



- Conjunto de métodos e técnicas para aplicação de conceitos matemáticos e estatísticos.
- Uso de modelagem preditiva e aprendizado de máquina.
- Interpretação e extração de conhecimento de grandes volumes de dados.
- Recurso necessário para as empresas no movimento *Data-Driven.*”
- Bigdata + Ciência de Dados = Bigdata Analytics



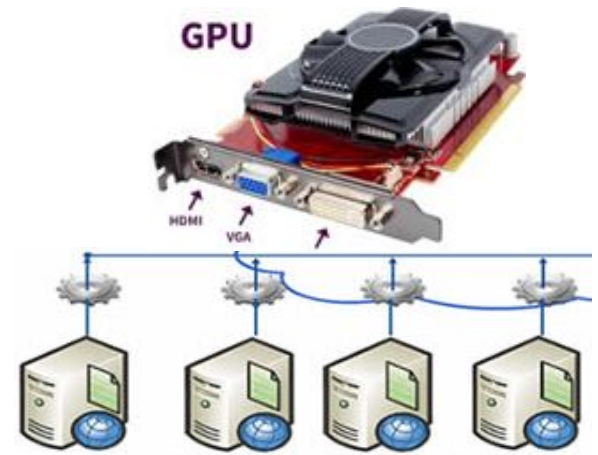
Crescimento
Exponencial na
geração de dados

A tecnologia trouxe a
Ciência de dados para
o centro das atenções.

A área cresce na
mesma proporção com
que os dados são
gerados.



Menor Custo de Storage
(Cloud Computing)

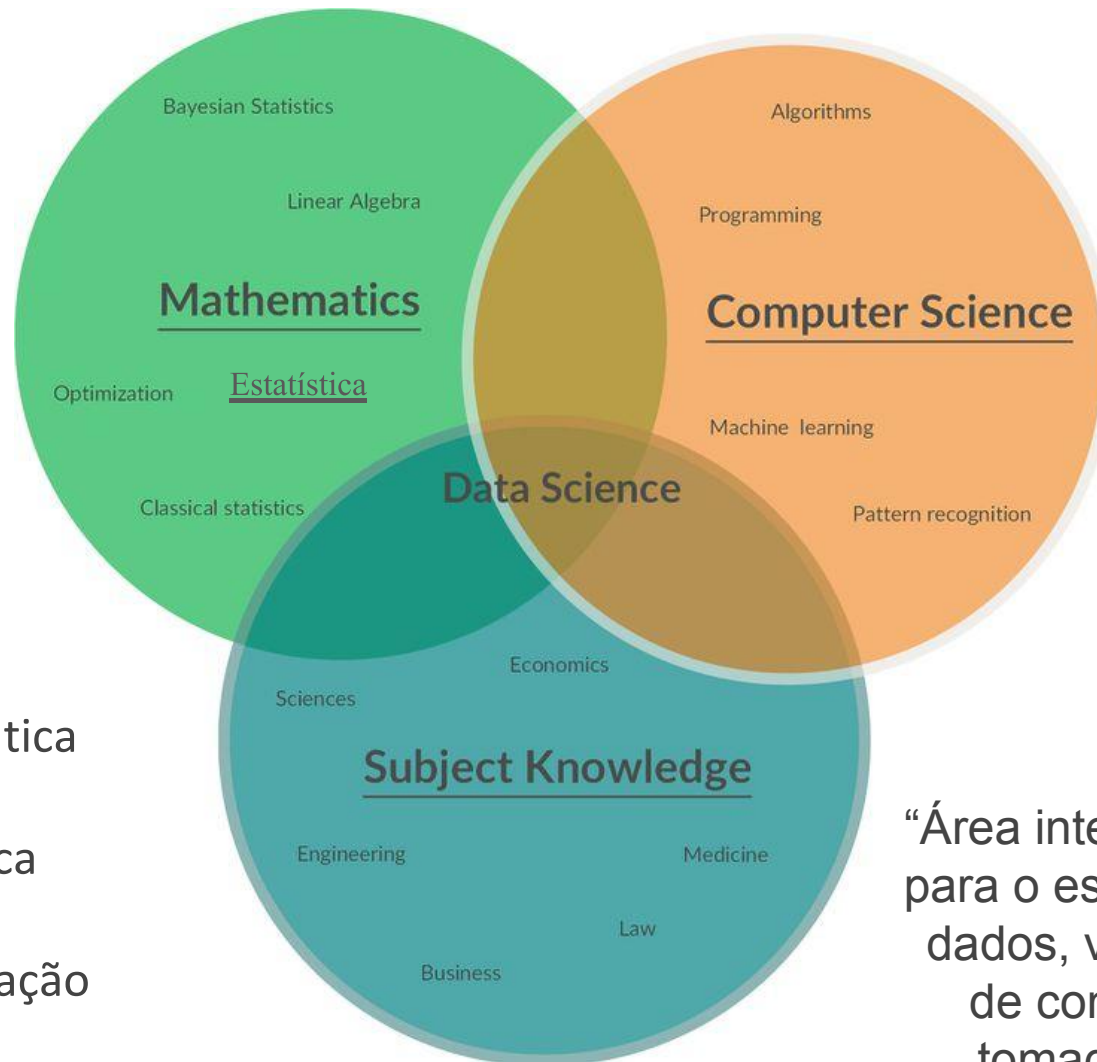


Maior poder de
Processamento

Novos métodos,
tecnologias e processos
para extrair informação.

Maior poder de
processamento, menor
custo de armazenamento
e serviços em Cloud.

