Uni-FACEF
Centro Universitário Municipal de Franca

PÓS GRADUAÇÃO Uni-FACEF

Especialização Desenvolvimento de Aplicações Web e Móveis Escaláveis

Turma 2021-2022

Big Data com Python

André Morais

andre.morais@luizalabs.com

09/2022

#83642646

Uni-FACEF
Centro Universitário Municipal de Franca

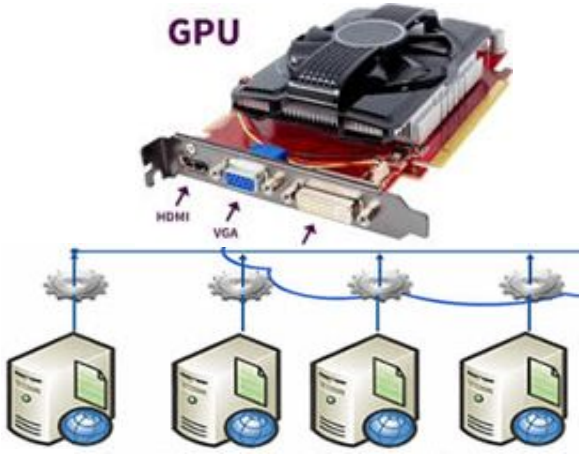
PÓS GRADUAÇÃO

Especialização Desenvolvimento de Aplicações Web e Móveis Escaláveis
Módulo: BIG DATA COM PYTHON - Professor: André Morais
Turma 2021-2022

Conceitos Introdução à Ciência de Dados

- 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*."
- Big Data + Ciência de Dados = Big Data Analytics

Crescimento Exponencial na geração de dados



Maior poder de Processamento

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

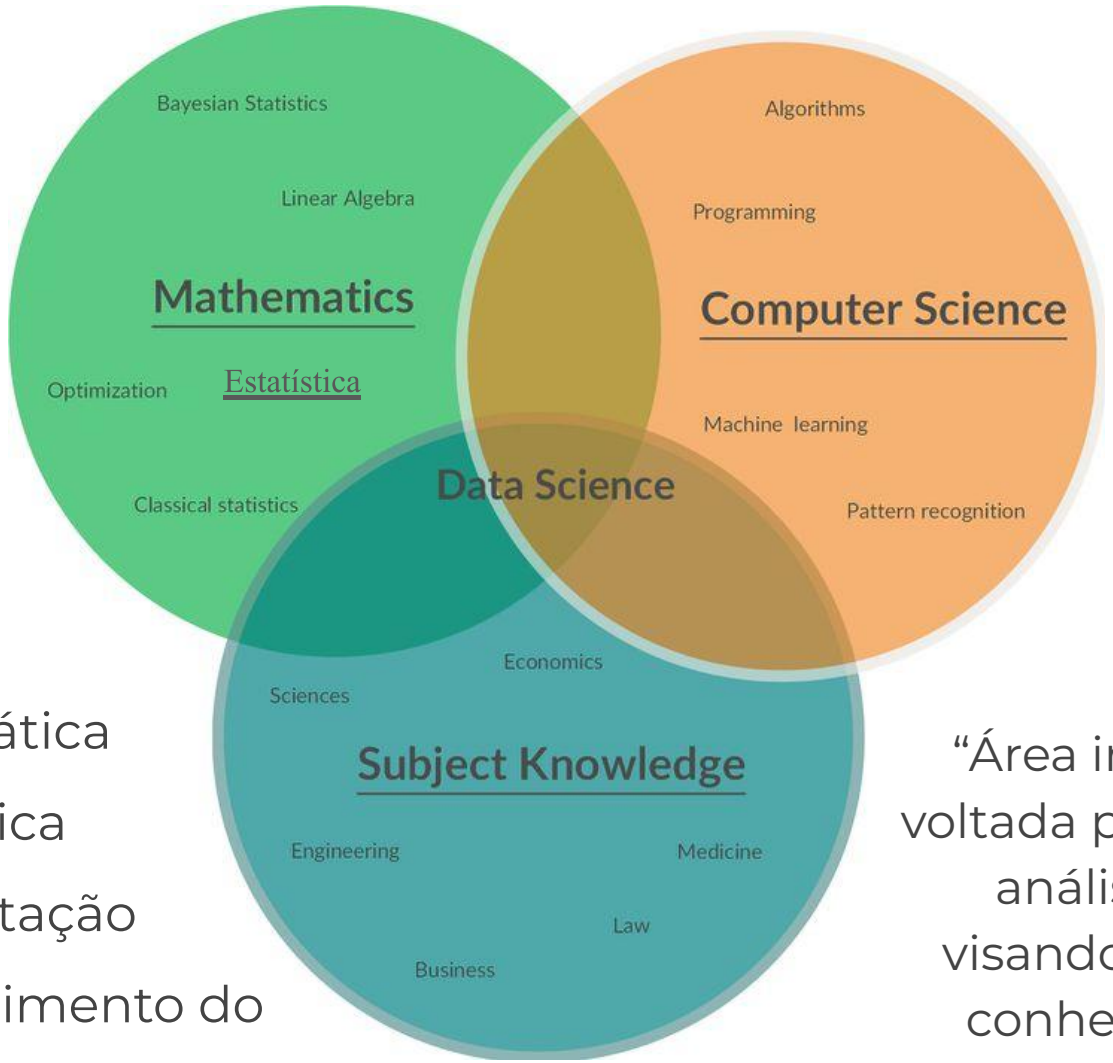
Novos métodos, tecnologias e processos para extrair informação.

