Clase 6

Business Intelligence
Métodos no Supervisados

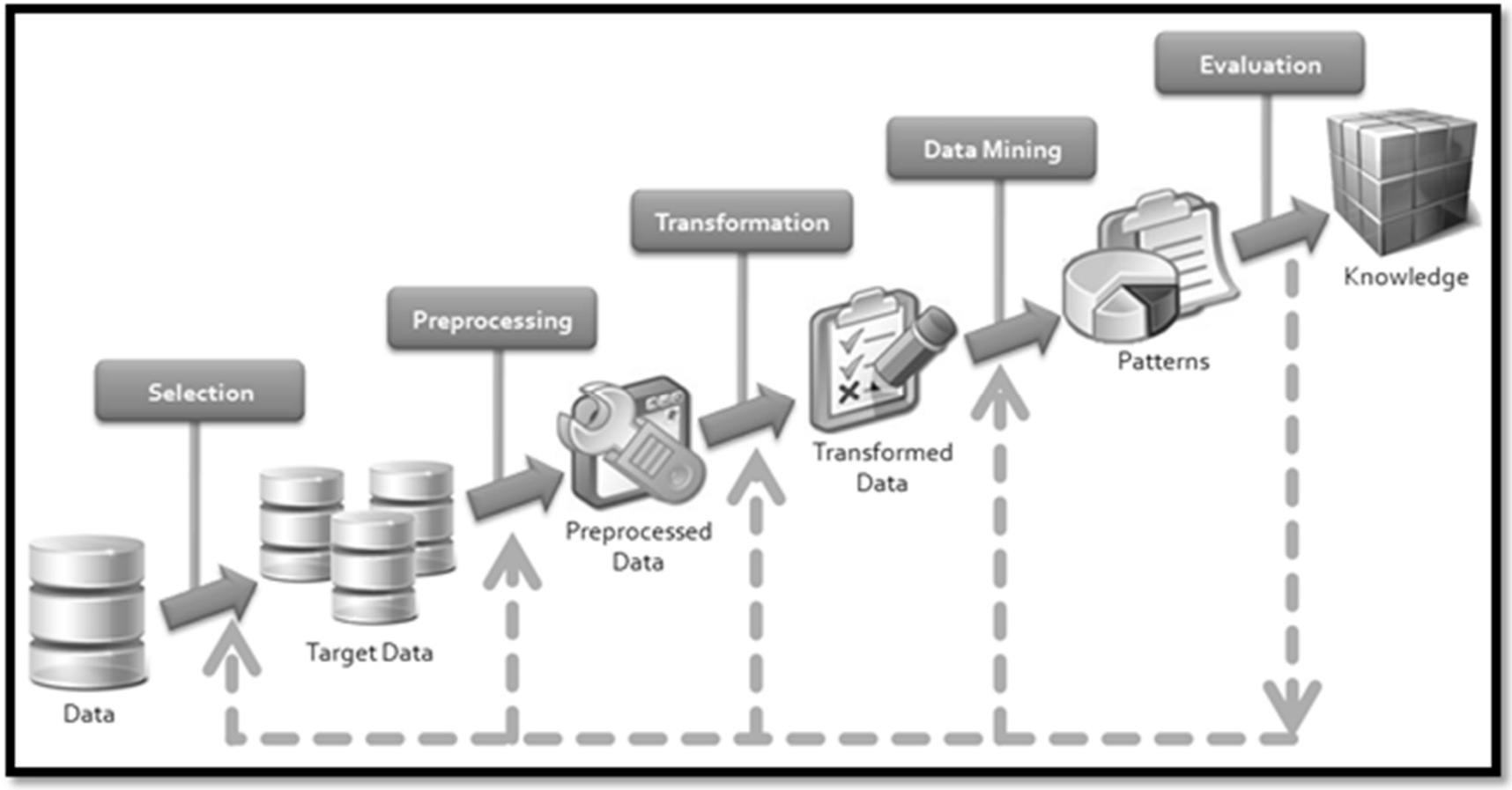
Sebastián Maldonado A.

Agenda

- Clustering
  - Proceso KDD (breve recuerdo)
  - Conceptos fundamentales
    - Aprendizaje supervisado v/s no supervisado
    - Medidas de similitud
    - Recomendaciones generales
    - Principales métodos de clustering
  - Actividad Taller Rapid Miner
    - Métodos de particionamiento
    - Métodos jerárquicos
    - Evaluación de resultados – determinación de la cantidad de segmentos

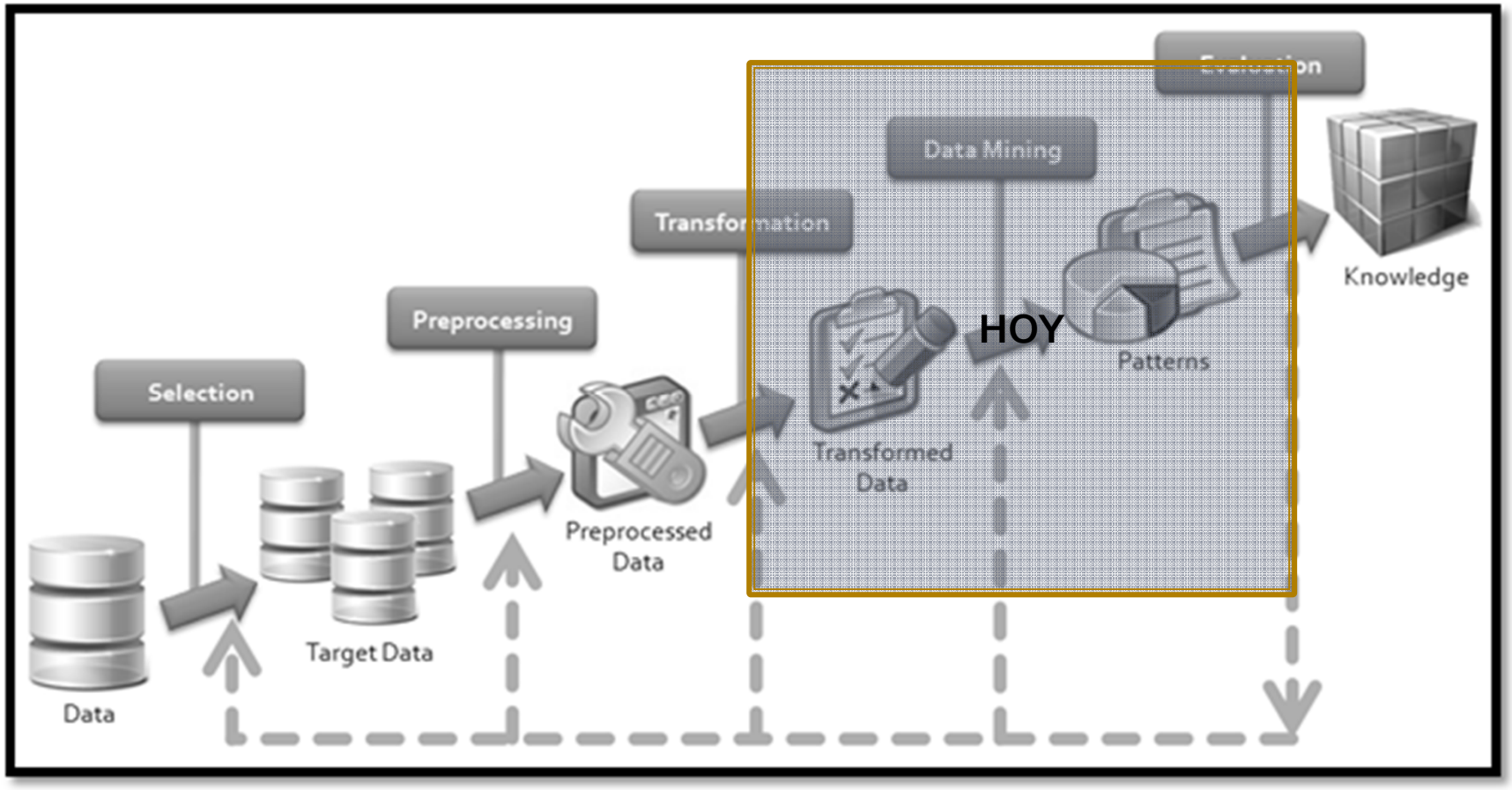
Clustering

Proceso KDD



Knowledge **D**iscovery in **D**atabases → **KDD**

Proceso KDD



Knowledge **D**iscovery in **D**atabases → **KDD**

RE-CAPITULANDO…

- **Aprendizaje supervisado**
  - Se utiliza el conocimiento *a-priori* del comportamiento de un conjunto de observaciones: Conjunto de Entrenamiento
  - Ejemplos: Arboles de Decisión, Redes Neuronales, Redes Bayesianas, Naïve Bayes, Support Vector Machines, etc…

- **Aprendizaje no-supervisado**
  - No se utiliza conocimiento *a-priori* del comportamiento de un conjunto de observaciones.
  - Ejemplos: **Fuzzy C-Means, k-Means**, Kohonen Self Organizing Maps, Expectation-Maximization, etc.

