Introduccion a la Bioinformática

Data Analysis | Data Clustering

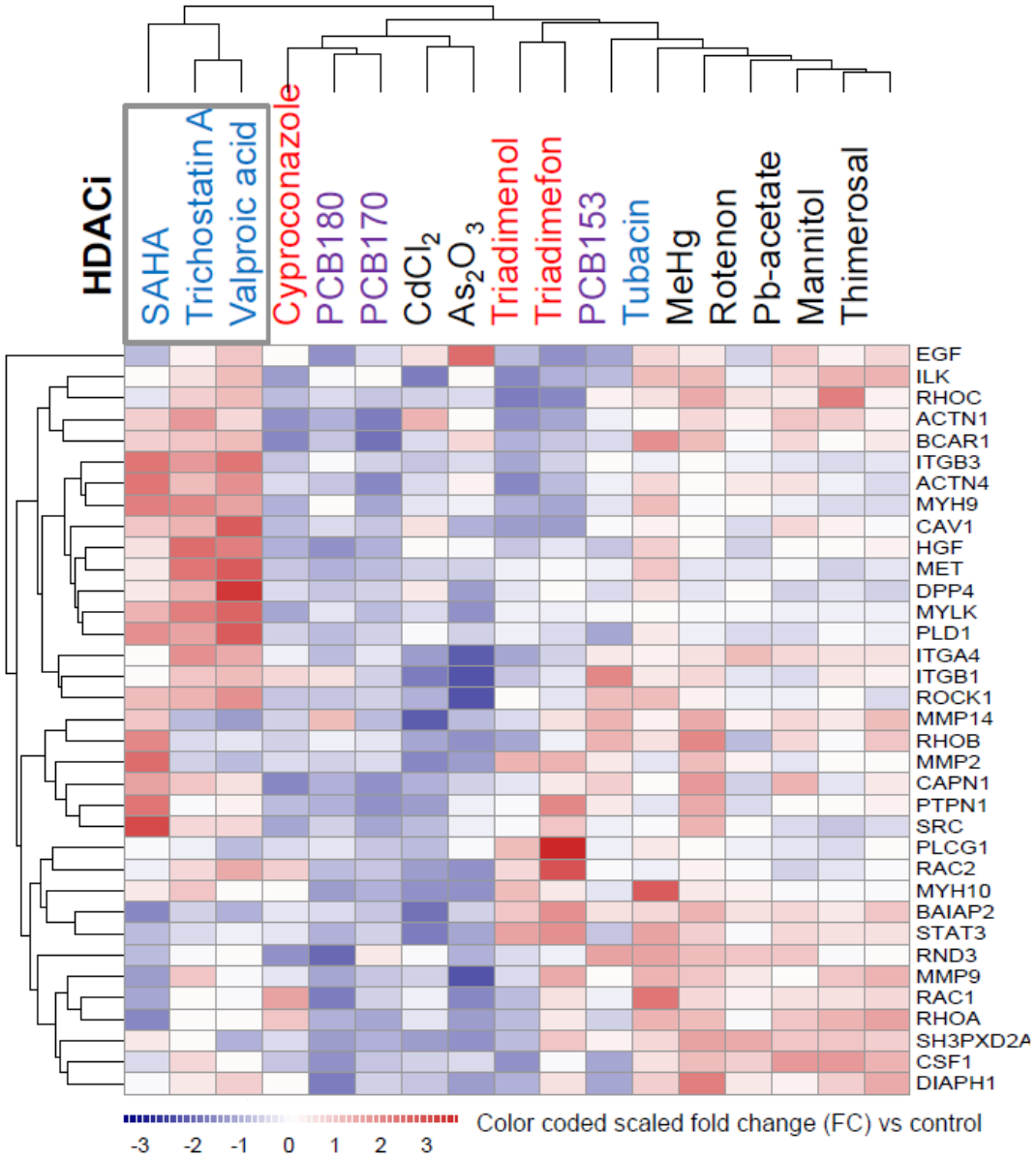
Fernán Agüero

Instituto de Investigaciones Biotecnológicas, UNSAM

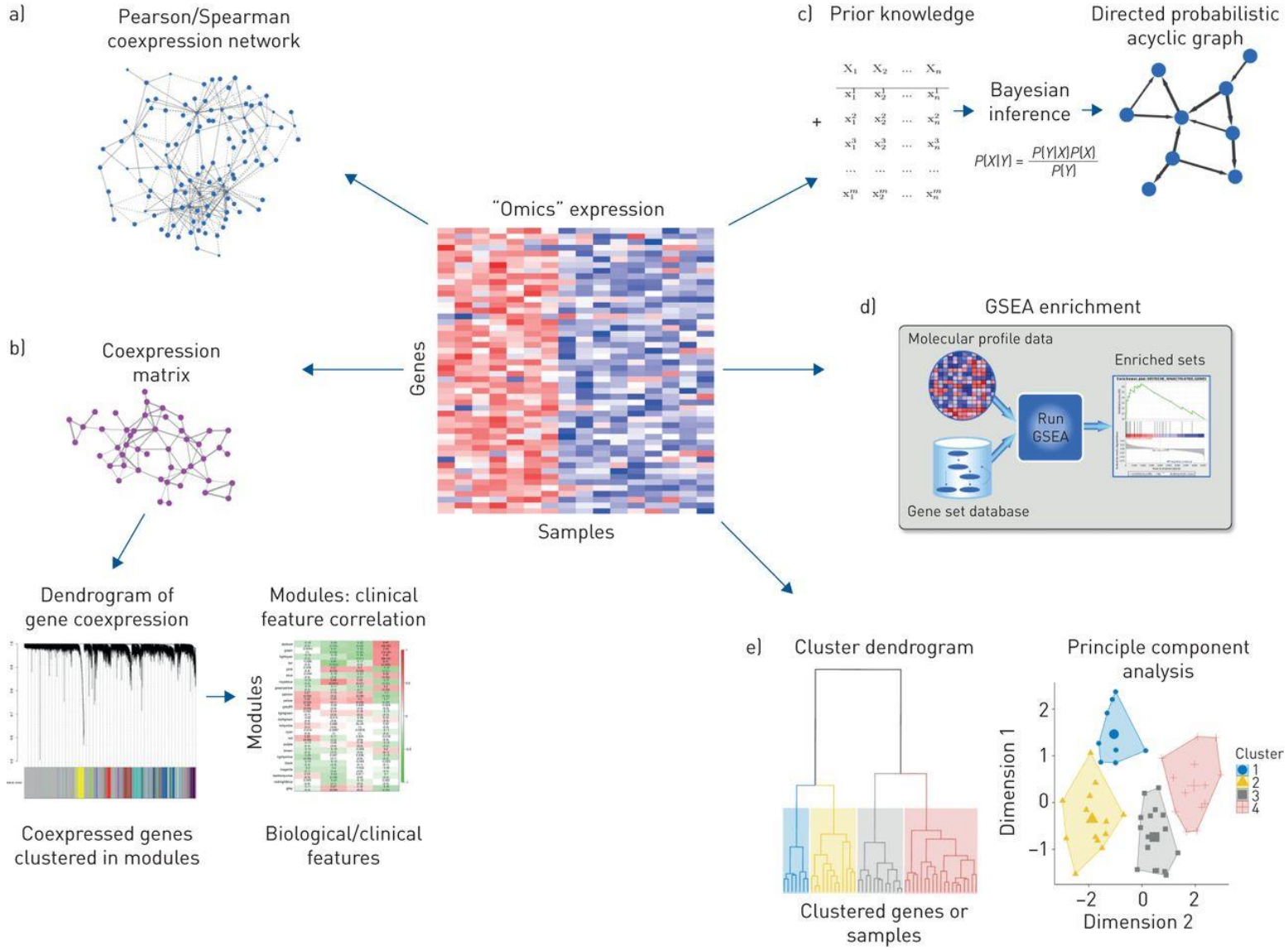
Por qué queremos analizar datos?

- Ayudar a la intuición | Construir intuición
  - Encontrar estructura / orden interno en datos altamente dimensionales
- Generación de hipótesis
  - Encontrar y caracterizar grupos similares de objetos en los datos
- Aprender de los datos | Generación de conocimiento
  - Reglas subyacentes, patrones recurrentes, tópicos salientes
- Resumir los datos | Comprimir
  - Reducir dimensionalidad, Generar resúmenes con información relevante (*"Sumarizar los datos"*)
- Visualización
  - Facilita entender los datos al cerebro humano

Por qué queremos analizar datos: generar hipótesis

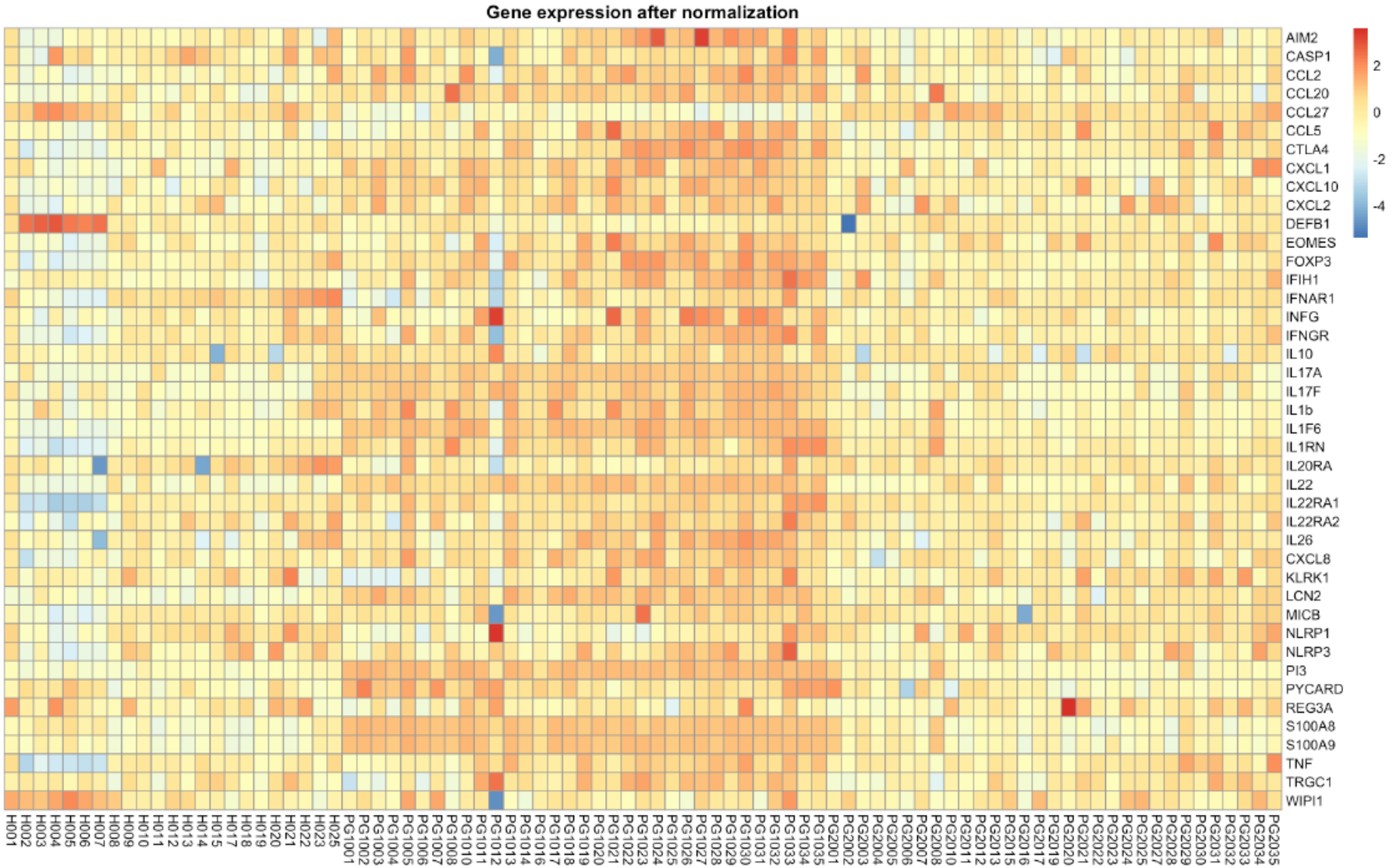


Por qué queremos analizar datos: aprender

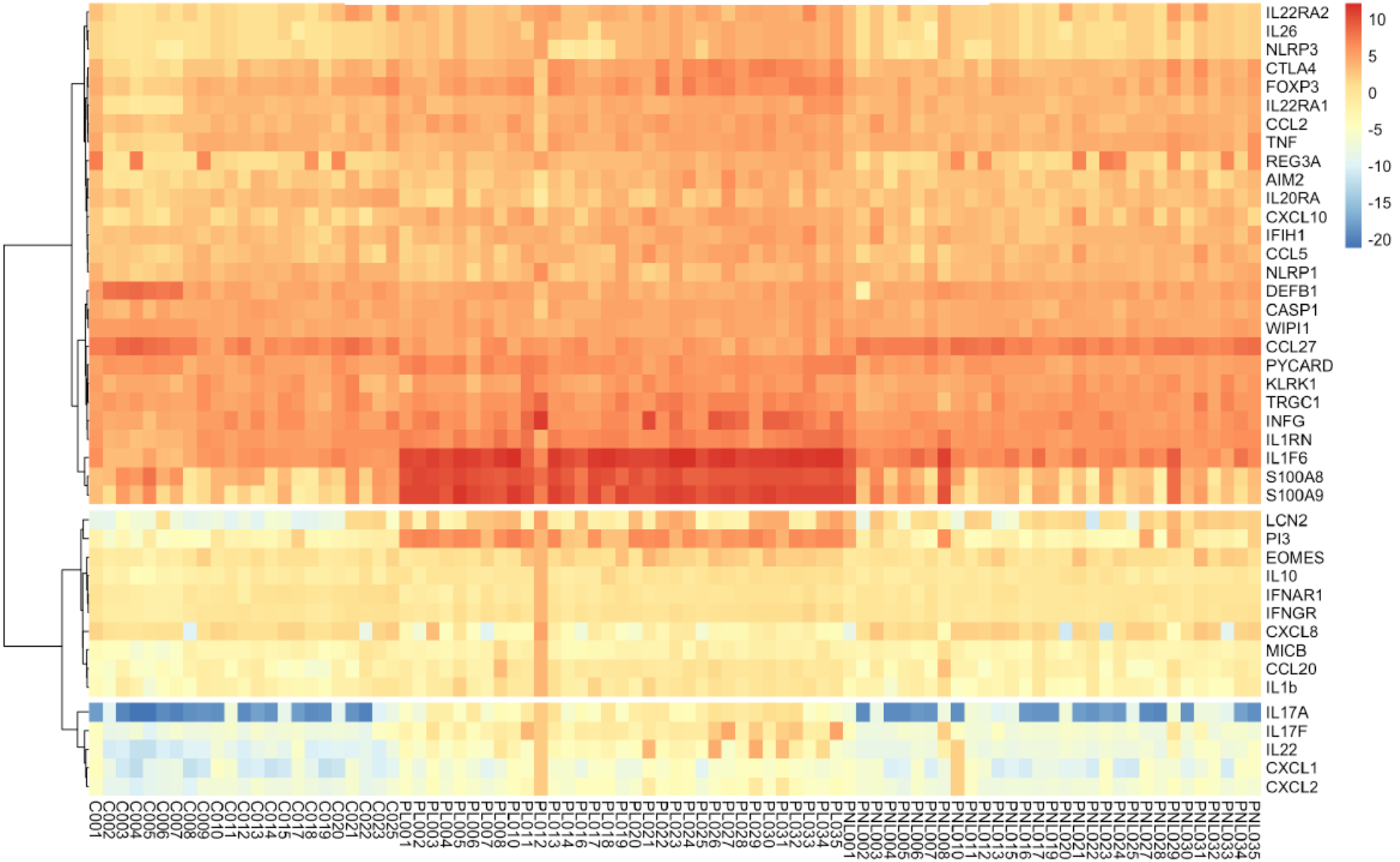


Guillaume Noell et al. Eur Respir Rev 2018;27:170110

Resumir los datos | Comprimir



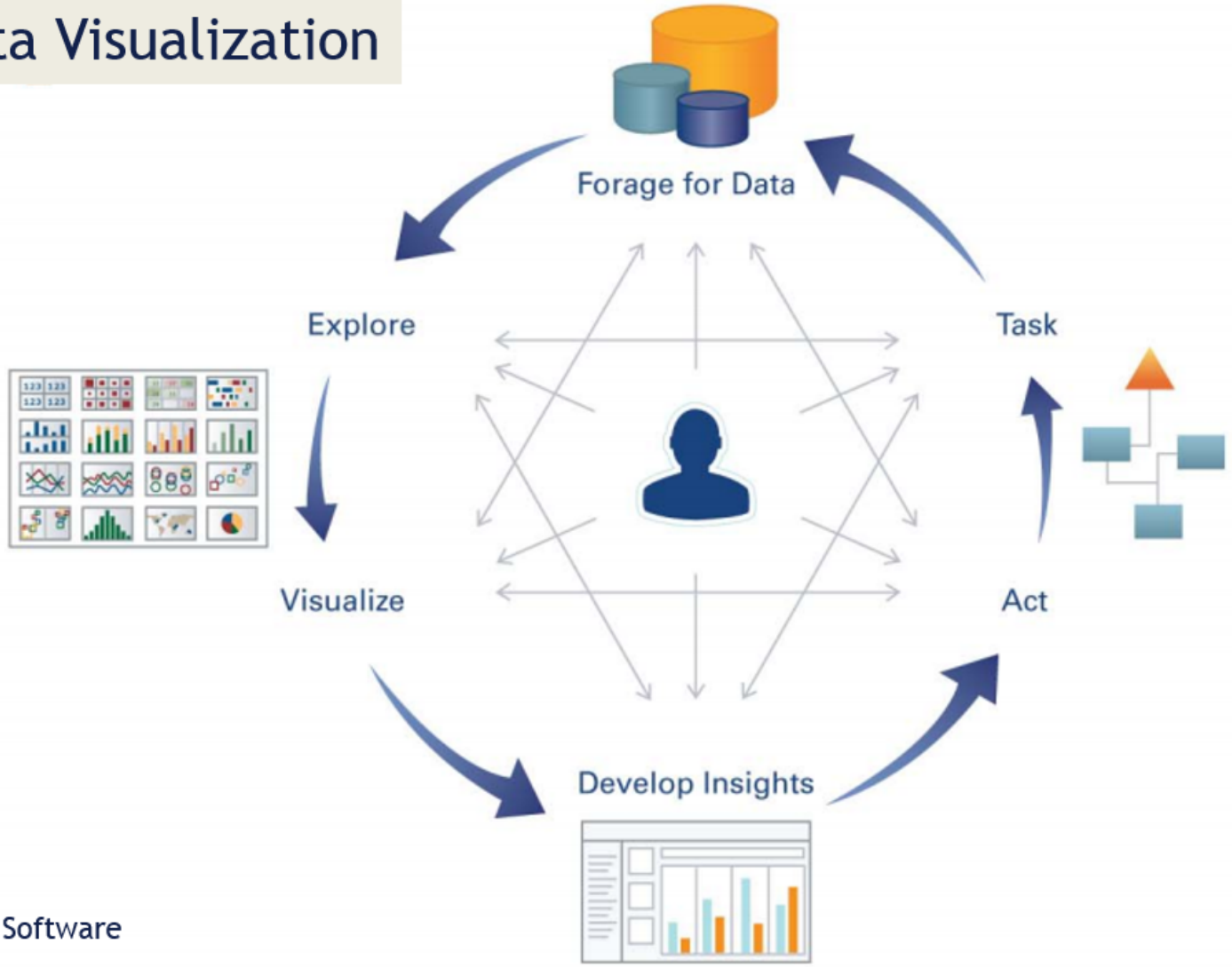
Resumir los datos | Comprimir



Por qué visualizar datos?

