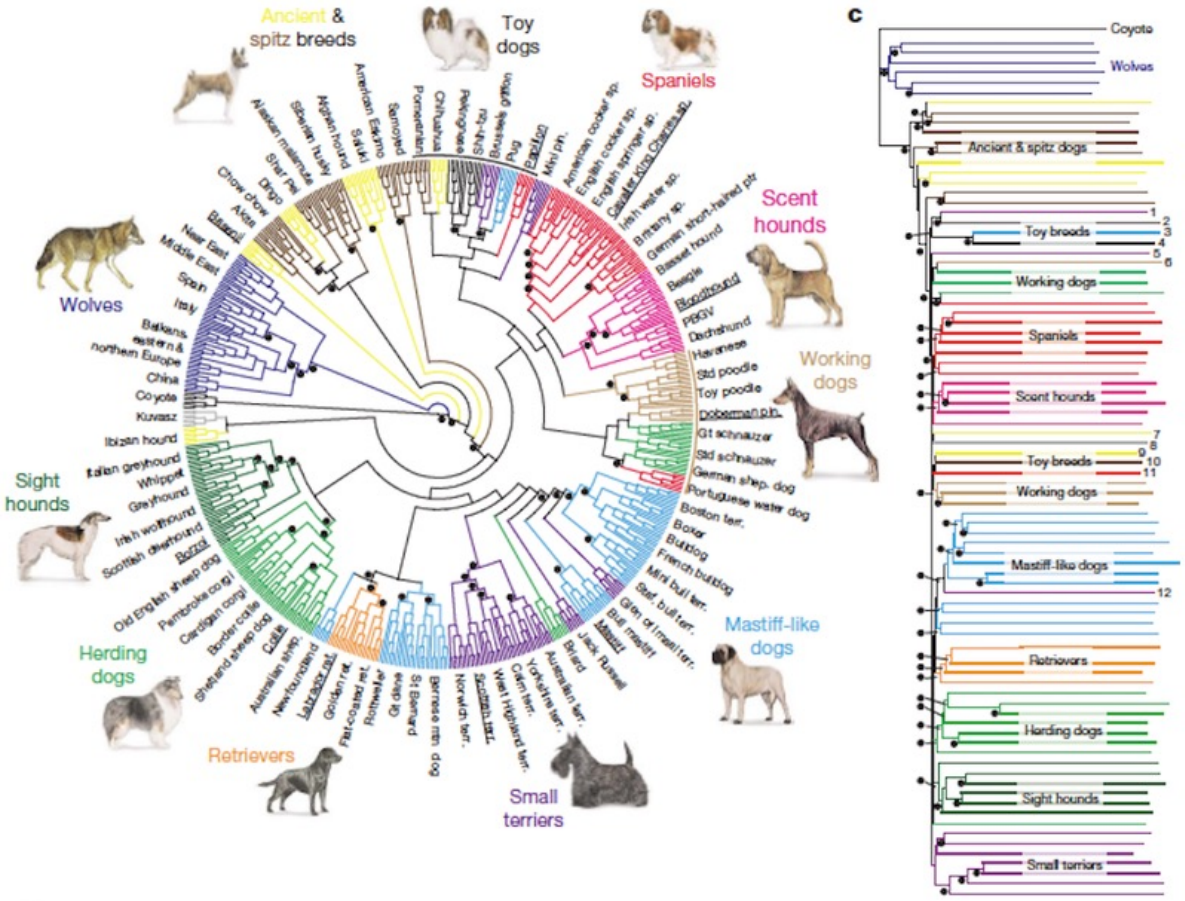
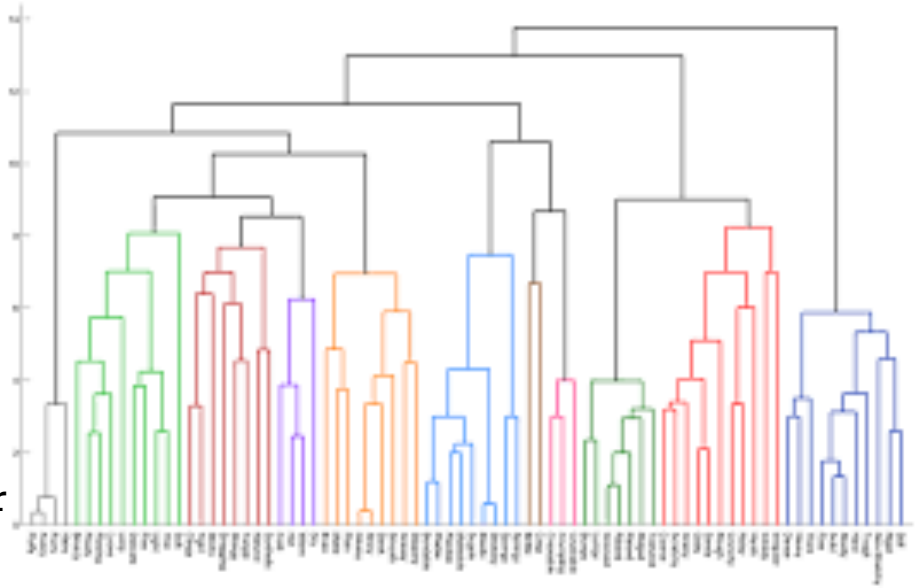
CLUSTERING JERÁRQUICO



UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO

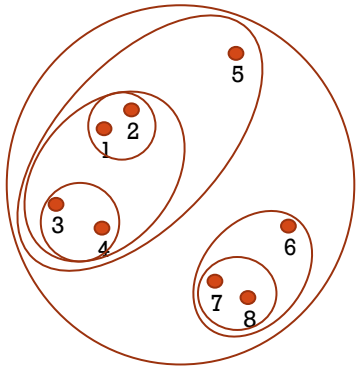
- Aproximación **bottom-up**
- Produce como resultado un **dendrograma**
  - Basado en las **distancias** entre instancias y entre clusters
  - Determina todas las segmentaciones posibles, permitiendo su visualización
- No se necesita repetir el proceso para diferentes valores de **K**
- Las instancias **excepcionales** pueden ser rápidamente identificadas

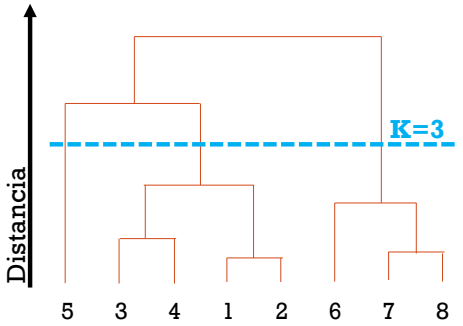


UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO

- Algoritmo (iterativo):
  1. Al inicio cada instancia es un cluster (n clusters)
  2. Se identifica el par de clusters más cercanos y se fusionan (n-1 clusters)
  3. Se repite el paso anterior hasta que queda un solo cluster con todas las instancias
  4. Se escoge un punto de corte
- Los clusters se pueden organizar en forma de **dendrograma**
- Es necesario definir como **fusionar** clusters y la **distancia** a utilizar

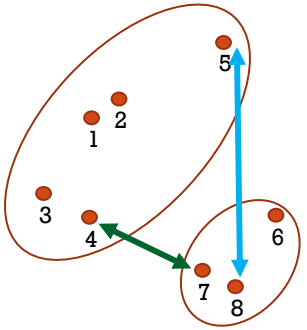




UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO

- **Fusión entre clusters** basadas en el cómputo de las distancias entre todos los pares de puntos de cada cluster:
  - **Single linkage**:
    - Distancia mínima entre dos puntos de los dos clusters.
    - Resultan clusters formados por "cadenas" de puntos, usualmente con fusiones consecutivas entre un cluster y un punto cercano
    - Sensible al ruido y a las excepciones
  - **Complete linkage**:
    - Distancia máxima entre dos puntos de los dos clusters.
    - Tiende hacia clusters esféricos con diámetros consistentes
  - **Average linkage**:
    - Promedio de las distancias entre todos los pares de puntos
    - Punto intermedio entre single y complete linkage
    - Menos afectado por las excepciones

