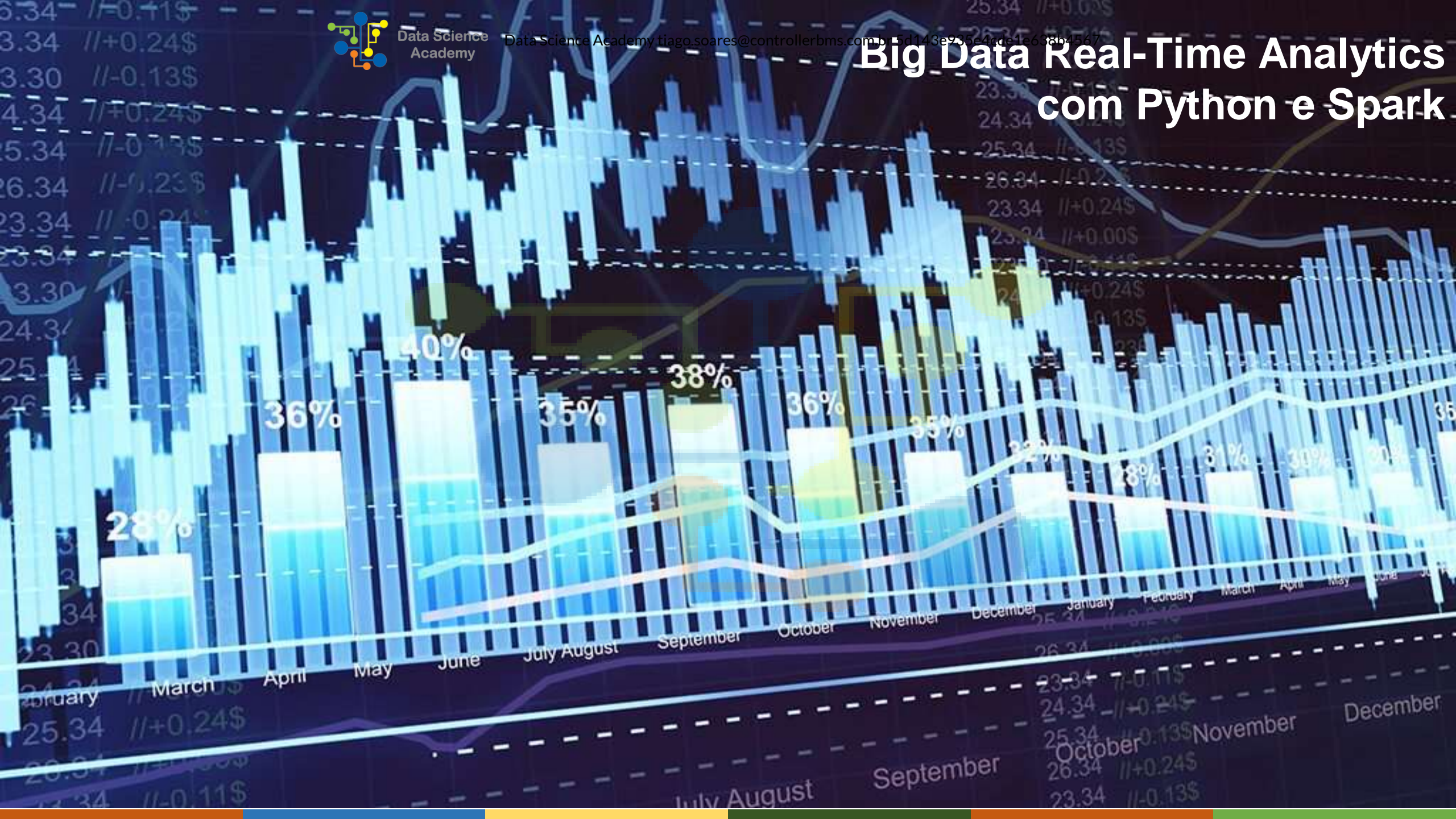




Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4ade1e638b4567

Big Data Real-Time Analytics com Python e Spark





Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Big Data Real-Time Analytics com Python e Spark

A large, faint, stylized network diagram in the background, featuring a central blue node connected to several other nodes in green, yellow, and orange, with lines representing connections.

Seja muito bem-vindo(a)!



Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

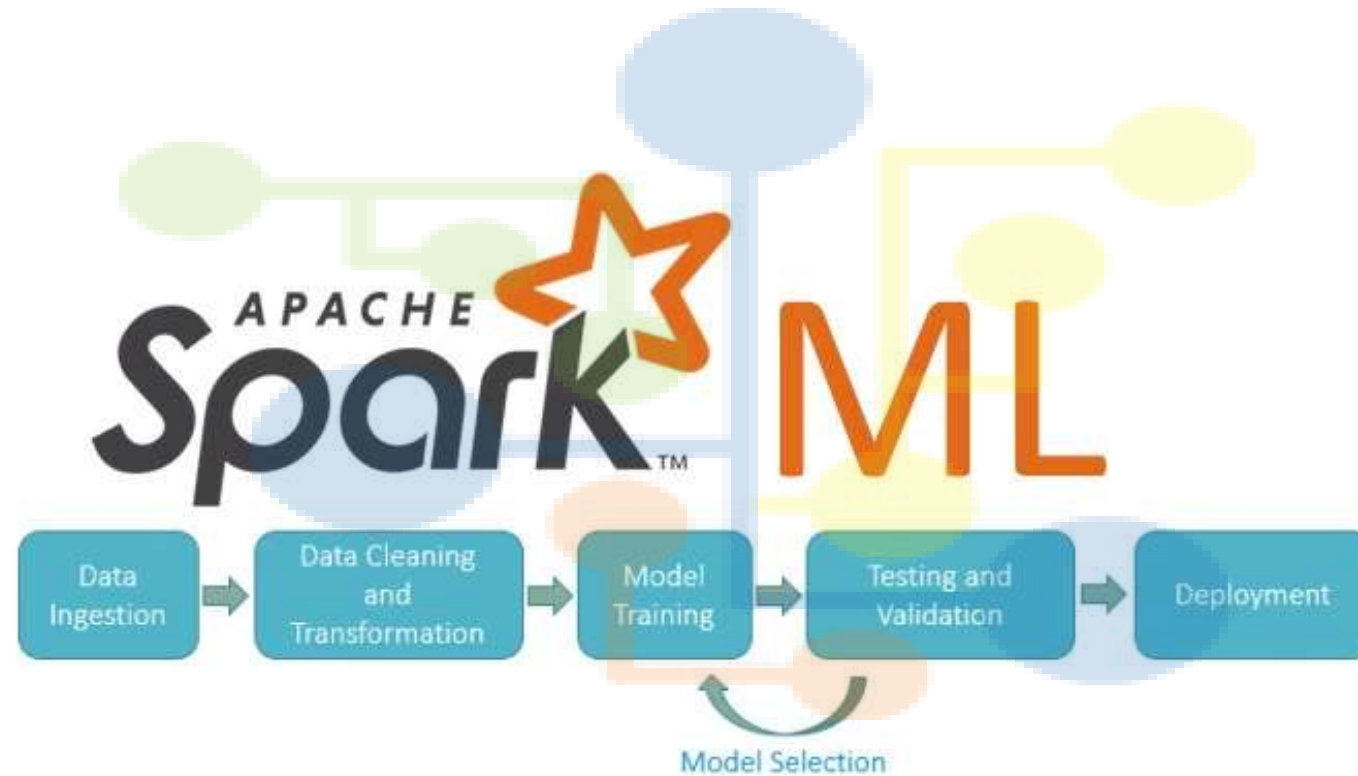
Big Data Real-Time Analytics com Python e Spark

Apache Spark Machine Learning





Apache Spark Machine Learning





Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Big Data Real-Time Analytics com Python e Spark

Machine Learning com Apache Spark





Data Science
Academy

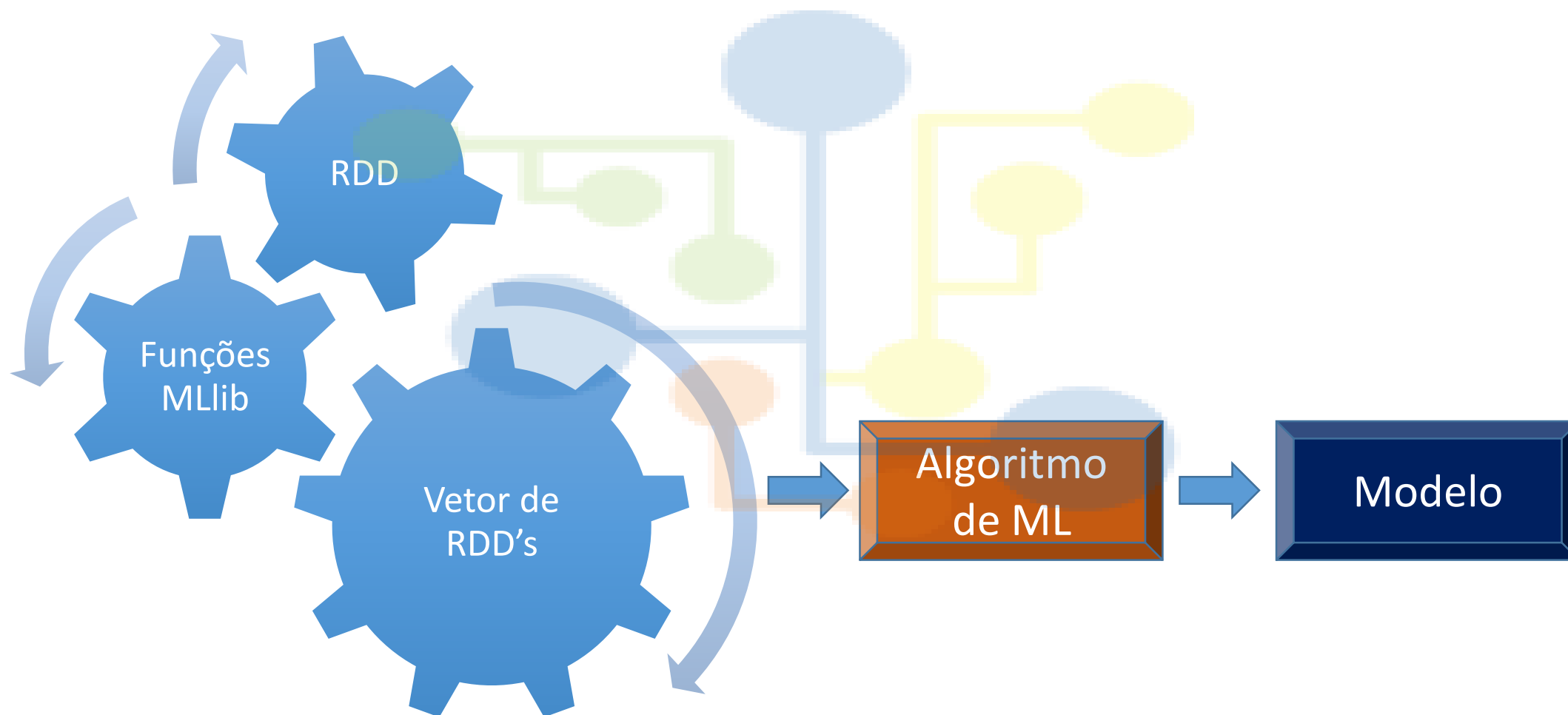
Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Machine Learning com Apache Spark



