Introduction to cluster analysis and classification: Performing clustering

Christophe Biernacki

Introduction to cluster analysis and classification: **Performing clustering**

C. Biernacki

Summer School on Clustering, Data Analysis and Visualization of Complex Data
May 21-25 2018, University of Catania, Italy

Preamble

What is this course?

- Be able to perform practically some methods
- Use them with discernment

What is not this course?

- Not an exhaustive list of clustering methods (and related bibliography)
- Do not make specialists of clustering methods

This preamble is valid for all four lessons:

1. Performing clustering
2. Validating clustering
3. Formalizing clustering
4. Bi-clustering and co-clustering

Lectures

- General overview of data mining (contain some pretreatments before clustering):
  Gérard Govaert et al. (2009). Data Analysis. Wiley-ISTE, ISBN: 978-1-848-21098-1.
  https://www.wiley.com/en-fr/Data+Analysis-p-9781848210981
- More advanced material on clustering:
  - Christian Hennig, Marina Meila, Fionn Murtagh, Roberto Rocci (2015). Handbook of Cluster Analysis. Chapman and Hall/CRC, ISBN 9781466551886, Series: Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

    https://www.crcpress.com/Handbook-of-Cluster-Analysis/Hennig-Meila-Murtagh-Rocci/p/book/9781466551886
  - Christophe Biernacki. Mixture models. J-J. Droesbeke; G. Saporta; C. Thomas-Agnan. Choix de modèles et agrégation, Technip, 2017.

    https://hal.inria.fr/hal-01252671/document
  - Christophe Biernacki, Cathy Maugis. High-dimensional clustering. Choix de modèles et agrégation, Sous la direction de J-J. DROESBEKE, G. SAPORTA, C. THOMAS-AGNAN Edition: Technip.

    https://hal.archives-ouvertes.fr/hal-01252673v2/document
- Advanced material on co-clustering:
  Gérard Govaert, Mohamed Nadif (2013). Co-Clustering: Models, Algorithms and Applications. Wiley-ISTE, ISBN-13: 978-1848214736.
  https://www.wiley.com/en-fr/Co+Clustering:+Models,+Algorithms+and+Applications-p-9781848214736
- Basic to more advanced R book: Pierre-Andre Cornillon, Arnaud Guyader, Francois Husson, Nicolas Jegou, Julie Josse, Maela Kloareg, Eric Matzner-Lober, Laurent Rouvière (2012). R for Statistics. Chapman and Hall/CRC, ISBN 9781439881453.
  https://www.crcpress.com/R-for-Statistics/
  Cornillon-Guyader-Husson-Jegou-Josse-Kloareg-Matzner-Lober-Rouviere/p/book/9781439881453

Take home message

cluster

clustering

define both!

Outline

Everything begins from data!

Today's data (1/2)

Today, it is easy to collect many features, so it favors

- data variety and/or mixed
- data missing
- data uncertainty (or interval data)

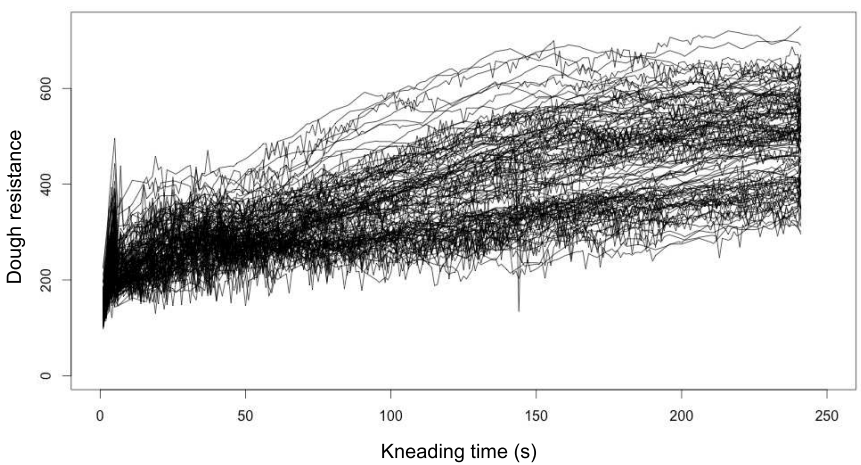
Mixed, missing, uncertain

Observed individuals $\boldsymbol{x}^O$			
?	0.5	?	5
0.3	0.1	green	3
0.3	0.6	{red,green}	3
0.9	[0.25 0.45]	red	?
↓	↓	↓	↓
continuous	continuous	categorical	integer

Today's data (2/2)

And also

- Ranking data (like Olympics games)
- Directional data (like angle of wind direction)
- Ordinal data (a score like $A > B > C$)
- Functional data (like time series)[1]
- Graphical data (like social networks, biological network)
- ...



[1] See a specific lesson for clustering times series

Todays features: full mixed/missing

categorical
Marital status
maried

integer
Children
3

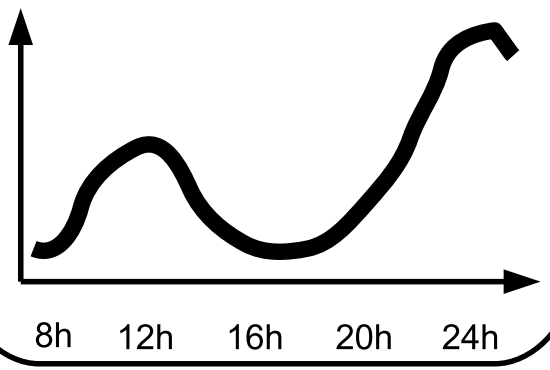
missing
Size (m)
?

rank
Drink preference
beer > soda > water

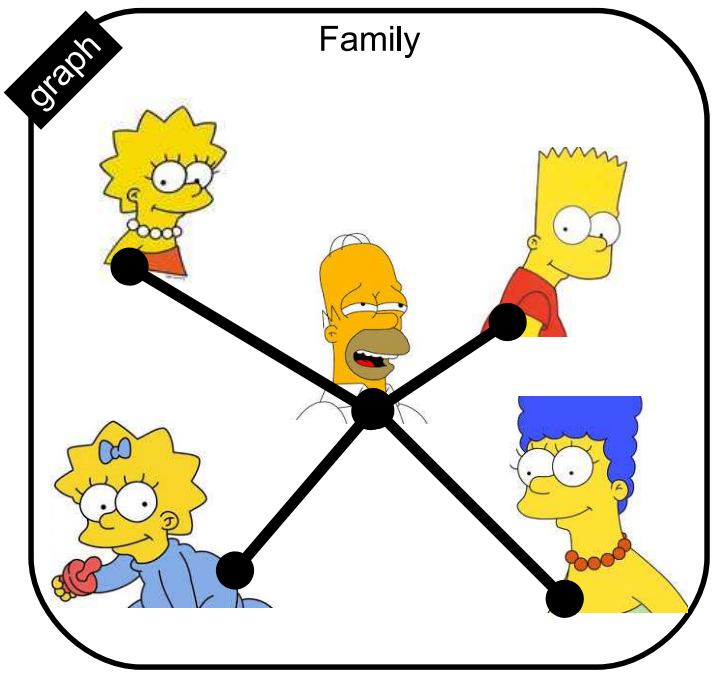
ordinal
Intelligence
low

continuous
Weight (kg)
119.5

functional

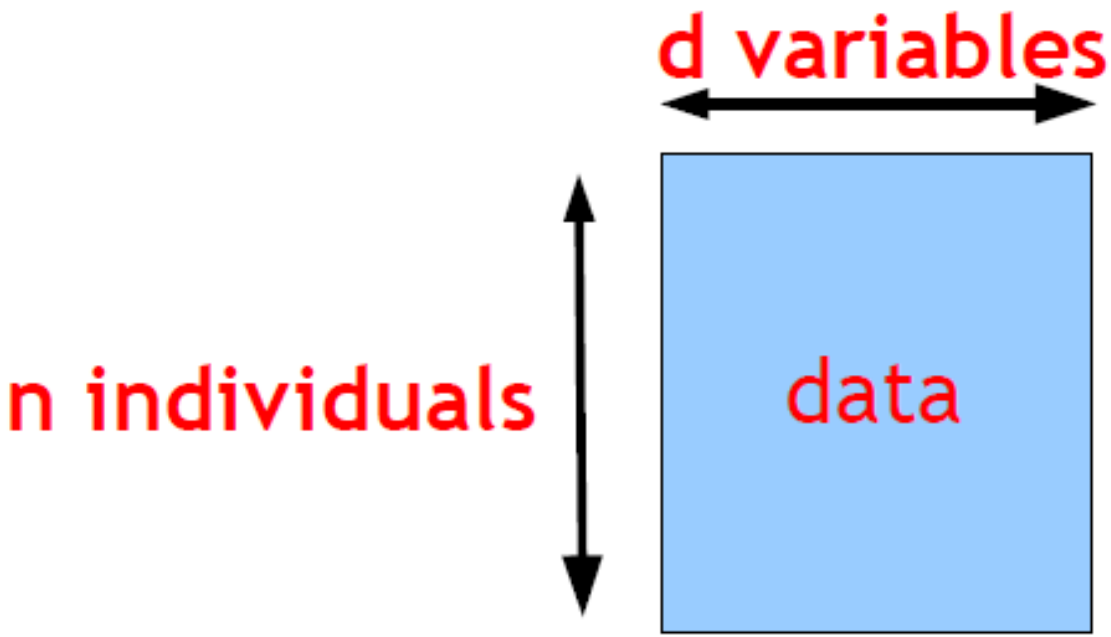


graph



And so on...

Data sets structure



Coding for data $\boldsymbol{x}$

- A set of n individuals

$$\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$$

with $\boldsymbol{x}_i$ a set of (possibly non-scalar) d variables

$$\boldsymbol{x}_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{id}\}$$

where $\boldsymbol{x}_{ij} \in \mathcal{X}_j$

- A n-uplet of individuals

$$\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

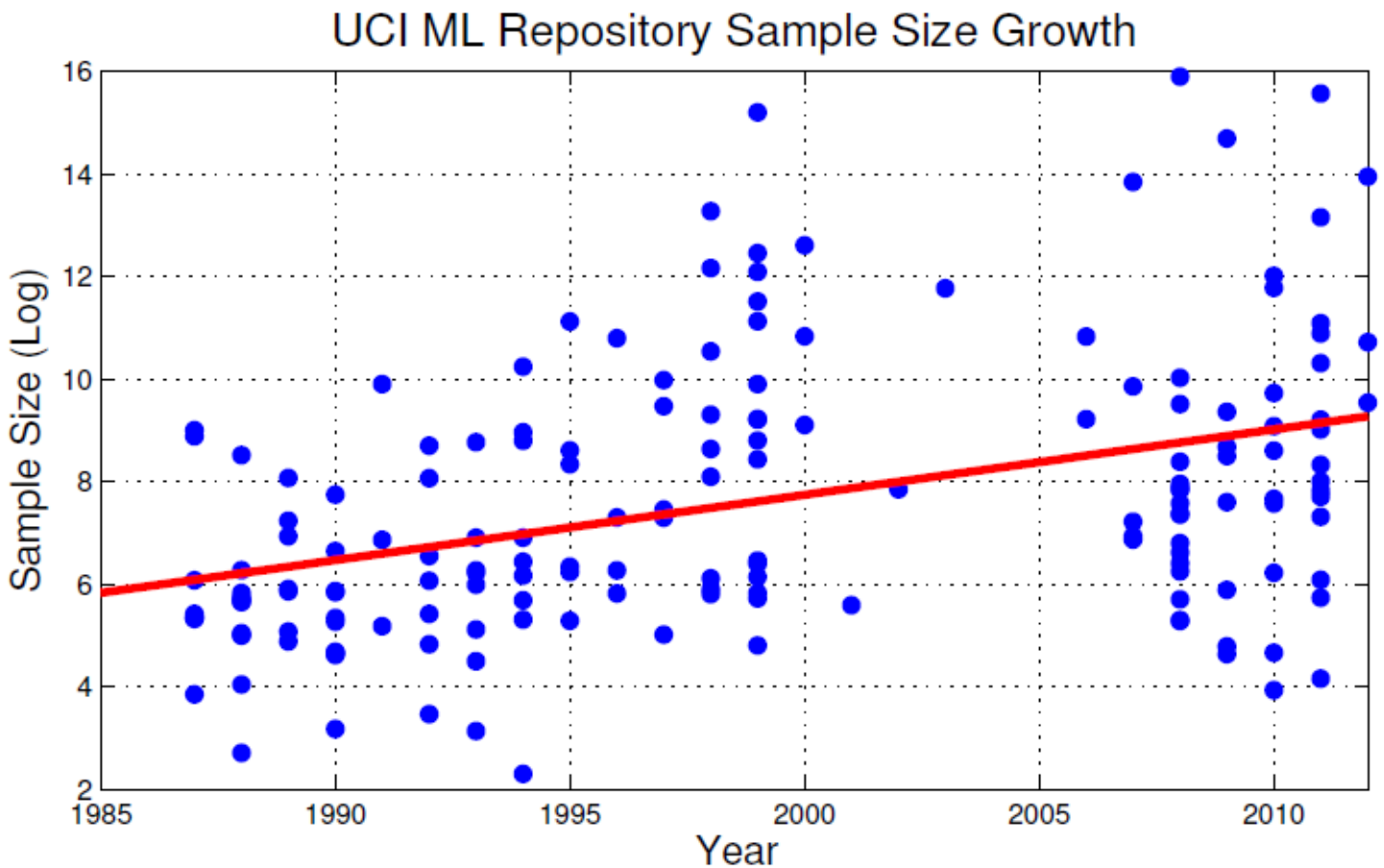
with $\boldsymbol{x}_i$ a d-uplet of (possibly non-scalar) variables

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{id}) \in \mathcal{X}$$

where $\mathcal{X} = \mathcal{X}_1 \times \ldots\ldots \mathcal{X}_d$

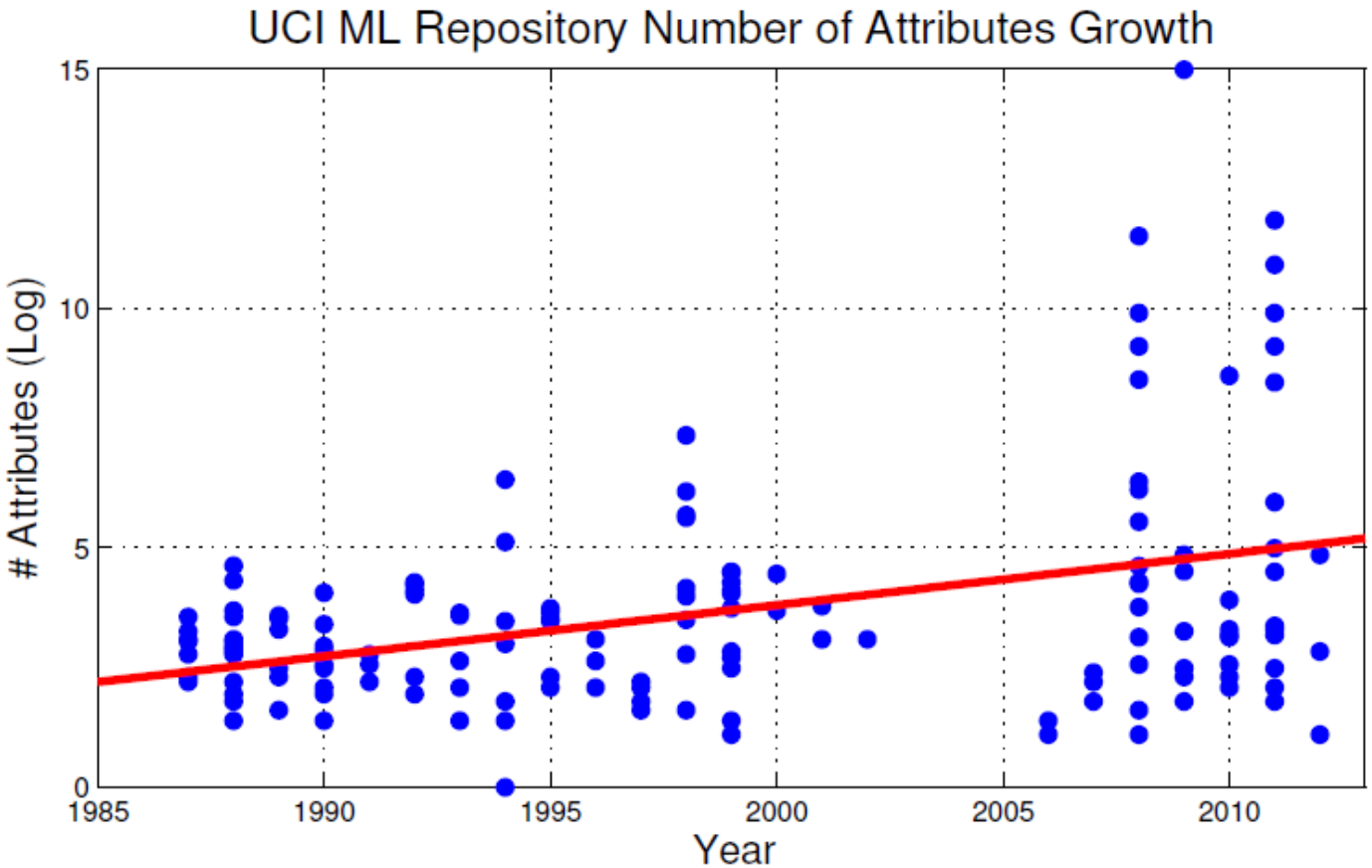
We will pass from a coding to another, depending of the practical utility (useful for some calculus to have matrices or vectors for instance)

