Data Science
Bootcamp

WOMAKERS CODE

Host Oficial

sãojudas universidade

MÓDULO #4.1

Clustering

Jéssica dos Santos
Vivian Yamassaki

Jéssica dos Santos

Cientista de Dados na NeuralMed

j3ssicaSant0s

jessica-santos-oliveira

Vivian Mayumi Yamassaki

Cientista de Dados na Creditas

vivianyamassaki

vivianyamassaki

Comentamos sobre clustering na aula passada....

Alguém lembra o que é?

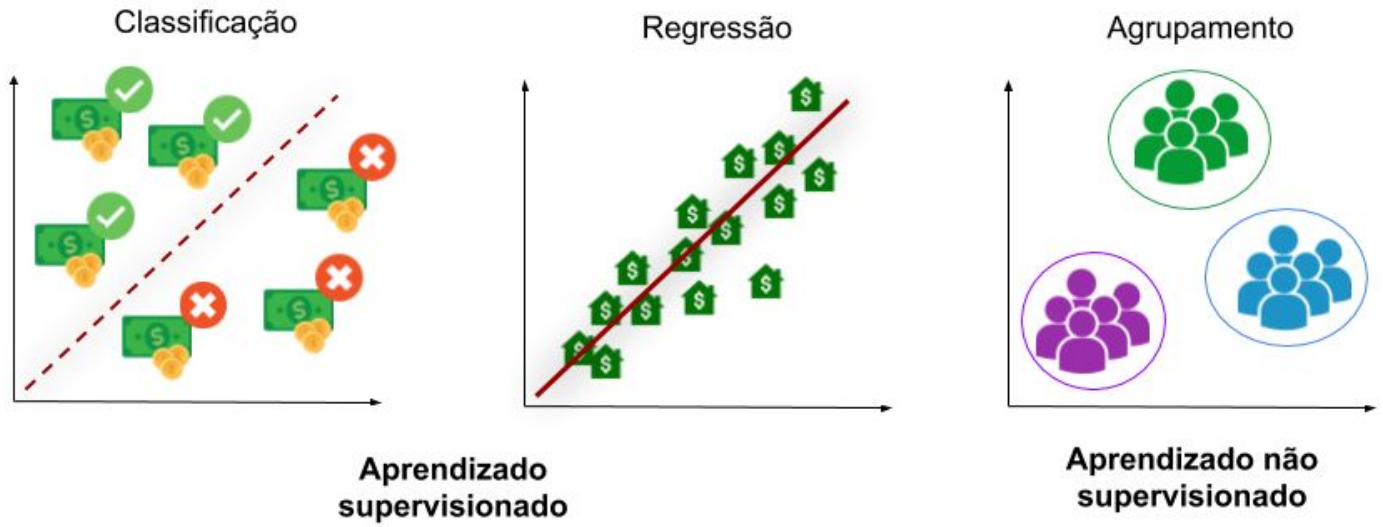
Aprendizado supervisionado e não supervisionado

Aprendizado supervisionado

- Regressão
- Classificação

Aprendizado supervisionado

- Clustering



Aprendizado supervisionado vs não supervisionado

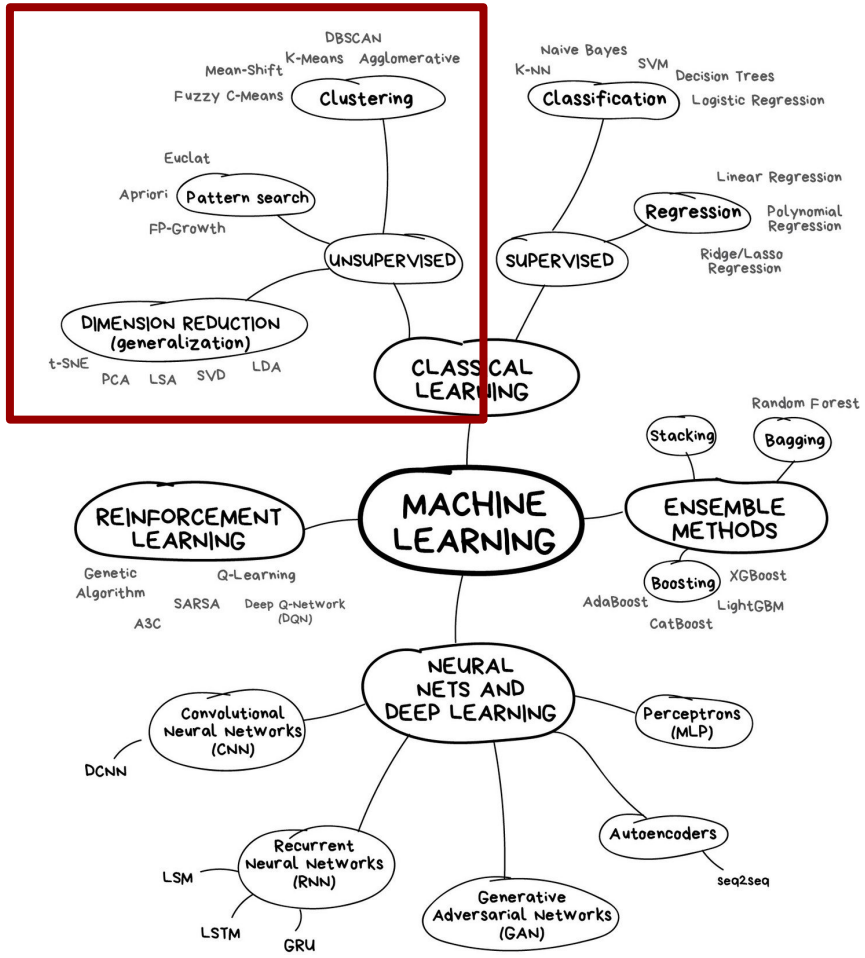


Diagrama retirado do [Machine Learning For Everyone]

