

INTELIGÊNCIA ARTIFICIAL

Aprendizado Descritivo
Clustering

Marcos de Souza
mso2@cesar.school



Marcos de Souza

Bacharel em Ciência da
Computação

Mestre em Ciência da
Computação (UFPE)

Doutorando em Ciência da
computação (UFPE)

DS & SE @ CESAR - 2017

Professor @ CESAR School -
2022



Linkedin: <https://www.linkedin.com/in/marcos-de-souza-msc-893758aa/>

Github: <https://github.com/marcosd3souza>

Lattes: <http://lattes.cnpq.br/6137784444858483>

João Canhoto

Técnico em Eletrônica (IFPE)

Bacharel em Engenharia da
Computação (UFPE)

DS & SE @ CESAR - 2017

Professor @ CESAR School -
2022



LinkedIn: <https://www.linkedin.com/in/jo%C3%A3o-lucas-canhoto-836670143/>

Antônio Júnior

Bacharel em Ciência da
Computação (UFAM)

Mestre em Ciência da
Computação (UFAM)

DS & SE @ CESAR - 2017

Professor @ CESAR School -
2022



Linkedin: <https://www.linkedin.com/in/antoniojsjunior/>

Lattes: <http://lattes.cnpq.br/6334165648715375>

Gabriel Calazans

Graduação em Engenharia da
Computação (UPE)

Mestrando em Engenharia da
Computação (UPE)

SE @ CESAR - 2020

Professor @ CESAR School -
2022



Linkedin: <https://www.linkedin.com/in/gabriel-c-159038141/>

Github: <https://github.com/gcalazansdm>

Lattes: <http://lattes.cnpq.br/7670369643263800>

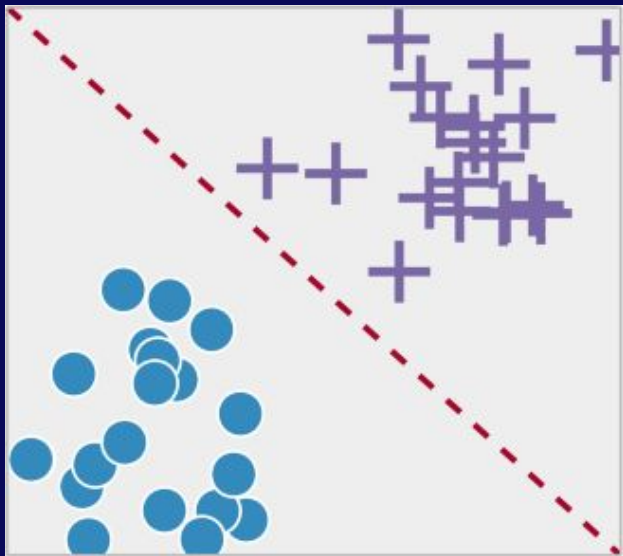


Aprendizado Supervisionado vs Não Supervisionado

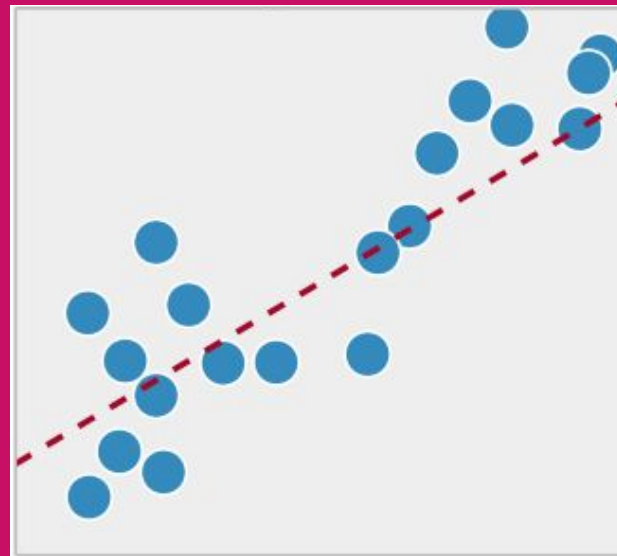
Aprendizado Supervisionado / Não Supervisionado

ref: <https://medium.com/opensanca/aprendizagem-de-maquina-supervisionada-ou-n%C3%A3o-supervisionada-7d01f78c>

Aprendizado Supervisionado (Classificação)



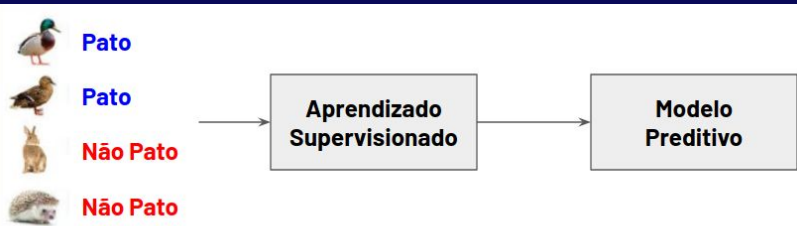
Aprendizado não supervisionado (Regressão)



Aprendizado Supervisionado / Não Supervisionado

ref: <https://www.venturus.org.br/machine-learning-para-leigos/>

Aprendizado Supervisionado (Classificação)



Aprendizado não supervisionado (Agrupamento)





Aprendizado Não-Supervisionado

- Exemplos fornecidos ao algoritmo não são acompanhados da **resposta esperada**
- Normalmente **agrupam** os exemplos semelhantes
- Extraíndo as suas **principais características (define um representante)**

Agrupamentos / Clustering

ref: https://www.cgee.org.br/documents/10195/734063/CGEE_Pan_Cie_Bra_2015-20.pdf

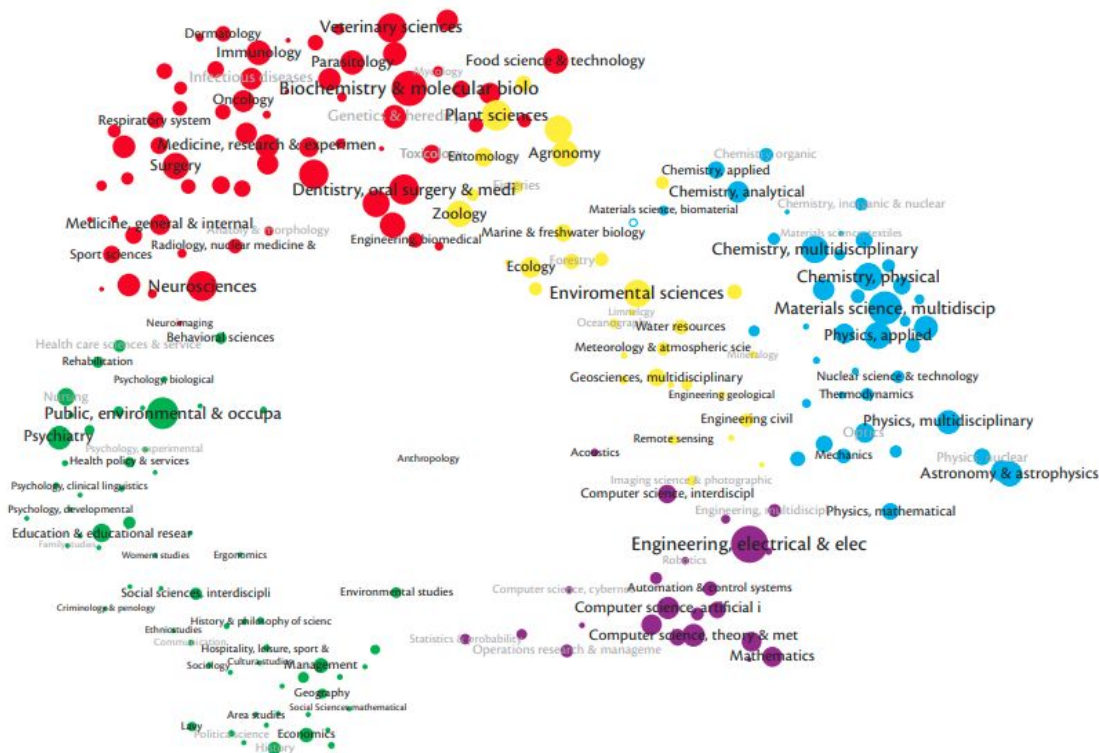


Figura 2 – Mapa da ciência da produção científica brasileira, *Web of Science* (2015-2020)

Fonte: *Web of Science*, dados extraídos em fevereiro de 2021.

Agrupamentos / Clustering

ref: https://www.cgee.org.br/documents/10195/734063/CGEE_Pan_Cie_Bra_2015-20.pdf

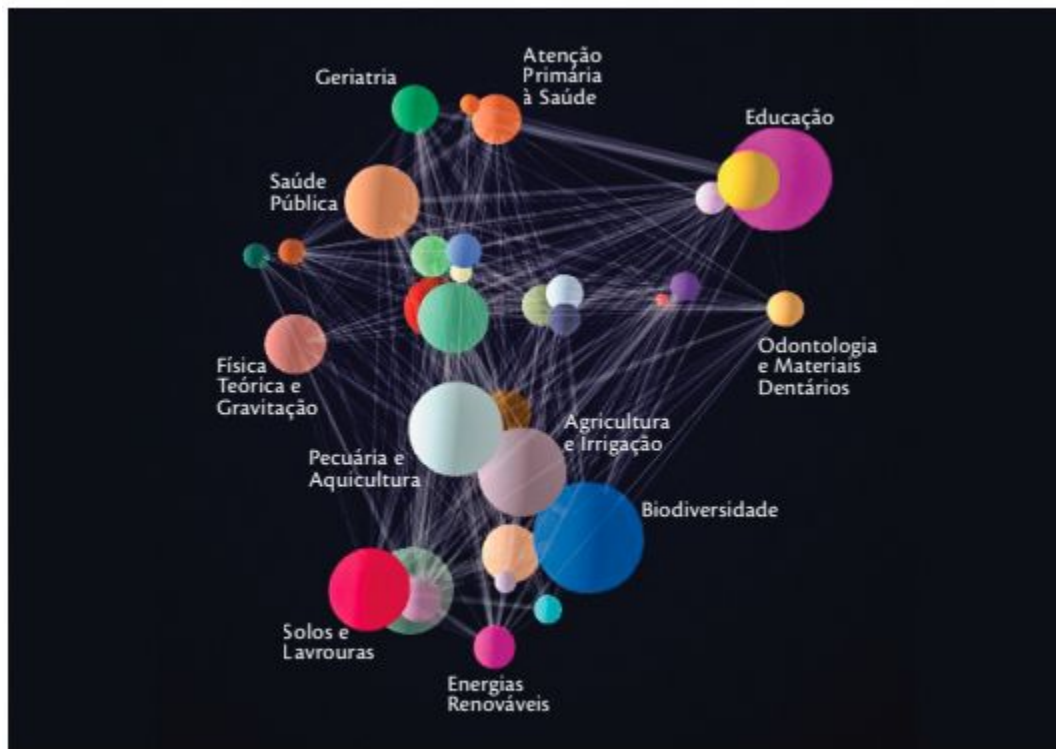


Figura 8 – Rede dos 35 agrupamentos temáticos formada a partir da rede de similaridade semântica da produção científica com participação brasileira

Fonte: Web of Science, dados extraídos em maio de 2020.

Agrupamentos / Clustering

Como podemos agrupar estas amostras de forma que faça sentido?

É razoável assumir - em problemas diversos - que amostras podem ser agrupadas 'naturalmente' por meio da similaridade de suas características?

Ornitorrinco



Pato



Baleia



Leão

Agrupamentos / Clustering

Como podemos agrupar estas amostras de forma que faça sentido?

É razoável assumir - em problemas diversos - que amostras podem ser agrupadas 'naturalmente' por meio da similaridade de suas características?

Com Bico

Ornitorrinco



Pato



Baleia



Leão

Sem Bico

Agrupamentos / Clustering

Como podemos agrupar estas amostras de forma que faça sentido?

É razoável assumir - em problemas diversos - que amostras podem ser agrupadas 'naturalmente' por meio da similaridade de suas características?

Ornitorrinco



Baleia

Água

Pato



Leão

Terra

Agrupamentos / Clustering

Como podemos agrupar estas amostras de forma que faça sentido?

