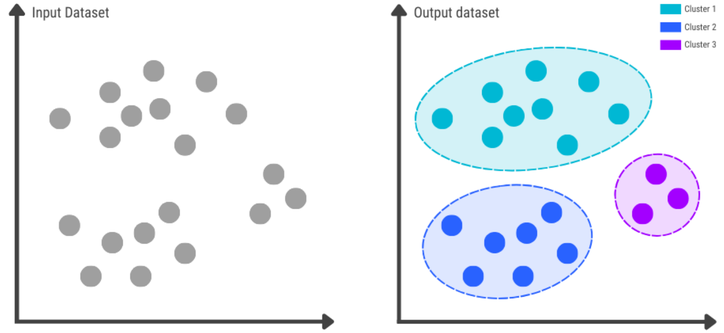
# Clustering

## Licenciatura en Inteligencia Artificial y Ciencia de Datos, CUGDL, Universidad de Guadalajara.

Guadalajara, Jal., agosto de 2025

# Introducción




- Un **clúster** es un conjunto de datos que presentan **alta similitud interna** y **baja similitud externa.** En otras palabras, los elementos dentro de un mismo grupo comparten características parecidas, mientras que los elementos de grupos distintos difieren de manera significativa.

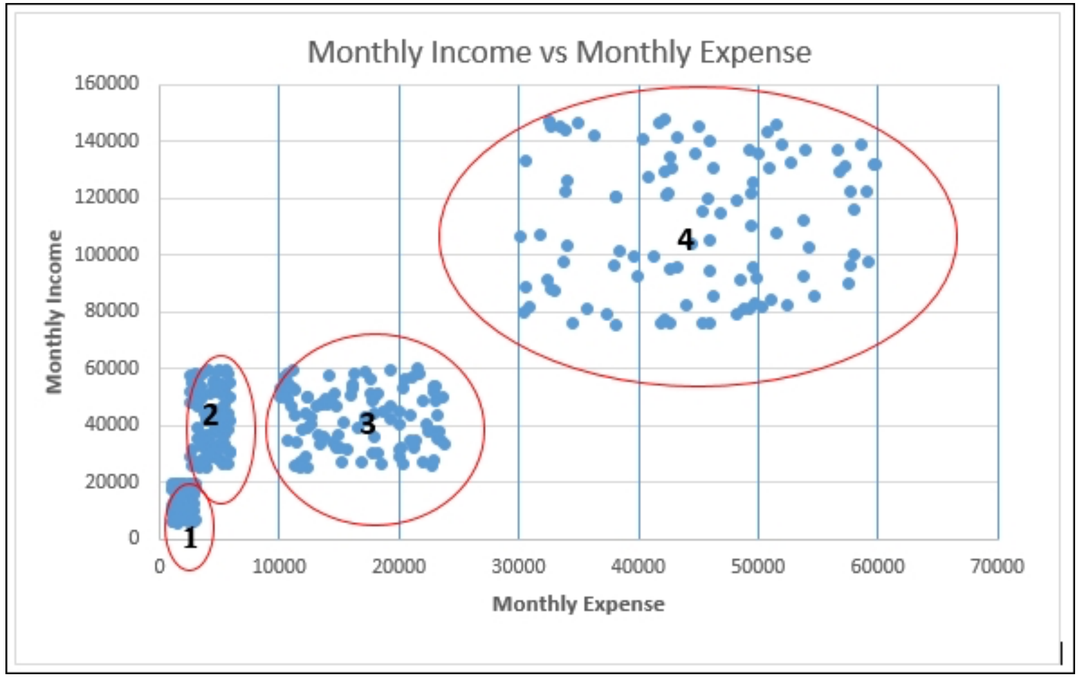
- El objetivo del **clustering** es identificar estas agrupaciones naturales dentro de los datos, sin utilizar etiquetas o categorías predefinidas, i.e., se trata de un método no supervisado. Este tipo de análisis permite describir patrones, estructuras ocultas o relaciones que no son evidentes a simple vista, siendo una de las técnicas fundamentales del aprendizaje automático no supervisado.

# Clustering

## Ejemplo

En el gráfico de la derecha tenemos los ingresos y gastos mensuales de diferentes individuos, donde podemos identificar 4 grupos,

