

# Minería de Datos

## IIC2433

Algoritmos de Clustering

Vicente Domínguez

# ¿Qué veremos esta clase?

- Diversos algoritmos de clustering

# Aprendizaje no supervisado

## *Clustering*

**Tarea para el computador:**

Identificar grupos de elementos similares

# Aprendizaje no supervisado

## *Clustering*

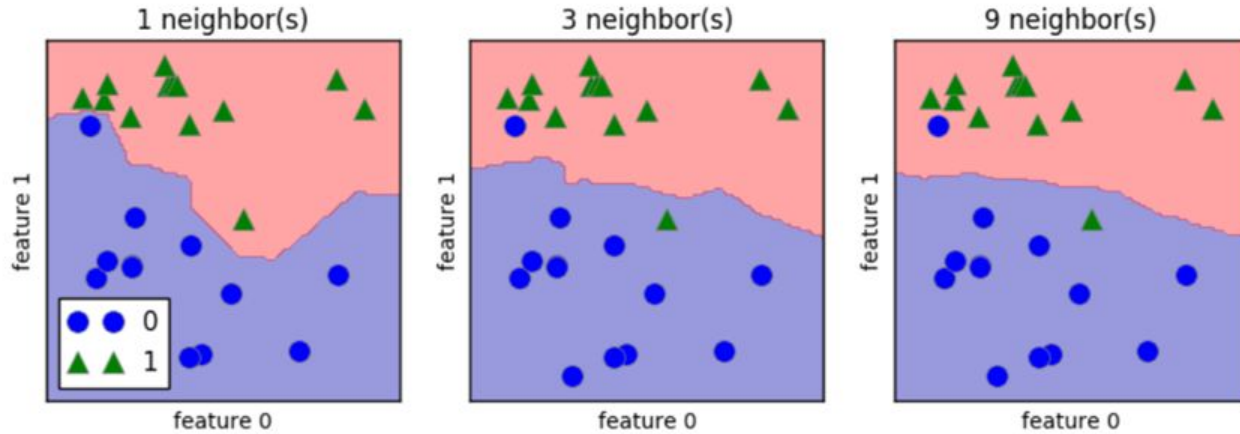
Conjunto de datos **no etiquetados**



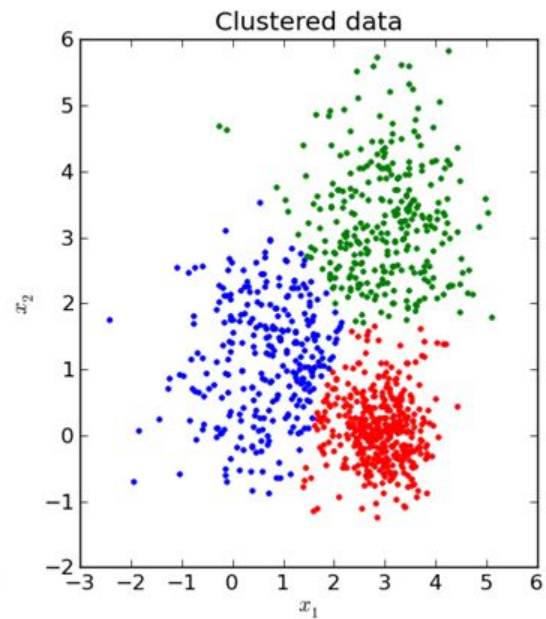
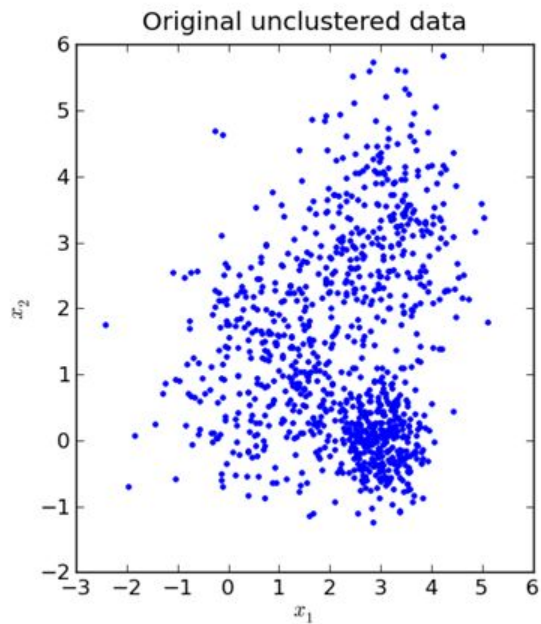
# Clustering

