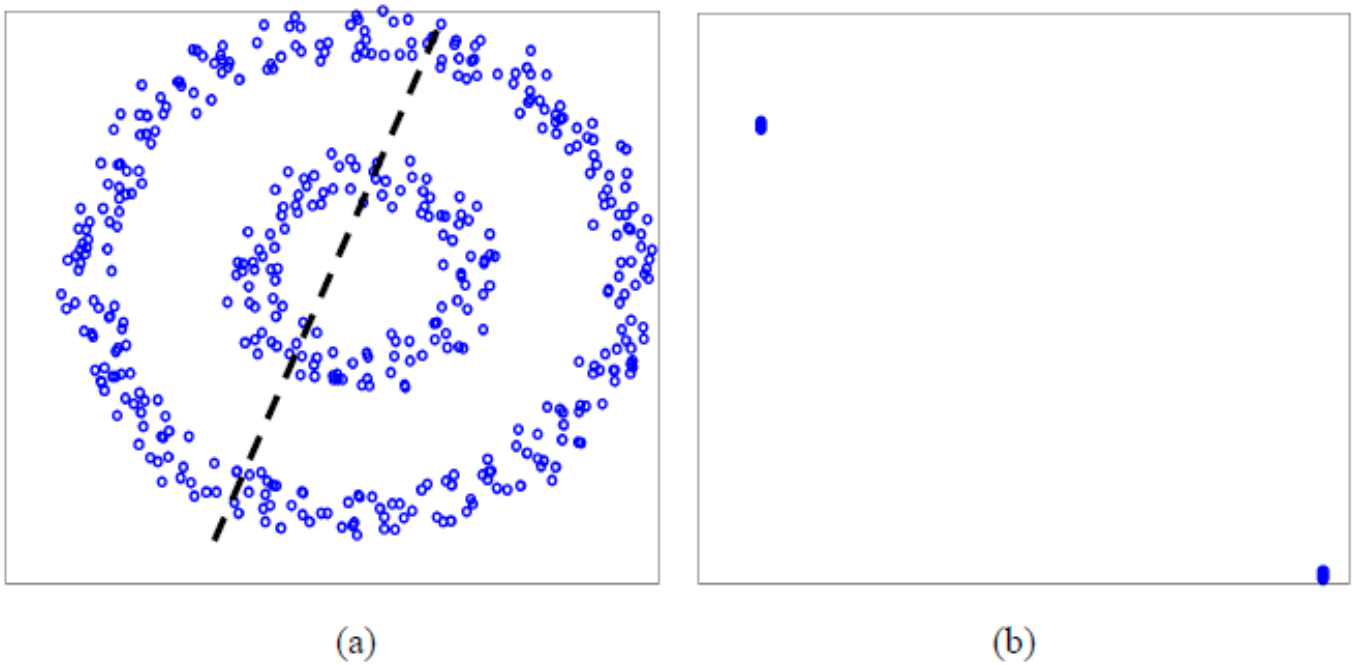


Figure 5 Importance of a good representation. (a) "Two rings" dataset where K-means fails to find the two "natural" clusters; the dashed line shows the linear cluster separation boundary obtained by running K-means with K = 2. (b) a new representation of the data in (a) based on the top 2 eigenvectors of the graph Laplacian of the data, computed using an RBF kernel; K-means now can easily detect the two clusters

[9] A.K. Jain (2008). Data Clustering: 50 Years Beyond K-Means.

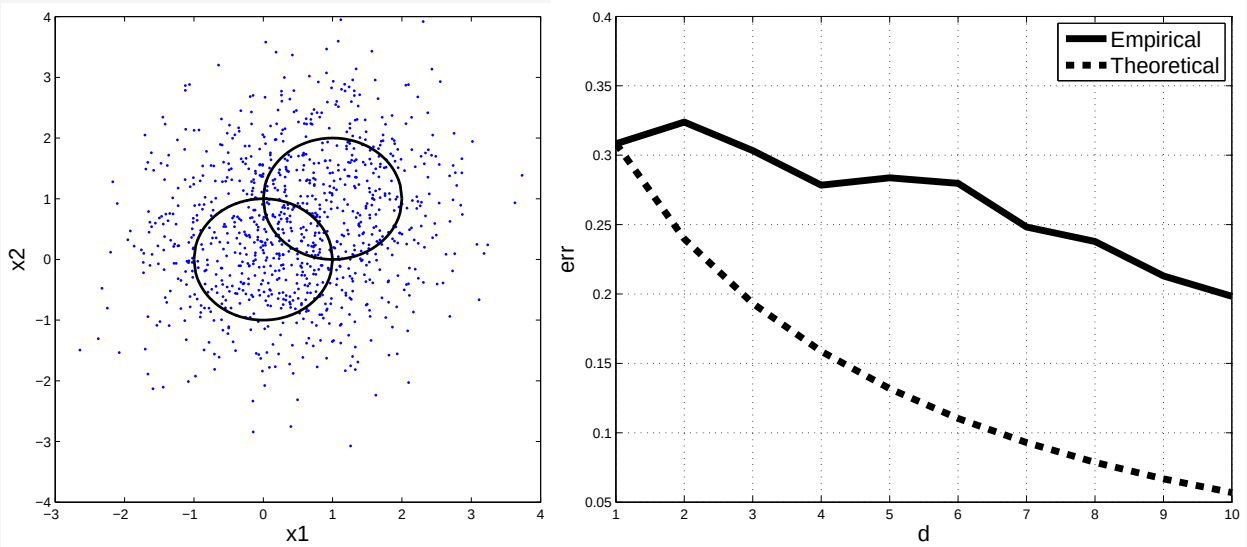
HD clustering: blessing (1/2)

A two-component d-variate Gaussian mixture:

$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \mathbf{X}_1|z_{11} = 1 \sim \mathrm{N}_d(\mathbf{0}, \mathbf{I}), \quad \mathbf{X}_1|z_{12} = 1 \sim \mathrm{N}_d(\mathbf{1}, \mathbf{I})$$

