Algoritmos Baseados em Instâncias − Agrupamento

Paulo Trigo Silva

Agrupamento *vs* Classificação − exemplo dos documentos

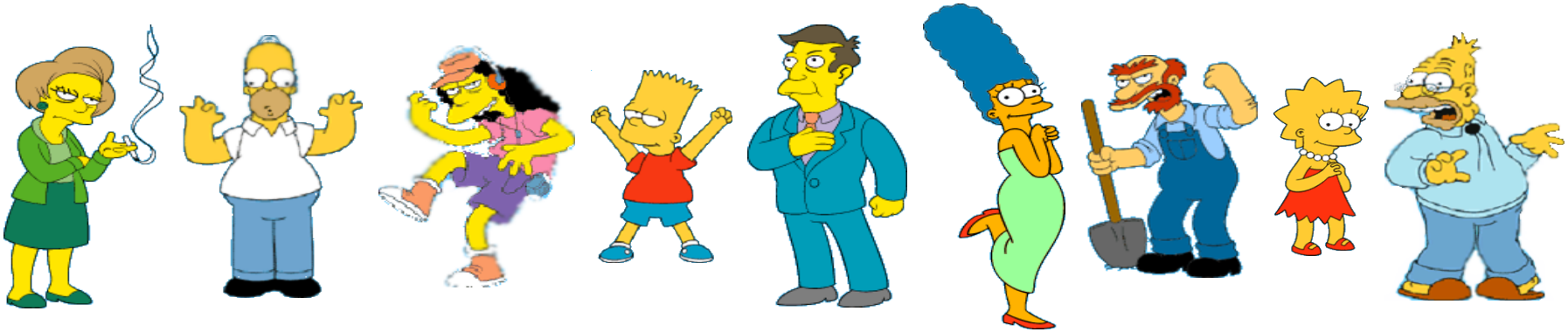
Agrupamento, processo **não-supervisionado** no qual as instâncias (documentos) são agrupadas de acordo com a sua semelhança.

Classificação, processo **supervisionado** (de **catalogação**) que atribui a cada objeto (documento) uma categoria pré-definida

	Categorias	Conhecimento	Decisão
Catalogação	estáticas	à-priori	baseada nos termos
Agrupamento	dinâmicas	após processo	baseada na “semelhança”

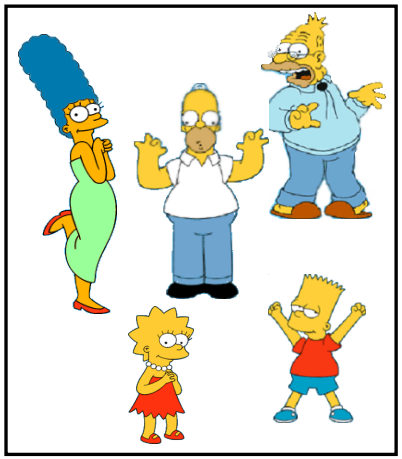
Paulo Trigo Silva

Agrupamento – qual a forma “natural” de agrupar objetos?



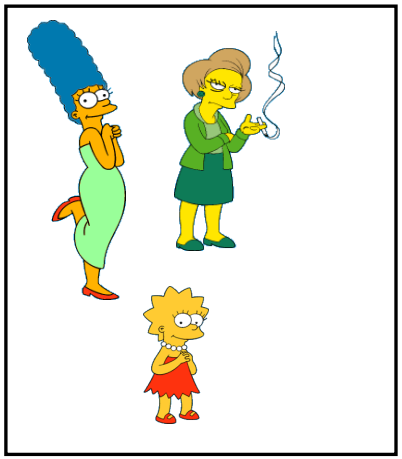
Qual a forma “natural” de agrupar?

Cada agrupamentos tem objetos “semelhantes”! noção “subjetiva”



Família Simpson

Empregados Escola



Feminino

Masculino

Paulo Trigo Silva

Semelhança – o que é?

> The quality or state of being similar; likeness; resemblance; as a similarity of features.

in Webster's Dictionary

Semelhança é difícil de definir, mas todos nós sabemos quando encontramos objetos semelhantes…

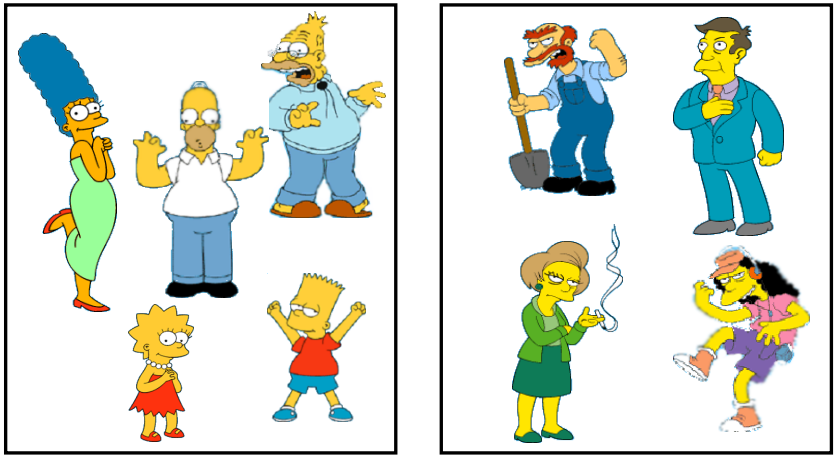
Paulo Trigo Silva

Agrupamento

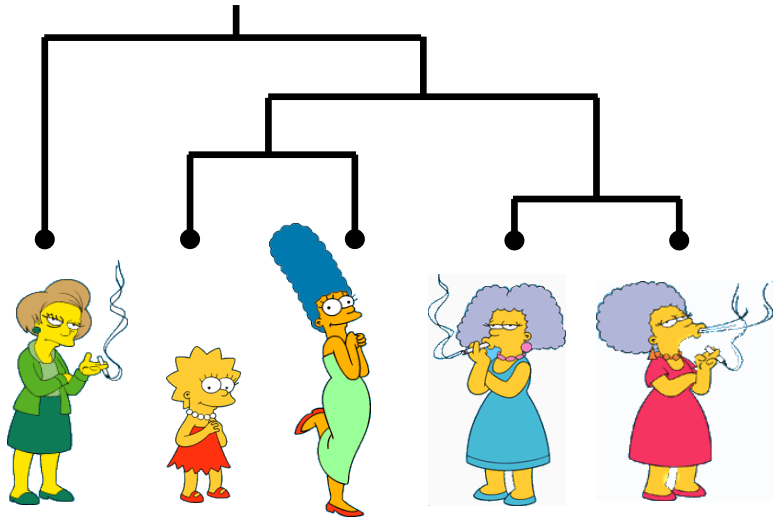
Não-Hierárquico (**partição**), criam-se partições mutuamente exclusivas de todo o espaço de objetos

Hierárquico (**decomposição**), cria-se uma árvore onde cada nível é constituído por partições mutuamente exclusivas

Não-Hierárquico (Partição)

