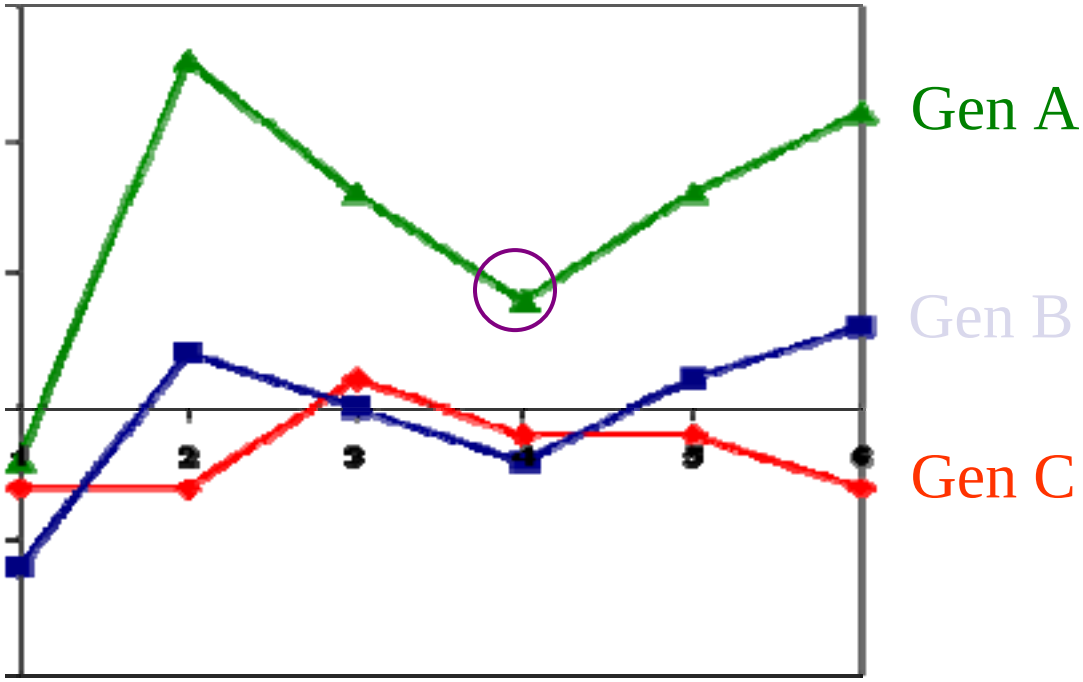
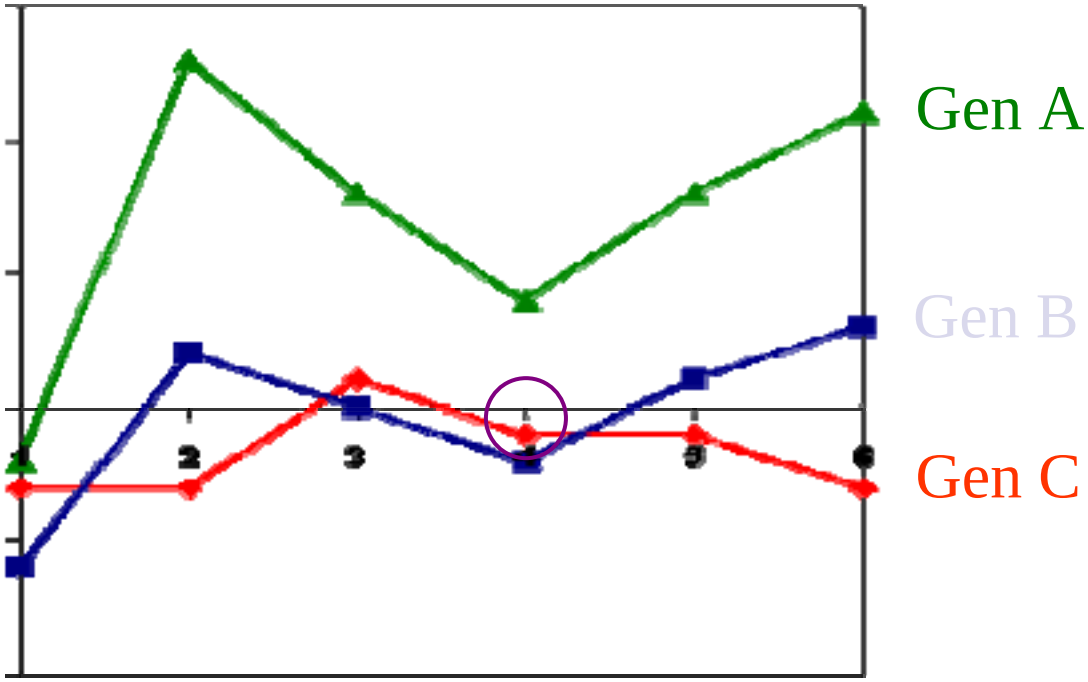
Métodos:

- **Fijar a un valor predeterminado: normalmente 0**
- **Sustituir el valor perdido por la media de toda la columna (experimento)**
- **Sustituir el valor perdido por la media de toda la fila (gen)**
- **Interpolación local pesada**



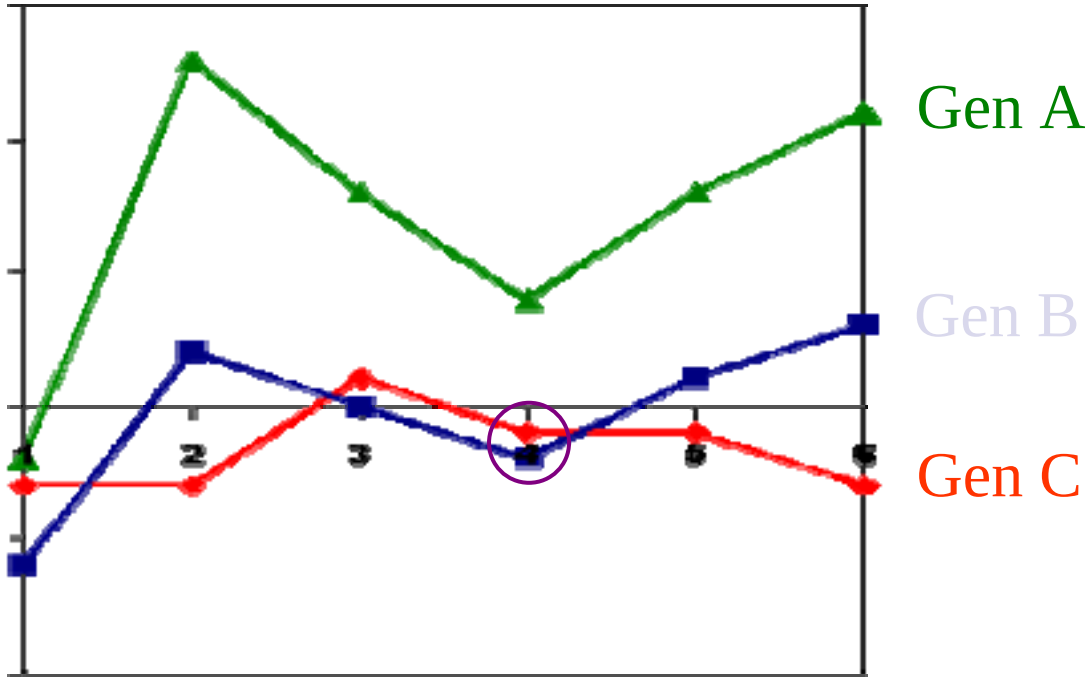
Métodos:

- **Fijar a un valor predeterminado: normalmente 0**
- **Sustituir el valor perdido por la media de toda la columna (experimento)**
- **Sustituir el valor perdido por la media de toda la fila (gen)**
- **Interpolación local pesada**



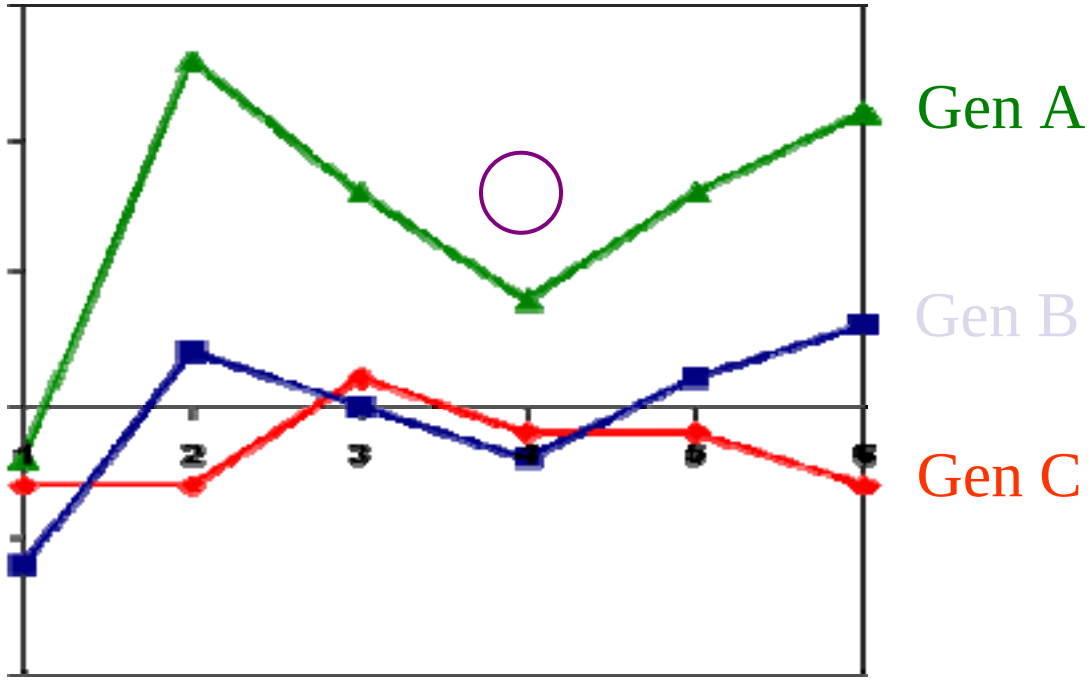
Métodos:

- **Fijar a un valor predeterminado: normalmente 0**
- **Sustituir el valor perdido por la media de toda la columna (experimento)**
- **Sustituir el valor perdido por la media de toda la fila (gen)**
- **Interpolación local pesada**

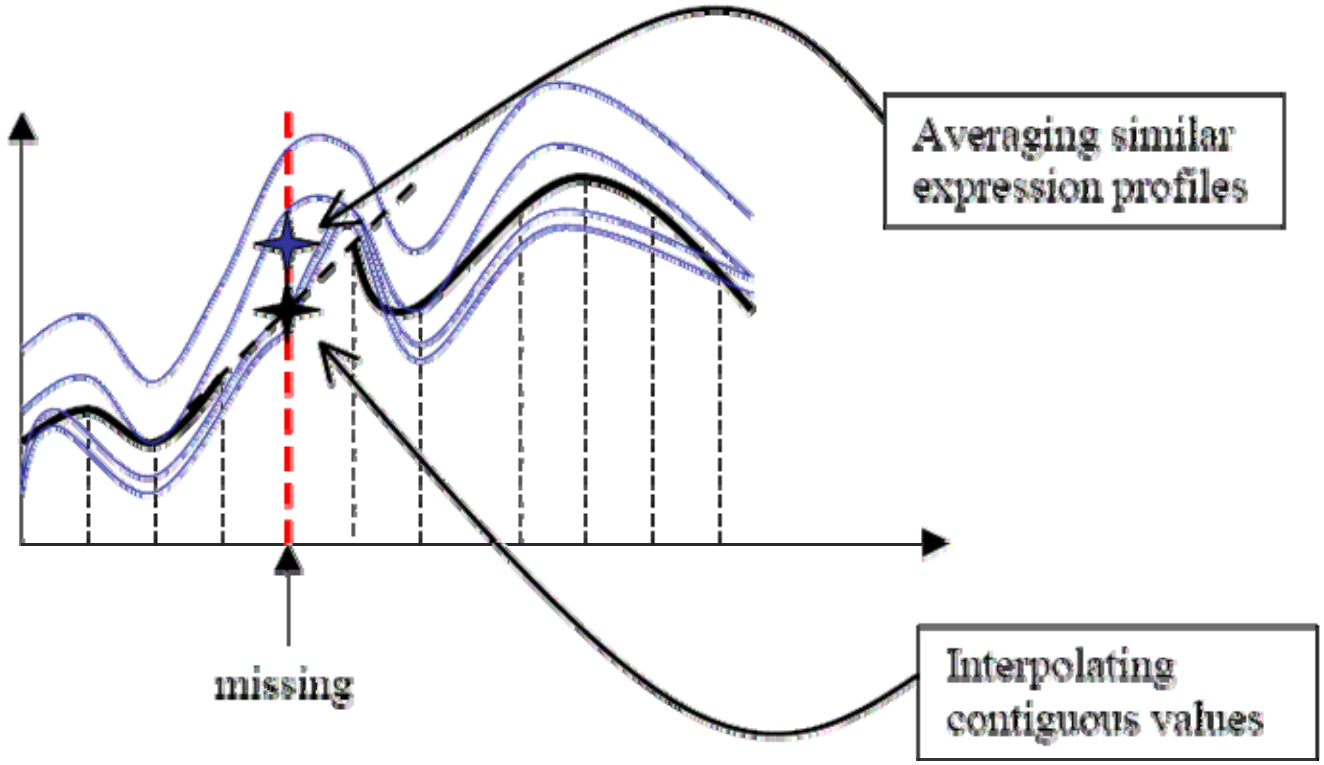


Métodos:

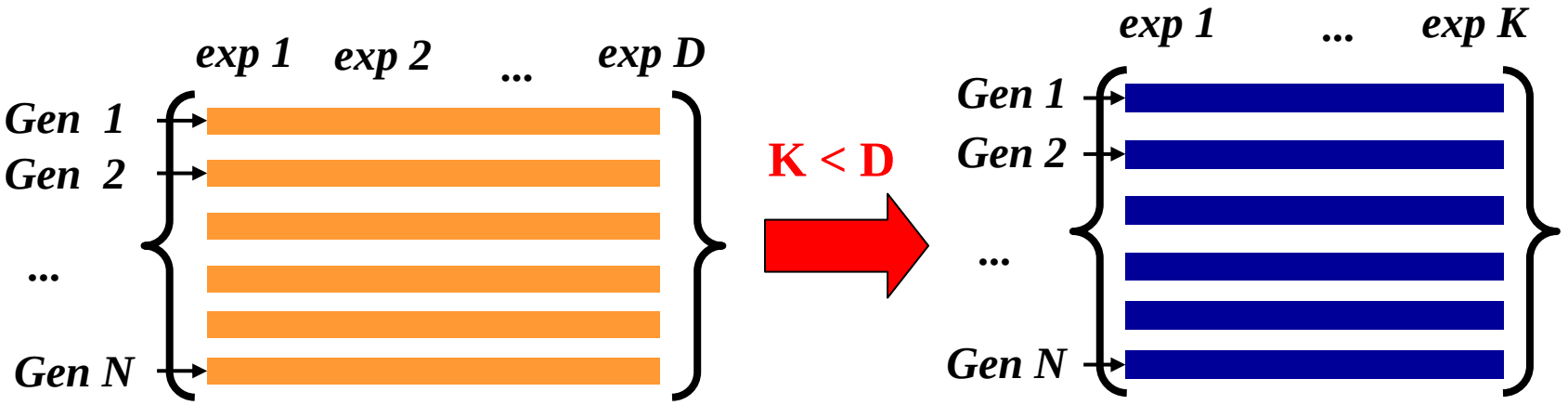
- **Fijar a un valor predeterminado: normalmente 0**
- **Sustituir el valor perdido por la media de toda la columna (experimento)**
- **Sustituir el valor perdido por la media de toda la fila (gen)**
- **Interpolación local pesada**



Interpolación pesada por k vecinos mas cercanos



Reducción de dimensionalidad



La idea es reducir el número de variables (descriptores) de los datos. En el caso de que los datos sean expresión génica, los genes se convertirán en un conjunto de "pseudo-genes", los cuales contendrán k nuevos experimentos, siendo k << número de experimentos originales.

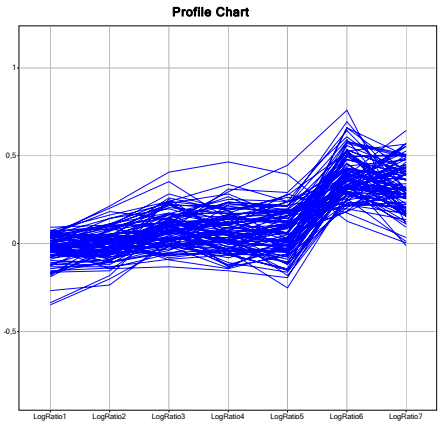
Principal Component Analysis (PCA)

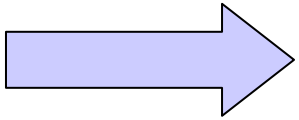
•Es una transformación ortogonal del sistema de coordenadas en el cual están representados los datos.

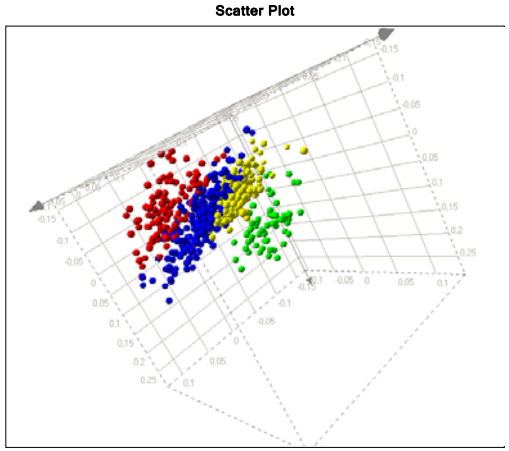
•Los nuevos valores de coordenadas mediante los cuales representamos los datos, se les llaman componentes principales.

•Los componentes principales no están correlacionados.

•Usualmente los primeros “nuevos” ejes correspondientes a los primeros componentes principales contienen gran parte de la información de los datos, así que el resto de componentes pueden ser eliminados.







