

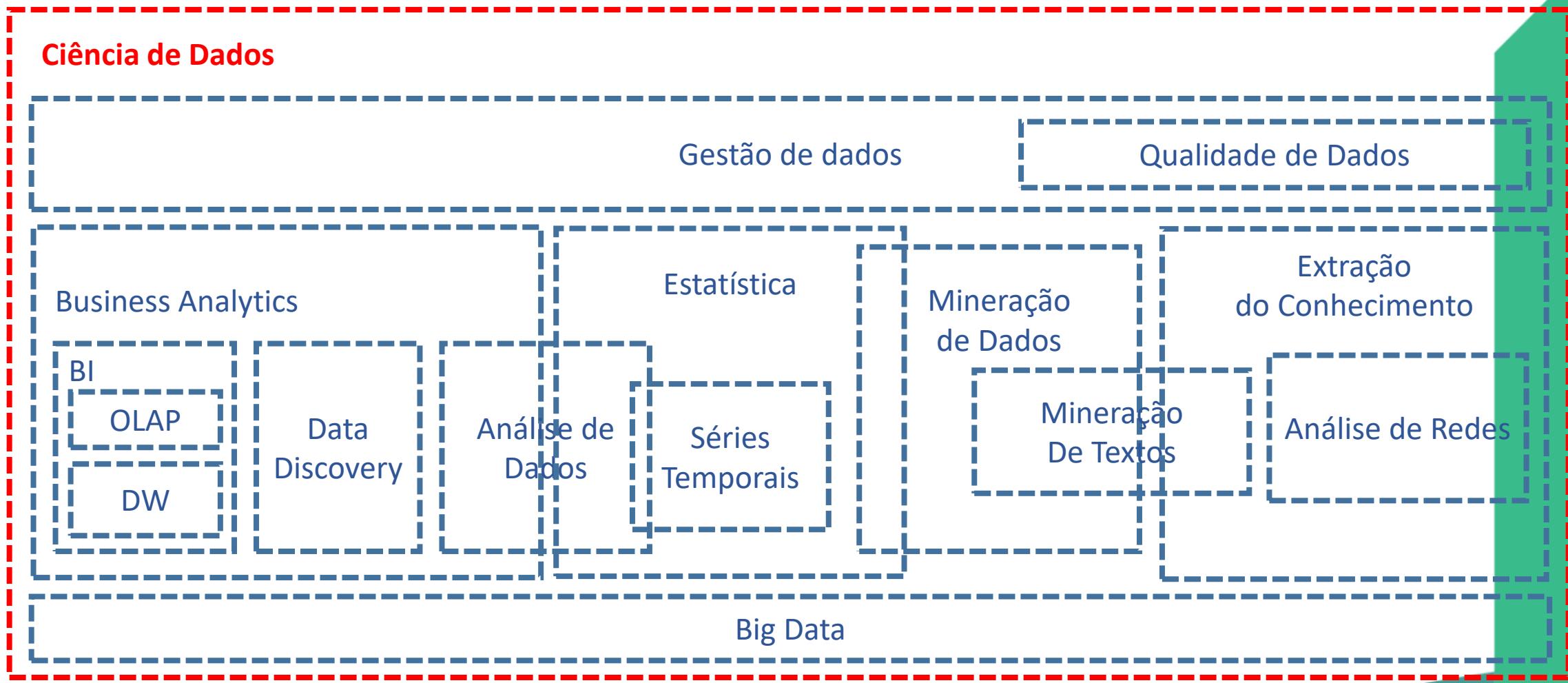
BIG DATA



Alvorada da Ciência de Dados



Abrangência



Método Científico



Produto de Dados

- É produzido através da prática de **Ciência de Dados**;
- É o conjunto dos **resultados, processos e tecnologias** envolvidas em Ciência de Dados direcionada à **geração de valor**;
- **Não é somente** uma análise ou uma recomendação para executivos, ou insight que direciona a melhoria de um processo de negócio.

Problema: como gerar valor com os dados?

- O que fazer com o **volume** e a **variedade** dos dados?
- O que eles **significam**?
- Como analisá-los em **tempo real**?
- O que isso pode gerar de **negócios, conhecimento, melhorias e transformações**?

Cientista de Dados

- Profissional com **habilidades técnicas e analíticas avançadas** para resolver **problemas complexos de dados**;
- Profissão mais **sexy** do século XXI.



Habilidades Plenas de um Cientista de Dados

- Conhecimento sobre:
 - ✓ **Negócios e comunicação;**
 - ✓ Economia, mercado e marketing;
 - ✓ Análise estatística avançada;
 - ✓ Gestão de Projetos e Gestão de Dados;
 - ✓ Business Intelligence e Data Warehouse;
 - ✓ Mineração de dados e extração de conhecimento de dados não estruturados;
 - ✓ Infra estrutura de TI, computação distribuída, computação de alta performance;
 - ✓ Programação avançada em diversas linguagens distintas.

Habilidades Plenas de um Cientista de Dados

- Conhecimento sobre:
 - ✓ Negócios e comunicação;
 - ✓ **Economia, mercado e marketing;**
 - ✓ Análise estatística avançada;
 - ✓ Gestão de Projetos e Gestão de Dados;
 - ✓ Business Intelligence e Data Warehouse;
 - ✓ Mineração de dados e extração de conhecimento de dados não estruturados;
 - ✓ Infra estrutura de TI, computação distribuída, computação de alta performance;
 - ✓ Programação avançada em diversas linguagens distintas.

Habilidades Plenas de um Cientista de Dados

- Conhecimento sobre:
 - ✓ Negócios e comunicação;
 - ✓ Economia, mercado e marketing;
 - ✓ **Análise estatística avançada;**
 - ✓ Gestão de Projetos e Gestão de Dados;
 - ✓ Business Intelligence e Data Warehouse;
 - ✓ Mineração de dados e extração de conhecimento de dados não estruturados;
 - ✓ Infra estrutura de TI, computação distribuída, computação de alta performance;
 - ✓ Programação avançada em diversas linguagens distintas.

Habilidades Plenas de um Cientista de Dados

➤ Conhecimento sobre:

- ✓ Negócios e comunicação;
- ✓ Economia, mercado e marketing;
- ✓ Análise estatística avançada;
- ✓ **Gestão de Projetos e Gestão de Dados;**
- ✓ Business Intelligence e Data Warehouse;
- ✓ Mineração de dados e extração de conhecimento de dados não estruturados;
- ✓ Infra estrutura de TI, computação distribuída, computação de alta performance;
- ✓ Programação avançada em diversas linguagens distintas.

