

Curso 2020

# MACHINE LEARNING UNSUPERVISED LEARNING

---

CRISANTO DE LOS SANTOS DURÁN





## **UNSUPERVISED LEARNING**

**CRISANTO DE LOS SANTOS DURÁN**

**CRISANTODLS@FACULTY.MIOTI.ES**

**Linkedin:**

**linkedin.com/in/crisantodelossantos**



ML2 Unsupervised Learning

## Unas palabras sobre nosotros...

Mi nombre es...  
Actualmente trabajo en ...

Mi formación es...

Me gustaría que ...

# Objetivos

- Diferencias entre Supervised y Unsupervised learning
- Principales técnicas en Unsupervised learning



ML2 Unsupervised Learning

# SUPERVISED LEARNING VS. UNSUPERVISED LEARNING

## SUPERVISED LEARNING

Descubrir patrones en los datos relacionados con un atributo (clase o etiqueta) target.

- Estos patrones son utilizados para predecir los valores del atributo target en nuevas instancias.

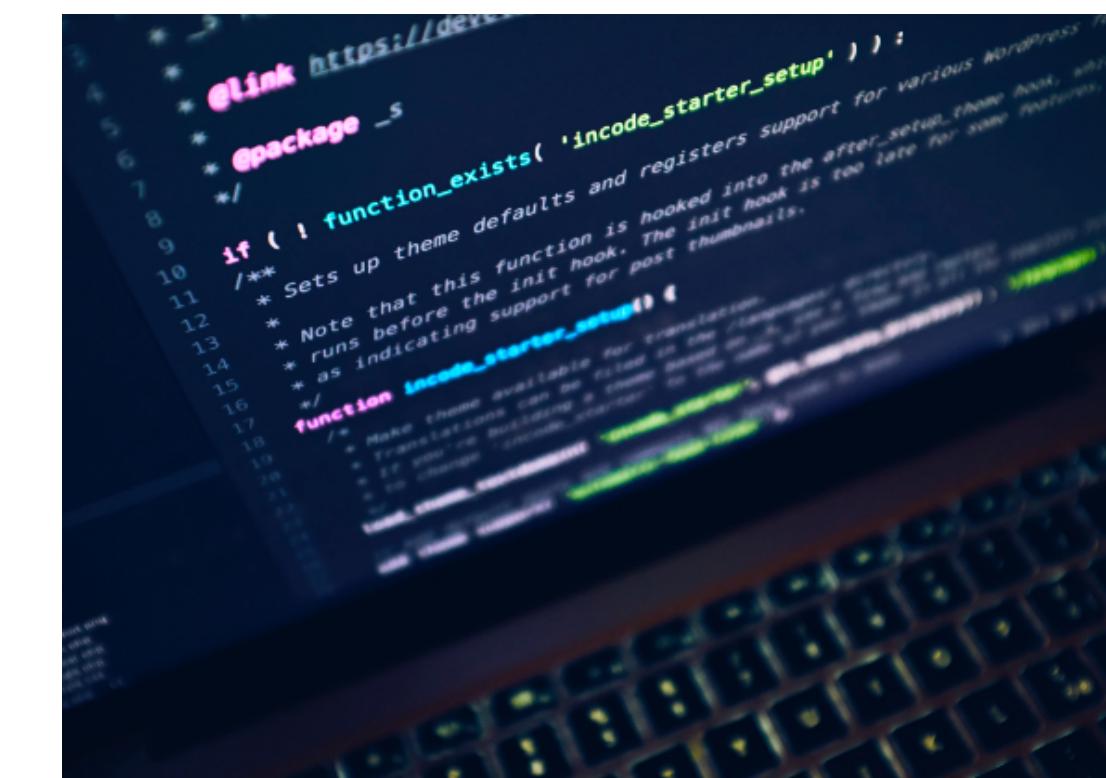
## UNSUPERVISED LEARNING

Los datos no tienen un atributo target (no disponemos de etiquetas).

- Exploramos los datos para encontrar algunas estructuras intrínsecas en ellos.



## SUPERVISED LEARNING

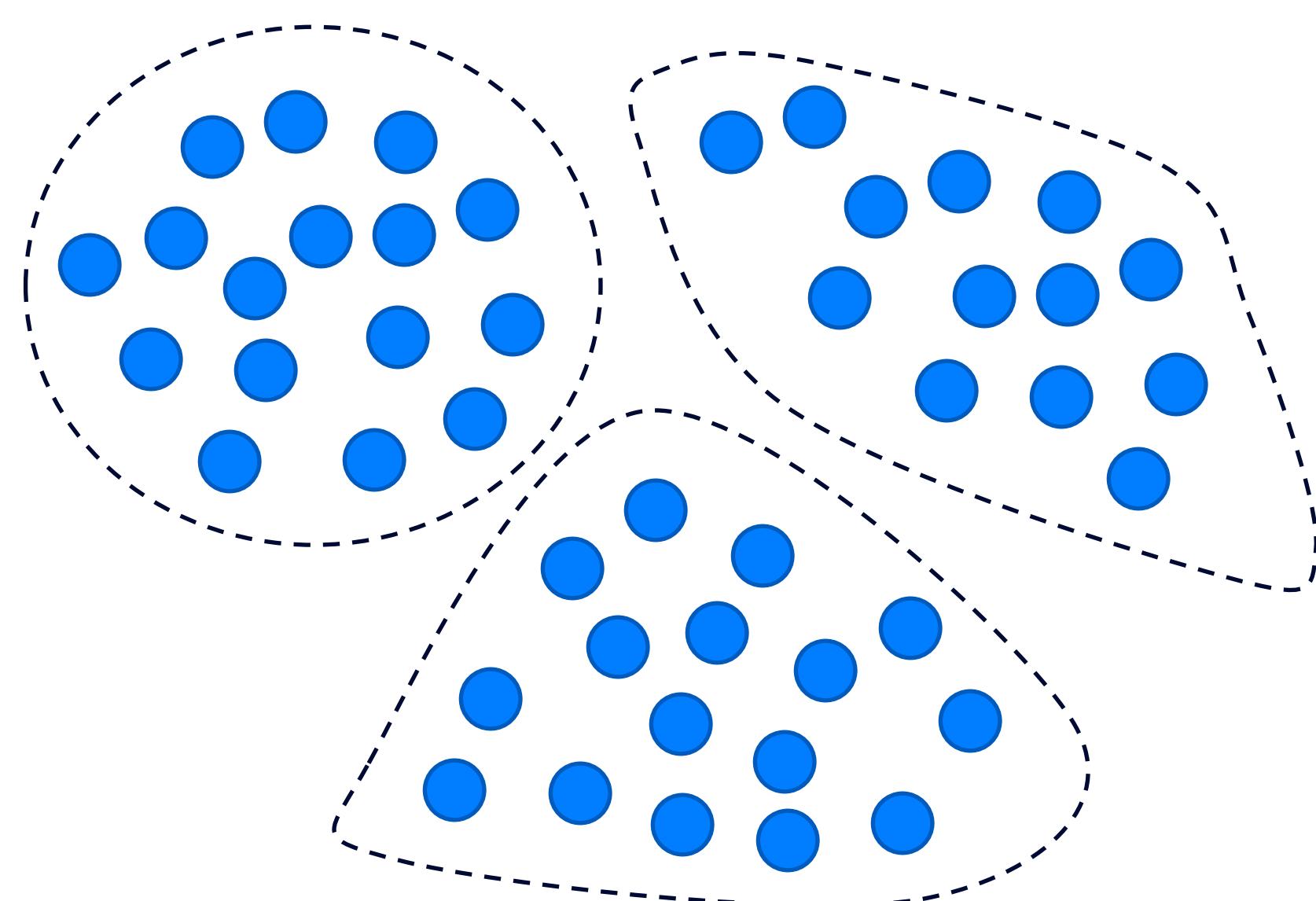
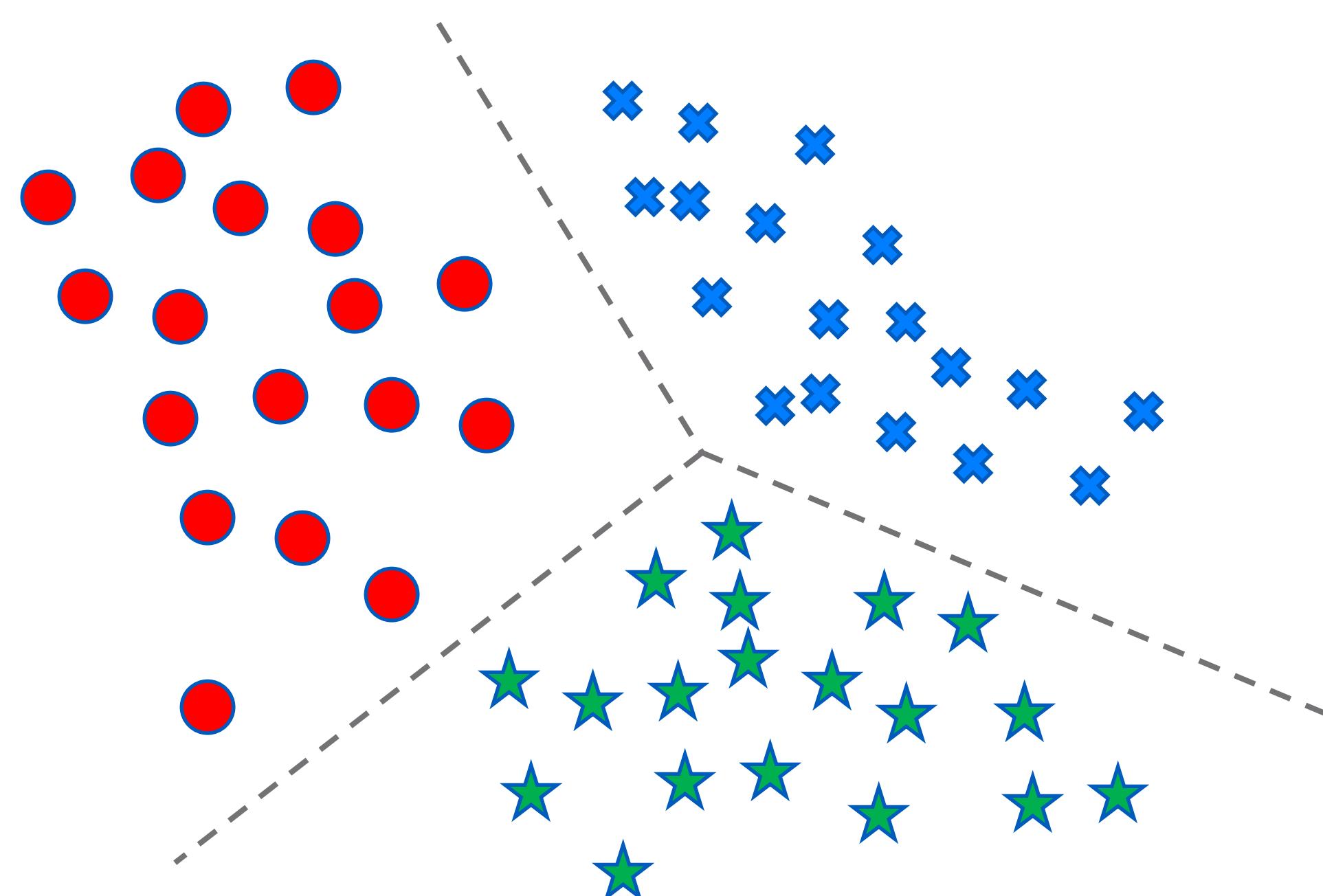


