

# Introducción a la Bioinformática

## Data Clustering

Fernán Agüero  
Instituto de Investigaciones Biotecnológicas, UNSAM

# Agrupamiento de datos / Data Clustering

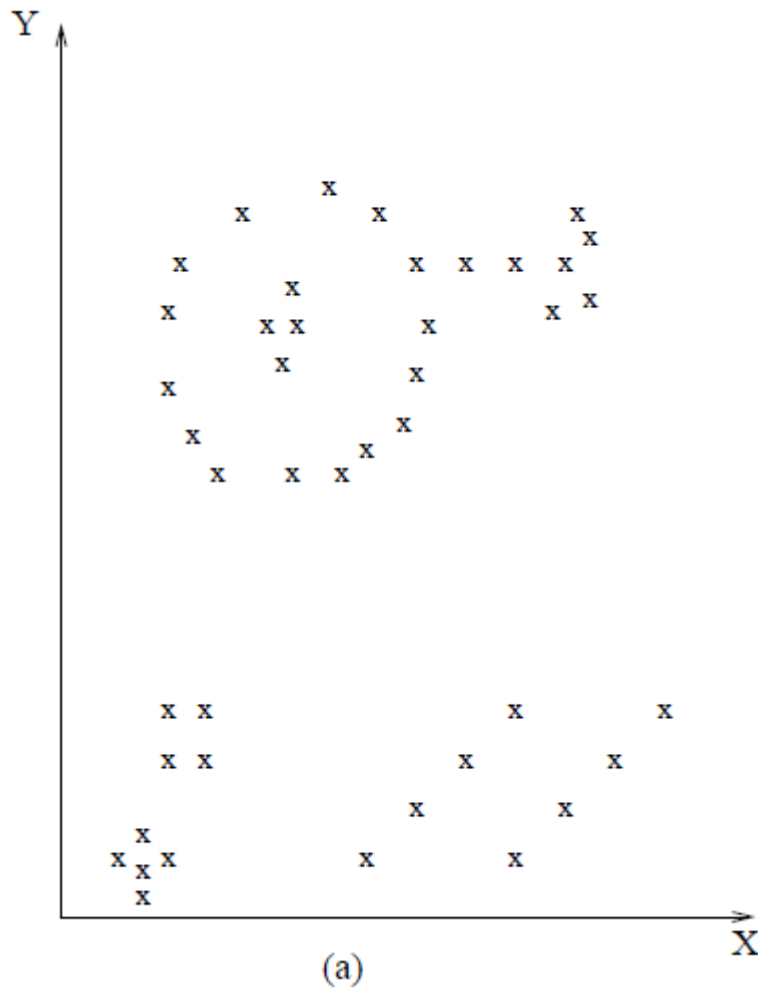
- El **agrupamiento de datos** consiste en la **clasificación** de objetos diferentes grupos, de manera que objetos similares son agrupados en el mismo grupo.
- Otra definición: particionar un conjunto de datos en subconjuntos o *clusters* de tal manera que estos tengan “algo en común”.
  - El problema: cuantificar “algo en común”
    - Proximidad
    - Similitud
- Es un tipo de aprendizaje **no supervisado**
- Es un problema combinatorio difícil

# Hay muchos tipos de datos ...

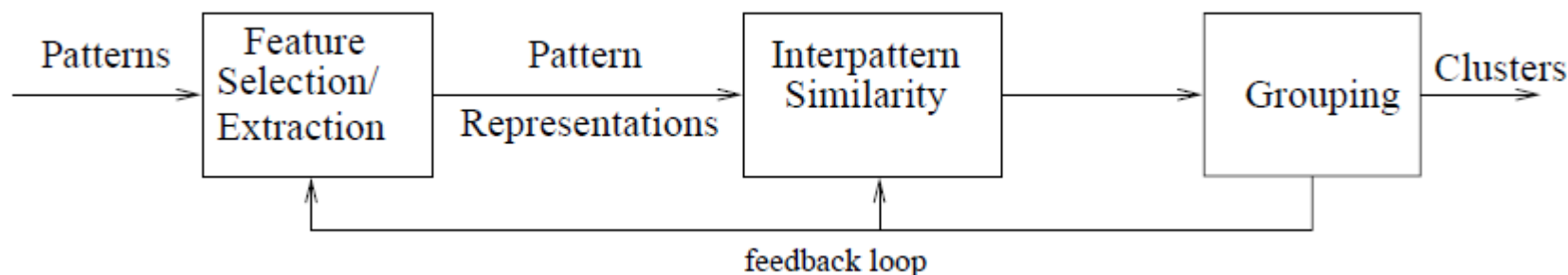
Se pueden agrupar:

- **Secuencias (DNA, RNA)**
  - Ej: Agrupar por similitud/identidad global
  - Ej: Agrupar por presencia de motivos o señales
- **Medidas de expresión de genes**
  - Ej: Agrupar todos los genes que tienen alta expresión
- **Abstracts en PubMed**
  - Ej: Agrupar abstracts en base a número de palabras compartidas
- **Marcadores morfológicos**
  - Ej: Puntos fluorescentes en una imagen de microscopía (por ej para delinar una membrana o cualquier otra estructura celular)
- **O todo a la vez**
  - **Vectores multidimensionales**

# Data clustering example



# Steps in data clustering



## Feature selection:

- Identificar en el dataset el subset de características (features) más informativo para agrupar objetos

## Pattern representations:

- La manera de representar una característica afecta directamente a las medidas de similitud

## Pattern proximity:

- Hay muchas maneras de medir proximidad (distancias). En general se calculan distancias de a pares, para todos los objetos a agrupar

## Clustering:

- Hay muchos algoritmos (estrategias) de clustering

## Cluster validation analysis

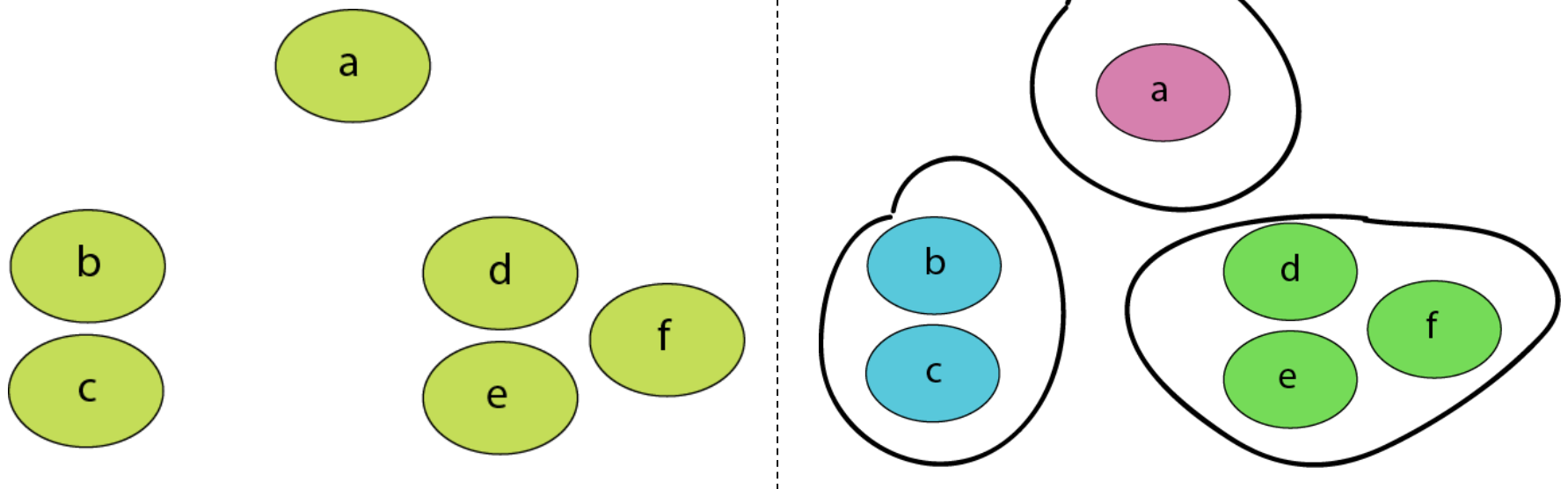
- La estructura de agrupamiento es válida si no puede obtenerse simplemente por azar o no es producto de un artefacto del método

