

# Unsupervised learning

The background of the slide is a close-up, low-angle shot of a complex electronic device, possibly a custom-built computer or a specialized piece of hardware. It features a dense arrangement of cables, connectors, and components. The scene is dramatically lit with a mix of cool blue and warm red/purple light, creating a futuristic and technical atmosphere. The lighting comes from above and the sides, casting strong highlights and deep shadows on the various parts of the device.

26 Octubre 2019

# Indice

Repaso General 1er mes

Unsupervised Learning

Diferencias con Supervised Learning

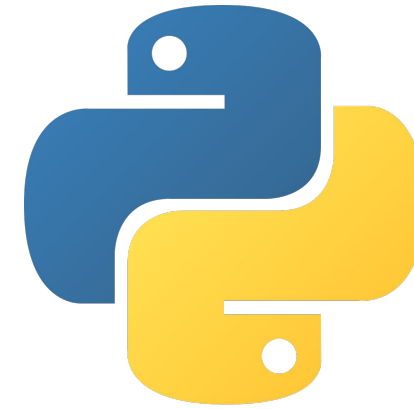
Tipos

K-Means ( Ejemplo Práctico )

PCA ( Ejemplo Práctico )



# Repaso Python & Colab

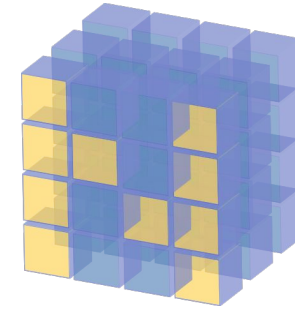
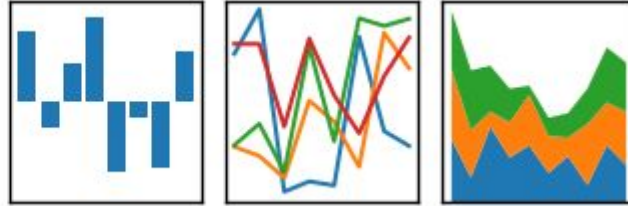


colab

# Repaso Numpy, Pandas & Matplot

pandas

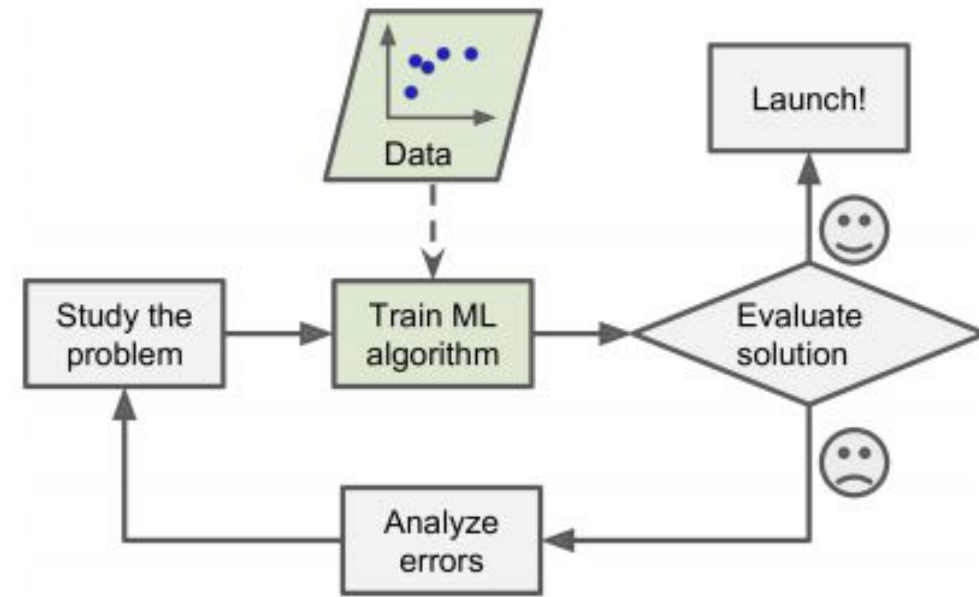
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



NumPy

matplotlib

# Repaso Machine learning



Preguntas de investigación

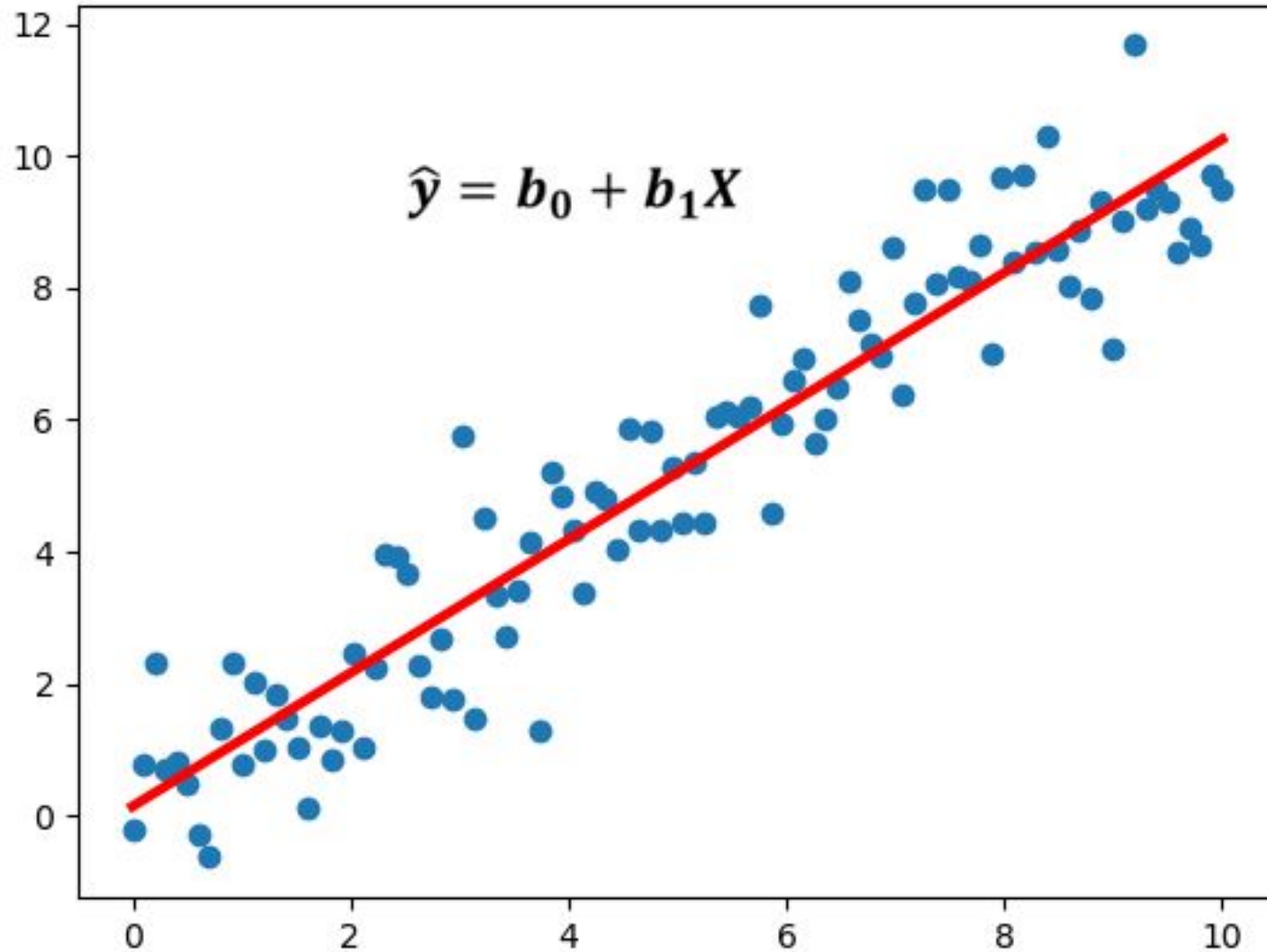
Limpiar datos

Exploración de datos

Construir modelos

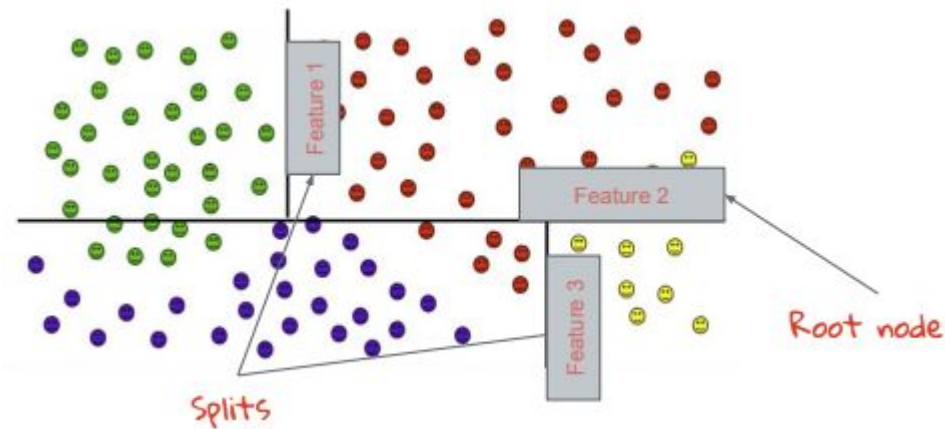
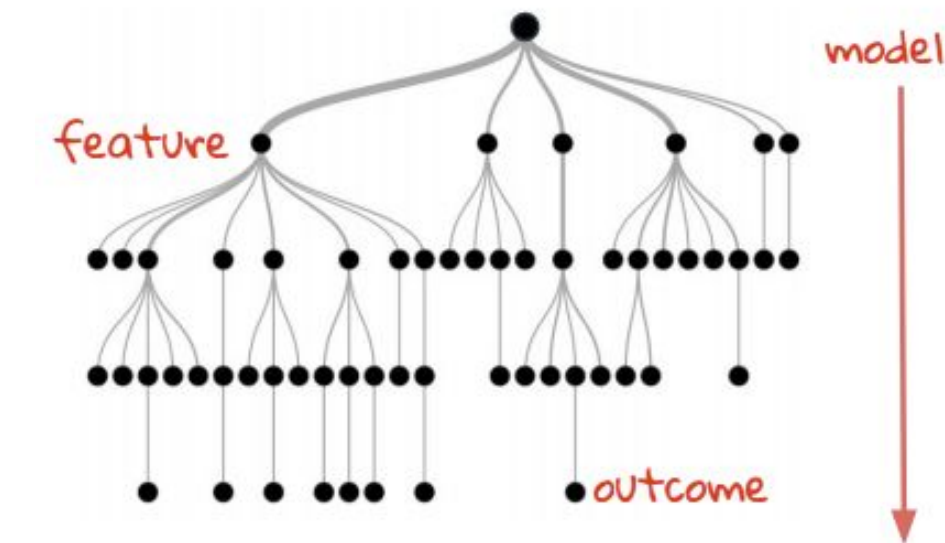
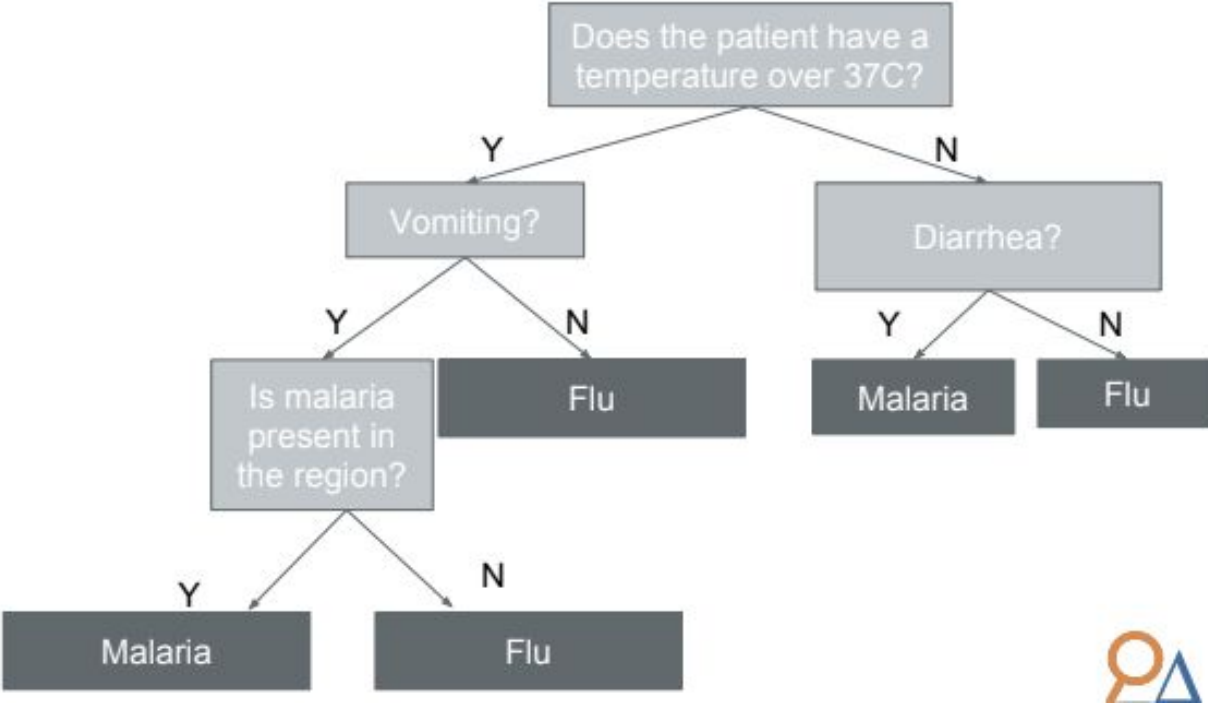
Performance