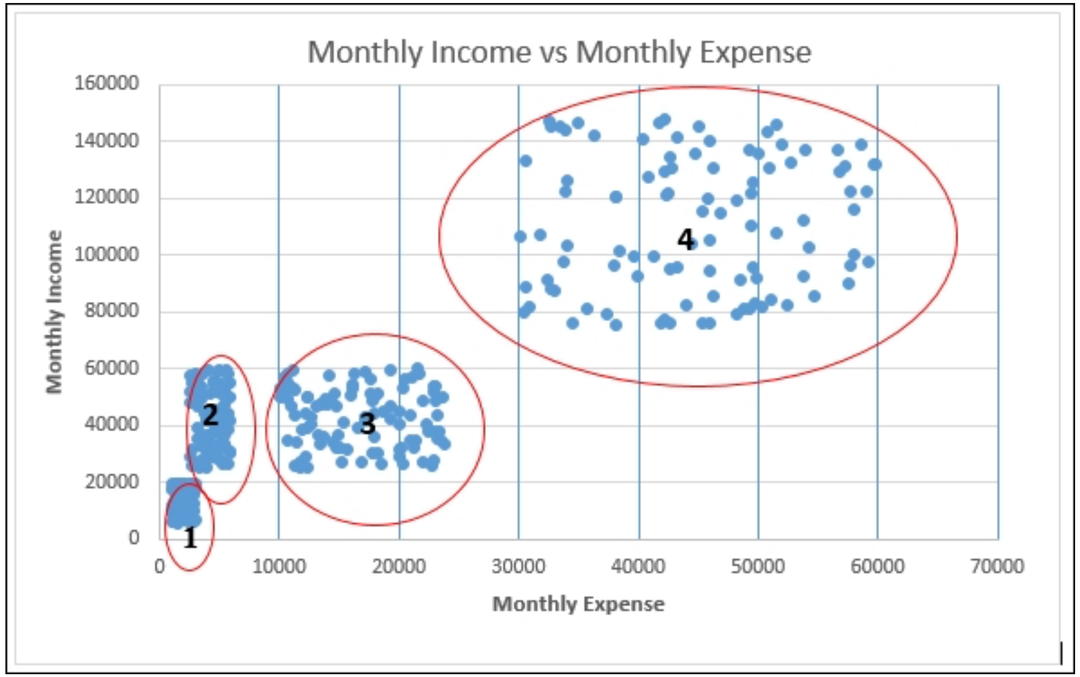
- Ganan y gastan poco
- Ingresos medios pero gastan poco
- Ingresos medios y gastos mayores
- Ganan y gastan mucho



# Clustering

## Ejemplo

- Este es sólo uno de los ejemplos donde el clustering puede ser ventajoso.
- En este caso es sencillo puesto que sólo tenemos dos atributos X y Y para clientes potenciales. Por esto podemos tener una gráfica 2D.
- Pero esto es más complicado para más dimensiones.
- Entonces, se busca generalizar una métrica de semejanza o diferencia entre las observaciones en un espacio de $n$ dimensiones más allá de hacer un gráfico y círculos.



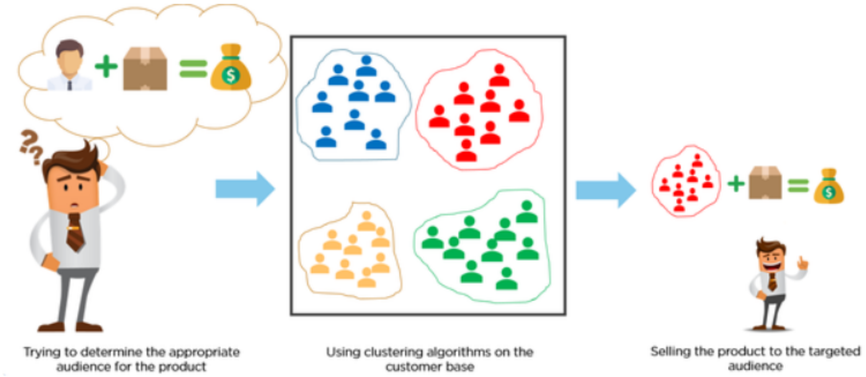
# Clustering

## Propiedades

- Los clústers deben ser **identificables de tamaño significativo**, de modo que representen grupos reales dentro de los datos.
- Los puntos que pertenecen al mismo clúster deben ser compactos, es decir, estar cercanos entre sí, y mostrar **mínima superposición** con los puntos de otros clústers.
- Cada clúster debe tener **coherencia dentro del contexto del análisis.** Los elementos agrupados deben compartir propiedades relevantes y tener sentido desde el punto de vista del problema o dominio estudiado.

# Clustering

## Aplicaciones




- El clustering y la segmentación son herramientas esenciales en muchos campos. Permiten descubrir patrones, dividir grandes conjuntos de datos en grupos significativos y tomar decisiones más informadas. Algunos ejemplos incluyen:
- **Marketing digital:** segmentar clientes según su comportamiento, intereses o demografía (edad, preferencias, clics, etc.) para diseñar campañas publicitarias más efectivas, por ejemplo.
- **Biología:** la ***taxonomía de especies*** es un ejemplo clásico de clustering jerárquico, donde los organismos se agrupan por similitud genética o morfológica.
- **Sismología:** detección de epicentros y zonas de riesgo sísmico mediante técnicas de agrupamiento espacial.
- **Imputación de datos faltantes:** estimar valores desconocidos utilizando las observaciones del clúster más cercano, en lugar de promediar todo el conjunto.
- **Planificación urbana:** agrupar viviendas o zonas según su ubicación geográfica, nivel socioeconómico o características de su infraestructura (fibra óptica, transporte, servicios).

# Clustering

## Matemáticas

- El *clustering* se fundamenta en **medir semejanzas y diferencias** entre las observaciones. Cuánto más parecidos sean dos elementos según ciertas características, mayor será la probabilidad de que pertenezcan al mismo grupo.

- **Ejemplo:** El sistema de recomendación de Netflix u otras plataformas de streaming. El algoritmo analiza las valoraciones, hábitos de visualización o preferencias de cada usuario para estimar qué tan probable es que le guste una película o serie determinada. A mayor cantidad de datos disponibles, más precisa será esta estimación.

# Clustering

## Matemáticas

- Supongamos que cada observación se representa como un punto en un espacio de $m$ dimensiones, es decir, $m$ variables en el dataset. Cada registro puede describirse mediante un vector,

$$\vec{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$$

- El grado de semejanza entre dos observaciones dependerá entonces de la **distancia** entres sus vectores en este espacio multidimensional.