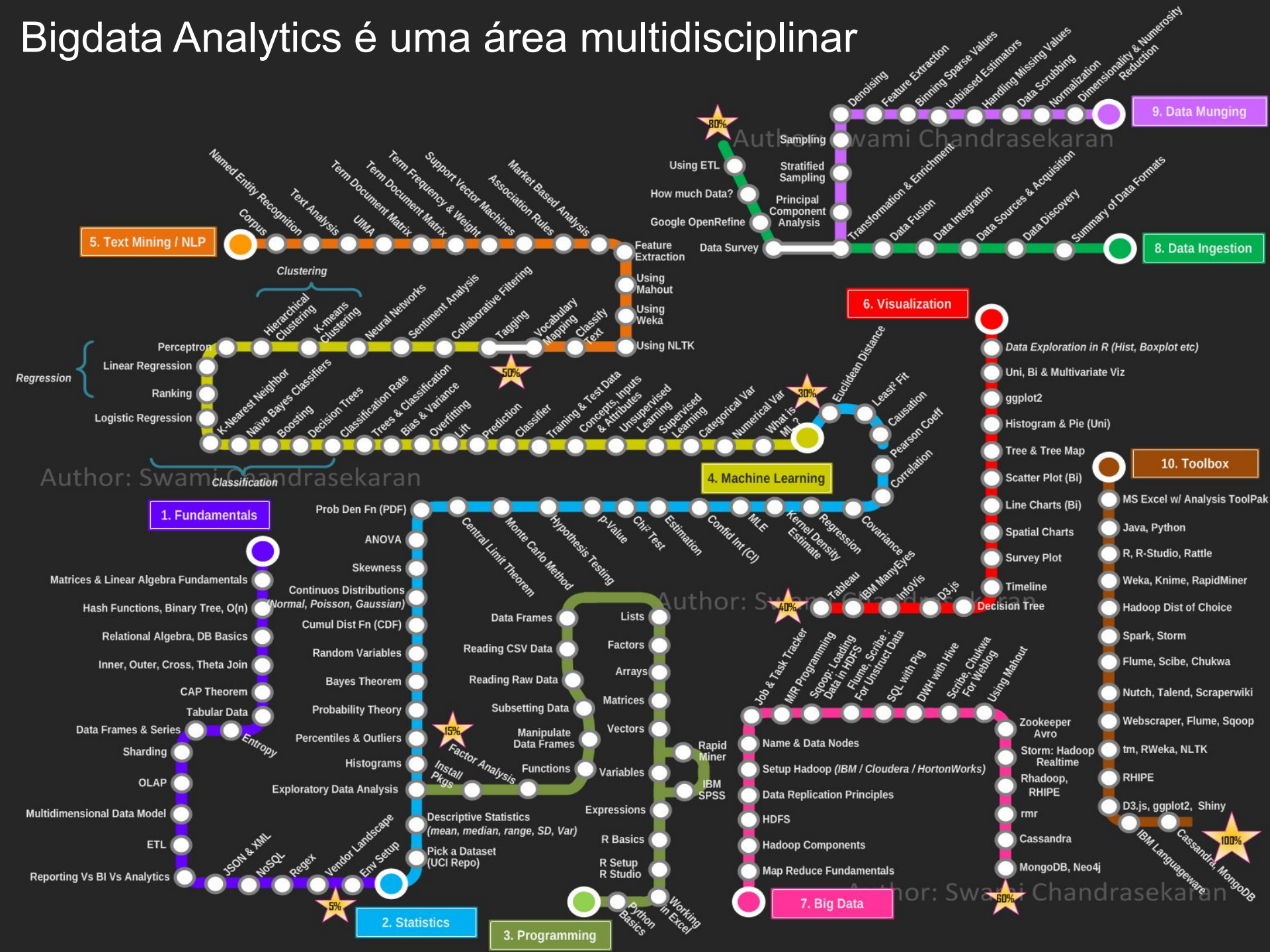
Áreas de Conhecimento em Data Science



- Matemática
- Estatística
- Computação
- Conhecimento do negócio

“Área interdisciplinar voltada para o estudo e a análise de dados, visando a extração de conhecimento para tomadas de decisão.”

Bigdata Analytics é uma área multidisciplinar



Áreas de Conhecimento em Data Science

Áreas de conhecimento	Habilidades em Data Science
Matemática e Estatística	Algebra linear, Testes de Hipótese, Estatística Descritiva, Média, Moda e Mediana, Análise Bayesiana, etc
Linguagem de Programação	Python, Java, Scala, SQL, R, Julia, etc
Base de dados	Banco de dados relacionais e No-SQL
Bigdata	Hadoop e Spark
Machine Learning	Modelagem, classificação, regressão, clusterização, etc
Visualização	D3.js, Tableau, Ggplot2, Matplotlib, etc
Conhecimento de Negócio	Varejo, Indústria, Saúde, Tecnologia, Finanças, etc

Business Intelligence X Data Science

Atuação	Analista de BI	Cientista de Dados
Foco	Tendências, indicadores e KPIs, relatórios e dashboards	Métodos científicos, estatísticas, correlações, modelagem preditiva e uso de algoritmos
Processo	Documental e comparativo	Exploratório, experimental e iterativo
Fonte de Dados	Data warehouses e bancos relacionais	Datalakes, bigdata, bancos No-Sql, bancos relacionais, streaming, arquivos
Qualidade de Dados	Alta, tratados processos de ETLs	Baixa ou média, requer tratamento dependendo da origem coletada. Muitas vezes os dados estão em seu formato bruto
Modelagem	Esquema e modelagem pré-definida	Esquema sobre demanda
Tipo de Análise	Descritiva e diagnóstica O que aconteceu?	Preditiva e Prescritiva O que pode acontecer?

O objetivo da Ciência de dados é converter dados brutos em inteligência de negócio, assim como o BI, porém com aplicação científica, testes de hipóteses, modelagem estatística e aprendizado de máquina.

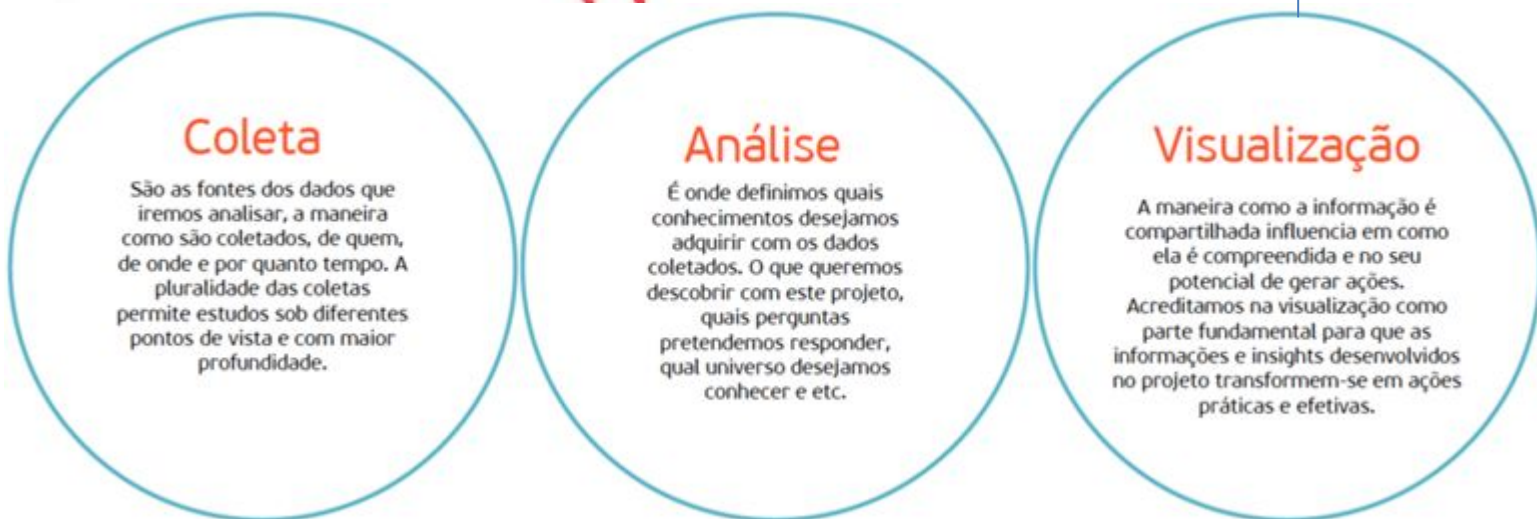
Aplicações da Ciência de Dados

- Recomendação de produtos e serviços
- Classificação de texto e Análise de Sentimento
- Detecção de Terremotos
- Análise de Crédito
- Marketing Personalizado
- Combate ao Crime e ao Terrorismo
- Sistemas de buscas mais eficientes
- Propensão de compras
- Carros automatizados
- Personalização do processo de aprendizagem
- Racionalização de custos
- Monitoramento de sinais vitais, etc

