

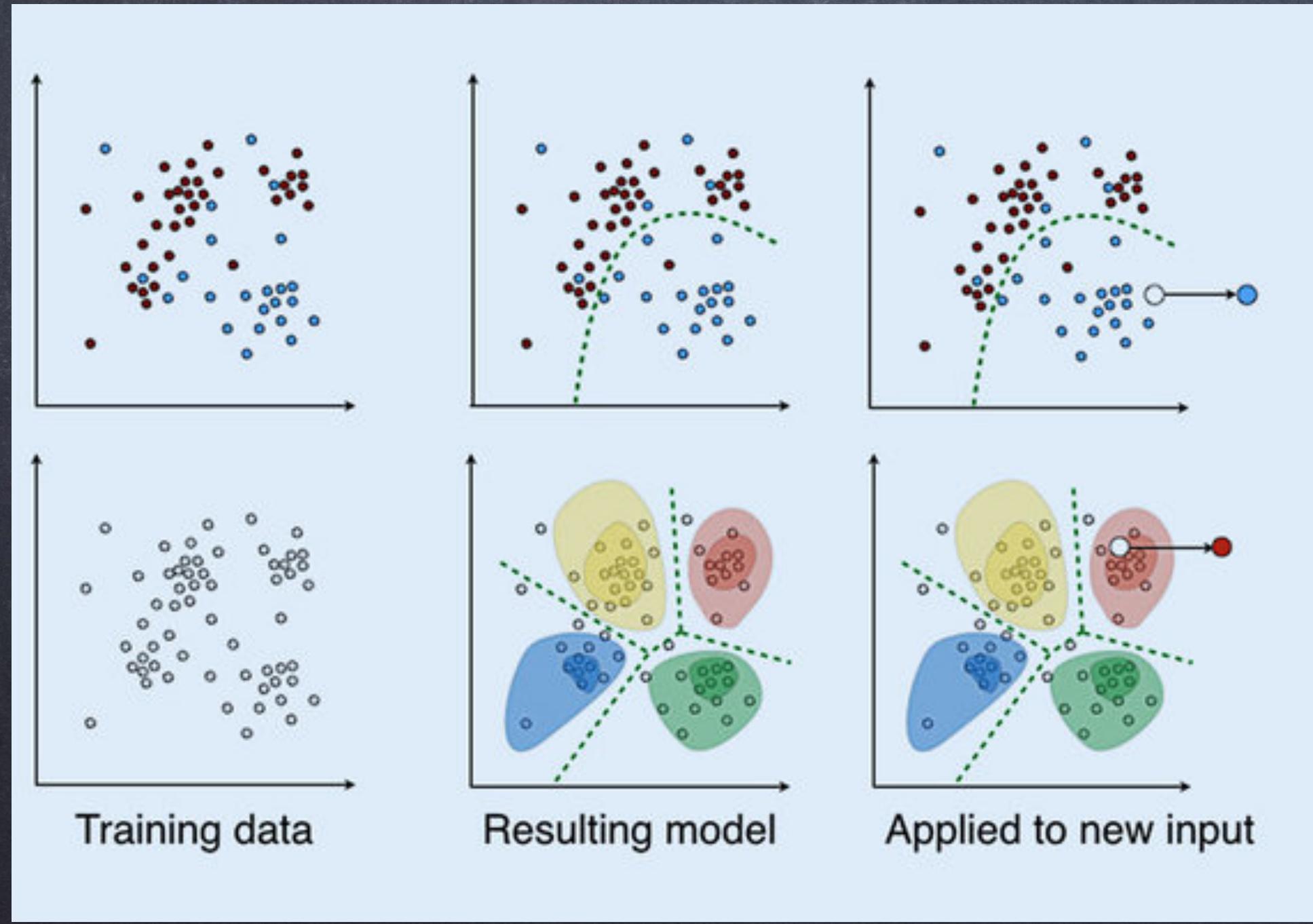
# Unsupervised Machine Learning

Física Computacional 2  
Ph.D. Santiago Echeverri A

