clusterAI 2020
ciencia de datos en ingeniería industrial
UTN BA
curso I5521

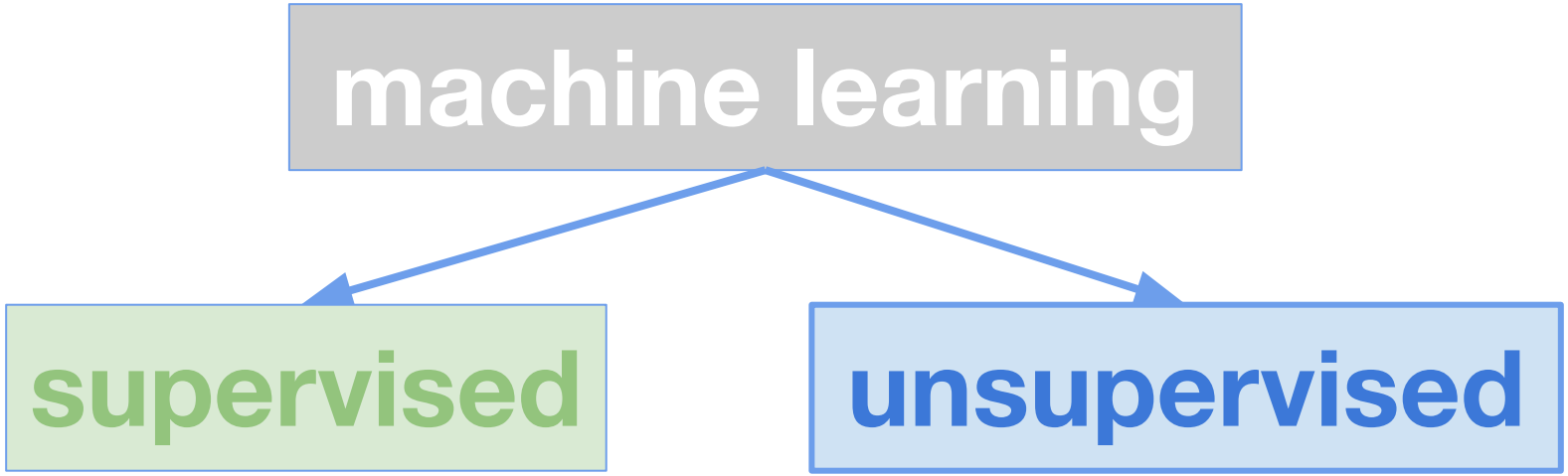
clase_05: Aprendizaje no supervisado

Docente: Martin Palazzo

agenda clase07

- aprendizaje no supervisado
- Dimension reduction
- cluster Analysis
  - K-means clustering
  - hierarchical clustering
- Python code lab

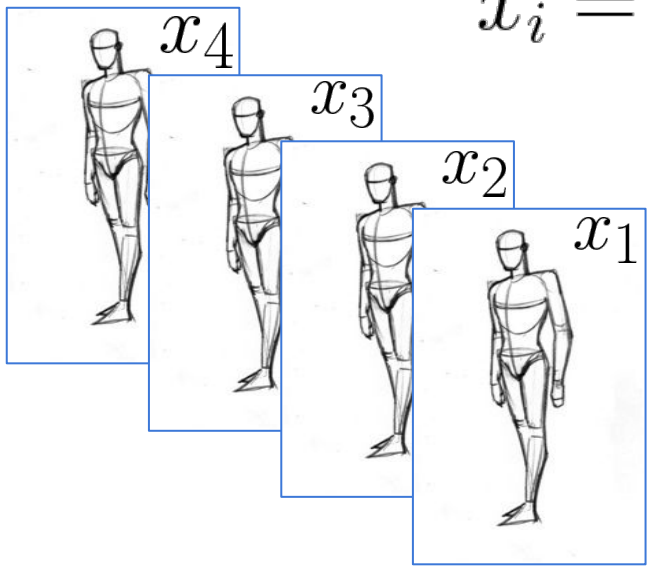
Learning approaches



Particularmente en este curso vamos a poner foco en el aprendizaje supervisado y el aprendizaje no-supervisado. Estos dos enfoques suelen ser los más populares y prácticos para la mayoría de los problemas.

samples and features

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{id},) \qquad \mathcal{X} \in \mathbb{R}^d$$



X1**1** (feat. 1 in sample 1)

X1**2** (feat. 2 in sample 1)

X1**3** (feat. 3 in sample 1)

X1**4** (feat. 4 in sample 1)

En este caso solo dispondremos de una matrix **X** caracterizada por “n” muestras y “d” features. Las filas de la matriz **X** representan vectores **x**, cada uno asociado a una muestra.

aprendizaje no supervisado

$$S = \{x_1, x_2, ..., x_n\}$$

Consiste en un aprendizaje sin un “supervisor” (aprender sin variable dependiente “Y” o etiqueta). El objetivo es inferir propiedades y **estructuras** de la distribución de los datos X, en muchos casos consiste en estimar la densidad de P(x).

*Ghahramani, Z. (2003, February). Unsupervised learning. In *Summer School on Machine Learning* .Springer, Berlin, Heidelberg.

aprendizaje no supervisado

$$S = \{x_1, x_2, \ldots, x_n\}$$

El aprendizaje supervisado intenta responder las siguientes preguntas:

- Si tuviese que agrupar las **n** muestras en **K** grupos: ¿a que grupo se asignará cada muestra? ¿qué criterio usaríamos para agruparlas? -> **clustering**.

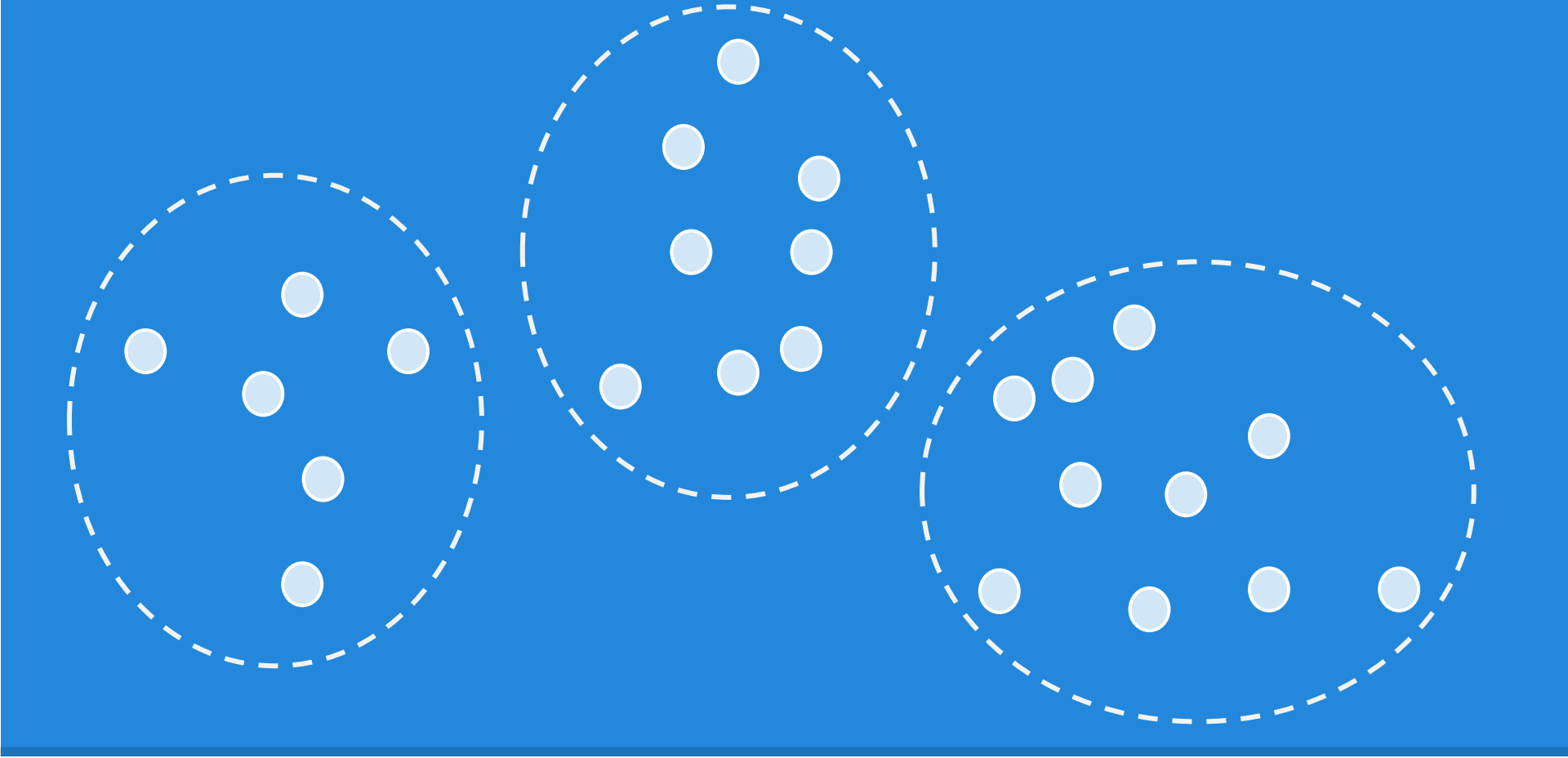
- ¿Cuál será el sub-espacio **z** que represente la estructura latente de los datos? ¿cómo llego a ese sub-espacio? -> **reducción de la dimensionalidad no-supervisada.**