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

Menor Custo de Storage (Cloud Computing)

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

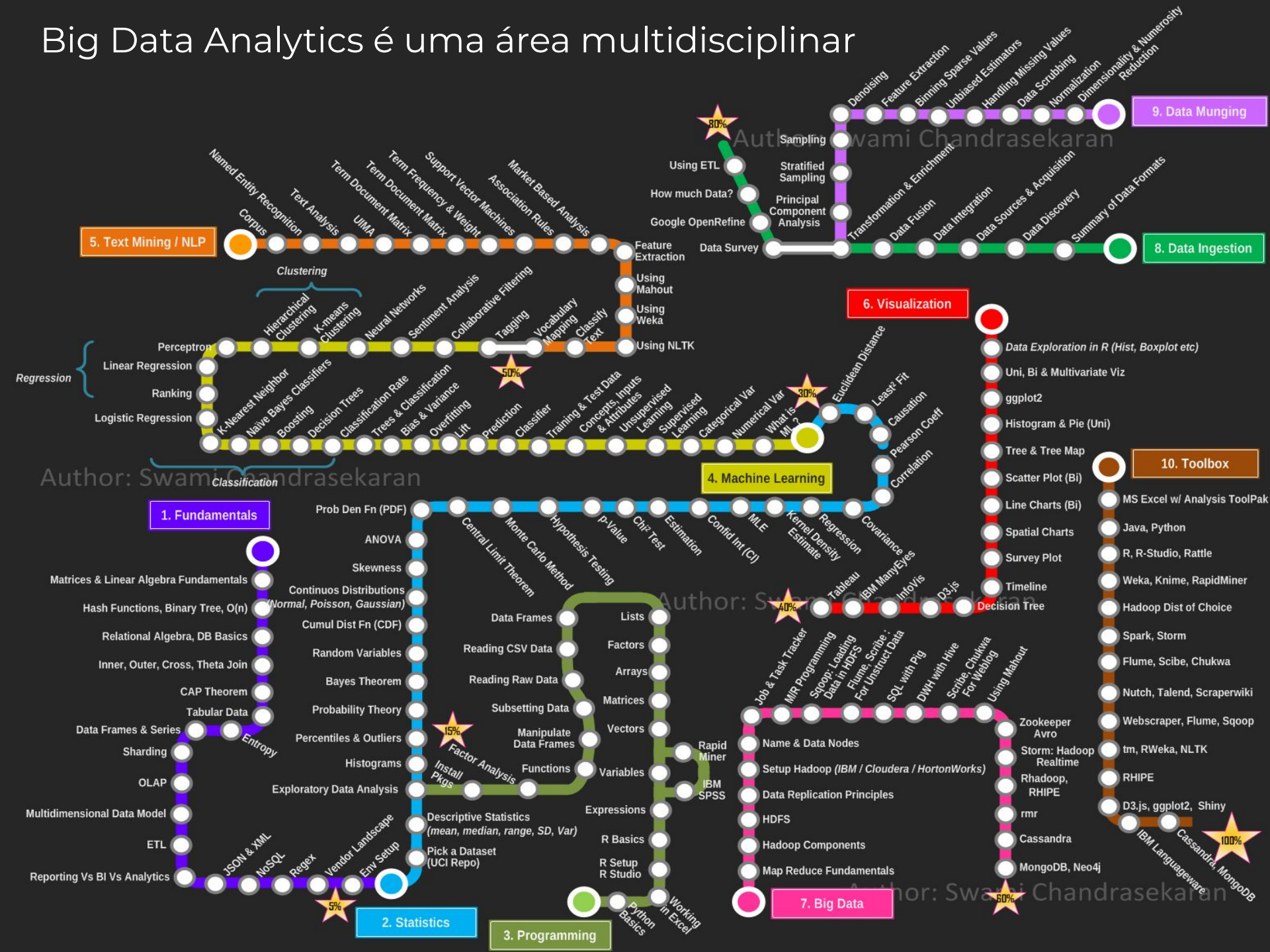
Conceitos Áreas de conhecimento em Data Science



- 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."

Big Data Analytics é uma área multidisciplinar



Conceitos 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, Industria, Saúde, Tecnologia, Finanças, etc

Conceitos Business Intelligence X Data Science

Atuação	Analista de BI	Cientista de Dados
Foco	Tendências, indicadores e KPIs, relatórios e dashboards	Métodos científicos, estatíticas, 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.

Conceitos 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

Ciclo da Análise de Dados

Mídias Sociais

Procedimentos Médicos

Notícias e Jornais

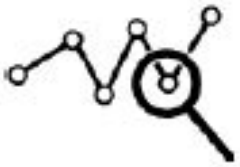
Coleta

Coleta

Imagens por satélite

Dados → Decisões → Ações

Correlação



Análise

Insights

E-commerce

TV & Video

Sensores e Sistemas de Monitoramento

Coleta

São as fontes dos dados que iremos analisar, a maneira como são coletados, de quem, de onde e por quanto tempo. A pluralidade das coletas permite estudos sob diferentes pontos de vista e com maior profundidade.

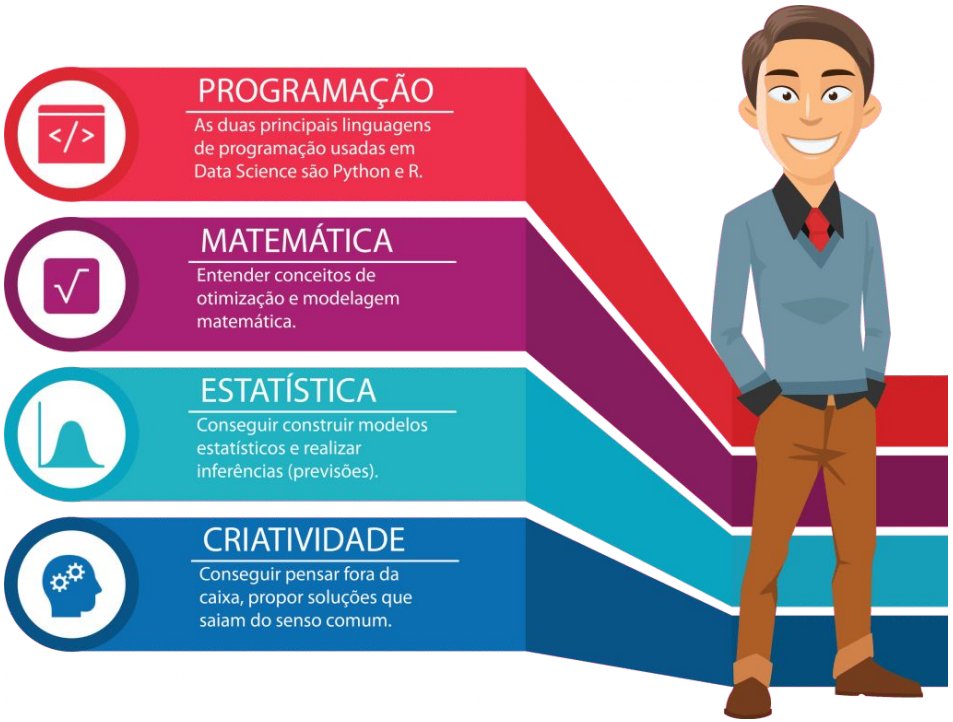
Análise

É onde definimos quais conhecimentos desejamos adquirir com os dados coletados. O que queremos descobrir com este projeto, quais perguntas pretendemos responder, qual universo desejamos conhecer e etc.