Habilidades Plenas de um Cientista de Dados

- Conhecimento sobre:
 - ✓ Negócios e comunicação;
 - ✓ Economia, mercado e marketing;
 - ✓ Análise estatística avançada;
 - ✓ Gestão de Projetos e Gestão de Dados;
 - ✓ **Business Intelligence e Data Warehouse;**
 - ✓ Mineração de dados e extração de conhecimento de dados não estruturados;
 - ✓ Infra estrutura de TI, computação distribuída, computação de alta performance;
 - ✓ Programação avançada em diversas linguagens distintas.

Habilidades Plenas de um Cientista de Dados

- Conhecimento sobre:
 - ✓ Negócios e comunicação;
 - ✓ Economia, mercado e marketing;
 - ✓ Análise estatística avançada;
 - ✓ Gestão de Projetos e Gestão de Dados;
 - ✓ Business Intelligence e Data Warehouse;
 - ✓ **Mineração de dados e extração de conhecimento de dados não estruturados;**
 - ✓ Infra estrutura de TI, computação distribuída, computação de alta performance;
 - ✓ Programação avançada em diversas linguagens distintas.

Habilidades Plenas de um Cientista de Dados

- Conhecimento sobre:
 - ✓ Negócios e comunicação;
 - ✓ Economia, mercado e marketing;
 - ✓ Análise estatística avançada;
 - ✓ Gestão de Projetos e Gestão de Dados;
 - ✓ Business Intelligence e Data Warehouse;
 - ✓ Mineração de dados e extração de conhecimento de dados não estruturados;
 - ✓ **Infra estrutura de TI, computação distribuída, computação de alta performance;**
 - ✓ Programação avançada em diversas linguagens distintas.

Habilidades Plenas de um Cientista de Dados

- Conhecimento sobre:
 - ✓ Negócios e comunicação;
 - ✓ Economia, mercado e marketing;
 - ✓ Análise estatística avançada;
 - ✓ Gestão de Projetos e Gestão de Dados;
 - ✓ Business Intelligence e Data Warehouse;
 - ✓ Mineração de dados e extração de conhecimento de dados não estruturados;
 - ✓ Infra estrutura de TI, computação distribuída, computação de alta performance;
 - ✓ Programação avançada em diversas linguagens distintas.

Quanto tempo para formação tradicional?

4 anos - Bacharelado em Ciência da Computação;

4 anos - Bacharelado em Estatística;

1,5 anos - MBA em Gestão Estratégica de Negócios;

1,5 anos - Especialização em Marketing e Comunicação Digital;

1,5 anos - Especialização em Computação Distribuída;

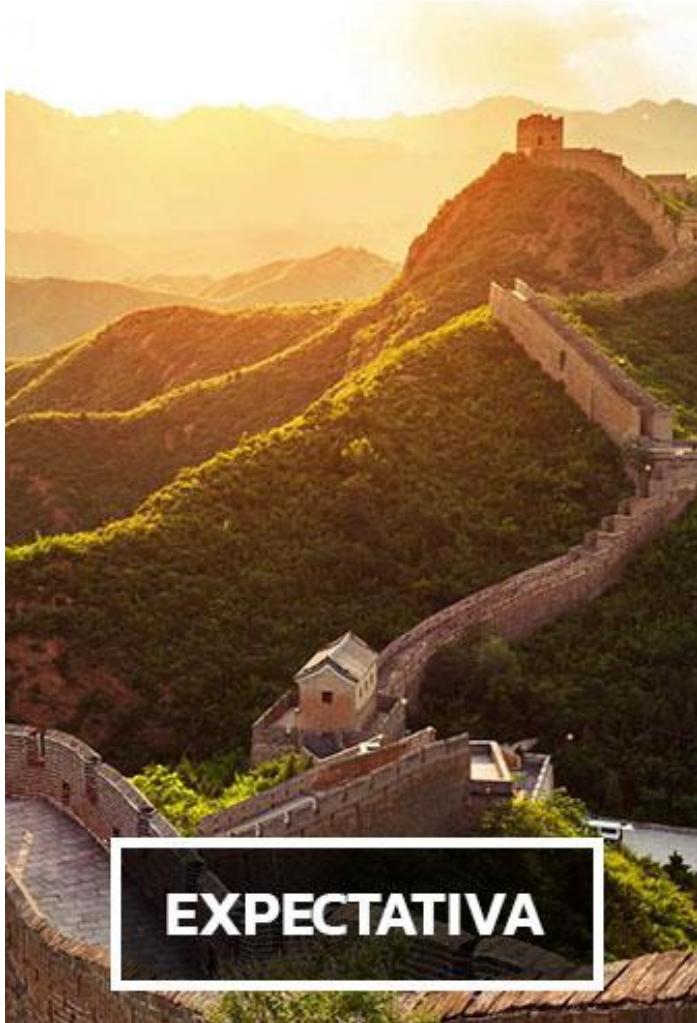
1 ano - Especialização em Programação e Padrões de Projeto;

2 anos - Mestrado em Gestão do Conhecimento;

? - Especialização técnica em diversas técnicas e ferramentas
(Soft Skills, Ferramentas de BI, Big Data, R e etc);