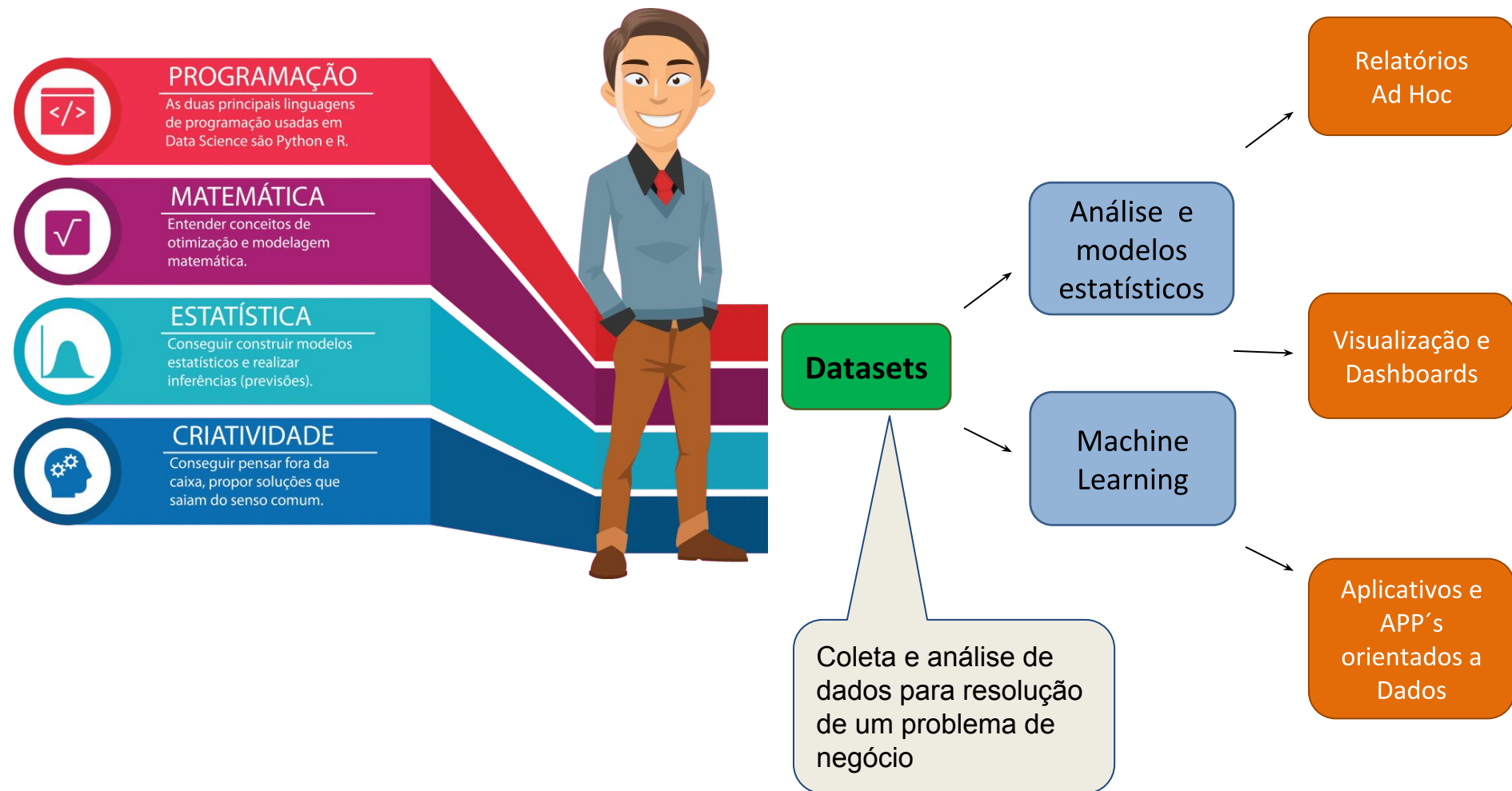



```

graph LR
    A[Dados] --> B[Decisões]
    B --> C[Ações]
  
```



Como trabalha um Cientista de Dados



O que são Datasets?

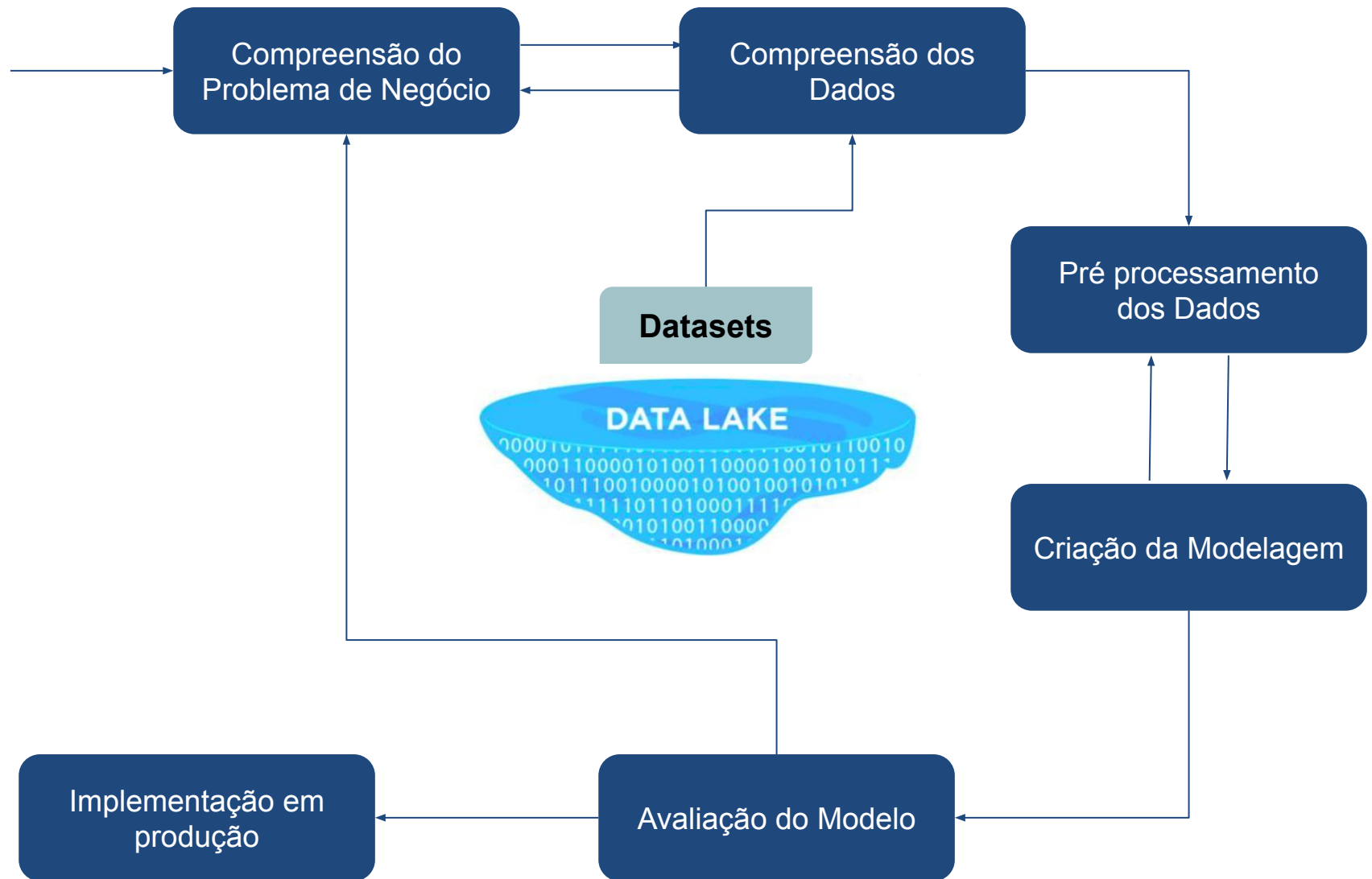
```
import pandas as pd
import numpy as np

# criação de um dataset aleatório
qtde = 20
df_cliente = pd.DataFrame(
    {
        'estado': [['SP', 'RJ', 'MG', 'GO'][i] for i in np.random.randint(0, 4, size=qtde)],
        'idade': np.random.randint(20, 50, size=qtde),
        'renda': [round(h, 2) for h in np.random.uniform(1e3, 9e3, size=qtde)],
        'comprou': [['Sim', 'Não'][i] for i in np.random.randint(0, 2, size=qtde)],
        'promocao': [['Eletronico', 'Eletro', 'Beleza'][i] for i in np.random.randint(0, 3, size=qtde)]
    }
)
df_cliente.head(10)
```

