Análisis de Datos y

Big Data

Sesión 4 : Análisis exploratorio de datos para Big Data

Presentan:

Dr. Ulises Olivares Pinto

Joshelyn Yanori Mendoza Alfaro

René Delgado Servín

BLOQUE

Innovación, Tecnología y Creatividad.

Contenido

1. Aplicación y ejercicios de análisis exploratorio de datos

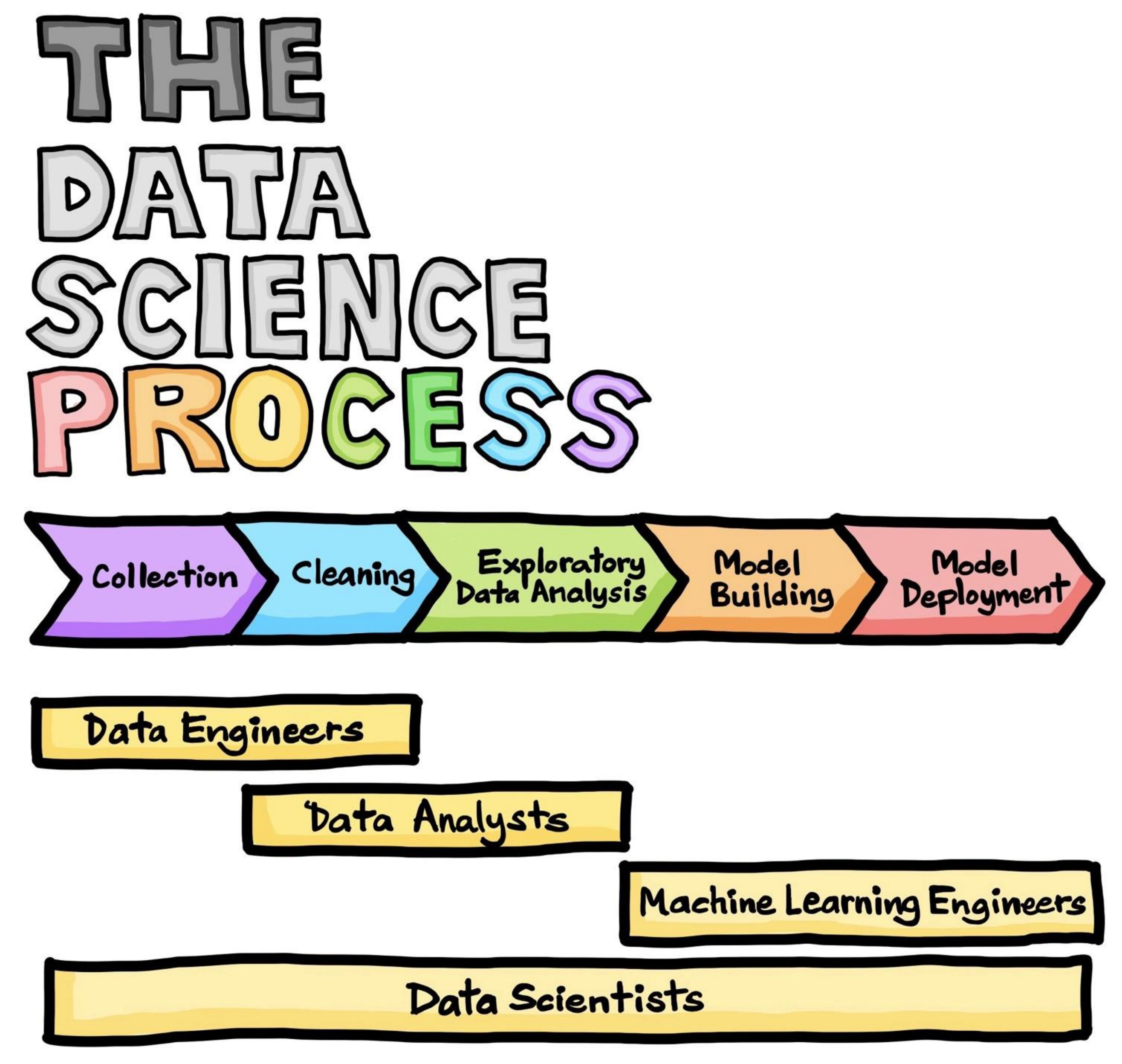
2. Reducción de dimensionalidad

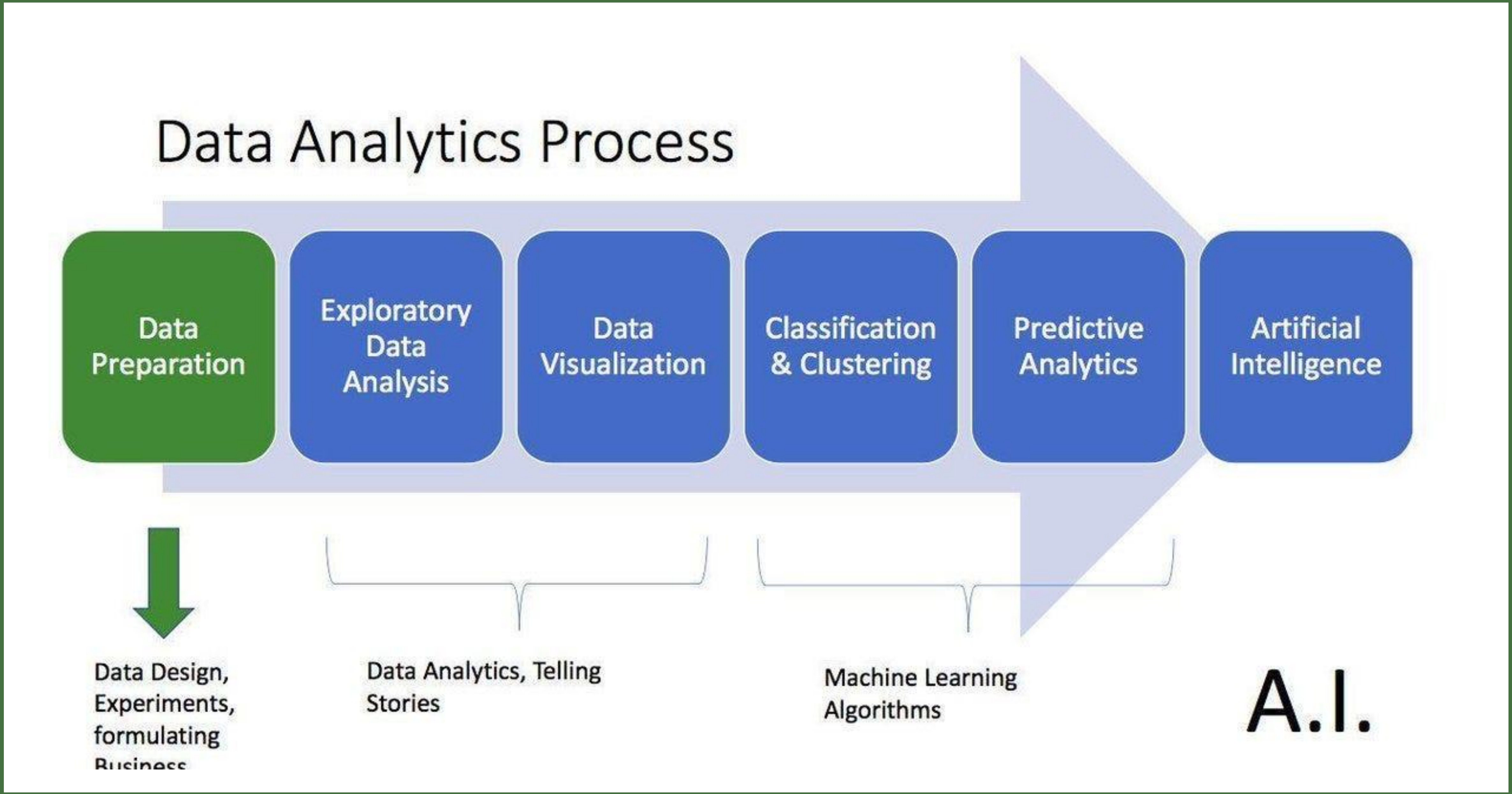
3. Procesamiento con Hadoop

1. Aplicación de análisis exploratorio de datos

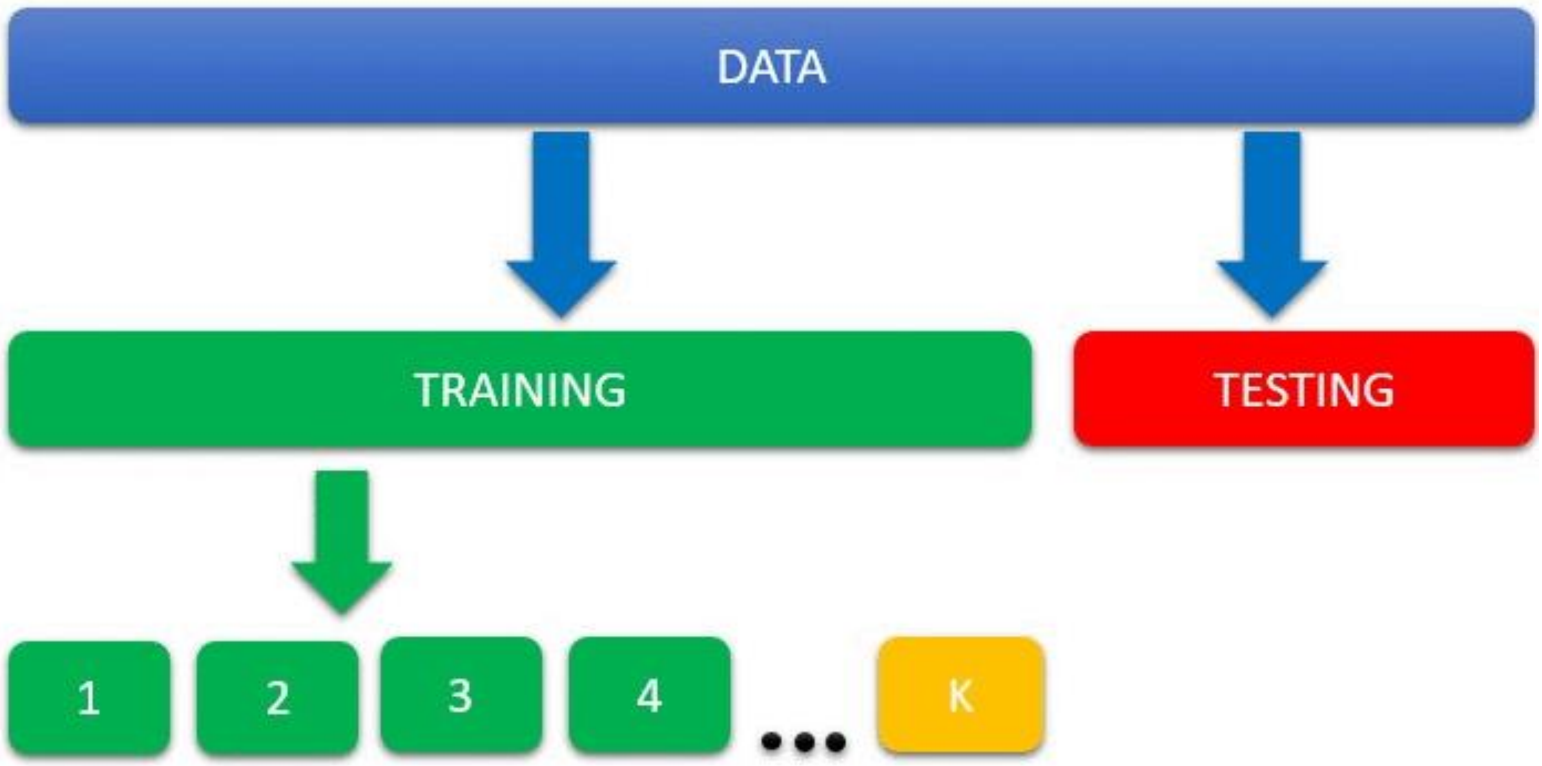
Casos de aplicación

Titanic

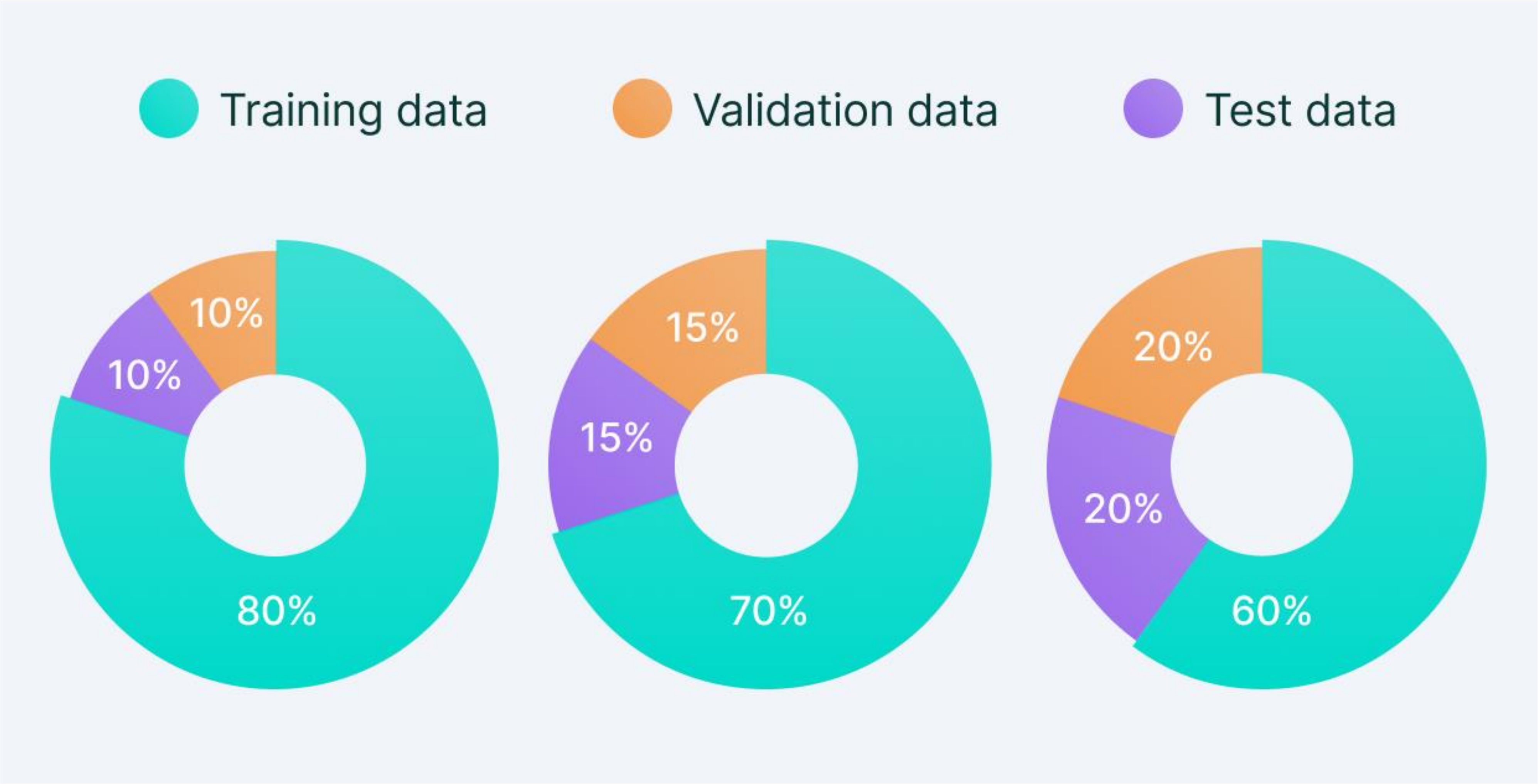

THE
DATA
SCIENCE
PROCESS
Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment
Data Engineers
Data Analysts
Machine Learning Engineers
Data Scientists


Data Analytics Process
Data Preparation
Exploratory Data Analysis
Data Visualization
Classification & Clustering
Predictive Analytics
Artificial Intelligence
Data Design, Experiments, formulating
Data Analytics, Telling Stories
Machine Learning Algorithms
A.I.

