

# Curso de Data Science

Charles Adriano dos Santos  
Rafael Roberto Dias



# Agenda

1 – Agenda

2 – R – Parte II

3 – Machine Learning

4 – Namorando Dados (SQL)

5 – Desafio Pessoal

6 – Desafio do Curso

7 – Bate Papo e Monitoria



# Manhã

---

## Horário Assunto

- 09:30 R – Parte II: Qualidade dos Dados e Variáveis Relevantes
- 11:00 Machine Learning: Teoria e Exemplos
- 12:30 Almoço

Tarde

---

## Horário Assunto

- 13:30 Namorando Dados com SQL: AgroXP
- 14:30 Desafio Pessoal: Extração de Características e ML
- 16:30 Desafio Curso: Aplicação da Solução
- 17:30 Bate Papo e Monitoria



# Nos Episódios Anteriores...



Profissão Data Science

Estatística & Ciência da Computação

Desafio Agro XP

- ETL
- Modelagem de Dados
- Banco de Dados
- Namorando Dados SQL
- Linguagem R / R Studio
- Linguagem Python

# Agenda

1 – Agenda

2 – R – Parte II

3 – Machine Learning

4 – Namorando Dados (SQL)

5 – Desafio Pessoal

6 – Desafio do Curso

7 – Bate Papo e Monitoria



# O Trabalho do Cientista de Dados > Desafio Curso

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção

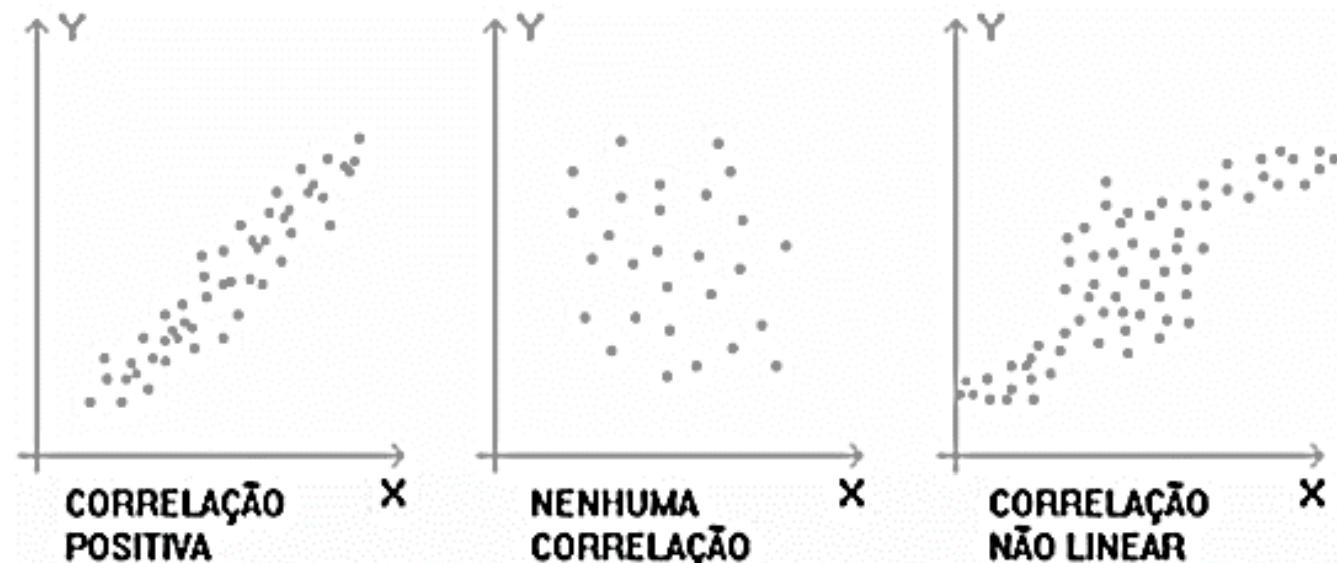


# Analizando a Qualidade dos Dados

- Objetivo nesta etapa do estudo é verificar a qualidade dos dados para entender quais tem potencial de fazer parte do estudo
- Foco maior em verificar se existem dados faltantes ou nulos que podem interferir no estudo
- Também aqui começa o entendimento de como cada variável ajuda a explicar o evento em estudo
- Aqui começam as **descobertas** do Cientista de Dados

# Variáveis Relevantes

- Objetivo nesta etapa do estudo é verificar a como as variáveis se relacionam entre si
  - **Foco maior aqui é entender a correlação entre as variáveis**
- O modelo ou a metodologia que será utilizada para responder as perguntas do estudo dependem dos achados desta etapa



# Agenda

1 – Agenda

2 – R – Parte II

3 – Machine Learning

4 – Namorando Dados (SQL)

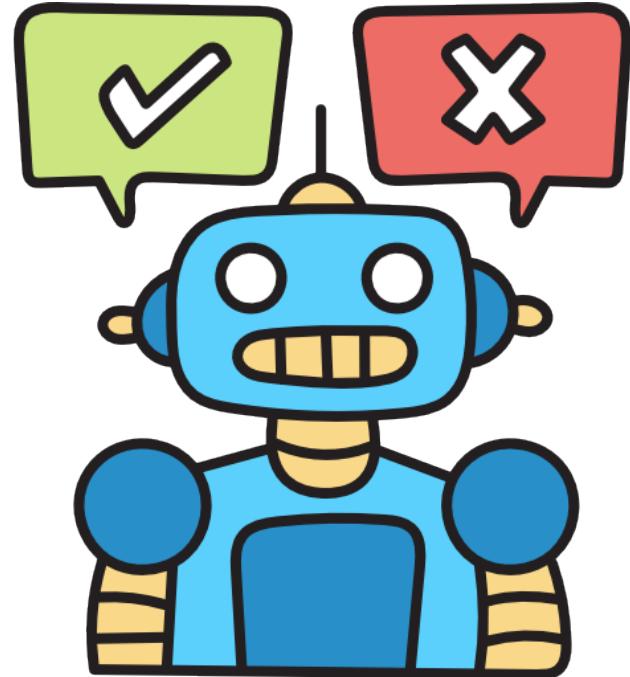
5 – Desafio Pessoal

6 – Desafio do Curso

7 – Bate Papo e Monitoria



# Machine Learning - Conceito

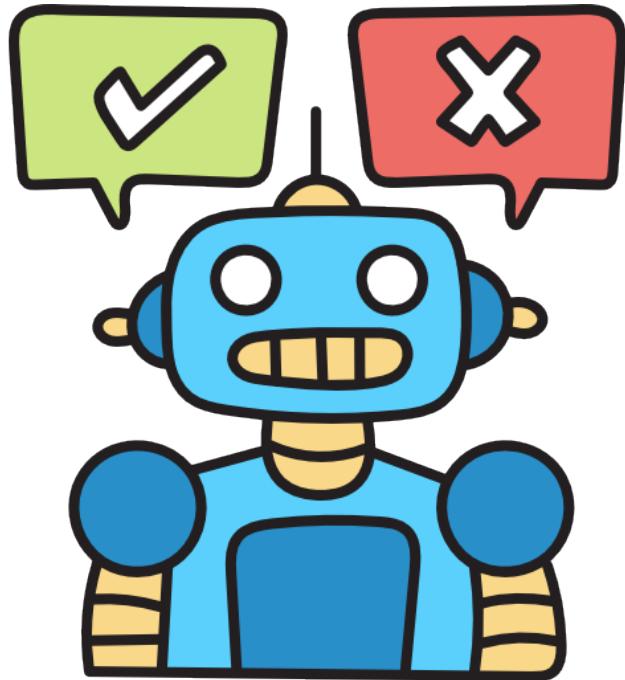


A máquina, através de algoritmos, obter padrões sobre características extraídas dos dados para, com um modelo gerado/criados, classificar as observações futuras de novos dados.

No conceito cada vez menos intervenção humana (conceito).

Pré-processamento e análise dos dados, além de realizar “grid” de valores para treinamento obterem maior acurácia  
(na prática)

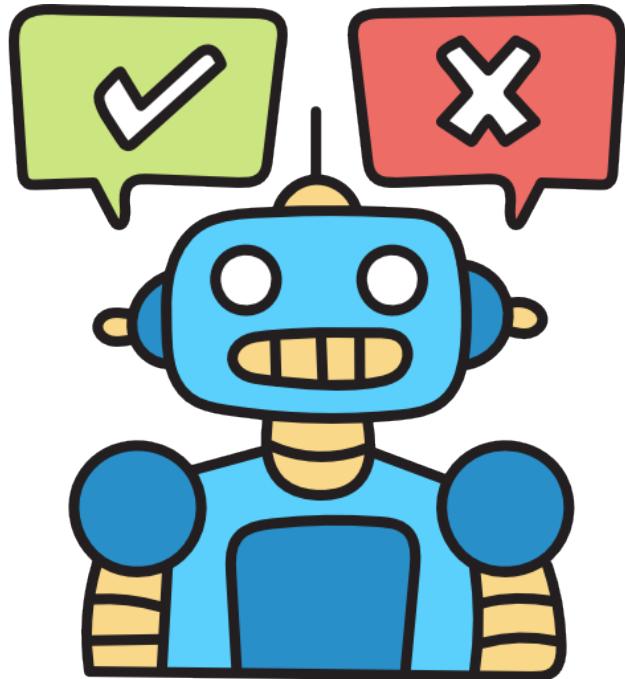
# Machine Learning - História



**1950 - IA: Computadores com habilidade de “pensar” -**  
Teste de Turing. Em 2014 chatbot enganou 10/30 juízes



# Machine Learning - História

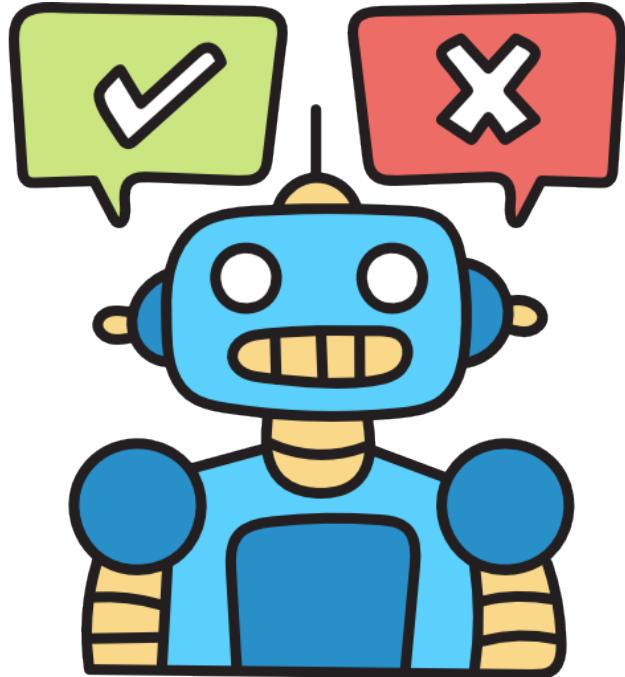


**1959 - ML: Aprender a partir dos dados** - Arthur Samuel

Aprender com a experiência que existe intrínseca aos dados.

