

TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

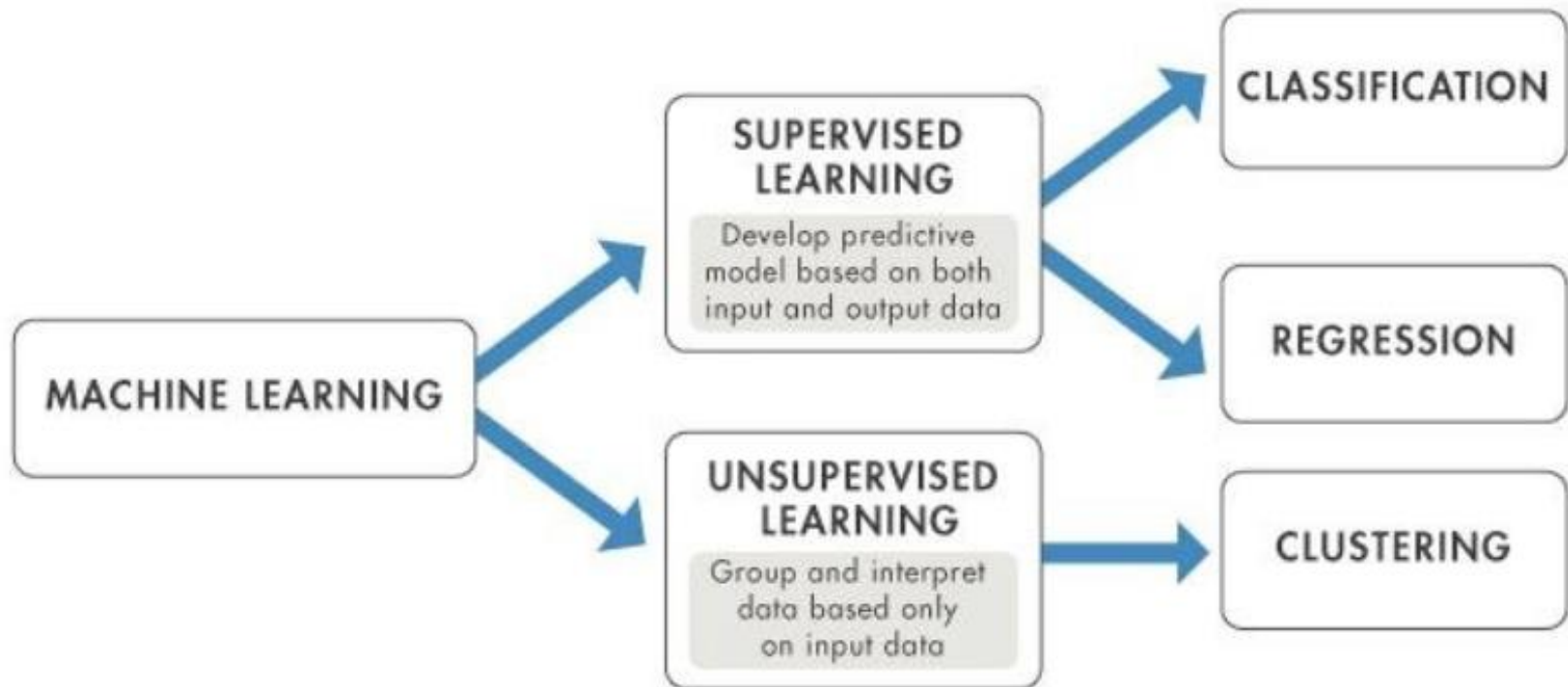
Aprendizado de Máquina Não Supervisionado – Agrupamento (Clustering)

Prof. André Tritiack

profandre.farias@fiap.com.br

Aprendizado de Máquina

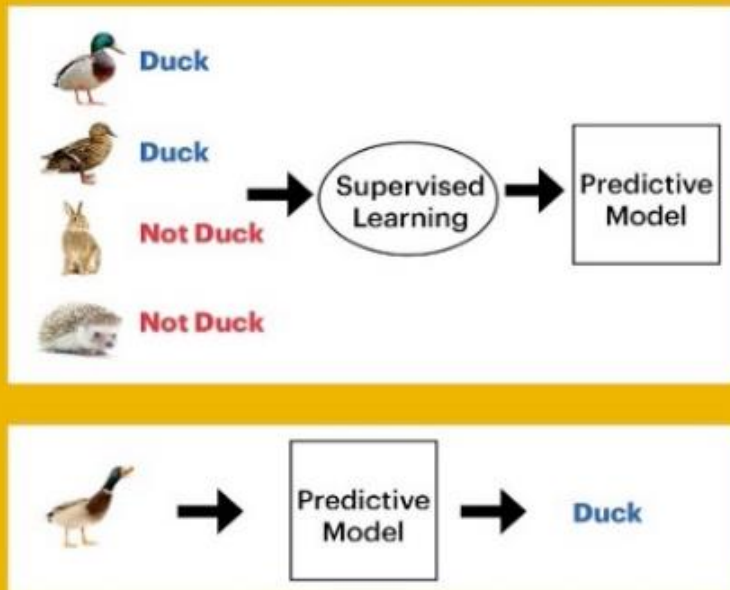
Para cada grupo de algoritmo (supervisionado, não supervisionado) existe tipos de **tarefas que os algoritmos realizam**. As três principais tarefas são Classificação, Regressão e Agrupamento.