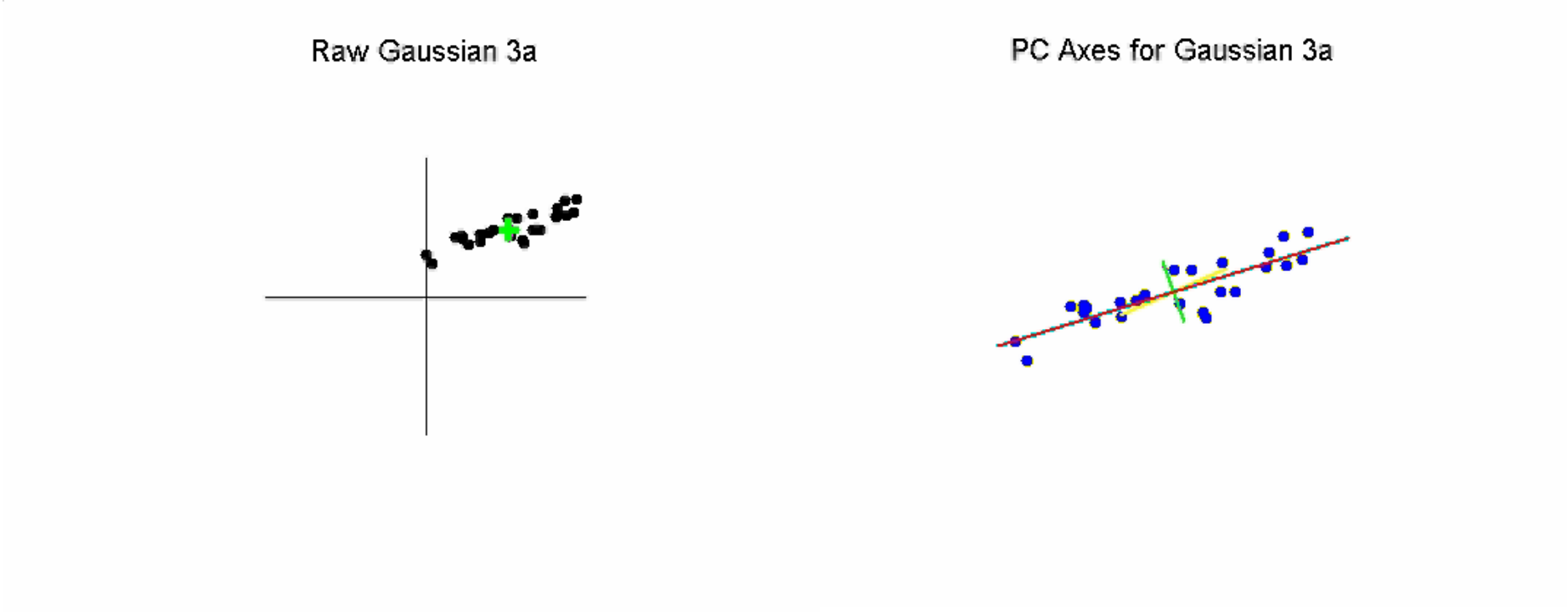
PCA

- Dado un conjunto de datos en un espacio d-dimensional, PCA describe la forma y la localización de la nube de puntos en este espacio d-dimensional.
- PCA se realiza en dos pasos:
  - Traslación: traslada la nube de puntos al origen
  - Rotación: Rota los alrededor del origen
- Otra forma de verlo es aplicar la rotación y traslación a los ejes de coordenadas en vez de a los datos.

Ejemplo de PCA

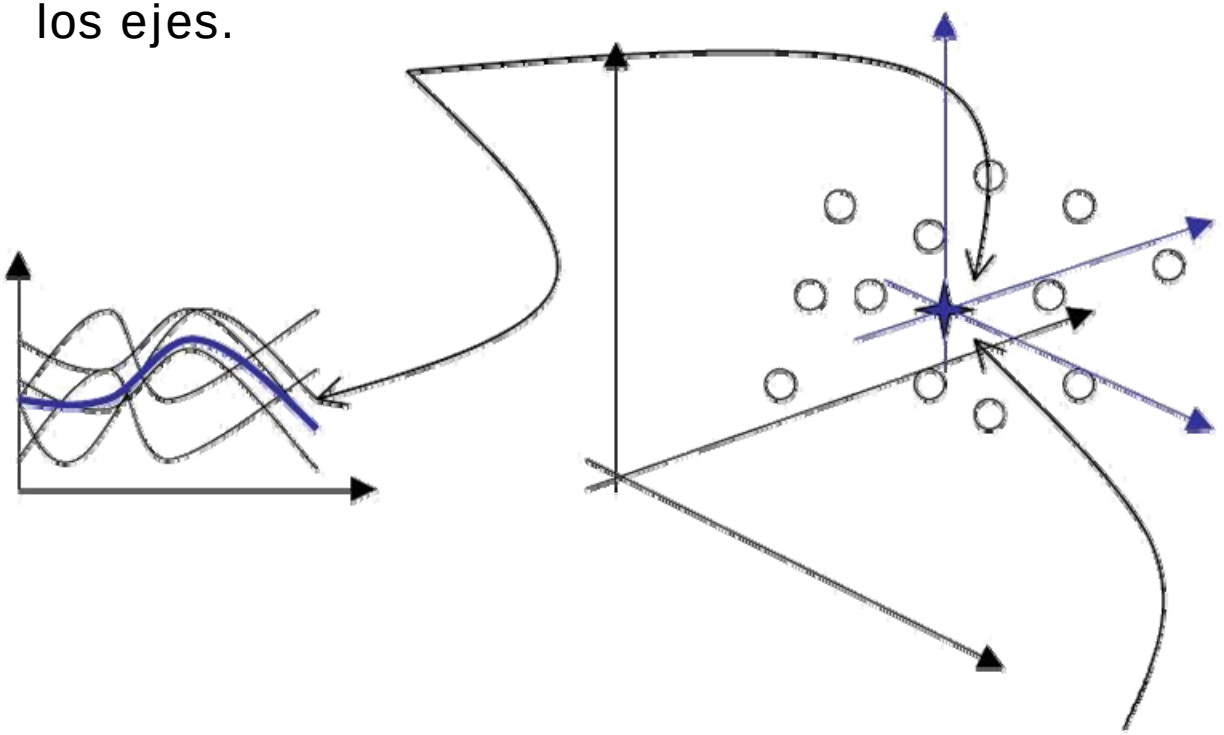
Datos “crudos”

Nuevos ejes: PCA

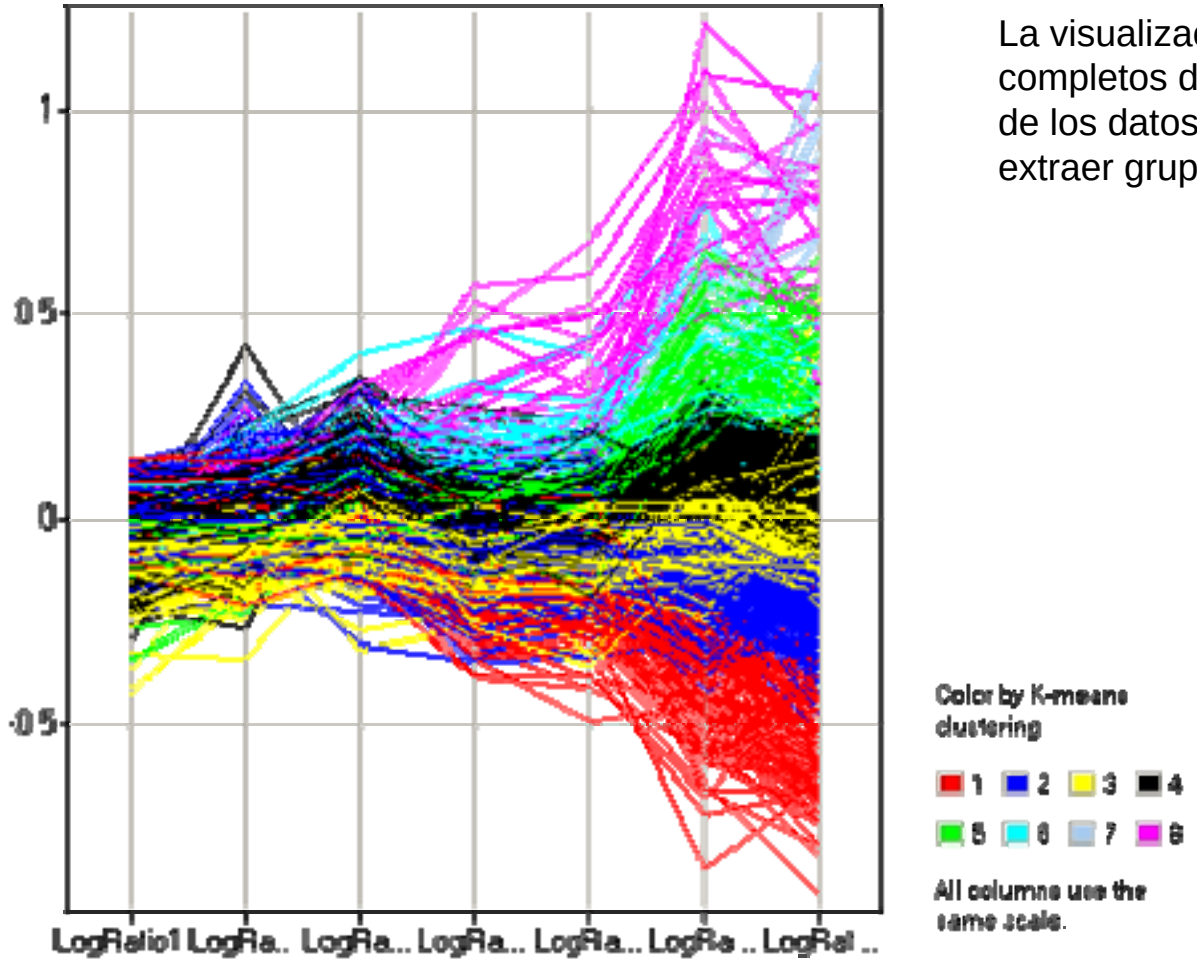


PCA en expresión génica

Transformación de los ejes.

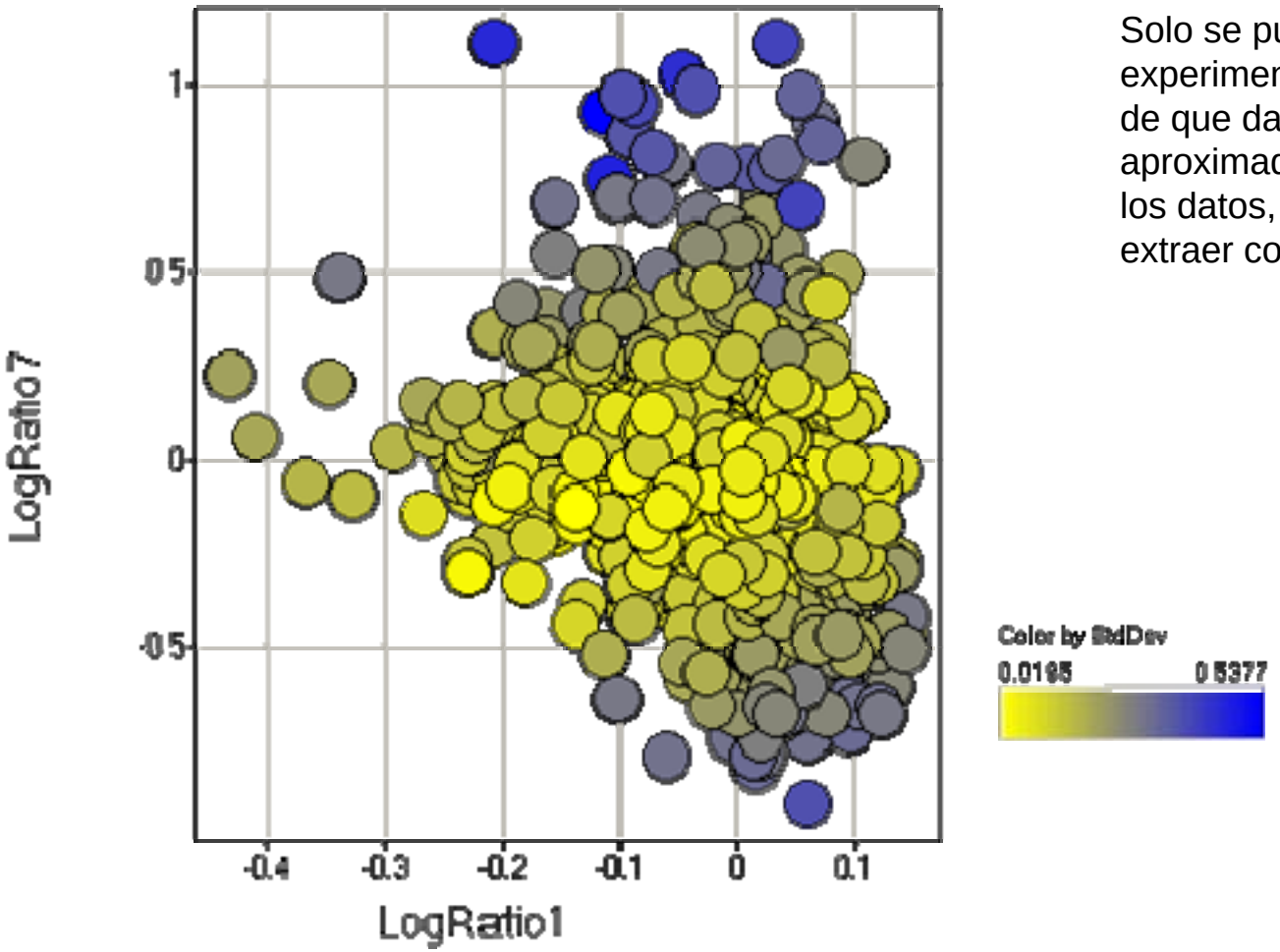


Visualización de perfiles



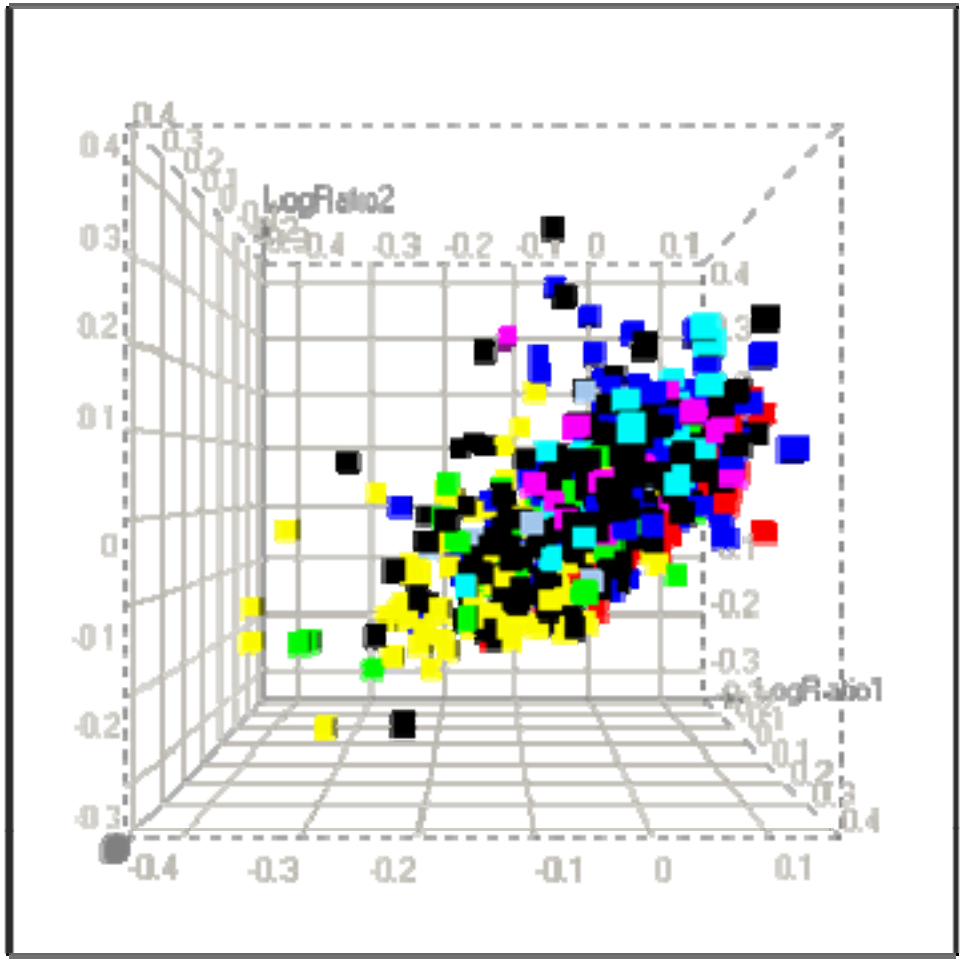
La visualización de perfiles completos da una idea de la forma de los datos, pero es muy difícil extraer grupos de manera visual.

Scatter Plot 2D



Solo se puede visualizar dos experimentos a la vez. A pesar de que da una idea aproximada de la estructura de los datos, no es suficiente para extraer conclusiones.

Scatter Plot (3D)

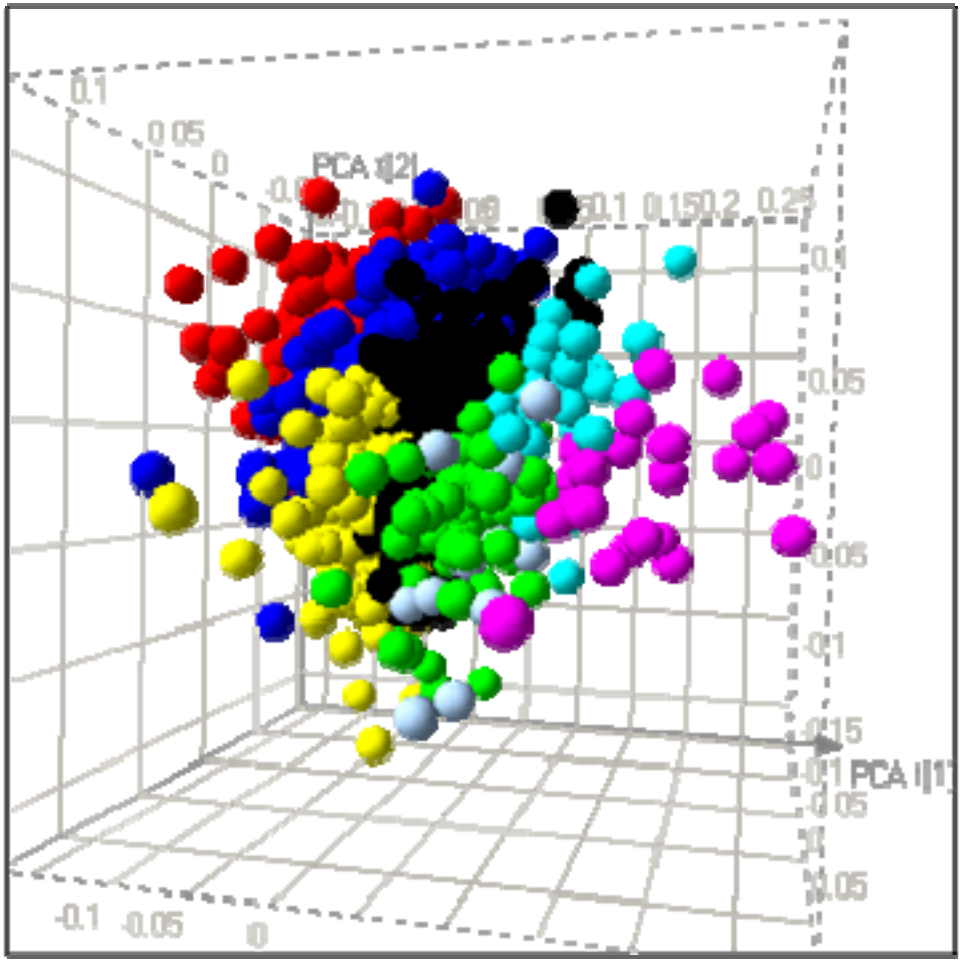


Solo se puede visualizar tres experimentos a la vez. A pesar de que da una idea aproximada de la estructura de los datos, no es suficiente para extraer conclusiones.