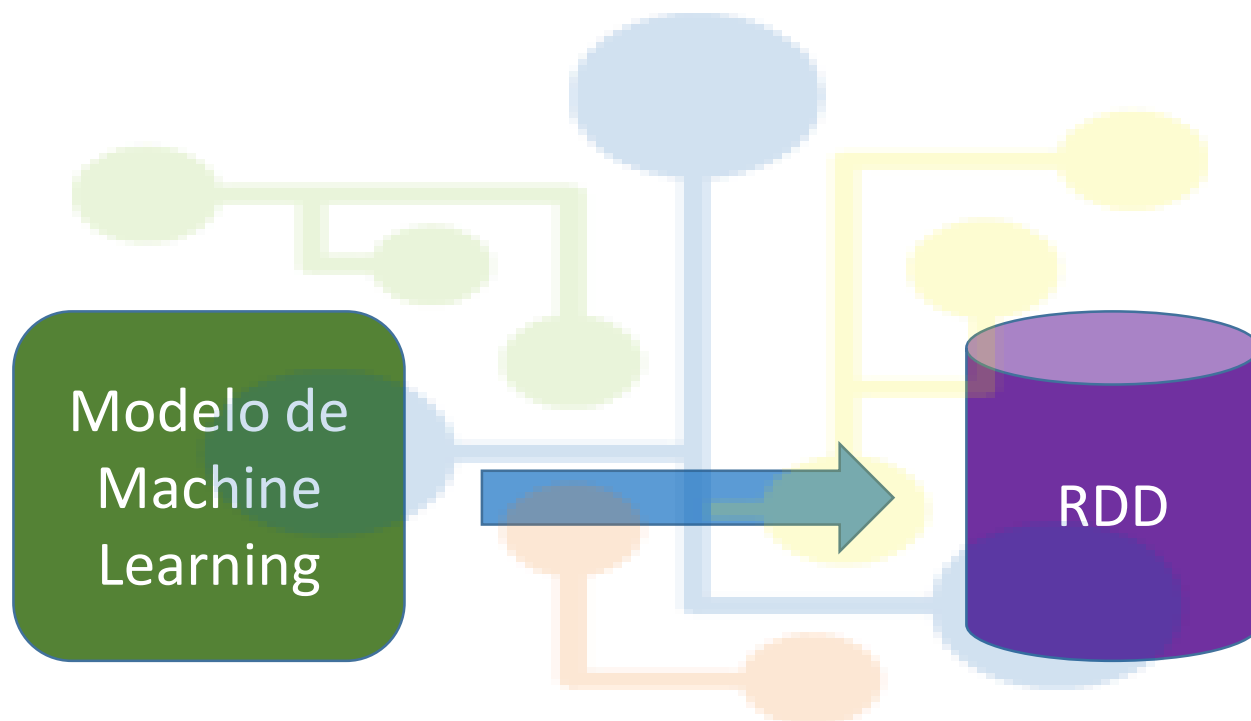


Machine Learning com Apache Spark





Machine Learning com Apache Spark



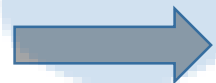


Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Machine Learning com Apache Spark

Spark
MLlib



Computer Clusters



Data Science
Academy

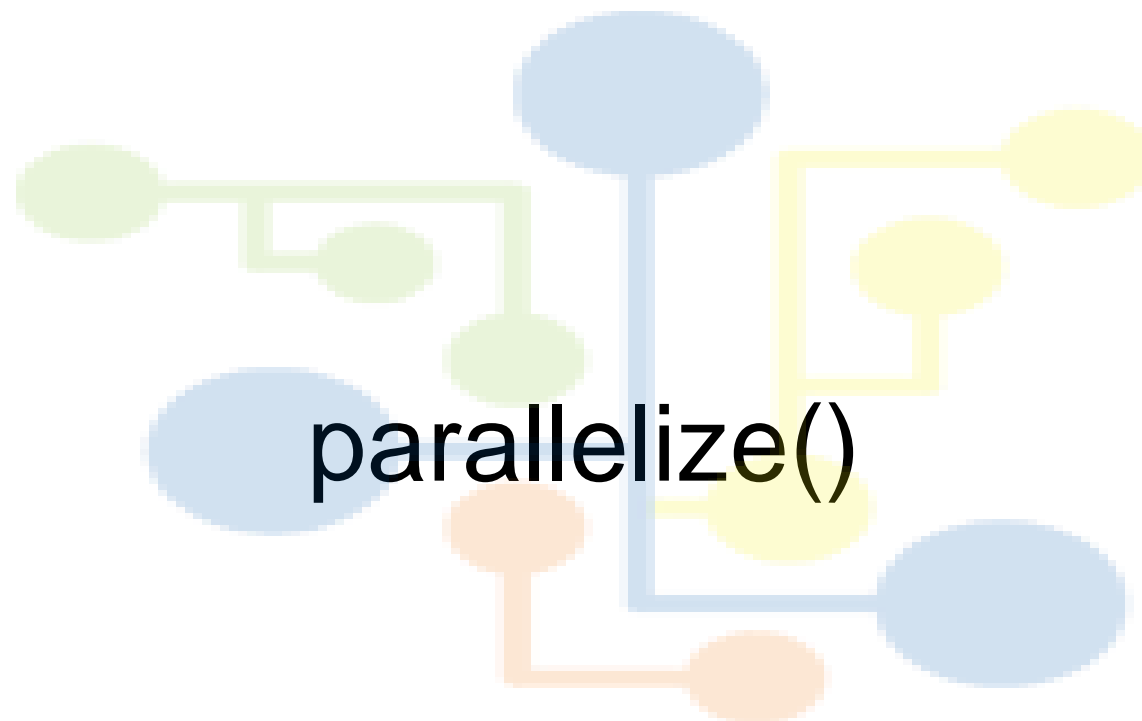
Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Machine Learning com Apache Spark





Machine Learning com Apache Spark



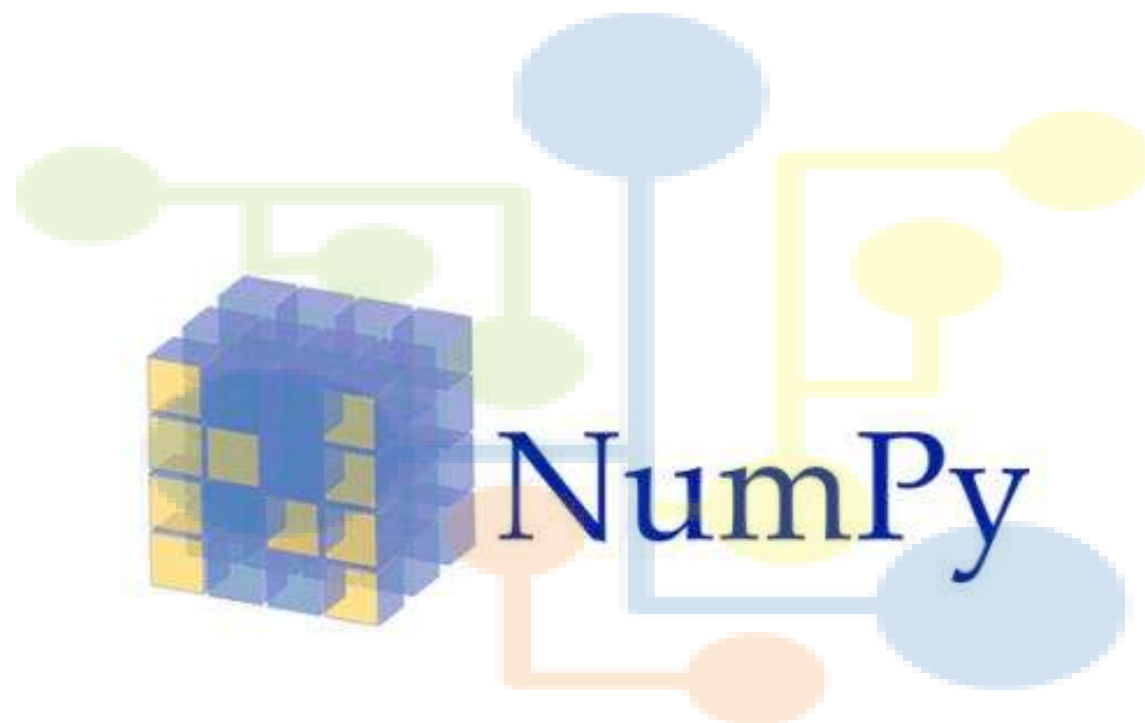
parallelize()



Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Machine Learning com Apache Spark





Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Big Data Real-Time Analytics com Python e Spark

Analytics e Dataficação





Analytics e Dataficação

Analytics



Datafication
(Dataficação)





Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Analytics e Dataficação



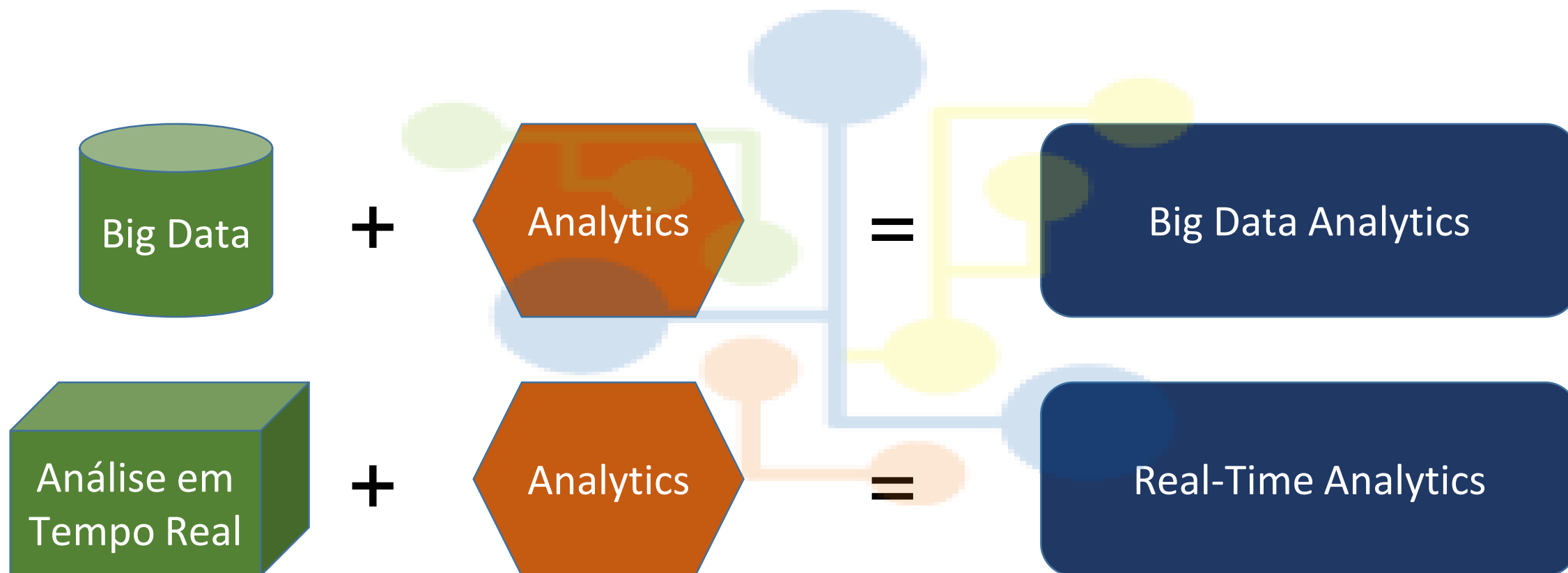


Analytics e Dataficação

Analytics é o processo de coletar dados e gerar insights para tomadas de decisões baseadas em fatos.



