

# Clustering

Sarajane Marques Peres

31 de outubro de 2020



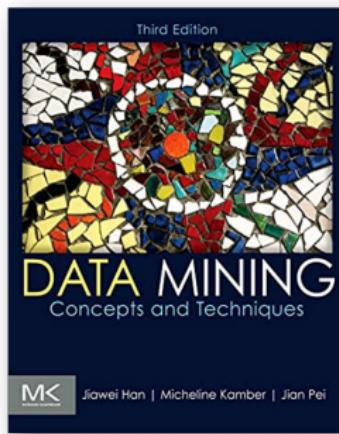
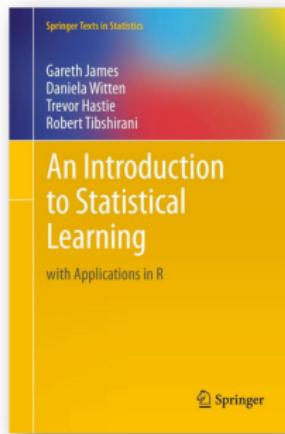
**WiMLDS**  
Women in Machine Learning & Data Science



**EACH** | campus capital  
Escola de Artes, Ciências e Humanidades  
Universidade de São Paulo

# Clustering

## Literatura recomendada



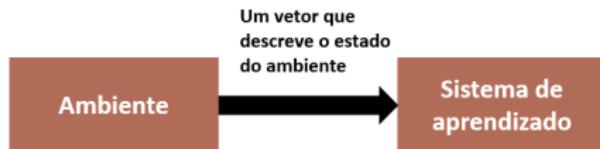
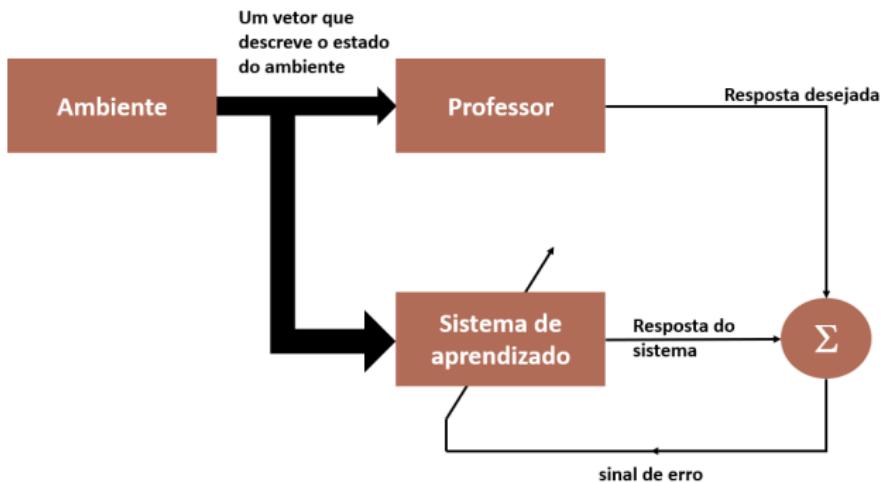
# Clustering & Aprendizado não supervisionado

\*Baseado em: Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall.



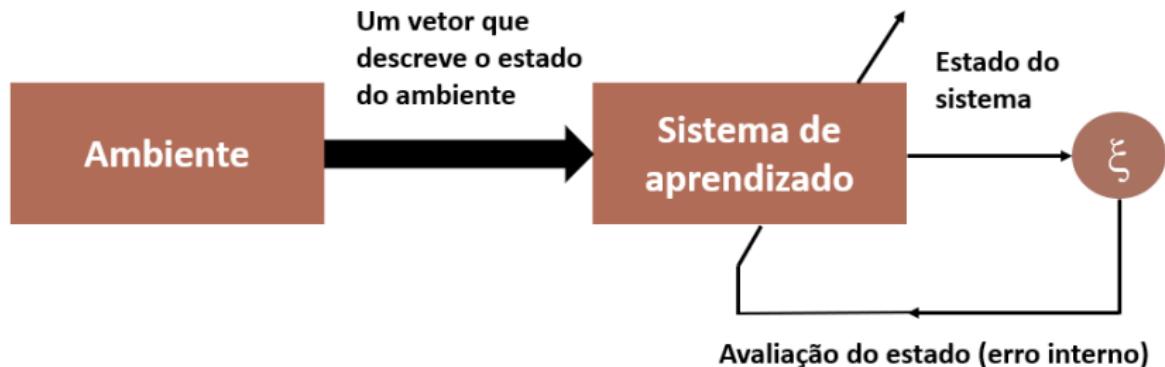
# Clustering & Aprendizado não supervisionado

\*Baseado em: Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall.



# Clustering & Aprendizado não supervisionado

\*Baseado em: Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall.



# Clustering

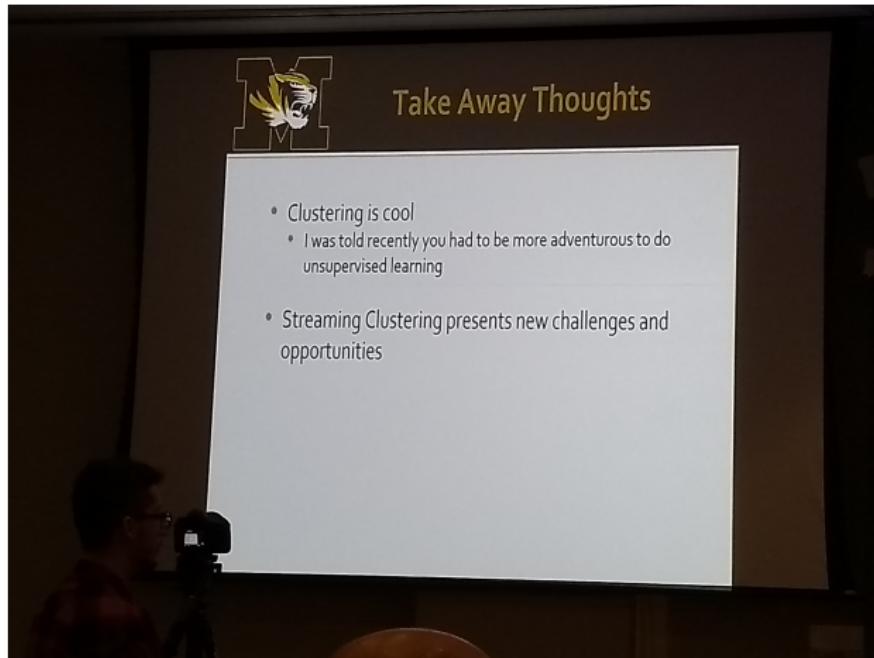
## Clustering - Agrupamento

O termo *grupo* deve ser usado quando não existe qualquer informação sobre como é a organização dos dados. Nesse caso, o trabalho de análise de dados é denominado *agrupamento (clustering)*, e tem por objetivo estudar as relações de similaridades entre os dados, determinando quais dados formam quais grupos.