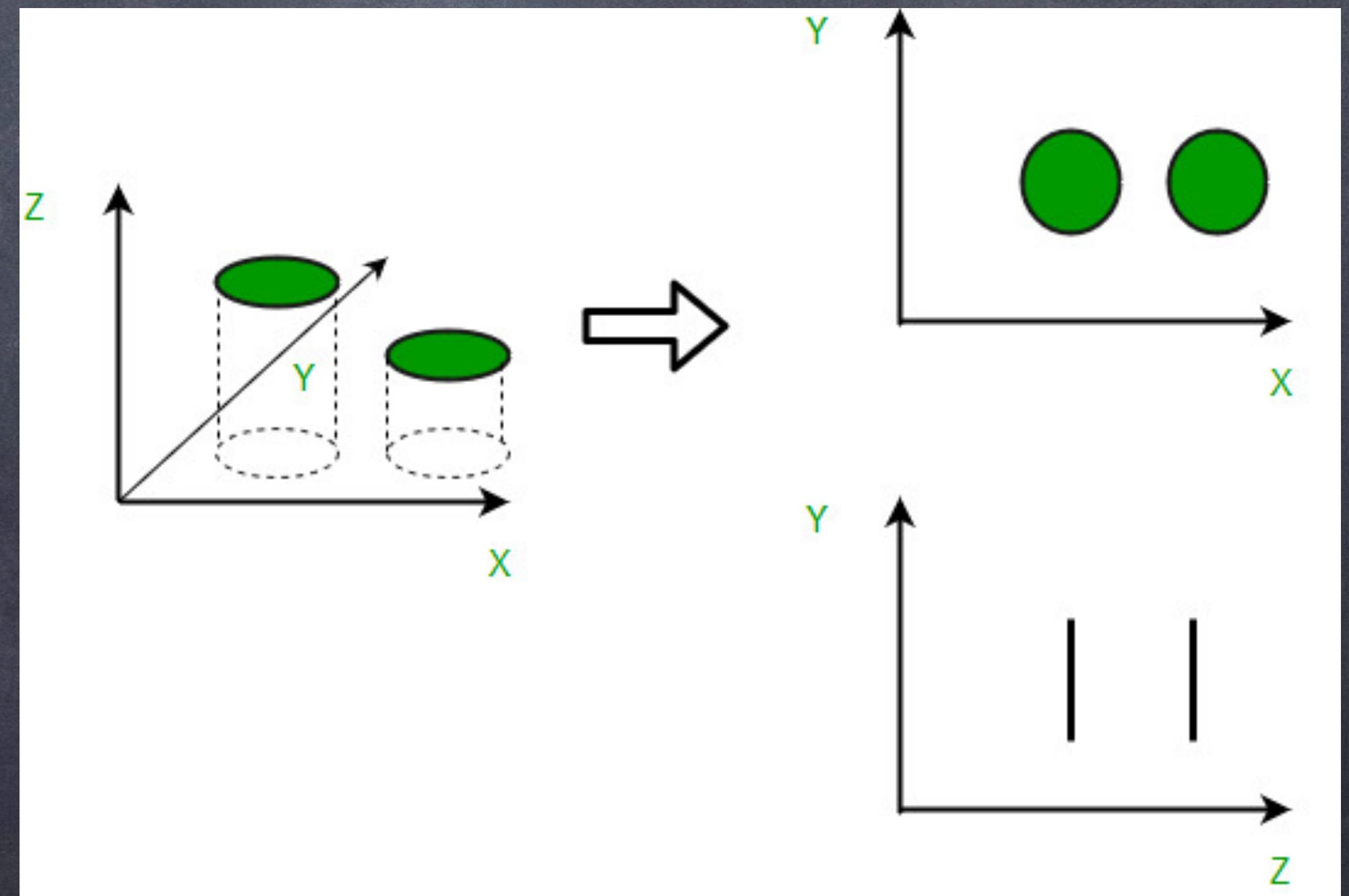
# ¿Qué significa no supervisado?

Puede ser una forma de EDA!

