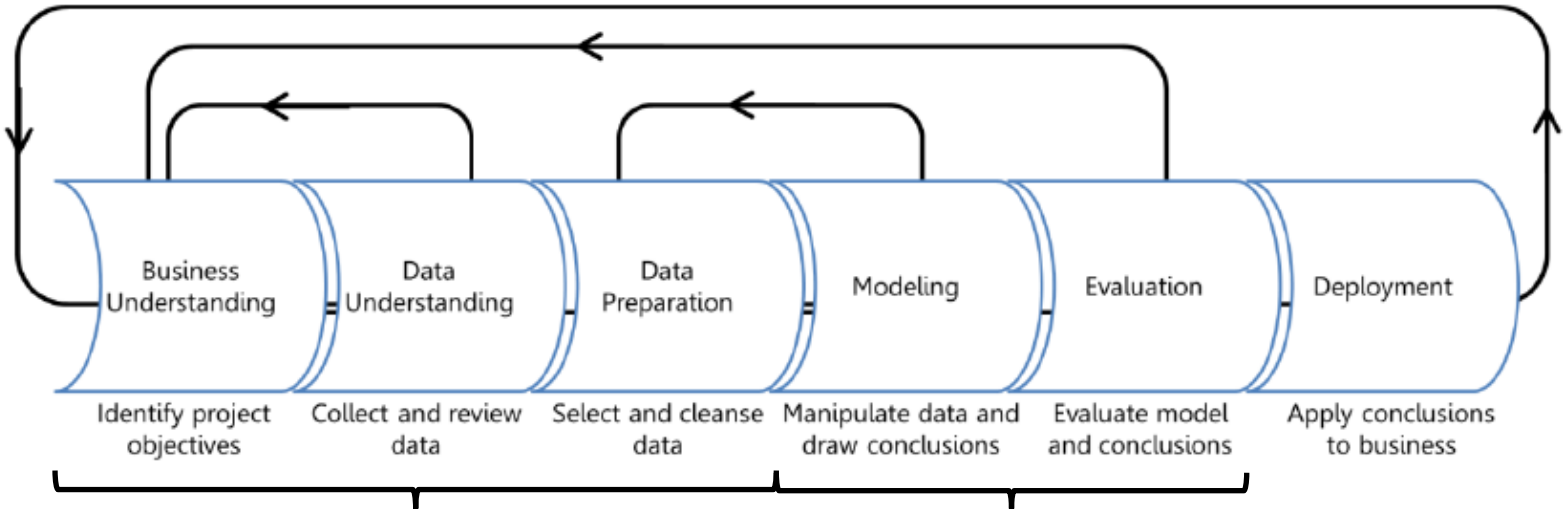
Análisis de Clusters

Dr. Raimundo Sánchez

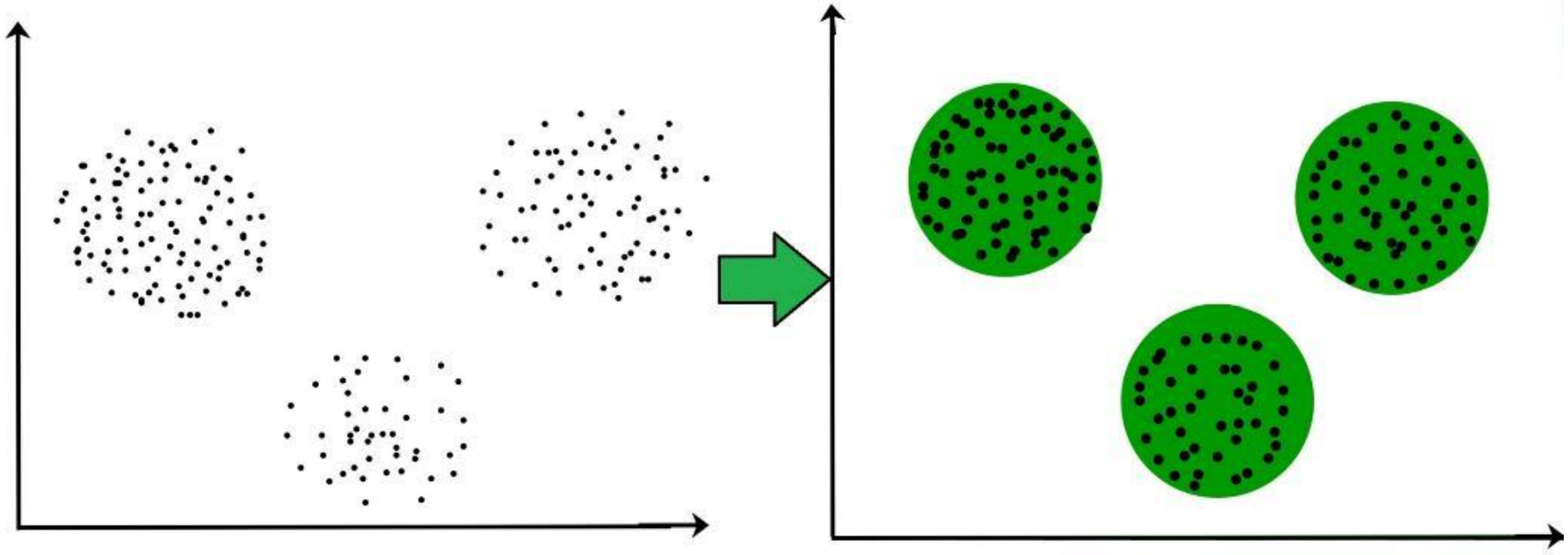
raimundo.sanchez@uai.cl

@raimun2

Proceso de análisis de datos (CRISP-DM)



Clusters



Clusters

La noción de clúster puede ser ambigua

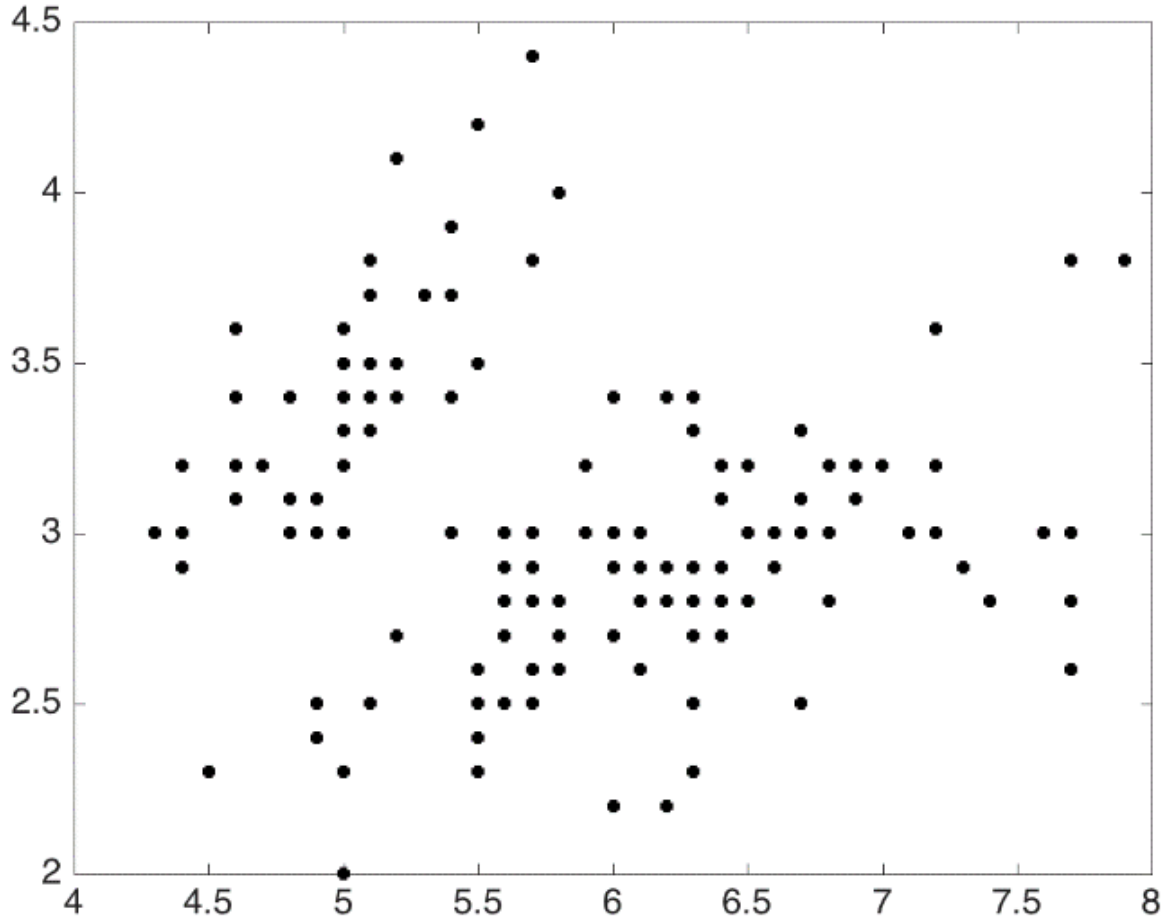
¿Cuántos clusters?