# Regression lineal

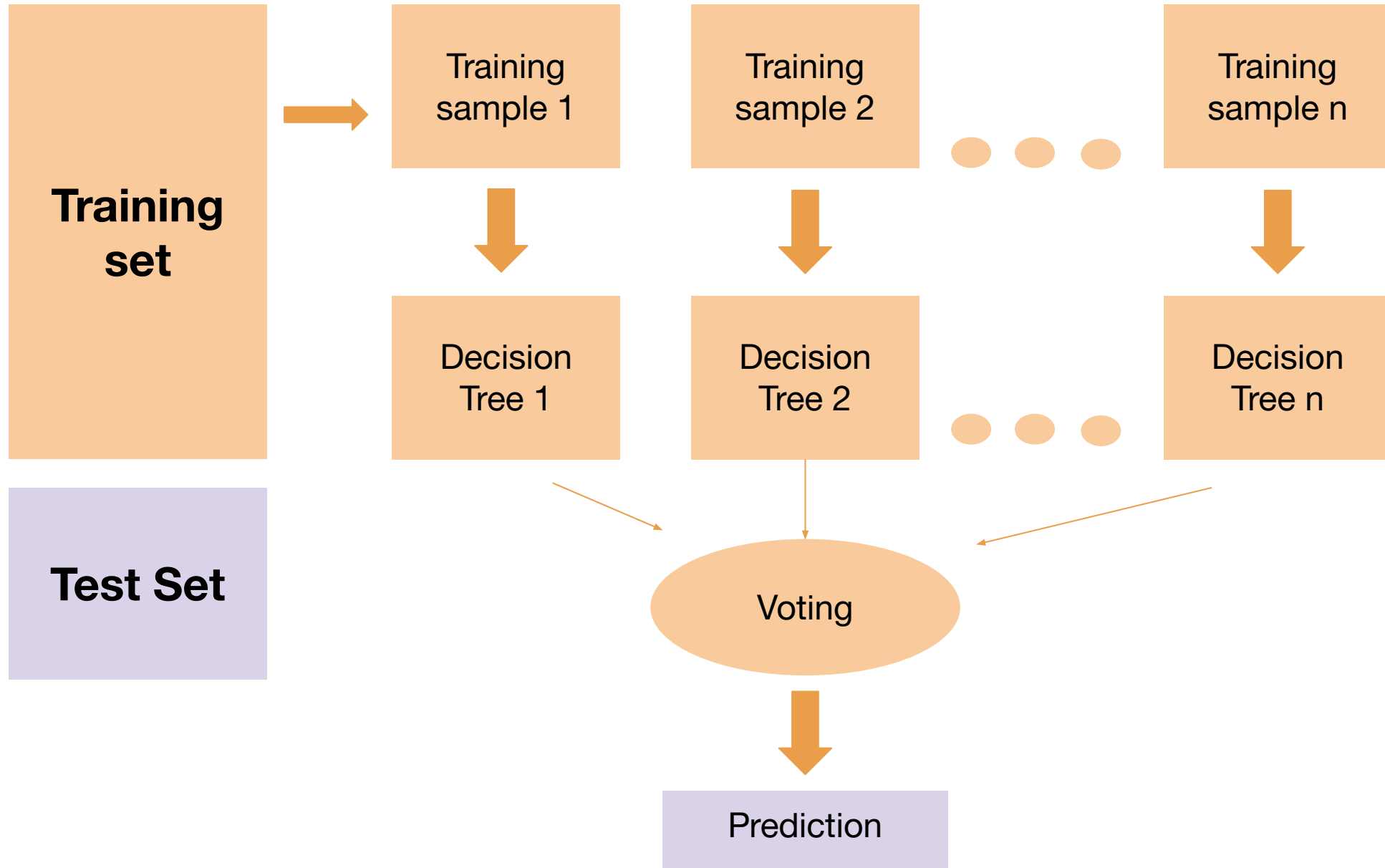


$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

# Árboles de decisión



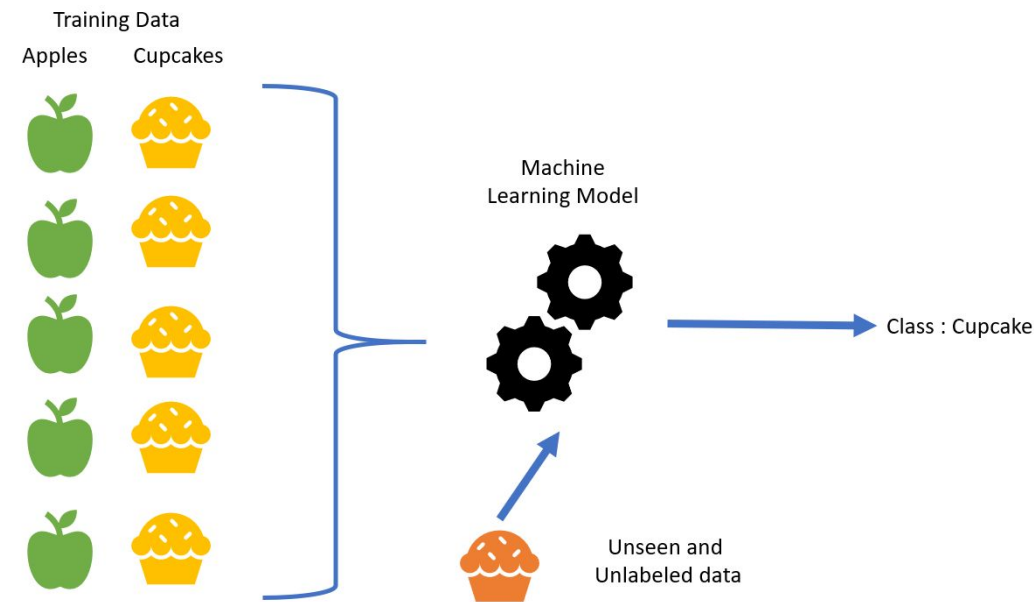
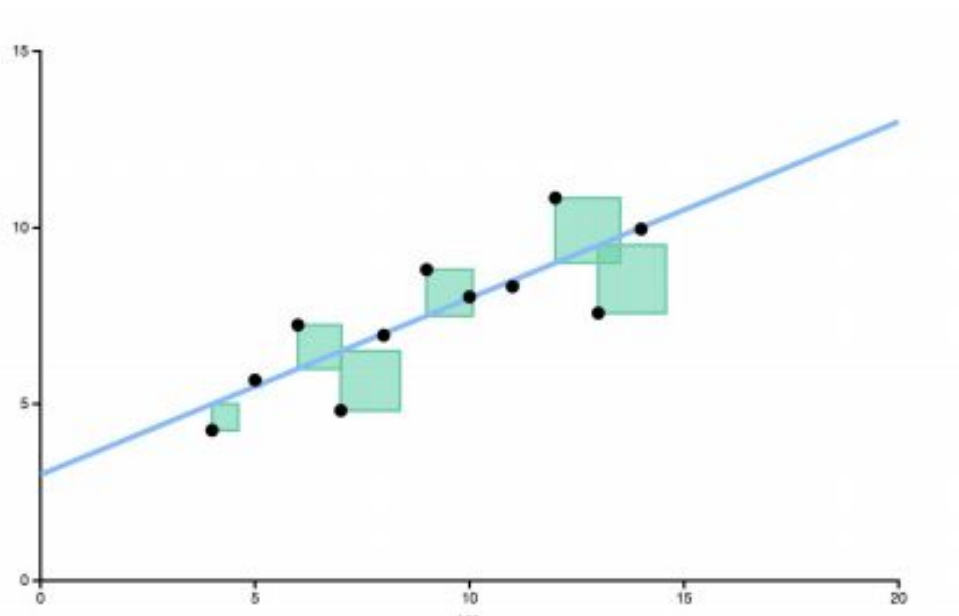
# Random Forest





# Supervised Learning

| Regresión          | Clasificación    |
|--------------------|------------------|
| Predecir un número | Predecir un tipo |



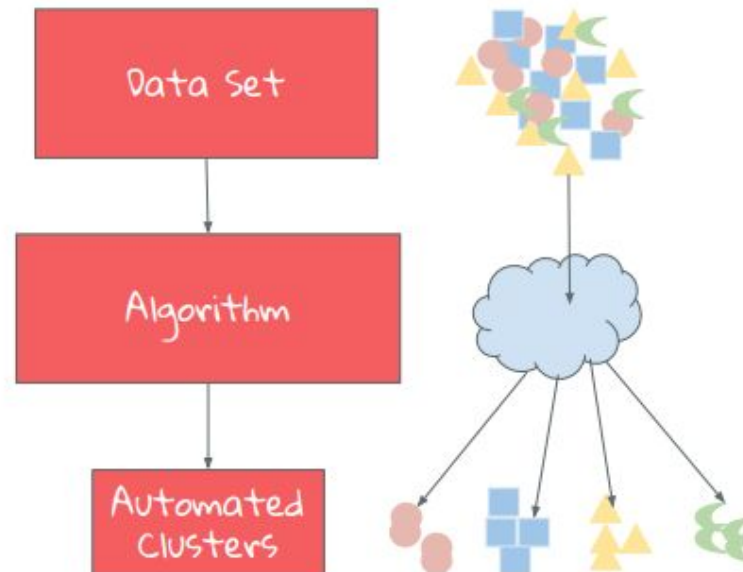
# Supervised vs UnSupervised Learning

## Supervised Learning:

- Por cada  $x$  hay una  $y$
- El objetivo es predecir  $y$  usando  $x$
- Muchos métodos que usamos en la práctica son supervised

## Unsupervised Learning:

- Por cada  $x$  no hay  $y$
- El objetivo no es predecir sino investigar  $x$
- Los métodos de unsupervised leen los datos antes y después sugieren que esquema de clasificación se puede aplicar



# UnSupervised Learning

**Task**

**Cual es el problema que el modelo quiere solucionar?**

**Que enfoque de Unsupervised queremos usar?**

**Por ejemplo, es el clustering jerárquico mejor que el K-means mejor para esta tarea?**

**Feature engineering & selection**

**Como decidimos que catacteristicas incluir en nuestro modelo?**

# UnSupervised Learning

**Metodología de  
aprendizaje**

**Los algoritmos de clustering son  
unsupervised, cómo esto afecta  
nuestro modelo?**

**Cómo aprenden los  
modelos de ML?**

**Cómo el modelo aprende solo**

**Proceso de optimización**

**Como los modelos de unsupervised  
learning se optimizan si no hay  
función de error?**

# Usos de UnSupervised Learning

- Datos con muchas dimensiones
- Detectar relaciones o patrones en nuestros datos
- En la parte de exploración, antes de un algoritmo supervisado
- Investigación de datos sin label
- Necesidad de información urgente



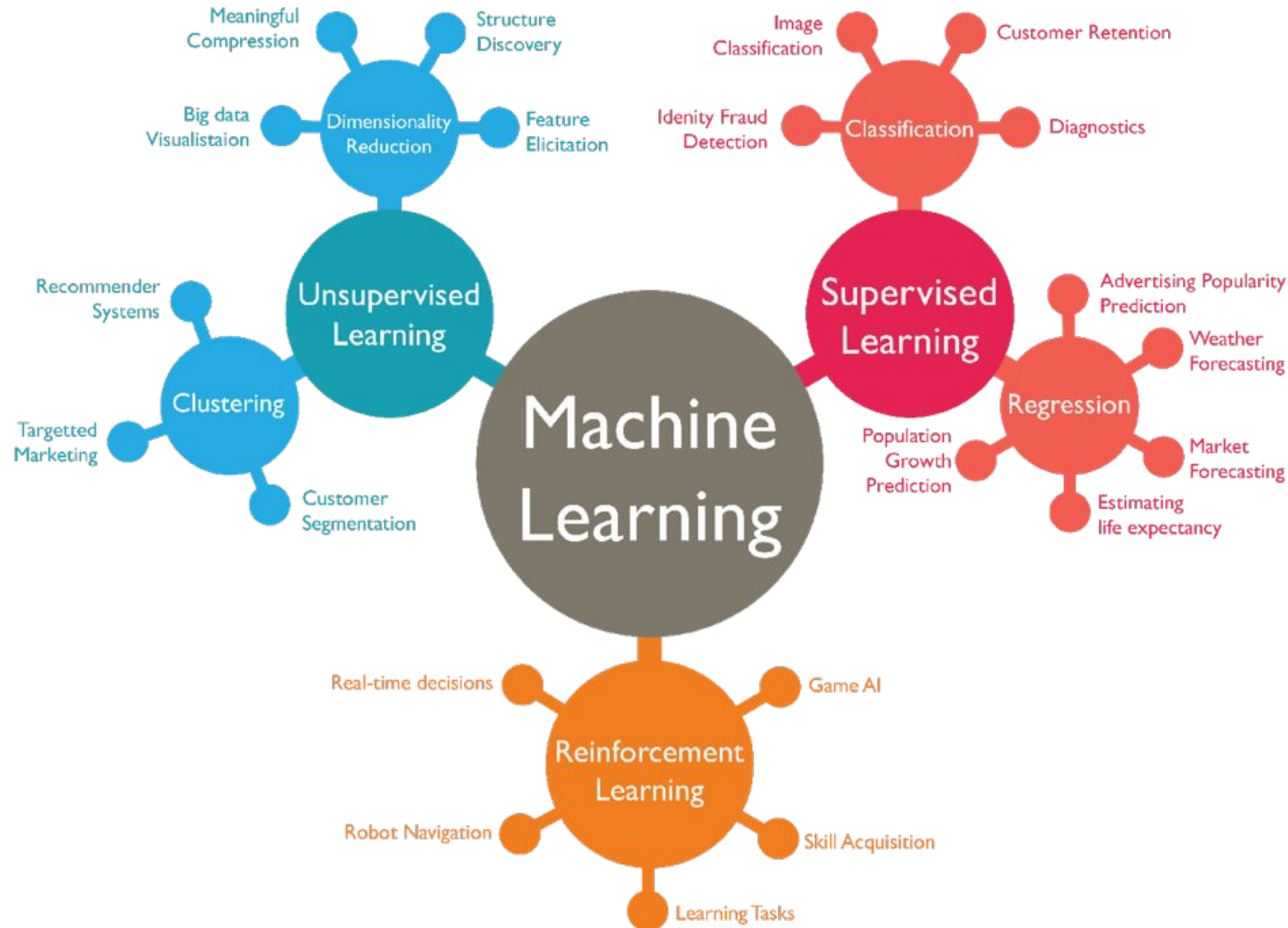
# UnSupervised Learning: el futuro del ML



