



# FUNDAMENTOS E TÉCNICAS EM CIÊNCIAS DE DADOS

PROF. JOSENALDE OLIVEIRA

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

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

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - UFRN

# UM POUCO SOBRE APRENDIZAGEM DE MÁQUINA

*Machine Learning - ML*



Ser humano estabelece **conexões** para lidar com coisas novas

Similaridade pode ser óbvio para o humano, mas não para computadores

Máquinas operam sobre **tarefas frequentes**, com alto volume e velocidade

Desafio: máquinas serem ensinadas e depois aprenderem 'sozinhas'

# UM POUCO SOBRE **APRENDIZAGEM DE MÁQUINA**

*Machine Learning - ML*

**Como uma criança aprende que ambos são dinossauros?**



**E aqui?**





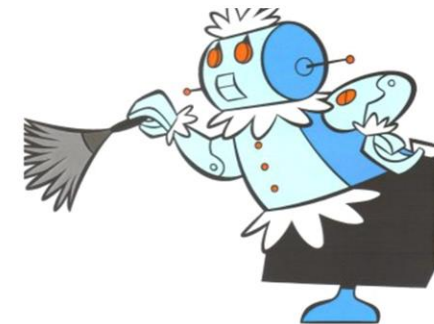
# UM POUCO SOBRE **INTELIGÊNCIA** ARTIFICIAL



The Rebellious Robots, R.U.R (1920 – Karel Capek)

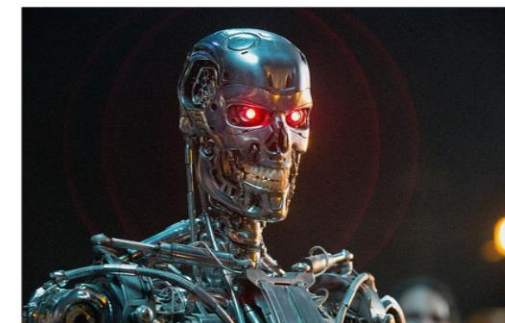


Eu, Robô (2004)  
Isaac Asimov (1950)



Rosie (Rosey) – Jetsons  
62-63 / 84-87

- **Inteligência artificial como meio e não como fim**
- **Inteligência artificial ajuda a encontrar respostas, mas ainda está longe de saber fazer as perguntas!**



Exterminador T-800 (1984)

# UM POUCO SOBRE INTELIGÊNCIA ARTIFICIAL

## QUE TIPOS DE IA EXISTEM?

Estreita Narrow A.I.

General A.I.

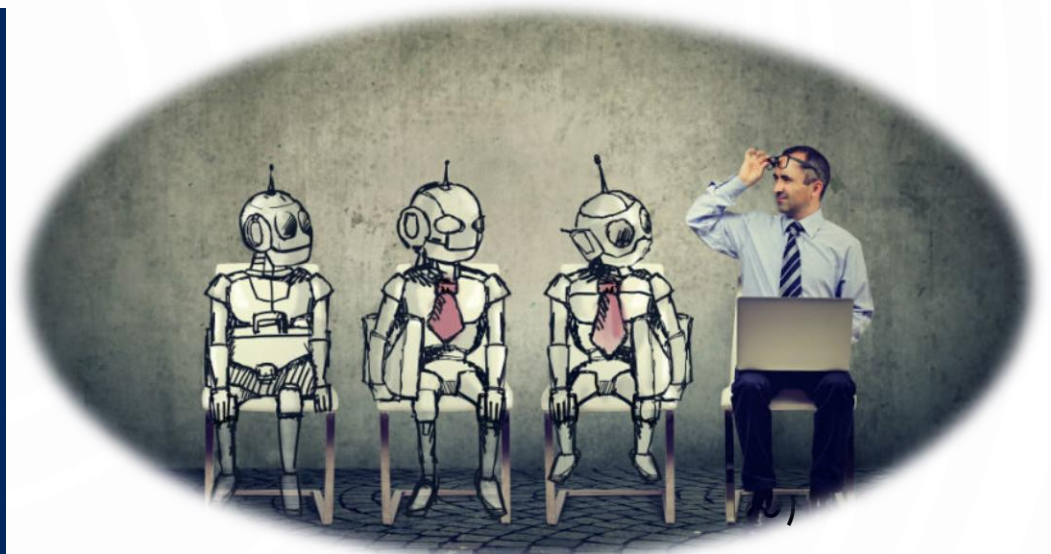
Hoje

Sistemas que pensam racionalmente	1	Sistemas que pensam como os humanos	3
Sistemas que agem racionalmente	2	Sistemas que agem como os humanos	4

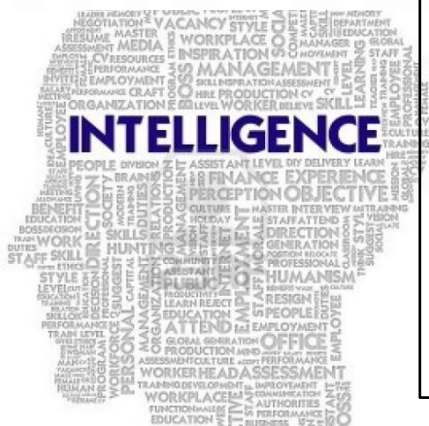
Futuro

Regras/Associações

Cognição/Emocional



Será que isto será possível?



- Prover as máquinas com capacidade de **tomarem decisões inteligentes**
  - (certa autonomia)
- Neste contexto, inteligência como capacidade de tomar a melhor decisão possível dada a informação disponível, com a capacidade de se adaptar a novas situações



# UM POUCO SOBRE **INTELIGÊNCIA ARTIFICIAL**

- Comparemos a autonomia num chão de fábrica, com regras Se...Então
- Com a autonomia necessária a um veículo autônomo