Large data sets $(n)^2$



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

High-dimensional data (d)[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"

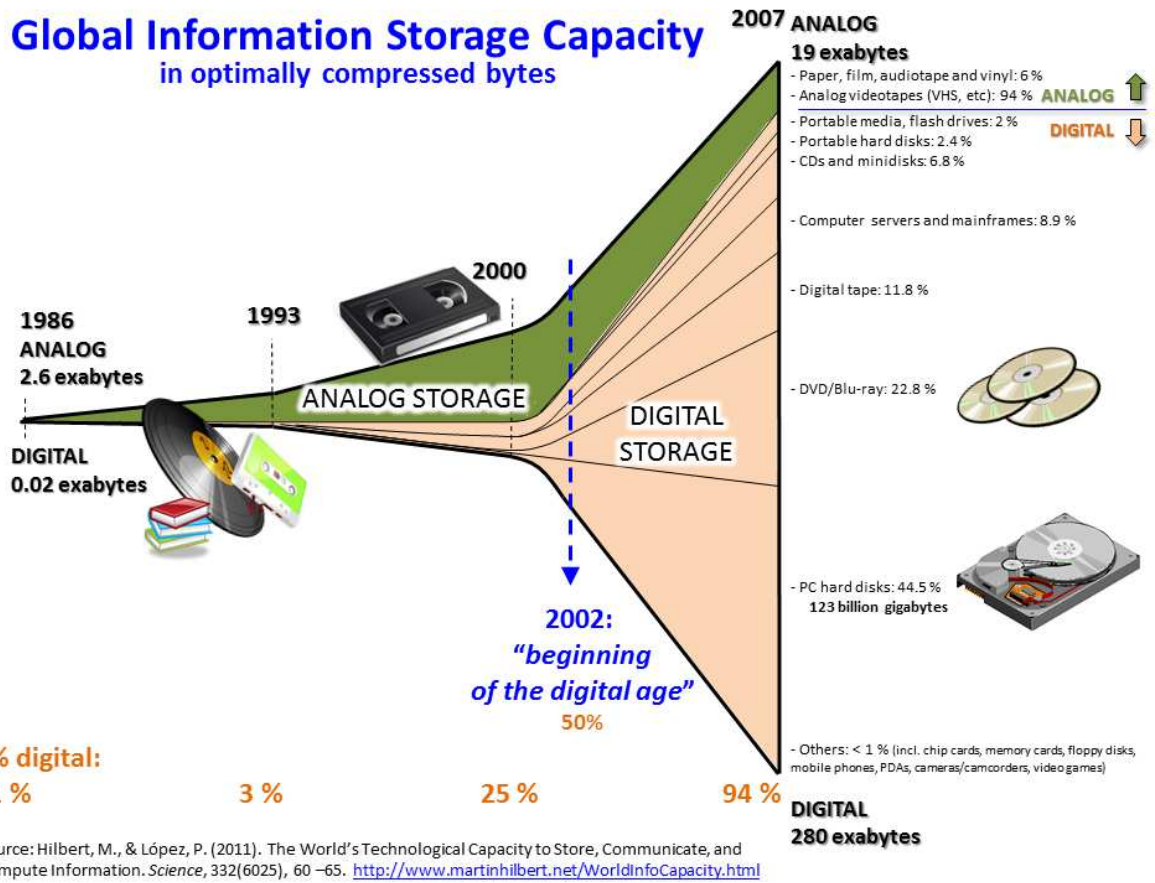
The Big Data phenomenon mainly originates in the increase of computer and digital resources at an ever lower cost

- 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)



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

[5] Exabyte

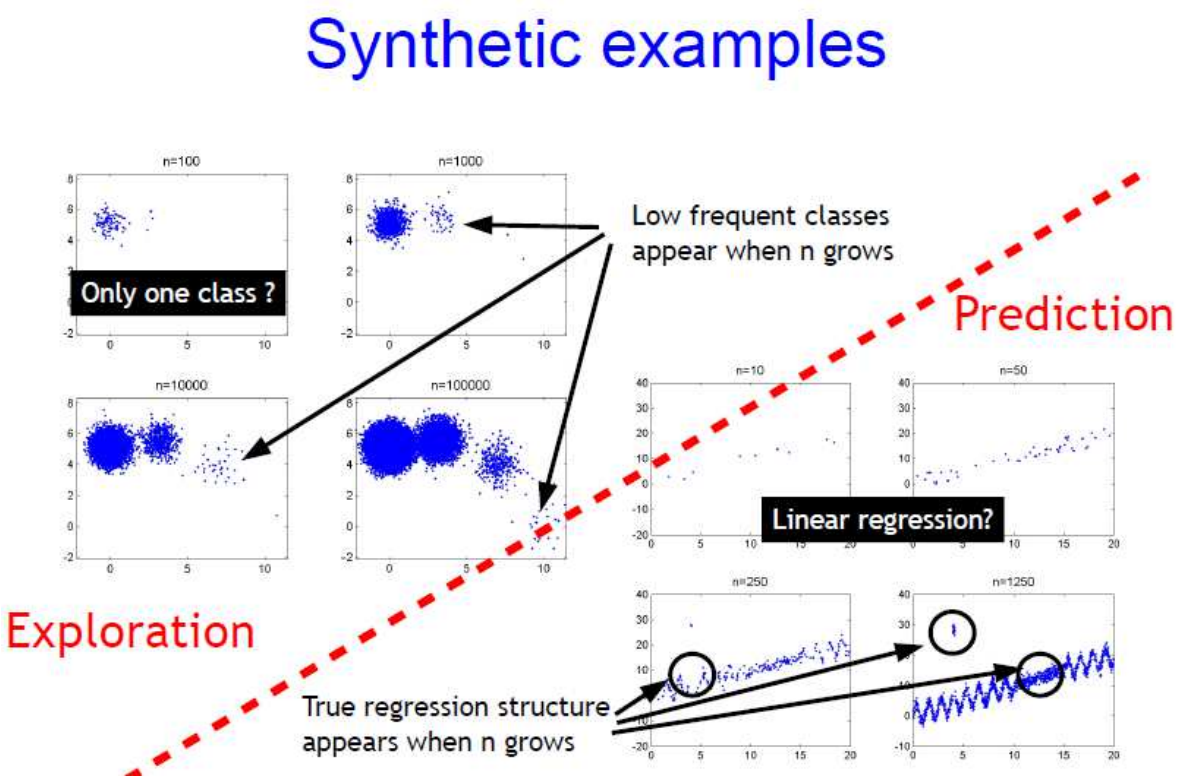
Societal phenomenon

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. . .

More data for what?

Opportunity to improve accuracy of traditional questionings



- Here is just illustrated the effect of n
- In further lessons will be illustrated the effect of d (be patient)

Outline

Supervised classification (1/3)

- Data: learning dataset $\mathcal{D} = \{\mathbf{x}^O, \mathbf{z}\}$
  - n individuals: $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \{\mathbf{x}^O, \mathbf{x}^M\}$, each $x_i \in \mathcal{X}$
  - Observed individuals x^O
  - Missing individuals x^M
  - Partition in K groups $G_1, \ldots, G_K$: $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$[6], $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$

$$\mathbf{x}_i \in G_k \quad \Leftrightarrow \quad z_{ih} = \mathbb{I}_{\{h=k\}}$$

- Aim: estimation of an allocation rule r from $\mathcal{D}$

$$\begin{array}{rccc} r: & \mathcal{X} & \longrightarrow & \{1, \ldots, K\} \\ & \mathbf{x}^O_{n+1} & \longmapsto & r(\mathbf{x}^O_{n+1}). \end{array}$$

Other possible coding for z

Sometimes it can be more convenient in formula to code $\mathbf{z} = (z_1, \ldots, z_n)$[a] where

$$\mathbf{x}_i \in G_k \quad \Leftrightarrow \quad z_i = k$$

[a] Sometimes it would be more convenient to use a set also: $\mathbf{z} = \{z_1, \ldots, z_n\}$

[6] Sometimes it would be more convenient to use a set also: $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$

Supervised classification (2/3)

1st coding for $\boldsymbol{z}$

Mixed, missing, uncertain

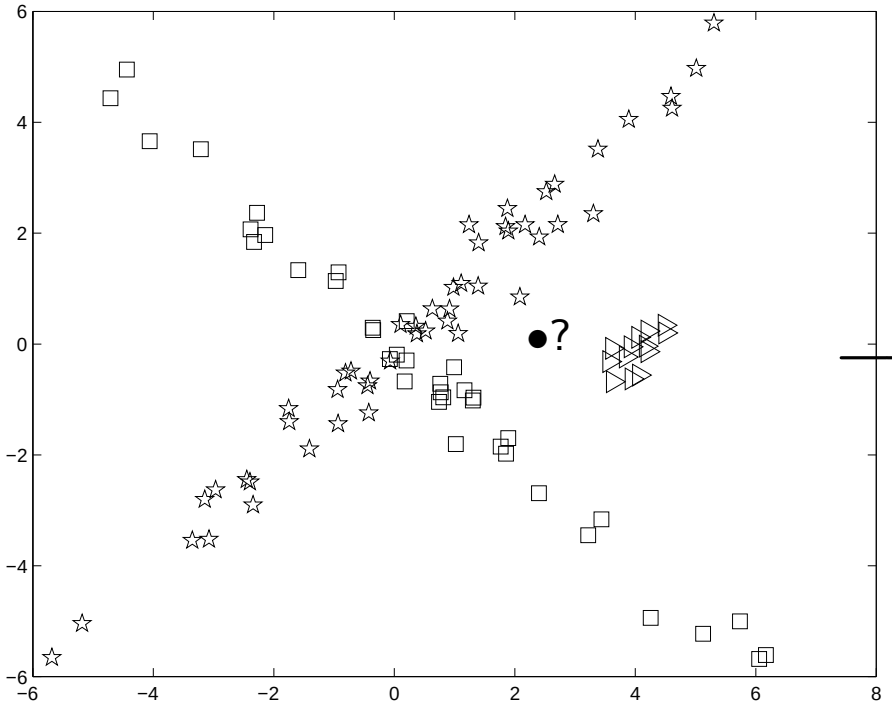
Individuals $\mathbf{x}^O$				Partition $\mathbf{z}$			$\Leftrightarrow$	Group
?	0.5	red	5	0	1	0	$\Leftrightarrow$	G_2
0.3	0.1	green	3	1	0	0	$\Leftrightarrow$	G_1
0.3	0.6	{red,green}	3	1	0	0	$\Leftrightarrow$	G_1
0.9	[0.25 0.45]	red	?	0	0	1	$\Leftrightarrow$	G_3
↓	↓	↓	↓					
continuous	continuous	categorical	integer					

2nd coding for $\boldsymbol{z}$

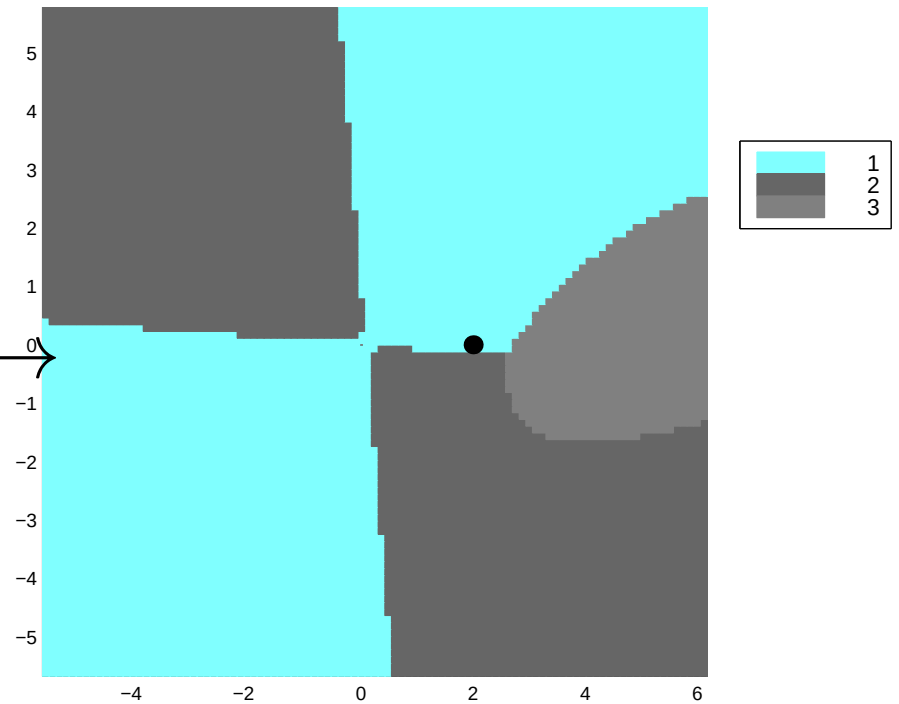
Mixed, missing, uncertain

Individuals $\mathbf{x}^O$				Partition $\mathbf{z}$	$\Leftrightarrow$	Group
?	0.5	red	5	2	$\Leftrightarrow$	G_2
0.3	0.1	green	3	1	$\Leftrightarrow$	G_1
0.3	0.6	{red,green}	3	1	$\Leftrightarrow$	G_1
0.9	[0.25 0.45]	red	?	3	$\Leftrightarrow$	G_3
↓	↓	↓	↓			
continuous	continuous	categorical	integer			

Supervised classification (3/3)



$\{\mathbf{x}^O, \mathbf{z}\}$ and $\mathbf{x}^O_{n+1}$



$\hat{r}$ and $\hat{\mathbf{z}}_{n+1}$

Semi-supervised classification (1/3)

- Data: learning dataset $\mathcal{D} = \{\mathbf{x}^O, \mathbf{z}^O\}$
  - n individuals: $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \{\mathbf{x}^O, \mathbf{x}^M\}$ belonging to a space $\mathcal{X}$
  - Observed individuals $\boldsymbol{x}^O$
  - Missing individuals $\boldsymbol{x}^M$
  - Partition: $\mathbf{z} = \{\boldsymbol{z}_1, \ldots, \mathbf{z}_n\} = \{\mathbf{z}^O, \mathbf{z}^M\}$
  - Observed partition $\boldsymbol{z}^O$
  - Missing partition $\boldsymbol{z}^M$
- Aim: estimation of an allocation rule r from $\mathcal{D}$

$$\begin{array}{rccc} r: & \mathcal{X} & \longrightarrow & \{1, \ldots, K\} \\ & \mathbf{x}^O_{n+1} & \longmapsto & r(\mathbf{x}^O_{n+1}). \end{array}$$

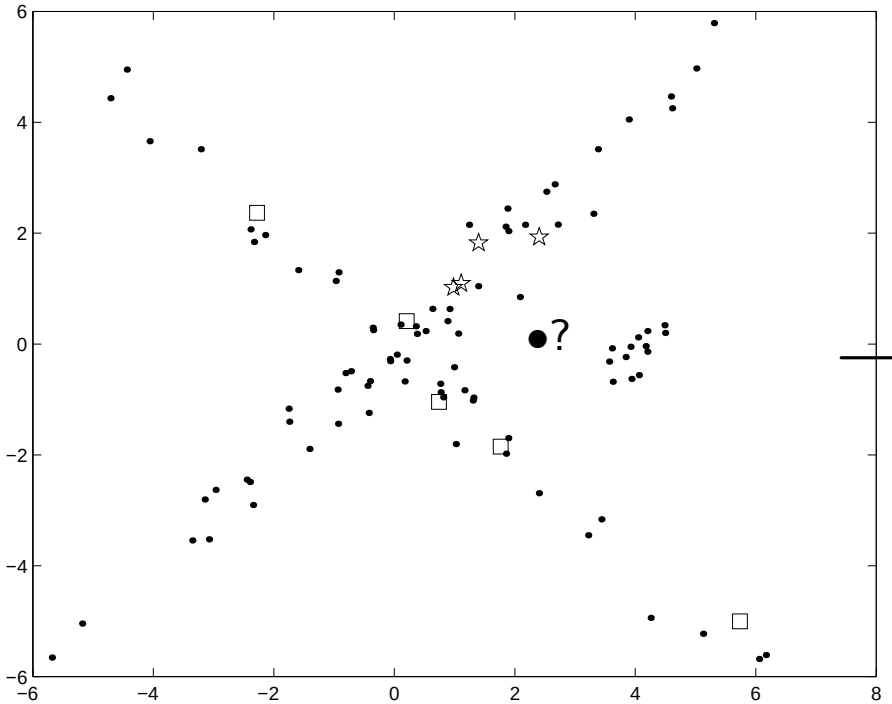
- Idea: $\mathbf{x}$ is cheaper than $\mathbf{z}$ so $\#\mathbf{z}^M \gg \#\mathbf{z}^O$

