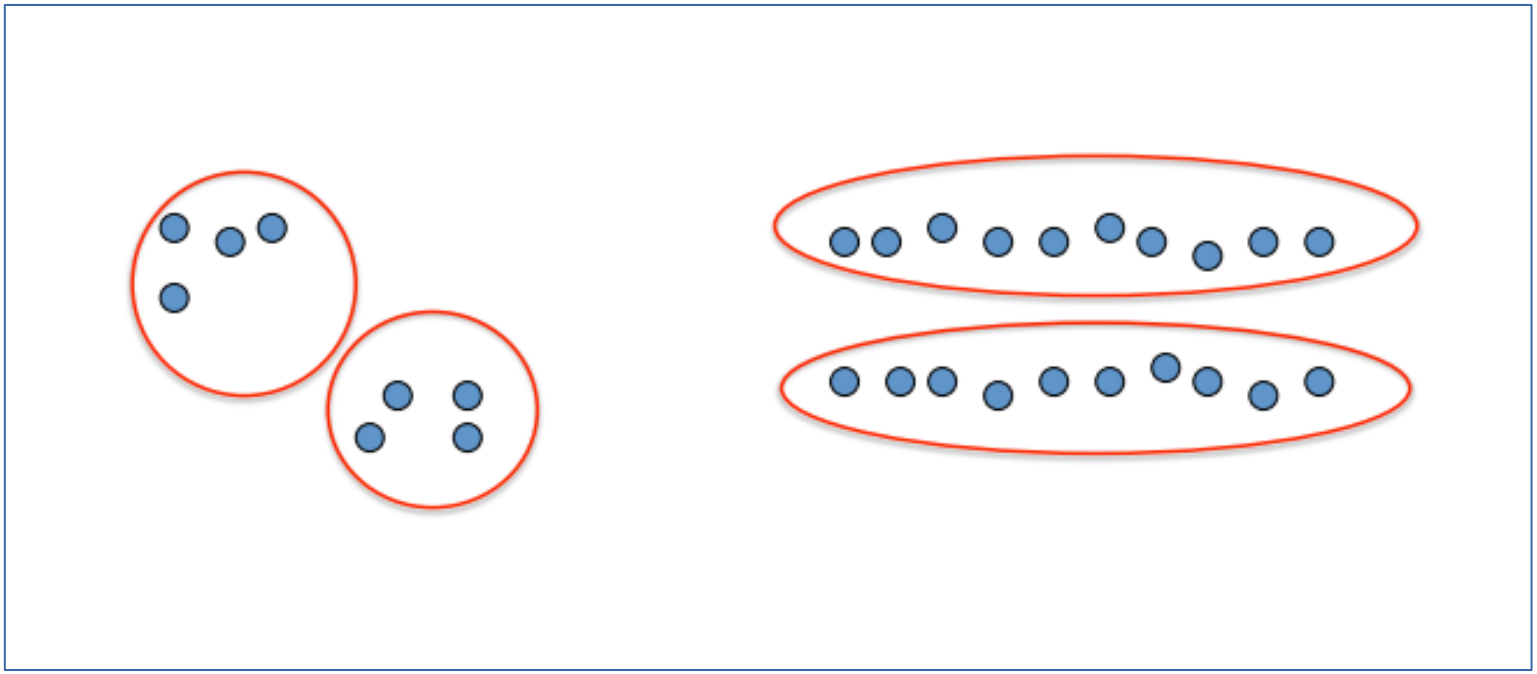
Mineração de Dados
Aula 2 – parte 2

Especialização em Ciência de Dados e suas Aplicações

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Agrupamento

- **Ideia básica:** agrupar instâncias similares



Espaço

Agrupamento

Biologia computacional

- Plantas/espécies
- Dados genéticos (família genética)

WWW

- Social *networks* (comunidades de usuários)
- Buscas similares (melhoraria de buscadores de páginas)

Marketing e *business*

- Compras similares
- Identificação de perfis

Agrupamento

Clusterização/agrupamento

É uma classificação não supervisionada: sem classes predefinidas

O que é um cluster?

Quantos clusters?

O que é um cluster?

Quantos clusters?

Seis Clusters

Dois Clusters

Quatro Clusters

O que é um cluster?

Quantos clusters?

Seis Clusters

Dois Clusters

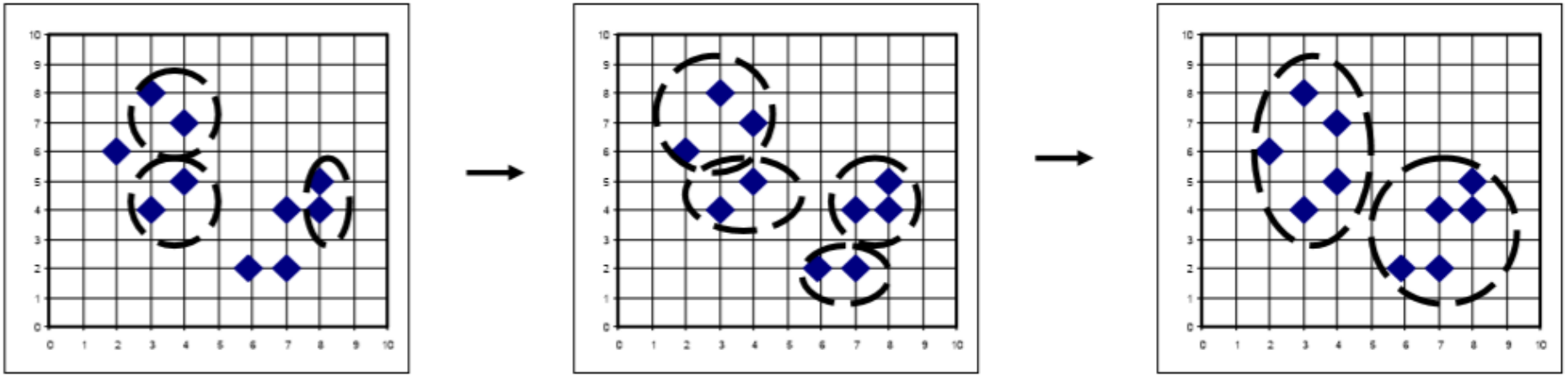
Quatro Clusters

Depende da natureza dos dados e dos resultados esperados

Agrupamento hierárquico (aglomerativo)

Agrupamento hierárquico (aglomerativo)

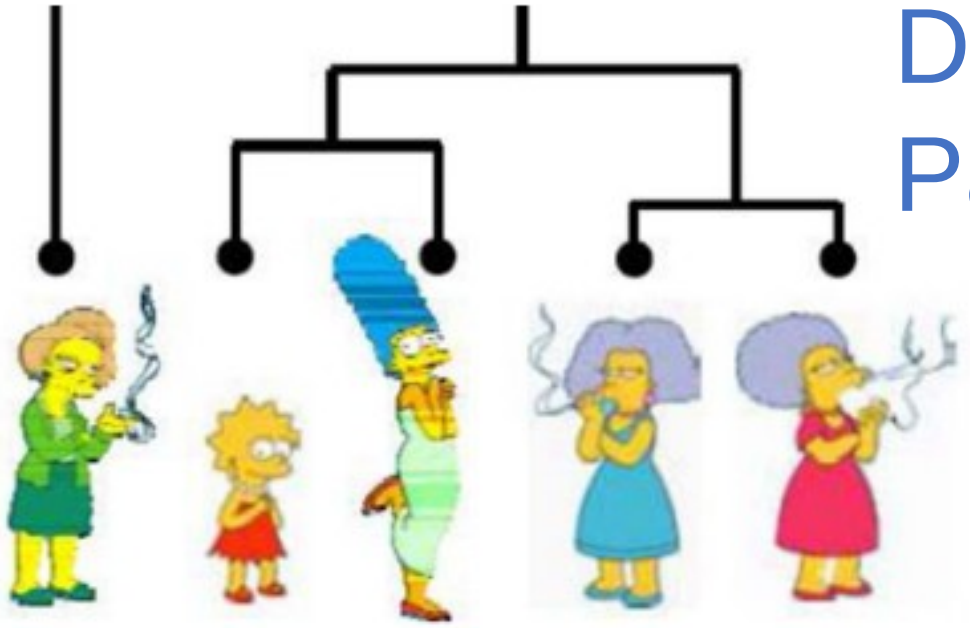
Aglomeração



Distância geográfica (haversine)

Imagem de Stanley R M Oliveira

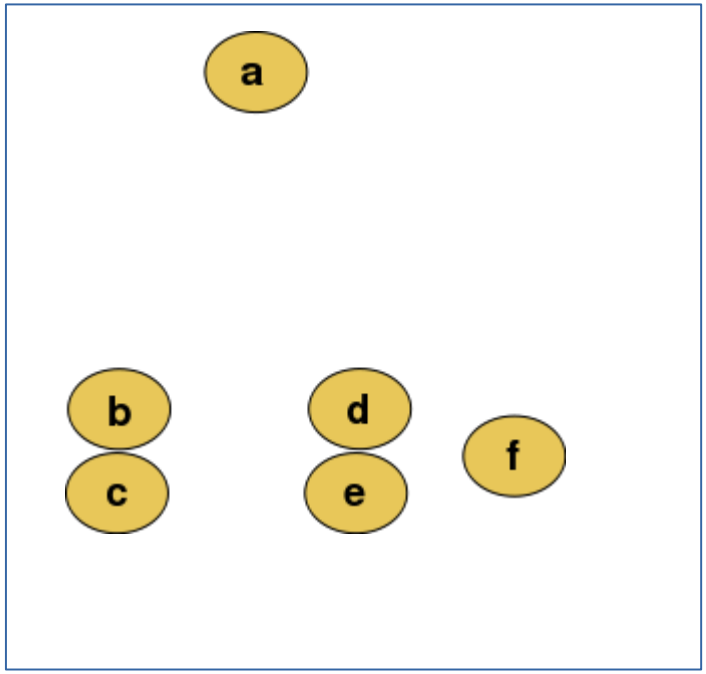
Agrupamento hierárquico (aglomerativo)

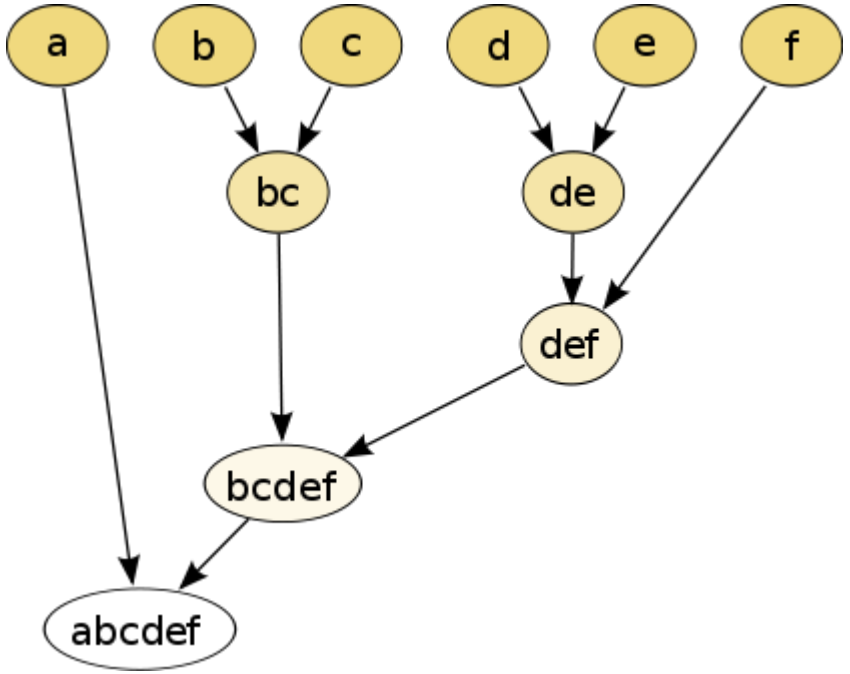


Distância
Parentesco

Imagem de David Sontag

Agrupamento hierárquico (aglomerativo)





Distância euclideana

Imagem da Wikipedia: https://en.wikipedia.org/wiki/Dendrogram

Agrupamento hierárquico (aglomerativo)

Algoritmo básico:

Computar a matriz de proximidade
Considere cada ponto como um cluster

Repetir
    Agrupar os dois clusters mais próximos
    Atualizar a matriz de proximidade
Até restar um único cluster

Operação chave!
Várias estratégias

Agrupamento hierárquico (aglomerativo)

Algoritmo básico:

Computar a matriz de proximidade
Considere cada ponto como um cluster

Repetir
Agrupar os dois clusters mais próximos
Atualizar a matriz de proximidade
Até restar um único cluster

Complexidade de tempo:

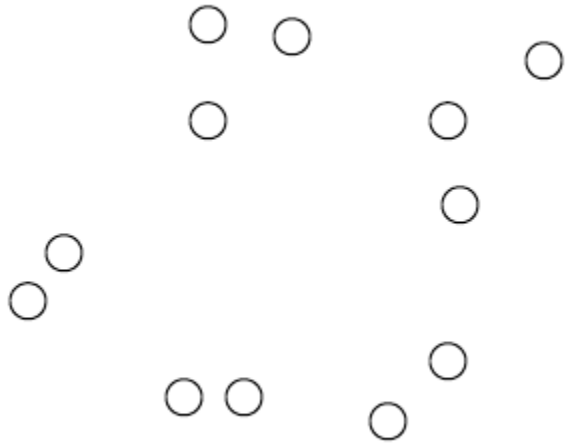
O (m^2 log m)

m = # de objetos

Operação chave!
Várias estratégias

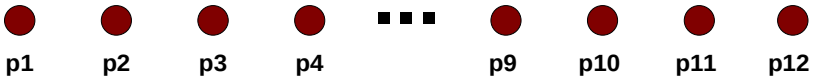
Situação inicial

- Clusters contendo pontos individuais e uma matriz de proximidade



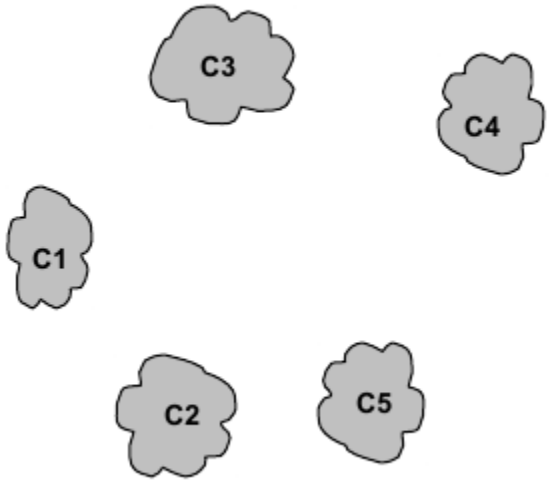
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
. . .						