	comprou	estado	idade	promocao	renda
0	Sim	GO	24	Beleza	2205.27
1	Sim	GO	27	Beleza	8137.72
2	Não	RJ	28	Eletro	8470.92
3	Sim	RJ	37	Beleza	6669.39
4	Não	SP	47	Eletronico	4908.34
5	Sim	RJ	49	Beleza	1242.19
6	Não	RJ	21	Beleza	6935.69
7	Não	GO	48	Beleza	8680.26
8	Não	GO	41	Beleza	1120.37
9	Sim	MG	47	Eletro	8194.13

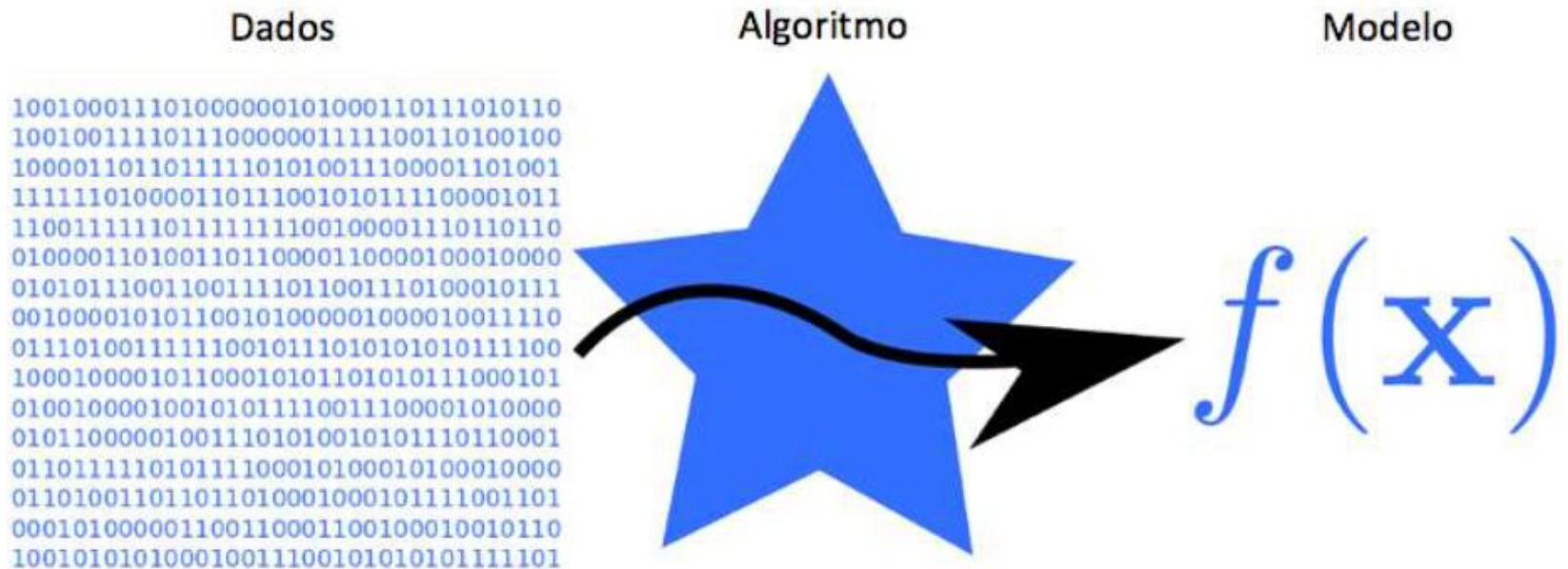
- ★ **Colunas = Variáveis.** São características de um evento.
- ★ **Linhas = Eventos.** São ocorrências ou acontecimentos.
- ★ **Tabela = Dataset.** É um conjunto de evento.