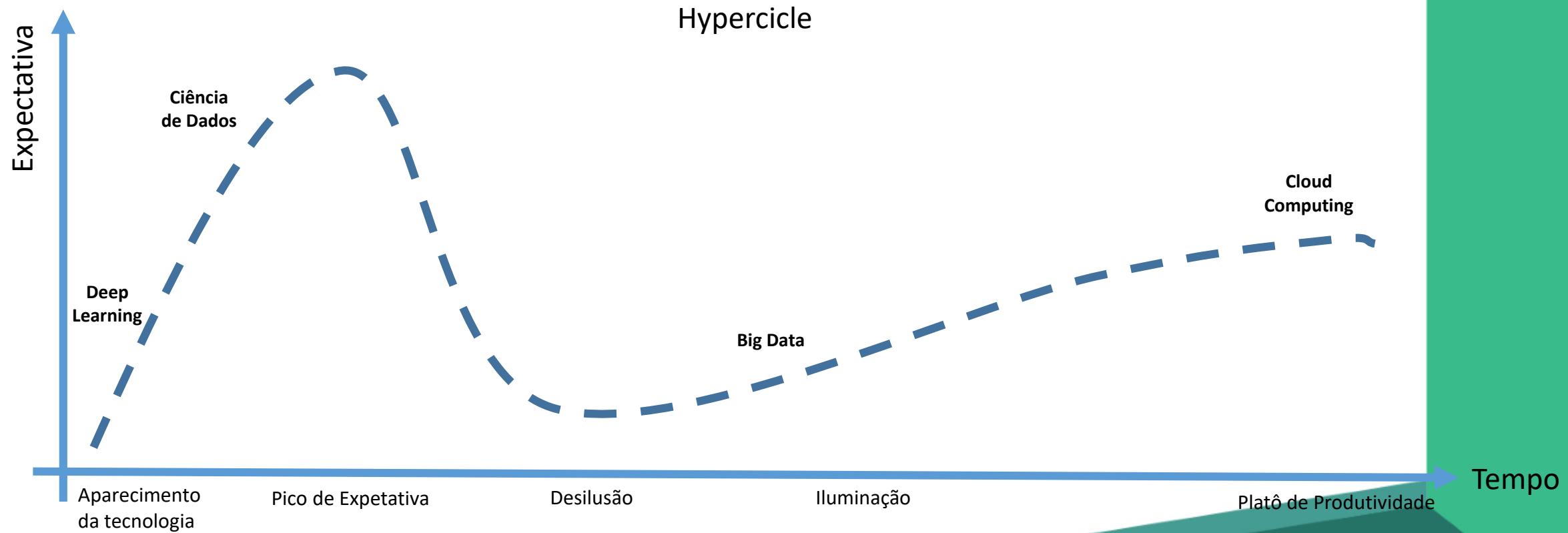
? - Tempo de carreira para entender plenamente o negócio a ponto
de identificar todas as oportunidades.

Expectativa x Realidade

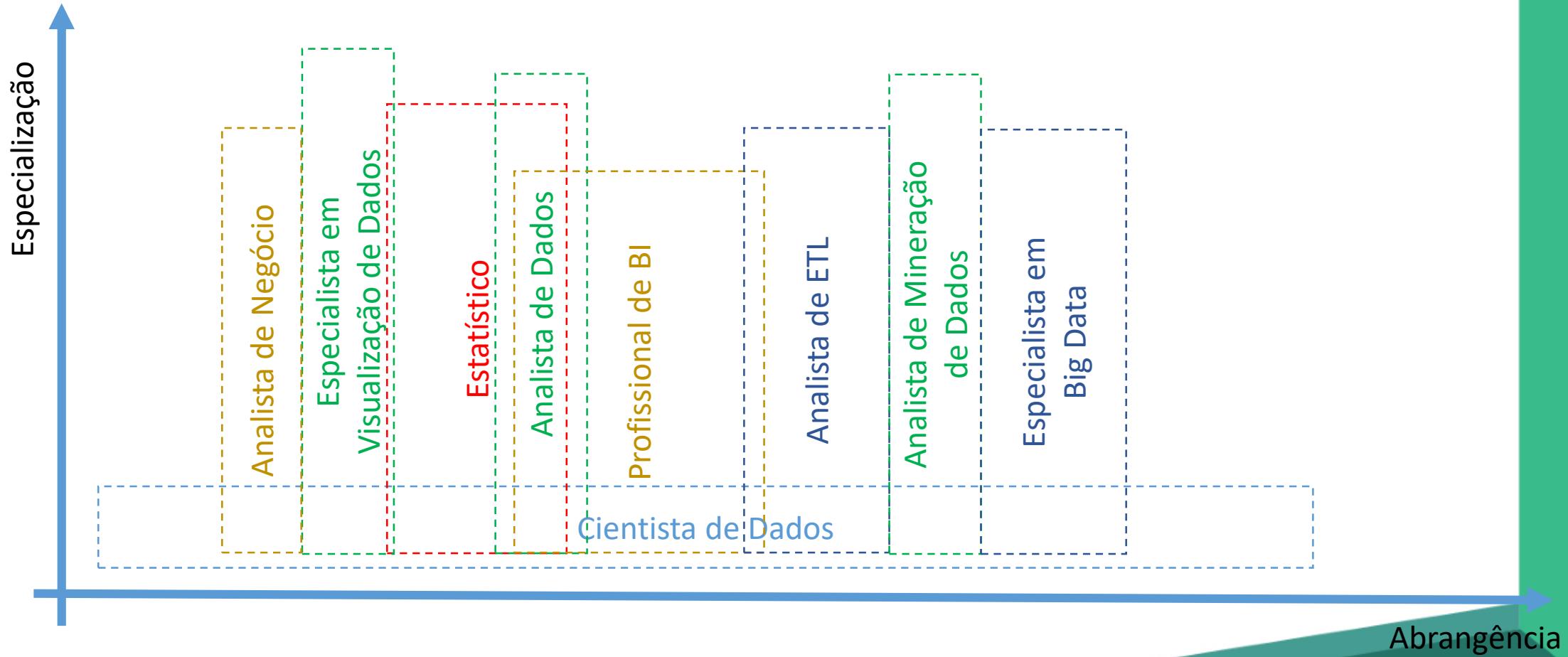


Buzzword (Big Data, Data Science...)