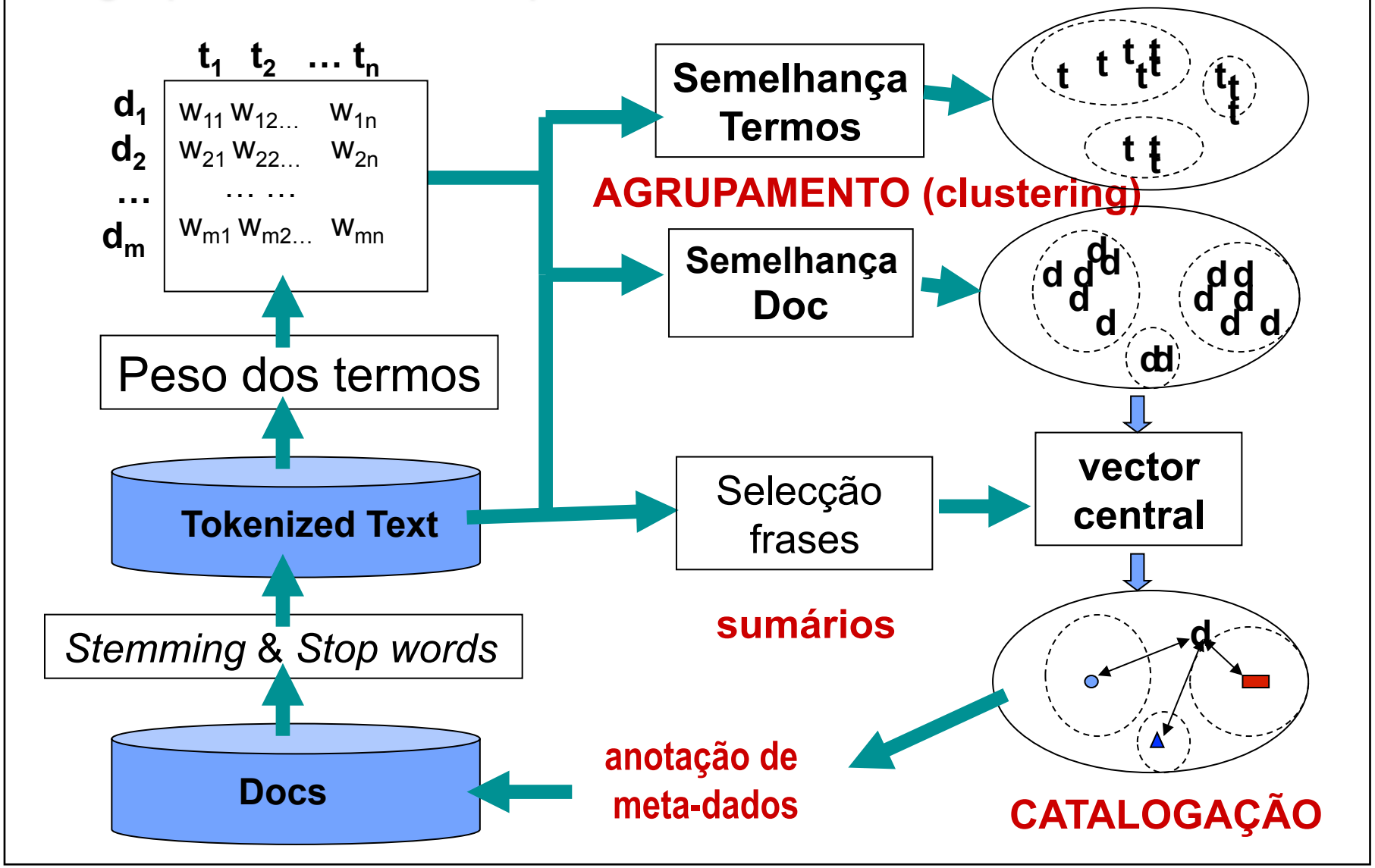


Hierárquico



Paulo Trigo Silva

Agrupamento – exemplo dos documentos



Agrupamento – baseados na “semelhança”

- Definir uma **função de semelhança**
  - e medir a semelhança entre dois objetos
- Objetivo da procura de partições
  - **minimizar** semelhança entre **agrupamentos distintos**
  - **maximizar** semelhança entre **indivíduos do mesmo agrupamento**
- ... duas abordagens para procurar as partições
  - iniciar de modo aleatório e adaptar iterativamente (e.g., **K-means**)
  - iniciar cada objeto como um grupo (“cluster”) e juntar de modo recursivo pares de “clusters” (e.g., **hierárquico aglomerativo**)

Paulo Trigo Silva

K-means (baseados na “semelhança”)

- Dada uma função de semelhança

$$d\left(\overrightarrow{x},\overrightarrow{y}\right)=\left[\sum_{i=1}^{n}\left(x_i-z_i\right)^2\right]^{\frac{1}{2}}$$ (distância Euclidiana)

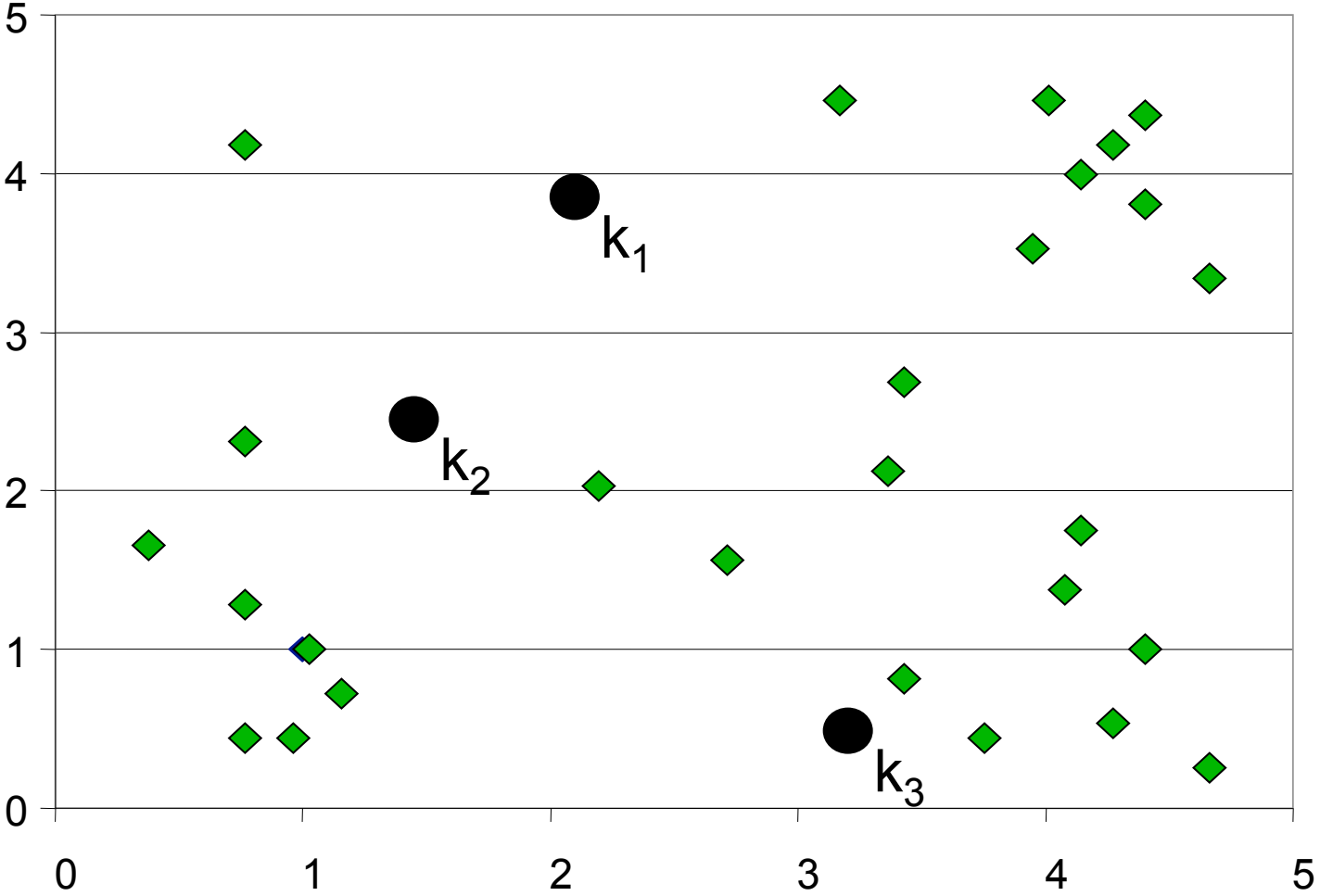
- Começar com k pontos escolhidos aleatoriamente (1º passo), assumindo que todos são centros de k agrupamentos
- Atribuir a cada indivíduo o ponto (de entre os k) que lhe é mais semelhante (que lhe está mais próximo)
- Recalcular o centro de cada agrupamento (centróide) $c_j=\frac{1}{|C_j|}\sum_{x\in C_j}x$
- Repetir o processo até o movimento dos centróides ser “desprezável”

função erro $E=\sum_{j=1}^{k}\sum_{x\in C_j}d(x,c_j)^2$ condição de paragem: $|E^{new}-E^{old}|<\varepsilon$

