

# Big Data Analytics com R e Microsoft Azure Machine Learning



Data Science Academy

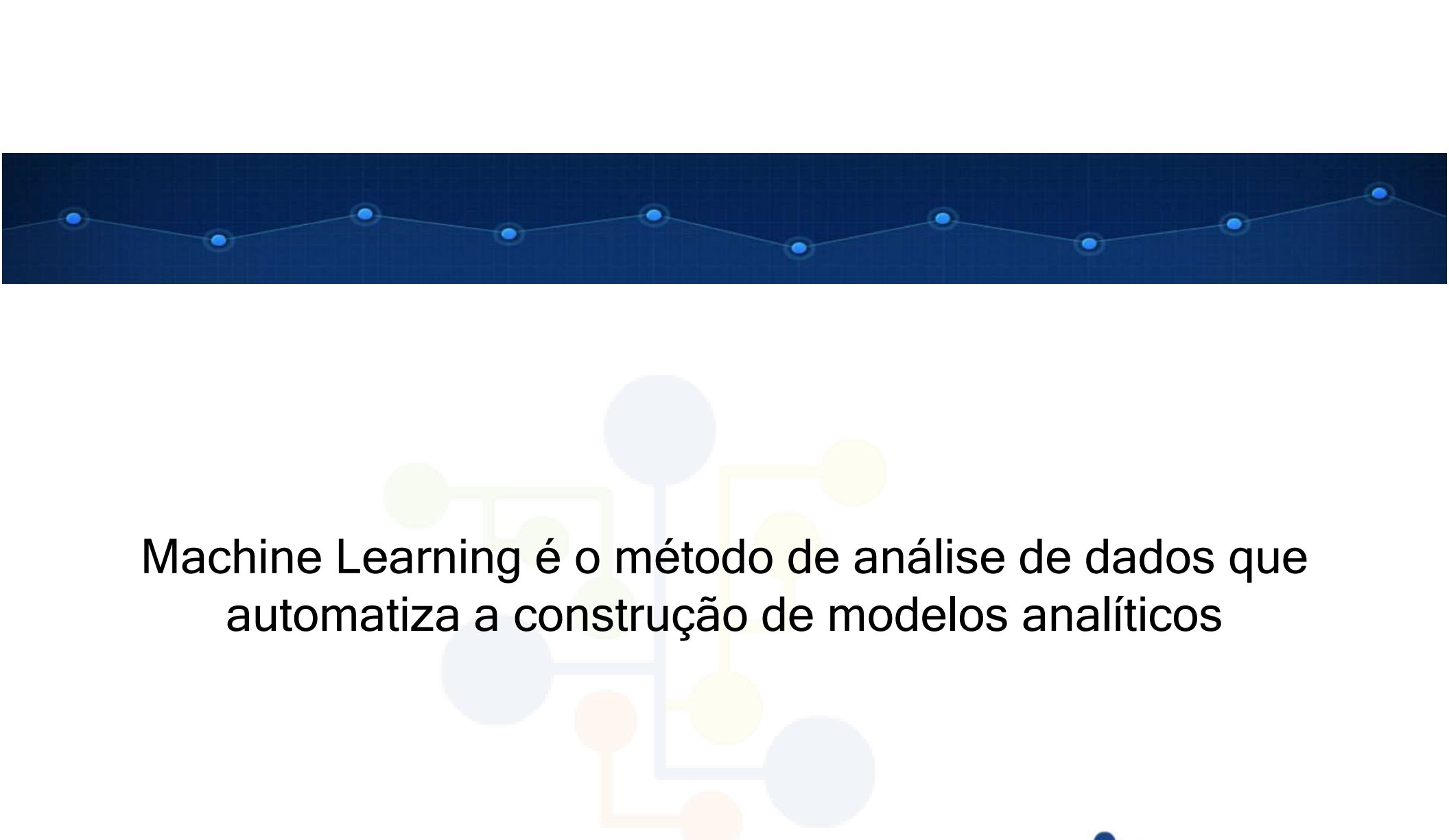


# Machine Learning



Data Science Academy

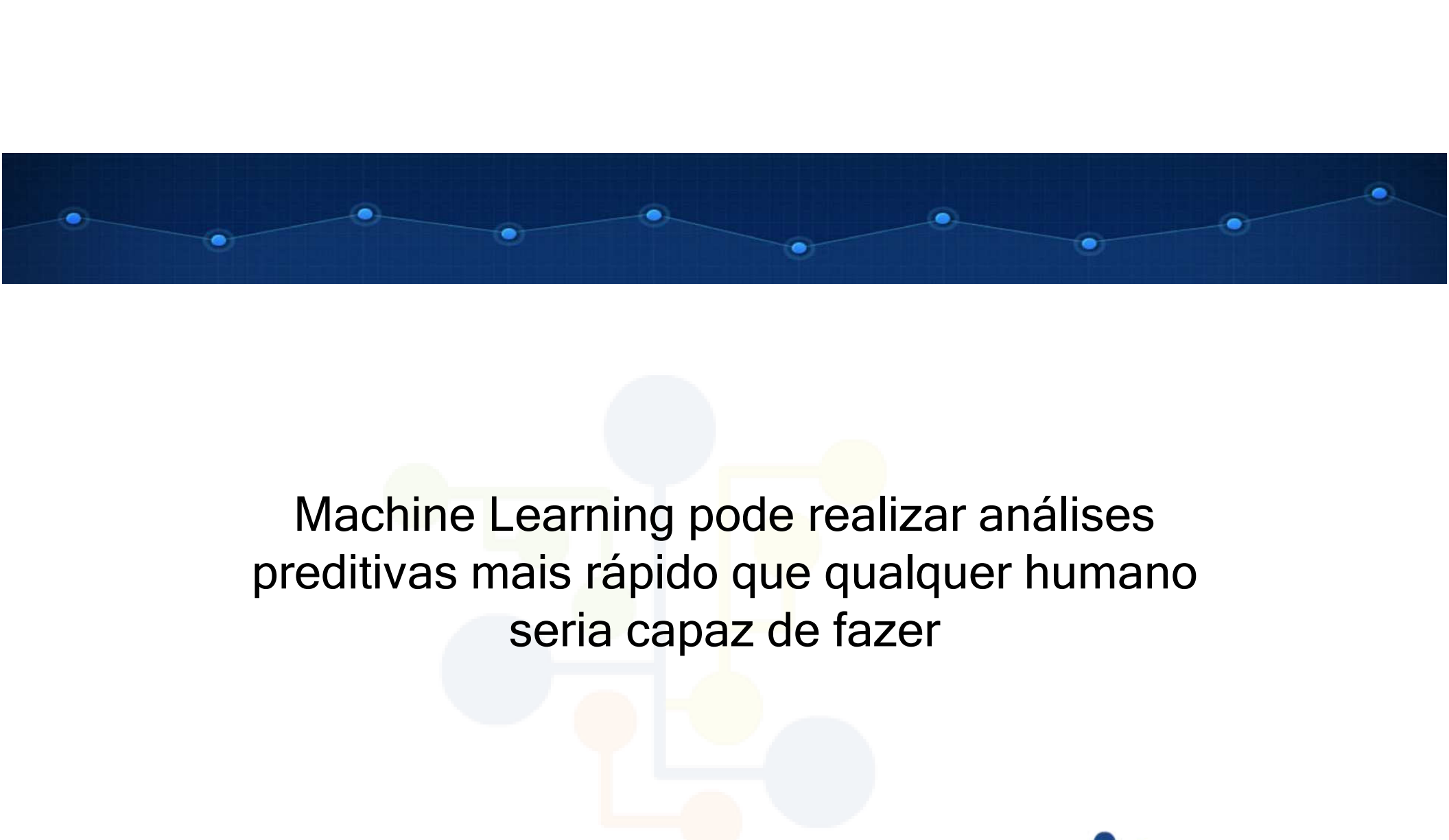
[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Machine Learning é o método de análise de dados que automatiza a construção de modelos analíticos



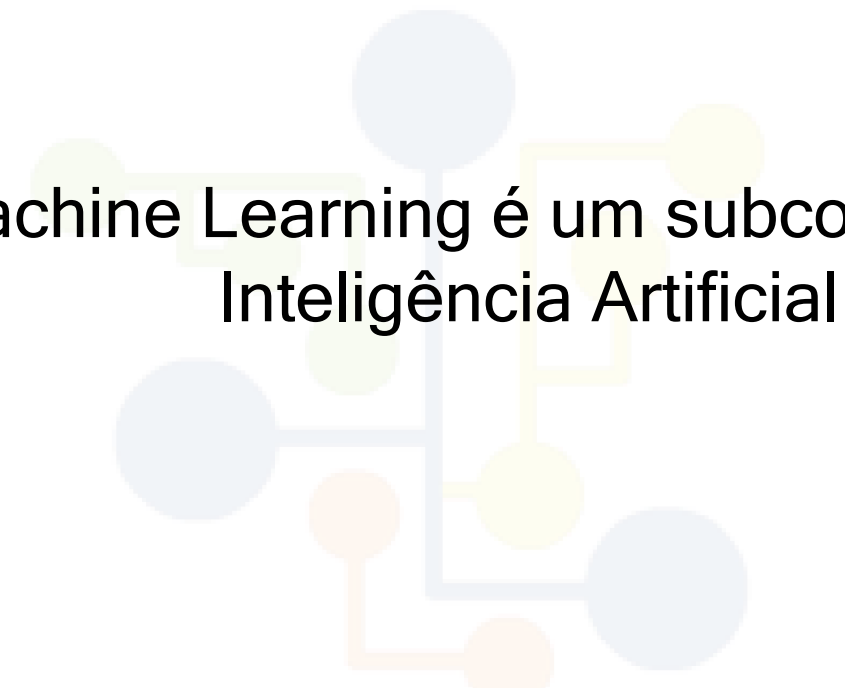

Data Science Academy



Machine Learning pode realizar análises  
preditivas mais rápido que qualquer humano  
seria capaz de fazer



Data Science Academy




Machine Learning é um subconjunto da  
Inteligência Artificial



Data Science Academy


[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Inteligência Artificial inclui Machine Learning,  
mas Machine Learning por si só não define  
Inteligência Artificial



Data Science Academy



# Inteligência Artificial é baseada em Machine Learning e Machine Learning é essencialmente diferente de Estatística

Mas é baseado na estatística



Data Science Academy



| Técnica                | Estatística   | Machine Learning  |
|------------------------|---|---|
| Entrada de Dados       | Os parâmetros interpretam fenômenos da vida real e trabalham a magnitude. | Os dados são randomizados e transformados para aumentar a acurácia de análises preditivas.                      |
| Tratamento de Dados    | Modelos são usados para previsões em amostras pequenas.                   | Trabalha com Big Data na forma de redes e gráficos. Os dados são divididos em dados de treino e dados de teste. |
| Resultado              | Captura a variabilidade e a incerteza dos parâmetros.                     | Probabilidade é usada para comparações e para buscar as melhores decisões.                                      |
| Distribuição dos Dados | Assumimos uma distribuição bem definida dos dados.                        | A distribuição dos dados é desconhecida ou ignorada antes do processo de aprendizagem.                          |
| Objetivos              | Assumimos um determinado resultado e então tentamos prová-lo.             | Os algoritmos aprendem a partir dos dados.  |



Data Science Academy





Machine Learning se baseia em alguns importantes conceitos da Matemática e da Estatística:




Manipulação de Matrizes

Teoria da Probabilidade

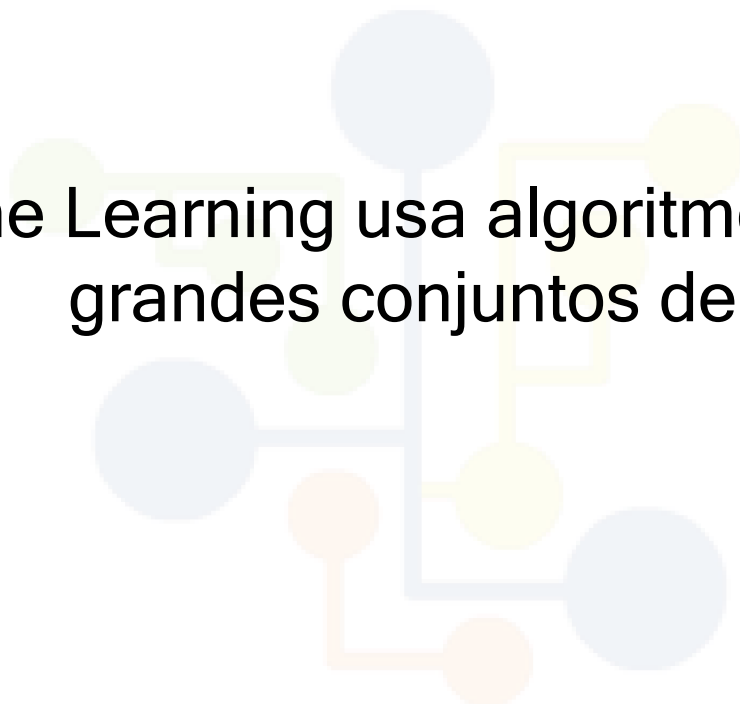
Teorema de Bayes



Data Science Academy

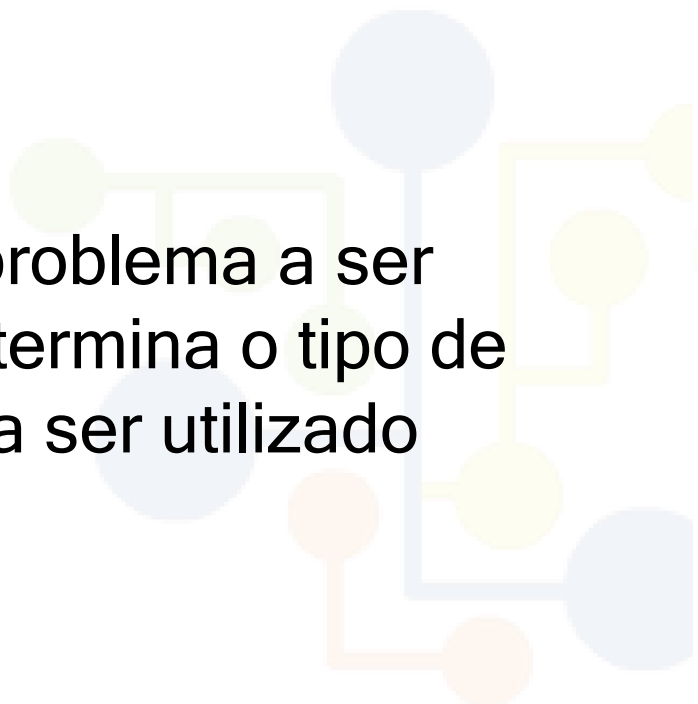



Machine Learning usa algoritmos para analisar  
grandes conjuntos de dados

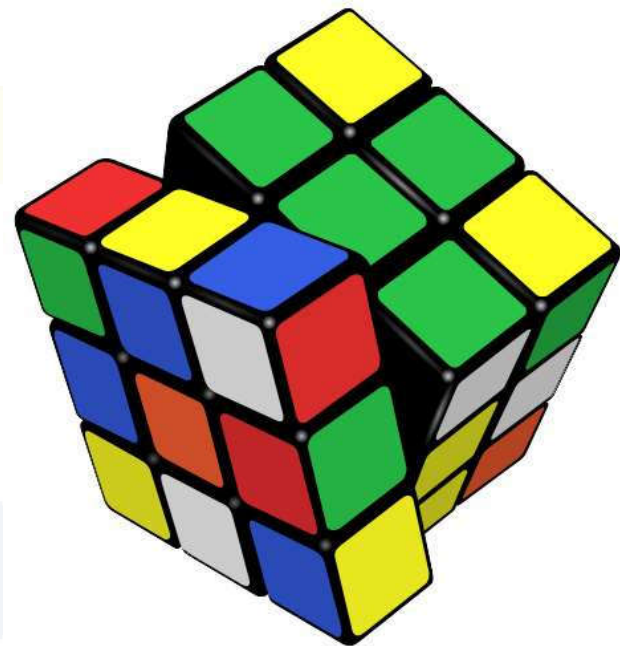


Data Science Academy

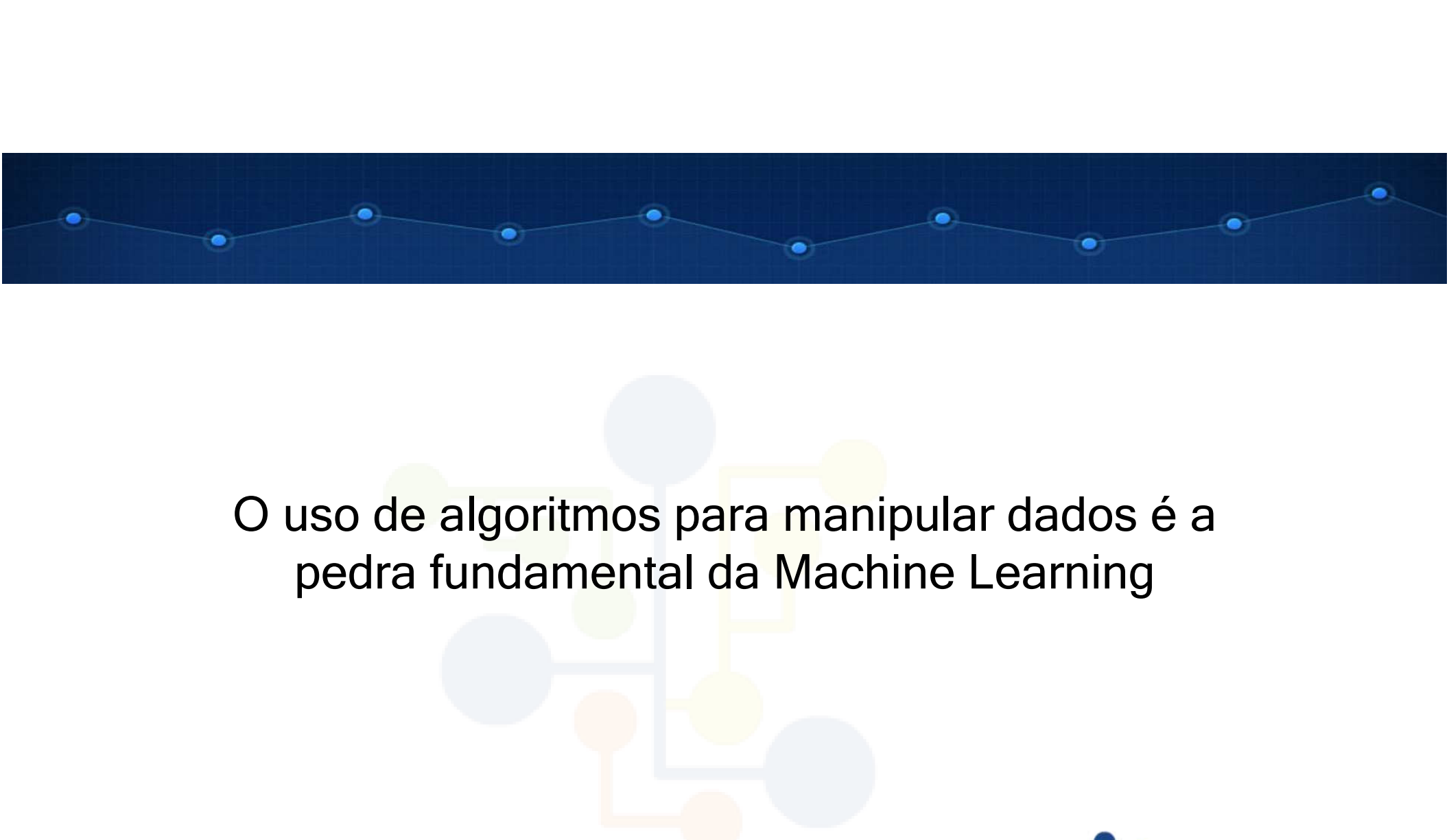
[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