Algoritmos de aprendizado de máquina analisam as correlações entre os atributos (variáveis) de um sistema (base de dados) a partir de dados amostrais (base de treinamento)

# Machine Learning - História



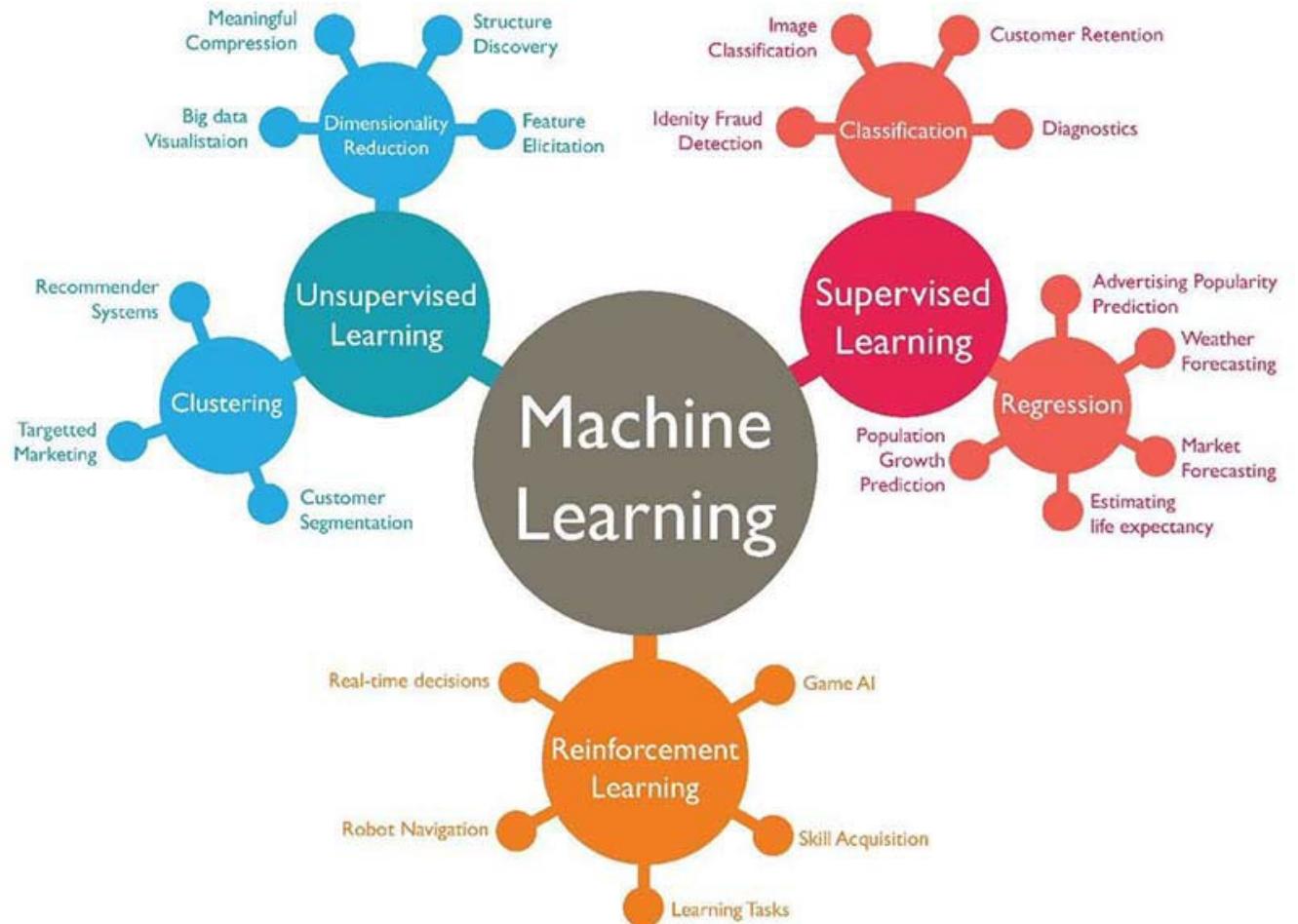
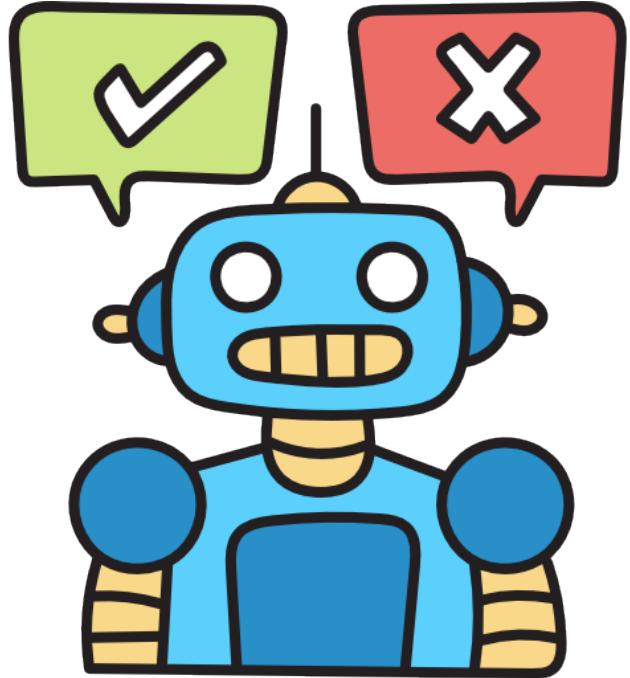
## 2012: DS – Entender os Dados

Ciência de dados utilizando probabilidade, estatística, álgebra linear e computação.

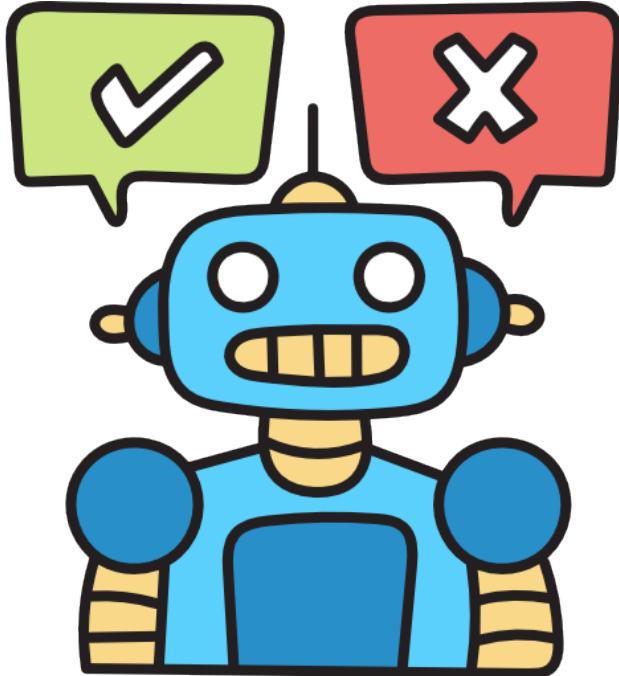
Conhecimentos de IA e ML

*“É a ciência (e arte) de programar computadores de tal forma que eles aprendam a partir de dados”*  
(Aurélien Géron, 2017)

# Machine Learning – Tipos de Aprendizado



# Machine Learning – Tipo de Aprendizado

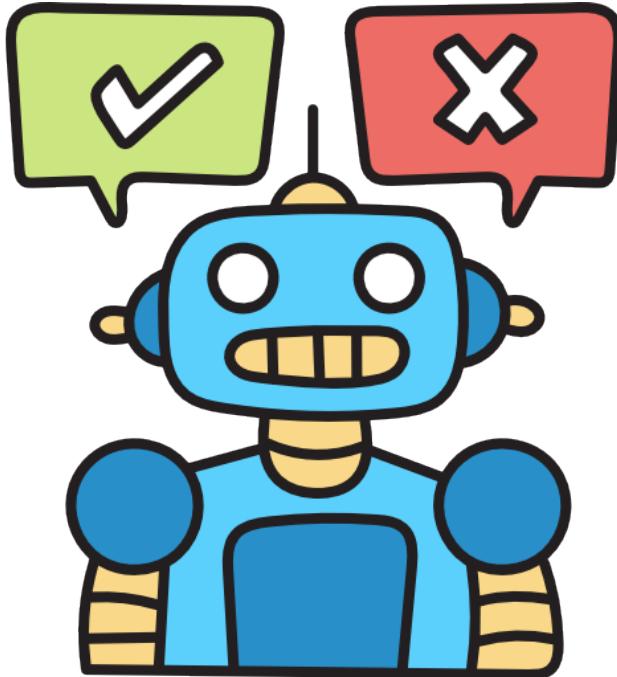


**Supervisionado** → rotulado com saídas esperadas. Modelo gera ao entrar com conjunto de características uma saída rotulada ([Classificação](#)) ou um valor futuro ([Predição](#)). Ex: Nosso desafio AgroXP.

**Não Supervisionado** → Não existe rótulo prévio. Analisa a rede de relacionamento entre os dados para agrupá-los por características similares. Ex: Categorização de Clientes

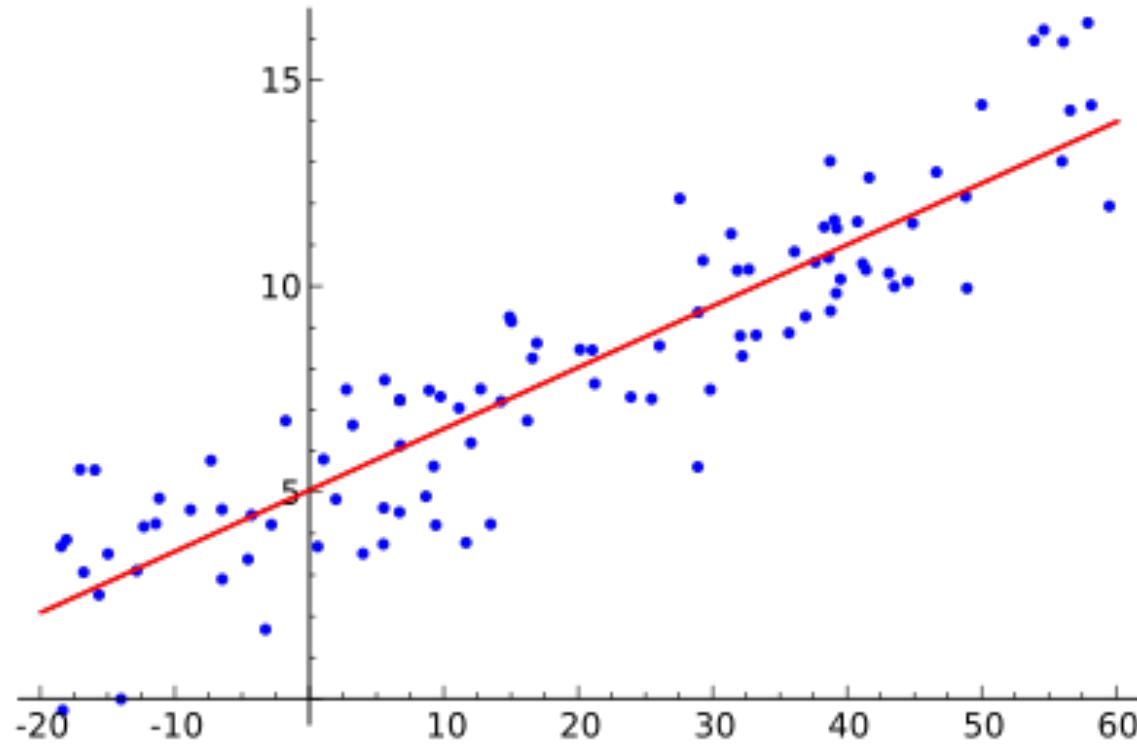
**Reforço** → Maximizar o resultado. Baseado em recompensa / punição. Com isso algoritmo encontrar a “política” que mapeia os dados. Ex: Personagens Jogos