Supervised learning es la guinda del pastel

UnSupervised learning es el pastel mismo

# Tipos de Unsupervised learning

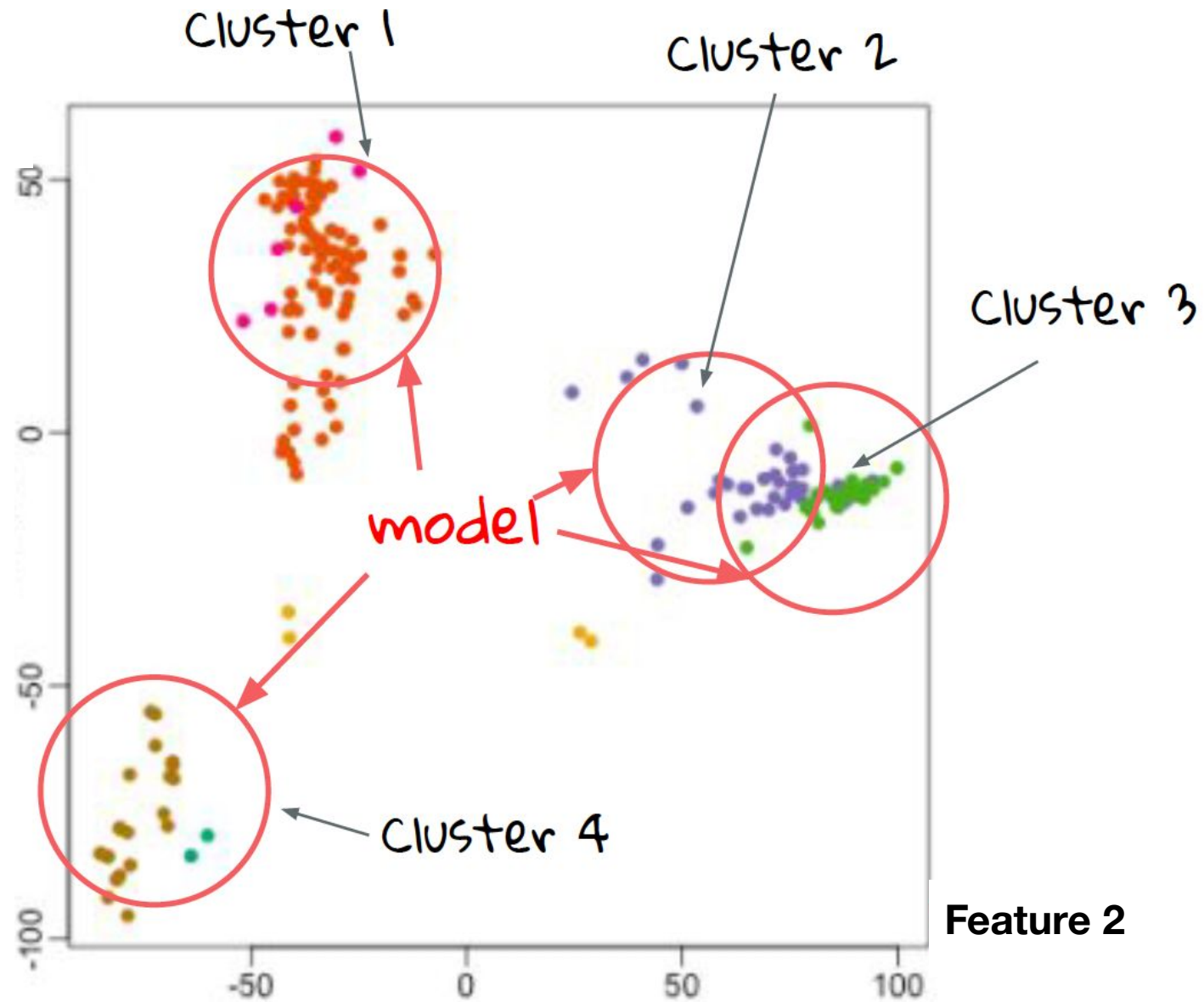


# Ejercicio supervised vs unsupervised

- Predecir el precio de un diamante en base a sus características
- Reducir features de un dataset de una fábrica
- Agrupar jugadores de baloncesto
- Detectar imagenes
- Categorizar artículos de un periódico

# Clustering

Feature 1



# Pros y Cons

## Pros

- Fácil de representar
- No asume distribuciones subyacentes
- Produce grupos intuitivos
- Puede ser usado en muchas dimensiones

## Cons

- Se pierde mucho tiempo en buscar en número optimal de clustering
- El future engineering lleva mucho tiempo, las características tienes que ser numeradas y normalizadas)

## Assumption

- Asume la existencia de grupos subyacentes





# Ejemplo

Clustering task

Definir los grupos

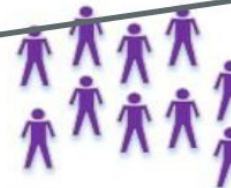


Dataset of  
customer  
features



???

???



???

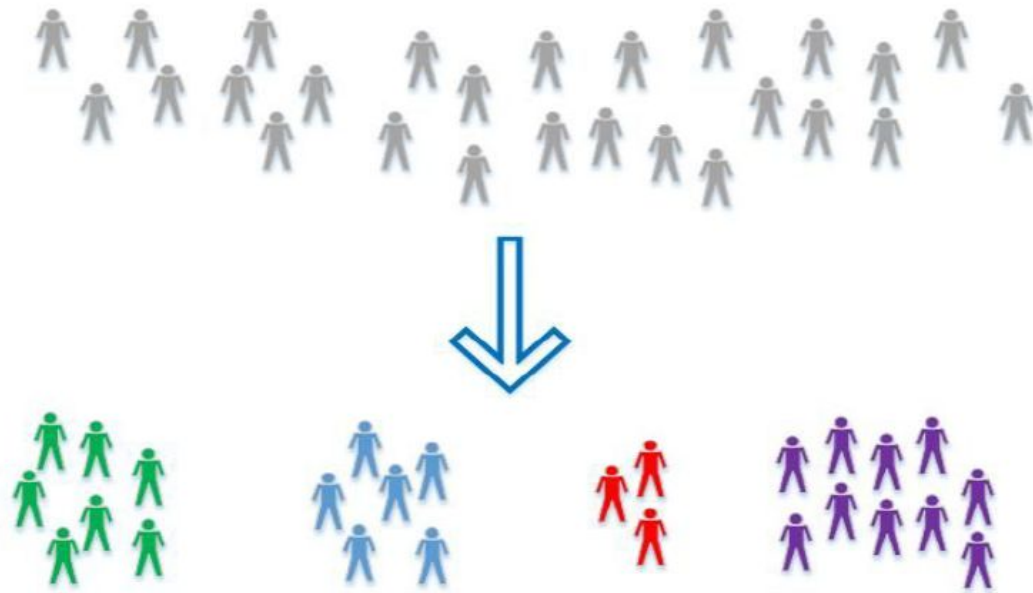
???



Saturdays.AI

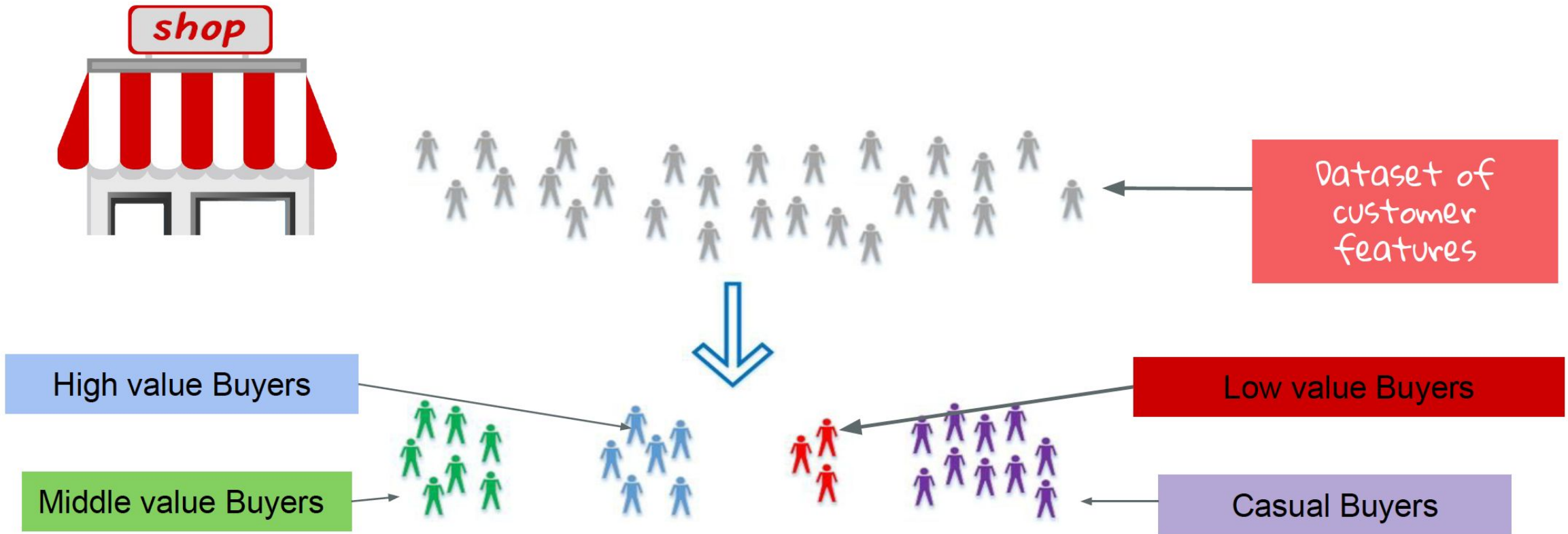
# Que datos podemos sacar?

Dataset of  
customer  
features

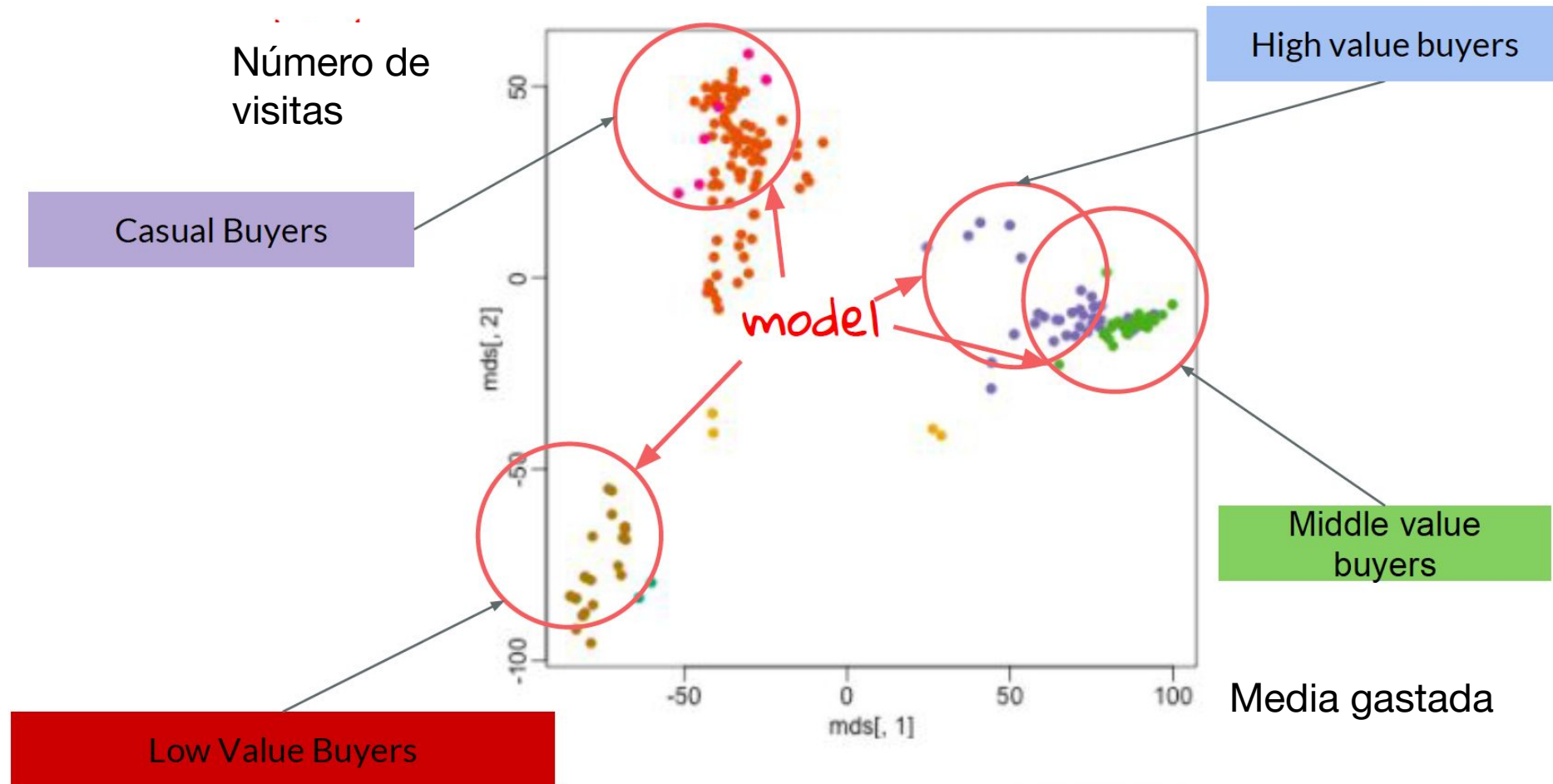


- Media de \$ gastado
- Numero de visitas
- Tiempo medio de visita
- Tipo de tráfico
- % de visita