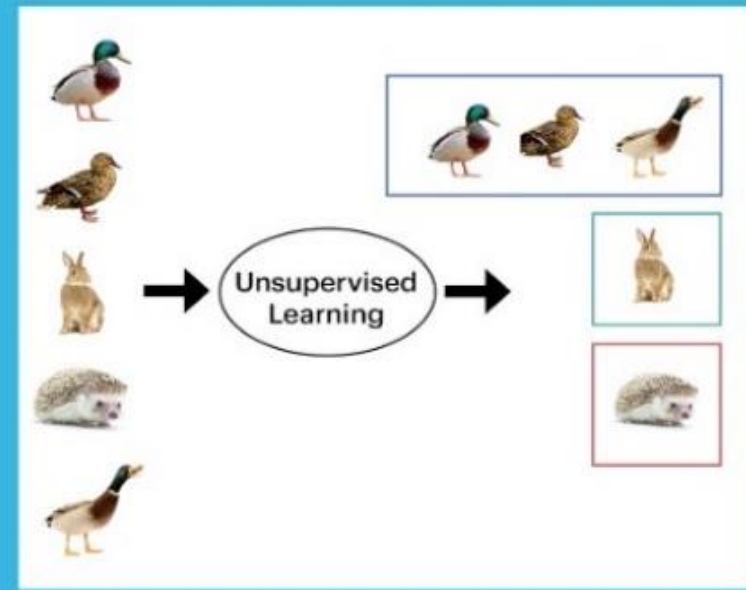
Aprendizado de Máquina

No aprendizado **supervisionado** temos rótulos para cada entrada de dado. No **não supervisionado**, não fornecemos nenhuma informação (rótulo) para o agrupamento.

Supervised Learning (Classification Algorithm)

