Instituto Andaluz Interuniversitario en
Data Science and **Computational Intelligence**

DaSCI

Universidad de Jaén

Seminario Permanente de Formación en Inteligencia Artificial aplicada a Defensa
SIADEF

Sesión 7: Aprendizaje no supervisado

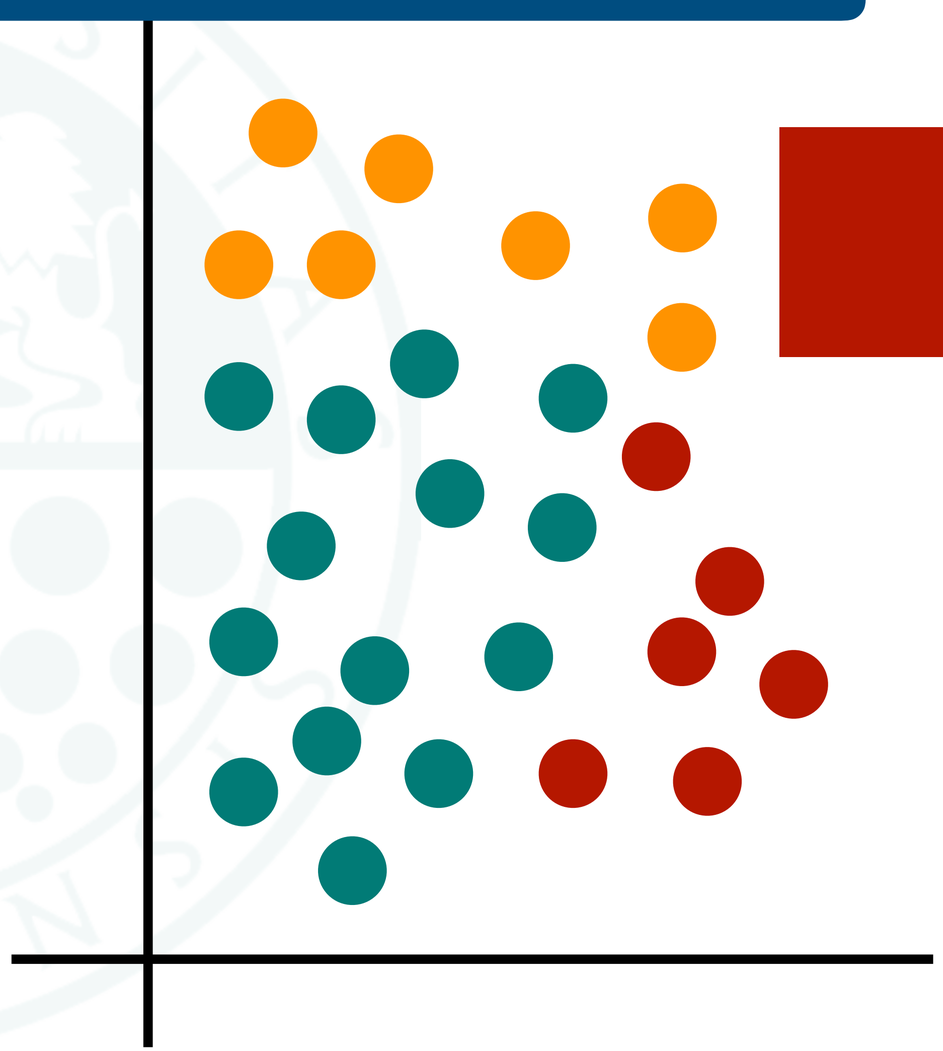
Cristóbal J. Carmona <ccarmona@ujaen.es>
Pedro González <pglez@ujaen.es>

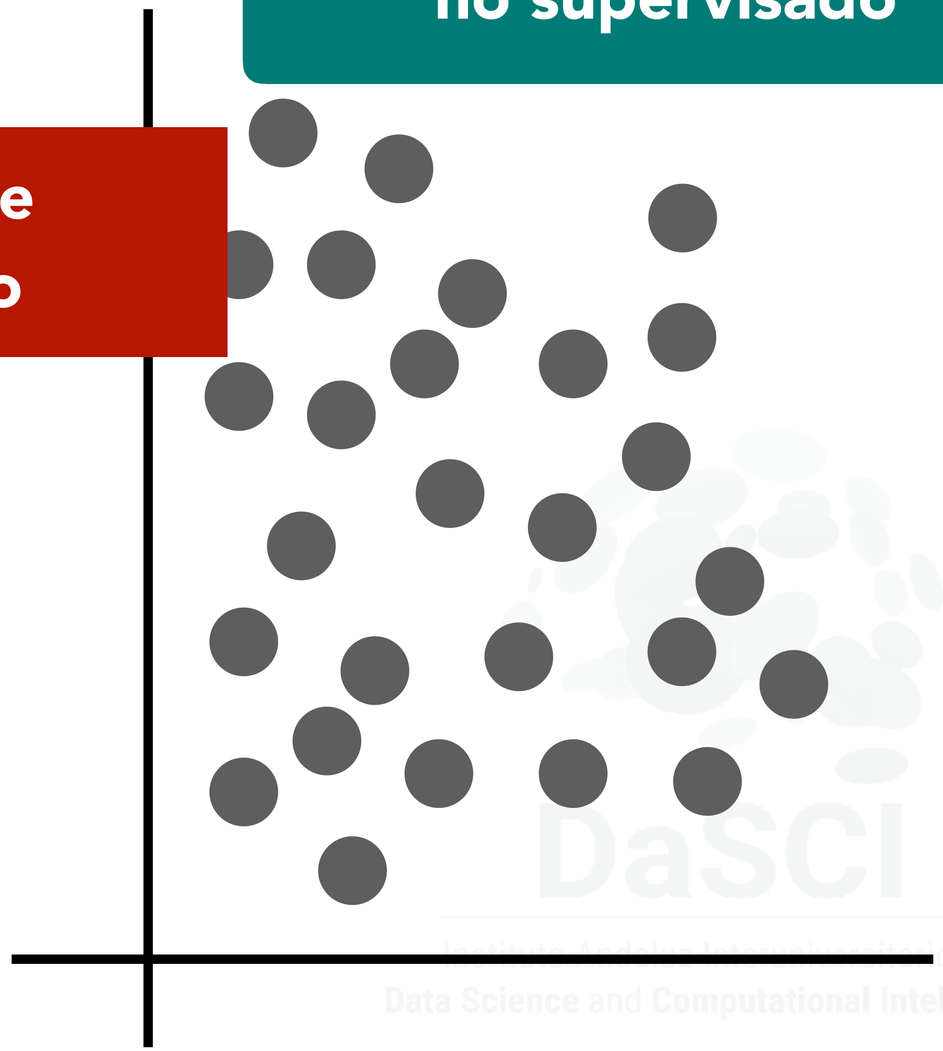
Aprendizaje automático

Aprendizaje supervisado

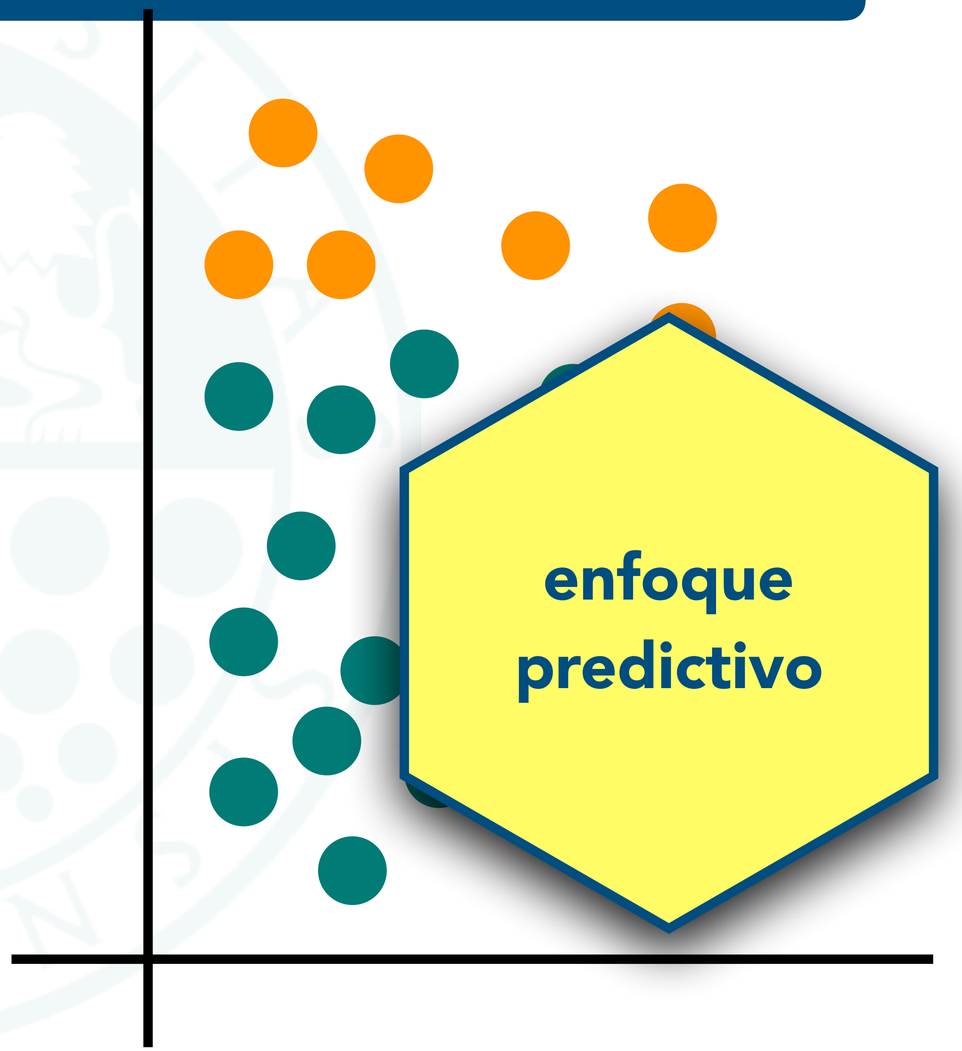
Aprendizaje automático

Aprendizaje no supervisado

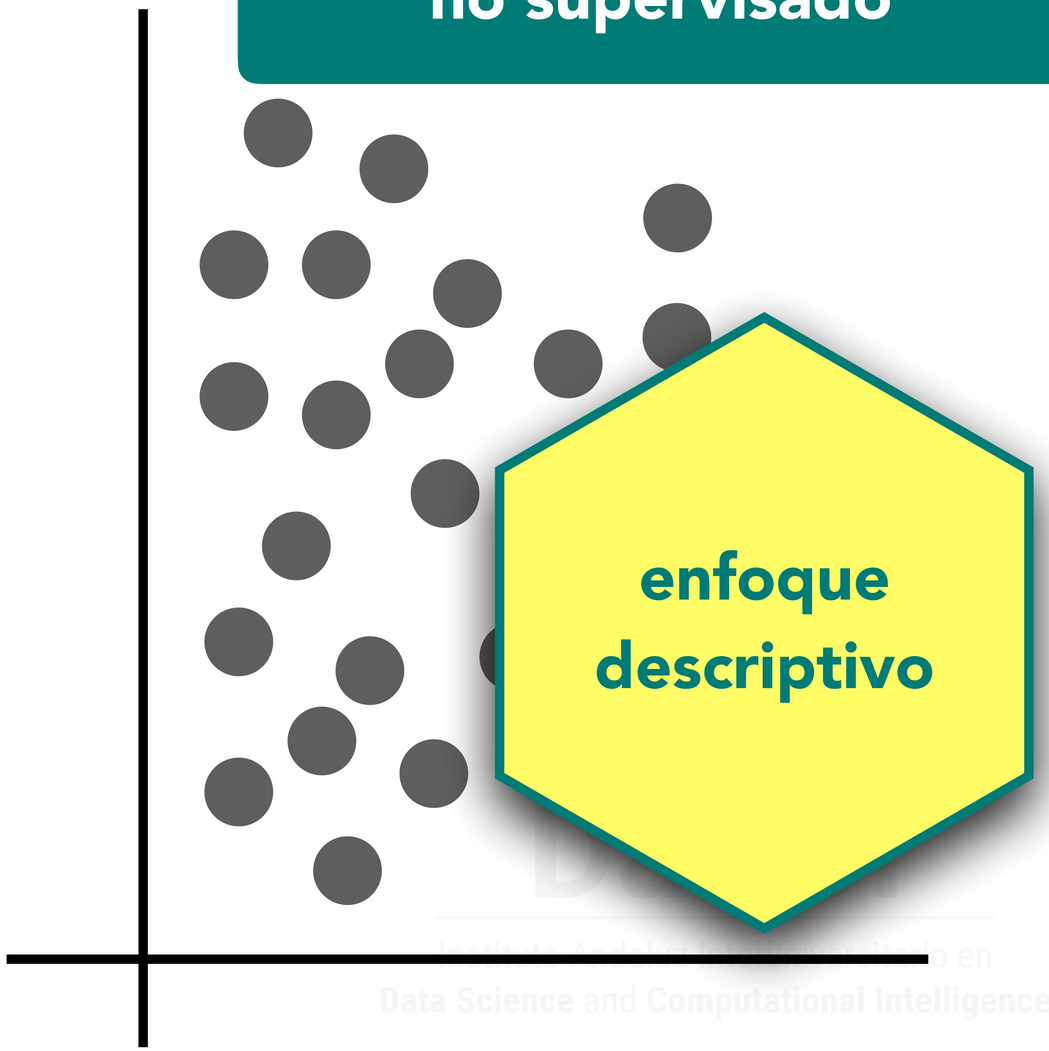




Aprendizaje supervisado



Aprendizaje no supervisado



Aprendizaje supervisado

- Clasificación
- Regresión
- Series temporales

enfoque predictivo

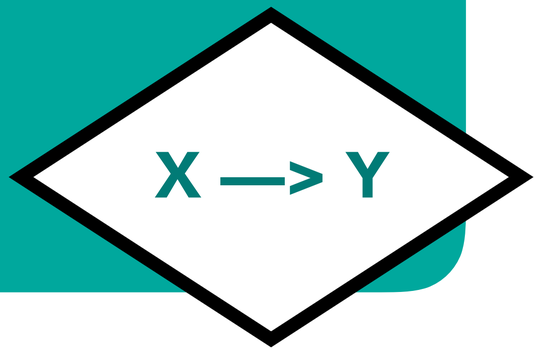
Aprendizaje no supervisado

- Asociación
- Agrupamiento

enfoque descriptivo

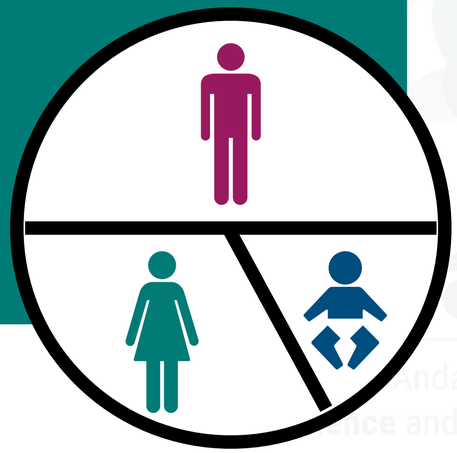
Asociación

conceptos básicos
ejemplos
librerías



Agrupamiento

conceptos básicos
ejemplos
paquetes en R



Asociación

Asociación

motivación

Para muchas empresas es un factor de éxito **comprender y actuar sobre los patrones de los clientes** utilizando los datos de las transacciones realizadas.

Sería muy interesante disponer de una herramienta que permita registrar y analizar las **transacciones** con los clientes para **revelar información valiosa sobre su comportamiento** respecto a la compra o consumo de artículos.

Asociación

definición

Dada una base de datos de transacciones, donde cada transacción es una lista de artículos (comprados por un cliente en la misma visita), encontrar todas las reglas que co-relacionen la presencia de un conjunto de artículos con otro conjunto de artículos

- **Ejemplos:**
  - Los estudiantes que cursan Inteligencia Artificial tienden a cursar también Minería de Datos.
  - El 98% de los clientes que compran neumáticos y accesorios para el automóvil, también adquieren servicios (cambio de neumáticos, …).

Asociación

conceptos básicos

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Itemset (Conjunto): colección de uno o más items

- Ejemplo: {Milk, Bread, Diaper}
- k-itemset: itemset con k items

Support count (σ) (conteo soporte): frecuencia de ocurrencia del itemset

- Ejemplo: σ ({Milk, Bread,Diaper}) = 2

Support (Soporte): fracción de transacciones que contienen un itemset

- Ejemplo: s ({Milk, Bread,Diaper}) = 2/5

Frequent Itemset (Conjunto frecuente): un itemset con soporte >= *minsup*.
En los conjuntos frecuentes los artículos están correlacionados positivamente

Asociación

conceptos básicos

Regla de asociación: *X* —> *Y*, donde X (antecedente) e Y (consecuente) son itemsets

- Ejemplo: {Milk, Diaper} —> {Beer} ("—>" es co-occurrencia, no causalidad)

Tipos de reglas de asociación:

- Asociaciones **Booleanas vs Cuantitativas** (tipo de los valores)

  compra (x,"PC") —> compra (x,"impresora") vs. ingresos (x,"2K..48K") —> compra (x,"PC")

- Asociaciones **Unidimensionales vs. Multidimensionales**

  A —> B vs. A & B & N —> D

- Análisis con **distintos niveles** de abstracción:

  Edad (x,"30..39") —> compra (x, "Tablet") vs. Edad (x,"30..39") —> compra (x,"Tablet Galaxy")

Asociación

conceptos básicos

Métricas de evaluación de reglas

- Soporte (s): fracción de las transacciones que contienen tanto a X como a Y
- Confianza (c): frecuencia con la que los items de Y aparecen en transacciones que contienen los items de X

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Ejemplo: $\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Asociación

conceptos básicos

Minería de reglas de asociación:

- Dado un conjunto *T* de transacciones, el objetivo de la minería de reglas de asociación es encontrar **TODAS** las reglas que cumplan:
  - Soporte ≥ *minsup*
  - Confianza ≥ *minconf*

Nota: encontrar estas reglas no significa que deba existir una relación entre antecedente y consecuente. Por lo tanto, un experto debería evaluarlas.