## UNSUPERVISED LEARNING



UNA APROXIMACIÓN VISUAL

# Supervised learning vs. Unsupervised learning



UNA APROXIMACIÓN VISUAL

# Cómo los agruparías?



UNA APROXIMACIÓN VISUAL

# Cómo los agruparías?



UNA APROXIMACIÓN VISUAL

# Cómo los agruparías?



UNA APROXIMACIÓN VISUAL

# Cómo los agruparías?



Y NUEVAS CATEGORÍAS DE APRENDIZAJE

# Supervised vs. Unsupervised learning

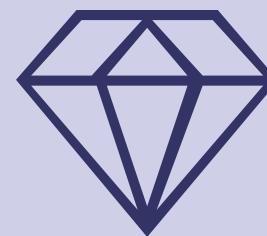
$$\{x_n \in R^d, y_n \in R\}_{n=1}^N$$

$$\{x_n \in R^d\}_{n=1}^N$$



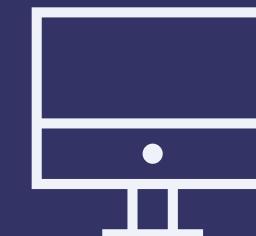
## SUPERVISED LEARNING

PREDICTION  
CLASSIFICATION



## UNSUPERVISED LEARNING

DIMENSION REDUCTION  