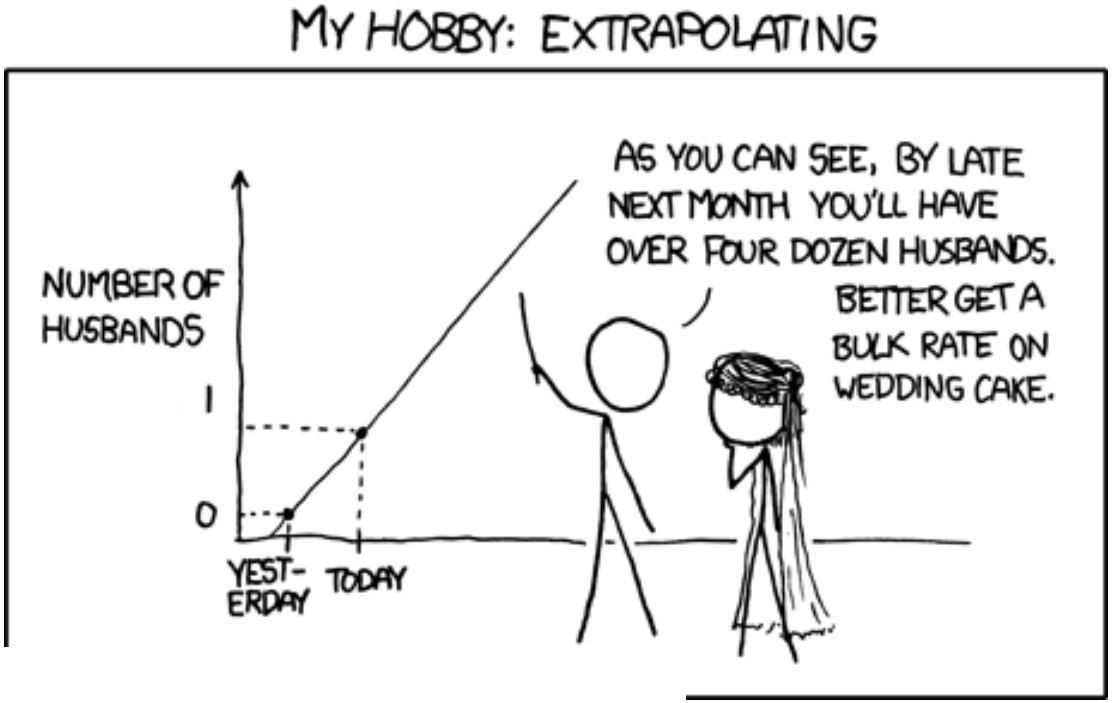
answer questions

With Data Visualization



Source: Tableau Software

Problemas en el análisis de datos



"If you torture the data long enough, it will confess to anything."

RONALD COASE

Economist, Nobel Prize (1991)

Agrupamiento de datos / Data Clustering

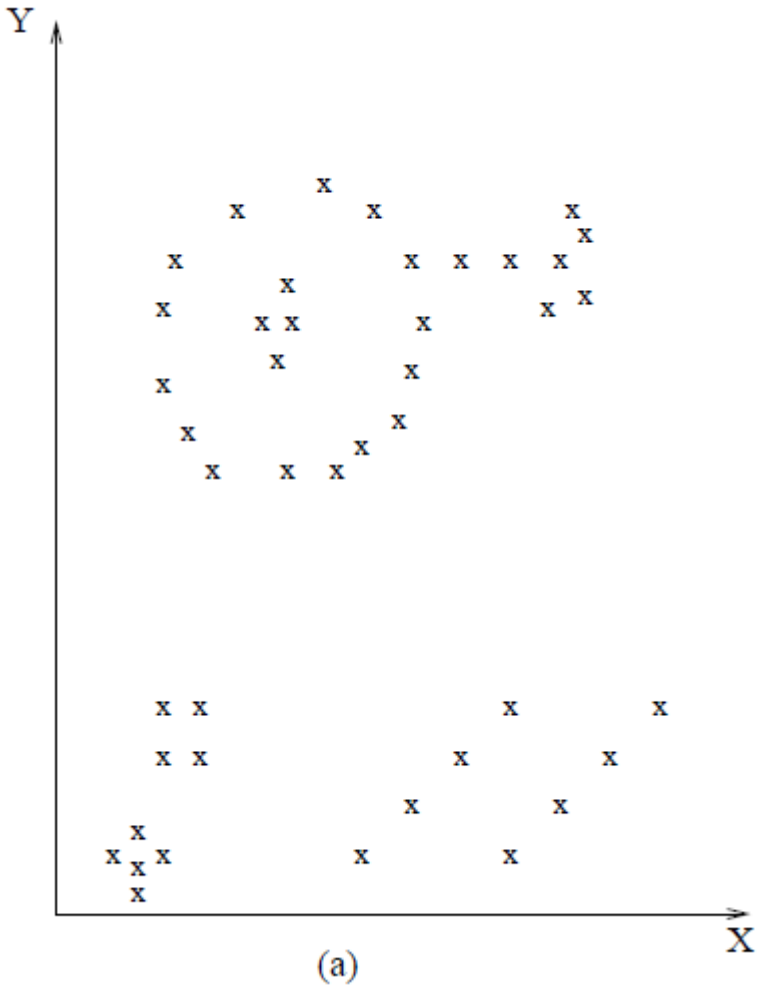
- **El *agrupamiento de datos* consiste en la clasificación de objetos diferentes grupos, de manera que objetos similares son agrupados en el mismo grupo.**
- **Otra definición: particionar un conjunto de datos en subconjuntos o *clusters* de tal manera que estos tengan "algo en común".**
  - **El problema: 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:

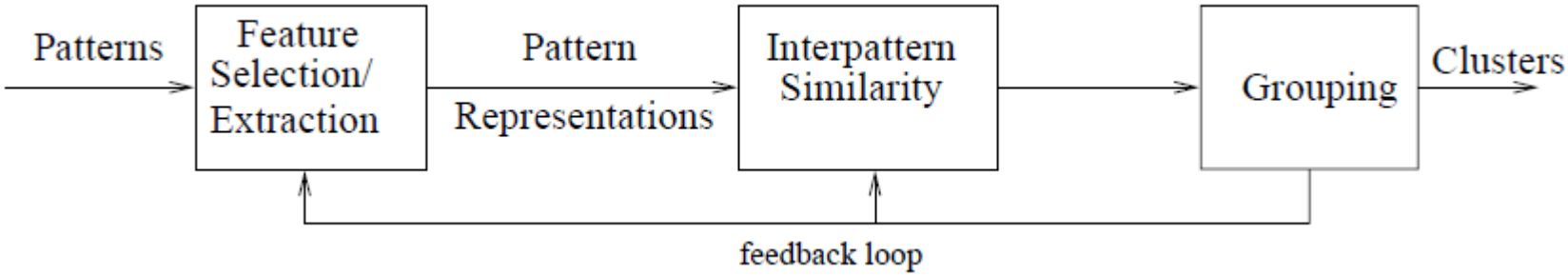
- Secuencias (DNA, RNA)
  - Ej: Agrupar por similitud/identidad global
  - Ej: Agrupar por presencia de motivos o señales
- Medidas de expresión de genes
  - Ej: Agrupar todos los genes que tienen alta expresión
- Abstracts en PubMed
  - Ej: Agrupar abstracts en base a número de palabras compartidas
- 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



(a)

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

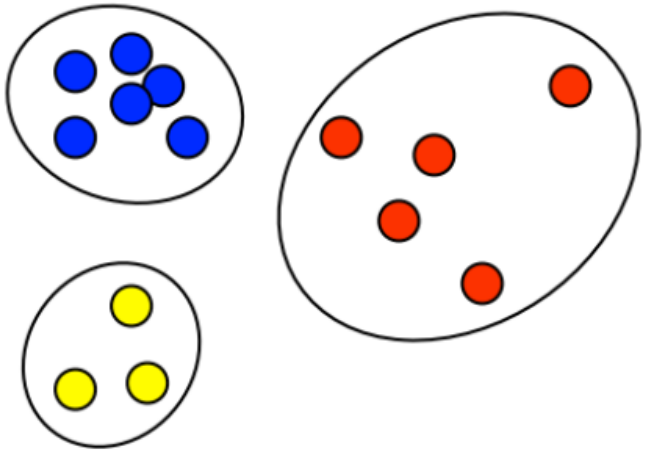
Clustering:

- Hay muchos algoritmos (estrategias) de clustering

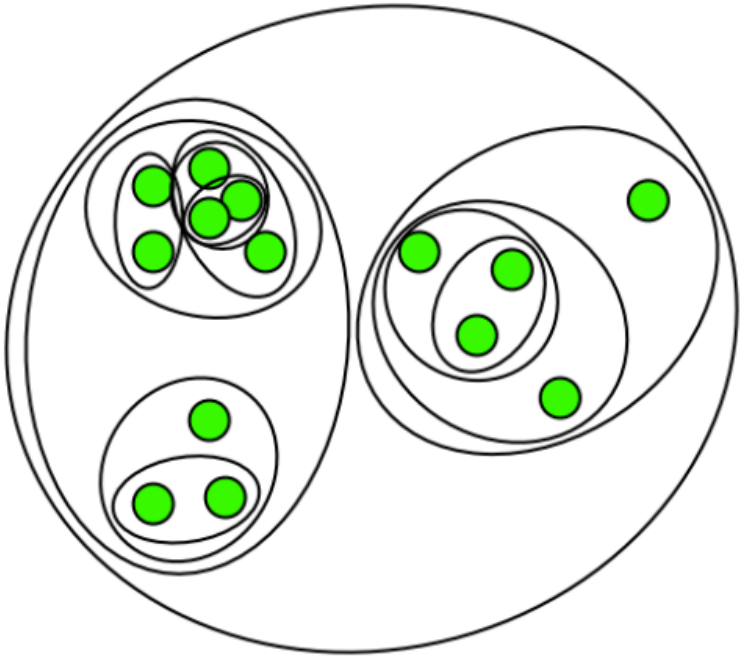
Cluster validation analysis

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

Particional vs Jerárquico



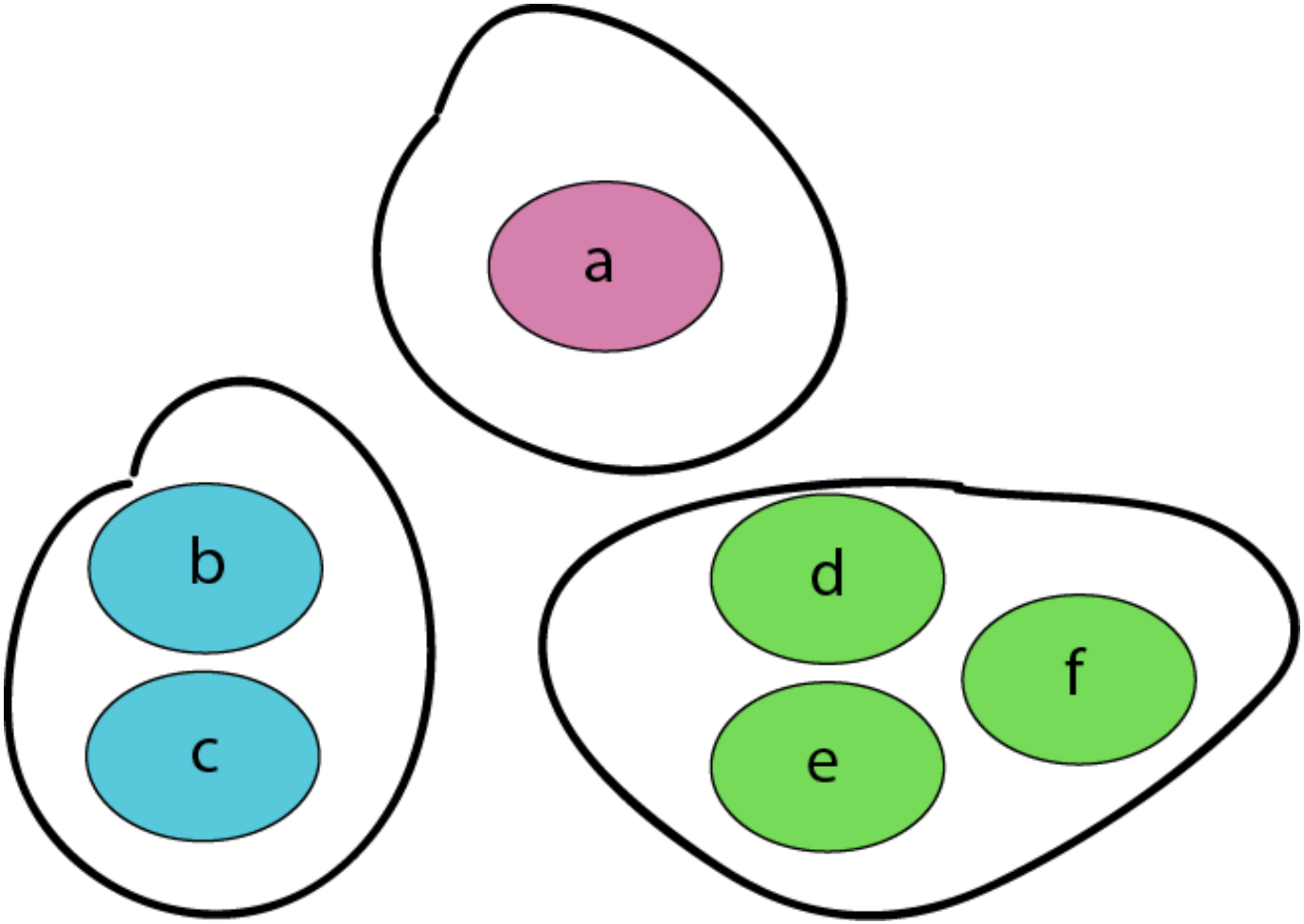
Each sample(point) is assigned to a unique cluster



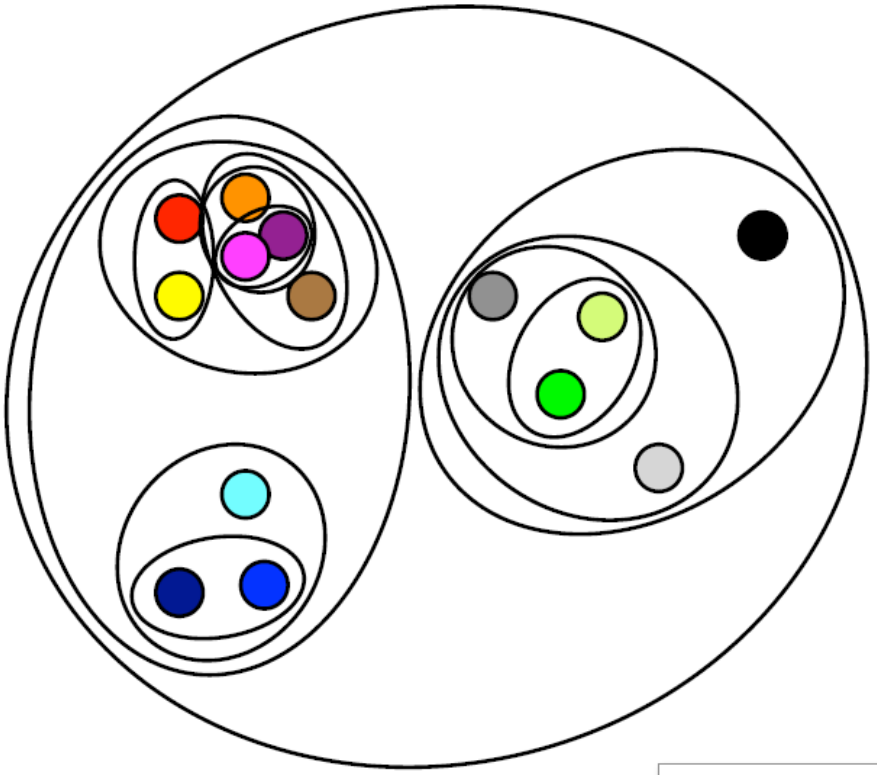
Creates a nested and hierarchical set of partitions/clusters

