VINI CARIDÁ

Human-Centered Data & AI

Vinicius Caridá, Ph.D.

Head of Digital Customer Service Platforms, PCP, WFM, Data and AI - Itaú Unibanco

MBA Professor - FIAP

Google Developer Expert – Machine Learning

Co-organizer TFUGSP and AWSUGSP

@vinicius caridá @vfcarida @vinicius caridá @vfcarida @vinicius caridá @vfcarida

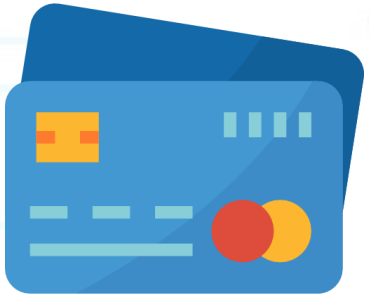
VINI CARIDÁ
Human-Centered Data & AI

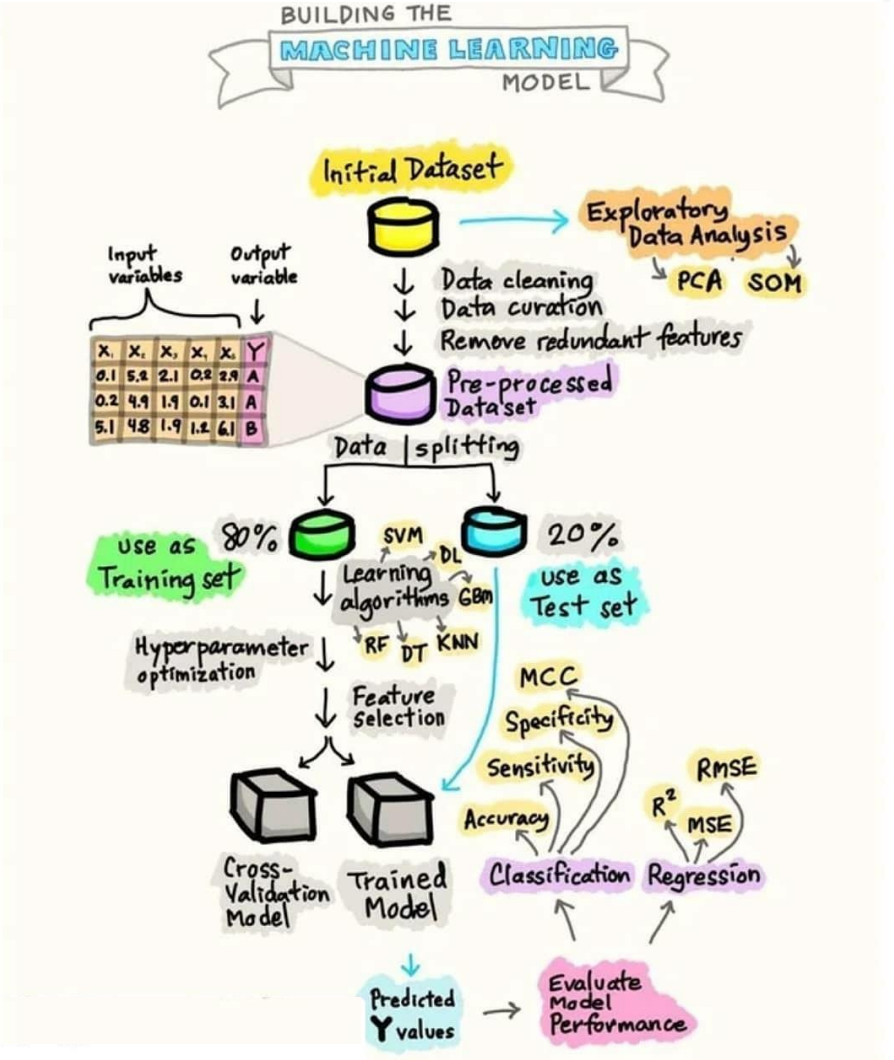
Zero to Hero
Machine Learning na AWS

Parte 3/5

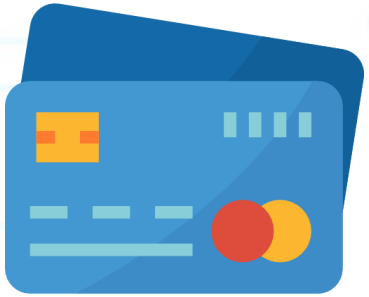
Vinicius F. Caridá

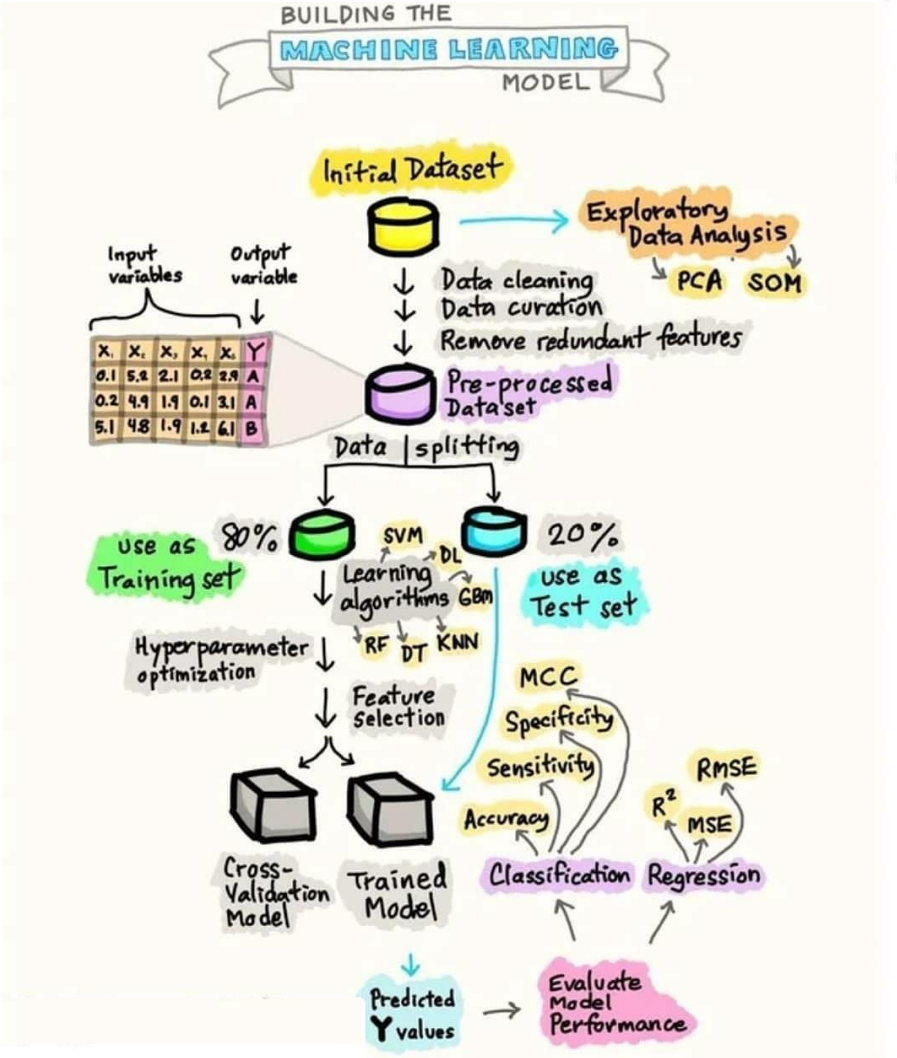
Classificação





Classificação

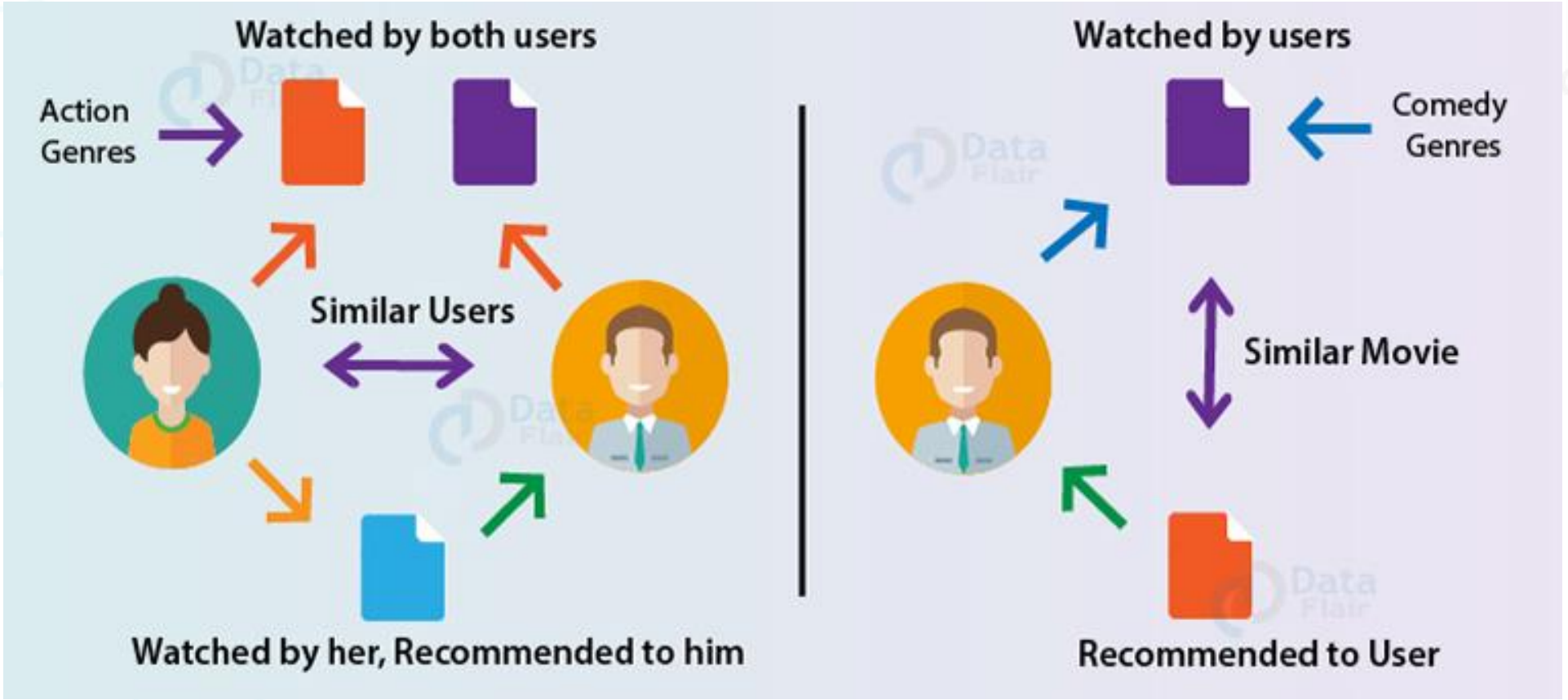




VINI CARIDÁ
Human-Centered Data & AI

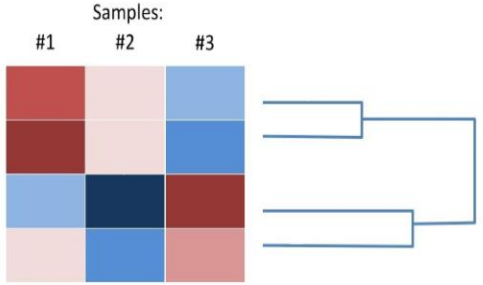
Agrupamento

Agrupamento



Agrupamento

Hierarchical Clustering, Clearly Explained!!!!



AGGLOMERATIVE CLUSTERING

All observations start as their own cluster. Clusters meeting some criteria are merged. This process is repeated, growing clusters until some end point is reached.

ChrisAlbon

K-MEANS CLUSTERING

1. k centerpoints are randomly initialized.
2. Observations are assigned to the closest centerpoint.
3. Centerpoints are moved to the center of their members.
4. Repeat steps 2 and 3 until no observation changes membership in step 2.

Chris Albon

DBSCAN

DBSCAN looks for densely packed observations and makes no assumptions about the number or shape of clusters.

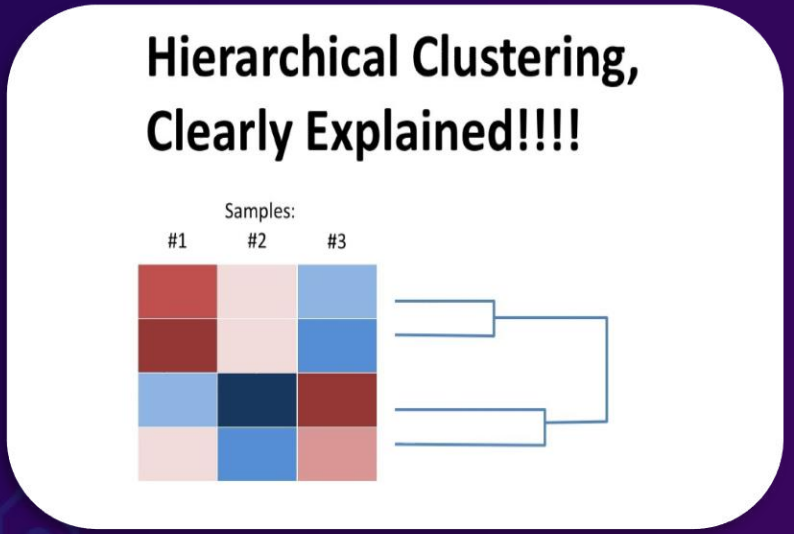
1. A random observation, x_i, is selected
2. If x_i has a minimum of close neighbors, we consider it part of a cluster.
3. Step 2 is repeated recursively for all of x_i's neighbors, then neighbors' neighbors etc... These are the cluster's core members.
4. Once Step 3 runs out of observations, a new random point is chosen

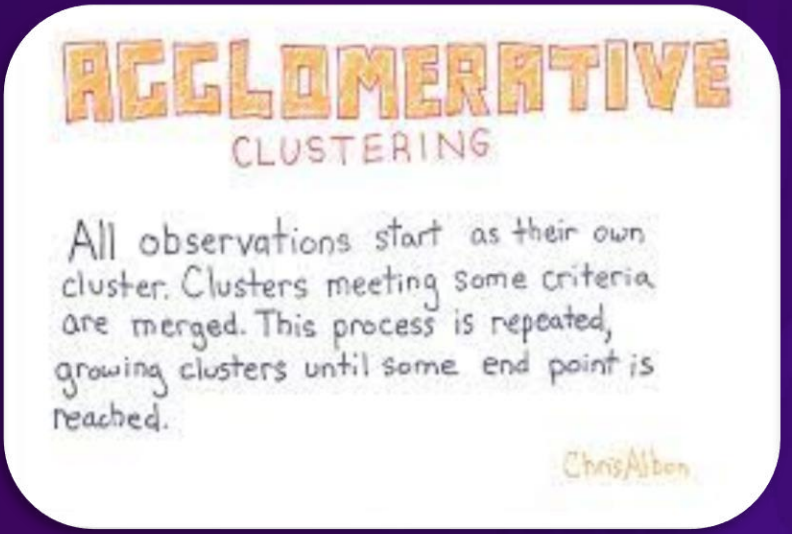
Afterwards, observations not part of a core are assigned to a nearby cluster or marked as outliers.

ChrisAlbon

VINI CARIDÁ
Human-Centered Data & AI

Hierárquico





Hierarquia: Conceitos Básicos

Hierarquias são comumente usadas para organizar informação

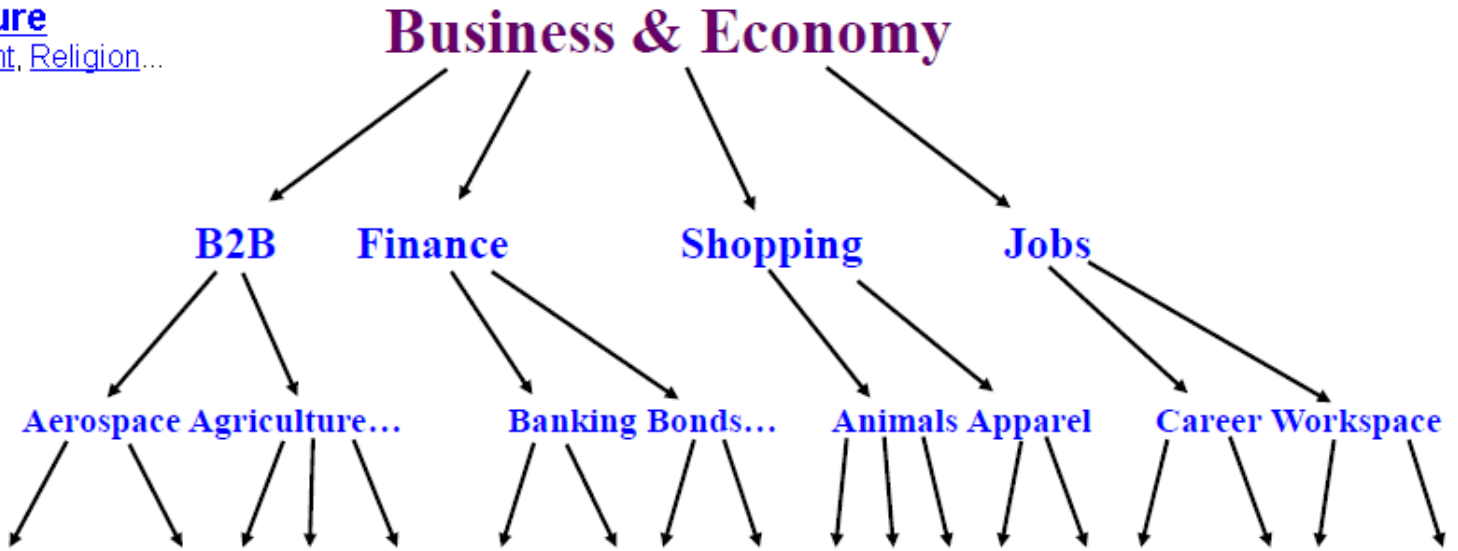
Web Site Directory - Sites organized by subject Suggest your site

Business & Economy
B2B, Finance, Shopping, Jobs...

Regional
Countries, Regions, US States...

Computers & Internet
Internet, WWW, Software, Games...

Society & Culture
People, Environment, Religion...