FINDING ASSOCIATION  
CLUSTERING  
PROBABILITY DISTRIBUTION  
ESTIMATION



## SEMI-UNSUPERVISED

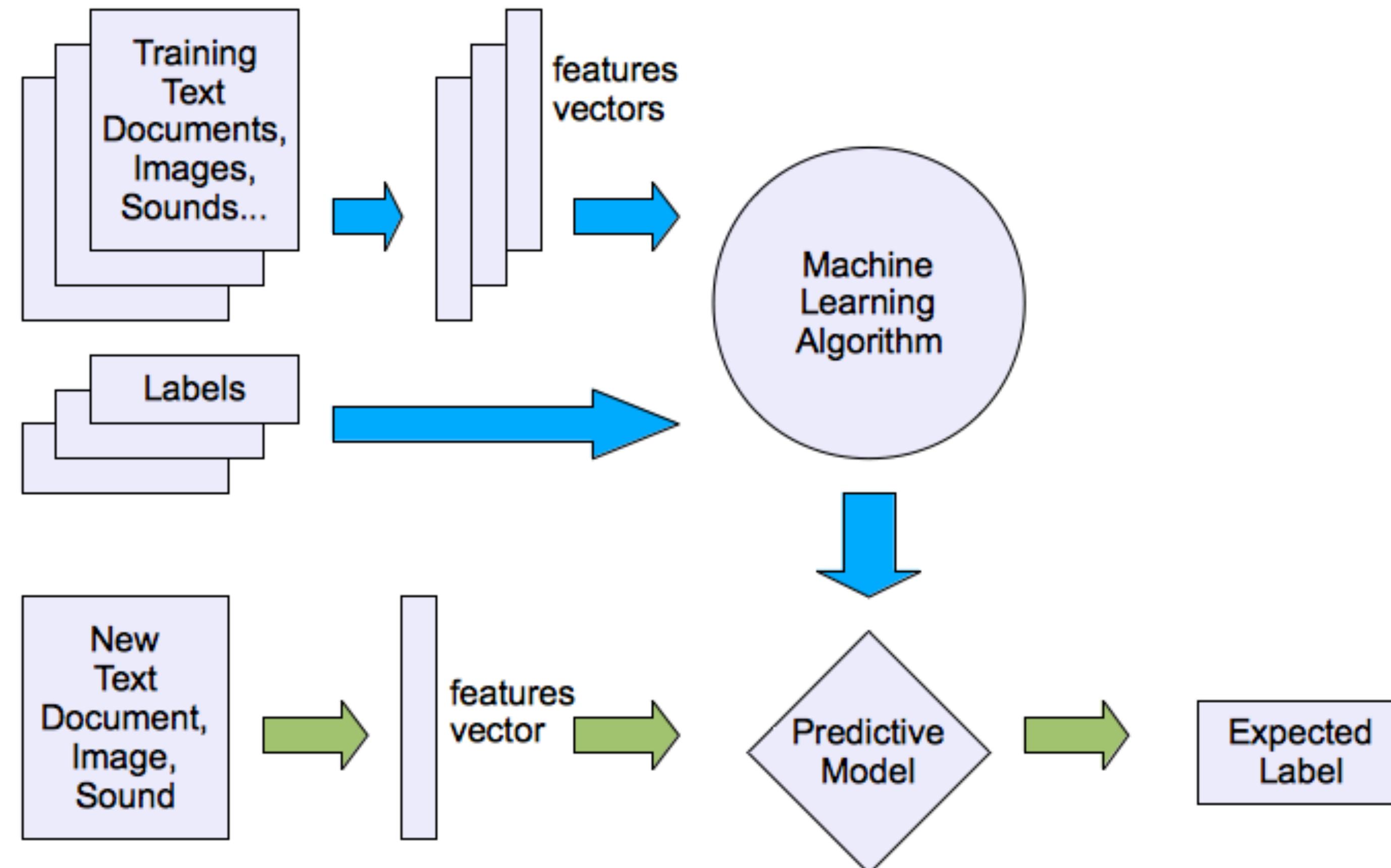
CLASSIFICATION  
DIMENSION REDUCTION  
CLUSTERING



## REINFORCEMENT

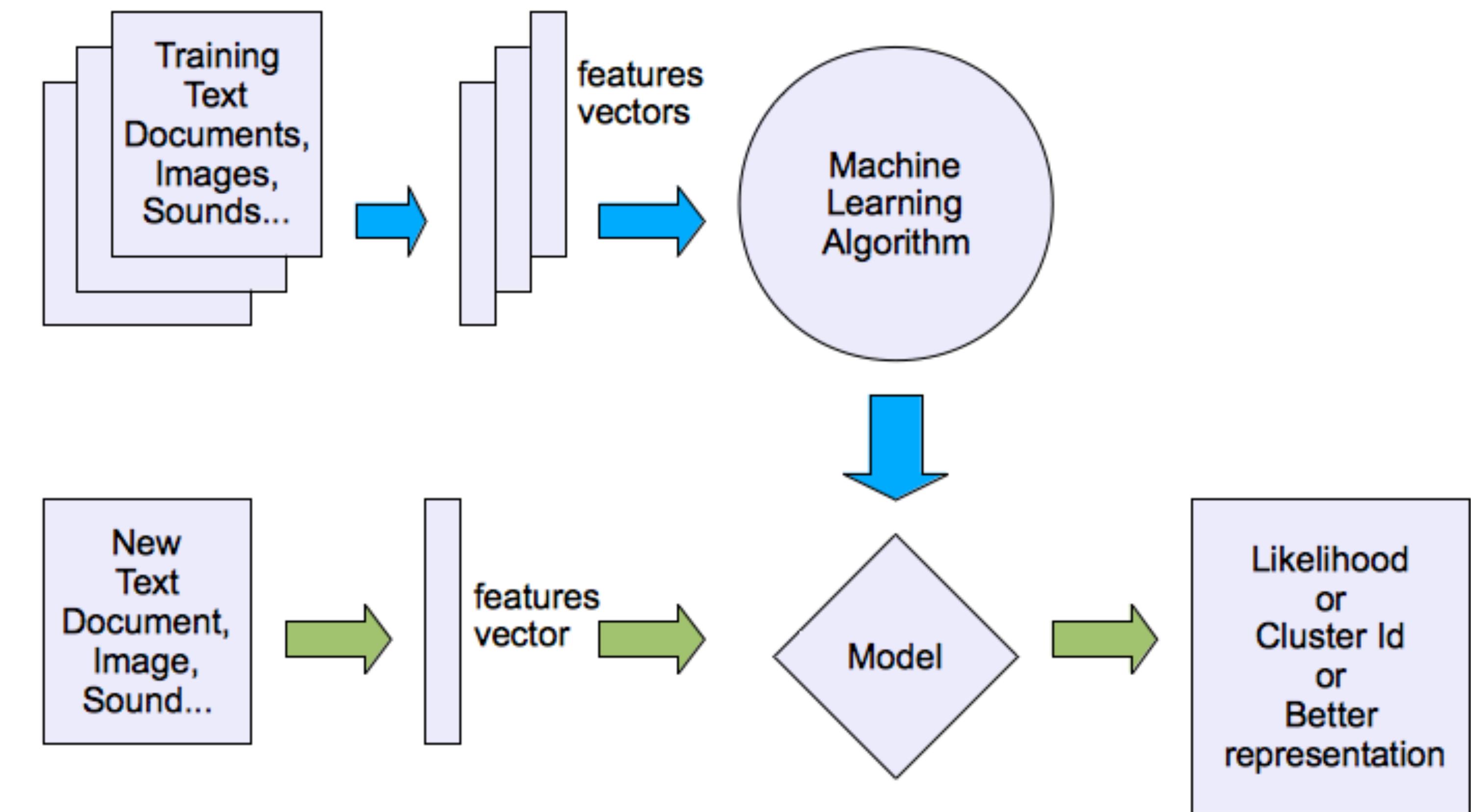
DECISION MAKING  
ROBOTS  
CHESS MACHINE

# Estructura Supervised Learning





# Estructura Unsupervised Learning



# Qué enfoque tenemos en cada caso

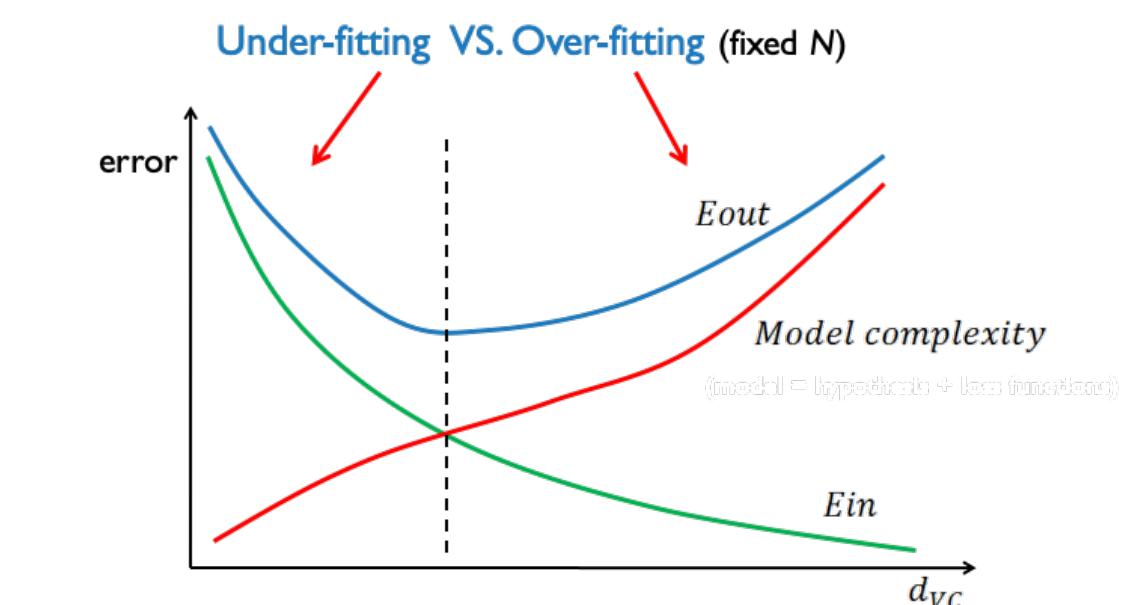


SUPERVISED: LOW E-OUT OR MAXIMIZE PROBABILISTIC TERMS

$$\text{error} = \frac{1}{N} \sum_{n=1}^N [y_n \neq g(x_n)]$$

$$Eout(g) \leq Ein(g) \pm O\left(\sqrt{\frac{d_{VC}}{N} \ln N}\right)$$

