

# Big Data Analytics com R e Microsoft Azure Machine Learning



Data Science Academy



# Machine Learning



Data Science Academy



Seja Bem-Vindo



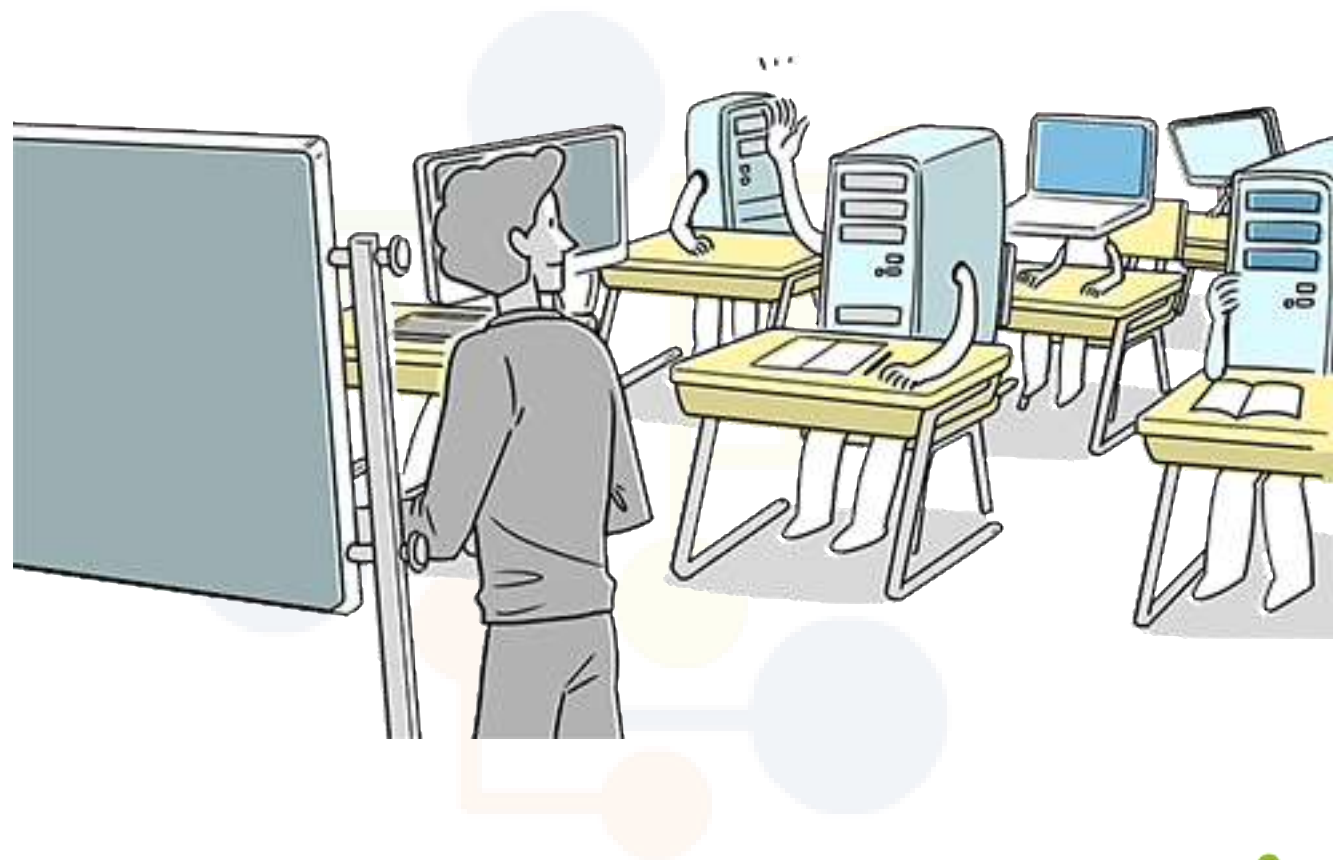
Data Science Academy



# Introdução



Data Science Academy



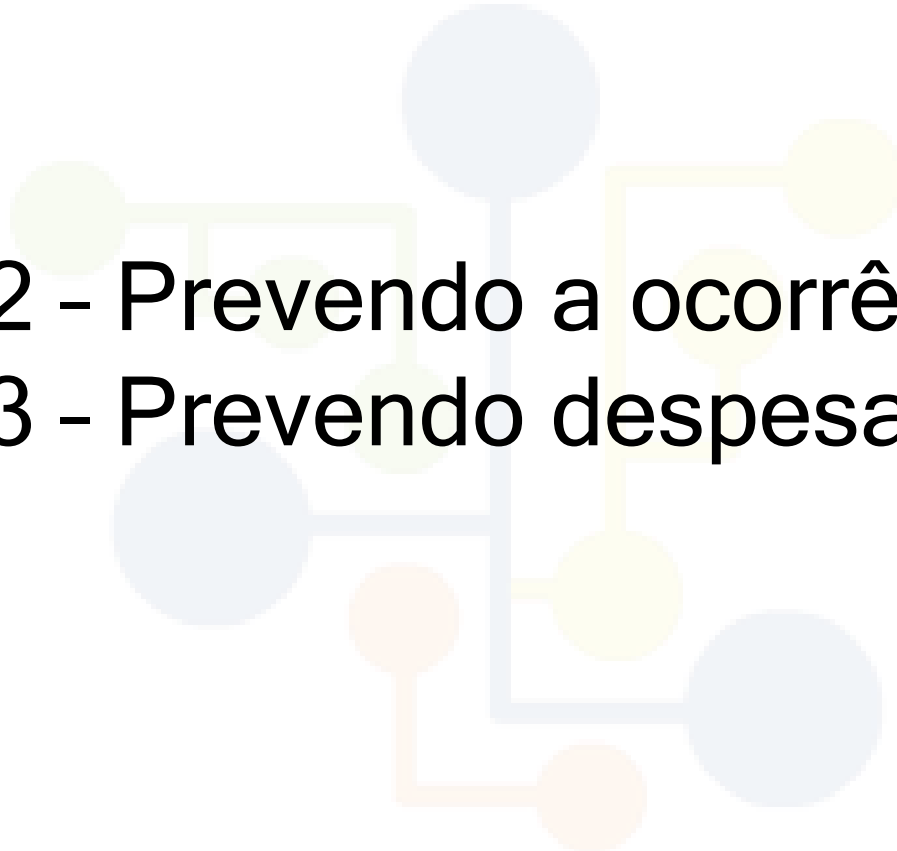

Data Science Academy

# O que veremos neste capítulo:

- Definição de Machine Learning
- Ética em Aprendizado de Máquina
- Frameworks de Machine Learning
- Processo de Aprendizagem
- Treinamento, Validação e Teste
- Modelos
- Algoritmos de Machine Learning
- Regressão, Classificação, Clustering, SVM e Redes Neurais
- Projetos



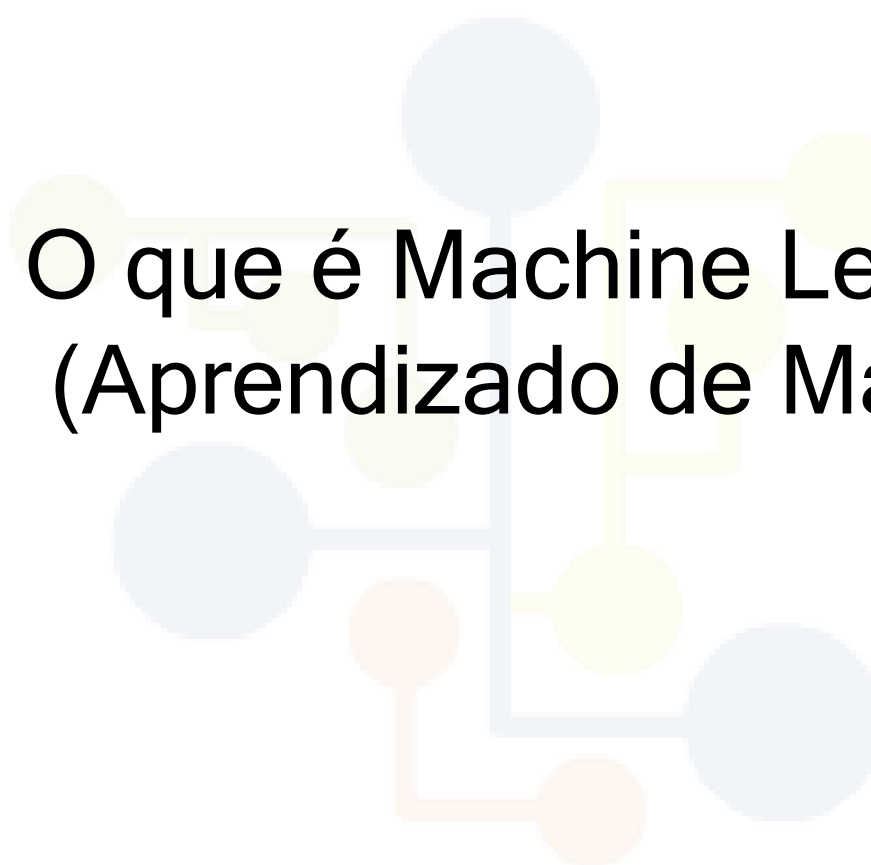

Data Science Academy



Projeto 2 - Prevendo a ocorrência de câncer  
Projeto 3 - Prevendo despesas hospitalares



Data Science Academy

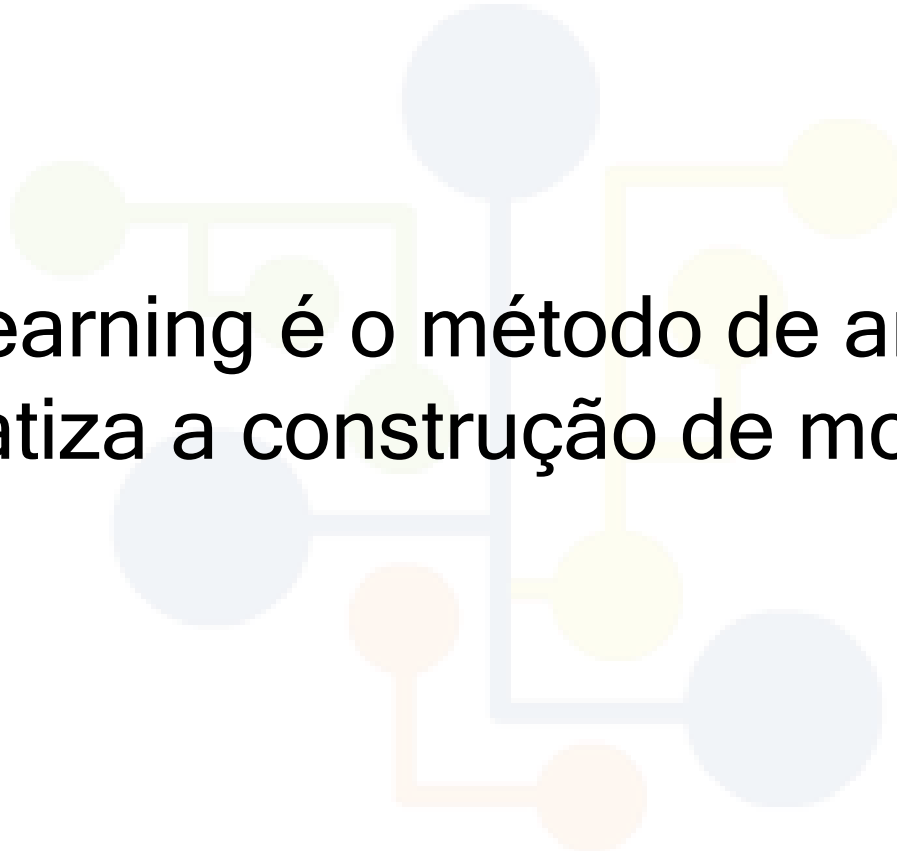



# O que é Machine Learning? (Aprendizado de Máquina)



Data Science Academy





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




Data Science Academy

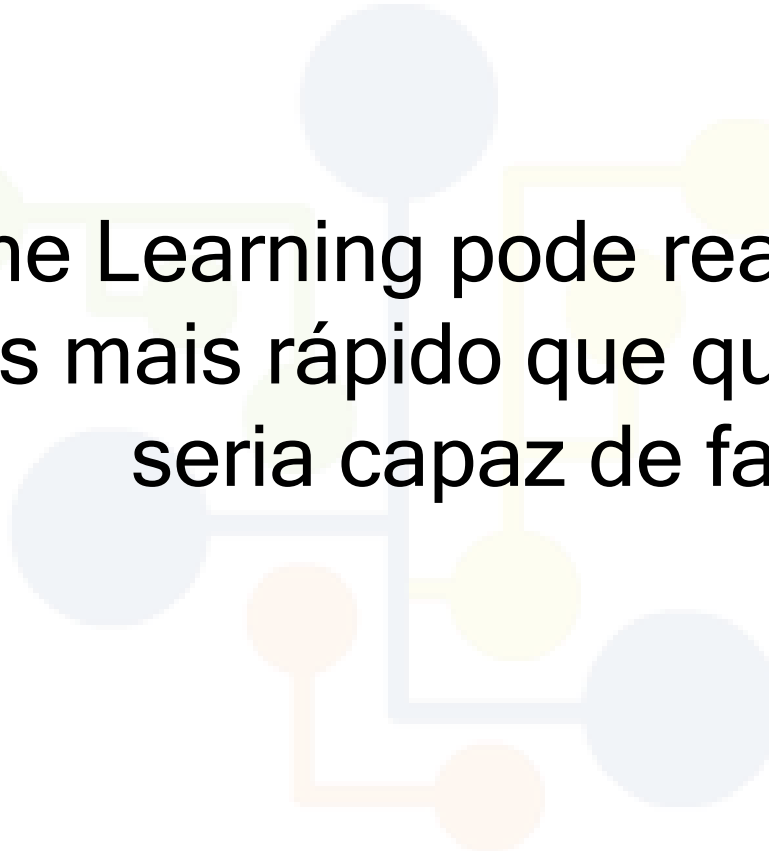
E como as máquinas  
aprendem?



Data Science Academy



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




Data Science Academy




Então, por que Machine Learning é tão dominante hoje em dia?




Data Science Academy




# Machine Learning é um subconjunto da Inteligência Artificial




Data Science Academy



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:



Data Science Academy



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



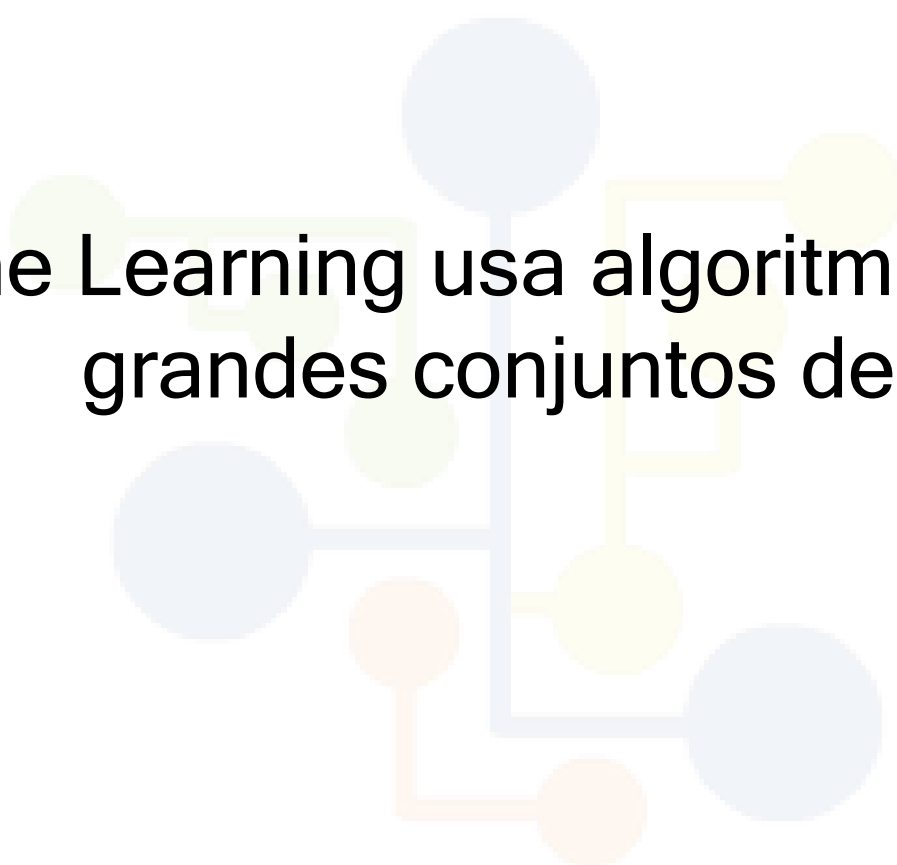

Manipulação de Matrizes

Teoria da Probabilidade

Teorema de Bayes



Data Science Academy

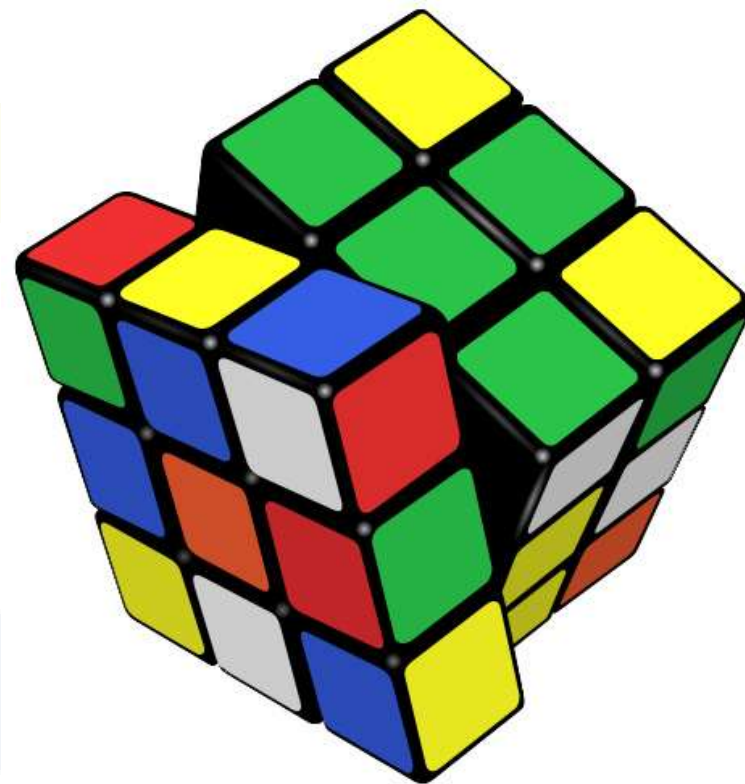


Machine Learning usa algoritmos para analisar  
grandes conjuntos de dados



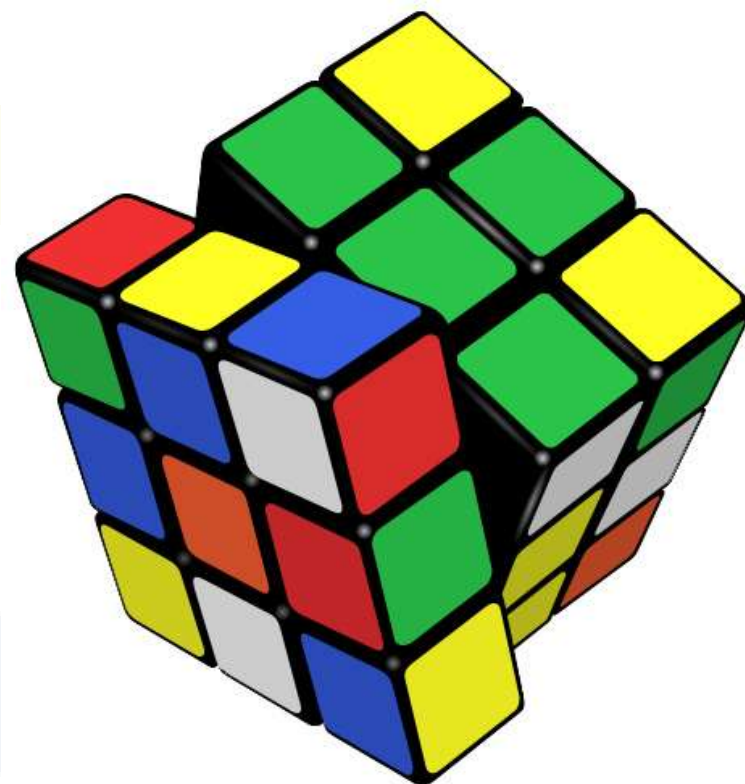
Data Science Academy

Ok entendi, mas o que são  
algoritmos?



Data Science Academy

Algoritmos são procedimentos  
ou fórmulas usados para  
resolver problemas



Data Science Academy

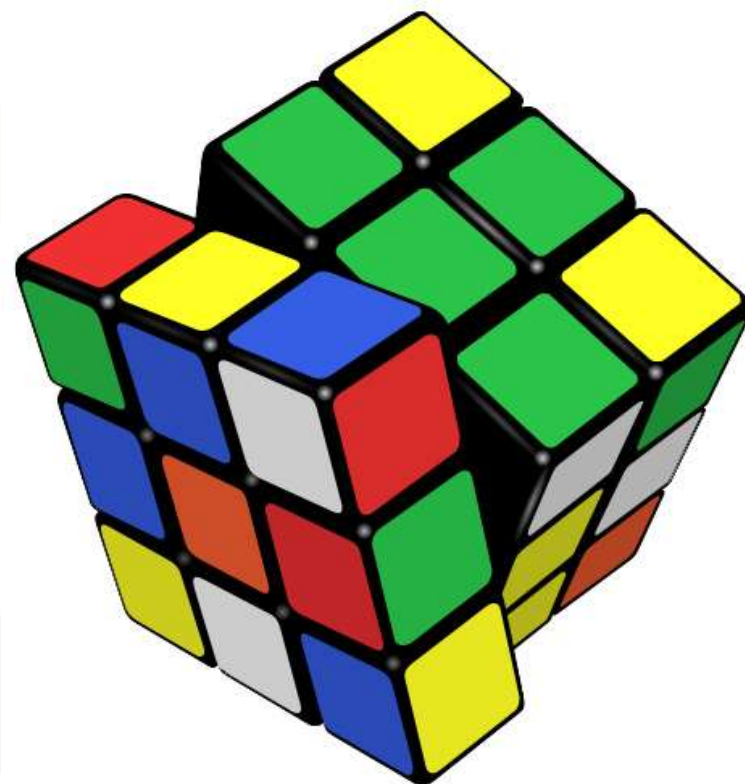
Algoritmos são procedimentos  
ou fórmulas usados para  
resolver problemas



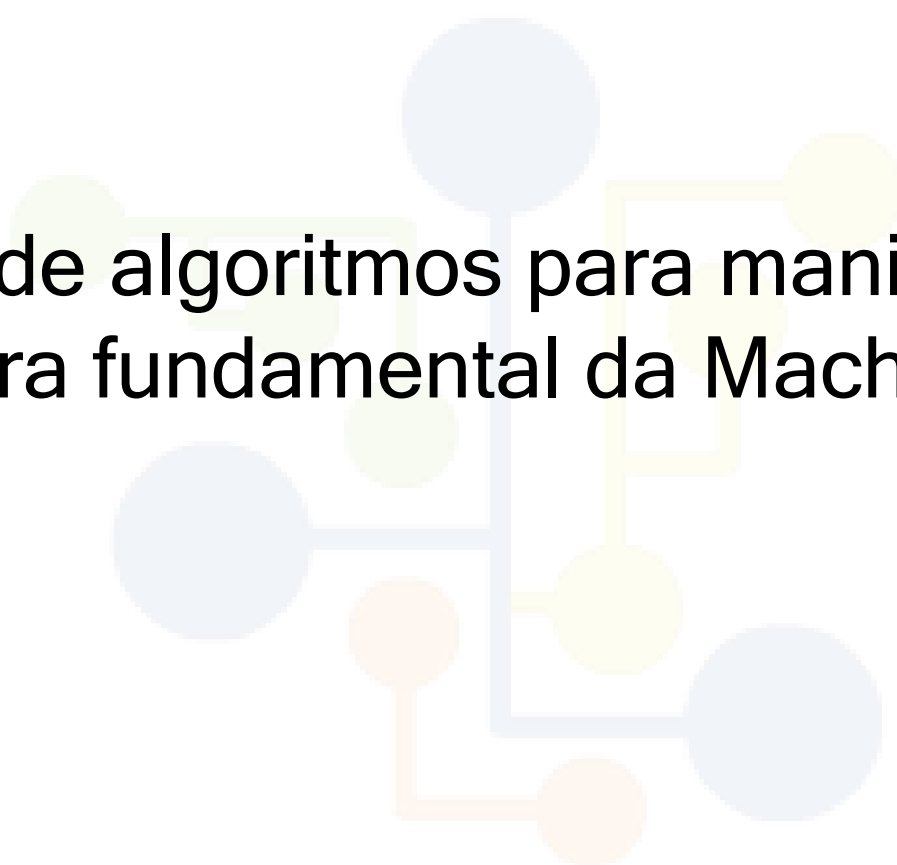

Data Science Academy



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



Falhas são mais comuns que sucesso em processos de Machine Learning



Data Science Academy




# Machine Learning x Data Mining

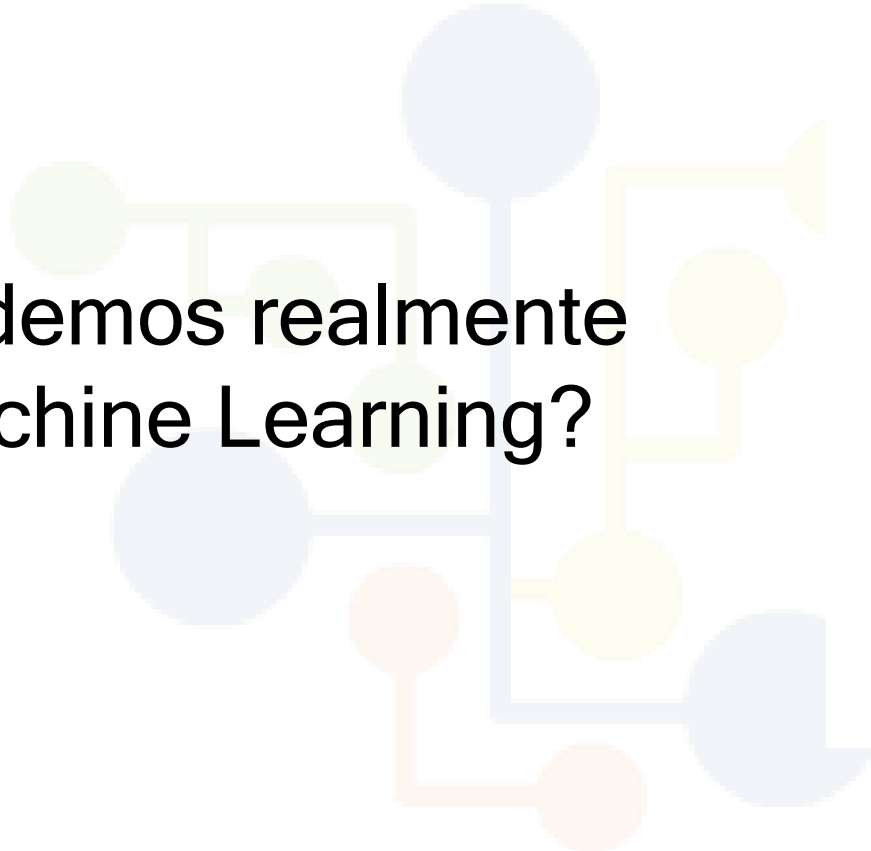
Foco é  
identificar  
padrões e  
solucionar  
problemas



Data Science Academy



E como podemos realmente  
utilizar Machine Learning?