CLUSTERING
Conceptos fundamentales

- **Cluster**: Agrupación o colección de objetos
  - Similares entre aquellos objetos del mismo cluster
  - Distintos a los objetos de otros clusters.
- **Análisis de Clusters**
  - Agrupar un conjunto de datos en clusters en base a los atributos definidos para determinar cuán "similares" son unos objetos de los otros.
- Un buen Clustering produce conjuntos con:
  - **Alto nivel de "similitud"** entre los **objetos de la misma clase**.
  - **Bajo nivel de "similitud"** entre las **distintas clases**.
- La bondad de los clusters dependen directamente de la **opinión** de los usuarios y los expertos del negocio
  - **No olvidar que es una técnica descriptiva**

CLUSTERING

- Medidas de "Similitud"
  - Similitudes o diferencias puede ser definidas con la ayuda medidas de **distancias**.
  - Se desea tener la posibilidad de comparar dos tipos de objetos $i = (x_{i,1}, ..., x_{i,n})$ y $j = (x_{j,1}, ..., x_{j,n})$
  - Propiedades:

$$d(i,j) \geq 0$$

$$d(i,i) = 0$$

$$d(i,j) = d(j,i)$$

$$d(i,j) \leq d(i,h) + d(h,j)$$

CLUSTERING (2)

- Medidas de "Similitud"
  - Distancia de Minkowski

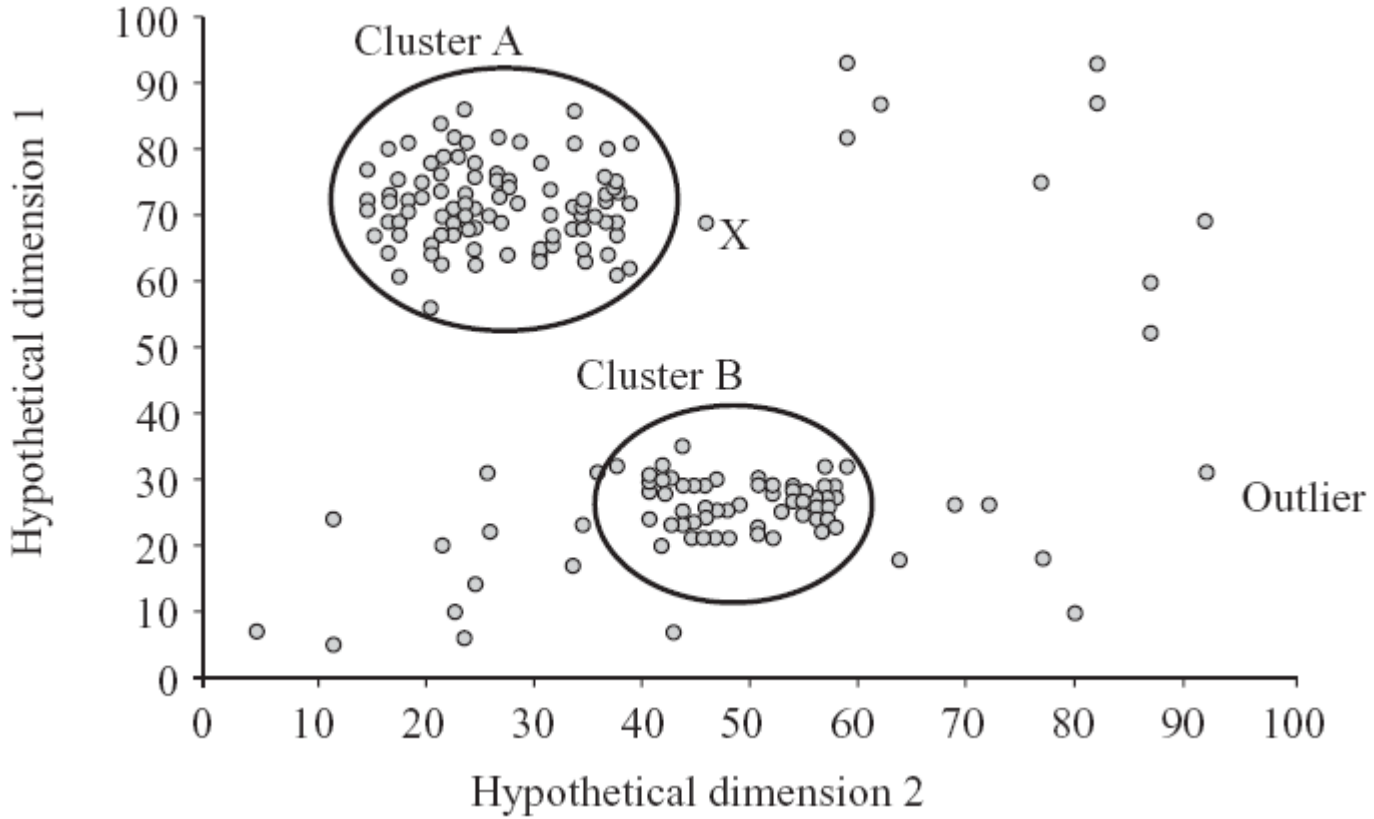
$$d(i,j) = \sqrt[q]{w_1 \cdot |x_{i,1} - x_{j,1}|^q + \ldots + w_n \cdot |x_{i,n} - x_{j,n}|^q}$$

  - Distancia euclidiana, q = 2, w = 1
  - Distancia "Manhattan" o "city block", q = 1, w = 1
  - Coseno del ángulo entre los vectores (texto)

- Otras distancias:
  - Distancia de Hamming (vectores binarios)
  - Distancia en grados de separación en un grafo (redes sociales)

CLUSTERING (4)

- La idea principal…



[Fuente: "Data Mining Techniques for Marketing, Sales and Customer Relationship Management" , Michael J.A. Berry, Gordon S. Linoff.

CLUSTERING
Tipos de Algoritmos

- **Métodos de Particionamiento:**
  - El número de segmentos se determina *a-priori*.
  - Partiendo de una segmentación inicial (casi siempre aleatoria) se distribuyen los objetos hasta cumplir un criterio de particionamiento.
  - Métodos de particionamiento difuso (e.g. fuzzy C-means)

- **Métodos Jerárquicos**
  - **Top-Down** (divisivo)
  - **Bottom-up** (aglomerativo)

- **Otros métodos**
  - Métodos basados en densidades (e.g. DBScan)
  - Métodos basados en modelos (e.g. EM algorithm)

PARTICIONAMIENTO
La idea general

- El problema de particionamiento puede ser formulado como un problema de optimización:

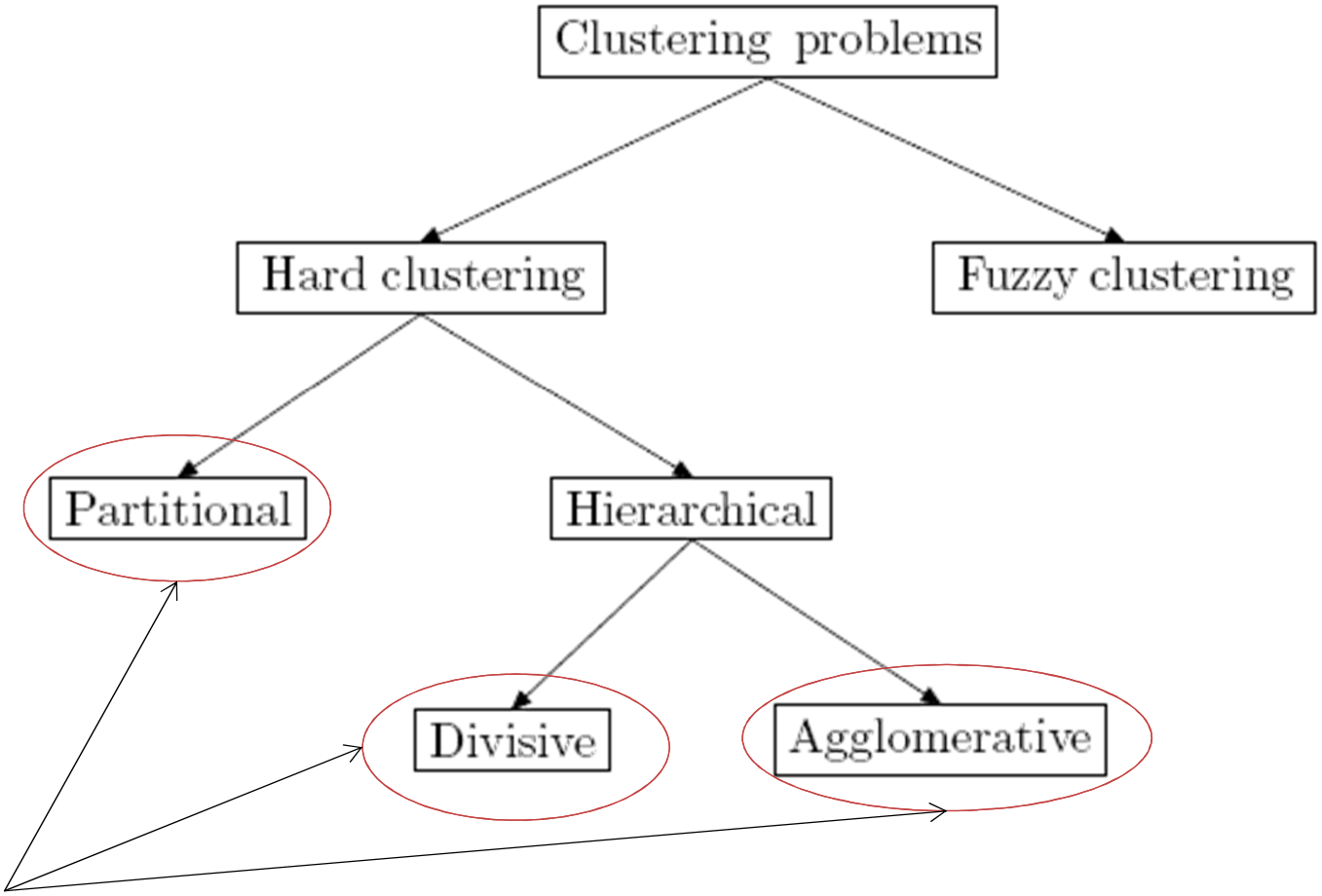
  "Dado un **K**, se desea encontrar los **K conjuntos** tal que **maximicen (o minimicen)** un **criterio de partición**."
  (Problema Computacionalmente Complejo)