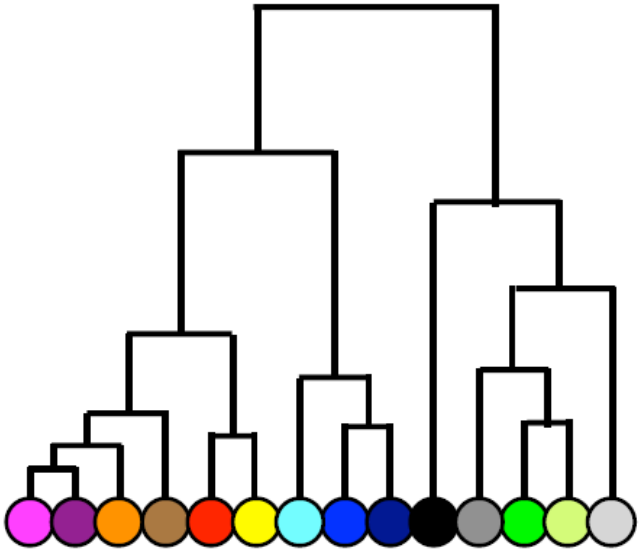
Estrategias de Clustering: clustering particional

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



Hierarchical clustering

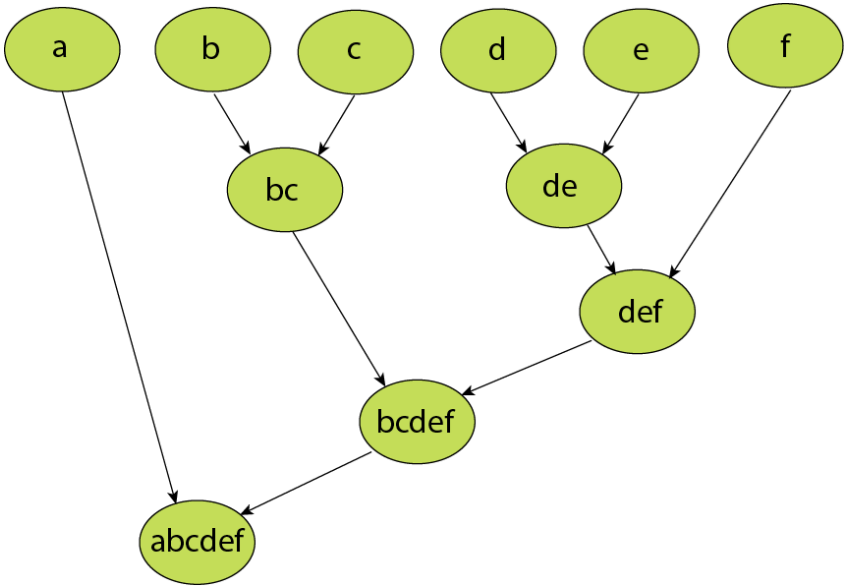




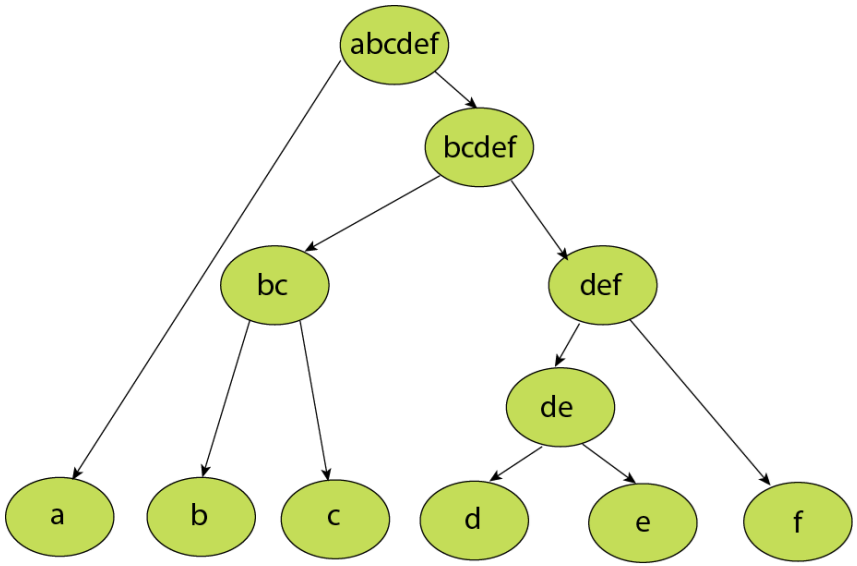
Hierarchical clustering is usually depicted as a dendrogram (tree)

Estrategias de Clustering: Clustering jerárquico

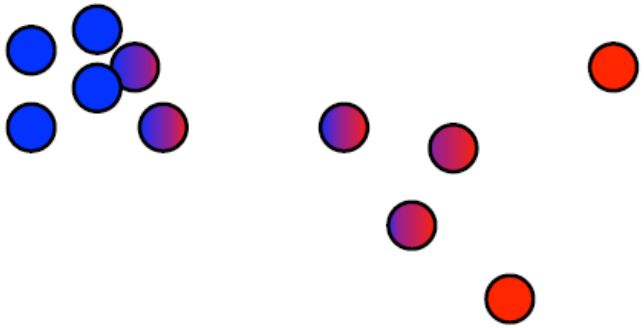
Aglomerativo



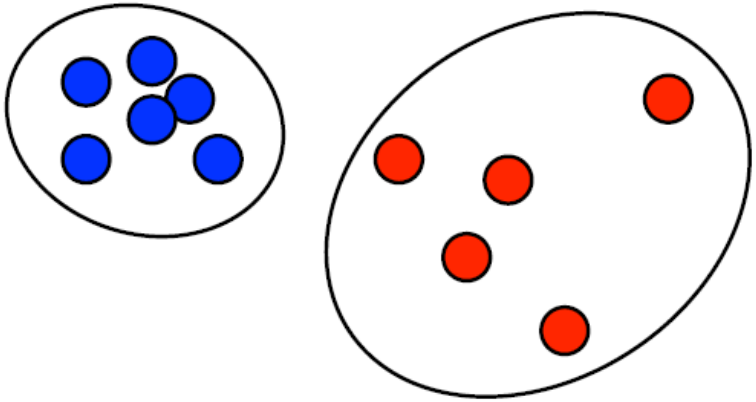
Divisible



Fuzzy vs Non-Fuzzy (Difuso vs No-Difuso)



Each object belongs to each cluster with some weight

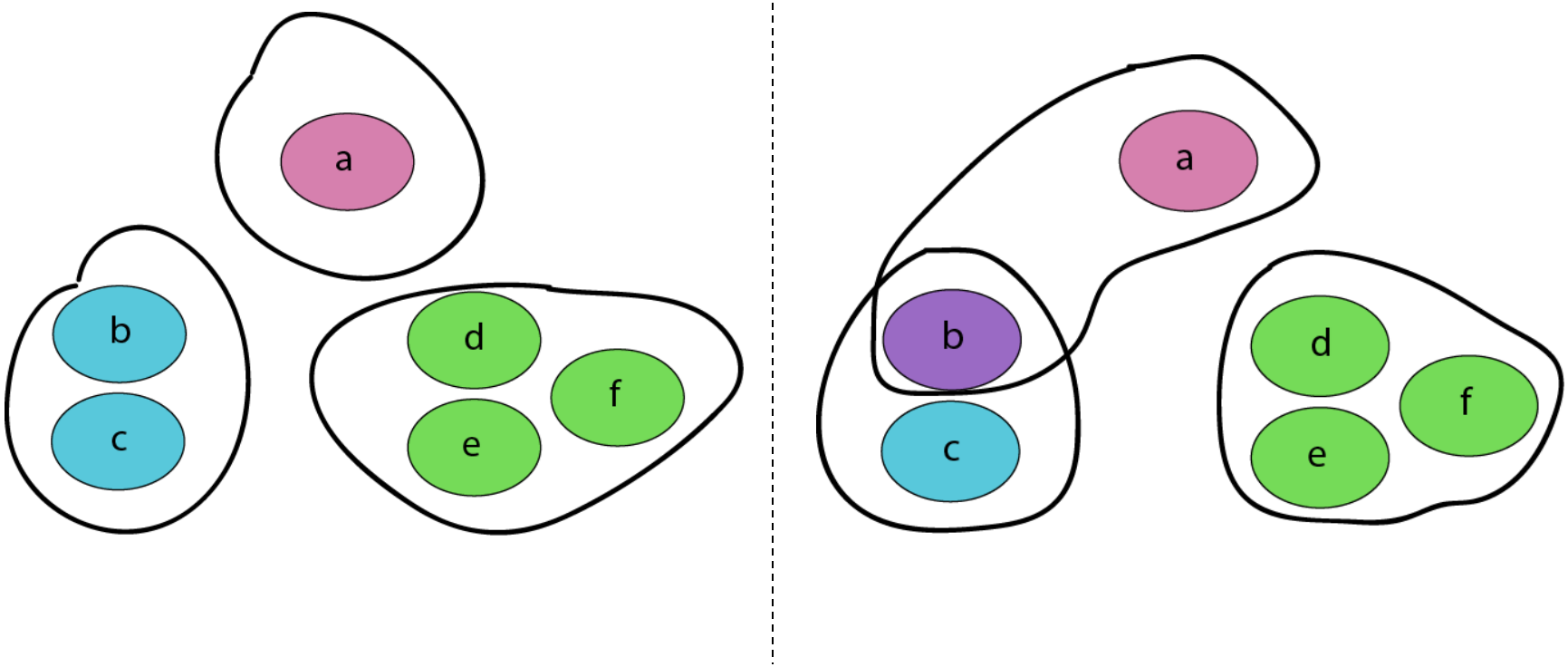


Each object belongs to exactly one cluster

Propiedades de los clusters

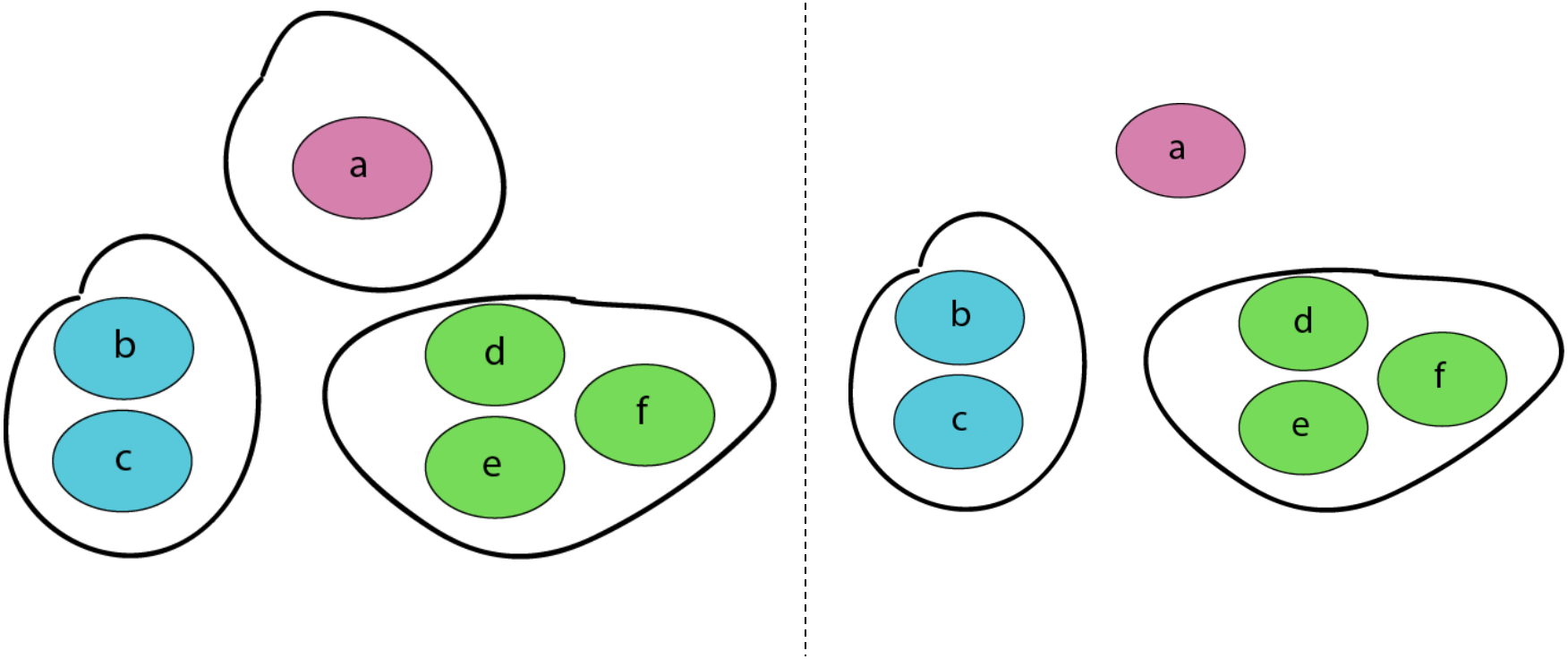
Disjuntos vs. No disjuntos

(hard) (fuzzy)