Semi-supervised classification (2/3)

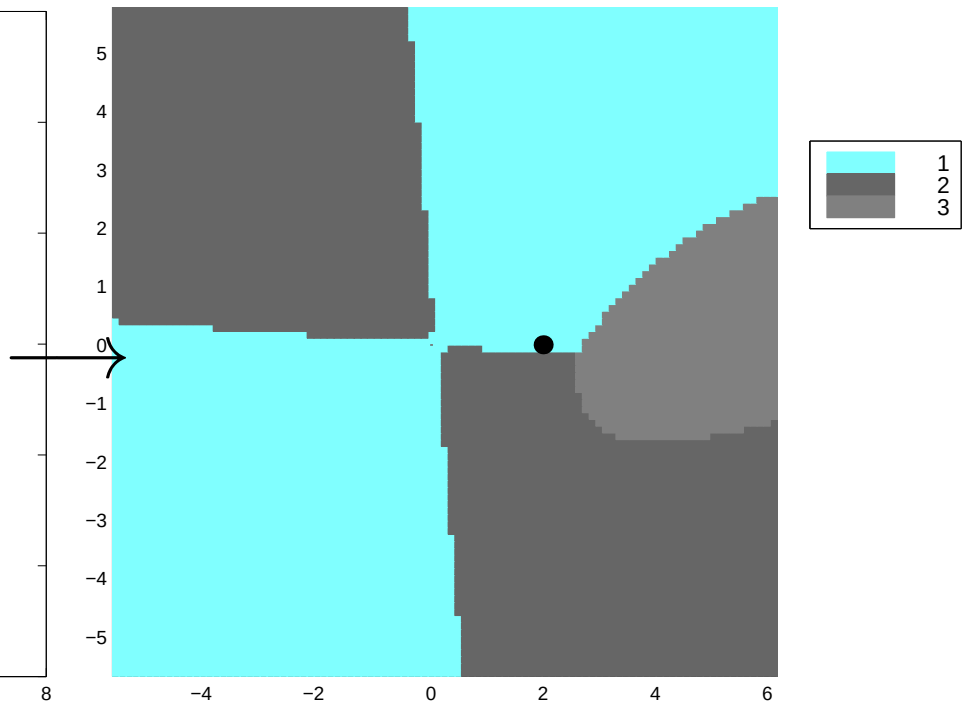
Mixed, missing, uncertain

Individuals $\mathbf{x}^O$				Partition $\mathbf{z}^O$			$\Leftrightarrow$	Group
?	0.5	red	5	0	**?**	**?**	$\Leftrightarrow$	G_2 or G_3
0.3	0.1	green	3	1	0	0	$\Leftrightarrow$	G_1
0.3	0.6	{red,green}	3	**?**	**?**	**?**	$\Leftrightarrow$	???
0.9	[0.25 0.45]	red	?	0	0	1	$\Leftrightarrow$	G_3
↓	↓	↓	↓					
continuous	continuous	categorical	integer					

Semi-supervised classification (3/3)



$\{\mathbf{x}^O, \mathbf{z}^O\}$ and $\mathbf{x}^O_{n+1}$



$\hat{r}$ and $\hat{\mathbf{z}}_{n+1}$

Unsupervised classification (1/3)

- Data: learning dataset $\mathcal{D} = \mathbf{x}^O$, so $\mathbf{z}^O = \emptyset$
- Aim: estimation of the partition $\mathbf{z}$[7] and the number of groups K
- Also known as: clustering

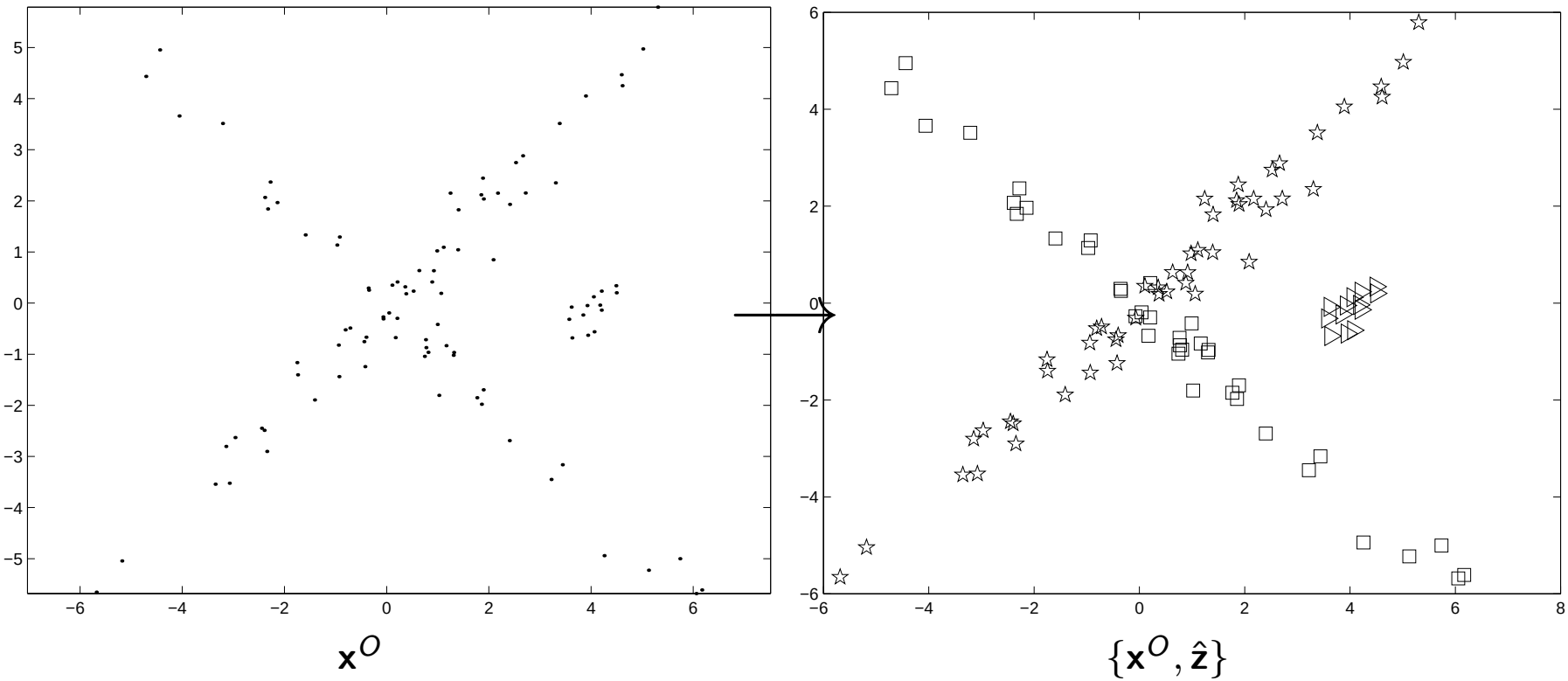
[7] We will see other clustering structures further, here to fix ideas with the main one

Unsupervised classification (2/3)

Mixed, missing, uncertain

Individuals $\mathbf{x}^O$				Partition $\mathbf{z}^O$			$\Leftrightarrow$	Group
?	0.5	red	5	?	?	?	$\Leftrightarrow$	???
0.3	0.1	green	3	?	?	?	$\Leftrightarrow$	???
0.3	0.6	{red,green}	3	?	?	?	$\Leftrightarrow$	???
0.9	[0.25 0.45]	red	?	?	?	?	$\Leftrightarrow$	???
↓	↓	↓	↓					
continuous	continuous	categorical	integer					

Unsupervised classification (3/3)



Traditional solutions (1/3)

Two main model-based frameworks[8]

- Generative models
  - Model $p(x, z)$
  - Thus direct model for $p(x) = \sum_z p(x, z)$
  - Easy to take into account some missing z and x
- Predictive models
  - Model $p(z|x)$ or sometimes $\mathbf{1}_{\{p(z|x)>1/2\}}$ or also ranking on $p(z|x)$
  - Avoid asumptions on $p(x)$, thus avoids associated error model
  - difficult to take into account some missing z and x

[8] Sometimes presented as distance-based instead of model-based: see some links in my 3th lesson

Traditional solutions (2/3)

No mixed, missing or uncertain data:

- Supervised classification[9]
  - Generative models: linear/quadratic discriminant analysis
  - Predictive models: logistic regression, support vector machines (SVM), k nearest neighbourhood, classification trees...
- Semi-supervised classification[10]
  - Generative models: mixture models
  - Predictive models: low density separation (transductive SVM), graph-based methods...
- Unsupervised classification[11]
  - Generative models: k-means like criteria, hierarchical clustering, mixture models
  - Predictive models: -

[9] Govaert *et al.*, Data Analysis, Chap.6, 2009

[10] Chapelle *et al.*, Semi-supervised learning, 2006

[11] Govaert *et al.*, Data Analysis, Chap.7-9, 2009

Traditional solutions (3/3)

But more complex with mixed, missing or uncertain data...

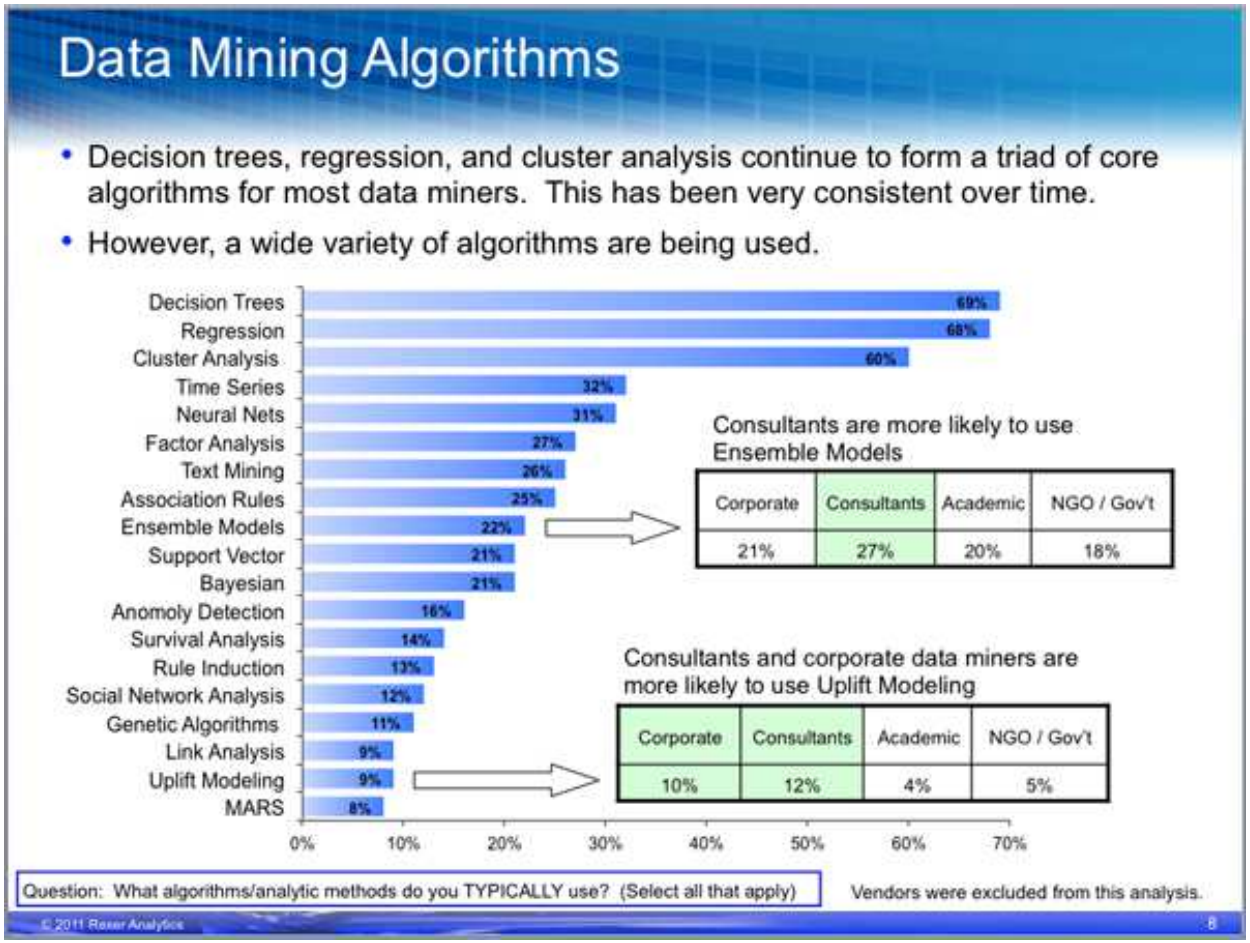
- Missing/uncertain data: multiple imputation is possible but it should ideally take into account the classification purpose at hand
- Mixed data: some heuristic methods with recoding

We will gradually answer along the lessons:

How to marry the classification aim with mixed, missing or uncertain data?

Outline

Clustering everywhere[12]

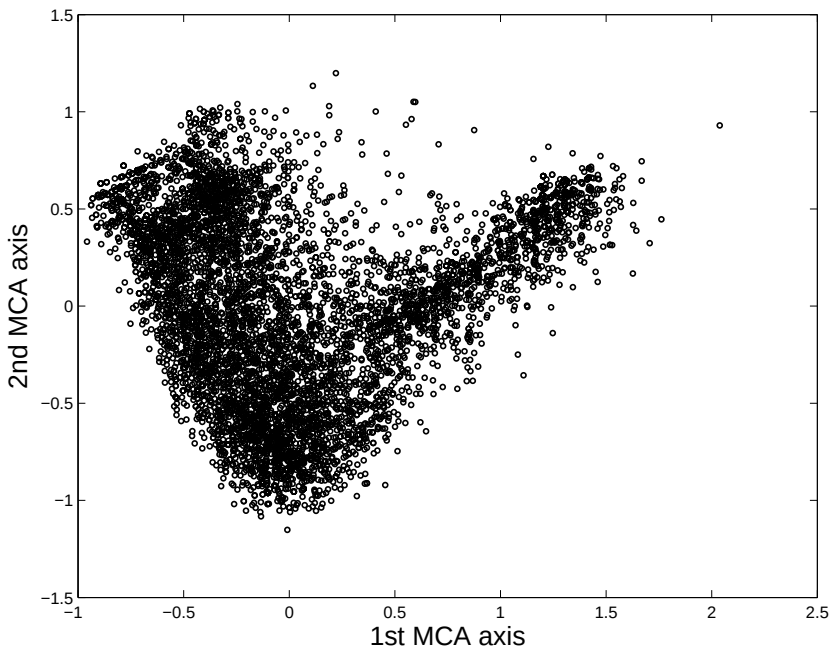


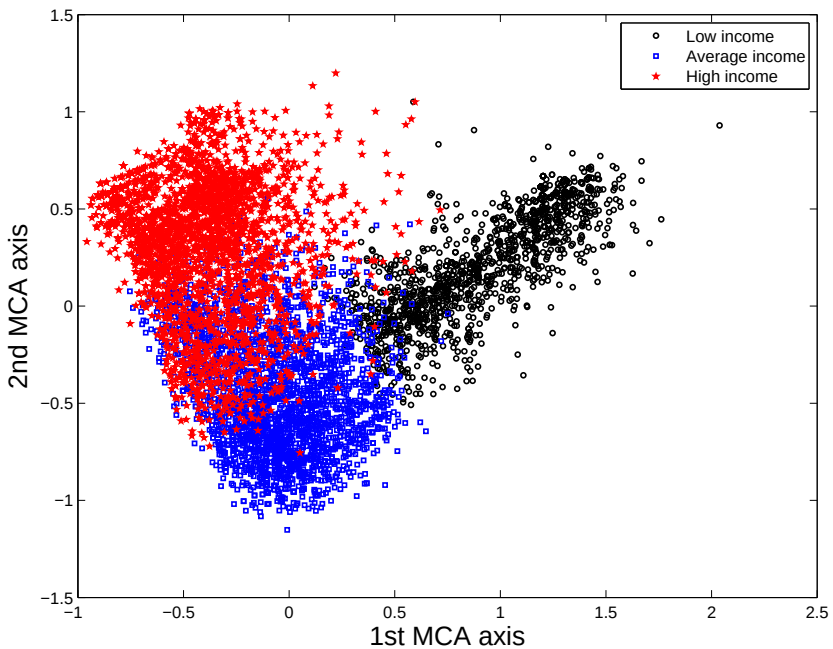
Why such a success despite its complexity relatively to other classification purposes?

