Sheet

(Module 3)

Science IFT3700 data /

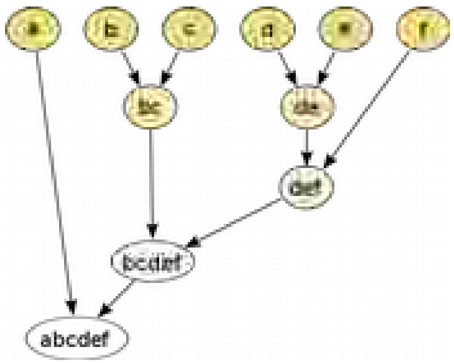
IFT6758 Fall 2018 ©

Alain Tapp

Contents

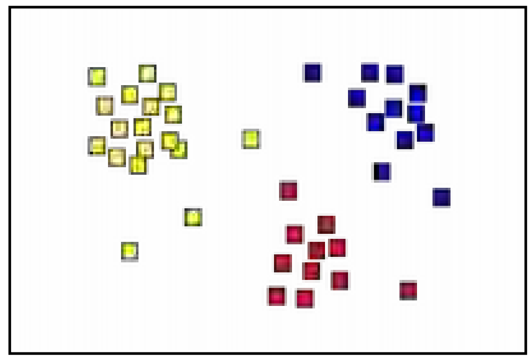
- Partitioning data
- Distance and similarity
- hierarchical clustering
- Expectation-maximization algorithm
  - -average
  - GMM
- dbscan

grouping data partitioning methods



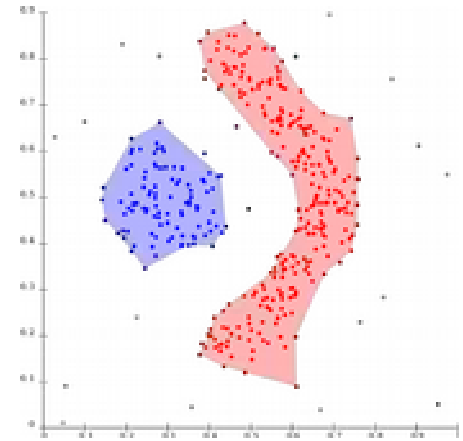
hierarchical

- Similarity
- by division
- by grouping



centroid

- Similarity or distance
- Average
- Representative



neighborhood

- Similarity distance
- neighborhood graph
- Density

Partitioning data

Given a set of points, with a notion of distance between points, group points to a number of groups, so that:

- The group members are close / similar to each other.
- Members of different groups are different.
- Points are soudants in a large space.
- The proximity is defined with the aid of a measure of similarity or distance.

Partitioning data

Clustering in two dimensions look easy.

Le regroupement en deux dimensions semble facile.

Many applications involve a large number of dimensions (MNIST,)

De nombreuses applications impliquent un grand nombre de dimensions (MNIST, $d = 784$)

In large areas, it is often noted that almost all pairs of points are about the same distance.

Dans les espaces de grande dimension, on remarque souvent que presque toutes les paires de points sont à peu près à la même distance.

Examples

galaxies

A catalog of 2 billion "celestial objects" represents objects by their radiation in 7 dimensions (frequency bands). How to group similar objects, such as galaxies, nearby stars, quasars, etc.

Music

The music is divided into categories and customers prefer certain categories. What is the best ranking. Represent parts of a set of customers who bought it.

Texts and articles

Representing the documents by all the significant words they contain.

distances

Distance

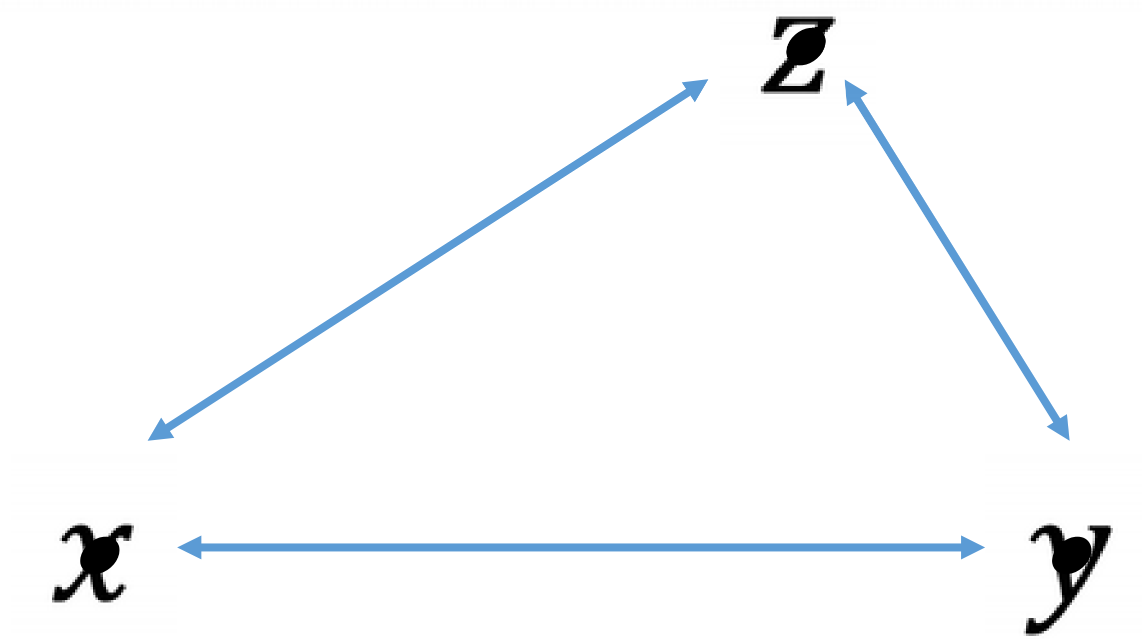
$D: E \times E \to \mathbb{R}^+$

Positivity: Positivité: $D(x, y) \geq 0$

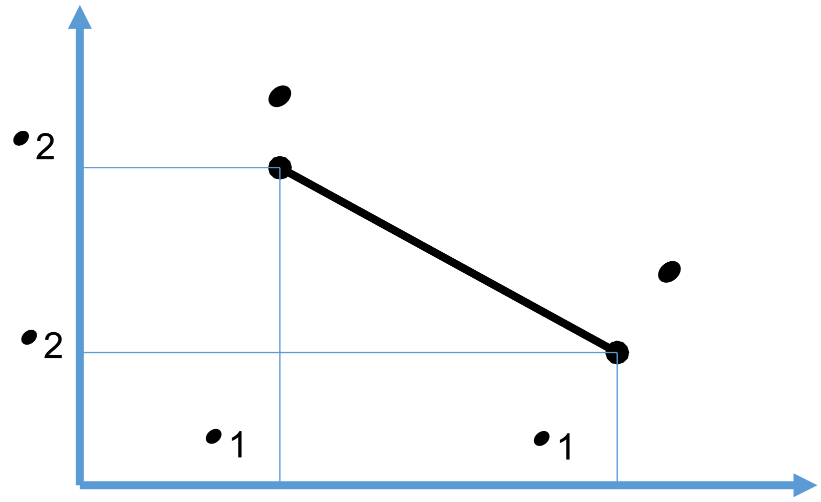
Symmetry: Symétrie: $D(x, y) = D(y, x)$

Separation: Séparation: $D(x, y) = 0 \rightarrow x = y$

triangle inequality: Inégalité triangulaire: $D(x, z) \leq D(x, y) + D(y, z)$

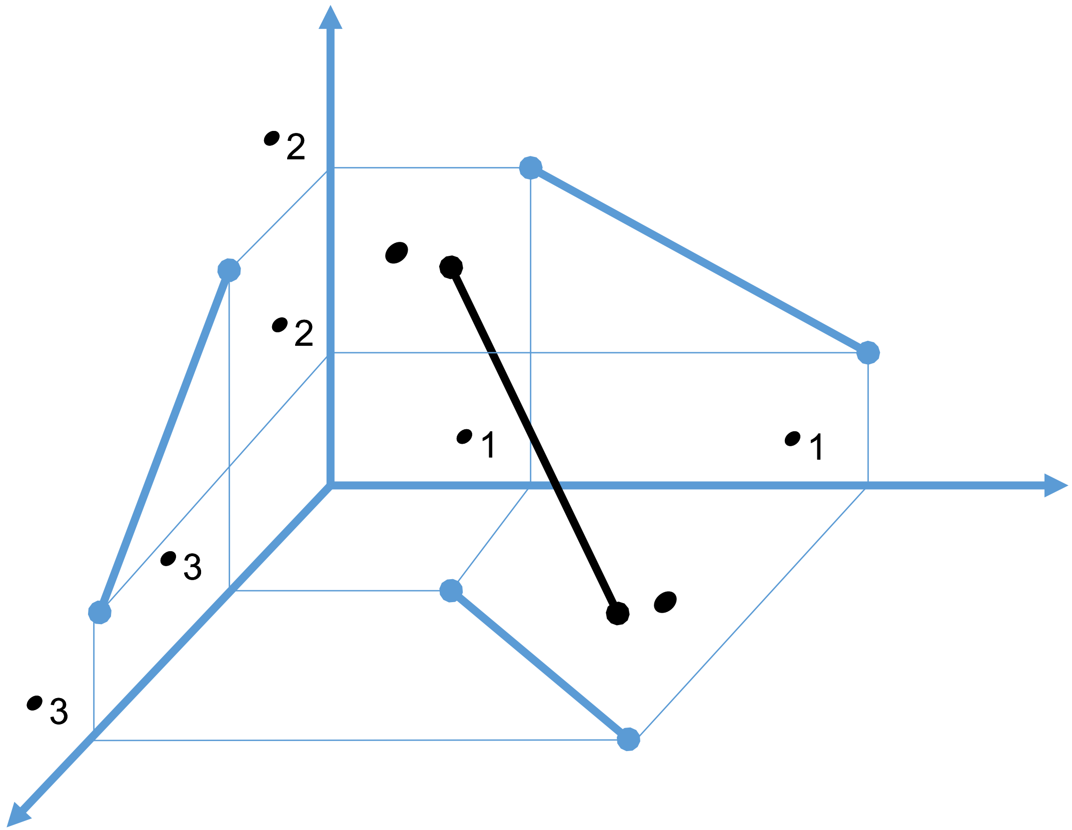


Euclidean Distance



•

Euclidean Distance



•

$\bullet = (\bullet_1\ \bullet_2\ \bullet_3)$ $\bullet = (\bullet_1\ \bullet_2\ \bullet_3)$

$$\bullet(\bullet, \bullet) = \sqrt{(\bullet_1 - \bullet_1)^2 + (\bullet_2 - \bullet_2)^2 + (\bullet_3 - \bullet_3)^2}$$

$\bullet = (\bullet_1\ \bullet_2\ \bullet_3\ \bullet_4)$ $\bullet = (\bullet_1\ \bullet_2\ \bullet_3\ \bullet_4)$

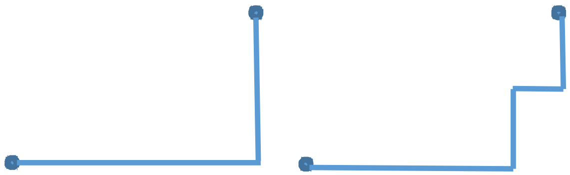
$$\bullet(\bullet, \bullet) = \sqrt{(\bullet_1 - \bullet_1)^2 + (\bullet_2 - \bullet_2)^2 + (\bullet_3 - \bullet_3)^2 + (\bullet_4 - \bullet_4)^2}$$

$\bullet = (\bullet_1\ \bullet_2\ \bullet_3 \cdots, \bullet_{784})$ $\bullet = (\bullet_1\ \bullet_2\ \bullet_3 \cdots, \bullet_{784})$

$$\bullet(\bullet, \bullet) = \sqrt{(\bullet_1 - \bullet_1)^2 + (\bullet_2 - \bullet_2)^2 + \cdots + (\bullet_{784} - \bullet_{784})^2}$$

$$L_1^1(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Manhattan

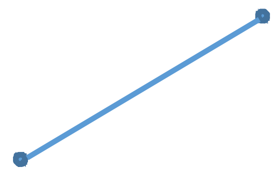


$$E_2^2(x,y) = \sum_{i=1}^{n} |x_i - y_i|^2$$

square error

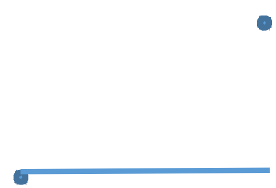
$$L_2^2(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$

Euclidean



$$L_\infty^\infty(x,y) = Max_i |x_i - y_i|$$

Max



cosine pseudodistance

$$x \cdot y = \sum_{i=1}^{n} x_i y_i$$

$$\|x\| = L_2(x, 0)$$

Similarité cosinus avec θ l'angle entre les vecteurs x et y.

