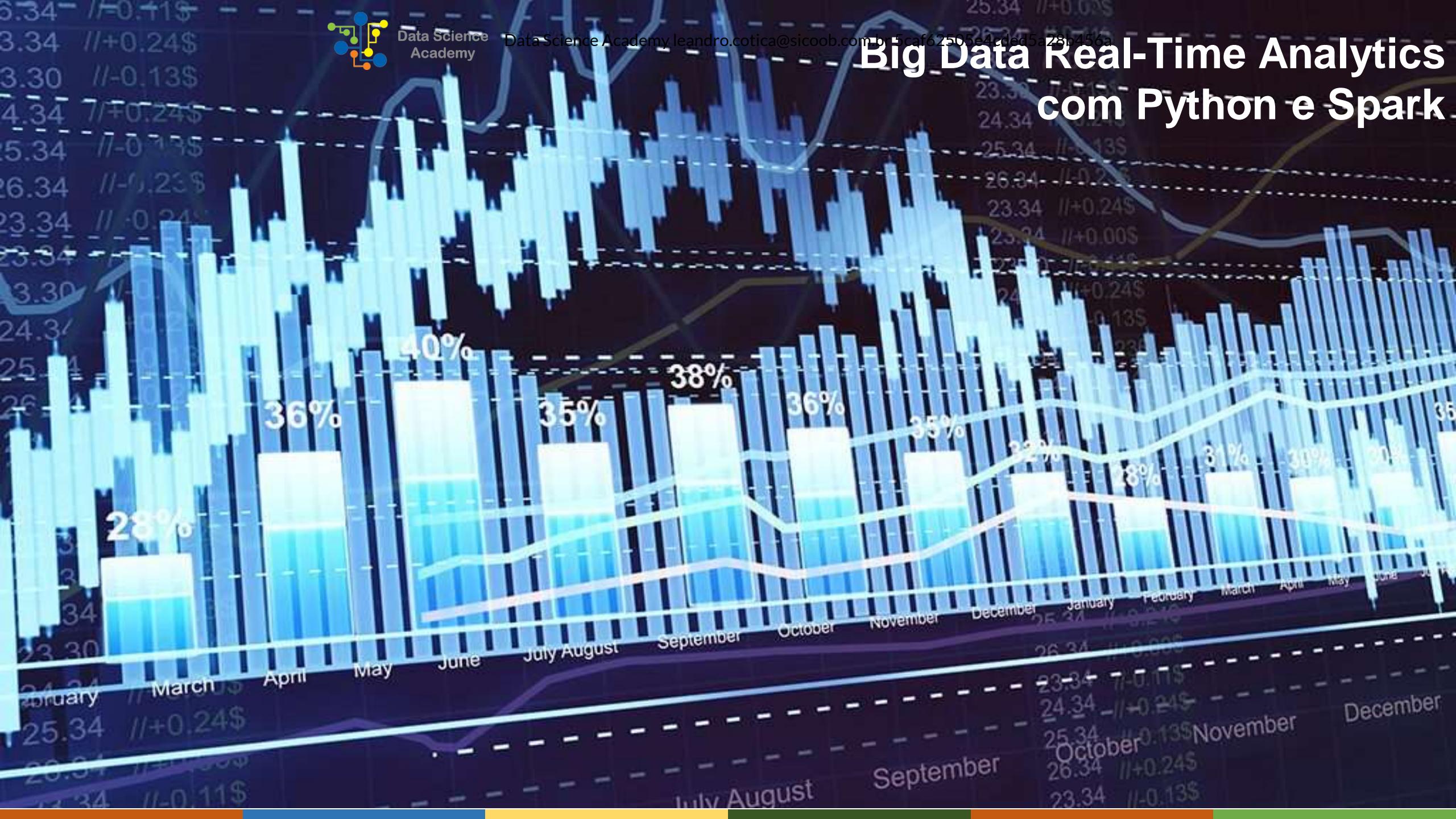


# Big Data Real-Time Analytics com Python e Spark





Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

**Seja muito bem-vindo(a)!**



Data Science  
Academy

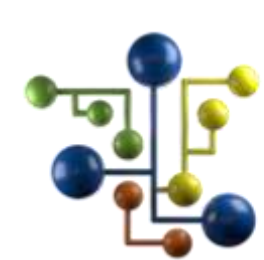
Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