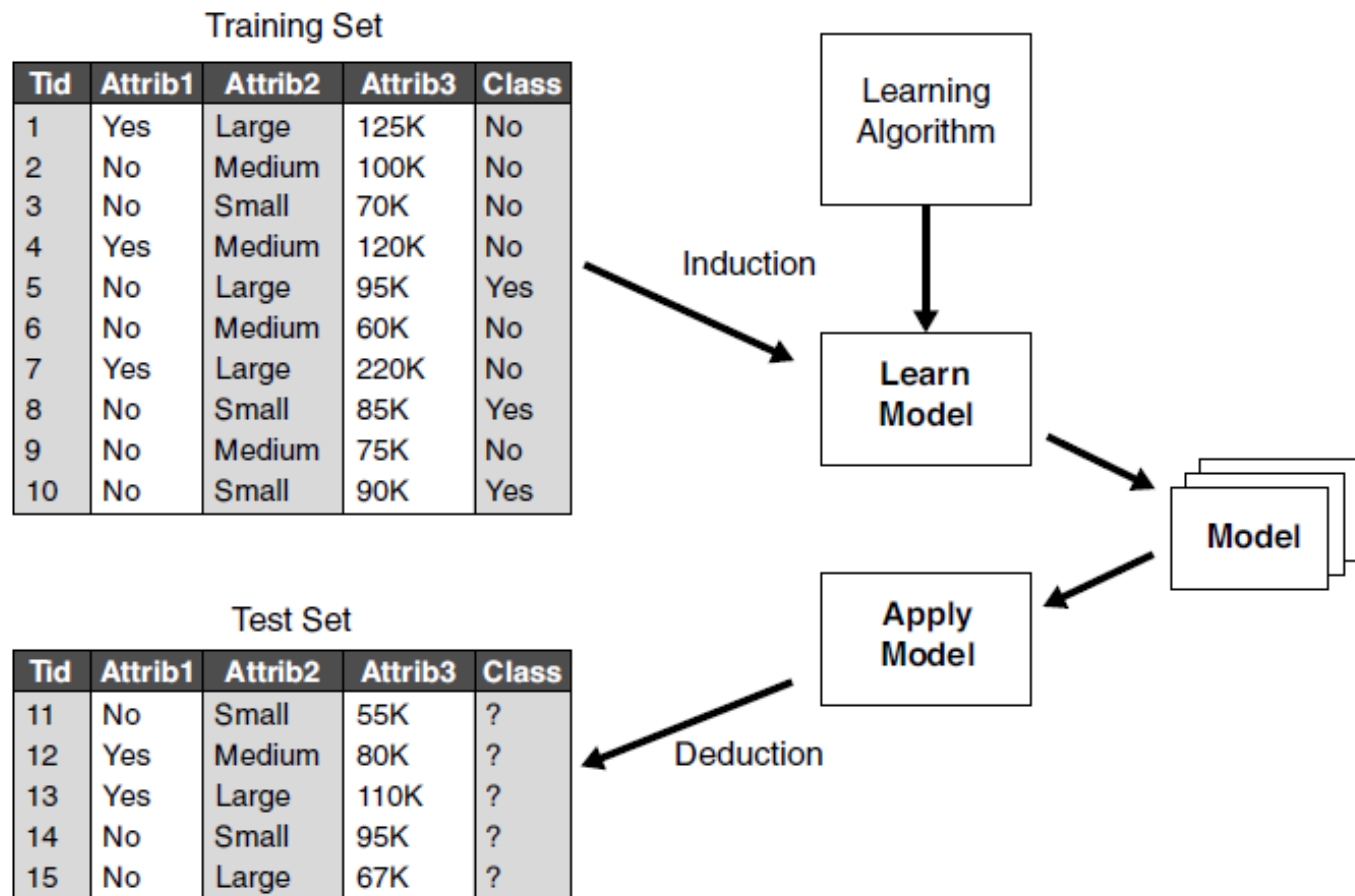


Unsupervised Learning (Clustering Algorithm)



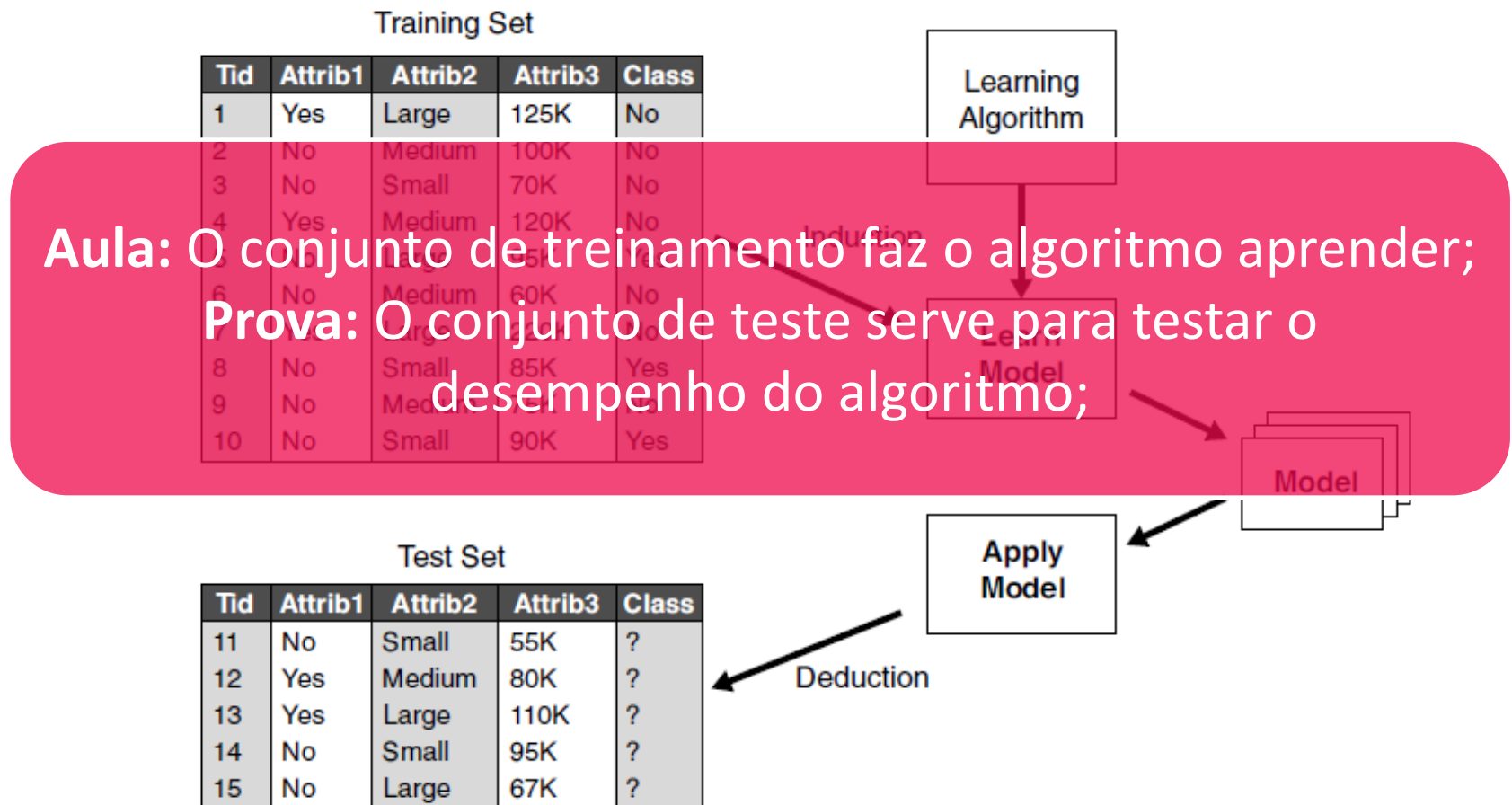
Aprendizado de Máquina - Supervisionado

No **aprendizado supervisionado** iremos separar nosso conjuntos de dados em dois grupos, o **conjunto de treinamento** e o **conjunto de teste**.