Etapas da Criação de um Modelo Preditivo



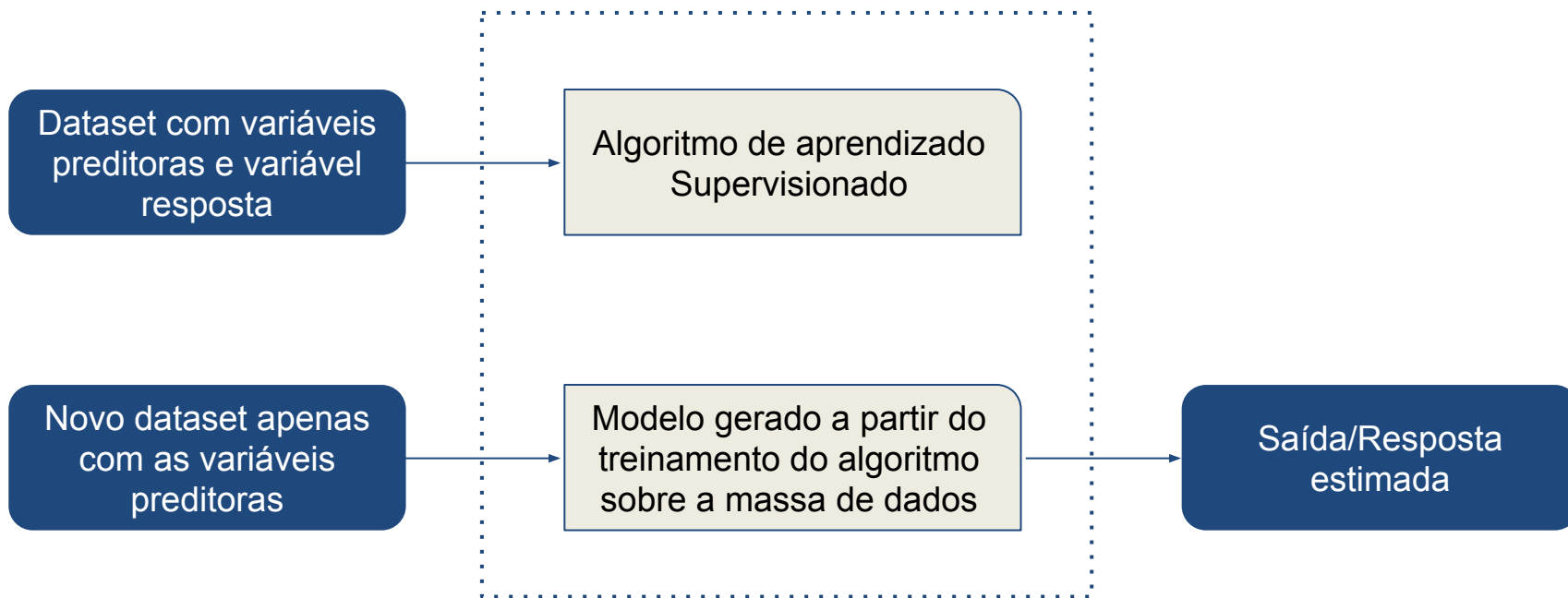
O que é um Modelo Preditivo?

De forma simplificada, consiste em uma função matemática que melhor representa a relação e padrões existente nos dados. O modelo aprende através do treinamento com uma massa de dados de amostragem. Então, o modelo passa a fazer previsões com novos dados, o que chamamos de generalização.



Modelo Preditivo

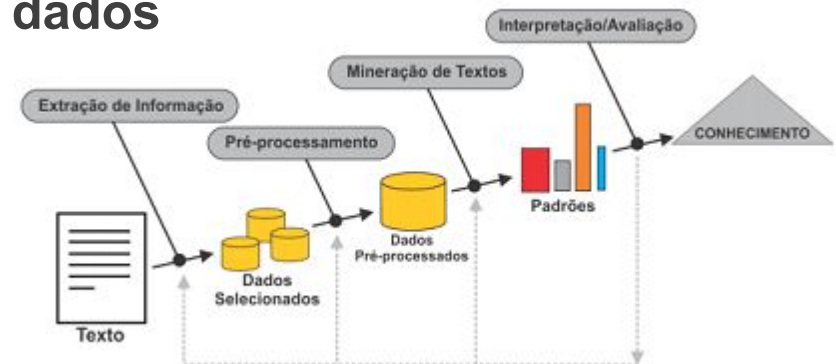
De um aprendizado Supervisionado



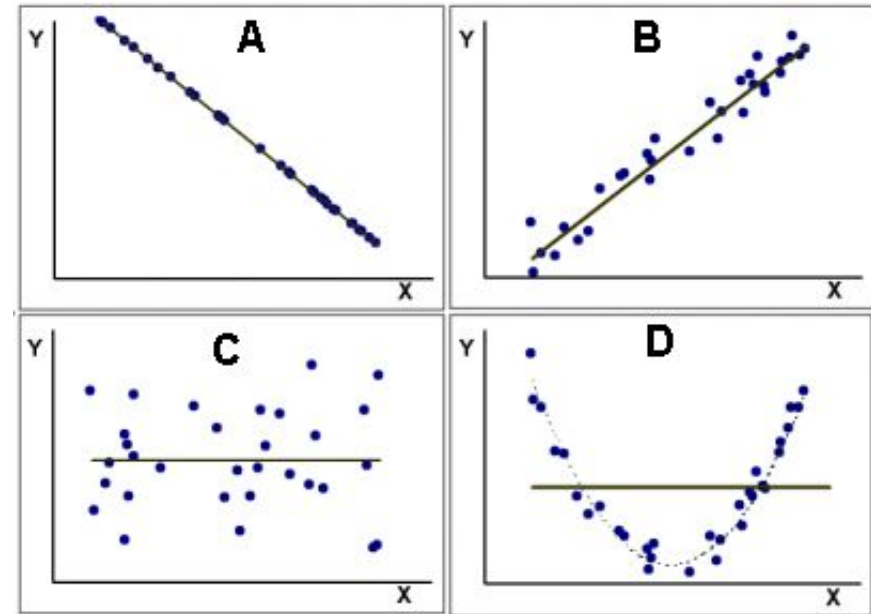
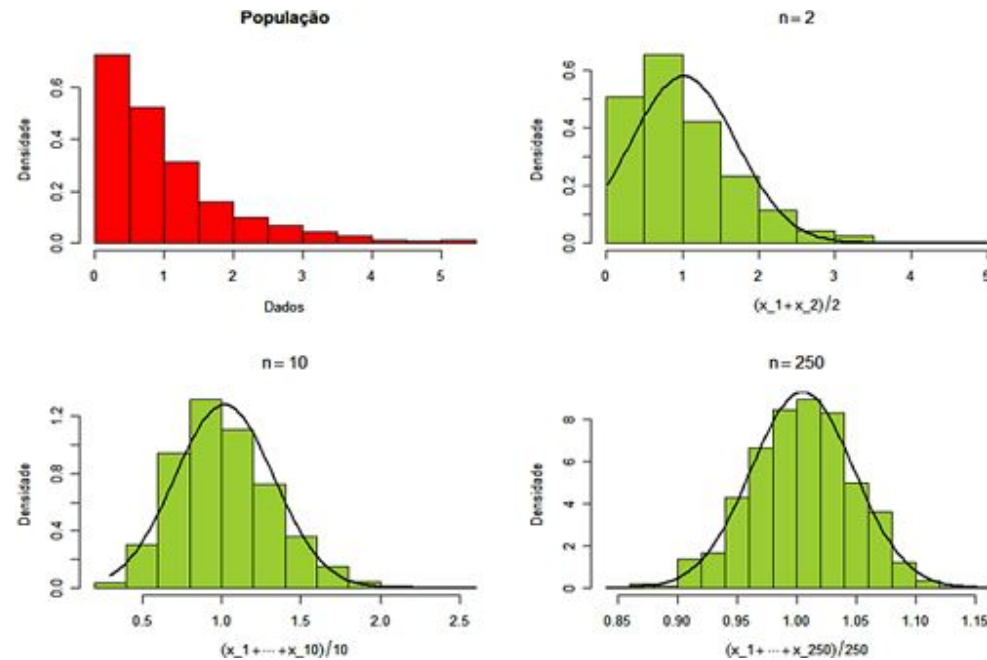
“Um modelo preditivo de Machine Learning é usado para resolver um problema específico. Cada problema é único, assim como a criação do modelo”