E-in: for training set  
E-out: for testing set

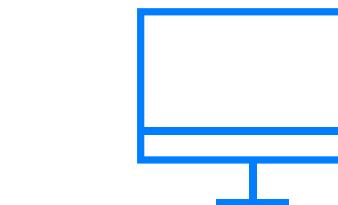
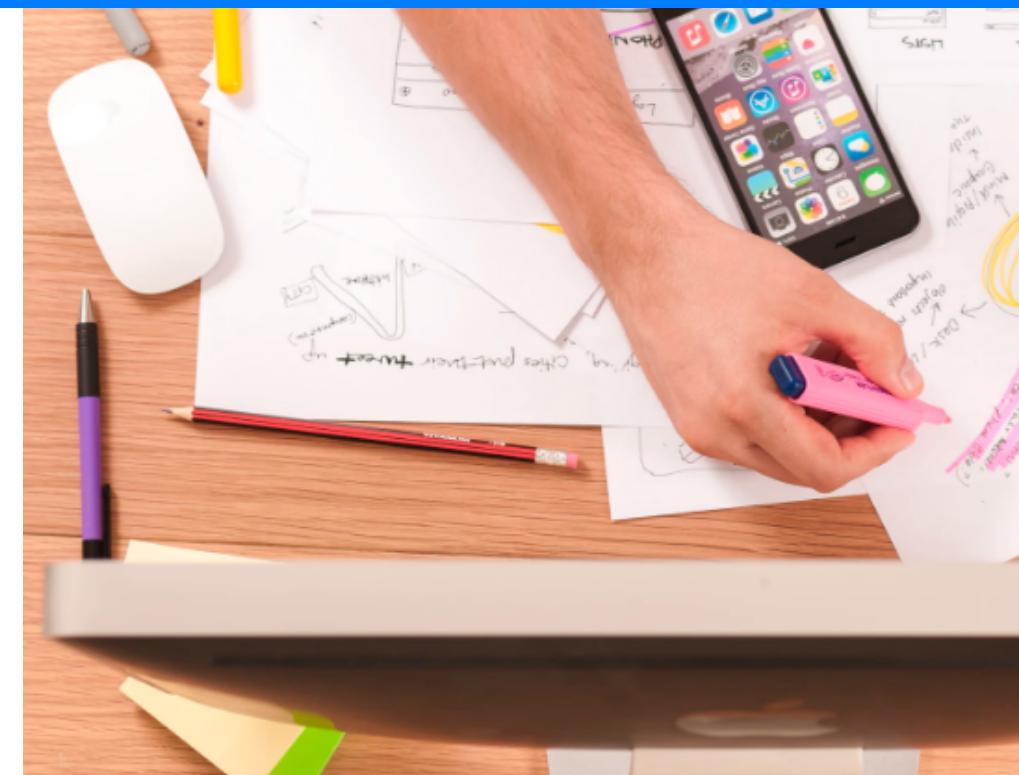
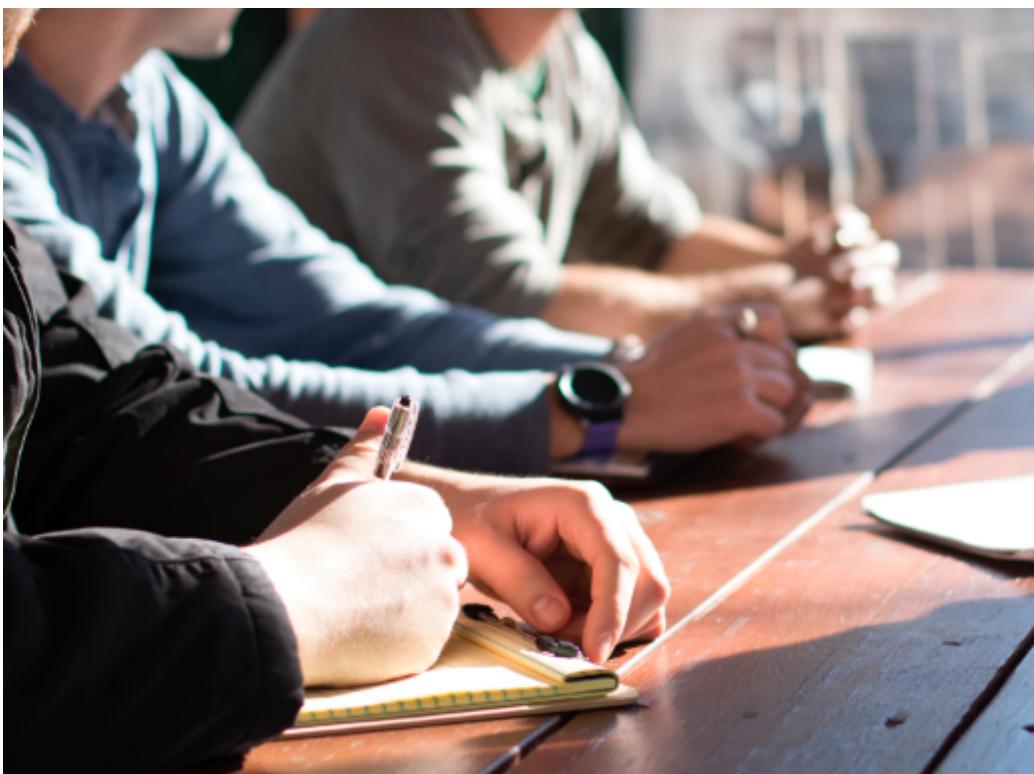


UNSUPERVISED: MÍNIMUM QUANTIZATION ERROR, MÍNIMUM DISTANCE, MLE (MÁXIMUM LIKELIHOOD ESTIMATION)

# Técnicas en Unsupervised learning



REDUCCIÓN



CLUSTERIZACIÓN

1

2

3

4

## REDUCCIÓN DE LA DIMENSIÓN

- Principal component analysis (PCA)
- Factor analysis
- t-Distributed Stochastic Neighbor Embedding (t-SNE)

## ANOMALY DETECTION

- Z-score
- Dbscan

## CLUSTERING

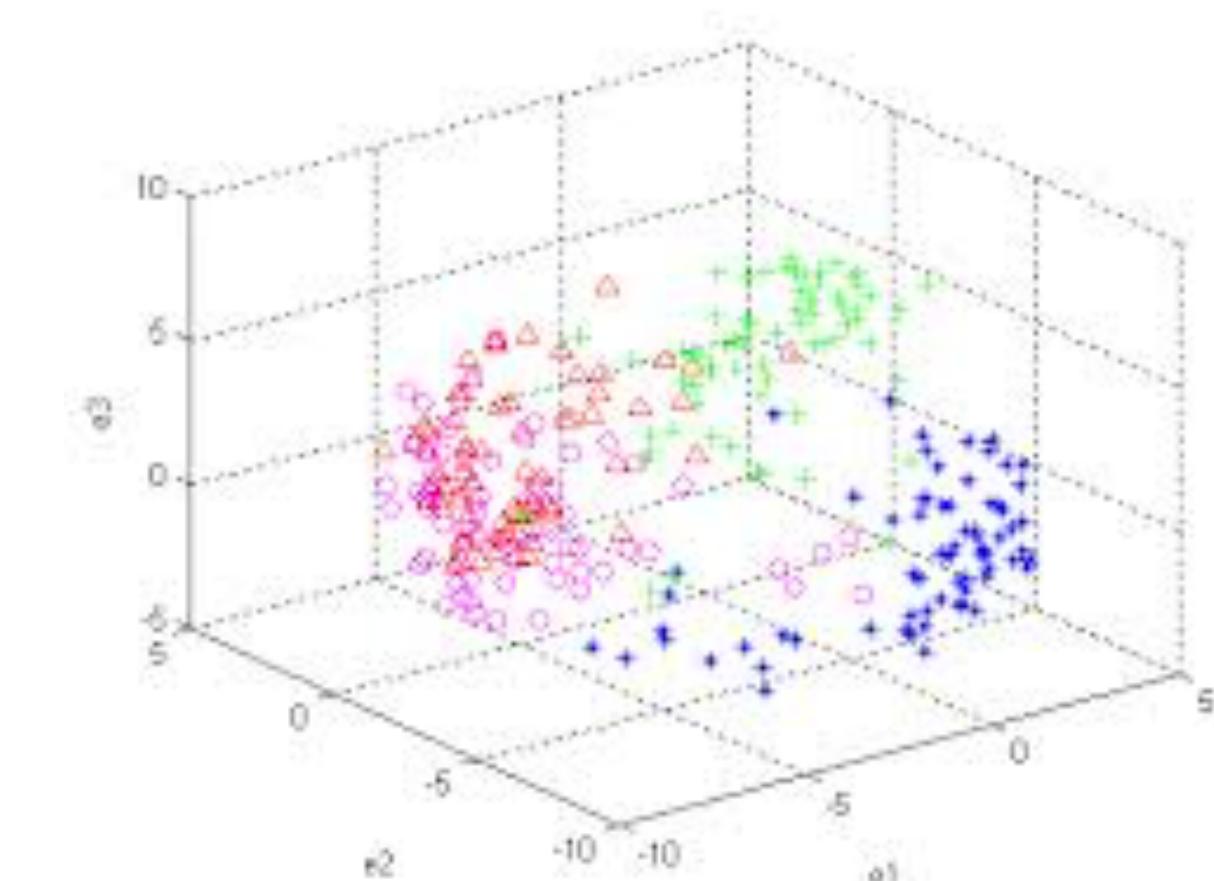
- K-means clustering
- Hierarchical clustering