Color by K-means clustering

1 2 3 4
5 6 7 8

PCA (3 componentes)



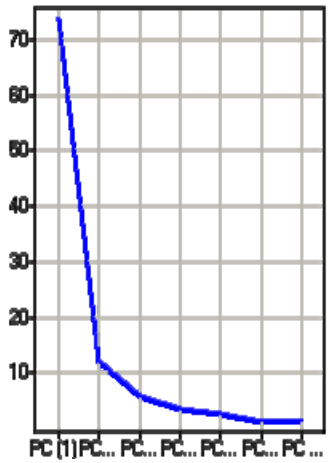
Los 3 primeros componentes principales se pueden mostrar en un gráfico 3D. Al representar los ejes de mayor varianza de los datos, su representación es mas completa y fiable.

Color by K-means clustering

1 2 3 4
5 6 7 8

Interpretación:

Principal Component	Eigenvalue	Eigenvalue (%)	Cumulative Eigenvalue (%)
PC (1)	0.149	73.661	73.661
PC (2)	2.4e-2	12.083	85.744
PC (3)	1.2e-2	5.765	91.509
PC (4)	7.3e-3	3.600	95.109
PC (5)	5.0e-3	2.499	97.608
PC (6)	2.4e-3	1.197	98.806
PC (7)	2.4e-3	1.194	100.000



Preguntas más frecuentes:

- ¿Qué genes están o no expresados?
  - En distintas células
  - En condiciones externas diferentes
  - En diferentes estados de enfermedades
- ¿En cuánto ha cambiado sus niveles de expresión?
- ¿El cambio en la expresión de los genes está correlacionado con otros parámetros externos?

Técnicas de estudio: Estadística descriptiva

Preguntas más frecuentes:

- Los proyectos de secuenciación están produciendo gran cantidad de genes cuya función se desconoce. ¿Se puede utilizar los datos de expresión génica para “predecir” las funciones de los nuevos genes?

	Gene Function	Exp 1	Exp 2	Exp 3	Exp4	Exp5
Exp Attributes		type I	type II	type III	type II	type III
Gene 1	kinase	0.21	0.56	0.72	0.38	0.69
Gene 2	kinase	0.69	0.64	0.55	0.67	0.79
Gene 3	protease	0.01	0.74	0.49	0.50	0.38
Gene 4	protease	0.37	0.91	0.98	0.31	0.25
Gene 5	protease	0.14	0.34	0.92	0.43	0.59
Gene 6	metabolism	0.28	0.45	0.60	0.60	0.46
Gene 7	metabolism	0.66	0.60	0.17	0.15	0.61
Gene 8	?	0.28	0.86	0.24	0.71	0.99
Gene 9	?	0.13	0.30	0.55	0.84	0.86
Gene 10	?	0.85	0.85	0.46	0.77	0.78
...	...					

Técnicas de estudio: Métodos de clasificación supervisados, redes neuronales, etc.

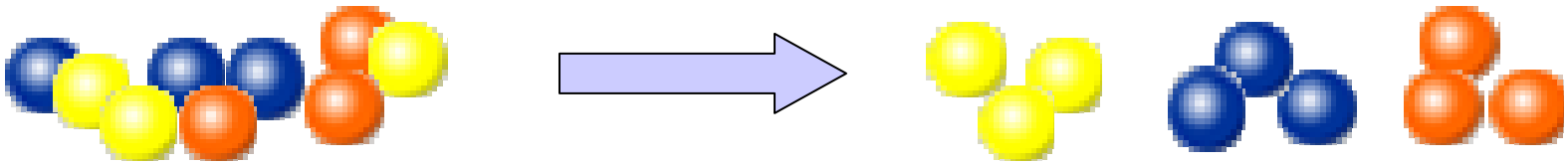
Preguntas más frecuentes:

- **¿Podemos utilizar los patrones de expresión de los genes para agrupar genes cuya función se desconoce?**
  - Clasificación funcional de genes cuya función se desconoce (patrones de expresión similares puede implicar funciones similares)
  - Clasificación molecular de muestras (por ejemplo subtipos de tumores indistinguibles morfológicamente).
  - Descifrar mecanismos de regulación mediante la identificación de grupos de genes que se co-expresen y que probablemente estén co-regulados.
  - Identificación de patrones de expresión de genes “diagnóstico” (cuya función se conoce)

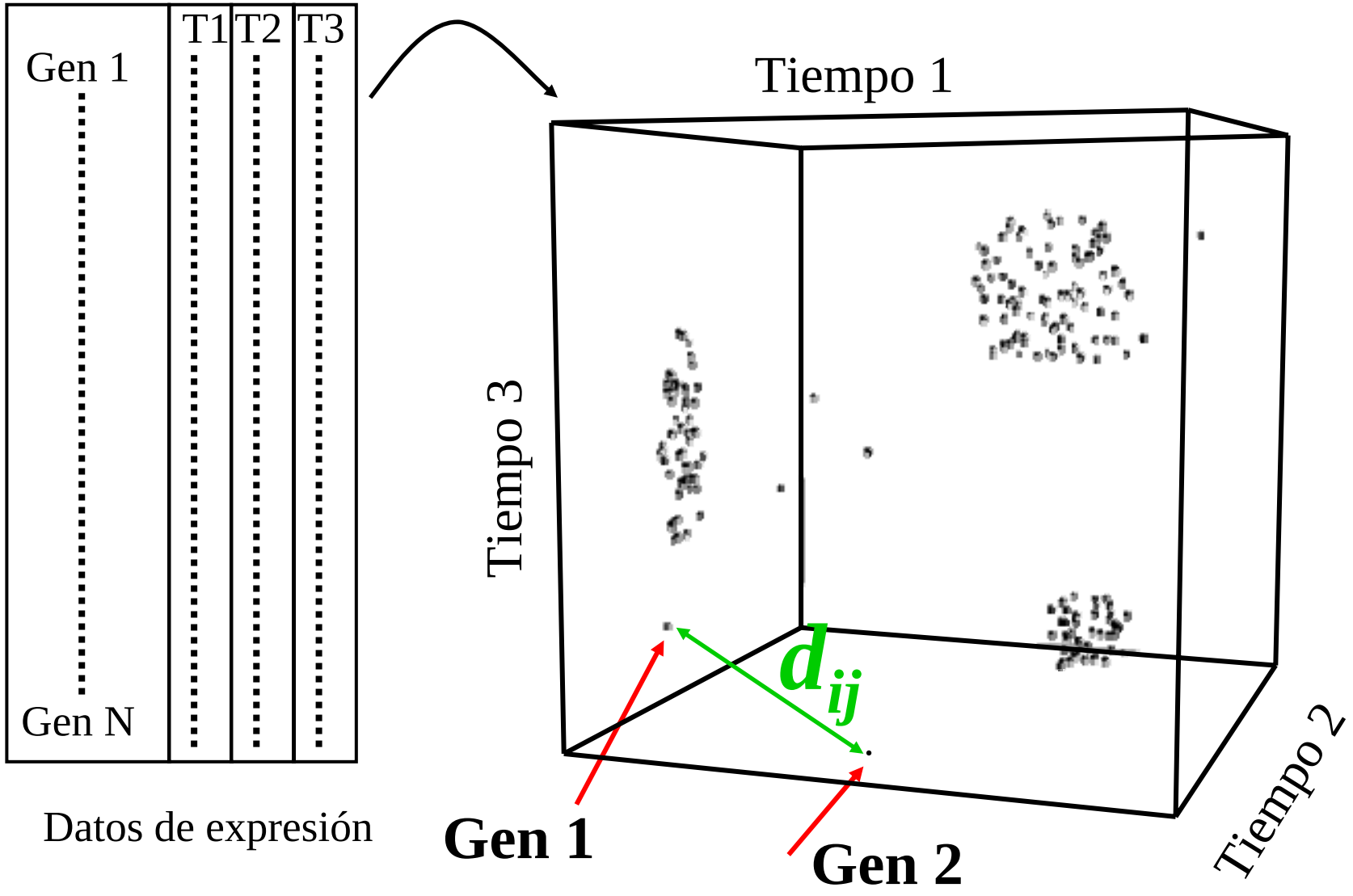
Técnicas de estudio: Análisis de agrupamiento

Análisis de agrupamiento (Cluster Analysis):

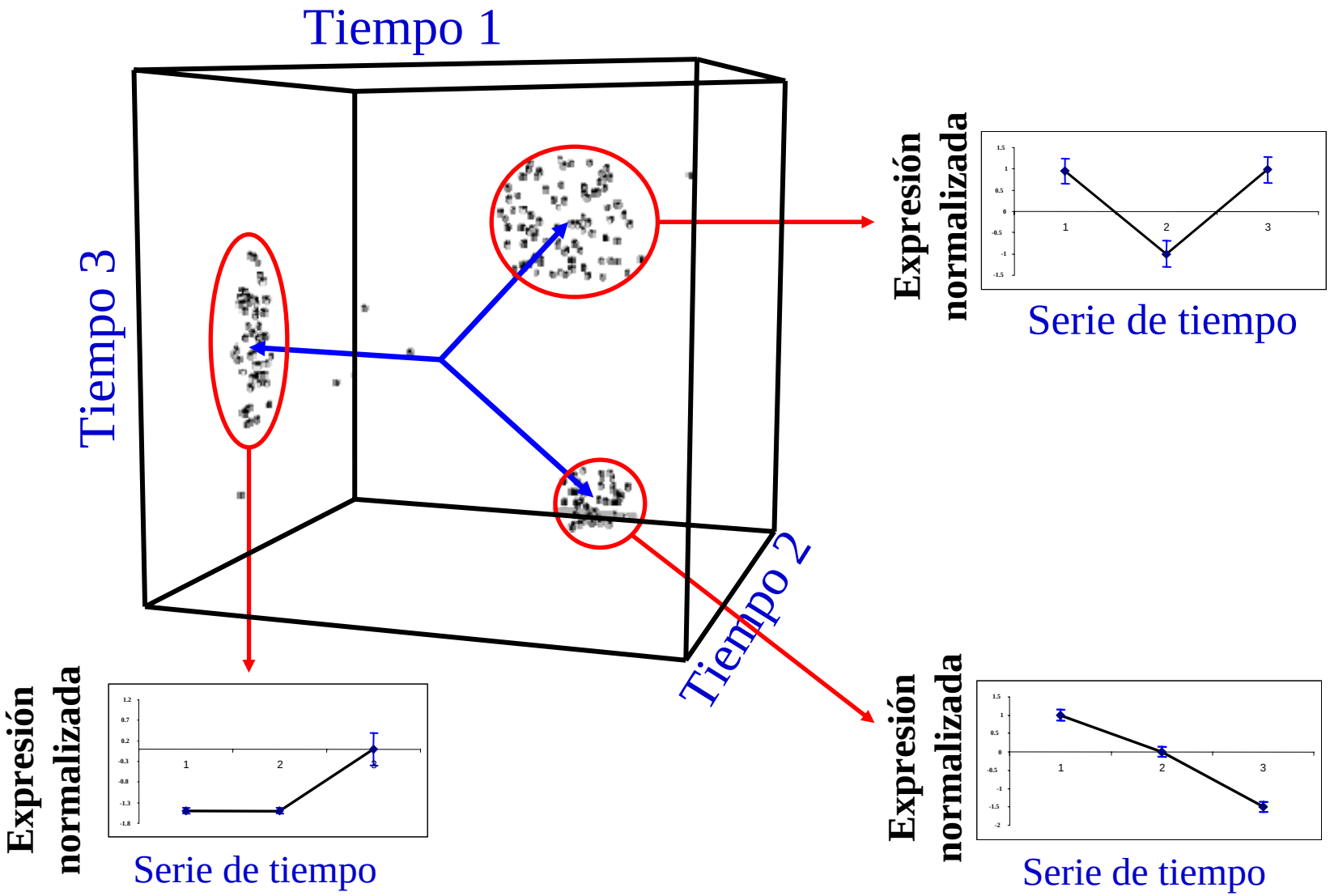
- Entrada: ***n*** datos, $\mathbf{X}_i$, i=1,2,…,N en un espacio ***p*** dimensional. Por ejemplo: 1000 (***n***) genes en 10 (***p***) condiciones experimentales.
- Objetivo: Encontrar grupos ó “clusters” naturales. Los datos en un mismo grupo o cluster deben ser “más similares”
- Nota importante: ¿Cuántos grupos tenemos?



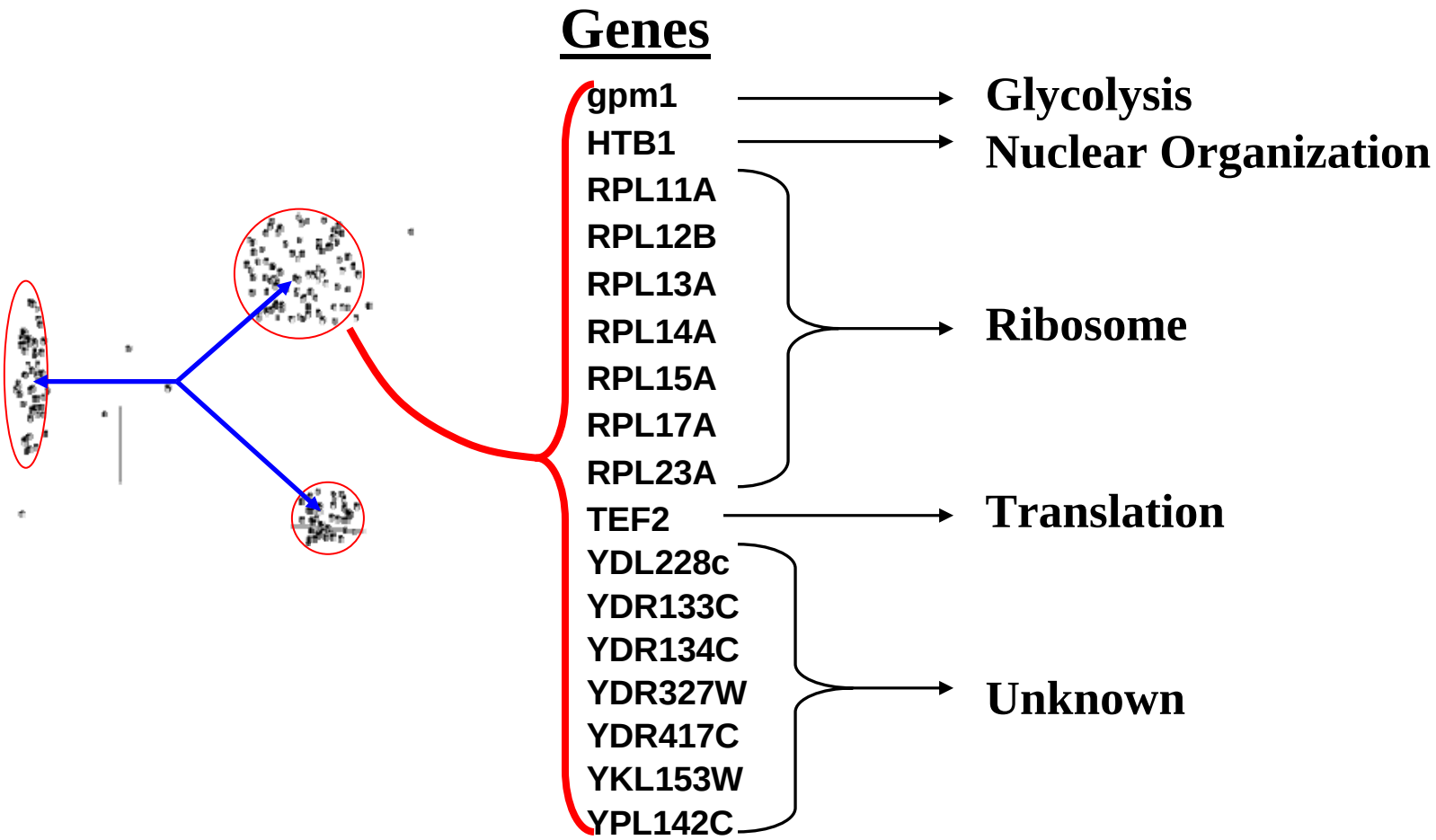
Representación de los datos de expresión:



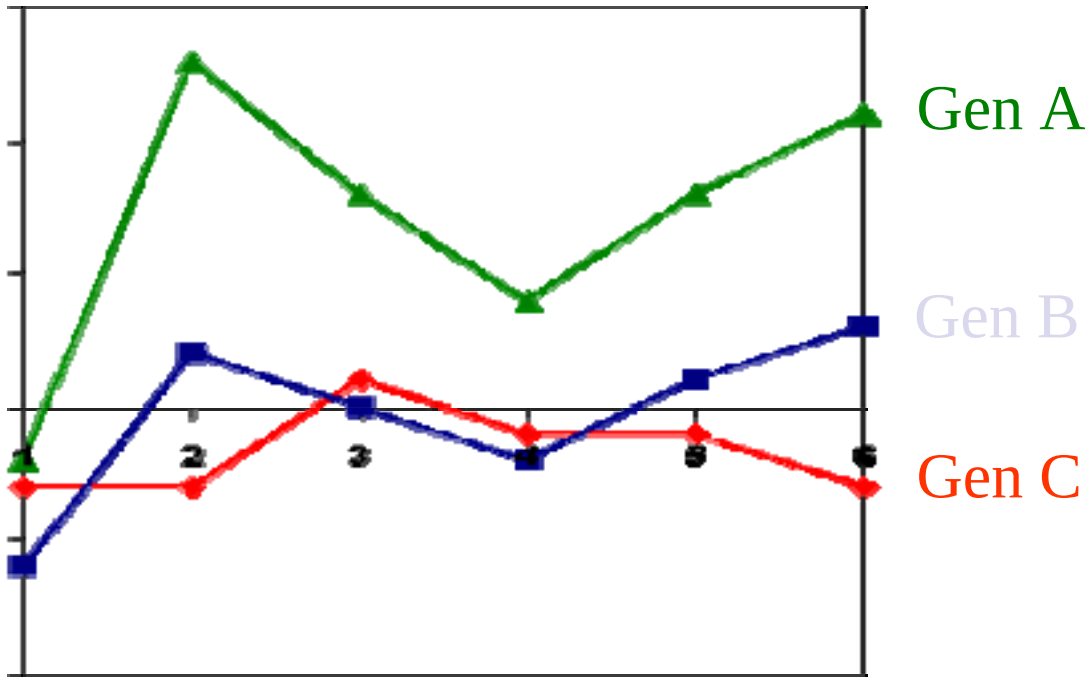
Identificación de patrones de expresión prevalentes (grupos de genes ó clusters)



Evaluación del contenido de los grupos:



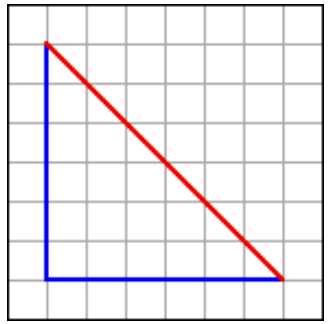
Distancia entre patrones de expresión: ¿Cuál es la más apropiada?