Aprendizado supervisionado vs não supervisionado

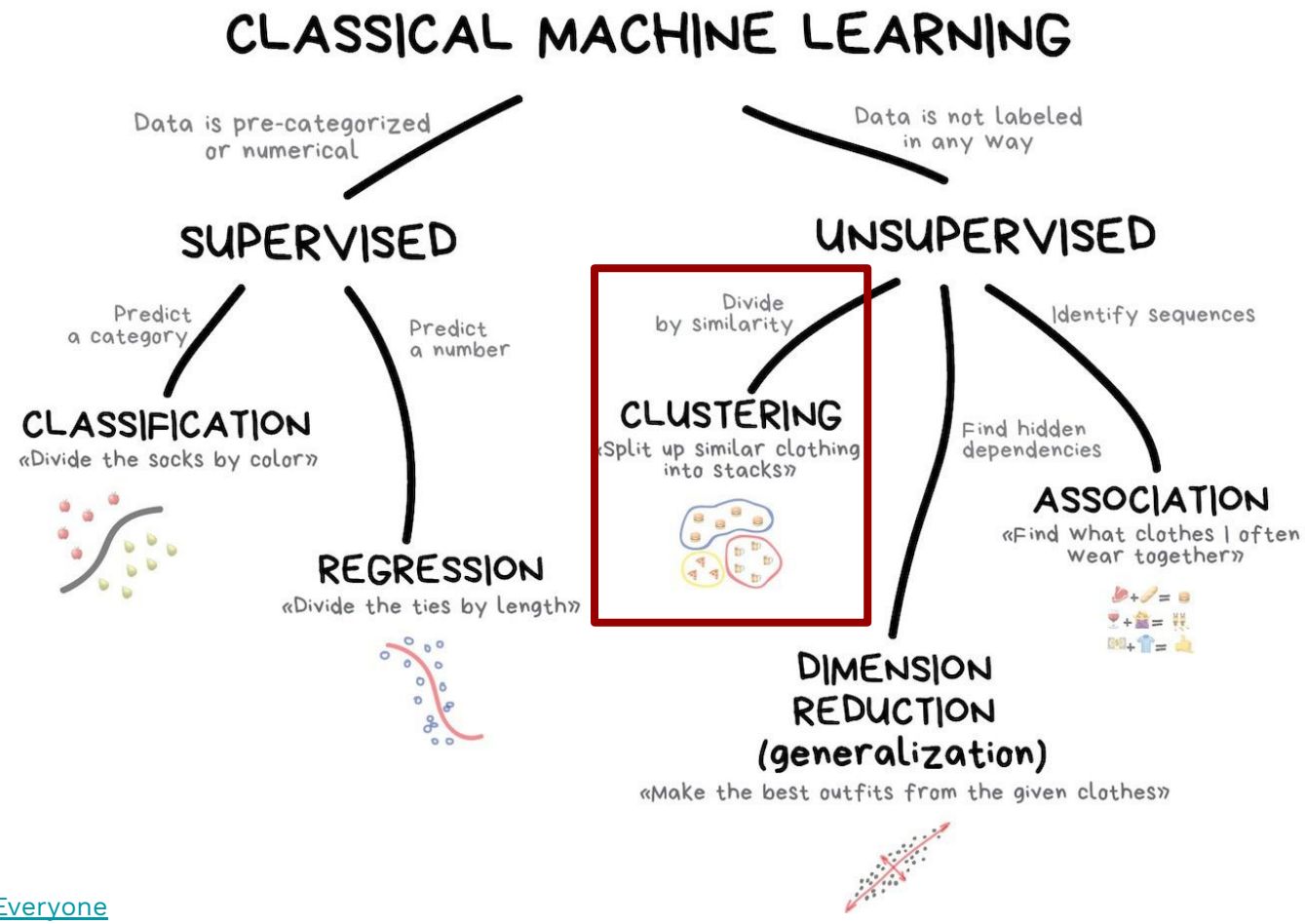


Diagrama retirado do [Machine Learning For Everyone]

O que é clustering?

(ou agrupamento/ clusterização)

Técnica para agrupar um conjunto de elementos tal que objetos agrupados em um mesmo grupo (também chamado de **cluster**) sejam *mais similares* entre eles do que com outros objetos pertencentes a outros clusters.

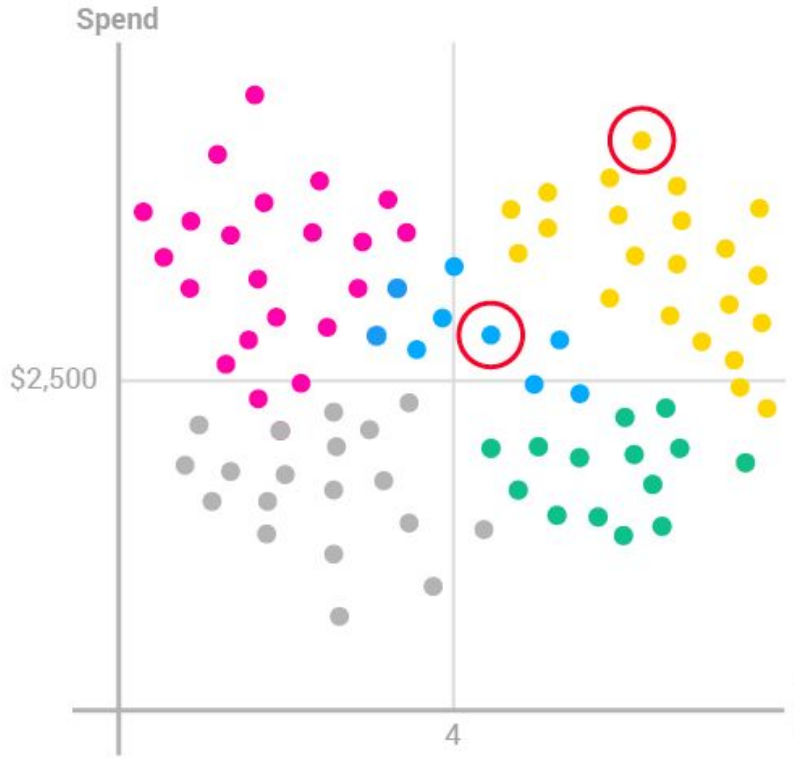
Sem clustering

O que é clustering?

(ou agrupamento/ clusterização)

Técnica para agrupar um conjunto de elementos tal que objetos agrupados em um mesmo grupo (também chamado de **cluster**) sejam *mais similares* entre eles do que com outros objetos pertencentes a outros clusters.

Com clustering



Como fica o nosso conjunto de dados para esses problemas?

Ao contrário do que acontece em problemas de classificação e regressão, o objetivo é que nosso modelo seja capaz de agrupar exemplos *similares* em um mesmo grupo **sem utilizar a classe** desses exemplos.

	Atributos (features)			Classe (target/label)
	Idade	**Renda**	**Possui dívidas**	**Cartão de crédito aprovado**
Exemplo	18	1000	Não	Não
	25	2500	Sim	Sim
	50	4500	Sim	Não
	42	10000	Não	Não
	33	6000	Não	Sim
	27	5700	Não	Não

Pra que fazer agrupamentos?

- Encontrar padrões
- Análises exploratórias
- Reduzir dimensionalidade

- Segmentação de clientes
- Detecção de anomalias e fraudes
- Segmentação de imagens
- Sistemas de recomendação
- Segmentação de documentos
- Análise de redes sociais

Quais são os tipos de clustering?

- Por partição
- Hierárquico
- Por densidade

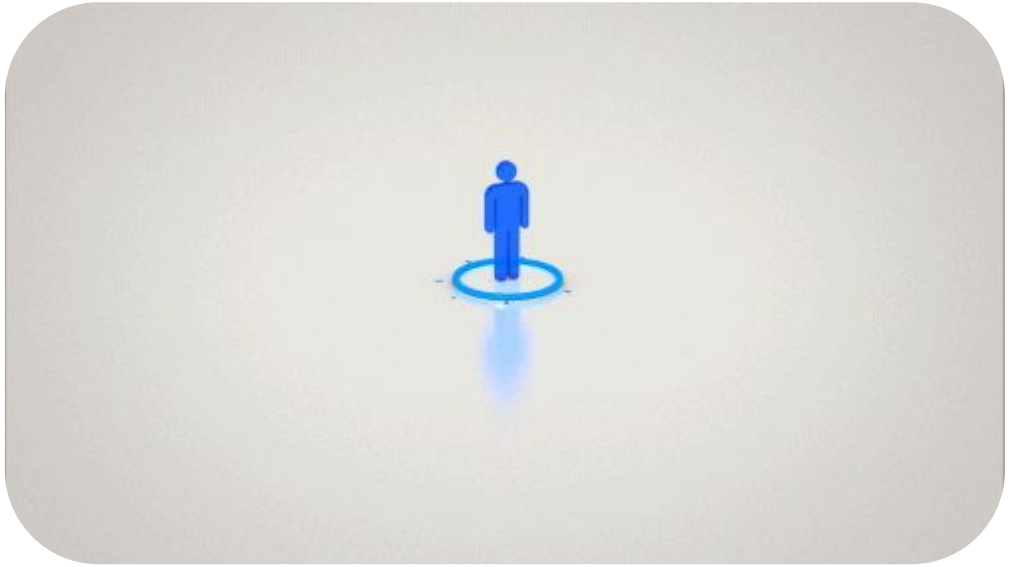
Agrupamento por partição

Vamos começar pelo de partição!