- Técnica utilizada para análisis y visualización de datos.
- No necesita labels o clases.
- Permite identificar grupos en los datos, también posibles outliers.

# Clasificación

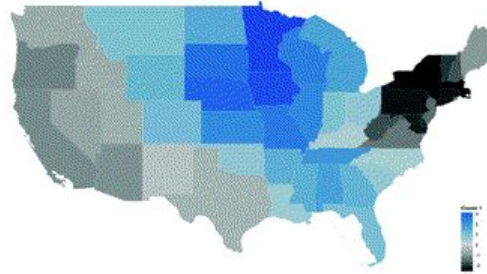


# Clustering

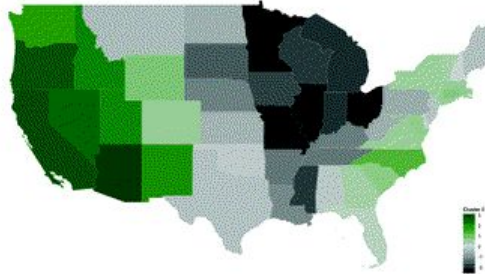


# Clustering en mapas

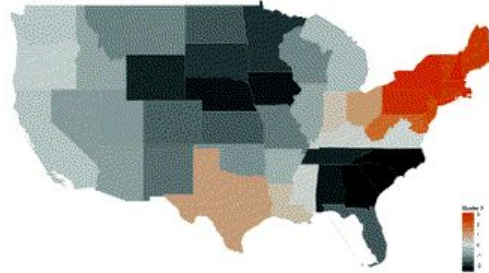
A. Cluster 1: Friendly & Conventional Region



B. Cluster 2: Relaxed & Creative Region



C. Cluster 3: Temperamental & Uninhibited Region





# Clustering de Galaxias

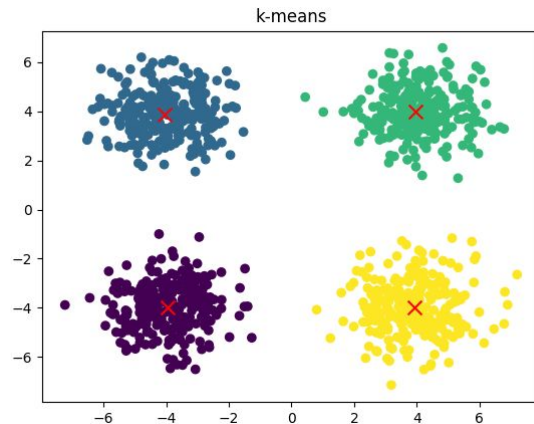


# Clustering de Imágenes



# K-Means

Buscar  $k$  centros y clusters, tales que cada centro sea la *media* o *centroide* de su respectivo cluster y cada elemento pertenezca a al cluster de su centro más cercano



# Vemos cómo calcular el centroides

Video de ejecución

<https://www.youtube.com/watch?v=5I3Ei69I40s>

# Algoritmo k-means

- Definir centros aleatorios
- Ejecutar los siguientes pasos iterativamente:
  - 1) Asignar cada elemento del dataset al cluster de su centro más cercano
  - 2) Recalcular los centros de cada cluster
- Repetir 1) y 2) hasta que los centros “dejen de moverse”

# Ejercicio

Suponga que tenemos 4 medicinas distintas y conocemos características para cada una de ellas (atributos).

