

# Clustering

---

Diplomatura en Ciencia de Datos,  
Aprendizaje Automático y sus Aplicaciones  
FaMAF-UNC  
agosto 2019

# Mapa de ruta

1. Reglas de Asociación
2. Aprendizaje Semi-supervisado
3. **Clustering**, y revisitar todos los conceptos que vimos hasta ahora
4. Embeddings
5. K-nn y recomendación

Entregables:

- Trabajos con sus mentores
- para ello van a tener a disposición notebooks para rehacer las figuras de la clase

# Mapa de ruta

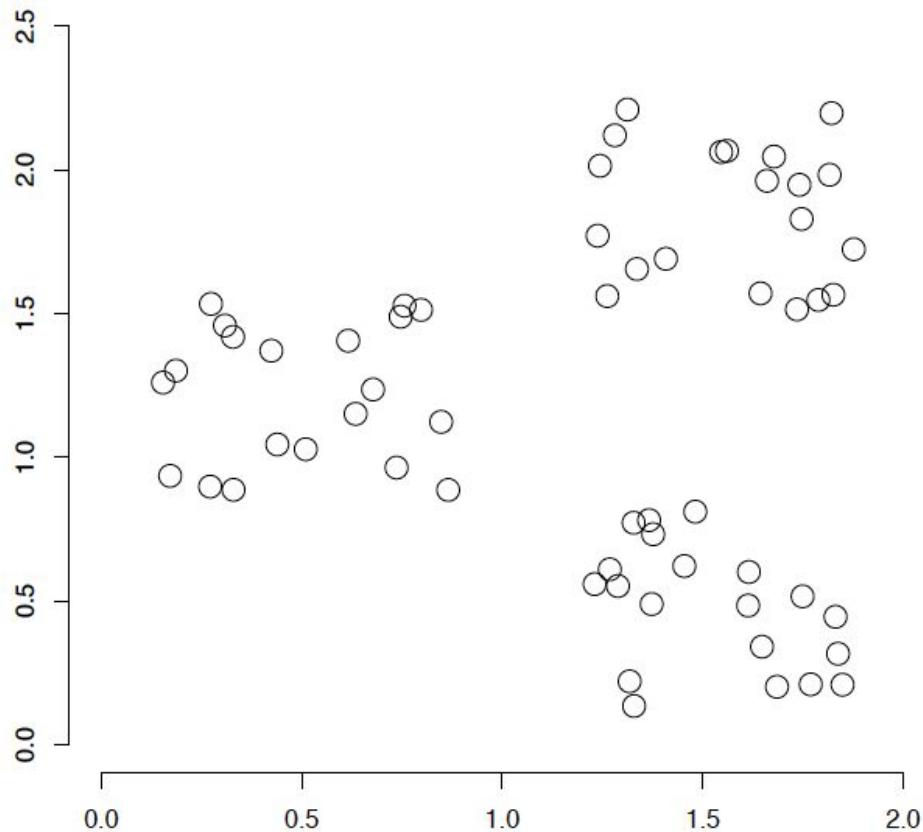
1. Cómo funciona
2. Evaluación
3. Qué puedo esperar
4. Metodología iterativa
5. Ejemplos con notebooks

# Cómo funciona clustering

Agrupar objetos semejantes

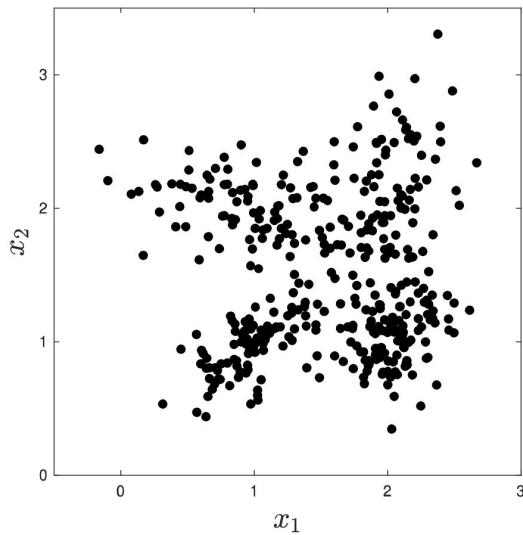
- Entrada: objetos en un espacio n-dimensional
- Salida: una **solución** con grupos (**clusters**) de objetos semejantes → cercanos en el espacio
  - Se minimiza la distancia entre los objetos de un mismo grupo
  - Se maximiza la distancia entre los objetos de distintos clusters
- Los centros de cada cluster son los **centroides**

# Dataset con clara estructura de clusters



¿Cómo sería un algoritmo para encontrar clusters en este espacio?

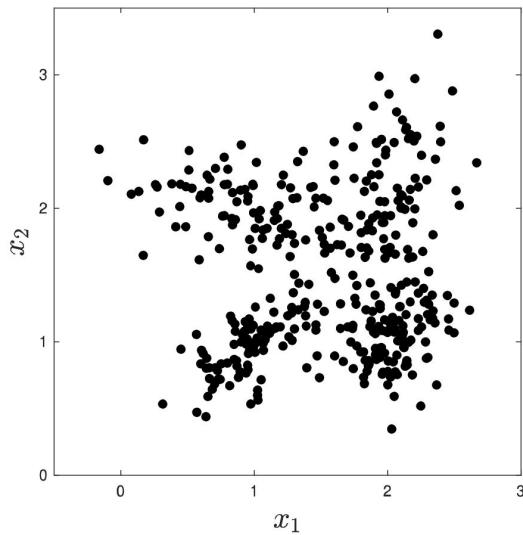
# Dataset con no tan clara estructura de clusters



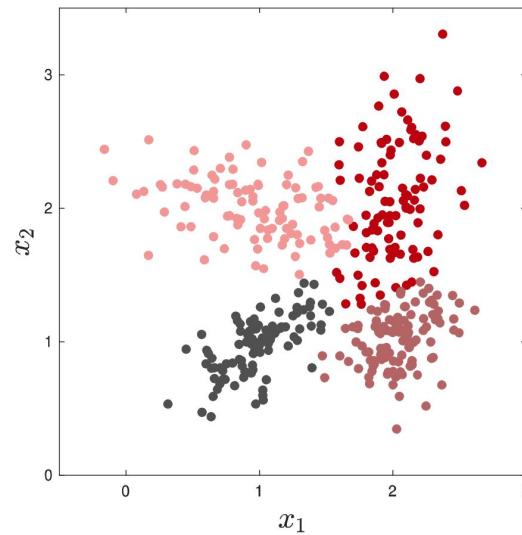
(a)

¿Cómo sería un algoritmo para encontrar clusters en este espacio?

# Dataset con no tan clara estructura de clusters



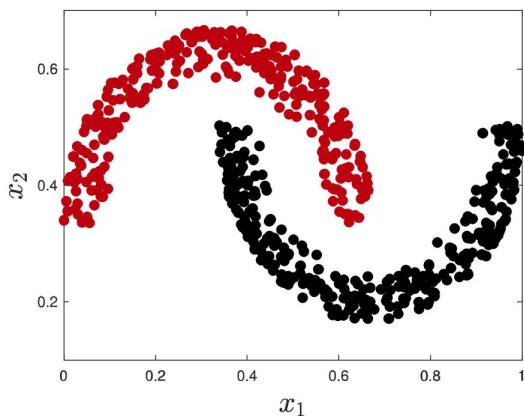
(a)



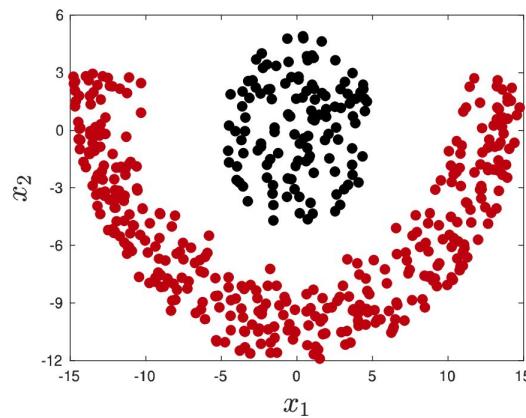
(b)

¿Cómo sería un algoritmo para encontrar clusters en este espacio?

# Dataset con clara estructura de clusters



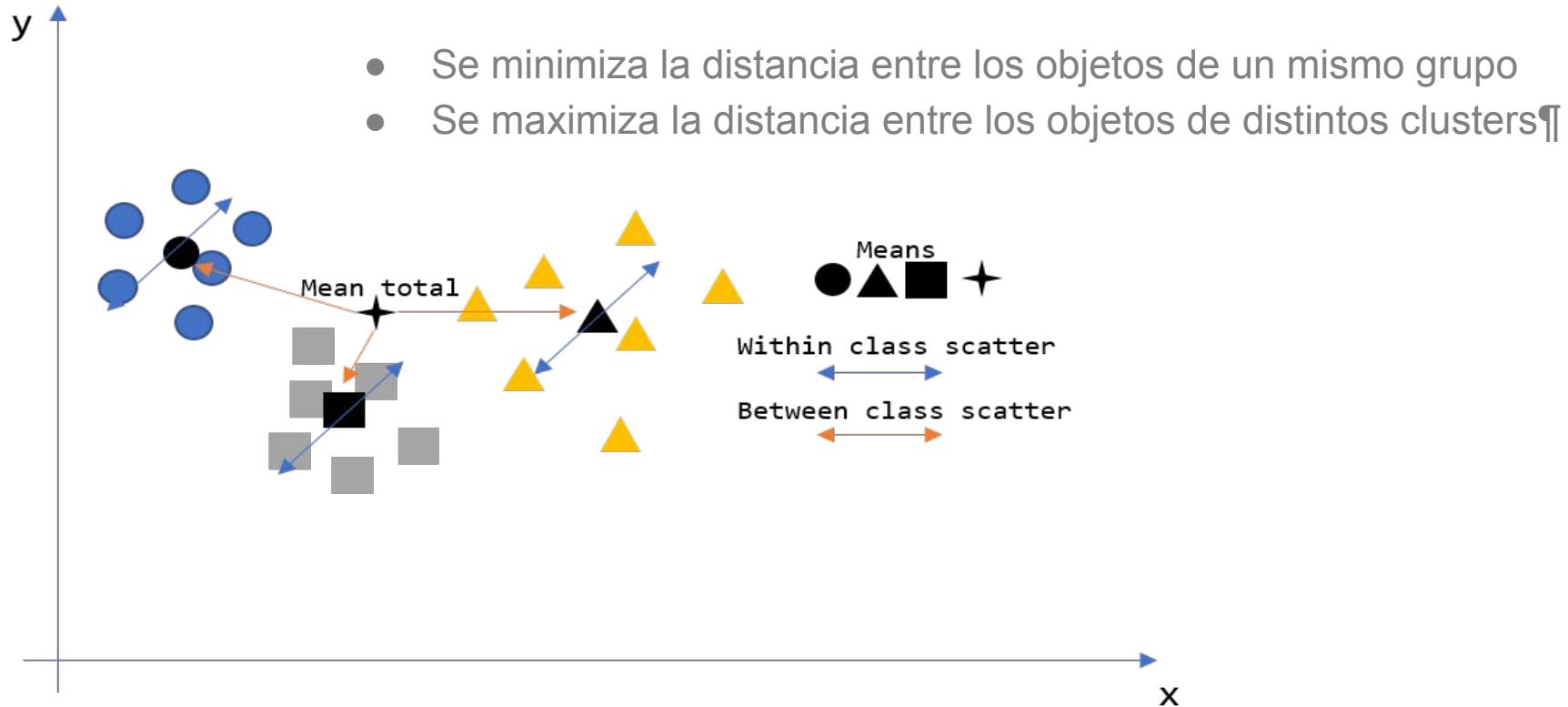
(a)



(b)

¿Cómo sería un algoritmo para encontrar clusters en este espacio?

# Cómo funciona clustering



# Cuestiones cruciales

- ❖ ¿Cómo es el espacio? ¿Cómo represento mis problemas?
- ❖ ¿Cómo se calcula la distancia (semejanza) en este espacio?
- ❖ ¿Cuántos clusters quiero distinguir?
- ❖ ¿Qué distribución tienen estos clusters? ¿Gaussiana? ¿En serie?
- ❖ ¿Busco una estructura jerárquica o plana?
- ❖ ¿Cómo veo qué hay en cada cluster?
- ❖ ¿Cómo evalúo la bondad de cada solución?

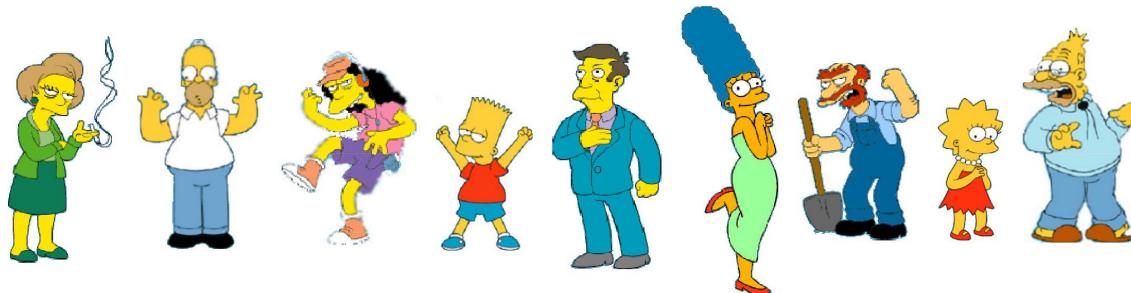
# ¿Cómo es el espacio? ¿Cómo represento mis problemas?

- Es multi dimensional?
- Mis datos son naturalmente categóricos? ordinales? continuos?
- Tengo información que me permite decir que debería encontrar grupos compactos?
- No sé nada y quiero usar clustering en forma exploratoria

# ¿Cómo se calculan las similaridades entre objetos en este espacio?

- ❖ Es un espacio Euclídeo? Métrica usual anda bien? Conviene usar ángulos en vez de distancias?
- ❖ No es un espacio Euclídeo? Similaridades ? Matriz de afinidad?
- ❖ Entender mi espacio me ayuda a elegir un método más razonable.
- ❖ Si mi método más razonable no me da nada, quizás sea porque no hay nada para ver...

# Datos



# Datos

id	sexo	fechnac	educ	catlab	salario	salini	T.emp	expprev	minoría
121	Mujer	6-agosto-1936	15	Administrativo	\$18.750	\$10.500	90	54	No
122	Mujer	26-septiembre-1965	15	Administrativo	\$32.550	\$13.500	90	22	No
123	Mujer	24-abril-1949	12	Administrativo	\$33.300	\$15.000	90	3	No
124	Mujer	29-mayo-1963	16	Administrativo	\$38.550	\$16.500	90	Ausente	No
125	Hombre	6-agosto-1956	12	Administrativo	\$27.450	\$15.000	90	173	Sí
126	Hombre	21-ene-1951	15	Seguridad	\$24.300	\$15.000	90	191	Sí
127	Hombre	1-septiembre-1950	12	Seguridad	\$30.750	\$15.000	90	209	Sí
128	Mujer	25-julio-1946	12	Administrativo	\$19.650	\$9.750	90	229	Sí
129	Hombre	18-julio-1959	17	Directivo	\$68.750	\$27.510	89	38	No
130	Hombre	6-septiembre-1958	20	Directivo	\$59.375	\$30.000	89	6	No
131	Hombre	8-febrero-1962	15	Administrativo	\$31.500	\$15.750	89	22	No
132	Hombre	17-mayo-1953	12	Administrativo	\$27.300	\$17.250	89	175	No
133	Hombre	12-septiembre-1959	15	Administrativo	\$27.000	\$15.750	89	87	No

# Medidas de similaridad

- ❖ A la hora de calcular la similaridad entre dos objetos
  - no tiene porqué usarse todos los datos
  - hay que tener cuidado con las magnitudes de cada variable
- ❖ No será posible que todas las variables tengan valores similares dentro de un mismo grupo, por lo que habrá que usar una medida de similaridad global entre elementos de un mismo grupo.

# ¿Cuántos clusters quiero distinguir?

- ❖ Tengo información? Hice varios experimentos? tengo varias databases de días diferentes y locales diferentes?
- ❖ Clustering exploratorio, debo estudiar los distintos agrupamientos para distintos números de clusters.
- ❖ Distintas técnicas para encontrar el mejor modelo de agrupamiento.

# ¿Busco una estructura jerárquica o plana?

- ❖ Si mis clusters estan anidados, tengo una estructura muy fuerte que explica los datos
- ❖ Si mis clusters son estructuras cercanas, las une sin remedio

# ¿Cómo veo qué hay en cada cluster?

- ❖ Visualización es la pesadilla. Rapido de correr, lento de analizar!!!
- ❖ Proyecciones en espacios de menor dimensión ayudan a visualizar los resultados.
  - Principal component analysis (PCA),
  - t-distributed Stochastic Neighbor Embedding (t-SNE)

# Datos

id	sexo	fechnac	educ	catlab	salario	salini	T.emp	expprev	minoría
121	Mujer	6-agosto-1936	15	Administrativo	\$18.750	\$10.500	90	54	No
122	Mujer	26-septiembre-1965	15	Administrativo	\$32.550	\$13.500	90	22	No
123	Mujer	24-abril-1949	12	Administrativo	\$33.300	\$15.000	90	3	No
124	Mujer	29-mayo-1963	16	Administrativo	\$38.550	\$16.500	90	Ausente	No
125	Hombre	6-agosto-1956	12	Administrativo	\$27.450	\$15.000	90	173	Sí
126	Hombre	21-ene-1951	15	Seguridad	\$24.300	\$15.000	90	191	Sí
127	Hombre	1-septiembre-1950	12	Seguridad	\$30.750	\$15.000	90	209	Sí
128	Mujer	25-julio-1946	12	Administrativo	\$19.650	\$9.750	90	229	Sí
129	Hombre	18-julio-1959	17	Directivo	\$68.750	\$27.510	89	38	No
130	Hombre	6-septiembre-1958	20	Directivo	\$59.375	\$30.000	89	6	No
131	Hombre	8-febrero-1962	15	Administrativo	\$31.500	\$15.750	89	22	No
132	Hombre	17-mayo-1953	12	Administrativo	\$27.300	\$17.250	89	175	No
133	Hombre	12-septiembre-1959	15	Administrativo	\$27.000	\$15.750	89	87	No

# Cómo evalúo la bondad de cada solución

- ❖ Para un método fijo,
  - Silhouette Score
  - Elbow method
  - BIC, AIC
- ❖ Varios métodos distintos
  - Rand measure
  - Mutual Information score
  - Contingency Matrix

# Semejanzas, Distancias y Afinidades

- ❖ La semejanza debería acercarse a las causas latentes
  - Entre documentos: semántica
  - Entre clientes: motivación para las compras
  - Entre imágenes: objetos físicos que representan
  - Entre propiedades inmobiliarias: elementos que otorgan valor
- ❖ Idealmente, debería calcularse de forma independiente para cada dimensión

# Normalización

- ❖ Atributos continuos
  - Para evitar que unas variables dominen sobre otras los valores de los atributos se normalizan a priori
  - - Desviación absoluta de la media
    - z-score
    -
- ❖ Atributos categoricos
  - encoding mediante transformaciones

# Distancias: datos continuos

- ❖ Euclídea
- ❖ Coseno → normalizado por longitud, producto punto → correlación!
- ❖ Distancia de Manhattan
- ❖ Distancia de mahalanobis
- ❖ Distancia de Edición (Levenshtein)

# Distancias: datos continuos

## Distancia de Minkowski

$$d_r(x, y) = \left( \sum_{j=1}^J |x_j - y_j|^r \right)^{\frac{1}{r}}, \quad r \geq 1$$

- Distancia de Manhattan ( $r=1$ ) / city block / taxicab

$$d_1(x, y) = \sum_{j=1}^J |x_j - y_j|$$

- Distancia euclídea ( $r=2$ ):

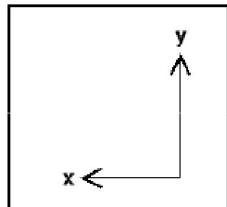
$$d_2(x, y) = \sqrt{\sum_{j=1}^J (x_j - y_j)^2}$$

- Distancia de Chebyshev ( $r \rightarrow \infty$ ) / dominio / chessboard

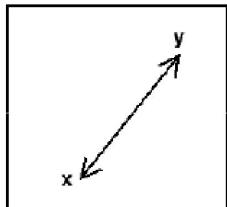
$$d_\infty(x, y) = \max_{j=1..J} |x_j - y_j|$$

# Distancias: datos continuos

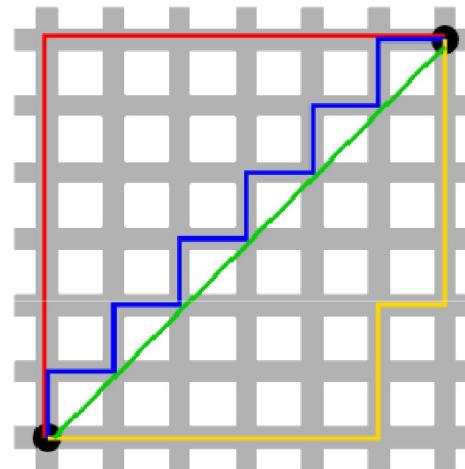
## Distancia de Minkowski



Manhattan



Euclidean



- Distancia de Manhattan = 12 (roja, azul o amarilla)
- Distancia euclídea  $\approx 8.5$  (verde - continua)
- Distancia de Chebyshev = 6 (verde - discreta)

# Distancias: datos continuos

## Distancia de Chebyshev

$$d_\infty(x, y) = \max_{j=1..J} |x_j - y_j|$$

También conocida como distancia de tablero de ajedrez (chessboard distance): Número de movimientos que el rey ha de hacer para llegar de una casilla a otra en un tablero de ajedrez.