# Qué características son relevantes para nosotros?



# Representación de datos

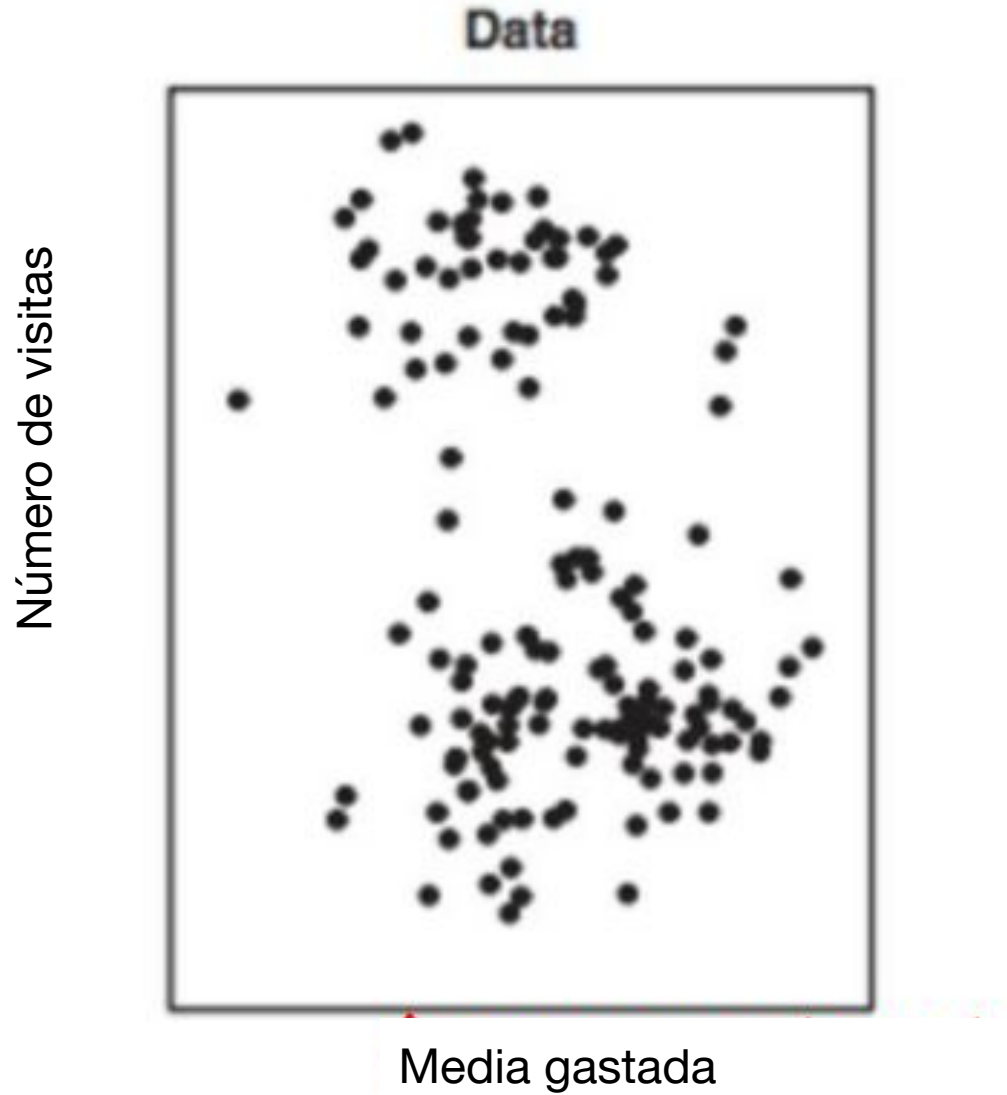


# Definir nuestros clientes



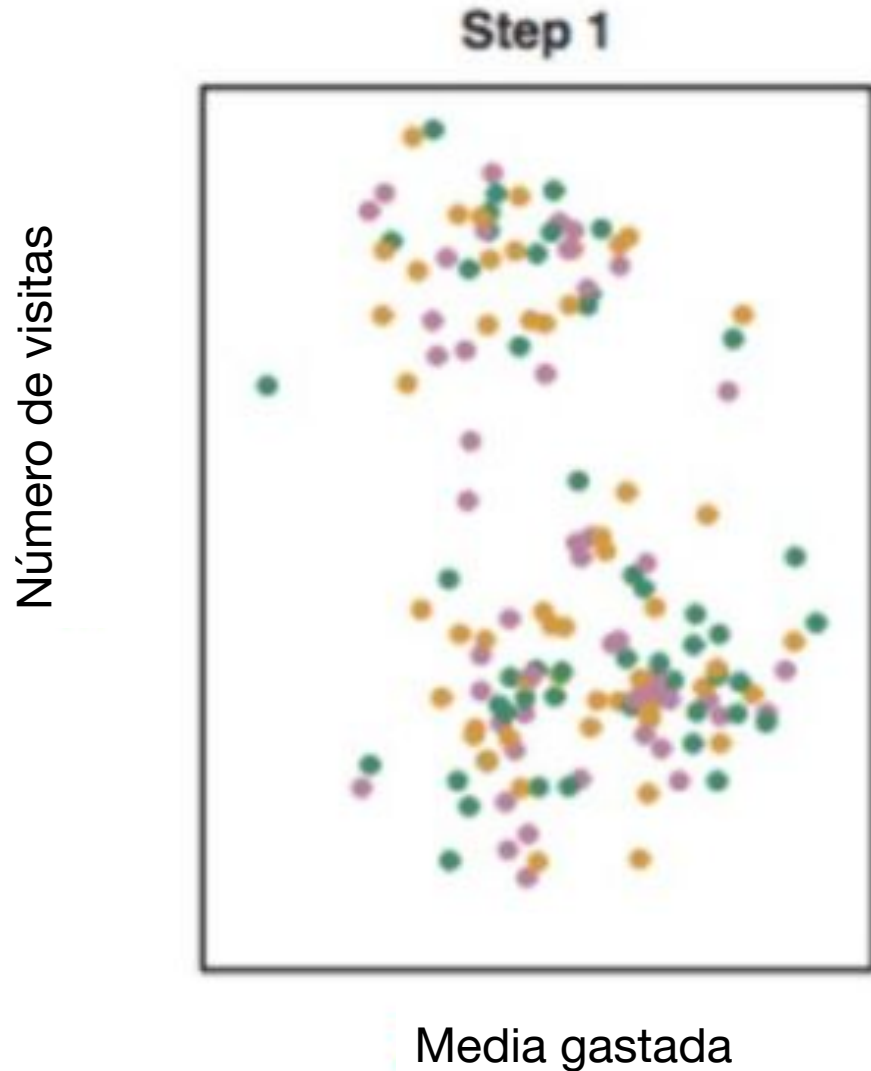


# Como aprende el modelo de ML?



$K = 3$

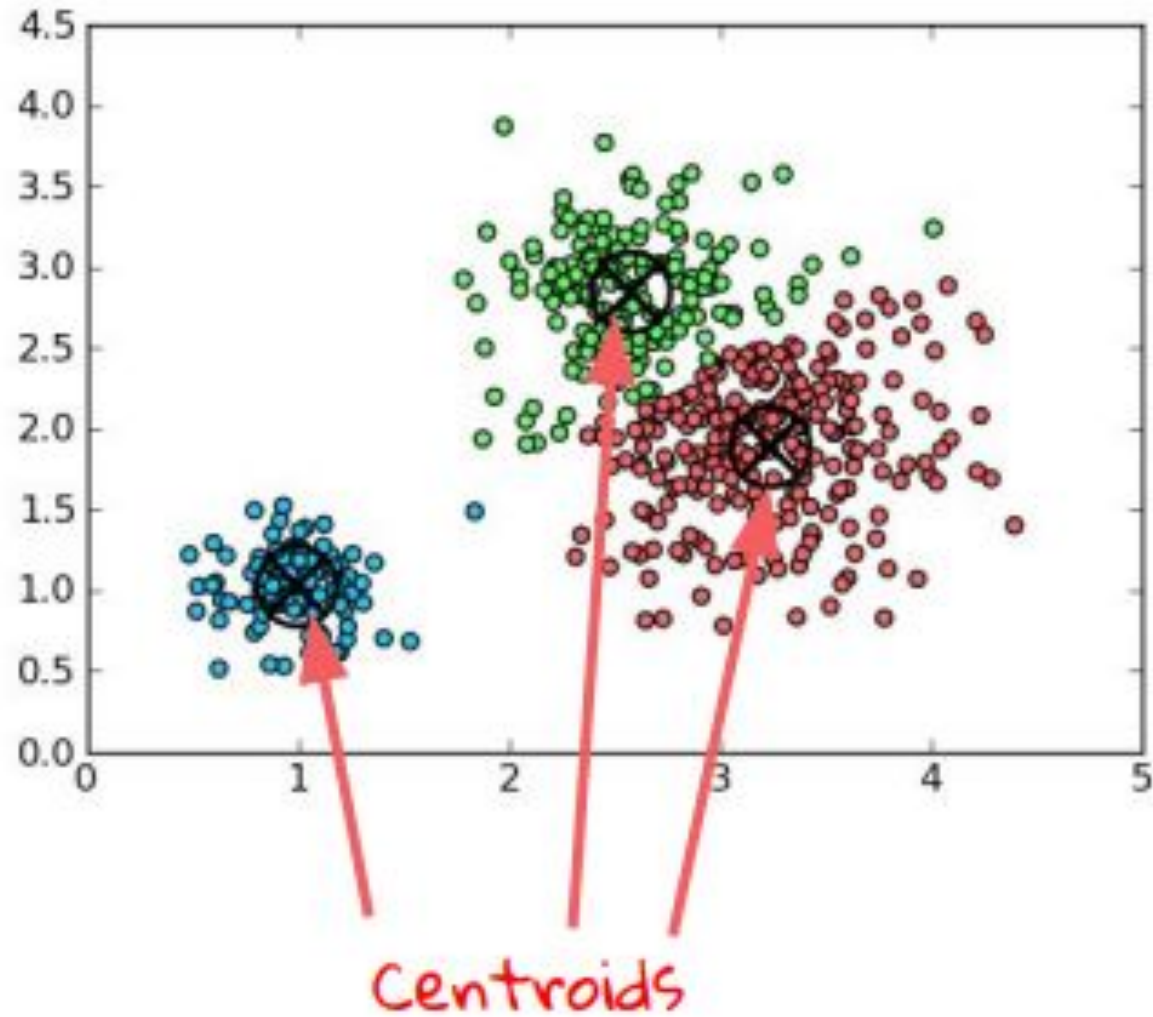
# Step 1



## 3 grupos de clientes:

- Gente que hace compras pequeñas pero constantes
- Gente que hace compras grandes frecuentes
- Gente que hace pocas compras