Análise Exploratória e Pré-Processamento nos Dados

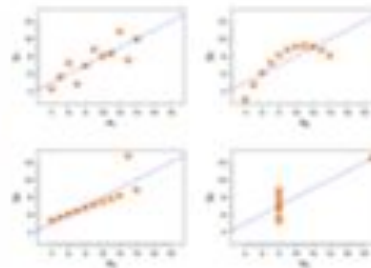
- **Feature Selection** - Seleção das variáveis preditoras mais relevantes. Diminui a dimensionalidade, melhora de performance, menor tempo de treinamento.
- **Correlação** - Verificar correlação entre as variáveis. E filtrar variáveis que representam a mesma informação.
- **Qualidade dos dados** - Dados faltantes e Outliers.
- **Distribuição e Amostragem dos dados** – Inferência estatística. Análise da média, mediana e moda.
- **Normalização e Padronização dos dados**
- **Redução de Dimensionalidade**
- **Agregações dos dados**



Análise Exploratória e Visualização dos Dados



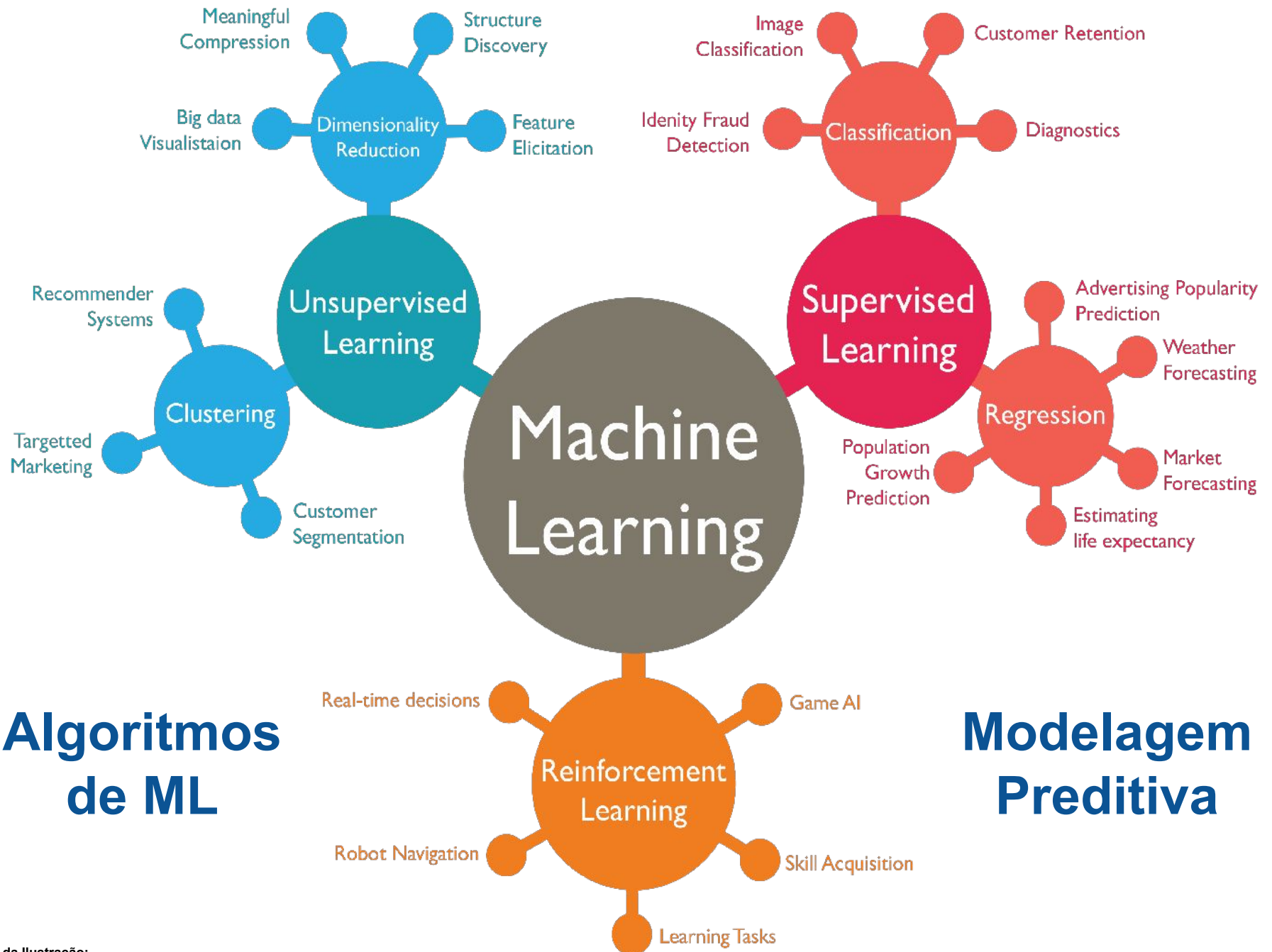
	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	2	1	4	3	6	5	8	7	9	10
3	3	4	1	2	5	6	7	8	9	10
4	4	3	2	1	6	5	7	8	9	10
5	5	6	5	6	1	2	3	4	9	10
6	6	5	6	5	2	1	3	4	8	10
7	7	8	7	8	3	4	1	2	9	10
8	8	7	8	7	4	3	2	1	9	10
9	9	9	9	9	9	9	9	9	1	10
10	10	10	10	10	10	10	10	10	10	1



Dado

Visualização

Informação



Aprendizado Supervisionado

Sua principal característica é que os dados que utilizamos para treiná-los contém a resposta desejada, isto é, o treinamento ocorre através das variáveis independentes, também conhecida como preditoras, juntamente da variável dependente, também conhecida como target ou resposta.

Entre as técnicas mais conhecidas para resolver problemas de aprendizado supervisionado estão:

- regressão linear,
- regressão logística,
- redes neurais artificiais,
- SVM - máquina de suporte vetorial (ou máquinas kernel)
- árvores de decisão,
- k-vizinhos mais próximos
- Bayes ingênuo

Aprendizado Não-Supervisionado

Muitas vezes não temos padrões conhecidos nos dados para tentar prever algo. É necessário antes descobrir algo mais, uma representação mais informativa dos dados que temos. Fazer uma mineração dos dados para encontrar padrões ocultos.

Dentre as técnicas mais conhecidas estão:

- Clusterização k-médias,
- PCA Análise de componentes principais
(Redução de dimensionalidade)
- redes neurais artificiais,
- Máquina de suporte vetorial (ou máquinas kernel)
- Clusterização Hierárquica
- Word2vec, entre outras

Aprendizado Por Reforço

No aprendizado por reforço, o algoritmo tenta aprender qual é a melhor ação a ser tomada, dependendo das circunstâncias na qual essa ação será executada.

