

CLUSTERING LARGE DATABASES



EL EQUIPO



HERNANDEZ AVILA
HERNAN



MONTESINOS
PEREZ REYNA
ISABEL



MORALES JIMENEZ
NATZIRY VANESSA



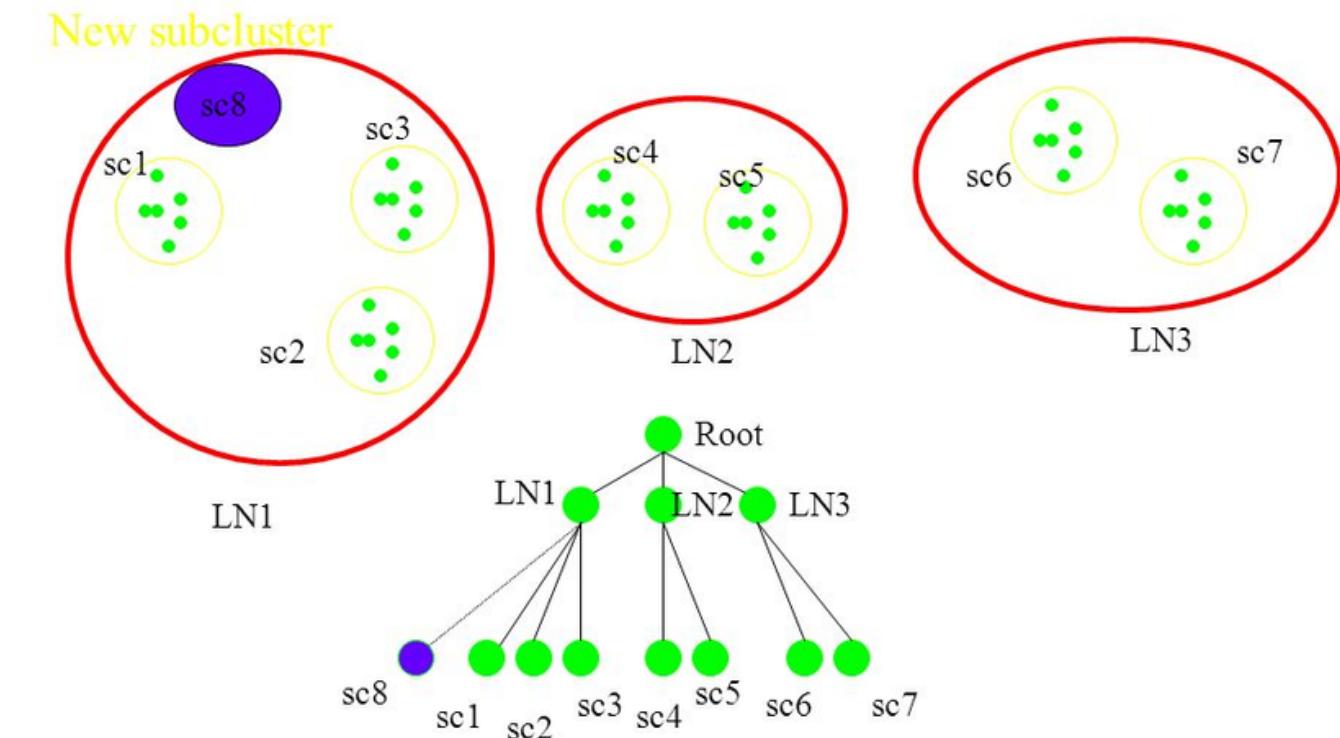
BIRCH

- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) es un algoritmo de clustering jerárquico desarrollado por Tian Zhang, Raghu Ramakrishnan y Miron Livny en 1996.



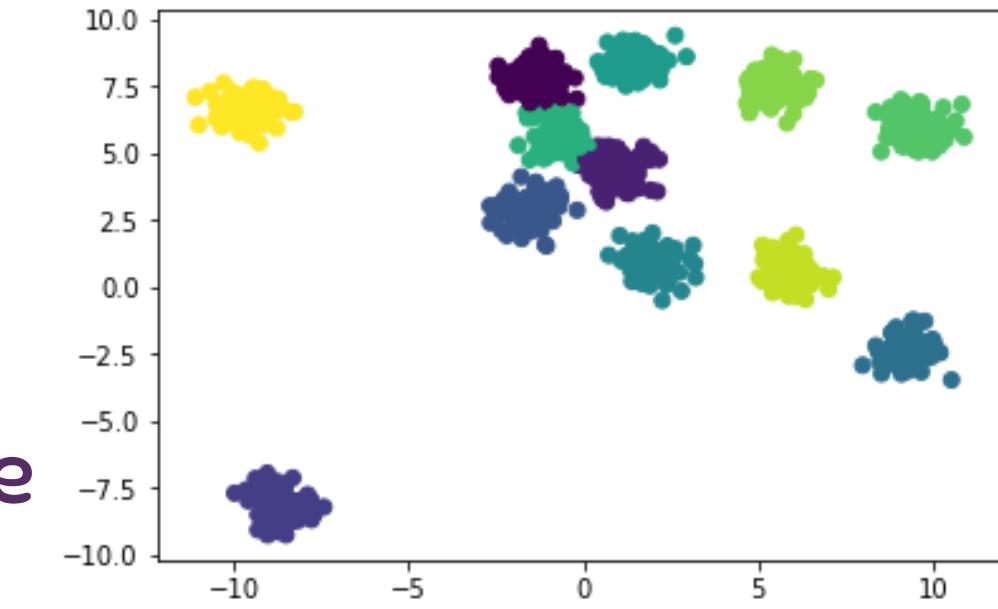
ESTÁ DISEÑADO PARA
MANEJAR GRANDES
CONJUNTOS DE DATOS Y
ES ESPECIALMENTE
EFICIENTE EN TÉRMINOS
DE USO DE MEMORIA.

Example of the BIRCH Algorithm



W W W BIRCH

- La principal característica de BIRCH es que construye un árbol de clústeres llamado Clustering Feature Tree (CFT) que captura la estructura jerárquica de los datos. El árbol se construye de manera incremental y utiliza técnicas de reducción y resumen de datos para reducir la complejidad computacional.



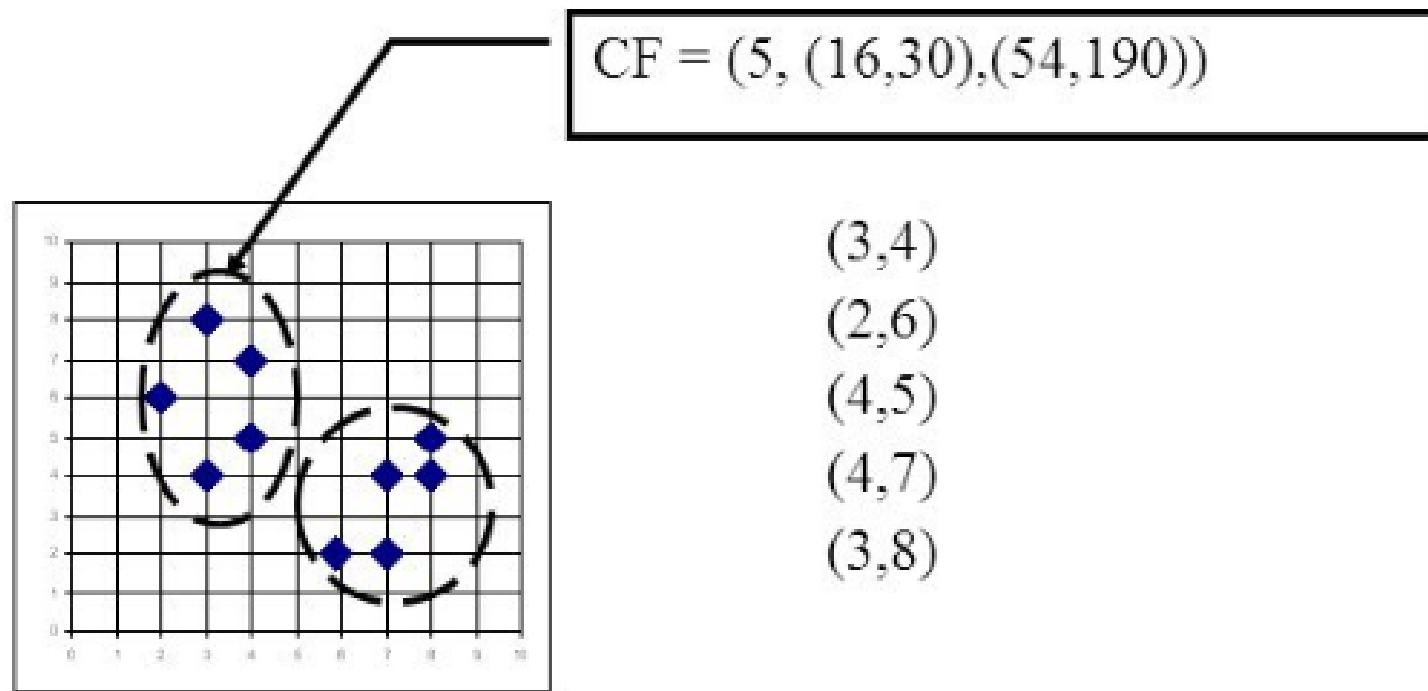
EL ALGORITMO BIRCH SE BASA EN DOS CONCEPTOS CLAVE: CF (CLUSTERING FEATURE) Y CF TREE.