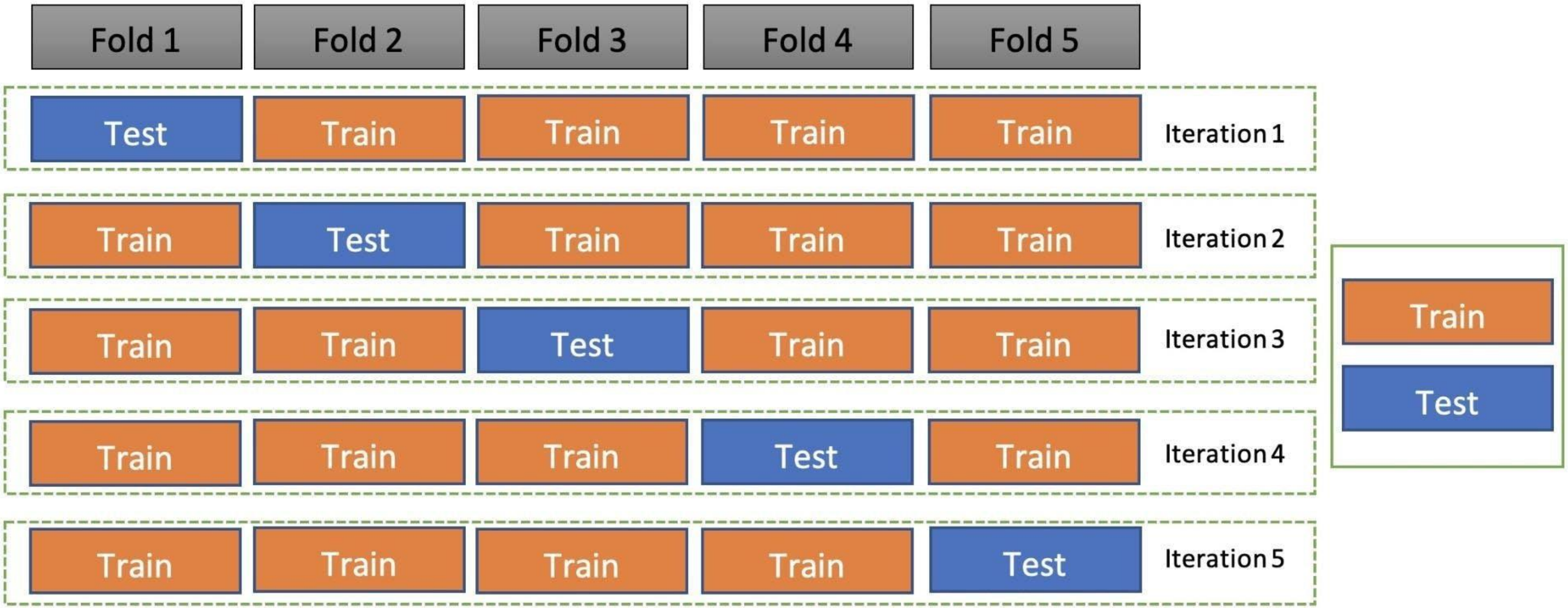
División de datos



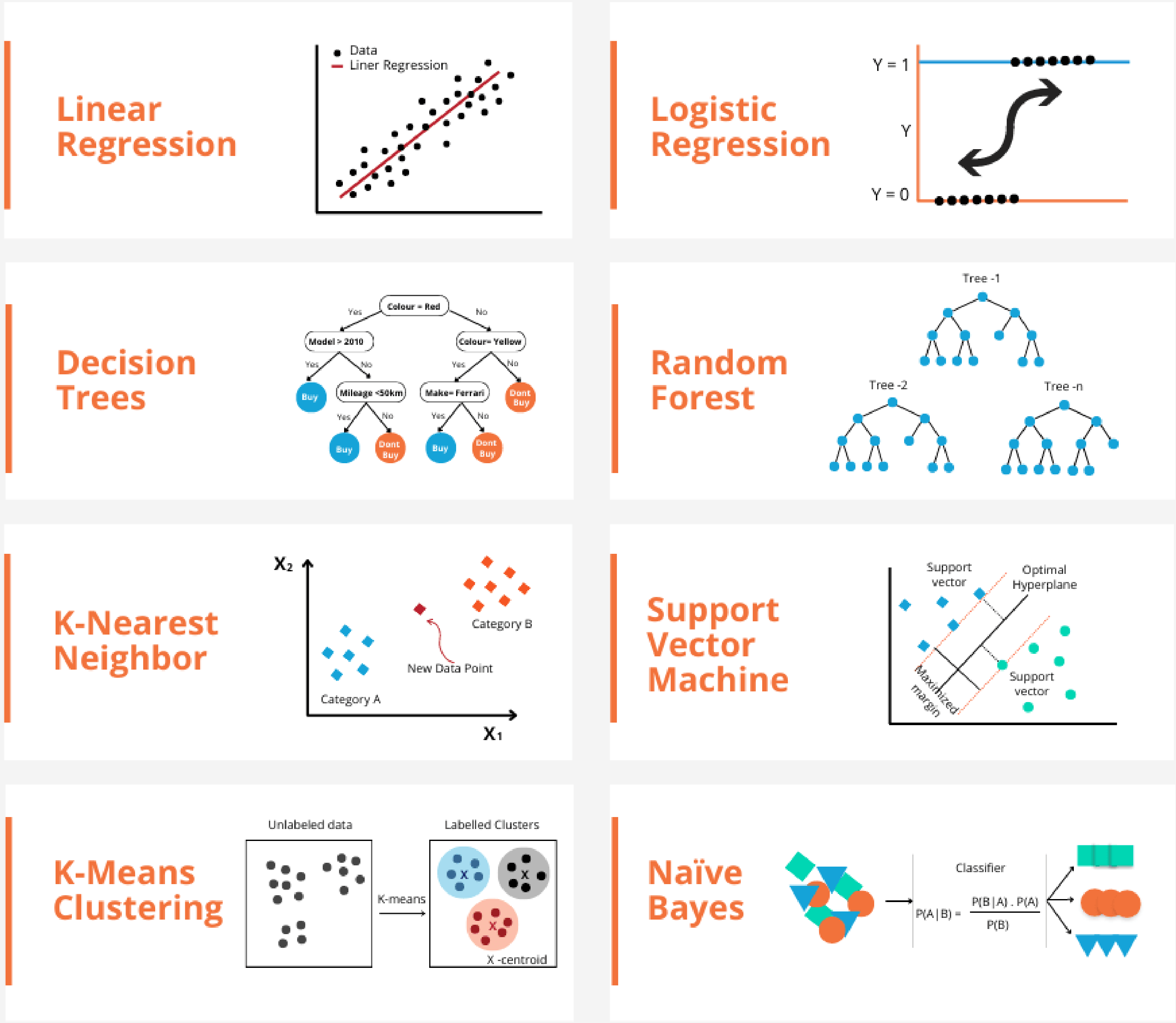
¿Cómo dividir mis datos?