Clustering



Reducción de dimensionalidad



# Clustering

- Identificar patrones de agrupamiento
- No hay variable independiente, por lo que no hay una función de pérdida (error) que nos permita medir la exactitud del modelo
- Para cada grupo hay que predecir el grupo al que pertenece pero ningún dato de la tabla tiene esa información
- Es útil para generar nuevas variables independientes que ayuden a mejorar en aprendizaje supervisado

# Distancia, disimilaridad, norma y $g^{ij}$

- Sea  $E$  un espacio vectorial dado. Toda aplicación  $E \times E \rightarrow \mathbb{R}^+$  se llama una **distancia** si verifica los tres axiomas siguientes:
  - $x = y \iff d(x, y) = 0$
  - $d(x, y) = d(y, x) \quad \forall x, y \in E$
  - $d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z \in E$
- Las aplicaciones que no cumplen el tercer axioma pero si los otros dos, se les conoce como medidas de **disimilaridad**
- Se llama **espacio métrico** a todo conjunto  $(E, d)$
- Sea  $E$  un espacio vectorial sobre  $K$ , Toda aplicación  $\|\cdot\|$  de  $E$  en  $\mathbb{R}^+$  se llama una **norma** si cumple las tres propiedades siguientes:
  - $\|x\| = 0 \iff x = 0$
  - $\|\lambda x\| = |\lambda| \|x\| \quad \forall x \in E, \lambda \in K$
  - $\|x + y\| \leq \|x\| + \|y\|$
- Todo espacio vectorial provisto de norma se llama **espacio vectorial normado**
- Si las variables no están escaladas, una variable en un rango amplio de valores va a generar distancias mucho mayores y por tanto dominará el resultado del agrupamiento.

# Distancias más comunes

- $d(x_i, x_j) = \left[ \sum_{k=1}^p |x_{ik} - x_{jk}|^l \right]^{1/l}$  para  $l = 1$  Manhattan (city block)  $l = 2$  Euclideana  $l > 2$  distancia  $l$  de Minkowski
- Distancia coseno  $d(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$
- Distancia de Jaccard: Sean  $A$  y  $B$  conjuntos, su distancia de Jaccard es  $1 - \frac{|A \cap B|}{|A \cup B|}$
- Mahalanobis distance TAREA OPCION 1 (Teoría y aplicación en ciencia de datos)
- "A general coefficient of similarity and some of its properties" TAREA OPCION 2

# K-Means



Contras:

- No es bueno para datos de densidad variable
- Sensible a outliers
- El sistema puede presentar múltiples soluciones (Depende condición inicial)



Pros: Escala bien

- Procedimiento
  - Cada punto pertenece al centroide más cercano.
  - Ajusta el centroide al nuevo promedio de los clusters.
  - Repetir hasta que no se presente una reasignación
- El primer punto se elige de forma aleatoria y los siguientes K-1 (**por defecto**) Los más lejanos usando el peso
$$\frac{d(x_i, C_k)}{\left[ \sum_{i=1}^N d(x_i, C_k) \right]^2}$$

# K-Means

- Si no es claro como seleccionar el cluster se puede usar el K que minimice una de las siguientes métricas
- Inercia: Sensible al número de puntos del cluster. Ayuda a analizar la entropía que se genera en los clusters  $\sum_{i=1}^N (x_i - C_k)^2$
- Distorsión: Soluciona la sensibilidad al número de puntos por cluster  $\frac{1}{N} \sum_{i=1}^N (x_i - C_k)^2$
- Si es importante la similaridad de los puntos se usa distorsión. Si es importante que los clusters tengan la misma cantidad de datos se usa Inercia