Paulo Trigo Silva

K-means (passo #1)

inicializar (aleatório) cada um dos *k* pontos

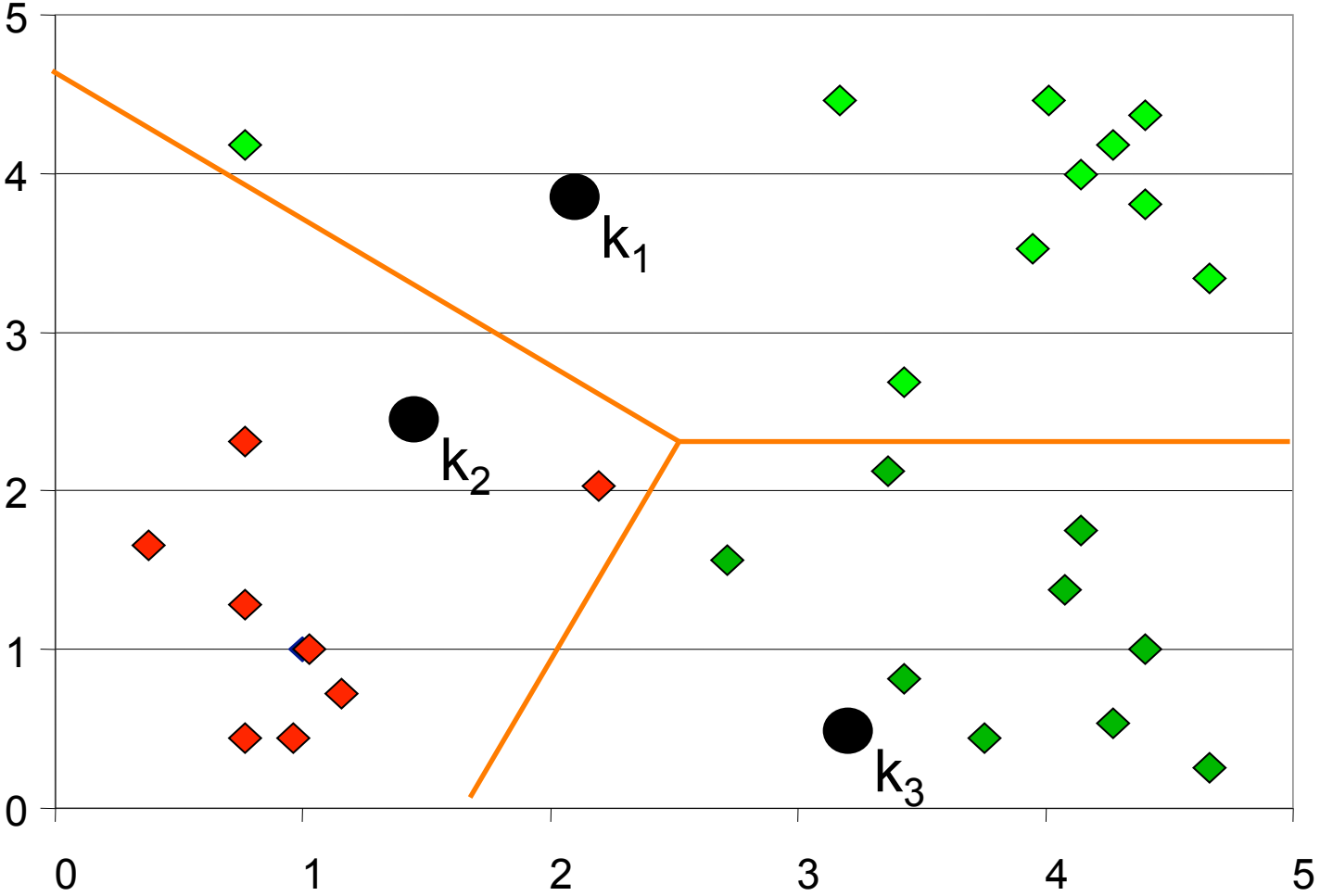


Paulo Trigo Silva

K-means (passo #2)

quais os membros de cada um dos *k* grupos?

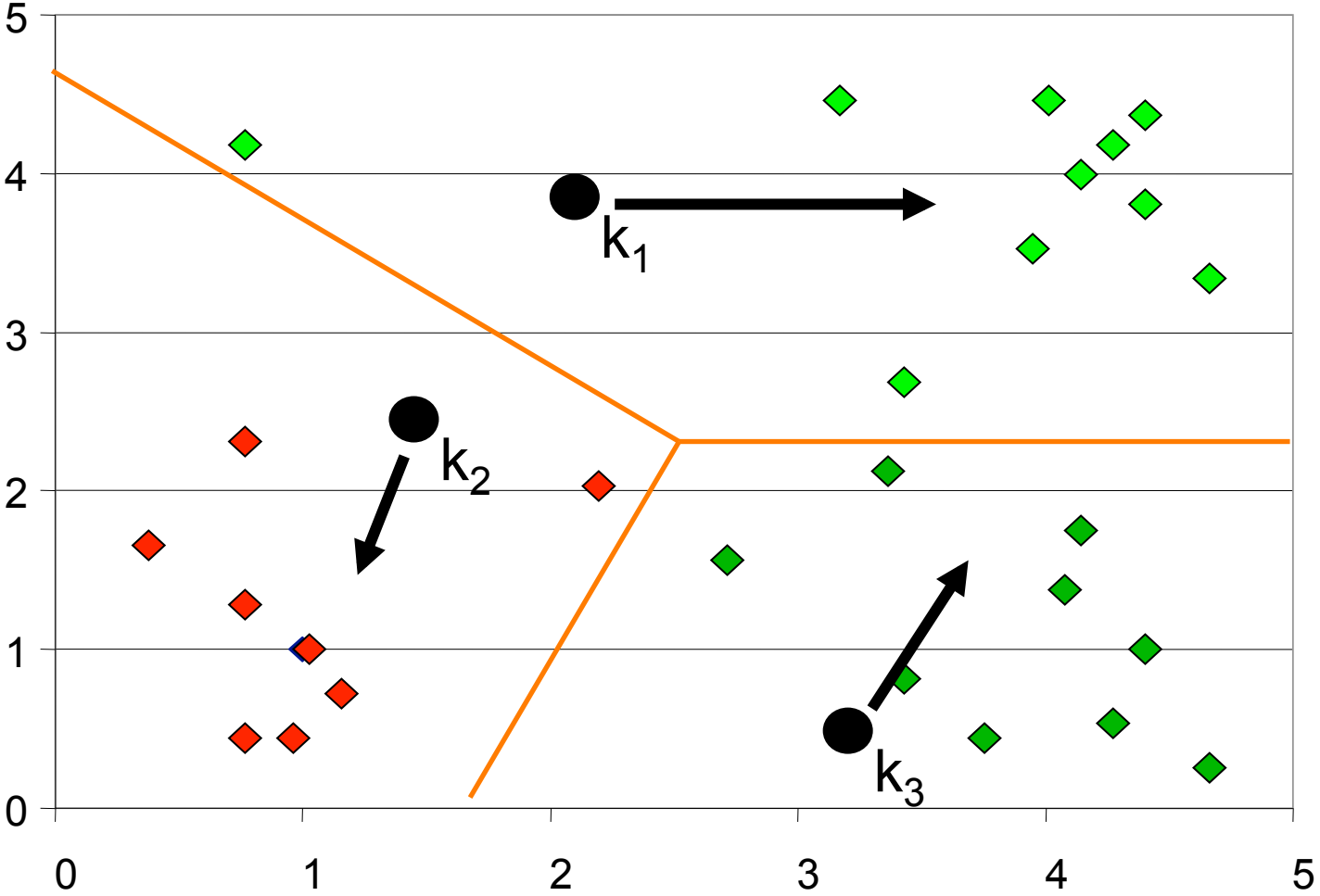
cada objeto "escolhe" o centróide (1 dos *k*) que lhe está mais "próximo"



Paulo Trigo Silva

K-means (passo #3)

mover cada um dos k pontos para o centróide do seu grupo
(estamos a usar usamos distância Euclidiana)