Analytics e Dataficação





Analytics e Dataficação

Podemos hoje analisar grandes conjuntos de dados ou dados gerados em tempo real e coletar insights que não podiam ser coletados há pouco tempo atrás.



Tipos de Analytics





Tipos de Analytics

Analytics	Descrição
Descritiva	Compreender o que aconteceu
Exploratória	Descobrir porque alguma coisa aconteceu
Inferencial	Compreender uma população a partir de uma amostra
Preditiva	Prever o que vai acontecer
Causal	O que ocorre com uma variável quando outra é alterada
Deep	Técnicas avançadas para compreender grandes conjuntos de dados de diversas fontes



Tipos de Analytics



Análise
Exploratória

The diagram features two large, overlapping hexagons. The left hexagon is green and contains the text 'Análise Exploratória'. The right hexagon is blue and contains the text 'Análise Preditiva'. Behind these hexagons is a faint, larger-scale network diagram with nodes and connecting lines in various colors (blue, green, yellow, orange).

Análise
Preditiva



Data Science
Academy

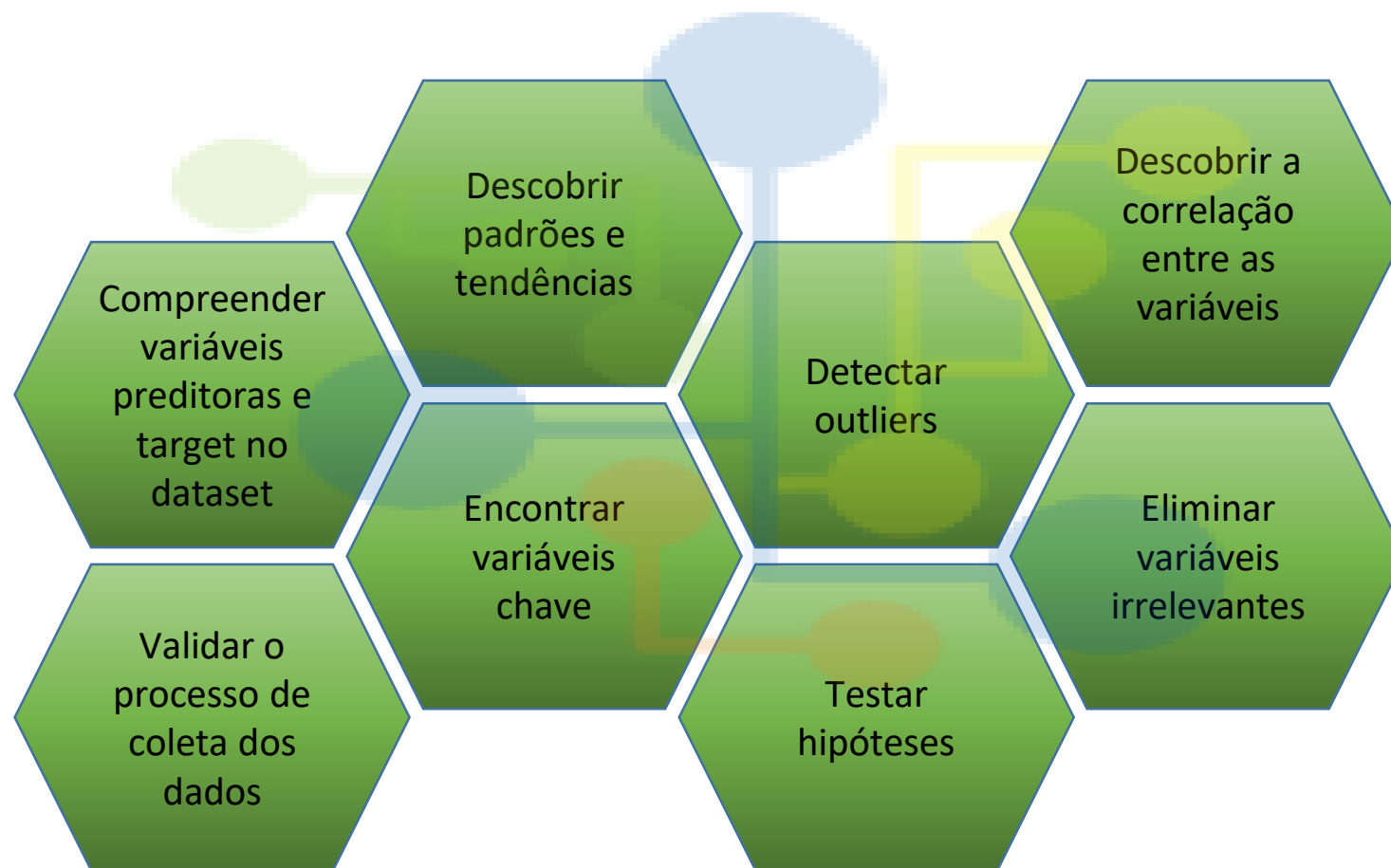
Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Análise Exploratória de Dados





Análise Exploratória de Dados





Análise Exploratória de Dados

Ferramentas usadas na
Análise Exploratória de Dados



Matriz de Correlação

Histogramas

Scatterplots

Boxplots

Principal Component Analysis

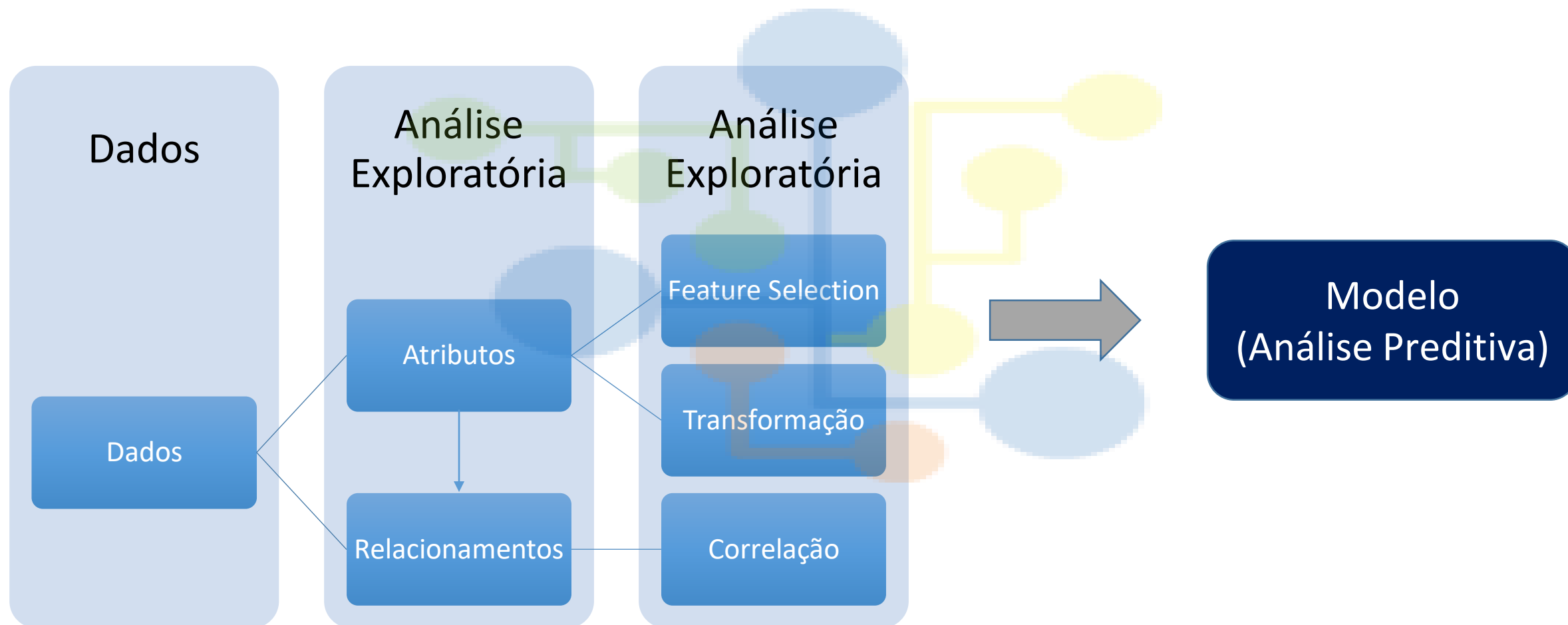


Análise Preditiva





Análise Preditiva





Análise Preditiva



Aprendizagem
Supervisionada

Aprendizagem
Não
Supervisionada



Análise Preditiva

- Tenta fazer previsões a partir do treinamento com dados de entrada e dados de saída.
- Os modelos são construídos em datasets de treino.
- Os modelos são usados para prever o futuro.

Pode ser:

- Regressão (dados numéricos e contínuos)
- Classificação (classes)



Aprendizagem
Supervisionada



Análise Preditiva

- Dados históricos contém variáveis preditoras e a variável alvo (target).
- O conjunto de dados é separado em dados de treino e dados de teste.
- Dados de treino são usados para treinar o modelo.
- Dados de teste são usados para testar e validar o modelo.
- Utilizamos uma métrica (como acurácia) para avaliar o modelo.
- Split 70/30 (treino/teste).
- Seleção aleatória dos dados em ambos os datasets.



Aprendizagem
Supervisionada


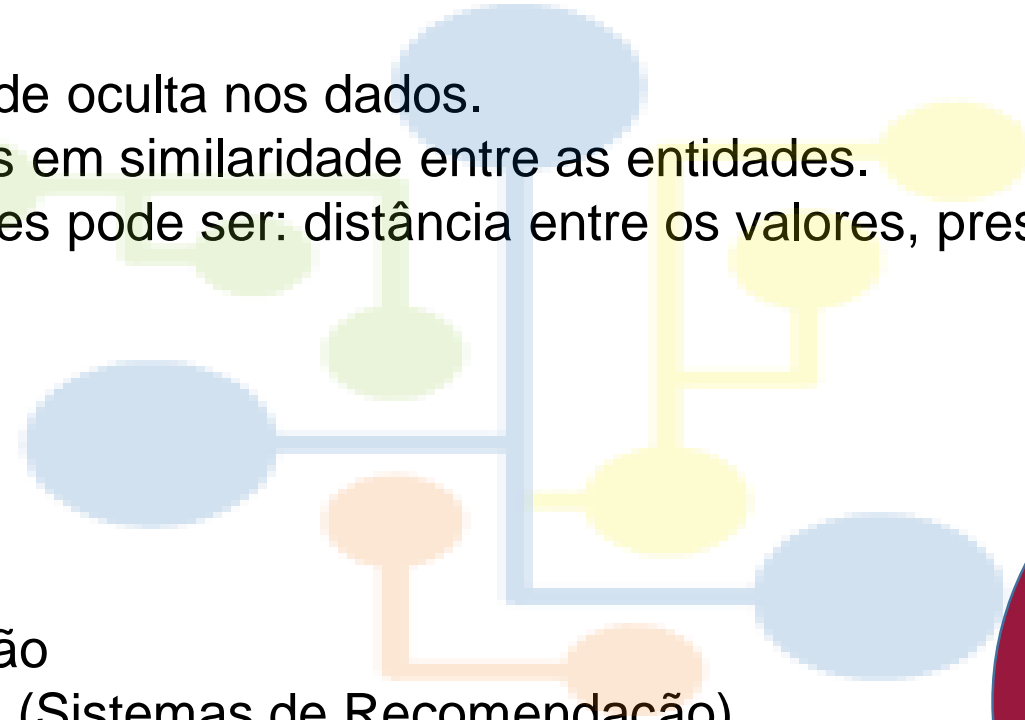


Análise Preditiva

- Busca estrutura ou similaridade oculta nos dados.
- Grupos observados baseados em similaridade entre as entidades.
- Similaridade entre as entidades pode ser: distância entre os valores, presença/ausência de atributos.

