Técnicas de MD

Data Mining - CC 3074 / 20

UVG

2021 - Ciclo I

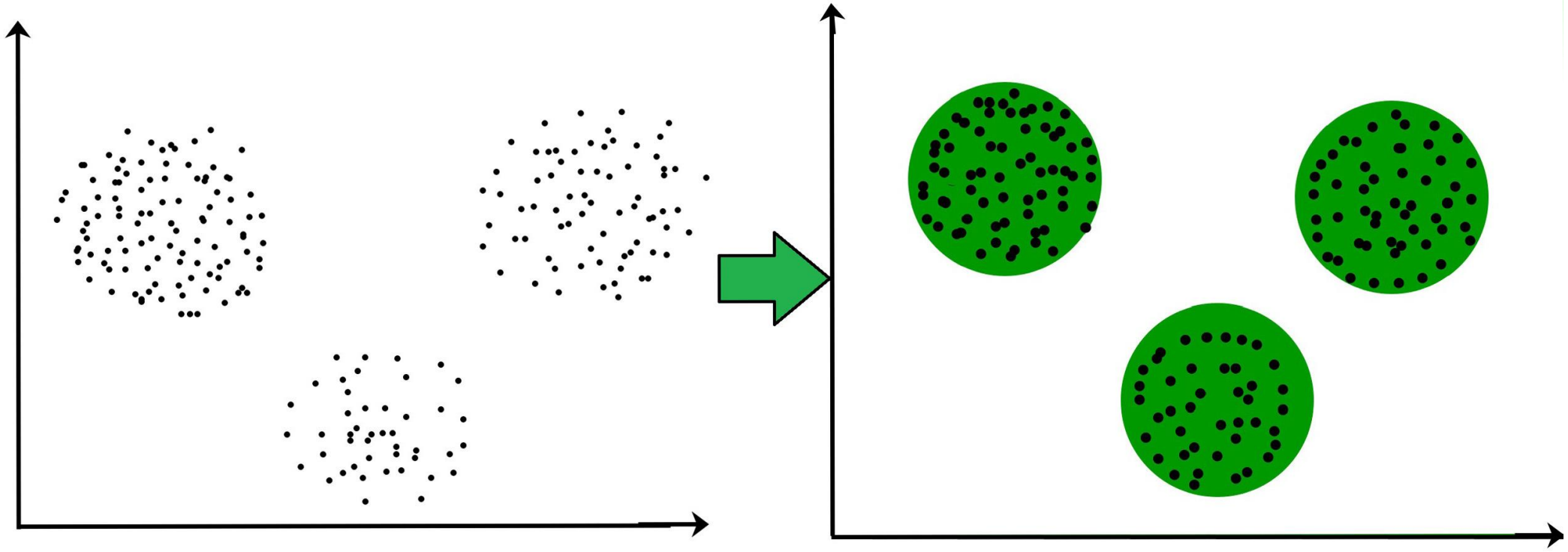
Clustering

Técnicas de MD

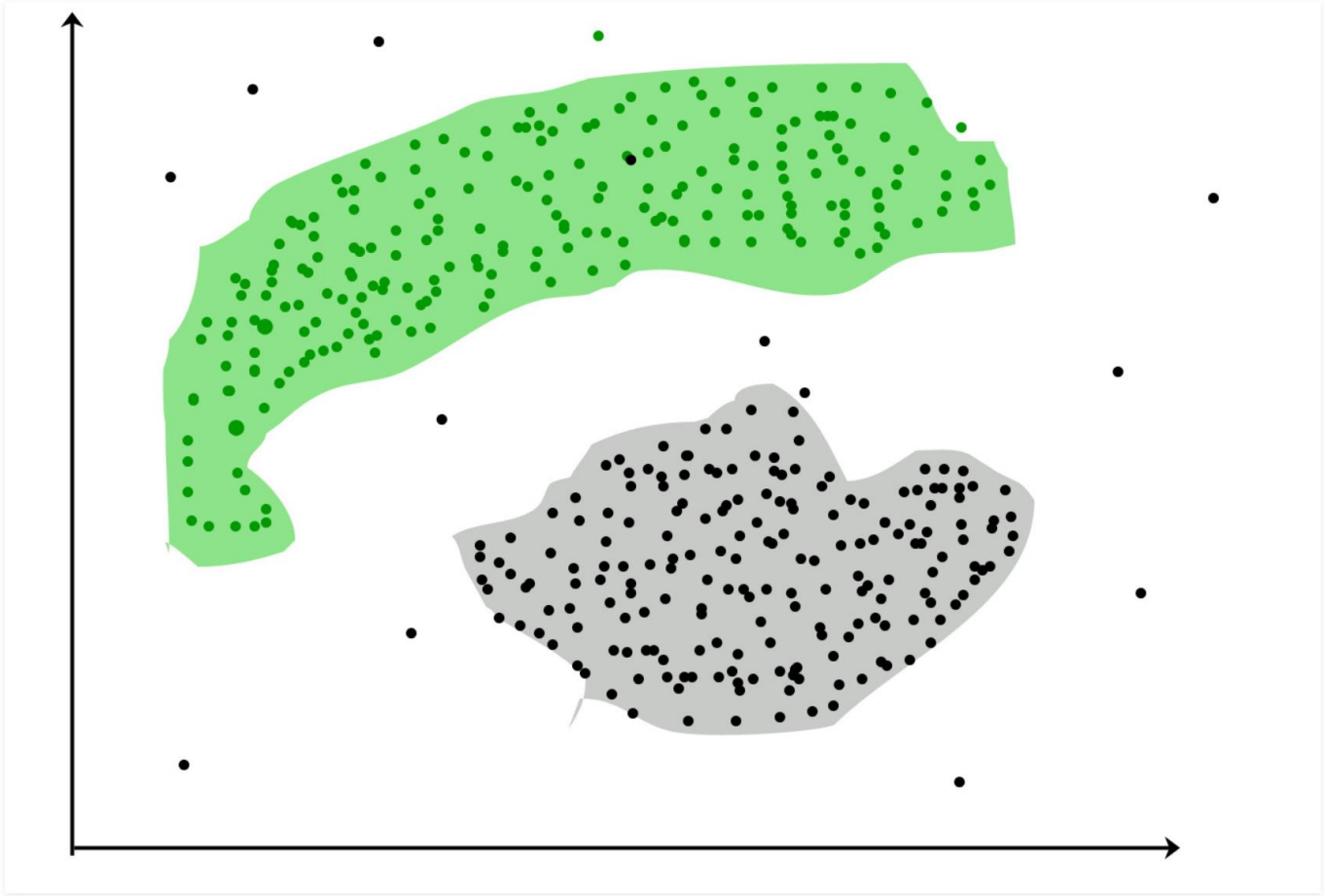
Definición de Clustering

- Método o Técnica de Aprendizaje No-Supervisado
- **Cluster analysis** or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). *–Wikipedia*
- The objective of ~~K-means~~ [Clustering] is simple: group similar data points together and discover underlying patterns. *–Andrey Bulezyuk*
- A cluster refers to a collection of data points aggregated together because of certain similarities. *–Dr. Michael J Garbade*