É razoável assumir - em problemas diversos - que amostras podem ser agrupadas 'naturalmente' por meio da similaridade de suas características?

Ornitorrinco



Pato



Ovíparo



Baleia

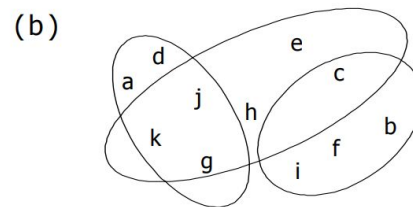
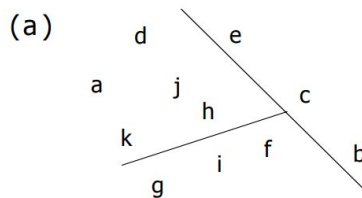
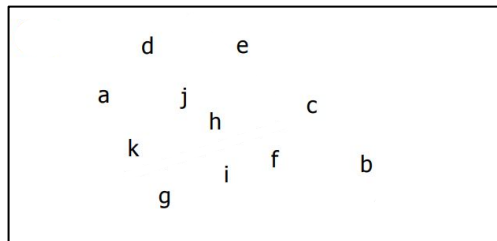


Leão

Mamífero

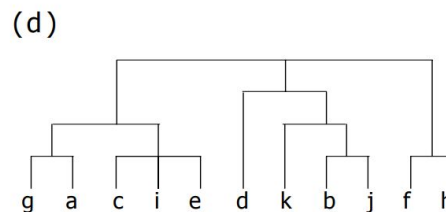
Agrupamentos / Clustering

Outras formas de agrupamento...



(c)

	1	2	3
a	0,4	0,1	0,5
b	0,1	0,8	0,1
c	0,3	0,3	0,4
d	0,1	0,1	0,8
e	0,4	0,2	0,4
f	0,1	0,4	0,5
g	0,7	0,2	0,1
h	0,5	0,4	0,1
...			



Aprendizado Não-Supervisionado (Descritivo)

O resultado normalmente são agrupamentos (clusters) dos exemplos da base de treinamento que podem ser utilizados para prever o grupo de um novo indivíduo.

Representação dos
Objetos

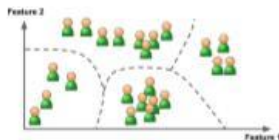
Clusters/
Agrupamentos

**Escolha/Definição
dos Objetos e seus
Atributos**

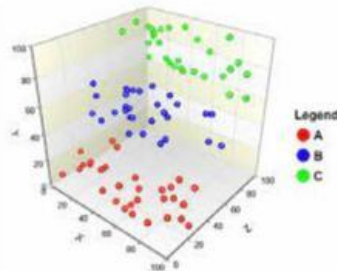


**Exemplos de
features:**
idade x escolaridade
latitude x longitude
altura x peso x idade

**Similaridade entre
Objetos**

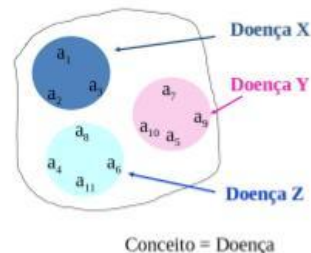


**Algoritmo de
Clustering**



Validação

Interpretação



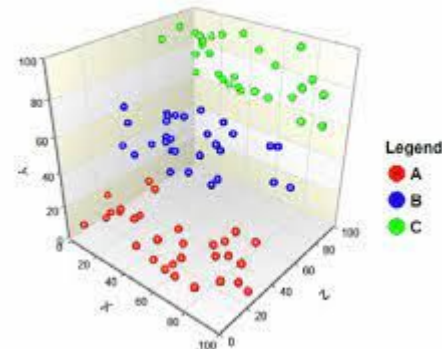
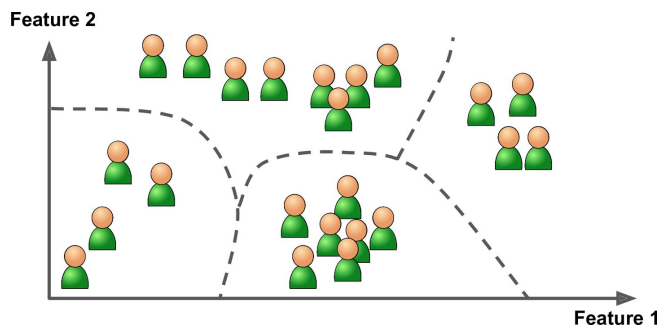
Obs: a normalização não altera o arranjo natural.

Aprendizado Não-Supervisionado (Descritivo)

O resultado normalmente são agrupamentos (**clusters**) dos exemplos da base de treinamento que podem ser utilizados para prever o grupo de um novo indivíduo.

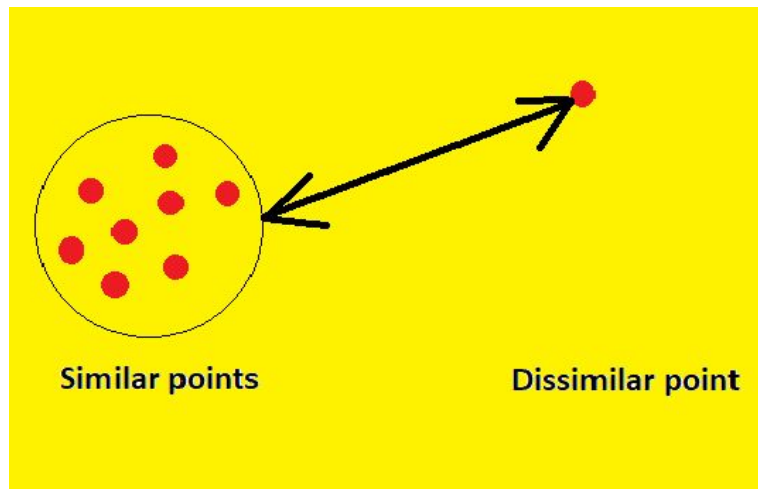
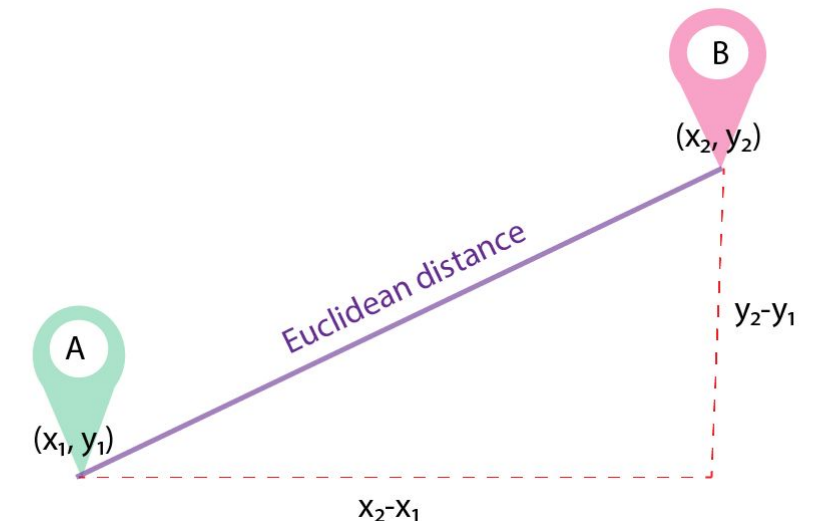
Exemplos de features:

idade x escolaridade
latitude x longitude
altura x peso x idade



Aprendizado Não-Supervisionado (Descritivo)

Uma medida de distância (como a euclidiana) é utilizada para definir o quão semelhantes os dados são

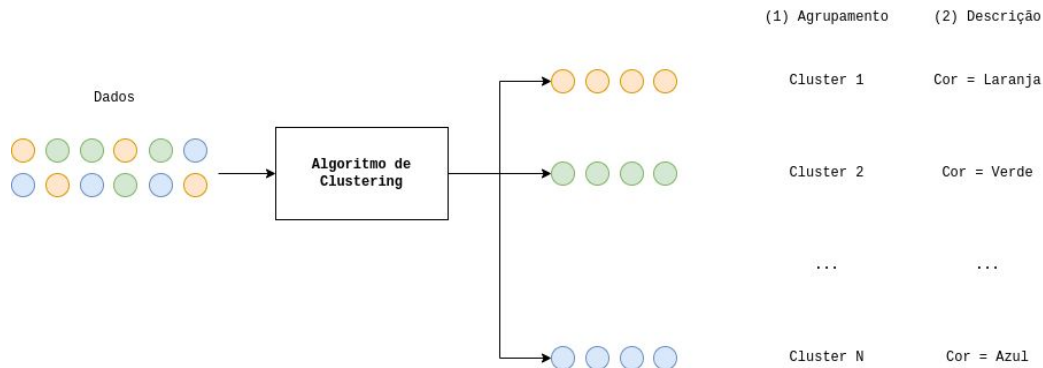


Aprendizado Não-Supervisionado (Descritivo)

Definição do problema de Agrupamentos / Clustering

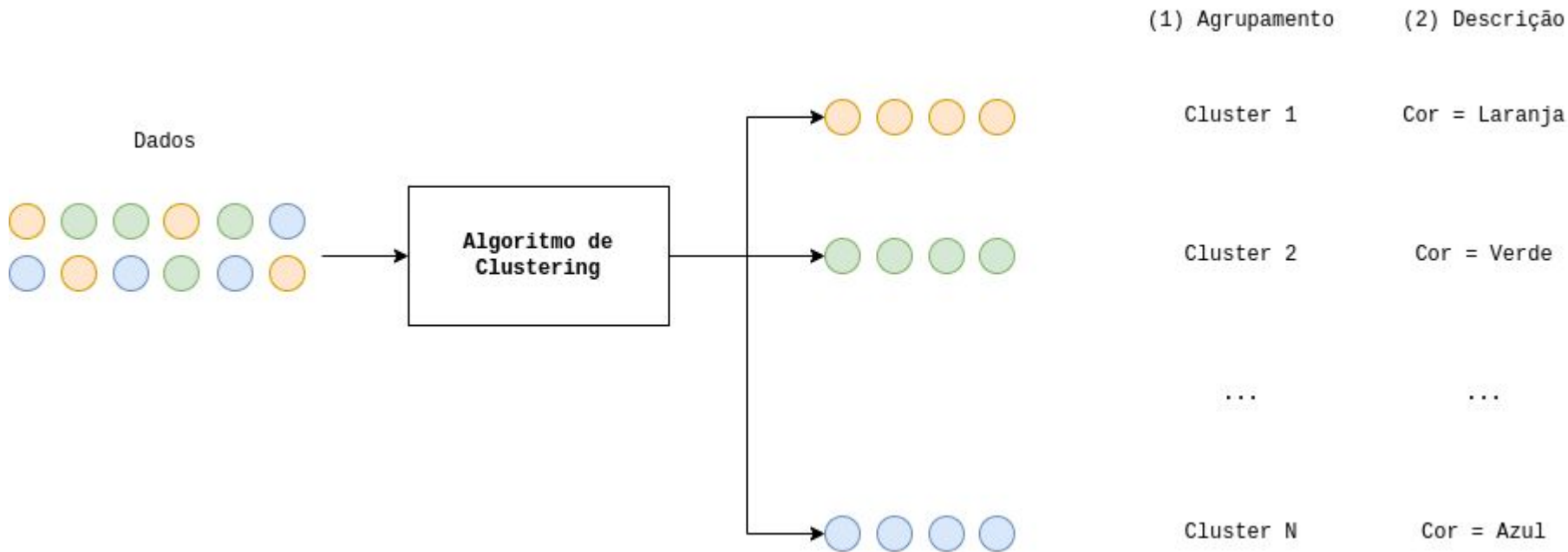
Considerando um conjunto de objetos descritos através de múltiplos valores (atributos), o algoritmo de agrupamento/clustering deve:

- 1) Atribuir grupos a objetos
 - a) Maximizar a similaridade entre objetos de um mesmo grupo/cluster
 - b) Minimizar a similaridade entre objetos de grupos distintos
- 2) Atribuir descrição para grupos para os agrupamentos descobertos



Aprendizado Não-Supervisionado (Descritivo)