Os grupos são formados de maneira a maximizar a similaridade entre os elementos de um grupo (similaridade intra-grupo) e minimizar a similaridade entre elementos de grupos diferentes (similaridade inter-grupos).



# Clustering



Mensagem de James M. Keller no WCCI 2018.

# Clustering

Para o contexto de nosso estudo, um conjunto de dados  $X$  é definido como:

$$X = \begin{matrix} \vec{x_1} & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ \vec{x_2} & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \dots & \dots & \dots & \dots & \dots \\ \vec{x_n} & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{matrix}$$

em que

- $\vec{x_j}$  é um vetor de  $p$  coordenadas. e
- $n$  é o número de elementos do conjunto de dados.

Cada vetor representa um dado desse conjunto e cada coordenada desse vetor representa um atributo descritivo do dado. O conjunto de dados  $X$  reside no espaço  $\mathbb{R}^p$ , e esse espaço é referenciado pelos algoritmos de análise de dados como “espaço dos dados”, “espaço de entrada” ou “espaço vetorial”.

# Clustering

Formalmente, dado um conjunto de dados de entrada  $\vec{x} \in \mathbb{R}^p$ , é encontrada uma função

$$\mathcal{G} : \mathbb{R}^p \times W \rightarrow C$$

em que  $W$  é um vetor de parâmetros ajustáveis, por meio de um algoritmo de aprendizado não supervisionado, que determina  $k$ -grupos em  $X$ ,  $C = C_1, \dots, C_k$ , ( $k \leq n$ ) tal que:

- $C_i \neq \emptyset, i = 1, \dots, k;$
- $\bigcup_{i=1}^k C_i = X;$
- $C_i \cap C_j = \emptyset, i, j = 1, \dots, k$  and  $i \neq j$ , assumindo a abordagem de agrupamento clássica.

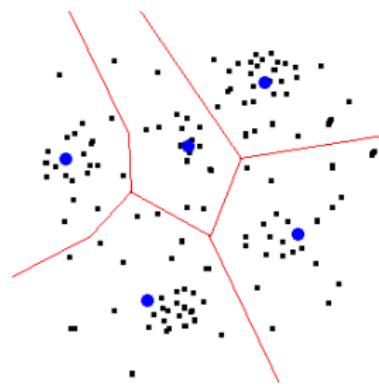
# Clustering

## Categorização de métodos de clustering

- Método por particionamento: dado um conjunto de dados com  $n$  instâncias, um método por particionamento constrói uma  $k$ -partição dos dados, na qual cada parte da partição representa um grupo e  $k \leq n$ . O método cria uma partição inicial e, então, usa uma técnica de realocação iterativa que tenta melhorar o particionamento.  
**Exemplo:** **k-Means** (ou k-Means), CLARANS.
- Métodos hierárquicos: cria uma decomposição hierárquica de um conjunto de dados. Os métodos hierárquicos podem ser *aglomerativos* ou *divisivos*, dependendo de como a decomposição hierárquica é formada - juntando decomposições ou dividindo composições. A cada passo, divisões ou junções são feitas. Podem representar seus resultados em dendogramas. **Exemplo:** AGNES/DIANA, BIRCH, ROCK, Chameleon.
- Métodos baseados em densidade: No caso dos métodos baseados em densidade, os grupos formados crescem de acordo com a densidade de dados em um "potencial" grupo. Para cada dado dentro de um grupo, a vizinhança em um raio tem que conter pelo menos um número mínimo de pontos. **Exemplo:** DBSCAN, OPTICS, DENCLUE. Também são considerados métodos por particionamento baseados em densidade.
- Métodos baseados em modelos: criam uma hipótese sobre um modelo para cada um dos grupos e encontram o melhor ajuste dos dados ao modelo. **Exemplo:** Self-Organizing Map (SOM), Expectation-Maximization (EM).
- Métodos baseados em *grid*: esses métodos quantizam o espaço de dados em um número finito de células que forma uma estrutura em *grid*. **Exemplo:** STING, WaveCluster

# Clustering

Por particionamento ....



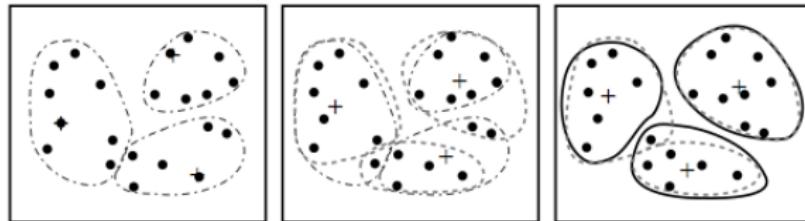
# Clustering - métodos por particionamento

## A K-Means Clustering Algorithm

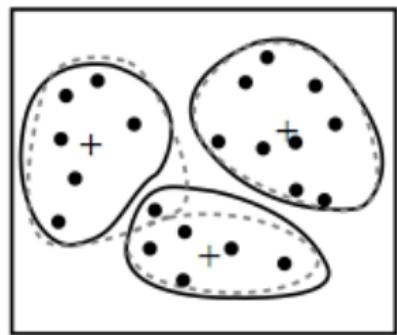
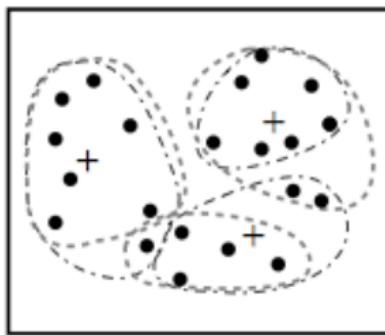
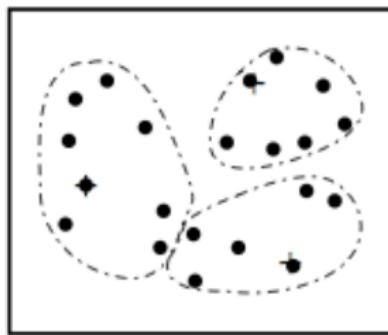
J. A. Hartigan and M. A. Wong. Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, n. 1, 1979. 100-108p.

### k-Means

Nesse algoritmo,  $k$  agrupamentos são representados como um conjunto  $\mathcal{C} = \{\vec{c}_1, \dots, \vec{c}_k\}$  de vetores chamados “protótipos”. Cada vetor protótipo sempre está associado à representação de um grupo do conjunto de dados e, para isso, deve residir no mesmo espaço  $\mathbb{R}^p$  que os dados do conjunto. O conjunto  $\mathcal{C}$  é representado por uma matriz de dimensão  $k \times p$ .