Matriz de proximidade



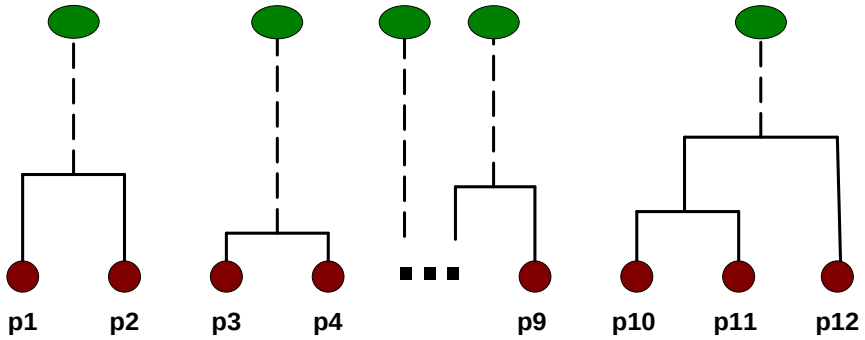
Situação intermediária

- Depois do passo de agrupamento, nós temos alguns grupos



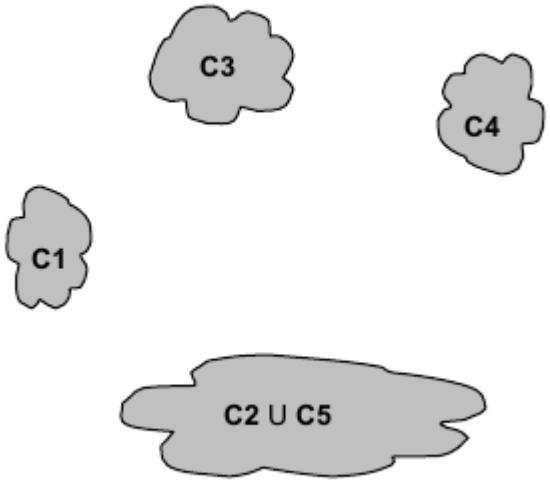
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de proximidade



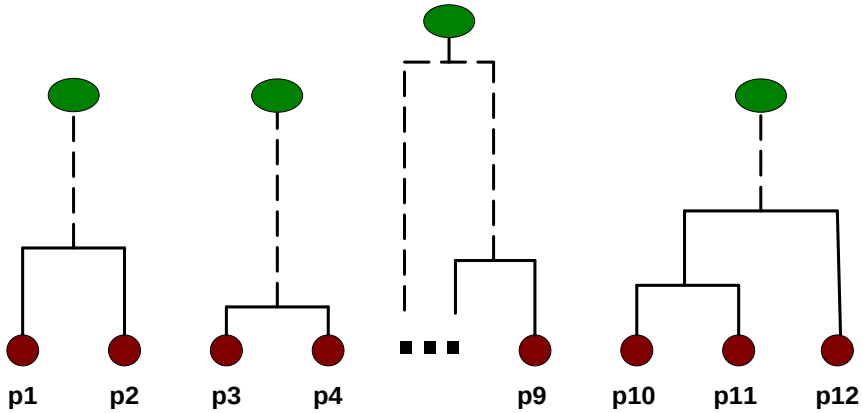
Depois do agrupamento

- Como atualizar a matriz de proximidade ???



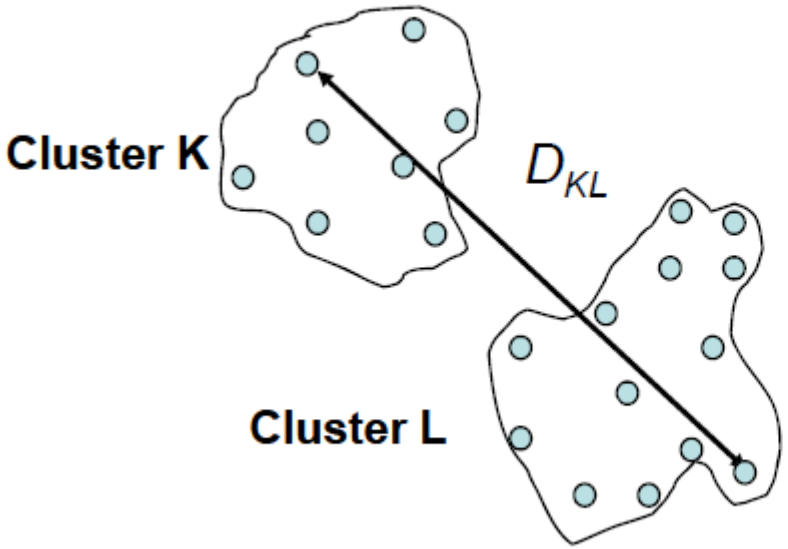
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matriz de proximidade



Distâncias entre clusters

Complete linkage – distância entre os pontos mais distantes

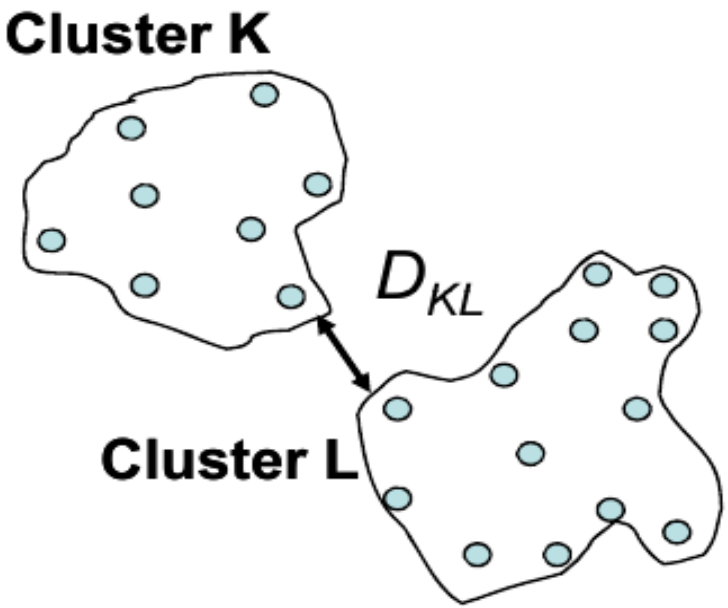


$$D_{KL} = \max_{\substack{i \in C_K \\ j \in C_L}} d(x_i, x_j)$$

Calcula para cada cluster e encontra a menor para agrupar

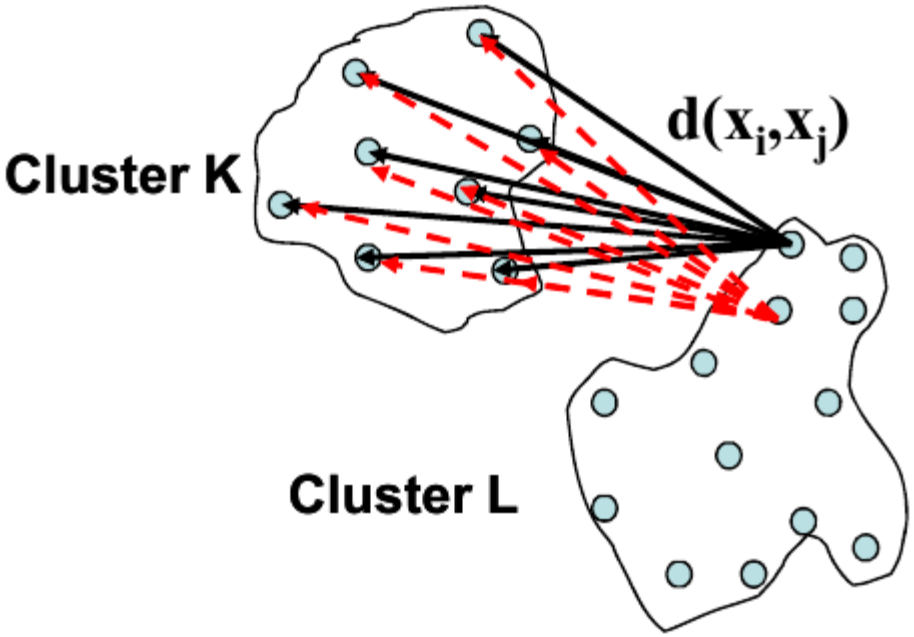
Distâncias entre clusters

Single linkage – distância entre os pontos mais perto



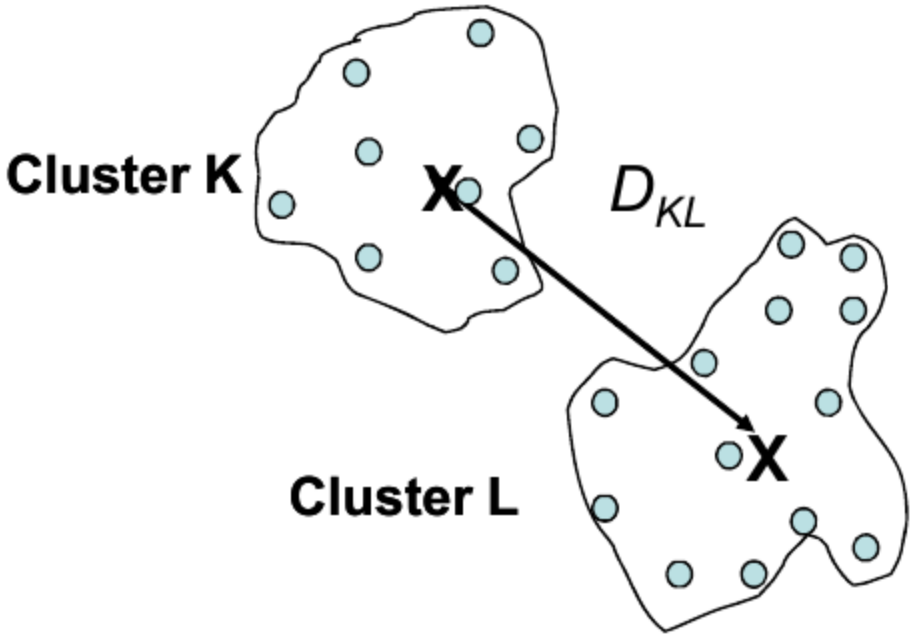
Distâncias entre clusters

Average-likage – distância média entre os pontos de um grupo com todos os pontos de outro grupo



Distâncias entre clusters

Centroid-linkage – distância entre os centroides dos clusters



Distâncias entre clusters

Ward's method – similaridade de dois clusters é baseada no aumento no erro ao quadrado (SSE) quando dois clusters são agrupados

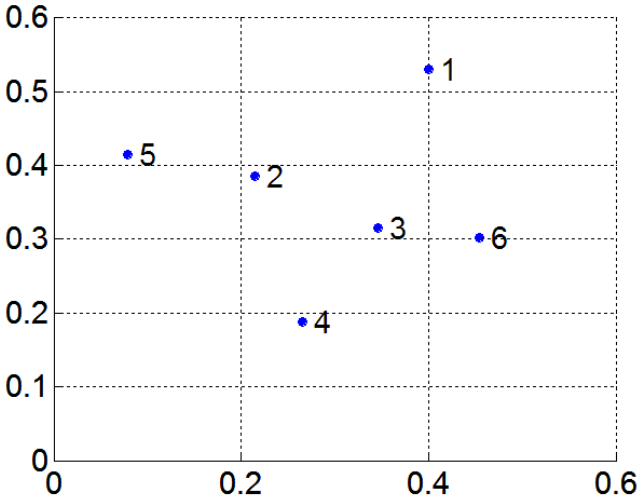
Passos:

- Considera o agrupamento de dois clusters
- Calcula o centroid
- Calcula o SSE

Menos suscetíveis a ruídos

Exemplo: Single Linkage

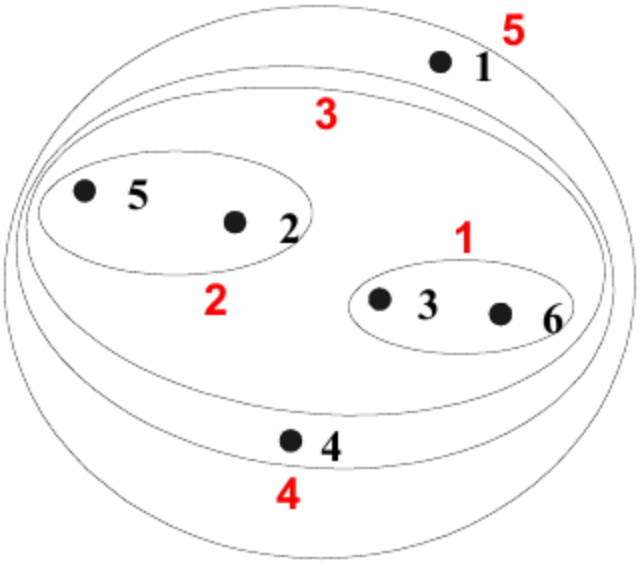
- Exemplo:

