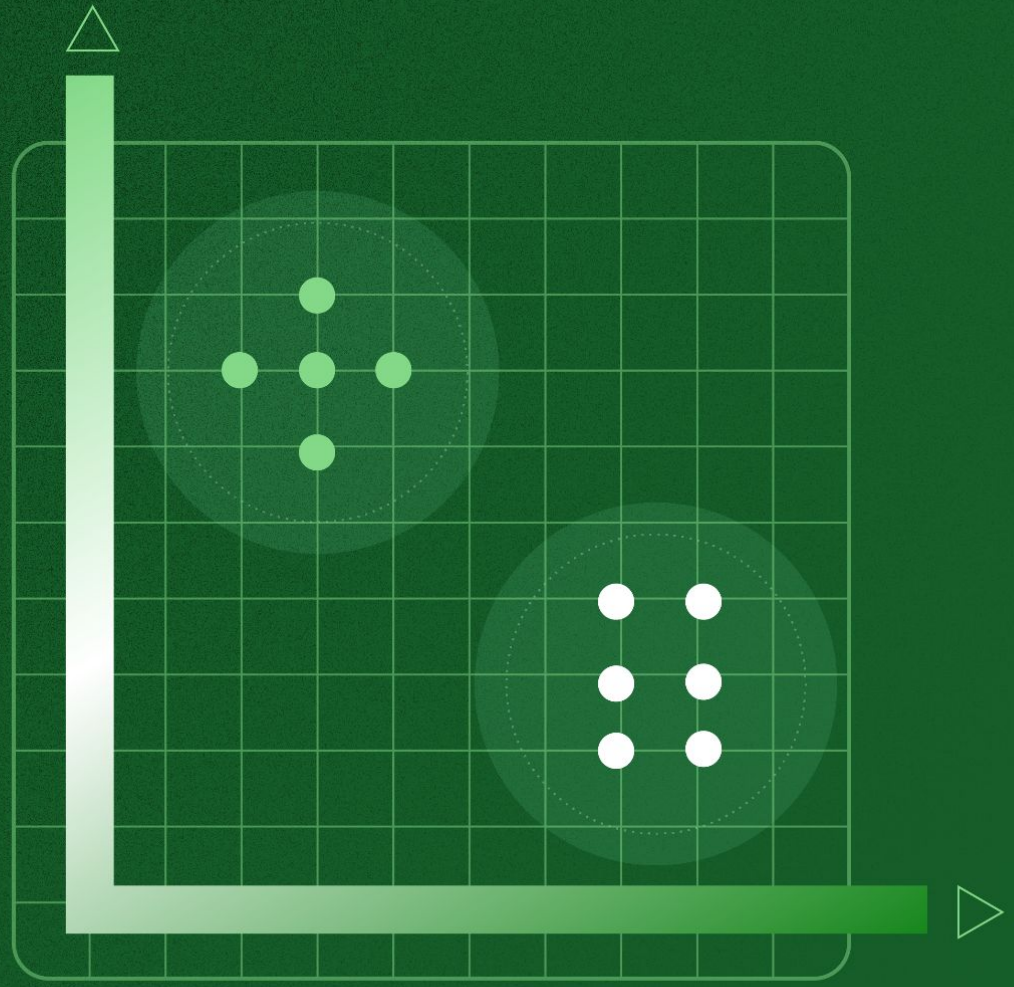
Curso de

Clustering con Python y scikit-learn

Carlos Alarcón



¿Quién es Carlos Alarcón?

🧑‍💼 Data Architect en Platzi.

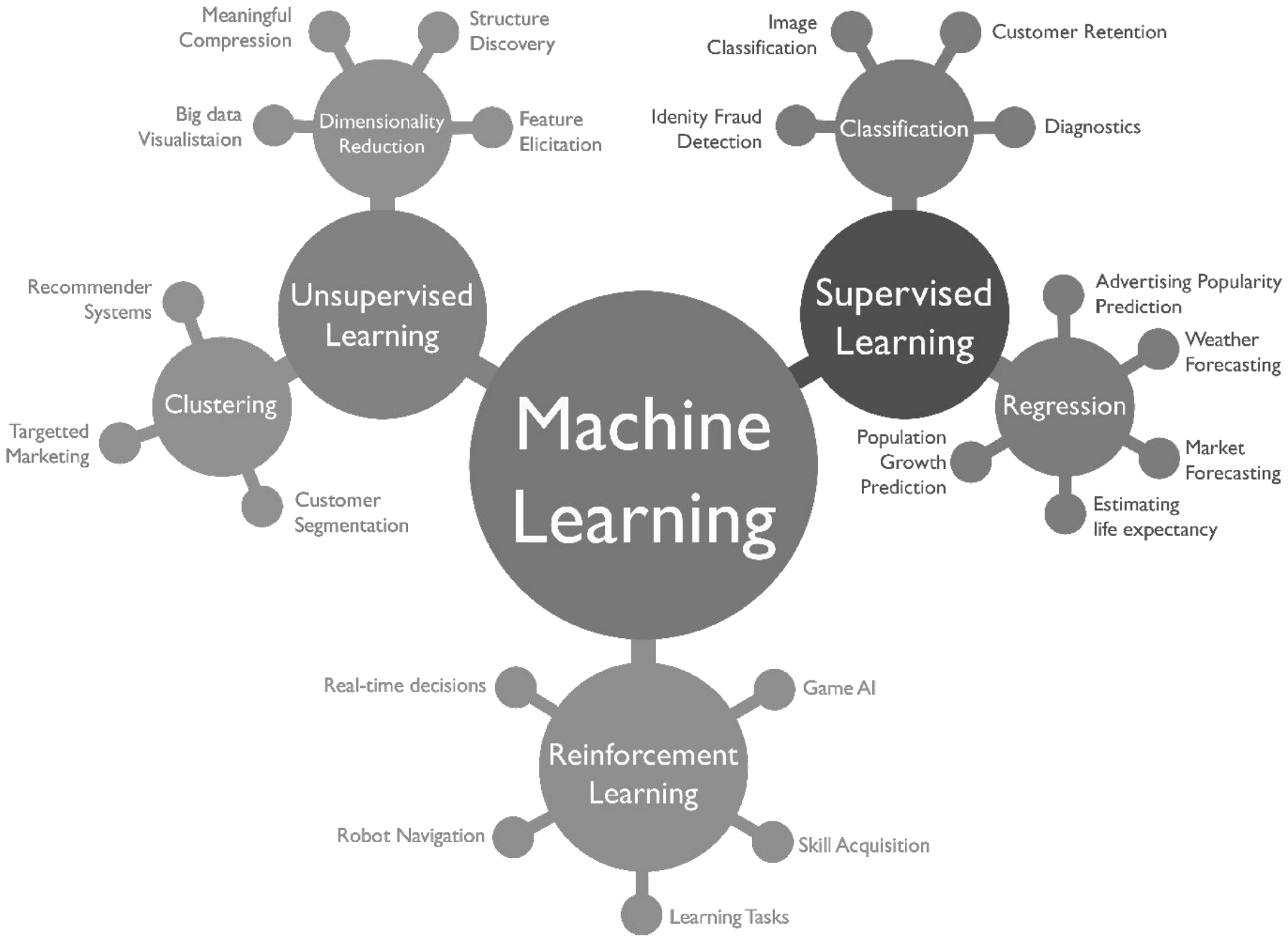
💻 Especialista en ciencia de datos, bases de datos y AI.

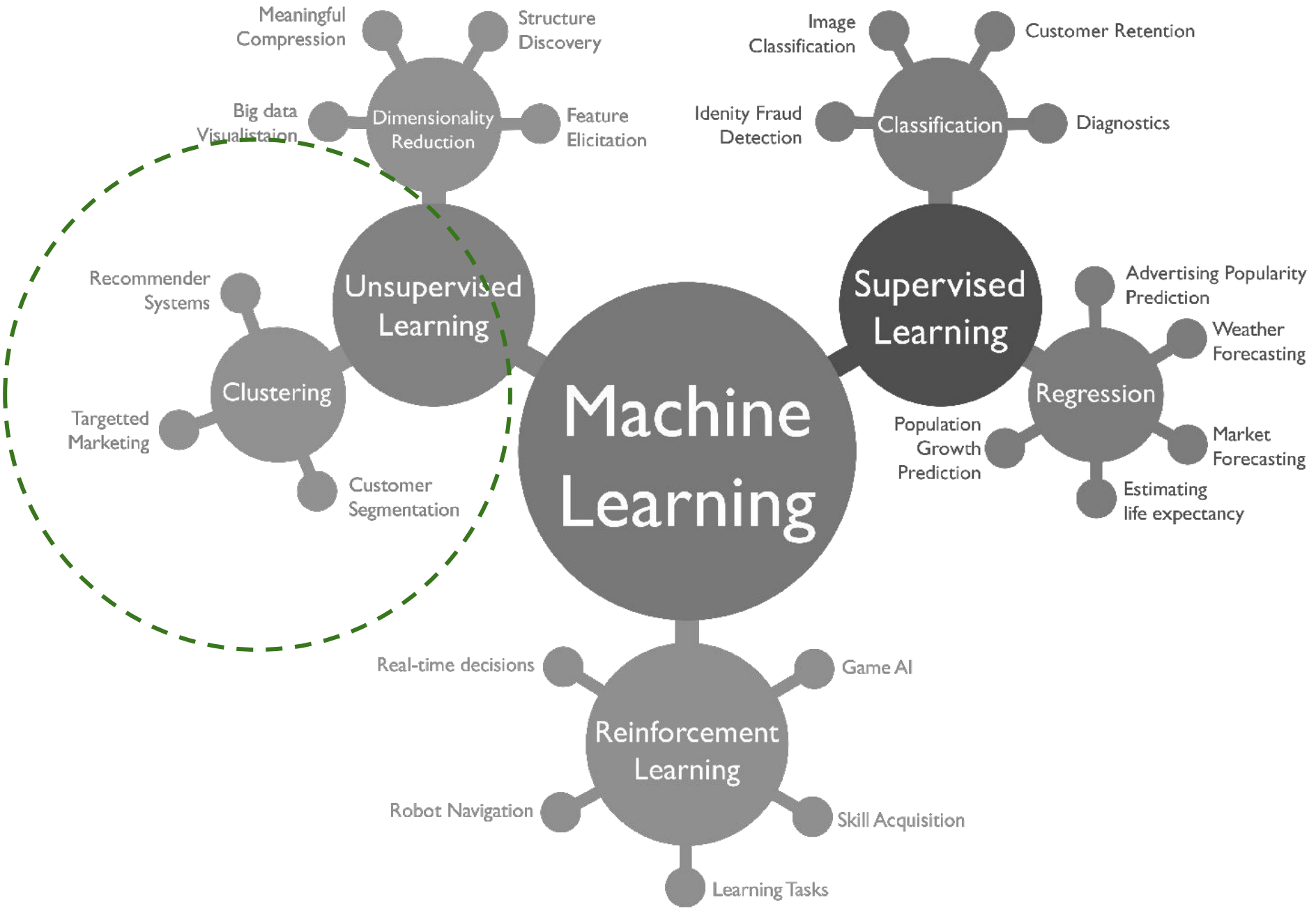
🧠 Profesor de data science y machine learning.

Requisitos previos

- Matemáticas para machine learning.
- Análisis exploratorio de datos con Python y Pandas.
- Visualización de datos con Matplotlib y Seaborn.
- Fundamentos de machine learning.

¿Qué es clustering?


Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Clustering
Classification
Regression
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Recommender Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks


Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Clustering
Classification
Regression
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Recommender Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Machine learning

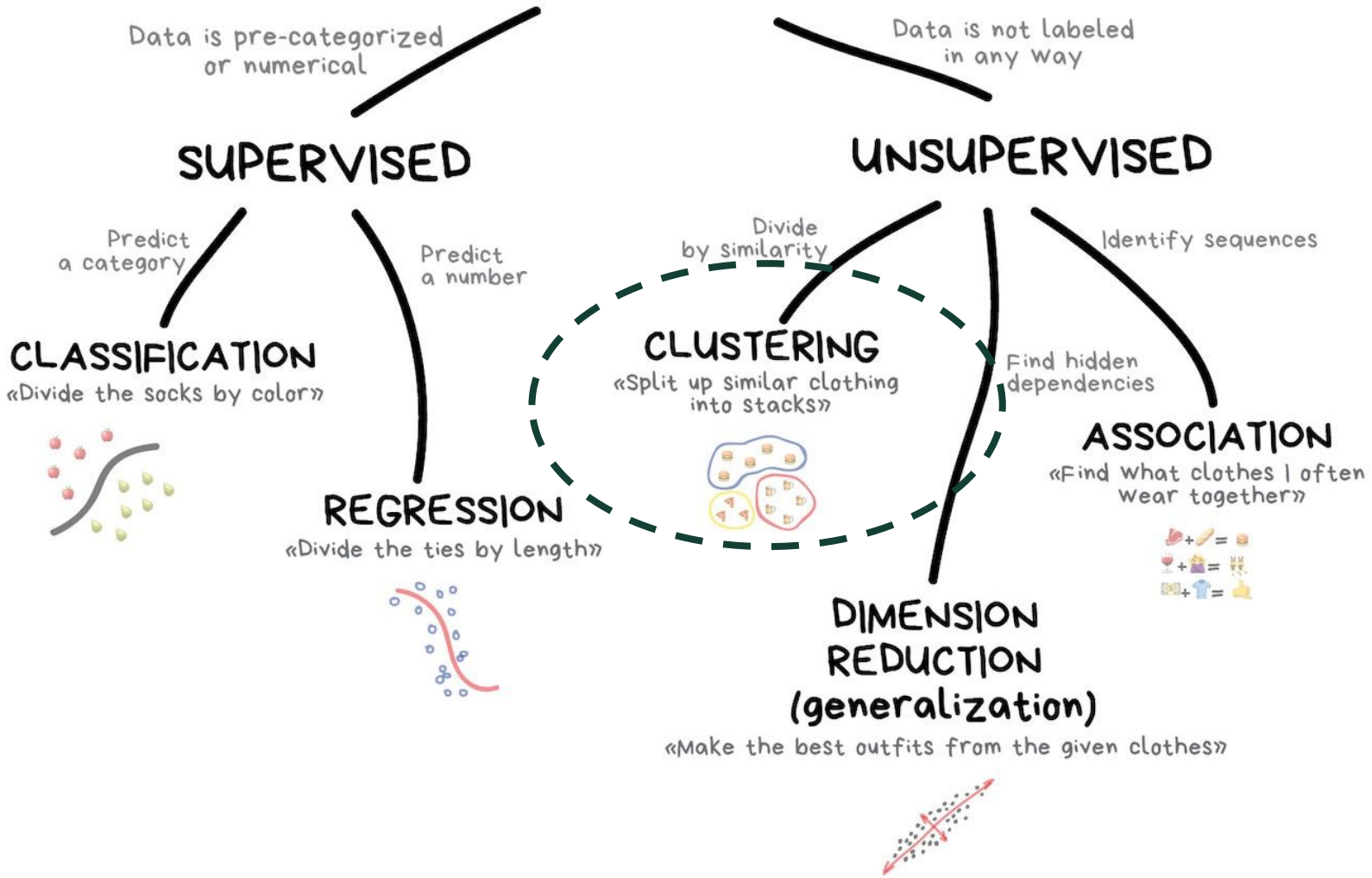
Supervised

X0	X1	X2	Y

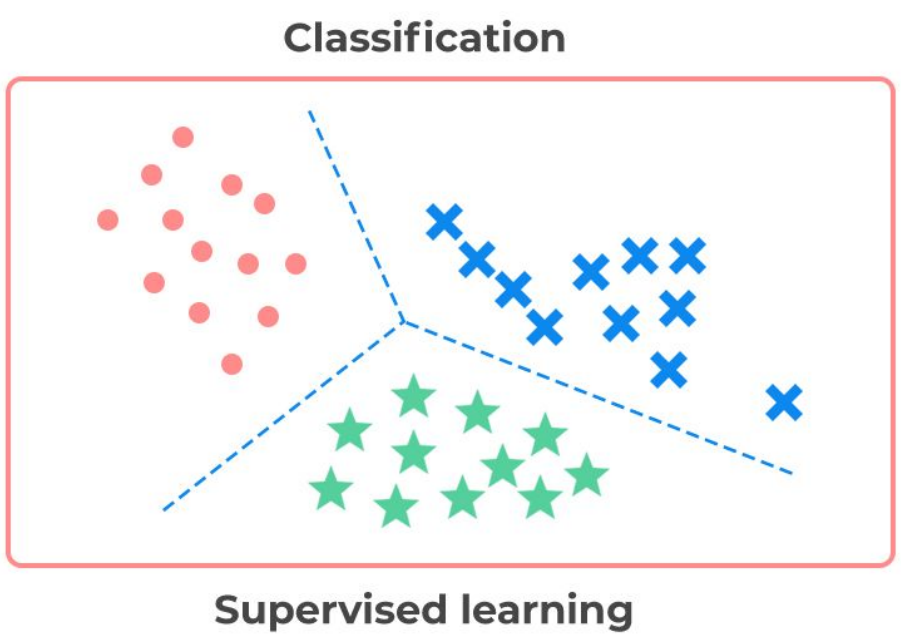
Unsupervised

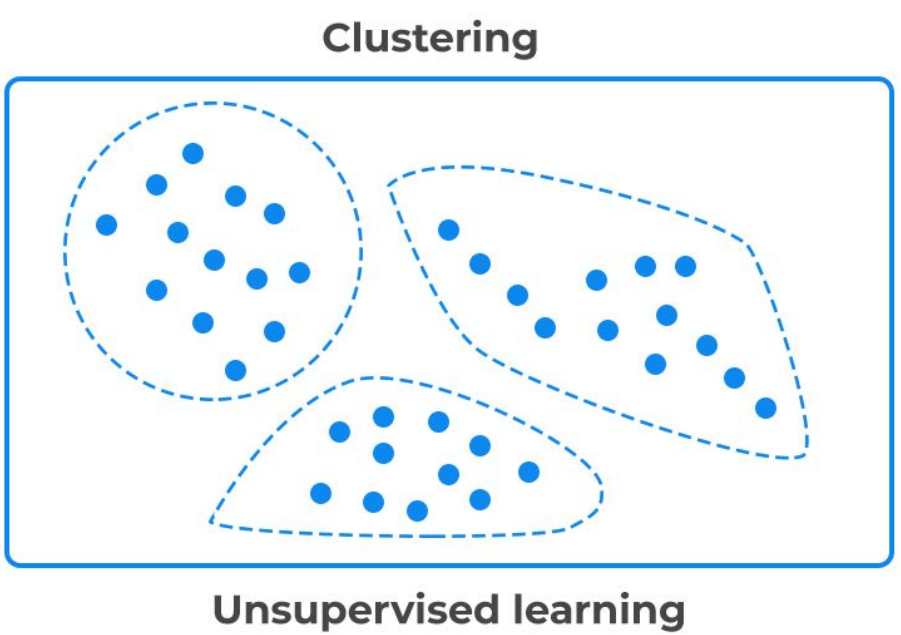
X0	X1	X2

Classical machine learning

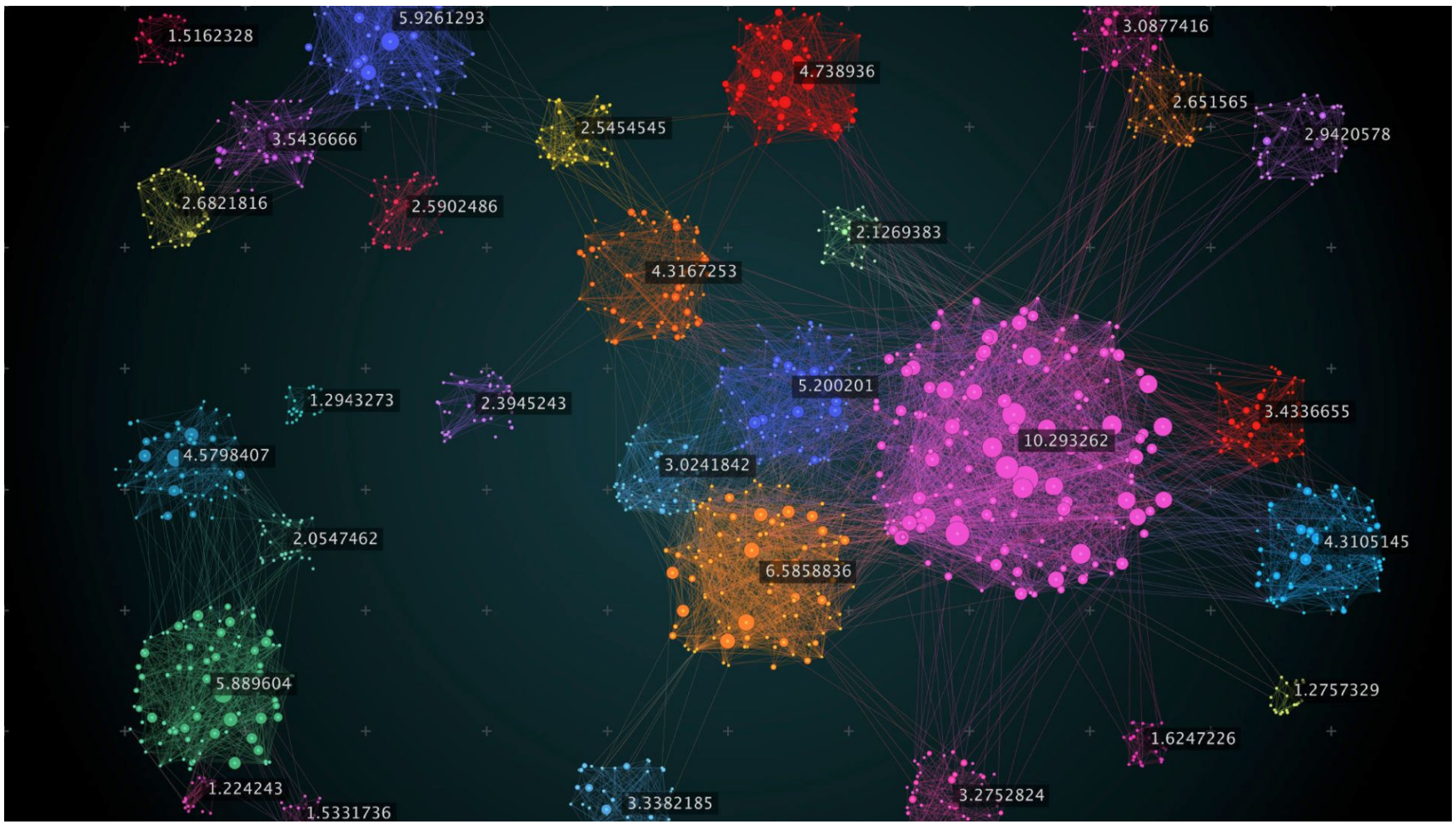


Machine learning





Clustering


1.5162328
5.9261293
4.738936
3.0877416
2.651565
2.9420578
3.5436666
2.5454545
2.6821816
2.5902486
2.1269383
4.3167253
5.200201
1.2943273
2.3945243
3.4336655
10.293262
4.5798407
3.0241842
2.0547462
4.3105145
6.5858836
5.889604
1.2757329
1.6247226
1.224243
1.5331736
3.3382185
3.2752824

Tu primer clustering con scikit-learn

¿Cuándo usar clustering?

¿Cuándo usar clustering?

- Mayor contexto.
- Detección de outliers.
- Clasificar/Agrupar (sin variable objetivo).
- Tareas manuales de crear etiquetas.

¿Qué puedo lograr?

- Clasificación de tráfico en una página.
- Segmentación de clientes por su perfil.
- Clasificación de contenido.
- Identificar comportamientos fraudulentos.
- Ciencia en los deportes.
- Muchas posibilidades.

