

CURSOS ANALYTICS

Machine Learning Advanced

Clustering

Docente: Manuel Montoya



Agenda

1. Clustering
2. Kmeans
3. Clustering Jerárquico

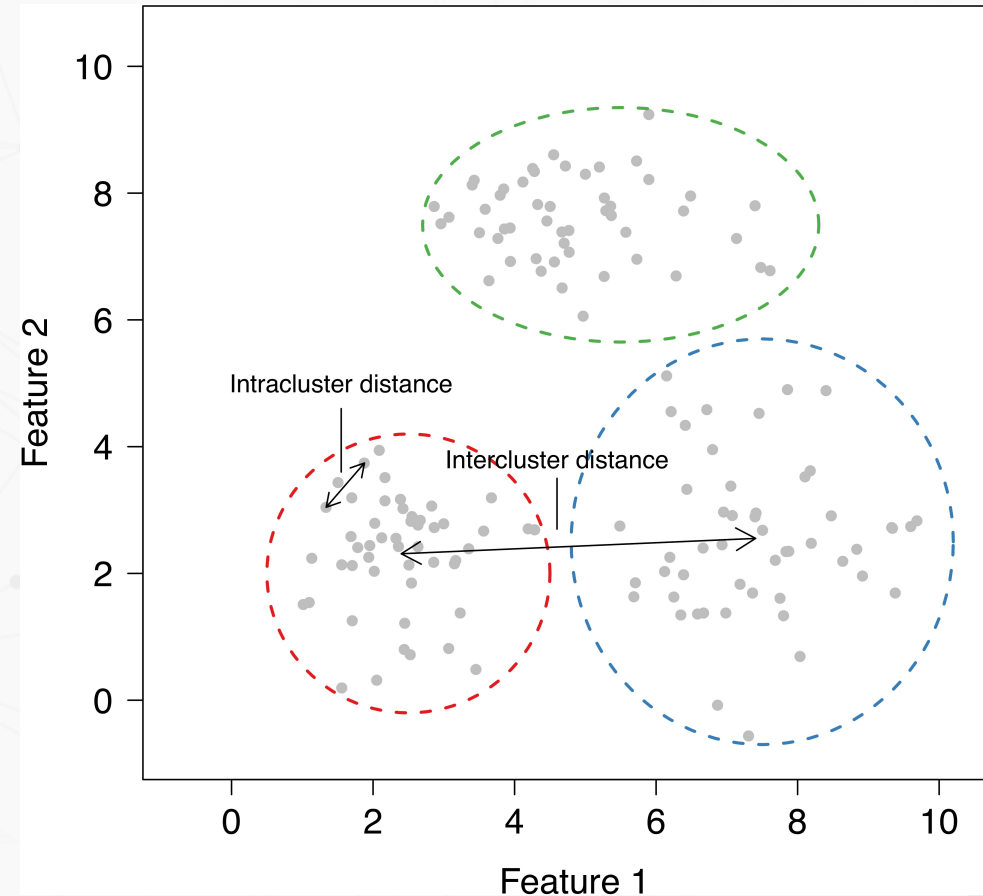
2. Clustering

Clustering

Es la tarea de encontrar agrupamientos (clusters) o grupos homogéneos dentro de un conjunto de datos

Se busca optimizar dos objetivos a la vez:

- Que los datos dentro de un mismo cluster sean muy **similares** entre sí.
- Que los datos de clusters **distintos** muy diferentes entre sí



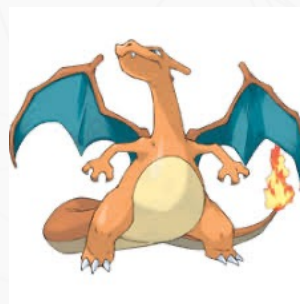
Ejemplos de aplicaciones de clustering

- Dado un conjunto de clientes, encontrar segmentos de mercado para aplicar estrategias de comunicación diferenciadas
- Dado un conjunto de noticias, identificar tópicos de información y agruparlas por su contenido
- Dentro de un conjunto de correos, identificar aquellos relacionados a spam o correos no deseados

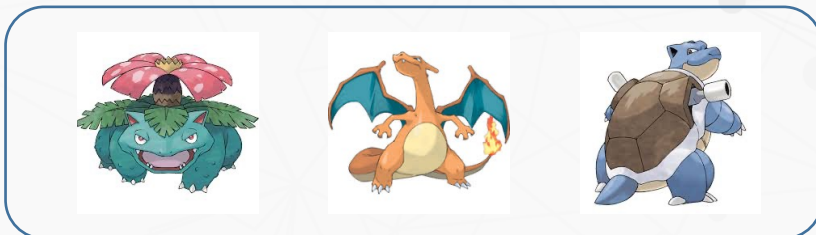
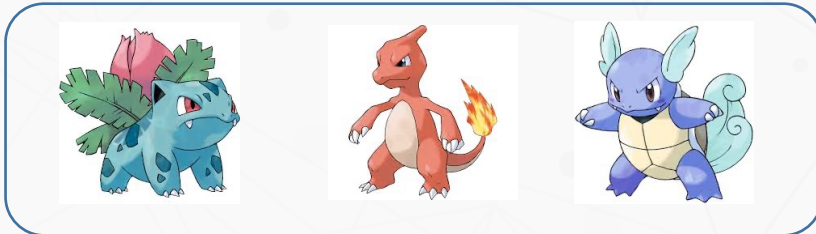


Segmentación de clientes

¿Cómo agrupar los datos?