Métricas más comunes:

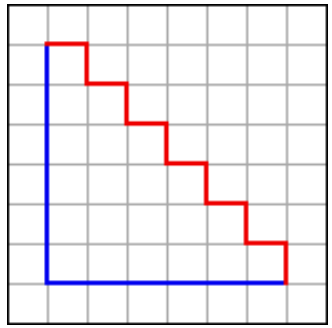
1, $r = 2$ (Distancia euclidea)

$$d(x, y) = \sqrt[2]{\sum_{i=1}^{p} | x_i - y_i |^2}$$



2, $r = 1$ (Distancia de Manhattan)

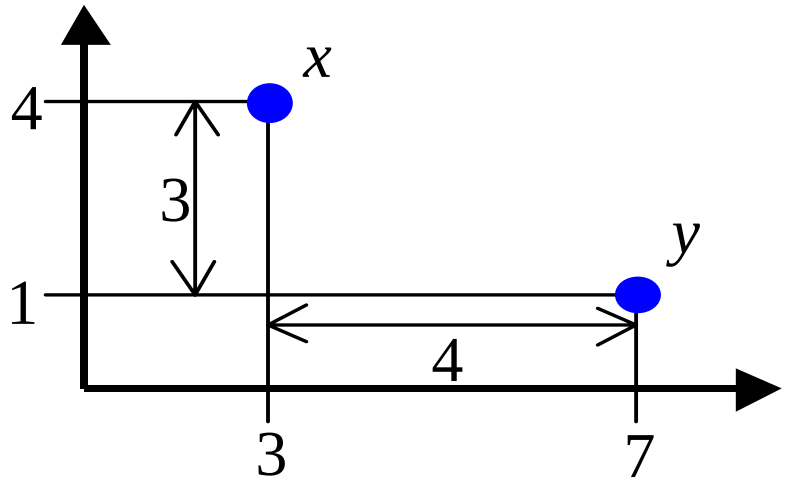
$$d(x, y) = \sum_{i=1}^{p} | x_i - y_i |$$



3, $r = +\infty$ (Distancia "sup")

$$d(x, y) = \max_{1 \le i \le p} | x_i - y_i |$$

Ejemplos:



1, Euclidean distance : $\sqrt[2]{4^2+3^2}=5.$

2, Manhattan distance : $4+3=7.$

3, "sup" distance : $\max\{4,3\}=4.$

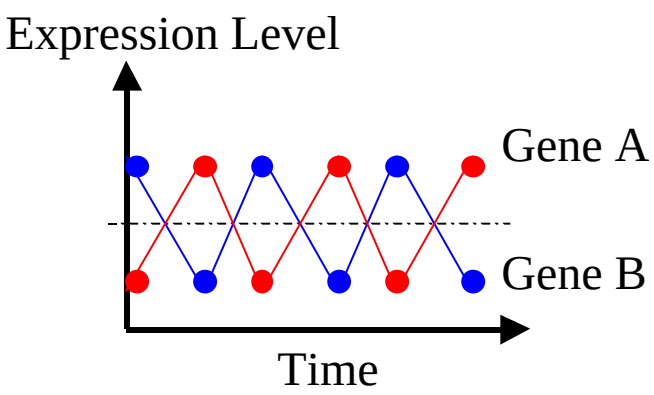
Otro método de similitud: coeficiente de correlación

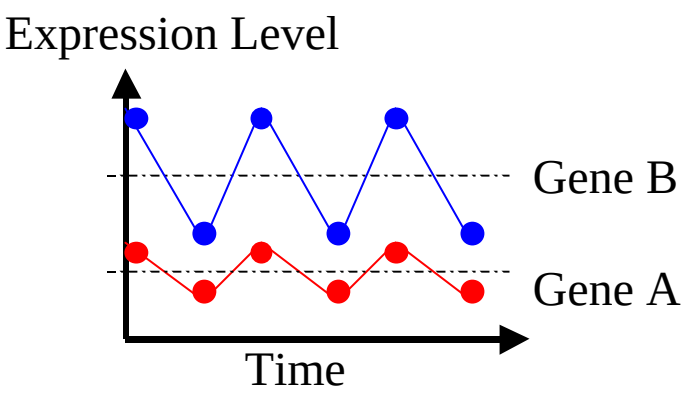
$$s(x,y) = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \times \sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

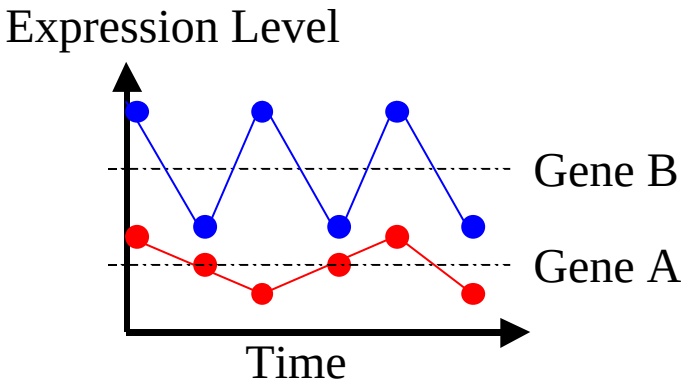
$$\text{averages:} \quad \bar{x} = \frac{1}{p}\sum_{i=1}^{p} x_i \quad \text{and} \quad \bar{y} = \frac{1}{p}\sum_{i=1}^{p} y_i .$$

$$\left| s(x,y) \right| \leq 1$$

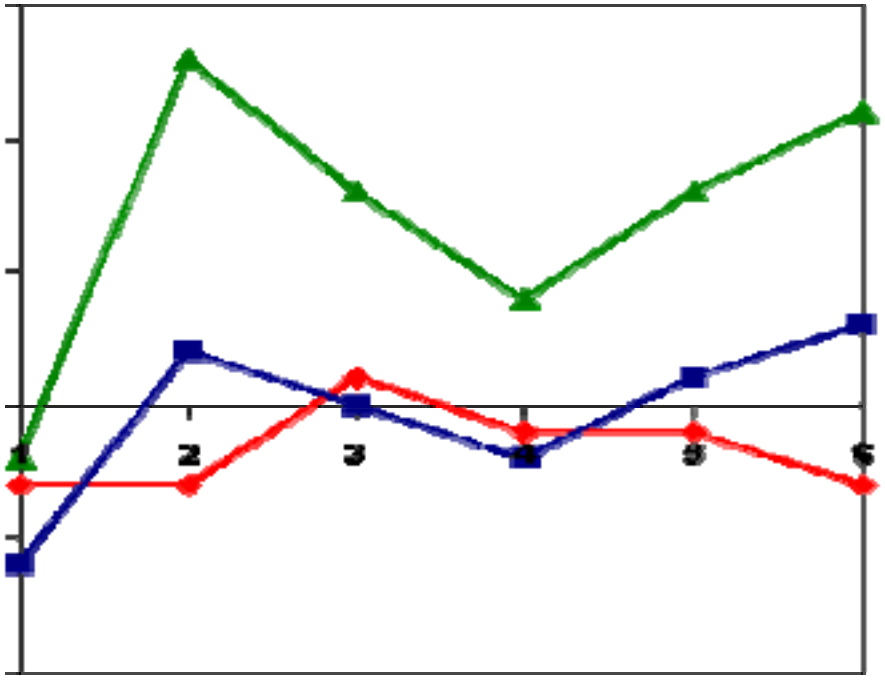
Ejemplos:


Expression Level
Gene A
Gene B
Time


Expression Level
Gene B
Gene A
Time


Expression Level
Gene B
Gene A
Time

Distancia entre genes:¿Euclidea ó Correlación?



Euclidea: $d_{x,y} = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$

Correlación: $d_{x,y} = 1 - \frac{1}{p} \sum_{i=1}^{p} \left[\frac{(x_i - \bar{x})}{\delta x} \right] \left[\frac{(y_i - \bar{y})}{\delta y} \right]$

Euclidea: Tiende a agrupar perfiles de acuerdo al valor absoluto de las diferencias entre los niveles de expresión. Rojo y Azul

Correlación: Tiende a agrupar perfiles de acuerdo a la tendencia de los mismos. Verde y Azul

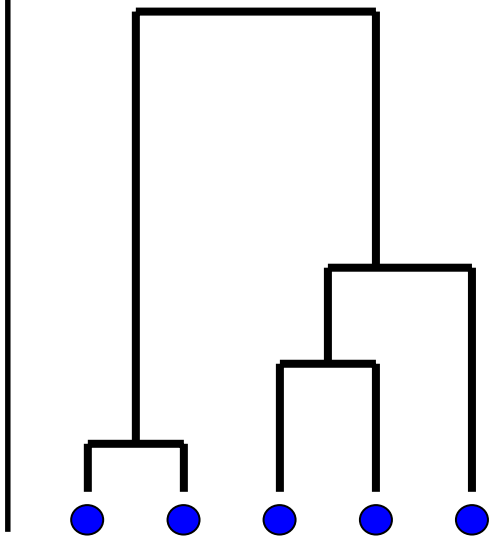
En DNA Arrays, la distancia de correlación usualmente tiene más significado biológico que la distancia Euclidea.

Métodos de clustering:

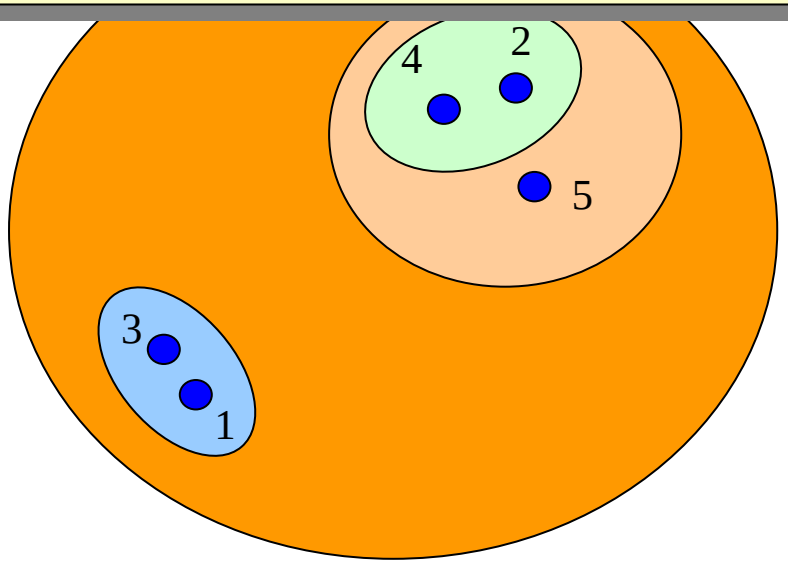
- ✓Jerárquicos:
  - **Aglomerativos (HAC)**
  - Divisivos
- ✓Particionales
  - **K-means**
  - Fuzzy K-means
- ✓Basados en modelos
- ✓Basados en Densidad (EM)
- ✓Basados en grid (CLIQUE)
- ✓Basados en grafos (CLICK)
- ✓**Redes neuronales (SOM)**

Se necesita calcular la distancia entre el nuevo cluster y los demás. Existen varios tipos:

Single Linkage: distancia entre el par de puntos más cercano.
Complete Linkage: distancia entre el par de puntos más alejado.
Average Linkage: distancia promedio entre todos los pares de puntos.
Centroids: distancia entre los centros de los clusters.
Ward: Une clusters que sean mas "compactos"



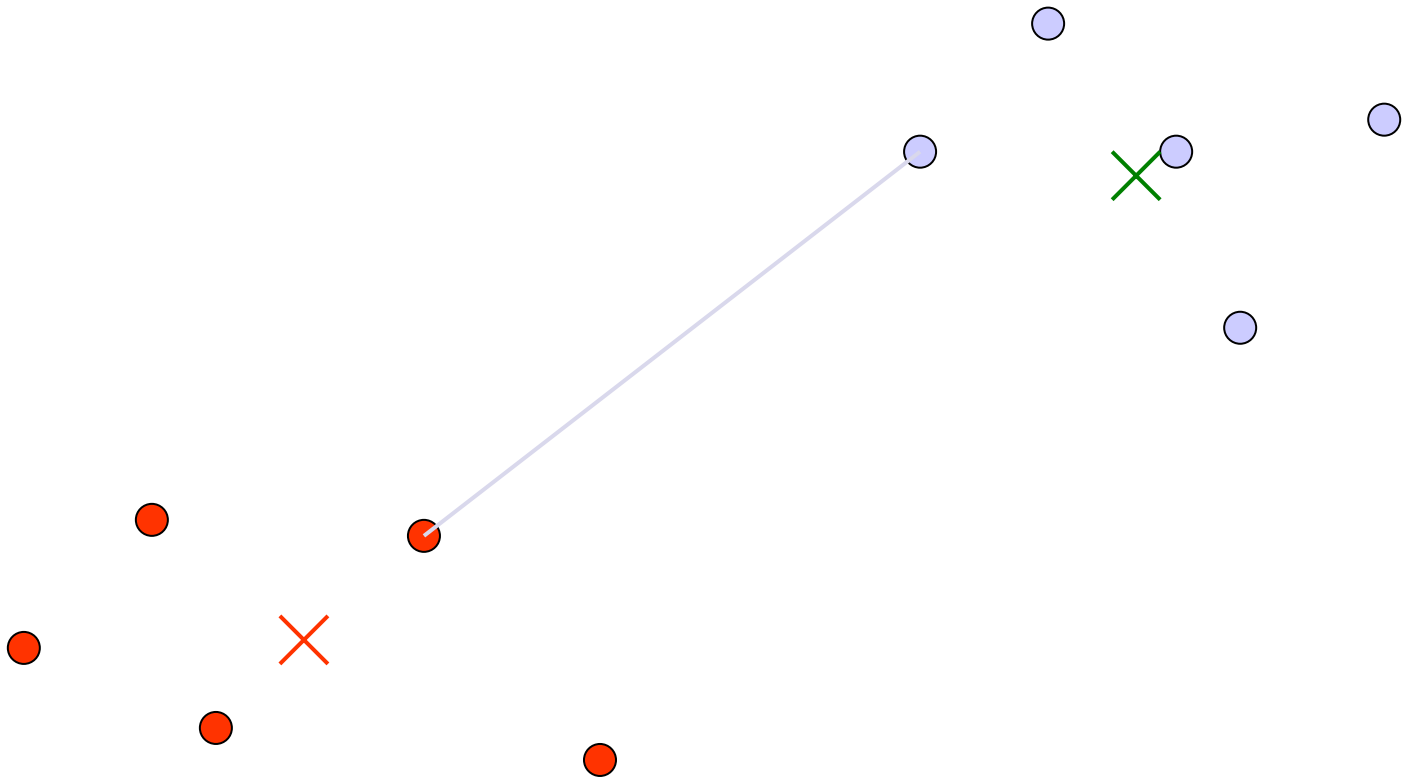
Dendrograma



El dendrograma induce un orden lineal de los datos.

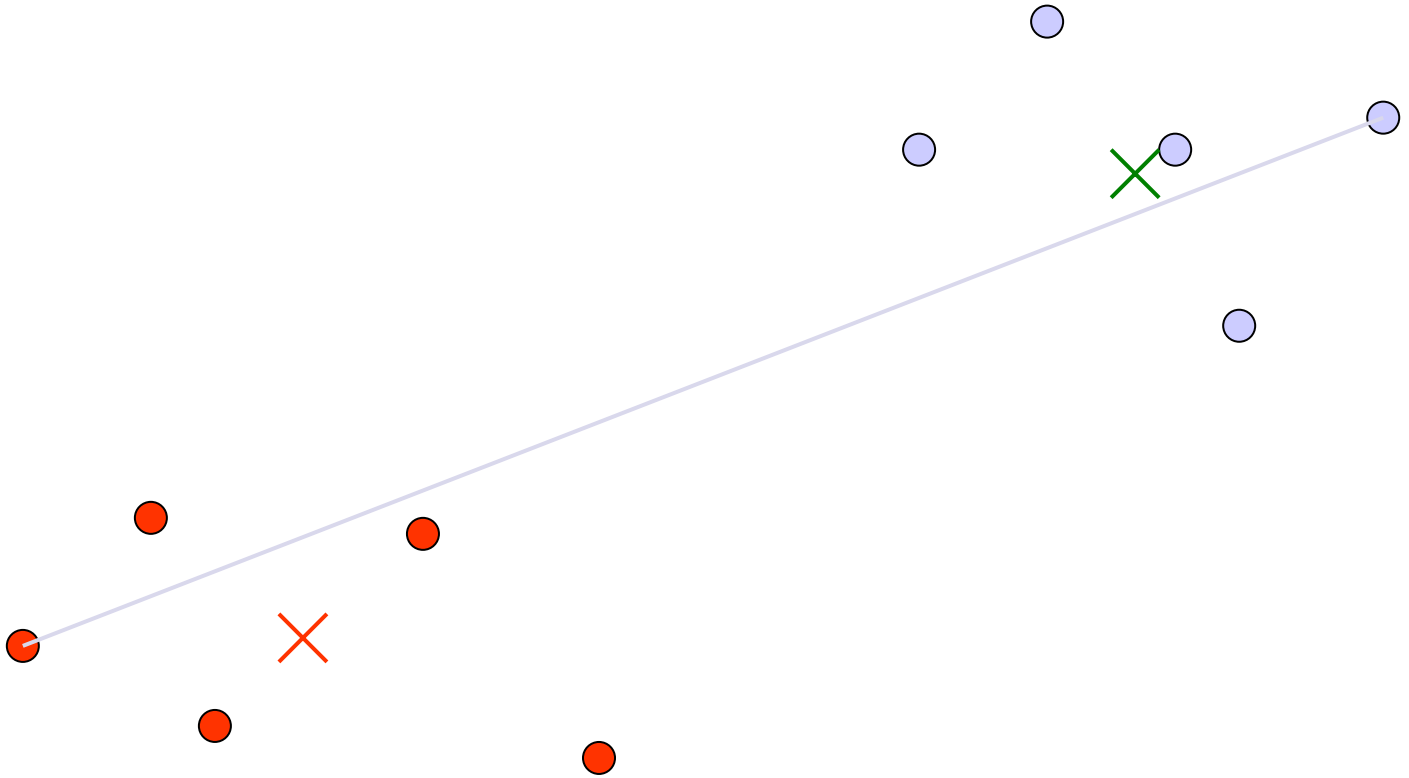
Single Linkage: distancia entre el par de puntos más cercano.