Clasificación y clustering

Clasificación	Clustering
Con variable objetivo.	Sin variable objetivo.
Más complejo.	Menos complejo.
Data de entrenamiento / test.	Un solo set de datos.
Asignar datos a determinada clase.	Encontrar similitudes en los datos.
Se conoce la cantidad de clases.	Se desconoce la cantidad de grupos.
Dos fases (entrenamiento-predicción).	Una fase.

Algoritmos de clustering

K-means	Hierarchical clustering	DBSCAN

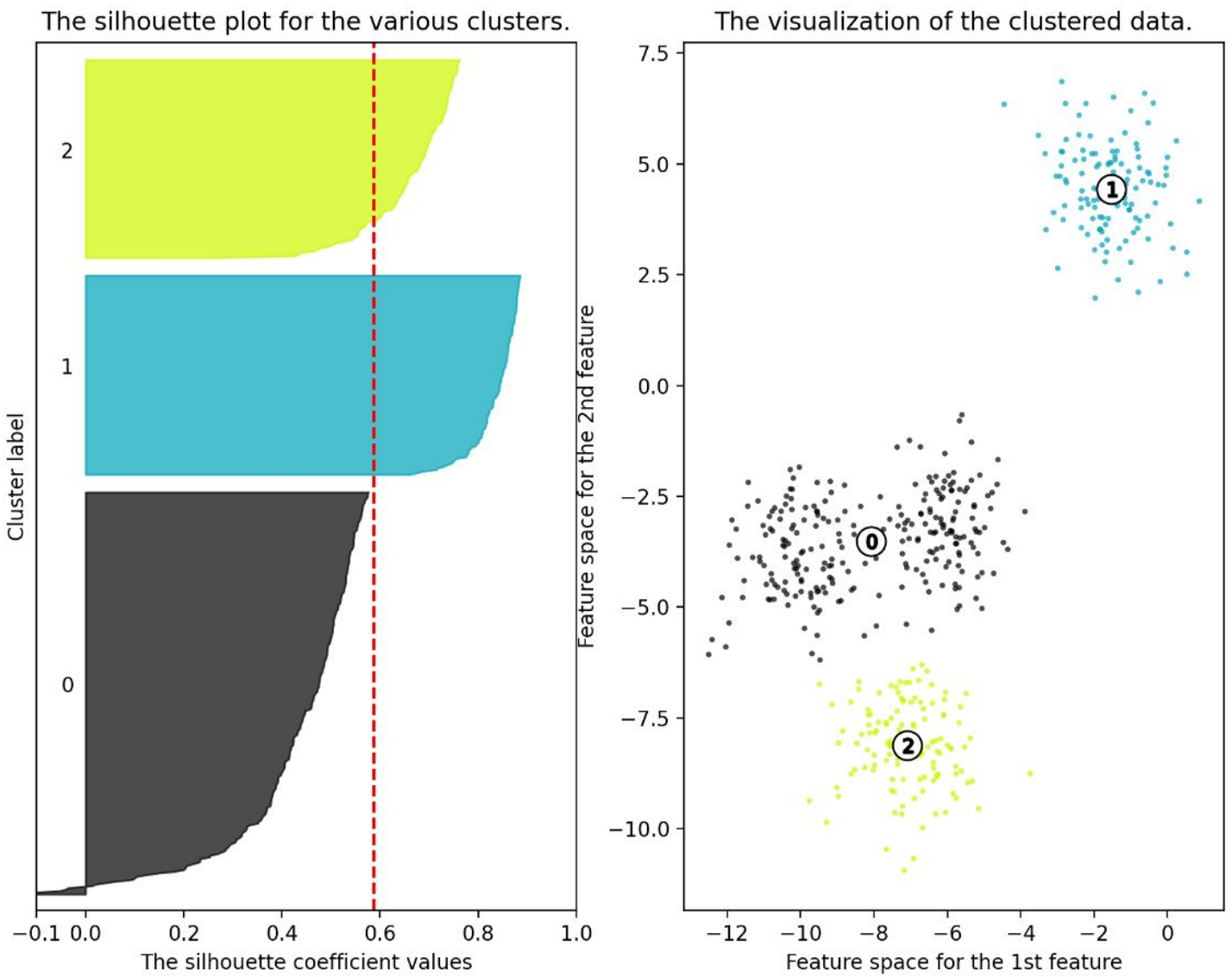
¿Cómo evaluar modelos de clustering?

Coeficiente de silueta

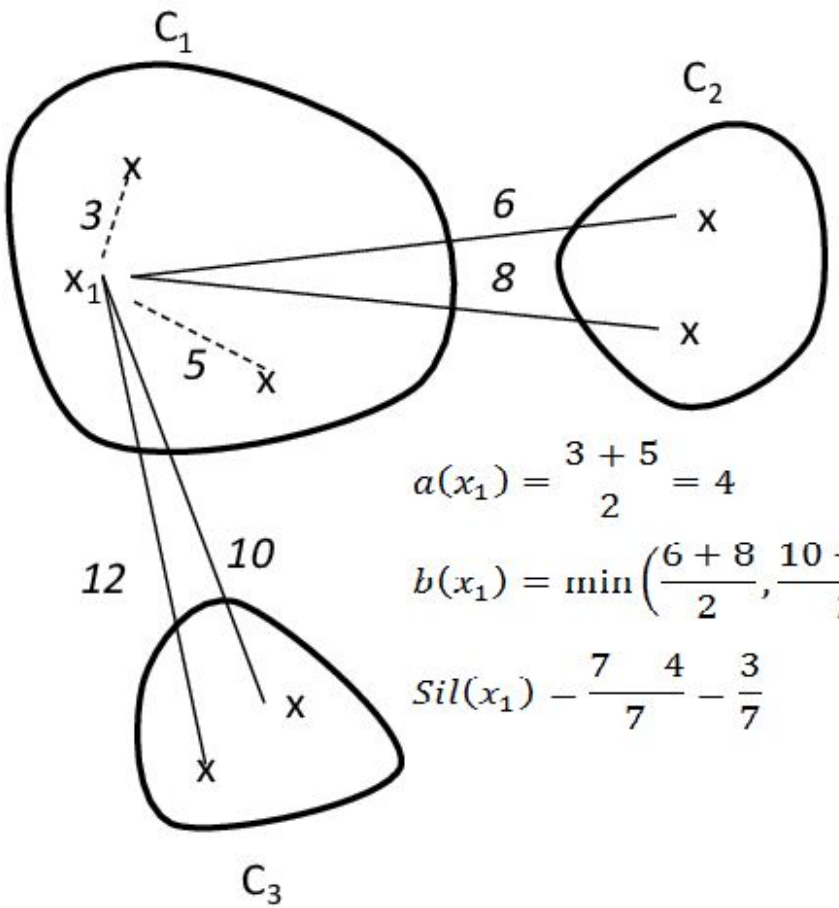
$$s(i) = b-a / max(a, b)$$

- **a** es el promedio de las disimilitudes (o distancias) de la observación **i** con las demás observaciones del cluster al que pertenece **i**.
- **b** es la distancia mínima a otro cluster que no es el mismo en el que está la observación **i**.

Ese cluster es la segunda mejor opción para **i** y se lo denomina vecindad de **i**.


The silhouette plot for the various clusters.
Cluster label
2
1
0
−0.1 0.0 0.2 0.4 0.6 0.8 1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12 −10 −8 −6 −4 −2 0
Feature space for the 1st feature
1
0
2

Coeficiente de silueta



$$a(x_1) = \frac{3+5}{2} = 4$$

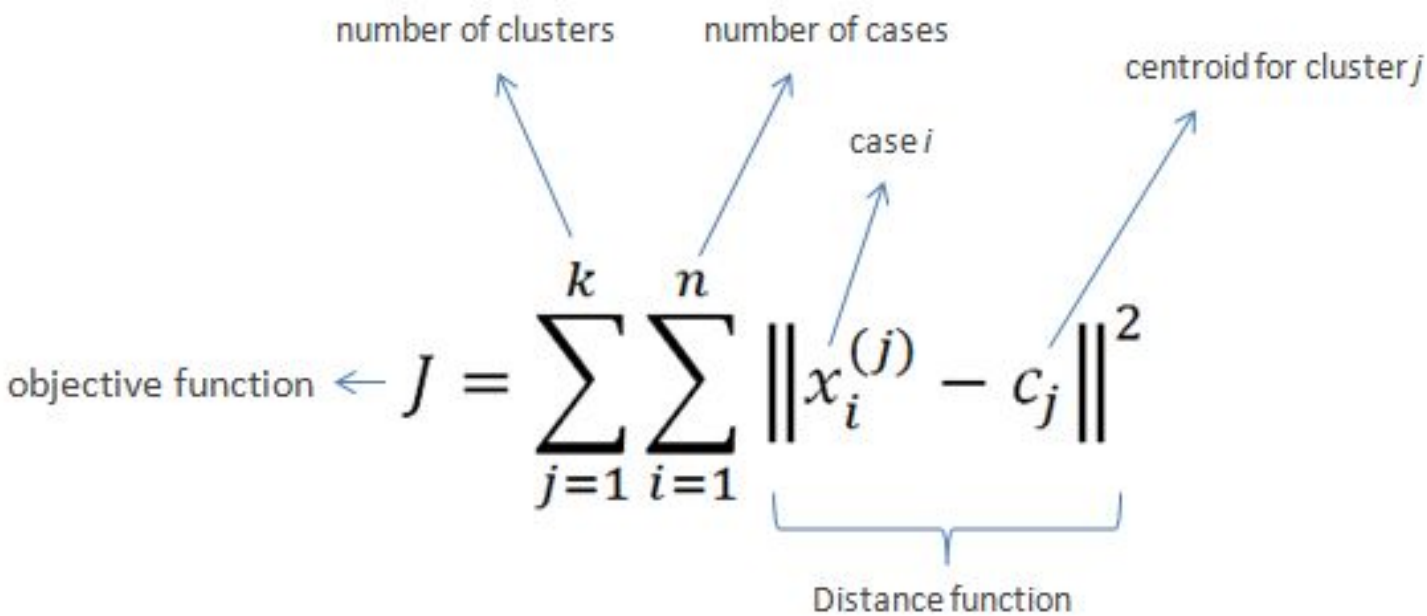
$$b(x_1) = \min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$$

$$Sil(x_1) - \frac{7}{7} \frac{4}{} - \frac{3}{7}$$

K-means

K-means

1. Indicar cantidad de clusters.
2. Ubicar centroides aleatoriamente.
3. Cada punto se asigna al centroide más cercano.
4. Recalcular centroides con el promedio.
5. Repetir paso 3 y 4 hasta que no se muevan los centroides.

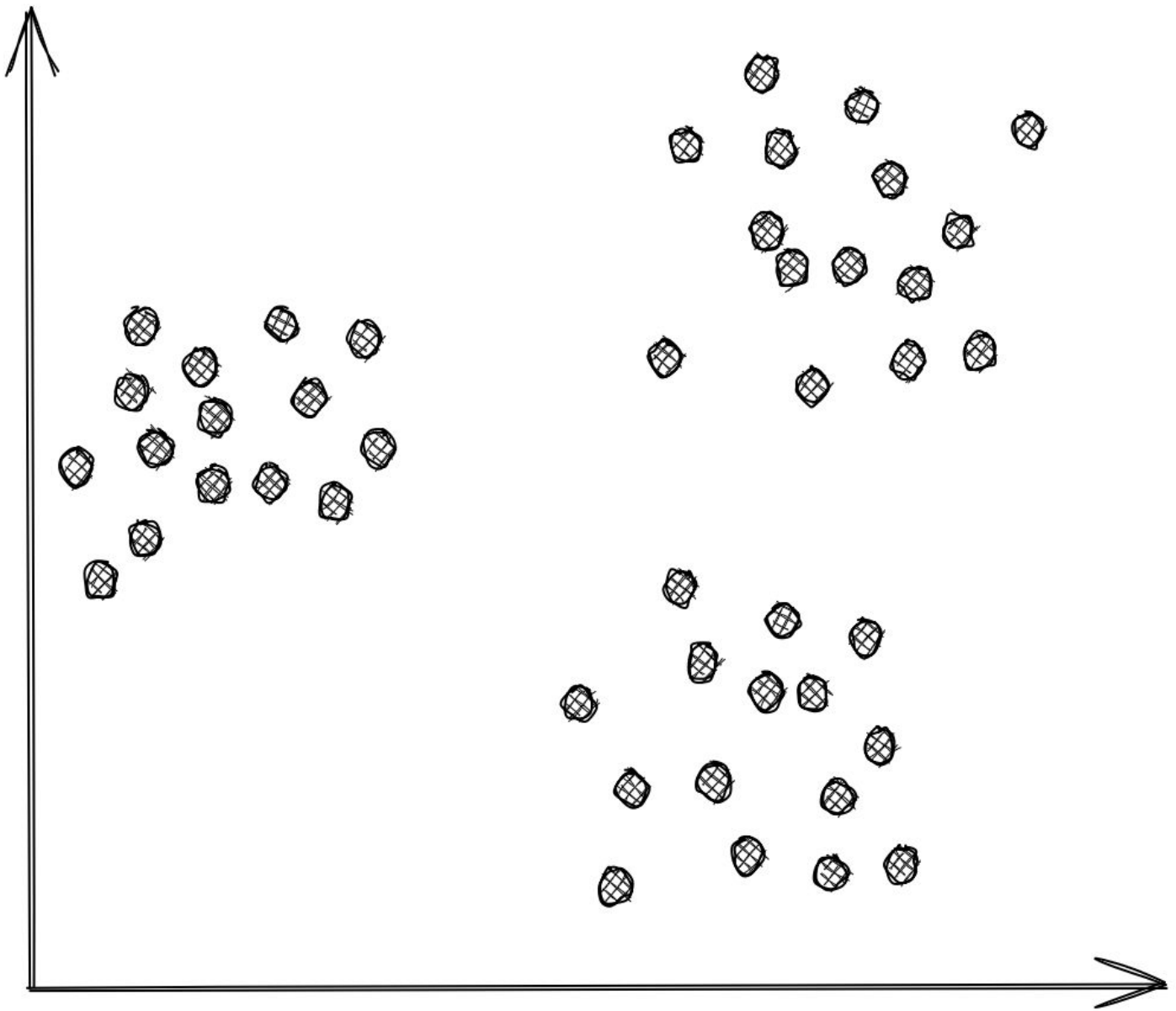

number of clusters
number of cases
centroid for cluster j
case i
objective function
$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$
Distance function

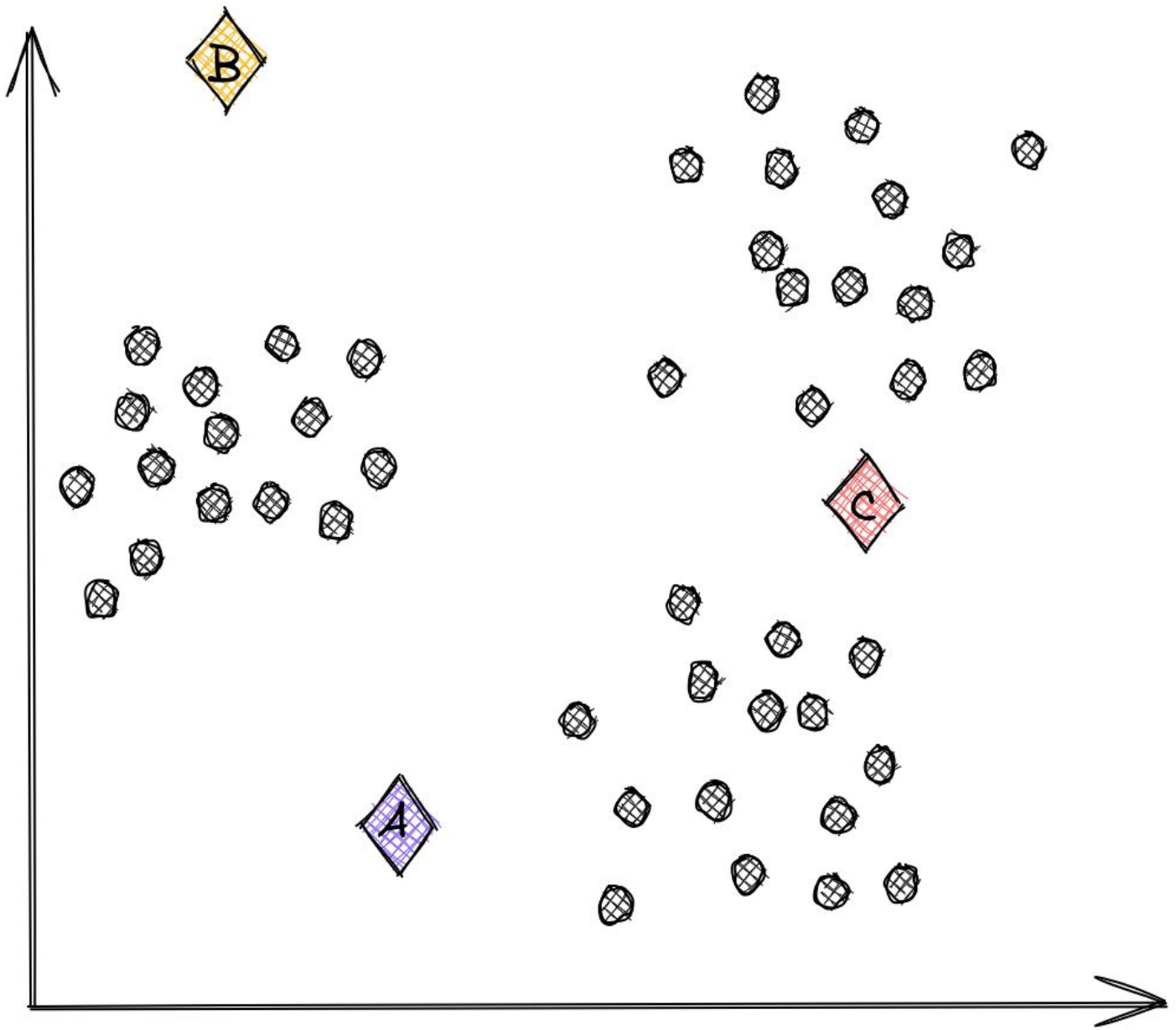
K-means

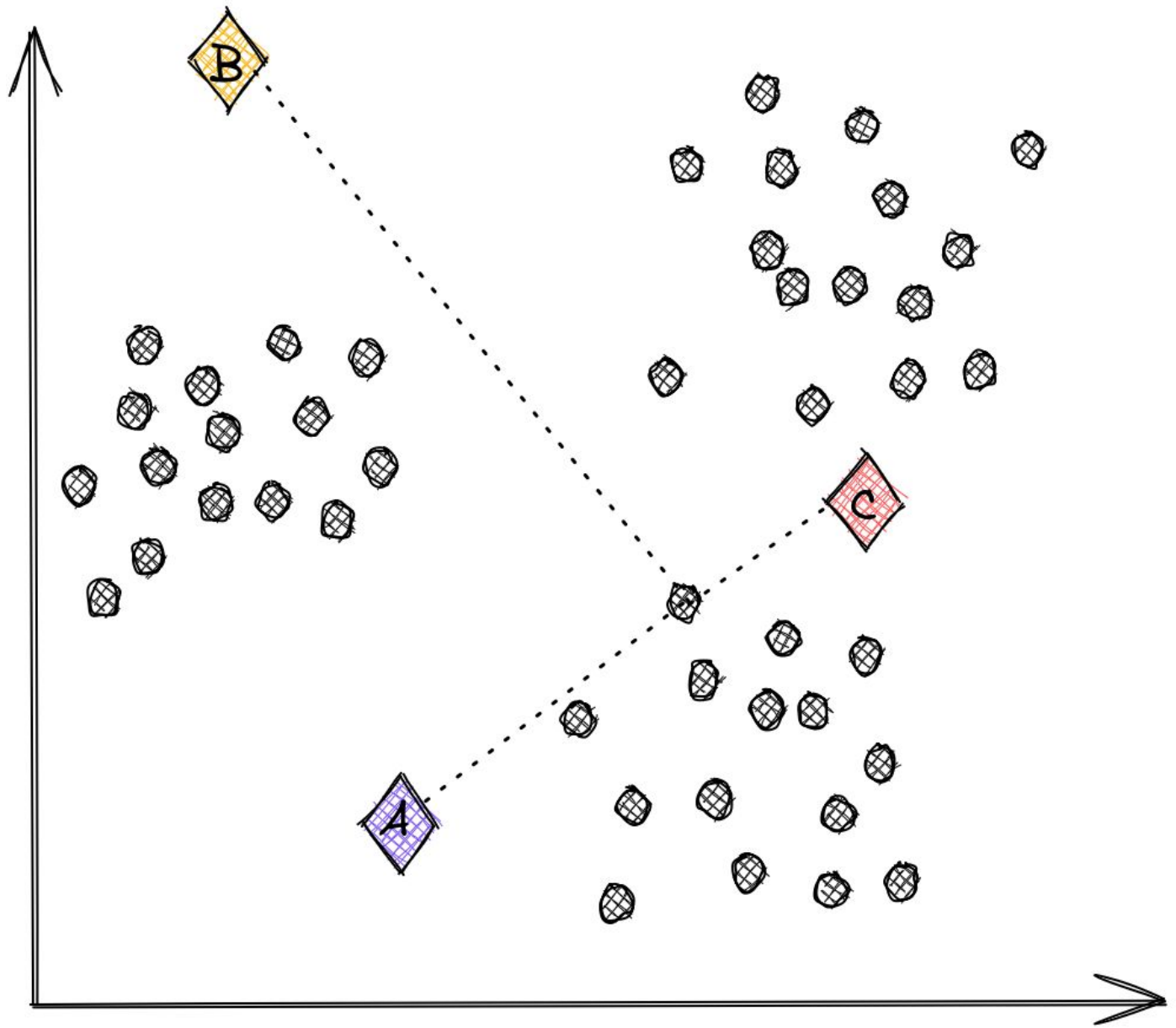
x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	16	22	1	7	1	
15	16	22	1	7	1	**15.33**
16	16	22	0	6	1	
19	16	22	9	3	2	
19	16	22	9	3	2	
20	16	22	16	2	2	
20	16	22	16	2	2	
21	16	22	25	1	2	
22	16	22	36	0	2	
28	16	22	12	6	2	
35	16	22	19	13	2	
40	16	22	24	18	2	**36.25**
41	16	22	25	19	2	
42	16	22	26	20	2	
43	16	22	27	21	2	
44	16	22	28	22	2	
60	16	22	44	38	2	
61	16	22	45	39	2	
65	16	22	49	43	2	