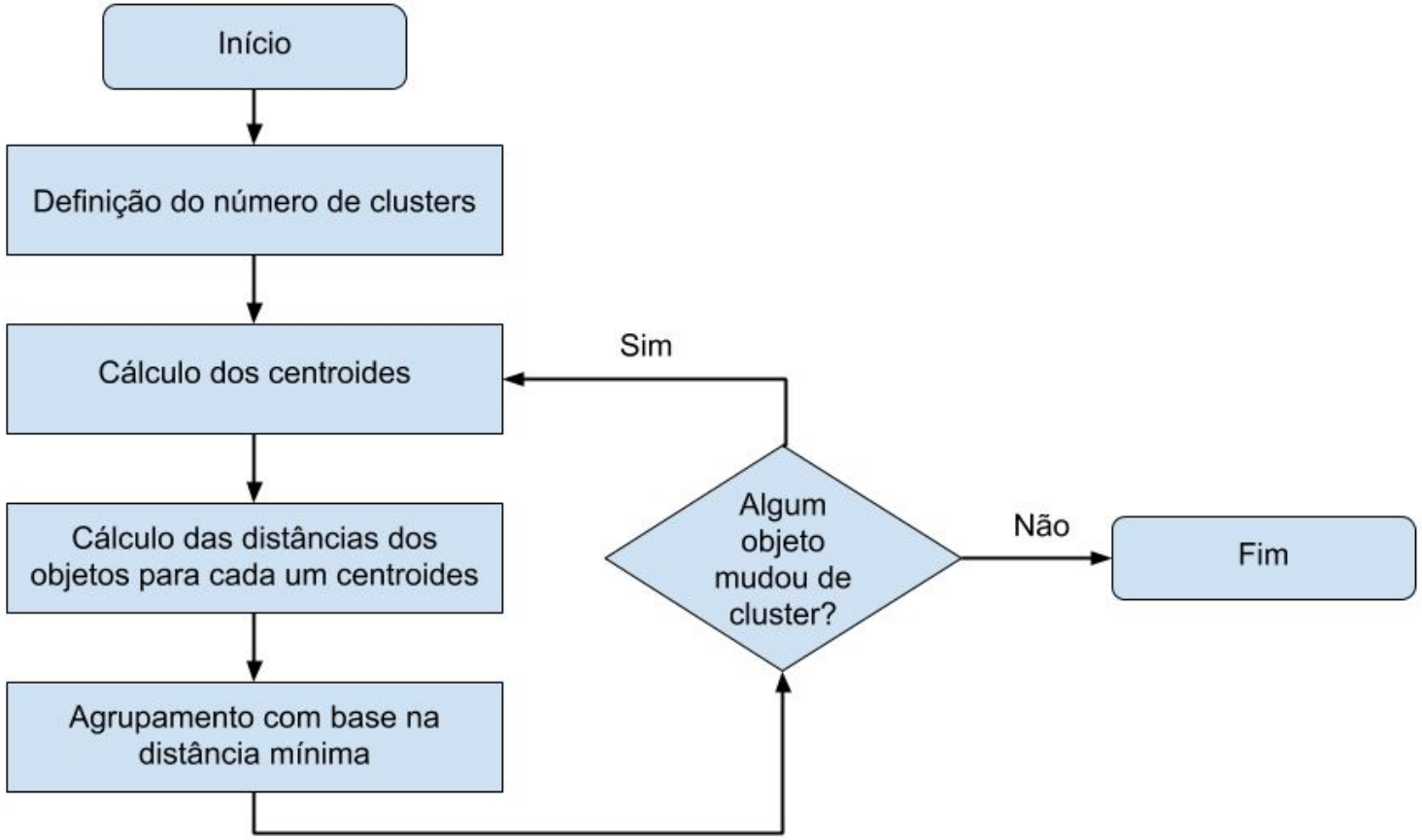
O algoritmo mais conhecido de partição é o **K-means**.

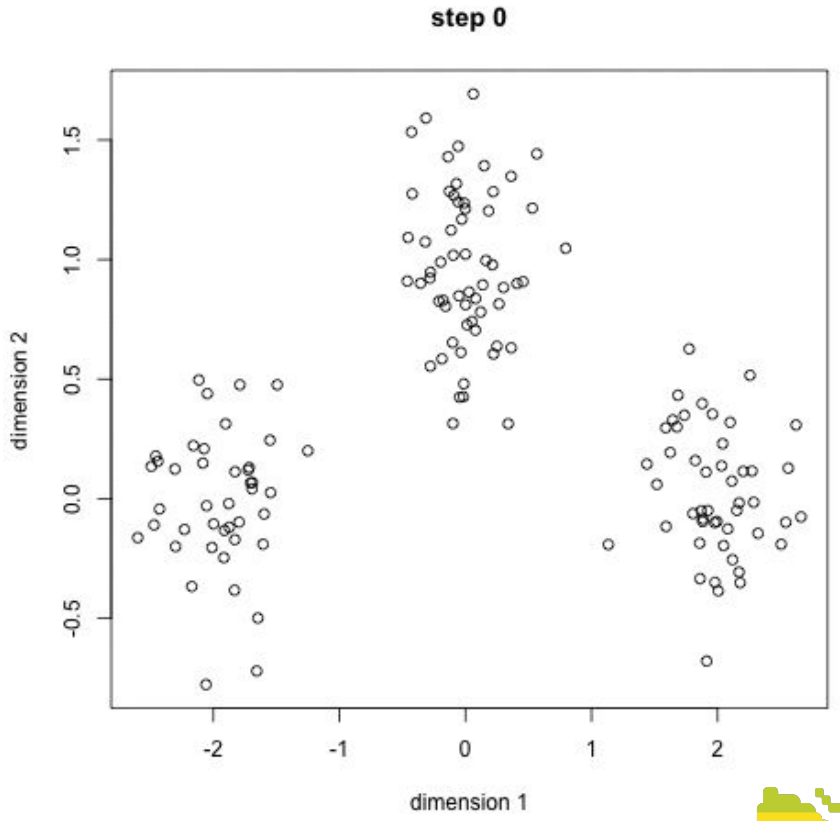
Ele tem como objetivo agrupar os objetos em *K* clusters.

Para isso, são elegidos representantes desses clusters, chamados de **centroides**.



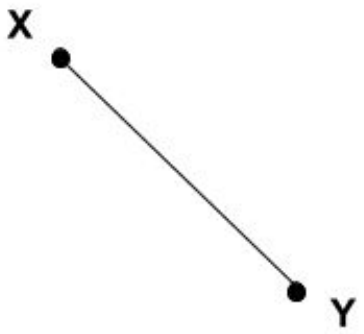
K-means


Início
Definição do número de clusters
Cálculo dos centroides
Cálculo das distâncias dos objetos para cada um centroides
Agrupamento com base na distância mínima
Algum objeto mudou de cluster?
Sim
Não
Fim


step 0
dimension 2
dimension 1
1.5
1.0
0.5
0.0
-0.5
-2
-1
0
1
2

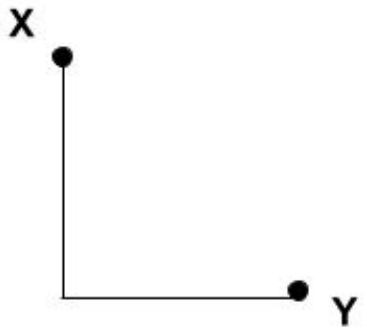
Distância? Como eu calculo essa distância?