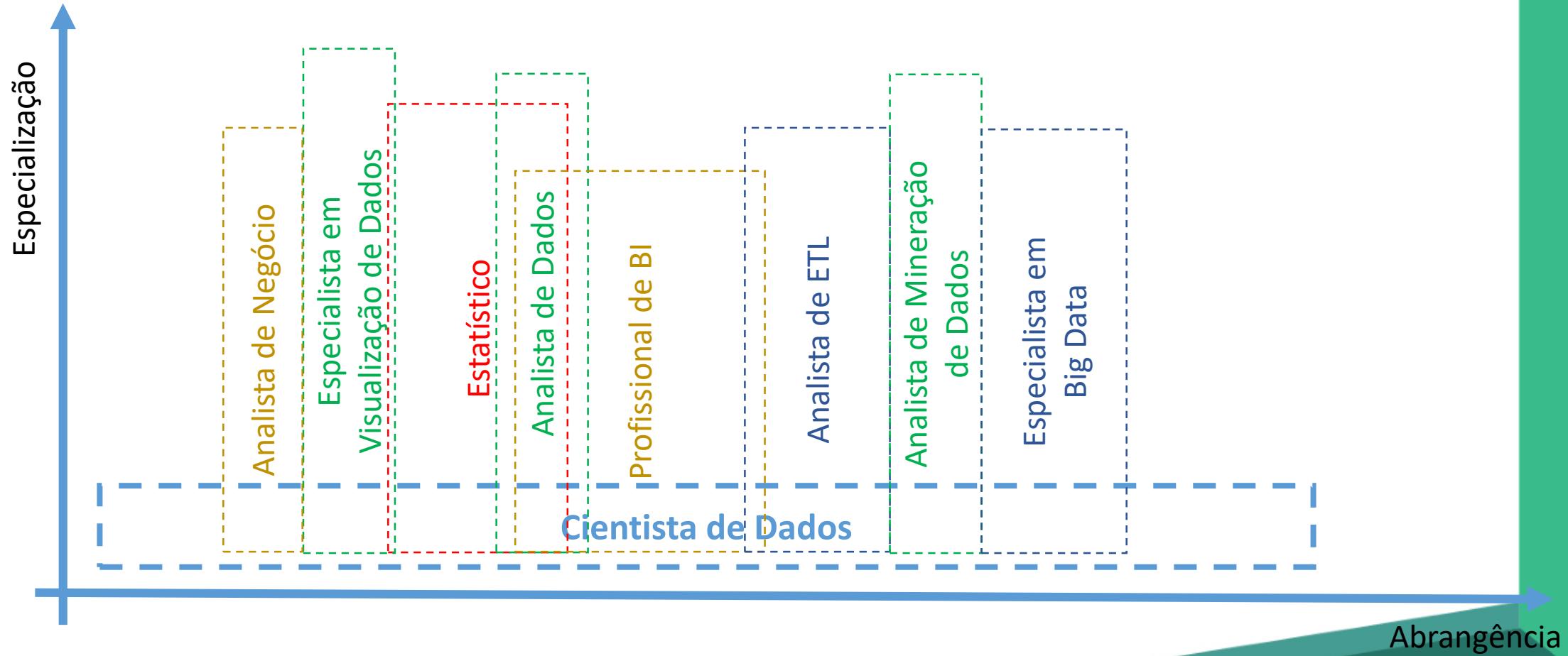
É novo termo, tratado com muito entusiasmo, que tem como base o aparecimento de uma palavra até então desconhecida