- Resolver el problema mediante métodos tradicionales no es factible → métodos heurísticos:
  - Pertenencia absoluta (e.g.**K-Means)**
  - Grados de pertenencia (e.g. **Fuzzy C-means**)

CLUSTERING
Preprocesamiento de datos

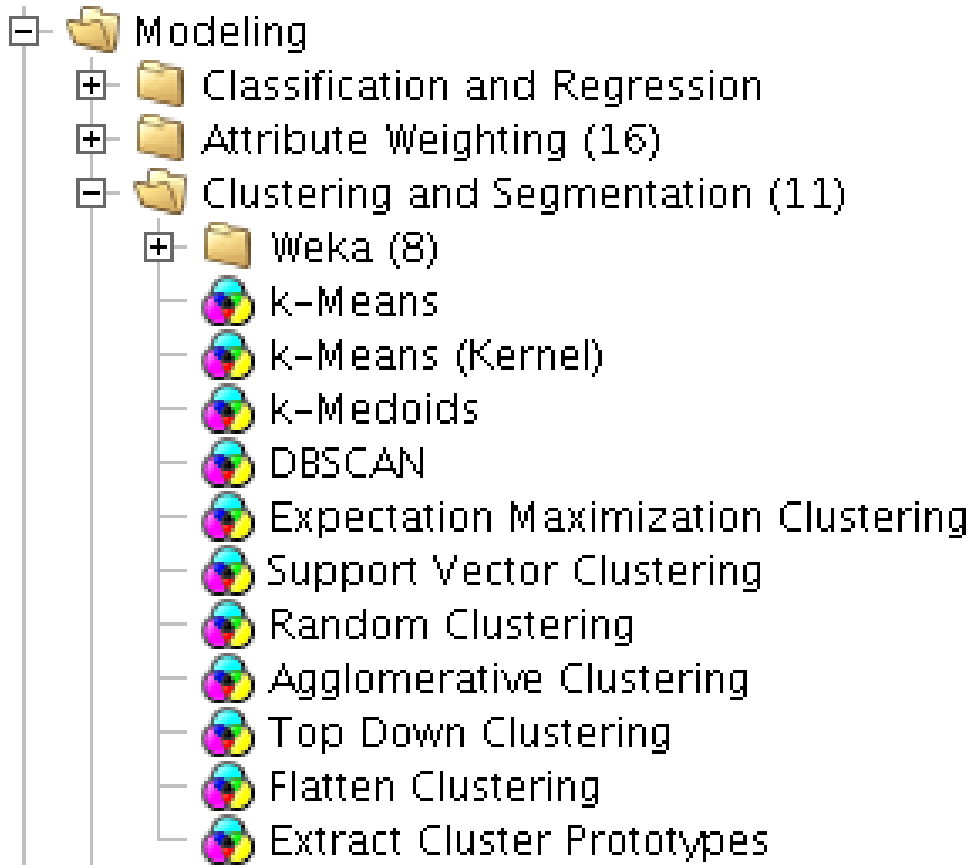
- Es relevante **escalar los valores** de las variables en un mismo rango:
  - Si no se normaliza o estandariza, algunos atributos pueden tomar mayor relevancia que otros en el modelo, afectando los resultados.
- Es importante **utilizar aquellos atributos que son relevantes** para disminuir la complejidad computacional:
  - Muchos atributos pueden no aportar información relevante y afectar los tiempos de cálculo de los segmentos.

CLUSTERING
Métodos de segmentación



Disponibles en RapidMiner v5.0

CLUSTERING
Métodos en RapidMiner 5



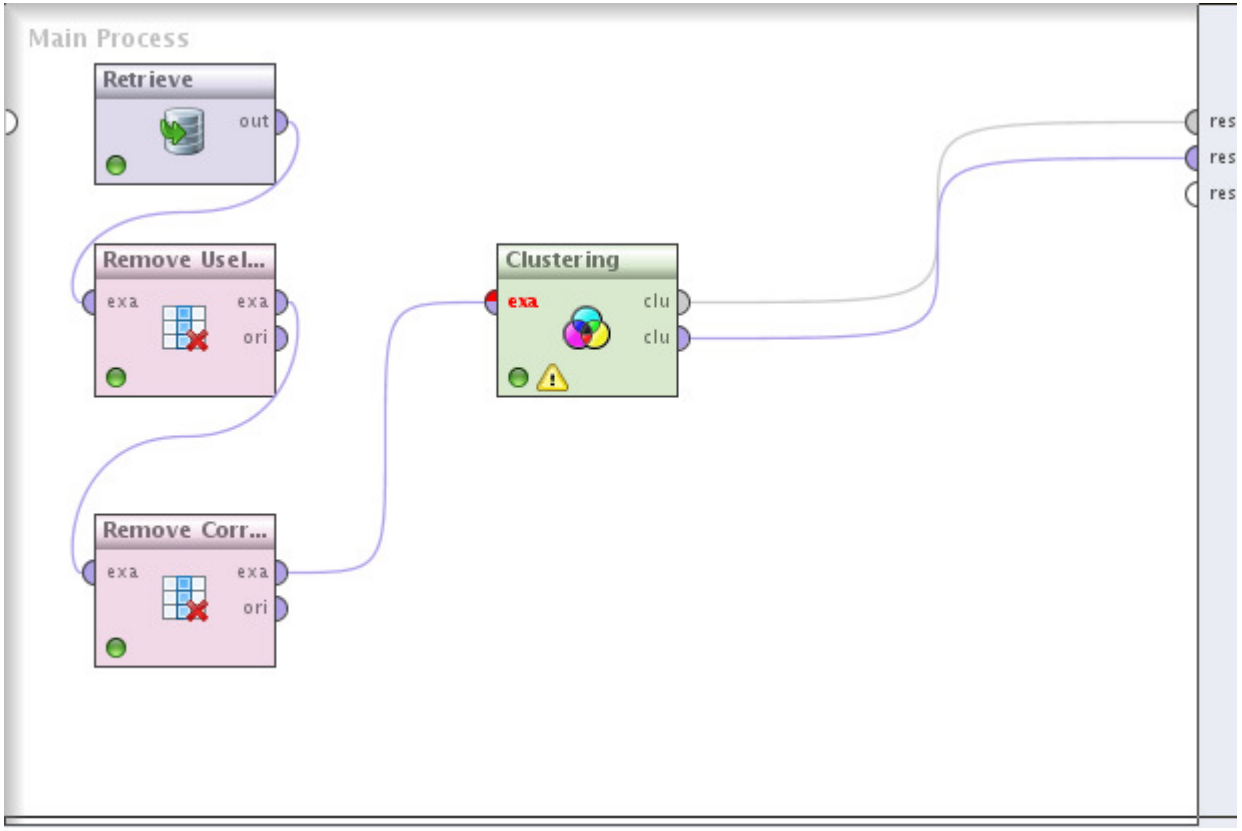
PARTICIONAMIENTO
K-Means

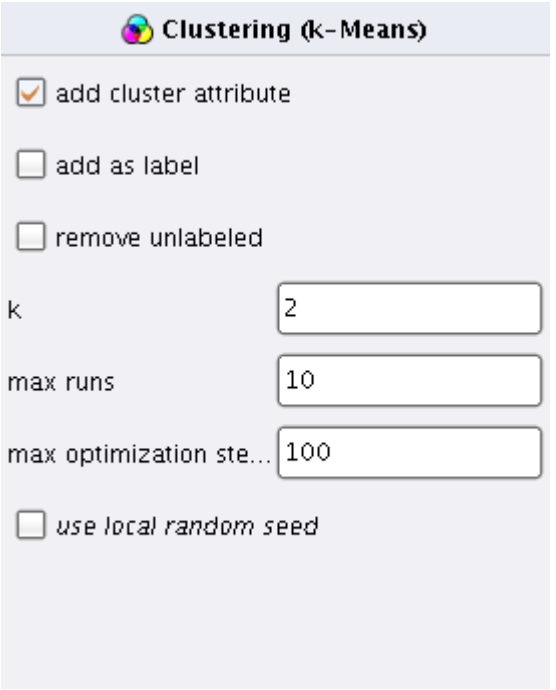
- Dado un conjunto de **objetos** y un número **K**,