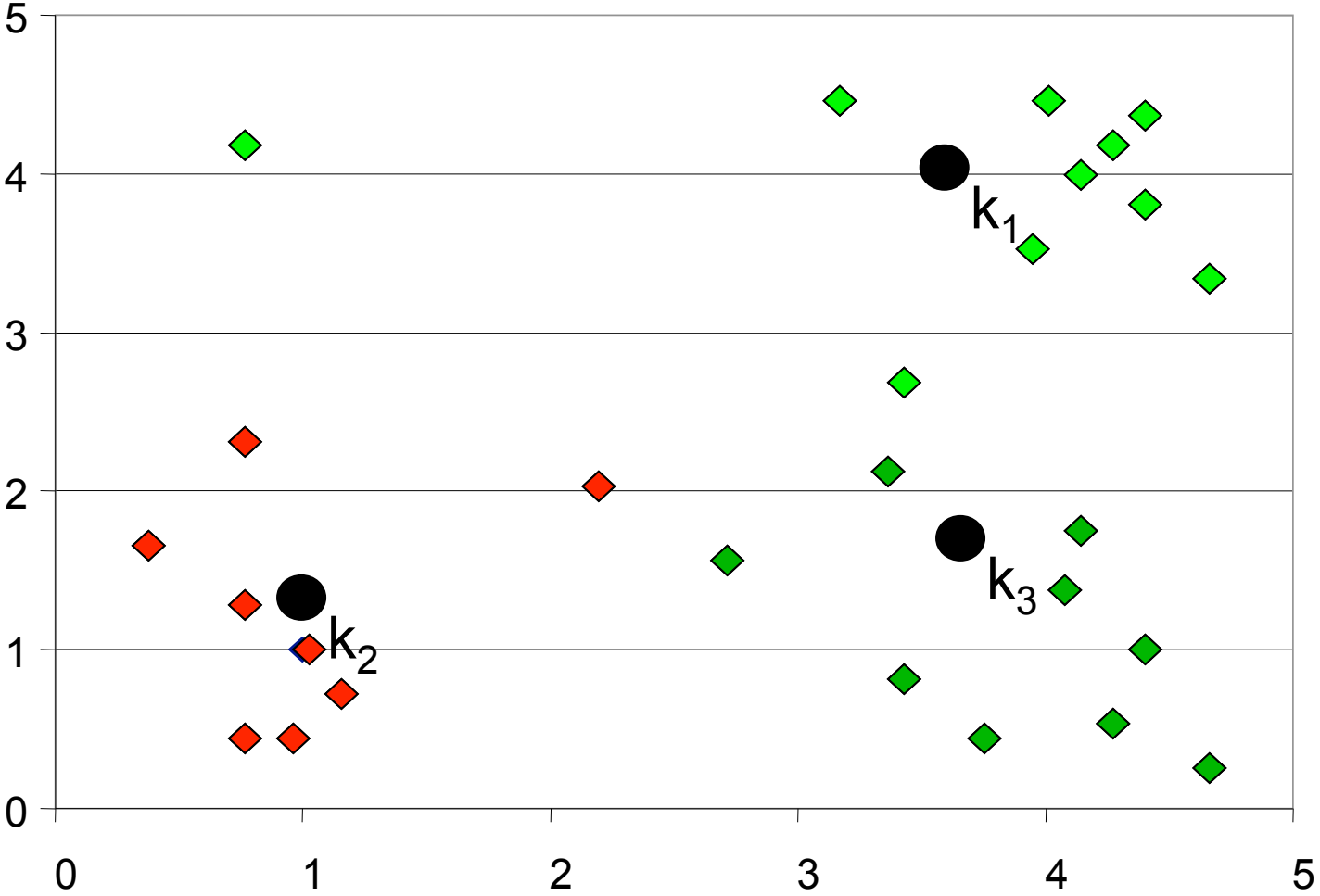


Paulo Trigo Silva

K-means (passo #4)

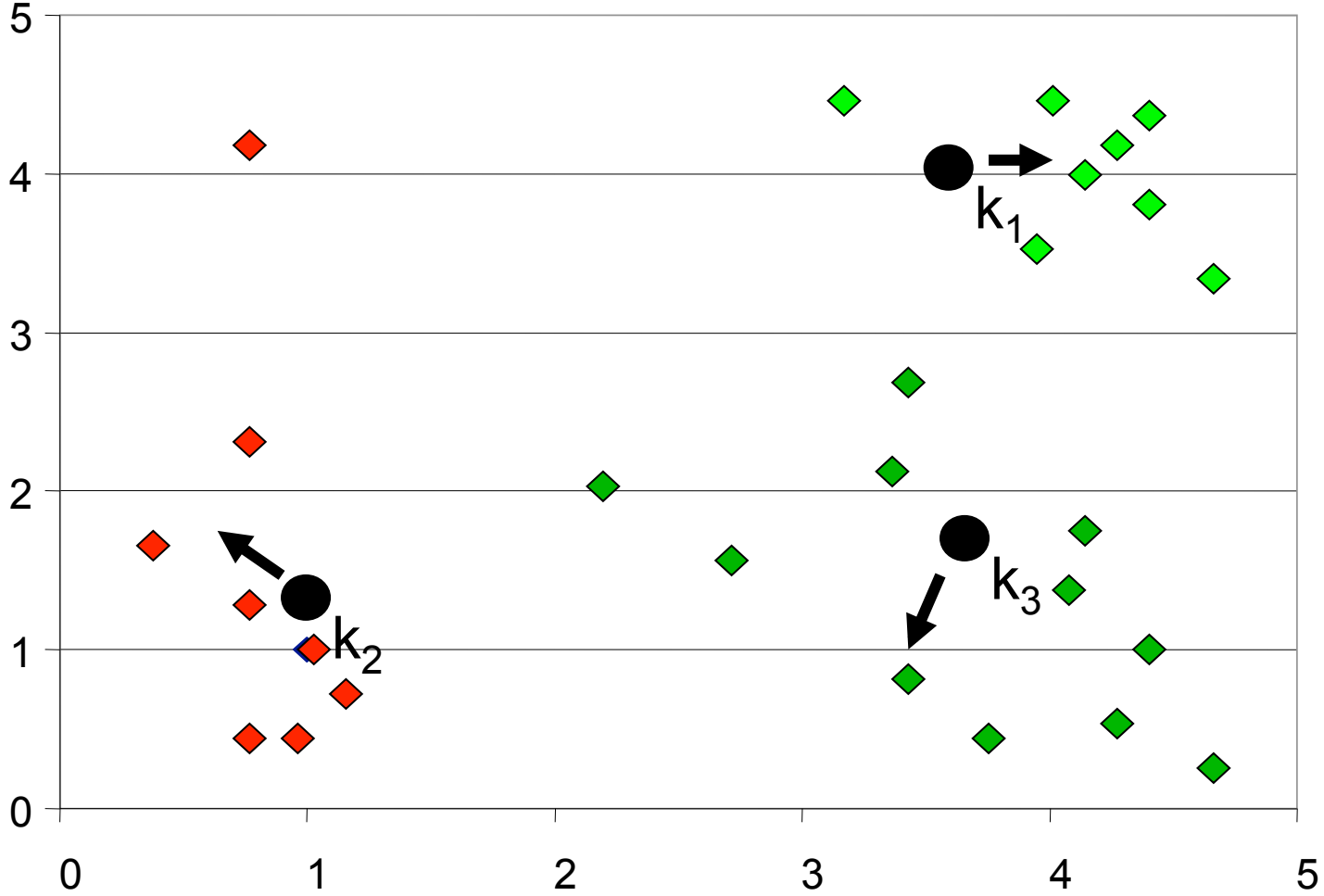
quais os membros de cada um dos *k* grupos?

cada objeto "escolhe" o centróide (1 dos *k*) que lhe está mais "próximo"



K-means (passo #5)