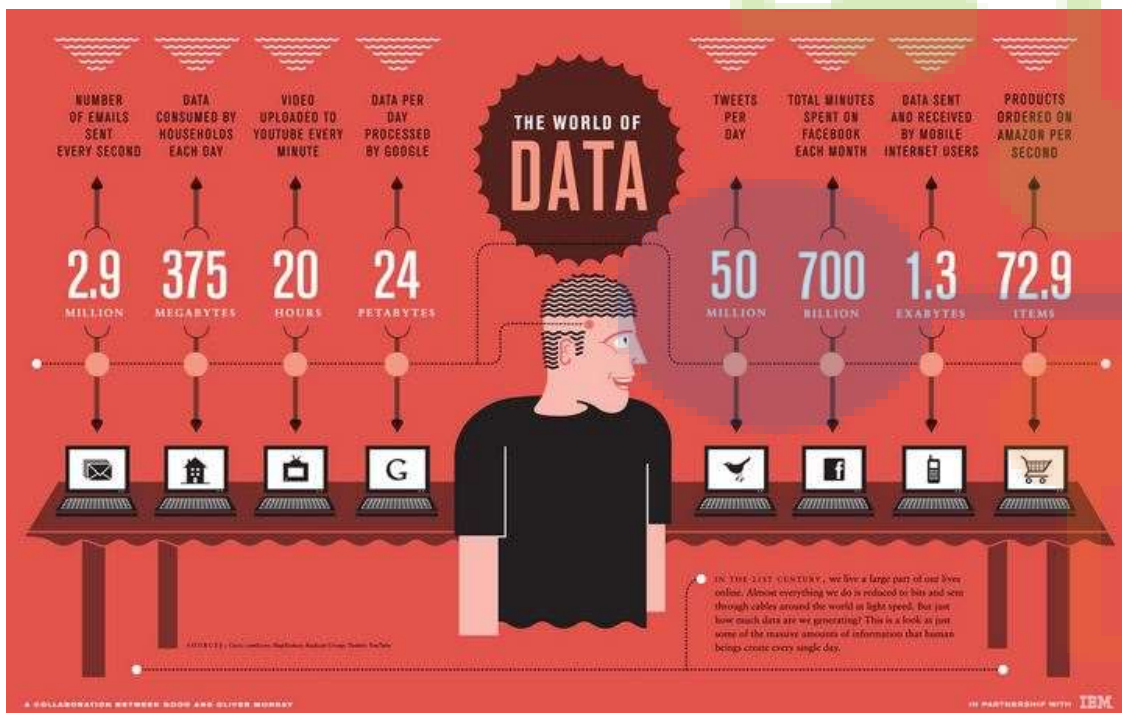
## Machine Learning em Python







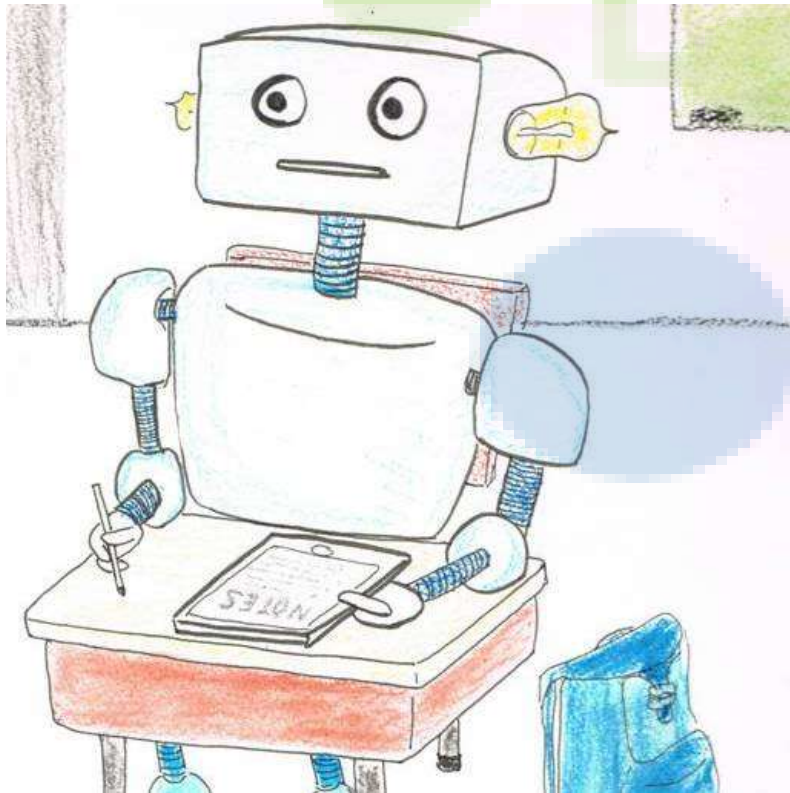
# Introdução



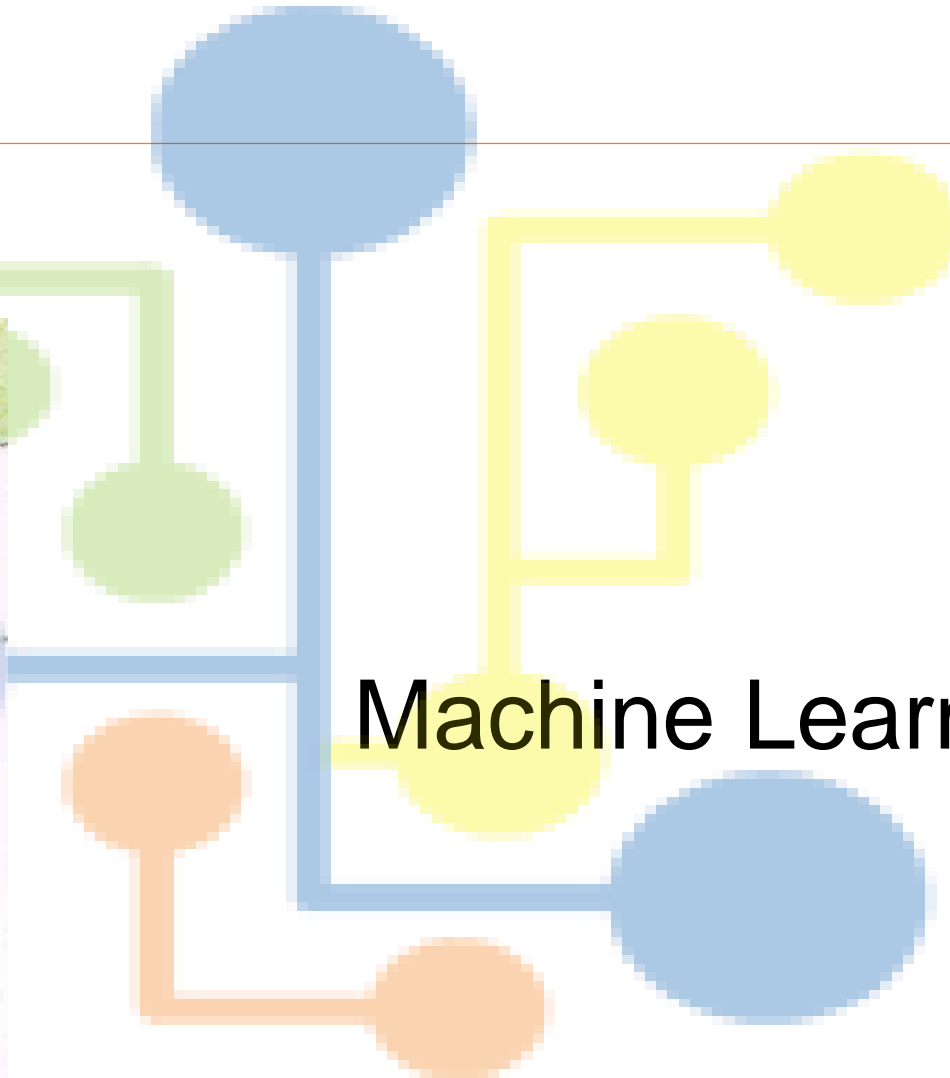
Estamos no meio de uma revolução, proporcionada pelo Big Data.



# Introdução



Machine Learning





# Introdução

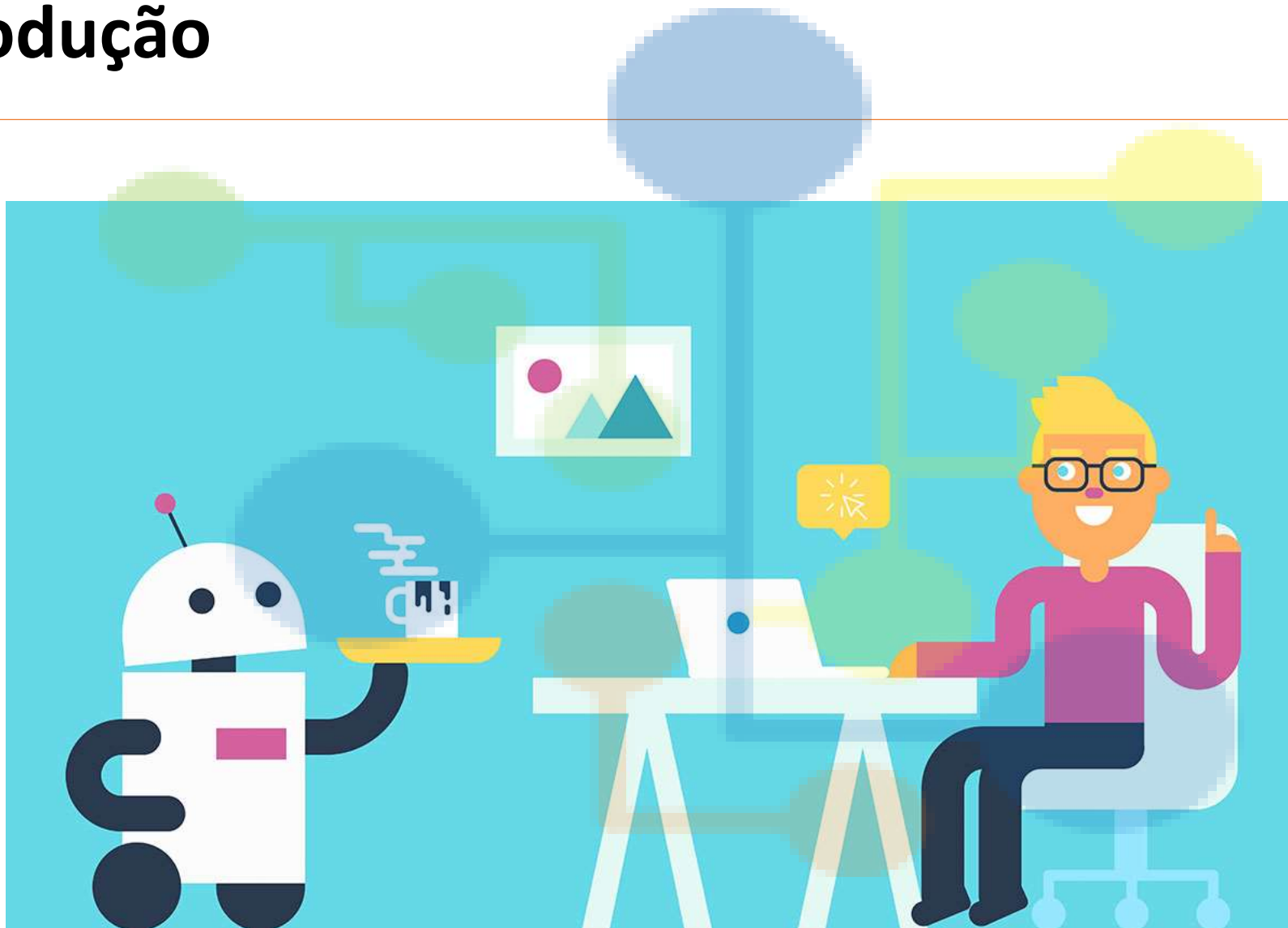
Considerar uma carreira em Machine Learning e aprender tudo que for possível sobre esse assunto é uma das decisões mais inteligentes que você pode tomar na sua carreira profissional.



Data Science  
Academy

Data Science Academy [leandro.cotica@sicoob.com.br](mailto:leandro.cotica@sicoob.com.br) 5caf62505e4cded5a28b456a

# Introdução





# Introdução



Computação Cognitiva





# Introdução

O que vamos estudar neste capítulo?

- Processo de Machine Learning
- Biblioteca Scikit-learn
- Coleta, Análise Exploratória e Pré-Processamento
- Feature Selection
- Algoritmos de Machine Learning – Classificação
- Algoritmos de Machine Learning – Regressão
- Métodos Ensemble
- Algoritmo XGBoost



# Introdução



**AVISO  
IMPORTANTE**

Este capítulo não é um curso de Machine Learning, mas sim uma introdução ao tema. O curso de Machine Learning é o número 4 da Formação Cientista de Dados e aborda o tema em profundidade. Nosso objetivo aqui é que você aprenda como aplicar Machine Learning, pois mais a frente no curso aplicaremos aprendizado de máquina em dados gerados em tempo real.



Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

## Algoritmos de Machine Learning





# Algoritmos de Machine Learning

Hoje, como Cientista de Dados, é possível construir sistemas que trituram dados com algoritmos complexos, tudo com baixo custo e alta capacidade de processamento.







# Algoritmos de Machine Learning

```
152 document.getElementById( 'bigimageDesc' ).innerHTML = descriptions[ page * 9 ];  
153 }  
154  
155 function updatePhotoDescription() {  
156   if ( descriptions.length > ( page * 9 ) + ( currentImage.substring( 0, 1 ) ) ) {  
157     document.getElementById( 'bigimageDesc' ).innerHTML = descriptions[ page * 9 ];  
158   }  
159 }  
160  
161 function updateAllImages() {  
162   var i = 1;  
163   while ( i < 10 ) {  
164     var elementId = 'foto' + i;  
165     var elementIdBig = 'bigimage' + i;  
166     if ( page * 9 + i - 1 < photos.length ) {  
167       document.getElementById( elementId ).src = 'images/' + photos[ page * 9 + i - 1 ];  
168       document.getElementById( elementIdBig ).src = 'images/' + photos[ page * 9 + i - 1 ];  
169     } else {  
170       document.getElementById( elementId ).src = '';  
171     }  
172     i++;  
173   }  
174 }
```

Machine Learning, ou seja a aplicação de algoritmos e ciência para extrair informações de dados, é um dos campos mais espetaculares da ciência da computação atualmente.