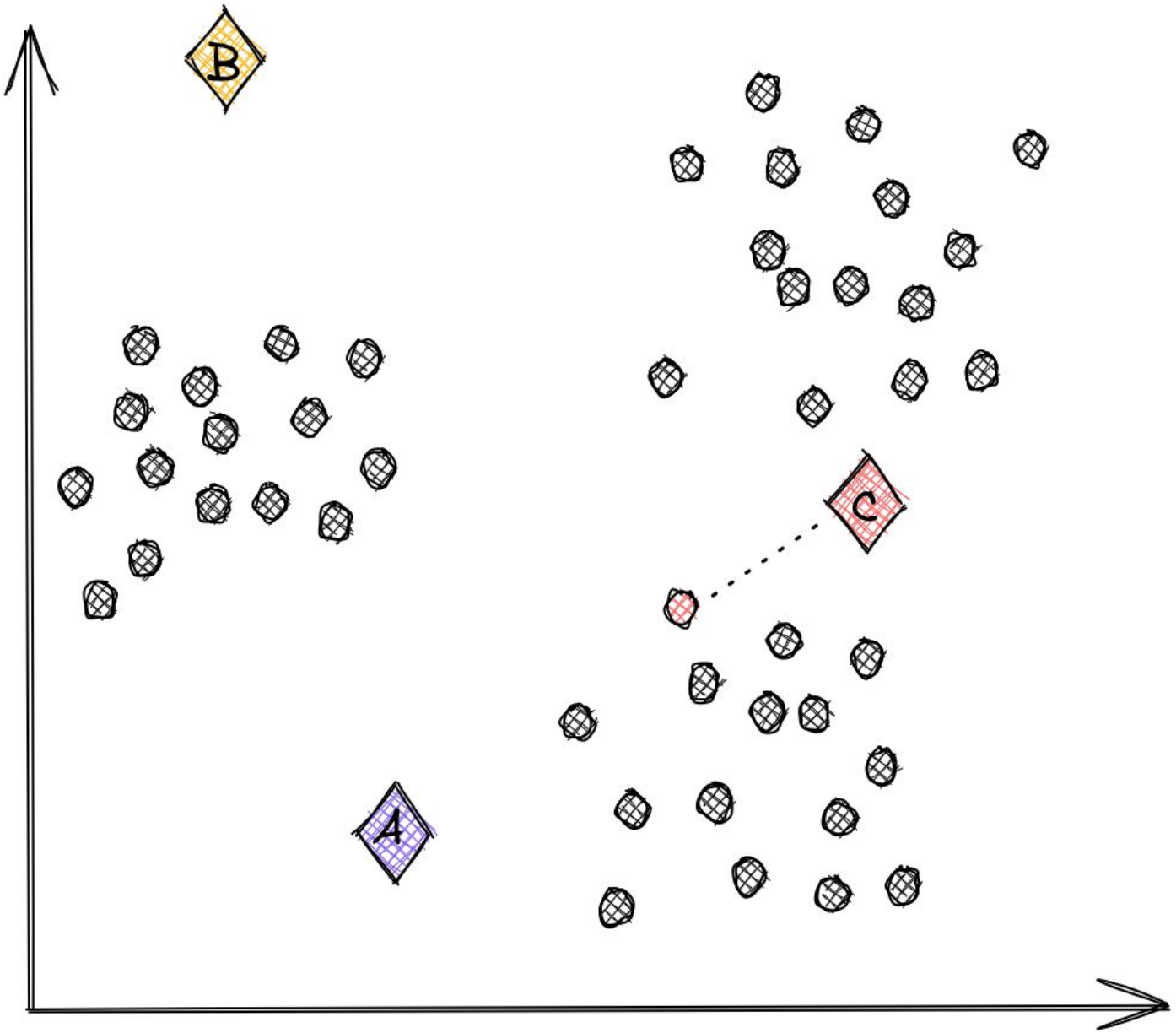
K-means

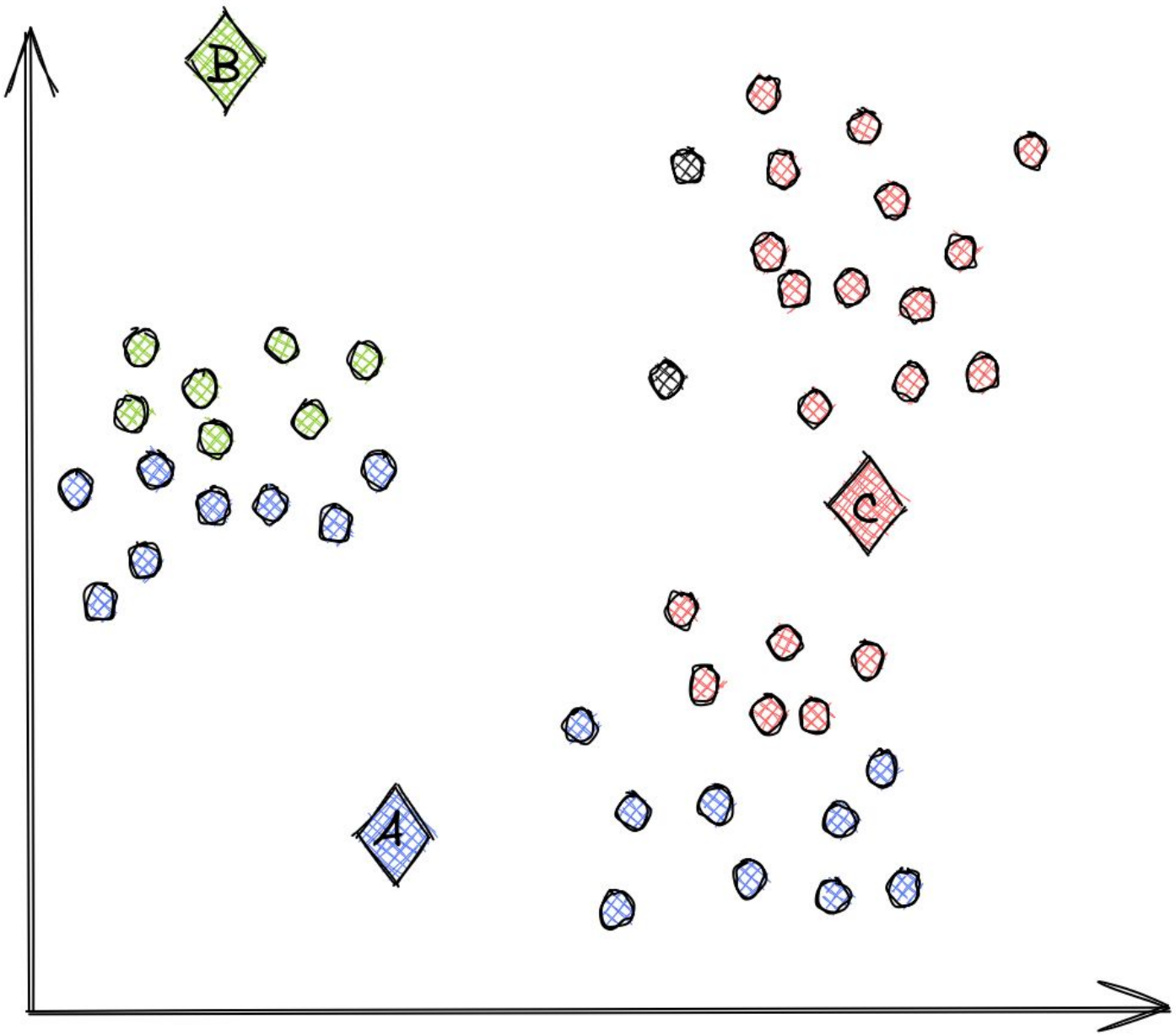
x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	15.33	36.25	0.33	21.25	1	
15	15.33	36.25	0.33	21.25	1	
16	15.33	36.25	0.67	20.25	1	
19	15.33	36.25	3.67	17.25	1	
19	15.33	36.25	3.67	17.25	1	**18.56**
20	15.33	36.25	4.67	16.25	1	
20	15.33	36.25	4.67	16.25	1	
21	15.33	36.25	5.67	15.25	1	
22	15.33	36.25	6.67	14.25	1	
28	15.33	36.25	12.67	8.25	2	
35	15.33	36.25	19.67	1.25	2	
40	15.33	36.25	24.67	3.75	2	
41	15.33	36.25	25.67	4.75	2	
42	15.33	36.25	26.67	5.75	2	
43	15.33	36.25	27.67	6.75	2	**45.9**
44	15.33	36.25	28.67	7.75	2	
60	15.33	36.25	44.67	23.75	2	
61	15.33	36.25	45.67	24.75	2	
65	15.33	36.25	49.67	28.75	2	

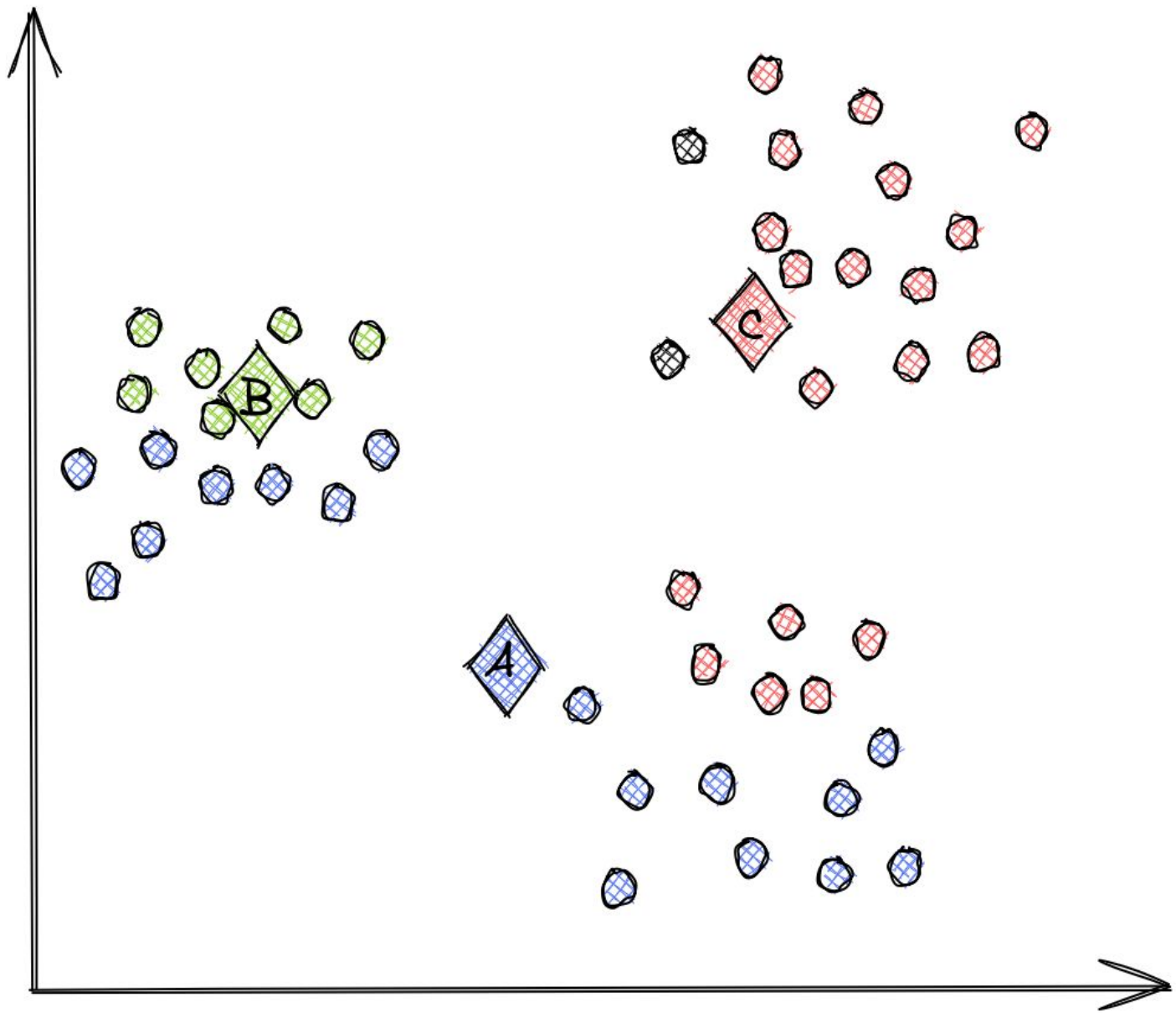


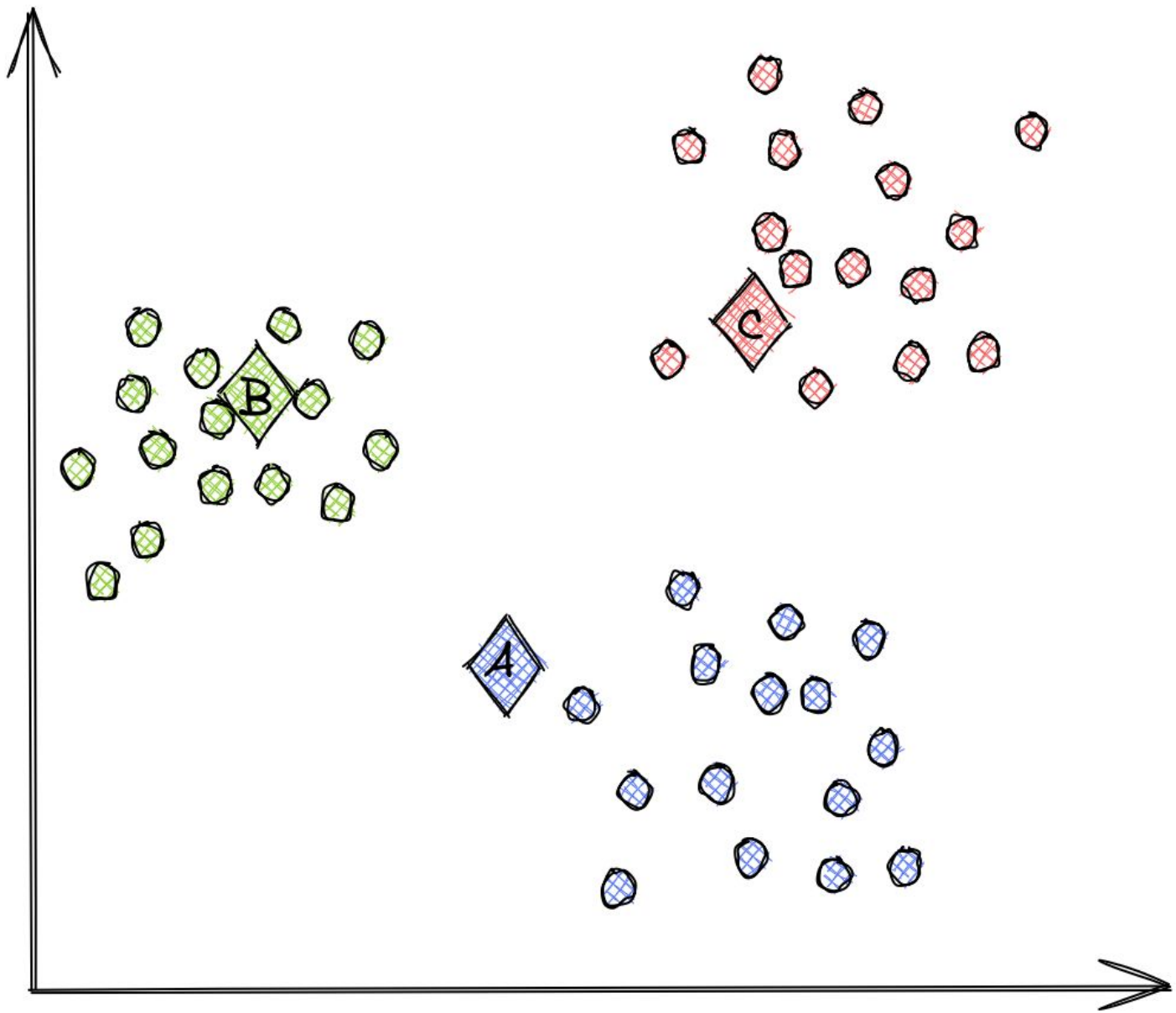

B
C
A


B
C
A


B
C
A


B
C
A


B
C
A


B
C
A

¿Cuándo usar K-means?

Ventajas de K-means

- Alto performance.
- Simple.
- Resultados interpretables.
- Garantiza convergencia.
- Se adapta a nuevos datos.

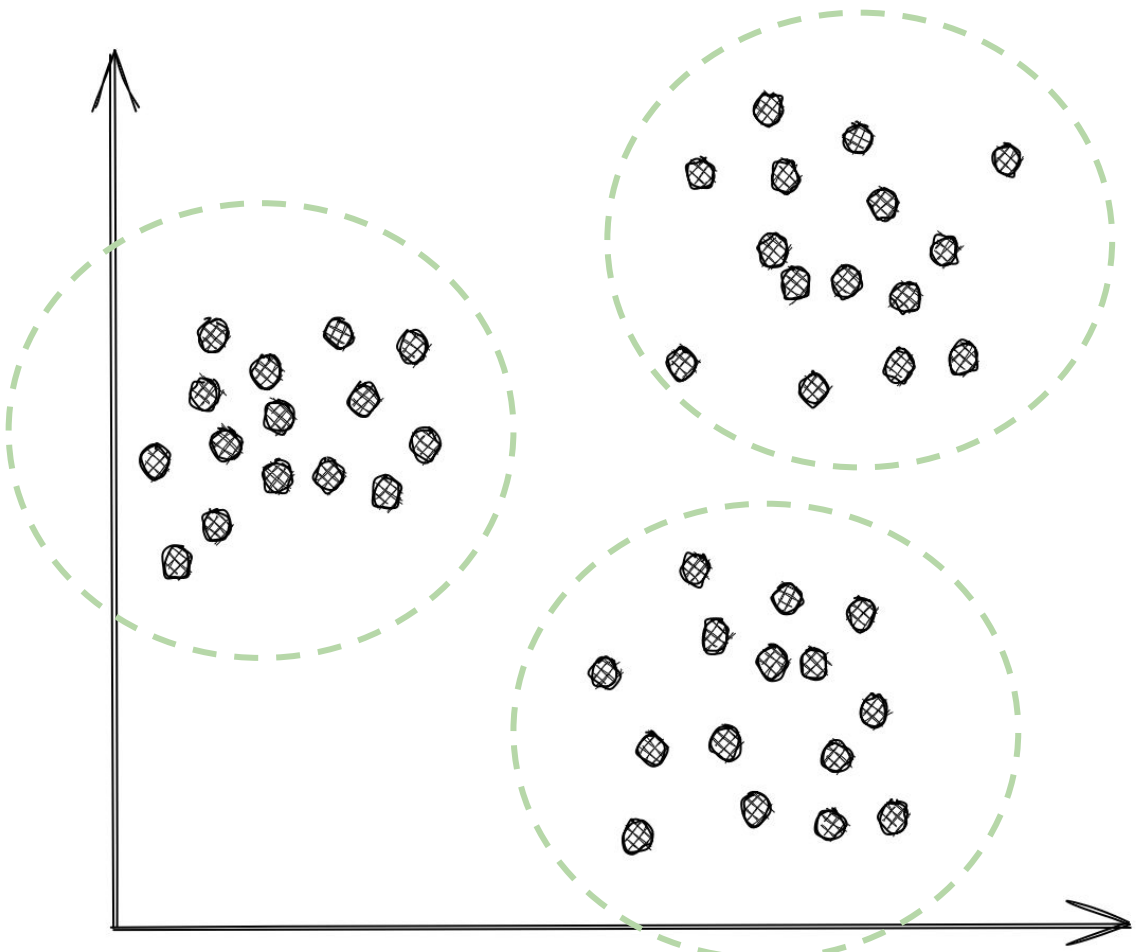
Desventajas de K-means

- Replicabilidad.
- Le afectan outliers.
- Mejor performance en datos esféricos.
- Debo elegir K manualmente.
- Afectado por alta dimensionalidad.

¿Cuándo implementar K-means?

- Conozco la cantidad de clusters que espero.
- Resultados rápidos.
- Resultados interpretables.
- Forma de datos esférica.
- Resultados escalables.

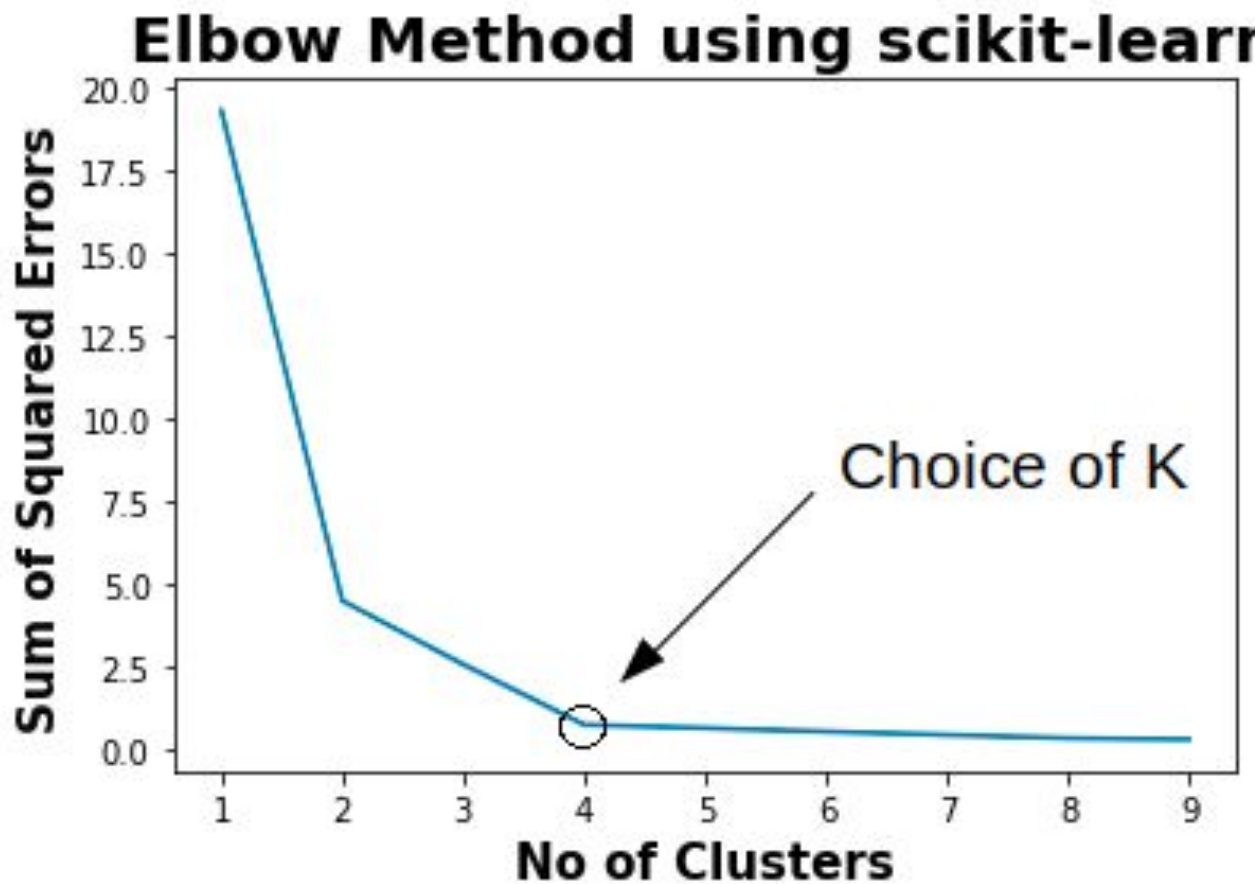
Formas esféricas



Implementando K-means

Encontrando K

Método del codo (Elbow)

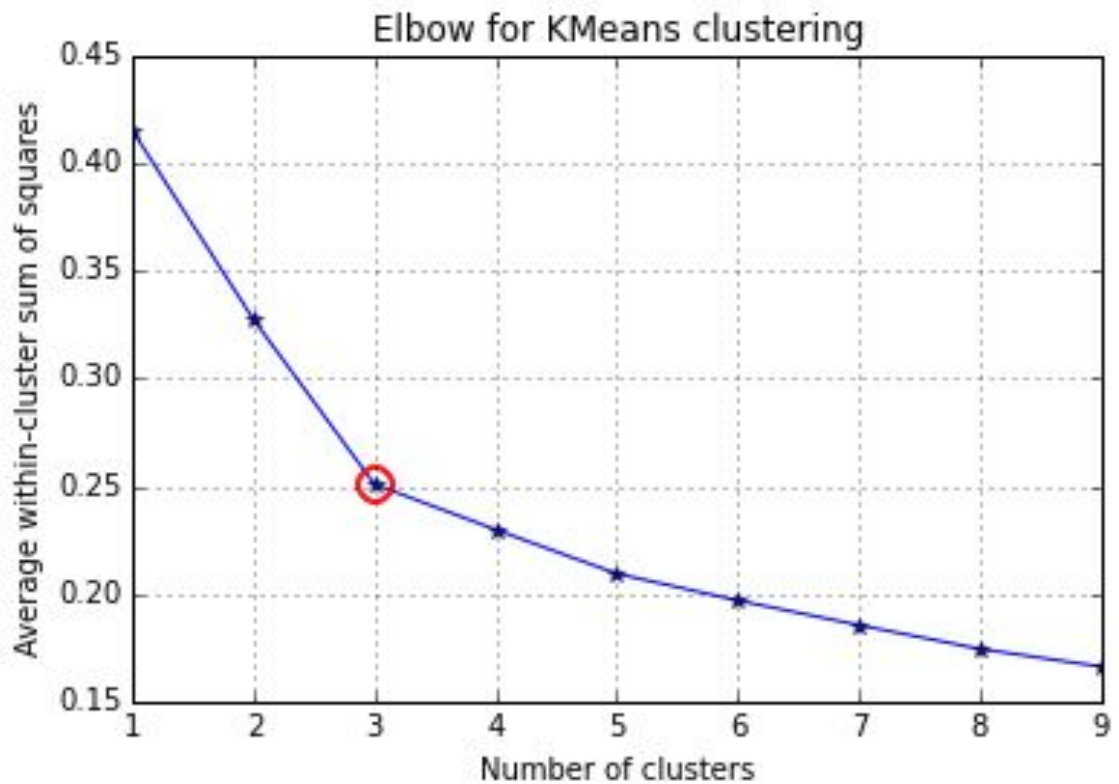


WCSS (Within-Cluster Sum of Square)