Distância Euclidiana



$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Distância Manhattan



$$|x_1 - x_2| + |y_1 - y_2|$$

Quais são as vantagens do K-means?

- Simples de entender e de ser implementado
- Eficiente
- Escalável

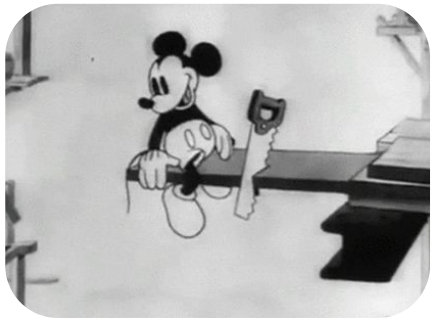
E quais são as desvantagens?

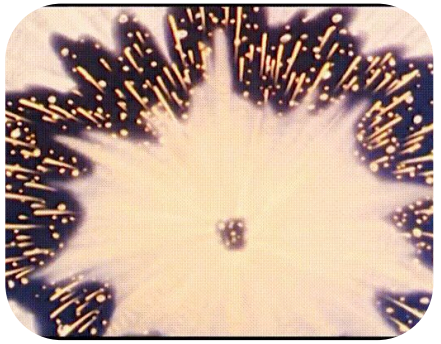
- Necessidade de escolher um número de clusters
- Pode sofrer com outliers
- Assume formatos esféricos para os clusters
- Funciona apenas para dados numéricos.

Vamos construir nosso primeiro modelo de clustering?

Seguiremos o seguinte fluxo (é o mesmo que seguimos na classificação):

Análise exploratória → **Feature Engineering** → **Modelagem** → **Avaliação**





Pré-processamento dos dados

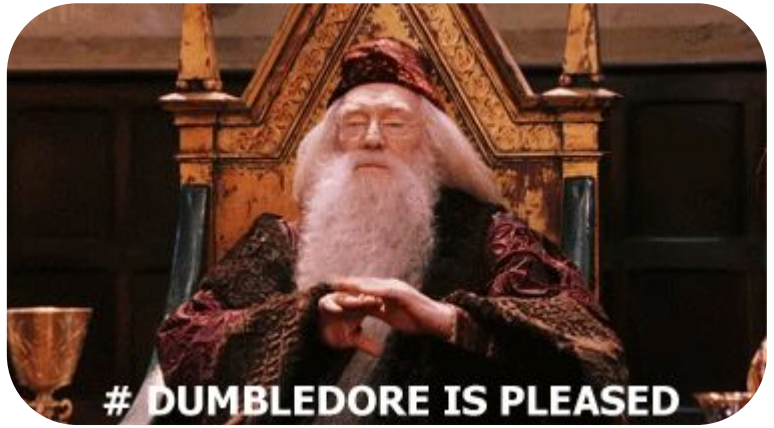

Vamos praticar!

Link para o notebook no Colab:

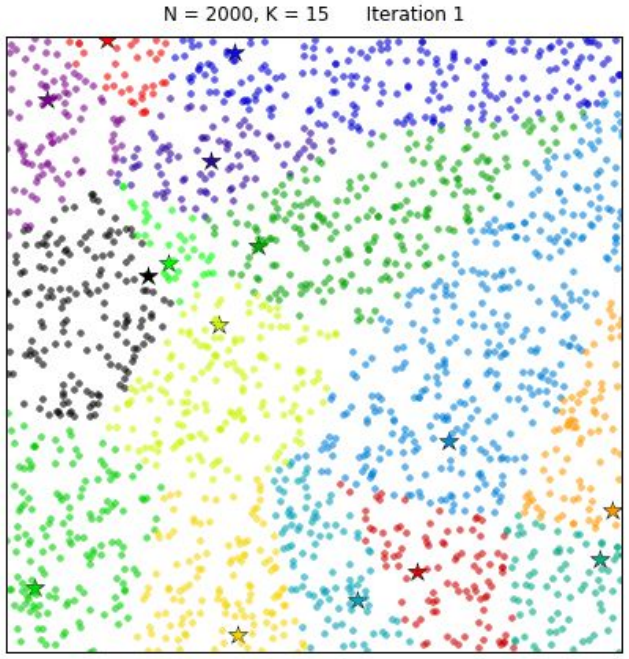
https://bit.ly/2CJMieq

Façam uma cópia para vocês conseguirem editar :D

UHUL! Fizemos nosso primeiro modelo de clustering



Foi fácil definir o número de clusters nesse caso?



...Mas quantos escolheríamos para o conjunto de dados ao lado?

Avaliação de clusters

Não podemos utilizar as métricas de avaliação da classificação...

Como os dados não são rotulados, não podemos usar uma matriz de confusão, por exemplo.

Para problemas de agrupamento, existem diversas métricas possíveis para avaliar o quão bons foram os agrupamentos encontrados. Hoje falaremos sobre uma delas: o **Elbow method**.

Elbow method ("método do cotovelo")

Fornece uma ideia de qual seria um bom número de clusters baseando-se na **inércia** entre os objetos e os centroides dos seus respectivos clusters.

Mas o que é essa ***inércia***?

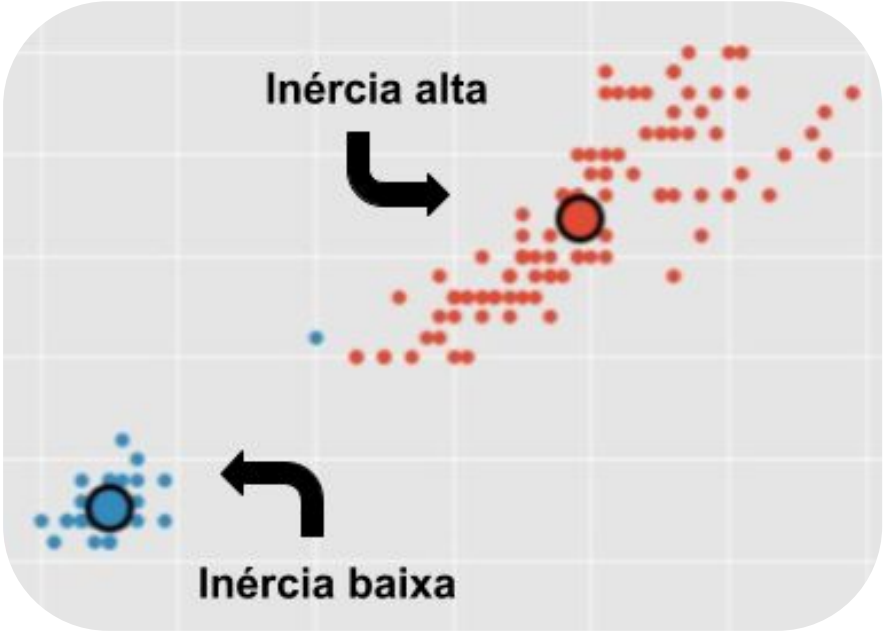
$$Inercia(k) = \sum_{j=1}^{k} \sum_{x_i \in cluster\ j} \|x_i - \bar{x}_j\|^2, \quad \text{onde } \bar{x}_j \text{ é o centroide do cluster j.}$$

Elbow method ("método do cotovelo")

*Mas o que é essa **inércia**?*

A inércia é uma medida calculada que se baseia na soma das distâncias quadráticas de cada objeto para os centroides de seus respectivos clusters.

Quanto maior for a inércia, maior será a dispersão dos clusters; quanto menor, mais os clusters estarão compactados.



Outras métricas de avaliação

Índices internos

Compara a estrutura de clusters descobertos com uma estrutura de grupos previamente conhecida.