Data Science Academy

## Aplicações de Machine Learning

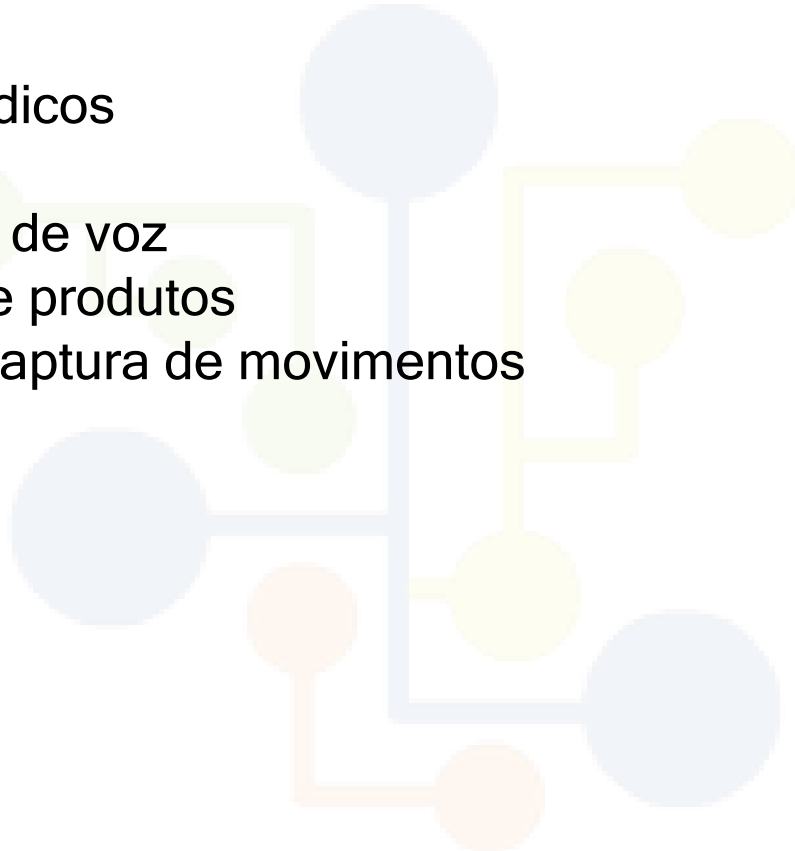
- Detecção de Fraudes
- Motores de Busca
- Advertising em tempo real em páginas web
- Score de crédito
- Predição de falhas em equipamentos
- Modelos de precificação
- Detecção de invasão de redes
- Sistemas de Recomendação
- Segmentação de Clientes
- Análise de Sentimentos em Textos
- Reconhecimento de imagens e Padrões
- Filtro de Spam
- Modelagem Financeira




Data Science Academy



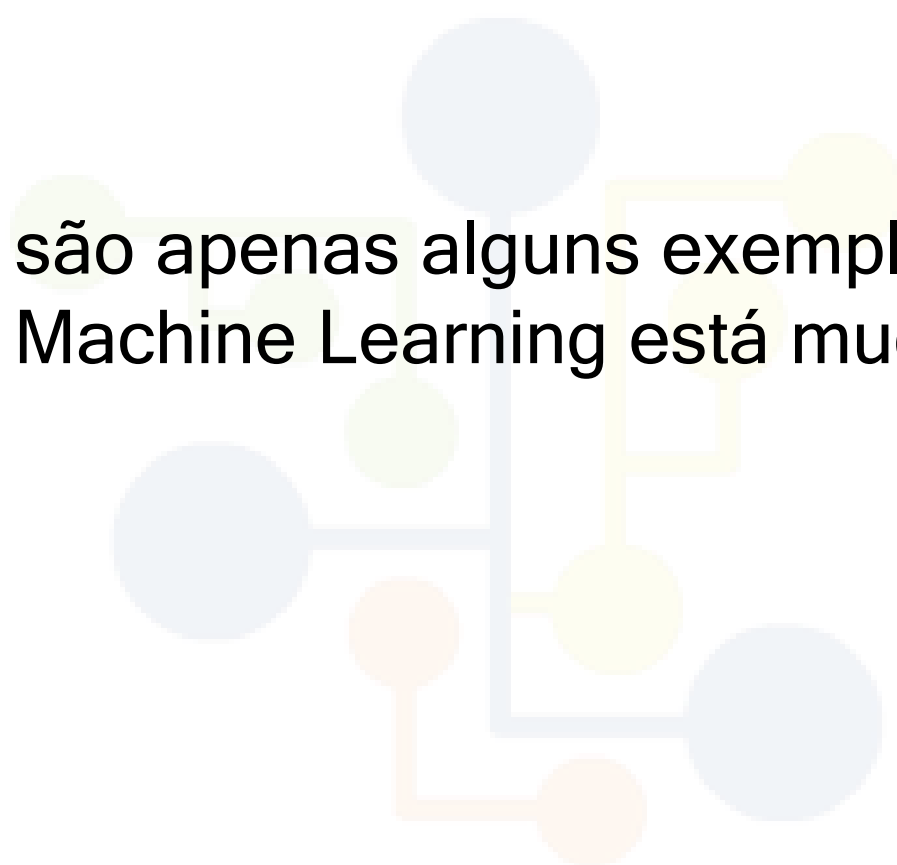
E tem mais...

- Diagnósticos médicos
  - Bioinformática
  - Reconhecimento de voz
  - Categorização de produtos
  - Tecnologias de captura de movimentos
- 





Esses são apenas alguns exemplos que mostram  
como a Machine Learning está mudando nossa vida



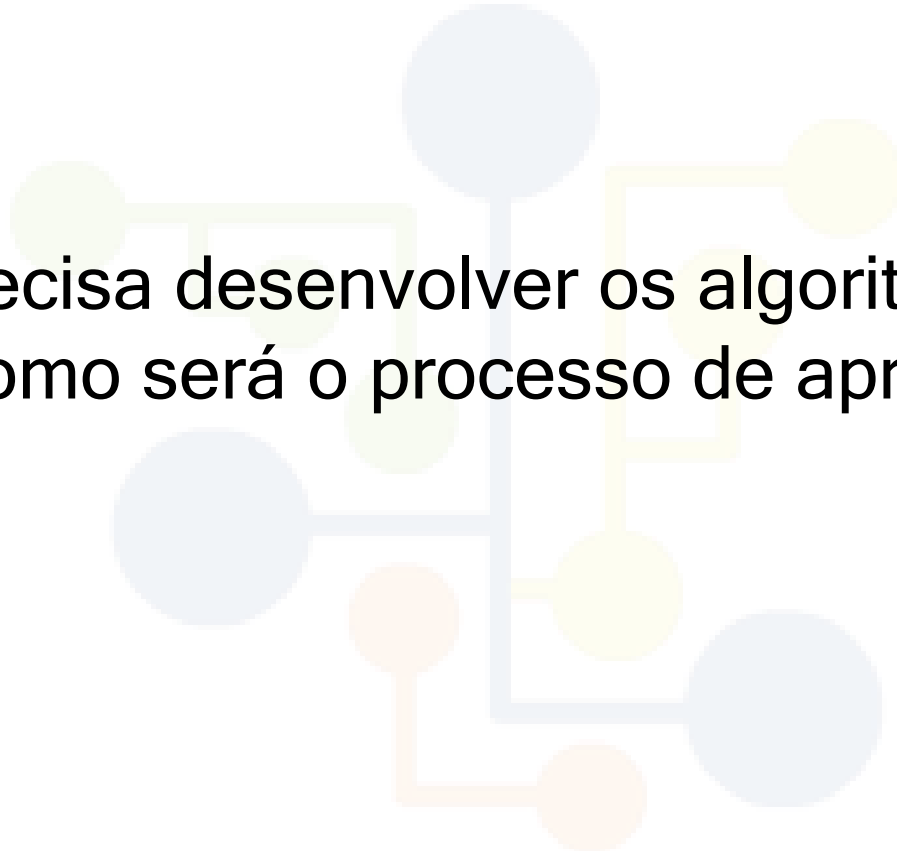

Data Science Academy



E como a Machine Learning é criada?



Data Science Academy



Você precisa desenvolver os algoritmos e dizer a eles como será o processo de aprendizagem.




Data Science Academy

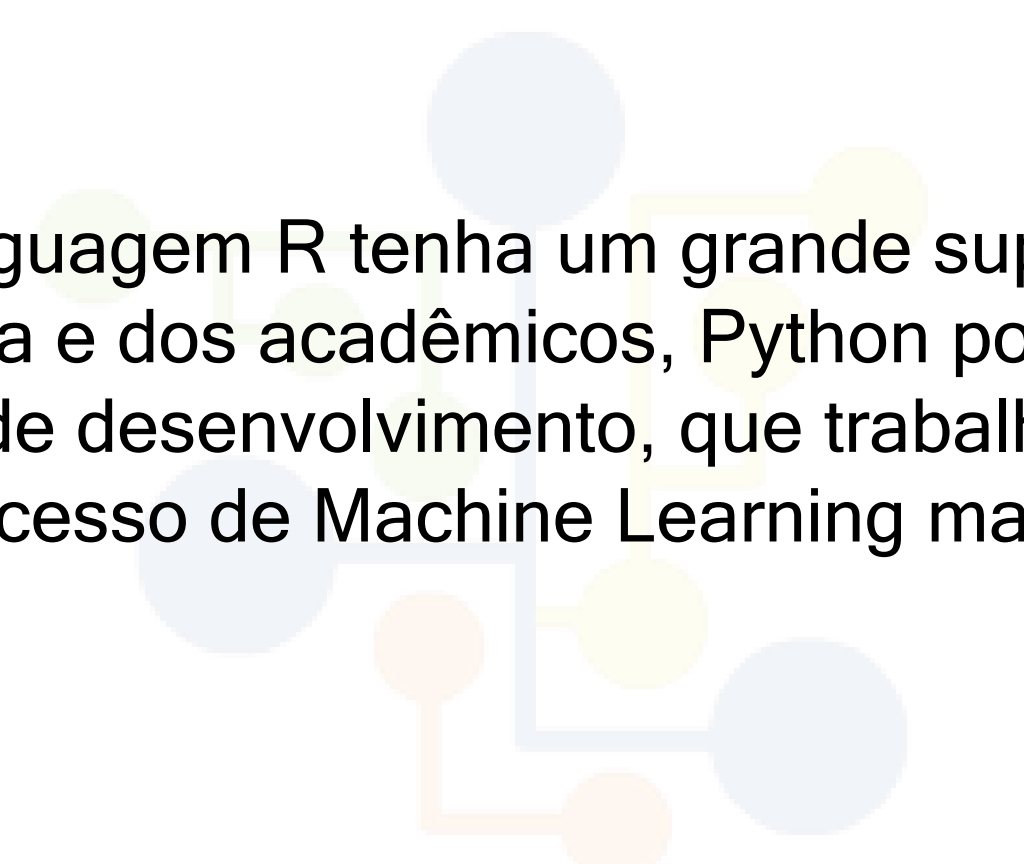




Data Science Academy

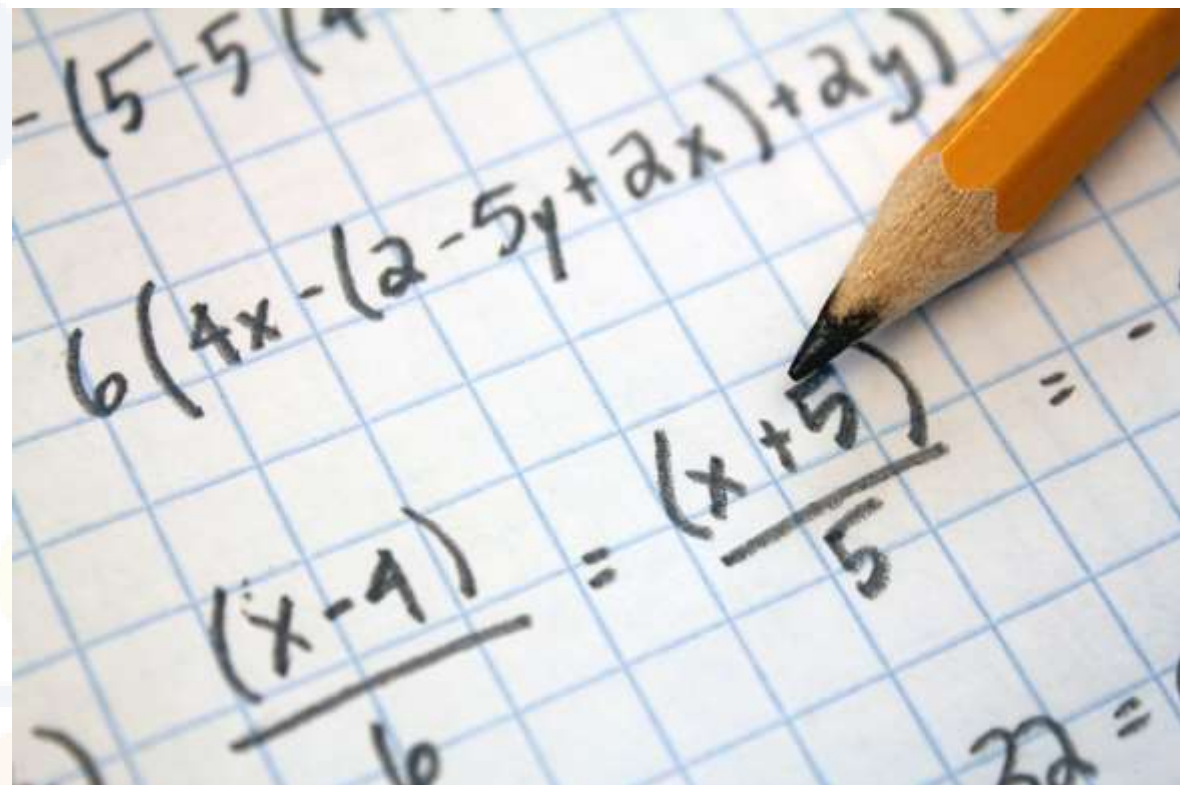


Embora a linguagem R tenha um grande suporte da comunidade Estatística e dos acadêmicos, Python possui uma grande comunidade de desenvolvimento, que trabalha intensamente para fazer o processo de Machine Learning mais fácil e amigável.




Data Science Academy

Matemática é a base  
da Machine Learning

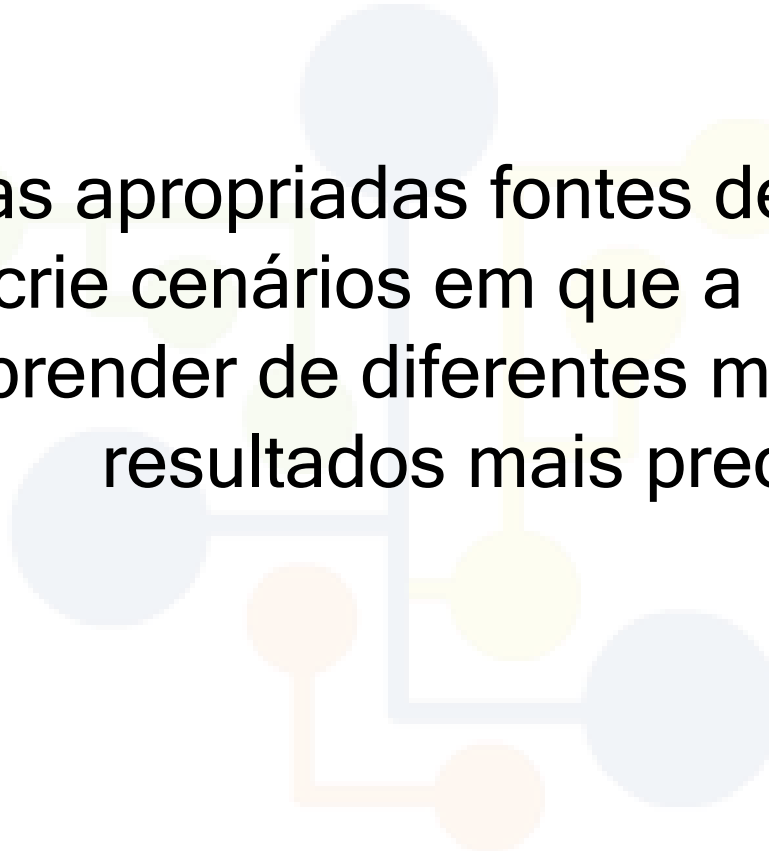


Data Science Academy

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



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



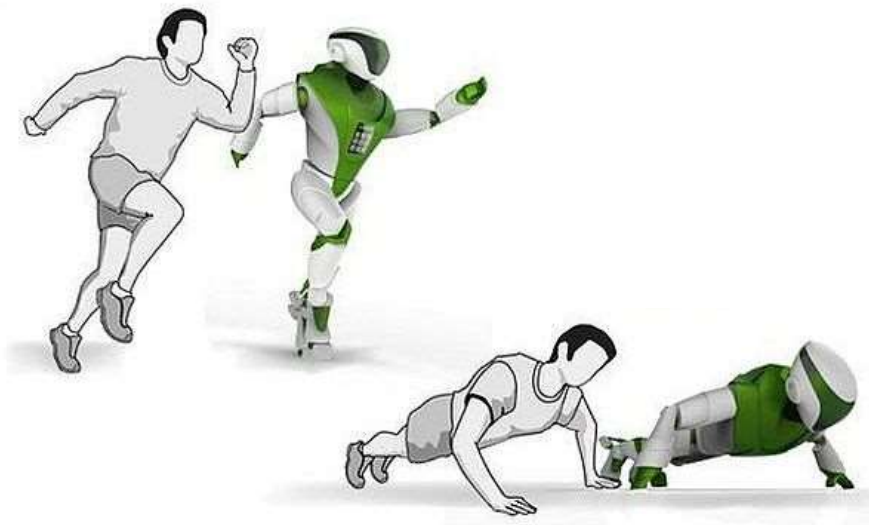
Data Science Academy

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



Data Science Academy

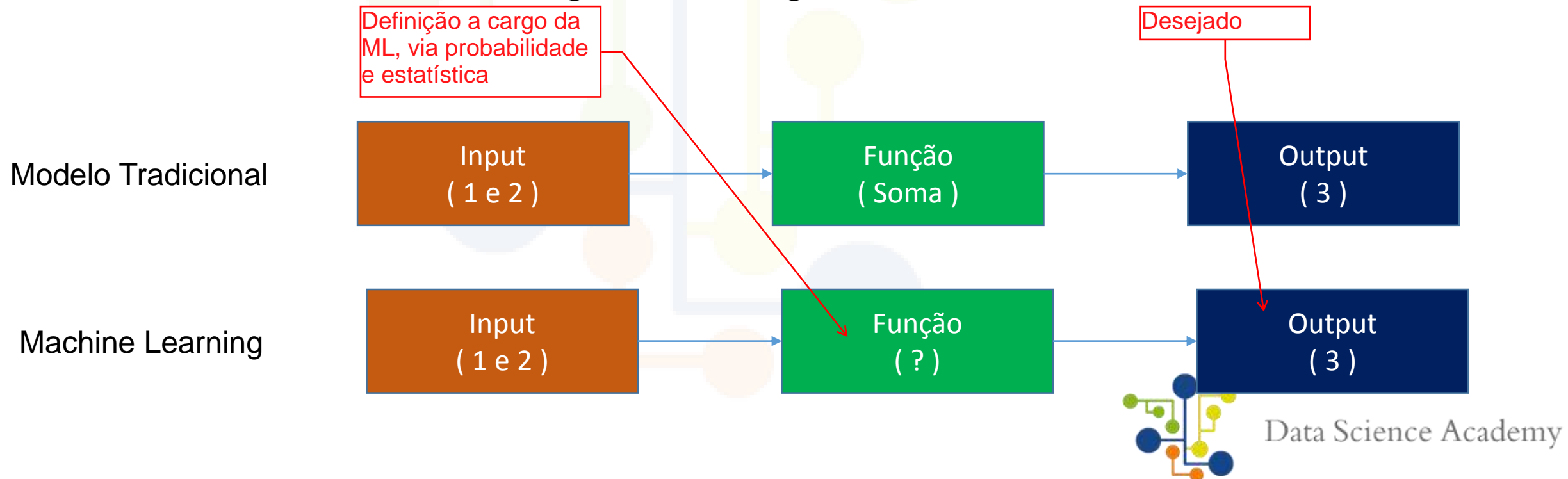
Mas antes que um  
algoritmo possa aprender,  
você precisa treiná-lo




Data Science Academy

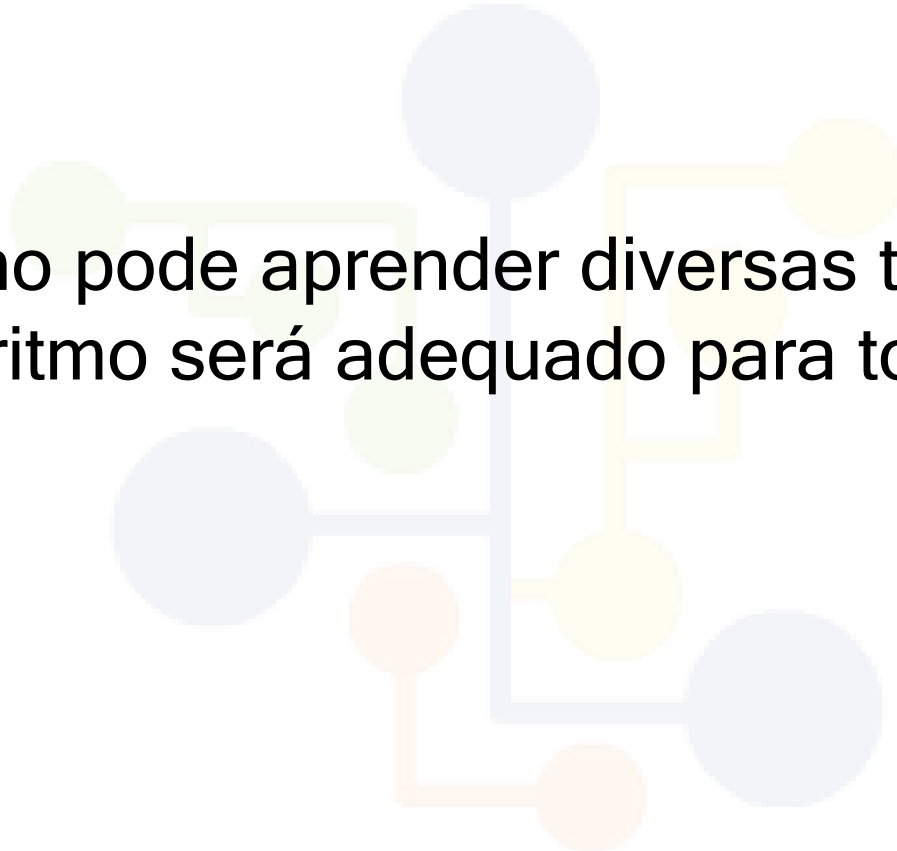


# O processo de treino de algoritmos de Machine Learning usa o seguinte conceito






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 ainda está  
no começo do que pode ser  
feito com esta tecnologia



Data Science Academy



Opa. Eu ouvi bem? Você está dizendo que o Big Data, esta imensidão de dados, ainda é pouco?



Data Science Academy



E o futuro?



Data Science Academy



## E o futuro?

É difícil prever quão longe a ML pode chegar, mas é certo que ela avançará cada vez mais rápido. As máquinas ainda não pensam sozinhas e precisam de seres humanos que digam a elas o que fazer. E acredite, este cenário não vai mudar tão cedo e talvez nossa geração não veja esta mudança. Tecnologias como Machine Learning, Inteligência Artificial e Computação Cognitiva estão na sua infância e há muito o que evoluir. Muito mesmo!! Os profissionais que se dedicarem a estas áreas, serão os verdadeiros unicórnios, que tanto se fala.



Data Science Academy





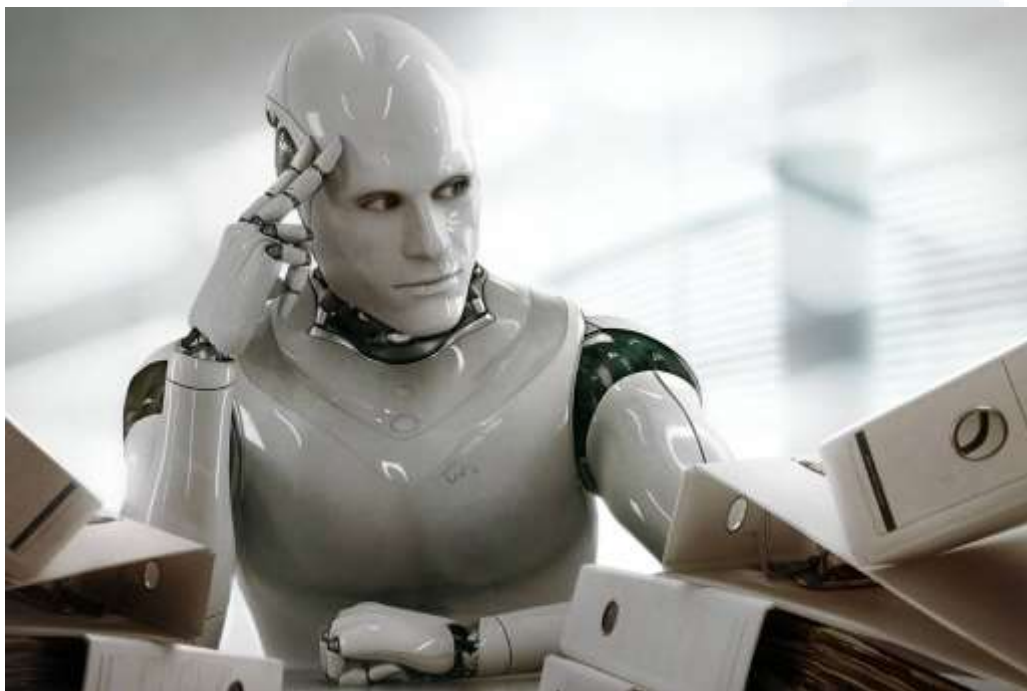
## E o futuro?

Outro ponto a se considerar e que reforçamos muito em nossos cursos, é que o profissional não deve focar apenas na parte técnica. Ela é importante, mas está em constante evolução. O profissional deve também adquirir conhecimento em negócios e por isso mesmo incluímos um curso inteiro sobre este tema na Formação Cientista de Dados, o curso de Business Analytics. Depois de adquirir o conhecimento técnico, vamos praticar a solução de problemas de negócio. Esse é o principal diferencial de um Cientista de Dados: ser capaz de identificar um problema e saber como resolvê-lo usando Analytics e Big Data.



Data Science Academy





Você pode ter uma  
máquina como chefe?

<http://www.hitachi.com/New/cnews/month/2015/09/150904.html>




Data Science Academy



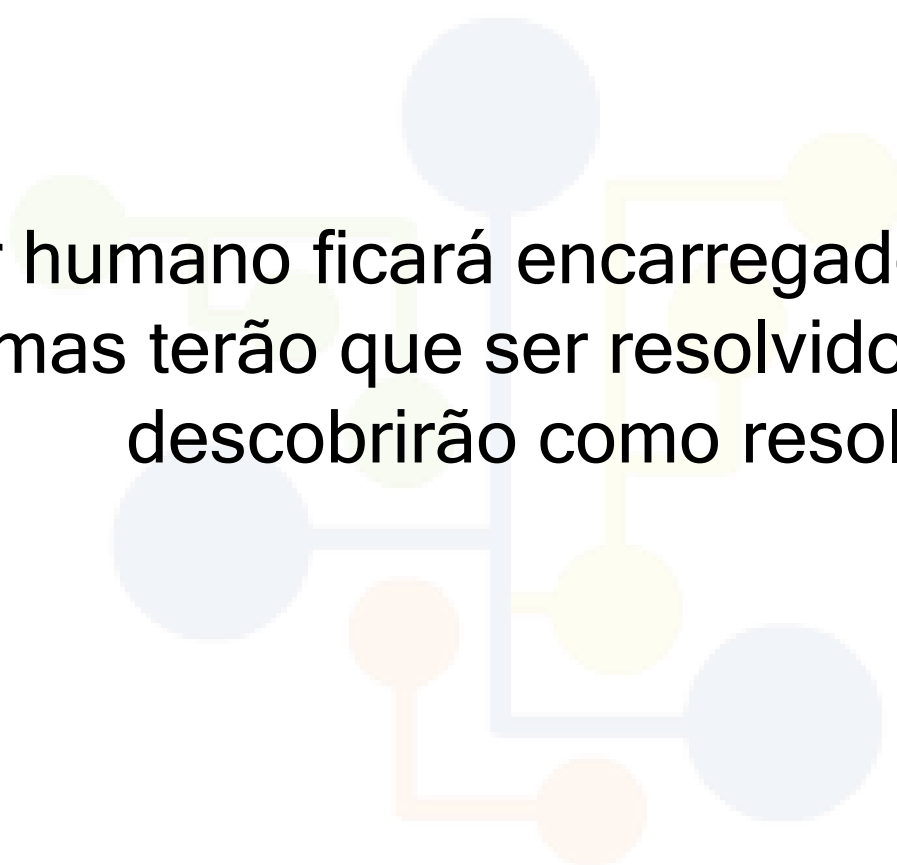
Mas claro, Machine Learning traz muitas  
oportunidades também



Data Science Academy



O ser humano ficará encarregado de definir que problemas terão que ser resolvidos e as máquinas descobrirão como resolvê-los




Data Science Academy

Para usar Machine Learning em todo seu potencial, as empresas vão precisar de profissionais capacitados



Data Science Academy



Máquinas não são boas em fazer perguntas  
(humanos são bons nisso)

Mas as máquinas são muito boas em responder perguntas  
(melhor e mais rápido que os humanos)



Data Science Academy



De que lado você pretende estar?

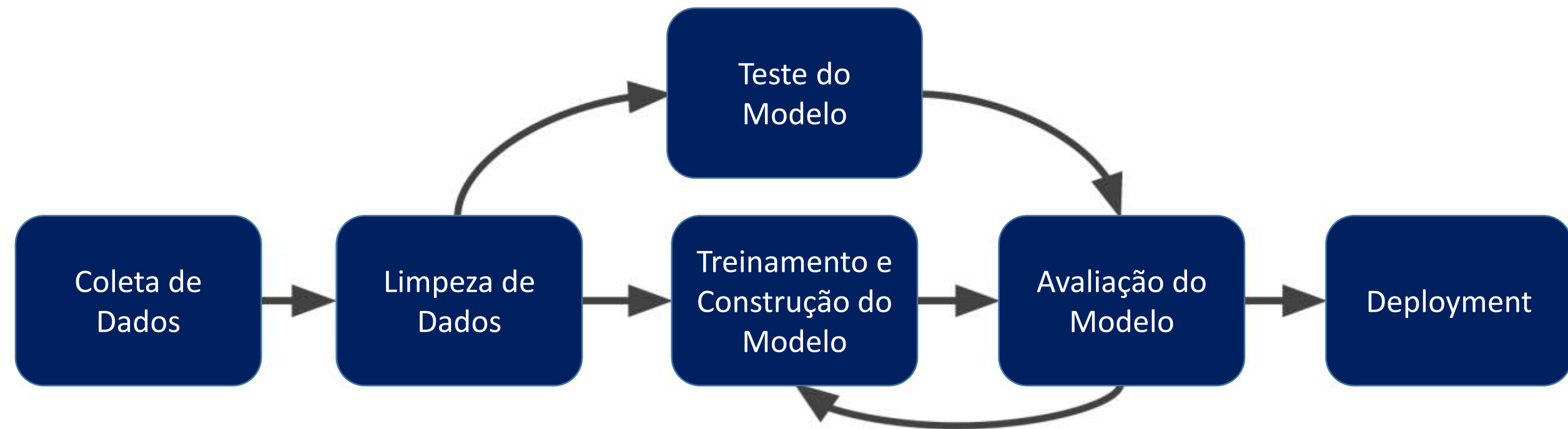
Do lado dos que receberão ordens de máquinas ou do lado dos  
que vão programá-las?