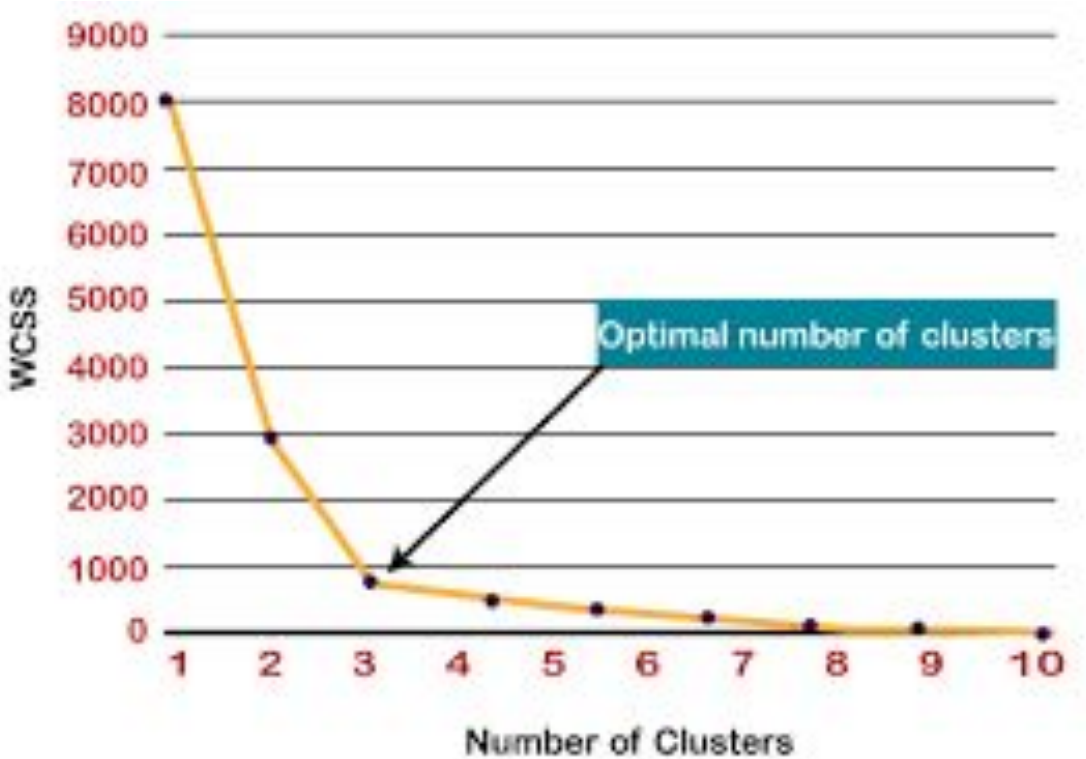
$$WCSS(C_j) = \sum_{p_i=1 \in C_j}^{p_m} distance(C_j, p_i)^2$$

La suma de la diferencia elevada al cuadrado de cada punto contra su centroide en el cluster.

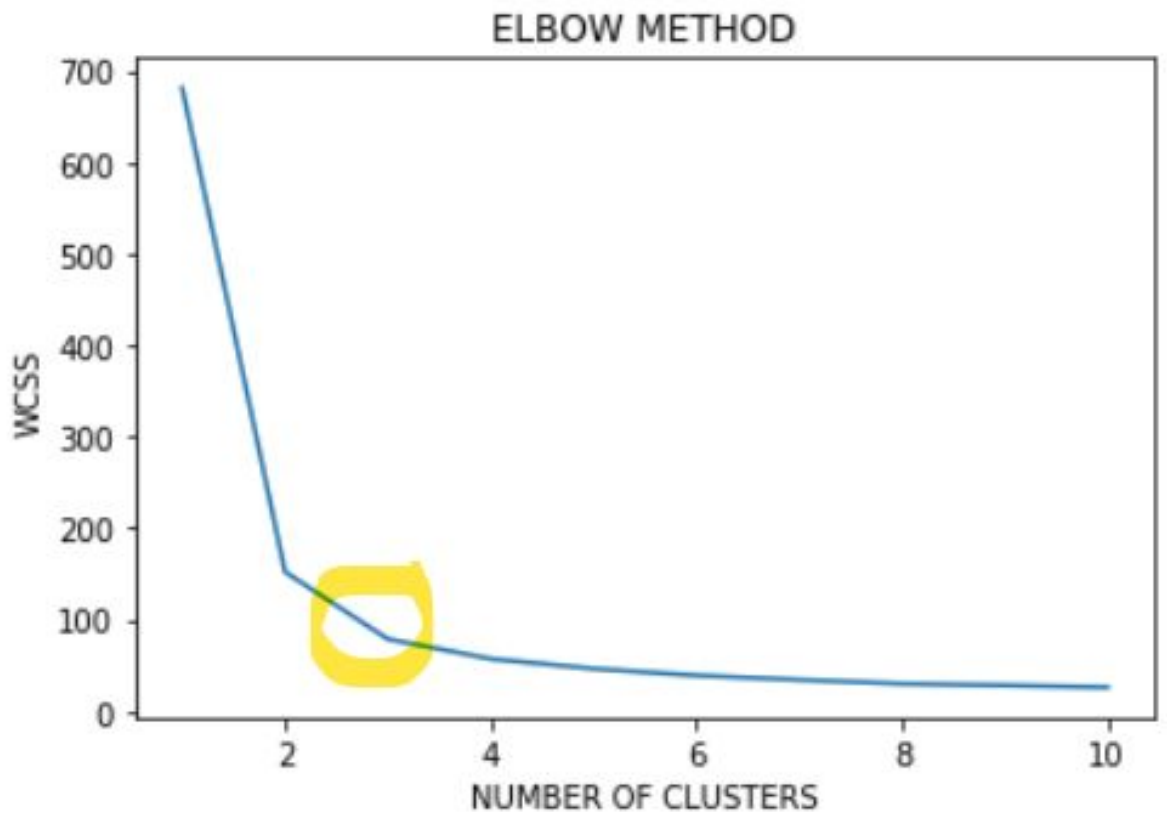
Método del codo (Elbow)



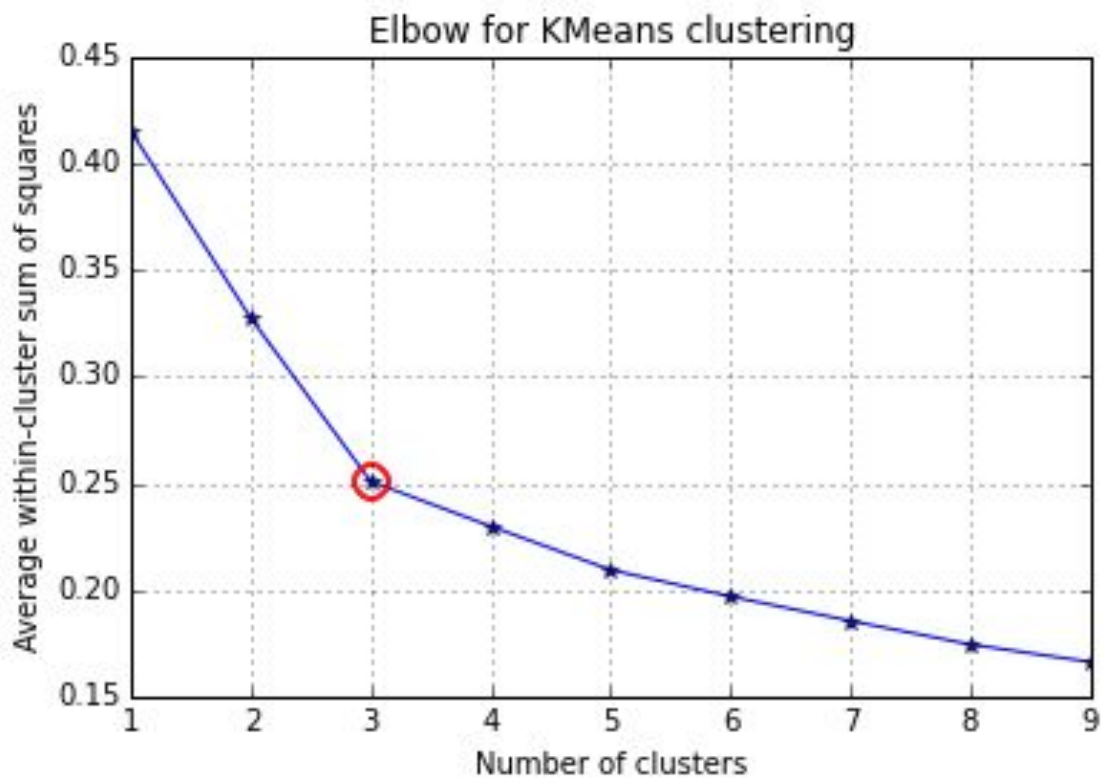
Método del codo (Elbow)



Método del codo (Elbow)



Método del codo (Elbow)



Coeficiente de silueta

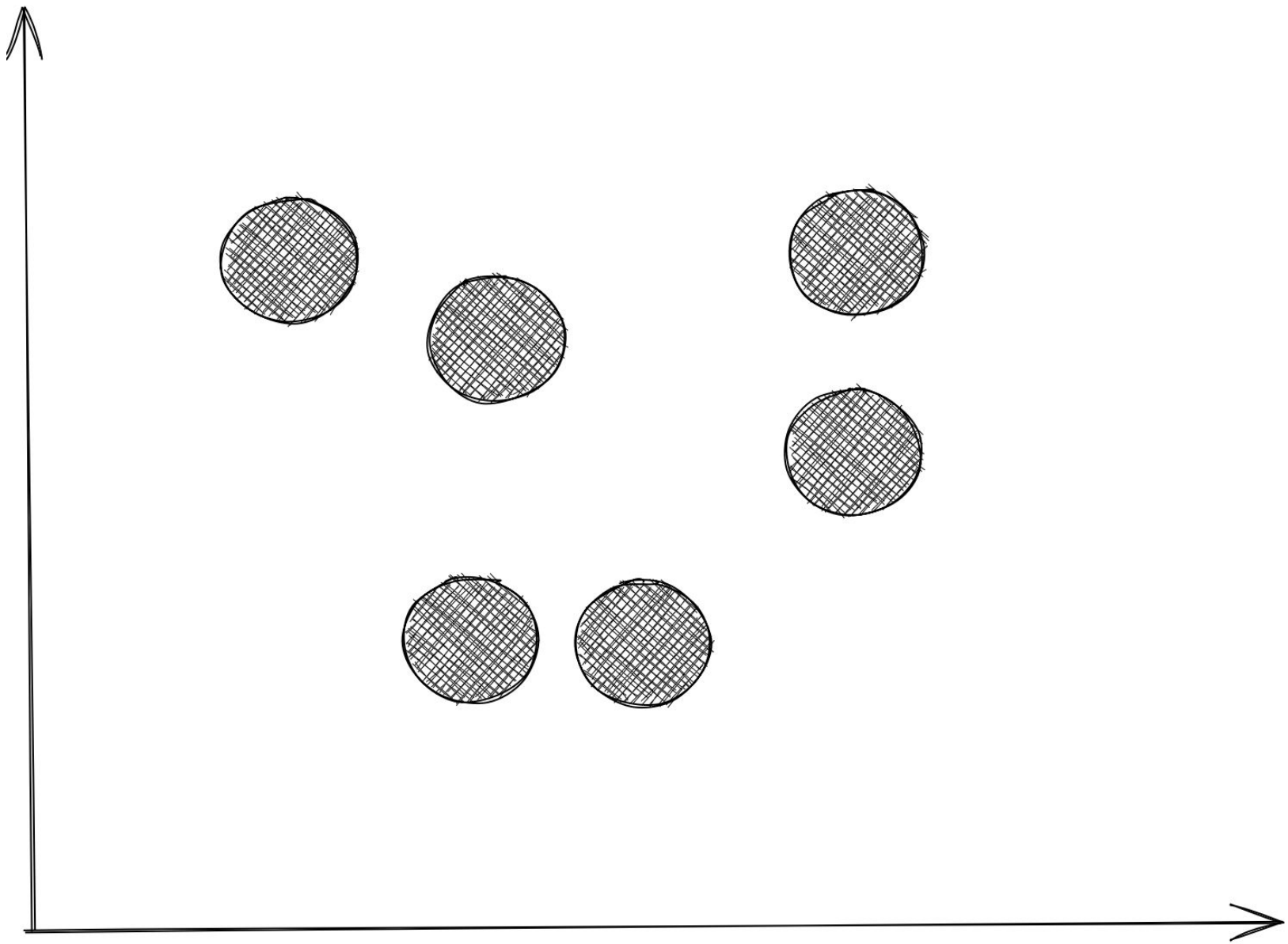
$$s(i) = b-a / max (a ,b)$$

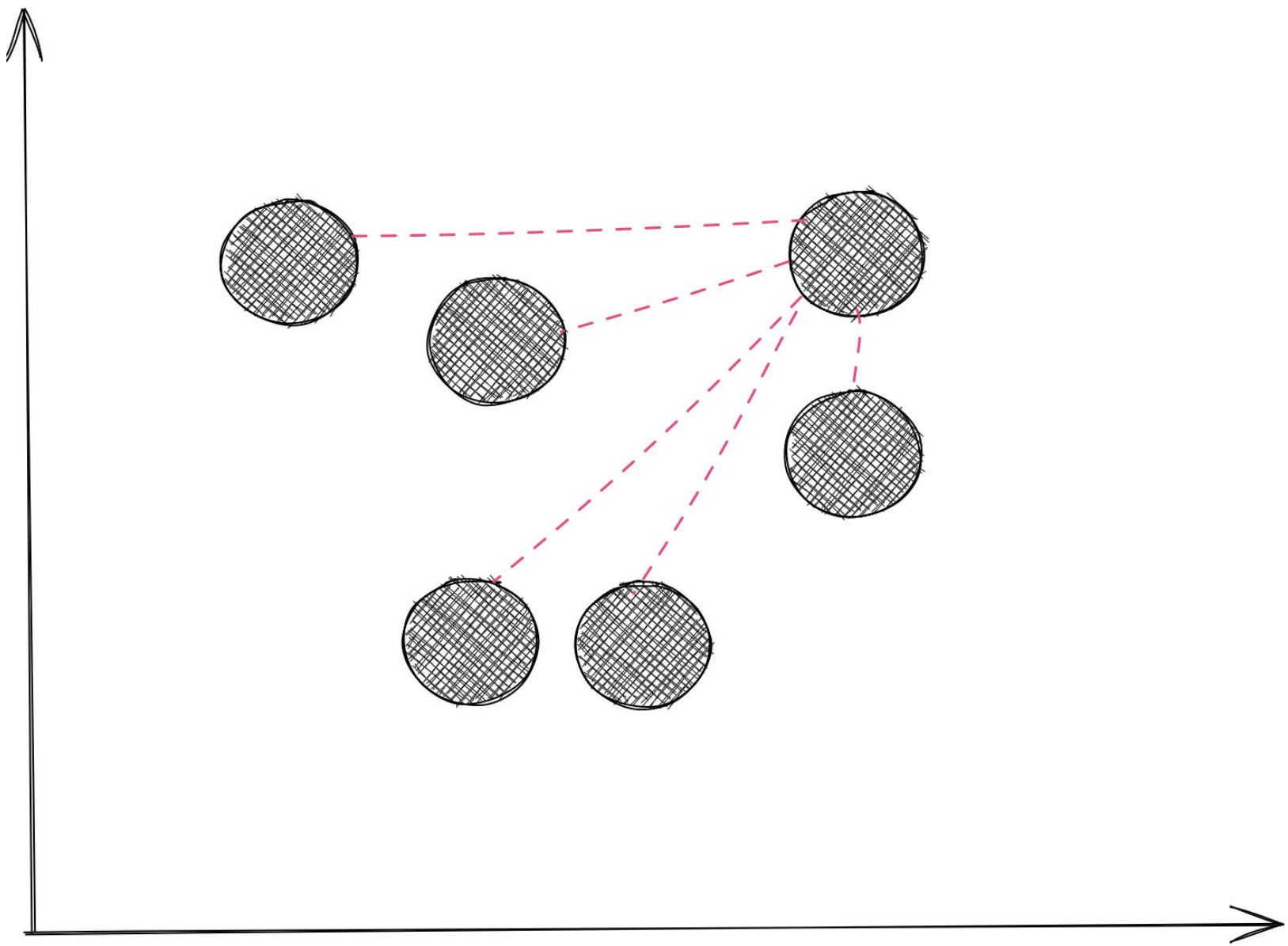
- **a** es el promedio de las disimilitudes (o distancias) de la observación **i** con las demás observaciones del cluster al que pertenece **i**.
- **b** es la distancia mínima a otro cluster que no es el mismo en el que está la observación **i**.

Ese cluster es la segunda mejor opción para **i** y se lo denomina vecindad de **i**.