# Objetivo

Clustering algorithm

Original data

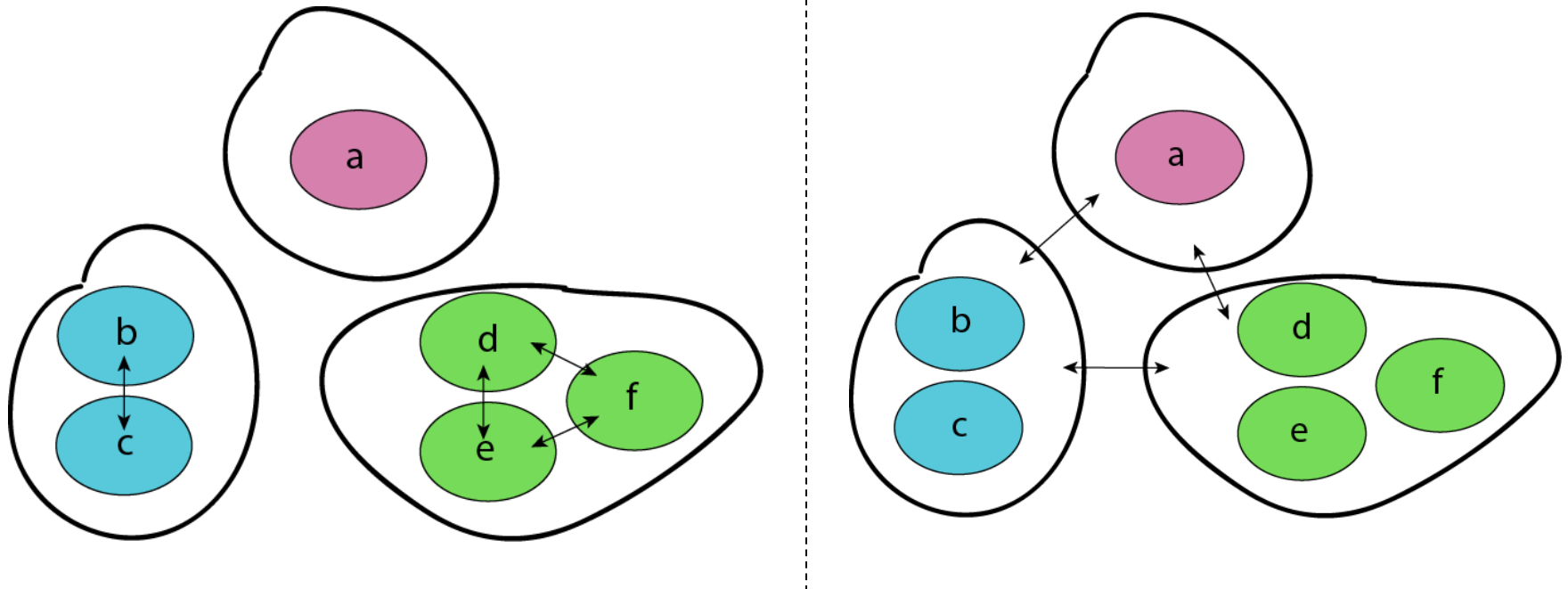
Clustered data



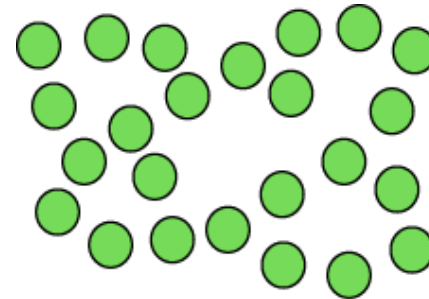
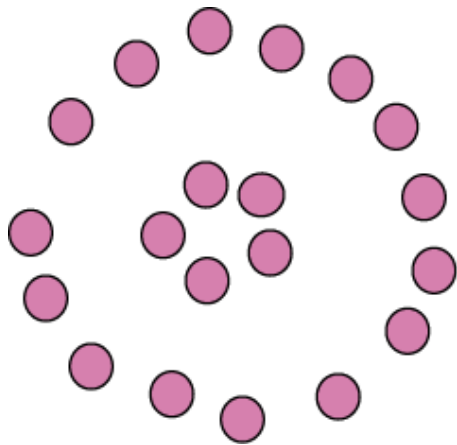
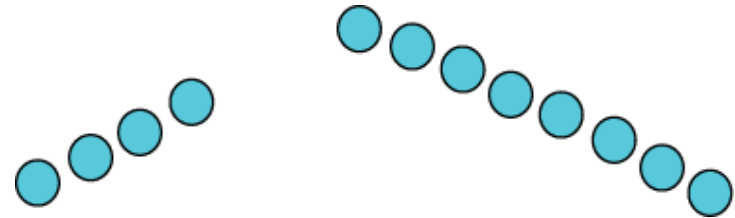
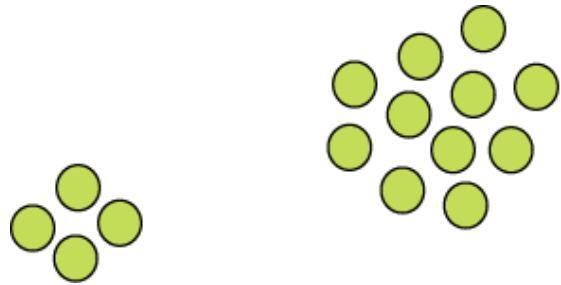
# Objetivo del algoritmo

Minimizar la distancia intracluster

Maximizar la distancia entre clusters

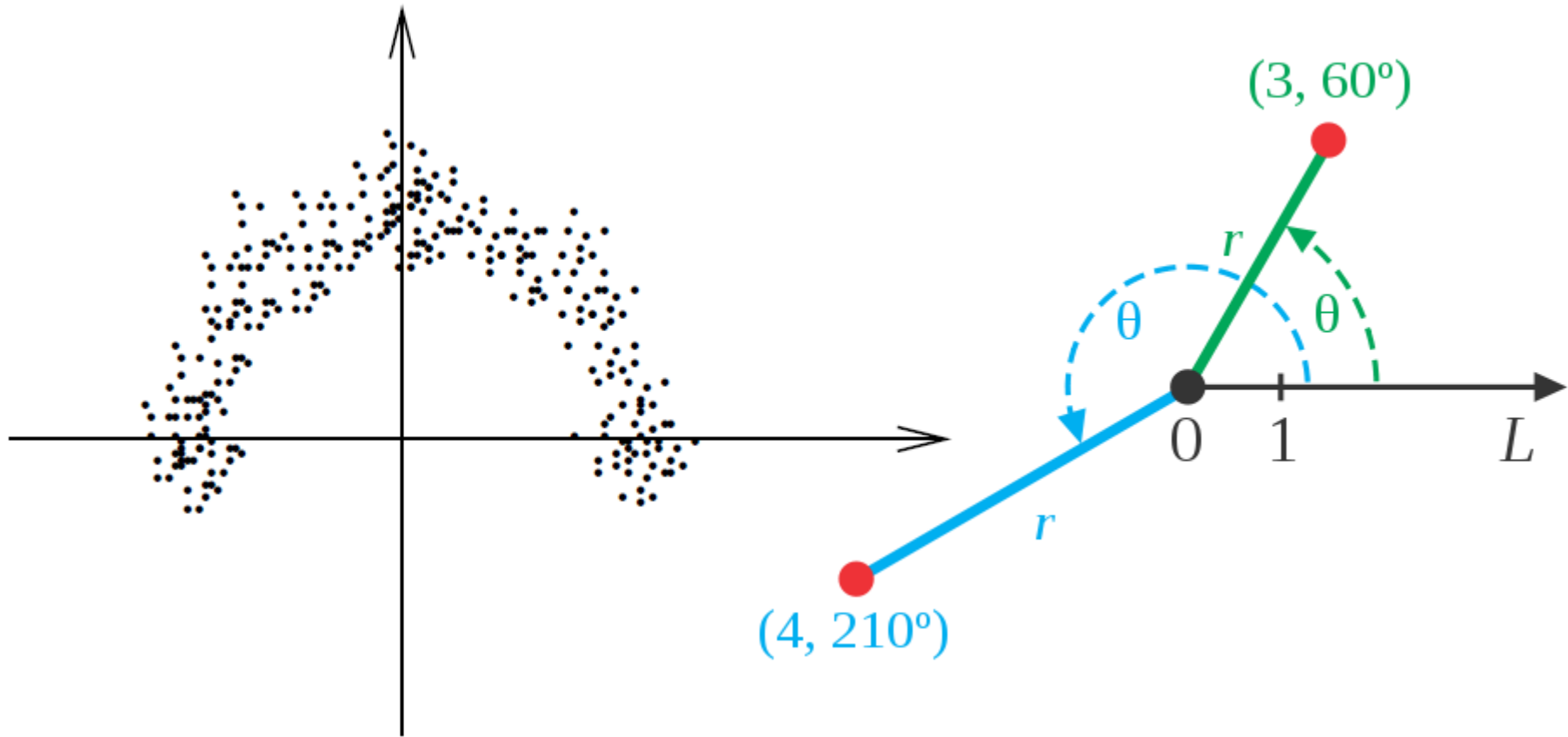


# Formas de los clusters



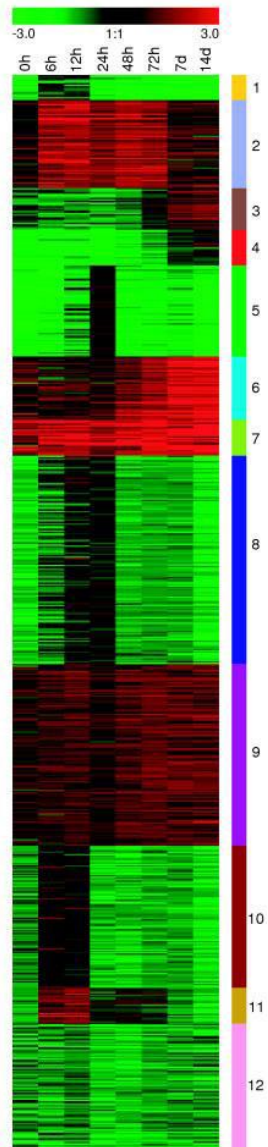


# Clustering: data representation example



**Cluster curvilíneo, donde los puntos están mas o menos equidistantes del origen.** Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review.

# Clustering in bioinformatics: expression data



Expresión de genes a lo largo de un experimento.

No importa tanto si los genes se expresan mucho o poco (ej agrupar por nivel de expresión no tiene sentido)

Importa el comportamiento de cada gen a lo largo de un tratamiento experimental.

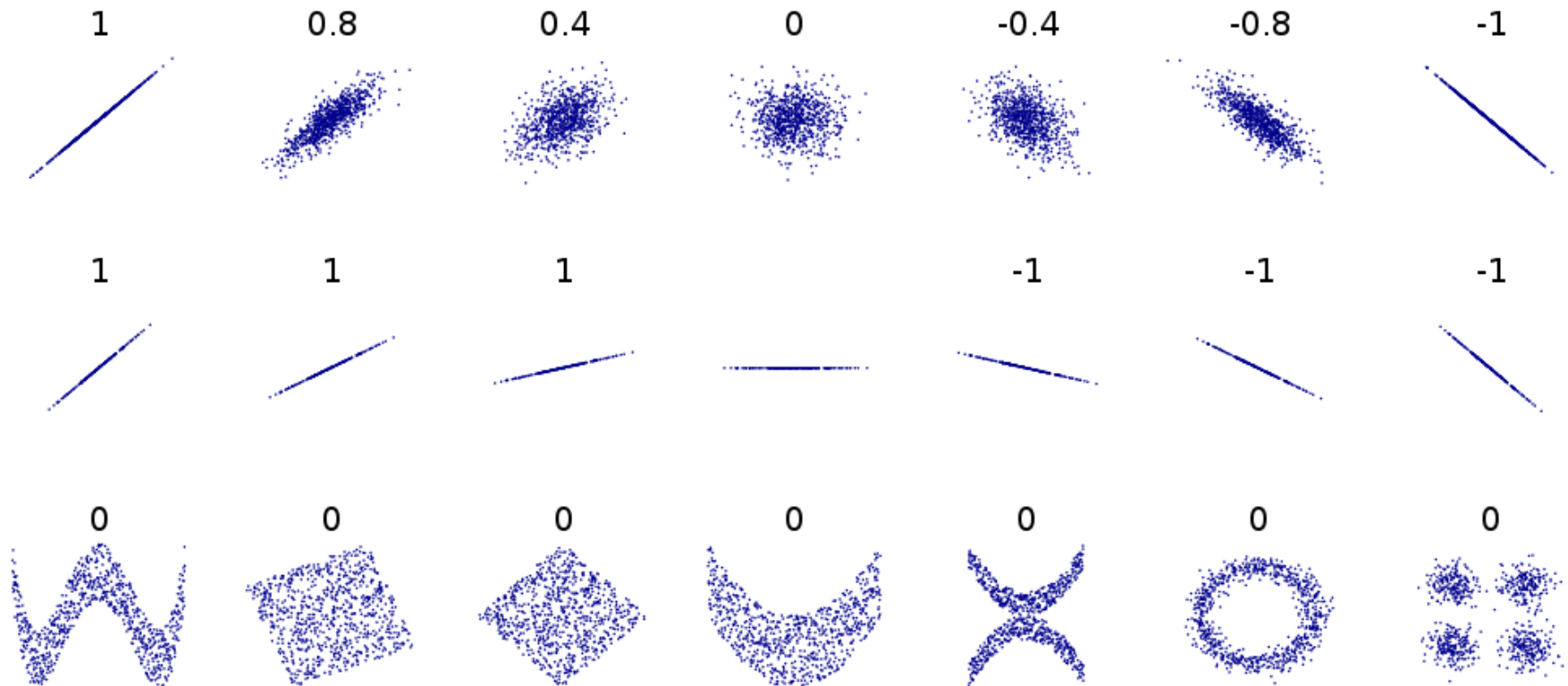
**Correlation distance.** Mide la dependencia entre datos.