# Machine Learning – Exemplos Algoritmos

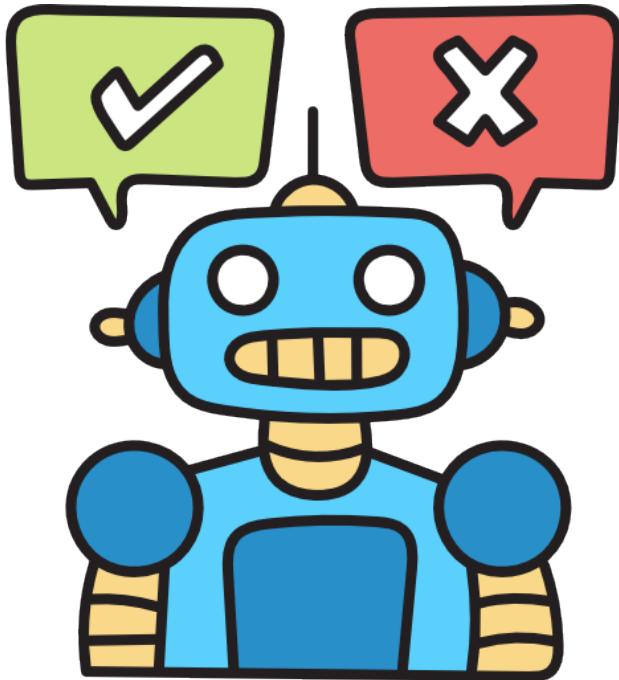


## Regressão Linear (Supervisionado – Predição)

Simples... Busca uma reta para se ajustar aos dados.  
Problemas de relação linear.

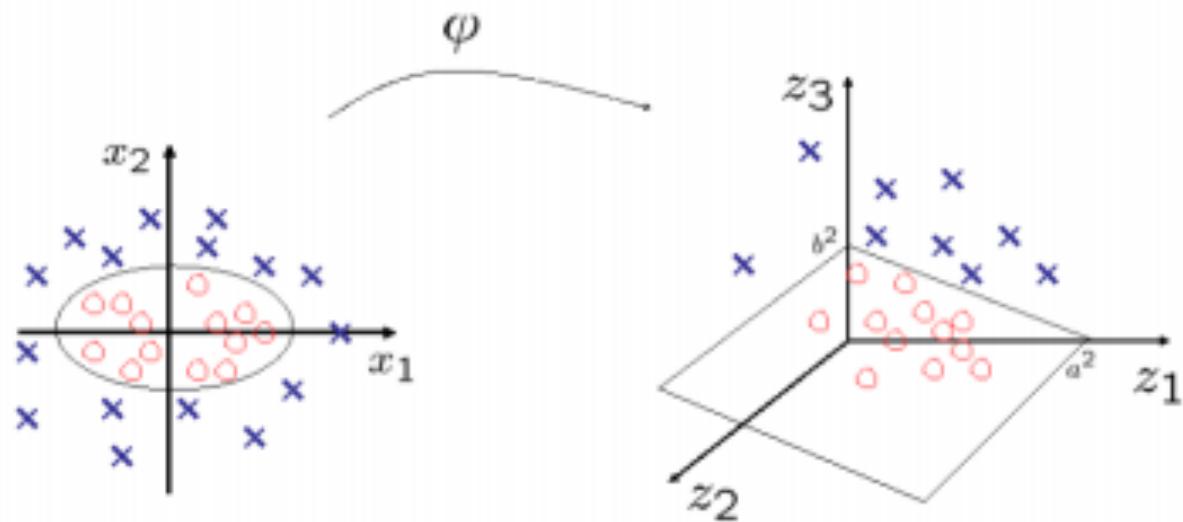


# Machine Learning – Exemplos Algoritmos

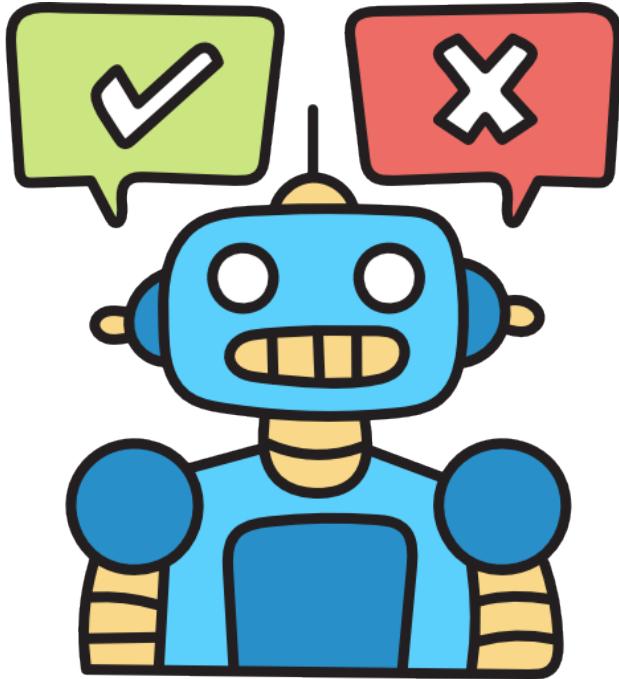


**SVM - Support Vector Machine (Supervisionado – Classificação) – Vapnik (1963)**

Distância das amostras da linha superfície de separação.  
Consegue trabalhar com dados não lineares com a premissa  
de que em alguma dimensão os dados terão linearidade.

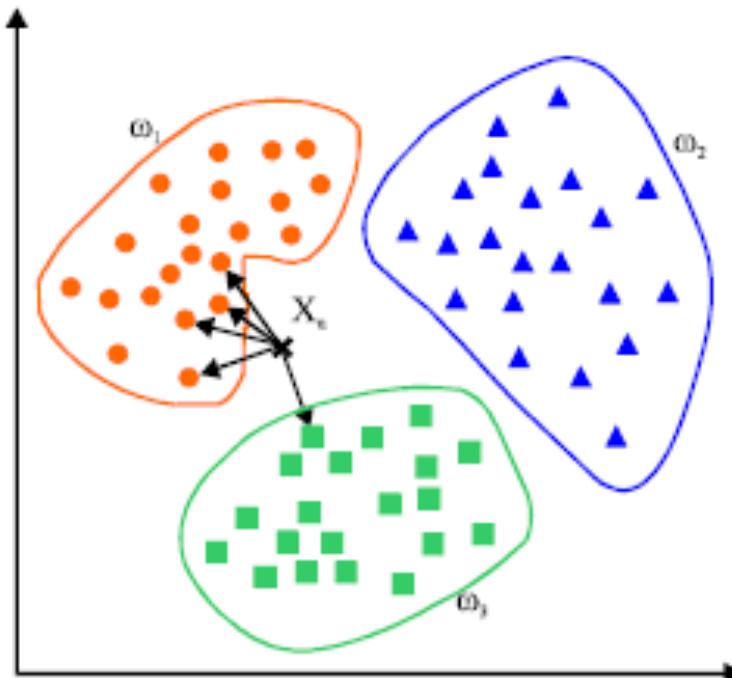


# Machine Learning – Exemplos Algoritmos

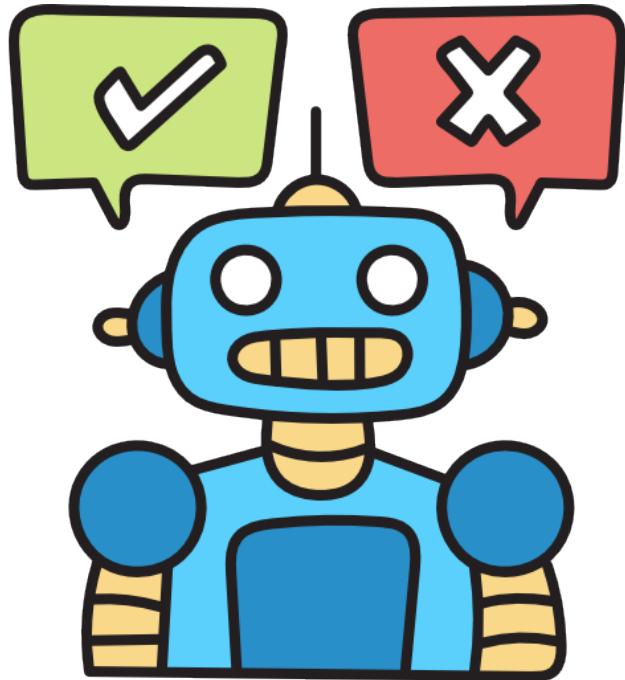


## KNN – K-Nearest Neighbors (Supervisionado – Classificação)

Baseado em encontrar o valor de K que consiga através de funções básicas de distância Euclidiana encontrar a melhor superfície de separação

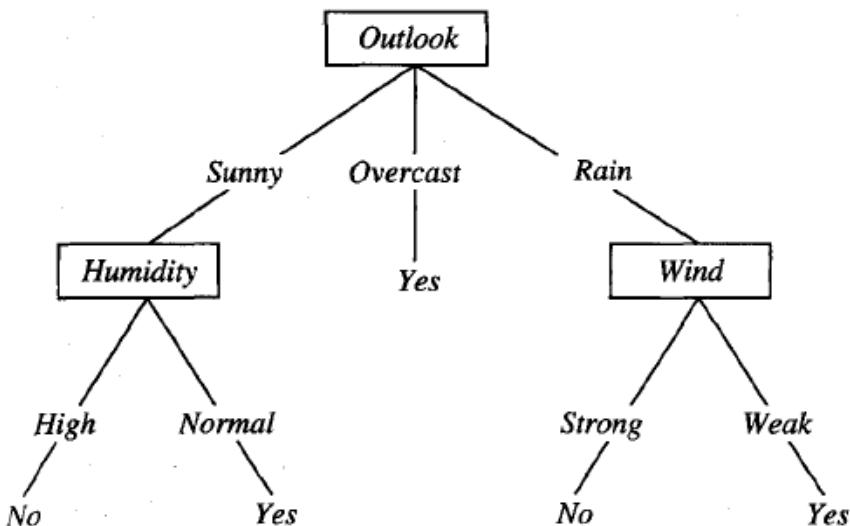


# Machine Learning – Exemplos Algoritmos

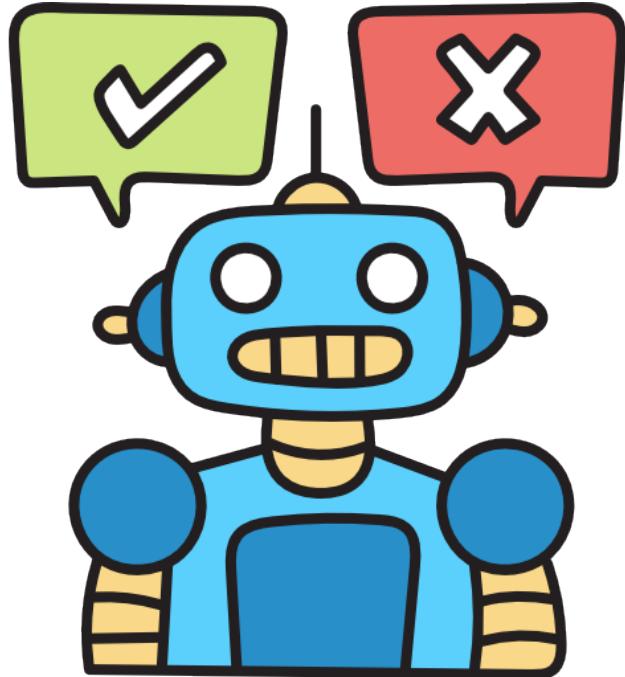


## Árvore de Decisão (Supervisionado – Classificação)

De fácil explicação do modelo obtido, este algoritmo utiliza a categorização utilizando técnicas referente a Ganho de Informação dos atributos (o quanto a variável sozinha classifica os exemplos de treinamento). Pode ser utilizado para dados numérico ou simbólicos.