Pense nisso!!

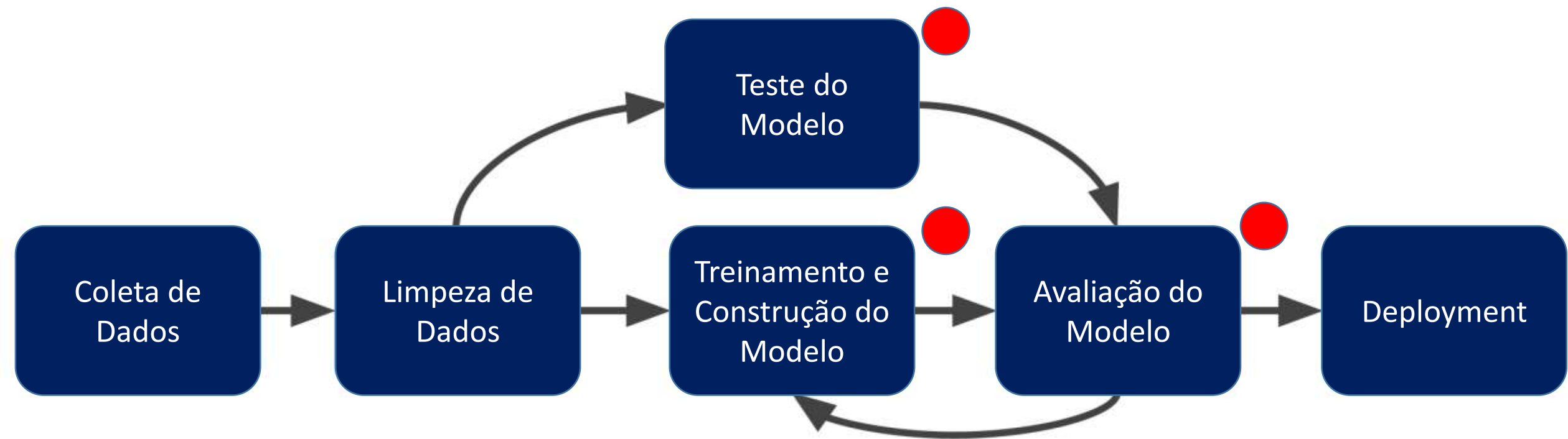


Data Science Academy





Data Science Academy



Data Science Academy






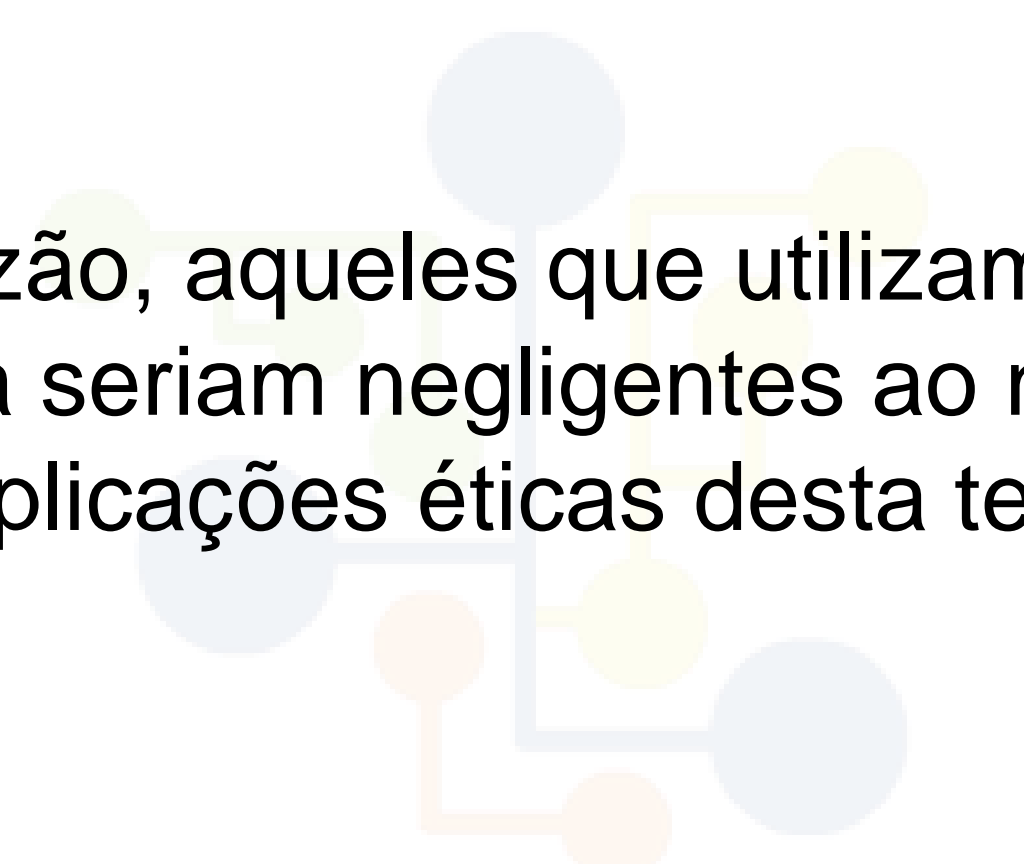
# Ética no Aprendizado de Máquina



Data Science Academy



Por esta razão, aqueles que utilizam a aprendizagem de máquina seriam negligentes ao não considerar as implicações éticas desta tecnologia



Data Science Academy

# Ética no Aprendizado de Máquina



Data Science Academy



Vejamos um exemplo



Data Science Academy





Data Science Academy

# Machine Learning em ação



Data Science Academy





## Machine Learning em ação



Data Science Academy

# Machine Learning em ação



Data Science Academy





# Machine Learning em ação



Data Science Academy




Data Science Academy



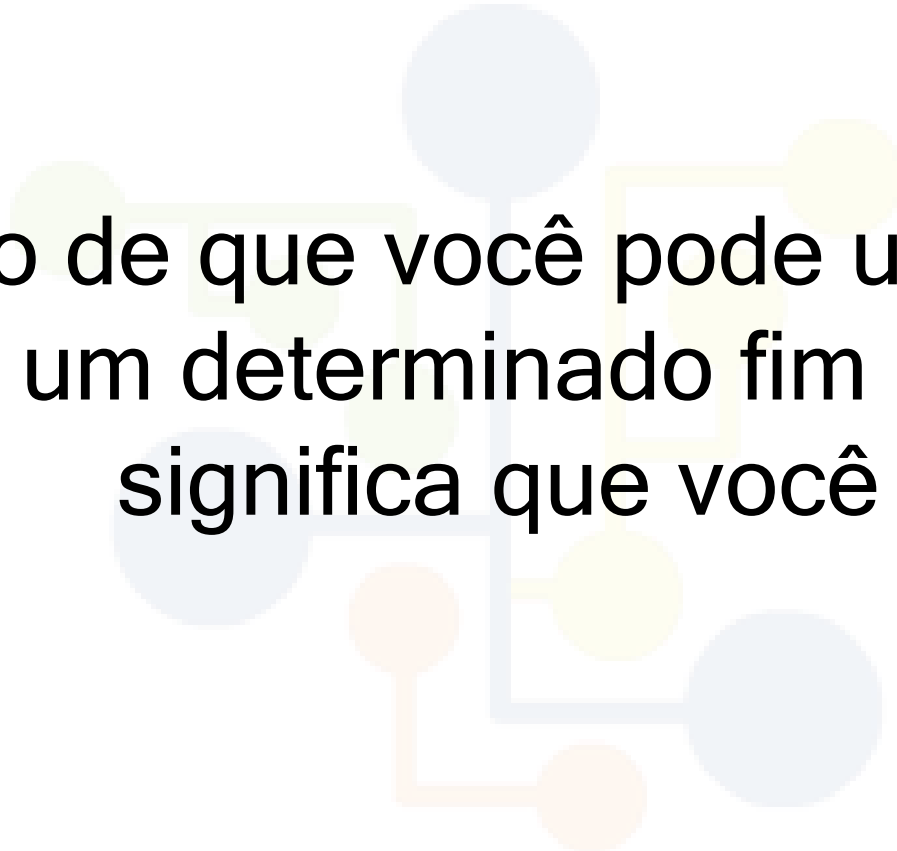
Algumas jurisdições podem impedir a utilização de dados raciais, étnicos, religiosos ou mesmo por razões comerciais. Tenha em mente que a exclusão de dados de sua análise pode não ser suficiente, porque algoritmos de aprendizado de máquina podem inadvertidamente aprender esta informação de forma independente. Por exemplo, se um determinado segmento de pessoas geralmente vive em uma determinada região, e compram um determinado produto ou ainda se comportam de uma maneira que os identifica de forma única como um grupo, alguns algoritmos de aprendizado de máquina podem inferir esta informação e identificar estes grupos. Em tais casos, pode ser necessário excluir quaisquer dados que potencialmente identifiquem a informação.




Data Science Academy



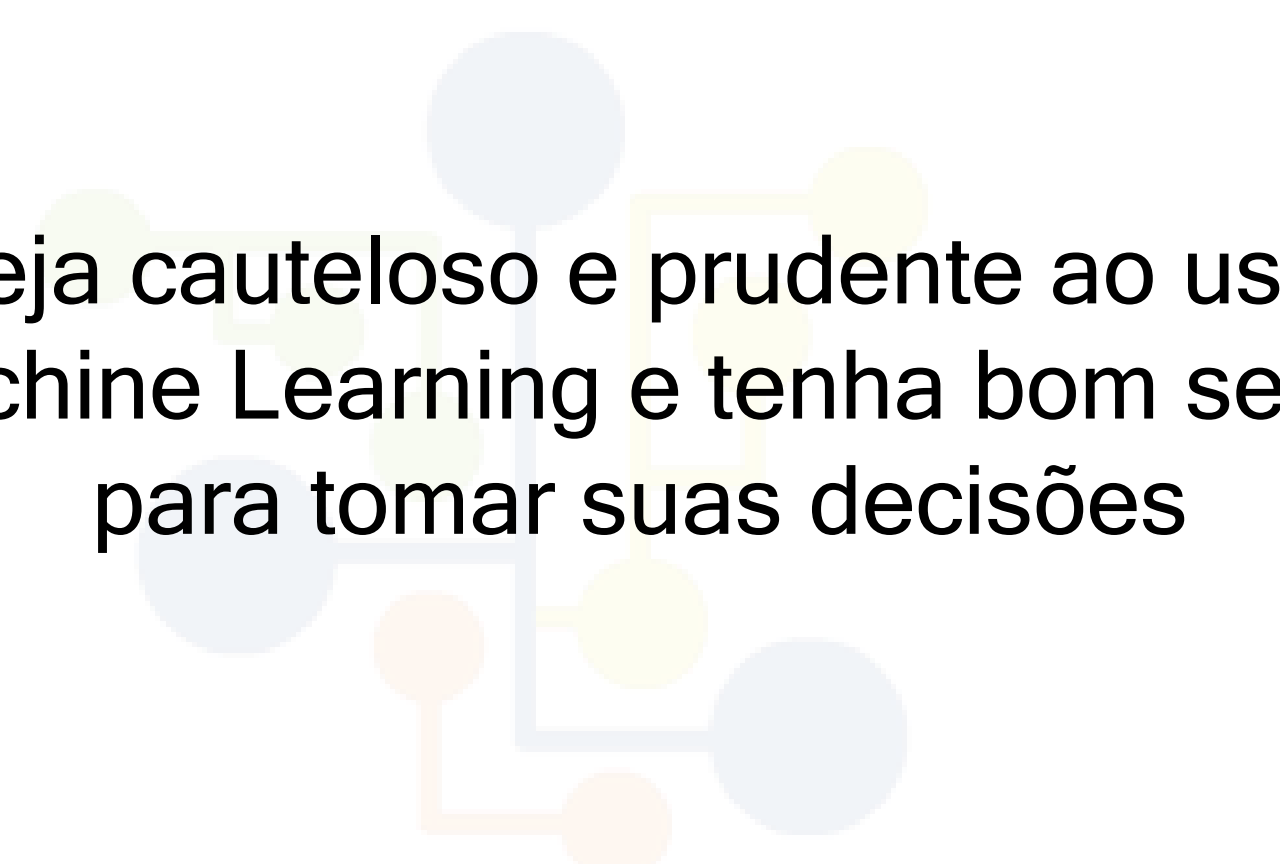
O fato de que você pode usar os dados  
para um determinado fim nem sempre  
significa que você deve



Data Science Academy



Seja cauteloso e prudente ao usar  
Machine Learning e tenha bom senso  
para tomar suas decisões



Data Science Academy



# Machine Learning Frameworks

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



Data Science Academy





# E por que usar Frameworks de Machine Learning?



Data Science Academy



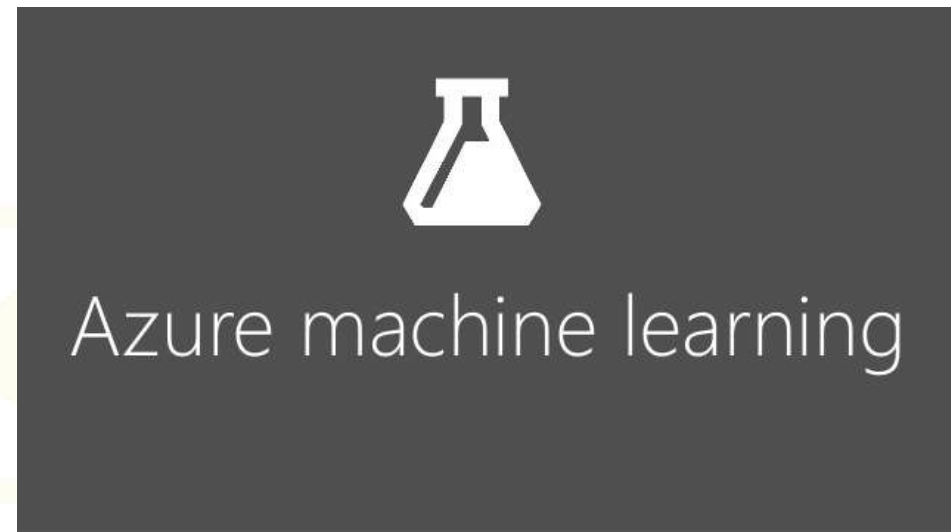
# E por que usar Machine Learning Frameworks?



Data Science Academy



# 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.



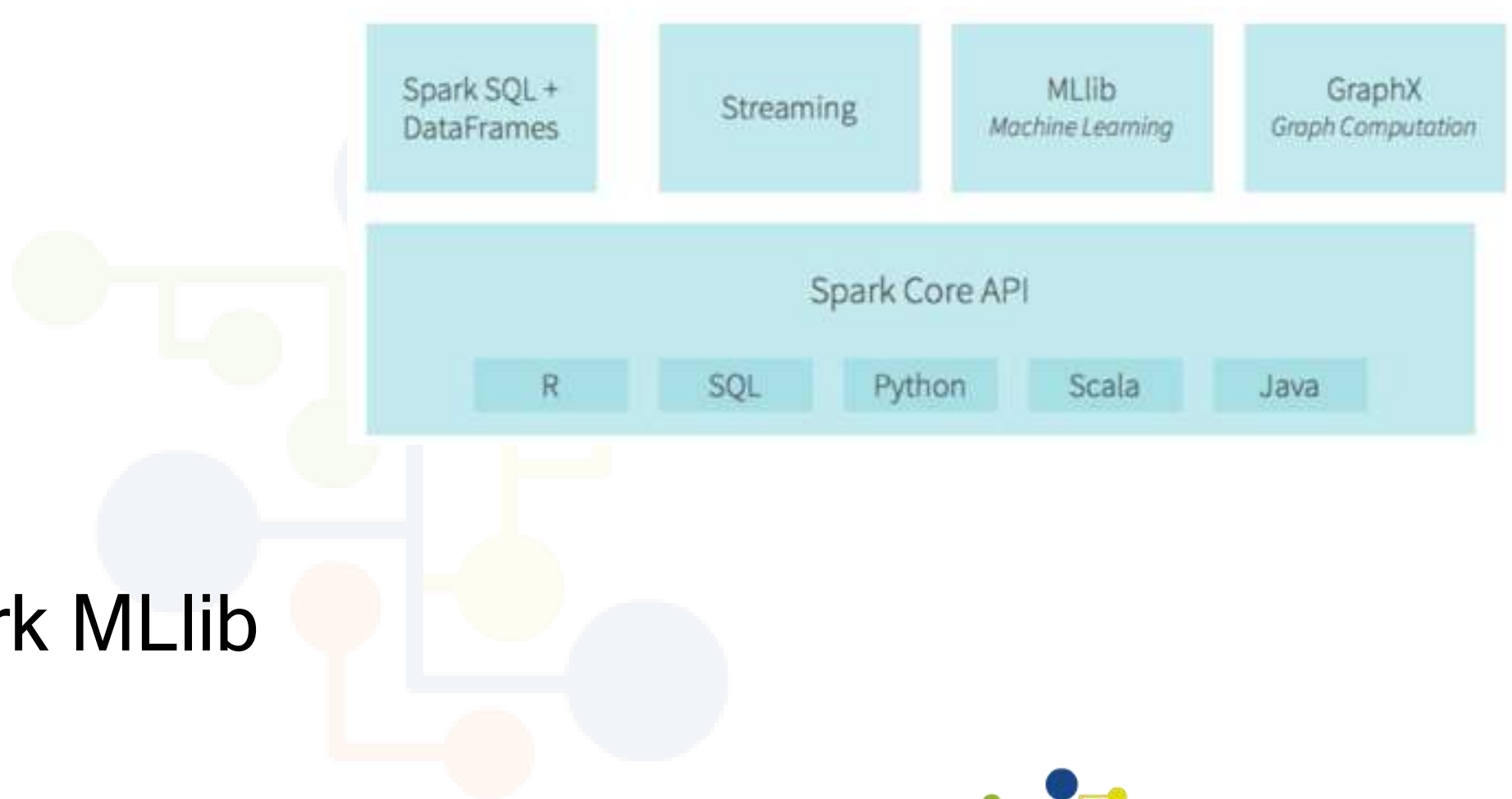
Data Science Academy

# Apache Spark MLlib



Data Science Academy

# Apache 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



# Caffe

Outro para deep learning, alta capacidade para  
processar muitas imagens



# Caffe



Data Science Academy



# Nervana

Nervana neon, para deep learning.  
Foco no hardware.



Data Science Academy





# Outras Ferramentas



Data Science Academy

# Weka

Waikato Environment for Knowledge Analysis (Weka)



Feito em Java. Uso em data mining



Data Science Academy



Data Science Academy



**rapidminer**



Data Science Academy



# O Processo de Aprendizagem



Data Science Academy







# Processo de Aprendizagem



Data Science Academy



# Processo de Aprendizagem



Data Science Academy

# Processo de Aprendizagem



Data Science Academy



# Processo de Aprendizagem

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

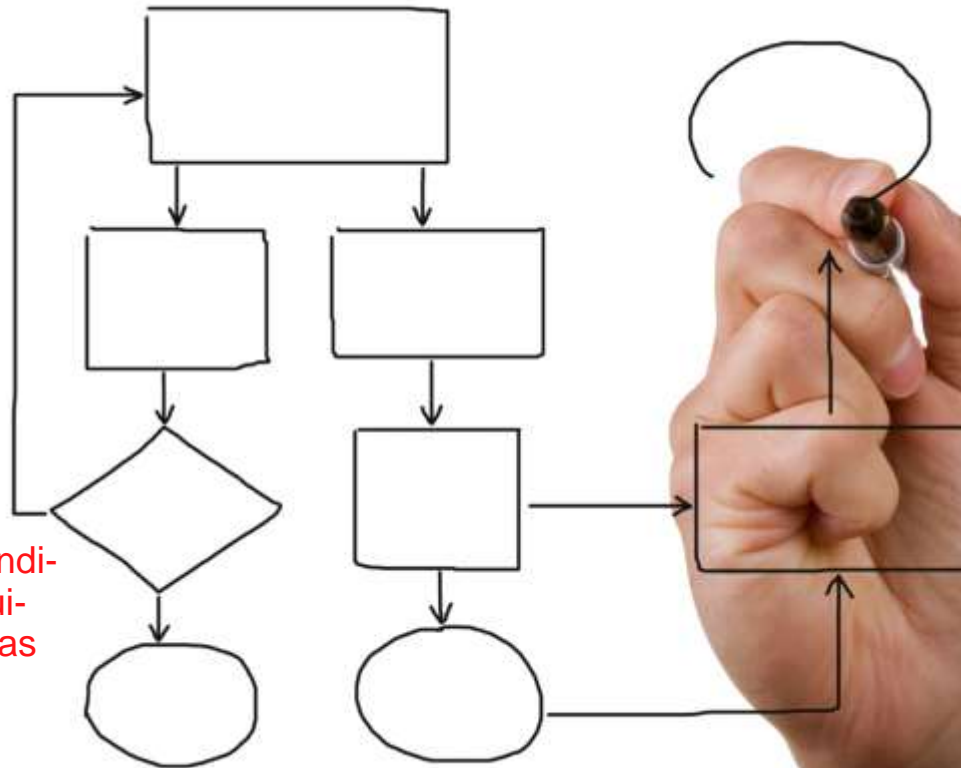
Aprendizagem Supervisionada, Aprendizagem Não Supervisionada e Reinforcement Learning



Data Science Academy

# Processo de Aprendizagem

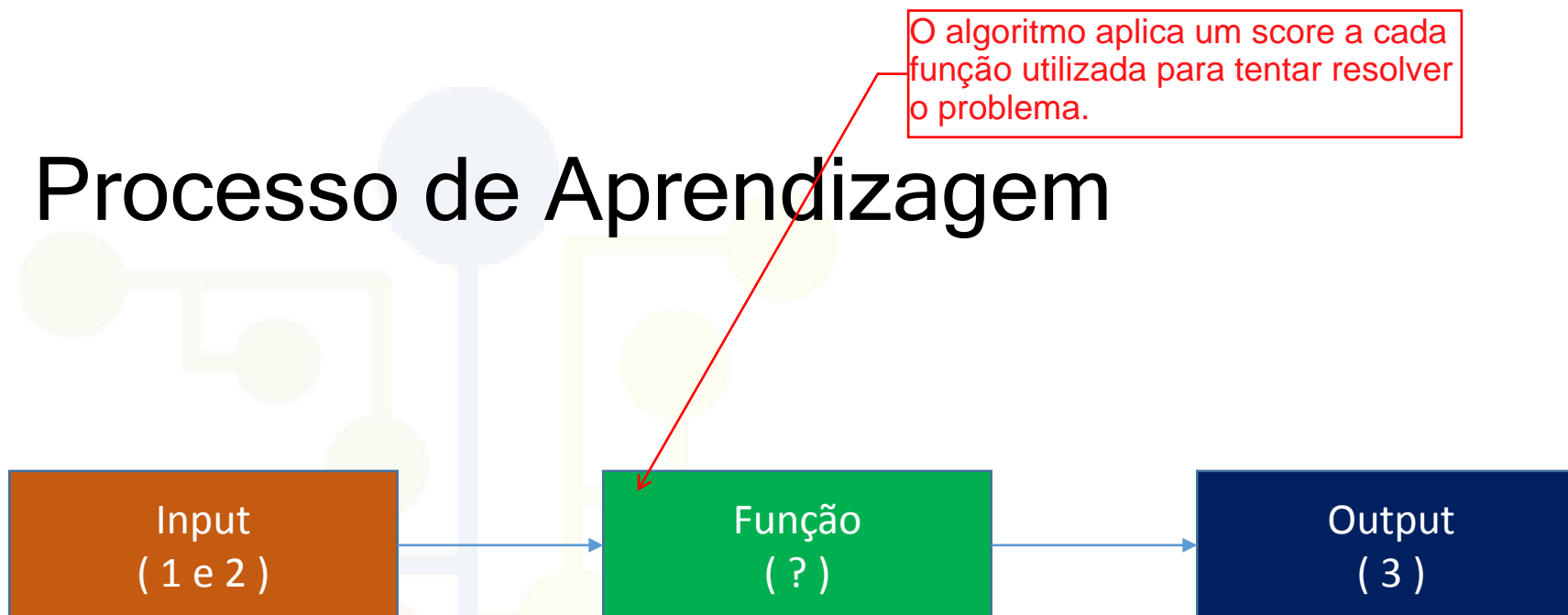
Do ponto de vista matemático, o processo de representação de aprendizagem de máquina pode ser expressado utilizando mapeamento equivalente. Mapeamento é a construção de uma função observando suas saídas.




Data Science Academy



Machine Learning



Data Science Academy




Vamos ver quais são os 3 tipos principais de aprendizagem e então voltaremos ao processo de aprendizagem.

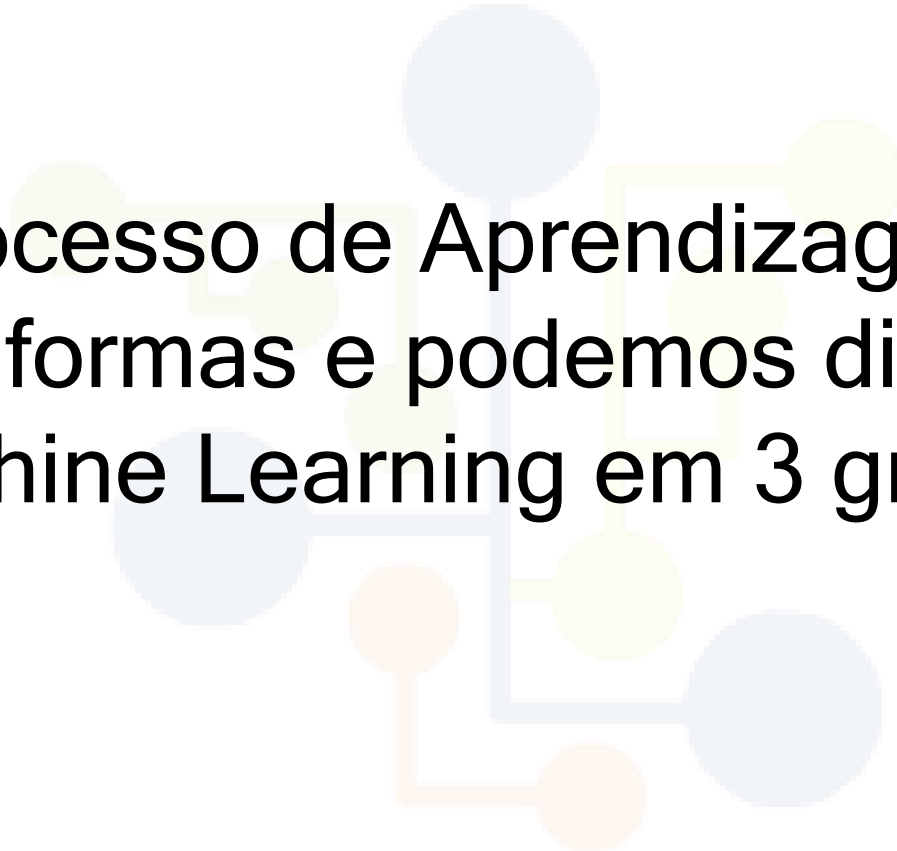
Ainda há muito a estudar sobre isso.