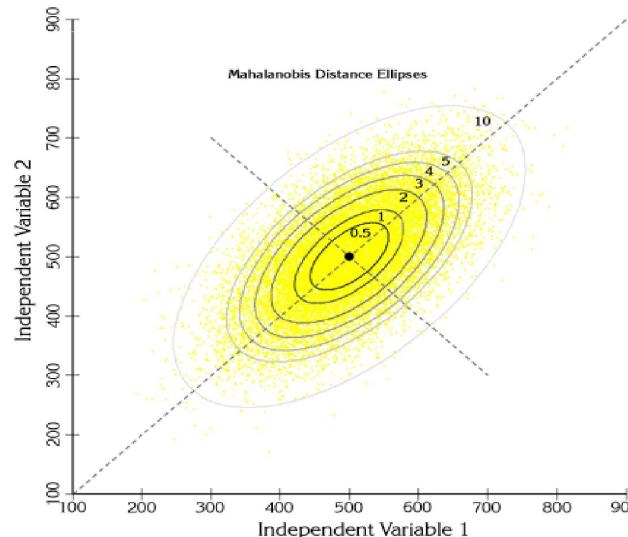
	a	b	c	d	e	f	g	h
8	5	4	3	2	2	2	2	2
7	5	4	3	2	1	1	1	2
6	5	4	3	2	1	1	1	2
5	5	4	3	2	1	1	1	2
4	5	4	3	2	2	2	2	2
3	5	4	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4
1	5	5	5	5	5	5	5	5
	a	b	c	d	e	f	g	h

# Distancias: datos continuos

## Distancia de Mahalanobis

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}.$$

- Considera las correlaciones entre variables.
- No depende de la escala de medida.



# Distancias: datos no continuos

## Distancia de edición = Distancia de Levenshtein

Número de operaciones necesario  
para transformar una cadena en otra.

$$d(\text{"data mining"}, \text{"data minino"}) = 1$$

$$d(\text{"efecto"}, \text{"defecto"}) = 1$$

$$d(\text{"poda"}, \text{"boda"}) = 1$$

$$d(\text{"night"}, \text{"natch"}) = d(\text{"natch"}, \text{"noche"}) =$$

Aplicaciones: Correctores ortográficos, reconocimiento de voz,  
detección de plagios, análisis de ADN...

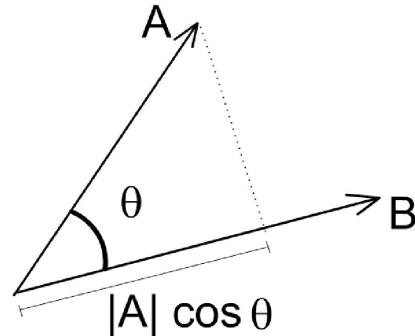
Para datos binarios: Distancia de Hamming

# Similaridades

## Medidas de correlación

- Producto escalar

$$S.(x, y) = x \cdot y = \sum_{j=1}^J x_j y_j$$



- “Cosine similarity”

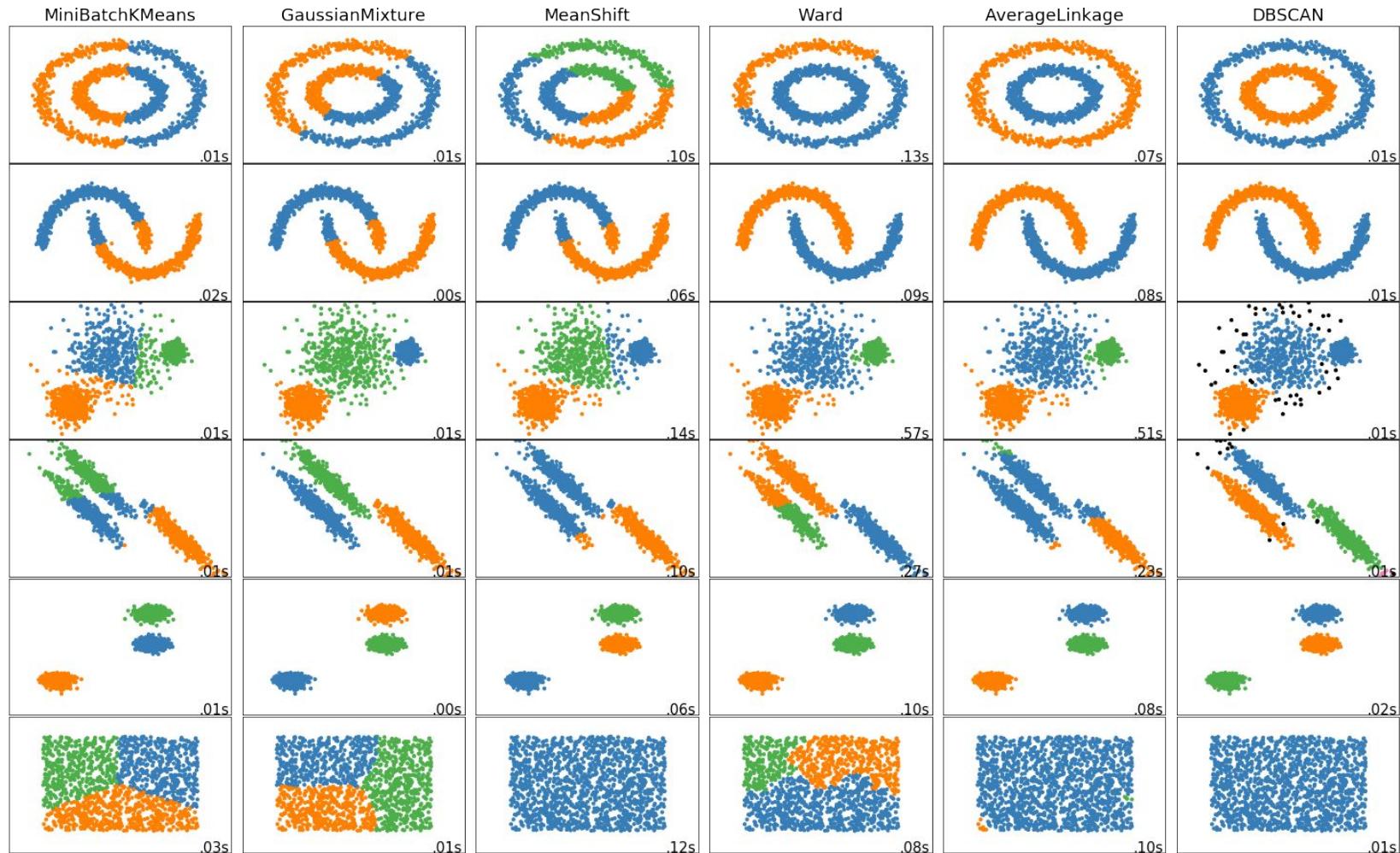
$$\cos(\vec{x}, \vec{y}) = \sum_i \frac{x_i \cdot y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$$

- Coeficiente de Tanimoto

$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{\vec{X}^t \cdot \vec{X} + \vec{Y}^t \cdot \vec{Y} - \vec{X}^t \cdot \vec{Y}},$$

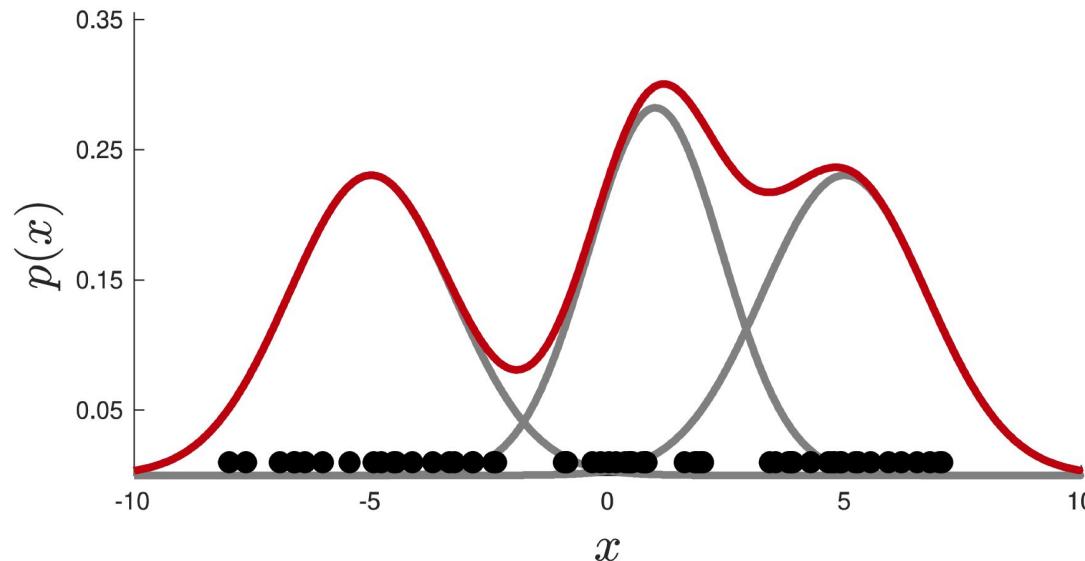
# Familias de algoritmos de clustering

- ❖ Métodos generativos
  - Mezcla de gaussianas, MeanShift
- ❖ Agrupamiento por particiones
  - k-Means, PAM/CLARA/CLARANS
- ❖ Métodos basados en densidad
  - DBSCAN, Optics, DenClue
- ❖ Clustering jerárquico
  - Ward, Diana/Agnes, BIRCH, CURE, Chameleon, ROCK
- ❖ Note\_fig1.ipynb crea la figura siguiente



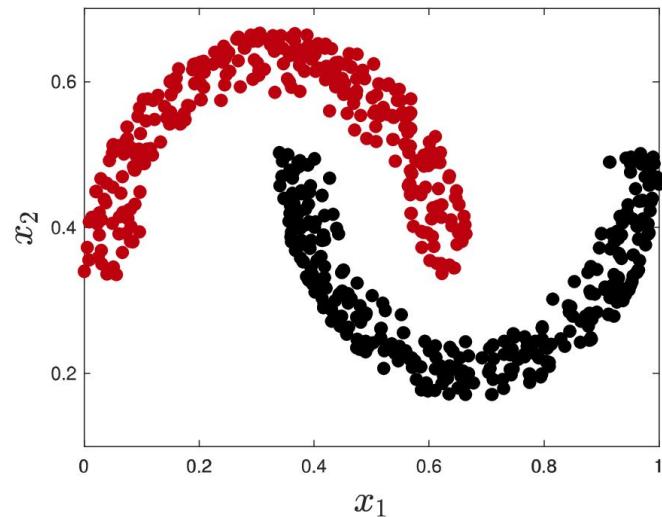
# Mezcla de Gaussianas

- ❖ Supongamos tener alguna información
  - Consideremos que estos datos son reales,
  - puedo trabajar con la distancia Euclídea.
  - datos producidos por una densidad mezcla de Gaussianas,

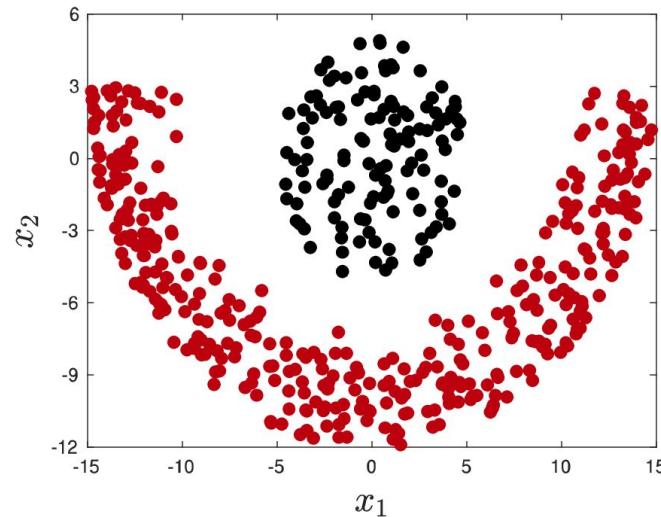


# Mezcla de Gaussianas

- ❖ Cualquier dato puede ser modelado con una mezcla de gaussianas?



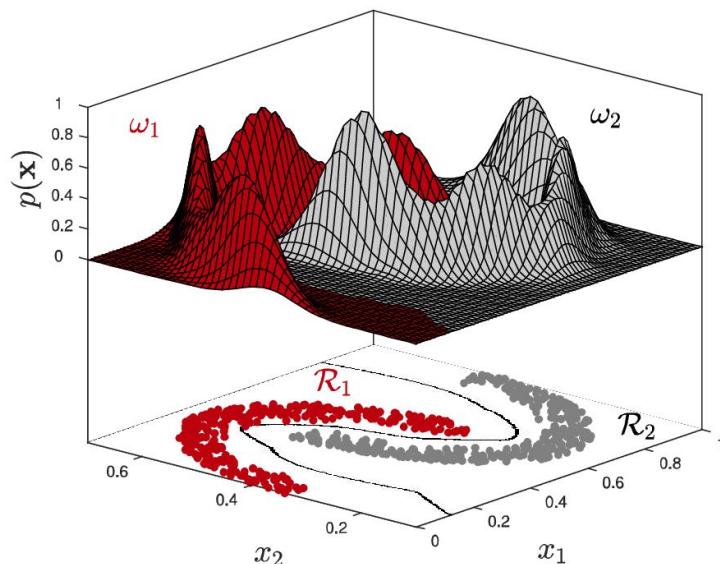
(a)



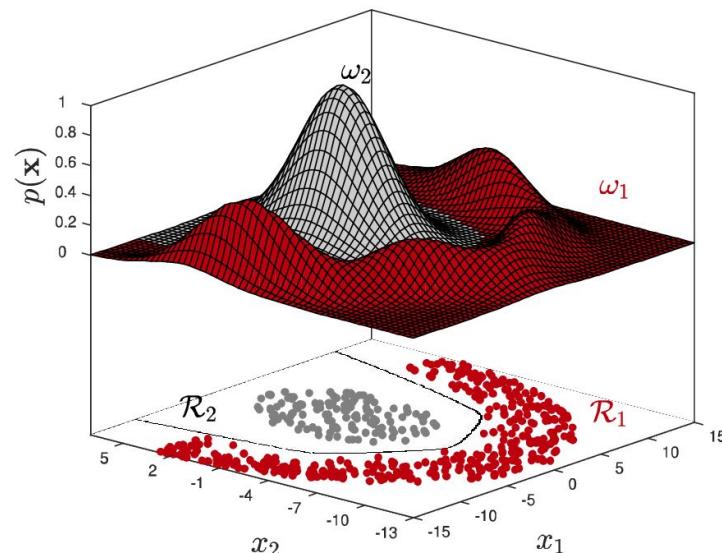
(b)

# Mezcla de Gaussianas

- ❖ No todos, pero muchos si se pueden modelar, si uno conoce la cantidad de gaussianas que forman la mezcla
- ❖



(a)



(b)

# Como funciona el GMM?

- ❖ Si uno fija la cantidad de gaussianas que uno considera que hay en la mezcla,
  - se estiman los parámetros de cada gaussiana y los parámetros de representación
  - se imputa cada dato como proveniente de la una de las componentes de la mezcla.
  - La estimación se realiza mediante el algoritmo Expectation Maximization.
-

# Como funciona el GMM?

Expectation Maximization Algorithm

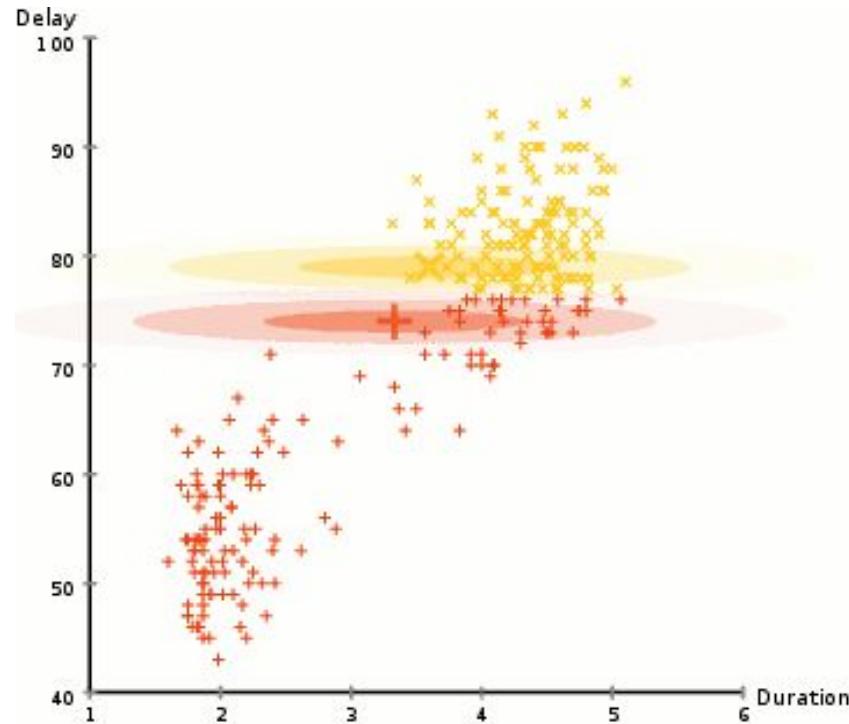
Input:  $\mathbf{X} = \{\mathbf{x}_i | i = 1, \dots, n\}$  datos,  $m$ =numero de componentes,  $\epsilon$  tolerancia