O sistema de inteligência artificial enfrenta uma situação, e utiliza tentativa e erro para encontrar uma solução para o problema. Para isso, o sistema recebe recompensas ou penalidades pelas ações que executa. Seu objetivo é maximizar a recompensa total.

Essa técnica é muito usada em Games e Robótica, e vem obtendo resultados cada vez melhores.

Que algoritmo de Machine Learning utilizar?

A resposta é “depende”. Tudo gira em torno do problema de negócio que se quer resolver. Cada caso é um caso, e o trabalho do cientista de dados é iterativo e exploratório. E a vários fatores a se considerar como mostrado acima.



Precisão



Tempo de Treinamento



Número de Recursos



Tipo de Problema a ser resolvido

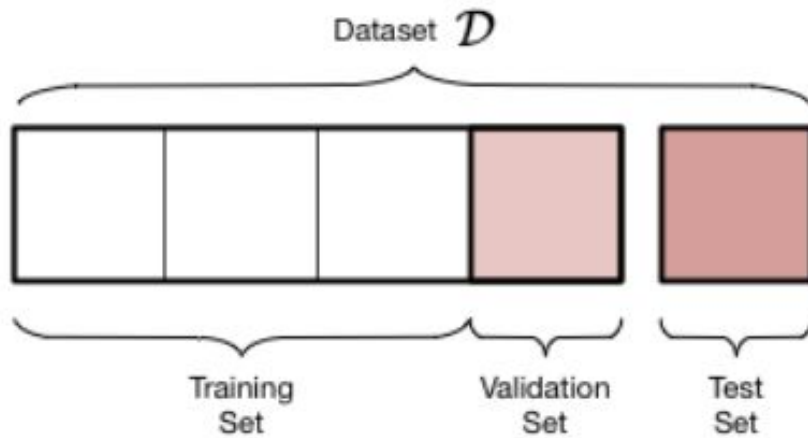


Linearidade

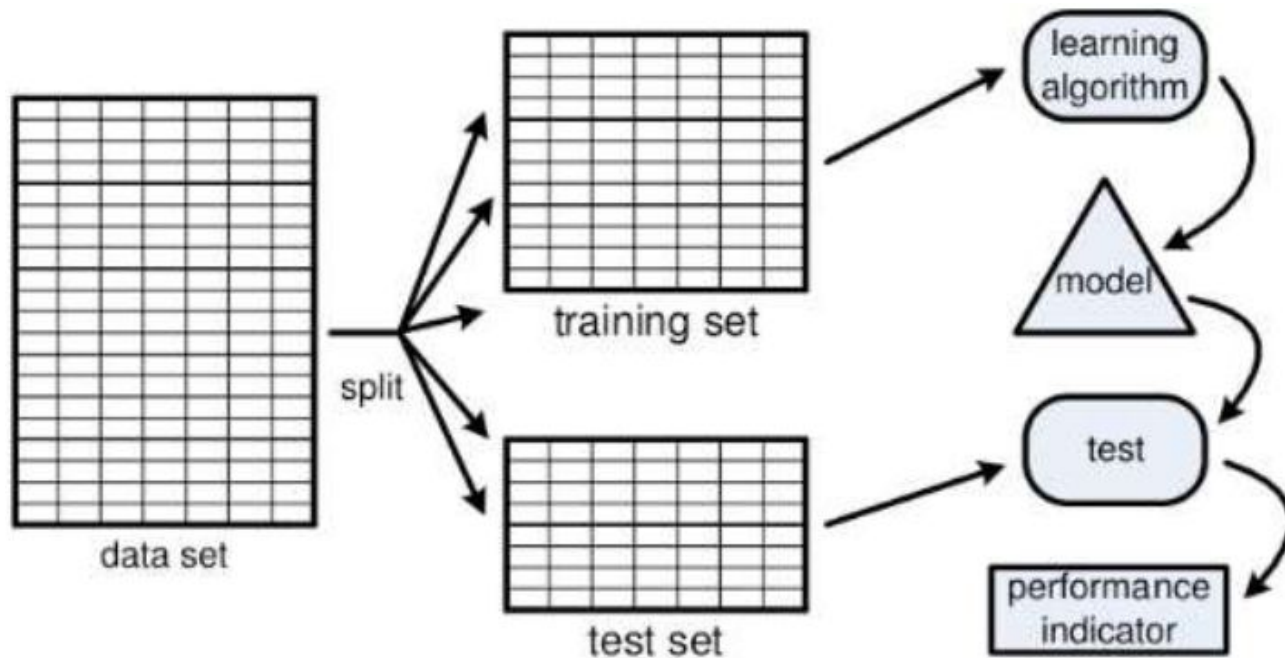


Número de Parâmetros

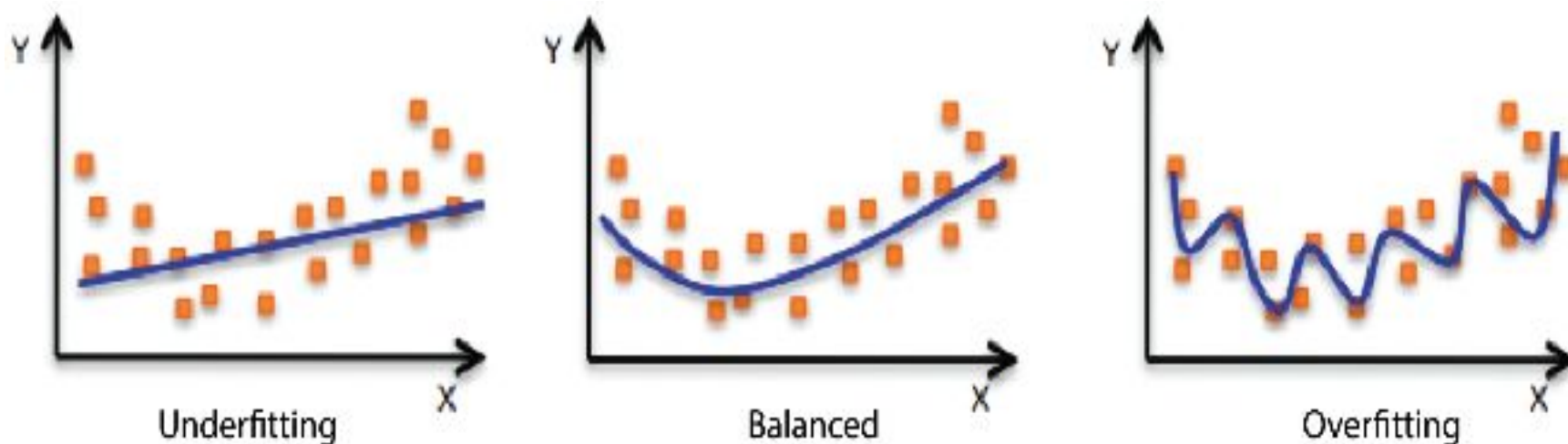
Treinamento e Avaliação do Modelo



75 a 70% - dados de treino
