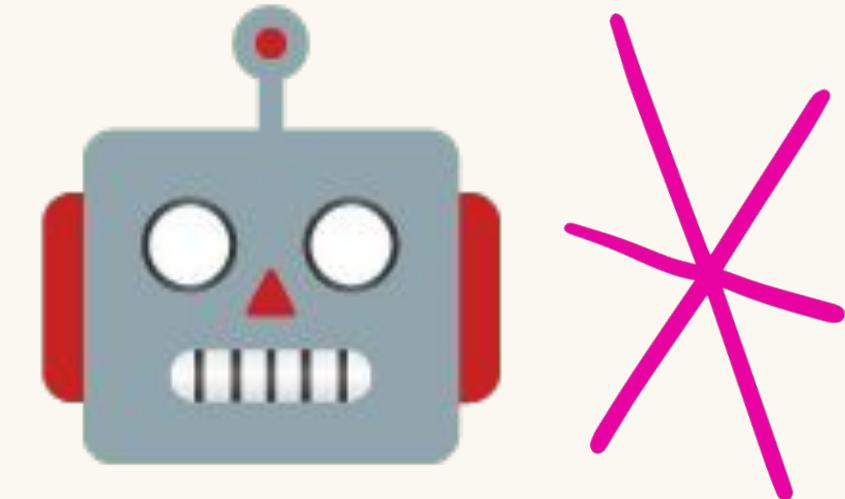


Desbloqueando el Poder de los Datos



# Inteligencia Artificial & Ciencia de Datos para todos

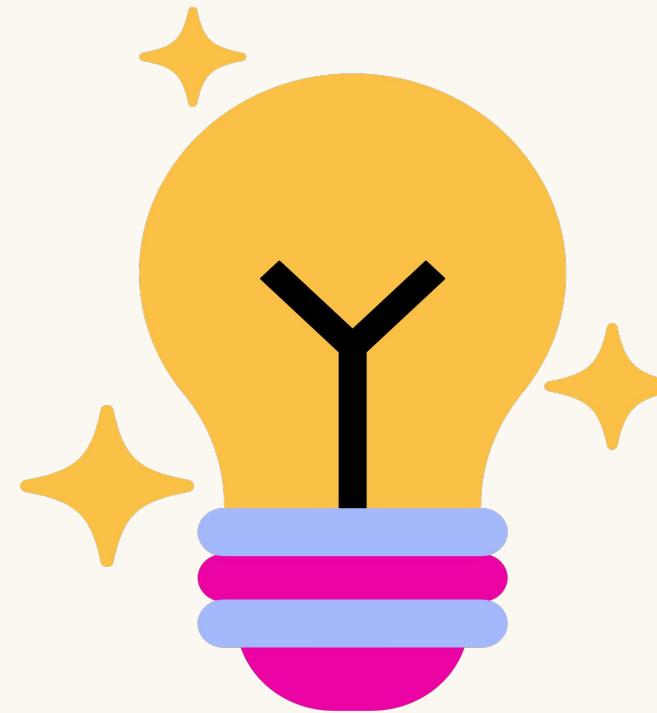


Descanzo. Regresamos a las: 8:05 a.m.

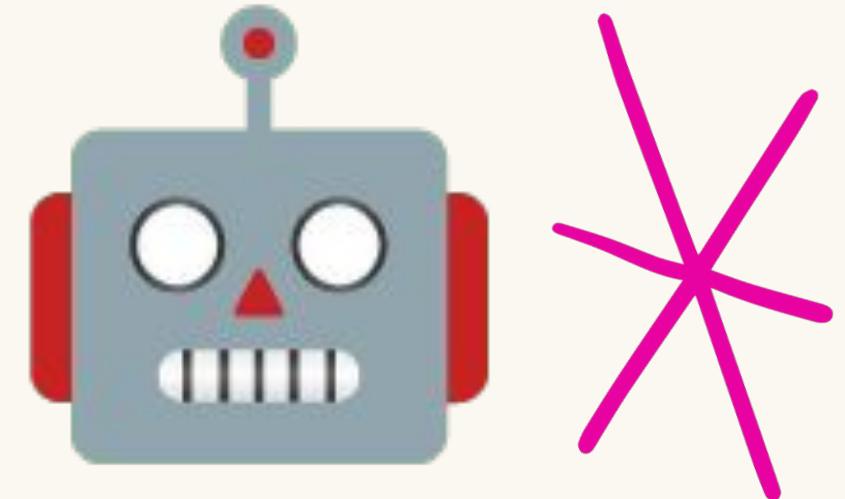
¿Te gustaría comenzar el día con alguna canción en específico?

Coméntala en el chat 

Desbloqueando el Poder de los Datos



# Inteligencia Artificial & Ciencia de Datos para todos



Comenzamos a las 7:05 a.m. en punto.

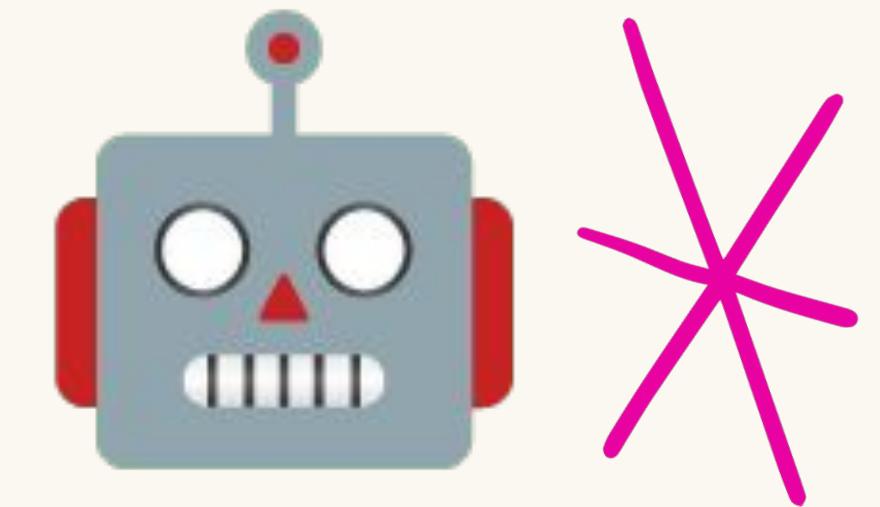
¿Te gustaría comenzar el día con alguna canción en específico?

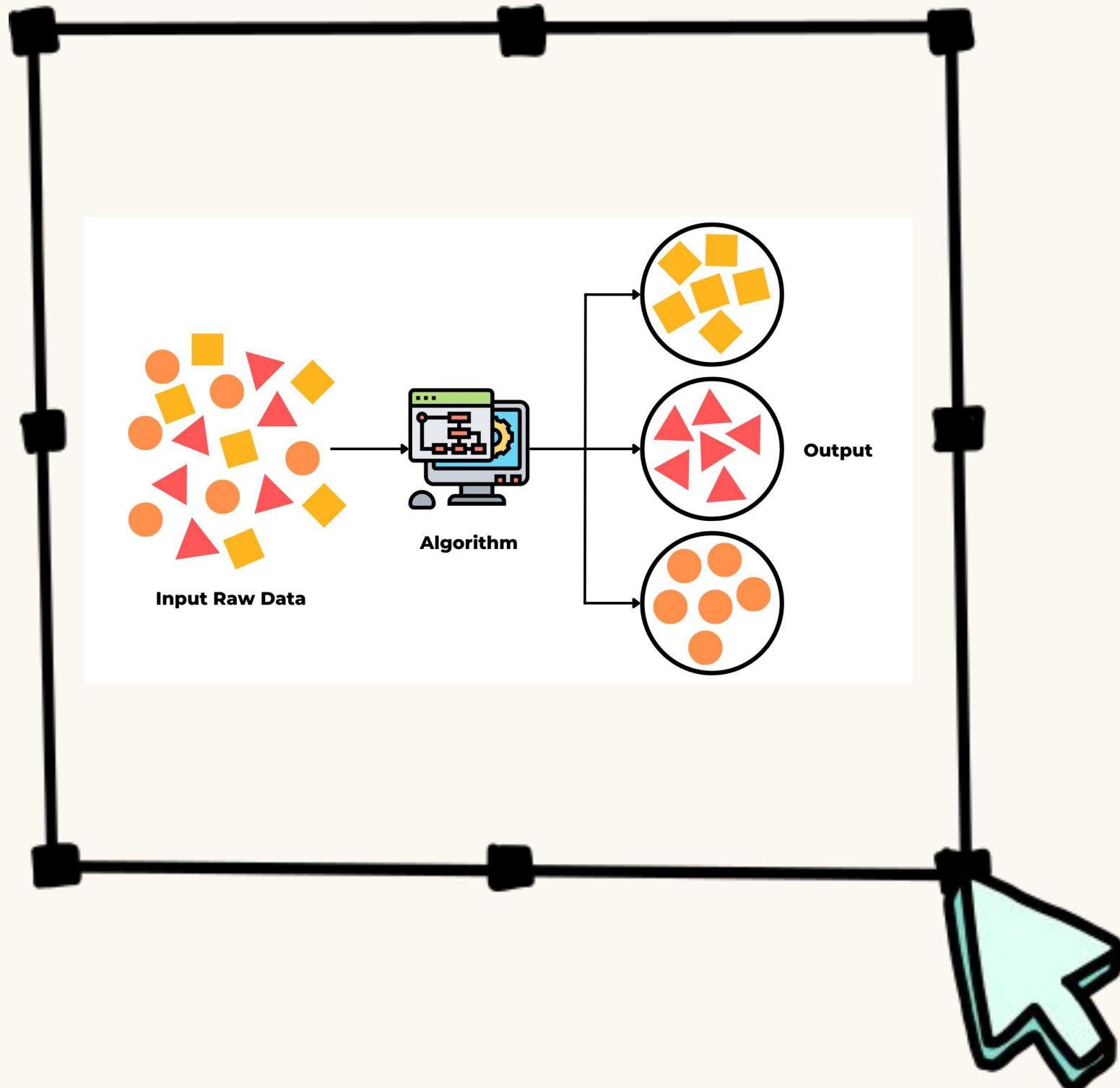
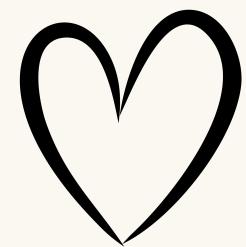
Coméntala en el chat 



Desbloqueando el Poder de los Datos

# Inteligencia Artificial & Ciencia de Datos para todos





# Modelos NO supervisados

Octubre 15, 2024



# Agenda

1. Repaso y solución del taller
2. Tema de hoy:
  - Modelos NO supervisados
    - K-Medias

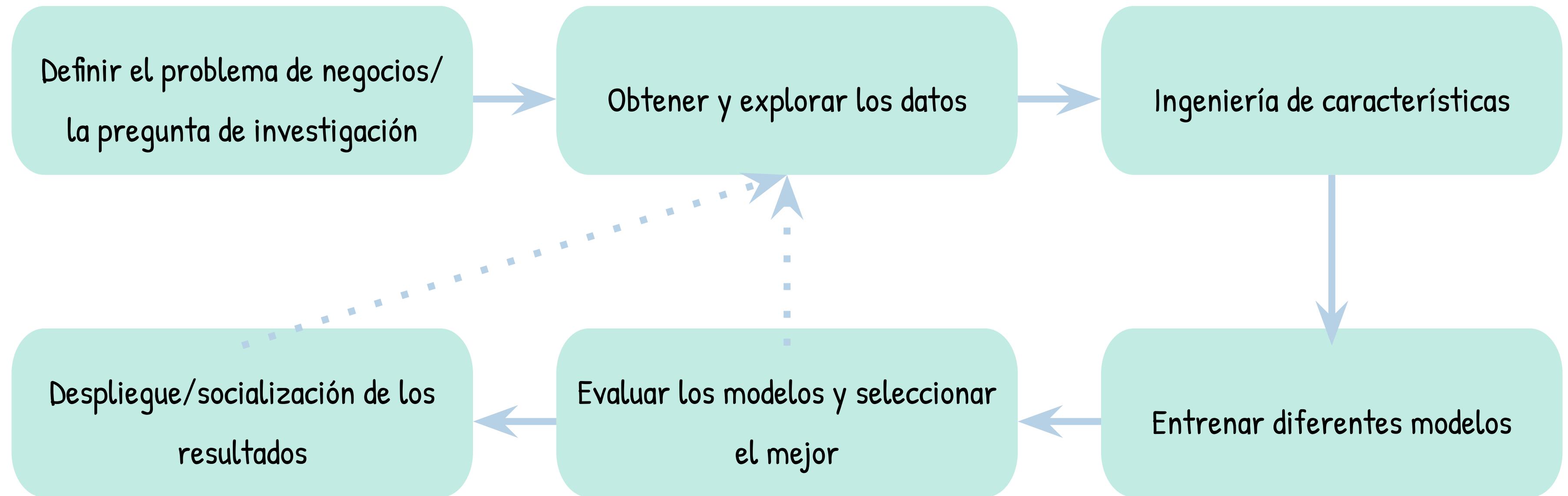


## Solución taller 10

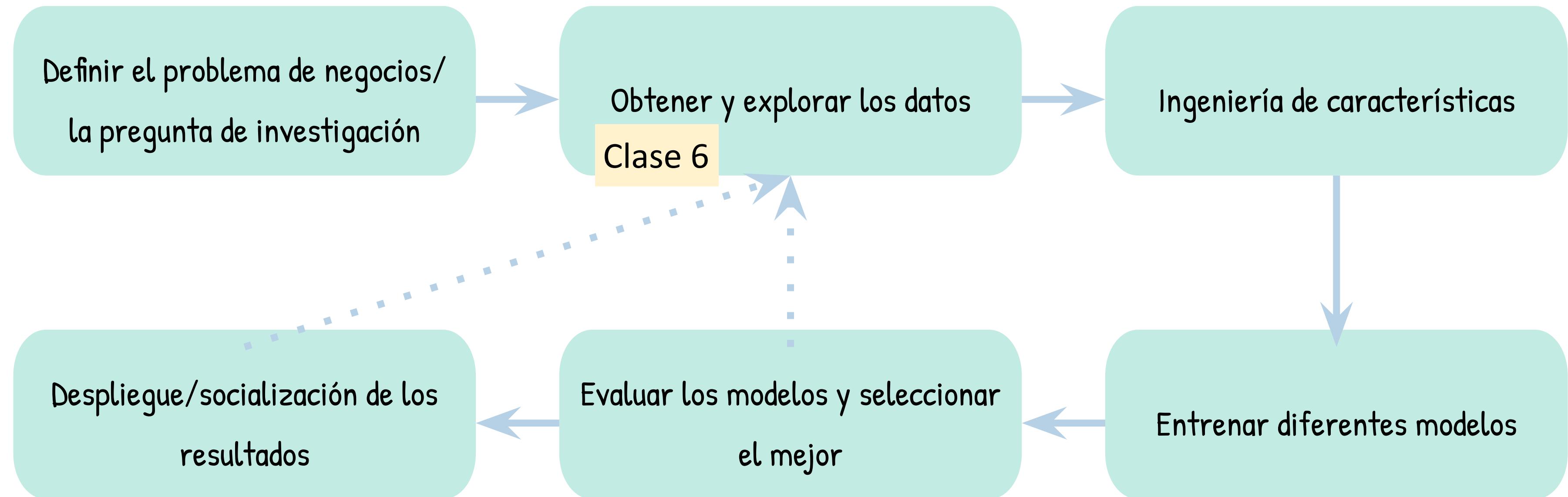
[https://colab.research.google.com/drive/1sgU2aft3A2FCDWt4X1v\\_Sx07jOUBfAf5?usp=sharing](https://colab.research.google.com/drive/1sgU2aft3A2FCDWt4X1v_Sx07jOUBfAf5?usp=sharing)