**Condicionais: Se obstáculo à frente...pare**  
**Se prateleira fazia...vá para próxima**  
**Ambiente controlado**

**Quais as necessidades para um carro autônomo?**  
**Desenvolvimento tradicional baseado em regras?**

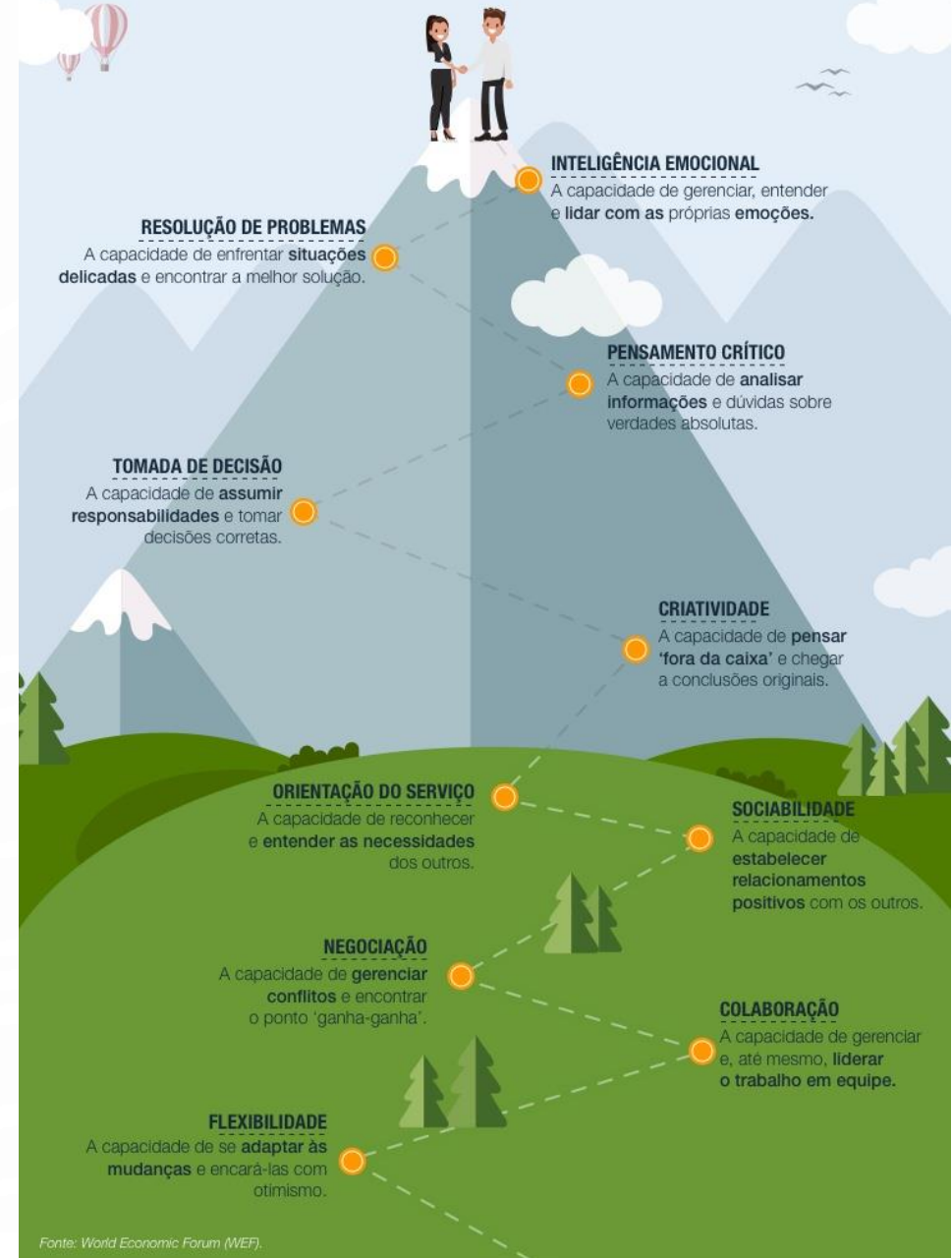
- Necessidades de reconhecimento/classificação
- **Predição**



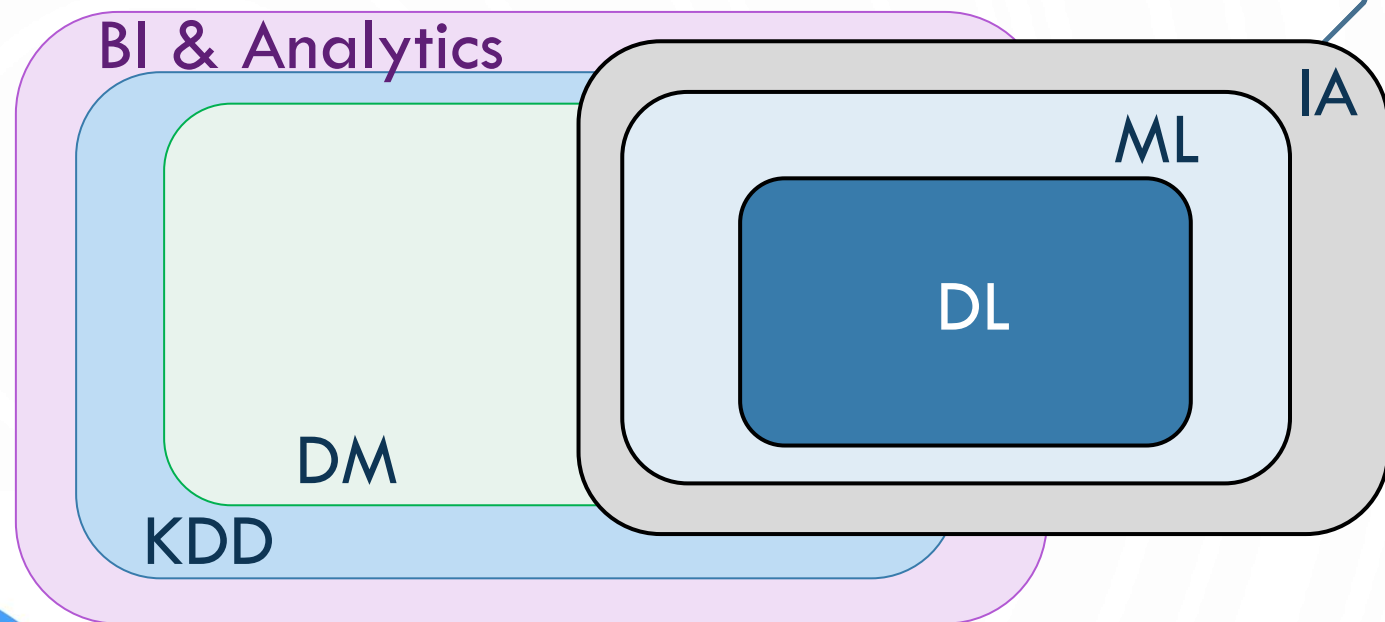
# UMA MÁQUINA COM **SOFT SKILLS**?

- Embora a computação otimize e amplie a adoção de análise (**analytics**) por meio de variadas técnicas de inferência e predição...
- O conceito aqui está na pessoa com capacidade analítica (negócios), de modo a, com base na manipulação dos dados obter insights que podem trazer algum tipo de vantagem competitiva – **competência**
- A modelagem pode ser automatizada, mas as pessoas definem os problemas

## As 'soft skills' mais procuradas



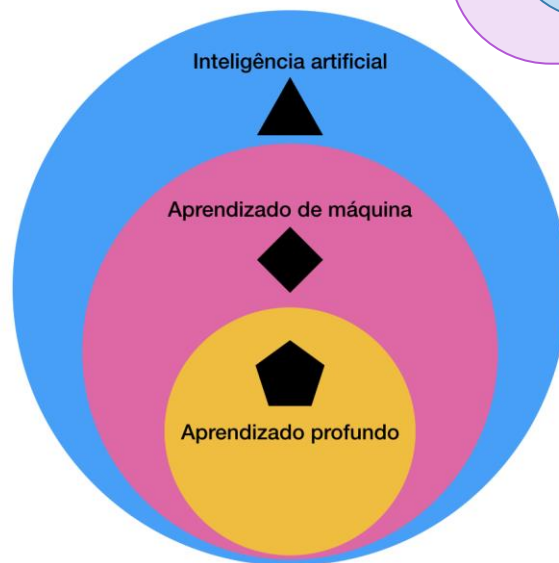
# SITUANDO...