Propiedades de los clusters

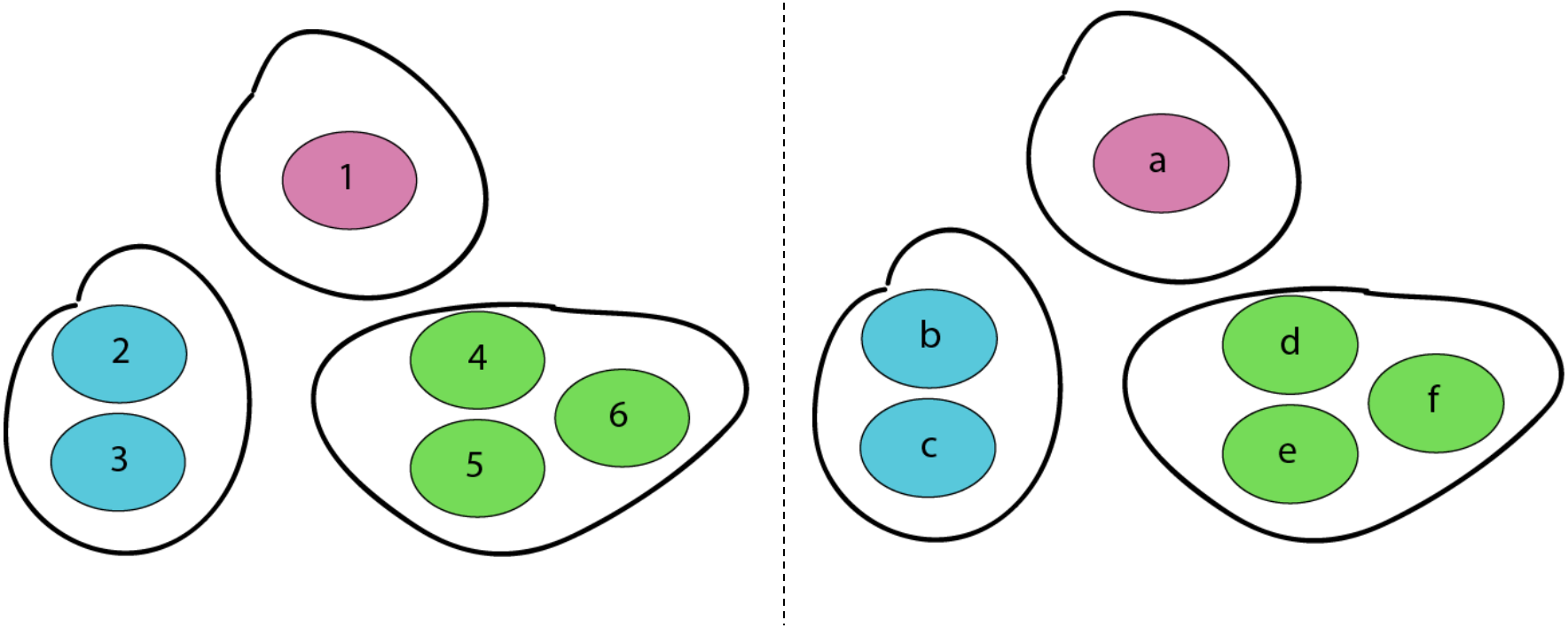
Completos vs. Incompletos



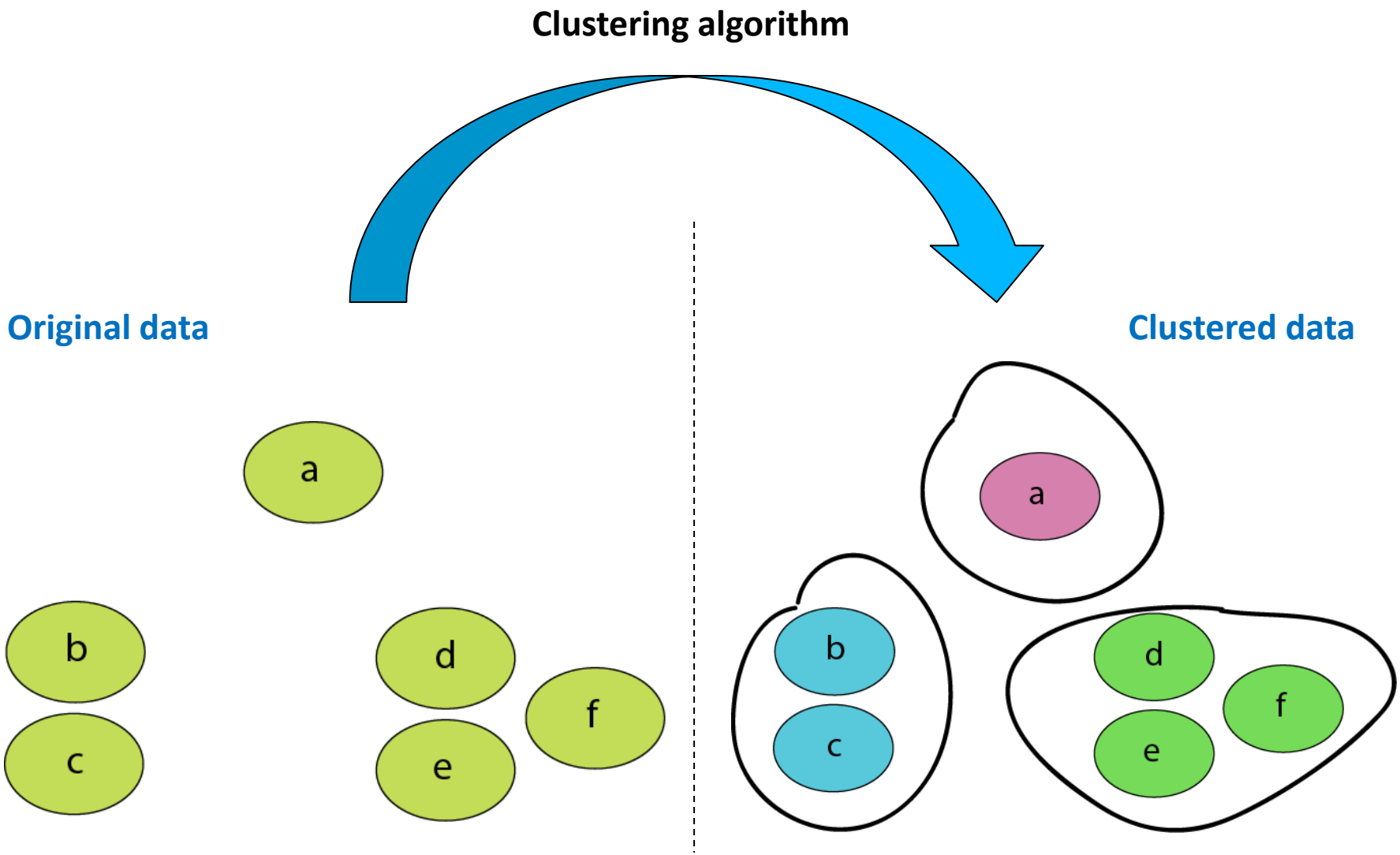
Propiedades de los clusters

Numéricos vs. Categóricos

Cómo calcular distancias entre objetos?



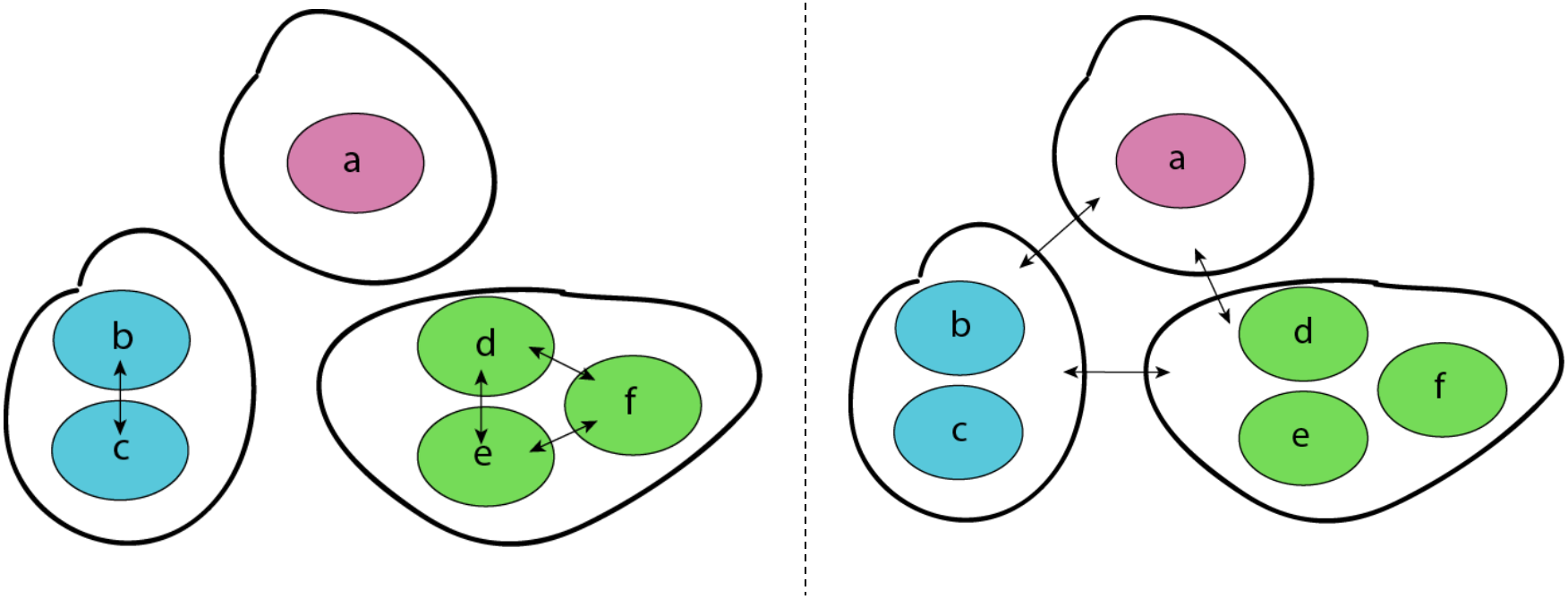
Objetivo


Clustering algorithm
Original data
Clustered data
a
b
c
d
e
f
a
b
c
d
e
f

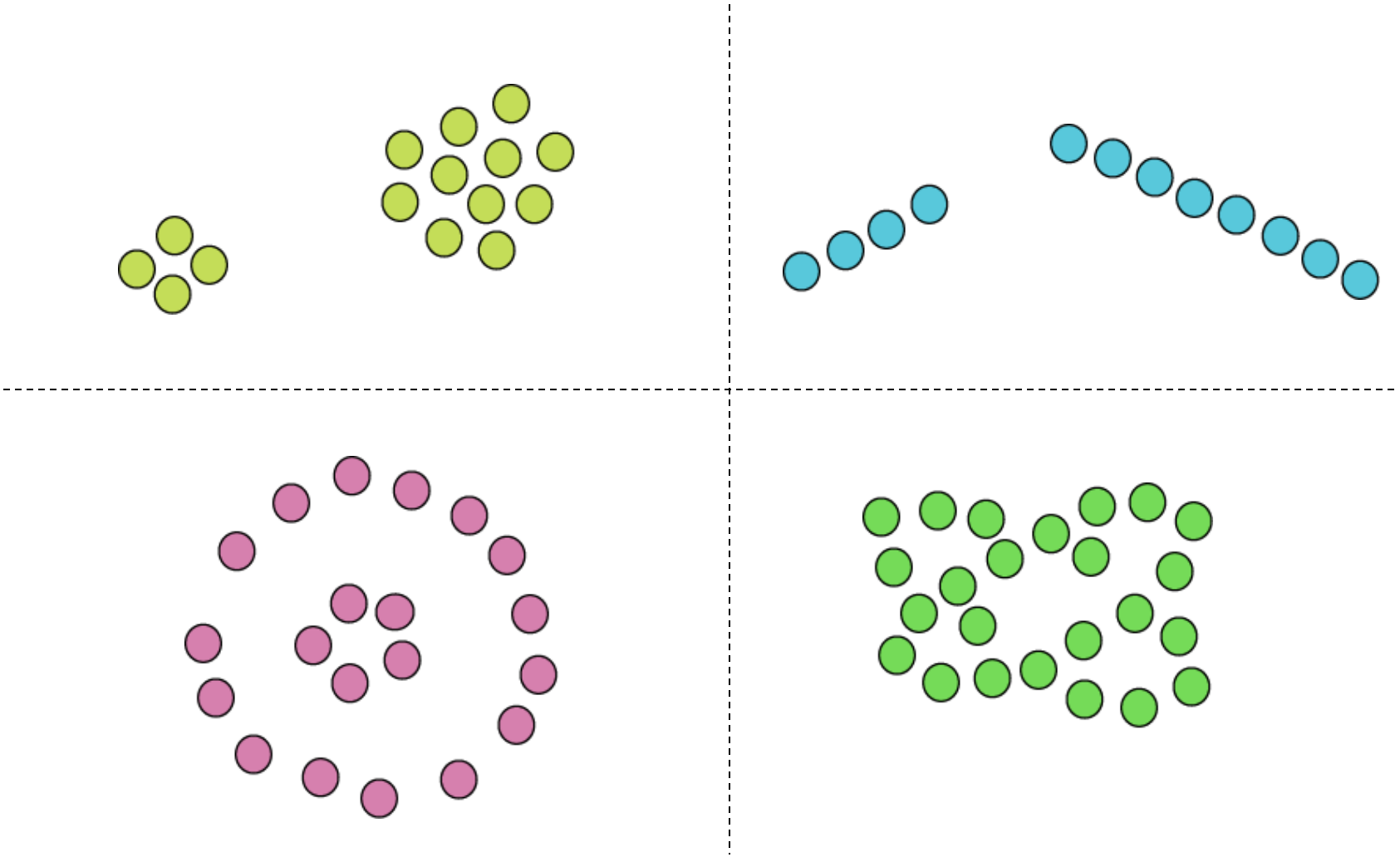
Objetivo del algoritmo

Minimizar la distancia intracluster

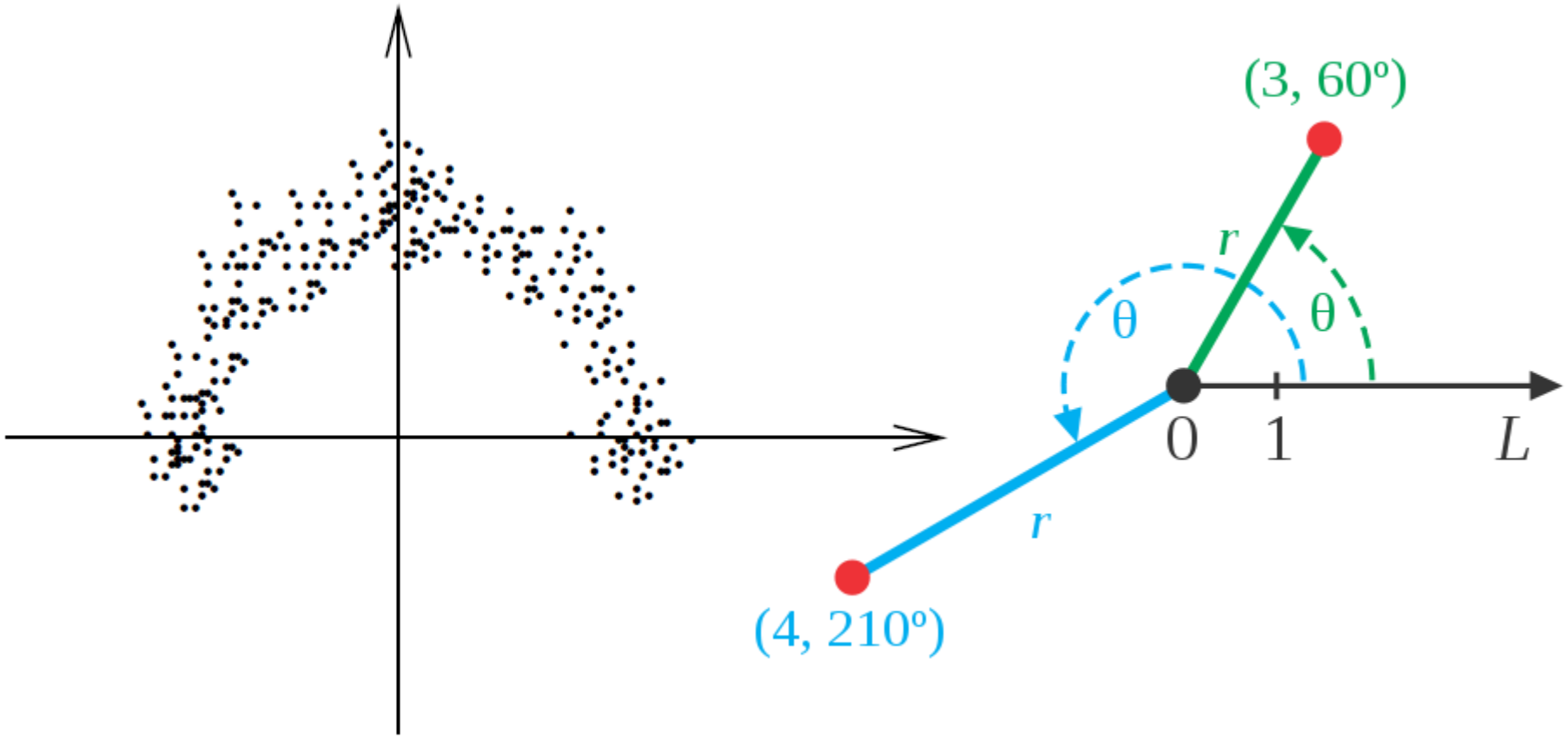
Maximizar la distancia entre clusters



Formas de los clusters

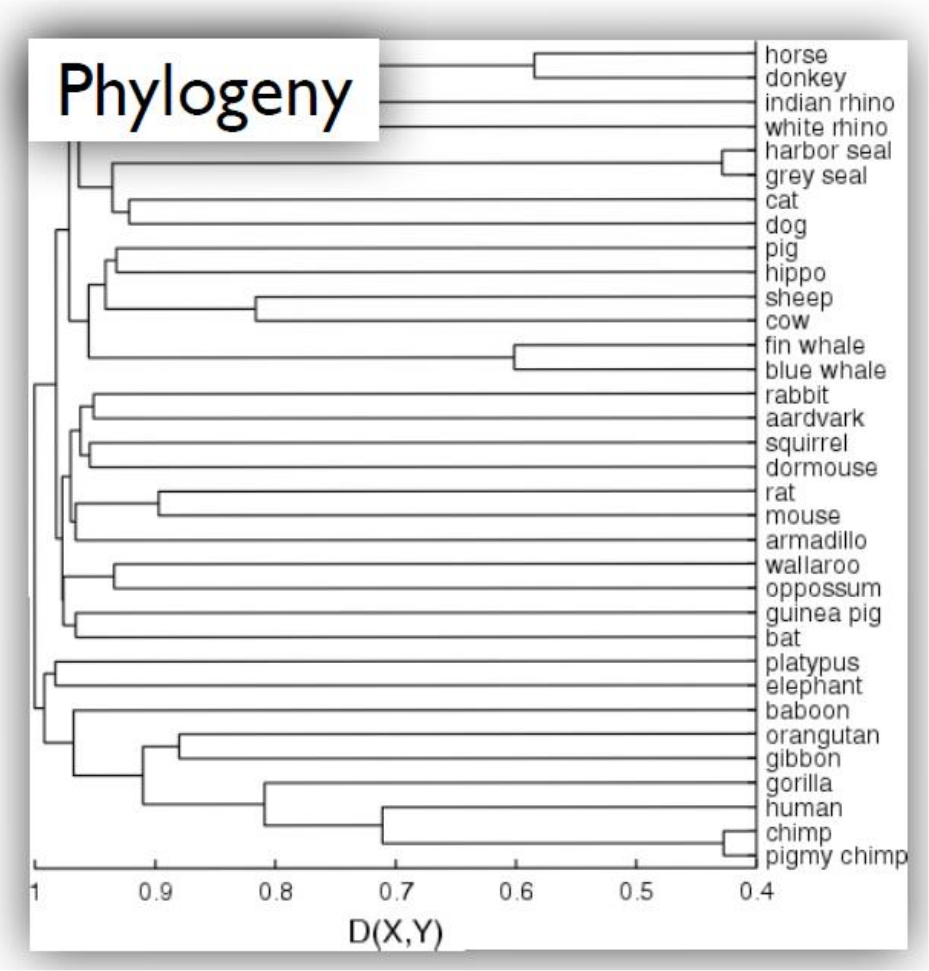


Clustering: data representation example

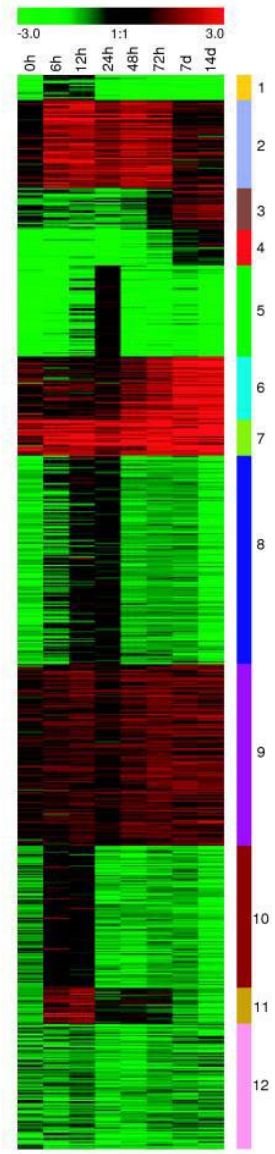


Cluster curvilineo, donde los puntos están mas o menos equidistantes del origen. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review.