- En el contexto de *clustering,* los elementos con **distancia pequeñas** tienden a agruparse dentro del mismo Clustering, mientras que aquellos con **distancias mayores** se ubican en grupos distintos. Esta noción de proximidad es la base de la mayoría de los algoritmos de agrupamiento.

# Clustering

## Ejemplo

| Usuarios | Star Wars | LOTR | Harry Potter |
|---|---|---|---|
| **1** | 1.2 | 4.9 | 2.1 |
| **2** | 2.1 | 8.1 | 7.9 |
| **3** | 7.4 | 3.0 | 9.9 |
| **4** | 5.6 | 0.5 | 1.8 |
| **5** | 1.5 | 8.3 | 2.6 |
| **6** | 2.5 | 3.7 | 6.5 |
| **7** | 2.0 | 8.2 | 8.5 |
| **8** | 1.8 | 9.3 | 4.5 |
| **9** | 2.6 | 1.7 | 3.1 |
| **10** | 1.5 | 4.7 | 2.3 |

# Clustering

## Métrica de Manhattan

- O también llamada distancia del taxi. Porque cuenta el número de calles (o manzanas, valor positivo) que separan un punto de otro. Distancias rectas, que es de las más alejadas.

$$D_1(x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}| = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

| Usuarios | Star Wars | LOTR | Harry Potter |
|---|---|---|---|
| **1** | 1.2 | 4.9 | 2.1 |
| **2** | 2.1 | 8.1 | 7.9 |
| **Resta** | -0.9 | -3.2 | -5.8 |
| **Valor absoluto** | 0.9 | 3.2 | 5.8 |

Entonces, ambos usuarios están a una distancia $D_1(x_1, x_2) = 0.9 + 3.2 + 5.8 = 9.9$

# Clustering

## Métrica euclidiana

- Esta sería la distancia en una línea recta.

$$D_2(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

| Usuarios | Star Wars | LOTR | Harry Potter |
|---|---|---|---|
| **1** | 1.2 | 4.9 | 2.1 |
| **2** | 2.1 | 8.1 | 7.9 |
| **Resta** | -0.9 | -3.2 | -5.8 |
| **Al cuadrado** | 0.81 | 10.24 | 33.64 |

Entonces, ambos usuarios están a una distancia $D_1(x_1, x_2) = \sqrt{0.81 + 10.24 + 33.64} = 6.68$

# Clustering

## Métrica de Minkowski

$$D_2(x_i, x_j) = ((x_{i1} - x_{j1})^p + (x_{i2} - x_{j2})^p + \cdots + (x_{in} - x_{jn})^p)^{\frac{1}{p}} = \left( \sum_{k=1}^{n} (x_{ik} - x_{jk})^p \right)^{\frac{1}{p}}$$

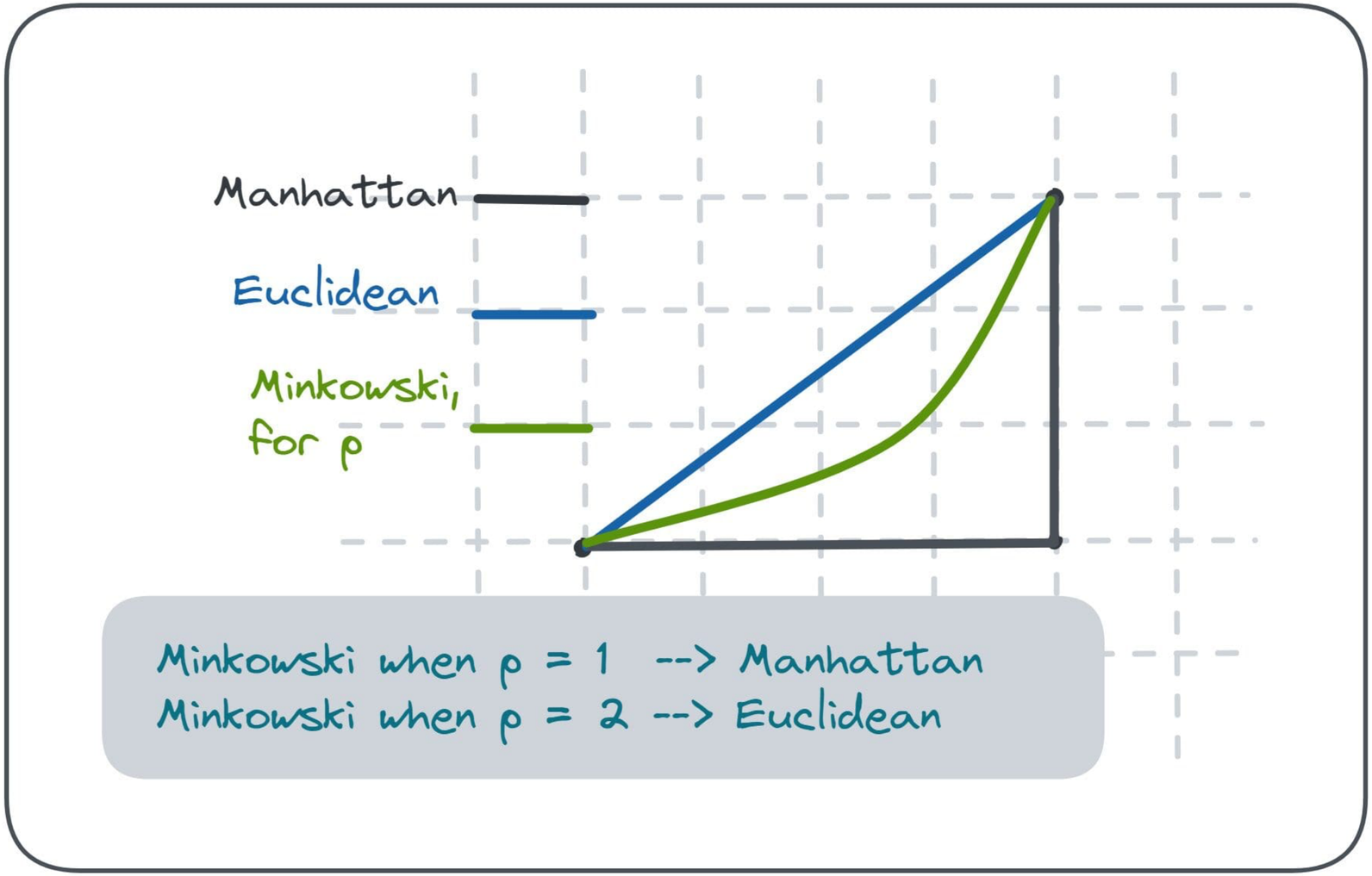
Es una forma general de medir la distancia entre dos puntos. Dependiendo del valor del parámetro $p$, esta métrica adopta distintas formas conocidas,