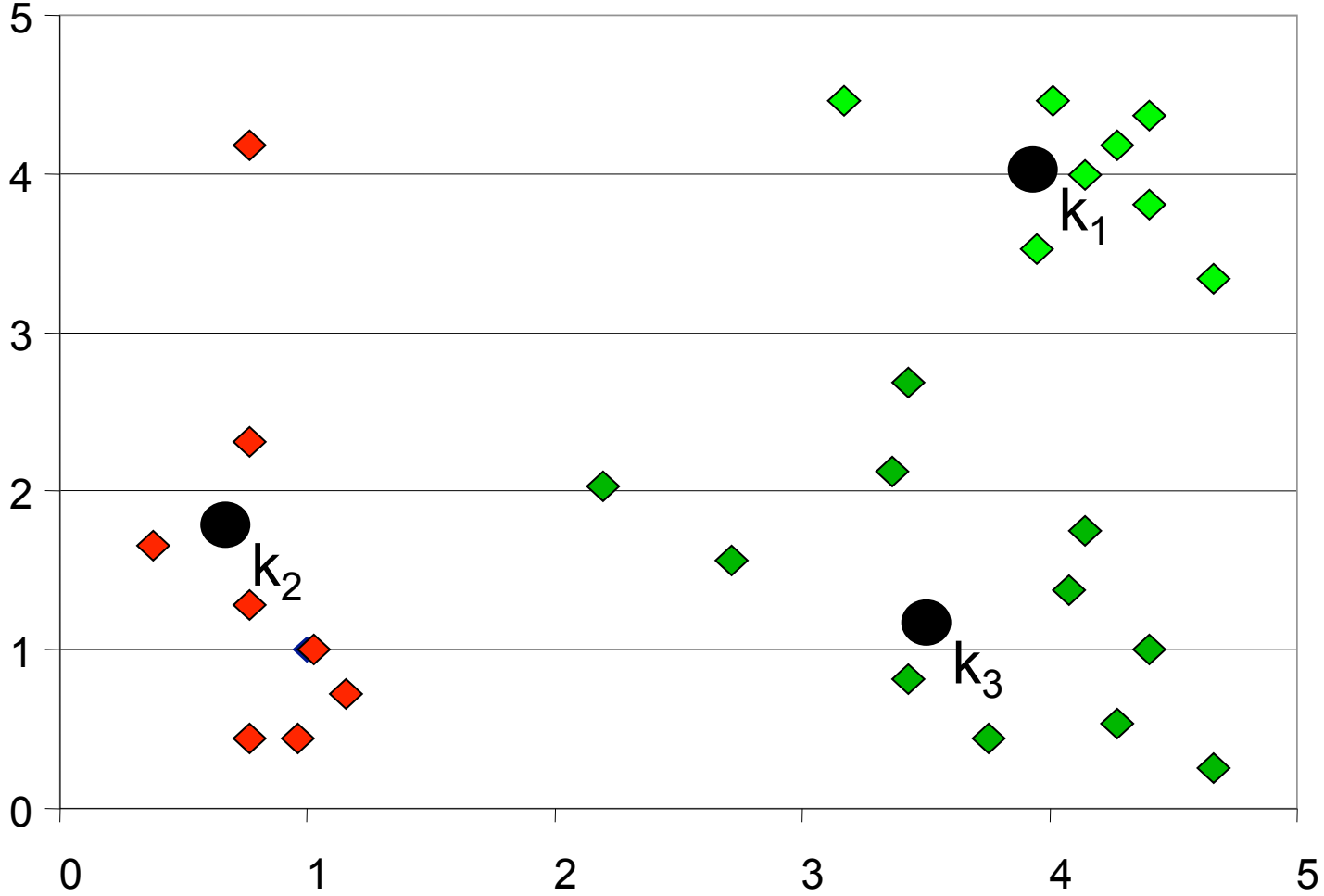
recalcular os centróides, e
mover cada um dos *k* pontos para o centróide do seu grupo



Paulo Trigo Silva

K-means (passo #6)

se não houver mudança de membros, terminar
caso contrário, repetir o passo #2



Paulo Trigo Silva

K-means – algumas características

- Vantagem
  - ordem de complexidade é O(K * N * T)
  - onde, K é #grupos, N é #objetos, e T é #iterações
  - ... habitualmente, K, T << N
  - ... o símbolo “<<“ significa “muito menor”

- Desvantagem
  - necessidade de especificar K
  - incapaz de lidar com ruído nos dados e pontos remotos
  - todos os grupos têm formas circulares (distância Euclidiana)

Paulo Trigo Silva

Agrupamento Hierárquico Aglomerativo

- construir a matriz de proximidade entre cada dois objetos
  - matriz quadrada N x N onde cada objeto tem a proximidade com todos os restantes (é matriz simétrica)
- considerar cada objeto como um grupo (“cluster”)
- juntar os 2 “clusters” mais próximos
- atualizar a matriz de proximidade
- repetir os dois passos anteriores até termos 1 único “cluster”

- ... existem diferentes formas de calcular semelhanças entre grupos(“clusters”) baseados na semelhança entre objetos

Paulo Trigo Silva

Agrupamento Hierárquico Aglomerativo (passo #1)

vamos admitir que temos 6 pontos (objetos): A, B, C, D, E F

considerar que cada ponto é um “cluster”: [A], [B], [C], [D], [E], [F]

calcular a matriz de distâncias (de cada “cluster” a todos os outros)

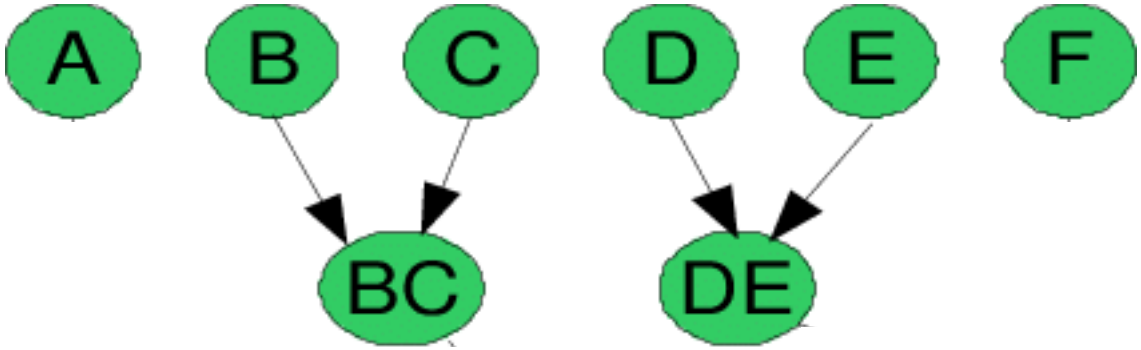
Paulo Trigo Silva

Agrupamento Hierárquico Aglomerativo (passo #2)

calcular a matriz de distâncias (de cada ponto a todos os outros)

vamos admitir que se vão juntar os “clusters” [B, C] e [D, E]

temos agora os 4 “clusters”: [A], [B, C], [D, E], [F]



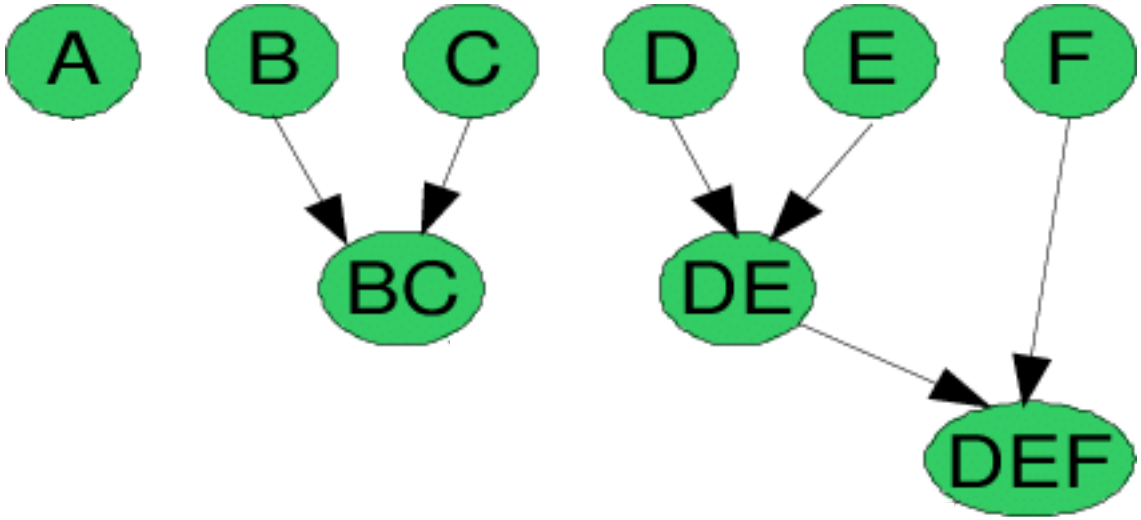
Paulo Trigo Silva

Agrupamento Hierárquico Aglomerativo (passo #3)