CD = 0 si los datos son independientes.

CD = 1 si tienen dependencia.

Clustering of ESTs found to be differentially expressed during fat cell differentiation. Shown is k-means clustering of 780 ESTs found to be more than twofold upregulated or downregulated at a minimum of four time points during fat cell differentiation. ESTs were grouped into 12 clusters with distinct expression profiles. Hackl *et al. Genome Biology* 2005 6:R108 doi:10.1186/gb-2005-6-13-r108

# Correlation distance: Pearson's correlation



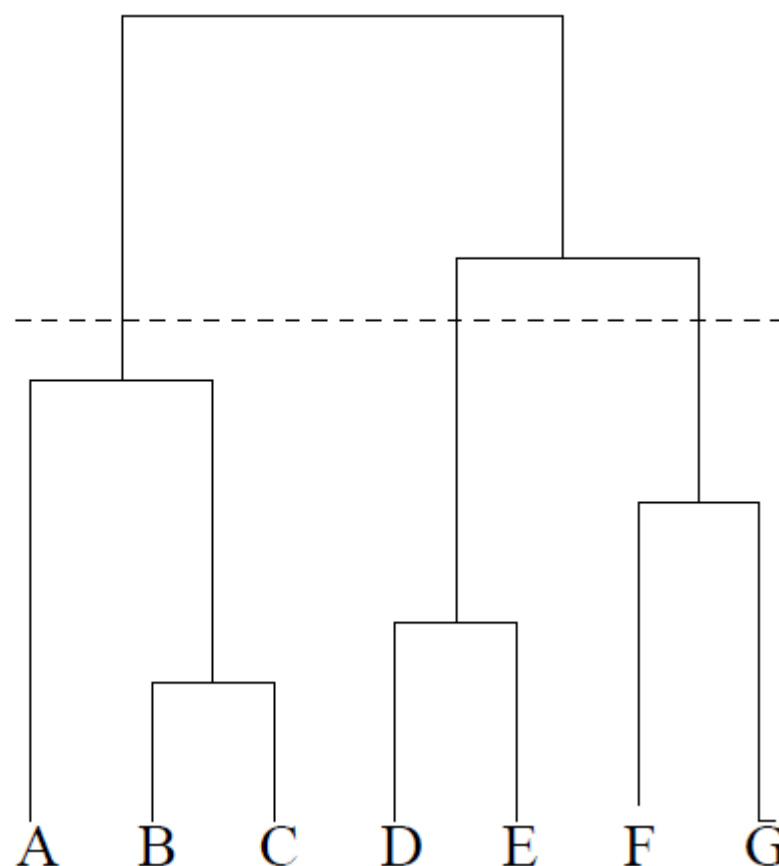
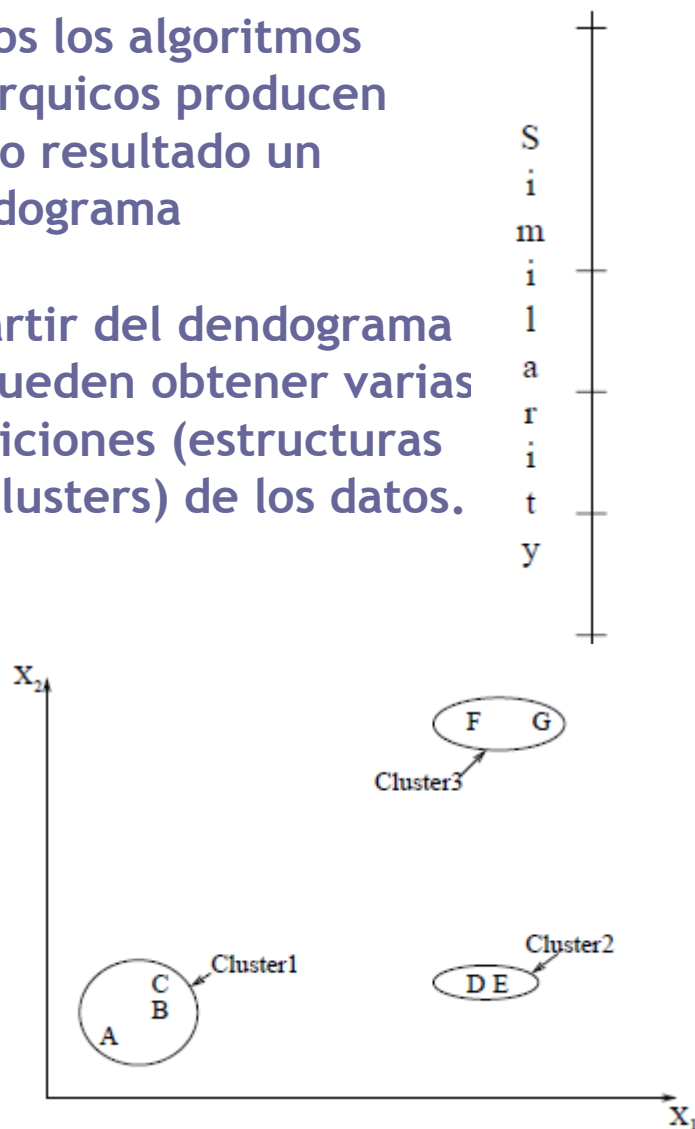
Several sets of  $(x, y)$  points, with the Pearson correlation coefficient of  $x$  and  $y$  for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of  $Y$  is zero.

<http://en.wikipedia.org/wiki/Correlation>

# Clustering algorithms: hierarchical clustering

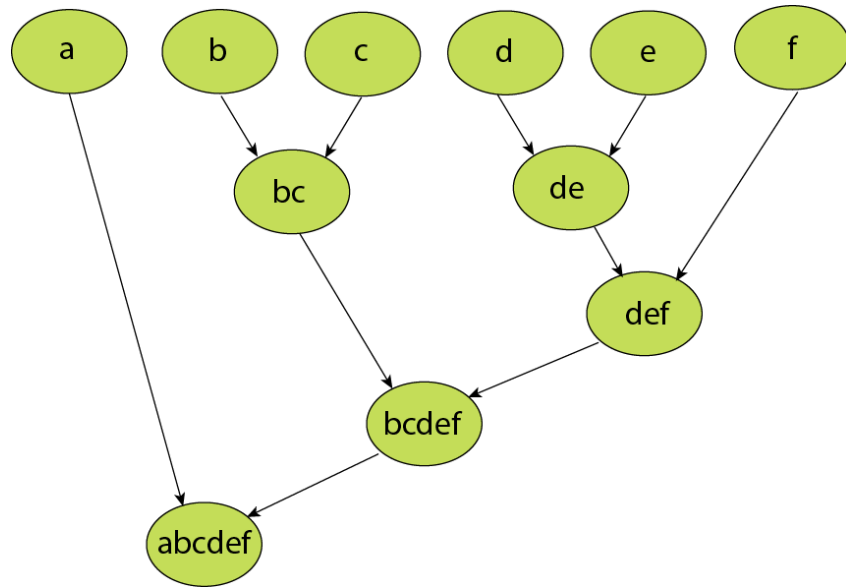
Todos los algoritmos jerárquicos producen como resultado un dendograma

A partir del dendograma se pueden obtener varias particiones (estructuras de clusters) de los datos.