Cosine similarity with the angle between the vectors and.

$$\cos(\theta) = S(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

$$-1 \leq S(x, y) \leq 1$$

Pseudorange cosine nonnegative vectors

$$D(x, y) = 1 - S(x, y)$$

Mahalanobis Distance

each coordinate is normalized according to the variance in that dimension.

On normalise chaque coordonnée en fonction de la variance dans cette dimension.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \left(\frac{x_i - y_i}{\sigma_i}\right)^2}$$

In general, is the covariance matrix of the distribution is obtained:

En général, S est la matrice de covariance de la distribution on obtient:

$$D(x,y) = \sqrt{(x-y)^T S^{-1} (x-y)}$$

Pseudodistance Jaccard

For sets can use the index of Jaccard.

Pour des ensembles on peut utiliser l'index de Jaccard.

For a whole is its cardinality.

Pour un ensemble x, $|x|$ est sa cardinalité.

$$\mathcal{J}(x,y) = \frac{|x \cap y|}{|x \cup y|} = \frac{|x \cap y|}{|x| + |y| - |x \cap y|}$$

$$0 \leq \mathcal{J}(x,y) \leq 1$$

$x = \{$☊, ♠, ♡, ♦, ֍, †, ∇, ∞, ∺, ⊛, ⌔, ⬠$\}$ $\qquad |x| = 12$

$y = \{$☊, ⏀, ♠, ♡, ♦, ∇, ∞, ∺, ⊛, ⋔, ⌔$\}$ $\qquad |y| = 11$

$x \cap y = \{$☊, ♠, ♡, ♦, ∇, ∞, ∺, ⊛, ⌔$\}$ $\qquad |x \cap y| = 9,$

$$\mathcal{J}(x,x) = \mathcal{J}(y,y) = 100\%$$

$$\mathcal{J}(x,y) = \frac{9}{12 + 11 - 9} = 64\%$$

Hamming distance and edit distance

For character sequences, bit or symbol of the same length, can be used **Hamming distance.**

Pour des séquences de caractère, bit ou symbole de même longueur, on peut utiliser la **distance de Hamming**.

$$H(x, y) = |\{i | x_i \neq y_i\}|$$

The **edit distance** counts the number of replacement, addition or deletion necessary to turn into.

La **distance d'édition** compte le nombre de remplacement, ajout ou effacement necessaire pour transformer x en y.

- $O_1: x_i \rightarrow a$
- $O_2: x_i \rightarrow x_i \cdot a$
- $O_3: x_{i-1} \cdot x_i \cdot x_{i+1} \rightarrow x_{i-1} \cdot x_{i+1}$

The edit distance can be calculated using dynamic programming time

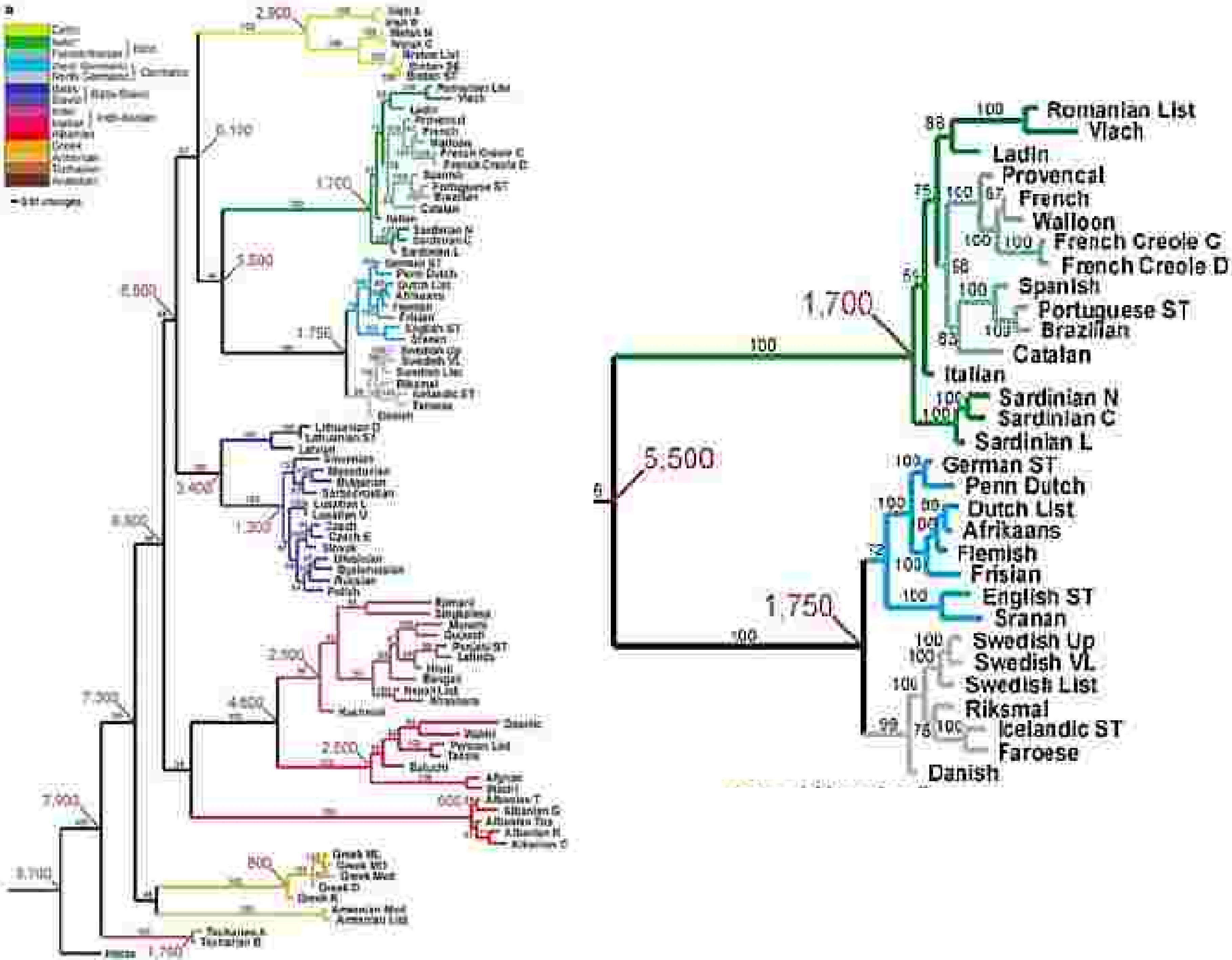
La distance d'édition peut être calculée avec la programmation dynamique en temps $O(n^2)$

$x = 01100101$
$y = 01101010$
$H(x, y) = 4$

$01100101 = x$
0110101
$01101010 = y$
$E(x, y) = 2$

hierarchical

Sheet



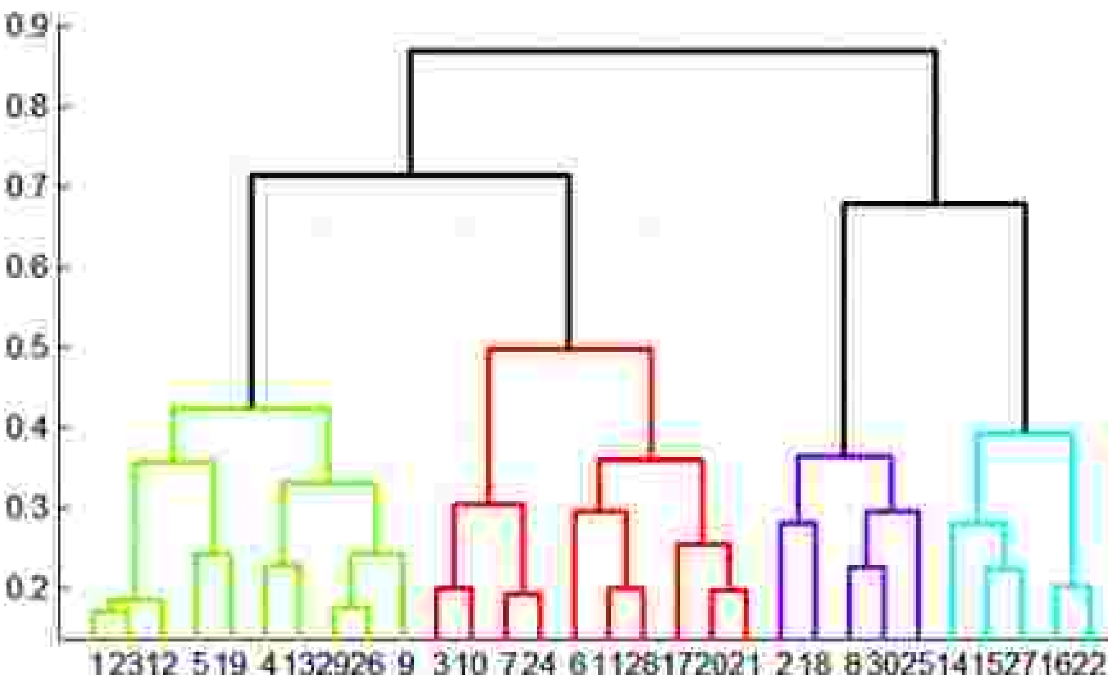
Romanian List
Vlach
Ladin
Provencal
French
Walloon
French Creole C
French Creole D
Spanish
Portuguese ST
Brazilian
Catalan
Italian
Sardinian N
Sardinian C
Sardinian L
German ST
Penn Dutch
Dutch List
Afrikaans
Flemish
Frisian
English ST
Sranan
Swedish Up
Swedish VL
Swedish List
Riksmal
Icelandic ST
Faroese
Danish
1,700
5,500
1,750
100

binary hierarchical partition

key operation: **Combine** the two closest groups.

Two important questions:

- How to represent a group?
- How do you determine the "closeness" of the groups?



hierarchical binary partition algorithm

$X = \{x_1, x_2, \dots, x_n\}$ avec les $x_i \in U$ Les éléments

Elements For

Pour $x_i \in U, I(x_i) = s_i \in S$ Initialisation

initialization

Pour $s_1, s_2 \in S, J(s_1, s_2) = s' \in S$ Joindre deux groupes

For Attach two groups

Pour $s_1, s_2 \in S, D(s_1, s_2) = d \in \mathbb{R}^+$ La distance entre deux groupes

For The distance between two groups

CBH(X)

CBH

$G = \{I(x_1), \dots, I(x_n\})$

Faire $n-1$ fois:

Make time:

Trouver $s_i, s_j \in G$ avec $D(s_i, s_j)$ minimale

Find minimally Show

Afficher (i, j)

$s_i = J(s_i, s_j)$

Retirer s_j de G

Take back from

hierarchical partition algorithm

binary

The naive implementation is in and an implementation using a priority queue can reduce the time but it's impractical for very large data sets.

L'implémentation naïve est dans $O(n^3)$ et une implémentation utilisant une file d'attente prioritaire peut réduire le temps à $O(n^2 \log n)$ mais cela reste impraticable pour les très gros jeux de données.