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Hierarquia: Conceitos Básicos

Exemplo: árvores filogenéticas em biologia

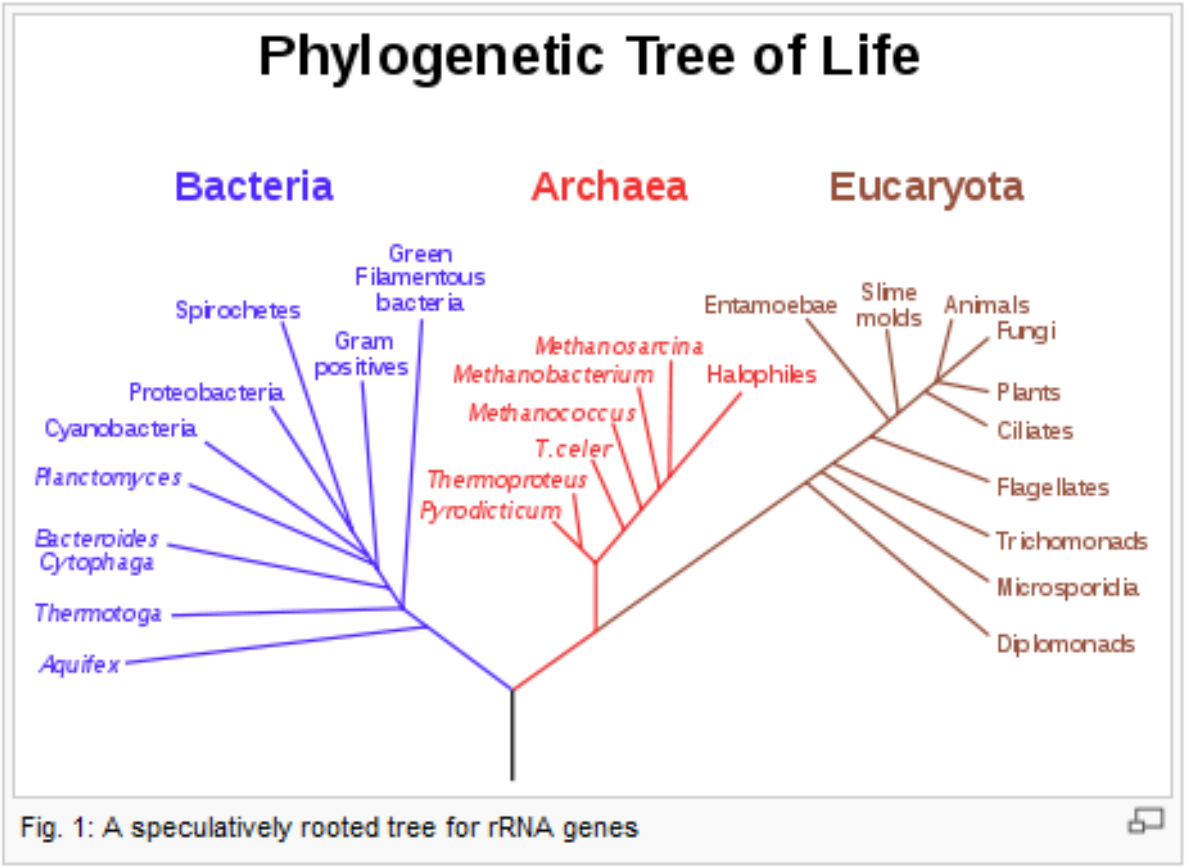
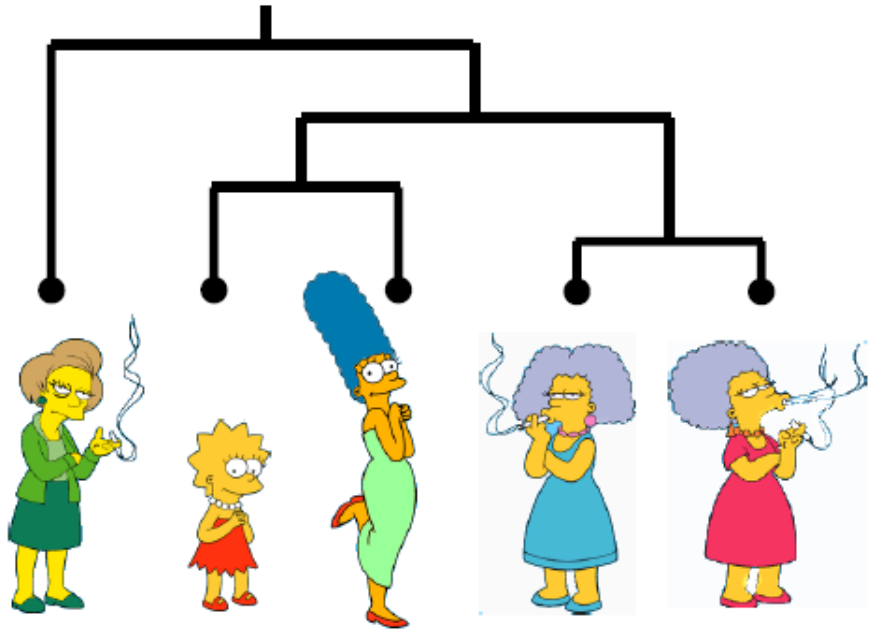


Fig. 1: A speculatively rooted tree for rRNA genes

http://en.wikipedia.org/wiki/Phylogenetic_tree

Métodos Clássicos para Agrupamento Hierárquico



Bottom-Up (aglomerativos):

- Iniciar colocando cada objeto em um *cluster*
- Encontrar o melhor par de *clusters* para unir
- Unir o par de *clusters* escolhido
- Repetir até que todos os objetos estejam reunidos em um só *cluster*

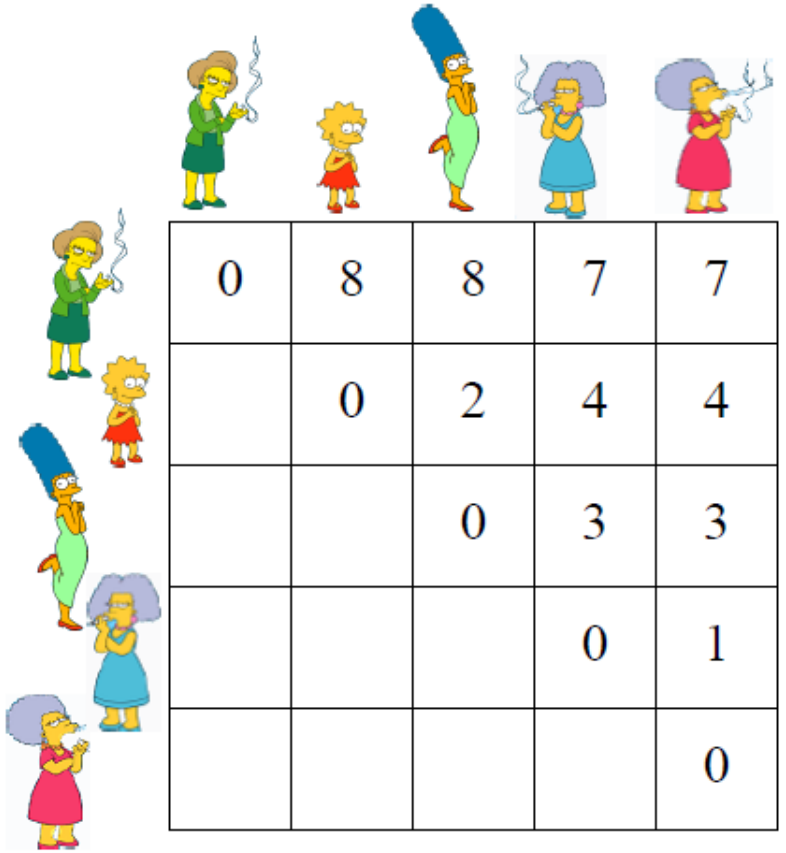
Top-Down (divisivos):

- Iniciar com todos objetos em um único *cluster*
- Sub-dividir o *cluster* em dois novos *clusters*
- Aplicar o algoritmo recursivamente em ambos, até que cada objeto forme um *cluster* por si só

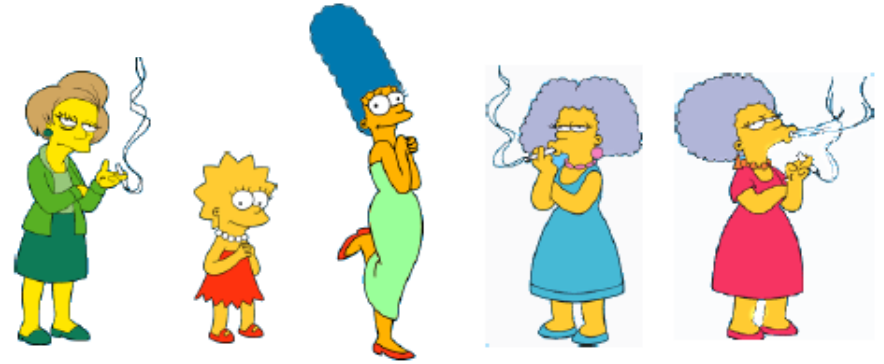
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Métodos Clássicos para Agrupamento Hierárquico

Algoritmos hierárquicos podem operar somente sobre uma matriz de distâncias.



0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0



$D(\,,\,) = 1$

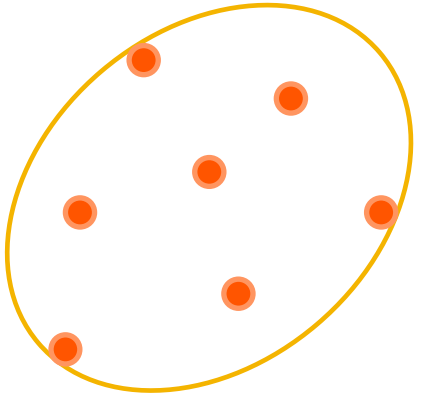
$D(\,,\,) = 8$

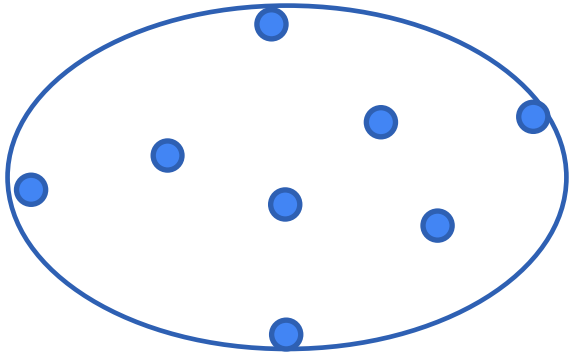
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Como definir Inter-Cluster (Dis)similaridade

Matriz de Distância

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						





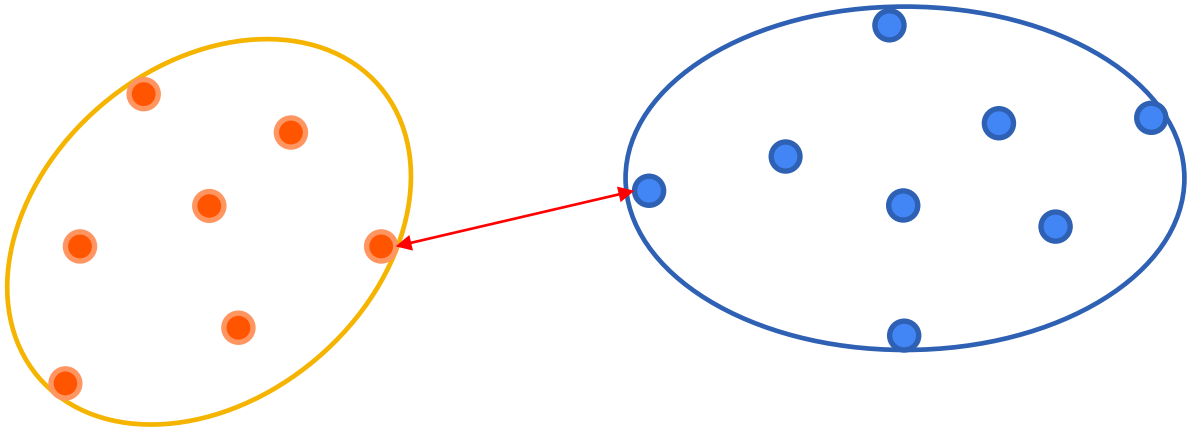
- MIN
- MAX
- Group Average
- ...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Como definir Inter-Cluster (Dis)similaridade

Matriz de Distância

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

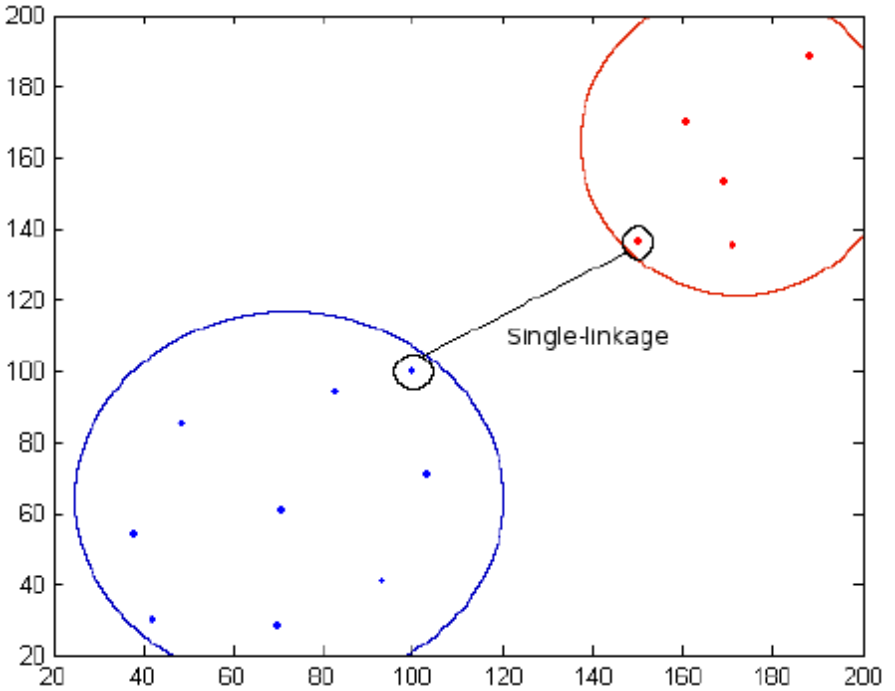


- MIN
- MAX
- Group Average
- ...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Single Linkage (Florek, 1951)

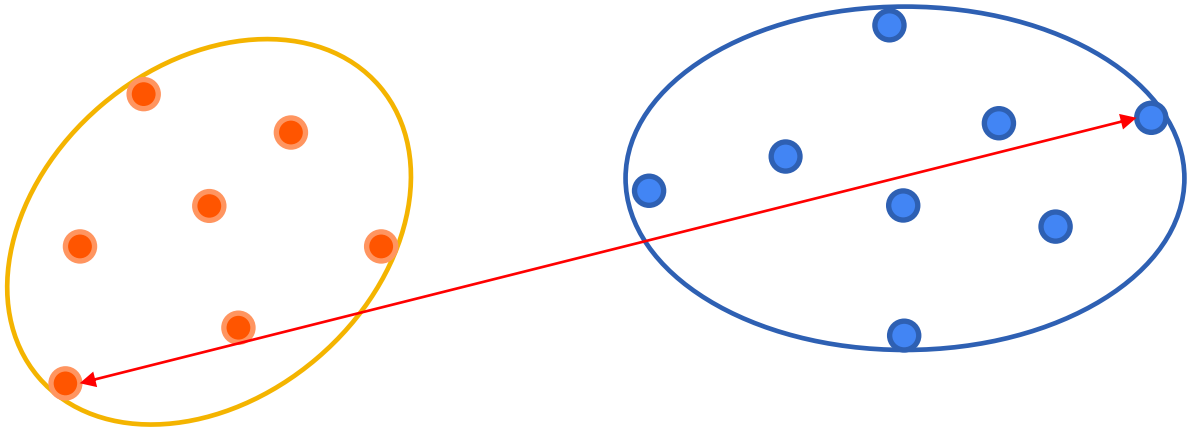
- Dissimilaridade entre clusters é dada pela **menor** dissimilaridade entre 2 objetos (um de cada cluster)
  - Originalmente baseado em Grafos: **menor** aresta entre dois vértices de subconjuntos distintos



Como definir Inter-Cluster (Dis)similaridade

Matriz de Distância

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

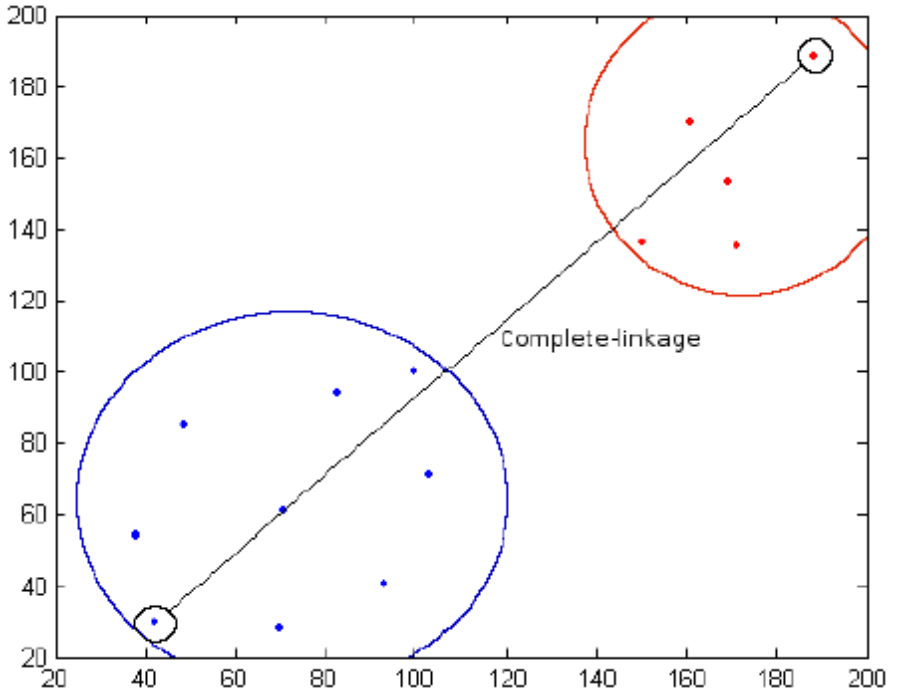


- MIN
- MAX
- Group Average
- ...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Complete Linkage (Sorensen, 1948)

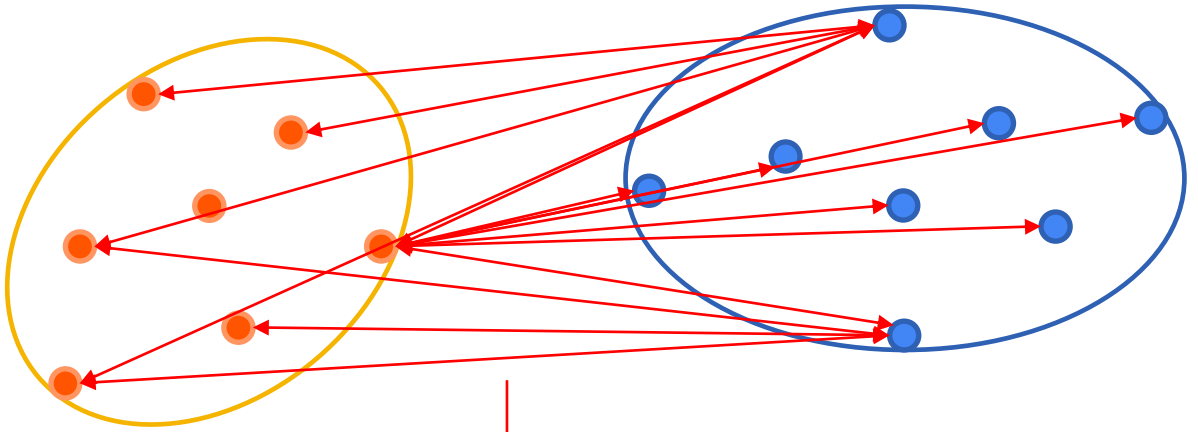
- Dissimilaridade entre clusters é dada pela **maior** dissimilaridade entre 2 objetos (um de cada cluster)
  - Originalmente baseado em Grafos: maior aresta entre dois vértices de subconjuntos distintos



