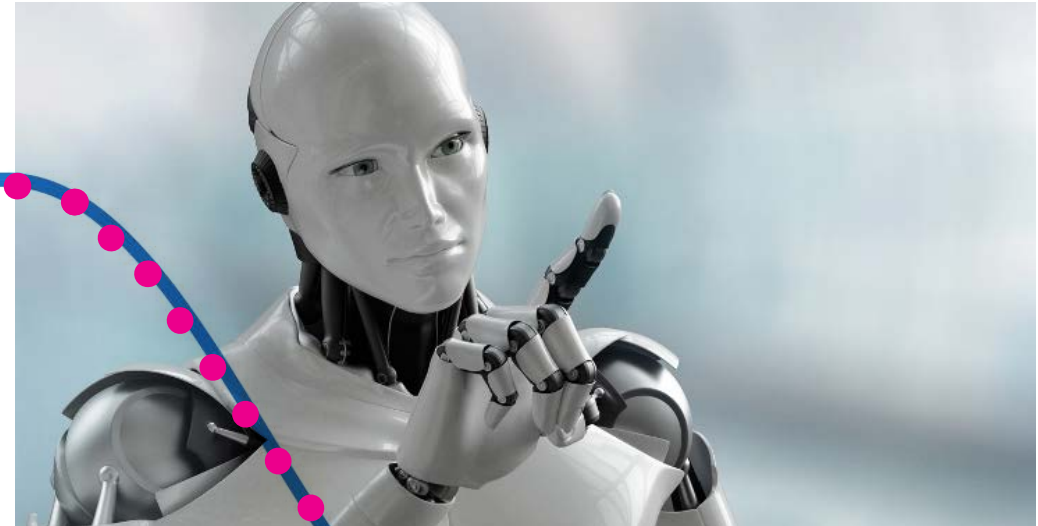
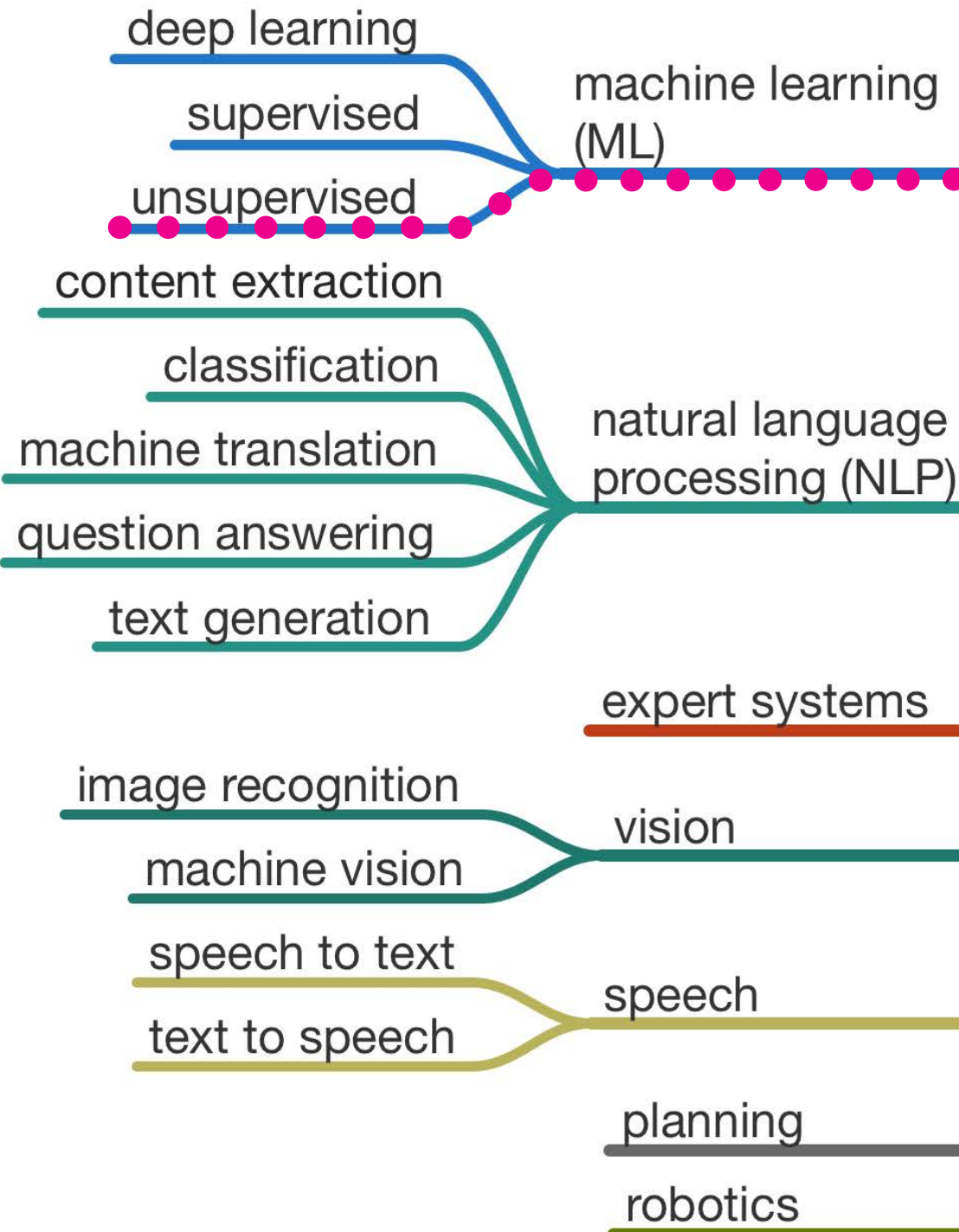




Empezando con Clusterización

Elena Rivas Ruzafa

CLUSTERING??



Artificial Intelligence
(AI)