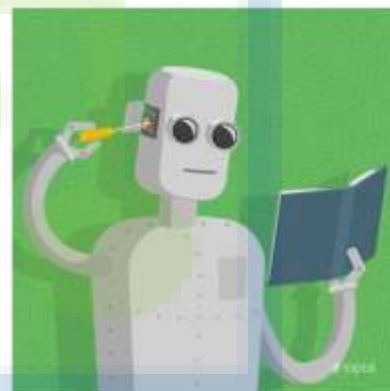


# Algoritmos de Machine Learning

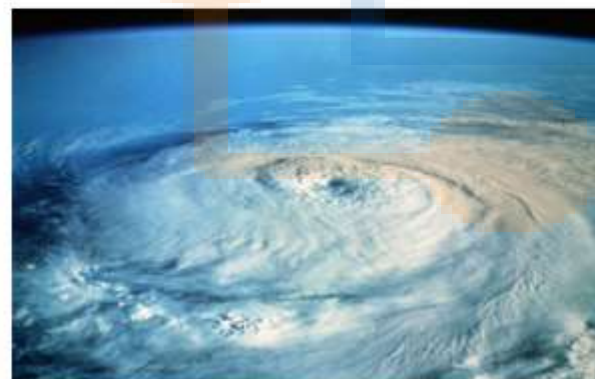
Big Data



Machine Learning

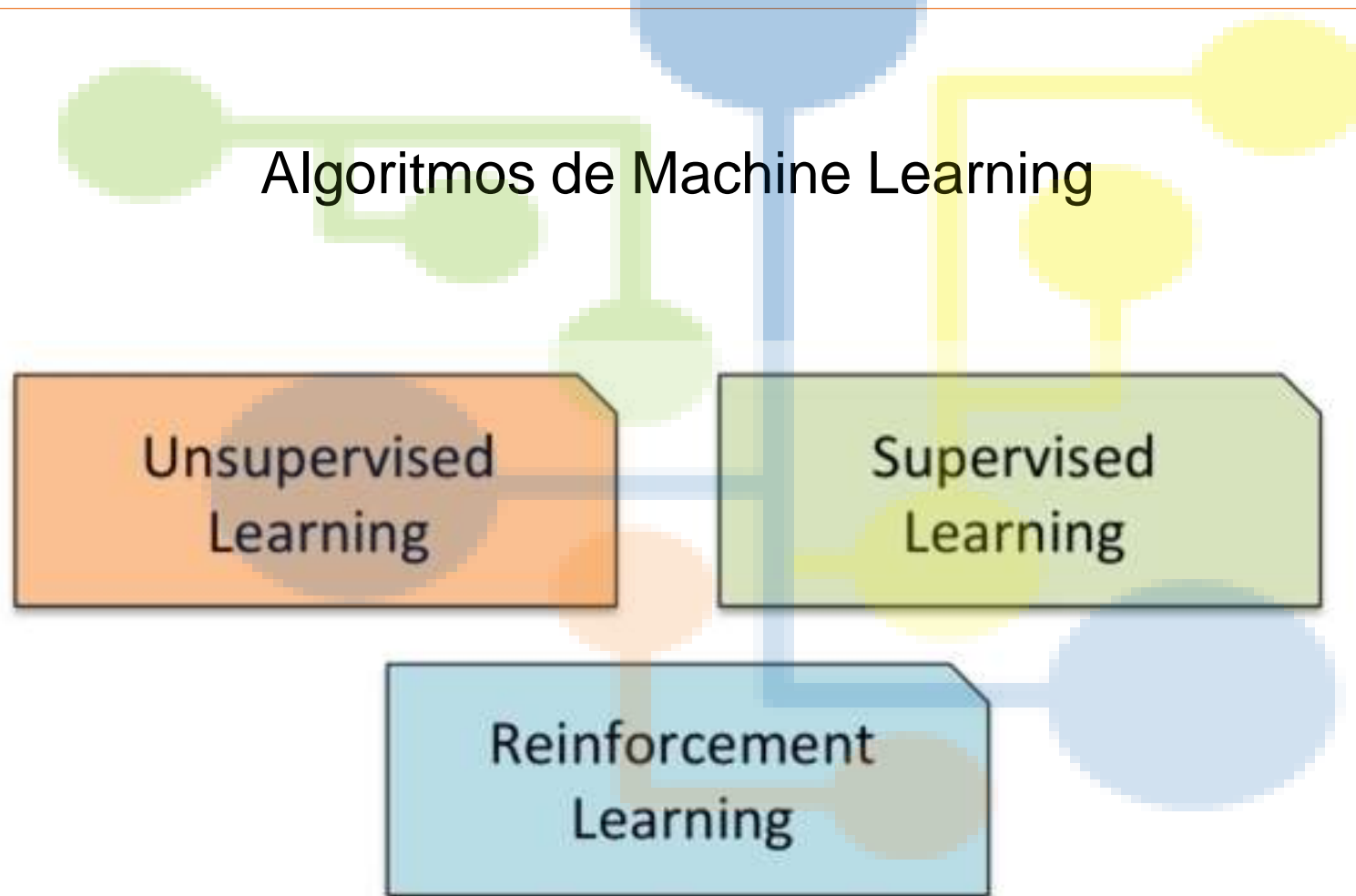


Capacidade Computacional





# Algoritmos de Machine Learning





Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

## Aprendizagem Supervisionada







# Aprendizagem Supervisionada



Como uma criança aprende?

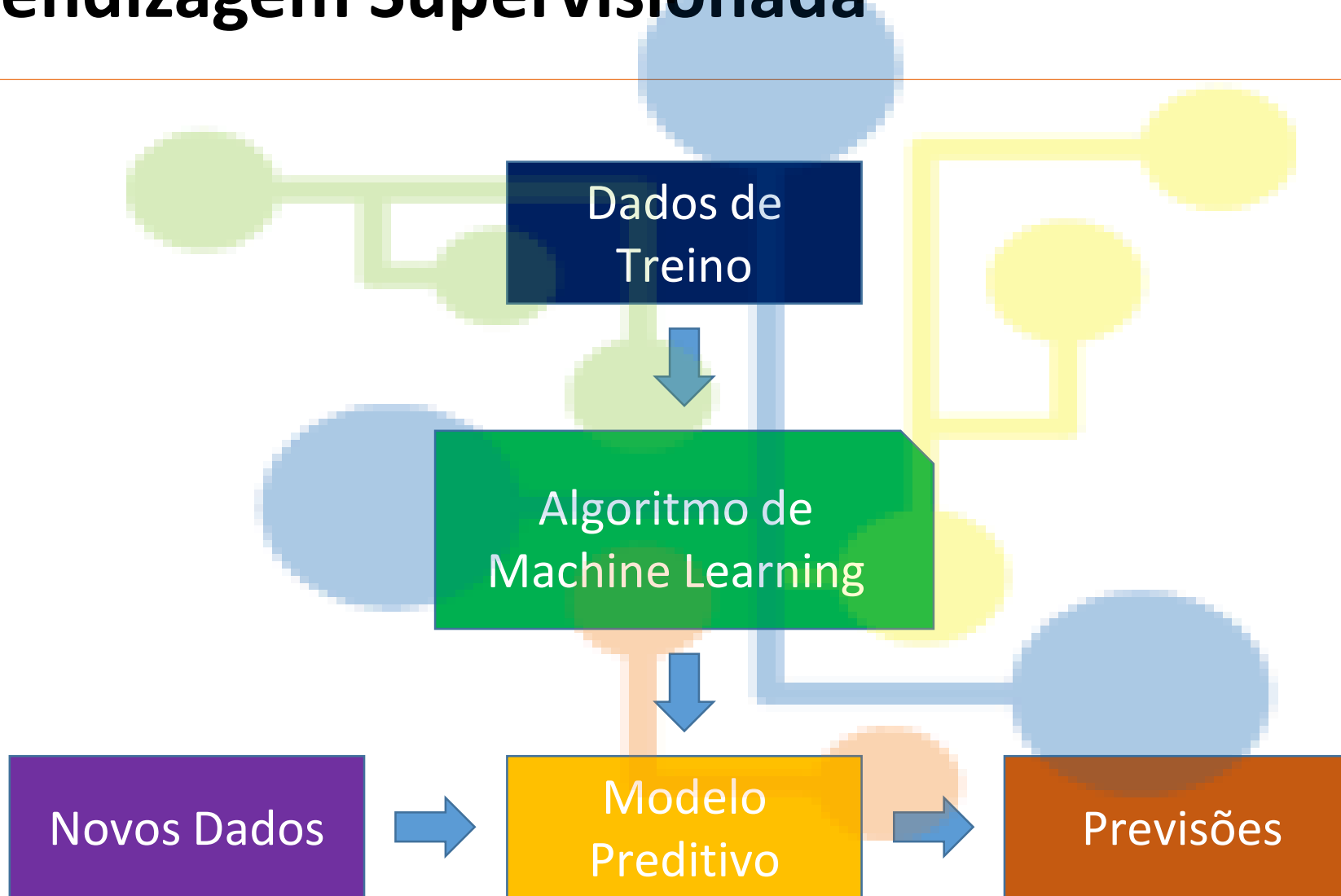
Um professor apresenta imagens, textos ou objetos informando para a criança o que aquilo representa.

Por exemplo, o professor apresenta a foto de um carro, explicando suas principais características.

Mais tarde, quando a criança encontrar algo com as mesmas características será capaz de reconhecer que se trata de um carro.



# Aprendizagem Supervisionada





Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

## Aprendizagem Não Supervisionada





# Aprendizagem Não Supervisionada

Na aprendizagem supervisionada, temos as entradas (as características) e temos as saídas. O algoritmo então aprende o relacionamento nos dados e um modelo é criado. Quando o modelo é apresentado a novos dados de entrada, é capaz de prever as saídas.

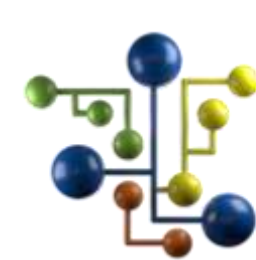




# Aprendizagem Não Supervisionada

Na aprendizagem supervisionada, temos as entradas (as características) e temos as saídas. O algoritmo então aprende o relacionamento nos dados e um modelo é criado. Quando o modelo é apresentado a novos dados de entrada, é capaz de prever as saídas.