Agrupe las medicinas en dos grupos distintos basándonos en las dos características:

Suponga que los valores iniciales de los centros son (1,1) y (2,1):

	x1	x2
Medicina A	1	1
Medicina B	2	1
Medicina C	4	3
Medicina D	5	4

# K-Means

- Los cluster finales son muy dependientes de los puntos iniciales de los centros
- Un centroide puede no ser un punto de la base de datos (como los centroides iniciales)
- Se puede ejecutar varias veces el algoritmo y ver cuales son los clusters que más aparecen en promedio

# K-Means

- ¿Siempre es bueno usar K-Means?
- ¿Cómo encuentro clusters que son tendencias en los datos?

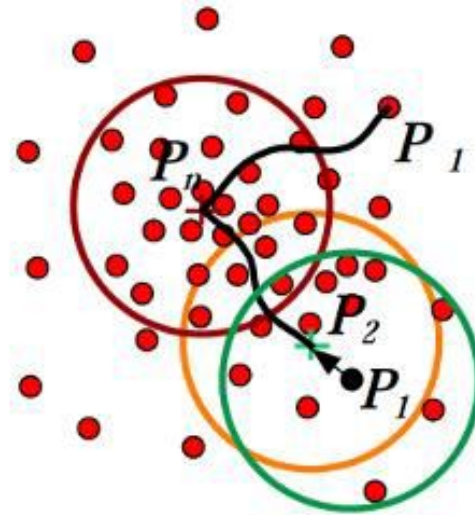


# Mean Shift

- Algoritmo que busca aglomeraciones de puntos que siguen una tendencia.
- No es necesario saber a priori la cantidad de clusters a encontrar pero si conocer la distribución de los datos
- Muy sensible a sus parámetros iniciales.

# Mean Shift

Este algoritmo considera una vecindad local a cada centro y mueve el centro en la dirección de mayor aumento de densidad



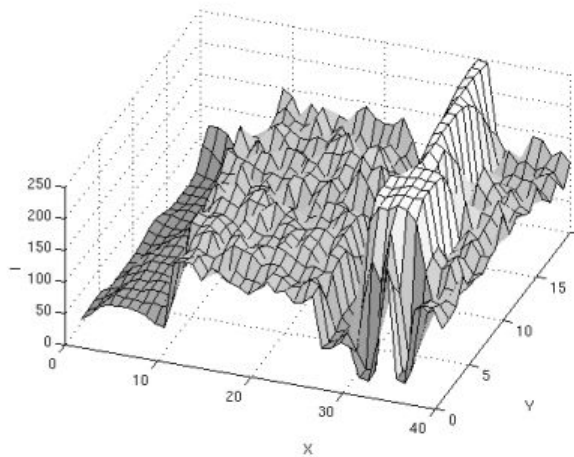
# Mean Shift Algoritmo

- Por cada punto, calculo su vecindad a una distancia dada.
- Calculo la media de la vecindad.
- Me muevo a esa nueva posición de la media y vuelvo al paso anterior.
- Repetir hasta que converger a un punto.
- Finalmente, todos los puntos que llegaron al mismo punto final son un cluster.