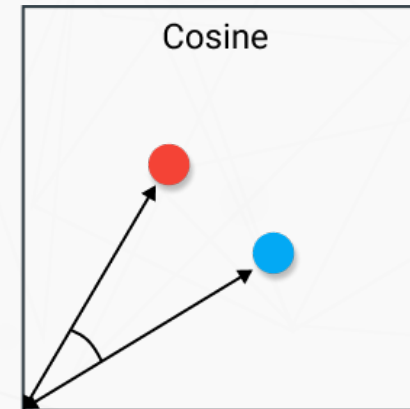
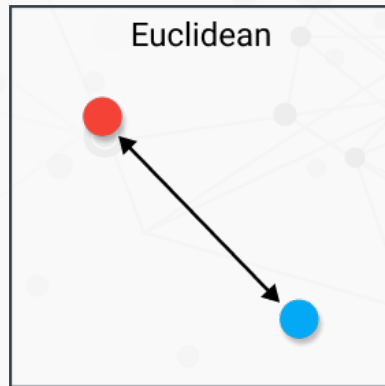
La agrupación es subjetiva



Etapa de evolución

Tipo de pokémon

Definición de distancia



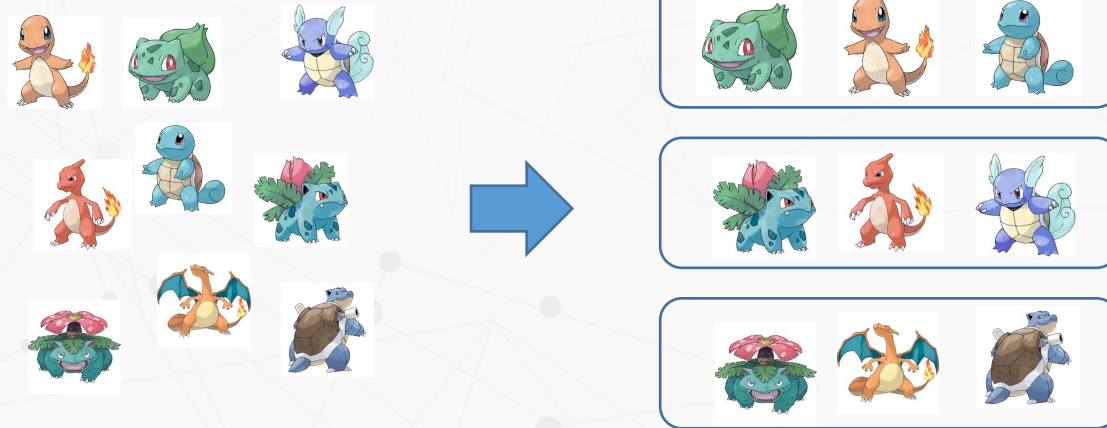
$$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

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

2.1 KMeans

KMeans

- Es un algoritmo de aprendizaje no supervisado
- Es una técnica de clusterización basada en centroides
- Cada elemento se asigna a uno de K Clusters
- Se debe indicar previamente el número de clusters deseados