Tipo de clusters: alongados, tipo cadenas.
Ventajas: Simples, eficientes, propiedades teóricas conocidas
Desventajas: No son adecuados para clusters no bien separados o esféricos



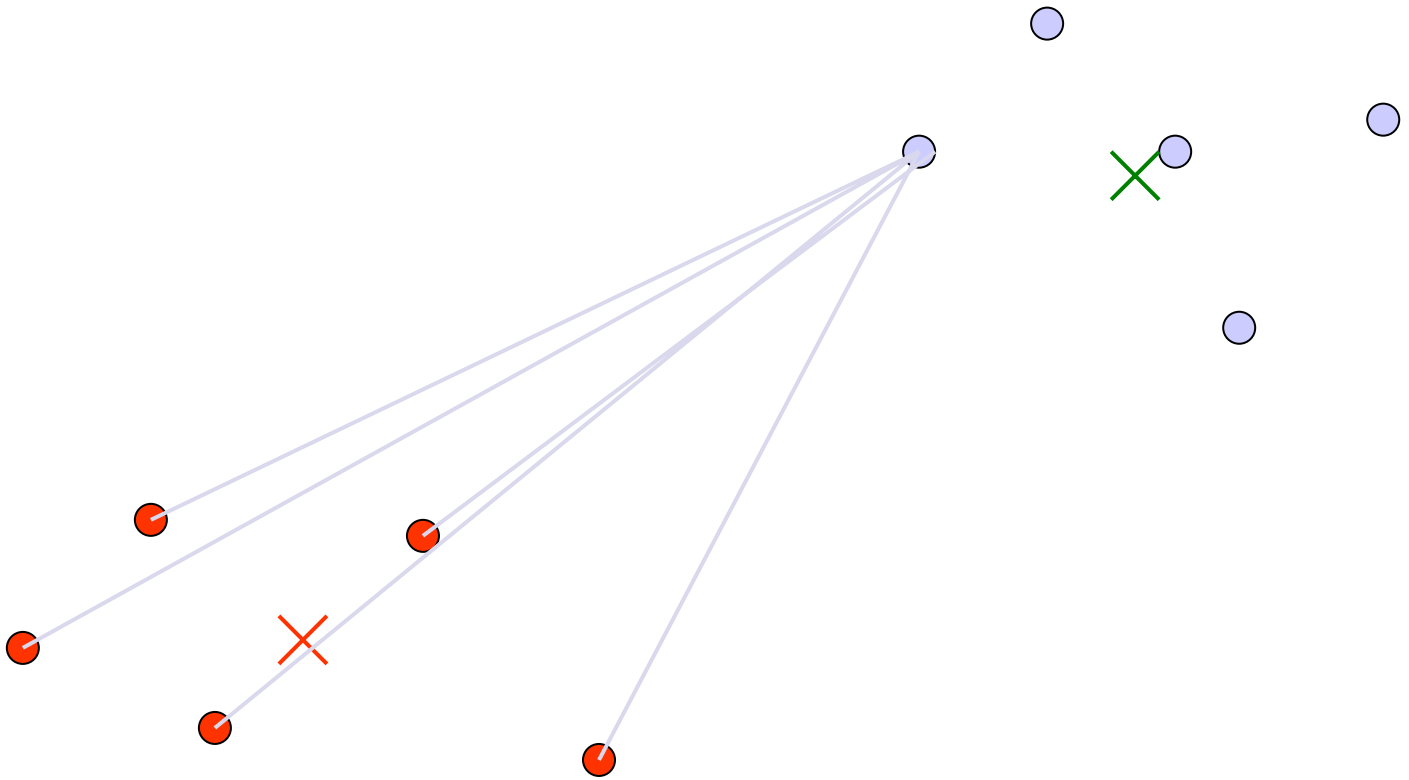
Complete Linkage: distancia entre el par de puntos más alejado.

Tipo de clusters: compactos, esféricos. Tiende a extraer “nubes de puntos”
Ventajas: Funcionan bien cuando los clusters son muy compactos.
Desventajas: No es eficiente en conjuntos grandes de datos.



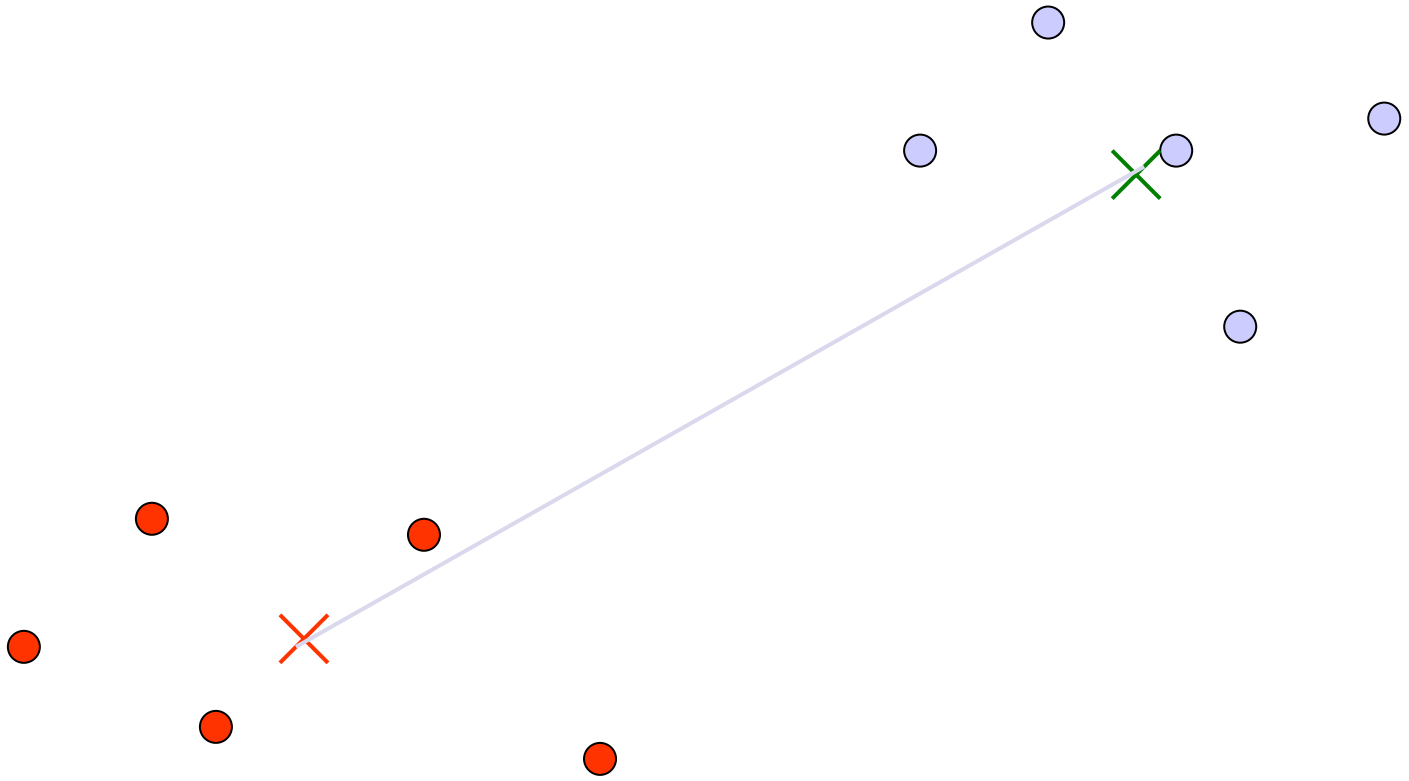
Average Linkage: distancia promedio entre todos los pares de puntos.

Tipo de clusters: Intermedio entre single y complete linkage.
Ventajas: Funciona bien tanto para clusters alongados como esféricos.
Desventajas: Muy costoso computacionalmente.

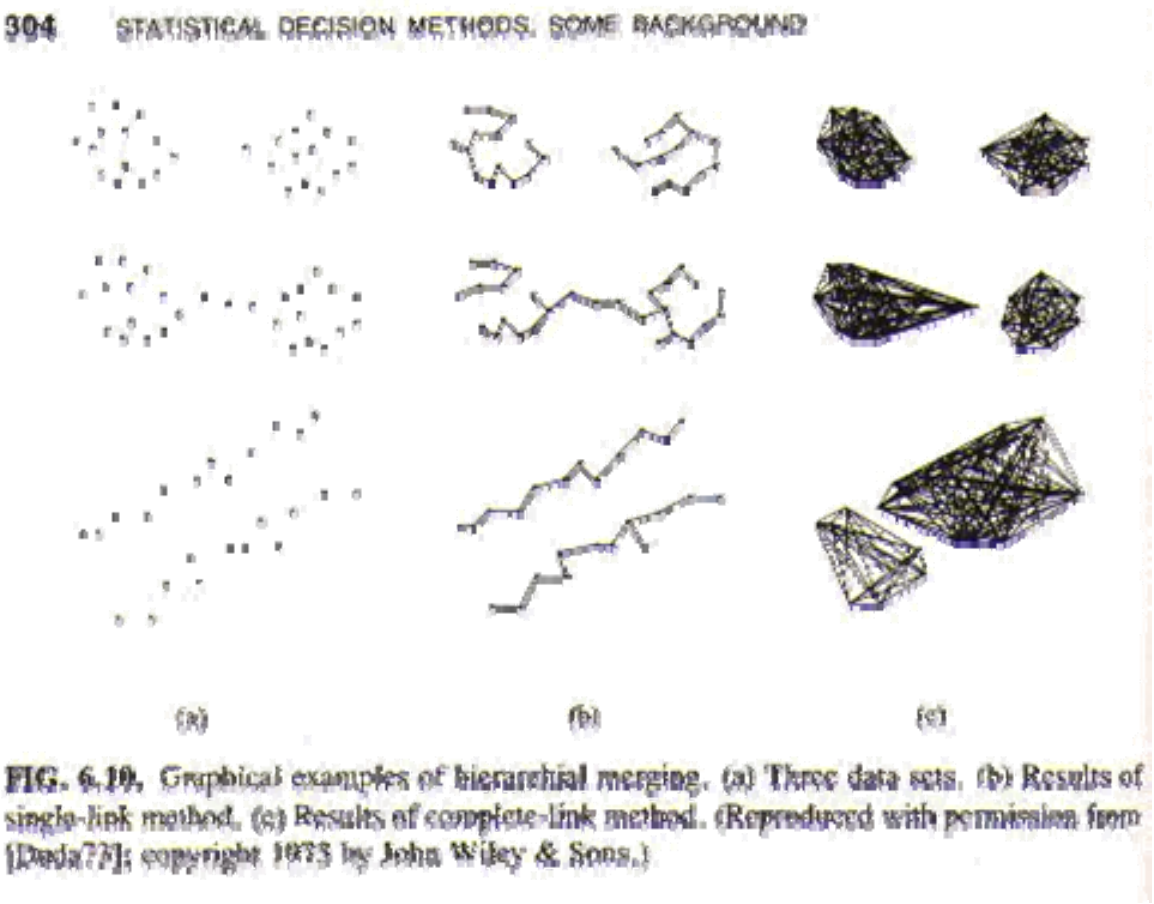


Centroids: **distancia entre los centros de los clusters.**

Desventajas: Actualizaciones en los datos pueden provocar la creación de jerarquías completamente diferentes.

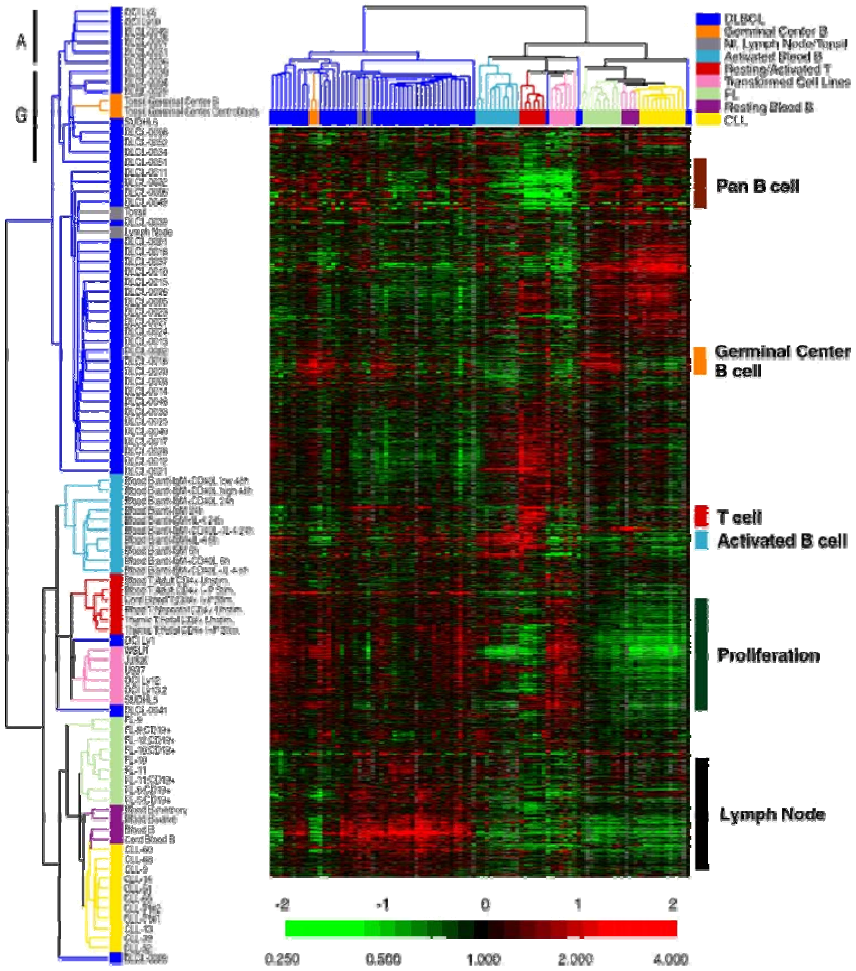


Efecto de los métodos de agregación:



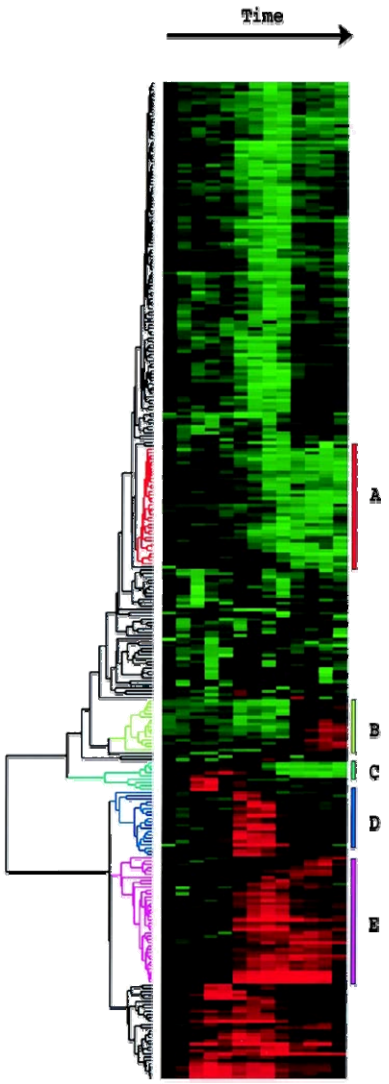
Nadler and Smith, Pattern Recognition Engineering, 1993

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., *et.al.* (2000)
Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.
Nature 403: 503-11.



Depicted are the ~1.8 million measurements of gene expression made on 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of relatedness of gene expression in each samples. The dendrogram is color coded based on the category of mRNA sample studied (see upper right key). Each row represents a separate cDNA clone on the microarray and each column a separate mRNA sample. The scale extends from fluorescence ratios of 0.25 to 4 (-2 to +2 in log base 2 units). Grey indicates missing or excluded data.

Eisen, M., Spellman, P.T., Botstein, D. & Brown, P.O. (1998). Cluster Analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863-14867.



Single time course data of a canonical model of the growth response in human cells: clustered data from serum simulation of primary human fibroblasts. Briefly, foreskin fibroblasts were grown in culture and were deprived of serum for 48 hours. Serum was added back and samples taken at time , 0, 15 min, 30 min, 1h, 2h, 3h, 4h, 8h, 12h, 16h, 20h and 24h. Five clusters were identify containing known genes involved in:

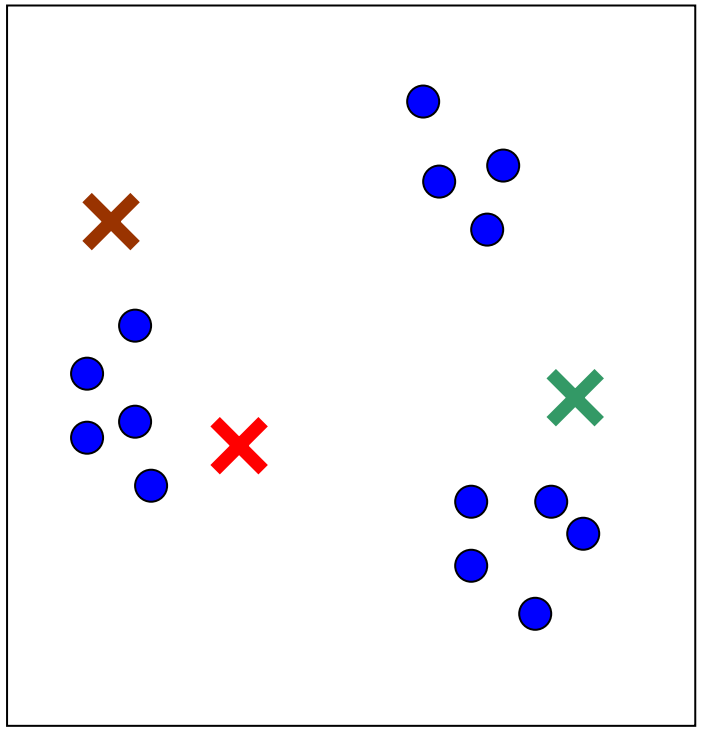
(A) cholesterol biosynthesis
(B) The cell cycle
(C) The immediate-early response
(D) Signaling and Angiogenesis
(E) Wound healing and tissue remodeling.

Agglomerative Hierarchical Clustering:

- Los resultados dependen mucho del método de distancia escogido:
  - Single Linkage: clusters alongados.
  - Complete Linkage: clusters esféricos.
- Proceso iterativo muy costoso.
- Fueron inicialmente diseñados para trabajar con datos que poseen estructura jerárquica, lo cual no garantiza que funcionen bien para todo tipo de datos.
- La naturaleza determinista del método y la imposibilidad de la reevaluación del clustering a posteriori puede causar que la agrupación sea basada en decisiones locales más que en globales.
- NO es robusto en presencia de ruido.
- La decisión del número de clusters es subjetiva.