Seis grupos

Dos grupos

Cuatro grupos

¿Cuántos clusters?

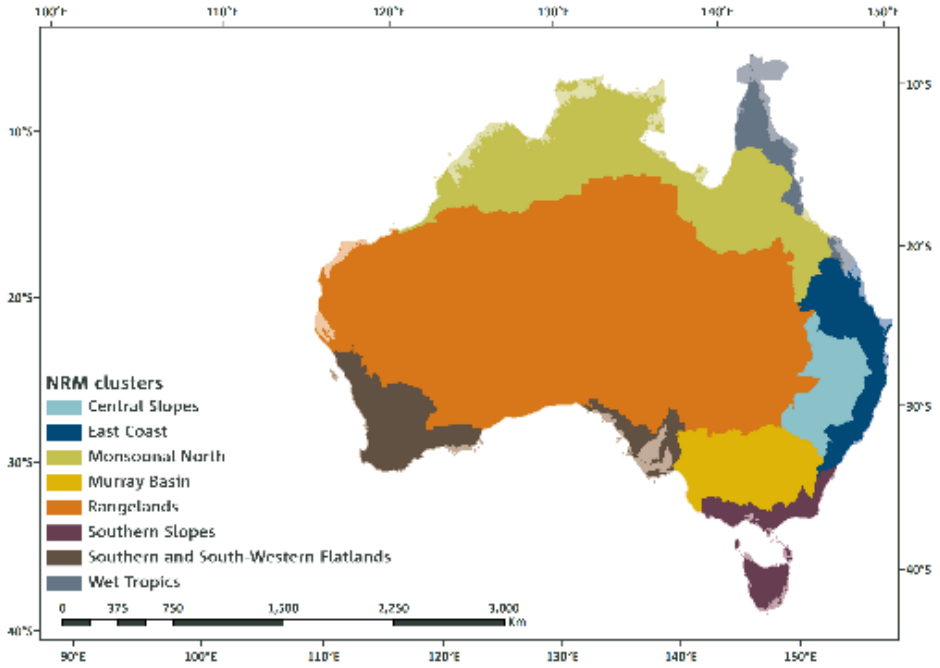


Objetivos

Encontrar **algorítmicamente** grupos de entidades tales que:

1. La similitud intragrupo es alta
2. La similitud entre grupos es baja

Las medidas de distancia y similitud son cruciales en este proceso



Criterios para agrupar

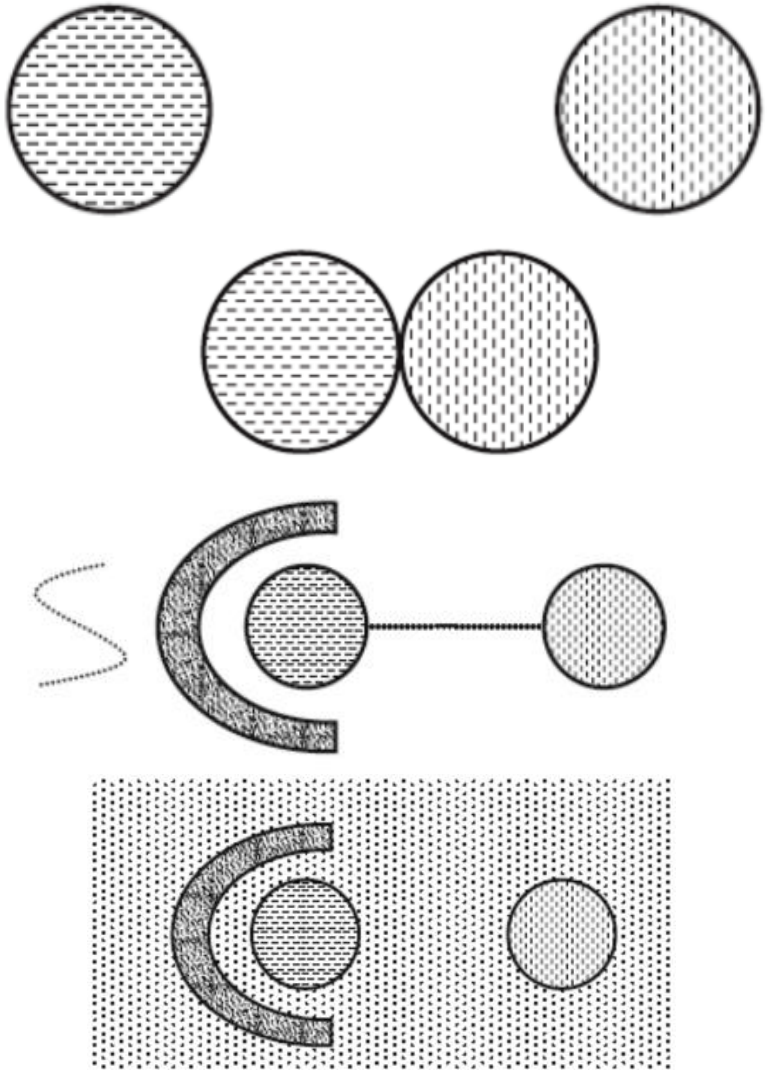
Existen varias nociones de clusters

Clústeres basados en distancias: Cada punto está más cerca de todos los puntos de su grupo que cualquier punto de otro grupo.

Clústeres basados en el centroide: Cada punto está más cerca del centro de su grupo que del centro de cualquier otro grupo.

Clústeres basados en contigüidad: Cada punto está más cerca de al menos un punto de su grupo que cualquier punto de otro grupo.

Clústeres basados en densidad: Los clusters son regiones de alta densidad separadas por regiones de baja densidad.



Enfoques al problema

Existen diferentes métodos de clusters

Cada método identifica clusters de diferentes "formas“

Clusters determinísticos:

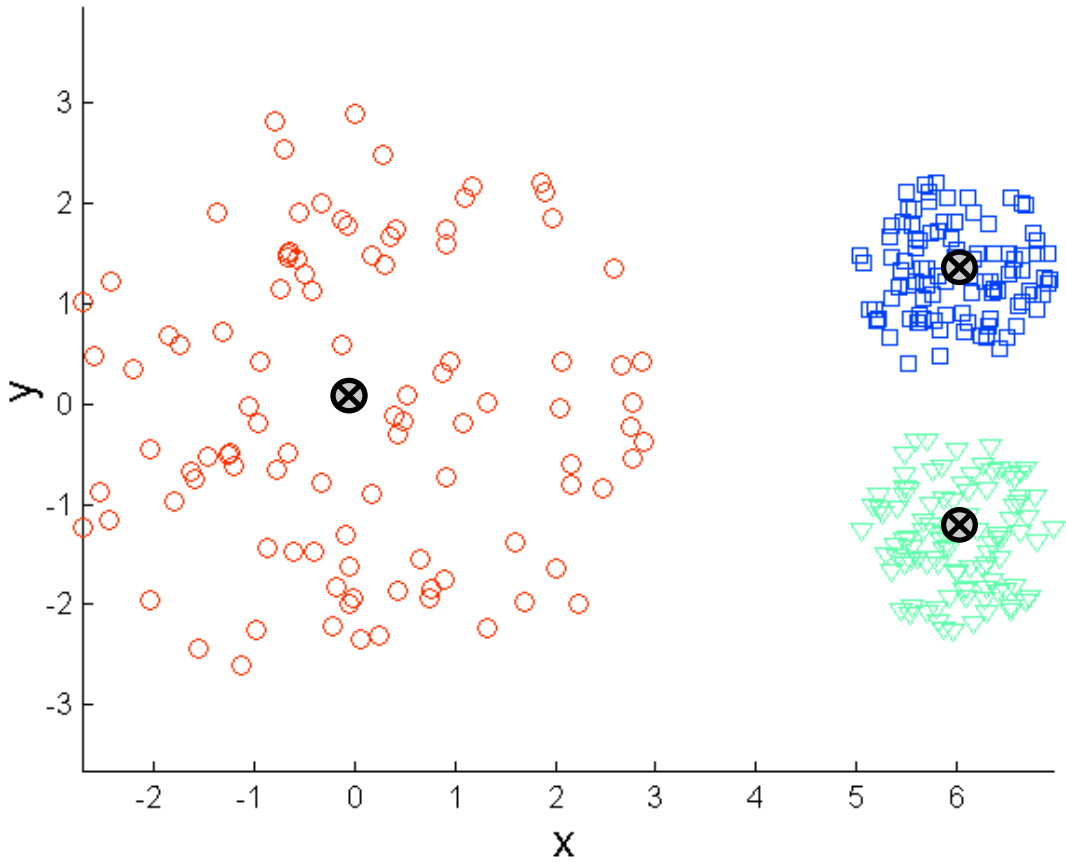
- Métodos basados en particiones
- Clusters jerárquicos

Clusters probabilísticos:

- Basados en modelos de probabilidad
- Método de particiones difusas

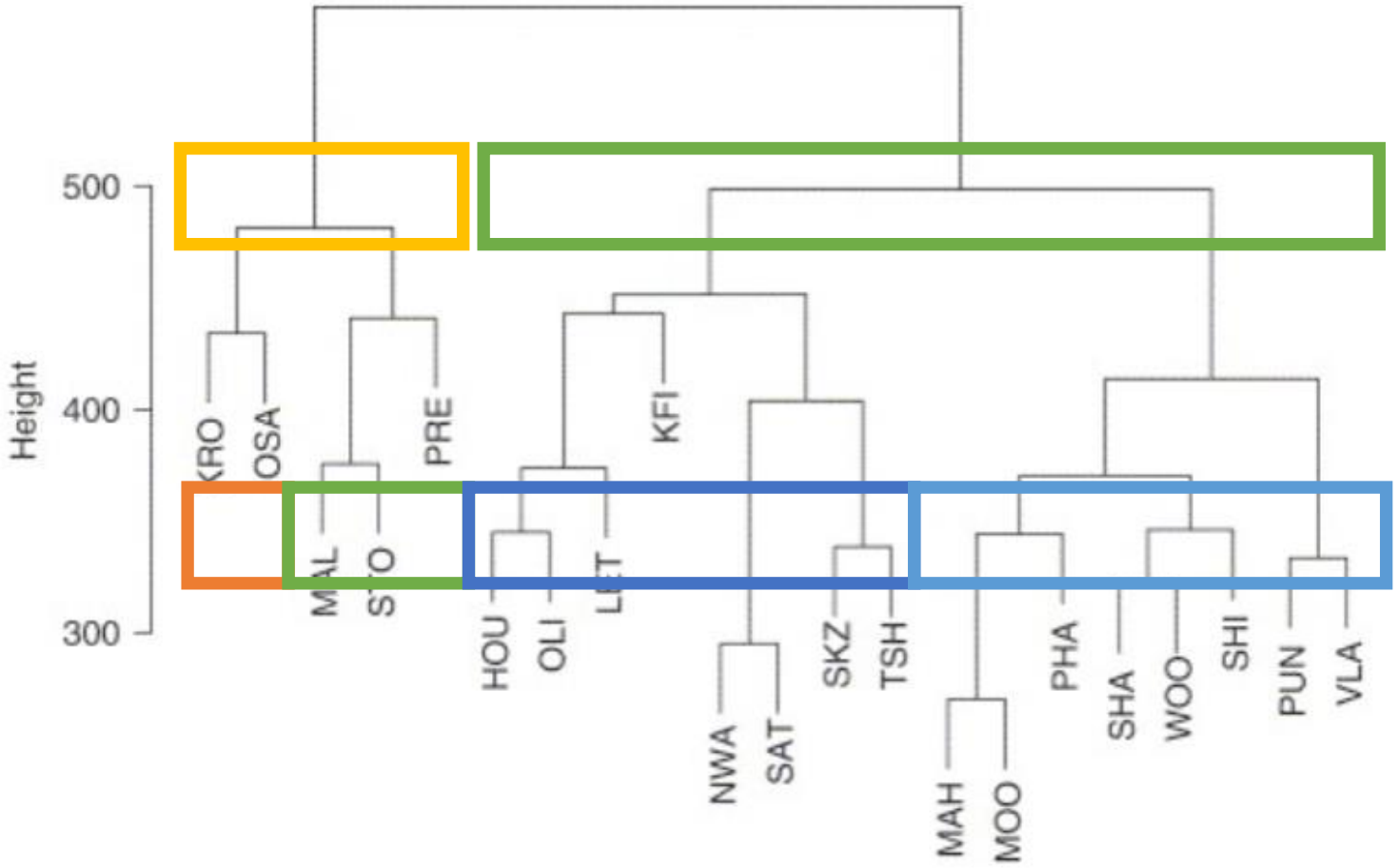
Métodos basados en particiones

Los datos se separan en grupos, a los que pertenece cada punto **exclusivamente** a un solo grupo.



Métodos jerárquicos

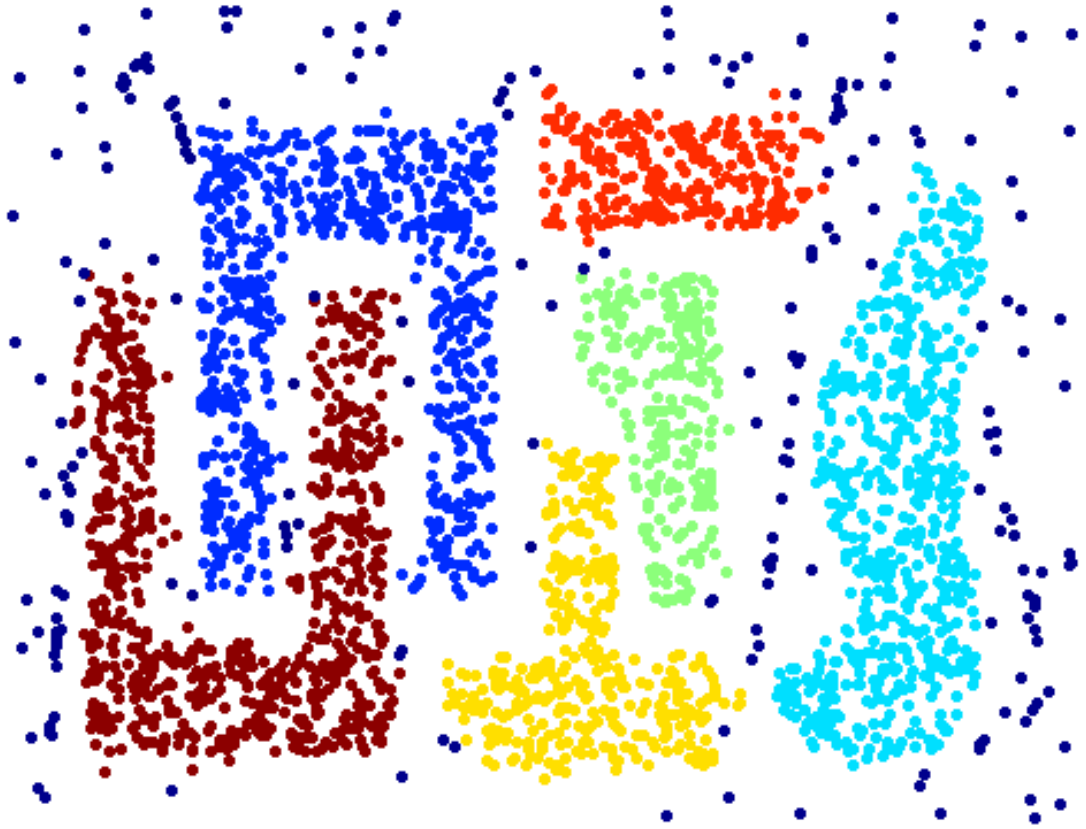
Las entidades se agrupan en una jerarquía de clústeres anidados.



Métodos basados en densidad

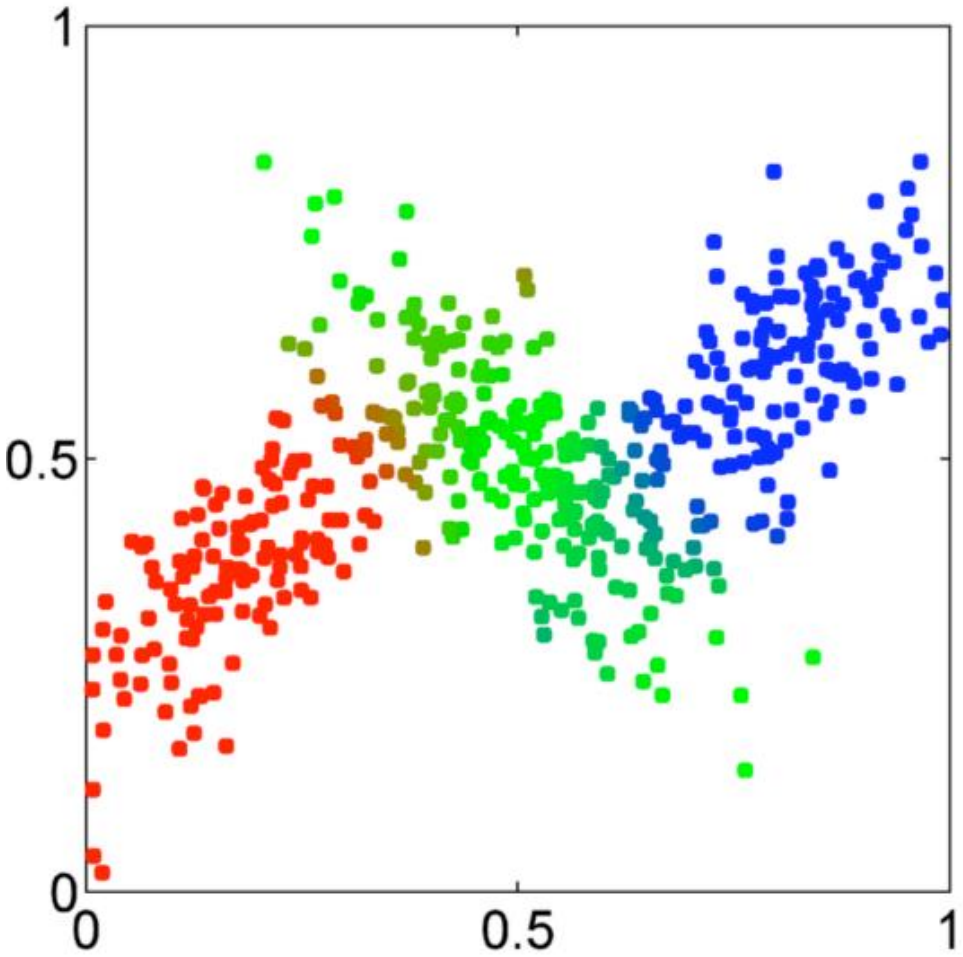
Identifica regiones densas en el espacio.