# Clustering - k-Means



# Clustering - k-Means

## k-Means

Para alcançar seu objetivo, o algoritmo realiza várias iterações na busca de uma configuração ótima de parâmetros para minimizar  $J_{CM}(U_h, C)$ , que é dado por:

$$J_{CM}(U_h, C) = \sum_{i=1}^k \sum_{j=1}^n u_{ij} d(\vec{C}_i, \vec{x}_j)^2 \quad (1)$$

em que  $d(\vec{C}_i, \vec{x}_j)$ , é a distância entre o vetor de dados  $\vec{x}_j$  e o protótipo do grupo  $\vec{C}_i$ ,  $k$  é o número de grupos a ser determinado pelo algoritmo,  $n$  é o número de dados no conjunto de dados e  $U_h$  é uma matriz binária chamada “matriz de partição”, de dimensões  $k \times n$ , definida como:

$$U_h = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ u_{i+1,1} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ u_{k,1} & u_{k,2} & \cdots & u_{k,n} \end{bmatrix}$$

# Clustering - k-Means

O processo de minimização deve (ou deveria) obedecer as seguintes restrições:

## Restrição 1

$$\sum_{i=1}^k u_{ij} = 1, \forall j \in 1, \dots, n.$$

garantindo que a soma das pertinências de um dado  $\vec{x}_j$  a todos os grupos em  $C$  seja igual a 1, ou seja, cada coluna da matriz de partição deve possuir o valor 1 em **uma e somente uma** célula.

## Restrição 2 - não garantida no procedimento padrão do k-Means

$$\sum_{j=1}^n u_{ij} \geq 1, \forall i \in 1, \dots, k.$$

tal que cada linha da matriz de partição deve possuir o valor 1 em **pelo menos** uma célula. Para garantir que todos os  $k$  grupos tenham, ao menos, um dado associado.

# Clustering - k-Means

No processo de minimização de  $J_{CM}$ , tanto  $U_h$  quanto  $\mathcal{C}$  devem ser atualizados:

## Atualização de $U_h$

$$u_{ij}^{t+1} = \begin{cases} 1, & \text{se } i = \arg \min_{i=1}^k d(\vec{C}_i, \vec{x}_j) \\ 0, & \text{caso contrário.} \end{cases} \quad (2)$$

em que  $t$  é o contador de iterações do processo de otimização e  $u_{ij}^{t+1}$  é o valor da pertinência do dado  $j$  ao grupo  $i$  na iteração  $t + 1$ . A atualização faz com que cada dado seja associado ao grupo cujo protótipo é o mais próximo a ele (possui a distância mínima) dentre todos os protótipos.

## Atualização de $\mathcal{C}$

$$\vec{C}_i^{t+1} = \frac{\sum_{j=1}^n u_{ij}^{t+1} \vec{x}_j}{\sum_{j=1}^n u_{ij}^{t+1}} \quad (3)$$

estabelece novos vetores protótipos para os grupos de acordo com a média de todos os vetores de dados associados a eles. O numerador soma, para cada grupo, os vetores de dados associados a eles. O denominador termina o processo de média.

# Clustering - k-Means

## Algoritmo k-Means

- a. determine o número de grupos (clusters -  $k$ );
- b. determine um valor pequeno e positivo para um erro máximo ( $\epsilon$ );
- c. inicialize o conjunto de protótipos  $\mathcal{C}$  aleatoriamente, escolhendo  $k$  vetores protótipos dentro do menor intervalor que contém todos os dados do conjunto  $X$ ;
- d. inicialize um contador de iterações  $t$  como  $t = 0$ ;
- e. **repita**
  - $t++$ ;
  - atualize  $U_h$  de acordo com (2);
  - atualize  $\mathcal{C}$  de acordo com (3);
- e. **até que**  $||\mathcal{C}^t - \mathcal{C}^{(t-1)}|| \leq \epsilon$

# Clustering - k-Means

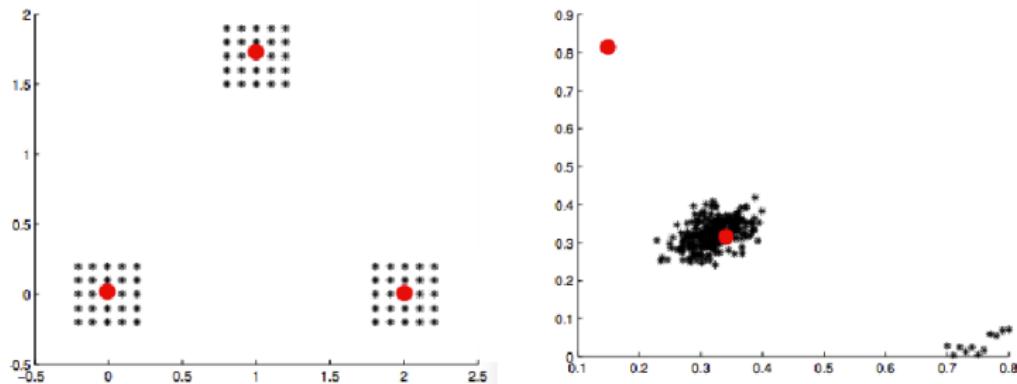
## Algoritmo k-Means

- a. determine o número de grupos (clusters -  $k$ );
- b. determine um valor pequeno e positivo para um erro máximo ( $\epsilon$ );
- c. inicialize o conjunto de protótipos  $\mathcal{C}$  aleatoriamente, escolhendo  $k$  vetores protótipos dentro do menor intervalor que contém todos os dados do conjunto  $X$ ;
- d. inicialize um contador de iterações  $t$  como  $t = 0$ ;
- e. repita
  - $t++$ ;
  - atualize  $U_h$  de acordo com (2);
  - atualize  $\mathcal{C}$  de acordo com (3);
- e. até que  $||\mathcal{C}^t - \mathcal{C}^{(t-1)}|| \leq \epsilon$

## Atualização passo a passo

Uma maneira alternativa de atualizar os centróides é fazer a atualização incremental, ou seja, atualizar os centróides depois de cada ponto ser associado a um grupo.