Como definir Inter-Cluster (Dis)similaridade

Matriz de Distância

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						



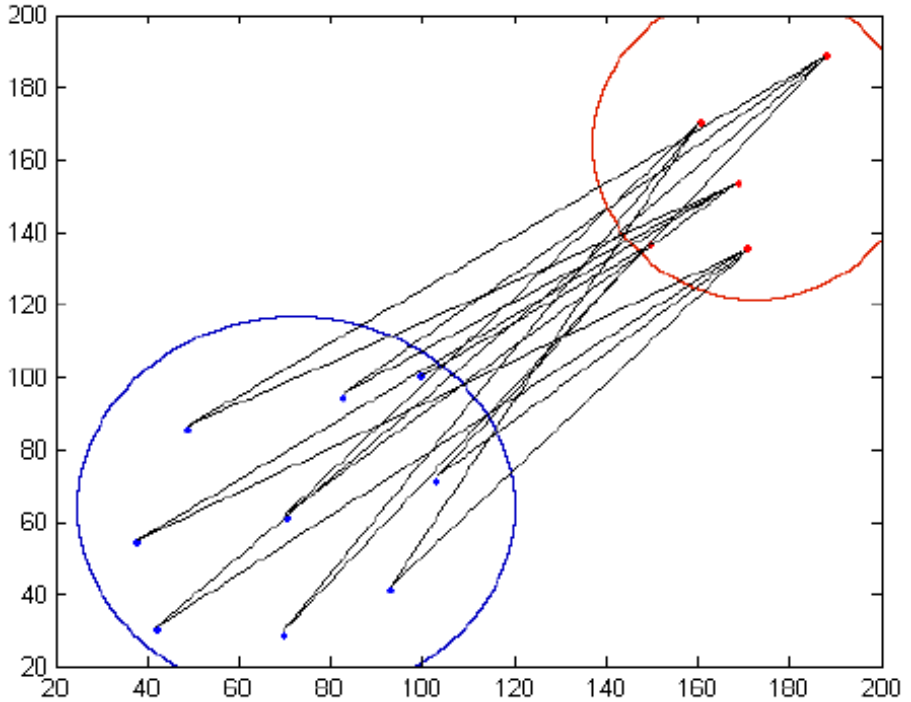
Todas as distâncias para-a-par

- MIN
- MAX
- Group Average
- ...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

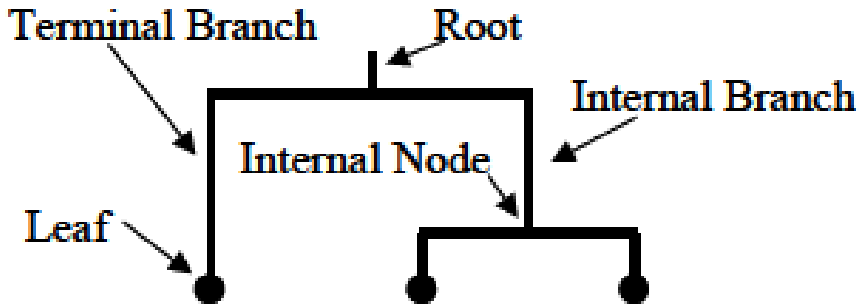
Complete Linkage (Sokal R and Michener C, 1958)

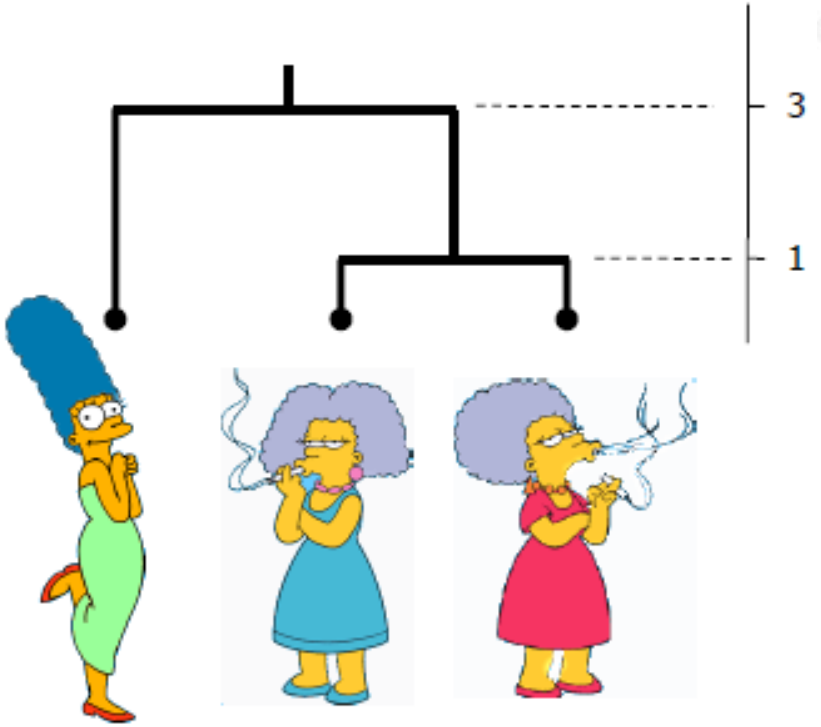
- Dissimilaridade entre clusters é dada pela **distância média** entre cada par de objetos (um de cada cluster)
- Também conhecido como UPGMA – Unweighted Pair Group Method using Arithmetic averages



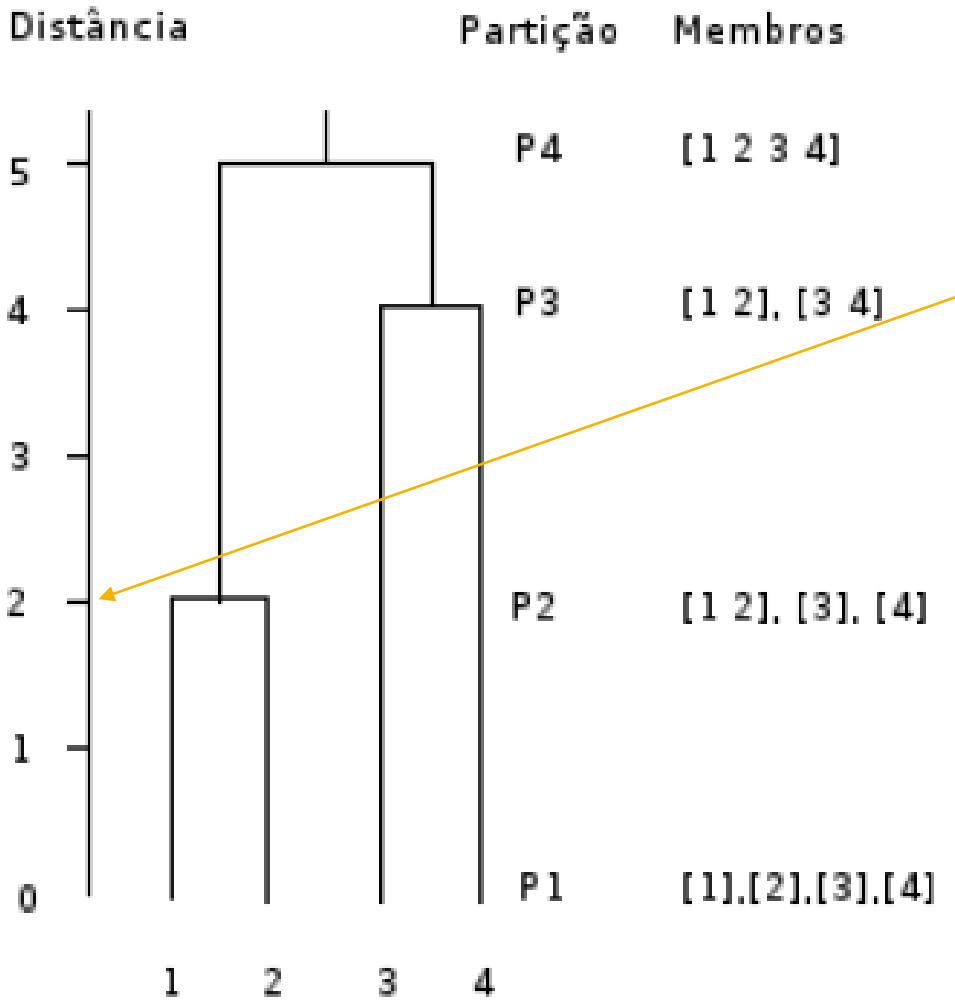
Dendrograma = Hierarquia + Dissimilaridade entre Clusters

A dissimilaridade entre dois clusters (possivelmente **singletons**) é representada como a altura do nó interno mais baixo compartilhado

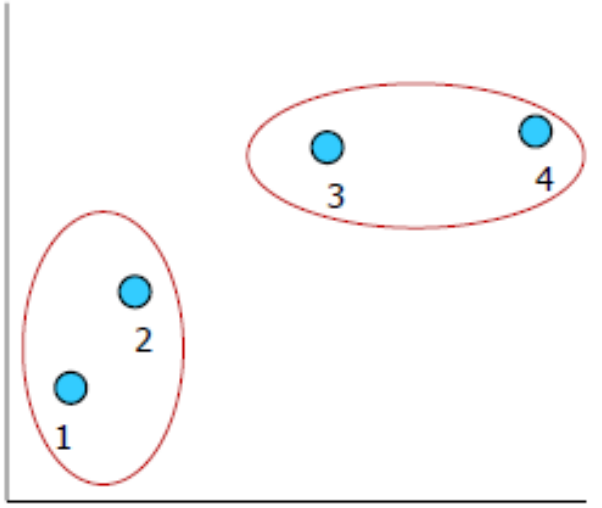




Dendrograma



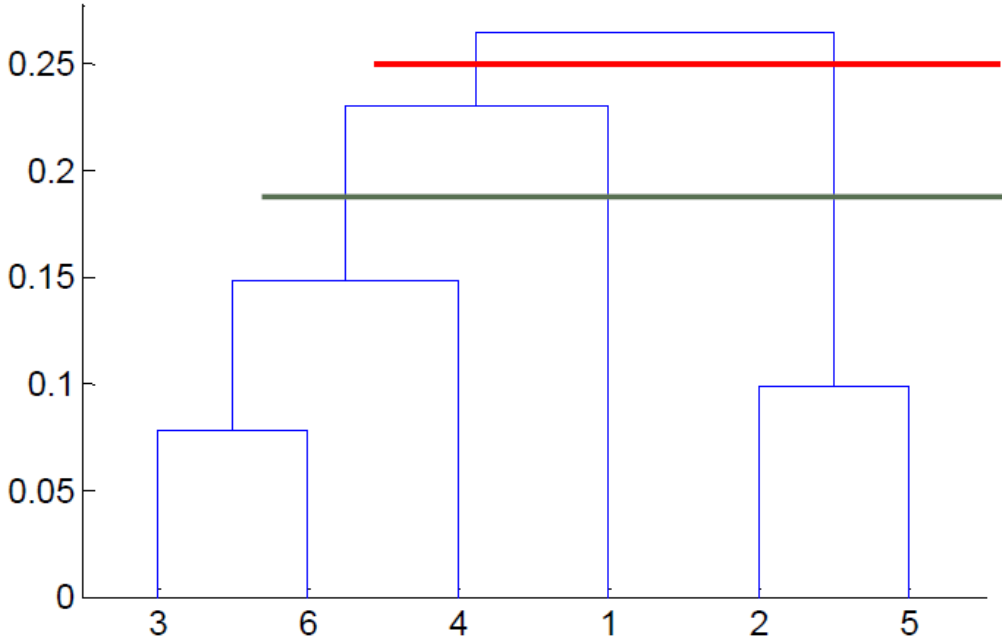
$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 2 & 7 & 13 \\ 2 & 0 & 5 & 10 \\ 7 & 5 & 0 & 4 \\ 13 & 10 & 4 & 0 \end{bmatrix} \end{array}$$



Dendrograma -> Grupos

Partições são obtidas via **cortes** no dendrograma

- cortes horizontais
- no. de grupos da partição = no. de interseções

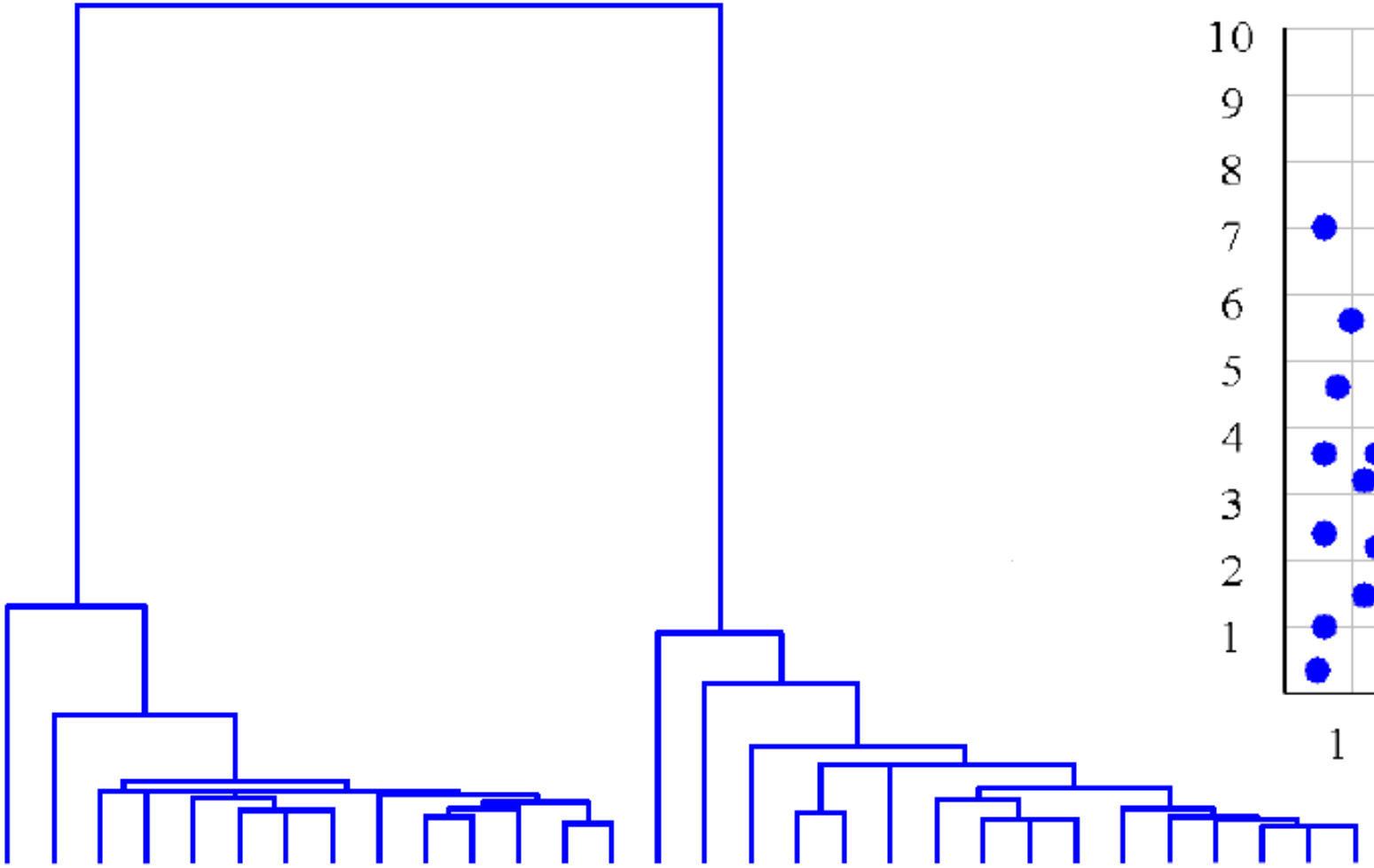


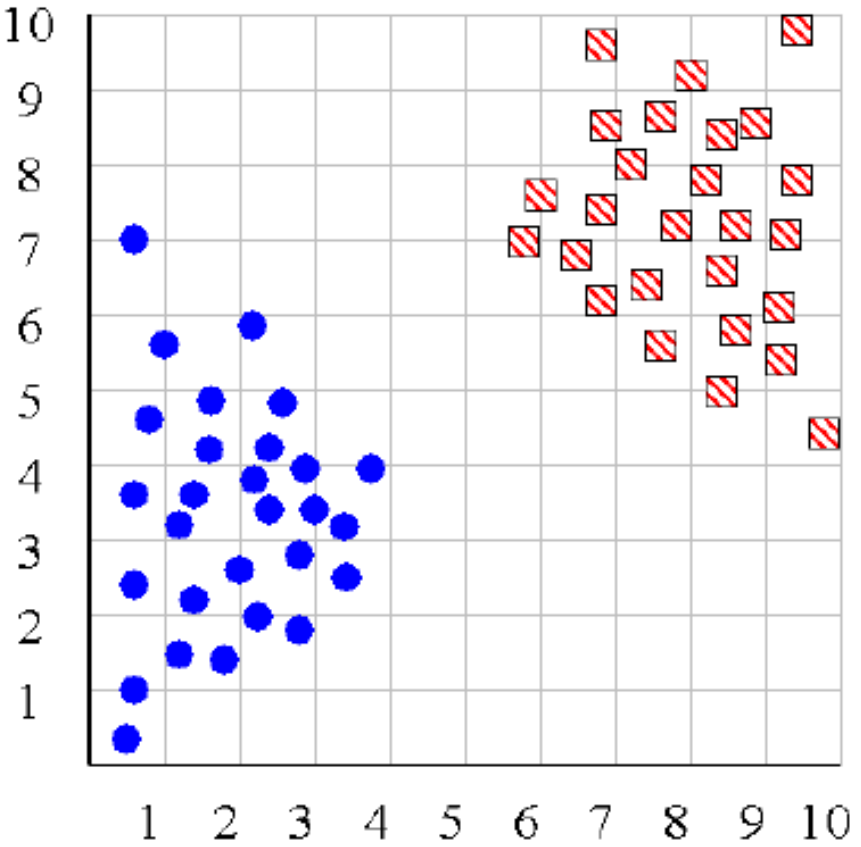
$$G_1 = \{ (x1, x3, x4, x6), (x2, x5) \}$$

$$G_2 = \{ (x1), (x3, x4, x6), (x2, x5) \}$$

Dendrograma -> Grupos

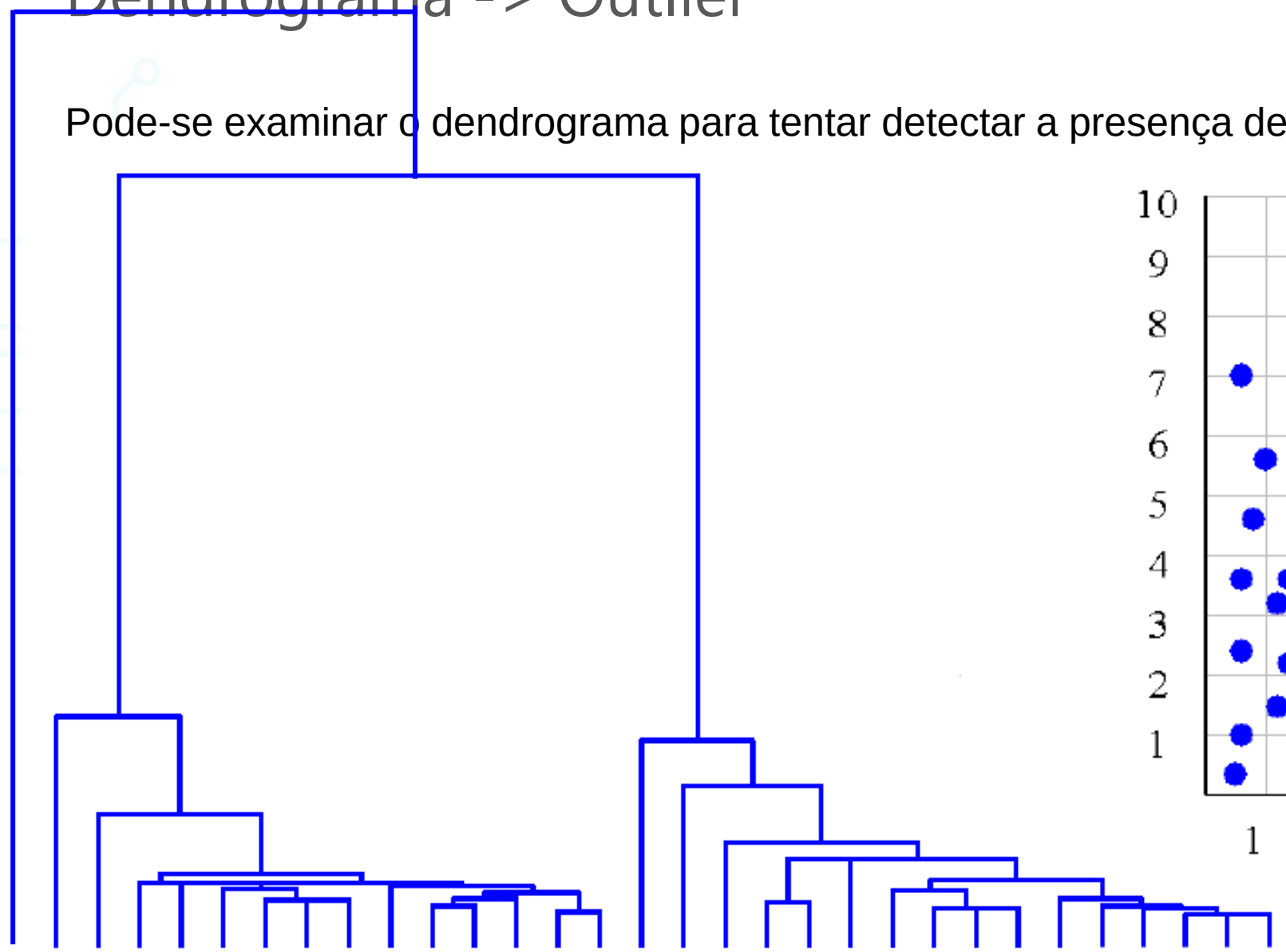
Pode-se examinar o dendrograma para tentar estimar o número mais natural de clusters.