# Ajuste de hiperparámetros

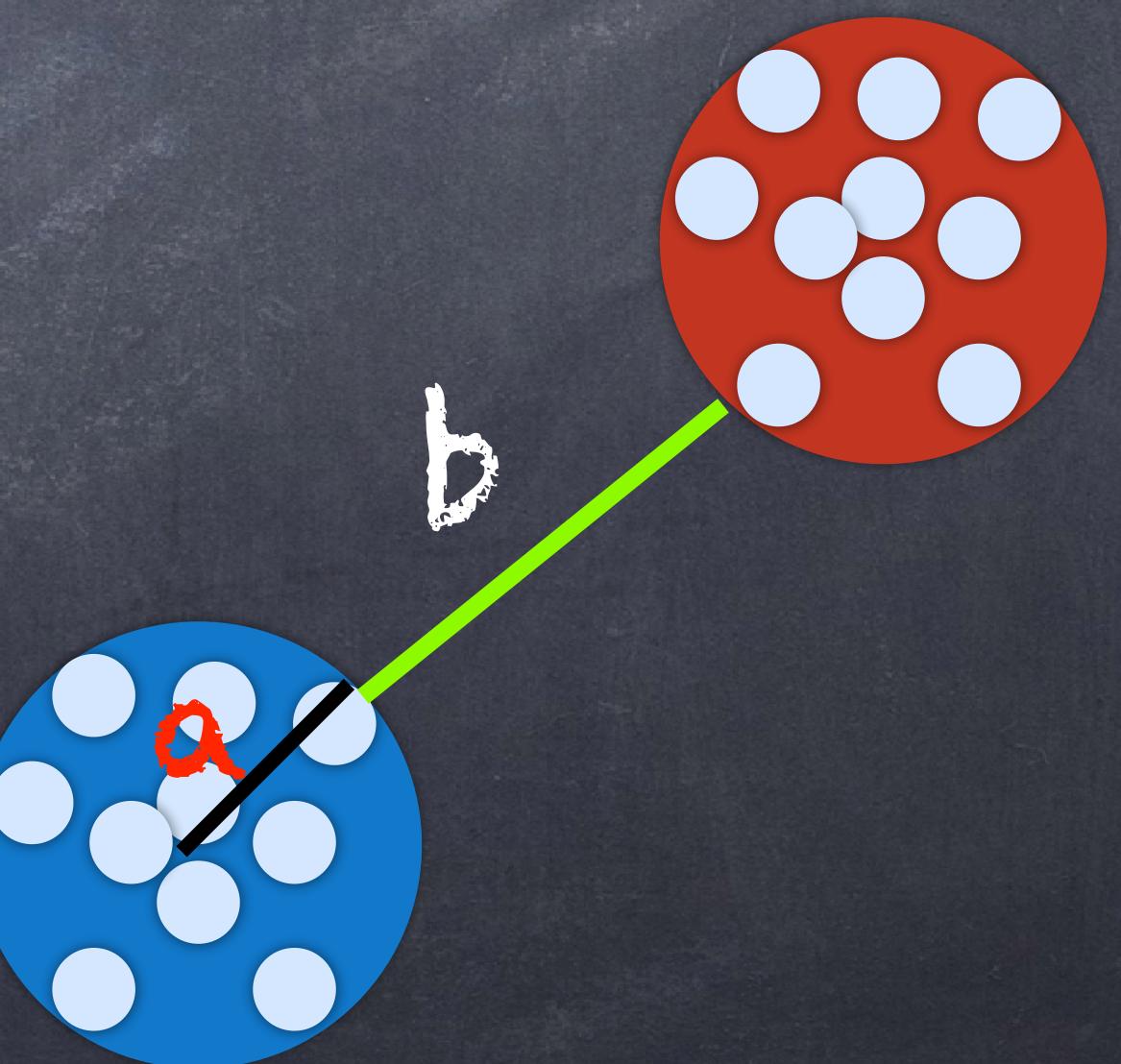
- Ajustar K por el método del codo (si no se sabe cuál es el K necesario)
- Usar el BIC (Bayesian information error) o el AIC (Akaike Information Criterion)

## TAREA

- Coeficiente de silueta: Si el método del hombro no es claro se usa la diferencia entre la similaridad entre los puntos en un cluster y los otros puntos respecto a los clusters circundantes (Valor entre -1 y 1)

$$\frac{b - a}{\max(b - a)}$$

- $a$ : distancia media en el cluster
- $b$ : promedio de la distancia mínima a otro cluster



# Mezcla de Gausianas (GMM)

- No solo da las etiquetas de los clusters sino la probabilidad de que pertenezca a uno de ellos
- Ajusta distribuciones Gaussianas a los datos usando MLE (maximum Likelihood)
- Se debe determinar el número de distribuciones Gaussianas K
- Hard Clustering: Cada punto pertenece a un solo Cluster
- Soft Clustering: Cada punto puede pertenecer a más de un cluster
- El método de GMM es soft clustering
- Combinación de K Gaussianas, cada una con un centro  $\mu_i$ , matriz de covarianza  $\sigma_i$  y a cada Guadiana se le asigna un peso  $\pi_i$
- Probability Density Function (PDF)  $p(x) = \sum_{i=1}^K \pi_i \mathcal{N}(x | \mu_i, \sigma_i)$  con  $\sum_{i=1}^N \pi_i = 1$