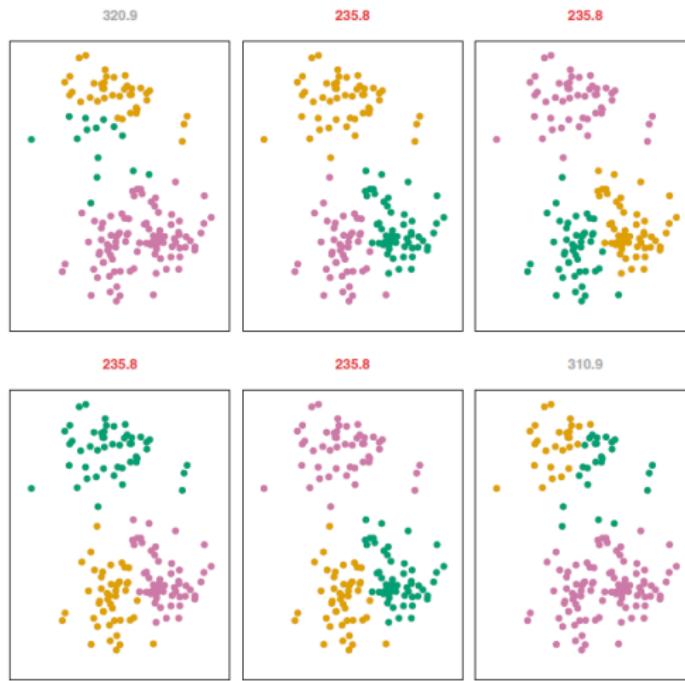
# Clustering - k-Means



Nos exemplos, o primeiro caso foi bem resolvido pelo k-Means, já o segundo caso não foi bem resolvido.

# Clustering - k-Means

Fonte: James, G; Witten, D.; Hastie, T.; Tibshirani, R. (2017) An Introduction to Statistical Learning with Applications in R. Springer.



# Clustering - k-Means ++

## *k-means++: The advantages of careful seeding*

David Arthur e Sergei Vassilvitskii<sup>a</sup>: propuseram uma variação para o k-Means que escolhe randomicamente os centros a partir do conjunto de dados, mas atribui pesos para os dados escolhidos de acordo com as suas distâncias do centro mais próximo já escolhido.

---

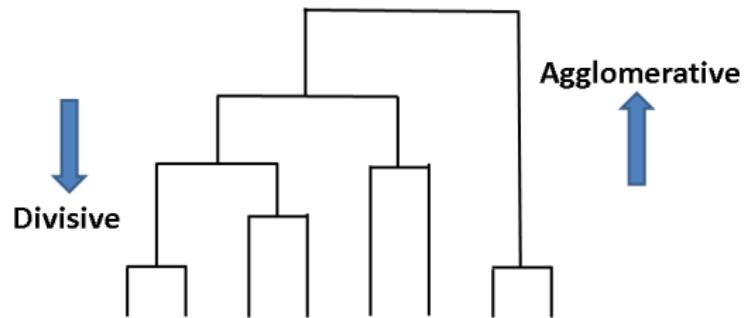
<sup>a</sup> Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035 (2007).

## k-Means ++

Melhorar o processo de agrupamento realizado pelo c-Means, por meio de um procedimento otimizado de inicialização dos protótipos (centróides). O processo melhora tanto em termos de velocidade (número de iterações necessárias para convergência) quanto em termos de qualidade de agrupamento.

# Clustering

Hierárquico ....



# Clustering - métodos hierárquicos

Neste método de agrupamento, os dados são agrupados em “dendogramas”. Os métodos podem ser **aglomerativos** ou **divisivos**, dependendo se a decomposição hierárquica é formada usando uma estratégia *bottom-up* (merge) ou *top-down* (*split*).

Esses algoritmos, em sua forma pura, sofrem do problema de não poderem executar ajustes uma vez que foi tomada uma decisão sobre juntar grupos ou dividir grupos. Isso pode levar à necessidade de alterar o método, mesclando-o com outras estratégias.

# Clustering - métodos hierárquicos

## Clustering Hierárquico Aglomerativo

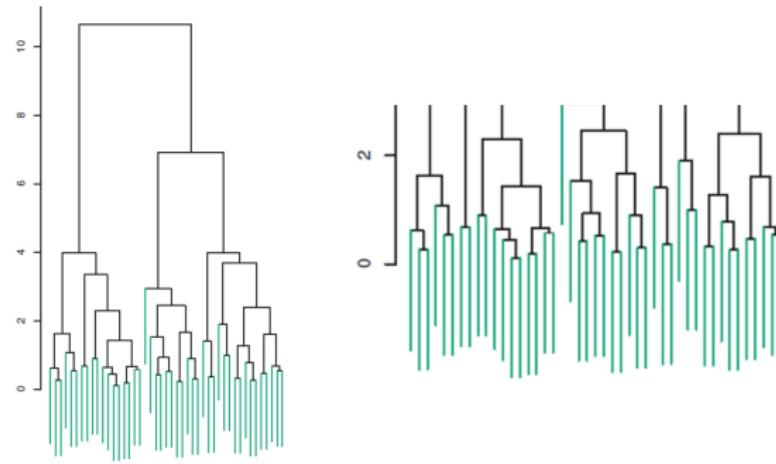
A estratégia *bottom-up* inicia pela alocação de cada objeto em seu próprio cluster. Então, junções destes clusters (atômicos) são realizadas, formando clusters cada vez maiores, até que todos os objetos sejam alocados em um único cluster, ou alguma condição de parada seja satisfeita.

## Clustering Hierárquico Divisivo

A estratégia *top-down* inicia com todos os objetos em um cluster. Então, divide o cluster em pedaços menores, até que cada objeto forme o seu próprio cluster, ou até que uma determinada condição de parada seja satisfeita (por exemplo, um número desejado de clusters, ou o diâmetro de cada cluster atingir um limiar).

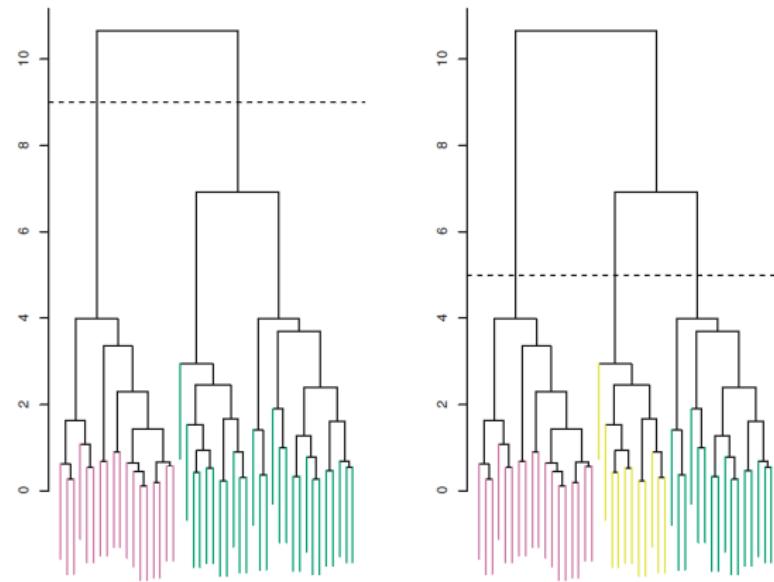
# Clustering - Um dendrograma

O quanto antes (mais próximo das folhas) a fusão de ramos (ou de folhas e ramos) ocorre, mais similares são os grupos de observações envolvidos. Por outro lado, as observações envolvidas em fusões mais próximas da raíz da do dendrograma podem ser muito diferentes entre si (James et al., 2017).