# Definición de centroid

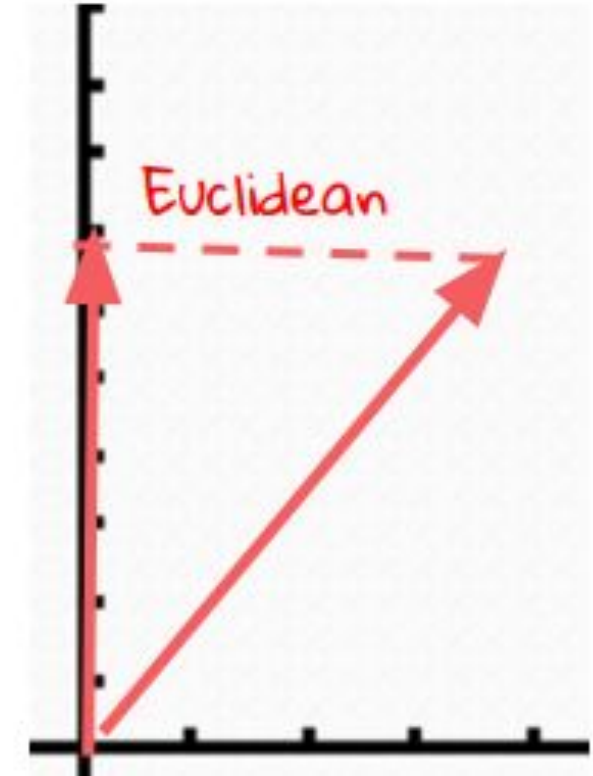


# Tipos de Unsupervised learning

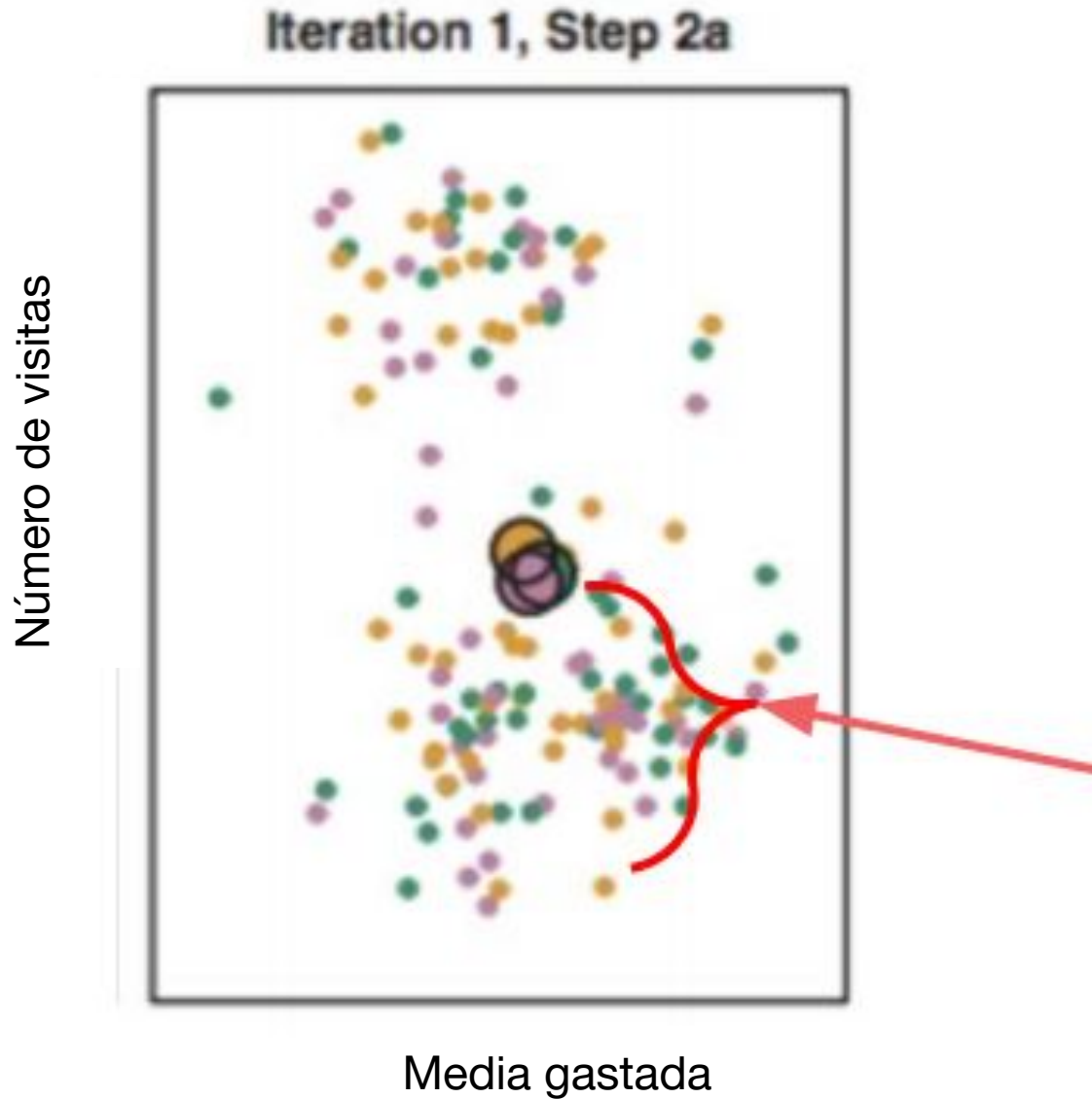
- La distancia es la distancia Euclidiana, donde la distancia entre 2 vectores U y V es:

$$d = \sqrt{\sum_n (u_i - v_i)^2}$$

- En nuestro ejemplo, la diferencia entre el cliente c6 (7,2) y el centro del cluster (4,7) sería la raíz cuadrada de  $(7-4)^2 + (2-7)^2 = 5.8$



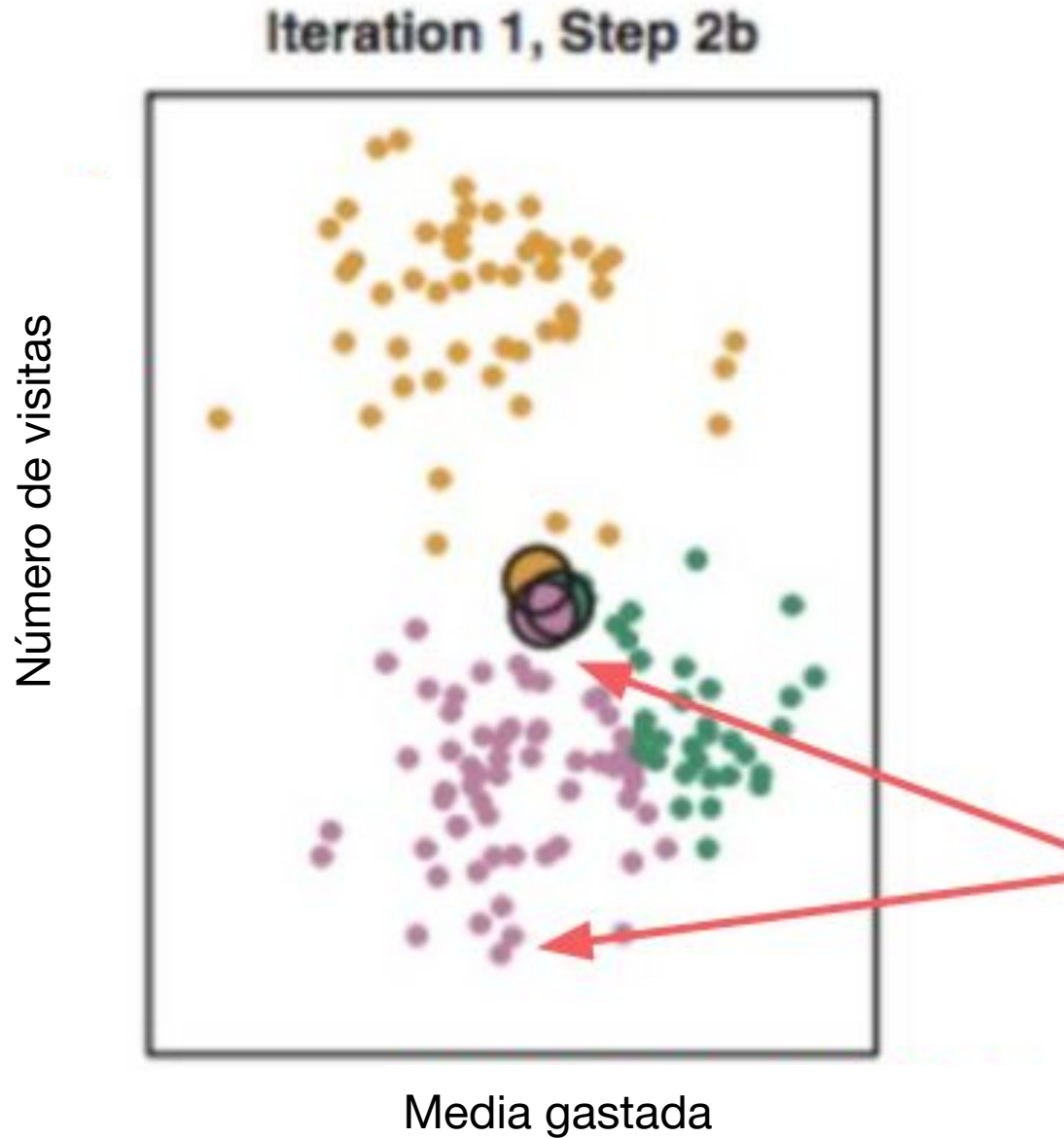
## Step 2



El objetivo es minimizar la distancia del centroide a cada una de las observaciones

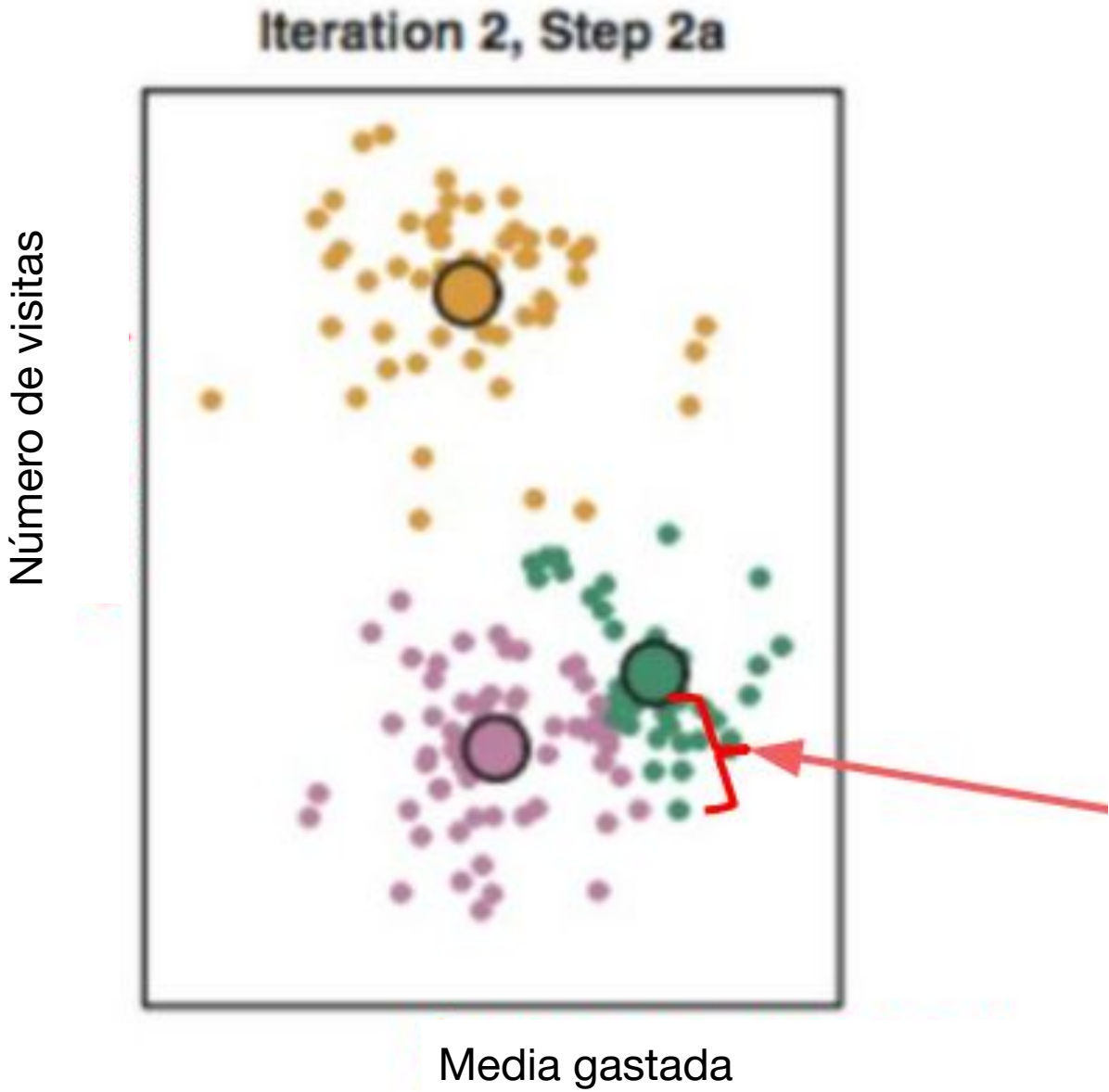


## Step 3



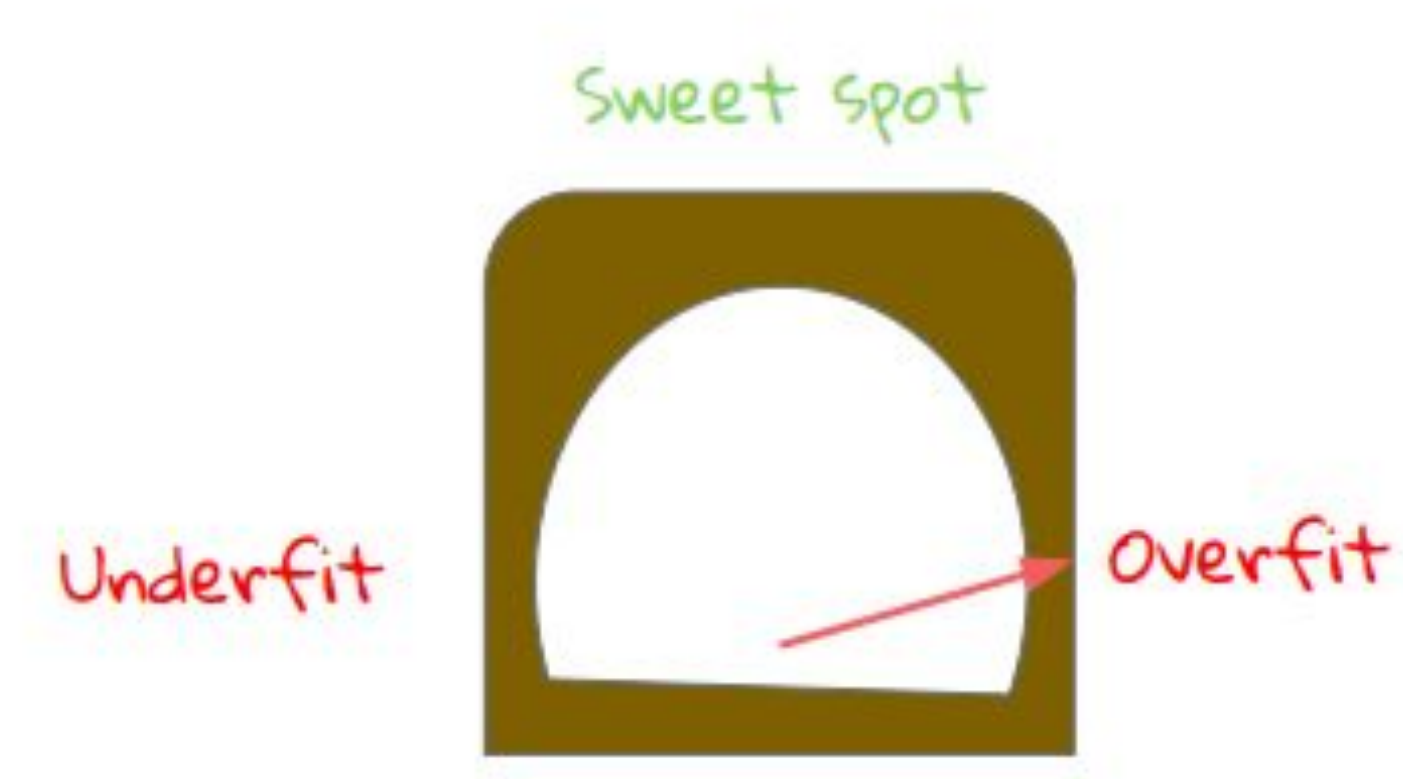
De los 3 centroides, la observación rosa es la más cerca, entonces lo asignamos a esta observación.