1. **Particionar** los **objetos** en **K** subconjuntos no vacíos.
2. Calcular los centroides de los subconjuntos.
3. → El centroide representa el **centro de cada cluster**.
4. **Asignar a cada objeto** al cluster cuyo centroide sea el **más cercano**.
5. Volver al paso 2, y detenerse cuando ya no sean necesarias las actualizaciones.

- Idea fundamental:
  - Asigna automáticamente un objeto a un determinado cluster.
  - No permite grados de pertenencia o probabilidades, solamente entrega las distancias a los distintos centroides.

CLUSTERING
Kmeans en RapidMiner 5

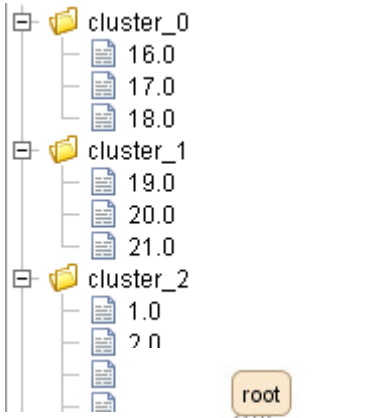


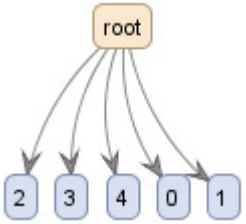


CLUSTERING
Kmeans en RapidMiner 5 [3]

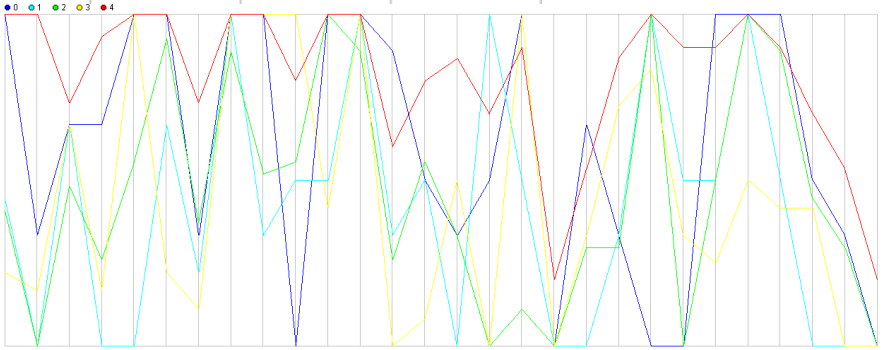
Cluster Model

```
Cluster 0: 3 items
Cluster 1: 3 items
Cluster 2: 9 items
Cluster 3: 6 items
Cluster 4: 5 items
Total number of items: 26
```





Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
a1	1	0.444	0.407	0.222	1
a2	0.333	0	0	0.167	1
a3	0.667	0.667	0.481	0.667	0.733
a4	0.667	0	0.259	0.167	0.933
a5	1	0	0.556	1	1
a6	1	0.667	0.926	0.222	1
a7	0.333	0.222	0.370	0.111	0.733
a9	1	1	0.889	1	1
a10	1	0.333	0.519	1	1
a15	0	0.500	0.556	1	0.800
a16	1	0.500	1	0.417	1
a18	1	1	0.889	1	1
a21	0.889	0.333	0.259	0	0.600
a23	0.500	0.50			
a24	0.333	0			
a25	0.500	1			
a31	1	0.50			
a33	0	0			
a34	0.667	0			
a40	0.333	0.33			



CLUSTERING JERARQUICO

- **Clustering Jerárquico Aglomerativo (Bottom Up)**
  - Se inicia con cada uno de los datos presentes en la base de datos como un cluster, y se van agrupando iterativamente hasta tenerlos todos incluídos en una jerarquía.
  - Problemas de implementación: Número muy grande de datos iníciales.

- **Clustering Jerárquico Divisivo (Top Down)**
  - Toda la base de datos se inicia como un cluster. Posteriormente se van definiendo los clusters iterativamente dentro del cluster inicial de manera jerárquica.

CLUSTERING JERARQUICO Aglomerativo

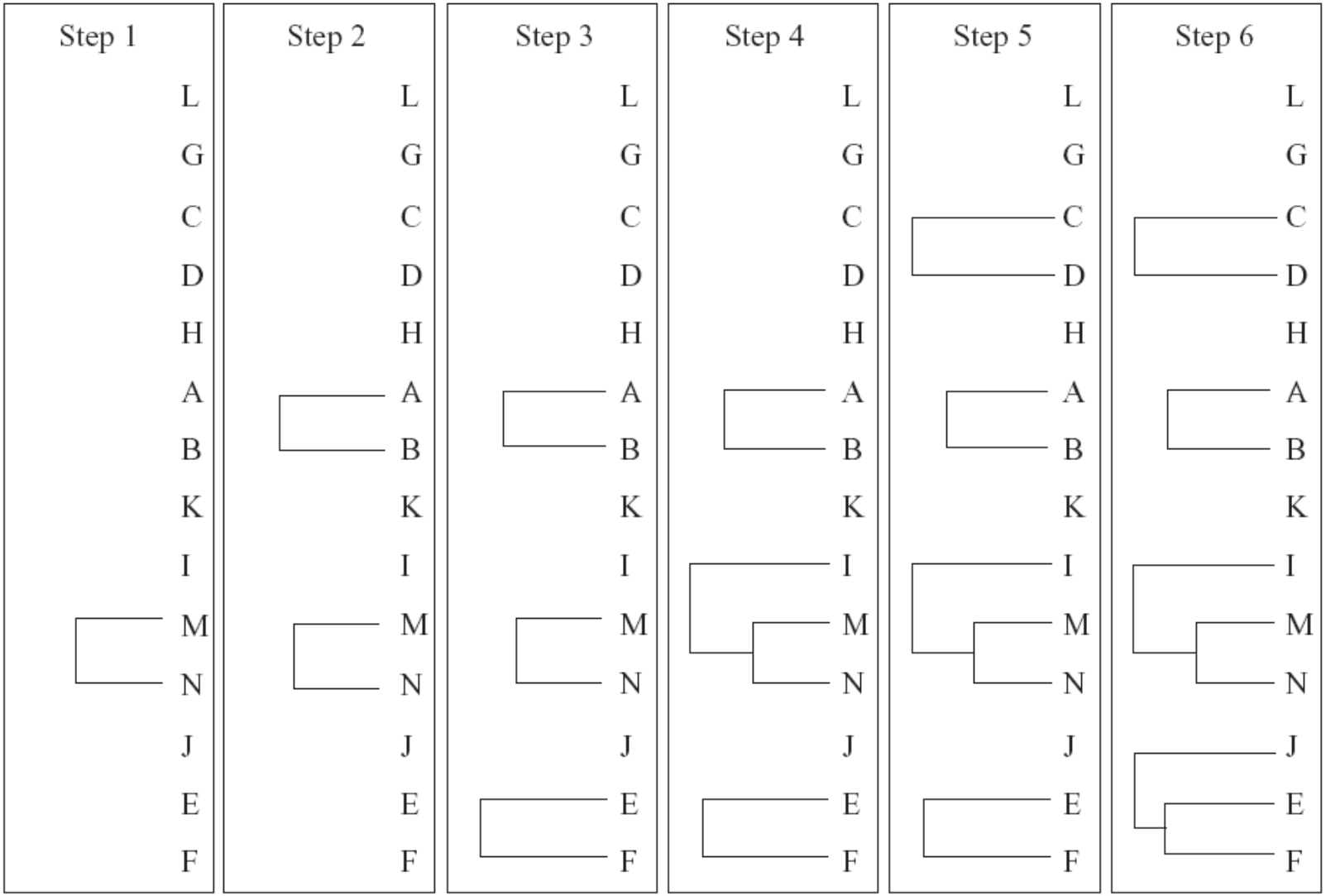
Cluster_Aglomerativo(D)