Output:  $\hat{\Theta}$

- ##  $\hat{\Theta}^0 \leftarrow \left[ \left( \hat{\theta}_1, \hat{p}_1 \right)^0, \dots, \left( \hat{\theta}_m, \hat{p}_m \right)^0 \right]$
  - ##  $t \leftarrow 0$
- ## do
- ##  $t \leftarrow t + 1$
  - ## **Paso-E:**  $Q(\Theta; \hat{\Theta}^t) \leftarrow \mathbb{E} \left[ \sum_{i=1}^{\infty} \ln \left( p(\mathbf{x}_i | j; \hat{\Theta}_j^t) p_j^t \right) \right]$
  - ## **Paso-M:**  $\hat{\Theta}^{t+1} \leftarrow \arg \max_{\Theta} Q(\Theta; \hat{\Theta}^t)$
  - ## until  $|Q(\Theta; \hat{\Theta}^t) - Q(\Theta; \hat{\Theta}^{t+1})| < \epsilon$

# Como funciona el GMM?

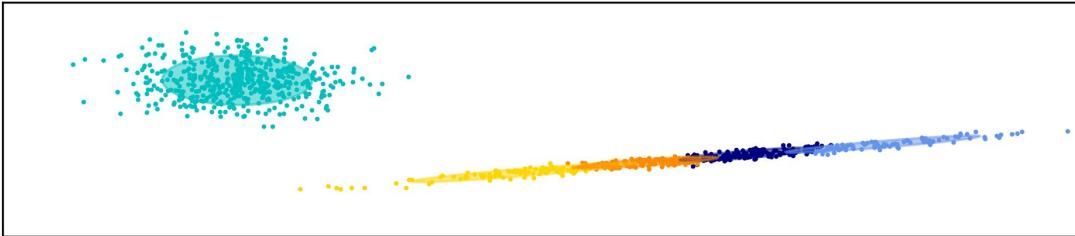
- ❖ Comenzamos con una partición aleatoria de la cual se sacan los parámetros de inicio y desde allí se itera



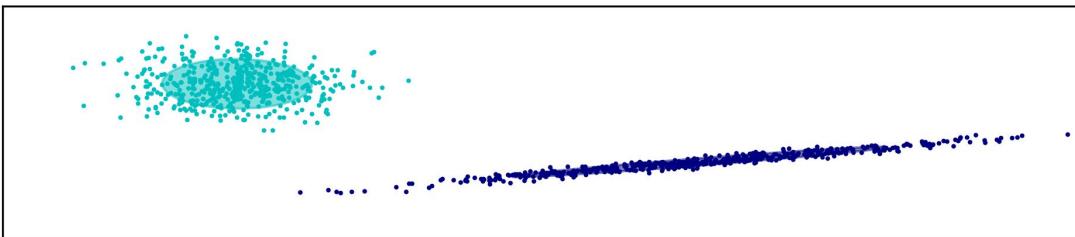
# Parámetros

- ❖ Gran problema de GMM es la determinación del número de componentes de la mezcla
- ❖ Si no se elige un buen número, el modelo partitiona de forma aglutinada pero los clusters pueden no tener sentido.
- ❖ La otra característica que puede ser forzada de inicio es el tipo de matriz de varianza covarianza.
- ❖ Este ejemplo (Note\_fig2.ipynb) ha sido realizado modelando matrices de covarianza full usando el módulo sklearn.

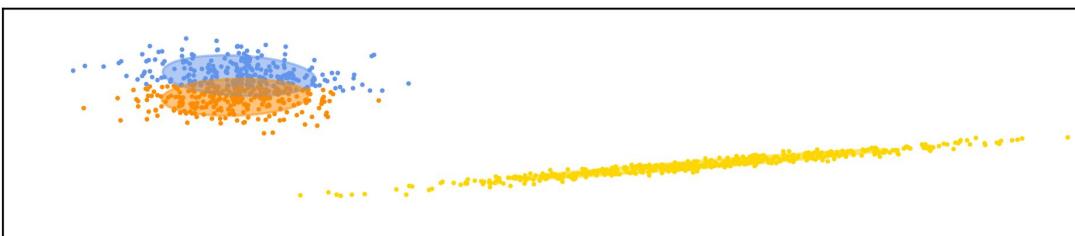
Gaussian Mixture K=5



Gaussian Mixture=2



Gaussian Mixture=3



# Detección automática de k

- ❖ Bayesian Information Criterium (BIC) da un score al modelo con m parámetros.

$$BIC = -2 \cdot \log(L(\Theta)) + \log(n)m$$

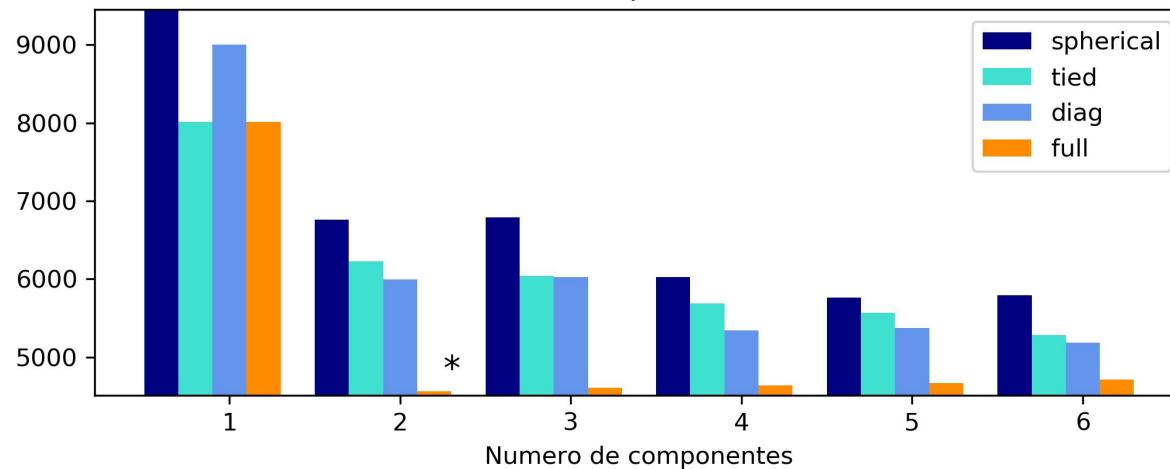
- ❖ Puede usarse otro índice llamado Akaike Information Criterium (AIC)

$$AIC = -2 \cdot \log(L(\Theta)) + 2m$$

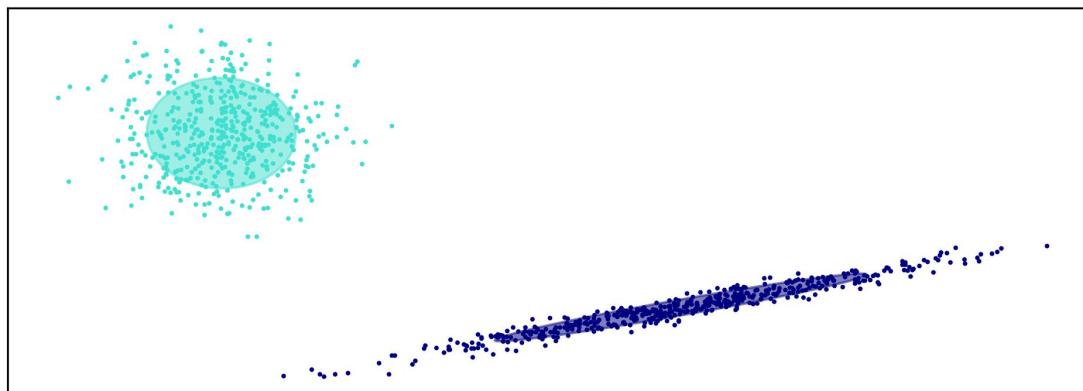
donde  $L(\Theta)$  es la verosimilitud,  $n$  el número de datos, y  $m$  el número de parámetros estimados, ( $k$ , el número de componentes, más las medias y entradas de la matriz de varianza covarianza.)

- ❖ La figura siguiente esta generada por el script Note\_fig3.ipynb

BIC score por modelo



GMM Seleccionado: modelo completo con 2 componentes



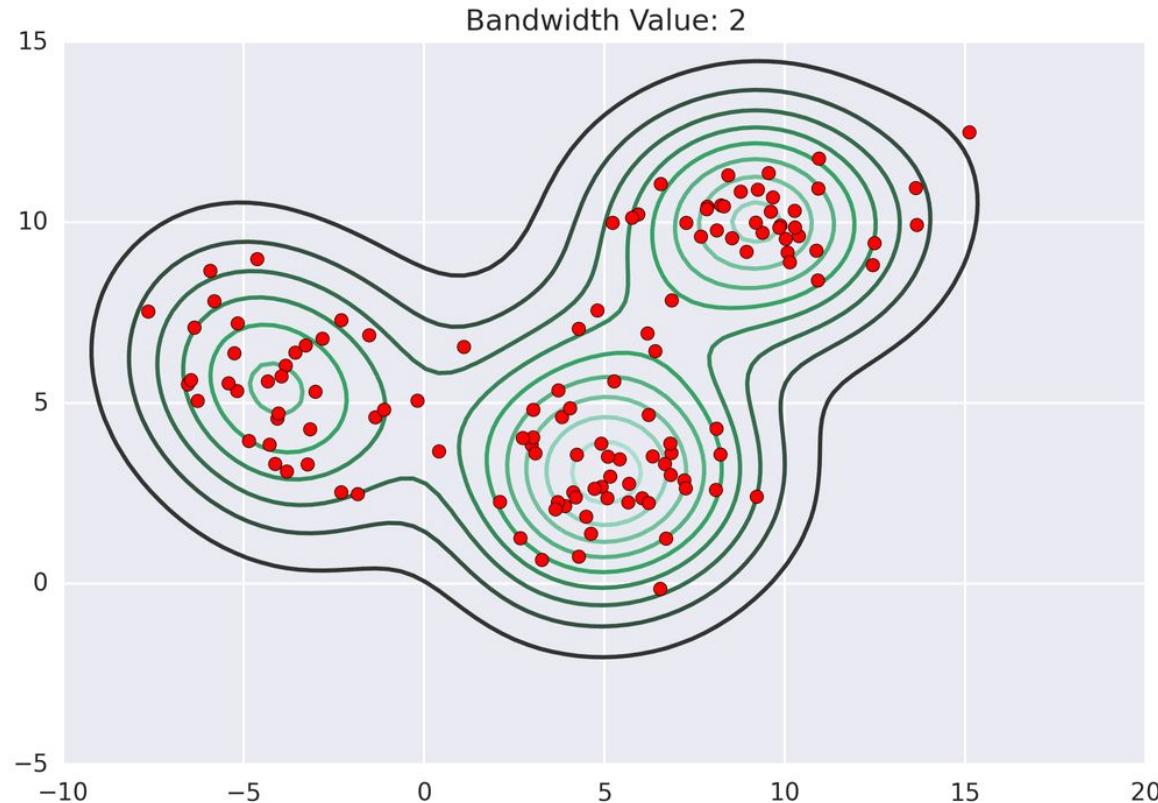
# Mean Shift Algorithm

- Mean shift se basa en el concepto de kernel density estimation (KDE)
- 
- Si los datos se suponen muestrados de una distribución de probabilidad, KDE es un estimador no paramétrico de la densidad asociada a dicha distribución.
- 
- KDE aplica un kernel, esto es, una función de peso, en una ventana alrededor del punto con un ancho de banda (bandwidth) determinado. Sumando todos las estimaciones individuales se obtiene el estimador de la densidad.

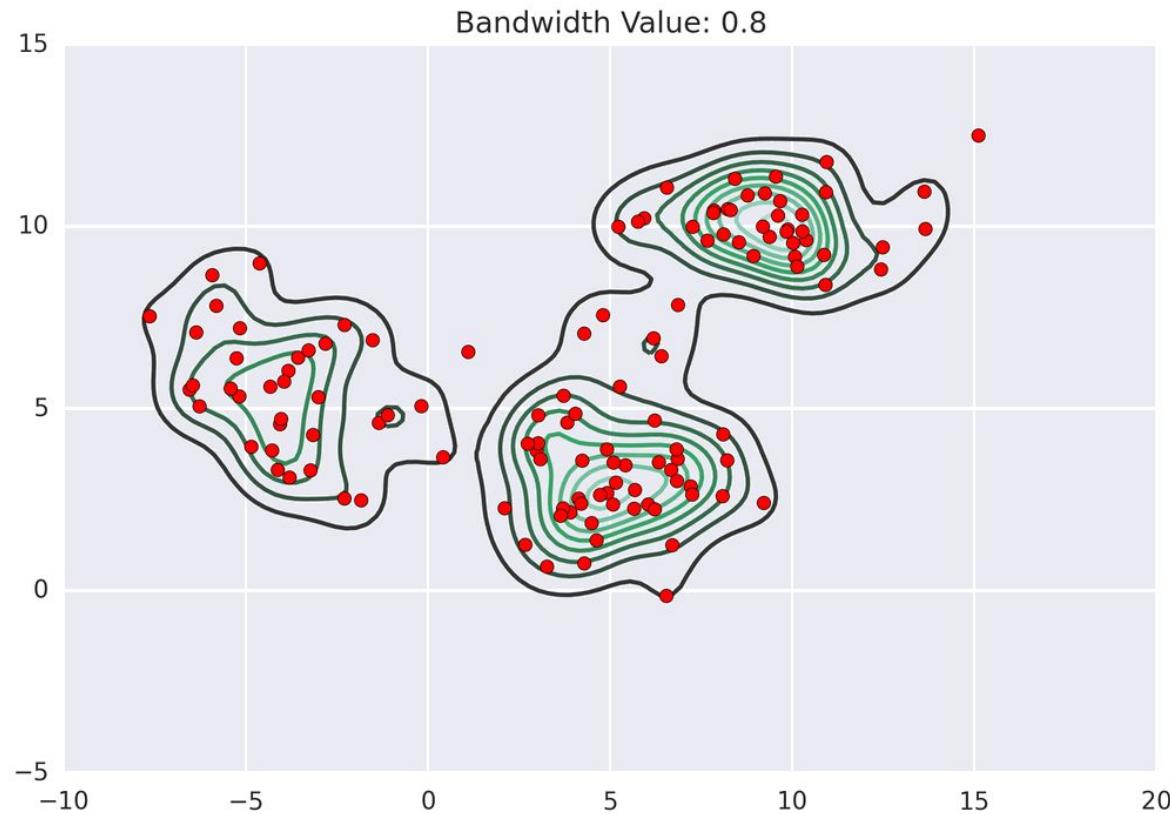
# Mean Shift Algorithm

- Para generar la partición, el algoritmo Mean-Shift Clustering va deslizando la ventana y computando el promedio de los datos pesados por el kernel, para localizar las áreas de alta densidad.
- ❖
- Cada moda de la densidad va a ser considerada un centroide, y los puntos de la partición van a ser asignados al centroide más próximo

# Mean Shift Algorithm



# Mean Shift Algorithm

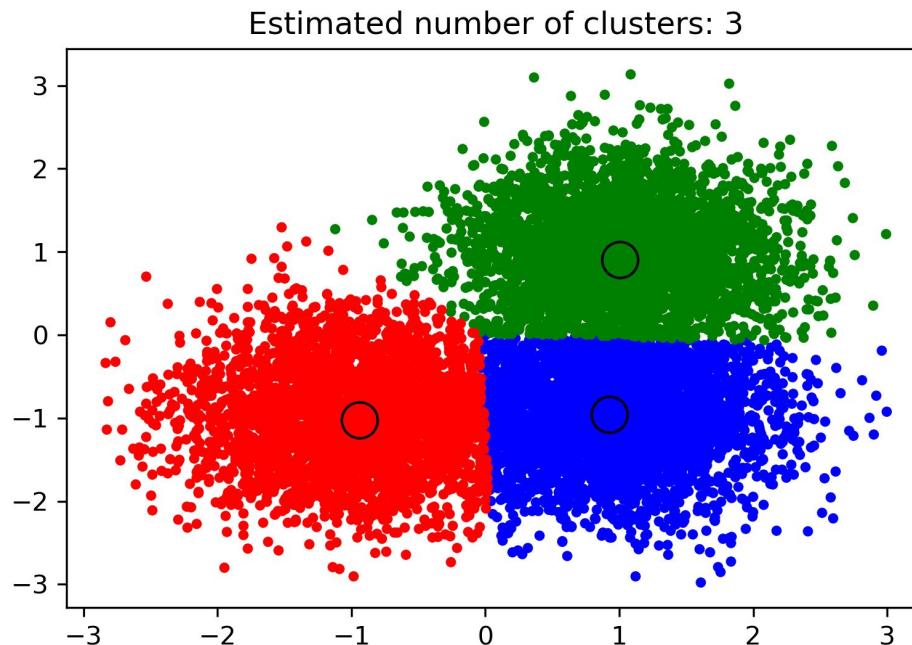


# Mean Shift Algorithm

- ❖ Parámetro bandwidth puede ser fijado a priori.
  -
- ❖ No tiene sentido usar BIC o AIC pues no se está fijando el modelo paramétrico
  -
- ❖ Puede ser estimado utilizando la teoría no paramétrica, dependiendo de que kernel se use.
  -
- ❖ Note\_fig4.ipynb tiene un ejemplo de Mean Shift automático y con k fijo.

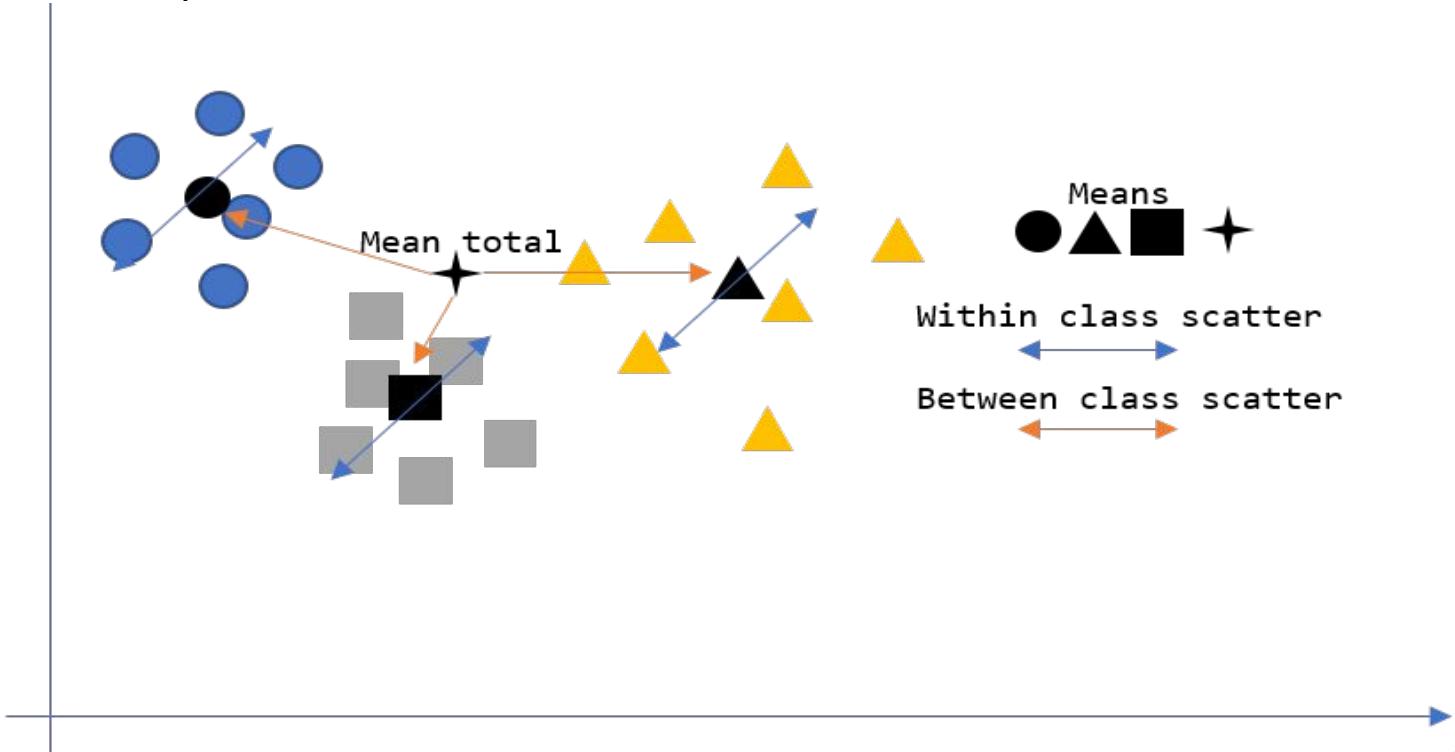
# Mean Shift Algorithm

Note\_fig4.ipynb tiene un ejemplo de Mean Shift automático y con k fijo.