Each variable provides equal and own separation information

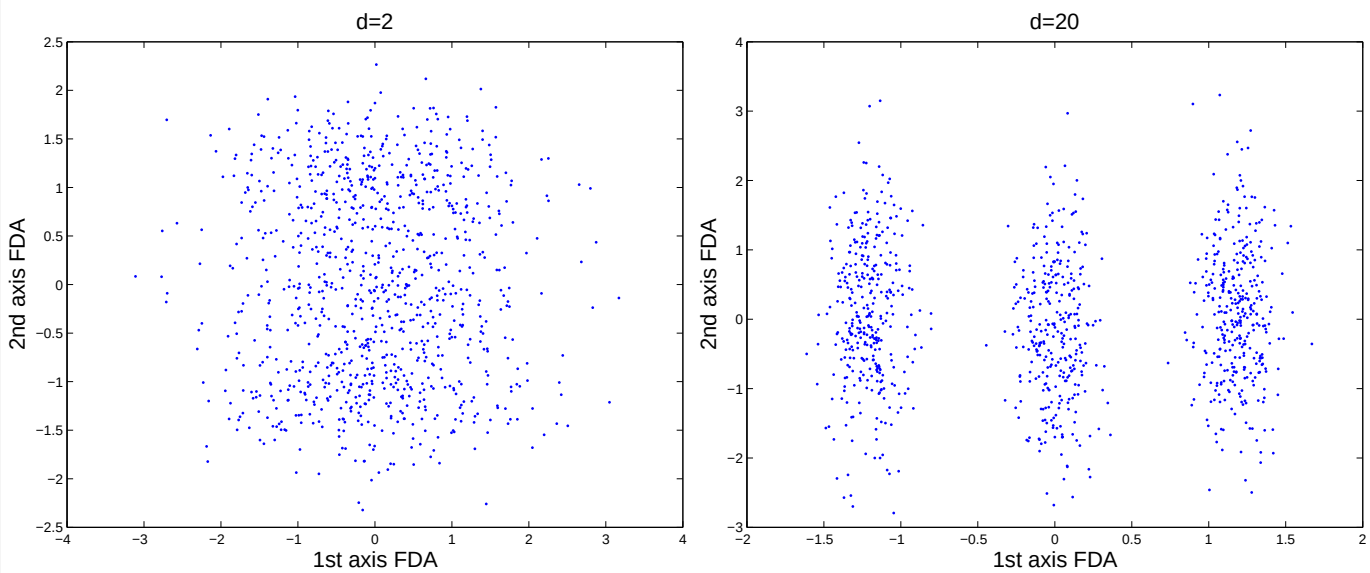
Theoretical error decreases when d grows: $\mathrm{err}_{theo} = \Phi(-\sqrt{d}/2)\ldots$



. . . and empirical error rate decreases also with d!

HD clustering: blessing (2/2)

FDA



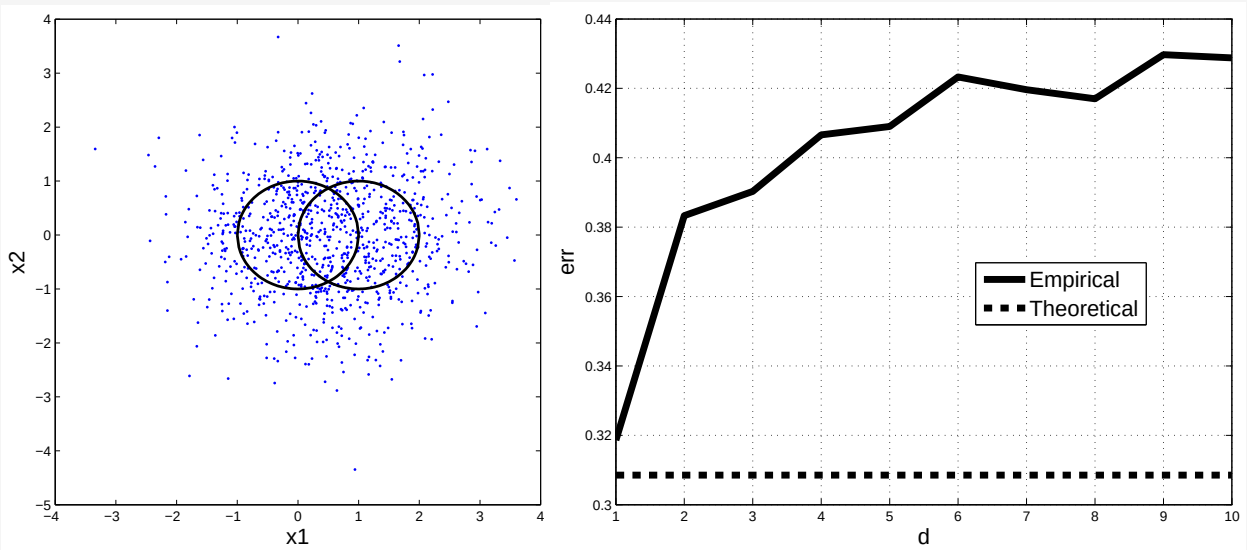
HD clustering: curse (1/2)

Many variables provide no separation information

Same parameter setting except:

$$\mathbf{X}_1 | z_{12} = 1 \sim \mathrm{N}_d((1\ 0\ \ldots\ 0)', \mathbf{I})$$

Groups are not separated more when d grows: $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{I}} = 1 \ldots$



. . . thus theoretical error is constant $(= \Phi(-\frac{1}{2}))$ and empirical error increases with d

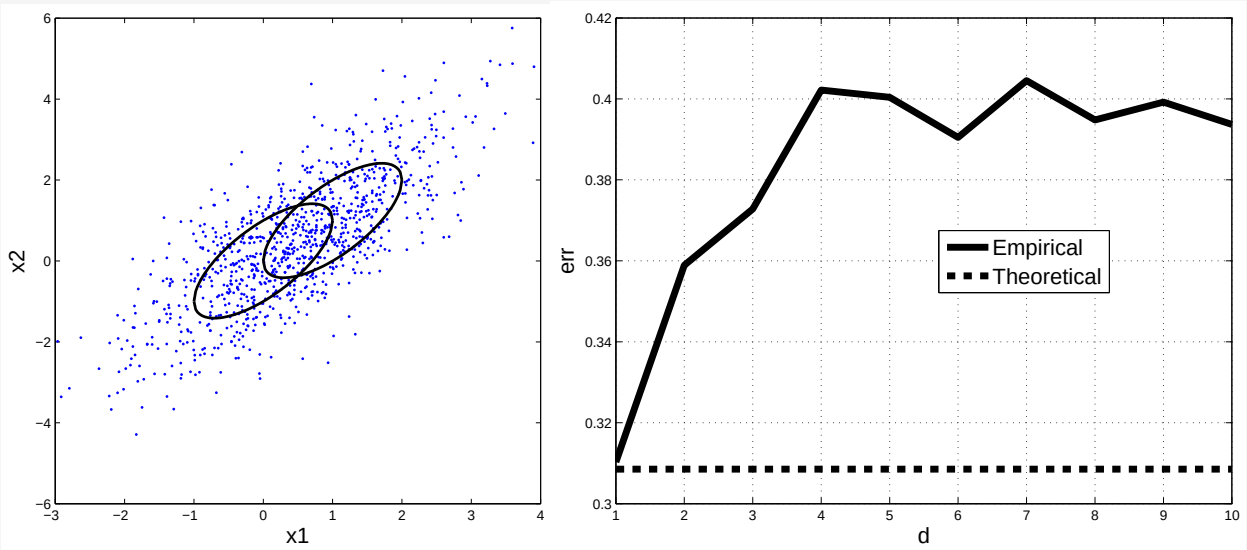
HD clustering: curse (2/2)

Many variables provide redundant separation information

Same parameter setting except:

$$\mathbf{X}_1^j = \mathbf{X}_1^1 + \mathrm{N}_1(0,1) \quad (j = 2, \ldots, d)$$

Groups are not separated more when d grows: $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}} = 1 \ldots$



. . . thus err_{theo} is constant $(= \Phi(-\frac{1}{2}))$ and empirical error increases (less) with d

Clustering: an ill-posed problem

Questions to be addressed:

- What is the best metric $\mathbf{M}_{(k)}$?
- How to choose the number K of clusters: $W_{\mathbf{M}}(\mathbf{z})$ decreases $K \ldots$
- clusters of different sizes are they well estimated?
- How to choose the data unit?
- How to select features in a high-dimensional context?
- How to deal with mixed data?
- ...