# Tipos de matrices de covarianza

- Son Gausianas N dimensionales!
- Los tipos de matrices de covarianza
  - Full: Cada componente tiene su propia matriz
  - Tied: Todos comparten la misma matriz
  - Diag: Cada componente tiene su propia matriz diagonal
  - Spherical: Cada componente tiene un único valor de varianza



8 componentes

2 componentes



# Clustering aglomerativo jerárquico

- Unir clusters hasta tener convergencia. Se identifican los puntos con distancia mínima y se unen en un cluster.
- El más cercano puede ser un punto o un cluster
- Hay que definir distancia entre clusters y distancia punto-cluster → **Linkage criterion**
- Se van identificando y uniendo puntos hasta el número de clusters deseado, o hasta que todos los clusters superen cierta distancia promedio de los clusters (radio)

# Criterios de ligado (Linkage)

- Single Linkage: Distancia de punto a punto mínima. Garantiza separación de clusters. No puede separar limpiamente si hay ruido (Outliers)
- Complete Linkage: Máxima distancia entre dos puntos. Separa mejor si hay puntos salados pero tiende a romper clusters grandes existentes
- Average: Distancia entre los centroides. Ventajas y desventajas como punto medio de los métodos anteriores
- Ward: Calcula la inercia y une los clusters que minimizan la inercia

# DBSCAN

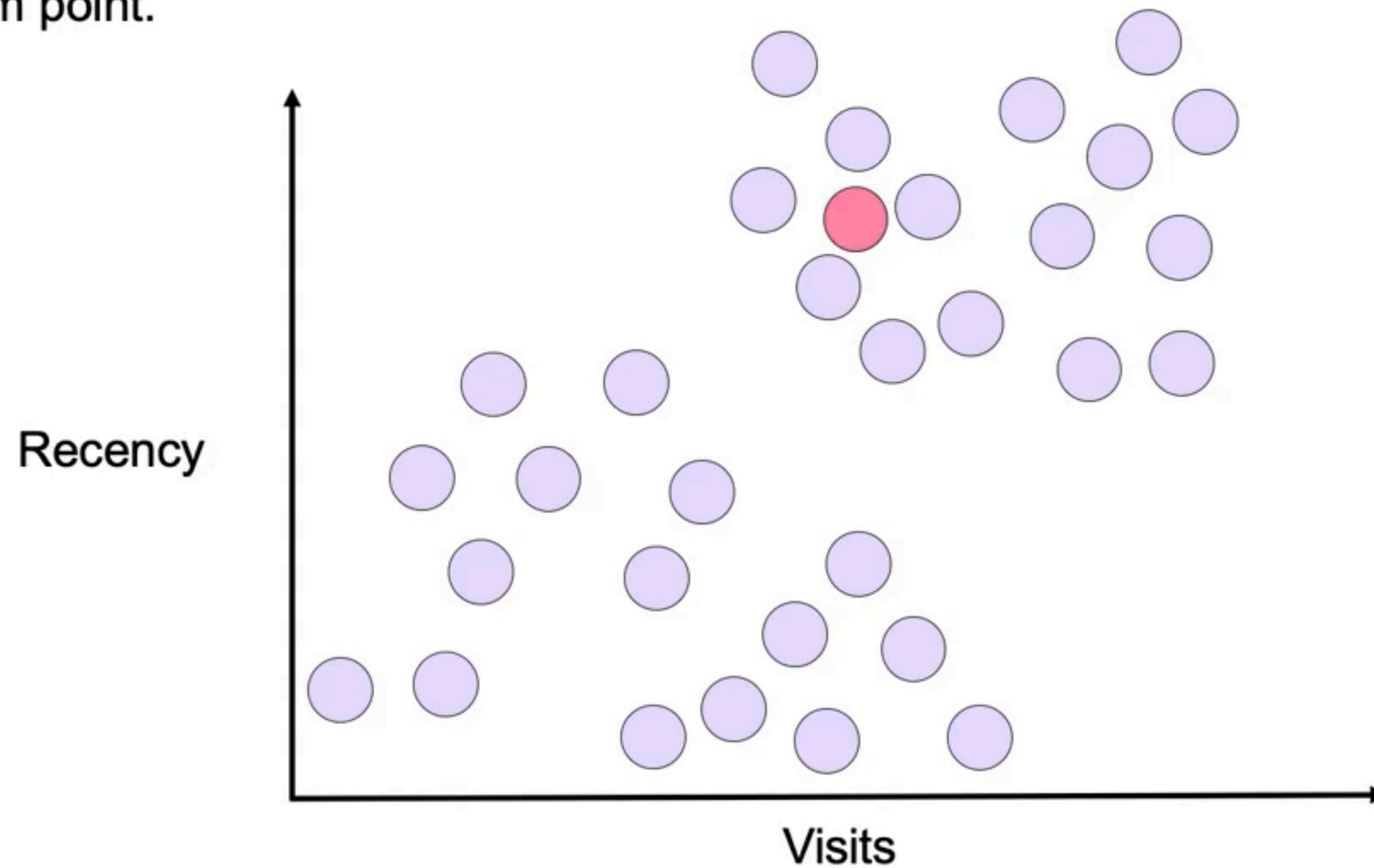
## UN VERDADERO MÉTODO DE CLUSTERING: Density-Based Spatial Clustering