aprendizaje no supervisado

Métodos y estrategias de aprendizaje no supervisado:

- Reducción de la dimensionalidad (PCA, t-SNE)[1][2]
- Cluster Analysis (k-means, spectral, dbscan)[3][4]
- Generative methods (GANs, VAEs)[5]

[1] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*,
[2] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*]
[3] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*,
[4] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*
[5] Mescheder, L., Nowozin, S., & Geiger, A. (2017). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks



Aprendizaje no supervisado: clustering - agrupamiento

Cluster analysis

Conocido también como “data segmentation” o “**community detection**”. Consiste en agrupar o segmentar un conjunto de muestras (samples) en subconjuntos, grupos o “**clusters**”.

Los clusters se construyen de manera tal que las muestras pertenecientes a un mismo cluster deben ser más similares entre si que muestras entre distintos clusters y simultáneamente hay baja similaridad entre muestras de distintos clusters.

El resultado de un algoritmo de clustering es un conjunto de etiquetas **C** asociadas a cada muestra.

$$C = \{c_1, c_2, \ldots, c_k\}$$

Cada etiqueta refiere a que cluster pertenece cada muestra.

Aprendizaje no supervisado: clustering

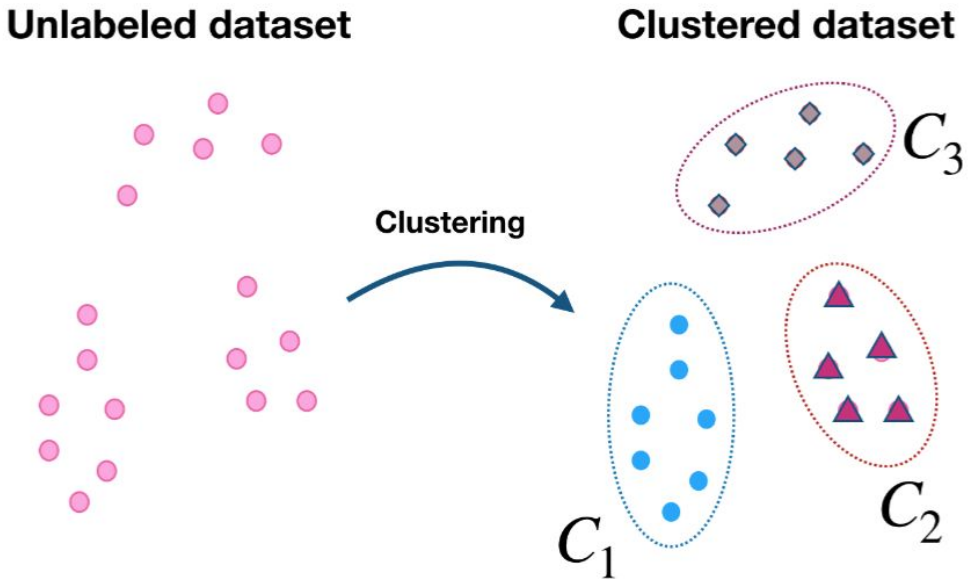
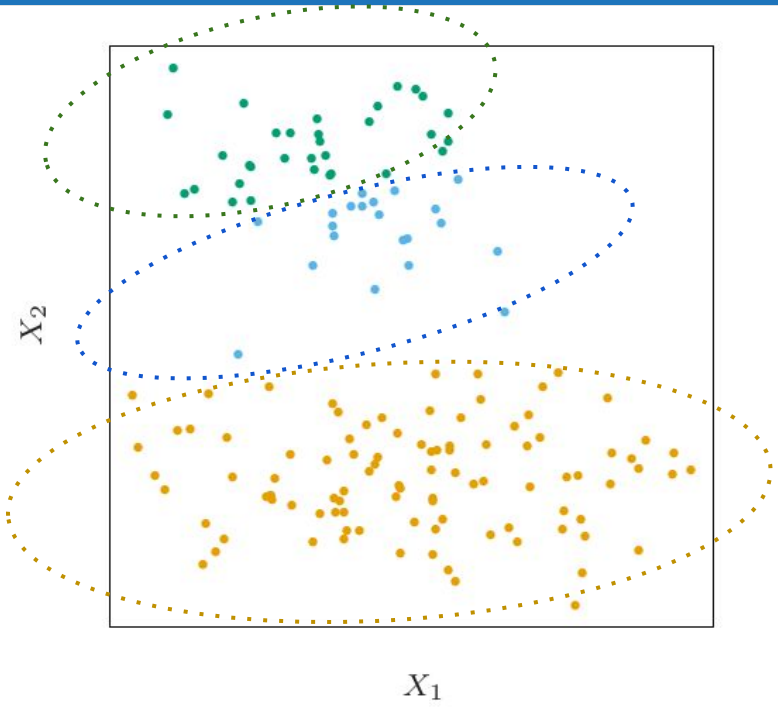


*Imagen: Interpretación de Señales Genómicas Humanas utilizando técnicas de Machine Learning para mejorar el Diagnóstico Médico , Martin Palazzo