- índice de Rand
- índice de Jaccard
- índice de Folkes e Mallows

Índices externos

Analisa a estrutura de clusters descobertos com relação a algum critério, como compacidade e separabilidade.

- índice Dunn
- índice Davies-Bouldin
- índice Silhouette

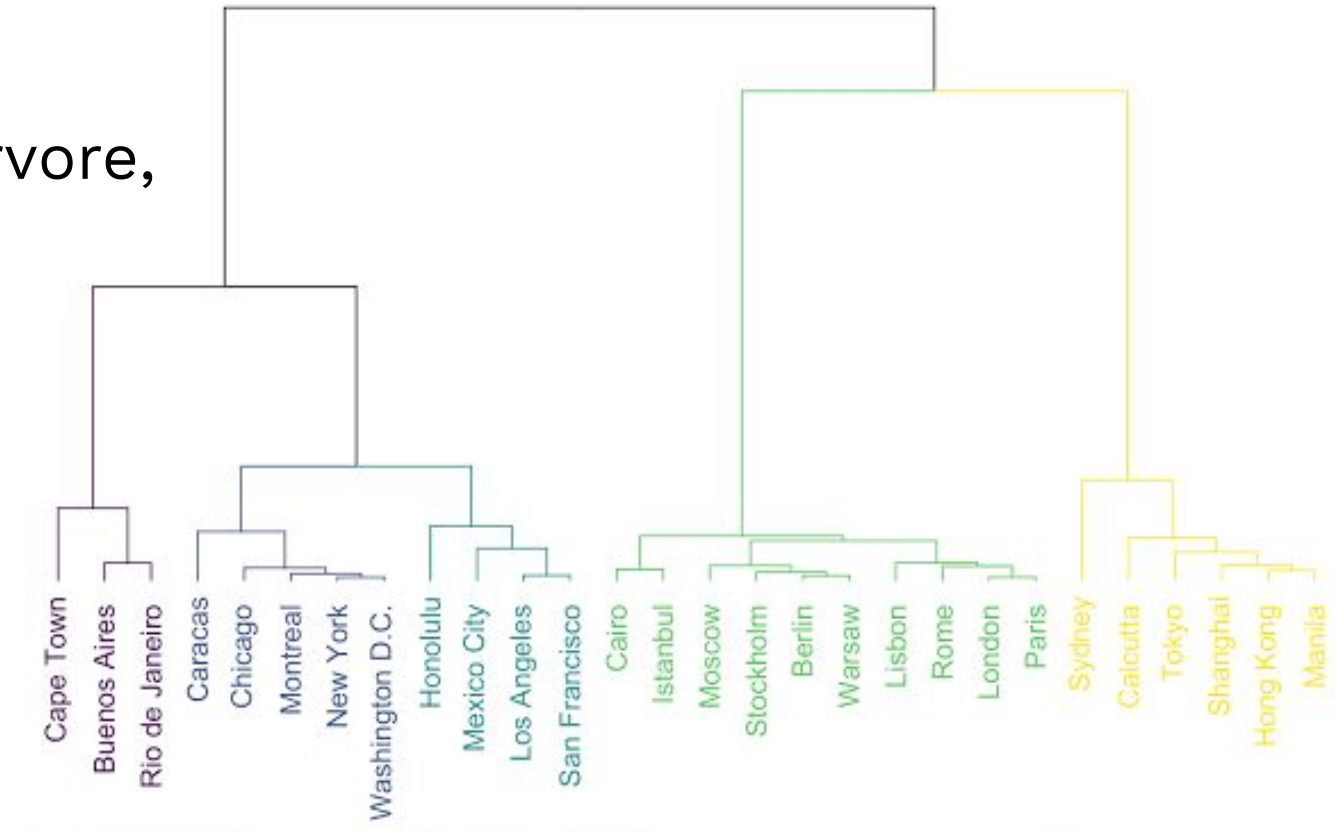
Apesar desses métodos fornecerem indícios do número de clusters ideal, também é importante ter um bom conhecimento sobre o domínio (ou envolver pessoas que o tenham no projeto!).

Aqui vocês conseguem encontrar essas métricas e mais outras implementadas no sklearn.


Vamos praticar!

Agrupamento hierárquico

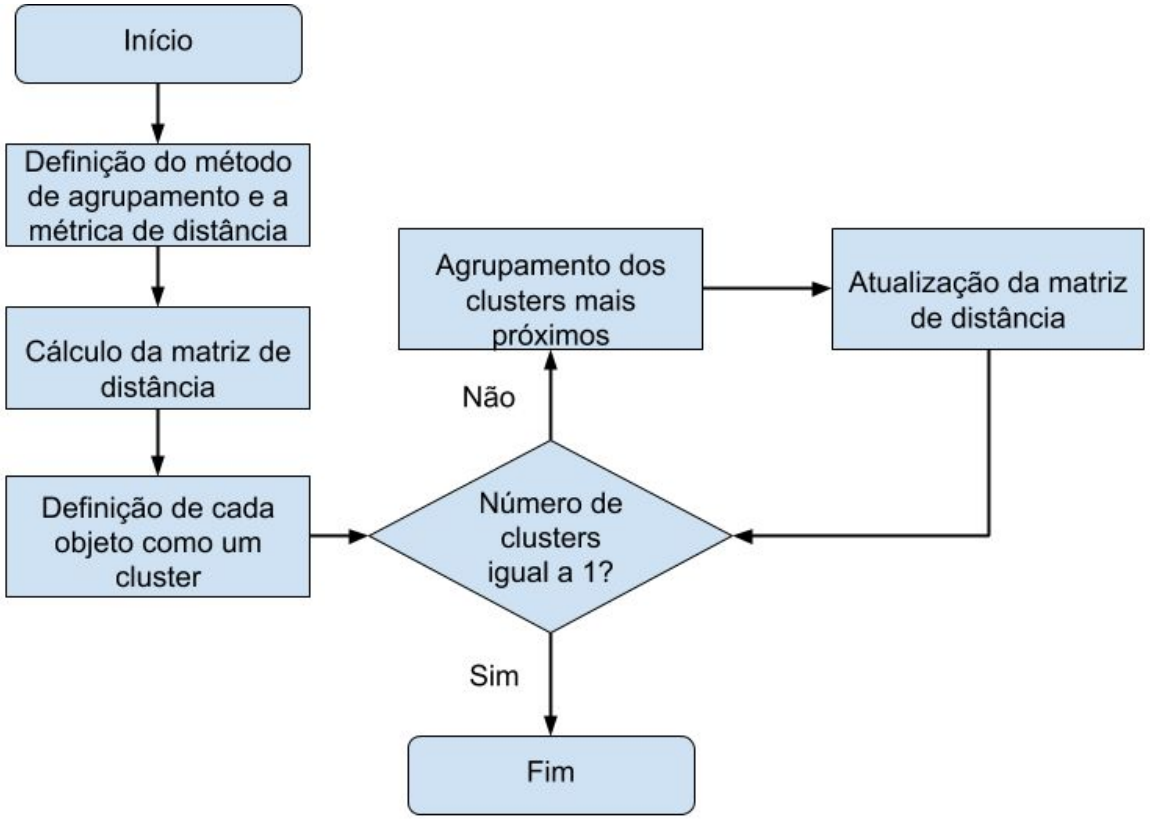
No agrupamento hierárquico, os clusters são representados hierarquicamente por meio de diagrama representando uma árvore, chamado de **dendrograma**.