## ASSOCIATION RULES

- A priori
- FP Growth

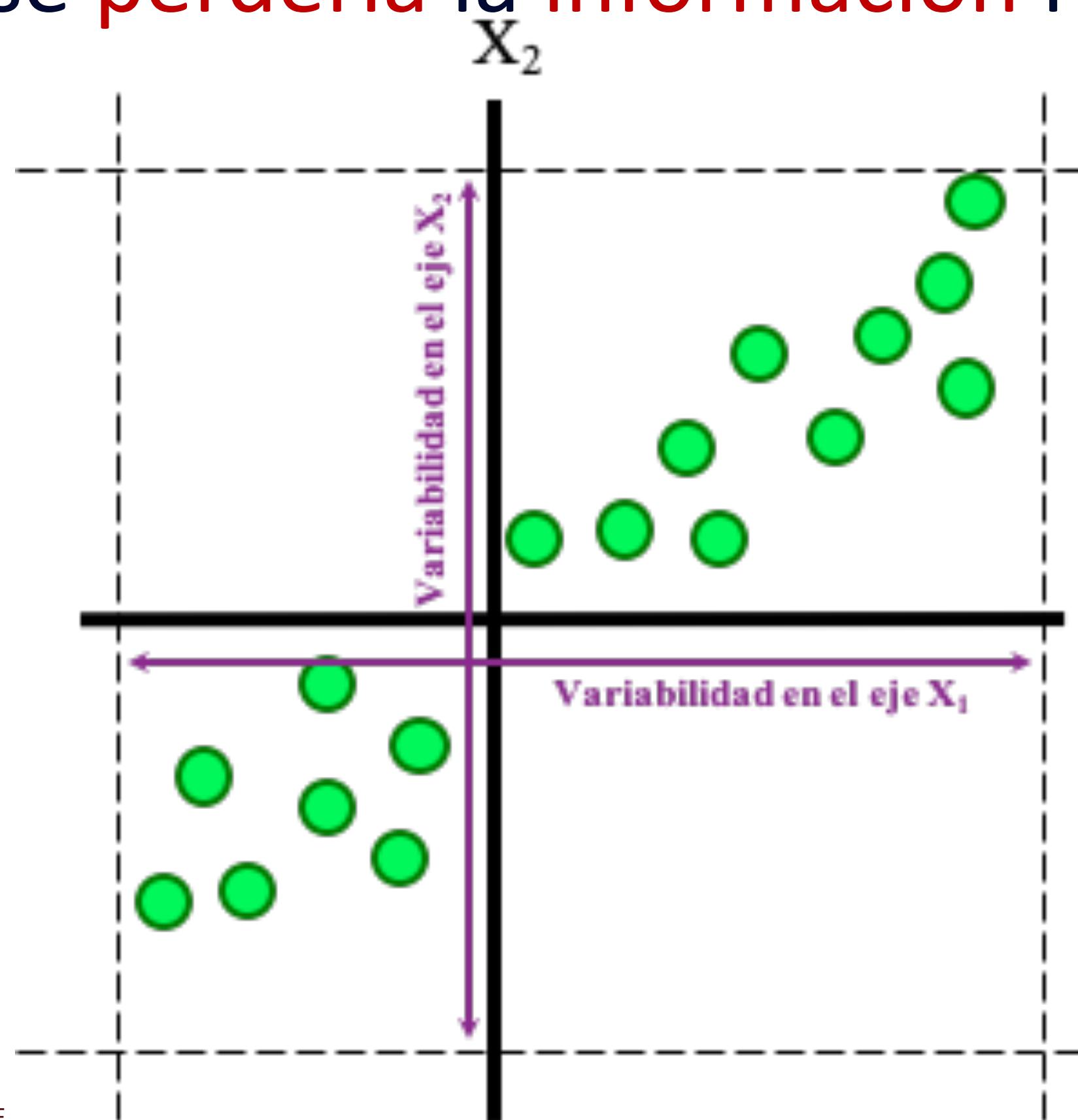
# PCA

- El **análisis de componentes principales (PCA)** tiene por **finalidad reducir la dimensionalidad** de un problema respecto al número de variables implicadas en él: pasar de  $p$  a  $q$  variables con  $q << p$ .
- La **información** contenida en las  $p$  variables está muy **relacionada** con su **variabilidad**: “Si los datos son muy distintos entre sí (mucha variabilidad), existe más información a explicar, y por tanto, se necesitarán más componentes principales”.
- La **finalidad** es que la **variabilidad explicada por las componentes sea lo mayor posible**: “Conseguir explicar con pocas variables un alto porcentaje de la variabilidad de las variables de partida”.