Asociación

algoritmos de descubrimiento de reglas de asociación

Objetivo: encontrar todas las reglas que superen los umbrales de soporte y confianza

Algoritmos:

- Apriori y AprioriTid (Agrawal & Srikant,1994)
- Opus (Webb,1996)
- Direct Hasing and Pruning (DHP) (Adamo, 2001)
- Dynamic Set Counting (DIC) (Adamo, 2001)
- Charm (Zaki & Hsiao, 2002)
- FP-growth (Han, Pei & Yin, 1999)
- Closet (Pei, Han & Mao, 2000)
- .....

Asociación

algoritmos de descubrimiento de reglas de asociación

Los algoritmos deben generar siempre el mismo conocimiento

¿Qué los hace diferentes?

- Forma en que los datos son cargados en memoria.
- Tiempo de procesamiento.
- Tipos de atributos (numéricos, categóricos).
- Forma en que generan los itemsets.
- Estructura de datos utilizada.

Asociación

algoritmos de descubrimiento de reglas de asociación

Enfoque habitual: descomponer el problema en dos pasos

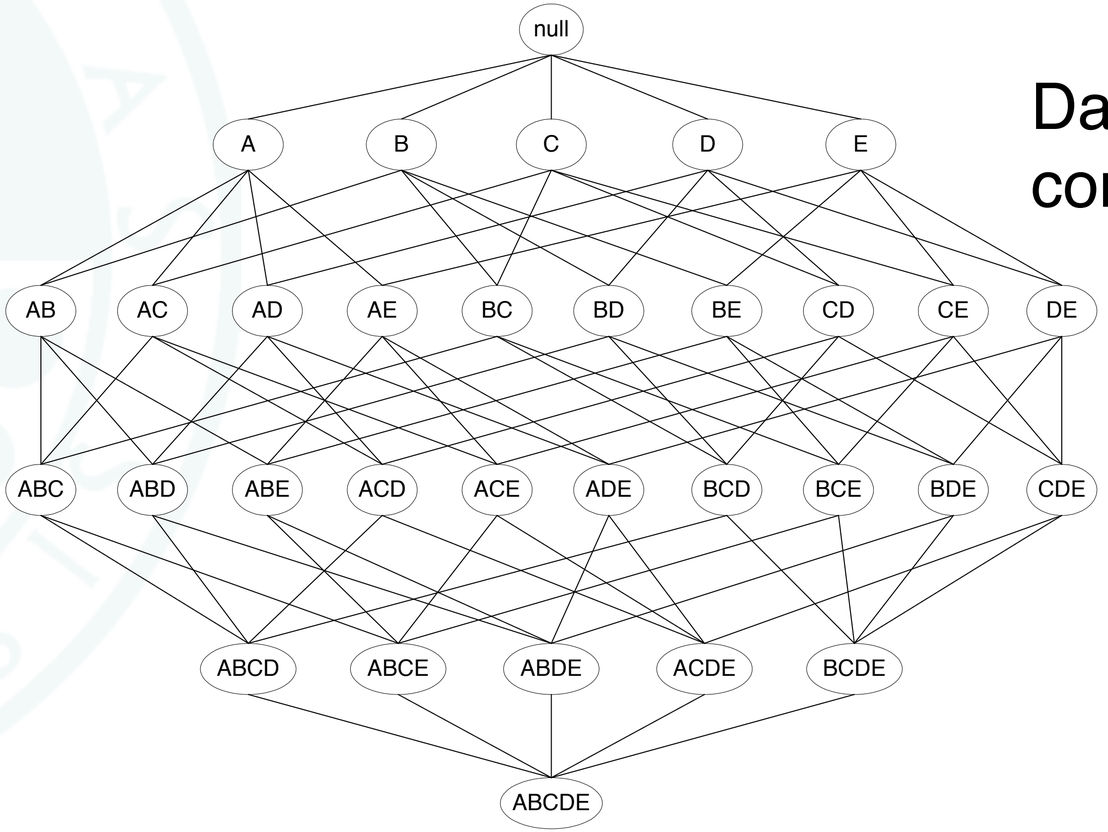
1. **Generar itemsets frecuentes**
   - Genera todos los conjuntos frecuentes con **soporte > minsup**.

2. **Utilizar itemsets frecuentes para generar reglas con fuerte asociación**
   - Generar reglas de alta confianza (**confianza > minconf**) de cada conjunto frecuente (que cada regla es una partición binaria de un conjunto frecuente).

Asociación

algoritmos de descubrimiento de reglas de asociación

La generación de conjuntos frecuentes es costosa



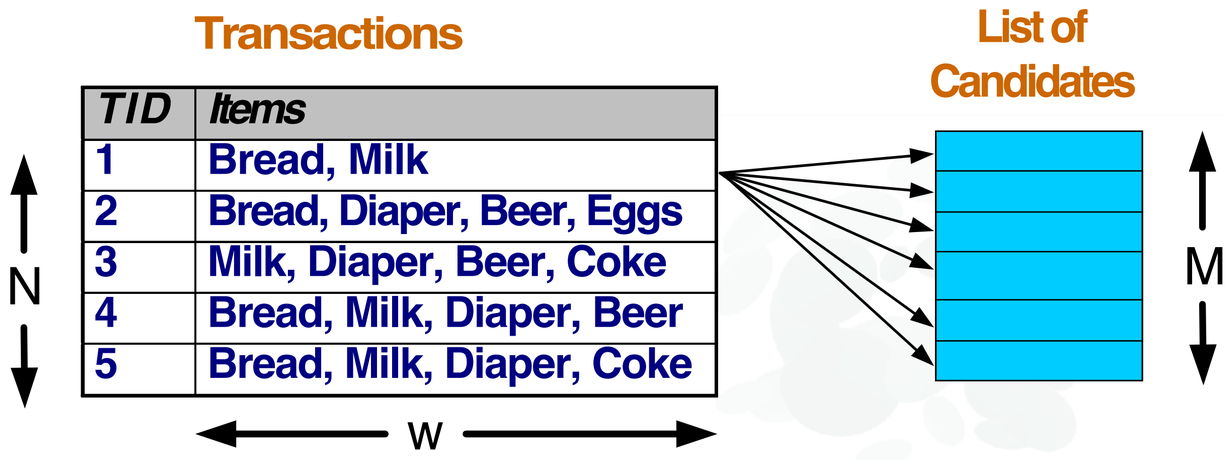
Dados d items, hay 2^d posibles conjuntos (itemsets) candidatos

Asociación

algoritmos de descubrimiento de reglas de asociación

La generación de conjuntos frecuentes es costosa

- Enfoque de fuerza bruta:
  - Cada Itemset del grafo es un conjunto frecuente candidato.
  - Calcular el soporte de cada candidato recorriendo la BD.
  - Comprobar cada transacción contra todos los candidatos.
  - Muy costoso: ~ O(NMw)



Asociación

algoritmos de descubrimiento de reglas de asociación

Estrategias de generación de conjuntos frecuentes

- Reducir el **número de candidatos** (M)
  - Búsqueda completa: $M=2^d$
  - Utilizar técnicas de poda para reducir M
- Reducir el **número de transacciones** (N)
  - Reducir el tamaño de N conforme aumenta el tamaño del Itemset
  - Lo utilizan los algoritmos DHP y de minado vertical.
- Reducir el **número de comparaciones** (NM)
  - Usar estructuras de datos eficientes para almacenar los candidatos o las transacciones.
  - No es necesario comprobar cada candidato contra cada transacción.

Asociación

`algoritmo Apriori`

Primer algoritmo de descubrimiento de reglas de asociación

Utiliza estrategias para optimizar el número de conjuntos frecuentes candidatos y de posibles reglas a generar

R. Agrawal, R. Srikant - Fast Algorithms for Mining Association Rules - Proc. of the 20th International Conference on Very Large Databases, Santiago, Chile, Sept. 1994.

Asociación

`algoritmo Apriori`

Reducción de conjuntos candidatos: principio Apriori

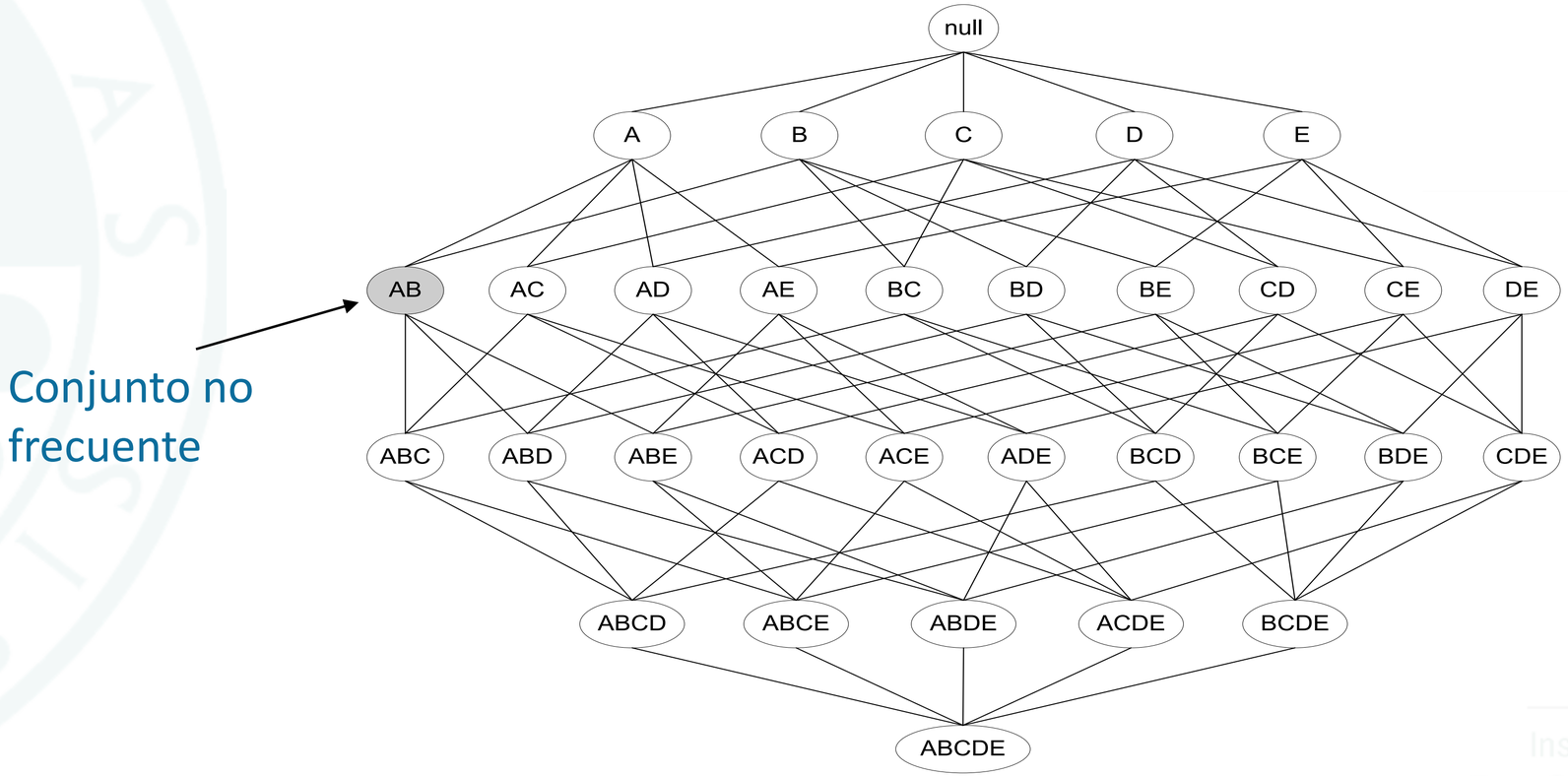
- Si un itemset es frecuente, TODOS sus subconjuntos serán frecuentes.
  - El soporte de un itemset nunca es superior al soporte de sus subconjuntos.
  - Esto se conoce como la propiedad de anti-monotonía del soporte.

Si un itemset NO es frecuente, ningún superconjunto suyo es frecuente

Asociación

algoritmo Apriori

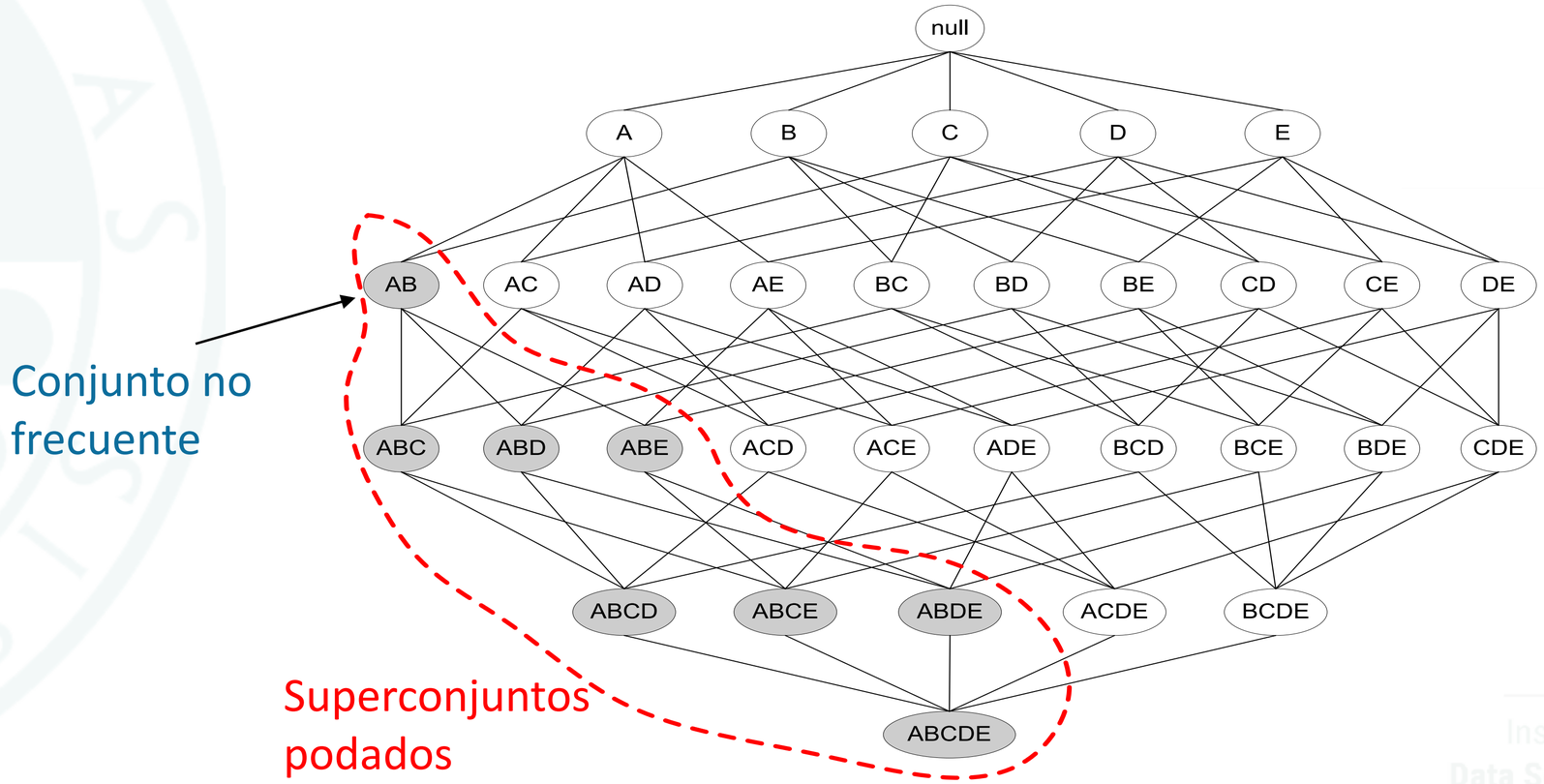
Reducción de conjuntos candidatos: principio Apriori



Asociación

`algoritmo Apriori`

Reducción de conjuntos candidatos: principio Apriori



Asociación

algoritmo Apriori

Reducción de conjuntos candidatos: principio Apriori

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Pairs (2-itemsets)

(No es necesario generar candidatos en que apareza Coke o Eggs)