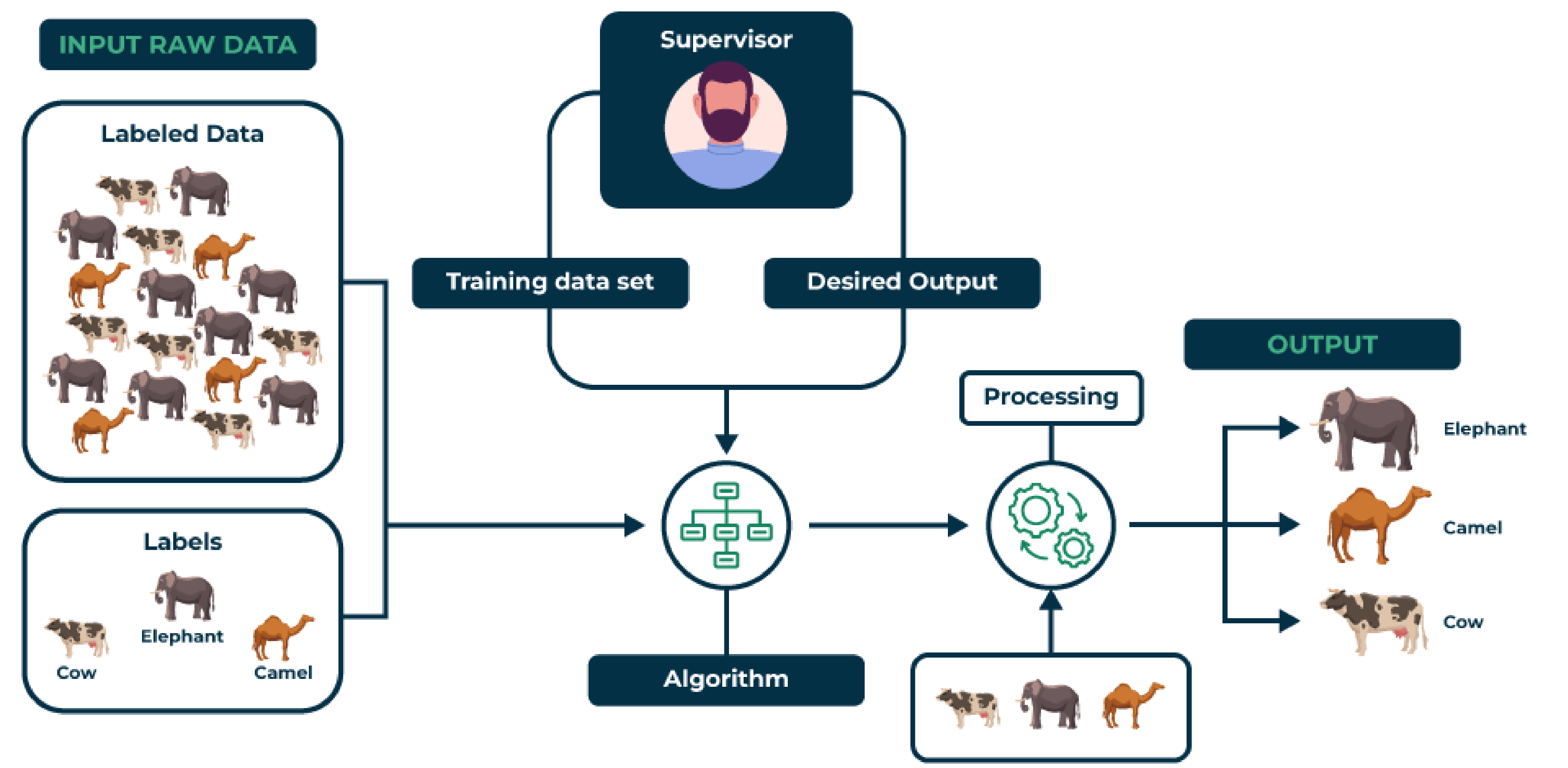
División e Iteraciones



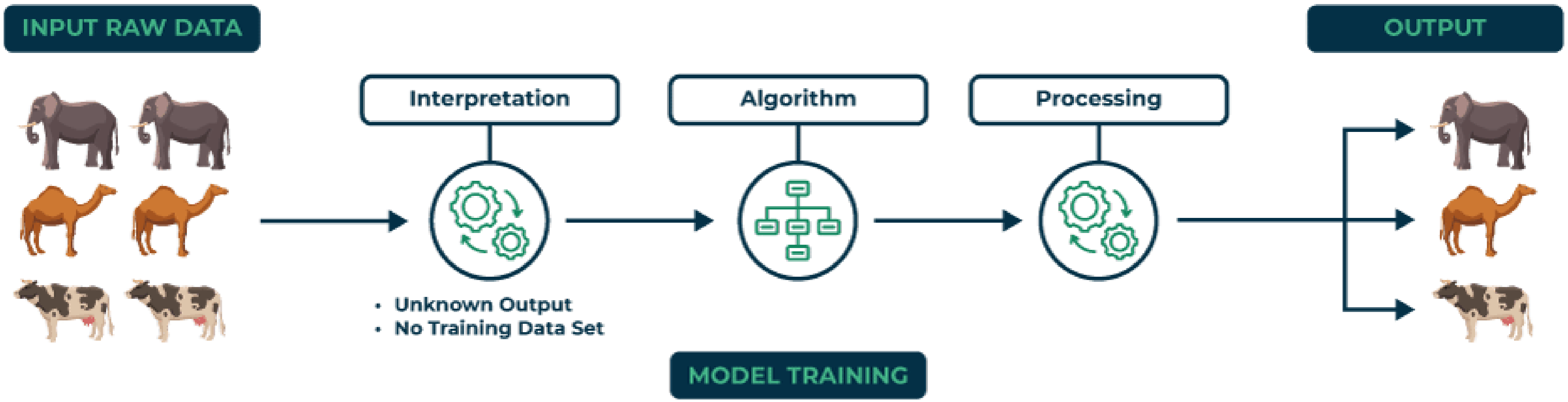
Algortmos de aprendizaje clásicos

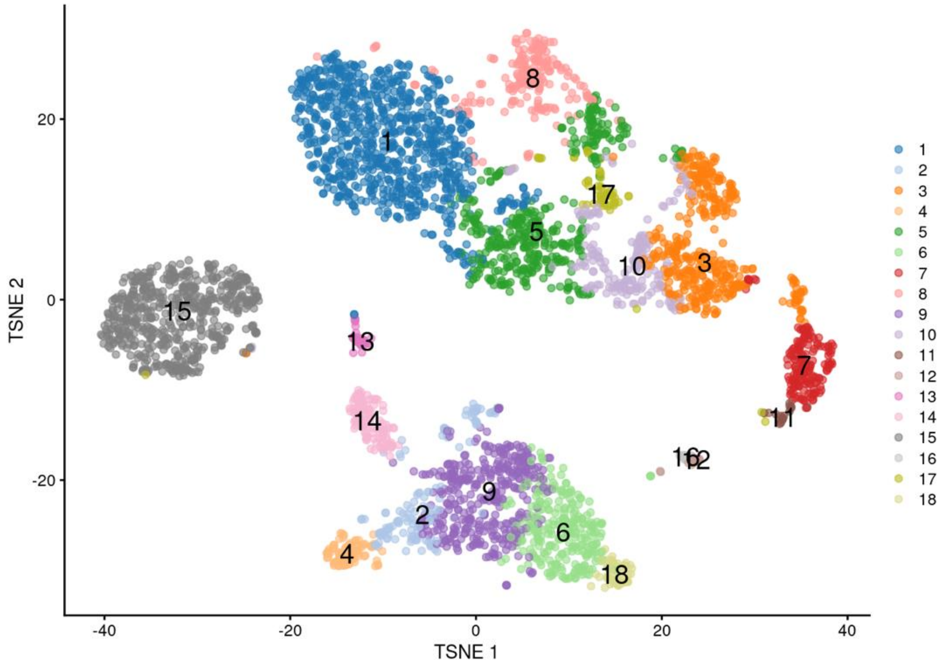


Supervised Learning

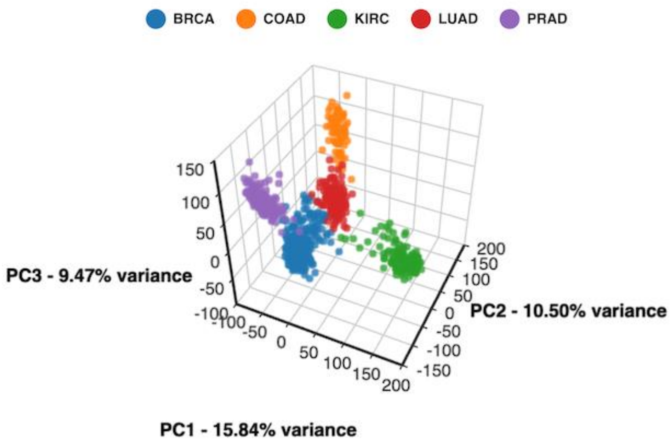


Unsupervised Learning



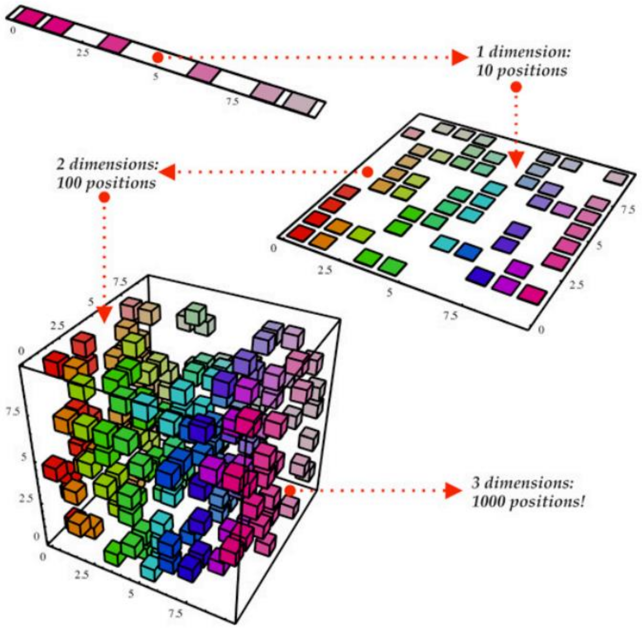


2. Algoritmos de Reducción de Dimensiones



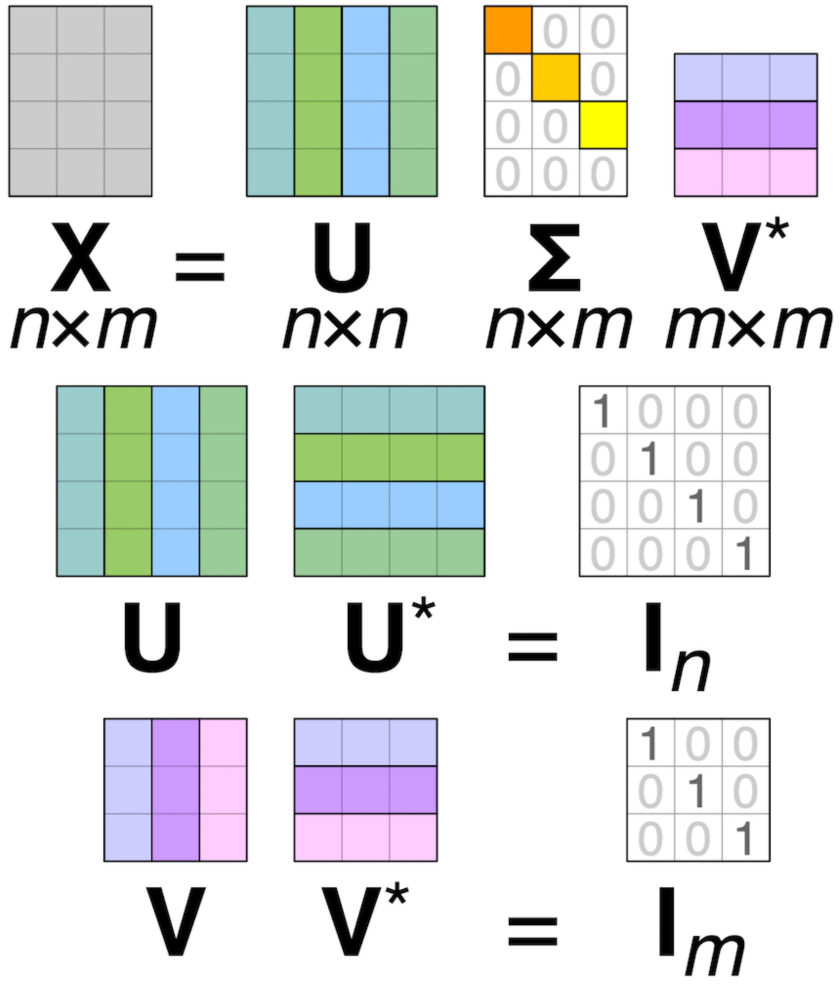
Conceptos, técnicas y aplicaciones prácticas

¿Qué es la Disminución de de Dimensiones?



Definición e Importancia

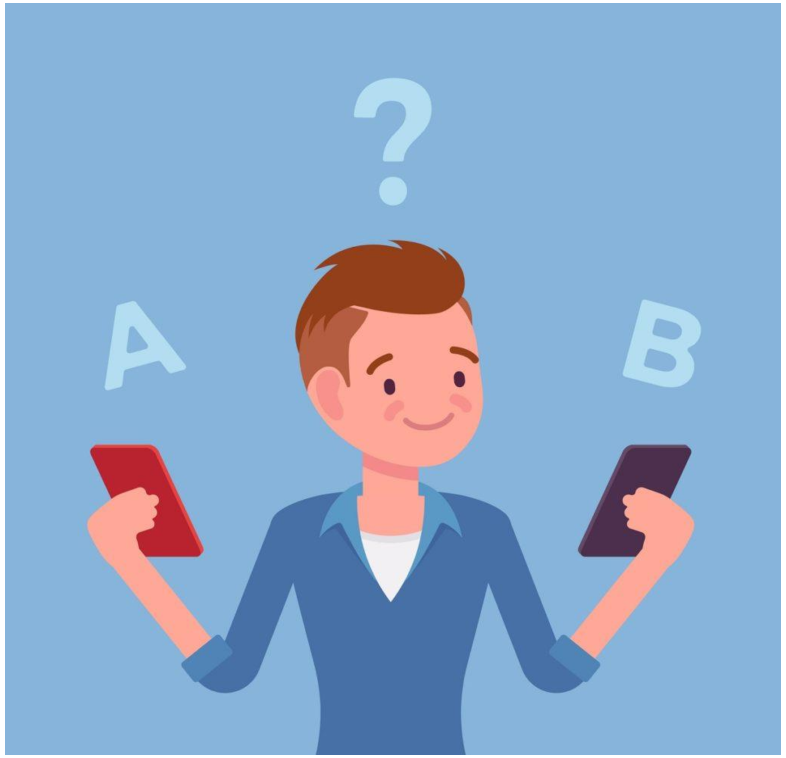
- Definición: Proceso de reducir el número de variables aleatorias bajo consideración, obteniendo un conjunto de de variables principales.
- Importancia: Facilita la visualización, reduce el costo computacional y mejora la eficiencia de los algoritmos de de aprendizaje automático.
