Aprendizaje no supervisado

Aprendizaje Automático Aplicado

Julio Waissman

Direrentes tipos de aprendizaje no supervisado

- Reduccion de la dimensionalidad
- Análisis aglomerativo (*clustering*)
- Entrenamiento de modelos con fines de transferencia del aprendizaje

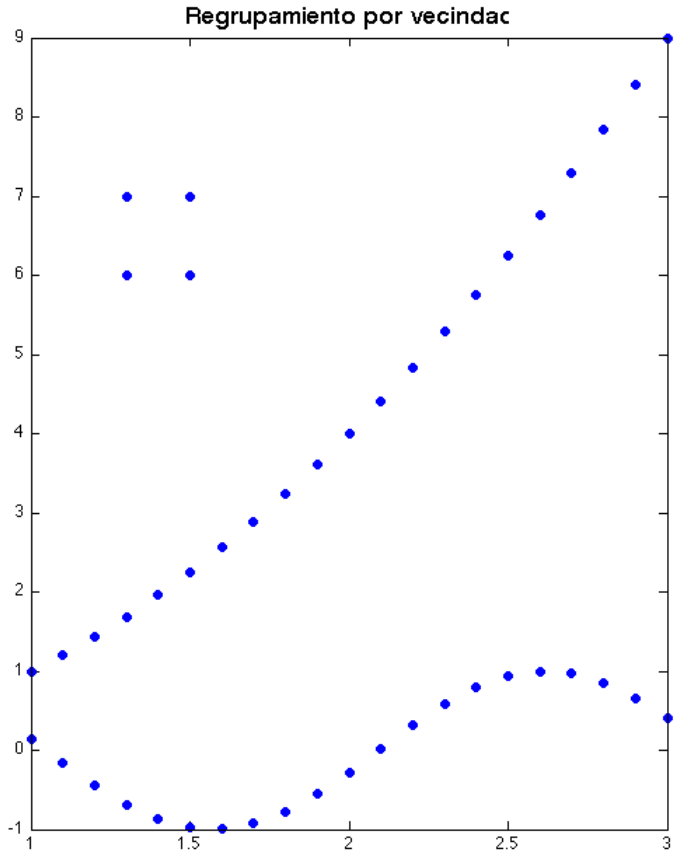
¿Que es el clustering?

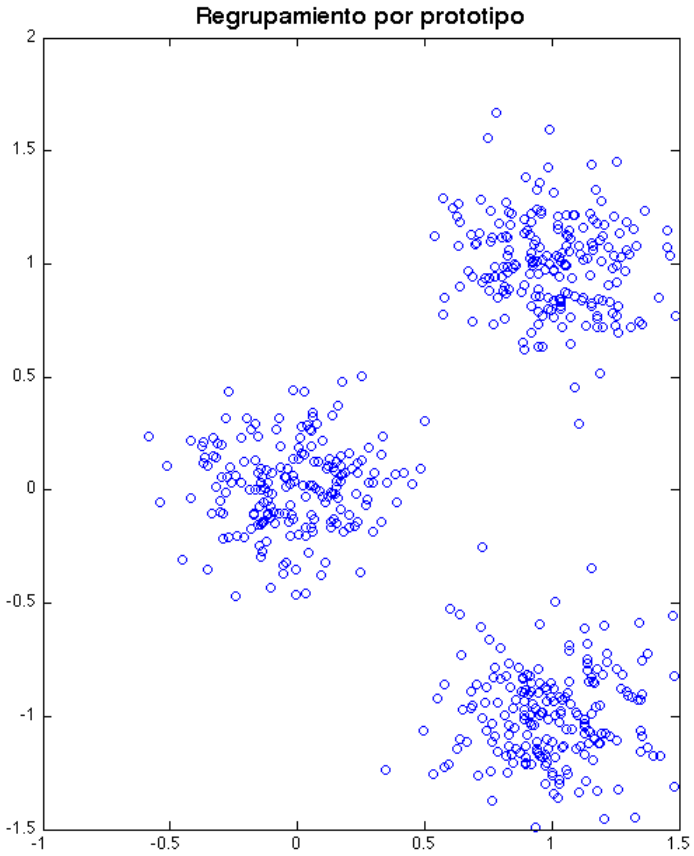
Cluster analysis

From Wikipedia, the free encyclopedia

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) to each other than to those in other groups (clusters). It is a main task of exploratory data analysis, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

¿Que es el clustering?





Algunas aplicaciones

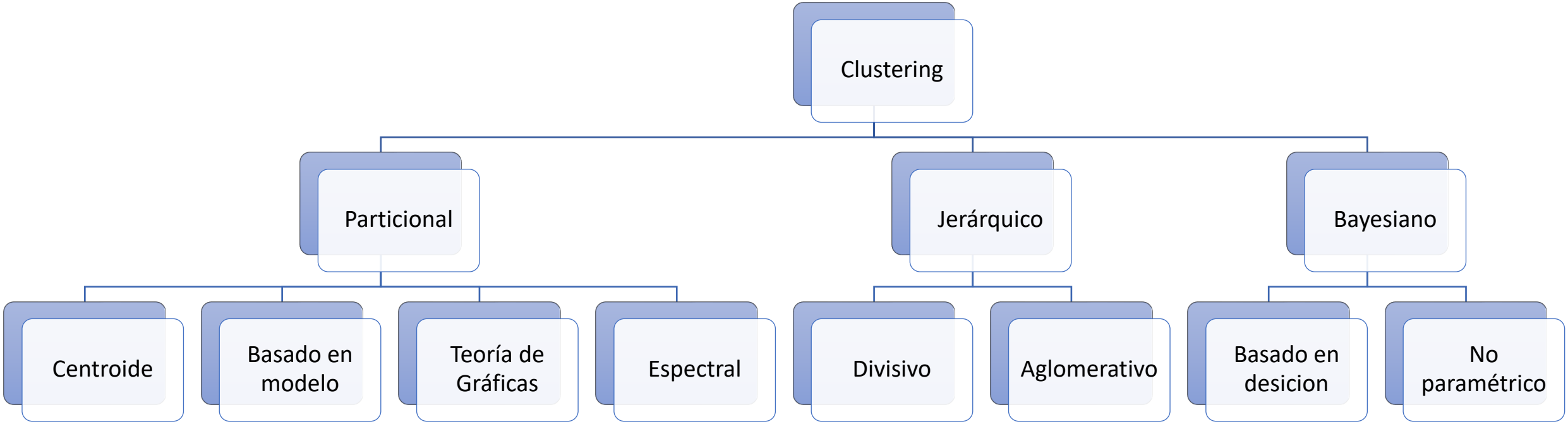
SEGMENTACIÓN DE IMÁGENES

ANÁLISIS EXPLORATORIO DE DATOS

BIOINFORMÁTICA

INVESTIGACIÓN DE MERCADOS

Métodos de clustering



¿Cómo medir la calidad de un algoritmo de clustering?

Para la mayoría de las aplicaciones la apreciación experta sigue siendo lo más importante

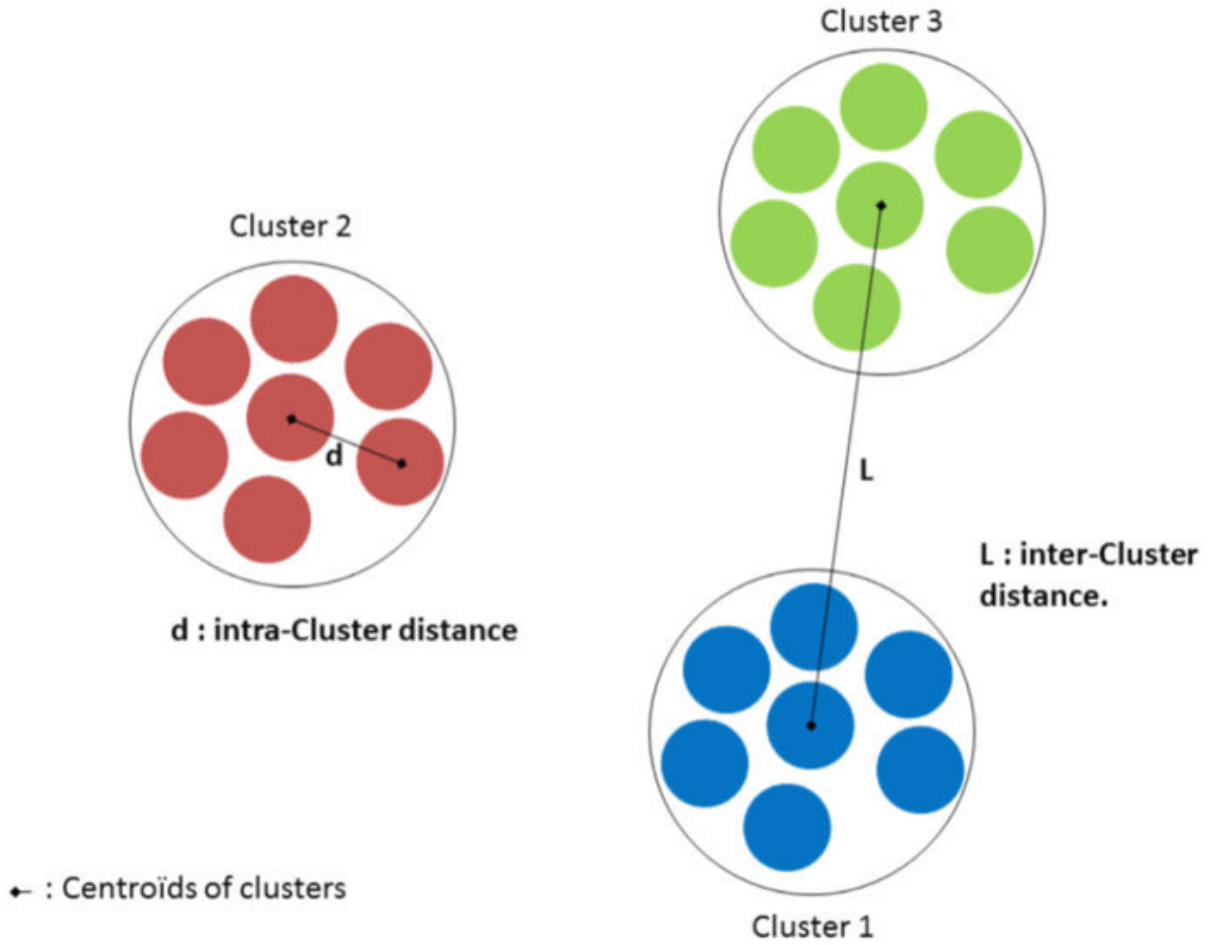
Posibilidad de tener una interpretabilidad de los regrupamientos

Diferentes medidas, todas subjetivas

Por mediciones intrinsecas en la formación de los regrupamientos

Si existe un conjunto de “clases” preasignadas a un conjunto relativamente pequeño de datos, medidas basadas en teoría de la información

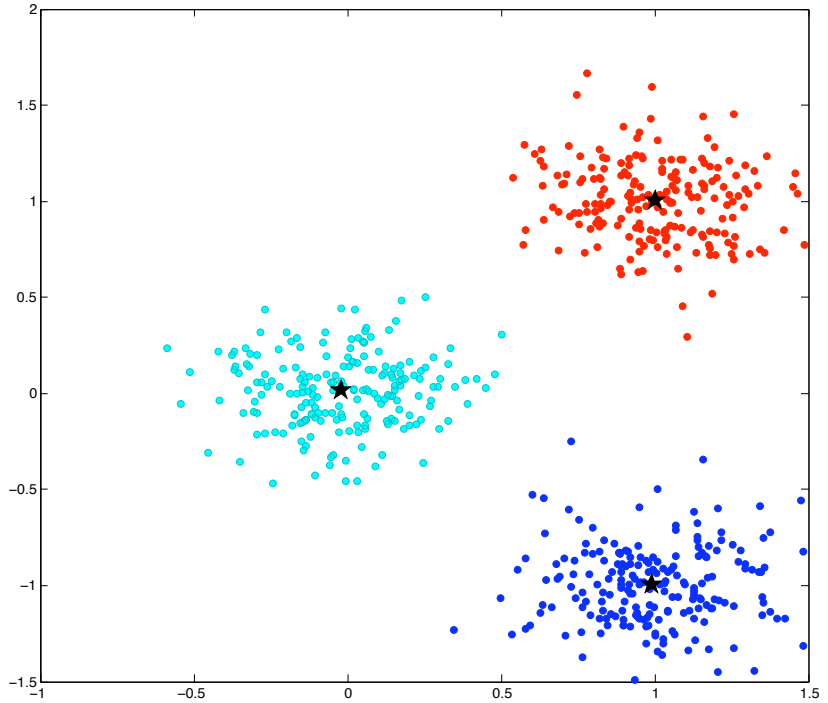
Medidas básicas



K-medias

Idea básica

Un objeto pertenece a un grupo si se encuentra más próximo al prototipo de esa clase que al de cualquier otra.