Example

Elements For

$X = \{x_1, x_2, \ldots, x_n\}$ avec les $x_i \in U$ Les éléments

initialization

Pour $x_i \in U, I(x_i) = s_i \in S$ Initialisation

For Attach two groups

Pour $s_1, s_2 \in S, J(s_1, s_2) = s' \in S$ Joindre deux groupes

For The distance between two groups

Pour $s_1, s_2 \in S, D(s_1, s_2) = d \in \mathbb{R}^+$ La distance entre deux groupes

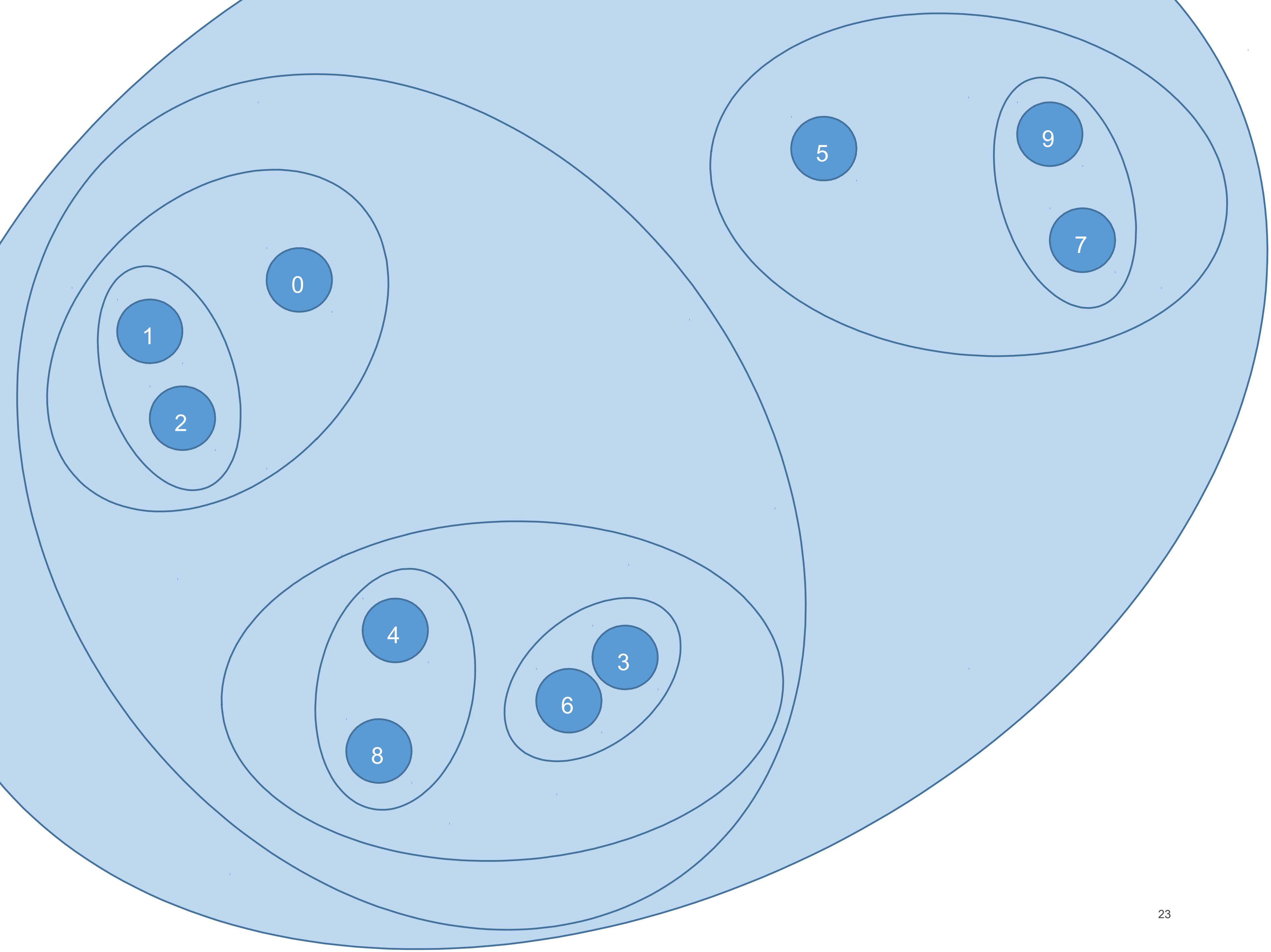
In the Euclidean case, a group can be represented by its cardinality and its mean.

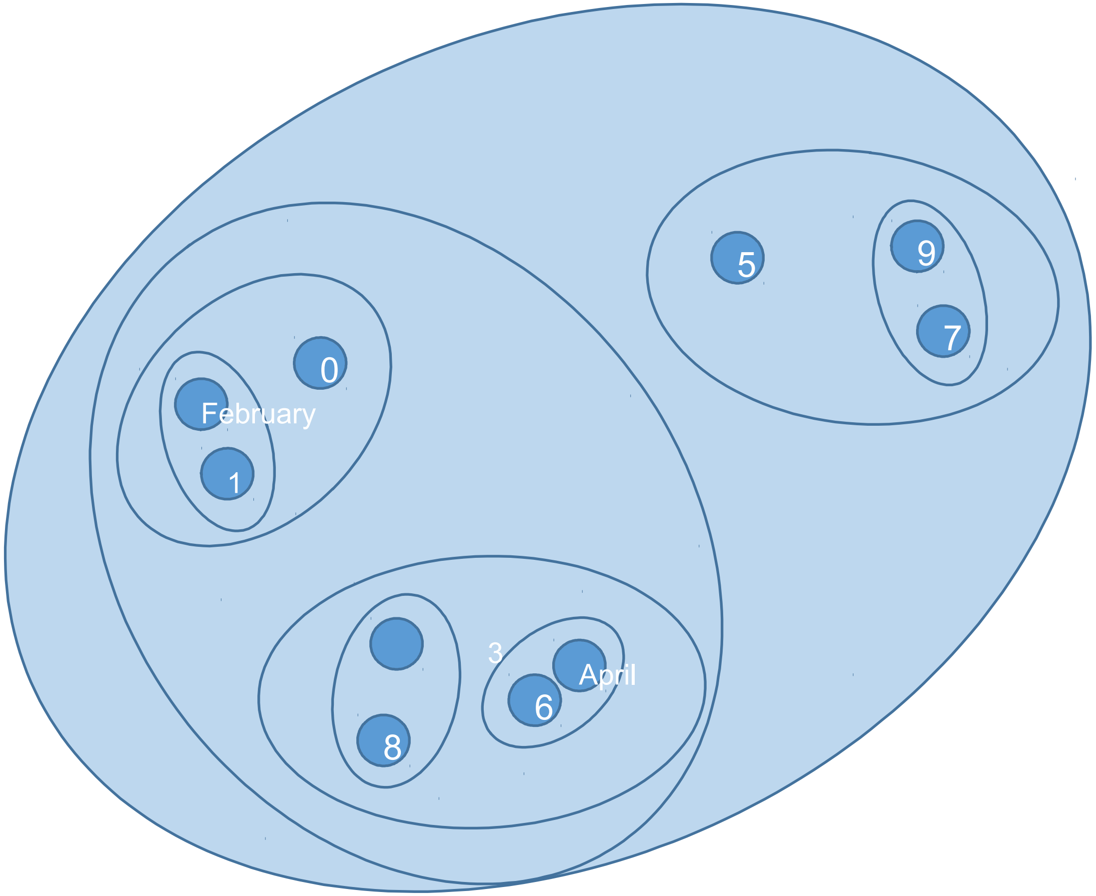
Dans le cas euclidien, un groupe peut être représenté par sa cardinalité et sa moyenne.

$$I(x_i) = [1, x_i]$$

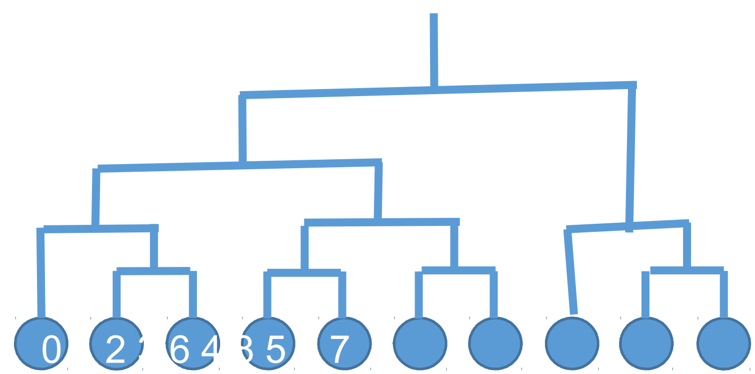
$$J([n_1, m_1], [n_2, m_2]) = \left[n_1 + n_2, \frac{n_1 m_1 + n_2 m_2}{(n_1 + n_2)} \right]$$

$$D([n_1, m_1], [n_2, m_2]) = L_2(m_1, m_2)$$


0
1
2
3
4
5
6
7
8
9



0
February
1
5
9
7
3
April
6
8



0
2
6
4
5
7

non-Euclidean case

We can not calculate average, but still has a notion of similarity between the elements.

approach 1

Choose a representative for each grouping. For example the one that minimizes the distance with the other members of the grouping. The similarity groups is obtained by calculating the similarity of Representatives.

approach 2

Use the mean similarity between peer element from each group.

approach 3

Select the combination that maximizes cohesion. Use the diameter of the merged group. The diameter is the maximum distance between the points of the group. Using the average distance between the points of the group. Density can also promote dividing by the number of points.

Partisions

Sheet

The set of elements is partitioned.

L'ensemble d'éléments à partitionné.

$X = \{x_1, x_2, \dots, x_n\}$

The sheet music for a chosen or learned.

Les partitions pour un k choisit ou appris.

$\{G_1, G_2, \dots, G_k\}$

$$X = \bigcup_{i=1}^{k} G_i$$

$\forall i \neq j, G_i \cap G_j = \emptyset$

Requires notion of similarity (or distance) between the elements.

Nécessite une notion de similarité (ou de distance) entre les éléments.

$S: X \times X \to R$

Useful to understand the data and that schematize and to make decisions.

Useful to apply a solution based on personality groups obtained.

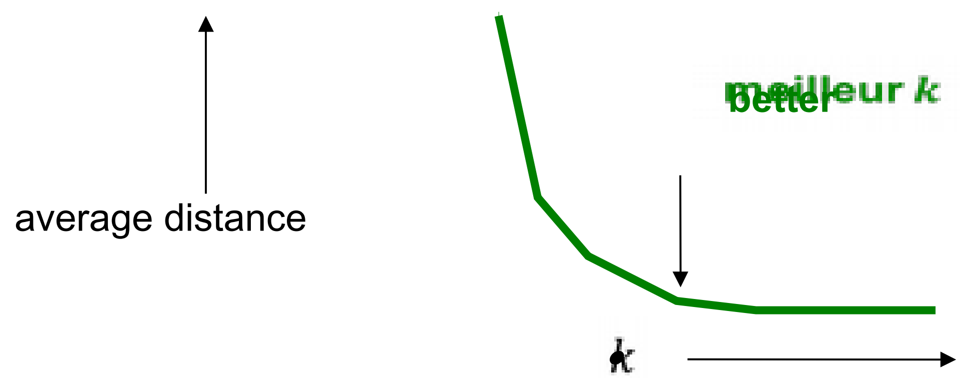
Utile pour comprendre les données et les schématiser ainsi que pour prendre des décisions. Utile pour appliquer une solution personnalité en fonction des groups obtenue.

Choose the number of group (k)

Use group cohesion.

Use the diameter of the merged group.

Using the average distance between the points of the groups. Using the average distance between the points of the group and their centers.



expectation-algorithm maximization

The expectation-maximization algorithm is an iterative algorithm to find the parameters of maximum likelihood of a probabilistic model when it depends on unobserved latent variables.

The expectation-maximization algorithm includes:

- an evaluation expectancy step (E), which calculates the likelihood expectancy taking into account the last observed variables,
- a maximization step (M), wherein the maximum likelihood parameter is estimated by maximizing the likelihood found in step E.

It then uses the settings found in M as the starting point of a new phase of assessment of hope, and it iterates well.

K-moyenne

-average

K-moyenne

- average

When the data belong to a Euclidean space, k-mean algorithm allows a set divided into groups for a given.

- Lorsque les données appartiennent à un espace euclidien, l'algorithme k-moyenne permet de divisé un ensemble en k groupes pour un k donné.

$X = \{x_1, x_2, \dots, x_n\}$ avec les $x_i \in \mathbb{R}^d$

$D(x, y)$ la distance euclidienne sur $\mathbb{R}^d$

Euclidean distance

Kmean ()

KMEAN(X, k)

To choose centers for groups.

111 to:

2. Return 1 to 1 if the **stop condition** is not satisfied. Return

Choisir $C = \{m_1, \dots, m_k\}$ avec les $c_i \in \mathbb{R}^d$ les centres pour les k groups.

1. Pour i de 1 à k : $G_i = \{x \mid \forall j \neq i, D(x, m_i) \leq D(x, m_j)\}$
2. Pour i de 1 à k : $m_i = \frac{1}{|G_i|} \sum_{x \in G_i} x$

Retourner a 1 si la **condition d'arrêt** n'est pas satisfaite.

Retourner $\{G_1, \dots, G_k\}$

- average

K-moyenne

To choose initializing centers

Choisir l'initialisation des centres

Choose items randomly and evenly assigned centers with these values.

Choisir k élément uniformément au hasard et affecter les k centres avec ces k valeurs.

Select an item at random, for next item, choose each time point that maximizes the distance to the points already selected.

