



# APRENDIZAGEM DE MÁQUINA

PROF. JOSENALDE OLIVEIRA

[josenalde.oliveira@ufrn.br](mailto:josenalde.oliveira@ufrn.br)

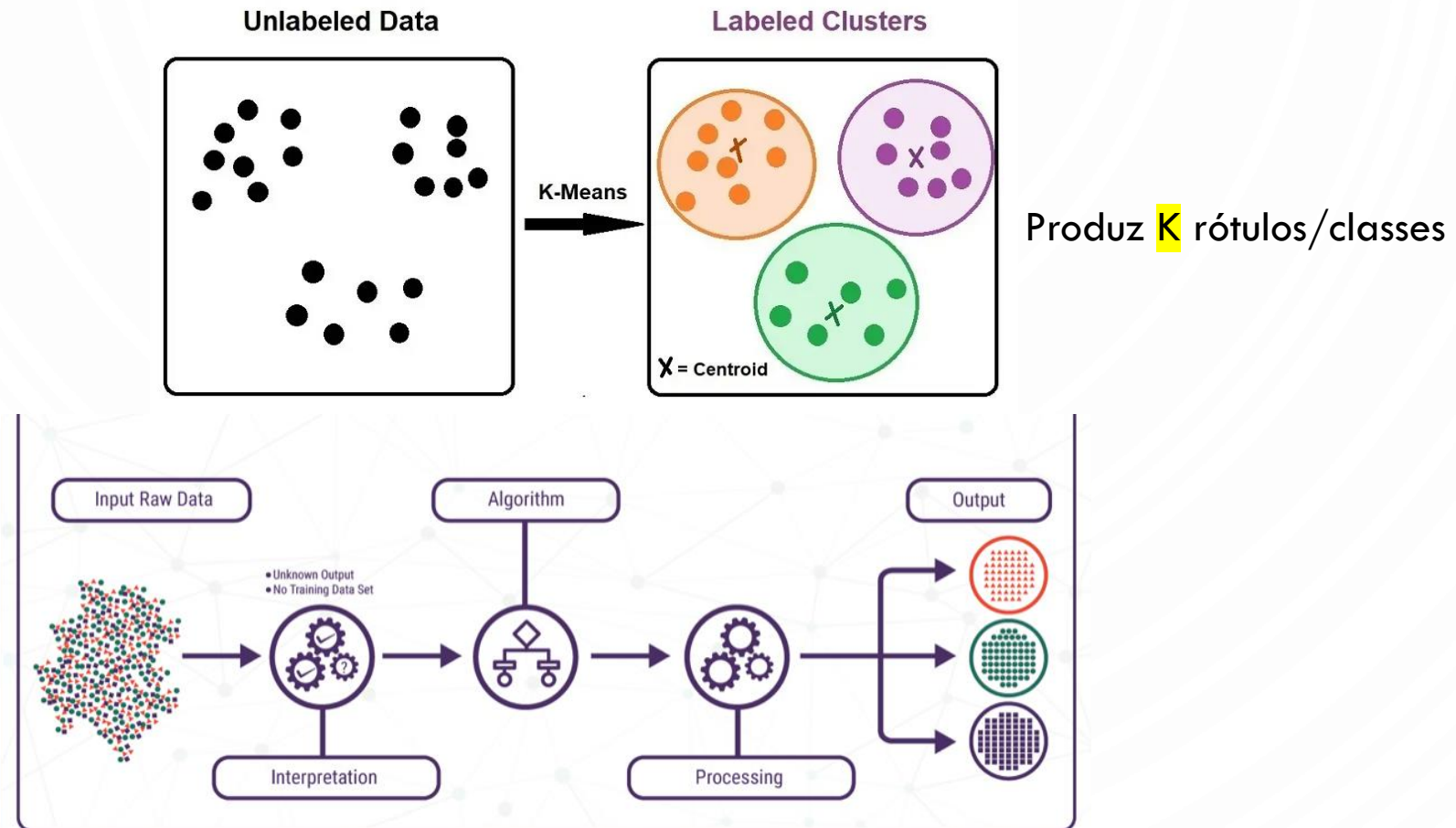
<https://github.com/josenalde/machinelearning>