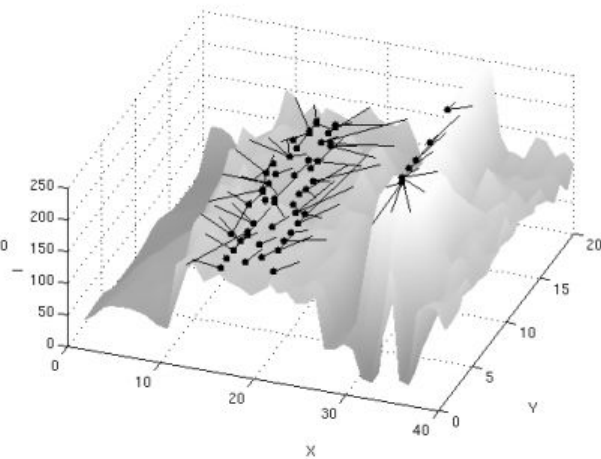
# Video Mean Shift

- <https://www.youtube.com/watch?v=TMPEujQrY70>

# Mean Shift



(a)



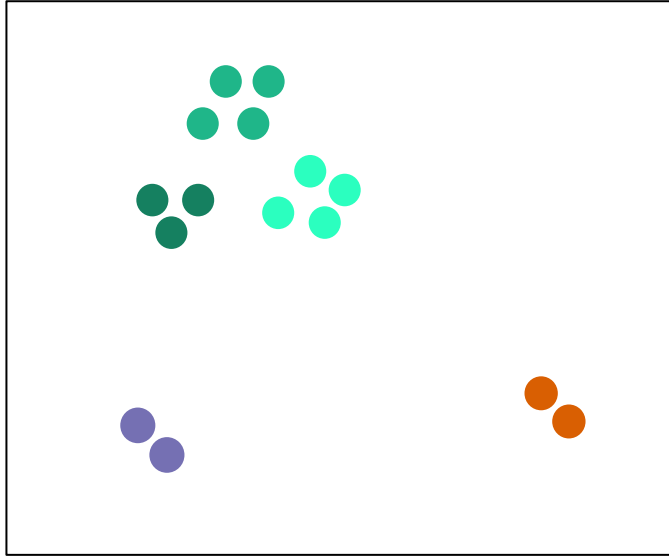
(b)

## Otro tipo de clustering

- Aparte de los algoritmos vistos anteriormente existen algoritmos que tratan de generar una jerarquía dentro de los datos
- Estos algoritmos son conocidos como clustering jerárquico

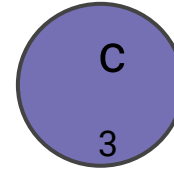
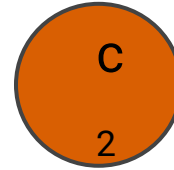
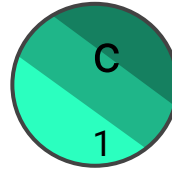
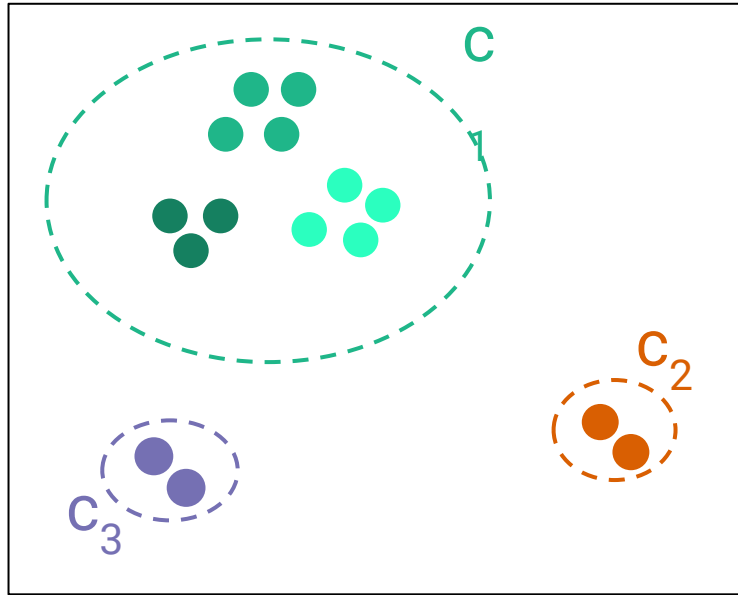
# Clustering Jerárquico