Definição do problema de Agrupamentos / Clustering

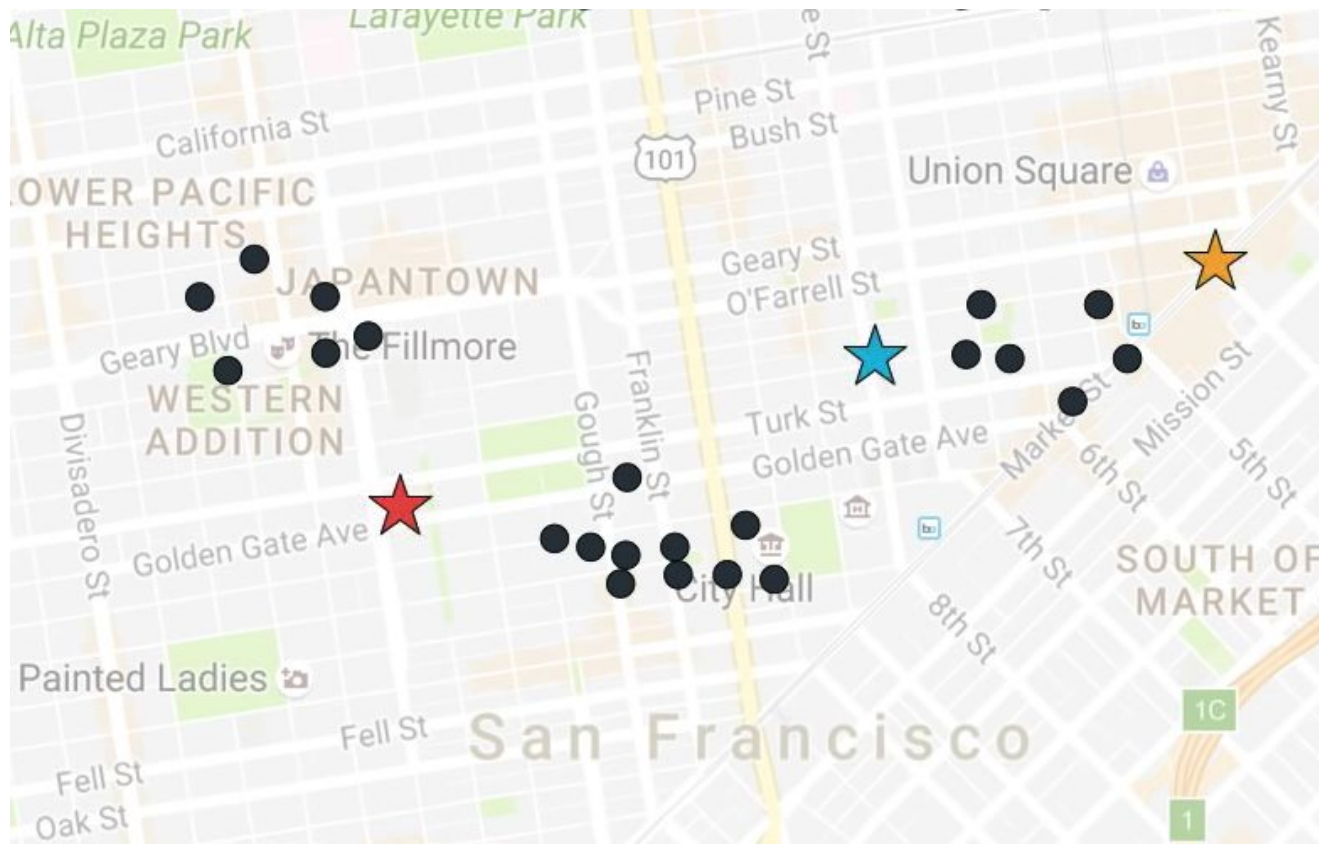


Agrupamentos / Clustering

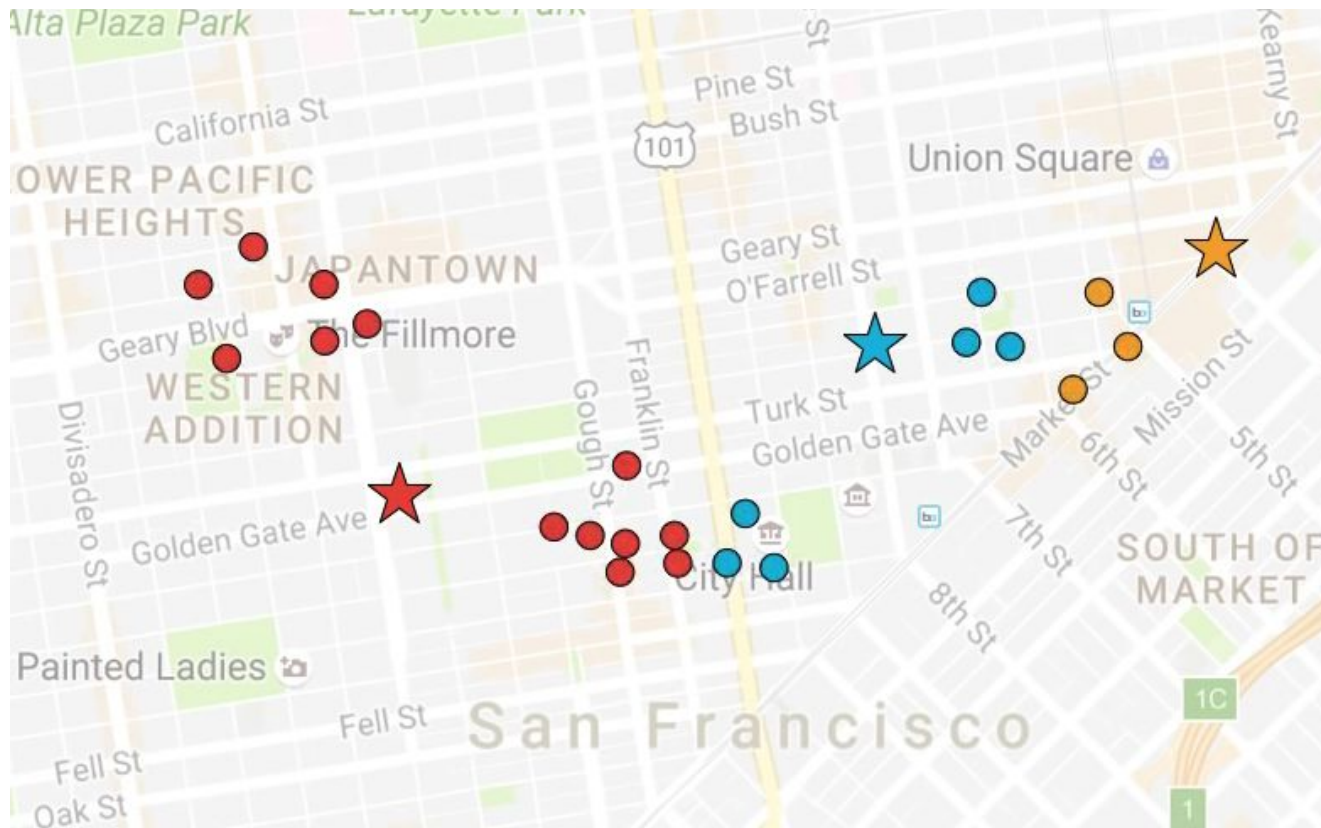
Algoritmos de agrupamentos, ou **clustering**, tentam descobrir o **arranjo 'natural'** de amostras quaisquer através de uma **medida de semelhança** entre eles.

- Tarefa de agrupar os exemplos sem conhecer sua classificação prévia (conhecimento do especialista)
- Existem diversas abordagens na literatura para realizar essa atividade, neste curso iremos focar no K-Means
- Aplicações:
 - Sistemas de recomendação e redes sociais
 - Biologia evolucionária, genética, bactérias.
 - Data labelling