- Para $p = 1$, se obtiene la **distancia de Manhattan.**
- Para $p = 2$, se obtiene la **distancia euclidea.**

De manera general, a medida que el valor de $p$ aumenta, se reduce la penalización por diferencias grandes, haciendo que las distancias más significativas tengan un menor impacto relativo en el cálculo total.

# Clustering

## Distancias



# Clustering

## Matriz de distancias

- Las distancias entre todas las observaciones de un conjunto de datos se pueden resumir en una **matriz de distancias.** Para construirla, se elige una métrica (por ejemplo, euclidiana o Manhattan) y se calcula la distancia entre **todos los pares posibles de puntos** del dataset.

- Si existen $n$ observaciones, el resultado es una **matriz cuadrada de orden** $n$**,** donde cada elemento $d_{ij}$ representa la distancia entre los puntos $x_i$ y $x_j$.

- Sin embargo, las distancias pueden resultar engañosas cuando las variables tienen **rango o escalas diferentes.** Para evitar esto, es común **normalizar** los valores de cada columna, de modo que todas las variables contribuyan de manera equivalente en el cálculo de distancias,

$$Z_{ij} = \frac{x_{ij} - \min(x_{*j})}{\max x_{*j} - \min(x_{*j})}$$

- Esta transformación (minmax) ajusta los valores a un rango común entre 0 y 1.

# Clustering jerárquico

# Clustering jerárquico

## Introducción

Los métodos jerárquicos buscan **agrupar o dividir** los datos de manera sucesiva, crean una estructura de árbol (dendrograma).

- En el enfoque **aglomerativo,** los clústers se van **fusionando** paso a paso.
- En el enfoque **divisivo,** un Clustering se **divide** en otros más pequeños.
- Ambos procesos se basan en **minimizar o maximizar** semejanzas entre observaciones.

Existen diversas estrategias para decidir cómo se combinan los grupos en cada etapa, y ningún procedimiento es universalmente óptimo. Por ello, la elección del método más adecuado dependerá del **criterio** del analista, y del **contexto** del problema que se esté estudiando.

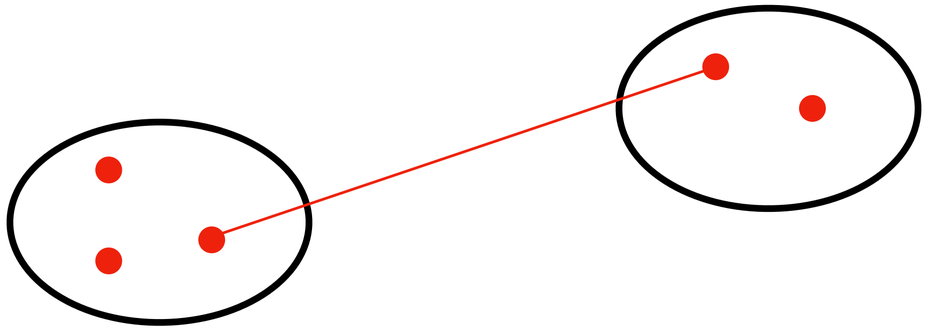
# Clustering jerárquico

## Enlace simple (Single Linkage)

- En este método, la **distancia entre dos clústers** se define como la **menor distancia** entre cualquier par de puntos pertenecientes a cada uno de ellos,

$$d(C_m, C_n) = \min(d(x_i, y_i)) \forall x_i \in C_m, y_j \in C_n$$

- En cada iteración, los clústers con la **distancia mínima** entre sí se **fusionan**. Luego, se recalculan las distancias entre los nuevos clústers de la misma forma.