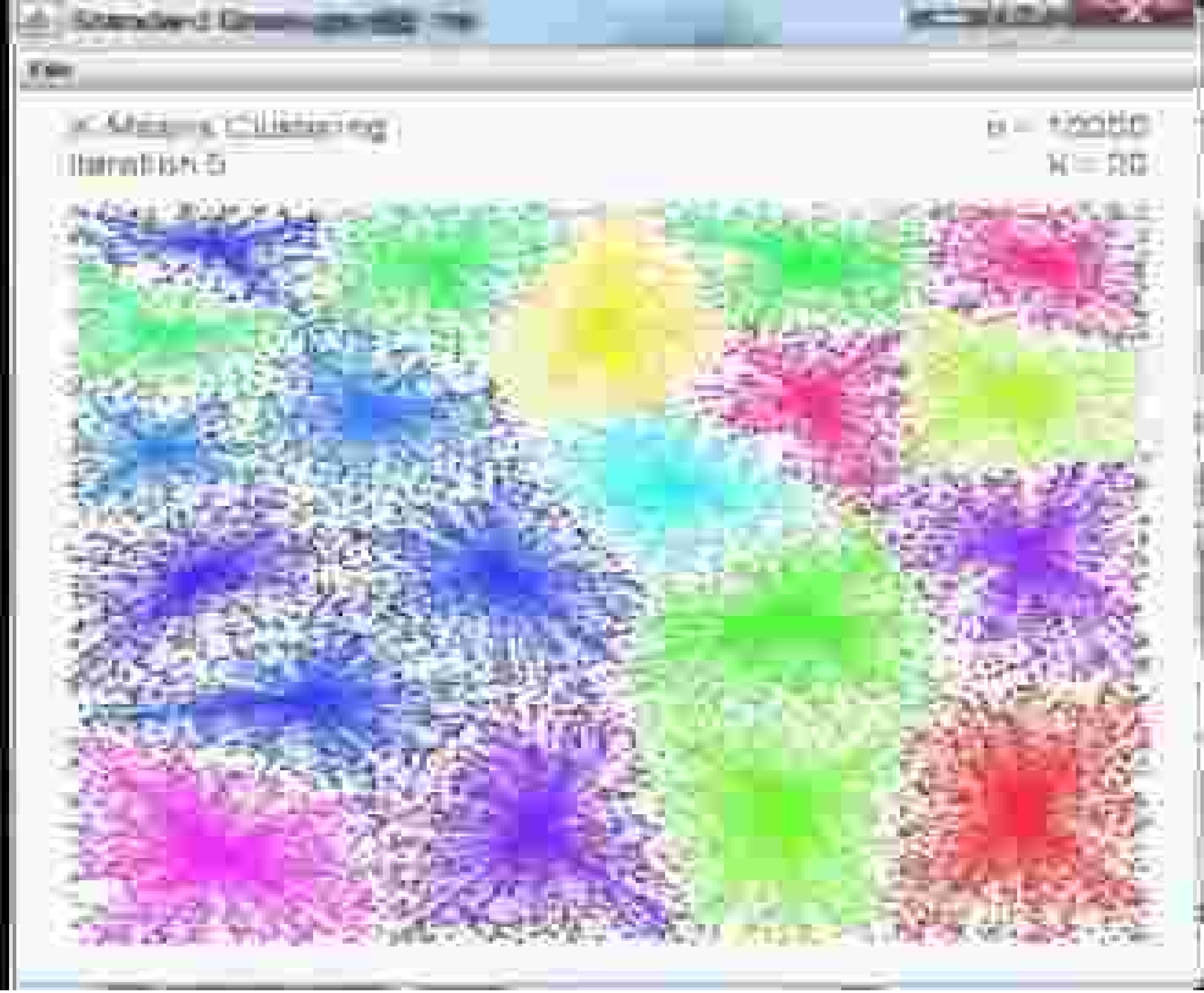
Choisir un élément au hasard, pour les $k-1$ élément suivant, choisir a chaque fois le point qui maximise la distance avec les points déjà choisis.

Stop condition of the algorithm.

Condition d'arrêt de l'algorithme.

The new centers are identical to those of the previous iteration.

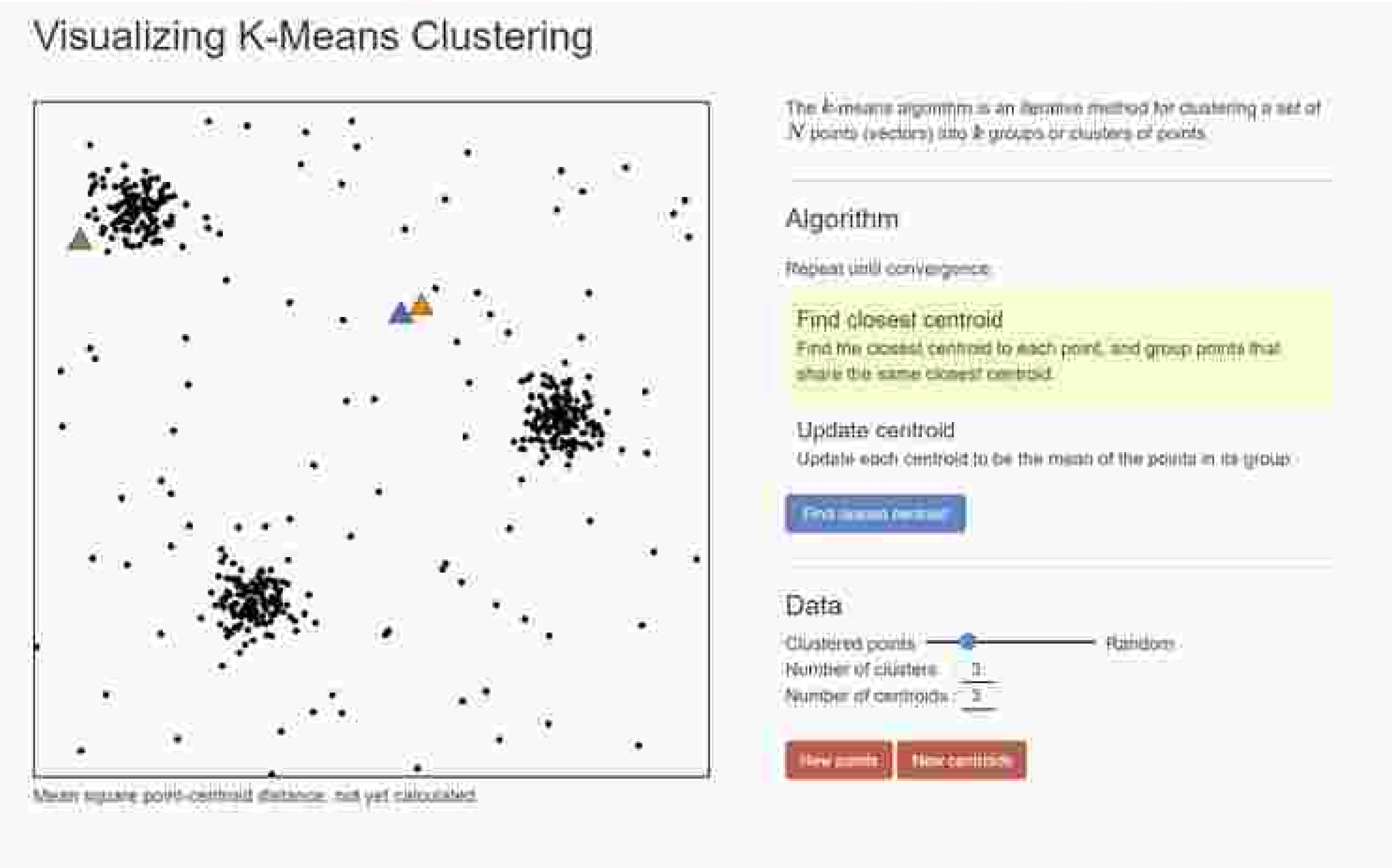
Les nouveaux centres sont identiques à ceux de l'itération précédente.



Bishma Stornelli

Published on March 22, 2014

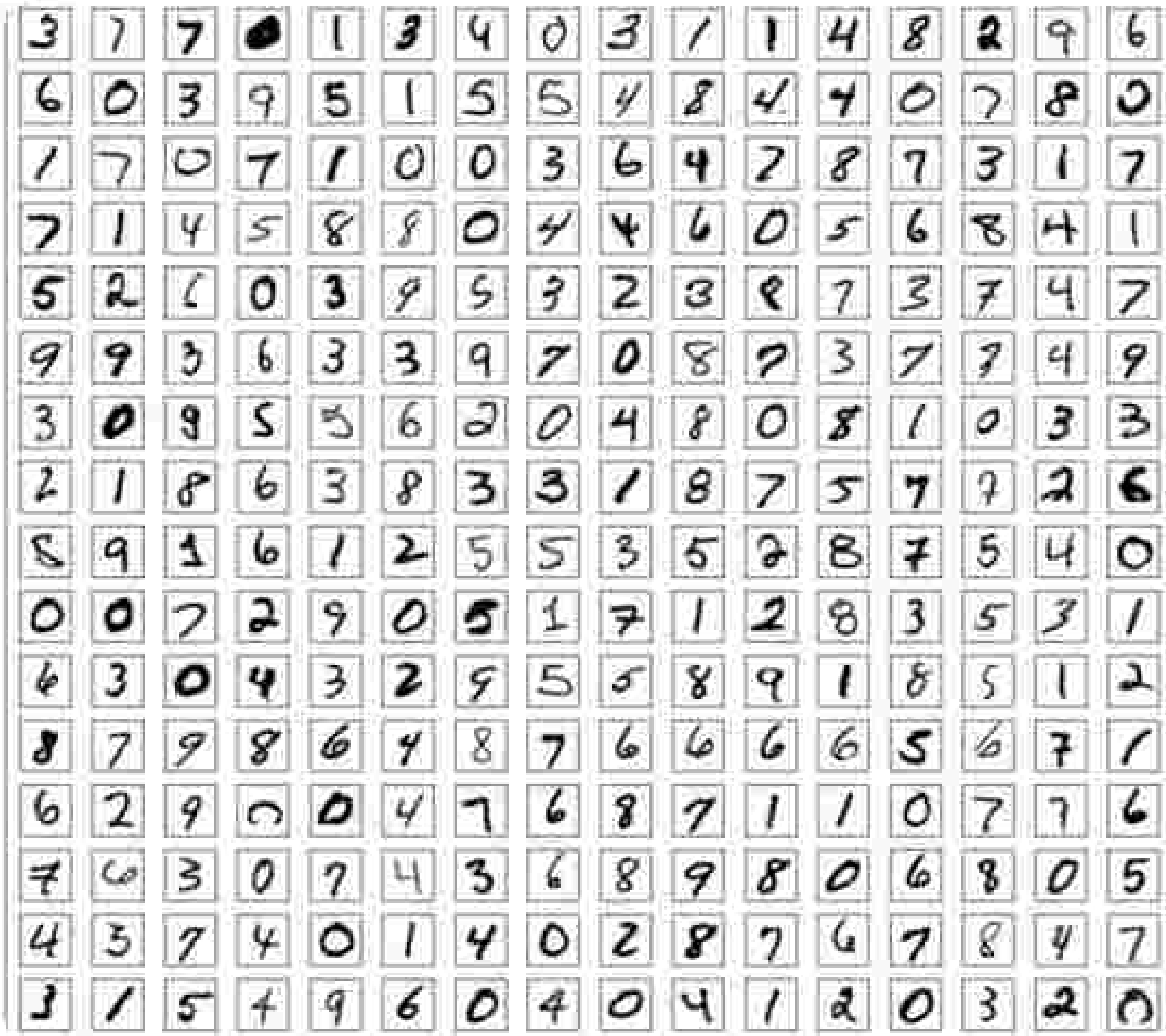
A simple example of a real-time simulation of the K-Means Clustering Algorithm using different values for n and k.