Aprendizado de Máquina - Supervisionado

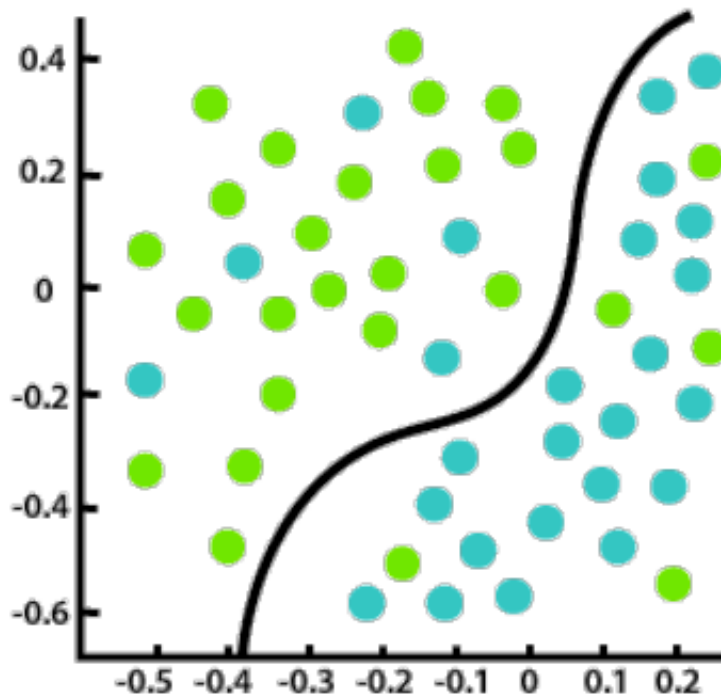
No **aprendizado supervisionado** iremos separar nosso conjuntos de dados em dois grupos, o **conjunto de treinamento** e o **conjunto de teste**.



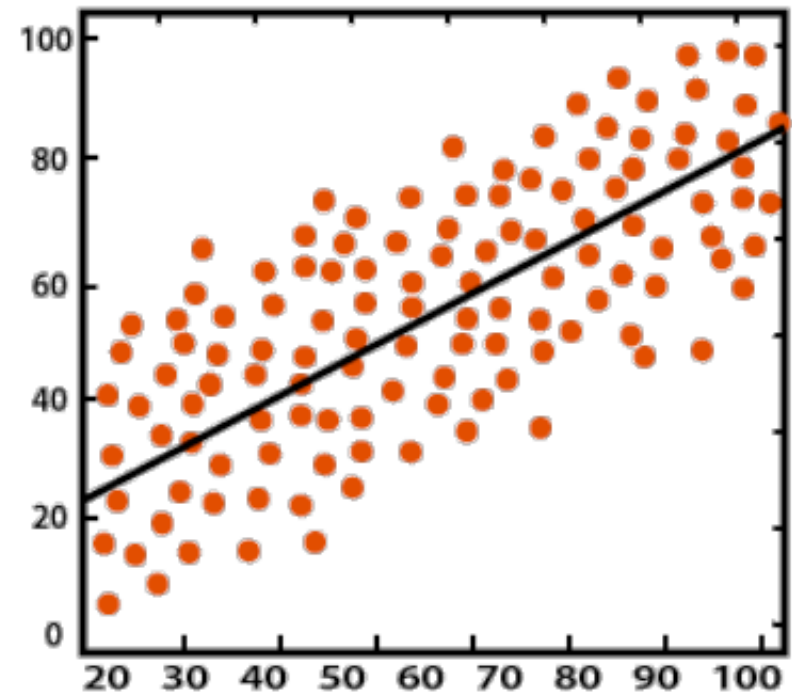
Aprendizado de Máquina - Supervisionado

Dentro do aprendizado supervisionado temos as tarefas de Regressão e Classificação:

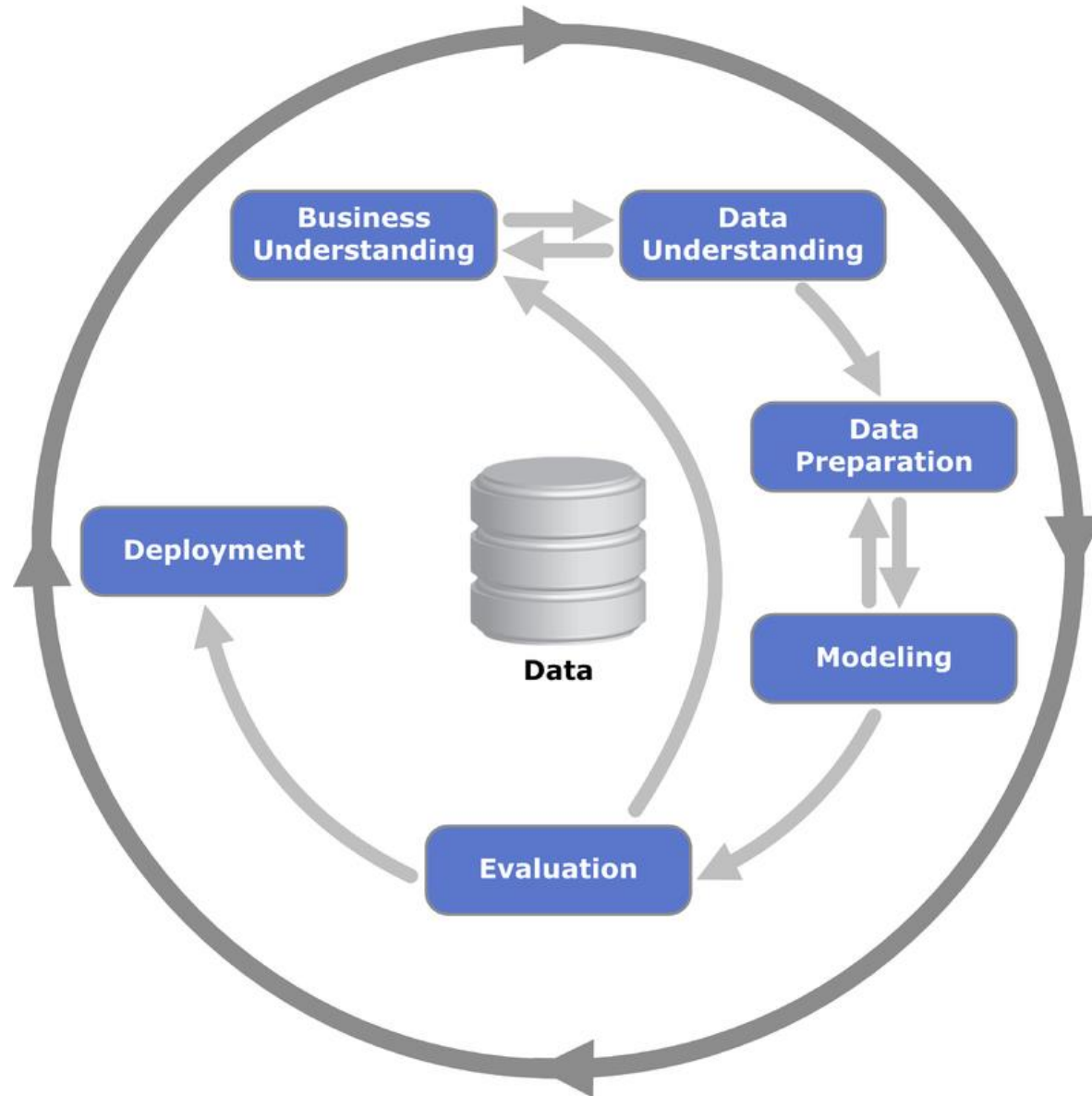
Classificação: predição de classe



Regressão: predição de valor

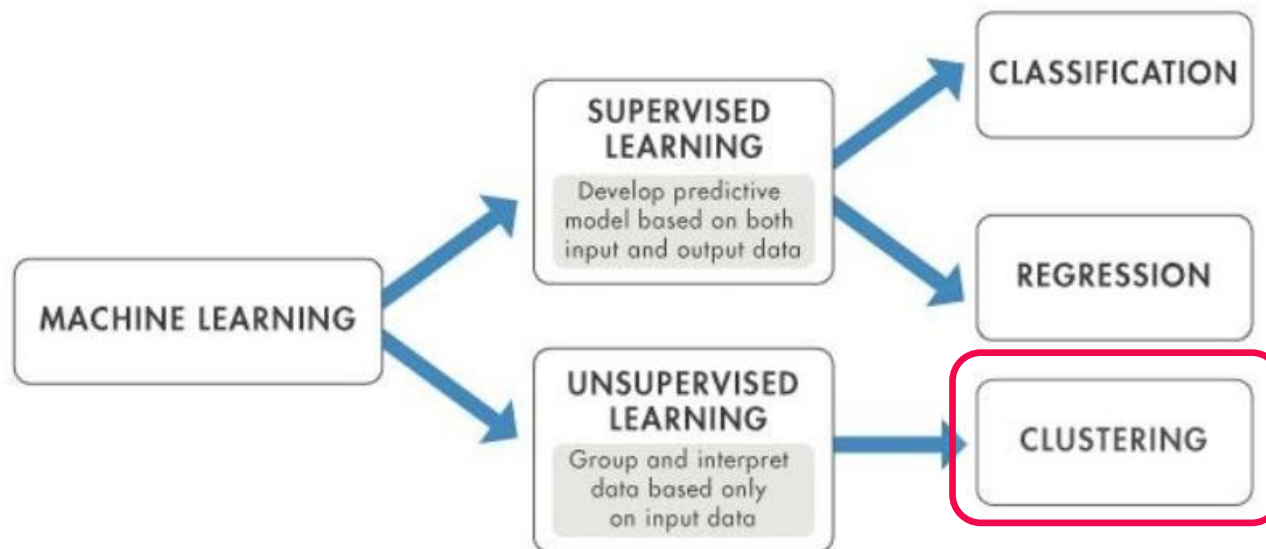


Fluxo de trabalho em ML



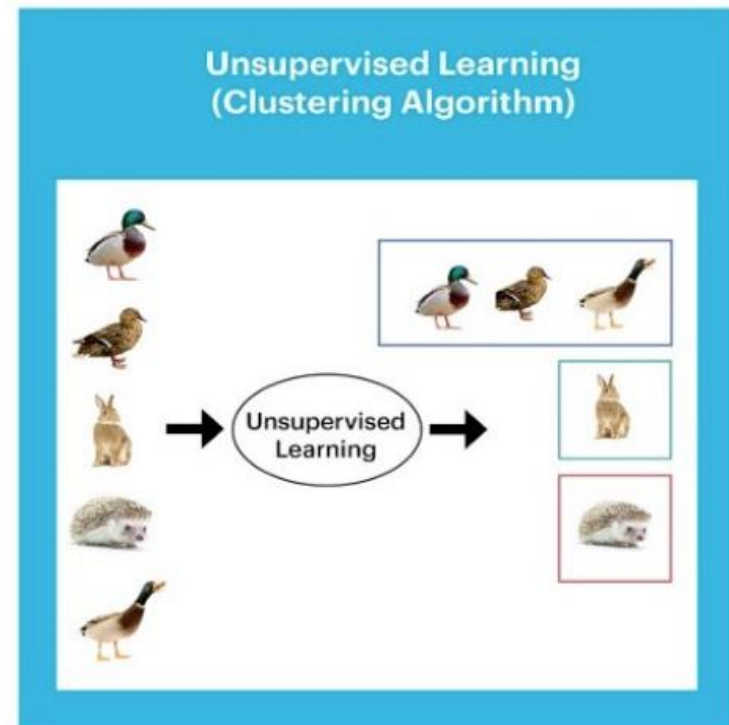
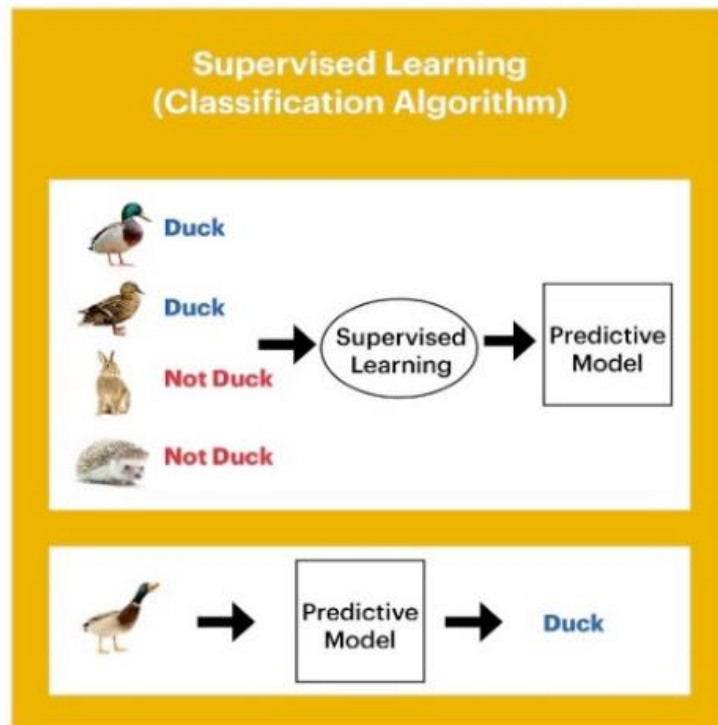
Aprendizado não supervisionado

- No aprendizado não supervisionado **não sabemos inicialmente os rótulos das classes.**
- Com essas técnicas gostaríamos de resolver problemas de:
 - **Agrupamento** (ou clusterização)
 - **Regras de Associação**
 - **Redução de Dimensionalidade**
 - **Detecção de Outliers**



Aprendizado não supervisionado

- Na **classificação** nosso atributo alvo é um objeto (string) previamente conhecido (dado rotulado);
- Na **clusterização** nosso atributo alvo é um número (int) previamente desconhecido (dado não rotulado);



Agrupamento

Agrupamento é uma
técnica não
supervisionada!

$$f(x_1, x_2, \dots, x_N)$$

Atributos
descritivos

Grupo alvo
desconhecido

Índice da linha	x1	x2	...	xn	\hat{y}
1	548.4	-9789	...	0.4875	?
2	689.4	-10235		-0.358	?
3	3154.8	-1031858		-0.1458	?
...	
k	803.54	-20000		1.054	?