Clustering Feature

Un Clustering Feature: es una representación compacta de un subconjunto de datos que contiene información estadística resumida, como la suma de los vectores de características, la suma de los cuadrados de los vectores de características y el número de puntos.

CF Tree

El CF Tree es una estructura de árbol que almacena y organiza los Clustering Features.



Clustering Feature

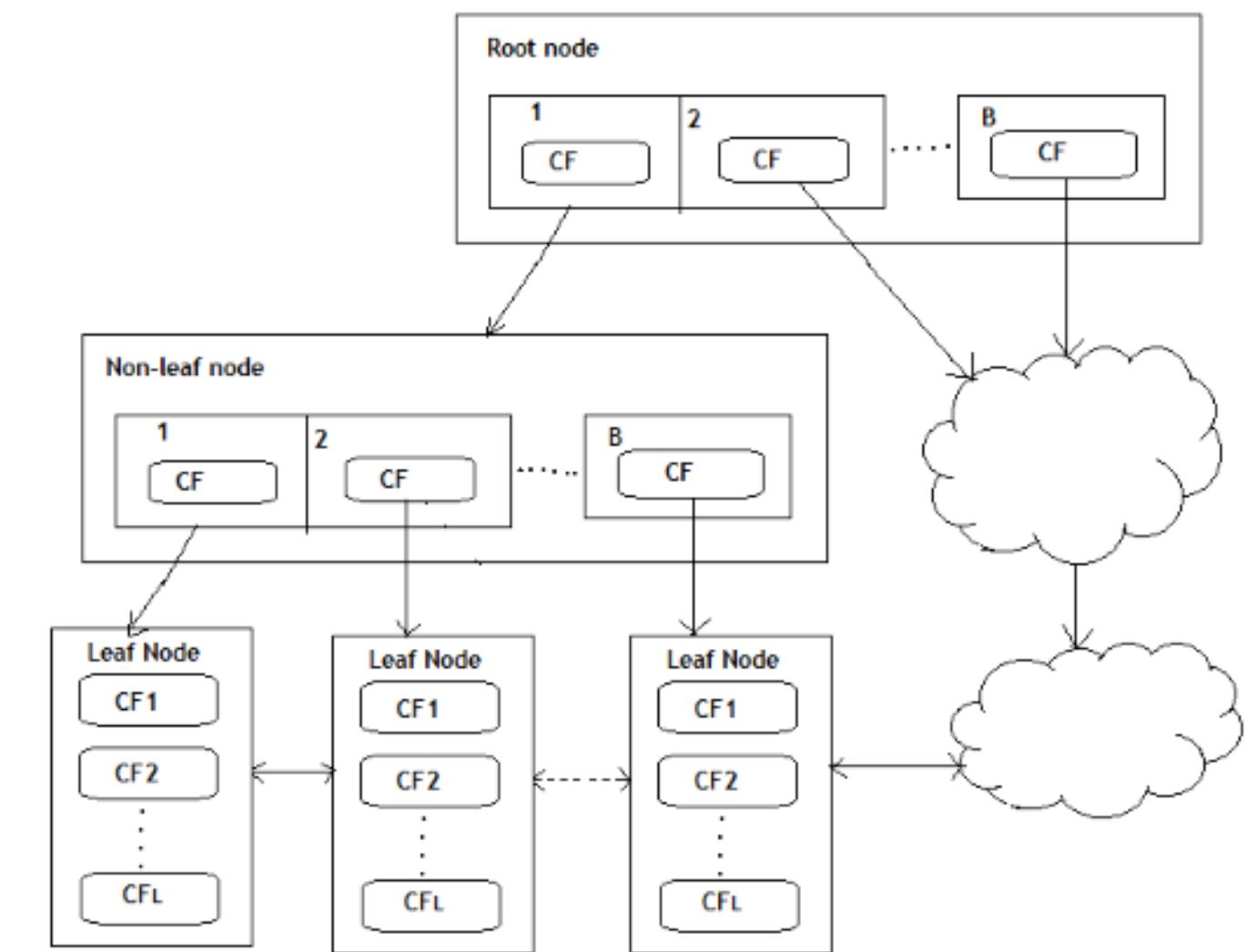


Fig: A CF tree structure

CF Tree



EL PROCESO DE CONSTRUCCIÓN DEL ÁRBOL BIRCH ES ITERATIVO Y CONSTA DE TRES FASES PRINCIPALES:



Construcción inicial del árbol:

Se construye un árbol inicial con una estructura básica utilizando los primeros datos del conjunto de entrenamiento. A medida que se agregan más datos, el árbol se va ajustando y refinando.



Fusión de clústeres:

A medida que los nuevos datos se agregan al árbol, se realiza una fusión de clústeres para mantener el equilibrio y controlar el tamaño del árbol. Esto implica combinar Clustering Features similares y ajustar los valores estadísticos.



Podado del árbol:

Se realiza un podado del árbol para eliminar clústeres redundantes o innecesarios. Esto se logra estableciendo umbrales de tamaño y distancia para determinar qué clústeres deben ser eliminados.

BIRCH

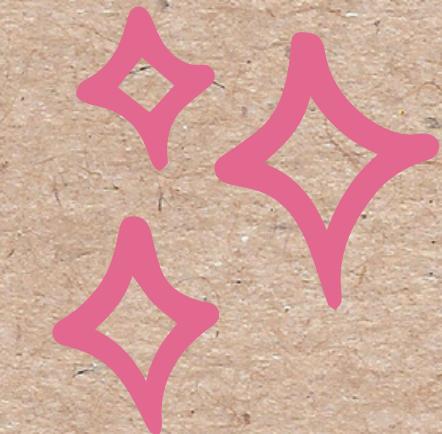
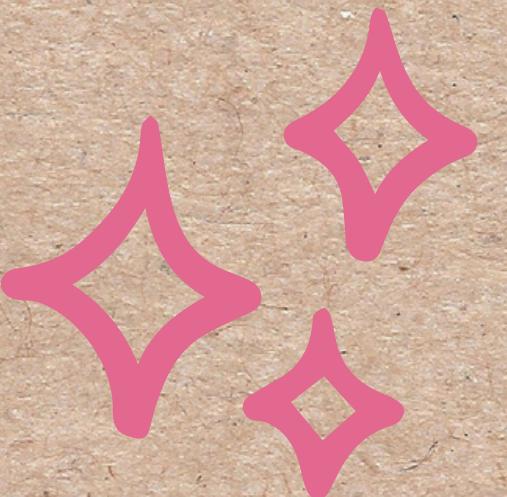
Una vez que se ha construido el árbol BIRCH, se puede realizar la asignación de clústeres a nuevos datos utilizando las estadísticas almacenadas en los Clustering Features y la estructura jerárquica del árbol.

BIRCH es lineal tanto en el espacio como en el tiempo I/O . La elección del valor umbral es imprescindible para una ejecución eficiente del algoritmo. De lo contrario, es posible que el árbol deba reconstruirse muchas veces para garantizar que pueda residir en la memoria.