http://web.stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

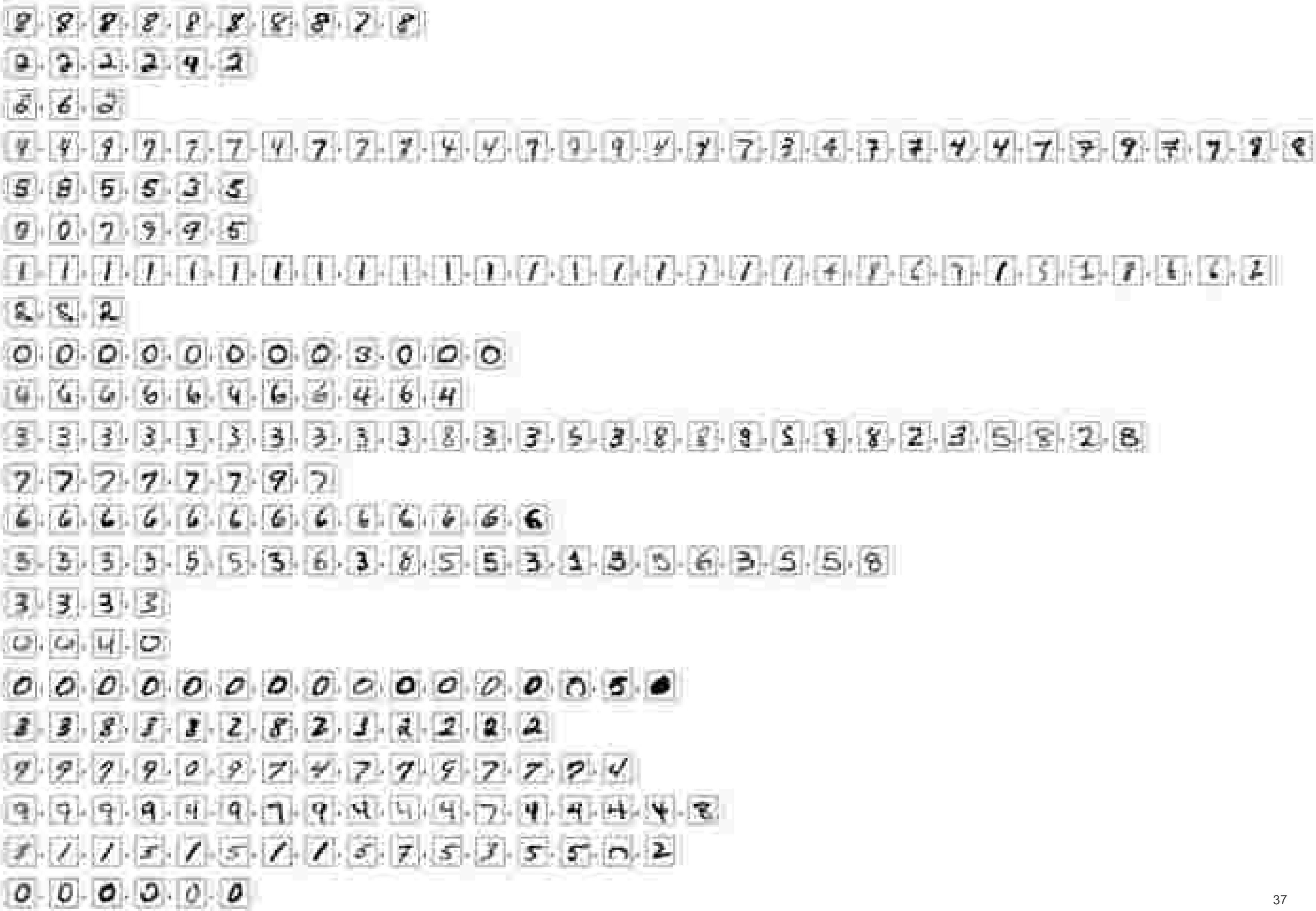
256 images dans $k = 22$ groupes

images in groups



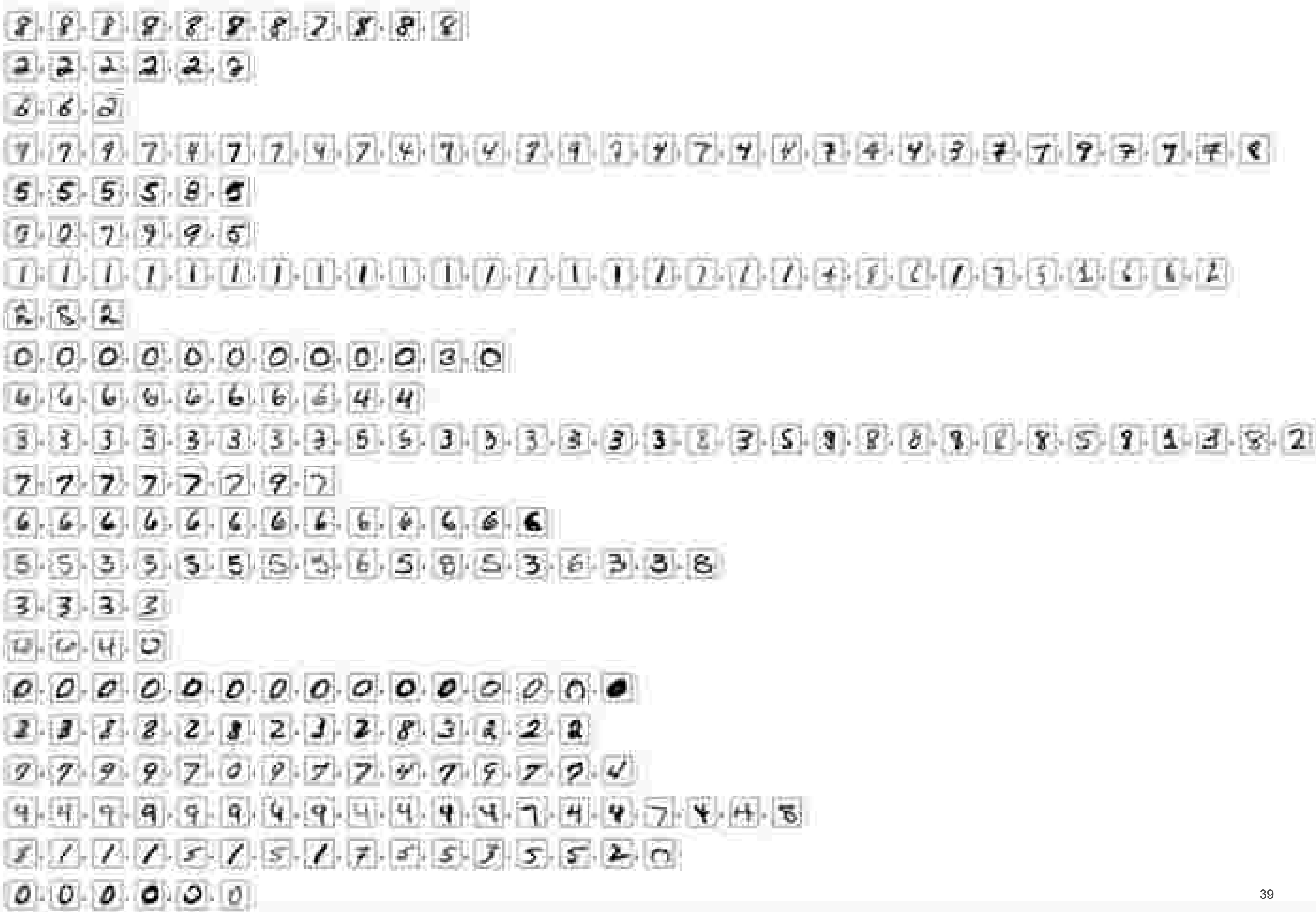
iteration 1

iteration 2



iteration 5

iteration 8

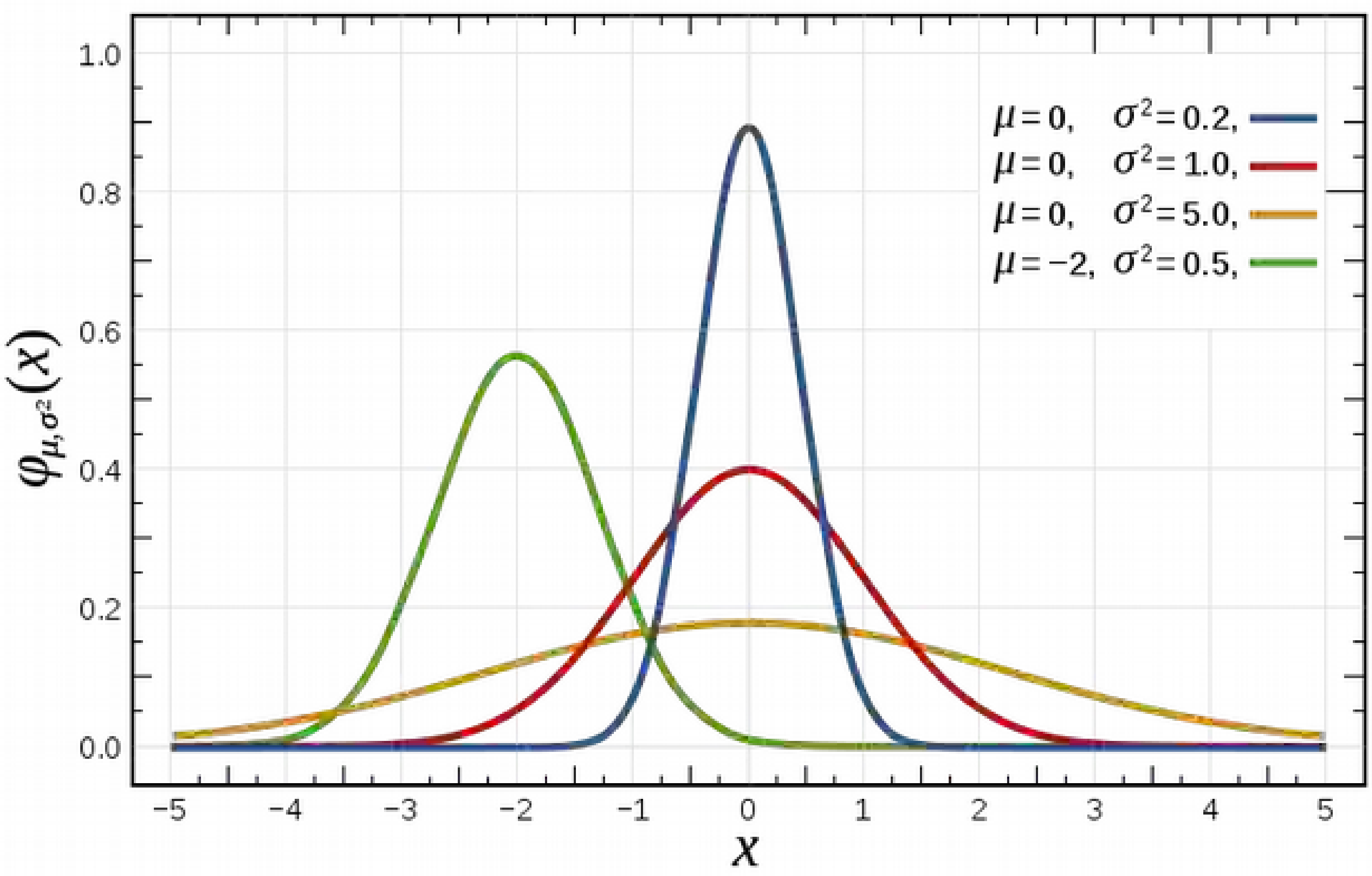


GMM

Gaussian Mixture Model

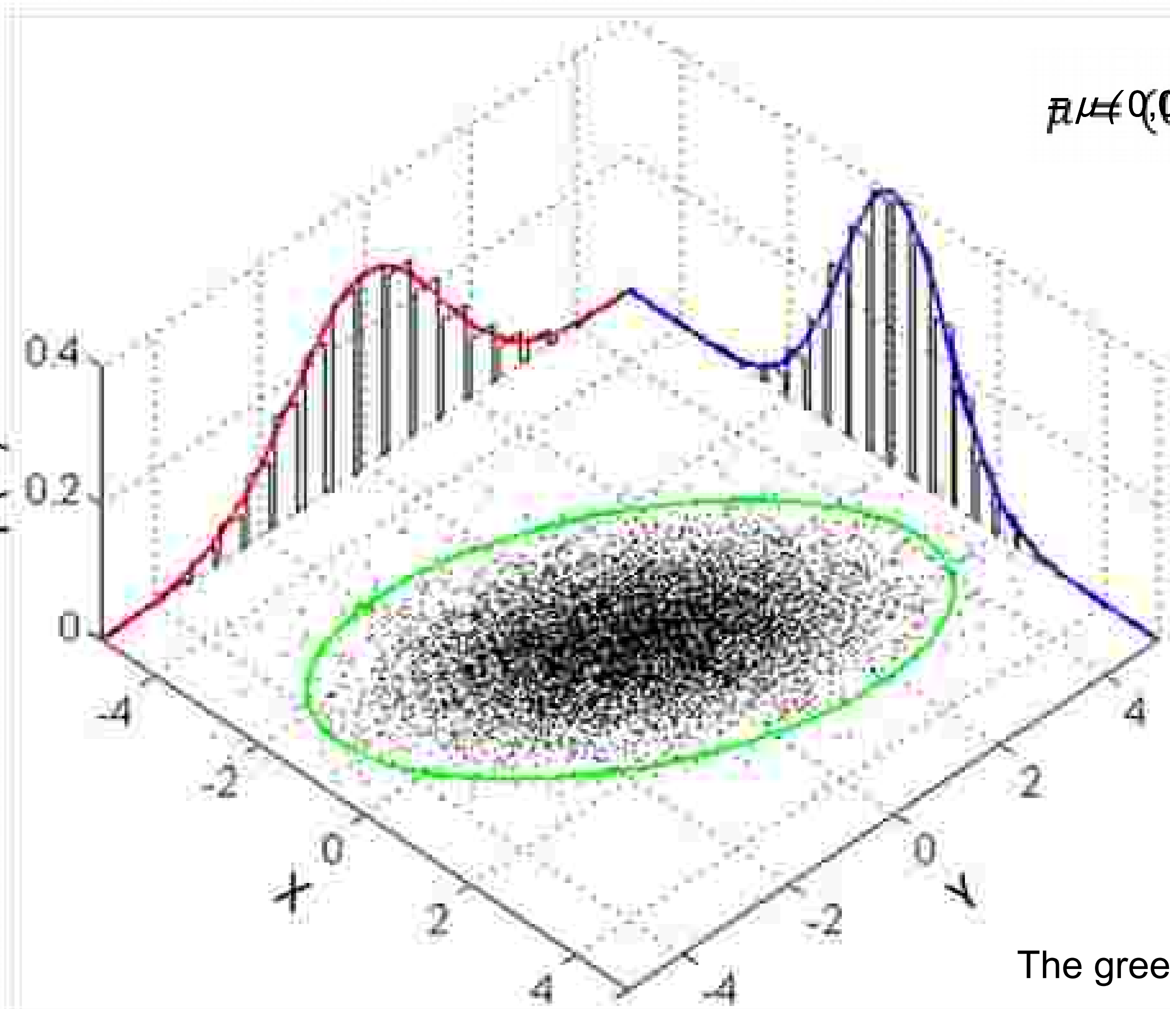
normal mixture

Normal law



$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \quad \sigma^2=0.5,$
$\varphi_{\mu,\sigma^2}(x)$
x
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5

multivariate normal distribution



$\mu = (0,0)$, $\Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 2 \end{pmatrix}$

The green ellipse is 3-sigma.

statistical Reminder

the data

$$X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^n$$
$$Y = \{y_1, y_2, \dots, y_n\} \in \mathbb{R}^n$$

The average