[12] 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)

A 1st aim: explanatory task

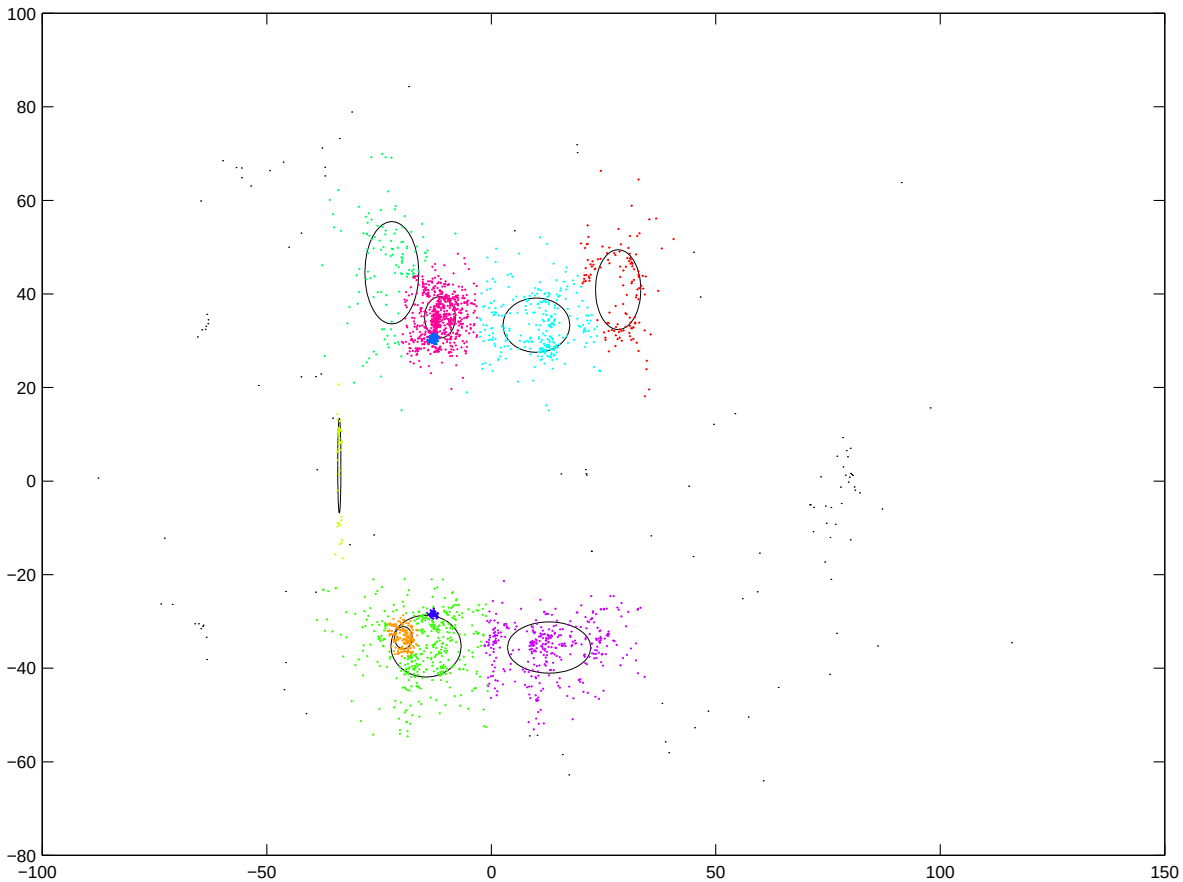
- A clustering for a marketing study
- Data: $d = 13$ demographic attributes (nominal and ordinal variables) of $n = 6\,876$ shopping mall customers in the San Francisco Bay (SEX (1. Male, 2. Female), MARITAL STATUS (1. Married, 2. Living together, not married, 3. Divorced or separated, 4. Widowed, 5. 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")





Another explanatory example: acoustic emission control

- Data: $n = 2\ 061$ event locations in a rectangle of $\mathbb{R}^2$ representing the vessel
- Groups: sound locations = vessel defects



A 2nd aim: preprocessing step (1/2)

A synthetic example in supervised classification: predict two groups (blue and black)[13]

- Logit model:
  - Not very flexible since linear borderline
  - Unbiased ML estimate but asymptotic variance $\sim (\mathbf{x}'\mathbf{w}\mathbf{x})^{-1}$ is influenced by correlations

$$\text{logit}(\text{p}(\text{blue}|\boldsymbol{x}_{n+1})) = \boldsymbol{\beta}'\boldsymbol{x}_{n+1}$$

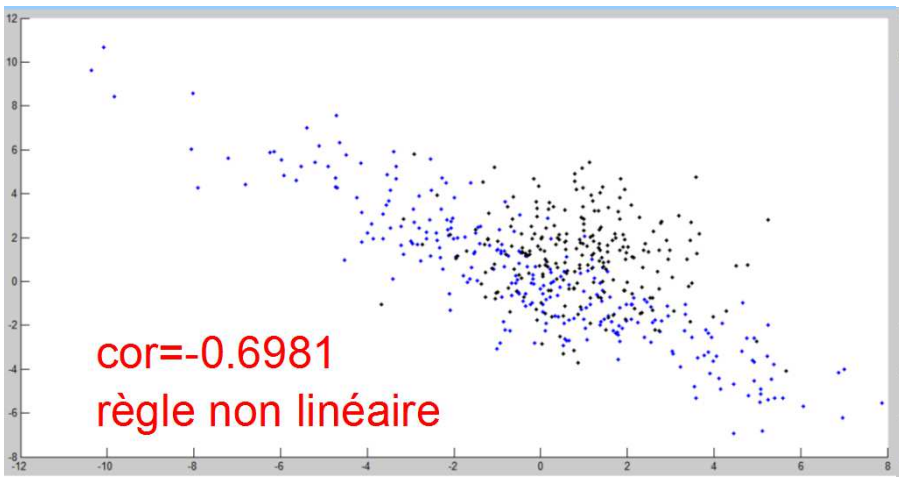
- A clustering may improve logistic regression prediction
  - More flexible borderline: piecewise linear
  - Decrease correlation so decrease variance

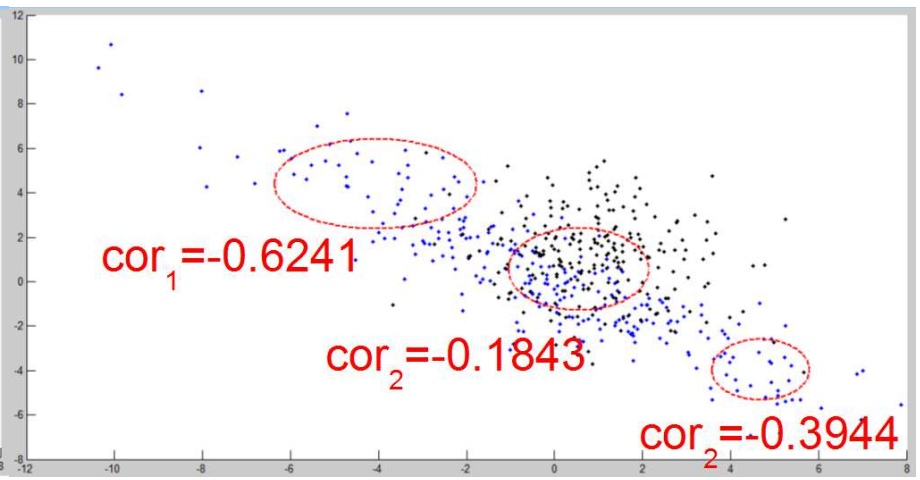
$$\text{logit}(\text{p}(\text{blue}|\boldsymbol{x}_{n+1})) = \boldsymbol{\beta}'_{z_{n+1}}\boldsymbol{x}_{n+1}$$

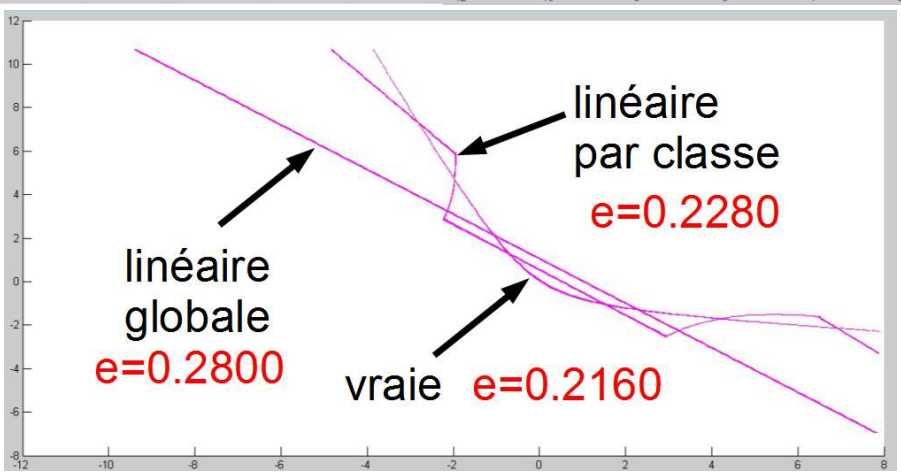
Do not confound groups blue and black and clusters z_{n+1}!!

[13] Another lesson will retain a similar idea through a mixture of regressions

A 2nd aim: preprocessing step (2/2)



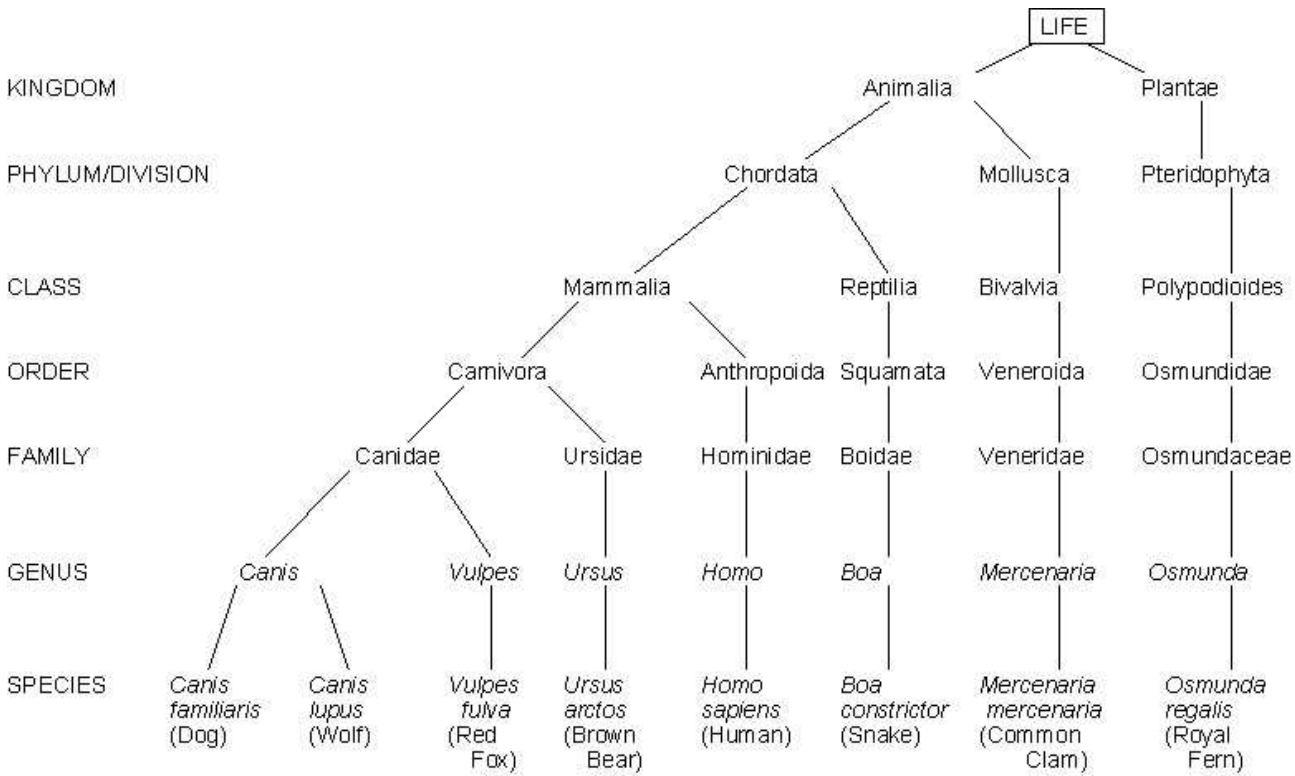




Outline

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)



Remind data we use

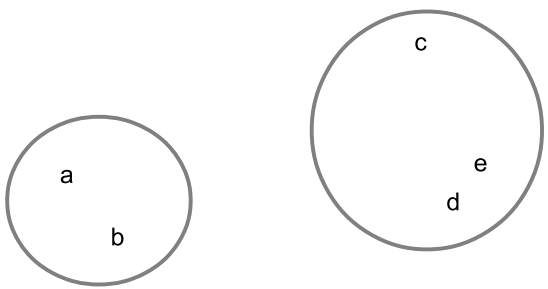
Data set of n individuals $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i$ described by d variables

Four main clustering structures: partition

Each object belongs to exactly one cluster

- Definition: a set of K non-empty parts of $\boldsymbol{x}$: $P = (G_1, \ldots, G_K)$:
  - For all $k \neq k'$, $G_k \cap G_{k'} = \emptyset$
  - $G_1 \cup \ldots \cup G_K = \boldsymbol{x}$
- Example: $\boldsymbol{x} = \{a, b, c, d, e\}$

$$P = \{\underbrace{\{a, b\}}_{G_1}, \underbrace{\{c, d, e\}}_{G_2}\}$$



- Notation: $\boldsymbol{z} = (z_1, \ldots, z_n)$, with $z_i \in \{1, \ldots, K\}$

Four main clustering structures: partition with outliers

Each object belongs to one cluster or less

- Definition: a set of K non-empty parts of $\boldsymbol{x}$: $P = (G_1, \ldots, G_K)$:
  - For all $k \neq k'$, $G_k \cap G_{k'} = \emptyset$
  - $G_1 \cup \ldots \cup G_K \subset \boldsymbol{x}$
- Example: $\boldsymbol{x} = \{a, b, c, d, e\}$

$$P = \{\underbrace{\{a, b\}}_{G_1}, \underbrace{\{d, e\}}_{G_2}\}, \quad c \text{ excluded}$$

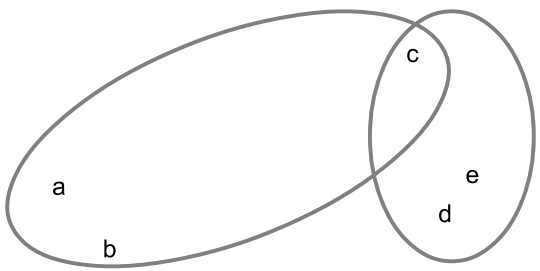
- See more in Lesson 3

Four main clustering structures: multiclass partition

Each object belongs to one cluster or more

- Definition: a set of K non-empty parts of $\boldsymbol{x}$: $P = (G_1, \ldots, G_K)$:
  - $G_1 \cup \ldots \cup G_K = \boldsymbol{x}$
- Example: $\boldsymbol{x} = \{a, b, c, d, e\}$

$$P = \{\underbrace{\{a, b, c\}}_{G_1}, \underbrace{\{c, d, e\}}_{G_2}\}$$



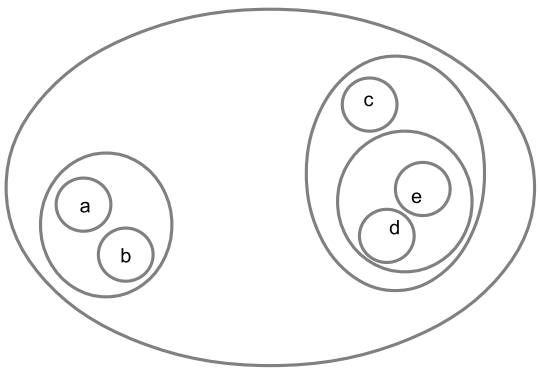
- Not considered in lessons

Four main clustering structures: hierarchy

Each object belongs to nested partitions

- Definition: H is a hierarchy of $\boldsymbol{x}$ if
  - $\boldsymbol{x} \in H$
  - For any $x_i \in \boldsymbol{x}$, $\{x_i\} \in H$
  - For any $G_1, G_2 \in H$, $G_1 \cap G_2 \in \{G_1, G_2, \emptyset\}$
- Example: $\boldsymbol{x} = \{a, b, c, d, e\}$

$$H = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{d, e\}, \{a, b\}, \{c, d, e\}, \{a, b, c, d, e\}\}$$