First, answer to... what is formally a cluster?

Model-based clustering solution

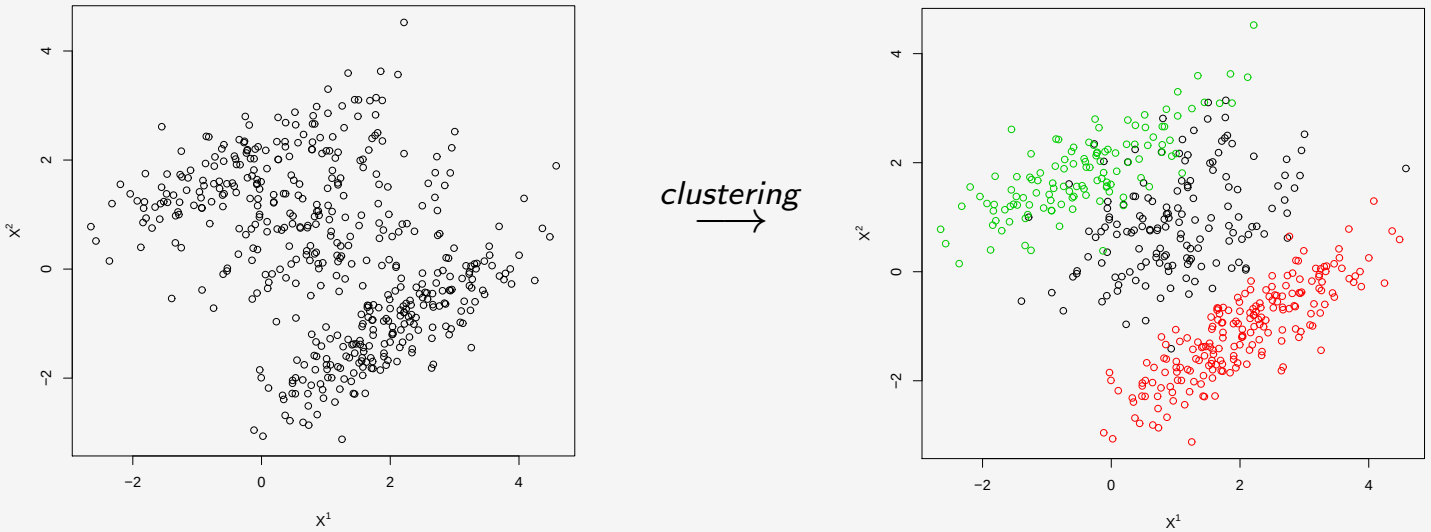
$$\text{a cluster} \iff \text{a distribution}$$

It recasts all previous problems into model design/estimation/selection

Outline

Model-based clustering

$\mathbf{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ $\hat{\mathbf{z}} = (\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_n)$, $\hat{K}$ clusters



Mixture model: well-posed problem

$$p(\mathbf{x}|K;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|K;\boldsymbol{\theta}_k) \quad \text{can be used for} \quad \begin{cases} \mathbf{x} \to \hat{\boldsymbol{\theta}} \to p(\mathbf{z}|\mathbf{x}, K; \hat{\boldsymbol{\theta}}) \to \hat{\mathbf{z}} \\ \mathbf{x} \to \hat{p}(K|\mathbf{x}) \to \hat{K} \\ \ldots \end{cases}$$

with $\boldsymbol{\theta} = (\pi_k, (\boldsymbol{\alpha}_k))$

Hypothesis of mixture of parametric distributions

- cluster k is modelled by a parametric distribution:

$$\mathbf{X}_{i|Z_{ik}=1} \overset{i.i.d.}{\sim} p(\cdot;\ \boldsymbol{\alpha}_k)$$

- cluster k has probability π_k with $\sum_{k=1}^{K} \pi_k = 1$:

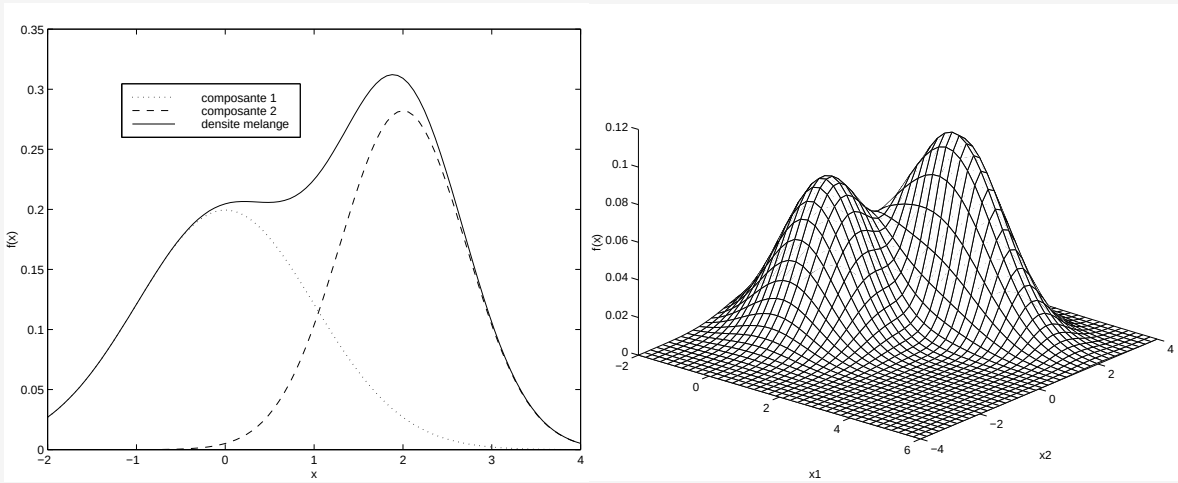
$$\mathbf{Z}_i \overset{i.i.d.}{\sim} \text{Mult}_K(1, \pi_1, \ldots, \pi_K)$$

The whole mixture parameter $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$

$$p(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_i; \boldsymbol{\alpha}_k)$$

Gaussian mixture

$$p(\cdot;\boldsymbol{\alpha}_k) = \mathsf{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{where} \quad \boldsymbol{\alpha}_k = (\underbrace{\boldsymbol{\mu}_k}_{\text{mean}}, \underbrace{\boldsymbol{\Sigma}_k}_{\text{covariance matrix}})$$



Parameter = summary + help to understand

cluster k is described by meaningful parameters:
cluster size (π_k), position ($\boldsymbol{\mu}_k$) et dispersion ($\boldsymbol{\Sigma}_k$).