Mas e quando não temos os dados de saída?



# Aprendizagem Não Supervisionada

Atributo 1	Atributo 2	Atributo 3	Saída
x1	x2	x3	Carro
x4	x5	x6	Avião
x7	x8	x9	?
x10	x11	x12	?

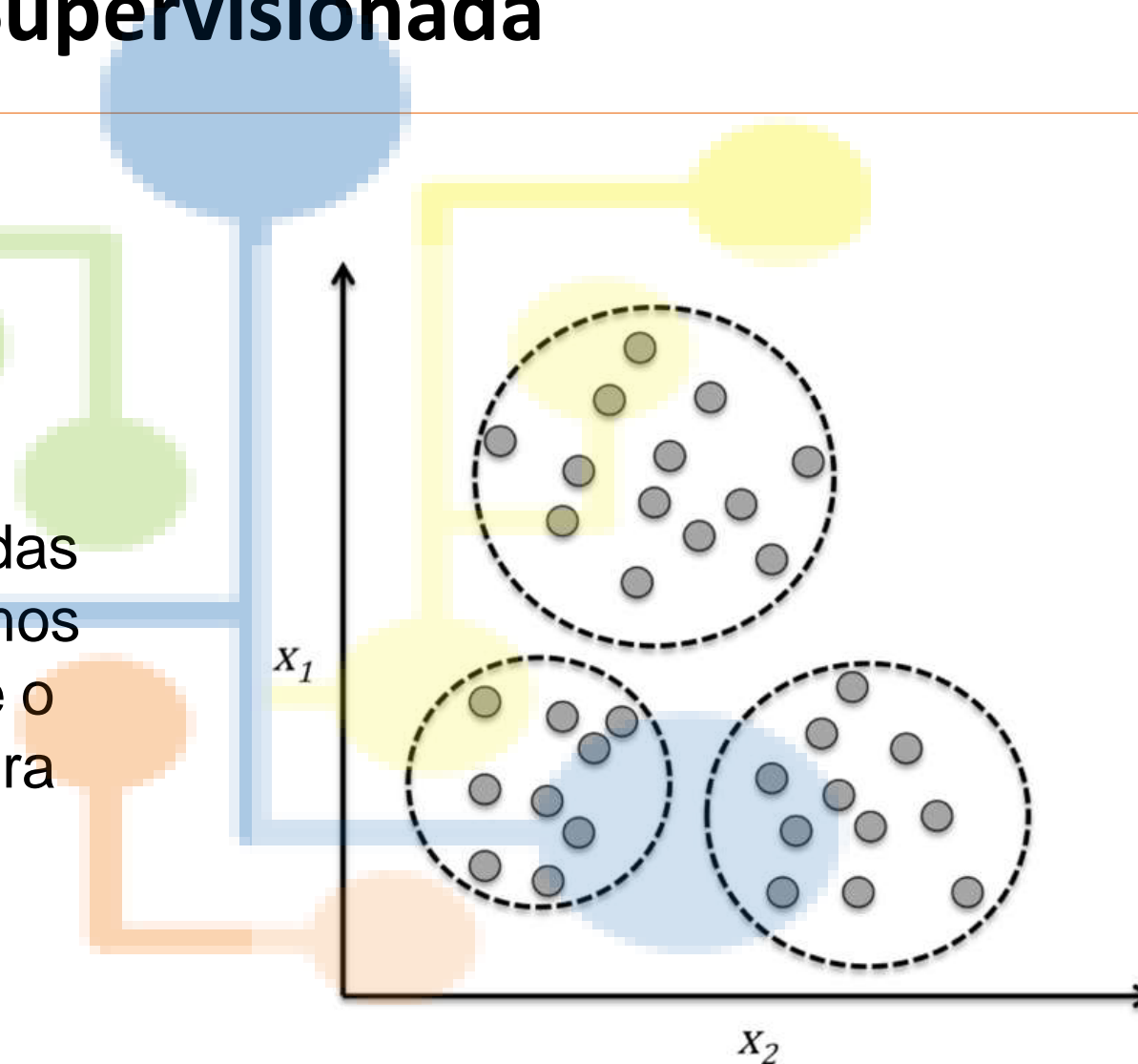
Aprendizagem Supervisionada

Aprendizagem Não Supervisionada



# Aprendizagem Não Supervisionada

Na aprendizagem não supervisionada, temos as entradas (as características) mas não temos as saídas. O algoritmo aprende o relacionamento nos dados e gera agrupamentos (clusters).





Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

## Aprendizagem Por Reforço







# Aprendizagem Por Reforço



Como alguém aprende a andar de bicicleta?

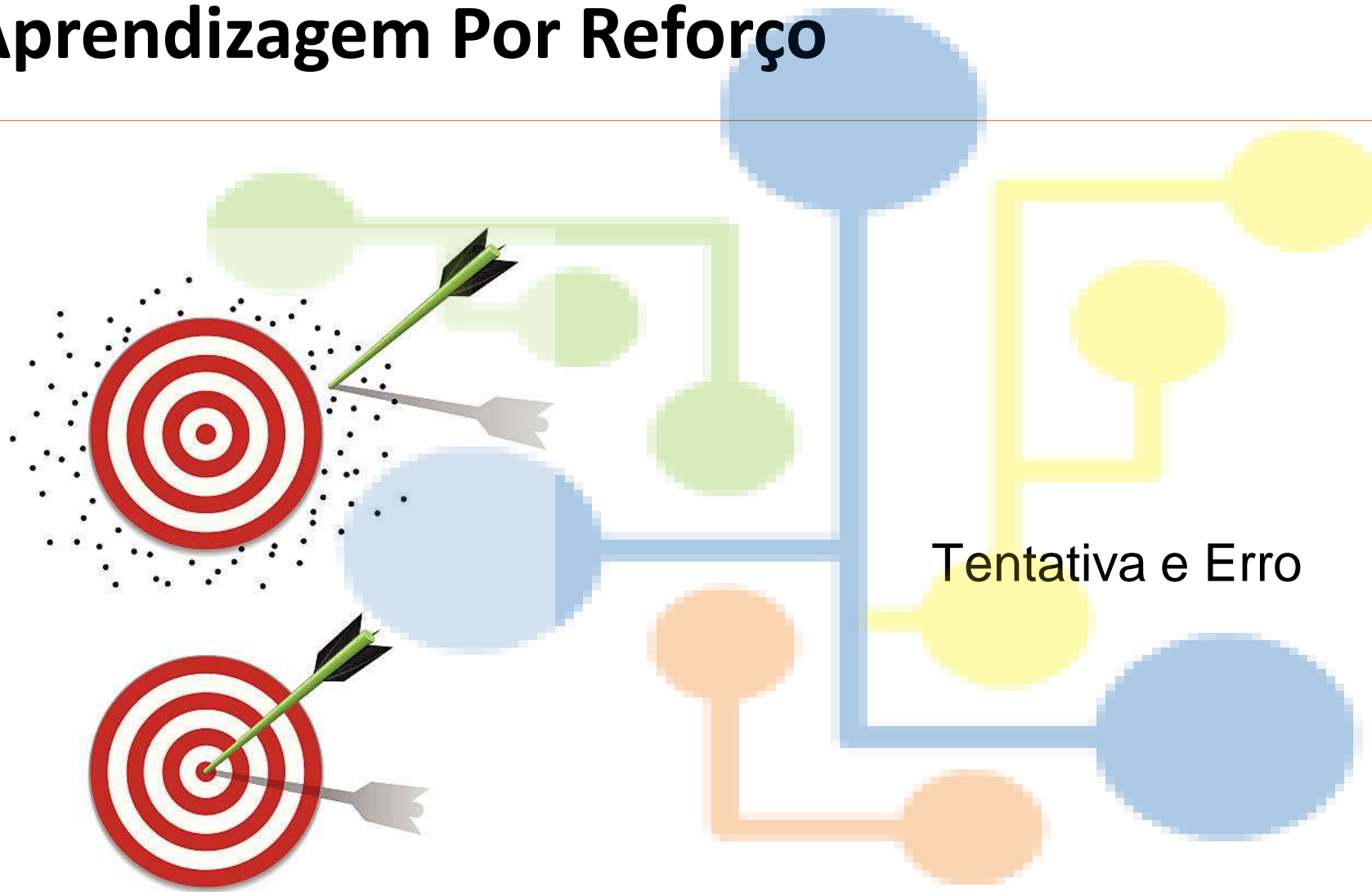
Podemos usar a aprendizagem supervisionada?

E a aprendizagem não supervisionada?

Qual seria o melhor método neste caso?