Ejemplos de clustering en bioinformática



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)

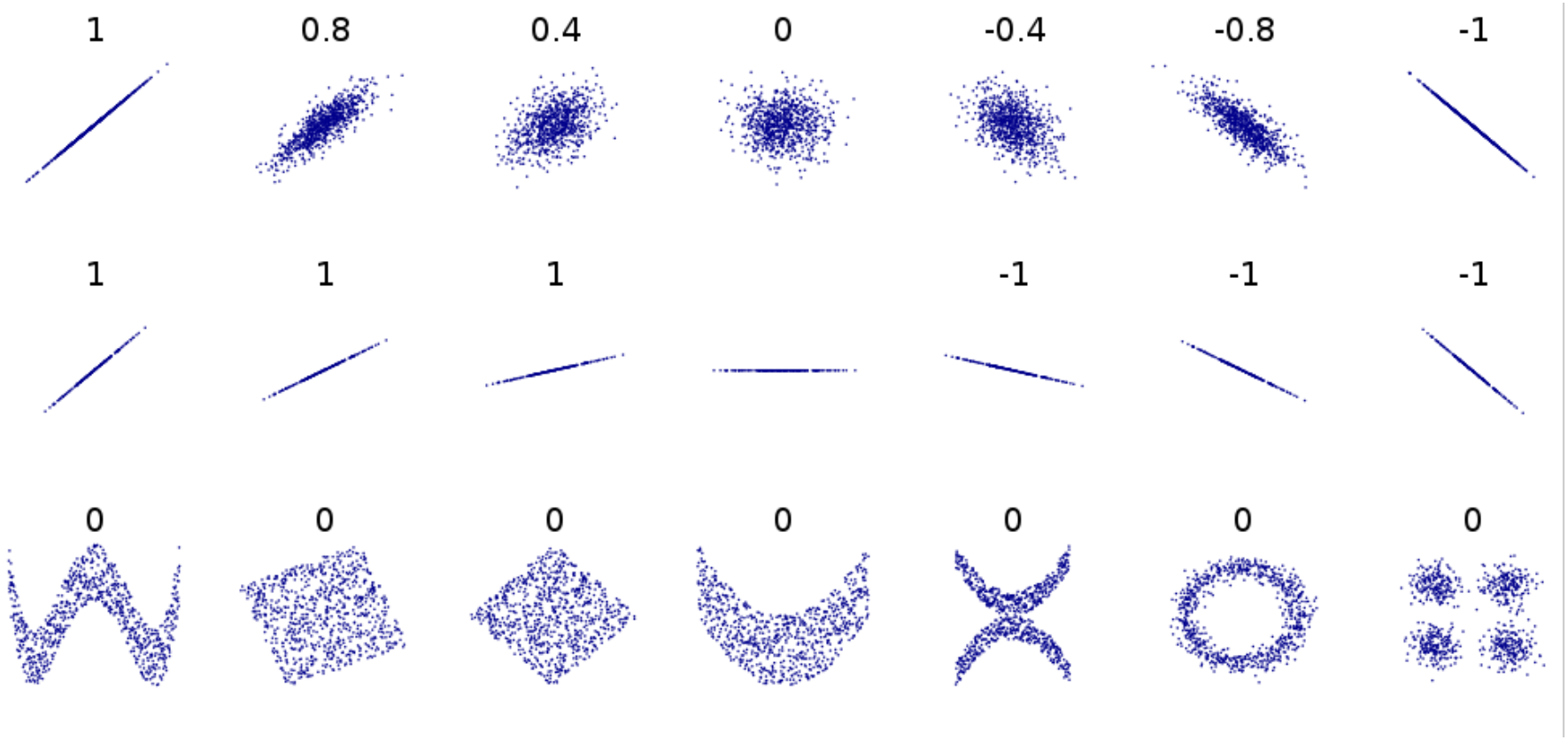
Importa el comportamiento de cada gen a lo largo de un tratamiento experimental.

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

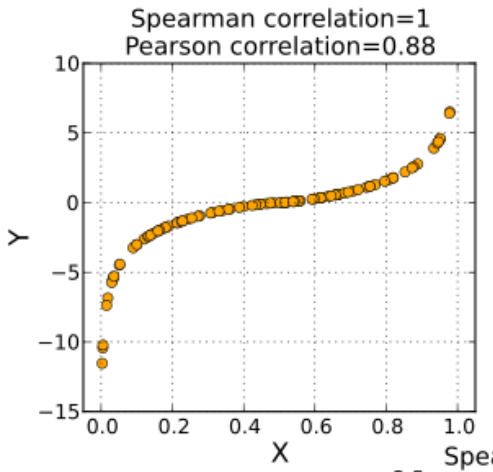


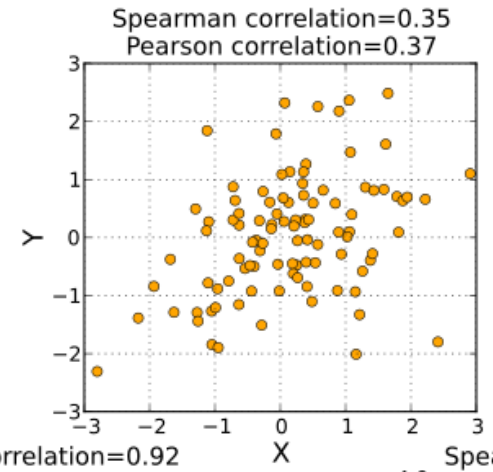
Several sets of (*x*, *y*) points, with the Pearson correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). 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

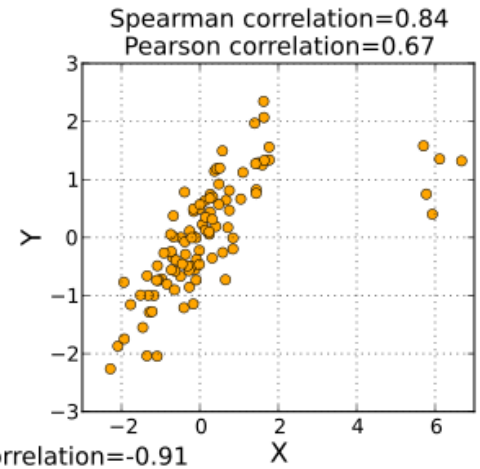
Correlation distance: Spearman's rank correlation

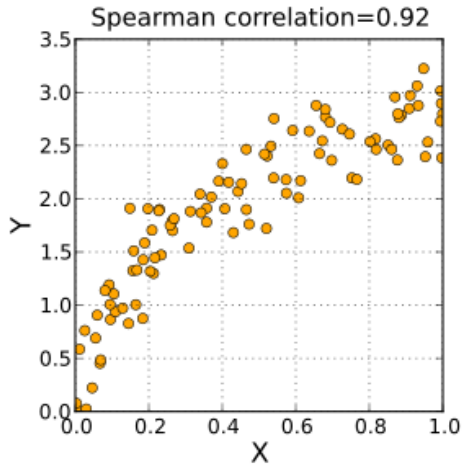
Spearman **Rank** Correlation Coefficient. Es una correlación de rankings (orden), entre dos variables. Resume la dependencia estadística de dos variables.

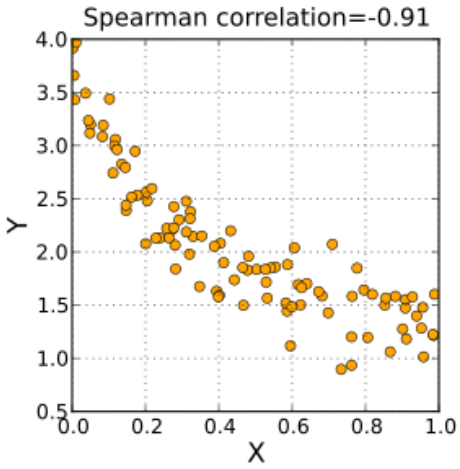
Mientras que la correlación de Pearson asume dependencia lineal entre las variables, la correlación de Spearman asume (si una variable sube, la otra sube también, y *vice versa*). **dependencia monotónica**







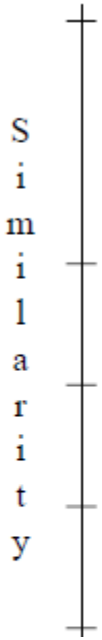


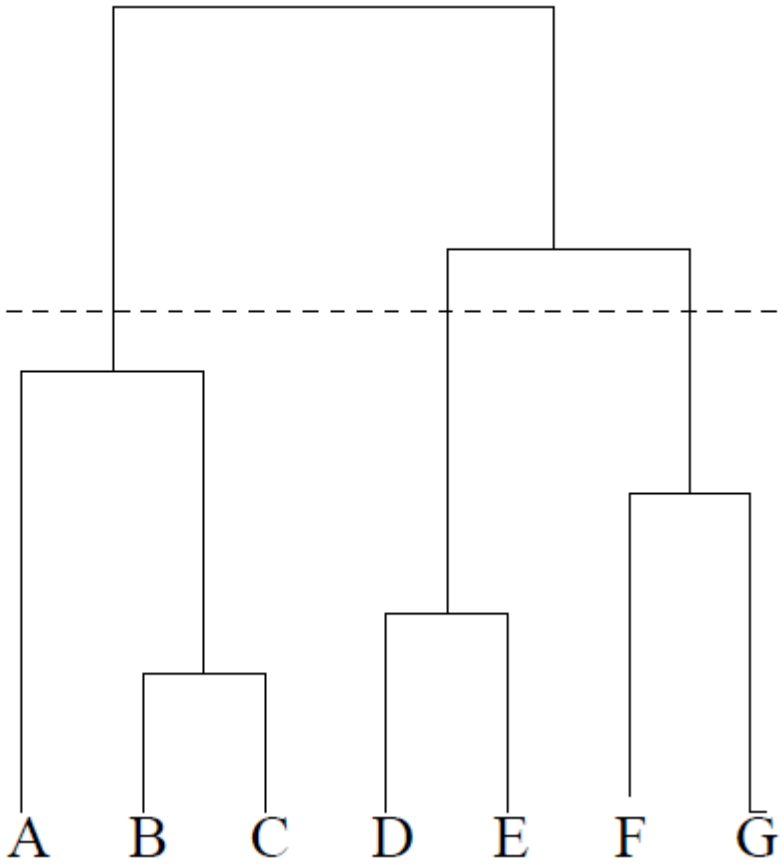


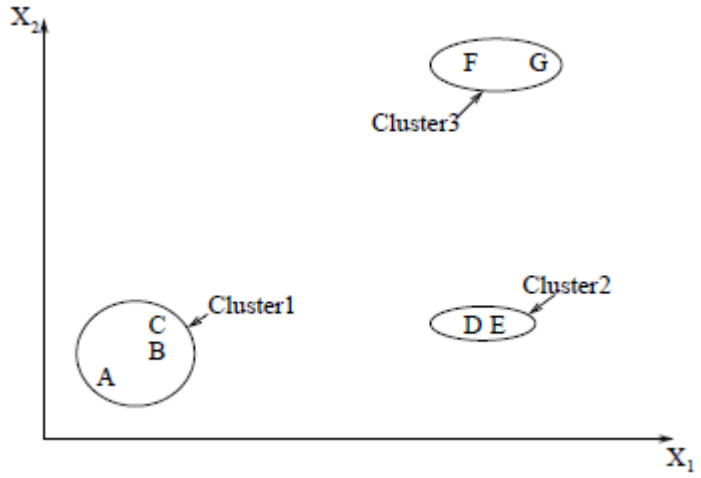
Clustering algorithms: hierarchical clustering

Todos los algoritmos jerárquicos producen como resultado un dendograma

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







Single-linkage, Complete-Linkage, Average-Linkage

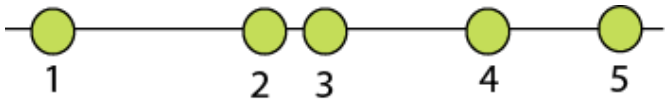
Clustering Jerárquico

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 x N:



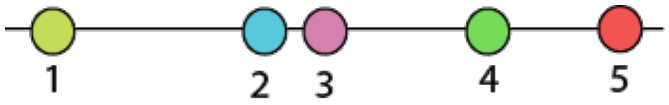
d	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

Hierarchical clustering

Comenzar por asignar cada item a un cluster.

Tenemos 5 clusters

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

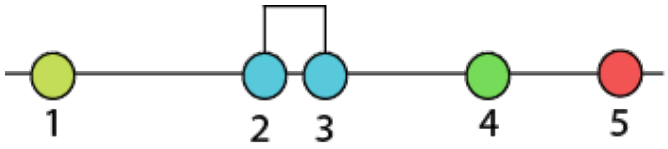


d	1	2	3	4	5
1	0	5	6	10	13
2	5	0	1	5	8
3	6	1	0	4	7
4	10	5	4	0	3
5	13	8	7	3	0

Hierarchical clustering

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

Tenemos 4 clusters



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

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

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

d	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

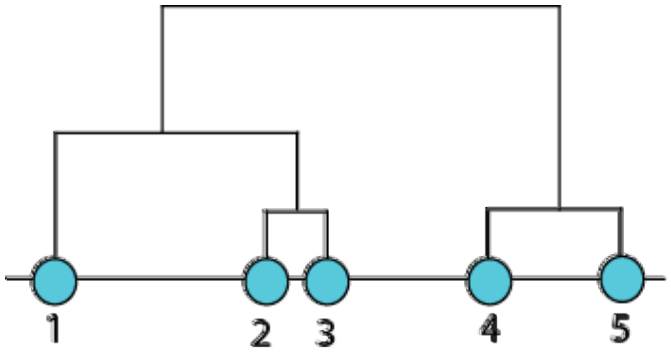
Y en **average-linkage?**

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

d	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

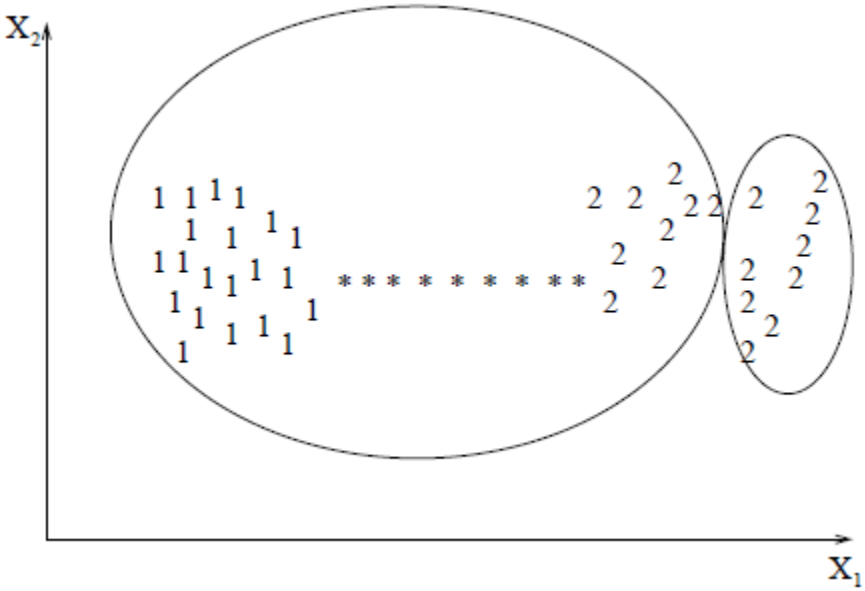
Single-linkage

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

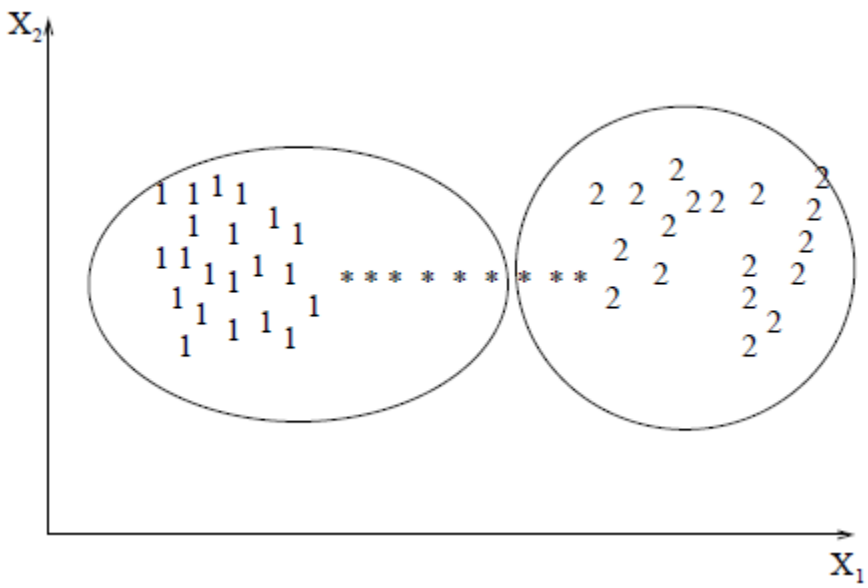


Diferencias entre single vs complete linkage

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.