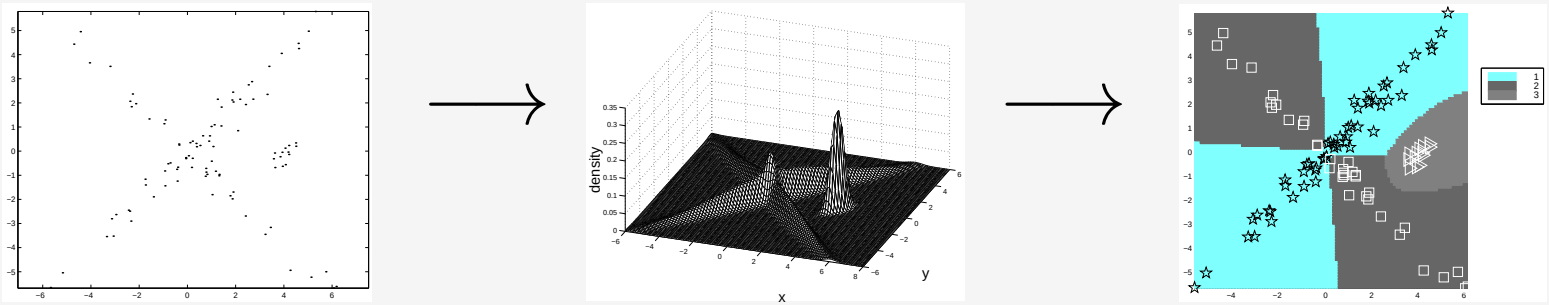
The clustering process in mixtures

1. Estimation of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$
2. Estimation of the conditional probability that $\mathbf{x}_i \in$ cluster k

$$t_{ik}(\hat{\boldsymbol{\theta}}) = p(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i; \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_k p(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)}{p(\mathbf{x}_i; \hat{\boldsymbol{\theta}})}$$

3. Estimation of $\mathbf{z}_i$ by *maximum a posteriori* (MAP)

$$\hat{z}_{ik} = \mathbb{I}_{\{k = \arg\max_{h=1,\ldots,K} t_{ih}(\hat{\boldsymbol{\theta}})\}}$$



Estimation of $\boldsymbol{\theta}$ by *complete*-likelihood

Maximize the complete-likelihood over $(\boldsymbol{\theta}, \mathbf{z})$

$$\ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \{\pi_k p(\mathbf{x}_i; \boldsymbol{\alpha}_k)\}$$

- Equivalent to tradition methods

Metric	$\mathbf{M} = \mathbf{I}$	$\mathbf{M}$ free	$\mathbf{M}_k$ free
Gaussian model	$[\pi\lambda I]$	$[\pi\lambda C]$	$[\pi\lambda_k C_k]$

- Bias of $\hat{\boldsymbol{\theta}}$: heavy if poor separated clusters
- Associated optimization algorithm: CEM (see later)
- CEM with $[\pi\lambda I]$ is strictly equivalent to K-means
- CEM is simple et fast (convergence with few iterations)

Estimation of $\boldsymbol{\theta}$ by *observe*-likelihood

Maximize the observe-likelihood on $\boldsymbol{\theta}$

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i; \boldsymbol{\theta})$$

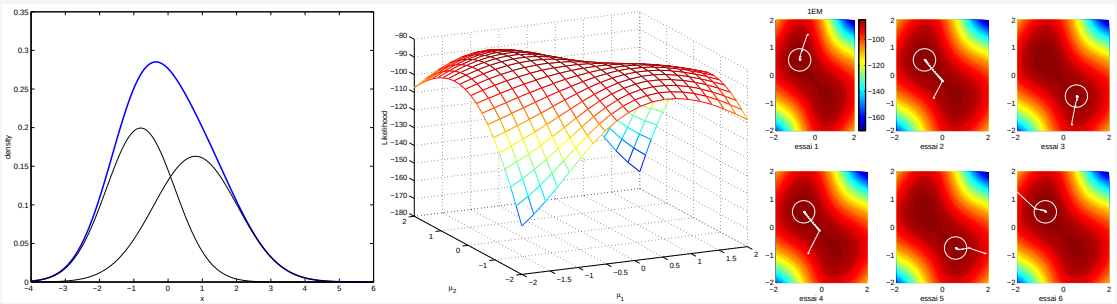
- Convergence of $\hat{\boldsymbol{\theta}}$
- General algorithm for missing data: EM
- EM is simple but slower than CEM
- Interpretation: it is a kind of fuzzy clustering

Principle of EM and CEM

- Initialization: $\boldsymbol{\theta}^0$
- Iteration n^o q:
  - Step E: estimate probabilities $\mathbf{t}^q = \{t_{ik}(\boldsymbol{\theta}^q)\}$
  - Step C: classify by setting $\mathbf{t}^q = \text{MAP}(\{t_{ik}(\boldsymbol{\theta}^q)\})$
  - Step M: maximize $\boldsymbol{\theta}^{q+1} = \arg\max_{\boldsymbol{\theta}} \ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}^q)$
- Stopping rule: iteration number or criterion stability

Properties

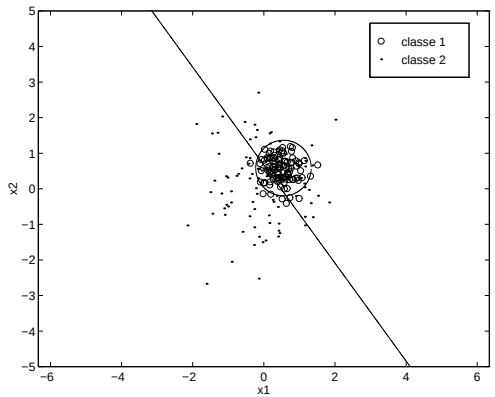
- $\oplus$: simplicity, monotony, low memory requirement
- $\ominus$: local maxima (depends on $\boldsymbol{\theta}^0$), linear convergence (EM)



Importance of model selection

Model = number of clusters + parametric structure of clusters

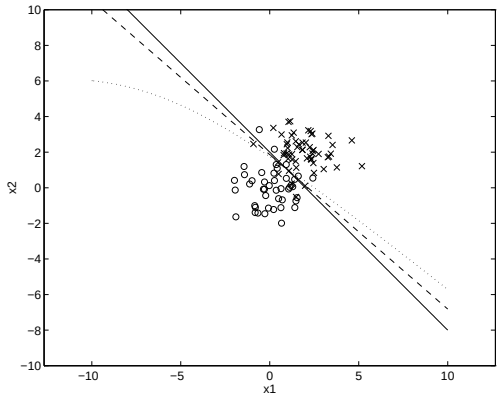
Too simple model: bias



true modèle: $[\pi\lambda_k I]$

too simple model: $[\pi\lambda I]$

Too complex model: variance



true borderline

borderline with $[\pi\lambda I]$

. . . borderline with $[\pi\lambda_k C_k]$

Model selection criteria

The most widespread principle

$$\underbrace{\text{Criterion}}_{\text{to be maximized}} = \underbrace{\text{maximum log-likelihood}}_{\text{model-data adequacy}} - \underbrace{\boxed{\text{penalty}}}_{\text{"cost" of the model}}$$