Ejemplo de Clustering



Ejemplo de Clustering



DBSCAN: Density-based Spatial Clustering of Applications with Noise









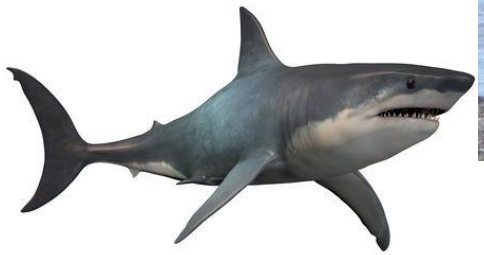




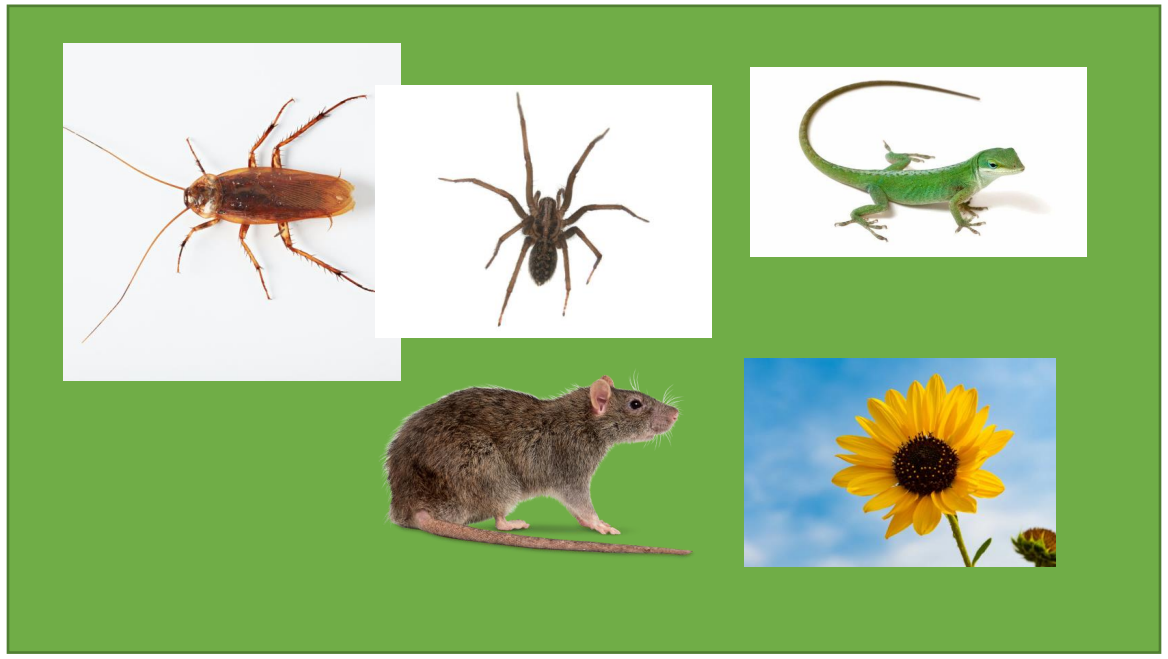






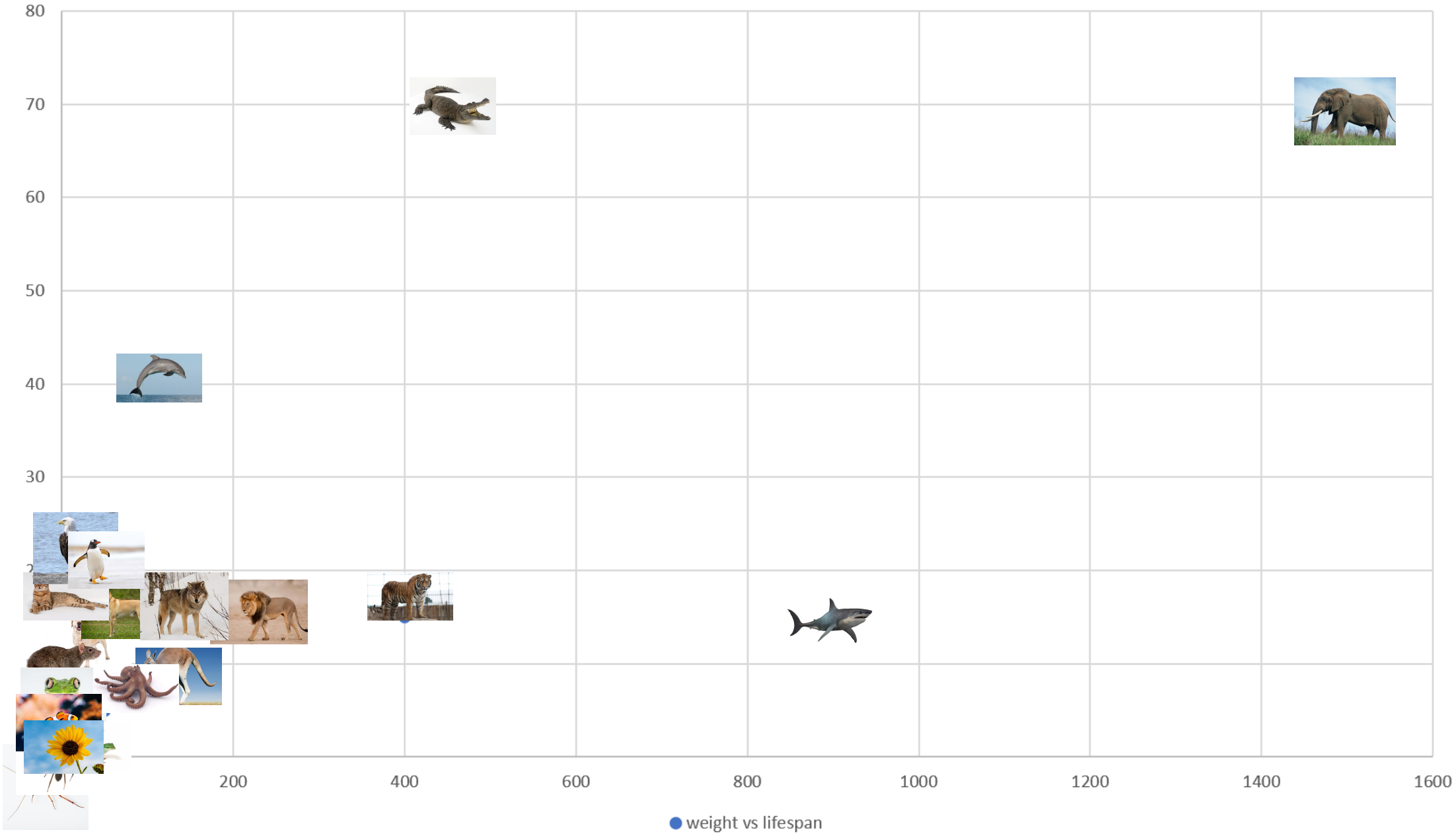



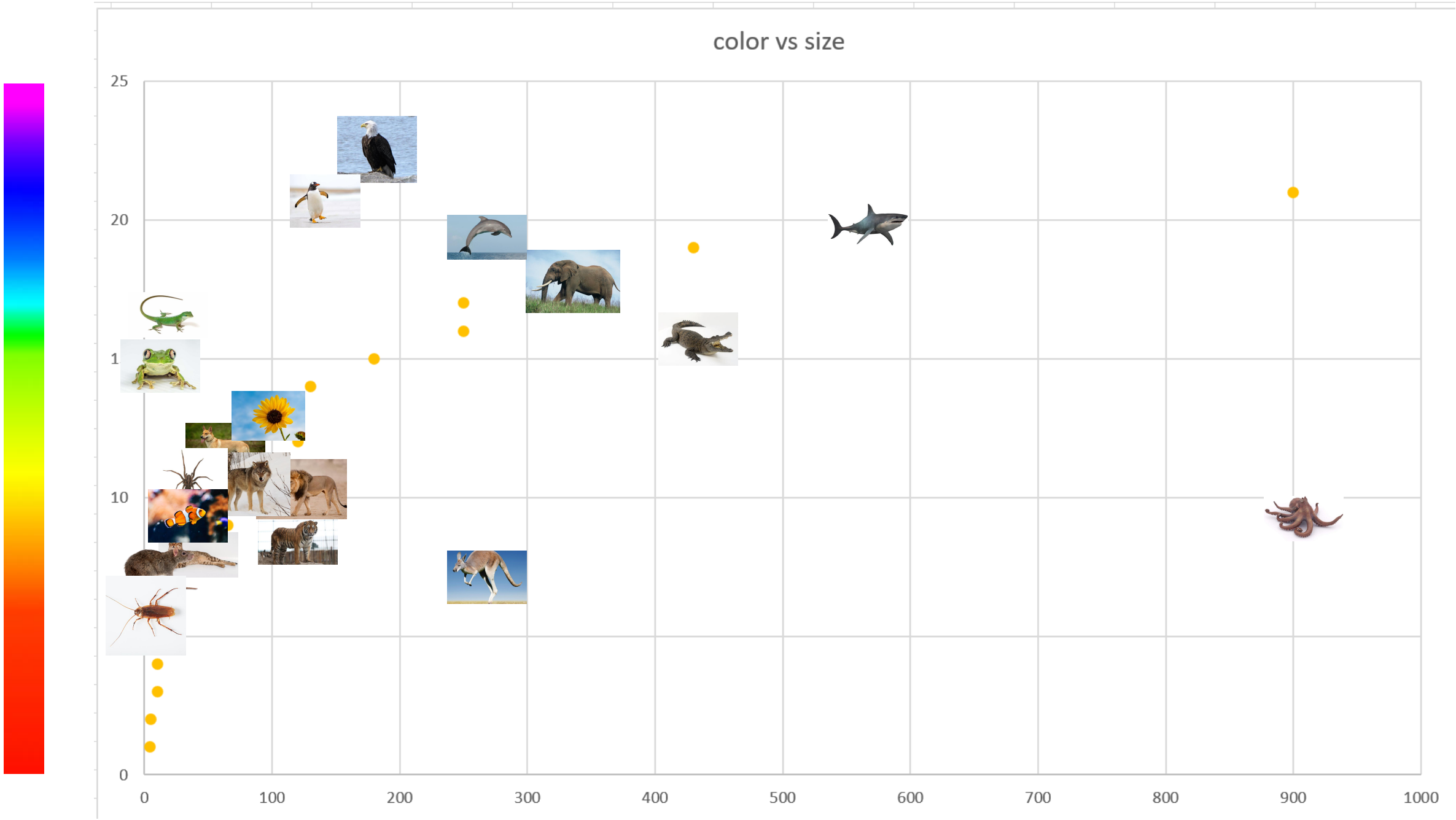

Hacer 3 agrupaciones:



Consideraciones

- ¿Qué características tomamos en cuenta para hacer el agrupamiento?
    - Reino, Familia, Clase
    - Peso
    - Tamaño
    - Tiempo de vida
    - Color
    - Número de dientes
    - Número de patas
- ¿Cuántos grupos queremos de resultado?
- ¿Cuántos atributos/dimensiones vamos a considerar?
- ¿De qué manera vamos a agrupar? (Algoritmos)
- ¿Qué hacemos con los *outliers* ?


weight vs lifespan
80
70
60
50
40
30
200
400
600
800
1000
1200
1400
1600
weight vs lifespan


color vs size
25
20
1
10
0
0
100
200
300
400
500
600
700
800
900
1000

K-means Clustering

- Es probablemente el algoritmo más simple y popular usado para Clustering
- El algoritmo K-means identifica *k* centroides y asigna cada punto de datos al clúster más cercano, intentando mantener los centroides lo más pequeño posible
- El *"means"* se refiere a las medias o promedios en la data; que es, el centroide
- Utilizado para clusterizar data numérica