Diferencias entre single vs complete linkage

Single linkage: ejemplo Mucinas *T. cruzi*

Cluster 358

TTTTTTTEAPTTTTTTTTQAPSTTTTETPTTTTTRAPSRLREIDGSLSSSAWVCAP
MKVSDSDAPSPTTTTTTTTTTTTTEAPINTAINTEAPTTTTTRAPSRLREVDGSLSSSAWVCVP
PSTDQTNTNADDSAKKNTAATINTTTTTTTTTAPEAPSTTTTEAPTTTTTRAPSRLREIDGTLSSSAWVCAP
TTTTTSTTEAPTTTTTTTEAPTTTTTRAPSRLREIDGSLSSSAWVCAP
TTGAPTTTTTRAPSRLREIDGSLSSSAWVFAR
TTTTTAAPEAPTTTTTRAPSRLREIDGSLSSPAWVC
TTTTTKAPTTTTTTTTTESPTTATTEVPTTTTTRAPSRLREIDGSLSS
VKKAEDAAATTTNTTKAPITTTTT
ADPTTTSARTPSRLREIDGSLGSSAWVC
PSTTTTRAPSLLRESDGSLS
FFGVWQTKPFEPSPLRSDGS
TTNTSAPSRLRSVDGSLSSS
EKLRQRRRSILREIDDVENHASQS
APSNTTMNTEAPTTTTSRAP
TATSTTTSTEAPTTTTTRAP
QQPSVSANPVQQIQKANAPT
RRECASTAADDSARKTYLRP
DHVNTNADDSAKKTTAATTT
…

Diferencias entre single vs complete linkage

Complete linkage: ejemplo Mucinas T. cruzi

Cluster 695

TTTTTTTEAPTTTTTTTTQAPSTTTTETPTTTTTRAPSRLREIDGSLSSSAWVCAP
TPTTTTTRAPSRLREIDGSLSSSAWVCA
TTAAPTTTTTRAPSRLREIDGSLSSSAWVCAP
TTTTEAPTTTTTRTPSRLREIDGSLSSS
TTTTEAPTTTTTRTPSRLREIDGSLSSS
APTTTITRTPSRLRESDGSLSSSAWVCA
SNTAMITEAPTTTTRTTTTEAPSRLREIDGSLSSSAWVCA

Cluster 393

TTGAPTTTTTRAPSRLREIDGSLSSSAWVFAR
TMNTEAPTTTTTTTTTTTRAPSRLREIDGSLSSSAWMCAP
PITTTTTAPEAPTTTTTTRAPTTTELPTTTTTRAPSRLREIDGSLSSS
EAPTTATTRAASRLREIDGSLSSS
APSNTTMNTEAPTTTTSRAP

Diferencias entre single vs complete linkage

Complete linkage: ejemplo Mucinas *T. cruzi*

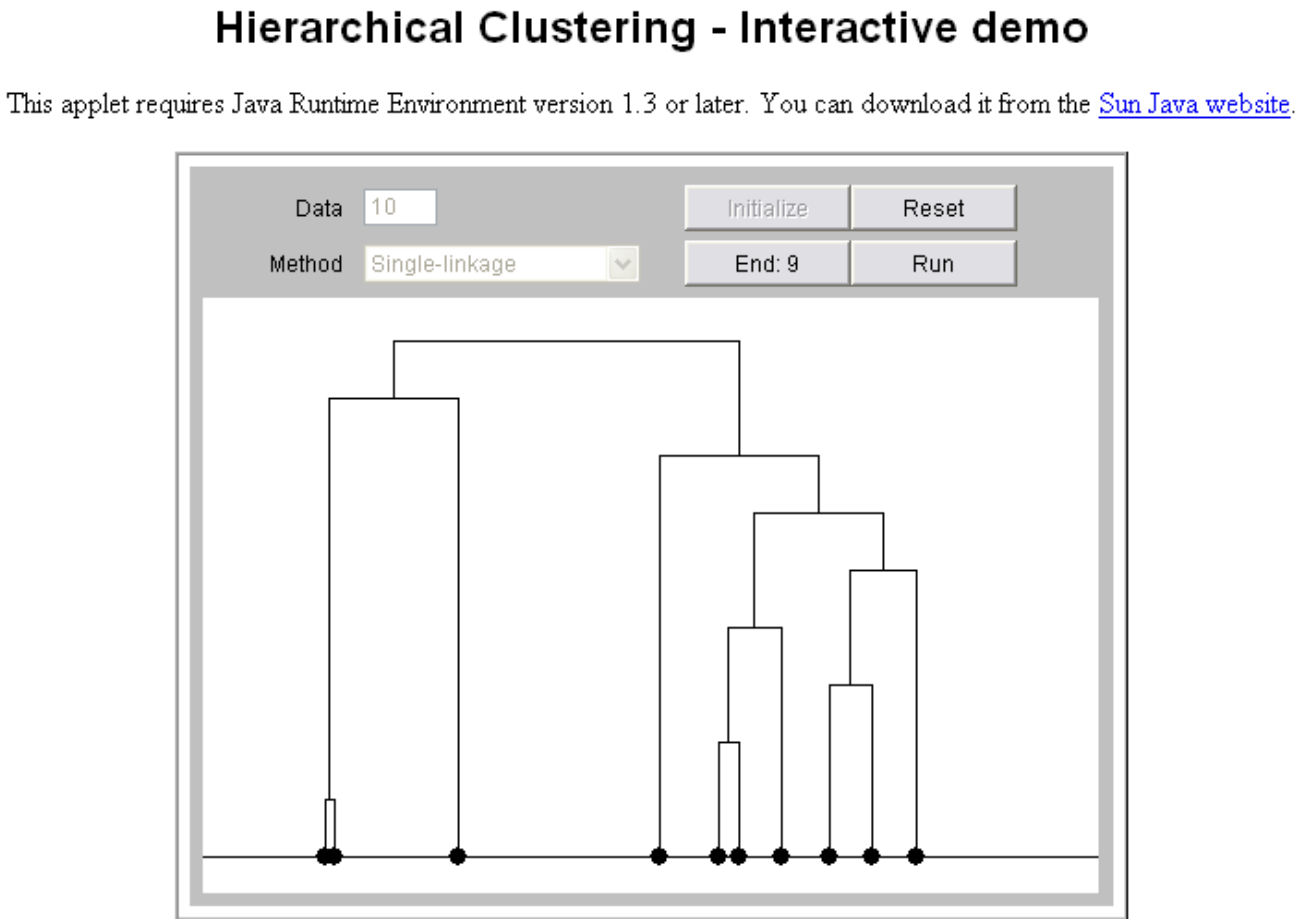
Cluster 695

```
TTTTTTTEAPTTTTTTTTQAPSTTTTETPTTTTTTRAPSRLREIDGSLSSSAWVCAP
--------------------------TPTTTTTRAPSRLREIDGSLSSSAWVCA
----------------------TTAAPTTTTTRAPSRLREIDGSLSSSAWVCAP
--------------------TTTTEAPTTTTTRTPSRLREIDGSLSSS
--------------------TTTTEAPTTTTTRTPSRLREIDGSLSSS
--------------------------APTTTITRTPSRLRESDGSLSSSAWVCA
--------------SNTAMITEAPTTTTRTTTTEAPSRLREIDGSLSSSAWVCA
```

Cluster 393

```
--------------------TTGAPTTTTTRAPSRLREIDGSLSSSAWVFAR
-------------TMNTEAPTTTTTTTTTTTRAPSRLREIDGSLSSSAWMCAP
PITTTTTAPEAPTTTTTTRAPTTTELPTTTTTRAPSRLREIDGSLSSS
----------------------EAPTTATTRAASRLREIDGSLSSS
--------------APSNTTMNTEAPTTTTSRAP
```

Hierarchical clustering: interactive demo



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)

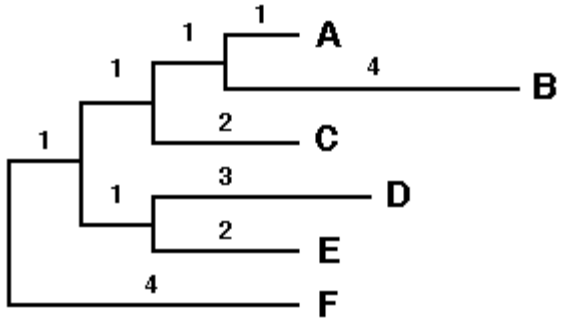
Distancias ultramétricas

Neighbor-joining

Usado para reconstruir filogenias

Usa la media aritmética

Las distancias son aditivas



Clustering algorithms: K-means

Es muy rapido!

Particional

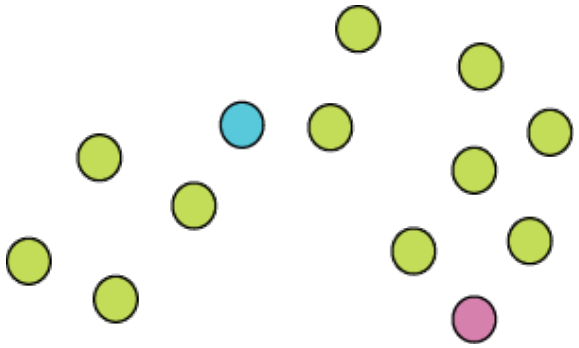
Usa Distancia euclídea

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

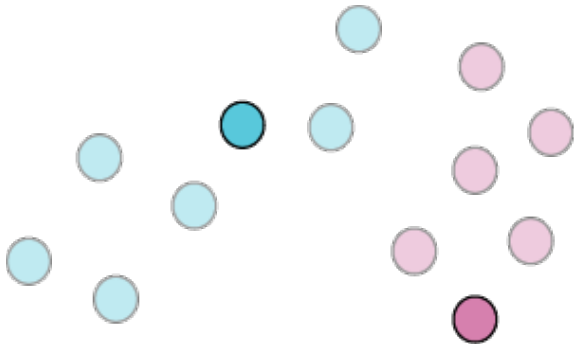
Util para búsqueda de prototipos

Sensible a outliers

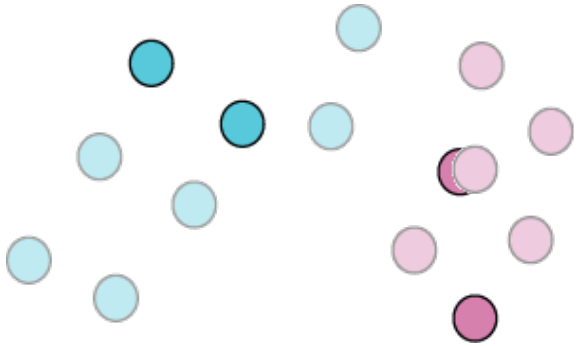
K-means



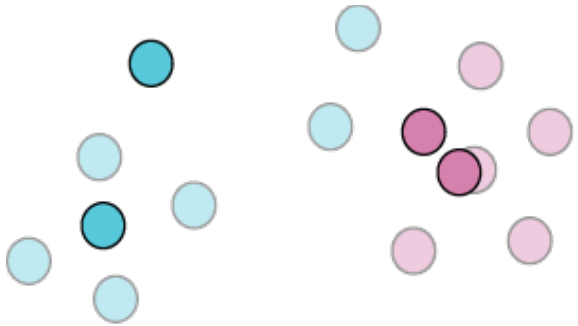
K-means



K-means



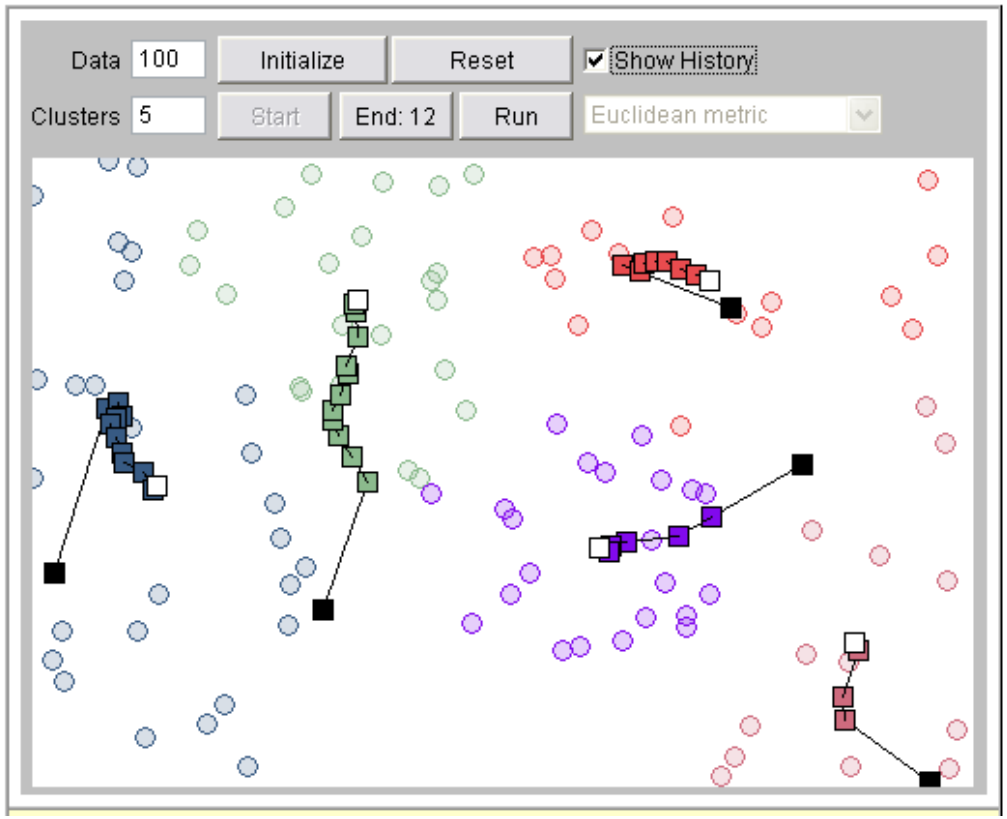
K-means

- 

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://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

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

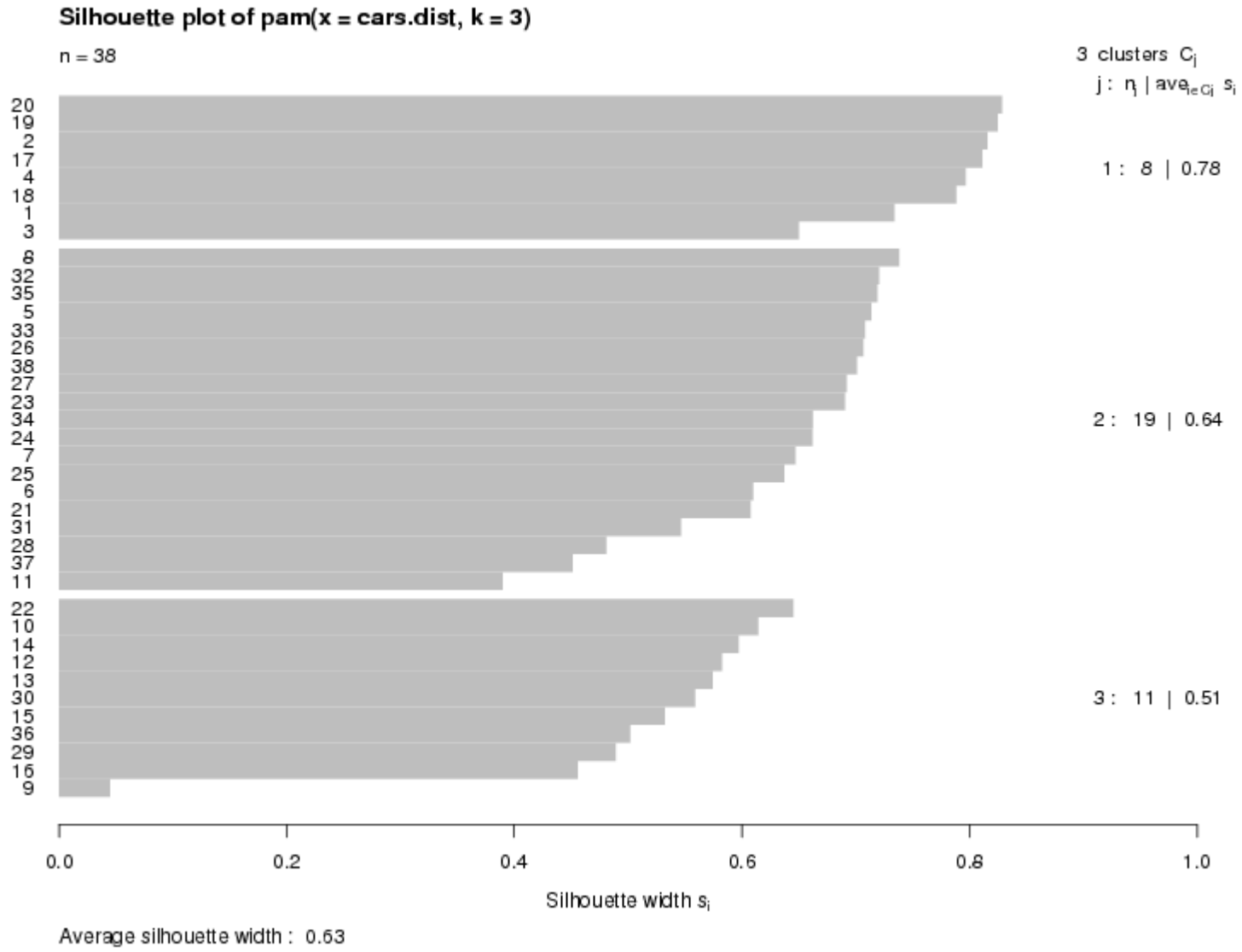
$a(\vec{x}_i)$ average distance from $\vec{x}_i$ to other instances in same cluster