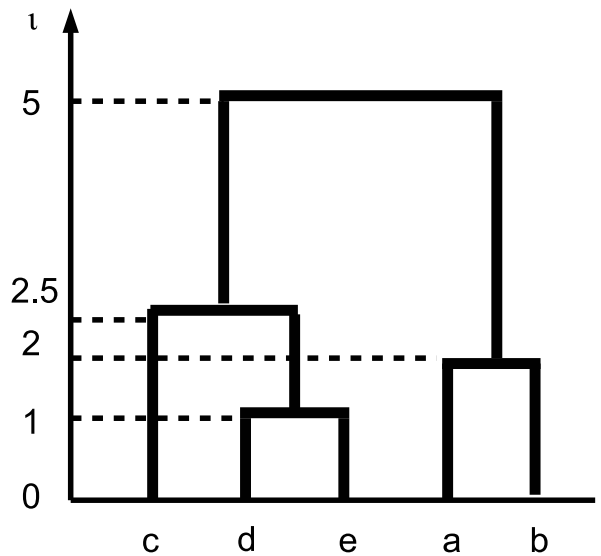
- Case of the species taxonomy

Four main clustering structures: indexed hierarchy

Each element of a hierarchy is associated to an index value

- Definition: $\iota : H \to \mathbb{R}^+$ is an index of a hierarchy H if
  - ι is a non-decreasing mapping: $(G_1, G_2) \in H^2,\ G_1 \subset G_2,\ G_1 \neq G_2 \Rightarrow \iota(G_2) < \iota(G_2)$
  - For all $x_i \in x$, $\iota(\{x_i\}) = 0$
- The couple (H, ι) is an indexed hierarchy and can be visualized by a dendrogram
- Example (continuated):

$$H = \{\underbrace{\{a\}}_{0}, \underbrace{\{b\}}_{0}, \underbrace{\{c\}}_{0}, \underbrace{\{d\}}_{0}, \underbrace{\{e\}}_{0}, \underbrace{\{d,e\}}_{1}, \underbrace{\{a,b\}}_{2}, \underbrace{\{c,d,e\}}_{2.5}, \underbrace{\{a,b,c,d,e\}}_{5}, \}$$



Interdisciplinary endeavour

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.

Limit of visualization (1/3)

Prostate cancer data[14]

- Individuals: 506 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)
- Some missing data: 62 missing values ($\approx 1\%$)

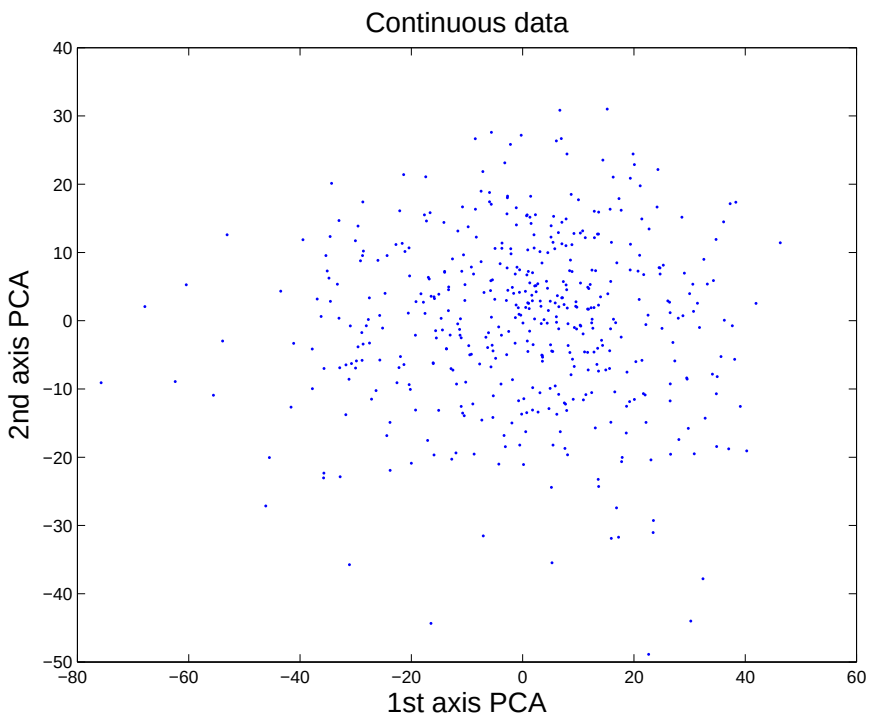
Questions

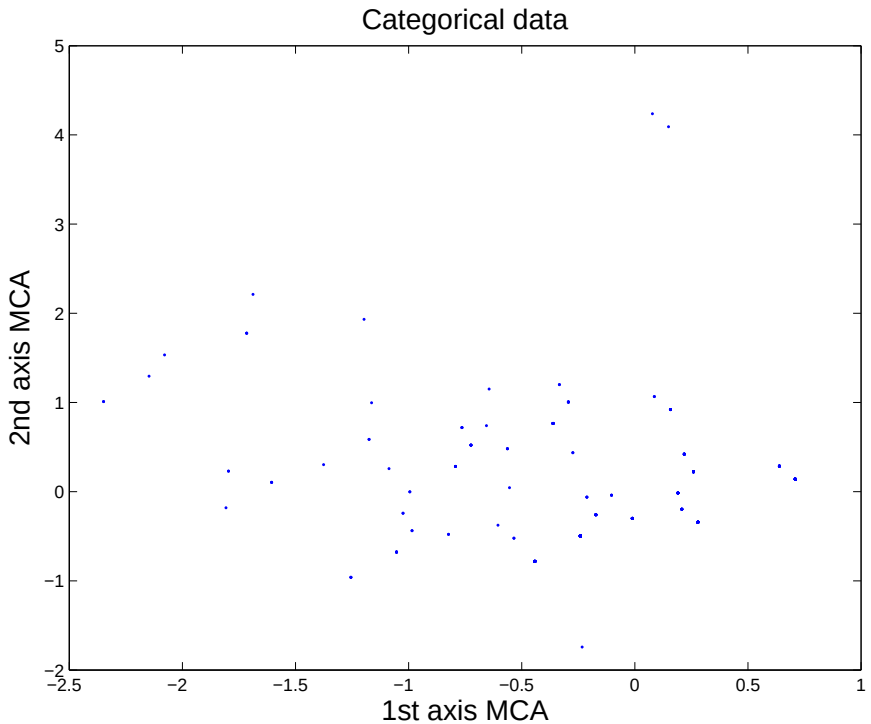
- Doctors say two stages of the disease are well separated with the variables
- We forget the classes (Stages of the disease) to retrieve clustering by visualizing

[14] Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

Limit of visualization (2/3)

- Perform PCA for continuous variables, MCA for categorical ones
- Discard missing values since not convenient for traditional PCA and MCA

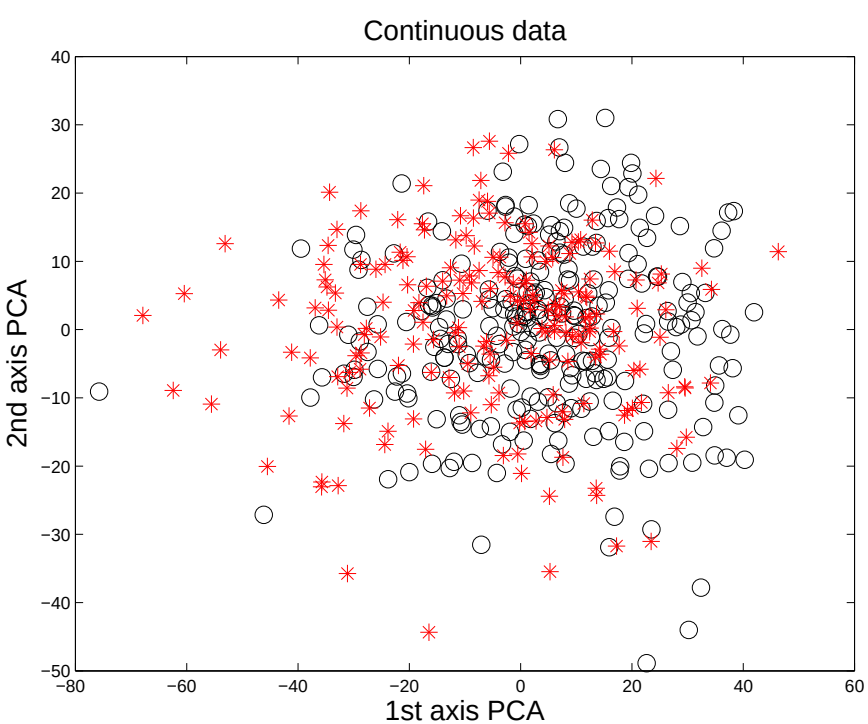


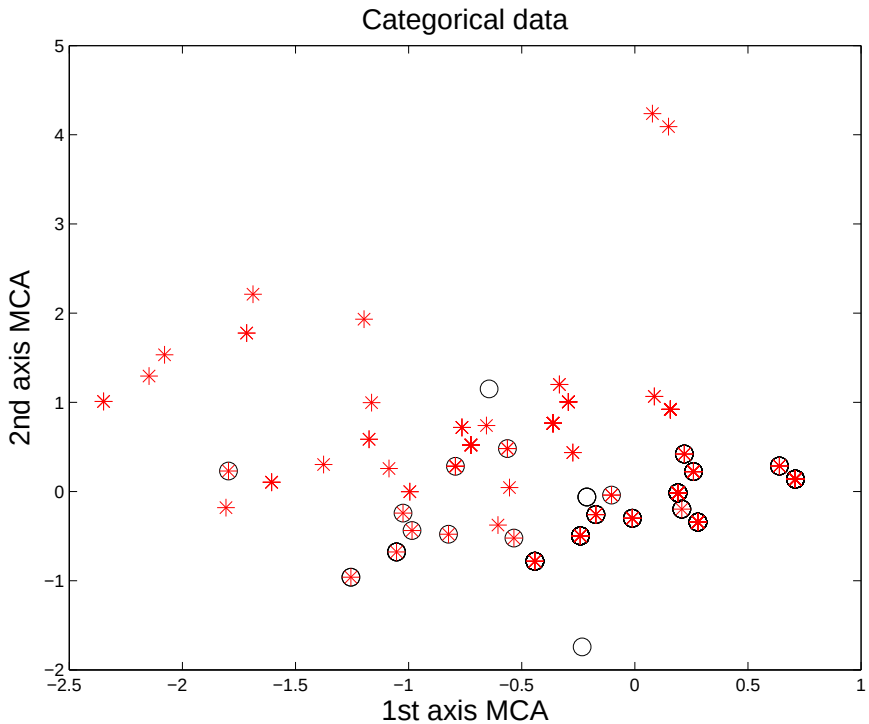


We do not "see" anything...

Limit of visualization (2/3)

If we have a look at the true partition, indeed visualizing was not enough. . .





Need to define more "automatic" methods acting directly in the space $\mathcal{X}$

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...

Outline

Clustering of clustering algorithms[15]

- 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.

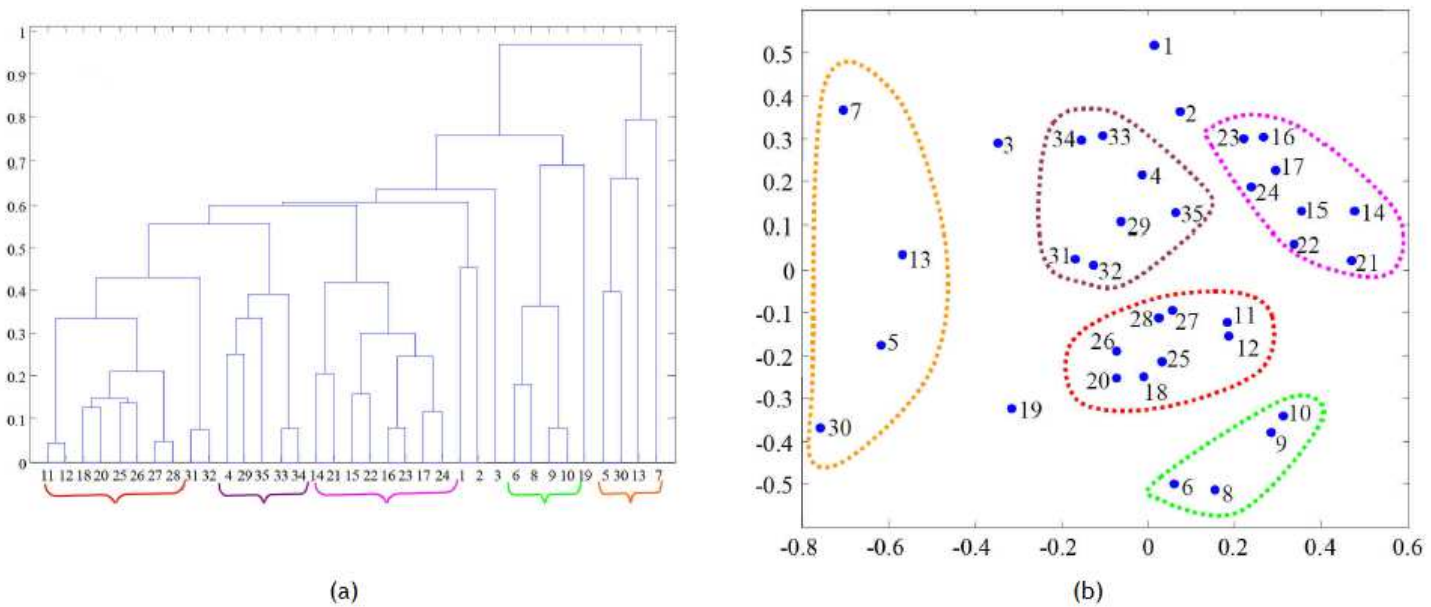


Figure 10 Clustering of clustering algorithms. (a) Hierarchical clustering of 35 different algorithms; (b) Sammon's mapping of the 35 algorithms into a two-dimensional space, with the clusters highlighted for visualization. The algorithms in the group (4, 29, 31-35) correspond to K-means, spectral clustering,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

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

Dissimilarities, distances, similarities: definitions

General idea

- Put in the same cluster individuals $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ which are similar
- Put in different clusters individuals $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ which are dissimilar

- Definition of a dissimilarity $\delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$
  - Separation: for any $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}^2$, $\delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0 \Leftrightarrow \boldsymbol{x}_1 = \boldsymbol{x}_2$
  - Symmetry: for any $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}^2$, $\delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \delta(\boldsymbol{x}_2, \boldsymbol{x}_1)$
- Definition of a distance $\delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$: it is a dissimilarity with a new property
  - Triangular inegality: for any $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \in \mathcal{X}^3$, $\delta(\boldsymbol{x}_1, \boldsymbol{x}_3) \leq \delta(\boldsymbol{x}_1, \boldsymbol{x}_2) + \delta(\boldsymbol{x}_2, \boldsymbol{x}_3)$
- Definition of a similarity (or affinity) $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$
  - For any $\boldsymbol{x}_1 \in \mathcal{X}$, $\delta(\boldsymbol{x}_1, \boldsymbol{x}_1) = s_{\max}$ with $s_{\max} \geq s(\boldsymbol{x}_1, \boldsymbol{x}_2)$ for any $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}^2$, $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$
  - Symmetry: for any $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}^2$, $s(\boldsymbol{x}_1, \boldsymbol{x}_2) = s(\boldsymbol{x}_2, \boldsymbol{x}_1)$

Often dissimilarities (or similarities) are enough for clustering

Similarities, dissimilarities, distances: examples for quantitative data

$$\mathcal{X} = \mathbb{R}^d$$

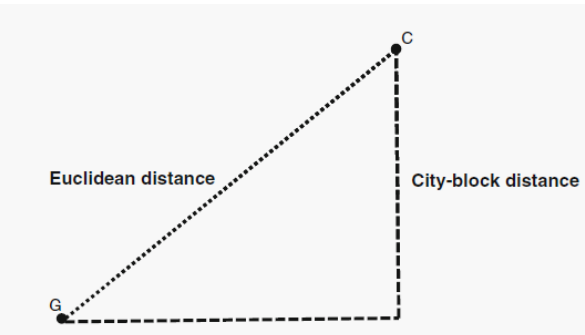
- Euclidean distance (or L_1 norm) with metric $\mathbf{M}$:

$$\delta_{\mathbf{M}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_{\mathbf{M}} = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)'\mathbf{M}(\mathbf{x}_1 - \mathbf{x}_2)}$$

  with a metric $\mathbf{M}$: symmetric and positive definite $d \times d$ matrix
- Manhattan distance or city-block distance or L_1 norm:

$$\delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{d} \|\boldsymbol{x}_{1j} - \boldsymbol{x}_{2j}\|$$

- ...