Visualização

A maneira como a informação é compartilhada influencia em como ela é compreendida e no seu potencial de gerar ações. Acreditamos na visualização como parte fundamental para que as informações e insights desenvolvidos no projeto transformem-se em ações práticas e efetivas.

Conceitos Como trabalha um Cientista de Dados



Datasets

Análise e modelos estatísticos

Machine Learning

Relatórios Ad Hoc

Visualização e Dashboards

Aplicativos e APP´s orientados a Dados

Coleta e análise de dados para resolução de um problema de negócio

Fonte da Ilustração:
https://sigmoidal.ai/como-se-tornar-um-cientista-de-dados/

Conceitos O que são Datasets?

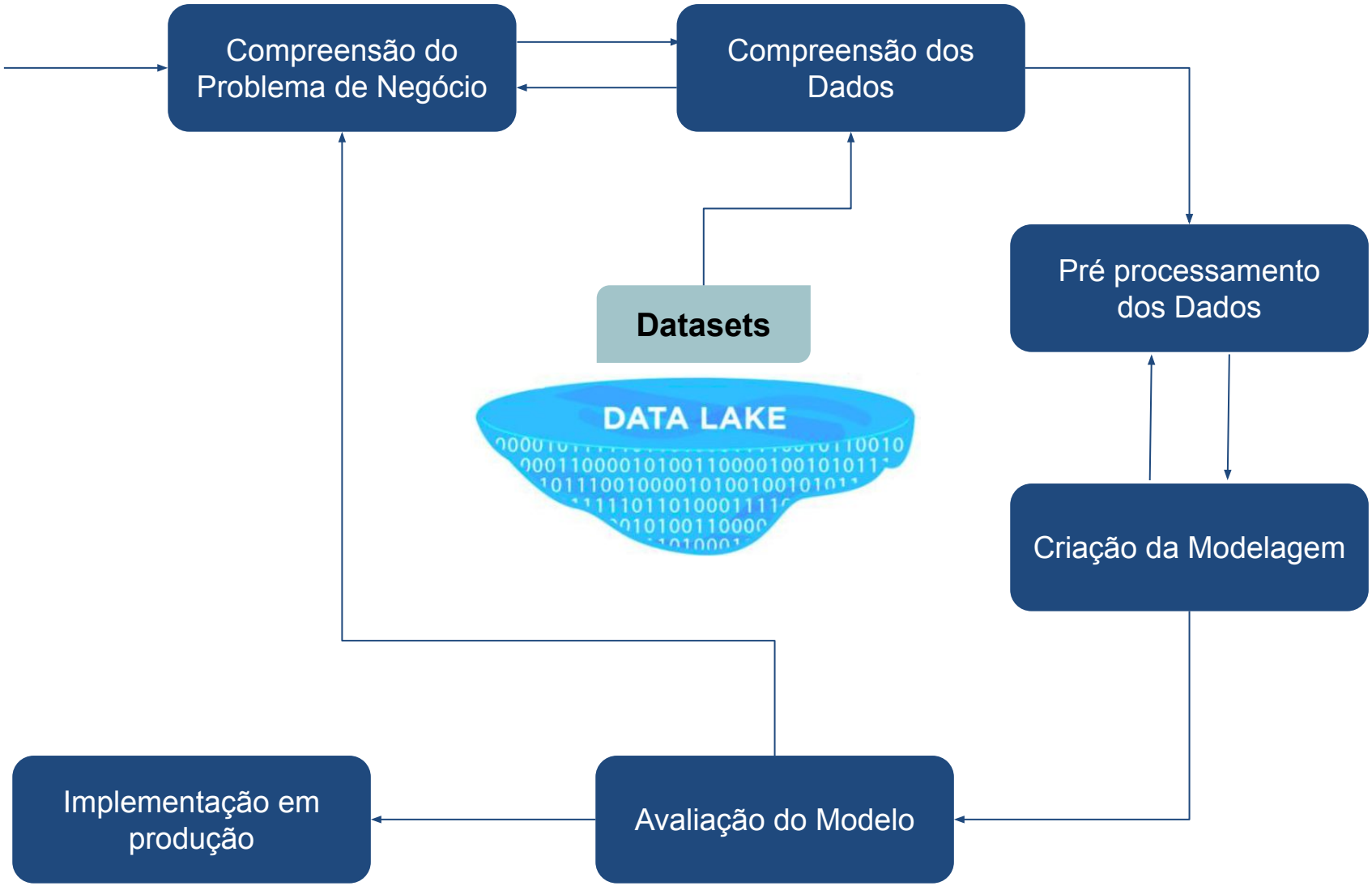