IA > APRENDIZAJE AUTOMÁTICO > CLÁSICO

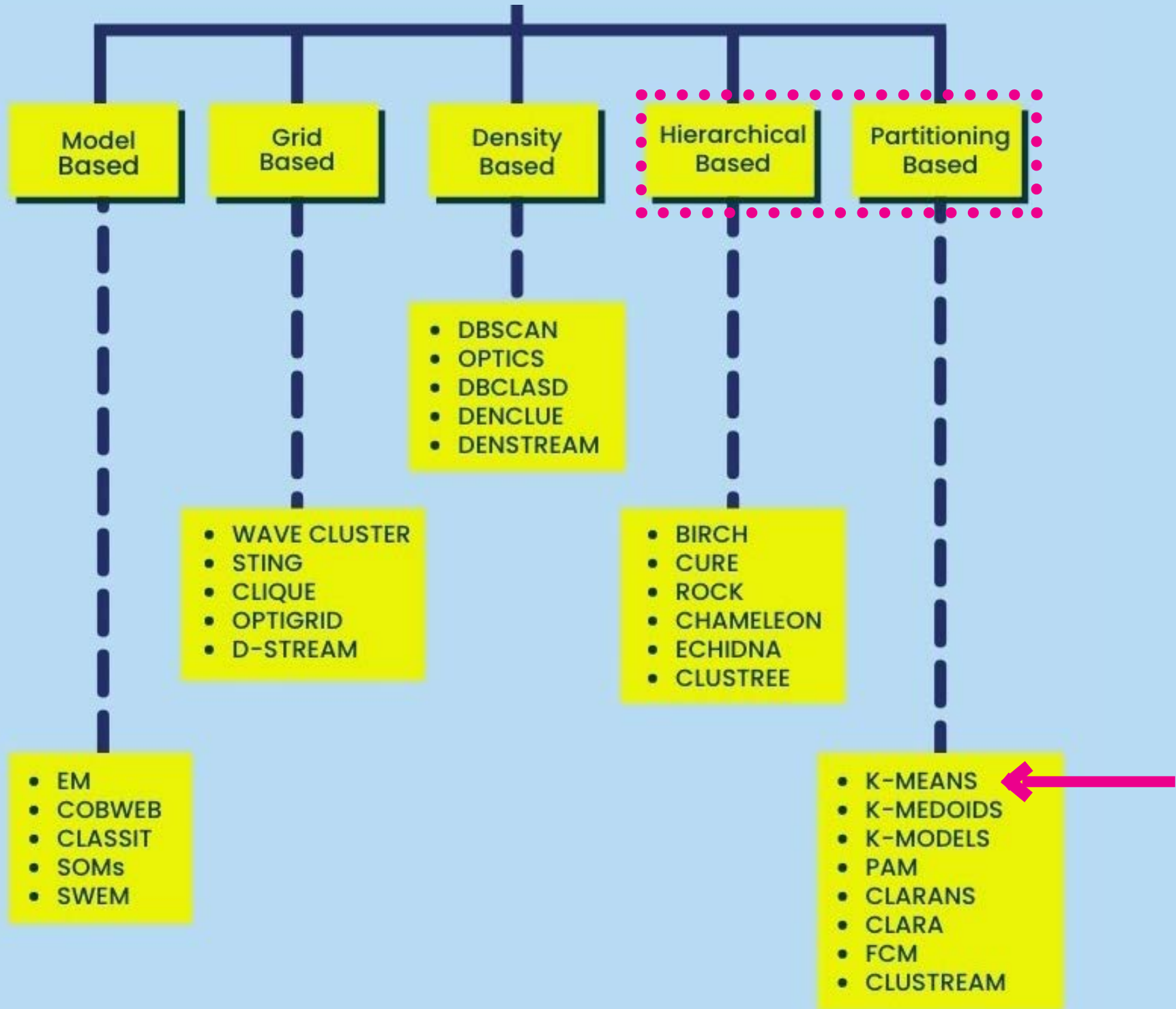
TIPOS PRINCIPALES DE APRENDIZAJE AUTOMÁTICO



C












ALGORITMOS DE CLUSTERING



Qué problema podemos resolver con Clusterización?

observaciones



By color					4
By shape					2
By size					2
etc...					

clusters

IDENTIFICAR FAKE NEWS

OBJETIVO: clasificar un artículo en 'Real' o 'Fake' usando las palabras



MOTORES DE RECOMENDACIÓN

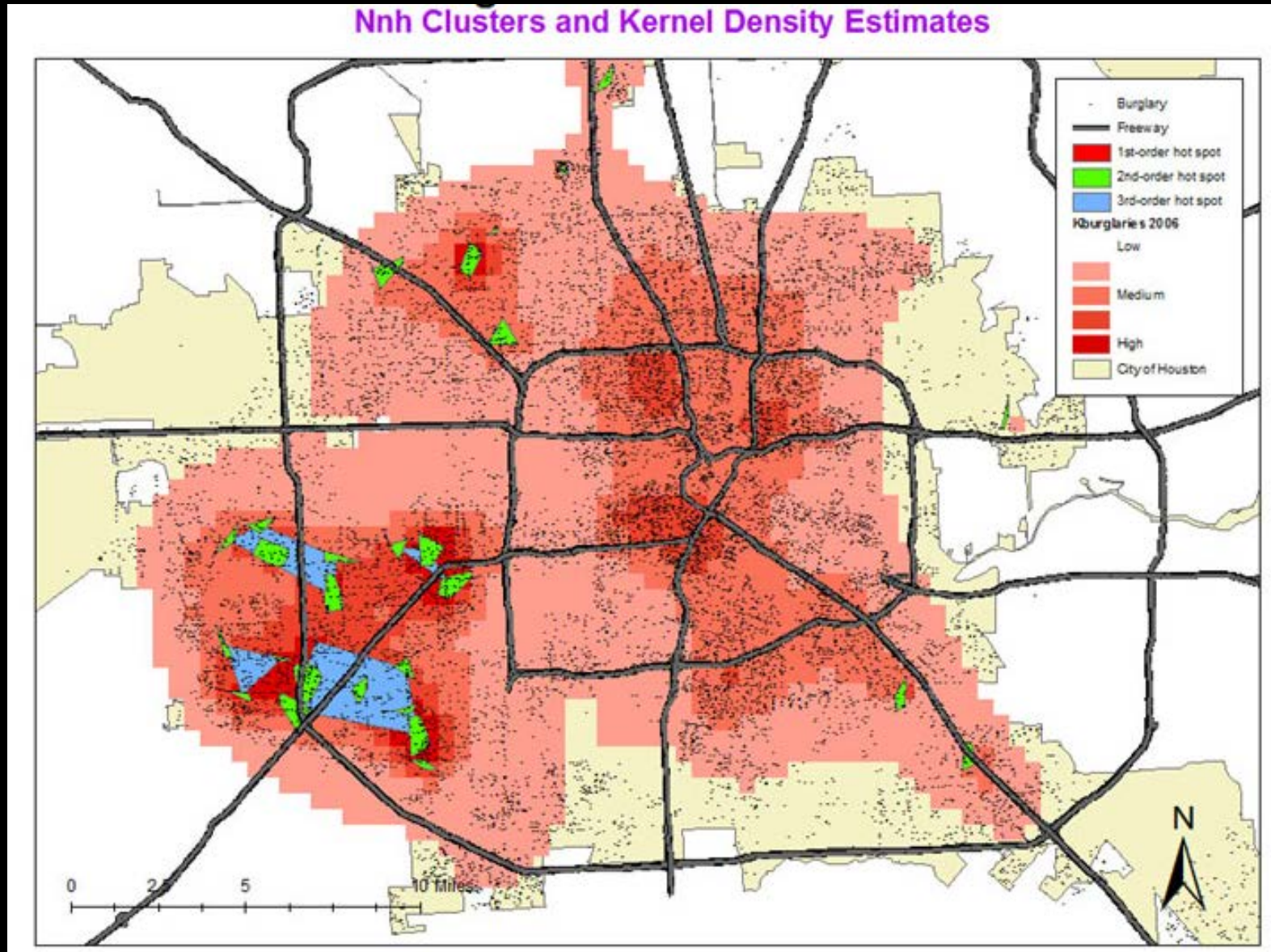
OBJETIVO: Obtener sugerencias de productos, servicios o información basándose en la identificación de usuarios similares