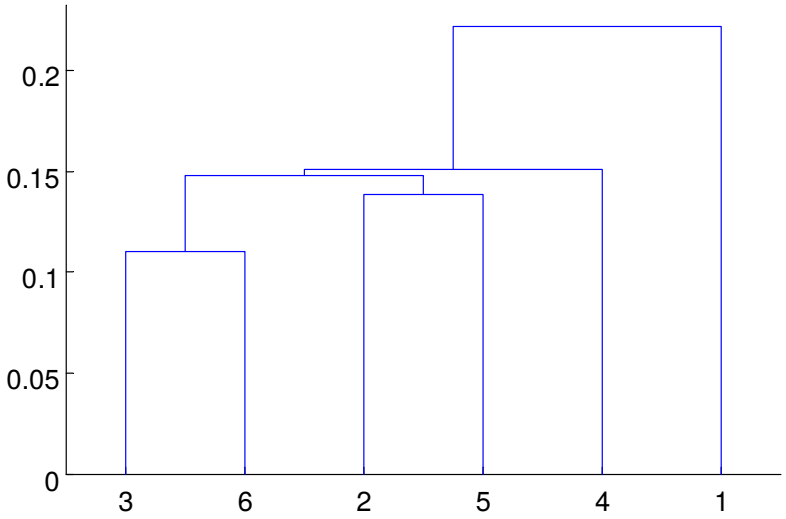


Matriz de distância

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Exemplo: Single Linkage

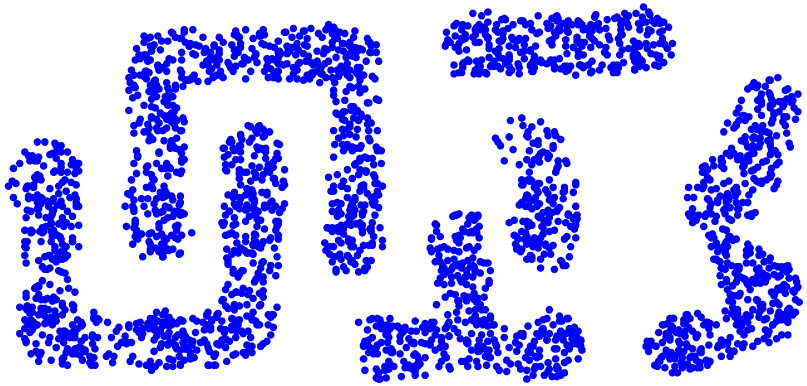




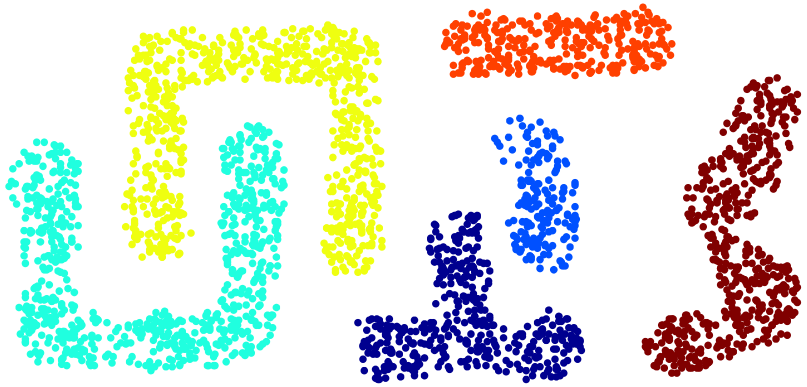
Clusters aninhados

Dendrograma

Vantagens do single linkage



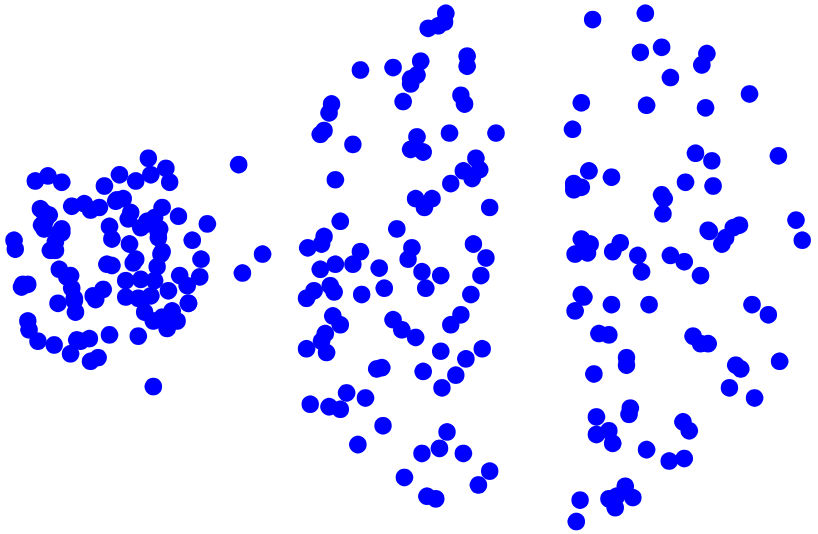
Original Points



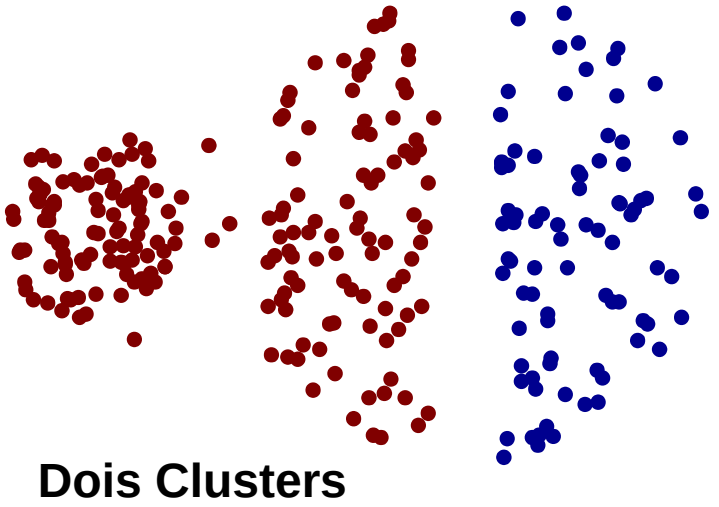
Six Clusters

- **Bom em agrupar formatos não-elípticos**

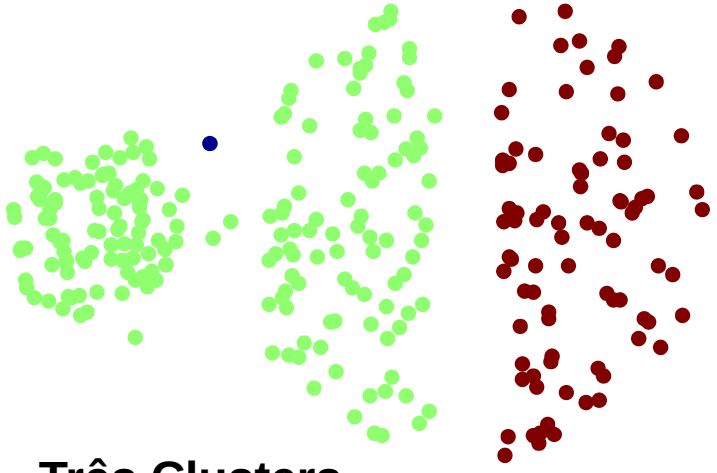
Limitações do single linkage



Original Points



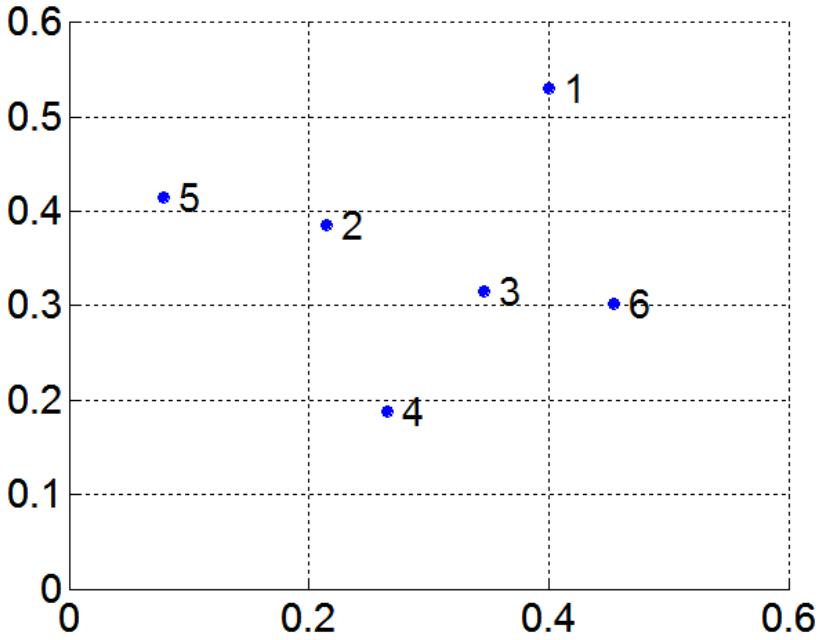
Dois Clusters



Três Clusters

Suscetíveis a ruídos

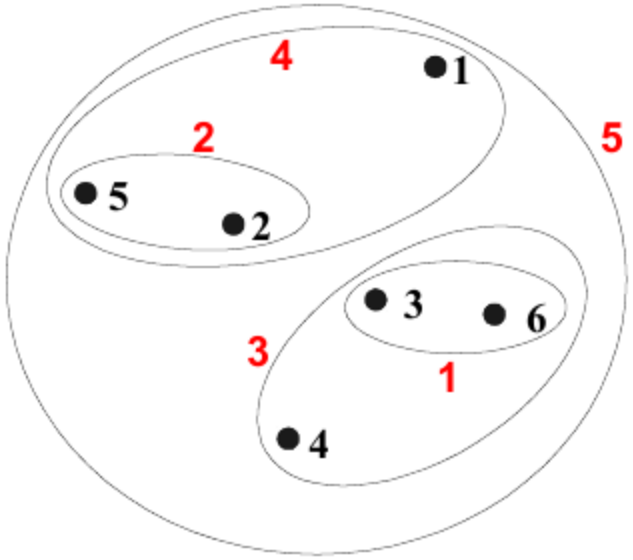
Exemplo: Complete Linkage



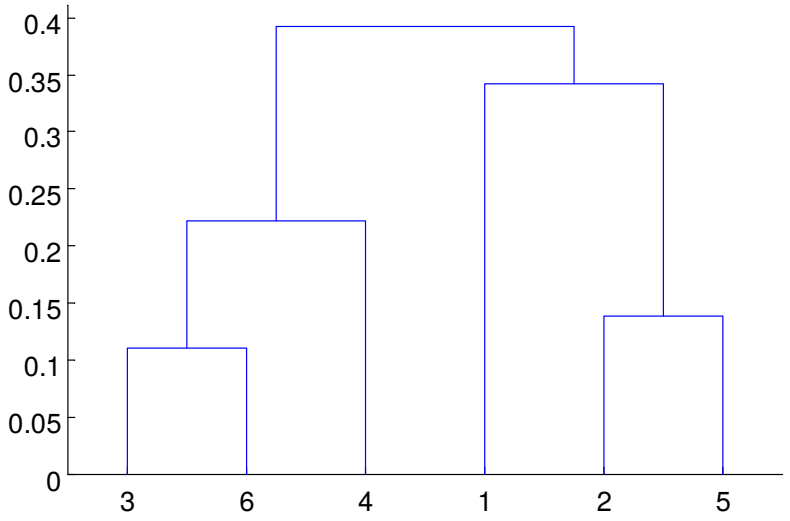
Matriz de distância:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Exemplo: Complete Linkage