- Aplicaciones: Desde la visualización de datos hasta el preprocesamiento en modelos de machine learning.



Principales Algoritmos de de Disminución de Dimensiones

PCA, SVD Truncado, Proyección Aleatoria Aleatoria

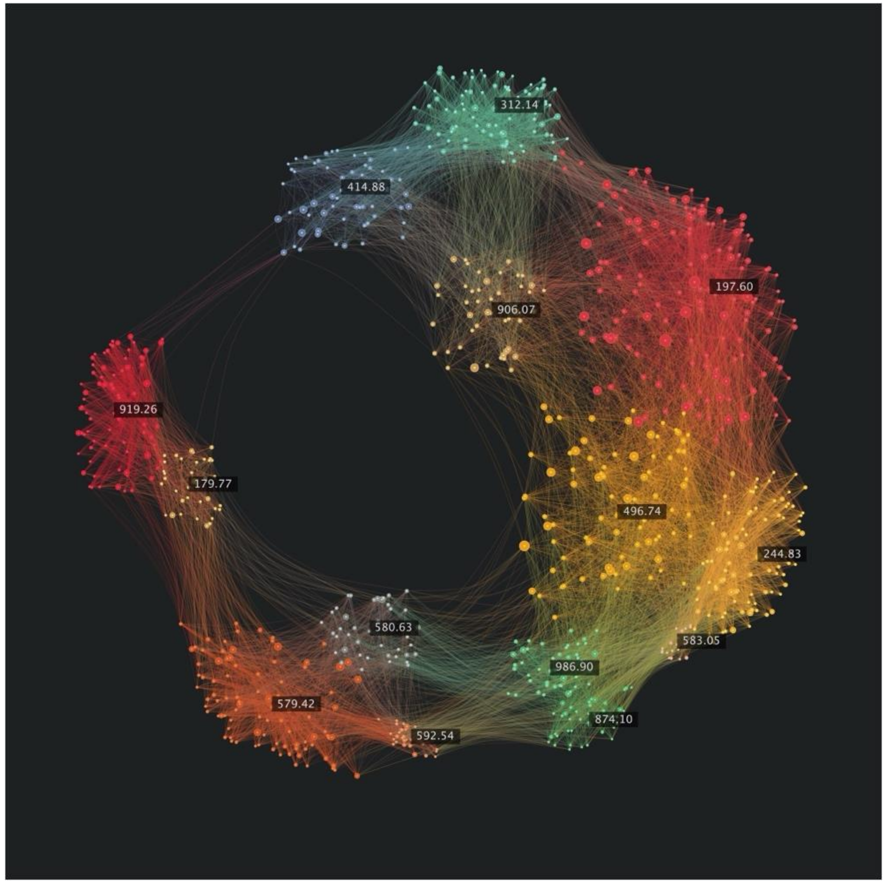
- PCA: Transforma variables originales en nuevas variables no variables no correlacionadas, maximizando la varianza retenida retenida en cada componente.
- SVD Truncado: Similar a PCA, aplicado a matrices dispersas, dispersas, descomponiendo una matriz en tres matrices más matrices más pequeñas.
- Proyección Aleatoria: Técnica no lineal que proyecta datos en datos en un espacio de menor dimensión usando matrices matrices aleatorias.
- Comparativa: Precisión, escalabilidad y facilidad de interpretación entre los tres algoritmos.