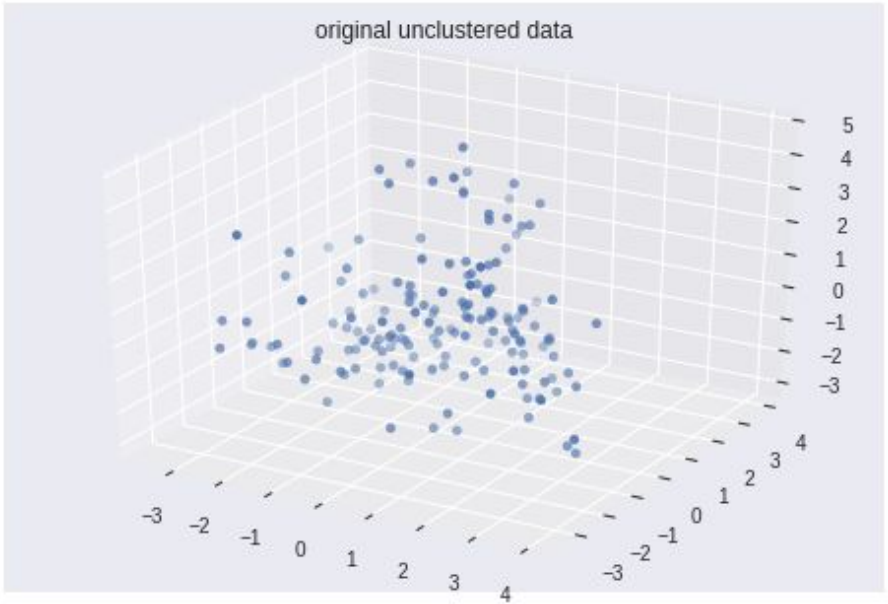
cluster analysis

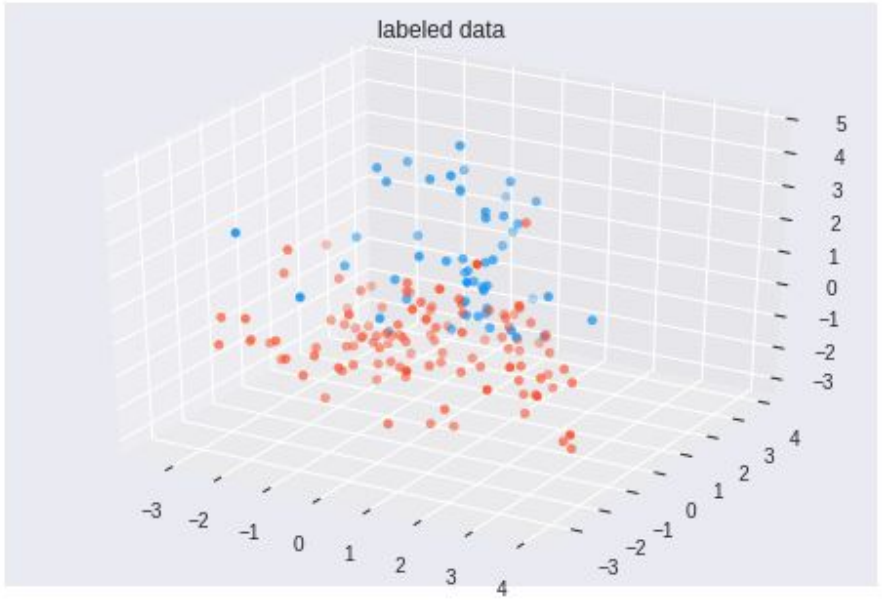
- Los métodos de clustering buscan medir la **similaridad** entre muestras.
- En base a la similaridad se definirán los distintos clusters.
- La similaridad puede ser construida por distancia entre muestras. Existen distintas medidas de distancia, una es la distancia euclídea.
- Los algoritmos de clustering asignan una etiqueta a cada muestra. Dicha etiqueta es el cluster al que pertenece la muestra.



* Elements of Statistical Learning, Tibshirani et Al.

cluster analysis





Cada instancia (sample) no posee etiqueta (izq). Los modelos de clustering buscan encontrar sub-grupos en los datos.