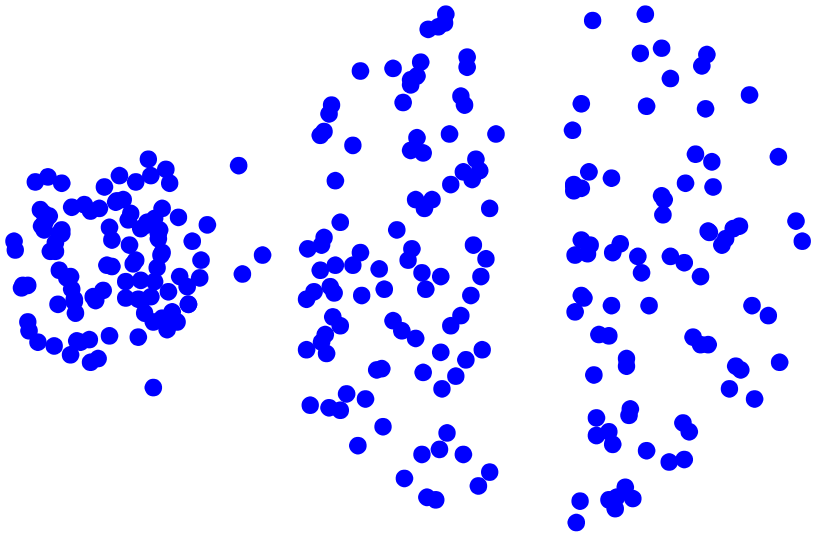


Clusters aninhados

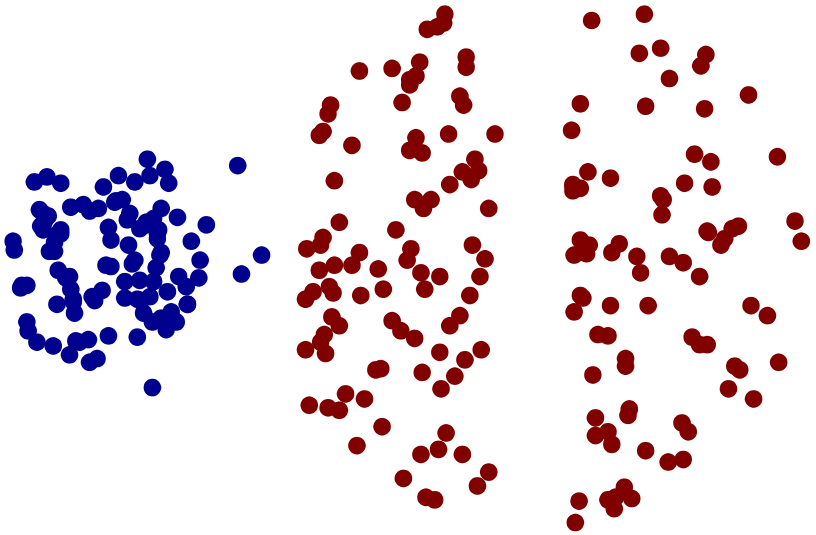


Dendrograma

Vantagens do complete linkage



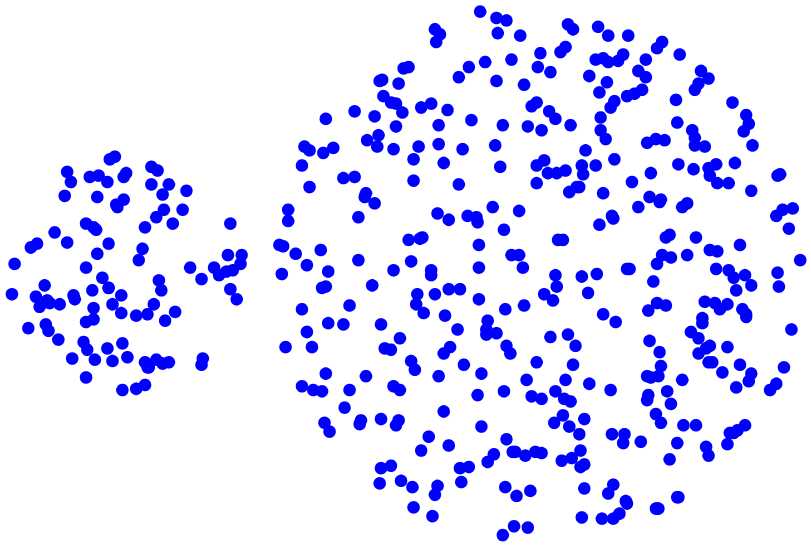
Original Points



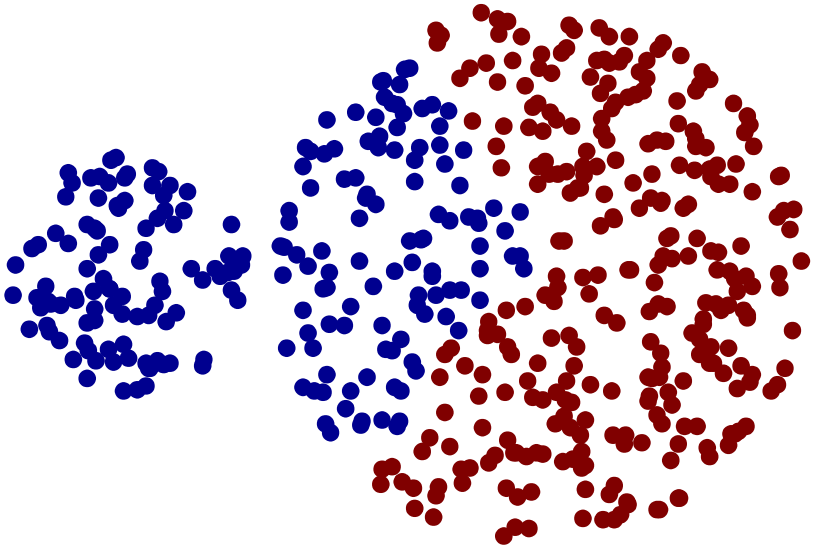
Two Clusters

- **Menos susceptível a ruídos**

Limitações do complete linkage



Original Points



Two Clusters

- **Tende a quebrar grupos largos**
- **Tende a formar clusters globulares (esféricos)**

Exemplo

Agrupando cidades similares

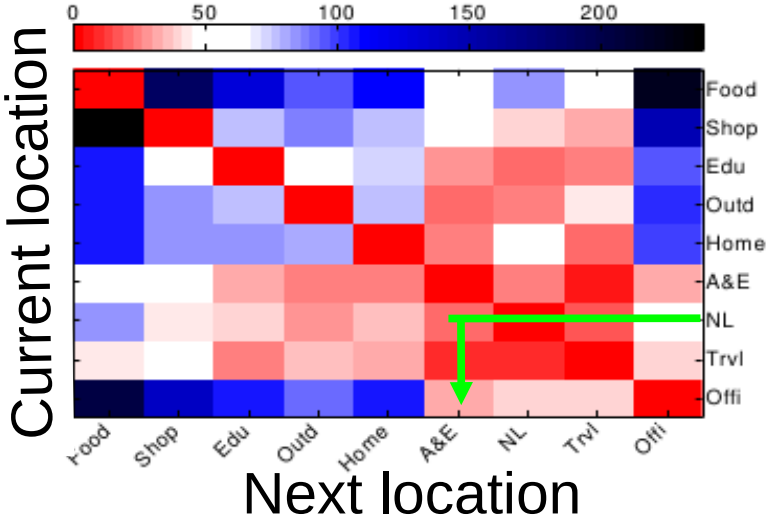
Normalizamos cada matriz de transição (City Image) e criamos um vetor {t1, t2, … t81}, onde cada posição é uma célula da matriz de transição

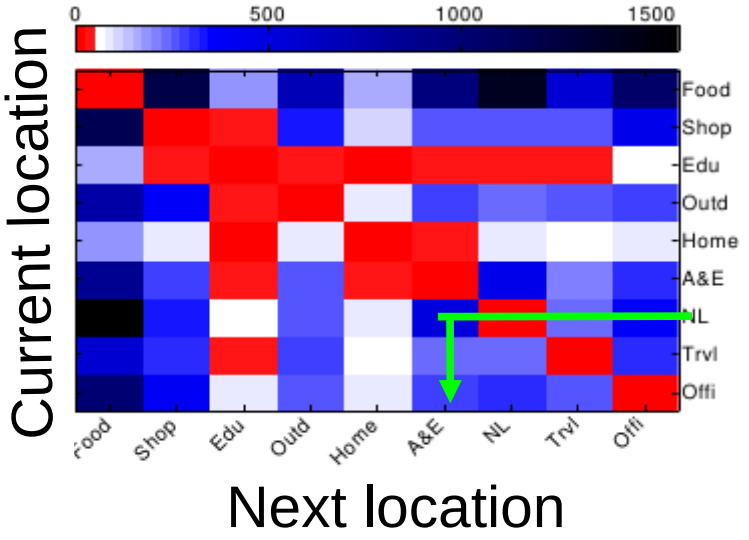
Calculamos a distância euclideana entre cada vetor

Realizamos uma clusterização hierárquica

Agrupamento

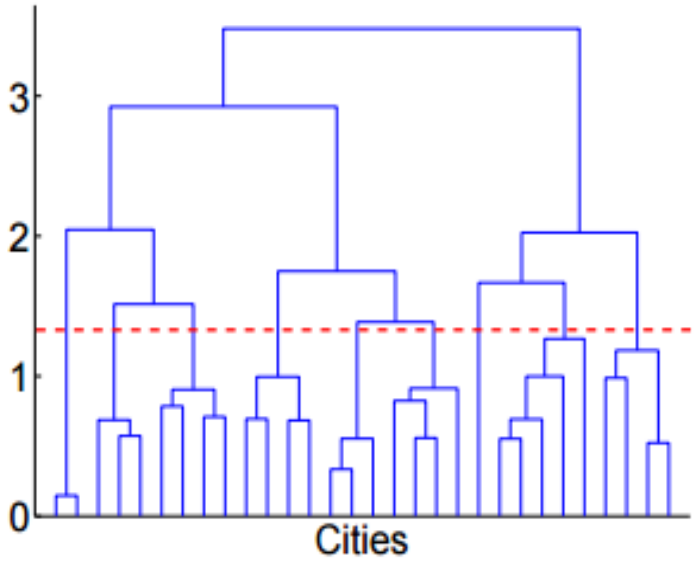
City image



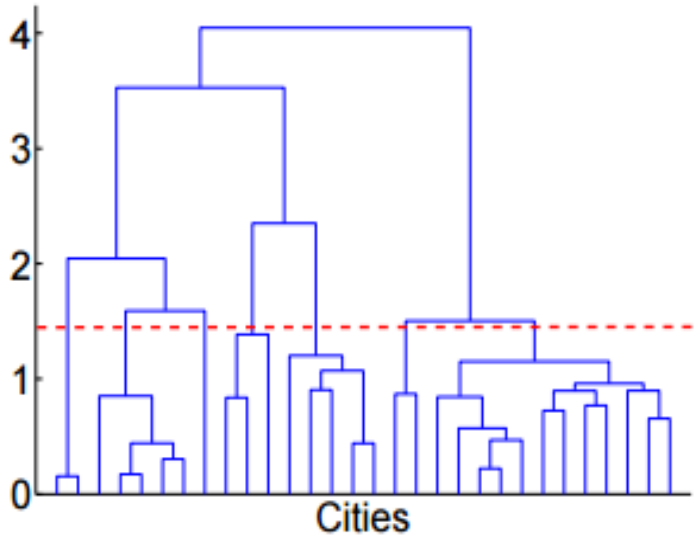


Agrupamento

Agrupando cidades similares



(a) Day – weekday



(b) Night – weekend

Agrupamento

Agrupando cidades similares

Table IV. Clustering results for weekend during the night.

Cluster	Cities
1	Kuwait, Singapore, Kuala Lumpur, Manila, Bangkok
2	Tokyo, Osaka
3	Seoul, Jakarta, Bandung, Semarang, Surabaya
4	Rio, Belo Horizonte, Sao Paulo
5	Istanbul, Moscow
6	Santiago
7	Los Angeles, Chicago, San Francisco, New York, Melbourne, Sydney, Paris, Madrid, London, Barcelona, Buenos Aires, Mexico City

Agrupamento baseado em protótipo: k-means

k-means

Algoritmo básico do k-means

- Inicialização: Escolher *k* pontos aleatórios para serem o centro do cluster
- Alternar:
  1. Atribuir os pontos para o centro mais próximo
  2. Mudar o centro do cluster para a *média* dos pontos atribuídos
- Parar quando não haver mudanças significativas

Tipicamente

k-means

Algoritmo básico do k-means

- Inicialização: Escolher *k* po
  centro do cluster
- Alternar:
  1. Atribuir os pontos para
  2. Mudar o centro do clus
     atribuídos
- Parar quando não haver mudanças significativas

Complexidade de tempo:

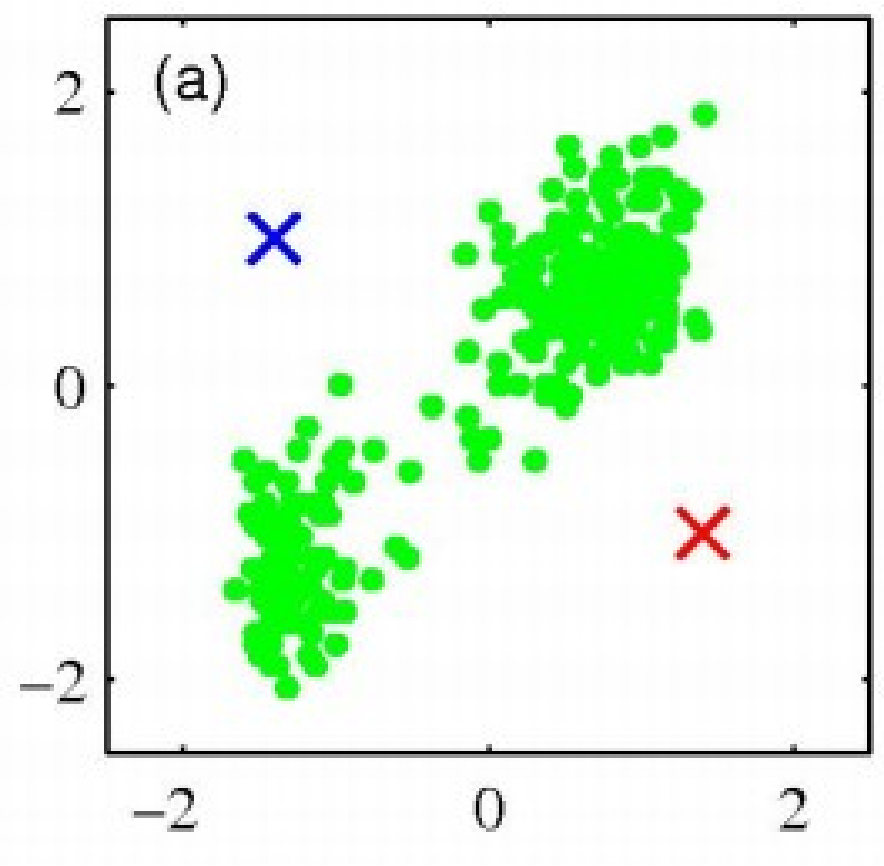
O(I x K x m x n)