# K-means

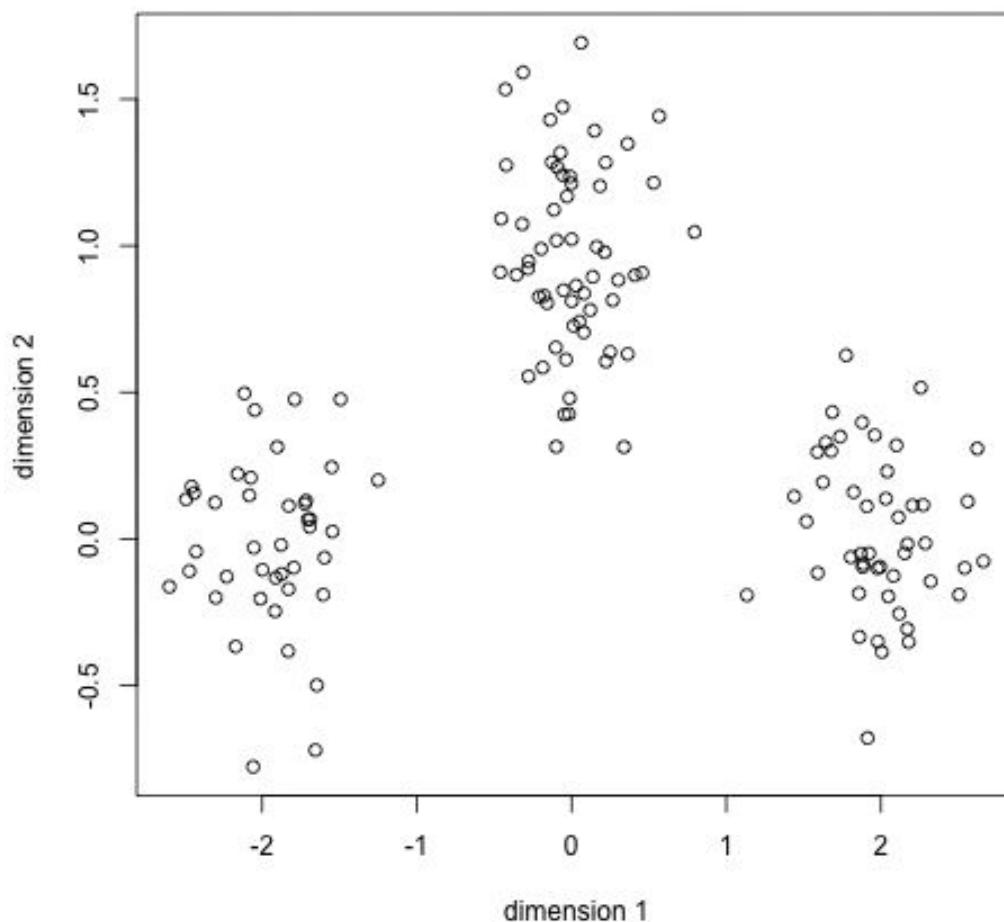
- Pensemos solo en particionar usando distancia, sin pensar en densidades ni distribuciones de probabilidad.



# K-means

- ❖ K medias es el algoritmo más usado para aglomerar datos.
  - K medias comienza por elegir k centros aleatorios.
  - Después, todos los puntos son asignados al centro más cercano basado en la distancia euclídea, lo cual genera una partición del espacio.
  - Luego los centros son re calculados usando la nueva partición y el ciclo comienza nuevamente.
  - Este proceso continúa hasta que no haya más cambios en la partición entre iteraciones.
- ❖ Este algoritmo genera una partición similar a la de la mezcla de Gaussianas Esféricas, esto es, con una matriz de varianza Covarianza múltiplo de la Identidad.

**step 0**



# K-means

## Problemas

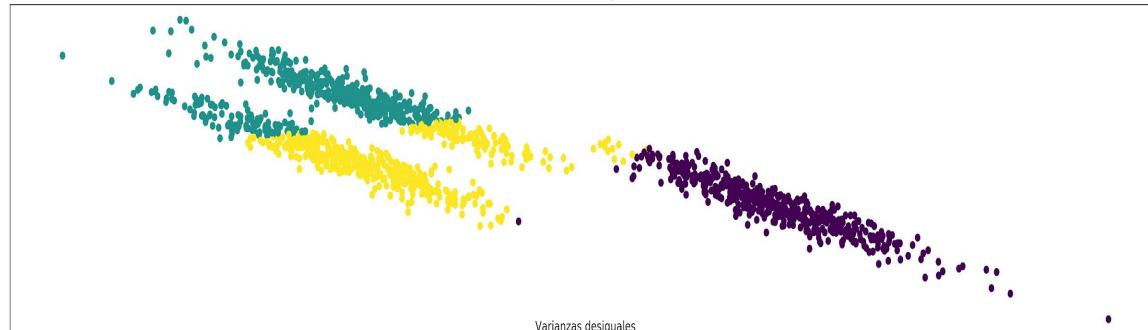
- Inestabilidad
- Mínimos locales (mucha sensibilidad a las semillas)
- Soluciones globales → sensibles a outliers
- El número de clusters k suele ser desconocido, Note\_fig5.ipynb

## Parámetros

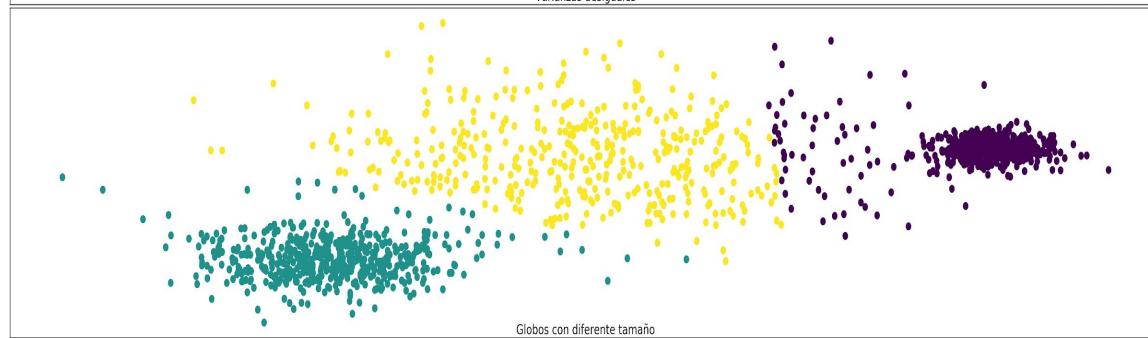
- Inicialización
- número de veces que se vuelven a tirar las semillas
- cuántas iteraciones hasta que termina la búsqueda

# K-means

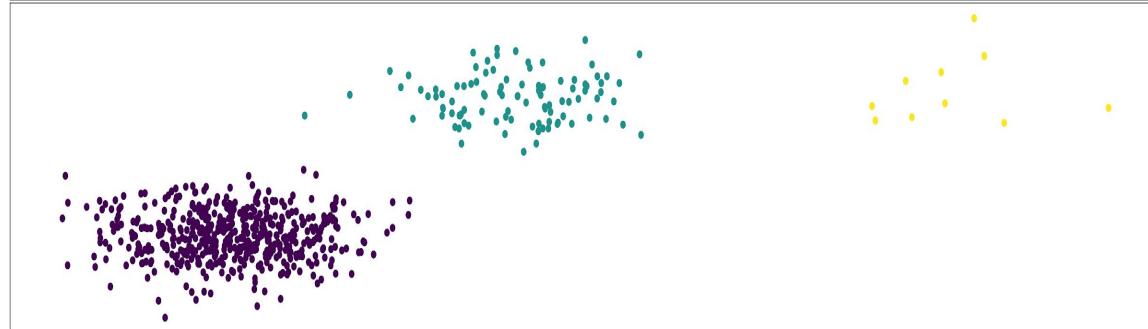
Globos Anisotropicos



Varianzas desiguales



Globos con diferente tamaño



# K-means: como definimos k?

- ❖ No podemos usar BIC, o MDL o AIC porque no usamos un modelo de verosimilitud para ajustar.
- ❖ Pero si podemos comparar entre diferentes modelos en función de k el valor de la inercia del modelo, esto es, la suma de distancias cuadradas dentro de cada cluster de la partición final.

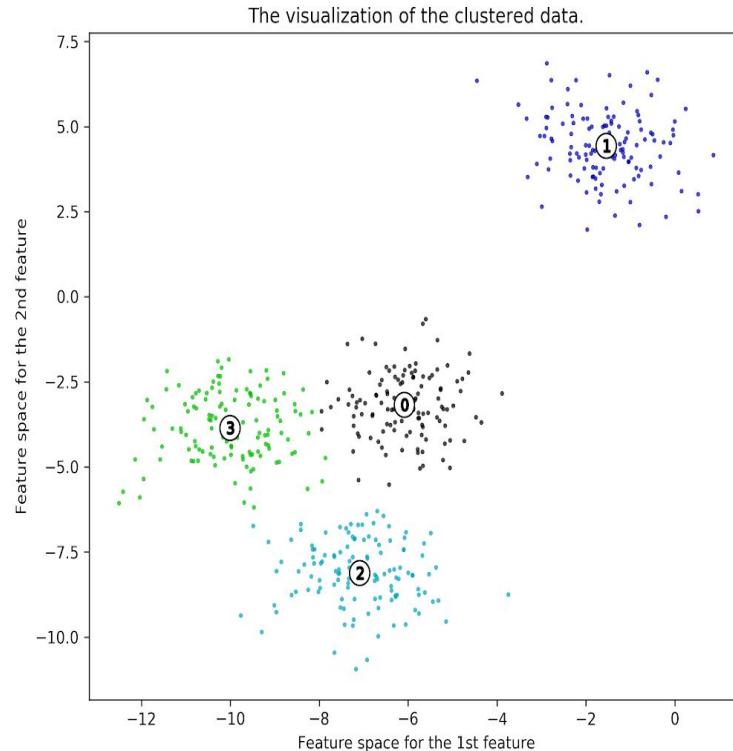
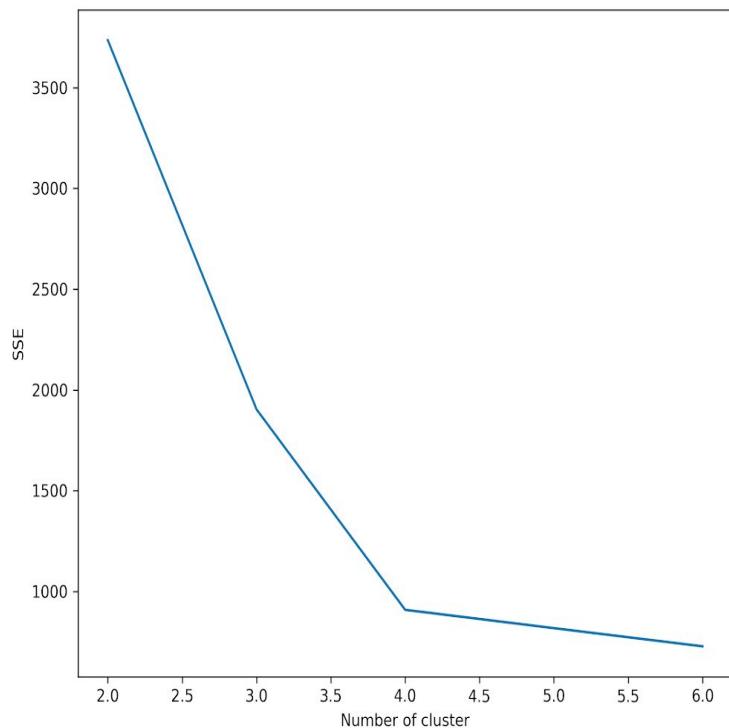
$$Inercia = SSE = \sum_{i=1}^K \sum_{x \in C_i} d^2(m_i, x)$$

# K-means: como definimos k?

- ❖ La inercia se considera una medida de cuán coherentes los clusters son,
- ❖ si se hace un gráfico de la inercia en función del k, se considera heurísticamente que el mejor valor se da cuando se desacelera la reducción de la inercia.
- ❖ La note\_fig6.ipynb muestra como elegir el k mas apropiado con el método del codo.

# K-means

Elbow method for KMeans clustering on sample data



# K-means: Análisis de siluetas

- ## Para cada punto  $i \in C_k$ , se crean los indices  $a(i)$  de similaridad promedio y  $b(i)$  de disimilaridad minima promedio

$$a(i) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, i \neq j} d(i, j) \quad b(i) = \min_{k \neq l} \frac{1}{|C_l|} \sum_{j \in C_l} d(i, j)$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

# K-means: Análisis de siluetas

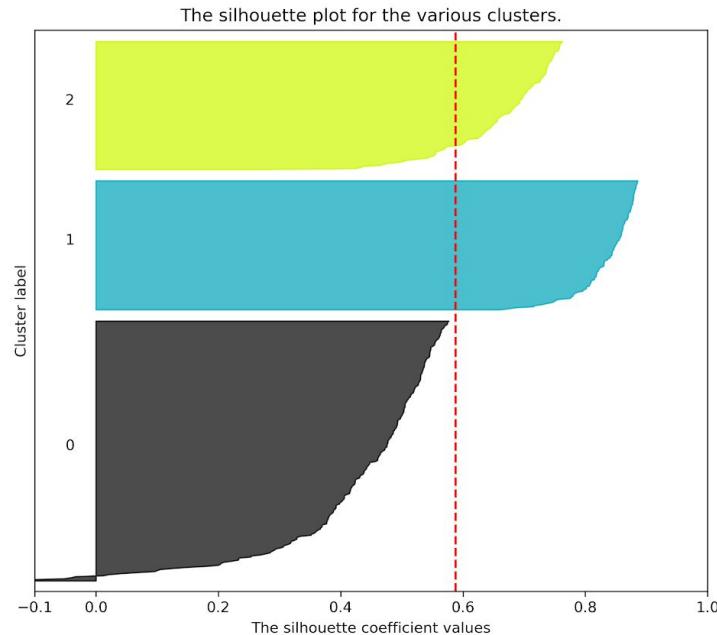
- ❖ La medida  $s(i)$  tiene valor entre -1 y 1.
- ❖
- ❖ Los coeficientes  $s(i)$  cercanos a +1 indican que la muestra está lejos de los clusters vecinos.
- ❖
- ❖ El valor 0 indica que la muestra está muy cerca del borde de decisión entre los clusters.
- ❖
- ❖ Un valor negativo indica que esos puntos deben haber sido asignados al cluster equivocado.

# K-means: Ejemplo

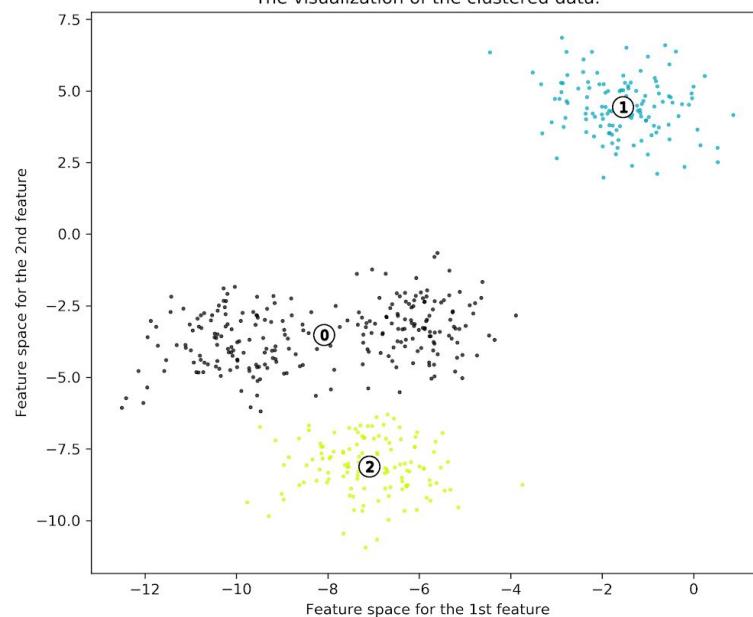
- ❖ Se simula un grupo de datos con cuatro gaussianas.
- ❖ Se calcula el gráfico de silueta para particiones de k medias con K=2,3,4,5,y 6.
- ❖ El gráfico de silueta para los valores de k =3, 5 and 6 muestran que esos k son una mala elección, dado que hay clusters por debajo del valor de silueta promedio y clusters con valores negativos.

# K-means: Ejemplo

Silhouette analysis for KMeans clustering on sample data with n\_clusters = 3

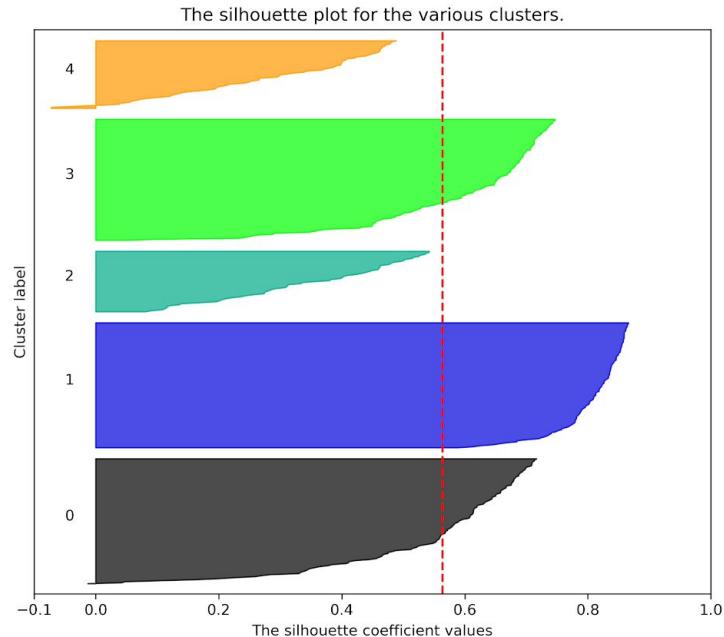


The visualization of the clustered data.