Soporte mínimo = 3

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	3

Si se consideran todos los subconjuntos,

$C_{6,1} + C_{6,2} + C_{6,3} = 41$

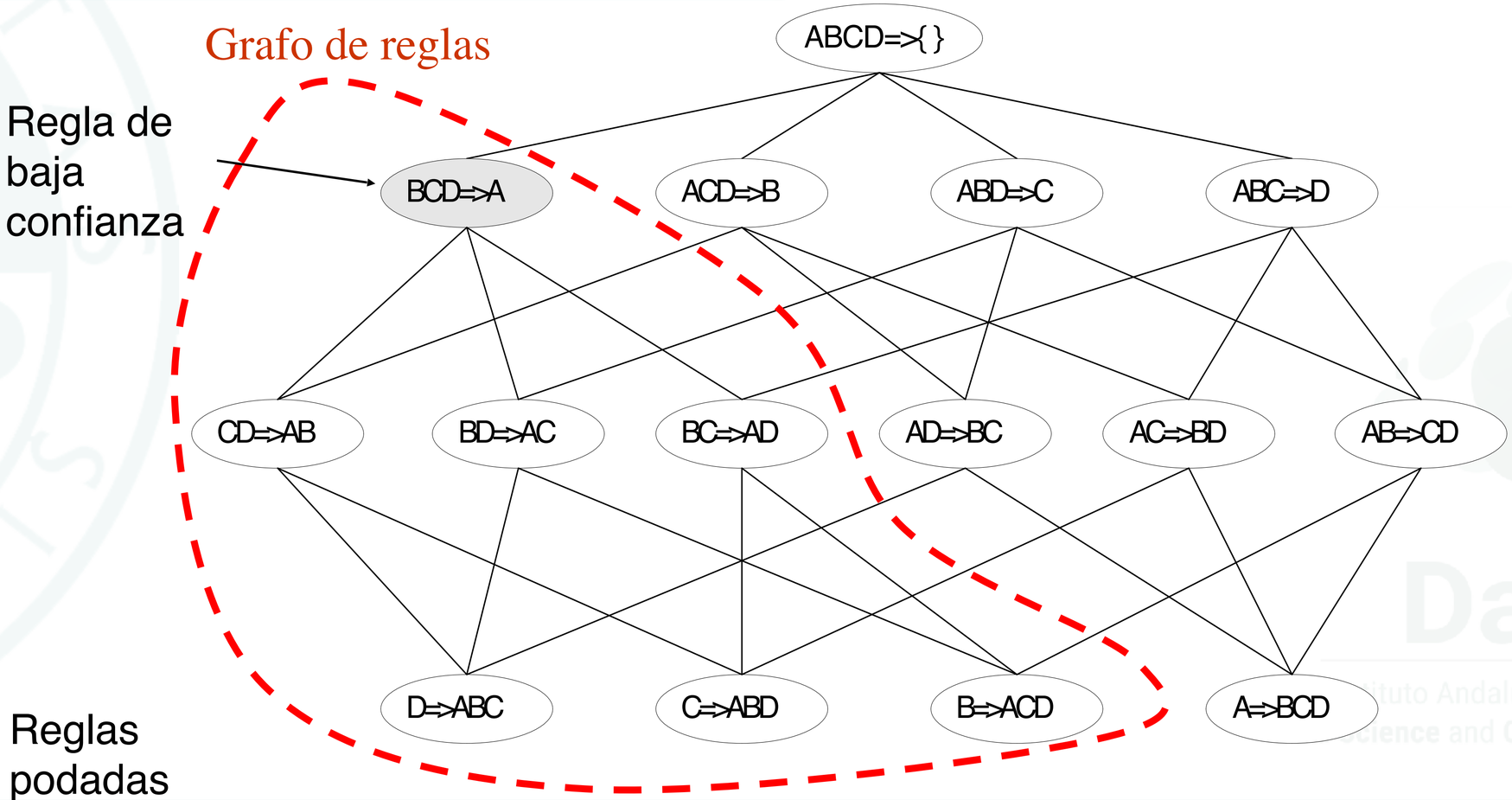
Con poda basada en soporte,

$6 + 6 + 1 = 13$

Asociación

algoritmo Apriori

Reducción de reglas candidatas

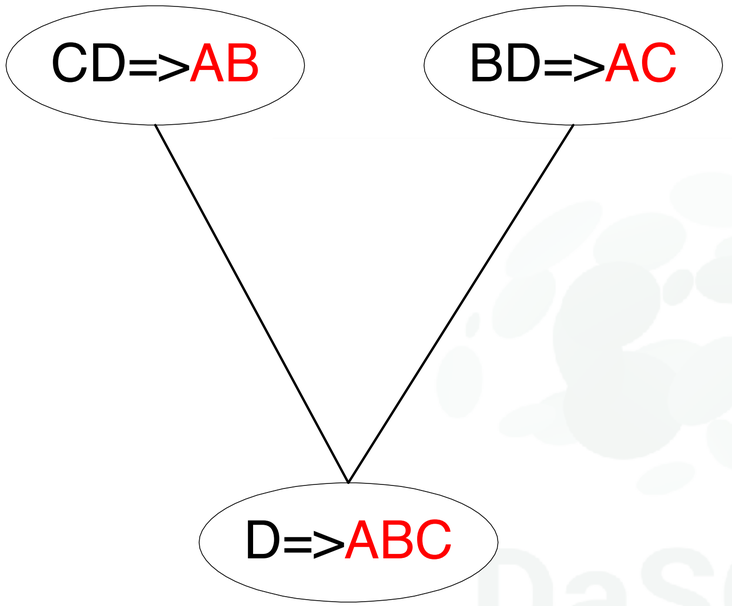


Asociación

algoritmo Apriori

Reducción de reglas candidatas

- La regla candidata se genera uniendo dos reglas que comparten el mismo prefijo en el consecuente de la regla
- join(CD=>AB,BD=>AC) produciría la regla candidata D => ABC
- Podar la regla D=>ABC si su subconjunto AD=>BC no tiene alta confianza



Asociación

`algoritmo FP-Growth`

Permite extraer reglas de asociación a partir de itemsets frecuentes sin necesidad de generar candidatos para cada tamaño.

Emplea una estructura de árbol (Frequent Pattern Tree) que almacena la información de las transacciones, comprimiéndola hasta 200 veces.

Después separa la estructura asociada a cada patrón frecuente para analizarlo de forma separada y concatenar los resultados.

En la mayoría de casos, FP-Growth es más rápido que Apriori.

Han, J.; Pei, J.; Yin, Y. Mining frequent patterns without candidate generation. In Proc. 2000 ACMSIGMOD Int. Conf. Management of Data (SIGMOD'00), 1–12. 2000.

Asociación

`evaluación de los patrones de asociación`

Algoritmos de reglas de asociación tienen a **generar demasiadas reglas**:

- Muchas de ellas no son interesantes o son redundantes.
- Aplicar **postprocesamiento** para seleccionar reglas valiosas.
- Ejemplo: $\{A,B,C\} \rightarrow \{D\}$ y $\{A,B\} \rightarrow \{D\}$
  - Son redundantes si tienen el mismo soporte y confianza.

Se pueden utilizar medidas de Interés para podar o calificar los patrones obtenidos (originalmente sólo se usaba soporte y confianza).

Asociación

evaluación de los patrones de asociación

La confianza no siempre es buena para buscar reglas interesantes:

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Regla: Tea → Coffee

Confianza = P(Coffee|Tea) = 0.75

- Por la alta confianza (0.75) parece una buena regla
- Pero es engañoso, porque P(Coffee) = 0.9
  - P(Coffee|Tea) = 0.75
  - P(Coffee|$\overline{\text{Tea}}$) = 0.9375

Asociación

evaluación patrones asociación

Medidas de interés

- **Medidas objetivas:**
  - Cataloga los patrones a partir de estadísticas calculadas sobre los datos
  - Ejemplo: 21 medidas de asociación (soporte, confianza, Laplace, Gini, información mutua, Jaccard, etc):

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left.P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

Asociación

evaluación patrones asociación

Medidas de interés

- **Medidas subjetivas:**
  - Cataloga los patrones a partir de la interpretación del usuario:
    - Un patrón es interesante subjetivamente si contradice las expectativas del usuario (Silberschatz & Tuzhilin).
    - Un patrón es interesante subjetivamente si permite actuar (Silberschatz & Tuzhilin).

Asociación

`aplicaciones`

- Recomendación de productos
- Recomendaciones de medios digitales
- Diagnóstico médico
- Optimización de contenidos
- Bioinformática
- Minería Web
- Análisis de datos científicos

Asociación

`aplicación: análisis de la cesta de la compra`

Un aplicación muy conocida es el **Análisis de la cesta de la compra** *(Market Basket Analysis)*, normalmente asociado con ventas de productos.

Asociación

`aplicación: análisis de la cesta de la compra`

Se utiliza para descubrir los patrones de co-ocurrencia de grupos específicos de artículos o items.

- Si alguien compra un grupo de artículos, es más probable (o menos) que también compre otro grupo de artículos
- Se trata de identificar grupos de artículos que son adquiridos en conjunto.

Intenta identificar reglas de la forma:

- {fideos, queso rallado} → {salsa}

Julander, C. (1992) Basket Analysis: A New Way of Analysing Scanner Data. International Journal of Retail & Distribution Management 20 (7), 10-18.

Asociación

aplicación: análisis de la cesta de la compra

Ejemplo: Compra de productos de alimentación

- Tenemos 10.000 recibos de compras de clientes.
  - Cada recibo es una transacción.
  - Cada línea de cada recibo es un artículo (no se repiten).
- Para guardar toda esa información, creamos una única tabla que contiene una fila por cada artículo de cada transacción:

Transaction	Items
A0001	citrus fruit
A0001	margarine
A0001	ready soups
A0001	semi-finished bread
A0002	coffee
A0002	tropical fruit

Asociación

aplicación: análisis de la cesta de la compra

Ejemplo: Compra de productos de alimentación

- Para usar esta información con los algoritmos de análisis de la cesta de la compra, primero debemos transformarlo a una **representación binaria** (0,1) indicando si un artículo concreto se compró en una transacción específica.
- Tendremos una única fila por transacción, y tantas columnas como artículos:

Transaction	Items
A0001	citrus fruit
A0001	margarine
A0001	ready soups
A0001	semi-finished bread
A0002	coffee
A0002	tropical fruit

Transaction	citrus fruit	margarine	ready soups	semi-finished bread	coffee	tropical fruit
A0001	1	1	1	1	0	0
A0002	0	0	0	0	1	1
A0003	0	0	0	0	0	0
A0004	0	0	0	0	0	0

Asociación

aplicación: análisis de la cesta de la compra

Ejemplo: Compra de productos de alimentación

- Hay que establecer los **parámetros** para la detección de patrones.
  - **Soporte mínimo:** 0.001 (muchos recibos y productos)
  - **Confianza mínima:** 0.70.
  - **Longitud de la regla:** máximo 3 elementos (máximo 2 en el antecedente)

Rules	Support	Confidence	Lift
{liquor, red/blush wine} => {bottled beer}	0.002	0.90	11.24
{cereals, yogurt} => {whole milk}	0.002	0.81	3.17
{butter, jam} => {whole milk}	0.001	0.83	3.26
{chocolate, pickled vegetables} => {whole milk}	0.001	0.86	3.35
{grapes, onions} => {other vegetables}	0.001	0.92	4.74
{hard cheese, oil} => {other vegetables}	0.001	0.92	4.74

Asociación

`aplicación: análisis de la cesta de la compra`

Ejemplo: Compra de productos de alimentación

- **Ajuste** de los parámetros del algoritmo:
  - La distribución de artículos por transacción será distinta de unos negocios a otros, siendo necesarios valores distintos de soporte y confianza.
  - Para determinar qué funciona mejor hay que experimentar con los parámetros:
    - Umbrales bajos —> más reglas: más difícil identificar las que tienen mayor impacto.
  - No hay certeza: una opción es comenzar experimentando con valores bajos de los parámetros e irlos subiendo en función de los resultados.

Asociación

aplicación: análisis de la cesta de la compra

Ejemplo: Compra de productos de alimentación

- Para aprovechar este conocimiento adicional, nos centramos en 3 artículos específicos: “Yogurt”, “Tropical Fruit” y “Bottled Beer”.
  - Para ello usamos los mismos umbrales y especificamos que aparezcan estos **términos en el consecuente**:

Rules	Support	Confidence	Lift
{pip fruit, sausage, sliced cheese} => {yogurt}	0.001	0.86	6.14
{butter, cream cheese , root vegetables} => {yogurt}	0.001	0.91	6.52
{butter, margarine, tropical fruit} => {yogurt}	0.001	0.85	6.07
{butter, curd, other vegetables, tropical fruit} => {yogurt}	0.001	0.83	5.97
{liquor, red/blush wine} => {bottled beer}	0.002	0.90	11.24
{citrus fruit, fruit/vegetable juice, grapes} => {tropical fruit}	0.001	0.85	8.06
{ham, other vegetables, pip fruit, yogurt} => {tropical fruit}	0.001	0.83	7.94

Asociación

aplicación: análisis de la cesta de la compra

Ejemplo: Compra de productos de alimentación

- También podemos utilizar estos **términos como antecedente** en la generación del conjunto de reglas:

Rules	Support	Confidence	Lift
{yogurt} => {whole milk}	0.056	0.40	1.57
{tropical fruit} => {other vegetables}	0.036	0.34	1.77
{yogurt} => {rolls/buns}	0.034	0.25	1.34
{tropical fruit} => {rolls/buns}	0.025	0.23	1.27
{bottled beer} => {soda}	0.017	0.21	1.21
{bottled beer} => {bottled water}	0.016	0.20	1.77
{tropical fruit} => {pip fruit}	0.020	0.19	2.57
{tropical fruit} => {citrus fruit}	0.020	0.19	2.29

Asociación

`aplicación: análisis de la cesta de la compra`

Uso del análisis para la toma de decisiones de negocio

- Antes de usar los datos para tomar ninguna decisión de negocio, es importante dar un paso atrás y recordar que:
  - La salida del análisis refleja la frecuencia de **co-ocurrencia** de los artículos en las transacciones. Esto es función tanto de la fortaleza de la asociación entre artículos como de la forma en que el negocio los ha presentado al cliente.
    - Es decir: los artículos pueden aparecer juntos no porque estén naturalmente conectados, sino porque los encargados de la organización los han presentado juntos.

Asociación

`aplicación: análisis de la cesta de la compra`

Uso del análisis para la toma de decisiones de negocio

- Los resultados del análisis de la cesta de la compra se pueden usar para realizar campañas de **marketing dirigido**:
  - Por ejemplo, para cada regla, elegimos algunos de los productos comprados con márgenes y retornos altos, y mandamos correos personalizados a los clientes.
  - La forma de usar el análisis es importante para el propio análisis:
    - Para alimentar a un sistema automático de recomendaciones, nos interesará obtener un conjunto amplio de reglas.
    - Para experimentar, tiene más sentido obtener pocas reglas de gran valor, y actuar sólo sobre ellas.

Asociación

reglas de asociación en R

Paquetes en R para descubrimiento de reglas de asociación

Paquete	Descripción
arules	algoritmos apriori y eclat
arulesViz	visualizaciones para exploración de reglas
rCBA	algoritmo fpgrowth

Asociación

referencias bibliográficas

- Adamo, J.M. Data Mining For Association Rules and Sequential Patterns Sequential and Parallel Algorithms. Springer. 2001
- Agrawal, R; Srikant, R. Fast Algorithms for Mining Association Rules - Proc. of the 20th International Conference on Very Large Databases, Santiago, Chile, Sept. 1994.
- Han, J.; Pei, J.; Yin, Y. Mining frequent patterns without candidate generation. In Proc. 2000 ACMSIGMOD Int. Conf. Management of Data (SIGMOD'00), 1–12. 2000.
- Julander, C. Basket Analysis: A New Way of Analysing Scanner Data. International Journal of Retail & Distribution Management 20 (7), 10-18, 1992.
- Pei, J; Han, J; Mao, R. CLOSET: An efficient algorithm for mining frequent closed itemsets. In DMKD 2000, pp. 11—20, 2000
- Silberschatz, A; Tuzhilin, A. On subjective measures of interestingness in knowledge Discovery. Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 275–281. 1995.
- Webb, G.I. OPUS: An efficient admissible algorithm for unordered search. Journal of Artificial Intelligence Research, 3:45--83, 1996.
- Zaki, M.J; Hsiao, C.J. CHARM: An Efficient Algorithm for Closed Itemset Mining. 2002
- Zhang, C; Zhang, S. Association rule mining: models and algorithms. Springer. 2002.

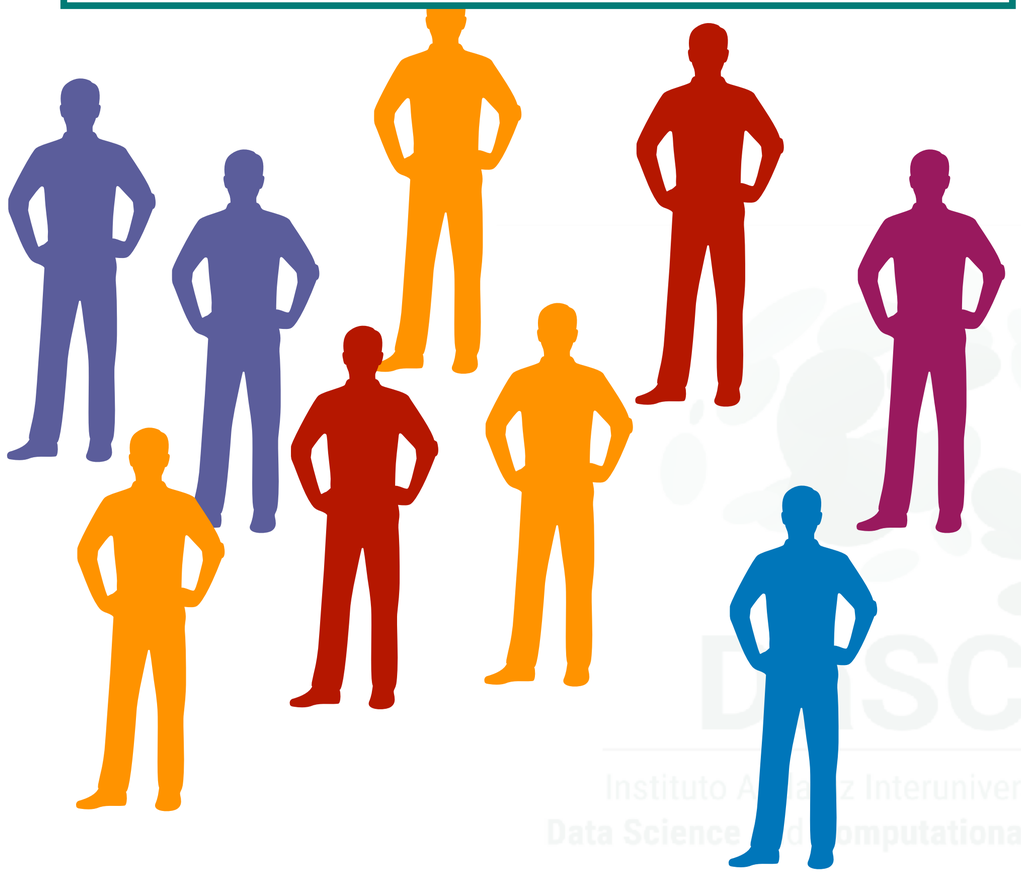
Agrupamiento

Agrupamiento

motivación

Crear grupos de instancias o ejemplos de similares características

Segmentación de clientes de una empresa



Agrupamiento

¿qué es un cluster? ¿qué es clustering?