# Iterar

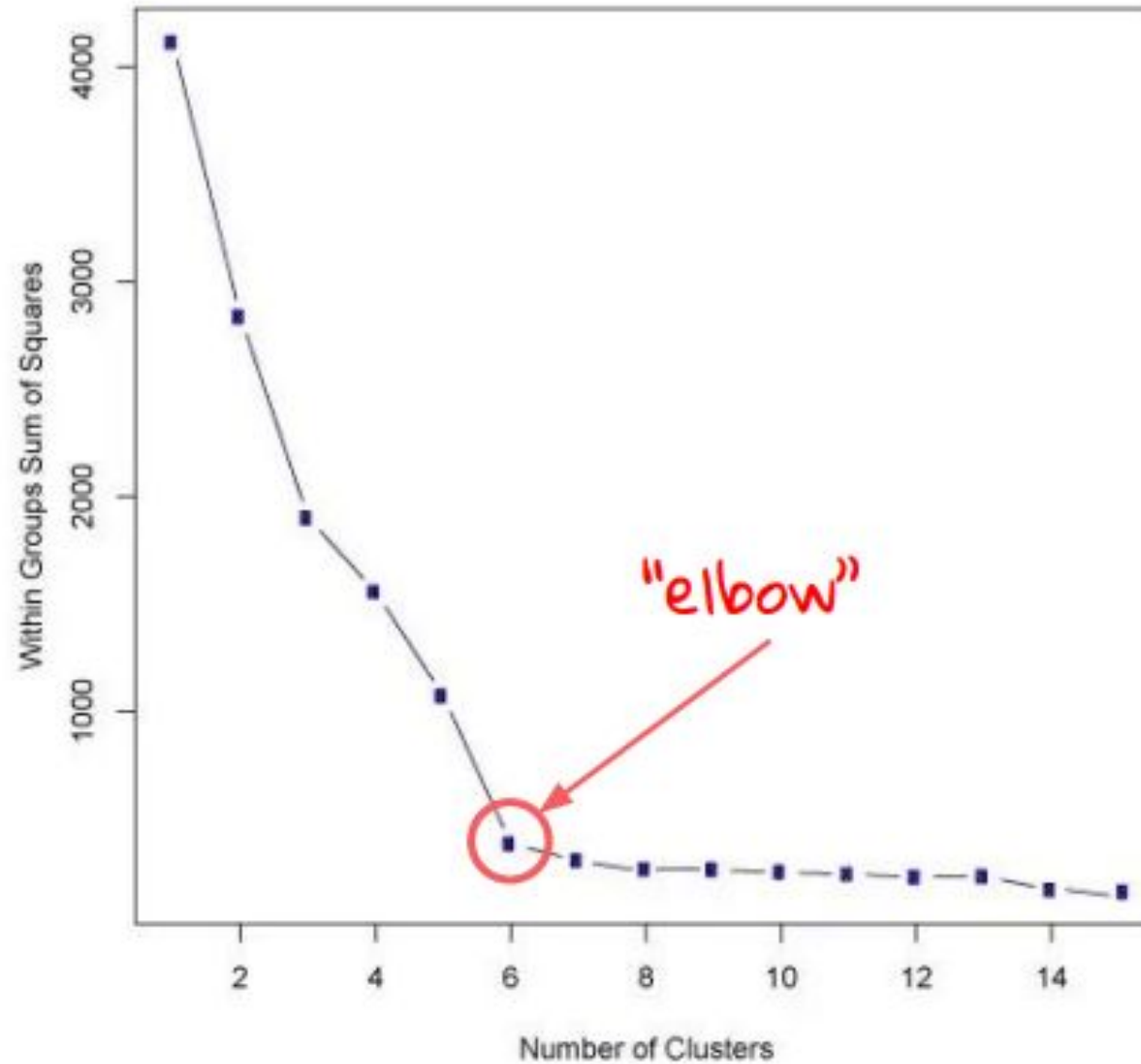


Ya las distancias se están acortando, nos estamos acercando!

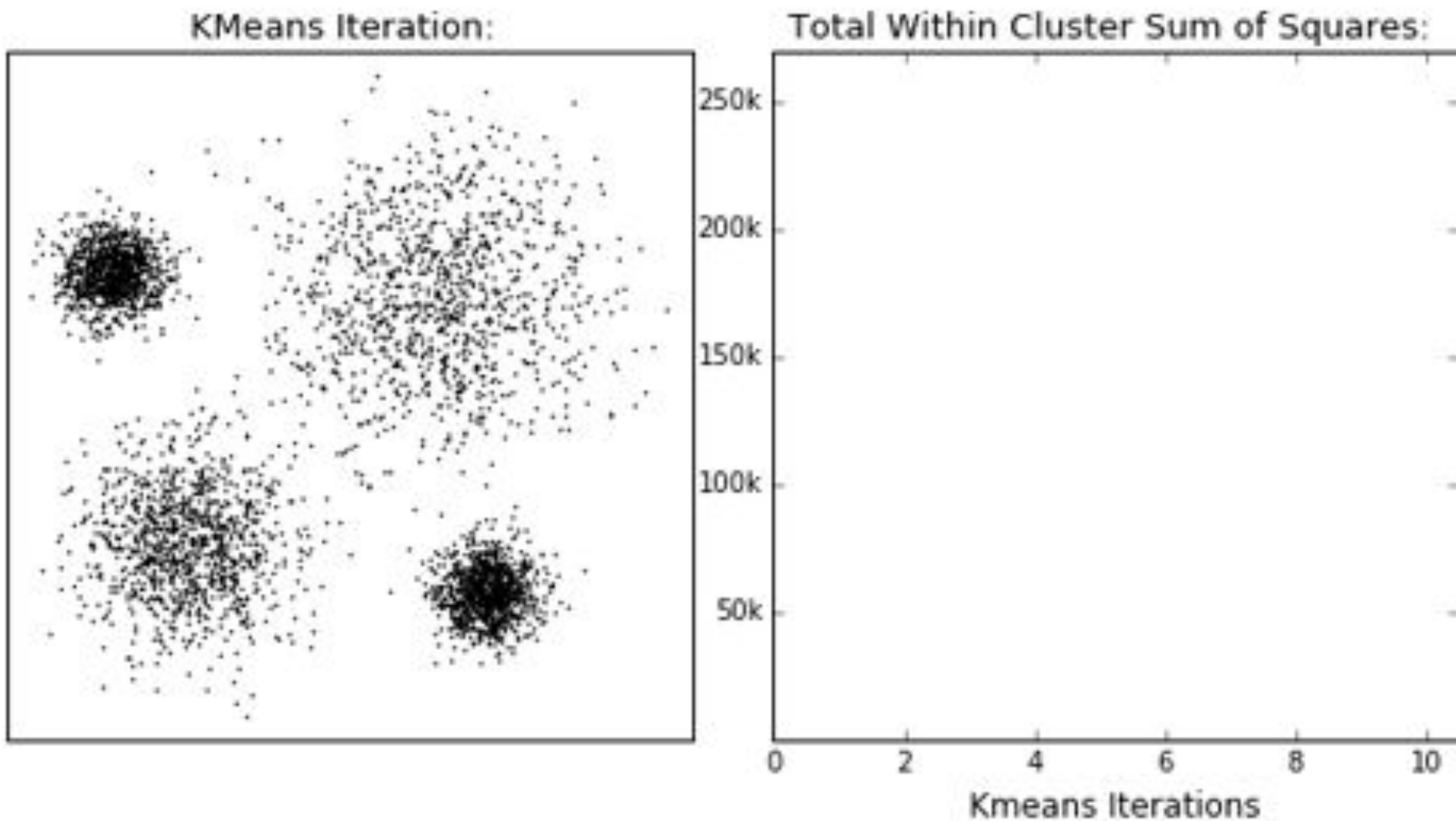
# Como elegir el número de clusters?



# Método de elbow



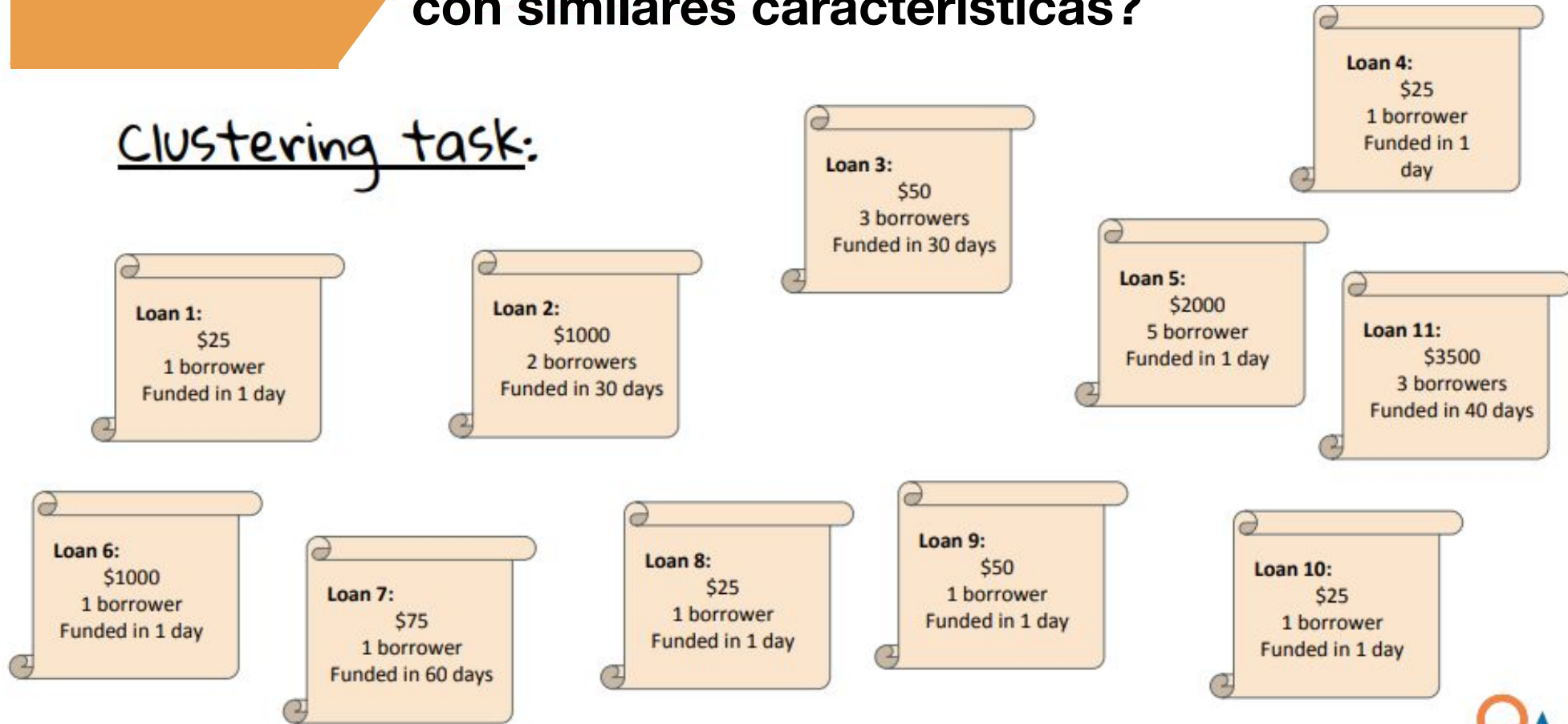
# Como funciona



# Ejercicio de clustering

Como clasificaremos estos préstamos entre grupos con similares características?