```python
import pandas as pd
import numpy as np

# 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

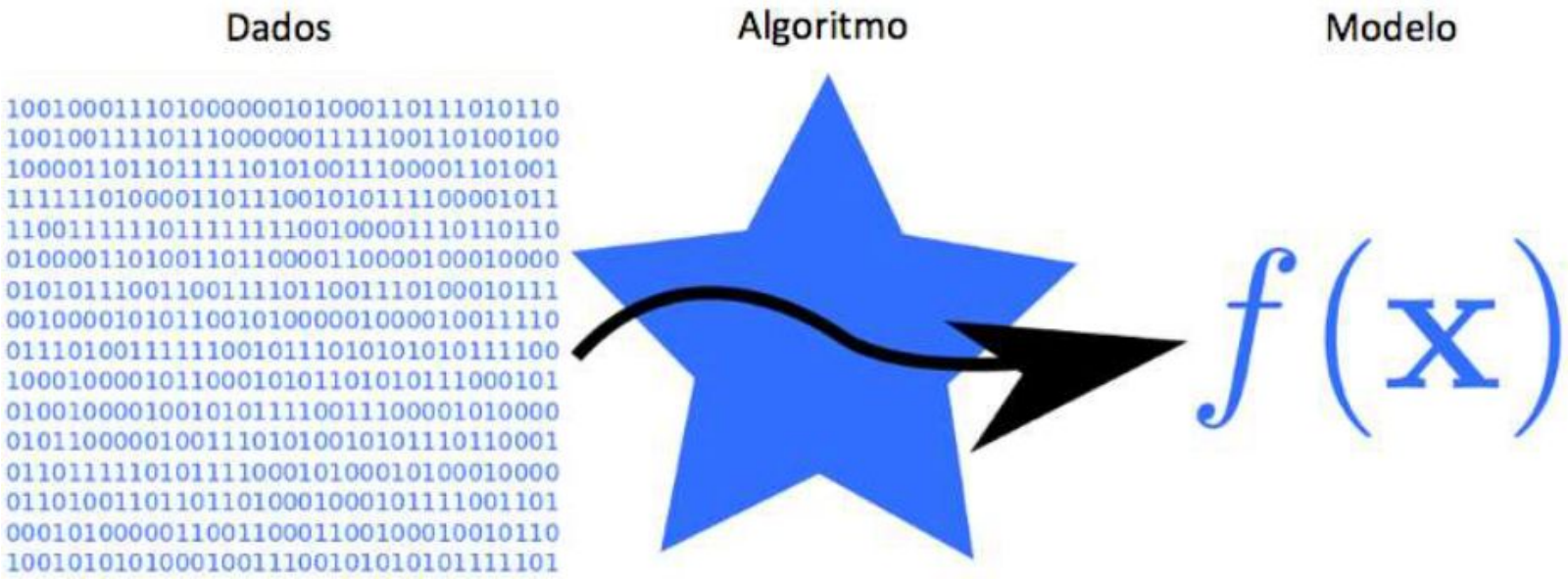
- **Colunas = Variáveis**. São características de um evento.
- **Linhas = Eventos**. São ocorrências ou acontecimentos.
- **Tabela = Dataset**. É um conjunto de eventos.

Conceitos Etapas da Criação de um Modelo Preditivo



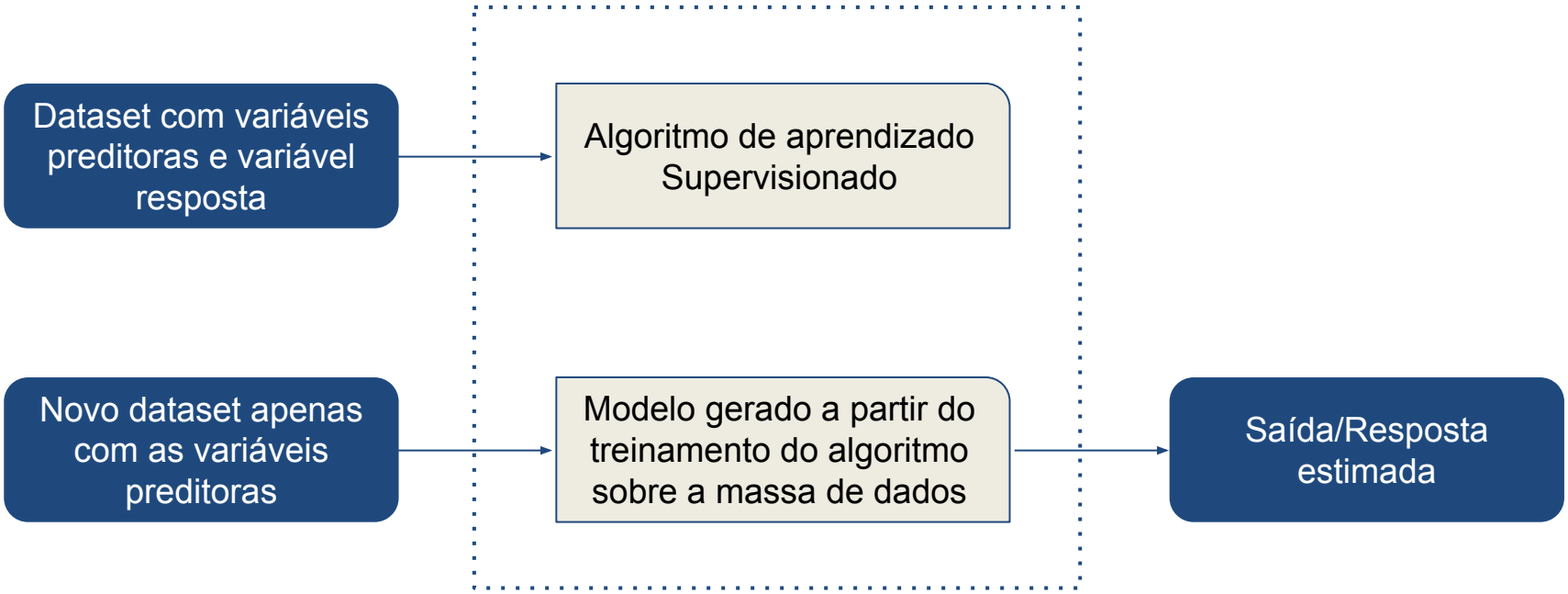
Conceitos O que é um Modelo Preditivo

De forma simplificada, consiste em uma função matemática que melhor representa a relação e padrões existentes 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.



Conceitos 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"