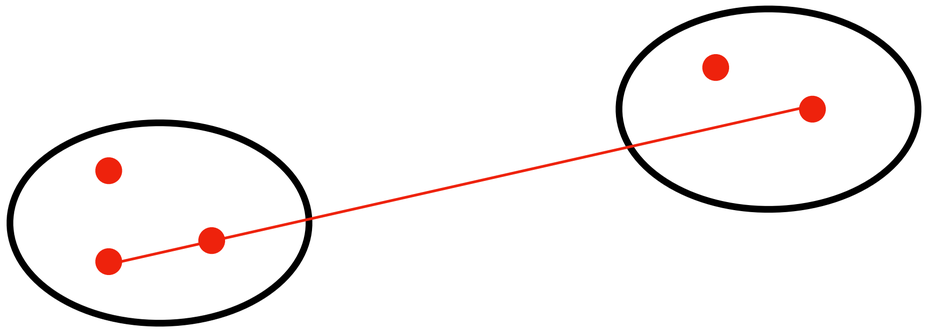


# Clustering jerárquico

## Enlace completo (Complete Linkage)

- En este método, la **distancia entre dos clústers** se define como la **mayor distancia** entre cualquier par de puntos pertenecientes a cada uno de ellos,

$$d(C_m, C_n) = \max(d(x_i, y_i)) \forall x_i \in C_m, y_j \in C_n$$

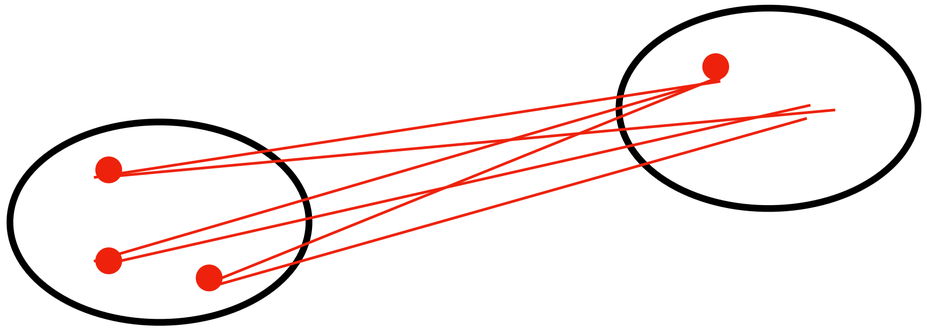


# Clustering jerárquico

## Enlace promedio (Average Linkage)

- En este método, la **distancia entre dos clústers** se define como el promedio de todas las distancias entre los puntos de ambos grupos,

$$d(C_m, C_n) = \text{mean}(d(x_i, y_i)) \forall x_i \in C_m, y_j \in C_n$$

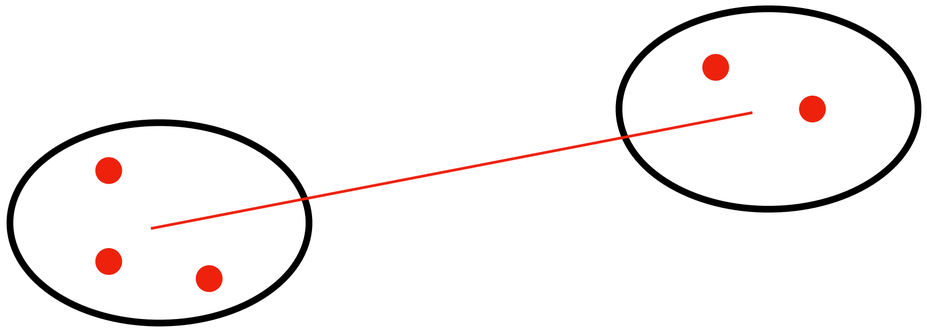


# Clustering jerárquico

## Enlace del centroide (Centroid Linkage)

- En este método, la **distancia entre dos clústers** se calcula como la distancia entre sus centroides, es decir, los puntos medios de cada grupo,

$$d(C_m, C_n) = d(\bar{C}_m, \bar{C}_n)$$



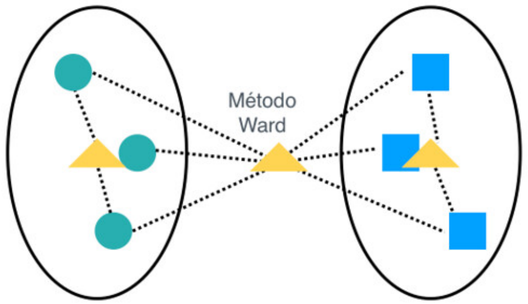
# Clustering jerárquico

## Enlace de Ward (Ward's linkage)

- El **método de Ward** fusiona en cada paso los dos clústers cuya unión produce el **menor incremento en la suma total del los cuadrados de las varianzas dentro de los grupos.**

- El criterio se basa en **minimizar la varianza interna** de los puntos dentro de cada clúster y, al mismo tiempo, mantener **baja la varianza global del conjunto de datos.** Es decir, se combinan de manera que el aumento del "error" (la pérdida de homogeneidad) sea mínimo.