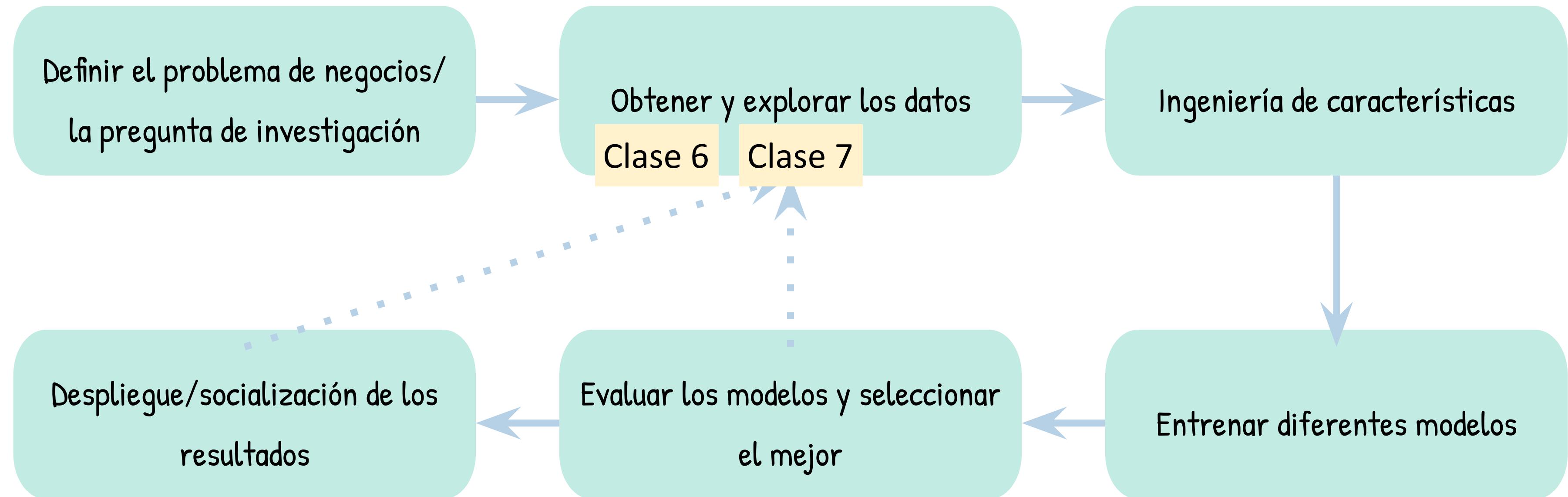
# Pasos en un proyecto de Machine Learning



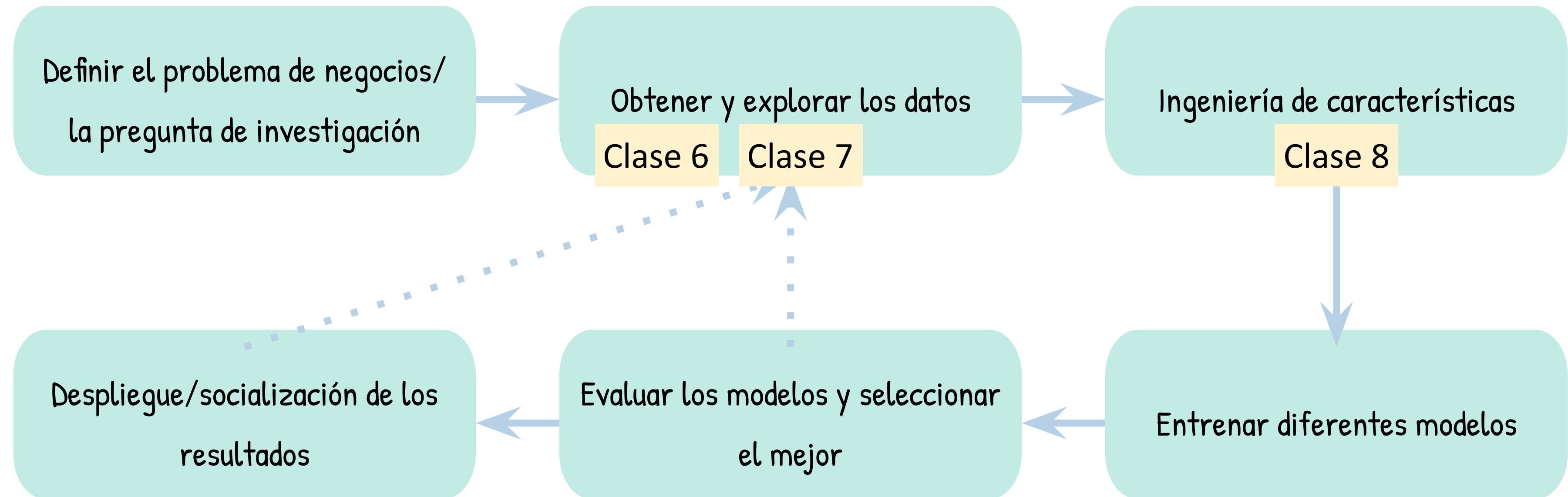
# Pasos en un proyecto de Machine Learning



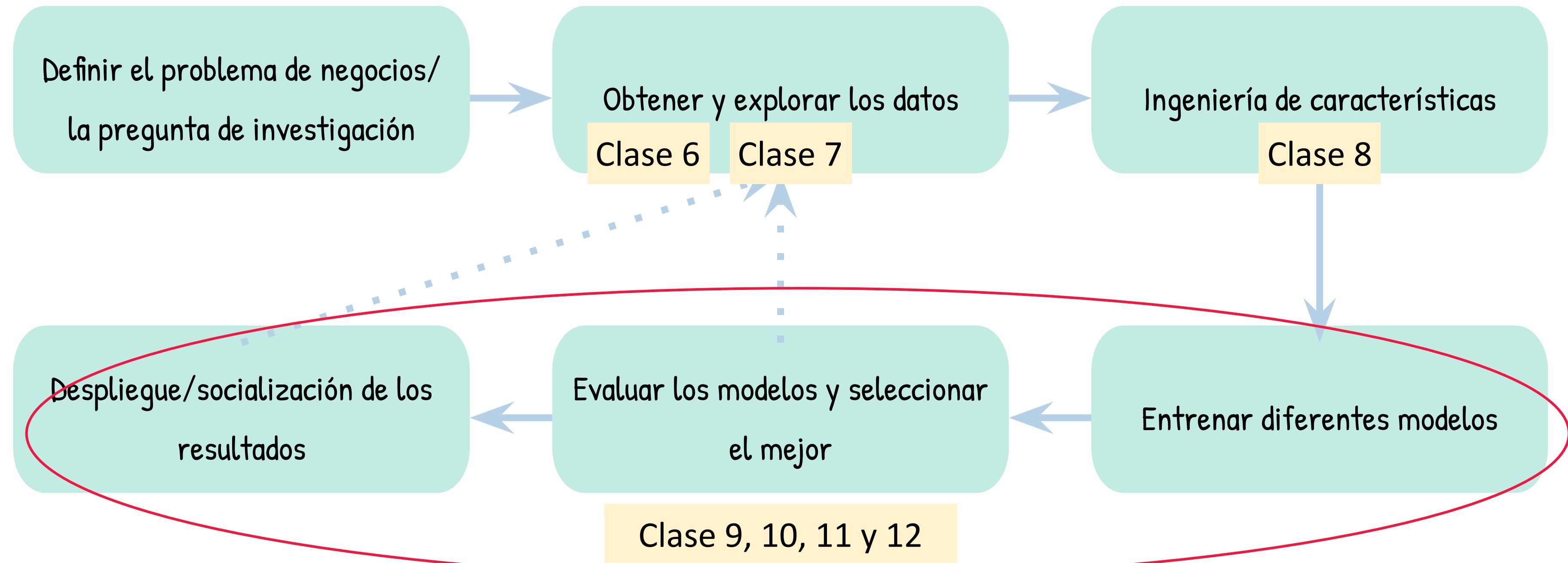
# Pasos en un proyecto de Machine Learning



# Pasos en un proyecto de Machine Learning



# Pasos en un proyecto de Machine Learning

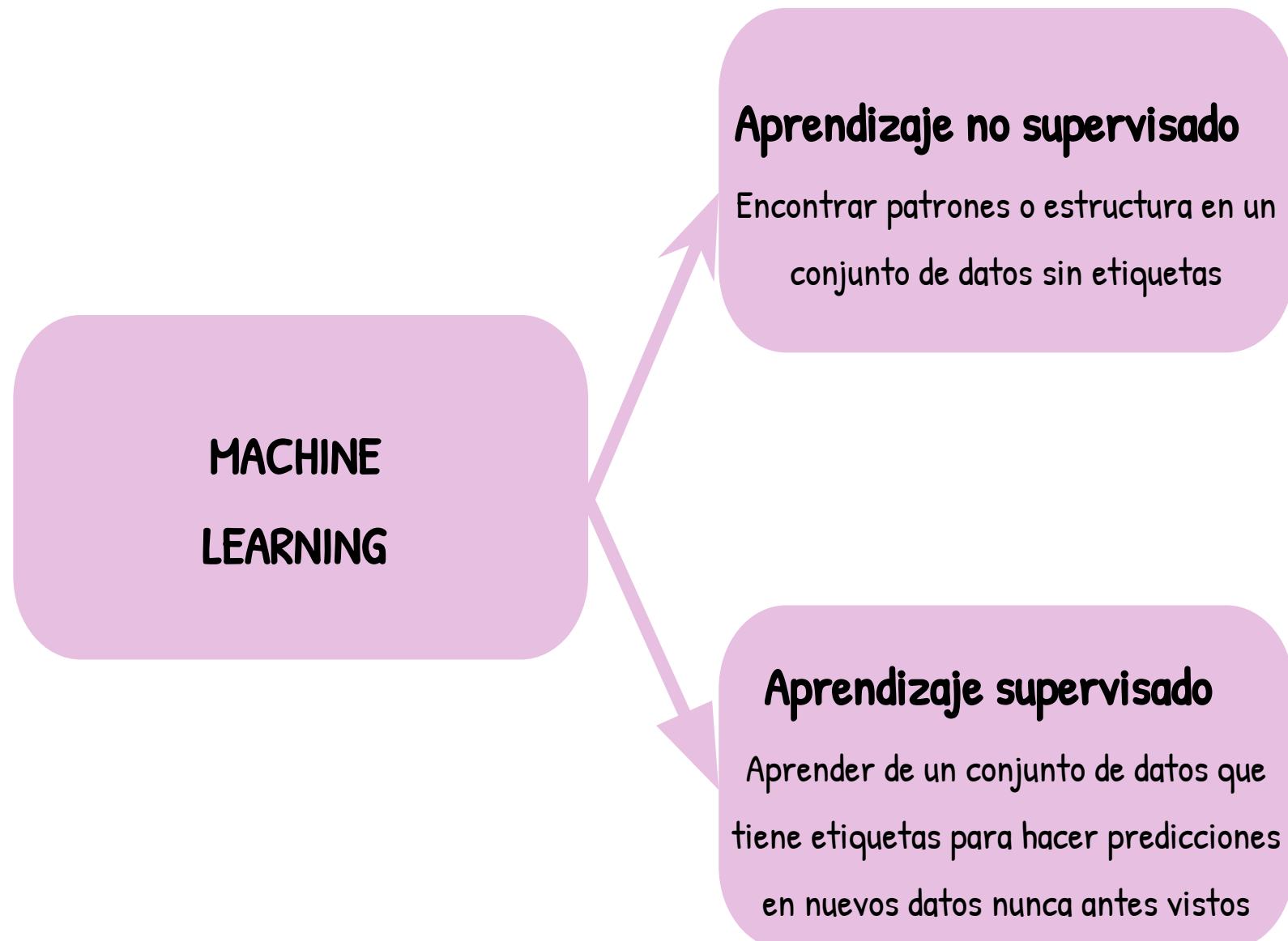


# Familia de modelos de Machine Learning

¿Qué herramienta usarías para clavar el clavo?



# 1. ¿Tenemos etiquetas?



# Aprendizaje NO supervisado



- El algoritmo aprende de datos no etiquetados
  - No necesitamos (*ni tenemos*) etiquetas.
  - **Objetivo:** Encontrar patrones, estructuras o relaciones subyacentes en los datos.
- Aplicaciones comunes:
  - Agrupamiento (Clustering)
  - Detección de anomalías
  - Reducción de dimensionalidad
  - y más...

# Aprendizaje NO supervisado



- **Beneficios**

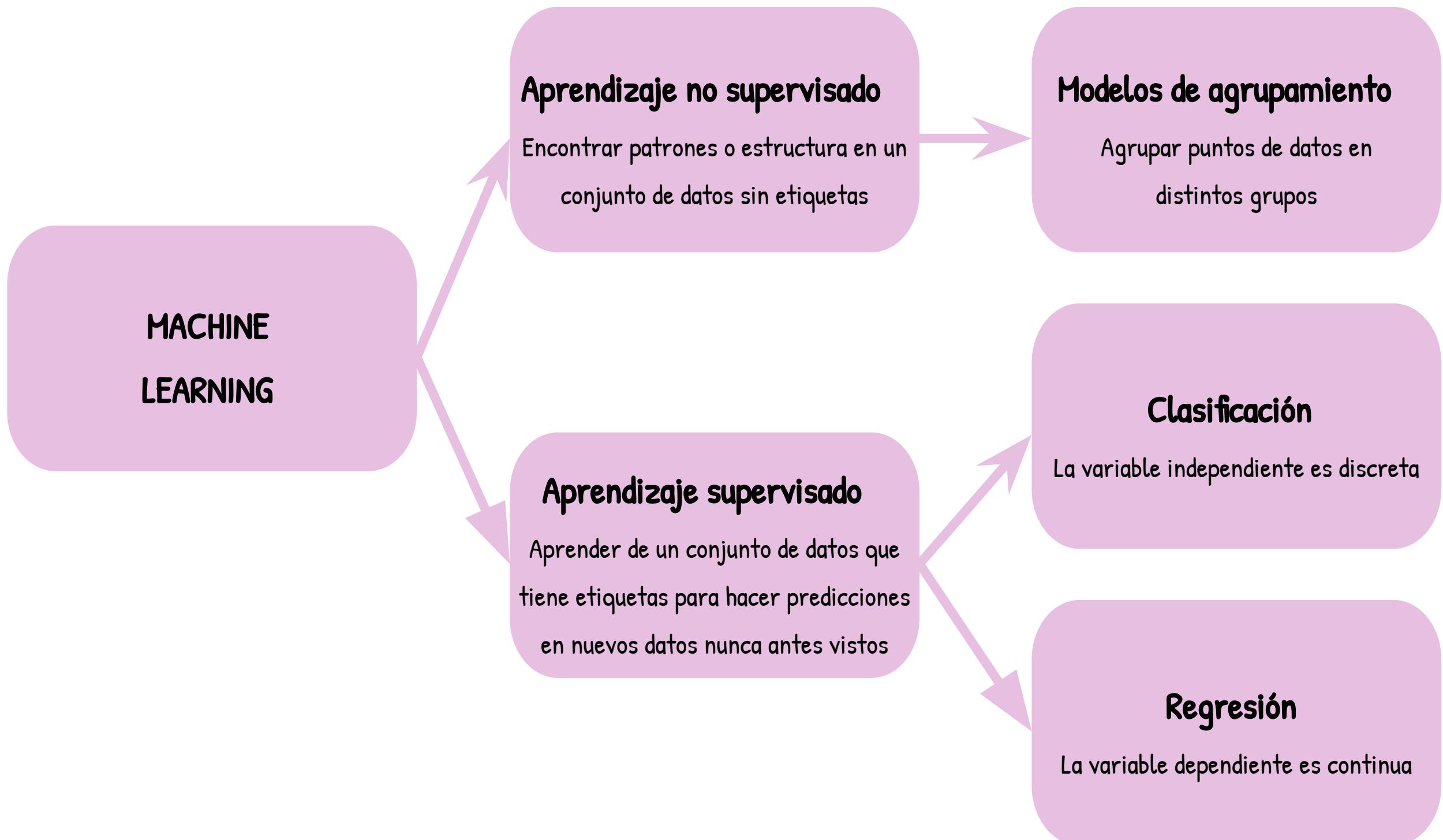
- Descubrir patrones ocultos en los datos que podrían no ser evidentes o intuitivos.
- Gran parte de los datos en el mundo real no están etiquetados.
- Métodos como PCA pueden reducir el número de características, haciendo que otras tareas de aprendizaje automático sean más eficientes.