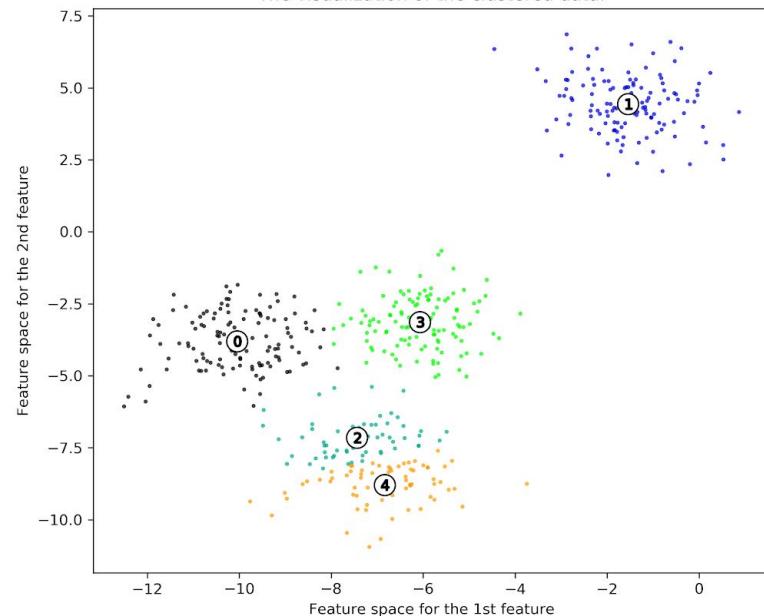


# K-means: Ejemplo

Silhouette analysis for KMeans clustering on sample data with `n_clusters = 5`

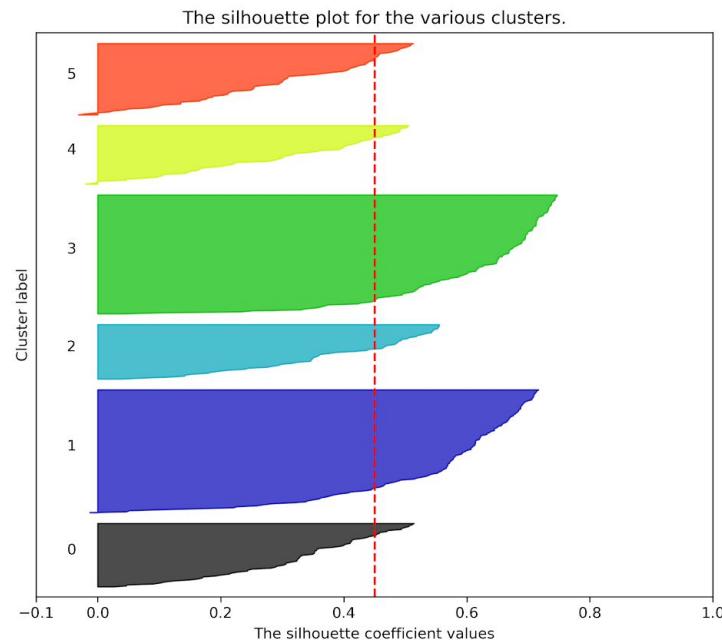


The visualization of the clustered data.

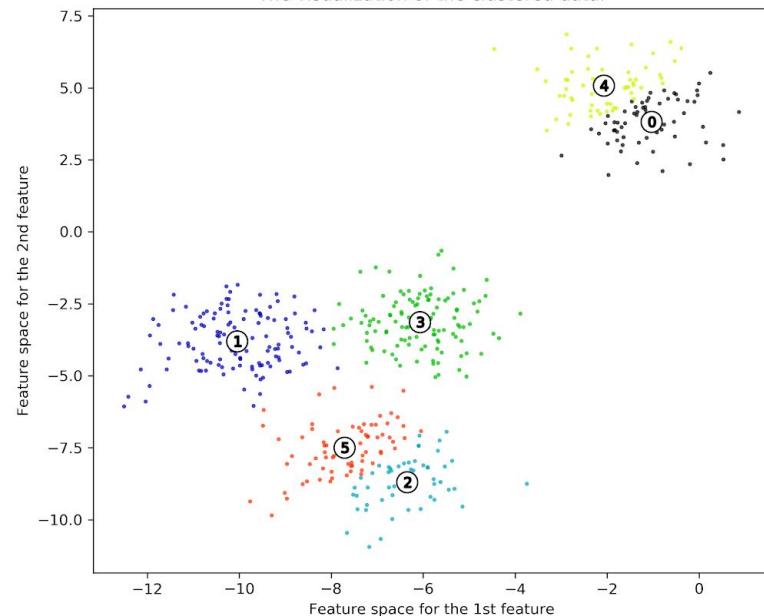


# K-means: Ejemplo

Silhouette analysis for KMeans clustering on sample data with `n_clusters = 6`



The visualization of the clustered data.

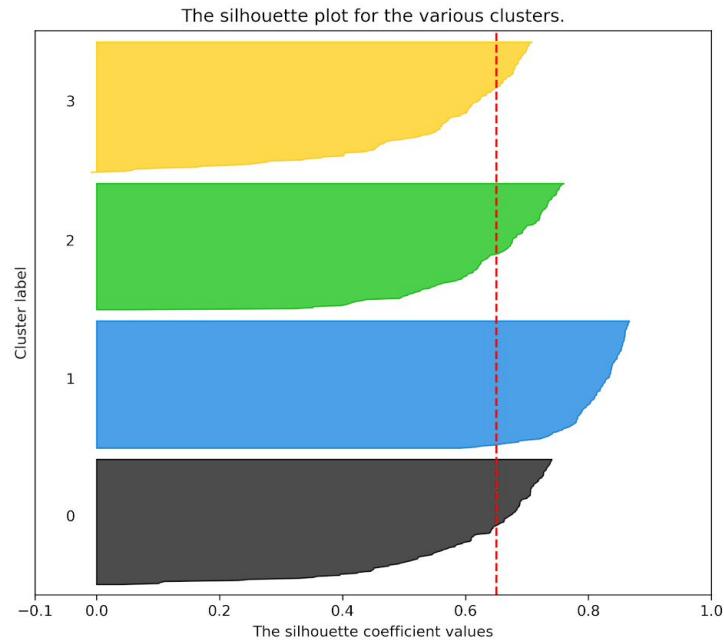


# K-means: Ejemplo

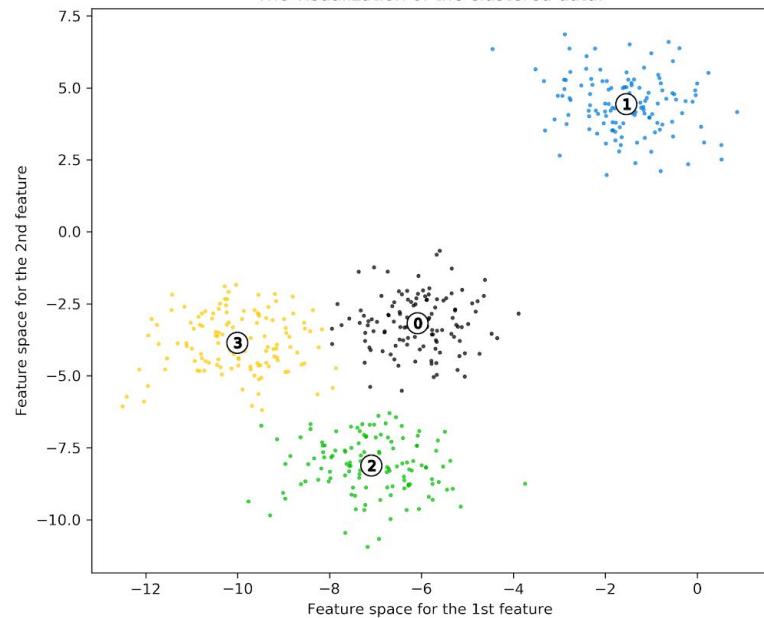
- ❖ Estas características no ocurren en el caso de  $k=2$  y  $4$
- ❖ Si se estudia el grosor de gráfico de silueta se ve que  $k=2$  produce una partición muy desbalanceada, dado que uno de los clusters absorbe tres clusters diferentes.
- ❖ Cuando  $k=4$  las siluetas están balanceadas, por lo cual este es el mejor  $k$

# K-means: Ejemplo

Silhouette analysis for KMeans clustering on sample data with `n_clusters = 4`

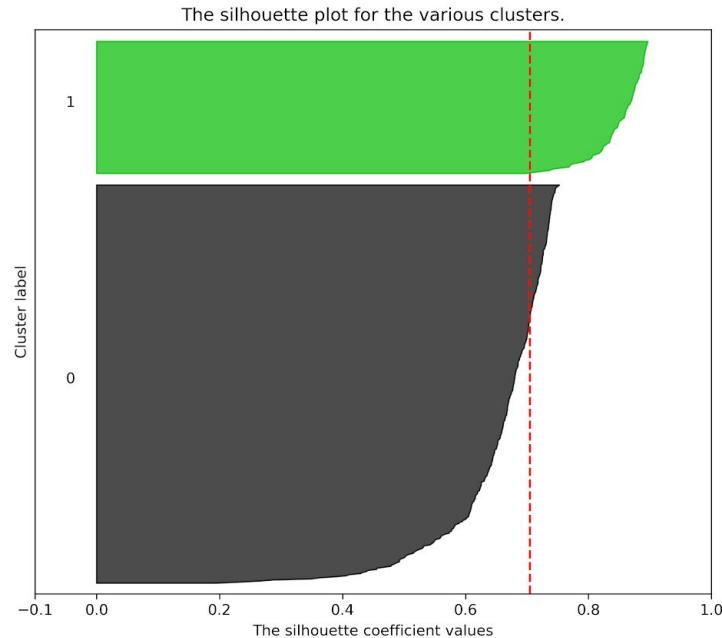


The visualization of the clustered data.

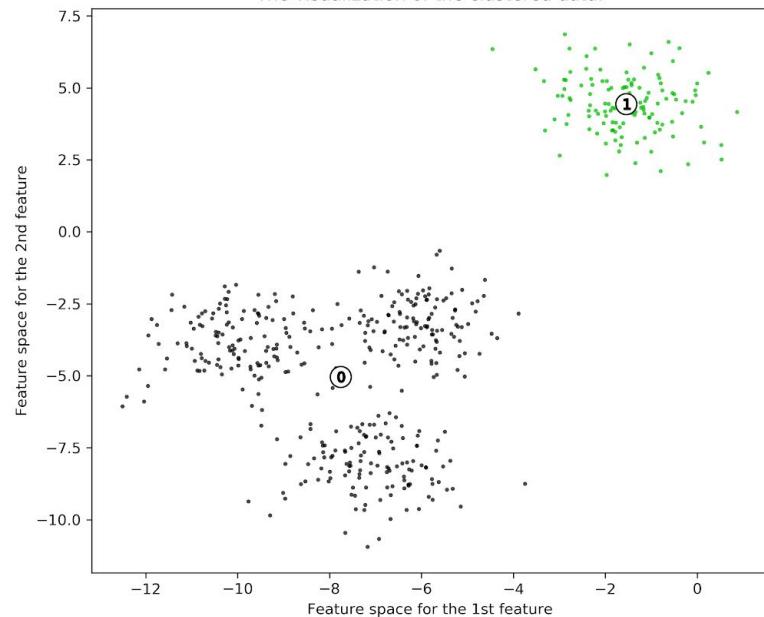


# K-means: Ejemplo

Silhouette analysis for KMeans clustering on sample data with `n_clusters = 2`



The visualization of the clustered data.



# DbSCAN

- ❖ El algoritmo DBSCAN define clusters como áreas de alta densidad separadas por áreas de baja densidad, que pueden ser de cualquier forma y tamaño.
- ❖ El concepto central del algoritmo es la clasificación y detección de puntos de núcleo (core samples), que son los puntos ubicados en zonas de alta densidad.
- ❖ Un cluster es un conjuntos de puntos de núcleo cercanos unos con otros, junto a un conjunto de puntos no núcleo que están cercanos a algún punto de núcleo.
- ❖ Hay dos parámetros en este algoritmo: min\_samples y eps, donde min\_samples son los datos mínimos requeridos en un entorno de radio eps, los cuales definen la noción de zona densa.

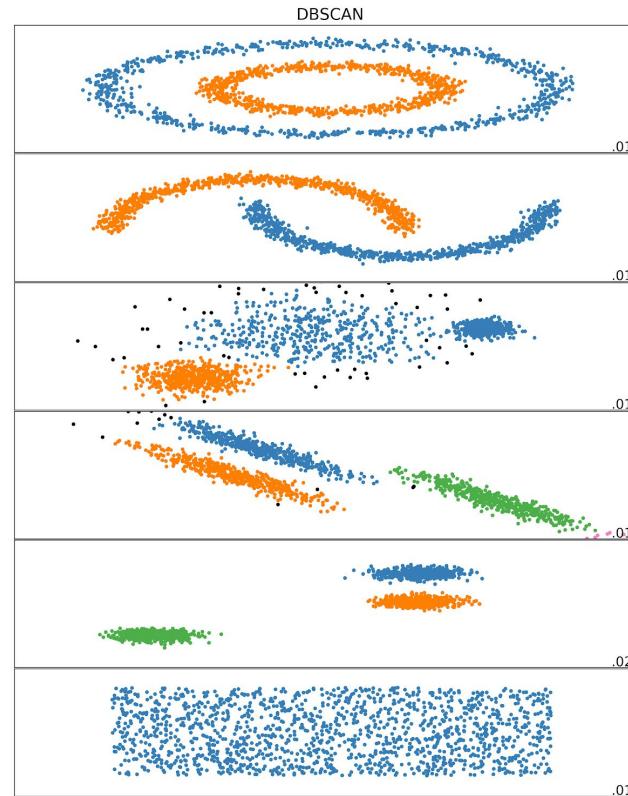
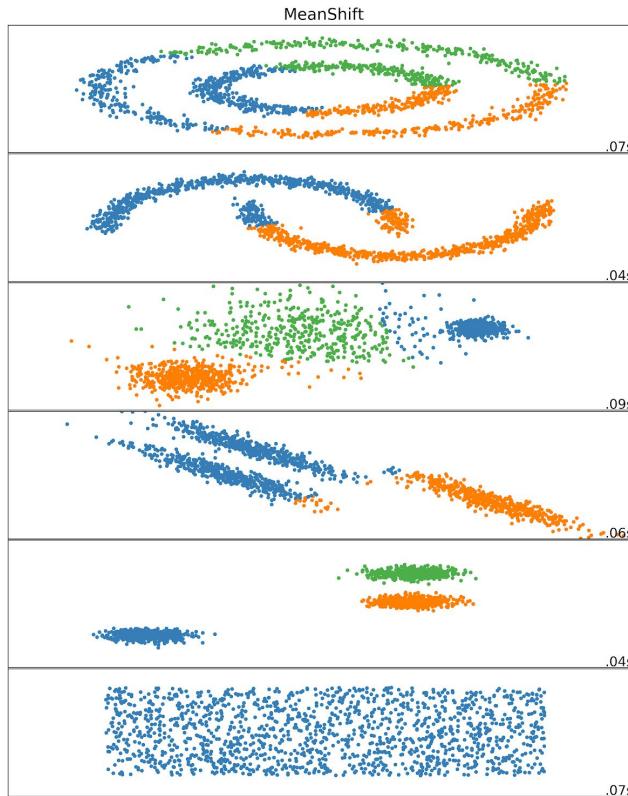
# Dbscan

- ❖ El algoritmo empieza con una muestra aleatoria y encuentra todos los puntos en el entorno de radio  $\text{eps}$ . Si el número de puntos es mayor a  $\text{min\_number}$  se etiqueta ese punto como un punto de núcleo, si no es un punto outlier.
- ❖
- ❖ Todos los puntos del entorno se etiquetan como puntos no núcleo de cluster. Se realiza el mismo procedimiento para cada uno de ellos, cambiando a punto core su etiqueta y agregando nuevos puntos, o marcando outliers.
- ❖
- ❖ Si no hay más puntos en un entorno  $\text{eps}$  de cada punto del cluster, se salta a otro punto aleatoriamente y se continúa hasta que todo punto es bien un punto de cluster o un outlier.

# Dbscan



# Mean shift Dbscan note\_fig8.ipynb



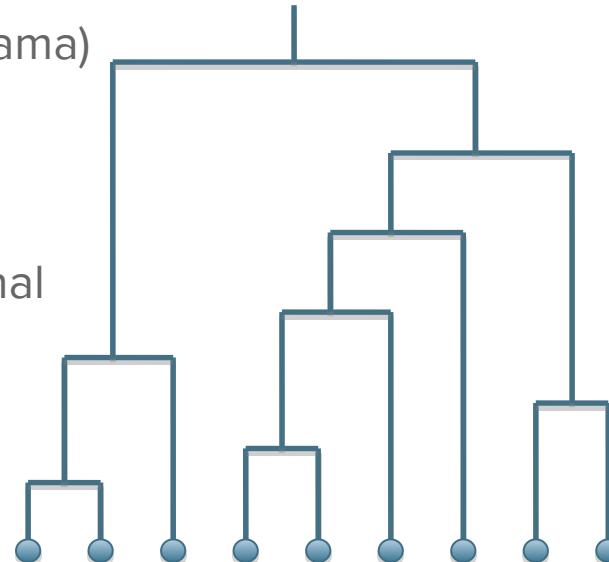
# Clustering jerárquico

Si no queremos especificar k...