K-medias

- Prototipo de clase $c_j = [c_{j,1}, \ldots, c_{j,m}]$.
- Los parámetros de aprendizaje son $\theta = c_1, \ldots, c_K$.
- Distancia euclidiana entre un objeto y un prototipo:

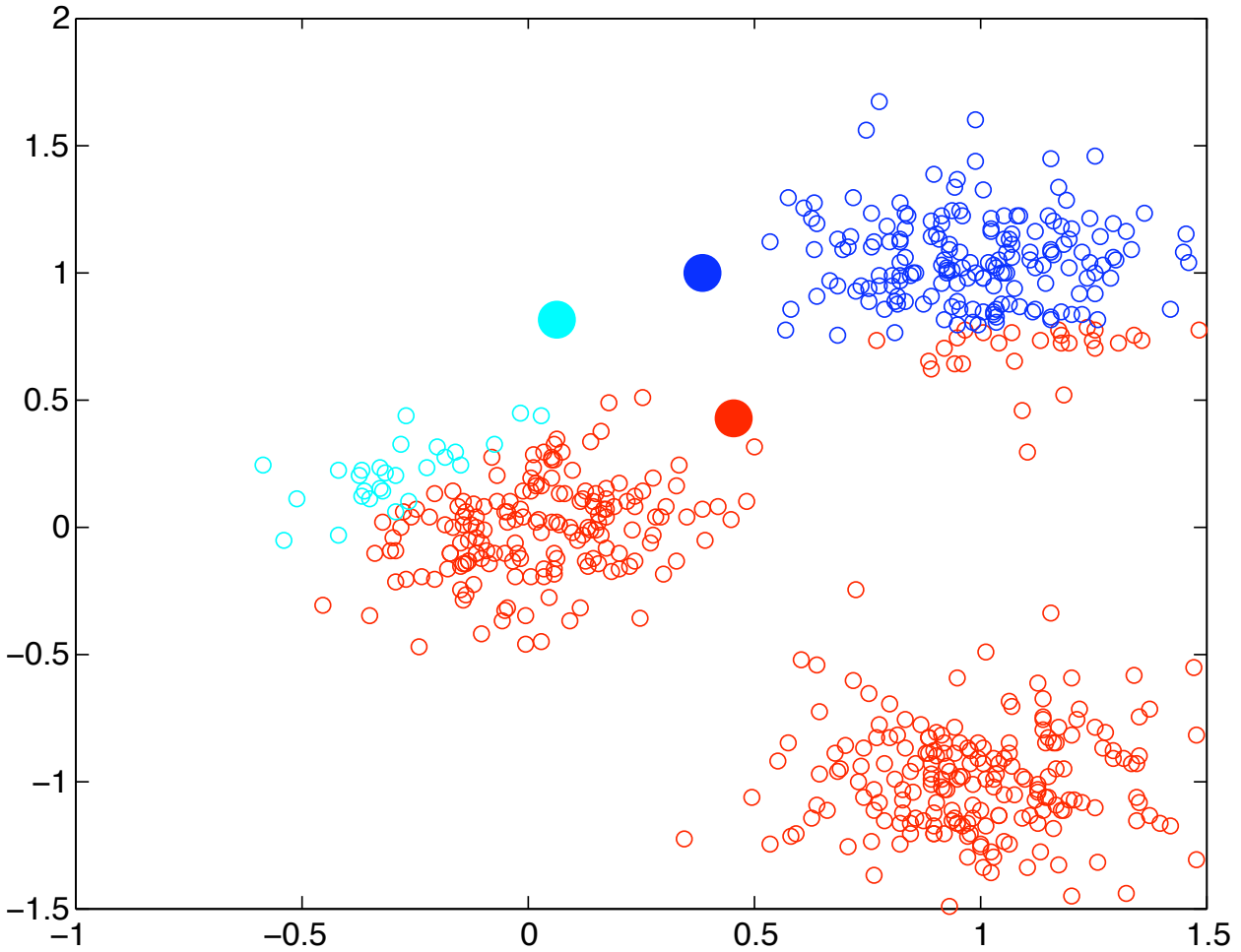
$$d(x_i, c_j) = \sqrt{\sum_{l=1}^{m} (x_{i,l} - c_{j,l})^2}.$$

- Minimizar:

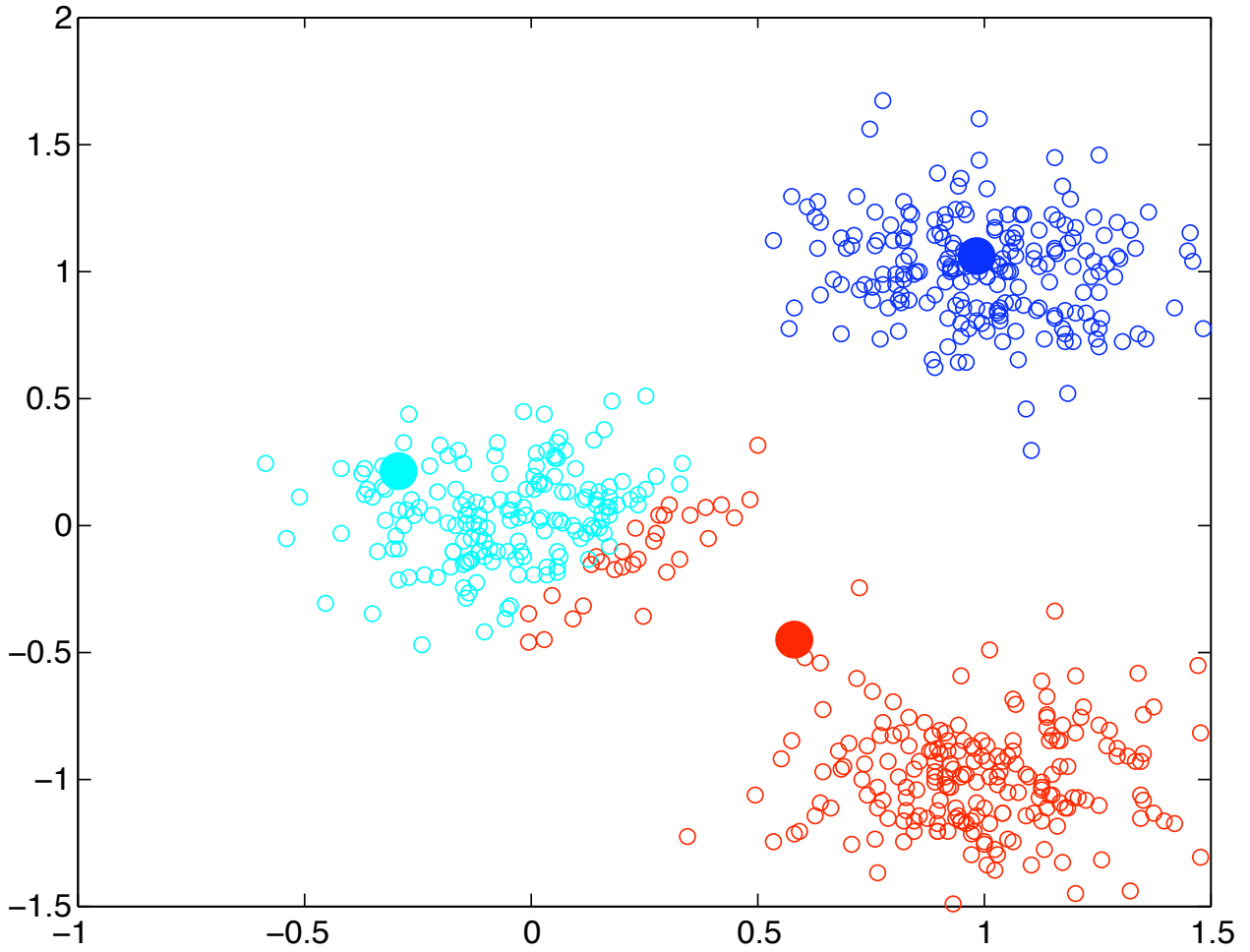
$$J(\theta) = \sum_{i=1}^{n} \min_k \left(d(x_i, c_k) \right).$$

- ¡Problema NP completo!