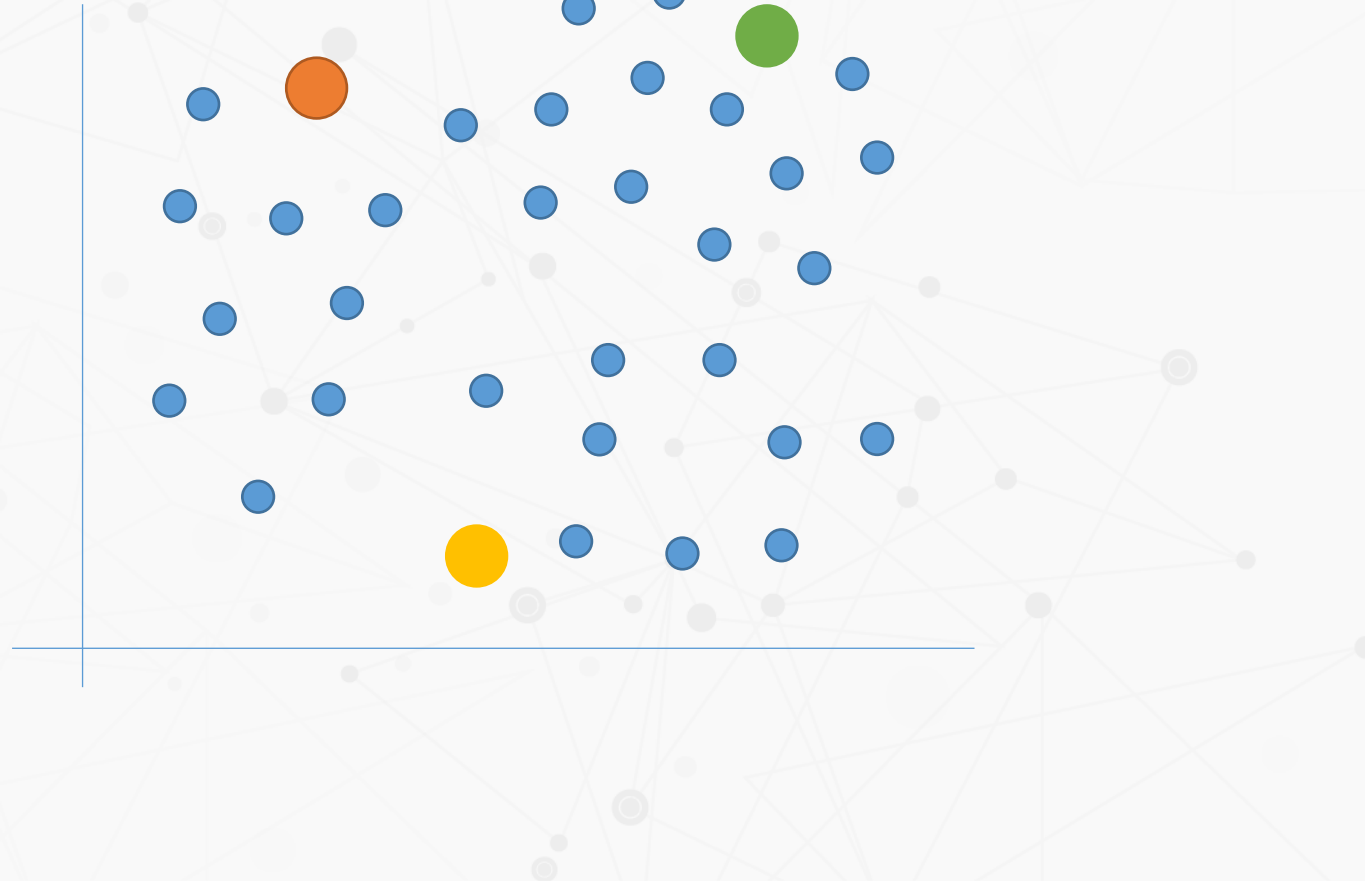


Kmeans: Algoritmo



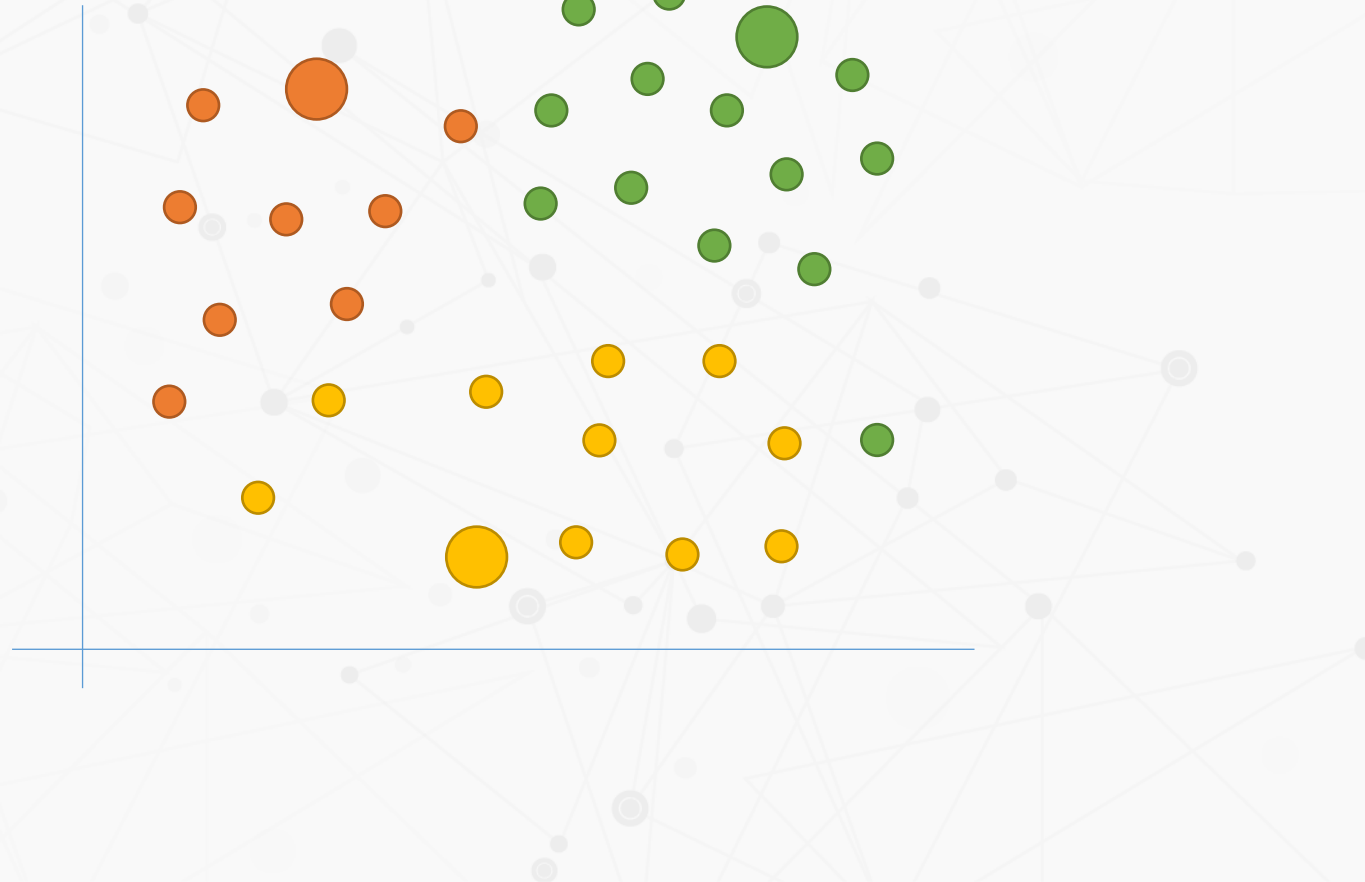
Kmeans: Algoritmo

- Seleccionar k puntos como centroides iniciales



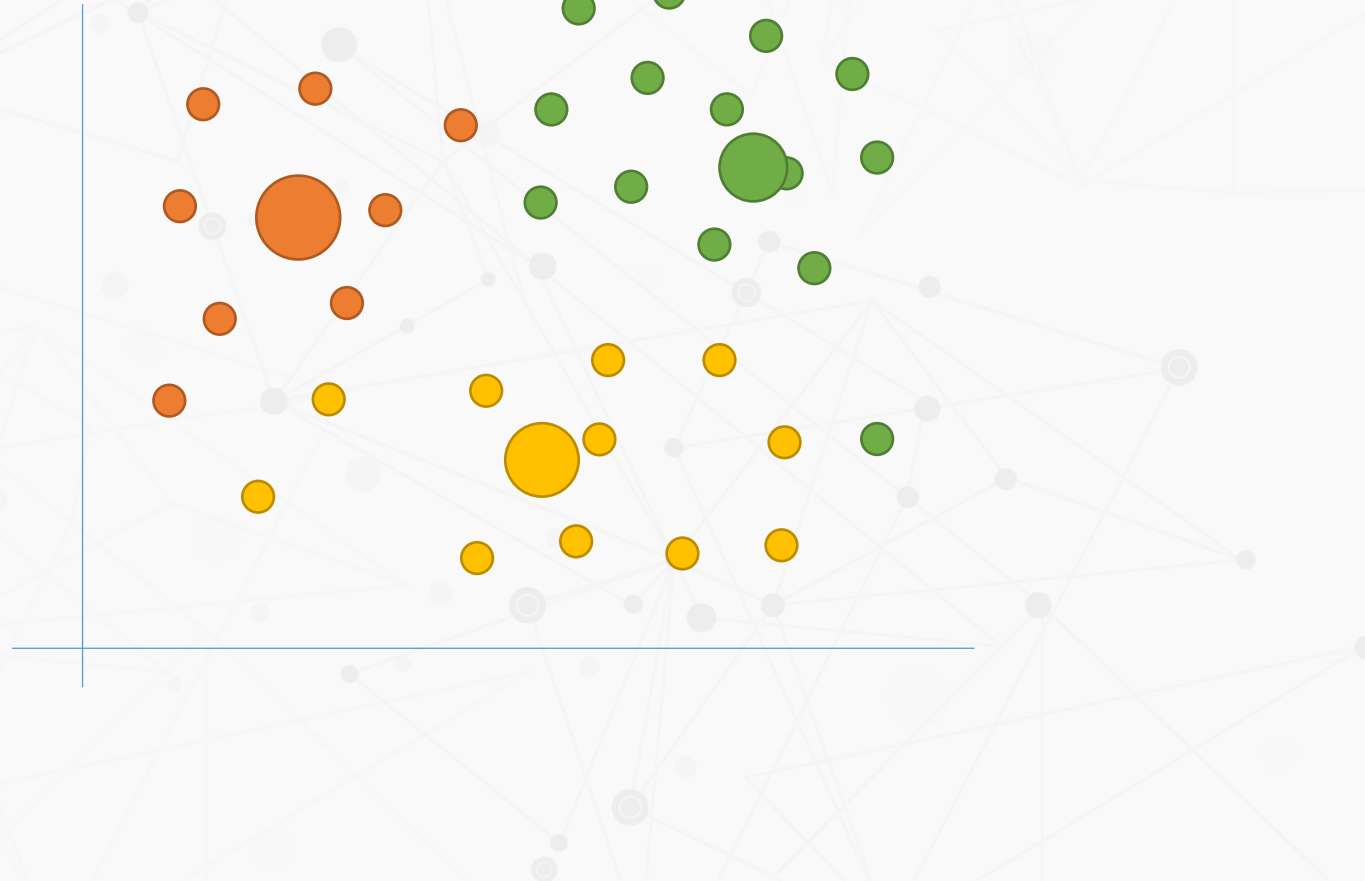
Kmeans: Algoritmo

- Asignar los puntos a cada cluster