BIRCH

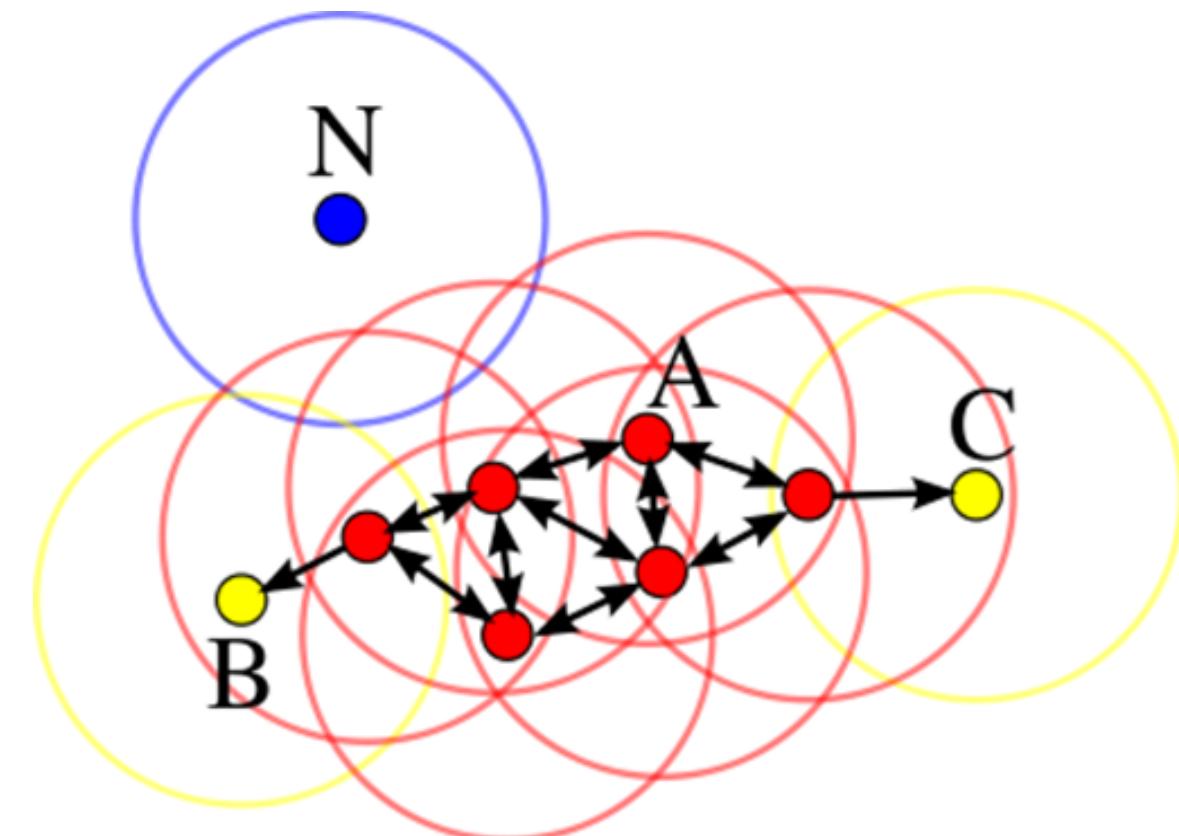


BIRCH es especialmente útil cuando se trabaja con grandes conjuntos de datos, ya que su estructura de árbol permite un acceso eficiente y rápido a los clústeres. Además, su capacidad para reducir y resumir los datos minimiza la necesidad de mantener todo el conjunto de datos en memoria, lo que hace que el algoritmo sea escalable.



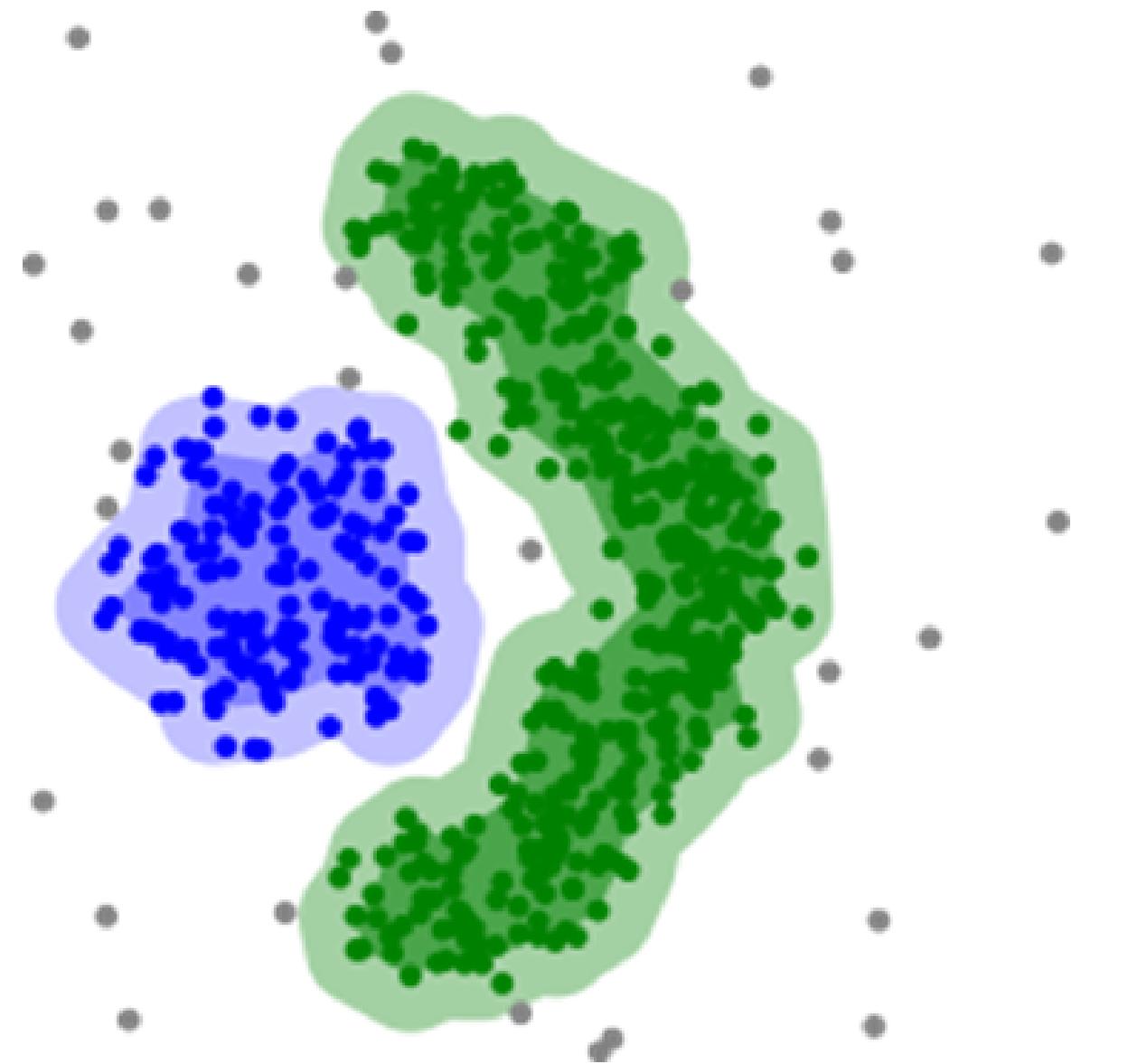
DBSCAN

EL ENFOQUE UTILIZADO POR DBSCAN (AGRUPACIÓN ESPACIAL BASADA EN LA DENSIDAD DE APLICACIONES CON RUIDO) ES CREAR CLUSTERS CON UN TAMAÑO Y DENSIDAD MÍNIMOS.



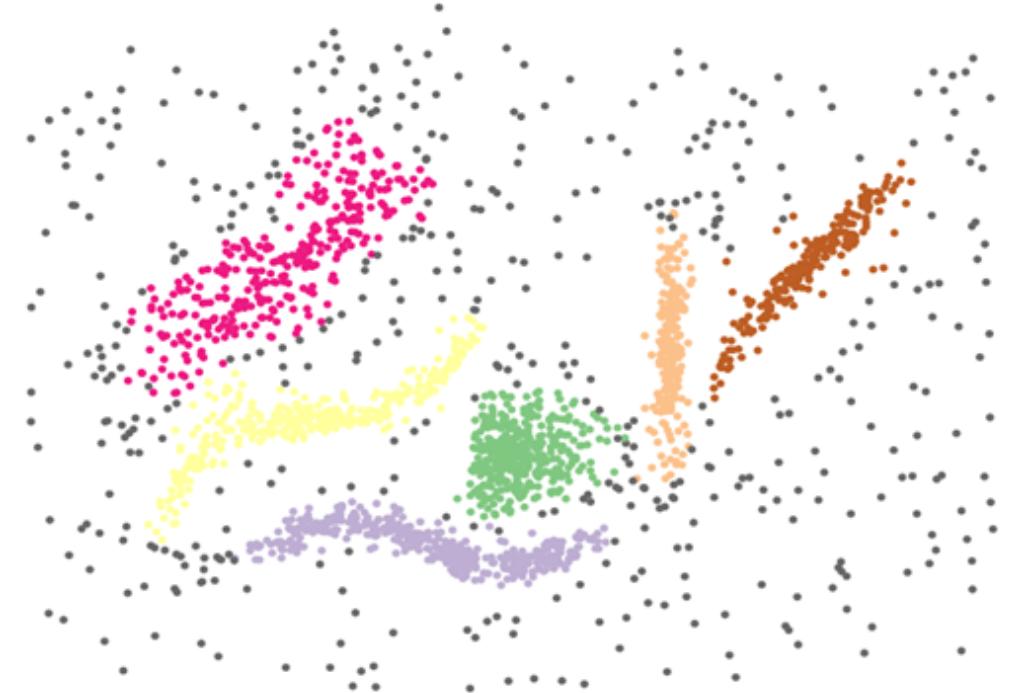
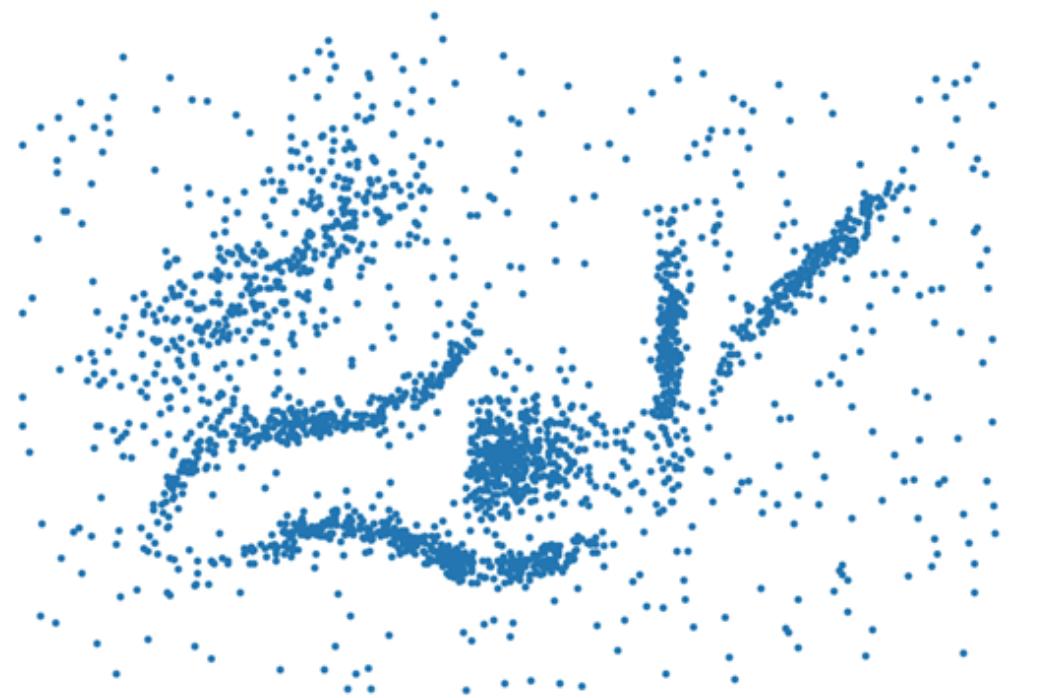
DENSIDAD