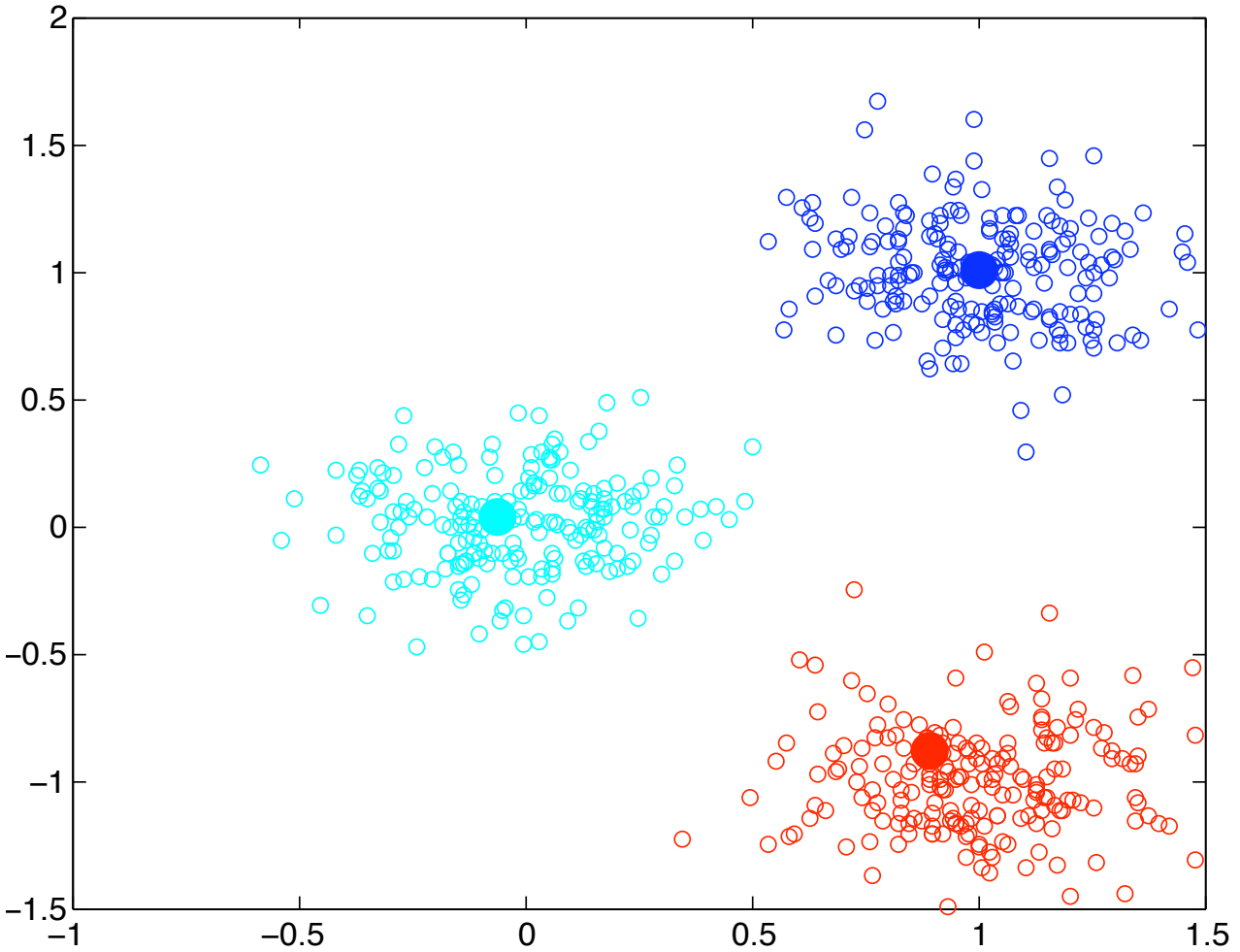
K-medias



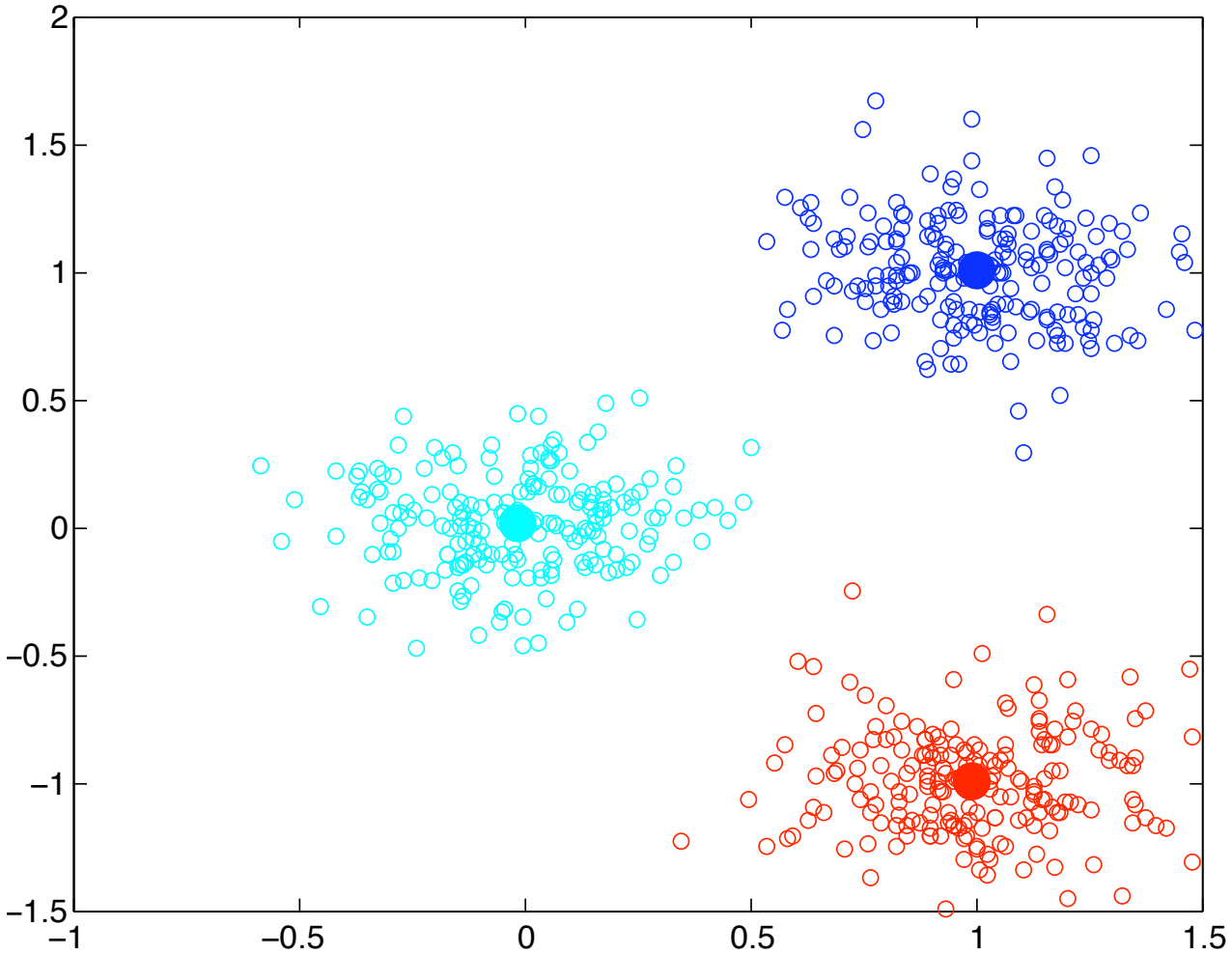
K-medias



K-medias



K-medias



K-medias

Entrada: $\{x_1, \ldots, x_n\}$, K, m.
Salida: $\{c_1, \ldots, c_K\}$.

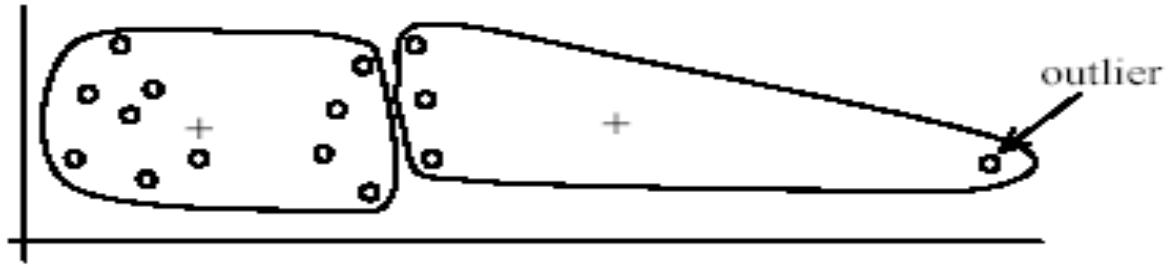
1: Inicializa $\{c_1, \ldots, c_K\}$ con valores aleatorios.
2: **repetir**
3: **para** i de 1 a n **hacer**
4: $x_i \in P_k$ si $\min_k(d(x_i, c_k))$.
5: **fin para**
6: **para** k de 1 a K **hacer**
7: **para** j de 1 a m **hacer**
8: $$c_{k,j} = \frac{\sum_{x_i \in P_k} x_{i,j}}{||P_k||}$$
9: **fin para**
10: **fin para**
11: **hasta** no varíe $c_{k,j}, \forall k, j$

Ventajas de las K-medias

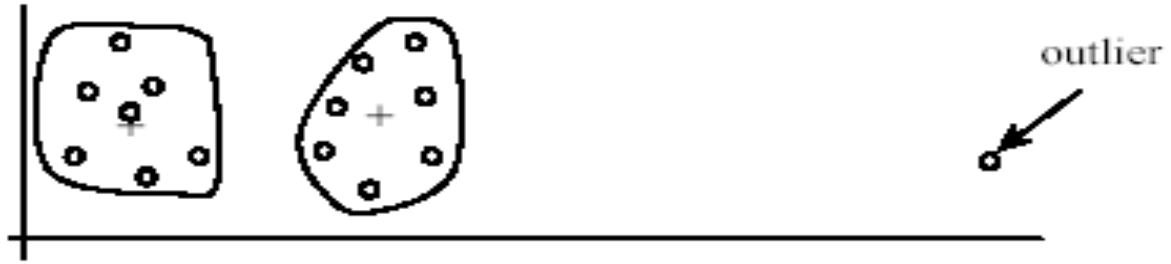
- Simple de implementar
- Eficiente: complejidad casi lineal
- Popular
- Mínimo local

Desventajas de las K-medias

Outliers

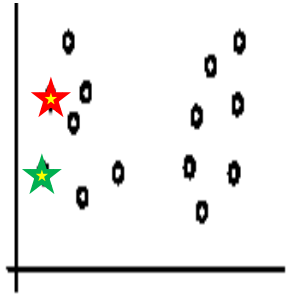


(A): Undesirable clusters

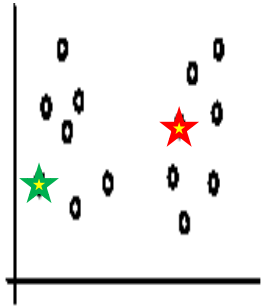


(B): Ideal clusters

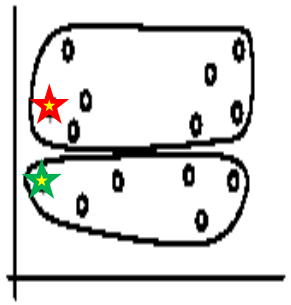
Desventajas de las K-medias



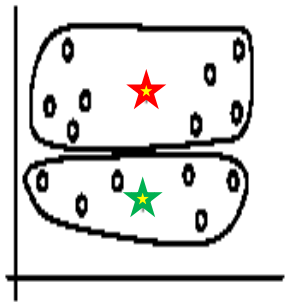
Random selection of seeds (centroids)



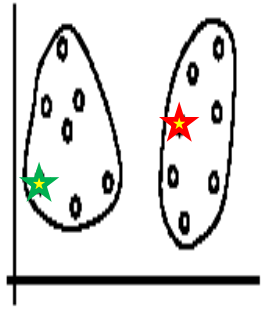
Random selection of seeds (centroids)