Algoritmos jerárquicos que generan una

taxonomía jerárquica de clusters (dendrograma)

- Interpretación más rica
- Más difícil de interpretar
- El corte del árbol tiene que ser ortogonal



# Clustering jerárquico aglomerativo

Bottom-up

- Cada objeto es su propio cluster
- Se unen en un solo cluster el par de clusters más semejantes
- La historia de uniones forma un árbol binario (jerarquía)

# Semejanza entre clusters

## Single-link

1. Para cada par de clusters A y B, el par de objetos a, b más cercanos tal que a pertenece a A y b pertenece a B
2. Se unen los clusters con el par de objetos más semejante

## Complete-link

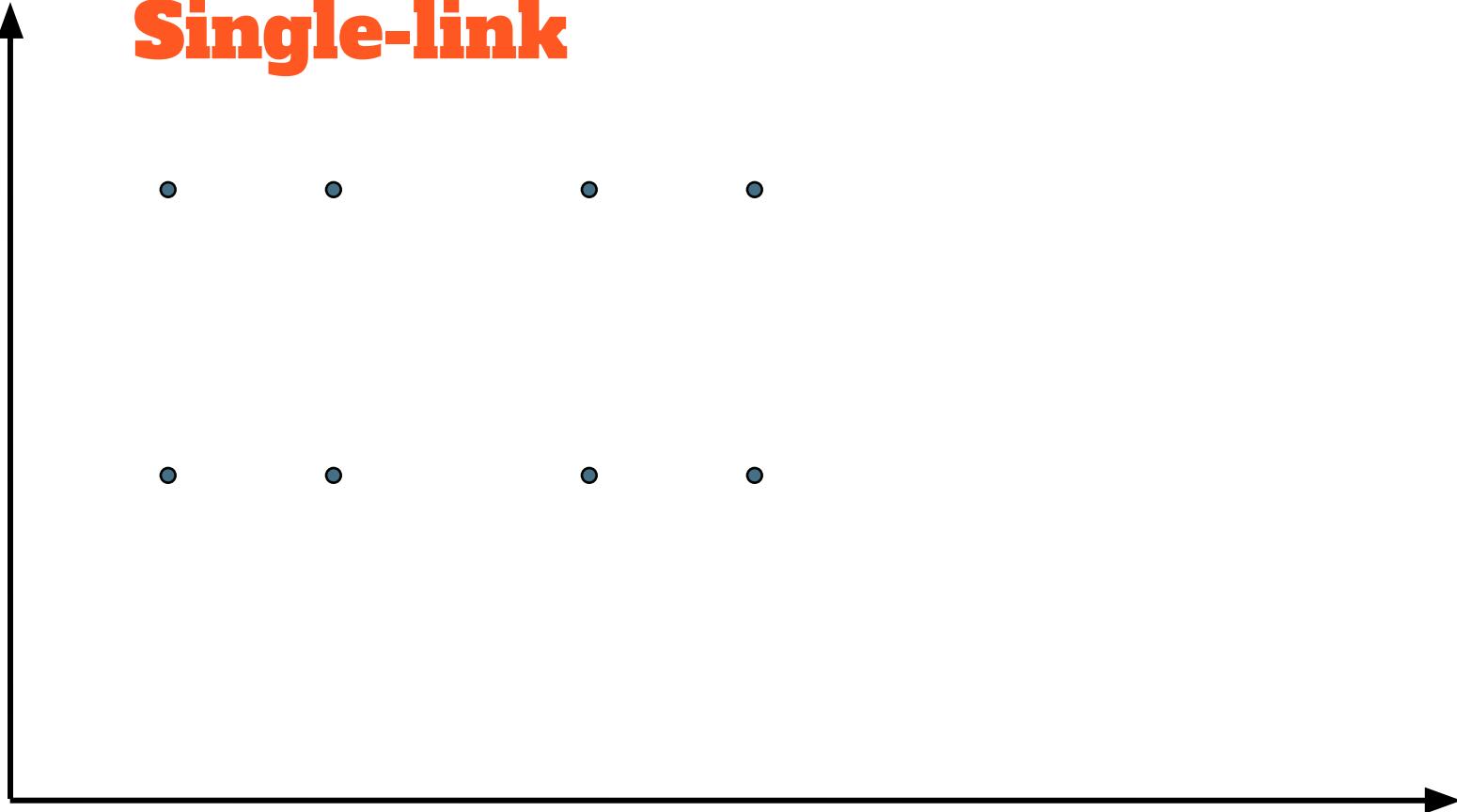
1. Para cada par de clusters A y B, el par de objetos a, b más distantes tal que a pertenece a A y b pertenece a B
2. Se unen los clusters con el par de objetos más semejante

## Average-link

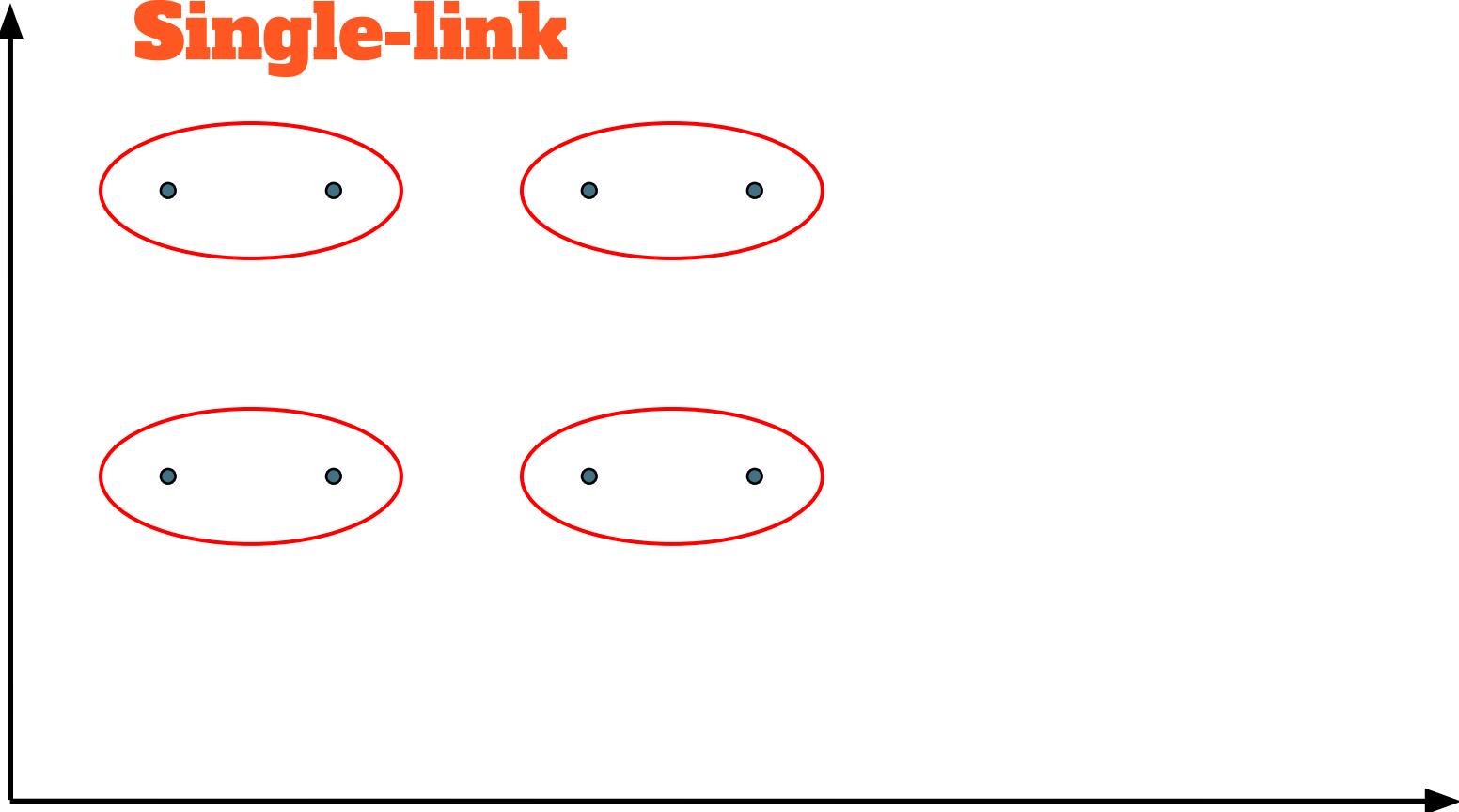
1. Para cada par de clusters A y B, se calcula la distancia entre todo par de objetos a, b tal que a pertenece a A y b pertenece a B
2. Se unen los clusters con el promedio de distancia más bajo