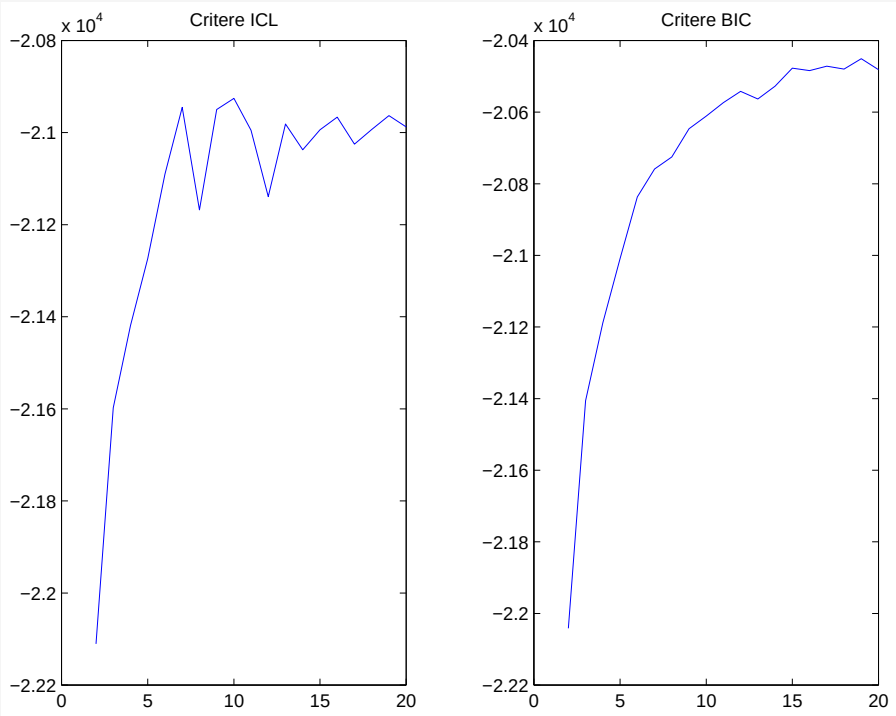
criterion	penalty	interpretation	user purpose
	general criteria in statistics		
AIC	ν	model complexity	prediction
BIC	$0.5\nu \ln(n)$	model complexity	identification
	specific criterion for the clustering aim		
ICL	$0.5\nu \ln(n)$ $- \sum_{i,k} \hat{z}_{ik} \ln t_{ik}(\hat{\boldsymbol{\theta}})$	model complexity + partition entropy	well-separated clusters

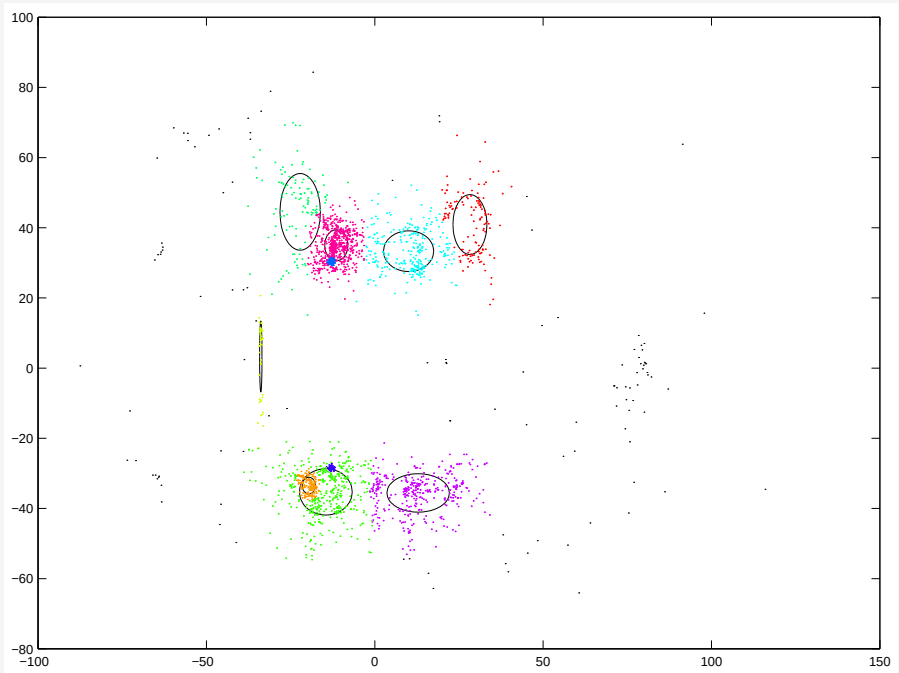
N.B.: in a prediction context, it is also possible to use the predictive error rate

Outline

ICL/BIC for acoustic emission control

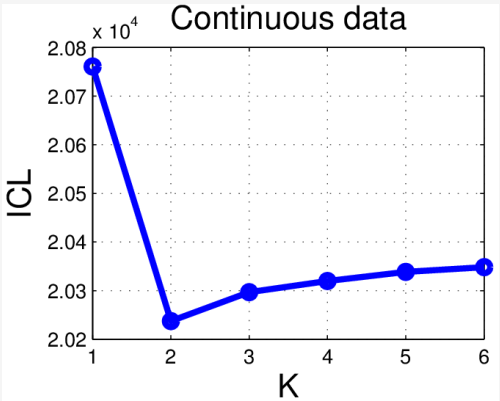
- Data: $n = 2\,061$ event locations in a rectangle of $\mathbb{R}^2$ representing the vessel
- Model: Diagonal Gaussian mixture + uniform (noise)
- Groups: sound locations = vessel defects