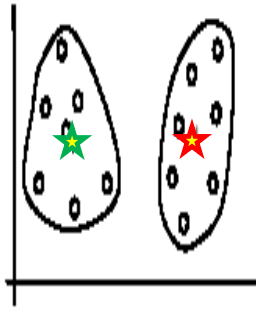
Iteration 1



Iteration 2

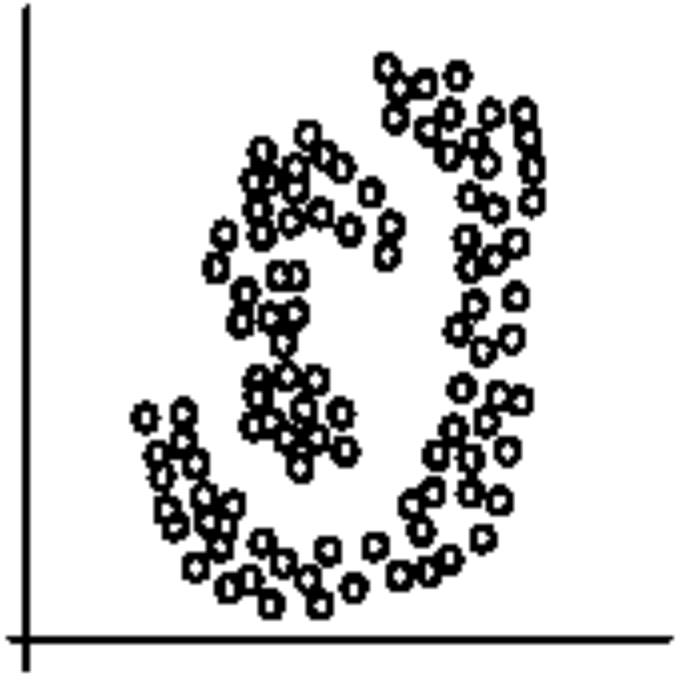


Iteration 1

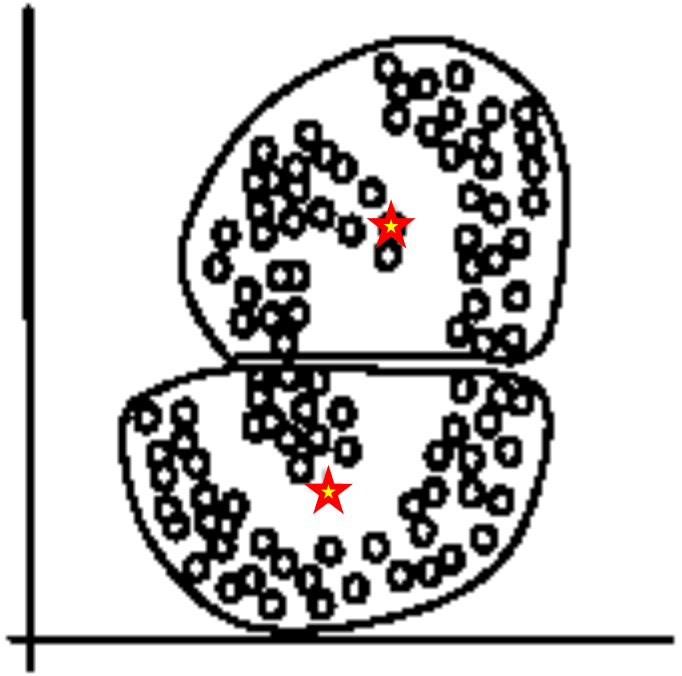


Iteration 2

Desventajas de las K-medias



(A): Two natural clusters



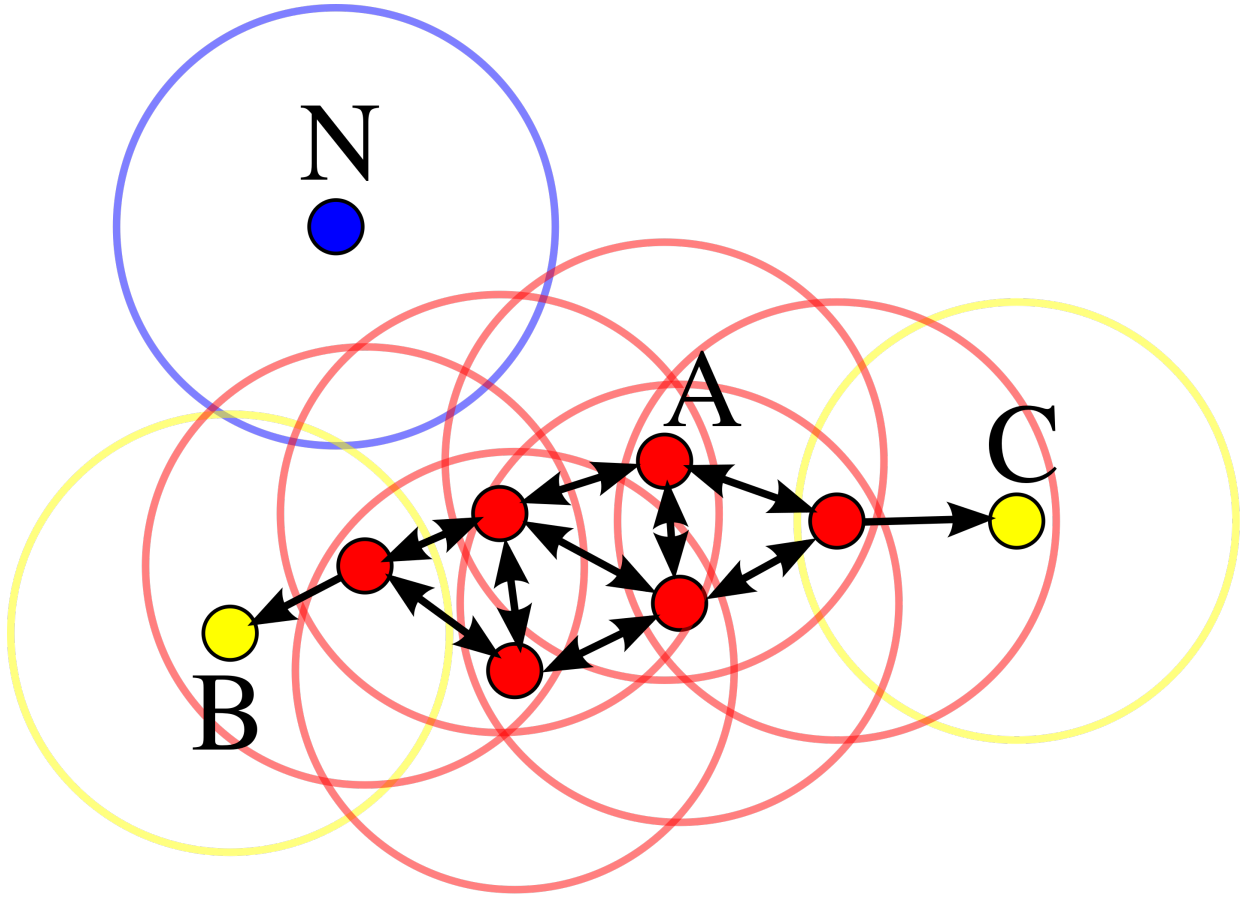
(B): *k*-means clusters

DBSCAN

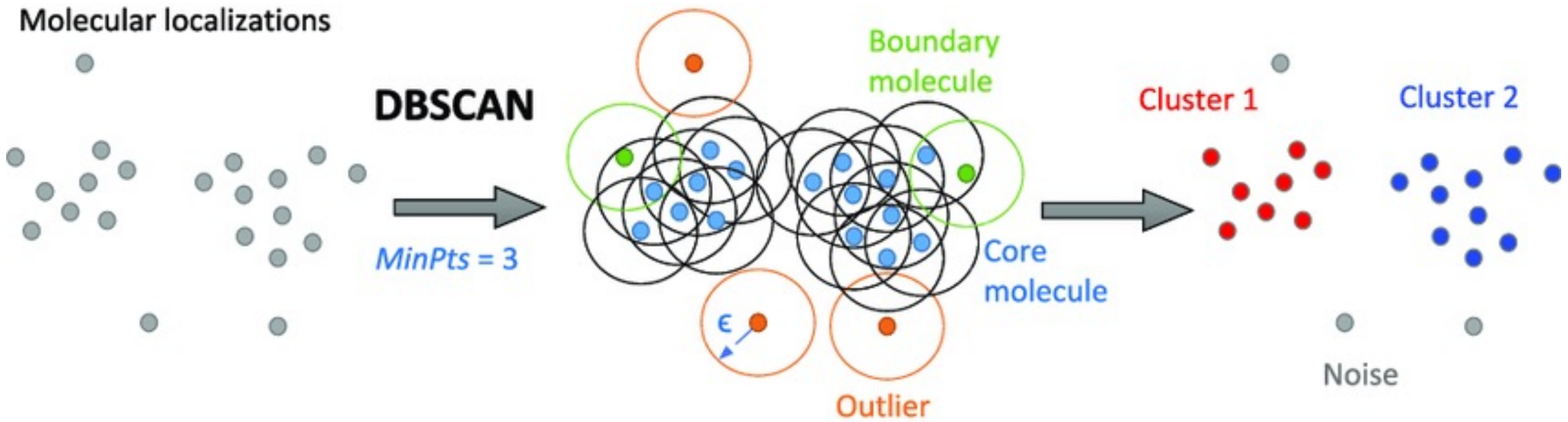
Density-based spatial clustering of applications with noise

- Un punto núcleo es un dato que al menos *minPts* puntos están a una distancia ε de él
- Un punto q es alcanzable desde p si existe una secuencia de puntos nucleo entre ellos y se encuentre a una distancia ε de alguno de ellos
- Un punto que no sea alcanzable desde cualquier otro punto es considerado ruido

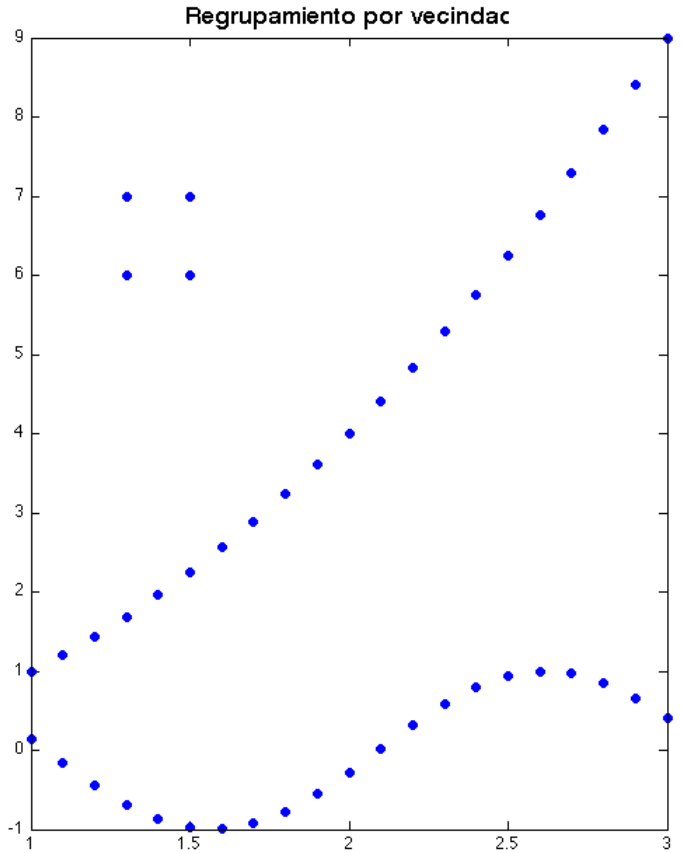
DBSCAN

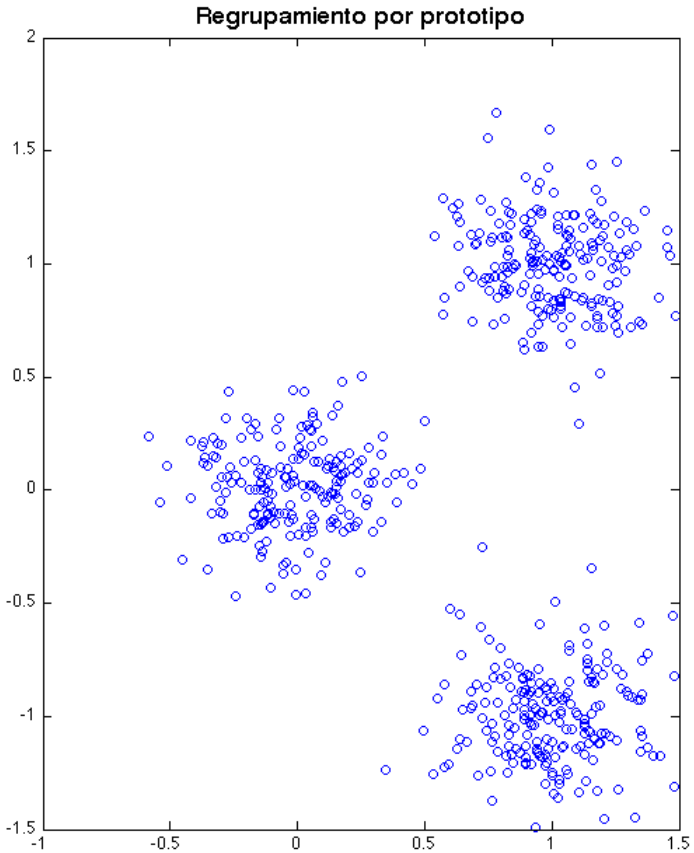


DBSCAN

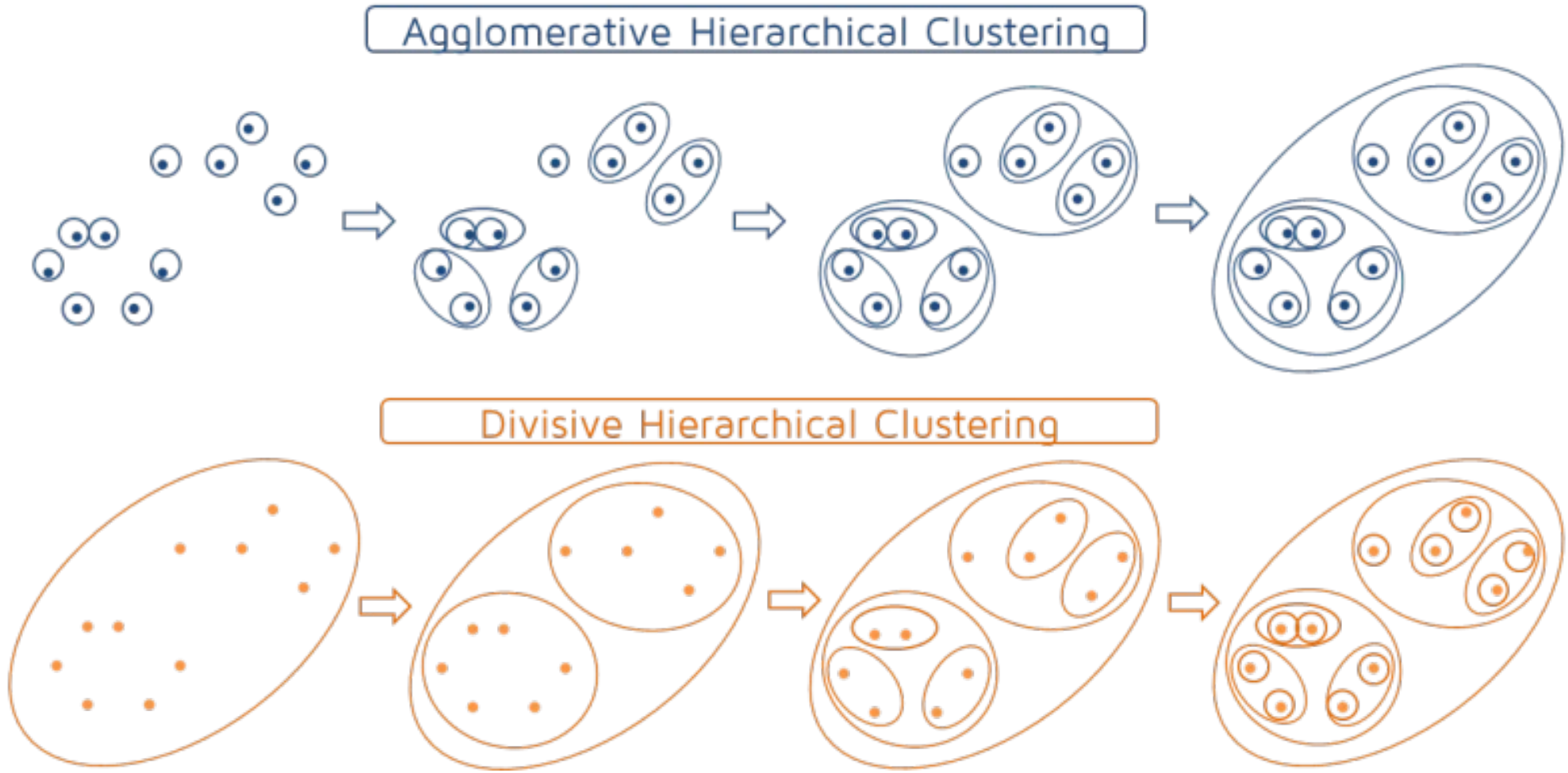

Molecular localizations
DBSCAN
MinPts = 3
Boundary molecule
Core molecule
Outlier
ϵ
Cluster 1
Cluster 2
Noise