*Clusters y subclusters*



# Clustering Jerárquico

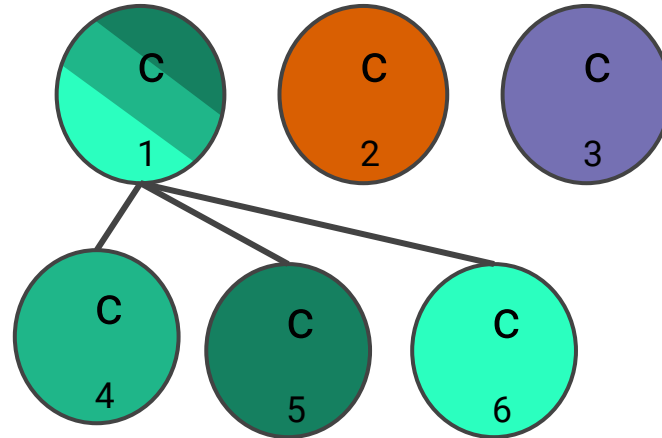
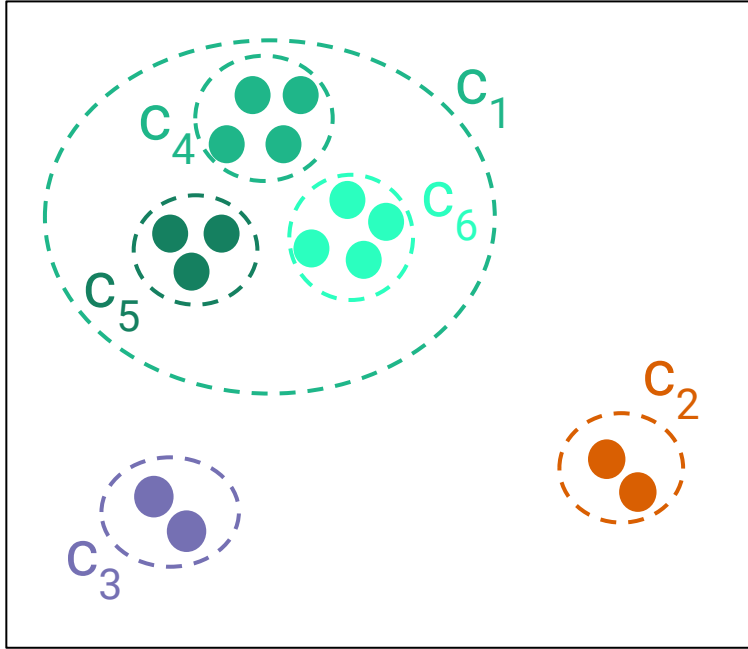
## *Clusters y subclusters*