Data Science Academy



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



Data Science Academy



- 
- 
- Aprendizagem Supervisionada
  - Aprendizagem Não Supervisionada
  - Aprendizagem por Reforço (Reinforcement Learning)



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

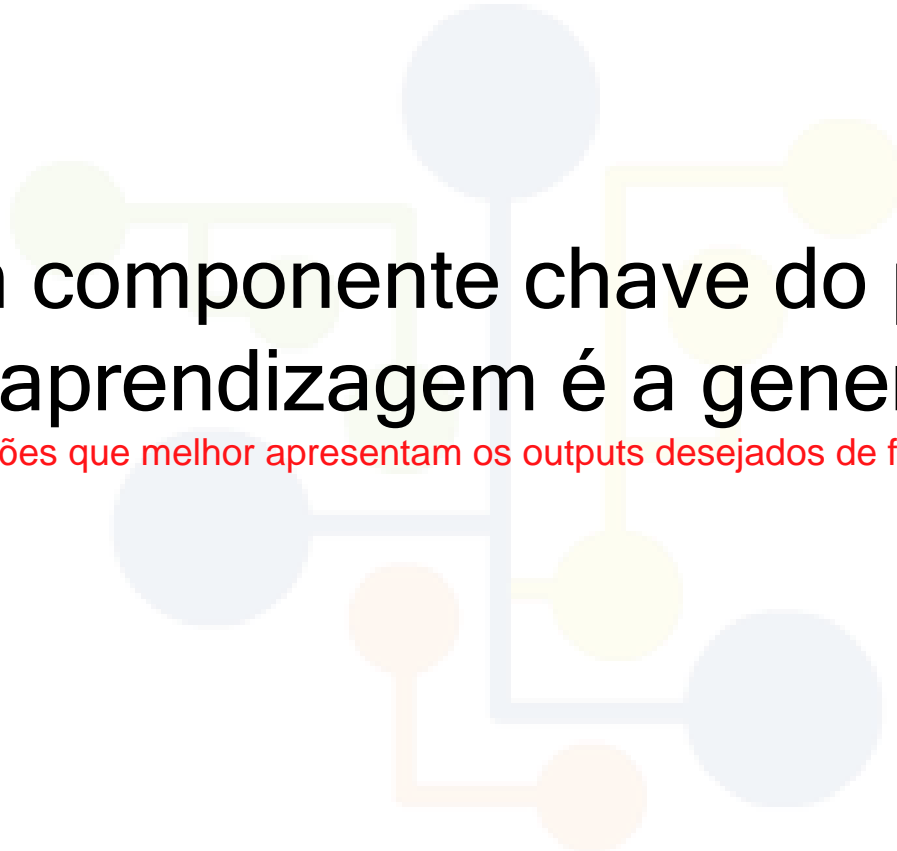

→ **Agente** tomador de decisão - próprio algoritmo

**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

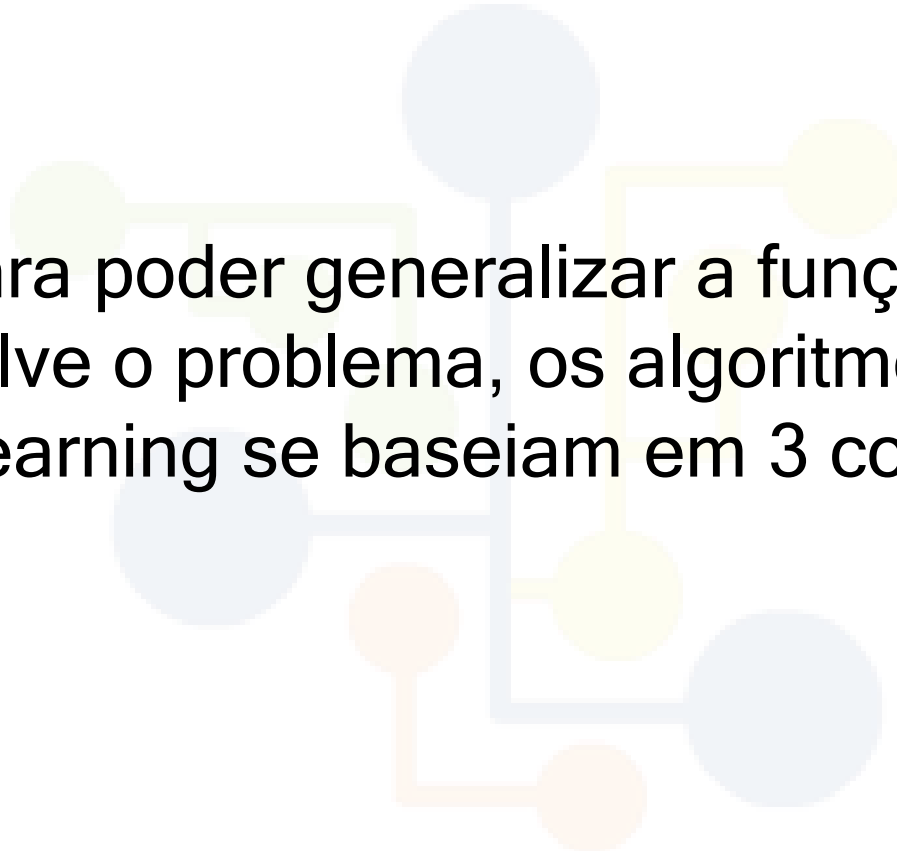
O objetivo é generalizar as funções que melhor apresentam os outputs desejados de forma que a mesma solução possa ser dada a outros conjuntos de dados




Data Science Academy



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



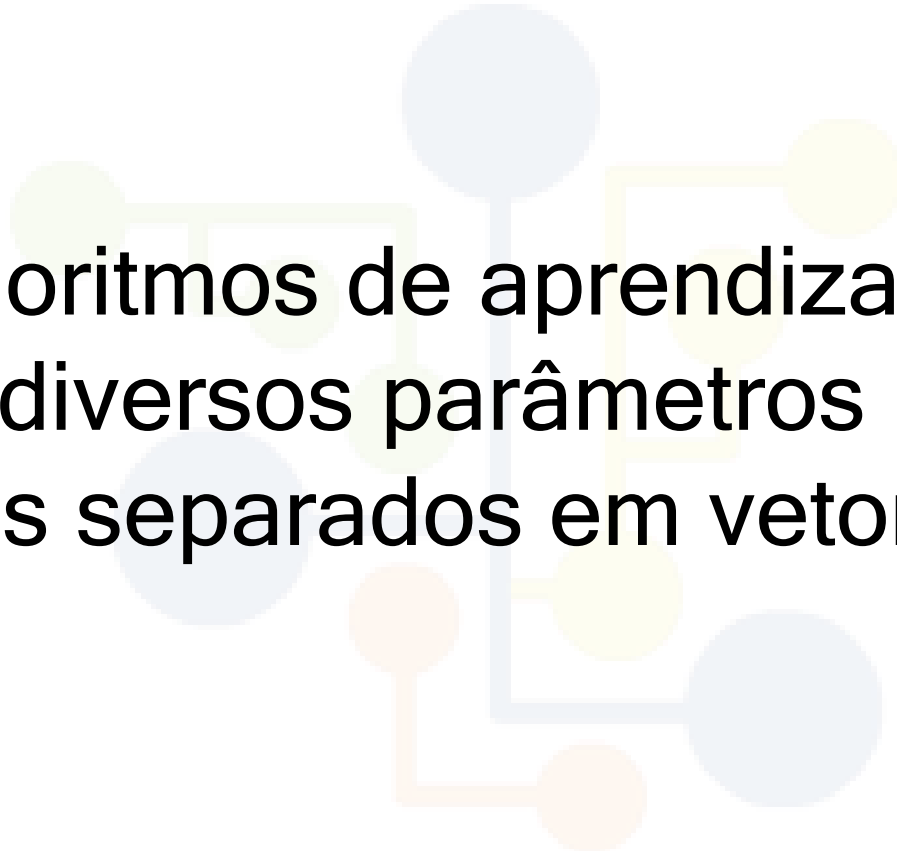

Data Science Academy

- 
- **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



Otimização



Data Science Academy



# Espaço de Hipótese

Contem as variações de parâmetros de ML



Data Science Academy

# Falso Positivo



Data Science Academy



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



Data Science Academy




False Positive → desperdício de tempo

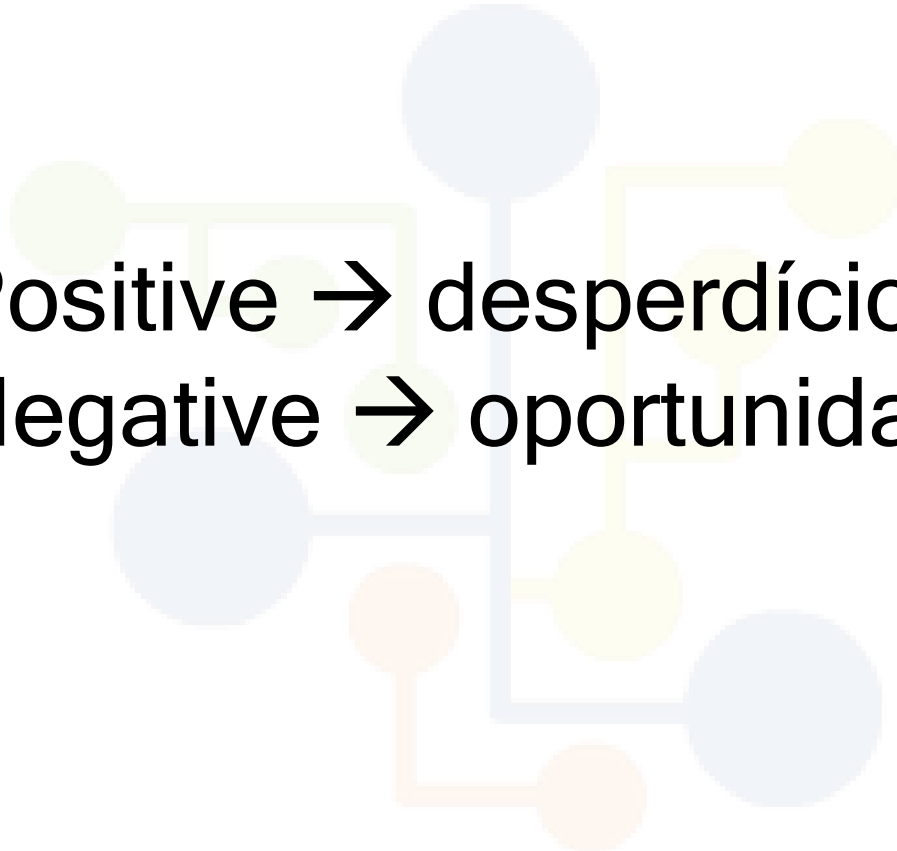


Data Science Academy






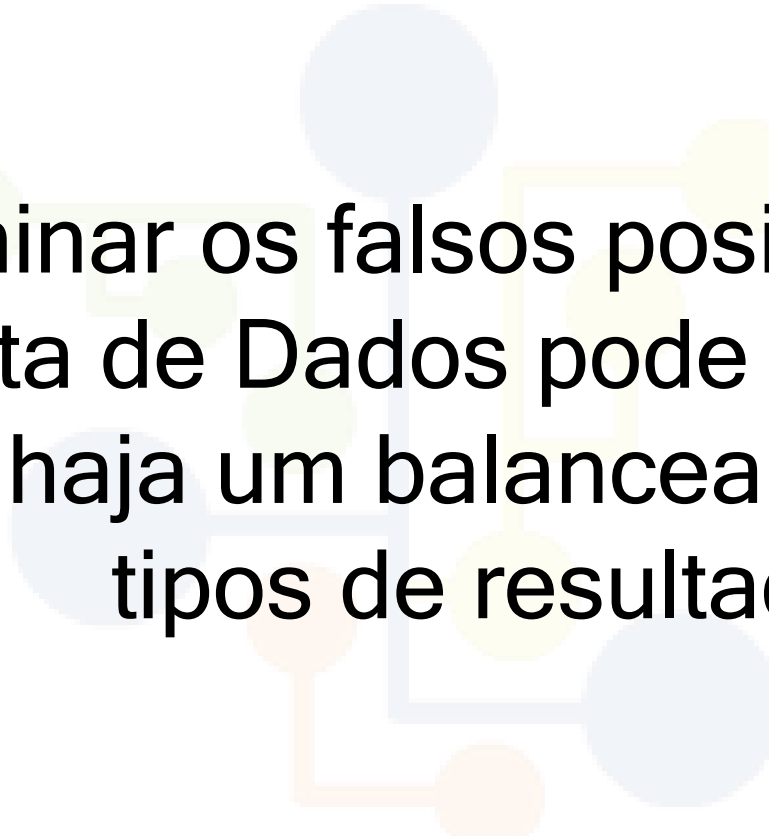
False Positive → desperdício de tempo  
False Negative → oportunidade perdida



Data Science Academy



É difícil eliminar os falsos positivos e negativos,  
mas o Cientista de Dados pode otimizar o algoritmo  
de forma que haja um balanceamento entre estes 2  
tipos de resultados



Data Science Academy

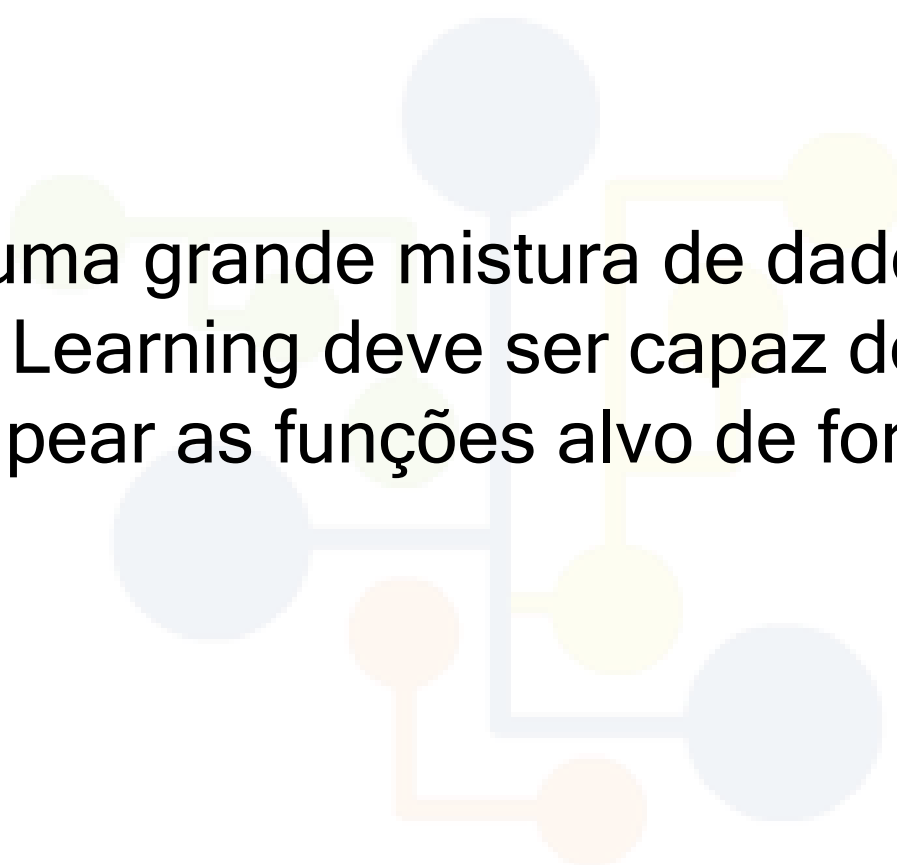

# Confusion Matrix ou matriz de erro.

		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



Big Data é uma grande mistura de dados. Um bom algoritmo de Machine Learning deve ser capaz de distinguir os sinais e mapear as funções alvo de forma eficiente.



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: minimize  $J(\theta_0, \theta_1)$   
 $\theta_0, \theta_1$



Data Science Academy

# Cost Function

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

Parameters:  $\theta_0, \theta_1$

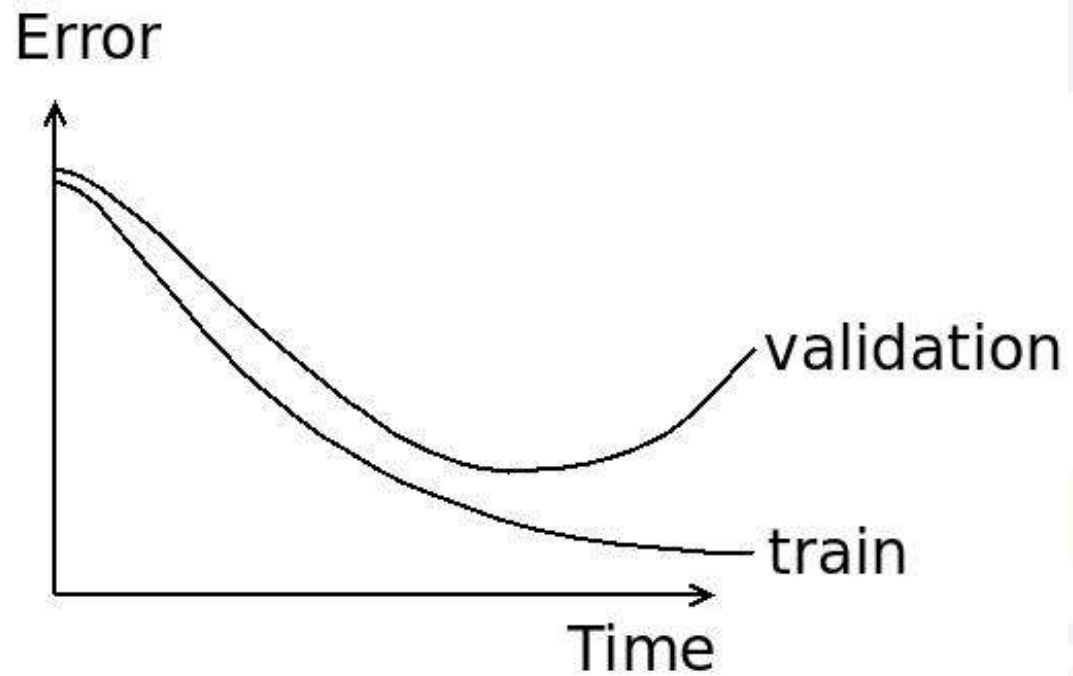
Cost Function:  $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

Goal: minimize  $J(\theta_0, \theta_1)$   
 $\theta_0, \theta_1$



Data Science Academy

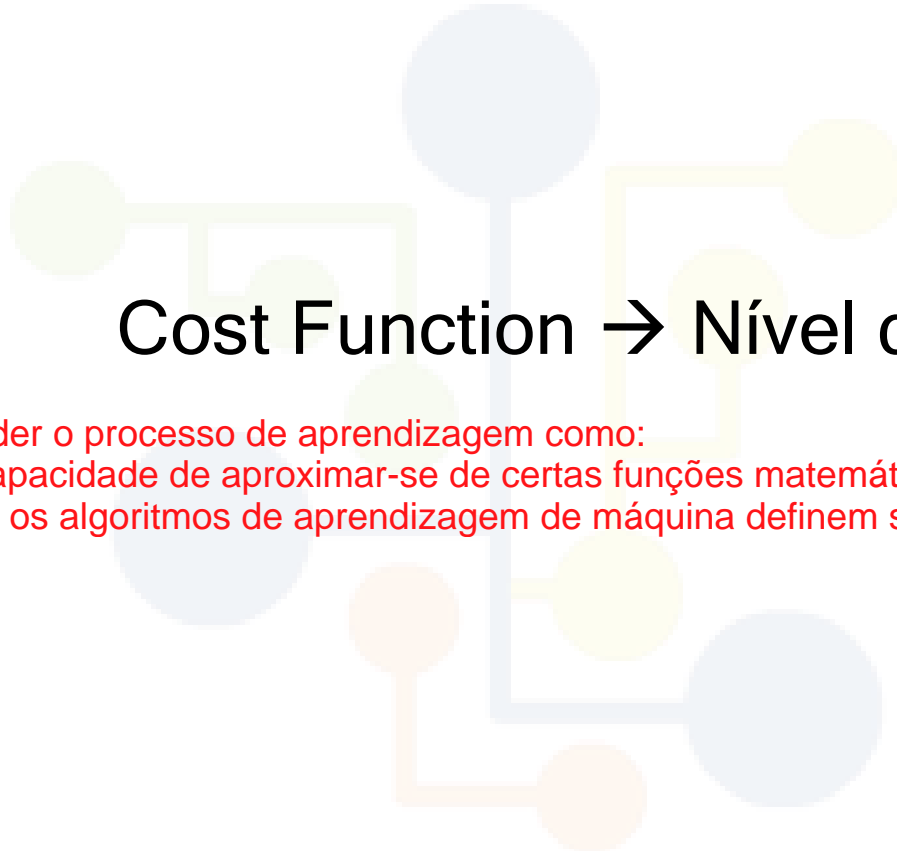





Definindo o Erro



Data Science Academy




## Cost Function → Nível de erro

Ajuda a compreender o processo de aprendizagem como:

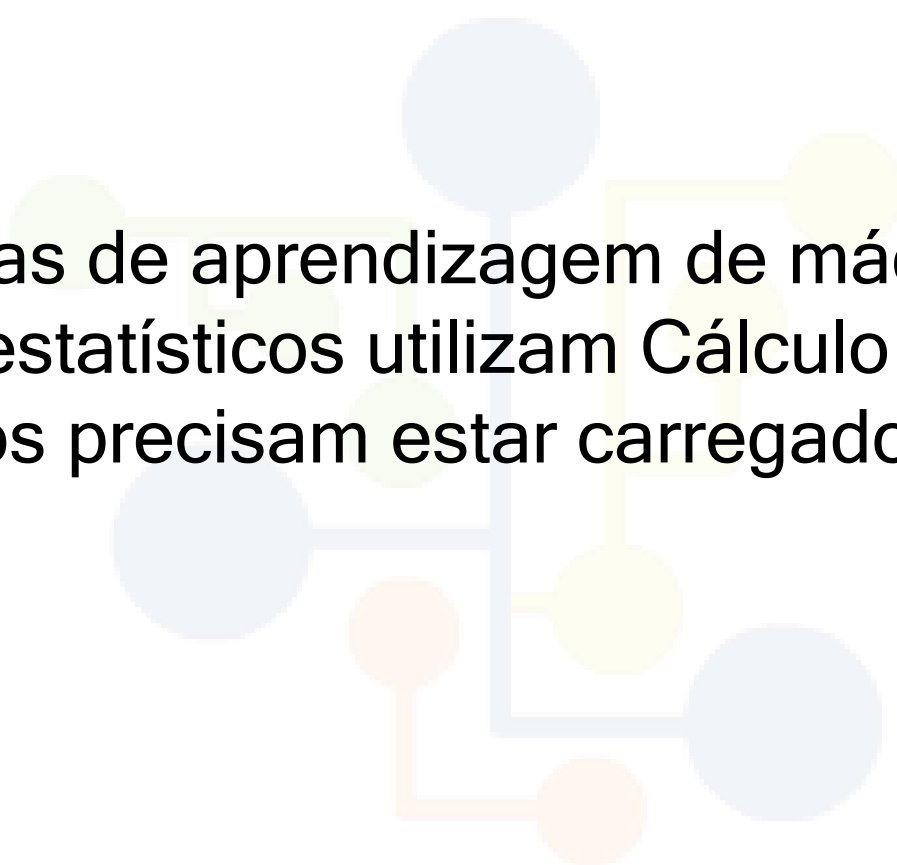
- representação: capacidade de aproximar-se de certas funções matemáticas
- otimização: como os algoritmos de aprendizagem de máquina definem seus parâmetros internos



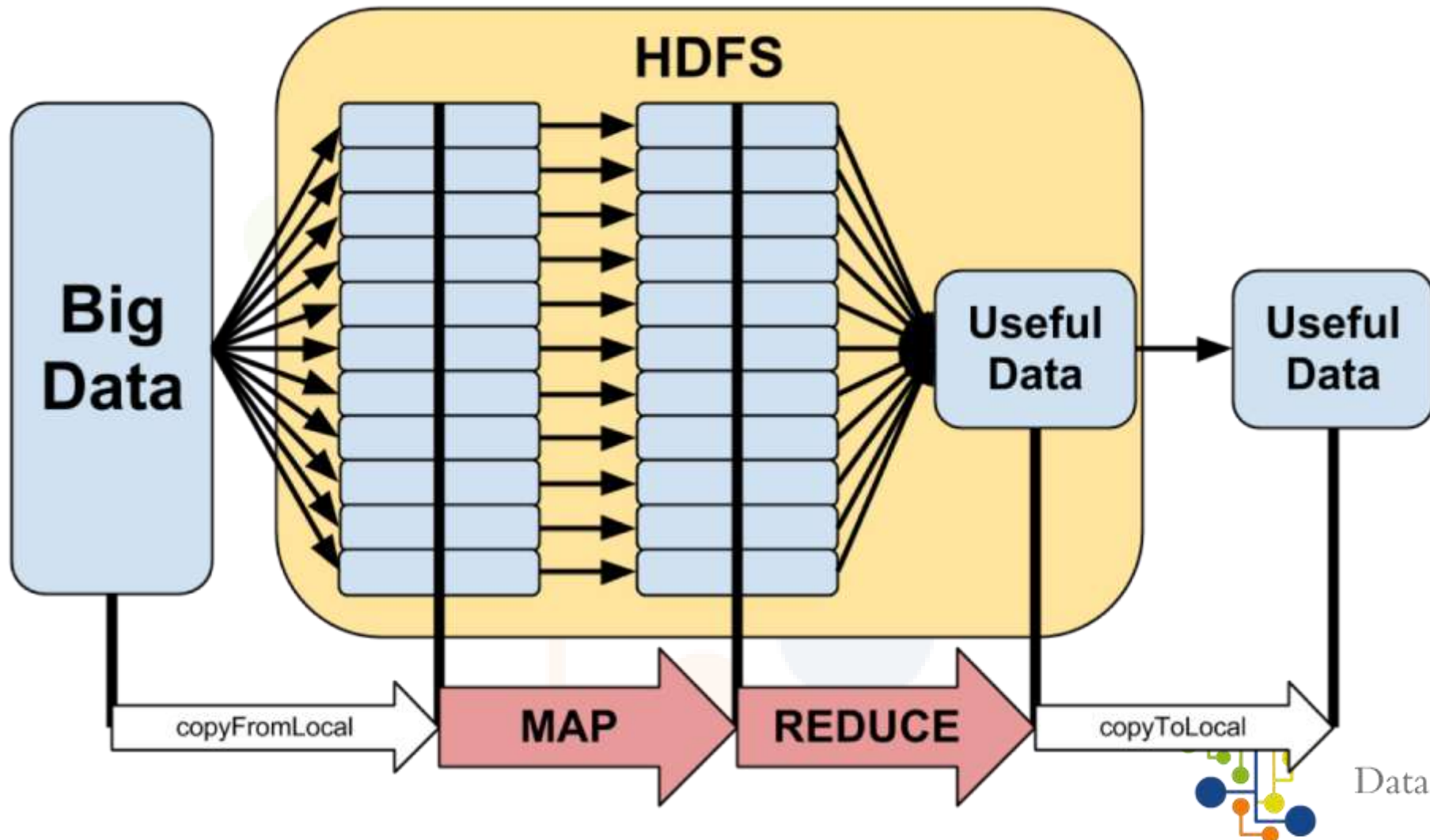
Data Science Academy



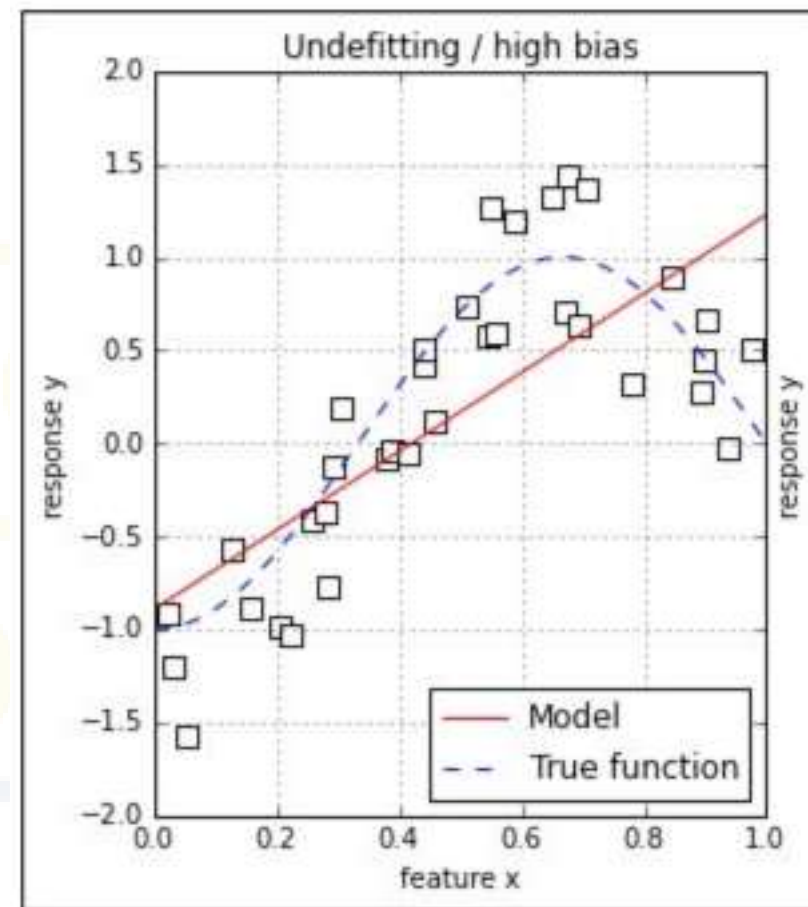
As técnicas de aprendizagem de máquina baseadas em algoritmos estatísticos utilizam Cálculo e Álgebra Linear e os dados precisam estar carregados em memória



Data Science Academy

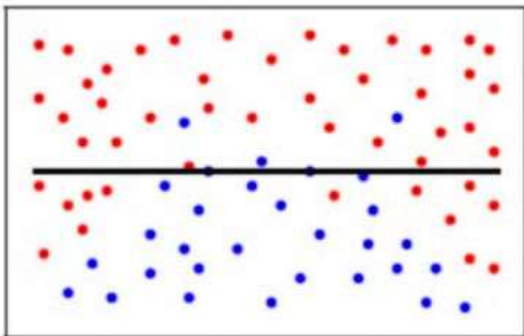


Perceba no gráfico que vai existir uma diferença entre o modelo preditivo (linha vermelha) e a função que resolve o problema (linha tracejada azul). Isso ocorre, por que o algoritmo tende a sistematicamente subestimar ou sobreestimar as regras do mundo real, que representam partes tendenciosas. Normalmente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