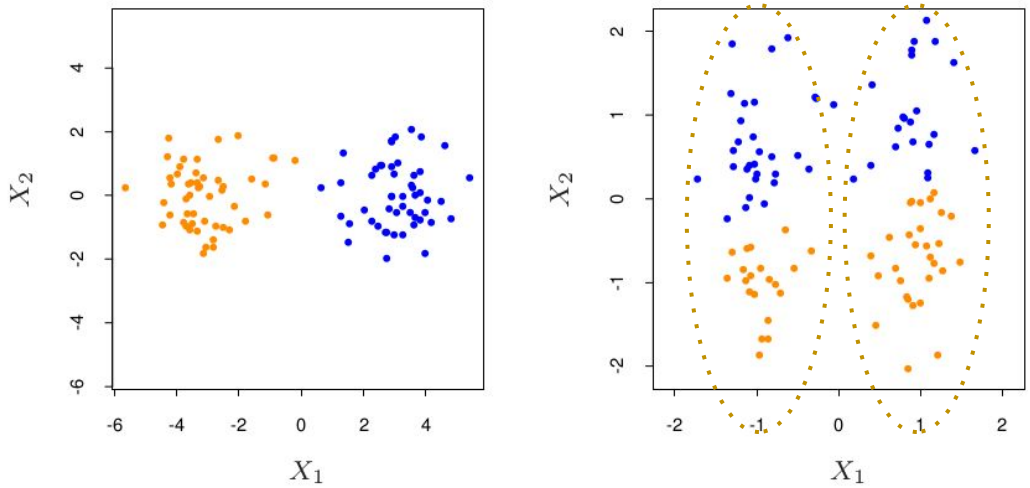
unsupervised

Supervised Analysis: Commercial

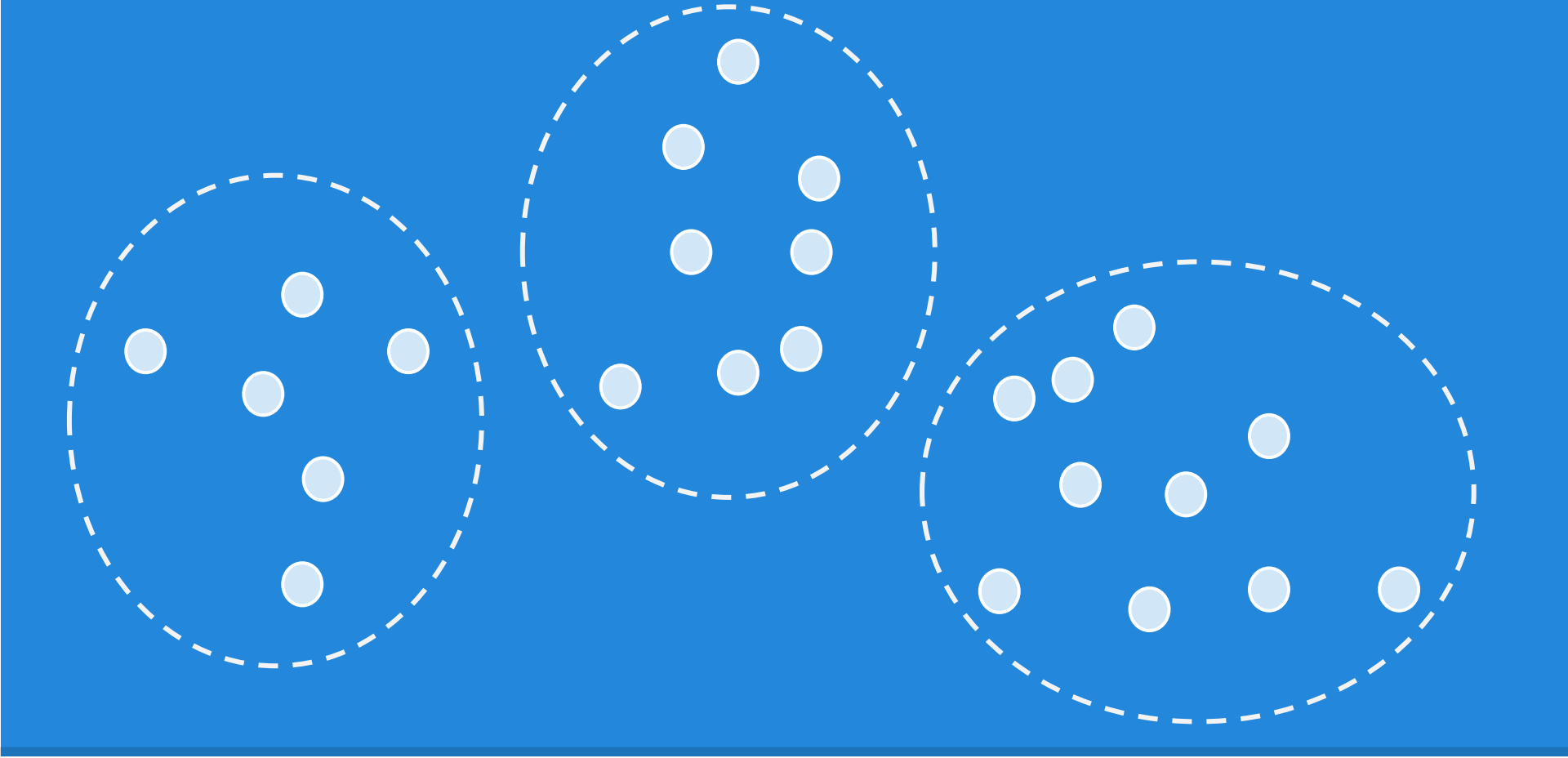
Unsupervised Analysis: Underground

https://towardsdatascience.com/unsupervised-omics-integration-688bf8fa49bf

cluster analysis



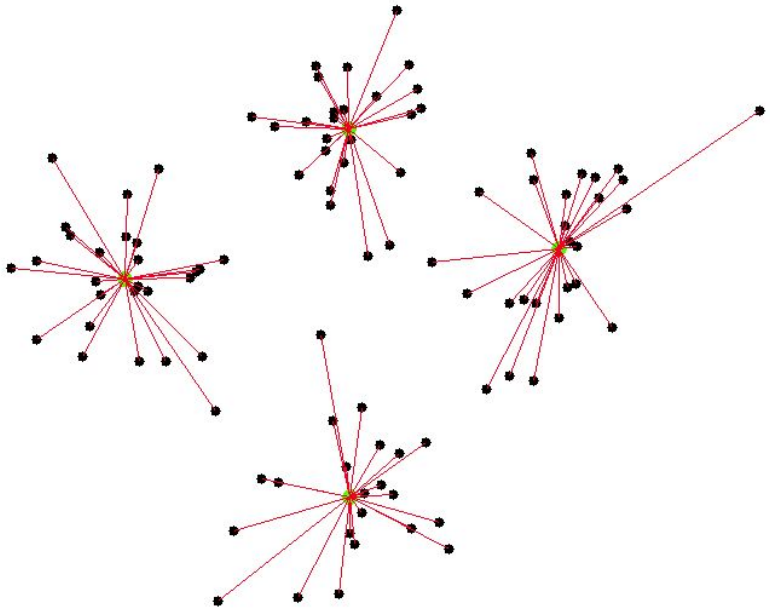
La manera en que los algoritmos de clustering computan similaridad y construyen los clusters puede diferir. Muchas veces la distancia euclídea no es la única (o mejor) manera de agrupar muestras en clusters.



Clustering: algoritmo de k-means

Clustering: algoritmo K-means

- Cada cluster estará identificado por un centroide.
- Una muestra será asignada al cluster cuyo centroide este mas cerca.
- El algoritmo de k-means es iterativo: durante el proceso de aprendizaje los centroides se re-calculan y en consecuencia la pertenencia de las muestras en los clusters.



* http://shabal.in/visuals/kmeans/6.html

Clustering: algoritmo K-means

Consideraciones del algoritmo de K-Means:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2$$

→ La distancia euclidiana cuadrática es la medida de similaridad entre muestras.

$$\min_{C, \{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2$$

→ Función objetivo: minimizar la distancia cuadrática total entre cada muestra “Xi” con el centroide “m” del cluster “k”.

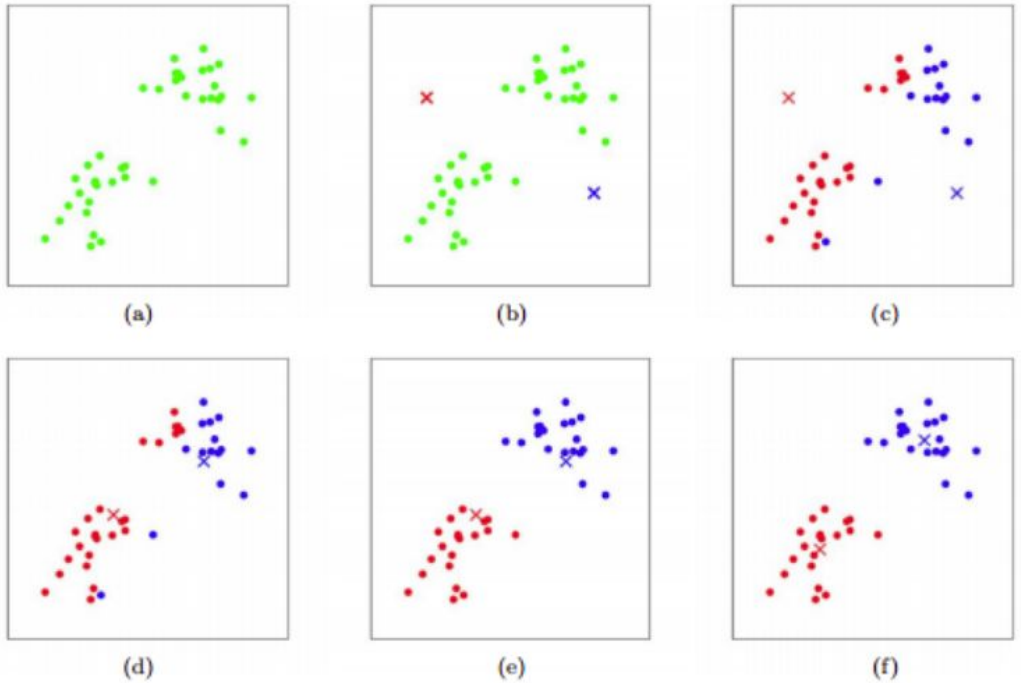
$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} ||x_i - m_k||^2$$

→ Cada muestra Xi será asignada al cluster “k” que presente la distancia cuadrática mas cercana con su centroide “m”.

* Libro: Elements of Statistical Learning, Tibshirani et Al.
* Paper: Data clustering: 50 years beyond K-means. Pattern recognition letters, (2010)

Paso a paso de K-means

1. Hiper Parámetro: número de K clusters
2. El algoritmo comienza asignando un centroide por cada Ki cluster de manera aleatoria.
3. Luego asigna a cada muestra el cluster correspondiente según la distancia al centroide mas cercano.
4. Con los clusters iniciales ya formados se recalculan los centroides de cada cluster para minimizar la distancia entre las muestras y sus respectivos centroides.
5. Con los nuevos centroides calculados se vuelve a computar la distancia entre cada muestra y los centroides actuales. Se re-calcula la pertenencia de cada muestra con cada cluster en función de los nuevos centroides.
6. Iterar hasta que no existan diferencias de de clusters entre iteración e iteración.



