Introducción a la ciencia de datos

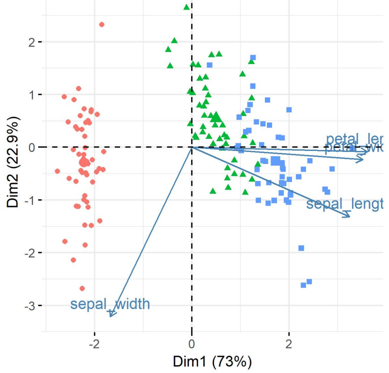
Clustering

Aprendizaje no supervisado

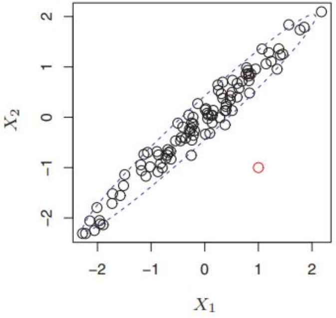
- Reglas de asociación

$$educ > 10 \wedge estado = Soltero$$
$$\Rightarrow ingreso > \$30.000$$

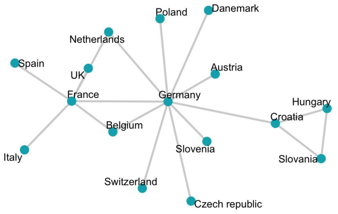
- Reducción de la dimensionalidad



- Detección de anomalías



- Grafos



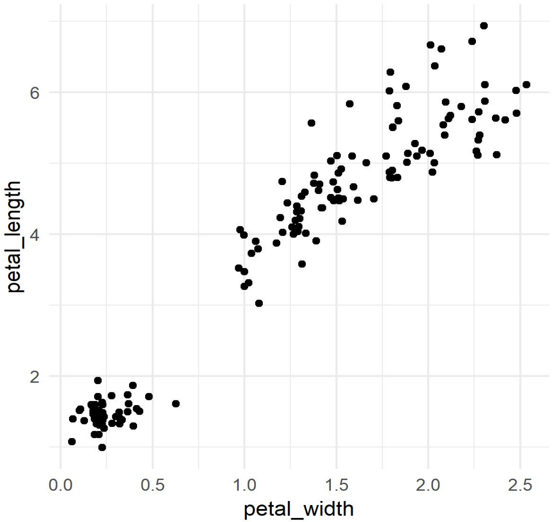
- Clustering

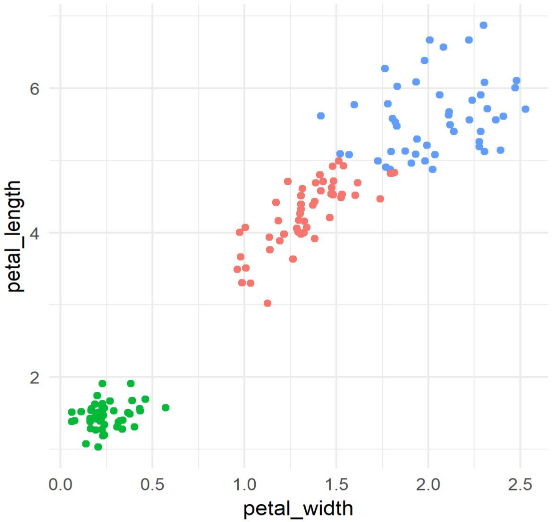
Clustering

Datos input: N observaciones - p variables

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2

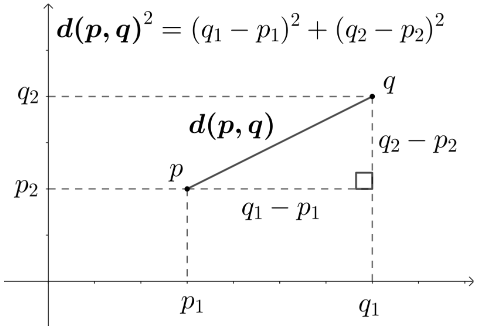
- Agrupar observaciones similares (***cohesión***)
- Separar observaciones distintas *(**separación**)*