Genera un modelo de densidad de probabilidad en cada una de estas regiones.



Clusters difusos

Cada punto tiene un **probabilidad** o pertenencia a pertenecer a cada clúster.



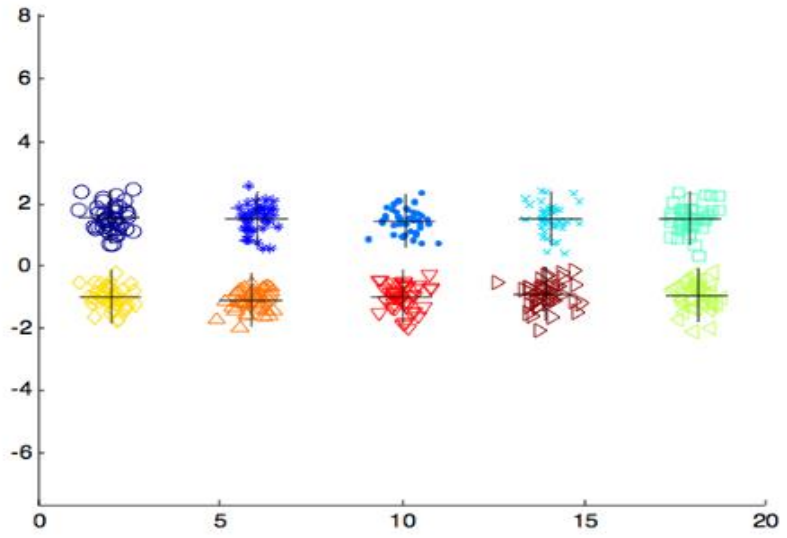
Métodos basados en particiones

K-medias

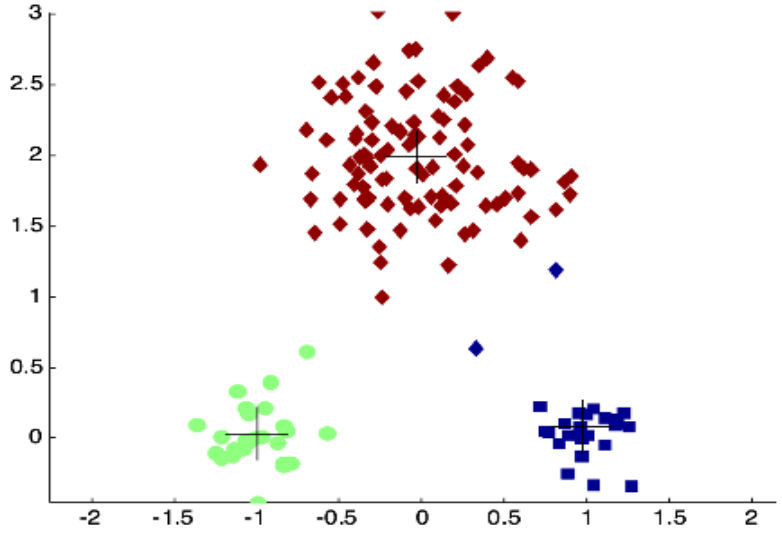
Uno de los algoritmos de clusters más simples.

Dado un número **K** de clusters (determinado por el usuario), cada **cluster** está asociado con un centroide y cada entidad se asigna al cluster con el centroide más cercano.

Variantes como K-medioides, o K-modas usan otros estadísticos como centroides



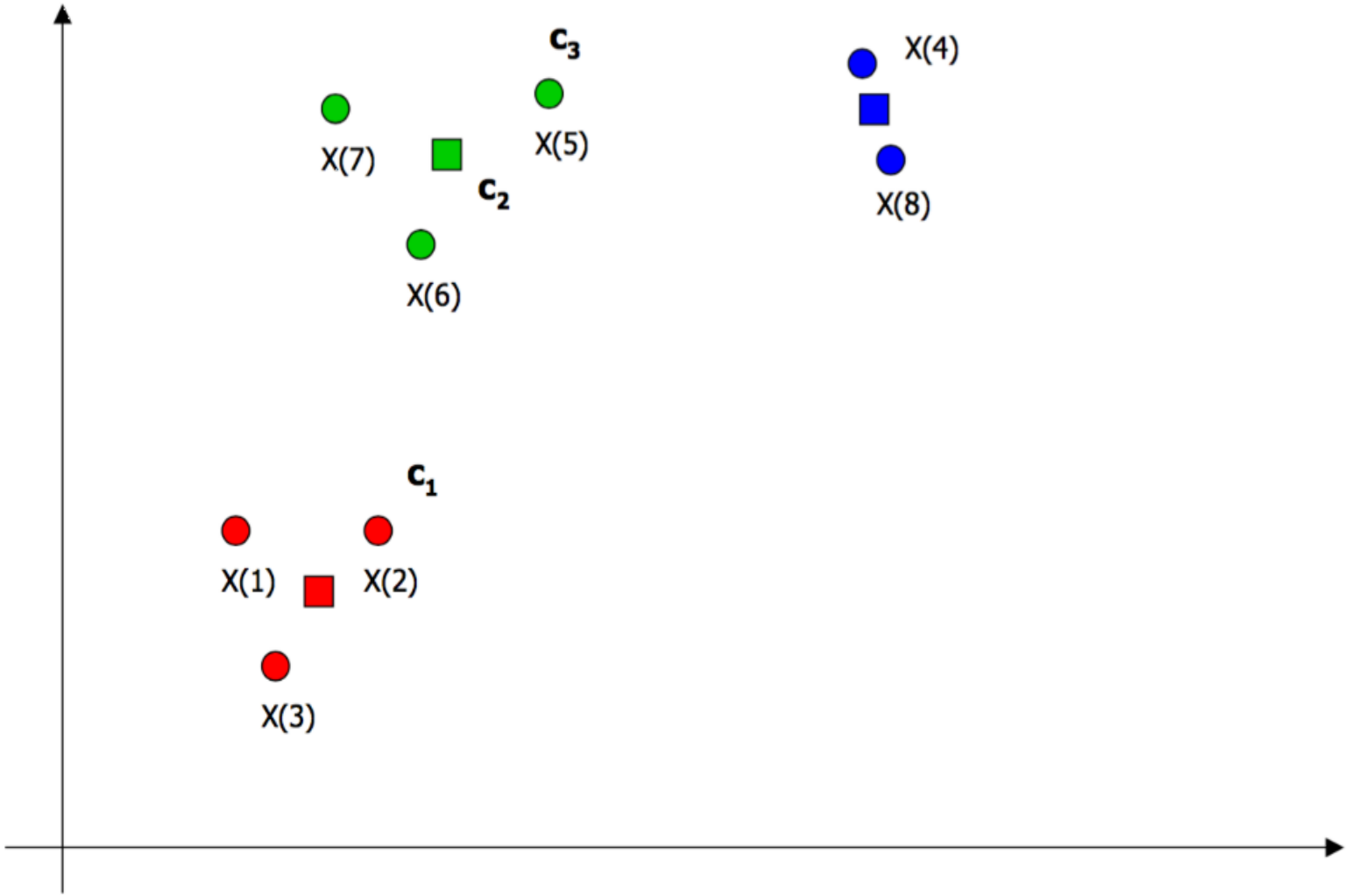
K = 10



K = 3

Algoritmo

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change



Fortalezas:

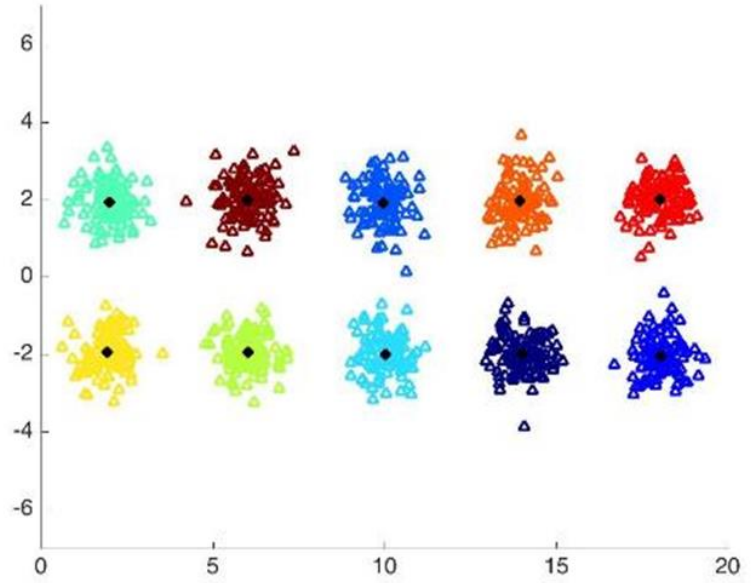
- Relativamente eficiente
- Encuentra grupos esféricos