Clustering particional: K-means

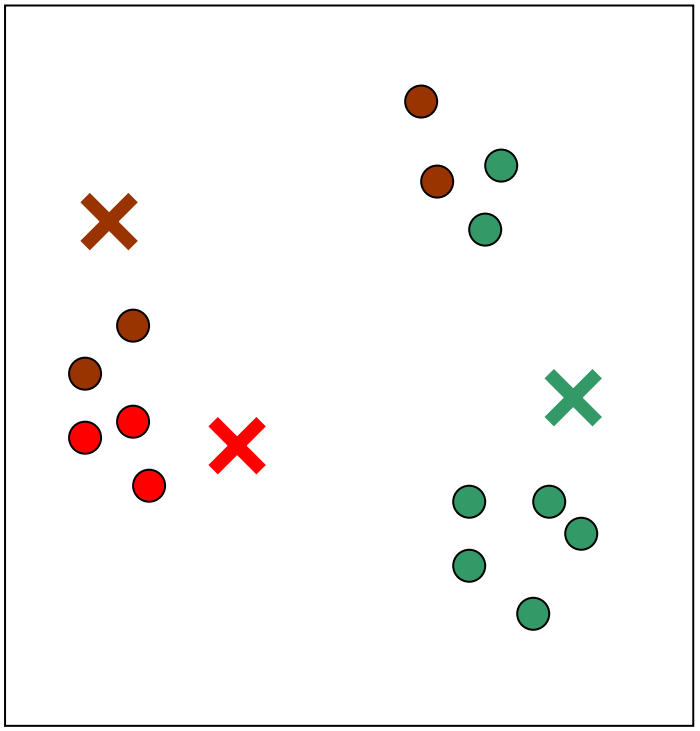
- ***Comenzar con una posición aleatoria para los centros de los clusters.***
- Iterar hasta que los centroides se estabilicen.
  - Asignar puntos a los centroides.
  - Mover los centroides hacia los “centros“ de los puntos asignados.



Iteración = 0

Clustering particional: K-means

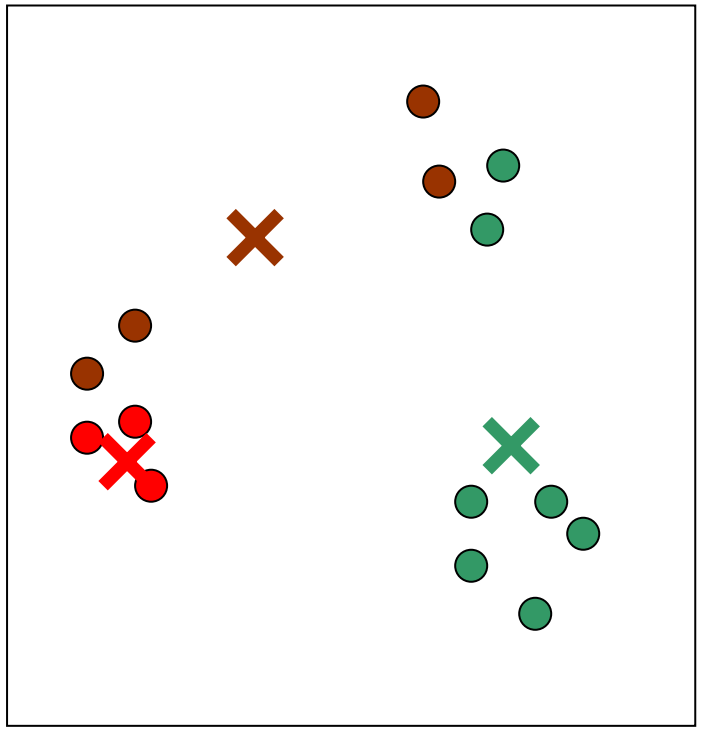
- Comenzar con una posición aleatoria para los centros de los clusters.
- Iterar hasta que los centroides se estabilicen.
  - ***Asignar puntos a los centroides.***
  - Mover los centroides hacia los “centros“ de los puntos asignados.



Iteración = 1

Clustering particional: K-means

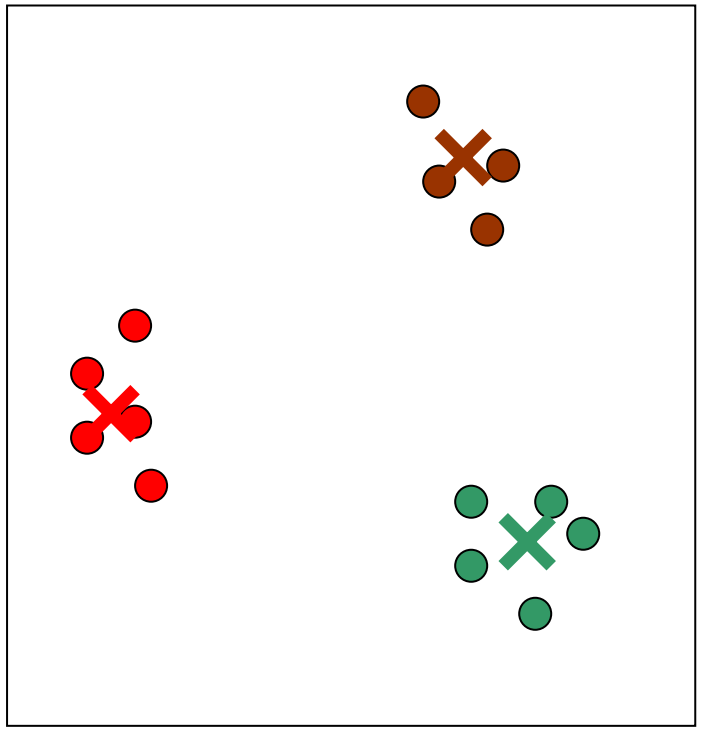
- Comenzar con una posición aleatoria para los centros de los clusters.
- Iterar hasta que los centroides se estabilicen.
  - Asignar puntos a los centroides.
  - ***Mover los centroides hacia los "centros" de los puntos asignados.***



Iteración = 1

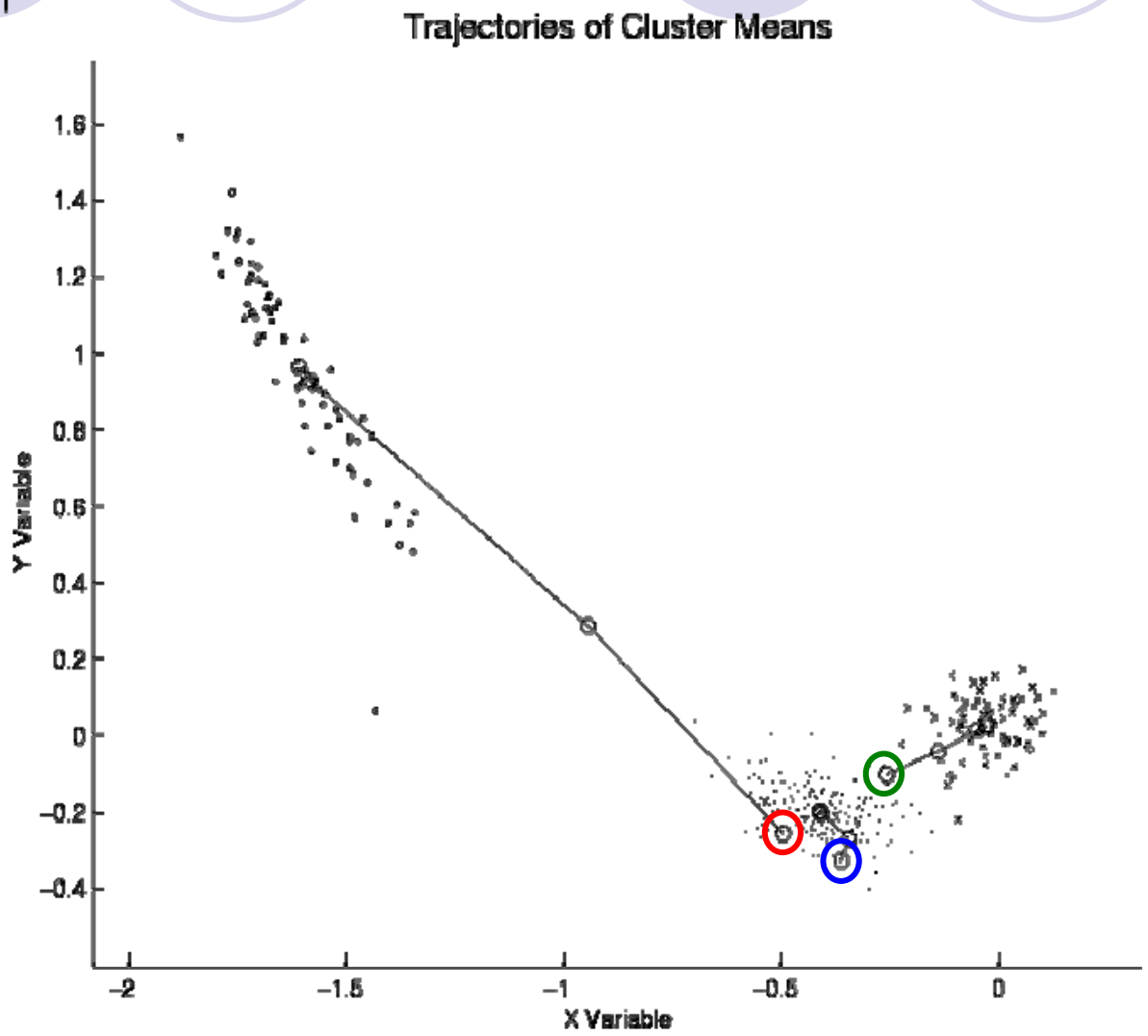
Clustering particional: K-means

- Comenzar con una posición aleatoria para los centros de los clusters.
- ***Iterar hasta que los centroides se estabilicen.***
  - Asignar puntos a los centroides.
  - Mover los centroides hacia los “centros“ de los puntos asignados.

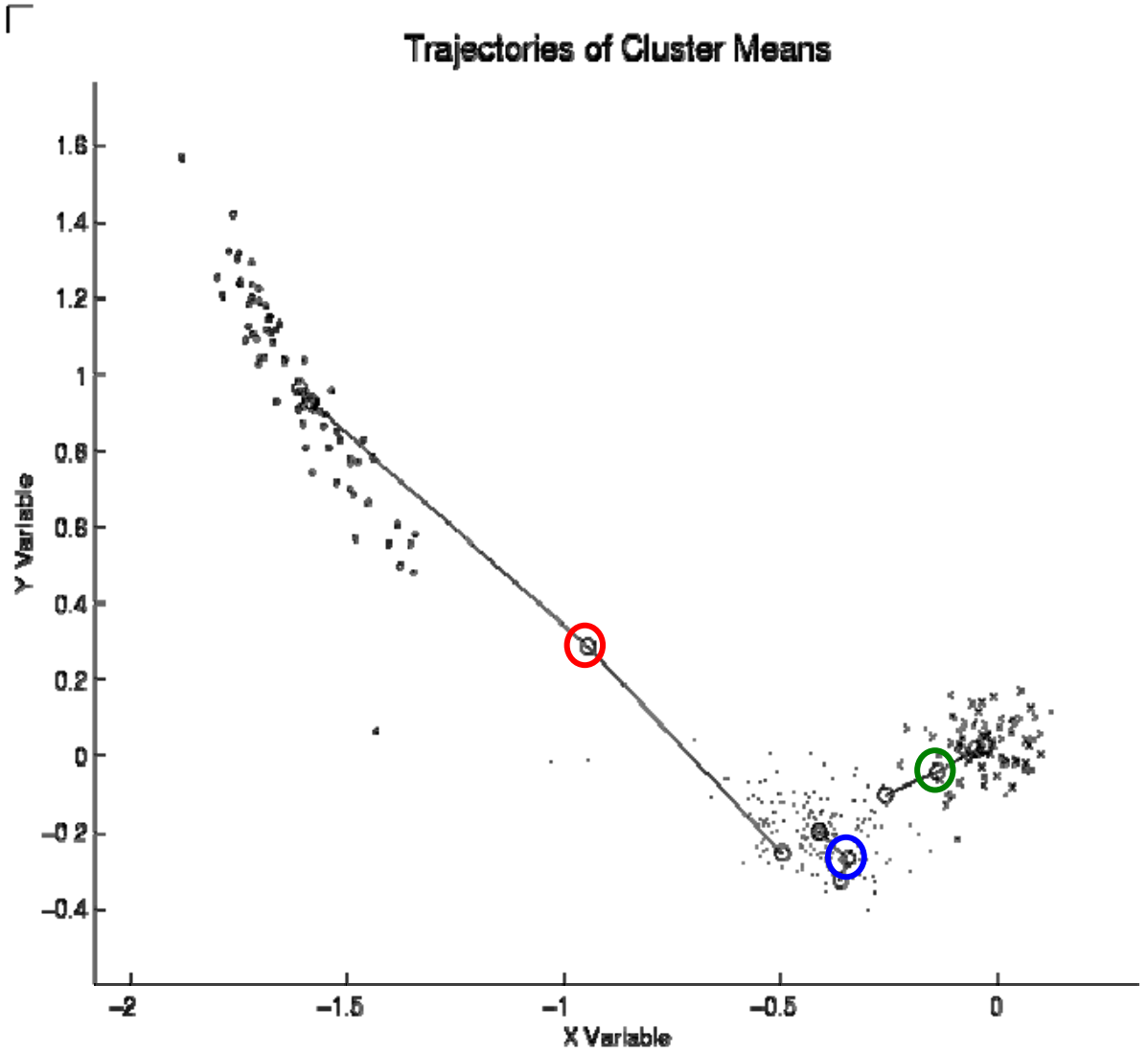


Iteración = 2

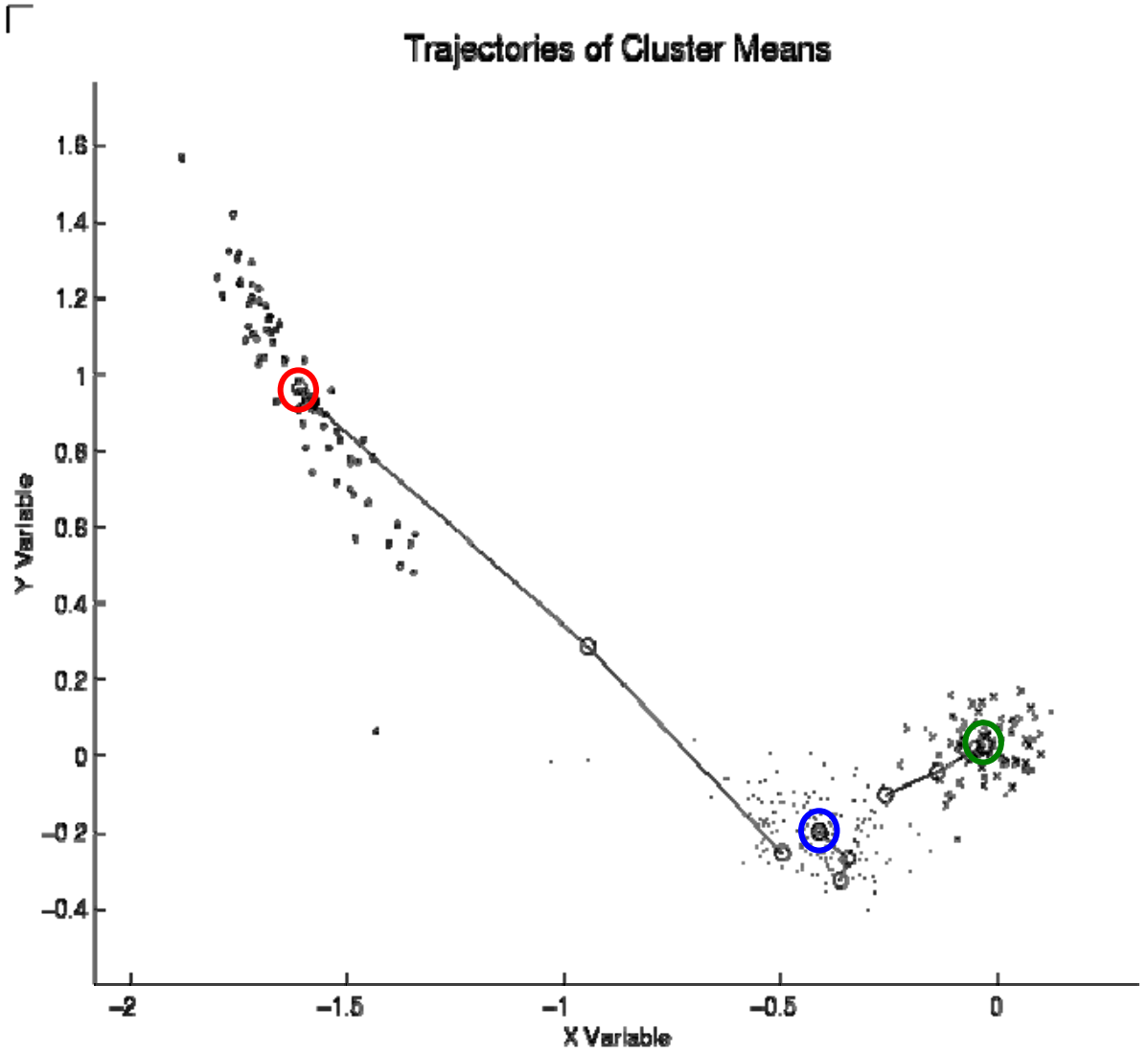
Ejemplo:


Trajectories of Cluster Means
Y Variable
X Variable
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−2
−1.5
−1
−0.5
0