La densidad se define como número mínimo de puntos dentro de una cierta distancia entre sí. Esto maneja el problema de valores atípicos asegurándose de que un valor atípico (o un pequeño conjunto de valores atípicos) no creará un grupo.

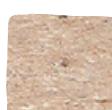


PARÁMETRO DE ENTRADA

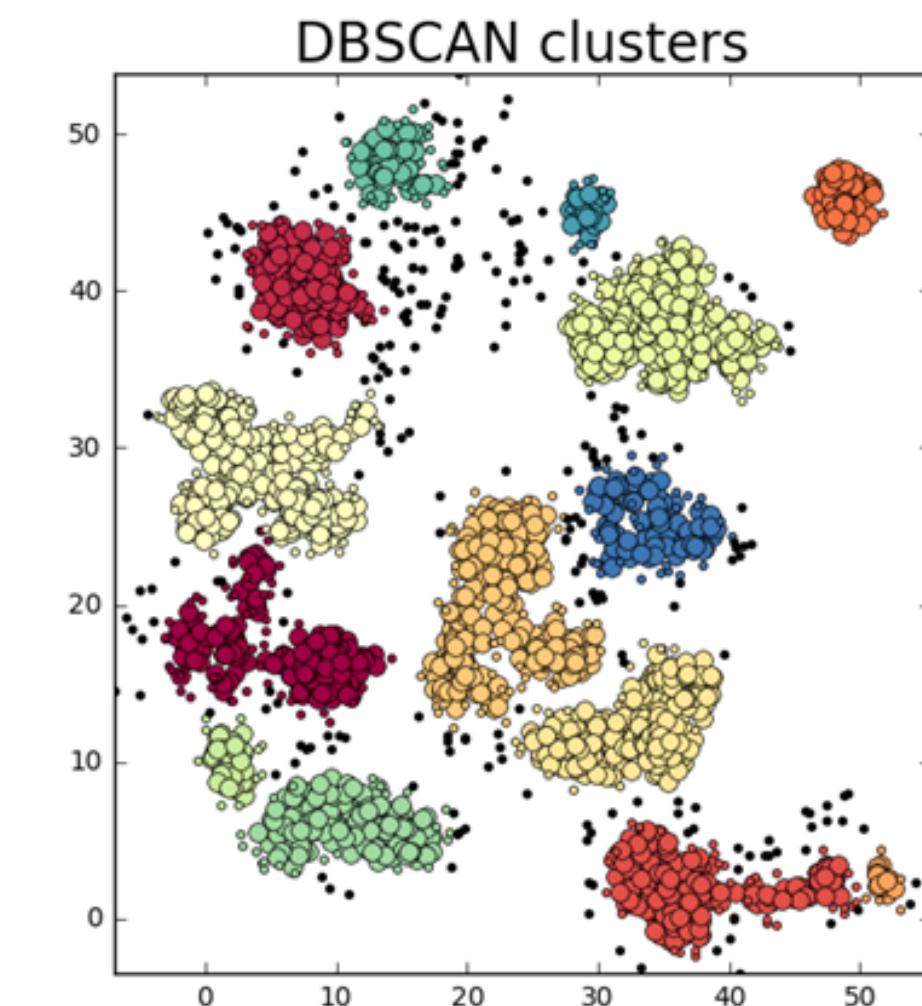
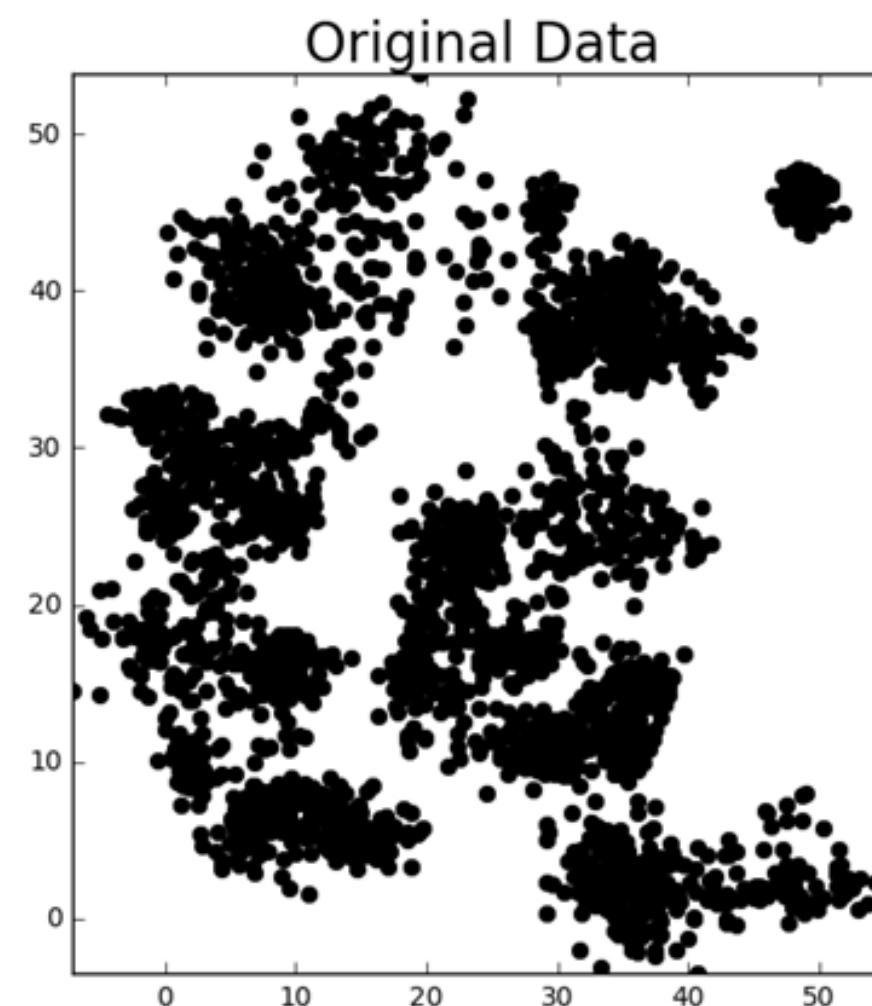
- Un parámetro de entrada, MinPts, indica el número mínimo de puntos en cualquier grupo. Además, por cada punto en un clúster debe haber otro punto en el clúster cuya distancia desde él es menor que un valor de entrada de umbral.