# Clustering Jerárquico

## *Clusters y subclusters*



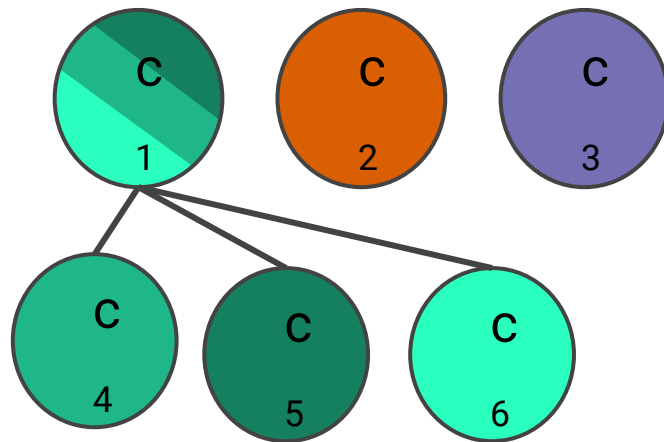
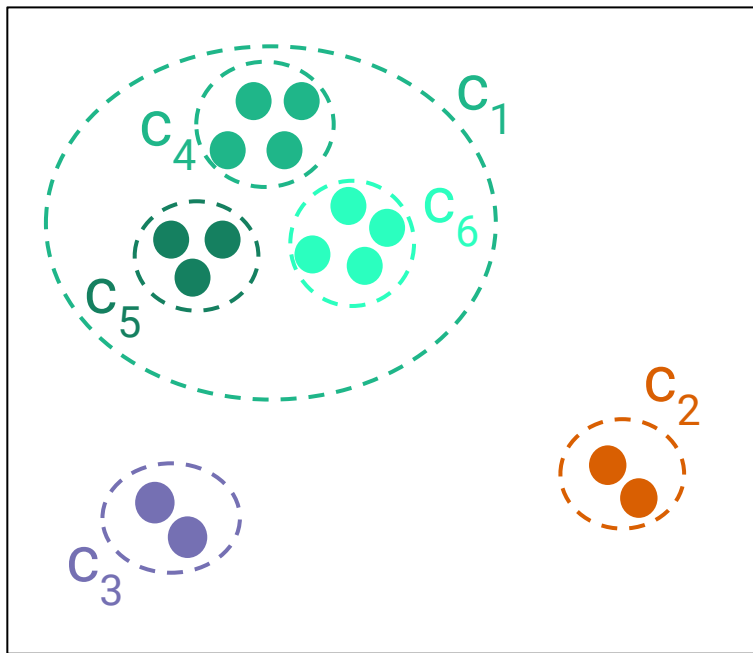
# Clustering Jerárquico