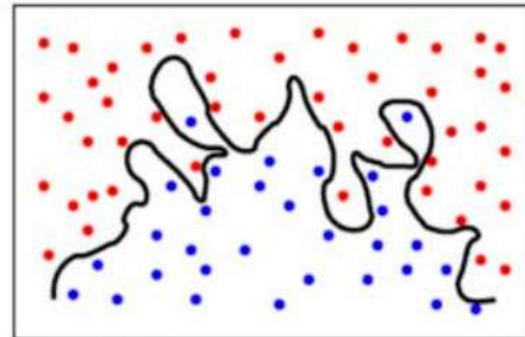


# Overfitting

Underfitting



Overfitting

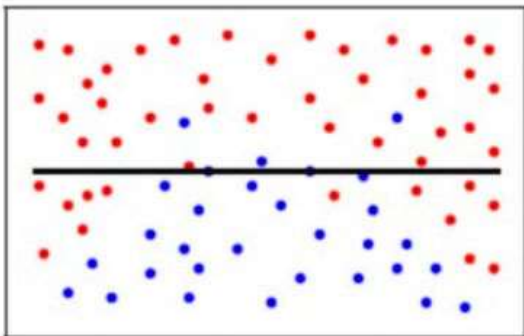


Data Science Academy

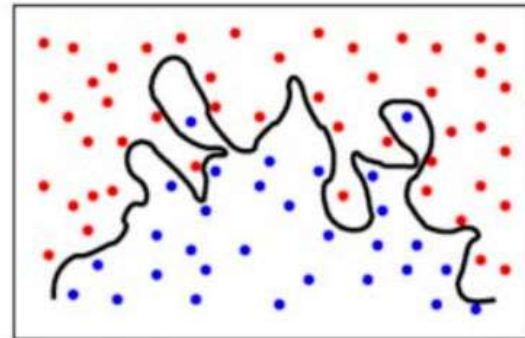


# Overfitting

Underfitting

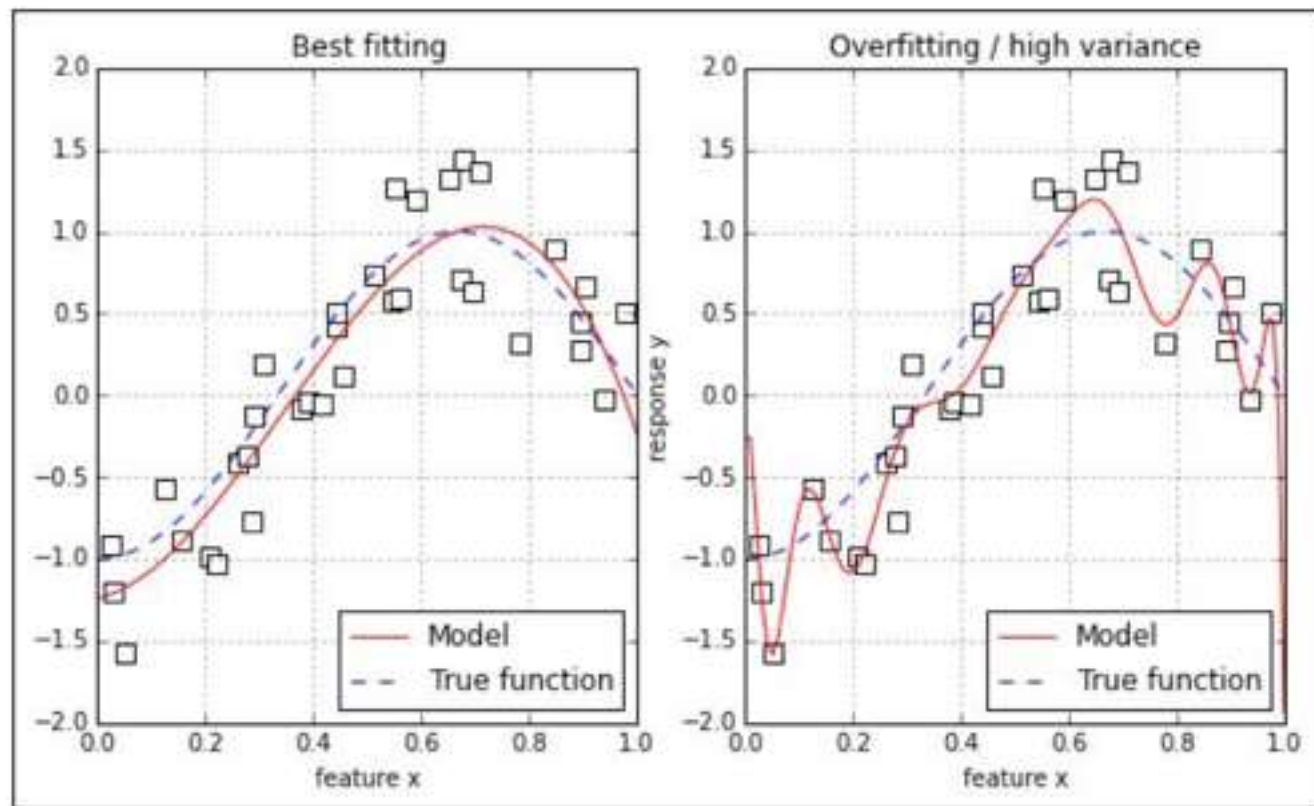


Overfitting



Data Science Academy

# Overfitting



Data Science Academy

Para atingir o equilíbrio e criar grandes soluções de Machine Learning, você terá que fazer escolhas

Simplicidade



Complexidade



Ponto Ideal



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



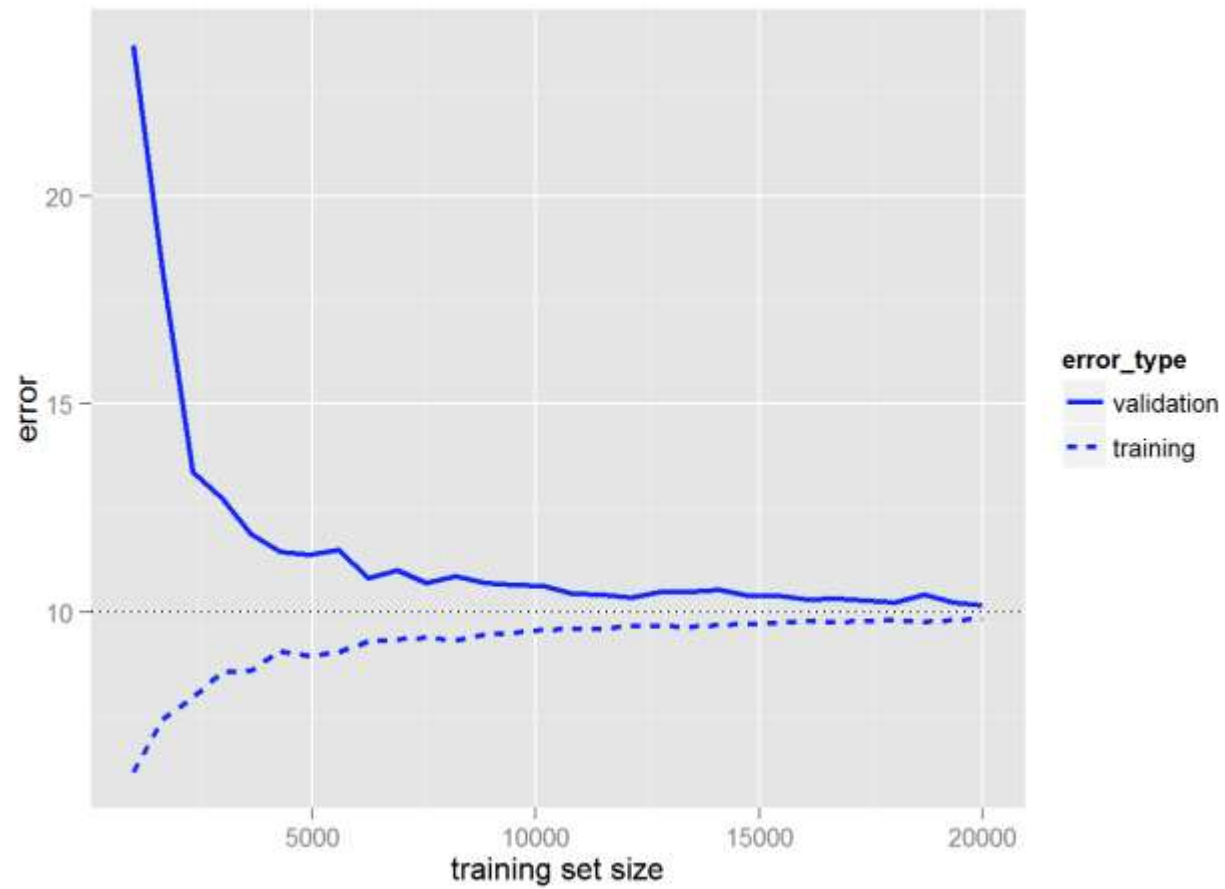
Data Science Academy

# 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.

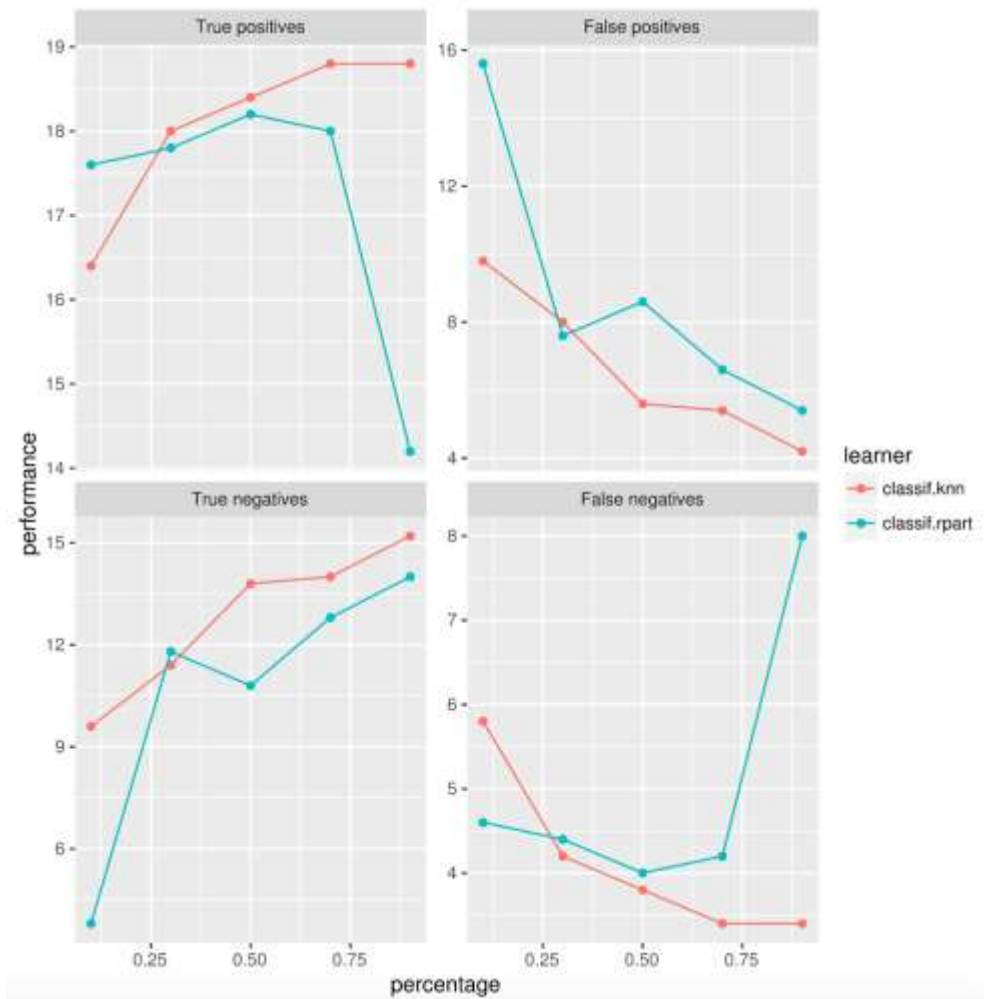


Data Science Academy



Data Science Academy






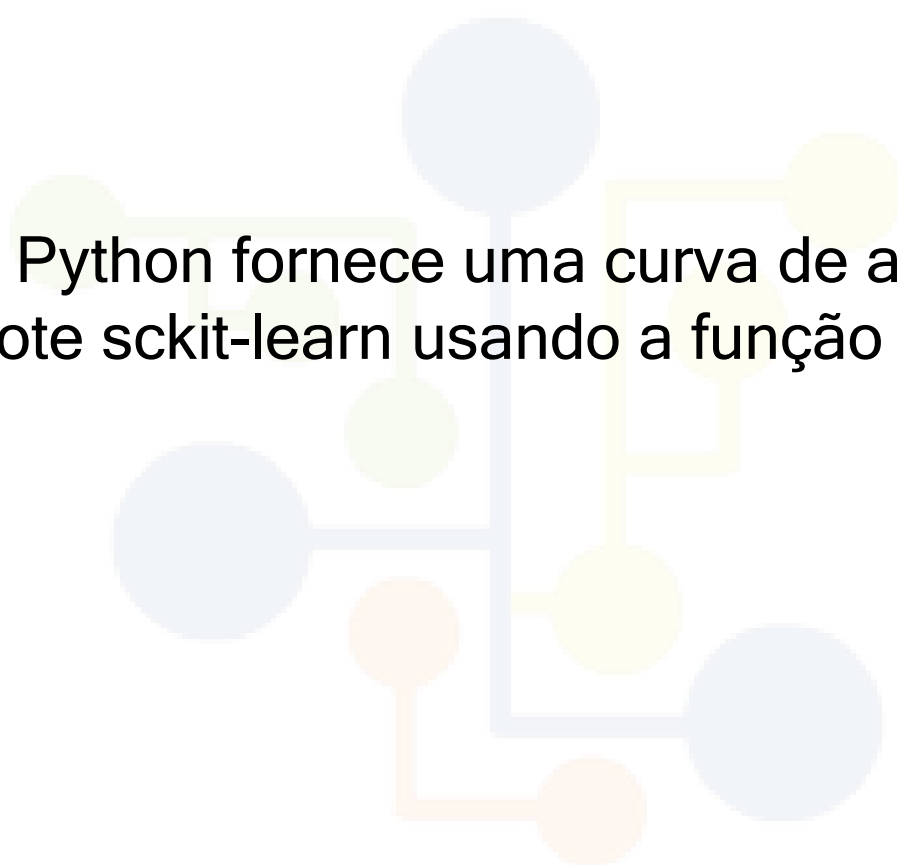
Podemos criar curvas de aprendizagem no R de diversas formas, usando os pacotes `mlr`, `caret` ou mesmo o `ggplot2`



Data Science Academy

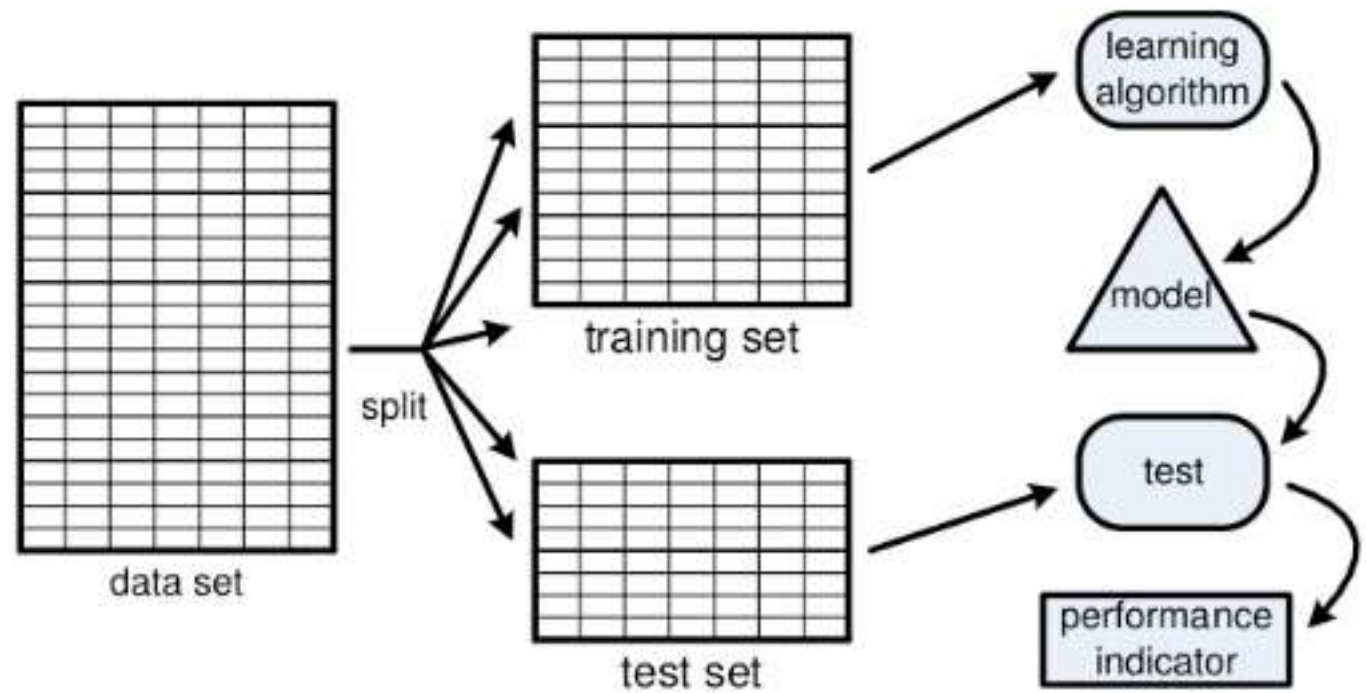


A linguagem Python fornece uma curva de aprendizagem através do pacote scikit-learn usando a função `learning_curve`.



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

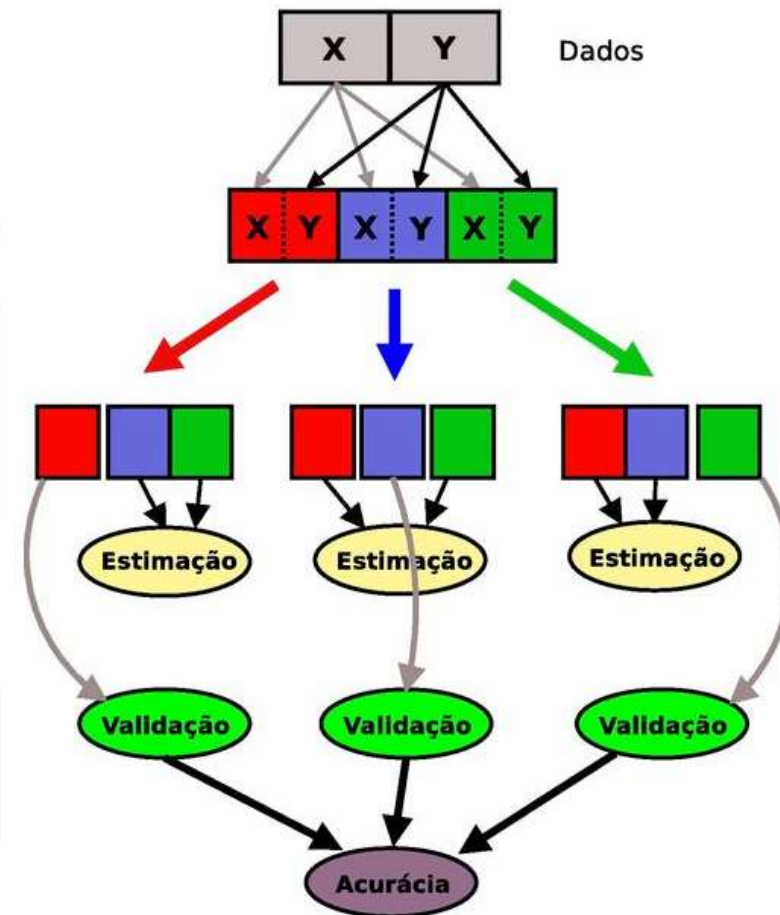
20% - dados de validação

10% - dados teste



Data Science Academy

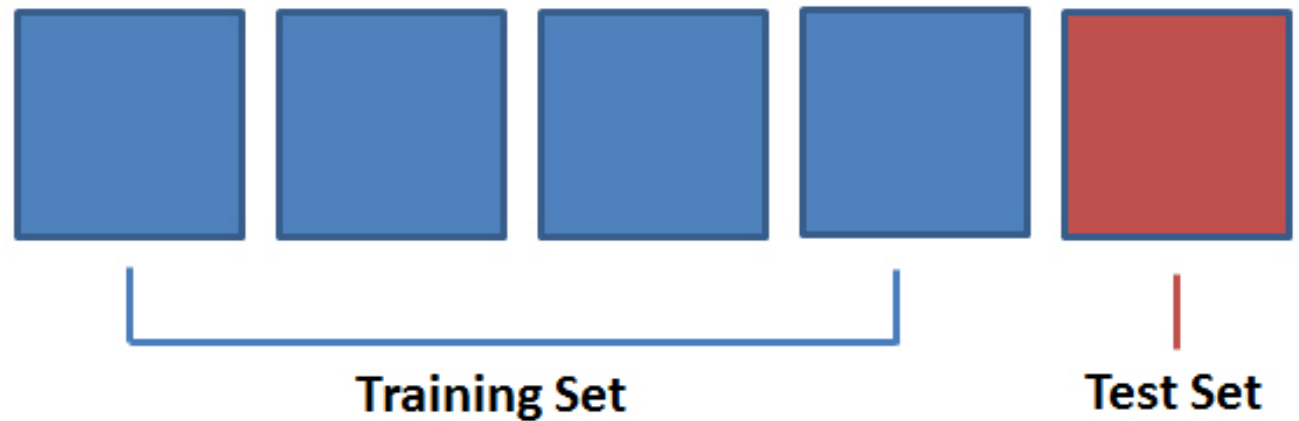
# Treinamento, Validação e Teste



Data Science Academy

# Treinamento, Validação e Teste

$n > 10.000$



Data Science Academy



# Cross-Validation



# Cross-Validation



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

## Slicing dos dados




Data Science Academy



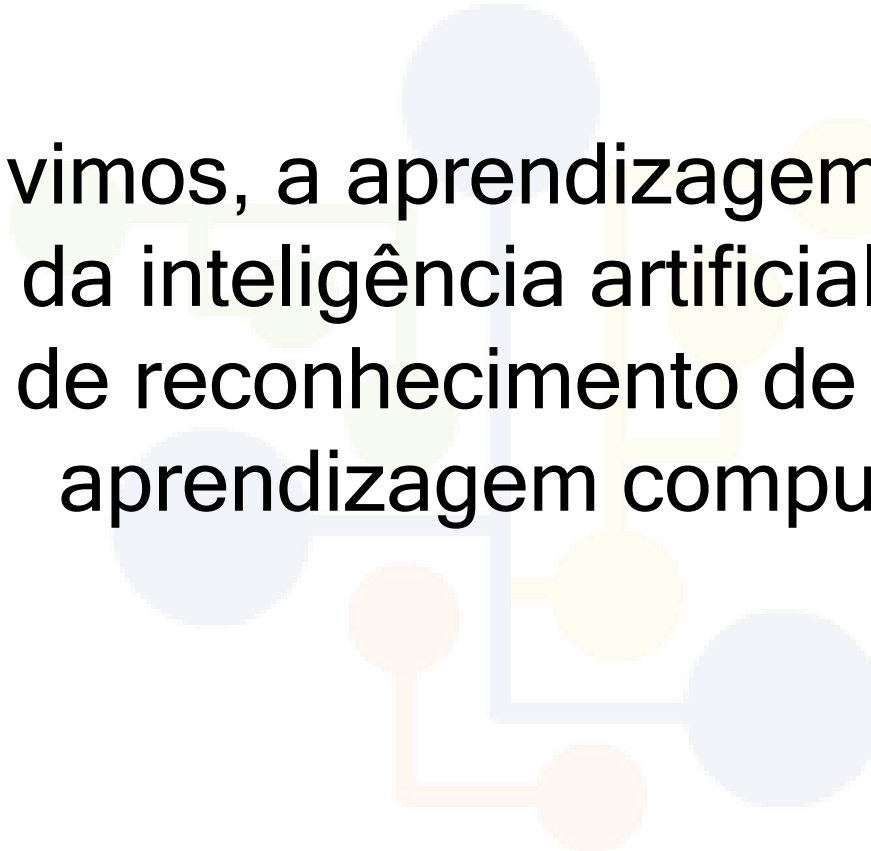
# Caret Package




Data Science Academy



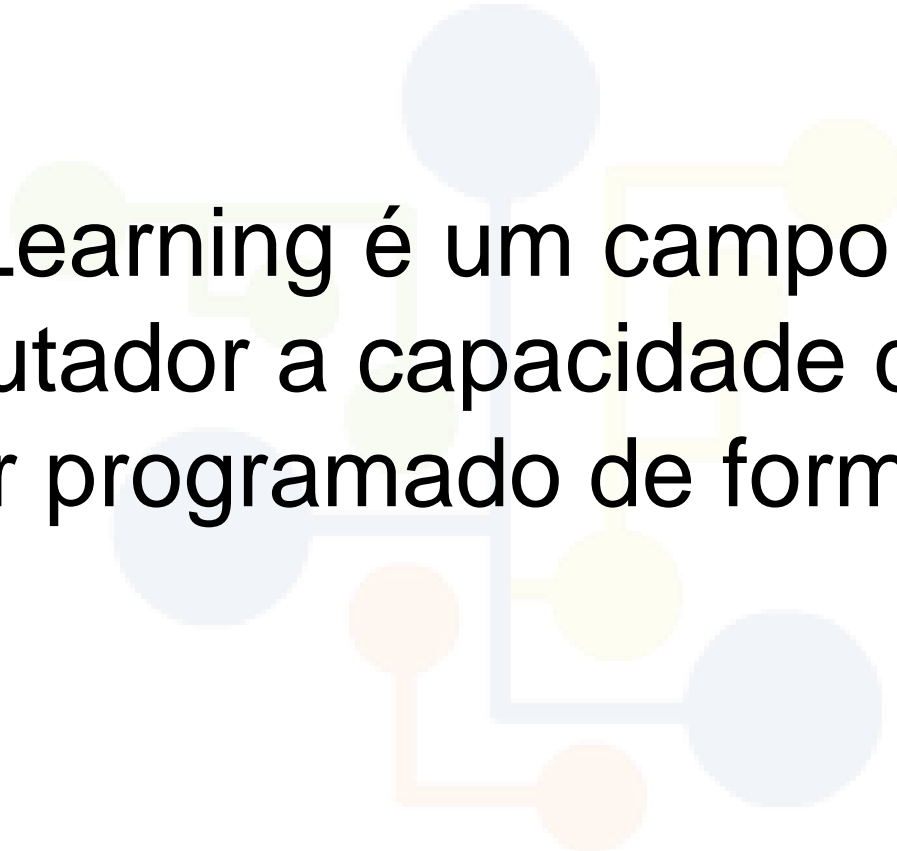
Como já vimos, a aprendizagem de máquina é um subcampo da inteligência artificial que evoluiu a partir do estudo de reconhecimento de padrões e teoria da aprendizagem computacional



Data Science Academy



Machine Learning é um campo de estudo que dá  
ao computador a capacidade de aprender, sem  
ser programado de forma explícita



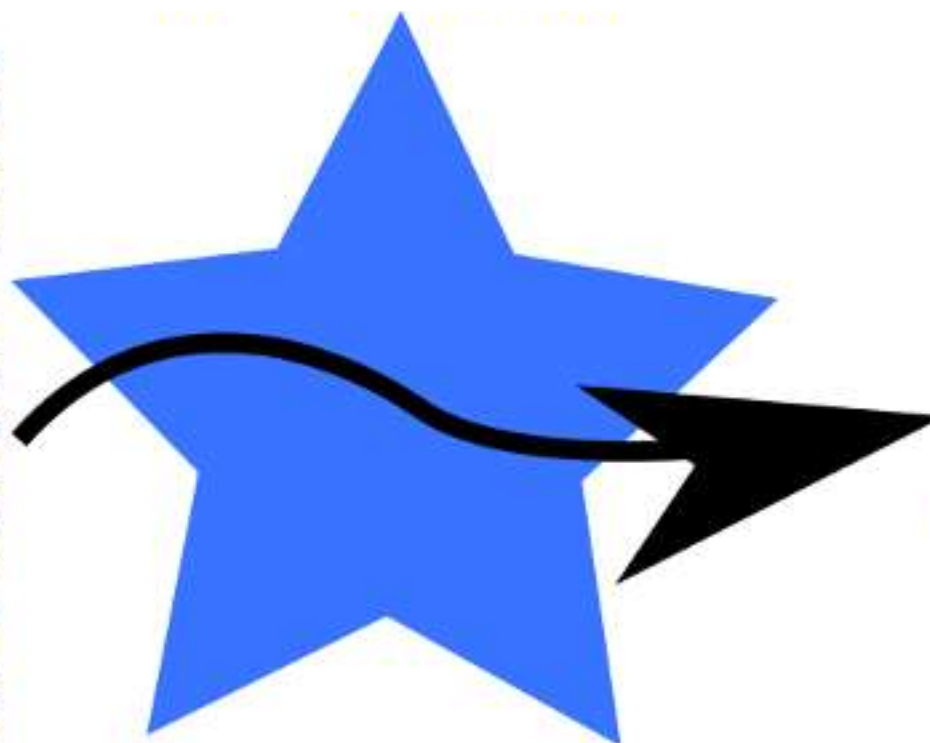
Data Science Academy



Dados

```
100100011101000000101000110111010110
100100111101110000001111100110100100
100001101101111101010011100001101001
111111010000110111001010111100001011
11001111110111111100100001110110110
010000110100110110000110000100010000
010101110011001111011001110100010111
001000010101100101000001000010011110
011101001111110010111010101010111100
100010000101100010101101010111000101
010010000100101011110011100001010000
010110000010011101010010101110110001
0110111111010111100010100010100010000
011010011011011010001000101111001101
000101000001100110001100100010010110
100101010100010011100101010101111101
```