BARRIO DE EPS

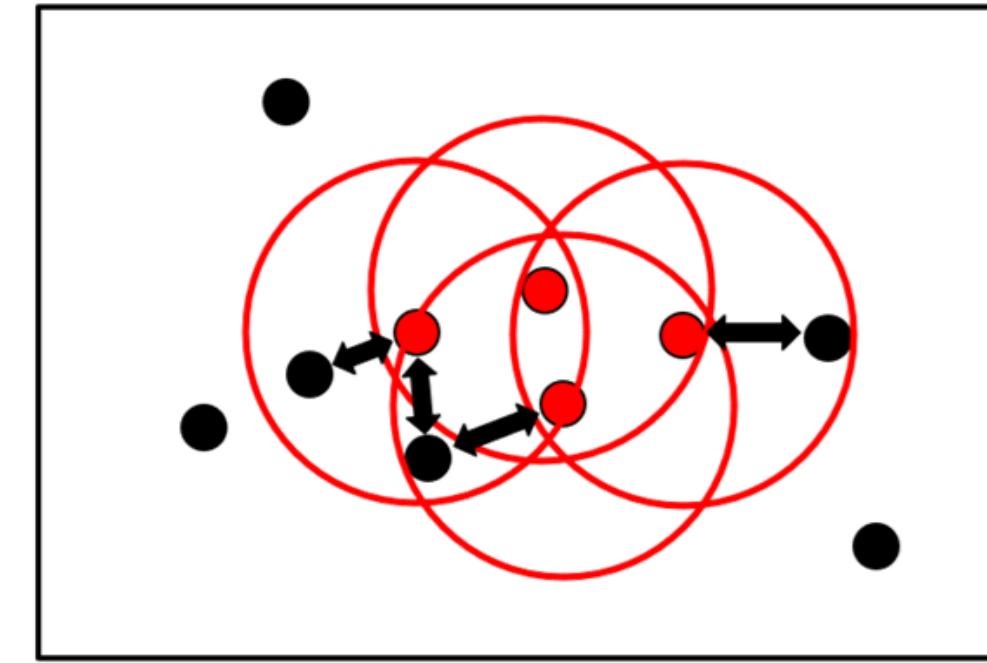
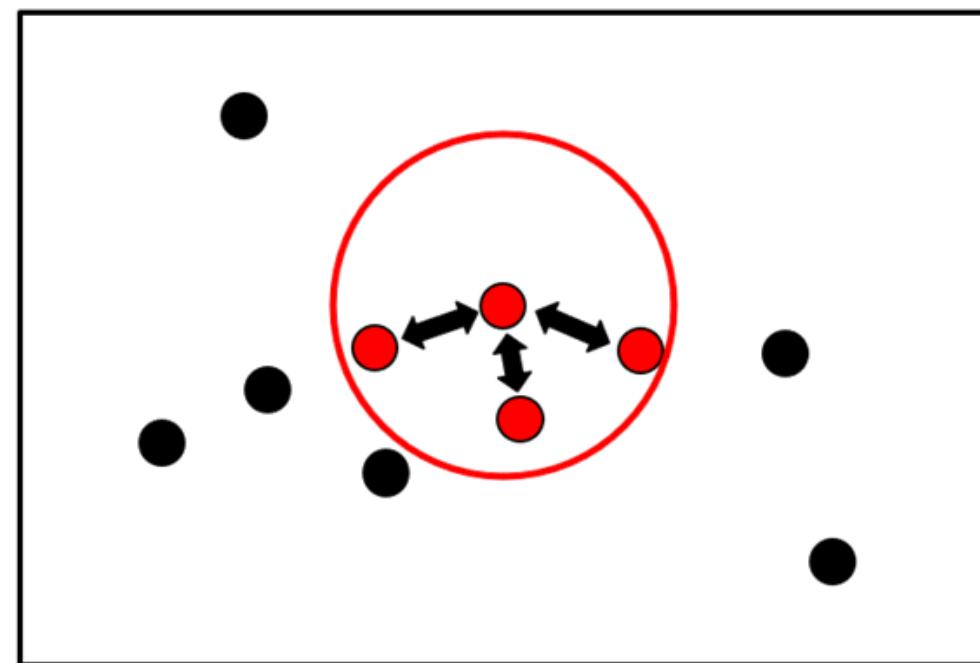


El barrio de Eps o vecindad de un punto es el conjunto de puntos dentro de una distancia Eps. el deseado el número de grupos, k, no se ingresa, sino que lo determina el propio algoritmo.



DBSCAN

La primera parte de la definición garantiza que el segundo punto esté "lo suficientemente cerca" del primero.



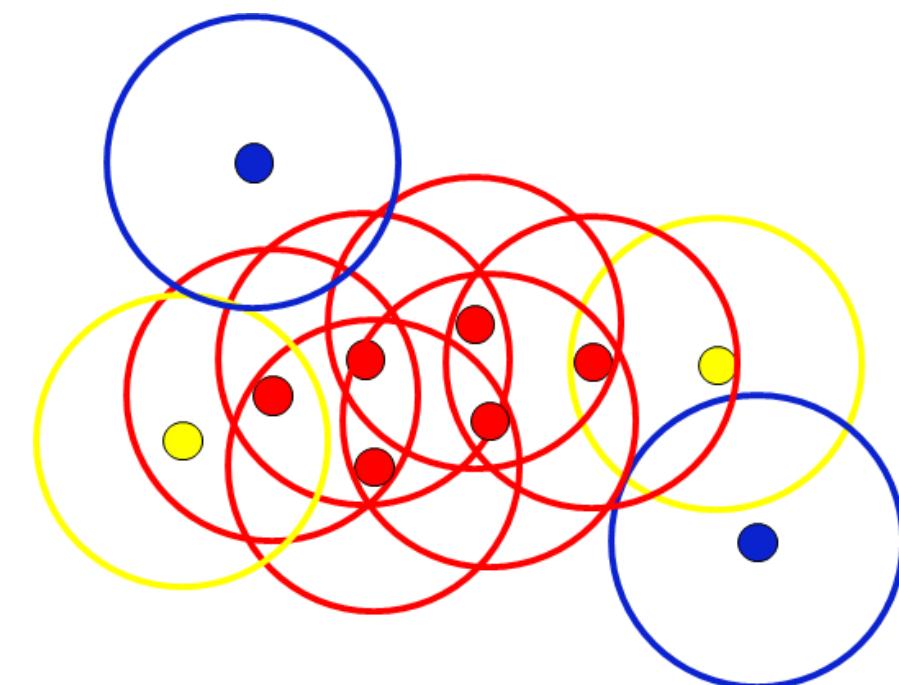
DBSCAN



La segunda parte de la definición asegura que hay suficientes puntos centrales lo suficientemente cerca cada uno de otro. Estos puntos centrales forman la parte principal de un grupo en el sentido de que todos están cerca entre sí.



Puntos centrales



DBSCAN



Un punto directamente alcanzable por densidad debe estar cerca de uno de estos puntos centrales, pero no necesita ser un punto central en sí mismo. En ese caso, se llama punto fronterizo.