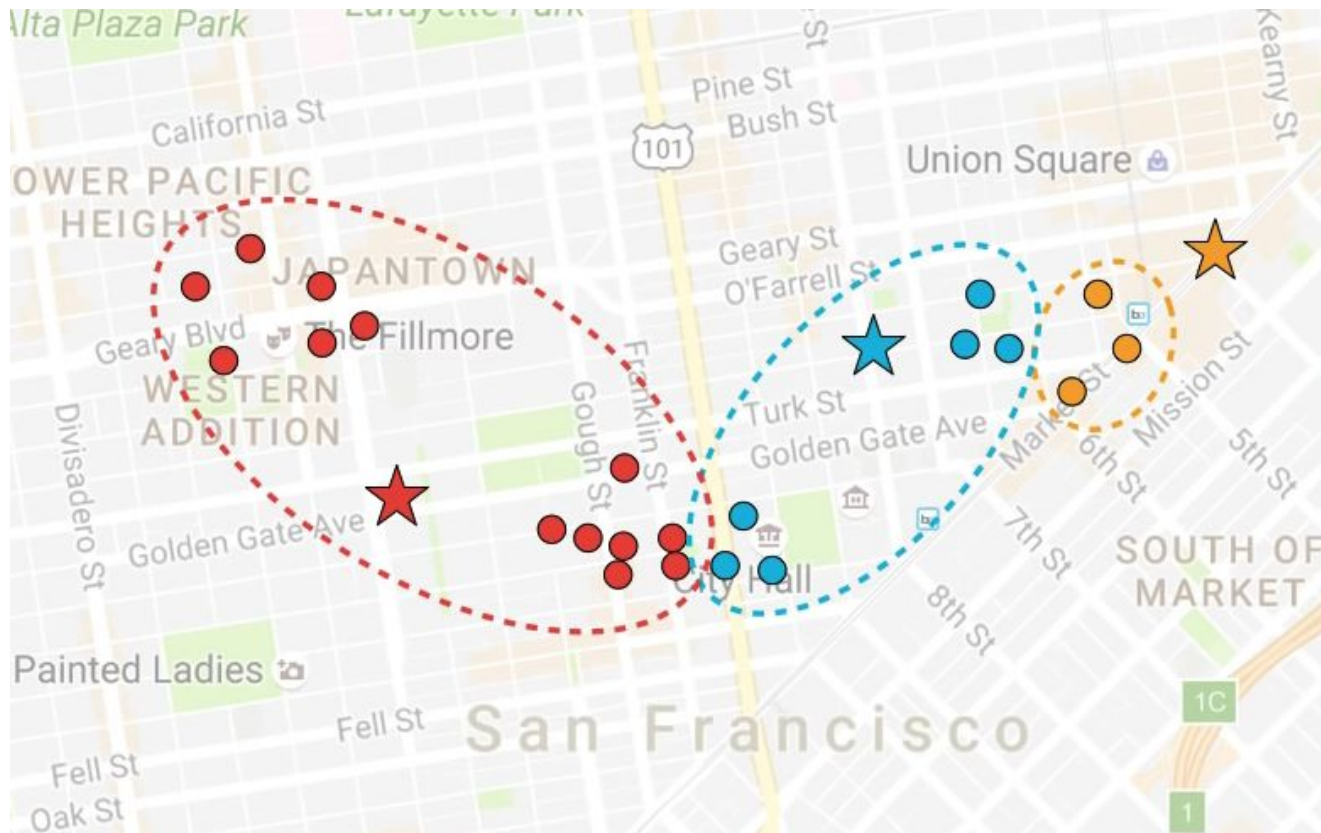
K-Means - Exemplo



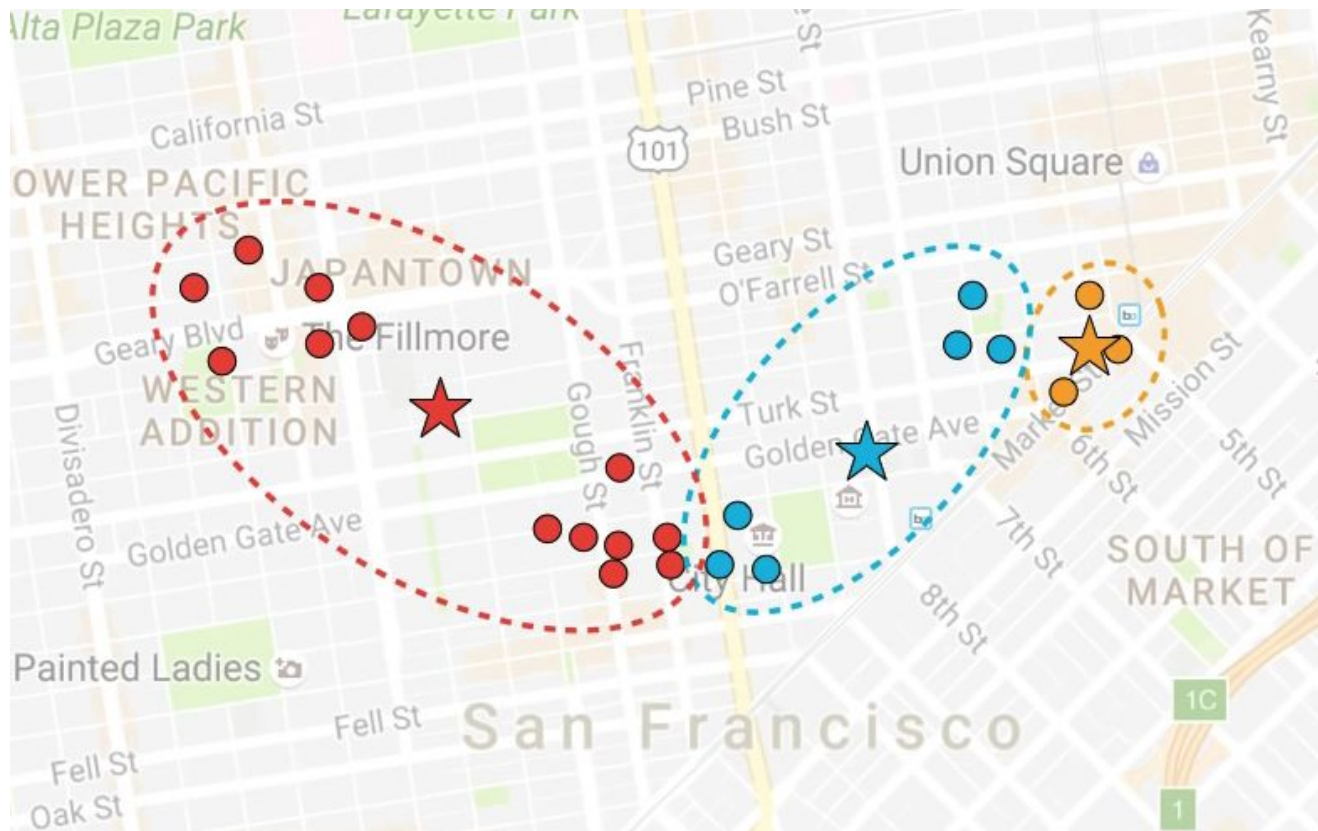
K-Means - Exemplo



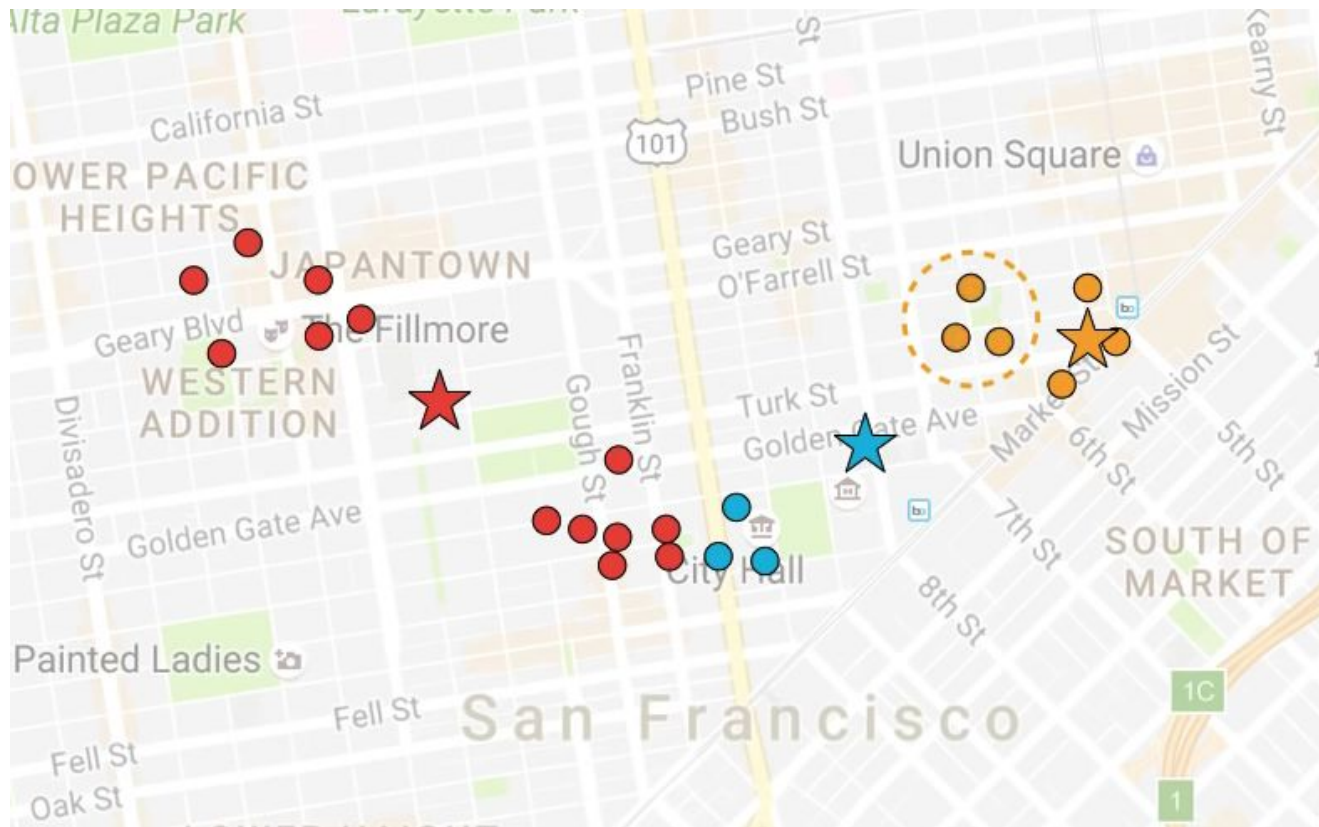
K-Means - Exemplo



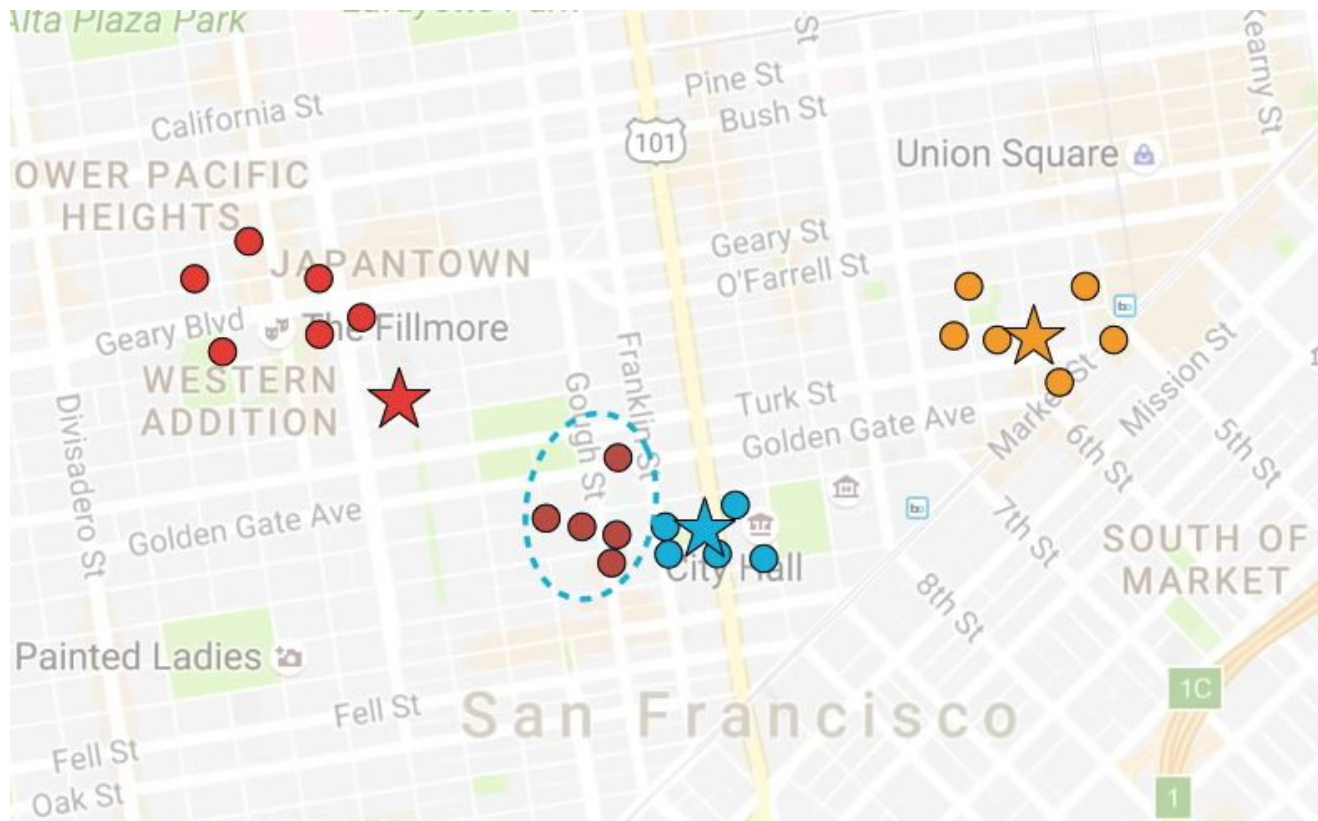
K-Means - Exemplo



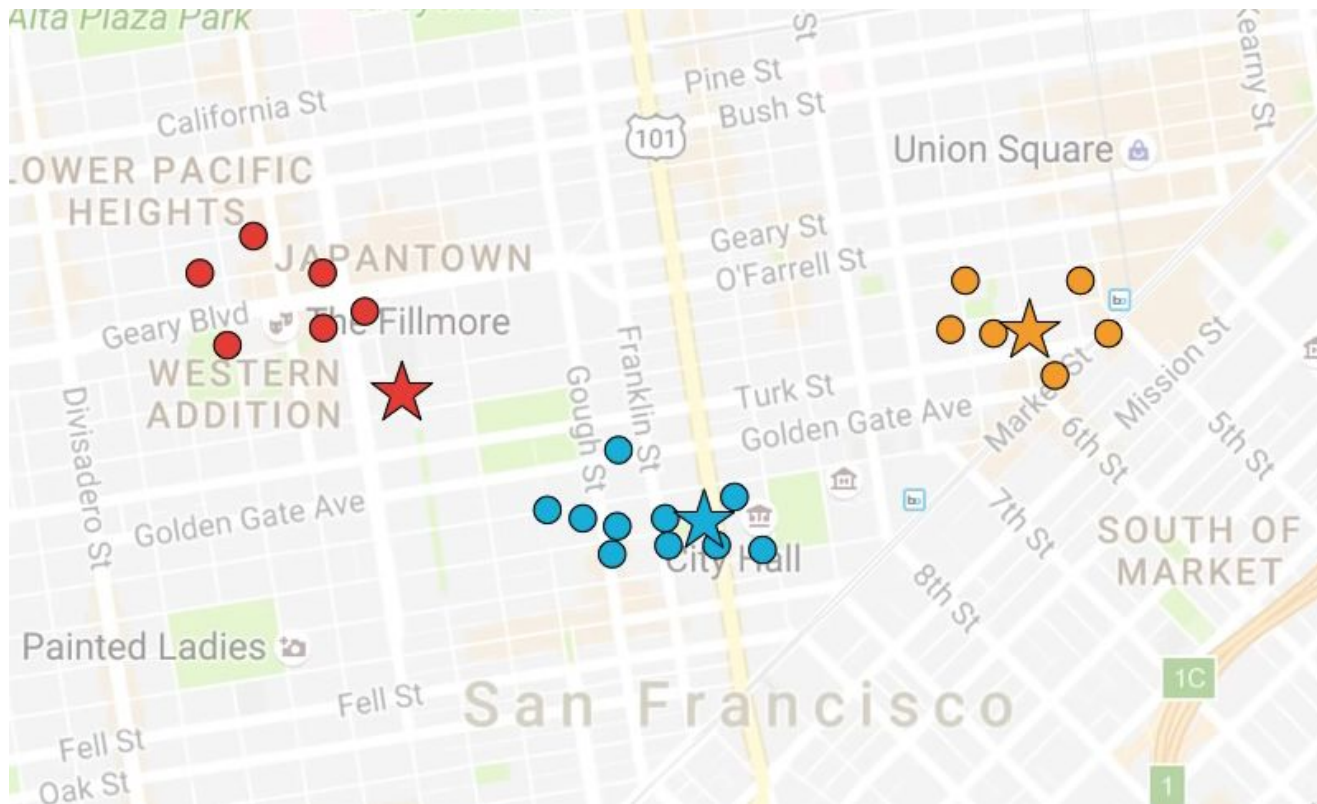
K-Means - Exemplo



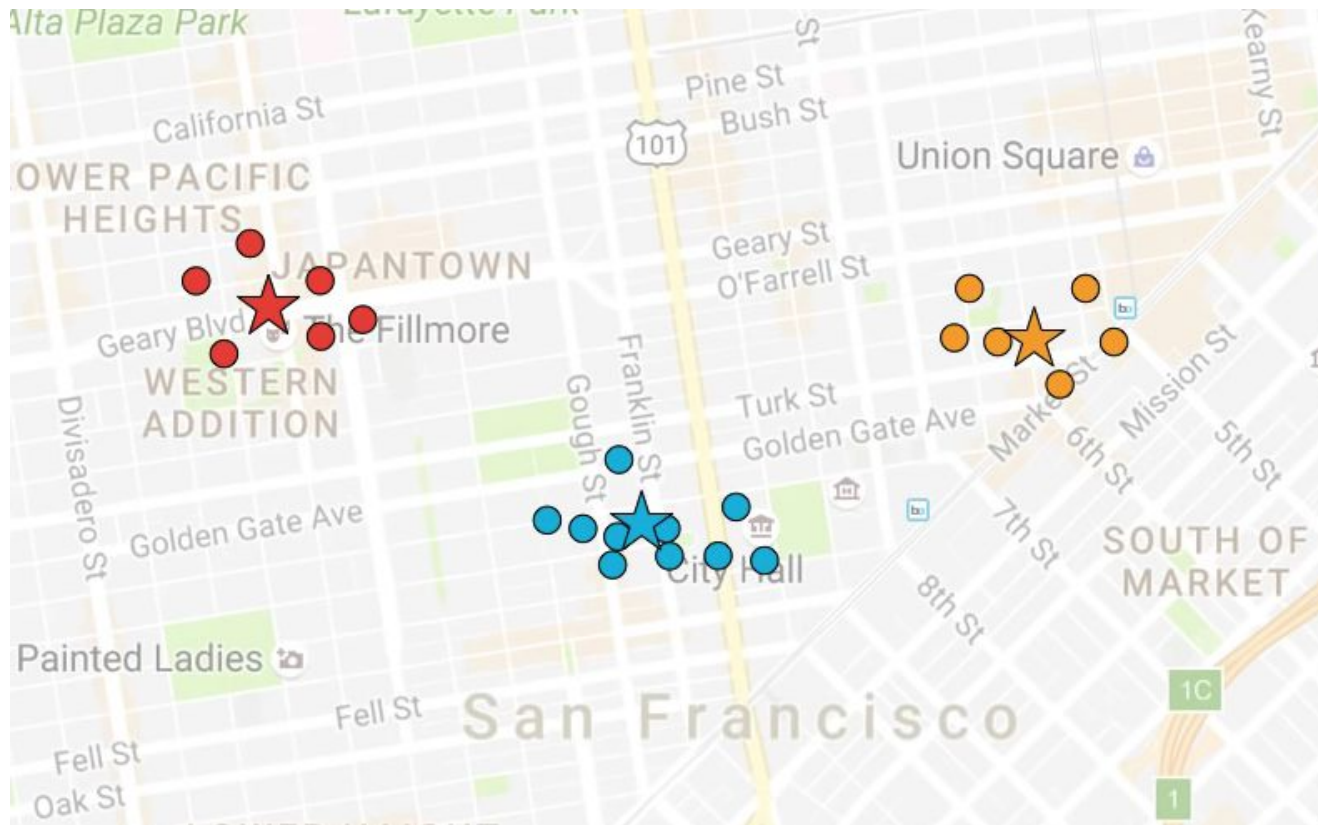