Pode ser:

- Clustering
- Regras de Associação
- Filtros Colaborativos (Sistemas de Recomendação)



Aprendizagem
Não
Supervisionada



Análise Preditiva





Viés e Variância

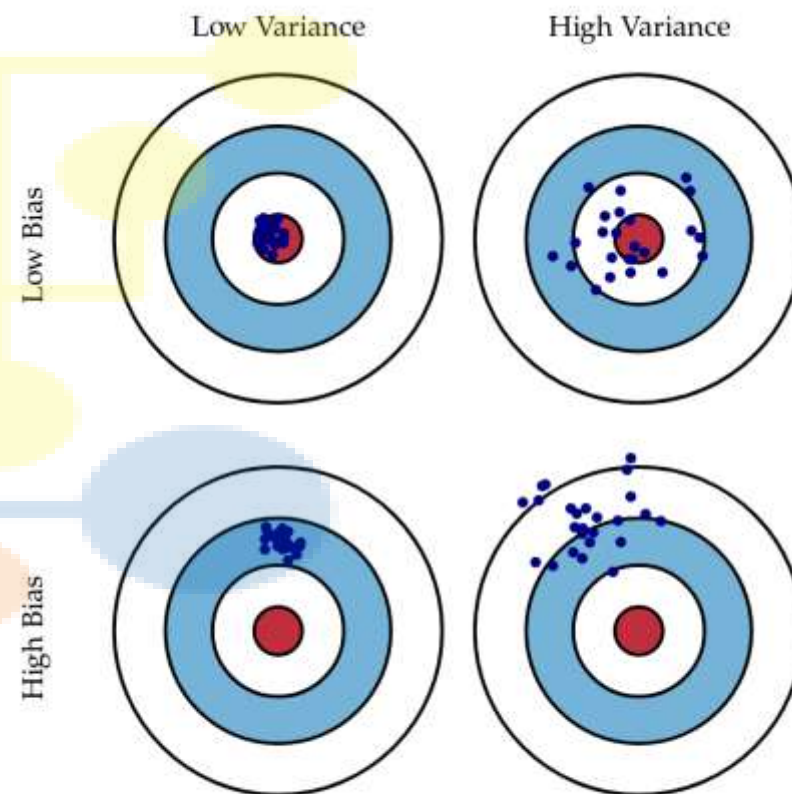
Bias e Variance Trade-off (Viés e Variância)





Viés e Variância

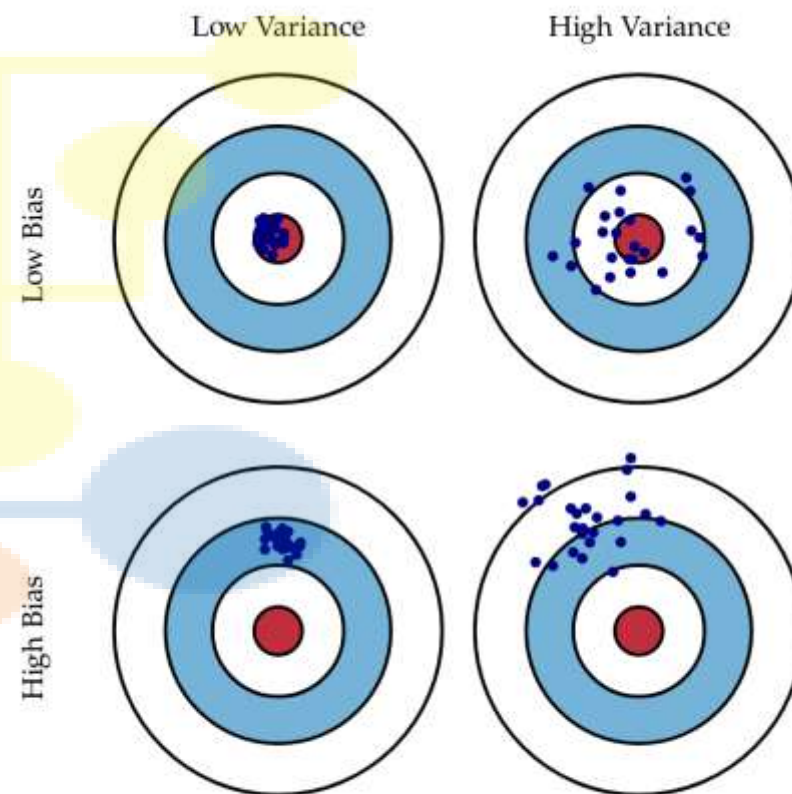
É comum, ao construirmos e escolhermos parâmetros para um modelo, nos depararmos com a seguinte questão: como reduzir o erro do modelo? Para respondermos essa pergunta de maneira correta, em primeiro lugar, devemos entender os 2 principais componentes do erro em nossas previsões: **bias** e **variance**.





Viés e Variância

É comum, ao construirmos e escolhermos parâmetros para um modelo, nos depararmos com a seguinte questão: como reduzir o erro do modelo? Para respondermos essa pergunta de maneira correta, em primeiro lugar, devemos entender os 2 principais componentes do erro em nossas previsões: **bias** e **variance**.





Viés e Variância

Bias (Viés)

É a diferença entre o valor esperado da predição do nosso modelo (média das predições) e o valor real que queremos prever.



Viés e Variância



Variância

É a variabilidade das predições.



Viés e Variância

De forma resumida: o bias está relacionado à habilidade do modelo em se ajustar aos dados, ou seja, se o seu problema é um *underfitting*, o seu modelo tem um alto bias. Já a variância está relacionada à habilidade do modelo se ajustar a novos dados, ou seja, se o seu problema é um *overfitting*, o seu modelo tem uma alta variância.

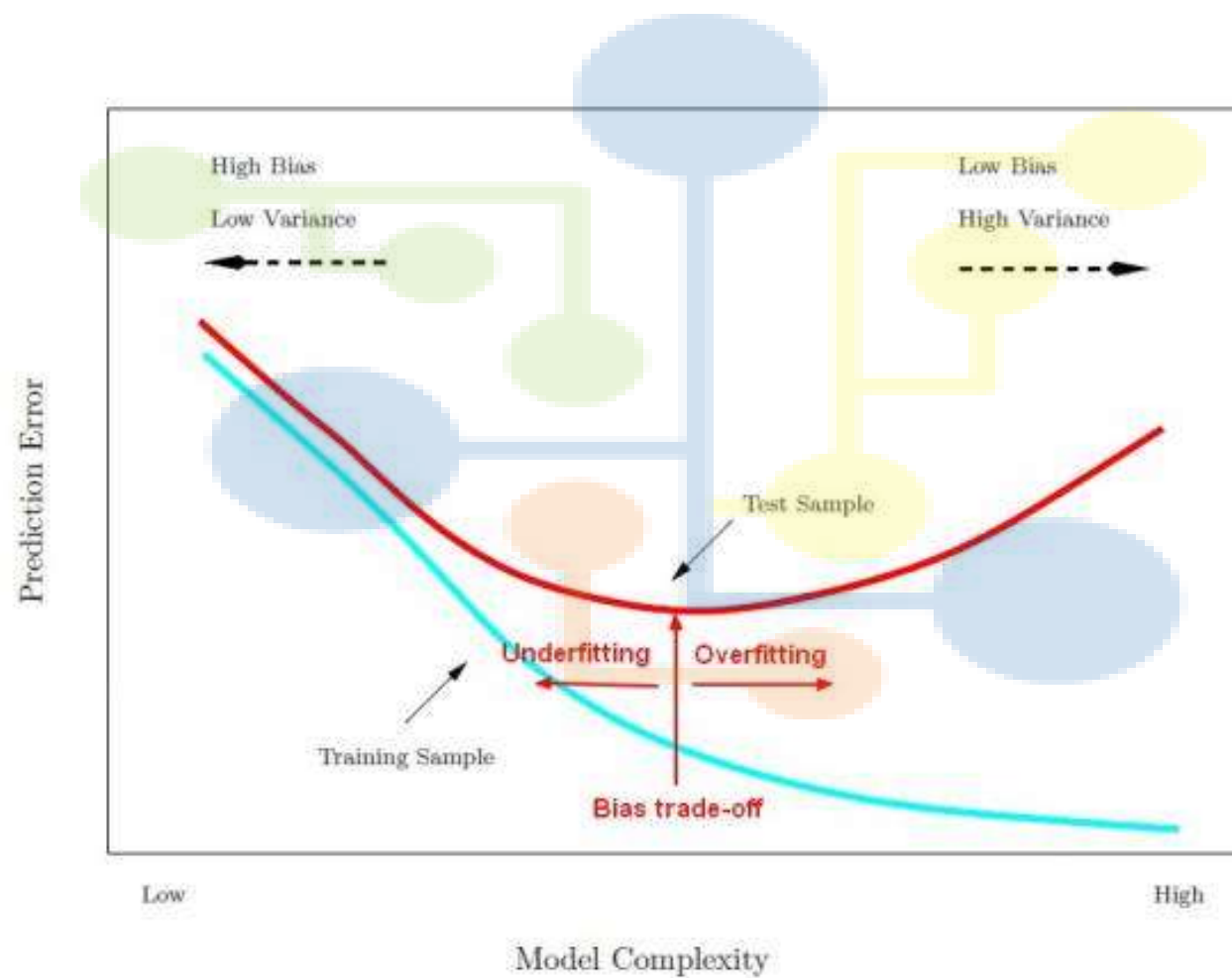


Viés e Variância

O nosso objetivo é reduzir o bias e a variância o máximo que pudermos, entretanto, nos deparamos com um *trade-off* entre *underfitting* e *overfitting*.