Conceitos O Análise Exploratória e Pré-Processamento nos Dados

- **Feature Selection** - Seleção das variáveis preditoras mais relevantes. Diminui a dimensionalidade, melhoria de performance, menor tempo de treinamento.

- **Correlação** - Verificar correlação entre as variáveis. E filtrar variáveis que representam a mesma informação.

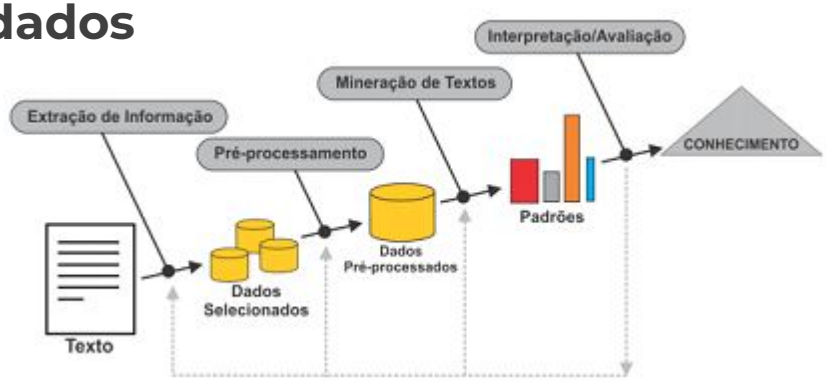
- **Qualidade dos dados** - Dados faltantes e Outliers.

- **Distribuição e Amostragem dos dados** – Inferência estatística. Análise da média, mediana e moda.

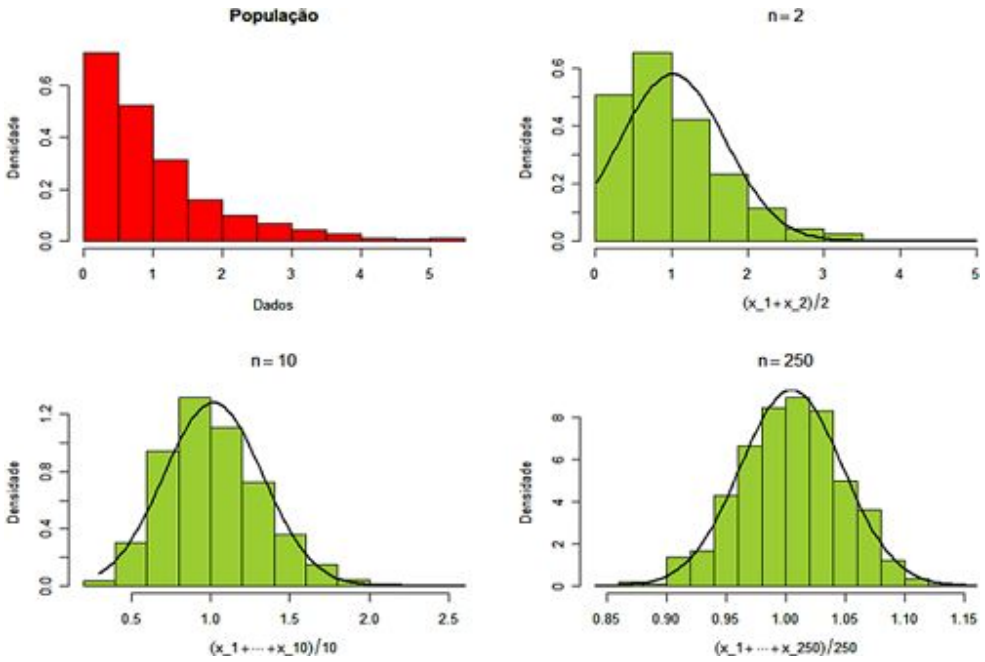
- **Normalização e Padronização dos dados**

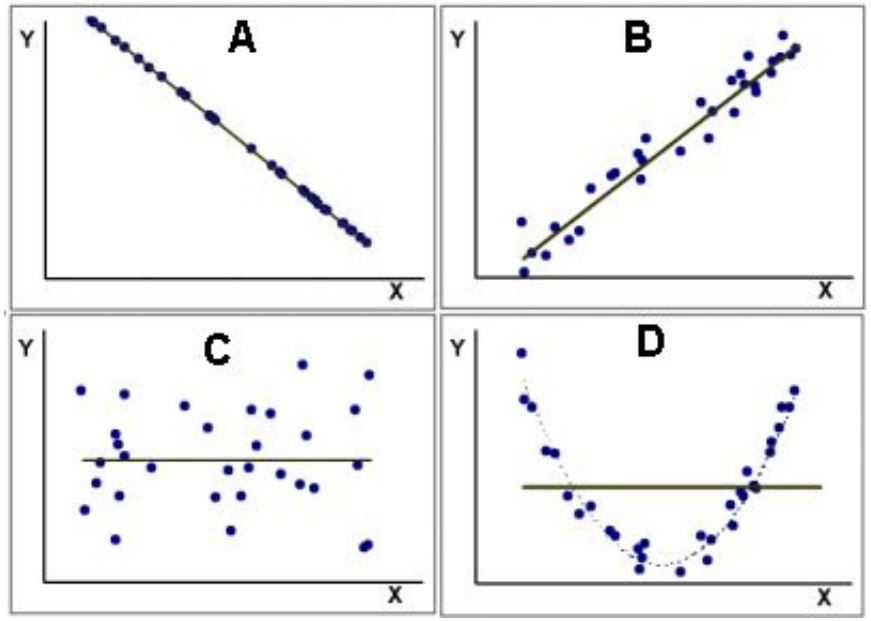