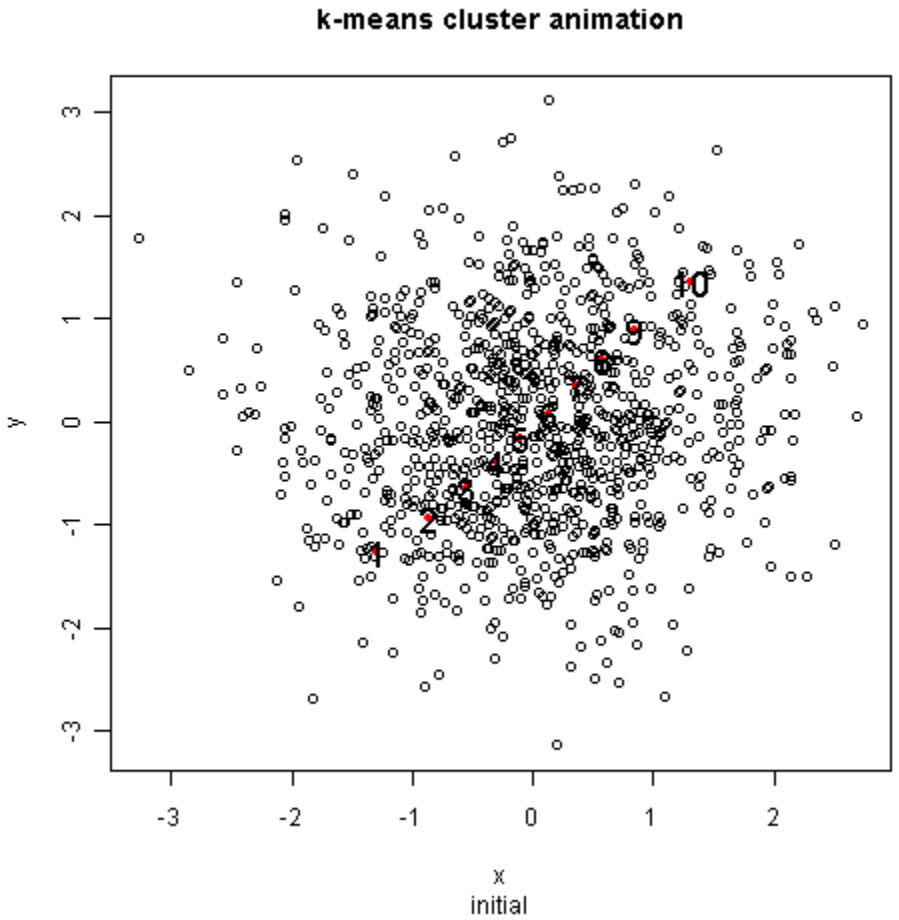
K-means: consideraciones

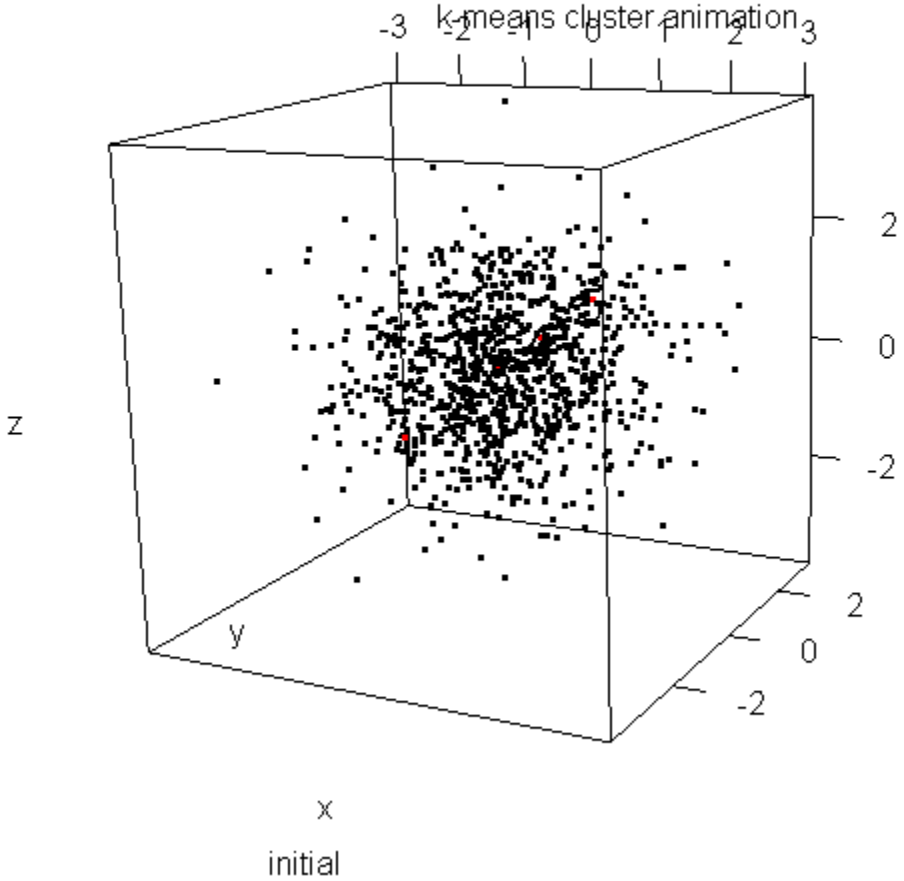
- Todos los atributos deben ser data numérica
- Existe una función de distancia entre dos puntos
  - Distancia no negativa
  - La distancia de una observación consigo misma es cero
  - La distancia entre la observación A y la observación B es la misma que la distancia entre la observación B y A
  - La distancia entre dos observaciones no puede ser mayor que la suma de la distancia entre esas dos observaciones y una tercera observación
- Se debe indicar *k* a priori

K-means: el proceso

1. Se seleccionan k centroides de manera aleatoria
2. Cada punto u observación es asignado al centroide más cercano
3. Se recalcula la posición de los centroides
4. Se repiten los pasos 2 y 3 hasta que no hayan cambios en las asignaciones

K-means: ejemplos





K-means: output del modelo

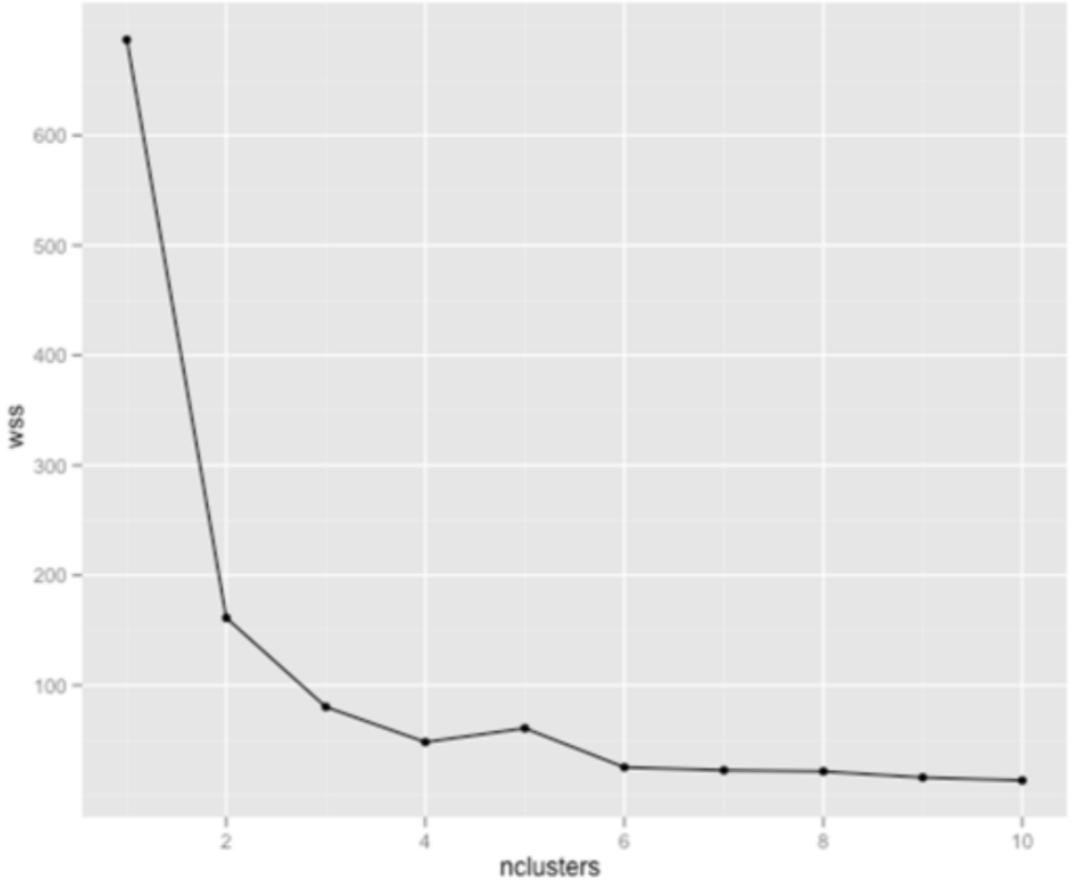
- Coordenadas de los centroides finales
- Asignación de cada observación a un clúster