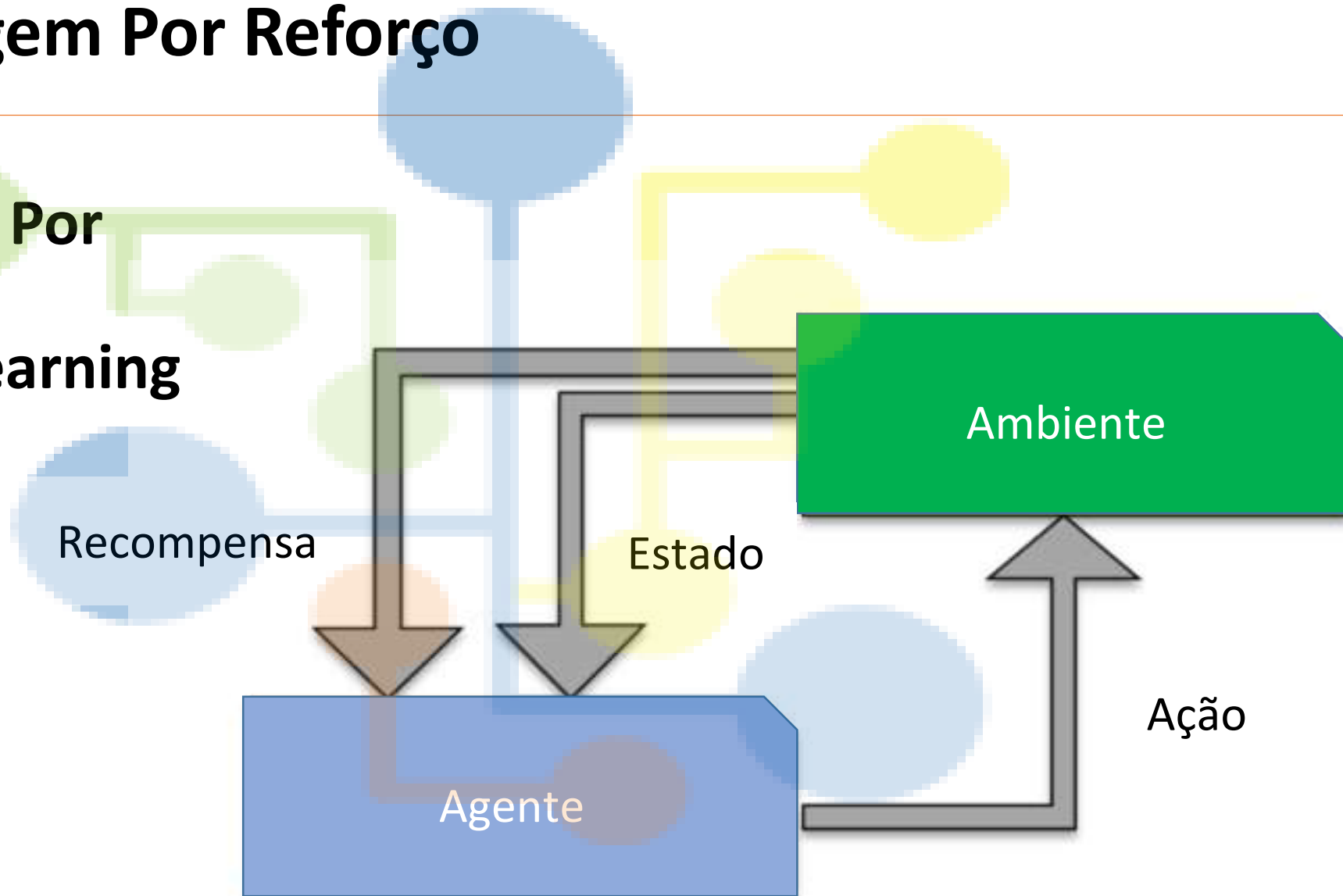
# Aprendizagem Por Reforço





# Aprendizagem Por Reforço

## Aprendizagem Por Reforço ou Reinforcement Learning





Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

## Principais Algoritmos de Machine Learning





# Principais Algoritmos de Machine Learning

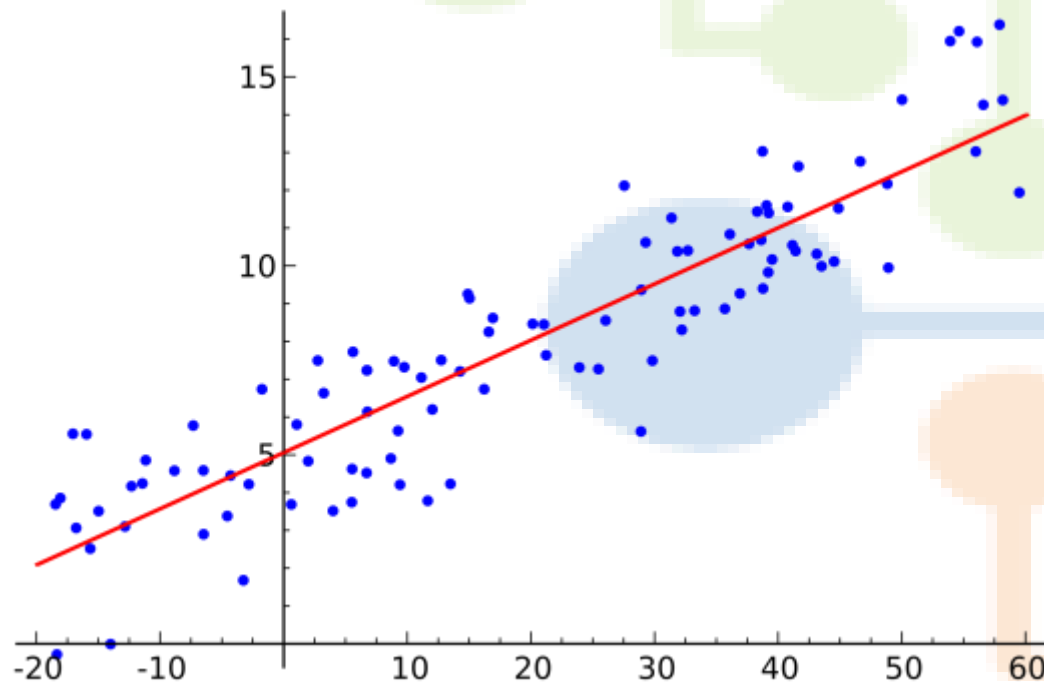
Quando trabalhamos com aprendizagem supervisionada temos basicamente 2 tipos de algoritmos:







# Principais Algoritmos de Machine Learning

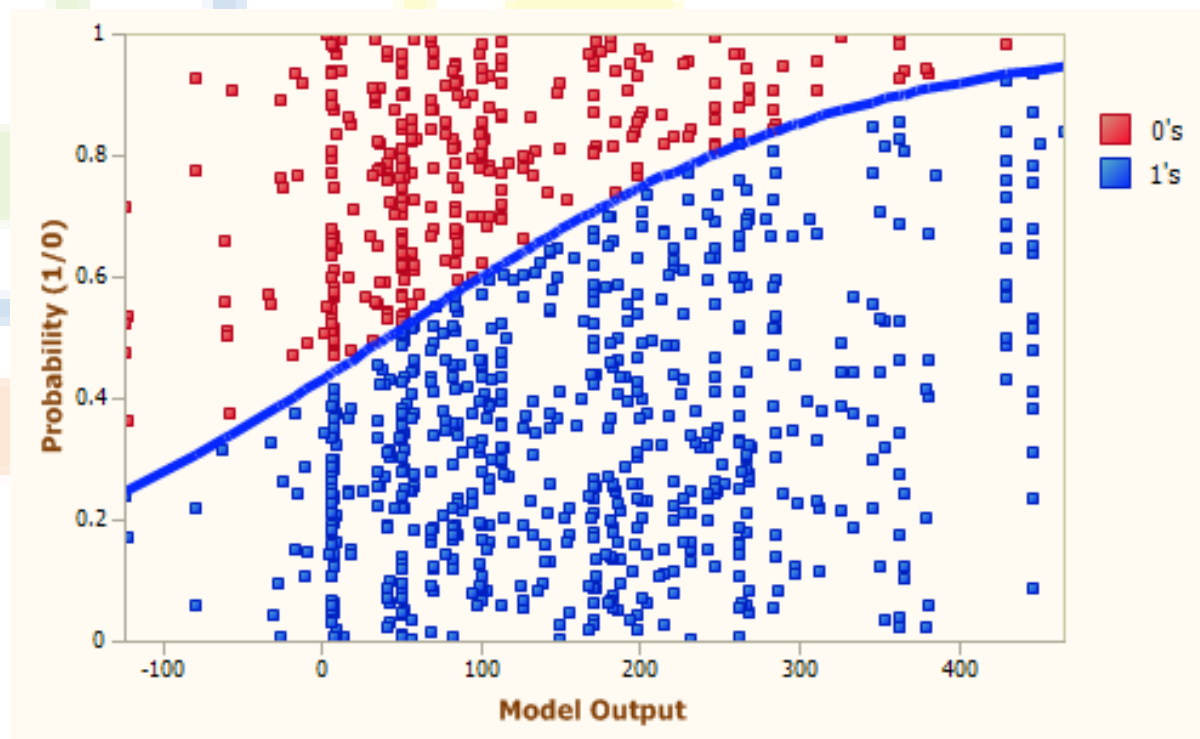


**Regressão Linear**



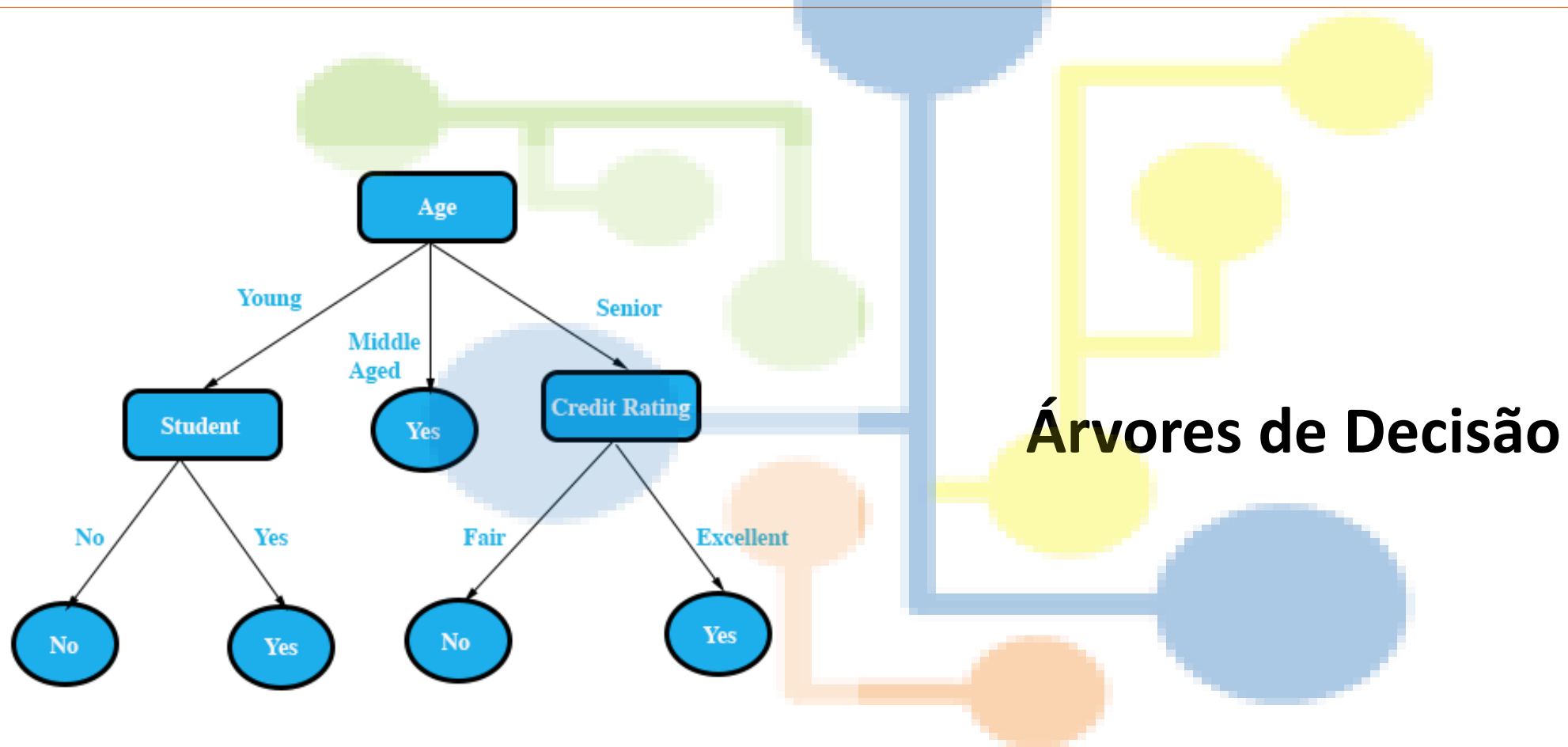
# Principais Algoritmos de Machine Learning

## Regressão Logística





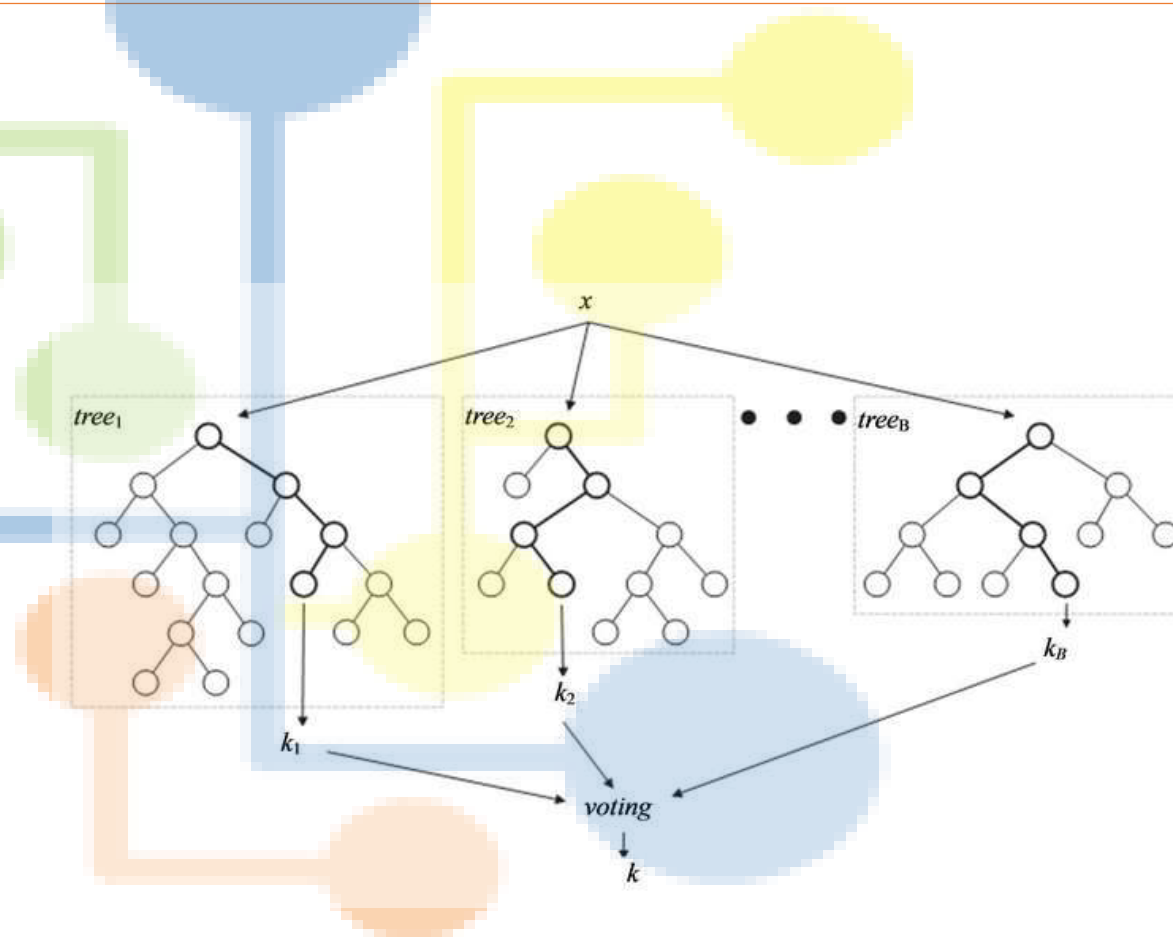
# Principais Algoritmos de Machine Learning





# Principais Algoritmos de Machine Learning

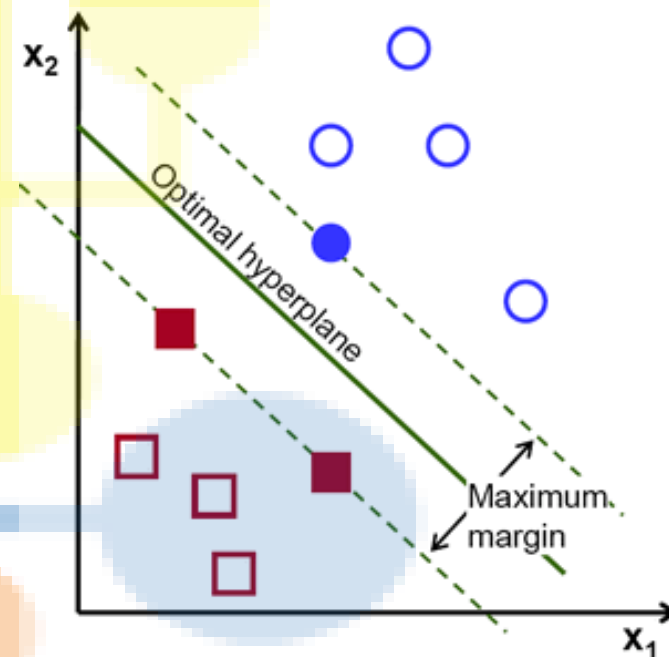
## Random Forest (Floresta Aleatória)





# Principais Algoritmos de Machine Learning

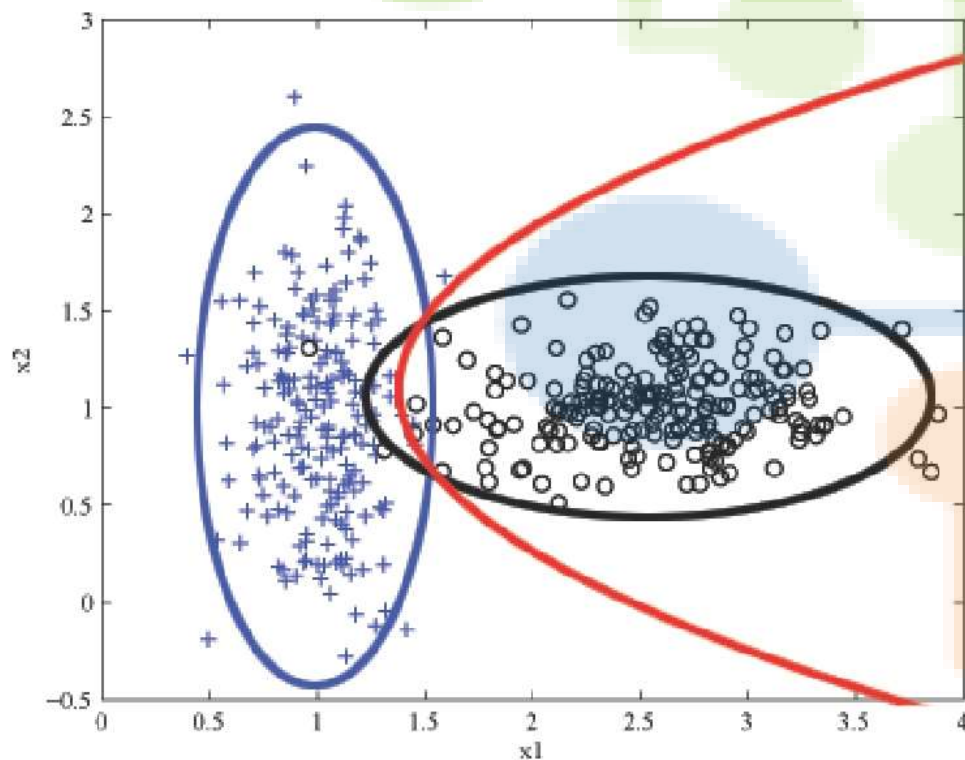
## Support Vector Machines (Máquinas de Vetor de Suporte)