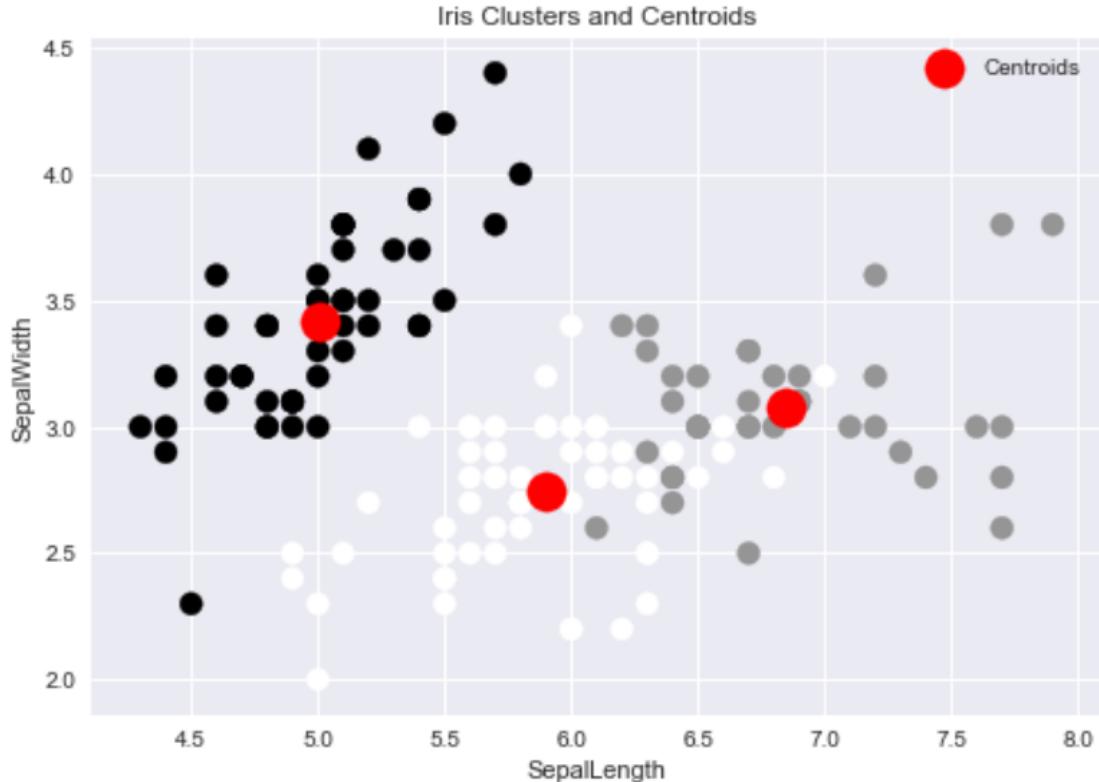


# Machine Learning – Exemplos Algoritmos

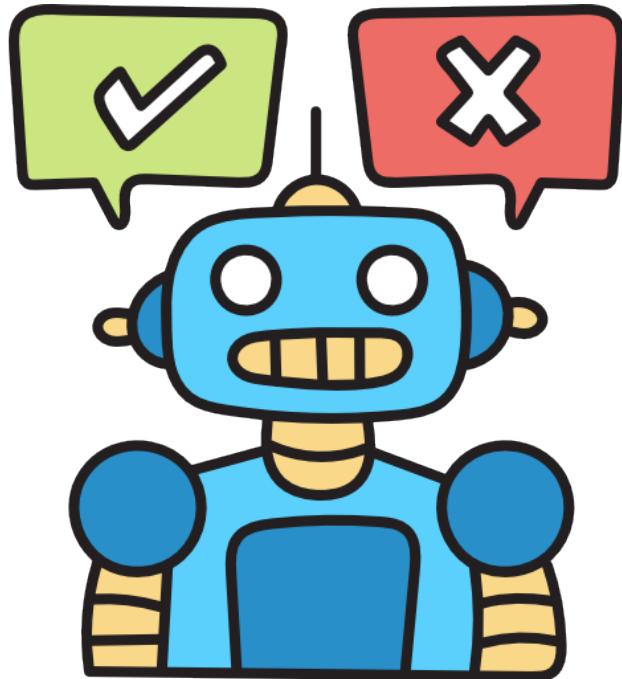


## K-Means – (Não Supervisionado)

Forma clusters que contêm pontos homogêneos aos dados.

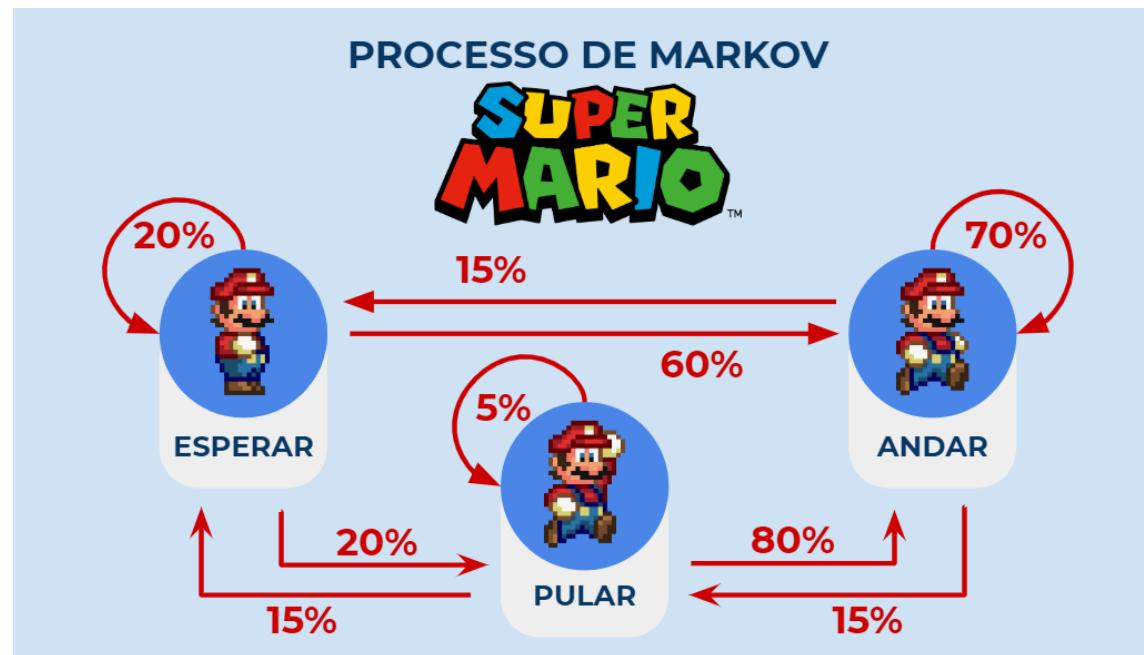


# Machine Learning – Exemplos Algoritmos

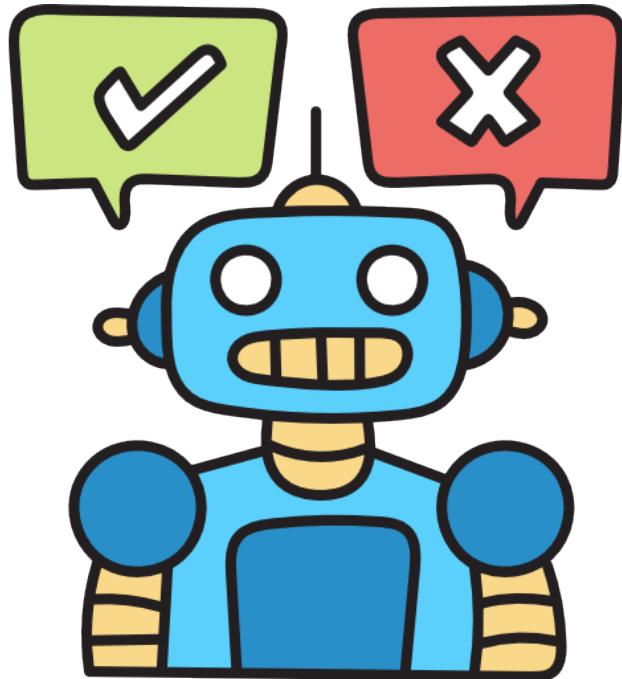


## Cadeia de Markov (Reforço)

Processo estocástico (futuro ← estado atual). Com base na cadeia e suas probabilidades o algoritmo toma uma decisão e, se houver recompensa, reforça a decisão tomada. Se houver uma punição rechaça.