- **Desafíos**

- Sin una métrica clara, a menudo es más difícil evaluar los resultados.
- Los resultados, como los grupos, pueden ser difíciles de interpretar. A menudo no es evidente de inmediato qué representan los grupos en términos de patrones o propiedades subyacentes.

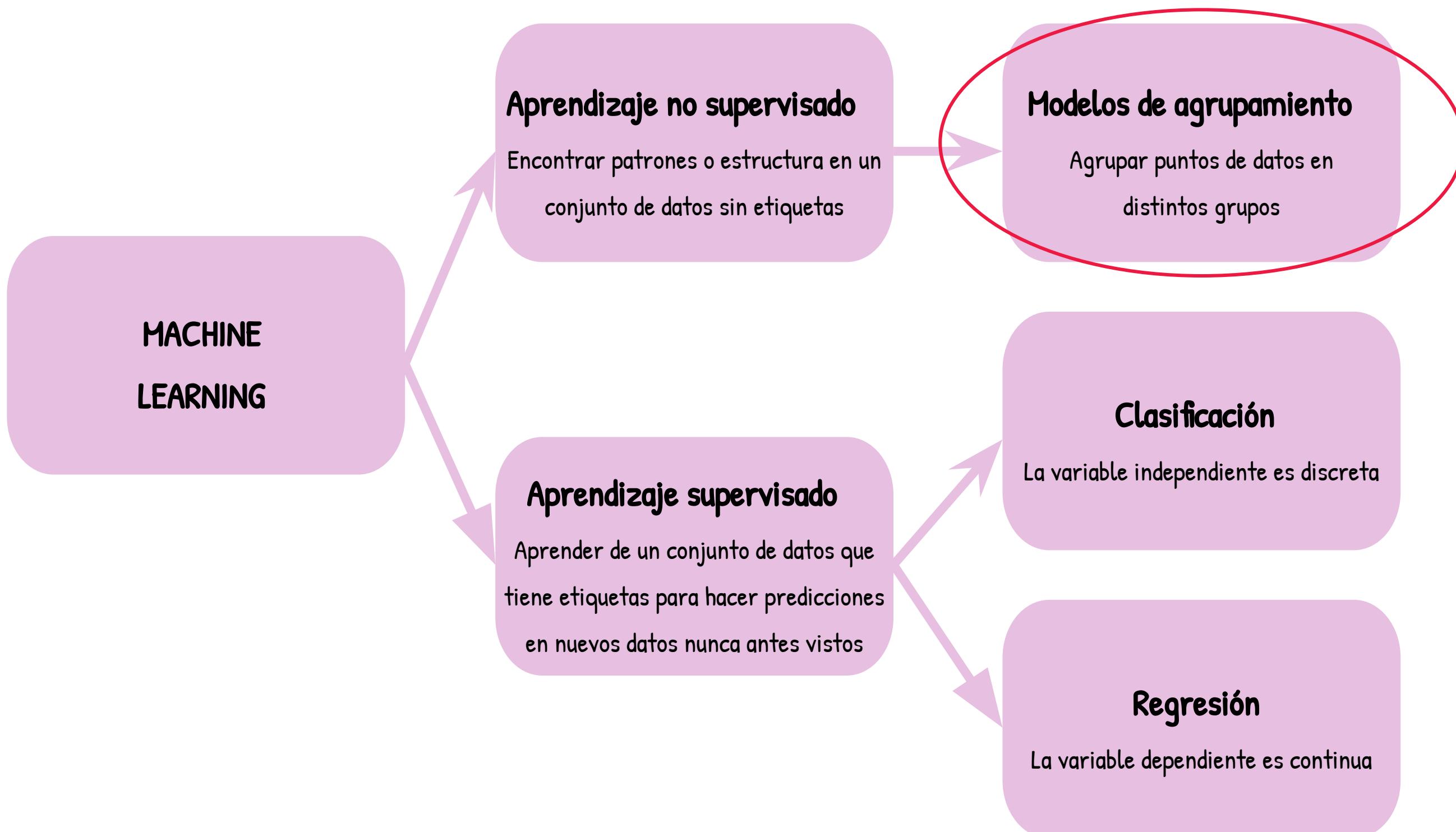
# 1. ¿Tenemos etiquetas?

## 2. ¿De qué tipo son nuestras etiquetas?



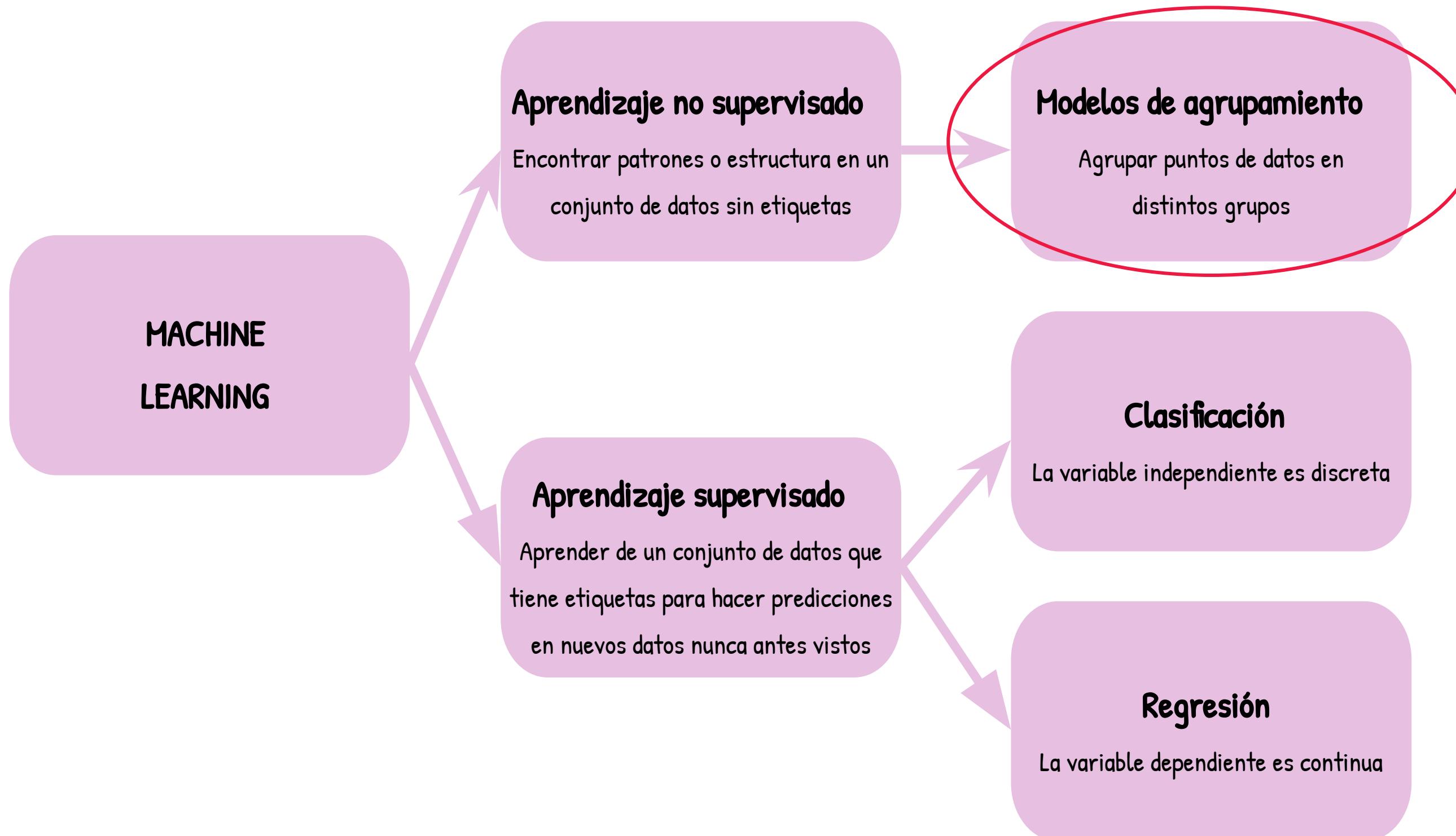
# 1. ¿Tenemos etiquetas?

## 2. ¿De qué tipo son nuestras etiquetas?



# 1. ¿Tenemos etiquetas?

## 2. ¿De qué tipo son nuestras etiquetas?



**Agrupamiento:**  
K-Means Clustering  
Hierarchical Clustering  
DBSCAN  
Gaussian Mixture Models (GMM)  
Spectral Clustering

**Detección de anomalías:**  
Isolation Forest  
One-Class SVM  
Local Outlier Factor

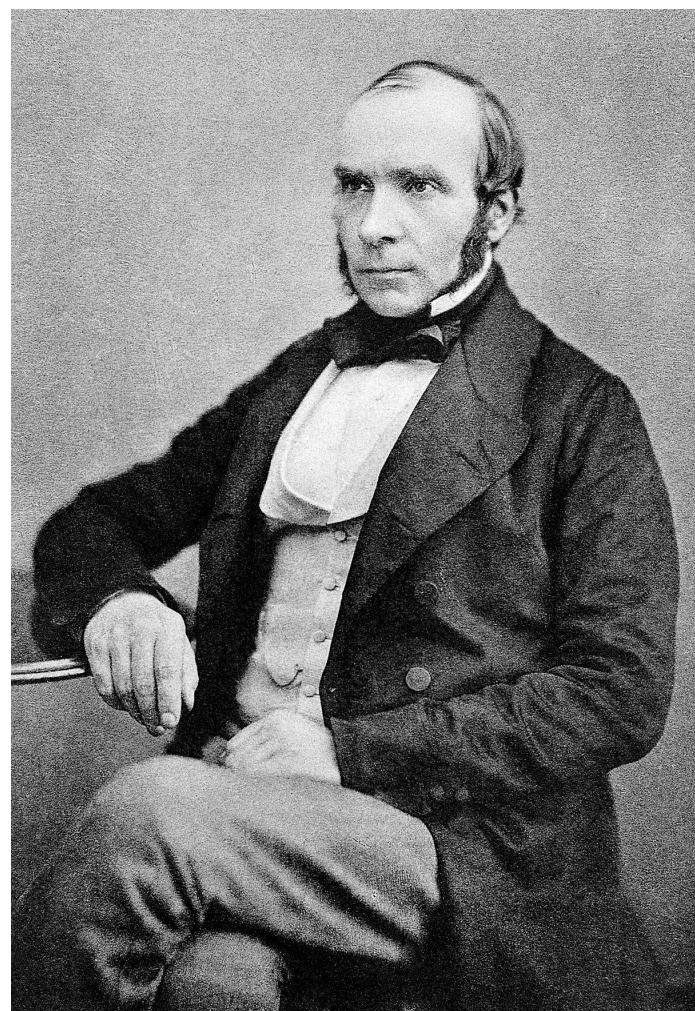
**Reducción de dimensión:**  
Principal Component Analysis (PCA)  
t-SNE (t-Distributed Stochastic Neighbor Embedding)  
Latent Dirichlet Allocation (LDA)

# Primera (?) aplicación de agrupamiento



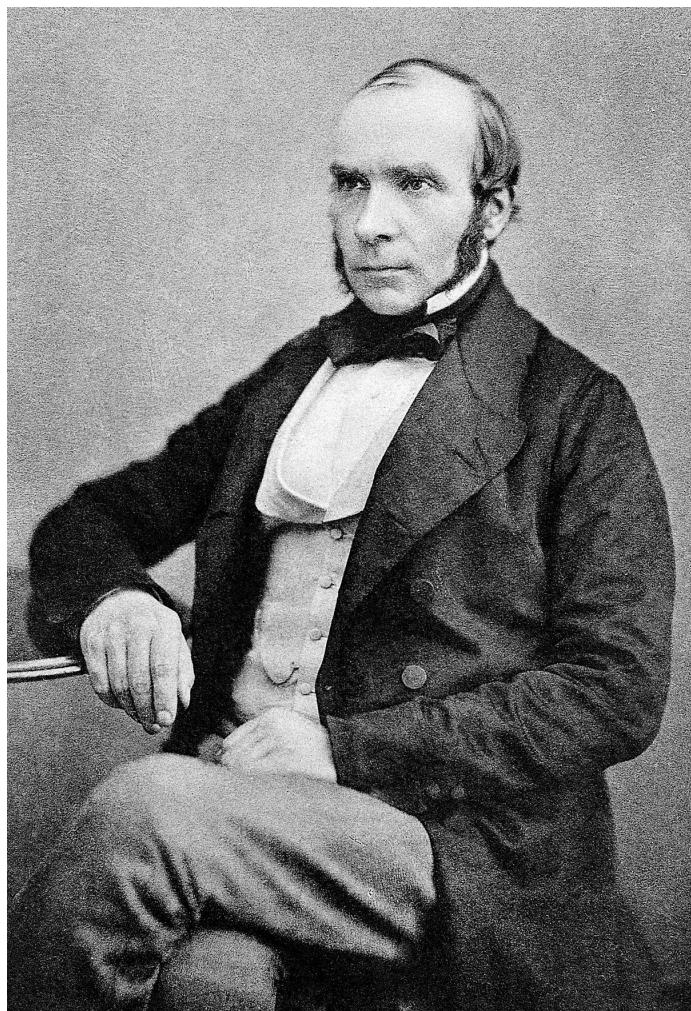
- En la década de 1850, un médico londinense llamado John Snow trazó la localización de las muertes por cólera en un mapa