- **Redução de Dimensionalidade**

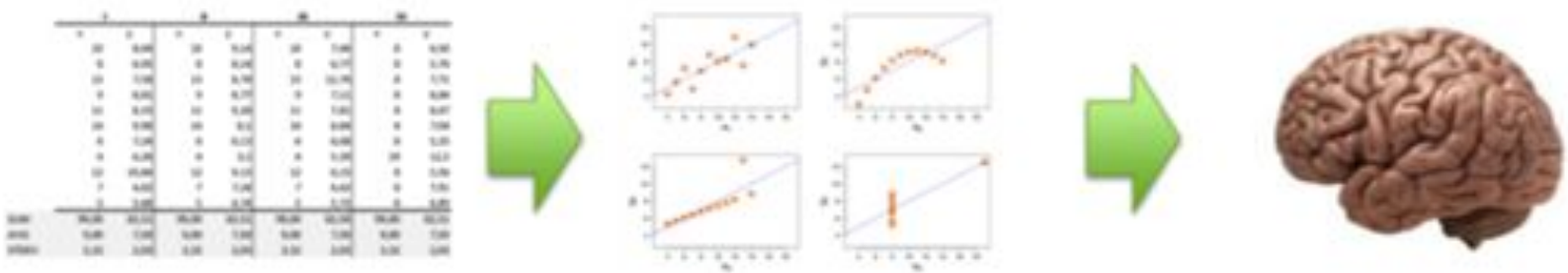
- **Agregações dos dados**



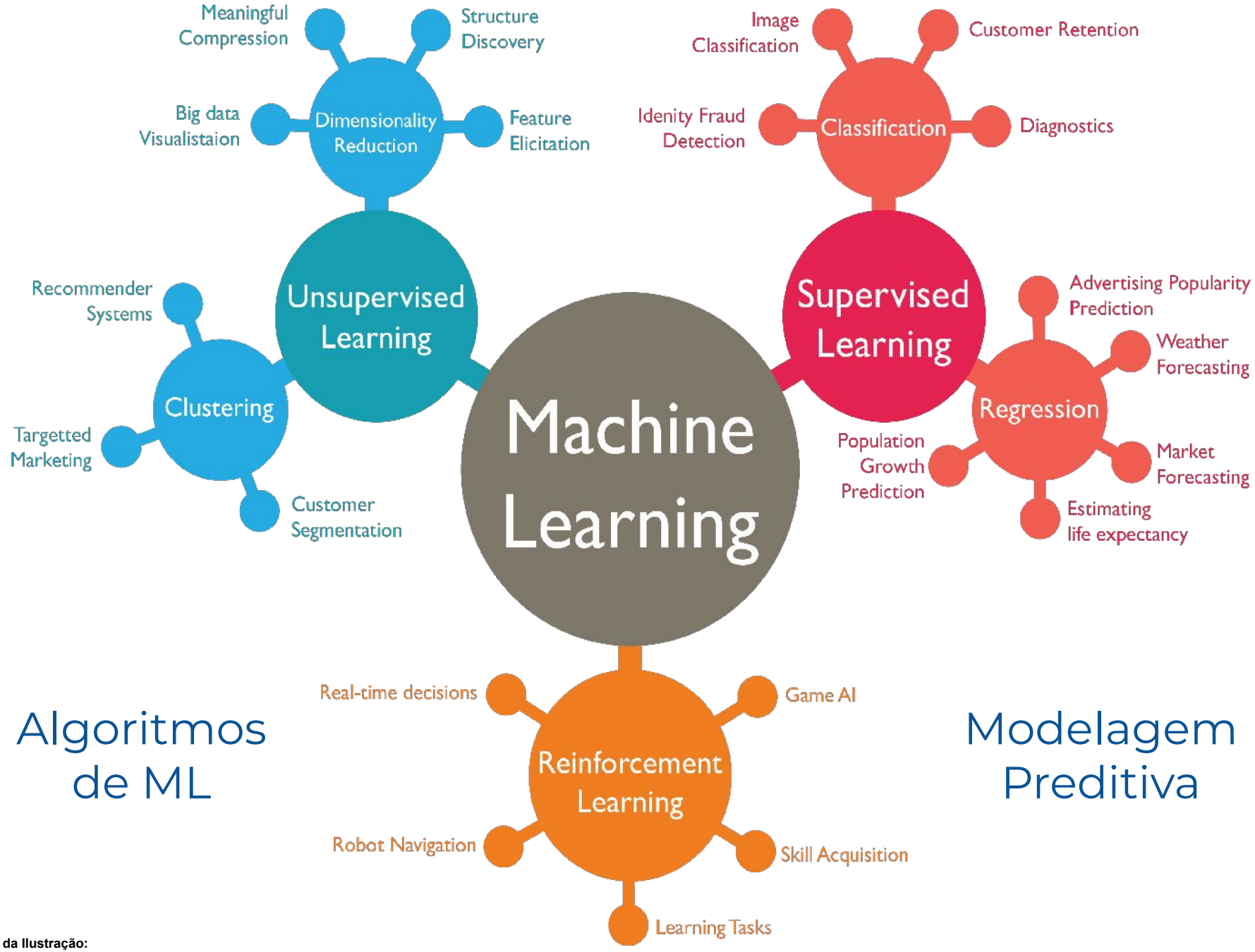
Conceitos O Análise Exploratória e Visualização dos Dados







Dado Visualização Informação



Algoritmos de ML

Modelagem Preditiva

Fonte da Ilustração:
https://www.houseofbots.com/news-detail/3736-4-the-biggest-issue-with-machine-learning-algorithms

Conceitos 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 conhecidas 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

Conceitos Aprendizado Não-Supervisionado

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

Conceitos Aprendizado Por Reforço

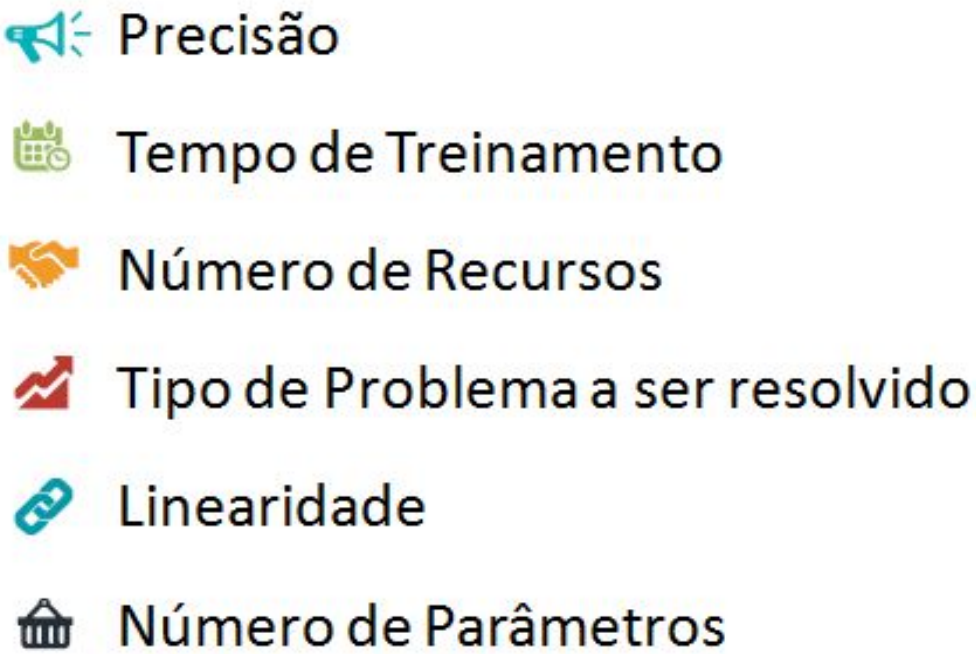