1. Se **inicializan** todos los puntos en **D** como un cluster independiente
2. Determinar las **distancias** entre los clusters.
3. **While** no hay más clusters que agrupar **do**
    1. **Combinar** 2 clusters que presenten la minima distancia entre ellos.
    2. **Actualizar** las distancias entre los clusters.
4. **End While**

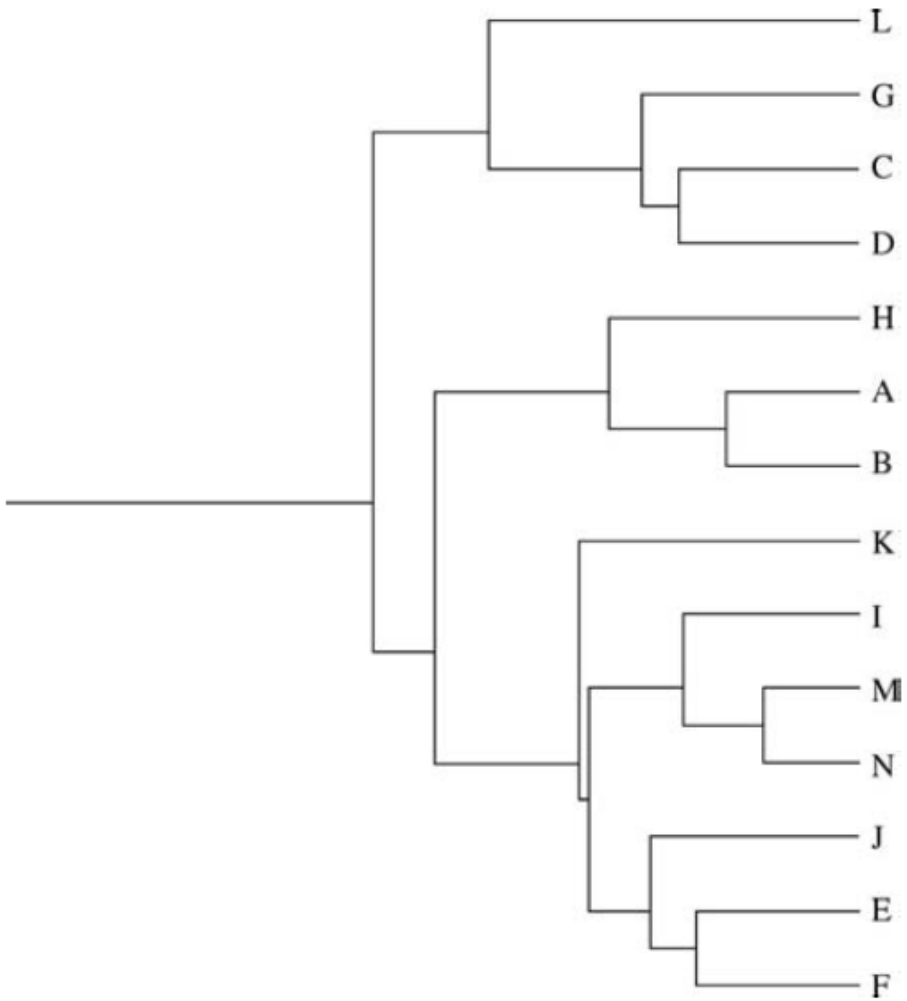
CLUSTERING JERARQUICO Aglomerativo [2]

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	7	8	4	5	2
B	6	8	5	4	2
C	8	9	7	8	9
D	6	7	7	7	8
E	1	2	5	3	4
F	3	4	5	3	5
G	7	8	8	6	6
H	8	9	6	5	5
I	2	3	5	6	5
J	1	2	4	4	2
K	3	2	6	5	7
L	2	5	6	8	9
M	3	5	4	6	3
N	3	5	5	6	3

CLUSTERING JERARQUICO
Aglomerativo [3]



CLUSTERING JERARQUICO
Aglomerativo [4]

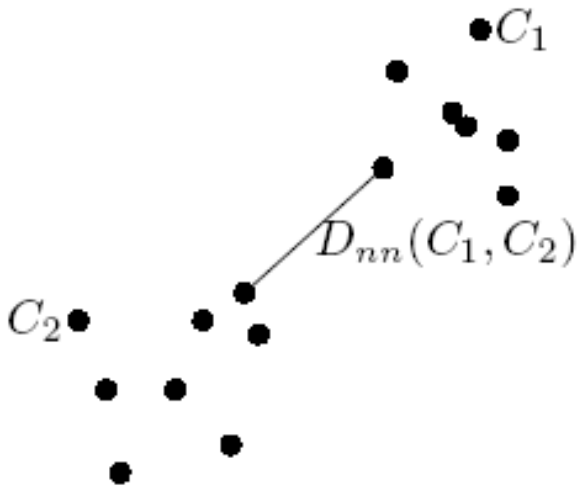


CLUSTERING JERARQUICO Aglomerativo [5]

- Determinación de la unión entre los clusters
  - **Enlace promedio (average link)**: Se calcula la **distancia media** entre los miembros del cluster y las observaciones en consideración a ser incluídas en el cluster.
  - **Enlace Singular (single link)**: Se calcula la distancia entre todos los miembros del cluster y las observaciones en consideración a incluir en el cluster, seleccionando **la menor distancia**.
  - **Enlace completo (complete link)**: Se calcula la distancia entre todos los miembros del cluster y las observaciones en consideración a incluir en el cluster, seleccionando **la mayor distancia**.

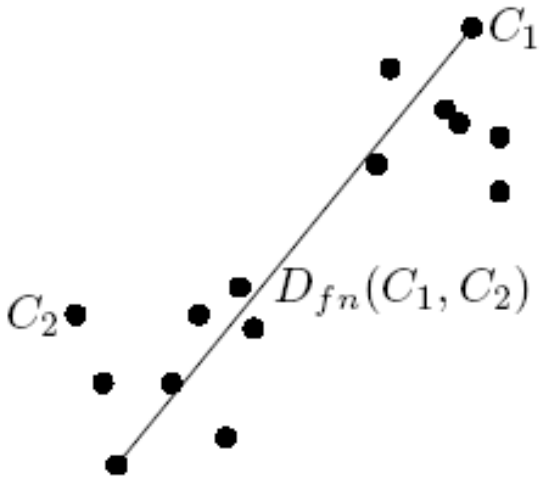
CLUSTERING JERARQUICO Aglomerativo [6]

Enlace Singular



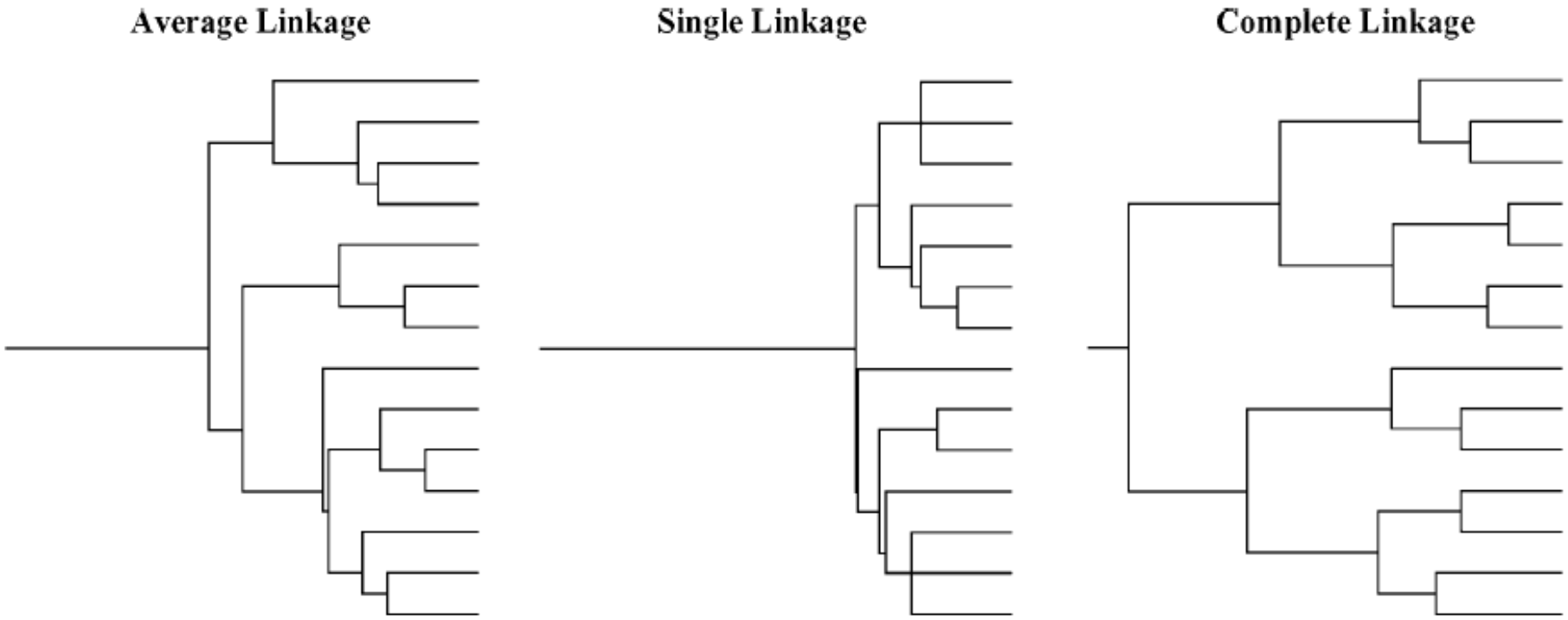
$$D_{nn}(C_1, C_2) = \min_{1 \le i \le r, 1 \le j \le s} d(\mathbf{y}_i, \mathbf{z}_j).$$