# Primera (?) aplicación de agrupamiento



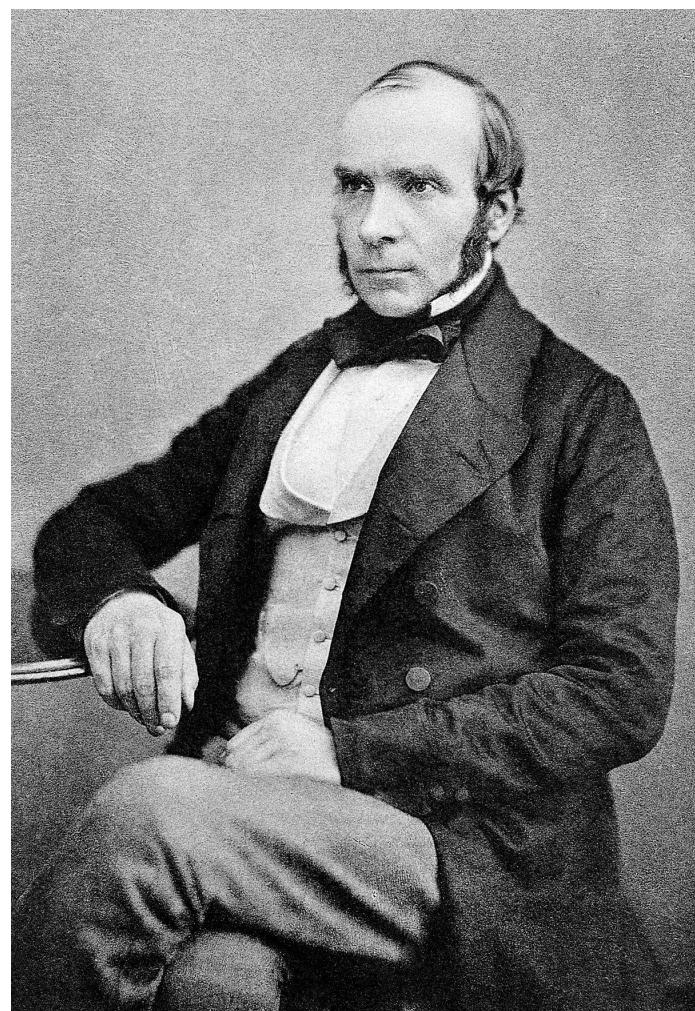
- En la década de 1850, un médico londinense llamado John Snow trazó la localización de las muertes por cólera en un mapa

# Primera (?) aplicación de agrupamiento

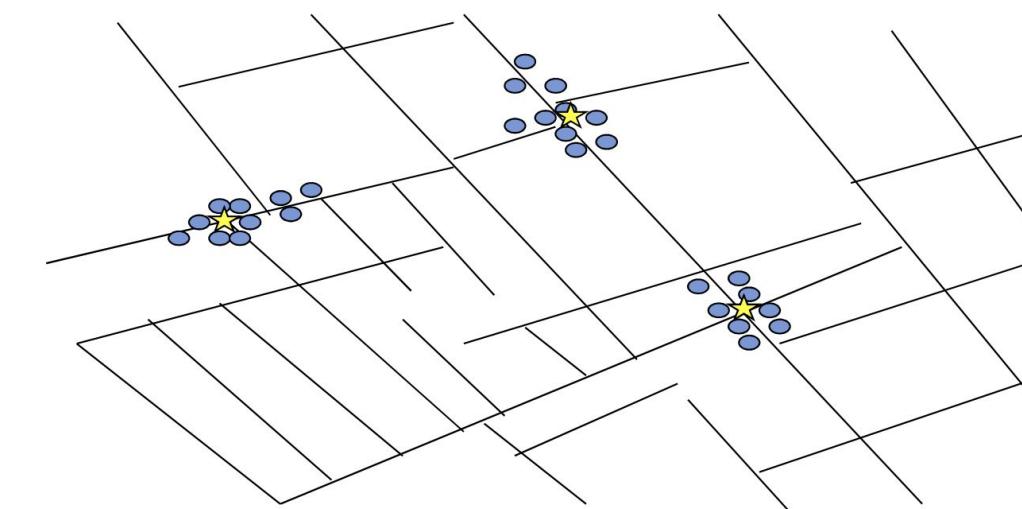


- En la década de 1850, un médico londinense llamado John Snow trazó la localización de las muertes por cólera en un mapa
- Las ubicaciones mostraron que los casos se agrupaban cerca de ciertas intersecciones donde había pozos contaminados, exponiendo así tanto el problema como la solución.

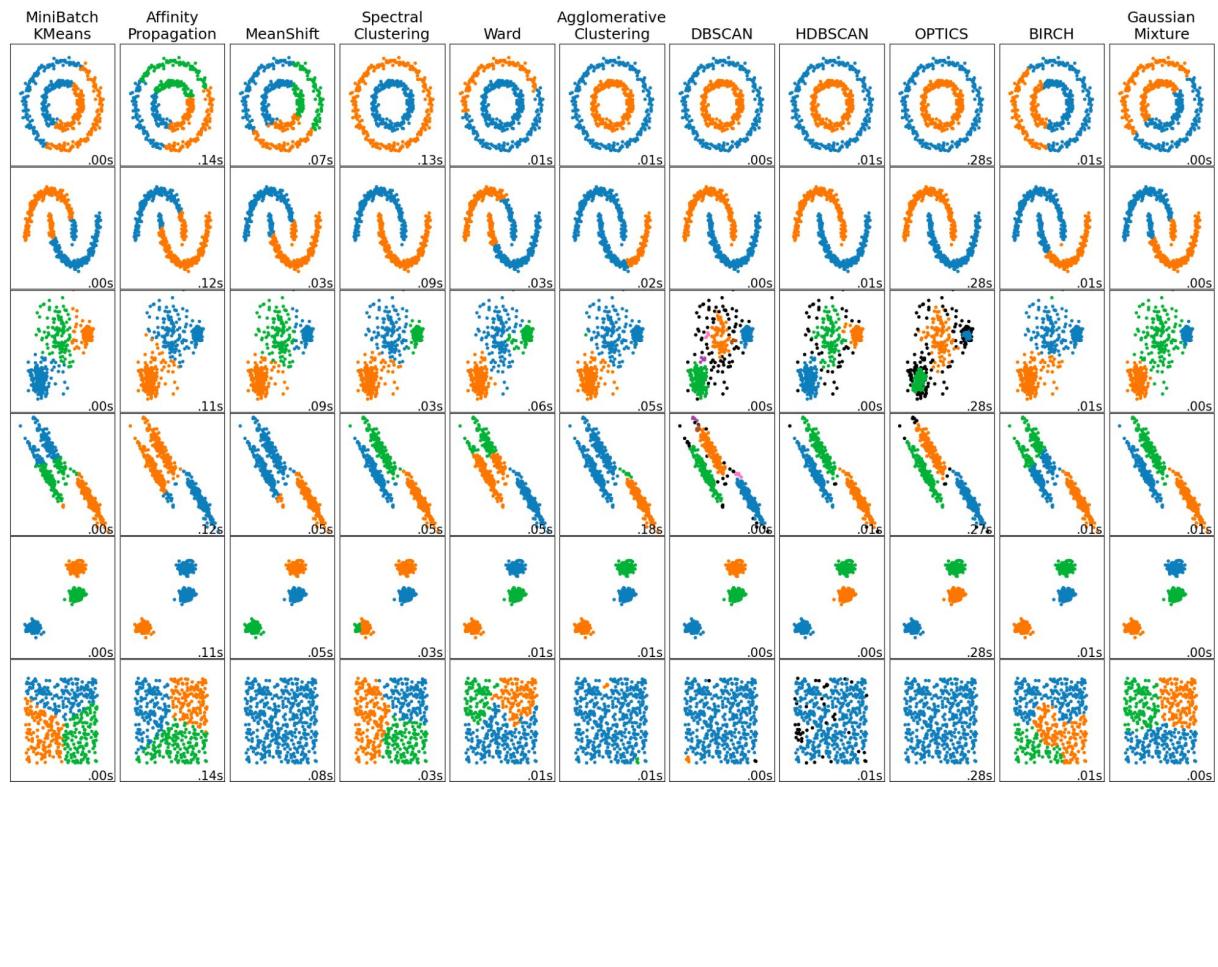
# Primera (?) aplicación de agrupamiento



- En la década de 1850, un médico londinense llamado John Snow trazó la localización de las muertes por cólera en un mapa
- Las ubicaciones mostraron que los casos se agrupaban cerca de ciertas intersecciones donde había pozos contaminados, exponiendo así tanto el problema como la solución.



# Algoritmos de agrupación



- **Objetivo:** Agrupar puntos de datos en distintos grupos de manera que los puntos dentro del mismo grupo sean más parecidos entre sí que a los de otros grupos.
- <https://scikit-learn.org/stable/modules/clustering.html>

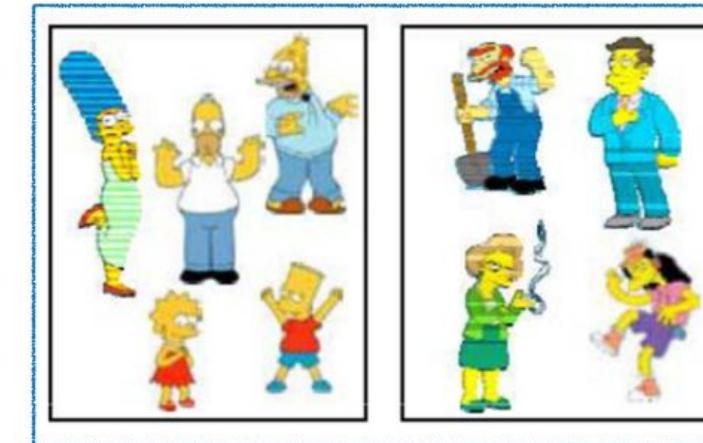
# Algoritmos de agrupación

- ¿Cuándo usarlos? Cuando no sabemos exactamente qué estamos buscando
- ... pero, **cuidado**, ¡puede volverse un caos! 
- El conjunto de datos debe tener:
  - Alta similitud dentro de cada clase
  - Baja similitud entre clases