Clustering jerárquico

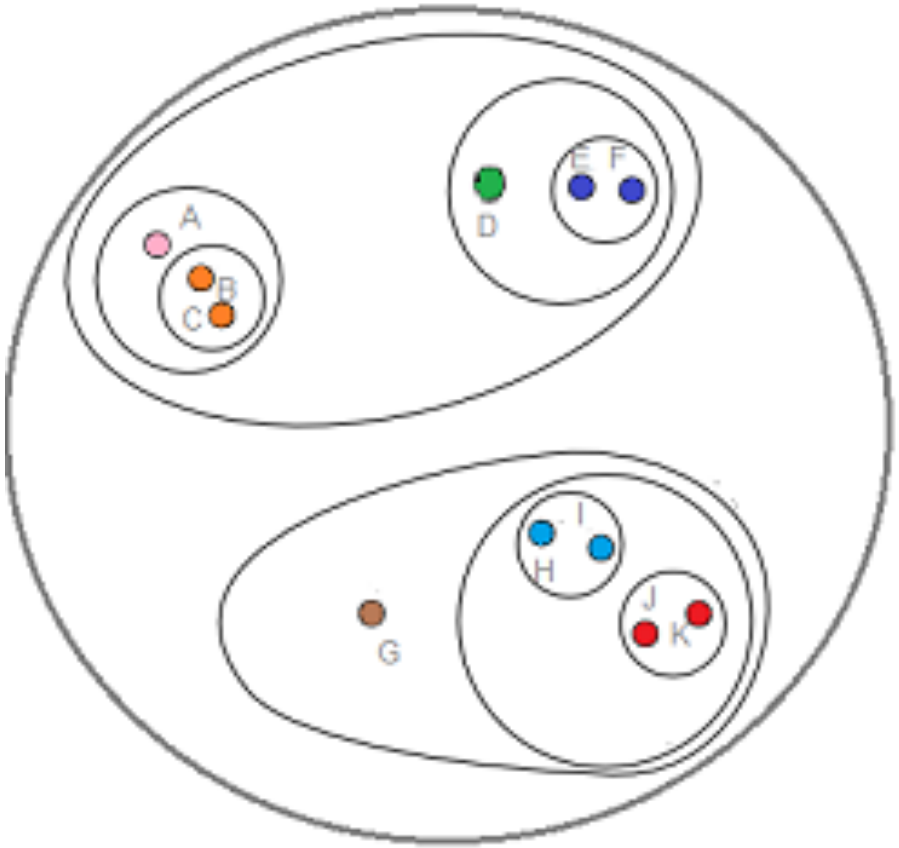


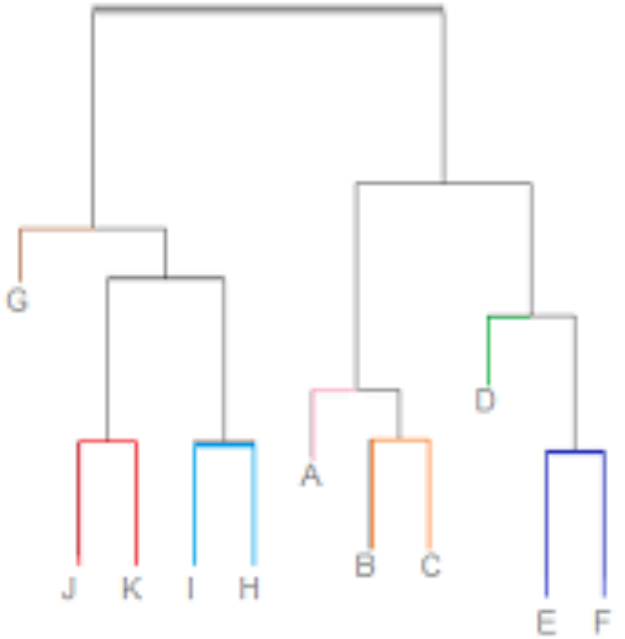


Dos métodos de clustering jerárquico

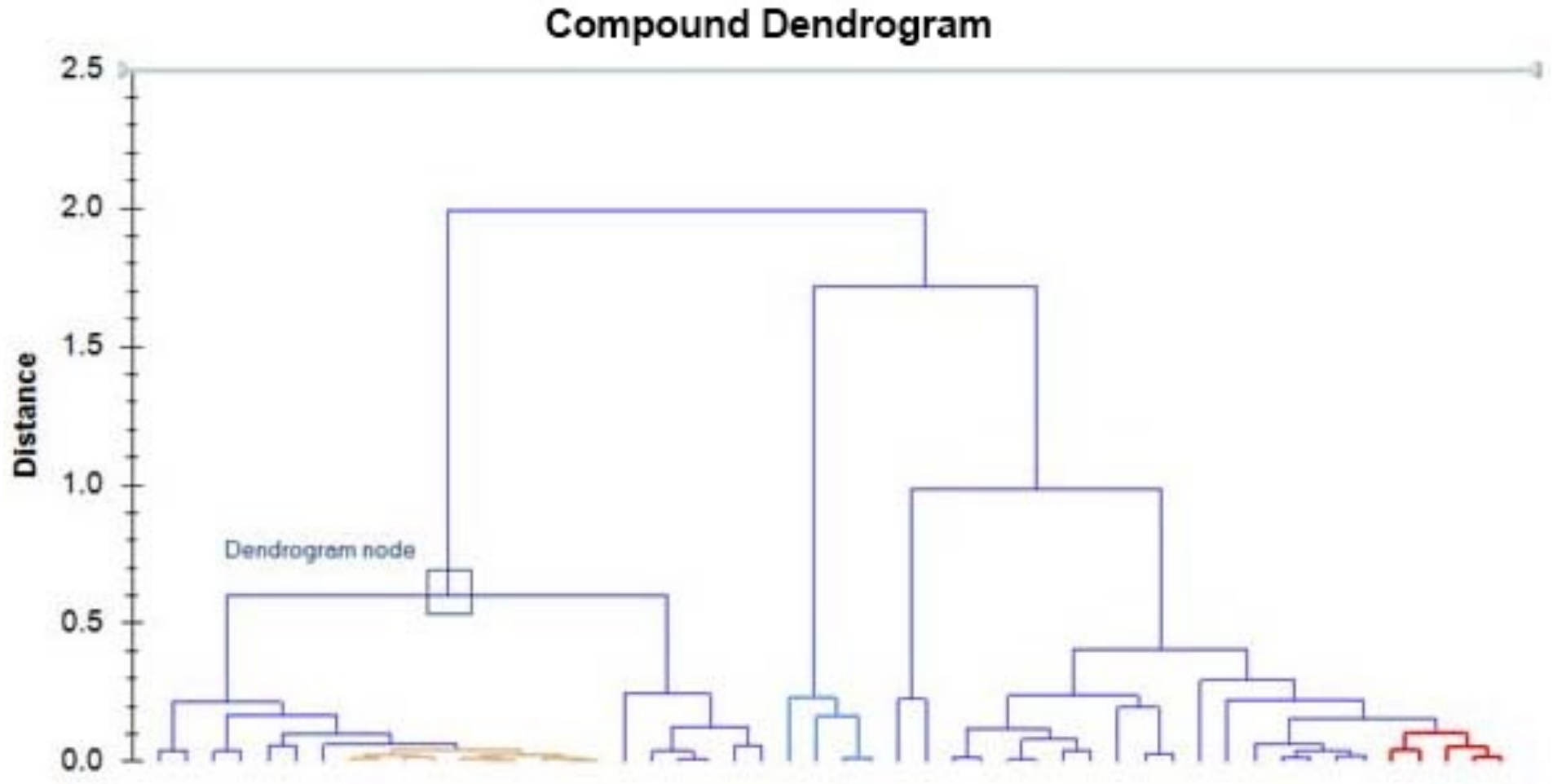


Método aglomerativo



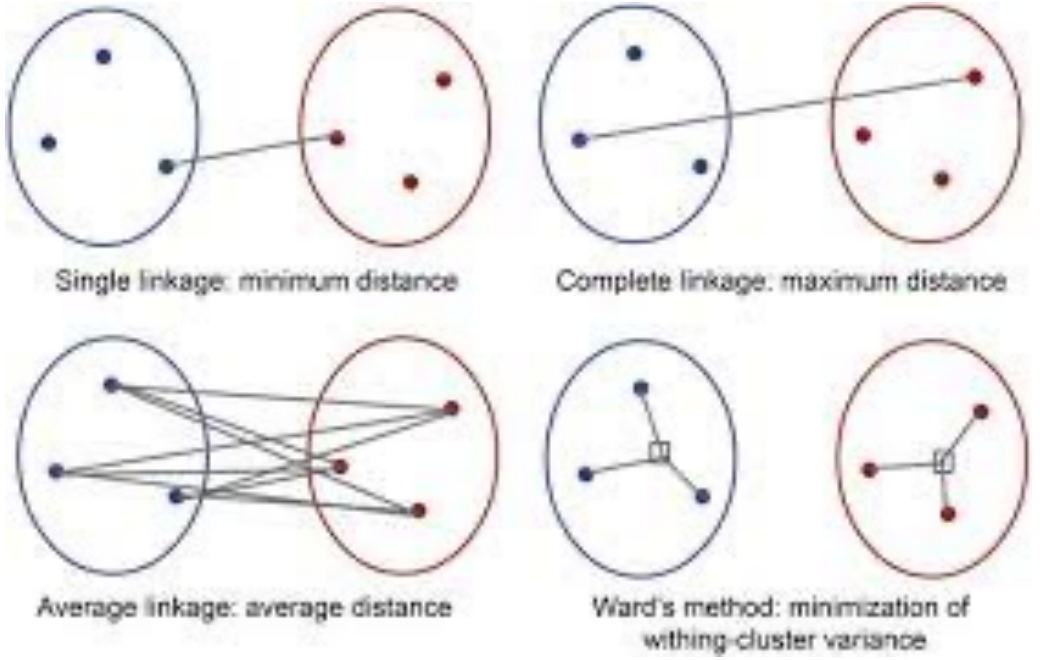


Dendograma


Compound Dendrogram
Distance
2.5
2.0
1.5
1.0
0.5
0.0
Dendrogram node

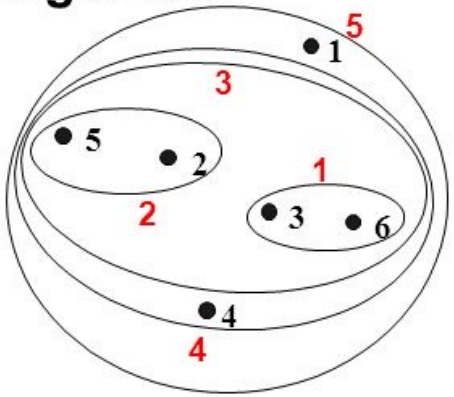
Enlazado

Basado en una medida de distancia entre instancias

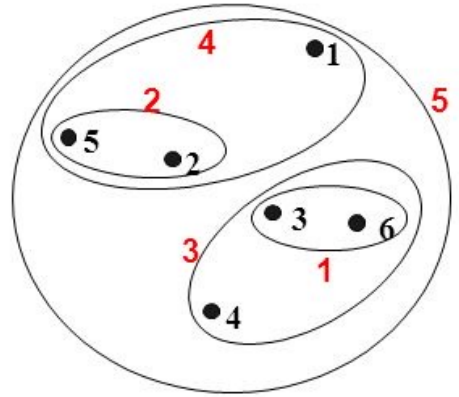


Enlazado

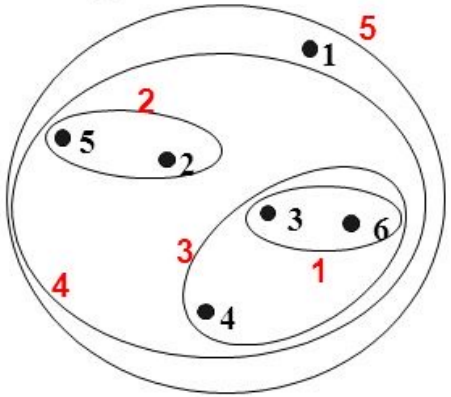
Single-link



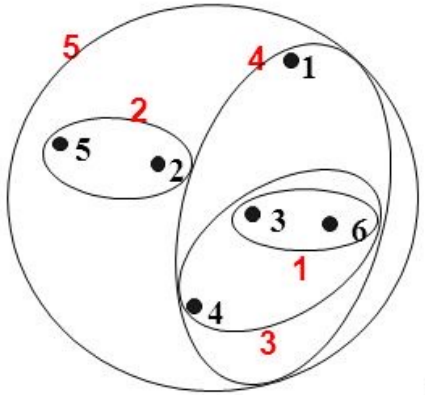