# Clustering - Um dendrograma

Para quaisquer duas observações, nós podemos procurar o ponto no dendrograma onde onde os ramos que contem as observações se fundiram. A altura dessa fusão, olhando para a medida no eixo vertical, indica o quanto diferentes essas observações são (James et al., 2017).



# Clustering - Métodos Hierárquicos

## Abordagens usadas nos algoritmos hierárquicos

Na abordagem **single linkage**, cada cluster é representado por todos os objetos nele contidos, e a similaridade entre dois clusters é representada pela distância entre pares de dados (objetos) mais próximos e pertencentes a clusters diferentes.

# Clustering - Métodos Hierárquicos

## Abordagens usadas nos algoritmos hierárquicos

Na abordagem **single linkage**, cada cluster é representado por todos os objetos nele contidos, e a similaridade entre dois clusters é representada pela distância entre pares de dados (objetos) mais próximos e pertencentes a clusters diferentes.

A abordagem **complete linkage** usa a maior distância entre dois grupos. É o método do vizinho mais distante. A distância entre dois clusters é determinada de acordo com a maior distância entre um par de dados, sendo cada dado pertencente a um cluster distinto.

# Clustering - Métodos Hierárquicos

## Abordagens usadas nos algoritmos hierárquicos

Na abordagem **single linkage**, cada cluster é representado por todos os objetos nele contidos, e a similaridade entre dois clusters é representada pela distância entre pares de dados (objetos) mais próximos e pertencentes a clusters diferentes.

A abordagem **complete linkage** usa a maior distância entre dois grupos. É o método do vizinho mais distante. A distância entre dois clusters é determinada de acordo com a maior distância entre um par de dados, sendo cada dado pertencente a um cluster distinto.

A abordagem **average linkage** usa a média entre as distâncias. Ou seja, é calculada a média das distâncias entre todos os pares de dados de dois grupos. Os pares de grupos que apresentarem a menor média são mais similares.

# Clustering - Métodos Hierárquicos

## Abordagens usadas nos algoritmos hierárquicos

Na abordagem **single linkage**, cada cluster é representado por todos os objetos nele contidos, e a similaridade entre dois clusters é representada pela distância entre pares de dados (objetos) mais próximos e pertencentes a clusters diferentes.

A abordagem **complete linkage** usa a maior distância entre dois grupos. É o método do vizinho mais distante. A distância entre dois clusters é determinada de acordo com a maior distância entre um par de dados, sendo cada dado pertencente a um cluster distinto.

A abordagem **average linkage** usa a média entre as distâncias. Ou seja, é calculada a média das distâncias entre todos os pares de dados de dois grupos. Os pares de grupos que apresentarem a menor média são mais similares.

**Centroid-linkage** usa o vetor protótipo do cluster para o cálculo da similaridade entre clusters. A similaridade entre os clusters é definida com base na distância euclidiana entre os protótipos dos clusters.

# Clustering - Métodos Hierárquicos

## Abordagens usadas nos algoritmos hierárquicos

Na abordagem **single linkage**, cada cluster é representado por todos os objetos nele contidos, e a similaridade entre dois clusters é representada pela distância entre pares de dados (objetos) mais próximos e pertencentes a clusters diferentes.

A abordagem **complete linkage** usa a maior distância entre dois grupos. É o método do vizinho mais distante. A distância entre dois clusters é determinada de acordo com a maior distância entre um par de dados, sendo cada dado pertencente a um cluster distinto.

A abordagem **average linkage** usa a média entre as distâncias. Ou seja, é calculada a média das distâncias entre todos os pares de dados de dois grupos. Os pares de grupos que apresentarem a menor média são mais similares.

**Centroid-linkage** usa o vetor protótipo do cluster para o cálculo da similaridade entre clusters. A similaridade entre os clusters é definida com base na distância euclidiana entre os protótipos dos clusters.

**Ward-linkage** diz que a distância entre dois clusters é o quanto o **erro de quantização** aumenta quando nós juntamos dois grupos. Em clustering hierárquico aglomerativo, o erro inicia em zero e cresce conforme juntamos os grupos. O método Ward mantém esse crescimento tão pequeno quanto possível.

# Clustering - Métodos Hierárquicos

AGNES (AGglomerative NESting) X DIANA (DIvisive ANAlysis)

**AGNES:** Inicialmente coloca cada objeto em seu próprio cluster, e depois junta os clusters, passo a passo, de acordo com algum critério (por exemplo, usando a abordagem *single-linkage*, cluster C1 e C2 se juntam se um objeto em C1 e um objeto em C2 possuem a distância Euclidiana mínima entre quaisquer dois objetos de clusters diferentes.) O processo de união (*merge*) continua até que só exista um cluster.

**DIANA:** Todos os objetos são usados para formar um cluster inicial. Para a divisão, o dado menos semelhante a todos os outros é selecionado para conduzir a formação de um novo cluster. Então, são buscados dentro do cluster original, os elementos que são mais semelhantes (de acordo com uma métrica de similaridade) ao novo cluster do que ao cluster original. Esses dados são transladados para o novo grupo.

# AGNES – AGglomerative NESting

A partir de um conjunto de dados, da escolha de uma medida de similaridade e da escolha de uma abordagem para cálculo da distância entre grupos:

- calcule uma matriz de similaridades entre os dados do conjunto de dados (para consulta);
- aloque cada instância (dado) do conjunto de dados em um grupo e assuma cada um deles como os nós folhas de uma árvore;
- realize a fusão de grupos enquanto for possível, da seguinte forma:
  - verifique qual é a similaridade entre cada par de grupos;
  - encontre o par de grupos mais similar e o transforme em um único grupo, criando um nó interno na árvore;

# DIANA - DIvisive ANAlysis

A partir de um conjunto de dados, da escolha de uma medida de similaridade e da escolha de uma abordagem para cálculo da distância entre grupos:

- calcule uma matriz de similaridades entre os dados do conjunto de dados (para consulta);
- aloque todos os exemplares em um único grupo criando a raiz de uma árvore;
- realize a divisão de grupos enquanto for possível, da seguinte forma:
  - selecione o grupo de maior dispersão;
  - dentro deste grupo encontre o exemplar menos similar a todos os demais;
  - crie um novo grupo com este exemplar e reorganize os dados no grupo de maior dispersão nesse novo, de acordo com a medida de similaridade;
  - organize os dois novos grupos na árvore, como subárvore do nó referente ao grupo original;

# Clustering

Avaliação ....