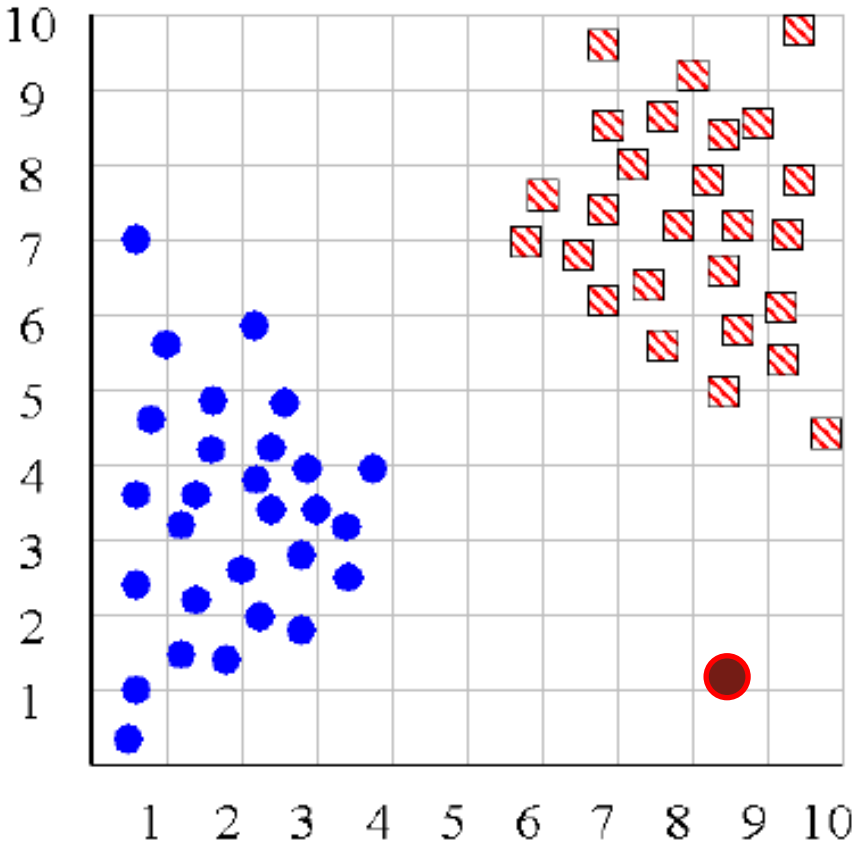




Dendrograma -> Outlier

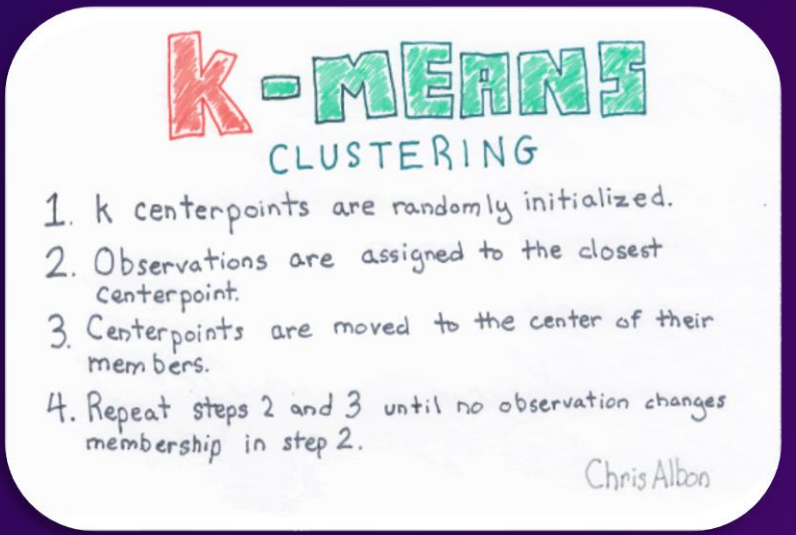
Pode-se examinar o dendrograma para tentar detectar a presença de outliers.





VINI CARIDÁ
Human-Centered Data & AI

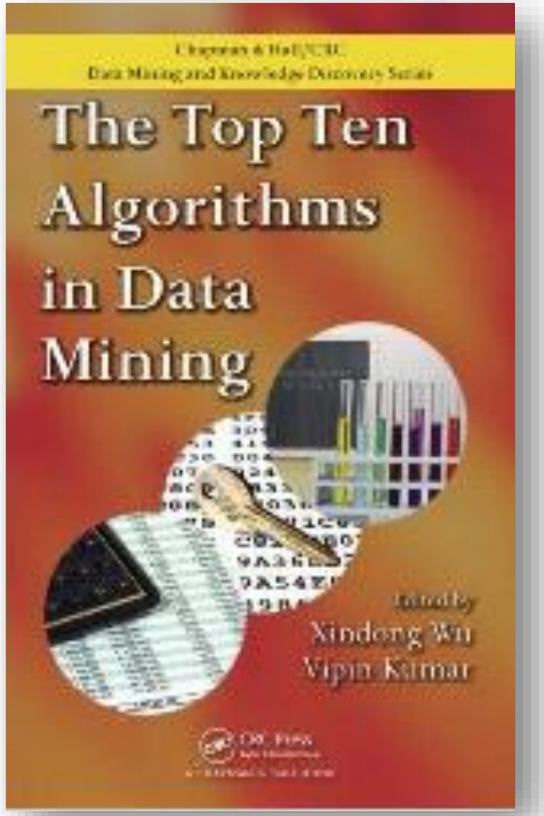
K-means



K-Means

Aqui veremos um dos algoritmos mais clássicos da área de mineração de dados em geral

- algoritmo das k-médias ou k-means
- listado entre os Top 10 Most Influential Algorithm



- Wu, X. and Kumar, V. (Editors), **The Top Ten Algorithms in Data Mining**, CRC Press, 2009

- X. Wu et al., "**Top 10 Algorithms in Data Mining**", Knowledge and Info. Systems, vol. 14, pp. 1-37, 2008

K-Means

Referência Mais Aceita como Original:

> J. B. MacQueen, Some methods of classification and analysis of multivariate observations, In Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, USA, 1967, 281–297

Porém...

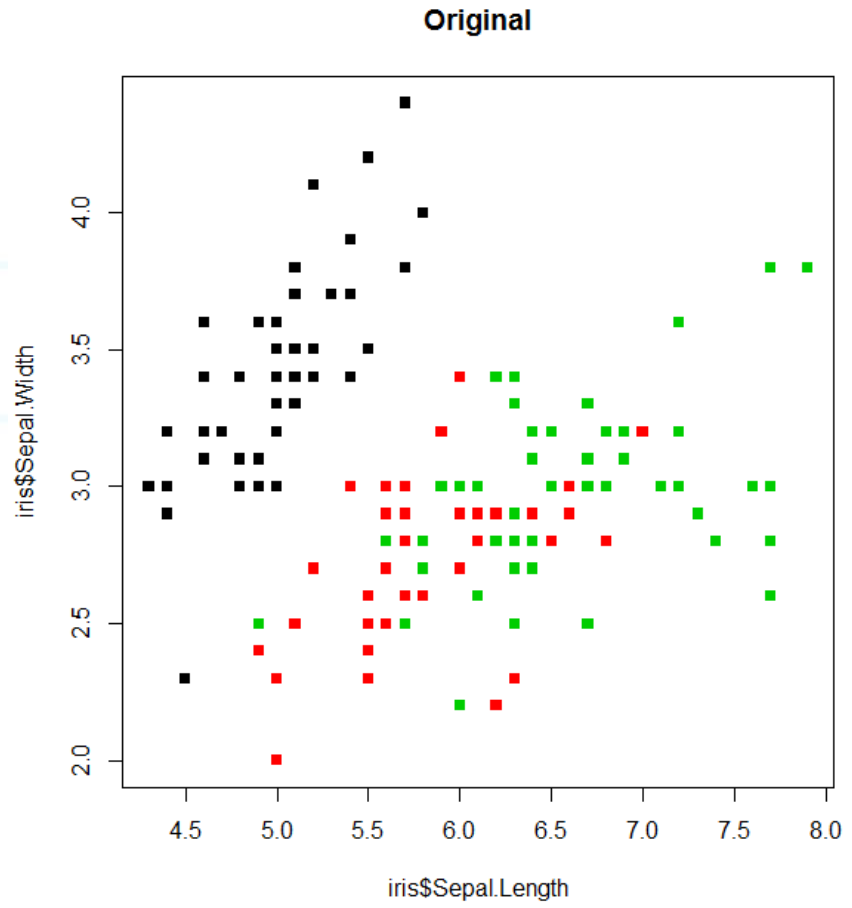
> “K-means has a rich and diverse history as it was independently discovered in different scientific fields by Steinhaus (1956), Lloyd (proposed in 1957, published in 1982), Ball & Hall (1965) and MacQueen (1967)” [Jain, Data Clustering: 50 Years Beyond K-Means, Patt. Rec. Lett., 2010]

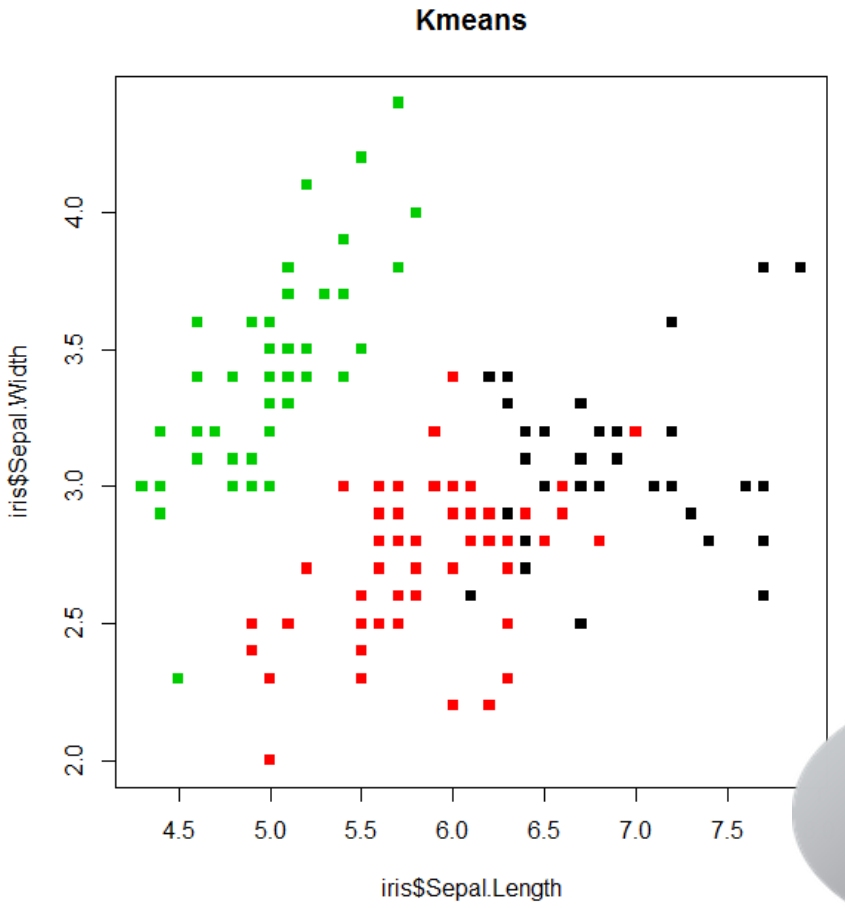
... e tem sido assunto por mais de meio século !

> Douglas Steinley, K-Means Clustering: A Half-Century Synthesis, British Journal of Mathematical and Statistical Psychology, Vol. 59, 2006

K-Means

```r
data(iris) #Carrega os dados
groups = kmeans(iris[1:4], center=3, iter.max=10)
```

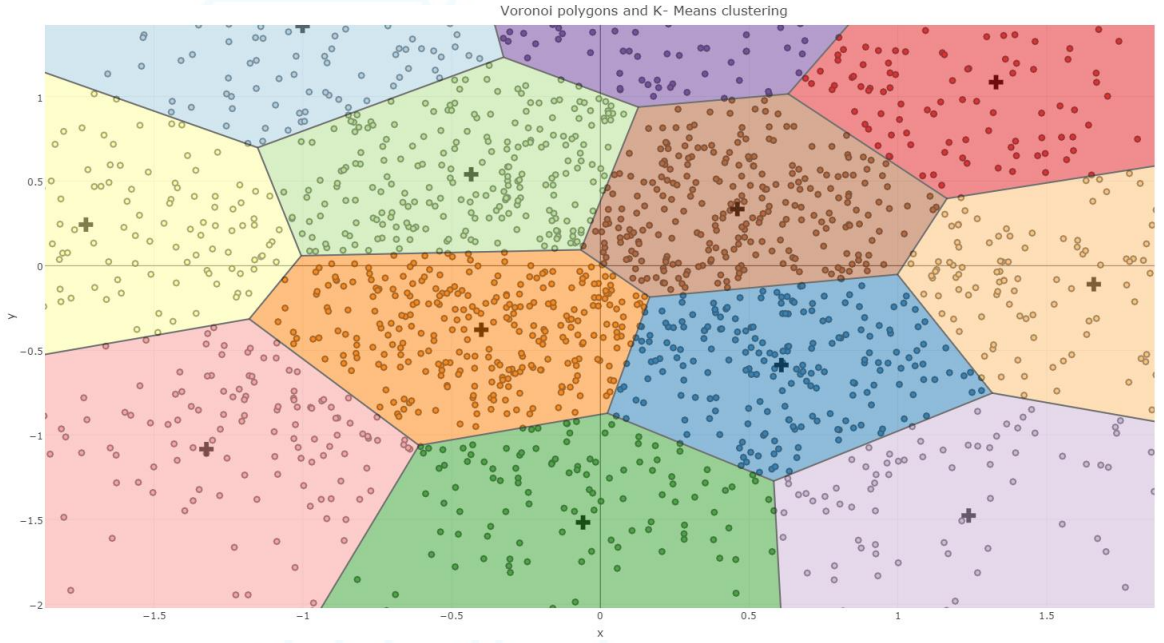


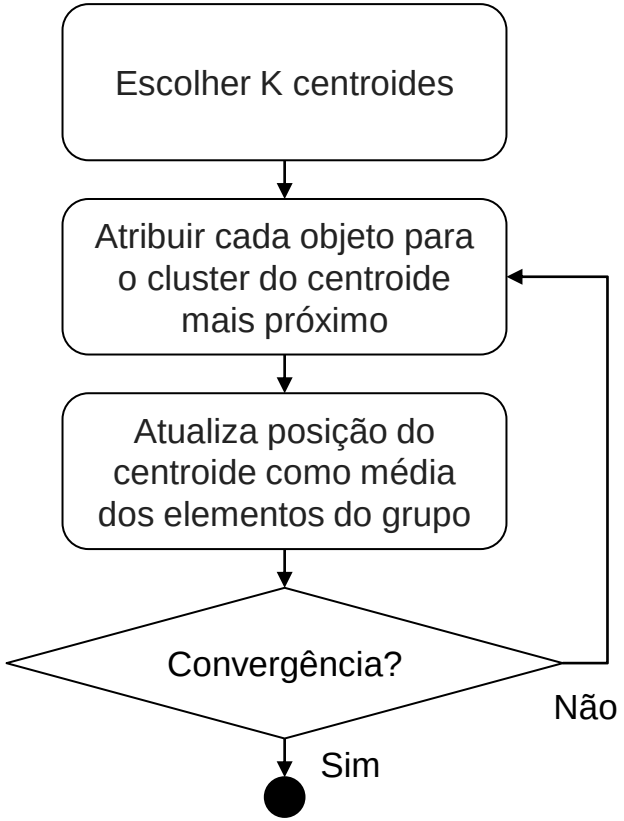


K-Means

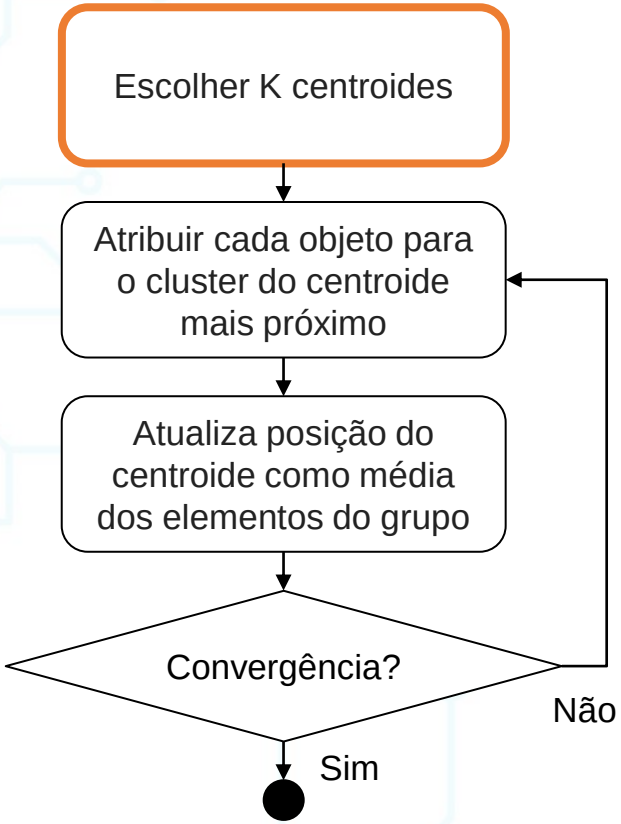
Objetiva particionar N observações dentre k grupos em que cada observação pertence ao grupo mais próximo da média. Isso resulta em uma divisão do espaço de dados em um Diagrama de Voronoi.