Enlace Completo

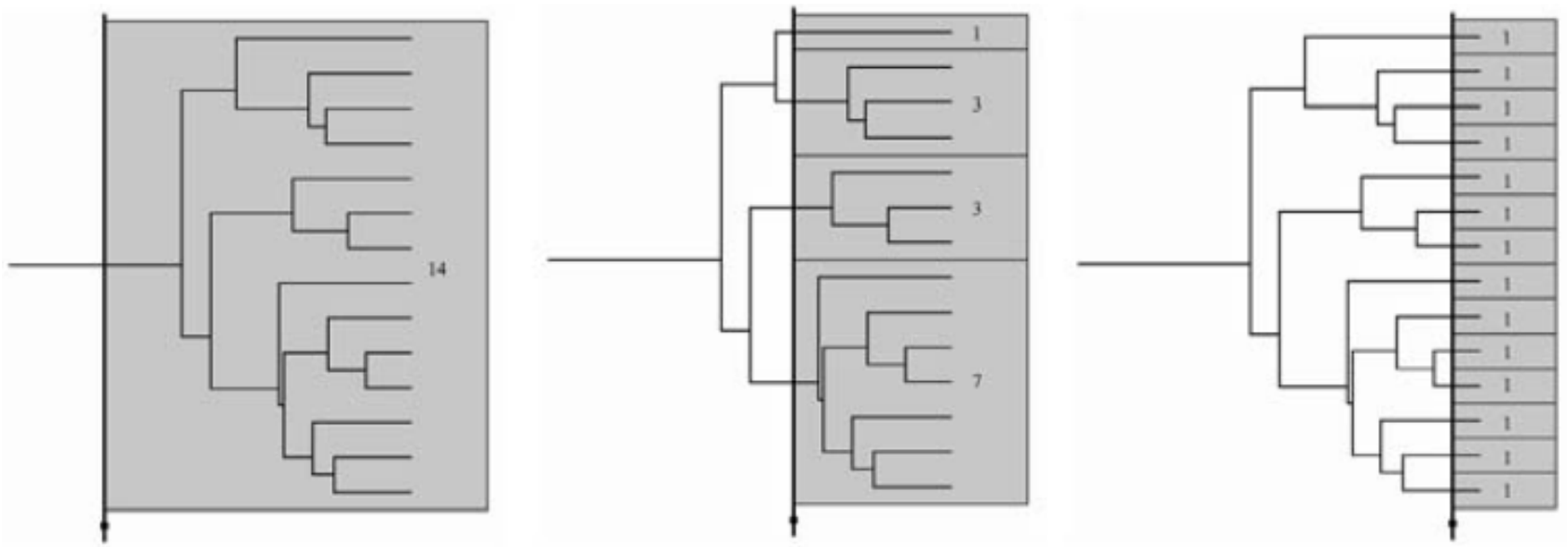


$$D_{fn}(C_1, C_2) = \max_{1 \le i \le r, 1 \le j \le s} d(\mathbf{y}_i, \mathbf{z}_j).$$

CLUSTERING JERARQUICO
Aglomerativo [7]



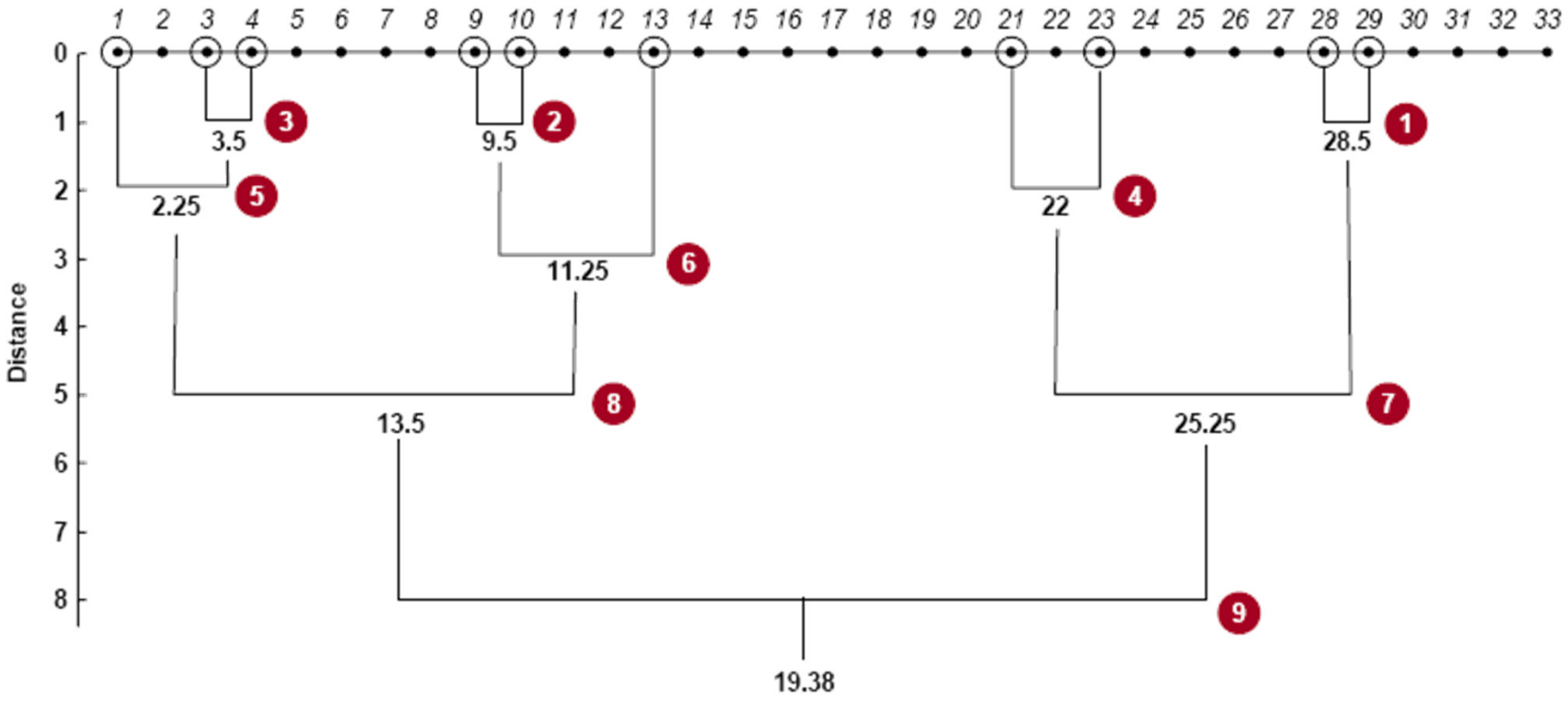
CLUSTERING JERARQUICO Aglomerativo [8]



- **Nota:** Se puede fijar el numero de clusters en base al criterio de la distancia entre los clusters (Recordar que es un método descriptivo).