On Clustering Validation Techniques Maria Halkidi, Yannis Batistakis,  
Michalis Vazirgiannis Journal of Intelligent Information Systems, 17:2/3,  
107-145 2001

# Motivação

## Avaliação → Validação

O processo de avaliação do resultado obtido a partir de um algoritmo de agrupamento é comumente chamado de validação.

## Objetivo

A pergunta a ser respondida é se o modelo de grupos descoberto é, de fato, a organização em grupos dos dados sob análise. **Porém, se não conhecemos a organização, como saber se o que descobrimos é o que deveríamos ter descoberto?**

## Estratégias

- analisar a compacidade: encontramos grupos que **maximizaram a similaridade intragrupo?**
- analisar a separabilidade: encontramos grupos que **minimizaram a similaridade intergrupos?**
- analisar conhecimento **a priori**: usar informações que já se tem sobre o conjunto de dados sob análise para validar os grupos encontrados.

# Clustering - Avaliação

- compacidade: os membros de cada grupo deveriam ser tão próximos entre si quanto possível. Uma medida comum para a compacidade é a variância, a qual deve ser minimizada.
- separação: os grupos devem ser largamente espaçados entre si. As medidas comuns para isso são: single linkage, complete linkage, average linkage e centroid-linkage.

# Índice Interno - Silhouette

## Índice Silhouette - conceito básico

$$I_{SIL}(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

em que

- $a(i)$  é a distância média do dado  $i$  a todos os demais dados do seu grupo;
- $b(i)$  é a distância média mínima do dado  $i$  a todos os dados de cada um dos demais grupos (excluindo o seu); i.e., é a distância média do exemplar  $i$  a todos os demais exemplares do grupo mais próximo ao seu.

O  $I_{SIL}(i)$  é calculado por dado e o  $I_{SIL}$  de um grupo é a média dos  $I_{SIL}(i)$  de todos os dados ( $i$ ) no grupo. E o  $I_{SIL}$  do agrupamento é a média dos  $I_{SIL}$  dos grupos. Quanto MAIOR o valor do índice MELHOR.

## Artigo original

Rousseeuw, J. Sillhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, Volume 20, 1987, pages 53-65.

# Índice Interno - Silhouette

O índice Sillhoute ( $I_{SIL}$ ) varia entre  $-1$  e  $1$ .

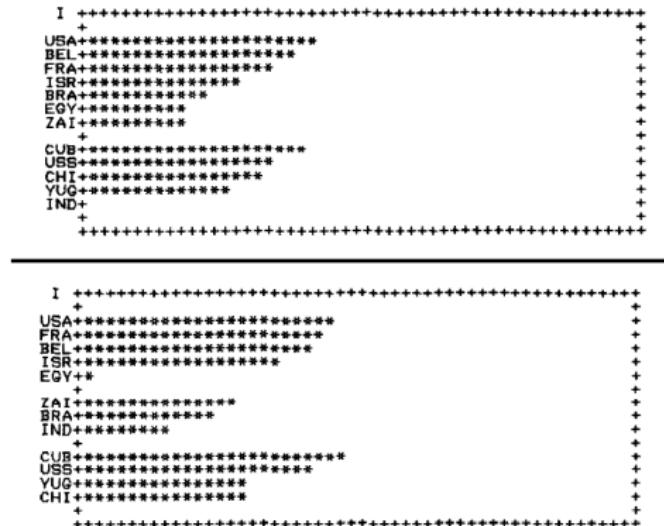
Significado do índice:

- Quando  $I_{SIL}(i)$  alcança o seu maior valor (1) significa que a dissimilaridade “dentro” do grupo ( $a(i)$ ) é muito menor do que a menor dissimilaridade “entre” grupos ( $b(i)$ ). Portanto, dizemos que o ponto  $i$  está bem agrupado.
- Quando  $I_{SIL}(i)$  se aproxima de 0,  $a(i)$  e  $b(i)$  são aproximadamente iguais. Isso significa que não está claro se o dado  $i$  deveria estar associado ao grupo  $A$  ou ao grupo  $B$ ;
- Quando  $I_{SIL}(i)$  alcança o seu menor valor ( $-1$ ), significa que a dissimilaridade “dentro” do grupo ( $a(i)$ ) é muito maior do que a menor dissimilaridade “entre” grupos ( $b(i)$ ). Portanto, dizemos que o ponto  $i$  está mal agrupado.

# Índice Interno - Silhouette

## Visualização gráfica

O Sillhouette do grupo  $A$  é a plotagem dos  $I_{SIL}(i)$ , ordenados de forma decrescente, para todos os objetos  $i$  em  $A$ . A silhueta mostra quais objetos “caem bem” dentro de seus grupos.



# Índice Interno - Silhouette

CLU	NEIG	S(I)	I
1	2	.43	USA+***** BEL+***** FRA+***** ISR+***** BRA+***** EGY+***** ZAI+***** CUB+***** USS+***** CHI+***** YUG+***** IND+
	2	.39	
	2	.35	
	2	.30	
	2	.22	
	2	.20	
	2	.19	
	2	.40	
	2	.34	
	2	.33	
2	.26		
2	-.04		

0 1  
0 0 0 1 1 2 2 2 3 3 4 4 4 4 5 5 6 6 6 6 7 7 8 8 8 8 9 9 0  
0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0

CLUSTER 1 HAS AVERAGE SILHOUETTE WIDTH .30  
CLUSTER 2 HAS AVERAGE SILHOUETTE WIDTH .26  
FOR THE ENTIRE DATASET, THE AVERAGE SILHOUETTE WIDTH IS .28

Fonte: Rousseeuw, J. (1987)