Viés e Variância

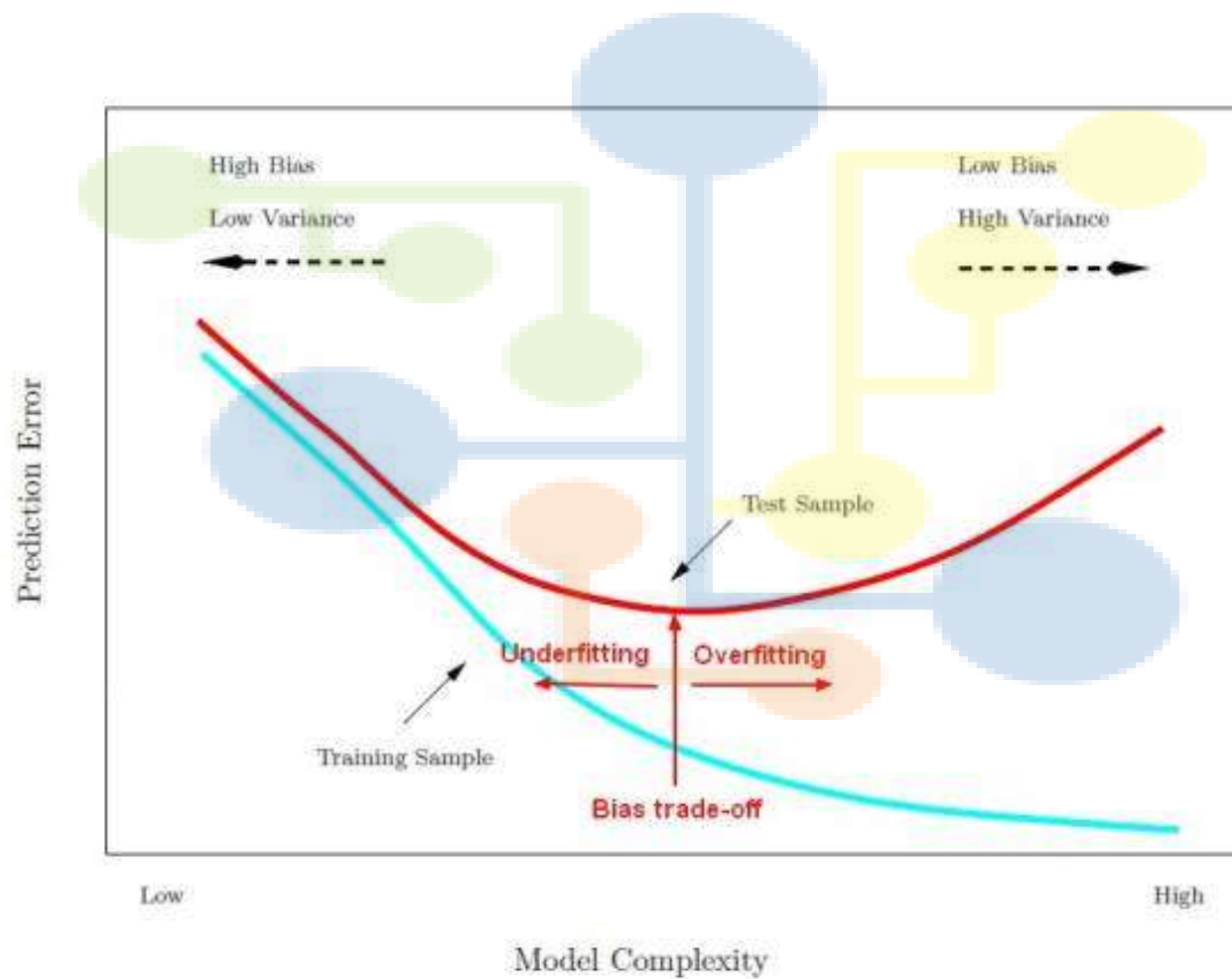


Overfitting



Viés e Variância

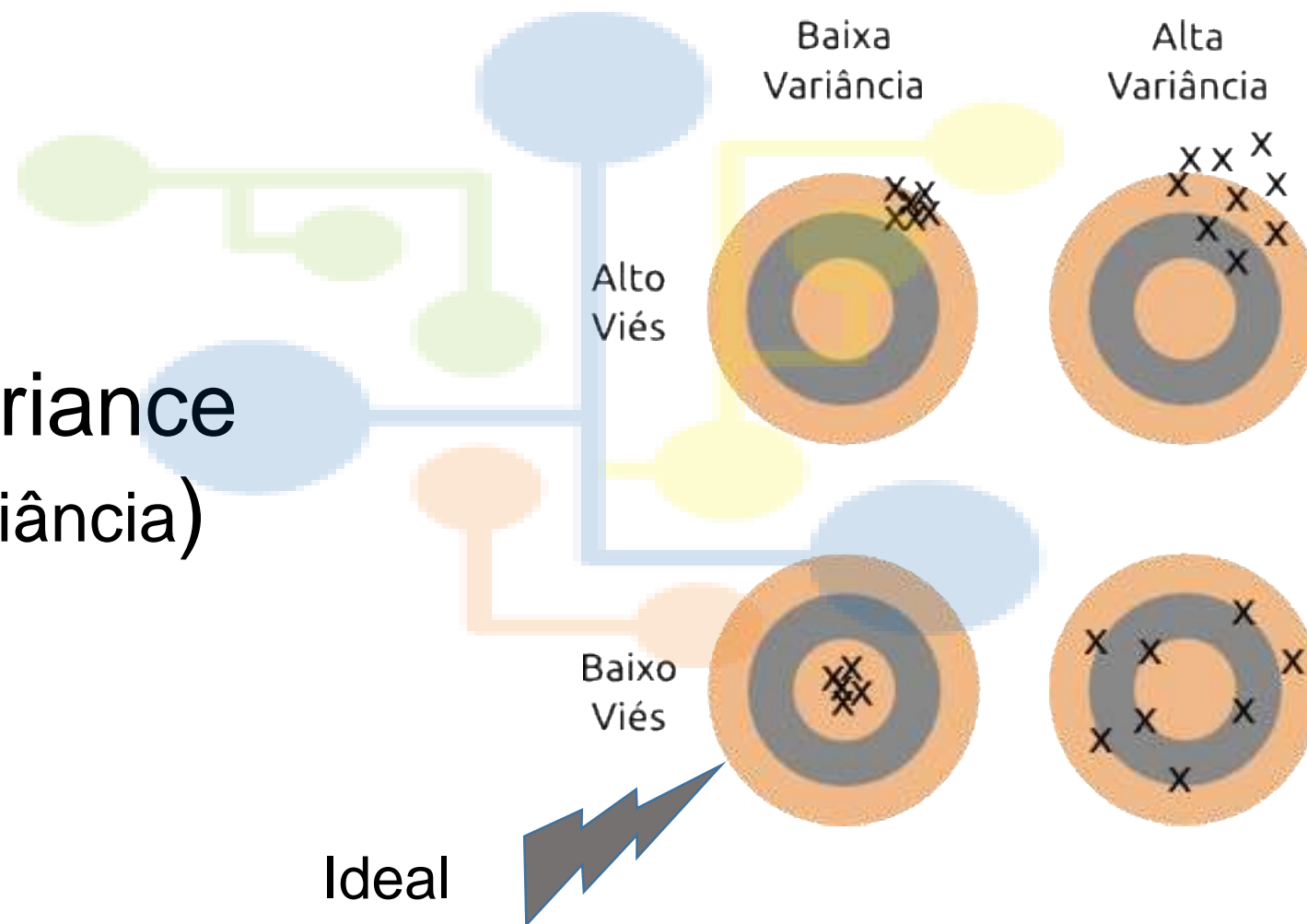
Underfitting





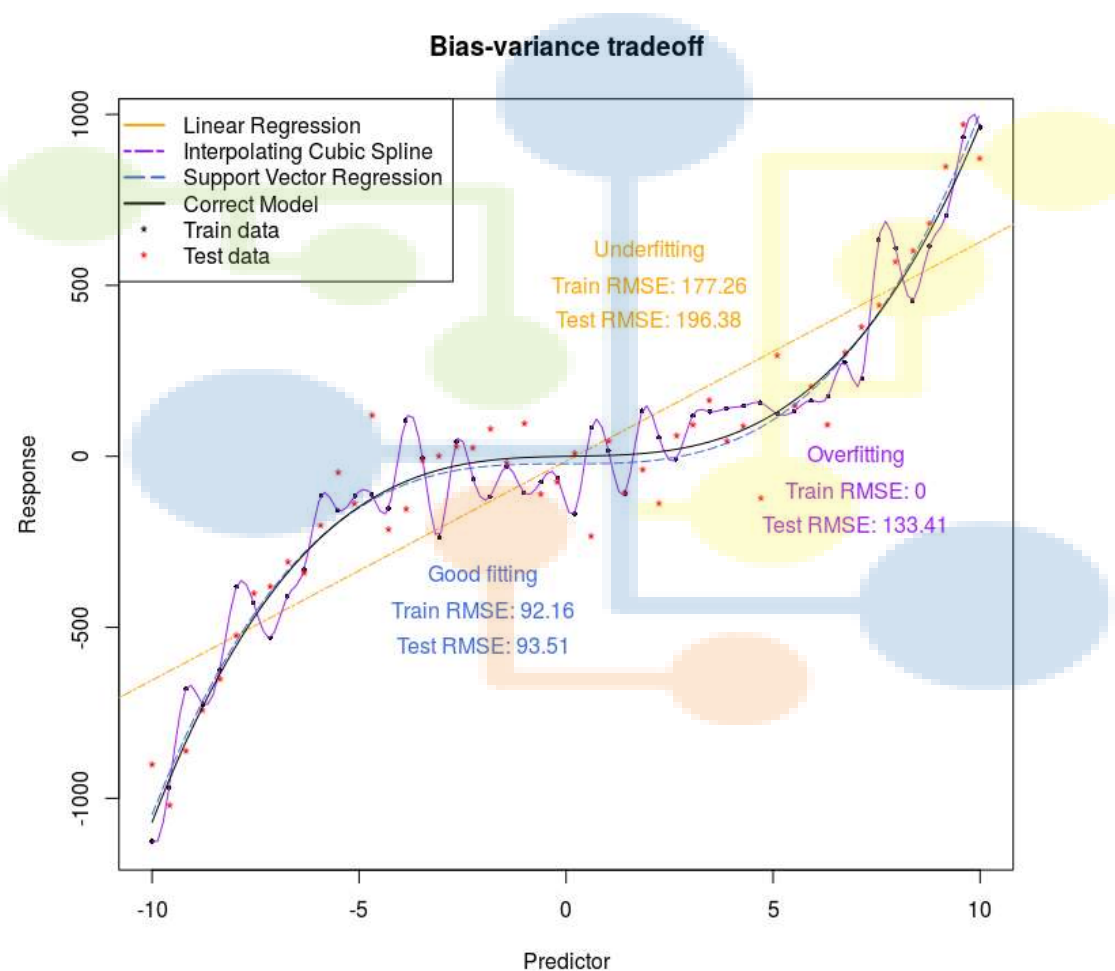
Viés e Variância

Bias e Variance (Viés e Variância)