Clustering

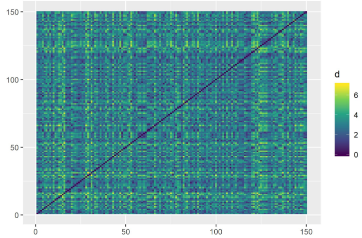
- ***Disimilitud / Distancia***



Distancia euclidiana

$$d_{euc}(p,q) = \sqrt{\sum_{j=1}^{k} (p_j - q_j)^2}$$

Matriz de distancias

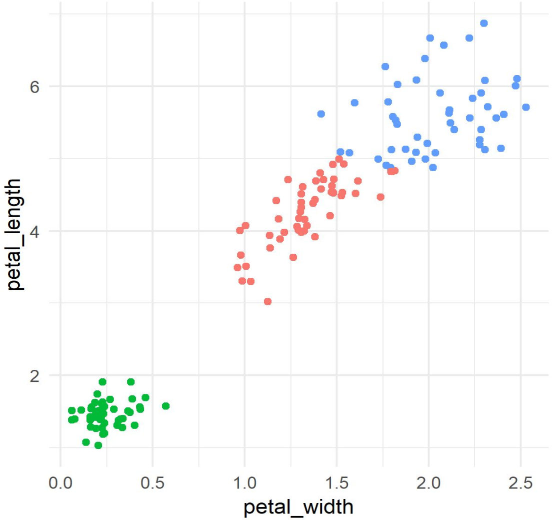


- ***Cohesión***
  - **Diámetro**

$$D(C) = \max(d_{ij})$$
$$i, j \in C$$

- ***Separación***
  - **Separación**

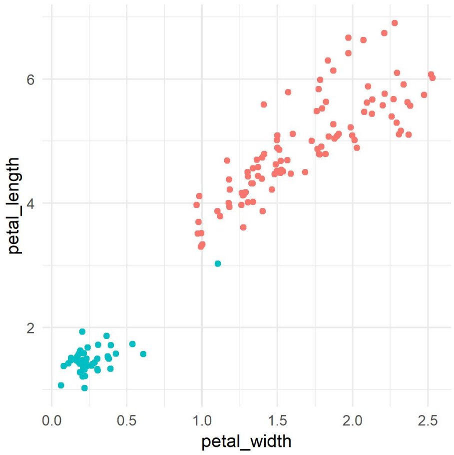
$$S(C) = \min(d_{ih})$$
$$i \in C, h \notin C$$



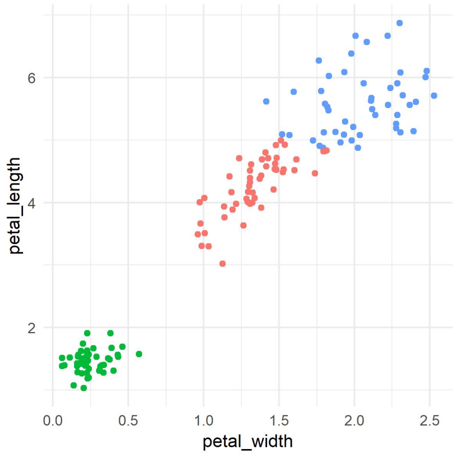
Clustering por partición

- Para *K* dado *C1,...,Ck* clusters ***excluyentes*** y ***colectivamente exhaustivos***
- Los clusters más pequeños ***no están anidados*** dentro de los más grandes