Debilidades:

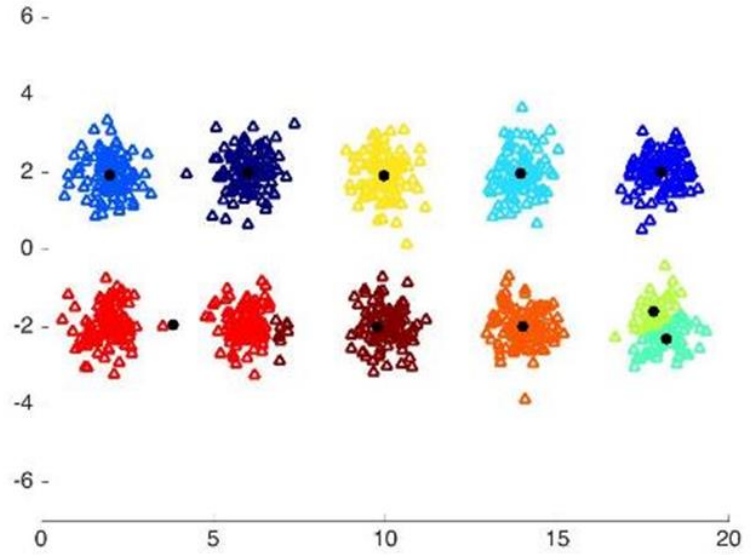
- Termina en el óptimo local
- Sensible a condiciones iniciales
- Aplicable solo cuando la media está definida (variables continuas)
- Necesita especificar K
- No funciona bien con grupos de diferente densidad
- Susceptible a valores atípicos

Estrategias de selección de centroides iniciales

1. Ejecutar k-medias con múltiples inicios aleatorios, elegir el resultado con mejor desempeño.
2. Seleccione K + N centroides iniciales y luego seleccione los K centroides más separados entre si.
3. Seleccione el primer centroide al azar y luego elija los puntos sucesivos que estén más lejos desde el punto anterior.



WC (C) = 491

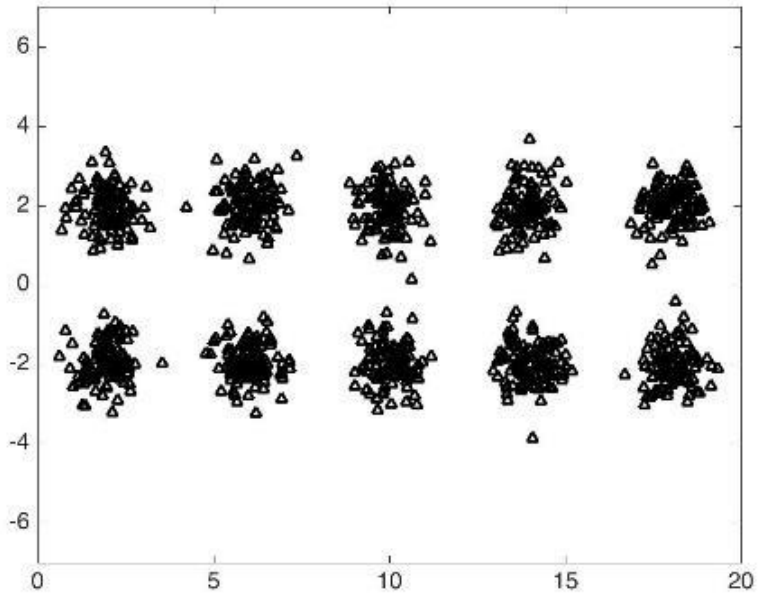


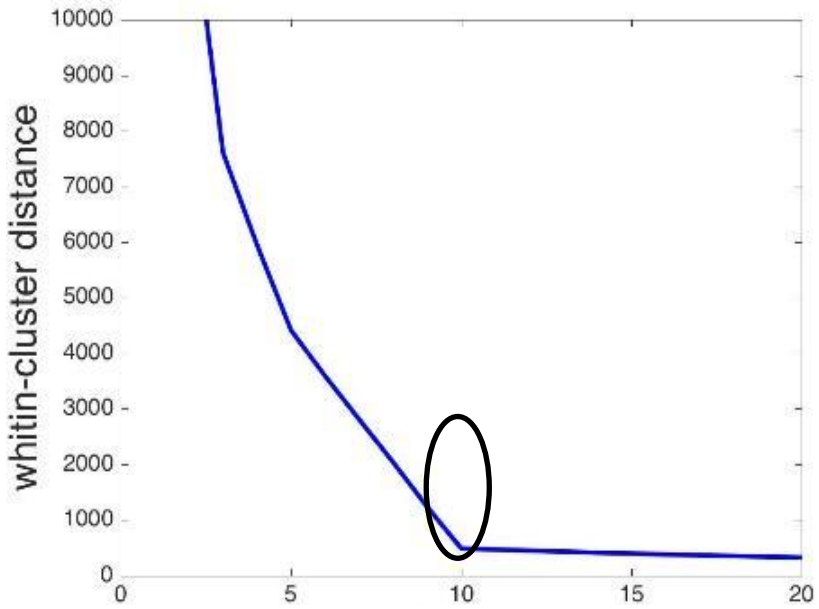
WC (C) = 1304

¿Cómo seleccionar K?

Una heurística para elegir K es aumentar el número de clusters desde 1 hasta N

Se grafica el **desempeño** del algoritmo como una función de K y se busca el punto de inflexión

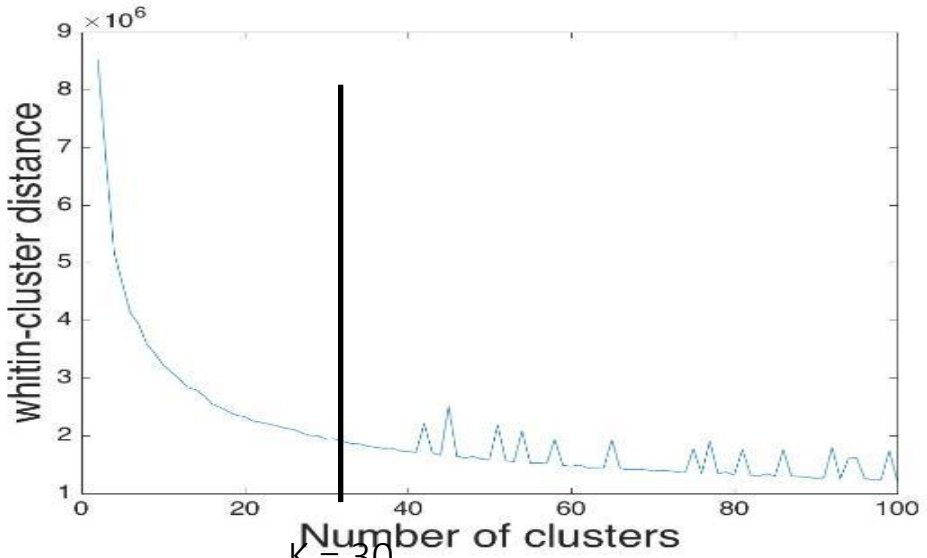




Ejemplo selección K

Cuantificación de color es un proceso que reduce la cantidad de colores distintos utilizados en una imagen,

La nueva imagen debe ser lo más similar posible a la imagen original.



K = 30

49633 colores

30 colores

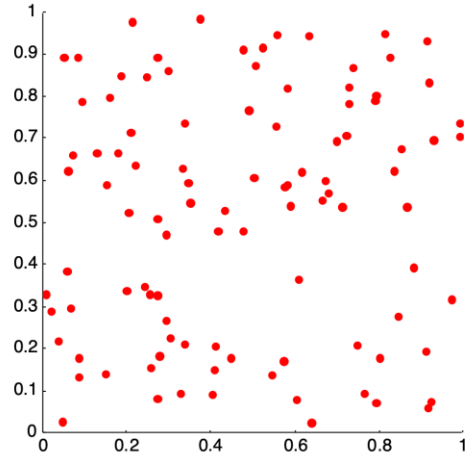
Evaluación de clusters

Evaluar la bondad de los clústeres resultantes

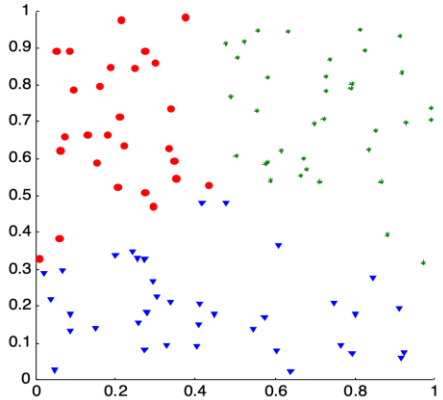
La evaluación de estructuras de clusters es la parte más difícil y frustrante del análisis de clústeres