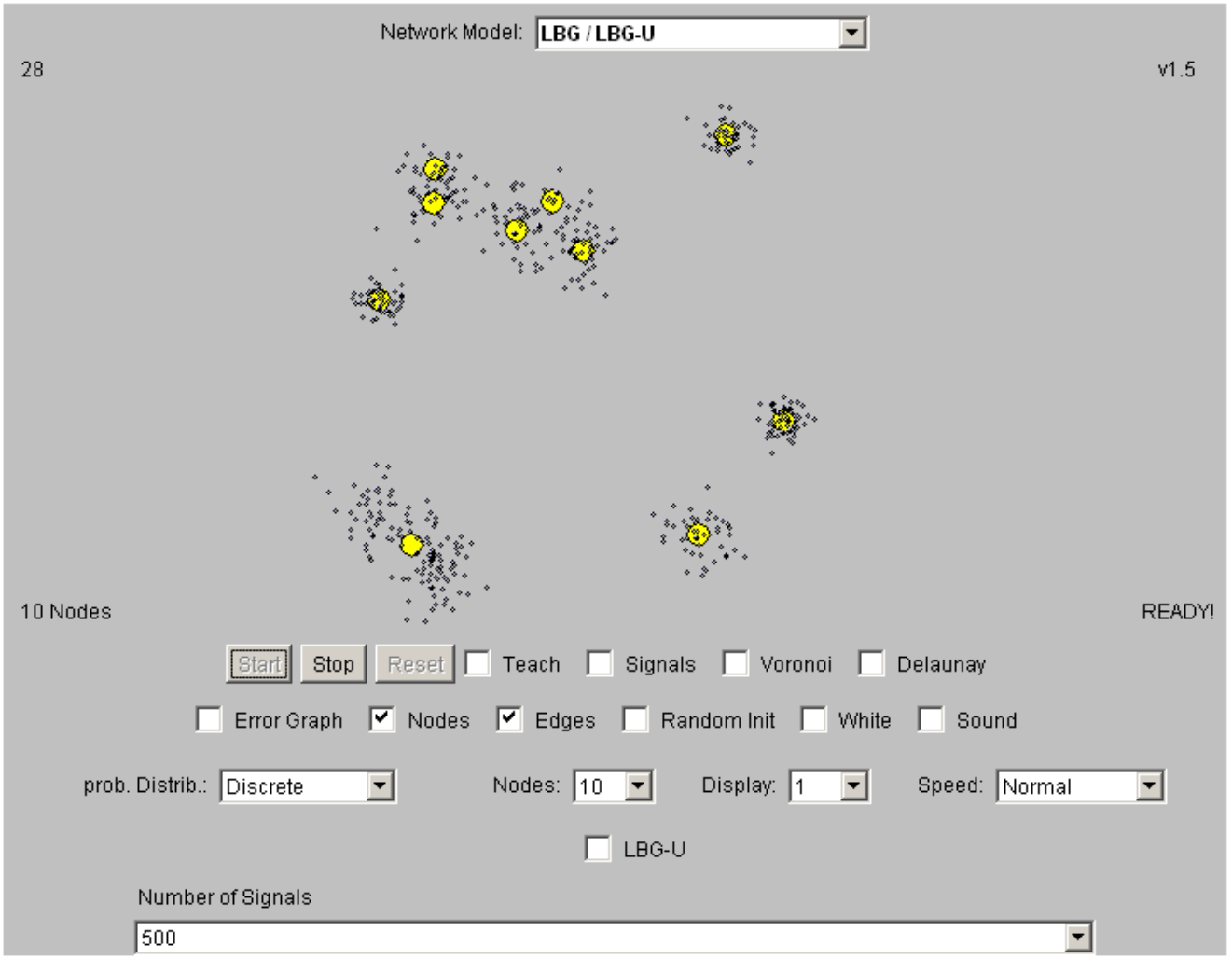
Ejemplo:


Trajectories of Cluster Means
Y Variable
X Variable
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−2
−1.5
−1
−0.5
0

Ejemplo:


Trajectories of Cluster Means
Y Variable
X Variable
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−2
−1.5
−1
−0.5
0

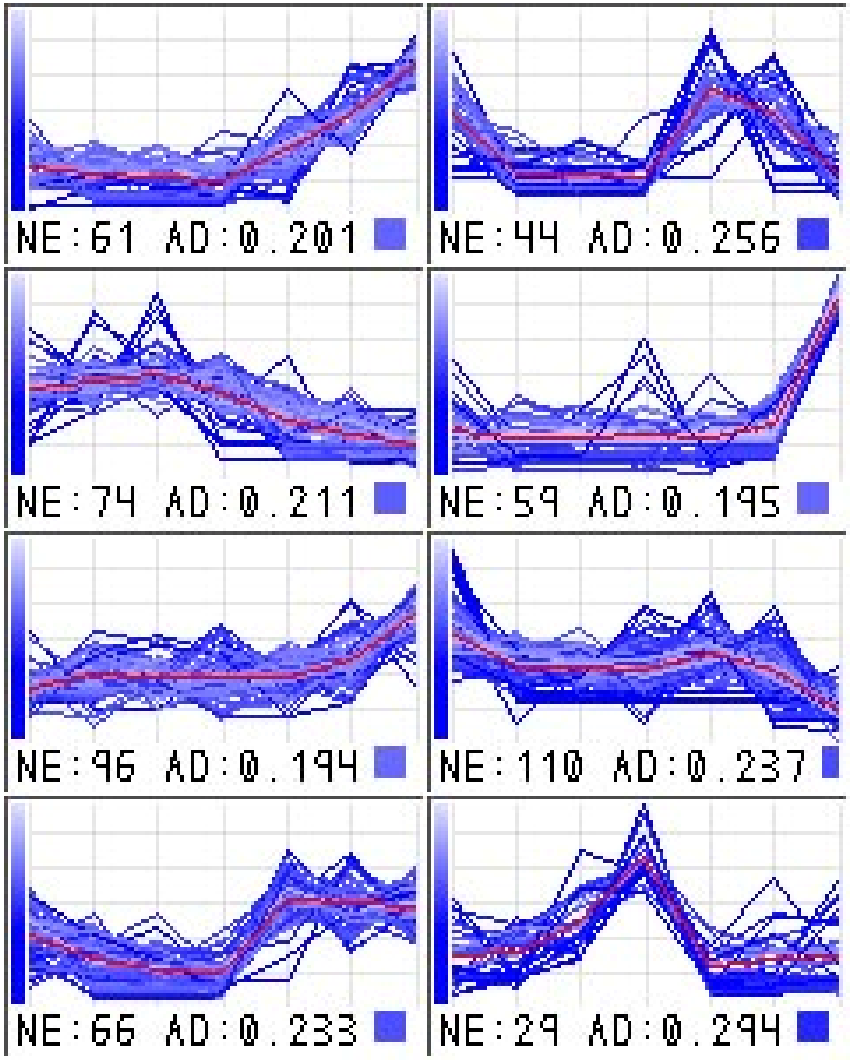
Demo de K-means:



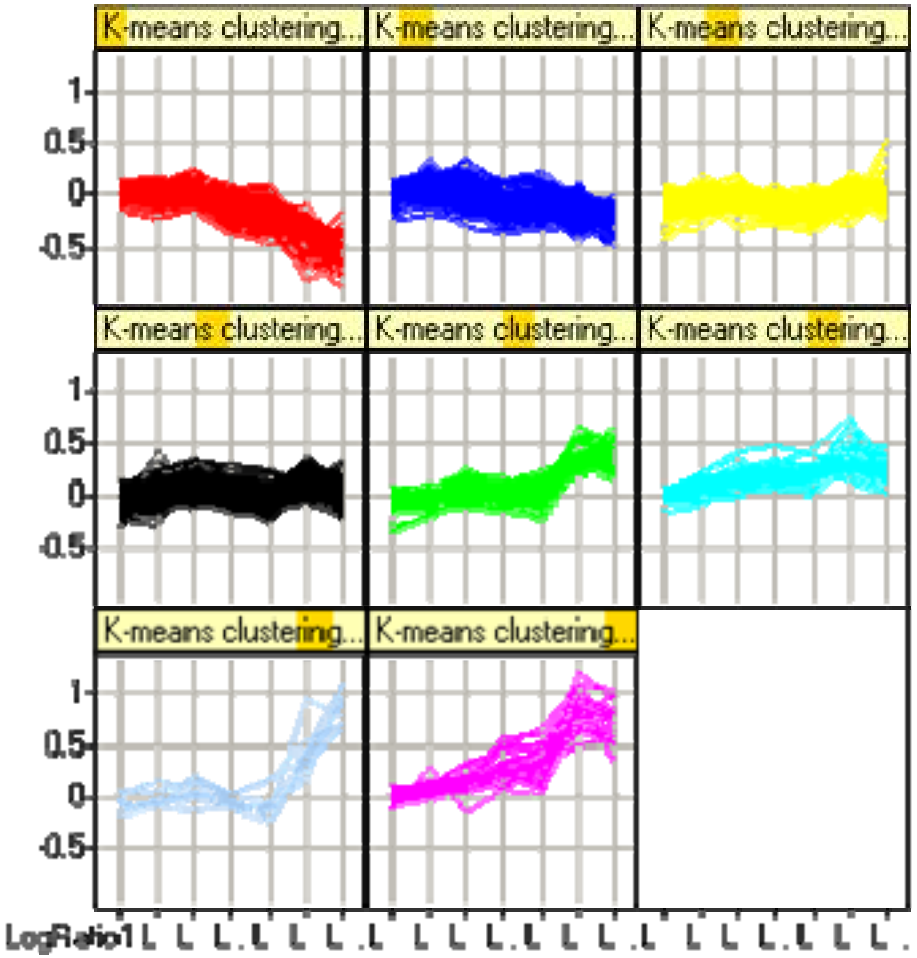
Clustering particional: K-means

- Los resultados dependen de las posiciones iniciales de los centroides.
- Algoritmo rápido: solo calcula las distancias de los puntos de datos a los centroides.
- El número de clusters hay que decidirlo de antemano (gran desventaja!)

Ejemplo de K-Means:

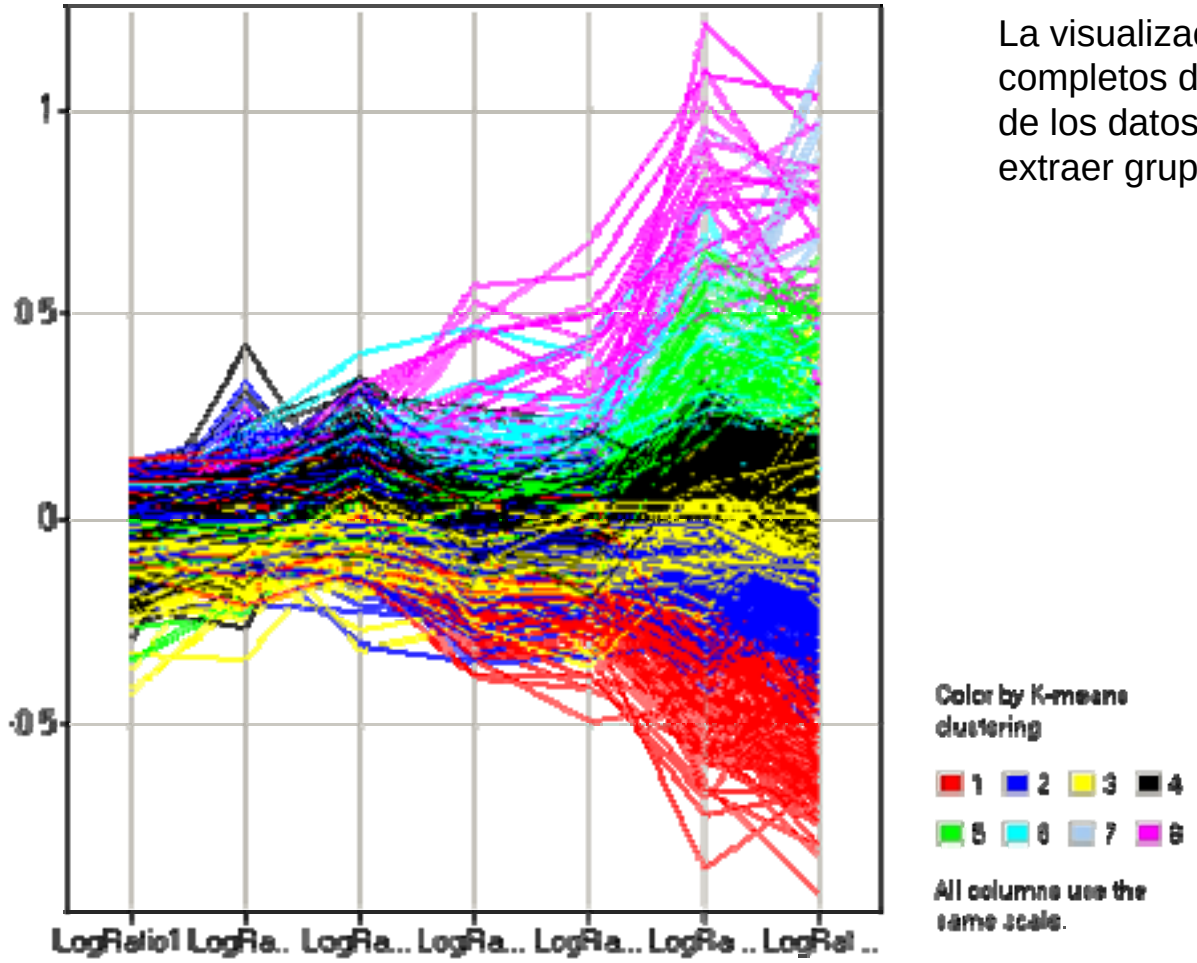


K-Means Clustering



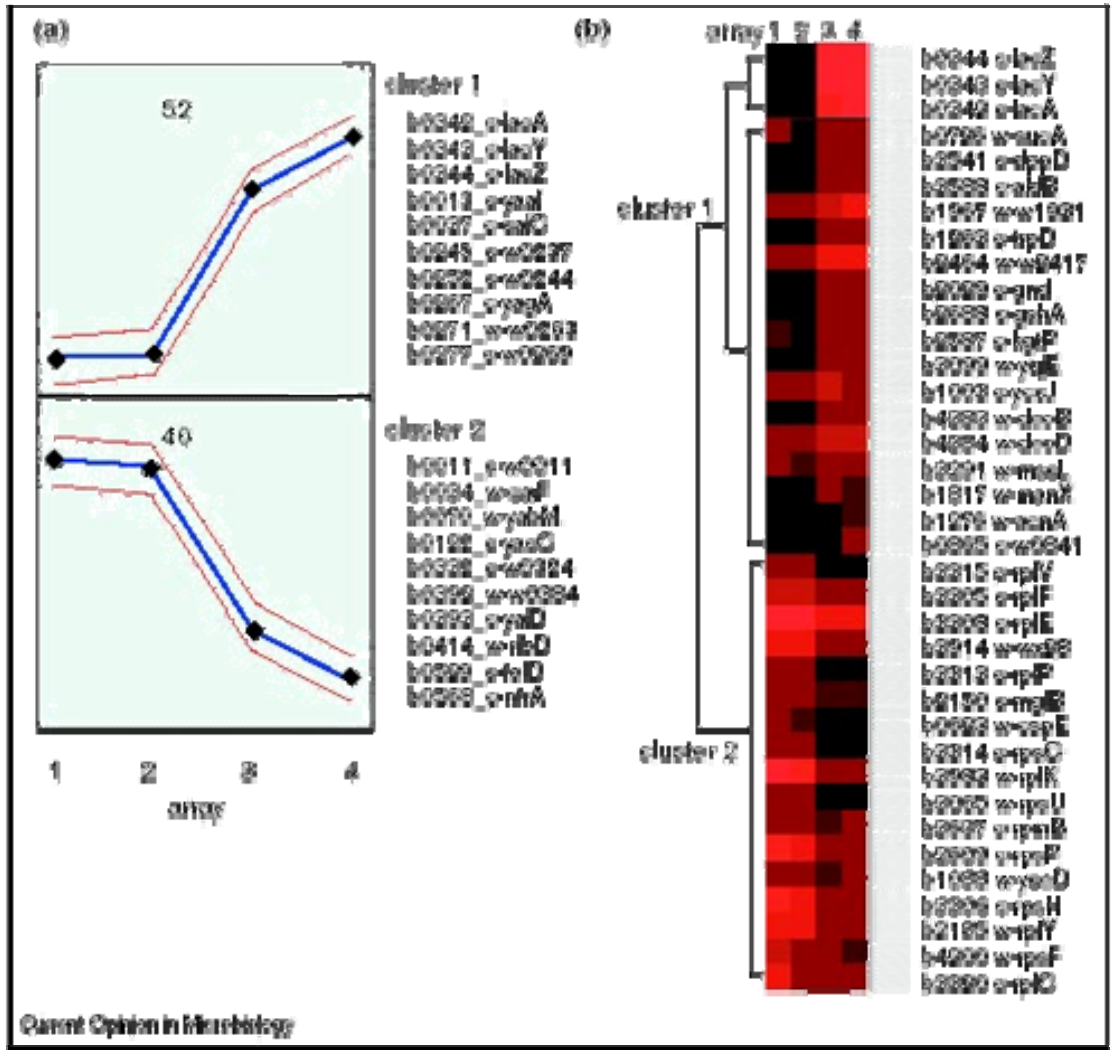
8 Grupos obtenidos utilizando k-means con distancia de correlación.

Visualización de perfiles



La visualización de perfiles completos da una idea de la forma de los datos, pero es muy difícil extraer grupos de manera visual.