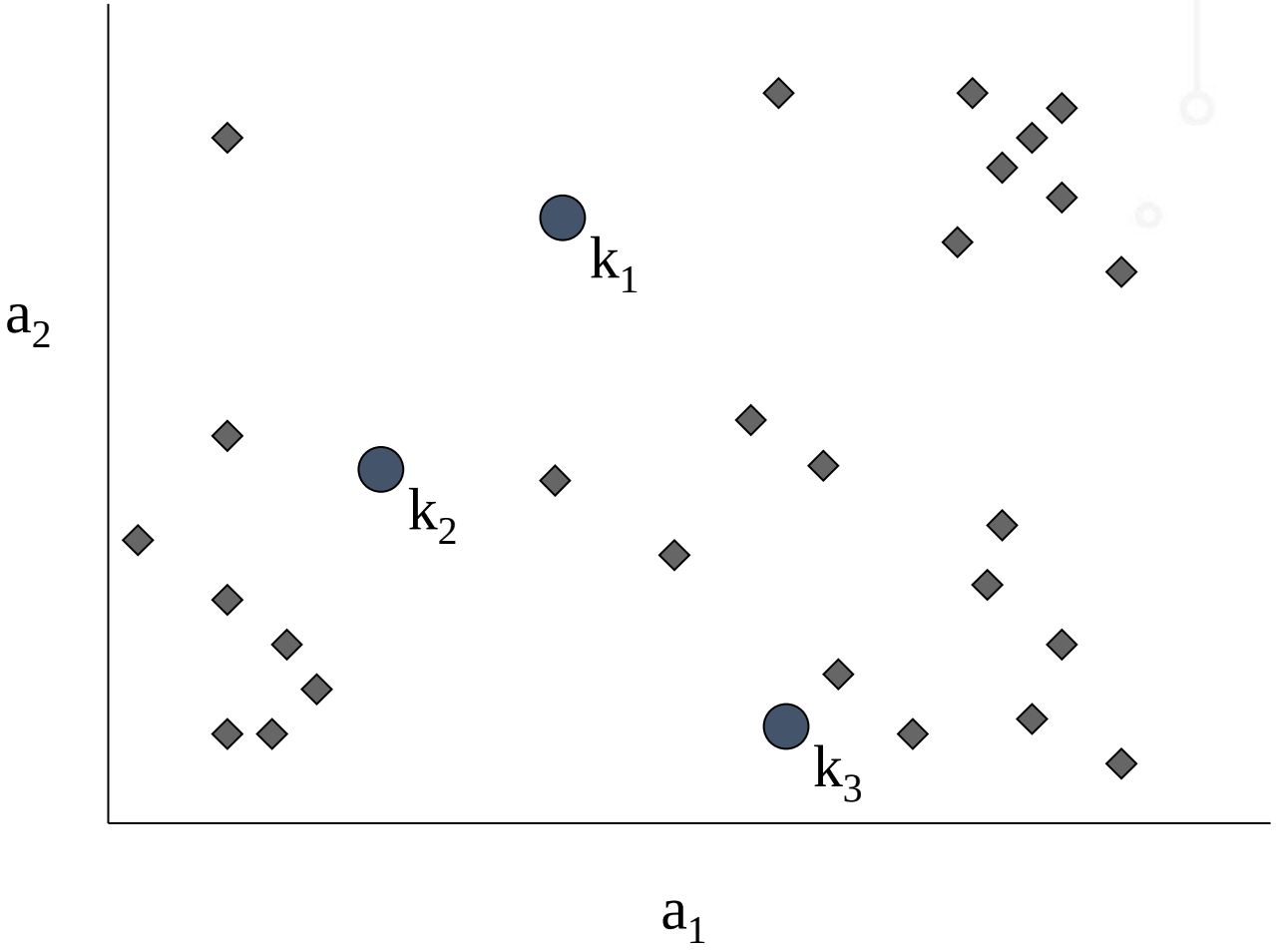
Calculado por meio da triangulação de Delaunay





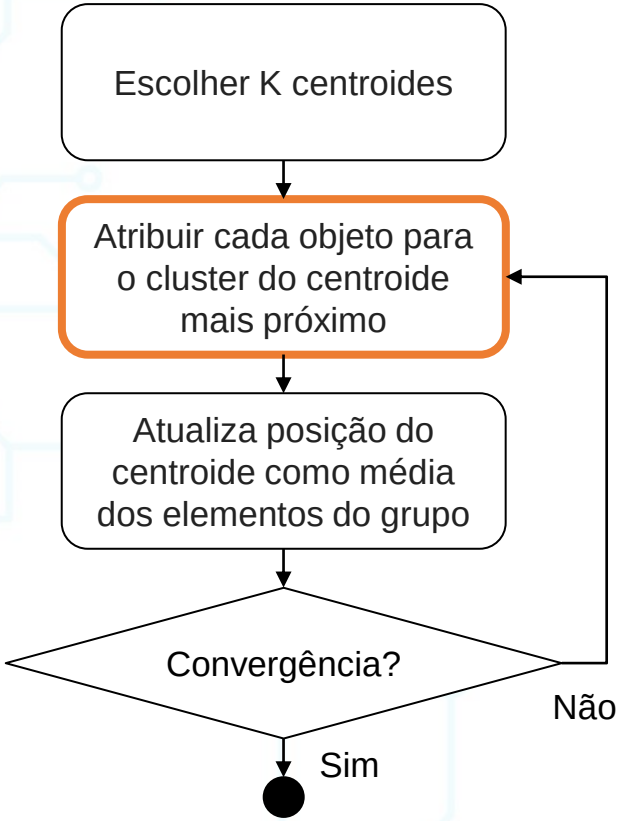
K-Means - Simulação

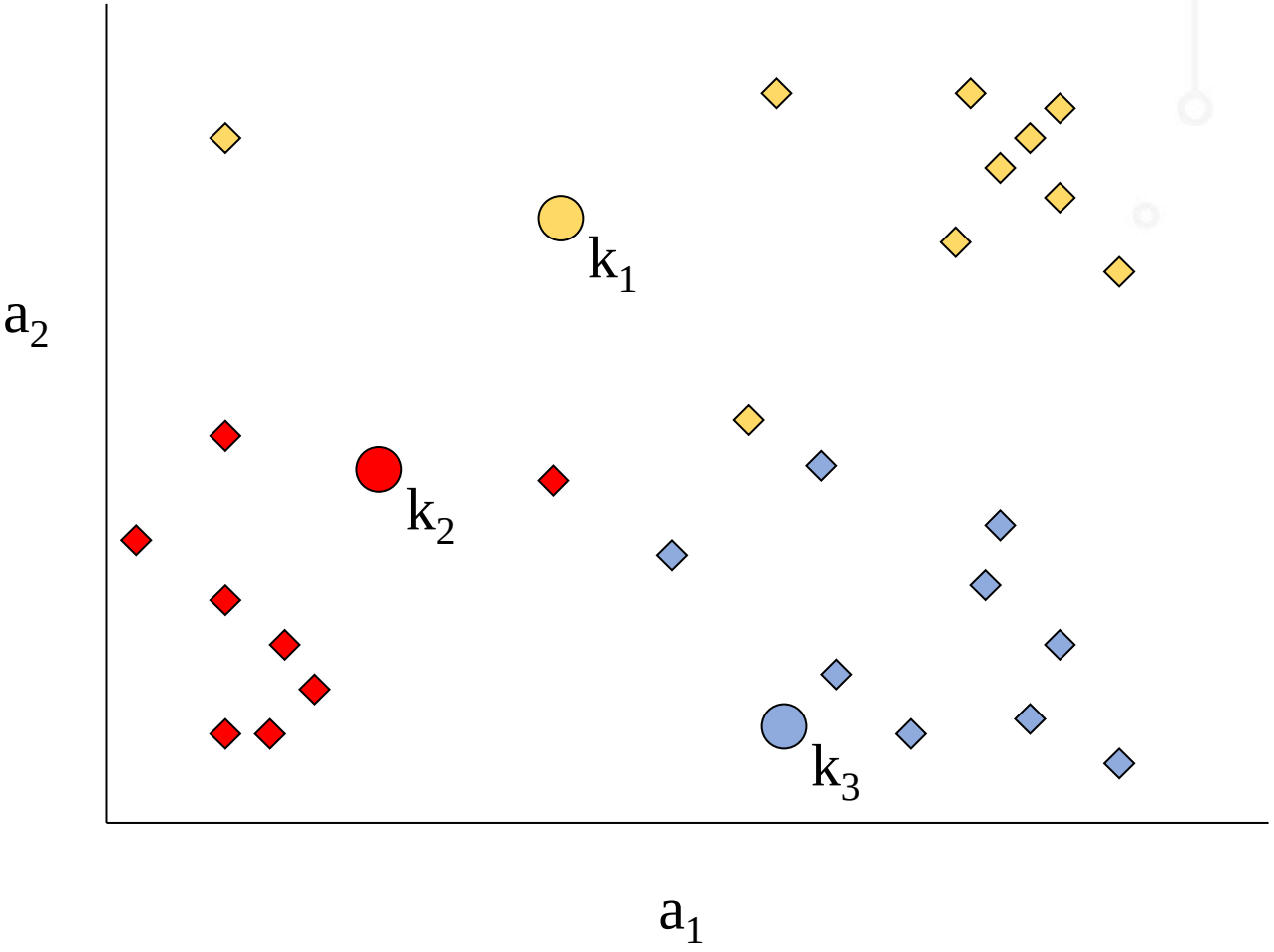




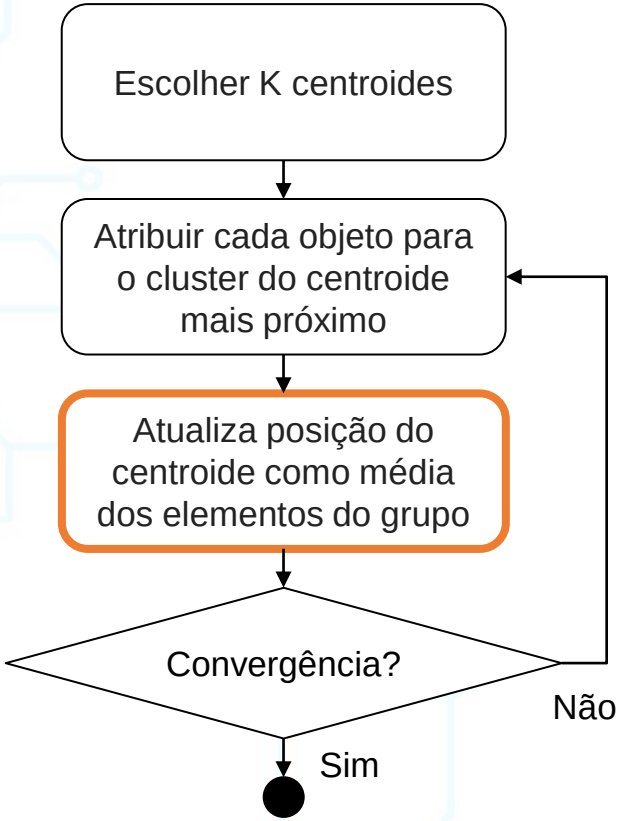
Slides desse exemplo são baseados no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