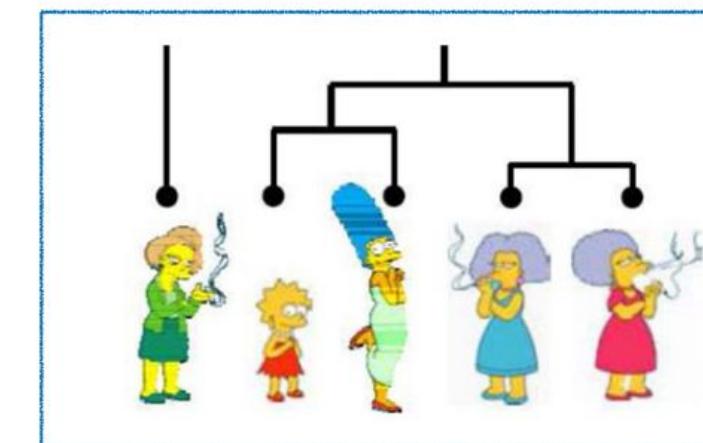
# Algoritmos de agrupación

Modelos de agrupamiento

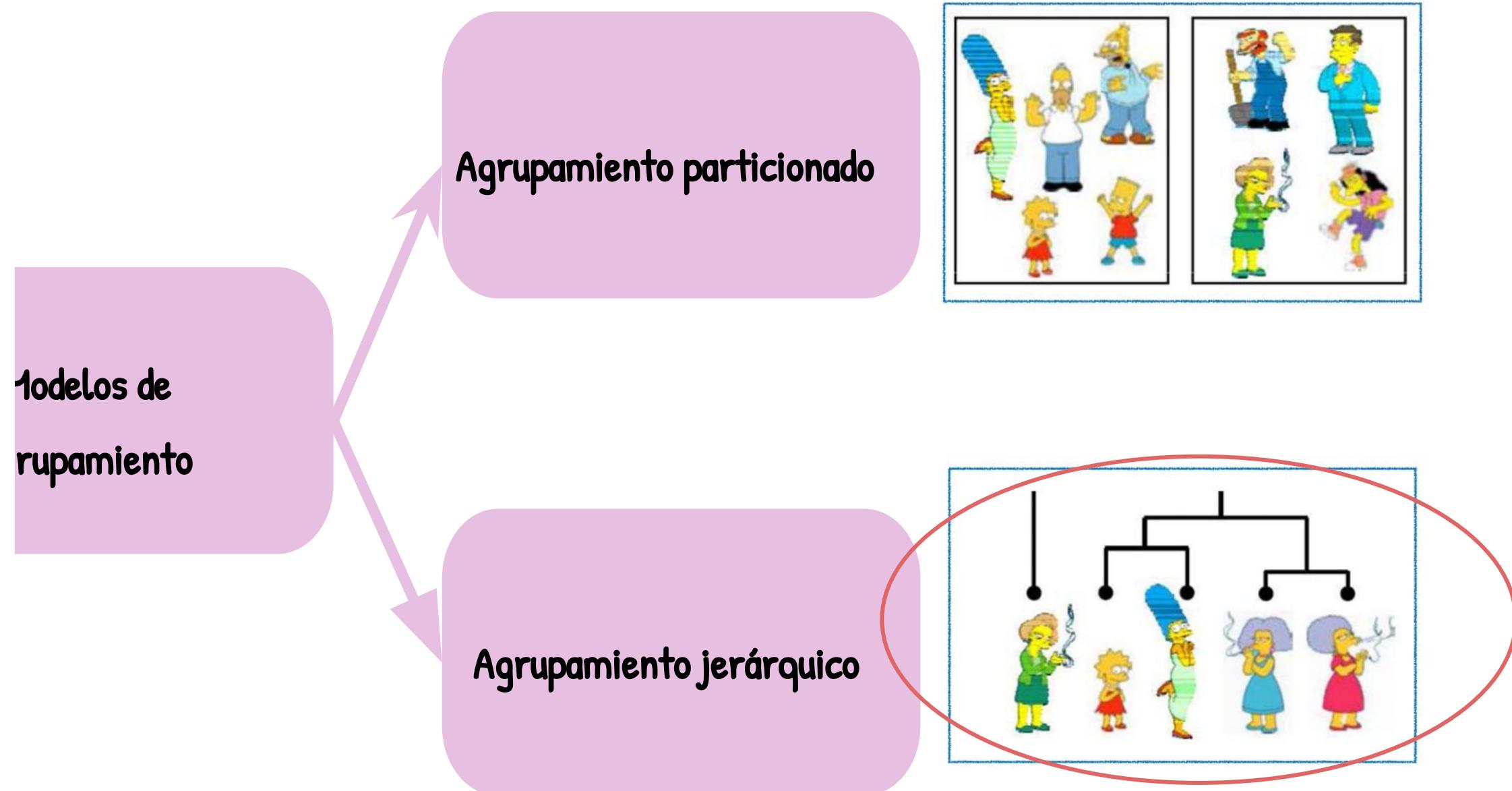
Agrupamiento particionado



Agrupamiento jerárquico



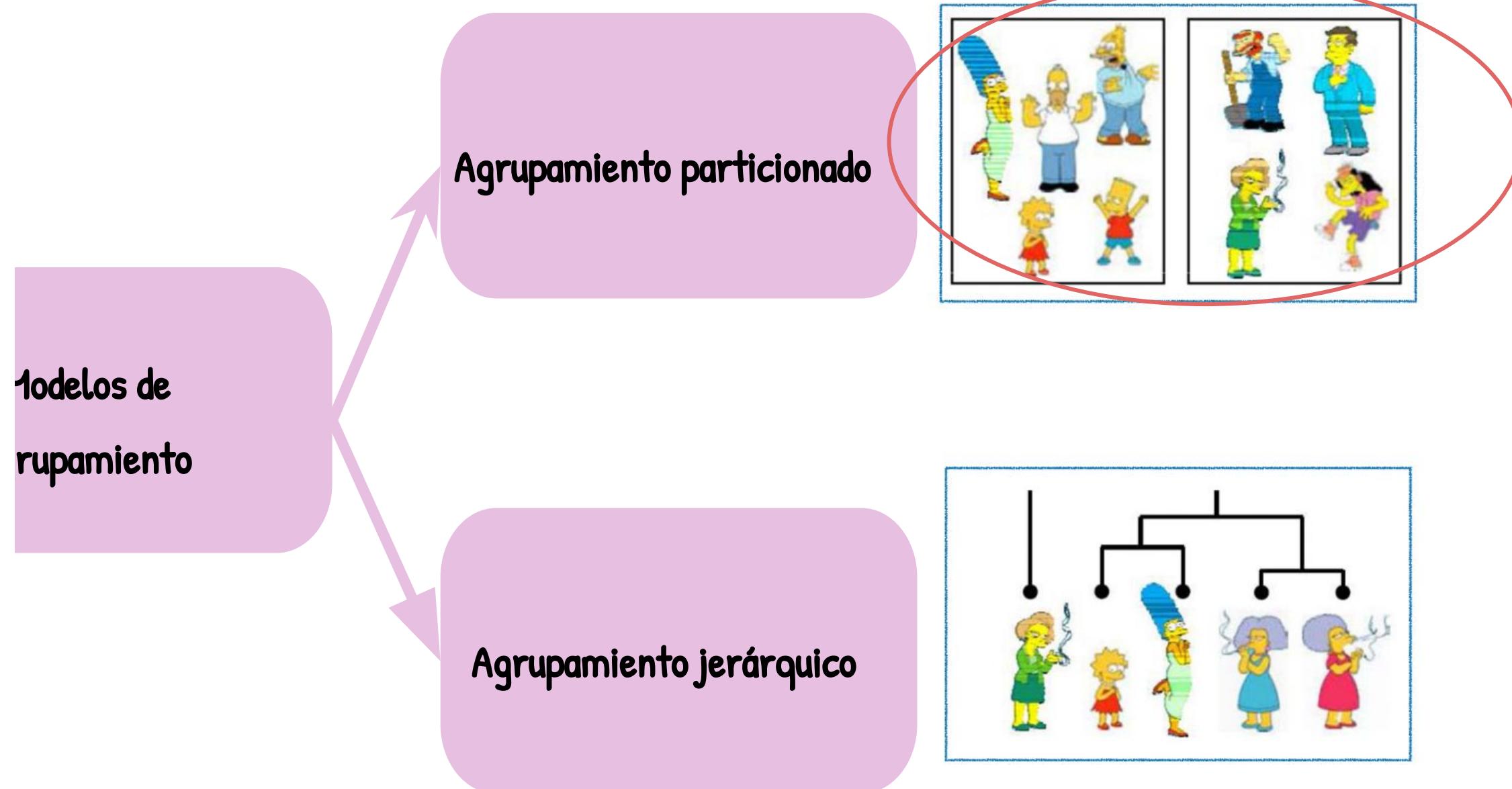
# Algoritmos de agrupación



# Modelos de agrupación: Jerárquicos

- El agrupamiento jerárquico crea un árbol de grupos, ofreciendo una jerarquía multnivel donde cada nivel representa un grado particular de granularidad en el agrupamiento.
- Algoritmos comunes:
  - Agrupamiento Jerárquico Aglomerativo
  - Agrupamiento Jerárquico Divisivo
  - y más...

# Algoritmos de agrupación



# Modelos de agrupación: Particionamiento

- El agrupamiento particional, también conocido como agrupamiento no jerárquico, divide un conjunto de datos en un conjunto de grupos que no se superponen. A diferencia del agrupamiento jerárquico, el agrupamiento particional no organiza los grupos en una estructura de árbol.
- Algoritmos comunes:
  - **K-Medias (K-Means)**
  - K-Medianas
  - DBSCAN
  - Agrupamiento espectral
  - y más...

**K-NN**  **K-Means**

**K-Vecinos**  
**NO es lo mismo**  
**que K-Medias**

# K-Medias

# K-Medias

- Modelo **NO** supervisado de Machine Learning
  - Agrupamiento
    - Particionamiento

# K-Medias

- Al igual que cualquier algoritmo de agrupamiento, K-Medias agrupa puntos de datos en un número determinado de clusters (grupos).

# K-Medias

- Al igual que cualquier algoritmo de agrupamiento, K-Medias agrupa puntos de datos en un número determinado de clusters (grupos).

K-means {        
 k → Número de grupos  
 means → Media aritmética

# K-Medias

- Al igual que cualquier algoritmo de agrupamiento, K-Medias agrupa puntos de datos en un número determinado de clusters (grupos).

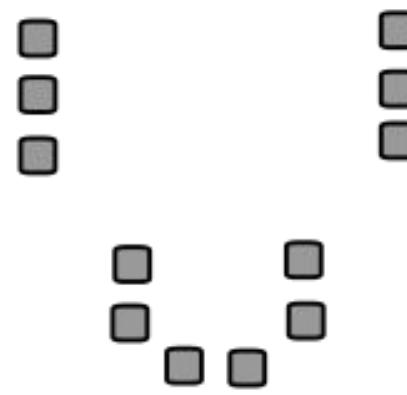
K-means {

k	→ Número de grupos
means	→ Media aritmética

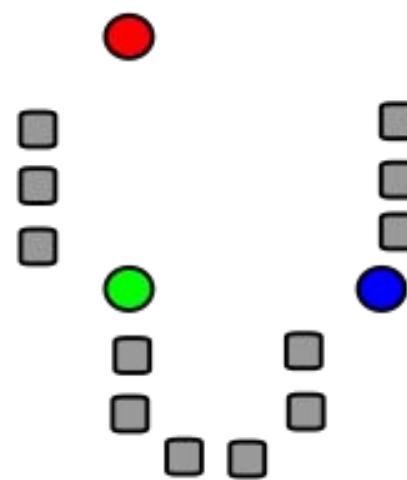
- K-Medias es un algoritmo **iterativo** cuyo objetivo es particionar un conjunto de **N** observaciones en **K** grupos en los que cada observación pertenece al grupo cuyo valor medio es más cercano.

# Paso 1: Inicializar

## Paso 1: Inicializar

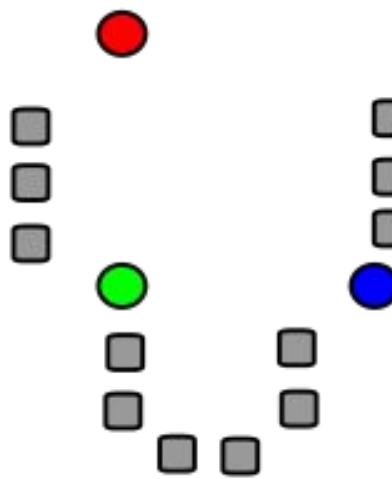


## Paso 1: Inicializar



- Elegir un número  $K$  de cúmulos
- Escoger aleatoriamente  $K$  puntos como centroides

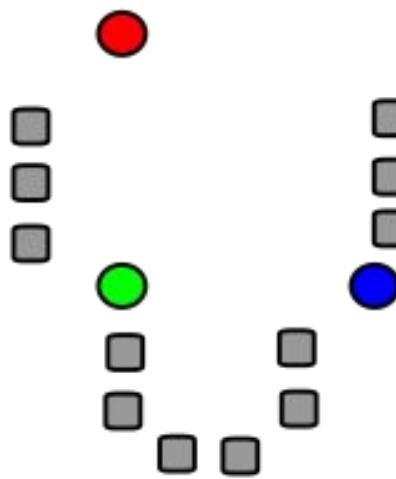
## Paso 1: Inicializar



- Elegir un número K de cúmulos
- Escoger aleatoriamente K puntos como centroides

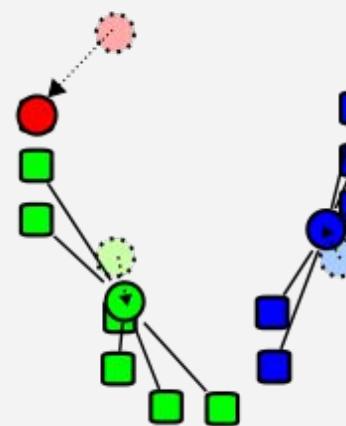
## Paso 2: Repetir

## Paso 1: Inicializar

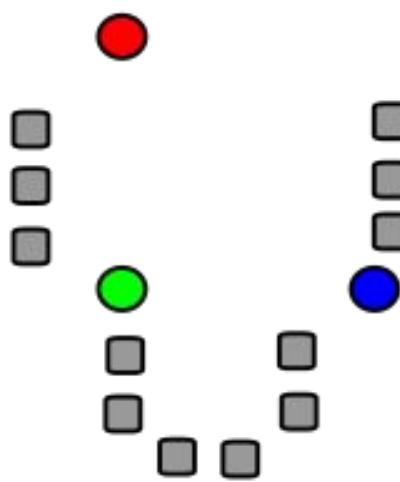


- Elegir un número K de cúmulos
- Escoger aleatoriamente K puntos como centroides

## Paso 2: Repetir

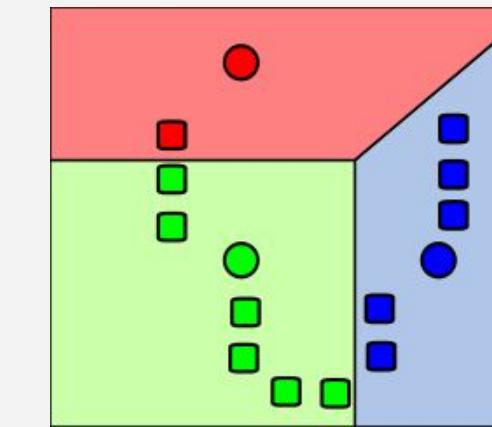
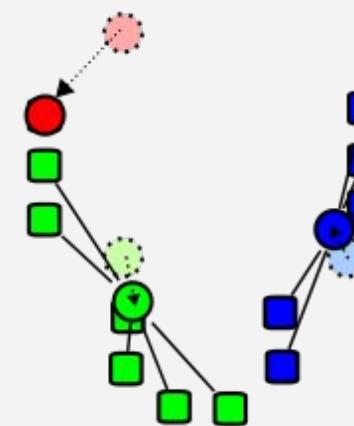


## Paso 1: Inicializar

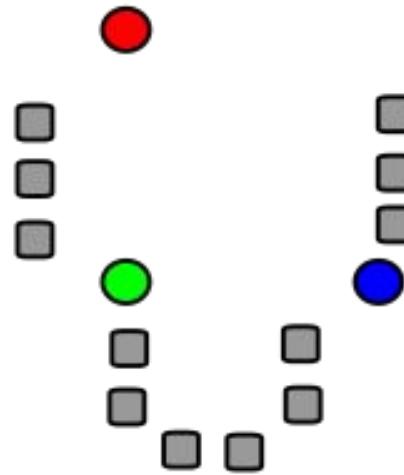


- Elegir un número K de cúmulos
- Escoger aleatoriamente K puntos como centroides

## Paso 2: Repetir

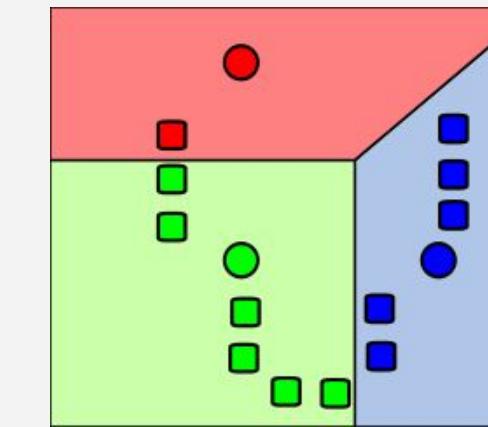
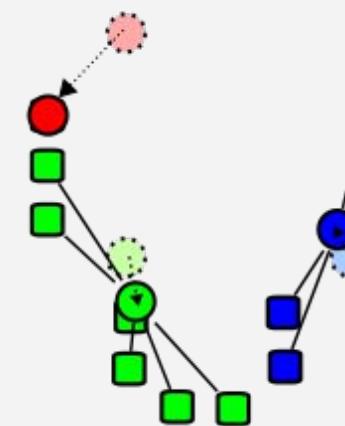


## Paso 1: Inicializar



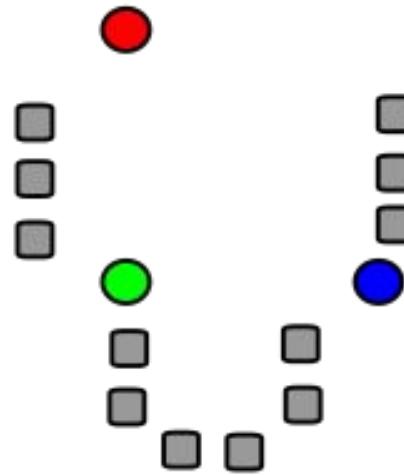
- Elegir un número K de cúmulos
- Escoger aleatoriamente K puntos como centroides

## Paso 2: Repetir



- Los K cúmulos se crean asociando cada observación con la media más cercana
- El nuevo centroide de cada uno de los K cúmulos es la media de sus observaciones

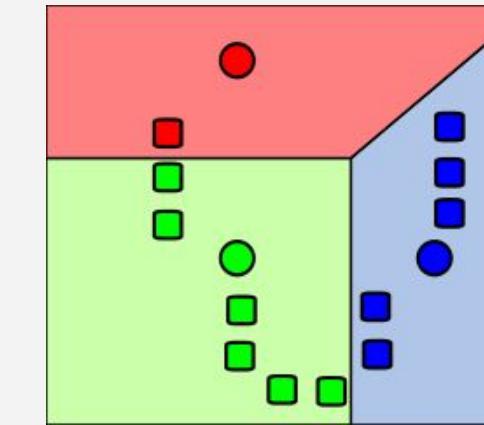
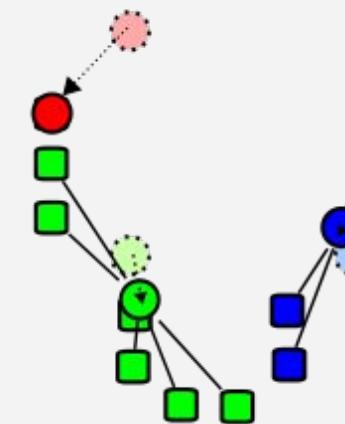
## Paso 1: Inicializar



- Elegir un número K de cúmulos
- Escoger aleatoriamente K puntos como centroides