Comparativa entre Algoritmos de Disminución de Dimensiones

PCA vs SVD vs Proyección Aleatoria

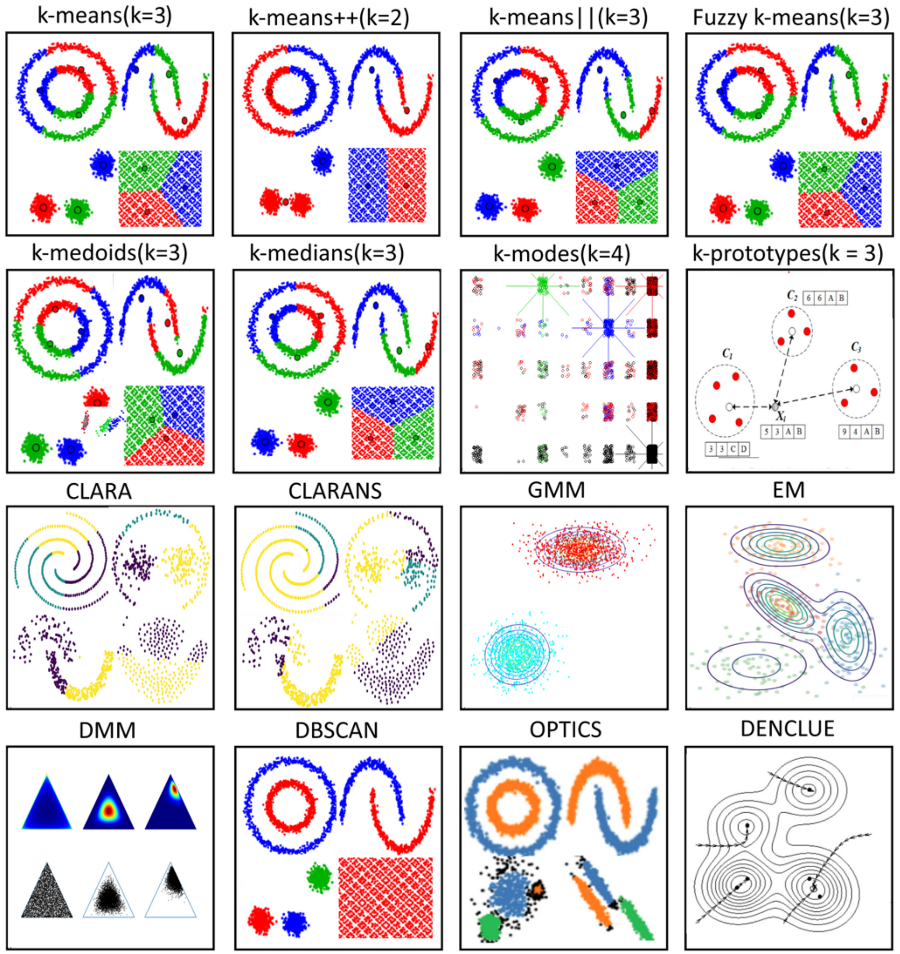
- Precisión: PCA > SVD > Proyección Aleatoria.
- Escalabilidad: Proyección Aleatoria > SVD > PCA.
- Facilidad de Interpretación: PCA > SVD > Proyección Aleatoria.



¿Qué es el Clustering?

Definición e Importancia

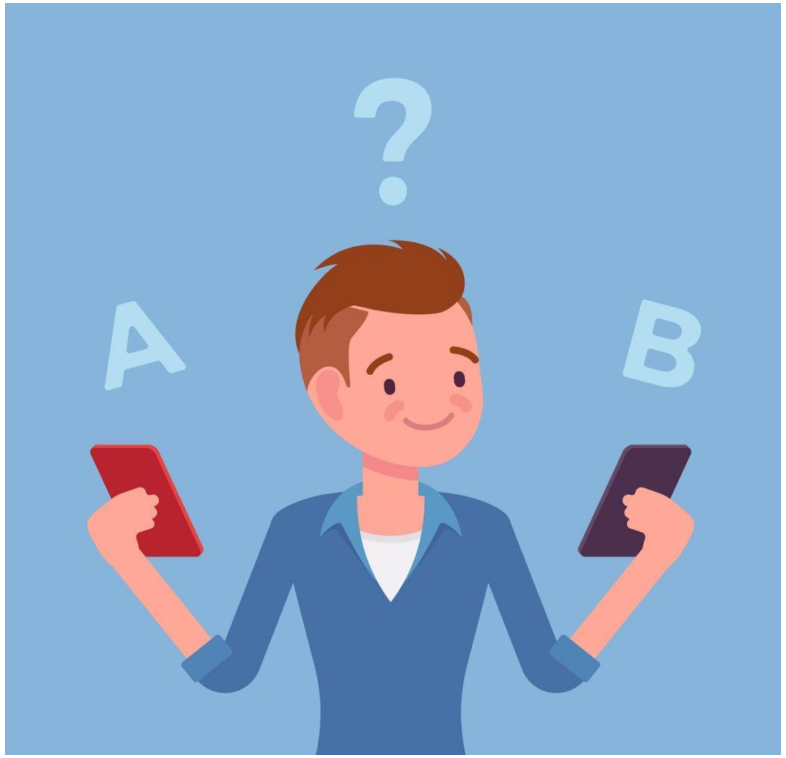
- Definición: Técnica de aprendizaje no supervisado que agrupa datos en clusters, donde los objetos en el mismo cluster son más similares entre sí que a los de otros clusters.
- Importancia: Útil para el análisis exploratorio, segmentación de clientes, detección de anomalías.
- Aplicaciones: Desde la segmentación de clientes hasta la detección de anomalías en datos de sensores.
-



Principales Algoritmos de Clustering

K-Means, Clustering Jerárquico, DBSCAN

- K-Means: Particiona los datos en k clusters, minimizando la suma de las distancias al centroide.
- Clustering Jerárquico: Construye una jerarquía de clusters utilizando una matriz de distancias, con enfoque aglomerativo (bottom-up) o divisivo (top-down).
- DBSCAN: Agrupa puntos que están densamente conectados, identificando outliers como ruido.
- Comparativa: Ventajas y limitaciones en diferentes contextos de aplicación.



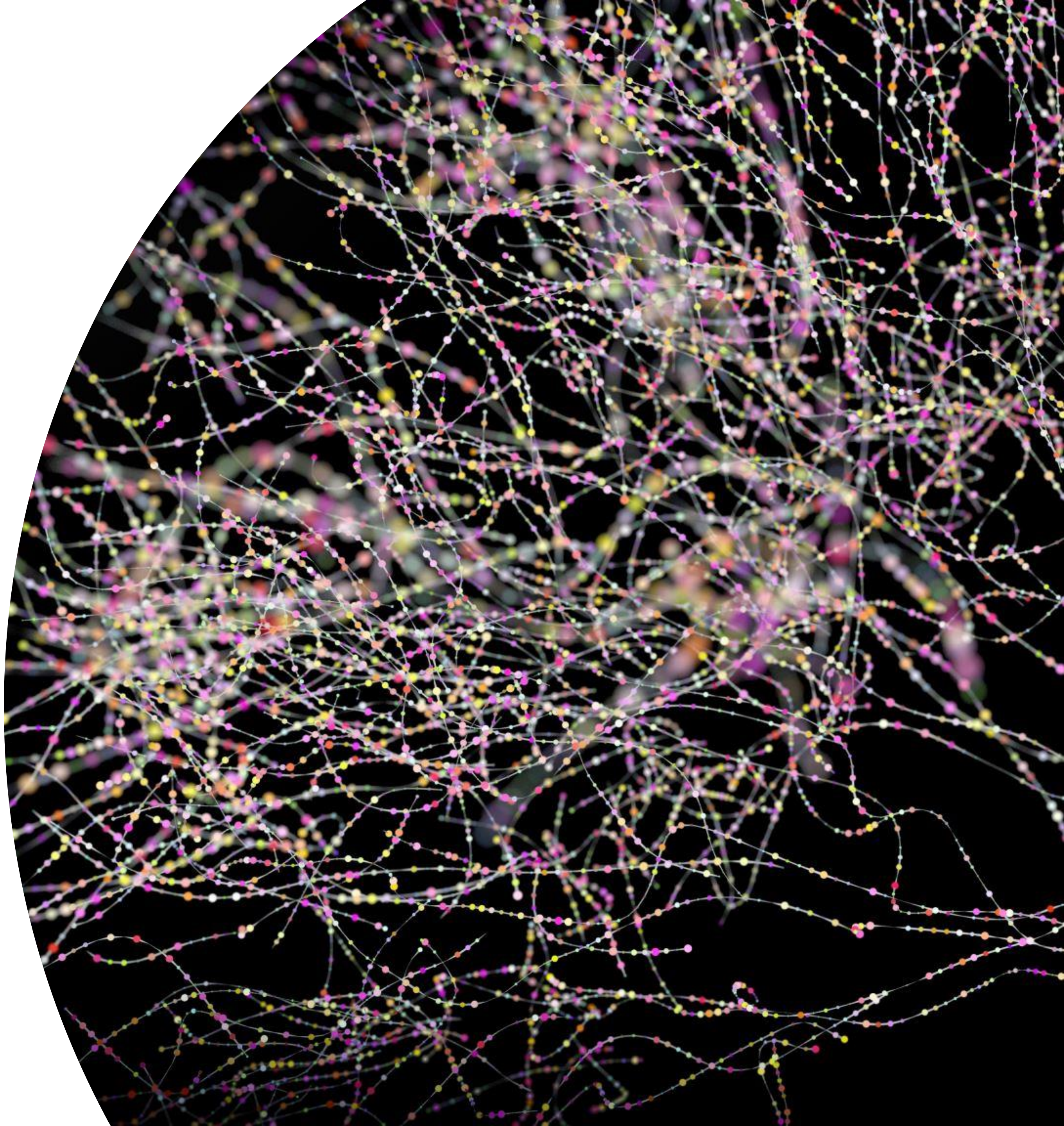
Comparativa entre Algoritmos de Clustering