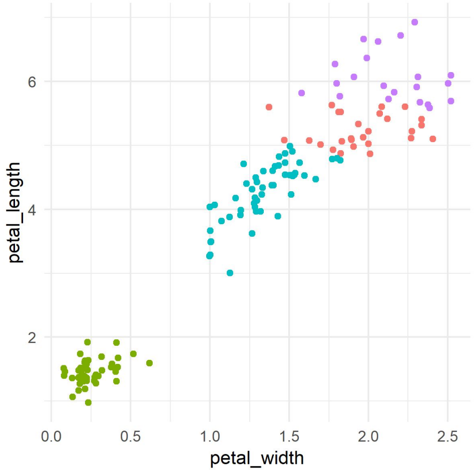
K=2



K=3



K=4



K-Medias

(0) Inicializar K centroides al azar

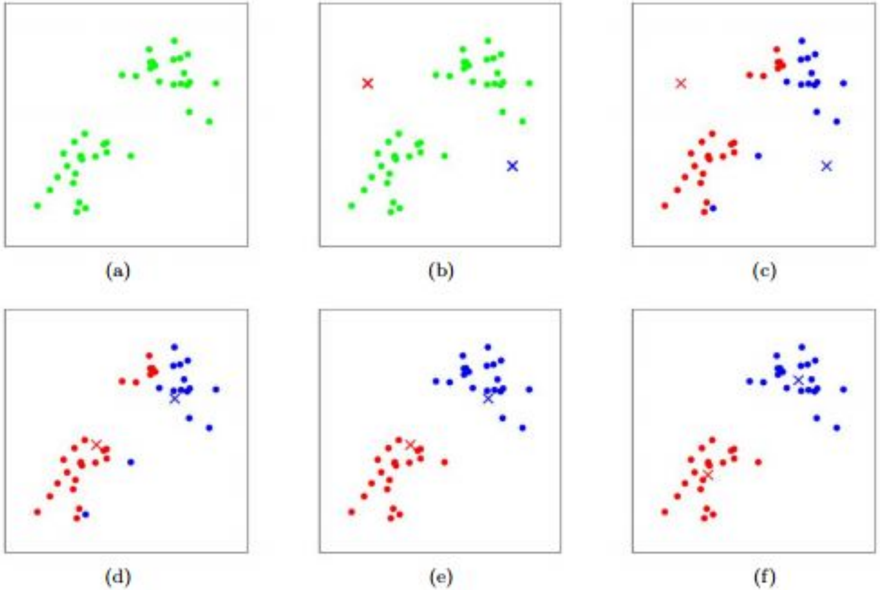
(1) Asignar cada objeto al cluster con **centroide más cercano**

(2) Computar el **centroide** de cada cluster

Repetir (1) y (2) hasta que no haya cambios en las asignaciones

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} VIC(C_k)$$

Variabilidad Intra-Cluster



Fuente: Stanford CS221

$VIC(C_k)$

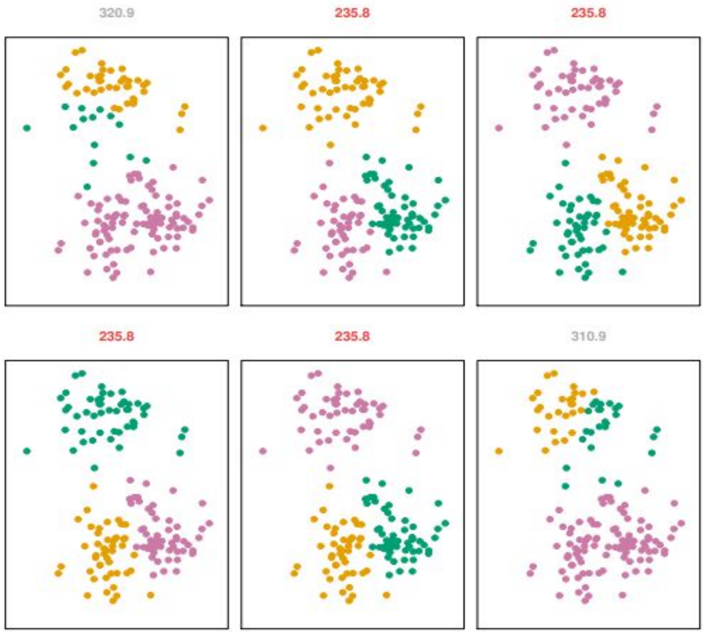
distancia euclidiana cuadrática

$$2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 = \overbrace{\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2}$$

"Within cluster Sum of Squares" **(WSS)**

K-Medias

El resultado depende de la inicialización



Fuente: An Introduction to Statistical Learning (2014)