http://stanford.edu/~cpiech/cs221/handouts/kmeans.html

k-means

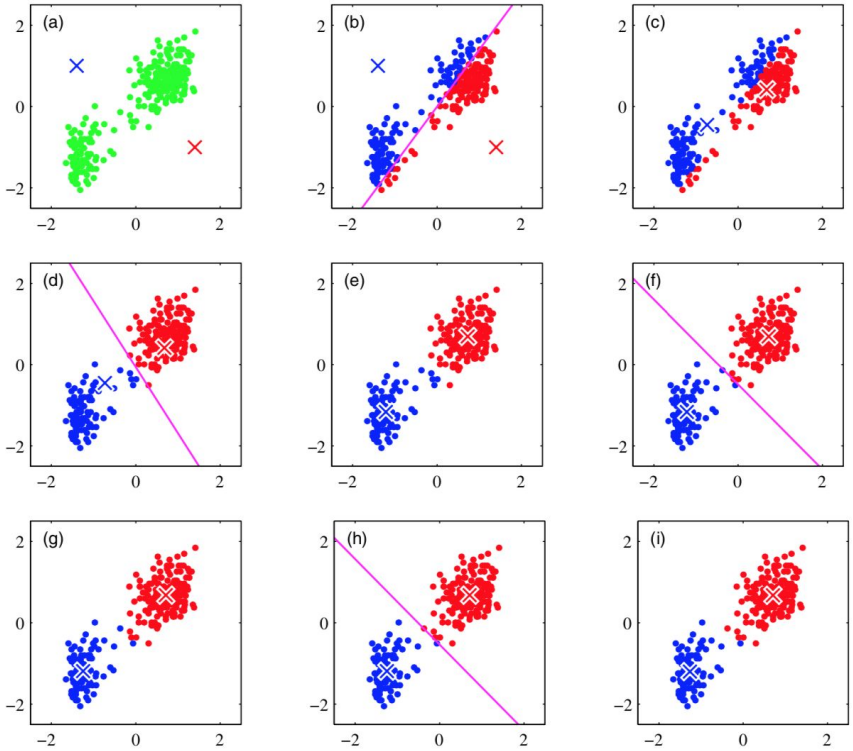
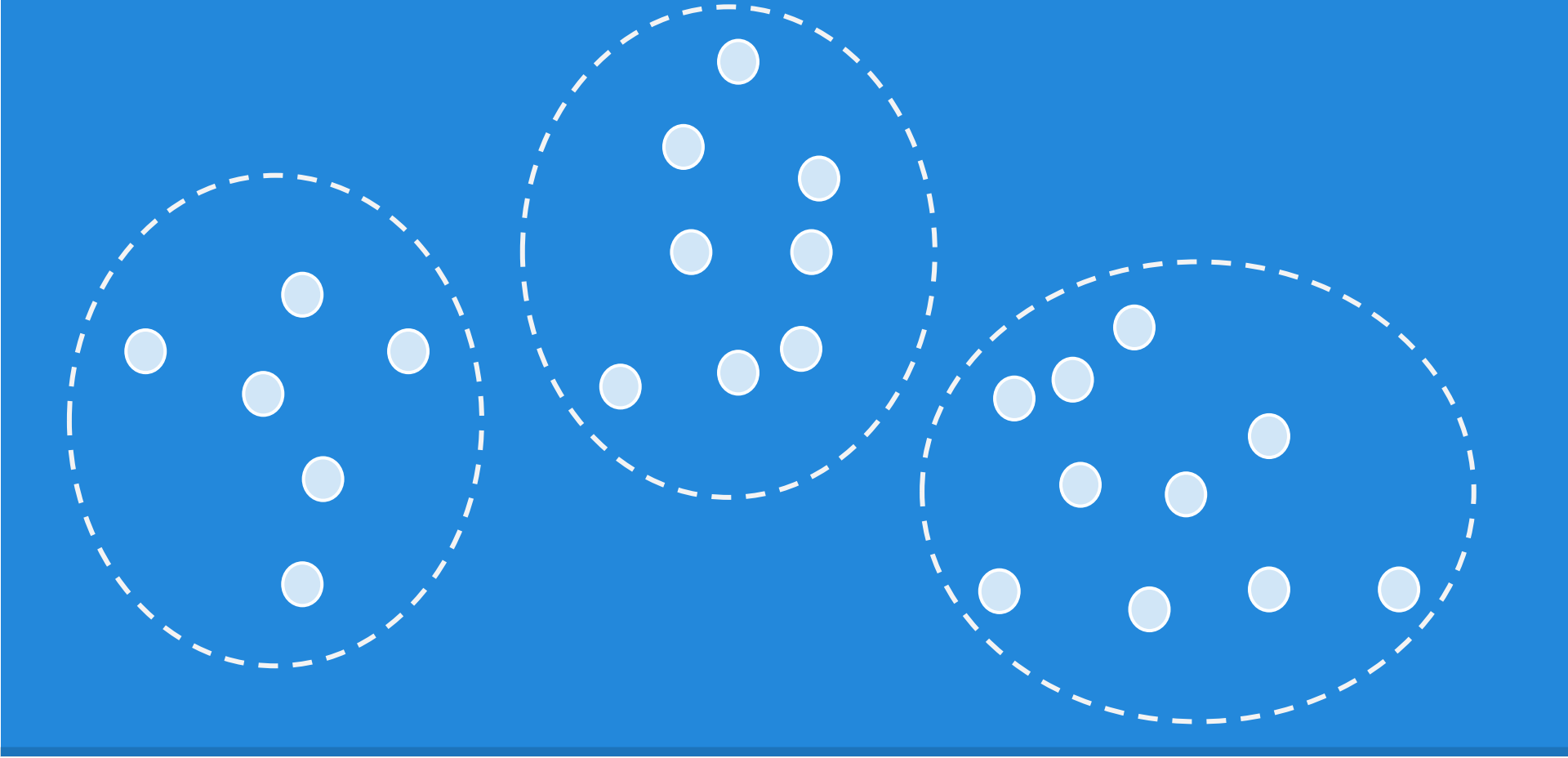


Imagen: Bishop, Machine Learning & Pattern Recognition

K-means clustering

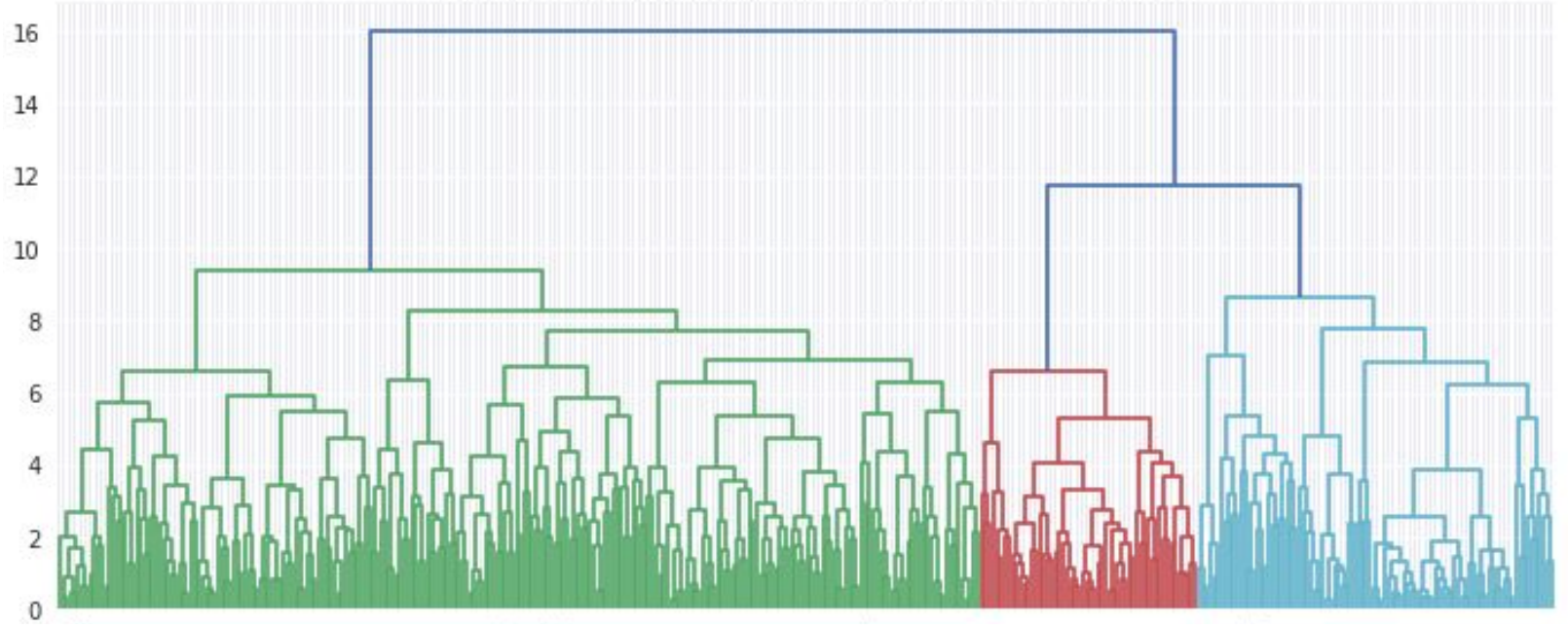
```
from sklearn.cluster import KMeans

km_clust = KMeans(n_clusters=8)
km_clust.fit(x)
cluster_labels = km_clust.labels_
```



Clustering: agrupamiento jerárquico

Hierarchical clustering



Hierarchical clustering

A diferencia del algoritmo K means, el resultado del clustering jerárquico no depende de determinar a priori la cantidad de clusters a formar, ni tampoco una asignación de centroides.