K-means: selección de K

- **Método Heurístico:**
  Encontrar el "codo"

$$wss = \sum_{i=1}^{k} \sum_{j=1}^{n_i} | \, x_{ij} \, - \, c_i \, |^2$$

  - k : número de clústers
  - n_i : puntos en el i-ésimo clúster
  - c_i : centroide del i-ésimo cúster
  - x_{ij} : j-ésimo punto del i-ésimo clúster



K-means: selección de K

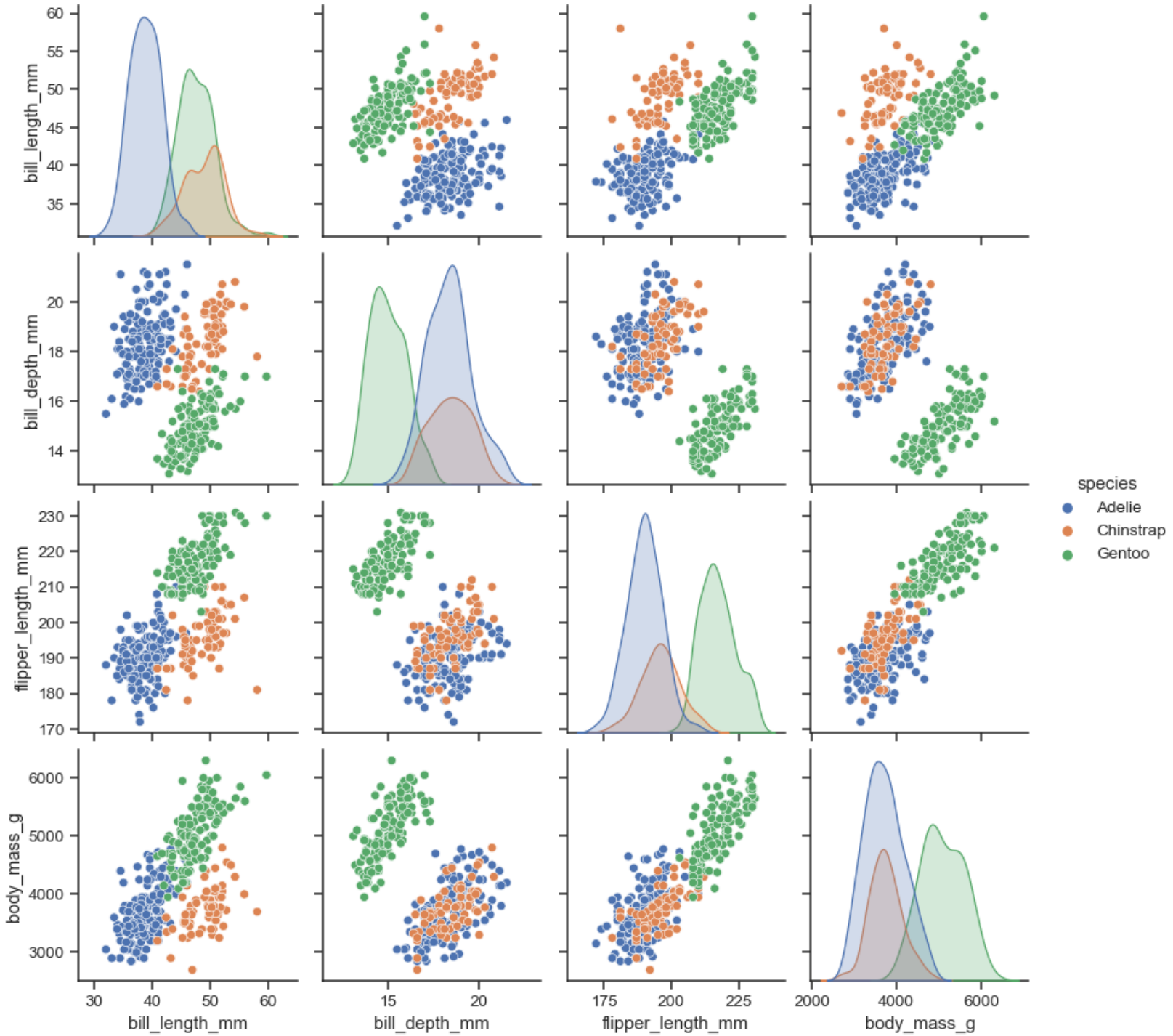
- Seleccionar el valor adecuado de K es posible saberlo de antemano si se posee conocimiento previo del tema

- Normalmente el valor de K adecuado es desconocido, por lo que se recomienda probar varios valores de K y comparar resultados

- WSS indica la homogeneidad de los clusters:
  - Entre más separados estén los clusters, y más pegados estén las observaciónes dentro de cada cluster, más homogéneos son