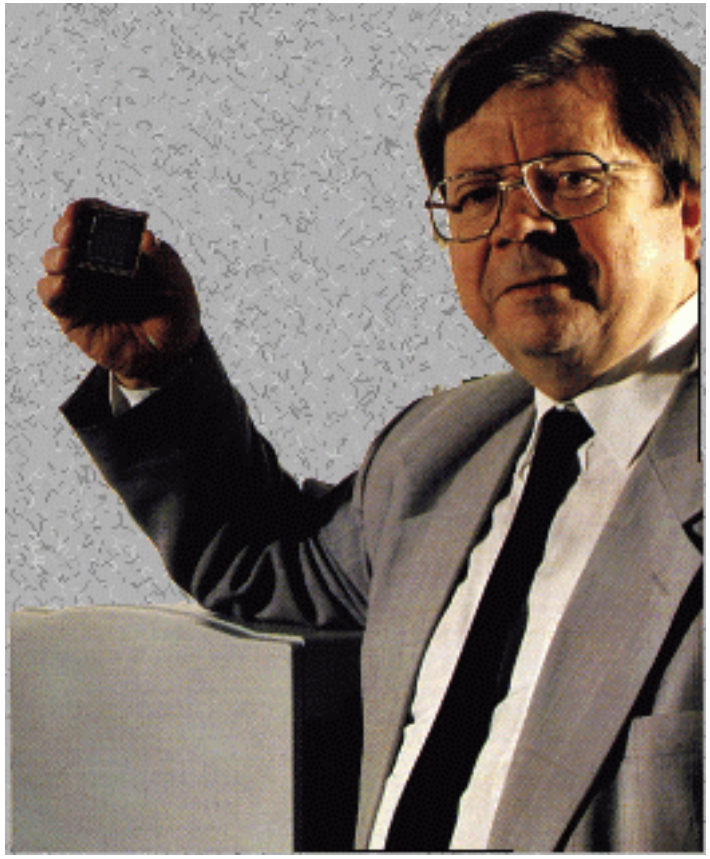
K-Means vs. HCA:



Redes Neuronales: Self-Organizing Maps (SOM)

- Es un modelo de red neuronal que simula la auto organización de las neuronas que se produce en el cortex del cerebro humano cuando le es presentado un estímulo.
- Tiene la capacidad de crear un conjunto mucho menos de datos que son fieles “representantes” de los datos originales.
- Conocidos como Mapas Auto-organizativos de Kohonen

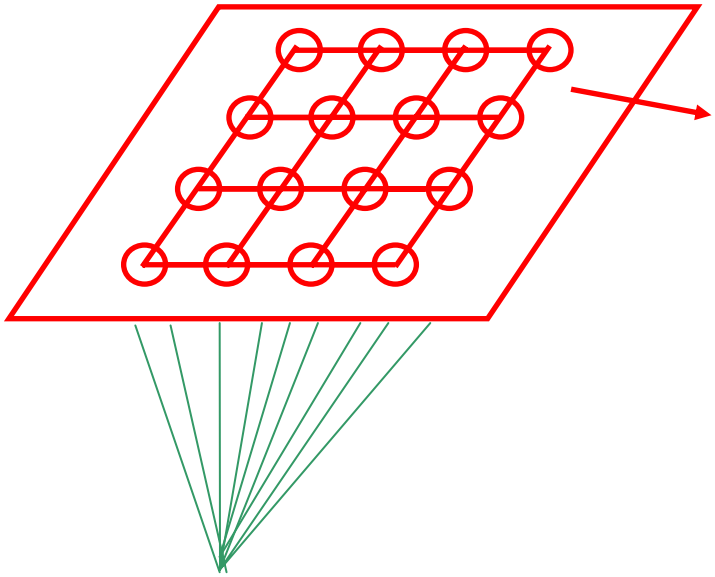
Teuvo Kohonen



Dr. Eng., Emeritus Professor of the Academy of Finland; Academician

Estructura del SOM:

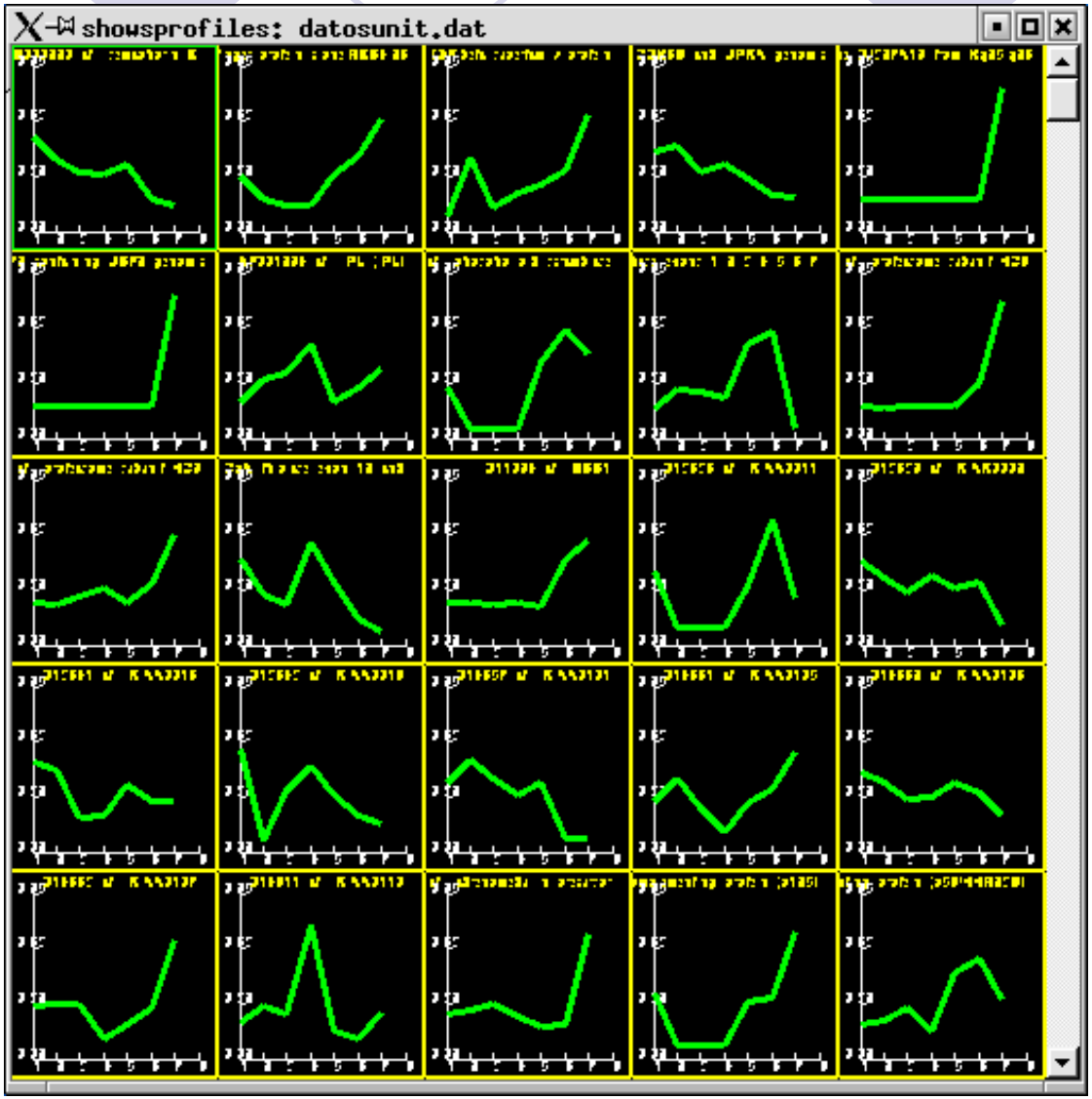
Self-Organizing Map



Neuronas

Datos de entrada: x_0, x_1, x_2,..,x_n

DNA Arrays. Datos originales:

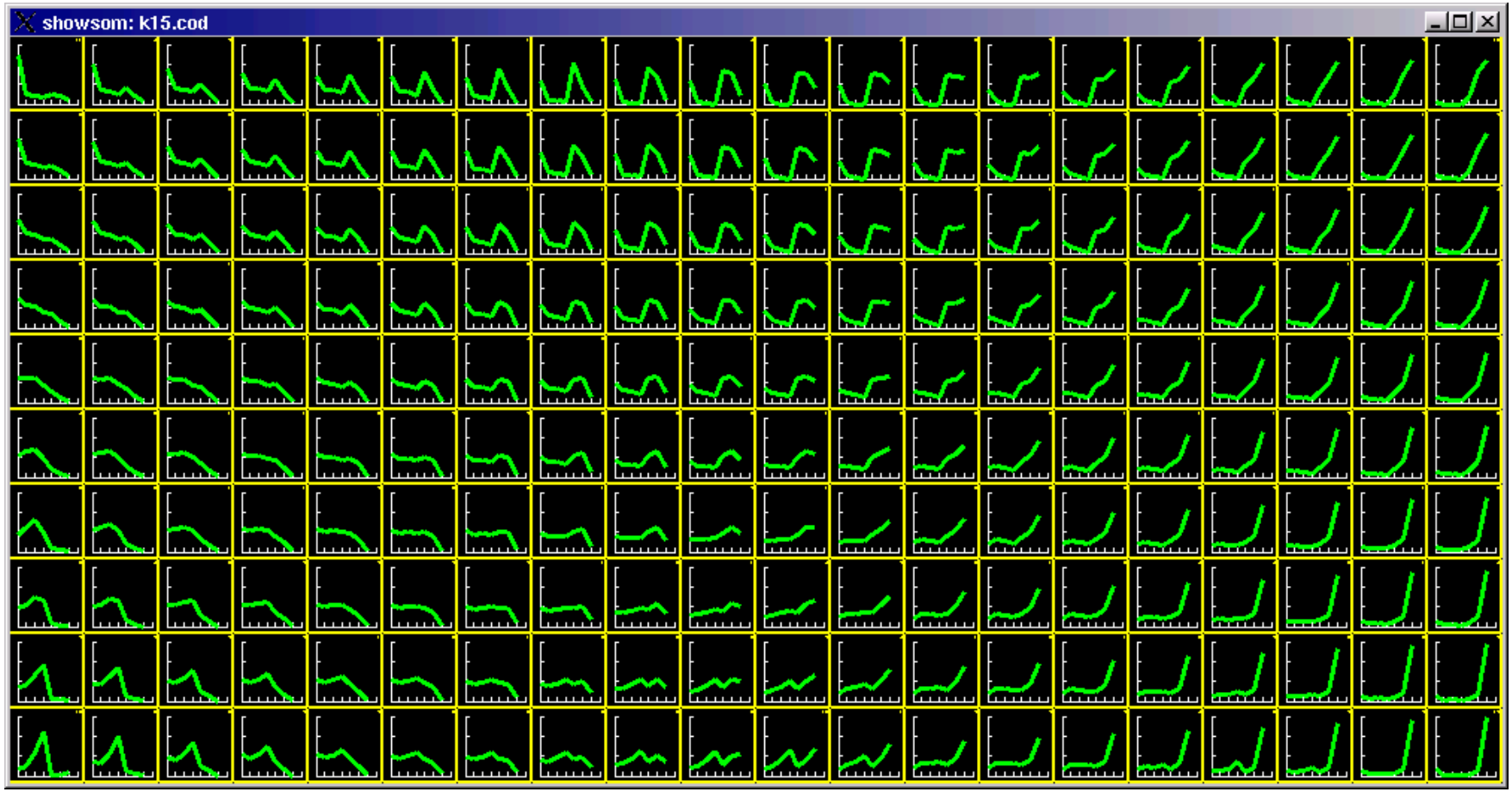


Gene expression behaviour of ultraviolet response on the skin (keratinocytes cells)

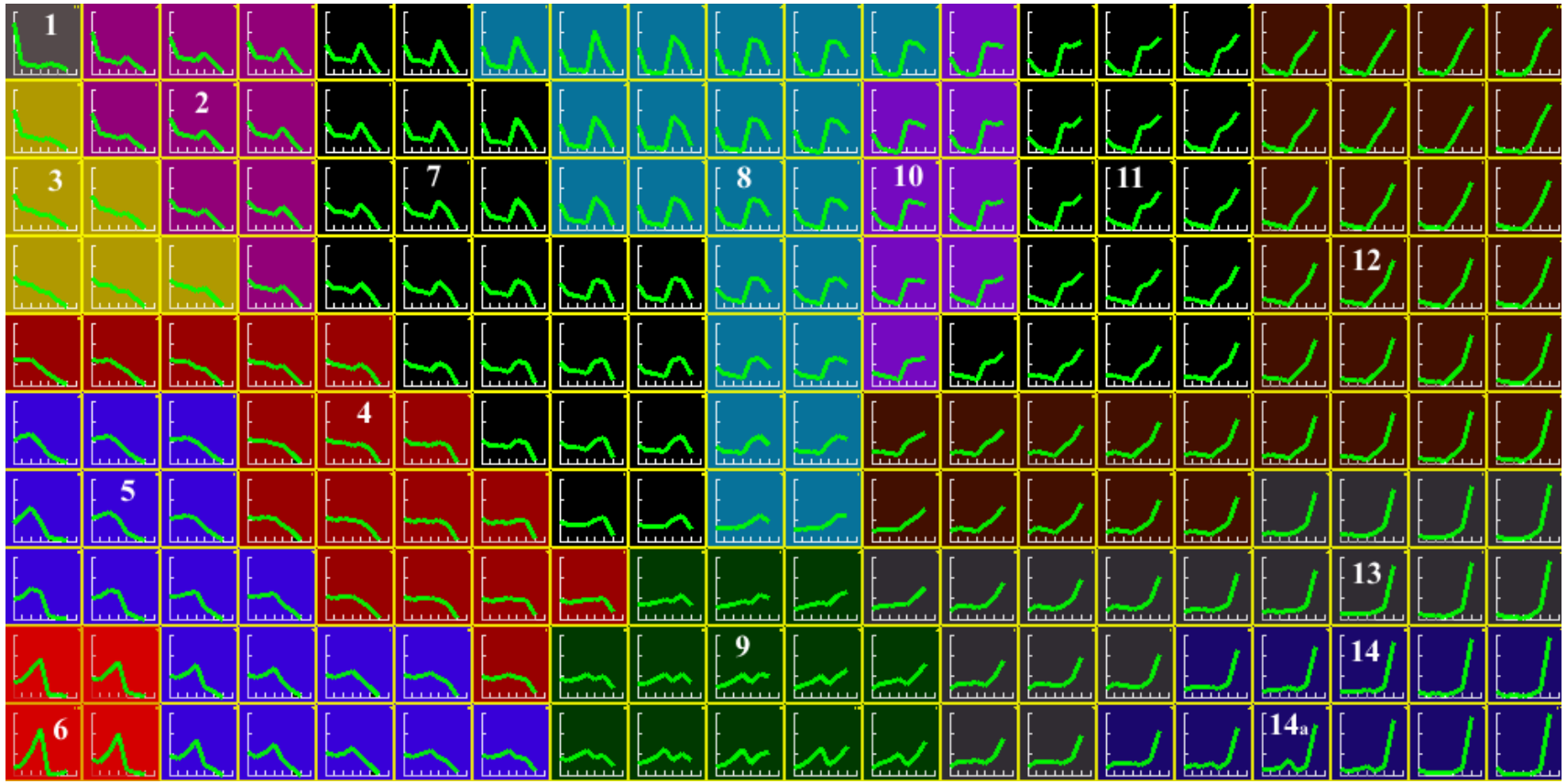
Experimental points:

1 Control
2 10 mJ/cm^2 4h
3 20 mJ/cm^2 4h
4 40 mJ/cm^2 4h,
5 10 mJ/cm^2 24h,
6 20 mJ/cm^2 24h
7 40 mJ/cm^2 24h.

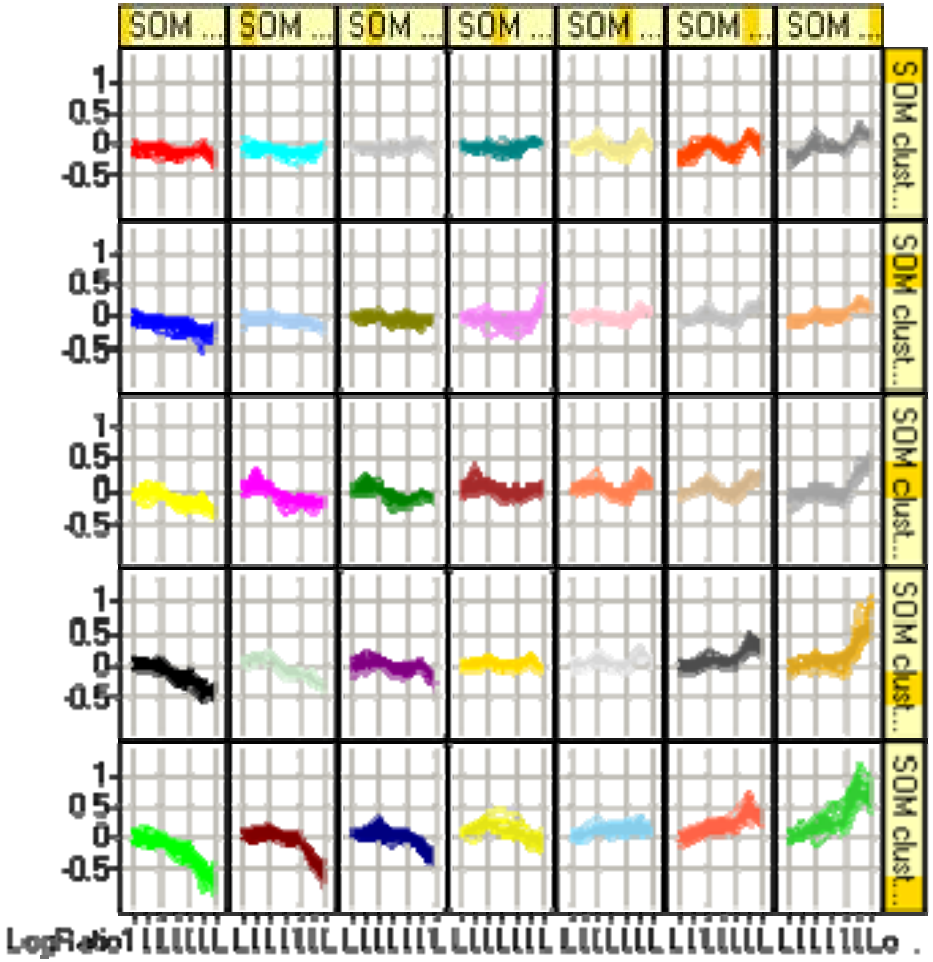
SOM: 20x10 neuronas



SOM agrupado:



Self-Organizing Maps



Generated by Self-Organizing Maps

Grid size (width x height): 7 x 5

Neighborhood function: Bubble
Radius (begin x end): 2.5 x 0

Learning function: Linear
Initial rate: 0.05

Number of training steps: 12500

Output parameters:
MappingPrecision: 6.823e-3
TopologyPreservation: 0.1

¿Qué método utilizar?

- Desgraciadamente **no existe un consenso** sobre cual es el mejor método a utilizar para hacer agrupamiento.
- No existe una manera simple de decidir cual es el mejor a partir de un conjunto de datos experimentales.
- Recomendación; Usar toda la información disponible y **utilizar varios métodos de agrupamiento y comparar**!

Paso final: Búsqueda de funciones de los genes

- Una vez creados los clusters, el paso final sería la búsqueda de las funciones de los genes que pertenecen a cada uno de los clusters.
- Generalmente los chips están compuestos por ESTs, lo que hace el proceso de búsqueda más largo.
- Bases de datos frecuentemente utilizadas para esto: GenBank, UniGen, OMIM, GeneCards, SwissProt, etc

Limitaciones:

Aunque los DNA microarrays son relativamente fácil de utilizar, existen ciertas limitaciones relacionadas con la información que estas técnicas brindan:

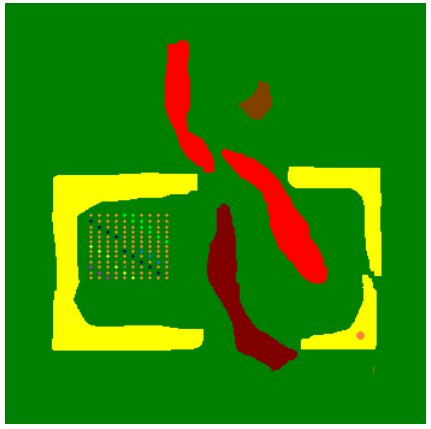
- El análisis del DNA **no puede predecir si las proteínas están en un estado activo**.
- A pesar de la correlación existente entre la cantidad de mRNA producido en la célula y la cantidad de proteína sintetizada, su cuantificación no es directa, por lo que **la cuantificación del RNA no siempre refleja los niveles correspondientes de proteínas**.
- **Múltiples proteínas pueden ser obtenidas de un mismo gen** cuando se tienen en cuenta la postraducción y el mRNA splicing.

Por lo tanto la técnicas de microarray solo permiten una estimación cualitativa del proteoma. Técnicas mas avanzadas se necesitan para el estudio del proteoma: **La proteómica**.

Software:

SpotFire DecisionSite for Functional Genomics

www.spotfire.com



Engene: Gene Expression Data Processing and Exploratory Data Analysis

www.biocomp.cnb.uam.es