Técnicas que habilitam os computadores a **resolver problemas** que seres considerados inteligentes resolvem. ▲

Subconjunto de técnicas da IA que usam métodos para permitir que as máquinas se aprimorem **a partir de experiências** (dados, por exemplo). ◆

Subconjunto de AM **que combina visão computacional e sistemas de aprendizagem** conjuntamente. ▴



BI: Business Intelligence

DM: Data Mining

KDD: Knowledge Discovery in Databases

DL: Deep Learning (Aprendizado Profundo)

ML: Machine Learning

IA: Inteligência Artificial



# KDD E FRAMEWORKS PARA MANIPULAÇÃO DE DADOS

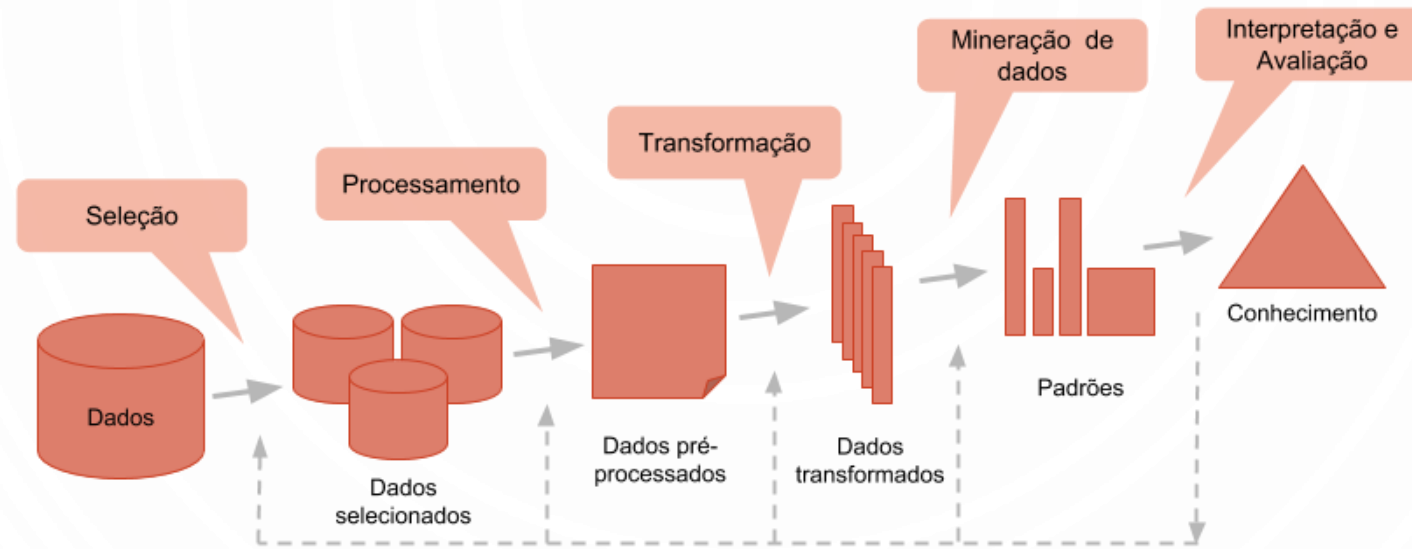
- ciclo de vida do dado que temos estudado é baseado em frameworks consolidados
- O termo BI remonta à 1958 no artigo de Luhn, H.P.

LUHN, H.P. **A Business Intelligence System**. *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 314-319, Oct. 1958. [doi: 10.1147/rd.24.0314](https://doi.org/10.1147/rd.24.0314)

Abstract: an automatic system is being developed to disseminate information to the various sections of any industrial, scientific or government organization. This intelligence system will utilize **data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the "action points" in an organization**. Both incoming and internally generated documents are automatically abstracted, characterized by **a word pattern**, and sent automatically to appropriate action points. This paper shows the flexibility of such a system in identifying known information, **in finding who needs to know it** and in disseminating it efficiently either in abstract form or as a complete document.

- O KDD (1996) pode ser visto como o framework geral para encontrar informação útil em massas de dados, e o DM é uma etapa do KDD, aplicando algoritmos para extrair padrões

# KDD E FRAMEWORKS PARA MANIPULAÇÃO DE DADOS



- Fazer o processo de forma estruturada é fundamental, seguindo **etapas, método**
- KDD é o mais antigo, mais conhecido, não foca em negócios ou modelos, mas na descoberta de padrões a partir dos dados
- CRISP-DM é uma das mais usadas e considerada mais completa – genérica e com maior link com o negócio alvo

# CRISP-DM

Processo Padrão Inter-Indústrias para Mineração de Dados (1996)  
(Daimler Chrysler, SPSS e NCR4)

➤ **Entendimento do negócio:** identificação do problema a ser resolvido – *business-oriented*

➤ *Artefatos: background (contexto, como projeto resolve), objetivos e critério de sucesso (métricas)*

➤ **Entendimento dos dados:** captura (coleta), descrição (o que se entende dos dados), exploração, **qualidade**

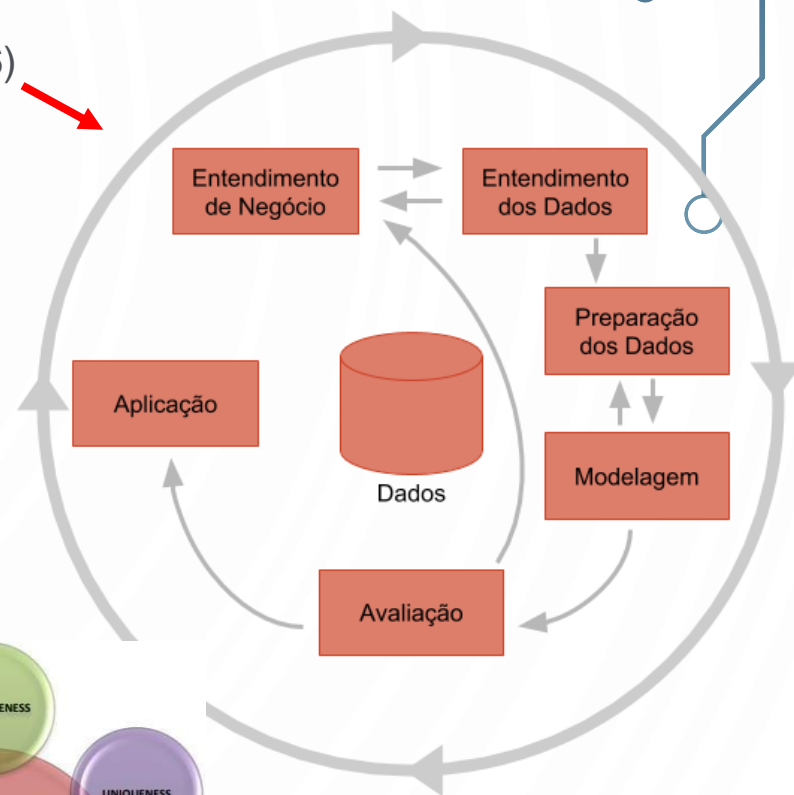
➤ **Preparação dos dados:** criação do *dataset* a partir de *raw data*, seleção, limpeza, transformação, integração – entrada da modelagem

➤ **Modelagem:** que técnica usar? Projetar testes, construir e avaliar modelos

➤ **Avaliar:** atende aos objetivos e vai para produção?

➤ **Aplicação:** é aqui que a empresa faz uso de toda a análise desenvolvida.