O tipo de problema a ser resolvido, determina o tipo de algoritmo a ser utilizado



Data Science Academy



O uso de algoritmos para manipular dados é a  
pedra fundamental da Machine Learning



Data Science Academy

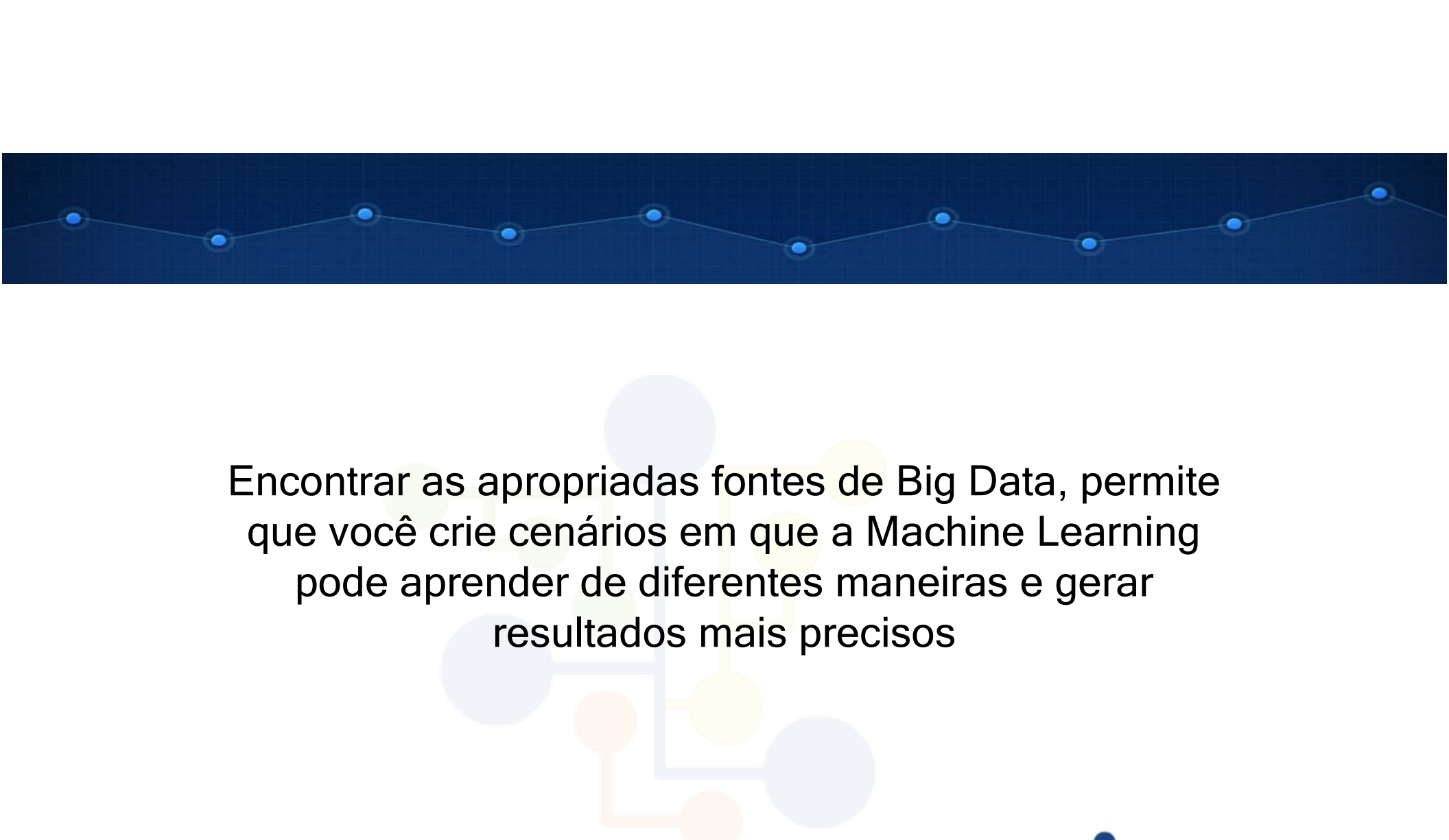


# Machine Learning x Data Mining

Foco é  
identificar  
padrões e  
solucionar  
problemas



Data Science Academy



Encontrar as apropriadas fontes de Big Data, permite  
que você crie cenários em que a Machine Learning  
pode aprender de diferentes maneiras e gerar  
resultados mais precisos




Data Science Academy



Big Data não é apenas um grande conjunto de dados, mas também uma grande variedade



Data Science Academy



Machine Learning é a chave que permite  
compreender o que está guardado no Big Data

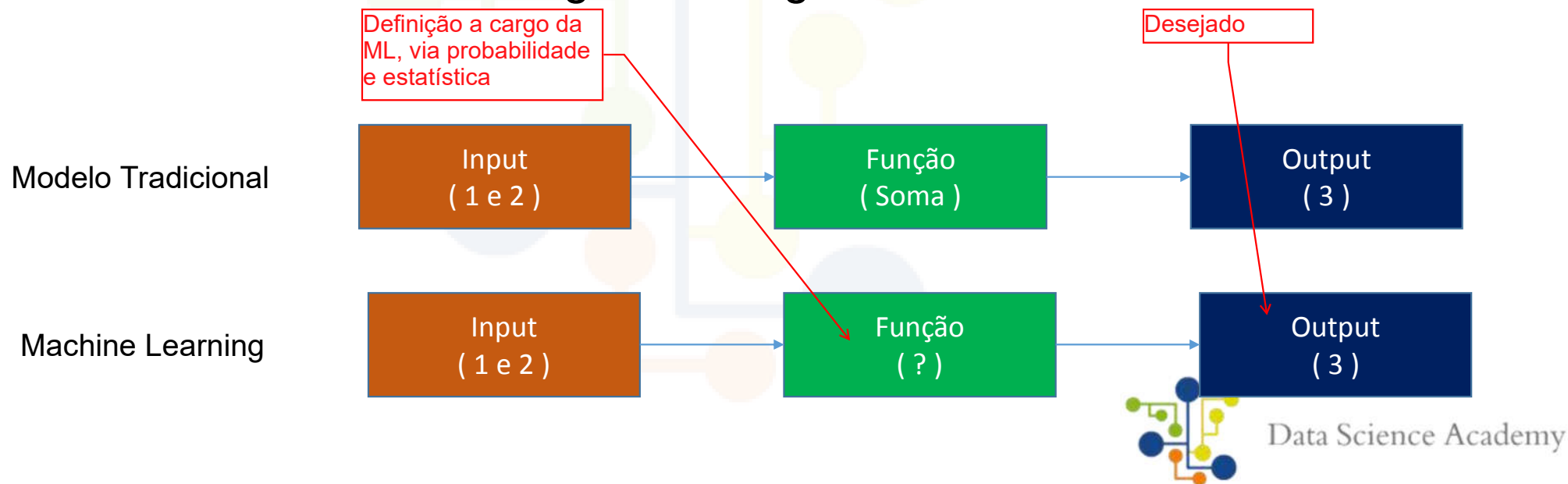


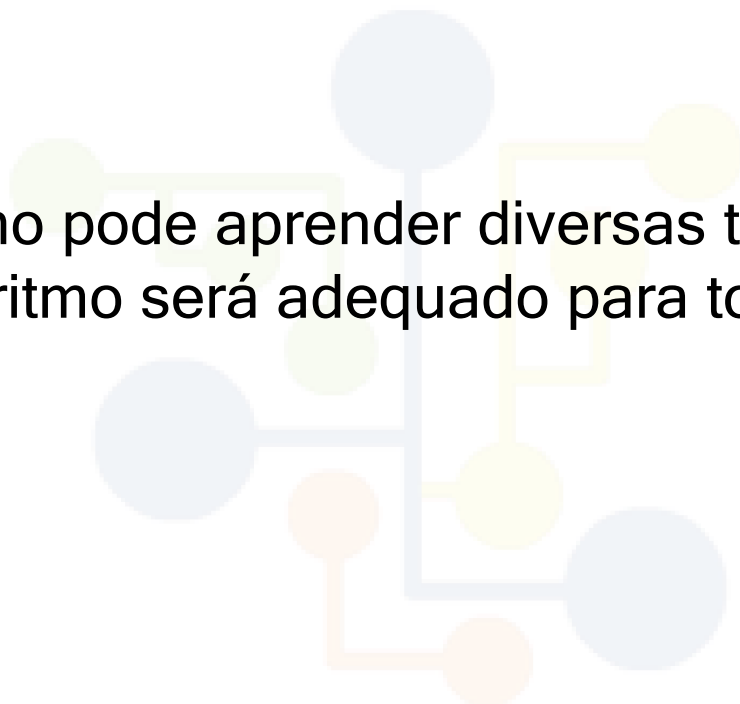

Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



## O processo de treino de algoritmos de Machine Learning usa o seguinte conceito

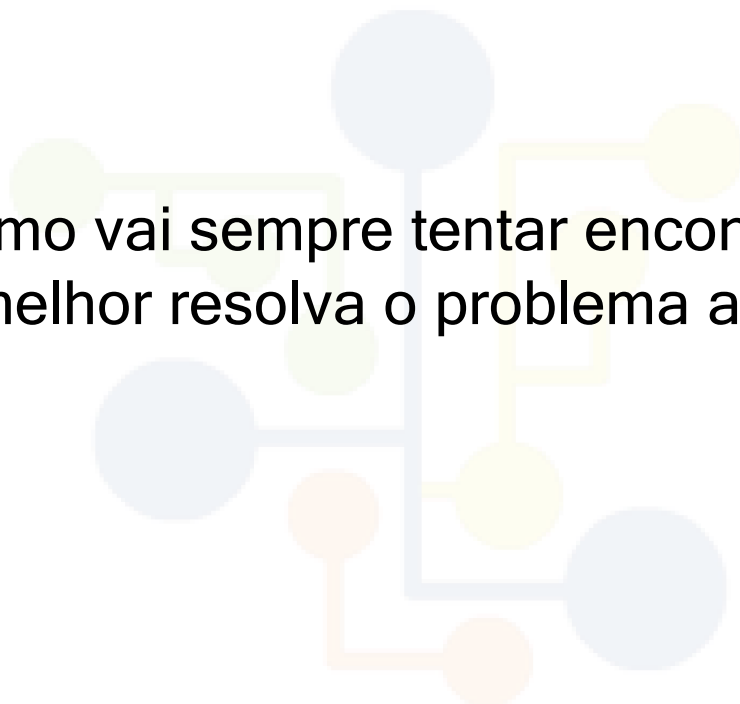




Um algoritmo pode aprender diversas tarefas, mas nem todo algoritmo será adequado para todas as tarefas



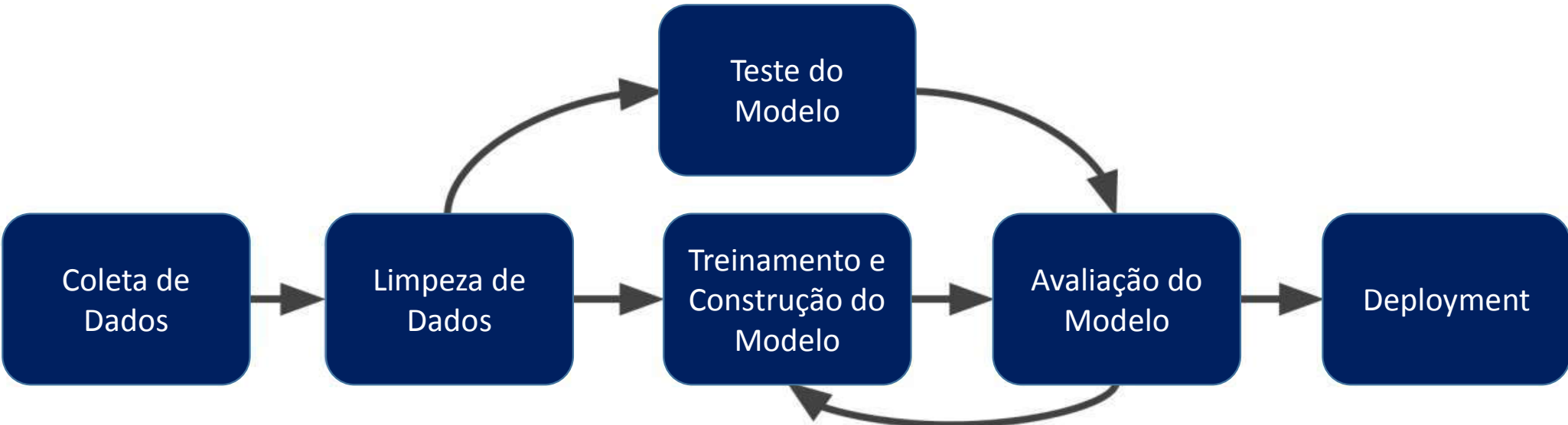
Data Science Academy



Um algoritmo vai sempre tentar encontrar uma função que  
melhor resolva o problema apresentado



Data Science Academy





# Machine Learning Frameworks

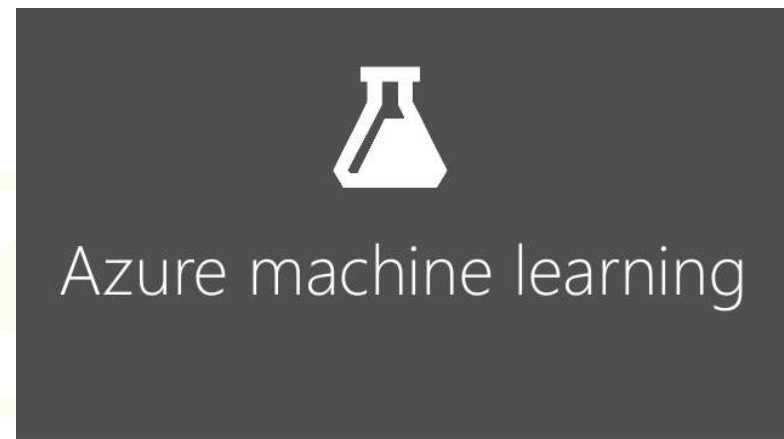
Um framework é um conjunto de softwares que produzem um resultado específico



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)