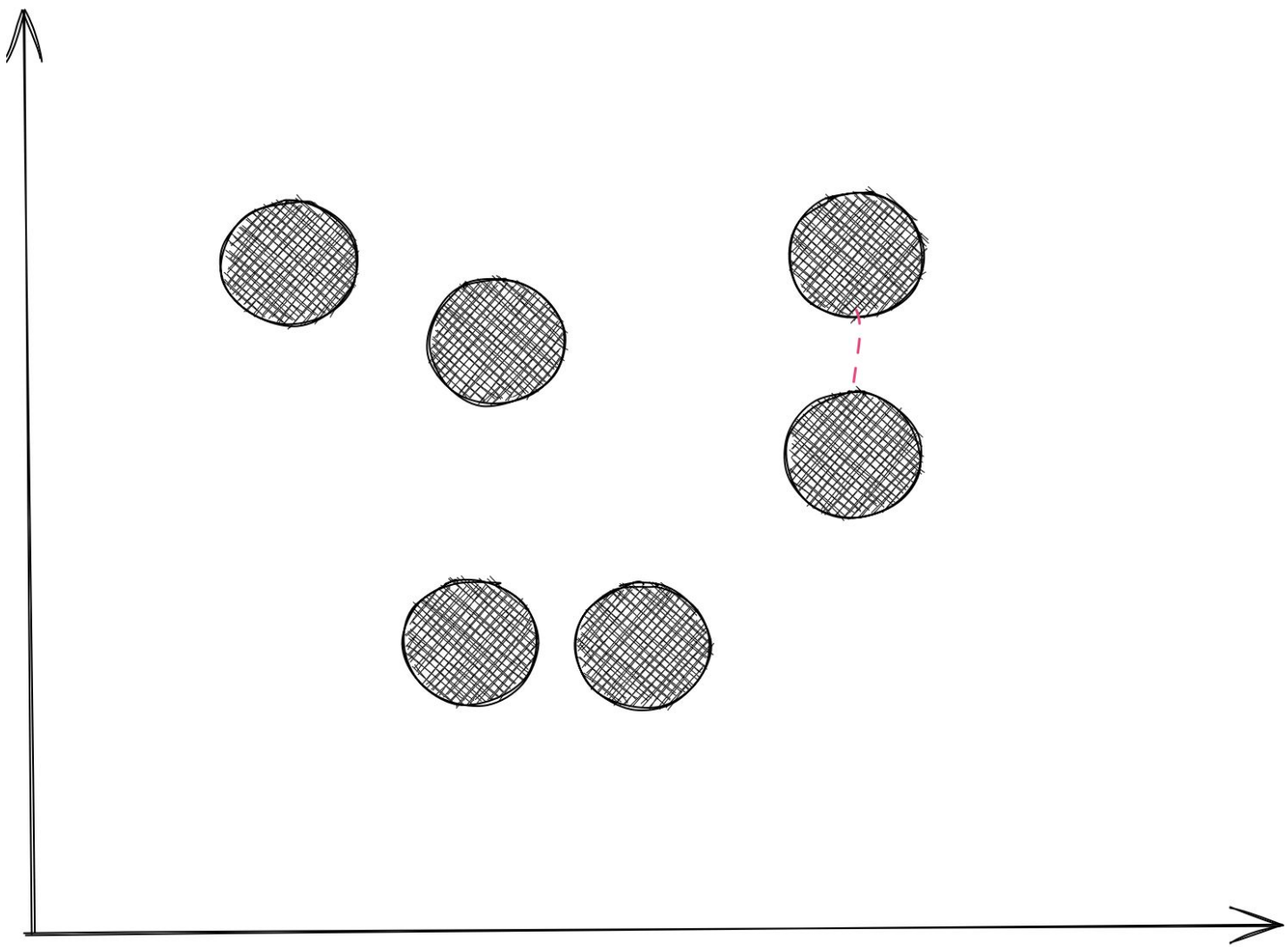
Evaluando clusters con K-means

Hierarchical Clustering

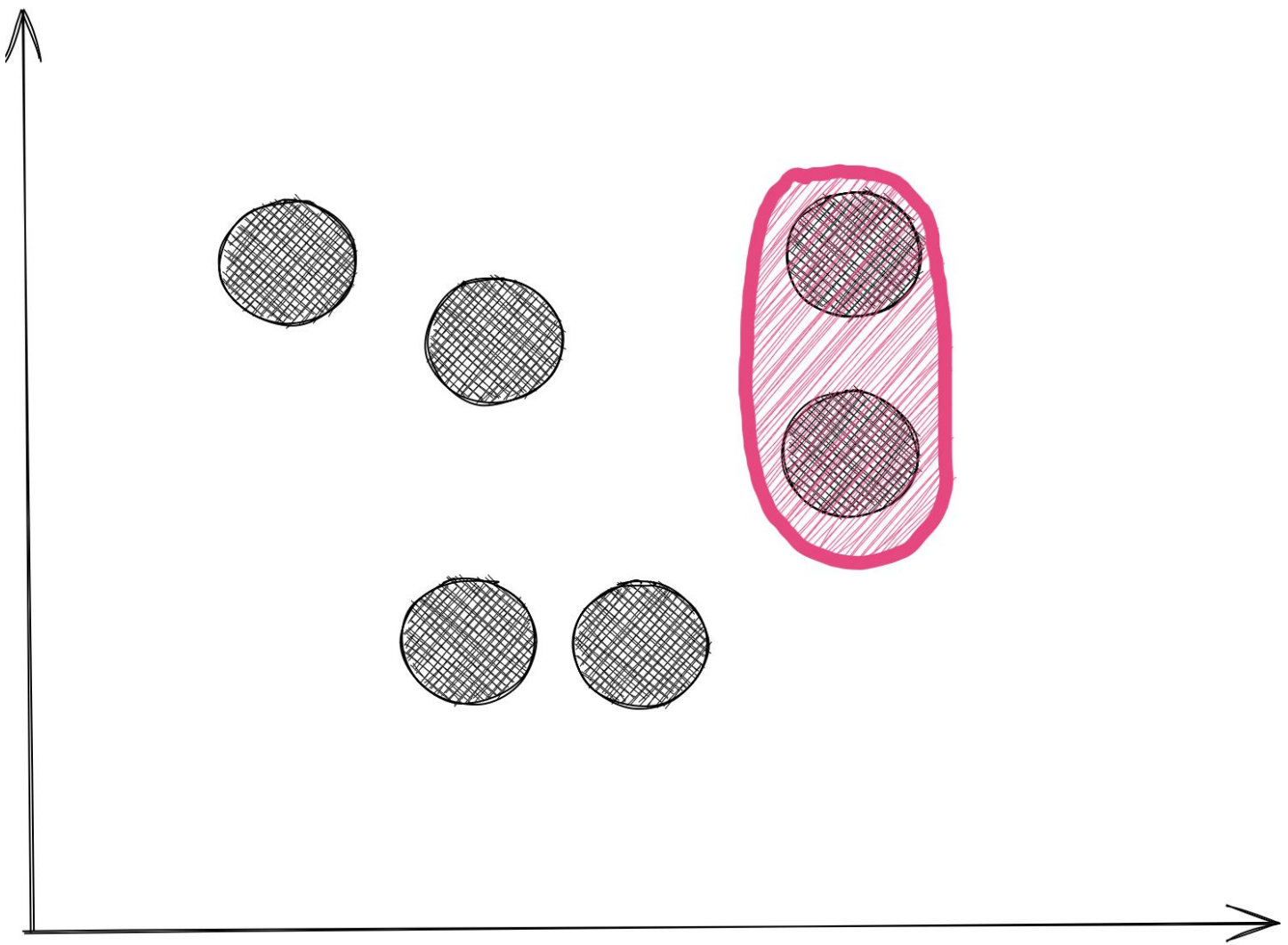




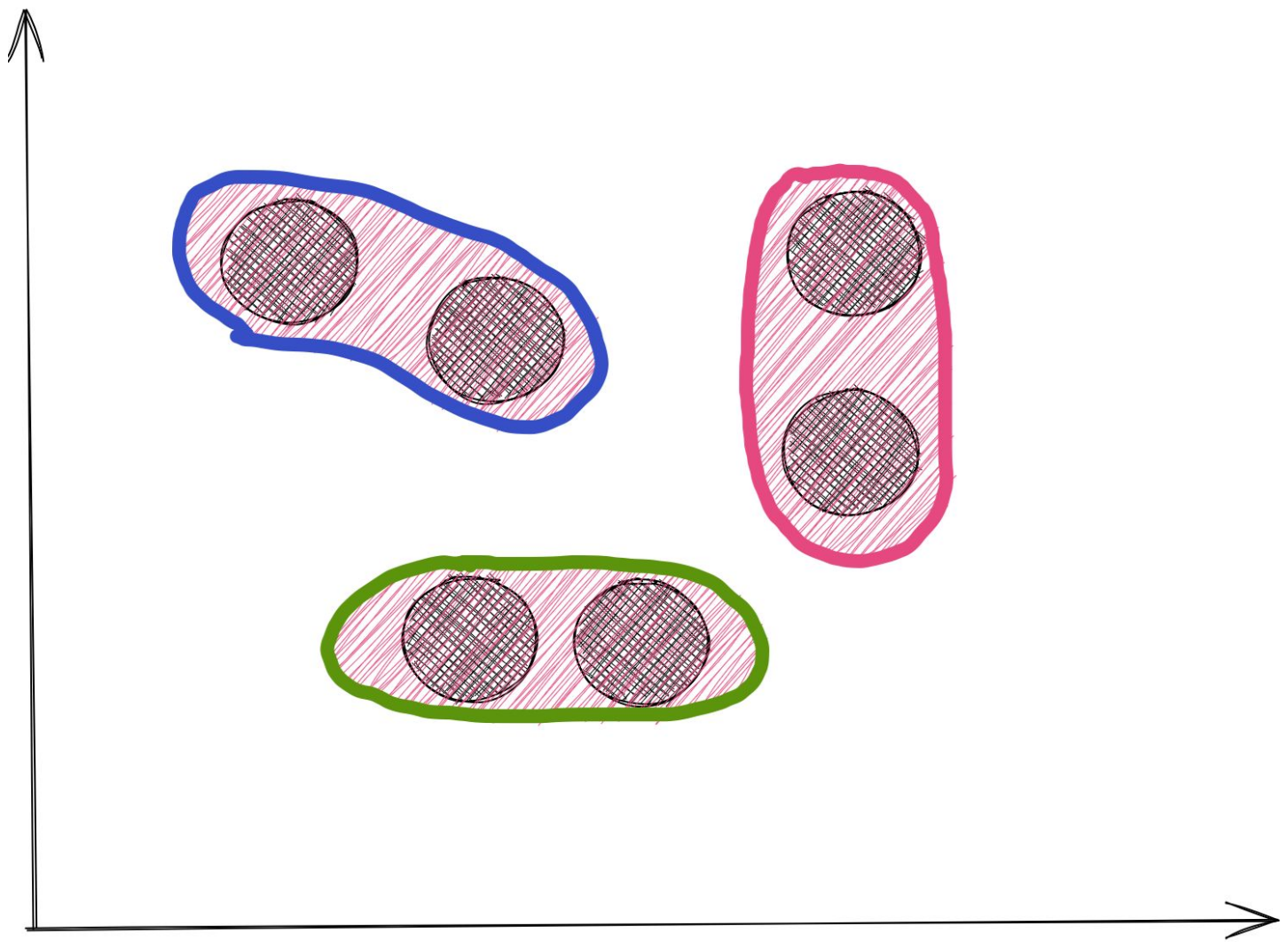
Hierarchical clustering



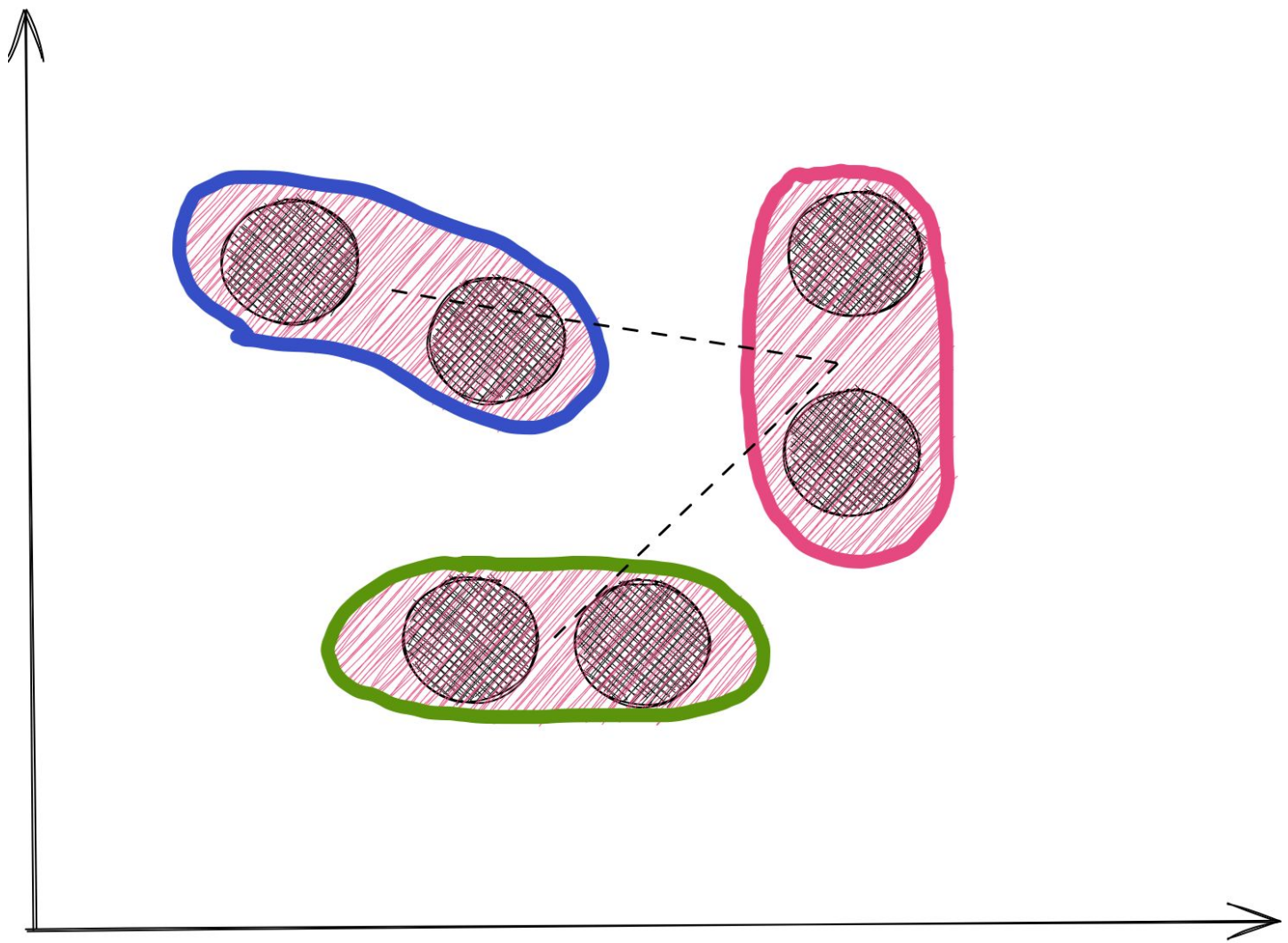
Hierarchical clustering



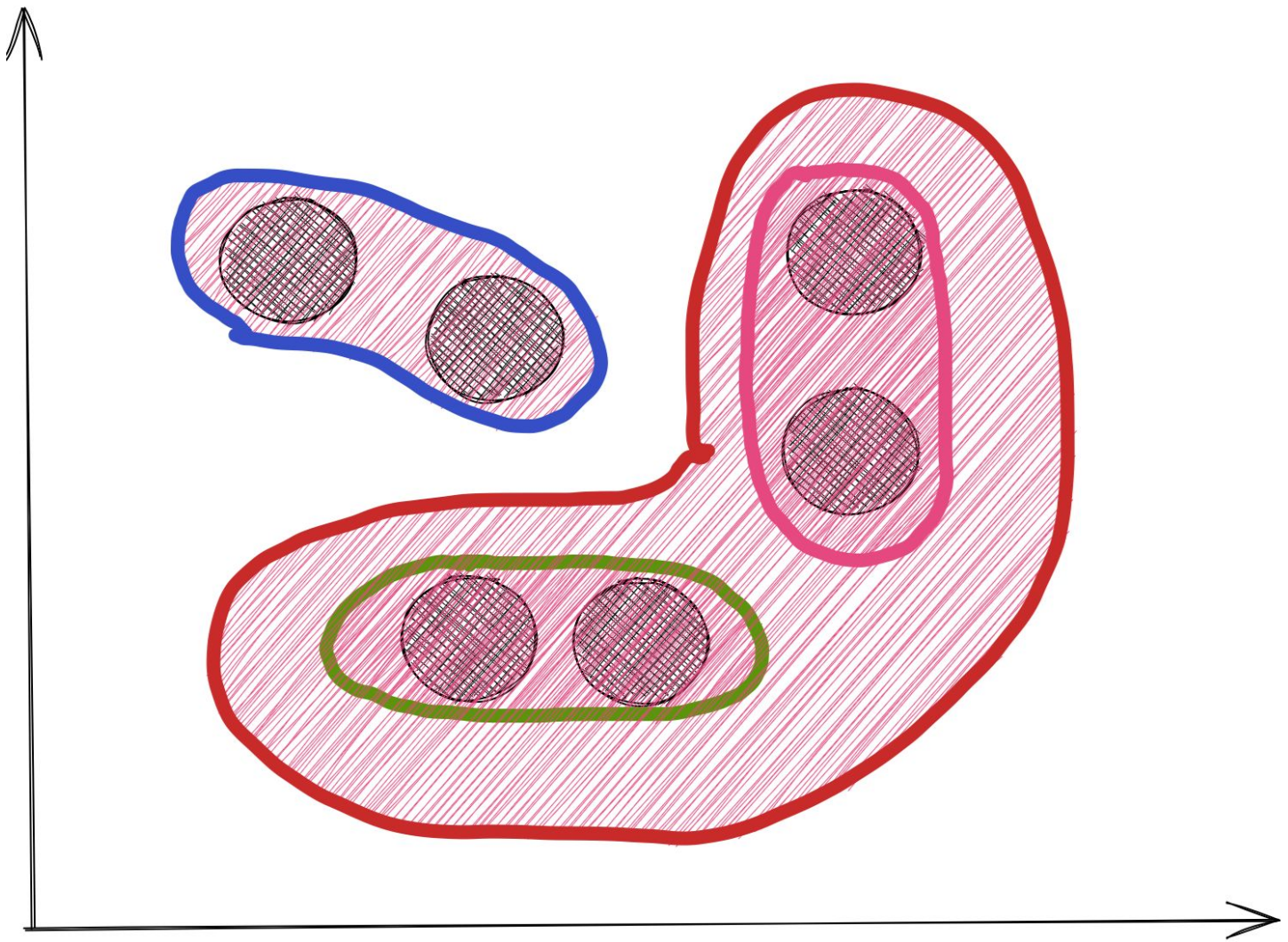
Hierarchical clustering



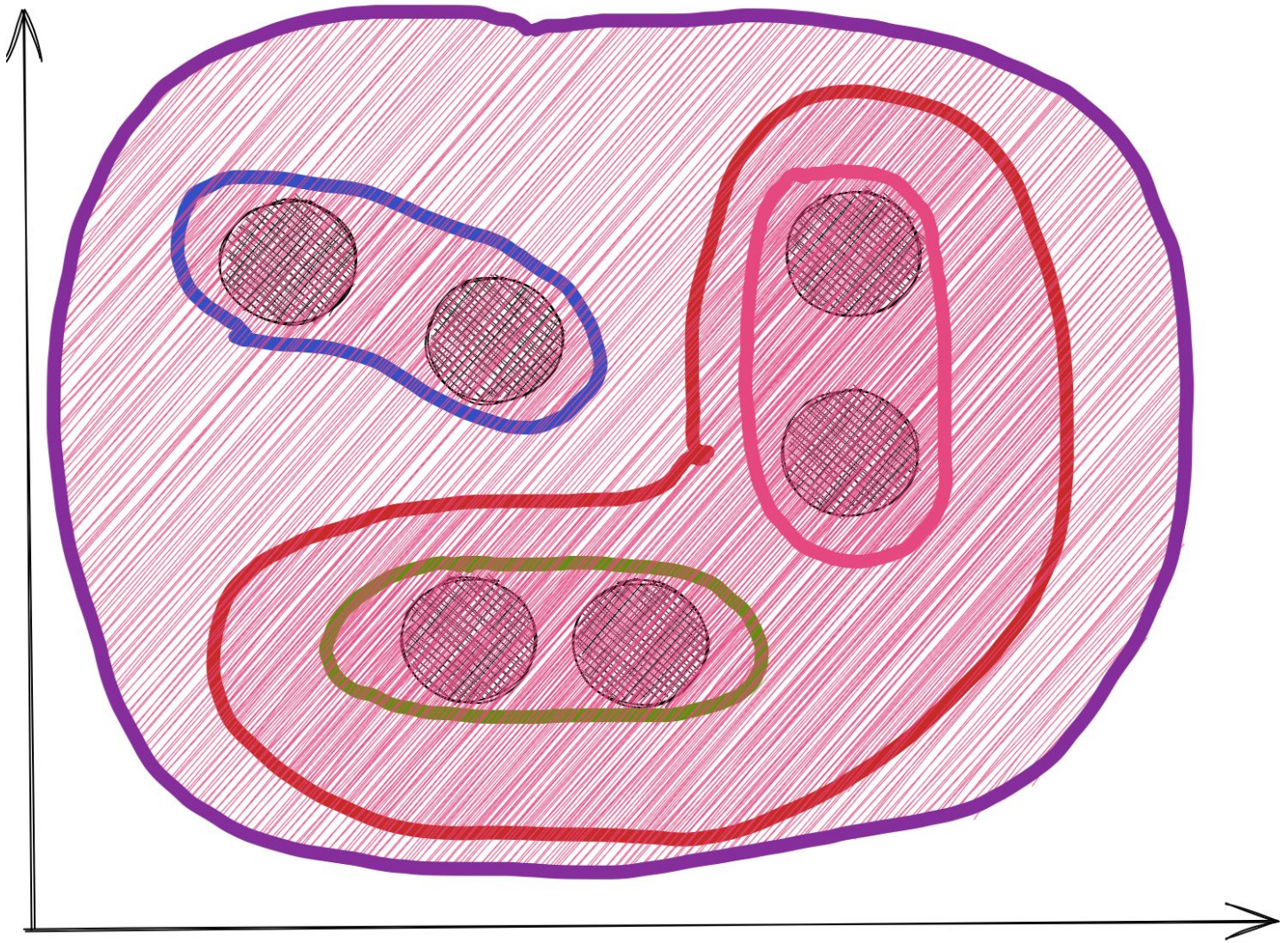
Hierarchical clustering



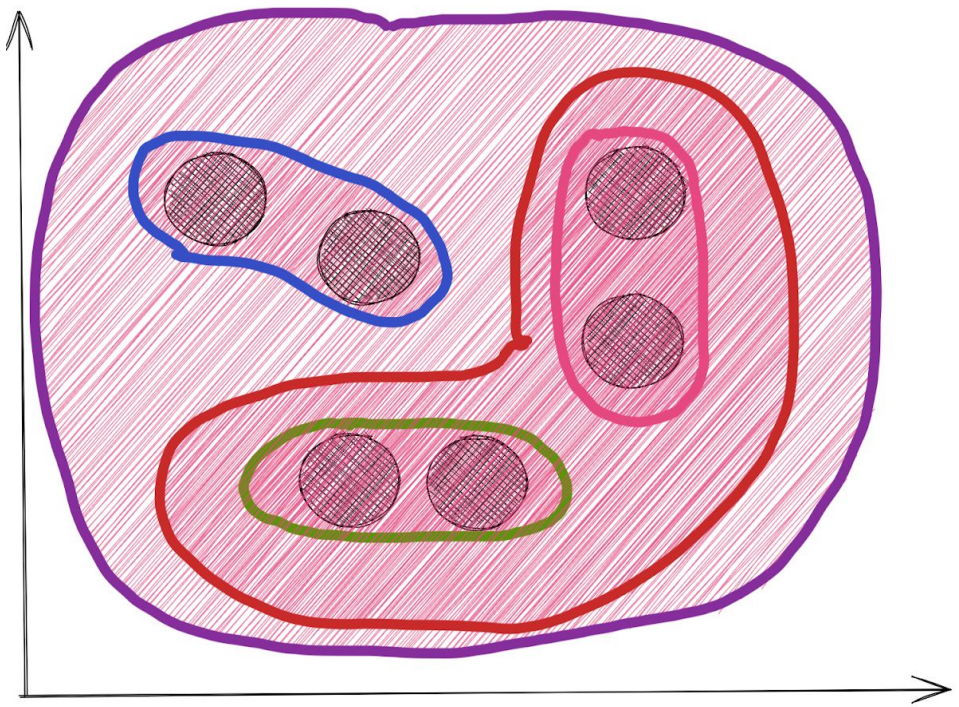
Hierarchical clustering

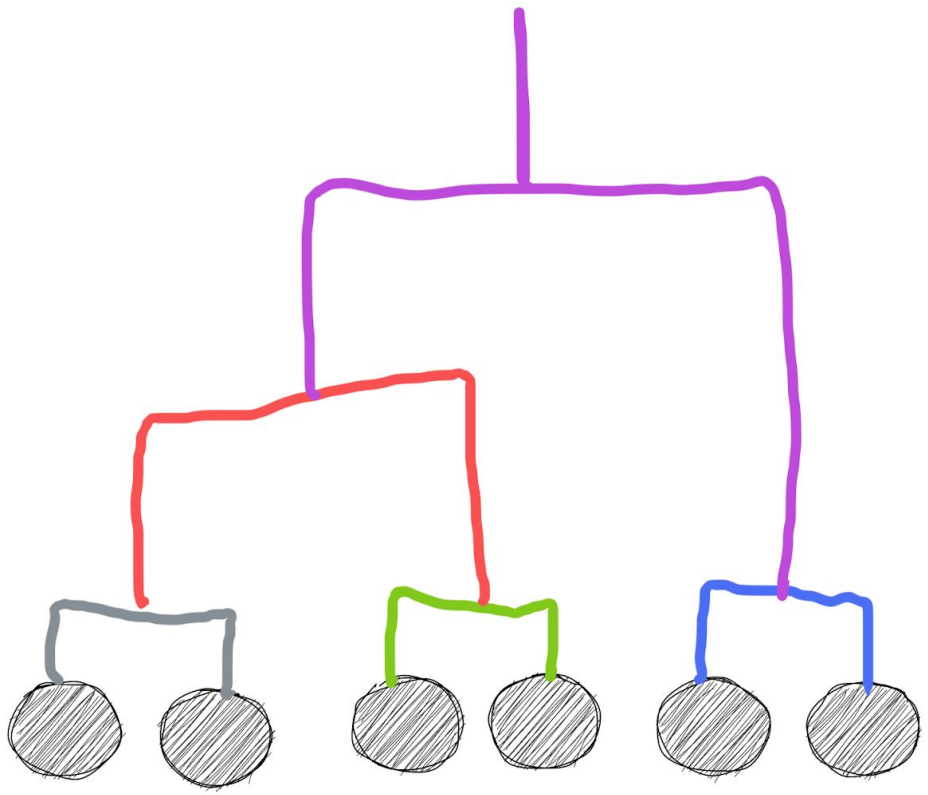


Hierarchical clustering



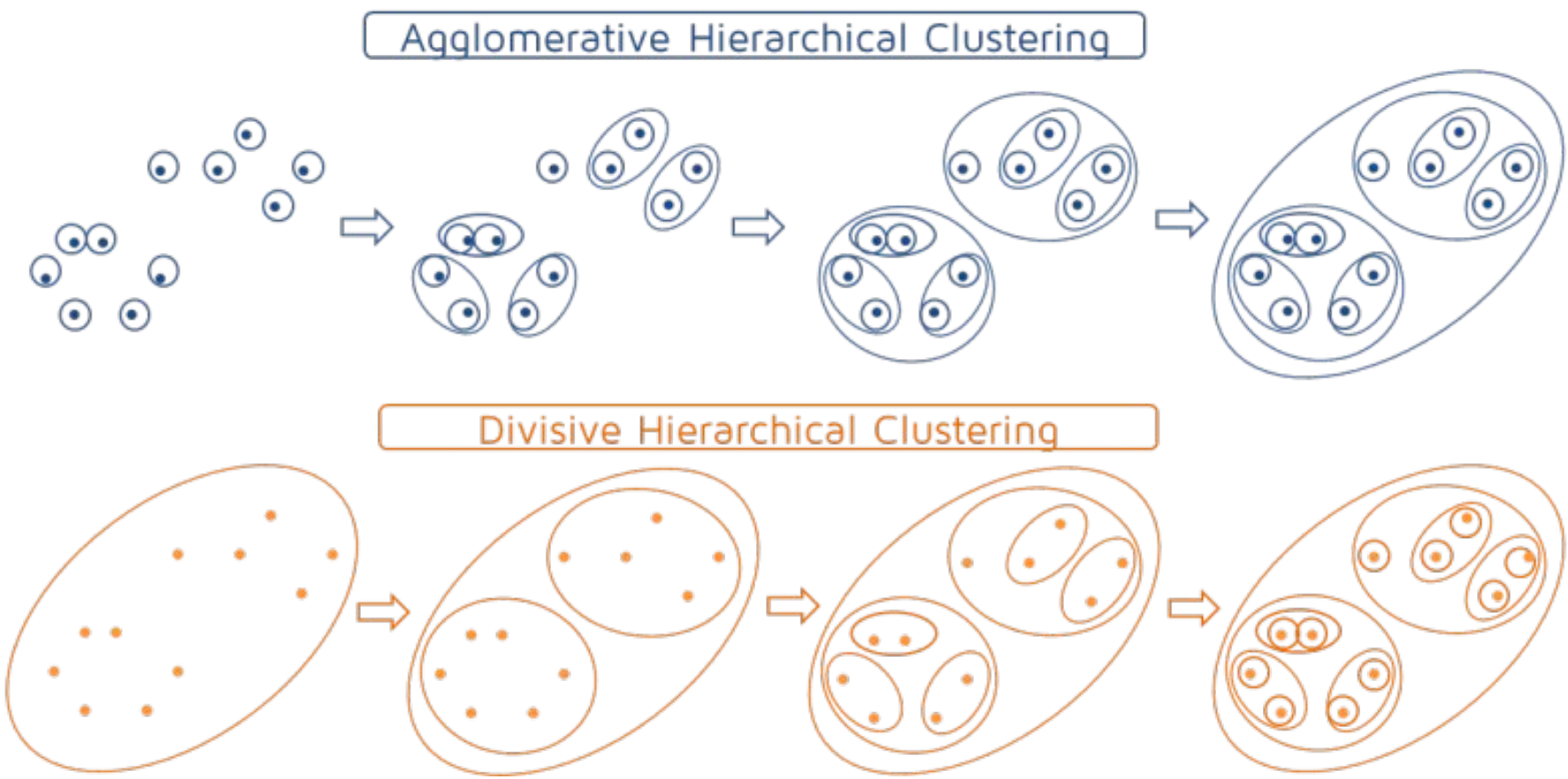
Hierarchical clustering





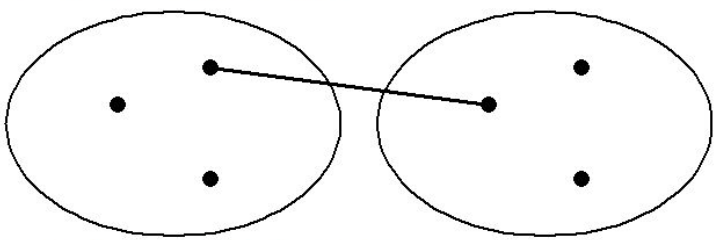
Dendrograma

Hierarchical clustering

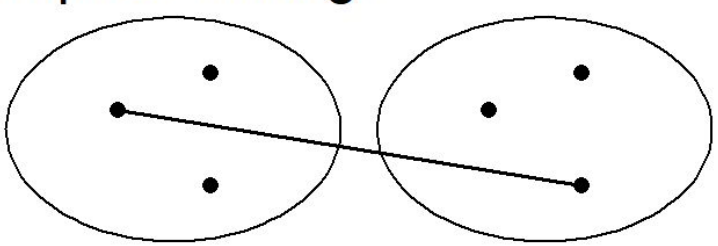


Linkage

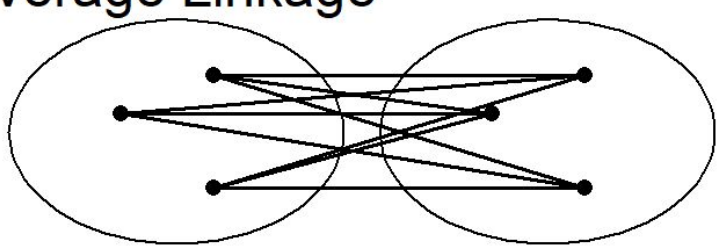
Simple Linkage



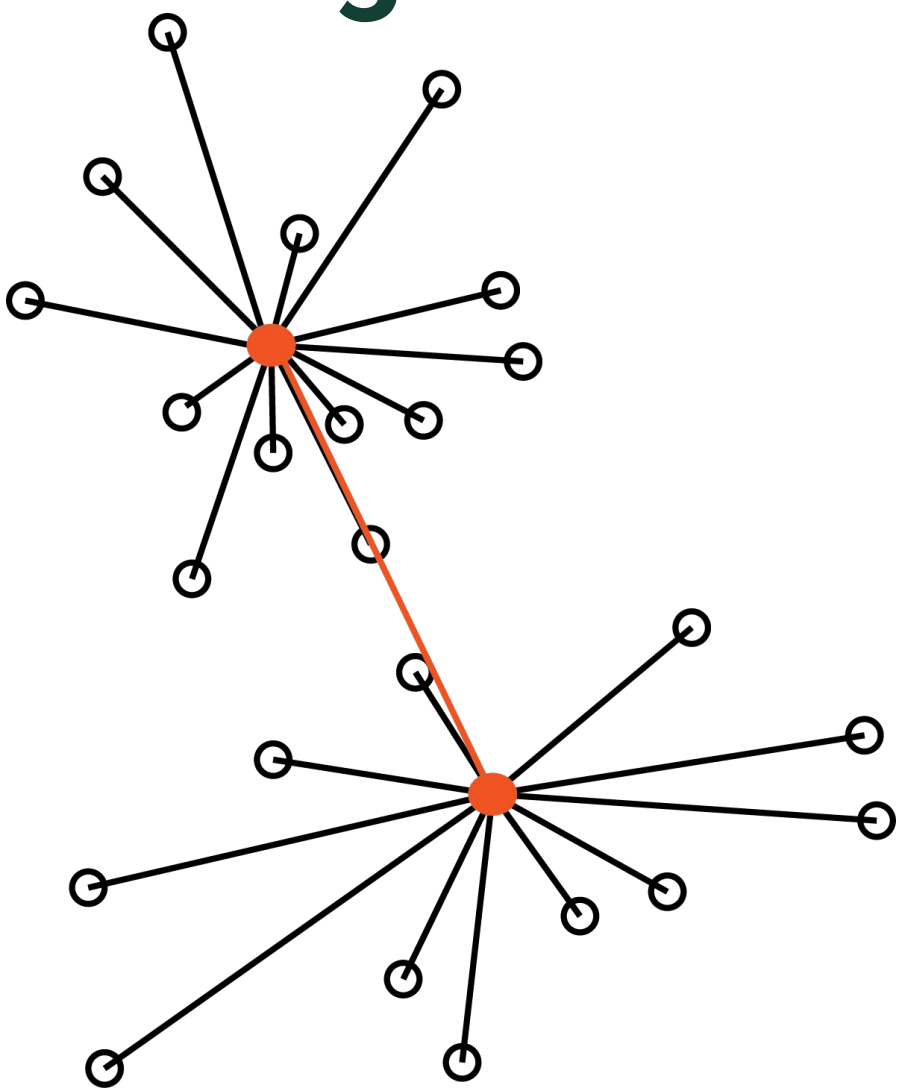
Complete Linkage



Average Linkage



Linkage - Ward



¿Cuándo usar hierarchical clustering?

Ventajas agglomerative clustering

- No necesito el número de clusters.
- Simple.