No aprendizado por reforço, o algoritmo tenta aprender qual é a melhor ação a ser tomada, dependendo das circunstâncias na qual essa ação será executada.

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.

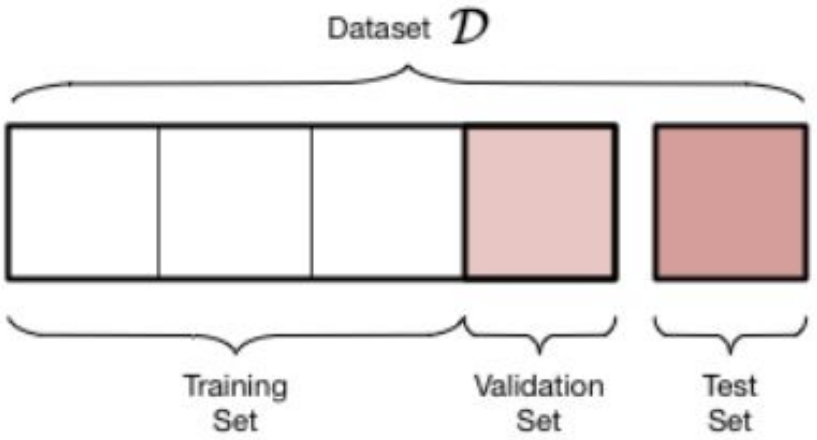
Essa técnica é muito usada em Games e Robótica, e vem obtendo resultados cada vez melhores.

Conceitos Que algoritmo de Machine Learning utilizar?

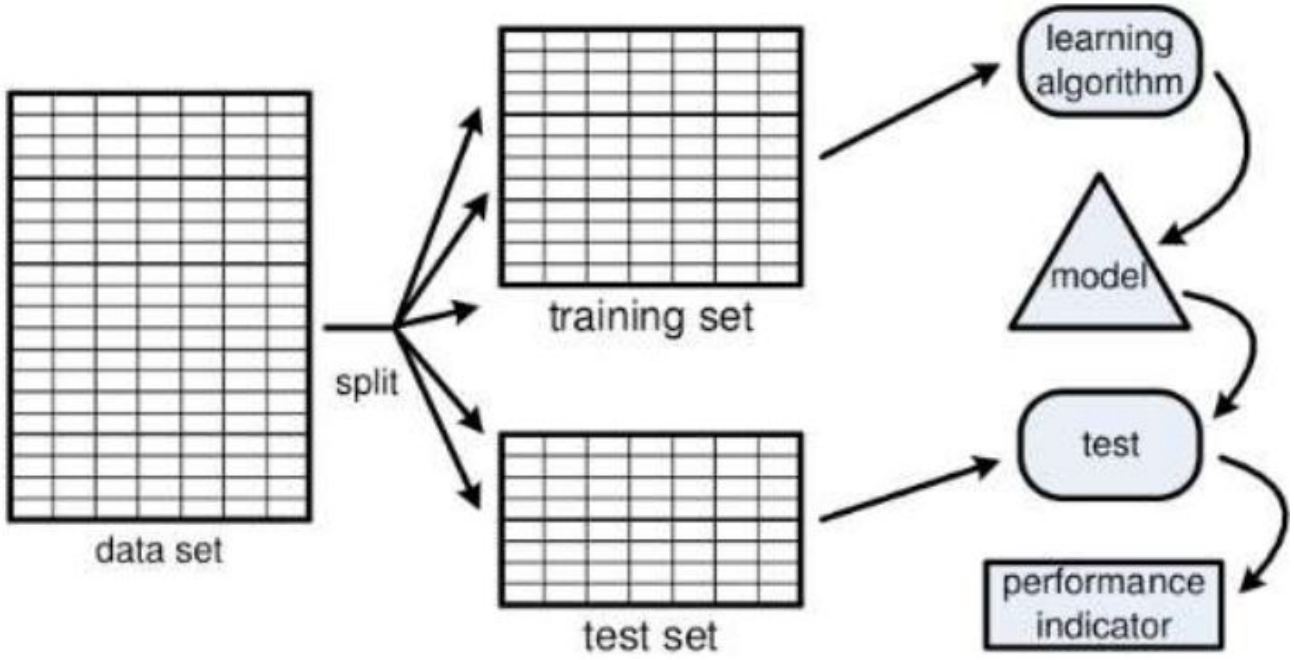
A resposta é “depende”. Tudo gira em torno do problema de negócio que se quer resolver. Cada caso é um caso, e o trabalho do cientista de dados é iterativo e exploratório. E a vários fatores a se considerar como mostrado acima.