K-Means - Exemplo



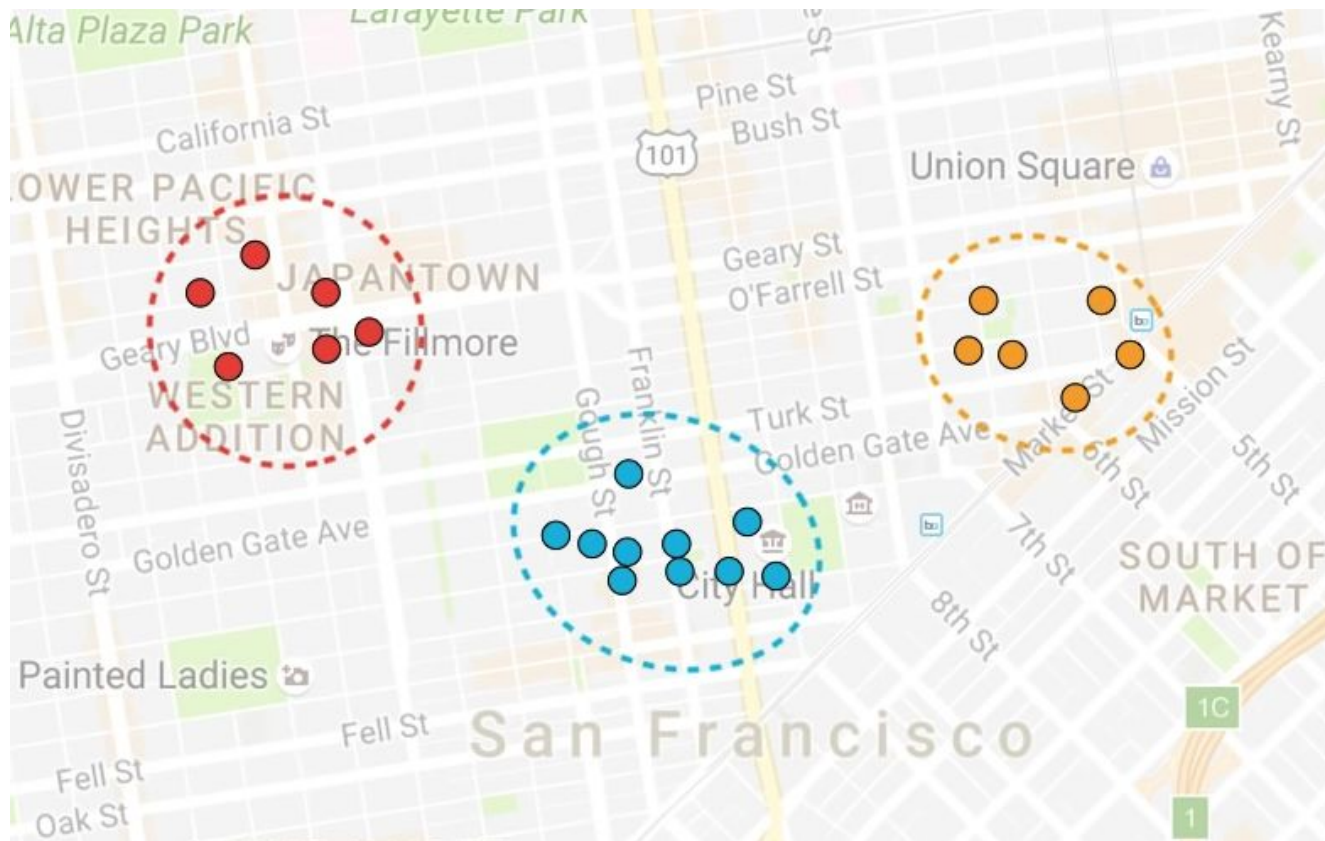
K-Means - Exemplo



K-Means - Exemplo



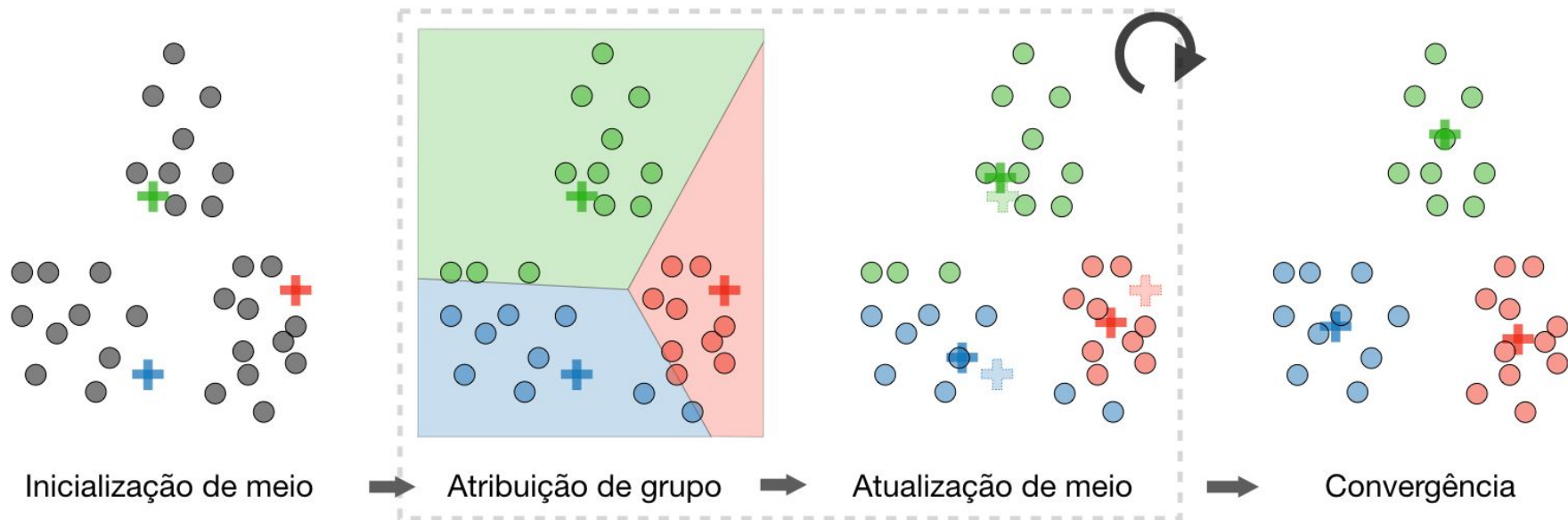
K-Means - Exemplo



K-Means - Exemplo

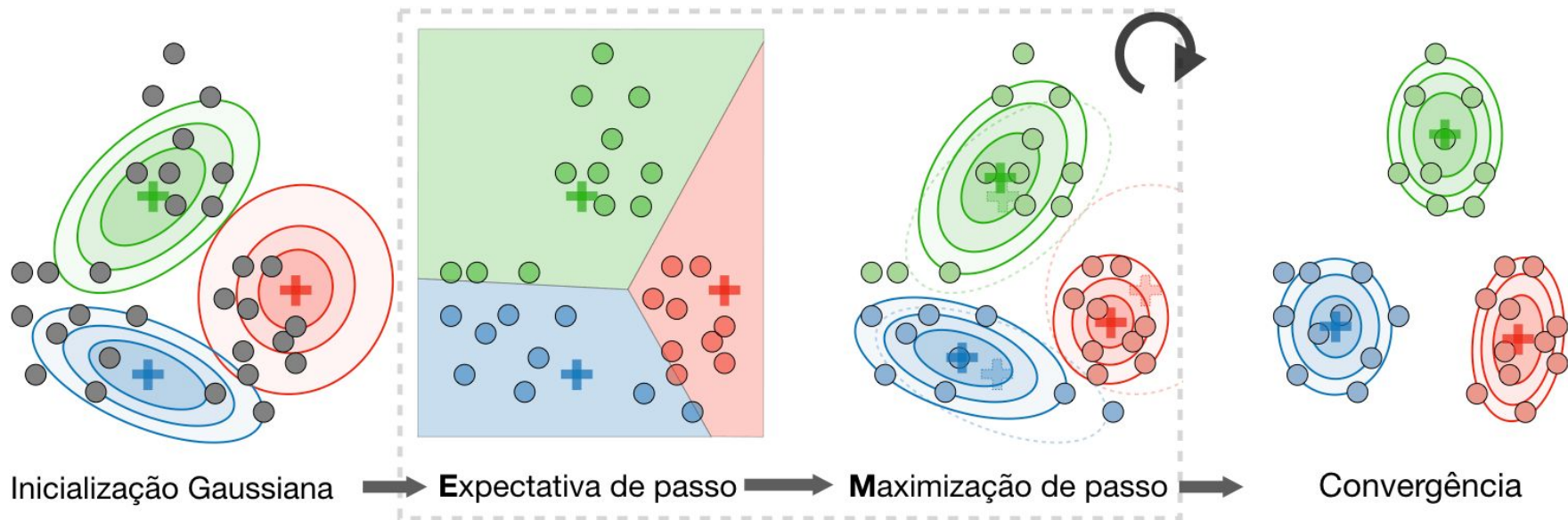
K-Means - Esquema

ref: <https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-aprendizado-nao-supervisionado>