$$\mu_X = \frac{1}{n}\sum_{I=1}^{n} x_i$$

The variance

$$\sigma_X = \frac{1}{n}\sum_{I=1}^{n} (x_i - \mu_X)^2$$

covariance

$$\sigma_{X,Y} = \frac{1}{n}\sum_{I=1}^{n} (x_i - \mu_X)(y_i - \mu_Y)$$

Correlation coefficient

$$r_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_y}$$

covariance matrix

$X = \{x_1, x_2, \dots, x_n\}$ avec les $x_i \in \mathbb{R}^d$

The $\mathrm{Cov}(X)_{\mathrm{ij}} = \sigma_{i,j}$

$d = 3$

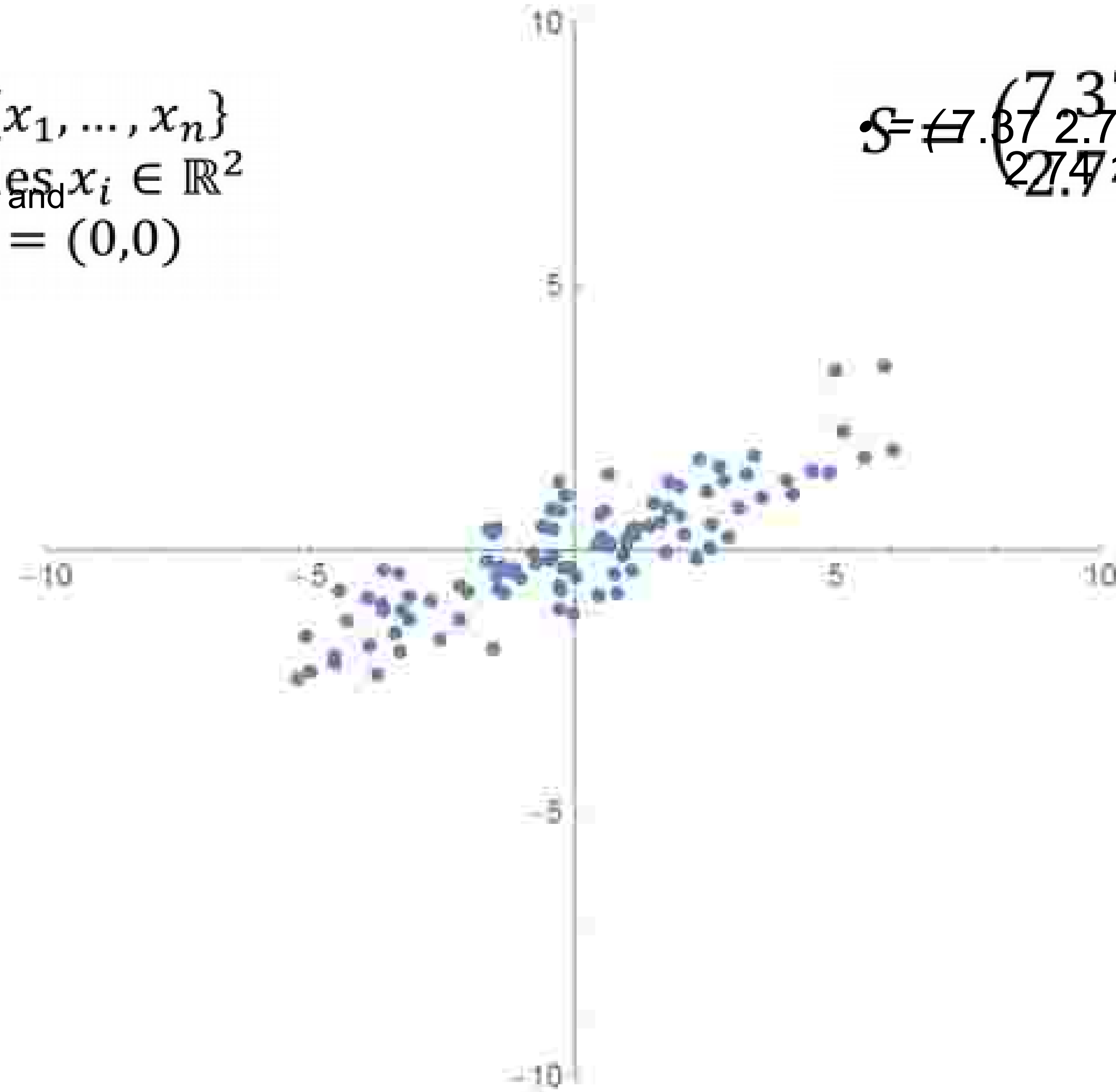
$$\begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_{3,3} \end{pmatrix} = \begin{pmatrix} \sigma_{1} & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_{2} & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_{3} \end{pmatrix}$$

covariance

$X = \{x_1, \dots, x_n\}$
avec les $x_i \in \mathbb{R}^2$
and
et $\mu_X = (0,0)$

$S = \begin{pmatrix} 7.37 & 2.74 \\ 2.74 & 1.39 \end{pmatrix}$

• S = (7.37 2.74
2.74 1.39)