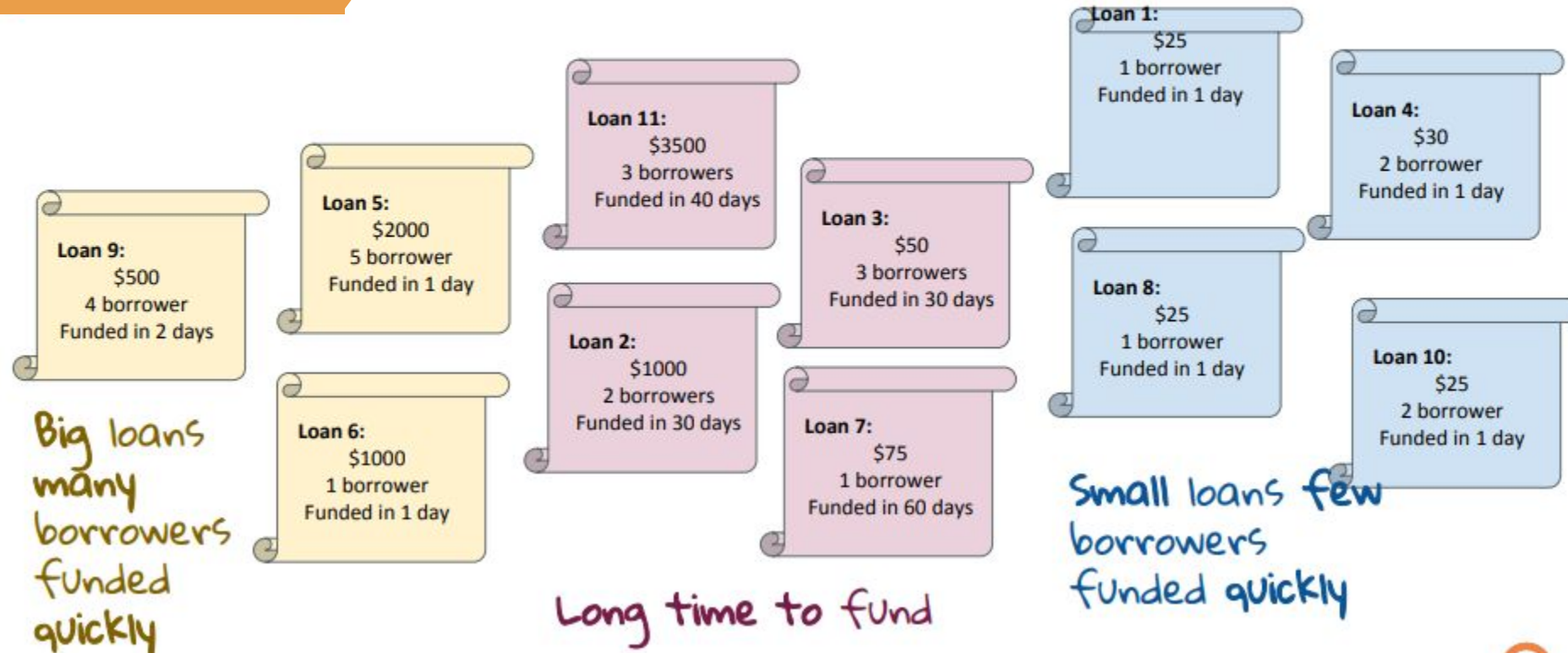
Clustering task:



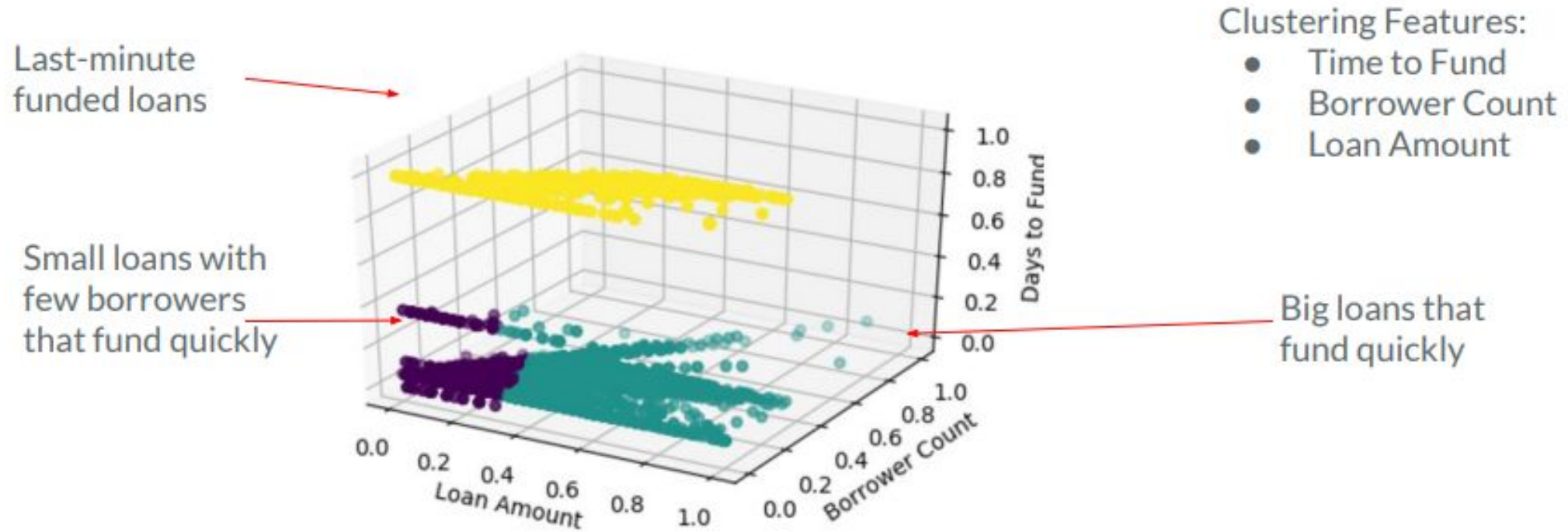


# Ejercicio de clustering

Podemos usar las características “tiempo del préstamo” y “tamaño del préstamo”

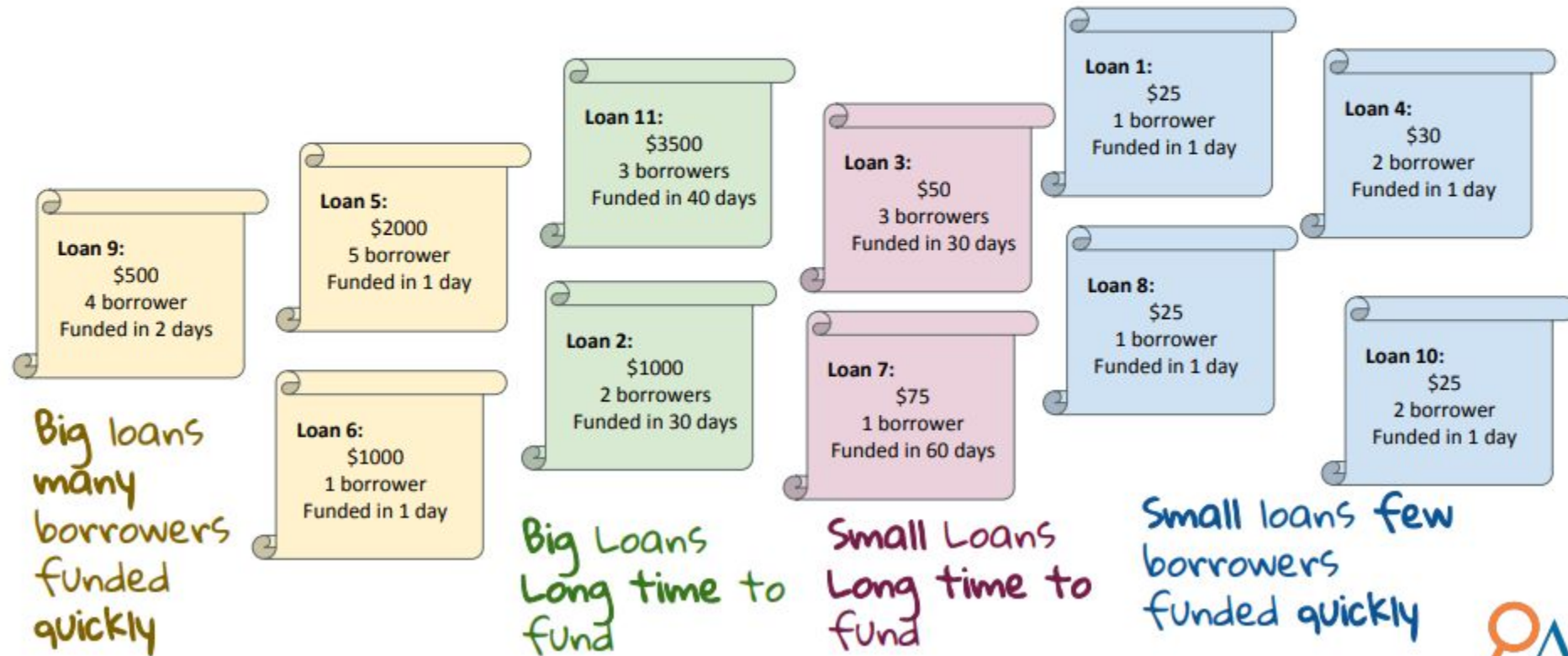


# Cluster con más dimensiones



# Ejercicio de clustering 2

Qué pasa si ponemos 4 grupos diferentes?



# Principal component analysis

- Algoritmo que nos ayuda a encontrar cuáles variables se relacionan entre ellas y así eliminarlas, dejándonos solo con las importantes.
- 
- Nos ayuda mucho en la parte de feature selection y engineering.

# Cómo PCA selecciona los datos

| Cumulative_GPA | Last_year_GPA | Test_scores | Attendance | TutoringNY |
|----------------|---------------|-------------|------------|------------|
| 3.55           | 3.65          | 89%         | 91%        | Y          |

Valores con más información

# Técnica de reducción de dimensiones

| # habitacione<br>s | Precio casa |
|--------------------|-------------|
| 1                  | 32000       |
| 2                  | 100000      |
| 4                  | 232000      |
| 2                  | 50000       |
| ....               | ....        |

| x    | y  | z | a        | ... |
|------|----|---|----------|-----|
| John | 31 | M | 21st ST. | ... |
| Jane | 42 | F | 3rd Ave  | ... |

**Cuando tenemos muchos datos, el PCA nos dice cuáles características son importantes.**

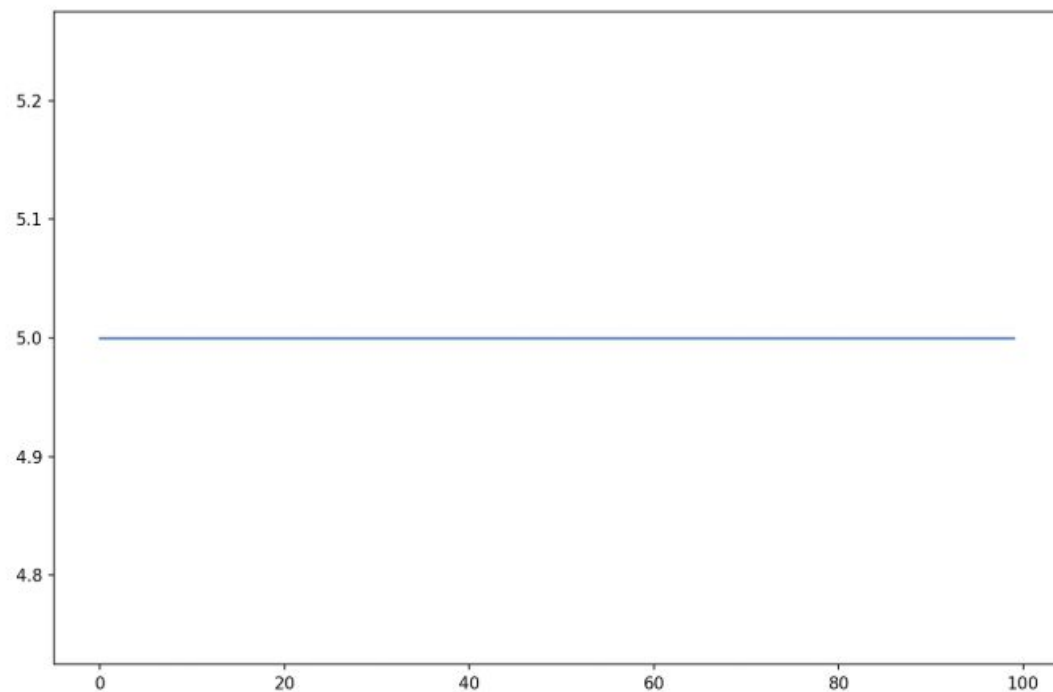




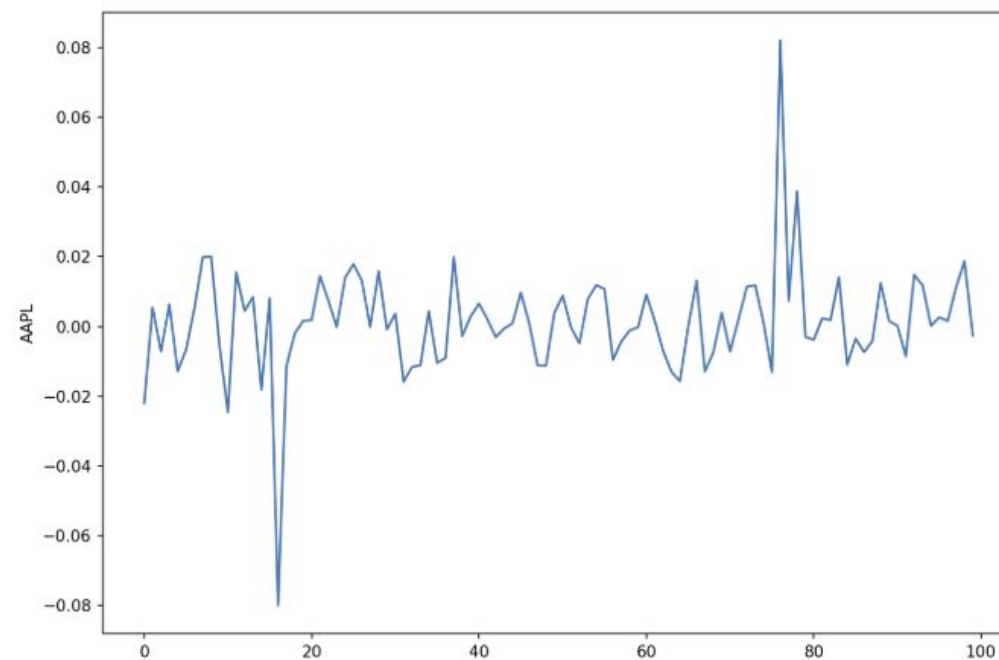
# Definir las variables importantes

| Cumulative_GPA | Last_year_GPA | Test_scores | Attendance | TutoringNY |
|----------------|---------------|-------------|------------|------------|
| 3.55           | 3.65          | 89%         | 91%        | Y          |
| 2.76           | 2.50          | 73%         | 90%        |            |