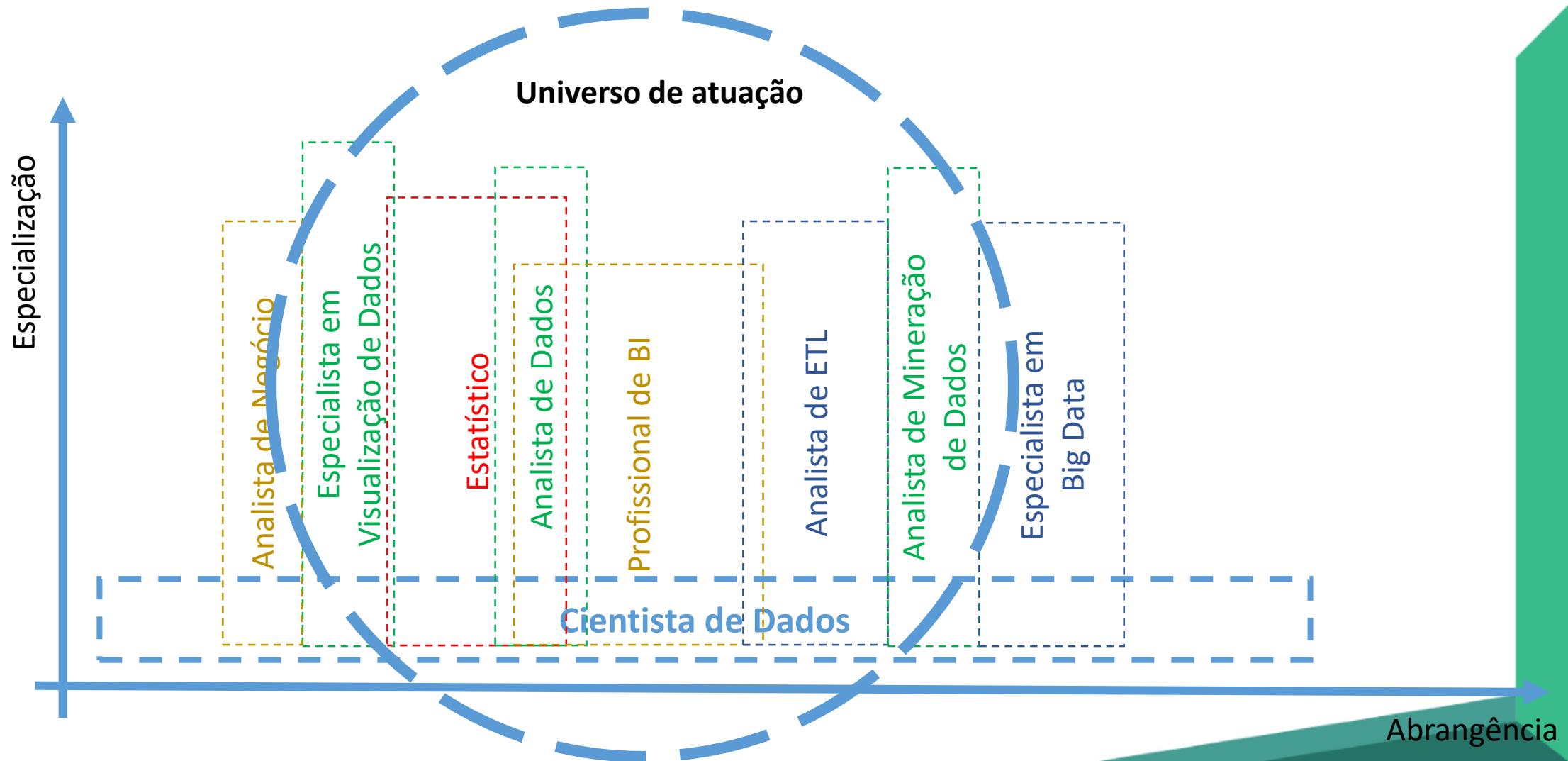
Conhecimento Generalista x Especialista



Conhecimento Generalista x Especialista



Conhecimento Generalista x Especialista



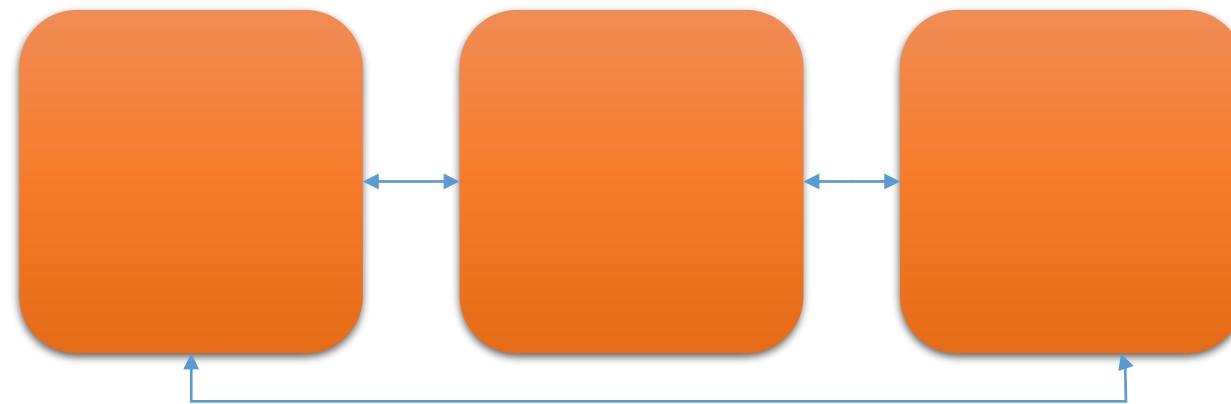
Multidisciplinaridade

- Existe uma temática comum;
- Não existe relação nem cooperação entre disciplinas.



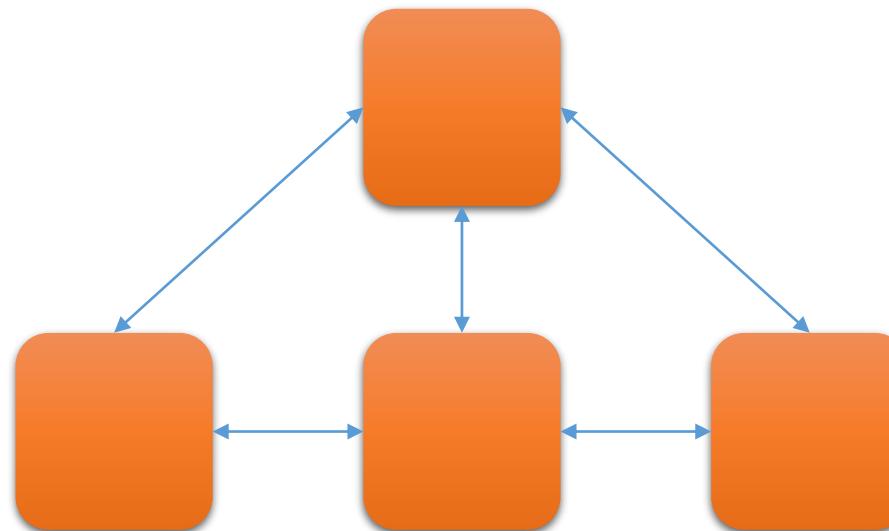
Plurisciplinaridade

- Existem uma temática comum;
- Existe uma relação de cooperação entre disciplinas.



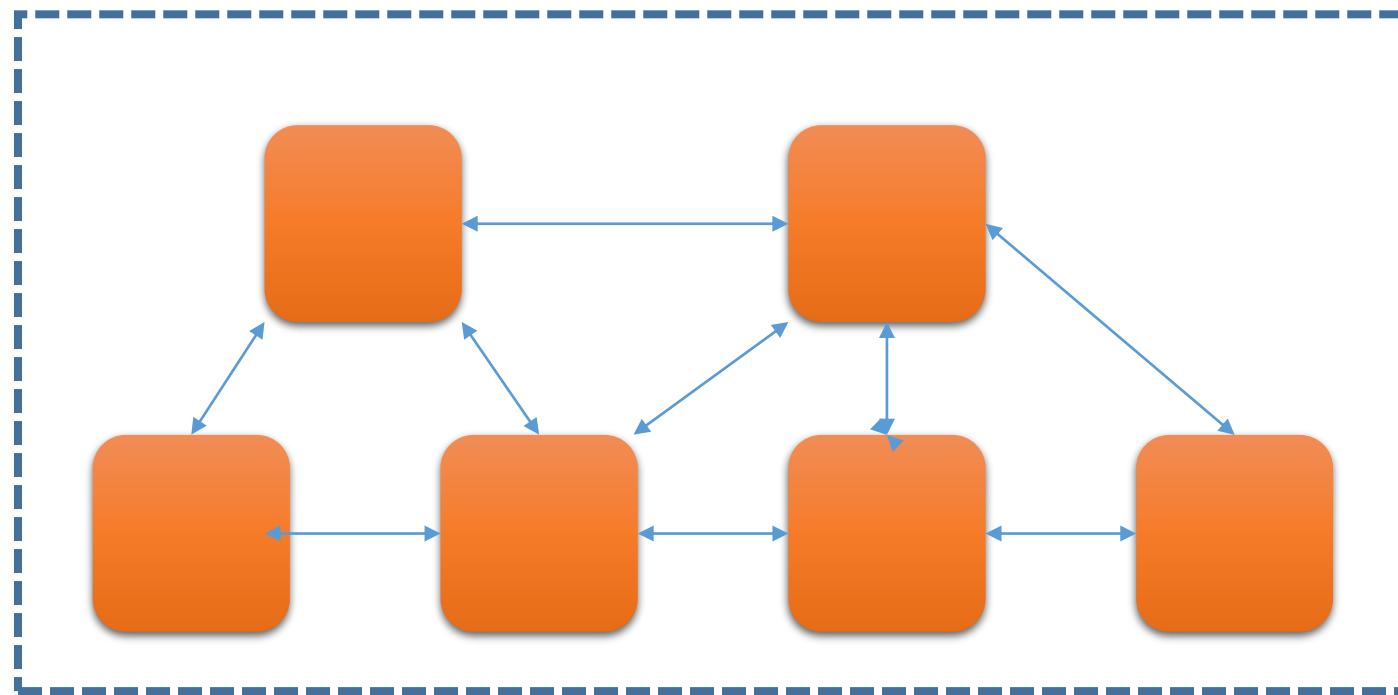
Interdisciplinaridade

- Existe cooperação e diálogo entre as disciplinas;
- Existe uma ação coordenada.