El algoritmo construye representaciones **jerárquicas** en donde los clusters de cada nivel de jerarquía son creados agrupando los clusters del nivel inmediatamente inferior. En el nivel más bajo posible, cada cluster contiene una sola muestra.

Se puede hacer clustering jerárquico **aglomerativo** (bottom up) o divisivo (top down).

El clustering jerárquico aglomerativo implica agrupar desde el nivel inferior de a dos clusters para formar uno en el nivel inmediato superior. El par de clusters seleccionados para agruparse en uno de mayor jerarquía es determinado por la menor “disimilaridad” existente entre pares de clusters.

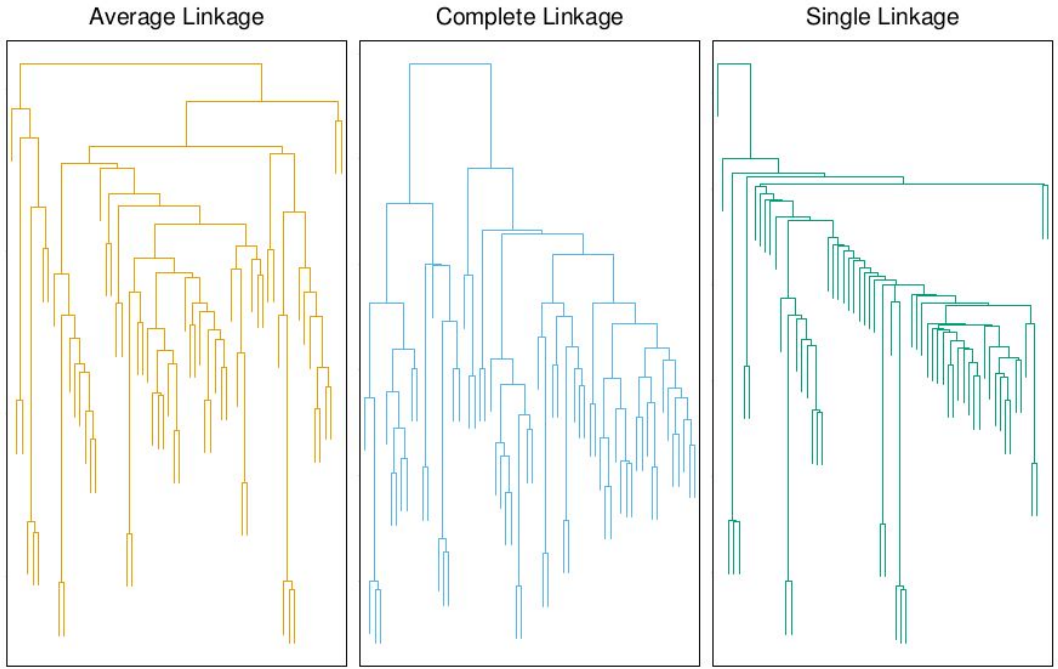
*** Paper: Survey of clustering algorithms (2005) (https://doi.org/10.1109/TNN.2005.845141)**

Hierarchical clustering: agglomerative

Hiperparámetros:

- Cantidad de clusters (que nivel jerárquico considerar).
- Tipo de distancia (euclidea, coseno, manhattan).
- Tipo de linkage (average, complete, single, ward).

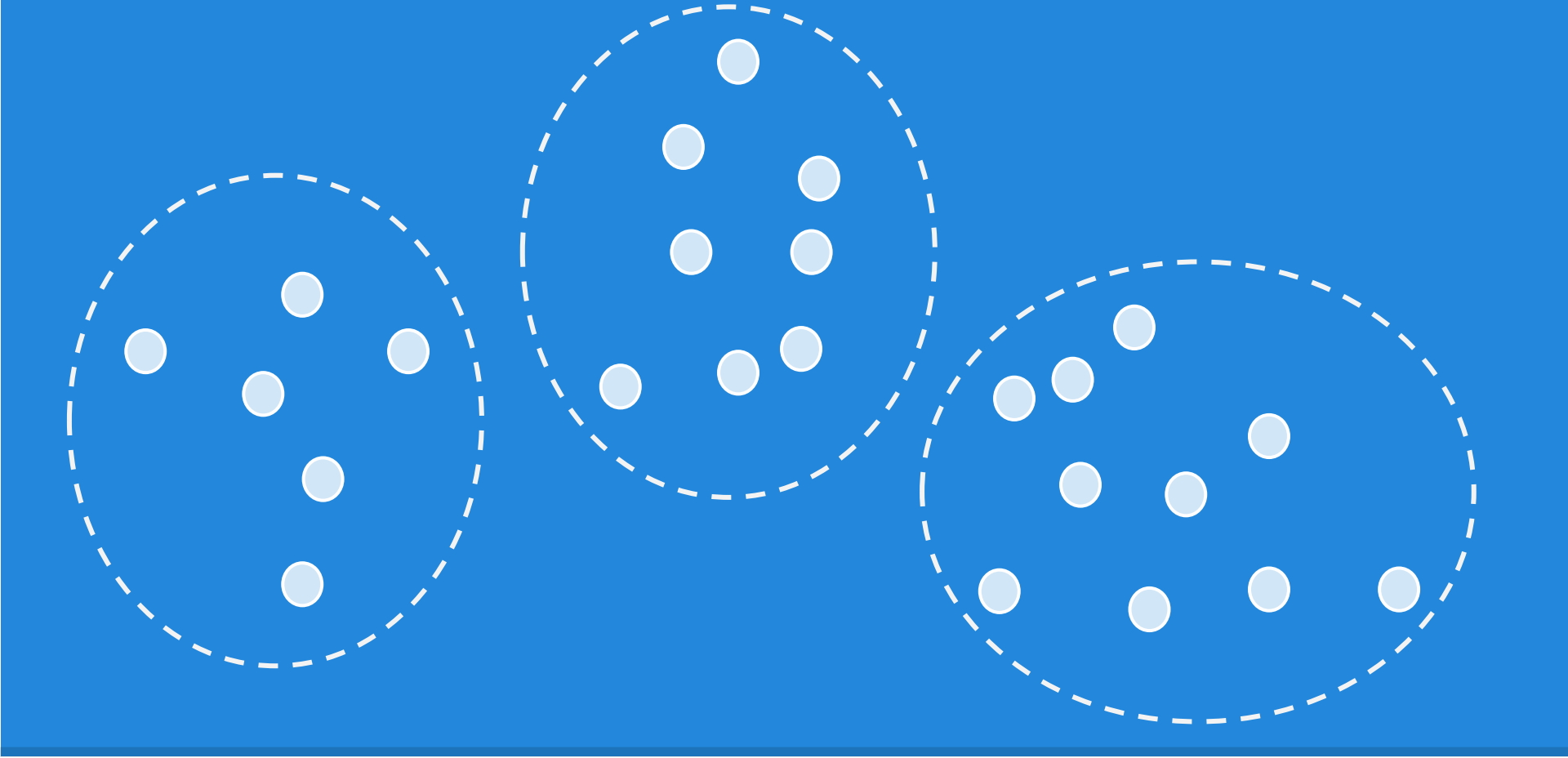
Una combinación distinta de hiperparámetros puede generar diferentes resultados.