The image is a collage of three screenshots illustrating recommendation engines:

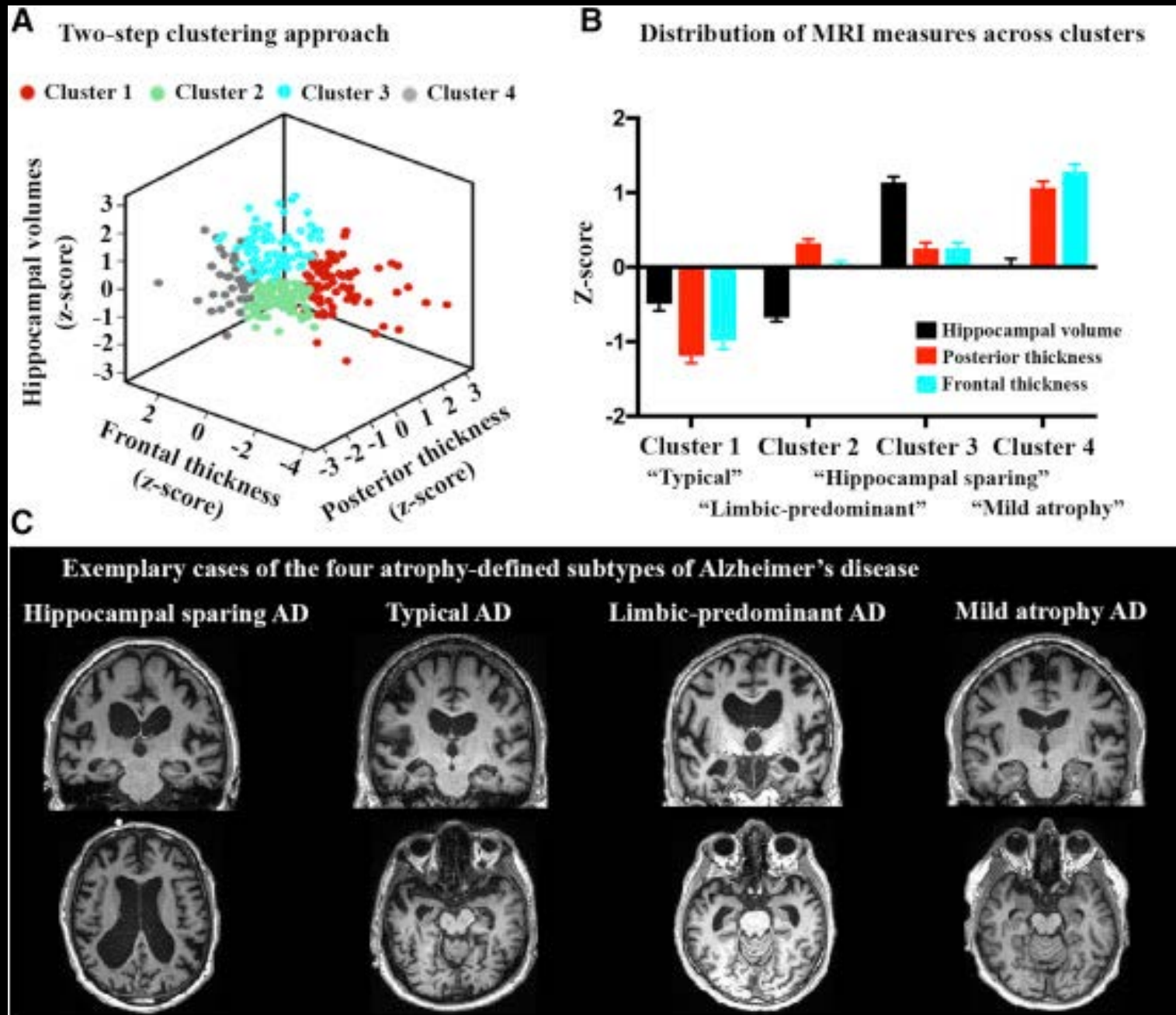
- Top Screenshot (Amazon.com):** Shows the "Recommended for You" section. It states: "Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own." Below this, there are three book covers for "Google Apps Administrator Guide" with a "LOOK INSIDE!" banner.
- Middle Screenshot (Spotify):** Shows the "Discover Weekly" playlist interface for user Moorissa Tjokro. The title is "Discover Weekly" and it is described as "Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your favourites!". It is made for Moorissa Tjokro by Spotify, containing 30 songs, 2 hr 6 min. The interface includes a "PAUSE" button, a "FOLLOWING" button, and a "FIND FRIENDS" button.
- Bottom Screenshot (YouTube):** Shows a video player for "PSY - GANGNAM STYLE (강남스타일) M/V". The video is at 1:00 / 4:13. To the right of the video is a "YouTube Mix" section with a list of related videos:
 - CALL ME 49 videos
 - PSY - GENTLEMAN M/V officialpsy 537,073,962 views (FEATURED)
 - PSY - GANGNAM STYLE (강남스타일) M/V Making officialpsy 61,634,457 views
 - PSY - GANGNAM STYLE (강남스타일) Teaser #1 officialpsy 19,360,319 views
 - PSY - GANGNAM STYLE (강남스타일) Teaser #2 officialpsy 4,624,309 views

IDENTIFICACIÓN DE ZONAS CON DIFERENTES NIVELES DE CRIMINALIDAD

OBJETIVO: A partir de datos históricos de diferentes tipos de crímenes, ser capaz de clasificar las áreas según la actividad criminal y el tipo.



MÉTODOS PARA DETERMINAR DISTINTOS TIPOS DE ALZHEIMER



Tipos de Clustering

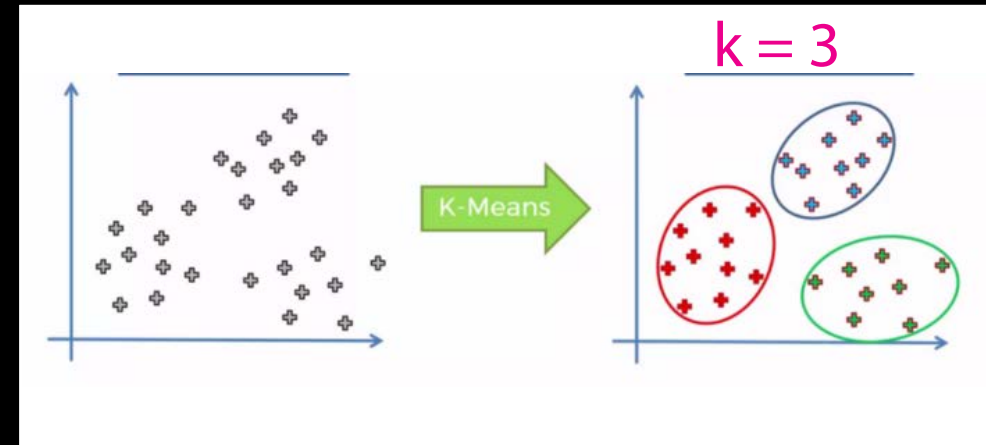
OBJETIVO > Encontrar patrones o grupos (clusters)

MÉTODO > No supervisado

A > Partitioning Clustering

Hay que especificar a priori el número de clusters que se van a crear.