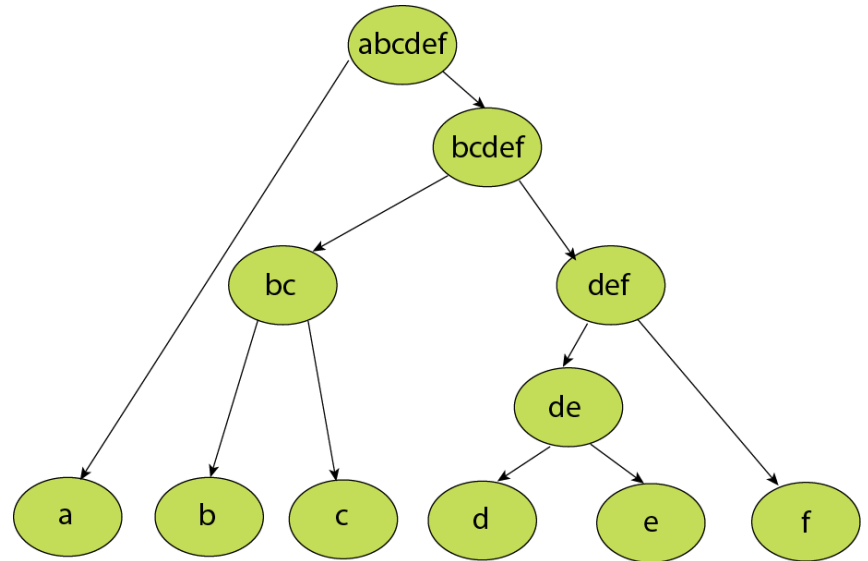


# Estrategias de Clustering: Clustering jerárquico

## Aglomerativo

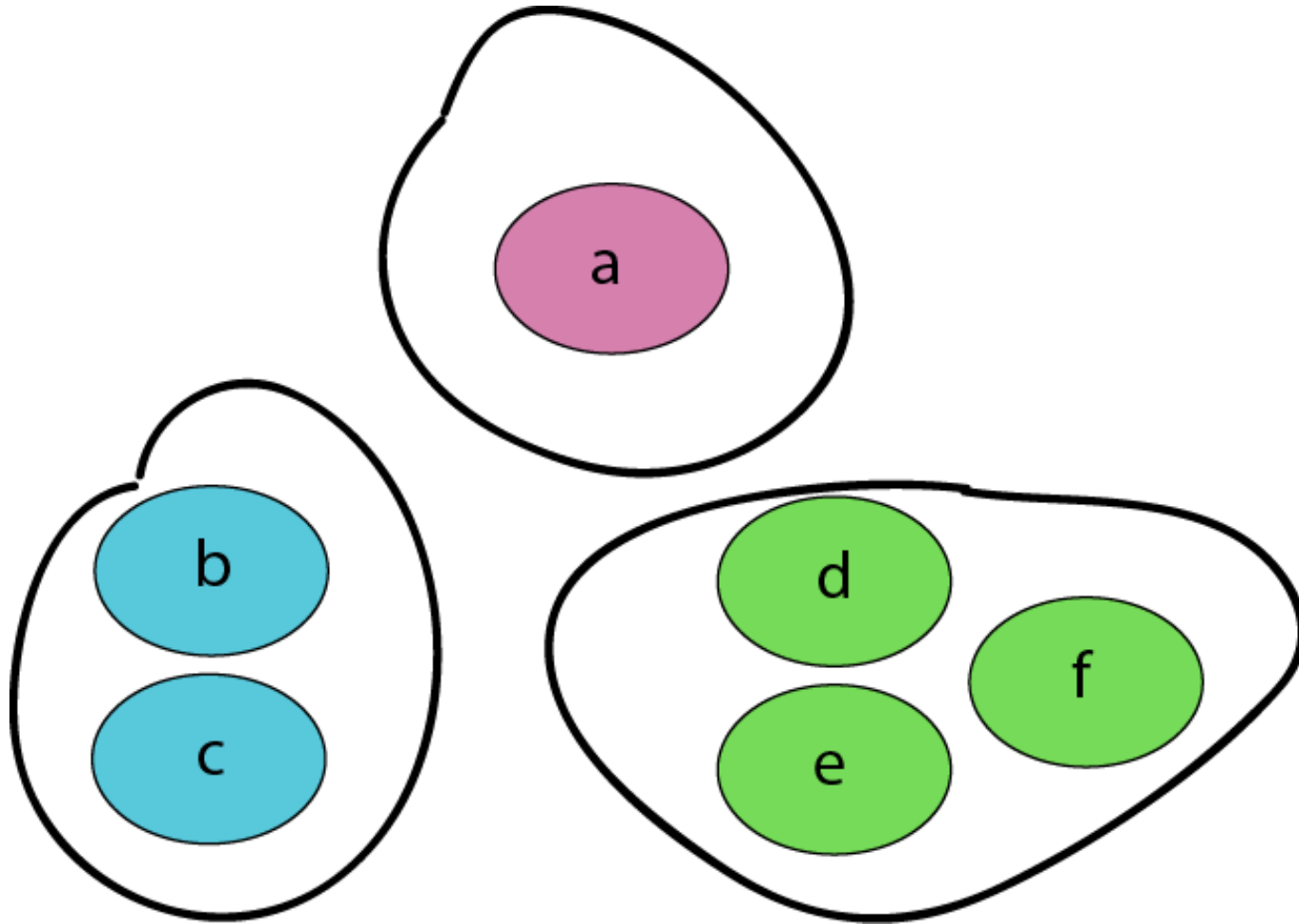


## Divisible



# Estrategias de Clustering: clustering particional

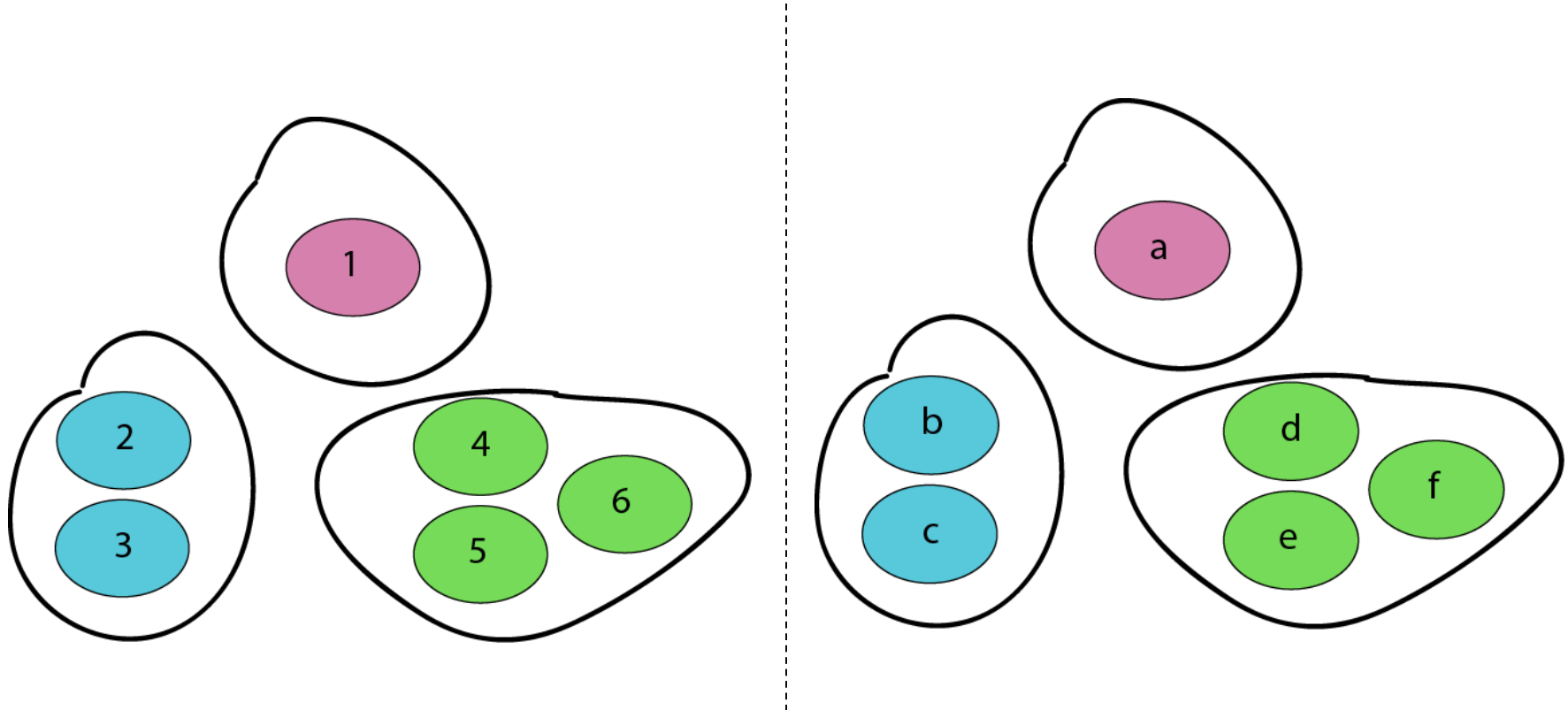
A diferencia de los algoritmos jerárquicos, se obtiene *una única partición de los datos* (una única estructura de clusters)



# Propiedades de los clusters

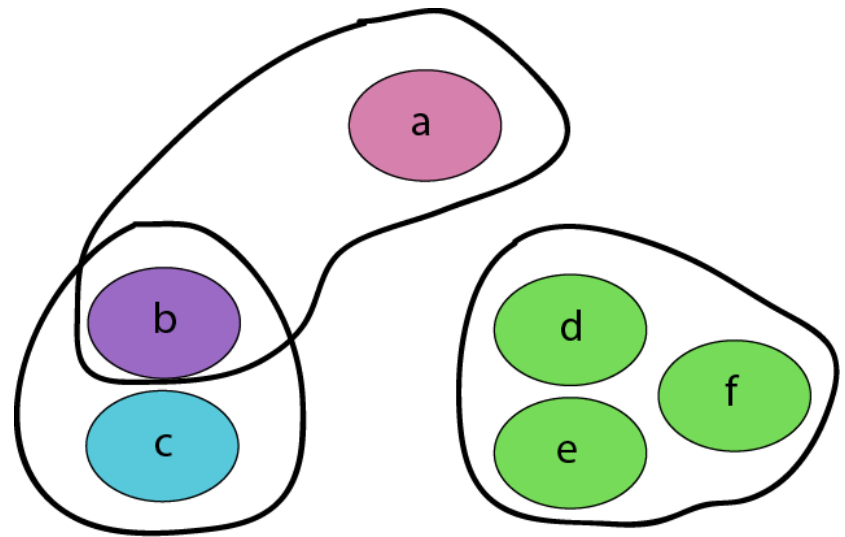
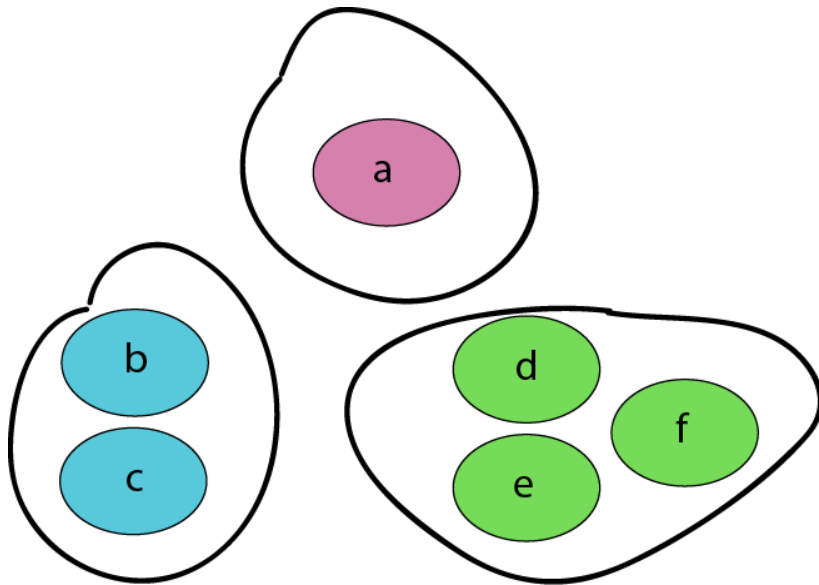
## Numéricos vs. Categóricos

**Cómo calcular distancias entre objetos?**



# Propiedades de los clusters

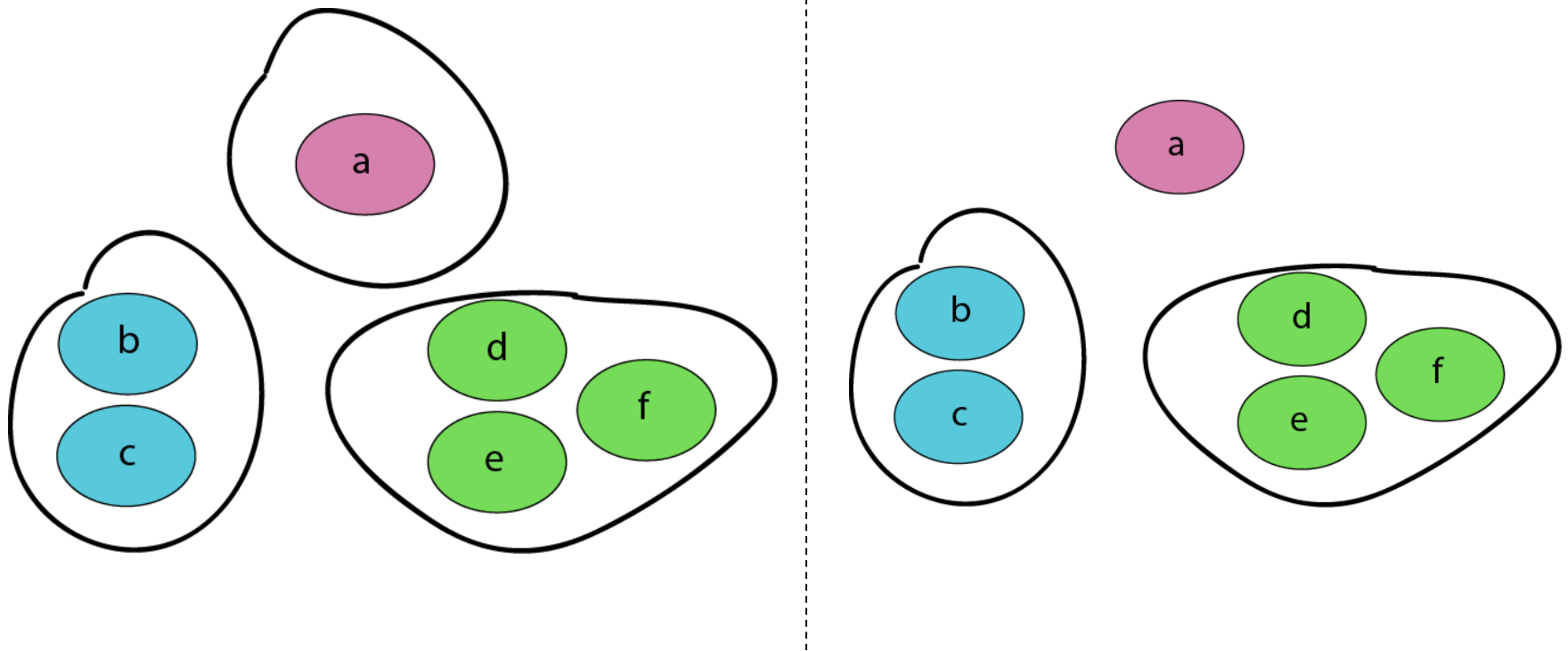
Disjuntos vs. No disjuntos  
(hard) (fuzzy)