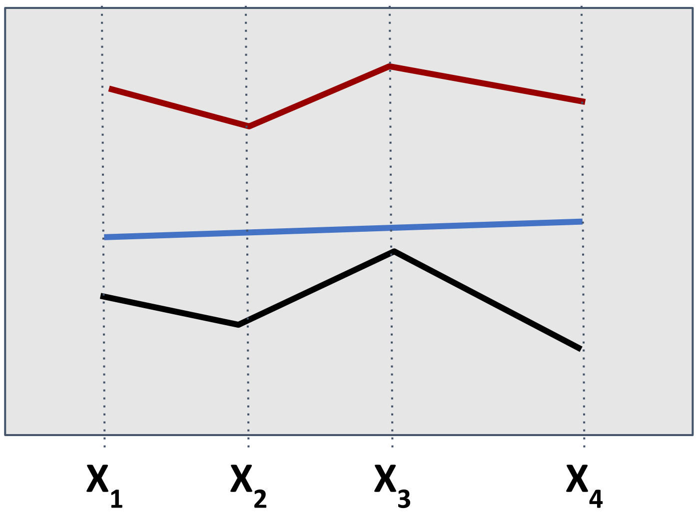
Distancia / (di)similitud

<u>*Variables cuantitativas*</u>

$$d_{euc}(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$$ **Euclidiana**

$$d_{man}(x,y) = \sum_{j=1}^{p}|(x_j - y_j)|$$ **Manhattan**

$$d_{cor}(x,y) = 1 - \frac{\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y})}{2\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2}} + \frac{1}{2}$$ **Correlación**



Es fundamental analizar si las variables deben ser **normalizadas**

Distancia / (di)similitud

Dados dos objetos i, j se computa la cantidad de de atributos para cada posible combinación

	$j = 1$	$j = 0$
$i = 1$	$n_{1,1}$	$n_{1,0}$
$i = 0$	$n_{0,1}$	$n_{0,0}$

Por ejemplo, $n_{1,1}$ indica la cantidad de atributos en los que ambos tienen 1

Fuente: Principles of Data Mining (2001)

*Variables binarias**

- Coeficiente de coincidencias $$\frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0}}$$
- Coeficiente de Jaccard $$\frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}}$$

Variables de tipo mixto

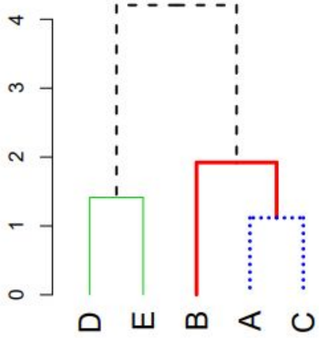
- Distancia de Gower

*Ver *A Survey of Binary Similarity and Distance Measures* (Choi et al, 2010)

Clustering jerárquico (aglomerativo)

0) *Inicialización: cada objeto es un cluster*
1) **Fusionar los dos clusters** ***más similares*** en un solo cluster

   Repetir (1) hasta tener un solo cluster



Dendrograma

Fuente: An Introduction to Statistical Learning (2014)

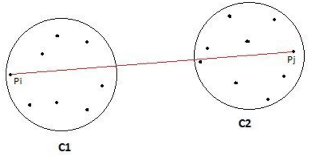
¿Qué es *similar*?

Métrica de distancia

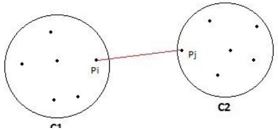
Euclidiana, Manhattan, correlación, etc.

Criterio de Linkage

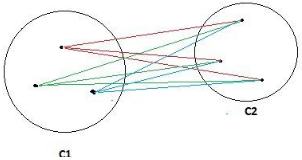
- **Complete** (disimilitud máxima)
- **Single** (disimilitud mínima)
- **Average** (disimilitud promedio)
- **Centroid** (disimilitud entre centroides)
- **Ward** (incremento de VIC)



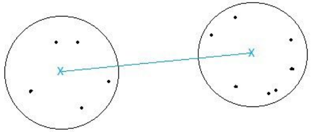
Complete



Single



Average



Centroid

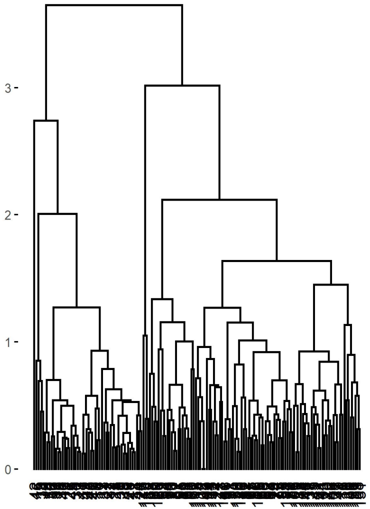
Fuente: towardsdatascience.com

¿Qué método elegir?