Complete-link



Average-link

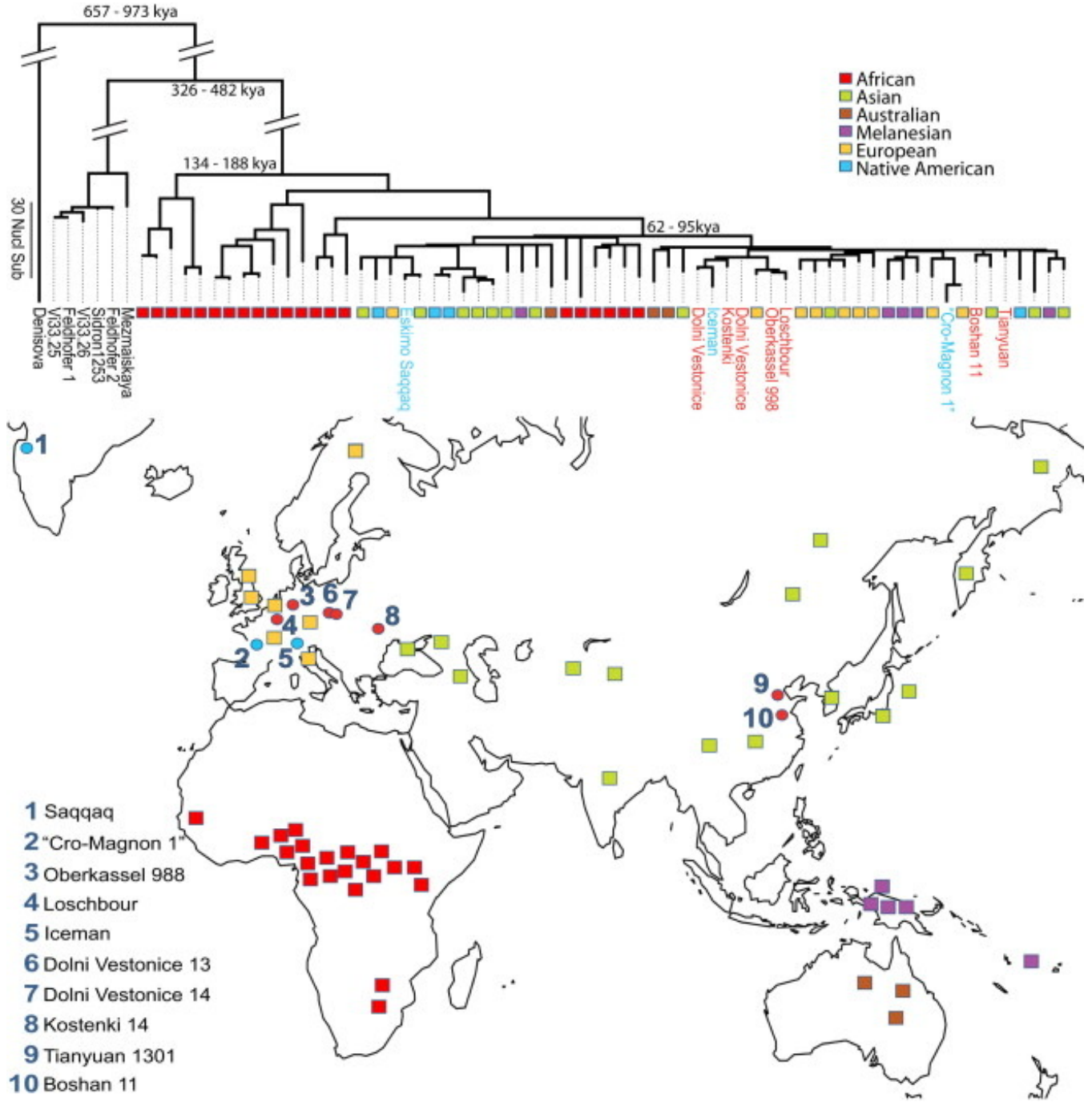


Centroid distance



Ejemplo:

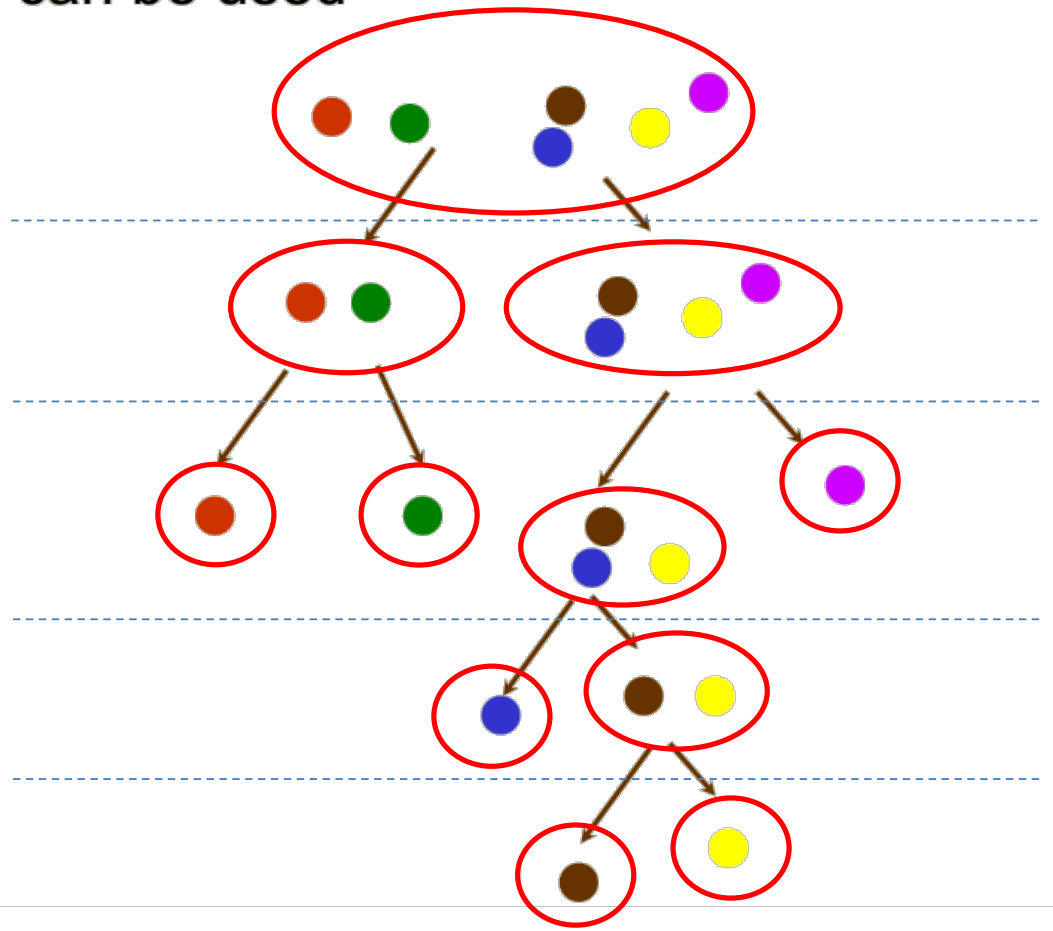
La Eva Mitocondrial



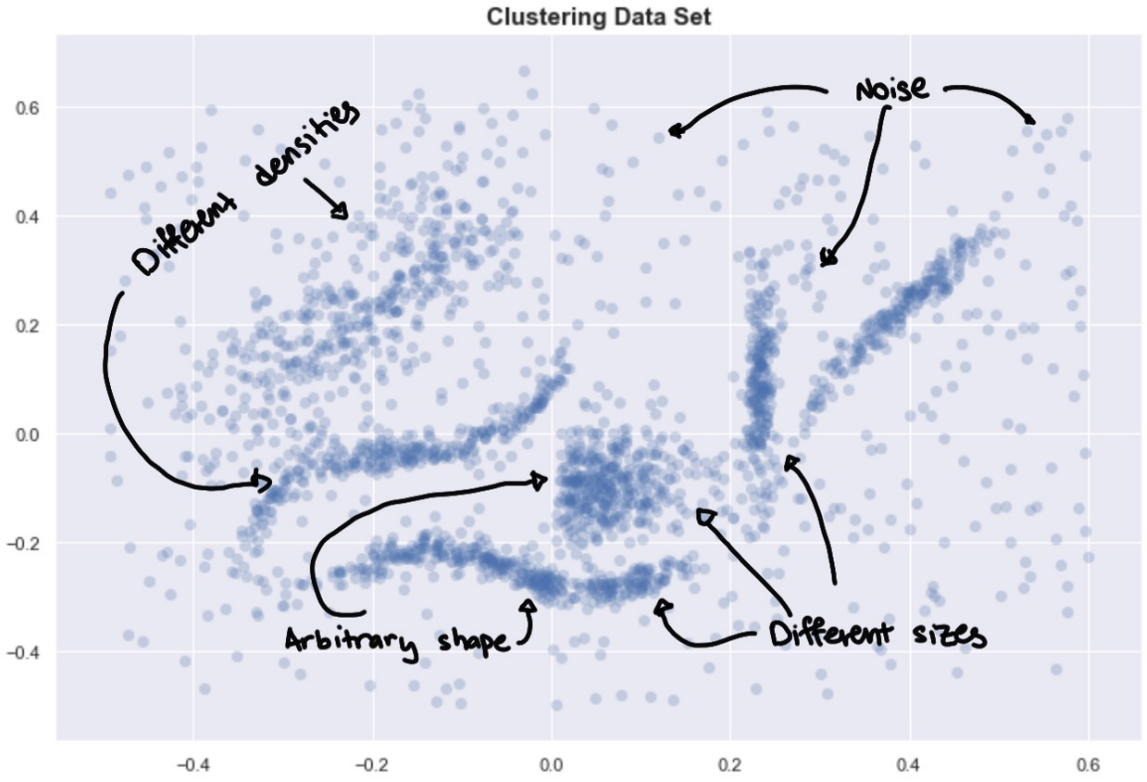
Método divisivo

Any “flat” algorithm which produces a fixed number of clusters can be used

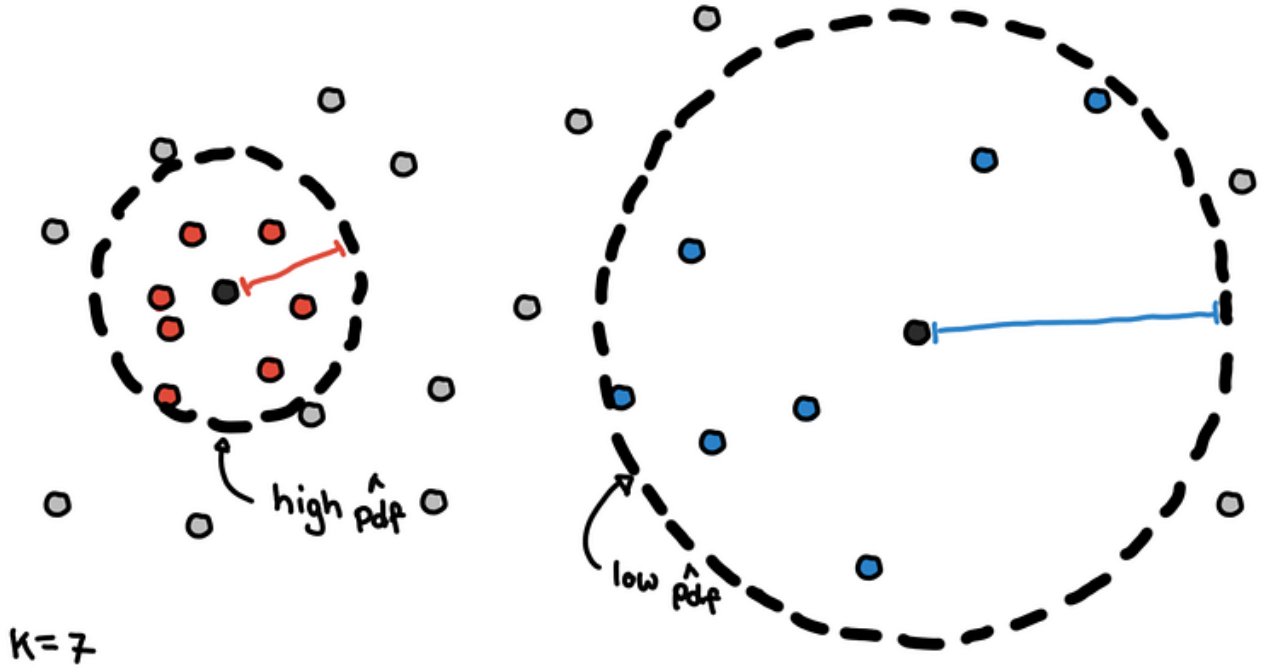
- set ***c*** = ***2***

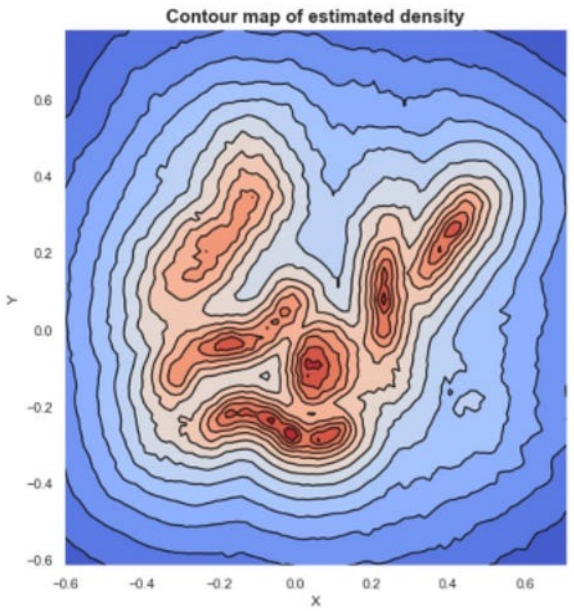


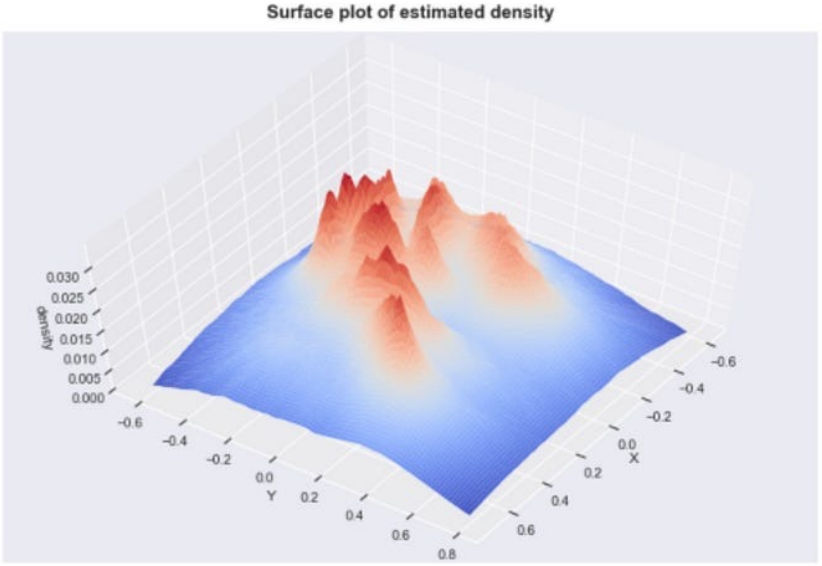
HDBSCAN


Clustering Data Set
Different densities
Noise
Arbitrary shape
Different sizes
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
0.6