# Propiedades de los clusters

## Completos vs. Incompletos



# Single-linkage, Complete-Linkage, Average-Linkage

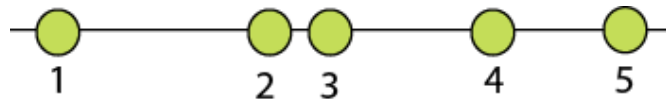
Clustering Jerárquico

Aglomerativo

Si hay un error en algún paso no se puede volver atrás ...

# Hierarchical clustering

Dado un conjunto de  $N$  (5) elementos a ser agrupado y una matriz de distancia (o similitud) de  $N \times N$ :



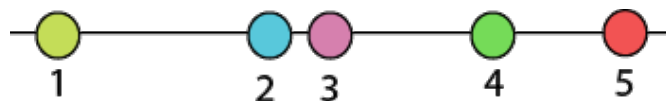
$d$	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

# Hierarchical clustering

Comenzar por asignar cada ítem a un cluster.

Tenemos 5 clusters

**En este paso**, las distancias entre los clusters son las mismas que entre los elementos de cada cluster

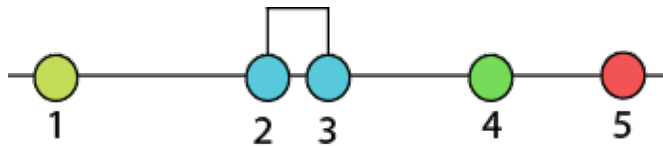


$d$	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

# Hierarchical clustering

Encontrar el par más cercano de clusters y unirlo en un único cluster.

Tenemos 4 clusters



<i>d</i>	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

# Hierarchical clustering

Calcular las distancias entre el nuevo cluster y los viejos clusters

En **single-linkage** la distancia que se usa es la **mínima** entre distintos elementos de un cluster

Los elementos se agrupan **siempre** encontrando la **mínima** distancia en la matriz

<i>d</i>	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

<i>d</i>	1	2-3	4	5
1	0	5	10	13
2-3	5	0	4	7
4	10	4	0	3
5	13	7	3	0

# Hierarchical clustering

En el algoritmo **complete-linkage** la distancia que se usa en la nueva matriz es la **máxima** entre distintos elementos de un cluster

Los elementos se agrupan **siempre** encontrando la **mínima** distancia en la matriz

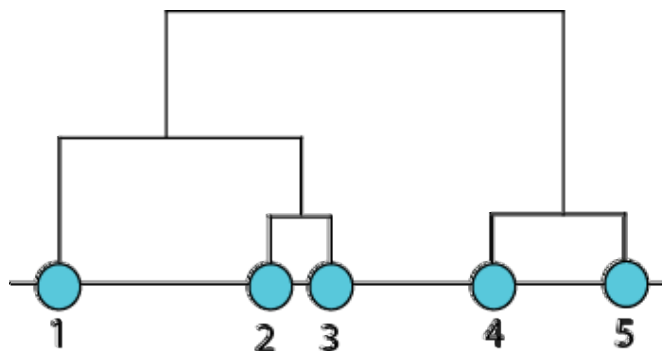
Y en **average-linkage**?

<i>d</i>	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

<i>d</i>	1	2-3	4	5
1	0	6	10	13
2-3	6	0	5	8
4	10	5	0	3
5	13	8	3	0

# Single-linkage

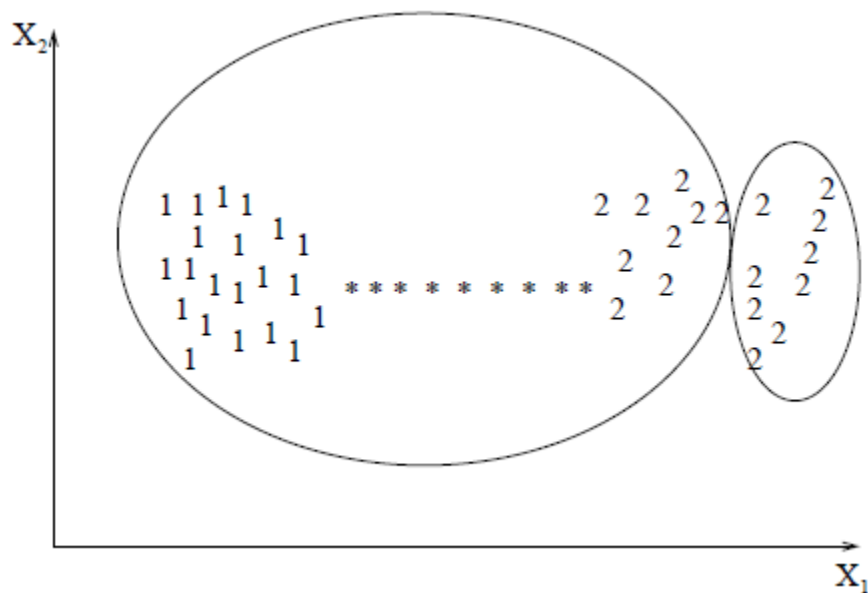
Repetir los pasos 2 y 3 hasta que todos los elementos se encuentren en el mismo cluster de tamaño  $N$  (5)



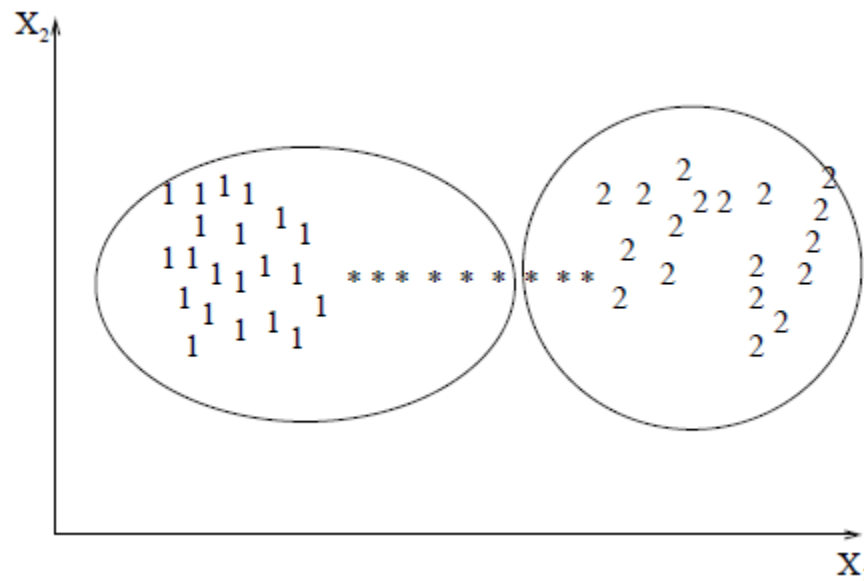


# Diferencias entre single vs complete linkage

Single-link



Complete-link



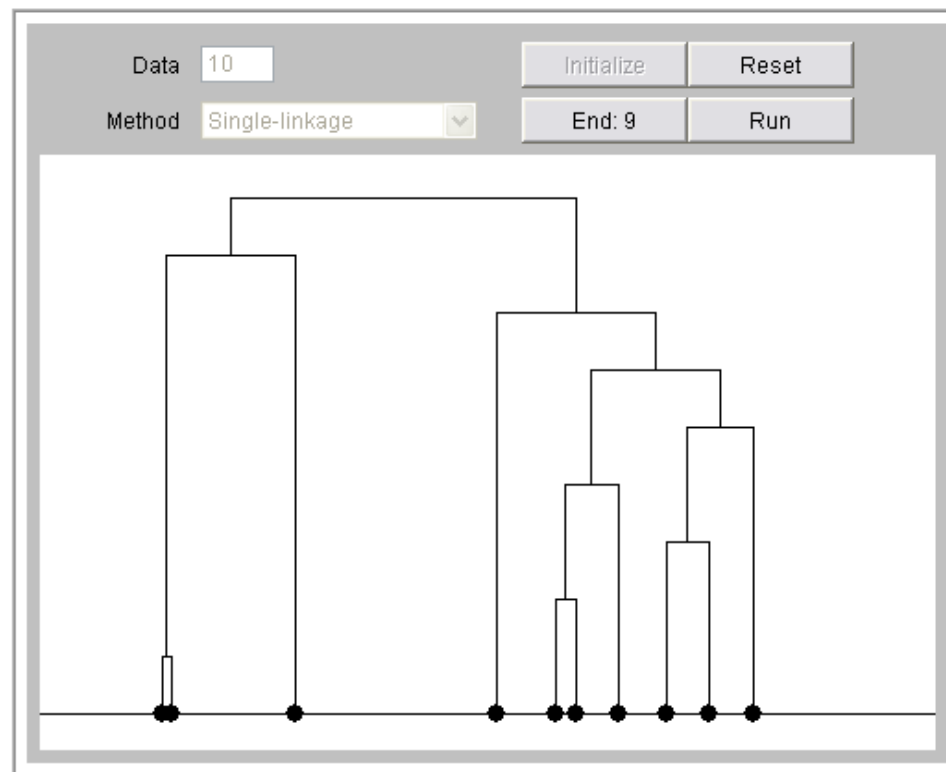
**Ejemplo: dataset compuesto por elementos pertenecientes a dos clases, conectadas por una cadena de datos ruidosos.** Tomado de Jain AK, Murty MN, Flynn PJ (1999)

Data clustering: a review.