## *Tipos*

- Clustering Jerárquico **divisivo**
- Clustering Jerárquico **aglomerativo**

# Clustering Jerárquico

*Ejemplo: k-means jerárquico*



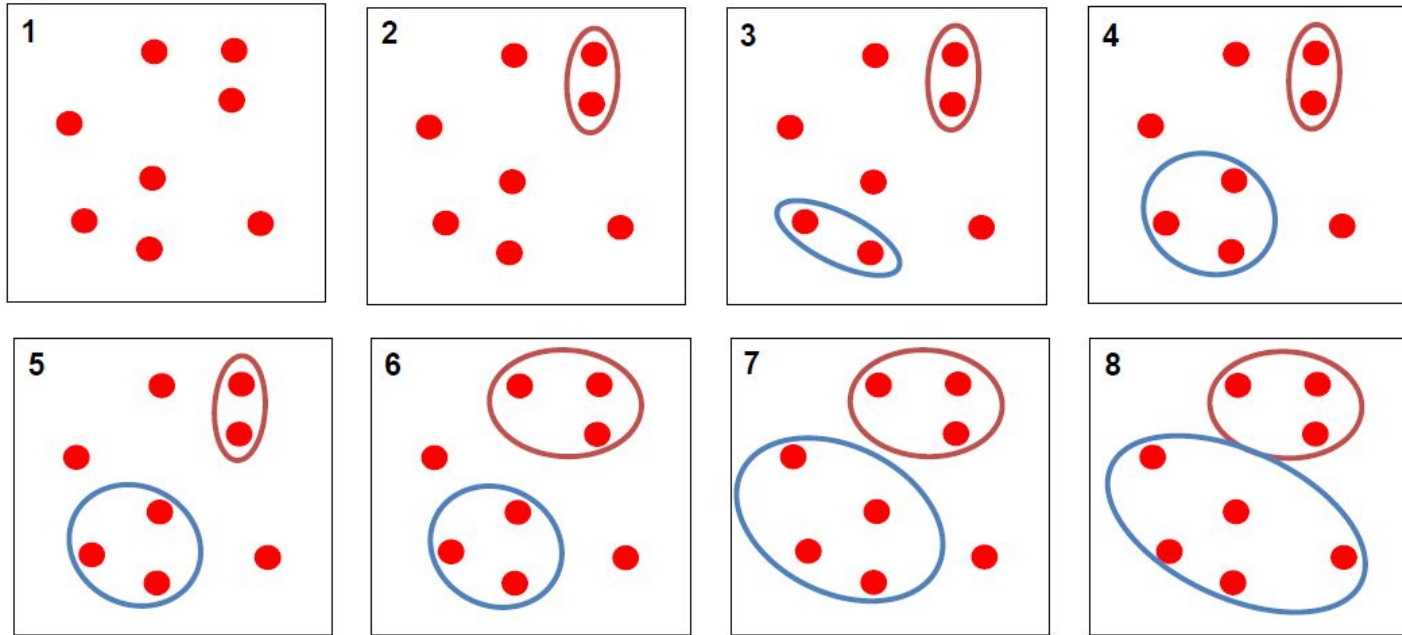
# Clustering jerárquico aglomerativo

- Inicialmente cada elemento se considera un cluster
- Iterativamente se van juntando los clusters más cercanos

# Clustering jerárquico aglomerativo

1. Calcular matriz de proximidad/similitud
2. Al comienzo cada punto es un cluster
3. Repetir:
  - a. Unir los dos clusters más cercanos
  - b. Actualizar la matriz de proximidad
4. Hasta: cuando la distancia entre los clusters a unir supere algún umbral predeterminado

# Clustering jerárquico aglomerativo



# Criterios de enlace

- Complete-linkage (enlace completo)

$$D(C_a, C_b) = \text{Max}\{d(i, j)\}, i \in C_a, j \in C_b$$

- Single-linkage (enlace único)

$$D(C_a, C_b) = \text{Min}\{d(i, j)\}, i \in C_a, j \in C_b$$

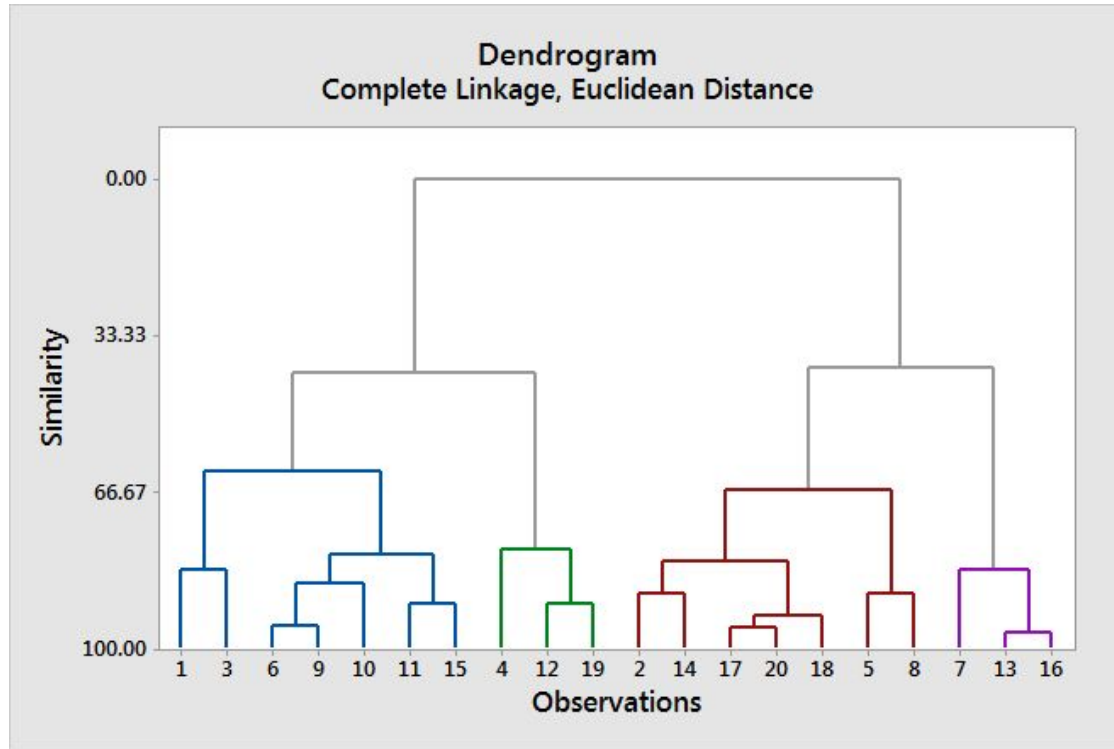
- Distancia entre medias (mean distance)