- Resultados interpretables.
- Única ejecución.
- Ayuda visual del dendrograma.

Desventajas agglomerative clustering

- Tarda en datasets largos.
- No tiene un objetivo matemático.
- Le afectan outliers drásticamente.
- Mayor necesidad de cómputo.

¿Cuándo usarlo?

- Comprender resultados manera visual.
- Tengo un dataset pequeño.
- Desconozco la cantidad de clusters por completo.
- Resultados rápidos.

Implementando hierarchical clustering

Evaluando nuestro cluster con hierarchical clustering

DBSCAN

Density-Based Spatial Clustering
of Aplication with Noise

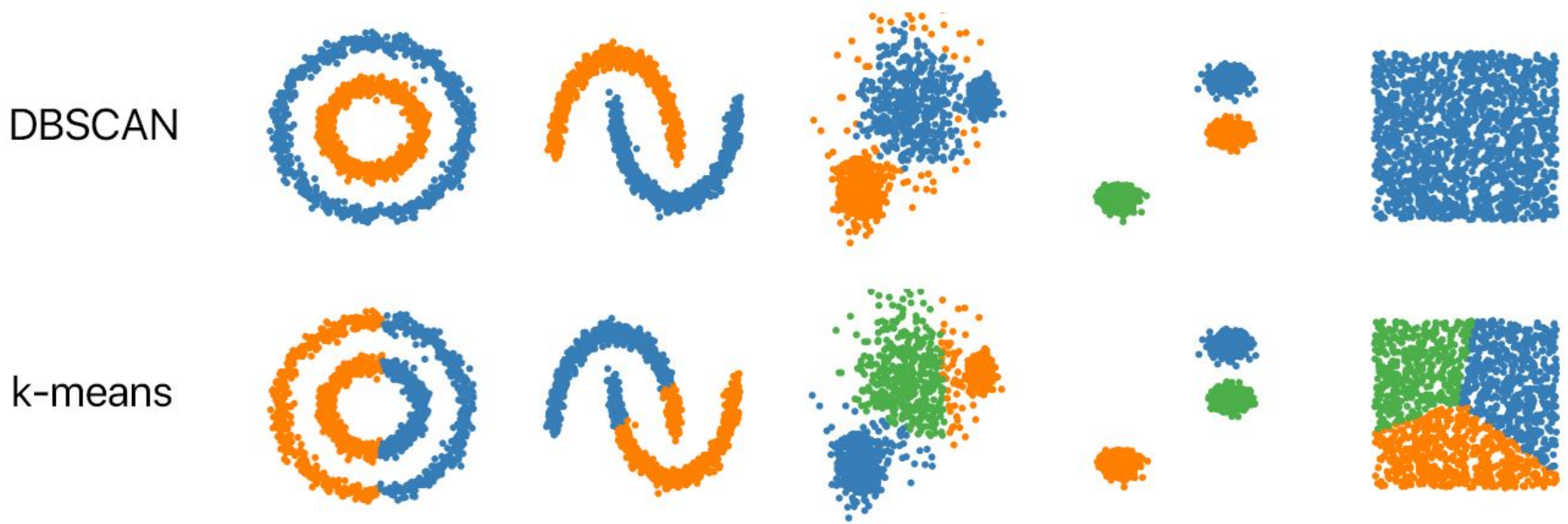
DBSCAN

- Clustering basado en densidad.
- Su principio se basa en que un cluster tiene alta densidad, mientras que la región que los separa no la tiene.

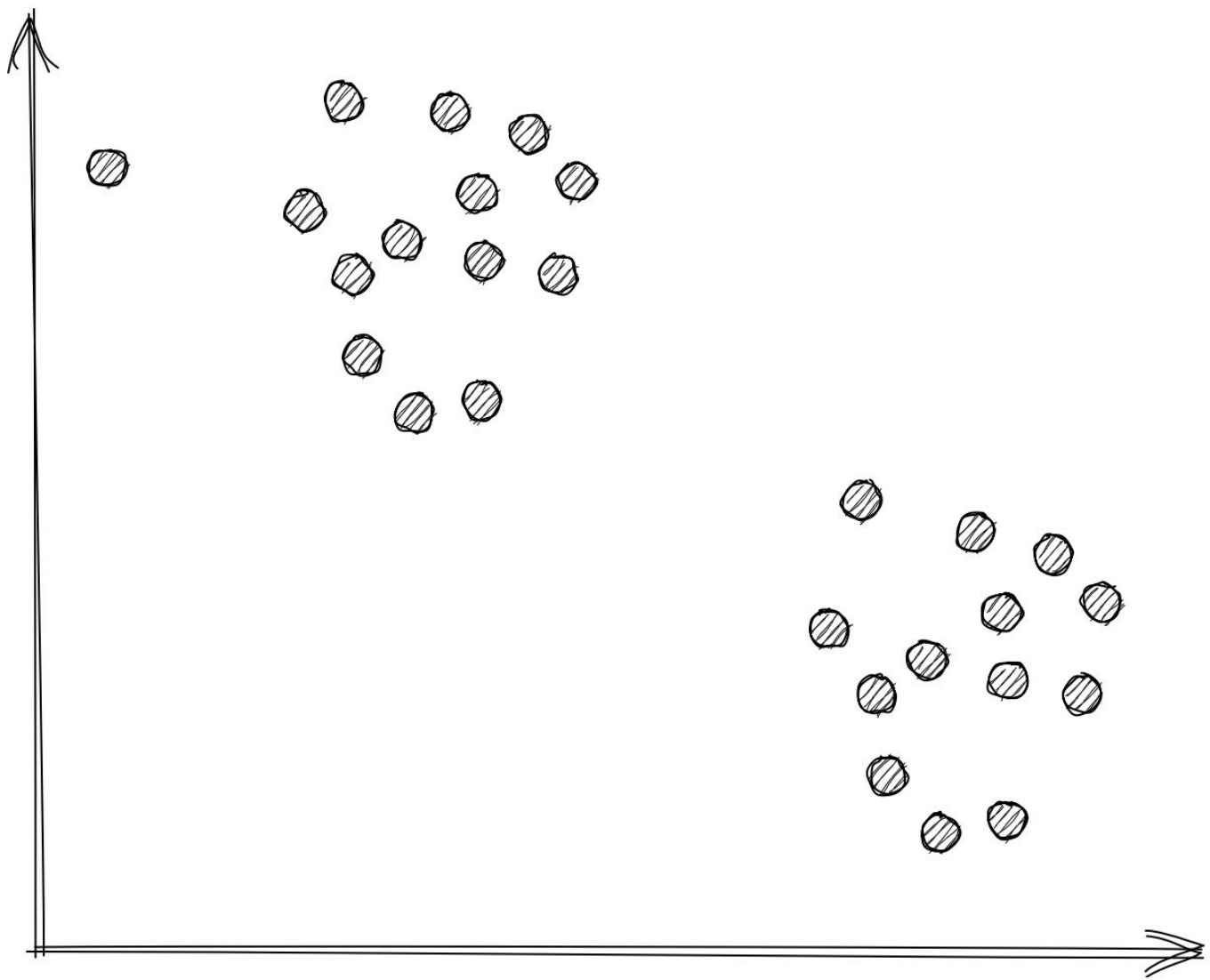
DBSCAN

- La idea principal es que cada punto en el cluster de acuerdo a un radio dado debe tener un mínimo de vecinos o puntos cercanos para encontrar densidad.
- Requiere dos parámetros:
  - Eps (Epsilon)
  - MinPts (Mínimo de puntos)