- > K-means
- > K-medoids
- > CLARA

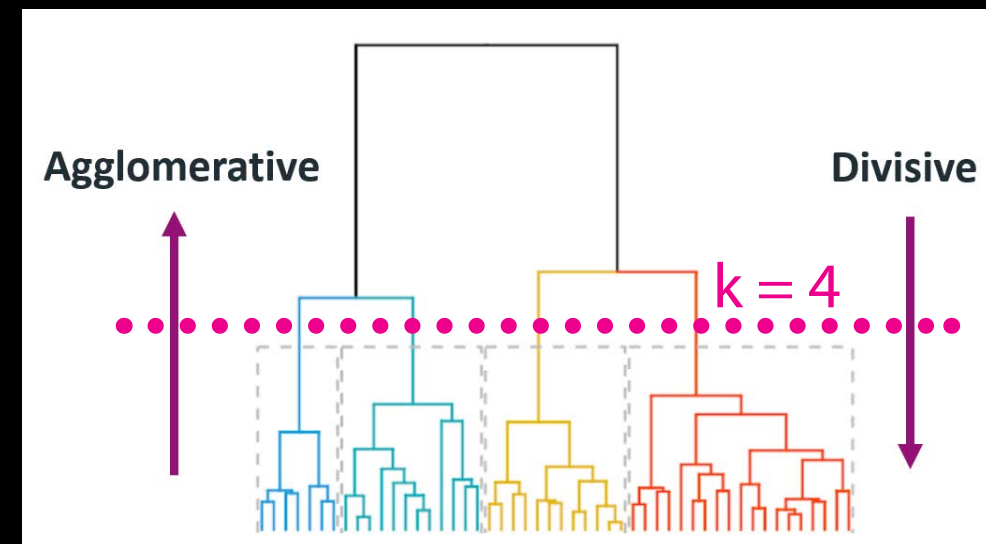


B > Hierarchical Clustering

NO hay que especificar a priori el número de clusters que se van a crear.

Representación en forma de árbol.

- > Agglomerative clustering (bottom-up)
- > Divisive clustering (top-down)



K-MEANS

CLAVES

- > No jerárquico
- > El número de clústeres se ha de establecer al inicio (conocer los datos es muy importante)
- > Particiona el set de datos en K clústeres distintos y no solapantes (ninguna observación puede pertenecer a más de un clúster)

ALGORITMO - iterativo

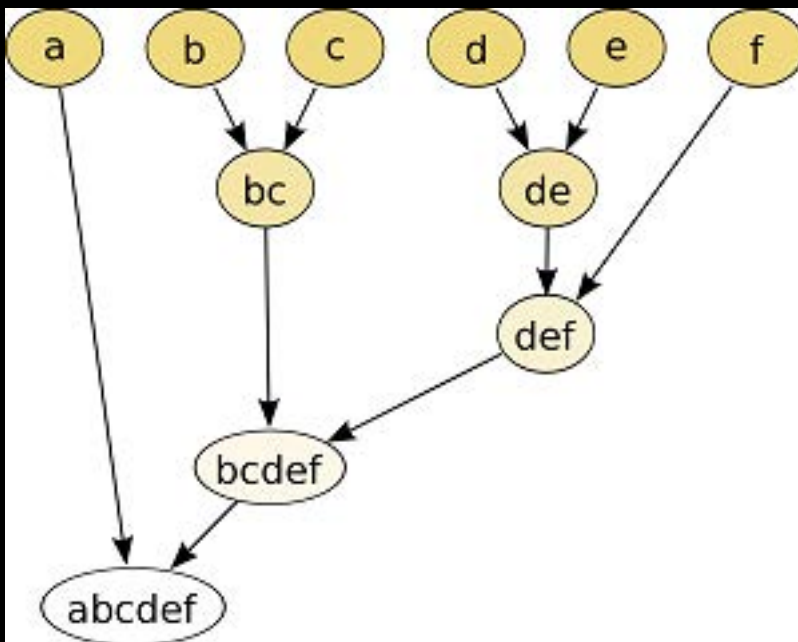
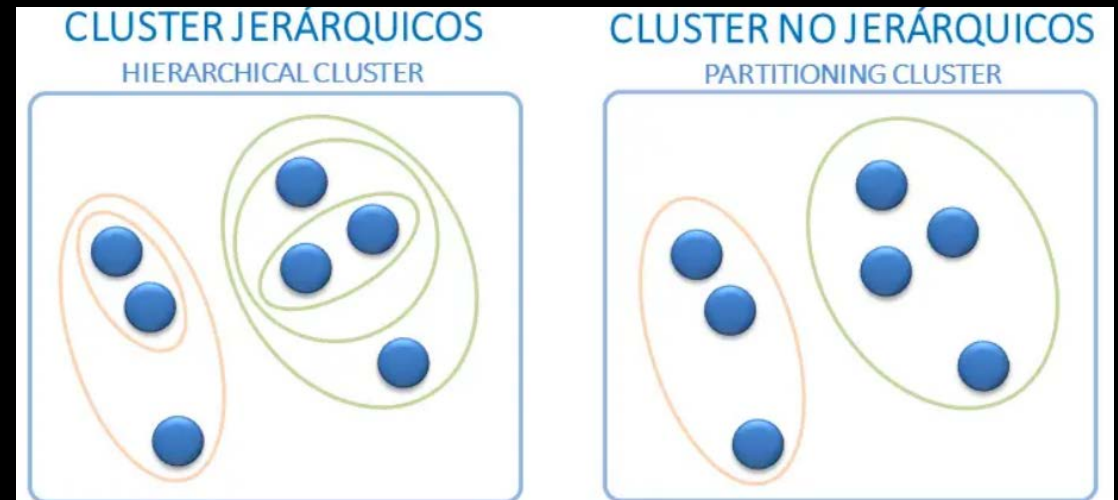
> MINIMIZAR:

Se busca que la varianza total dentro de cada clúster, sumada sobre todos los K clústeres, sea lo más pequeña posible

- > 1. Especificar el número de clusters (k)
- > 2. Colocar de manera aleatoria k elementos como centroides
- > 3. Asignar cada observación al clúster cuyo centroide esté más próximo
- > 4. Los centroides se desplazan a la media de las muestras más cercanas y se repite 3.
- > 3. Iterar hasta que la asignación de cada clúster deje de cambiar, al maximizar la distancia entre los distintos grupos y minimizar la distancia intragrupo

CLUSTERING JERÁRQUICO

> Este método insinúa el número de cluster
> Hay jerarquía de agrupaciones, de manera que se particiona el set de datos en K clústeres distintos y no solapantes en cada nivel, pero muestra diferentes niveles de clústering en los que los clusters inferiores pertenecen a grupos superiores