K-Means - Esquema de Agrupamento

ref: <https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-aprendizado-nao-supervisionado>

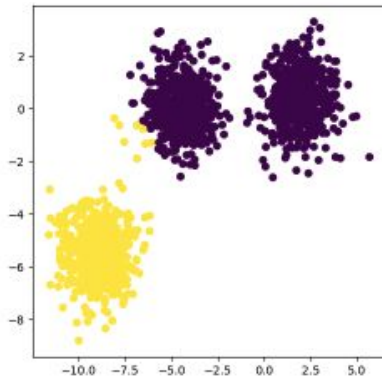


K-Means

Particiona as observações em **k** grupos, nos quais cada observação vai pertencer ao grupo que, na média, é mais similar a ele

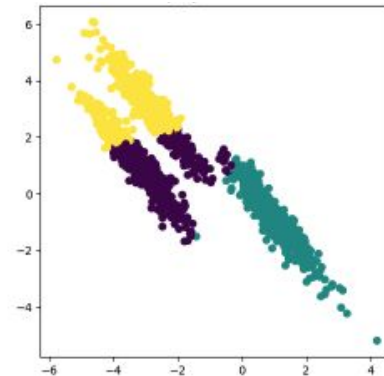
Prós

- Simples e eficaz
- Computacionalmente eficiente e escalável
- Gera protótipos que representam os grupos





Contras

- 'k' precisa ser determinado
- Sensível a ruído e outliers
- Não é determinístico
- Pressupõe distribuição gaussiana



K-Means

ref: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

[Install](#) [User Guide](#) [API](#) [Examples](#) [More](#) 

[Prev](#) [Up](#) [Next](#)

scikit-learn 1.0.2
[Other versions](#)

Please [cite us](#) if you use the software.

[sklearn.cluster.KMeans](#)
Examples using
[sklearn.cluster.KMeans](#)

sklearn.cluster.KMeans

```
class sklearn.cluster.KMeans(n_clusters=8, *, init='k-means++', n_init=10, max_iter=300, tol=0.0001, verbose=0,
                             random_state=None, copy_x=True, algorithm='auto')
```

[\[source\]](#)

K-Means clustering.

Read more in the [User Guide](#).

Parameters:

- n_clusters : int, default=8**
The number of clusters to form as well as the number of centroids to generate.
- init : {'k-means++', 'random'}, callable or array-like of shape (n_clusters, n_features), default='k-means++'**
Method for initialization:
 - 'k-means++': selects initial cluster centers for k-mean clustering in a smart way to speed up convergence. See section Notes in `k_init` for more details.
 - 'random': choose `n_clusters` observations (rows) at random from data for the initial centroids.
- If an array is passed, it should be of shape `(n_clusters, n_features)` and gives the initial centers.
- If a callable is passed, it should take arguments `X`, `n_clusters` and a random state and return an initialization.

K-Means

ref: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>



K-Means em Python

Code use example

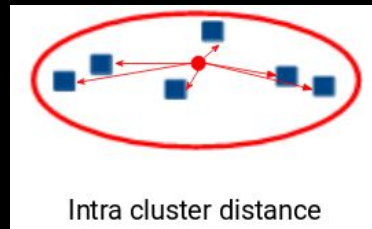
```
# importa o pacote cluster
from sklearn.cluster import KMeans

# Cria o modelo
model = KMeans(n_clusters=2).fit(X)

# grupos encontrados
model.labels_

# centroides
model.cluster_centers_

# Soma do quadrado das distâncias das
# amostras para o centro do seu cluster.
# Também pode ser utilizado para o Elbow method!
model.inertia_
```





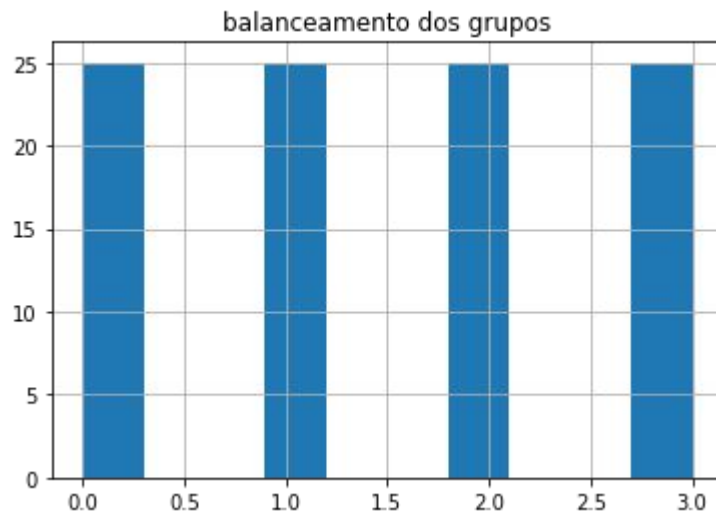
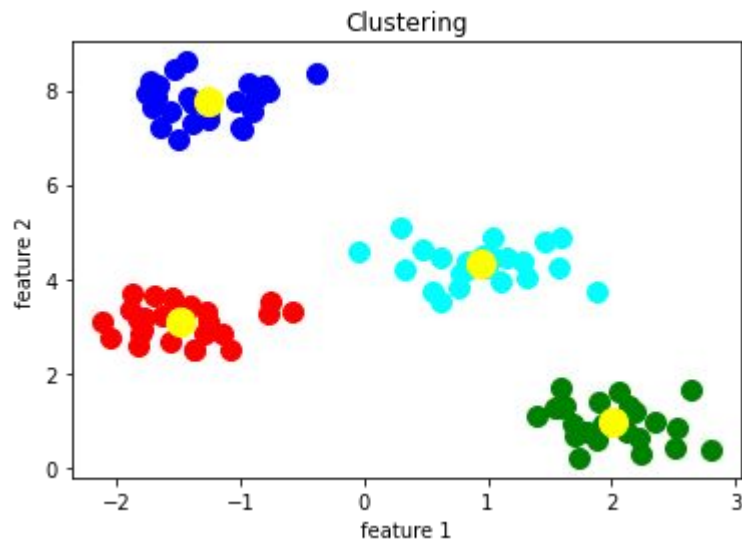
Hands On!

Experimentos em Machine Learning com Python

Hands On: Resultado Esperado

Silhueta: 0.7697826124517921

NMI: 1.0

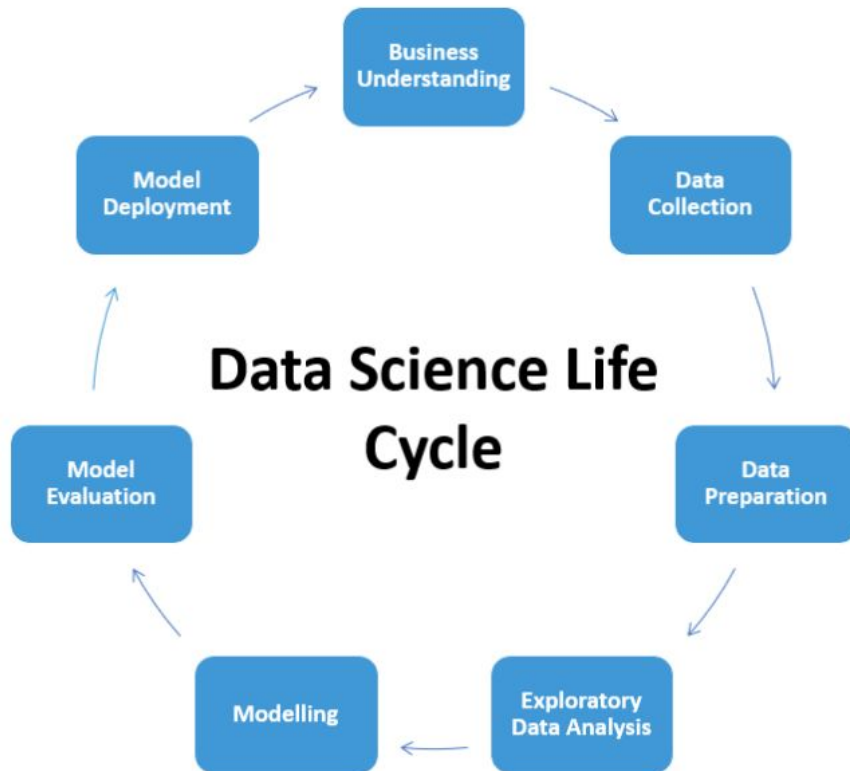
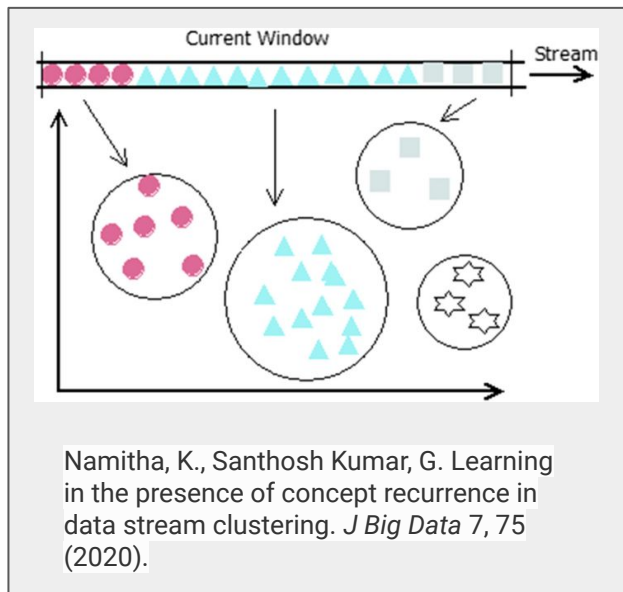




Avaliação de Agrupamentos

Data Pipeline

Ref: <https://laptrinhx.com/complete-life-cycle-of-a-data-science-machine-learning-project-3256392522/>



mas antes ...