Similarities, dissimilarities, distances: examples for binary data

$$\mathcal{X} = \{0,1\}^d \text{ thus } \boldsymbol{x}_{ij} \in \{0,1\}$$

- The Hamming distance counts the dimensions on which both vectors are different:

$$\delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{d} \mathbb{I}_{\{x_{1j} \neq x_{2j}\}}$$

- The matching coefficient counts dimensions on which both vectors are non-zero:

$$\delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{d} \mathbb{I}_{\{x_{1j} = x_{2j} = 1\}}$$

- ...

A lot of dissimilarities and distances...

For instance, in the binary case[16]:

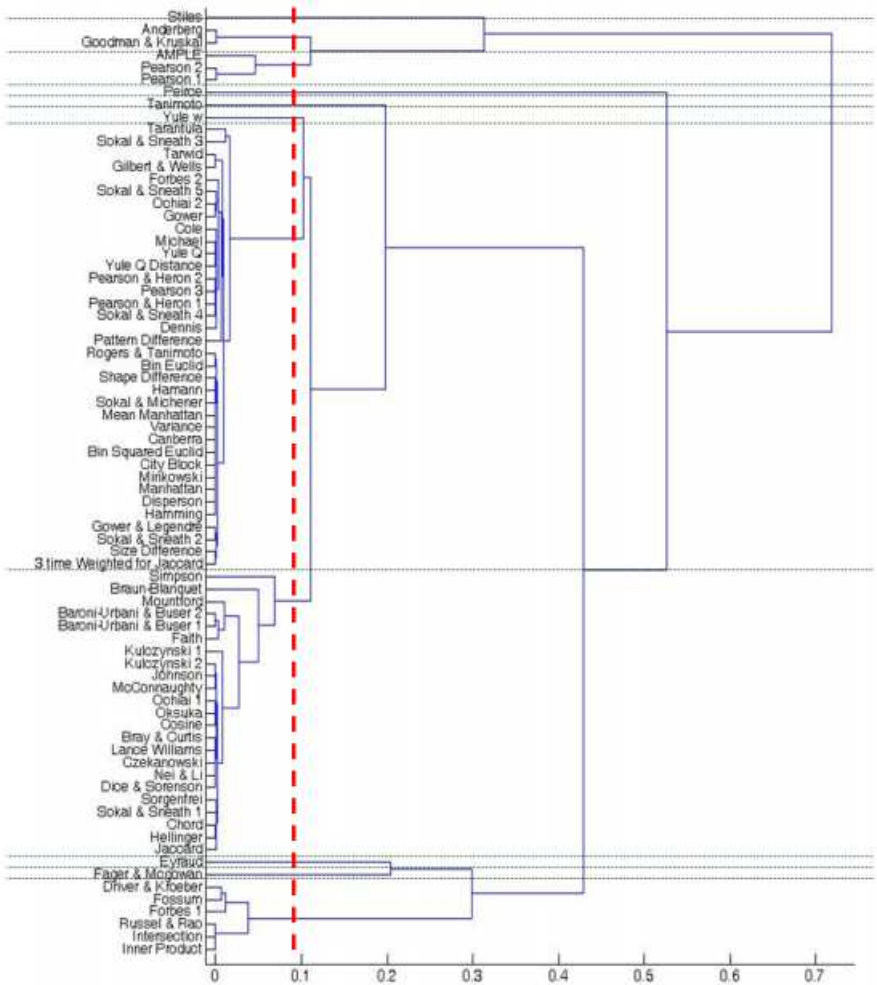


Figure 2 Hierarchical Clustering Result of Random Binary Data Set

[16] S.-S. Choi, S.-H. Cha, C. C. Tappert (2010). A Survey of Binary Similarity and Distance Measures. Systemics, Cybernetics and ININformatics, 8, 1, 43–48.

Similarities, dissimilarities, distances: similarities proposals

- It is easy to construct similarities from dissimilarities or distances
- A typical example:

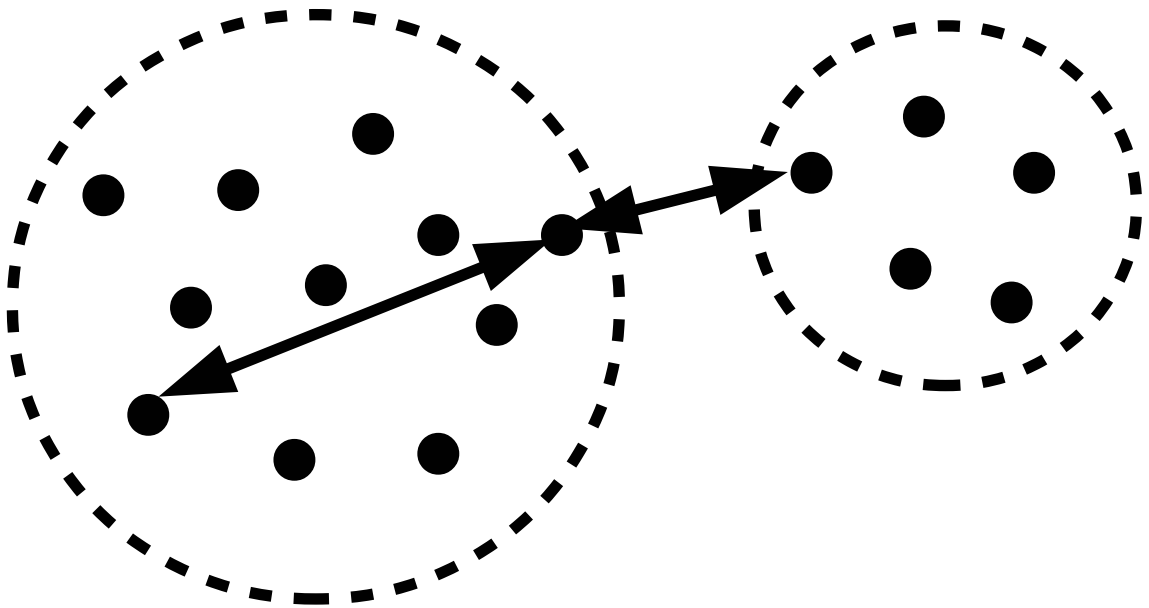
$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\delta(\boldsymbol{x}_1, \boldsymbol{x}_2))$$

  Thus $s_{\max} = 1$
- The Gaussian similarity is an important case ($\sigma > 0$):

$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{2\sigma^2}\right)$$

Weaken a two restrictive requirement

The property that two individuals in the same cluster are closer than any other individual in another cluster is two strong



Relax this local criterion by a global criterion

Within-cluster inertia criterion

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

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

- Look for compact clusters (individuals of the same cluster are close from each other)
- $\|\cdot\|_{\mathbf{M}}$ is the Euclidian distance with metric $\mathbf{M}$ in $\mathbb{R}^d$
- $\bar{\mathbf{x}}_k$ is the mean (or center) 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

Between-cluster inertia criterion

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

$$B_{\mathbf{M}}(\mathbf{z}) = \sum_{k=1}^{K} n_k \|\bar{\mathbf{x}} - \bar{\mathbf{x}}_k\|^2_{\mathbf{M}}$$

- Look for clusters far from each other
- $\bar{\mathbf{x}}$ the mean (or center) of the whole data set $\boldsymbol{x}$:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

- It is in fact equivalent to minimize $W_{\mathbf{M}}(\mathbf{z})$ since

$$B_{\mathbf{M}}(\mathbf{z}) + W_{\mathbf{M}}(\mathbf{z}) = \text{cste}$$

A combinatorial problem

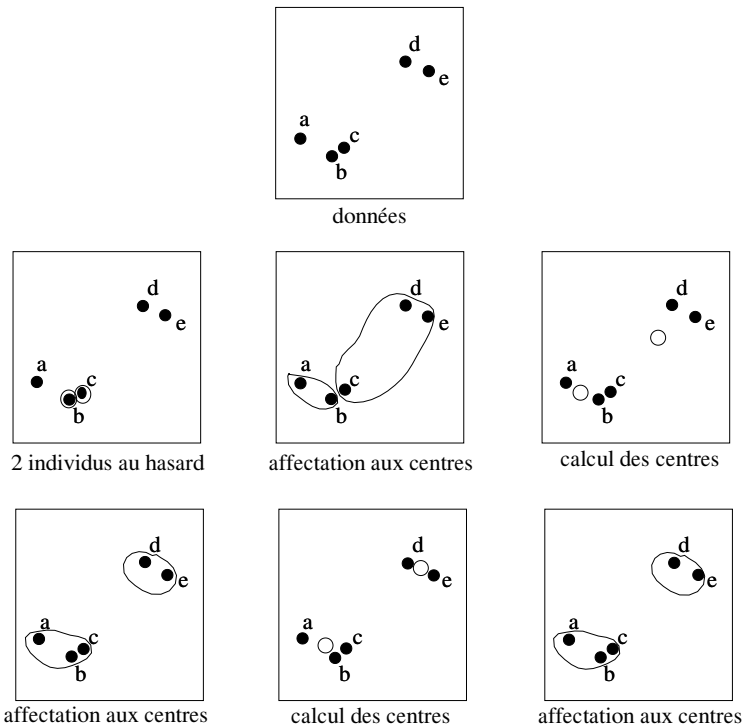
- Number of possible partitions $\boldsymbol{z}$: $\#\{\boldsymbol{z}\} \approx \frac{K^n}{K!}$
- For instance : $n = 40$, $K = 3$, $\#\{\boldsymbol{z}\} \approx 2.10^{18}$ (two billion of billion), thus about 64.000 years for a computer calculating one million of partitions per second
- Thus impossible to minimize directly $W_{\mathbf{M}}(\boldsymbol{z})$
- But equivalent to minimize on $\boldsymbol{z}$ and $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ with $\boldsymbol{\mu}_k \in \mathbb{R}^d$, the criterion

$$\tilde{W}_{\mathbf{M}}(\mathbf{z}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_{\mathbf{M}}^2$$

It leads to the famous K-means algorithm

K-means algorithm

Alternating optimization between the partition $\boldsymbol{z}$ and the center $\boldsymbol{\mu}$ of clusters



Algorithme des centres mobiles

- Start from a random $\boldsymbol{\mu}$ among $\boldsymbol{x}$ (avoid to start from a random $\boldsymbol{z}$)
- Each iteration decreases the criterion $\tilde{W}_{\mathsf{M}}(\boldsymbol{z}, \boldsymbol{\mu})$ (so also $W_{\mathsf{M}}(\boldsymbol{z})$)
- Converge into a finite (and often low) number of iterations
- Memory complexity $O((n + K)d)$, time complexity $O(\text{nb.iter.}Kdn)$

Fuzzy within-cluster inertia criterion

Select the fuzzy partition $\mathbf{c}$ minimizing the criterion

$$W_{\mathbf{M},m}^{\text{fuzzy}}(\mathbf{c}) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}^{m}\|\mathbf{x}_i - \bar{\mathbf{x}}_k\|_{\mathbf{M}}^2$$

- Fuzzy partition $\mathbf{c} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$, $\mathbf{c}_k = (c_{1k}, \ldots, c_{nk})$, $c_{ik} \in [0, 1]$ defined by

$$c_{ik} = \frac{1}{\sum_{k'=1}^{K}\left(\frac{\|\mathbf{x}_i - \bar{\mathbf{x}}_k\|_{\mathbf{M}}}{\|\mathbf{x}_i - \bar{\mathbf{x}}_{k'}\|_{\mathbf{M}}}\right)^{\frac{2}{m-1}}}$$

- $m \geq 1$ determines the level of cluster fuzziness: m large produces fuzzier clusters; $m = 1$ (in the limit) produces hard clustering (m converges to 0 or 1)
- $\bar{\mathbf{x}}_k$ is the fuzzy mean (or center) of the kth cluster

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

  and $n_k = \sum_{k=1}^{n} c_{ik}$ indicates the fuzzy number of individuals in cluster k
- Optimize $W_{\mathbf{M},m}^{\text{fuzzy}}(\mathbf{c})$ (through $\tilde{W}_{\mathbf{M},m}^{\text{fuzzy}}(\mathbf{c}, \boldsymbol{\mu})$) by the fuzzy K-means algorithm

Kernel clustering: the idea

- Original (input) space: non-linearly separable clusters
- Feature space: a non-linear mapping ϕ to a higher-dimensional space

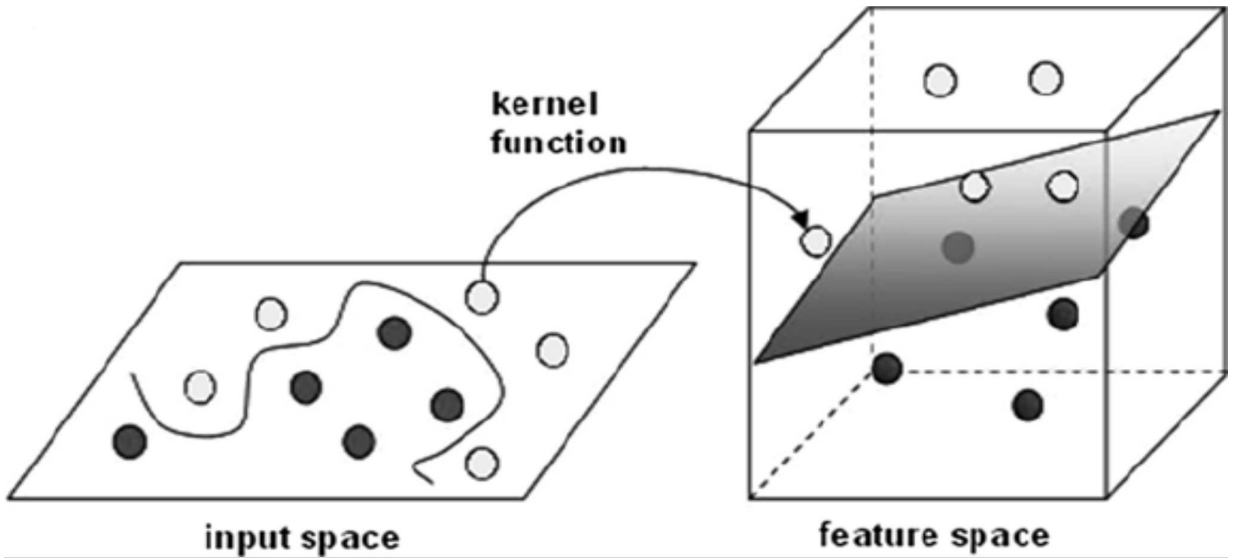
$$\begin{array}{rccc} \phi: & \mathcal{X} & \mapsto & \mathcal{Y} \\ & \boldsymbol{x}_i & \to & \phi(\boldsymbol{x}_i) \end{array}$$

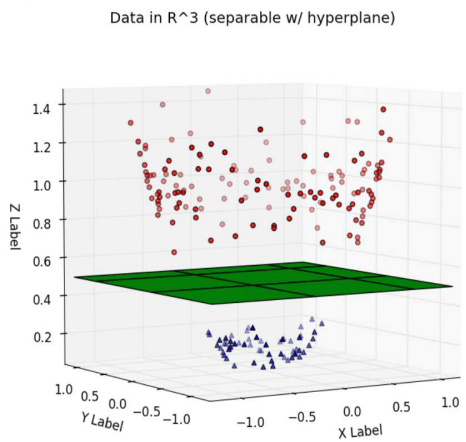
- Expected property: linearly separable clusters in the feature space
- Difficulty: define ϕ
- The "kernel trick": not necessary to define ϕ, define just an inner product