I = # de iterações para convergir
m = # de objetos
n = # de atributos

I é tipicamente pequeno, assim k-means é considerado linear em m

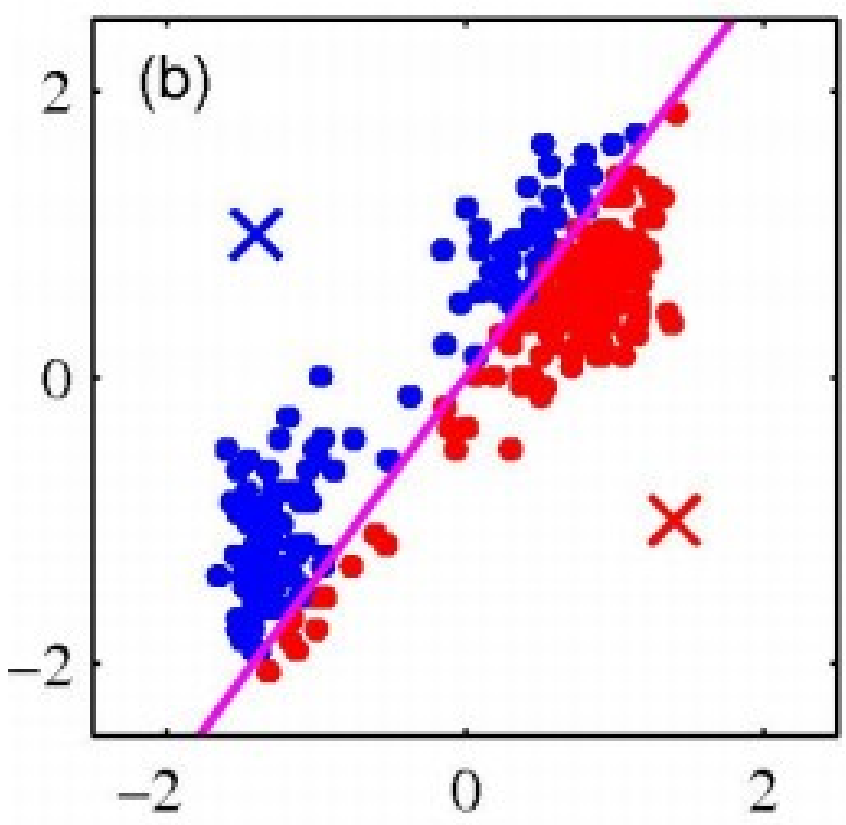
Tipicamente

k-means



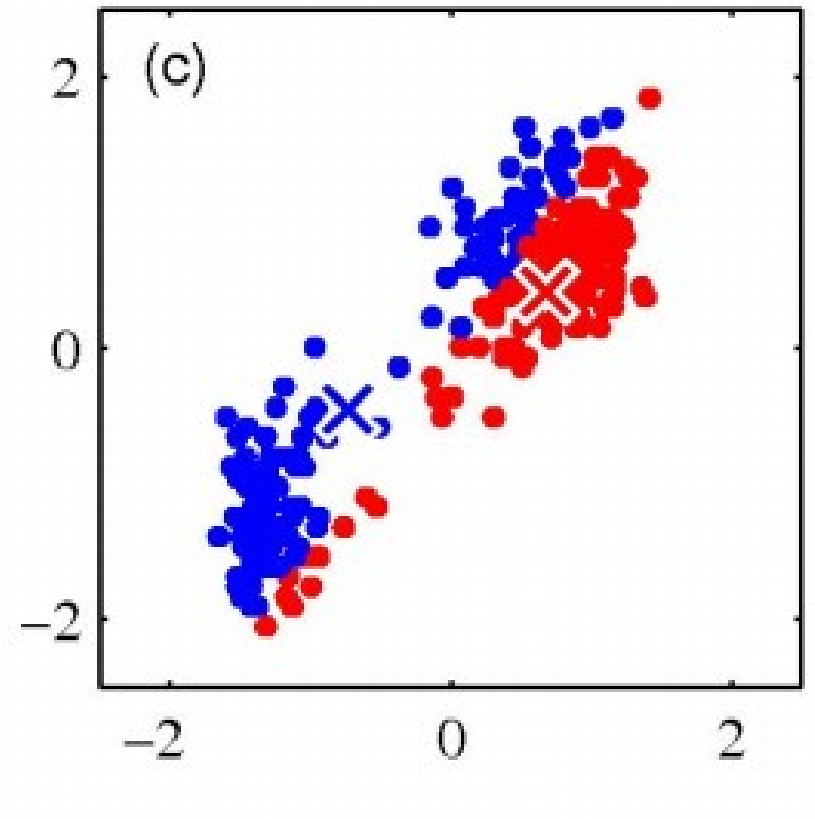
Escolher k=2
pontos aleatórios

k-means



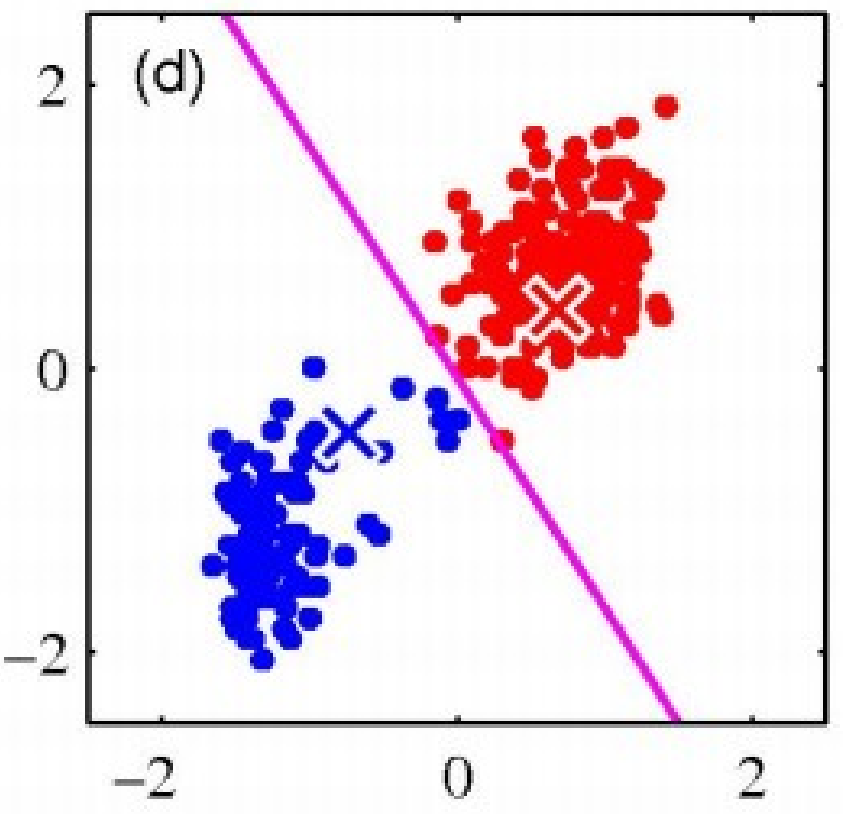
Atribuir os pontos para o centro mais próximo

k-means



Mudar o centro do cluster para a média dos pontos atribuídos

k-means



Repetir até convergir

Considerações sobre o K-means

Os centroides iniciais são escolhidos, tipicamente, aleatoriamente

O centroid é, tipicamente, a média de todos os pontos do cluster

‘Proximidade’ é medida pela distância euclideana, similaridade do cosseno, etc.

Limitações do K-means

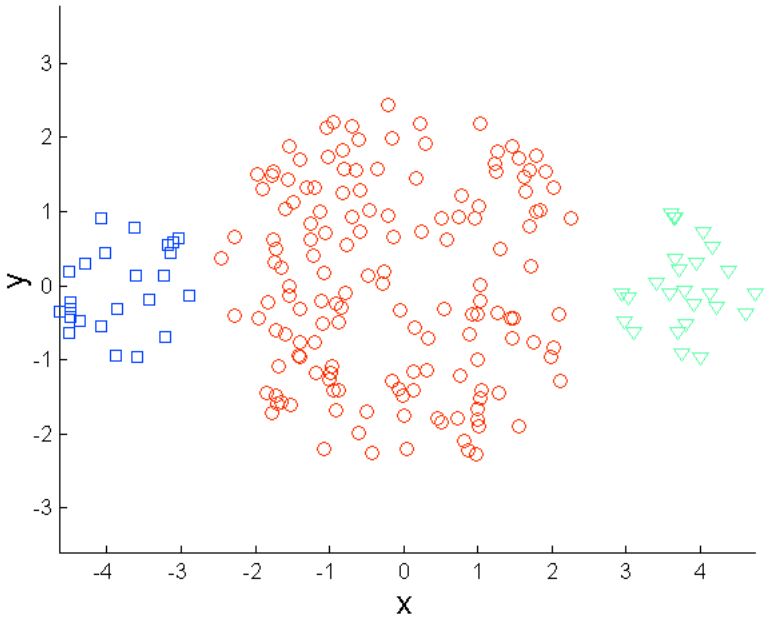
- K-means apresenta problemas quando clusters possuem diferentes:
  - Tamanhos
  - Densidades
  - Formatos não-esféricos

- K-means apresenta problemas com ruídos.

Limitações do K-means

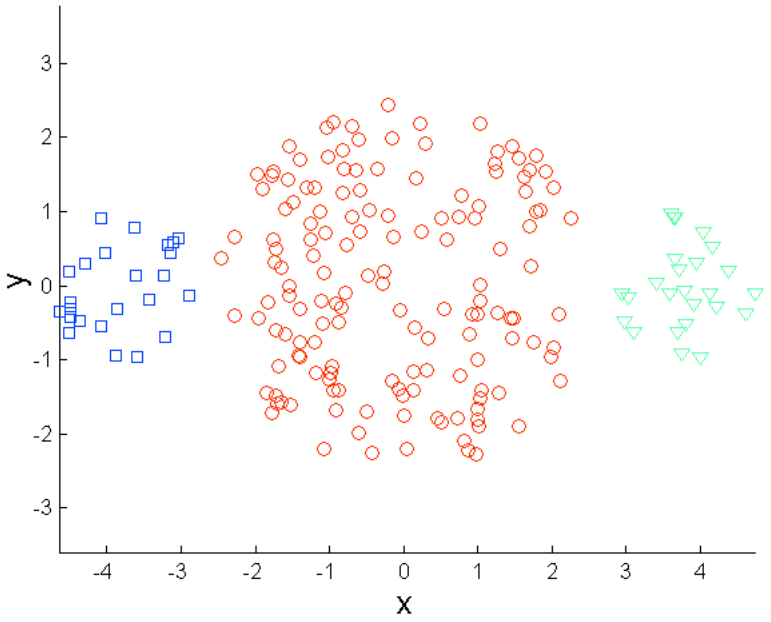
- A cada rodada o resultado pode ser diferente
- Na prática é comum inicializar o algoritmo várias vezes (50 – 100), para encontrar uma rodada com melhor resultado

Limitações: tamanhos diferentes

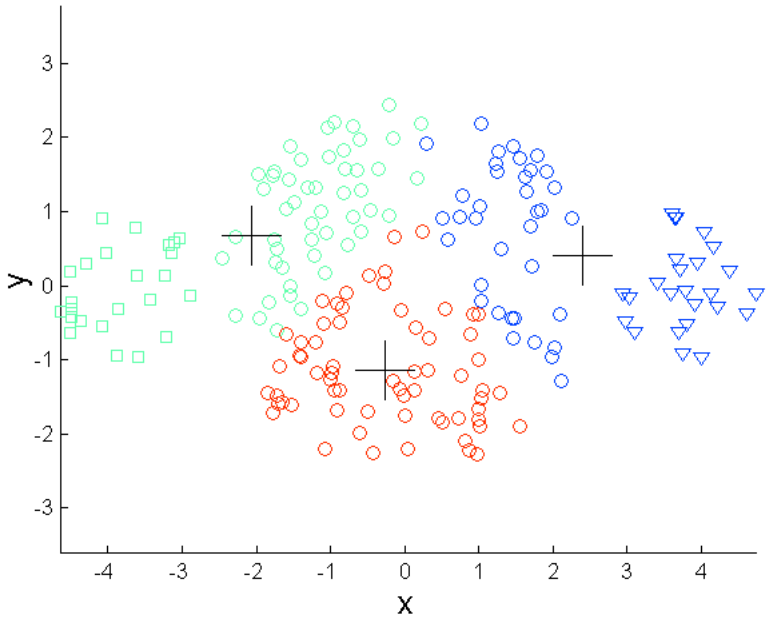


Pontos originais

Limitações: tamanhos diferentes

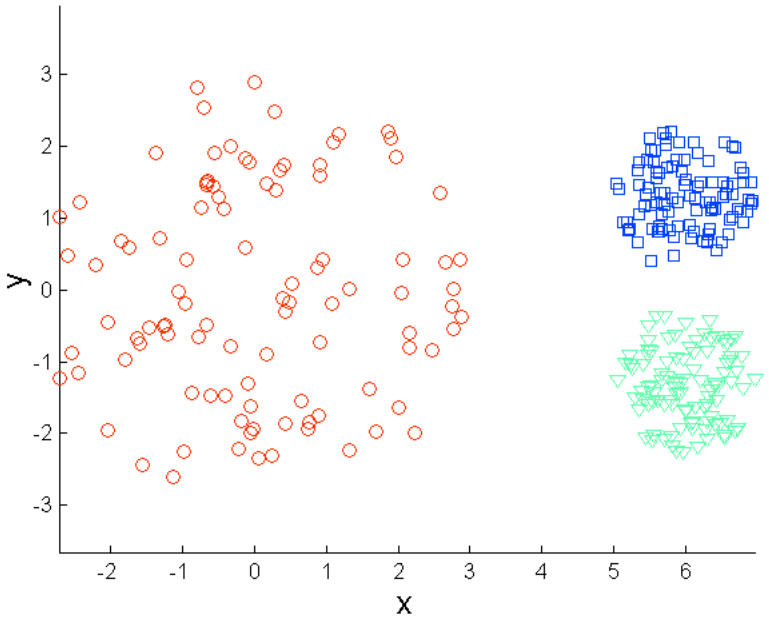


Pontos originais



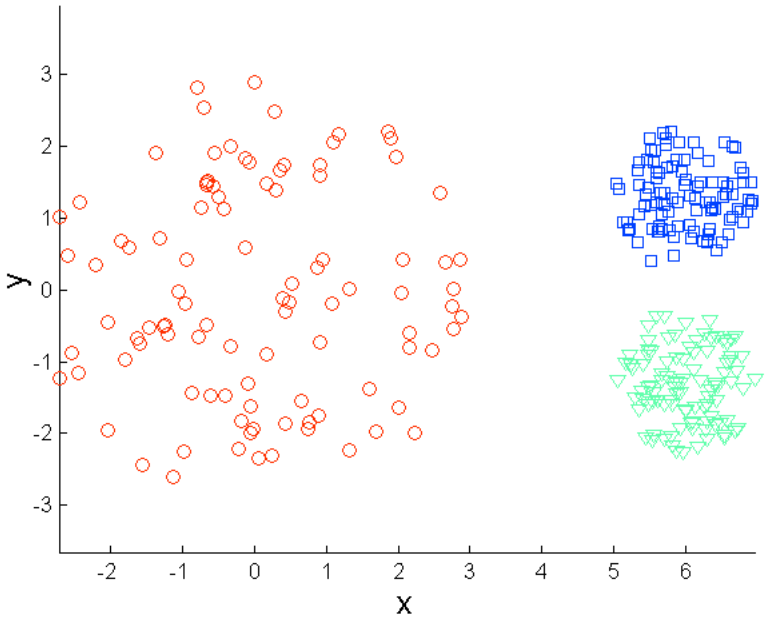
K-means (3 Clusters)