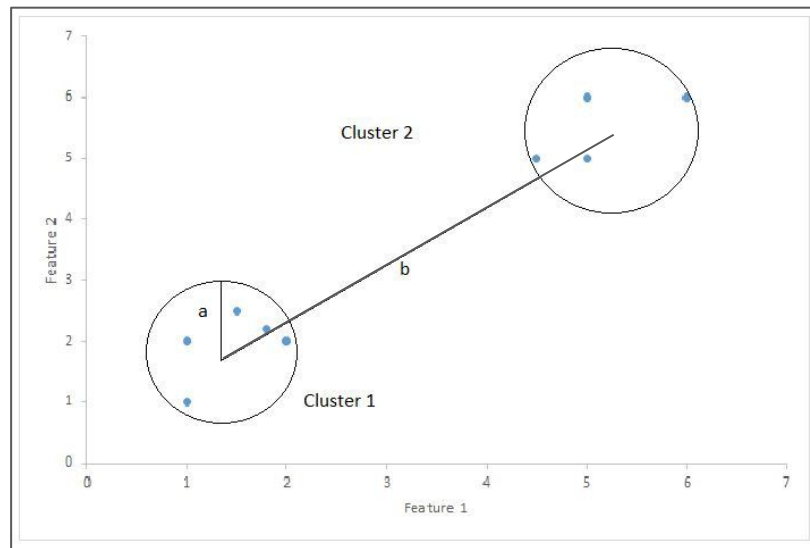
Avaliação de Agrupamentos

Principais características

- Alta densidade **em um mesmo** grupo (homogeneidade)
- Alta separabilidade **entre** grupos (heterogeneidade)

Outras características

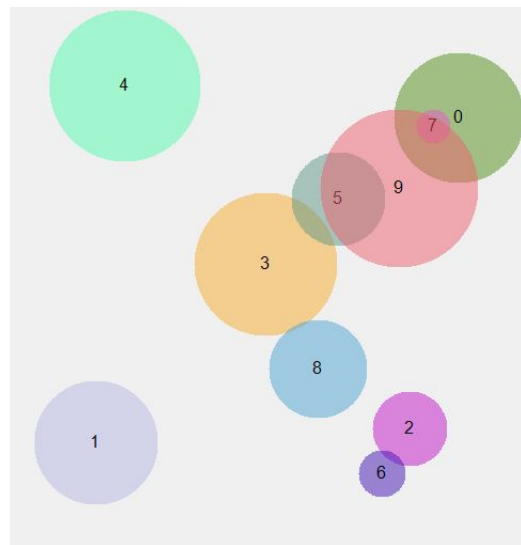
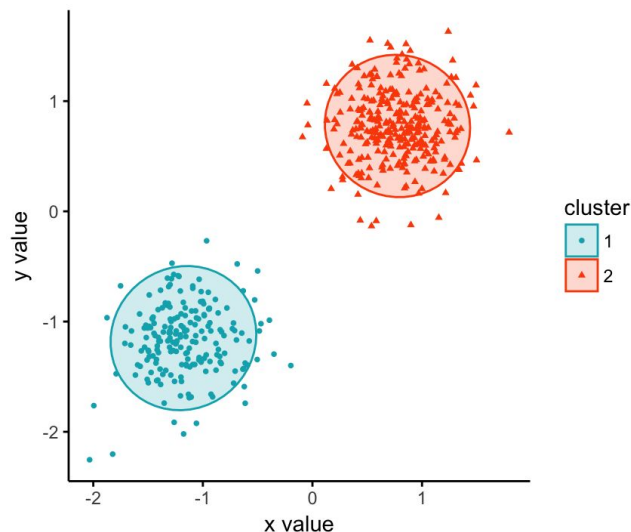
- Balanceamento de itens por grupo



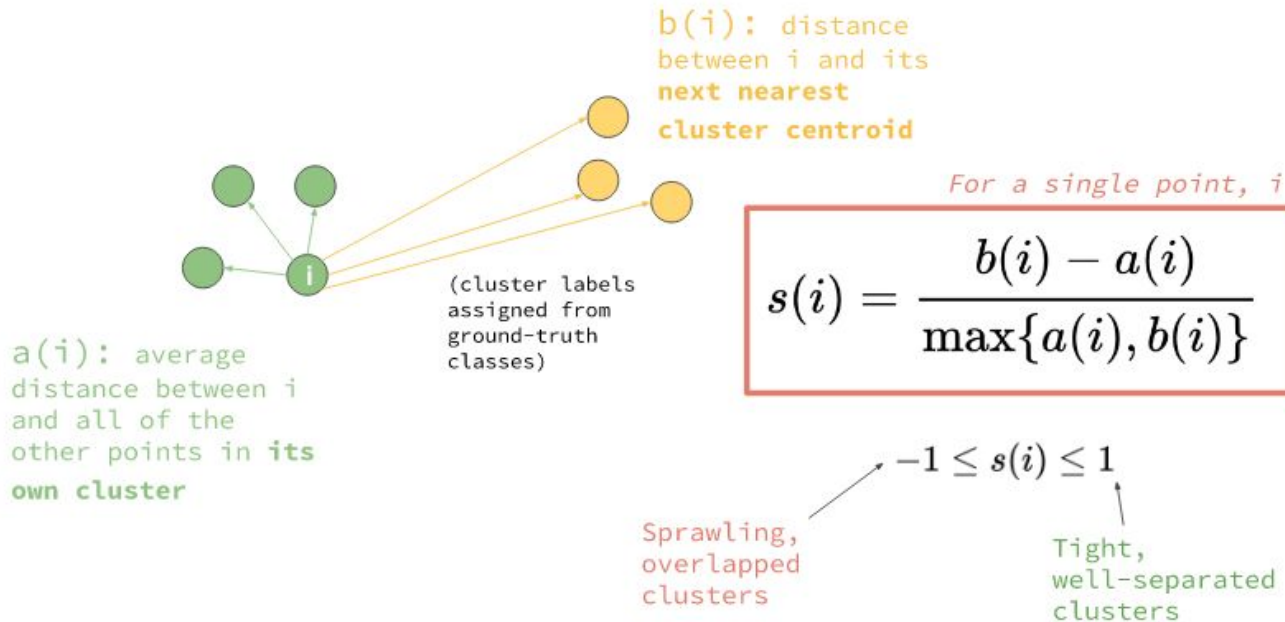
O que seria um “bom” agrupamento ?

Um resultado esperado de uma atividade de clustering é a geração de grupos

- com **alta densidade** (menor distância entre os pontos de um mesmo grupo)
- e **maior separabilidade** (maior distância entre os grupos).



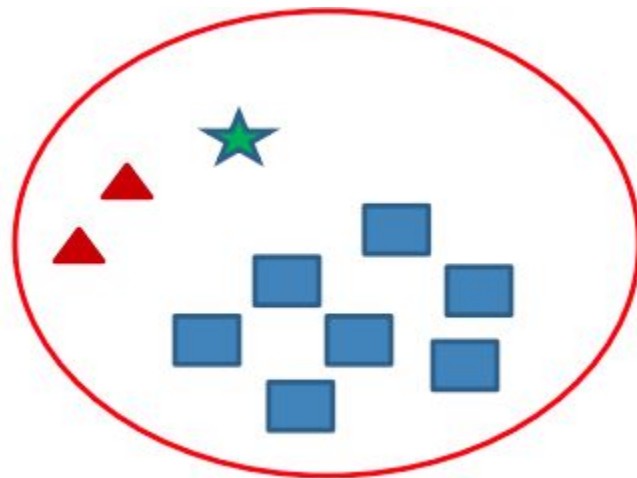
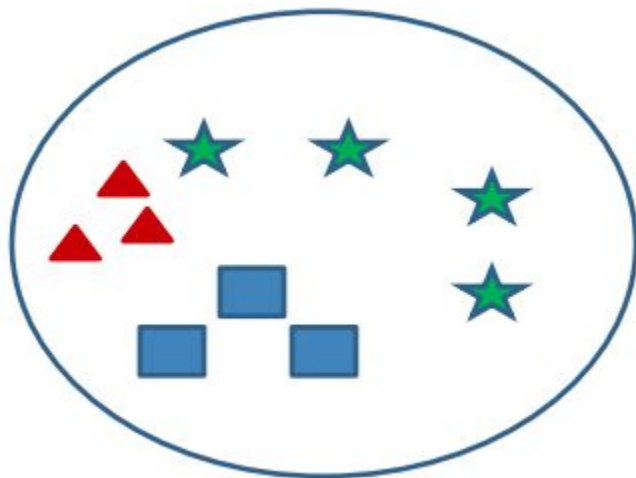
Silhouette



Normalized Mutual Information (NMI)

ref: https://course.ccs.neu.edu/cs6140sp15/7_locality_cluster/Assignment-6/NMI.pdf

- Quanto maior (max=1) melhor.
- Avalia o quão “puro” são os grupos (baixa entropia)



```
# importa o pacote cluster
from sklearn import metrics

# silhouette
sil = metrics.silhouette_score(X, labels, metric='euclidean')

# NMI
nmi = metrics.normalized_mutual_info_score(y_true, labels)
```

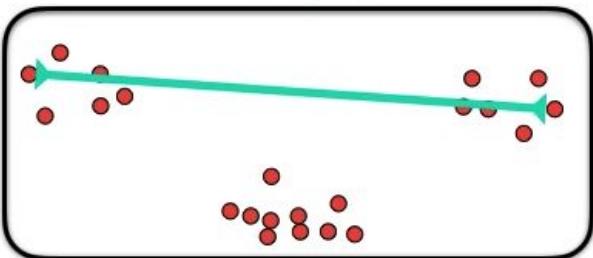


Elbow Method

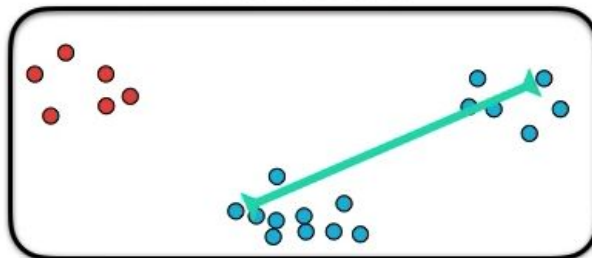
Como definir k?

Calcular o K-Means para diversos valores de k

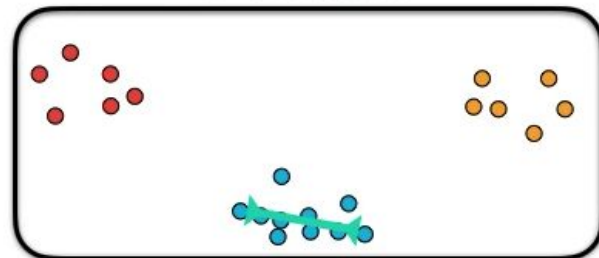
1 cluster



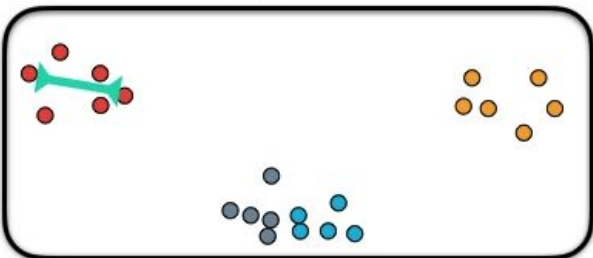
2 clusters



3 clusters



4 clusters



5 clusters



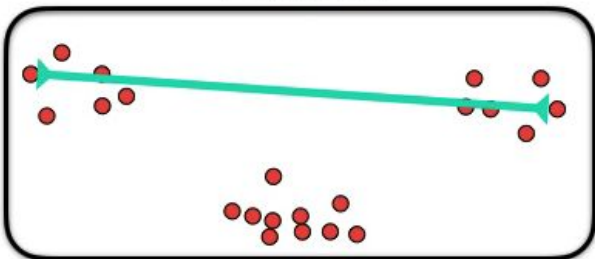
6 clusters



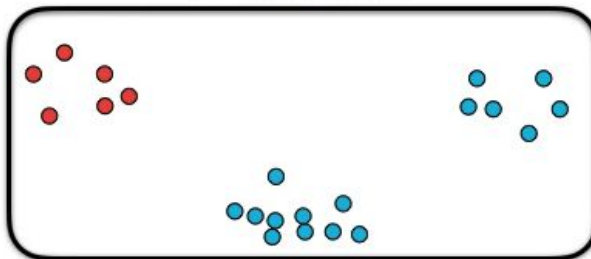
Como definir k?