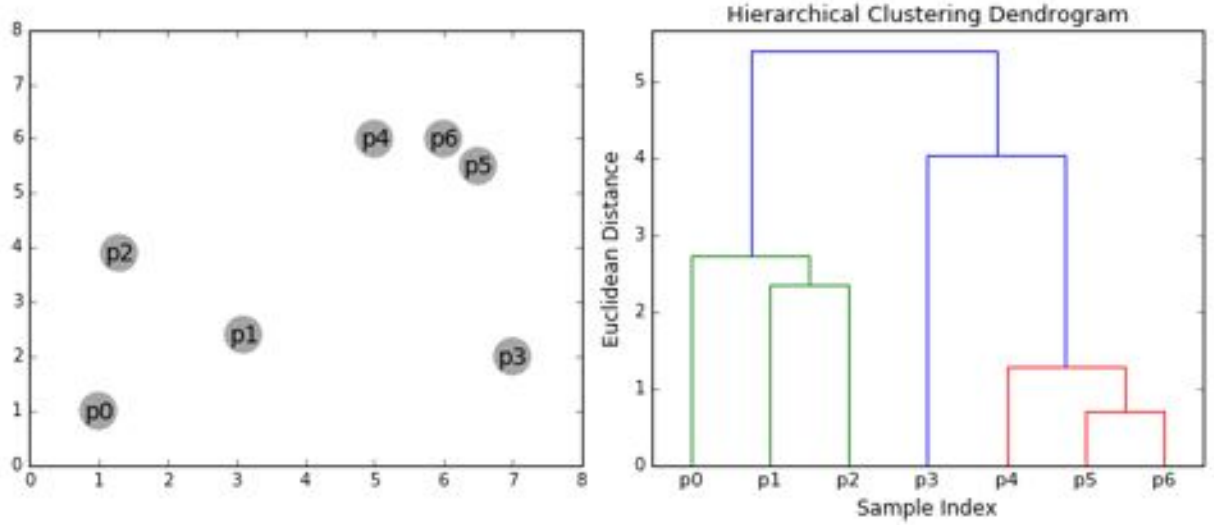


Tipos

- Aglomerativo
- Divisivo

Agrupamento hierárquico aglomerativo



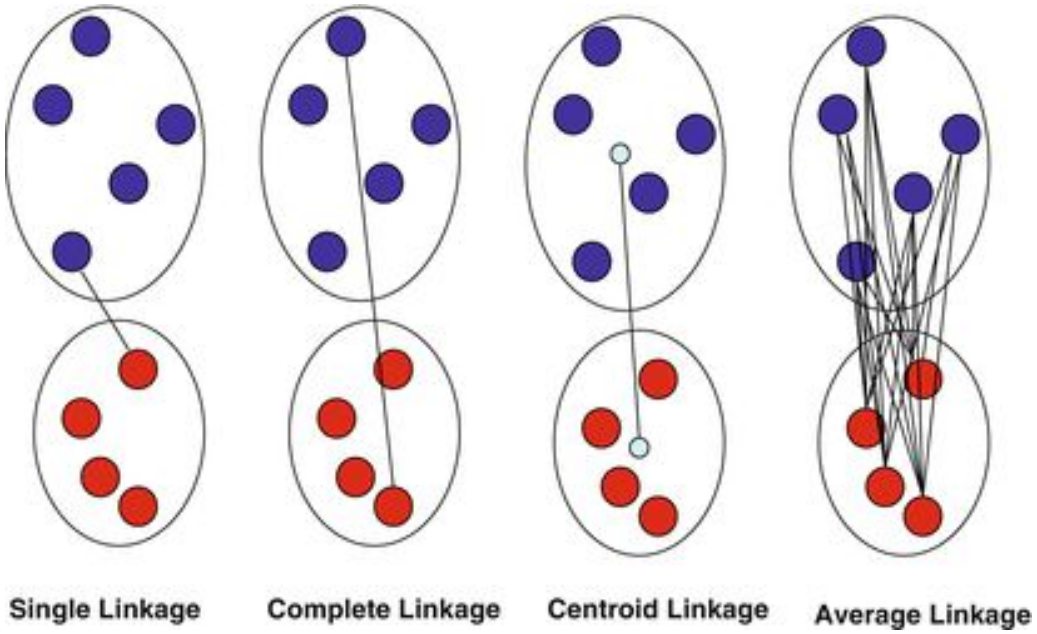


Quais critérios para fazer os agrupamentos?

Uma medida de distância

- Distância euclidiana
- Distância Manhattan
- Outras métricas aceitas

Um método de agrupamento



- Outras abordagens aceitas

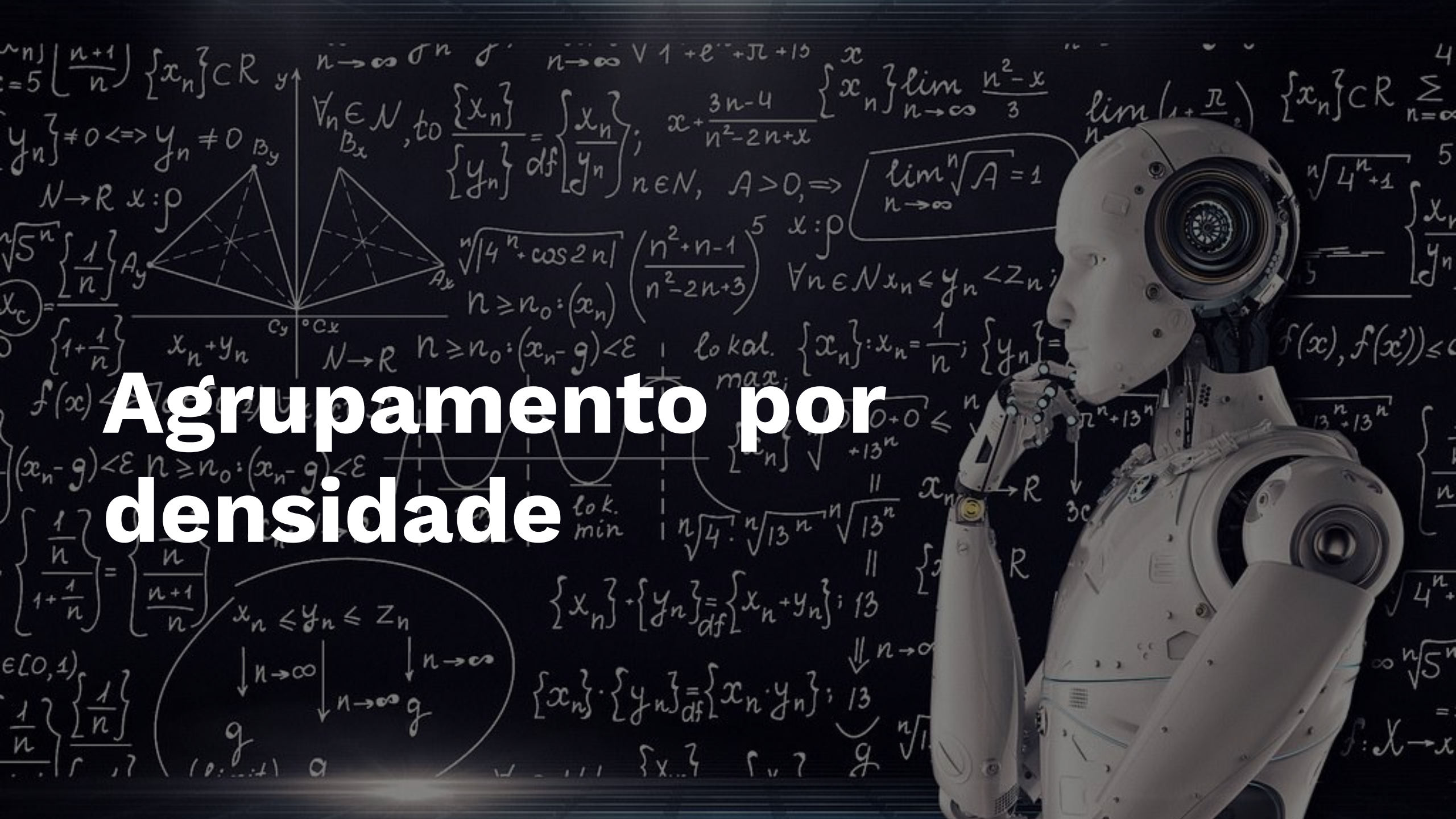
Quais são as vantagens do agrupamento hierárquico?