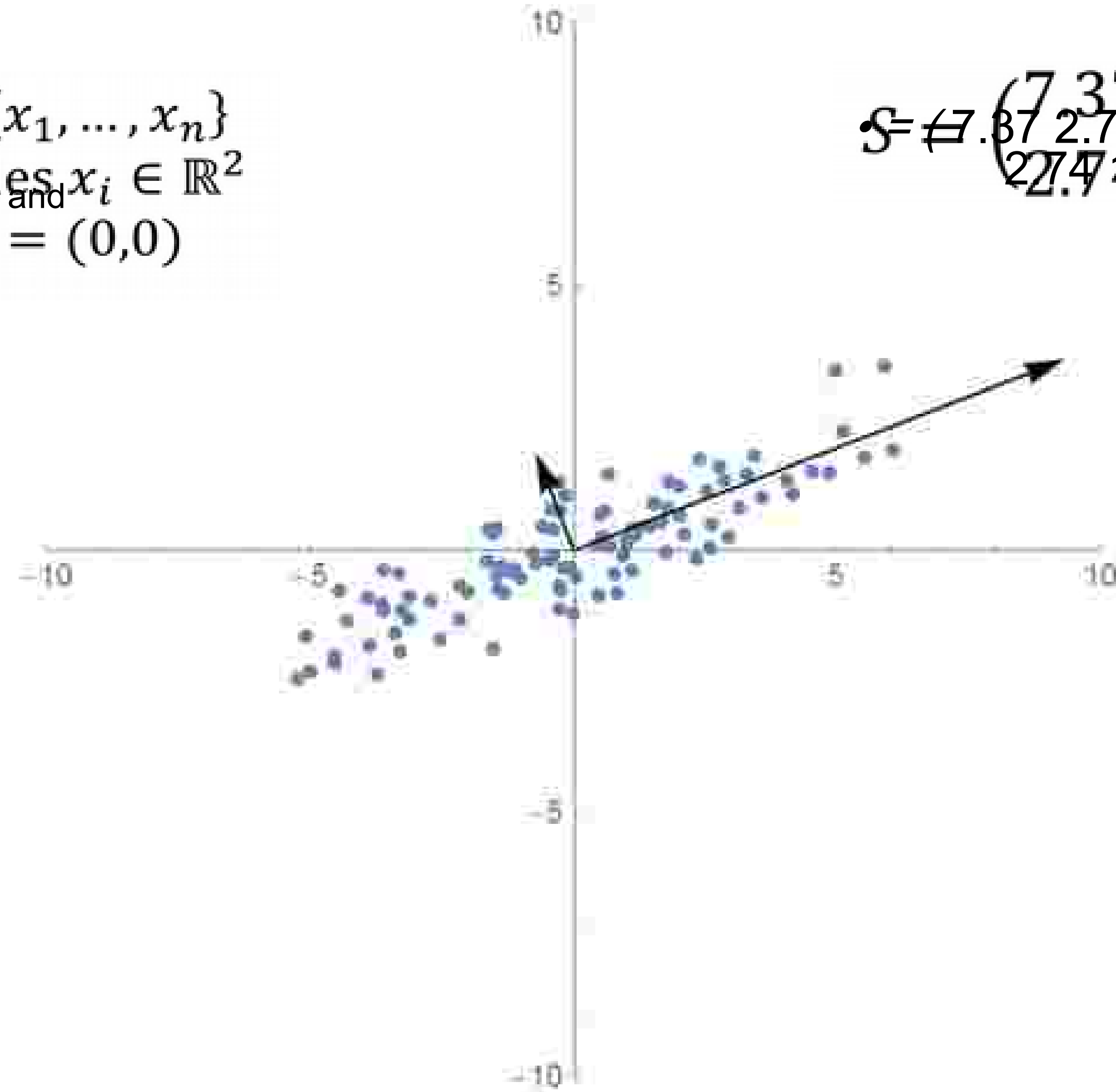
Les vecteurs proper de S

The proper vectors

$X = \{x_1, \dots, x_n\}$
avec les $x_i \in \mathbb{R}^2$
and
et $\mu_X = (0,0)$

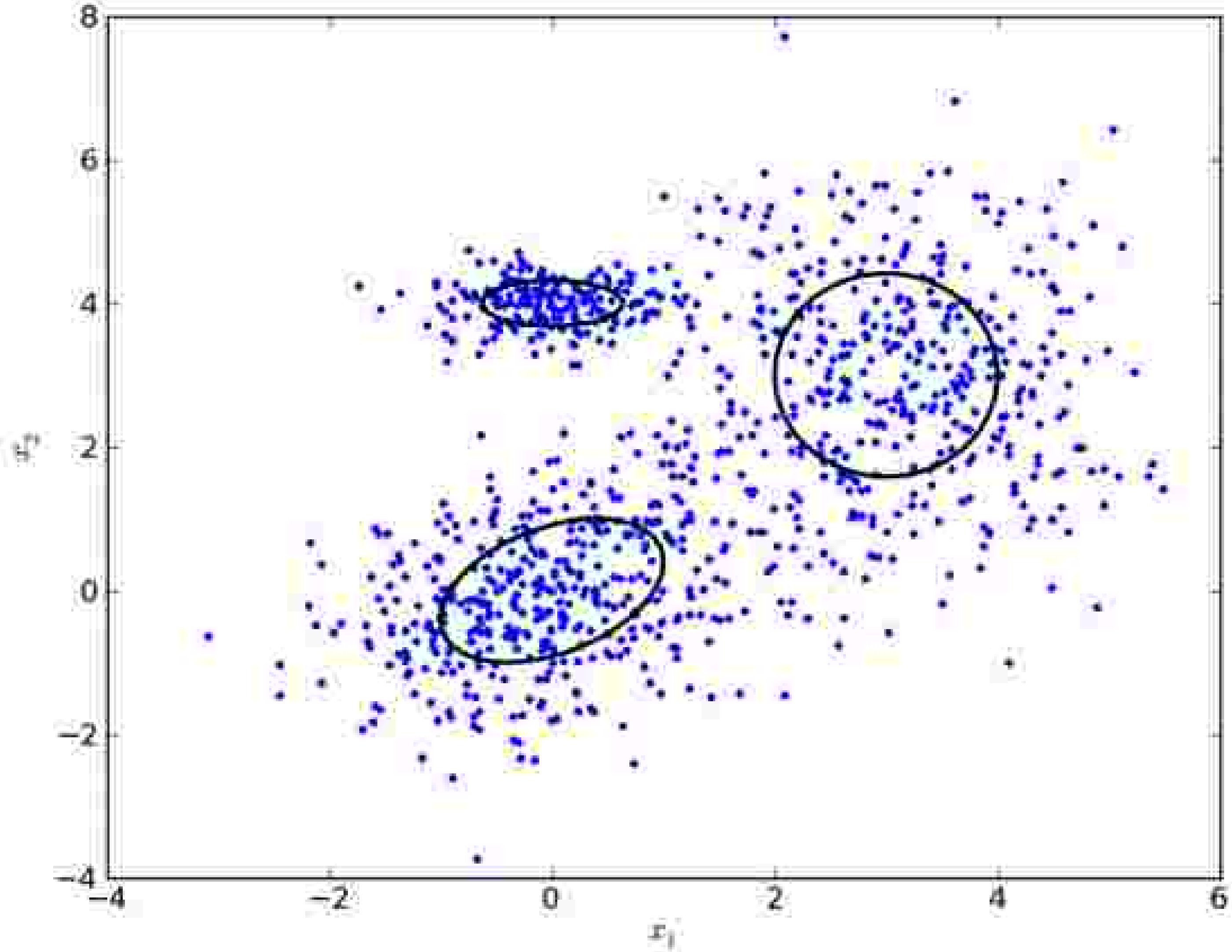
$S = \begin{pmatrix} 7.37 & 2.74 \\ 2.74 & 1.39 \end{pmatrix}$

• $S = \begin{pmatrix} 7.37 & 2.74 \\ 2.74 & 1.39 \end{pmatrix}$

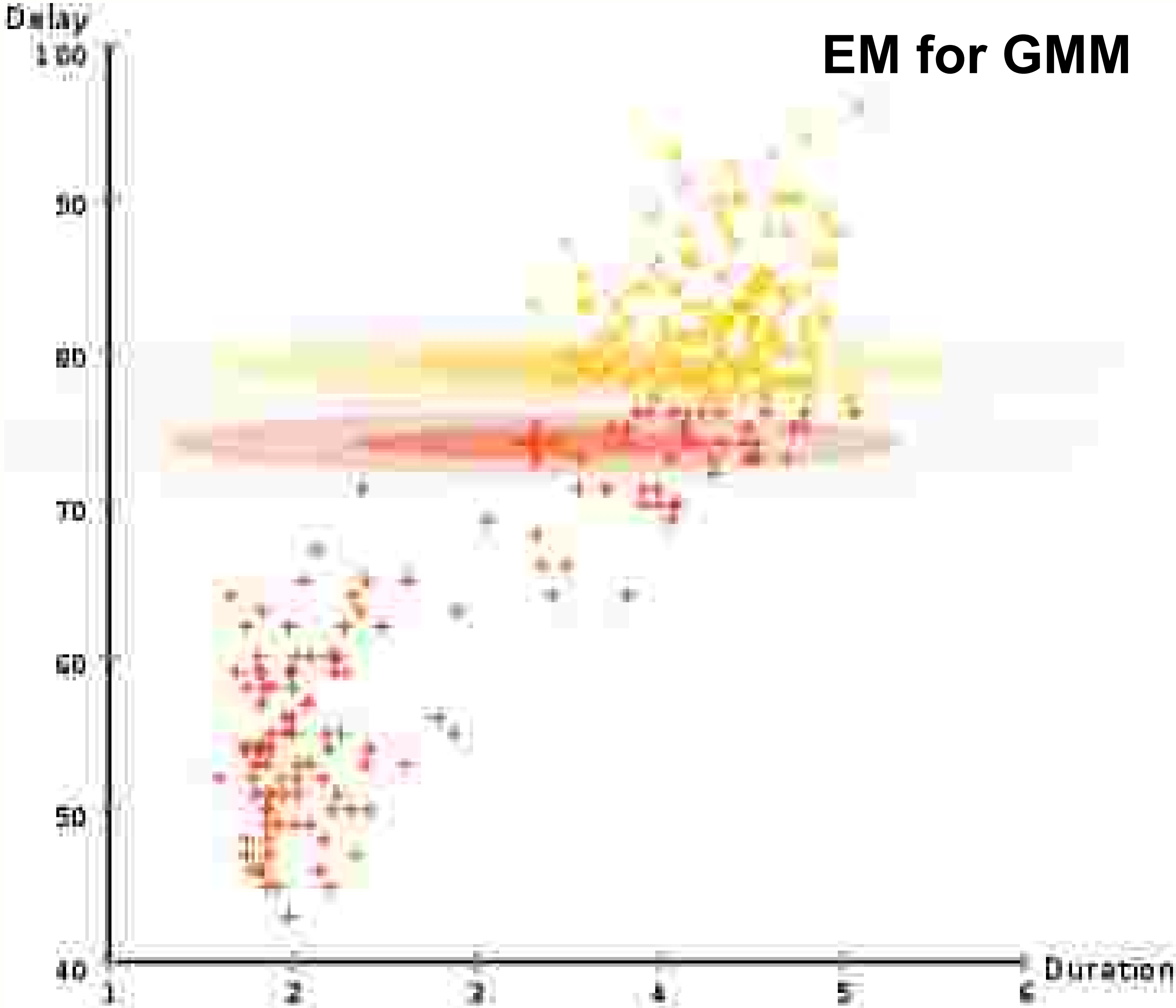


Gaussian Mixture Model

normal mixture



EM for GMM



GMM

$X = \{x_1, x_2, \dots, x_n\}$ avec les $x_i \in \mathbb{R}^d$

$\mathcal{G} = \{(p_1, \mu_1, S_1), (p_2, \mu_2, S_2), \dots, (p_k, \mu_k, S_k)\}$

$$\sum_{i=1}^{k} p_i = 1$$

$$\mathcal{N}(x|\mu, S) = \frac{1}{\sqrt{(2\pi)^d |S|}} exp(-\frac{1}{2}(x-\mu)^T S^{-1}(x-\mu))$$

Vraisemblance d'un élément

Likelihood of an element

$$v(x) = \sum_{i=1}^{k} p_i \mathcal{N}(x|\mu_i, S_i)$$

Trouver $\mathcal{G}$ qui maximise la vraisemblance des donnés

$\sum_{i=1}^{n} v(x_i)$

Find that maximizes the likelihood given

La probabilité pour un point x_i d'appartenir au groupe g est donné par

$$p_{i,g} = \frac{p_i \mathcal{N}(x_i|\mu_g, S_g)}{v(x_i)}$$

The probability for a point to belong to the group is given by

EM for GMM

$X = \{x_1, x_2, \dots, x_n\}$ avec les $x_i \in \mathbb{R}^d$

$\mathcal{G} = \{(p_1, \mu_1, S_1), (p_2, \mu_2, S_2), \dots, (p_k, \mu_k, S_k)\}$

GMM(X, k)

(GMM)

Choisir uniformément $(\mu_1, \mu_2 \dots, \mu_k)$, un sous-ensemble de X

Initialiser les $S_i = S$ et les $p_i = \frac{1}{k}$

uniformly choosing a subset of the Initialiser and

Espérance

Hope Pour tout x_i et g calculer $\gamma_{i,g}$ la probabilité que x_i soit dans g.

For and calculate the probability that either.

$$\gamma_{i,g} = \frac{p_g \mathcal{N}(x_i | \mu_g, S_g)}{\nu(x_i)}$$

Maximisation

Maximization

Réévaluer les paramètres de $\mathcal{G}$

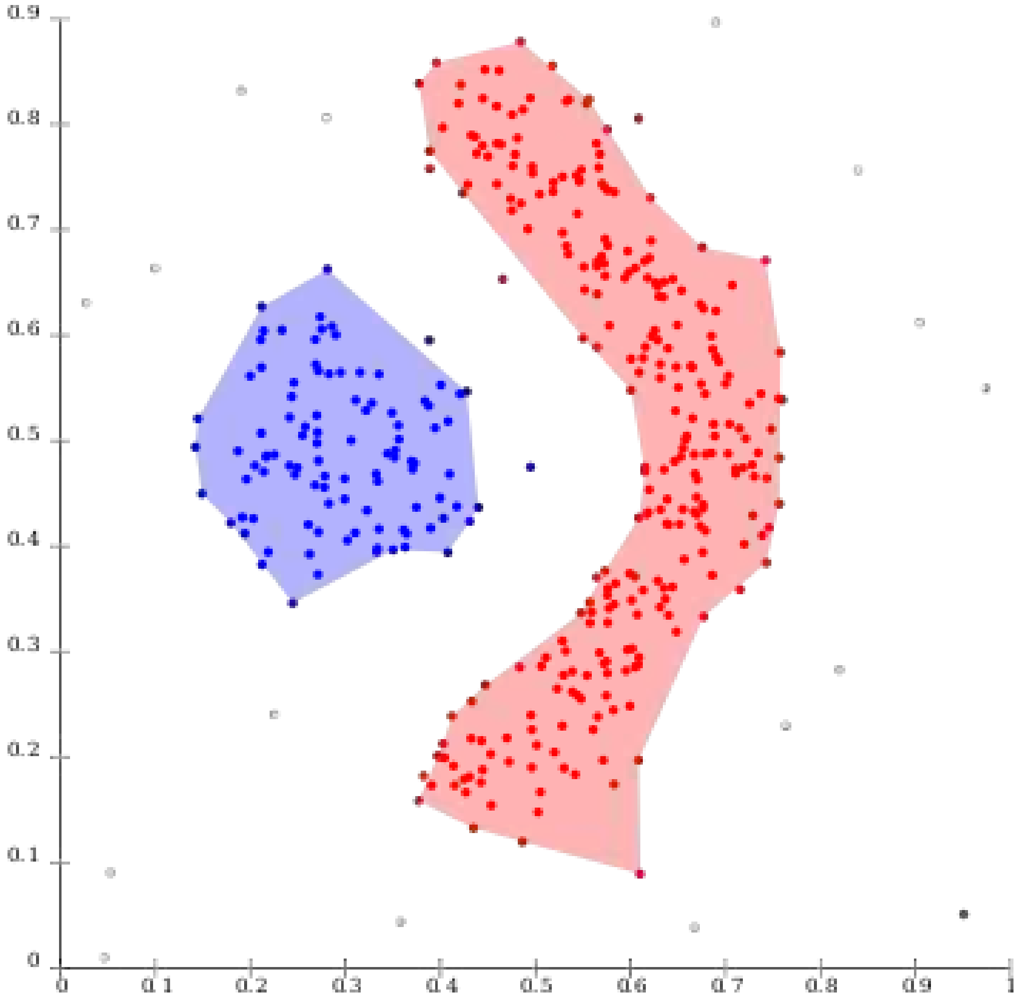
Reassess parameters For For For

Pour tout g, $p_g = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ig}$

Pour tout g, $\mu_g = \frac{\sum_{i=1}^{n} \gamma_{ig} x_i}{\sum_{i=1}^{n} \gamma_{ig}}$

Pour tout g, $S_g = \frac{\sum_{i=1}^{n} \gamma_{ig} (x_i - \mu_g)(x_i - \mu_g)^T}{\sum_{i=1}^{n} \gamma_{ig}}$

dbscan algorithm



Dbscan can find separable groups with disparate forms. This data set can not be partitioned appropriately with Kmean GMM.