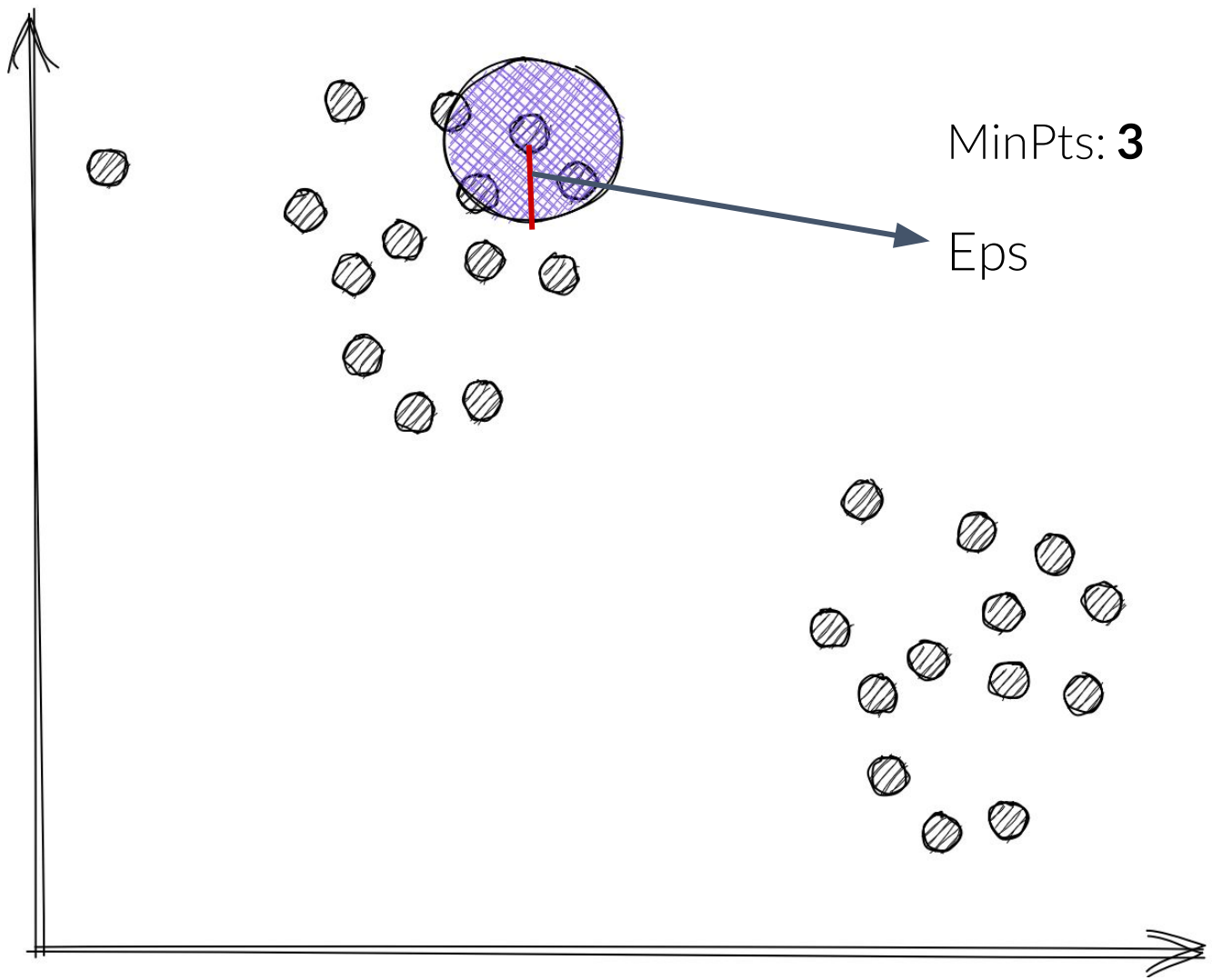
DBSCAN



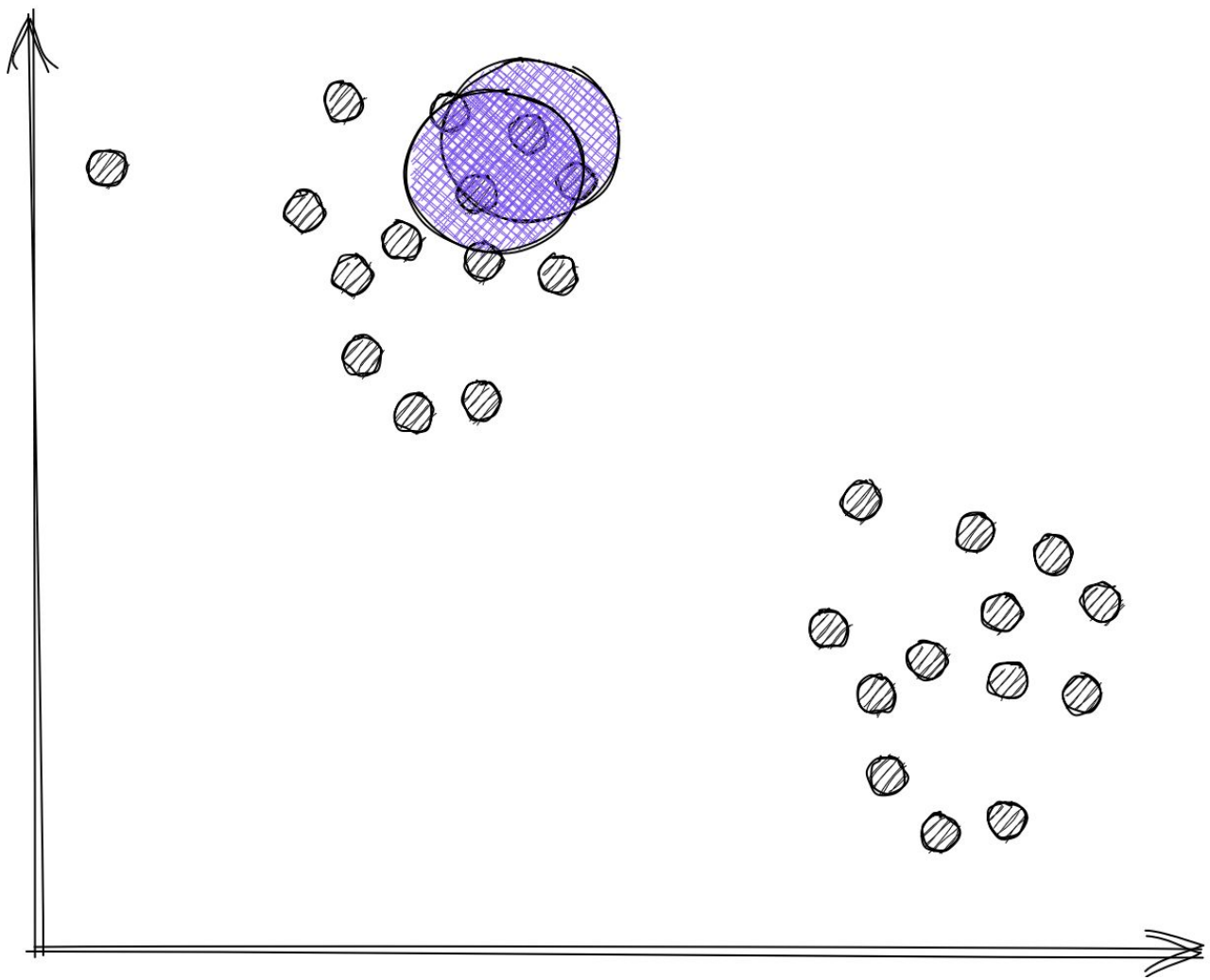
DBSCAN



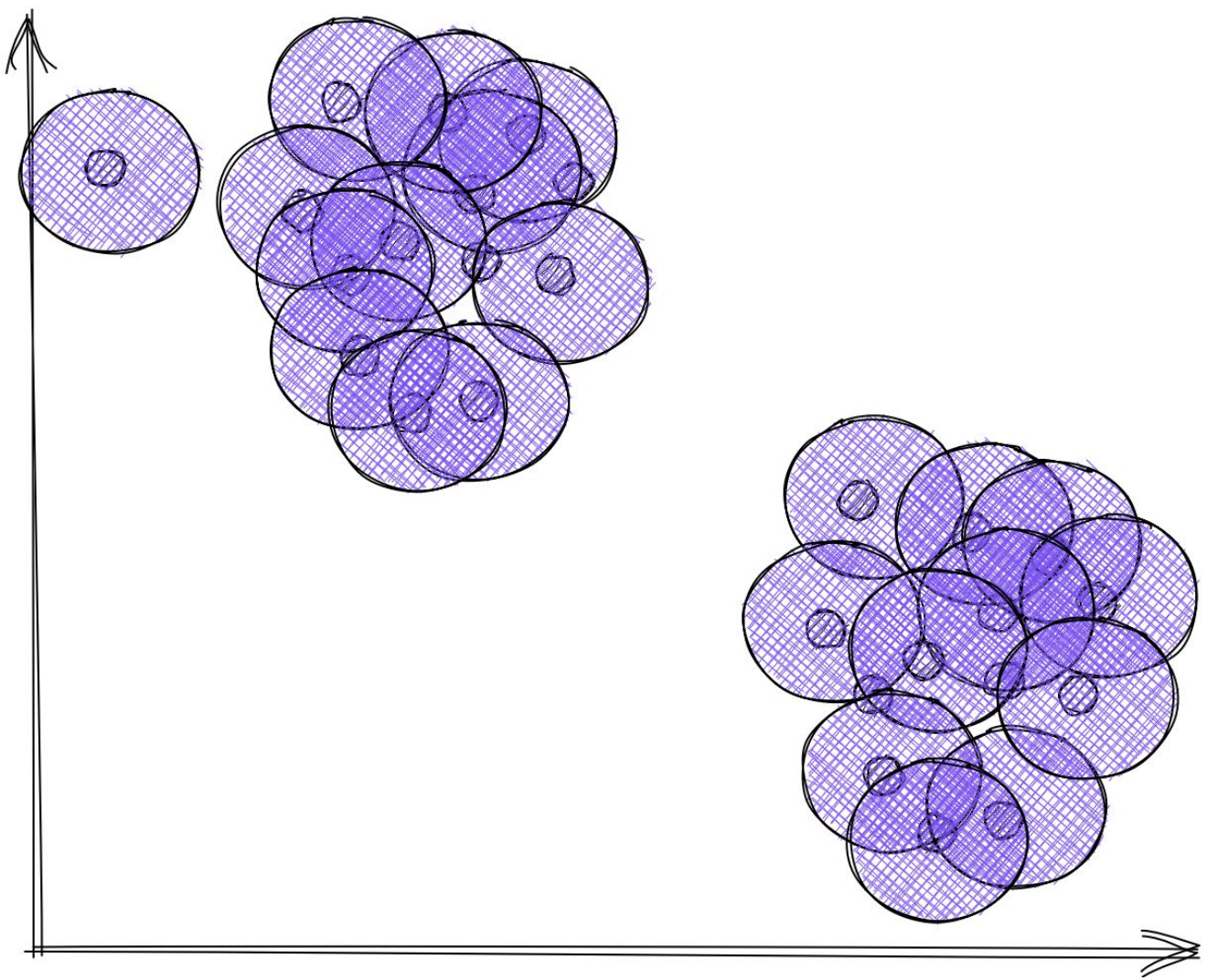
DBSCAN



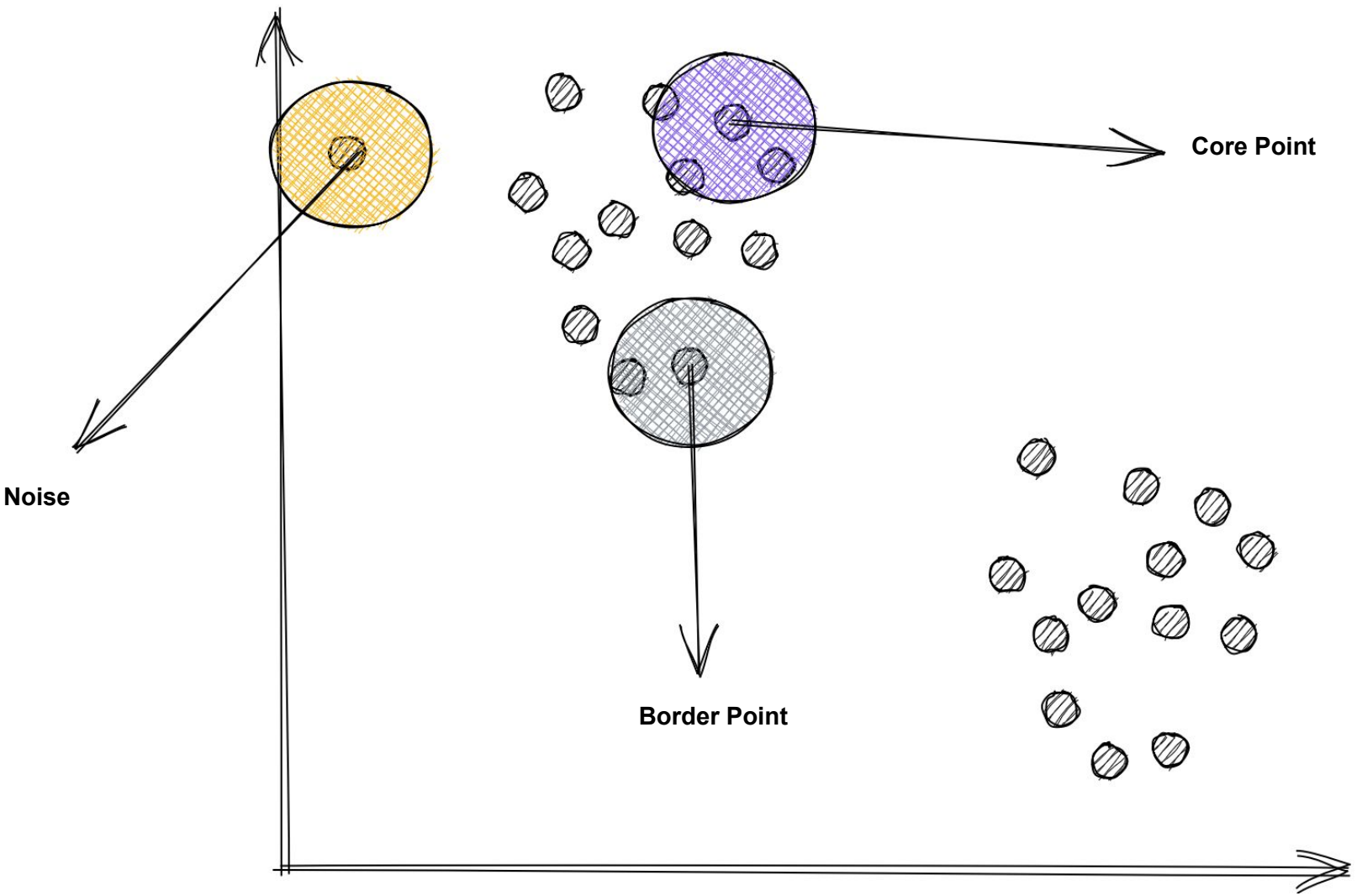
DBSCAN



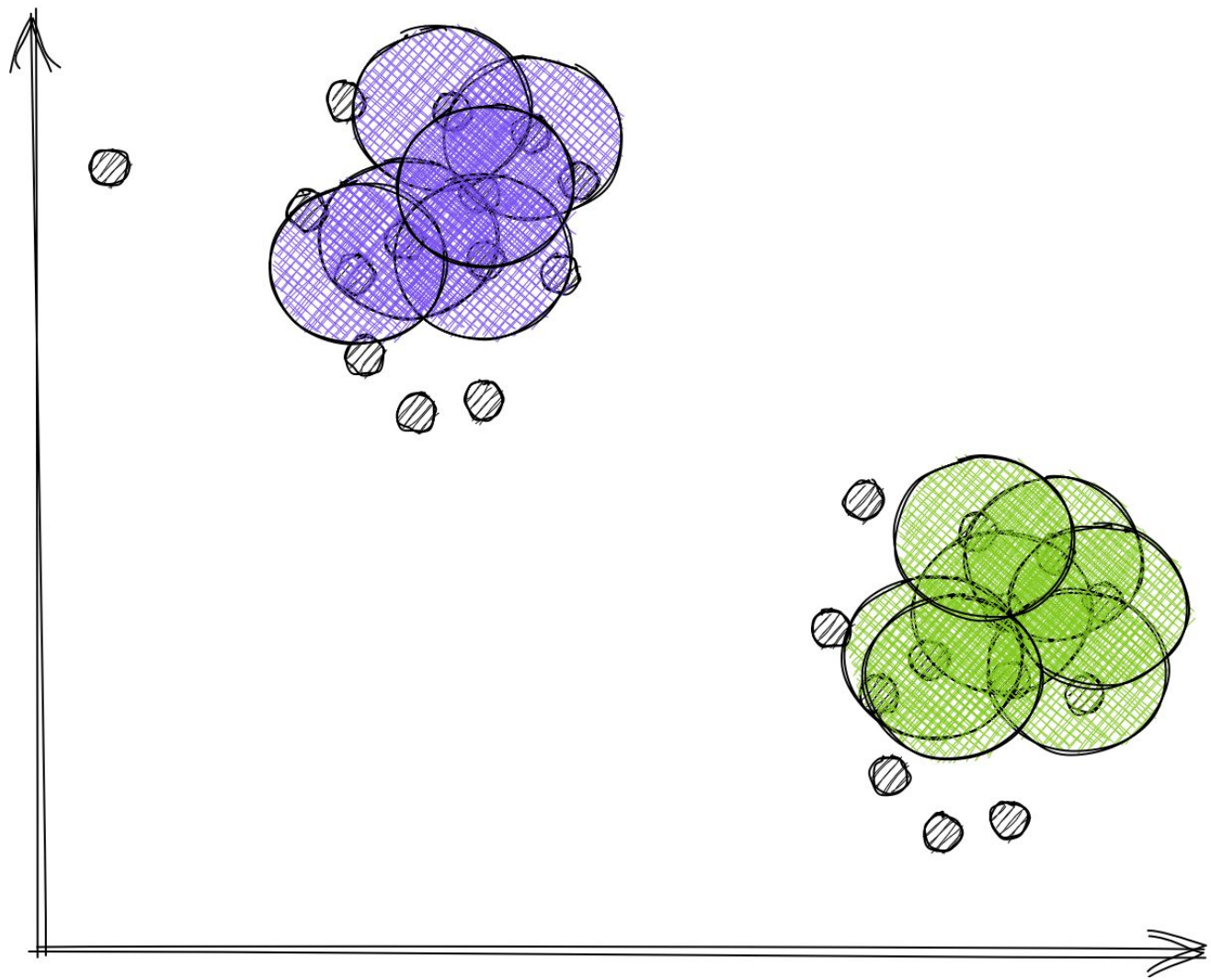
DBSCAN



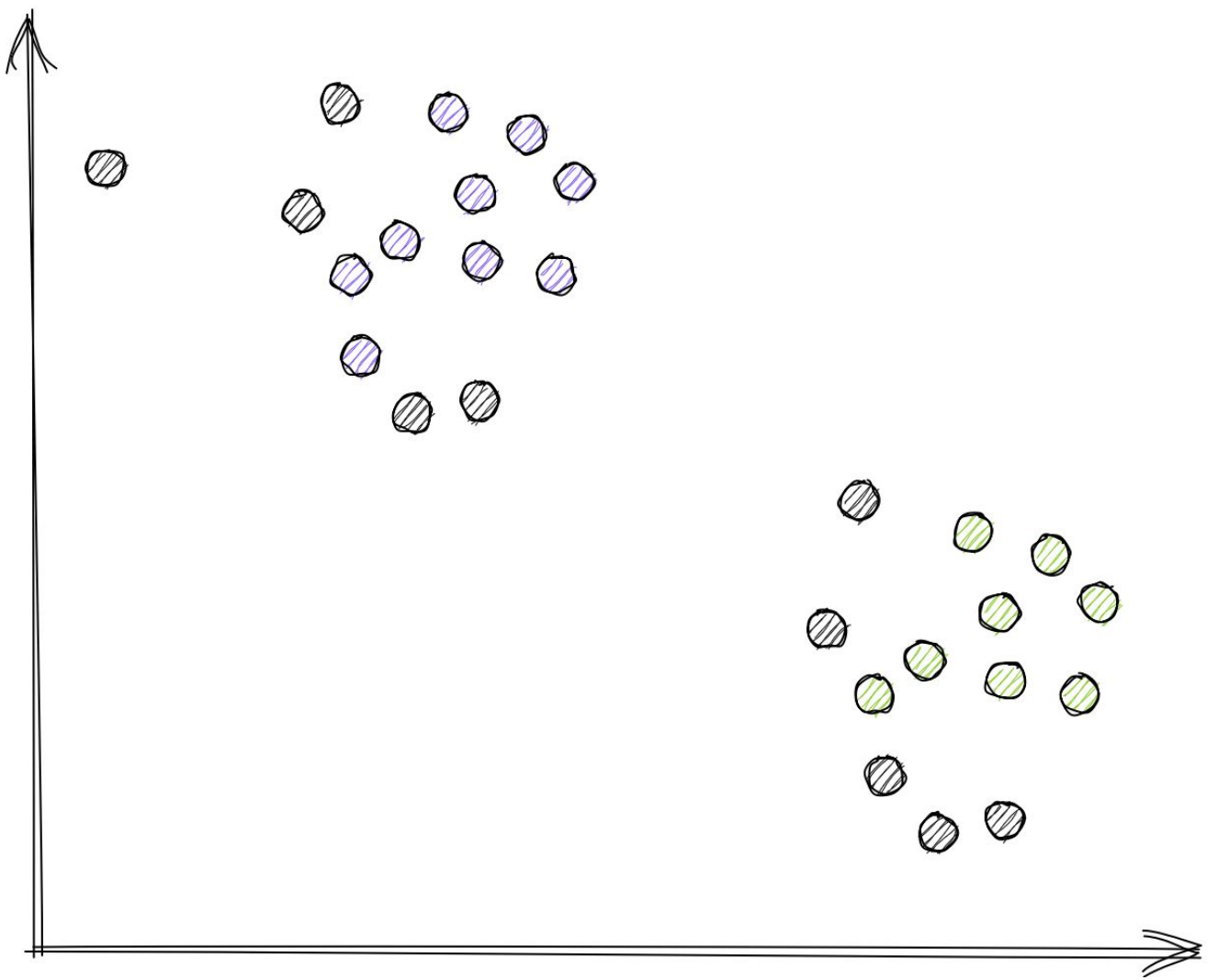
DBSCAN



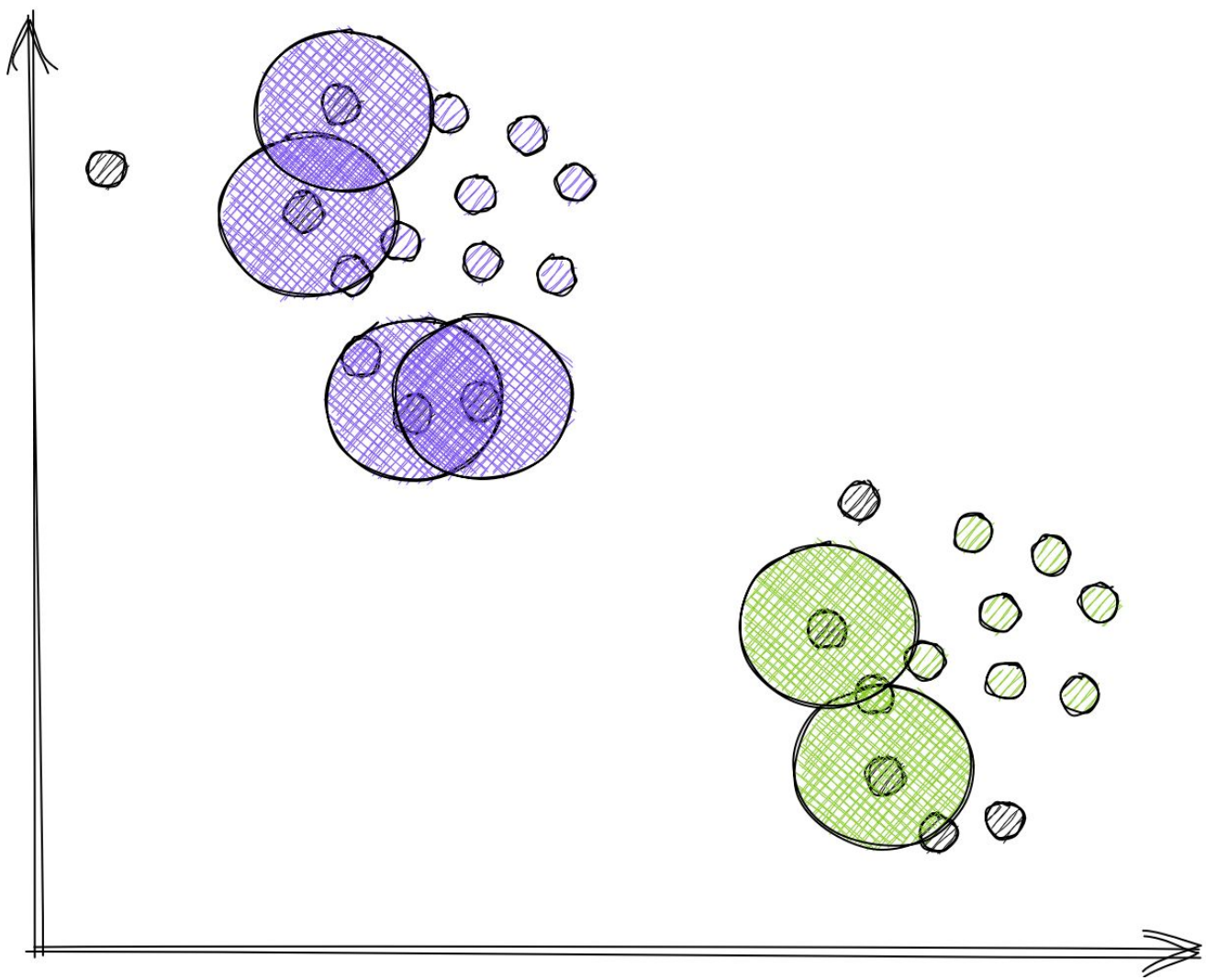
DBSCAN



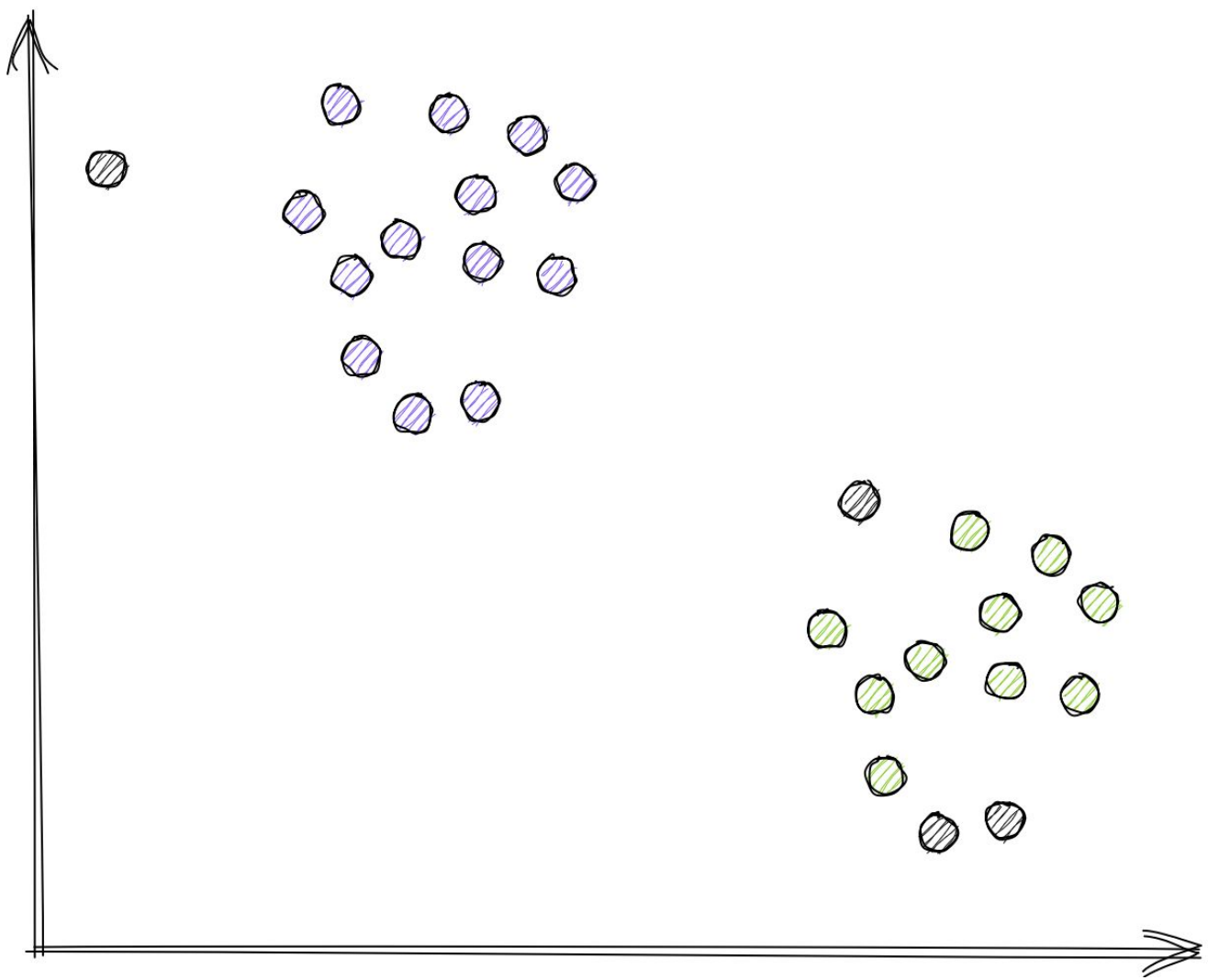
DBSCAN



DBSCAN



DBSCAN



¿Cuándo usar DBSCAN?

Ventajas DBSCAN

- No requiere especificar el número de clusters.
- Es capaz de detectar outliers o ruido.
- Puede encontrar clusters en formas y tamaños arbitrarios.

Desventajas DBSCAN

- Los híper-parámetros son muy determinantes, algunas combinaciones no funcionan igual para todos los grupos con distintas densidades.
- Los puntos fronterizos a los que se puede acceder desde más de un cluster pueden formar parte de cualquier cluster.

¿Cuándo usarlo?

- Desconozco la cantidad de clusters.
- No uso formas esféricas.
- Densidades similares entre clusters.

Implementando DBSCAN

Encontrar híper-parámetros

Evaluando resultados de DBSCAN

Preparar datos para clusterizar

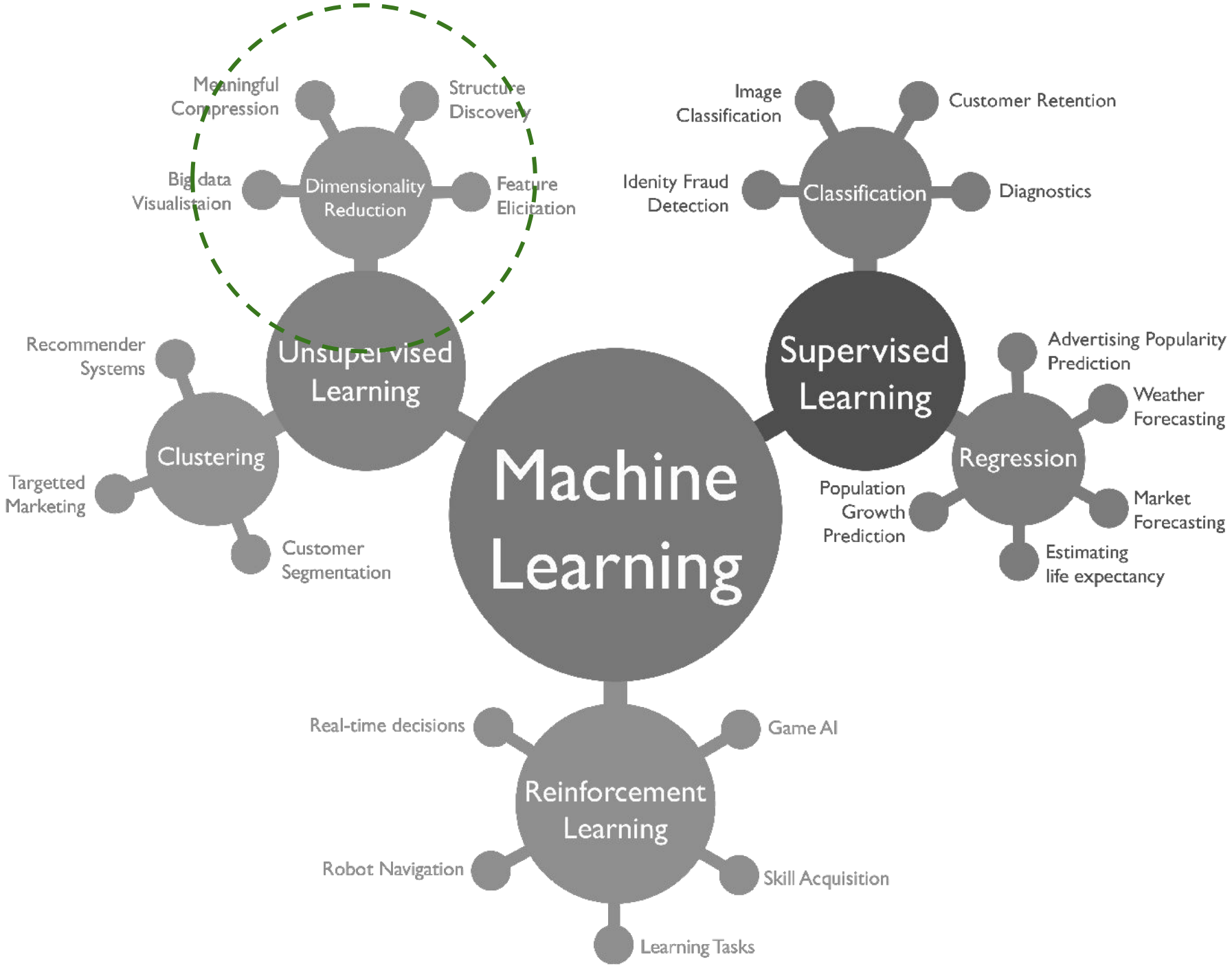
Preparar datos

- Eliminar duplicados.
- Remover columnas innecesarias.
- Remover outliers.
- Escalar data.