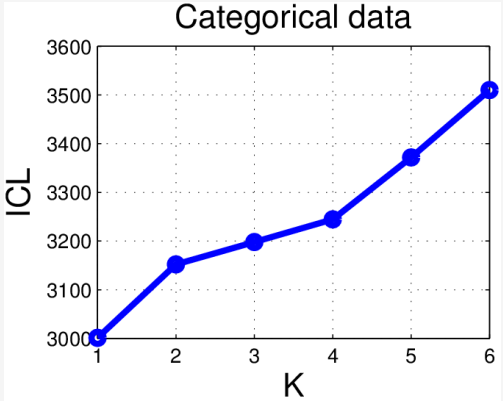


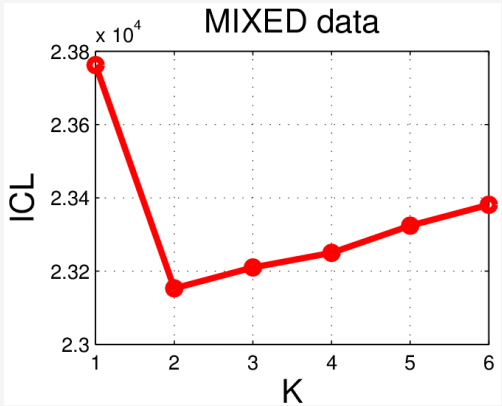


ICL for prostate cancer data (1/2)

- Individuals: $n = 475$ patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease
- Variables: $d = 12$ pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases)
- Model: cond. indep. $\mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k) = \mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k^{cont}) \cdot \mathrm{p}(\mathbf{x}_1; \boldsymbol{\alpha}_k^{cat})$

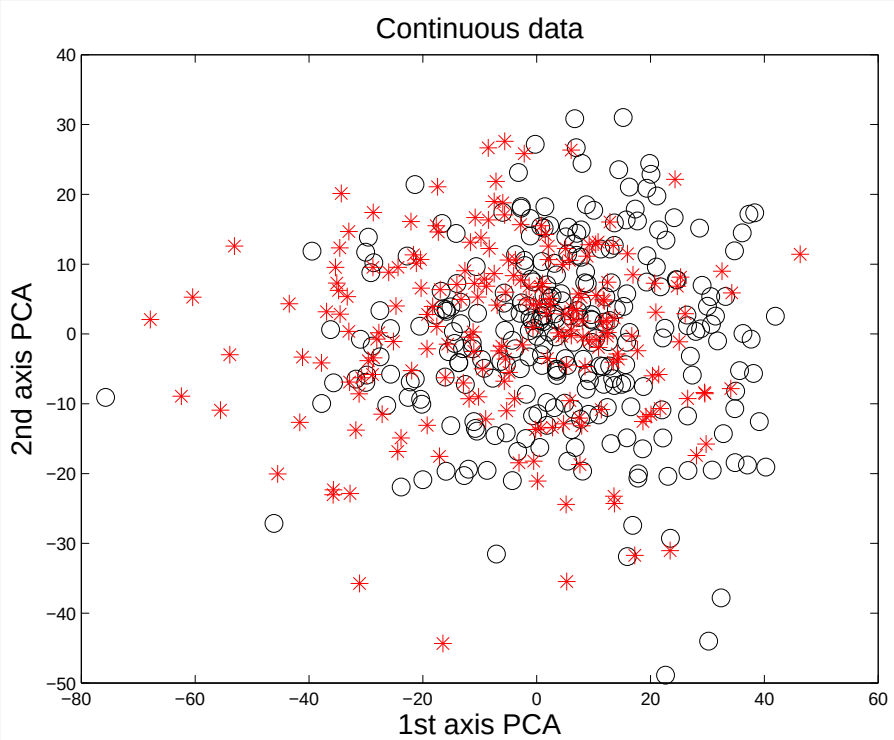


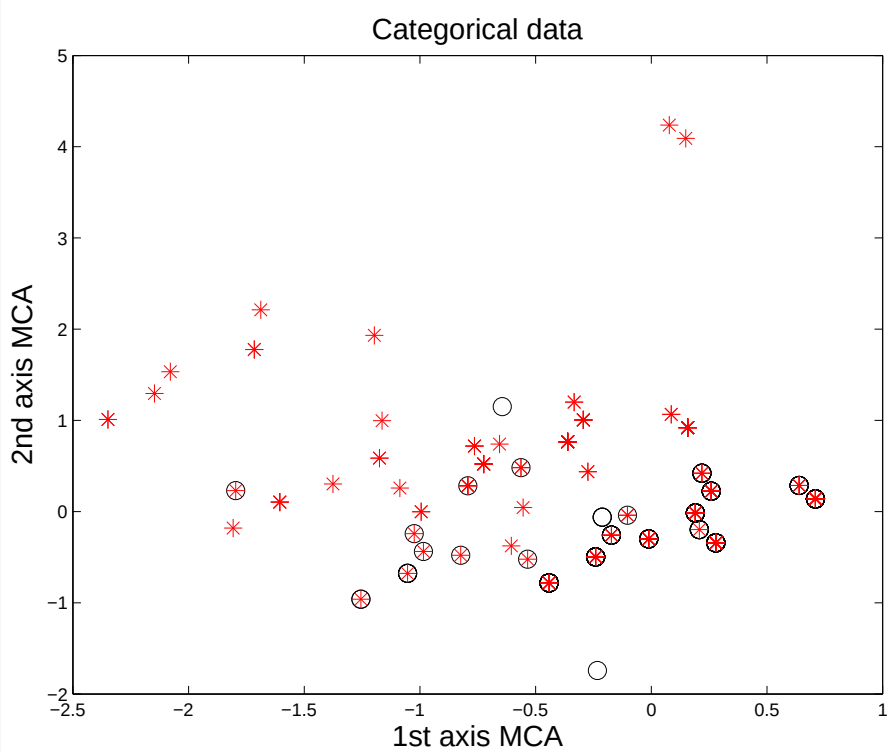




ICL for prostate cancer data (2/2)

Variables	Continuous		Categorical		Mixed	
Error (%)	9.46		47.16		8.63	
True \ estimated group	1	2	1	2	1	2
Stage 3	247	26	142	131	252	21
Stage 4	19	183	120	82	20	182