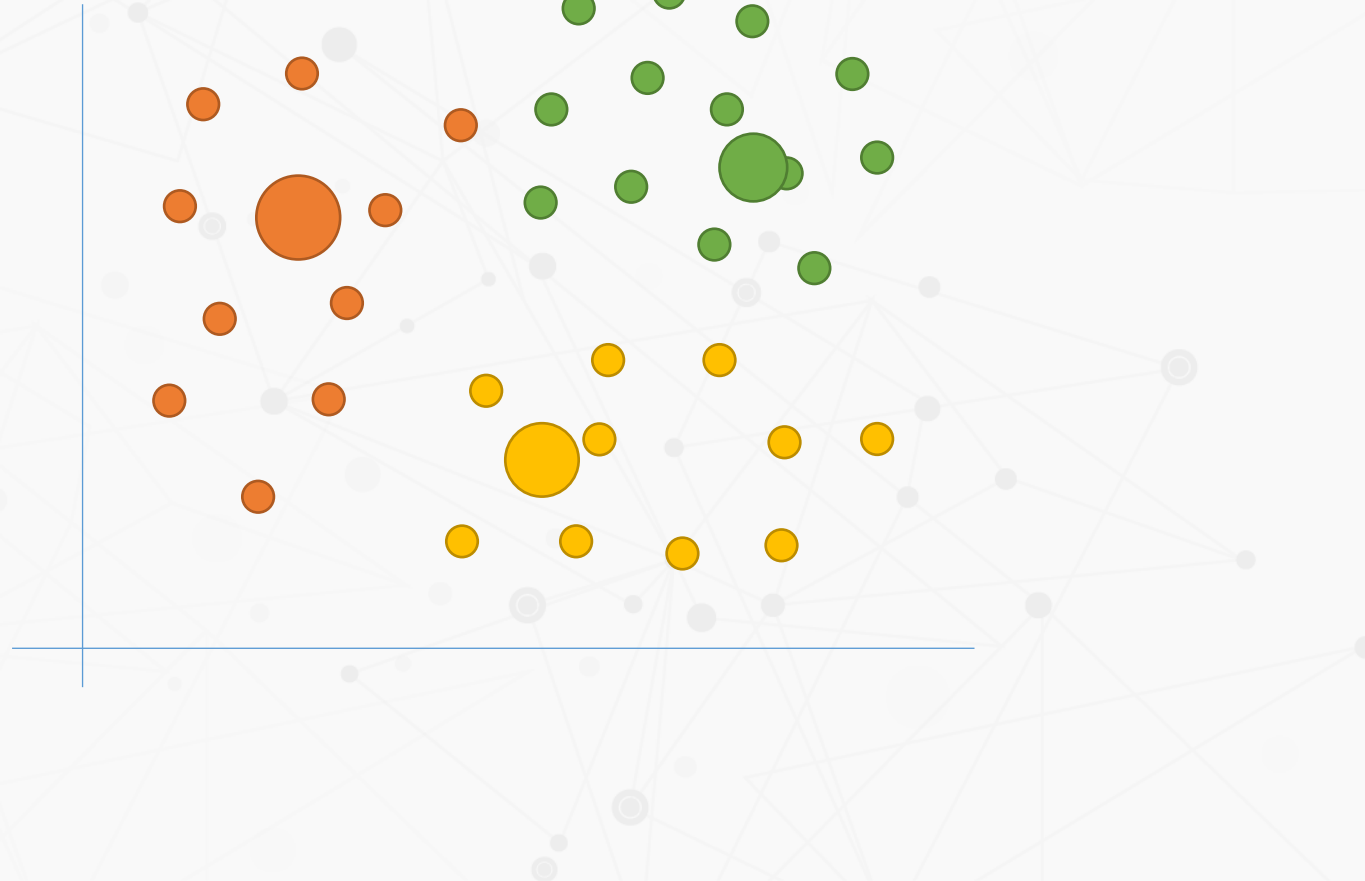
Kmeans: Algoritmo

- Recalcular los centros de cada cluster



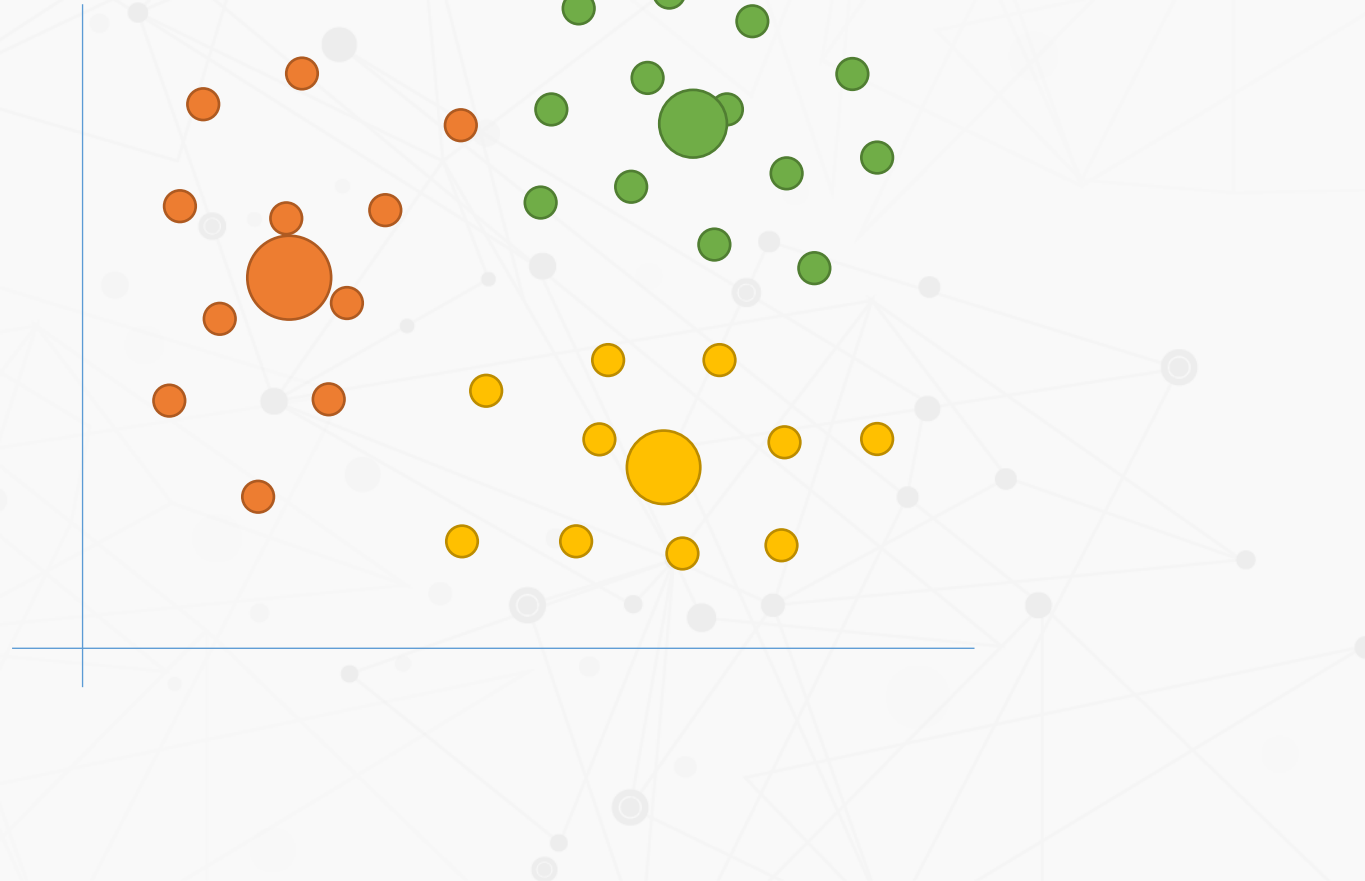
Kmeans: Algoritmo

- Reasignar los elementos a cada cluster hasta que el centroide no cambie



Kmeans: Algoritmo

- Reasignar los elementos a cada cluster hasta que el centroide no cambie



Kmeans

Ventajas

- Simple, entendible
- Los elementos son asignados automáticamente a los clústers

Desventajas

- No se sabe a priori el número de clusters
- Todos los elementos se deben asignar a un cluster
- Los resultados pueden variar de acuerdo a la asignación inicial de los centroides
- Es muy sensible a valores extremos

Kmeans

Visualización de Kmeans: <https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>

Variaciones:

- Kmedoids
- Kmedians
- Kmeans ++

2.2 Clustering Jerárquico

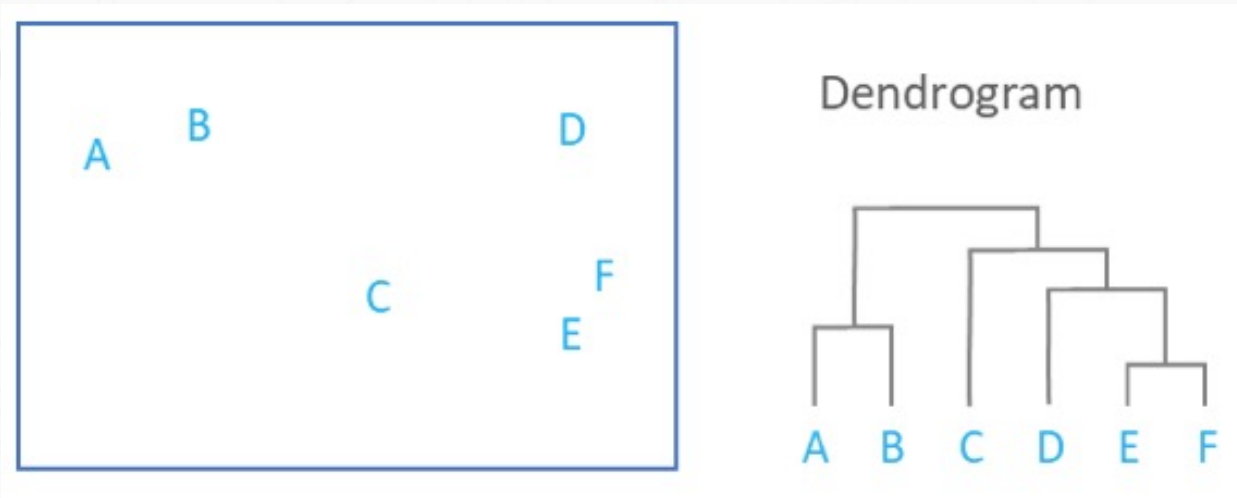
Clustering Jerárquico

Bottom-Up (aglomerativo):

Empezar con cada elemento en su propio cluster, encontrar el mejor par y unirlo en un nuevo cluster. Repetir hasta que todos los clusters estén unidos

Top-Down (divisorio):

Empezar con todos los datos en un único cluster, considerar cada forma posible para dividir el cluster en dos. Repetir el proceso para cada cluster obtenido hasta que todos los elementos estén separados













Matriz de distancia

Se inicia con una matriz de distancia que contiene las distancias entre cada par de objetos en la base de datos

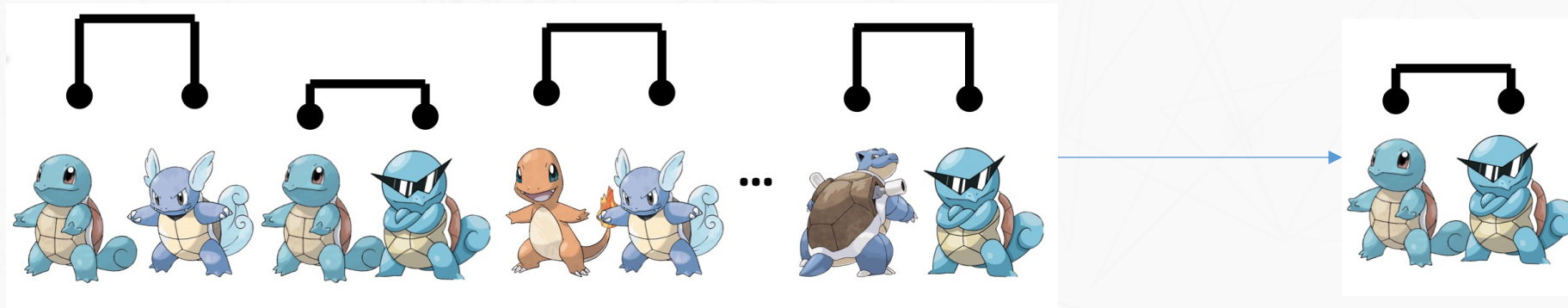
$$D(\text{Charmander}, \text{Blastoise}) = 8$$

$$D(\text{Squirtle}, \text{Blastoise}) = 1$$

					
	0	8	8	7	7
		0	2	4	4
			0	3	3
				0	1
					0

Bottom-Up (aglomerativo):