Algoritmo



Modelo

$$f(x)$$



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

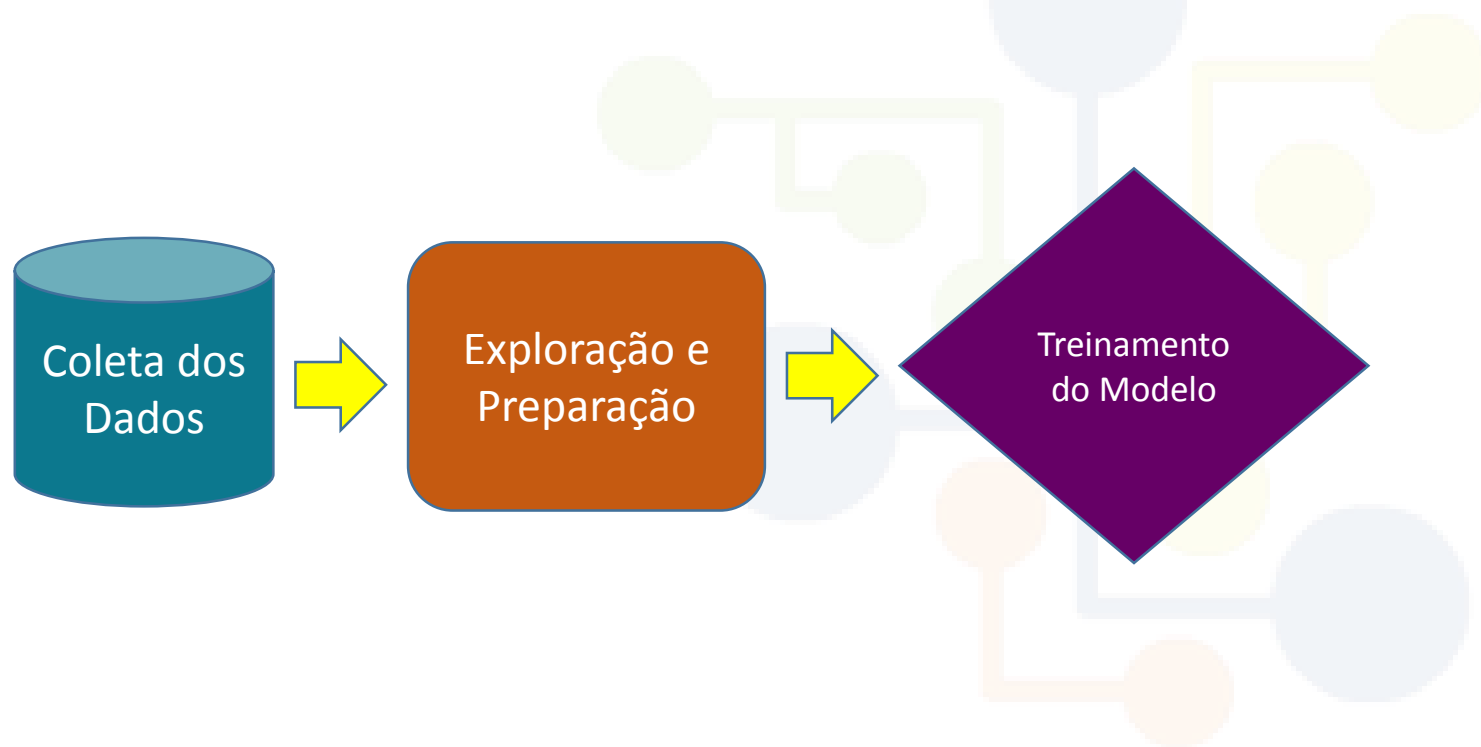
Modelo

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



Data Science Academy

# Criação do Modelo



Data Science Academy



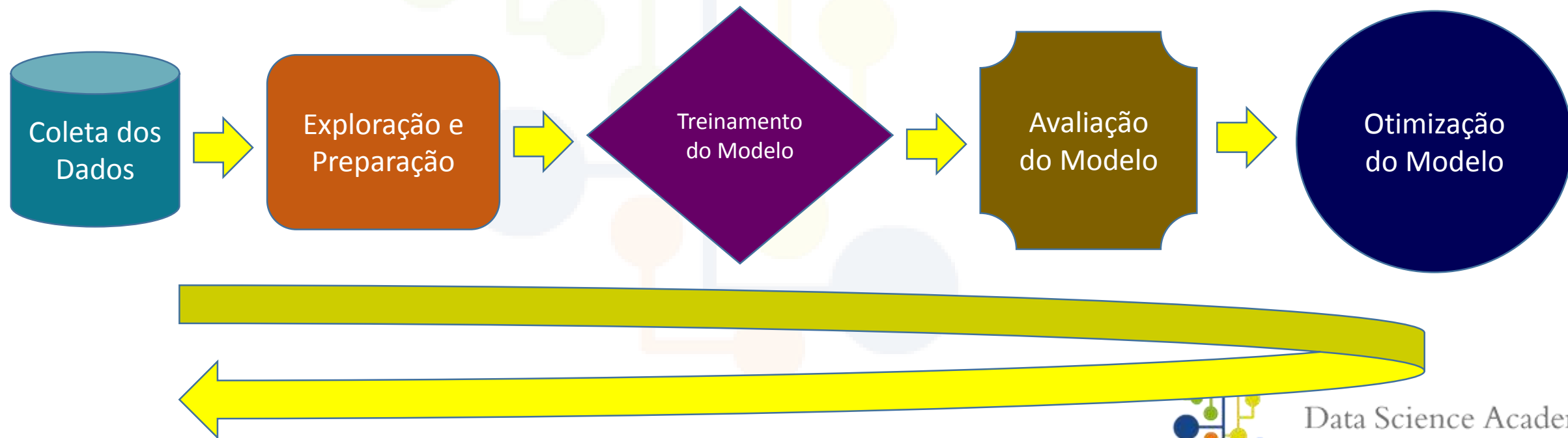
# Modelo

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



Data Science Academy

# Criação do Modelo



Data Science Academy





Este é um trabalho iterativo e assim como um surfista está sempre em busca da onda perfeita, seu trabalho como Cientista de Dados é buscar sempre o melhor modelo possível para suas previsões.



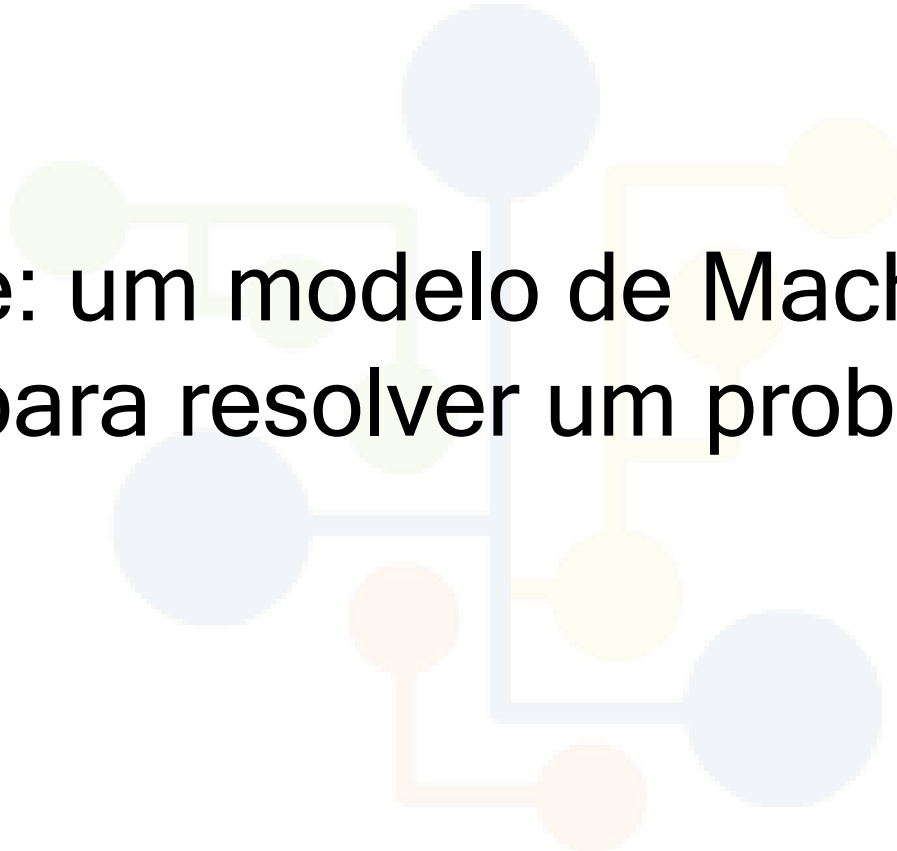

Data Science Academy



# Machine Learning na Prática




Data Science Academy



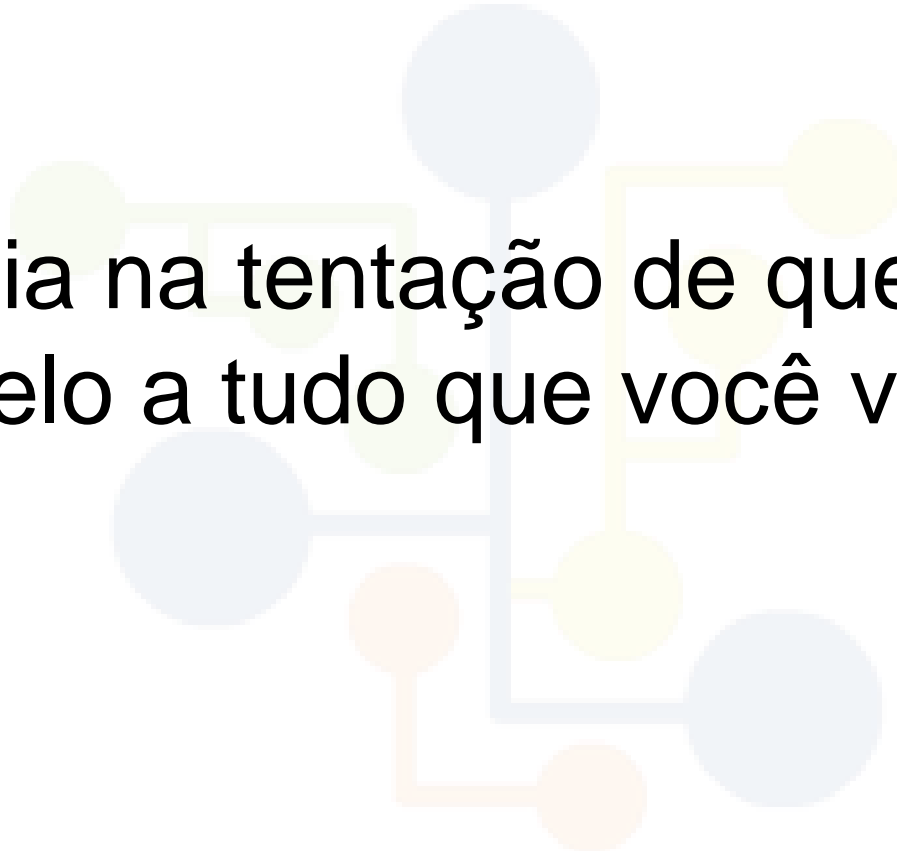
Lembre-se: um modelo de Machine Learning será usado para resolver um problema específico



Data Science Academy



Não caia na tentação de querer aplicar seu modelo a tudo que você vê pela frente



Data Science Academy



# Algoritmos de Machine Learning



Data Science Academy



## Aprendizagem Supervisionada

- Classificação
- Regressão

## Aprendizagem Não Supervisionada

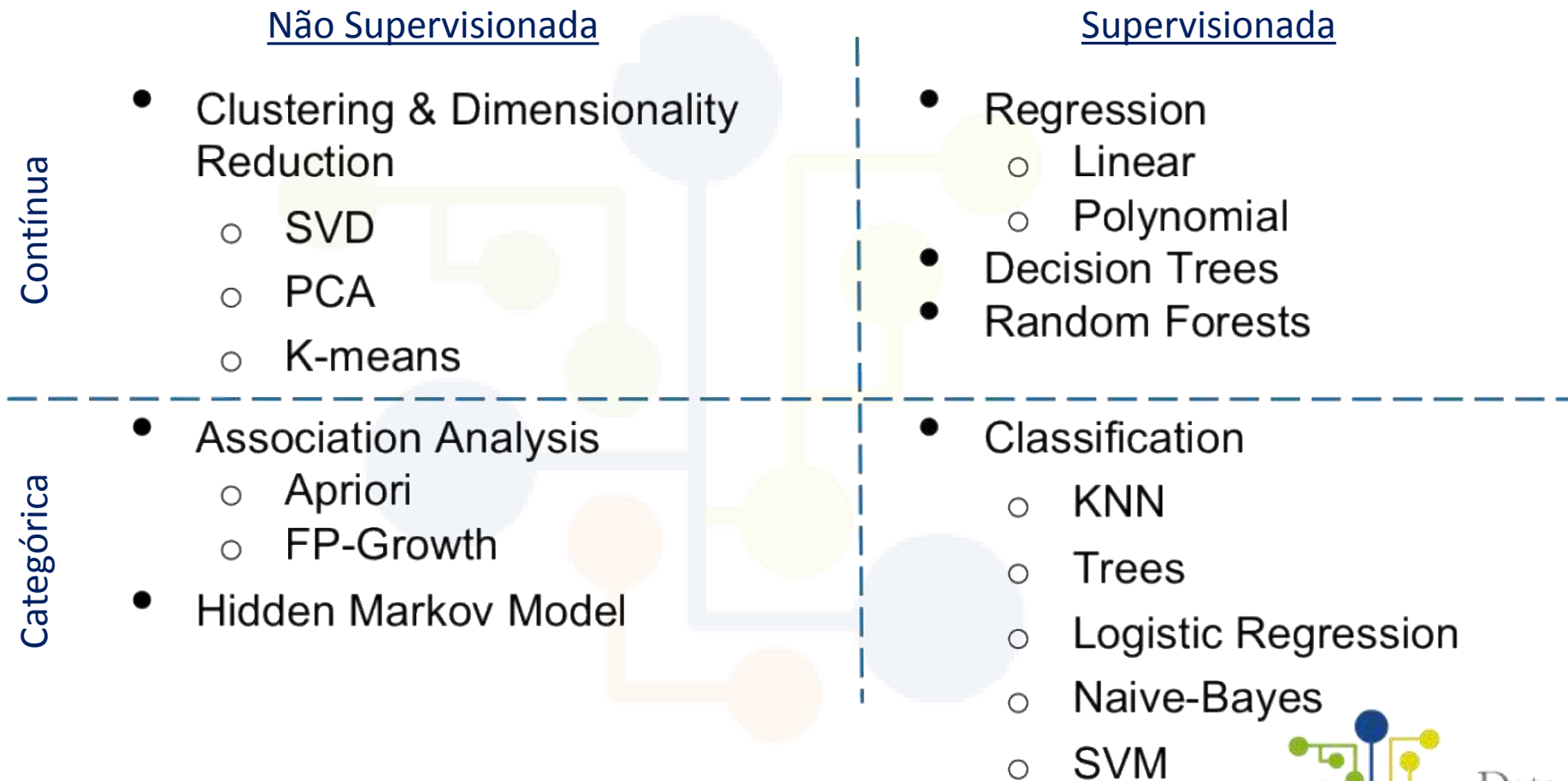

- Clustering
- Segmentação
- Redução de Dimensionalidade

## Aprendizagem por Reforço

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



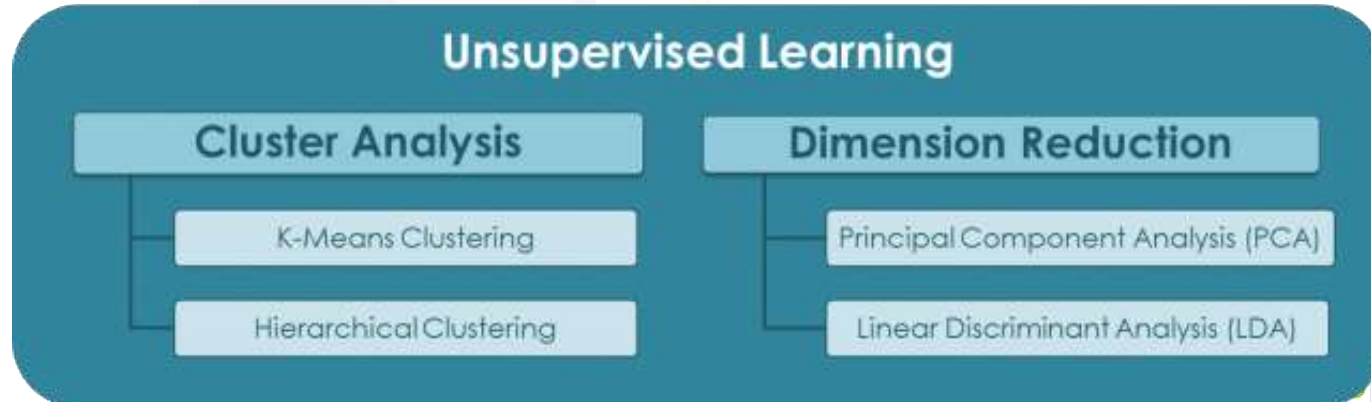
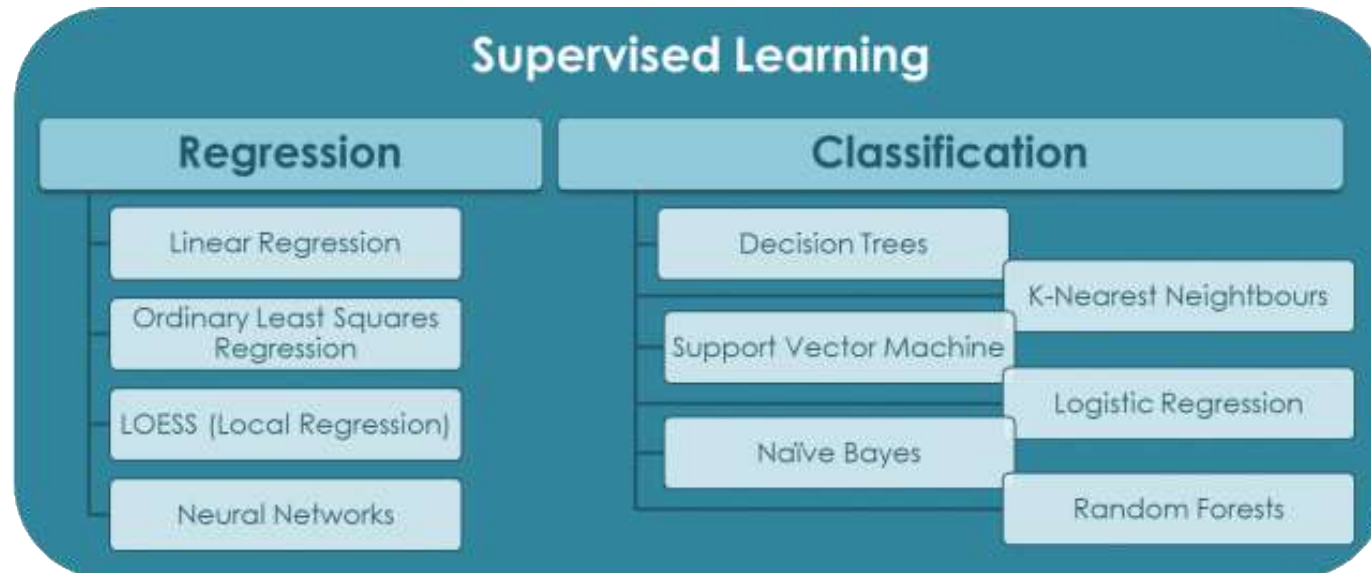
Data Science Academy



	<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



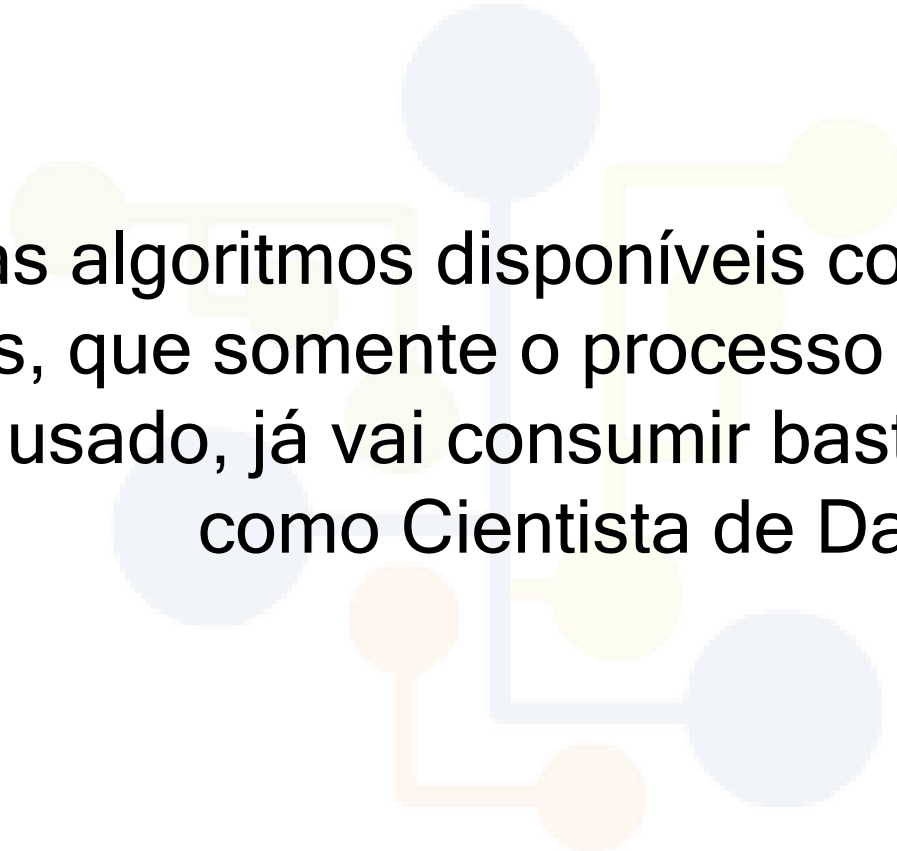




Data Science Academy



Há tantos algoritmos disponíveis com tantos métodos diferentes, que somente o processo de escolha de qual deve ser usado, já vai consumir bastante do seu tempo como Cientista de Dados




Data Science Academy



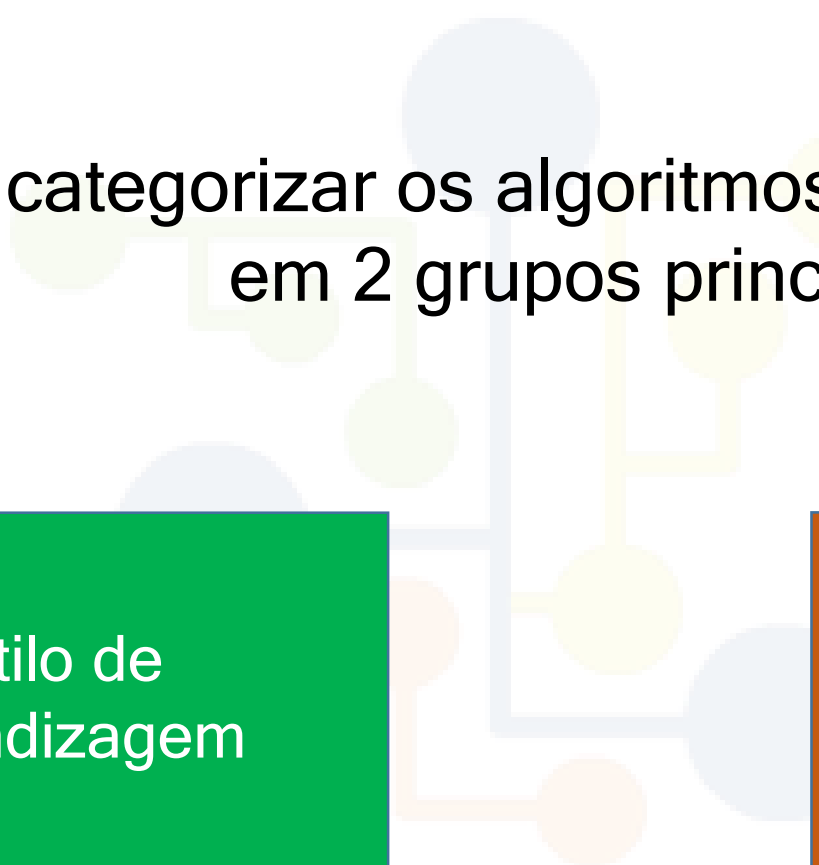
Podemos categorizar os algoritmos de Machine Learning  
em 2 grupos principais:



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



## 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

# 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



# 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

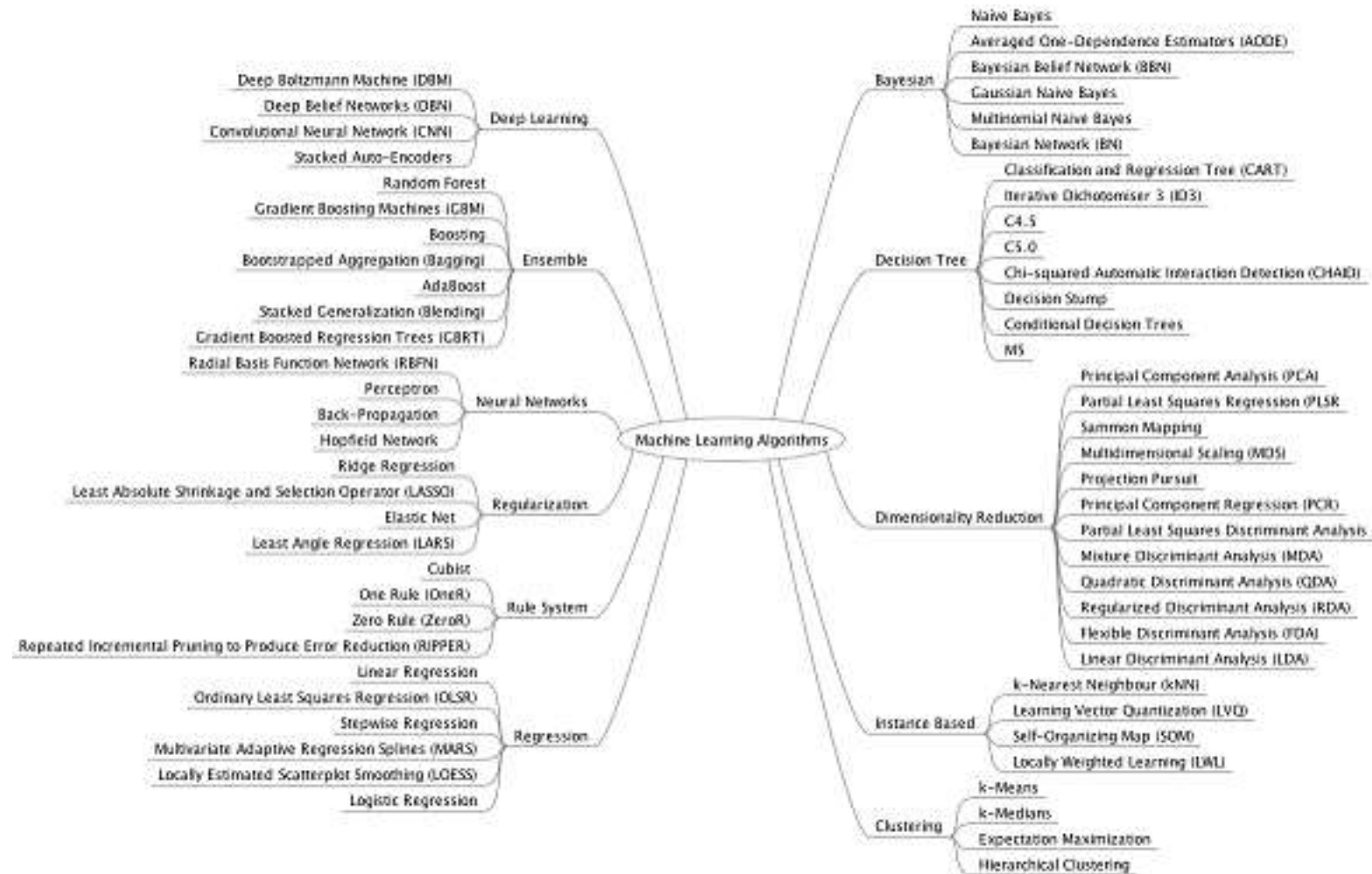
- Support Vector Machines
- Computer Vision (CV)
- Natural Language Processing (NLP)
- Recommender Systems
- Graphical Models