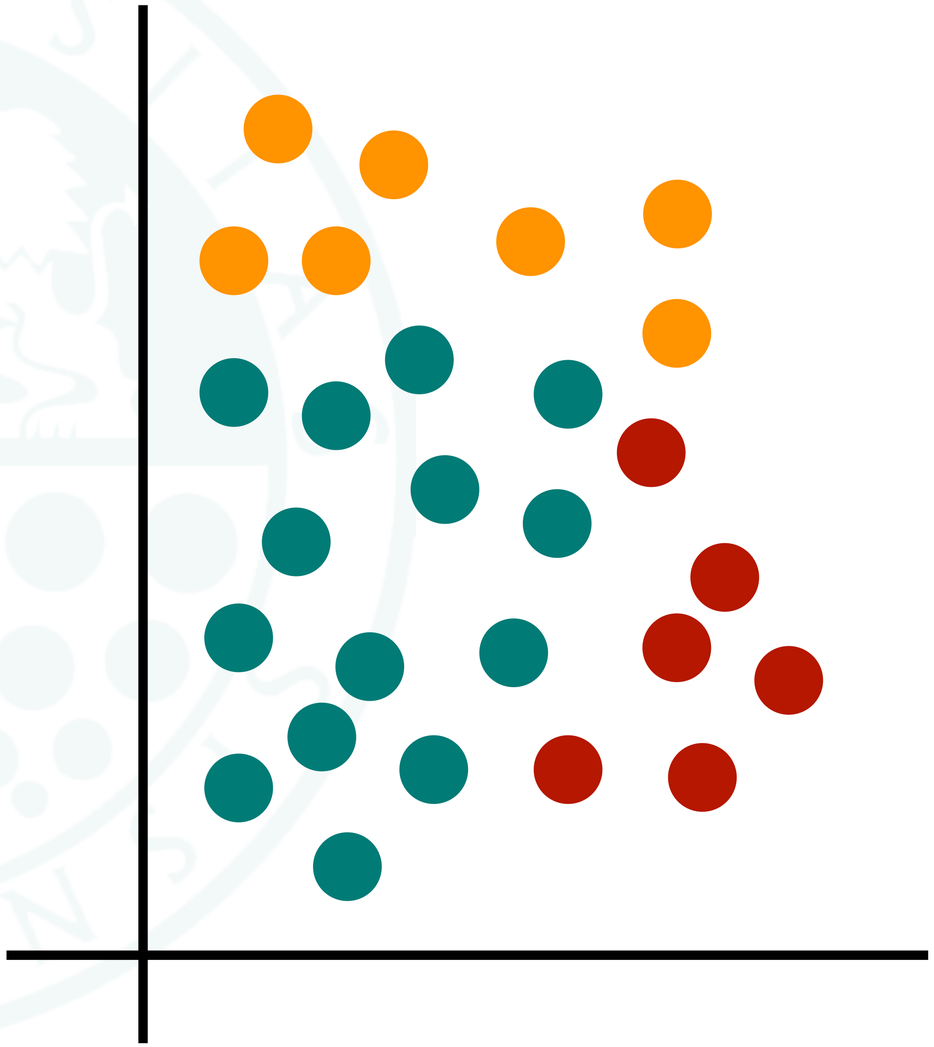
Por definición, **un *cluster* es un grupo o conjunto de objetos**.

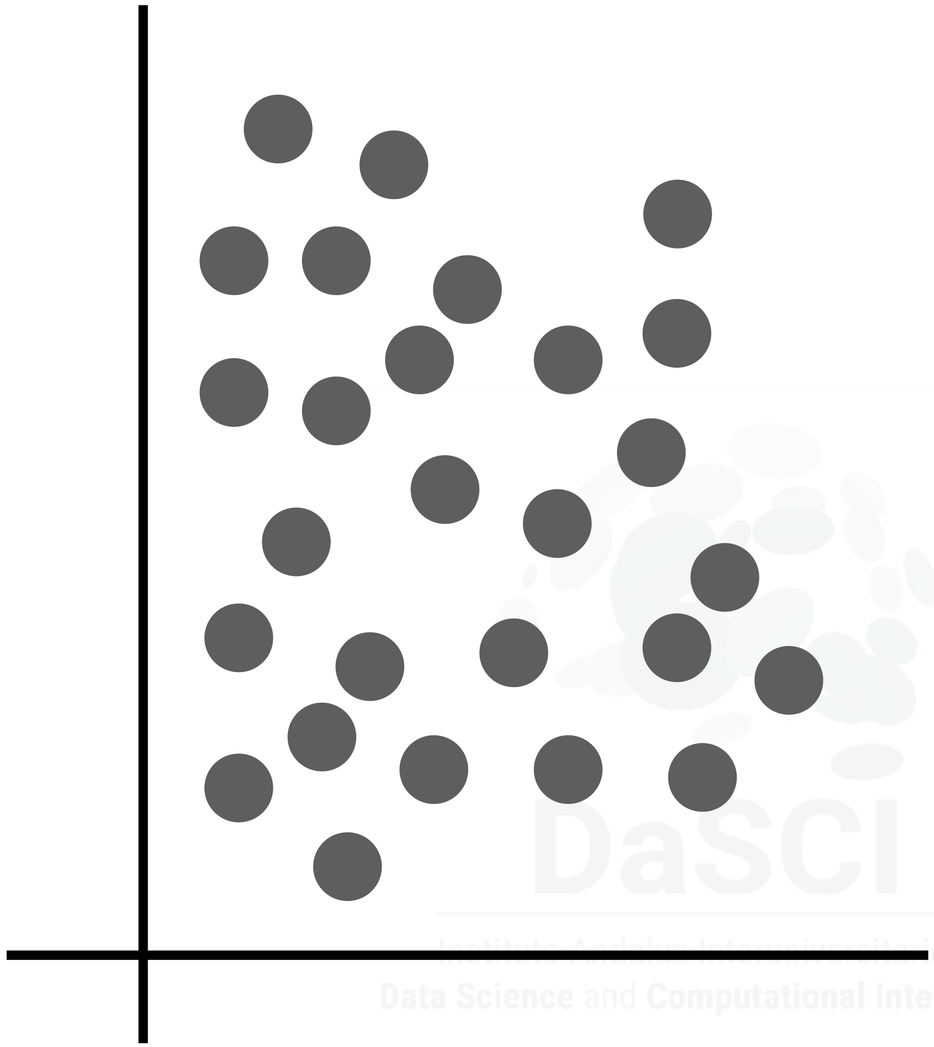
- Similares a cualquier otro incluido en el mismo *cluster*.
- Distintos a los objetos incluidos en otros grupos.

Clustering es la técnica de segmentar o agrupar una población heterogénea en un número de subgrupos homogéneos o clusters.

Agrupamiento

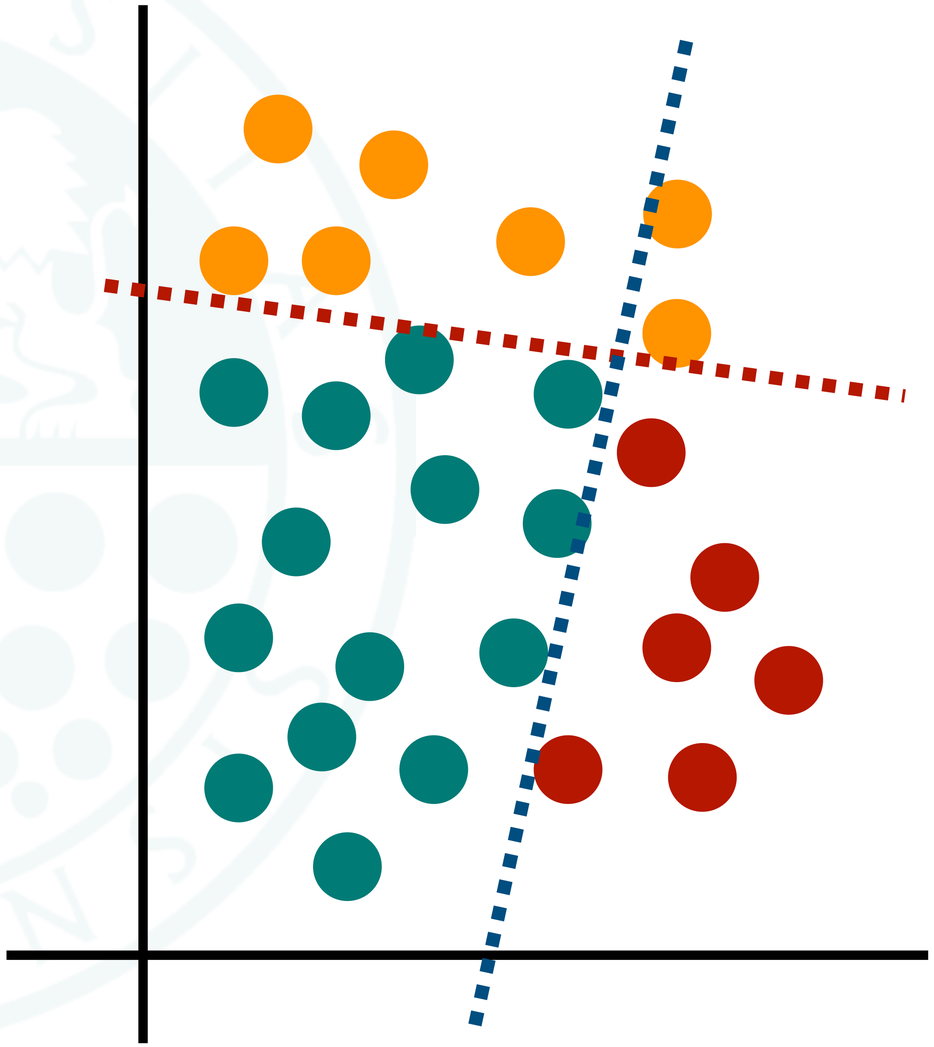
clasificación versus clustering

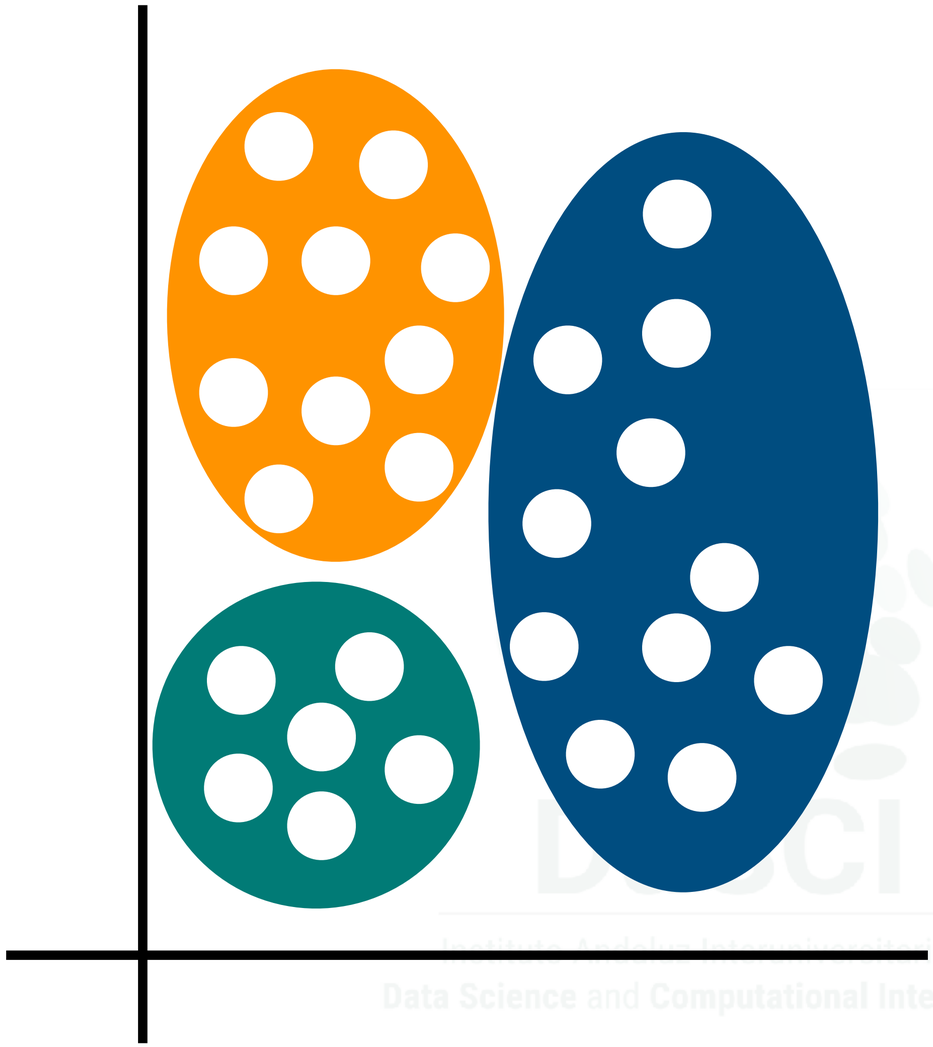




Agrupamiento

clasificación versus clustering





Agrupamiento

aplicaciones típicas

- Para tareas de preprocesamiento antes de aplicar otra técnica de descubrimiento del conocimiento

- Técnica de descubrimiento del conocimiento para obtener información acerca de la distribución de los datos

Agrupamiento

`problemas reales`

Datos o instancias con valores atípicos respecto del conjunto completo

- Dificultad en los manejos de outliers
- En bases de datos dinámicas implica que la pertenencia a clusters varía en el tiempo
- Interpretar el significado de cada cluster puede ser difícil
- **NO tenemos una única solución para un problema. El número de clusters es difícil de determinar**

Agrupamiento

aplicaciones

Segmentación de clientes de una empresa para usos publicitarios

Agrupamiento

aplicaciones

Identificación de áreas con usos similares a partir de observaciones en cultivos

Agrupamiento

aplicaciones

Agrupar asegurados con similares características para ofertar otros productos que estos clientes ya tienen contratados

Santander

Agrupamiento

aplicaciones

Grupos de viviendas de acuerdo a su tipo, valor o situación geográfica

Agrupamiento

aplicaciones

Agrupar documentos, analizar ficheros .log, patrones similares de comportamiento en clientes

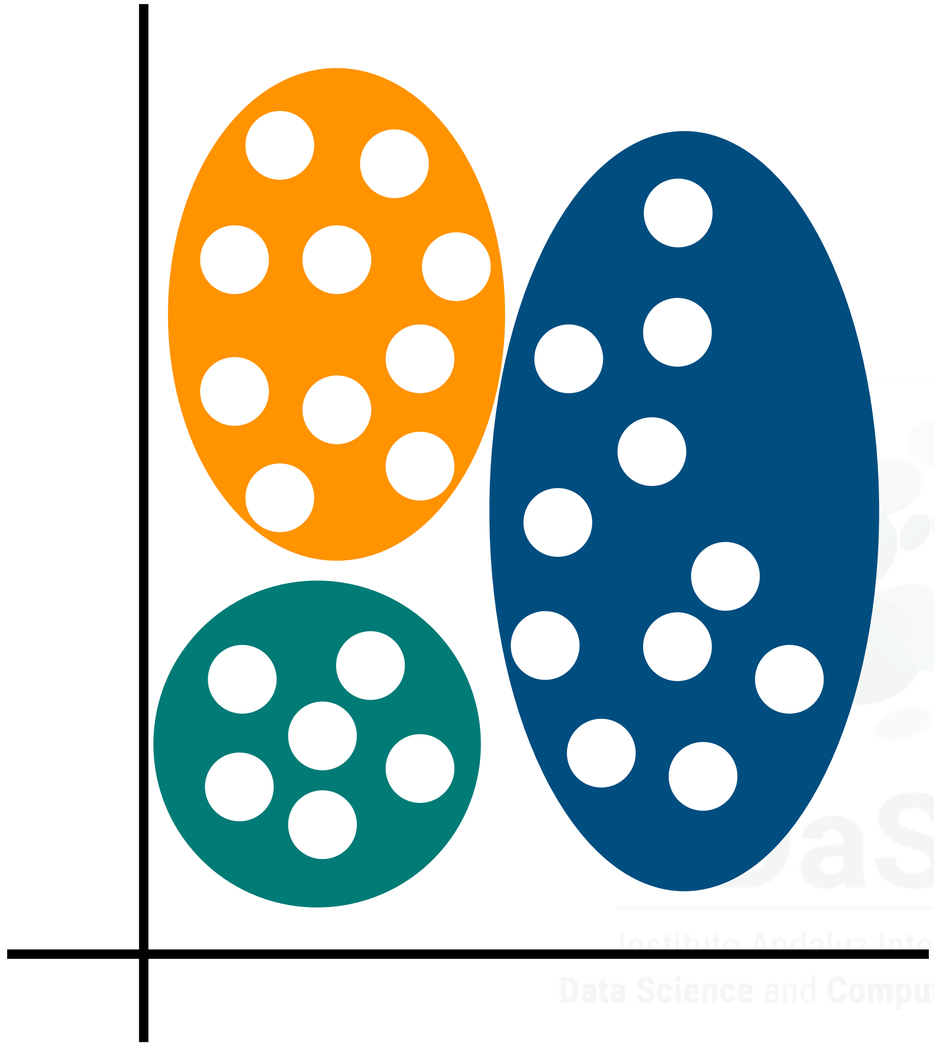
Agrupamiento

bondad de un método de clustering

Dos conceptos clave

MINIMIZAR la similaridad intra-cluster

MAXIMIZAR la similaridad inter-cluster



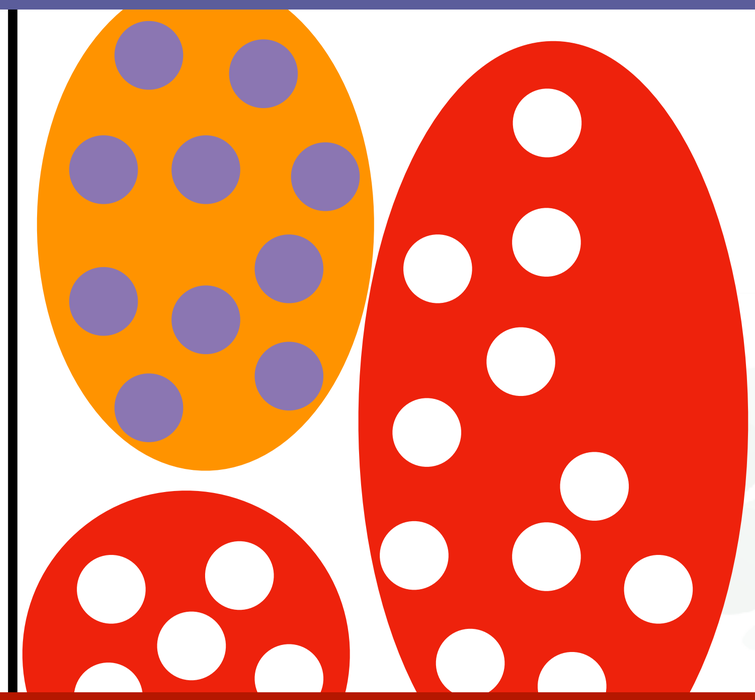
Agrupamiento

bondad de un método de clusteri

Dos conceptos clave

MINIMIZAR la similaridad intra-cluster

Los elementos de un mismo cluster deben ser lo más similares posibles



MAXIMIZAR la similaridad inter-cluster

Los elementos de distintos clusters deben ser lo menos similares posibles

Agrupamiento

`bondad de un método de clustering`

Las medidas de similaridad/disimilaridad se asocian normalmente a una medida de distancia:

- Son funciones muy sensibles al tipo de variable usadas (intervalos, binarias, booleanas, categóricas, ordinales)
- Posible asignar peso a variables por criterios del experto o la propia aplicación
- Complicado de dar definiciones para términos de "suficientemente similar"

Agrupamiento

subjetividad

¿Cómo agruparíais a este grupo de personas?