# Hierarchical clustering: interactive demo

## Hierarchical Clustering - Interactive demo

This applet requires Java Runtime Environment version 1.3 or later. You can download it from the [Sun Java website](http://www.sun.com/java/javase/1.3/download.html).



[http://home.dei.polimi.it/matteucc/Clustering/tutorial\\_html/AppletH.html](http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html)

# Otras Variantes

**Hierarchical clustering techniques applied to phylogenetic reconstruction:**

**UPGMA (Unweighted Pair Group Method with Arithmetic Mean)**

Usado para reconstruir filogenias

Usa la media aritmética (average-link)

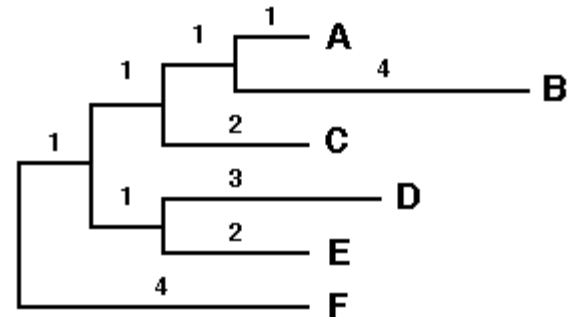
Distancias ultramétricas

**Neighbor-joining**

Usado para reconstruir filogenias

Usa la media aritmética

Las distancias son aditivas



# Clustering algorithms: K-means

Es muy rapido!

Particional

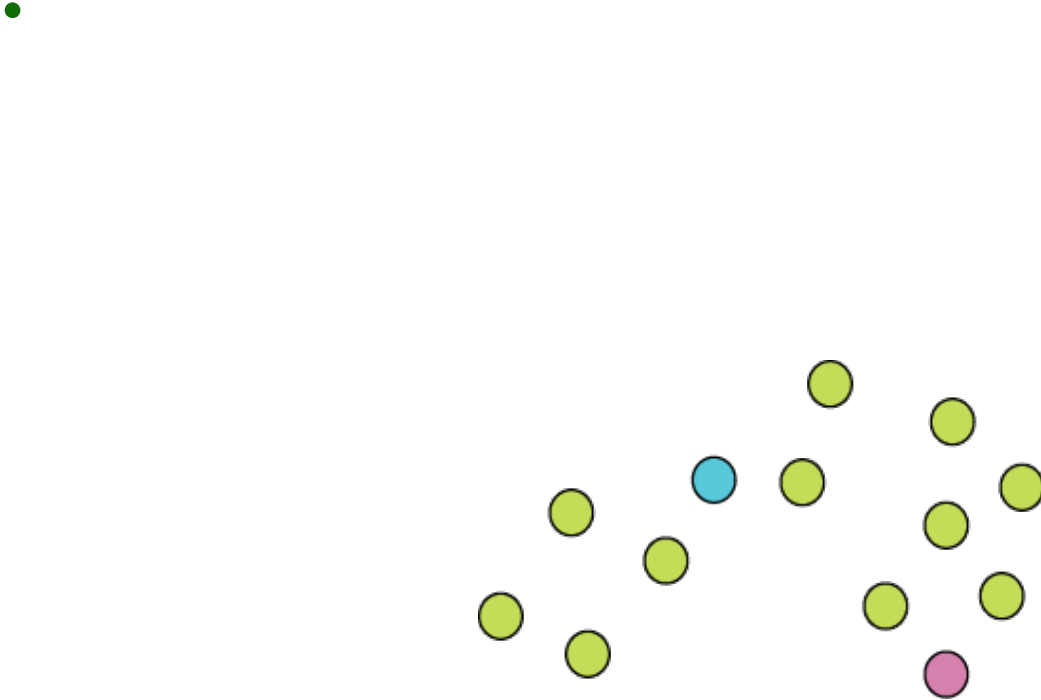
Usa Distancia euclídea

Necesita el valor de  $k$  (Nro de clusters)

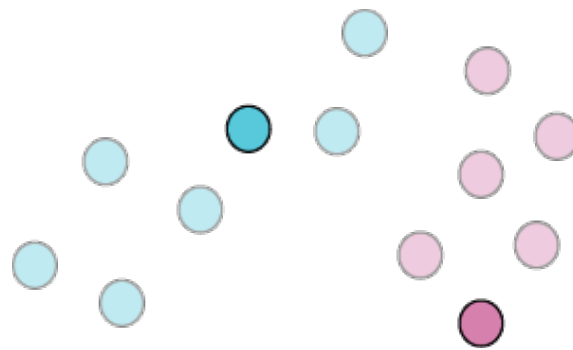
Util para búsqueda de prototipos

Sensible a outliers

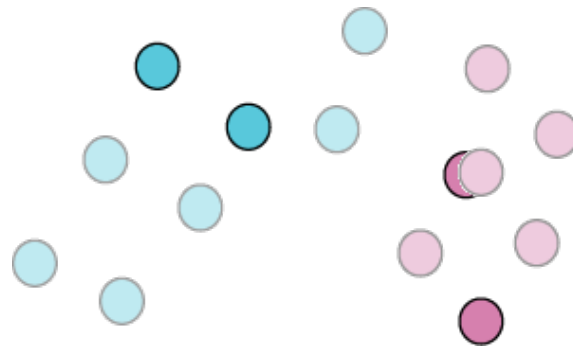
# K-means



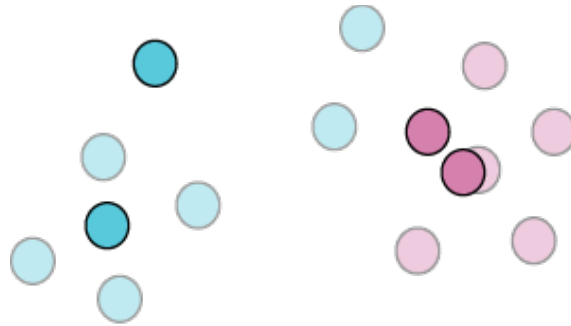
# K-means



# K-means



# K-means

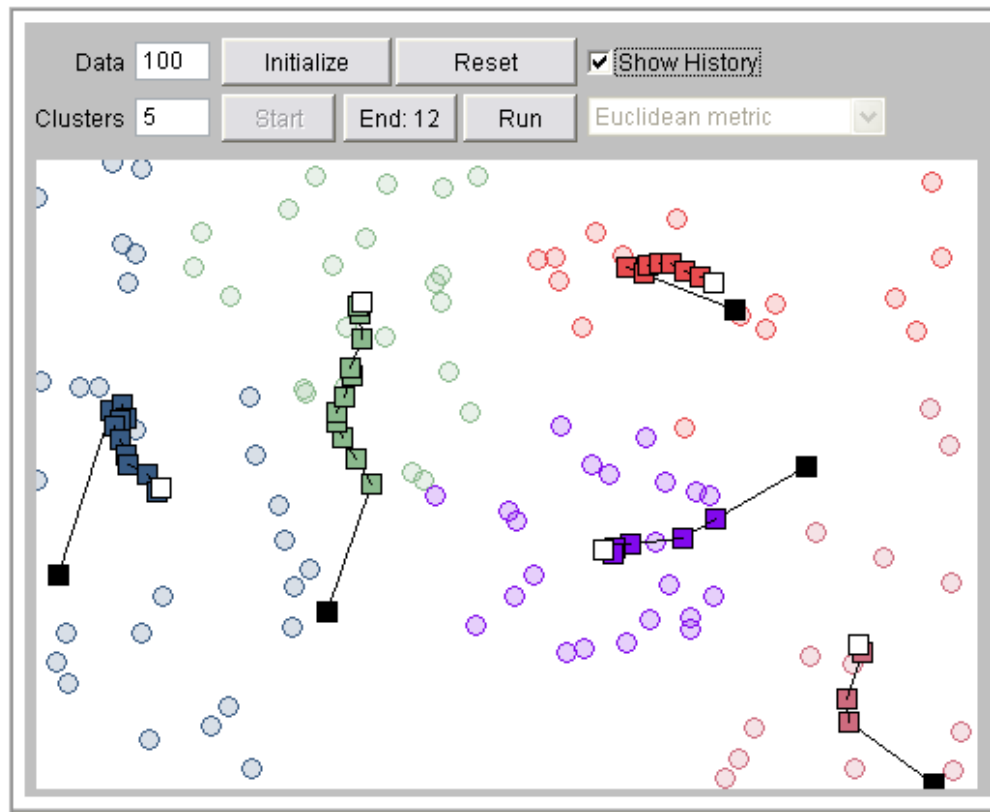




# K-means

## K-means - Interactive demo

This applet requires Java Runtime Environment version 1.3 or later. You can download it from the [Sun Java website](http://www.sun.com/javase/6/other/javase_downloads/javase6_jre_downloads/javase6_jre6-windows-i586-jdk-6u5-windows-i586-jre.exe).



[http://home.dei.polimi.it/matteucc/Clustering/tutorial\\_html/AppletKM.html](http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html)

# Clustering: hay que analizar los resultados!

Dado un set de datos al azar, **sin ninguna estructura**, los algoritmos de clustering siempre encuentran agrupamientos!

**Gold Standard:** los agrupamientos, corresponden a categorías naturales? (Validación externa)

Cuán bien están maximizados y minimizados la similitud intra-cluster y la disimilaridad inter-cluster? (Validación interna)

# Validación interna: Silhouette Index

$$\frac{1}{k} \sum_k \left( \frac{1}{|c_k|} \sum_{\vec{x}_i \in c_k} \frac{b(\vec{x}_i) - a(\vec{x}_i)}{\max[b(\vec{x}_i), a(\vec{x}_i)]} \right)$$