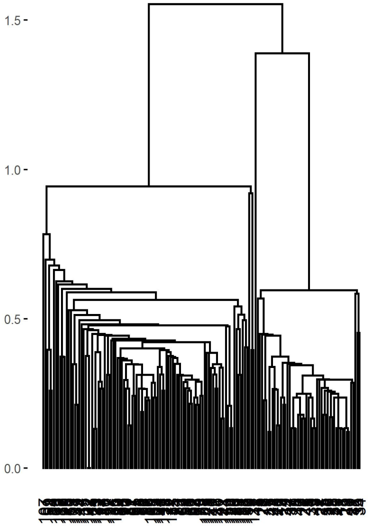
Coeficiente de correlación cofenético

- Comparación entre **distancias reales** y **distancias cofenéticas**

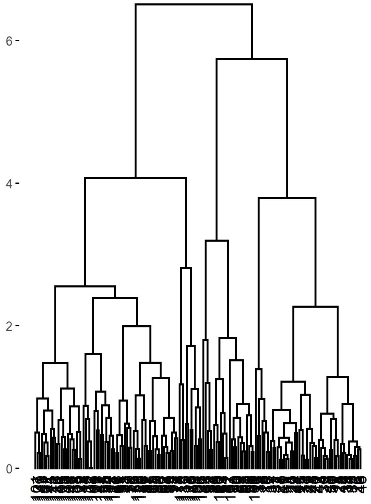
Average



Single



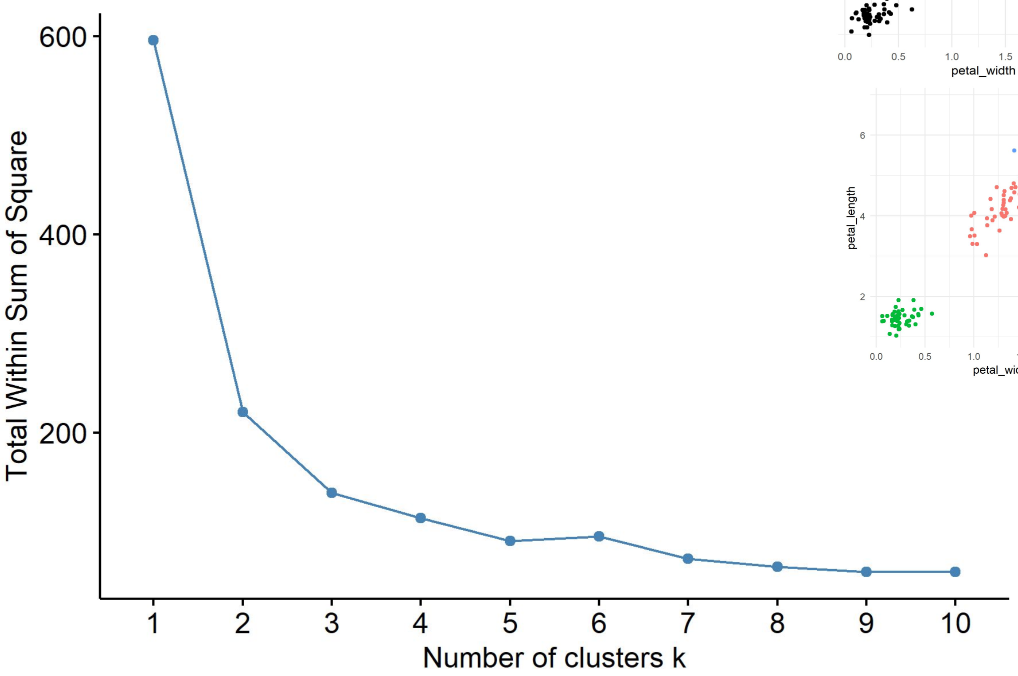
Complete