# Microsoft Azure Machine Learning

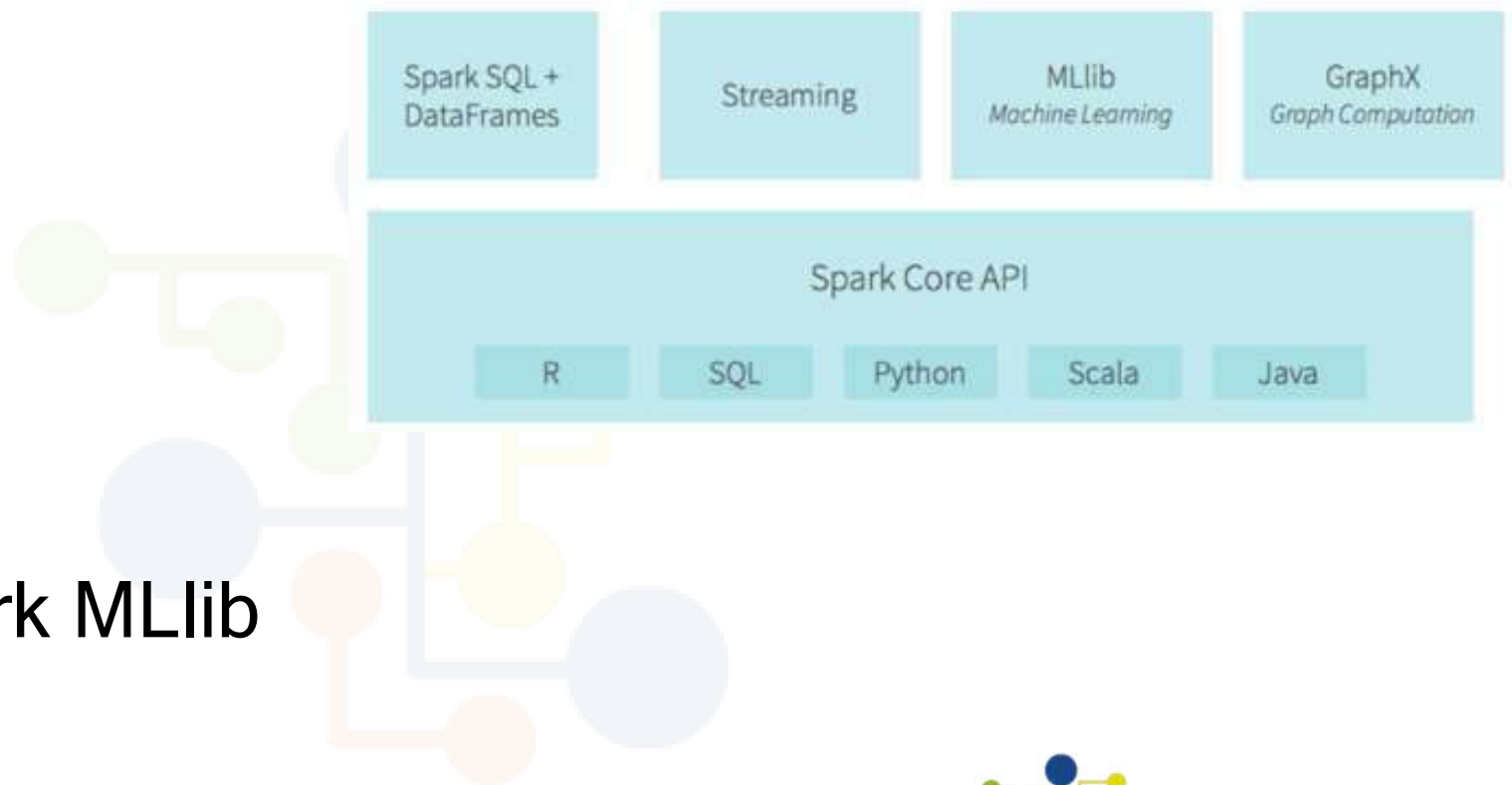


O Microsoft Azure ML será alvo de estudo nos próximos capítulos. Ele é um serviço em nuvem (Cloud) que tem como objetivo implementar modelos de Machine Learning de forma rápida e fácil. Com o Azure Machine Learning é possível construir modelos de análise preditiva, usando datasets de treino das mais variadas fontes e então fazer o deploy destes modelos através de web services com o serviço Cloud da Microsoft. Com o Azure Machine Learning Studio, é possível criar experimentos de dados, usando os módulos disponíveis ou construindo seus próprios modelos usando R, Python e SQL por exemplo.



Data Science Academy

# Apache Spark MLlib



Data Science Academy

# Apache Singa

Para deep learning, usado em processamento de linguagem natural e reconhecimento de imagens. Pode ser um pouco lento



Data Science Academy



# Google Tensor Flow



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# Caffe

Outro para deep learning, alta capacidade para  
processar muitas imagens

# Caffe



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# Nervana

Nervana neon, para deep learning.  
Foco no hardware.



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# Outras Ferramentas



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# Weka

Waikato Environment for Knowledge Analysis (Weka)



Feito em Java. Uso em data mining



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# O Processo de Aprendizagem



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)





# Processo de Aprendizagem

O Processo de Aprendizagem ocorre de diferentes formas e podemos dividir os algoritmos de Machine Learning em 3 grupos principais:

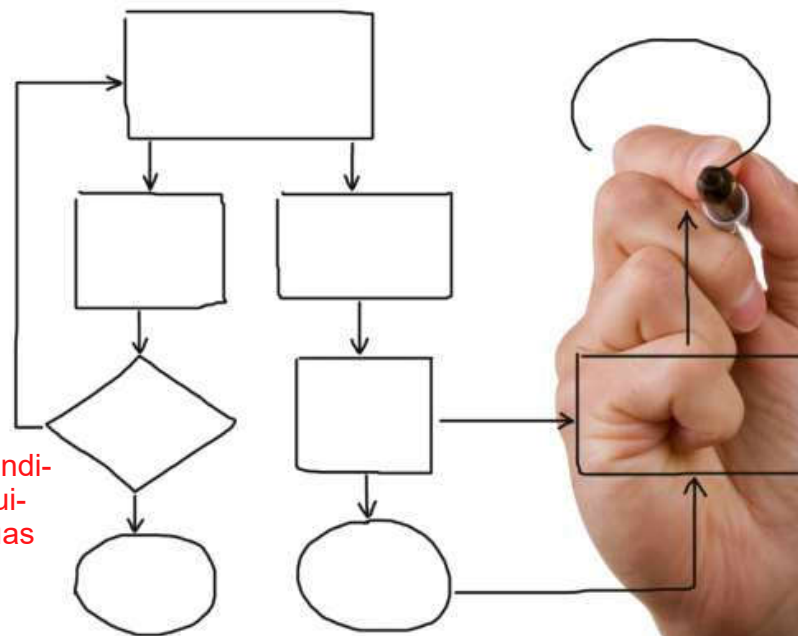
Aprendizagem Supervisionada, Aprendizagem Não Supervisionada e Reinforcement Learning



Data Science Academy

# Processo de Aprendizagem

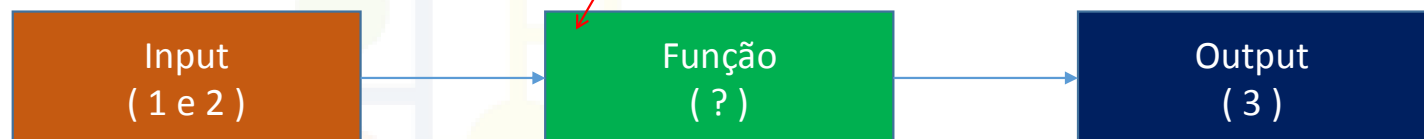
Do ponto de vista matemático, o processo de representação de aprendizagem de máquina pode ser expressado utilizando mapeamento equivalente. Mapeamento é a construção de uma função observando suas saídas.



Data Science Academy

Machine Learning


## Processo de Aprendizagem



O algoritmo aplica um score a cada função utilizada para tentar resolver o problema.



Data Science Academy



# Aprendizagem Supervisionada

O algoritmo aprende a partir de dados de exemplos de inputs e possíveis outputs, que podem ser valores quantitativos ou qualitativos



Data Science Academy



# Aprendizagem Supervisionada

Para variável qualitativa/categórica

Classificação

Atribui um rótulo

Alvo é valor numérico, segue espectro contínuo

Regressão



Data Science Academy



## Aprendizagem Supervisionada

É o termo usado sempre que o programa é “treinado” sobre um conjunto de dados pré-definido



Data Science Academy



# Aprendizagem Supervisionada

O algoritmo de aprendizagem recebe um conjunto de entradas, juntamente com as saídas corretas correspondentes e o algoritmo aprende comparando a sua saída real com as saídas corretas para então encontrar erros. Em seguida, o algoritmo ajusta o modelo de acordo com seu processo de aprendizagem



Data Science Academy



# Aprendizagem Supervisionada

A aprendizagem supervisionada é normalmente usada em aplicações onde dados históricos preveem eventos futuros



Data Science Academy





# Aprendizagem Não Supervisionada

O algoritmo aprende com exemplos simples, sem resposta associada. Os padrões de dados são determinados a cargo do algoritmo



Data Science Academy



## Aprendizagem Não-Supervisionada

Termo usado quando um programa pode automaticamente encontrar padrões e relações em um conjunto de dados



Data Science Academy



## Aprendizagem Não-Supervisionada

Os exemplos mais comuns são o K-Means, o Singular Value Decomposition (SVD) e o Principal Component Analysis (PCA)



Data Science Academy



# Reinforcement Learning

Parecido com aprendizagem não supervisionada

Similar ao que chamamos de aprender por tentativa e erro



Data Science Academy



# Reinforcement Learning

Aprendizagem por tentativa e erro

Neste caso existem e componentes envolvidos:

Tem como objetivo escolher as ações que maximizam a premiação esperada sobre um espaço de tempo

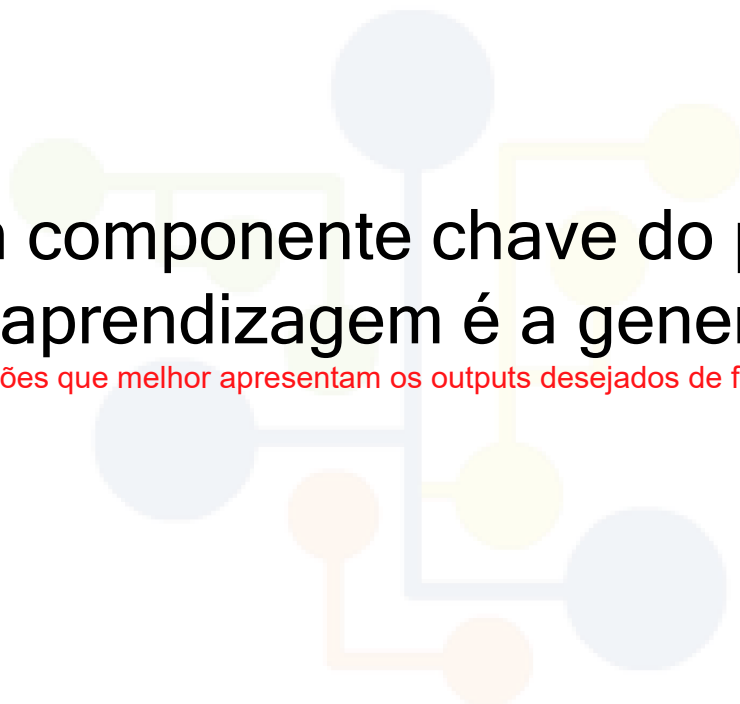

→ **Agente** tomador de decisão - próprio algoritmo

**Ambiente** onde ocorre a interação com o agente

**Ações** o que o agente pode fazer



Data Science Academy

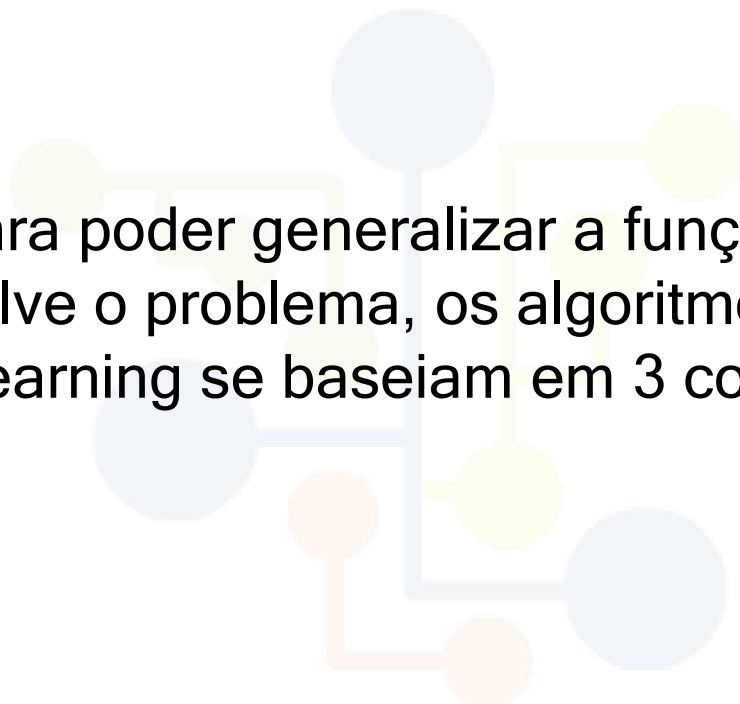



# Um componente chave do processo de aprendizagem é a generalização

O objetivo é generalizar as funções que melhor apresentam os outputs desejados de forma que a mesma solução possa ser dada a outros conjuntos de dados



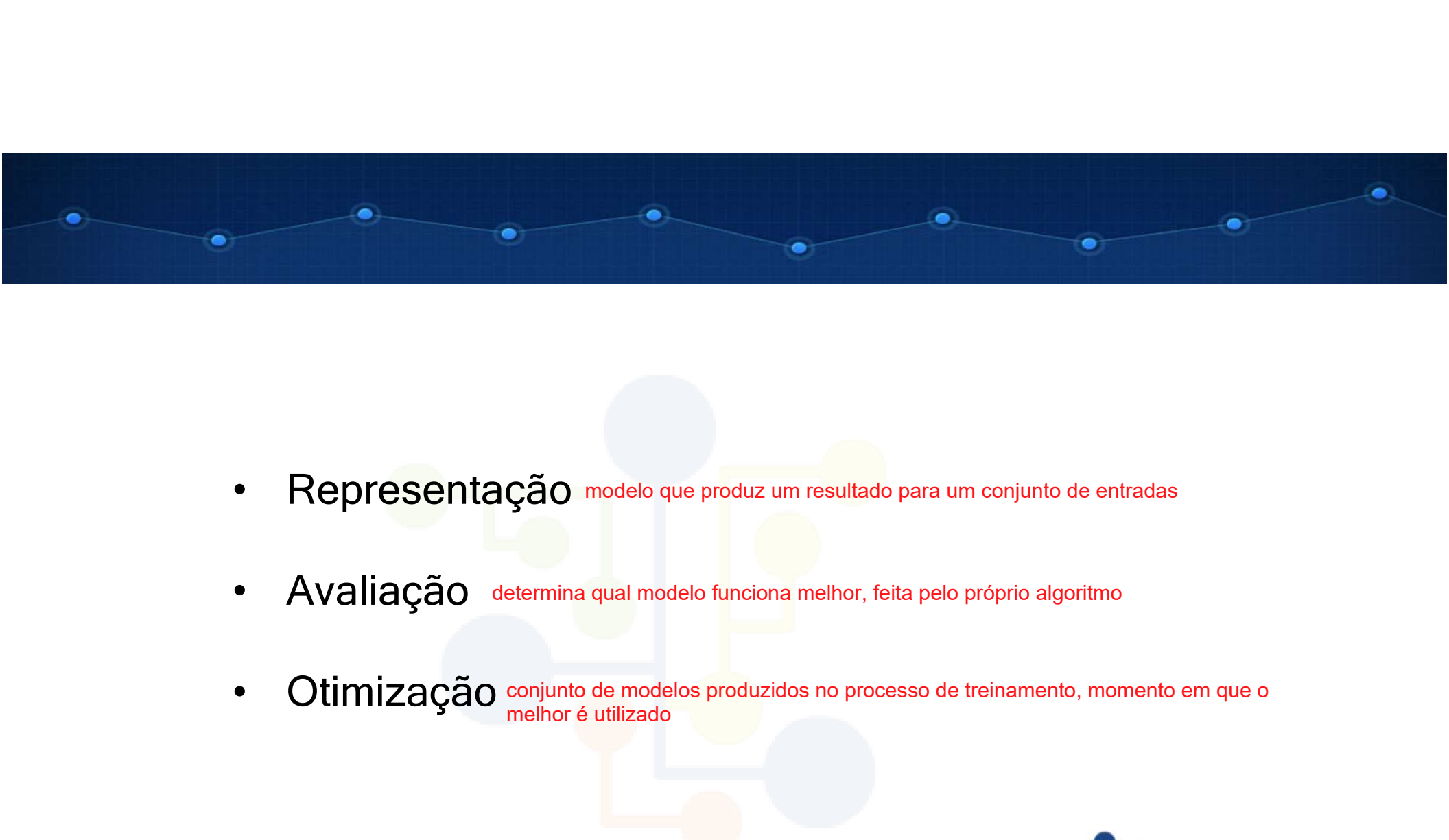
Data Science Academy



E para poder generalizar a função que melhor resolve o problema, os algoritmos de Machine Learning se baseiam em 3 componentes:




Data Science Academy

- 
- **Representação** modelo que produz um resultado para um conjunto de entradas
  - **Avaliação** determina qual modelo funciona melhor, feita pelo próprio algoritmo
  - **Otimização** conjunto de modelos produzidos no processo de treinamento, momento em que o melhor é utilizado

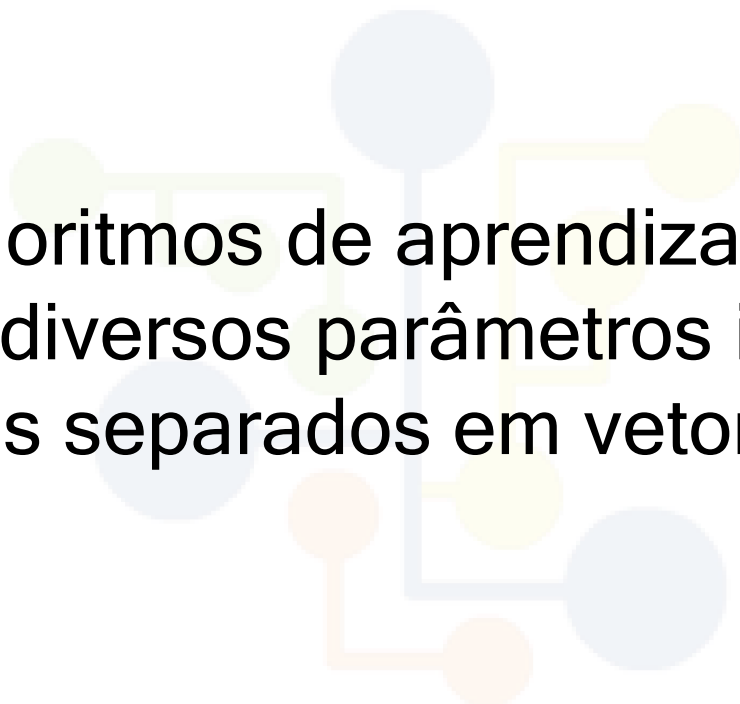


Data Science Academy





Os algoritmos de aprendizagem possuem  
diversos parâmetros internos  
(valores separados em vetores e matrizes)



Data Science Academy



# Espaço de Hipótese

Contem as variações de parâmetros de ML



Data Science Academy



- True Positive
- True Negative
- False Positive
- False Negative



Data Science Academy



False Positive → desperdício de tempo  
False Negative → oportunidade perdida



Data Science Academy

## Confusion Matrix ou matriz de erro.

|       |          | Truth                         |                |  |
|-------|----------|-------------------------------|----------------|--|
|       |          | true                          | false          |  |
| Guess | positive | true positive                 | false positive | $precision = \frac{tp}{tp + fp}$               |
|       | negative | false negative                | true negative  |  |
|       |          | $recall = \frac{tp}{tp + fn}$ |                | $accuracy = \frac{tp + tn}{tp + tn + fp + fn}$ |

Cada coluna representa as instâncias de uma classe prevista. As linhas representam as instâncias de uma classe real (valores reais).