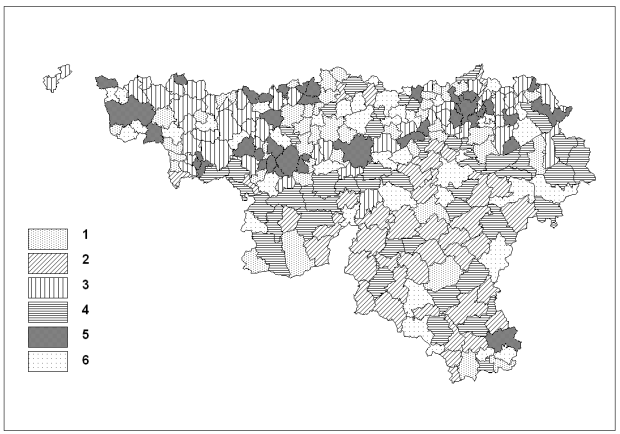


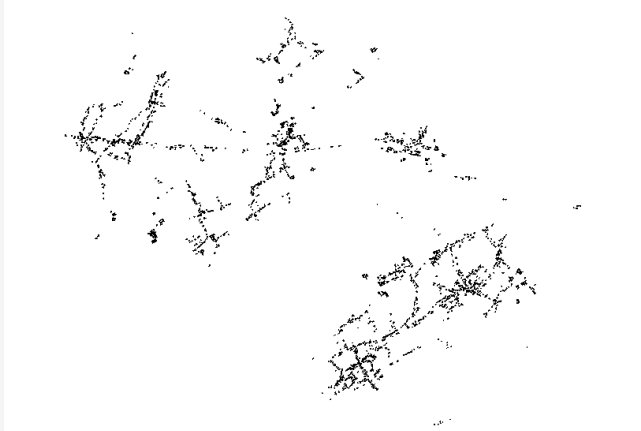


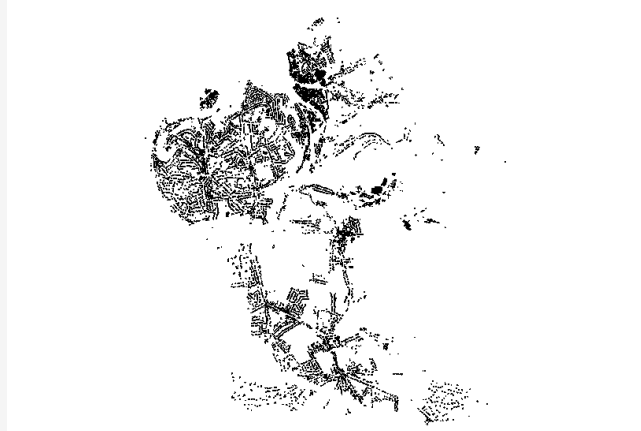
BIC for partitioning communes of Wallonia

- Data: $n = 262$ communes of Wallonia in terms of $d = 2$ fractals at a local level
- Model:
  - Data unit: one to one transformation $\mathbf{g}(\mathbf{x}) = (g(x_i^j), i = 1, \ldots, n\ j = 1, \ldots, d)$ of the initial data set. Typically, standard transformations are $g(x_i^j) = x_i^j$ (identity), $g(x_i^j) = \exp(x_i^j)$ or $g(x_i^j) = \ln(x_i^j)$
  - Mixture: $K = 6$ (fixed) but all 28 Gaussian models
  - Result: 6 meaningful groups with $g(x_i^j) = \exp(x_i^j)$ (natural for fractals...)
- Model criterion:

$$\text{BIC}_{\mathbf{g}} = \ell(\hat{\boldsymbol{\theta}}_{\mathbf{g}}; \mathbf{g}(\mathbf{x})) - \frac{\nu}{2} \ln n + \ln |\underbrace{\mathbf{H}_{\mathbf{g}}}_{\text{Jacobian}}|.$$







BIC for Gaussian "variable selection"[10]

Definition

$$p(\mathbf{x}_1;\boldsymbol{\theta}) = \underbrace{\left\{\sum_{k=1}^{K}\pi_k p(\mathbf{x}_1^S;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}}_{\text{clustering variables}} \times \underbrace{\left\{p(\mathbf{x}_1^U;\mathbf{a}+\mathbf{x}_1^R\mathbf{b},\mathbf{C})\right\}}_{\text{redundant variables}} \times \underbrace{\left\{p(\mathbf{x}_1^W;\mathbf{u},\mathbf{V})\right\}}_{\text{independent variables}}$$

where

- all parts are Gaussians
- S: set of variables useful for clustering
- U: set of redondant clustering variables, expressed with $R \subseteq S$
- W: set of variables independent of clustering

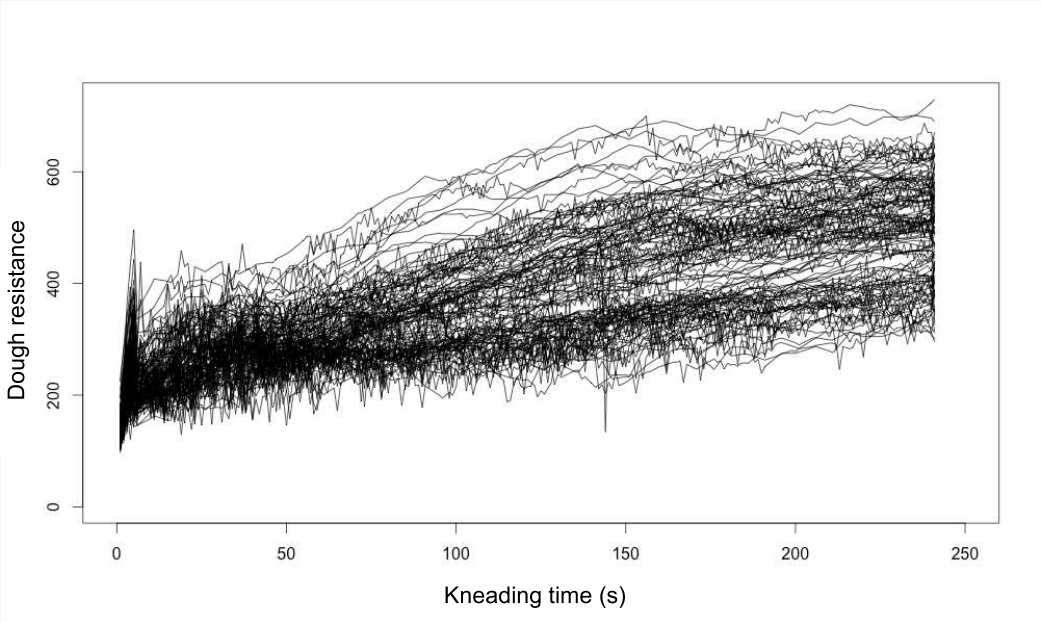
Trick

Variable selection is recasted as a particular model selected by BIC

[10] Raftery and Dean (2006), Maugis *et al.* (09a), Maugis *et al.* (09b)

Curve "cookies" example (1/2)