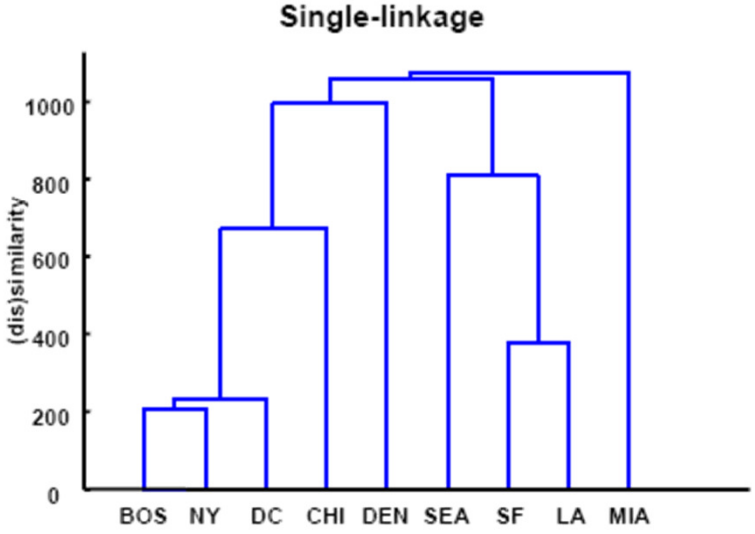
CLUSTERING JERARQUICO Aglomerativo [9]

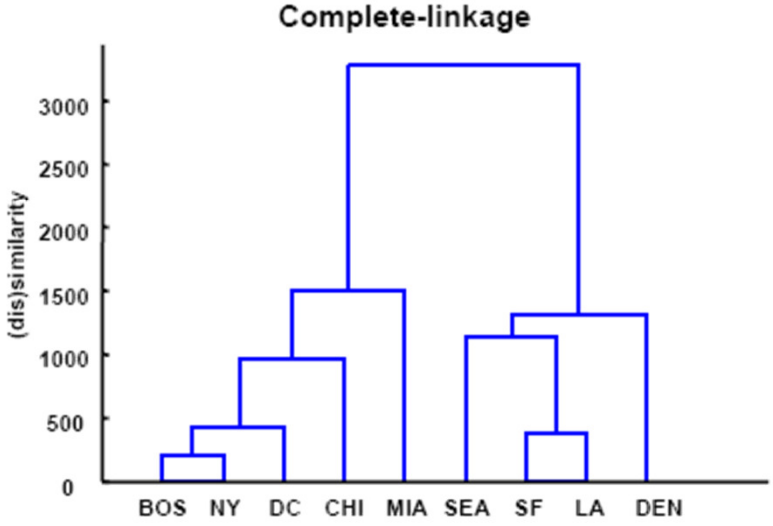
- Supongamos X = {1,3,4,9,10,13,21,23,28,29}



CLUSTERING JERARQUICO
Aglomerativo [10]

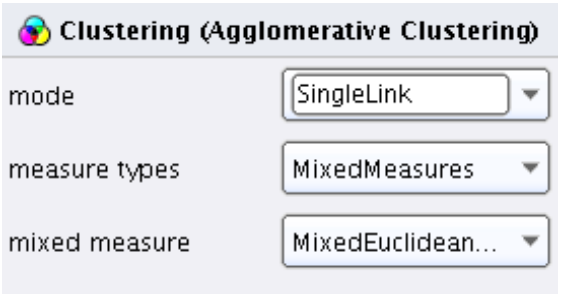
	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0



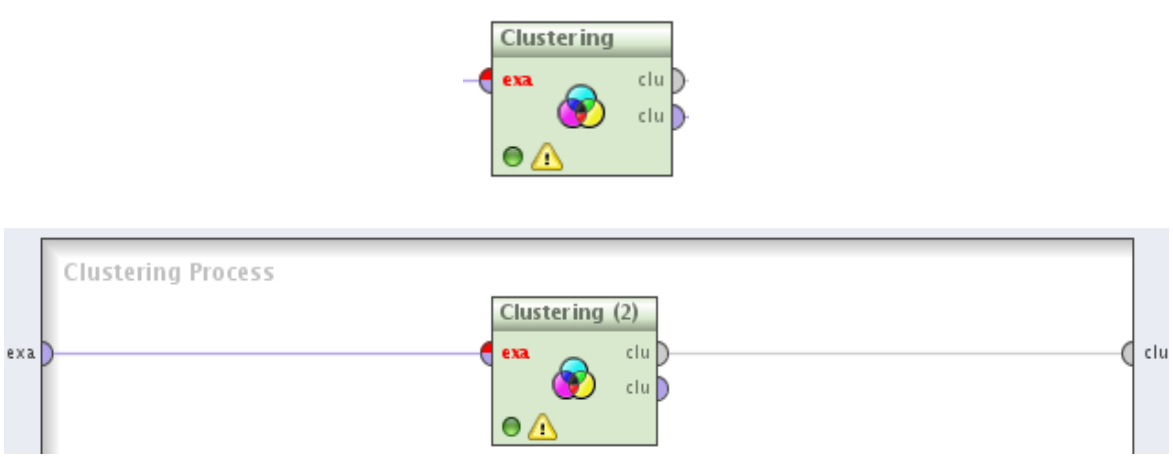


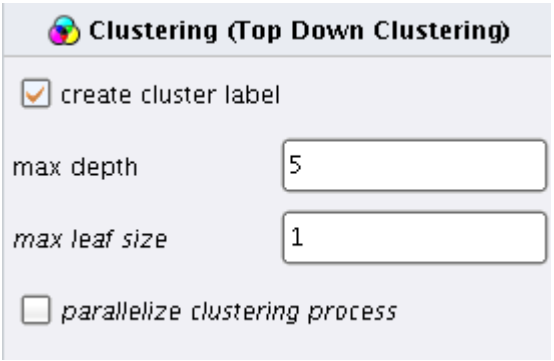
CLUSTERING
Jerárquico en RapidMiner 5

- Jerarquico Aglomerativo → Bottom up



- Jerarquico Divisivo → Top Down





CLUSTERING
Jerárquico en RapidMiner 5 [2]

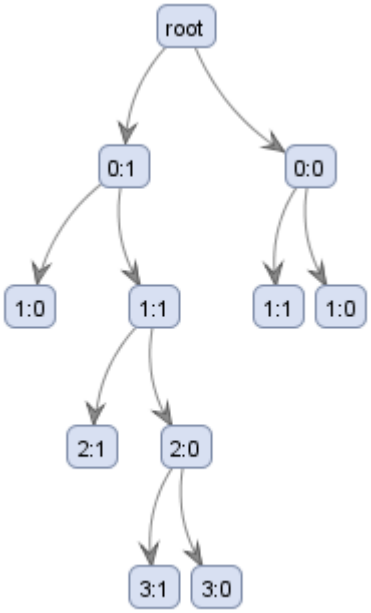
- Jerarquico Aglomerativo → Bottom up
  - Opciones SingleLink, CompleteLink y AverageLink
  - Measures_types: depende del tipo de atributos utilizados. Recomendación: utilizar MixedMeasures.

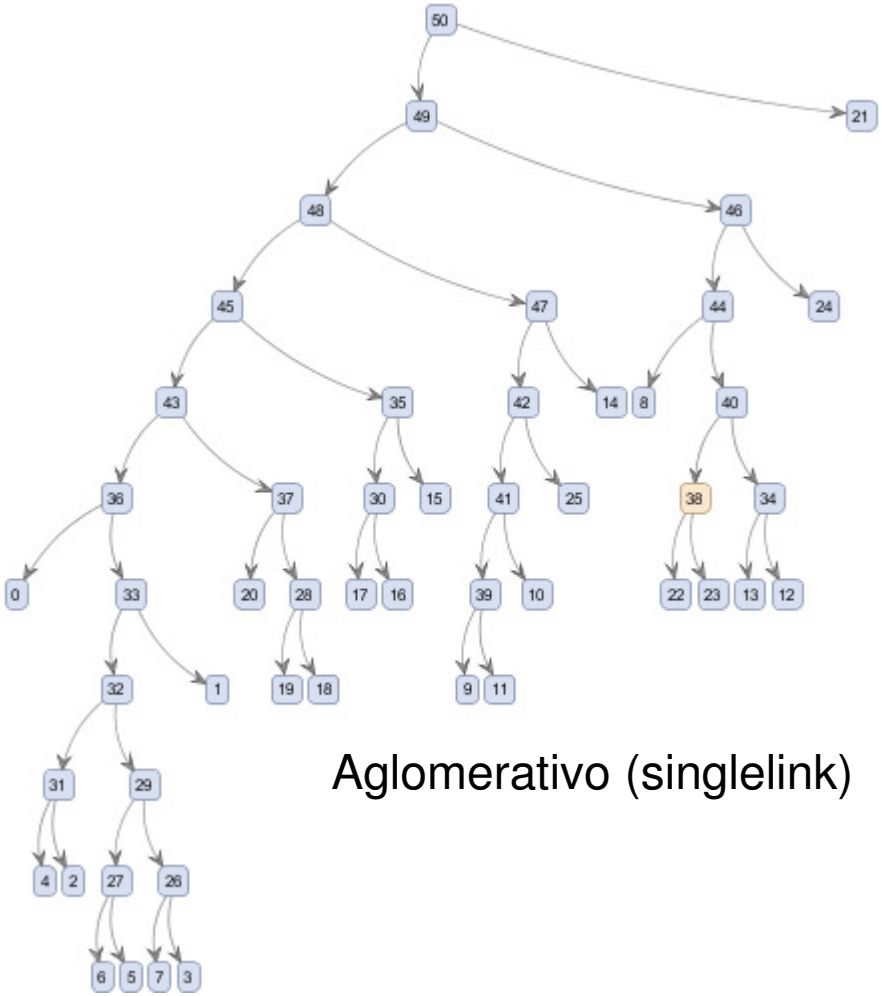
CLUSTERING
Jerárquico en RapidMiner 5 [2]