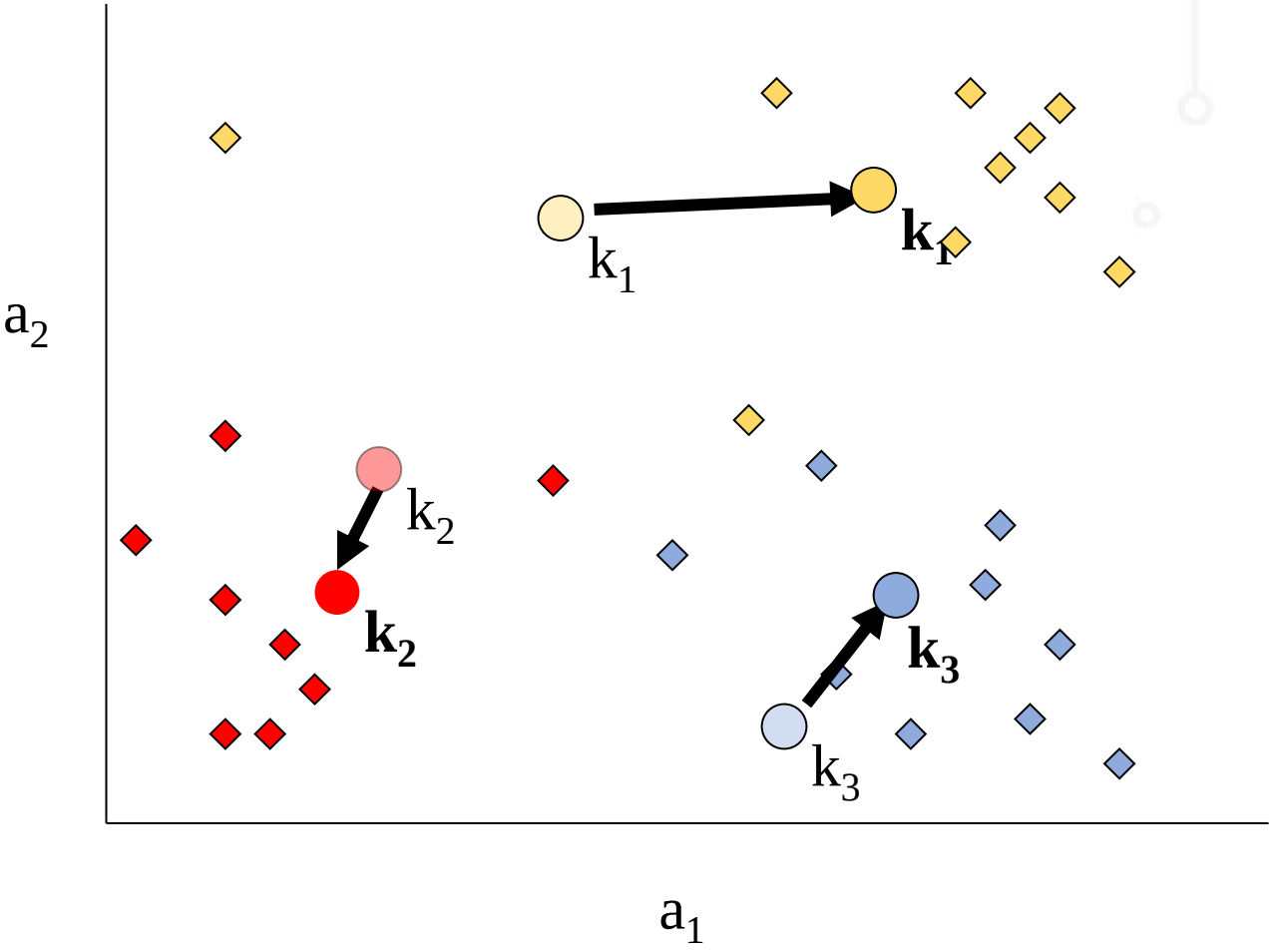
K-Means - Simulação



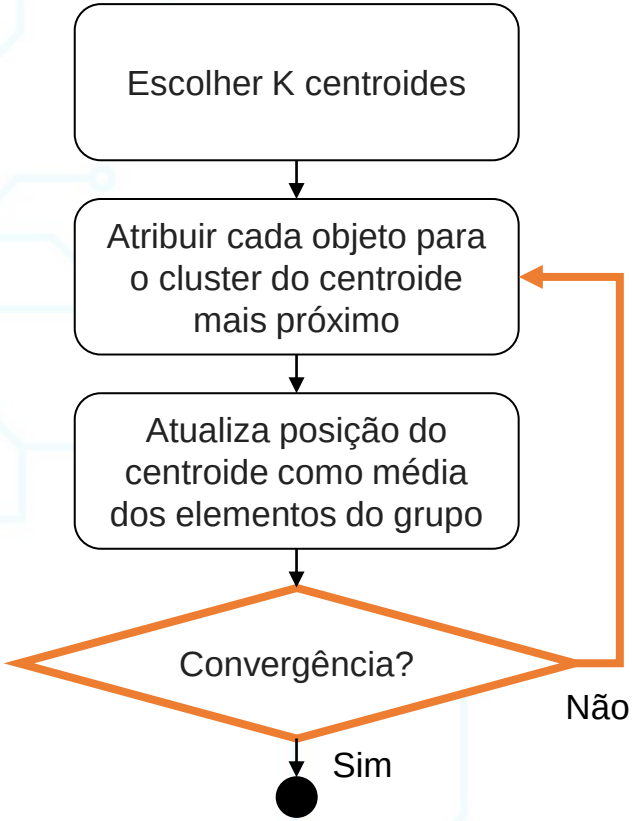


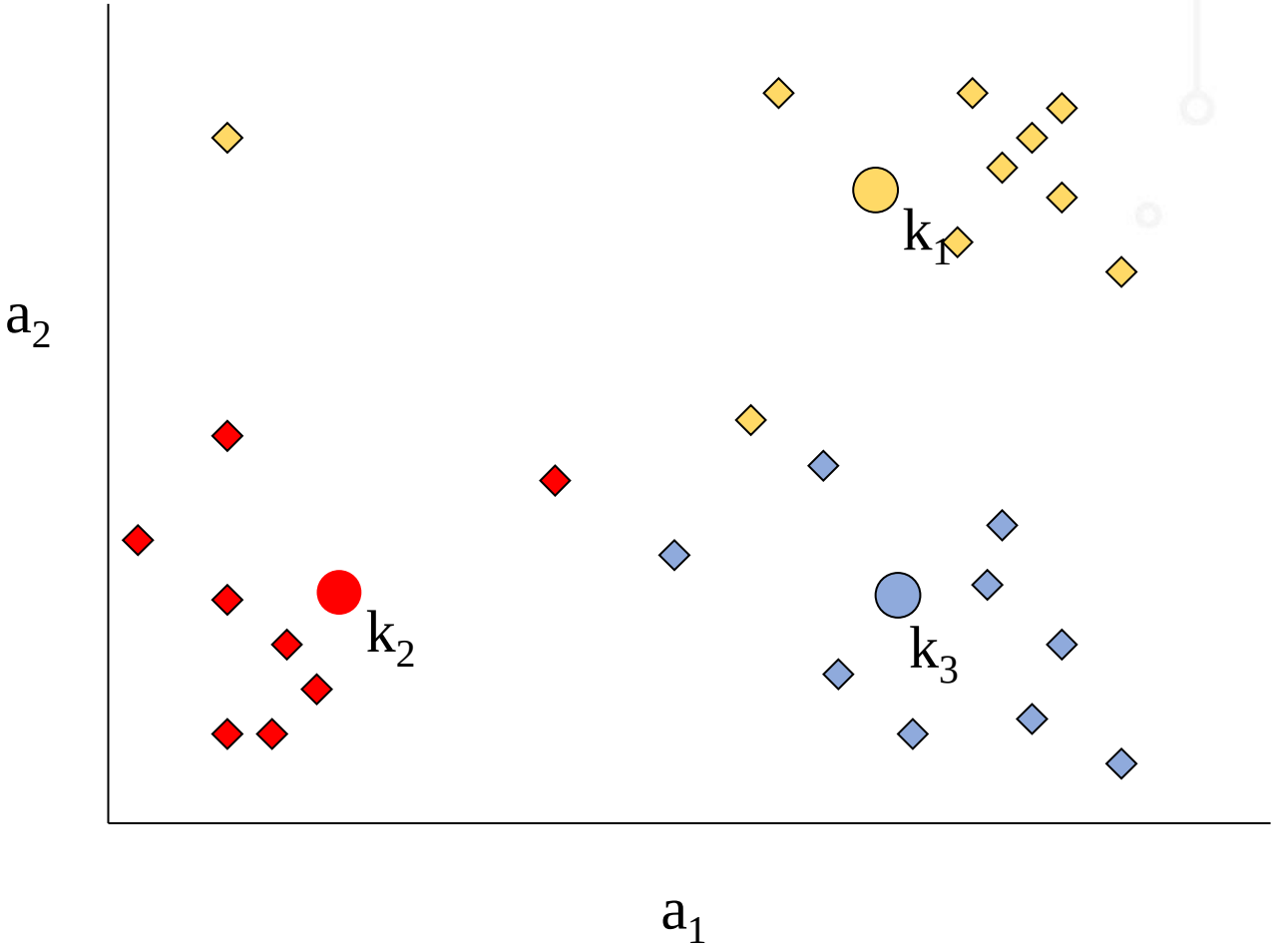
K-Means - Simulação



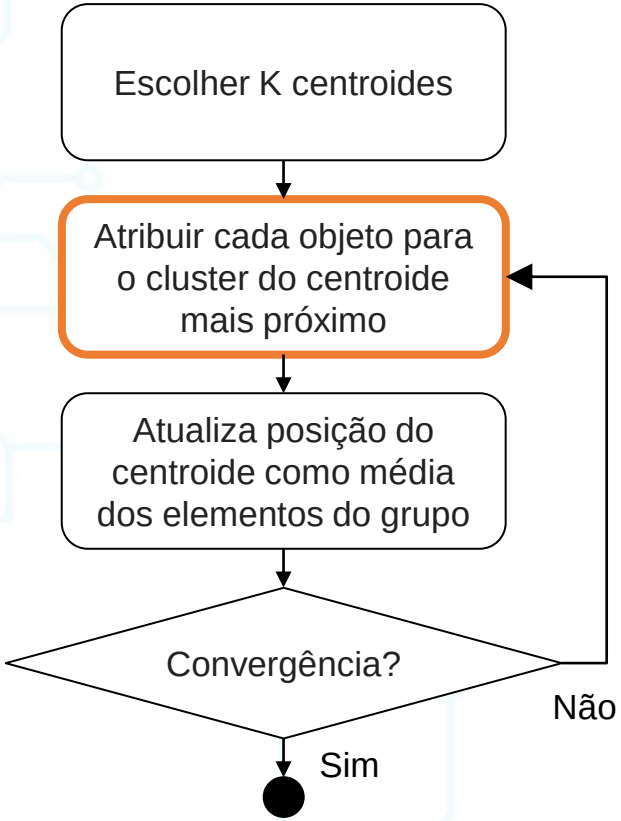


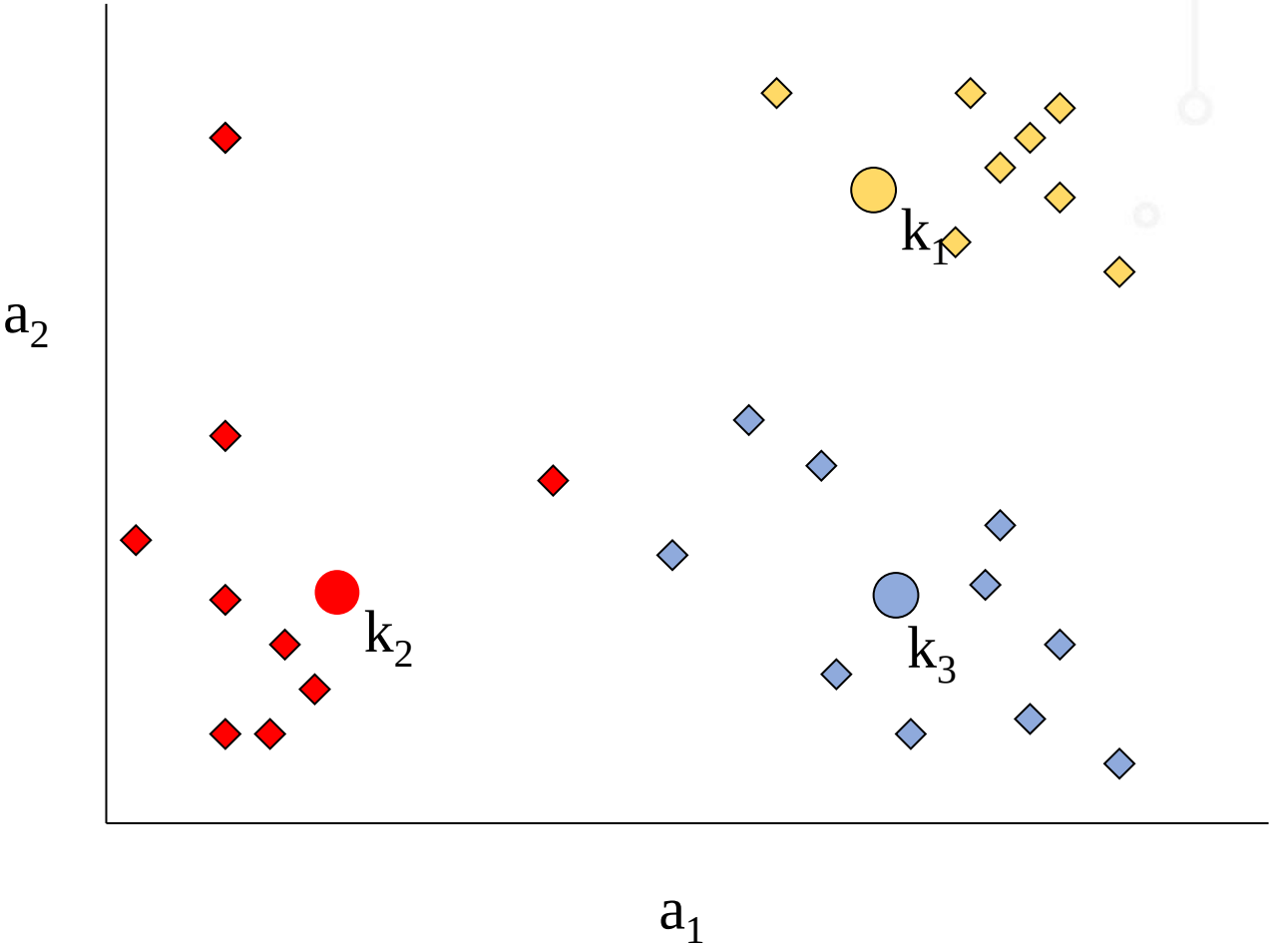
K-Means - Simulação



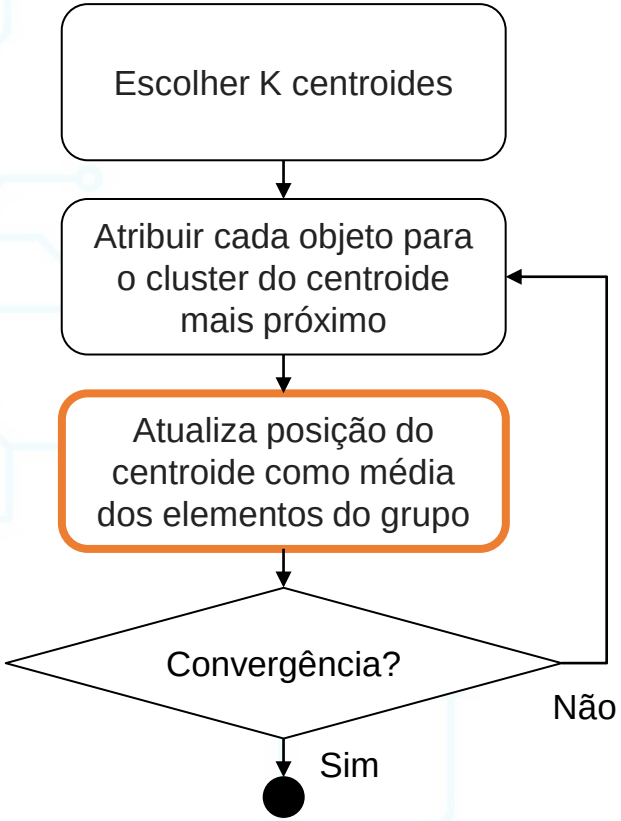


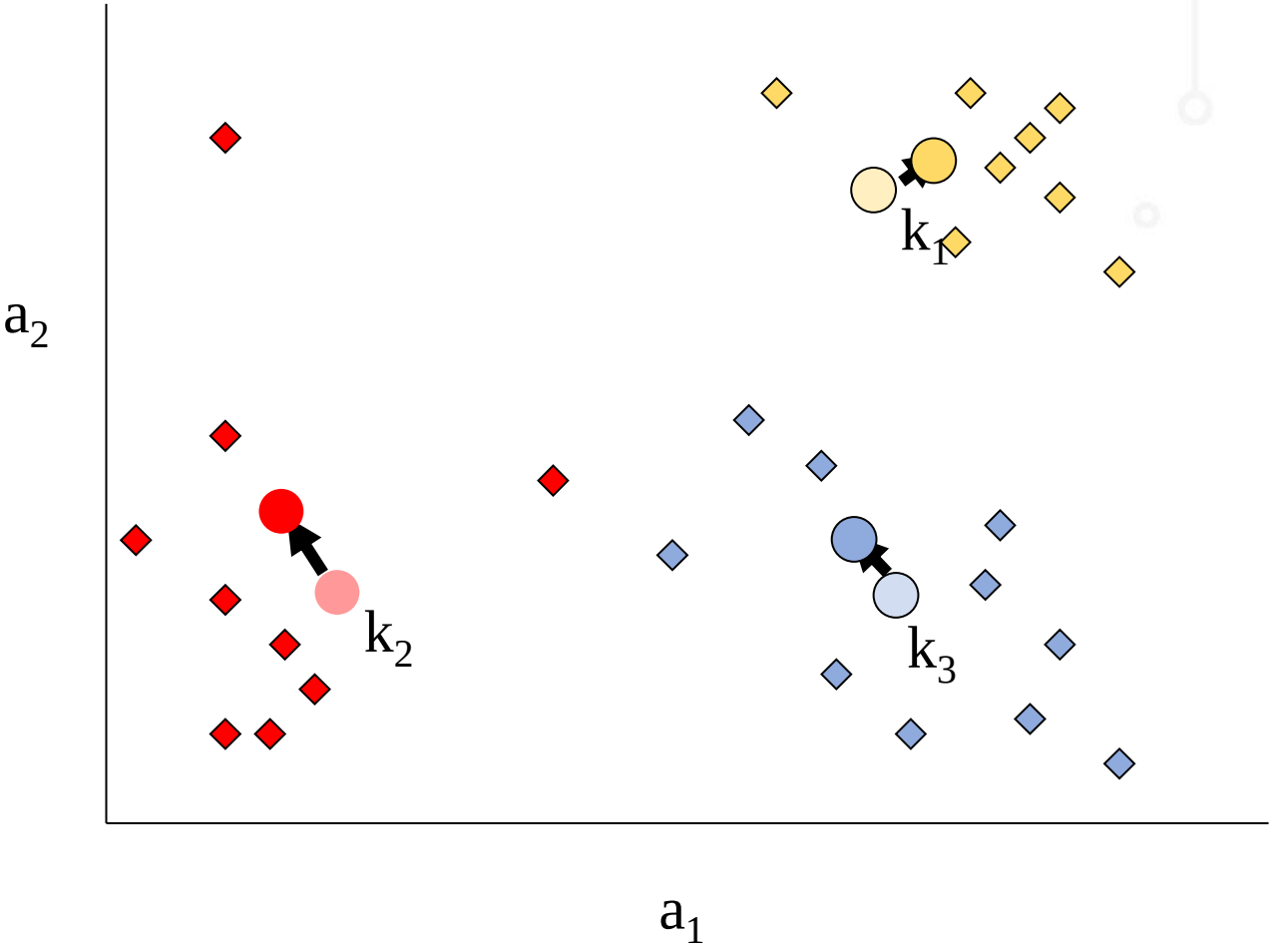
K-Means - Simulação



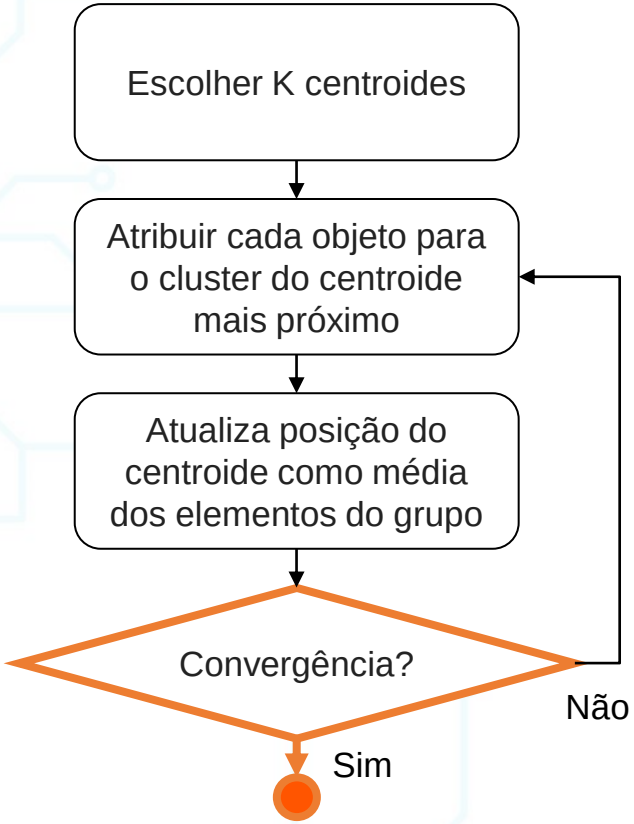


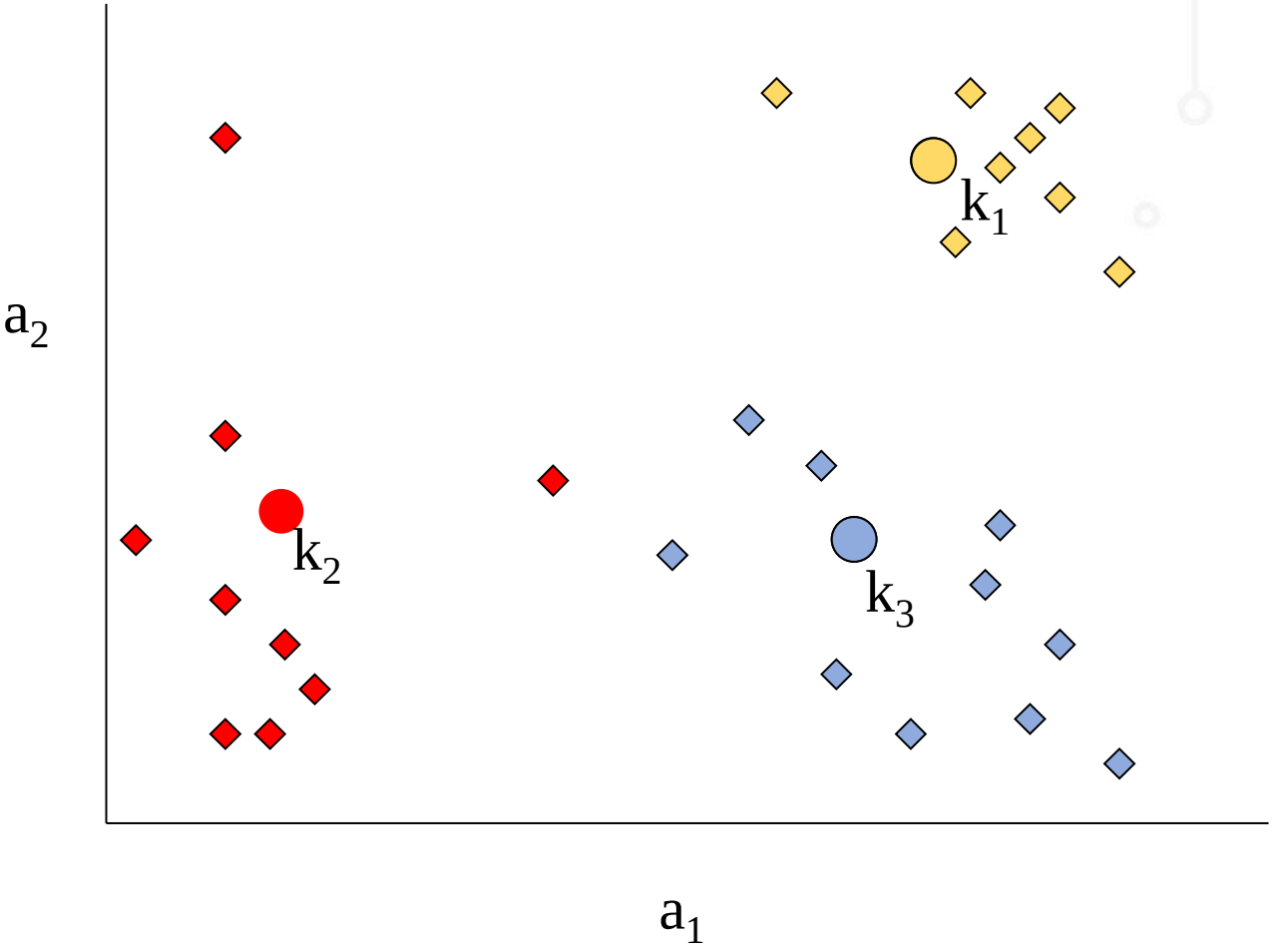
K-Means - Simulação





K-Means - Simulação





K-Means - Animação

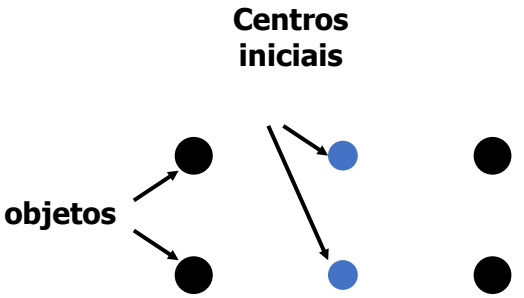
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-Means