K-means: evaluación del modelo

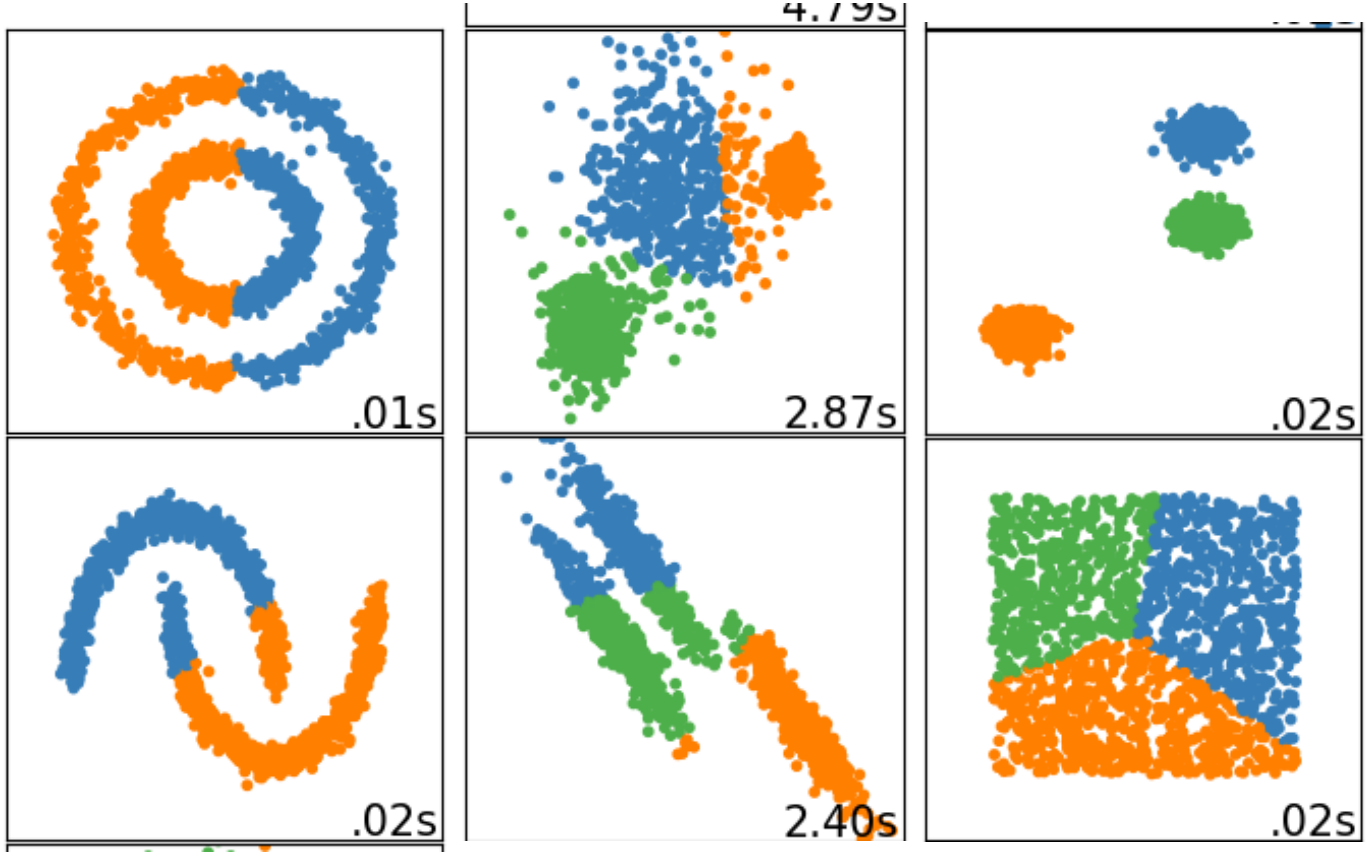
- ¿Los clústers se ven separados en algunas gráficas cuando se visualizan las gráficas pair-wise de los clusters?
- ¿Se generaron clústers con muy pocas observaciones?
  - Intentar reduciendo K
- ¿Hay divisiones que esperaría ver en las variables pero no se ven?
  - Intentar aumentando K
- Los centroides están muy cercanos entre sí?
  - Intentar reduciendo K


bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
species
Adelie
Chinstrap
Gentoo

K-means: consideraciones

- No maneja variables categóricas (todos los atributos deben ser numéricos)
- Todas las dimensiones deben ser normalizadas a una misma escala
- Sensible a la inicialización (random)
- Tiende a producir clústers redondos o esféricos (o hiperesféricos)

K-means: no es universal



Métodos de Clústering

- There are no criteria for a good clustering
- It depends on the user, what is the criteria they may use which satisfy their need

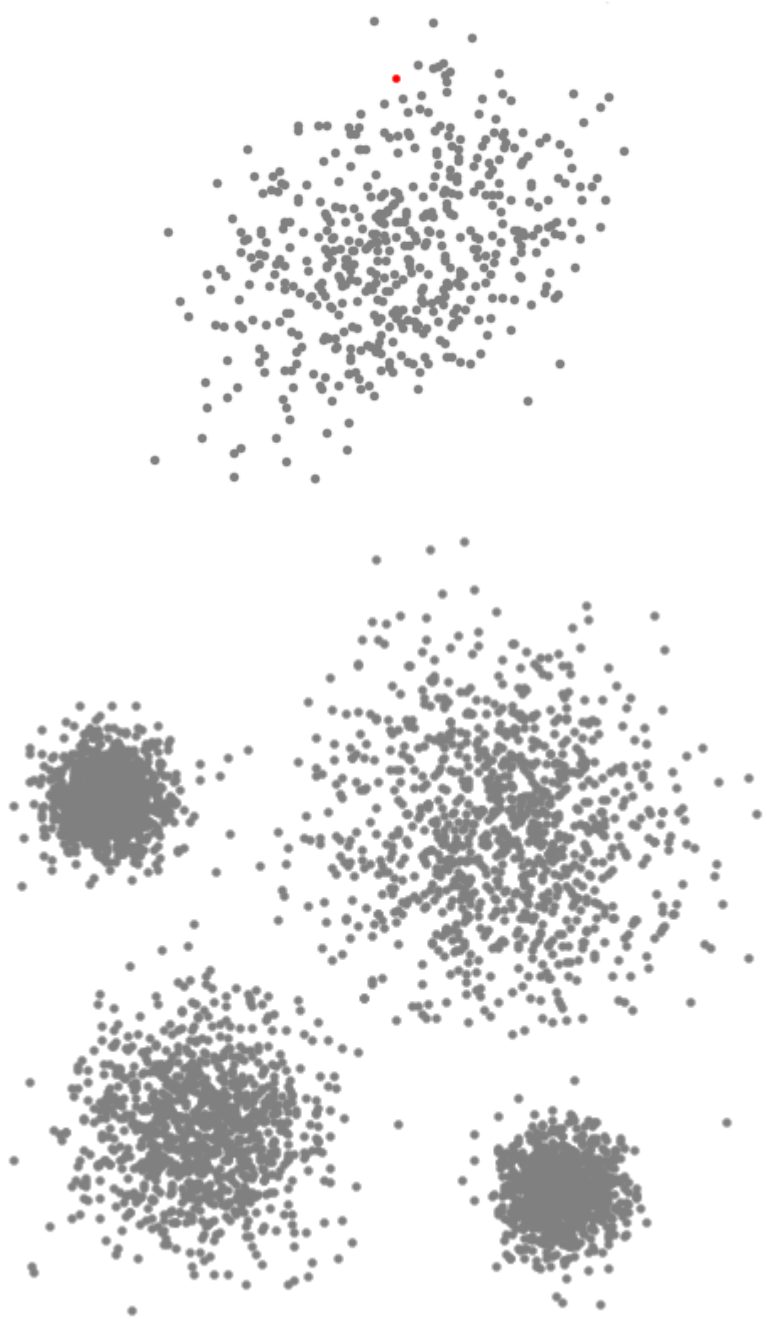
    - We could be interested in finding representatives for homogeneous groups (data reduction)
    - We could be interested in finding “natural clusters” and describe their unknown properties (“natural” data types)
    - We could be interested in finding useful and suitable groupings (“useful” data classes)

Métodos de Clústering

- **Density-Based Methods:** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space.
- **Hierarchical Based Methods**: The clusters formed in this method forms a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one.
- **Partitioning Methods**: These methods partition the objects into k clusters and each partition forms one cluster.
- **Grid-based Methods**: In this method the data space is formulated into a finite number of cells that form a grid-like structure.