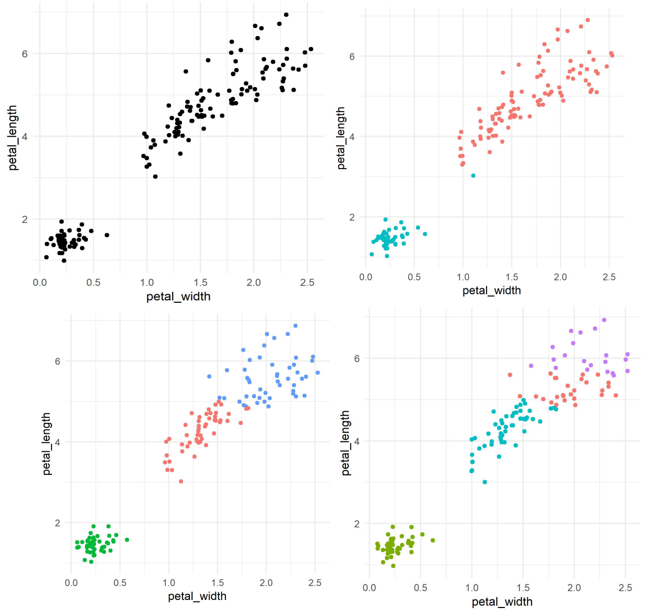


	Cophenetic Corr.
complete	0.7514592
average	0.8543606
single	0.8300050

¿Cómo elegir K?

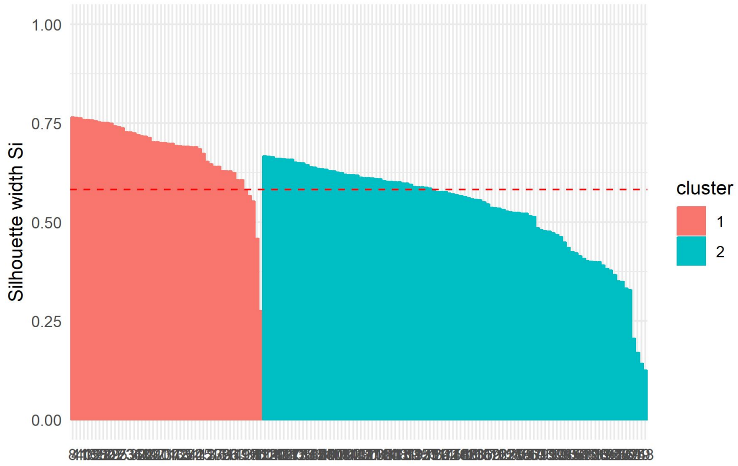
(1) Punto de quiebre en **VIC total**



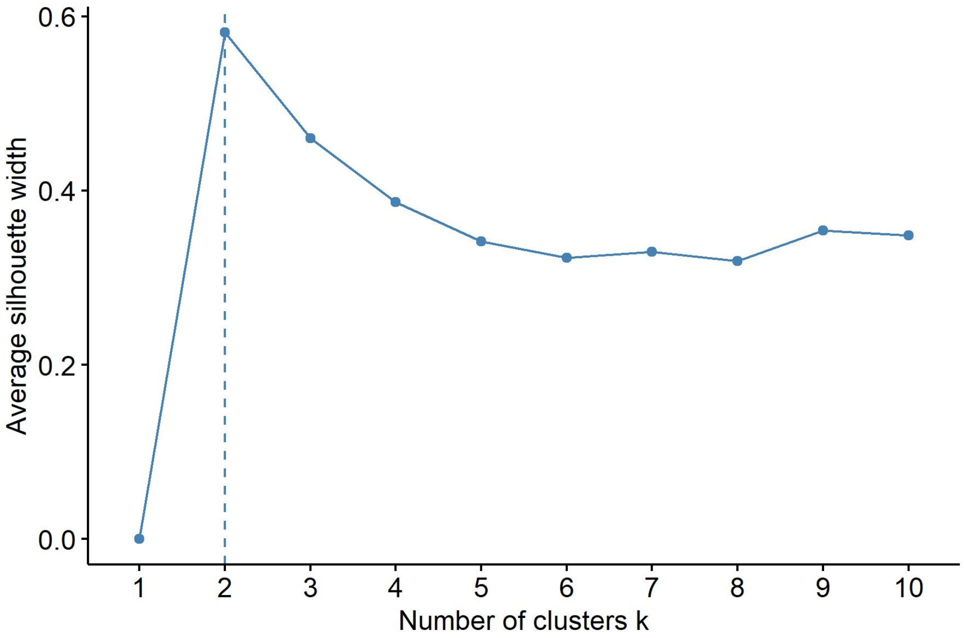


¿Cómo elegir K?