**Centroid:** Se unen los clusters con los centroides más cercanos

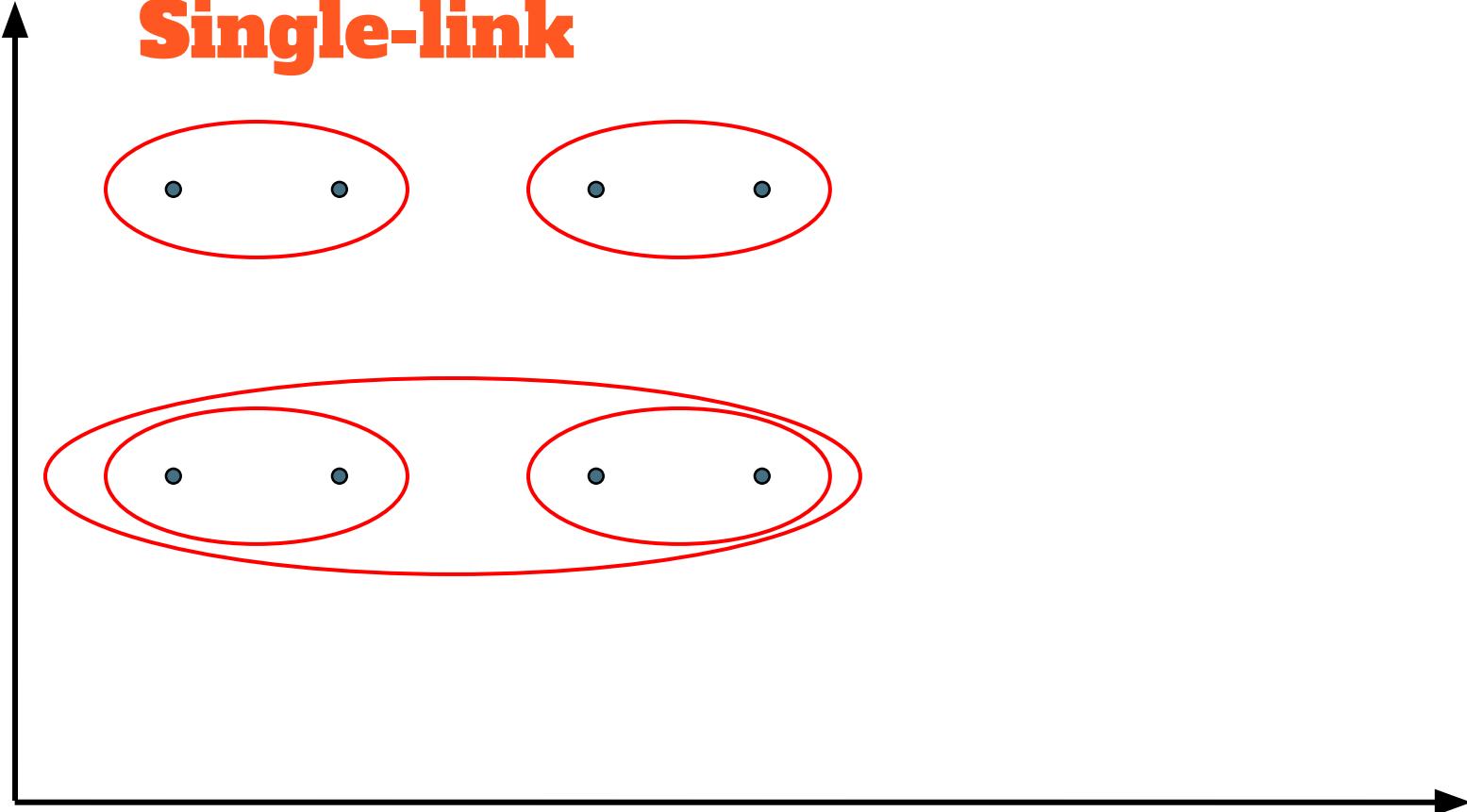
# Single-link



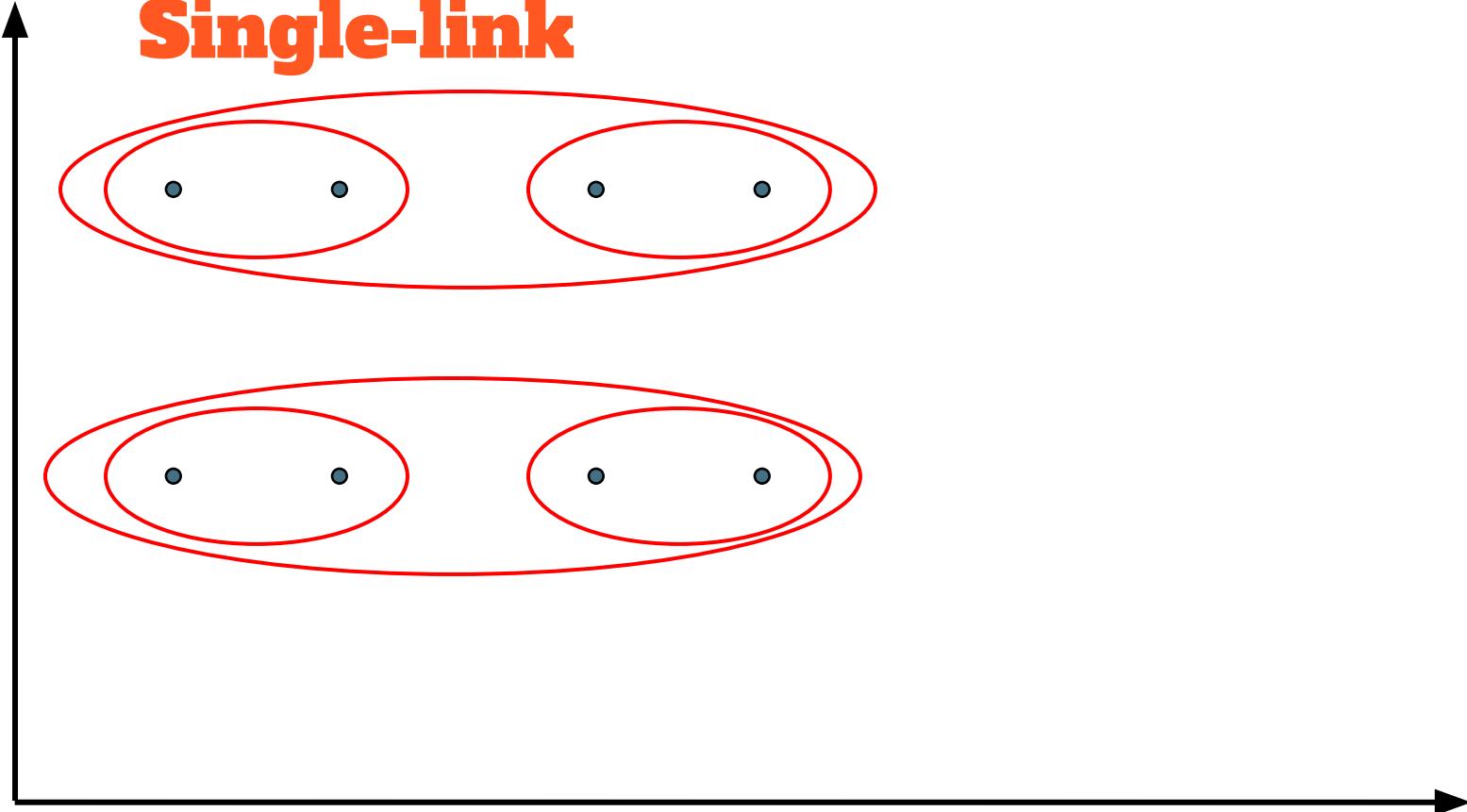
# Single-link



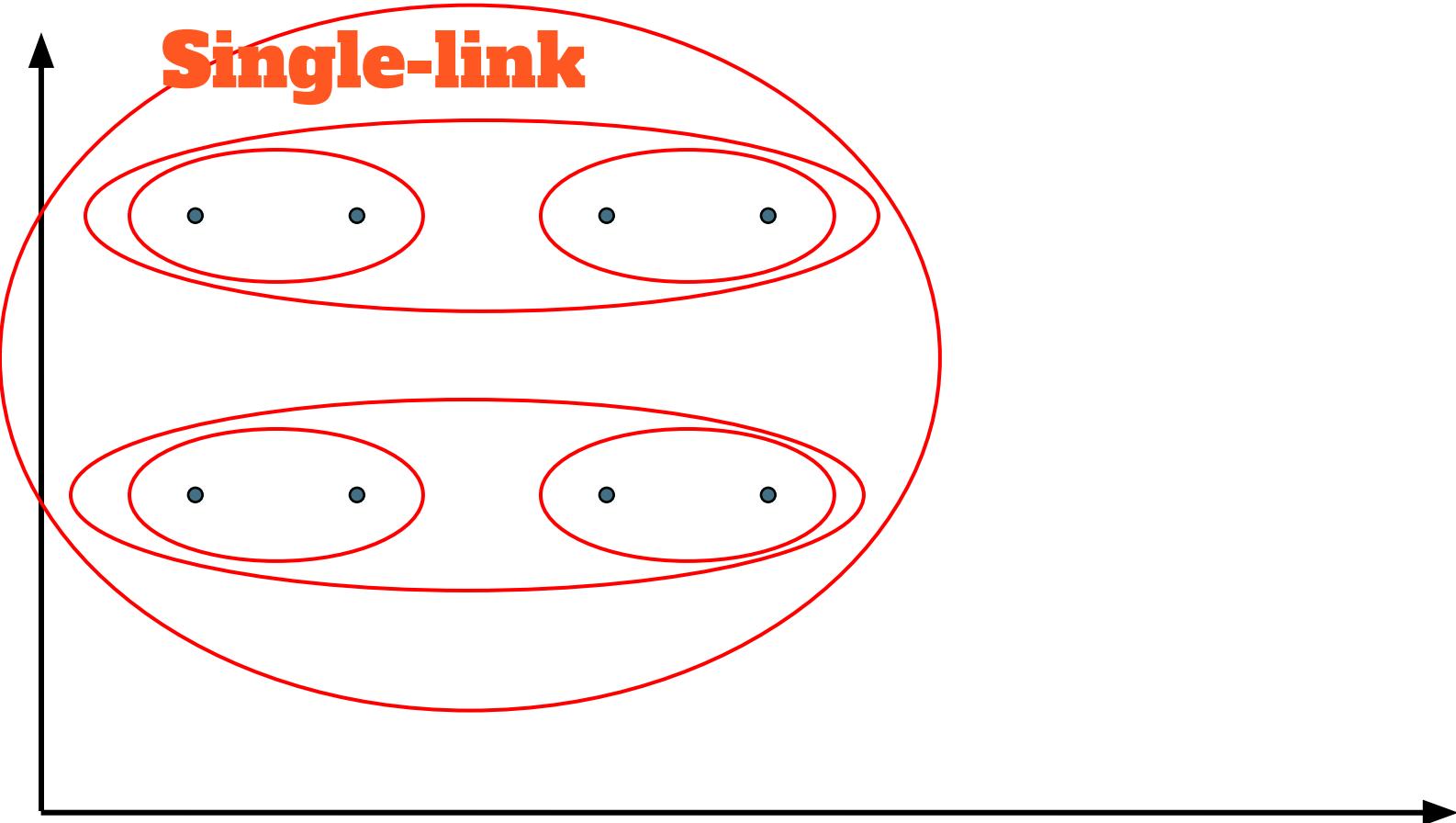
# Single-link



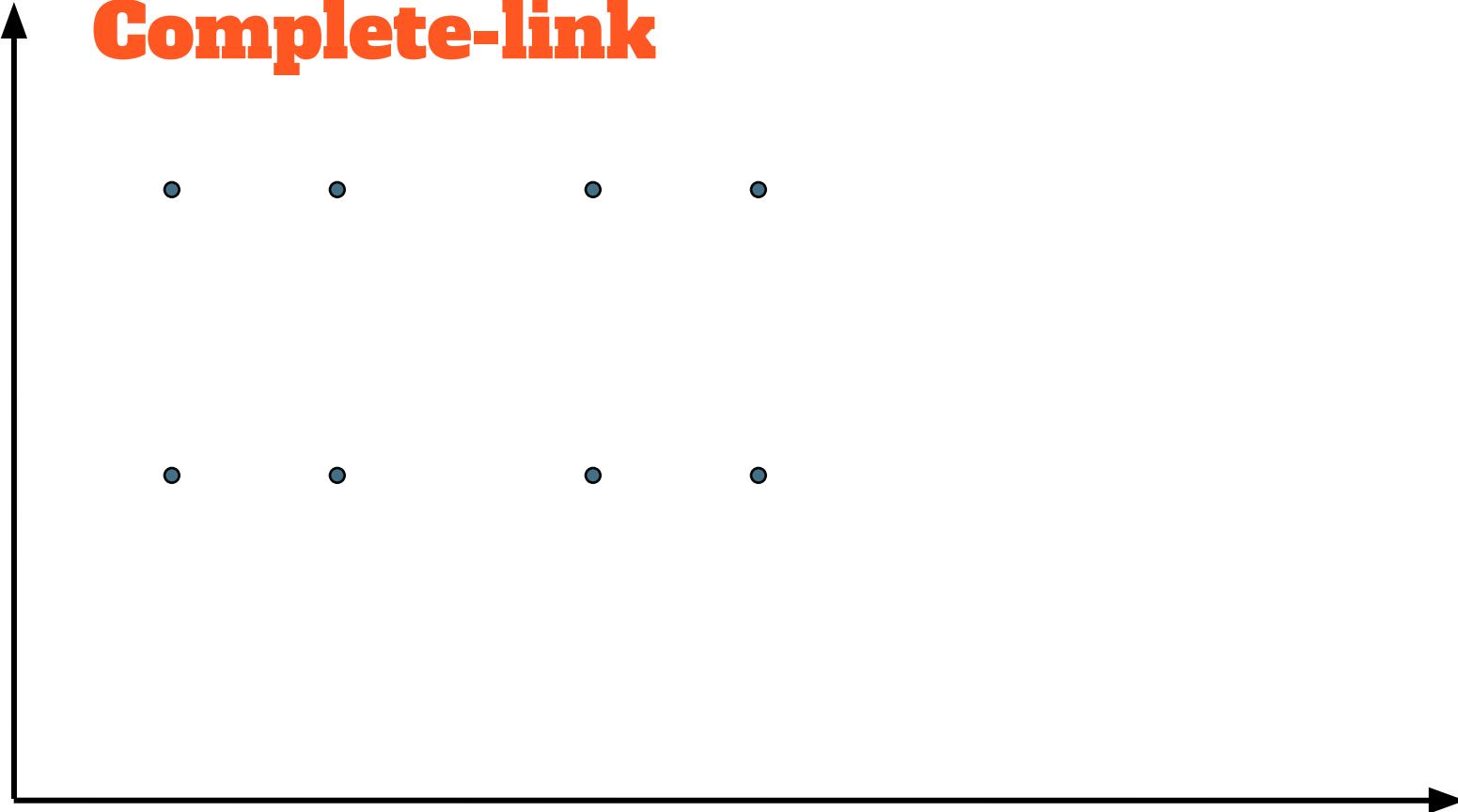
# Single-link



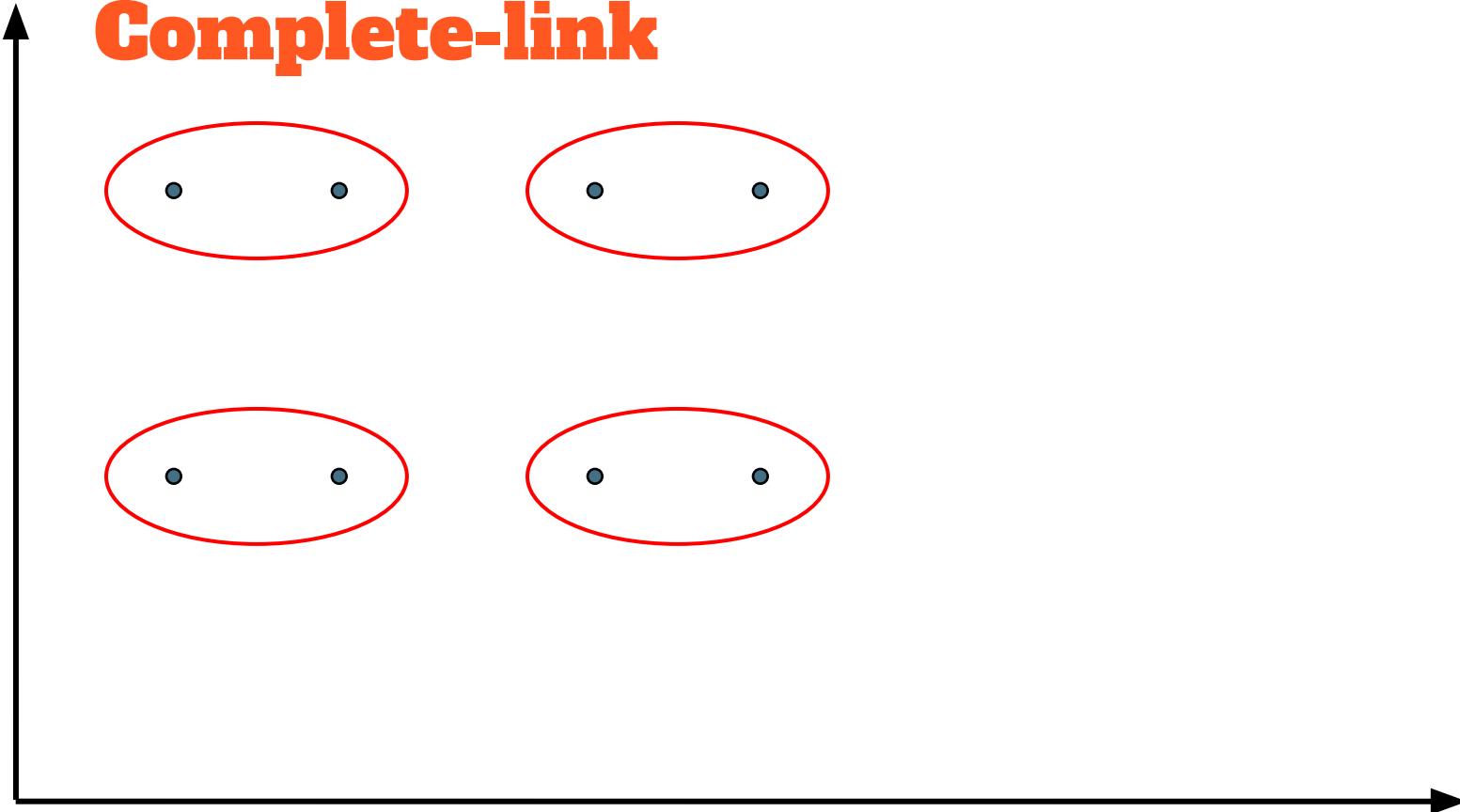
# Single-link



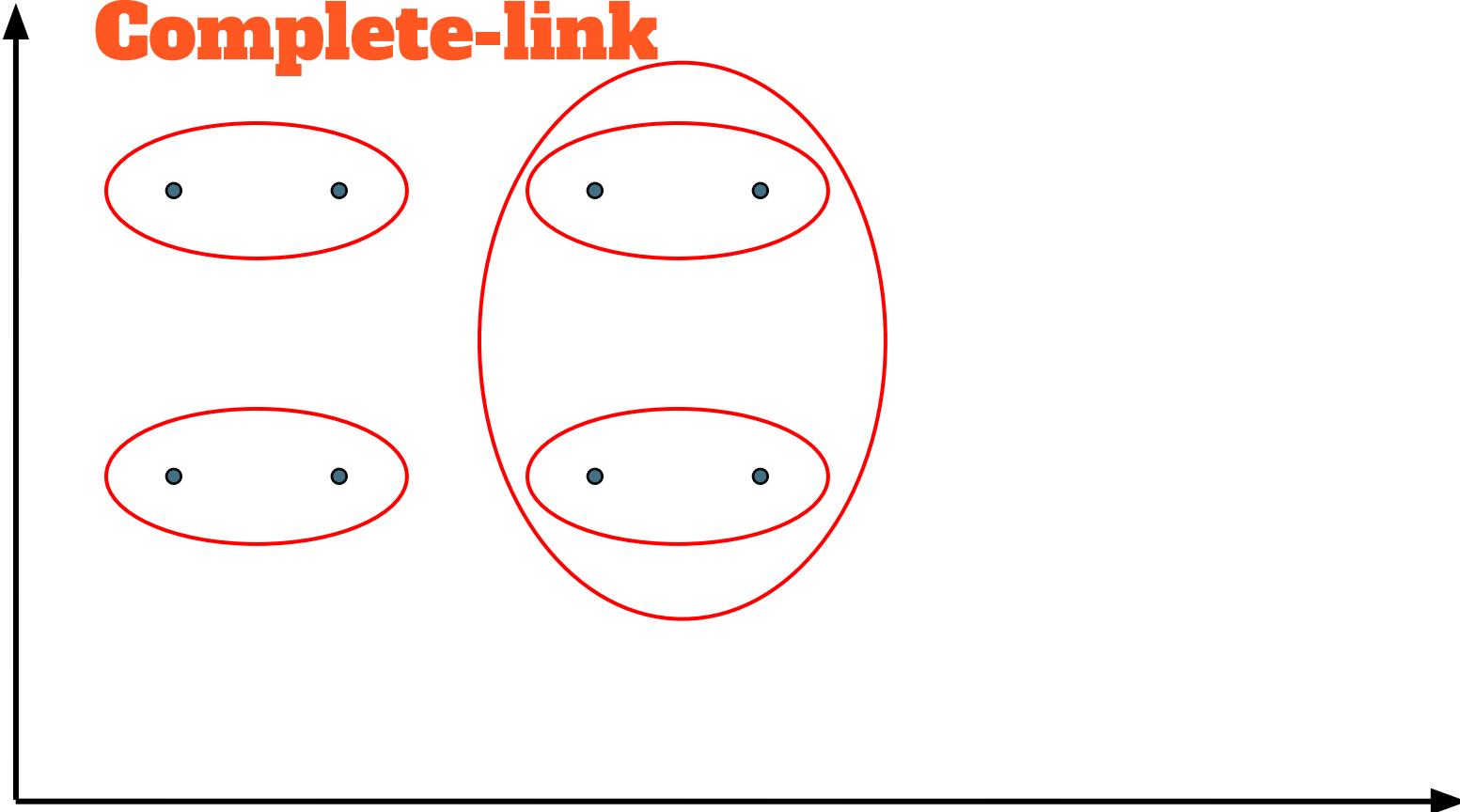
# Complete-link



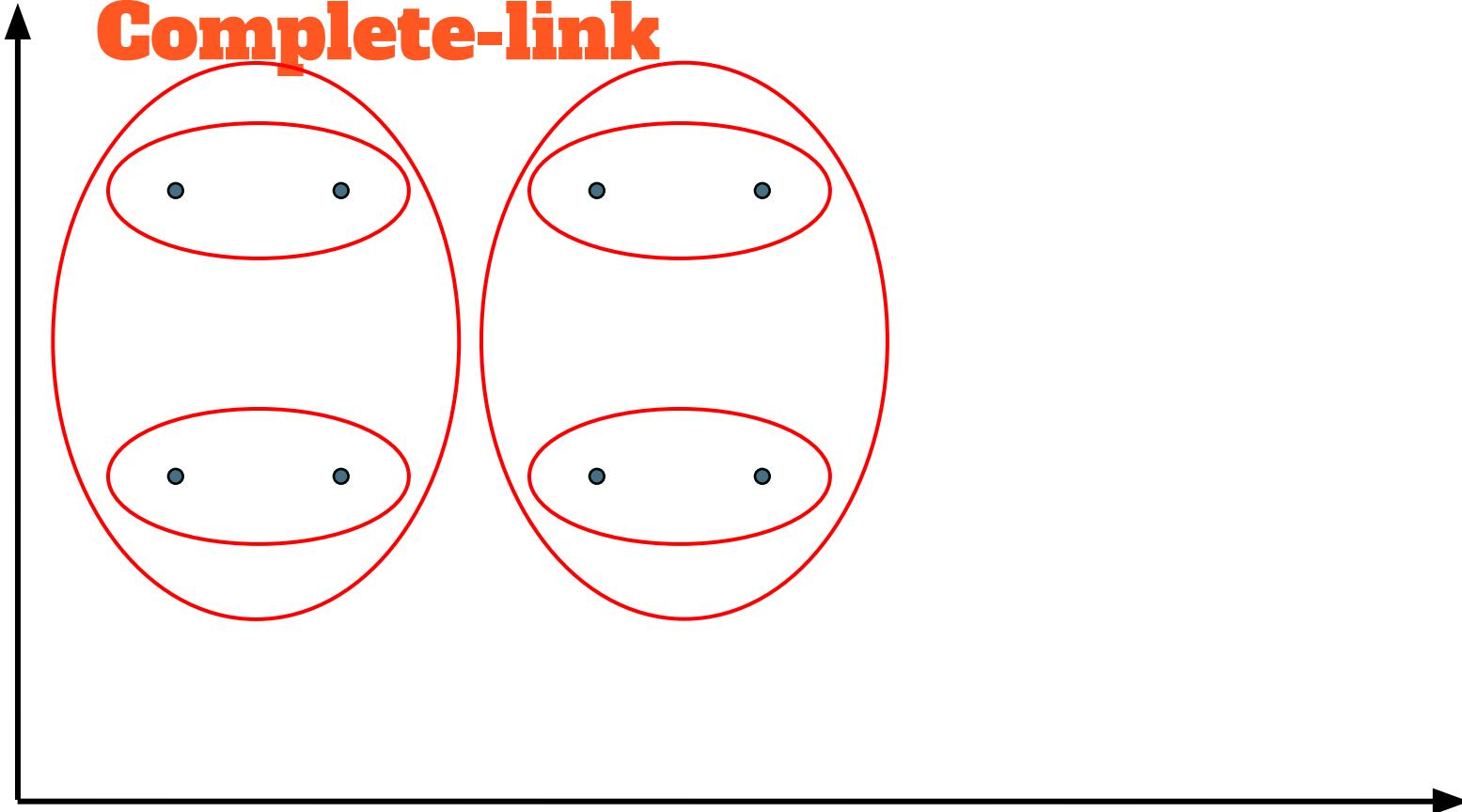
# Complete-link



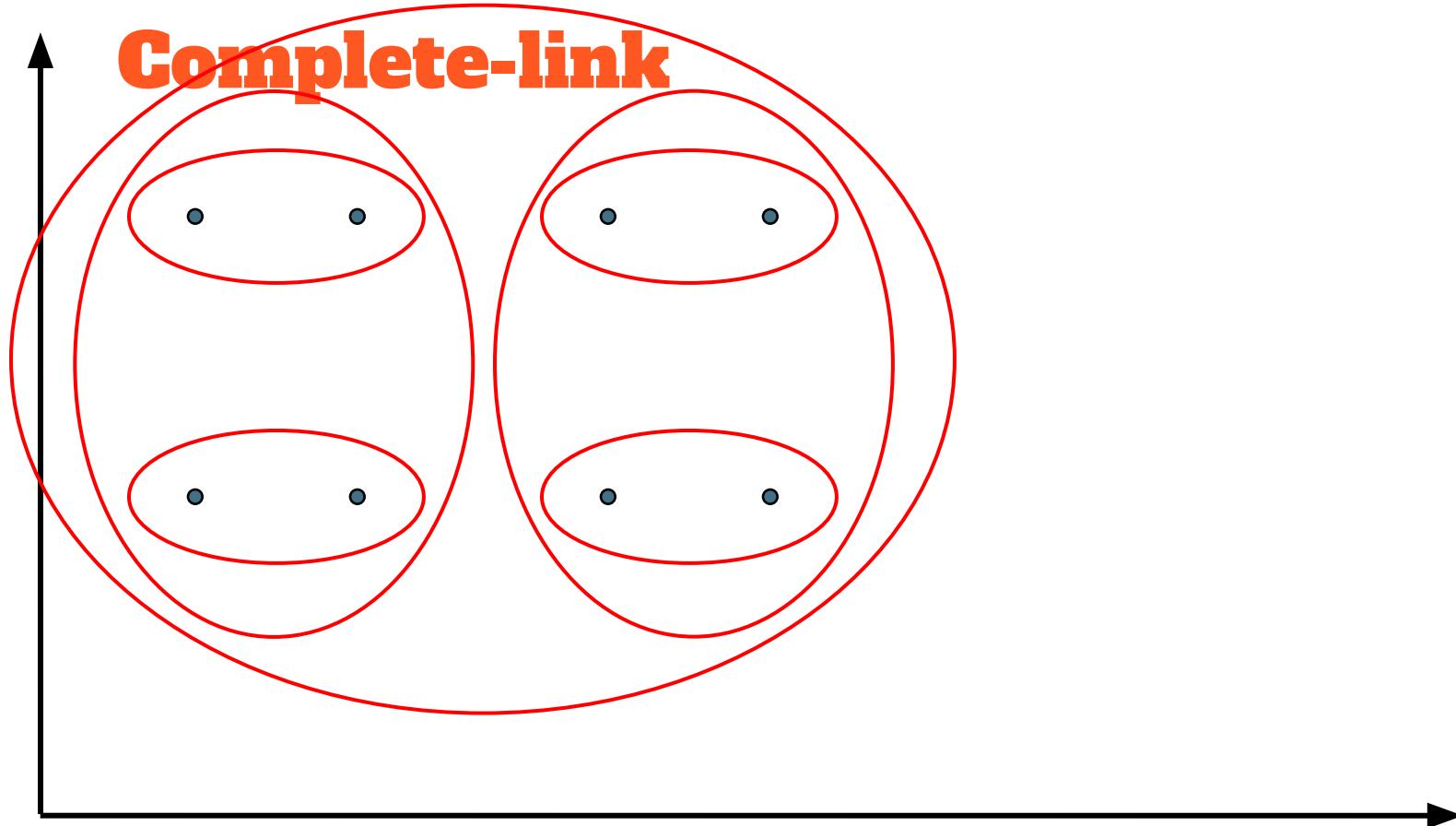
# Complete-link



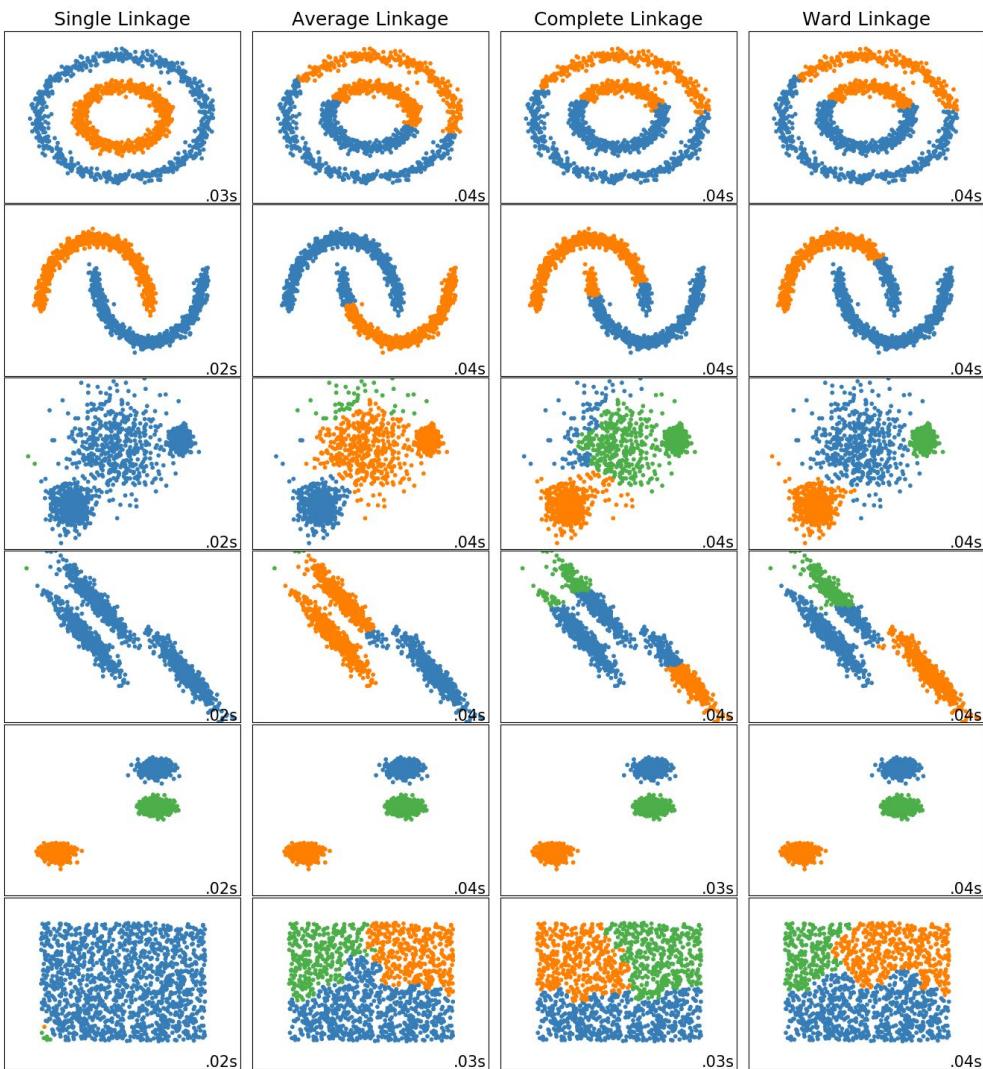
# Complete-link



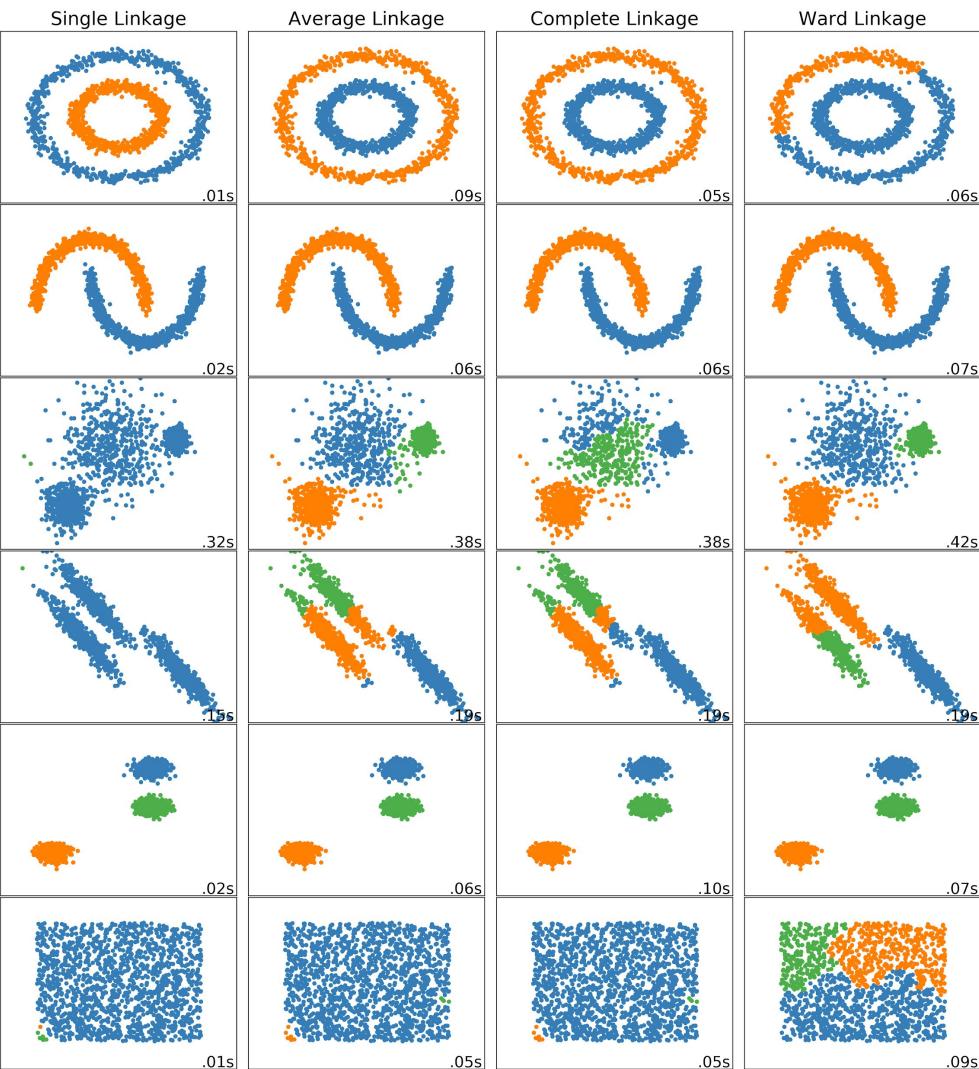
# Complete-link



# Clustering jerárquico Aglomerativo note\_fig9.ipynb



# Clustering jerárquico Aglomerativo con restricción de conectividad



# Evaluación

Un experto de dominio **interpreta** los clusters y encuentra información valiosa

¿Cómo mostrar el contenido de los clusters?

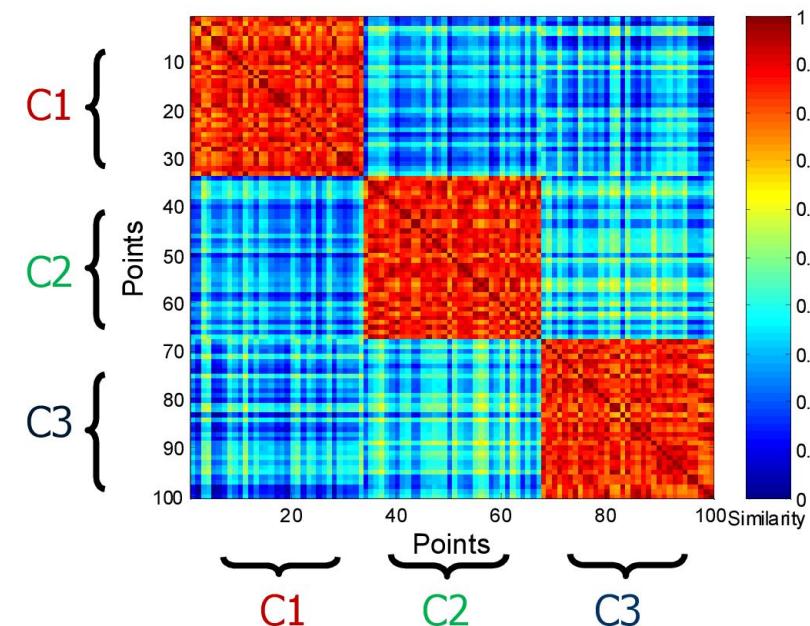
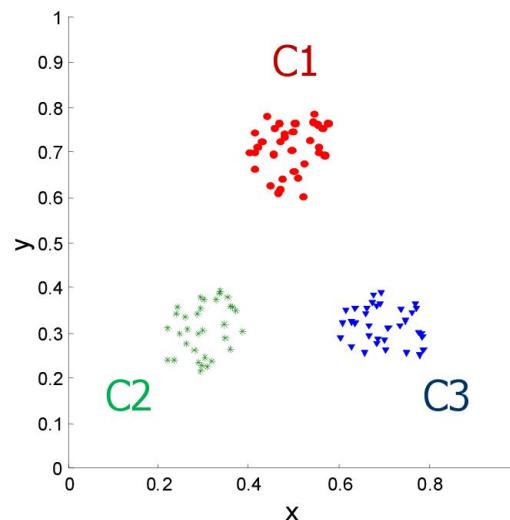
- Centroides (medoides)
- Resumen de características
- Características más distintivas de cada cluster
- Aplicar un algoritmo de aprendizaje automático interpretable (Decision Tree)

# Evaluación intrínseca

- ❖ Coeficiente Silhouette
  - Mide la semejanza de cada objeto al cluster al que se asigna (cohesión), comparada con otros clusters (separación).
  - Si el valor es bajo o negativo, el número de clusters puede ser inadecuado.
  -
- ❖ Gráfico de codo
  - Se observa la inercia en función del número de clusters.

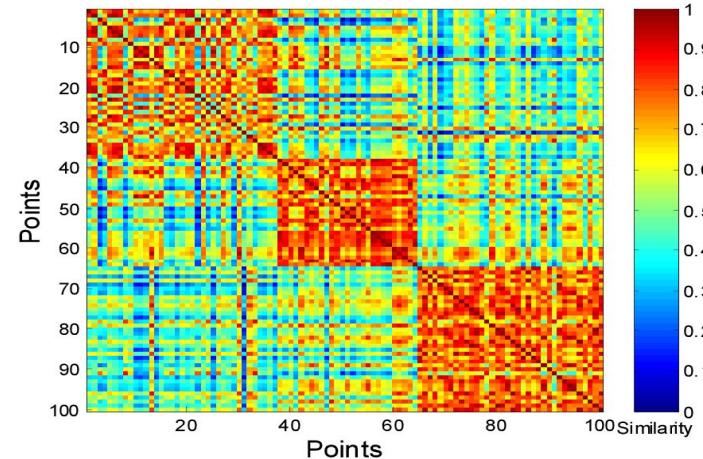
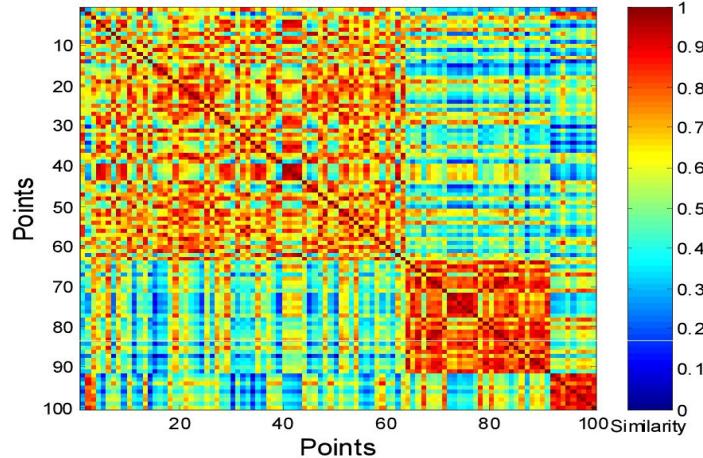
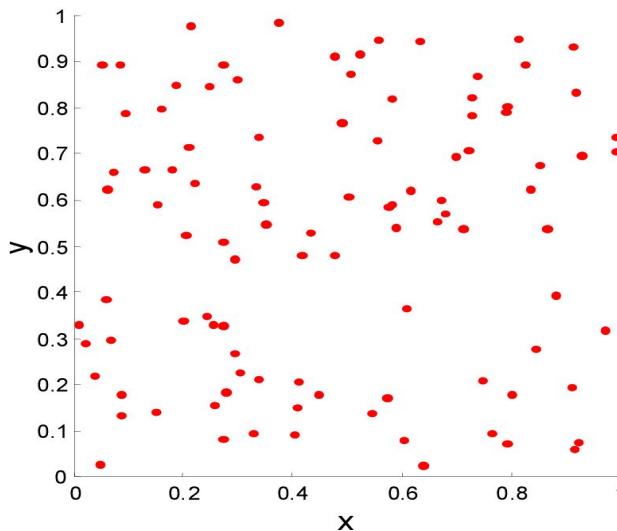
# Matriz de similitud

Ordenamos los datos en la matriz de similitud con respecto a los clusters en los que quedan los datos e inspeccionamos visualmente...



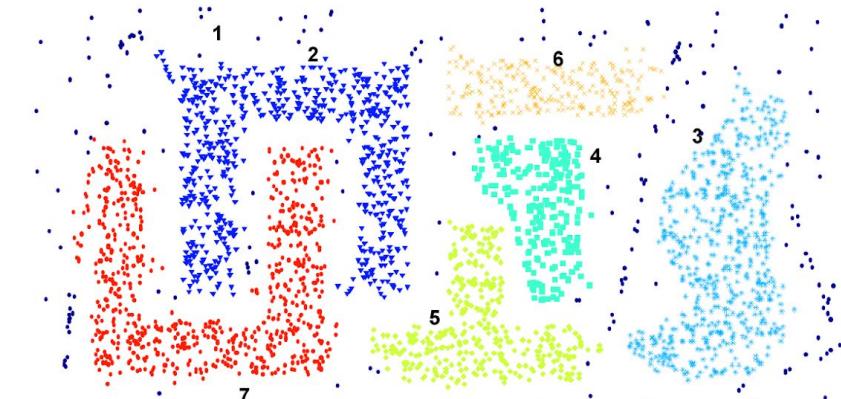
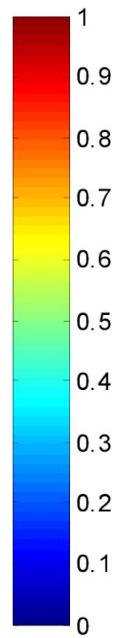
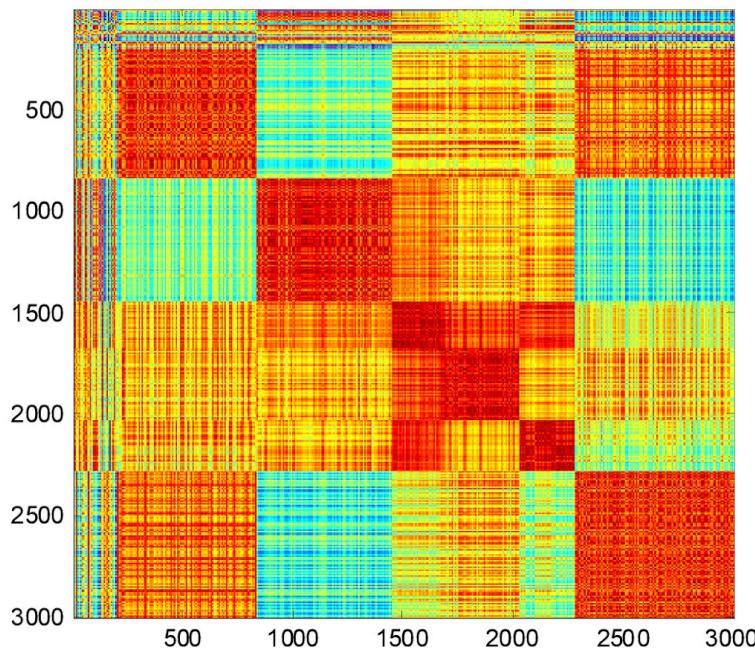
# Problema

Incluso en datos aleatorios,  
si nos empeñamos,  
encontramos clusters:  
DBSCAN (arriba) y  
k-Means (abajo)



# Matriz de similitud

DBSCAN



# Evaluación con clases

Si los objetos tienen alguna etiqueta, observamos su distribución en los clusters, estamos cerca de el problema semi supervisado!!