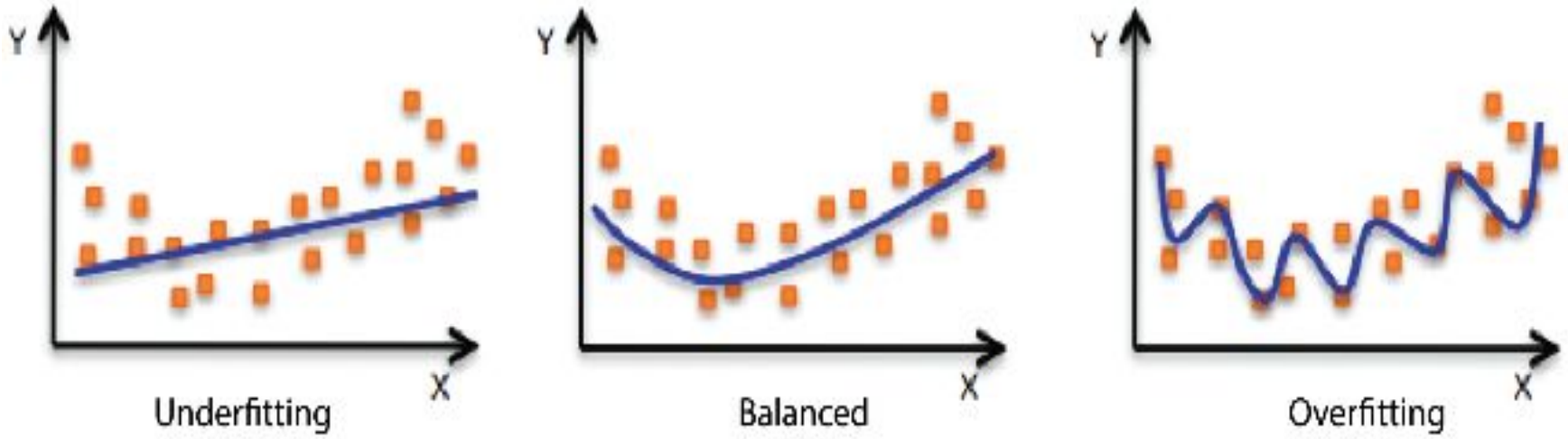
Conceitos Treinamento e Avaliação do Modelo



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



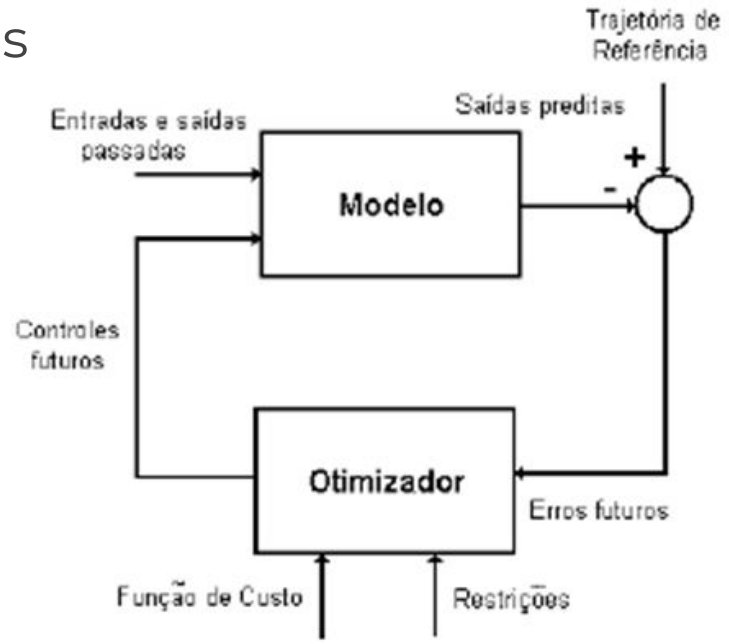
Conceitos Problemas de um Modelo Preditivo



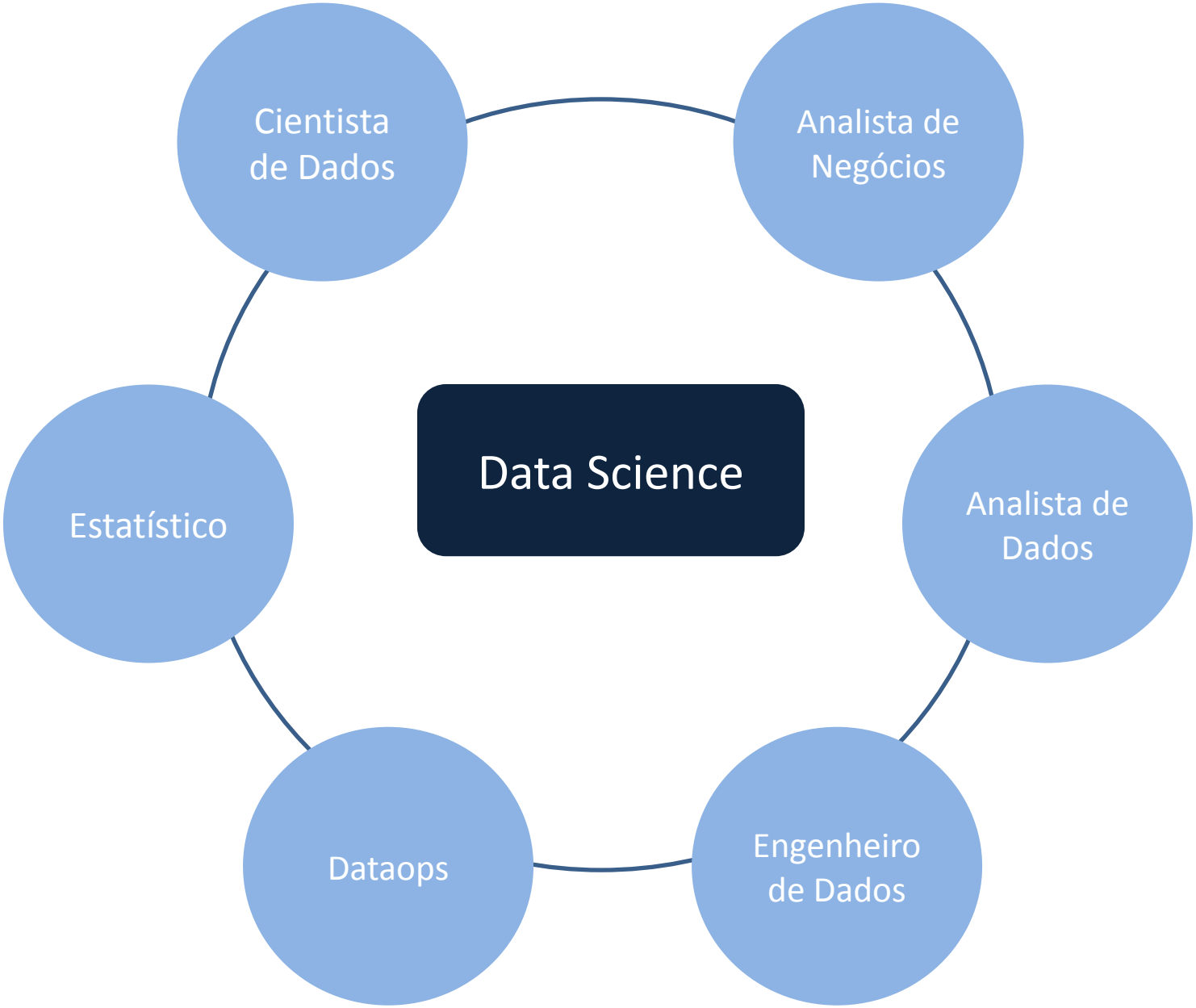
"O principal objetivo de um Modelo Preditivo é a **generalização** com novos dados, isto é, a capacidade de acerto com novos dados"

Conceitos 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.



Conceitos Carreiras em Big Data e Data Science



Tecnologias

No próximo Capítulo veremos:

- Um Case de Modelo Preditivo

"Os dados são importantes ativos para as empresas, pois são a bases para construção de Conhecimento"