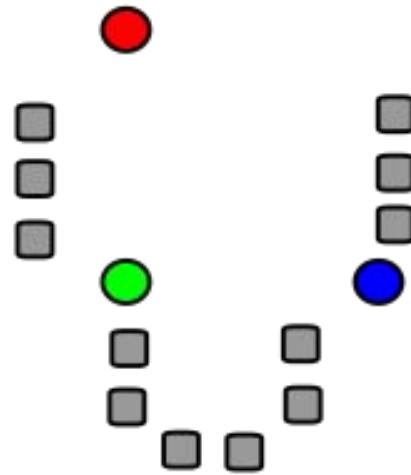
## Paso 3: Parar

## Paso 2: Repetir



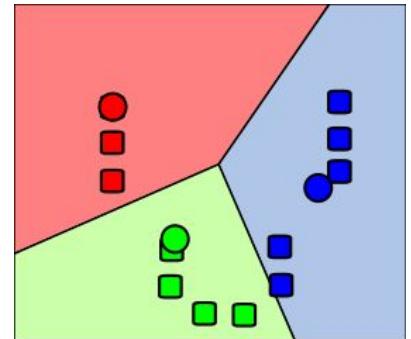
- Los K cúmulos se crean asociando cada observación con la media más cercana
- El nuevo centroide de cada uno de los K cúmulos es la media de sus observaciones

## Paso 1: Inicializar

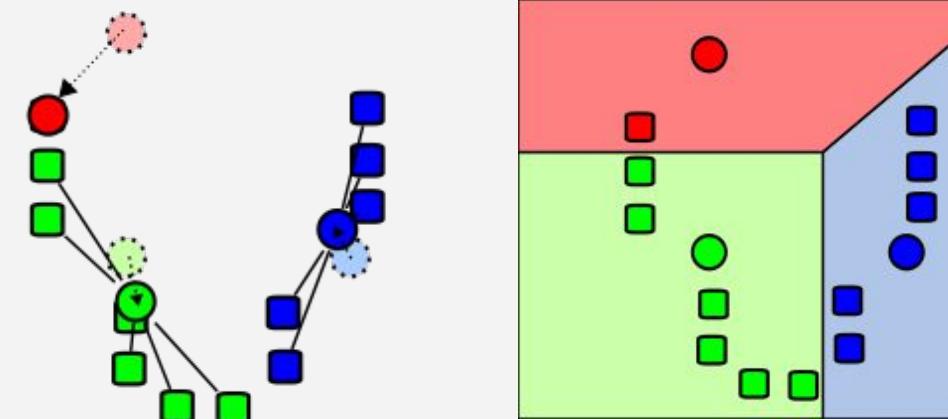


- Elegir un número K de cúmulos
- Escoger aleatoriamente K puntos como centroides

## Paso 3: Parar

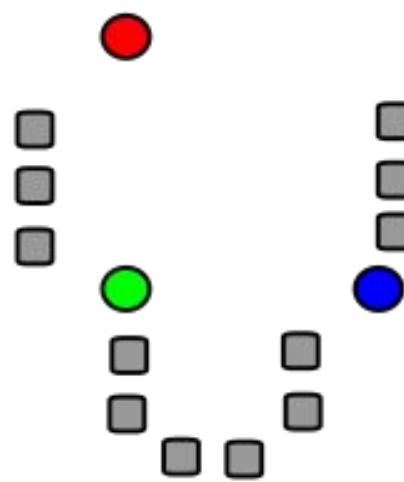


## Paso 2: Repetir



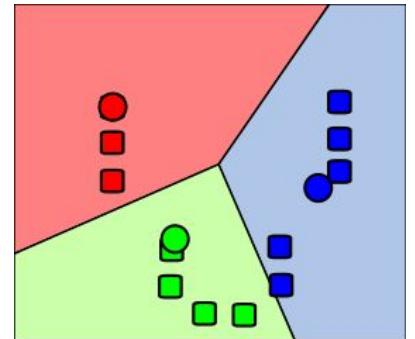
- Los K cúmulos se crean asociando cada observación con la media más cercana
- El nuevo centroide de cada uno de los K cúmulos es la media de sus observaciones

## Paso 1: Inicializar

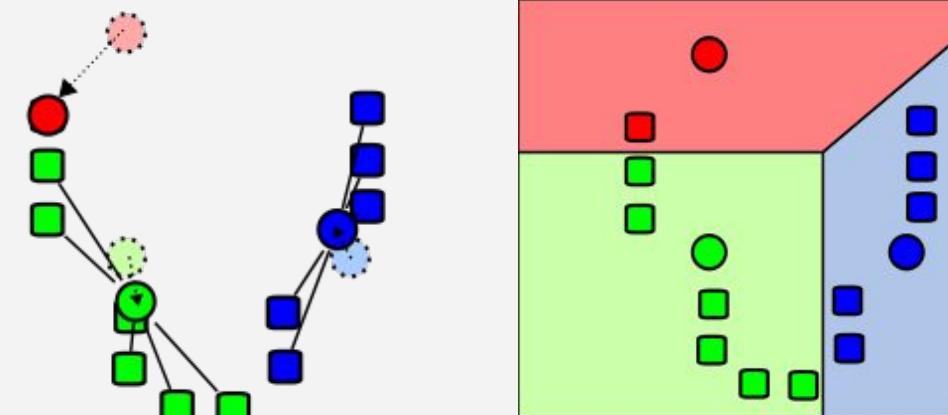


- Elegir un número K de cúmulos
- Escoger aleatoriamente K puntos como centroides

## Paso 3: Parar



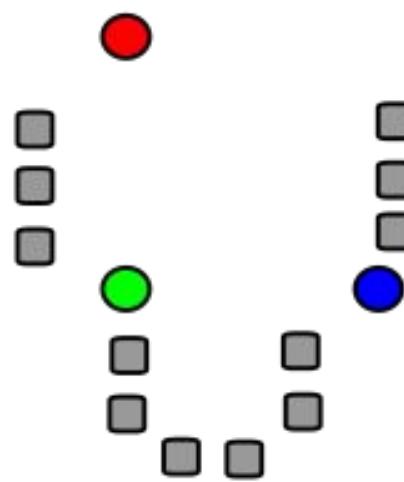
## Paso 2: Repetir



- Los K cúmulos se crean asociando cada observación con la media más cercana
- El nuevo centroide de cada uno de los K cúmulos es la media de sus observaciones

- Repetir pasos 1 y 2
- El algoritmo acaba cuando ya no hay cambio en los centroides de los cúmulos, las observaciones de los cúmulos siguen siendo las mismas, o el máximo número de iteraciones es alcanzado

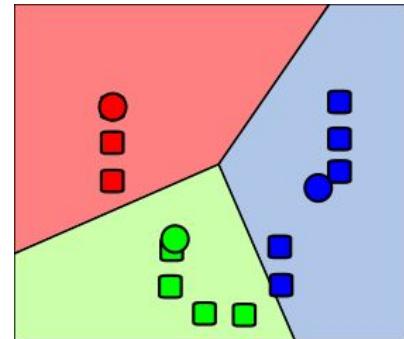
## Paso 1: Inicializar



- Elegir un número K de cúmulos
- Escoger aleatoriamente K puntos como centroides

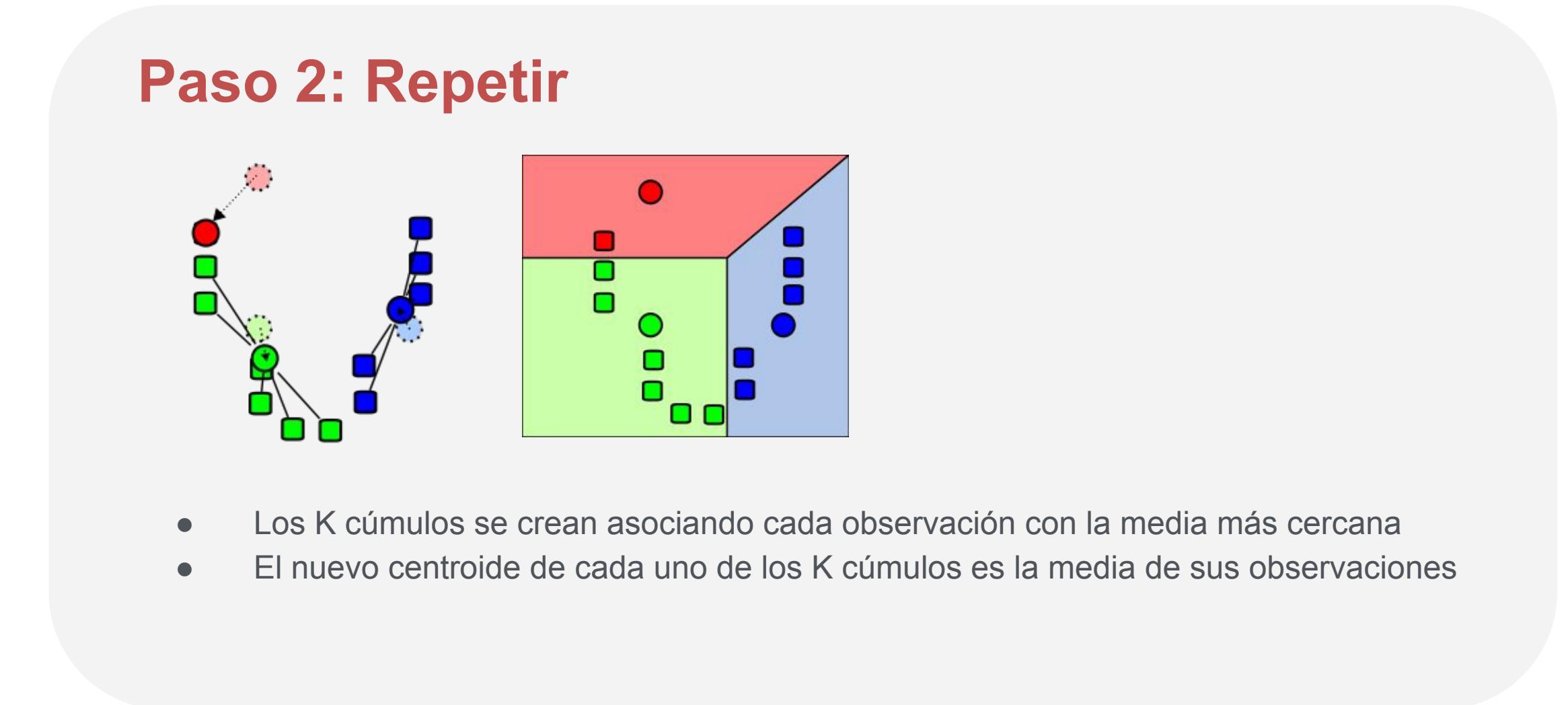
- [La matemática detrás del algoritmo](#)
- [Video](#)

## Paso 3: Parar



- Repetir pasos 1 y 2
- El algoritmo acaba cuando ya no hay cambio en los centroides de los cúmulos, las observaciones de los cúmulos siguen siendo las mismas, o el máximo número de iteraciones es alcanzado

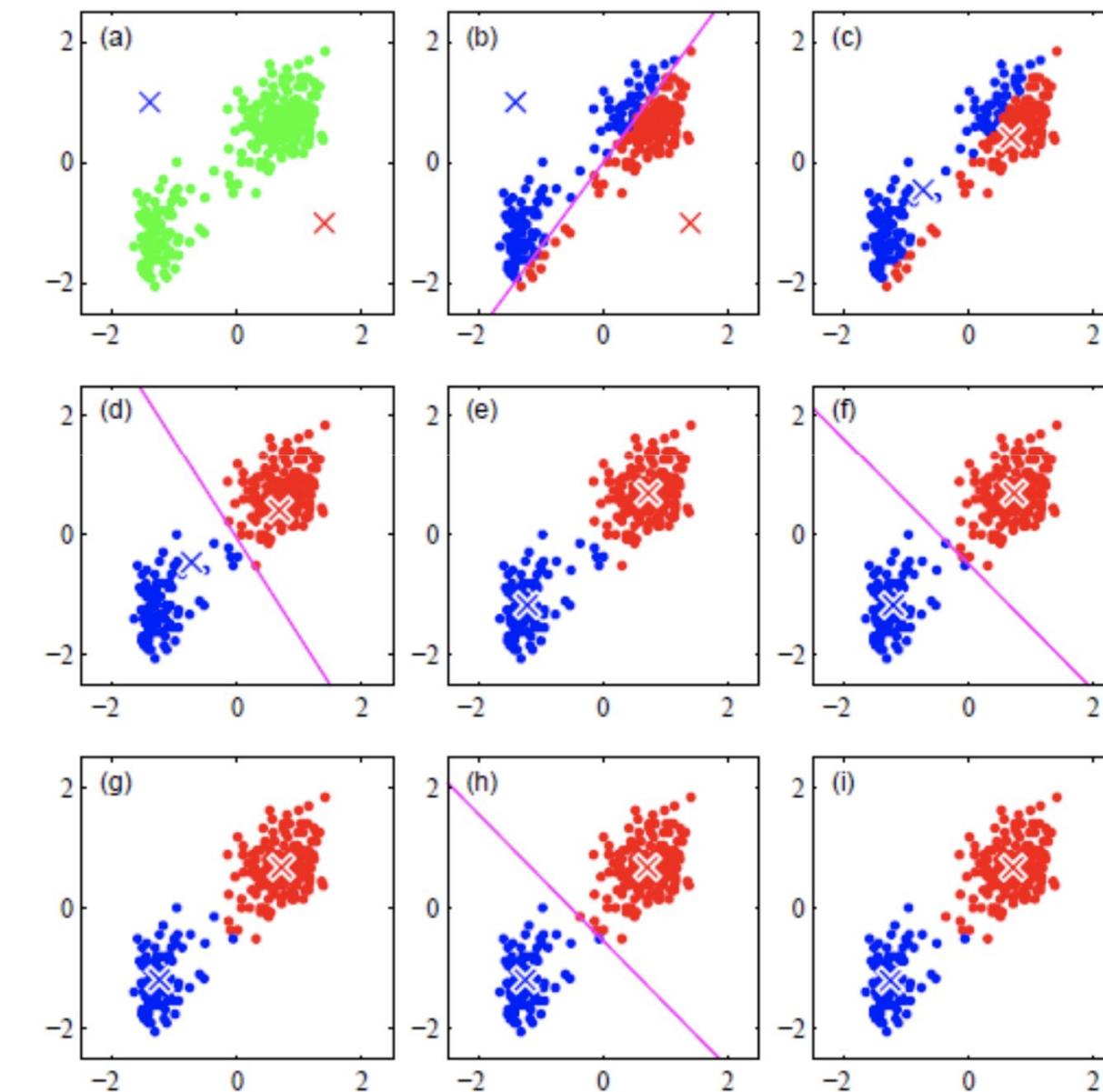
## Paso 2: Repetir





# Punto de control

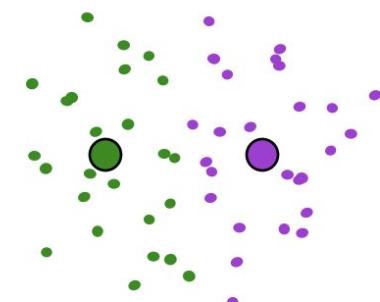
1. ¿Cuál es el número de K?
2. ¿Qué está pasando en cada paso?