Agrupamiento

subjetividad

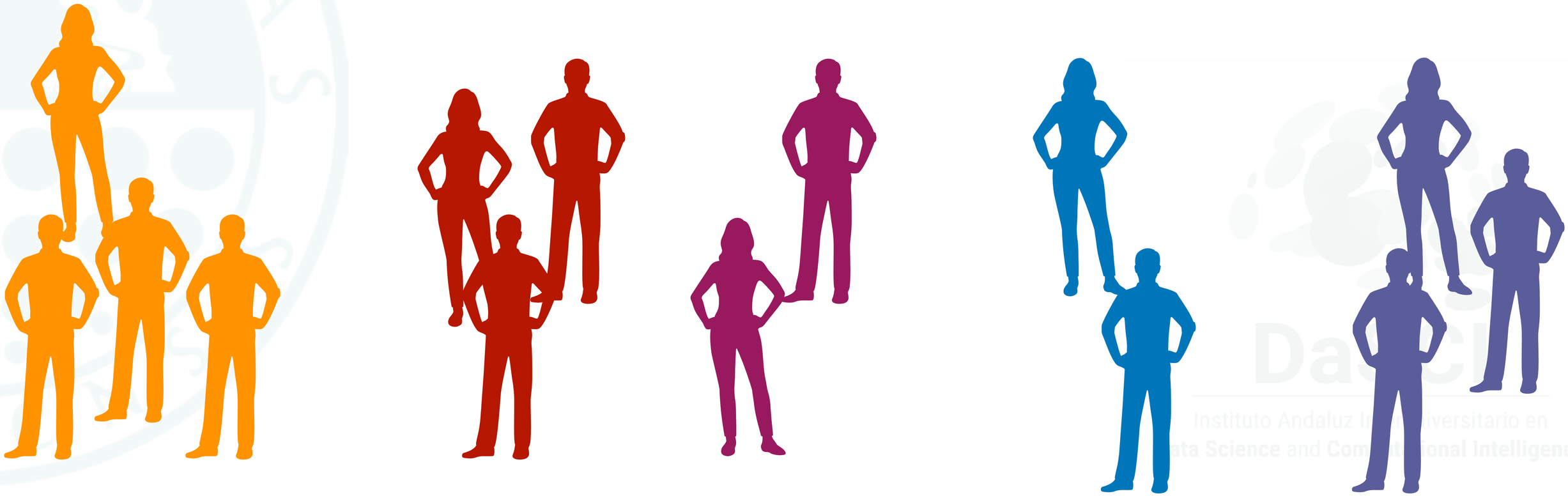
¿Cómo agruparíais a este grupo de personas?
Hombres/Mujeres

Agrupamiento

subjetividad

¿Cómo agruparíais a este grupo de personas?

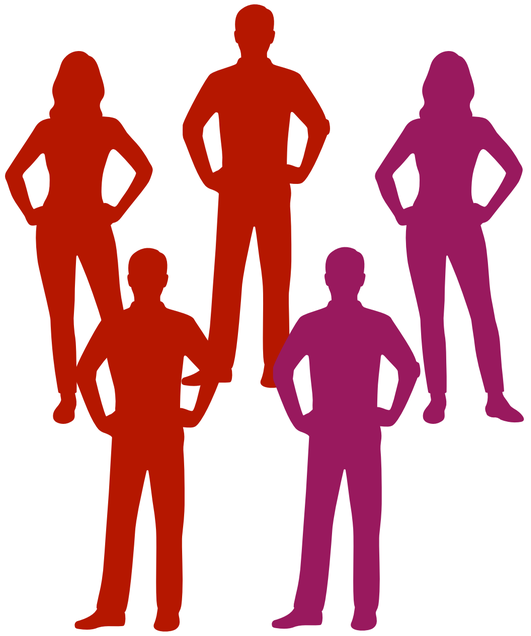
Raza

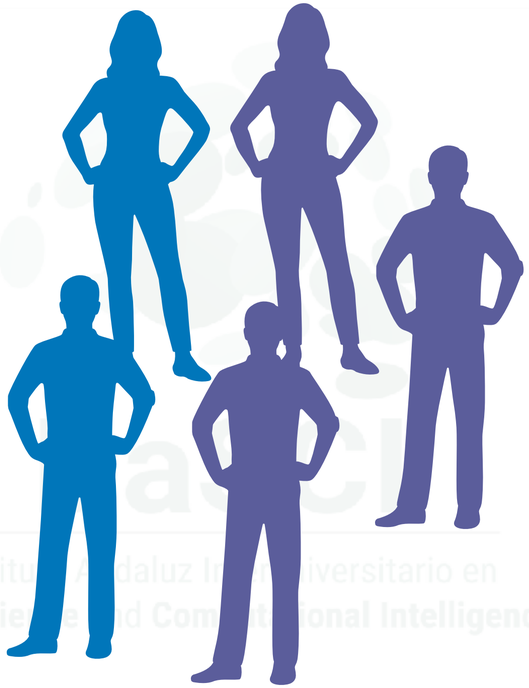


Agrupamiento

subjetividad

¿Cómo agruparíais a este grupo de personas?
Por número de clusters, por ejemplo, TRES





Agrupamiento

`propiedades deseables`

insensible al orden de registros de entrada

escalables

descubrir clusters formas arbitrarias

tratar distintos tipos de variables

Incorporar restricciones del usuario

ruido y outliers

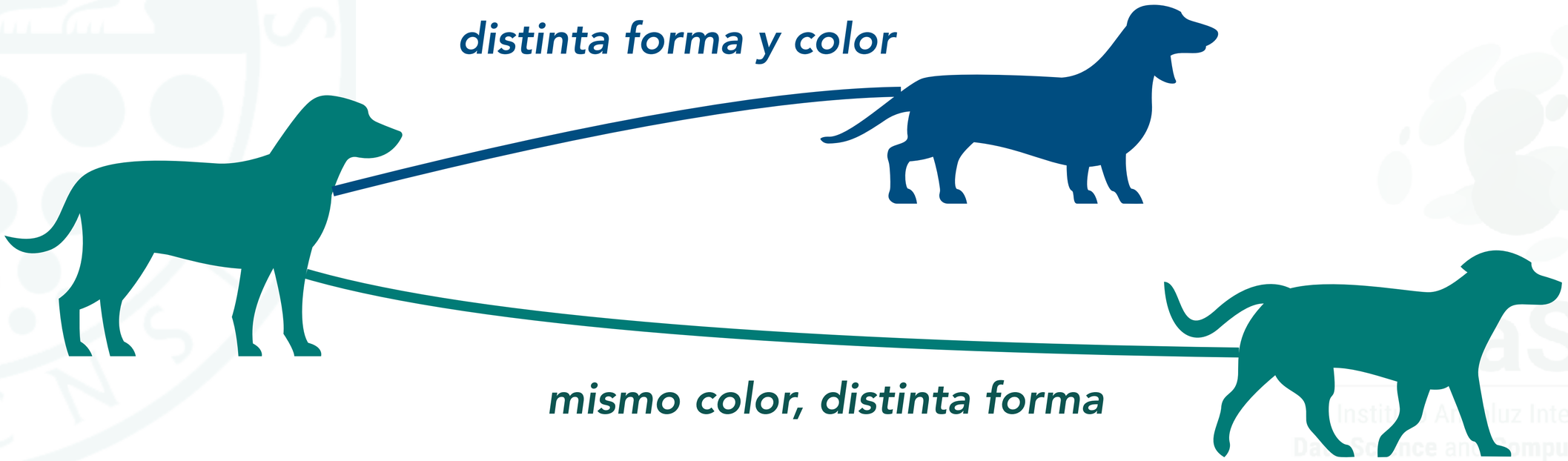
alta dimensionalidad

resultados interpretables

requisitos mínimos del problema

Medidas de distancia y similitud

¿Qué es la similitud? Se define por la RAE como semejanza (cualidad de semejante (que semeja o se parece a alguien o algo (tener determinada apariencia o aspecto)))



Medidas de distancia y similitud

¿Qué es la similitud? Se define por la RAE como semejanza (cualidad de semejante (que semeja o se parece a alguien o algo (tener determinada apariencia o aspecto)))

Medidas de distancia y similitud

id	sexo	nacimiento	categoria	salario	experiencia	minoría
25	Mujer	6-10-36	Administ	18750	54	No
26	Mujer	26-9-65	Administ	38550	22	No
27	Hombre	6-10-60	Administ	27450	173	Sí
28	Hombre	21-1-51	Seguridad	24300	191	Sí
29	Hombre	1-9-50	Seguridad	30750	209	Sí
30	Mujer	25-7-46	Directivo	68750	38	No
31	Hombre	18-7-59	Administ	19650	229	Sí
32	Hombre	6-9-58	Directivo	59375	6	No

Medidas de distancia y similitud

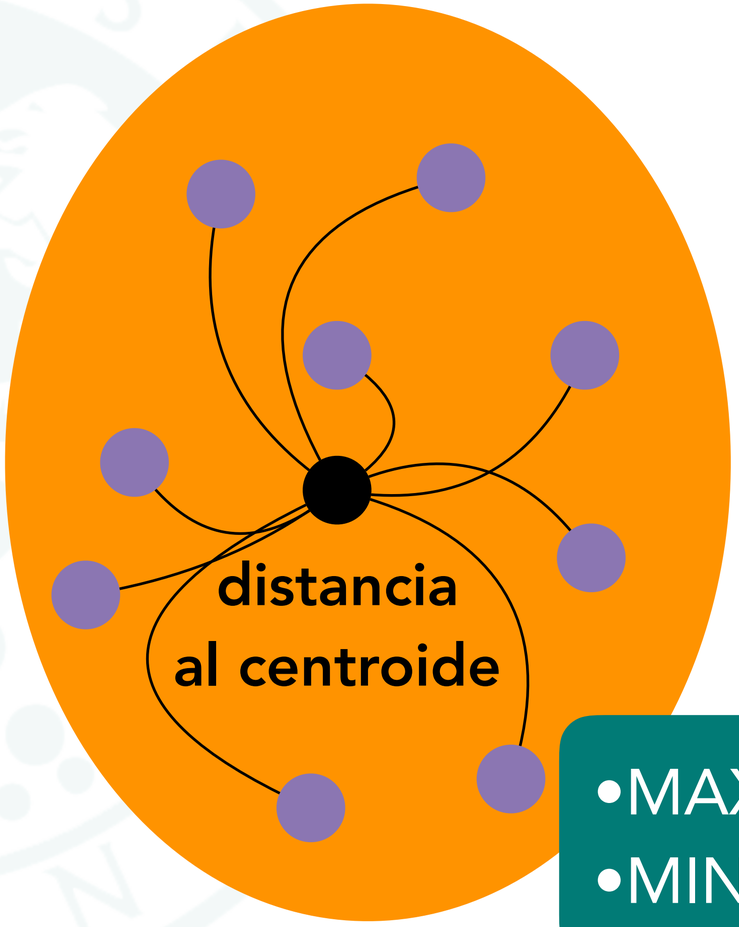
id	sexo	nacimiento	categoria	salario	experiencia	minoría
25	Mujer	6-10-36	Administ	18750	54	No
26	Mujer	26-9-65	Administ	38550	22	No
27	Hombre	6-10-60	Administ	27450	173	Sí
28	Hombre	21-1-51	Seguridad	24300	191	Sí
29	Hombre	1-9-50	Seguridad	30750	209	Sí
30	Mujer	25-7-46	Directivo	68750	38	No
31	Hombre	18-7-59	Administ	19650	229	Sí
32	Hombre	6-9-58	Directivo	59375	6	No

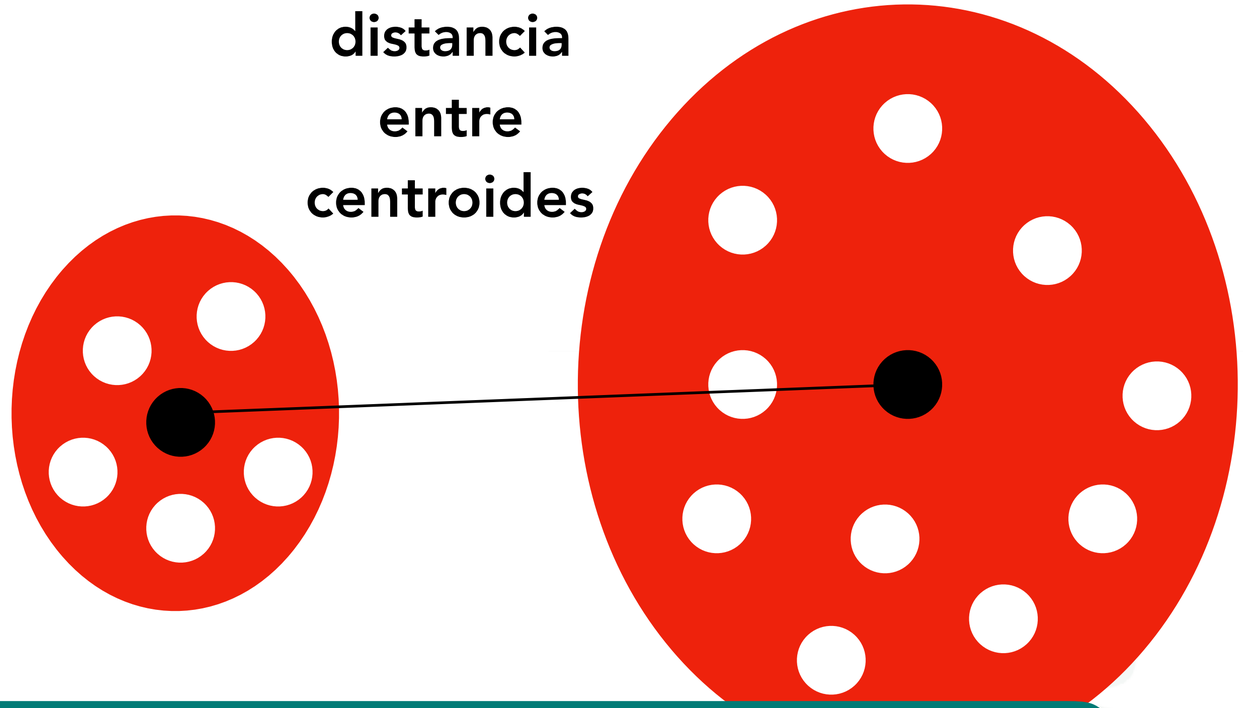
Medidas de distancia y similitud

id	sexo	racimiento	categoria	salario	experiencia	minoría
25	Mujer	6-10-36	Administ	18750	54	No
26	Mujer	26-9-65	Administ	38550	22	No
27	Hombre	6-10-60	Administ	27450	173	Sí
31	Hombre	18-7-59	Administ	19650	229	Sí

- Dentro de un mismo grupo tendríamos variables que nos podrían dividir el grupo en otros subgrupos o categorías
- No es necesario utilizar todos los atributos para agrupar objetos o elementos

Medidas de distancia y similitud





- MAXIMIZAR la distancia entre clústers
- MINIMIZAR la distancia entre elementos de un clúster

Medidas de distancia y similitud

La definición de la medida de distancia suele depender del tipo de variable:

- Variables intervalares
- Variables binarias o booleanas (verdadero/falso)
- Variables nominales o categóricas
- Variables ordinales

Medidas de distancia y similitud

`variables numéricas`

Son las más sencillas de calcular

Distancia (X, Y) = A(X) - A(Y)

Cuando tenemos n dimensiones se utiliza la famosa

distancia Euclidea

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_i-q_i)^2 + \cdots + (p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

Medidas de distancia y similitud

`variables nominales`

Son también muy sencillas de calcular porque la distancia se fija a 1 si los valores son diferentes, y a 0 si son iguales

Una variación sería ponderar las variables según su importancia en el problema

$$d(i,j) > d(i,k)$$

Medidas de distancia y similitud

`variables nominales`

Son también muy sencillas de calcular porque la distancia se fija a 1 si los valores son diferentes, y a 0 si son iguales