K-Means vs Clustering Jerárquico vs DBSCAN

- Escalabilidad: K-Means > DBSCAN > Clustering Jerárquico.
- Detección de Outliers: DBSCAN > K-Means, Clustering Jerárquico.
- Flexibilidad en la Forma de los Clusters: DBSCAN > Clustering Jerárquico > K-Means.
- Ejemplo: Aplicación de cada algoritmo en diferentes escenarios de datasets.

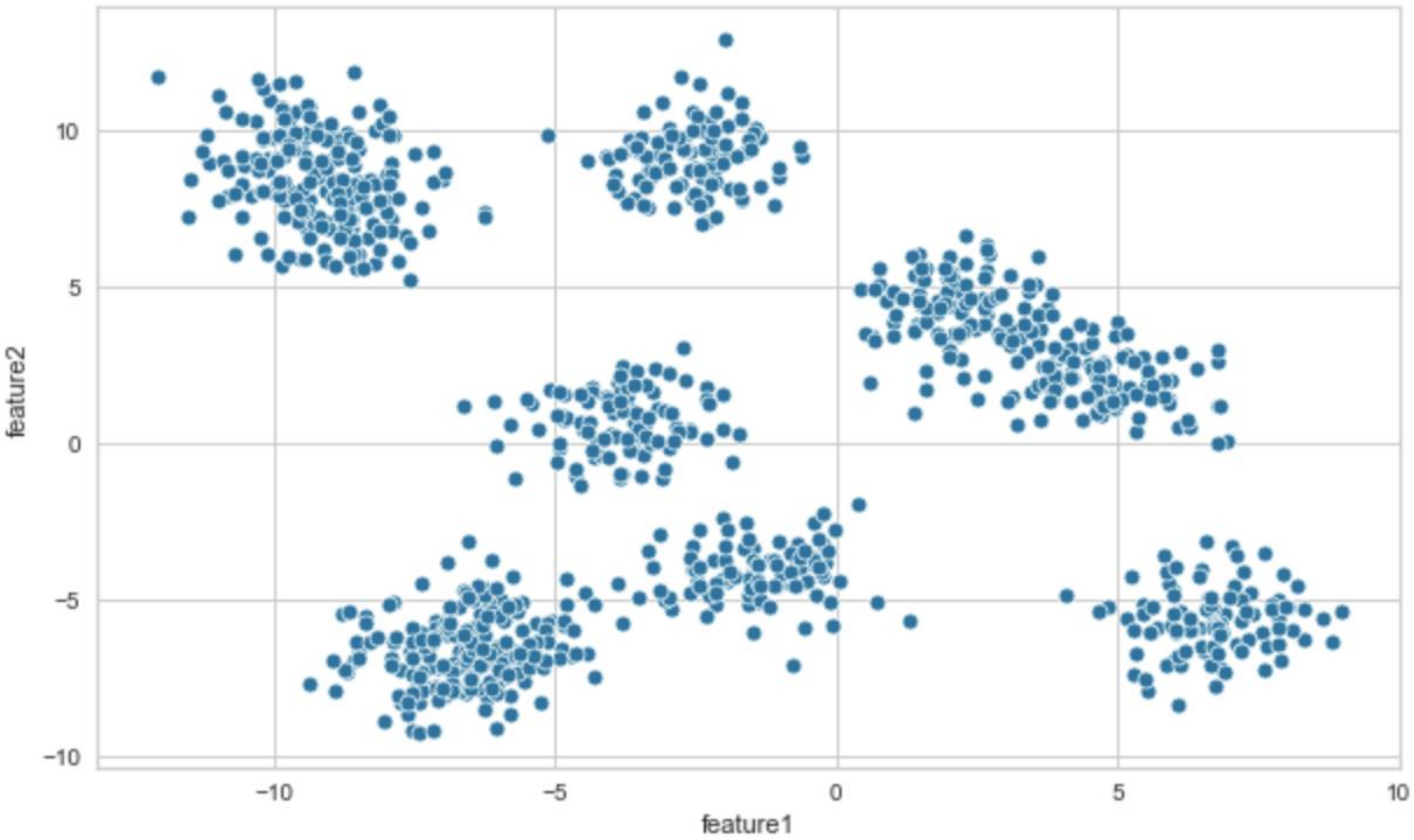
Algoritmos de agrupamiento

Kmeans



Encontrar los k clústeres que mejor describan a los datos

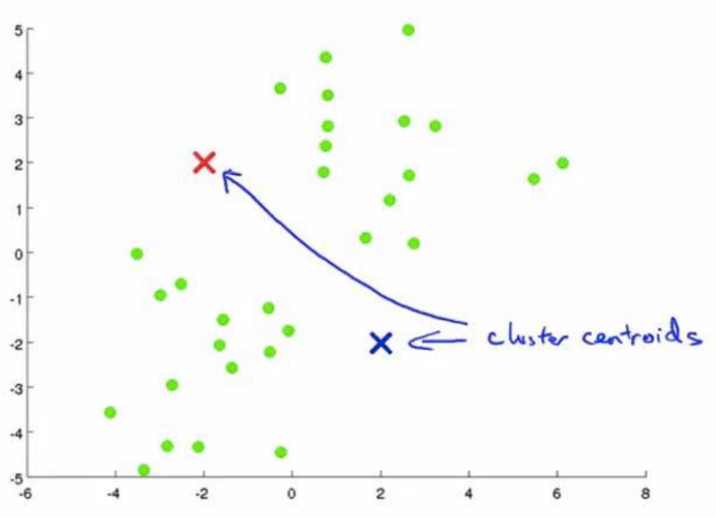
Algoritmo K-means



Algoritmo K-means

Número de clusters k = 2

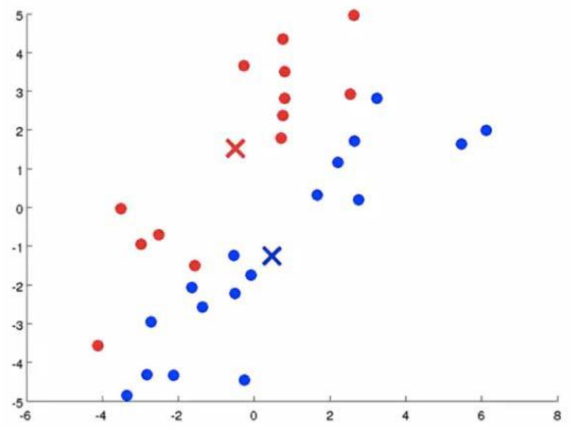
- Se inicializan los centroides de forma aleatoria

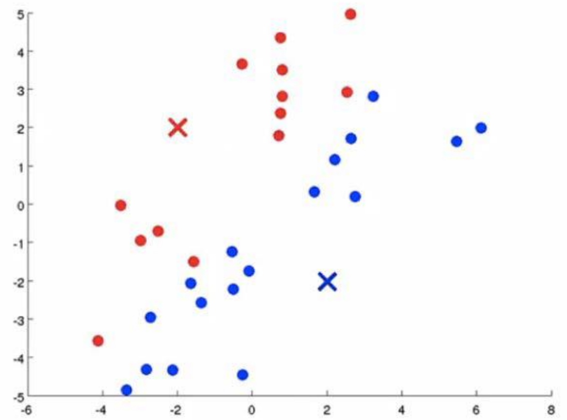


Algoritmo K-means

Número de clusters k = 2

- Asignar mebresia para cada clúster
- Actualizar el centroide de cada cluster (promedio del los puntos para cada cluster)

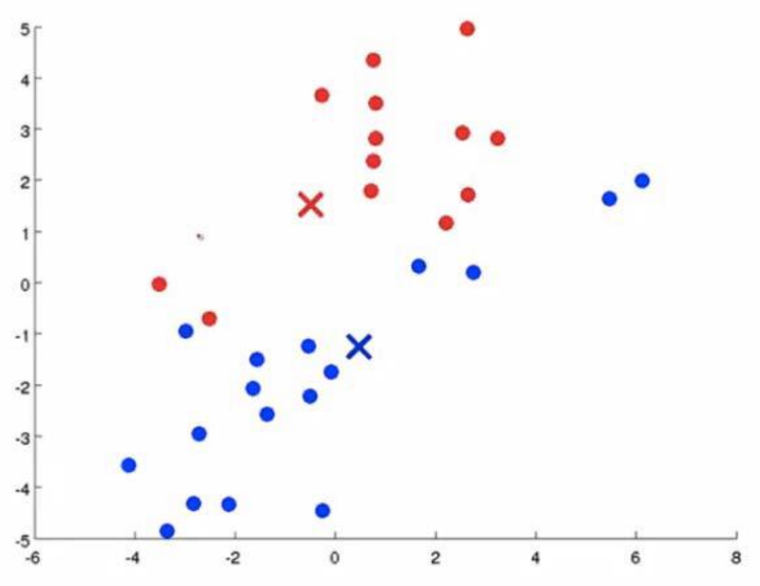




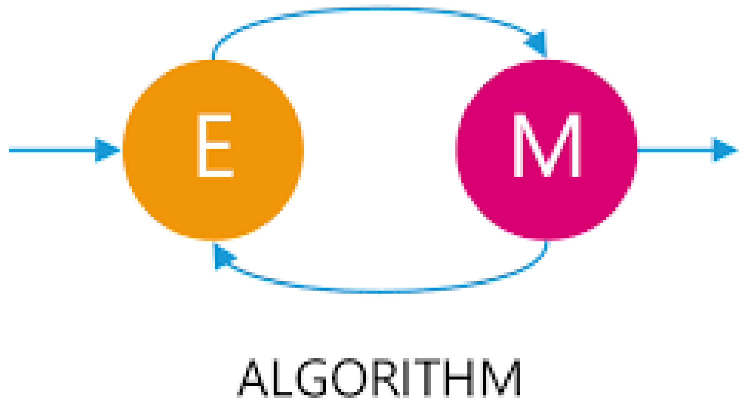
Algoritmo K-means

Número de clústeres k = 2

- Actualizar la membresía del cluster
- Actualizar los valores hasta que no existan cambios en la membresía