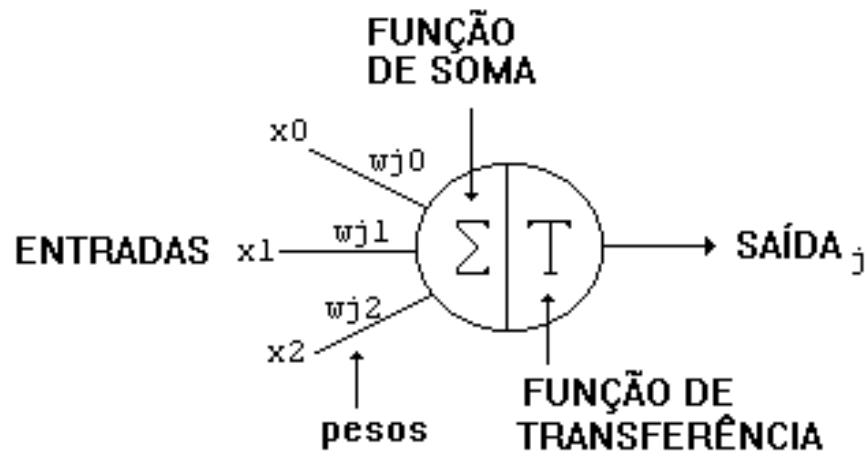


# Machine Learning – Exemplos Algoritmos

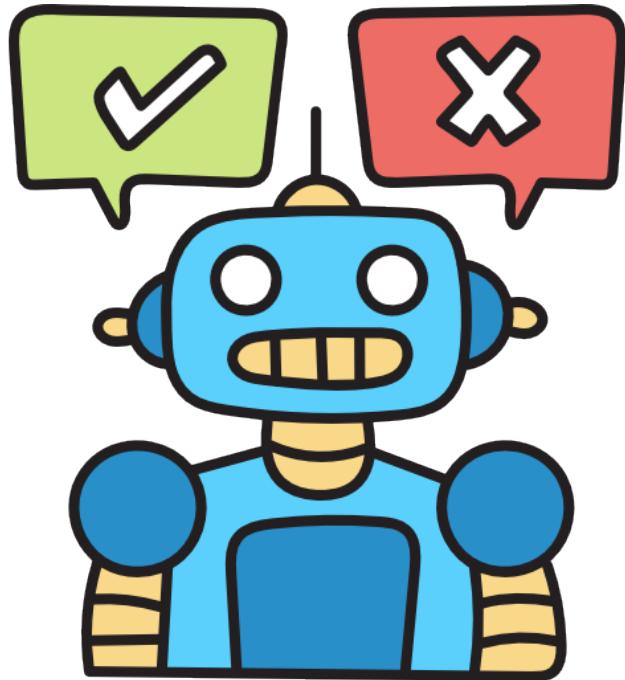


## Redes Neurais (Supervisionado – Classificação)

Baseado no conceito matemático e computacional (1943) que visa descrever o modelo artificial para um neurônio biológico. Responde “ligando/desligando” os vários neurônios interligada e com isso classifica as características de entrada no rótulo predito pelo modelo.

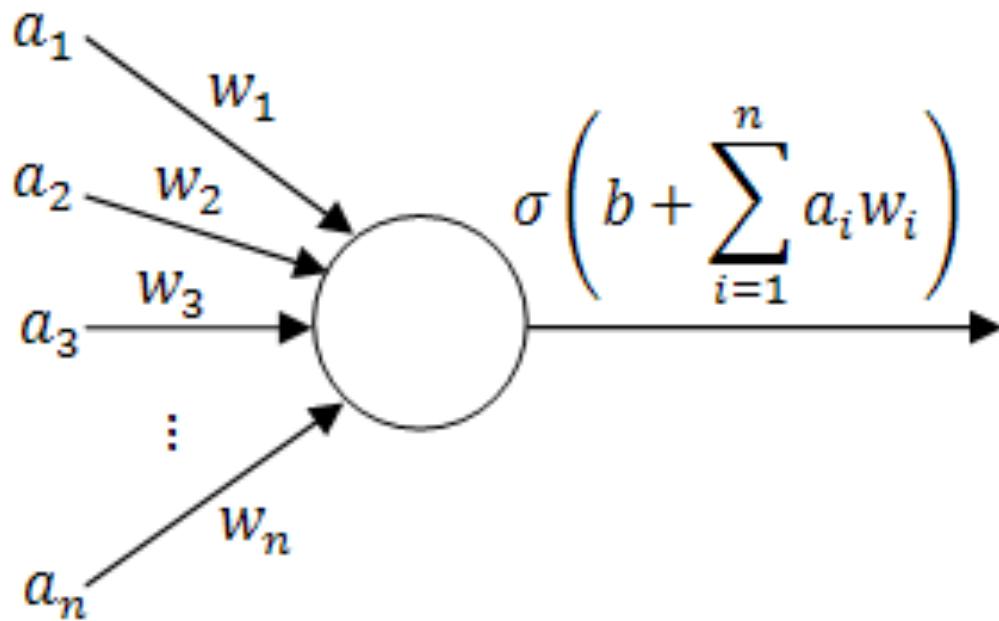


# Machine Learning – Exemplos Algoritmos

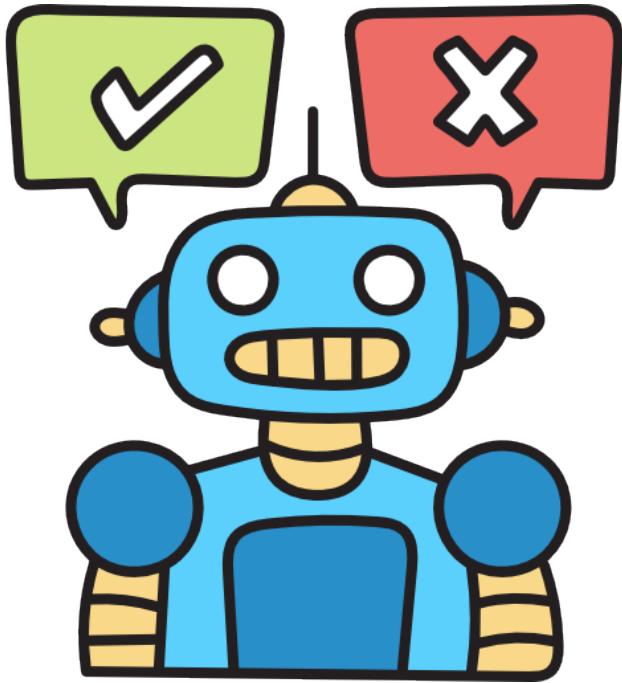


## Redes Neurais (Supervisionado – Classificação)

Perceptron → Tipo básico de rede neural. Demonstrou em 1957 a possibilidade de simulação de um neurônio biológico.

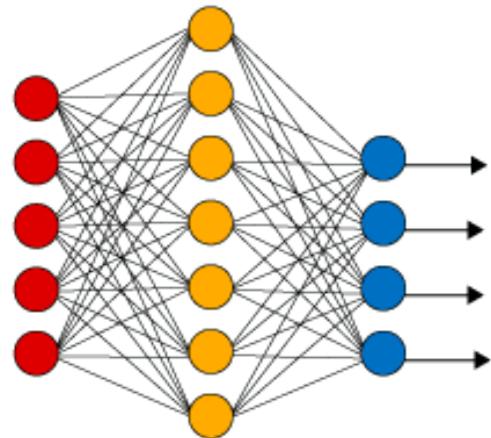


# Machine Learning – Exemplos Algoritmos



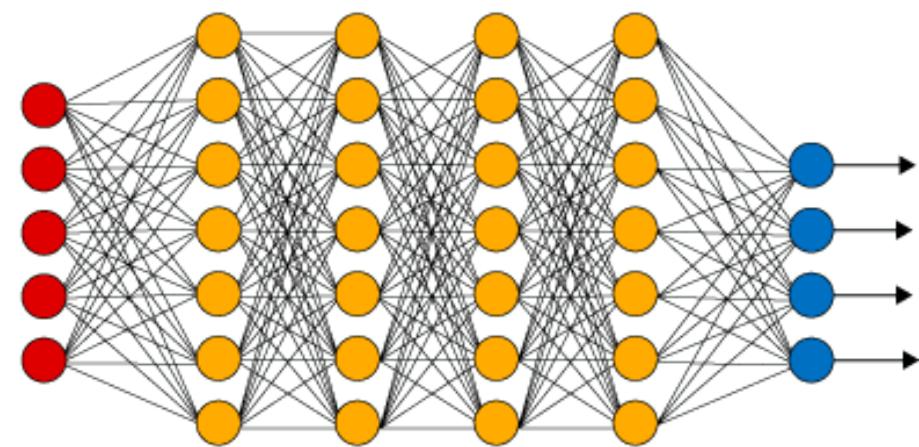
## Redes Neurais (Supervisionado – Classificação)