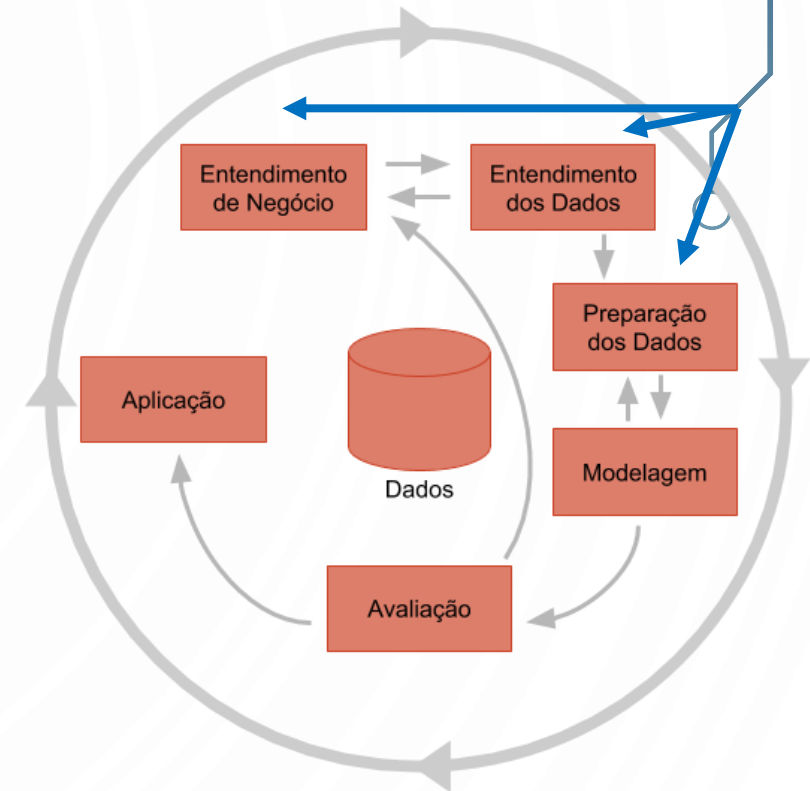
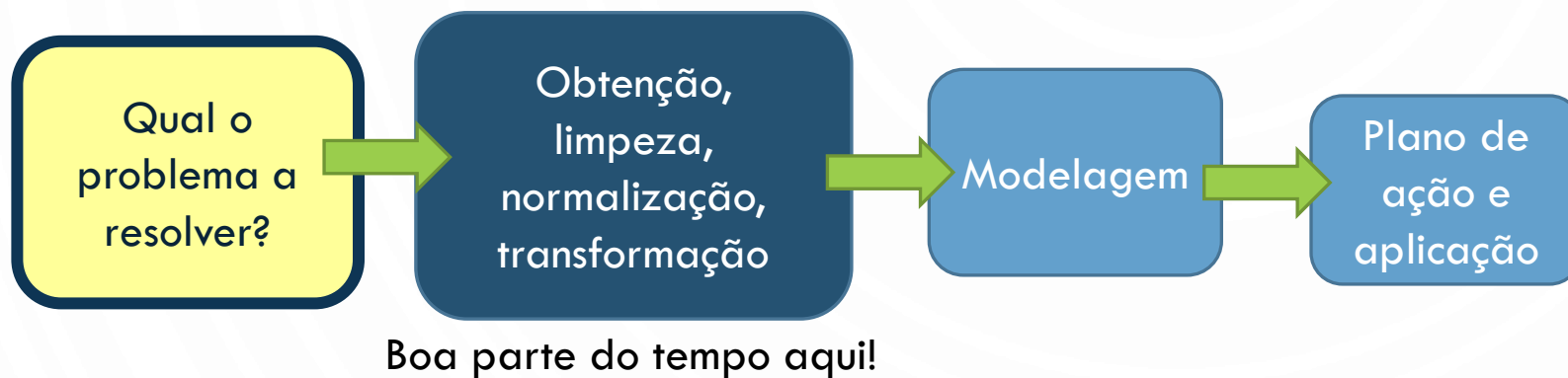
Apresentação dos resultados da modelagem para a tomada de decisão



DAMA Framework



# O PESO DAS FASES...



# BIG PICTURE DO PROJETO DE DATA SCIENCE

Identificação de disponibilidade, fonte de coleta e tipos de dados

Mais CONCRETO	Mais FÁCIL de obter		Mais SUBJETIVO
	Mais DIFÍCIL de obter		
<ul style="list-style-type: none"><li>• Idade</li><li>• Sexo</li><li>• Est. Civil</li><li>• Salário</li><li>• Tempo na organização</li><li>• Nível de qualificação</li><li>• Cargo</li><li>• Localidade de trabalho</li></ul>	<ul style="list-style-type: none"><li>• Nível de risco</li><li>• Horas extras</li><li>• Absenteísmo</li><li>• Dependentes</li><li>• Cargo</li><li>• Regime de contratação</li><li>• Salário Atual x inicial</li><li>• Salário Atual x salário há 1 ano</li></ul>	<ul style="list-style-type: none"><li>• Resultado feedback atual x anterior</li><li>• Resultado de pesquisa de clima</li></ul>	<div>CUIDADOS ESPECIAIS COM OS ATRIBUTOS ABAIXO: Raça, Cor, Opção sexual Religião, Política</div> <ul style="list-style-type: none"><li>• Nível de pressão</li><li>• Variedade atividades</li><li>• Indicação interna</li></ul>
<ul style="list-style-type: none"><li>• Renda familiar</li><li>• Resultado de feedack do funcionário</li><li>• Resultado de feedack do gestor</li><li>• Salário atual x salário emprego anterior</li><li>• Salário atual x salário médio do mercado</li></ul>	<ul style="list-style-type: none"><li>• Família em outra localidade</li><li>• Tempo permanência médio últimos 3 empregos</li><li>• Distância do trabalho</li><li>• Tempo desde a última promoção</li></ul>	<ul style="list-style-type: none"><li>• Nível de satisfação no trabalho</li><li>• Significância da atividade</li><li>• Nível de autonomia</li><li>• Viagens</li><li>• Amigos/Familiar es na organização</li></ul>	

Exemplo baseado em Canvas – Fabiano Castello@IA

## BIG PICTURE DO PROJETO DE DATA SCIENCE

Qual o problema a ser resolvido?

1

Fontes dos Dados

2

Target

4

Produtos de Dados

5

Direcionadores PITCH

- Qual a oportunidade?
- O que está buscando?
- Quais vantagens do produto de dados?