Una variación sería ponderar las variables según su importancia en el problema

$$d(i,j) > d(i,k)$$

el elemento i es más parecido a k que al elemento j

Medidas de distancia y similitud

`variables continuas`

Para evitar que unas variables dominen sobre otras, los valores de los atributos se "normalizan" a priori:

- Desviación absoluta media

$$s_f = \frac{1}{n}(|(x_{1f} - m_f| + \ldots + |(x_{nf} - m_f|) \qquad m_f = \frac{1}{n}(x_{1f} + \ldots + x_{nf})$$

- z-score (medida estandarizada)

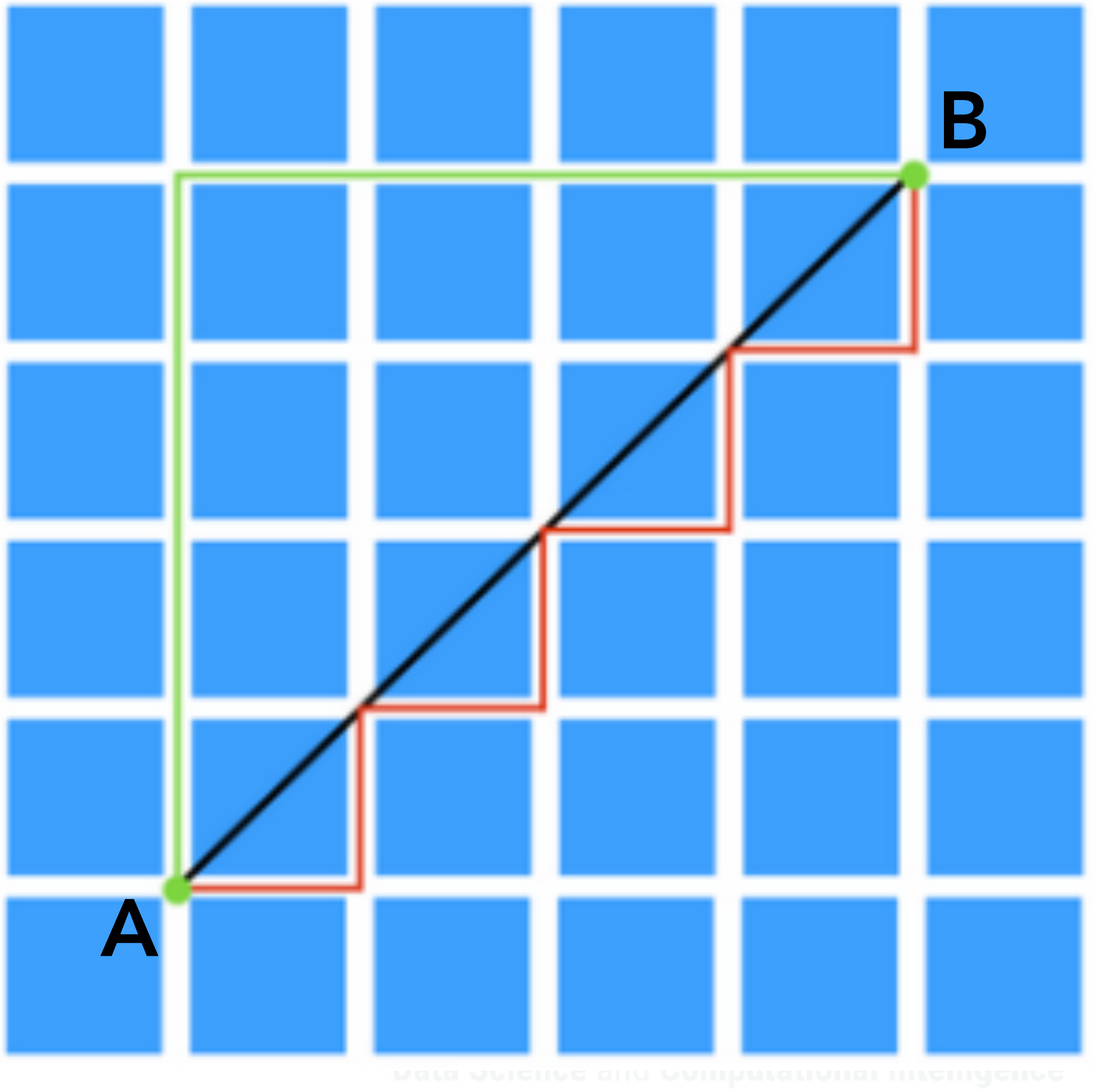
$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Medidas de distancia y similitud

`distancia de Minkowski`

- La distancia entre dos puntos (AB) se calcula mediante la distancia **Euclidea** (línea negra)

- Si consideramos un mapa cuadriculado como el de las calles de la isla de Manhattan hace que el camino más corto posible en taxi sea exactamente la distancia de **Manhattan** (línea roja y verde)



Medidas de distancia y similitud

distancia de Minkowski

$$d_r(x,y) = \left(\sum_{j=1}^{J} |x_j - y_j|^r\right)^{\frac{1}{r}}$$

- Si r=1 es Distancia de Manhattan

$$d_1(x,y) = \sum_{j=1}^{J} |x_j - y_j|$$

- Si r=2 es Distancia Euclidea

$$d_2(x,y) = \sqrt{\sum_{j=1}^{J} |x_j - y_j|^2}$$

- Si r=∞ es Distancia de Chebyshev

Medidas de distancia y similitud

`otras distancias`

Distancia de Mahalanobis

- Correlación entre variables y no depende de la escala de medida

Distancia de Levenshtein

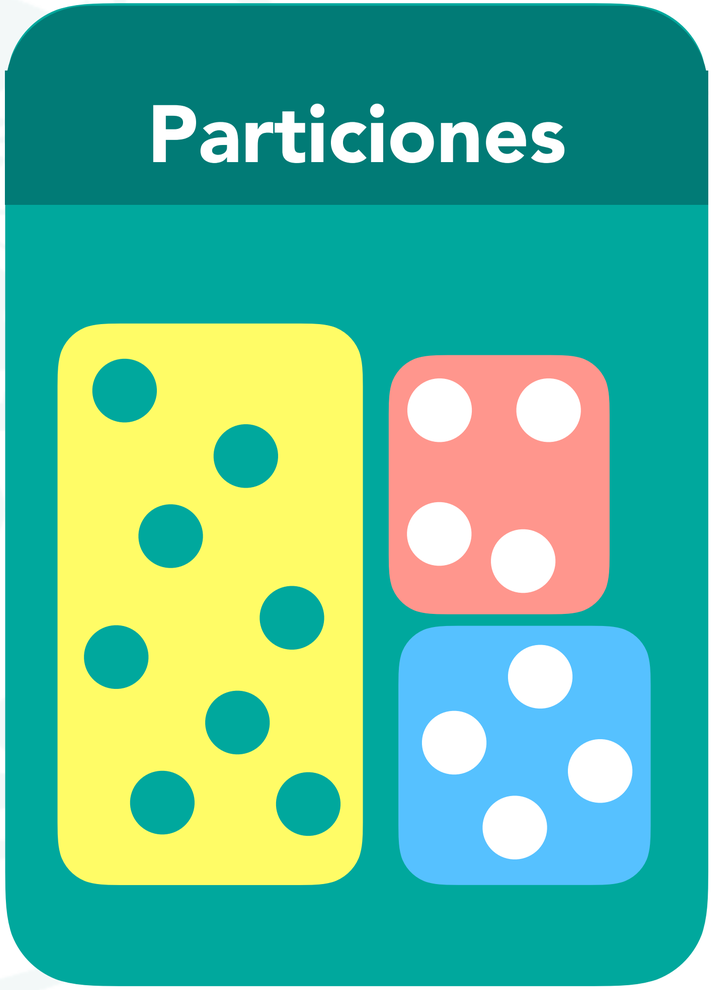
- Empleada en correctores ortográficos, reconocimiento de voz, etc.

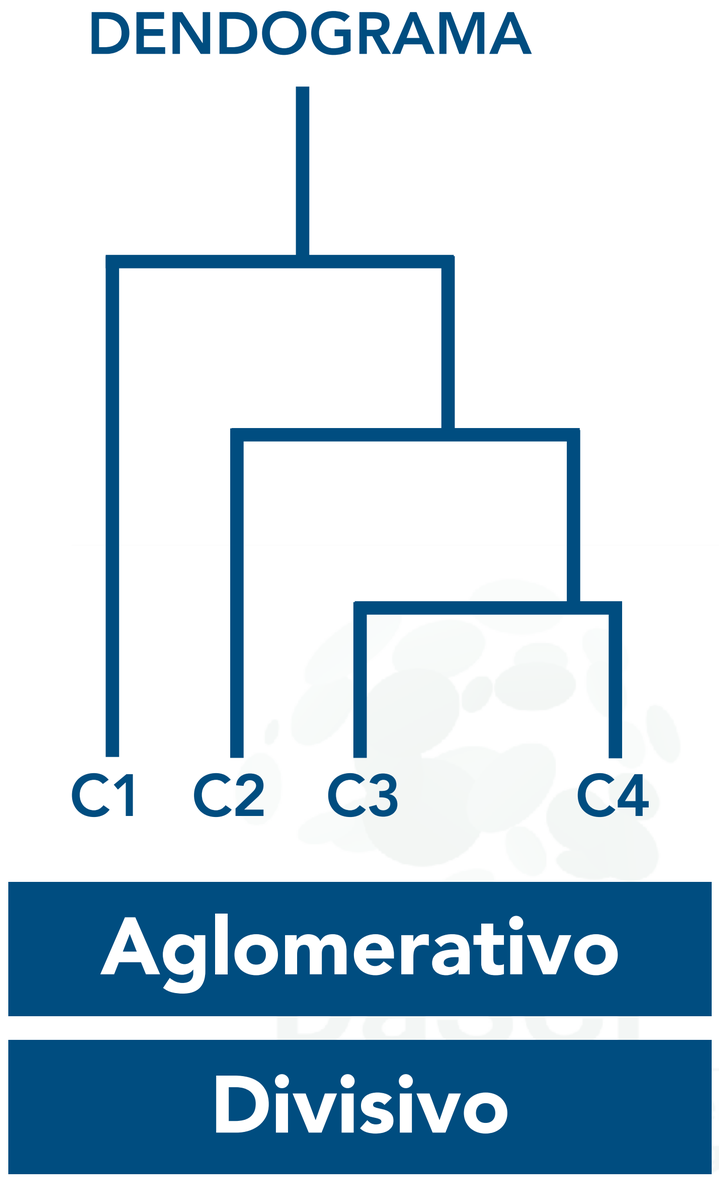
Modelos basados en teorías de conjuntos

Distancia de vecinos compartidos

Medidas de correlación

Clasificación de algoritmos de clustering





Clasificación de algoritmos de clustering

Exclusiva versus no exclusiva

- No exclusivos los elementos pueden pertenecer a varios grupos, y pueden representar múltiples “fronteras”

Difuso versus no difuso

- A un elemento se le asocia un peso entre 0 y 1, y todos los pesos del cluster deben sumar 1.
- El agrupamiento probabilístico tiene características similares

Parcial versus completa

Heterogéneo versus homogéneo

Método basado en particionamiento clásico

k-Means

Se basa en la idea de ir moviendo entre clusters hasta que se alcanza el número de clusters deseado

Cada cluster se representa por el centro del mismo, también conocido como CENTROIDE

Parámetro de entrada (k) que es el número de clusters

OBJETIVO: Minimizar distancia euclidea total entre cada punto y su representante de cluster más cercano de forma iterativa

Método basado en particionamiento clásico

Algorithm 2.1 The `k-means` algorithm

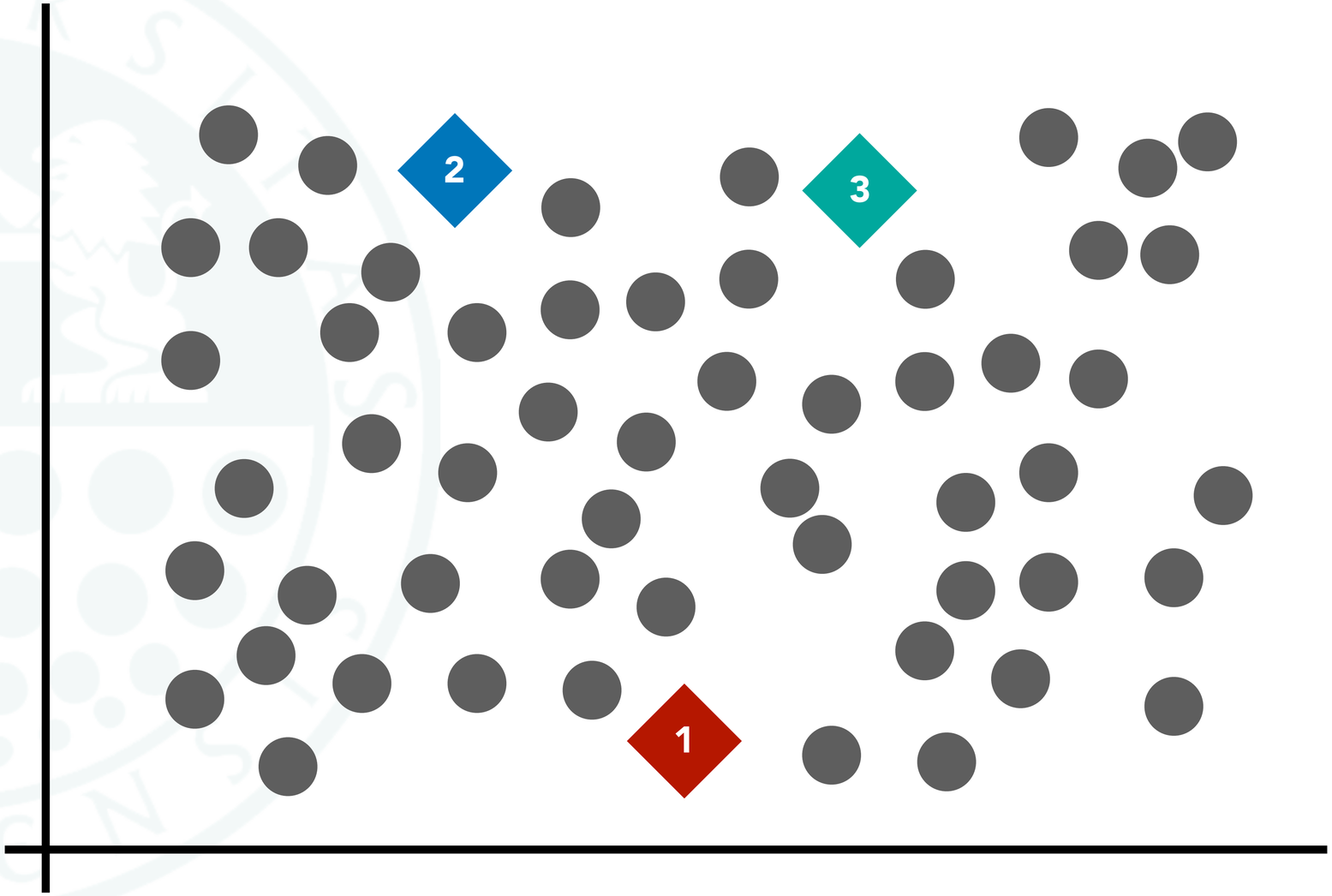
Input: Dataset D, number clusters k
Output: Set of cluster representatives C, cluster membership vector $\mathbf{m}$

```
    /* Initialize cluster representatives C */
    Randomly choose k data points from D
 5: Use these k points as initial set of cluster representatives C
    repeat
        /* Data Assignment */
        Reassign points in D to closest cluster mean
        Update m such that m_i is cluster ID of ith point in D
10:     /* Relocation of means */
        Update C such that c_j is mean of points in jth cluster
    until convergence of objective function sum_{i=1}^{N} (argmin_j ||x_i - c_j||_2^2)
```

k-Means

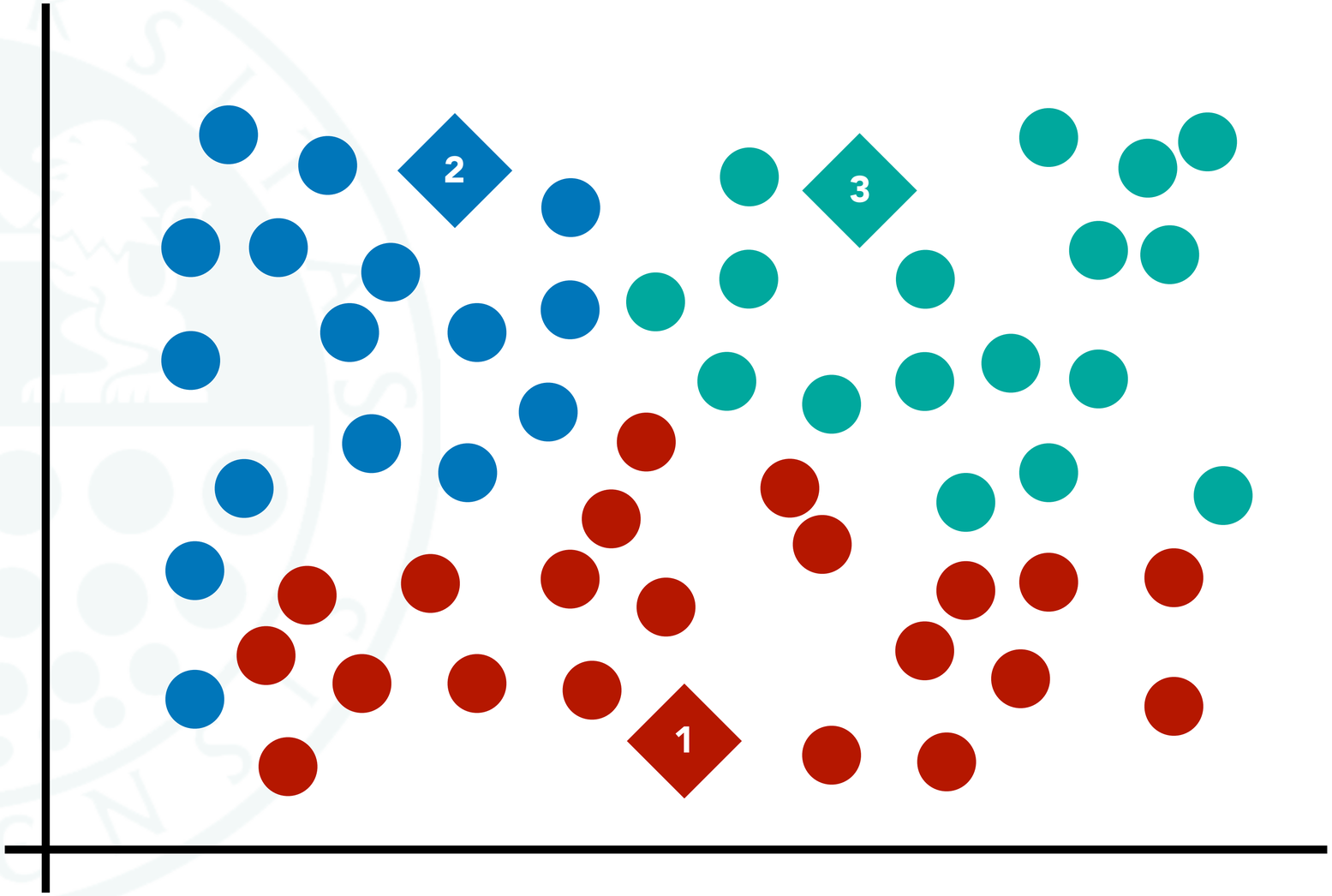
Método basado en particionamiento clásico