Simple Neural Network



● Input Layer

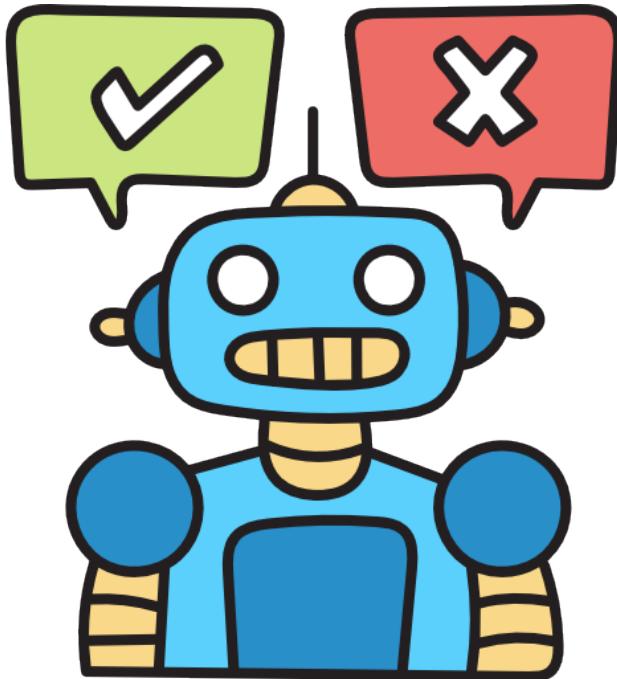
Deep Learning Neural Network



● Hidden Layer

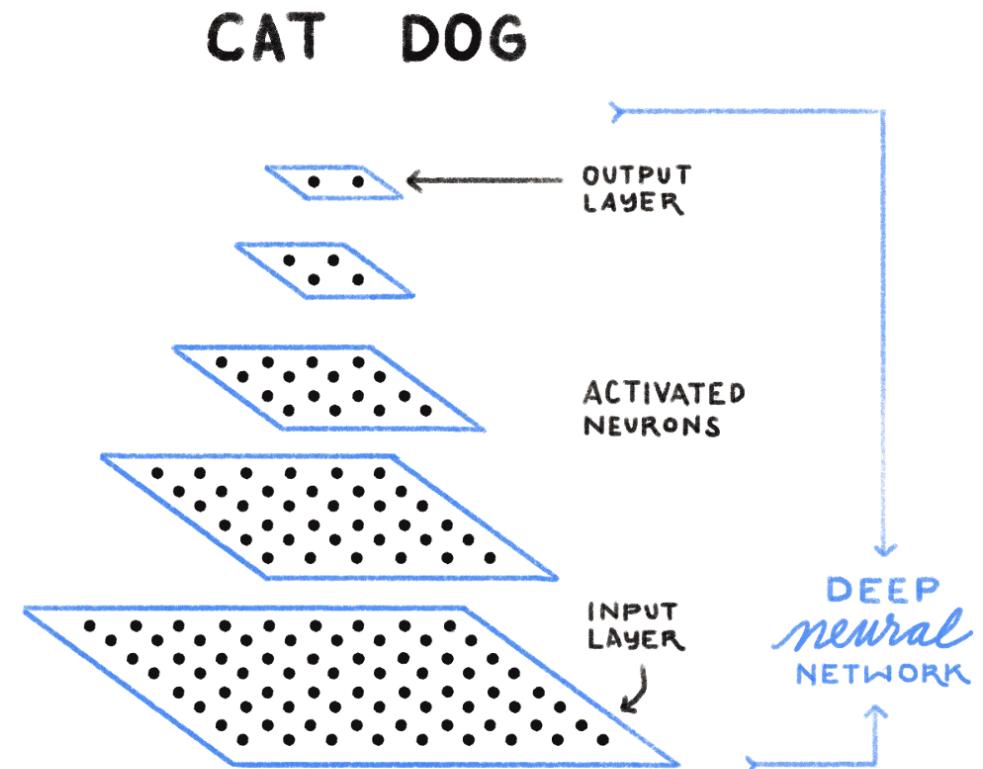
● Output Layer

# Machine Learning – Exemplos Algoritmos

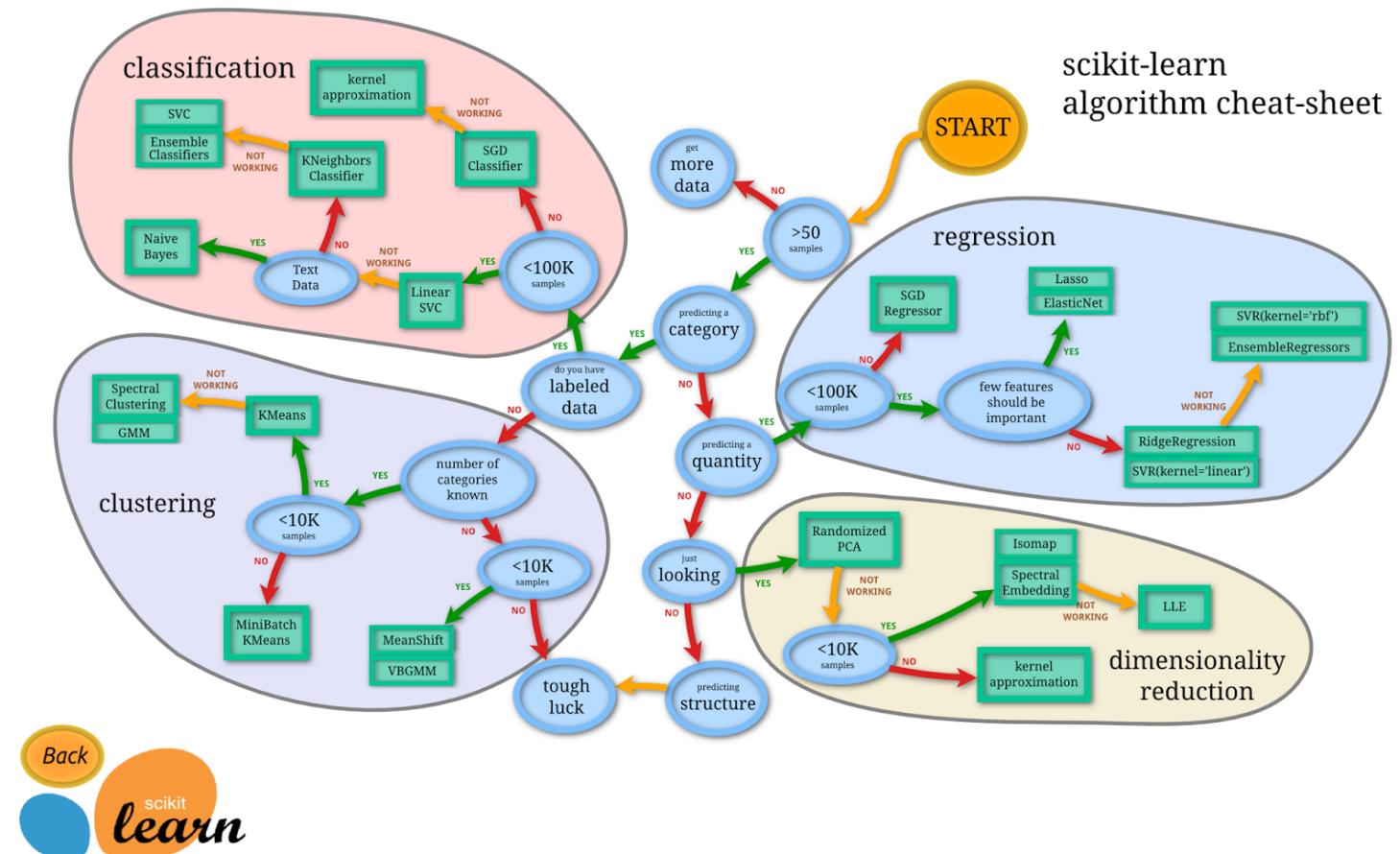
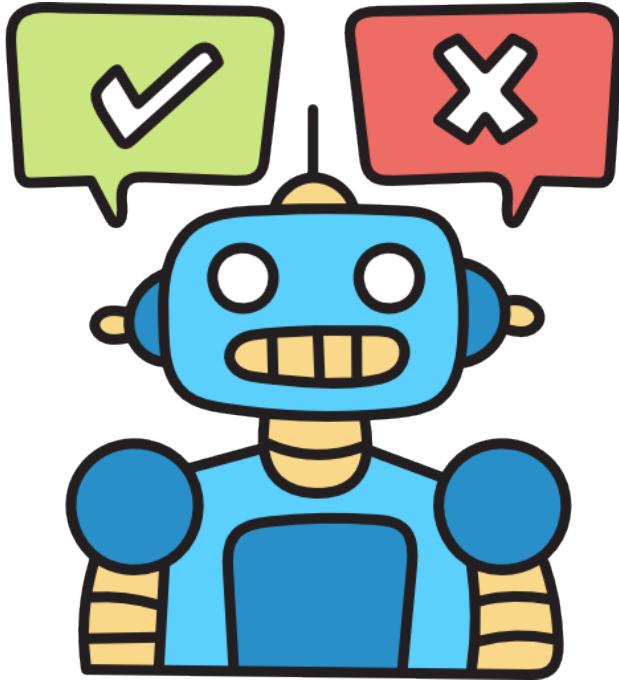


## Redes Neurais (Supervisionado – Classificação)

IS THIS A  
**CAT or DOG?**



# Machine Learning – Algoritmo x Características Dados



Fonte: [https://scikit-learn.org/stable/tutorial/machine\\_learning\\_map/](https://scikit-learn.org/stable/tutorial/machine_learning_map/)

# Agenda

1 – Agenda

2 – R – Parte II

3 – Machine Learning

4 – Namorando Dados (SQL)

5 – Desafio Pessoal

6 – Desafio do Curso

7 – Bate Papo e Monitoria



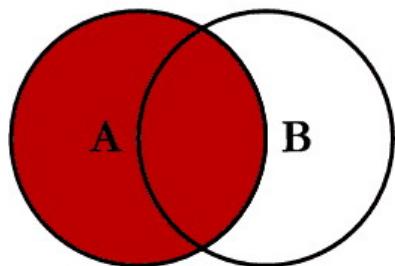
# O Trabalho do Cientista de Dados > Desafio Curso

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção

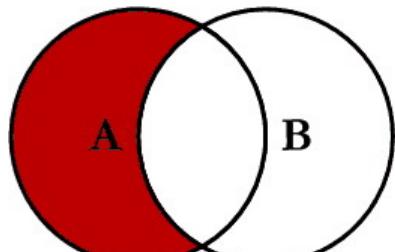


# Namorando os Dados (Queries SQL)

## SQL JOINS

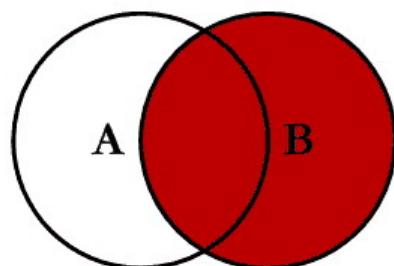


```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```

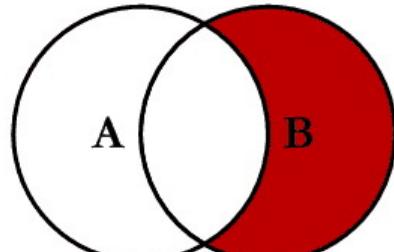