Hacerlo nos ayudará a:

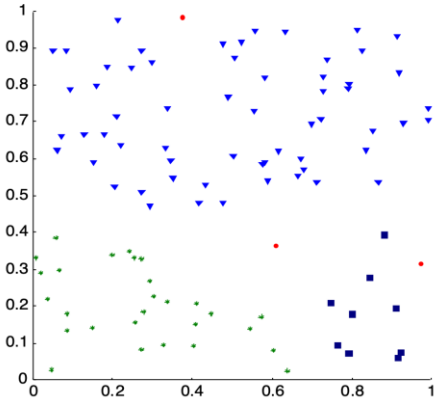
- Evitar encontrar patrones espurios
- Comparar algoritmos de agrupación en clústeres
- Comparar dos conjuntos de clústeres



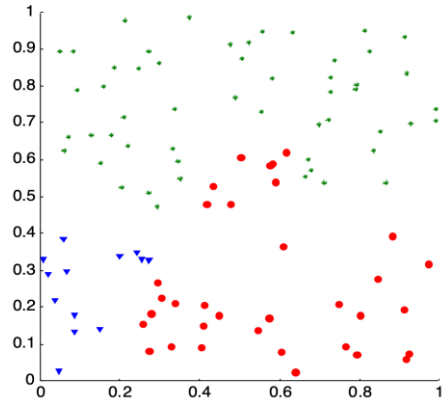
Random Points



K-means



DBSCAN



Complete Link

Evaluación de modelos no-supervisados

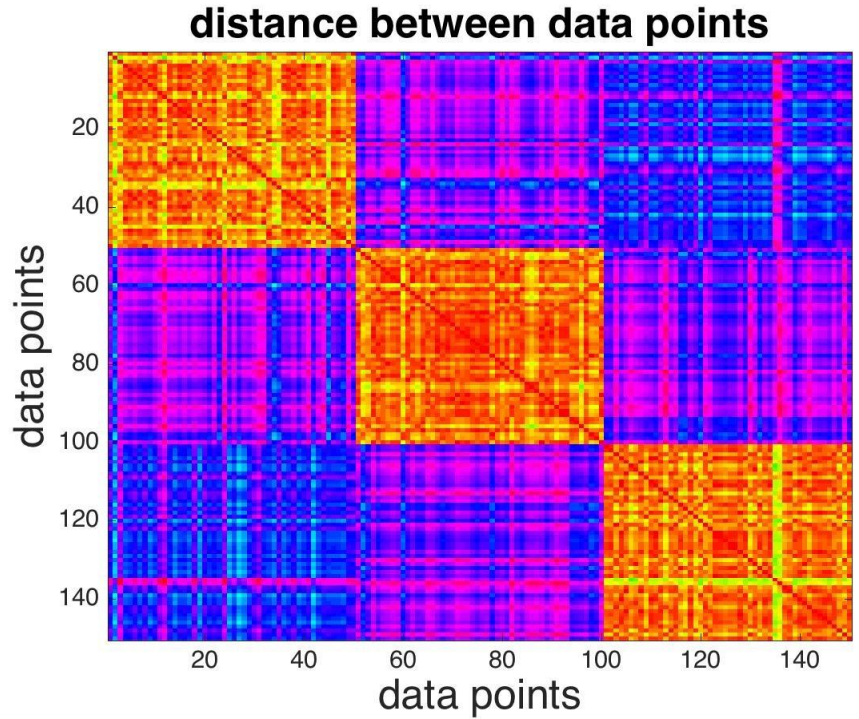
Existen diferentes métodos para evaluar clústeres:

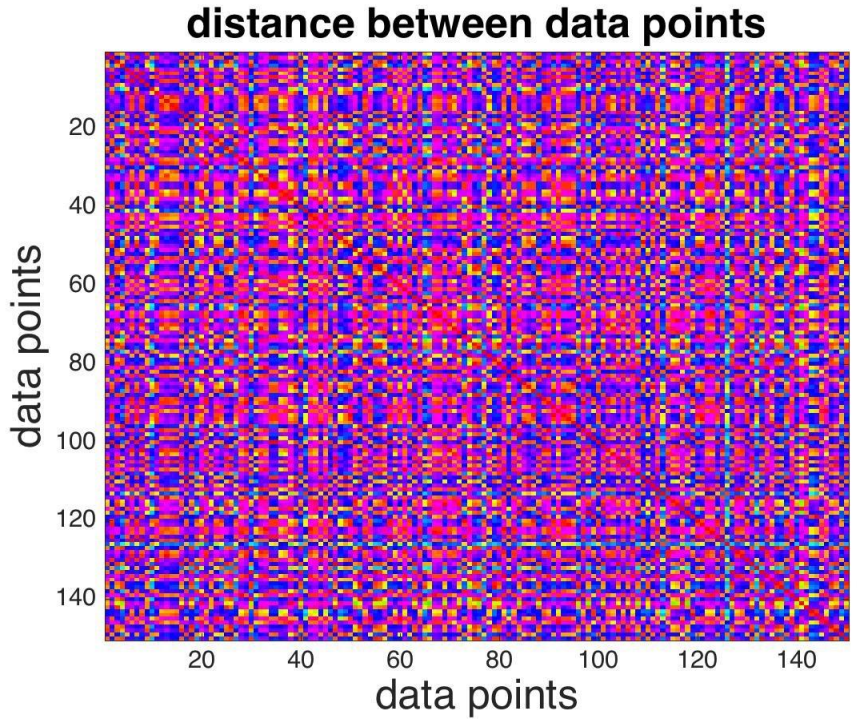
- Inspección visual basada en la matriz de proximidad
- Correlación entre similitud y resultados de agrupación en clústeres
- Estadístico de Hopkins
- Medidas internas: Coeficiente de cohesión, separación y silueta.

Inspección visual

1. Crear la matriz de distancia
2. Ordene la matriz en función de las etiquetas de clúster obtenidas.
3. Inspeccione visualmente

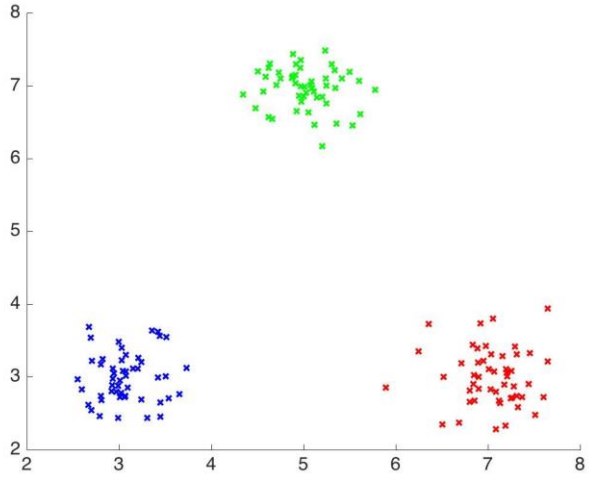
Las buenas agrupaciones exhiben un patrón de bloque claro con "mismo color"



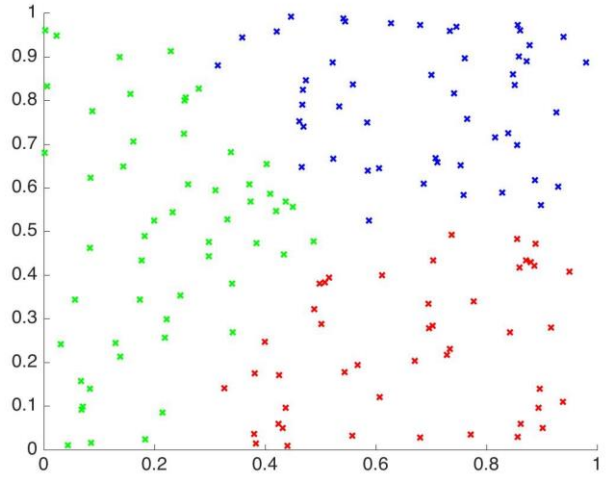


Coeficiente de correlación

1. Construir la matriz de similitud inicial entre todos los puntos s(i,j)=1/(1+d(i,j))
2. Construye la matriz de similitud "ideal" basada en la pertenencia al clúster
3. Calcule la correlación entre la matriz de similitud inicial y la matriz de similitud "ideal" (los ejes X e Y son la similitud inicial/ideal respectivamente).
4. La alta correlación indica que los puntos del mismo clúster están cerca el uno del otro



corr=0.95



corr=0.59

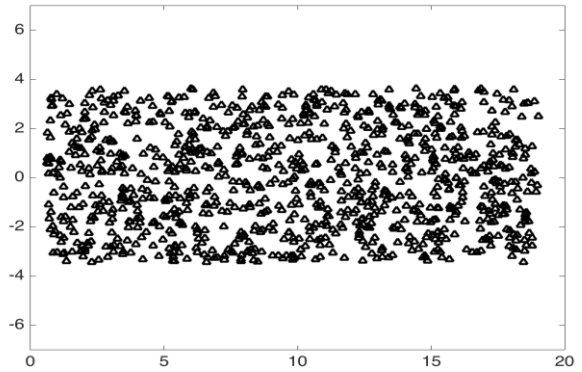
Estadístico de Hopkins

Evalúa la tendencia de los clusters
Mide si un conjunto de datos contiene clusters naturales
Utiliza un test estadístico para la aleatoriedad espacial