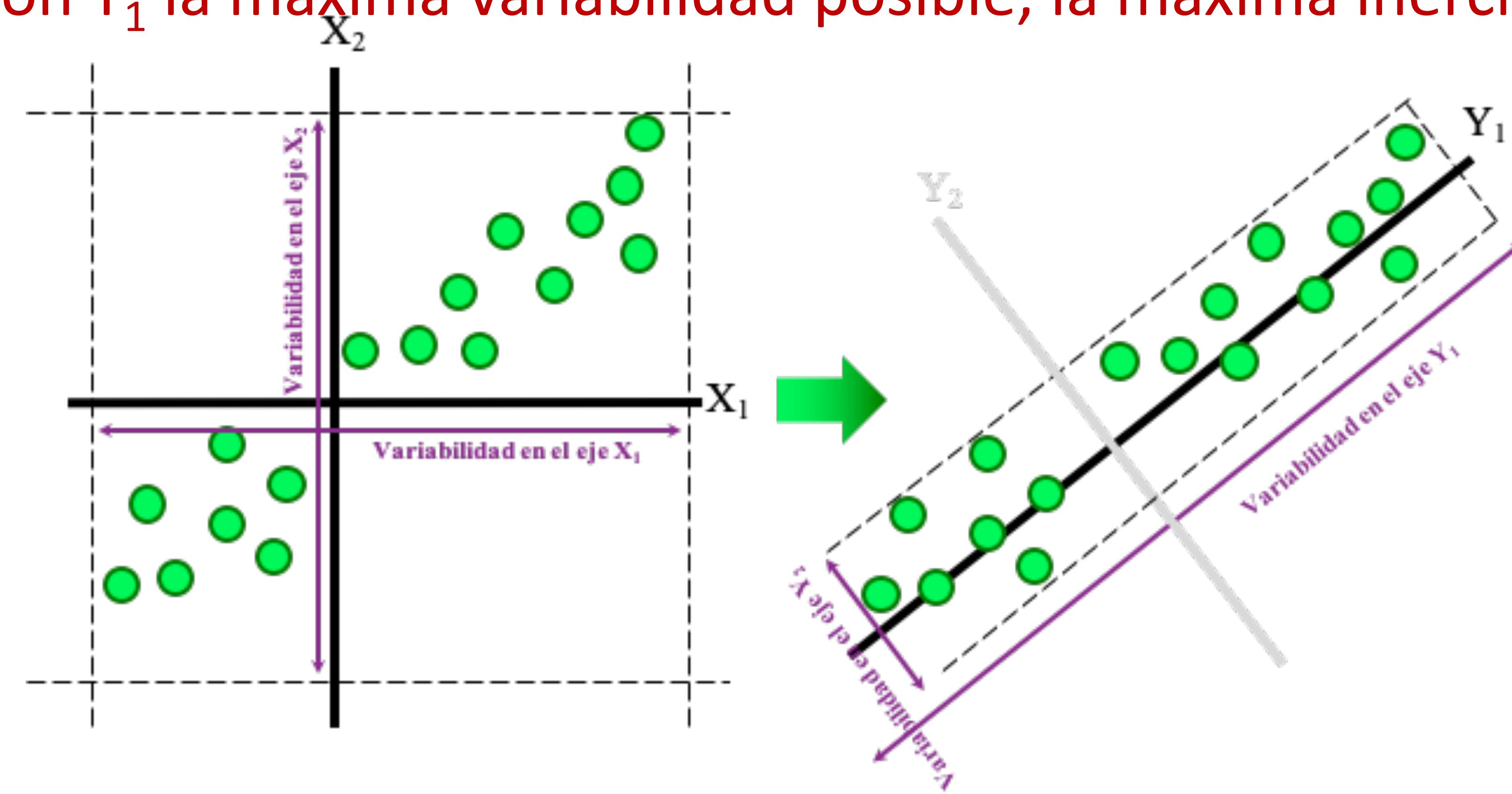
# PCA

- En el **caso bidimensional**, el objetivo sería reducir a **una única dimensión**.
- Si nos quedáramos **únicamente** con la información relativa a la variable  $X_1$ , se perdería la **información** relativa a la ordenada del punto.

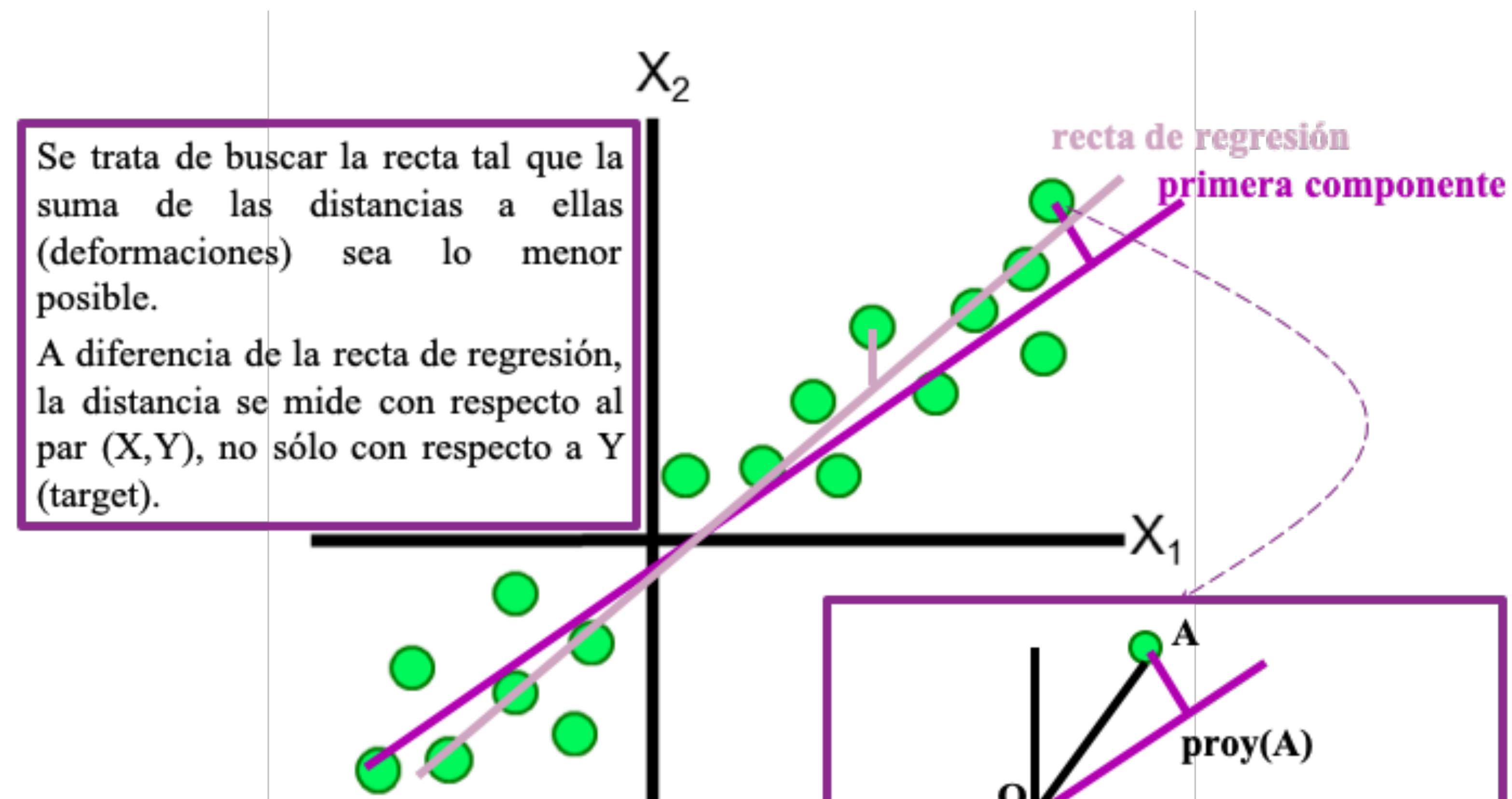


# PCA

- Se trata de buscar una rotación de  $X_1$  a  $Y_1$  de forma que la diferencia entre el máximo y el mínimo valor al proyectar sobre  $Y_1$  sea lo mayor posible: **capturar con  $Y_1$  la máxima variabilidad posible, la máxima inercia.**



# PCA



Se trata de buscar la recta tal que la suma de las distancias a ellas (deformaciones) sea lo menor posible.

A diferencia de la recta de regresión, la distancia se mide con respecto al par (X,Y), no sólo con respecto a Y (target).

El problema de minimizar la suma de deformaciones a una recta se convierte en un problema de maximizar inercia.

$$\overline{proy(A)A} = \overline{OA} - \overline{Oproy(A)}$$

$$\text{Min } \{\overline{proy(A)A}\} \Leftrightarrow \text{Max } \{\overline{Oproy(A)}\}$$

# Aplicaciones PCA

- Usos:
  - Data Visualization
  - Data Reduction
  - Data Classification
  - Trend Analysis
  - Factor Analysis
  - Noise Reduction
- Oportunidad de responder a:
  - Cuantos sub-conjuntos únicos tenemos?
  - Como son de similares / diferentes?
  - Cuales son los factores subyacentes que influyen en la muestra?
  - Qué tendencias temporales están correlacionadas?
  - Qué mediciones necesitamos diferenciar?
  - Como presentar mejor lo que es “interesante”?
  - A qué sub-conjunto pertenece esta nueva muestra?