20% - dados de validação
10% - dados de teste



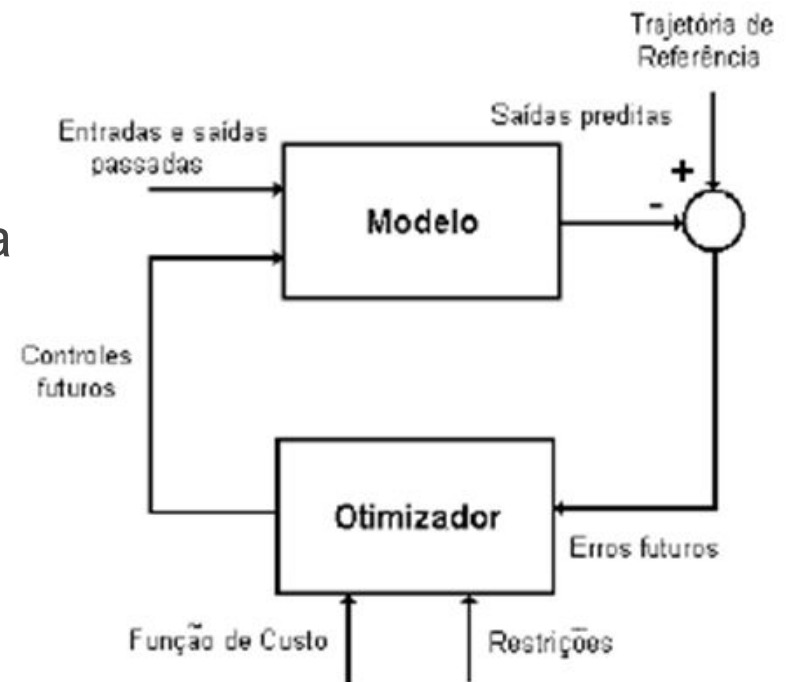
Problemas de um Modelo Preditivo



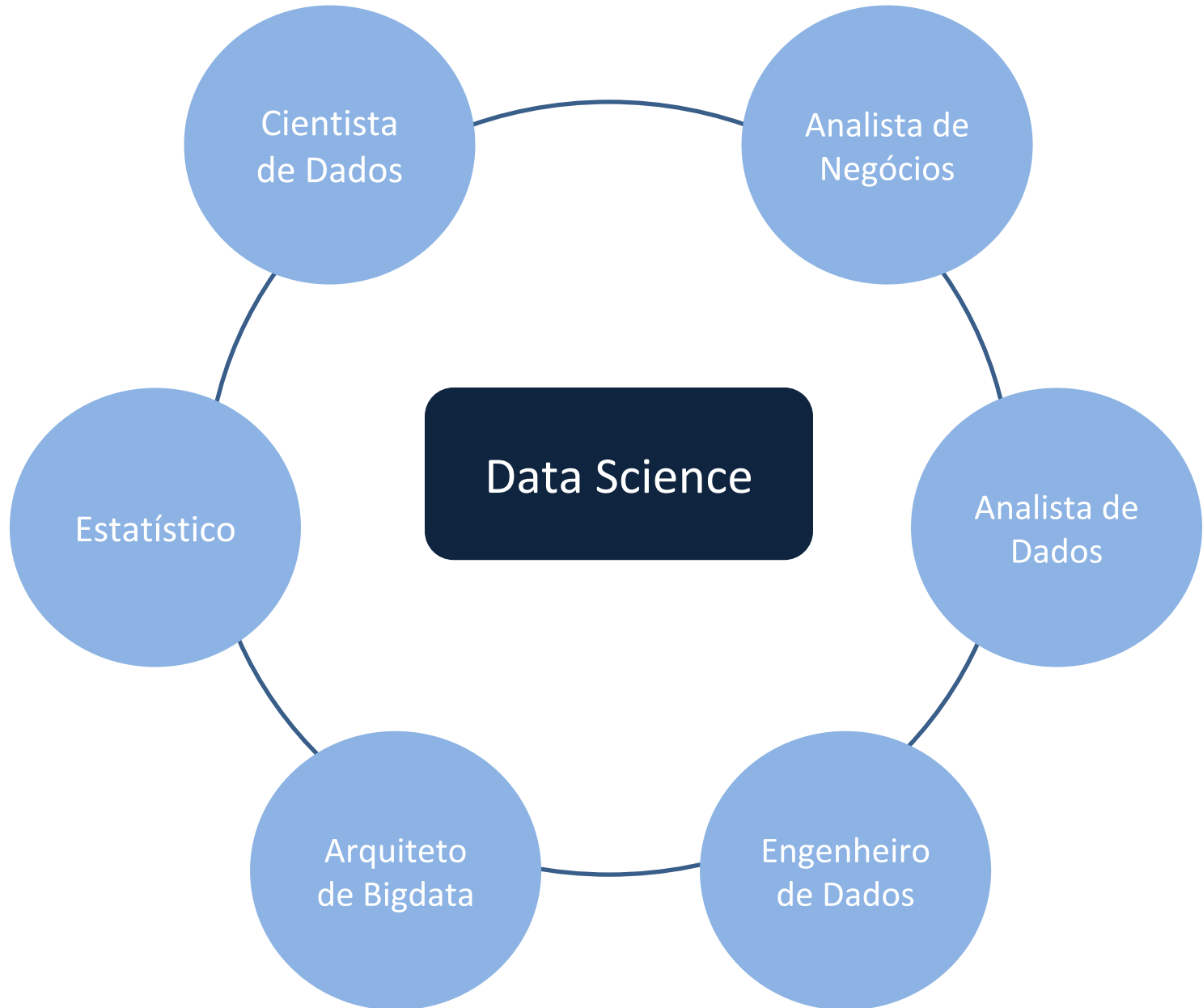
“O principal objetivo de um Modelo Preditivo é a generalização com novos dados, isto é, a capacidade de acerto com novos dados”

Melhoria e Otimização do Modelo Preditivo

- Aumentar a quantidade dados e exemplos;
- Melhorar a distribuição da amostra, etc;
- Melhorar o pré-processamento dos dados;
- Rever a abordagem nos dados para tratamento do problema de negócio;
- Testar novos algoritmos e parâmetros
- Utilização de Métodos e Algoritmos de otimização para identificar o ponto que gera melhor resultado, minimizando uma função de custo para melhor o resultado final do modelo.



Carreiras em Big Data e Data Science



Tecnologias



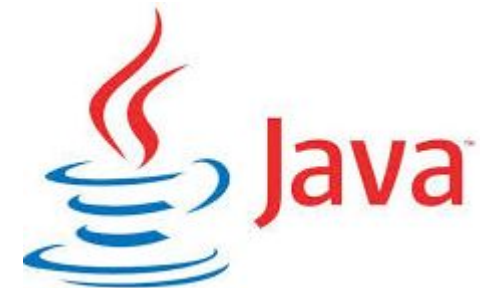
ANACONDA®



python™

Spark

MLlib



julia





Criando um modelo preditivo