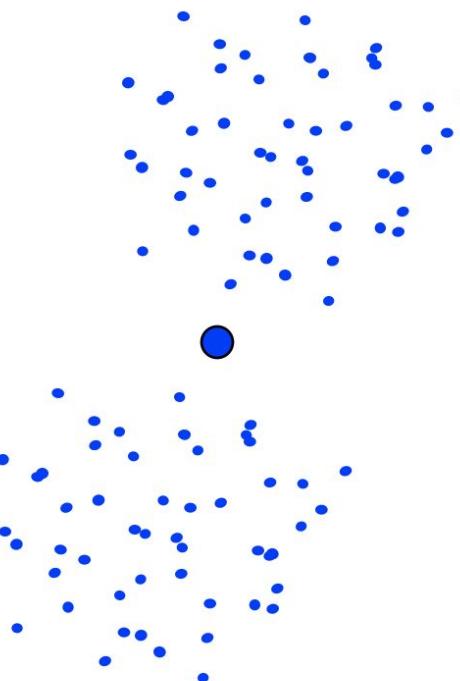


# ¿Cómo escogemos K?

- Es importante escoger un buen número para K  
A local optimum:



Would be better to have  
one cluster here

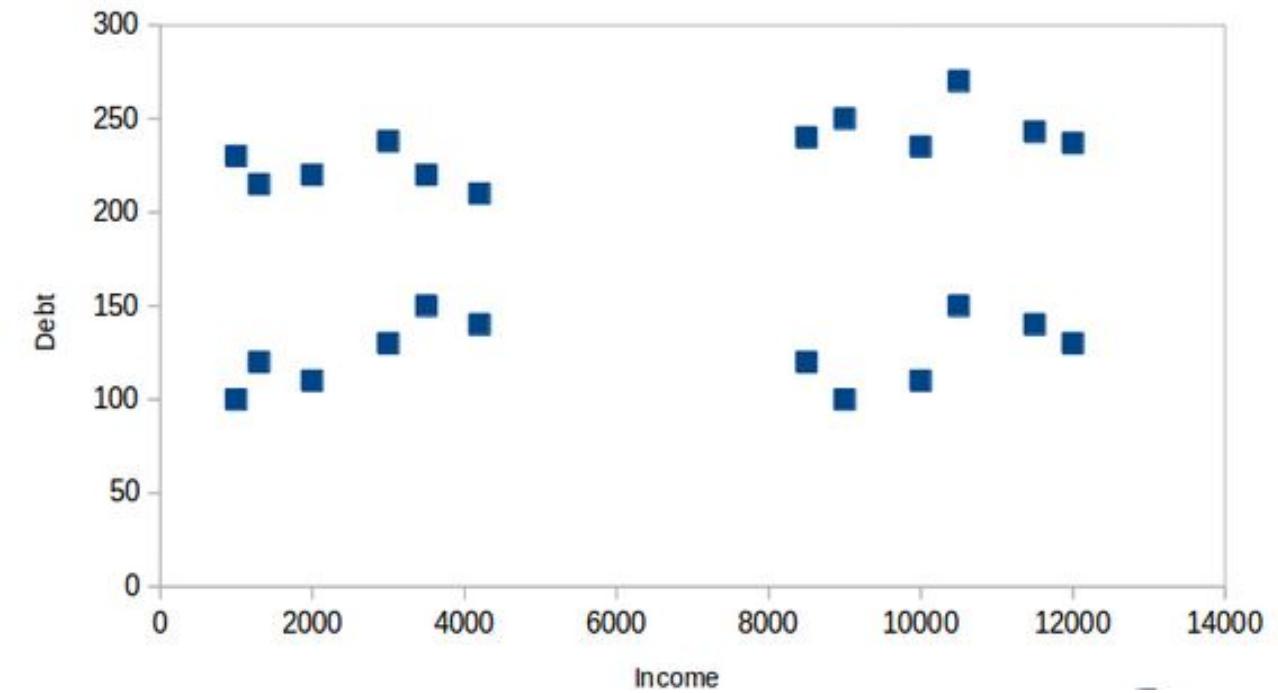


... and two clusters here



# ¿Cómo escogemos K?

- ¿Cuántos grupos tenemos aquí?





# Punto de control

1. ¿Cuál es el número mínimo posible para K?

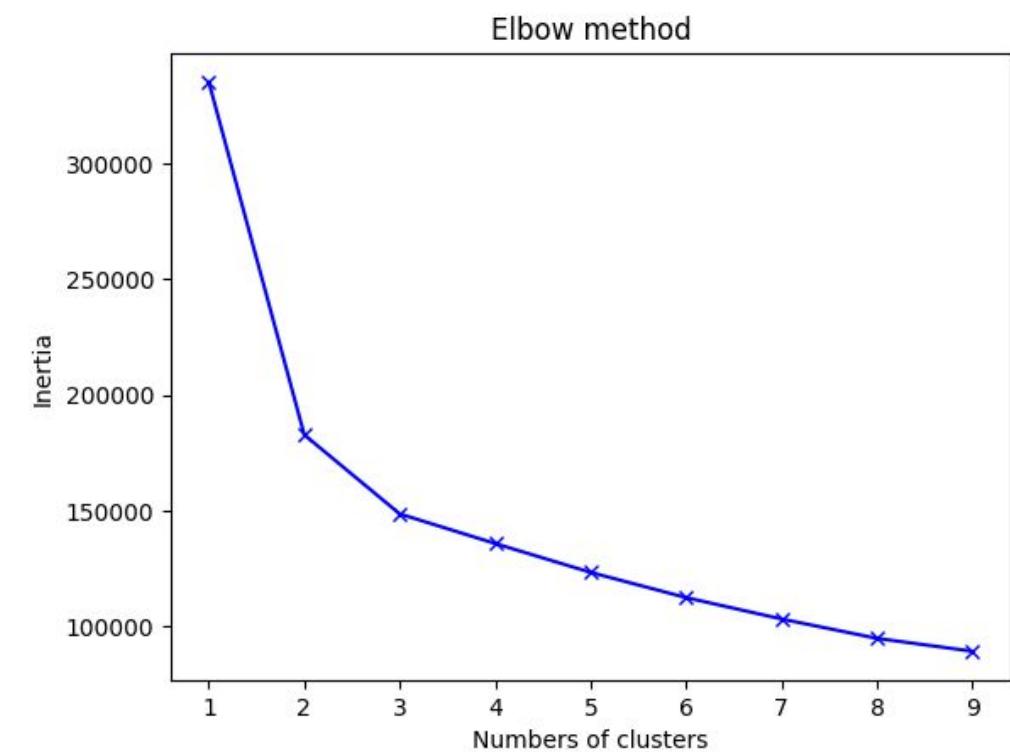


# Punto de control

1. ¿Cuál es el número mínimo posible para K?
2. ¿Cuál es el número máximo posible para K?

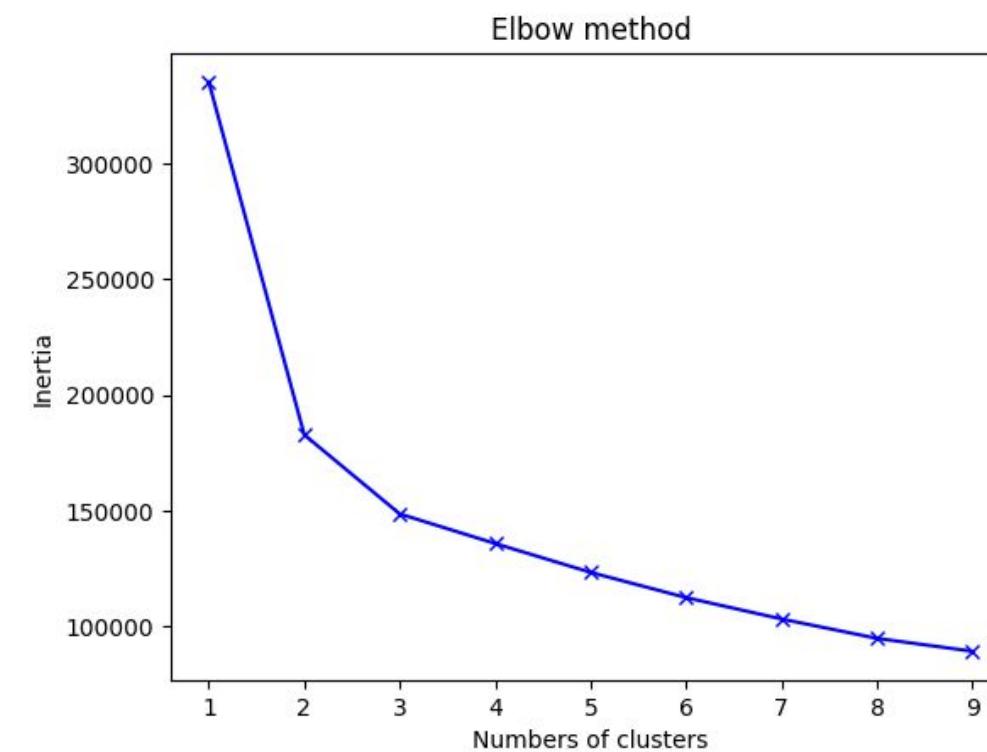
# El método del codo

- Un método utilizado para seleccionar el mejor valor de K es conocido como el método del codo.
- Esta decisión puede ser subjetiva y es crucial, ya que el número de clusters puede influir significativamente en los conocimientos obtenidos del proceso de agrupamiento.



# El método del codo

- Un método utilizado para seleccionar el mejor valor de K es conocido como el método del codo.
- Esta decisión puede ser subjetiva y es crucial, ya que el número de clusters puede influir significativamente en los conocimientos obtenidos del proceso de agrupamiento.



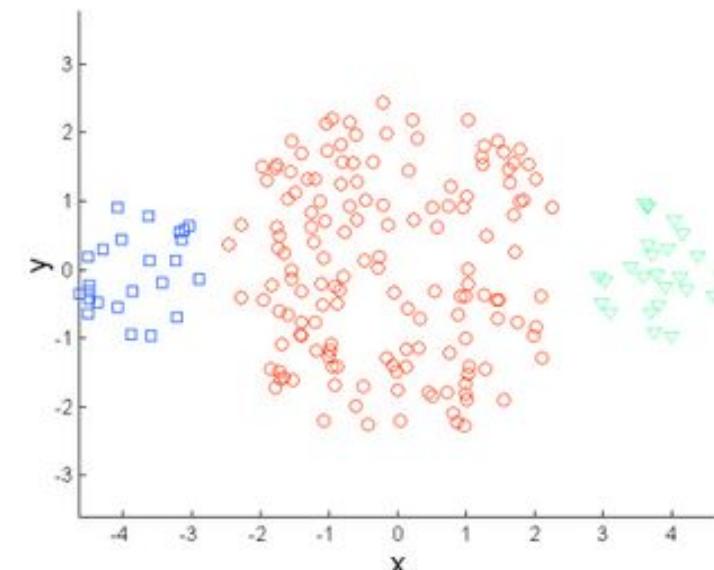
1. Calcular el agrupamiento K-means para diferentes valores de K
2. Calcular la inercia (suma de las distancias al cuadrado entre cada punto de datos y el centroide del grupo al que está asignado)
3. Graficar la curva del codo
4. Ubicar el 'codo'"



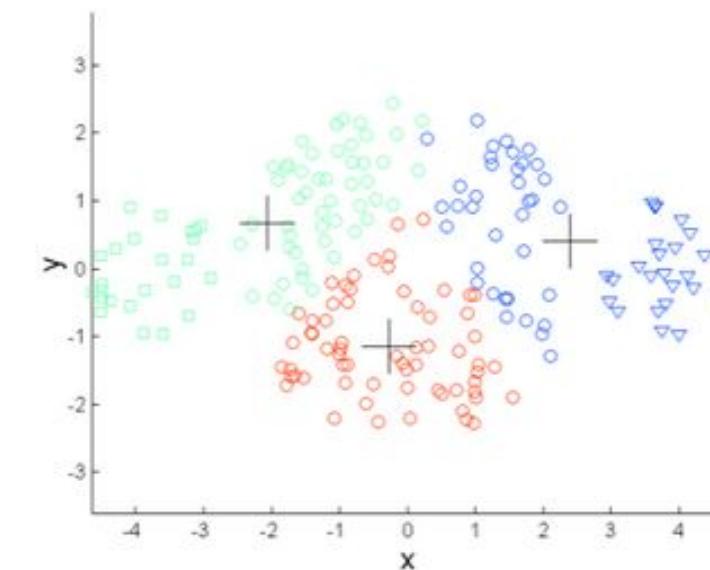
## ¿Qué puede salir mal?

- Desventajas de K-Medias:
  - Intenso computacionalmente
  - Cada observación corresponde a un sólo grupo
  - Muy sensible a observaciones atípicas
  - No puede modelar relaciones complejas