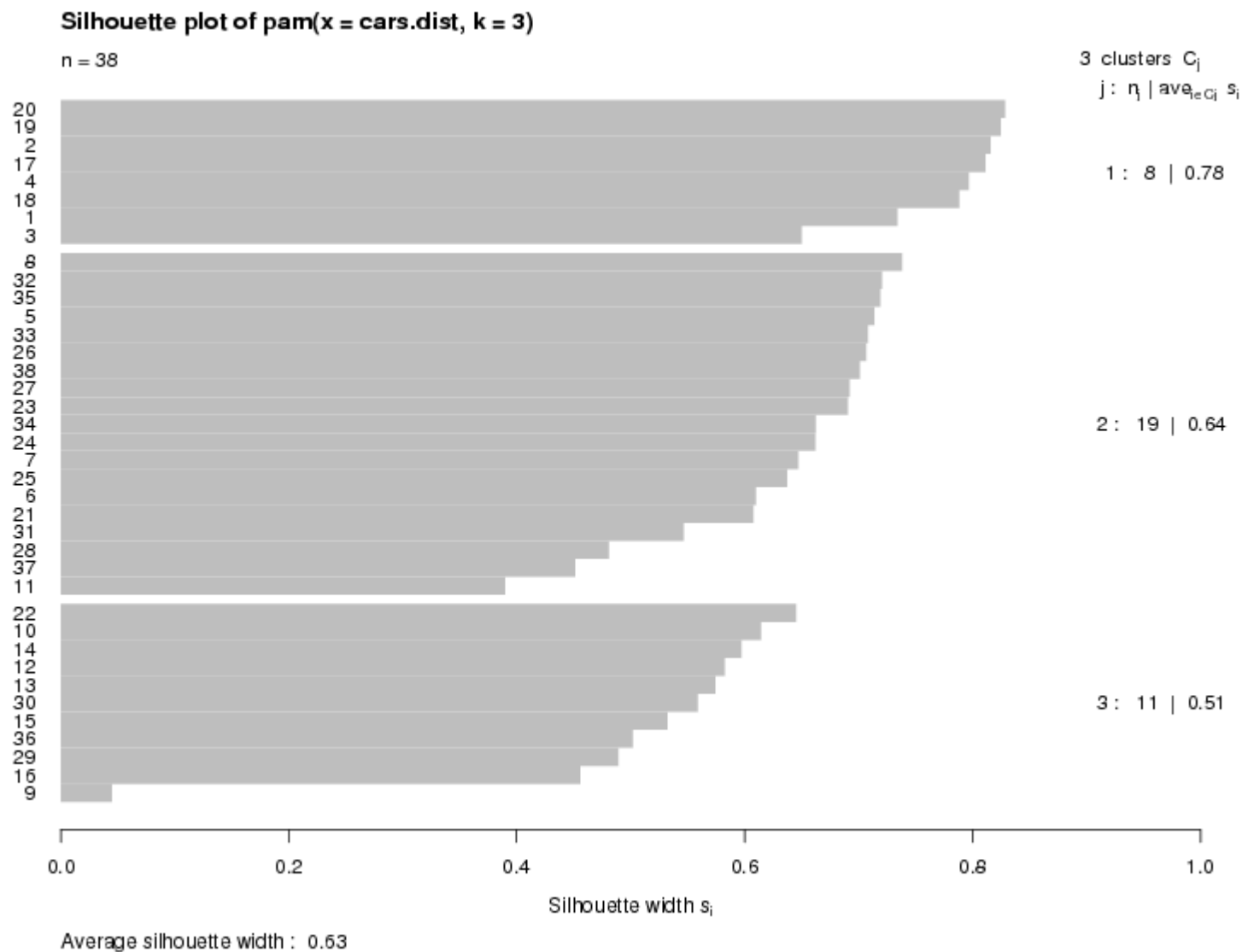
$a(\vec{x}_i)$  average distance from  $\vec{x}_i$  to other instances in same cluster

$b(\vec{x}_i)$  average distance from  $\vec{x}_i$  to instances in next closest cluster

**SI = 1 means element is well placed in its cluster**

**SI = 0 means element might well be placed in another cluster**

# Validación interna: Silhouette Index



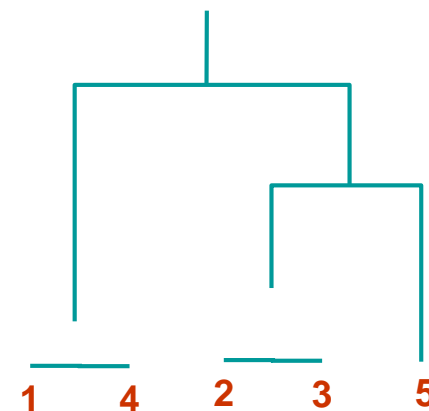
- **Agradecimientos: Dra. Rocío Romero Zaliz**  
(Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada)
- **Bibliografía adicional:**
  - Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. (PDF disponible en la página de la materia)
  - Witten IH, Frank E, Hall MA (2011) Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition.

**Filogenia**  
**Reconstrucción filogenética**  
**Inferencia de filogenias**

**Fernán Agüero**  
**Instituto de Investigaciones Biotecnológicas**  
**UNSAM**

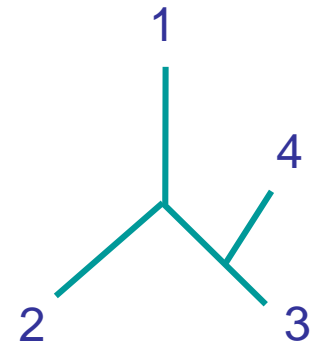
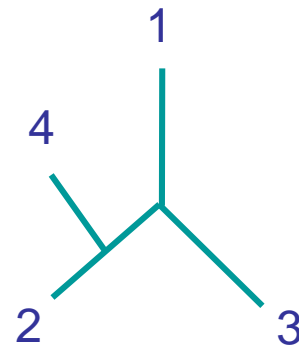
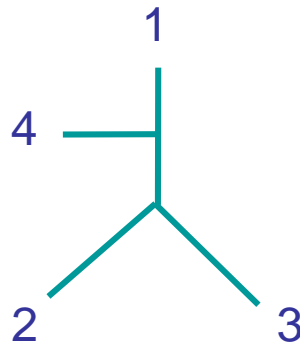
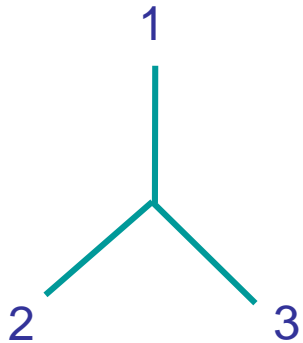
# Filogenia

- Una filogenia es un árbol que describe la secuencia de eventos que llevó a producir los caracteres que observamos en la actualidad
- Es una hipótesis!
- Los eventos pasados son desconocidos. *Se infieren*
- Un árbol es un grafo
  - Nodos y ejes
- En particular:
  - Los nodos exteriores (hojas del árbol) son los eventos observados (especies actuales)
  - Los nodos internos son los eventos (ancestros) postulados
  - La longitud de los ejes (ramas) representa el tiempo de evolucion entre nodos



# Espacio de árboles posibles

Taxa	Rooted trees	Unrooted trees
3	3	1
4	15	3
5	105	15
-	-	-
7	10,395	954





- **Basados en**
  - **Distancias**
  - **Parsimonia**
  - **Verosimilitud (likelihood)**

- **Cómo inferir la filogenia?**

- Definir los caracteres a seguir
- Construir una matriz de distancias
- Seleccionar un algoritmo para reconstruir la filogenia a partir de los datos de distancias

- **Caracteres y estados**

- Los caracteres deben evolucionar en forma independiente
- Los estados observados comparten un origen común

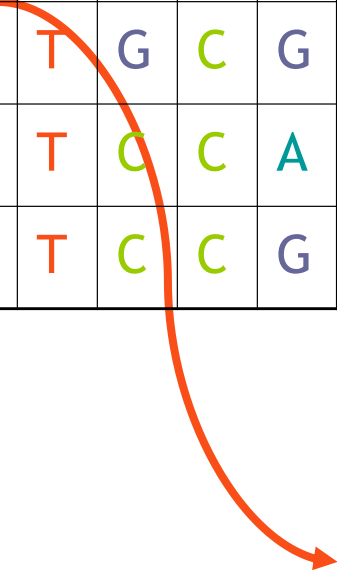
Para secuencias de ADN un caracter corresponde a una posición en la secuencia y los estados posibles, son los nucleótidos A, T, C, G.

# Tipos de caracteres

<b>MORFOLÓGICOS</b> Medidas Corporales Medidas Parciales Presencia de estructuras	<b>MOLECULARES</b> Hibridación DNA-DNA RFLP Secuencias (DNA ó Proteínas)
<b>CONTINUOS</b> Medidas Corporales Medidas Parciales Hibridación de DNA-DNA	<b>DISCRETOS</b> Presencia de estructuras RFLP Secuencias (DNA ó Proteínas)

# Matriz de caracteres

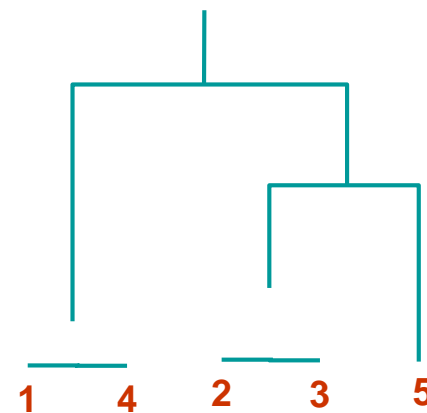
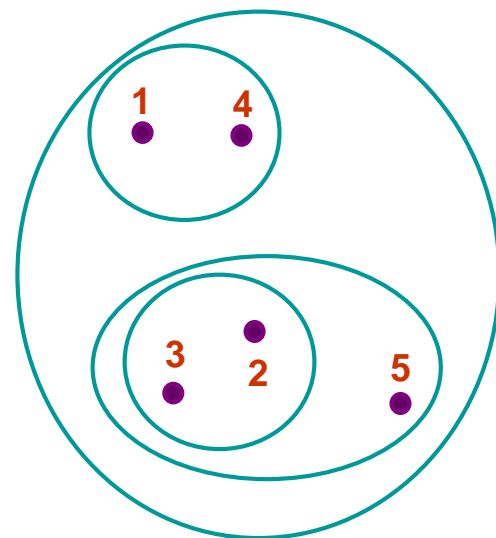
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G



	Sp. 1	Sp. 2	Sp. 3	Sp. 4
Sp. 1	0			
Sp. 2	4	0		
Sp. 3	5	5	0	
Sp. 4	6	4	2	0

# Algoritmos basados en distancias

- Los pares de secuencias más cercanos (neighbors) comparten un ancestro común y están unidos a él por ramas
- El objetivo del método es encontrar un árbol que acomode a todos los *vecinos* correctamente
- El largo de las ramas tiene que concordar con los datos de distancia
- Usan métodos de *clustering* para agrupar *vecinos*

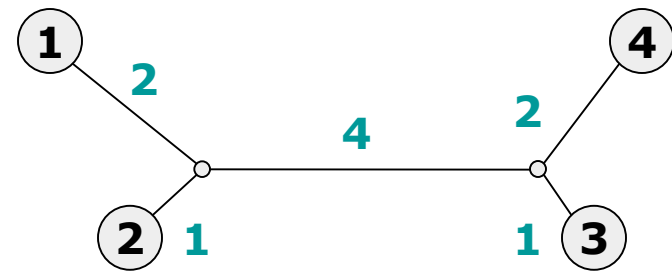


# Distintos tipos de distancias

- **Aditivas**

- La suma de las longitudes de las ramas de dos especies con su nodo ancestral es igual a la distancia calculada entre las especies

	Sp. 1	Sp. 2	Sp. 3	Sp. 4
Sp. 1	-			
Sp. 2	3	-		
Sp. 3	7	6	-	
Sp. 4	8	7	3	-



- **Ultramétricas**

- Cada ancestro común está equidistante de sus descendientes
- Util para visualizar similitud en contextos no evolutivos



# Máxima parsimonia

- Predicen el árbol (o árboles) que minimizan el número de cambios (o pasos) que es necesario hacer para generar la variación observada entre las secuencias
- También conocido como *método de evolución mínima*