(re)calcular a matriz de distâncias (de cada ponto a todos os outros)

vamos admitir que se vão juntar os “clusters” [D, E] e [F]

temos agora os 3 “clusters”: [A], [B, C], [D, E, F]



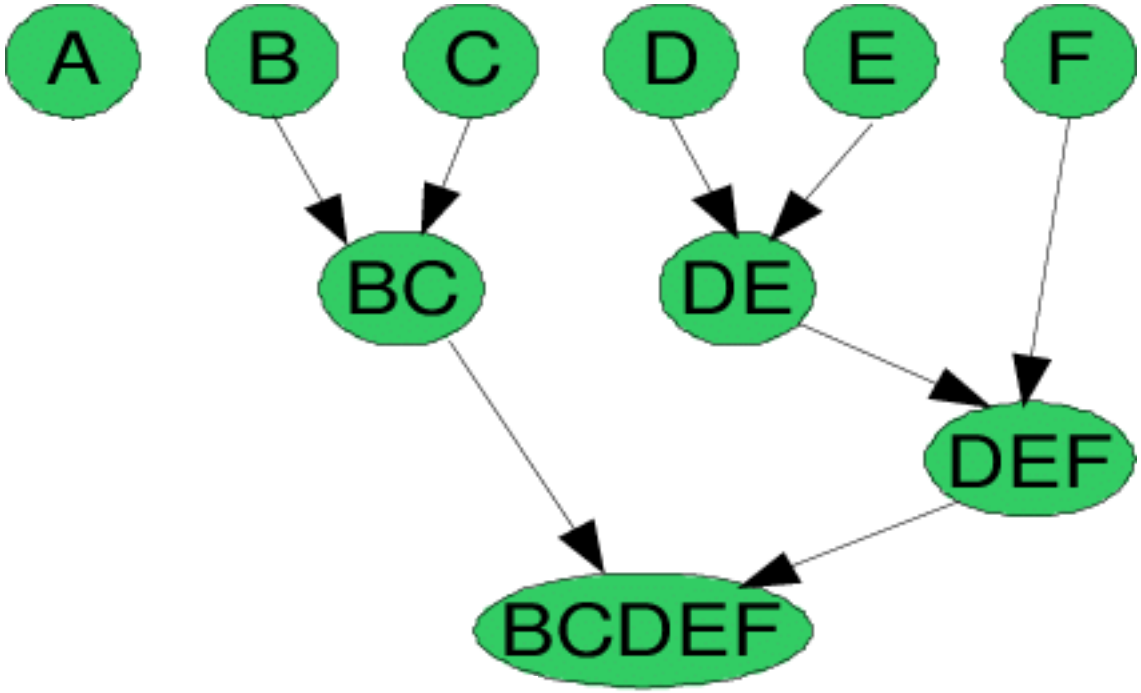
Paulo Trigo Silva

Agrupamento Hierárquico Aglomerativo (passo #4)

(re)calcular a matriz de distâncias (de cada ponto a todos os outros)

vamos admitir que se vão juntar os “clusters” [B, C] e [D, E, F]

temos agora os 2 “clusters”: [A], [B, C, D, E, F]

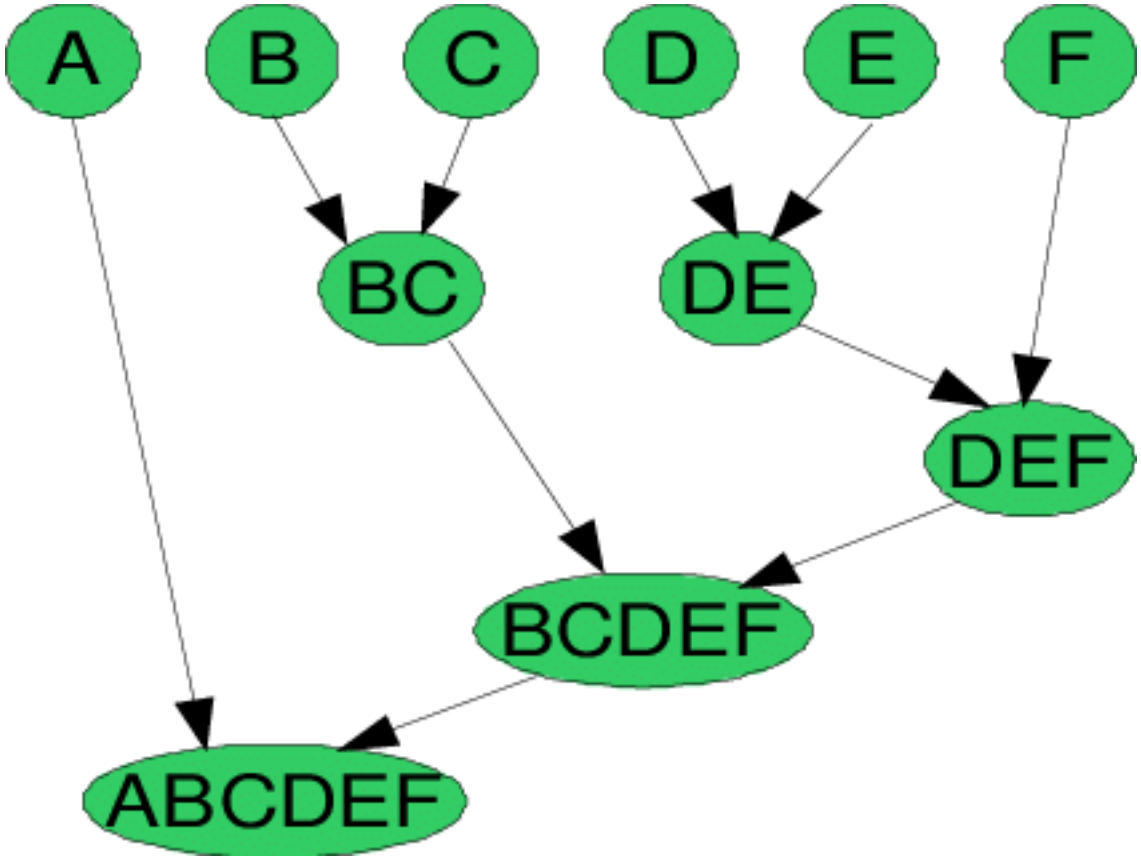


Paulo Trigo Silva

Agrupamento Hierárquico Aglomerativo (passo #5)

(re)calcular a matriz de distâncias (de cada ponto a todos os outros)

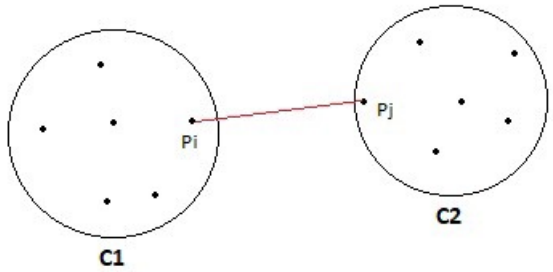
por fim vão-se juntar os “clusters” [A] e [B, C, D, E, F]
que constitui o “cluster” raiz da árvore: [A, B, C, D, E, F]



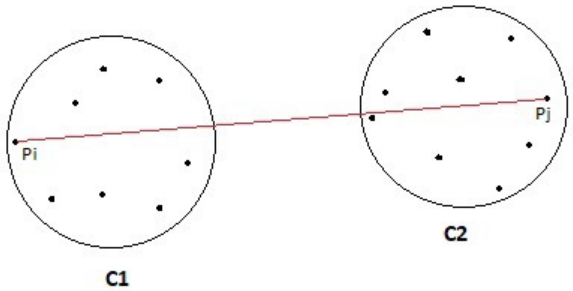
Paulo Trigo Silva

Como calcular semelhança entre 2 grupos (“clusters”)?