Data Science Academy

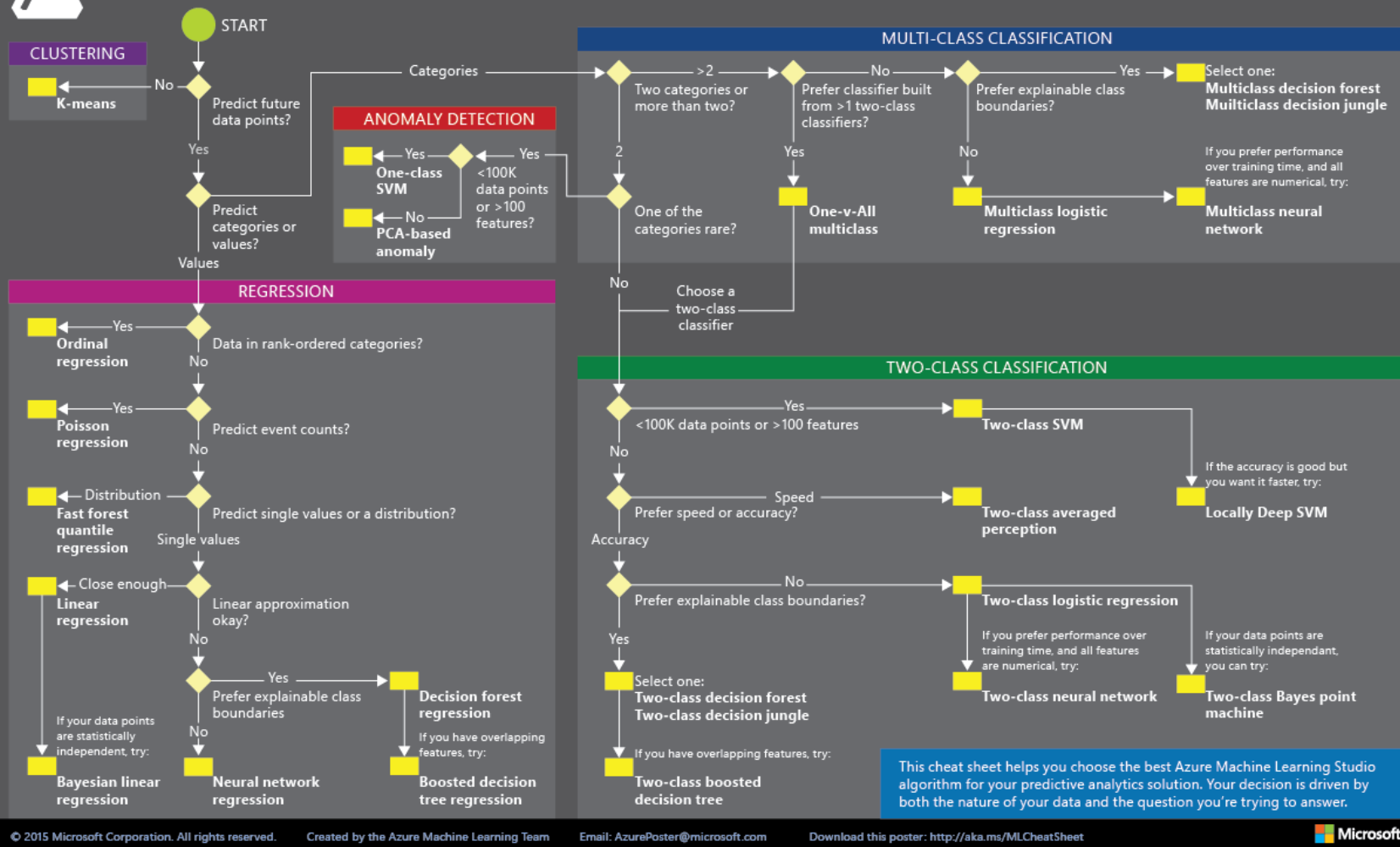


Data Science Academy

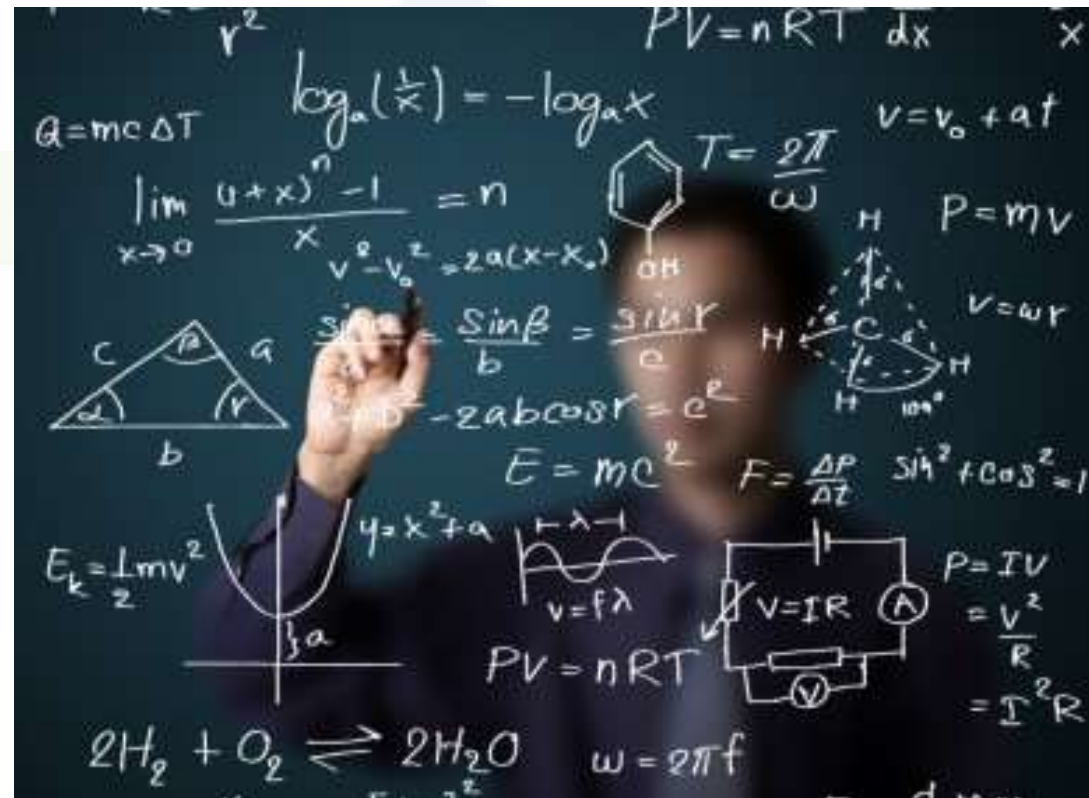




# Microsoft Azure Machine Learning: Algorithm Cheat Sheet







Data Science Academy



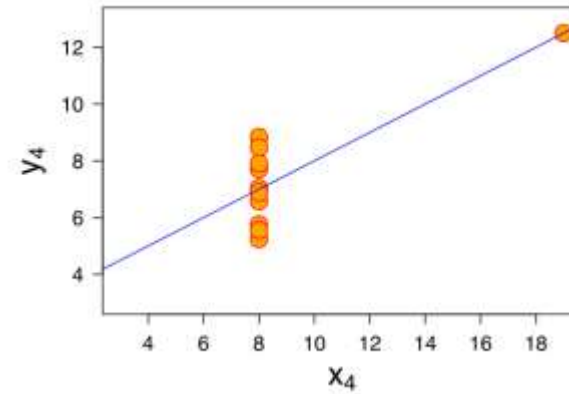
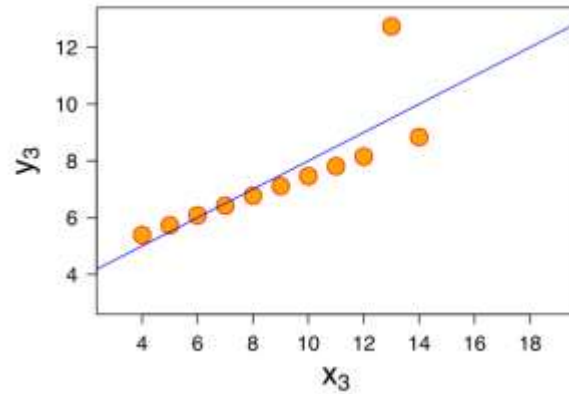
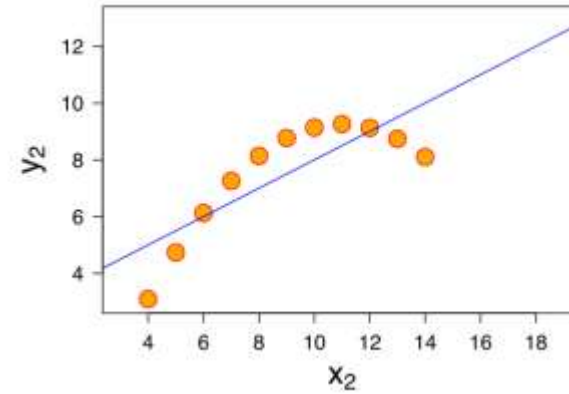
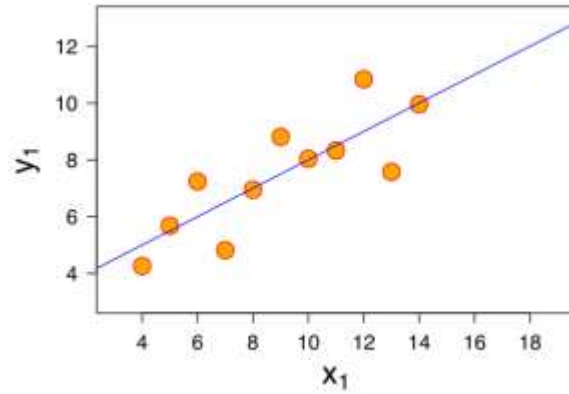
Data Science Academy



Data Science Academy



Data Science Academy



Data Science Academy



# Regressão Linear



Data Science Academy



# Machine Learning



O que a sociedade acha  
que você faz!



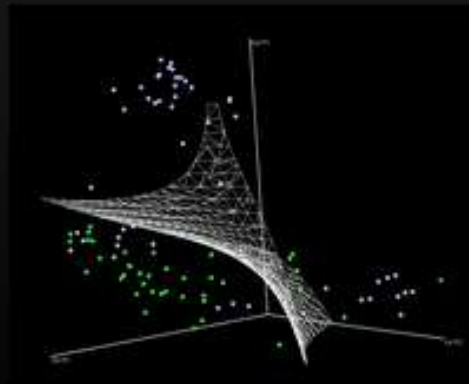
O que seus amigos  
acham que você faz!



O que seus pais acham  
que você faz!

$$\begin{aligned} L_p &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_i \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_i \alpha_i \\ \alpha_i &\geq 0, \forall i \\ \mathbf{w} &= \sum_i \alpha_i y_i \mathbf{x}_i, \sum_i \alpha_i y_i = 0 \\ \nabla \hat{g}(\theta_t) &= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t), \\ \theta_{t+1} &= \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t) \\ \mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] &= \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t), \end{aligned}$$

O que outros  
profissionais acham  
que você faz!



O que você acha que  
você faz!

```
➤ dados_treino <- subset(df)
➤ dados_teste <- subset(df)
➤ modelo <- lm(varY ~ varX, treino)
```

O que você realmente  
faz!





# 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



# Regressão Linear



Data Science Academy



# Como a Regressão pode ser usada?



- Investigação Científica
- Relações Causais
- Identificação de Padrões



Data Science Academy

# Compreendendo a Regressão

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

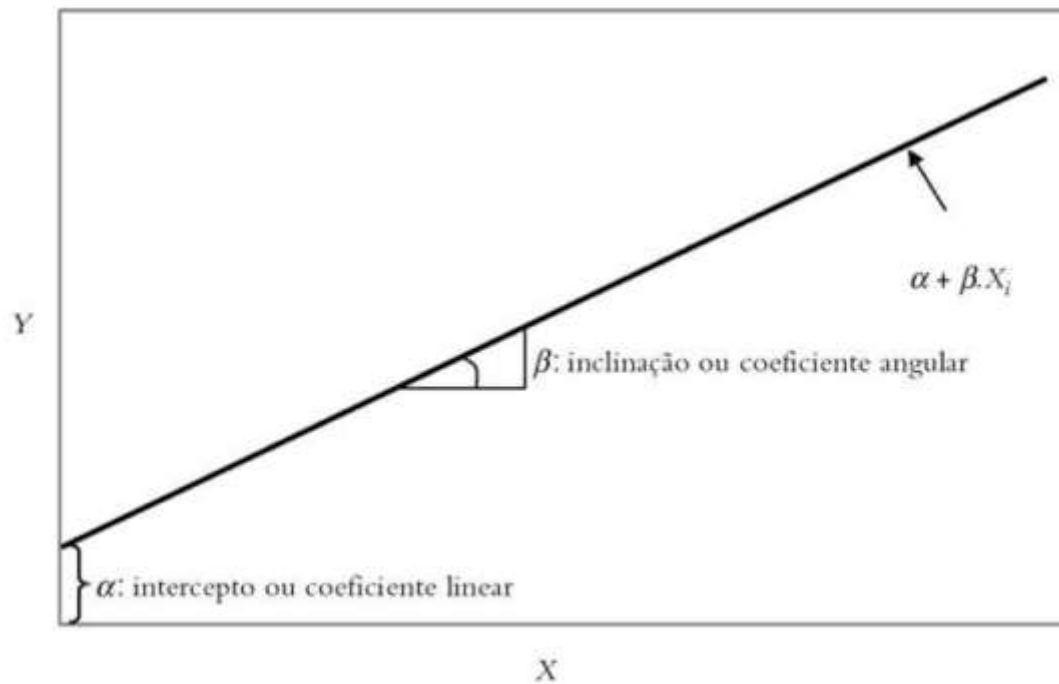
Onde:

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

$x$  = variável independente

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

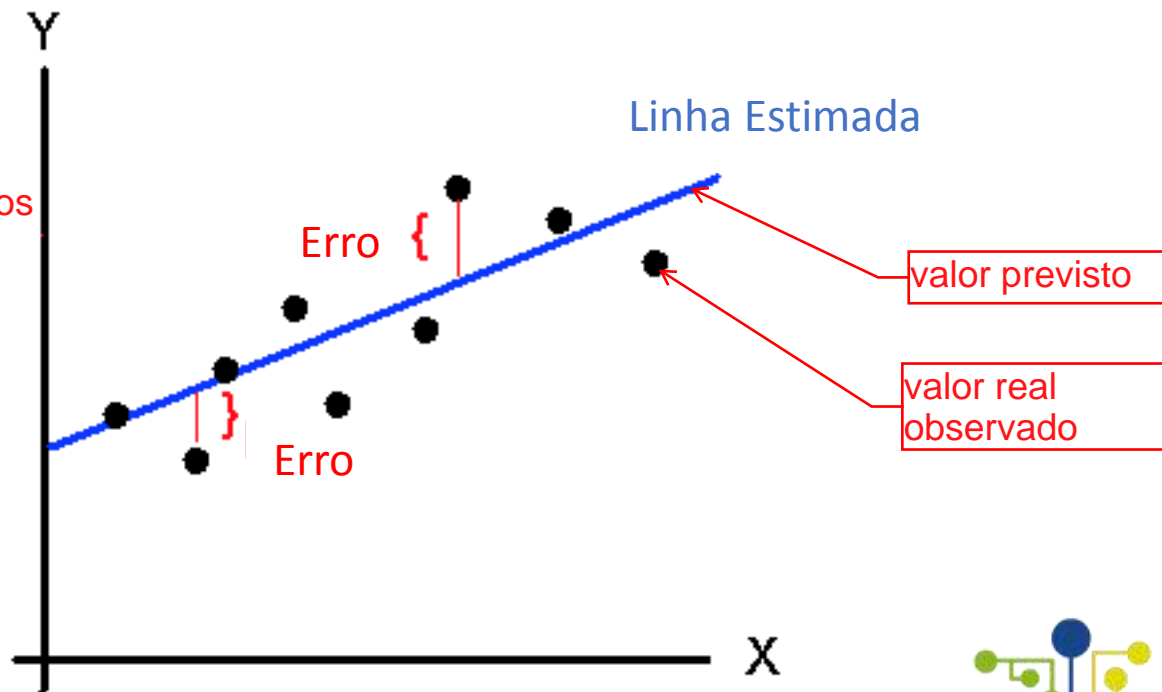
$b$  = inclinação da linha reta




Data Science Academy

# Estimativa dos Mínimos Quadrados

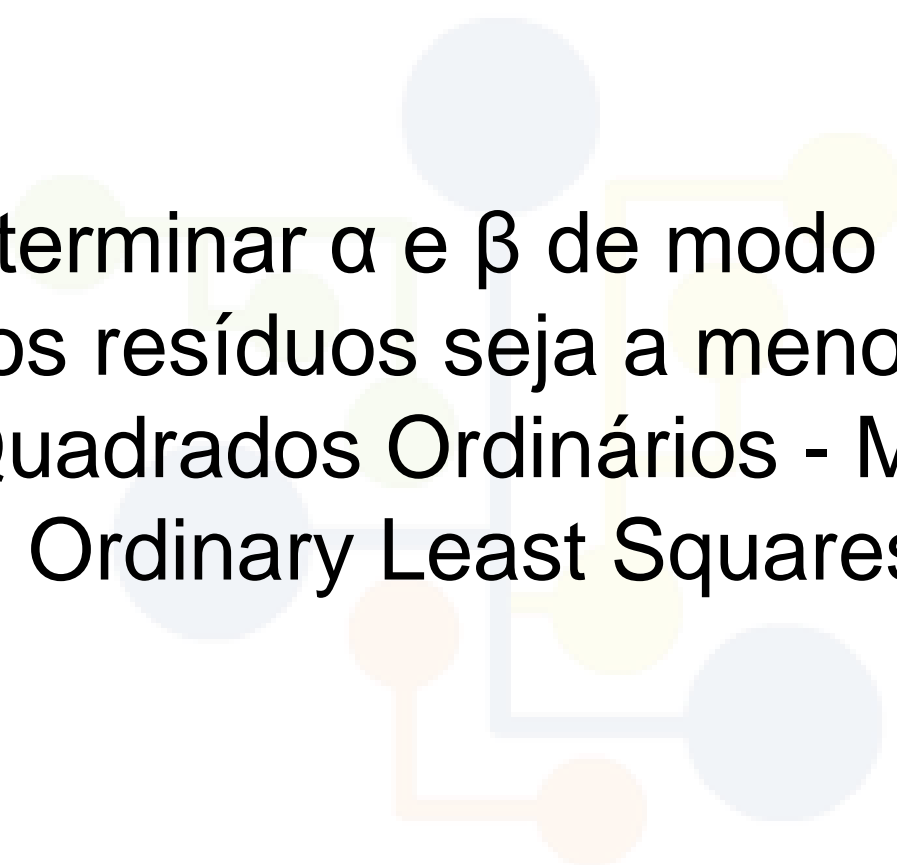
Fornece os valores de  $a$  e  $B$  da fórmula anterior, que minimizam a soma dos quadrados dos resíduos, ou seja, que minimizam a distância entre os valores observados e os valores estimados pelo modelo, indicados pela reta.



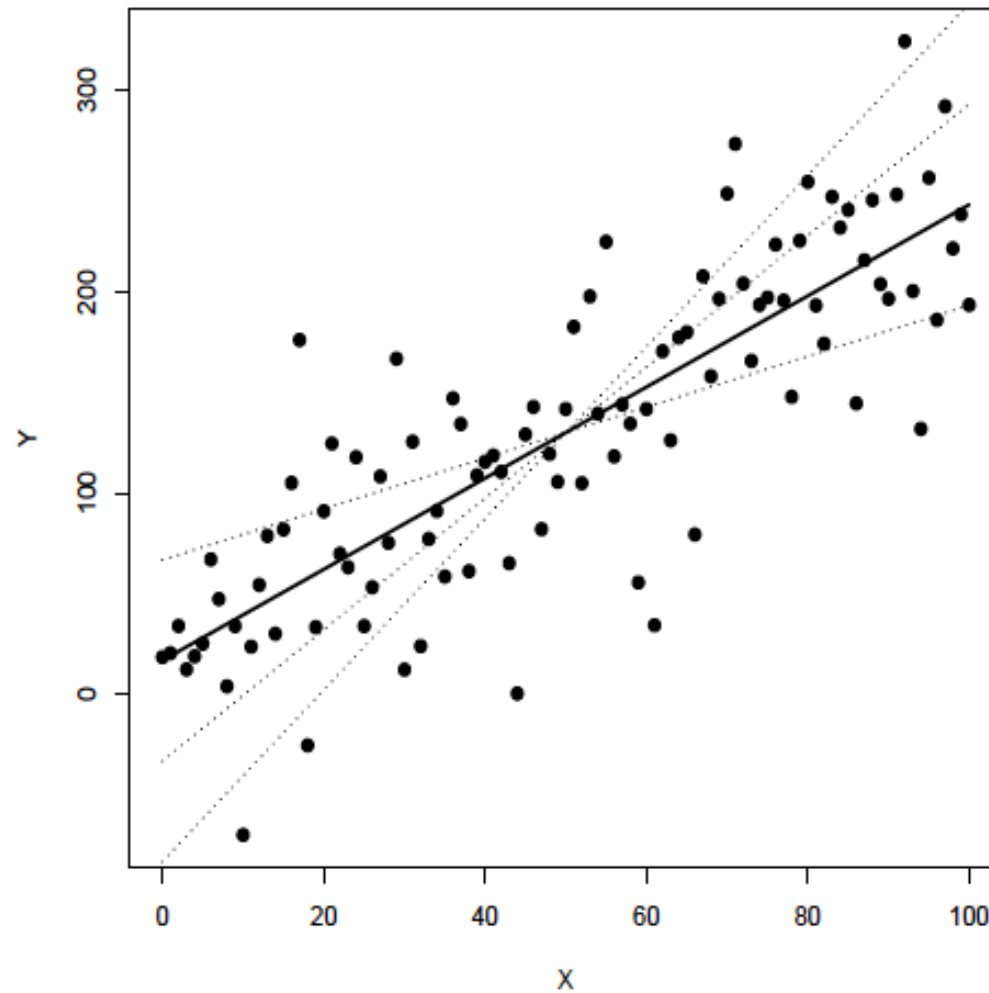
Data Science Academy



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

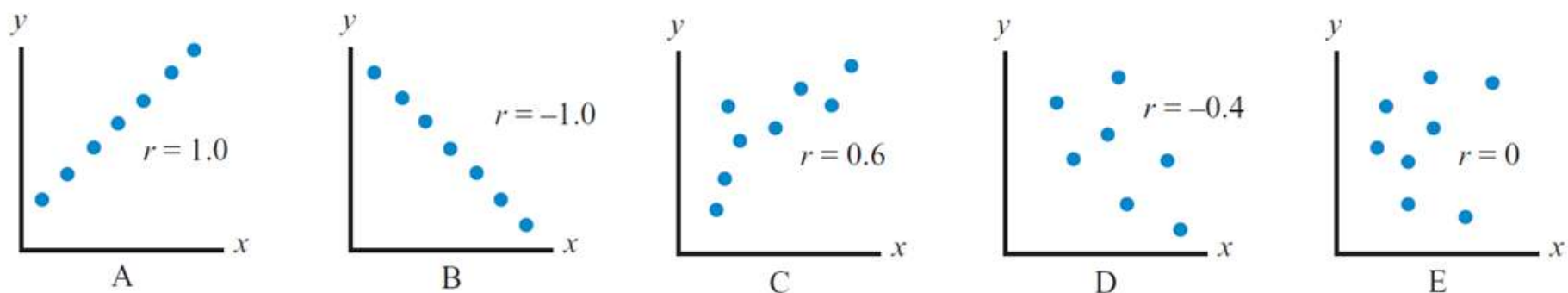


Data Science Academy



A correlação entre duas variáveis indica o quão perto sua relação segue uma linha reta.

# Correlação



**Gráfico A ( $r = 1.0$ ):**

correlação positiva perfeita entre  $x$  e  $y$

**Gráfico B ( $r = -1.0$ ):**

correlação negativa perfeita entre  $x$  e  $y$

**Gráfico C ( $r = 0.6$ ):**

relação positiva moderada:  $y$  tende a aumentar se  $x$  aumenta, mas não necessariamente na mesma taxa observada no Gráfico A

**Gráfico D ( $r = -0.4$ ):**

relação negativa fraca: o coeficiente de correlação é próximo de zero ou negativo:  $y$  tende a diminuir se  $x$  aumenta

**Gráfico E ( $r = 0$ ):**

Sem relação entre  $x$  e  $y$



Data Science Academy



# Correlação

Os valores de  $r$  variam entre **-1.0** (uma forte relação negativa) até **+1.0**, uma forte relação positiva.



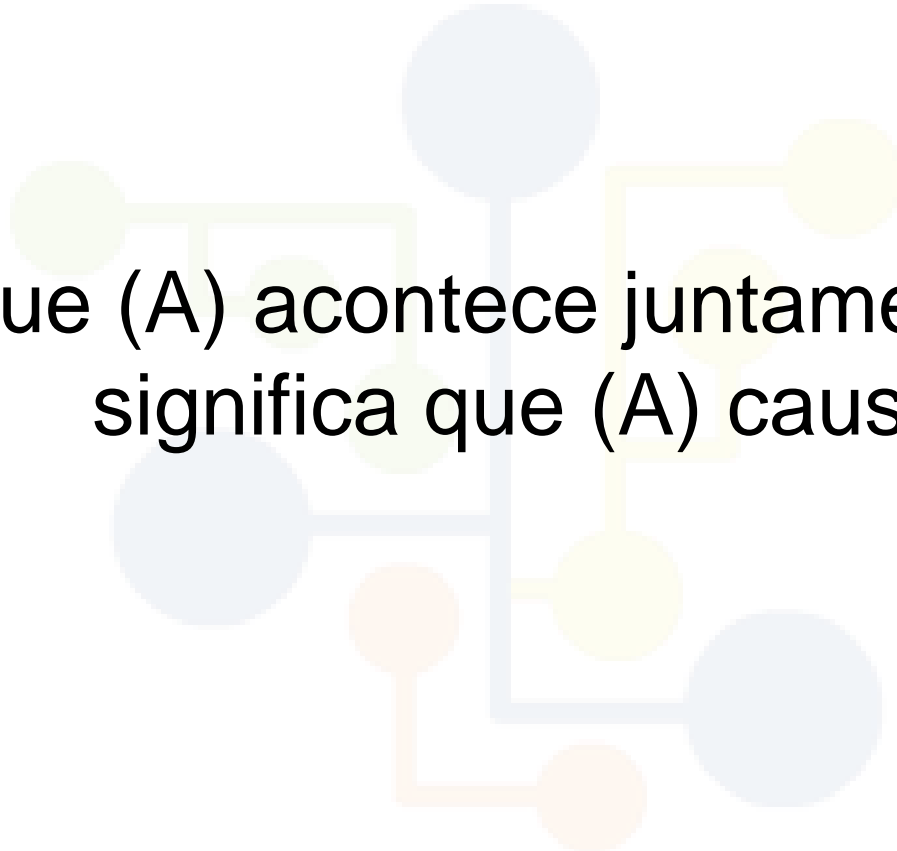

Data Science Academy



# Correlação Não Implica Causalidade




Data Science Academy



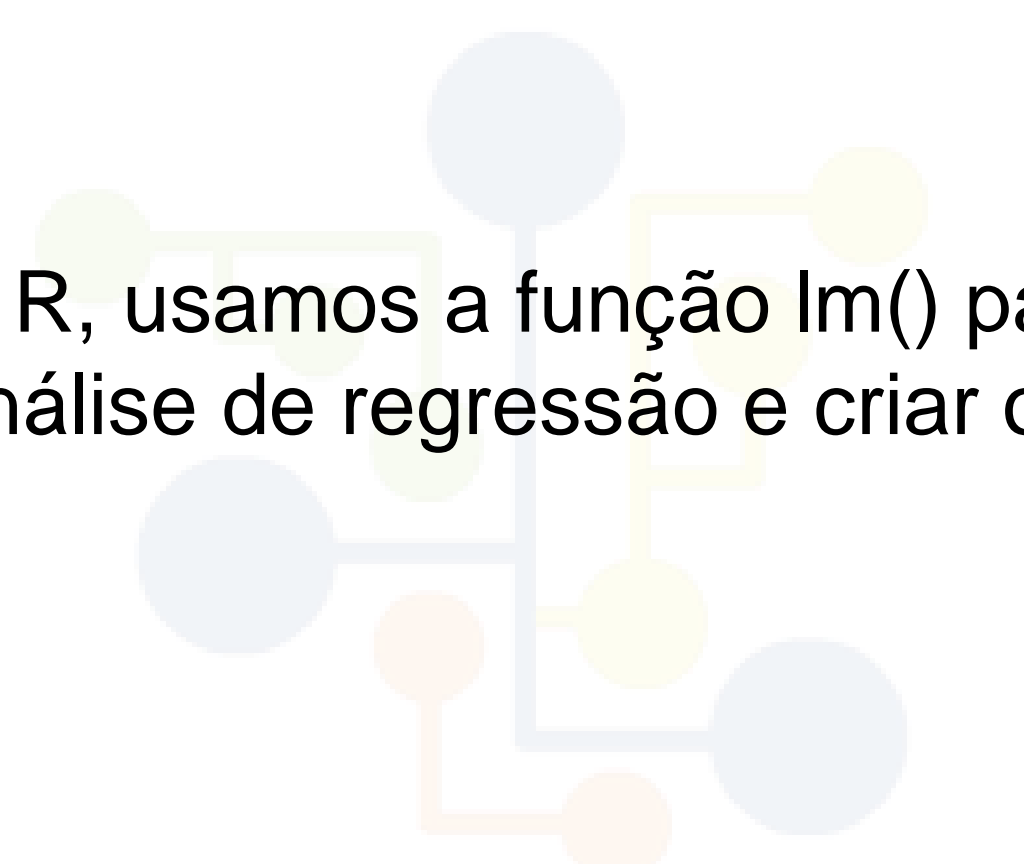
Só porque (A) acontece juntamente com (B) não significa que (A) causa (B).