$$D(C_a, C_b) = D(\mu_a, \mu_b)$$

- Distancia promedio entre pares ( average pairwise distance)

$$D(C_a, C_b) = \text{avg}\{d(i, j)\}, i \in C_a, j \in C_b$$

# Dendrograma / Dendrograma





# Clustering jerárquico aglomerativo

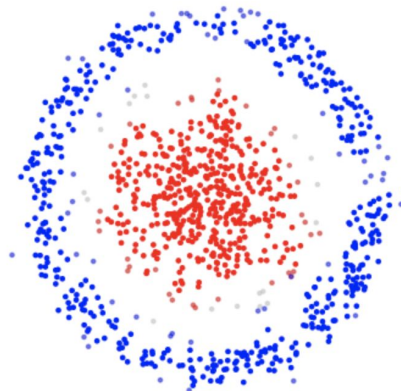
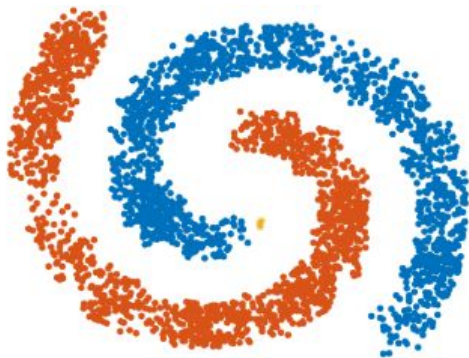
- Inicialmente cada elemento se considera un cluster
- Iterativamente se van juntando los clusters más cercanos

Ejemplo:

	x1	x2
Perro 1	1	1
Perro 2	2	1
Gato 1	4	3
Gato 2	5	4
Iguana 1	15	16
Iguana 2	15	15

# ¿Qué algoritmo usar?

¿Qué algoritmo utilizarían para este ejemplo?

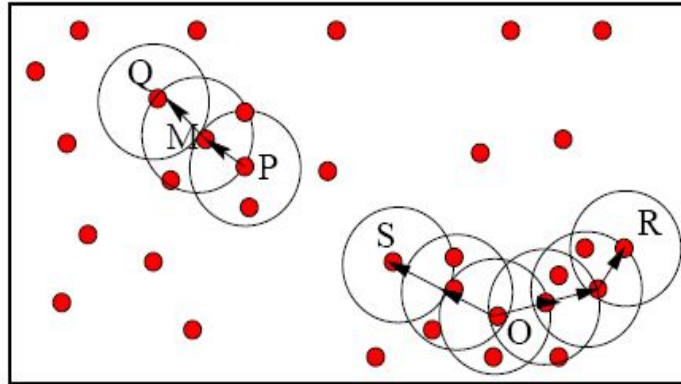


# DBScan

- Forma clusters de formas arbitrarias
- Se define un radio y un nivel de densidad de puntos en ese radio
- Si hay un conjunto de puntos que sean alcanzable en el radio y mantienen la densidad mínimo, me voy moviendo entre ellos.
- Un cluster está dado por el conjunto de puntos alcanzables en cadena.

# DBScan

- Si hay un conjunto de puntos que sean alcanzable en el radio y mantienen la densidad mínimo, me voy moviendo entre ellos.
- Un cluster está dado por el conjunto de puntos alcanzables en cadena.



# Comparación de algoritmos de clustering