K-Means: Sensibilidade em relação à inicialização

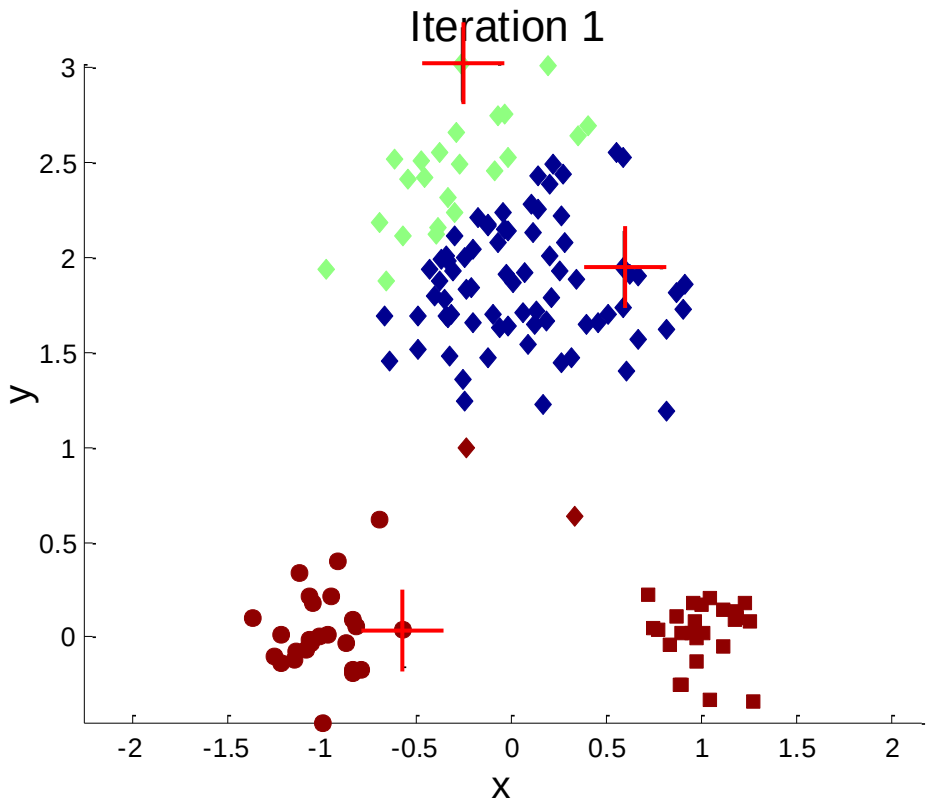
Resultado pode variar significativamente dependendo da escolha das sementes (protótipos) iniciais

k-means pode “ficar preso” em ótimos locais:



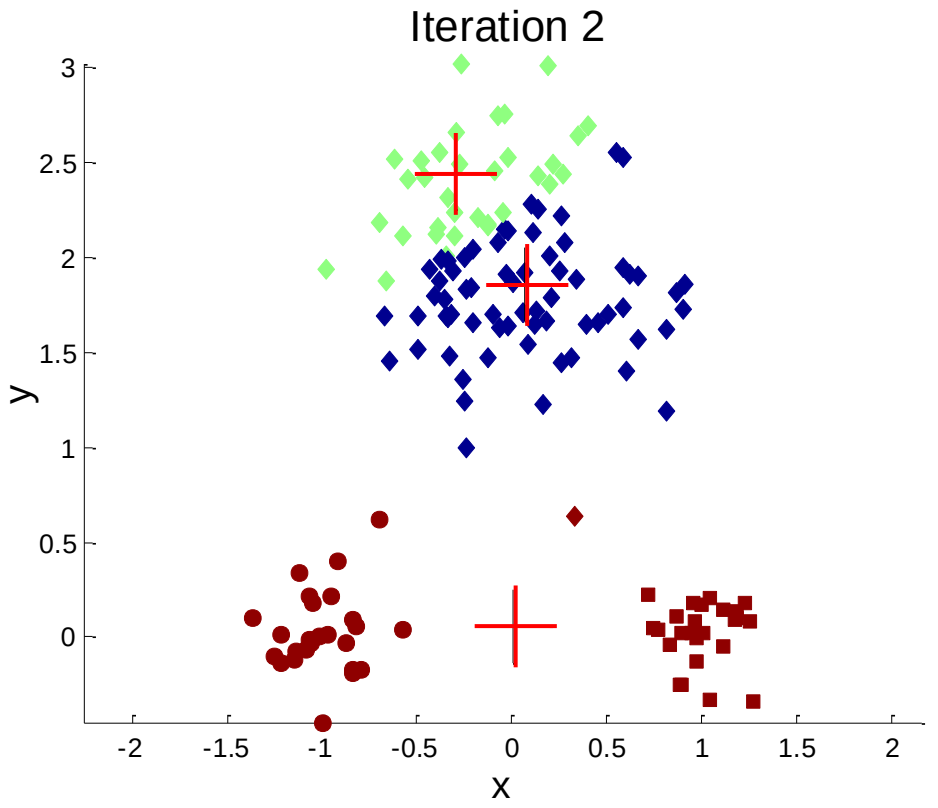
Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

K-Means: Sensibilidade em relação à inicialização



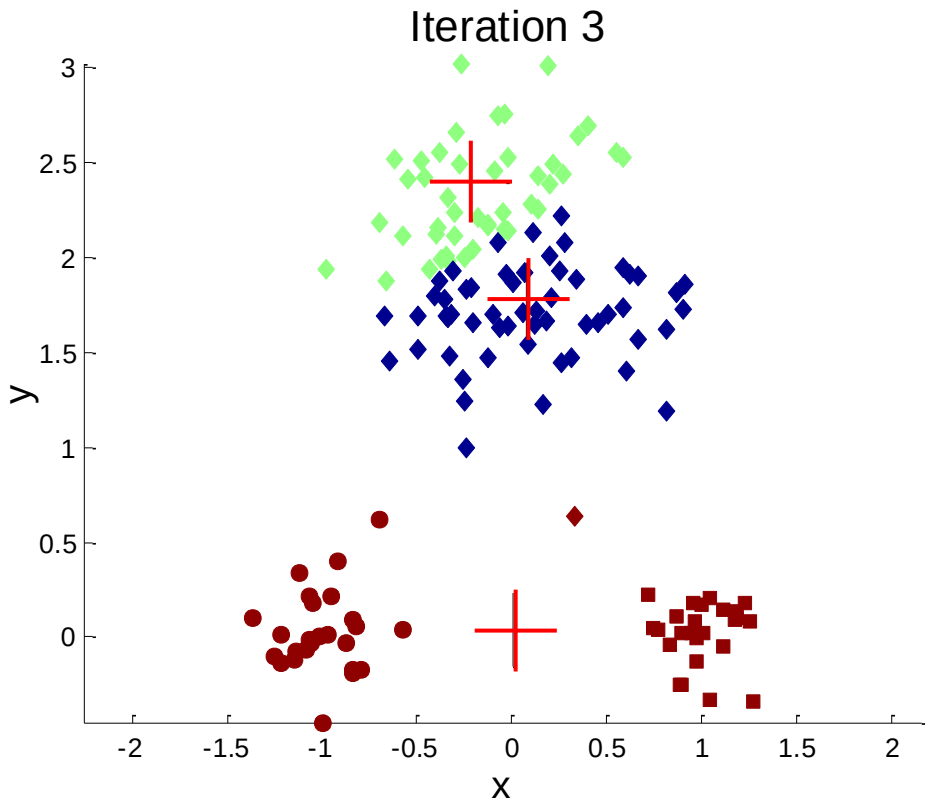
Tan, Steinbach & Kumar, Introduction to Data Mining, 2006.

K-Means: Sensibilidade em relação à inicialização



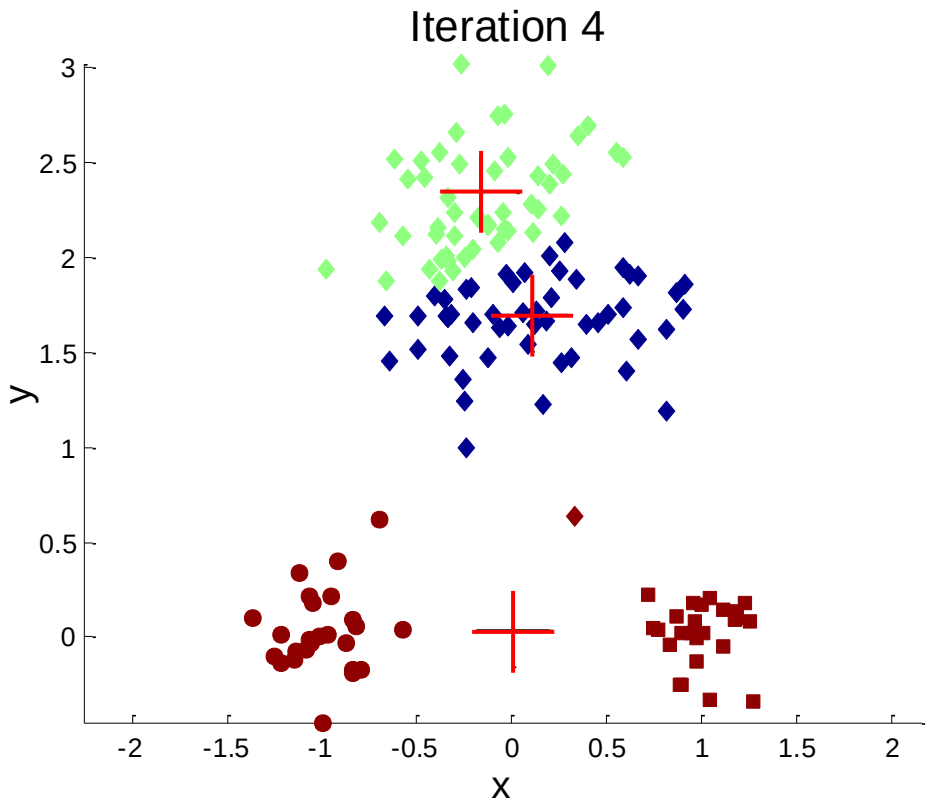
Tan, Steinbach & Kumar, Introduction to Data Mining, 2006.

K-Means: Sensibilidade em relação à inicialização



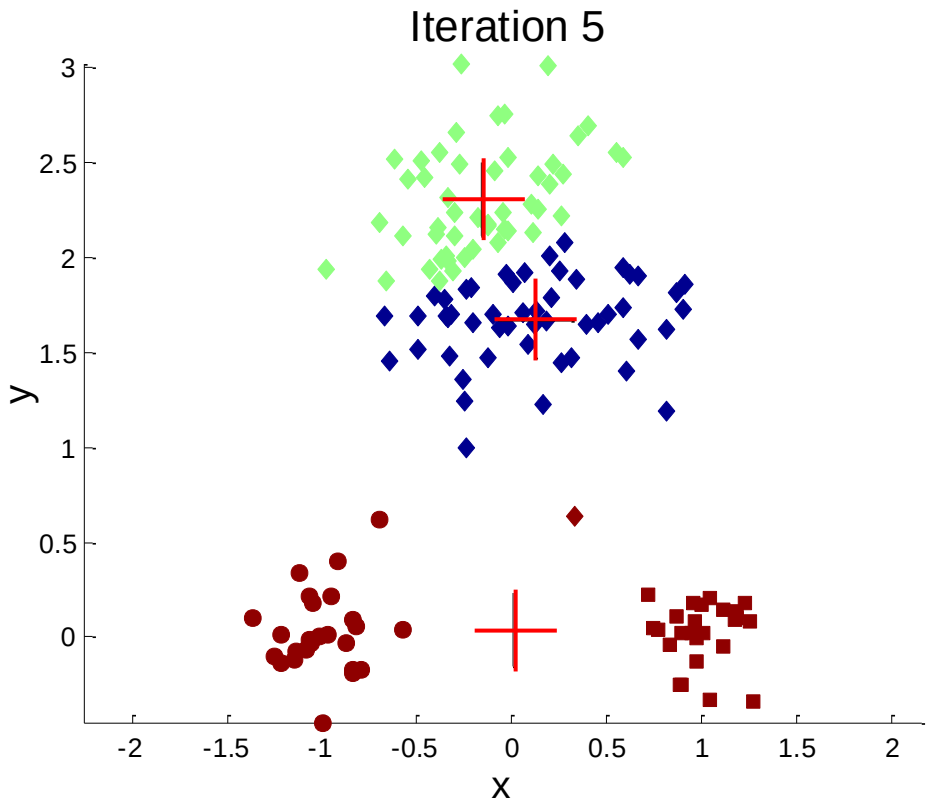
Tan, Steinbach & Kumar, Introduction to Data Mining, 2006.

K-Means: Sensibilidade em relação à inicialização



Tan, Steinbach & Kumar, Introduction to Data Mining, 2006.

K-Means: Sensibilidade em relação à inicialização



Tan, Steinbach & Kumar, Introduction to Data Mining, 2006.

K-Means: Sensibilidade em relação à inicialização

Múltiplas Execuções (inicializações aleatórias):

- Funciona bem em muitos problemas;
- Pode demandar muitas execuções (especialmente com *k* alto).

Agrupamento Hierárquico:

- agrupa-se uma amostra dos dados para tomar os centros da partição com *k* grupos.

Seleção “informada” em uma amostra dos dados:

- Tomar o 1º protótipo como um objeto aleatório ou como o centro dos dados (*grand mean*);
- Sucessivamente escolhe-se o próximo protótipo como o objeto mais distante dos protótipos correntes.

Busca Guiada:

- X-means, *k*-means evolutivo, ...

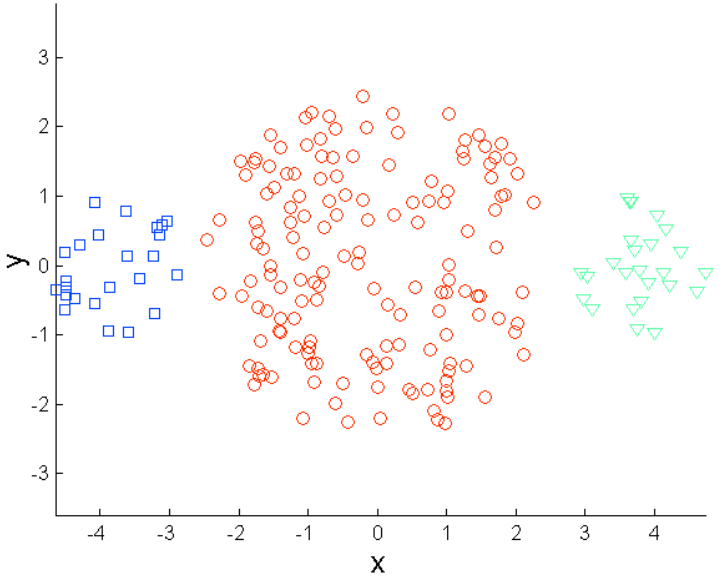
K-Means

K-Means: Problemas estruturais

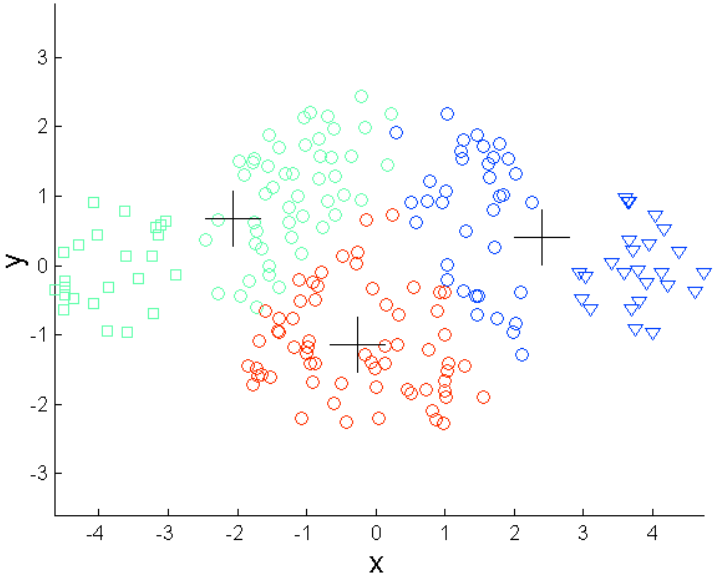
Algoritmo *k*-means funciona bem se:

- Clusters são (hiper)esféricos e bem separados
- Clusters de volumes aproximadamente iguais
- Cluster com quantidades de pontos semelhantes
- Formas Globulares

K-Means: Problemas estruturais

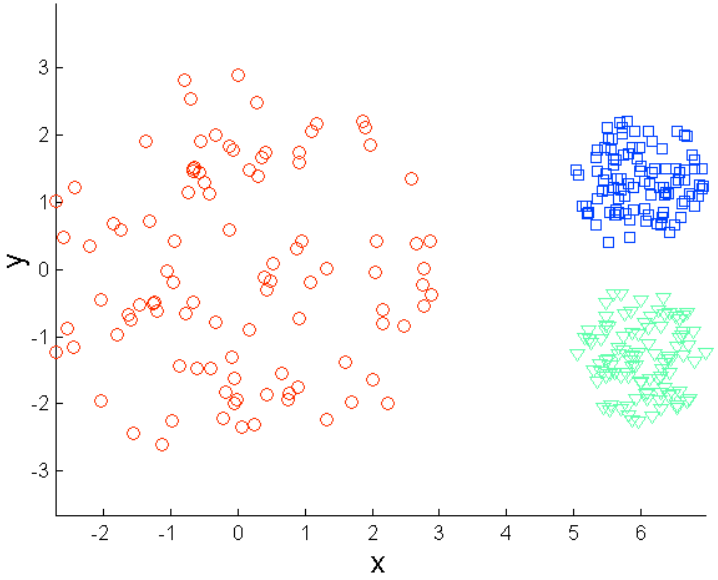


Estrutura correta

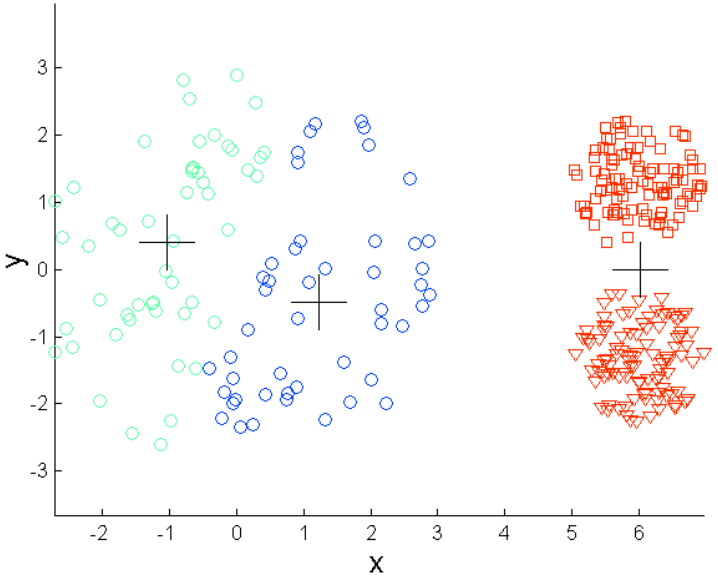


k-means (3 Clusters)