6

Principais Atributos (variáveis)

3

Mais CONCRETO

Mais SUBJETIVO

Mais FÁCIL de obter

Mais DIFÍCIL de obter

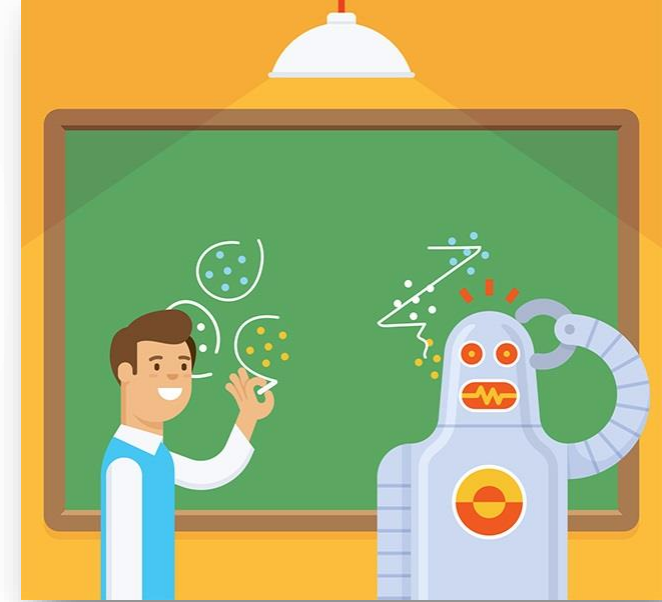
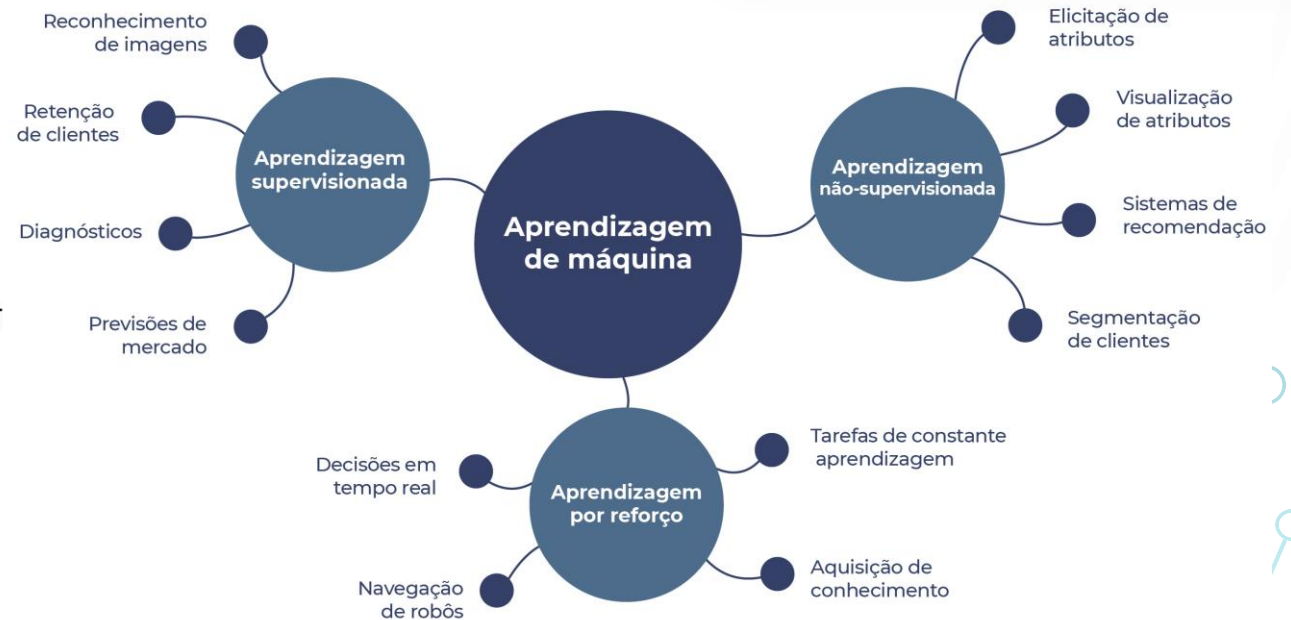
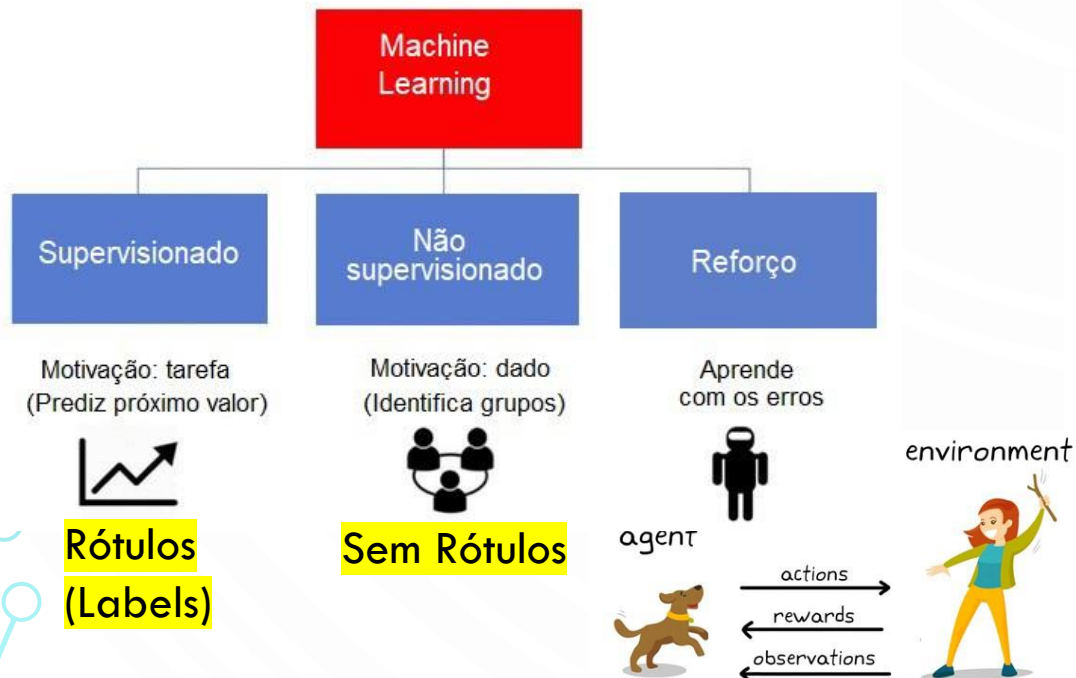
14