# Índice Interno - Silhouette

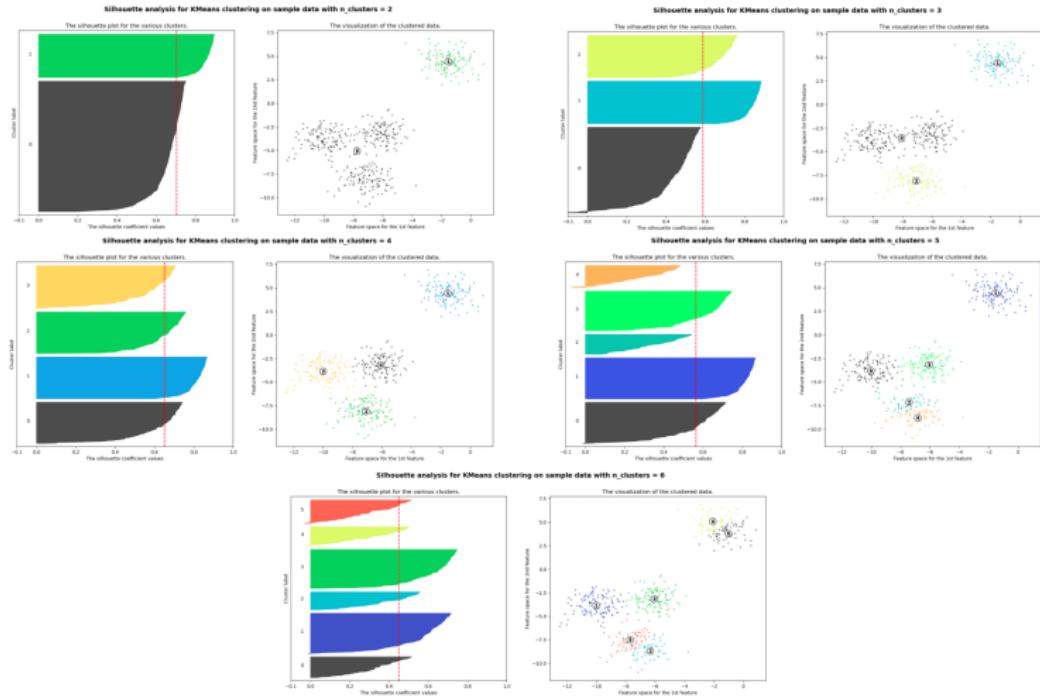
CLU	NEIG	S(I)	I
1	2	.47	USA+***** FRA+***** BEL+***** ISR+***** EGY++
1	20	.44	
1	21	.42	
1	22	.37	
1	23	.02	
2	1	.28	ZAI+*****
2	2	.25	BRA+*****
2	3	.17	IND+*****
3	2	.48	CUB+*****
3	1	.44	USS+*****
3	1	.31	YUG+*****
3	2	.31	CHI+*****
			+ +++++
			0 1
			0 0 0 1 1 2 2 2 3 3 4 4 4 5 5 6 6 6 7 7 8 8 8 9 9 0
			0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0 4 8 2 6 0
CLUSTER 1 HAS AVERAGE SILHOUETTE WIDTH .34			
CLUSTER 2 HAS AVERAGE SILHOUETTE WIDTH .24			
CLUSTER 3 HAS AVERAGE SILHOUETTE WIDTH .38			
FOR THE ENTIRE DATASET, THE AVERAGE SILHOUETTE WIDTH IS .33			

Fonte: Rousseeuw, J. (1987)

# Escolhendo o número de grupos (k)

[http://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_silhouette\\_analysis.html](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)

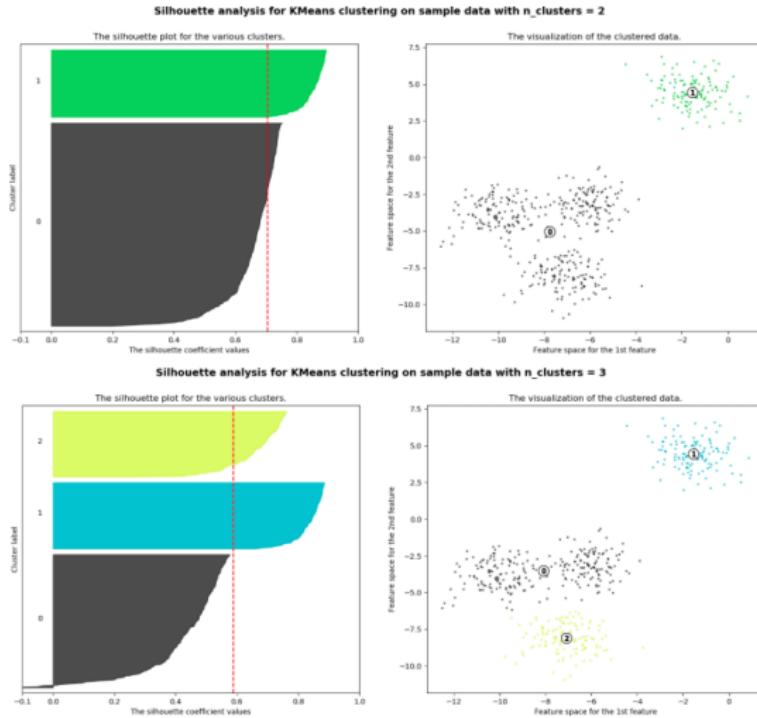
As situações mais favoráveis são aquelas com número de grupos igual a 2 e 4.



# Escolhendo o número de grupos (k)

[http://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_silhouette\\_analysis.html](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)

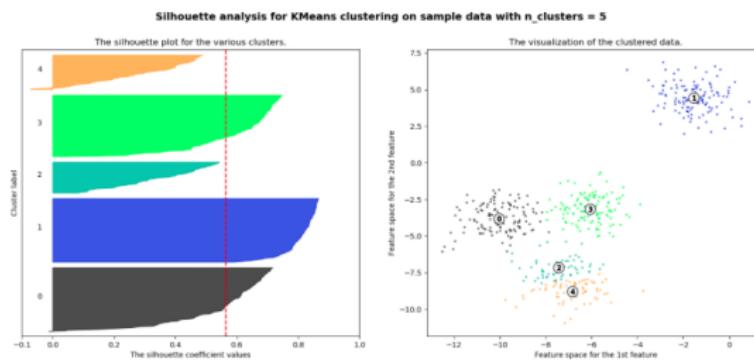
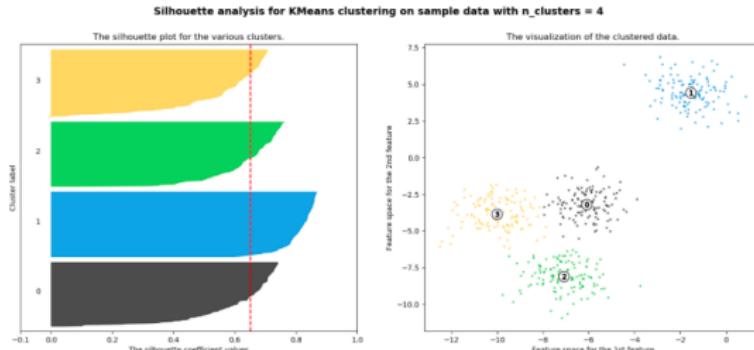
As situações mais favoráveis são aquelas com número de grupos igual a 2 e 4.