```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```

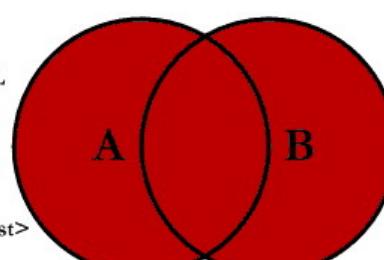
```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```



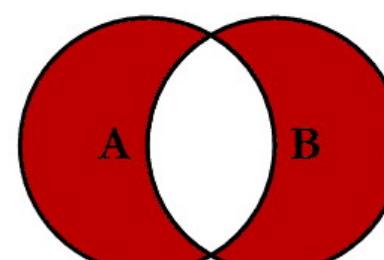
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```



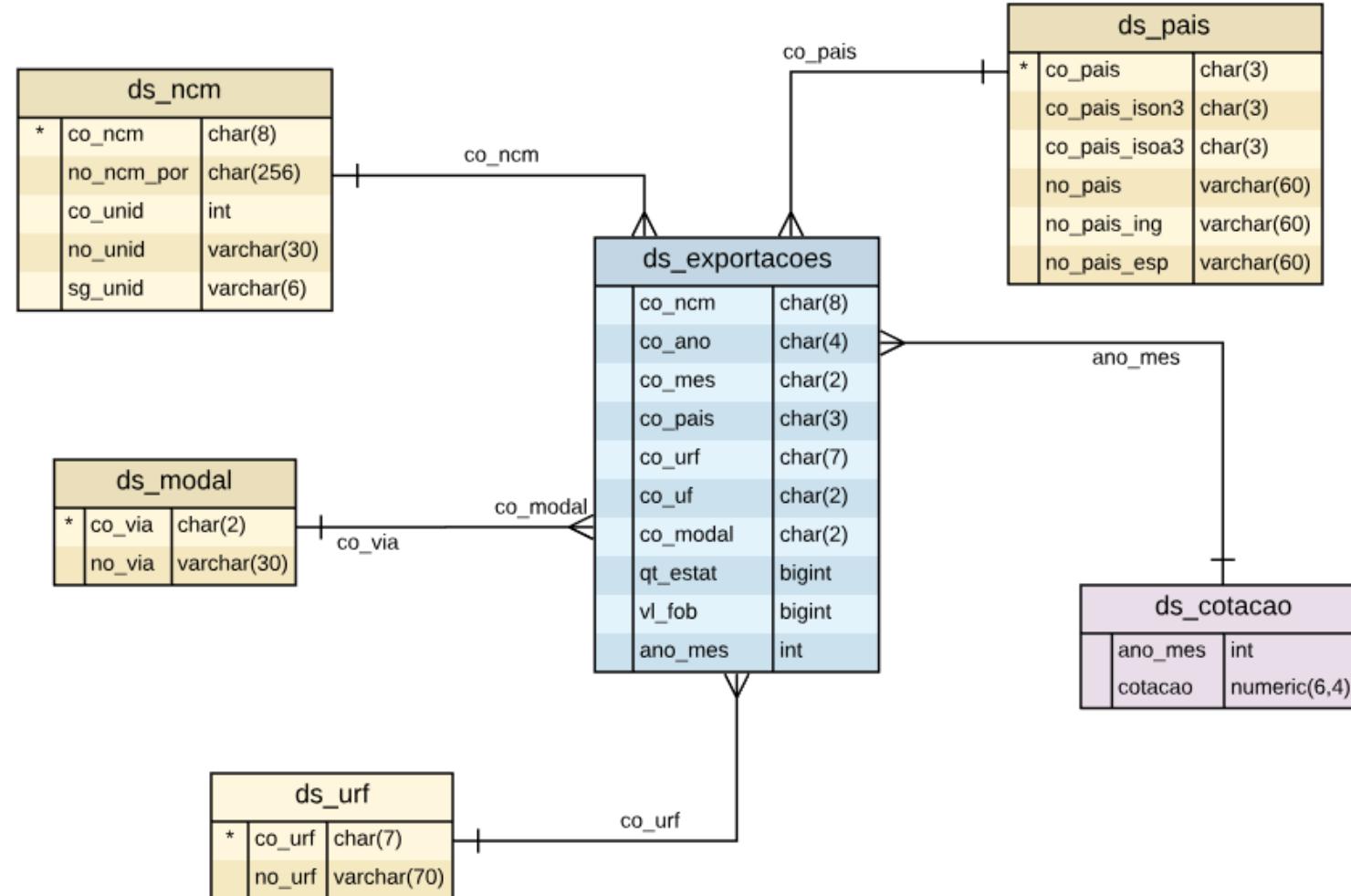
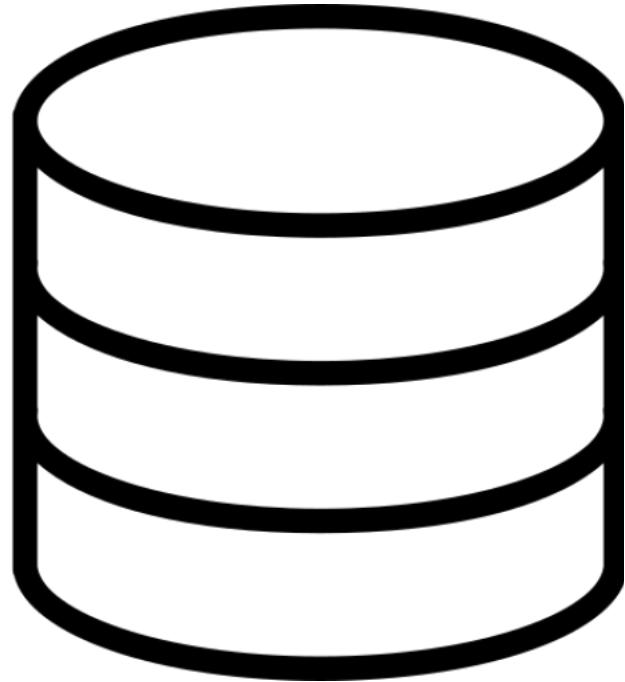
© C.L. Moffatt, 2008



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

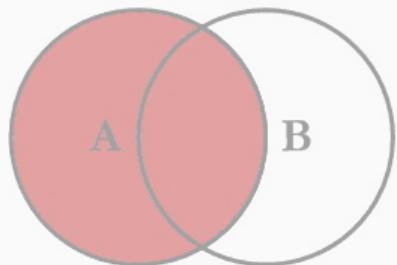


# Desafio – Modelo de Dados



# Namorando os Dados (Queries SQL)

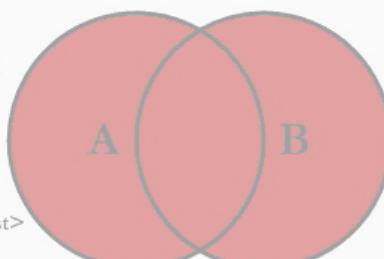
## SQL JOINS



```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```

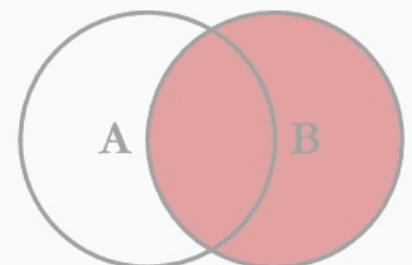


```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```



```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```

## SQL JOINS

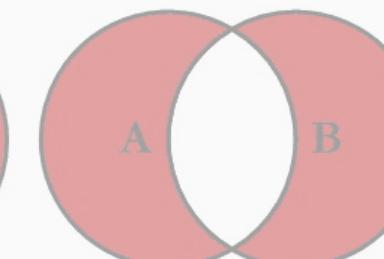


```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```

```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

© C.L. Moffatt, 2008



# Agenda

1 – Agenda

2 – R – Parte II

3 – Machine Learning

4 – Namorando Dados (SQL)

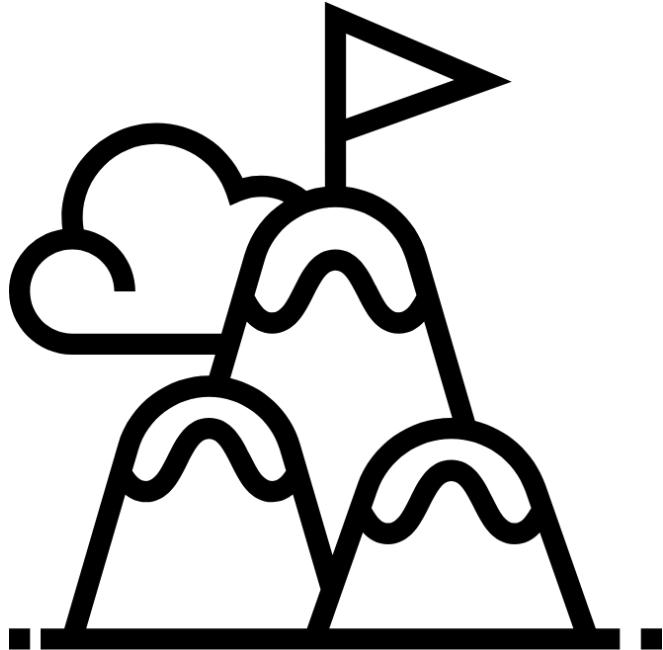
5 – Desafio Pessoal

6 – Desafio do Curso

7 – Bate Papo e Monitoria



# Desafio Pessoal



## Machine Learning

O Melhor Algoritmo Vence

→ Maior Acurácia

→ Menor Tempo Processamento

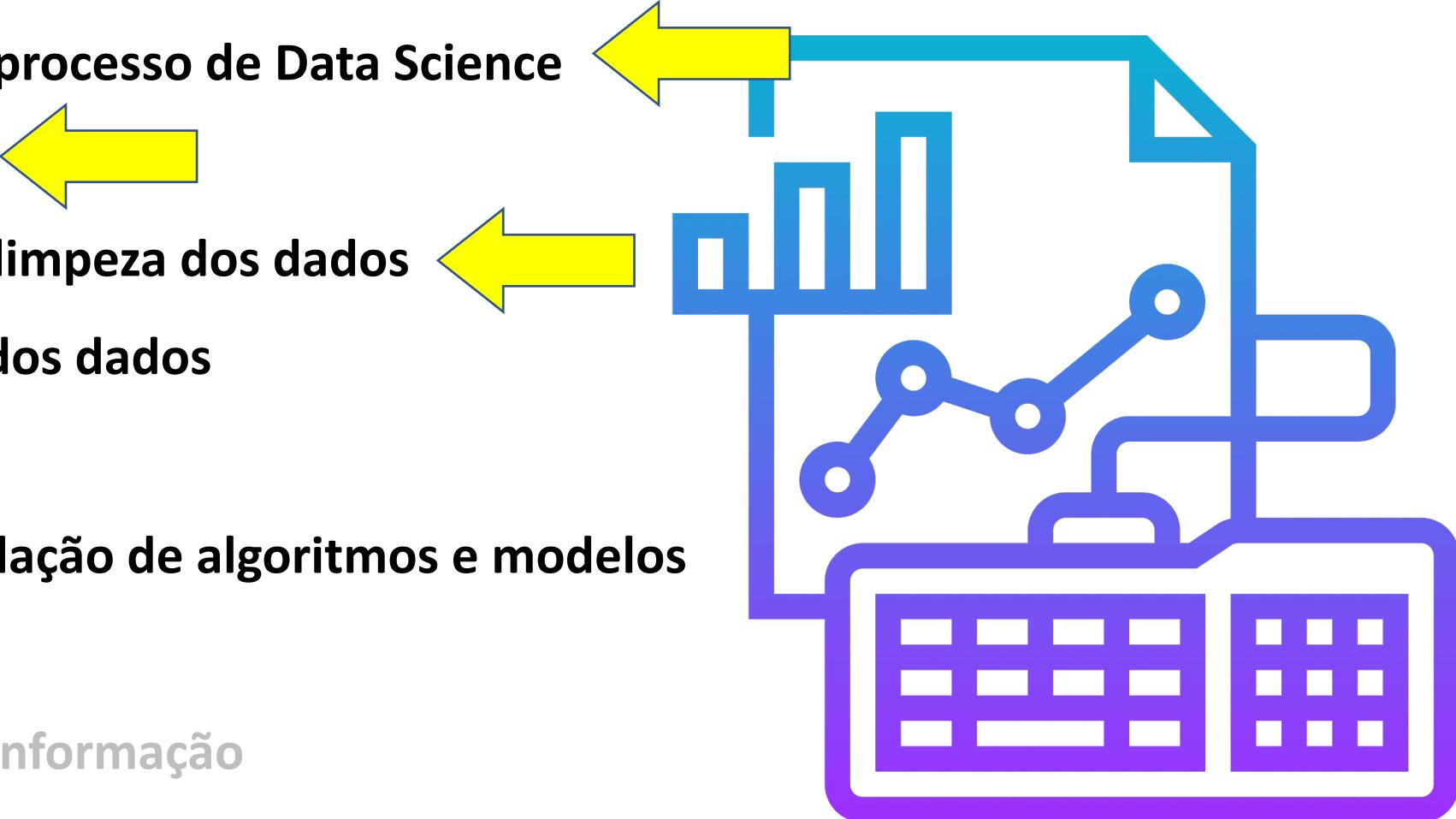
Entrega (Código .py ou .R) → até **30/09/2019**