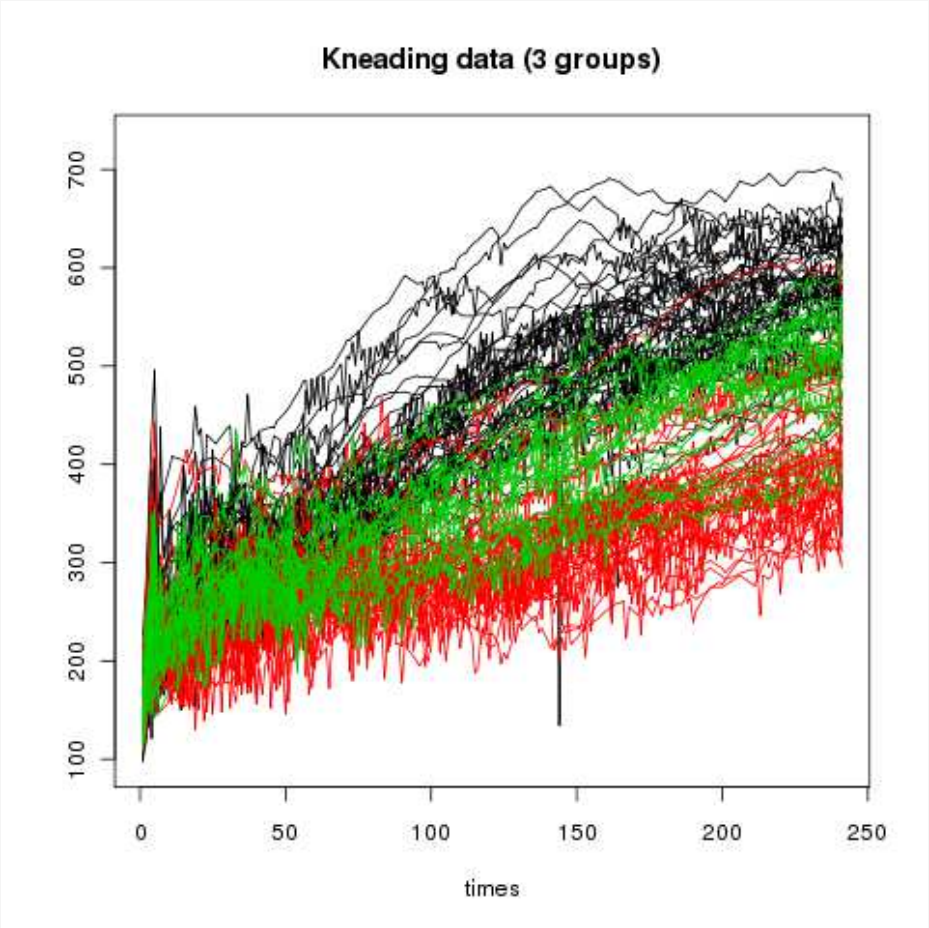
The Kneading dataset comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process[11]. There are 115 different flours for which the dough resistance is measured during the kneading process for 480 seconds. One obtains 115 kneading curves observed at 241 equispaced instants of time in the interval [0; 480]. The 115 flours produce cookies of different quality: 50 of them have produced cookies of good quality, 25 produced medium quality and 40 low quality.



[11] Lévéder *et al* (2004)

Curve "cookies" example (2/2)

Using a basis functional model-based design for functional data[12]



[12] Jacques and Preda (2013)

Co-clustering (1/2)

Contingency table: document clustering

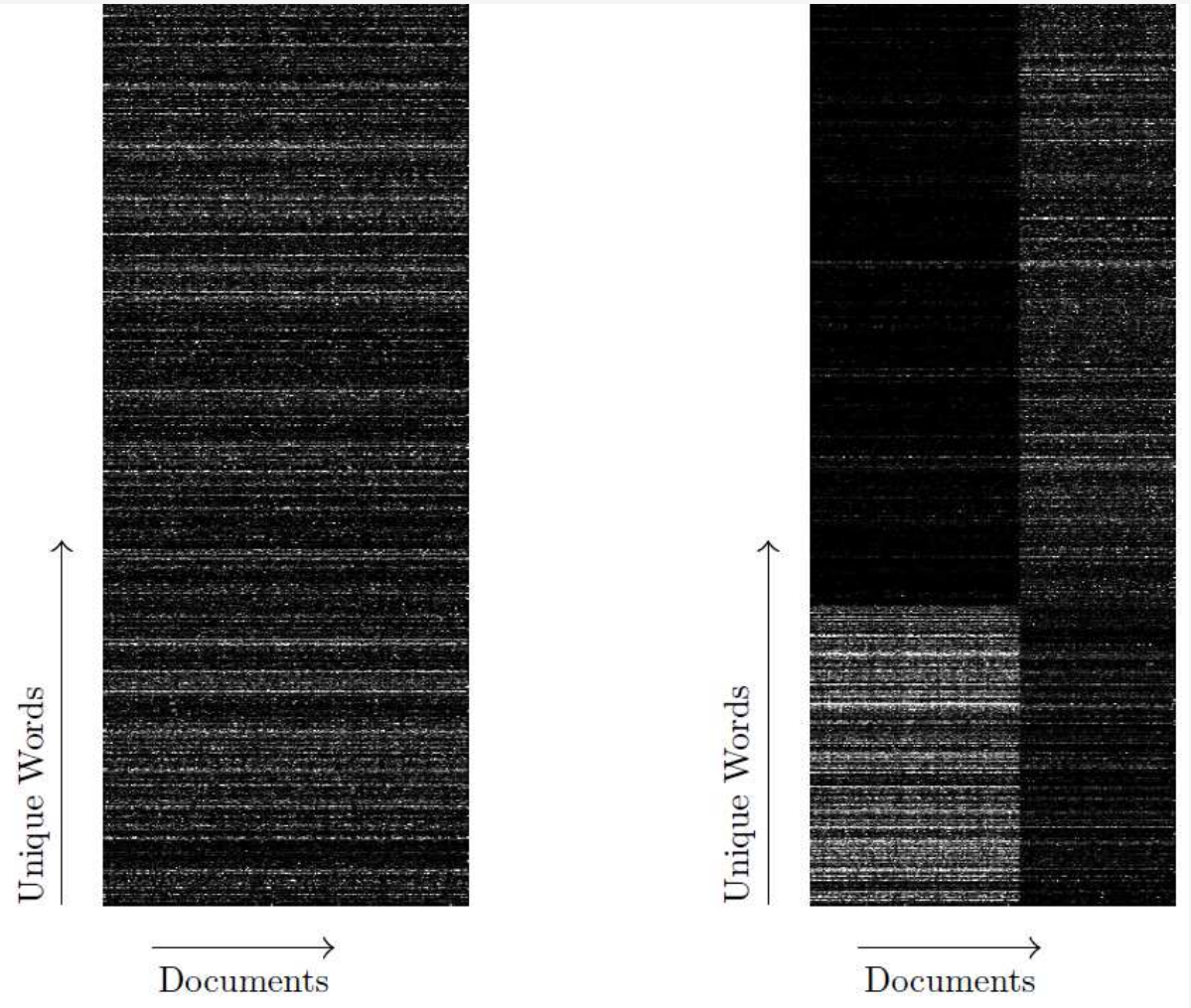
- Mixture of Medline (1033 medical summaries) and Cranfield (1398 aeronautics summaries)
- Lines: 2431 documents
- Columns: present words (except stop), thus 9275 unique words
- Data matrix: cross counting document×words
- Poisson model[13]

Résults with 2×2 blocs

	Medline	Cranfield
Medline	1033	0
Cranfield	0	1398

[13] G. Govaert and M. Nadif (2013). Co-clustering. Wiley.

Co-clustering (2/2)



Outline

Three kinds of challenges, linked to the user task

- Model design: depends on data, should incorporate user information
- Model estimation: find efficient algorithms along the user requirement
- Model selection (validation): depends again on the user purpose

Model-based clustering

- It is a just a confortable and rigorous framework
- The user keep its freedom in this word, because high flexibility at each level