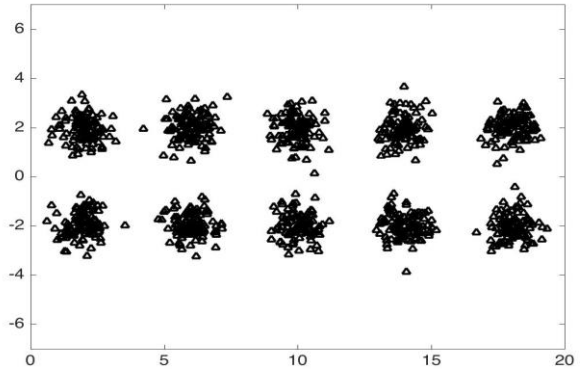
1. Muestrea p puntos a partir del conjunto de datos
2. Generar p puntos aleatorios en el mismo espacio
3. Calcula H, donde:
   - wi: distancia desde el punto aleatorio hasta el vecino más cercano en los datos originales
   - ui: distancia desde el punto de muestra hasta el vecino más cercano en los datos originales

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

Los valores de H cercanos a 0,5 indican datos aleatorios, a 1 indica datos altamente agrupados y a 0 indica una distribución uniforme.



H=0.4975
p=50

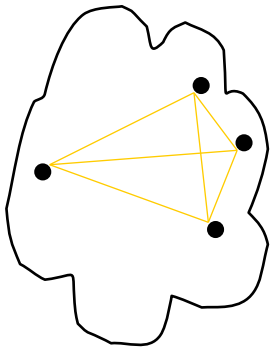


H=0.8191
p=50

Coeficiente de cohesión

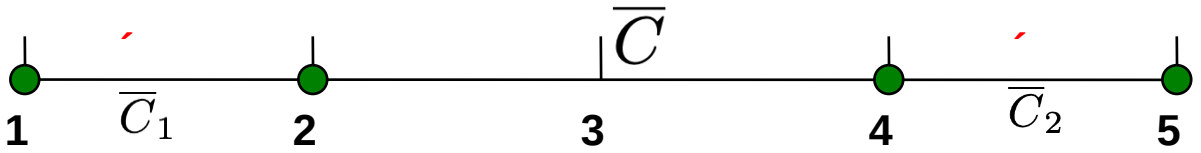
Mide cuán estrechamente relacionados están los objetos dentro de cada clúster.

Suma de errores cuadrados (SSE) es la suma de la distancia cuadrada de un punto al centroide de su clúster.



cohesion

$$SSE_{total} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} (\mathbf{x} - \overline{C}_i)^2$$



$$\text{K=1} \Rightarrow SSE_{total} = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

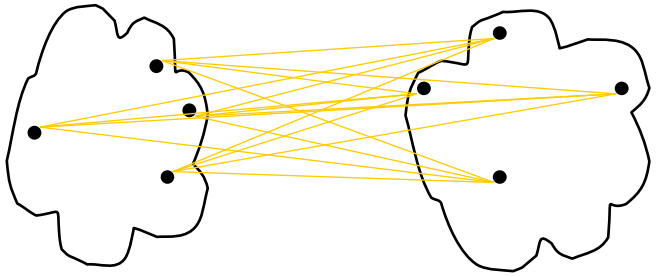
$$\text{K=2} \Rightarrow SSE_{total} = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

Coeficiente de separación

Mide cuán distinto es un clúster de los otros clusters.

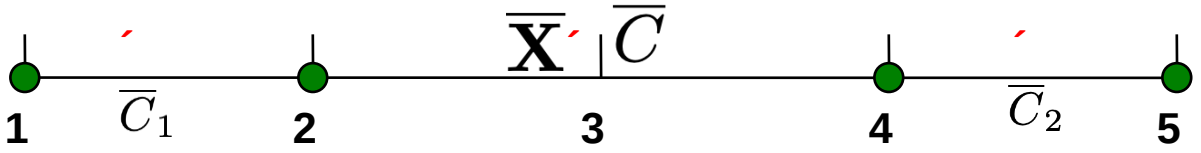
La suma de cuadrados intra grupo (SSB) es la suma de la distancia cuadrada de un centroide de clúster a la media general.

Minimizar la cohesión equivale a maximizar la separación.



separación

$$SSB_{total} = \sum_{k=1}^{K} |C_i|(\overline{C}_i - \overline{\mathbf{X}})^2$$



$$\text{K=1} \Rightarrow SSB_{total} = 4 * (3-3)^2 = 0$$

$$\text{K=2} \Rightarrow SSB_{total} = 2 * (1.5-3)^2 + 2 * (4.5-3)^2 = 9$$

Coeficiente de silueta

Combina cohesión y separación.

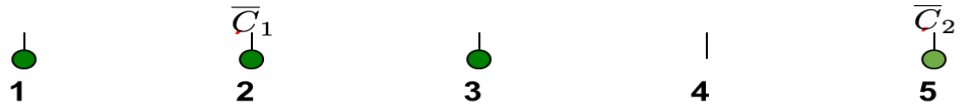
Normalmente varía entre -1 y 1, donde

- Valores cercanos a 1 implican una mejor agrupación en clústeres.
- Valor negativo implica que el punto i está más cerca de otro clúster

Para un punto individual i:

1. Calcular ai como la distancia media de i a puntos en el mismo clúster
2. Calcule bij como la distancia media del punto i a todos los puntos del clúster j.
3. Calcular bi como el bij mínimo excluyendo el clúster propio.
4. El coeficiente de silueta para el punto i es Si = (bi-ai) / max(ai,bi)

El coeficiente de silueta de un clúster es el promedio de los coeficientes de silueta de los puntos pertenecientes.

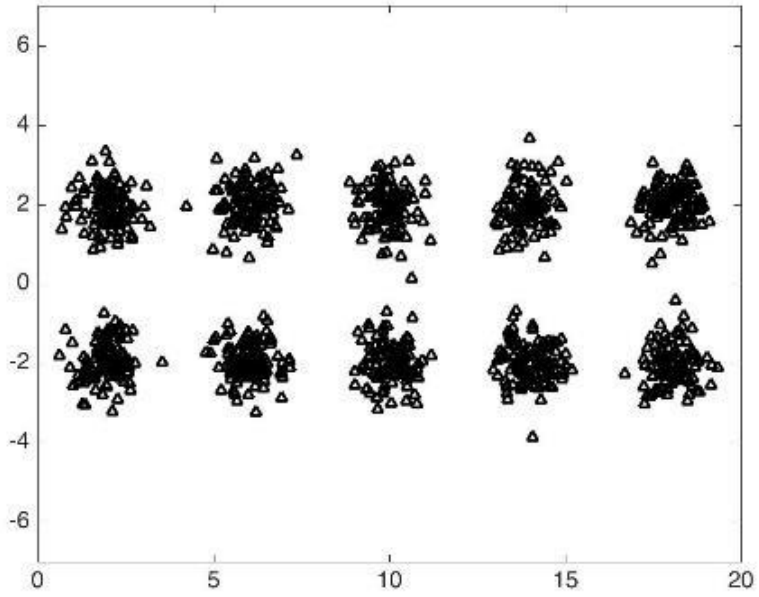


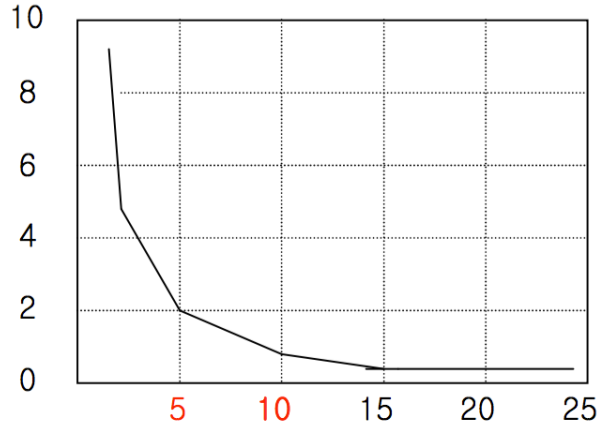
	a_i	b_{i1}	b_{i2}	b_i	s_i
1	1.5	—	4.0	4.0	2.5/4
2	1.0	—	3.0	3.0	2/3
3	1.5	—	2.0	2.0	0.5/2
5	0.0	3.0	—	3.0	1.0

¿Cómo seleccionar K?