# DBSCAN

- Permite identificar puntos que no pertenecen a ningún cluster
- Consideración: Puntos en un cluster tienen que estar cerca unos a otros dentro de un vecindario
- Se seleccionan puntos aleatorios de zonas densas y se expanden los clusters incluyendo SOLO los puntos que están a cierta distancia de los puntos que se han incluido iterativamente en el cluster.
- Termina cuando no hay puntos a una distancia mínima
- Inputs:
  - Métrica
  - $\epsilon$ : Radio del vecindario local
  - $N_{clu}$  Umbral de densidad
- Los puntos pueden:
  - **Puntos del Core:** Aquellos puntos que tienen más vecinos que  $N_{clu}$  incluyendo.
  - **Puntos del borde:** Pueden ser alcanzados por un punto de Core pero tiene menos de  $N_{clu}$  vecinos
  - **Noise:** Puntos que no tienen puntos de core en su vecindario
- **PROS:** No se debe introducir el número de clusters, permite ruido, puede manejar clusters de forma arbitraria
- **CONS:** Requiere dos parámetros, encontrar valores apropiados de  $\epsilon$  y  $N_{clu}$  puede ser difícil, No se desempeña bien en clusters de diferente densidad

# DBSCAN

Start at a random point.



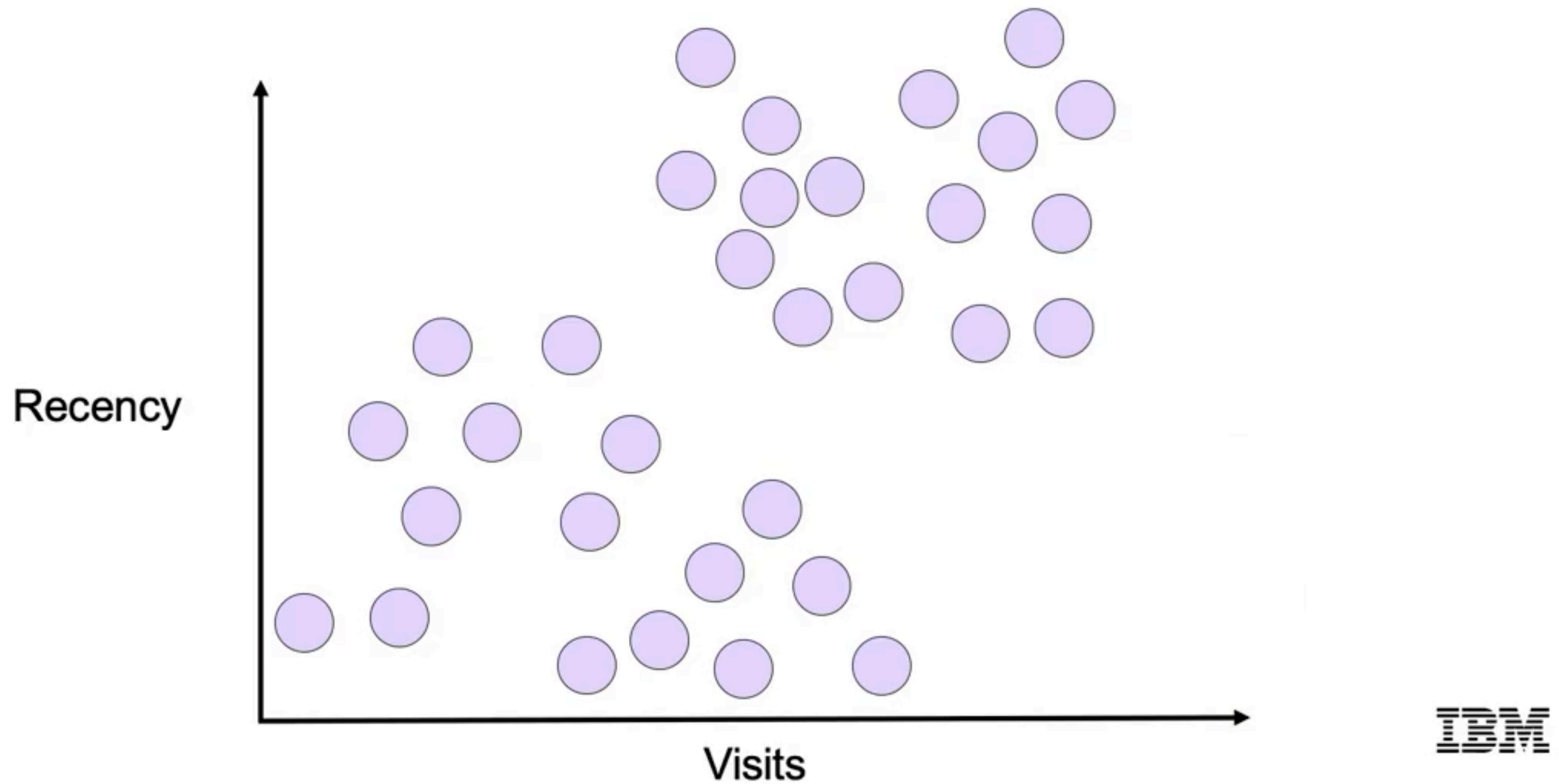
# Mean Shift

- Similar a K-Means en términos de que se dividen los puntos de acuerdo al centroide más cercano.
- El centroide se toma como el punto de mayor densidad local.
- Términa cuando todos los puntos son asignados a un cluster