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - UFRN

# APRENDIZADO NÃO SUPERVISIONADO

## AGRUPAMENTO (CLUSTERING)

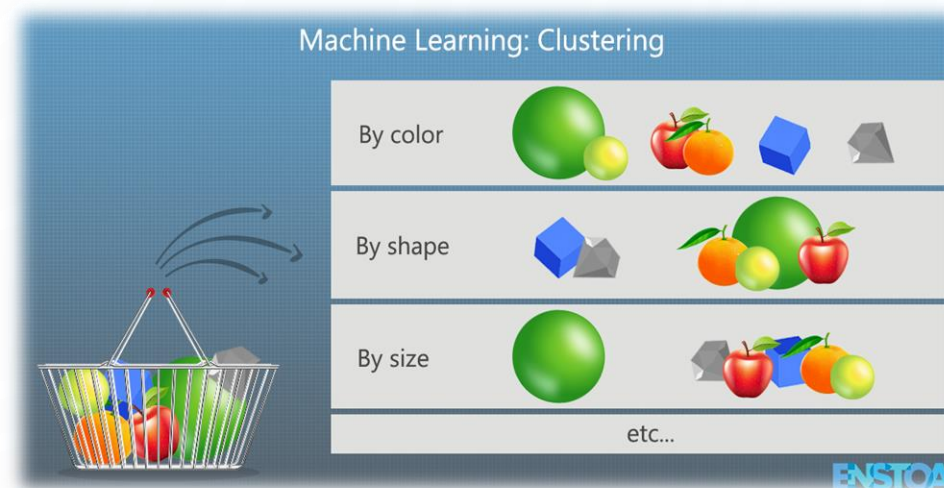
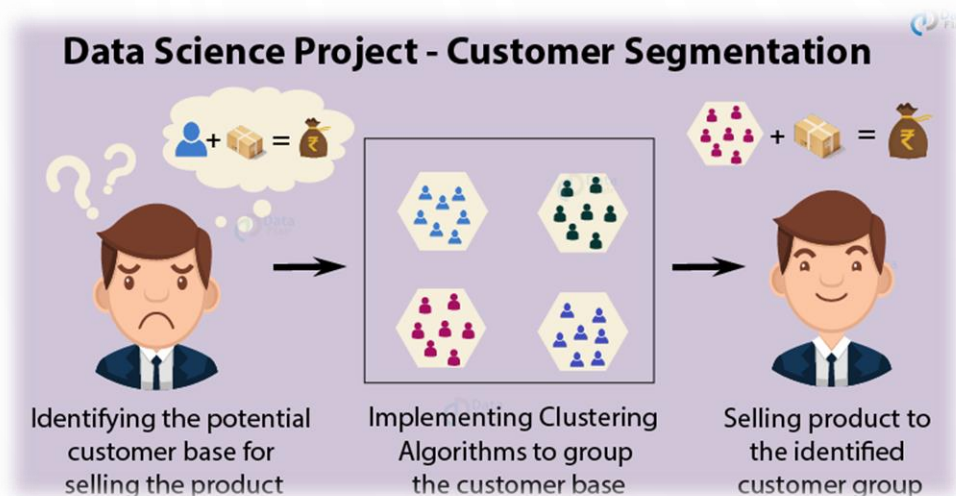
A partir de dataset sem considerar coluna alvo (target) – unlabeled data, explorar padrões entre as features e agrupar por similaridade, de modo a manter próximo 'similares' e afastar 'discrepantes', criando clusters



# APRENDIZADO NÃO SUPERVISIONADO

## AGRUPAMENTO (CLUSTERING)

### *Algumas aplicações*



Projeto de pesquisa (exemplo): extrair features (feature engineering) de imagens com técnicas PDI e alimentar algoritmo de clusterização: [https://github.com/josenalde/Projeto\\_Pesquisa\\_PalmaS/blob/main/Projeto\\_Segmentado.ipynb](https://github.com/josenalde/Projeto_Pesquisa_PalmaS/blob/main/Projeto_Segmentado.ipynb)

Projeto de pesquisa (exemplo): lotes de frangos de corte mistos – agrupar M/F por medidas de peso, de modo a obter conjuntos suportes e aplicar média ponderada de peso – corrigir previsão de peso final de abate

# APRENDIZADO NÃO SUPERVISIONADO

## AGRUPAMENTO (CLUSTERING)

### *Algumas aplicações*

A própria técnica de redução de dimensionalidade (como PCA por exemplo) é uma tarefa não supervisionada.