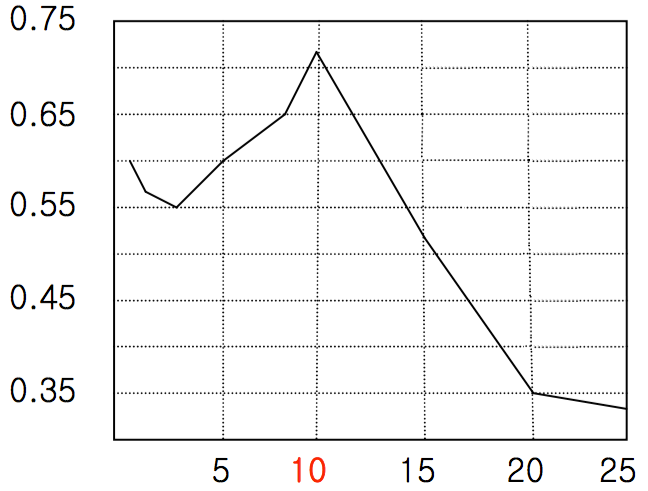
Para determinar el mejor valor de K, se evalúa el índice de silueta y el SSE sobre un rango de K

¿Cual punto de inflexión es mas evidente?





SSE



Silhouette

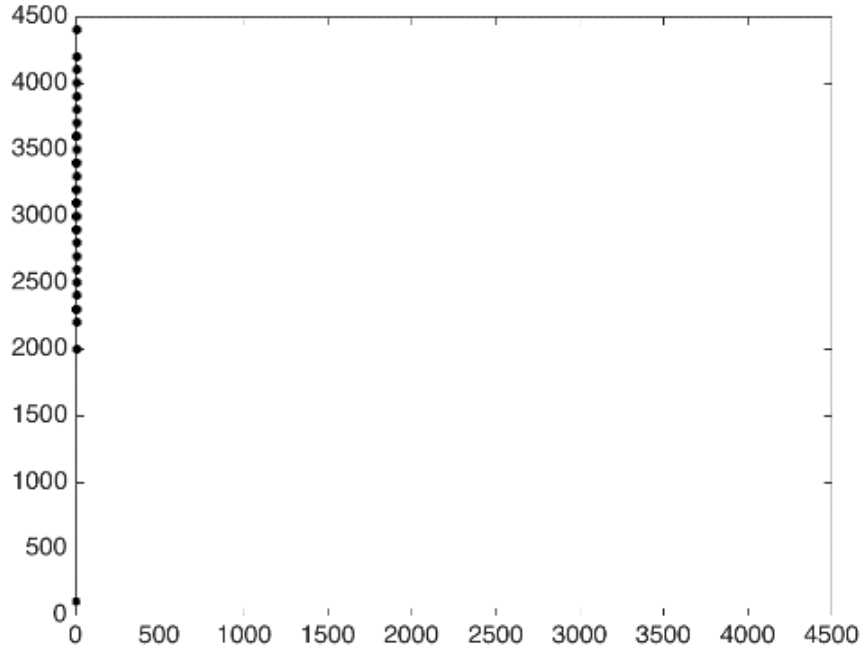
Pre procesamiento

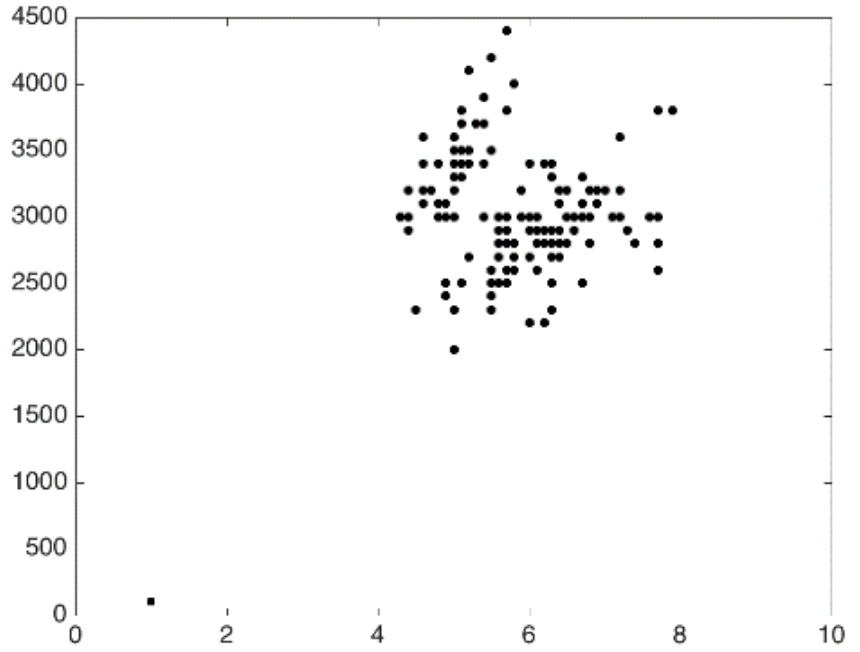
Los valores atípicos afectan el desempeño de la mayoría de los modelos.

La escala de los datos también puede jugar en contra

Por lo tanto, es necesario pre procesar los datos:

- Normalizar los datos
- Lidiar con valores atípicos





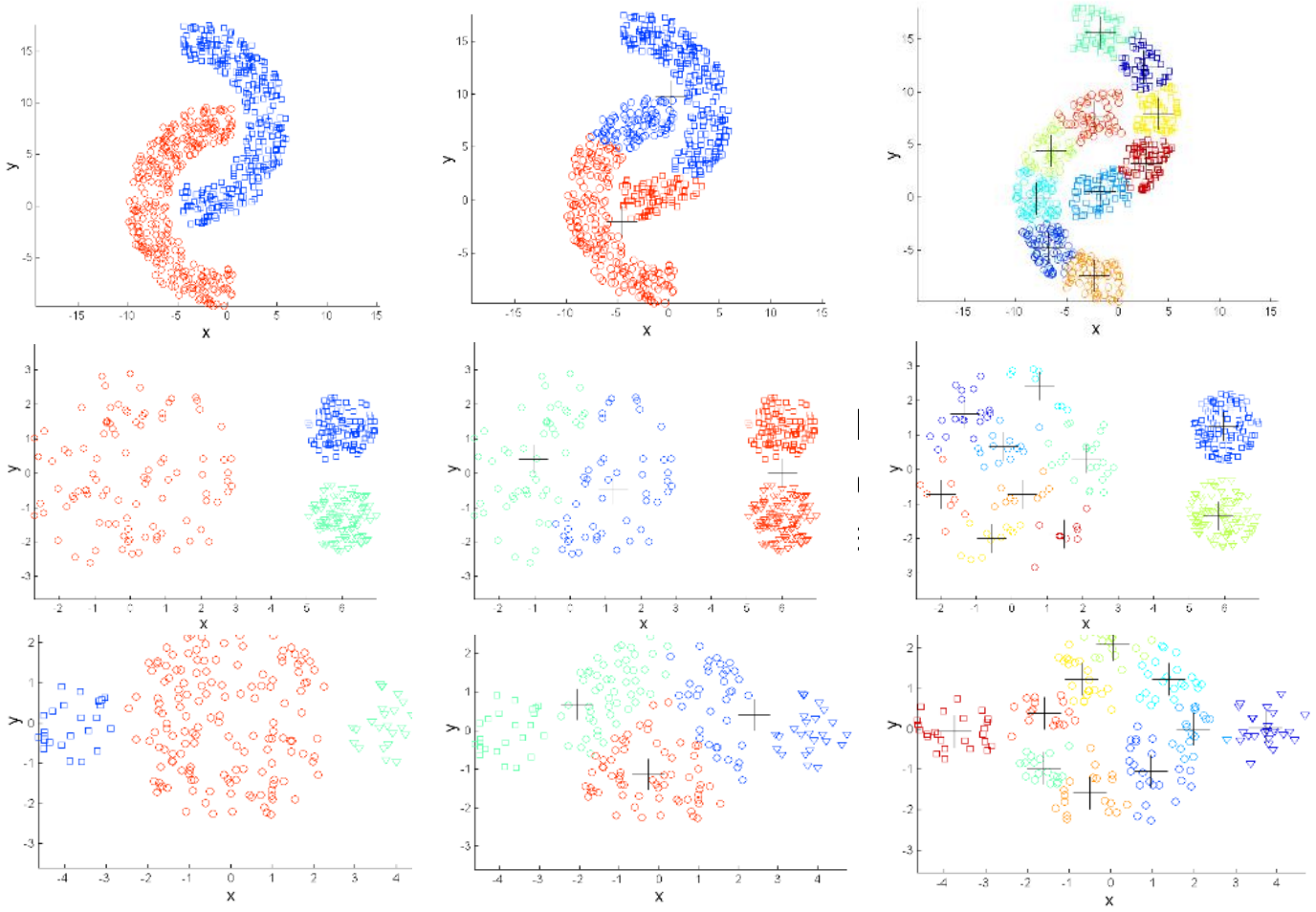
Post procesamiento

Algunos problemas se pueden resolver a través de fusión y división de clusters.

Métodos tradicionales generan clústeres esféricos

La fusión de clústeres mas pequeños podría mitigar este problema.

División de clústeres mas dispersos también mejora el desempeño



Análisis de Clusters

Dr. Raimundo Sánchez

raimundo.sanchez@uai.cl

@raimun2