- **MIN** (“single linkage”), semelhança (*sem*) entre C1 e C2 é:
  - igual ao mínimo das distâncias entre quaisquer pontos Pi e Pj,
  - onde, Pi pertence ao “cluster” C1 e Pj pertence ao “cluster” C2
  - ... *sem*(C1, C2) = min *sem*(Pi, Pj), para todos Pi ∈ C1, Pj ∈ C2



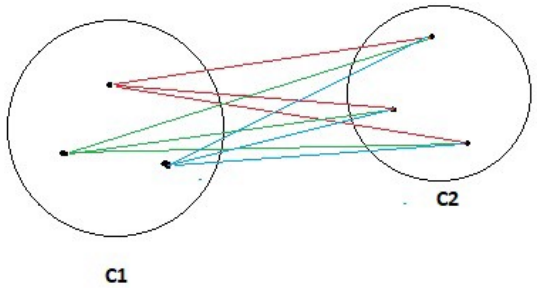
- **MAX** (“complete linkage”), considera MAX em vez de MIN



Paulo Trigo Silva

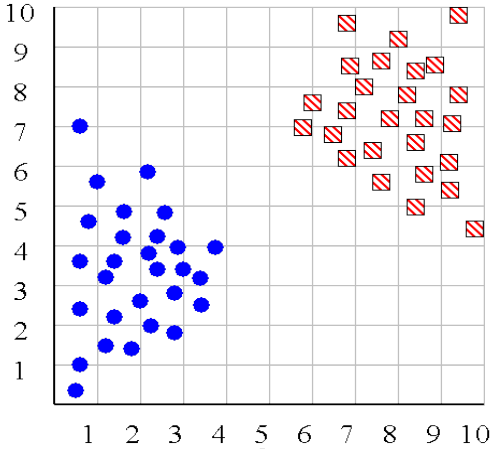
... como calcular semelhança entre 2 grupos (“clusters”)?

- **Group Average**, semelhança (*sem*) entre C1 e C2 é:
  - considerar a semelhança entre cada ponto de C1 com todos os pontos de C2 e calcular a média de todas as semelhanças
  - ... *sem*(C1, C2) = `sum` [*sem*(Pi, Pj)] / (|C1| * |C2|)

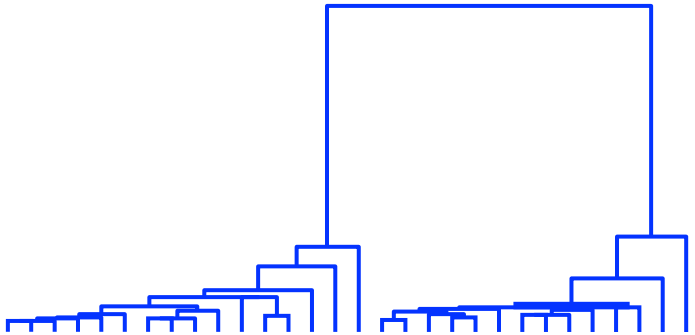


- **Ward’s Method**, Group Average com quadrados das semelhanças
  - ... *sem*(C1, C2) = `sum` [*sem*(Pi, Pj) ^{2}] / (|C1| * |C2|)

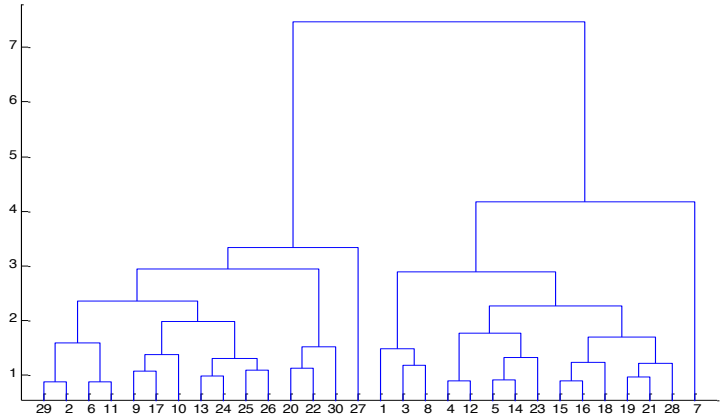
... “efeito” das medidas de “semelhança”



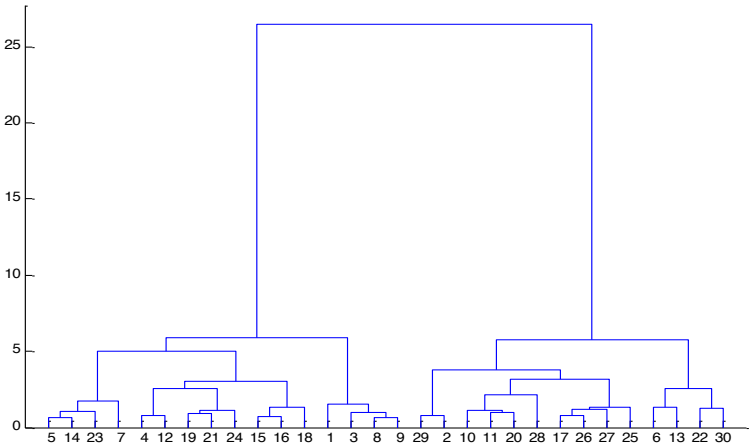
as instâncias



método MIN



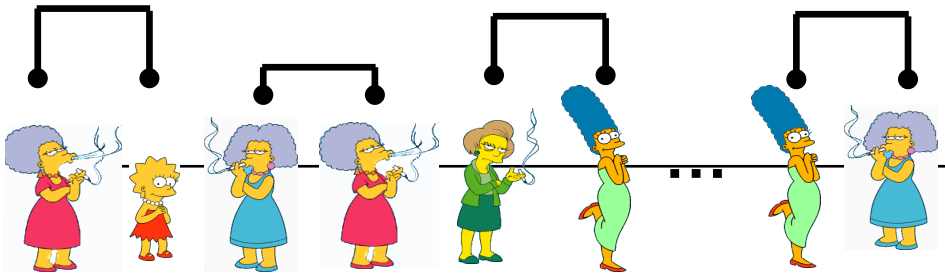
método Group Average



método Ward’s

... outro exemplo – nível 0 (zero)