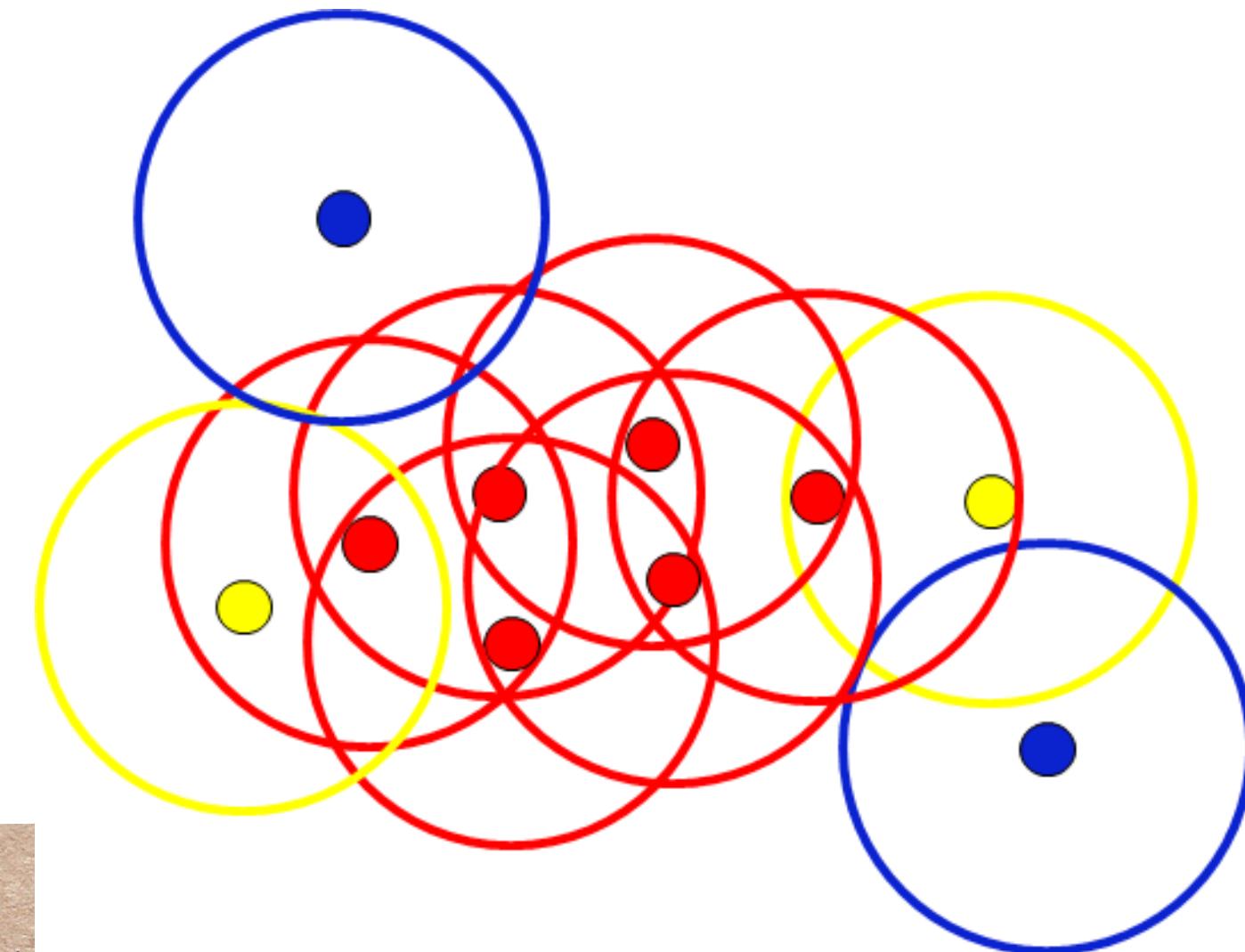
Puntos CENTRALES



Puntos FRONTERIZOS

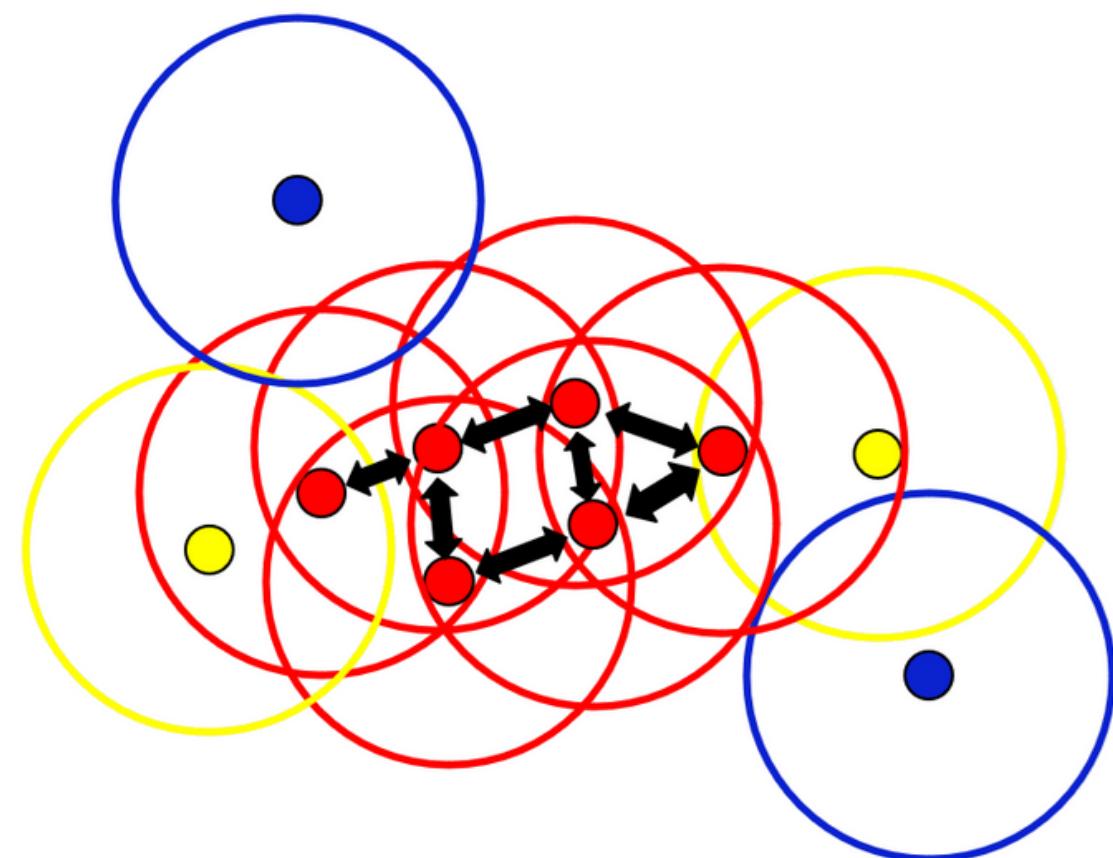


Puntos ATÍPICOS



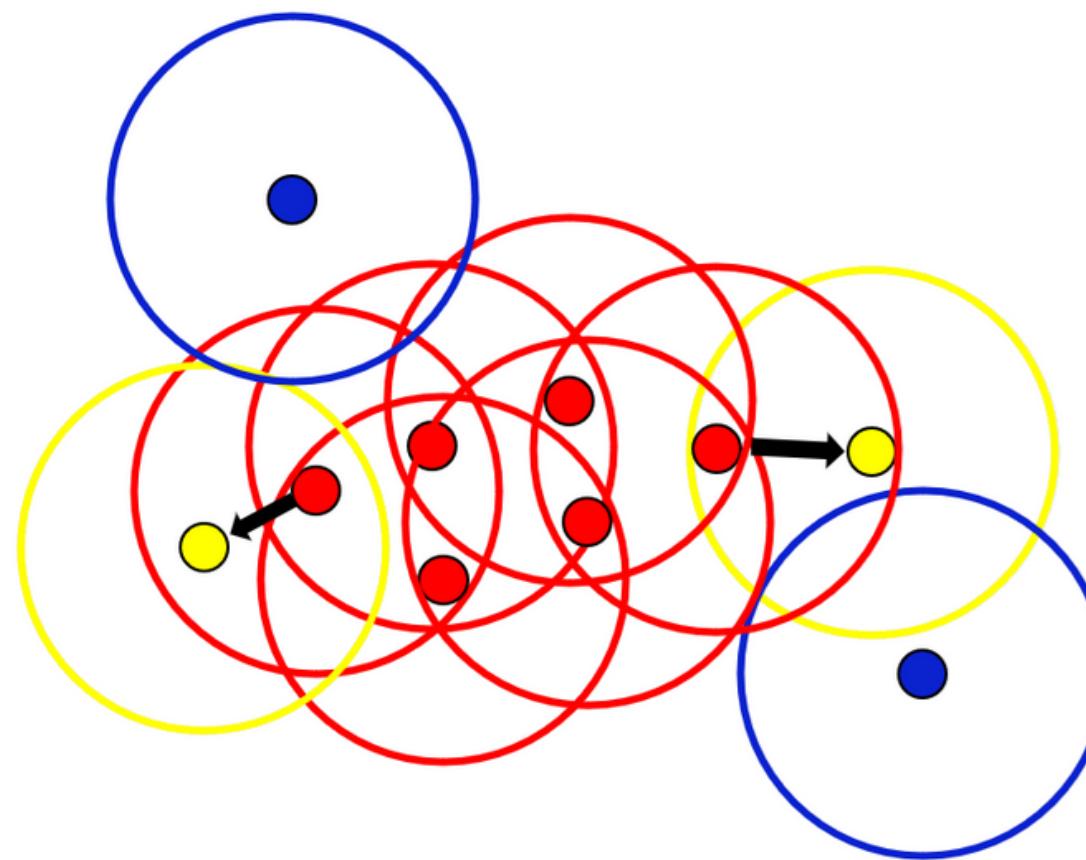
DBSCAN

Un punto es se dice que es alcanzable por densidad desde otro punto si hay una cadena de uno a otro que contiene solo puntos que son directamente alcanzables por densidad desde el punto anterior.