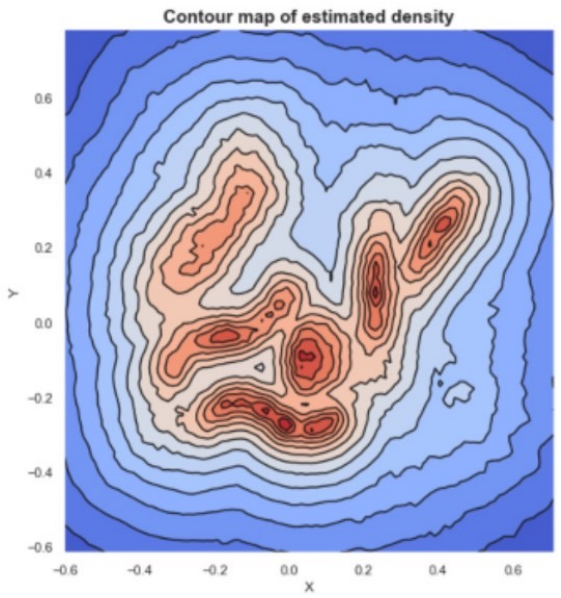
Core Distance

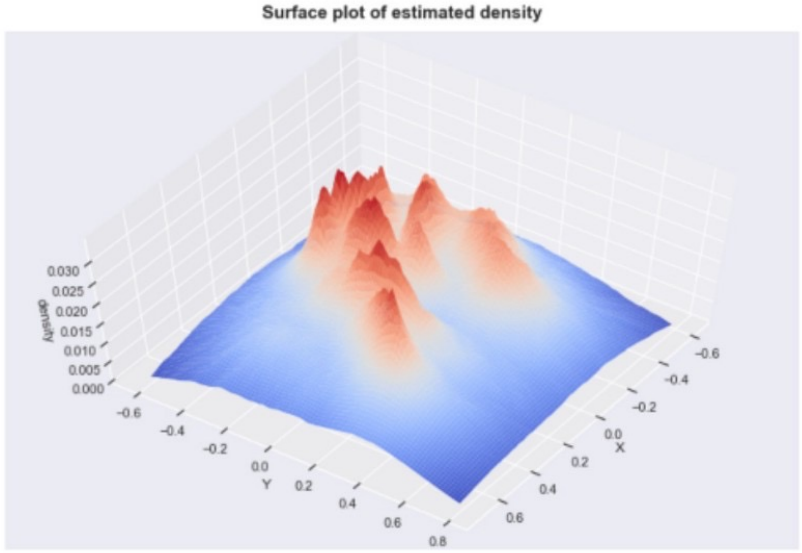


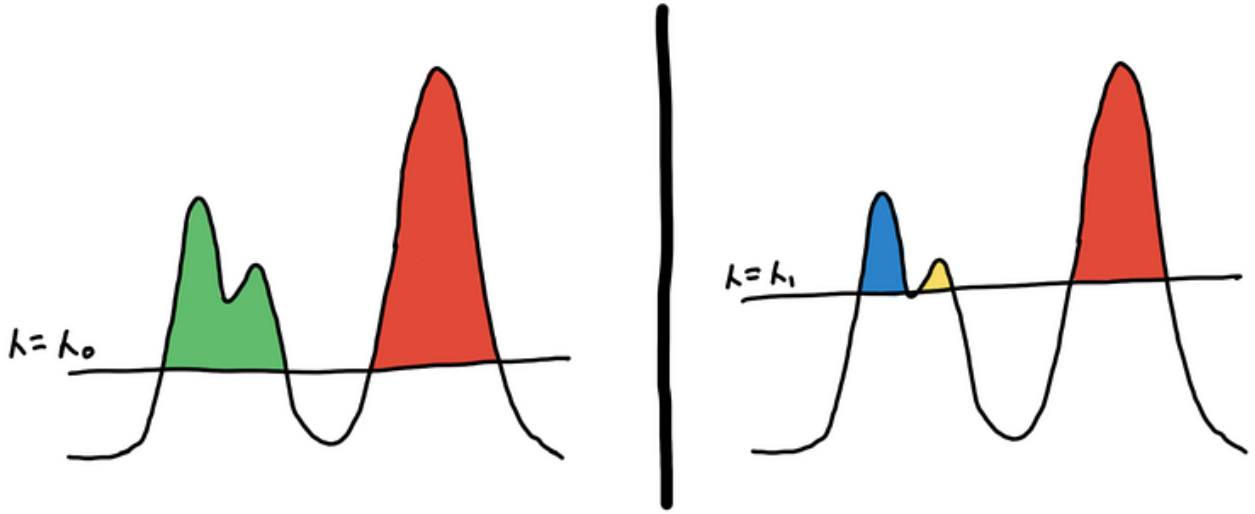




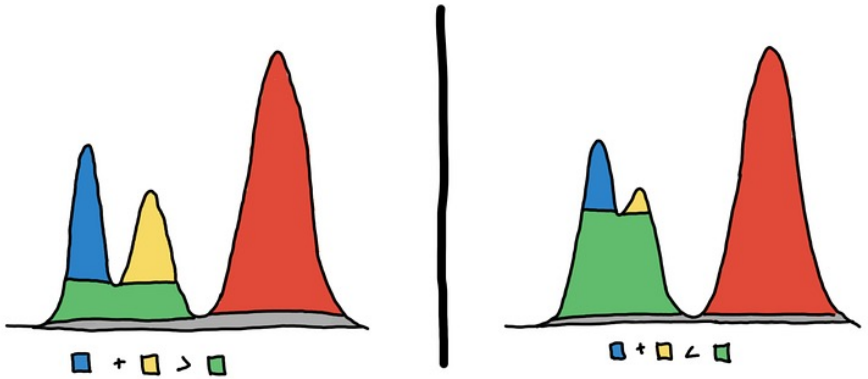
Cluster Selection

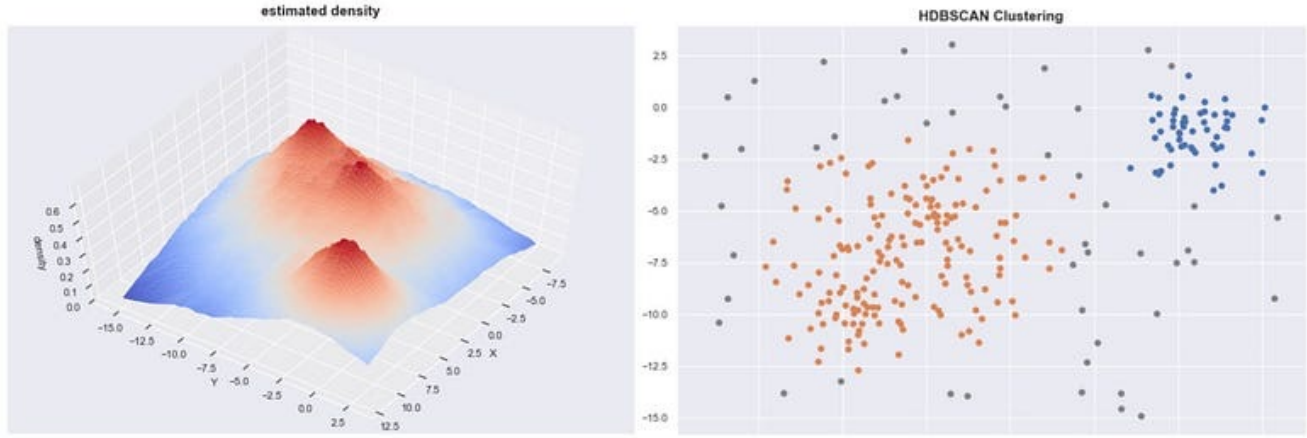


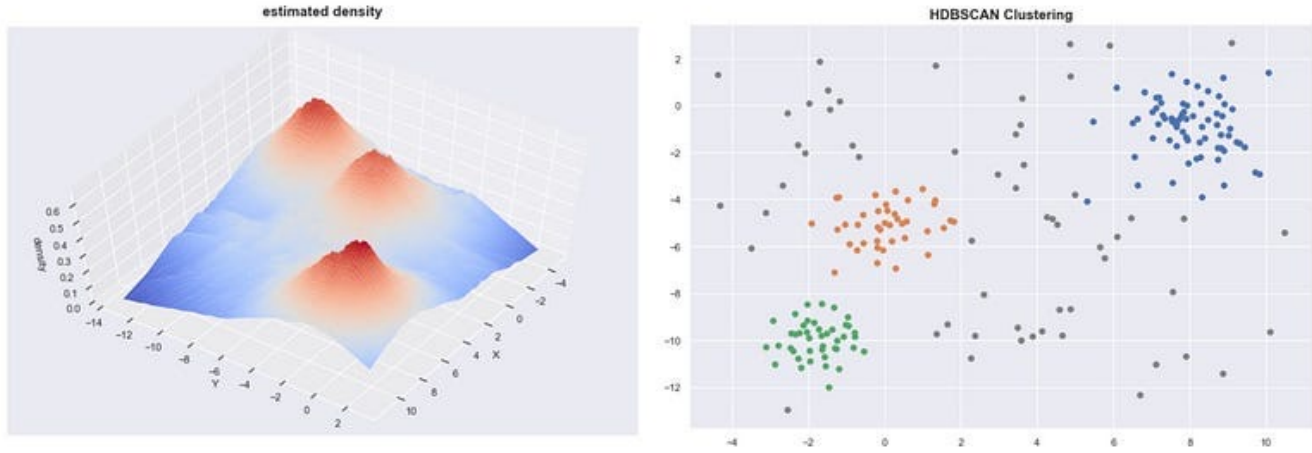




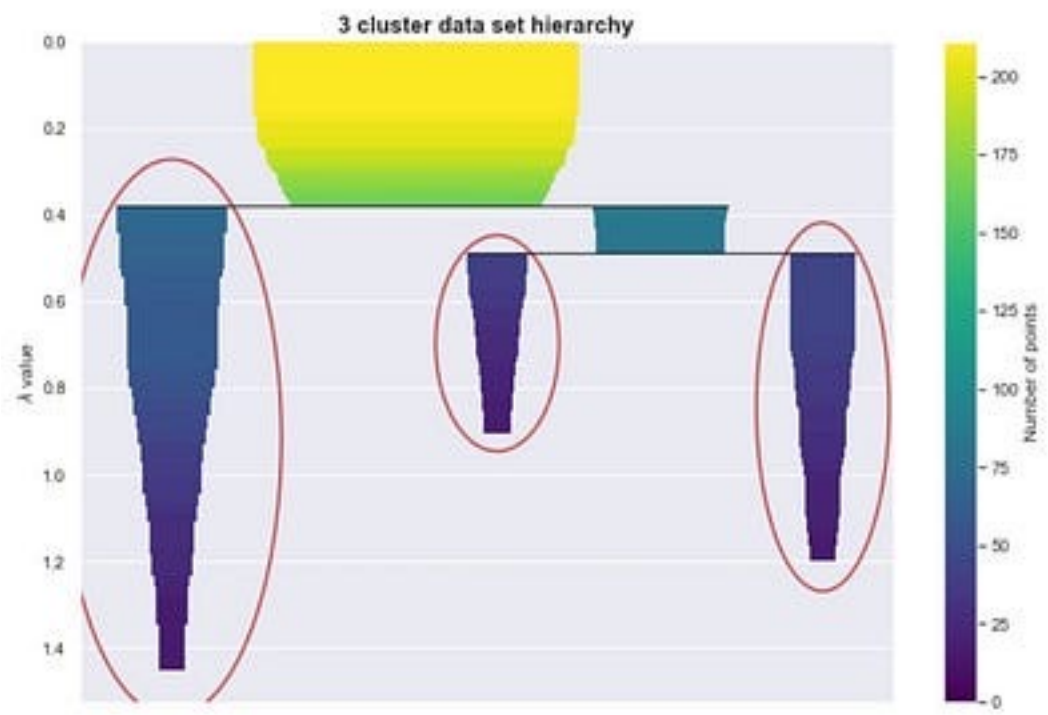
Escogiendo clusters

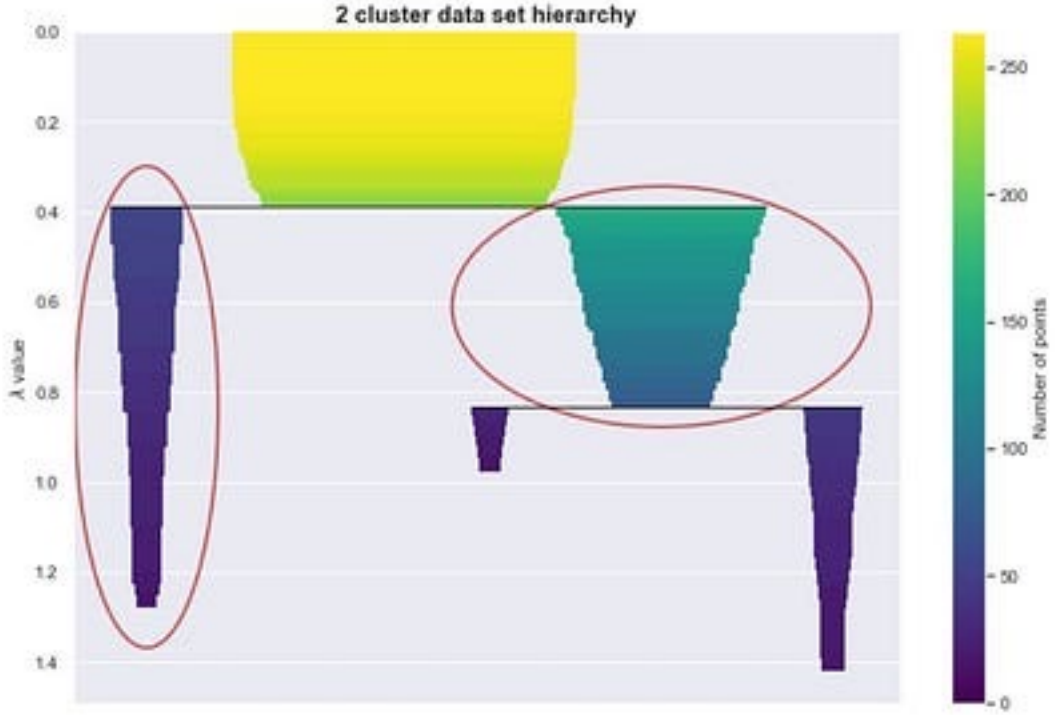




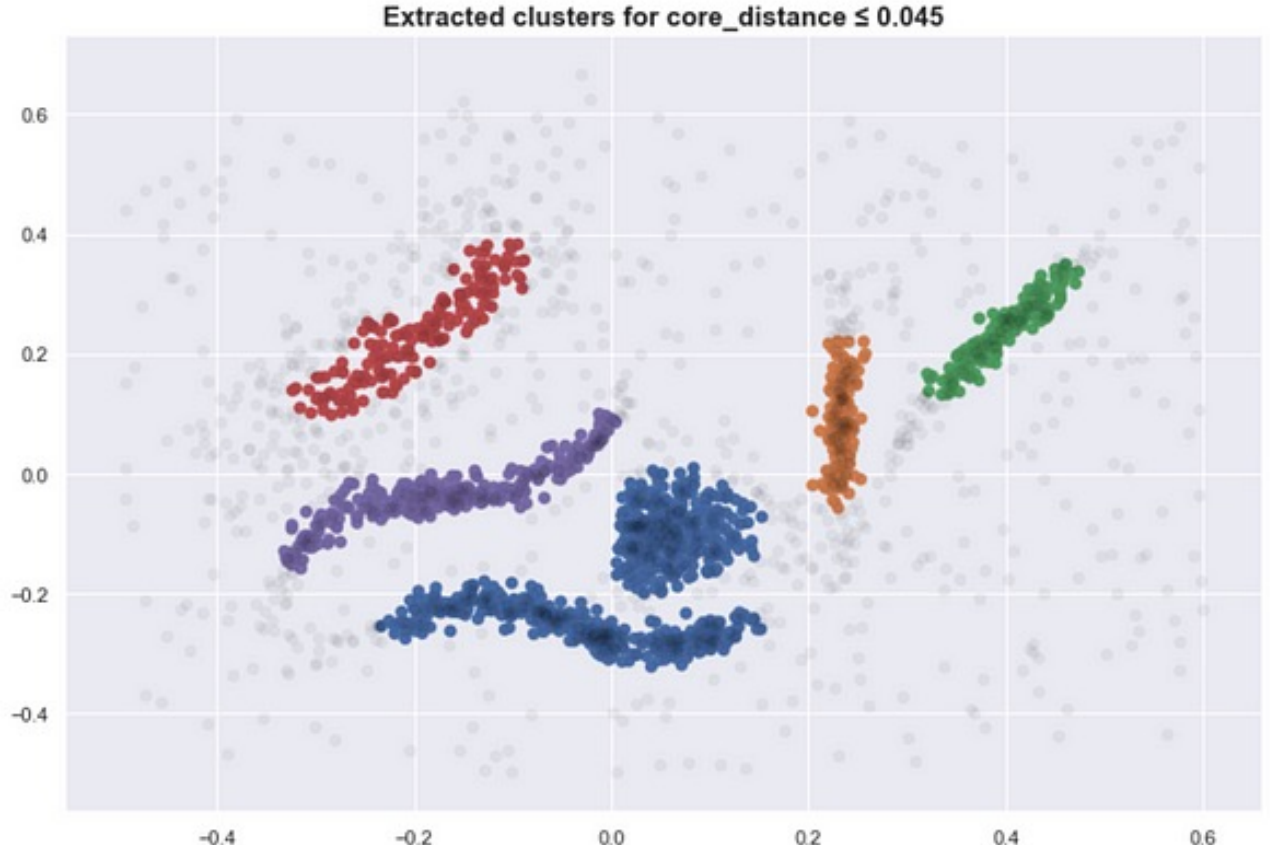


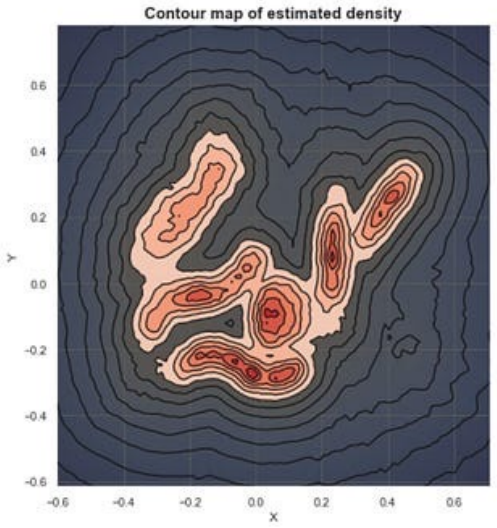
Clusters estables