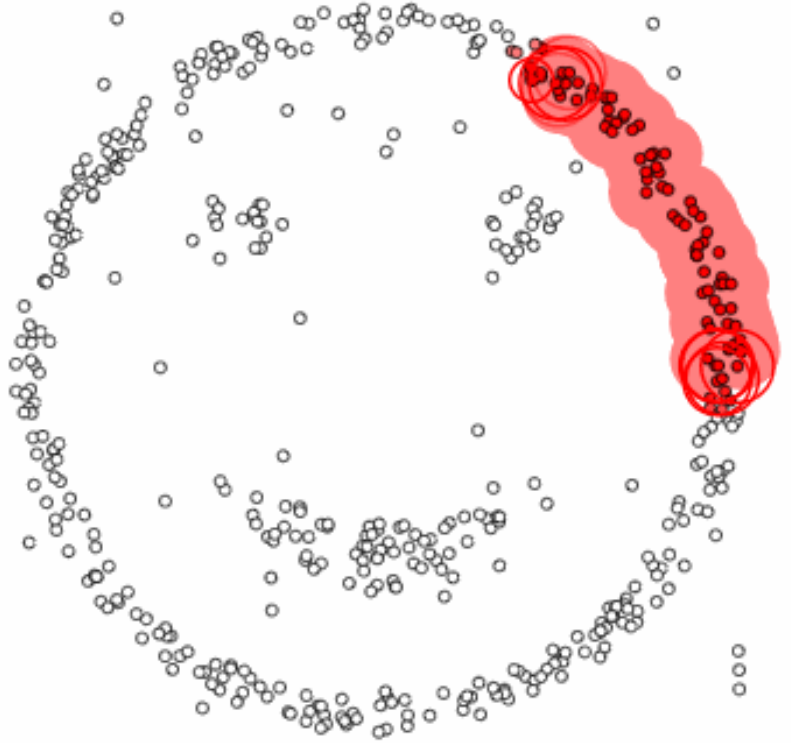
Mean-Shift Clustering

1. Se selecciona una ventana circular de radio r centrada en un punto aleatorio C
2. En cada iteración la ventana se moverá a una región con mayor densidad, moviendo el centro hacia el centroide de los puntos cubiertos por la ventana
3. Se continúa moviendo la ventana hasta una iteración en la que ya no encuentre una posición con mayor densidad
4. Los pasos 1-3 se ejecutan para la cantidad de *sliding windows* deseadas.



Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

1. Se inicia con un punto aleatorio del dataset que no haya sido visitado y se calcula su vecindario a un radio ε
2. Si hay un mínimo de puntos en el vecindario se comienza el proceso de clustering. El punto inicial se asigna al clúster.
3. Se asignan al cluster los puntos que se encuentren en el vecindario de los puntos previamente asignados al cluster.
4. Cuando no hay más puntos en ningún vecindario del clúster se busca otro punto que no haya sido visitado y se repite desde el punto 2