# ASPECTOS DE MACHINE LEARNING

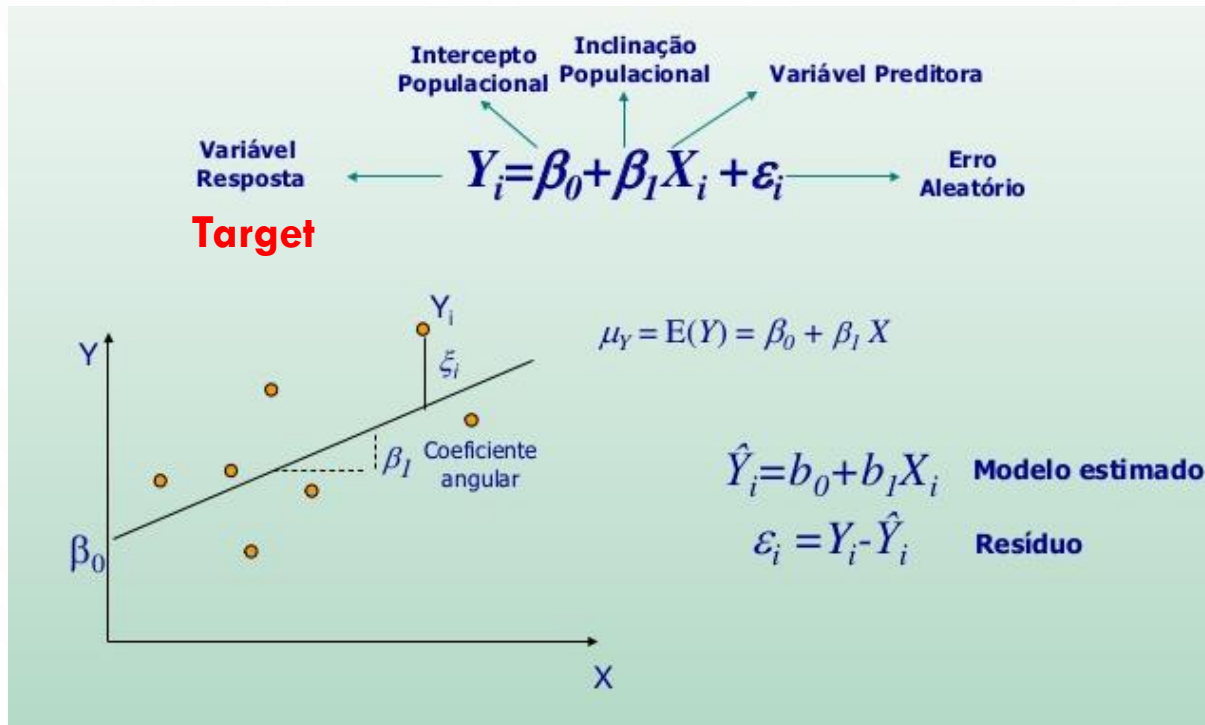
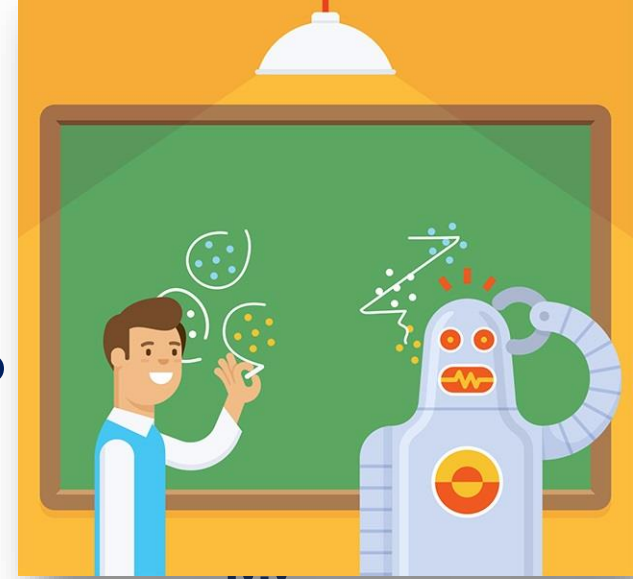
- O foco da ML está em **ANÁLISE PREDITIVA**
- Ferramentas/Técnicas/Algoritmos que fazem 'predições', de forma mais rápida, barata, assertiva e que captam bem o 'novo'

## Tipos de Machine Learning



# ASPECTOS DE MACHINE LEARNING

- Um bom projeto de ML tem boa capacidade preditiva
- Acurácia nas decisões (acertos)
- Performance preditiva! Nem sempre a interpretação do processo é simples (pois decisões envolvem processos complexos)
- Em Inferência (regressão por exemplo), a relação entre as variáveis é melhor interpretável, mas usualmente pior performance preditiva

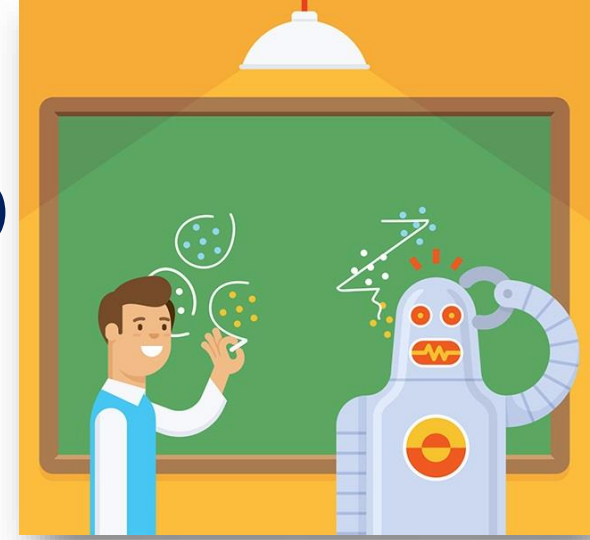
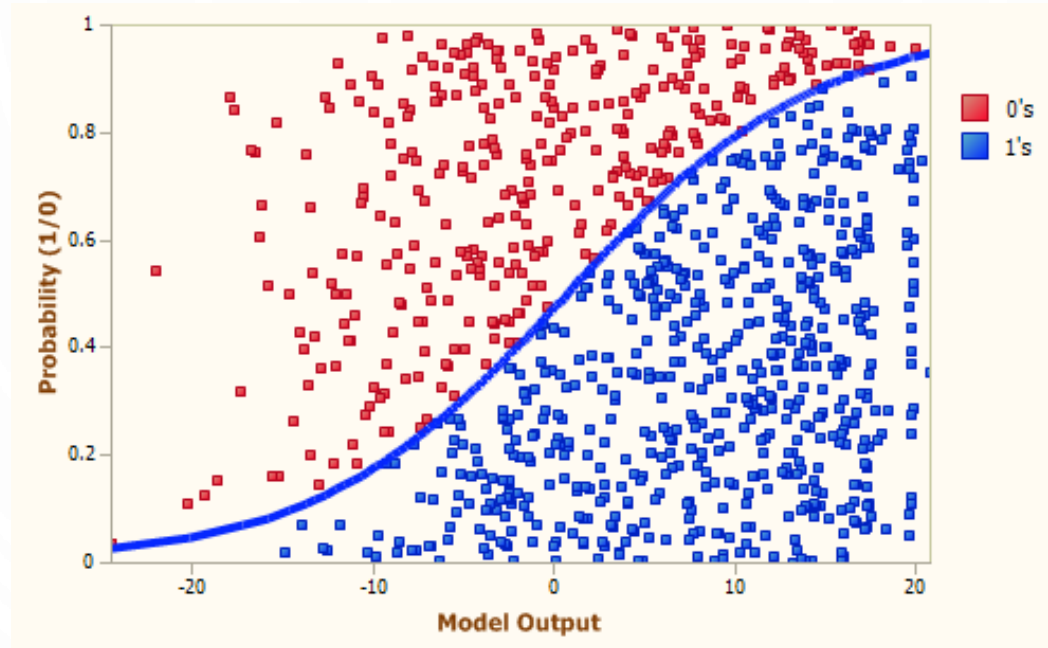


**Regressão/Inferência:** target quantitativo  
**Classificação:** target qualitativo  
**Regressão logística:** target quantitativo é probabilidade de um evento ocorrer como função de outros fatores.