Viés e Variância





Data Science
Academy

Data Science Academy tiago.soares@controllerbms.com.br 5d143e935e4cde1e638b4567

Big Data Real-Time Analytics com Python e Spark

APIs de Machine Learning do Apache Spark





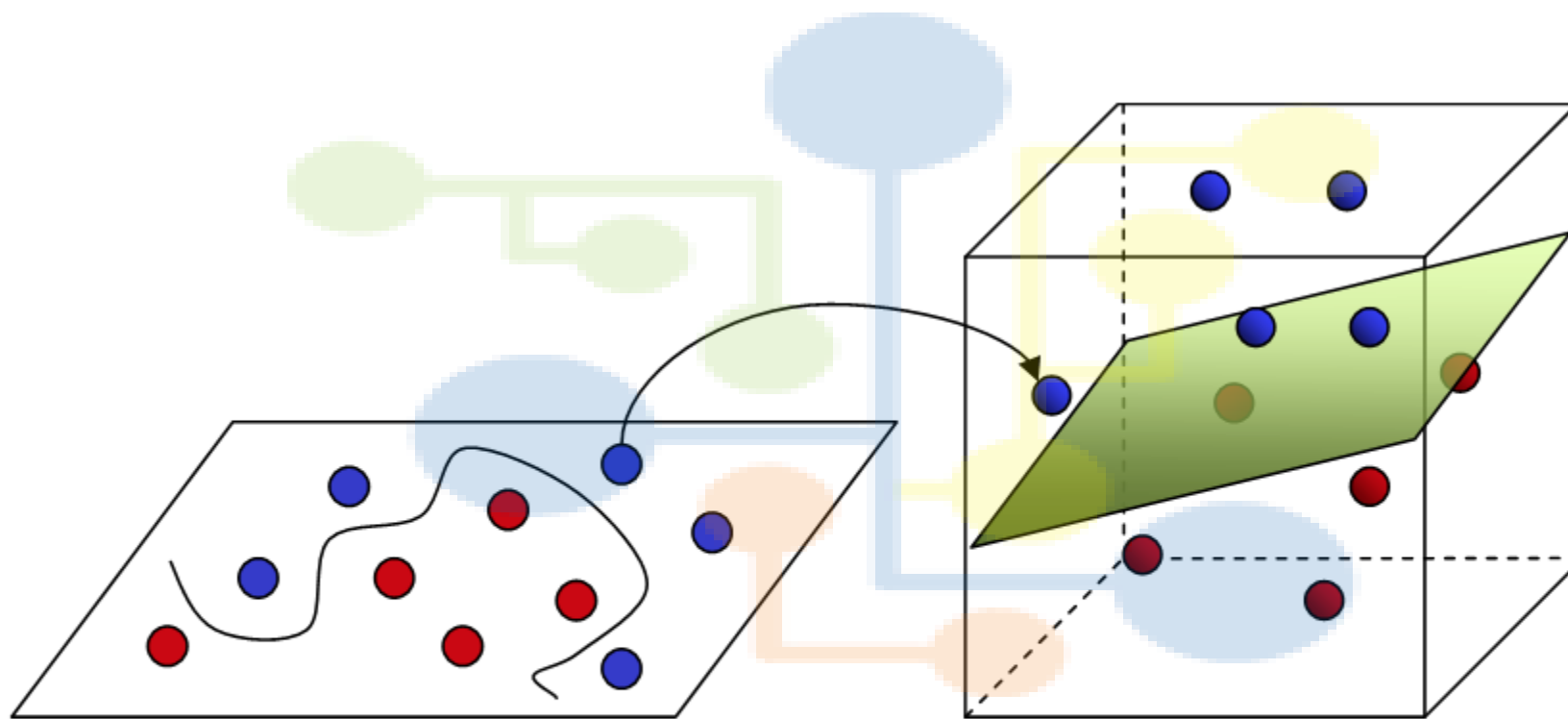
APIs de Machine Learning do Apache Spark

Apache Spark MLlib (Machine Learning Library)

- `spark.mllib` → API original construída para trabalhar com RDD's.
- `spark.ml` → Nova API construída para funcionar também com Dataframes e SparkSQL.



Big Data Real-Time Analytics com Python e Spark

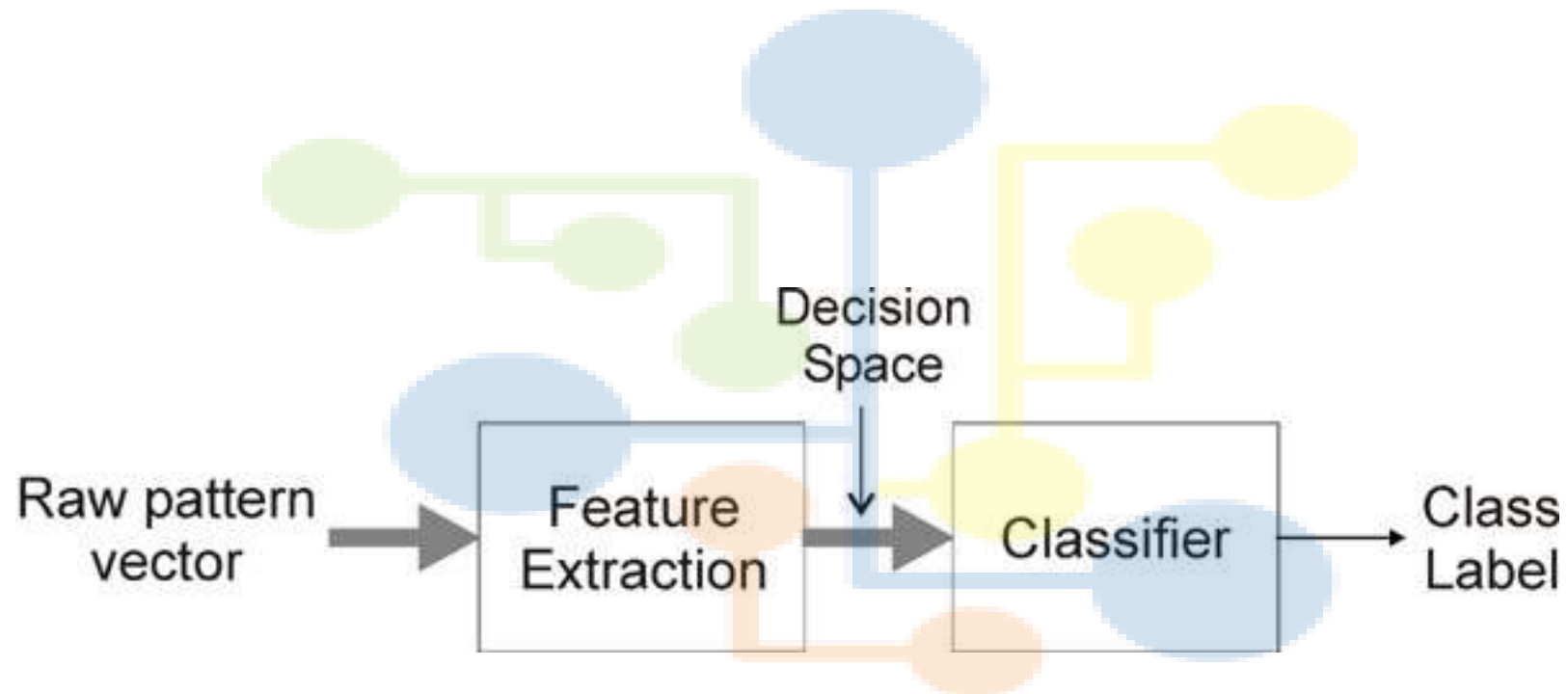


Input Space

Feature Space



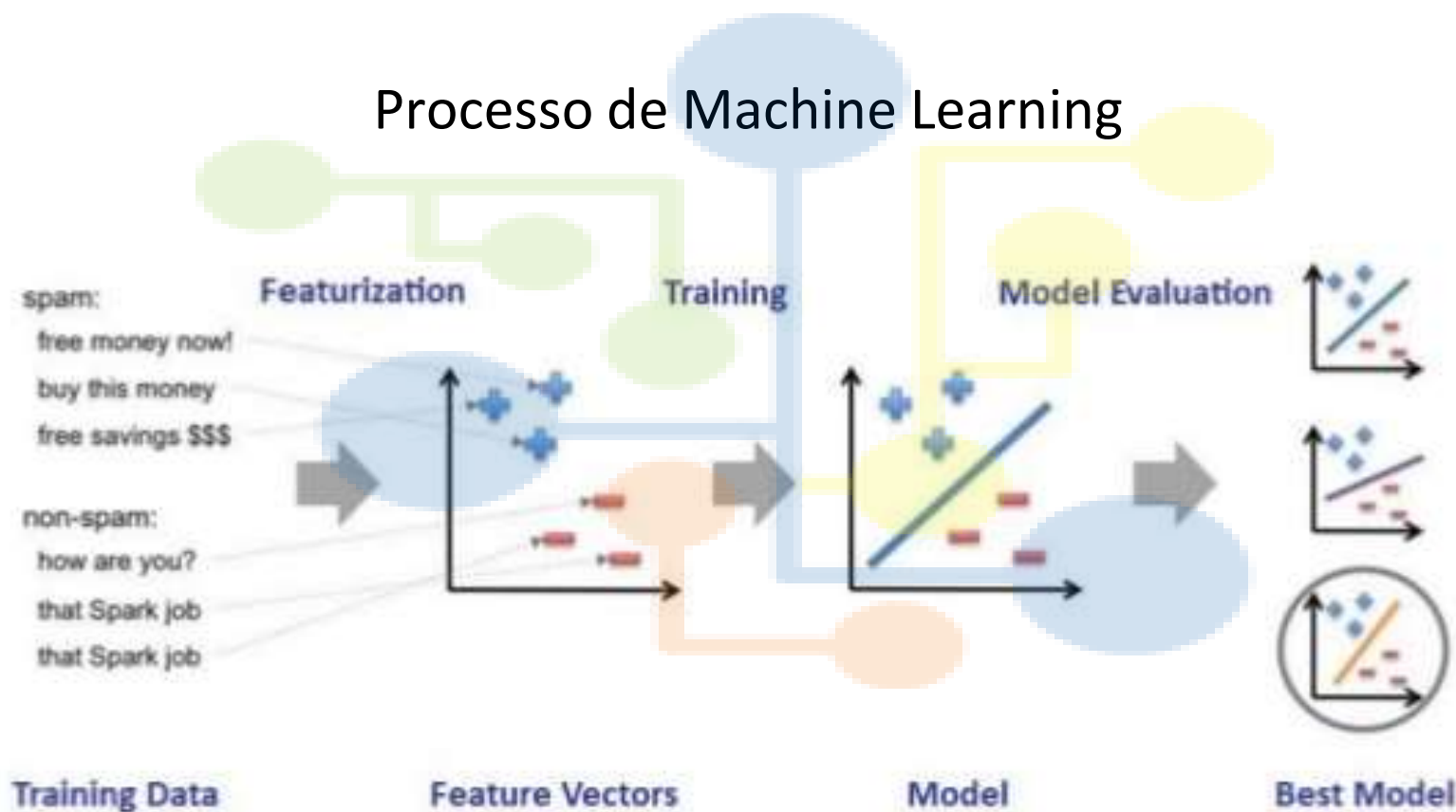
Big Data Real-Time Analytics com Python e Spark