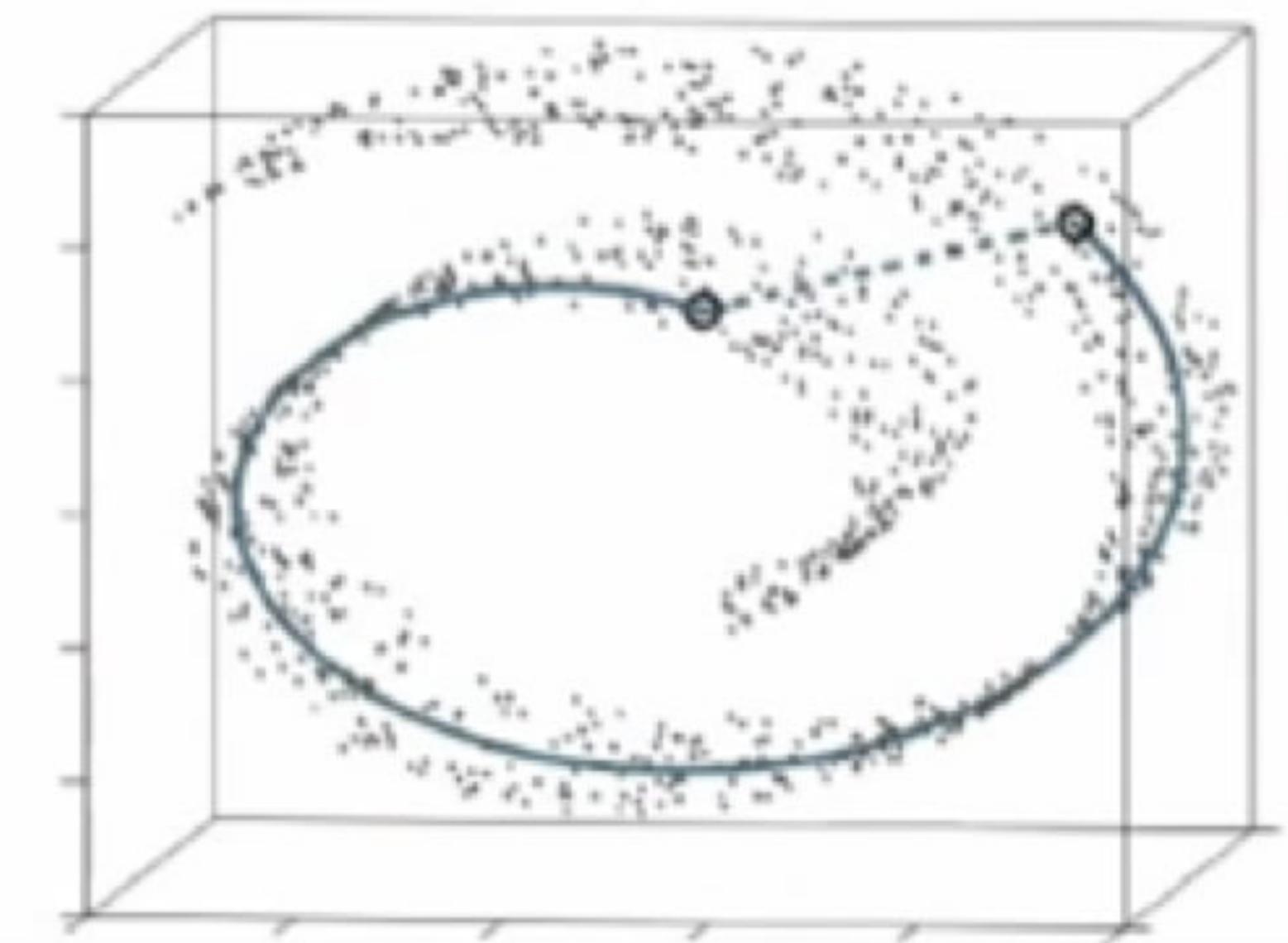
# t-SNE

- **Qué proporciona**

- Un modo intuitivo para entender como están organizados los datos en un espacio de dimensión grande. Fue desarrollado por Laurens van der Maaten y Geoffrey Hinton en 2008.

- **Diferencias con PCA**

- PCA fue desarrollado en 1933 y t-SNE en 2008
- Preserva pequeñas distancias o similitudes locales frente preservar grandes distancias para maximizar la varianza