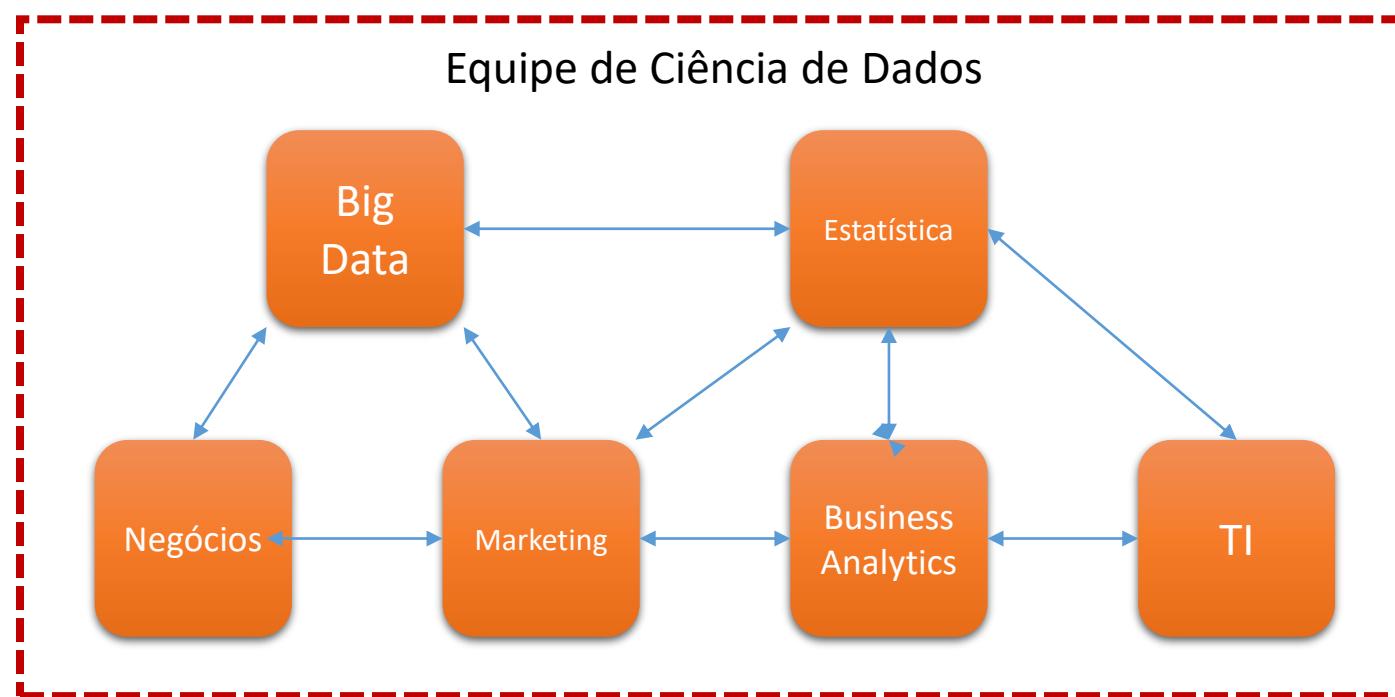
Transdisciplinaridade

- Cooperação entre todas as disciplinas e interdisciplinas.

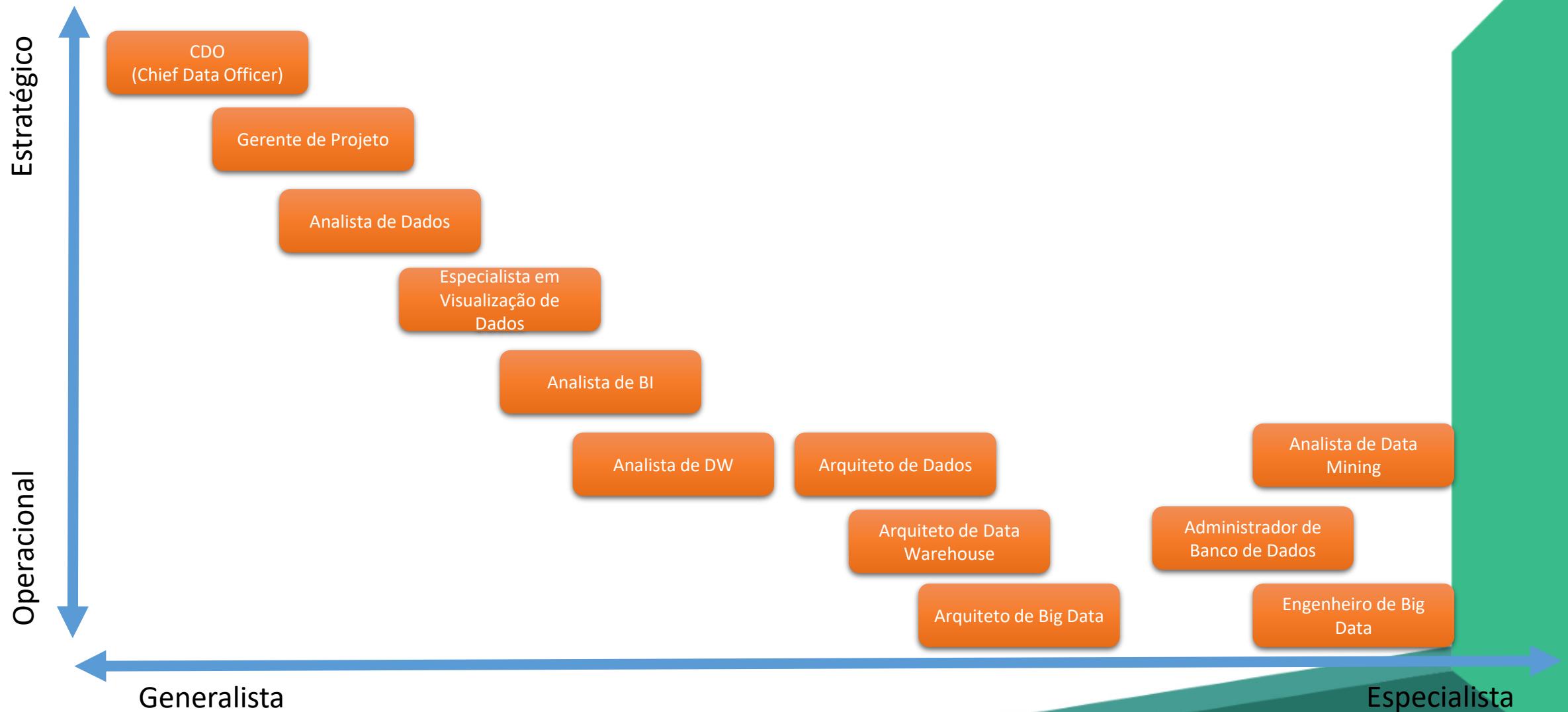


Transdisciplinaridade

- Cooperação entre todas as disciplinas e interdisciplinas.