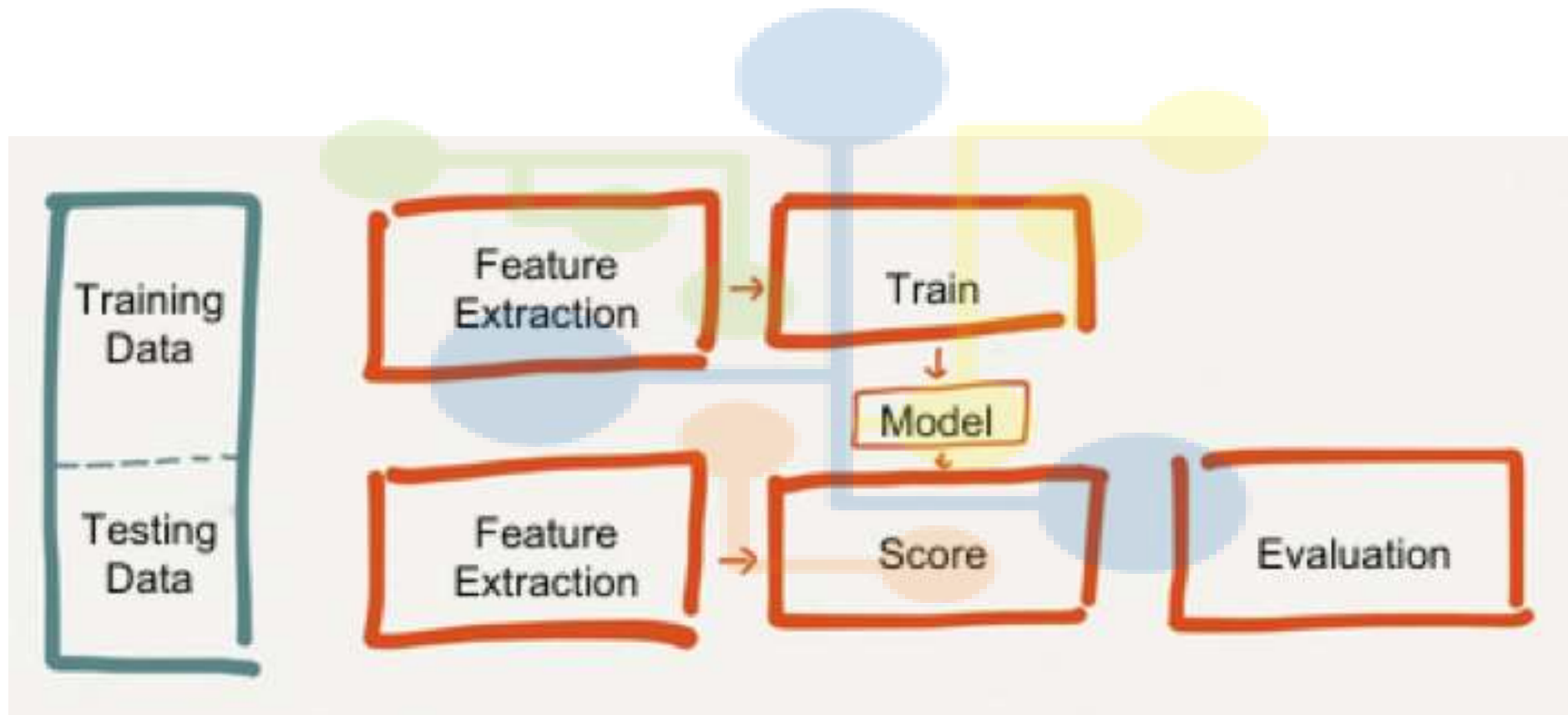
Big Data Real-Time Analytics com Python e Spark

Processo de Machine Learning





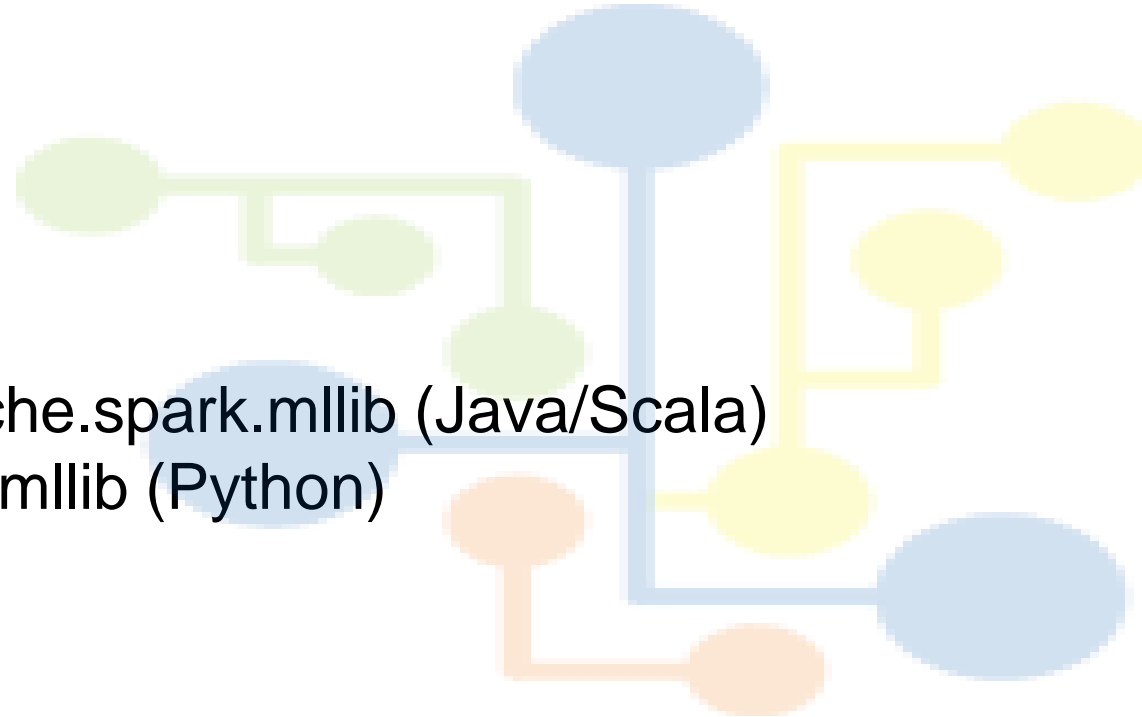
Big Data Real-Time Analytics com Python e Spark





Tipos de Dados

- org.apache.spark.mllib (Java/Scala)
- pyspark.mllib (Python)





Tipos de Dados

Tipo de Dado	Pacote
Vetor	ml.linalg.Vectors

Vetor Denso

(2.0, 4.0, 8.5)

Vetor Esparso

Original (1.0, 0.0, 0.0, 2.0, 0.0)

Representação (5, (0,3), (1.0, 2.0))



Tipos de Dados

Tipo de Dado	Pacote
Vetor	<code>mllib.linalg.Vectors</code>
LabeledPoint	<code>mllib.regression</code>
Rating	<code>mllib.recommendation</code>