Data Science Academy



# Cost Function

Mede quão bem o algoritmo mapeia a função alvo

Hypothesis:  $h_{\theta}(x) = \theta_0 + \theta_1 x$

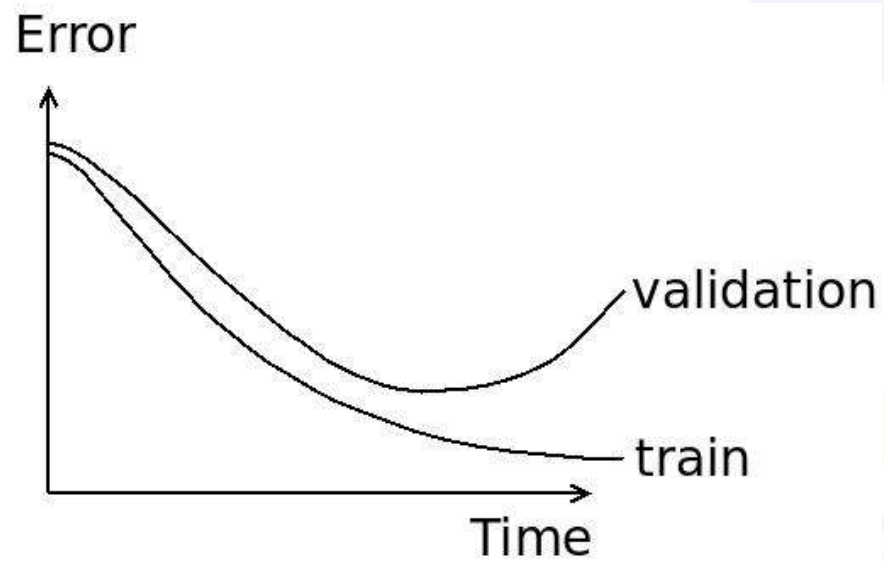
Parameters:  $\theta_0, \theta_1$

Cost Function:  $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

Goal:  $\underset{\theta_0, \theta_1}{\text{minimize}} J(\theta_0, \theta_1)$



Data Science Academy



Definindo o Erro



Data Science Academy



## Cost Function → Nível de erro

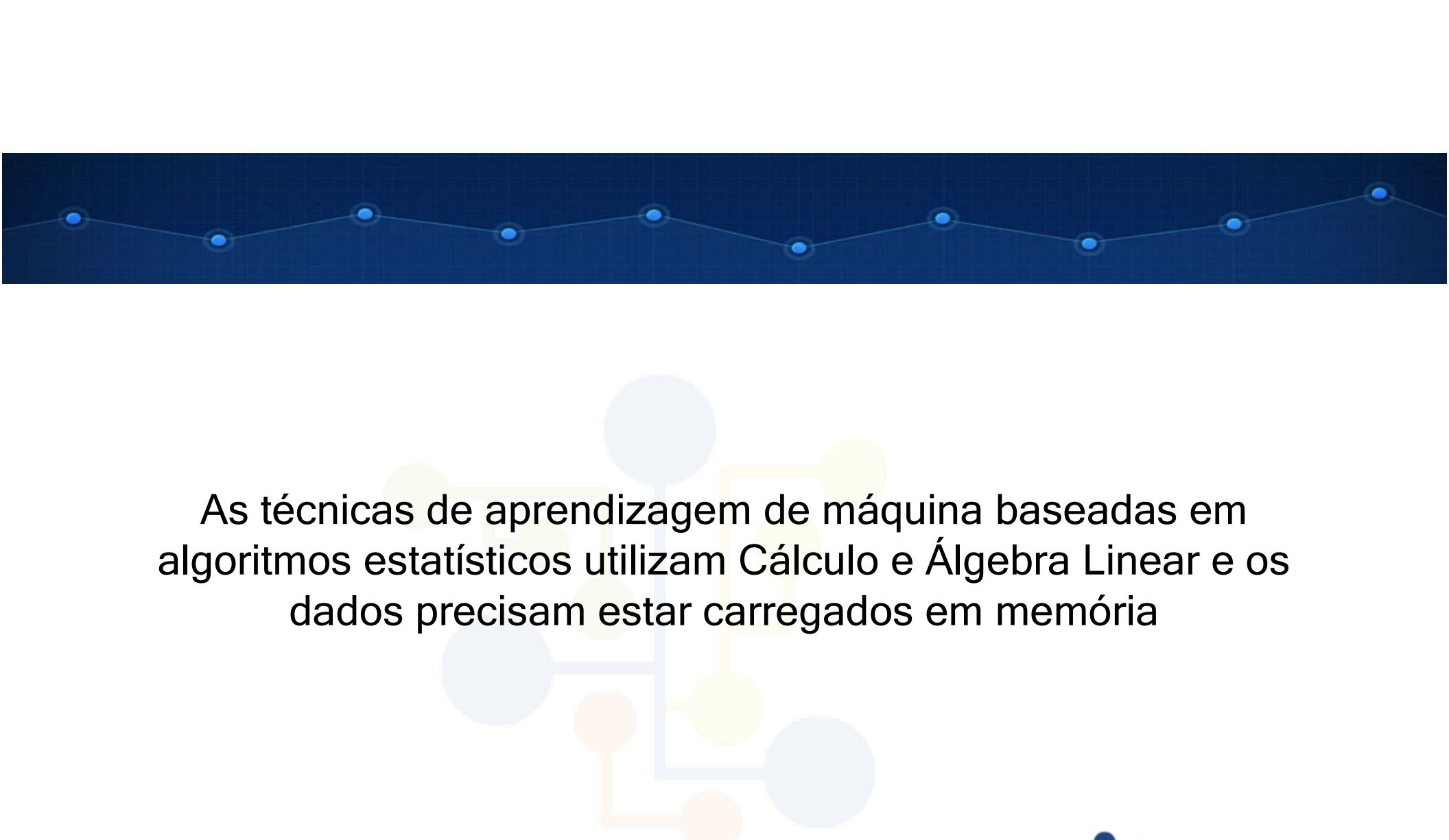
Ajuda a compreender o processo de aprendizagem como:

- representação: capacidade de aproximar-se de certas funções matemáticas
- otimização: como os algoritmos de aprendizagem de máquina definem seus parâmetros internos



Data Science Academy

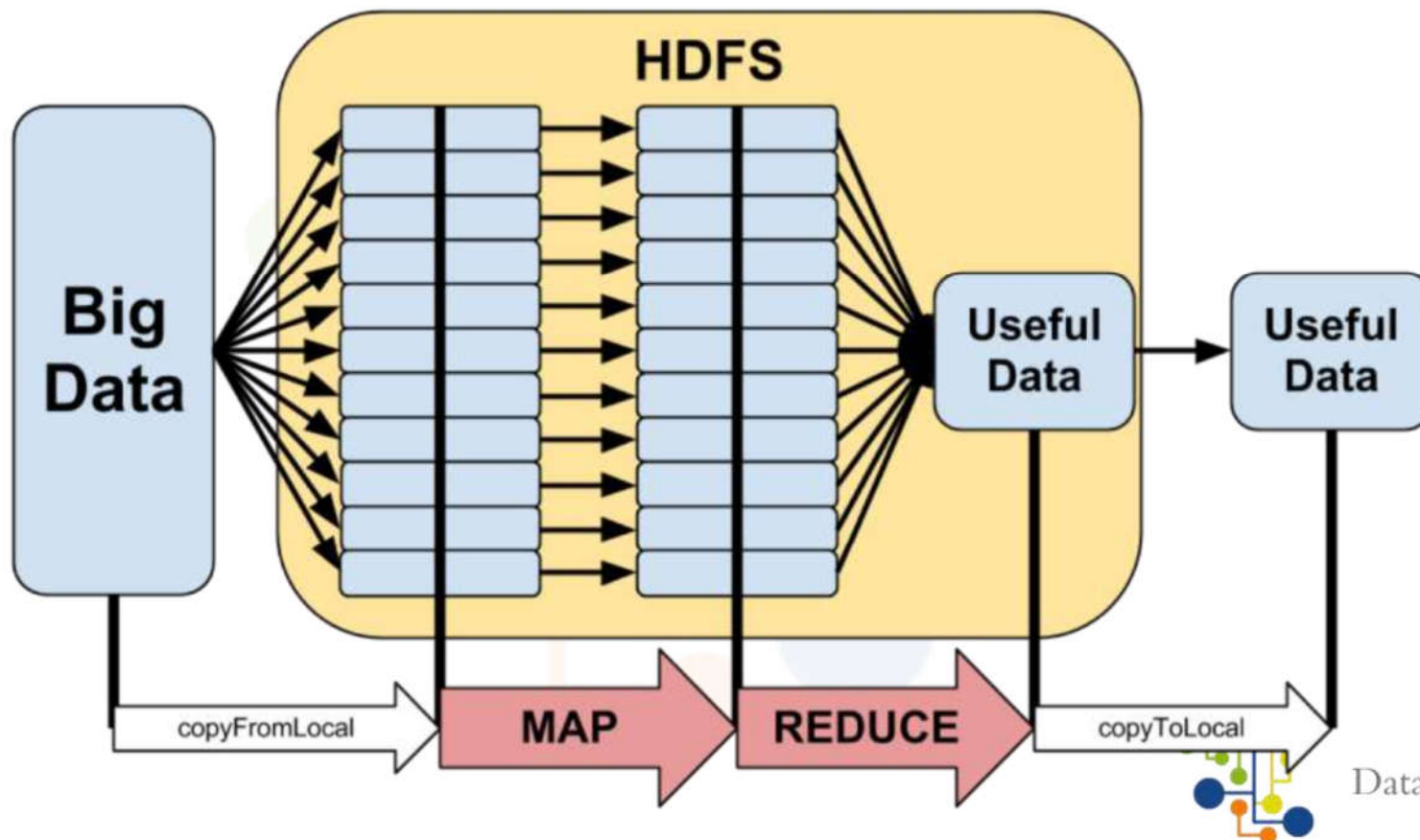




As técnicas de aprendizagem de máquina baseadas em algoritmos estatísticos utilizam Cálculo e Álgebra Linear e os dados precisam estar carregados em memória

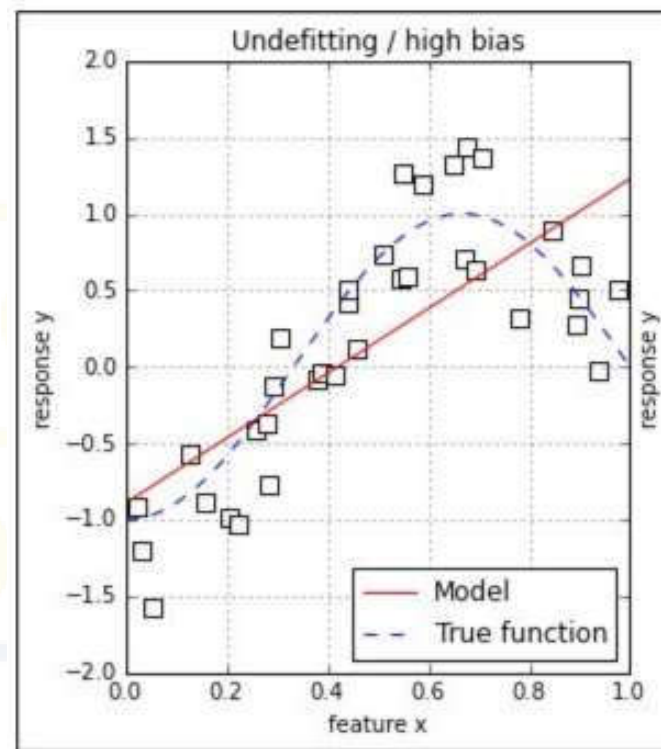


Data Science Academy



Data Science Academy

Perceba no gráfico que vai existir uma diferença entre o modelo preditivo (linha vermelha) e a função que resolve o problema (linha tracejada azul). Isso ocorre, por que o algoritmo tende a sistematicamente subestimar ou sobreestimar as regras do mundo real, que representam partes tendenciosas. Normalmente isso ocorre com algoritmos que não são capazes de expressar problemas matemáticos complexos.

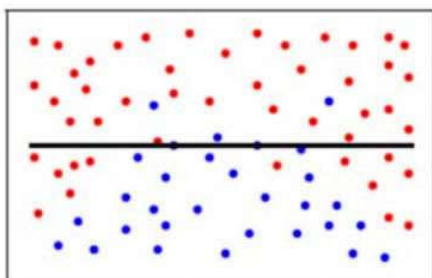


Data Science Academy

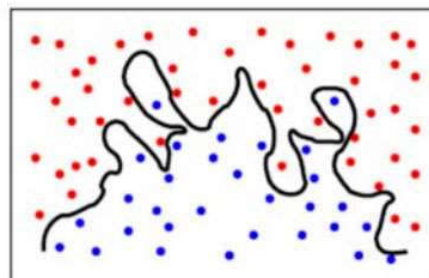
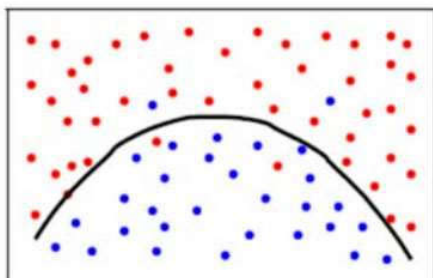


# Overfitting


Underfitting




Overfitting



Data Science Academy



Para visualizar se os seus algoritmos de Machine Learning estão sofrendo algum tipo de força tendenciosa, você pode usar um gráfico chamado *Curva de Aprendizagem*



mostra a performance do algoritmo



Data Science Academy

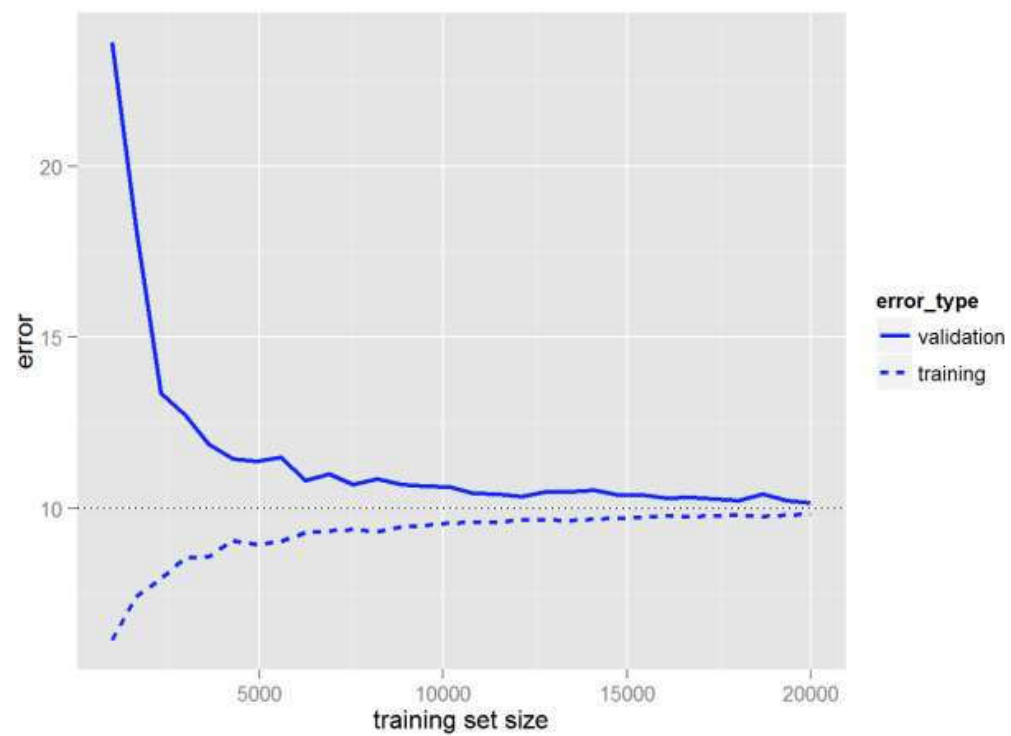


## Para usar uma curva de aprendizagem, você precisa:


- 1- Dividir seus dados em amostras, chamadas dados de treino e dados de teste (uma divisão 70/30 funciona bem e permite cross-validation).
- 2- Criar porções dos seus dados de treino, com tamanhos diferentes a cada passagem de treino. Conceito de amostragem
- 3- Treinar seus modelos com os diferentes subsets. Registrar a performance.
- 4- Gerar um gráfico com os resultados. Atenção aos intervalos de confiança e ao desvio padrão.