- $$\Delta E = \sum_{i \in (C_m \cup c_n)} ||\mathbf{x}_i - \vec{\mathbf{x}}_{mn}||^2$$



# Clustering jerárquico

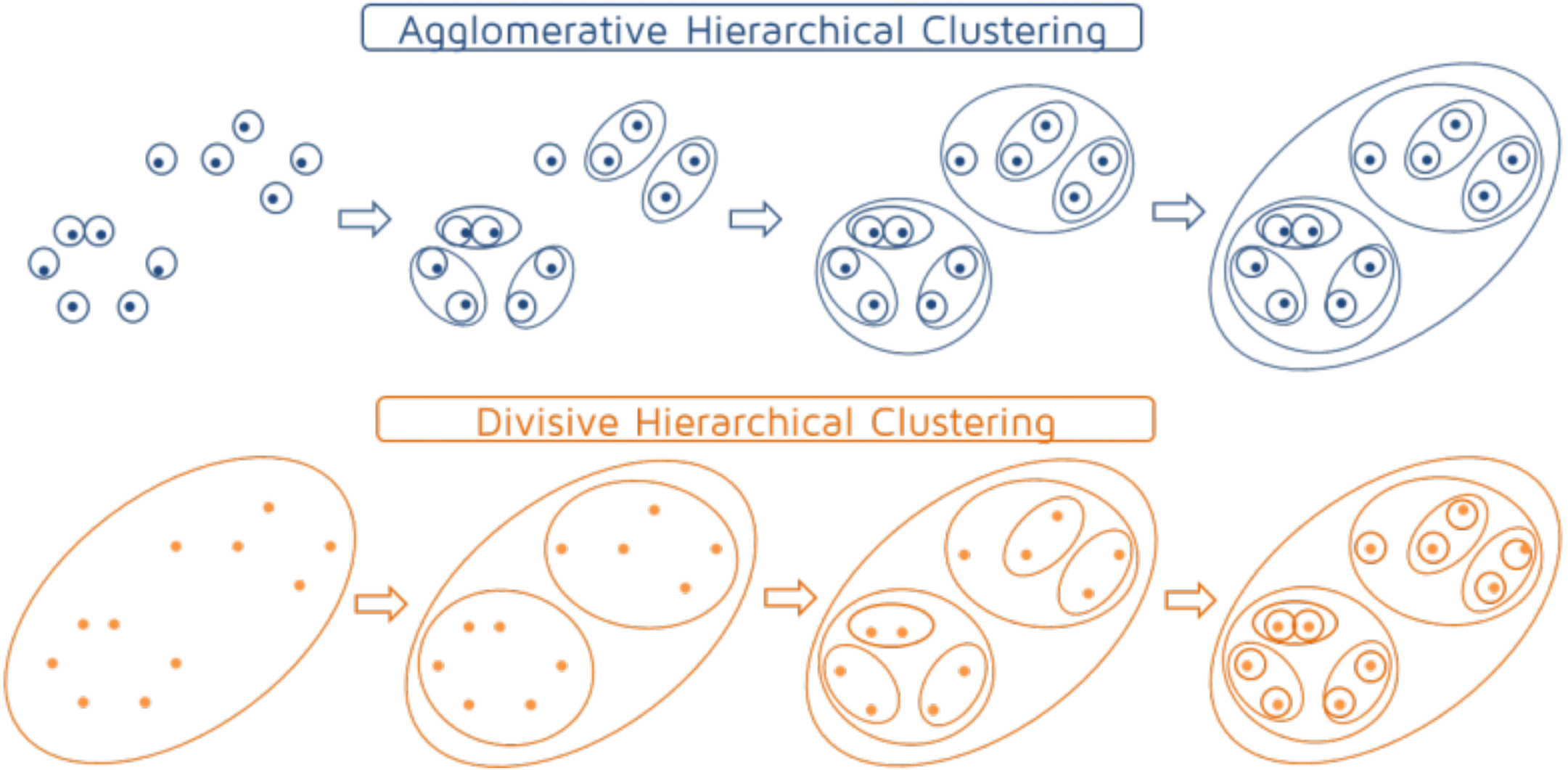
## Algoritmo

Es un **método jerárquico** en el que cada observación comienza siendo un clúster individual. En cada iteración, los clústers más similares se **fusionan progresivamente** hasta formar un único grupo con todas las observaciones. El procedimiento es el siguiente:

1. Cada observación inicia como un clúster (se tienen $N$ clústers).
2. Se identifica la **menor distancia** en la matriz de distancias y se unen los dos clústers correspondientes.
3. Se **recalculan** las distancias entre el nuevo clúster y los existentes, utilizando alguno de los **método de enlace** mencionados previamente (simple, completo, promedio, centroide, Ward).
4. Se repite el proceso hasta que todas las observaciones quedan unidas en un único clúster.

El resultado puede representarse mediante un **dendrograma**, que muestra visualmente cómo se van fusionando los clústers a lo largo del proceso.