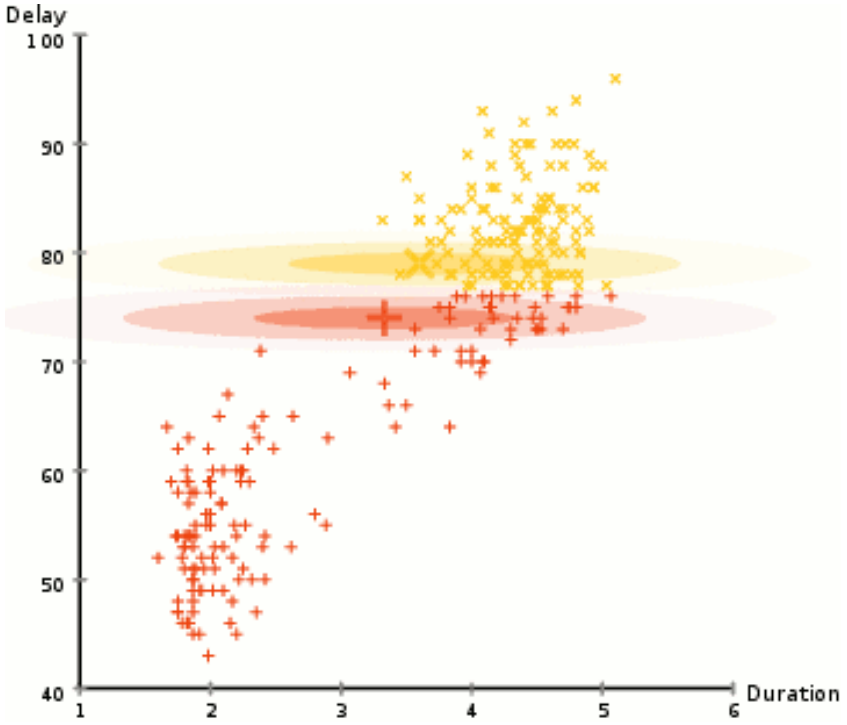


Restart

Pause

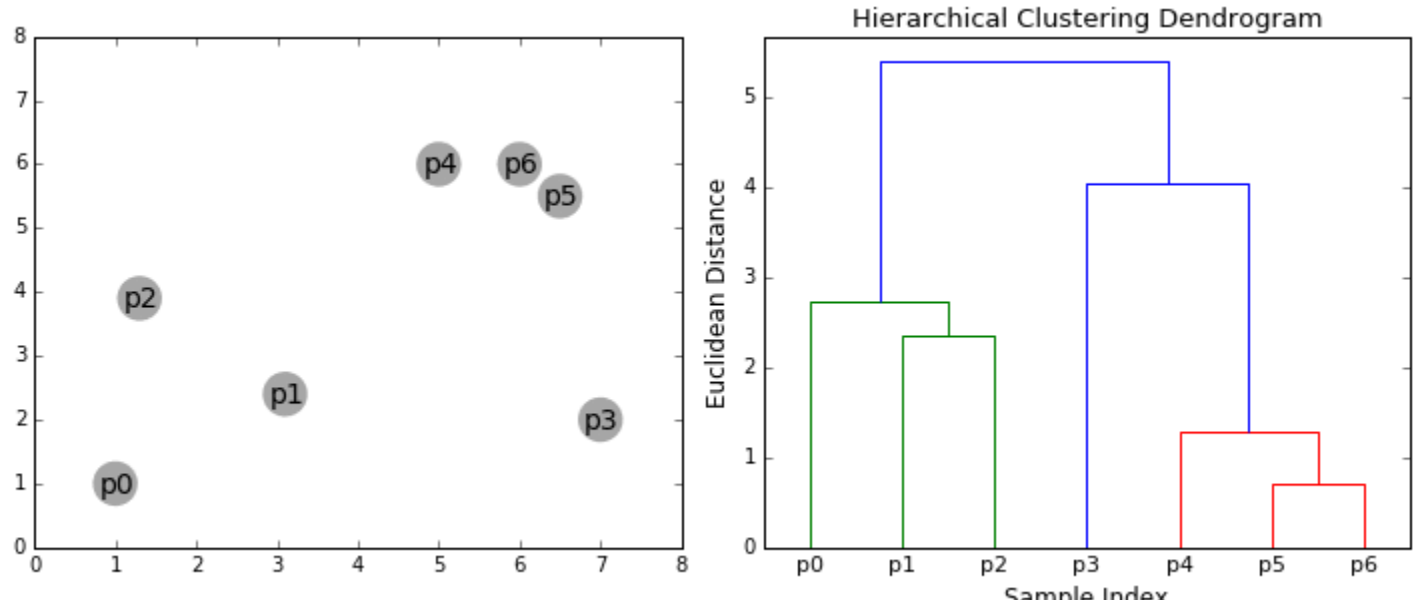
Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

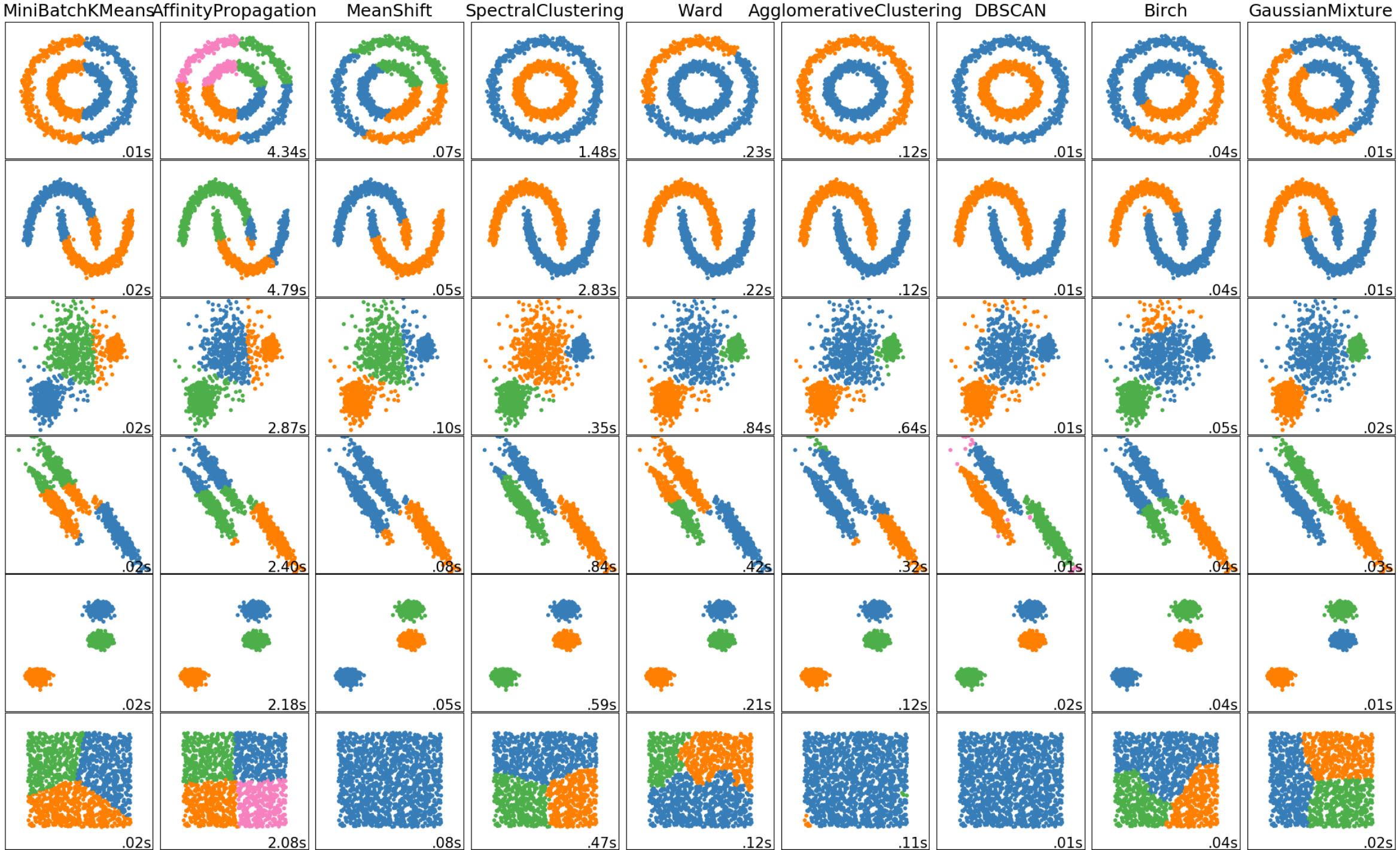
1. Se selecciona el número de clústers y se inicializan los parámetros de distribución gaussiana para cada uno
2. Se calcula la probabilidad de cada punto de pertenecer a cada clúster
3. Basado en esta asignación se calcula una nueva distribución gaussiana para cada clúster
4. Se repiten los pasos 2 y 3 hasta llegar a una convergencia



Agglomerative Hierarchical Clustering

1. Se considera a cada punto del dataset como un clúster
2. En cada iteración se seleccionan los dos clústers con menor distancia entre sí y se combinan en uno. (Depende de la métrica de distancia entre clústers que se considere)
3. Se repite el paso 2 hasta tener la cantidad de clústers deseados o hasta que sea sólo uno que abarque toda la data.




MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s
4.34s
.07s
1.48s
.23s
.12s
.01s
.04s
.01s
.02s
4.79s
.05s
2.83s
.22s
.12s
.01s
.04s
.01s
.02s
2.87s
.10s
.35s
.84s
.64s
.01s
.05s
.02s
.02s
2.40s
.08s
.84s
.42s
.32s
.01s
.04s
.03s
.02s
2.18s
.05s
.59s
.21s
.12s
.02s
.04s
.01s
.02s
2.08s
.08s
.47s
.12s
.11s
.01s
.04s
.02s