- Se calcula el punto más denso encontrando el promedio pesado alrededor de cada punto (se asigna más peso a los puntos más cerca al punto original en una ventana)
- PROS: No asume número de clusters o su forma, solo se tiene un parámetro, es robusto a outliers.
- CONS: Resultado depende del tamaño de la ventana (hay ayuda que demora  $n^2$ ), complejidad proporcional a  $mn^2$  con  $m$  iteraciones y  $n$  número de puntos

# Mean Shift

1. Elegir un punto y una ventana  $W$
2. Calcular el promedio pesado  $m = \frac{\sum_{x_i \in W} x_i K(x_i - x)}{\sum_{x_i \in W} K(x_i - x)}$
3. Mueva el centroide al nuevo promedio
4. Repetir 2 y 3 hasta la convergencia (no se muevan los centroides)  $\rightarrow$  Se encuentra un Modo
5. Repetir 1-4 para todos los puntos
6. Los puntos que conducen al mismo Modo pertenecen al mismo cluster

# Mean Shift



# Comparando métodos de Clustering

MÉTODO	K-MEANS	MEAN SHIFT	AGLOMERATIVO JERÁRQUICO	DBSCAN
PARÁMETROS	Número de clusters	Tamaño de ventana	Número de clusters	$\epsilon, N_{clu}$
ESCALABILIDAD	Muy alta (Mini Batch) con k bajo/medio	No escalable	Alta	Muy alta con numero de clusters medio
USO GENERAL	Uso general Geometría circular	Muchos clusters Tamaño y geometría desigual	Muchos clusters Posibles ligaduras de conexión	Geometría no plana Tamaño desigual Detección de outliers
APLICACIONES	Clusters de igual tamaño	Identificar número de clusters Usado en video	Clusters diferente tamaño	Vision computacional Detección de outliers