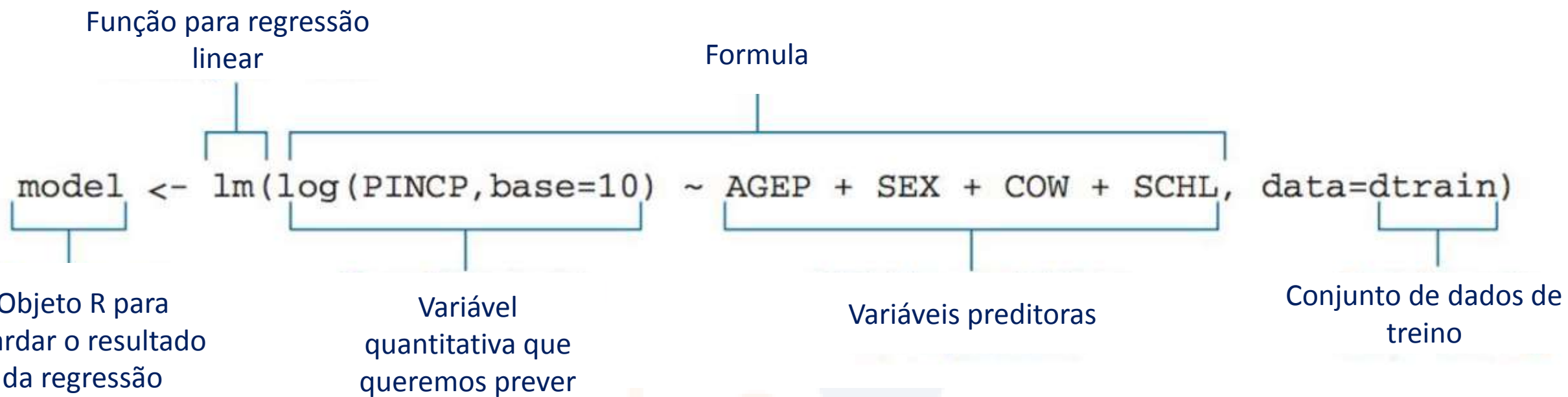
Data Science Academy

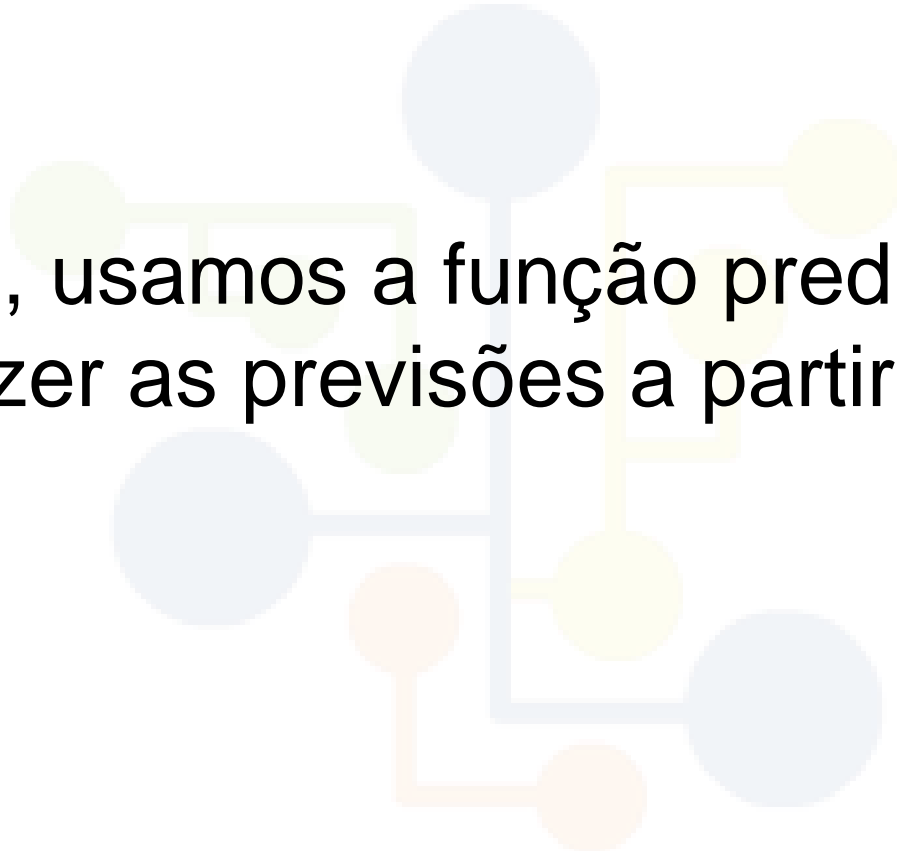



Em R, usamos a função `lm()` para fazer a análise de regressão e criar o modelo



Data Science Academy



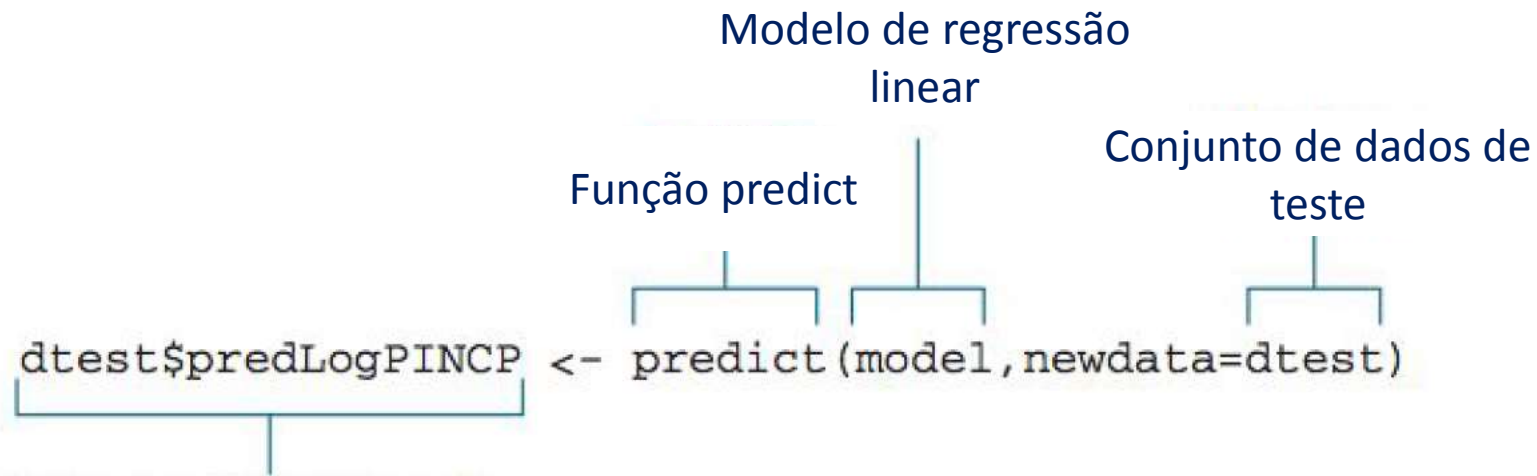


Em R, usamos a função `predict()` para fazer  
fazer as previsões a partir do modelo



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



Data Science Academy

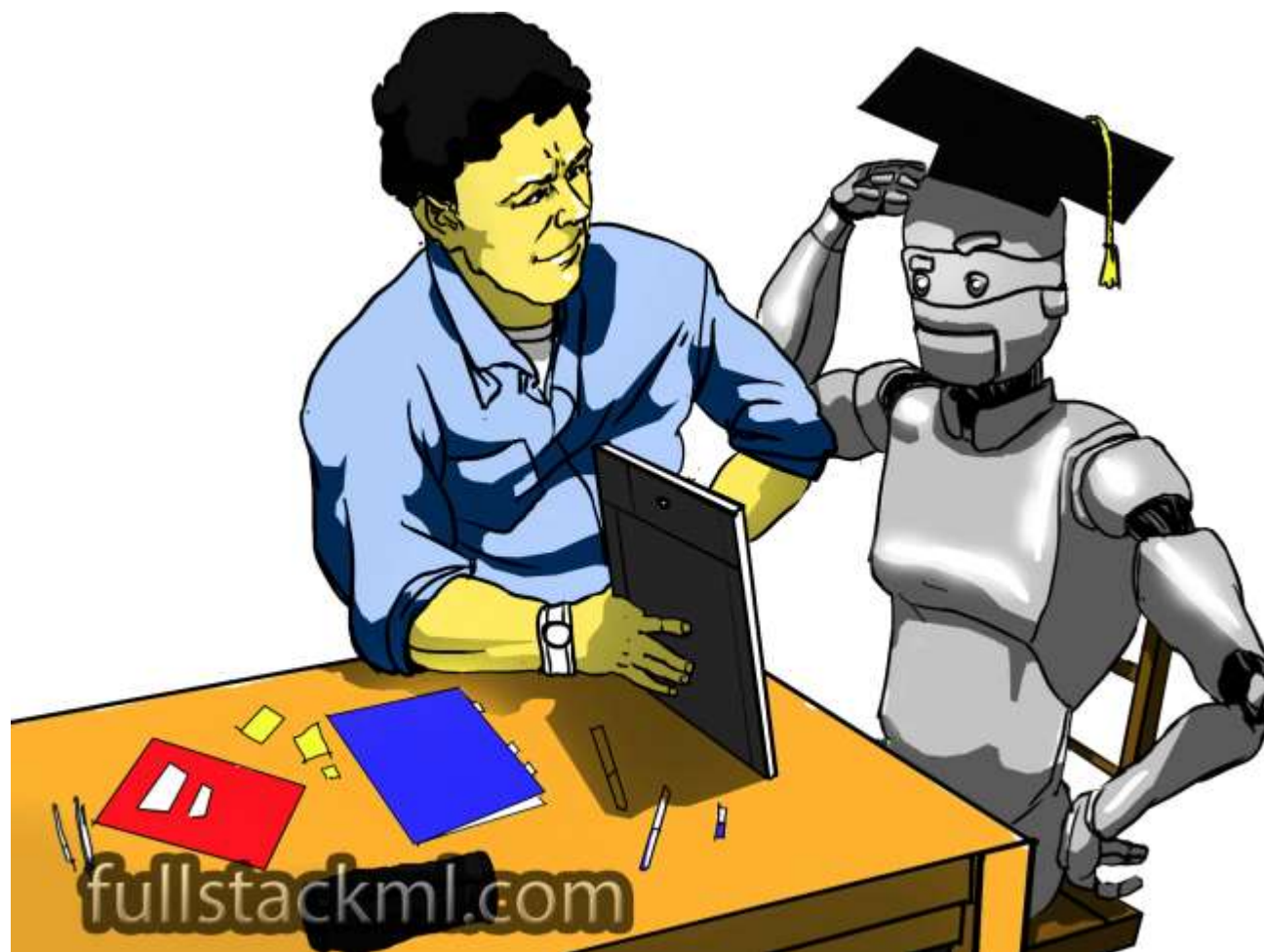


Data Science Academy



Data Science Academy





Data Science Academy



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

Exemplo: determinar o diagnóstico de uma doença em um paciente, observando as características similares em outros grupos de pacientes



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

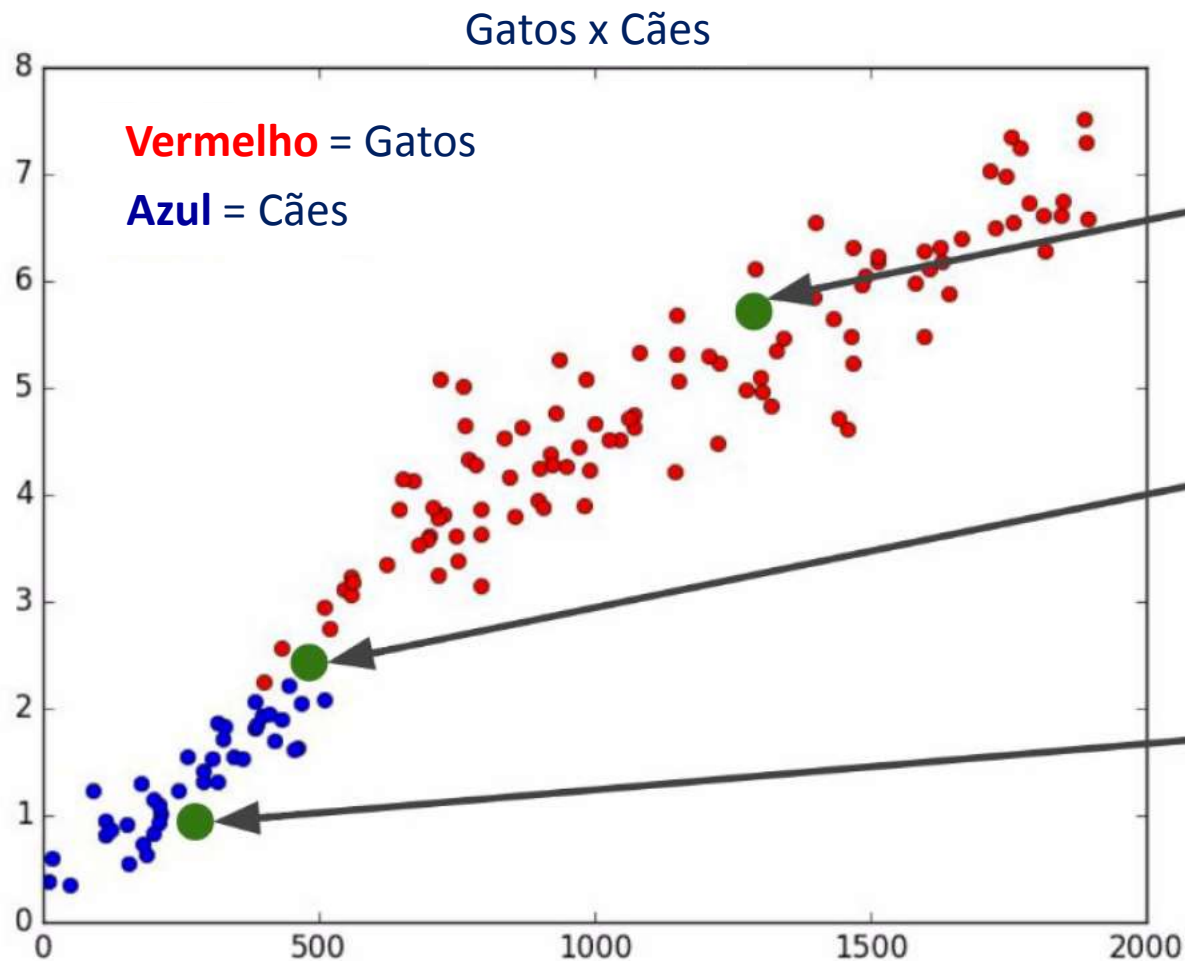


# Classificador kNN

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



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?

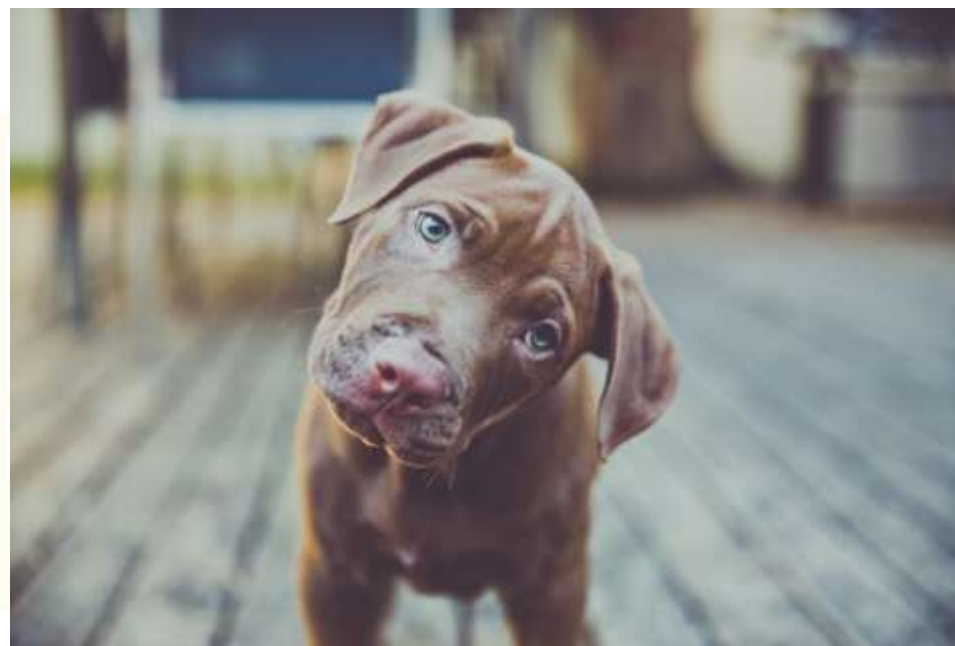
Este novo ponto de dado representa um gato ou um cachorro?



Data Science Academy

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

Achou que nunca mais fosse  
ouvir falar no Teorema de  
Pitágoras, não é?



Data Science Academy



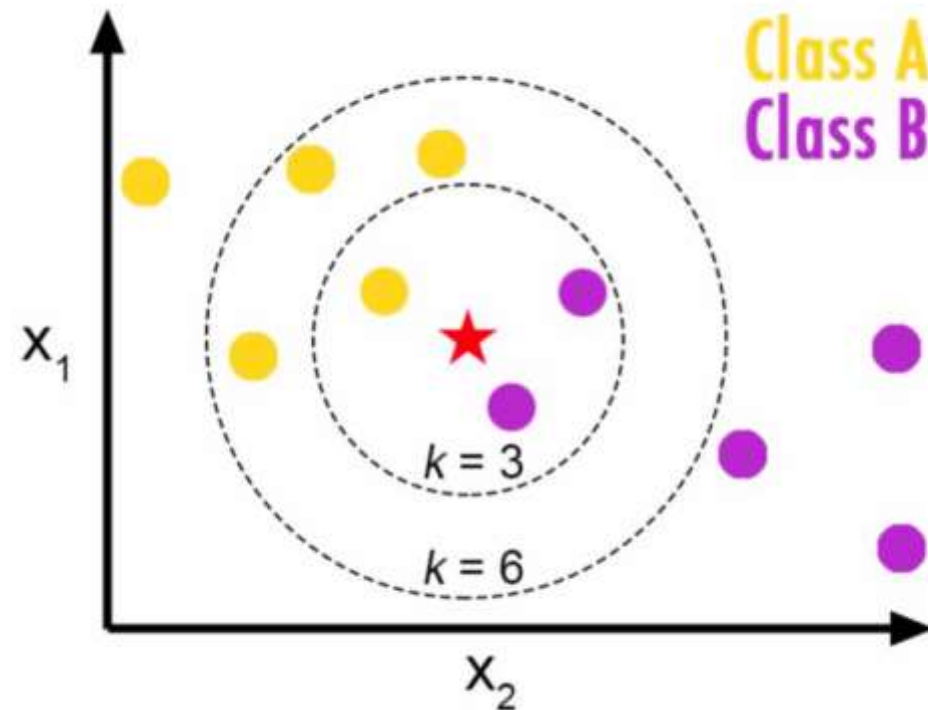
# Classificação kNN

- 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

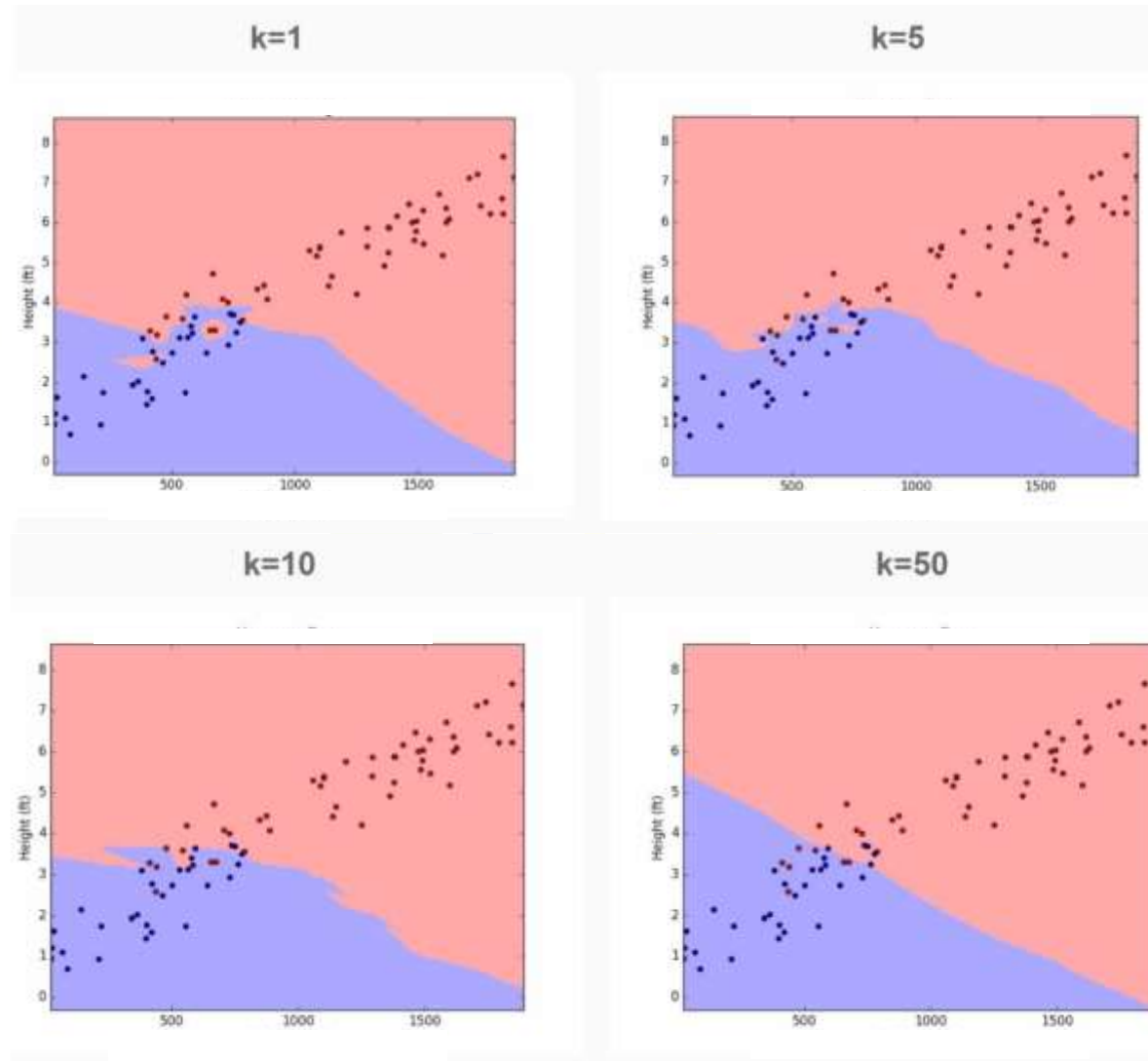


Data Science Academy

## O valor de k faz a diferença







Data Science Academy





# Support Vector Machine



Data Science Academy



# Support Vector Machine



Data Science Academy



# Support Vector Machine



Data Science Academy

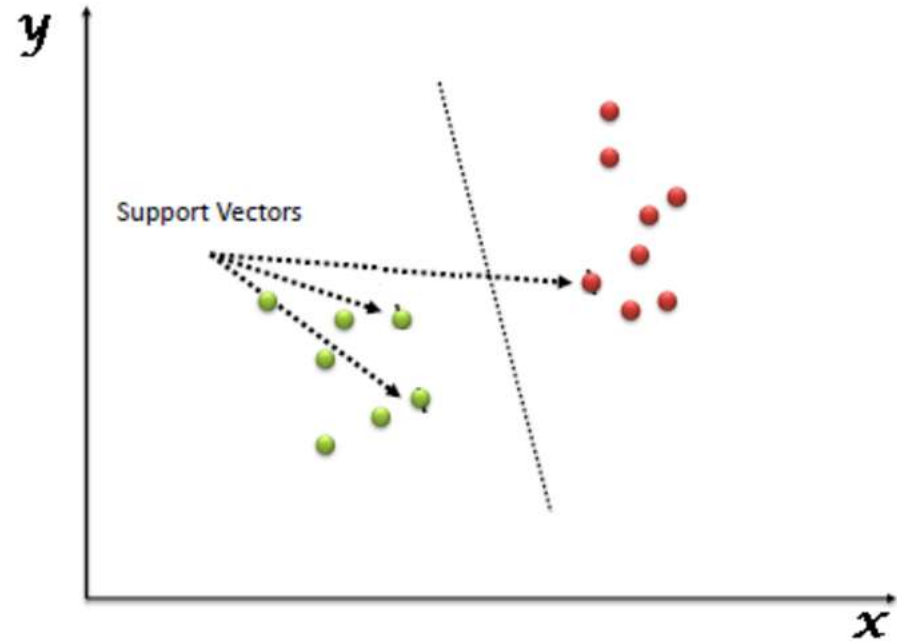


# Support Vector Machine




Data Science Academy

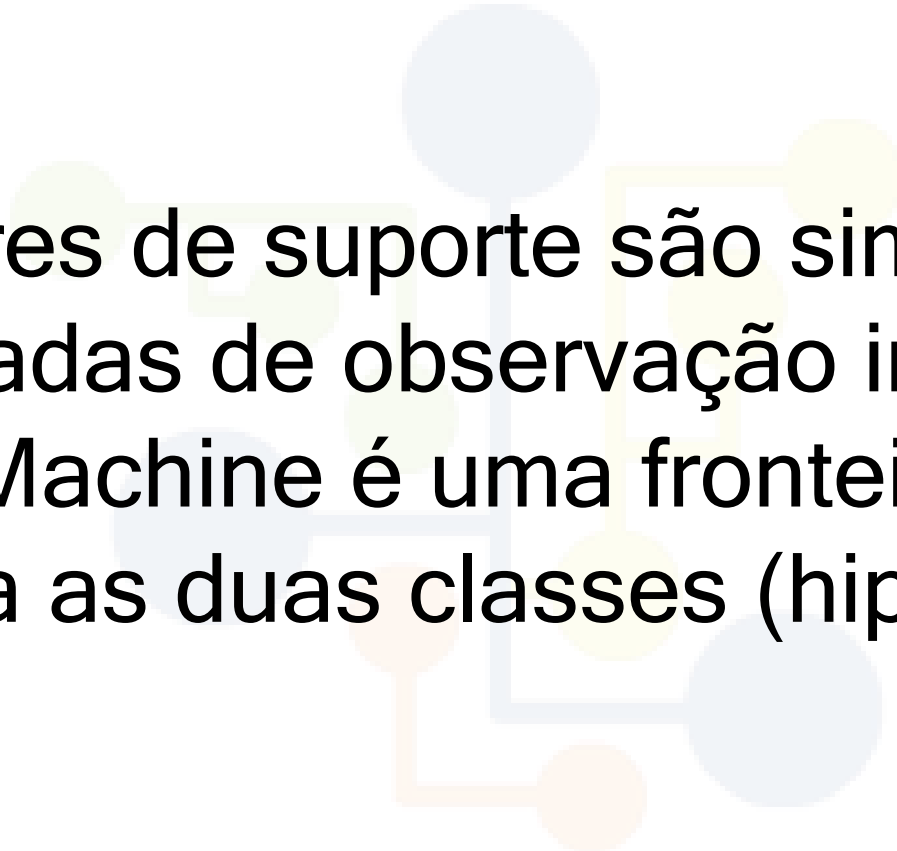
# Support Vector Machine



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

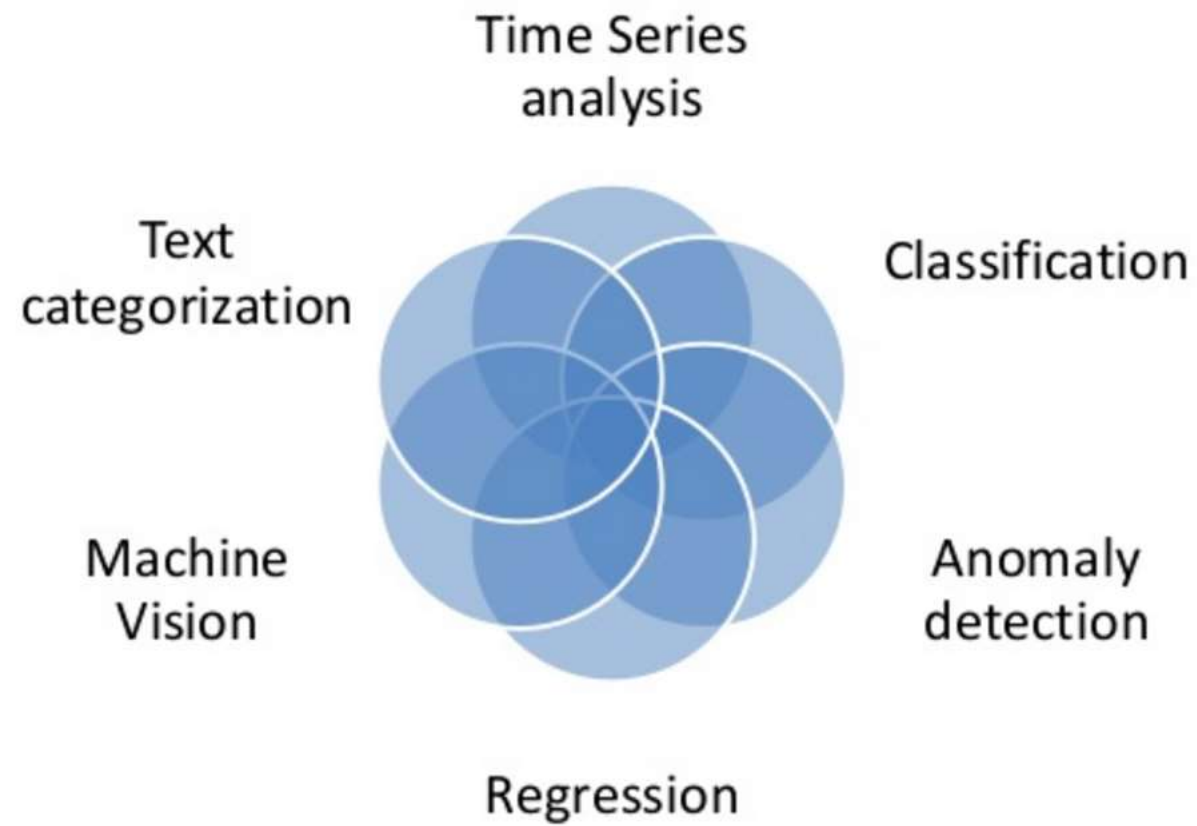


# Support Vector Machine



Data Science Academy





Data Science Academy



Obrigado!



Data Science Academy