Expectation Maximiztion

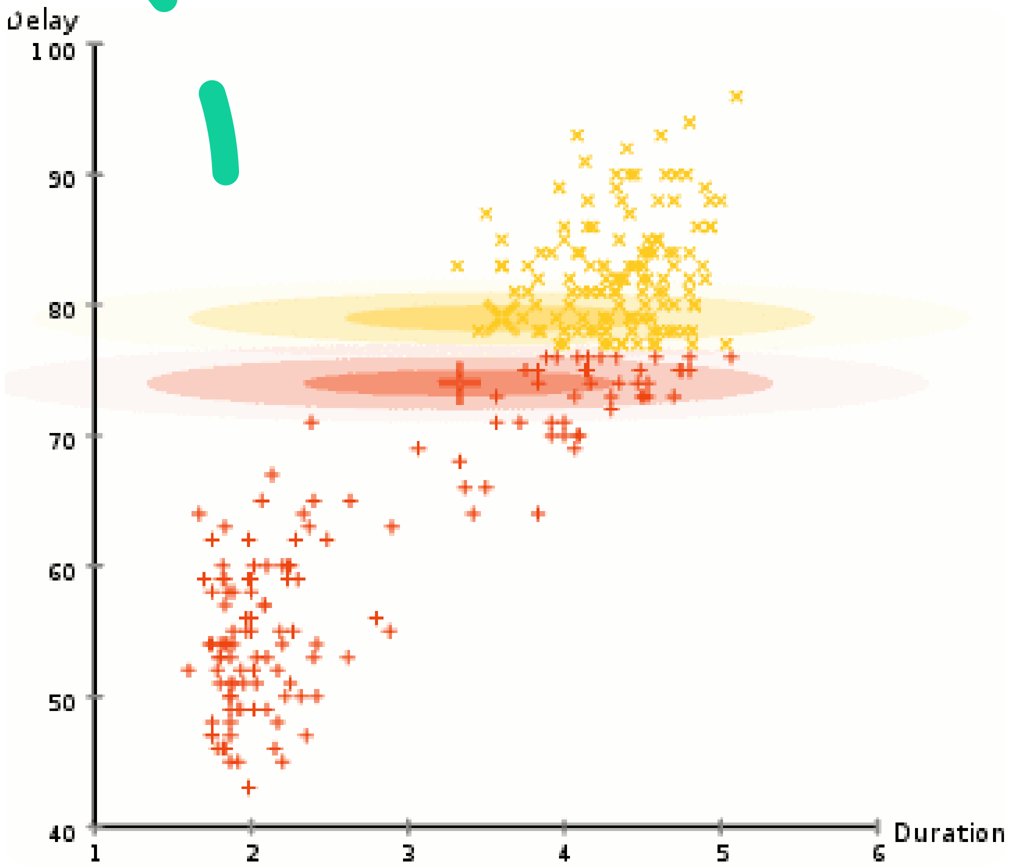


Expectation Maximization (EM)

- Es un algoritmo poderoso que se puede emplear en múltiples contextos.

1. Designa puntos iniciales de centroides (aleatorios).

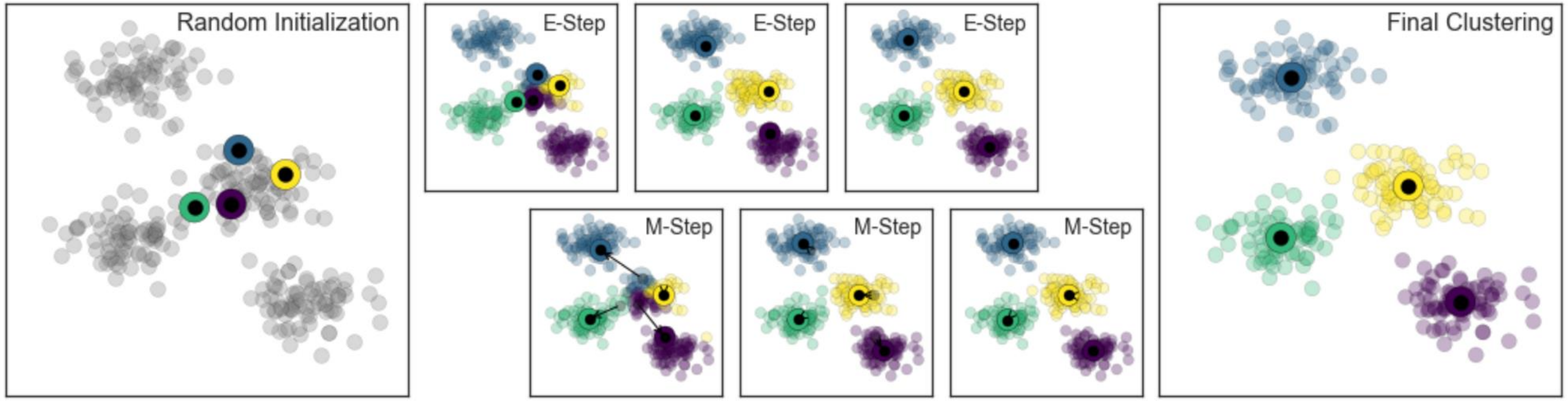
2. Este algoritmo consiste de dos partes
   - Paso E: Asignar puntos al centroide más cercano.
   - Paso M: Establecer los centroides usando promedios.



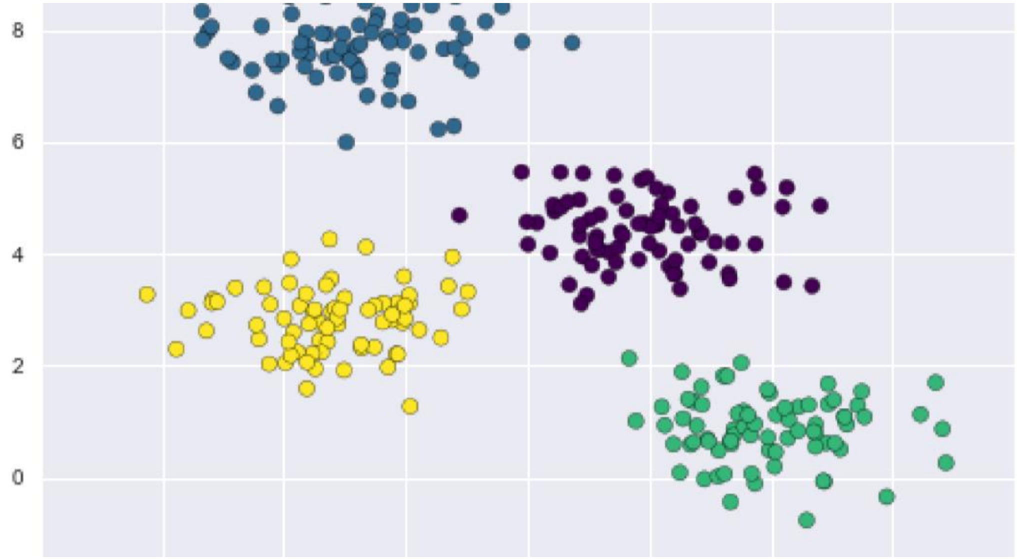
Expectation Maximization (EM)

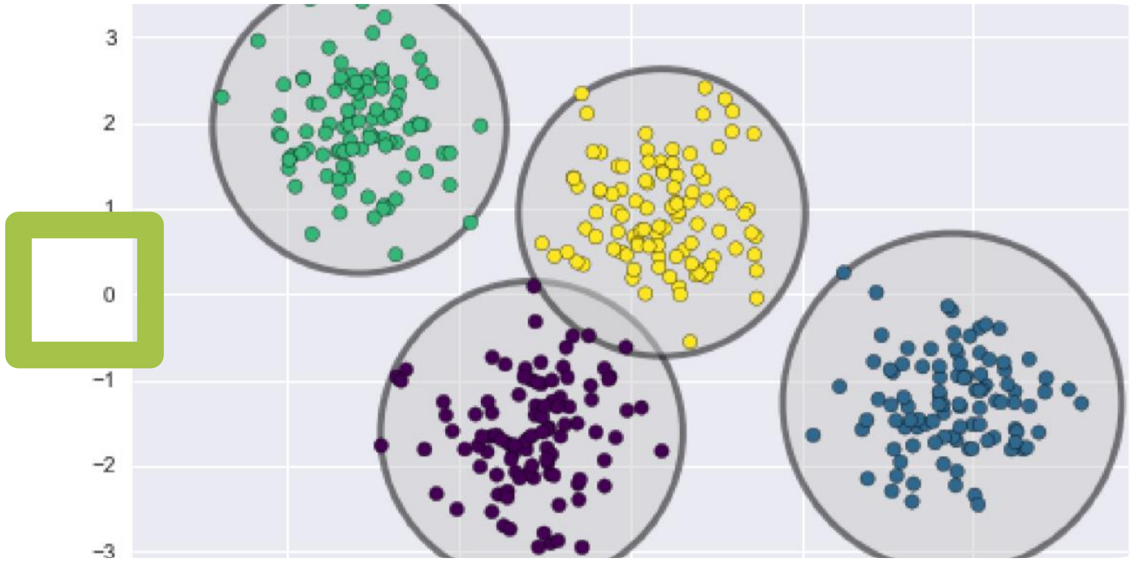
- En el paso E (Expectation), consiste en actualizar la pertenencia de un punto a un centroide.