ALGORITMO > Aglomerativo

- > 1. calculamos la proximidad de individuales y consideramos cada observación como un cluster individual
- > 2. los grupos similares se fusionan y se forman como un solo grupo y repetimos 1
- > 3. Repetir hasta que todos los elementos se fusionen en un sólo cluster

Diferentes cálculos de matriz de distancia y de cálculo de similitud

Medidas de Distancia

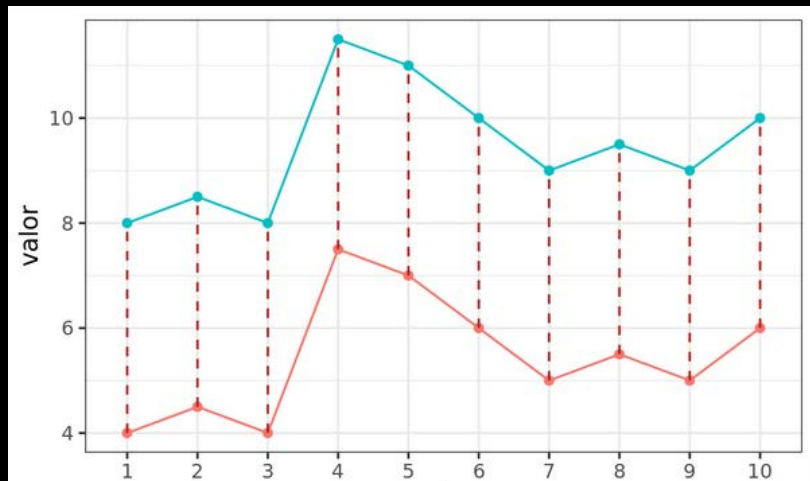
- Agrupar > MEDIR/CUANTIFICAR la SIMILITUD/DIFERENCIA entre las observaciones
 > observaciones que minimicen la distancia
 > En todas las variables (n)
 > MATRIZ DE DISTANCIAS

TIPOS

- > Distancia euclídea
- > Distancia de Manhattan (menos sensible a outliers)
- > Correlación
- > Jackknife correlation (sensible a outliers)
- > Simple matching coefficient (variables binarias)
- > Índice Jaccard
- > Distancia coseno
- ...

No existe una única medida de distancia que sea mejor que las demás, sino que, dependiendo del contexto, una será más adecuada que otra.

Medidas de Distancia: Distancia Euclídea



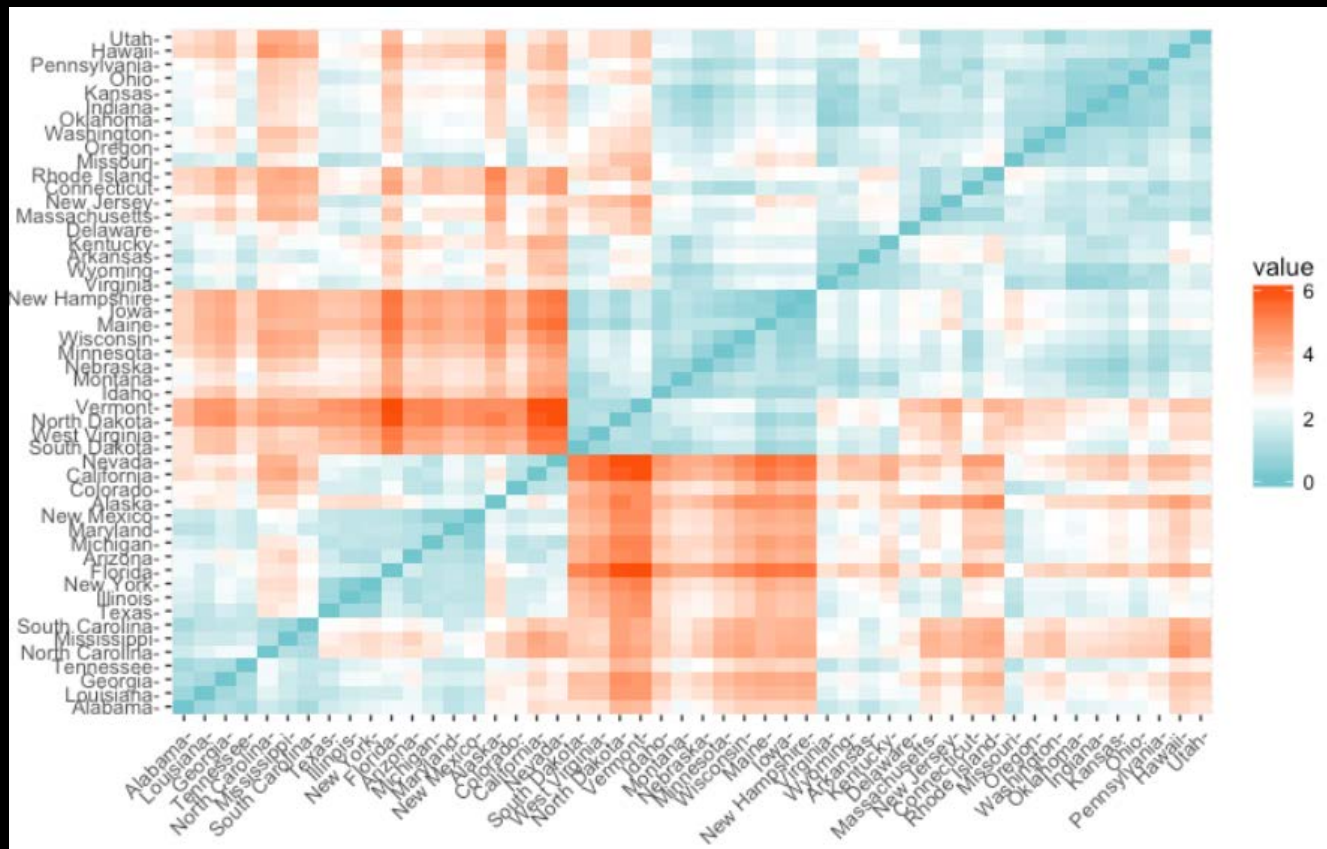
raíz cuadrada de la suma de las longitudes de los segmentos rojos que unen cada par de puntos

Euclidean distance:

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

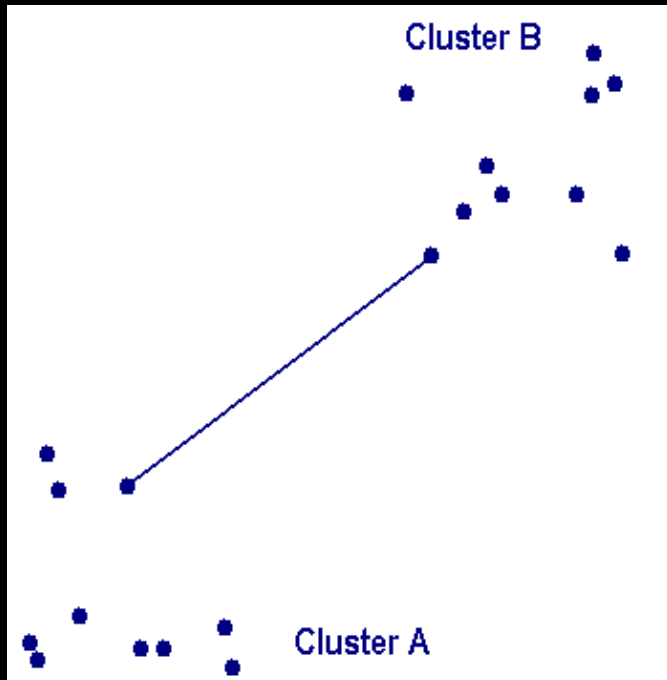
Manhattan distance:

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



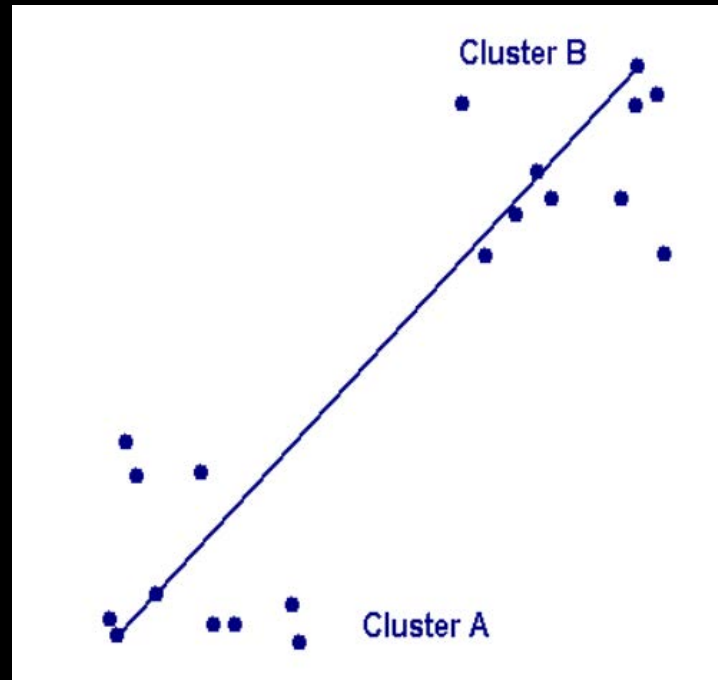
> MATRIZ DE DISTANCIAS

DISTANCIA ENTRE CLUSTERS (Jerárquico)



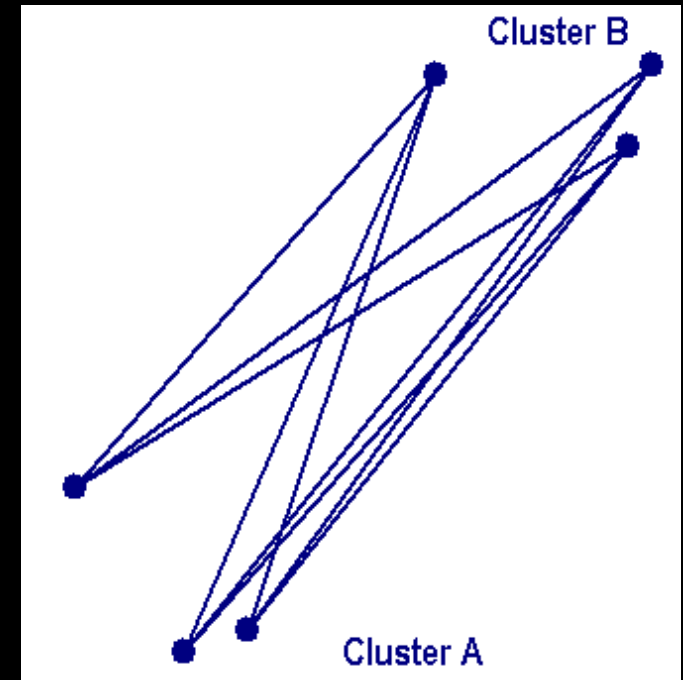
Single linkage clustering

Uno de los más sencillos métodos conocido también como la técnica de los vecinos cercanos



Complete linkage clustering

Llamado también como de los vecinos más alejados, es el opuesto al anterior



Average linkage clustering

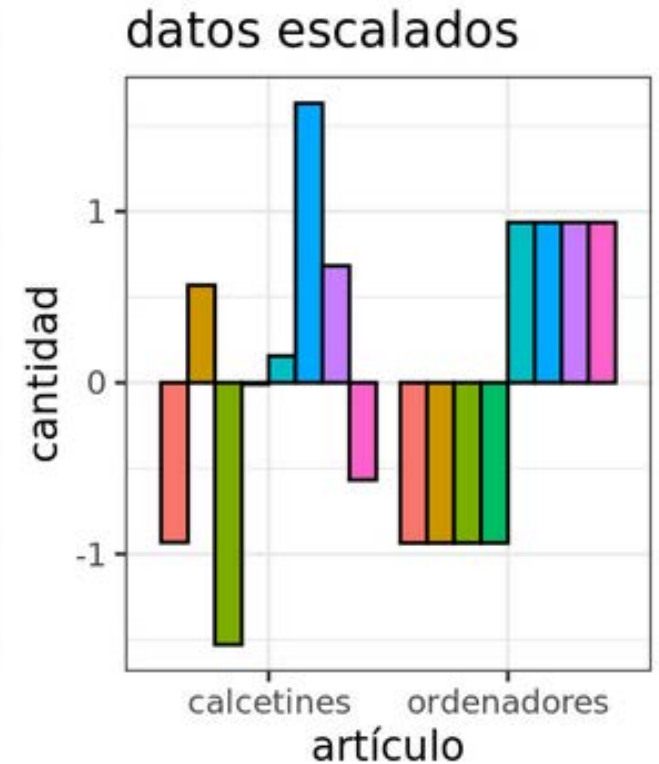
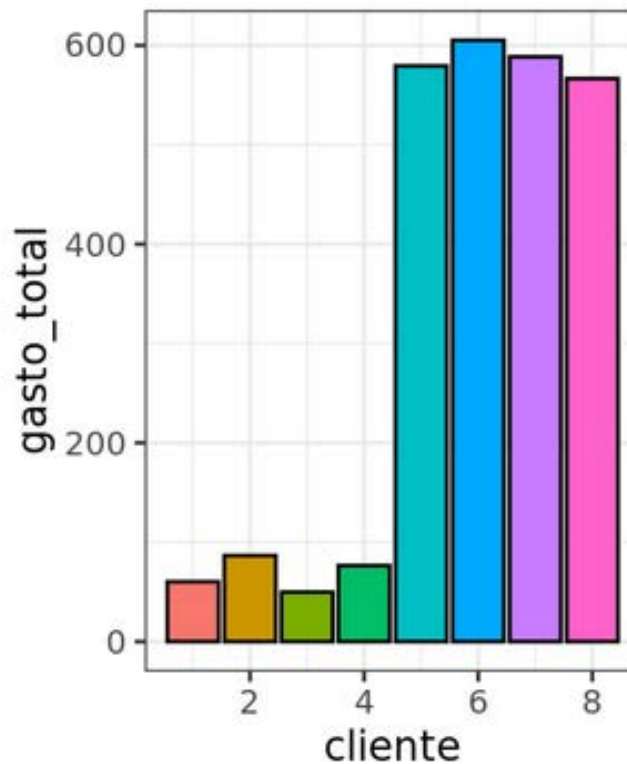
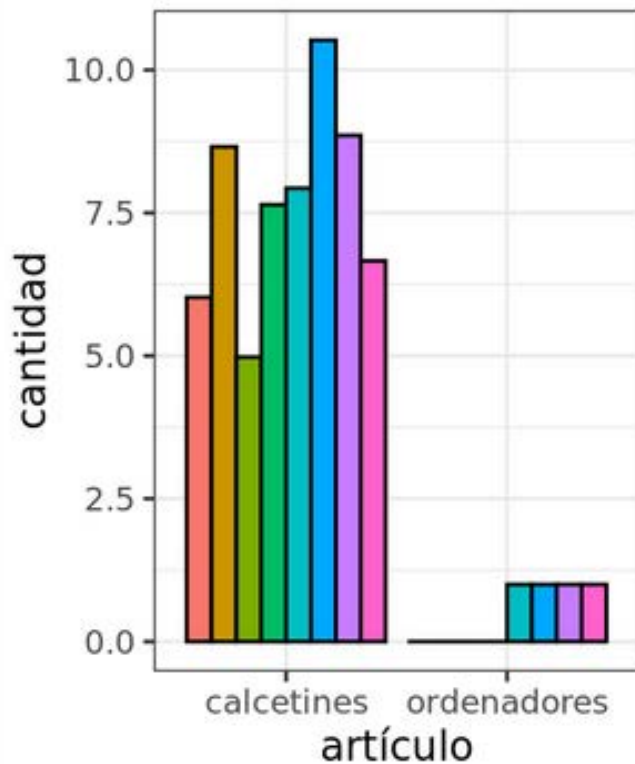
Se calcula como la media de distancias entre todos los pares de objetos