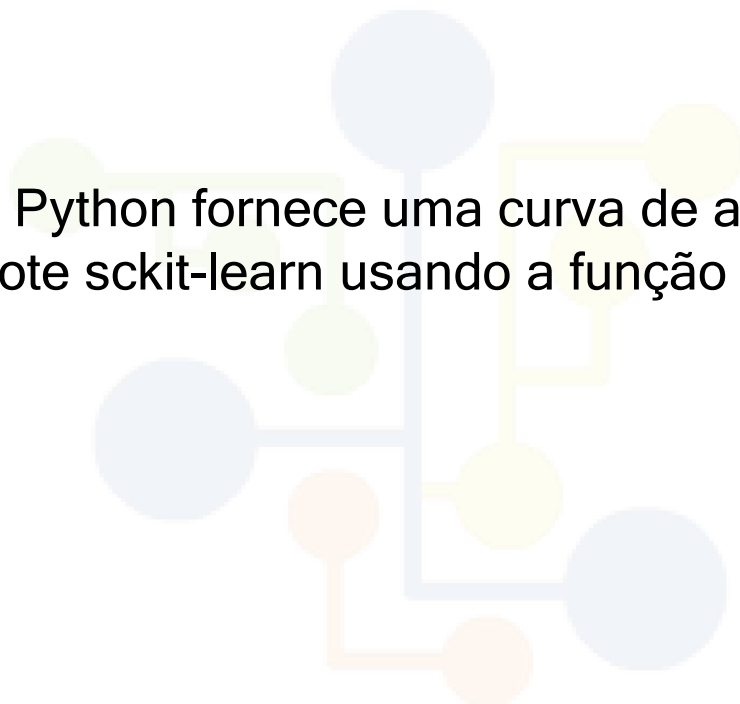
Data Science Academy



Data Science Academy



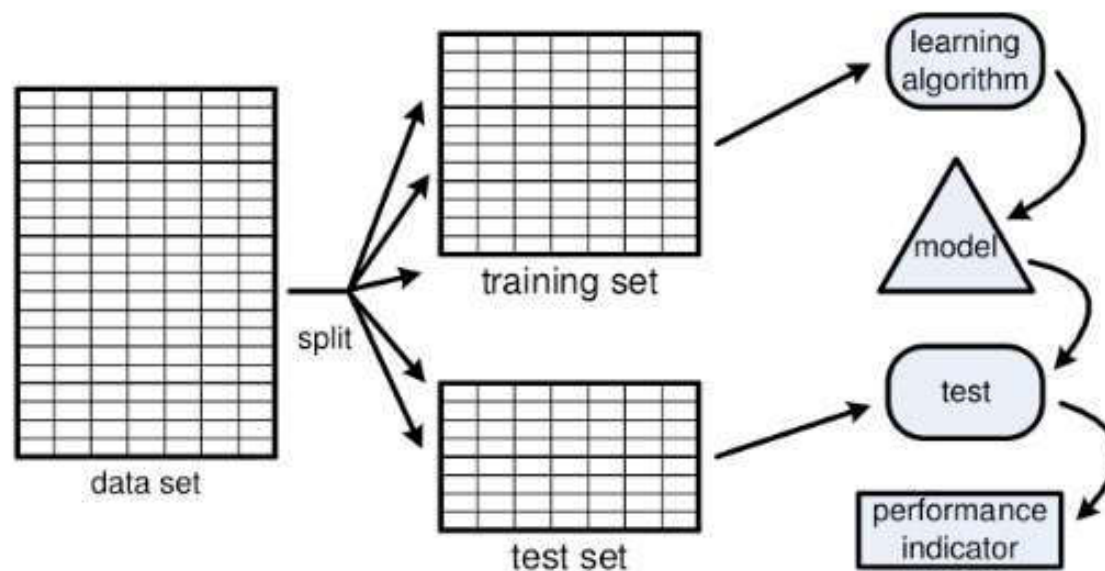
A linguagem Python fornece uma curva de aprendizagem através do pacote scikit-learn usando a função `learning_curve`.



Data Science Academy



# Treinamento, Validação e Teste



Data Science Academy



## Treinamento, Validação e Teste

75 a 70% - dados de treino

25 a 30% - dados de teste



Data Science Academy



## Treinamento, Validação e Teste

70% - dados de treino

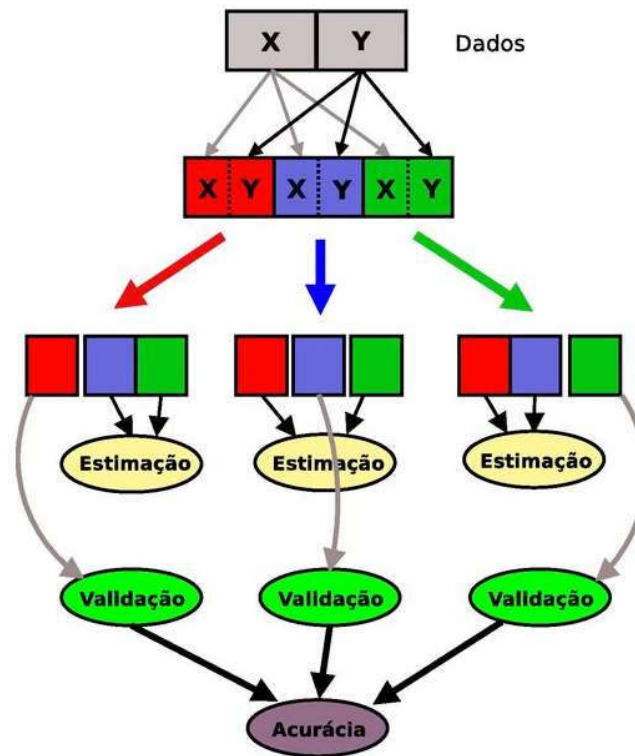
20% - dados de validação

10% - dados teste



Data Science Academy

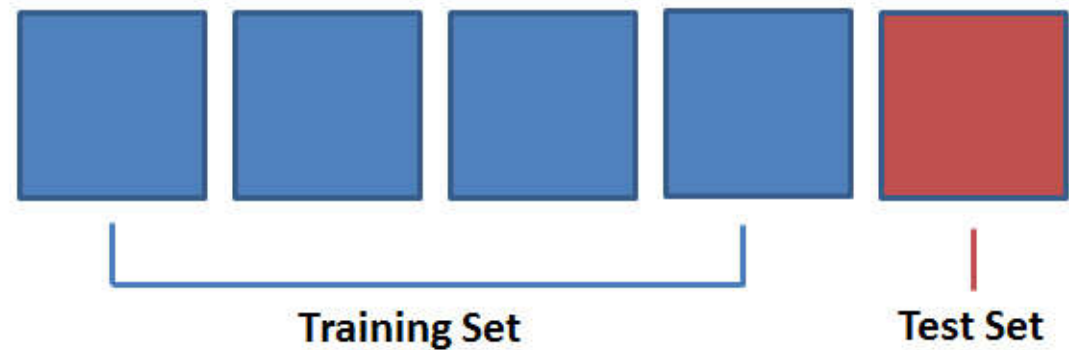
# Treinamento, Validação e Teste



Data Science Academy

# Treinamento, Validação e Teste

$n > 10.000$



Data Science Academy



# Cross-Validation

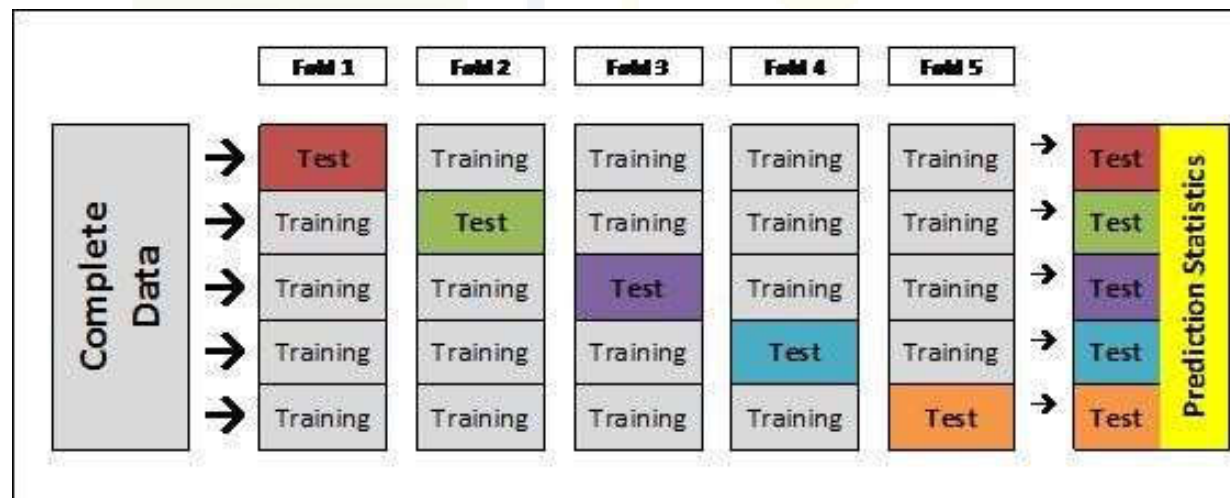


O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutualmente exclusivos, e posteriormente, utiliza-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo



Data Science Academy

# Cross-Validation



Data Science Academy

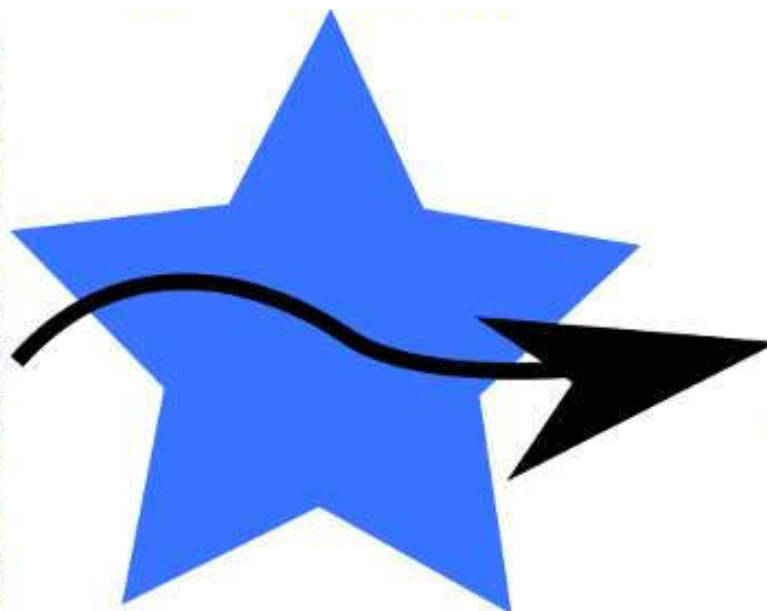




Dados

```
100100011101000000101000110111010110
100100111101110000001111100110100100
100001101101111101010011100001101001
111111010000110111001010111100001011
11001111110111111100100001110110110
010000110100110110000110000100010000
010101110011001111011001110100010111
001000010101100101000001000010011110
011101001111110010111010101010111100
100010000101100010101101010111000101
010010000100101011110011100001010000
010110000010011101010010101110110001
011011111010111100010100010100010000
011010011011011010001000101111001101
000101000001100110001100100010010110
100101010100010011100101010101111101
```

Algoritmo



Modelo

$$f(\mathbf{x})$$



Data Science Academy





# Modelo

Existem muitos tipos diferentes de modelos. Você pode já estar familiarizado com alguns. Os exemplos incluem:

- Equações matemáticas
- Diagramas relacionais
- Agrupamentos de dados, conhecidos como clusters



Data Science Academy

# Modelo

Observações



Dados

| Distance | Time |
|----------|------|
| 4.9m     | 1s   |
| 19.6m    | 2s   |
| 44.1m    | 3s   |
| 78.5m    | 4s   |

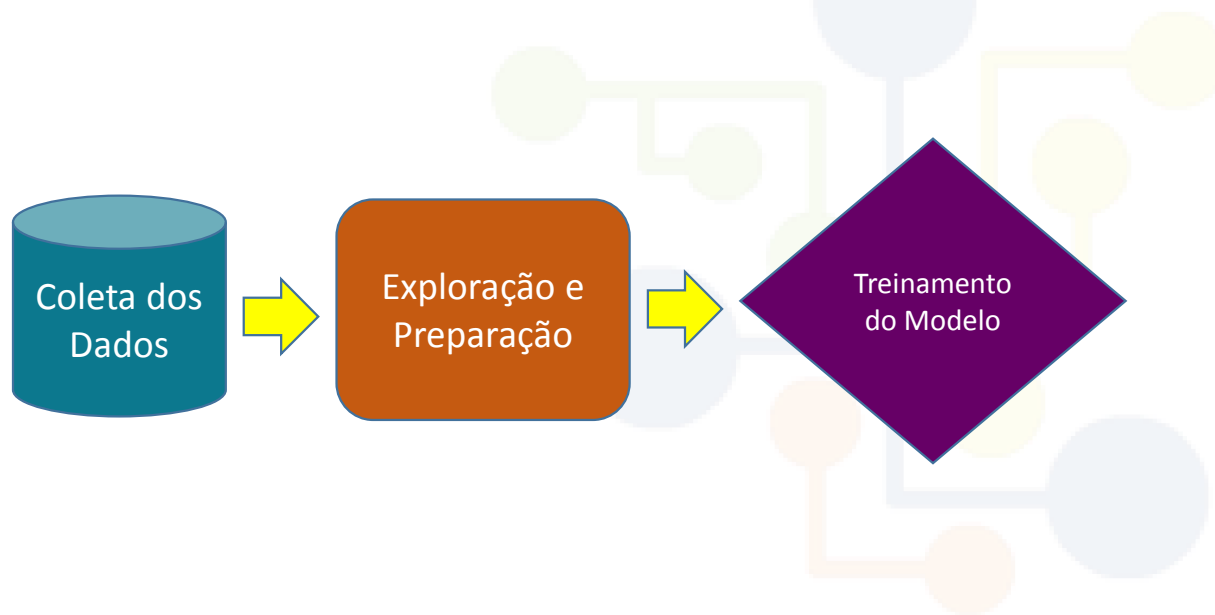
Modelo

$$g = 9.8m/s^2$$



Data Science Academy

# Criação do Modelo



Data Science Academy



# Modelo

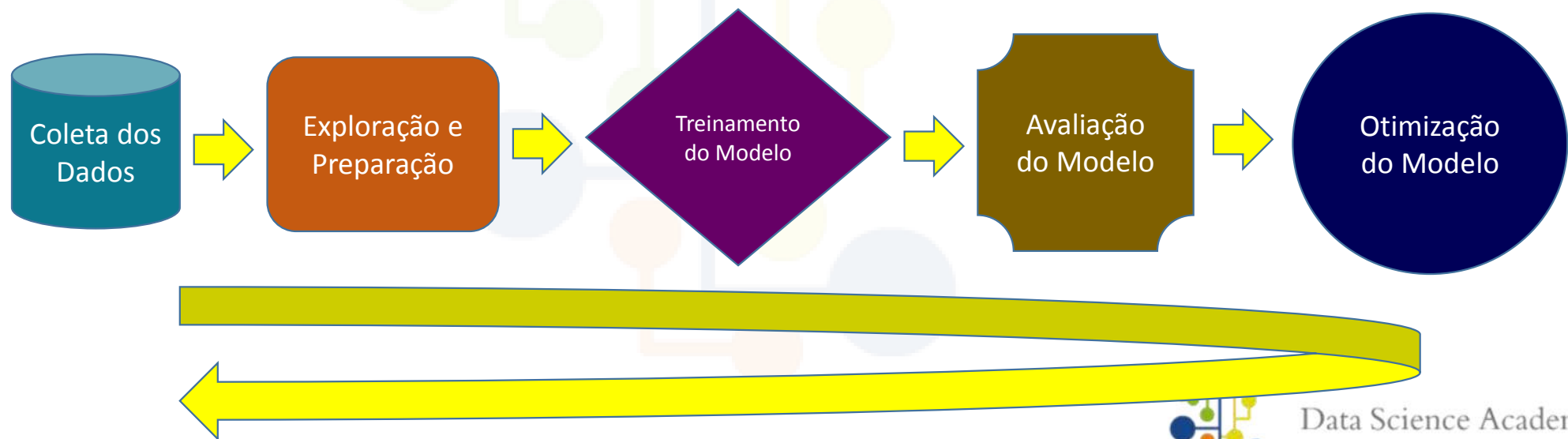


O processo de "fitting" um modelo a um dataset é chamado de treinamento do modelo