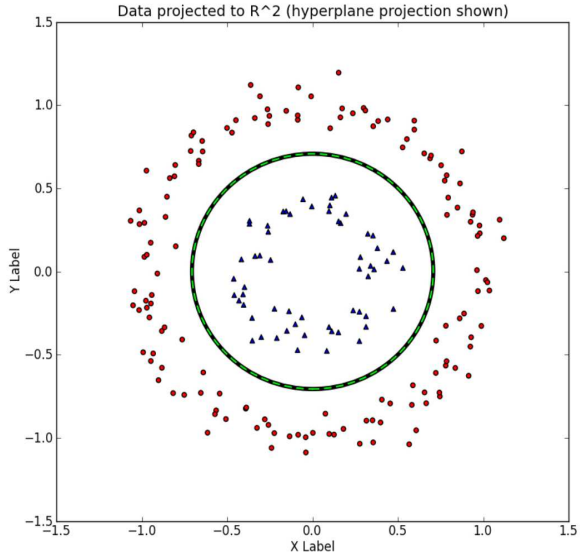
$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_{i'})$$

  κ is so-called a kernel
- Lesson 4: more arguments on the discriminant interest of higher dimension

Kernel clustering: illustration







Kernel clustering: fuzzy within-cluster rewriting

- W^{fuzzy} with the data set $\phi(\boldsymbol{x}) = \{\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)\}$ (we drop $\mathbf{M}$ and m)

$$W^{\text{fuzzy}}(\mathbf{c}) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}^m \|\phi(\mathbf{x}_i) - \bar{\phi}_k(\boldsymbol{x})\|^2$$

with

$$\bar{\phi}_k(\mathbf{x}) = \frac{1}{n_k}\frac{\sum_{i=1}^n c_{ik}^m \phi(\mathbf{x}_i)}{\sum_{i=1}^n c_{ik}^m} \text{ and } c_{ik} = \frac{1}{\sum_{k'=1}^K \left(\frac{\|\phi(\mathbf{x}_i) - \bar{\phi}_k(\mathbf{x})\|}{\|\phi(\mathbf{x}_i) - \bar{\phi}_{k'}(\mathbf{x})\|}\right)^{\frac{2}{m-1}}}$$

- After some algebra, only appears inside an inner product since:

$$\begin{aligned}
\|\phi(\mathbf{x}_i) - \bar{\phi}_k(\boldsymbol{x})\|^2 &= \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_i) - 2\frac{\sum_{i=1}^n c_{ik}^m \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_{i'})}{\sum_{i'=1}^n c_{i'k}^m} + \frac{\sum_{i'=1}^n \sum_{i''=1}^n c_{i'k}^m c_{i''k}^m \phi(\boldsymbol{x}_{i'})'\phi(\boldsymbol{x}_{i''})}{(\sum_{i'=1}^n c_{i'k}^m)^2} \\
&= \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2\frac{\sum_{i=1}^n c_{ik}^m \kappa(\boldsymbol{x}_i, \boldsymbol{x}_{i'})}{\sum_{i'=1}^n c_{i'k}^m} + \frac{\sum_{i'=1}^n \sum_{i''=1}^n c_{i'k}^m c_{i''k}^m \kappa(\boldsymbol{x}_{i'}, \boldsymbol{x}_{i''})}{(\sum_{i'=1}^n c_{i'k}^m)^2}
\end{aligned}$$

Consequently, need only to define κ for performing the (fuzzy) K-means...

Kernel clustering: kernel examples

- Polynomial

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = (\boldsymbol{x}_i' \boldsymbol{x}_{i'} + a)^b$$

- Gaussian

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|/(2\sigma^2))$$

- Sigmoid

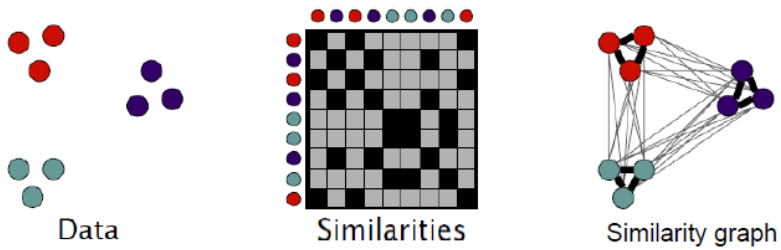
$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \tanh(a\boldsymbol{x}_i' \boldsymbol{x}_{i'} + b)$$

Spectral clustering: the idea[17]

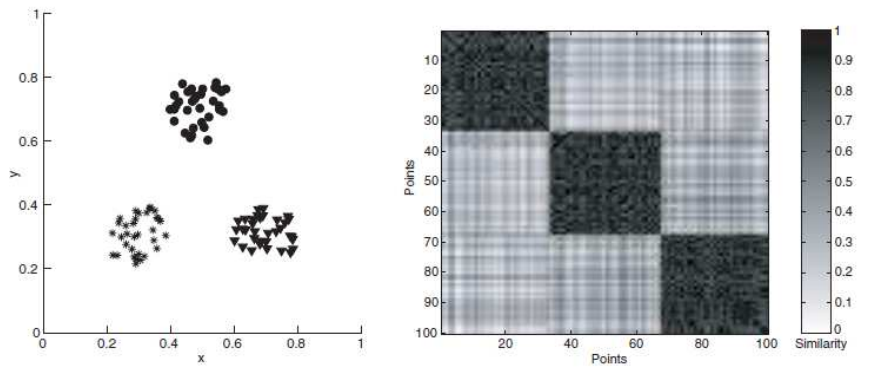
- Build a similarity matrix $\mathbf{S}$ $n \times n$ from $\boldsymbol{x}$:

$$\mathbf{S}_{ii'} = s_{ii'} \text{ if } i \neq i', \quad \mathbf{S}_{ii} = 0$$

- Similarities can be viewed as proximities between individuals (nodes) in a graph



- If some disconnected sub-graphs, $\mathbf{S}$ can be reordered as block diagonal



(a) Well-separated clusters. (b) Similarity matrix sorted by cluster labels.

- We "can see" the clusters thus clustering is expected to be easy

[17] Figures from [A. Jain *at al.*. Data Clustering: A Review.]

Spectral clustering: Laplacian matrix

- For technical reasons, we study (unnormalized graph) Laplacian **L** instead of **S**

$$\mathbf{L} = \mathbf{D} - \mathbf{S}$$

  where $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$ is the diagonal degree matrix with $d_i = \sum_{i'=1}^{n} \mathbf{S}_{ii'}$
- Have in mind **S** if easier ; the structure of **L** is similar

Properties of **L**

1. For every vector $\mathbf{u} = (u_1, \ldots, u_n') \in \mathbb{R}^n$, we have

$$\mathbf{u}'\mathbf{L}\mathbf{u} = \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} s_{ii'} (u_i - u_{i'})^2$$

2. **L** is symmetric and positive semi-definite
3. The smallest eigenvalue of **L** is 0, the corresponding eigenvector is the constant one vector $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^n$
4. **L** has n non-negative, real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$

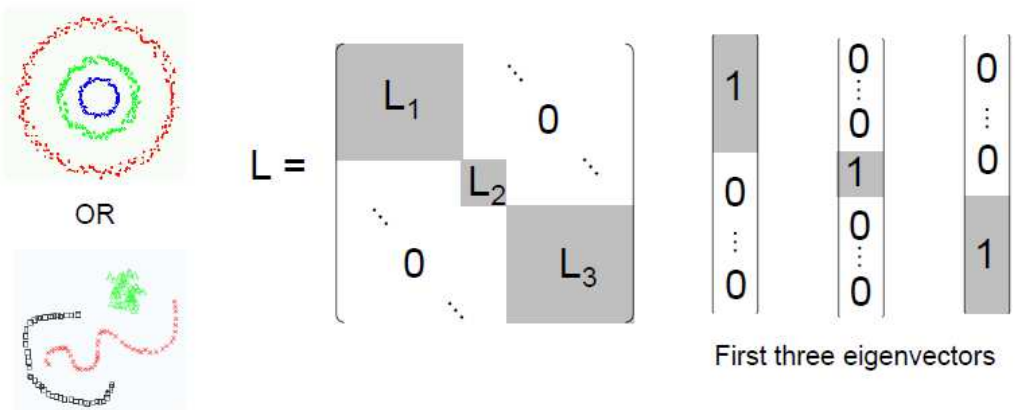
Spectral clustering: specific spectral structure of **L**

Linking clusters and **L**

1. The multiplicity K of the eigenvalue 0 of $\mathbf{L}$ equals the number of clusters $G_1, \ldots, G_K$ (or the connected components in the graph)
2. The eigenspace of eigenvalue 0 is spanned by the indicator vectors $\mathbf{1}_{G_1}, \ldots, \mathbf{1}_{G_K}$ of those components, where $\mathbf{1}_{G_k}(i) = 1$ if $\boldsymbol{x}_i \in G_k$, zero otherwise

[A. Singh (2010). Spectral Clustering. Machine Learning.]

- If graph is connected, first Laplacian evec is constant (all 1s)
- If graph is disconnected (k connected components), Laplacian is block diagonal and first k Laplacian evecs are:



Spectral clustering: eigenvectors as new data to cluster

- Define a new data set $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \in \mathbb{R}^K$ where

$$\mathbf{y}_i = (\mathbf{1}_{G_1}(i), \ldots, \mathbf{1}_{G_K}(i))$$

- From the previous property, $\mathbf{y}_i$s in the same clusters are very similar...
- ...whereas $\mathbf{y}_i$s in different clusters are very different
- Thus just run a K-means (for instance) to finish the clustering job!

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

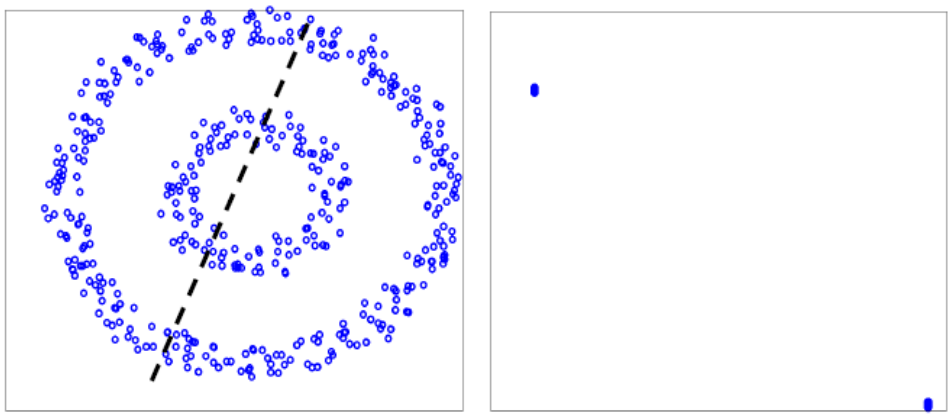
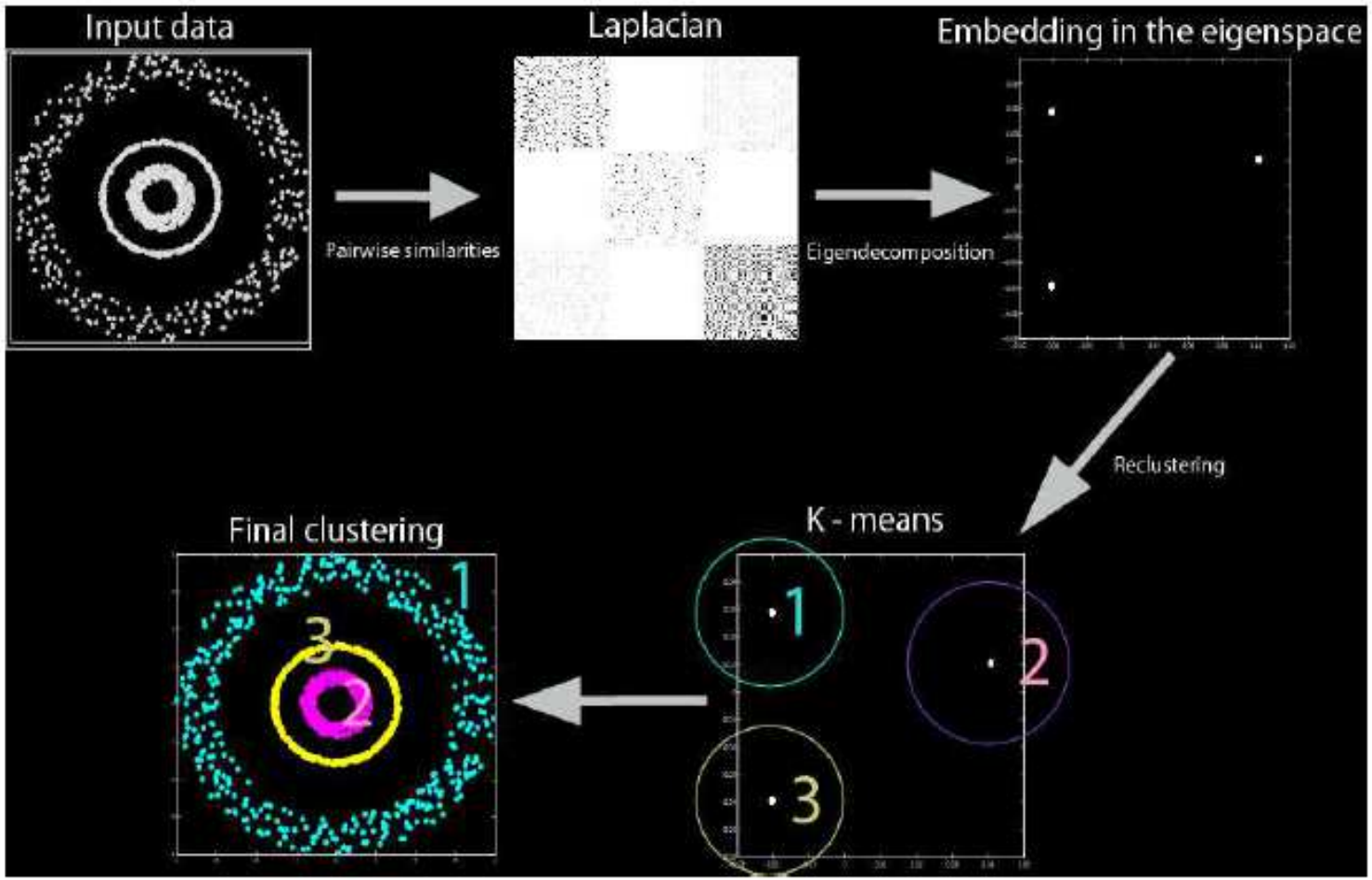


Figure 5 Importance of a good representation. (a) "Two rings" dataset where K-means fails to find the two "natural" clusters; the dashed line shows the linear cluster separation boundary obtained by running K-means with K = 2. (b) a new representation of the data in (a) based on the top 2 eigenvectors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

Spectral clustering: summary[18]



[18] [J. Suykens and C. Alzate (2011). Kernel spectral clustering: model representations, sparsity and out-of-sample extensions.]

Spectral clustering: some last elements

- Other Laplacian choices are possible, as the normalized one

$$\mathbf{L}^{\text{norm}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

- Its overall complexity is $O(n^2)$ thus problem when n large
- It is in fact equivalent to kernel K-means modulo relaxed version[19]

[19] See details in [I.S. Dhillon *et al.* (2004). Kernel kmeans, Spectral Clustering and Normalized Cuts. KDD'04.]