Limitações: diferentes densidades

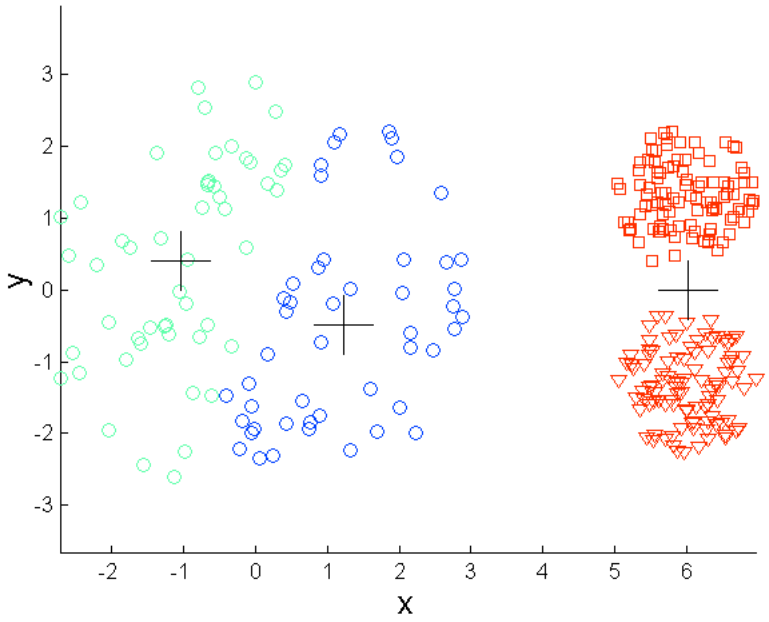


Pontos originais

Limitações: diferentes densidades

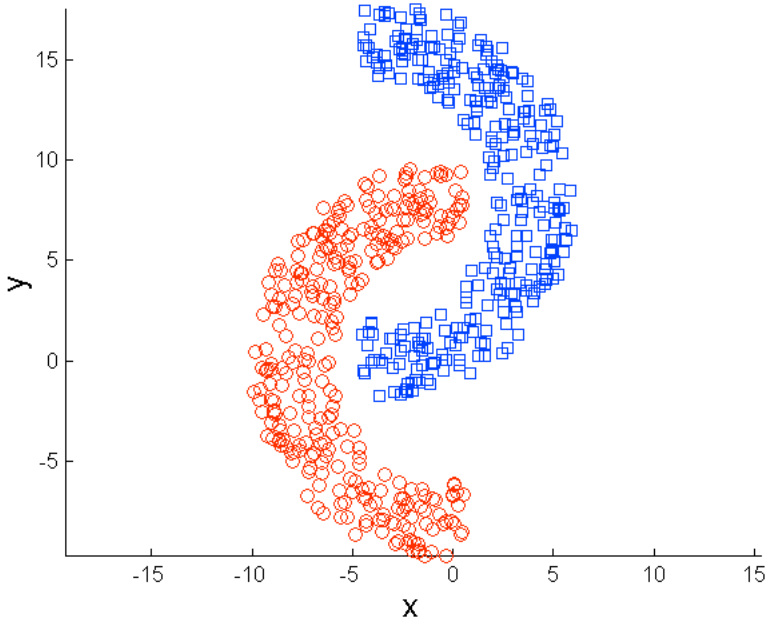


Pontos originais



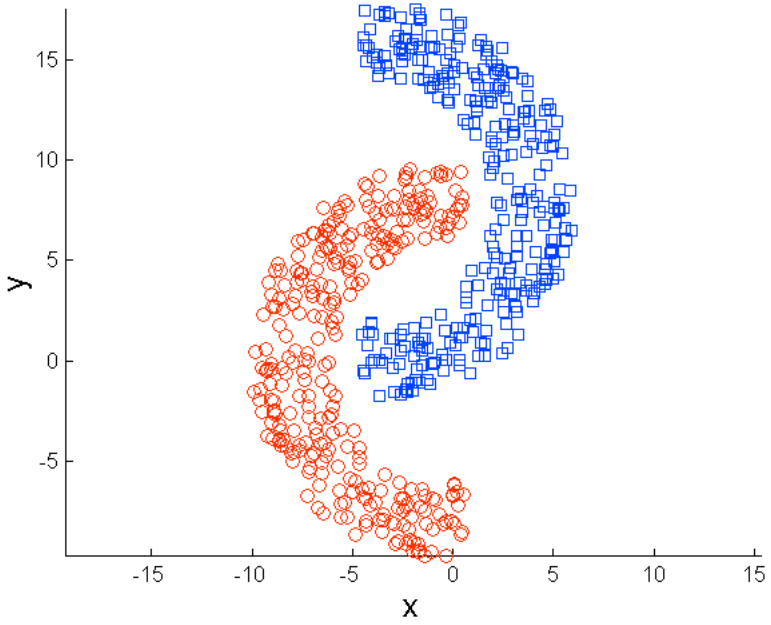
K-means (3 Clusters)

Limitações: formatos não-esféricos

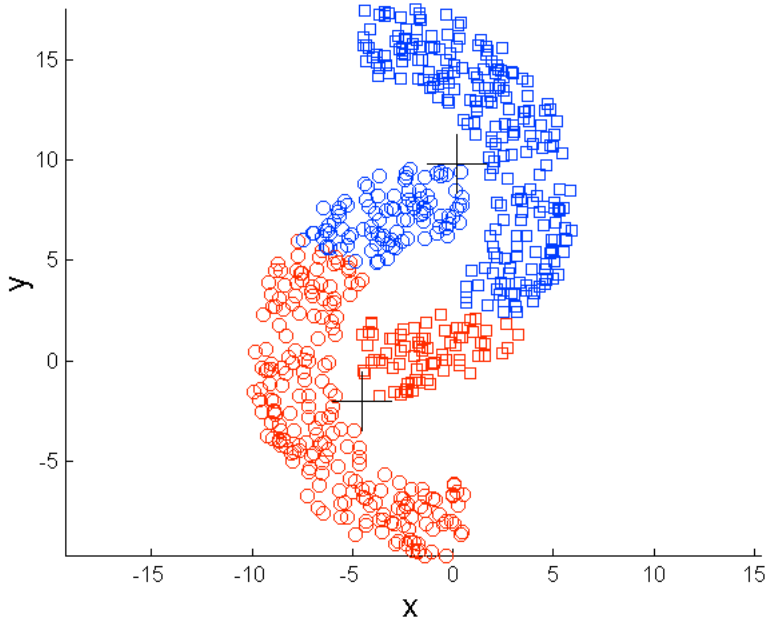


Pontos originais

Limitações: formatos não-esféricos



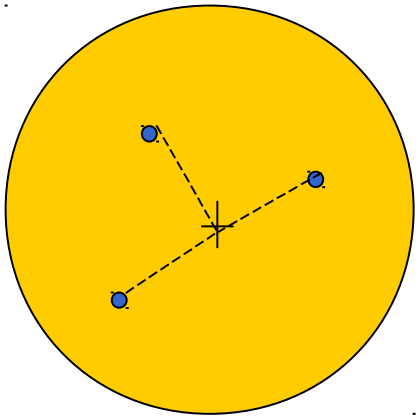
Pontos originais

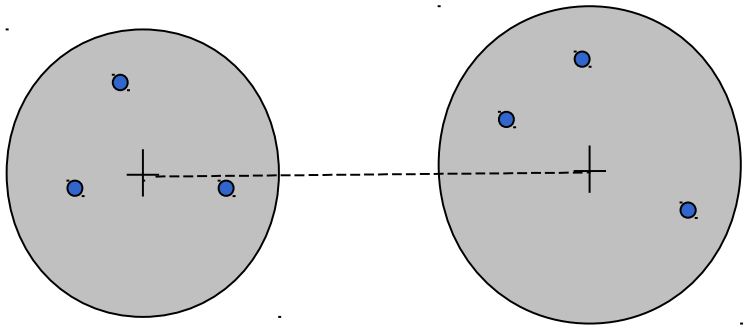


K-means (2 Clusters)

Grupos baseados em protótipos

- **Medidas de Avaliação** = coesão e separação





Coesão (individual)
Mede o quanto os objetos dentro de um grupo se aglomeram perto do **centro do grupo**

Separação (inter grupos)
Mede o quanto os centros dos grupos estão bem separados entre si

Grupos baseados em protótipos

Coesão(C_i) = $\sum_{x \in C_i}$ *proximidade*(x, m_i)

m_i = centroide de C_i

Exemplo:

$$SSE(C_i) = \sum_{x \in C_i} \textbf{\textit{dist}}(x, m_i)^2$$

Separação(C_i,C_j) = *proximidade*(c_i,c_j)

Proximidade : noção que pode variar dependendo da aplicação

Coesão e separação

Como utilizar coesão e separação para “melhorar” a clusterização

- Um cluster com baixo grau de coesão pode ser dividido em 2 subclusters.
- Dois clusters que têm boa coesão mas que não tem bom grau de separação podem ser juntados para formar um único cluster.

Coeficiente de silhueta

- Medida que combina coesão e separação

Coeficiente de silhueta

Coeficiente silhueta de um objeto t:

Dado um conjunto de Clusters C = {C_1,...,C_k} e um objeto t

Calcule a_t -> distância média de t a todos os objetos de seu cluster

Calcule b_t ->

Para cada cluster C' não contendo t, calcule t(C') a distância média entre t e todos os objetos de C'

b_t = min {t(C') | C' não contém t }

Coeficiente Silhueta (t) = (b_t – a_t) / max(a_t , b_t)

Coeficiente de silhueta

Coeficiente de Silhueta varia de -1 a 1

Valores negativos: $a_t > b_t$ (não desejados)

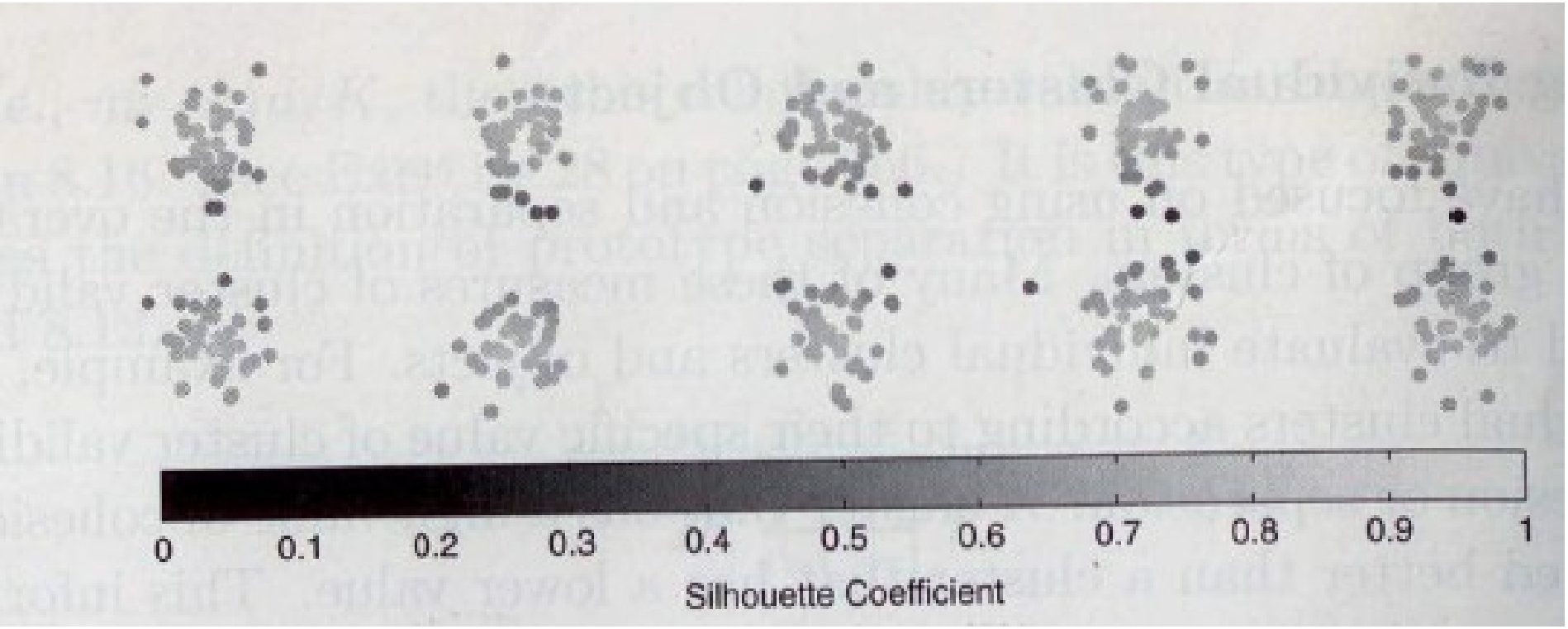
- Pois distância média de t a objetos de seu cluster é maior que distância média de t a objetos de outros clusters

Valores Ideais

- Valores positivos
- a_t bem próximo de zero
- Coeficiente de silhueta bem próximo de 1

Coeficiente de silhueta

Dados agrupados em 10 clusters e os coeficientes de silhueta dos pontos



Escolhendo o valor de k

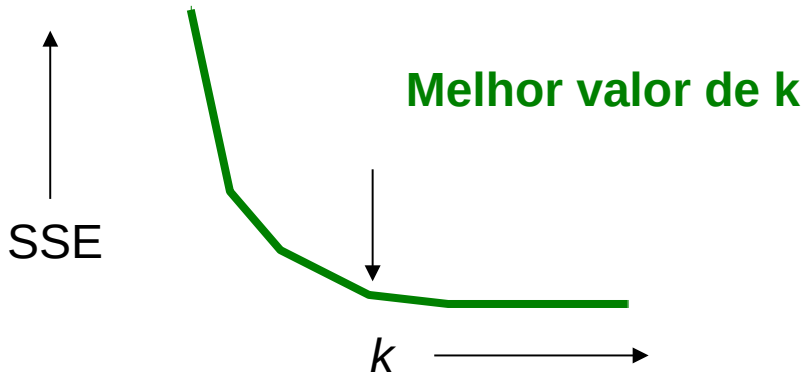
Como selecionar o k no k-means?