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

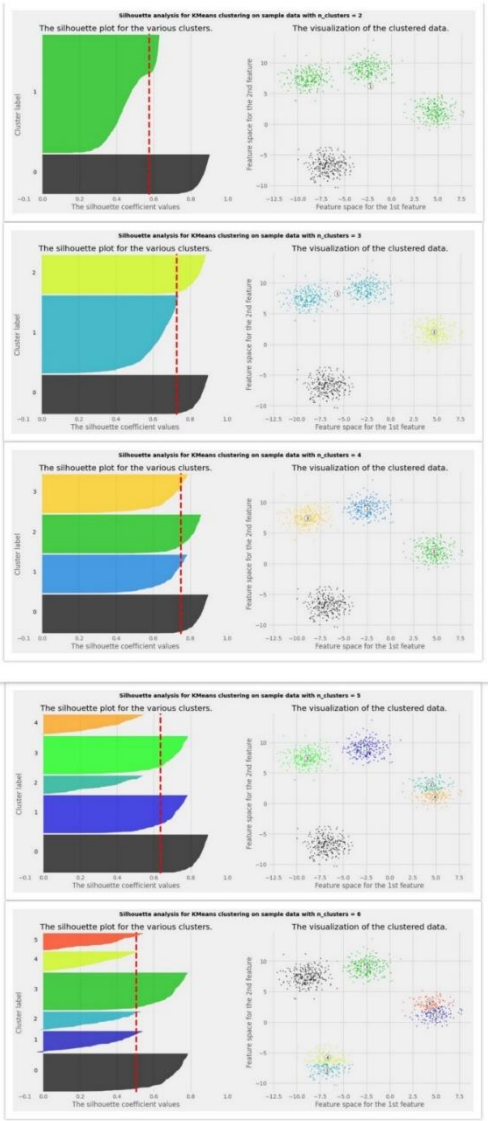
SI = 1 means element is well placed in its cluster

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

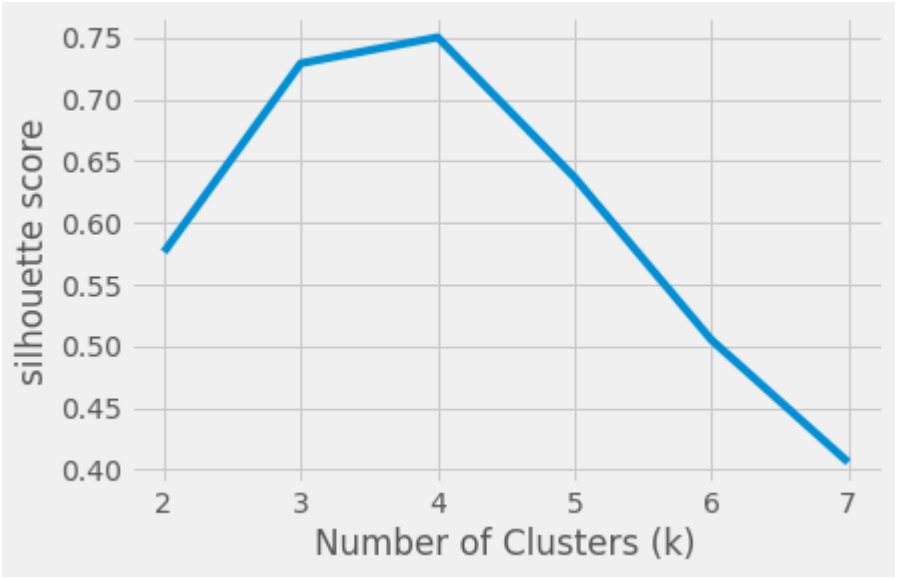
Validación interna: Silhouette Index



Combinando K-means + Silhouette



Modo exploratorio: Acá la idea no es combinar para validar clustering final, sino para ***encontrar el mejor k***.



Fuente: https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

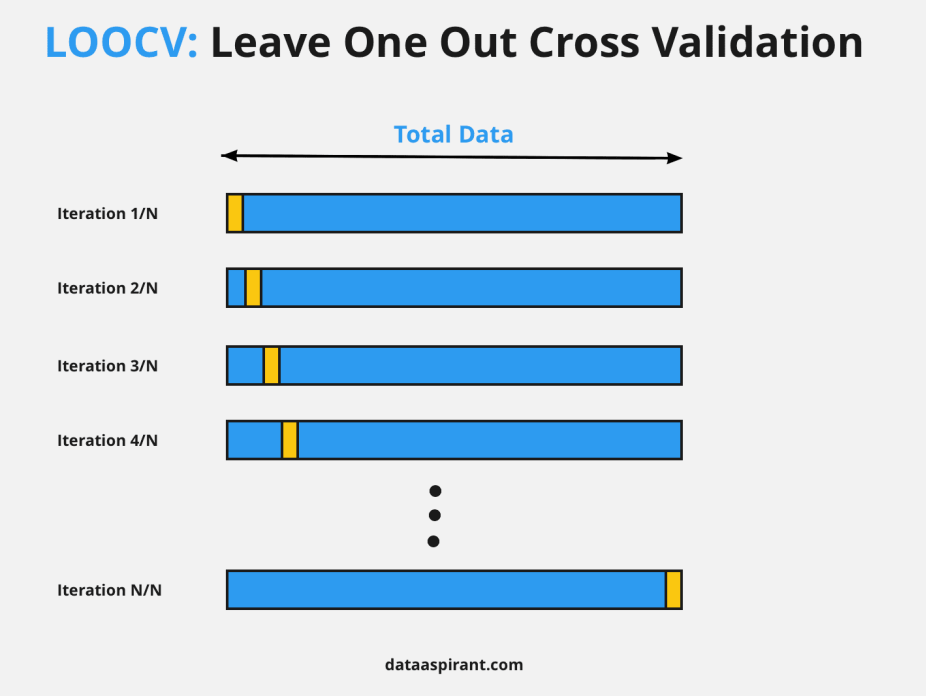
Combinando Silhouette Index con Cross-Validation

Qué es cross-validation?

Cross-validation es el proceso de crear varias distribuciones de los datos, a partir de un solo dataset. Se usa en machine learning para generar pares de distribuciones para aprendizaje y testeo. En cross-validation los datos se particionan en *k* subsets, $S_1 \ldots S_k$, cada uno se lo llama un *fold*. Los folds son usualmente del mismo tamaño approximadamente.

Qué es Leave-One-Out cross-validation?

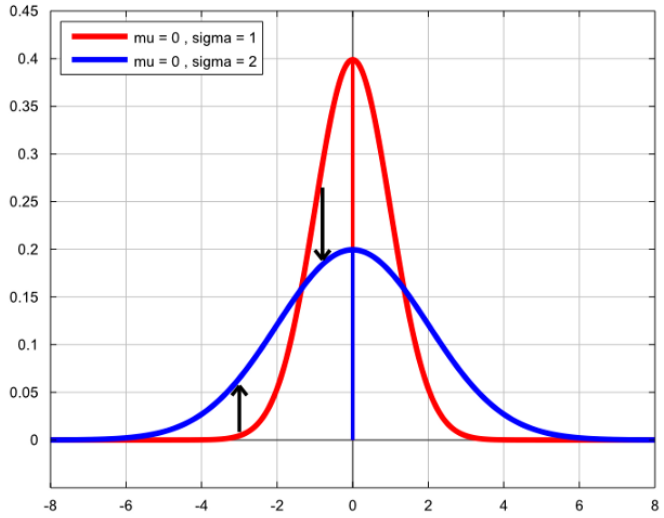
Es un caso especial de cross-validation. En este caso el numero de folds (subsets) es igual al numero de casos (tamaño total del dataset). Es muy cercano al metodo estadístido jack-knife.

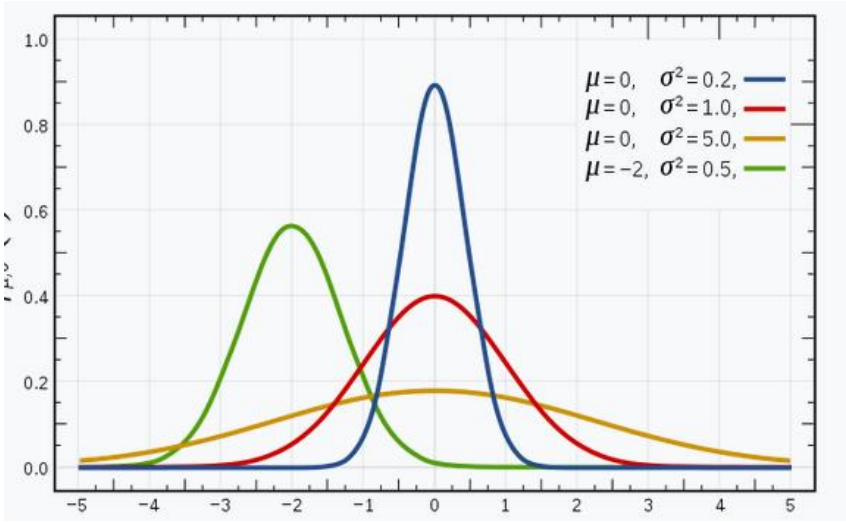


Que son los Gaussian Mixture Models (GMMs)?

Estos modelos asumen que hay un ***determinado número*** de distribuciones Gausianas, y que cada una de ellas representa un ***cluster***.

Los modelos de *mixturas Gausianas* pueden usarse para etiquetar y agrupar datos de la misma manera que k-means. Pero, mientras que k-means no usa la dispersión de los datos, los GMMs si.

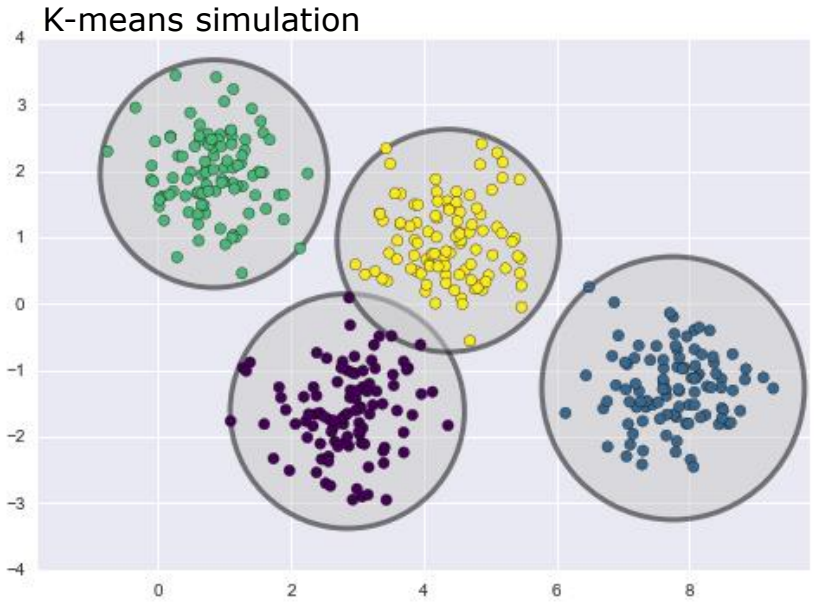




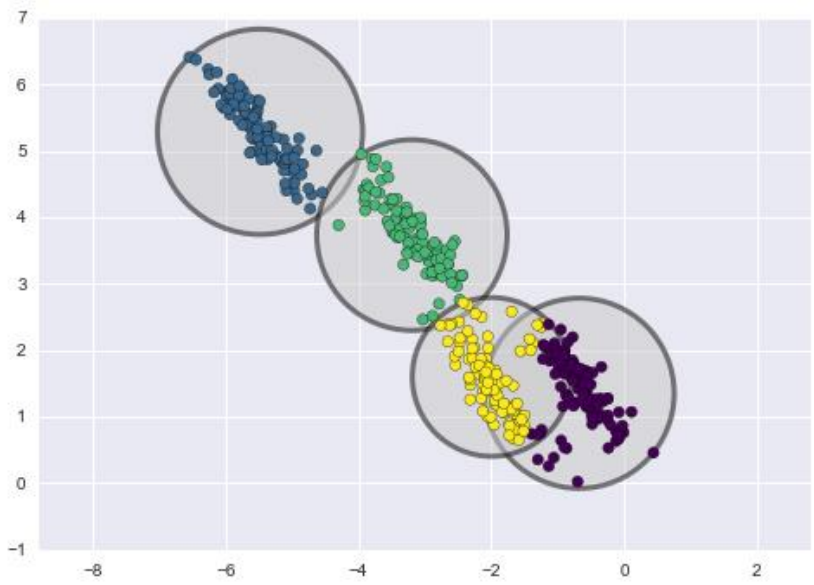
https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

Clustering usando Gaussian Mixture Models

Una manera de pensar K-means es que forma clusters con un punto central imaginario (centroide) y un radio definido por el elemento más distante de cada cluster. En 2D eso genera ***círculos.***



Datos "circulares"



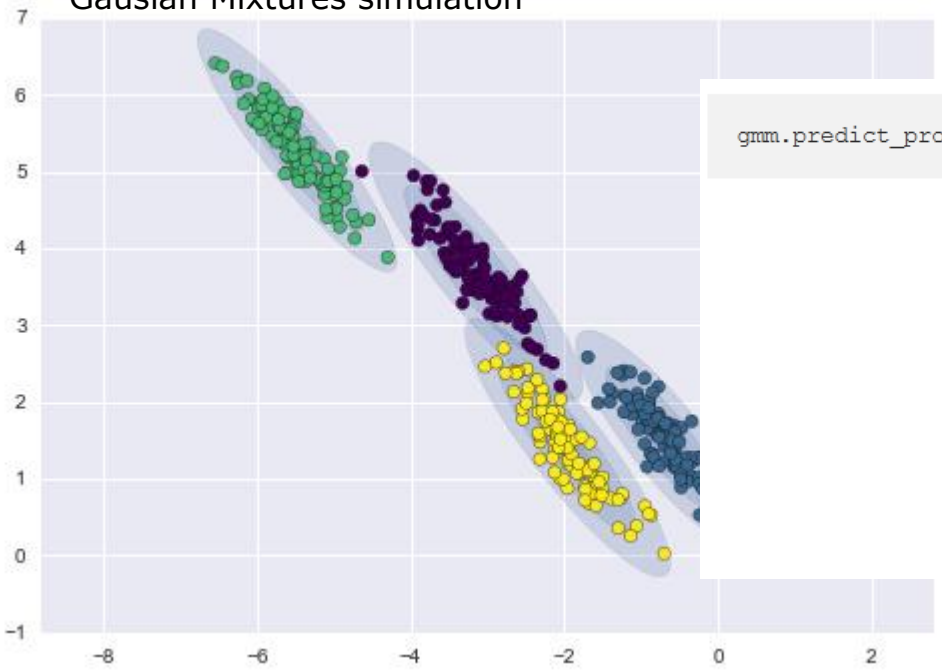
Datos "oblongos"

Clustering usando Gaussian Mixture Models

K-means asigna puntos a cada cluster en forma estricta (***hard classification***)

Mientras que las ***Mixturas Gaussianas*** son ***modelos probabilísticos***. Así que además de asignar puntos a un cluster, podemos calcular la probabilidad de que ese punto pertenezca a ese cluster (o a otro!).