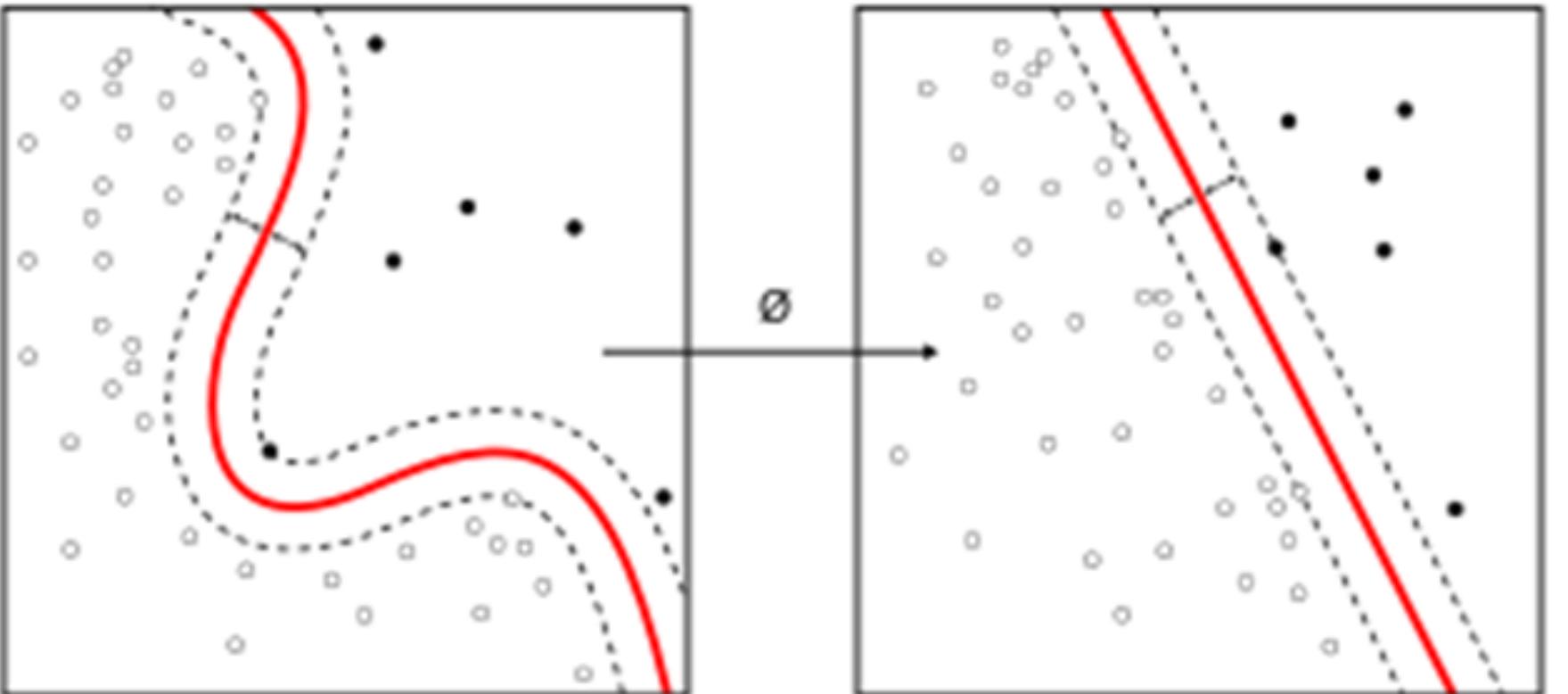


DBSCAN

Este garantiza que cualquier clúster tendrá un conjunto central de puntos muy cerca de un gran número de otros puntos (puntos centrales) y luego algunos otros puntos (puntos fronterizos) que son suficientemente cerca de al menos un punto central.



CLUSTERING USING REPRESENTATIVES



CURE ALGORITHM

Uno de los objetivos del CURE ALGORITHM es el buen manejo de los errores atípicos. Se basa en la selección de más de un elemento representativo de cada clúster.

Jerárquico/Particional

Aglomerativo

Tamaño de grupos



PORQUÉ USAR CURE?

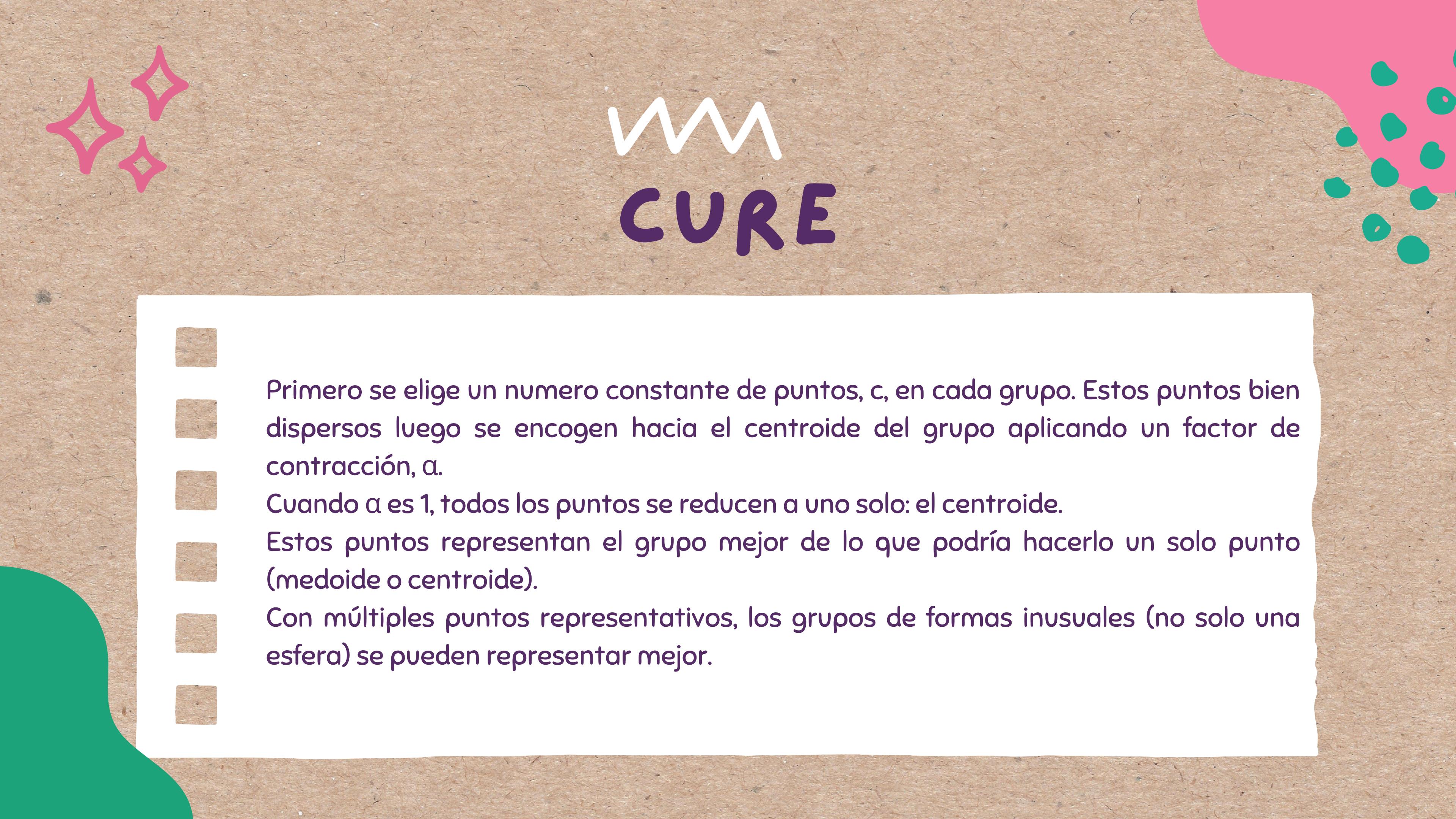
Es una extensión de k-means para clústers y tamaños arbitrarios.

Problema

Clústers están normalmente distribuidos en cada dimensión (ejes fijos).

Solución

Distancia euclídea, Se permite que los clústers adopten cualquier forma



W CURE



Primero se elige un numero constante de puntos, c , en cada grupo. Estos puntos bien dispersos luego se encogen hacia el centroide del grupo aplicando un factor de contracción, a .



Cuando a es 1, todos los puntos se reducen a uno solo: el centroide.