# Principais Algoritmos de Machine Learning

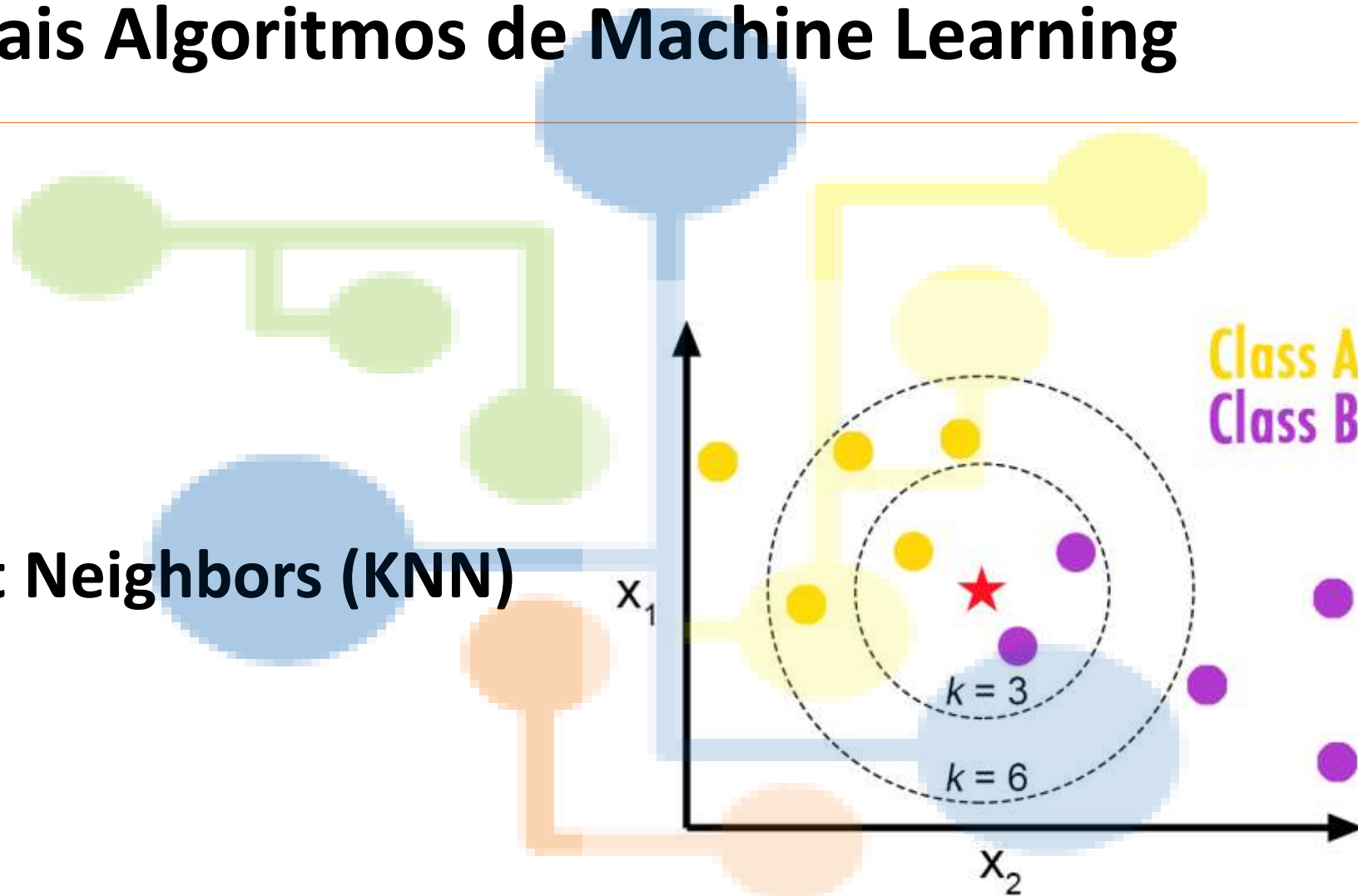


**Naive Bayes**



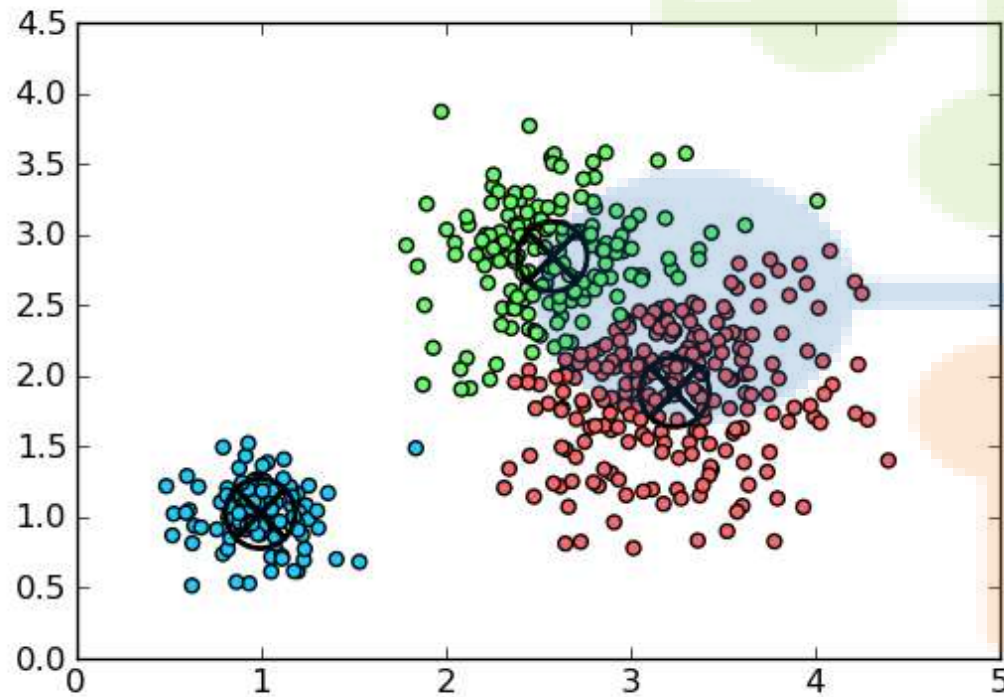
# Principais Algoritmos de Machine Learning