K-Means: Problemas estruturais

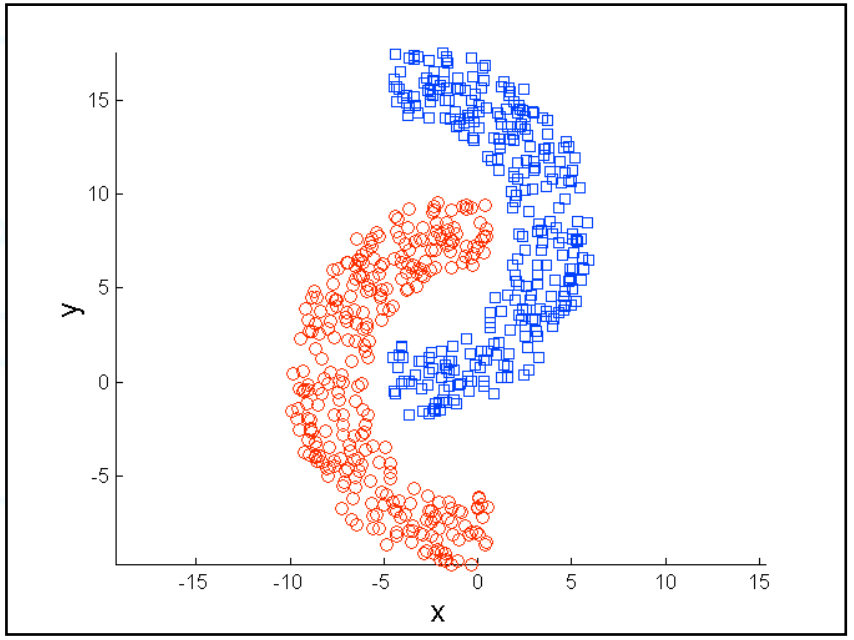


Estrutura correta

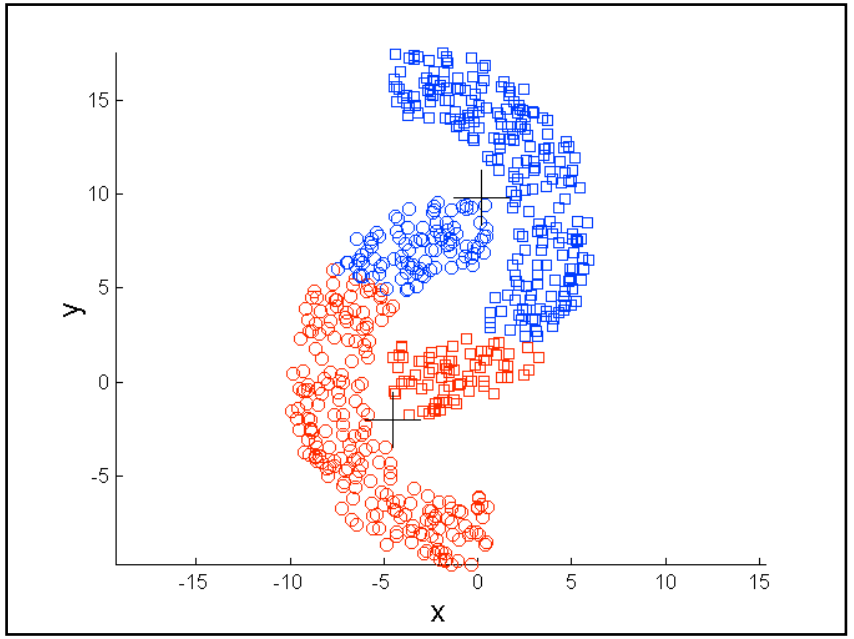


K-means (3 Clusters)

K-Means: Problemas estruturais



Estrutura correta



K-means (3 Clusters)

(Tan, Steinbach, Kumar)

Nota: na prática, esse problema em geral não é crítico, i.e., há pouco interesse na maioria das aplicações de mundo real.

K-Means: Custo Computacional

Complexidade (assintótica) de tempo:

$$O(i \cdot K \cdot N \cdot n)$$

- O que isso significa?

O que dizer sobre a constante de tempo?
→ Computar Distância Euclidiana via aproximações sucessivas (Newton-Raphson) custa caro.

Se também tenho problema de espaço em memória...
→ Solução aproximada (*sampling*).
→ Paralelizar (mesmo computador) ou distribuir (e.g., map-reduce) o processamento.

Resumo das (des)vantagens do k-means

Vantagens

- Simples e intuitivo
- Complexidade **linear** em todas as variáveis críticas
- Eficaz em muitos cenários de aplicação
- Resultados de interpretação simples

Desvantagens

- k = ?
- Sensível à inicialização dos protótipos (mínimos locais de J)
- Limita-se a encontrar clusters volumétricos / globulares
- Cada item deve pertencer a um único cluster (**partição rígida**)
- Limitado a atributos numéricos
- Sensível a *outliers*

K-Medianas

K-medianas: Substituir as médias pelas medianas

- Média de 1, 3, 5, 7, 9 é 5
- Média de 1, 3, 5, 7, 1009 é 205
- Mediana de 1, 3, 5, 7, 1009 é 5

Vantagem: menos sensível a outliers

Desvantagem: implementação mais complexa
cálculo da mediana em cada atributo...

K-Medóides

K-medóides: Substituir cada centróide por um objeto representativo do cluster, denominado **medóide**

- Medóide = objeto mais próximo aos demais objetos do cluster mais próximo em média (empates resolvidos aleatoriamente)

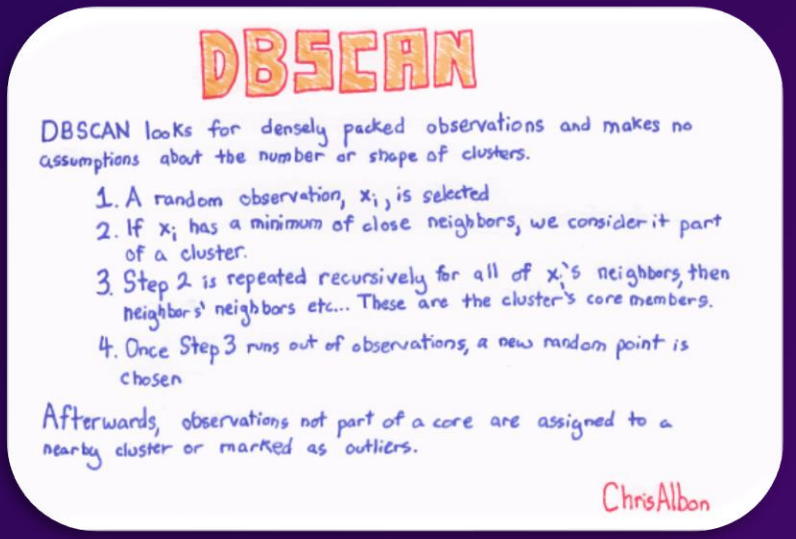
Vantagens:

- menos sensível a outliers
- permite cálculo relacional (apenas matriz de distâncias)
  - logo, pode ser aplicado a bases com atributos categóricos
- convergência assegurada com qualquer medida de (dis)similaridade

Desvantagem: Complexidade quadrática com no. de objetos (N)

VINI CARIDÁ
Human-Centered Data & AI

DB Scan

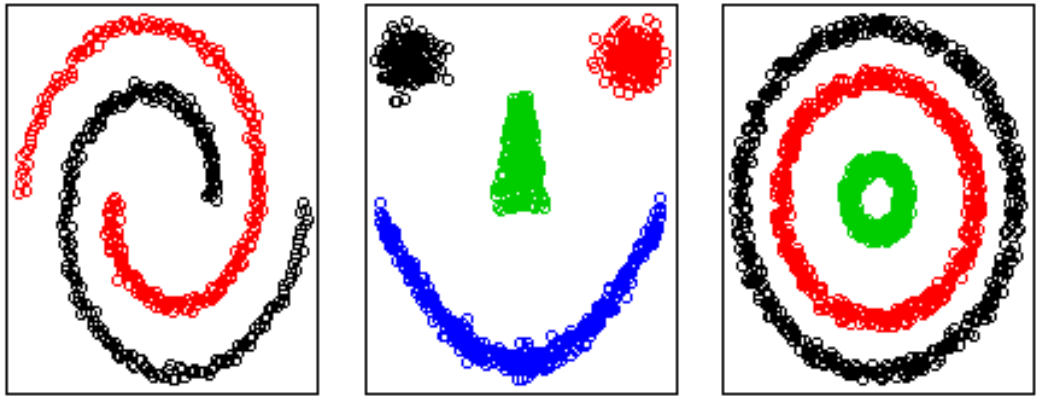


Algoritmos Baseados em Densidade

Paradigma de Agrupamento por Densidade

- Clusters como regiões de alta concentração de objetos separadas por regiões de baixa concentração de objetos
- Paradigma alternativo àquele baseado em protótipos: K-means e variantes, EM, etc

Existem vários algoritmos, veremos a seguir um dos mais conhecidos: **DBSCAN**

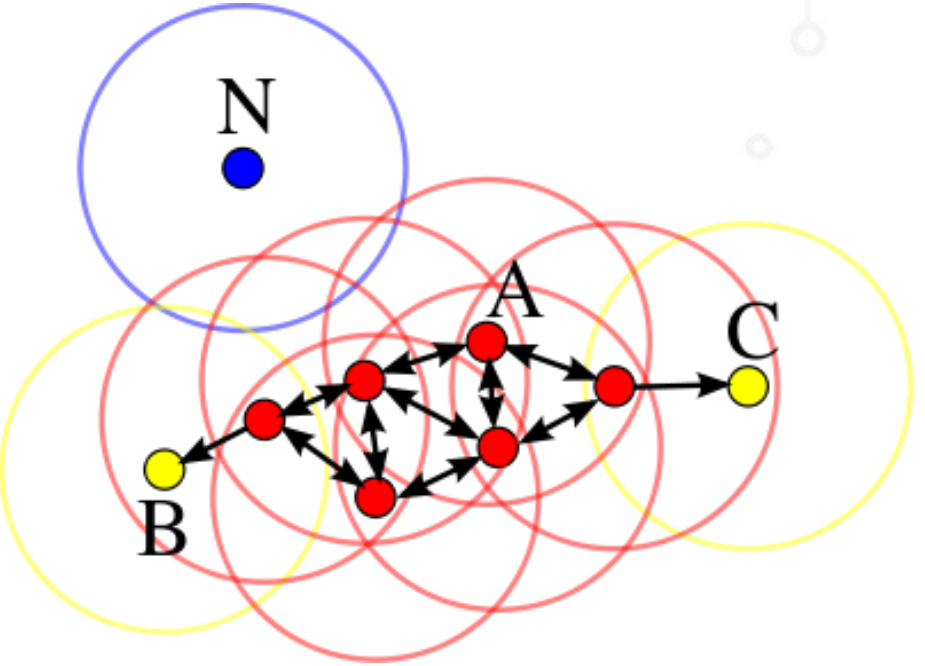


DBScan: definições

- A point is a **core** point if it has at least a specified number of points (MinPts) within the radius Eps (including the point itself)
  - These are points that are in the interior of a cluster

- A **border** point has fewer than MinPts within Eps, but is in the neighborhood (within the radius) of at least 1 core point

- A **noise** point is neither a core point nor a border point



DBScan: algoritmo

Algoritmo Conceitual:

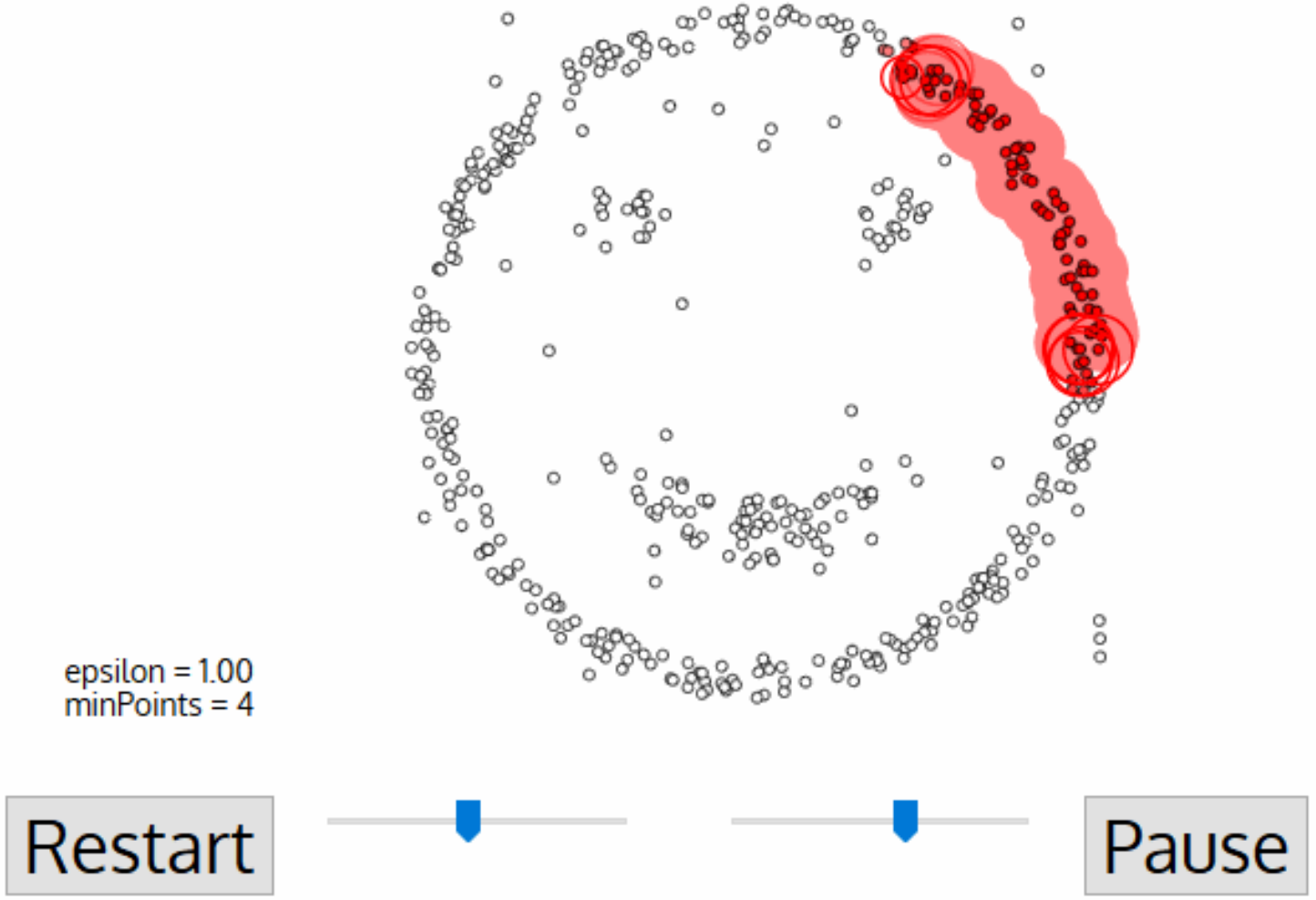
1. Percorra a BD e rotule os objetos como core, border ou noise
2. Elimine aqueles objetos rotulados como **noise**
3. Insira uma aresta entre cada par de objetos **core** vizinhos

- 2 objetos são vizinhos se um estiver dentro do raio Eps do outro

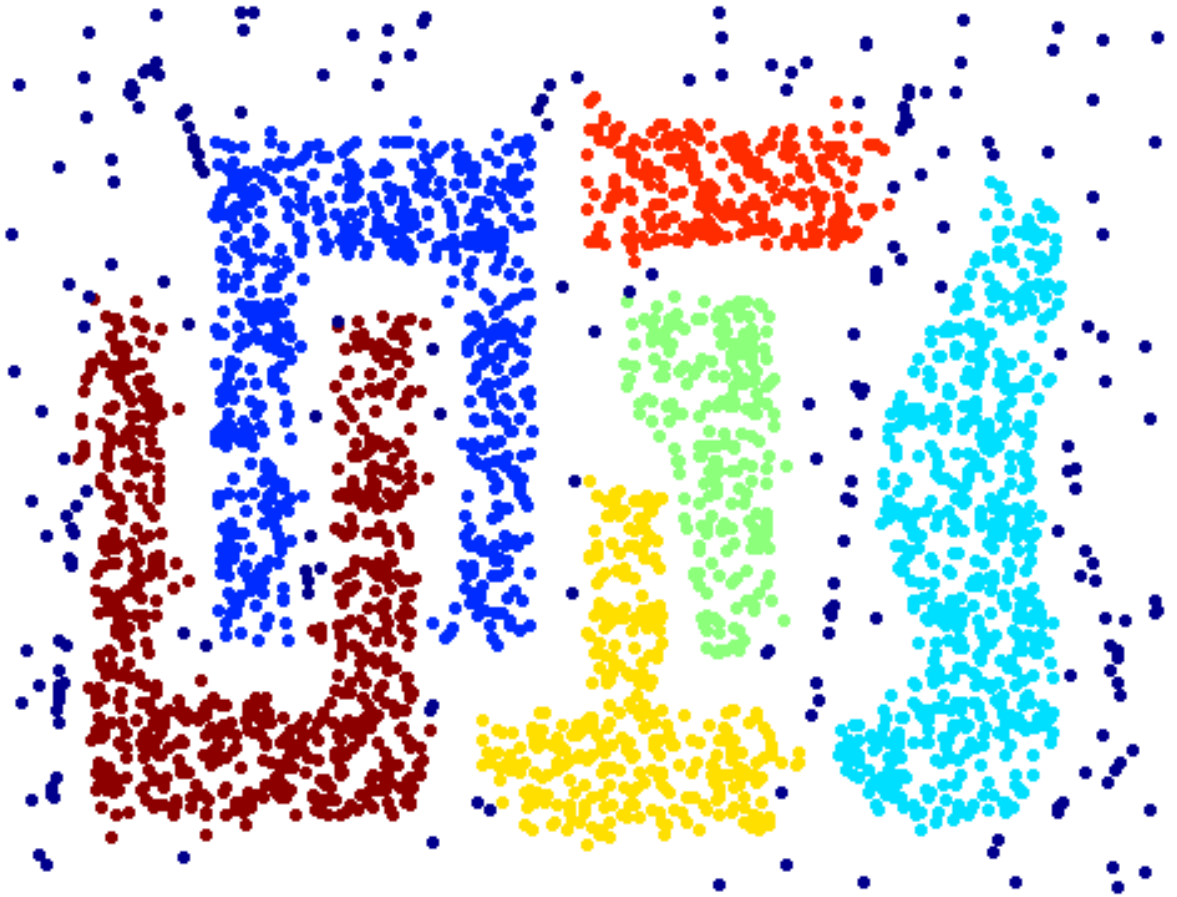
4. Faça cada componente conexo resultante ser um cluster
5. Atribua cada **border** ao cluster de um de seus core associados

- Resolva empates se houver objetos core associados de diferentes clusters

DBScan: algoritmo



DBScan: Exemplo



Point types: **core**, **border** and **noise**

Resumo das (des)vantagens do DBScan

VINI CARIDÁ
Human-Centered Data & AI

Vantagens

- Não necessita do número de clusters a priori
- Consegue encontrar clusters com formatos arbitrários
- Tem uma definição de ruído e é robusto a outliers
- Necessita de apenas dois parametros:
  - Raio
  - Número de vizinhos para virar core (minpts)

Desvantagens

- Extremamente sensível aos parametros Raio e minPts
- Depende da distância utilizada para determiner se um ponto está ou nao presente dentro do raio. (tipicamente se utiliza euclidiana)
- Não consegue clusterizar dados com grupos com grandes diferenças de densidades
- Se a escala dos dados não for conhecida, desterminar o raio pode ser difícil

Thanks !

Vinicius Fernandes Caridá

vfcarida@gmail.com


@vinicius caridá


@vfcarida


@vinicius caridá


@vfcarida


@vinicius caridá


@vfcarida