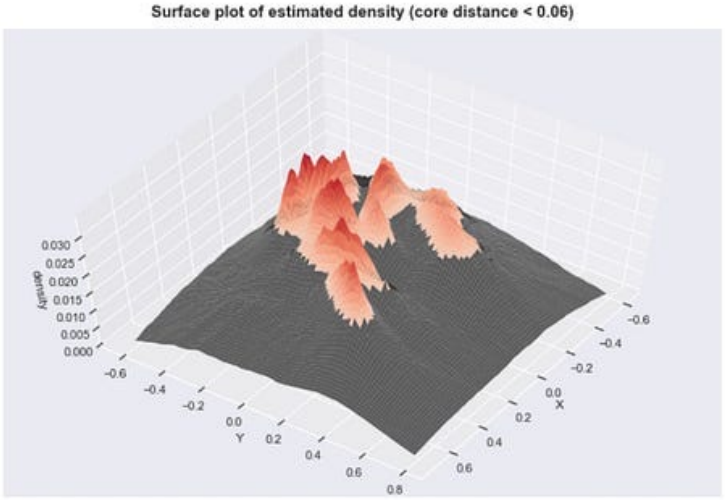




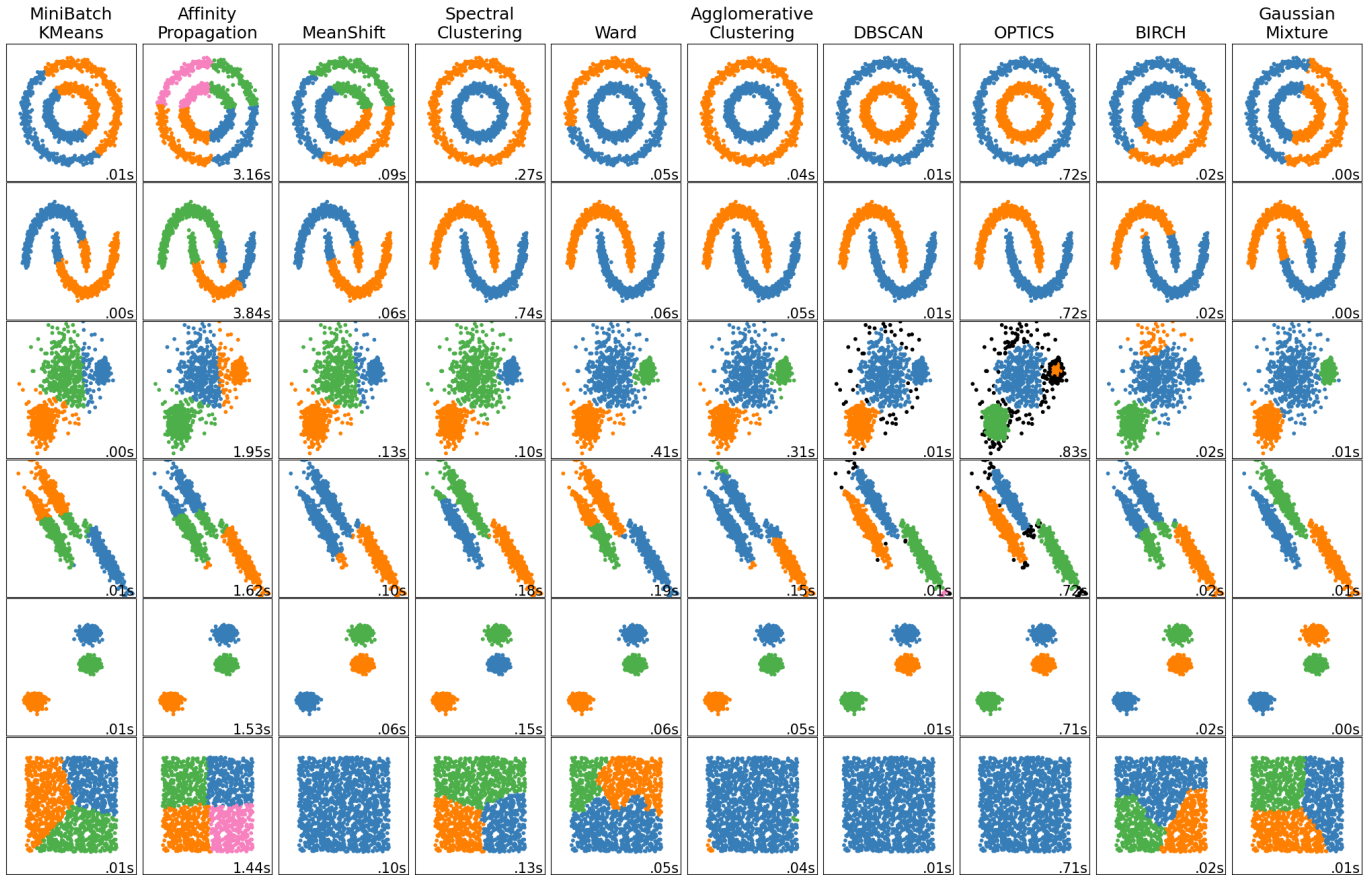
HDBSCAN


Extracted clusters for core_distance ≤ 0.045
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
0.6


Contour map of estimated density
Y
X


Surface plot of estimated density (core distance < 0.06)
density
X
Y

Y todo esto lo podemos ver en Sci-Kit Learn



A modo de discusión final…

¿Para qué sirve el análisis aglomerativo?

¿Qué tipos de clustering existen?

¿Porqué usar diferentes técnicas?

¿Cómo saber si un método es bueno?