- Não precisamos escolher um número inicial de clusters
- Dendrogramas são ótimos para visualização dos clusters
- Gera uma hierarquia entre os clusters

E quais são as desvantagens?

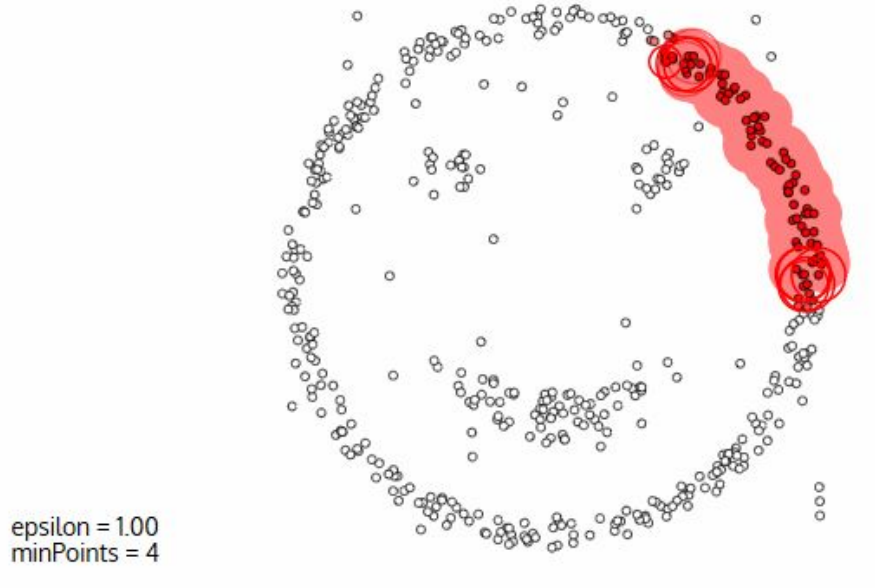
- Complexidade
- Dependendo dos dados, pode ser difícil escolher o número de clusters
- Um agrupamento feito erroneamente não pode ser desfeito

Agrupamento por densidade



Dos agrupamentos por densidade, o mais conhecido é o **DBSCAN**.

Ele é um método de clustering por densidade que busca por clusters definidos como regiões com alta densidade de objetos, separados por regiões de baixa densidade.

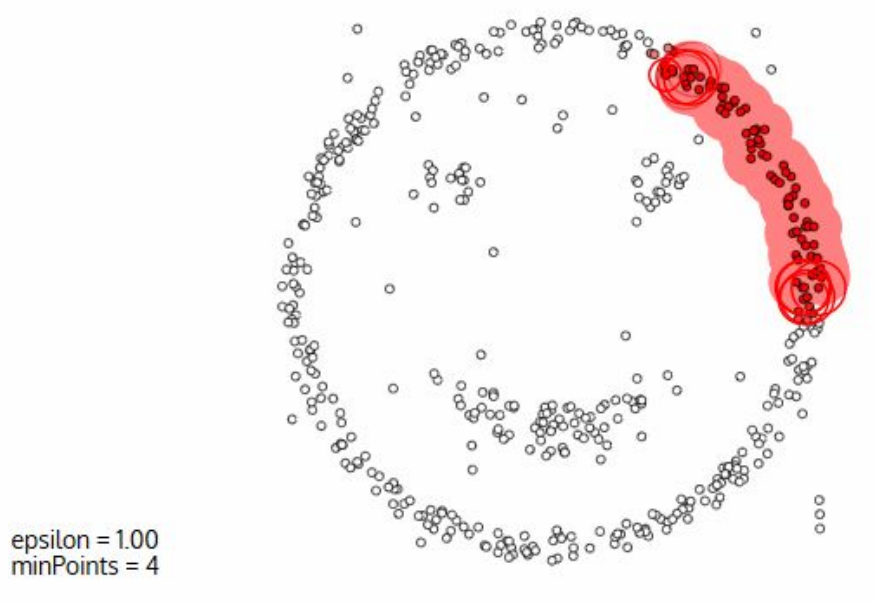


Ele necessida dos seguintes parâmetros:

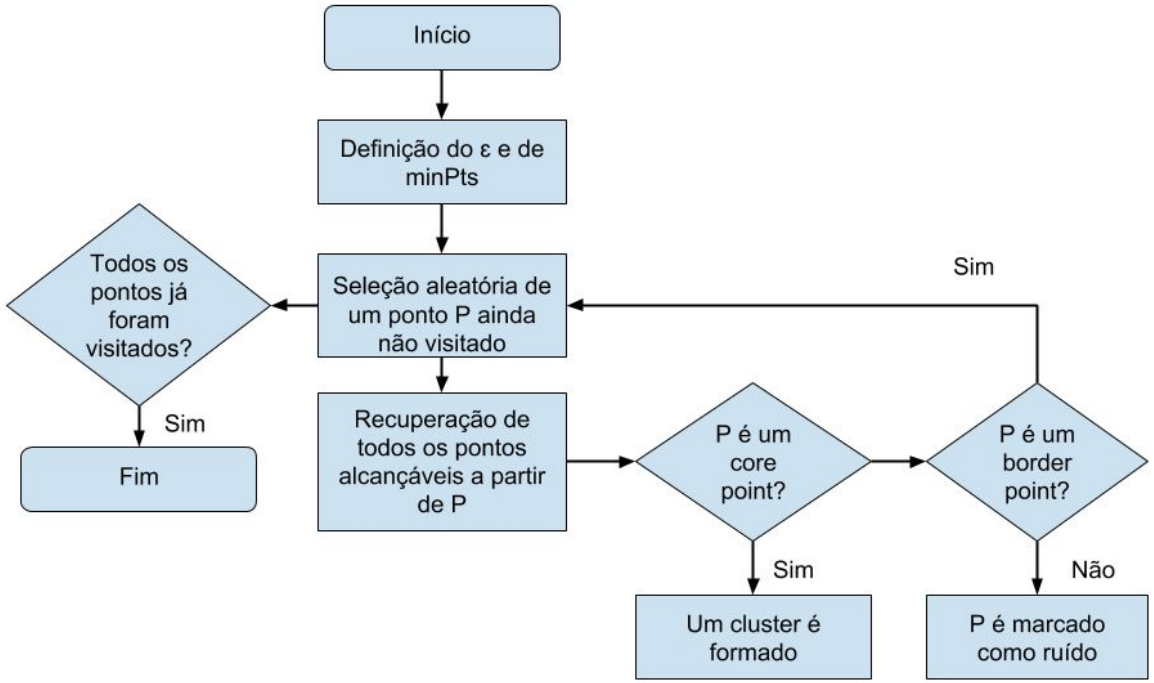
- **ε**: raio da vizinhança ao redor do ponto P
- **minPts**: número mínimo de pontos na vizinhança para que seja definido um cluster