Resultado → **02/10/2019**

Prêmio: Super Cupom Madeira Madeira

# O Trabalho do Cientista de Dados > Desafio Pessoal

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção

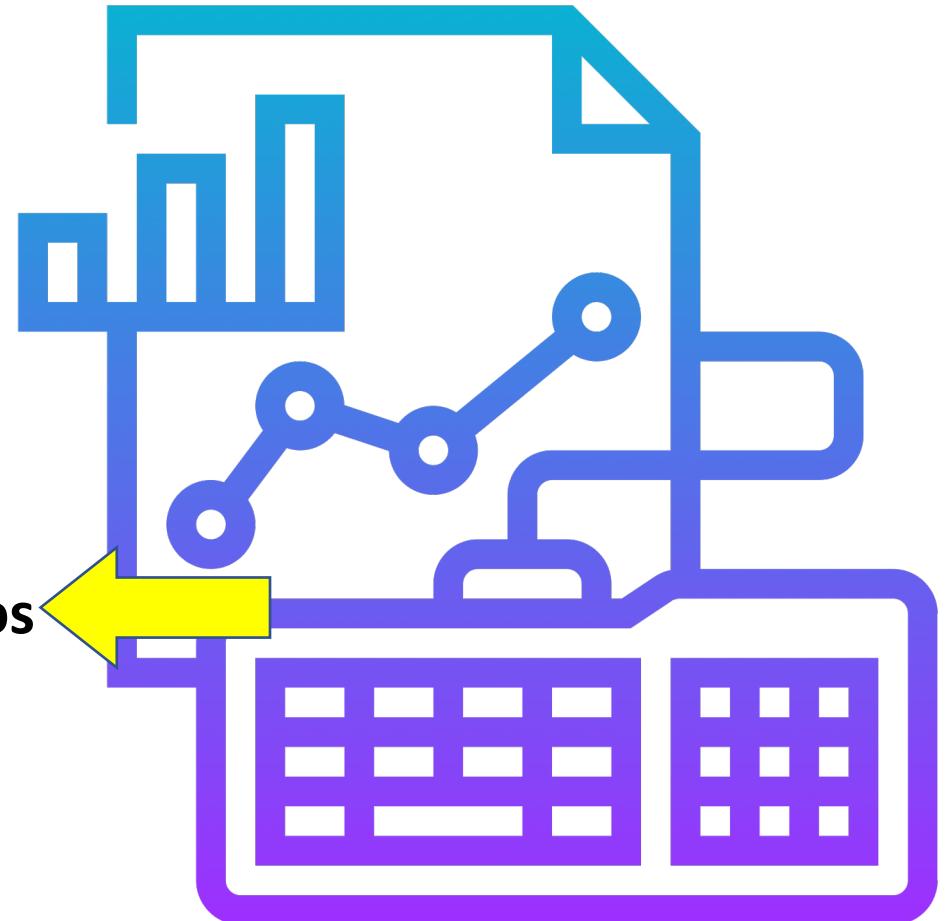


# Desafio Pessoal > Extração de Características



# O Trabalho do Cientista de Dados > Desafio Pessoal

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



# Desafio Pessoal > Treinando Modelo



# Agenda

1 – Agenda

2 – R – Parte II

3 – Machine Learning

4 – Namorando Dados (SQL)

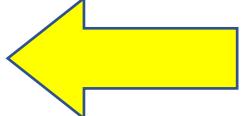
5 – Desafio Pessoal

6 – Desafio do Curso

7 – Bate Papo e Monitoria



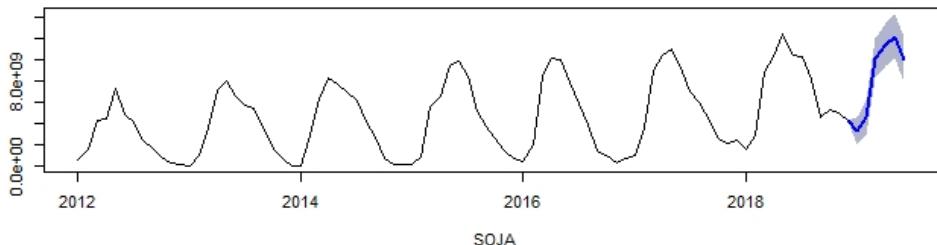
# O Trabalho do Cientista de Dados > Desafio Curso

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization 
9. Disseminação da informação
10. Colocar modelo em produção

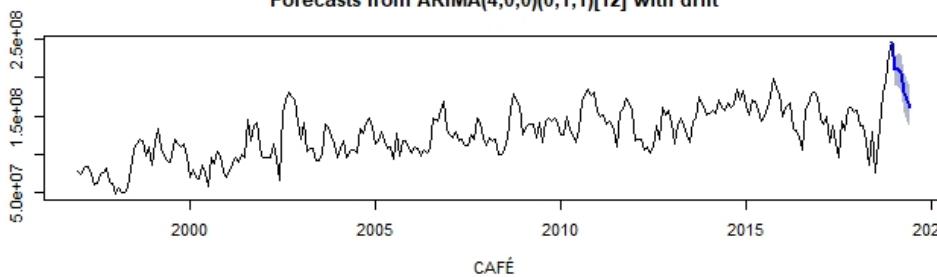


# Agro XP Brazil - Solução

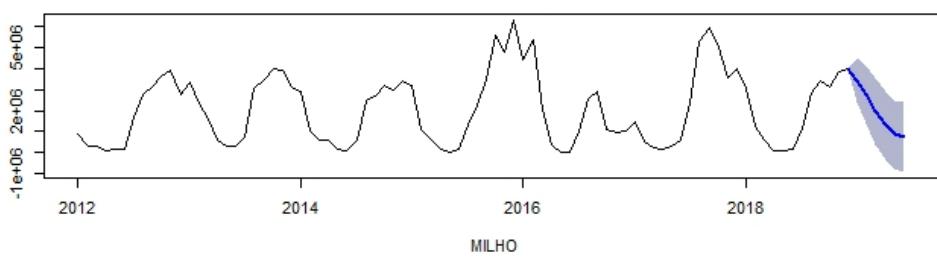
Forecasts from ARIMA(1,0,0)(2,1,0)[12]



Forecasts from ARIMA(4,0,0)(0,1,1)[12] with drift

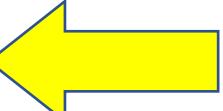


Forecasts from ARIMA(3,0,1) with non-zero mean



- **Proposta:** Verificar qual é a previsão para os próximos 4 meses para cada um dos grãos
- E decidir em qual commodities iremos investir no 1º semestre/2019
- Utilizaremos técnicas de **Séries Temporais**

# O Trabalho do Cientista de Dados > Desafio Curso

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação 
10. Colocar modelo em produção 



# Agenda

1 – Agenda

2 – R – Parte II

3 – Machine Learning

4 – Namorando Dados (SQL)

5 – Desafio Pessoal

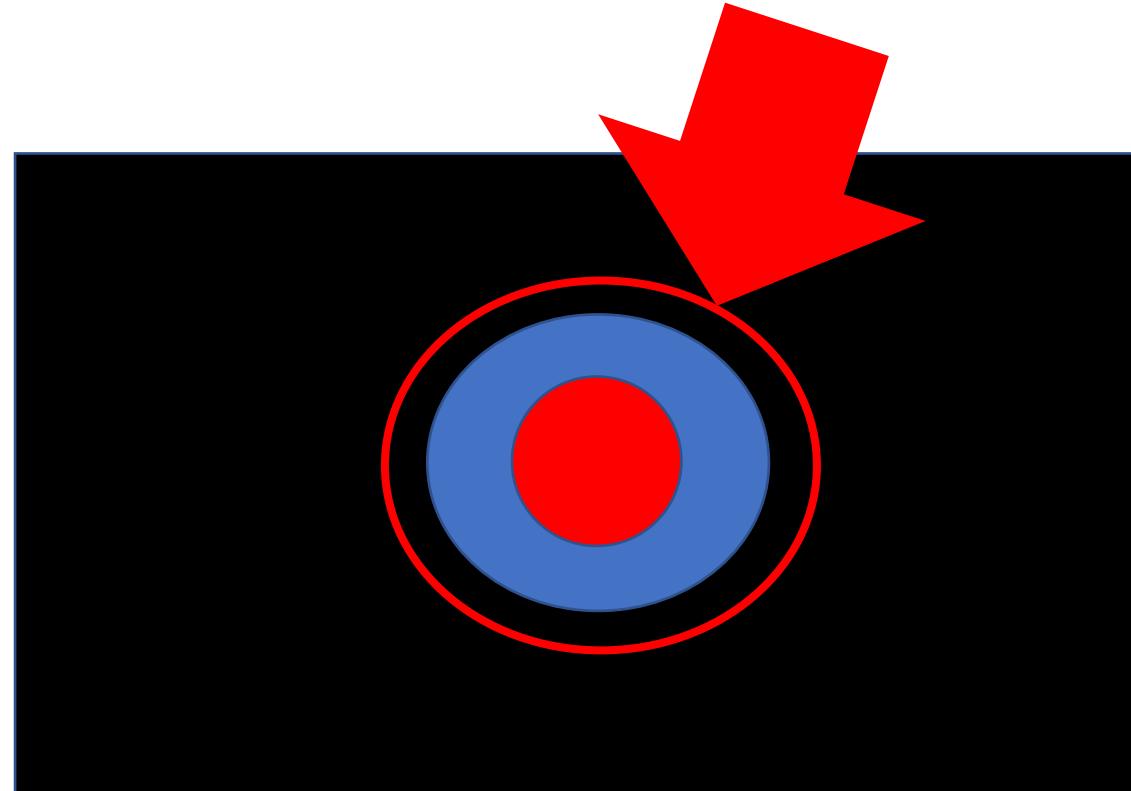
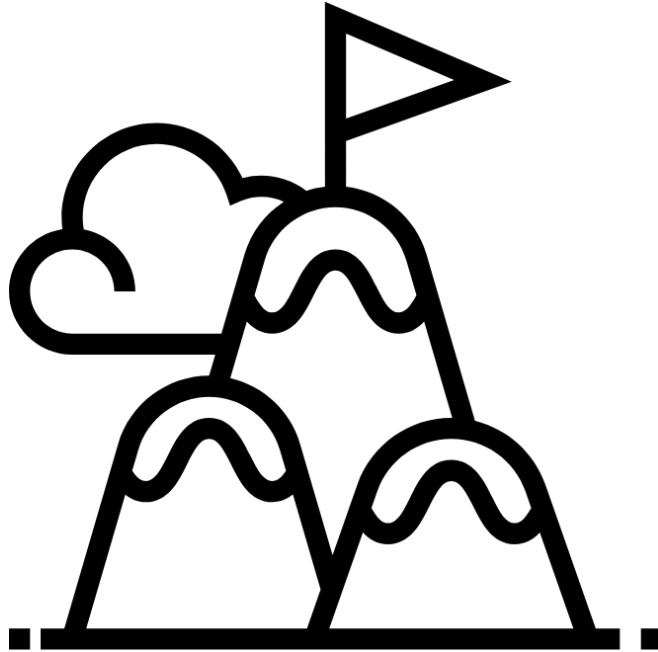
6 – Desafio do Curso

7 – Bate Papo e Monitoria



# Bate Papo e Mentoría > Objetivo Tema do Curso

**Objetivo sobre o tema do curso:**



# Bate Papo e Mentoría > Metas / Próximos Passos



E a partir de agora?

Estudo

Colocar em Prática / Projetos

Insights

# Obrigado !

 Charles Adriano dos Santos  
 charles.a.santos@caelis.it  
 chadri  
 41 99144 6663

 Rafael Roberto Dias  
 rafael.dias@madeiramadeira.com.br  
 rafael-roberto-dias-00b39123  
 41 99672 7170