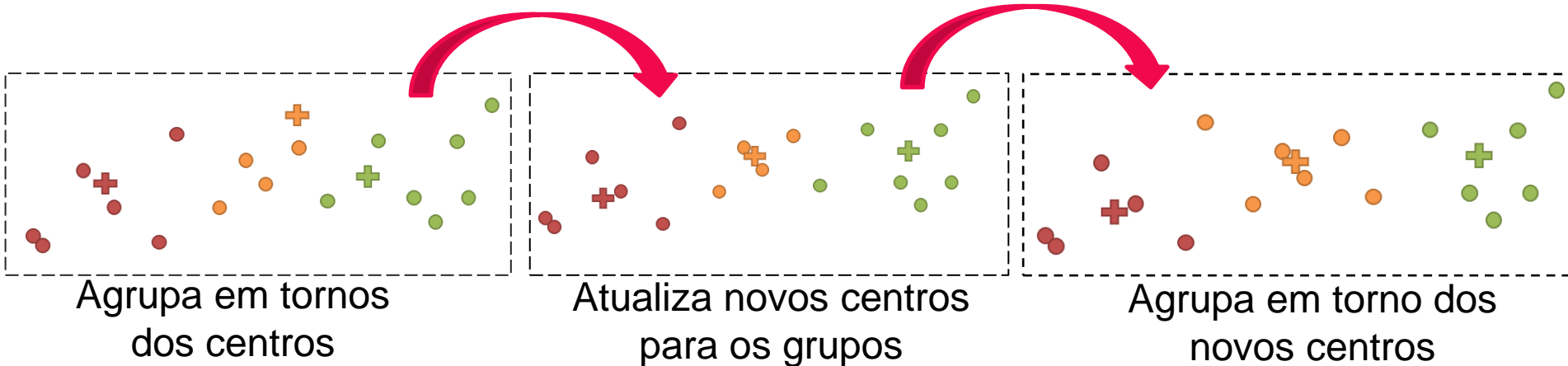
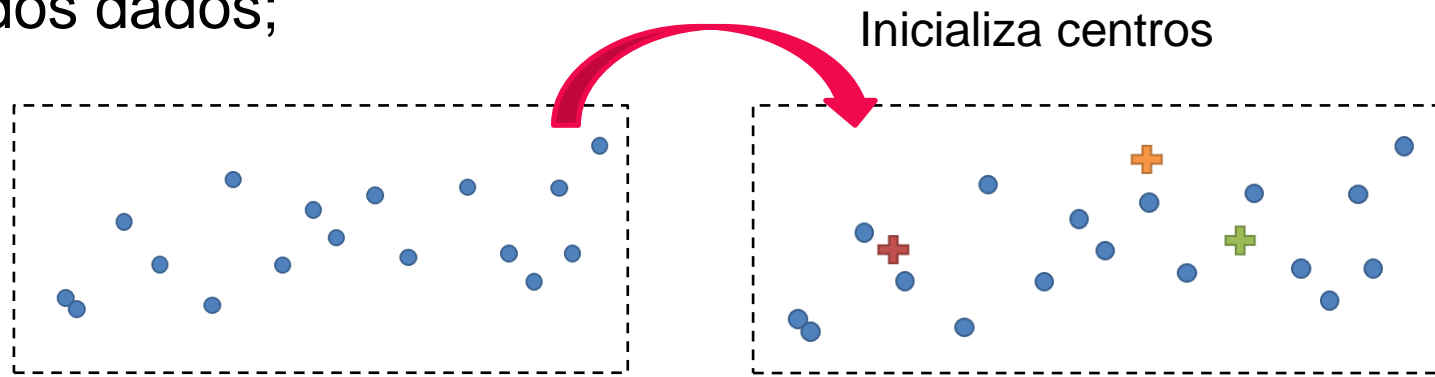
Queremos
criar essa
coluna!



ALGORITMOS DE AGRUPAMENTO (*CLUSTERING*)

Baseado em Partição – k-Means

- O k-Means é um dos métodos mais antigos (referências originais datam de 1956, 1965 e 1967) e mais utilizados;
- Ele é simples e intuitivo, baseado na ideia de se quebrar o espaço multidimensional em partições a partir do centroide dos dados;



O pseudocódigo do k-Means pode ser sumarizado como:

1. Escolher aleatoriamente k centros para os clusters;
2. Atribuir cada objeto para o cluster de centro mais próximo, segundo alguma métrica de distância (ex: euclidiana);
3. Mover cada centro para a média (centroide) dos objetos do cluster correspondente;
4. Repetir os passos 2 e 3 até que algum critério de convergência seja atendido (ex: número máximo de interação, limiar mínimo de mudança nos centroides).

k-Means – Algoritmo / Métrica

- Precisamos usar uma métrica de distância entre os centroides e os pontos de dados;
- Podemos usar diferentes métricas. A mais comum é a distância Euclidiana:

$$d(A, B) = \sqrt{\sum_{i=1}^n (A_{x_i} - B_{x_i})^2}$$

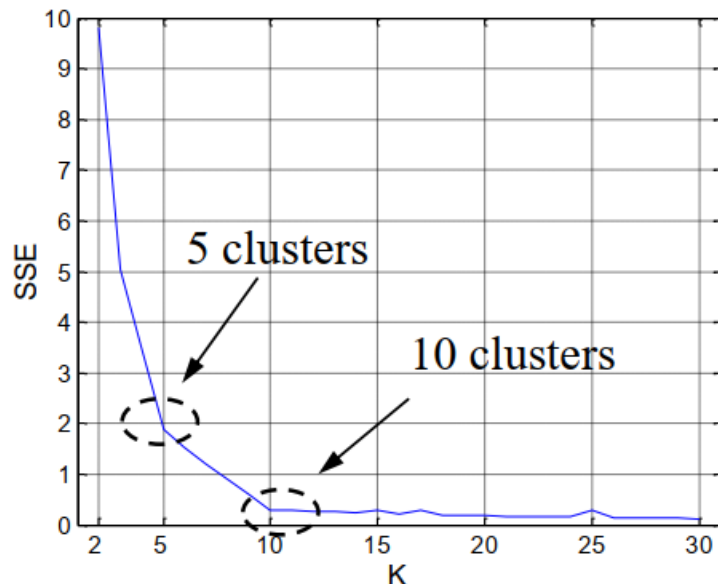
identifier	class name	args	distance function
"euclidean"	EuclideanDistance	•	<code>sqrt(sum((x - y)^2))</code>
"manhattan"	ManhattanDistance	•	<code>sum(x - y)</code>

Outras métricas:

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html>

k-Means – Algoritmo / Hiperparâmetro

- O k-Means tem o hiperparâmetro k que é o número de grupos (clusters);
- Como saber qual é o melhor número de k?
- Podemos usar a Soma dos Erros Quadráticos (SSE) em relação ao centroide para encontrar o “cotovelo” da curva de otimização:



$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist(c_i, x)^2$$

- *dist* é a distância euclidiana;
- c_i é o centro do i-ésimo agrupamento;
- x são os dados pertencentes ao i-ésimo agrupamento.

k-Means – Vantagens e Desvantagens

Vantagens:

- Implementação simplificada.
- Facilidade em lidar com qualquer medida de similaridade e por consequência, qualquer tipo de atributo.

Desvantagens:

- Dificuldade na definição do valor de “k”.
- Suscetível a outliers e a ausência de normalização.



MÉTRICAS DE DESEMPENHO DE AGRUPAMENTO

Silhouette

O *Silhouette Score* é uma métrica que avalia o “formato” dos clusters obtidos.

Ele é obtido calculando a distância média entre um dado de um agrupamento com todos os outros dados do mesmo cluster (a) e com a média desse mesmo dado com todos os dados do agrupamento mais próximo.

Essa métrica é definida como:

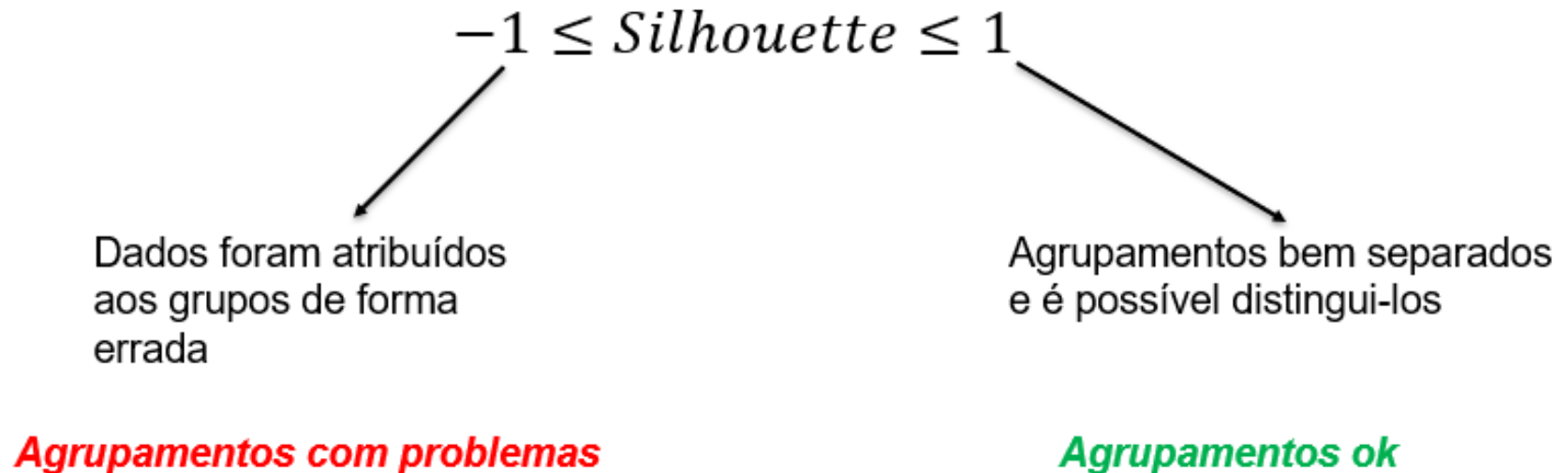
$$Silhouette(i) = \frac{b_i - a_i}{\max(a_i, b_i)}$$



a: distância média entre i e todos os pontos de seu agrupamento

b: distância entre i e o agrupamento mais próximo

A média do *Silhouette Score* de todos os dados nos define o quão bom é o nosso agrupamento:



Copyright © **2023** Prof. Henrique Ferreira dos Santos

Colaboração e adaptação: Prof. André Tritiack

Todos direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem o consentimento formal, por escrito, do Professor (autor).