(2) **Silhouette** promedio máximo



$$s(i) = \frac{\overbrace{\min_{j \neq l(i)} \bar{d}(i, C_j)}^{b} - \overbrace{\bar{d}(i, C_{l(i)})}^{a}}{\max(\underbrace{\bar{d}(i, C_{l(i)})}_{a}, \underbrace{\min_{j \neq l(i)} \bar{d}(i, C_j)}_{b})}$$



a : distancia promedio de *i* a objetos de su cluster

b: mínima distancia promedio de *i* a otro cluster

Validación

- Estadístico de Hopkins*

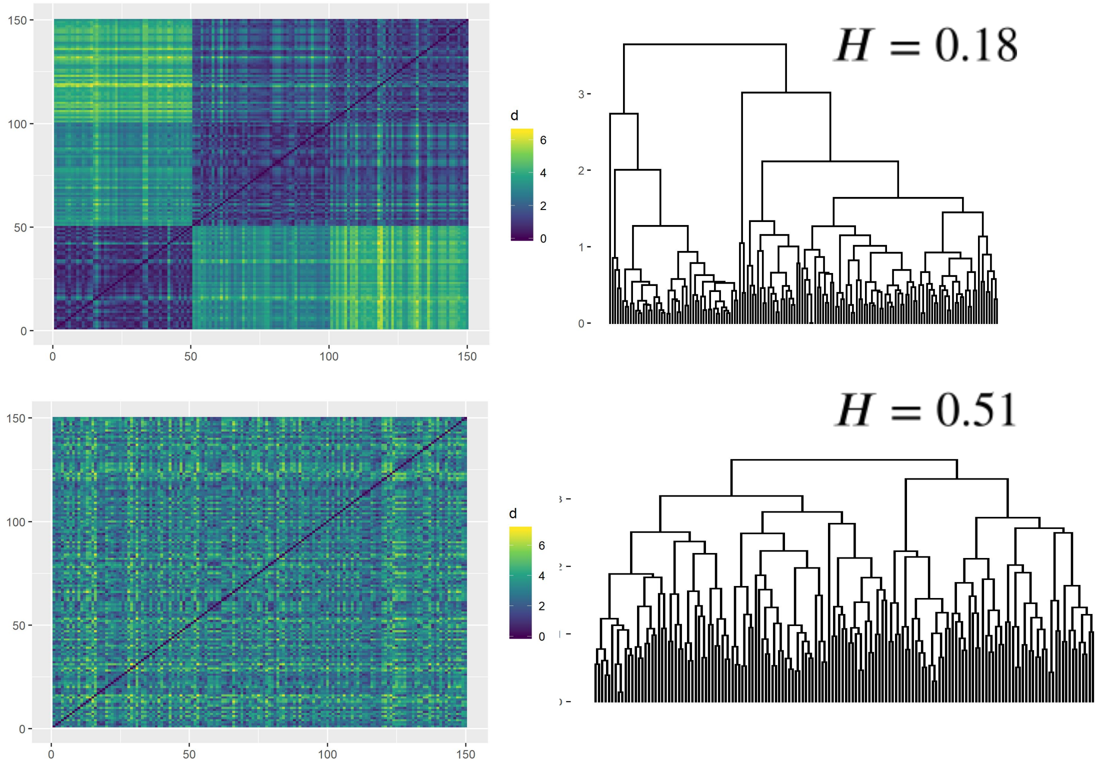
$$H = \frac{\sum_{i=1}^{s} x_i}{\sum_{i=1}^{s} x_i + \sum_{i=1}^{s} z_i}$$

- Dataset X
- Muestra de *s* objetos:

x_i : distancia entre objeto *i* y vecino más cercano en X

- Muestra de *s* objetos de un dataset con distribución uniforme:

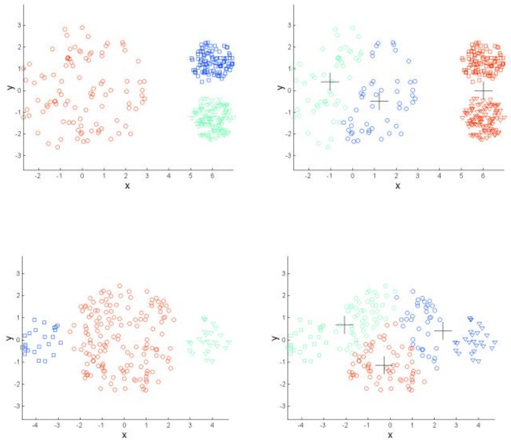
z_i : distancia entre objeto *i* y vecino más cercano en X