- El paso M (Maximization) consiste en maximizar una función fitness, la cual define la ubicación de los centroides.



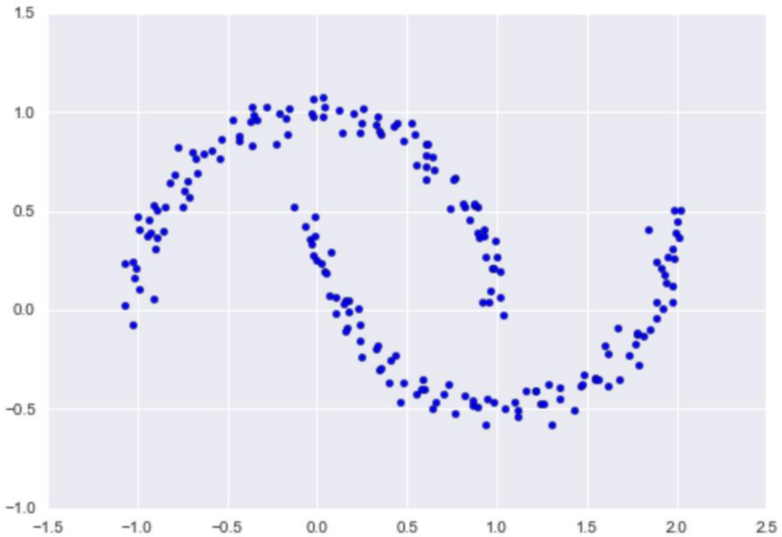
Expectation Maximization (EM)

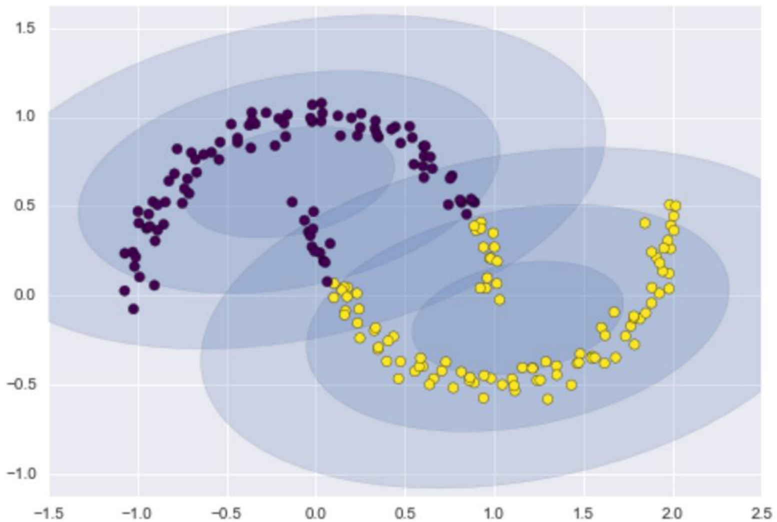


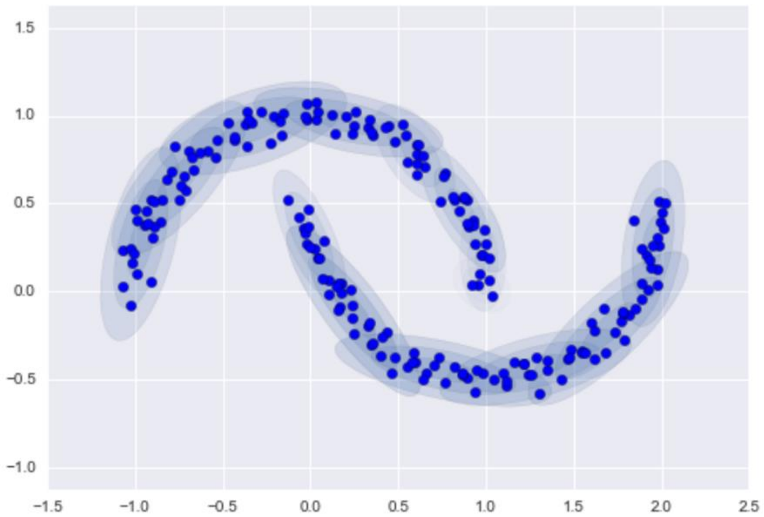


- Proceso de clusterizado

Expectation Maximization (EM)



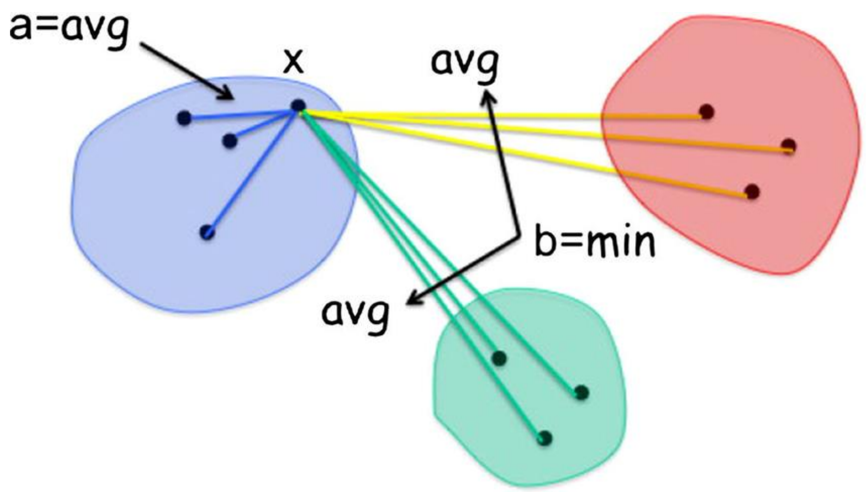




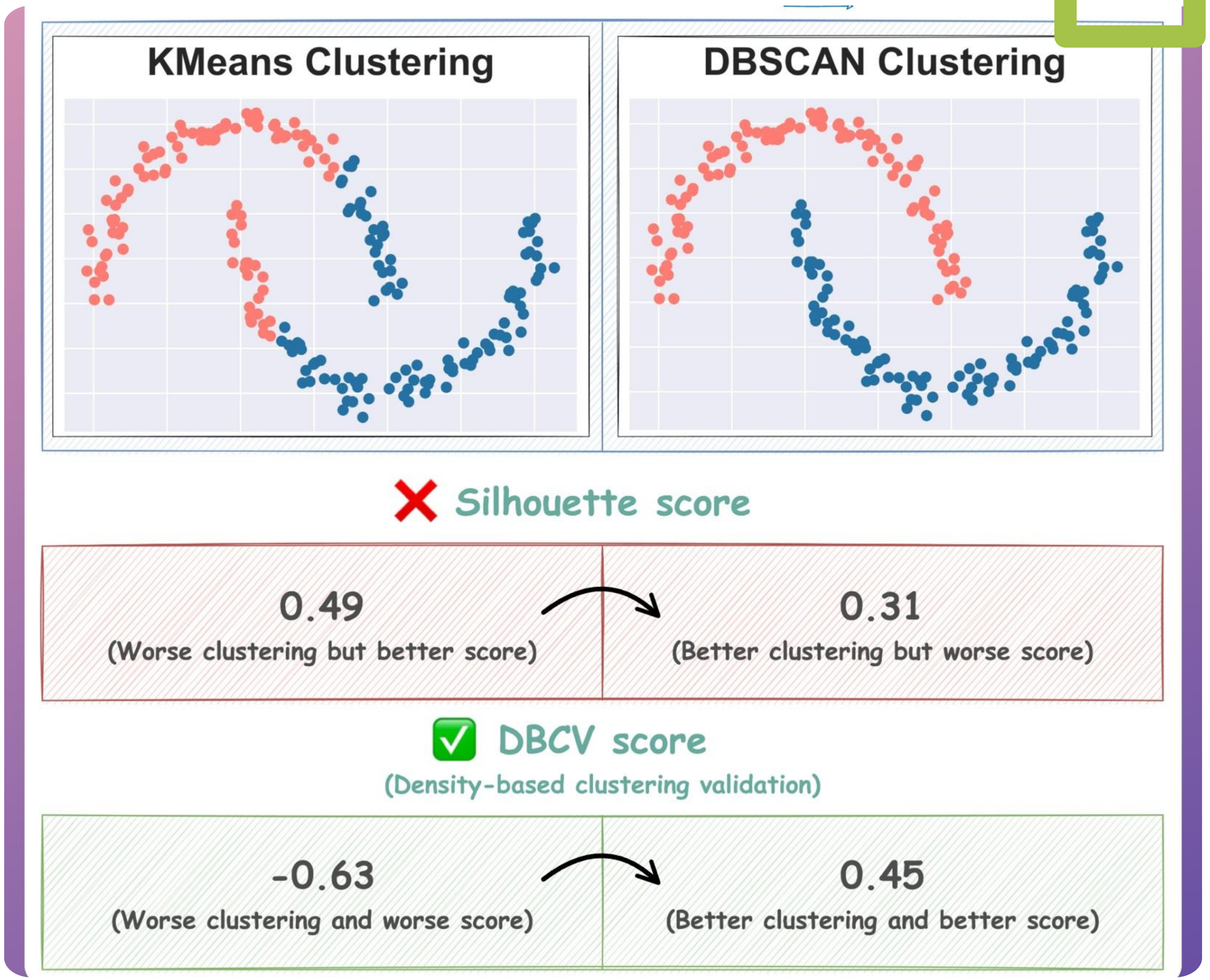
Métricas de evaluación

$$s = \frac{b - a}{\max(a, b)}$$

- El **Silhouette Score** es una métrica utilizada para evaluar la calidad de los clusters creados por un algoritmo de clustering. Esta métrica toma en cuenta tanto la cohesión dentro de los clusters como la separación entre los clusters.



¿Cómo evaluar algoritmos de clustering?



3. Hadoop y Python

Aplicaciones