Data Science Academy

# Criação do Modelo



Data Science Academy



## Aprendizagem Supervisionada

- Classificação
- Regressão

## Aprendizagem Não Supervisionada

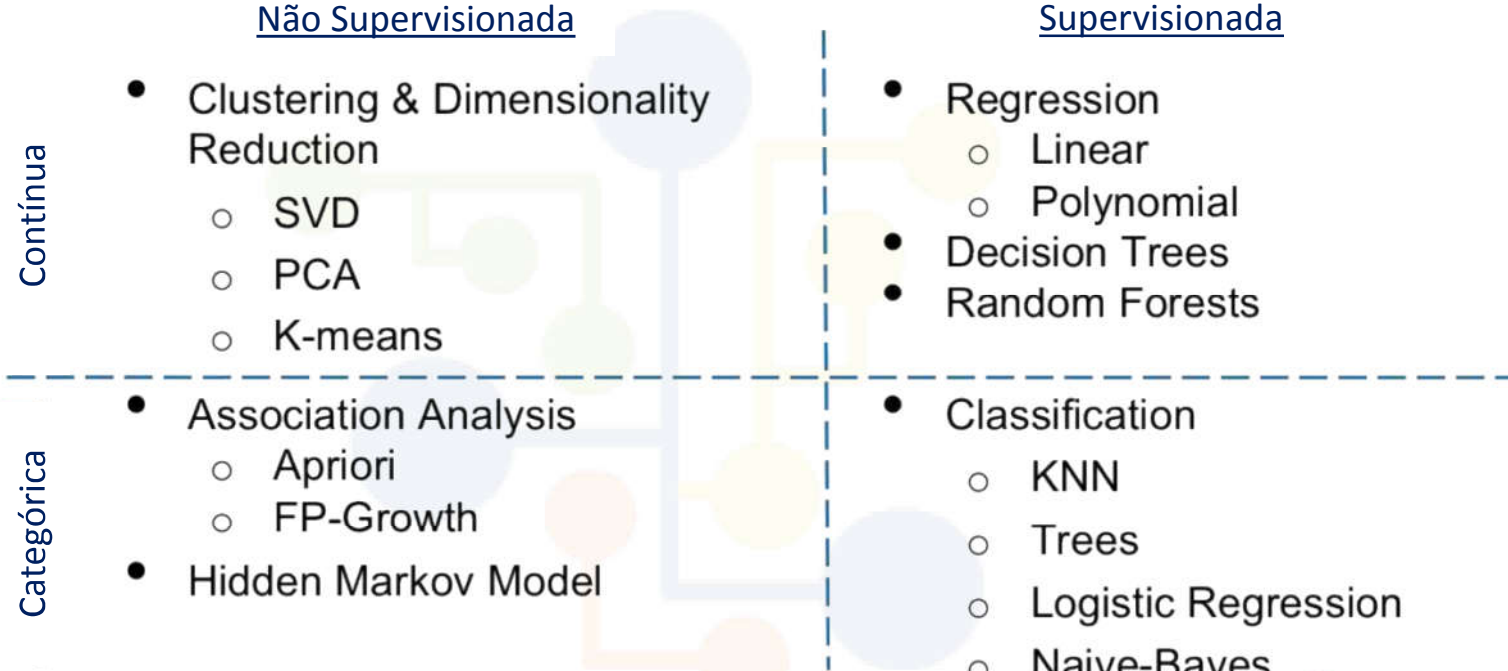

- Clustering
- Segmentação
- Redução de Dimensionalidade

## Aprendizagem por Reforço

- Sistemas de Recomendação
- Sistemas de Recompensa
- Processo de Decisão



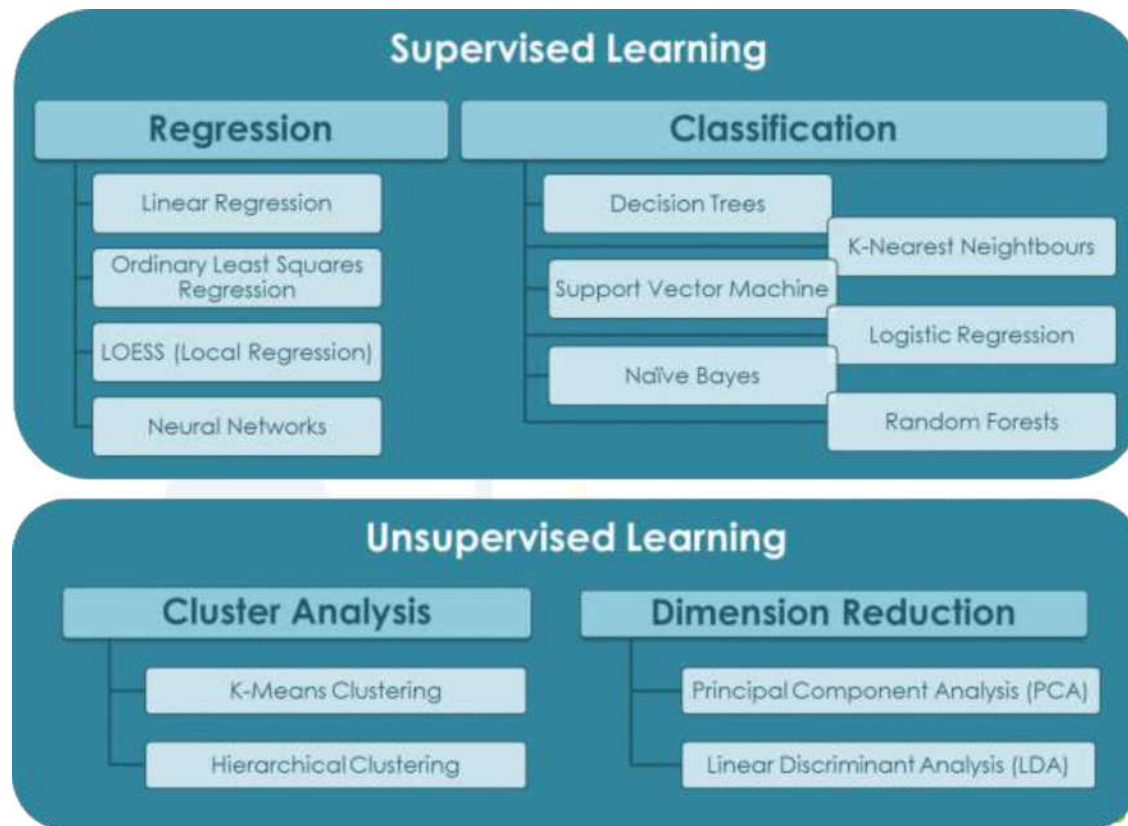
Data Science Academy



|            | <u>Não Supervisionada</u>   | <u>Supervisionada</u>   |
|------------|---|---|
| Contínua   | <ul style="list-style-type: none"><li>• Clustering &amp; Dimensionality Reduction<ul style="list-style-type: none"><li>○ SVD</li><li>○ PCA</li><li>○ K-means</li></ul></li></ul>  | <ul style="list-style-type: none"><li>• Regression<ul style="list-style-type: none"><li>○ Linear</li><li>○ Polynomial</li></ul></li><li>• Decision Trees</li><li>• Random Forests</li></ul>             |
| Categórica | <ul style="list-style-type: none"><li>• Association Analysis<ul style="list-style-type: none"><li>○ Apriori</li><li>○ FP-Growth</li></ul></li><li>• Hidden Markov Model</li></ul> | <ul style="list-style-type: none"><li>• Classification<ul style="list-style-type: none"><li>○ KNN</li><li>○ Trees</li><li>○ Logistic Regression</li><li>○ Naive-Bayes</li><li>○ SVM</li></ul></li></ul> |



Data Science Academy

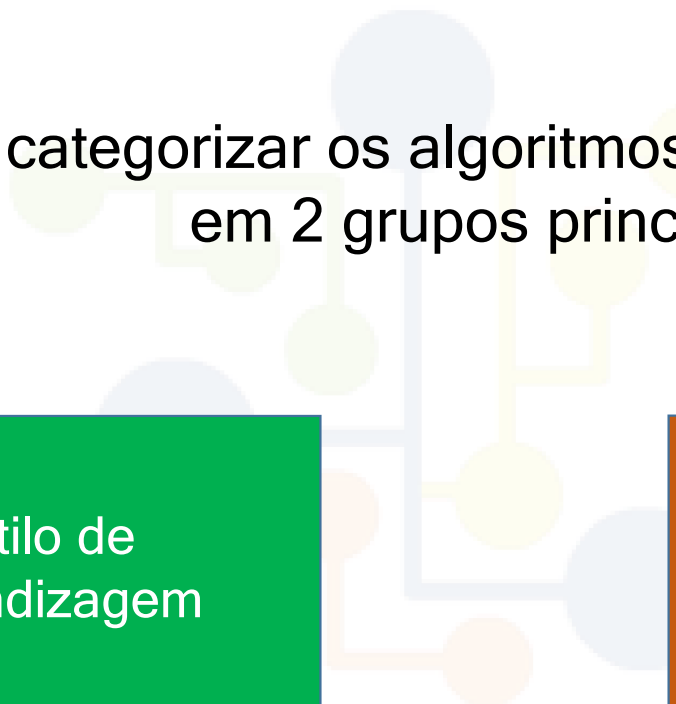


Data Science Academy





Podemos categorizar os algoritmos de Machine Learning  
em 2 grupos principais:




Estilo de  
Aprendizagem

Similaridade  
(Funcionamento)



Data Science Academy



Podemos categorizar os algoritmos de Machine Learning em 2 grupos principais:

Estilo de  
Aprendizagem

- Aprendizagem Supervisionada
- Aprendizagem Não Supervisionada
- Reinforcement Learning



Data Science Academy



# Algoritmos de Regressão

Regressão refere-se a modelar a relação entre variáveis, ajustando as medidas de erro nas previsões feitas pelo modelo.

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

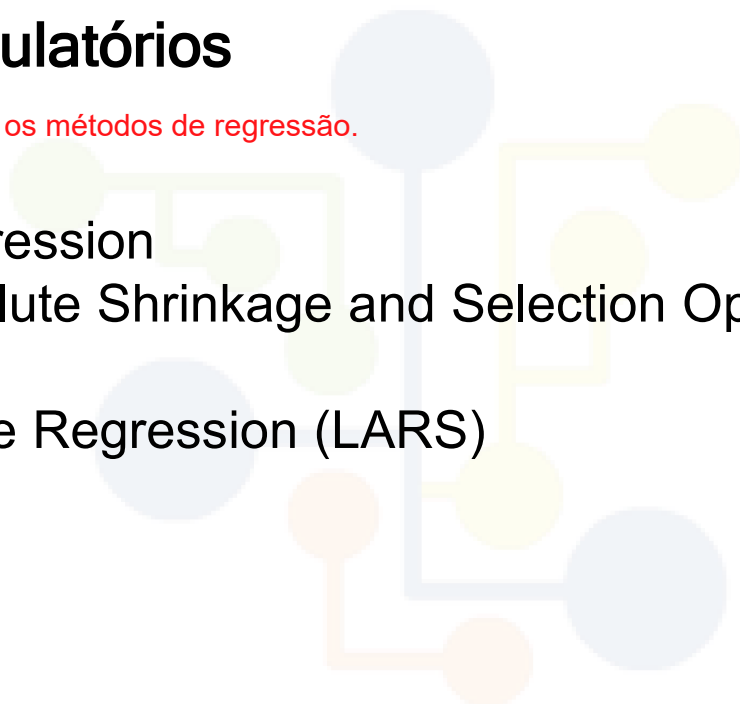


Data Science Academy



# Algoritmos Regulatórios

Geralmente são extensão para os métodos de regressão.

- Ridge Regression
  - Least Absolute Shrinkage and Selection Operator (LASSO)
  - Elastic Net
  - Least-Angle Regression (LARS)
- 



Data Science Academy



## Algoritmos Baseados em Instância (Instance-based)

Constroem banco de dados de exemplo e comparam novos dados com esse banco por similaridade.

- k-Nearest Neighbour (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)



Data Science Academy



## Algoritmos de Árvore de Decisão

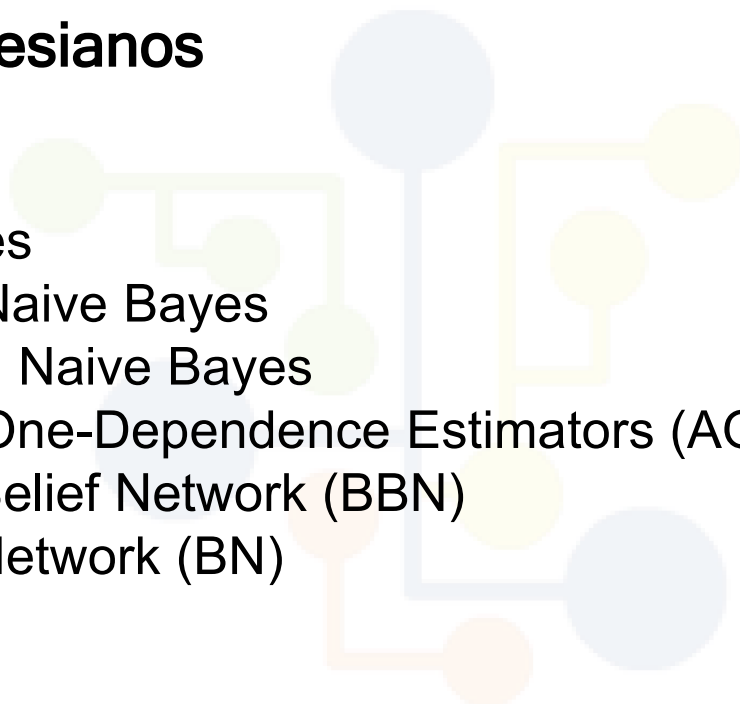
- Classification and Regression Tree (CART)
- Conditional Decision Trees
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5



Data Science Academy



## Algoritmos Bayesianos

- Naive Bayes
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Averaged One-Dependence Estimators (AODE)
  - Bayesian Belief Network (BBN)
  - Bayesian Network (BN)
- 




Data Science Academy



# Algoritmos de Clustering

Dados organizados em clusters

- k-Means
  - k-Medians
  - Expectation Maximisation (EM)
  - Hierarchical Clustering
- 

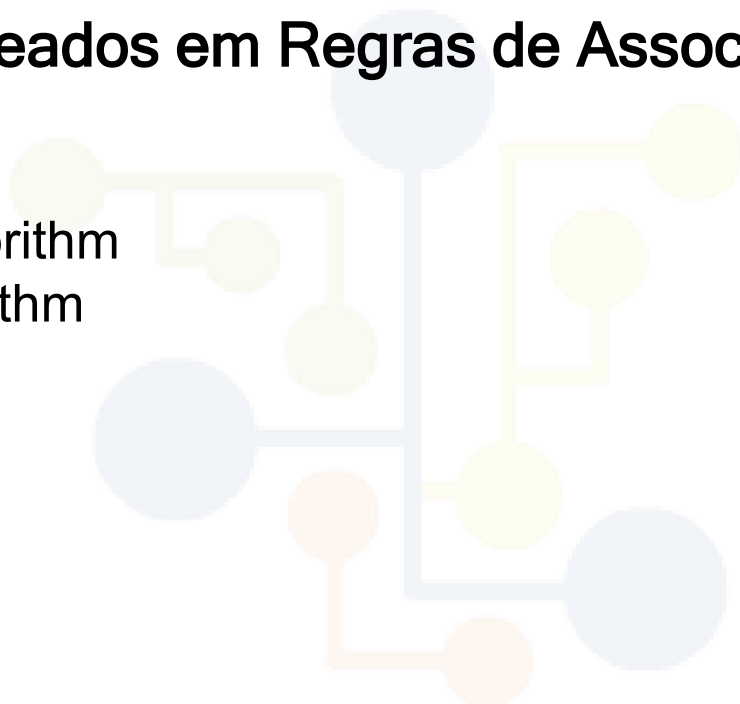


Data Science Academy





## Algoritmos Baseados em Regras de Associação

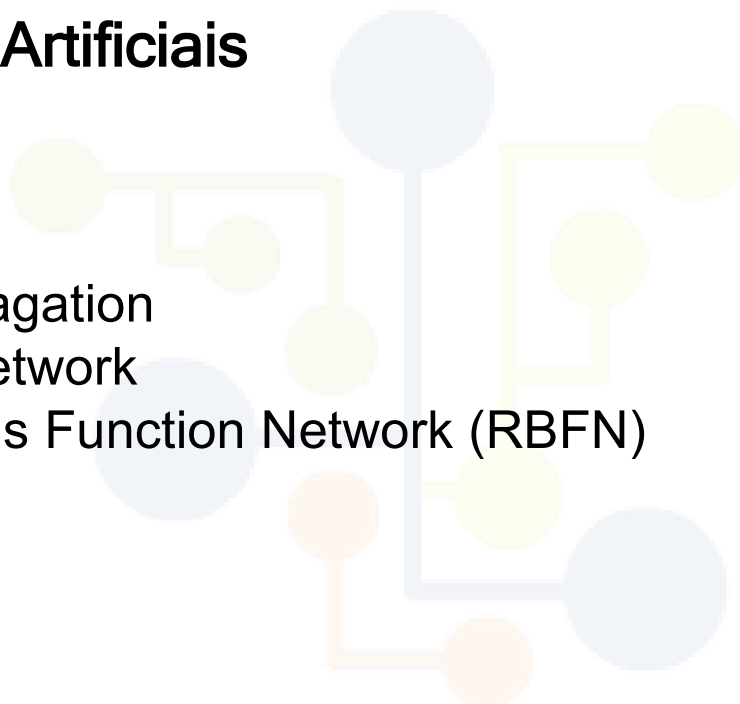
- Apriori algorithm
  - Eclat algorithm
- 



Data Science Academy



## Redes Neurais Artificiais

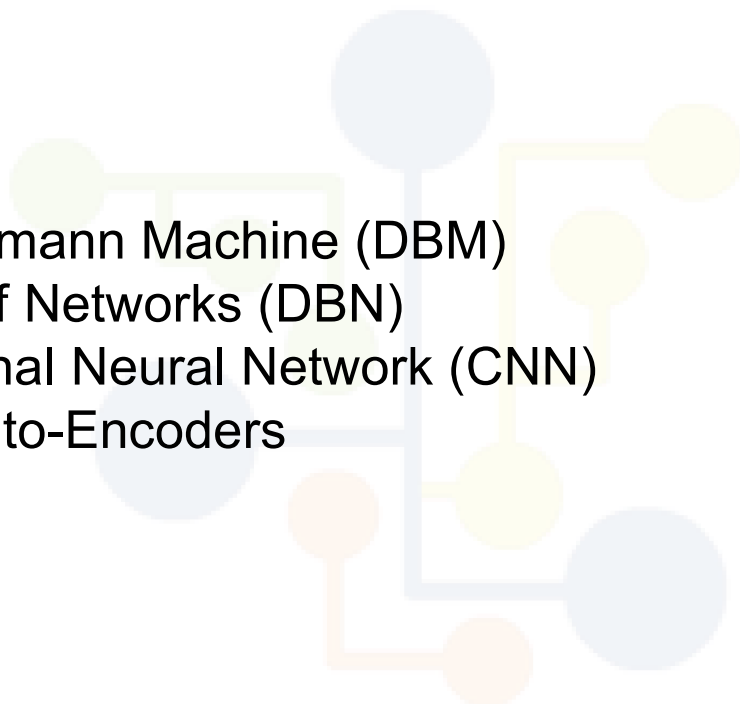
- Perceptron
  - Back-Propagation
  - Hopfield Network
  - Radial Basis Function Network (RBFN)
- 