k-Means



Selección aleatoria de k=3 centroides

Método basado en particionamiento clásico

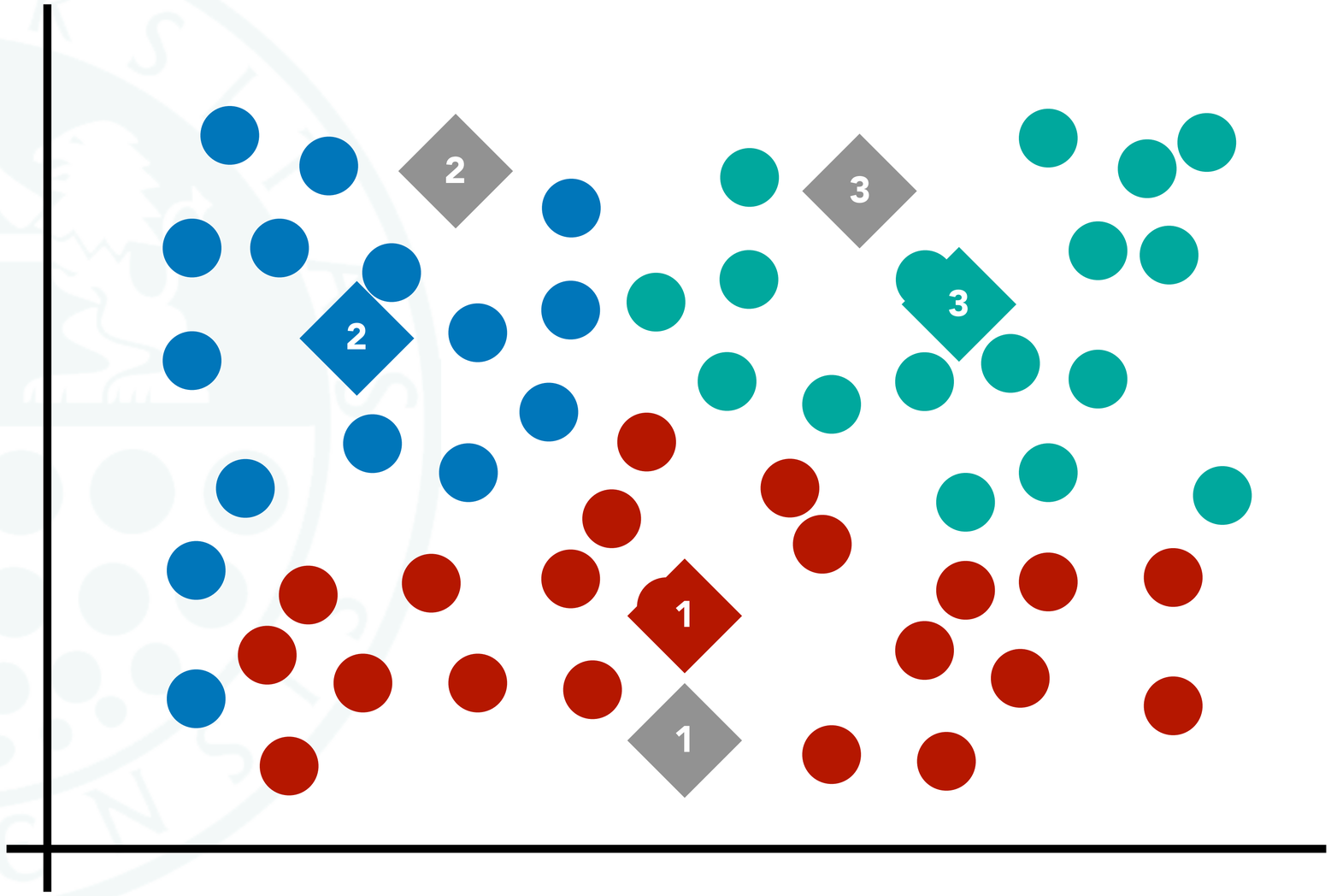


k-Means

Selección aleatoria de k=3 centroides

Asignar puntos a cada centroide

Método basado en particionamiento clásico



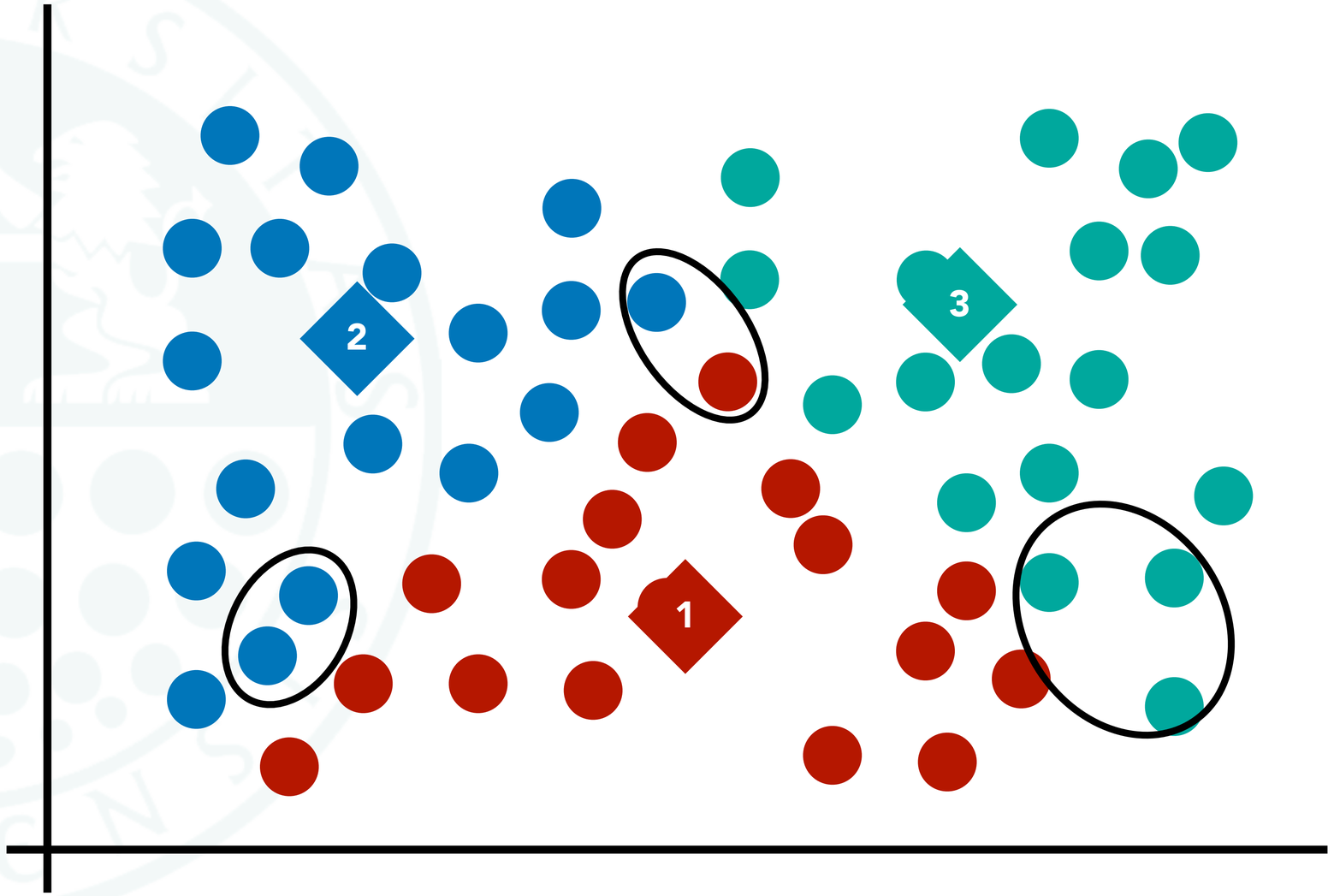
k-Means

Selección aleatoria de k=3 centroides

Asignar puntos a cada centroide

Re-colocación de medias

Método basado en particionamiento clásico



k-Means

Selección aleatoria de k=3 centroides

Asignar puntos a cada centroide

Re-colocación de medias

0 asignaciones

Método basado en particionamiento clásico

k-Means

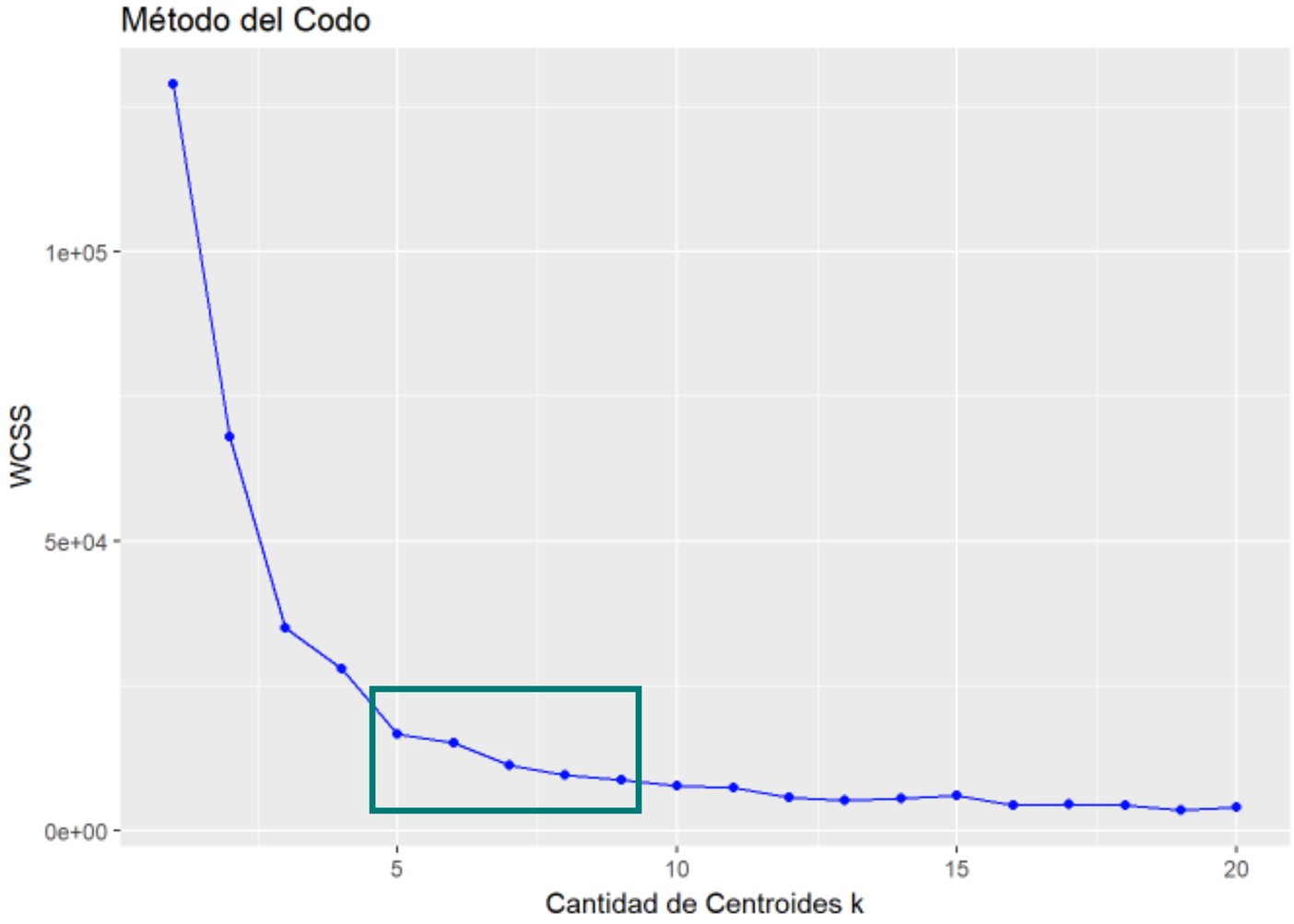
VENTAJAS

- Relativamente eficiente O(tkn) — t son iteraciones, k es clusters y n objetos
- Finaliza de forma frecuente en un óptimo local

DESVENTAJAS

- Aplicable si se puede definir el concepto de media
- Necesidad de fijar clusters
- Débil en ruido/outliers
- Sólo indicado en clusters convexos

Método basado en particionamiento clásico

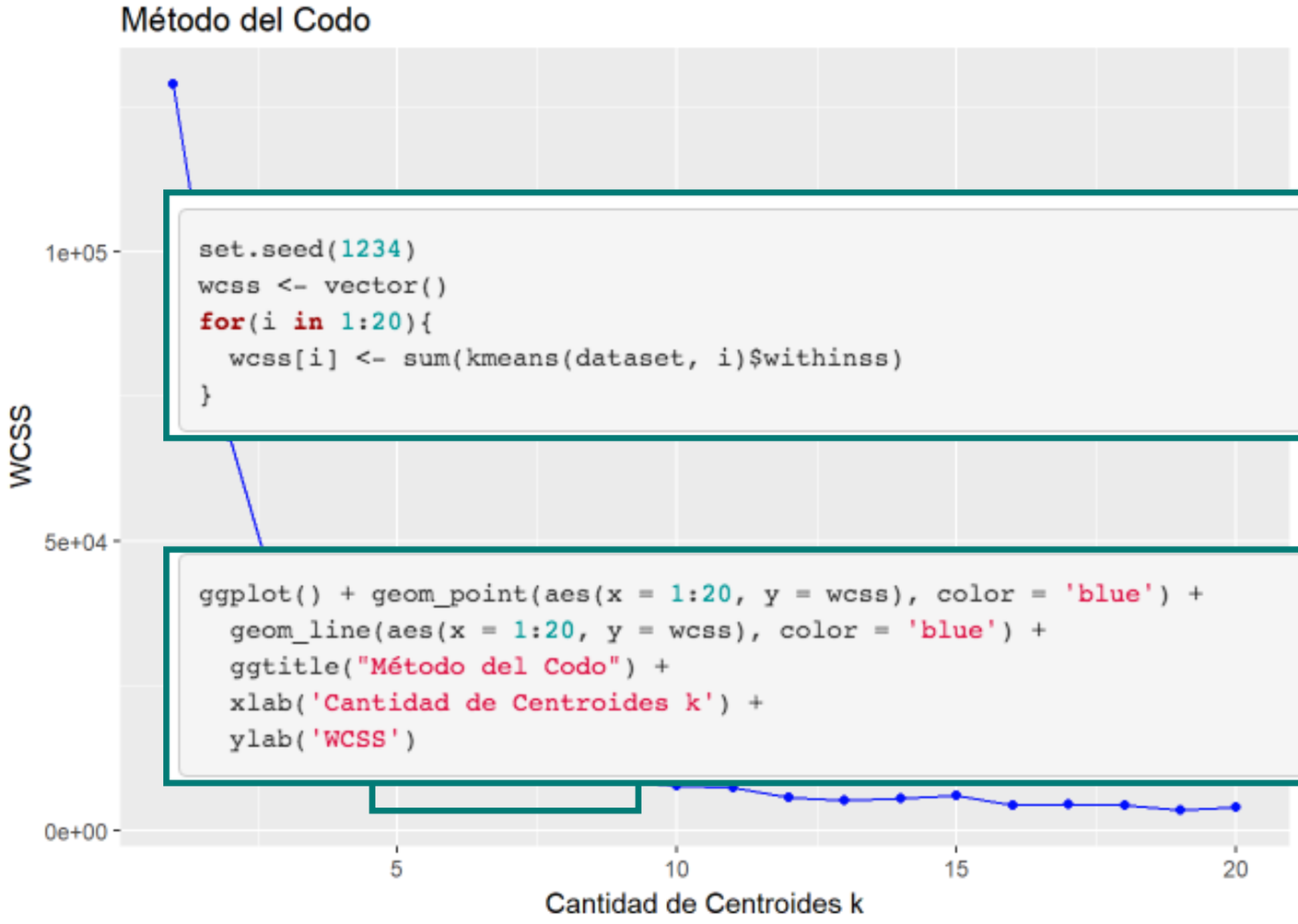


k-Means
Método del codo

- Al aumentar los centroides, disminuye el WCSS
- Se genera una forma de codo.
- El valor óptimo de k se escoge donde ya no se dejan de producir variaciones importantes del valor de WCSS (Within Clusters Summed Squares).
- A partir de k >= 5, por lo que sería interesante evaluar los resultados con valores como 5, 6 y 7.