Hierarchical agglomerative clustering: algorithm

Aim

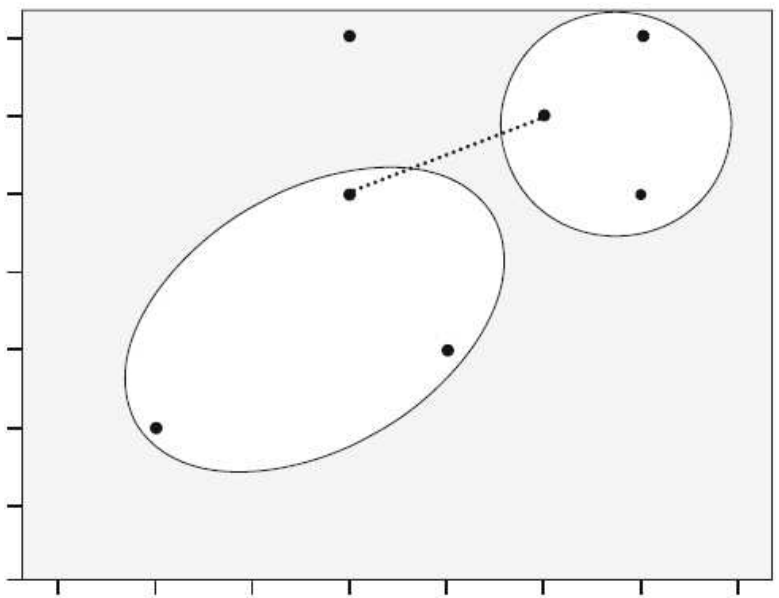
Design a procedure to build an indexed hierarchy

- Start: make a dissimilarity matrix from δ between singletons $\{\boldsymbol{x}_i\}$ of $\boldsymbol{x}$
- Iterations:
  - Merge two clusters that are the closest from a linkage criterion Δ
  - Compute dissimilarities between new clusters
- End: as soon as a unique cluster

- Need to define $\Delta : H \times H \to \mathbb{R}^+$ from δ
- If Δ is non-decreasing, it defines an index on the final hierarchy

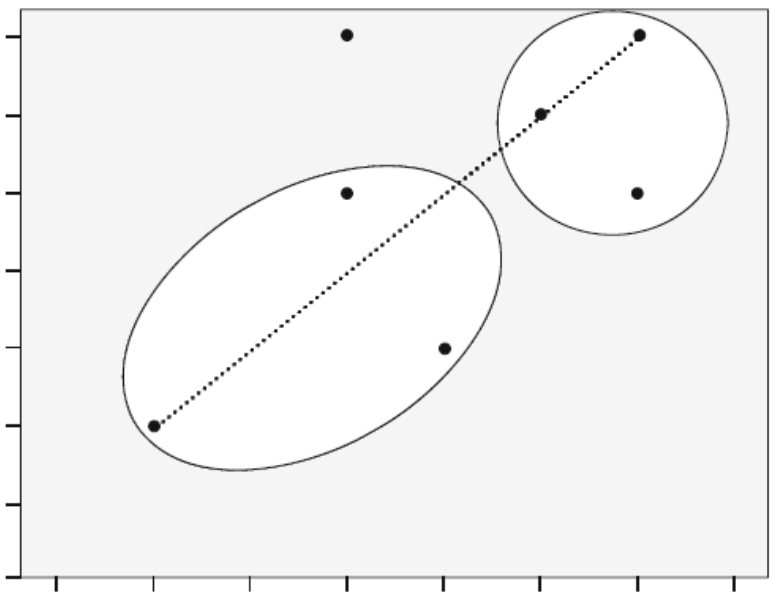
Single linkage criterion

$$\Delta(G_1, G_2) = \min\{\delta(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) : \boldsymbol{x}_i \in G_1, \boldsymbol{x}_{i'} \in G_2\}$$



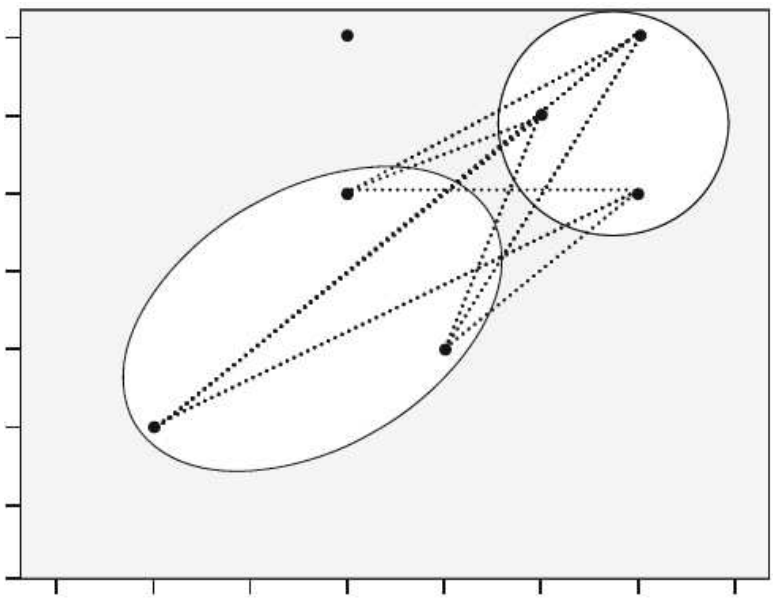
Complete linkage criterion

$$\Delta(G_1, G_2) = \max\{\delta(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) : \boldsymbol{x}_i \in G_1, \boldsymbol{x}_{i'} \in G_2\}$$



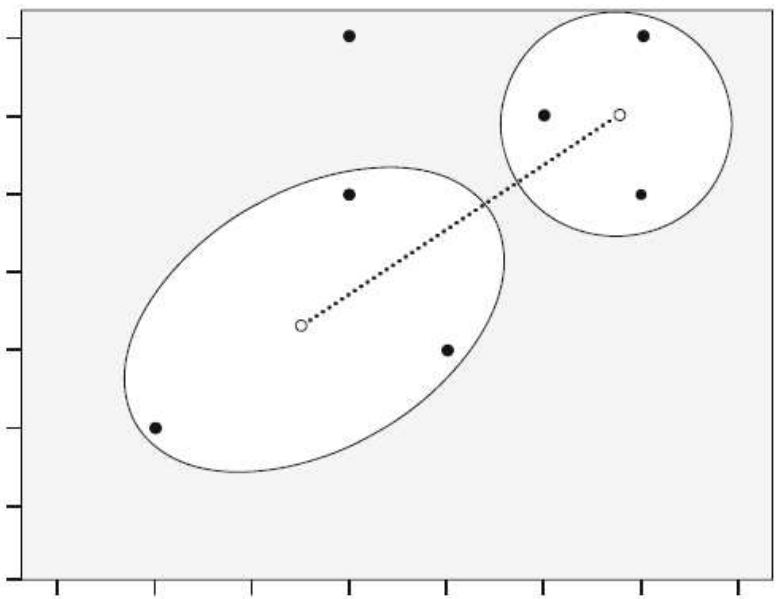
Average linkage criterion

$$\Delta(G_1, G_2) = \frac{1}{n_1 n_2} \sum_{\boldsymbol{x}_i \in G_1} \sum_{\boldsymbol{x}_{i'} \in G_2} \delta(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$$

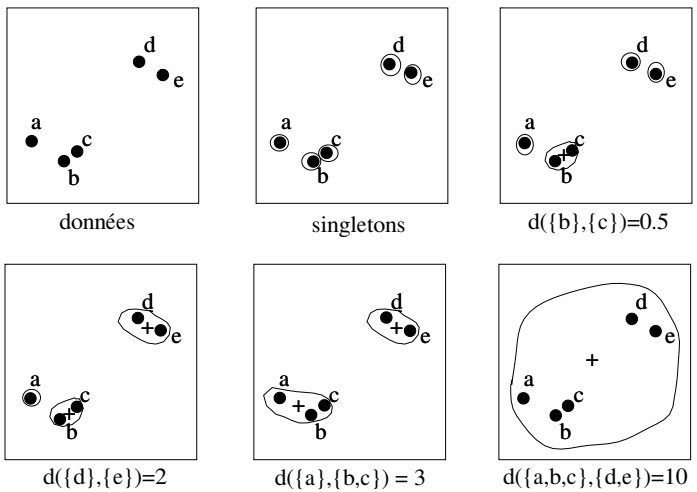


Ward linkage criterion

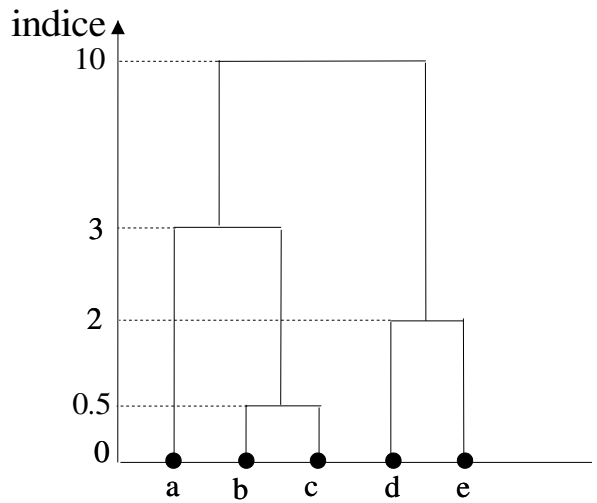
$$\Delta(G_1, G_2) = \frac{n_1 n_2}{n_1 + n_2} \delta^2(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$$



Hierarchical agglomerative clustering: Ward example



Classification hiérarchique ascendante
(méthode de Ward)



Dendogramme

- A partition is obtained by cuting the dendrogram
- A dissimilarity matrix between pairs of individuals is enough
- Memory complexity $O(n^2 \ln n)$, space complexity $O(n^2)$ (do not scale well...)
- Suboptimal optimisation of $W_{\mathbf{M}}(\cdot)$ (see next slide)

Link between Ward and K-means

Two successive Ward hierarchical algorithm iterations minimize W

- Partition at a given iteration: $P = \{G_1, \ldots, G_K, G_{K+1}, G_{K+2}\}$

$$W(P) = \sum_{k=1}^{K} \sum_{x_i \in G_k} \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + \sum_{x_i \in G_{K+1}} \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_{K+1}) + \sum_{x_i \in G_{K+2}} \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_{K+2})$$

- Partition at a the next iteration: $P^+ = \{G_1, \ldots, G_K, \underbrace{G_{K+1} \cup G_{K+2}}_{G^+_{K+1}}\}$

$$W(P^+) = \sum_{k=1}^{K} \sum_{x_i \in G_k} \delta(\boldsymbol{x}_i, \boldsymbol{\mu}_k) + \sum_{x_i \in G_{K+1} \cup G_{K+2}} \delta(\boldsymbol{x}_i, \boldsymbol{\mu}^+_{K+1})$$

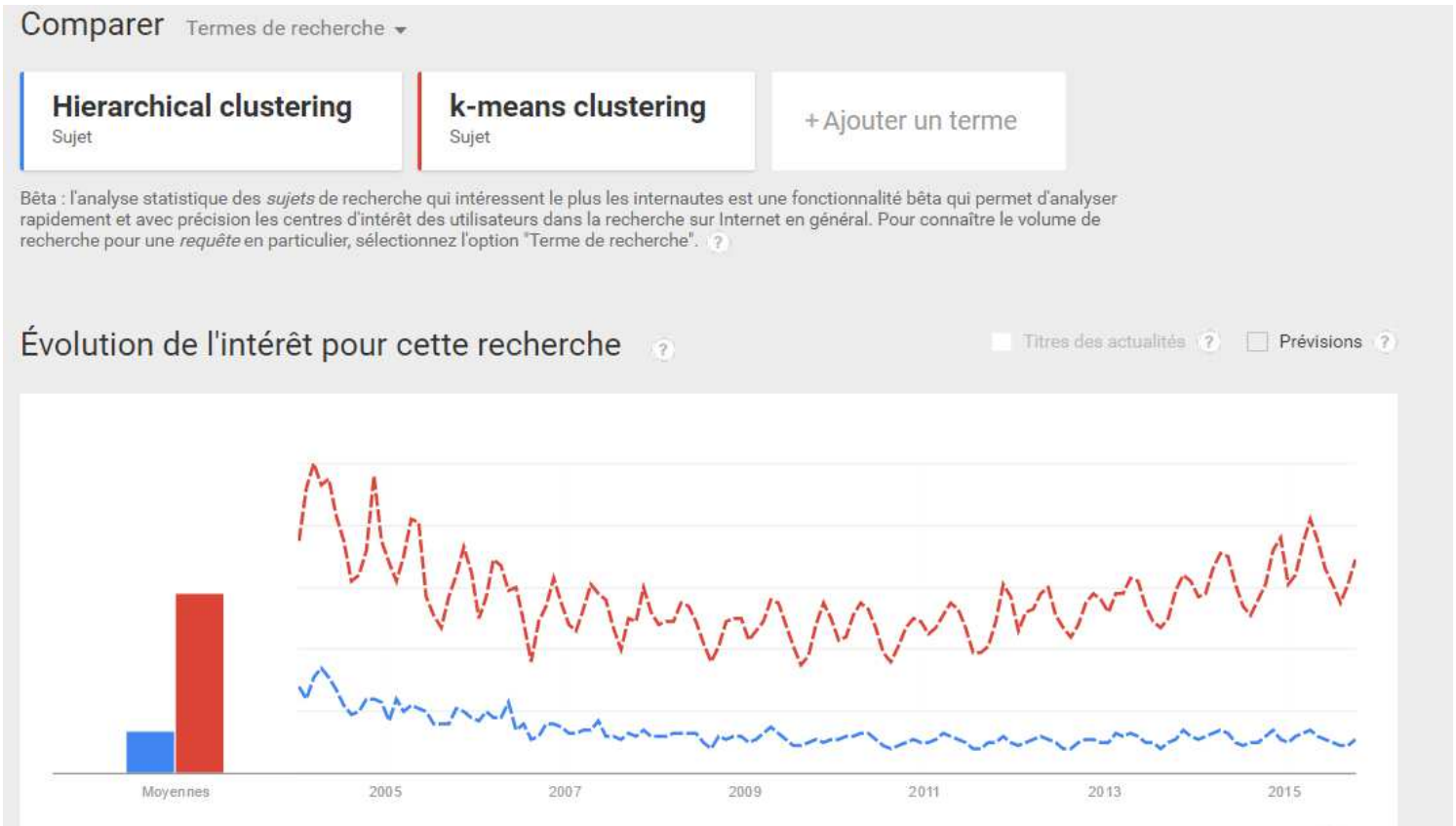
- After some algebra and using the Koenig-Huygens theorem, we obtain

$$W(P^+) - W(P) = \frac{n_{K+1} n_{K+2}}{n_{K+1} + n_{K+2}} \delta(\boldsymbol{\mu}_{K+1}, \boldsymbol{\mu}_{K+2})$$

- Thus, it is like performing K-means under (partial) partition P constraints

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)



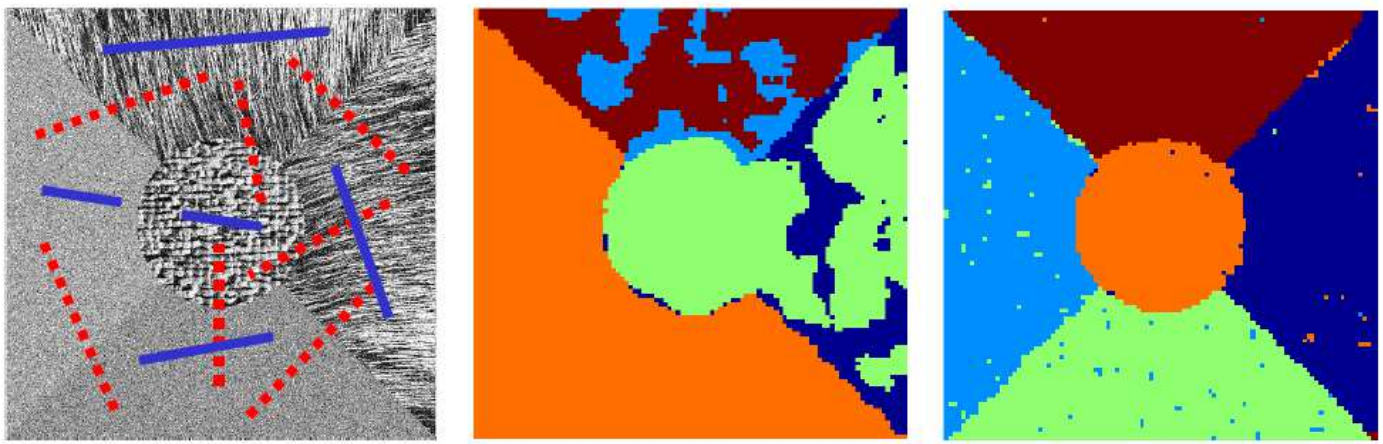
Outline

Packages

- R (statistical community): see practical sessions to use some of them
- Python (machine learning community): have a look also at the scikit-learn library

Semi-supervised clustering[20]

- 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



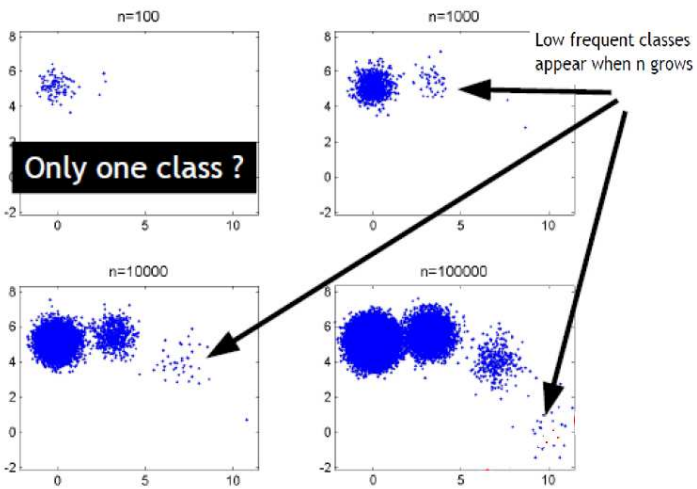
(a) Input image and constraints (b) No constraints (c) 10% pixels in constraints

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].

[20] 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



Next lesson

Introduction to cluster analysis and classification:
Evaluating clustering