Escala de las Variables

ESCALAR O NO ESCALAR? > Depende del problema a resolver

QUÉ IMPLICA > media 0 y desviación estándar 1 antes de calcular la matriz de distancias

OBJETIVO > las variables tengan el mismo peso/importancia cuando se realice el clustering

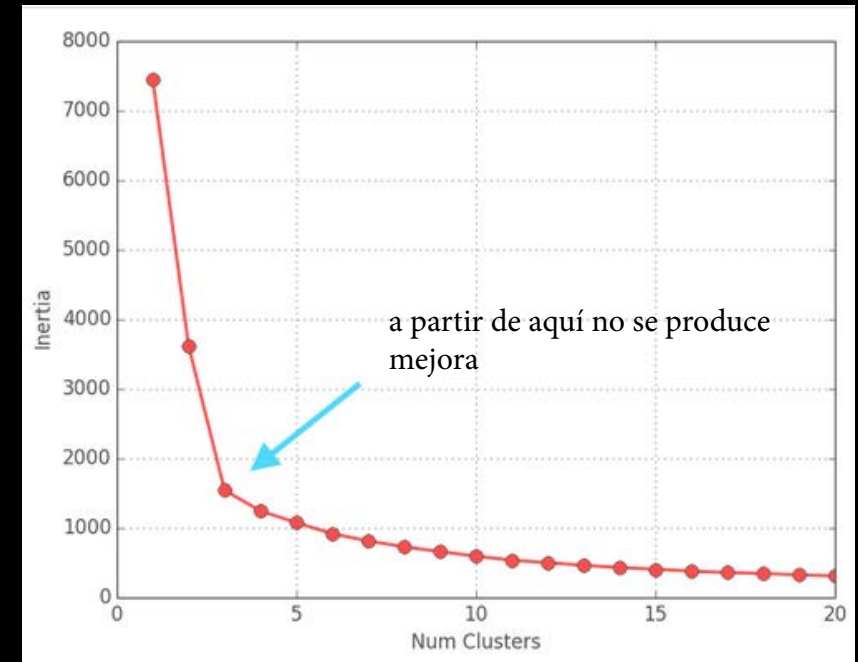


OBTENER NÚMERO OPTIMO DE CLUSTERS

Métodos

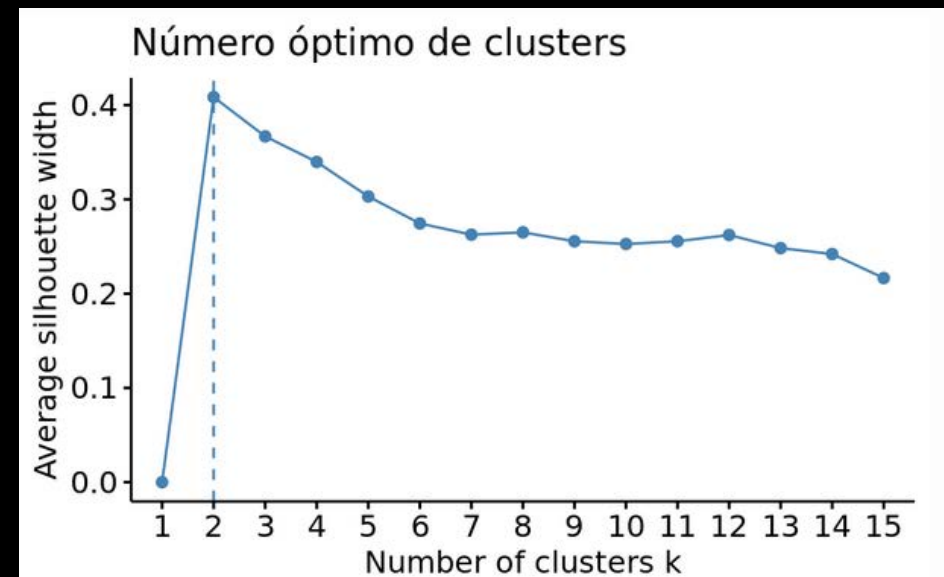
> Elbow method/ método del codo :

se basa en el cálculo de la inercia tras la obtención de los clusters (la inercia la suma de las distancias al cuadrado de cada objeto del Cluster a su centroide). Buscamos el valor a partir del cual deja de producirse mejora aumentar el número de clusters.



> Silhouette method :

en lugar minimizar, se maximiza la media de los silhouette coefficients. Éste mide cómo de buena es la asignación que se ha hecho de una observación comparando su similitud con el resto de observaciones de su cluster frente a las de los otros clusters.



CASO PRÁCTICO:

CLASIFICAR ZONAS POR SU CALIDAD DEL AIRE EN BASE A LAS MEDICIONES TOMADAS EN DIFERENTES ESTACIONES METEOROLÓGICAS DISTRIBUÍDAS POR LA CIUDAD



DATOS:

Portal de datos abiertos del Ayuntamiento de Madrid

 **MADRID**

datos abiertos

¿Qué estás buscando?