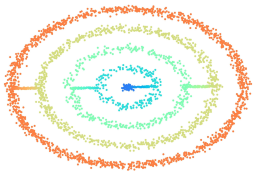


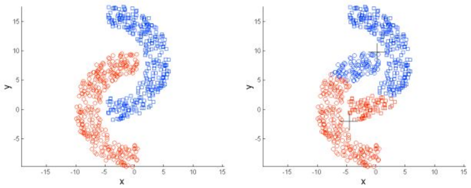
*En algunas implementaciones se usa *z* en el numerador — esto invierte la interpretación pero no altera las conclusiones

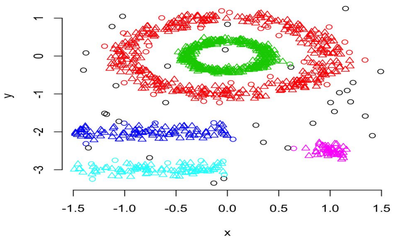
Extensiones

- **Fuzzy/Soft** clustering: Soft K-Means
- **Density-based** clustering: DBSCAN - Clustering Espectral
- **Large databases**: CURE - BFR
- **ETC**









Fuente: Western Michigan University

<u>Lecturas recomendadas</u>:
Kaufman y Rousseeuw (1990) - Finding Groups in Data
Gan et al (2007) - Data clustering: Theory, algorithms and applications

Conclusión

Selección de variables, **normalización** de variables, métrica de **distancia**, **algoritmo** de clustering, criterio de **linkage**, etc....

"... **decisions can have a strong impact on the results** obtained. In practice, we **try several different choices**, and look for the one with the most useful or interpretable solution. With these methods, **there is no single right answer**—any solution that exposes some interesting aspects of the data should be considered"

"Most importantly, we must be careful about how the results of a clustering analysis are reported. These **results should not be taken as the absolute truth about a data set**. Rather, they should constitute a starting point for the development of a scientific hypothesis and further study"

Introduction to Statistical Learning (2014)

Bonus Track

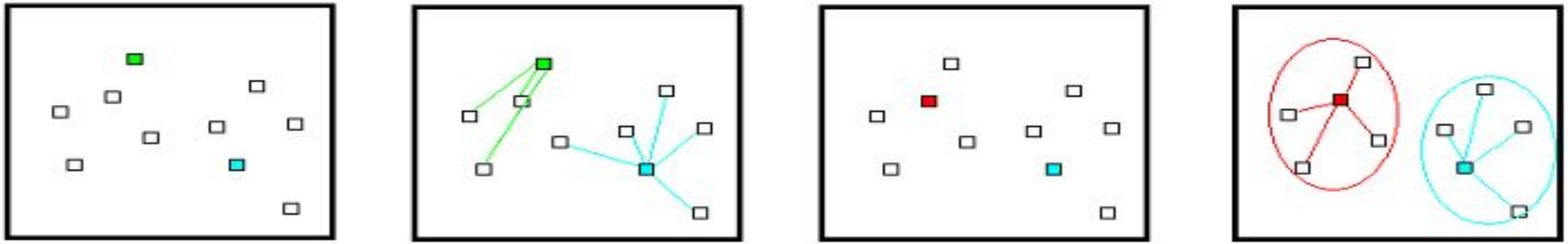
PAM (K-Medoides)

- Clustering por partición

1) Seleccionar K **medoides** al azar
2) Generar K clusters asignando cada objeto a su medoide más cercano
3) Para cada cluster, evaluar la caída en el costo (C) de intercambiar el medoide por otro objeto. Realizar el intercambio que maximiza la caída en el costo.
4) Si realizó al menos algún intercambio en (3), volver a (2) — caso contrario, finalizar.

$$C = \sum_{j=1}^{k} \sum_{i=1}^{n_j} d(o_i \ , m_i)$$

PAM (K-Medoides)



Fuente: https://blog.naver.com/asus1984/120065317344

¿Cómo elegir K?

(3) **Gap**

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B}\log(W_{kb}^{*}) - \log(W_k)$$

- W = **Variabilidad Intra Cluster**
- Se simulan **B** datasets sin agrupamientos (**distribución uniforme**) que tienen **W***
- Para cada K: diferencia promedio entre escenario real y simulado
- Se elige el menor K tal que **Gap(K) > Gap(K+1) - sd(K+1)**