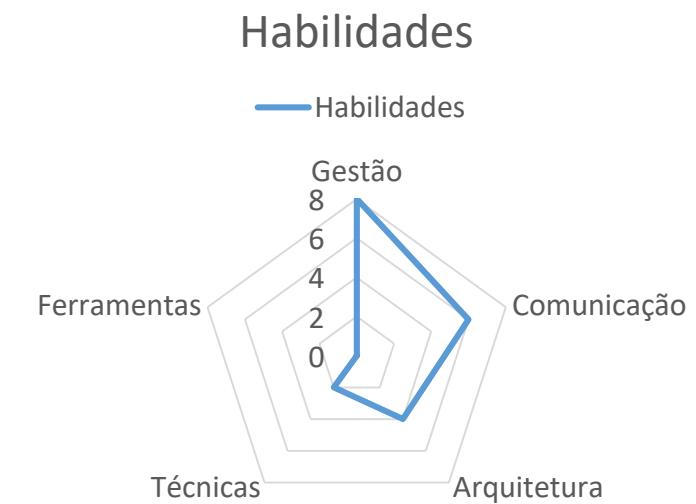


Equipe de Ciência de Dados



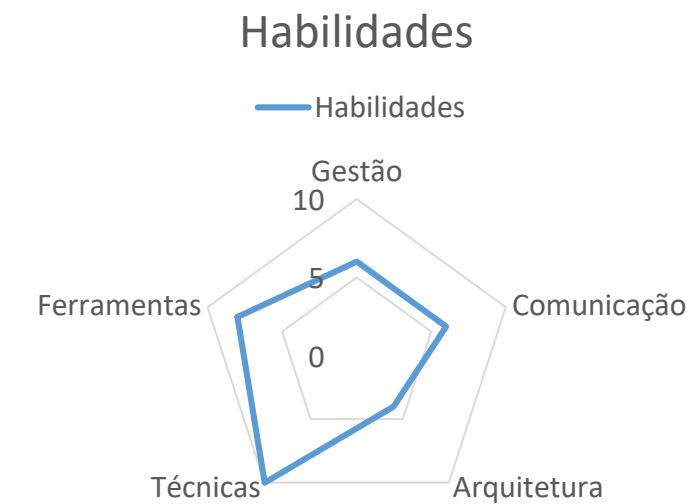
Chief Data Officer

- Trabalhar com executivos, provedores de dados e data stewards para atingir os objetivos estratégicos dos clientes internos e externos.



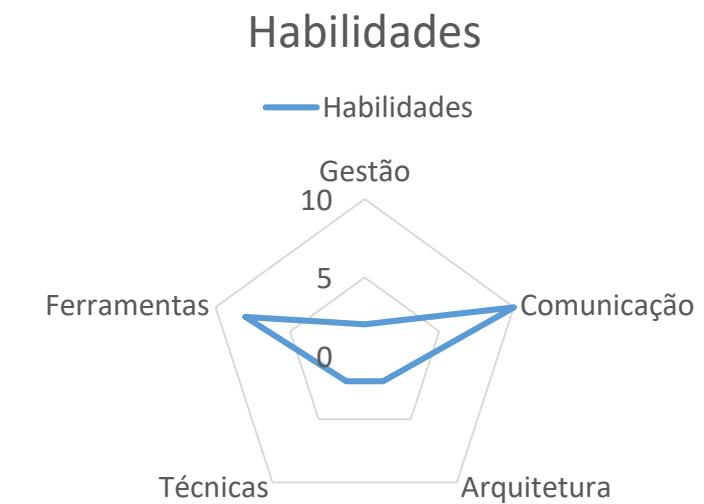
Analista de Dados

- Coordenar o prover suporte para todas as análises de dados;
Apresentar os produtos de dados;
- Conhecer o negócio.



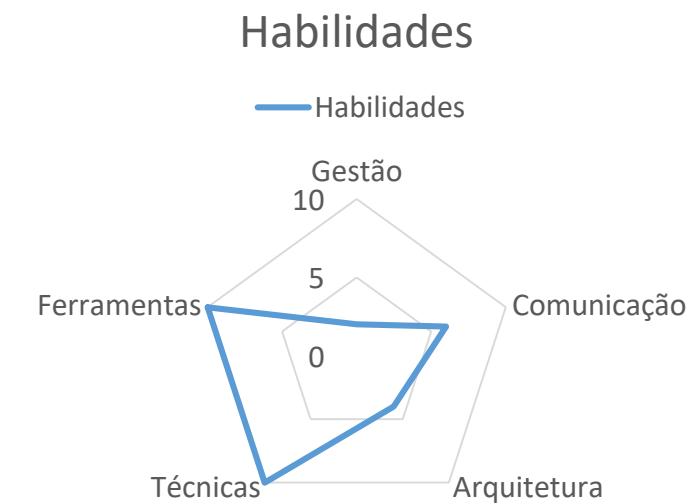
Especialista em Visualização de Dados

- Traduzir uma informação complexa em informação simples de ser entendida;
- Projetar infográficos, painéis e apresentações.



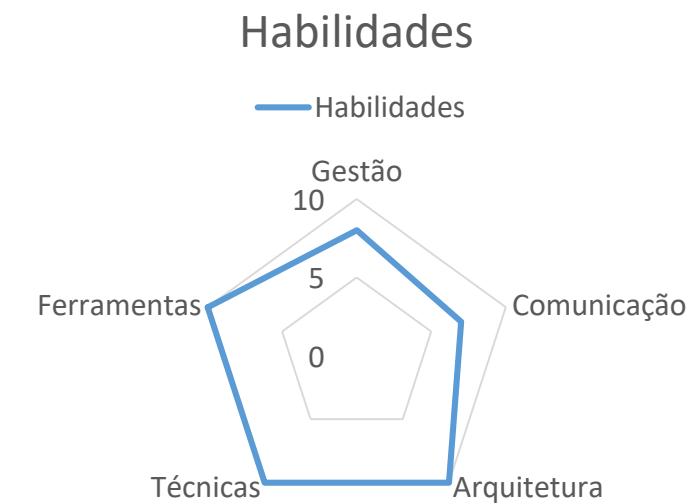
Analistas de BI e DW

- Implementar Produtos de Dados por meio de ferramentas OLAP, Data Discovery, Reports, Dashboards;
- Implementar modelos multidimensionais;
- Implementar processos de carga de ETL.