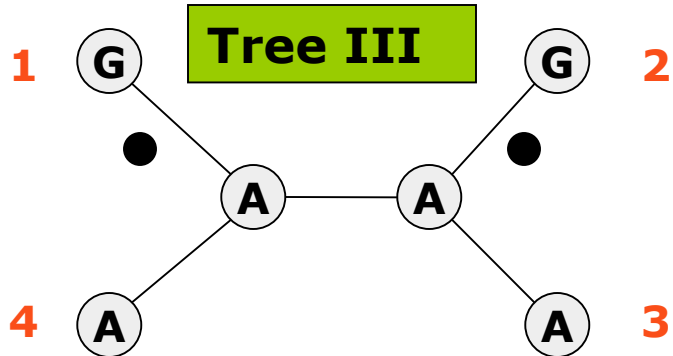
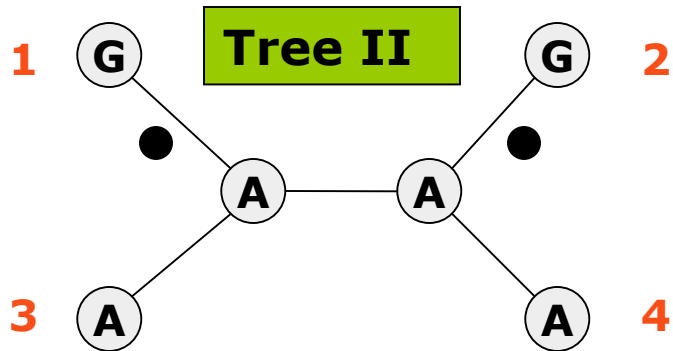
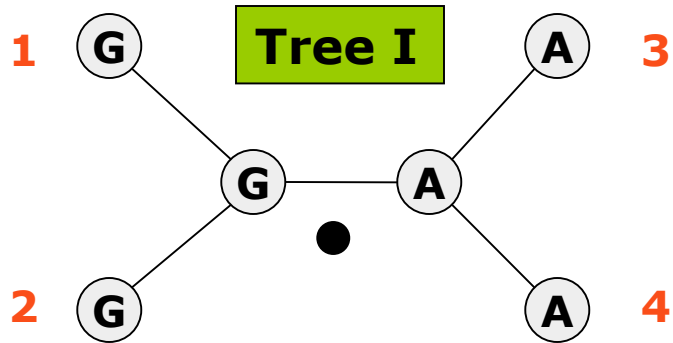
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

- **Ejemplo**

- Para ser informativo un sitio debe tener dos estados presentes en al menos dos especies
- Sitios no informativos: 1, 2, 3, 4, 6 y 8
- Sitios informativos: 5, 7 y 9
- Sólo se analizan los sitios informativos



# Máxima parsimonia: ejemplo



- Hay 3 árboles posibles (sin raíz) para describir la evolución de 4 especies
- Menor número de cambios para explicar la evolución: árbol 1 (1 cambio)
- El mismo análisis se repite para cada uno de los sitios informativos
- El resultado es el árbol que provee el menor número de pasos para acomodar los datos en los sitios informativos (el más parsimonioso)

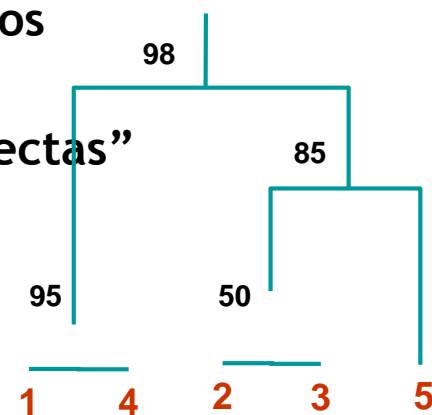
# Máxima parsimonia: detalles

- Asume que la velocidad de evolución es similar en todas las ramas
  - La inferencia obviamente falla cuando esto no se cumple
  - Ejemplo: cambio de G a A en forma independiente en dos especies
    - Especie 1: G > A
    - Especie 2: G > C > T > G > C > A
- Se pueden asignar puntajes a los árboles
  - En lugar de contar cambios se pueden asignar distintos valores a los cambios (por ejemplo usando una matriz)
- A diferencia de los métodos de distancia, el método permite obtener la secuencia postulada de cualquier ancestro

- **Maximum likelihood**
  - Similar al método de máxima parsimonia: usa todas las columnas del alineamiento, considera todos los árboles posibles
  - Usa probabilidades

- **Bootstrap test**

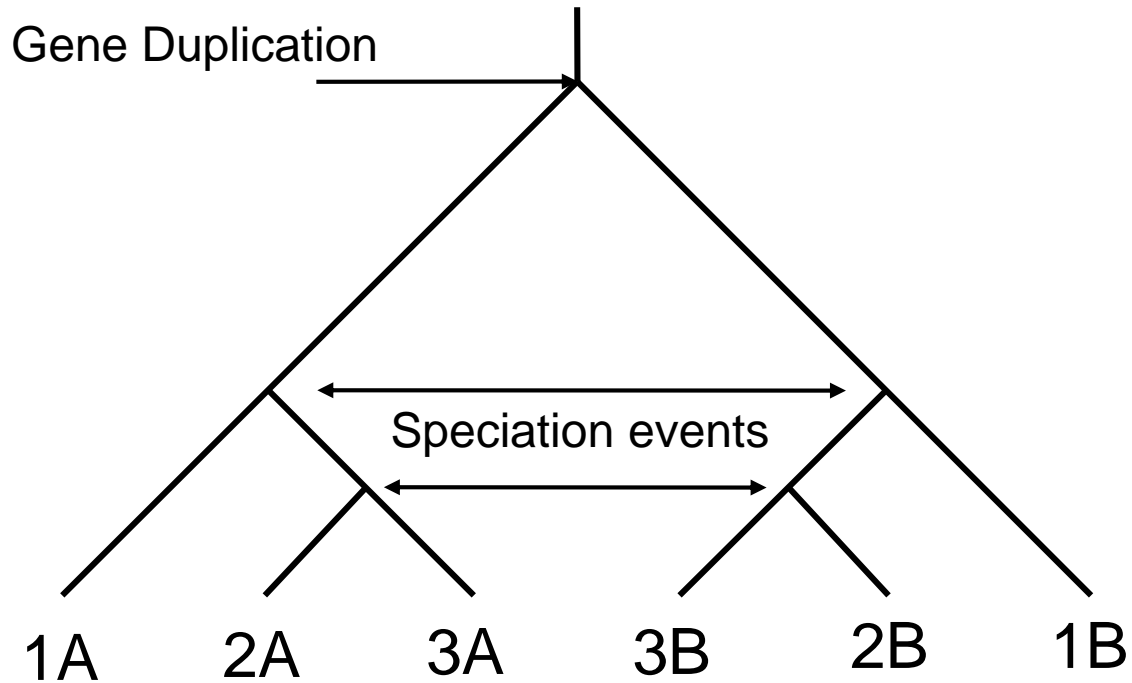
- Bootstrap resampling technique (Efron 1982)
- Dado un número de secuencias  $M$  de longitud  $N$  (un alineamiento), y un árbol calculado por un método cualquiera, se genera un nuevo set de secuencias  $M'$  en el cual  $N'$  bases/residuos elegidos al azar son reemplazados, también al azar.
- En base a este nuevo set  $M'$  se recalcula el árbol utilizando el mismo método y se comparan las topologías del árbol.
- Esto se repite varias veces (100, 1000 repeticiones) y se calcula, para cada rama un valor de bootstrap
- Bootstrap value: % de veces que la rama aparece en los distintos árboles
- Bootstrap values  $\geq 95\%$  corresponden a ramas “correctas”



- **Jackknife**

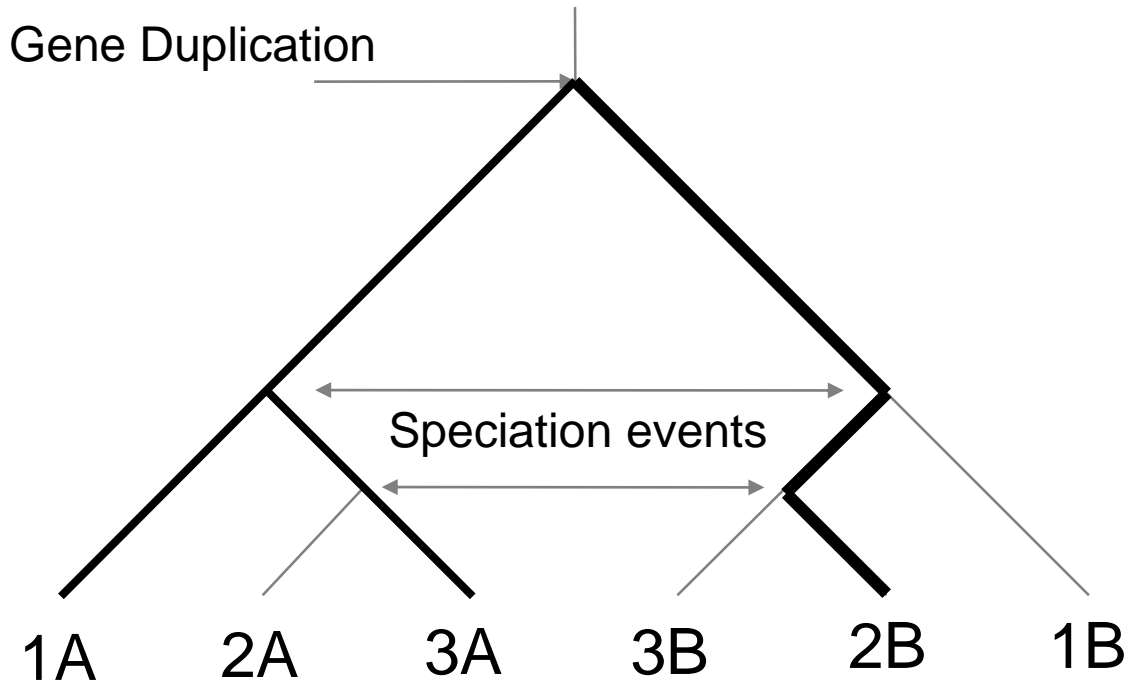
- Muy similar al test de bootstrapping
- Se generan nuevos data sets por muestreo parcial del original
- Usualmente se muestrea el 50% de los datos originales
- Se rehacen los árboles y se verifica la topología
- Se hacen varios re-muestreos (100-1000 veces)
- Se construye un árbol consenso con valores de confianza para cada rama

# Selección de secuencias: parálogos



# Problemas con parálogos

Si sólo usamos



# Hill-climbing



- **Phylip**
  - Unix, linea de comando. Gratuito.
  - DNA, Proteinas,
  - Distancias, Parsimonia
  - Bootstrap, Jackknife
- **PAUP**
  - Similar a Phylip. Comercial. Interfase gráfica, linea de comando.
- **PhyML**
  - Maximum likelihood