Data Science Academy



## Deep Learning

- Deep Boltzmann Machine (DBM)
  - Deep Belief Networks (DBN)
  - Convolutional Neural Network (CNN)
  - Stacked Auto-Encoders
- 



Data Science Academy



## Algoritmos de Redução de Dimensionalidade

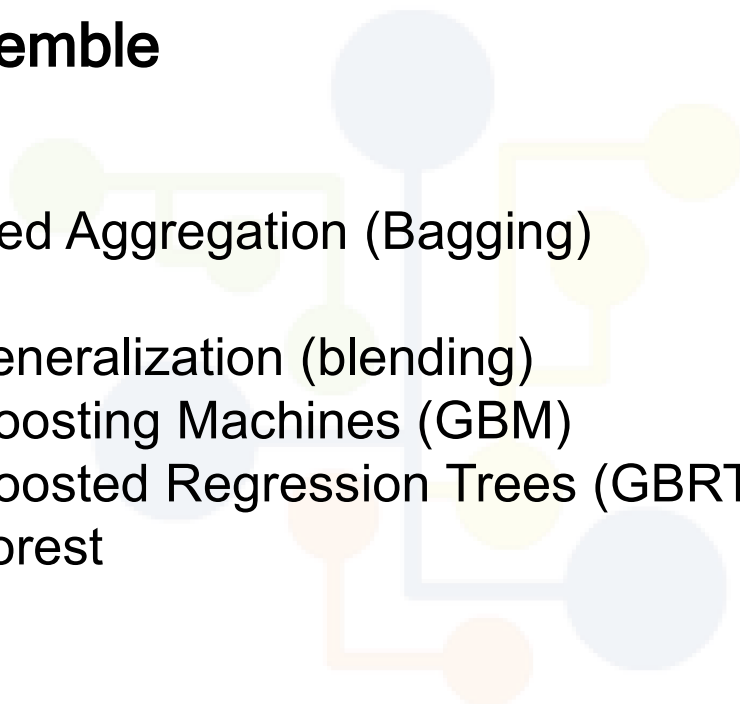
- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Flexible Discriminant Analysis (FDA)



Data Science Academy



## Algoritmos Ensemble

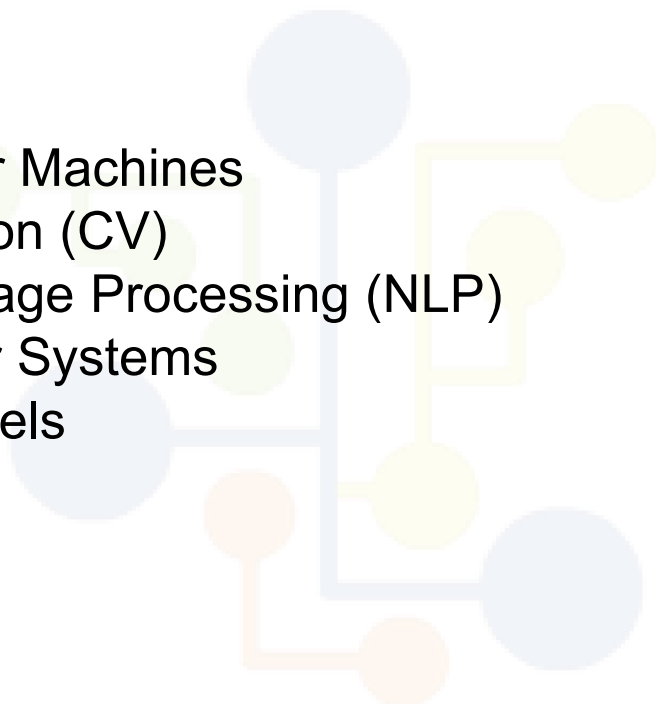
- Boosting
  - Bootstrapped Aggregation (Bagging)
  - AdaBoost
  - Stacked Generalization (blending)
  - Gradient Boosting Machines (GBM)
  - Gradient Boosted Regression Trees (GBRT)
  - Random Forest
- 



Data Science Academy



## Outros Algoritmos

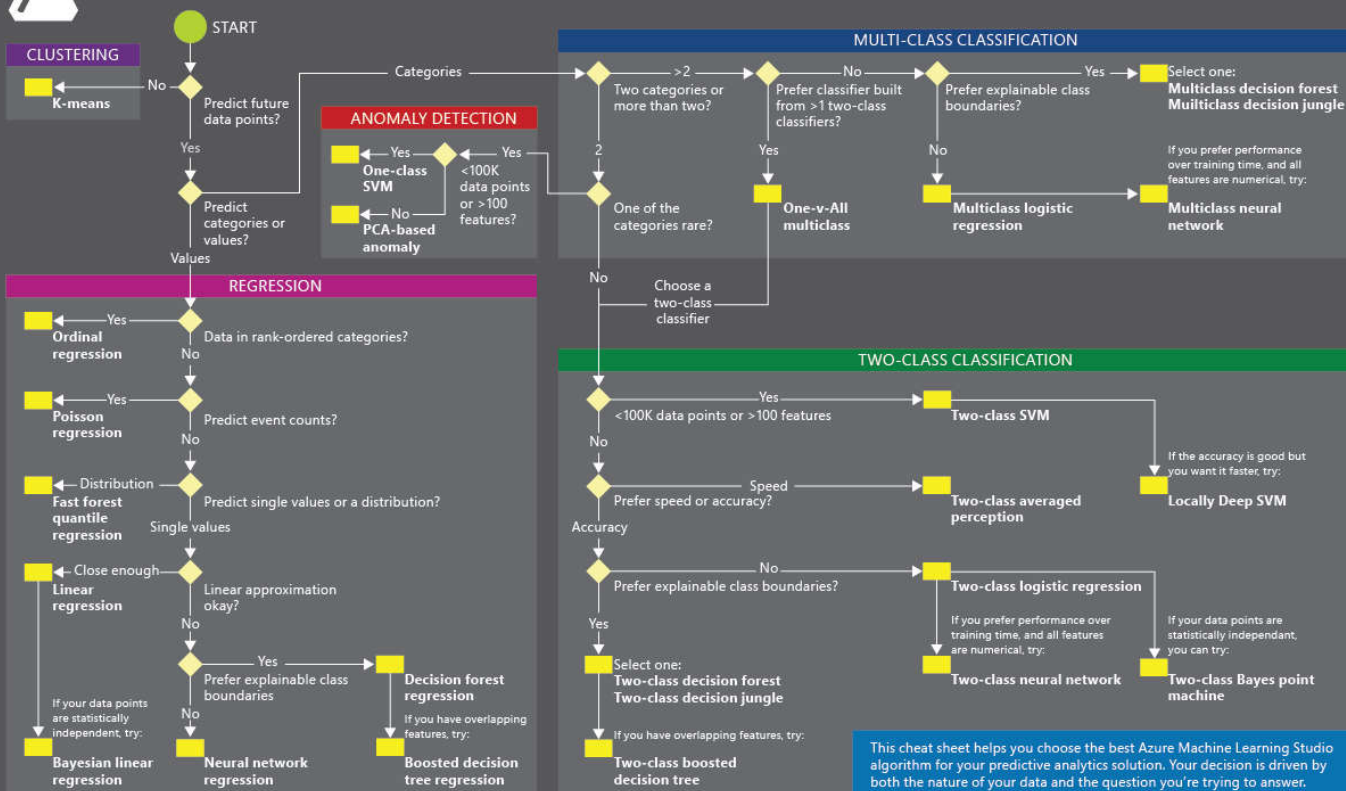
- Support Vector Machines
  - Computer Vision (CV)
  - Natural Language Processing (NLP)
  - Recommender Systems
  - Graphical Models
- 



Data Science Academy



## Microsoft Azure Machine Learning: Algorithm Cheat Sheet



© 2015 Microsoft Corporation. All rights reserved.

Created by the Azure Machine Learning Team

Email: AzurePoster@microsoft.com

Download this poster: <http://aka.ms/MLCheatSheet>



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



# Regressão Linear



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



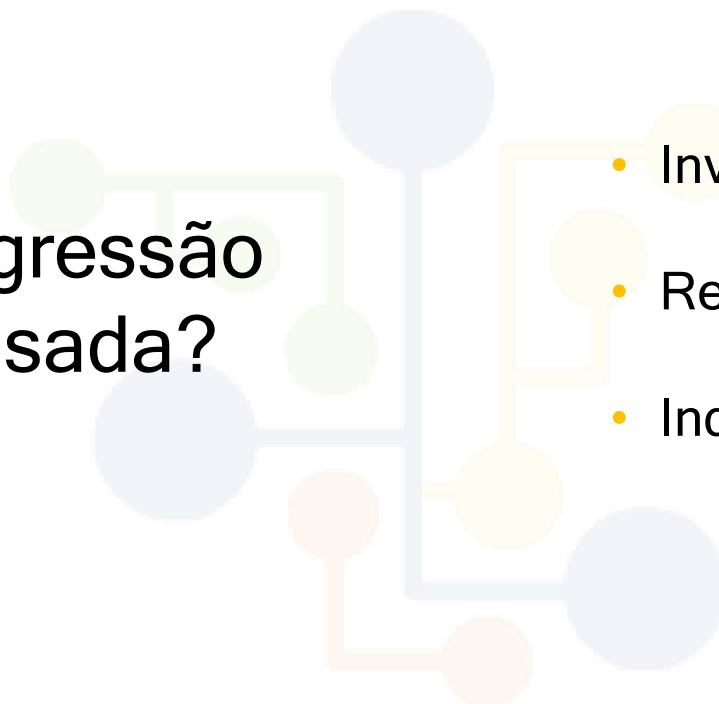



# Regressão Linear Simples

Um estudo de regressão linear simples busca, essencialmente, associar uma variável  $Y$  (denominada variável resposta ou variável dependente) a uma outra variável  $X$  (denominada variável explanatória ou variável independente)



Data Science Academy



## Como a Regressão pode ser usada?

- Investigação Científica
- Relações Causais
- Identificação de Padrões



Data Science Academy

# Compreendendo a Regressão

$$\hat{y} = a + bx$$

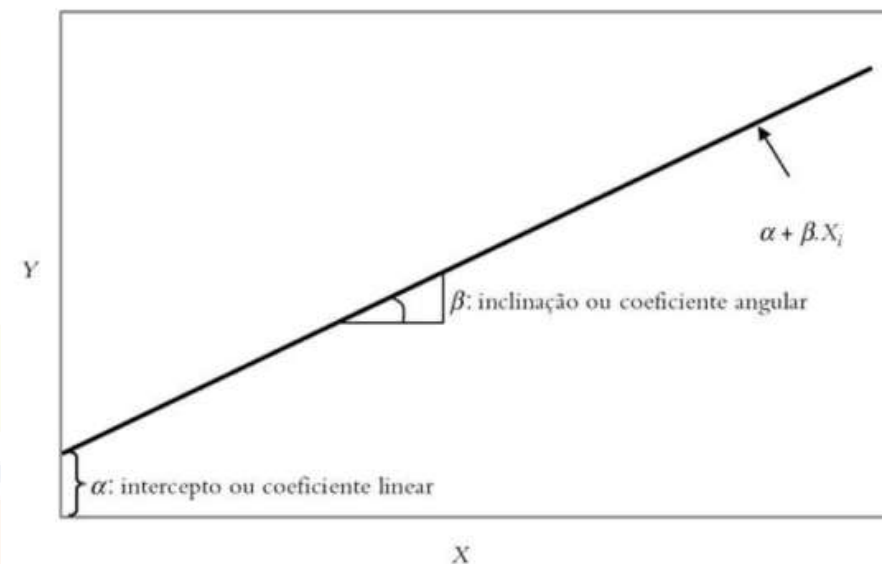
Onde:

$\hat{y}$  = valor previsto de  $y$  dado um valor para  $x$

$x$  = variável independente

$a$  = ponto onde a linha intercepta o eixo  $y$

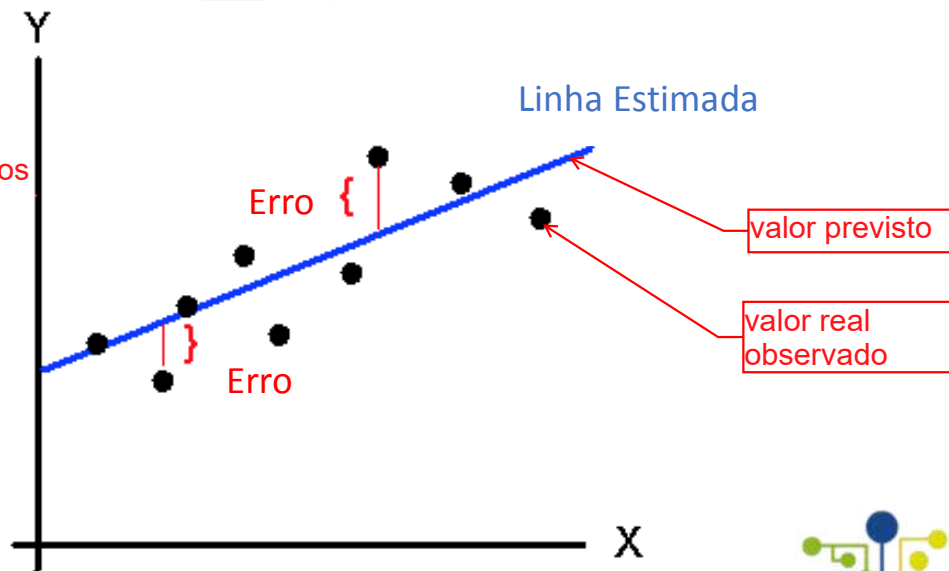
$b$  = inclinação da linha reta




Data Science Academy

# Estimativa dos Mínimos Quadrados

Fornece os valores de  $a$  e  $B$  da fórmula anterior, que minimizam a soma dos quadrados dos resíduos, ou seja, que minimizam a distância entre os valores observados e os valores estimados pelo modelo, indicados pela reta.



Data Science Academy



Deve-se determinar  $\alpha$  e  $\beta$  de modo que a somatória dos quadrados dos resíduos seja a menor possível (método de Mínimos Quadrados Ordinários - MQO, ou, em inglês, Ordinary Least Squares - OLS)



Data Science Academy



Função para regressão  
linear

Formula

```
model <- lm(log(PINCP, base=10) ~ AGEP + SEX + COW + SCHL, data=dtrain)
```

Objeto R para  
guardar o resultado  
da regressão

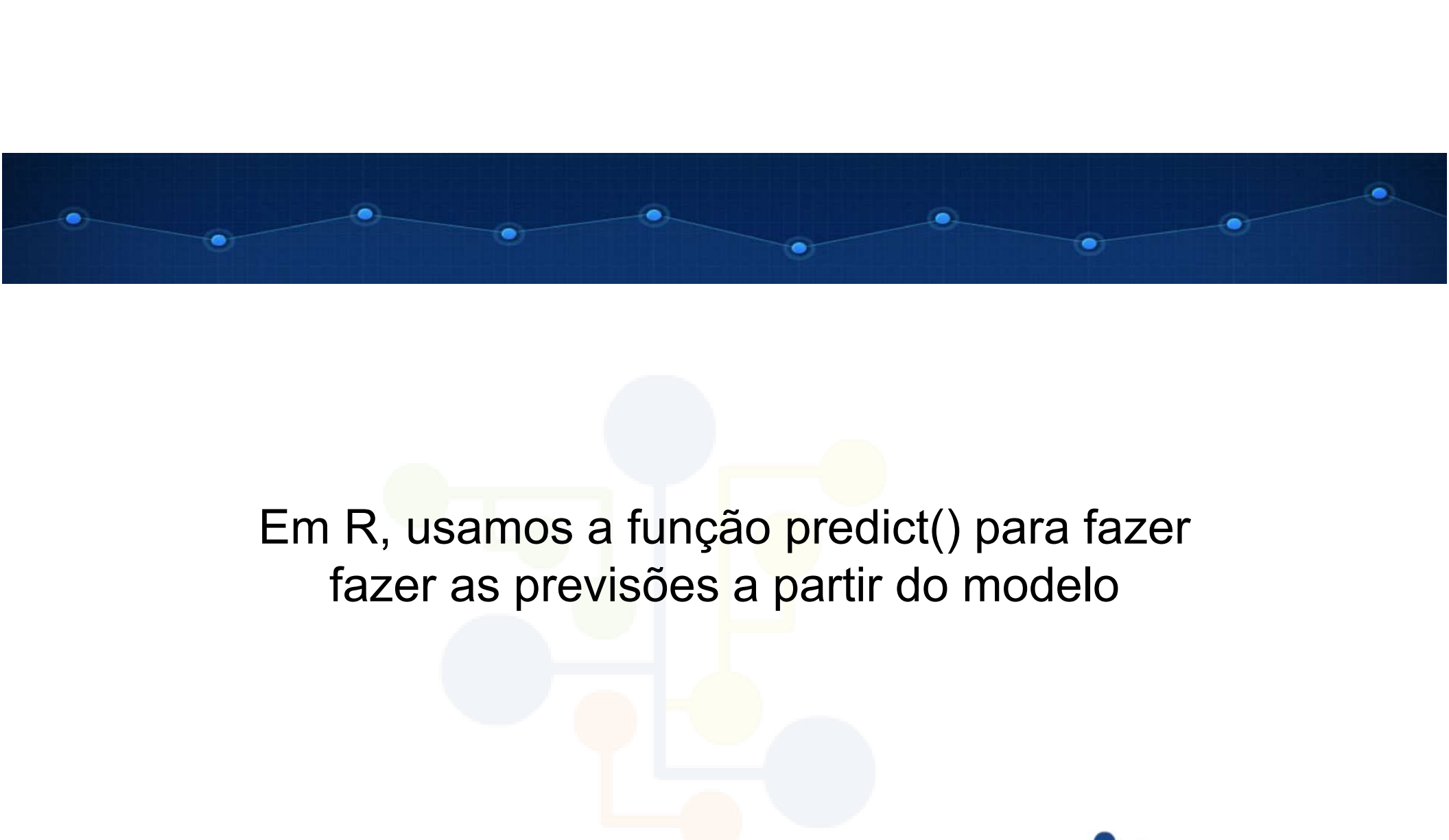
Variável  
quantitativa que  
queremos prever