Método basado en particionamiento clásico

k-Means
Método del codo




```
set.seed(1234)
wcss <- vector()
for(i in 1:20){
  wcss[i] <- sum(kmeans(dataset, i)$withinss)
}
```

```
ggplot() + geom_point(aes(x = 1:20, y = wcss), color = 'blue') +
  geom_line(aes(x = 1:20, y = wcss), color = 'blue') +
  ggtitle("Método del Codo") +
  xlab('Cantidad de Centroides k') +
  ylab('WCSS')
```

- Al aumentar los centroides,
- A partir de k >= 5, por lo que sería interesante evaluar los resultados con valores como 5, 6 y 7.

Clustering en R

paquetes más citados

Nombre	Citas
apcluster	3682
akmedoids	1220
clustMixType	925
mclust	885
conclust	532
EMMIXgene	338

View: Cluster

https://cran.r-project.org/web/views/Cluster.html

Listado de paquetes

_clustering_cran_r.xlsx

Clustering en R

`paquete CLUSTERING`

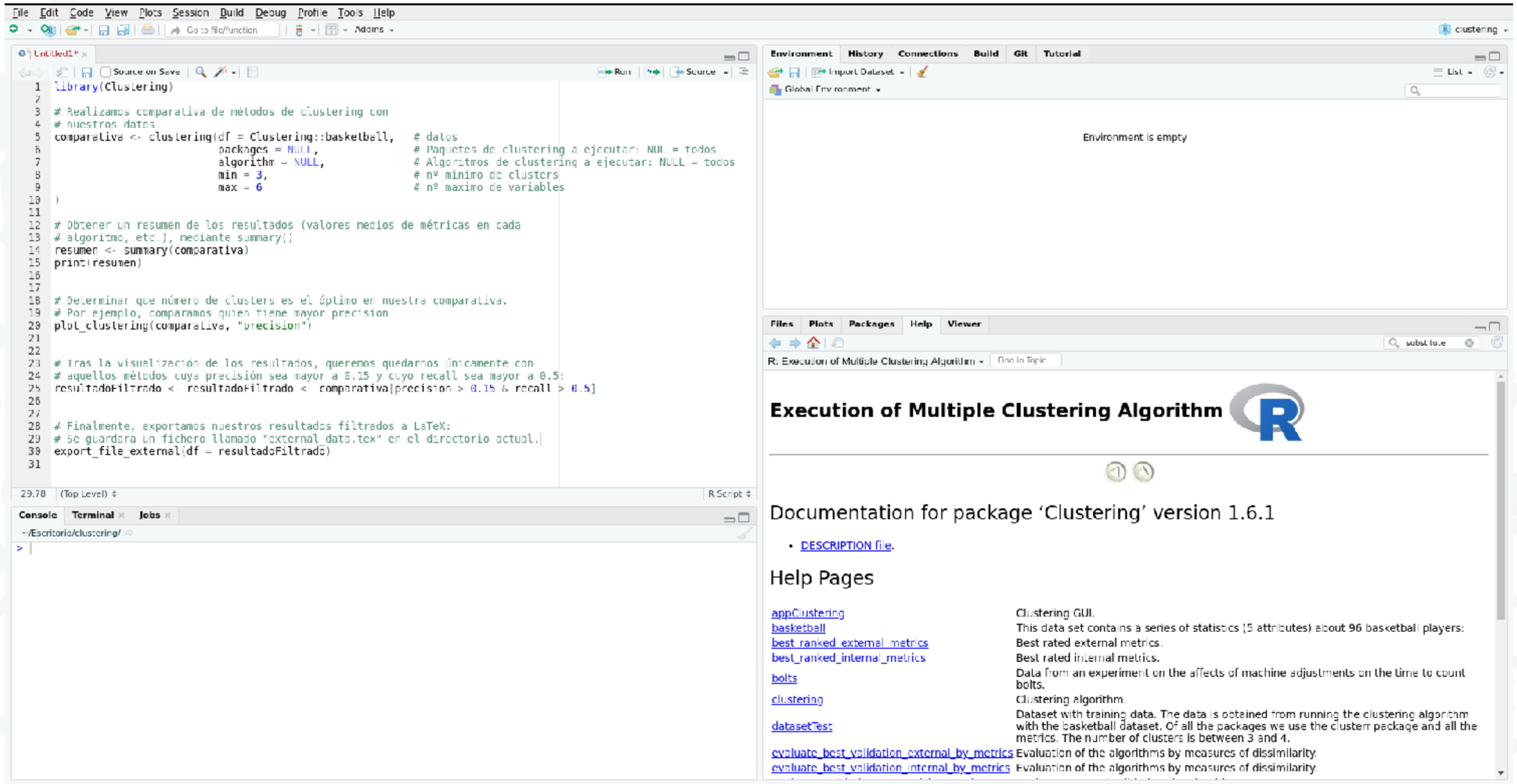
Es un nuevo paquete que hemos desarrollado que permite analizar datos no etiquetados (aprendizaje no supervisado) con distintos objetivos:

- Comparativa de algoritmos de la literatura
- Ranking de las variables que más influyen a la hora de agrupar los datos por diferentes medidas
  - Externas: Purity, Entropy, Recall, Precision, etc
  - Internas: Dunn, Connectivity y Silhoutte

Clustering en R

paquete CLUSTERING

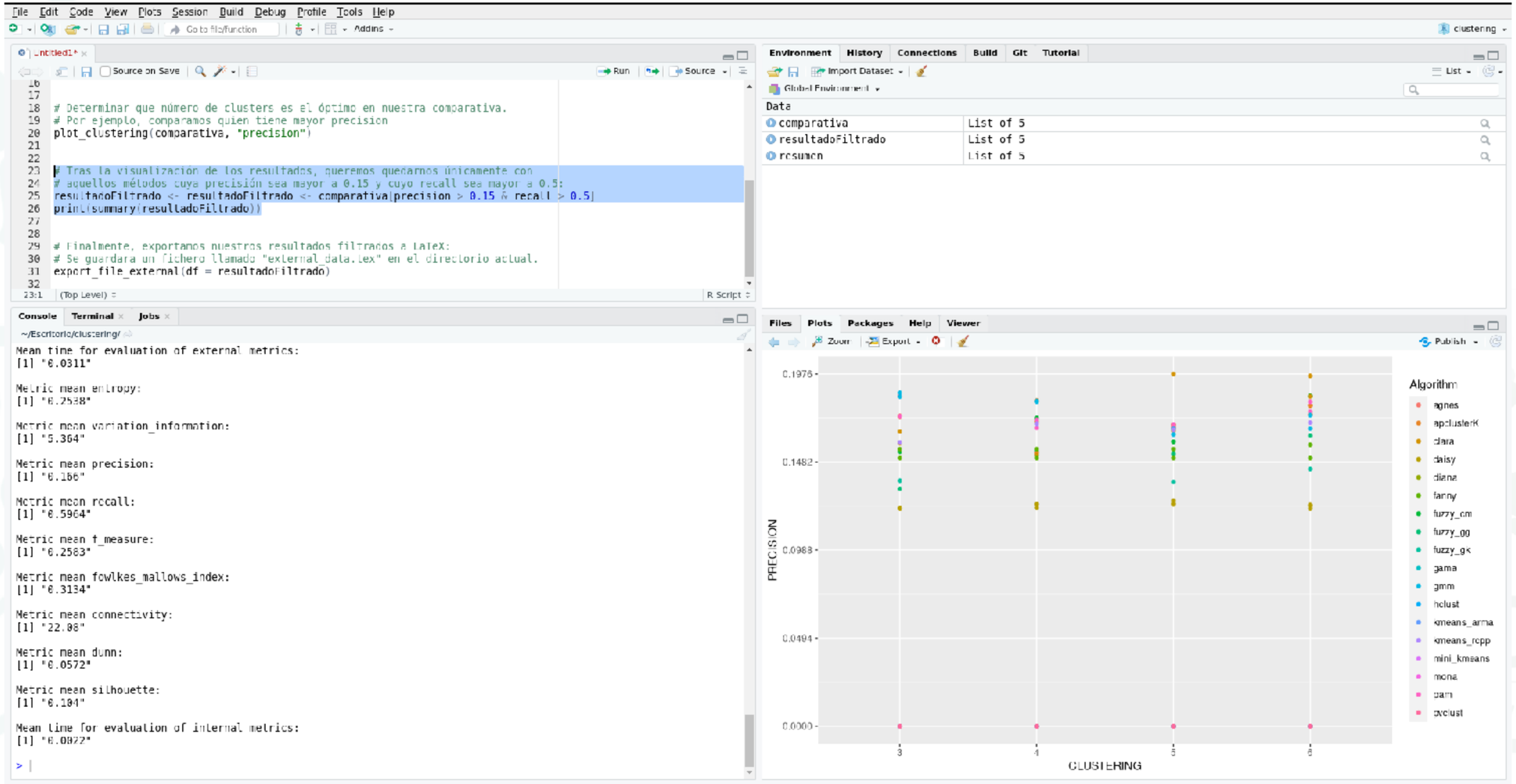
https://cran.r-project.org/web/packages/Clustering/Clustering.pdf



Clustering en R

paquete CLUSTERING

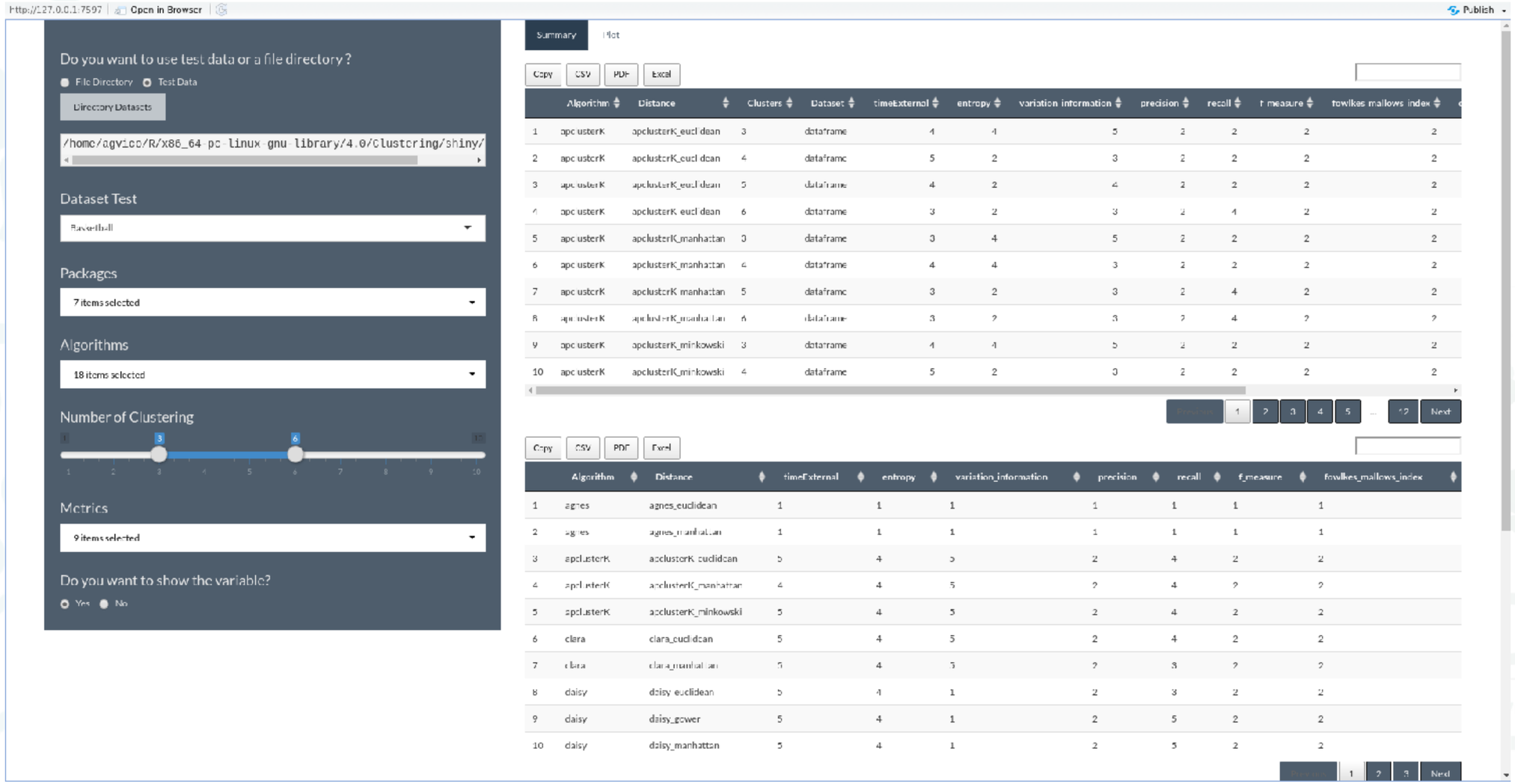
https://cran.r-project.org/web/packages/Clustering/Clustering.pdf



Clustering en R

paquete CLUSTERING

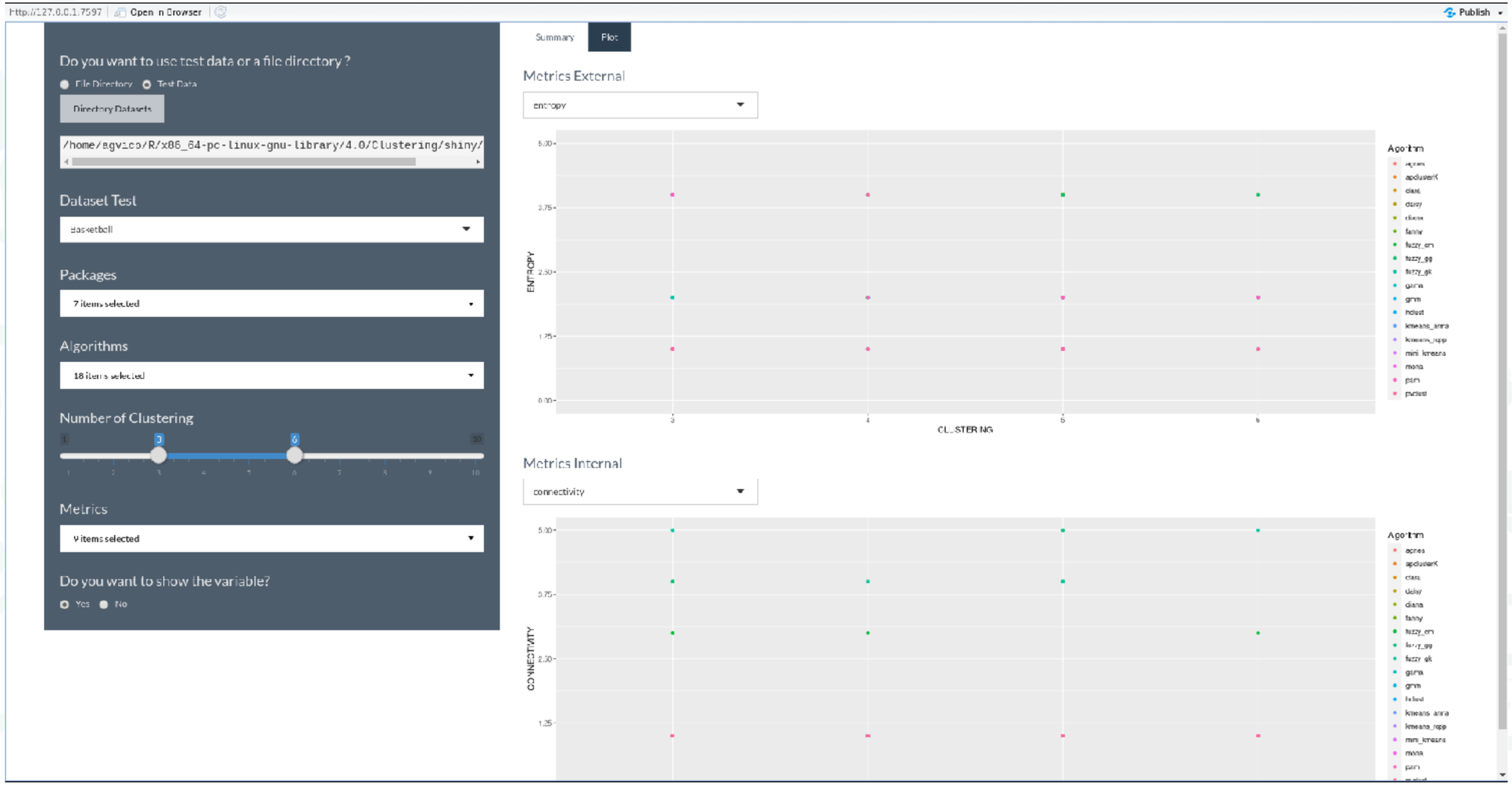
https://cran.r-project.org/web/packages/Clustering/Clustering.pdf



Clustering en R

`paquete CLUSTERING`

https://cran.r-project.org/web/packages/Clustering/Clustering.pdf



Clustering

referencias bibliográficas

- Pang-Ning Tan, Michael Steinbach & Vipin Kumar. Introduction to Data Mining, 2006.
- Jiawey Han. Data mining: concepts and techniques, 2006.
- Charu C. Aggrawal & Chandan K. Reddy (editors): Data Clustering: Algorithms and Applications. Champan & Hall / CRC Press, 2014.
- Lloyd, S. P. (1957). Least squares quantization in PCM. Technical Report RR-5497, Bell Lab, September 1957.
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297). California: University of California Press.

Instituto Andaluz Interuniversitario en
Data Science and **Computational Intelligence**

DaSCI

Universidad de Jaén

Seminario Permanente de Formación en Inteligencia Artificial aplicada a Defensa
SIADEF

Sesión 7: Aprendizaje no supervisado

Cristóbal J. Carmona <ccarmona@ujaen.es>
Pedro González <pglez@ujaen.es>