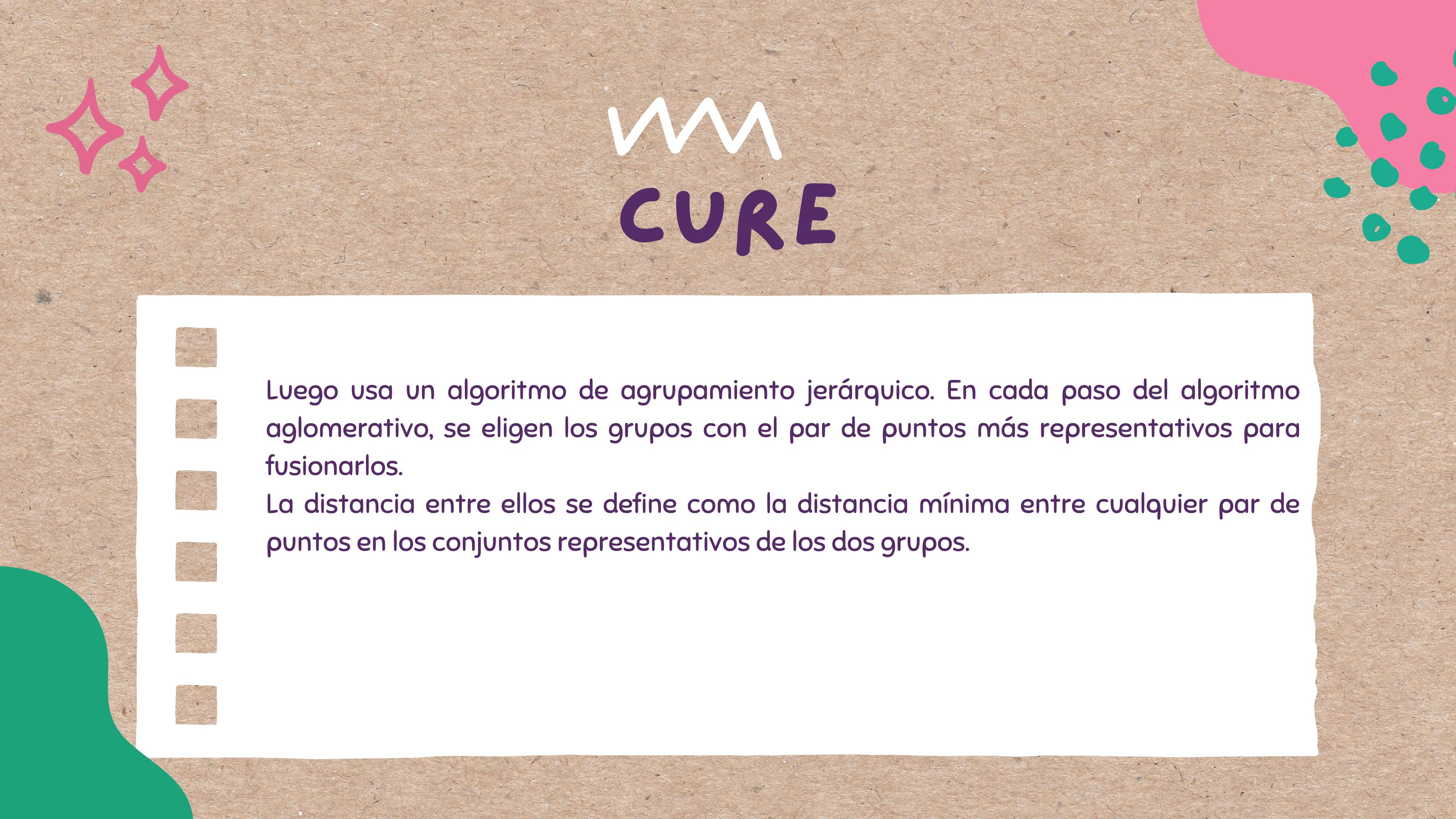


Estos puntos representan el grupo mejor de lo que podría hacerlo un solo punto (medoide o centroide).



Con múltiples puntos representativos, los grupos de formas inusuales (no solo una esfera) se pueden representar mejor.





W CURE



Luego usa un algoritmo de agrupamiento jerárquico. En cada paso del algoritmo aglomerativo, se eligen los grupos con el par de puntos más representativos para fusionarlos.

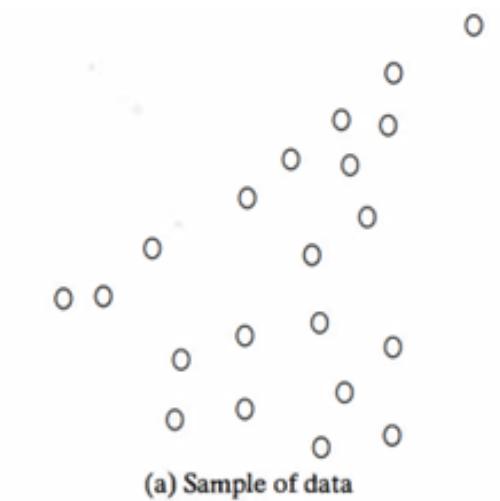


La distancia entre ellos se define como la distancia mínima entre cualquier par de puntos en los conjuntos representativos de los dos grupos.

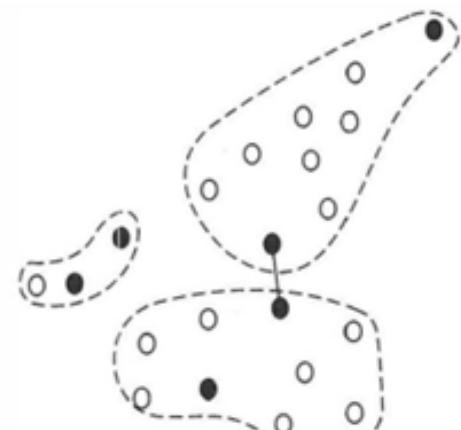


ENFOQUE BÁSICO

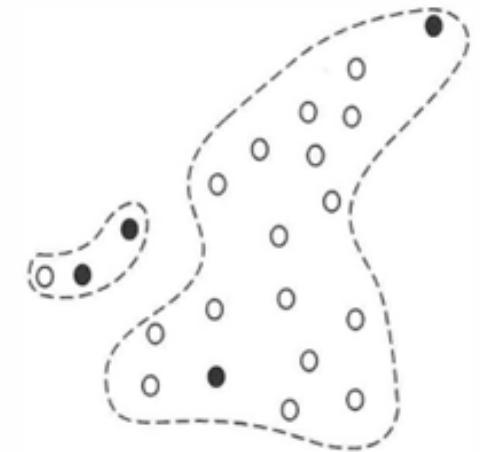
1. Es una muestra de datos
2. Hay tres grupos con dos puntos representativos por cada uno.
3. Se fusionan dos de los grupos y se eligen dos nuevos puntos.
4. Los puntos se contraen hacia la media del conglomerado.



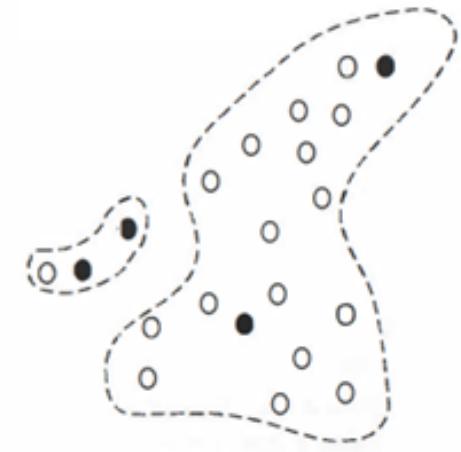
(a) Sample of data



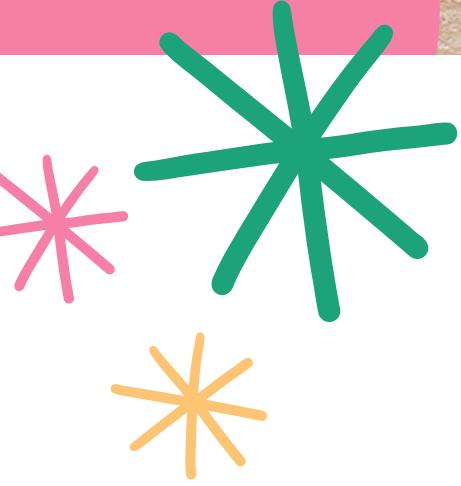
(b) Three clusters with representative points



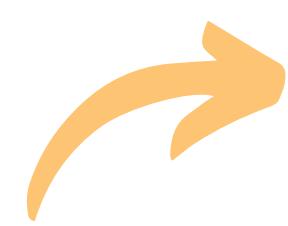
(c) Merge clusters with closest points



(d) Shrink representative points



1. PASO
Obtener una muestra de datos



2. PASO
Divida la muestra en p particiones de tamaño n/p



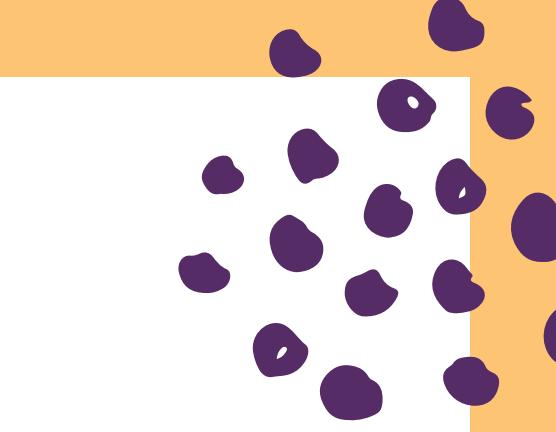
3. PASO
Agrupe parcialmente los puntos en cada partición usando el algoritmo jerárquico El número de grupos es n/pq para alguna constante



4. PASO
Eliminar valores atípicos. Los valores atípicos se eliminan mediante el uso de dos técnicas diferentes



5. PASO
Agrupe por completo todos los datos de la muestra



PASOS

PASOS

6. PASO

Agrupe toda la base de datos en el disco usando c puntos para representar cada grupo. Un elemento de la base de datos se coloca en el grupo que tiene el punto representativo más cercano.

BIBLIOGRAFIA

1.-Dunham, M. H 2002 Data Mining
Introductory and advanced Topics
Capitulo 5.6



GRACIAS

A todos y cada uno de vosotros
por formar parte de este
proyecto.