# Cuándo NO usar K-Medias

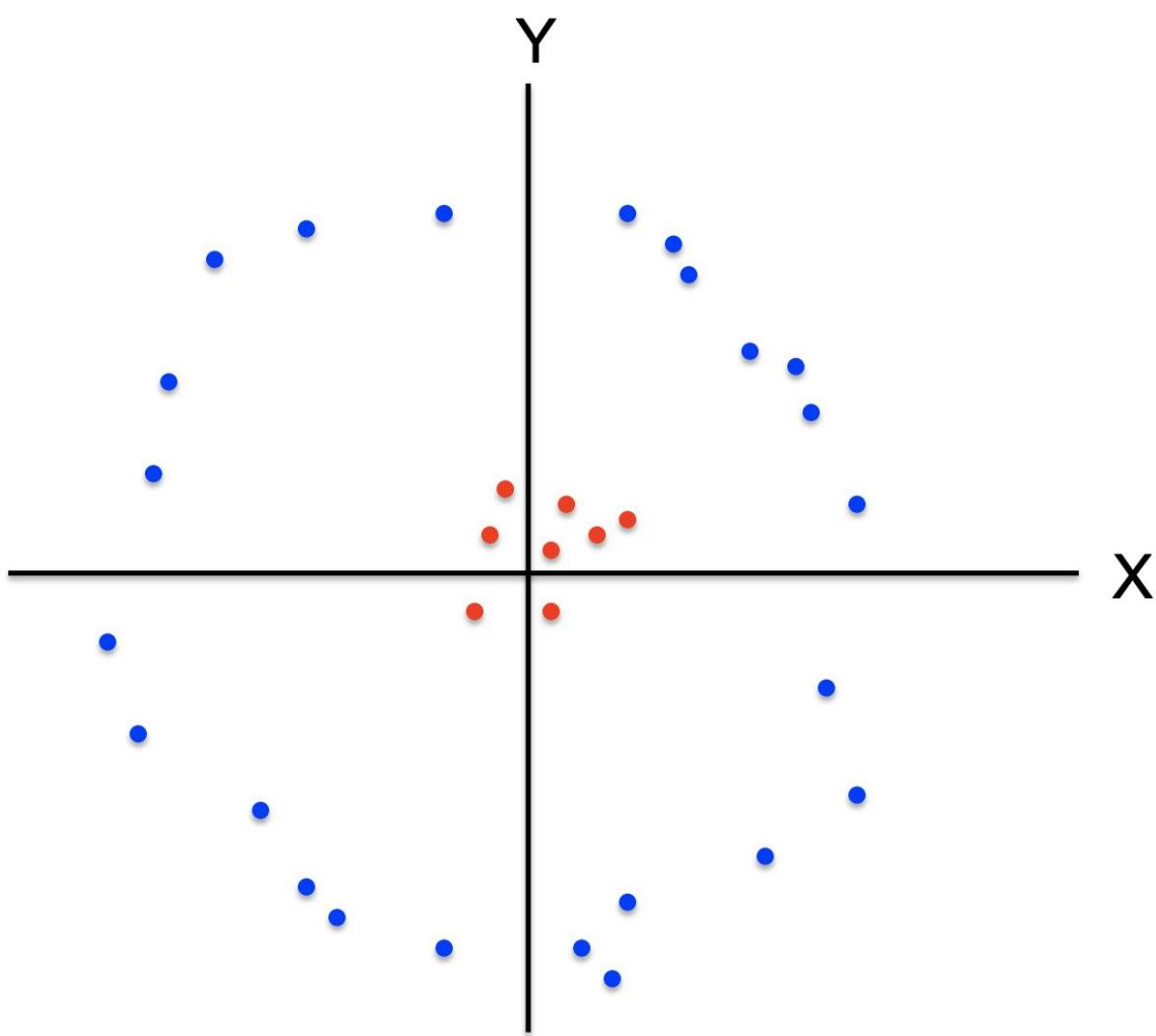


Original Points

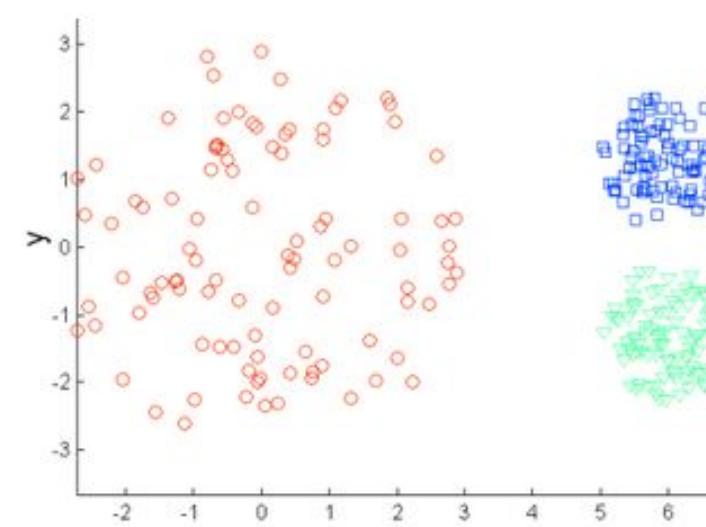


K-means ( $k = 3$ )

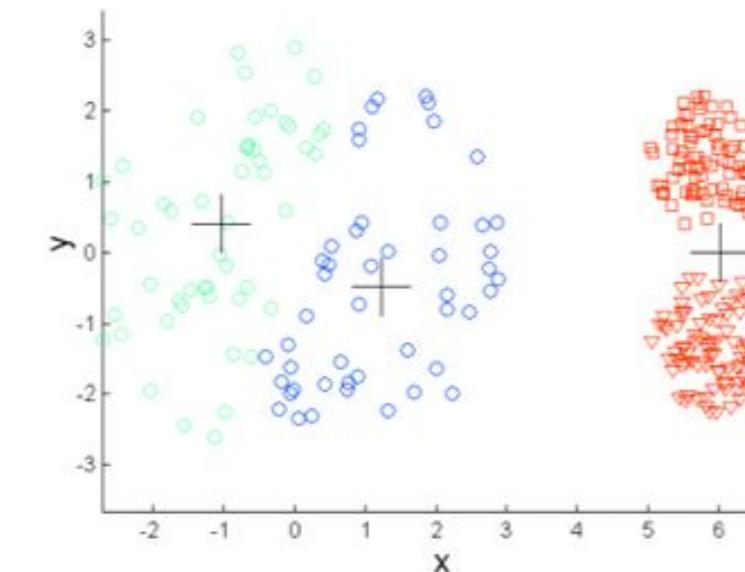
# Cuándo NO usar K-Medias



# Cuándo NO usar K-Medias



Original Points



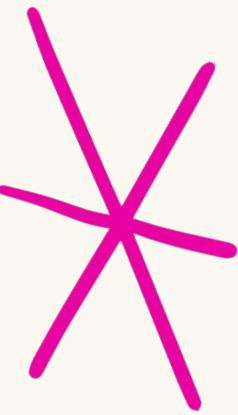
K-means ( $k = 3$ )



Notebook de hoy

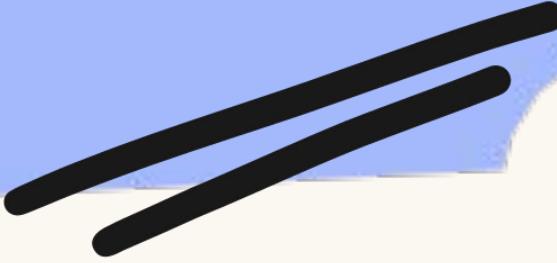
[https://colab.research.google.com/drive/1CnrDudTWnrHlU\\_WhOPUWBLcEz3zLNmnG?usp=sharing](https://colab.research.google.com/drive/1CnrDudTWnrHlU_WhOPUWBLcEz3zLNmnG?usp=sharing)





# Agrupación

(Taller # 11)



Fecha de entrega: Octubre 21, 2024

Aplicar K-Means a un conjunto de datos de su elección.

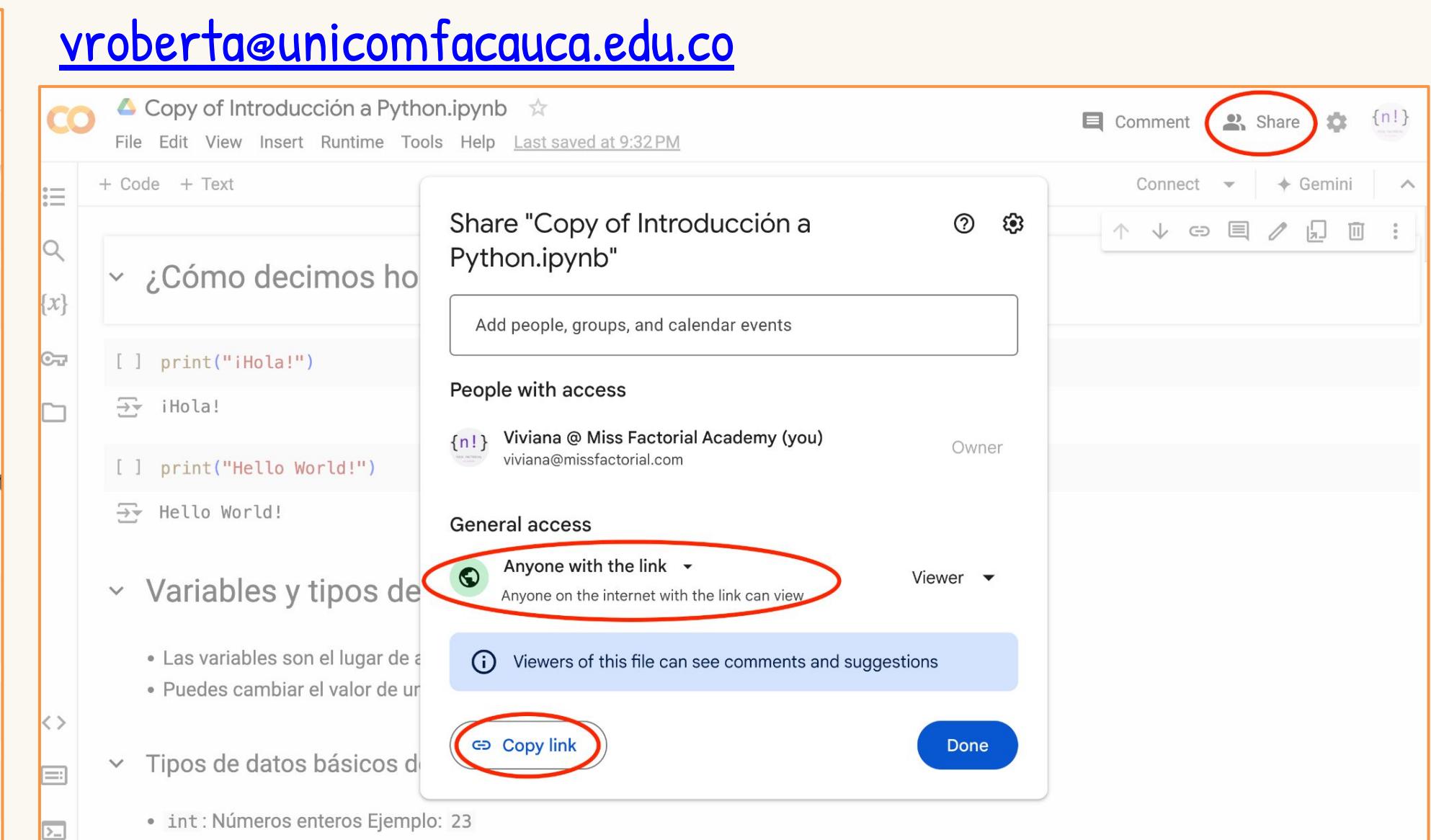
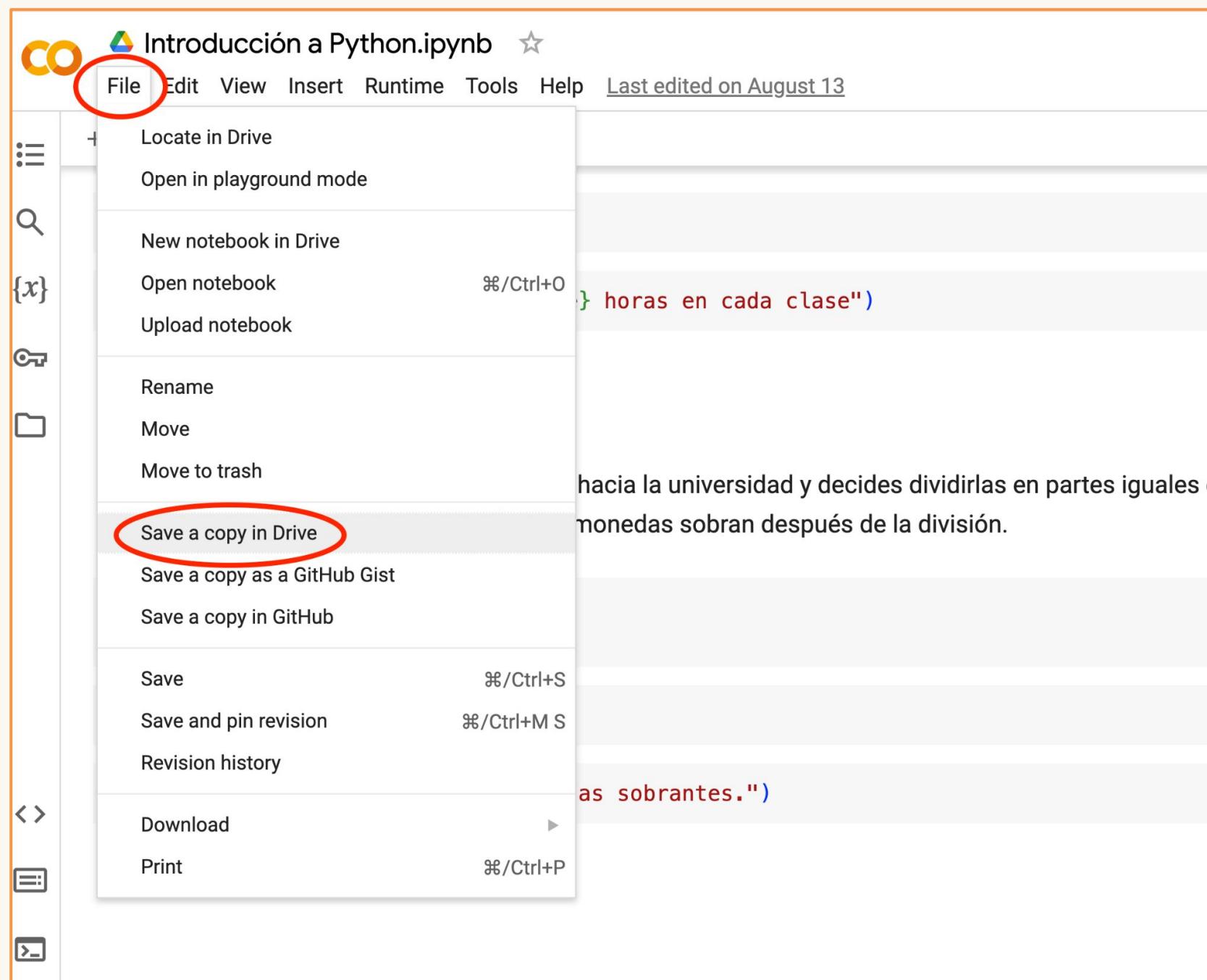
Por favor NO usar ninguno de los datos que ya hemos usado durante el semestre en clase.

Incluir:

1. Objetivo
2. Análisis Exploratorio de Datos
3. Ingeniería de características de los datos (sólo si es necesario para los datos que escogiste)
4. Usar el método del codo para escoger K
5. Entrenar el modelo usando ese K
6. Visualización (sólo si se tienen dos características)
7. Recomendaciones de negocio

# Para enviar los talleres de código

- ❑ Hacer click en **archivo** → **guardar copia en mi Drive** para que les quede una copia en su cuenta, de lo contrario, los resultados no serán guardados.
- ❑ En la copia creada, hacer click en **compartir** , asegurarse que el enlace sea visible a **cualquier persona** , copiar el enlace y enviarlo.

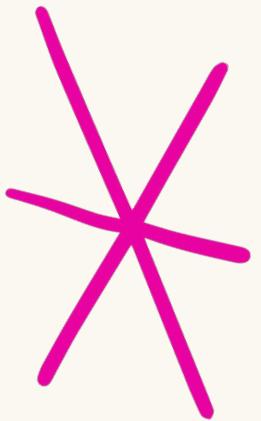


## Socialización de notas segundo corte:

- A lo largo del semestre les ido respondiendo a cada taller con su nota, si no han recibido esto, por favor contactárme.
- El próximo martes le enviaré a cada uno de ustedes un email con sus notas del segundo corte y el número de asistencias. Por favor revisar el correo institucional.
- El último día para enviar reclamos y/o aclaraciones es el 18 de octubre.

Taller #	Descripción	Enlace	Fecha de entrega	Porcentaje en el segundo corte	Porcentaje en el curso
Taller # 6	Adquisición de datos (opcional)	<a href="#">Enlace</a>	Septiembre 16, 2024	0%	0%
Taller # 7	Exploración de datos (mandatorio)	Diapositiva 35	Septiembre 23, 2024	25%	7.5%
Taller # 8	Ingeniería de características (mandatorio)	Diapositiva 46	Septiembre 30, 2024	25%	7.5%
Taller # 9	Regresión Lineal (mandatorio)	<a href="#">Enlace</a>	Octubre 7, 2024	25%	7.5%
Taller # 10	Modelos de supervisión (mandatorio)	<a href="#">Enlace</a>	Octubre 14, 2024	25%	7.5%

# ¡Gracias!



¿Dudas? Email de la profe:

[vroberta@unicomfacauba.edu.co](mailto:vroberta@unicomfacauba.edu.co)

Página web del curso con toda la info:

<https://github.com/vivianamarquez/unicomfacauba-ai-2024>