Calcular a maior distância entre exemplos de um mesmo grupo

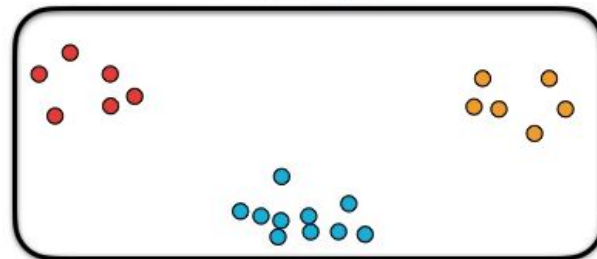
1 cluster



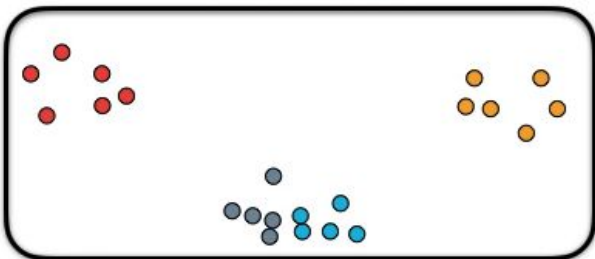
2 clusters



3 clusters



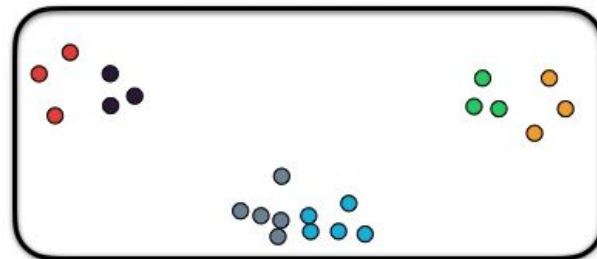
4 clusters



5 clusters



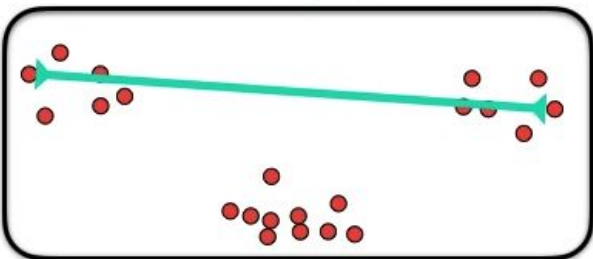
6 clusters



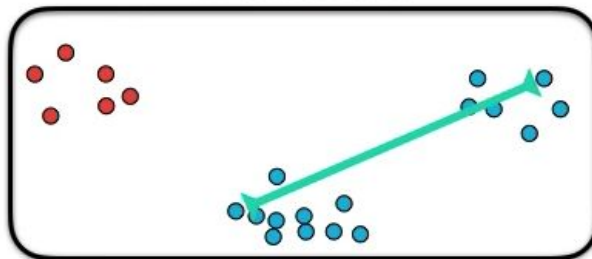
Como definir k?

Calcular a maior distância entre exemplos de um mesmo grupo

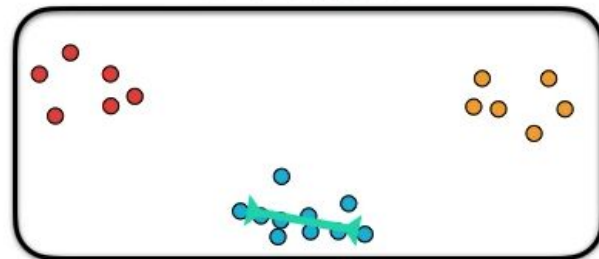
1 cluster



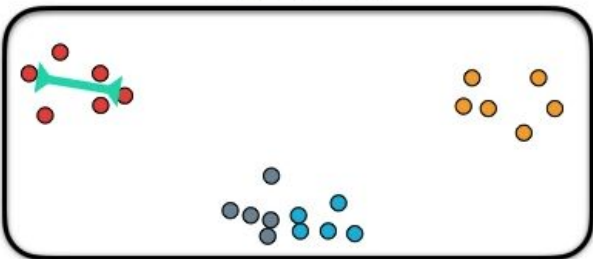
2 clusters



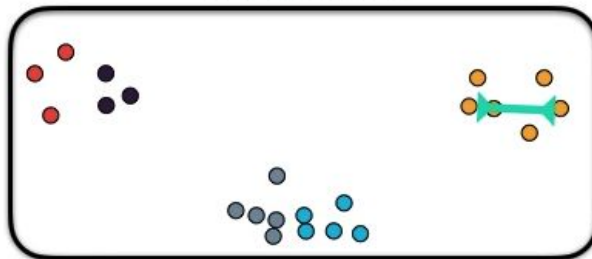
3 clusters



4 clusters



5 clusters



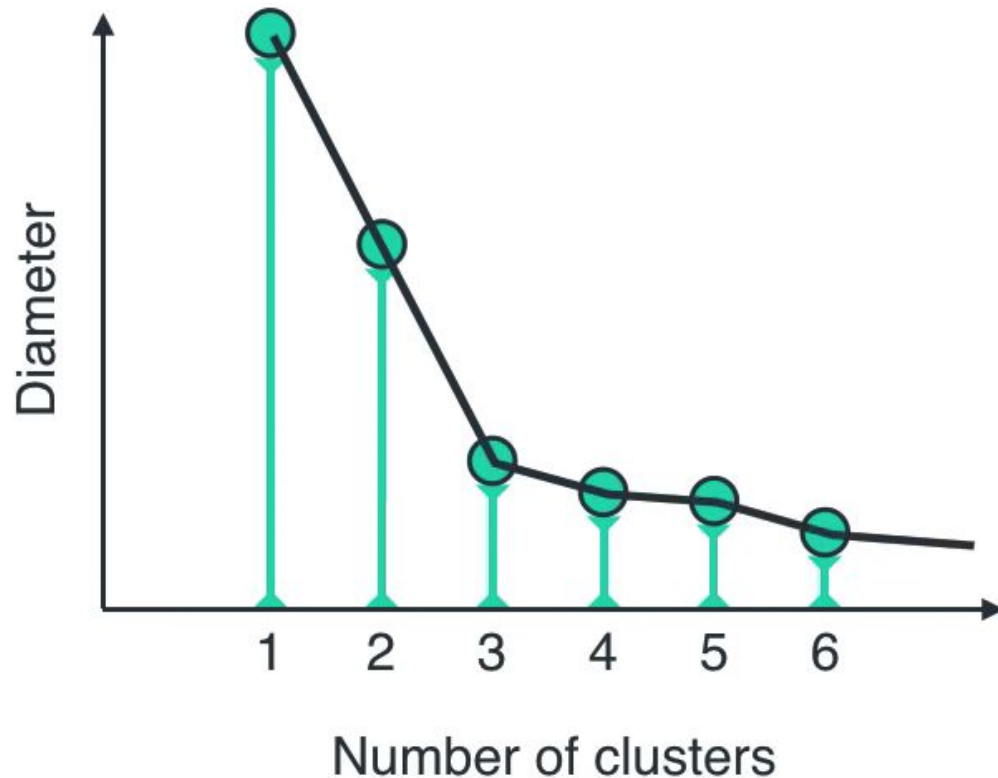
6 clusters



Como definir k?

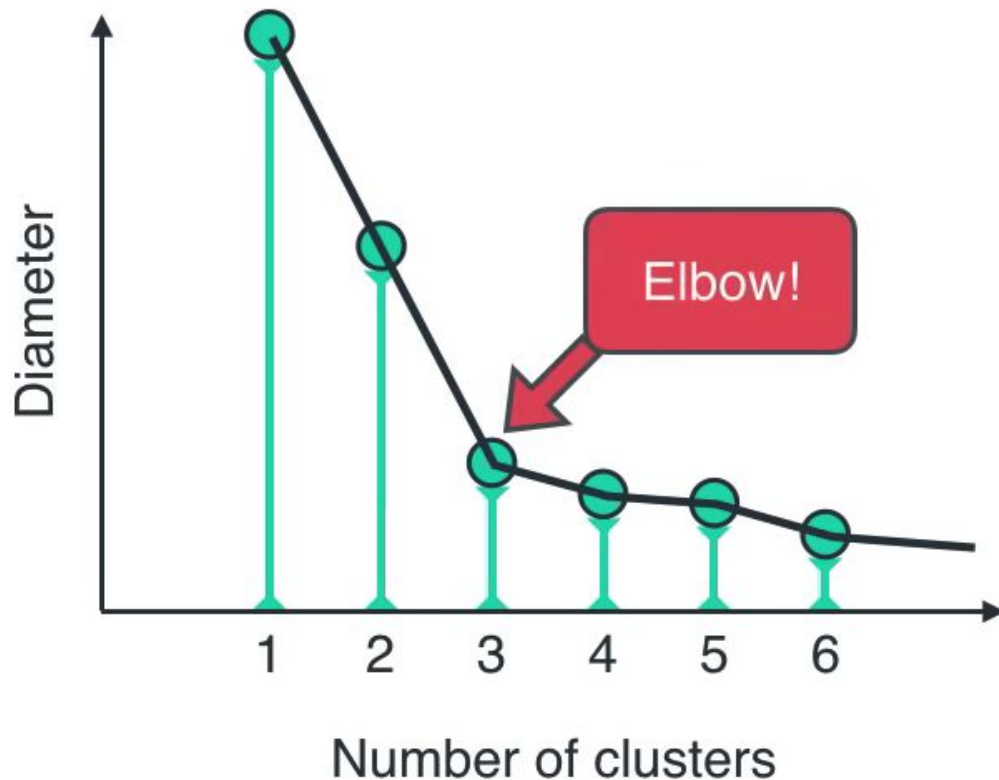
Plotar a variação intra-clusters pelo número de clusters (k).

Procurar pelo 'cotovelo' da curva!



Como definir k?

O melhor valor de k será definido pelo ponto no qual um aumento no número de clusters não significa uma grande diminuição no diâmetro.



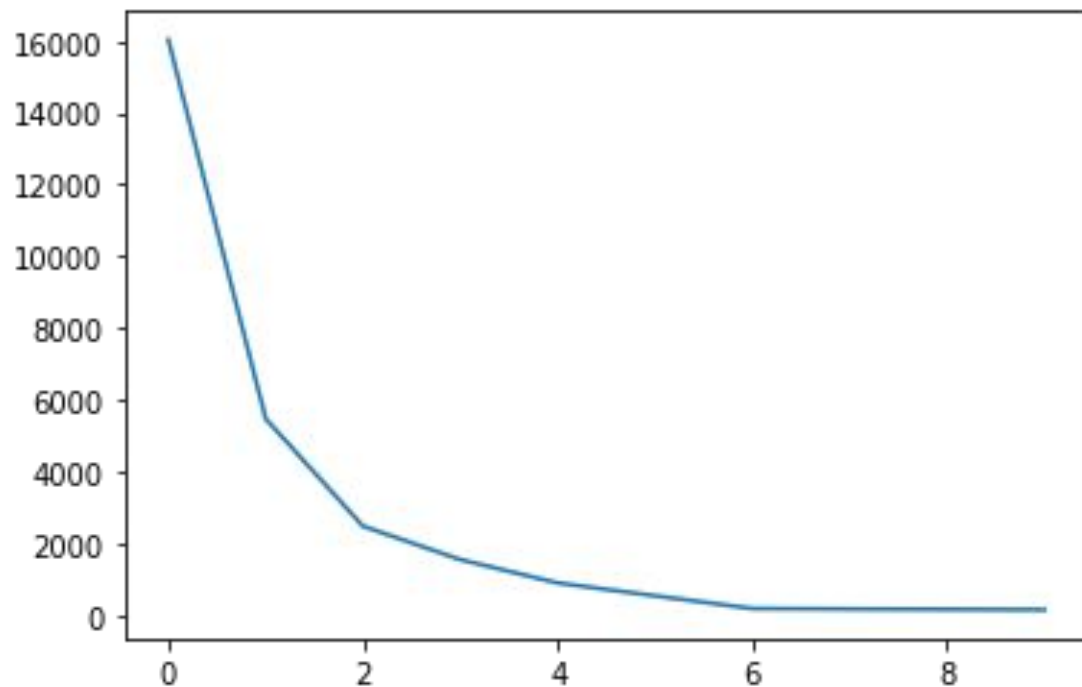


Hands On!

Experimentos em Machine Learning com Python

Hands On: Resultado Esperado

best K: 7



Para continuar:

- [StatQuest: K-Means clustering](#)
- [Clustering: K-Means and Hierarchical](#)



C . E . S . A . R

Pessoas impulsionando inovação.
Inovação impulsionando negócios.

NOSSO CONTATO

mso@cesar.org.br

mso2@cesar.school