Escolhendo o valor de k

Como selecionar o k no k-means?

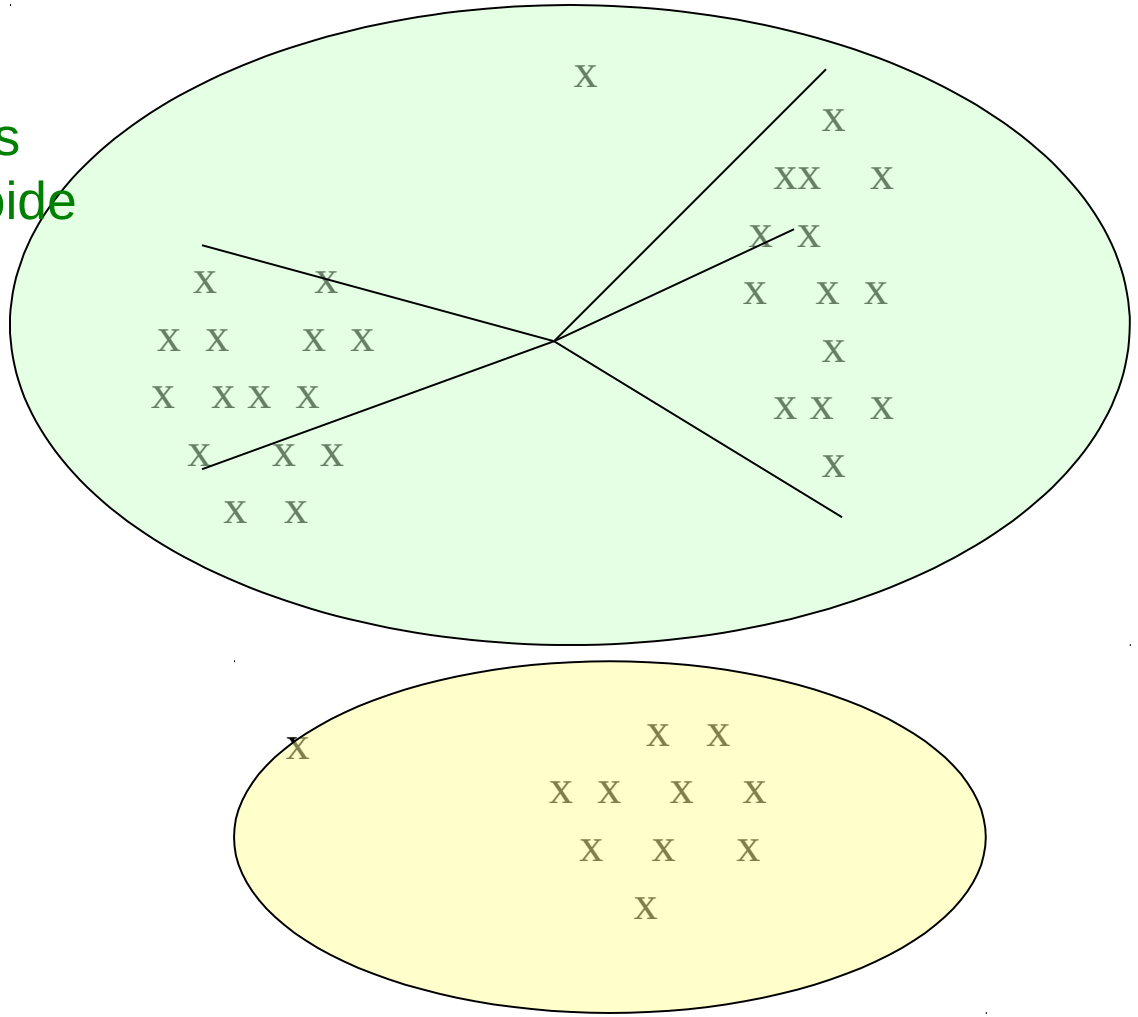
- Teste diferentes **k**, olhando as alterações na distância média ao centroide com o aumento de k (SSE)
- A média cai rapidamente até o k ideal, após muda pouco

$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2$$



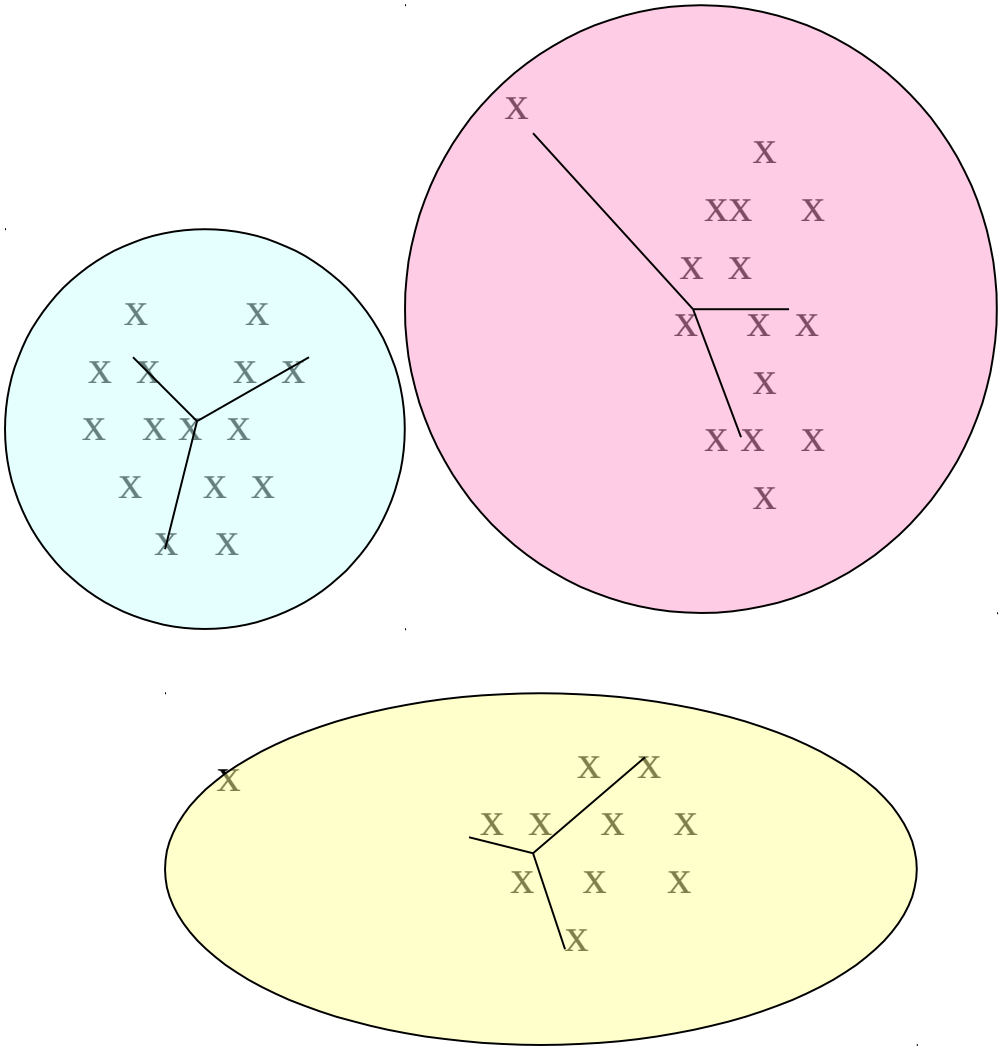
Escolhendo o valor de k

Muito pouco;
Muitas distâncias
longas ao centroide

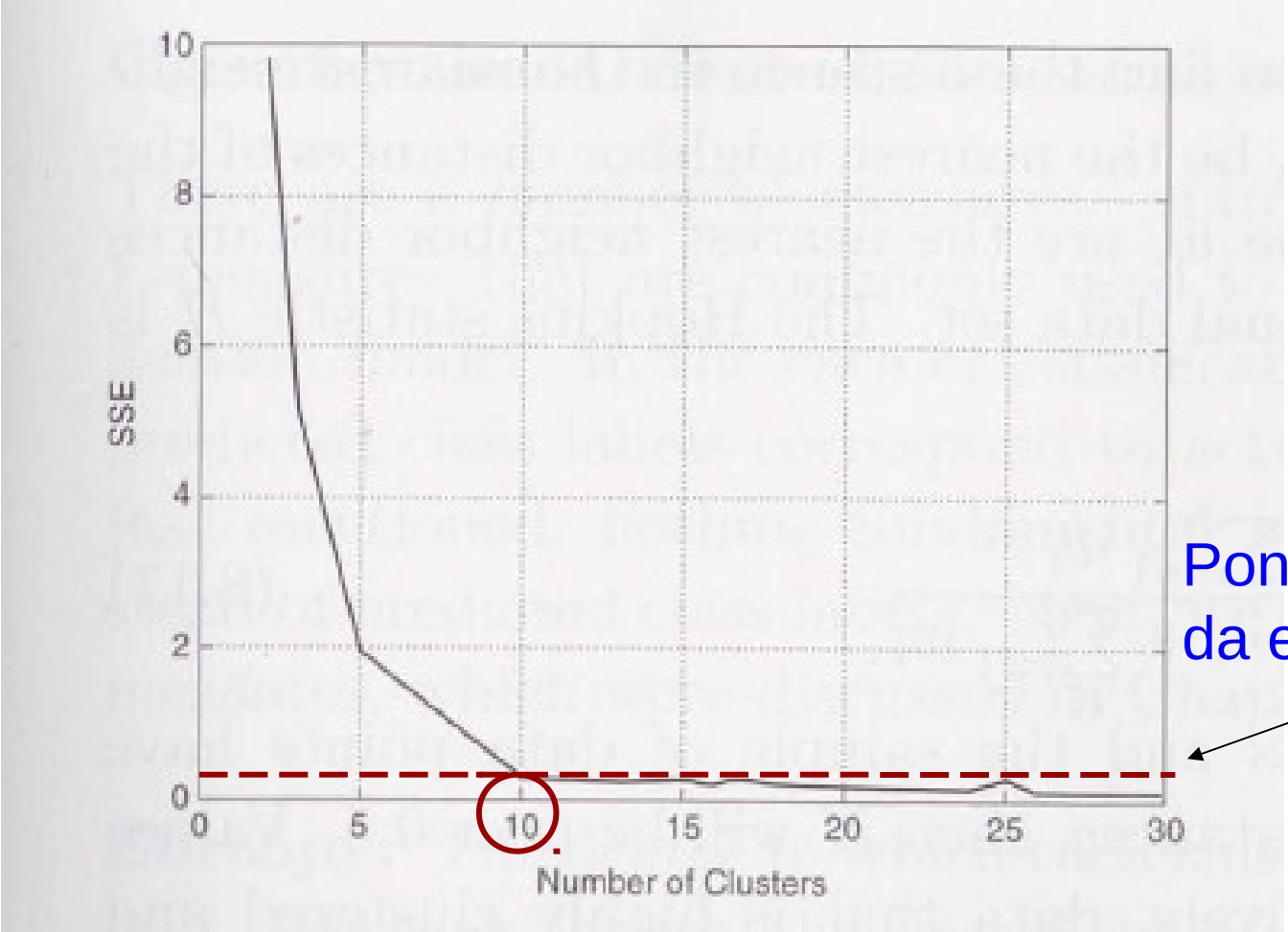


Escolhendo o valor de k

Número ideal;
distâncias curtas



Escolhendo o valor de k



Ponto mínimo antes da estabilização

Escolhendo o valor de k

Técnica 2

Executar o algoritmo K-means diversas vezes com diferentes números de clusters.

Calcular o **coeficiente de silhueta global** de cada clusterização obtida.

Plotar os valores dos coeficientes de silhueta (eixo y) por número de clusters (eixo x)

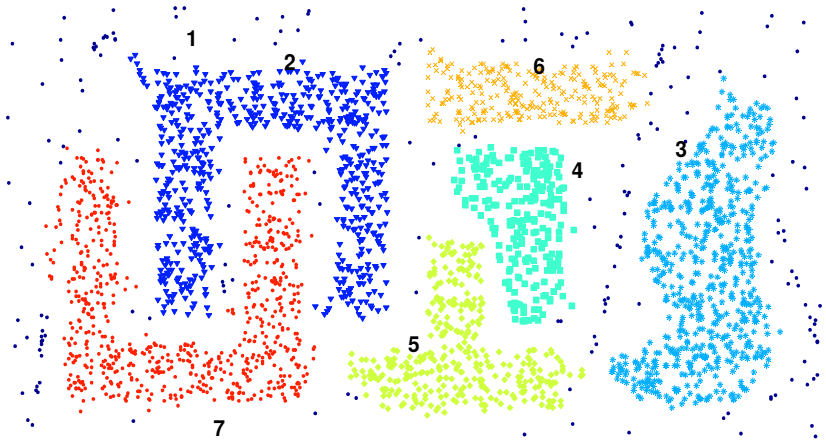
O número ideal de clusters corresponde a um momento onde se atinge um pico no gráfico.