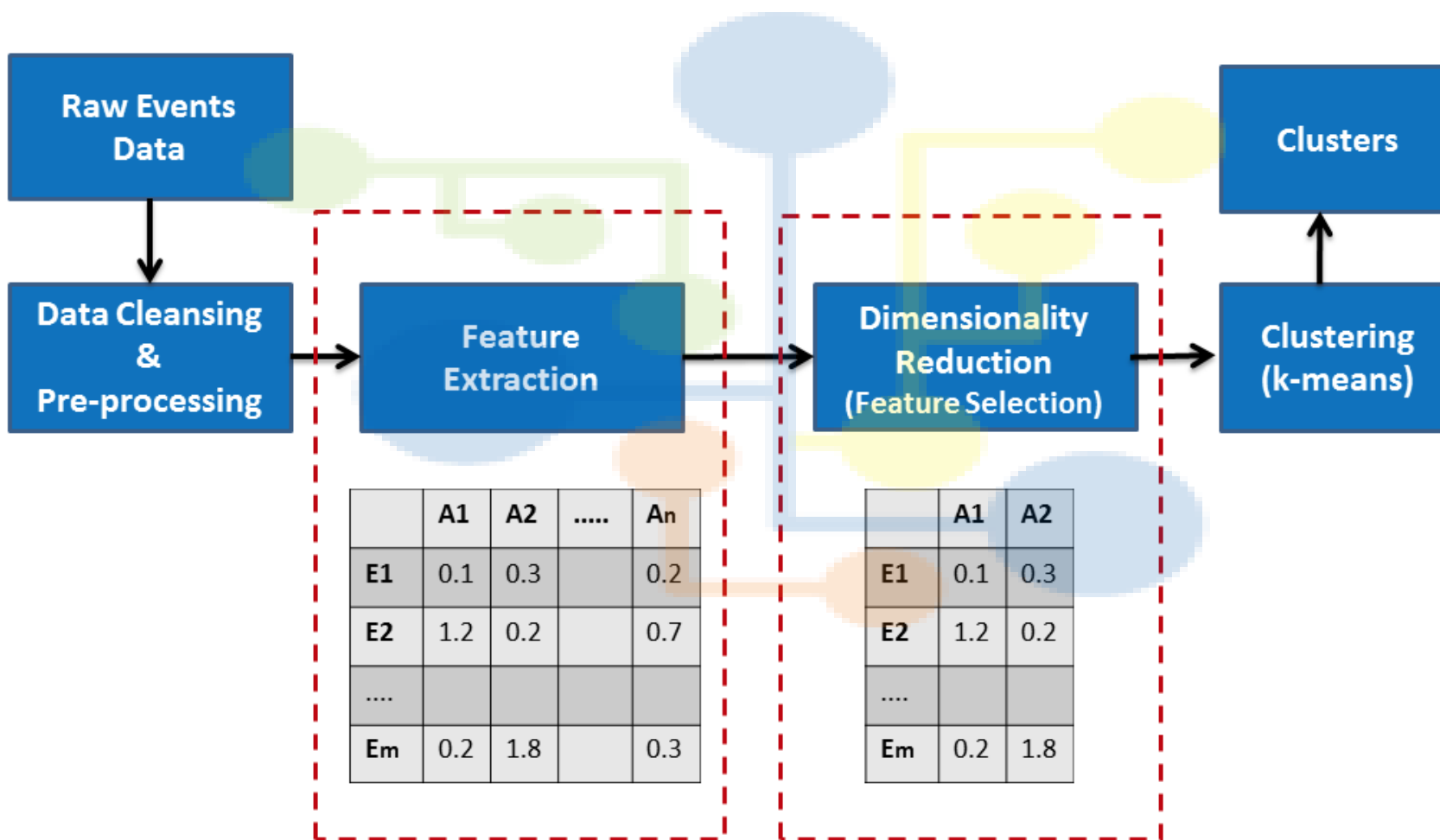
Pipelines

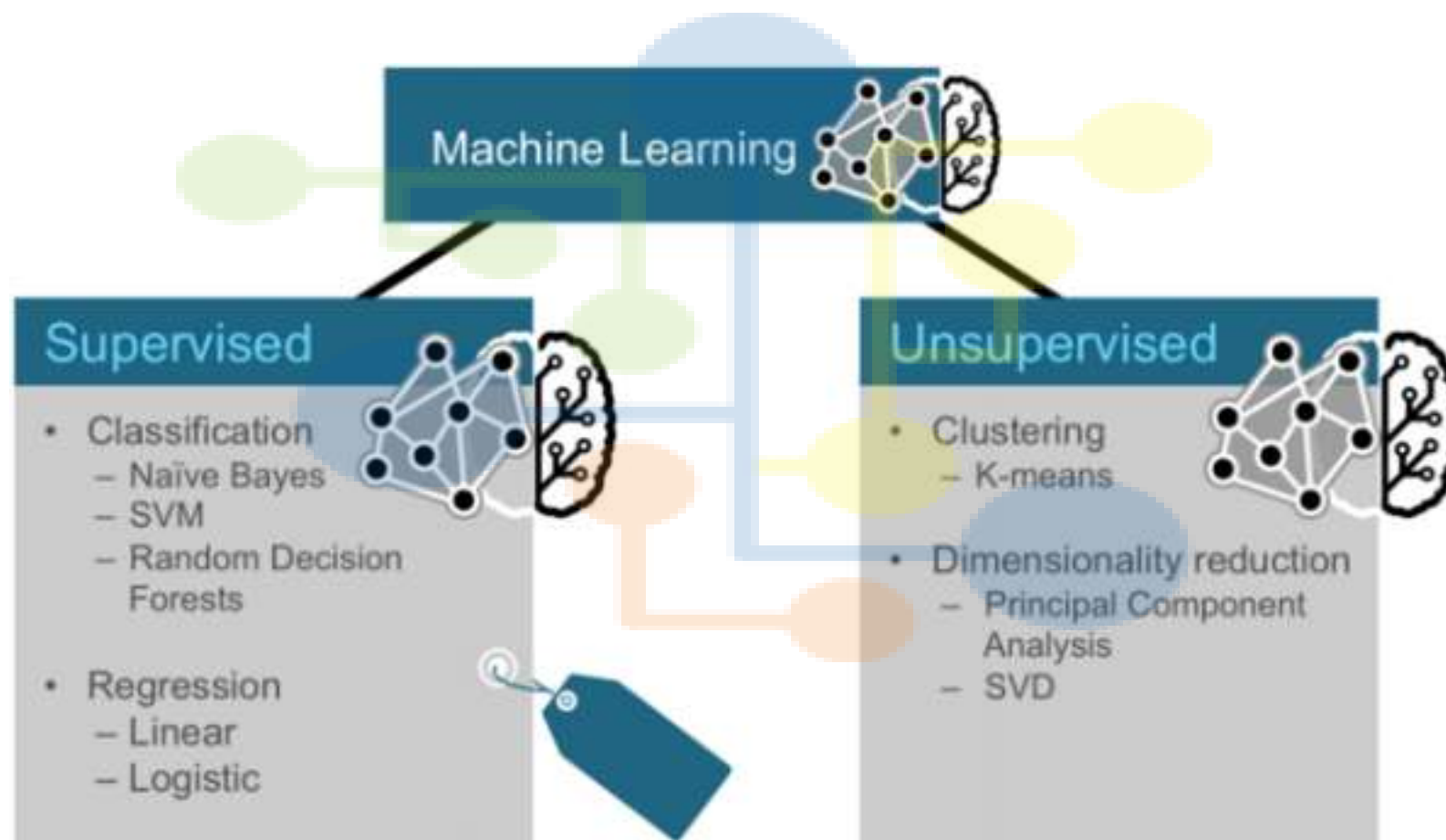
Pipeline consiste de uma série de transformações e ações que precisam ser realizadas para criar um modelo

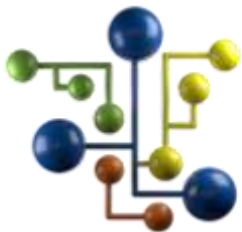


Mllib – Outras Funcionalidades

Funcionalidade (Feature Extraction)	Funções (importadas a partir do pacote mllib.feature)
TF-IDF (Term Frequency – Inverse Document Frequency)	HashingTF() e IDF ()
Escala	StandardScaler()
Normalização	Normalizer()
Word2Vec	Word2Vec()
Estatística	colStats(), corr(), chiSqTest(), mean(), stdev(), sample()

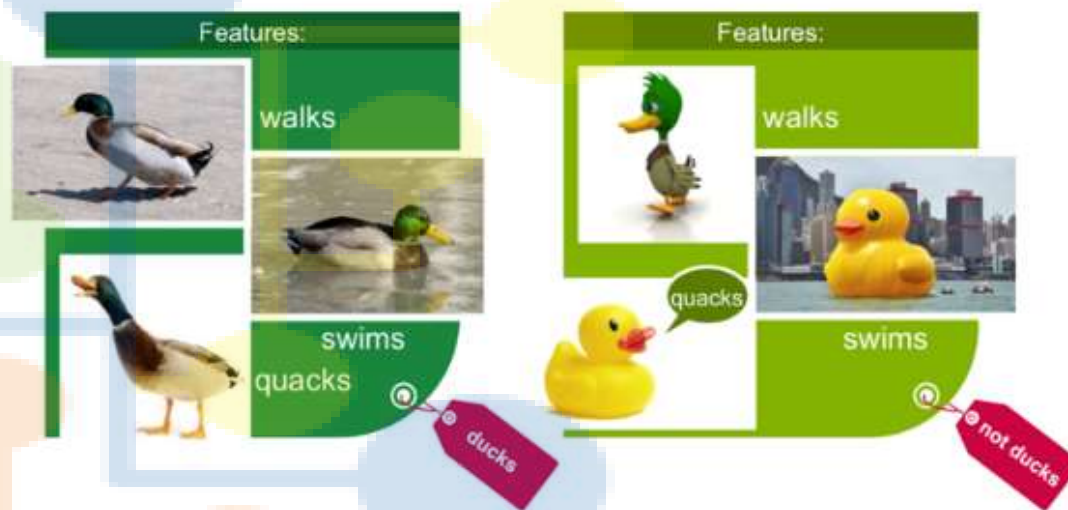






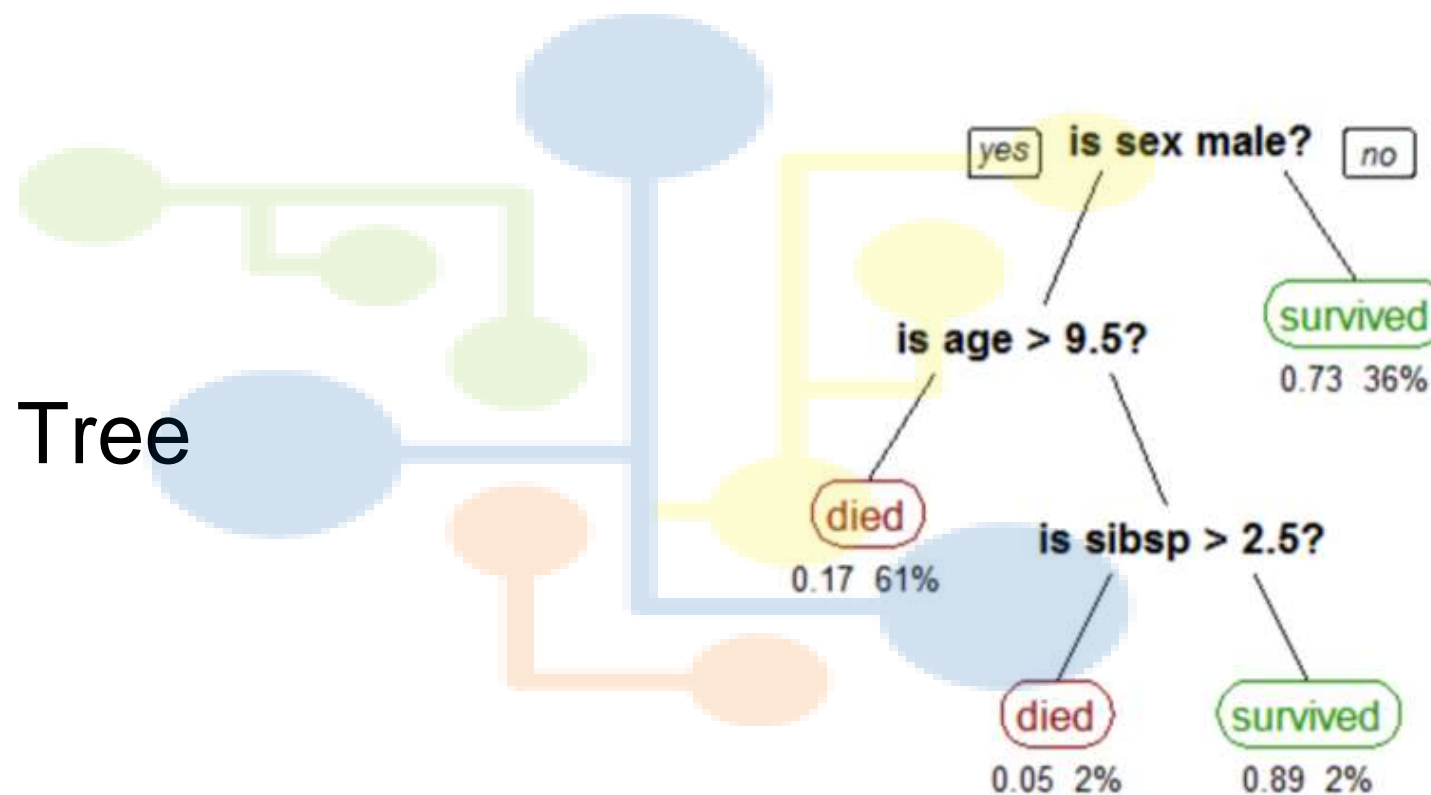
Classificação

If it Walks/Swims/Quacks Like a Duck Then It Must Be a Duck





Decision Tree



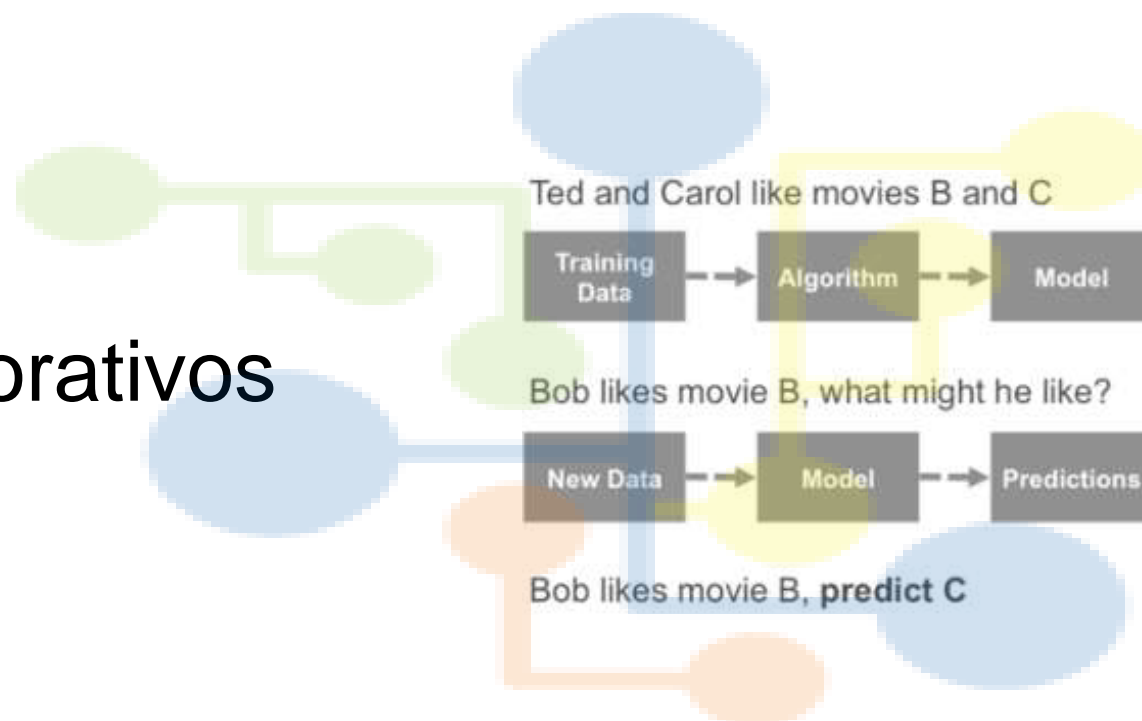


Clustering





Filtros Colaborativos



User Item Rating Matrix

	A	B	C
Ted	4	5	5
Carol		5	5
Bob		5	?



Data Science
Academy

Data Science Academy tiago.soares@contato.com.br 5d143e935e4cde1e638b4567



Tenha uma Excelente Jornada de Aprendizagem.

Muito Obrigado por Participar!

Equipe Data Science Academy