## K-Nearest Neighbors (KNN)





# Principais Algoritmos de Machine Learning

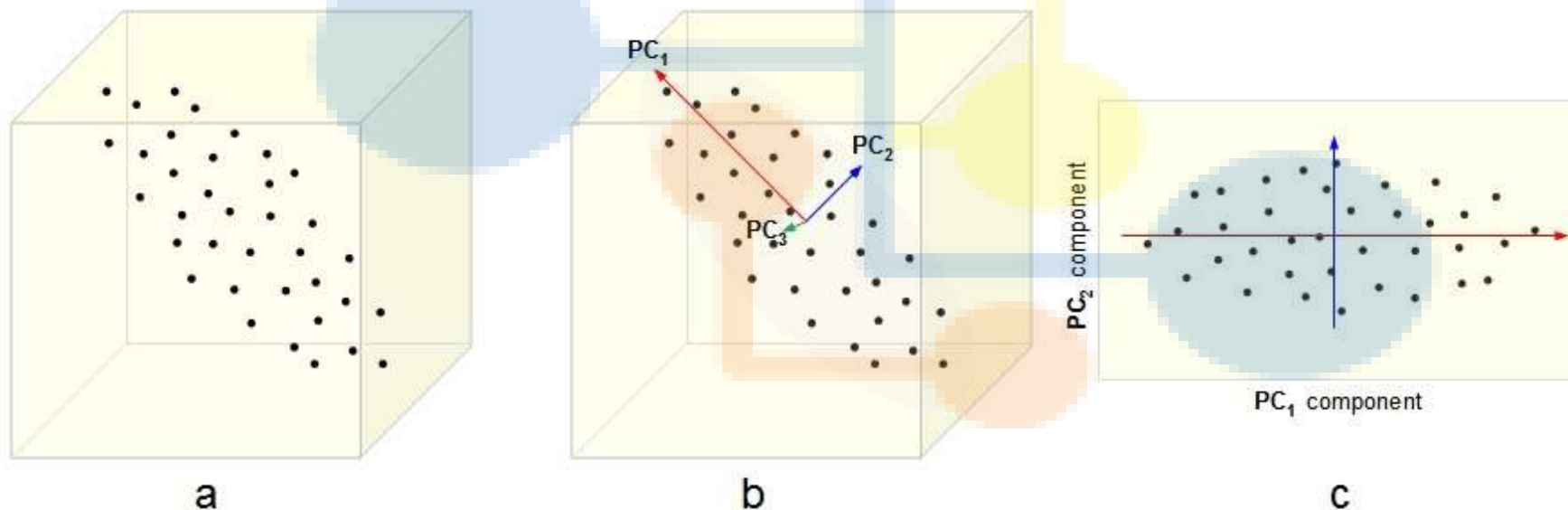


**K-Means**



# Principais Algoritmos de Machine Learning

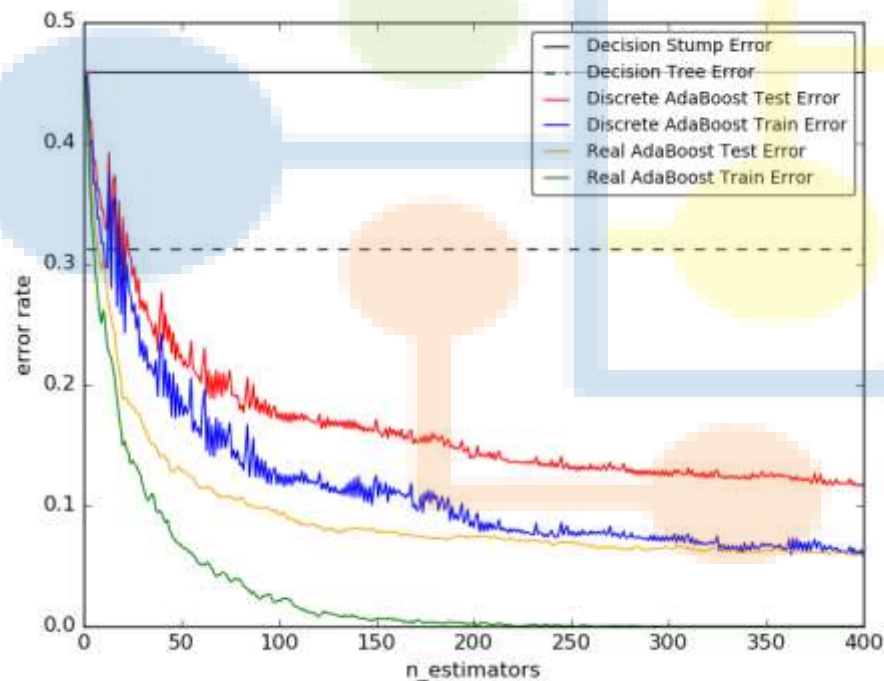
## Algoritmo para Redução de Dimensionalidade





# Principais Algoritmos de Machine Learning

## Gradient Boosting & AdaBoost







Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a





Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

## O Processo de Construção de Modelos de Machine Learning

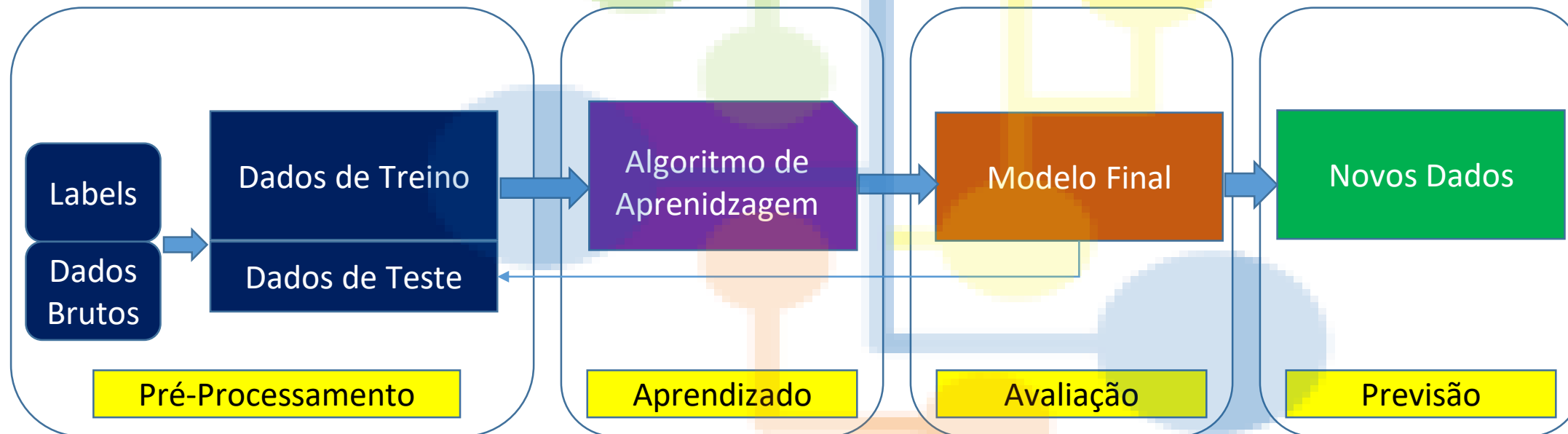




# O Processo de Construção de Modelos de Machine Learning

- Transformação de Variáveis
- Feature Selection
- Redução de Dimensionalidade
- Amostragem

- Validação do Modelo
- Otimização



- Seleção do Modelo
- Cross-Validation
- Métricas de Performance
- Otimização



Data Science  
Academy

Data Science Academy leandro.cotica@sicoob.com.br 5caf62505e4cded5a28b456a

# Big Data Real-Time Analytics com Python e Spark

## Soluções de Machine Learning





# Soluções de Machine Learning

Podemos construir modelos de Machine Learning de duas formas principais:

Desenvolvendo todo o algoritmo a partir do zero usando uma linguagem de programação.

Utilizando um framework pronto, onde os principais algoritmos já estão implementados.



# Soluções de Machine Learning

## Soluções de Machine Learning

Principais linguagens de programação para ML:

- Python
- Linguagem R
- Scala
- Java
- JavaScript
- Go
- C++ / C#

Principais frameworks para ML:

- Scikit-learn (Python)
- Caret (R)
- TensorFlow (Python, R, Java, C++)
- Apache Mahout (Python, Java)
- Spark Mllib (Scala, Java, Python, R)
- H2O (Java, Python)
- Weka (Java, Python)
- PyTorch, CNTK, MXNet (Python, C++, Java)





# Soluções de Machine Learning

A linguagem Python oferece duas vantagens principais sobre todas as outras soluções. Primeiro, por se tratar de uma linguagem de uso geral, ela pode ser usada com qualquer uma destas soluções. Segundo, Python possui uma das mais poderosas soluções gratuitas de Machine Learning, o Scikit-learn, que estudaremos neste capítulo.

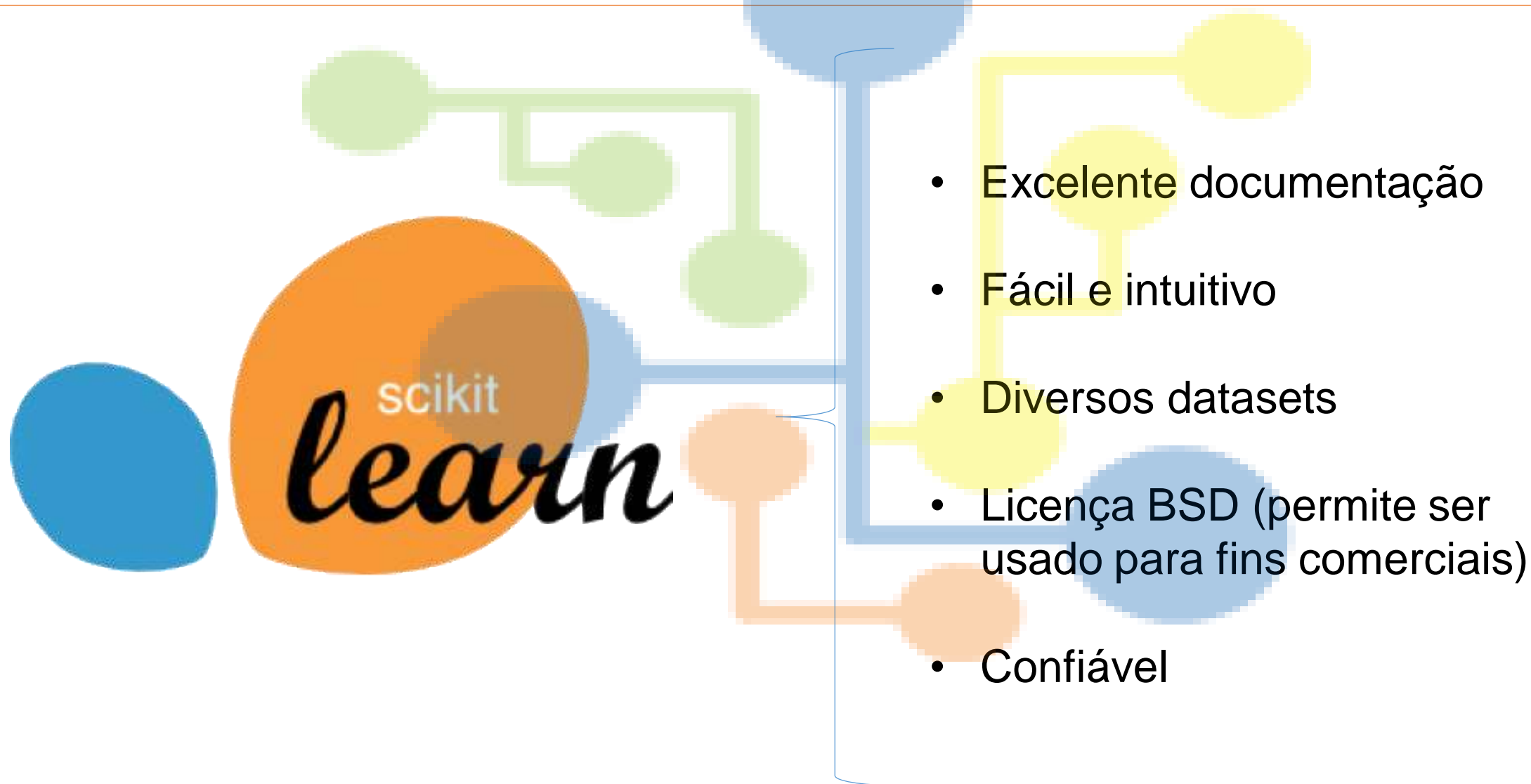


# Soluções de Machine Learning





# Soluções de Machine Learning





Data Science  
Academy

Data Science Academy leandro.cotica@data-science-academy.com.br 5caf62505e4cded5a28b456a



Tenha uma Excelente Jornada de Aprendizagem.

Muito Obrigado por Participar!

Equipe Data Science Academy