* Elements of Statistical Learning, Tibshirani et Al.

Otros algoritmos de clustering

- Spectral Clustering
- DBSCAN
- Gaussian Mixtures



Medicion de la performance de los metodos de clustering

Medir la calidad de los clusters

Una manera de poder medir cuán buenos son los resultados de los algoritmos de clustering es computando el **Silhouette Index (S)**.

El valor de Silhouette mide cuán similar es una muestra respecto a su propio cluster (cohesión) comparado con el resto de los clusters (separación). Este indicador varía entre 1 y -1.

Cuando S es cercano a 1 decimos que los clusters han sido asignados a grupos de muestras bien separadas y definidas. Cuando S = 0 hablamos de clusters que están superpuestos y es dificultoso encontrar grupos bien definidos y compactos. Cuando S = -1 decimos que el algoritmo de clustering asigno erroneamente las etiquetas con respecto a la estructura/distribución de los datos.

*Paper “Model-based evaluation of clustering validation measures”. *Pattern recognition*, *40*(3), 807-824 (2007).

Silhouette Index: Medir la calidad de los clusters

a(i) indica cuán bien fue asignada la muestra “i” al cluster “k”. Cuanto más chico sea mejor es la asignación. Queremos que la similaridad entre muestras de intra-cluster sea alta es decir que la distancia sea baja.

$$a(x_i) = \frac{1}{n_k - 1} \sum_{x_j \in C_k, x_j \neq x_i} d(x_i, x_j)$$

b(i) indica cuán grande es la “disimilaridad” entre la muestra “i” y las muestras de los clusters más próximos a “k”. Queremos que la similaridad inter-cluster, entre muestras de distintos clusters sea baja).

$$b(x_i) = \min_{v=1,\ldots,K, v \neq k} \left[\frac{1}{n_v} \sum_{x_j \in C_v} d(x_i, x_j) \right]$$

Luego para cada muestra “i” calculamos el coeficiente S(i) y finalmente calculamos el Silhouette Score Sx para todas las muestras del dataset **X**. Cuanto mas alto sea el Silhouette Score (mas cerca de 1) quiere decir que los clusters asignados por el algoritmo de clustering están bien separados y definidos. Cuanto más cerca de 0 decimos que los clusters son muy difusos y poco definidos.

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max[b(x_i), a(x_i)]}$$

$$S_X = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{n_k} \sum_{x \in C_k} S(x_i) \right]$$

Silhouette con sklearn

```
from sklearn import metrics

metrics.silhouette_score(X, cluster_labels, metric='sqeuclidean')
```

Rand Index

El Rand Index es una métrica utilizada cuando las etiquetas de las muestras (ground truth) están disponibles únicamente para validar a posteriori la calidad de los clusters obtenidos.

El Rand Index compara la “pureza” de los clusters en función de la homogeneidad de clases que encuentre cada uno. Un cluster donde todas sus muestras pertenecen a la misma clase asigna un rand index mayor.

El rand index es una manera de plantear el ‘accuracy’ de clustering. Cuidado! En ningún momento se utilizan las etiquetas para el entrenamiento. Solo para la validación final.

Rand Index

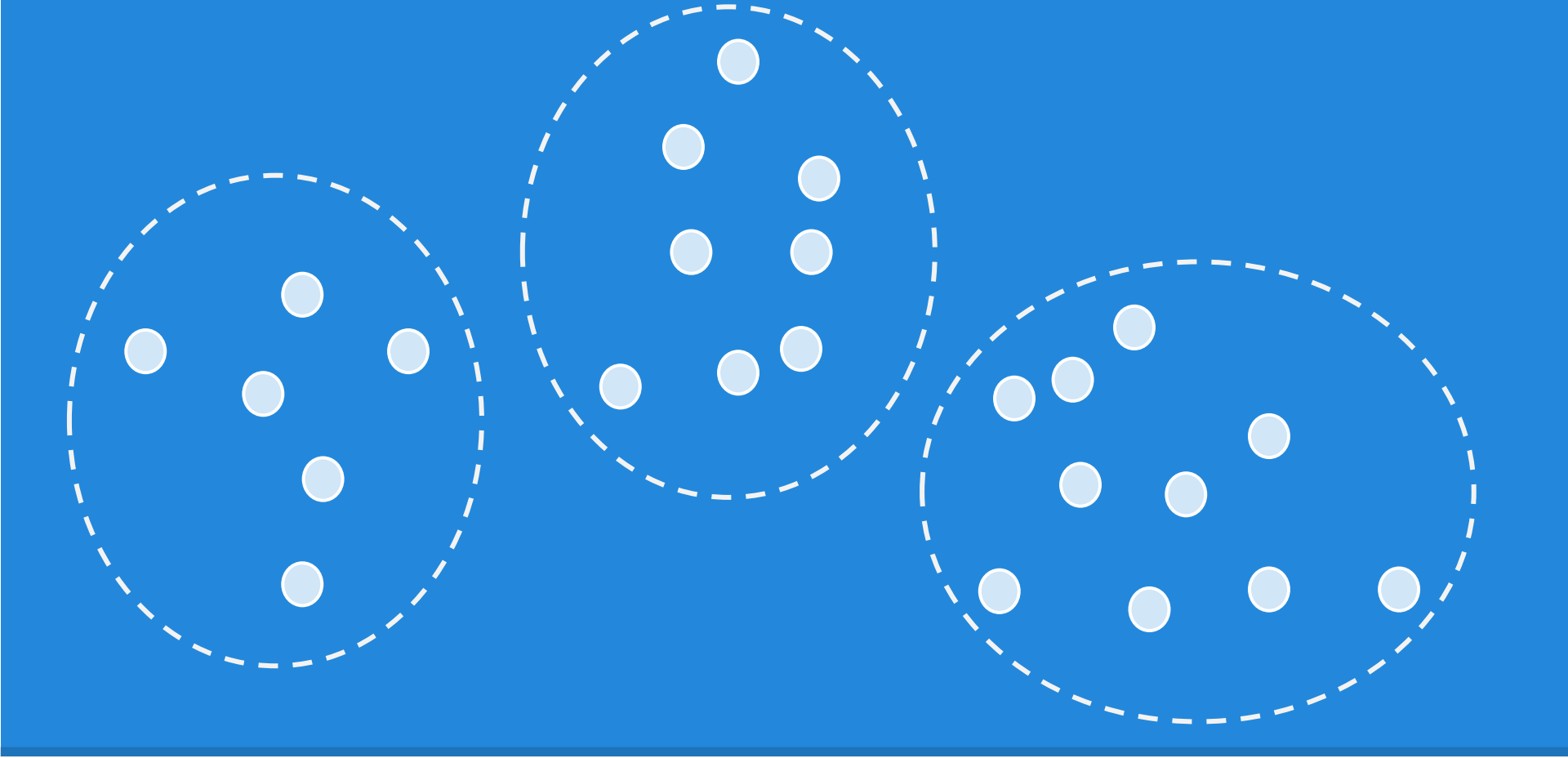
$$\text{Rand Index} = \frac{A + B}{A + B + C + D}$$

A: número de pares de muestras asignadas al mismo cluster y simultáneamente pertenecientes a la misma clase.

B: número de pares de muestras asignadas a distintos clusters y simultáneamente pertenecientes a distintas clases.