PCA para clustering

Preparar datos

- Eliminar duplicados.
- Remover columnas innecesarias.
- Remover outliers.
- Escalar data.
- **Reducción de dimensionalidad.**


Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Clustering
Classification
Regression
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Recommender Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Resolviendo con K-means

Resolviendo un problema con clustering

Resolviendo con hierarchical clustering

Resolviendo un problema con clustering

Resolviendo con DBSCAN

Resolviendo un problema con clustering

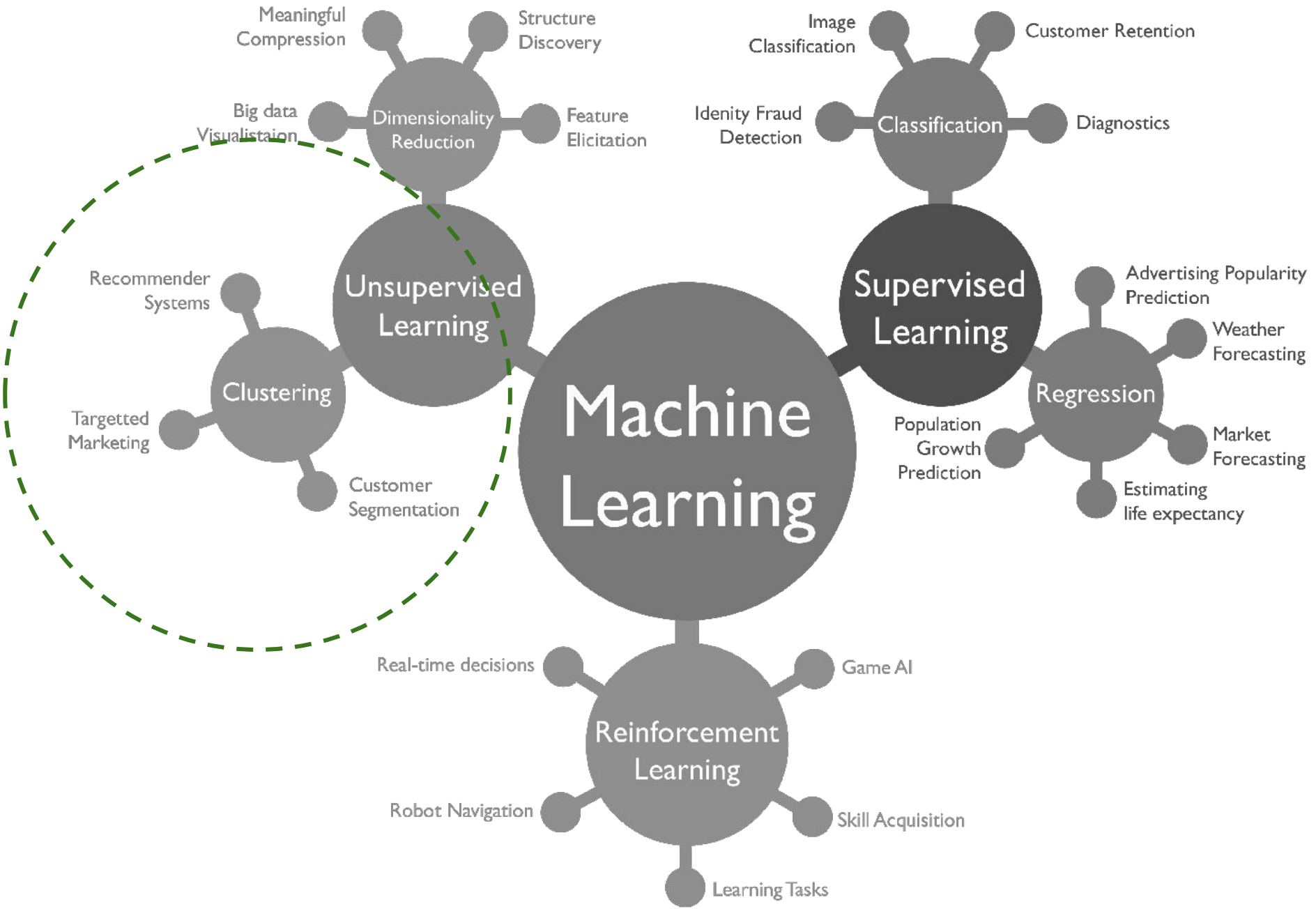
Resolviendo con DBSCAN (sin PCA)

Resolviendo un problema con clustering

Evaluación de resultados

Resolviendo un problema con clustering

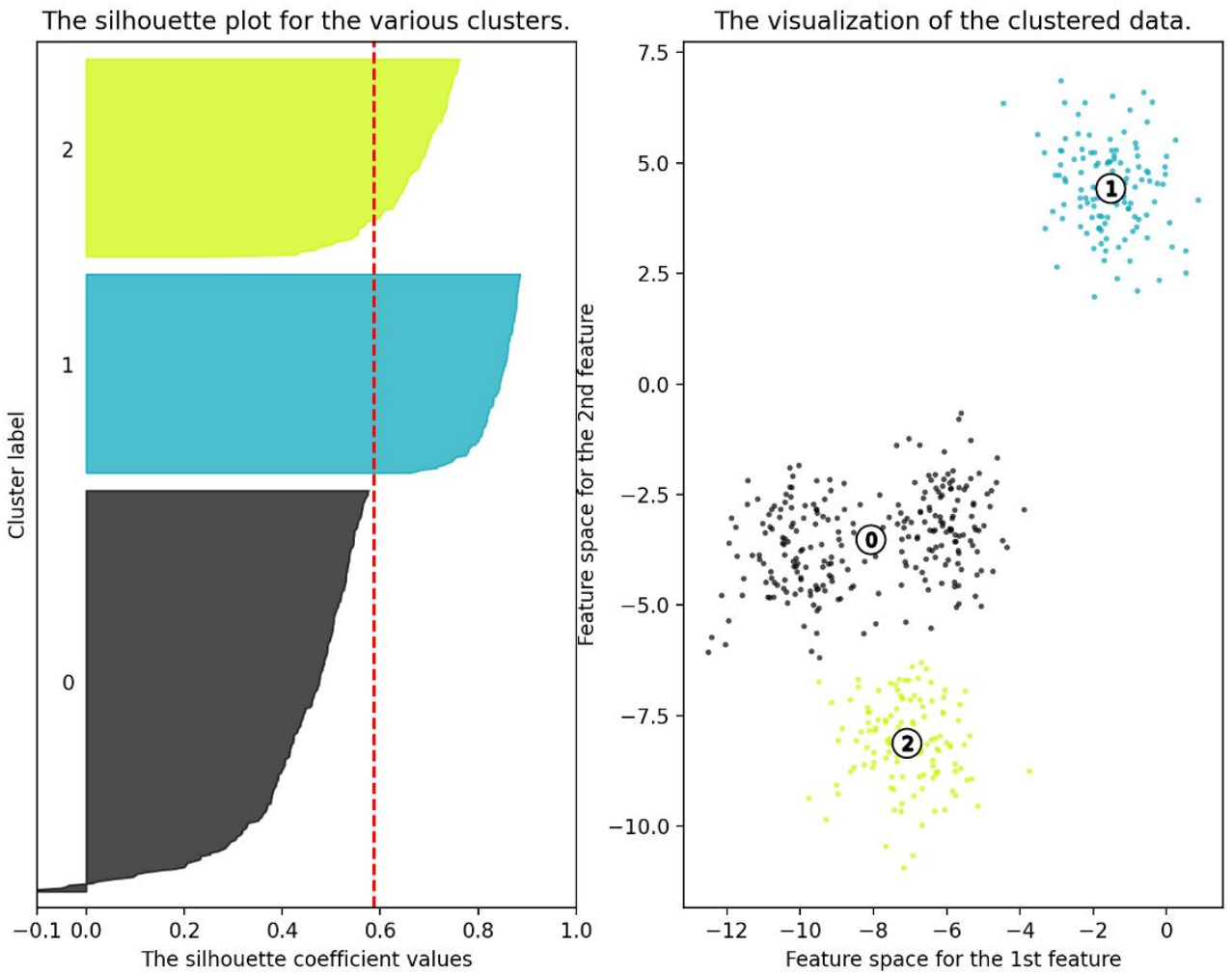
Proyecto final y cierre


Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Clustering
Classification
Regression
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Recommender Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

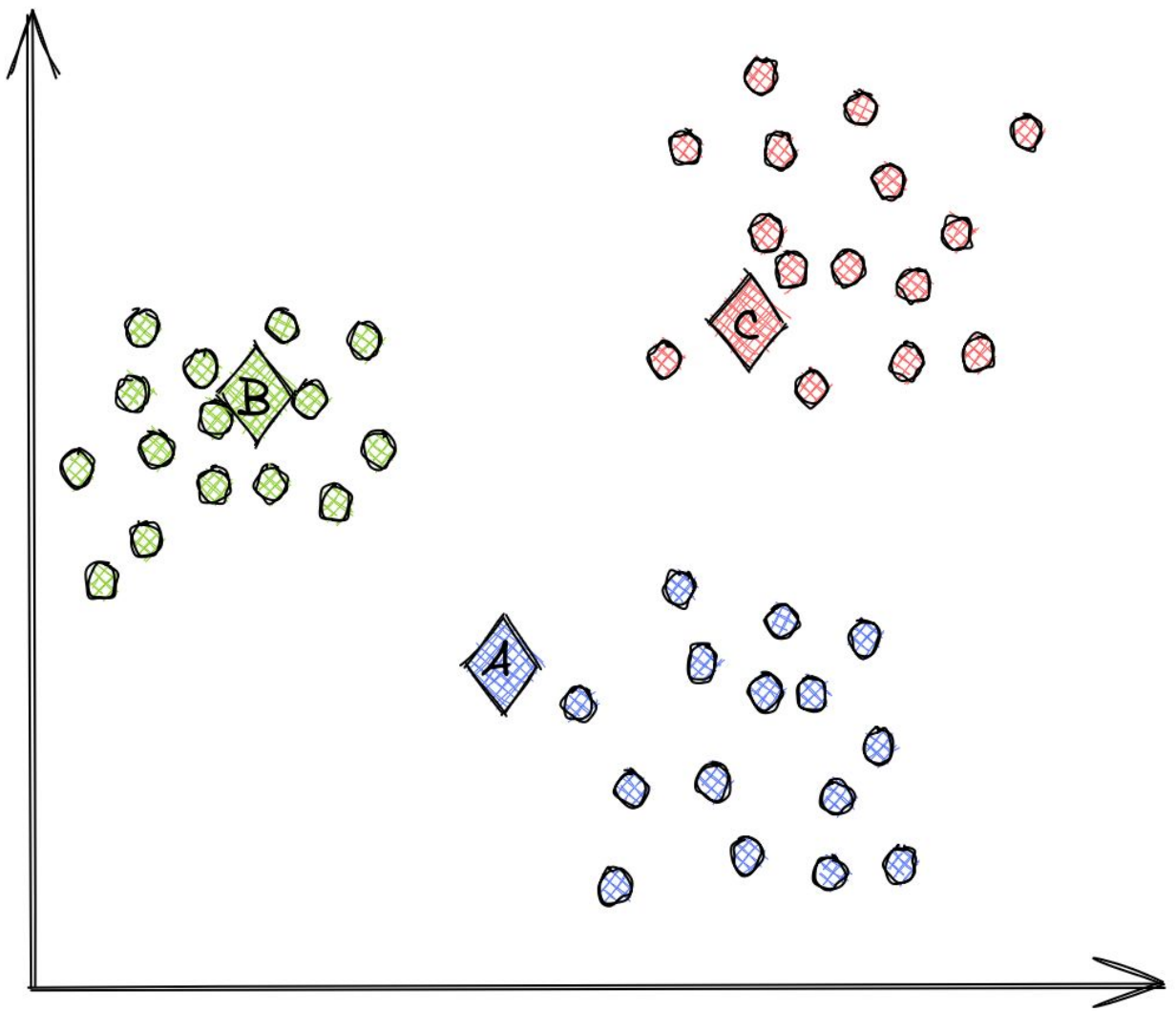
Preparar datos

- Eliminar duplicados.
- Remover columnas innecesarias.
- Remover outliers.
- Escalar data.
- Reducción de dimensionalidad.

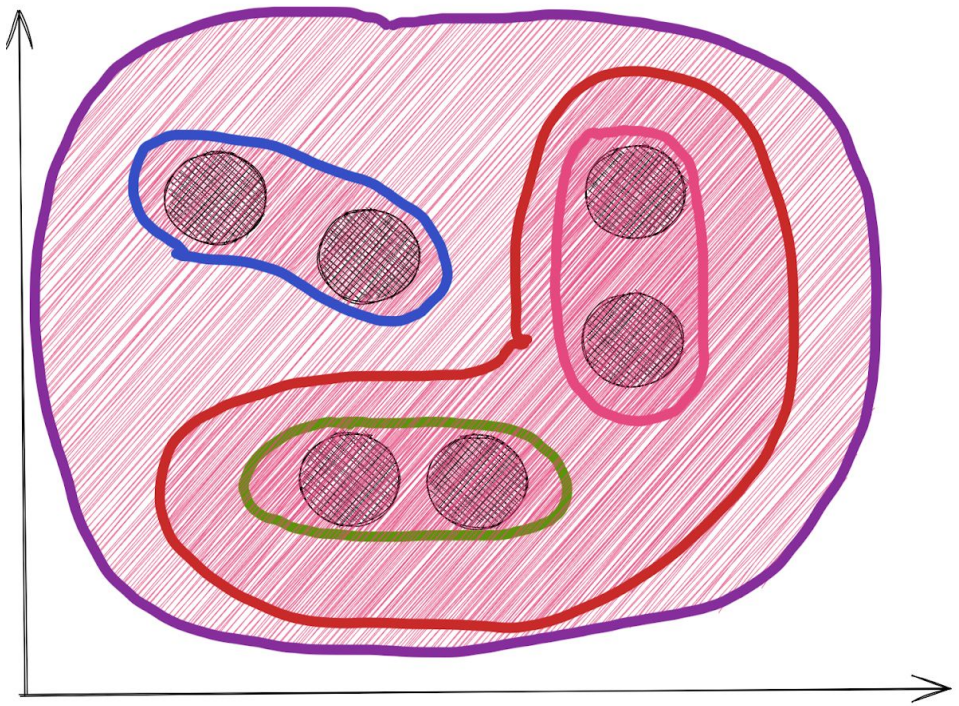
Coeficiente de silueta

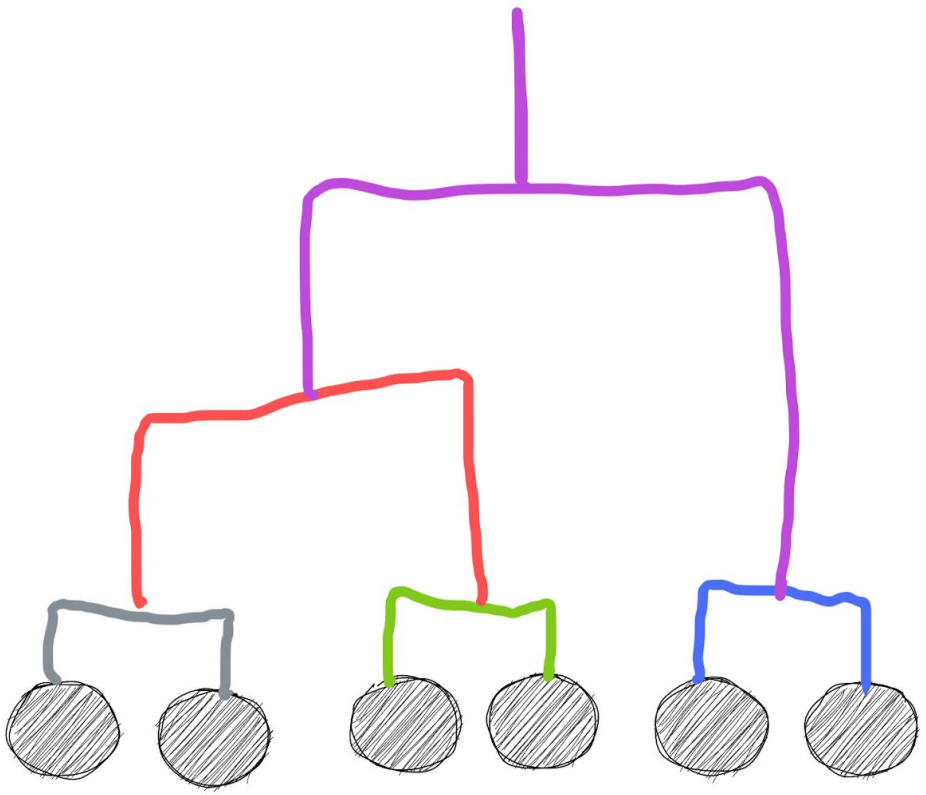


K means

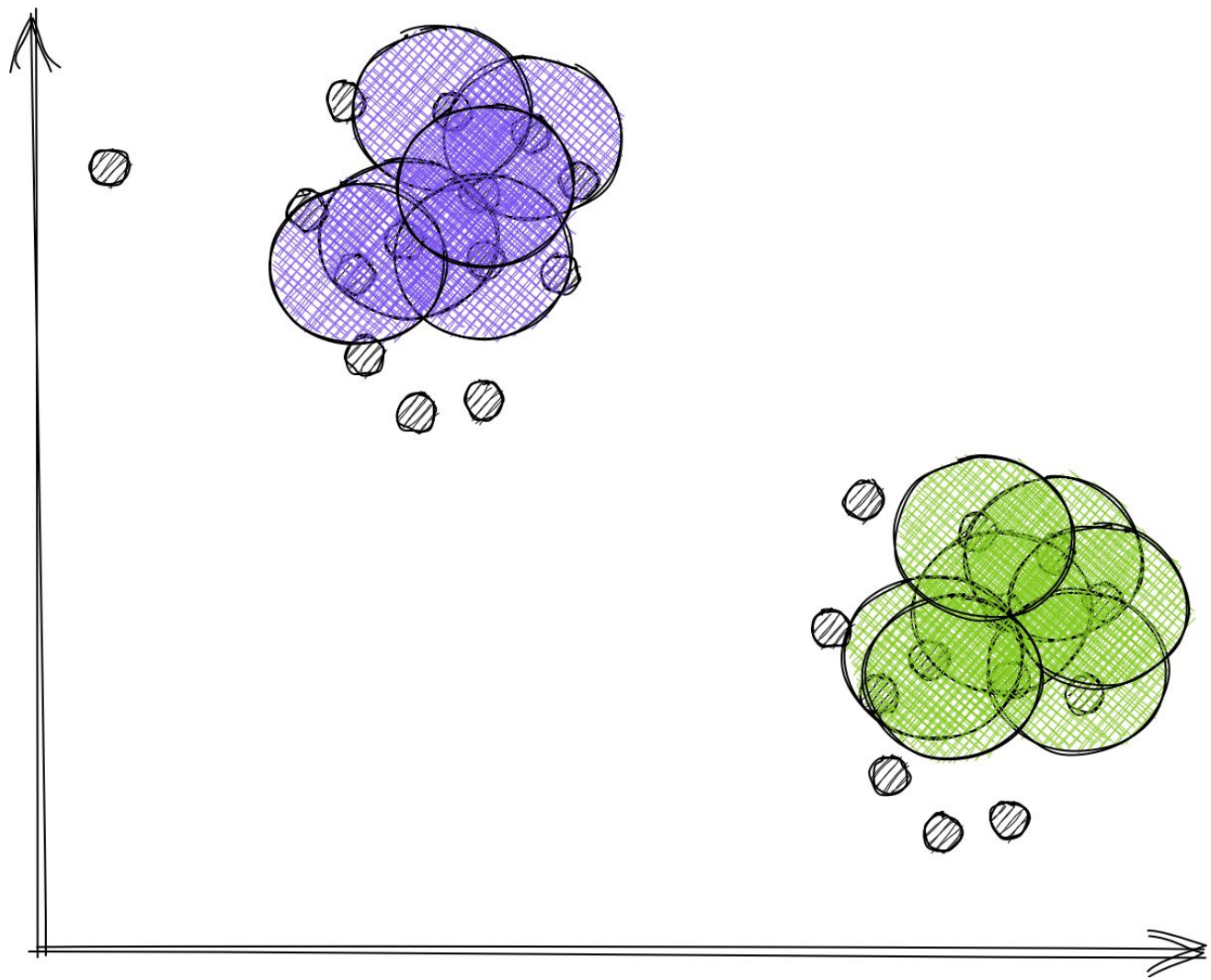


Hierarchical clustering





DBSCAN



Tu proyecto

Customer Personality Analysis

Analysis of company's ideal customers

Data | Code (247) | Discussion (24) | Metadata

About Dataset

Context

Problem Statement

Customer Personality Analysis is a detailed analysis of a company’s ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers.

Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company’s database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

Content

Usability ⓘ
9.71

License
CC0: Public Domain

Expected update frequency
Never

Tu proyecto

Activity Overview

ACTIVITY STATS

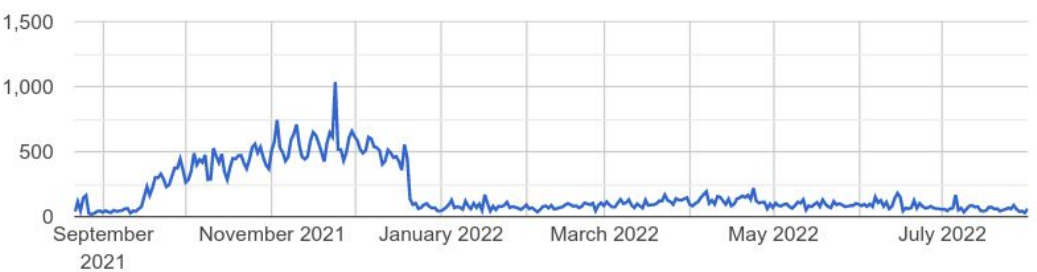
VIEWS
427285

DOWNLOADS
66313

DOWNLOAD PER VIEW RATIO
0.16

TOTAL UNIQUE CONTRIBUTORS
258

Downloads



NOTEBOOKS STATS

NOTEBOOKS
247

NOTEBOOK COMMENTS
1122

UPVOTE PER NOTEBOOK RATIO
12.69

NOTEBOOK UPVOTES
3134

TOP CONTRIBUTORS

Karnika Kapoor

Sonali Singh

Akash Patel

DISCUSSION STATS

TOPICS
23

TOTAL COMMENTS
64

UPVOTE PER POST RATIO
2.78

DISCUSSION UPVOTES
178

Carlos Andrés Alarcón

@Alarcon7a