As principais motivações para reduzir dimensionalidade são:

- Acelerar um algoritmo de treinamento posterior (remover ruído, características redundantes...)
- Visualizar os dados e obter insights sobre as features mais importantes
- Economizar espaço (compactação)

Mas há também desvantagens, como:

- Perder algumas informações, talvez com prejuízo para desempenho dos algoritmos de treinamento
- Em termos computacionais, pode ser custoso
- Adiciona um pouco de complexidade aos pipelines ML
- As features transformadas costumam ser difíceis de interpretar

# APRENDIZADO NÃO SUPERVISIONADO

## AGRUPAMENTO (CLUSTERING)

### *Algumas aplicações*

- **Segmentação de clientes:** perfis com base em atividades em site, por exemplo. Muito usado em sistemas de recomendação
- **Análise de dados:** perceber padrões em dados
- **Redução de dimensionalidade**
- **Detecção de anomalias** (outliers): instância com baixa afinidade com todos os clusters provavelmente será anomalia. Uso em detecção de defeitos, falhas, fraudes...
- **Aprendizado semisupervisionado:** caso tenha apenas alguns rótulos, poderá executar a clusterização e propagar os rótulos para todas as instâncias do mesmo cluster, aumentando o número de rótulos para supervisionado subsequente
- **Mecanismos de busca:** busca de imagens semelhantes por meio de uma imagem de referência. Suponha um imageset. Imagens semelhantes estariam no mesmo cluster e quando o usuário fornece imagem de referência, encontra-se o cluster desta imagem
- **Segmentação de imagens:** clusterizar pixels de acordo com a cor e substituir a cor de cada pixel pela cor média do cluster, reduzindo a quantidade de cores diferentes na imagem. Usado em detecção e rastreamento de objetos, facilitando detectar contornos

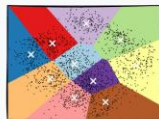
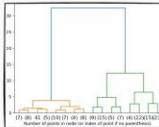
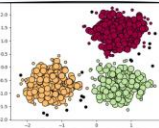
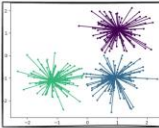
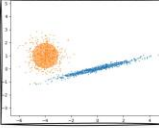
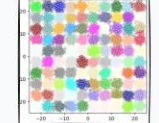
# CLUSTERING – TIPOS DE ALGORITMOS

## AGRUPAMENTO (CLUSTERING)

Algumas aplicações

### 6 Types of Clustering Algorithms in Machine Learning

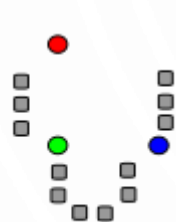
 [blog.DailyDoseofDS.com](http://blog.DailyDoseofDS.com)

| Clustering Algorithm Type                                                            |                    | Clustering Methodology                                                          | Algorithm(s)                                         |
|--------------------------------------------------------------------------------------|--------------------|---------------------------------------------------------------------------------|------------------------------------------------------|
|    | Centroid-based     | Cluster points based on proximity to centroid                                   | KMeans<br>KMeans++<br>KMedoids                       |
|    | Connectivity-based | Cluster points based on proximity between clusters                              | Hierarchical Clustering (Agglomerative and Divisive) |
|    | Density-based      | Cluster points based on their density instead of proximity                      | DBSCAN<br>OPTICS<br>HDBSCAN                          |
|   | Graph-based        | Cluster points based on graph distance                                          | Affinity Propagation<br>Spectral Clustering          |
|  | Distribution-based | Cluster points based on their likelihood of belonging to the same distribution. | Gaussian Mixture Models (GMMs)                       |
|  | Compression-based  | Transform data to a lower dimensional space and then perform clustering         | BIRCH                                                |

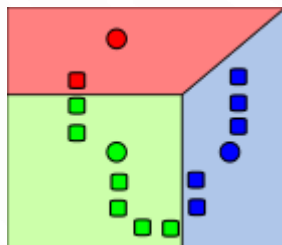


# CLUSTERING – K-MEANS (K MÉDIAS)

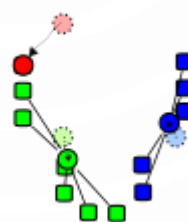
- Proposto em 1957 – Stuart Lloyd (Bell Labs); 1965 Edward Forgy também publica (conhecido como Lloyd-Forgy). Conhecido fora da Bell em 1982. James MacQueen cunhou o termo “k-means” em 1967.



1. K médias iniciais são inicializadas aleatoriamente, definindo centroides



2. Calculando as distâncias de cada instância com o centroide, formam-se K clusters. Estas partições se chamam diagrama Voronoi



3. Os centroides de cada cluster se tornam a nova média



4. Repetem-se os passos 2 e 3 até convergência

- O método de inicialização dos centroides é importante para o desempenho
- As features precisam ser escalonadas antes de executar o K-means
- Vamos analisar alguns notebooks Jupyter de clusterização