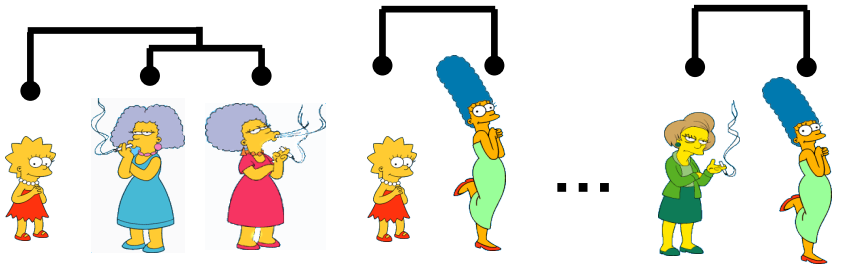
Considerar todas as hipóteses…



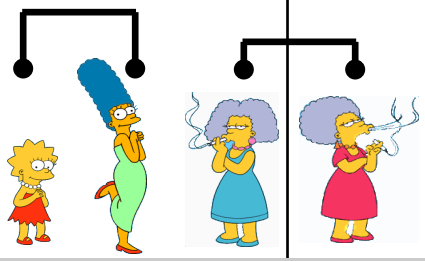
Escolher o melhor

... outro exemplo – nível 1

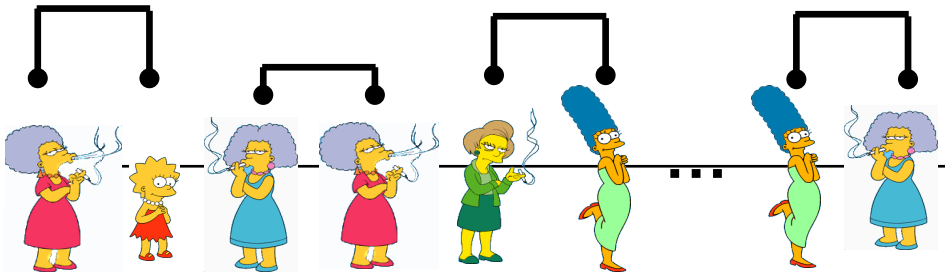
Considerar todas as hipóteses…



Escolher o melhor



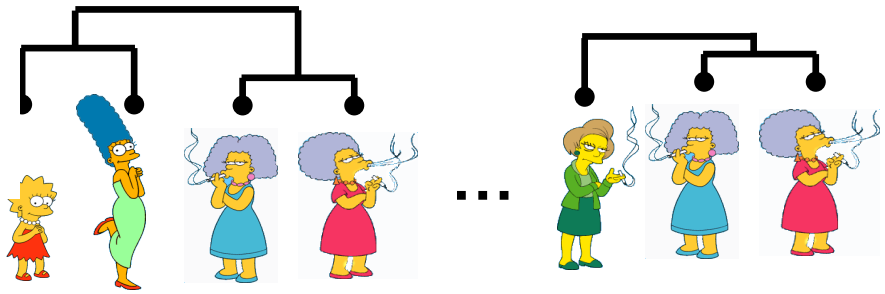
Considerar todas as hipóteses…



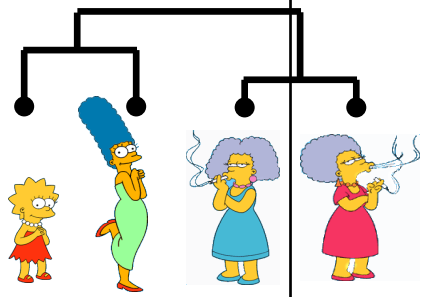
Escolher o melhor

... outro exemplo – nível 2

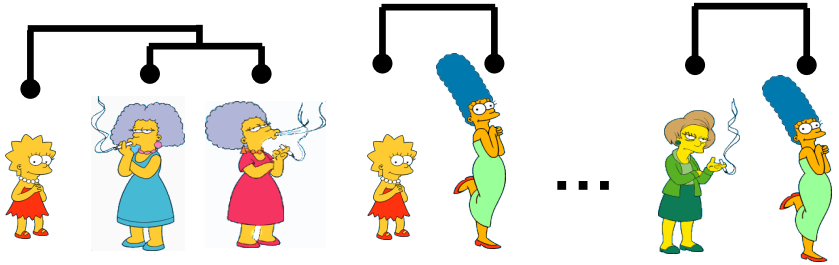
Considerar todas as hipóteses…



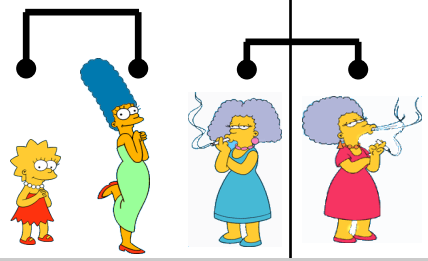
Escolher o melhor



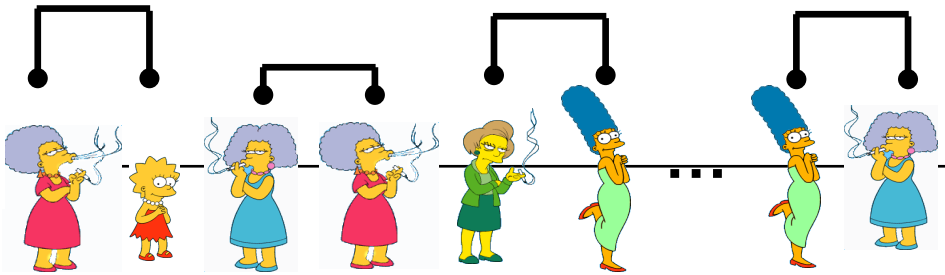
Considerar todas as hipóteses…



Escolher o melhor

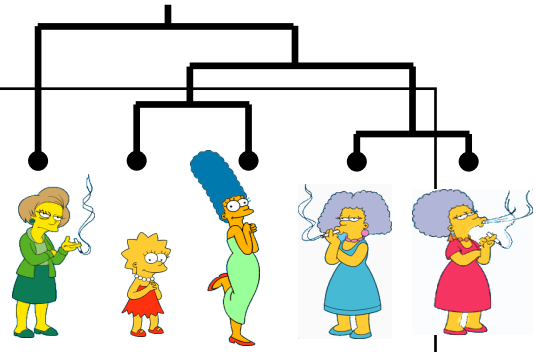


Considerar todas as hipóteses…

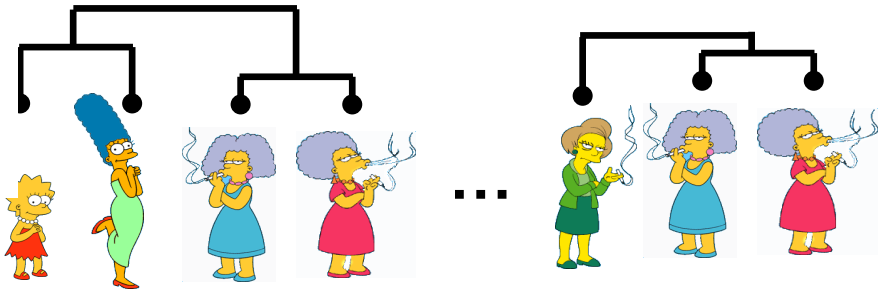


Escolher o melhor

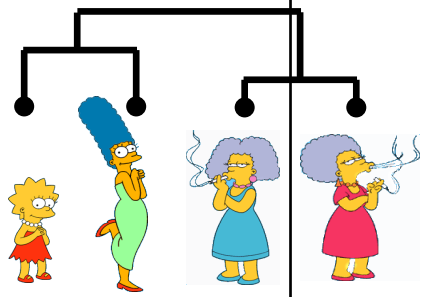
... outro exemplo – nível 3 ...



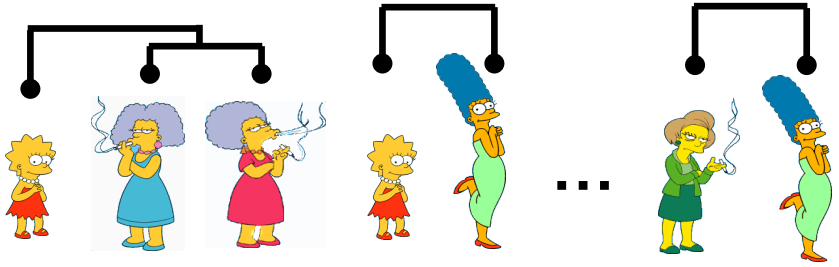
Considerar todas as hipóteses…



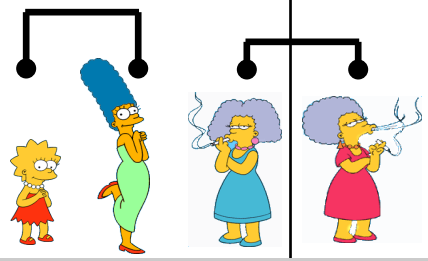
Escolher o melhor



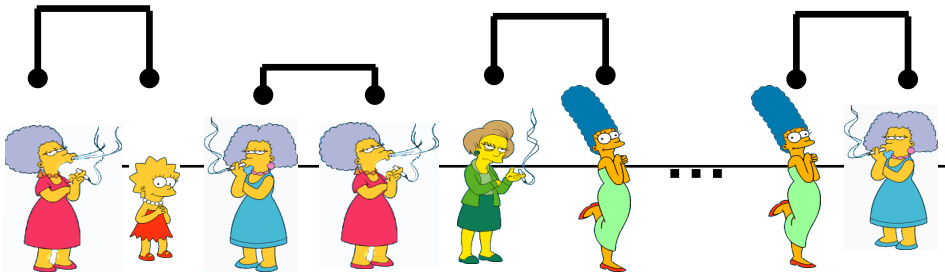
Considerar todas as hipóteses…



Escolher o melhor



Considerar todas as hipóteses…



Escolher o melhor

Agrupamento Hierárquico Aglomerativo – características

- não é necessário definir o número de grupos
- elevada complexidade computacional
  - da ordem $O(N^2)$, onde N é o número total de objetos
- de fácil percepção humana
- ... com diferentes possíveis interpretações dos resultados

Paulo Trigo Silva

Sumário dos Métodos de Agrupamentos

- Todos os métodos necessitam da medida de semelhança (“bias”)
  - baseado em função de semelhança dada explicitamente
  - baseado em modelos – função de semelhança é implícita

- Decidir o número óptimo de agrupamentos
  - é um problema complicado em todos os métodos!

- O melhor método depende do problema em causa!
  - aproximação reflete a nossa perspectiva da precisão dos agrupamentos?
  - aproximação feita é capaz de lidar com o problema?