Tu ciudad más cerca

Gracias a nuestra plataforma de datos abiertos podrás encontrar todos los datos de Madrid que necesitas para tu proyecto

En portada

Acerca de Datos Abiertos

Catálogo de datos

Colabora

Lo más visto 

Contenedores de ropa autorizados ... / Zonas del Servicio de Estacionami... / Listas de espera de aparcamientos...



Catálogo de datos > Conjuntos de datos

  a+ a- 

Calidad del aire. Datos horarios años 2001 a 2021

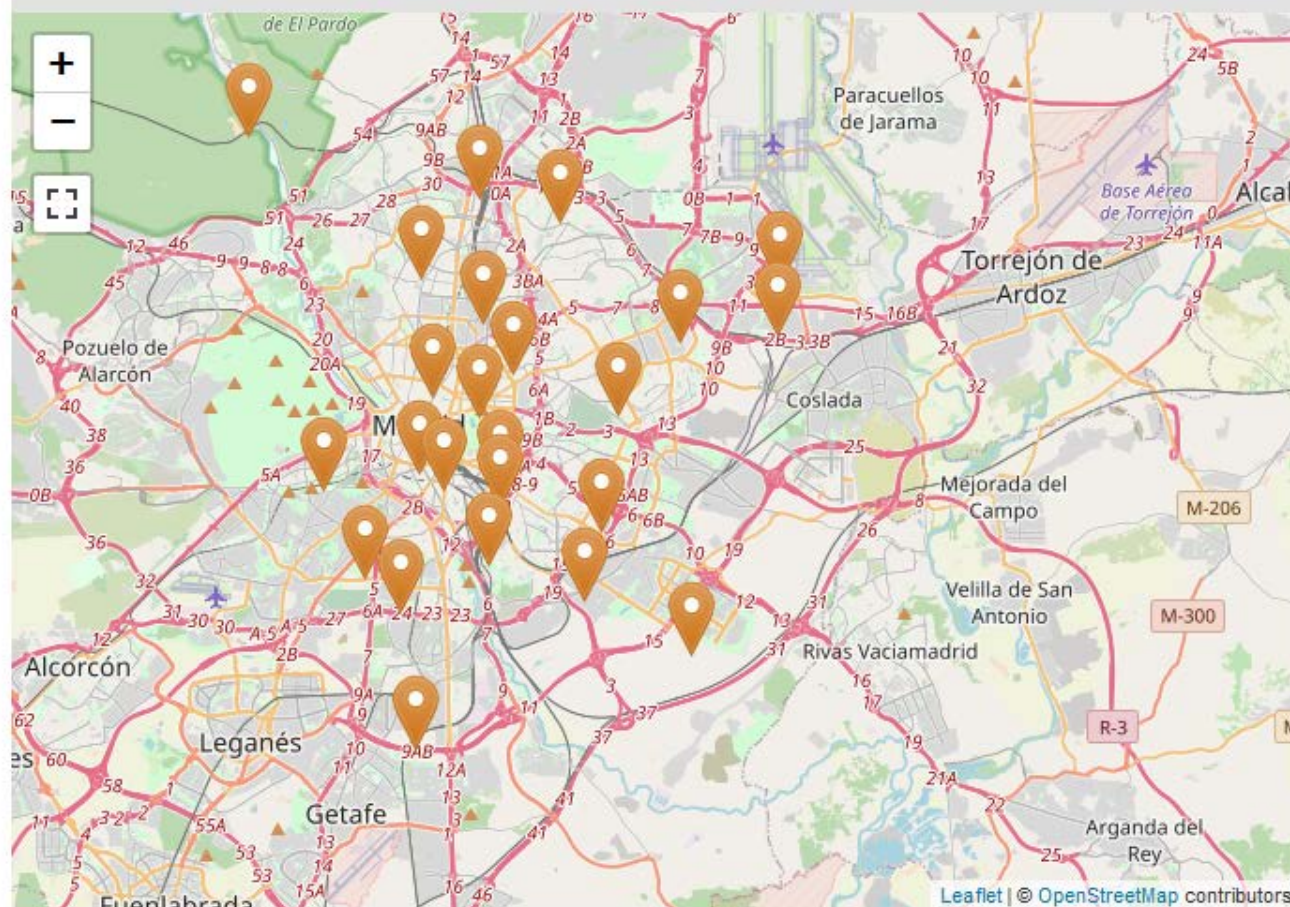
← Volver

Conjuntos de datos

API

El Sistema Integral de la Calidad del Aire del Ayuntamiento de Madrid permite conocer en cada momento los niveles de contaminación atmosférica en el municipio. En este conjunto de datos puede obtener la información recogida por las estaciones de control de calidad del aire, con los datos horarios por anualidades de 2001 a

Ver resultados en el mapa



Descargas

Calidad del aire: estaciones de control



Descargar fichero

XLS, 35 Kbytes - 68.266 descargas

Magnitud		Abreviatura o fórmula	Unidad medida	Técnica de medida	
01	Dióxido de Azufre	SO ₂	µg/m ³	38	Fluorescencia ultravioleta
06	Monóxido de Carbono	CO	mg/m ³	48	Absorción infrarroja
07	Monóxido de Nitrógeno	NO	µg/m ³	08	Quimioluminiscencia
08	Dióxido de Nitrógeno	NO ₂	µg/m ³	08	Id.
09	Partículas < 2.5 µm	PM2.5	µg/m ³	47	Microbalanza
10	Partículas < 10 µm	PM10	µg/m ³	47	Id.
12	Óxidos de Nitrógeno	NOx	µg/m ³	08	Quimioluminiscencia
14	Ozono	O ₃	µg/m ³	06	Absorción ultravioleta
20	Tolueno	TOL	µg/m ³	59	Cromatografía de gases
30	Benceno	BEN	µg/m ³	59	Id.
35	Etilbenceno	EBE	µg/m ³	59	Id.
37	Metaxileno	MXY	µg/m ³	59	Id.
38	Paraxileno	PXY	µg/m ³	59	Id.
39	Ortoxileno	OXY	µg/m ³	59	Id.



Dióxido de nitrógeno



Este artículo o sección necesita [referencias](#) que aparezcan en una [publicación acreditada](#).

Este aviso fue puesto el 22 de diciembre de 2015.

Para otros usos de este término, véase [Óxido de nitrógeno \(IV\)](#) (desambiguación).

El **dióxido de nitrógeno** u **óxido de nitrógeno (IV)**² (NO₂), es un [compuesto químico](#) formado por los [elementos nitrógeno](#) y [oxígeno](#), uno de los principales contaminantes entre los varios [óxidos de nitrógeno](#).

El dióxido de nitrógeno es de color [marrón-amarillento](#). Se forma como subproducto en los procesos de [combustión](#) a altas temperaturas, como en los [vehículos motorizados](#) y las [plantas eléctricas](#). Por ello es un contaminante frecuente en zonas urbanas.

MUCHAS GRACIAS!!

erivasruzafa@gmail.com