Variáveis preditoras

Conjunto de dados de  
treino



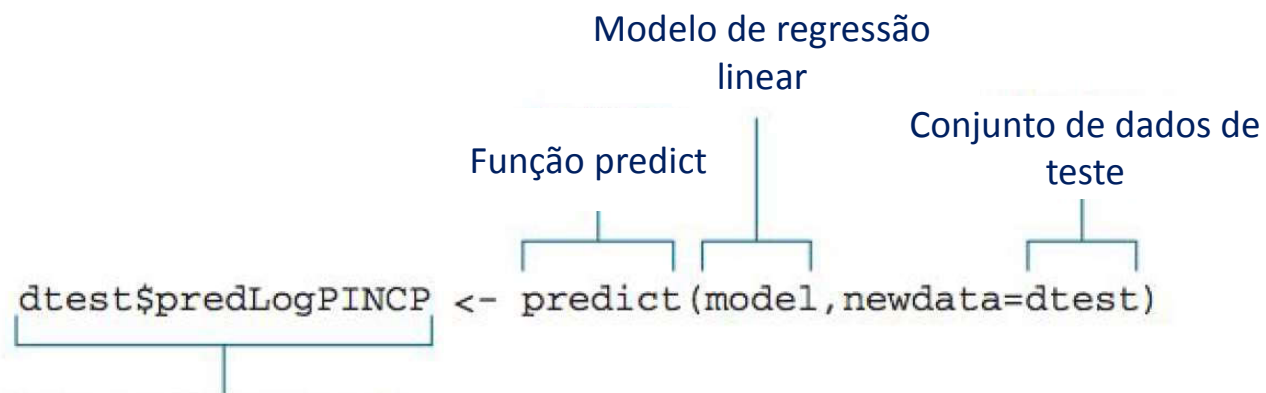
Data Science Academy



Em R, usamos a função `predict()` para fazer  
fazer as previsões a partir do modelo



Data Science Academy



Podemos armazenar a previsão em outra coluna no dataset de teste

```
dtrain$predLogPINCP <- predict(model, newdata=dtrain)
```

E podemos fazer a mesma operação no dataset de treino



Data Science Academy





# Classificação

É o processo de identificar a qual conjunto de categorias uma nova observação pertence, com base em um conjunto de dados de treino contendo observações (ou instâncias) cuja associação é conhecida



Data Science Academy



## Classificação

K Nearest Neighbors (kNN) é um algoritmo que armazena e então classifica os dados de acordo com os dados mais próximos de suas características



Data Science Academy



## Classificação

O kNN é um algoritmo não paramétrico, que pode ser usado para classificação ou para regressão



Data Science Academy



## Classificação

Não paramétrico significa que o algoritmo não conhece previamente os dados e suas distribuições



Data Science Academy



## Classificação

O kNN é um dos algoritmos mais simples de Machine Learning, mas que tem sido muito utilizado em diversos segmentos



Data Science Academy



# Classificação

- Aplicações de reconhecimento de imagens e reconhecimento facial, tanto em imagens quanto em vídeos.
- Previsão se uma pessoa irá gostar da recomendação de filmes ou músicas.
- Identificação de padrões em dados genéticos, detectando doenças específicas.



Data Science Academy



O classificador KNN é indicado para tarefas de classificação onde o relacionamento entre as variáveis e as classes, ou grupos de variáveis, são numerosas, complexas e difíceis de compreender, embora os itens dessas classes sejam homogêneos. Ou seja, usamos classificação quando o conceito é difícil de explicar, mas fácil de definir depois de encontradas algumas características.



Data Science Academy



# Classificador kNN

| Vantagens                                     | Desvantagens   |
|---|--|
| Simple e efetivo                              | Não produz um modelo, limitando a compreensão como as características das classes de dados se relacionam |
| Cria suposições sobre a distribuição de dados | Requer a apropriada seleção do valor de k  |
| Fase de treinamento bastante veloz            | Fase de classificação é lenta  |

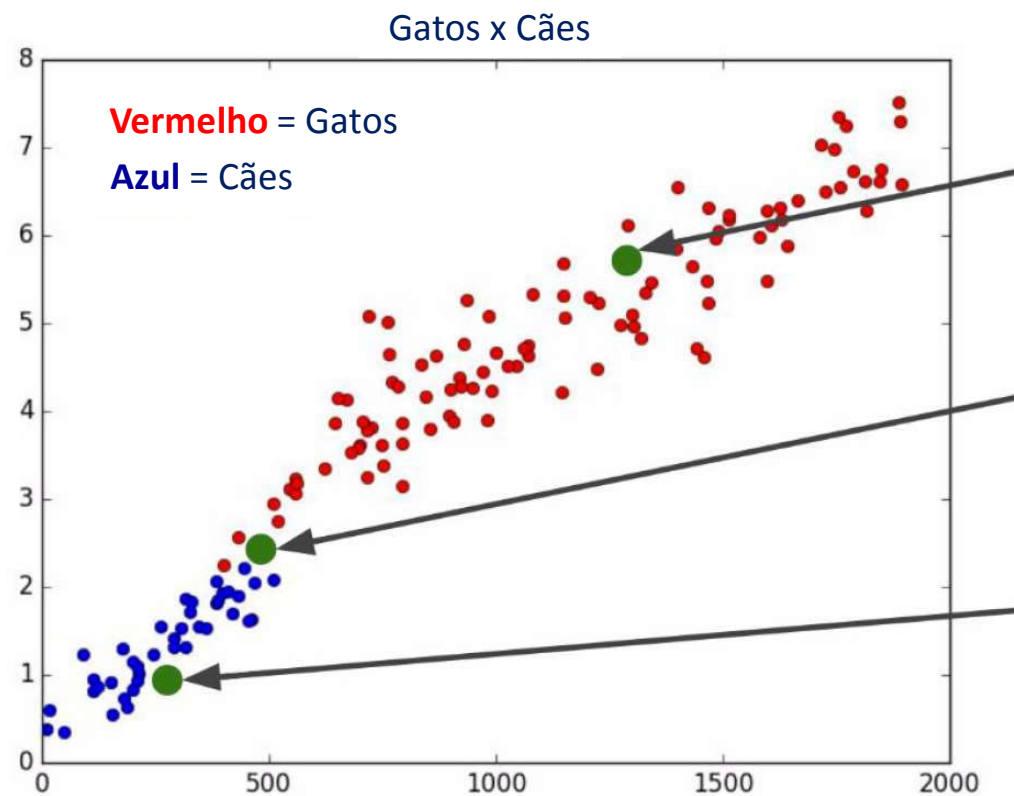


Data Science Academy





Normalmente as observações mais próximas são definidas como aquelas com a menor distância euclidiana ao ponto de dados em consideração.



Este novo ponto de dado representa um gato ou um cachorro?

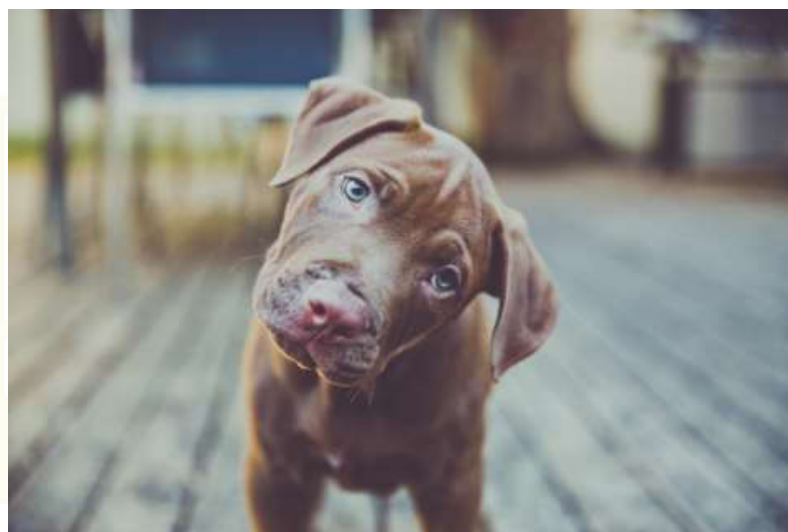
Este novo ponto de dado representa um gato ou um cachorro?

Este novo ponto de dado representa um gato ou um cachorro?



Distância euclidiana, ou distância métrica, é a distância entre dois pontos que pode ser provada pela aplicação repetida do teorema de Pitágoras. Aplicando esta fórmula como distância o espaço euclidiano torna-se o espaço métrico.

Eucli o que?



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



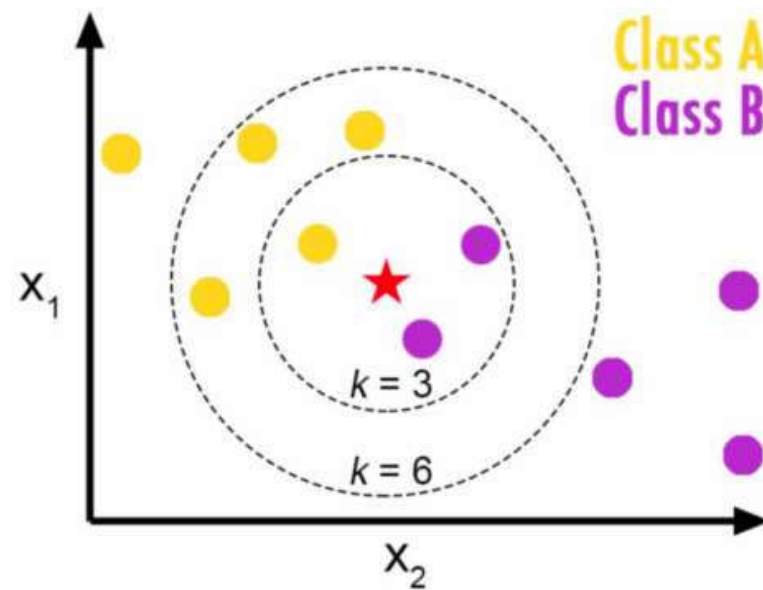
## Classificação kNN

- Armazena todos os dados
- Calcula a distância de  $x$  para todos os pontos de dados
- Ordena os pontos dentro dos seus dados aumentando a distância para  $x$
- Prevê a maioria de valores de "k" próxima aos pontos

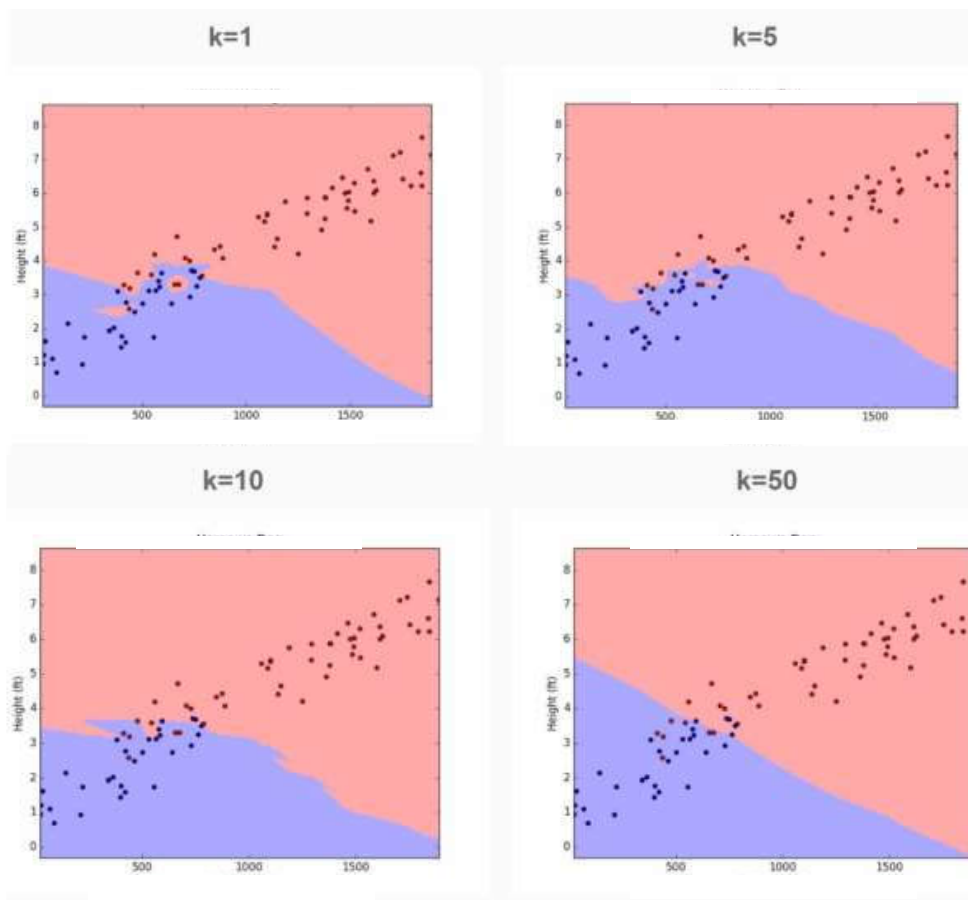


Data Science Academy

O valor de k faz a diferença



Data Science Academy



Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)



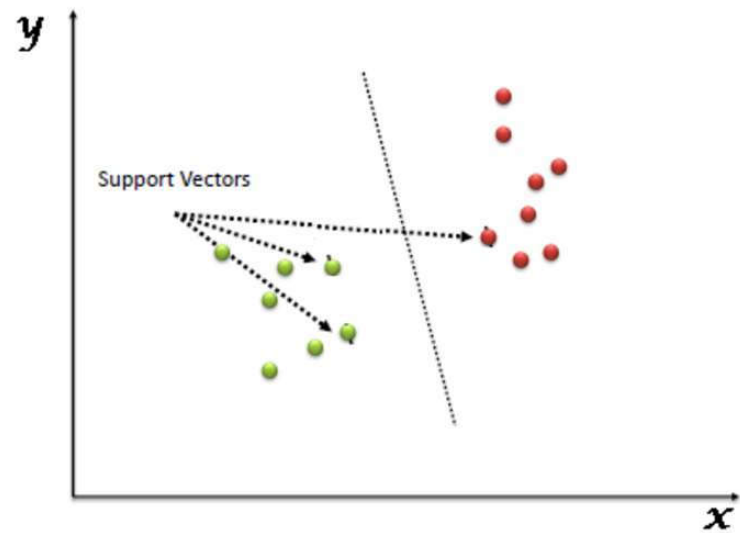
# Support Vector Machine




Data Science Academy

[www.datascienceacademy.com.br](http://www.datascienceacademy.com.br)

# Support Vector Machine



Data Science Academy

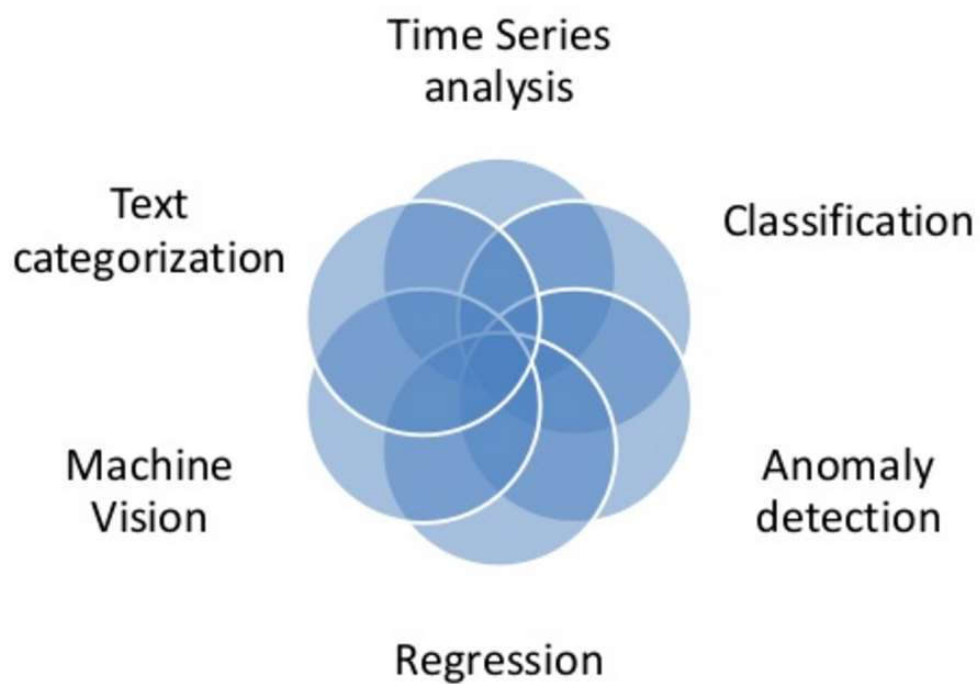


Vetores de suporte são simplesmente as coordenadas de observação individual. Support Vector Machine é uma fronteira que melhor se segrega as duas classes (hiper-plano / linha).



Data Science Academy





Data Science Academy