# ASPECTOS DE MACHINE LEARNING

- **Regressão logística: target categórico e binário (s/n, 0/1, é/não é)**
  - **Sucesso/Fracasso, Desligado? (s/n)**

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$
$$\Rightarrow P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



## Diferenças

	Linear	Logística
Reta	Reta	Curva - S
Variável Dependente	Continua	Categórica
Interpretação	$\hat{y}$	$\ln\left(\frac{p}{1+p}\right)$

[http://neilsoncrepalde.github.io/2019-11-25-regressao\\_logistica\\_python/](http://neilsoncrepalde.github.io/2019-11-25-regressao_logistica_python/)

[https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

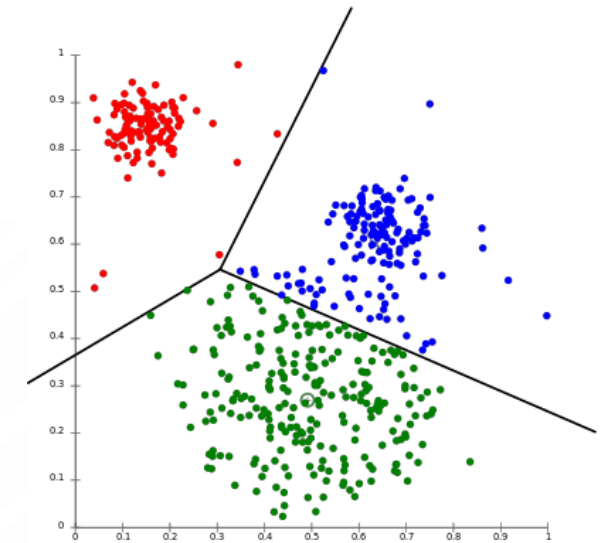
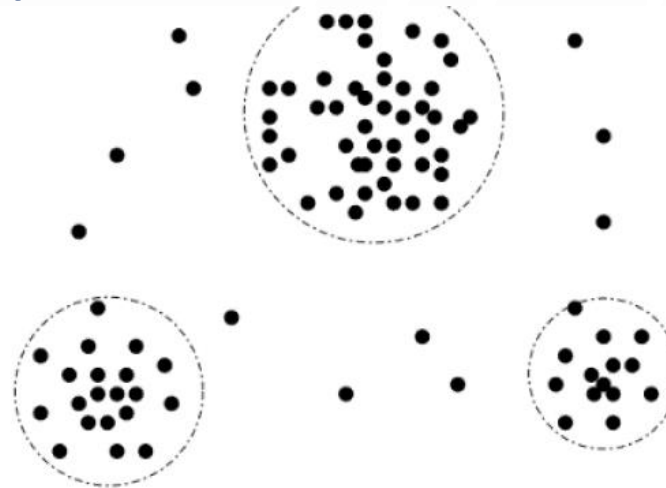
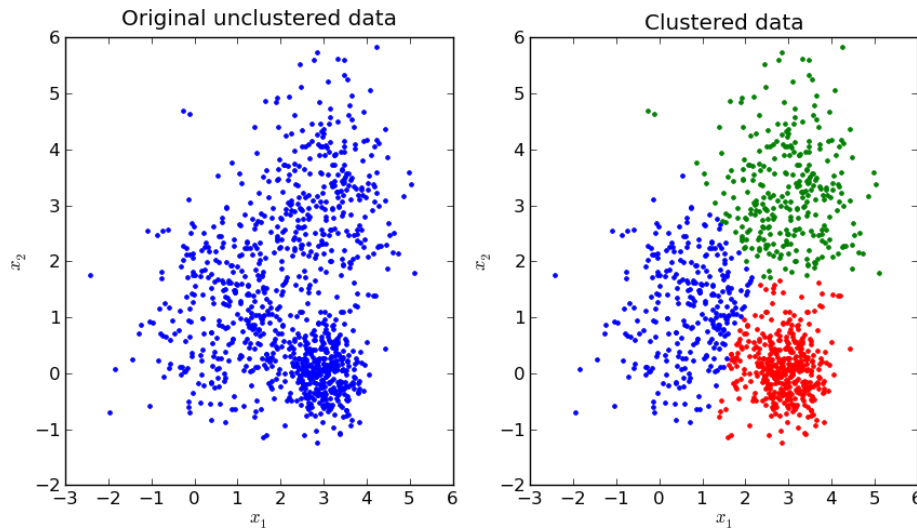


# ASPECTOS DE MACHINE LEARNING

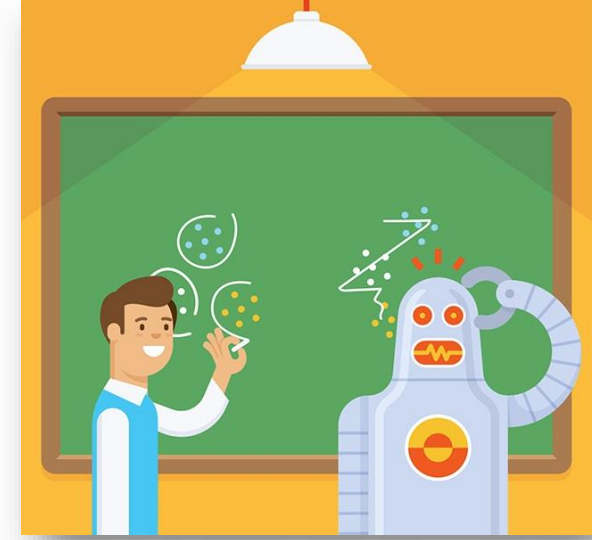
- **Agrupamentos: não supervisionado – maximizar semelhanças (minimizar distâncias) dentro do cluster e maximizar diferenças (maximizar distâncias) entre clusters**

**Método K-means é o mais usado**

**Mas existem vários outros métodos e variantes: Hierarquicos, aglomerativos, incremental etc.**

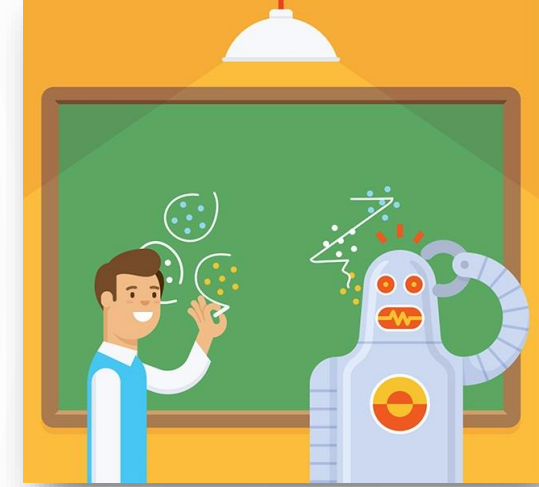


[https://github.com/josenalde/datascience/blob/main/notebooks/nb\\_kmeans1.ipynb](https://github.com/josenalde/datascience/blob/main/notebooks/nb_kmeans1.ipynb)