# Varianza




A Flat Line Has Zero Variance



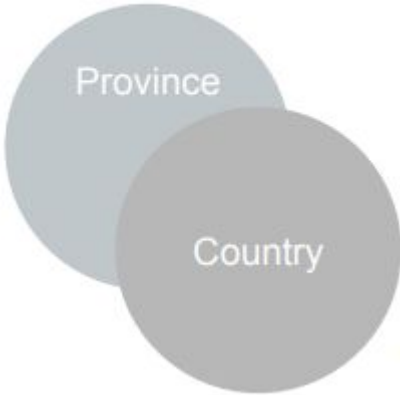
100 Days of AAPL Daily Returns



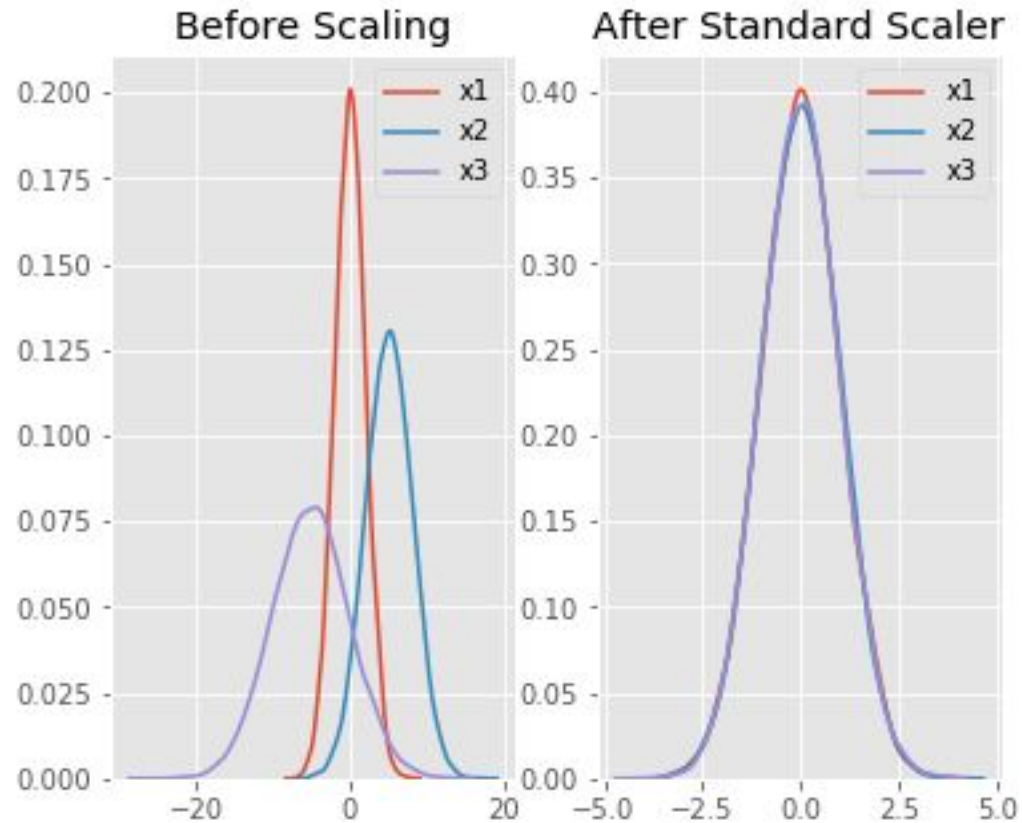
# Ordenar las características



| studentid | cumulative_gpa | last_year_gpa | test_scores | attendance | tutoringYN |
|-----------|----------------|---------------|-------------|------------|------------|
| 1         | 3.55           | 3.65          | 89%         | 91%        | Y          |
| 2         | 2.76           | 2.50          | 73%         | 90%        | Y          |
| ...       | ...            | ...           | ...         | ...        | ...        |

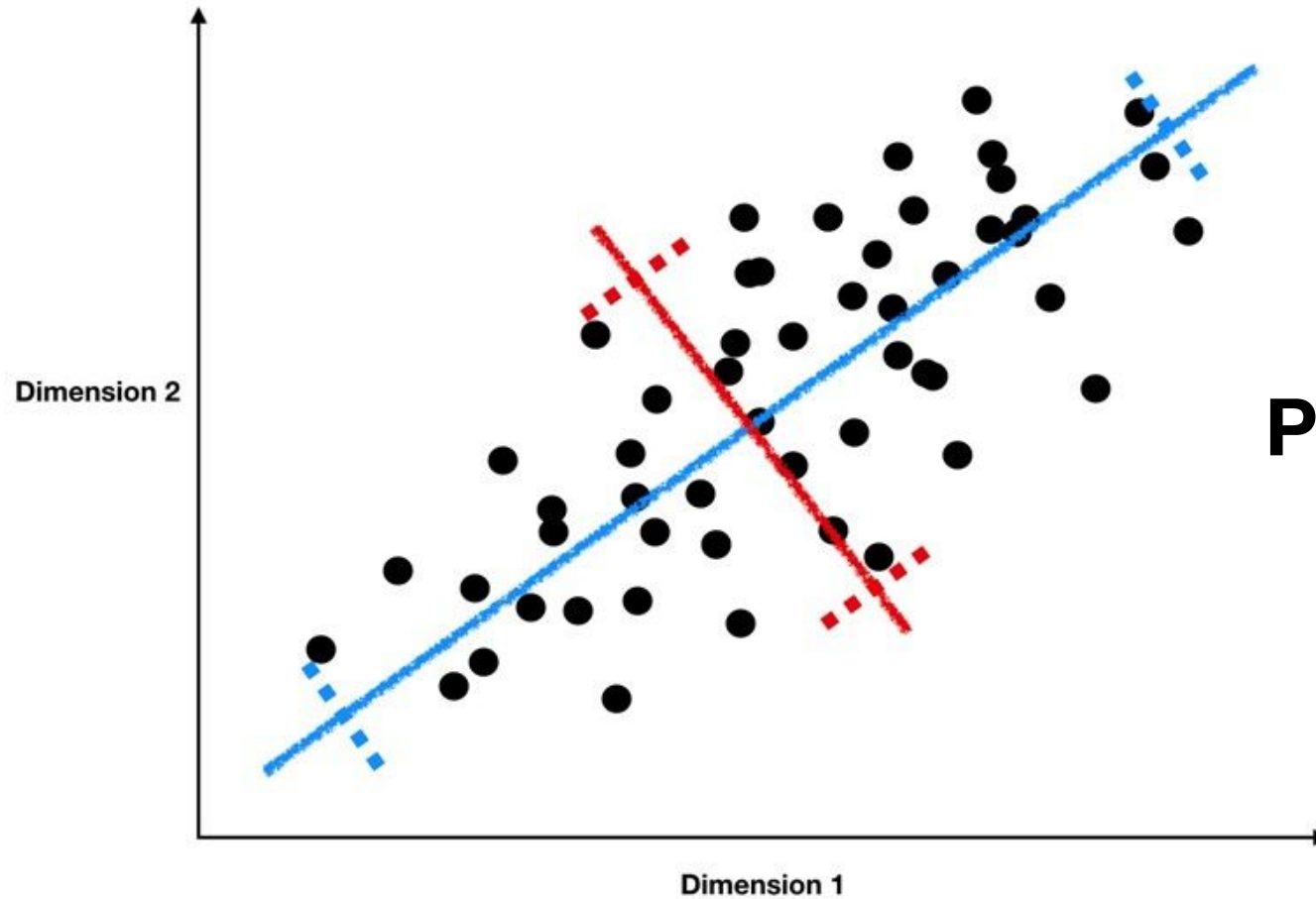


# Como funciona este algoritmo?



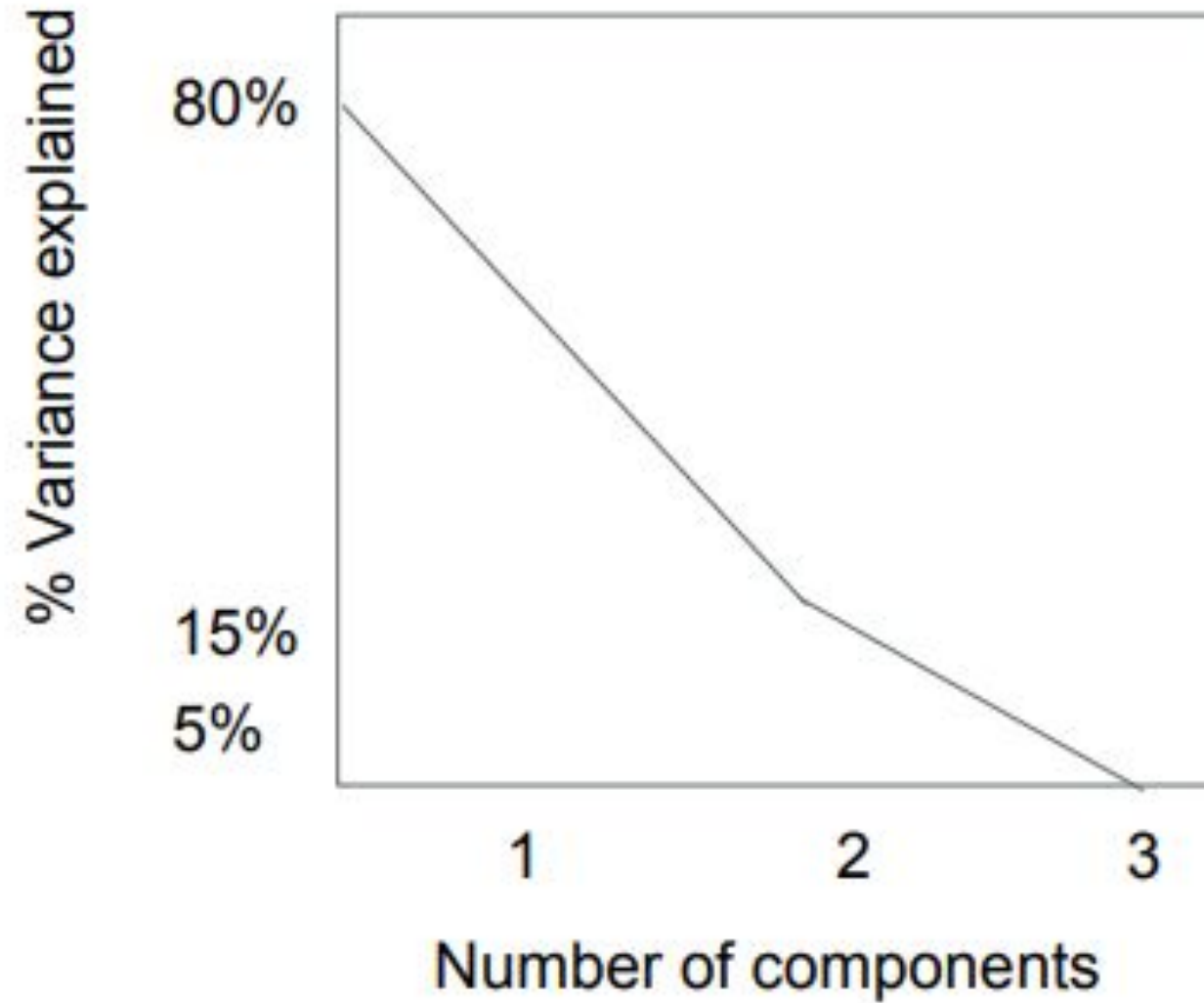
**Estandarizar  
nuestros datos**

# Buscar el trend principal



$$PC1 = c1 (XA) + c2 (XB) + \dots$$

# Método de elbow





# Como un ordenador lee las imágenes

- Por cada uno de estos  $321 \times 261$  píxeles de la imagen, cada pixel es una característica que se traduce en  $321 \times 261 = 83781$  características por sólo 4 observaciones.



- Este es un dataset de gran dimensión, pero podemos usar PCA para simplificarlo.

Imagen Original  
+83k features



Imagen recreada con 4  
componentes principales



# GRACIAS!

Eslem Alzate

[j.eslem03@gmail.com](mailto:j.eslem03@gmail.com)