Gausian Mixtures simulation

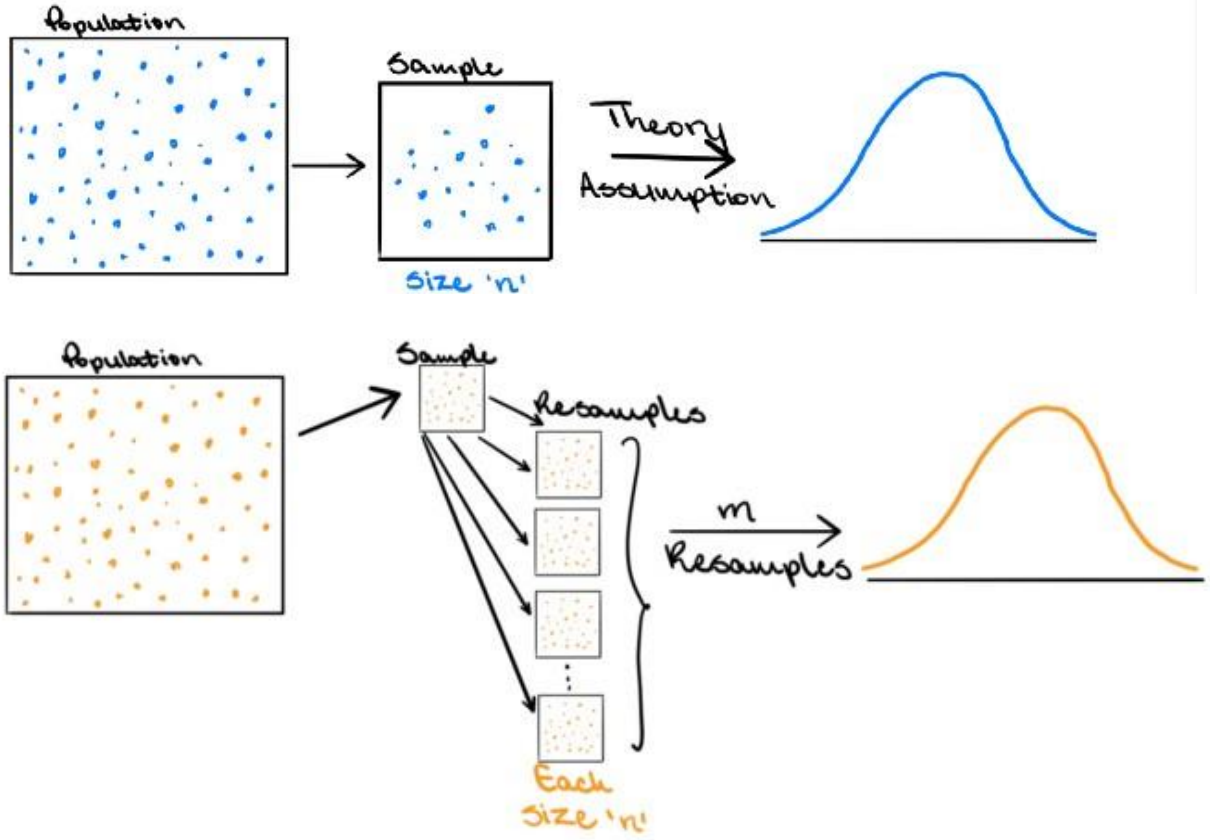



```
gmm.predict_proba(X)
```

```
array([[0.026, 0.000, 0.972, 0.002],
       [0.000, 1.000, 0.000, 0.000],
       [1.000, 0.000, 0.000, 0.000],
       ...,
       [1.000, 0.000, 0.000, 0.000],
       [0.000, 1.000, 0.000, 0.000],
       [0.000, 0.000, 0.000, 1.000]])
```

Bootstrapping

- Simulaciones con Bootstrapping
  - **Bootstrap resampling technique (Efron 1982)**
  - **Remuestreo de un conjunto de datos para generar nuevos conjuntos de muestras de esos datos.**

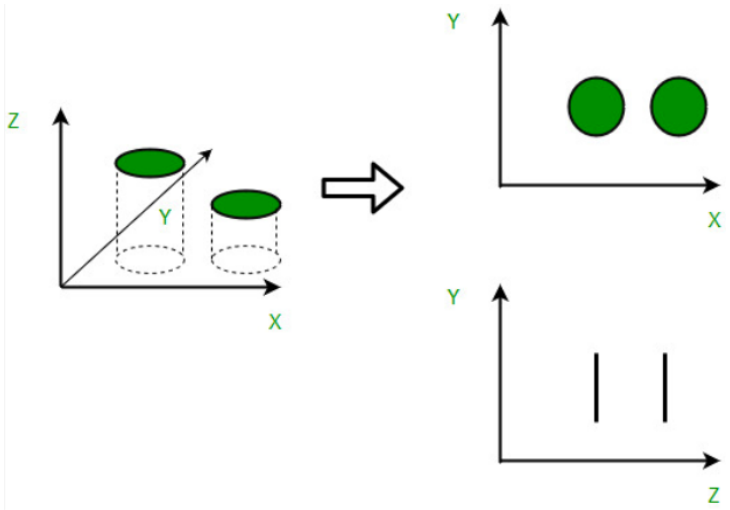


Técnicas de Reducción de Dimensionalidad

La ***dimensionalidad de los datos*** simplemente se refiere al **número de variables** (features) en un dataset.

Pero no todas las variables (dimensiones) son igualmente útiles!

- **Sin datos**: algunas variables no tienen datos para todos los objetos
- **Baja varianza:** algunas variables no muestran cambios
  - Ojo, la varianza depende del rango! Normalizar antes!
- **Alta correlación:** algunas variables pueden estar altamente correlacionadas
  - tiene sentido dejar solo una de ellas!

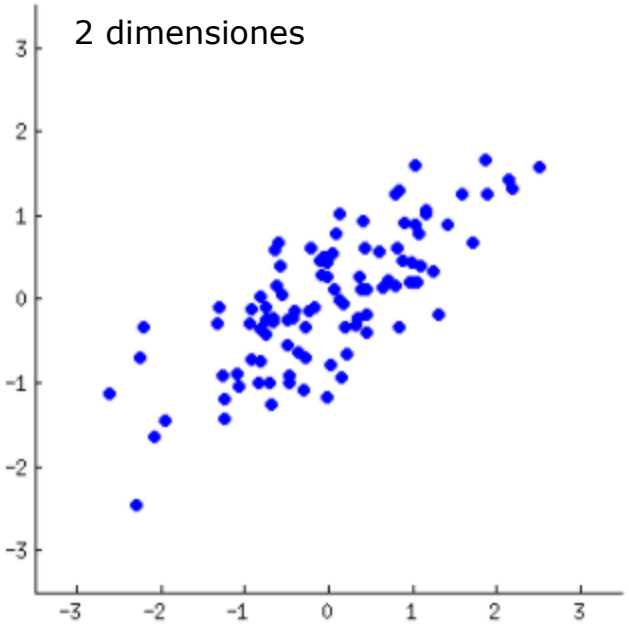


PCA = Principal Component Analysis

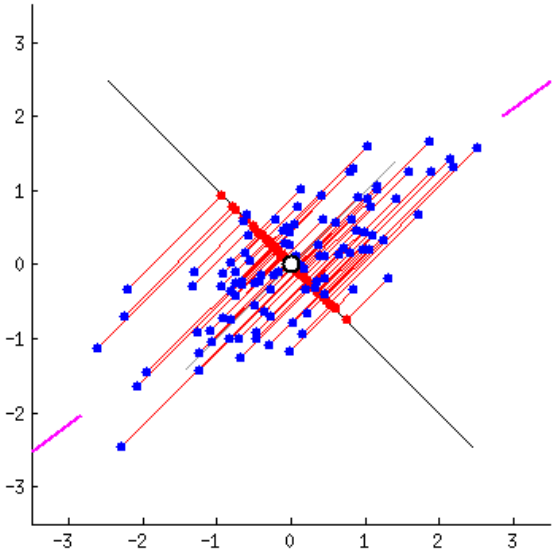
Procedimiento para transformar en forma ortogonal los datos

A partir de n coordenadas, se obtienen un nuevo conjunto de n coordenadas, pero transformadas en su componente principal.

2 dimensiones



1 dimension



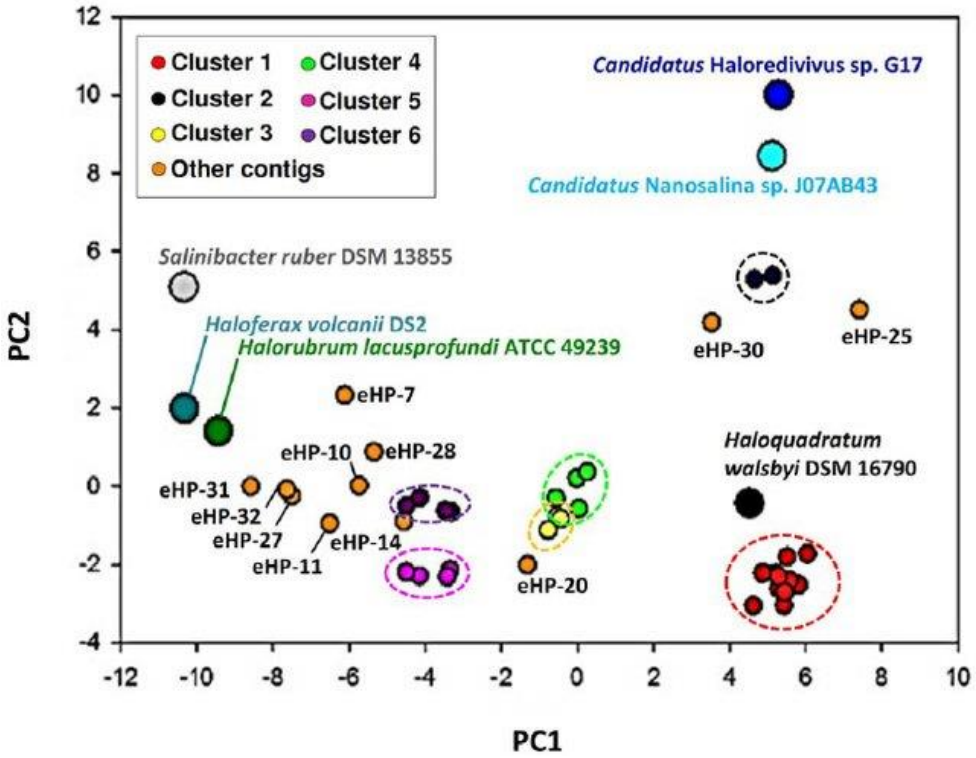
https://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html

En bioinformática, ejemplo de uso

Uso de codones en genes codificantes de proteínas:

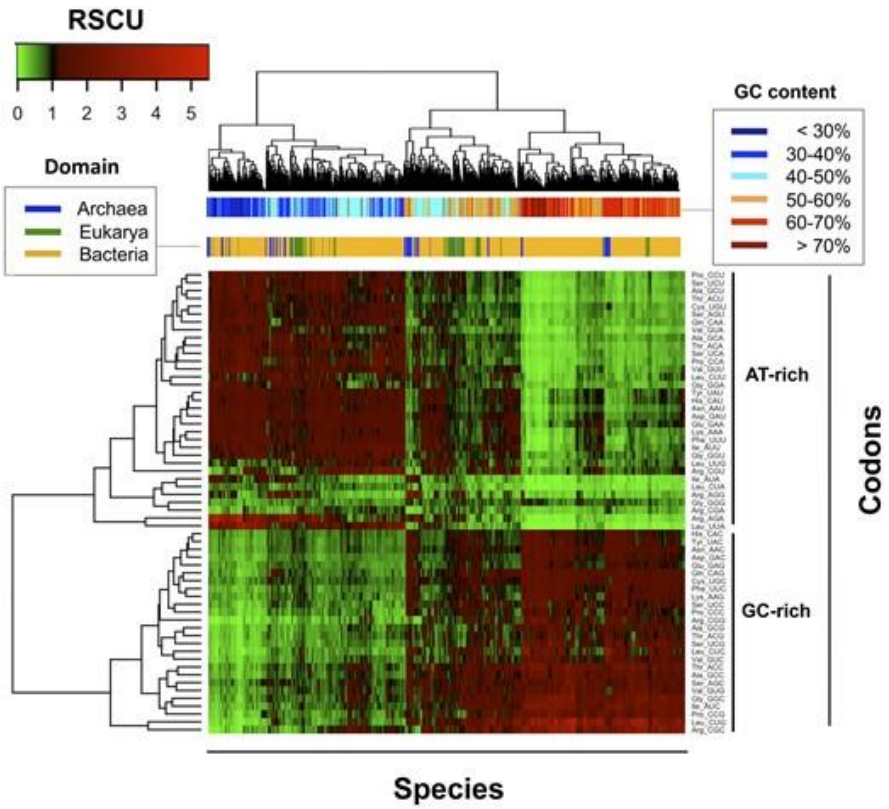
- Hay 64 codones (variables / dimensiones)
  - No todos son igualmente informativos

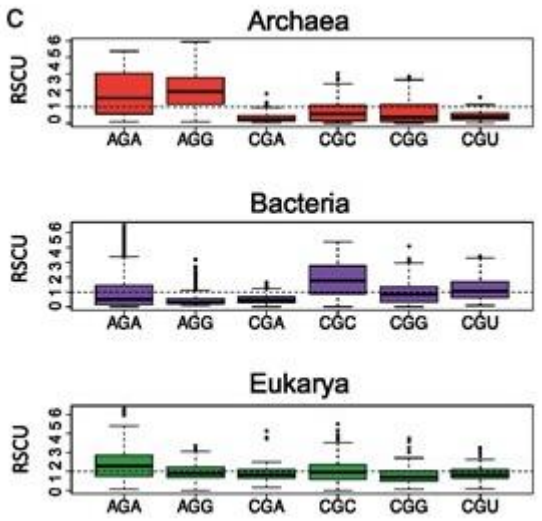
Reducción a 2 dimensiones



Garcia-Heredia I, Martin-Cuadrado A-B, Mojica FJM, Santos F, Mira A, Antón J, et al. (2012) Reconstructing Viral Genomes from the Environment Using Fosmid Clones: The Case of Haloviruses. PLoS ONE 7(3): e33802. https://doi.org/10.1371/journal.pone.0033802

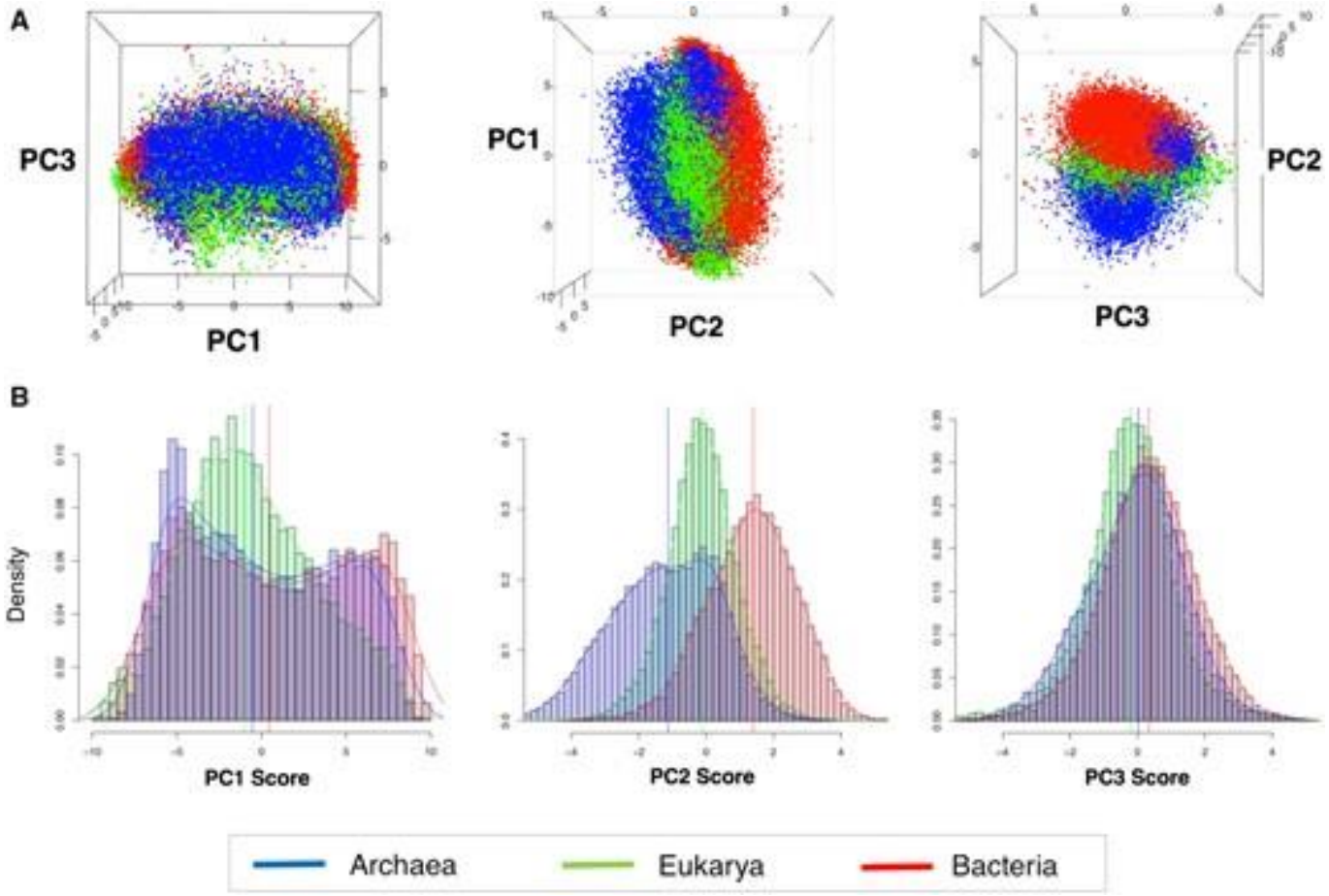
Uso de codones: archaea vs eukarya vs bacteria





Eva Maria Novoa, Irwin Jungreis, Olivier Jaillon, Manolis Kellis, Elucidation of Codon Usage Signatures across the Domains of Life, *Molecular Biology and Evolution*, Volume 36, Issue 10, October 2019, Pages 2328–2339, https://doi.org/10.1093/molbev/msz124

Análisis de las 3 componentes principales (en 2D)



Eva Maria Novoa, Irwin Jungreis, Olivier Jaillon, Manolis Kellis, Elucidation of Codon Usage Signatures across the Domains of Life, *Molecular Biology and Evolution*, Volume 36, Issue 10, October 2019, Pages 2328–2339, https://doi.org/10.1093/molbev/msz124

Final

- **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.**