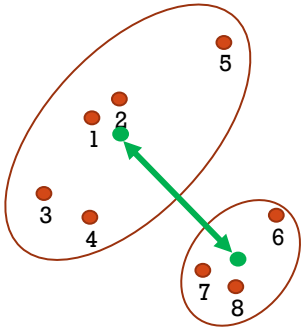


→ Complete y average se prefieren sobre single linkage

UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO

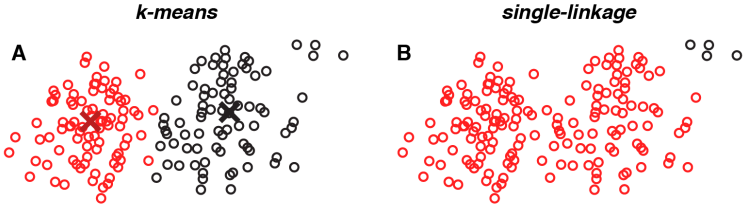
- Otros tipos de **fusión entre clusters**
  - **Centroide**:
    - Distancia entre los centroides de los clusters
    - Sufre de *inversiones*, cuando el punto de fusión de dos clusters en el dendrograma es inferior al de alguno de los clusters fusionados

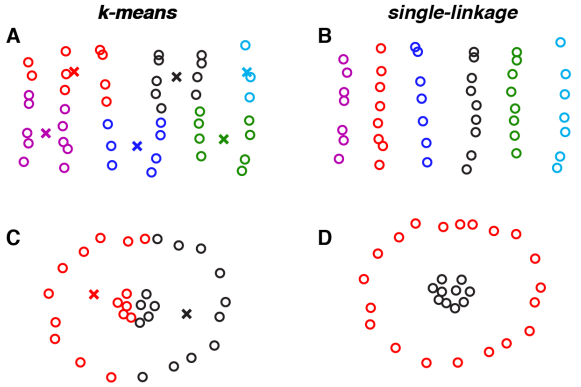


  - **Ward**: Para cada fusión se analiza el cambio en la varianza
    - Con cada fusión, la varianza global del conjunto de clusters aumenta
    - Se escoge la fusión cuyo aumento de varianza es mínimo

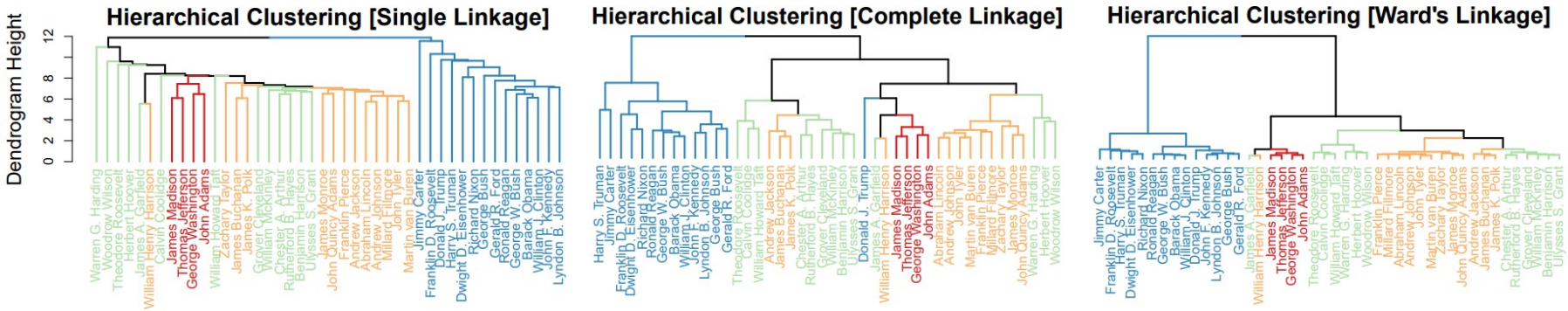
UNIVERSIDAD ICESI

CLUSTERING JERÁRQUICO





http://alexhwilliams.info/itsneuronalblog/2015/09/11/clustering1/



https://arxiv.org/pdf/1901.01477.pdf

CLUSTERING JERÁRQUICO

- Consideraciones
  - La pertenencia de las instancias a los clusters es absoluta
  - Requiere poder de cálculo computacional grande
  - Una vez una fusión se decide, no hay vuelta atrás
  - Dependiendo de la distancia utilizada y al tipo de fusión:
    - Sensible al ruido y a excepciones
    - Dificulta gestionar clusters de tamaños diferentes o no convexos
    - Puede llegar a particionar clusters grandes
  - Influencia de las unidades de los atributos utilizados → estandarización
  - Qué punto de corte (k) escoger?

UNIVERSIDAD ICESI

TALLER: CLUSTERING JERÁRQUICO

Realizar el taller 10-SYNTH- HClust-STUD.html con datos sintéticos

UNIVERSIDAD ICESI