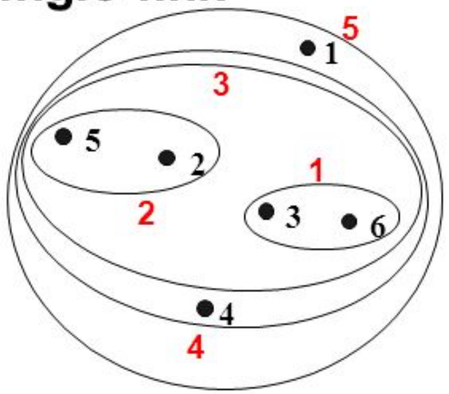
# Clustering jerárquico



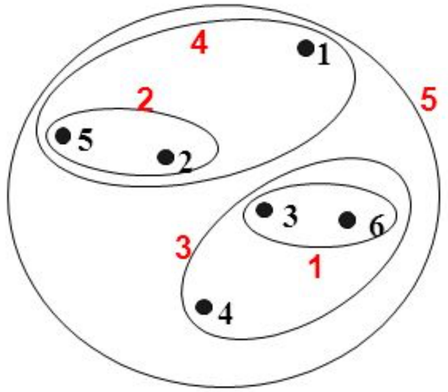
# Clustering jerárquico

## Hierarchical Clustering: Comparison

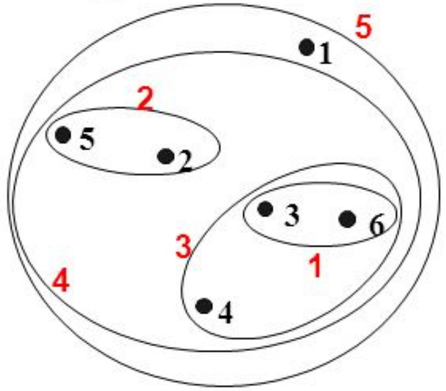
**Single-link**



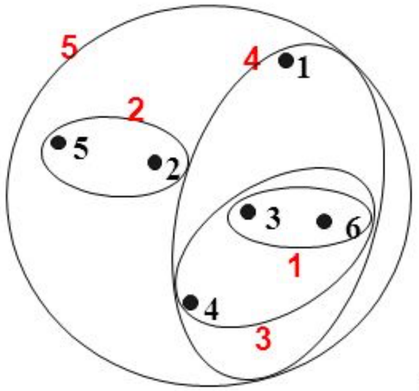

**Complete-link**




**Average-link**




**Centroid distance**



# Clustering con k-means

# Clustering por k-means

## Algoritmo

En el mismo contexto, este modelo no supervisado busca dividir un conjunto de observaciones en un número fijo de grupos *k*, el cual debe **definirse previamente.**

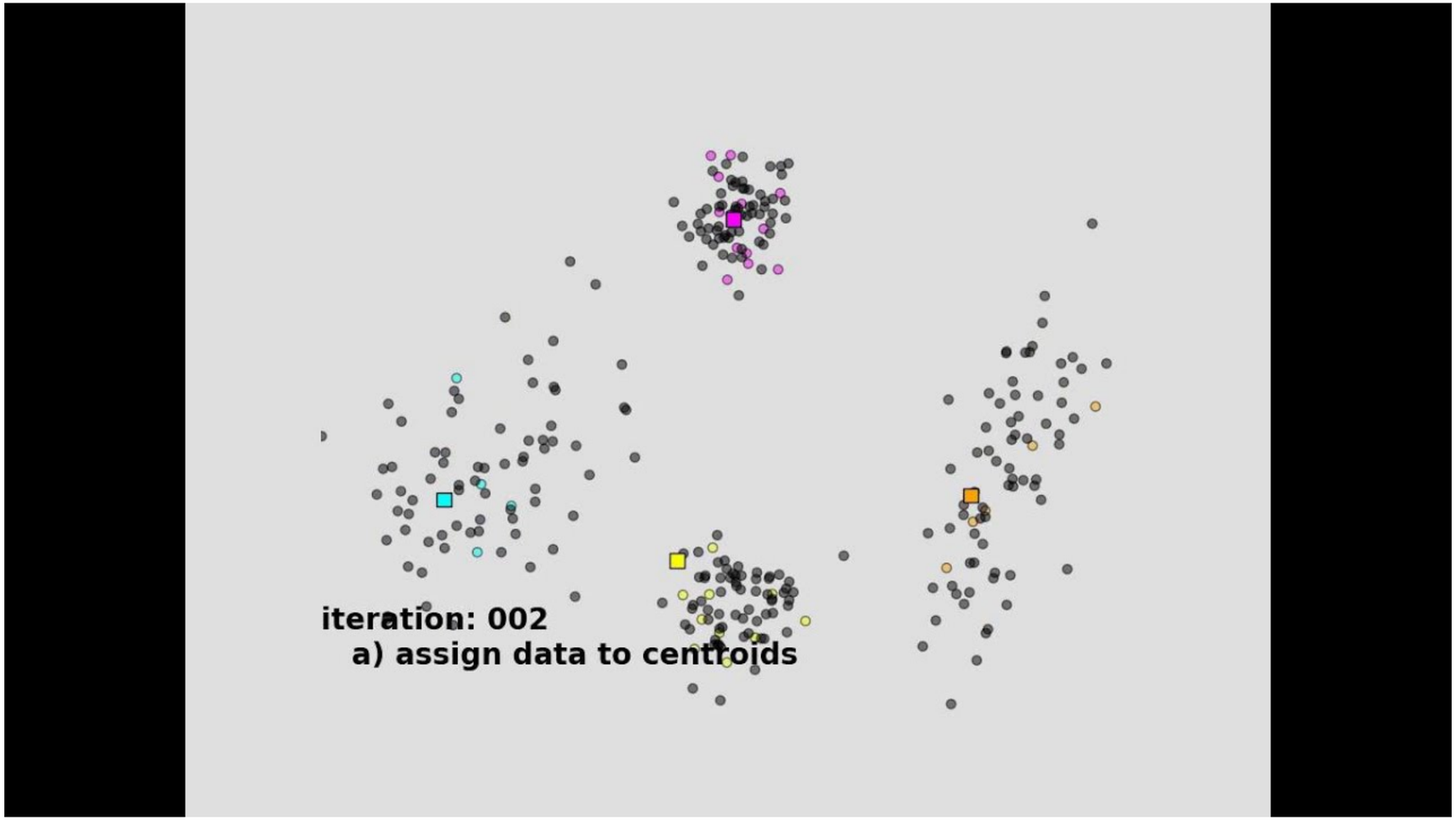
**Procedimiento**

1. Se eligen *k* **centroides** de forma aleatoria o determinada por el usuario.
2. Cada punto del dataset se **asigna** al centroide más cercano.
3. Con los puntos de cada clúster, se **recalcula el centroide** como el promedio de sus elementos.
4. Se repite el proceso de asignación y actualización hasta que los centroides dejan de cambiar significativamente.