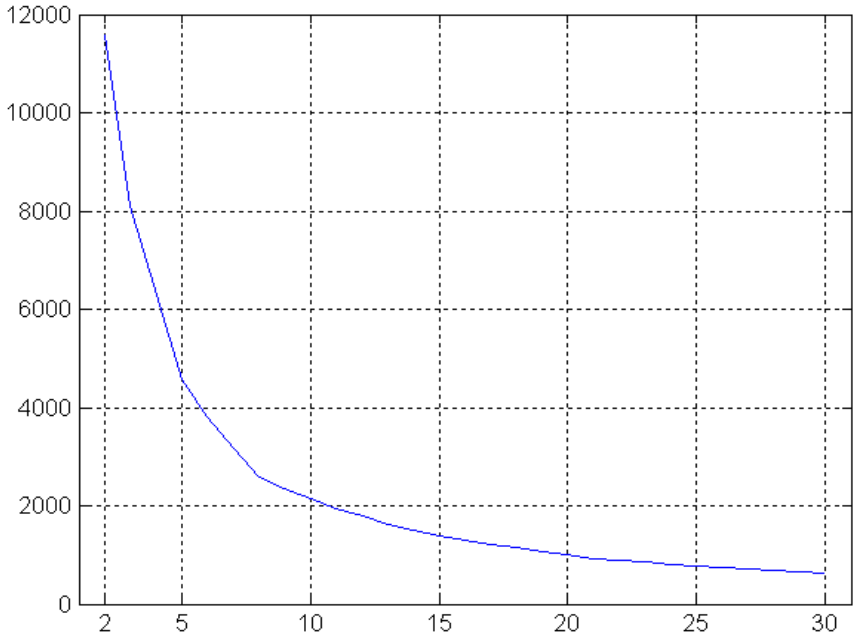
Escolhendo o valor de k

Ponto de Pico

Escolhendo o valor de k

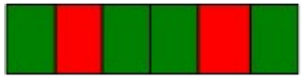
Curva SSE para um dataset mais complicado

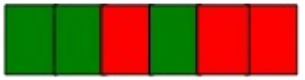




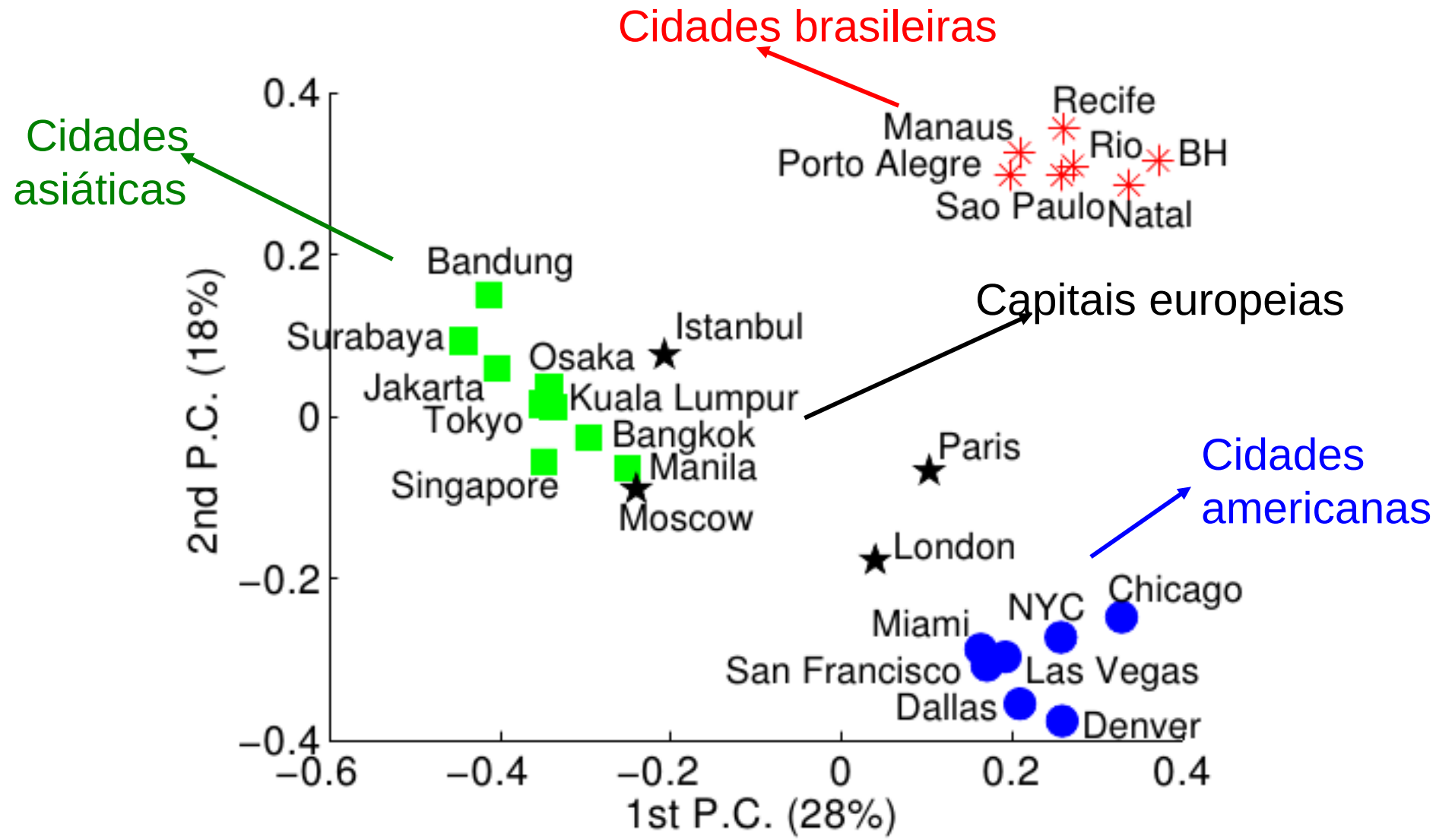
SSE de clusters usando o K-means

Correspondem a clusters reais?



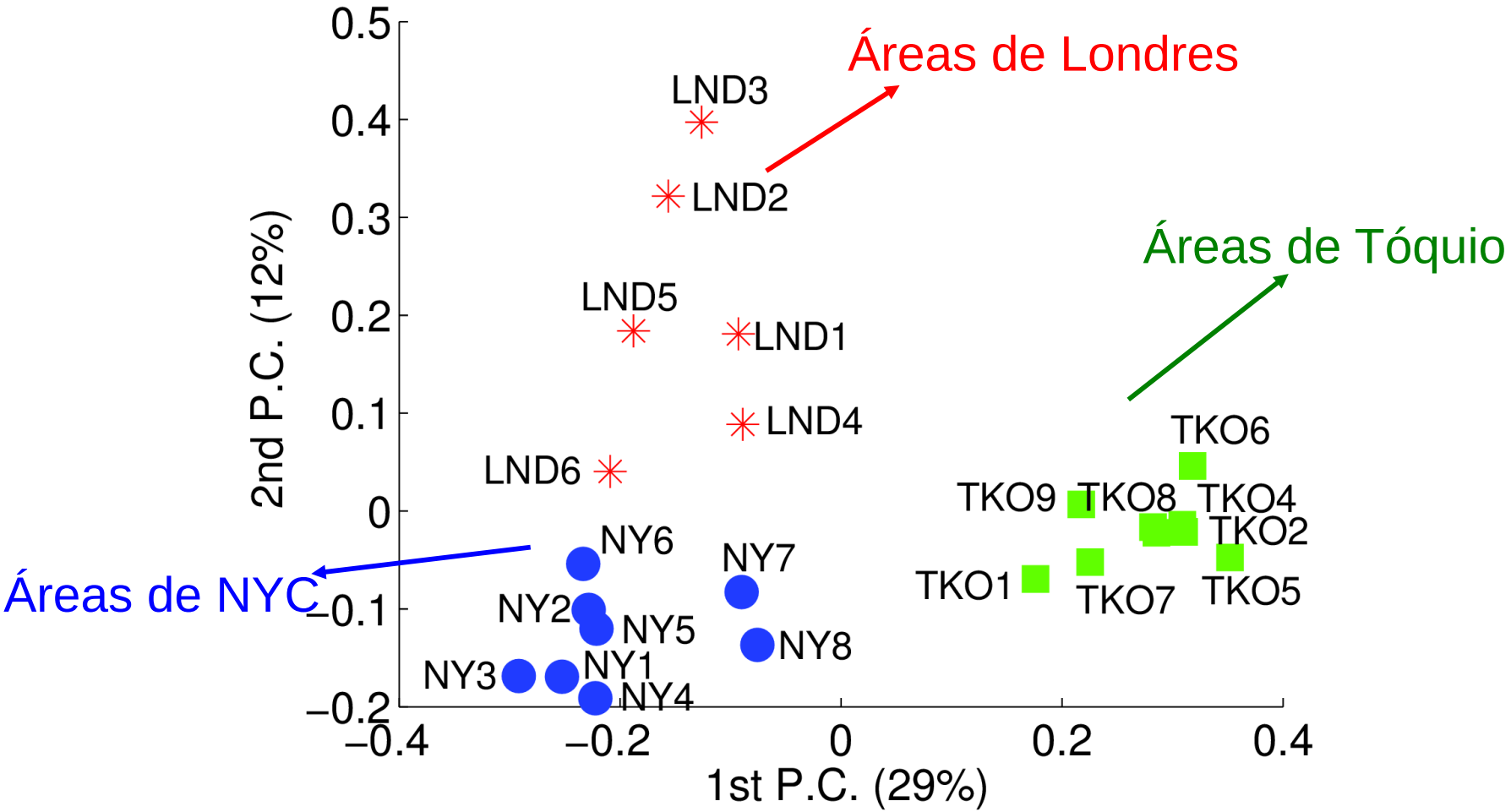


Correspondem a clusters reais?

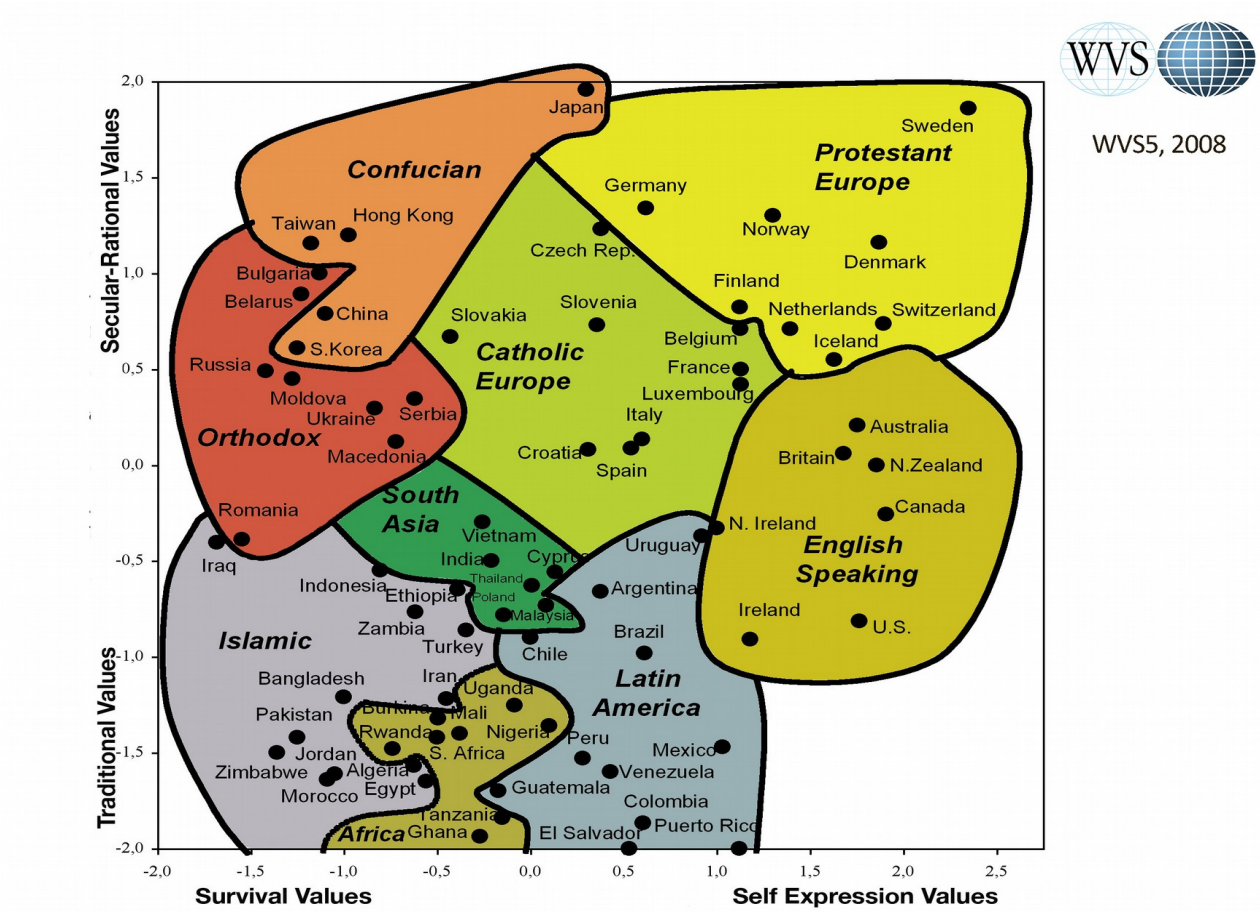


Correspondem a clusters reais?

Agrupamento de áreas dentro da cidade



Correspondem a clusters reais?



Comparando com dados do World Value Survey

A similaridade é muito boa!

Agradecimentos

Alguns slides foram derivados/inspirados em:

- Livro Introduction to Data Mining - Tan, Steinbach, Kumar.