- Jerarquico Divisivo → Top Down
  - Es necesario considerar un operador interno para realizar el particionamiento de los conjuntos
    - Operador recomendado: kmeans
  - Max_leaf_size: Cantidad de elementos máximos por hoja (depende de la base de datos en uso)
  - Max_depth: Máxima profundidad del árbol de particionamiento

CLUSTERING
Jerárquico en RapidMiner 5 [3]

Divisivo (kmeans)



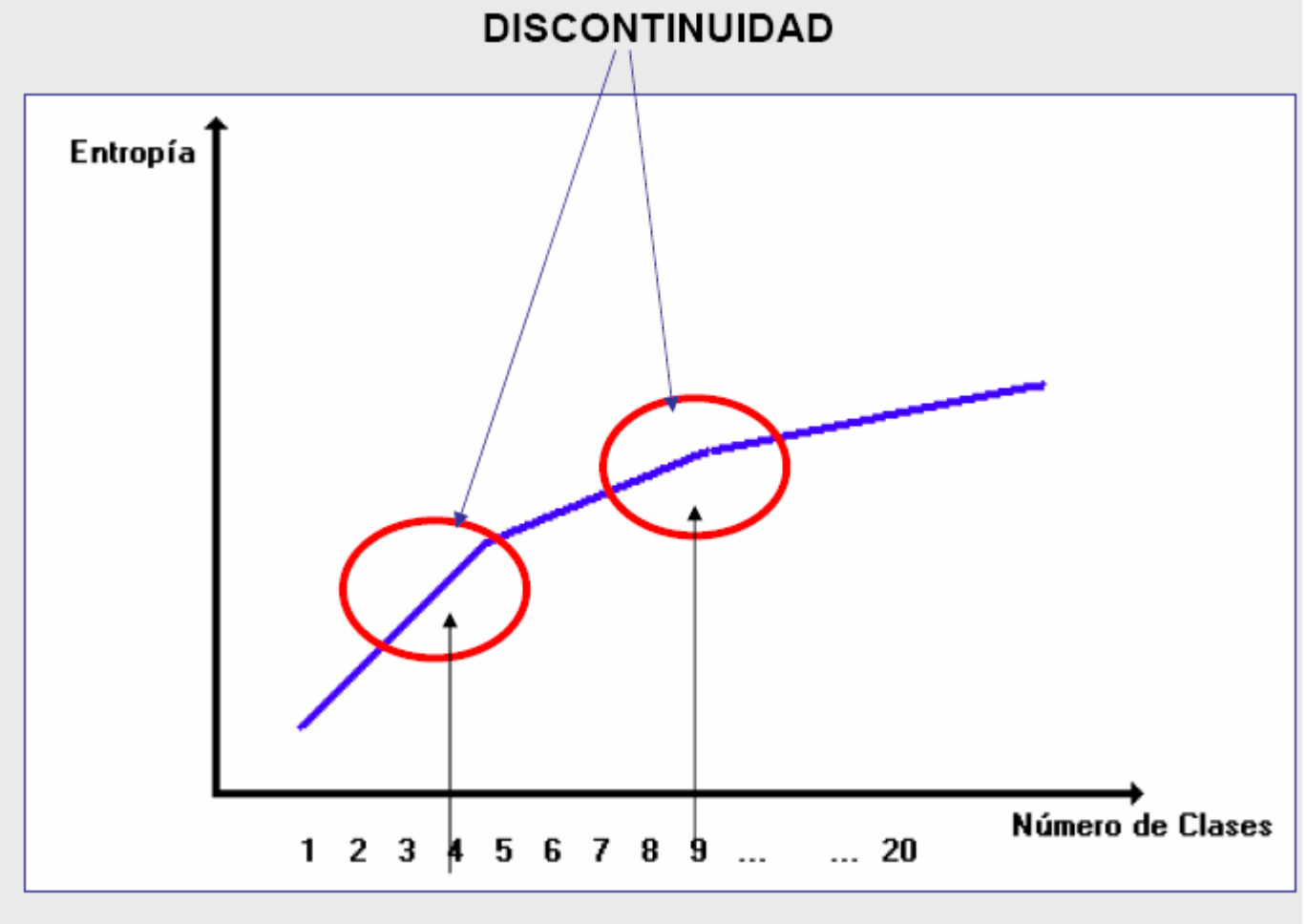


Aglomerativo (singlelink)

CLUSTERING
Evaluación del parametro "k"

- Existe una gran variedad de reglas y metodologías:
  - Regla **k = sqrt(n/2)**, con n = cantidad de objetos
  - Regla del codo
  - Índice Davis-Boudin (varianza intra e inter cluster)
  - Gráficos Silhouette
  - Índices de Criterio de Información (AIC, BIC, etc.)
  - "k" en base a interpretación clustering jerárquico por expertos del negocio.

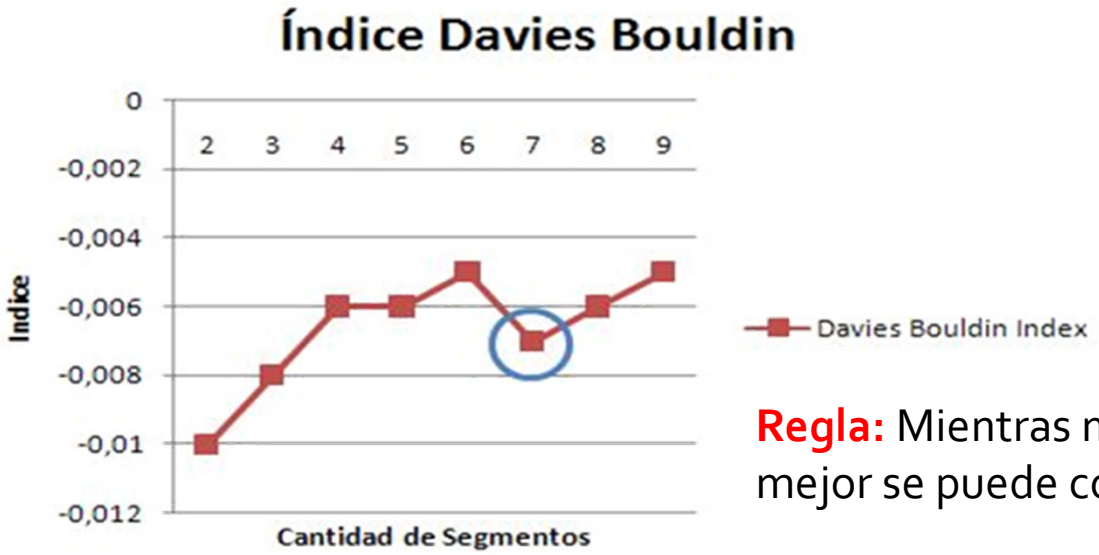
CLUSTERING
Regla del Codo



CLUSTERING
Índice de Davies Bouldin

$$DB = \frac{1}{K}\sum_{i=1}^{K} \max_{i \neq j} \left[\frac{Sk(Q_i) + Sk(Q_j)}{S(Q_i + Q_j)} \right]$$

- **Sk(Qi)** = distancia media entre los objetos y su centroide
- **S(Qi + Qj)** = distancia media entre los centroides
- **K** = número de clusters



Regla: Mientras menor es el valor del índice DB, mejor se puede considerar la segmentación

Clustering – Ventajas

- Flexible:
  - Existen varias alternativas de implementación con un bajo nivel de parámetros.
  - Es posible utilizar distintas medidas de distancia para distintas aplicaciones, dependiendo de la aplicación.
- Aproximaciones Jerárquicas y no Jerárquicas.
  - Es posible organizar el conjunto de datos de manera jerárquica, lo que permite identificar entre otras cosas el numero de clusters presentes.
  - Es posible definir un ajuste de la base de datos en función de un número inicial de clusters que se deseen comprobar.
- Aproximaciones paramétricas
  - Permiten estimar las probabilidades de pertenencia de los objetos a los distintos clusters

Clustering – Desventajas

- Interpretación:
  - Siempre es necesario un análisis con expertos en el tema para validar los resultados obtenidos.
  - Es difícil definir la medida de "similitud" entre objetos, depende de la interpretación.
  - Gran variedad de métodos para analizar la segmentación. No es fácil decidirse por un solo método, es necesario evaluar la mayor cantidad y apoyarse a la vez de conocimiento experto.
- Implementación:
  - Muchas técnicas de clustering son costosas en implementación,
  - El definir adecuadamente los clusters en una determinada base de datos puede ser sumamente costoso computacionalmente.

Referencias

- "Data Mining Techniques for Marketing, Sales and Customer Relationship Management" , Michael J.A. Berry, Gordon S. Linoff.
- "Making Sense of Data", Glenn J. Myatt
- "Data Mining – Challenges, Models, Methods and Algorithms", Markus Hegland
- "Data Clustering: Theory, Algorithms and Applications", Guojun Gan, Chaoqun Ma, Jianhong Wu