Este método tienen a formar clústers **compactos y esféricos**, siendo uno de los más utilizados por su simplicidad y eficiencia.

# Clustering por k-means

## Algoritmo



# Clustering por k-means

## Distancia intracluster

- La **distancia intraclúster** mide la **cohesión** dentro de cada grupo, i.e., qué tan cercanos están los puntos de su centroide. Para un clúster $C_j$, se define como,

$$SS_w(C_j) = \sum_{x \in C_j} (x - c_j)^2$$

- Para evaluar la eficiencia global del modelo, se utiliza la **distancia intracluster normalizada** respecto a la variabilidad total,

$$S\tilde{S}_W = \sum_{j=1}^{k} \frac{SS_W(C_j)}{SS_T}, \text{ donde } SS_T = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Valore pequeños de $\tilde{S}_W$ indican **clústers más compactos y mejor definidos.**

# Clustering por k-means

## Distancia intracluster

- El objetivo del algoritmo k-means es encontrar una configuración de **centroides** que **minimice la suma total de las distancias cuadráticas** entre los puntos y su centroide correspondiente,

$$SS_W(k) = \sum_{j=1}^{k} SS_W(C_j) = \sum_{j=1}^{k} \sum_{x_i \in C_j} (x_i - c_j)^2$$

Donde,

- $k$ es el número de clústers definidos.
- $x_i$ son los puntos que pertenecen al cluster $C_j$.
- $c_j$ es el centroide del clúster $j$

# Clustering por k-means

## Evaluación: Coeficiente de la silueta

El **coeficiente de la silueta** mide qué tan bien asignado está un punto dentro de su clúster.

- $a(i)$, promedio de la distancia del punto $i$ a todos los puntos de su mismo clúster.
- $b(i)$, menor distancia promedio del punto $i$ a los puntos de otro clúster, es decir, el clúster más cercano.

El coeficiente se define como

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \Rightarrow S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{si a(i)} < \text{b(i)} \\ 0 & \text{si a(i)} = \text{b(i)} \\ \frac{b(i)}{a(i)} - 1 & \text{si a(i)} > \text{b(i)} \end{cases}$$

# Clustering por k-means

## Evaluación: Coeficiente de la silueta

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \Rightarrow S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{si a(i) < b(i)} \\ 0 & \text{si a(i) = b(i)} \\ \frac{b(i)}{a(i)} - 1 & \text{si a(i) > b(i)} \end{cases}$$

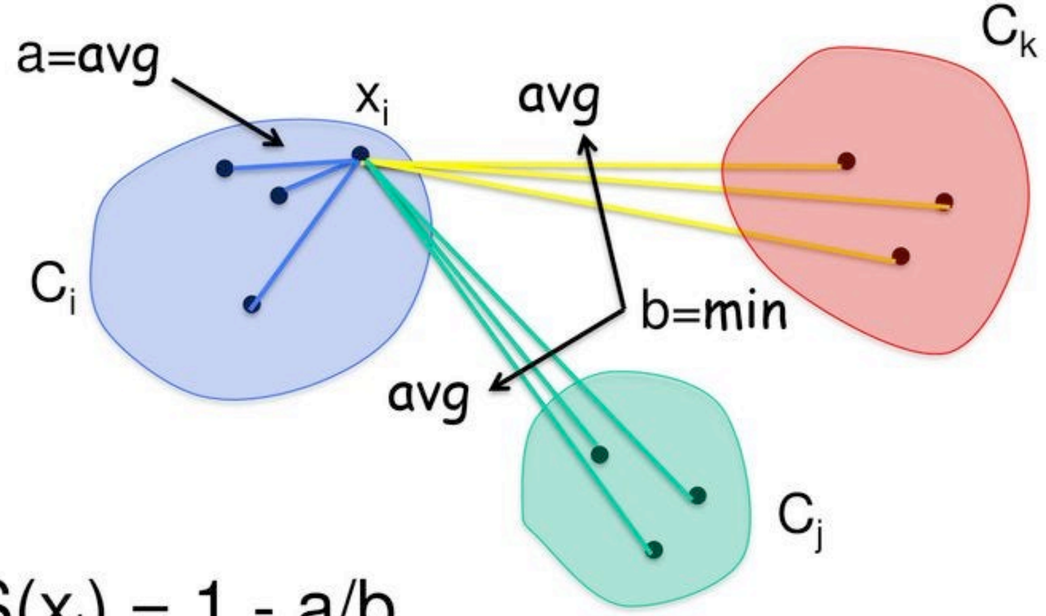
- Si $S(i)$ está cerca de 1, el punto está bien agrupado.
- Si $S(i)$ es cercano a 0, el punto está en el límite entre dos clústers.
- Si $S(i)$ es negativo, el punto probablemente fue mal asignado.
- El promedio de $S(i)$ sobre todos los puntos indica la calidad general del agrupamiento y puede usarse como criterio para elegir el número óptimo de clústers.

# Clustering por k-means

## Evaluación: Coeficiente de la silueta

### Silhouette Coefficient

- The idea...




- Usually, $S(x_i) = 1 - a/b$