# Escolhendo o número de grupos (k)

[http://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_silhouette\\_analysis.html](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)

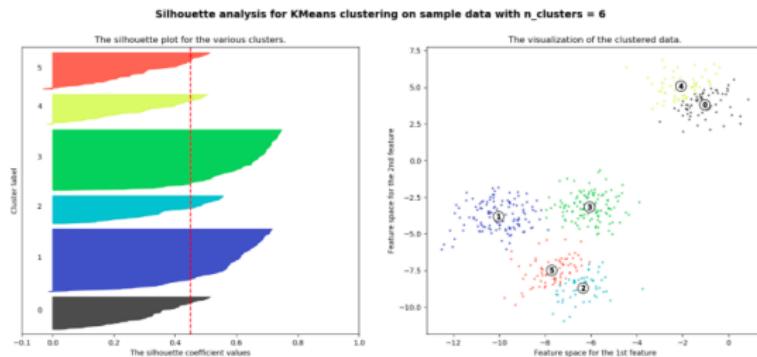
As situações mais favoráveis são aquelas com número de grupos igual a 2 e 4.



# Escolhendo o número de grupos (k)

[http://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_silhouette\\_analysis.html](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)

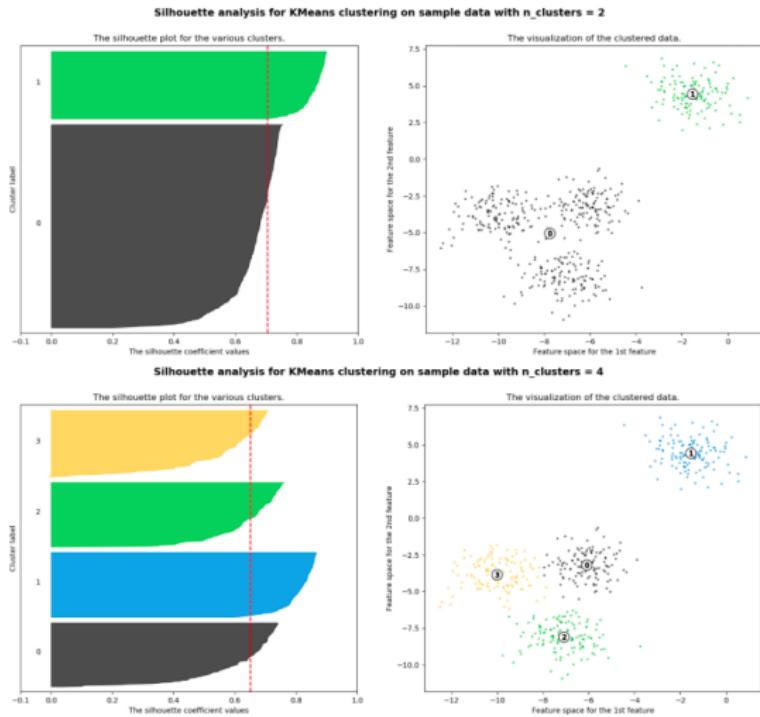
As situações mais favoráveis são aquelas com número de grupos igual a 2 e 4.



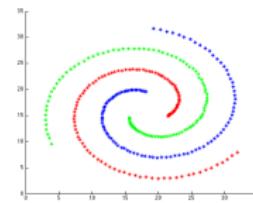
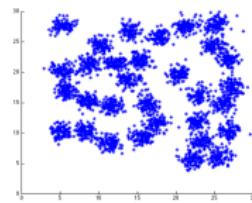
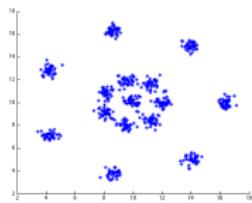
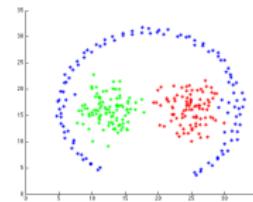
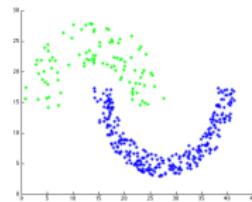
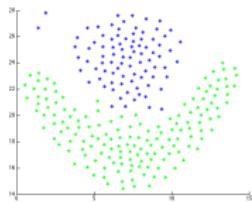
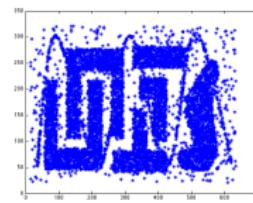
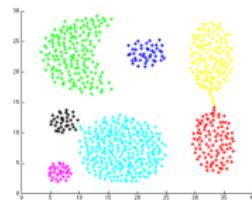
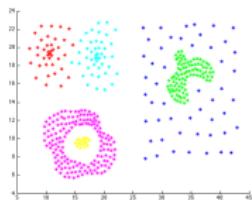
# Escolhendo o número de grupos (k)

[http://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_silhouette\\_analysis.html](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)

As situações mais favoráveis são aquelas com número de grupos igual a 2 e 4.



# Exemplos de organizações em grupos



# Escolha do melhor modelo de agrupamento

A melhor partição/clustering pode estar relacionada à qualidade dos grupos encontrados e/ou à quantidade de grupos encontrados. Muito provavelmente, a melhor qualidade estará relacionada com quantidade ideal.

## Estratégia

- crie vários modelos de agrupamento para o conjunto de dados sob análise, variando, sistematicamente, o número de grupos e os demais parâmetros do algoritmo;
- para cada modelo de agrupamento compute um índice de qualidade;
- selecione o modelo de agrupamento que gerou o MELHOR valor para o índice de qualidade.

- Davies, D. L.; Bouldin, D. W. A Cluster Separation Measure. In: IEEE Transaction on Pattern Analysis and Machine Intelligence, v.1, no 2, p. 224-227, 1979.
- Desgraupes, B. Clustering Indices. Package clusterCrit for R. University Paris Ouest - Lab Modal'X, 2013.
- Dunn, J. C. A Fuzzy Relative of the ISODATA Process and its Use in Detection Compact Well-Separate Clusters. In. Journal of Cybernetics, v. 3, no 3, p. 32-57, 1973.
- Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On Clustering Validation Techniques. In: Journal of Intelligent Information Systems, v. 17, no 2-3, p. 107-145, 2001.
- Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. In: Journal of Computational and Applied Mathematics, v.20, no 1, p. 53-65, 1986.



## Clustering

- Sarajane Marques Peres - [sarajane@usp.br](mailto:sarajane@usp.br)
- [each.usp.br/processmining](http://each.usp.br/processmining)
- [www.linkedin.com/in/sarajane-marques-peres](https://www.linkedin.com/in/sarajane-marques-peres)

Escola de Artes, Ciências e Humanidades - EACH  
Universidade de São Paulo - USP