- ❖ Homogeneidad: cada cluster contiene sólo miembros de una clase
- ❖ Completitud: todos los miembros de una clase están en el mismo cluster
- ❖ V-measure: media armónica de los anteriores
- ❖ Adjusted Rand index: semejanza entre las etiquetas originales y las asignadas
- ❖ Información Mutua entre etiquetas originales y asignadas

# Evaluación con clases: [note\\_fig10.ipynb](#)

Estimated number of clusters: 3

Estimated number of noise points: 18

Homogeneity: 0.953

Completeness: 0.883

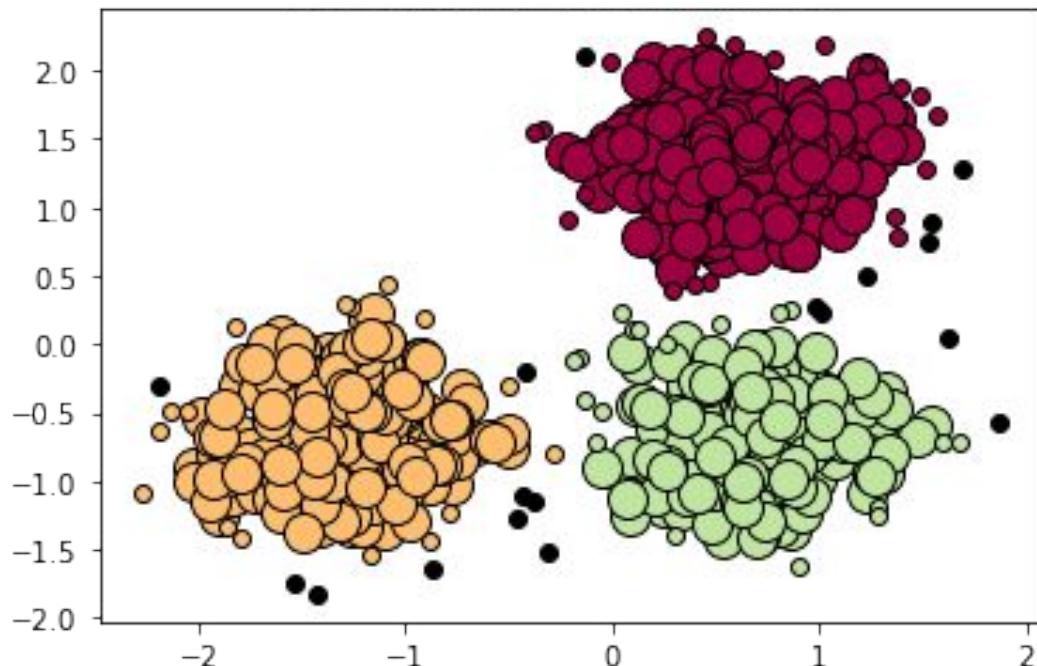
V-measure: 0.917

Adjusted Rand Index: 0.952

Adjusted Mutual Information: 0.916

Silhouette Coefficient: 0.626

Estimated number of clusters: 3



# Evaluación con testigos

1. Se seleccionan aleatoriamente pares de objetos del dataset
2. Un experto del dominio decide si tienen que estar en el mismo cluster o en diferentes clusters
3. Observamos el grado de acuerdo entre cada solución y los testigos
4. Se seleccionan aleatoriamente objetos del dataset
5. Se los etiqueta
6. Se observa cómo se distribuyen en el dataset

# Indicadores de malas soluciones

En general, las malas soluciones se deben a malas características

- Una clase muy grande y el resto mucho más chicas → la mayoría de objetos son no diferenciables con esas características o distancia
- Clases con uno o pocos elementos → el número de clases es demasiado grande para el dataset
- Clusters con las mismas características, poco distinguibles
- Soluciones muy diferentes con diferentes inicializaciones, número de clusters

# **Clustering no es clasificación**

No vamos a obtener clases bien diferenciadas, sino más bien mucho ruido

Es fuertemente sensible a las características de los objetos, a los parámetros, a los outliers

La mayor parte de aproximaciones son muy inestables

La primera aproximación suele ser inservible, hay que refinar características e iterar

# Aplicaciones

- Segmentación de clientes, usuarios... para marketing personalizado
- Encontrar temas → topic detection
- Imágenes de los mismos objetos → gatitos, tumores (imágenes médicas), tormentas (imágenes satelitales), plagas (imágenes de cultivos)
- Agrupamiento de productos
- Detección de anomalías
- Taxonomías de plantas y otros organismos
- Detección de clases con significados semejantes