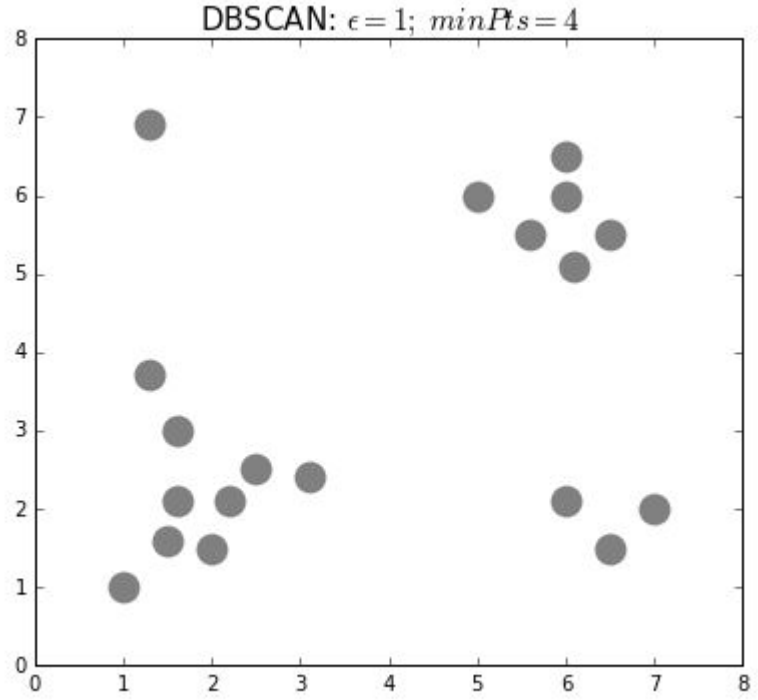
Com base nesses dois parâmetros, o DBSCAN categoria os pontos em três categorias:

- **Core Points**: um ponto P é um core point se sua vizinhança contém ao menos minPts
- **Border Points**: um ponto Q é um border point se sua vizinhança contem menos pontos que minPts, mas se Q é alcancável por algum core point P.
- **Outlier**: um ponto O é um outlier se não for nem um core point e nem um border point



DBSCAN


Início
Definição do ε e de minPts
Seleção aleatória de um ponto P ainda não visitado
Todos os pontos já foram visitados?
Sim
Fim
Recuperação de todos os pontos alcançáveis a partir de P
P é um core point?
Sim
Um cluster é formado
P é um border point?
Não
P é marcado como ruído
Sim


DBSCAN: $\epsilon = 1;\ minPts = 4$
0
1
2
3
4
5
6
7
8

Quais são as vantagens do DBSCAN?

- Não é necessário especificar um número inicial de clusters
- Lida bem com outliers
- Consegue encontrar clusters com formatos diferentes

E quais são as desvantagens?

- A escolha dos dois parâmetros iniciais pode não ser muito intuitiva
- Tem dificuldade para encontrar clusters se a densidade dos dados variar muito
- Exige mais processamento

Outros métodos de clustering

Por partição

- K-medians
- K-modes
- K-prototypes

Por densidade/hierárquico

- HDBSCAN

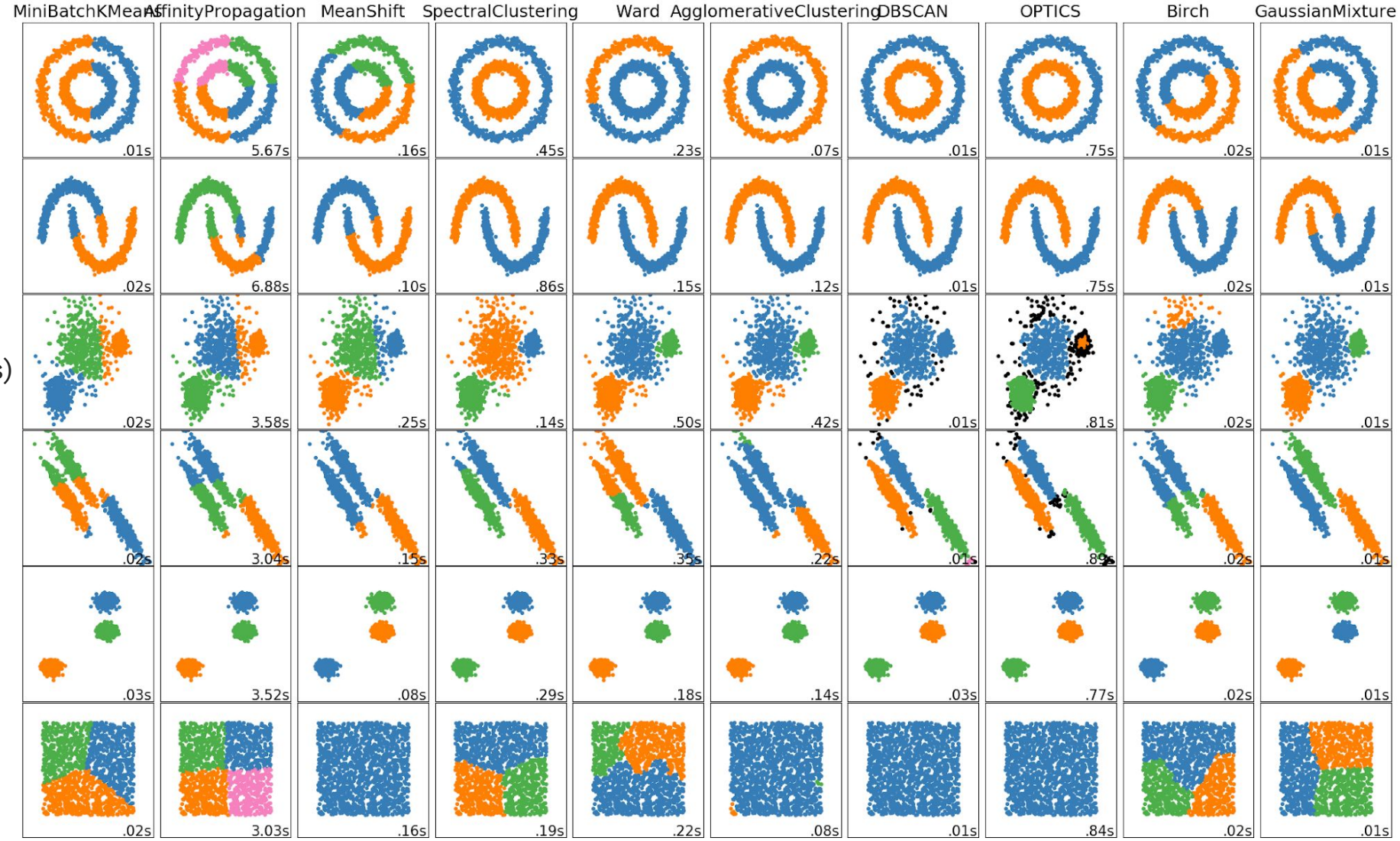
Por distribuição

- Gaussian Mixture Models (GMMs)

Redes neurais

- Self Organizing Map (SOM)

O sklearn conta com mais alguns algoritmos de clustering e também tem uma comparação entre eles para vários conjuntos de dados:



Para praticar...

- O repositório do UCI contém alguns datasets para realizar clustering
- Também pode-se retirar a classe de datasets existentes para problemas supervisionados e aplicar técnicas de aprendizado não supervisionado!

Por hoje, é isso!

No próximo sábado iremos falar sobre Deep Learning e NLP \o/

Vamos preencher o formulário de feedback???

http://bit.ly/bootcamp-ds-sp-feedback-19

Obrigada!

Dúvidas?
Podem nos procurar! :D