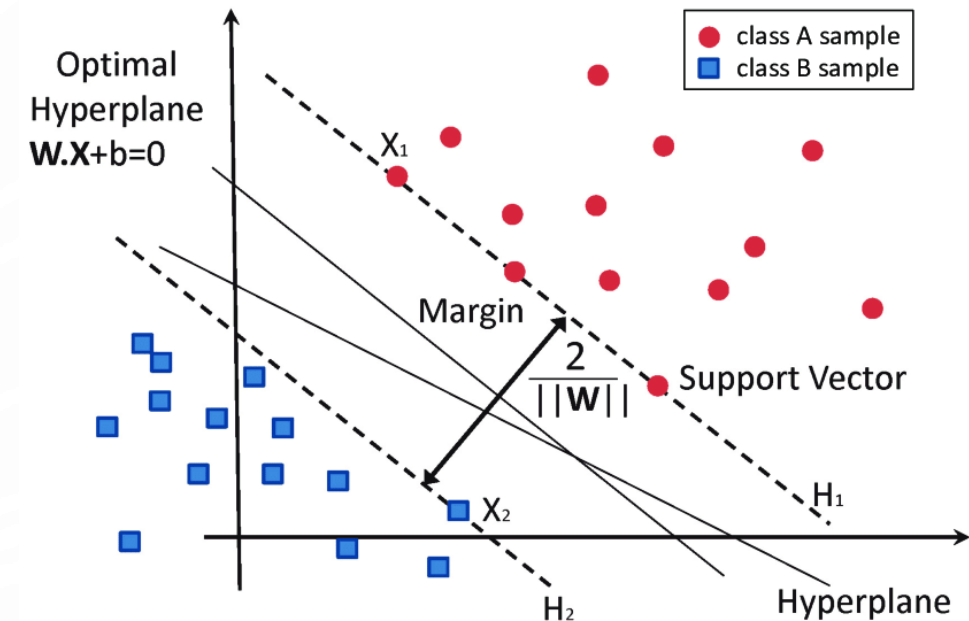
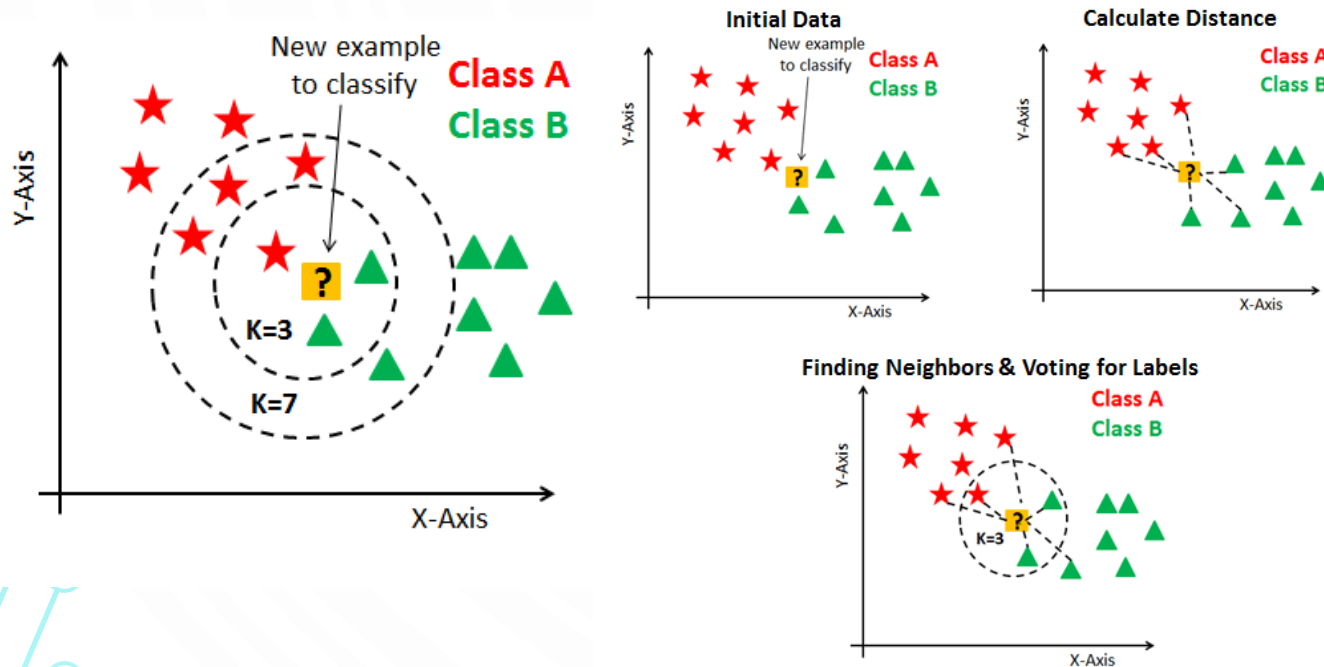


# ASPECTOS DE MACHINE LEARNING

- **Classificação: supervisionado** – conjuntos de treino / teste, métricas de avaliação mais bem definidas, por comparar com o *ground truth*, matriz de confusão etc.



**Métodos mais comuns: KNN, SVM, Árvore de Decisão, Redes Neurais**



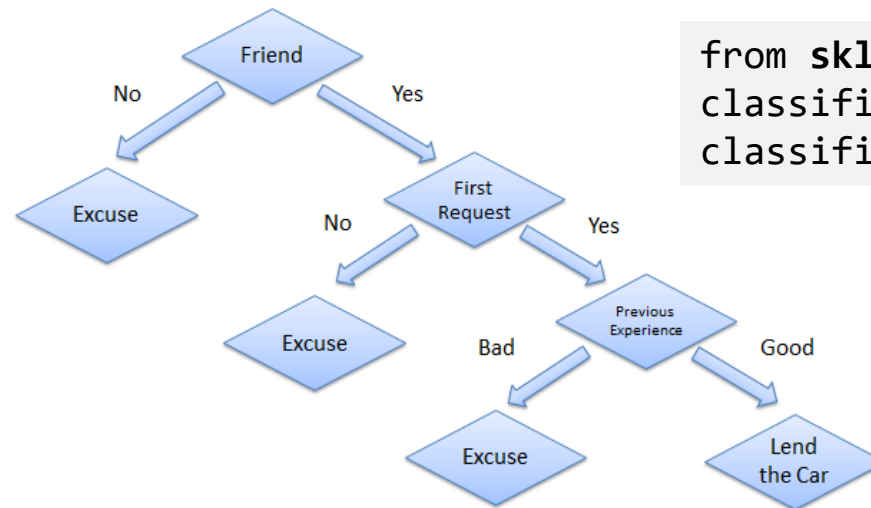
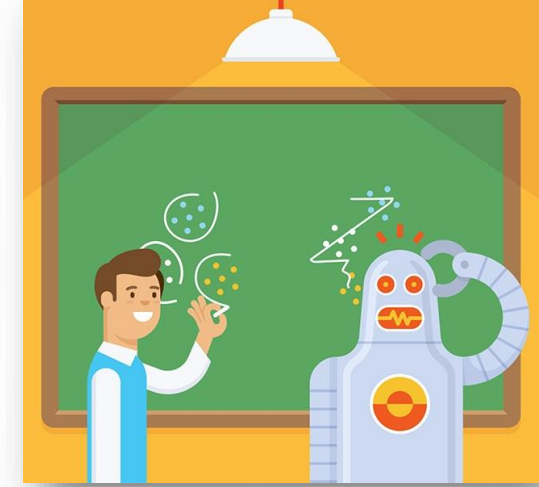
[https://github.com/josenalde/datascience/blob/main/notebooks/nb\\_knn1.ipynb](https://github.com/josenalde/datascience/blob/main/notebooks/nb_knn1.ipynb)

[https://github.com/josenalde/datascience/blob/main/notebooks/nb\\_svm1.ipynb](https://github.com/josenalde/datascience/blob/main/notebooks/nb_svm1.ipynb)

# ASPECTOS DE MACHINE LEARNING

- **Classificação: supervisionado** – conjuntos de treino / teste, métricas de avaliação mais bem definidas, por comparar com o *ground truth*, matriz de confusão etc.

Métodos mais comuns: KNN, SVM, Árvore de Decisão, Redes Neurais



```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier()
classifier.fit(X_train, y_train)
```

<https://www.datacamp.com/community/tutorials/decision-tree-classification-python>