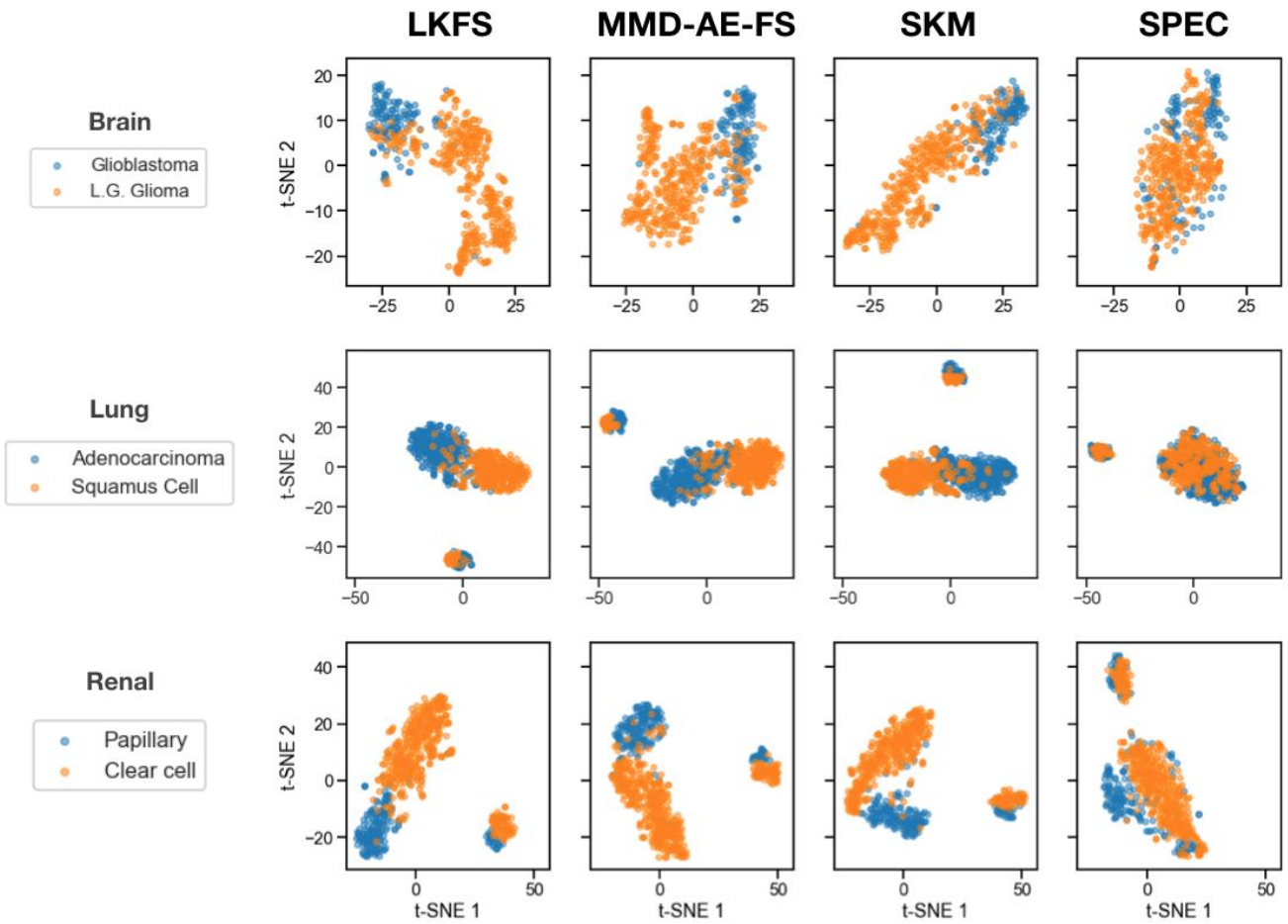
C: número de pares de muestras asignadas al mismo cluster y pertenecientes a distintas clases.

D: número de pares de muestras asignadas a distintos clusters y pertenecientes a la misma clase.



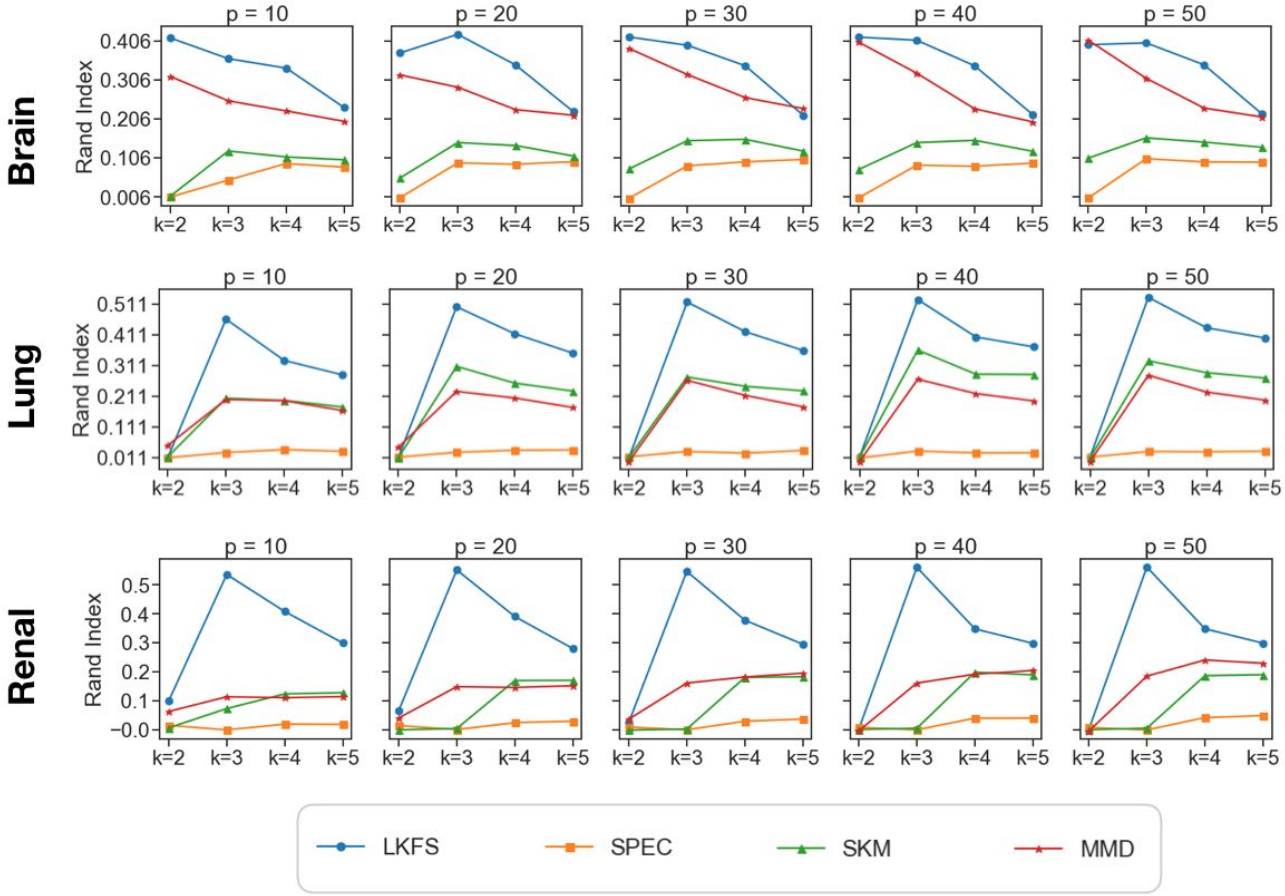
Caso de aplicación de clustering: cancer genomics

Clustering en perfiles Genomicos tumorales



Palazzo, M., Beauseroy, P., & Yankilevich, P. (2020). Unsupervised Feature Selection for Tumor Profiles using Autoencoders and Kernel Methods. https://arxiv.org/pdf/2007.06106.pdf

Clustering en perfiles Genomicos tumorales



Palazzo, M., Beauseroy, P., & Yankilevich, P. (2020). Unsupervised Feature Selection for Tumor Profiles using Autoencoders and Kernel Methods.
https://arxiv.org/pdf/2007.06106.pdf


CON QUE TIPO DE APRENDIZAJE AGRUPAMOS DATOS?
NO SUPERVISADO
CLUSTERING
CLASIFICACION?