Scores based on density

Dbscan is an approach to the partition based on the density. Main characteristics:

- Discover arbitrary form groups
- Manage Noise
- Effective
- Depends density parameters

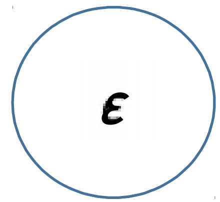
dbscan

- is the radius of the neighborhoods
- Density is the number of points in a neighborhood.
- is the critical density
- Points **internal** s, is a point with the critical density in its vicinity.
- A point **border** is not an internal point, but has an internal point in its vicinity.
- The other points are points **noise**.

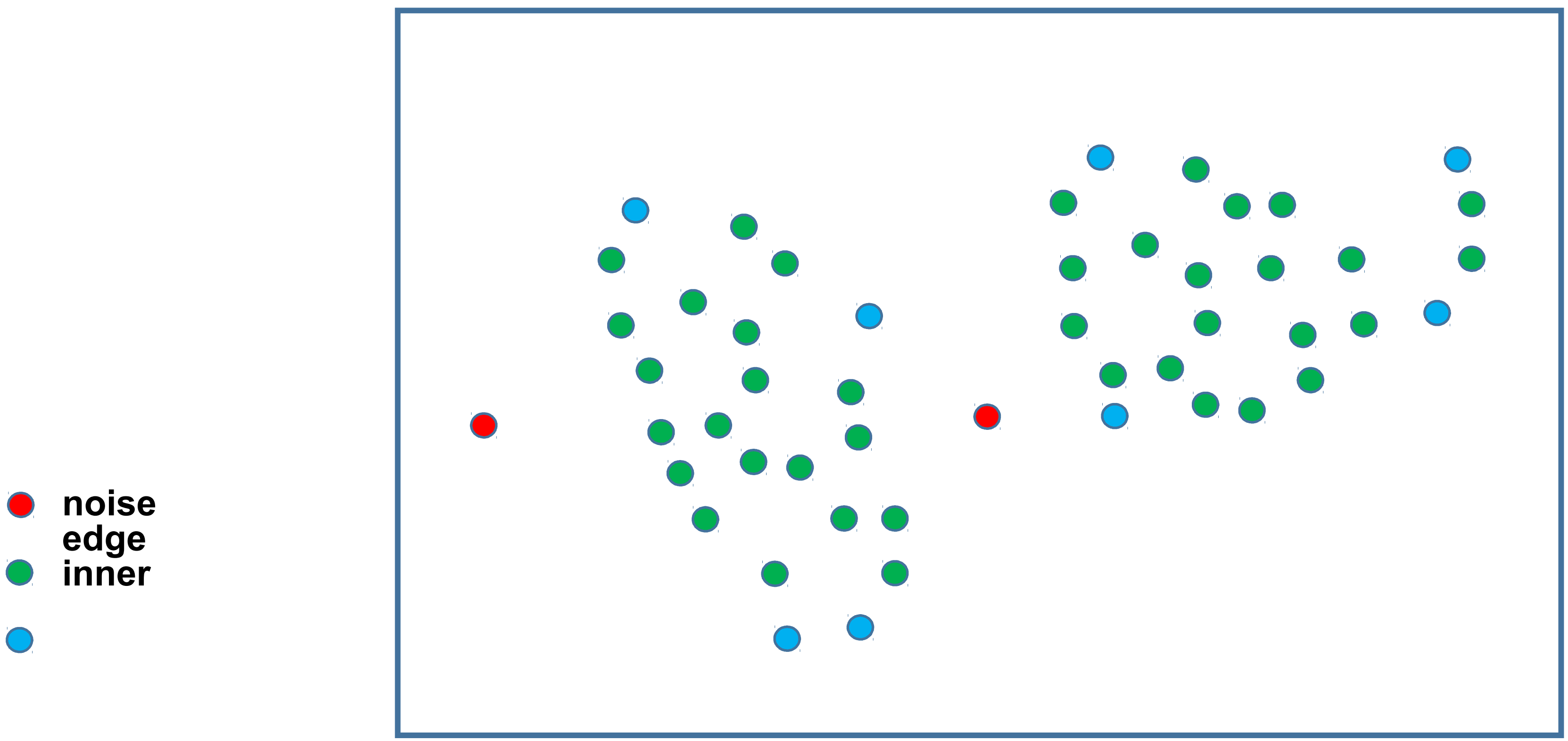
- ε est le rayon des voisinages
- La densité est le nombre de points dans un voisinage.
- Δ est la densité critique
- Points **internes**, est un point avec la densité critique dans son voisinage.
- Un point de **bordure** n'est pas un point interne, mais possède un point interne dans son voisinage.
- Les autres points sont des points de **bruit**.

Dbscan: internal, border, and noise

$\Delta = 4$



2 groups

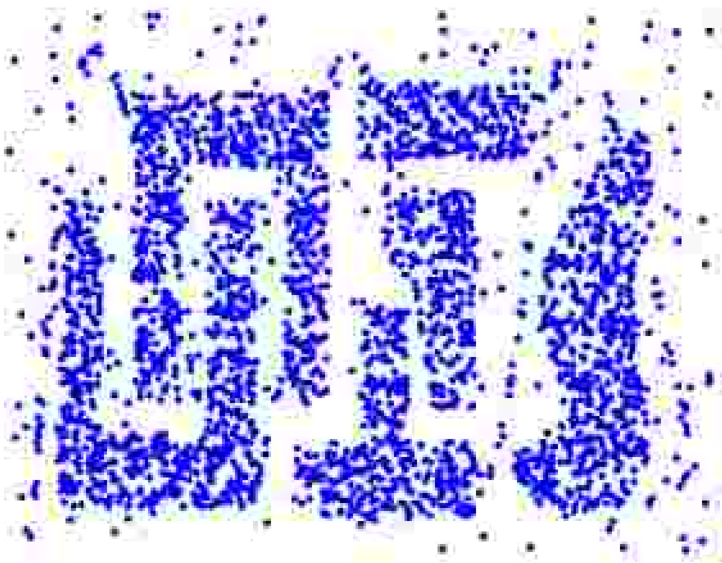


Dbscan algorithm (simplified)

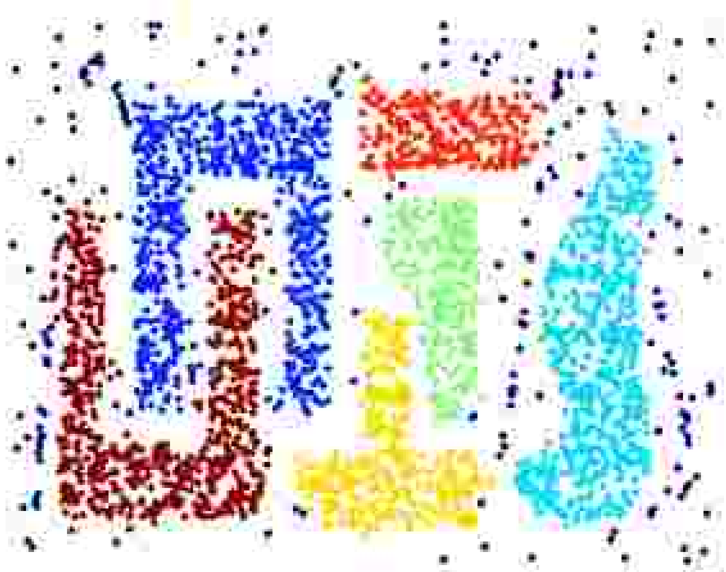
1. Select and.
2. Create a graph whose nodes are the points to regroup.
3. For each item, make a stop at each element in the vicinity.
4. Until all elements
   - Select an item. If type noise is removed. Or create a group with all the elements in its vicinity and remove the group overall.

1. Choisir Δ et ϵ.
2. Créer un graphe dont les nœuds sont les points à regrouper.
3. Pour chaque élément x, créez une arrête de x à chaque élément dans le voisinage.
4. Tant qu'il reste des éléments
   - Choisir un élément, s'il est de type bruit on l'enlève. Sinon créer un groupe avec tous les éléments dans son voisinage et retirer le groupe de l'ensemble.

Example where dbscan works well

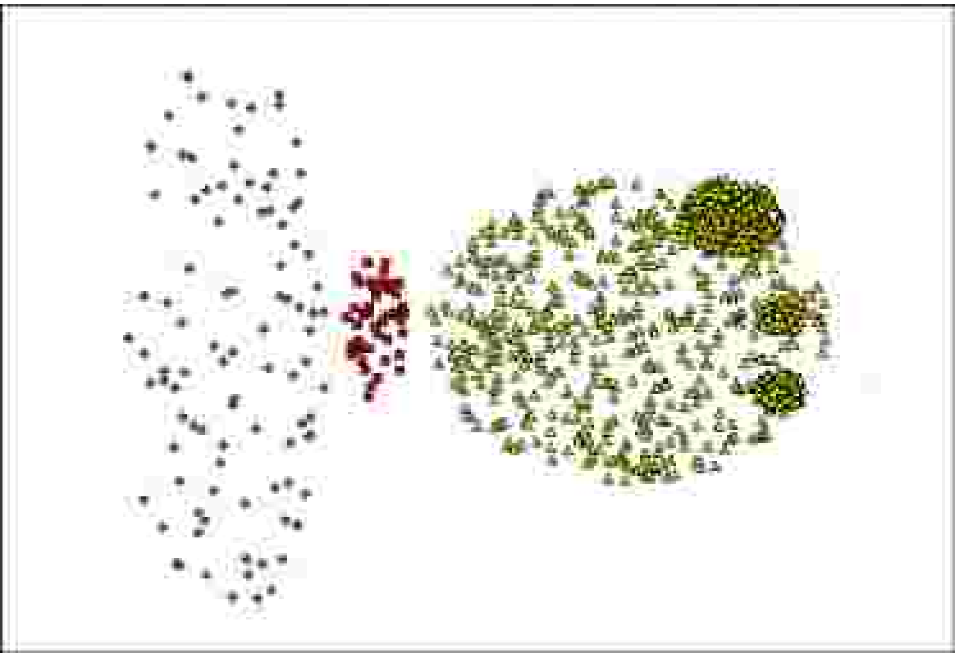


internal , border and noise

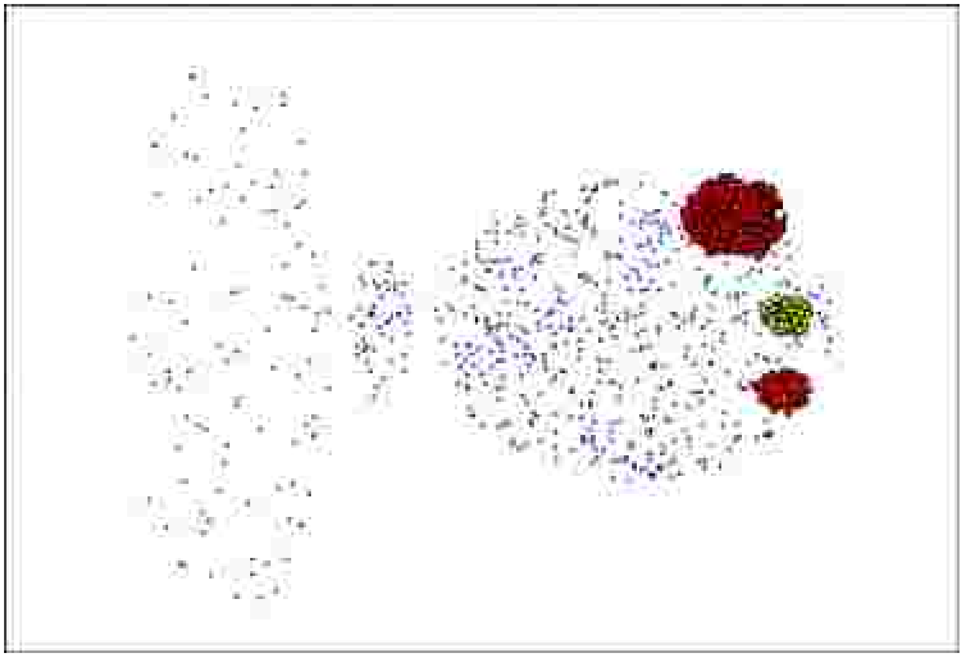


6 groups and noise

Example where dbscan works not good



$\Delta = 4, \varepsilon = 9.75$



$\Delta = 4, \varepsilon = 9.92$