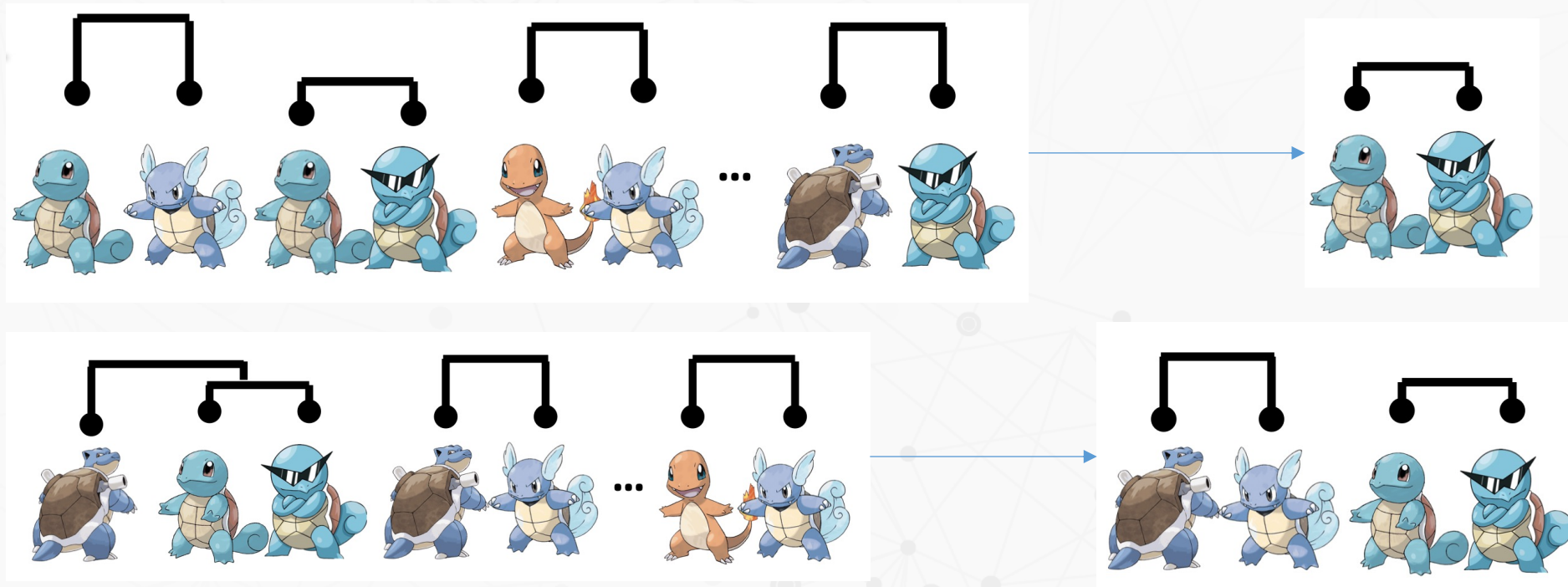
Empezar con cada elemento en su propio cluster, encontrar el mejor par y unirlo en un nuevo cluster. Repetir hasta que todos los clusters esten unidos



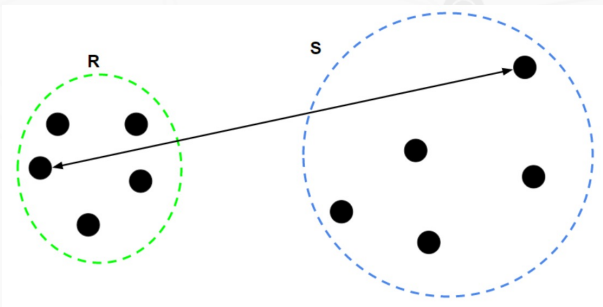
Considerar todas las posibles uniones y elegir la mejor

Bottom-Up (aglomerativo):

Empezar con cada elemento en su propio cluster, encontrar el mejor par y unirlo en un nuevo cluster. Repetir hasta que todos los clusters esten unidos

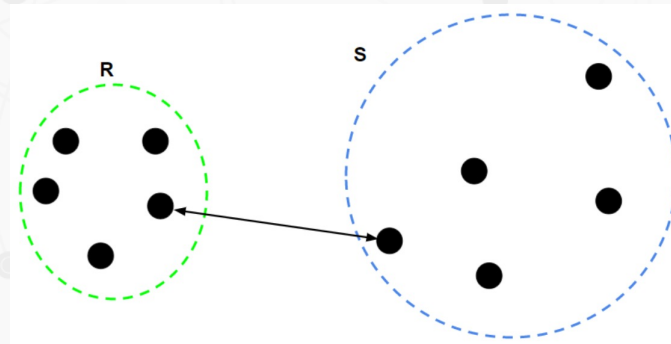


Criterios de similitud



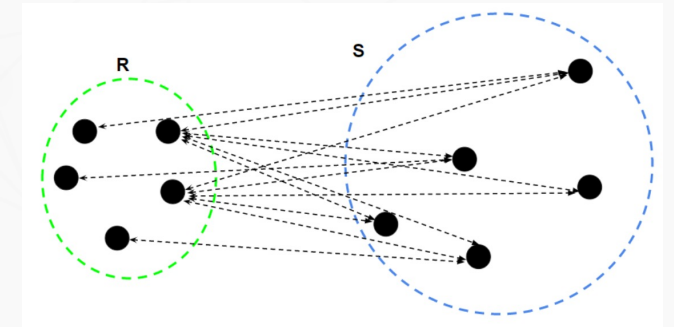
Complete Linkage

Distancia máxima entre dos puntos de distintos clusters



Single Linkage

Distancia mínima entre dos puntos de distintos clusters



Average Linkage

Promedio de las distancias entre los puntos de distintos clusters

Clustering jerárquico aglomerativo

Ventajas

- No requiere un número de *clusters* predefinido
- Permite establecer jerarquías entre *clusters* y elementos
- Un dendrograma es fácilmente interpretable

Desventajas

- Carece de una función objetivo global
- Es costoso en tiempo y espacio de almacenamiento
- La decisión de mezclar *clusters* es irreversible: *“Once the damage is done, it can never be repaired”* (Kaufman, 1990).
- Lo anterior representa un problema en conjuntos de datos con mucho ruido y de muchas dimensiones



¡Gracias!