Swiss Roll Dataset

# t-SNE

- **Paso 1**

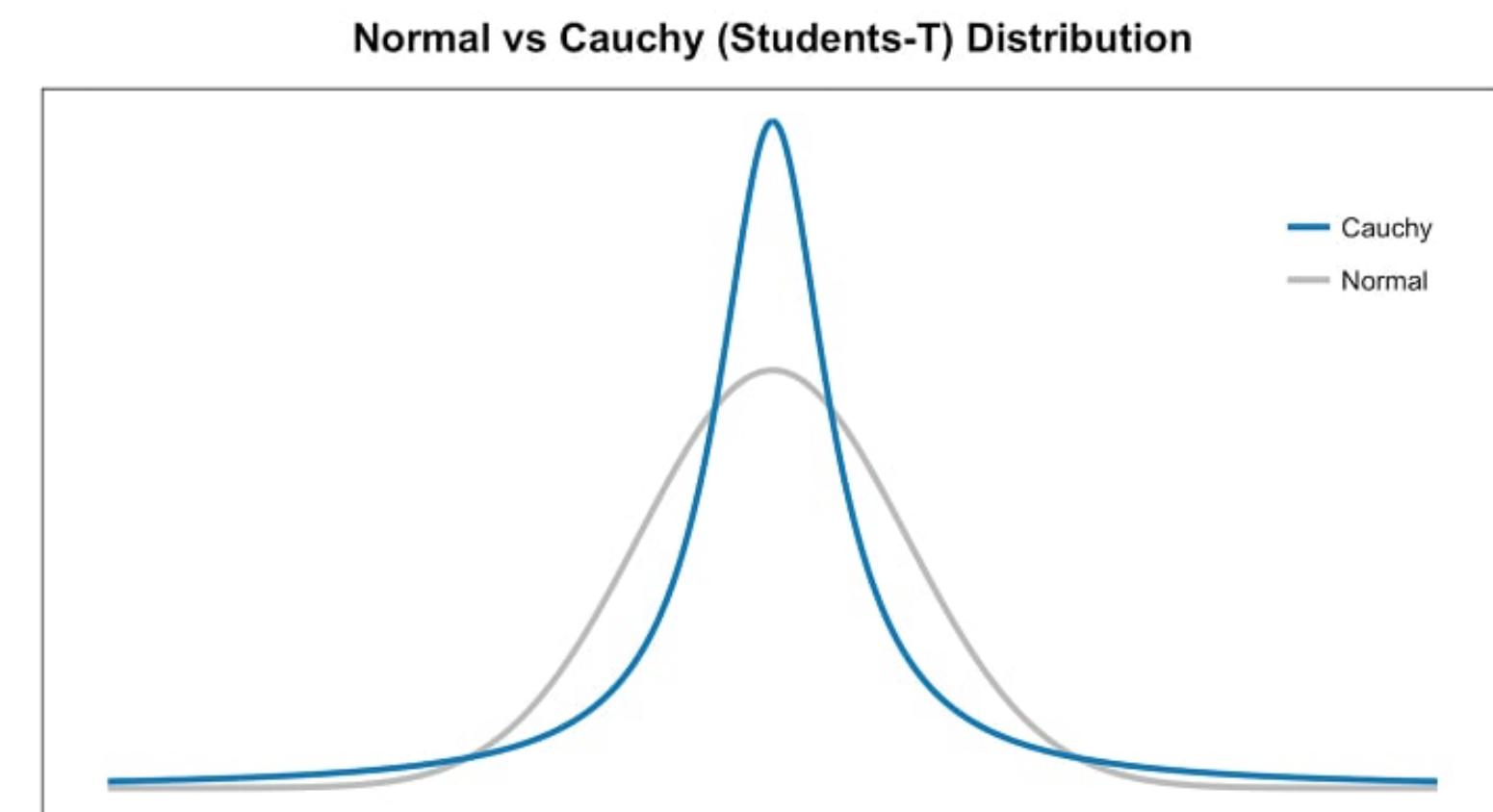
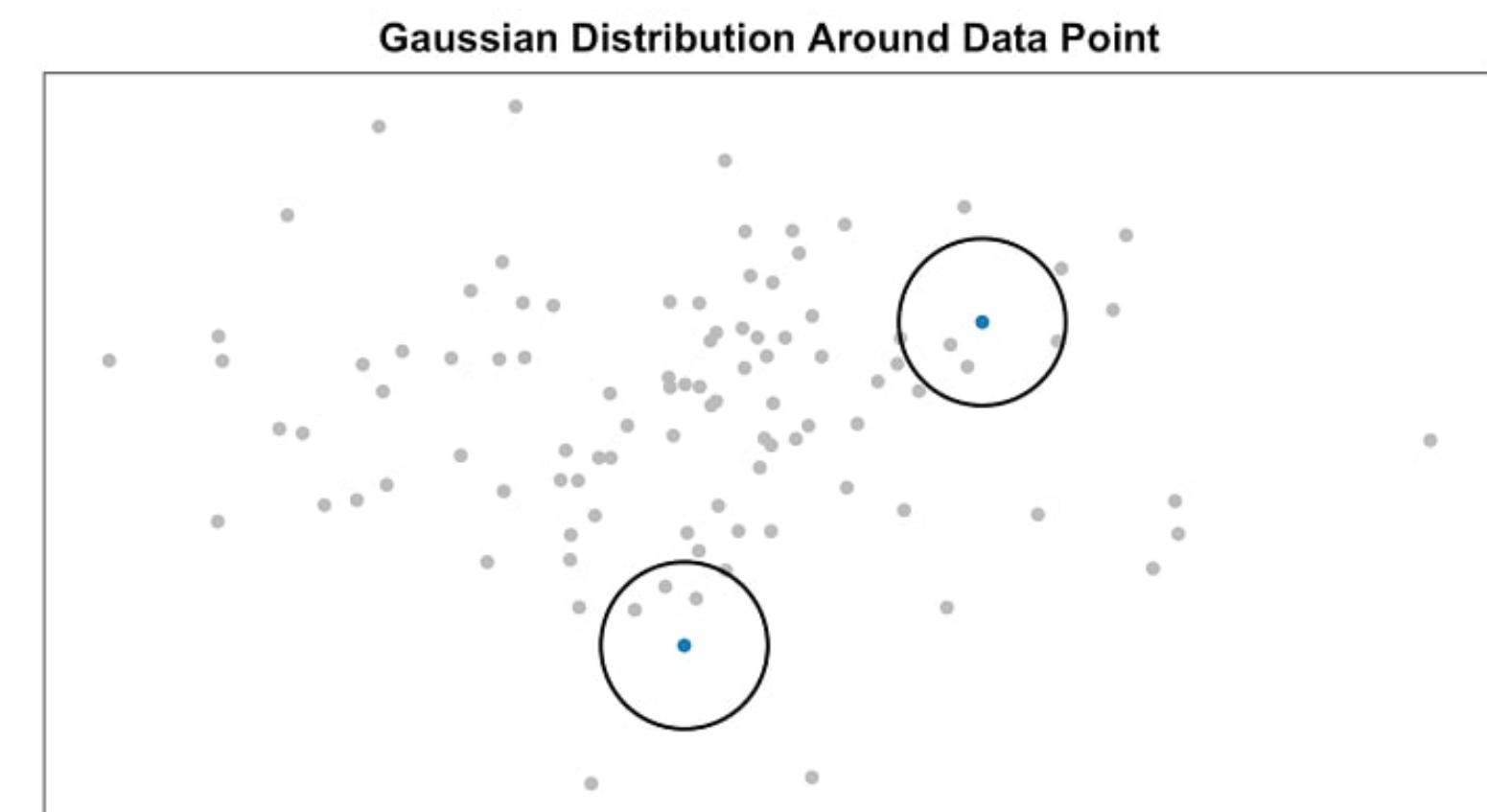
- Medir similaridades entre puntos en el espacio dimensión alto utilizando distribución Gaussiana.

- **Paso 2**

- Utilizar distribución t-Student en una dimensión mas pequeña

- **Paso 3**

- Comparar ambos conjuntos de probabilidades utilizando Kullback-Liebler divergence (KL).
- Finalmente utilizamos descenso del gradiente para minimizar la función de coste KL



# Reglas de asociación

- Permite encontrar relaciones de asociación entre los datos
  - Análisis de la cesta de la compra (*market-basket analysis*). Ejemplo: cerveza y pañales (ó leyenda urbana)
- Las reglas de asociación generalmente se escriben de la forma:
  - $\{\text{Antecedente}\} \Rightarrow \{\text{Consecuente}\}$
- La fuerza y el sentido de la relación se mide con diferentes indicadores:
  - el soporte: mide la fuerza de la regla.
  - la confianza:
- y la mejora de la confianza:

$$\text{conf}(\{\text{Antecedente}\} \Rightarrow \{\text{Consecuente}\}) = \frac{\text{conf}(\{\text{Antecedente, Consecuente}\})}{\text{conf}(\{\text{Antecedente}\})}$$

$$\text{conf}(\{\text{Antecedente}\} \Rightarrow \{\text{Consecuente}\}) = \frac{\text{conf}(\{\text{Antecedente, Consecuente}\})}{\text{conf}(\{\text{Antecedente}\}) \times \text{conf}(\{\text{Consecuente}\})}$$

- La técnica Apriori es la más emblemática para resolver estos problemas.

# Aspectos del Clustering

- Algoritmos de clustering
  - Partitional clustering
  - Hierarchical clustering
  - ...
- Una función de distancia (similaridad, o disimilaridad)
- Calidad del Clustering
  - Inter-clusters distance  $\Rightarrow$  maximizado
  - Intra-clusters distance  $\Rightarrow$  minimizado
- La calidad del resultado de un clustering depende del algoritmo utilizado, la función de distancia y la aplicación.

## K-means

- K-means es un algoritmo de **partitional clustering (clustering divisible)**
- Sea el conjunto de puntos (o instancias)  $D$

$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ,

donde  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ir})$  es un **vector** en un espacio real  $X \subseteq R^r$ , y  $r$  es el número de atributos (dimensiones) en los datos.

- El algoritmo  $k$ -means divide el conjunto de datos en  $k$  clusters.
  - Cada cluster tiene un **centro del cluster**, llamado **centroide**.
  - $k$  es especificado por el usuario.

## Algoritmo K-means

- Dado  $k$ , el algoritmo *k-means* funciona de la siguiente manera:
  - 1) Aleatoriamente elegir  $k$  puntos (*seeds*) para inicializar los **centroides**, centros de los clusters
  - 2) Asignar cada punto al **centroide** mas próximo
  - 3) Recalcular los **centroides** utilizando los elementos del cluster actual.
  - 4) Si no se cumple el criterio de convergencia, ir a 2).

## Criterio de convergencia/parada

1. nº (o minimo) reasignaciones de puntos a diferentes clusters,
2. nº (o minimo) cambio de centroides, o
3. Minimo decrecimiento en la **suma de errores cuadráticos (SSE)**,

$$SSE = \sum_{j=1}^k \sum_{\mathbf{x} \in C_j} dist(\mathbf{x}, \mathbf{m}_j)^2$$

- $C_j$  es el  $j$ -ésimo cluster,  $\mathbf{m}_j$  es el centroide del cluster  $C_j$  (el vector de medias de todos los puntos en  $C_j$ ), y  $dist(\mathbf{x}, \mathbf{m}_j)$  es la distancia entre el punto  $\mathbf{x}$  y el centroide  $\mathbf{m}_j$ .

Curso 2020

## UNSUPERVISED LEARNING ML2

---

CRISANTO DE LOS SANTOS DURÁN

[CRISANTODLS@FACULTY.MIOTI.ES](mailto:CRISANTODLS@FACULTY.MIOTI.ES)