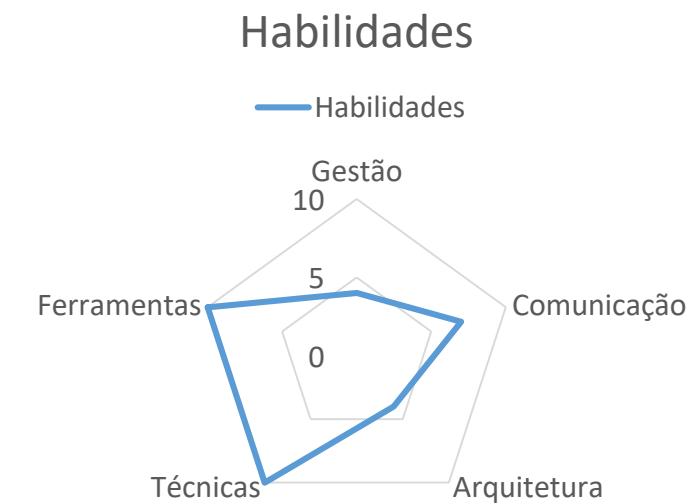
Arquiteto de Dados

- Projetar modelos de dados, processos de integração de dados;
- Mapear origens de dados potencialmente úteis.



Analista de Data Mining

- Projetar e coordenar o desenvolvimento de modelos de análise de dados avançadas.



Representação do Conhecimento

- Como representar o conhecimento?
- Existem melhor forma representar alguma coisa?

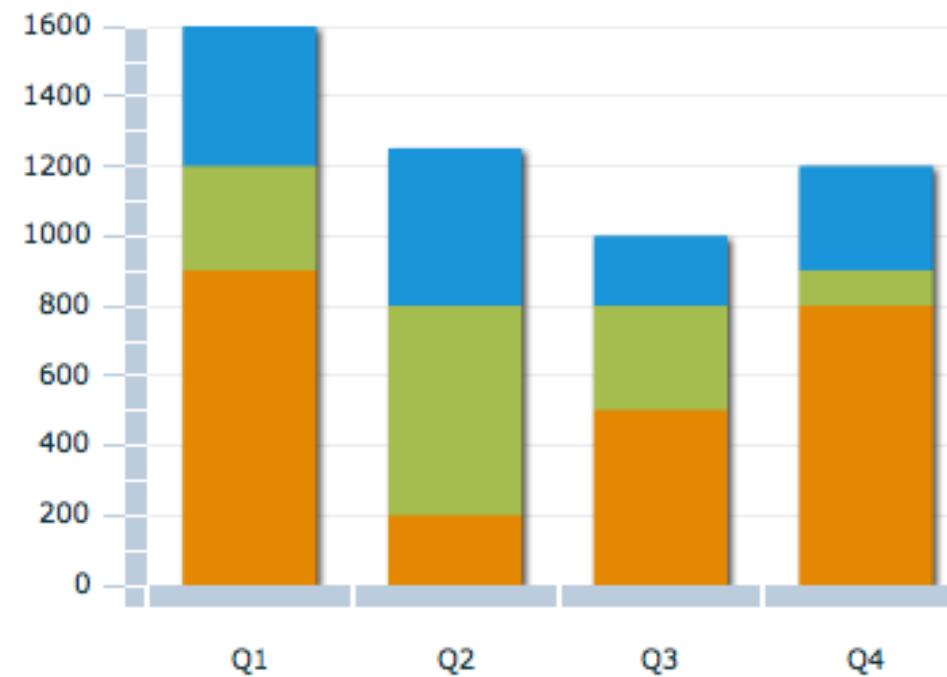


Análise Categórica - Unidimensional

Estatística	Visualização	Descrição
Contagem	Gráfico de Barras	Comparar valores diferentes em série
Porcentagem	Gráfico de Pizza	Permite ter noção sobre a proporção das categorias analisadas sobre um assunto

Análise Bi Variada (Categórico e Categórico)

- O gráfico de colunas empilhadas é útil para visualizar o relacionamento entre duas variáveis categóricas. Exemplo:



Analise Bi Variada (Categórico e Categórico)

- Um gráfico de combinação usa dois ou mais tipos de gráficos para enfatizar diferentes tipos de informação;
- O gráfico de combinação é o melhor método de visualização para demonstrar o poder de previsibilidade de um preditor (eixo X) contra um alvo (eixo Y).

Gráfico de dois eixos Y

- Compara valores sobre várias séries coletadas sobre os dados.

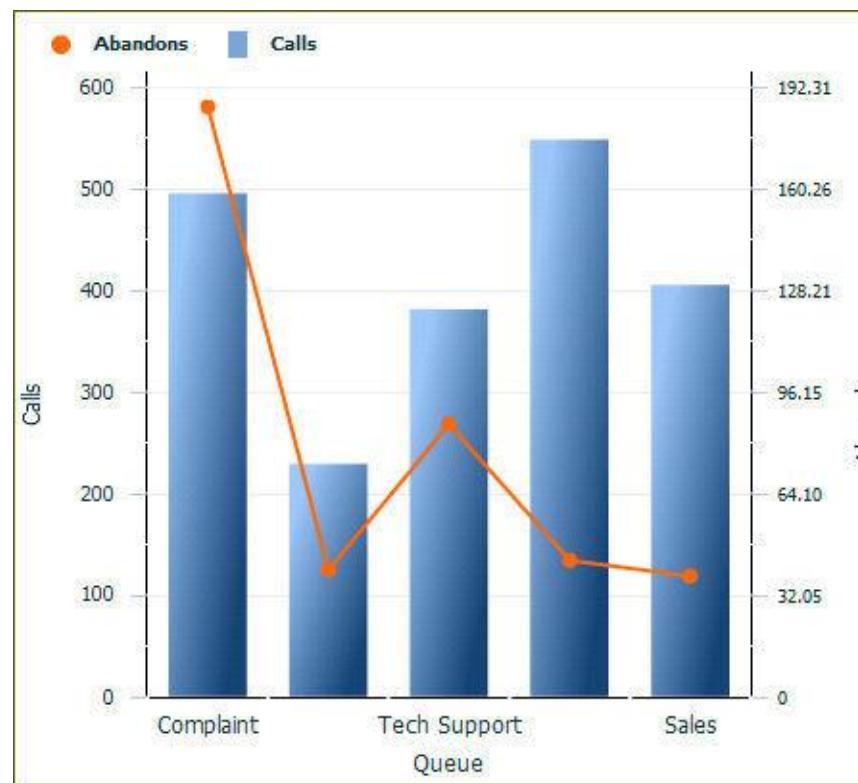


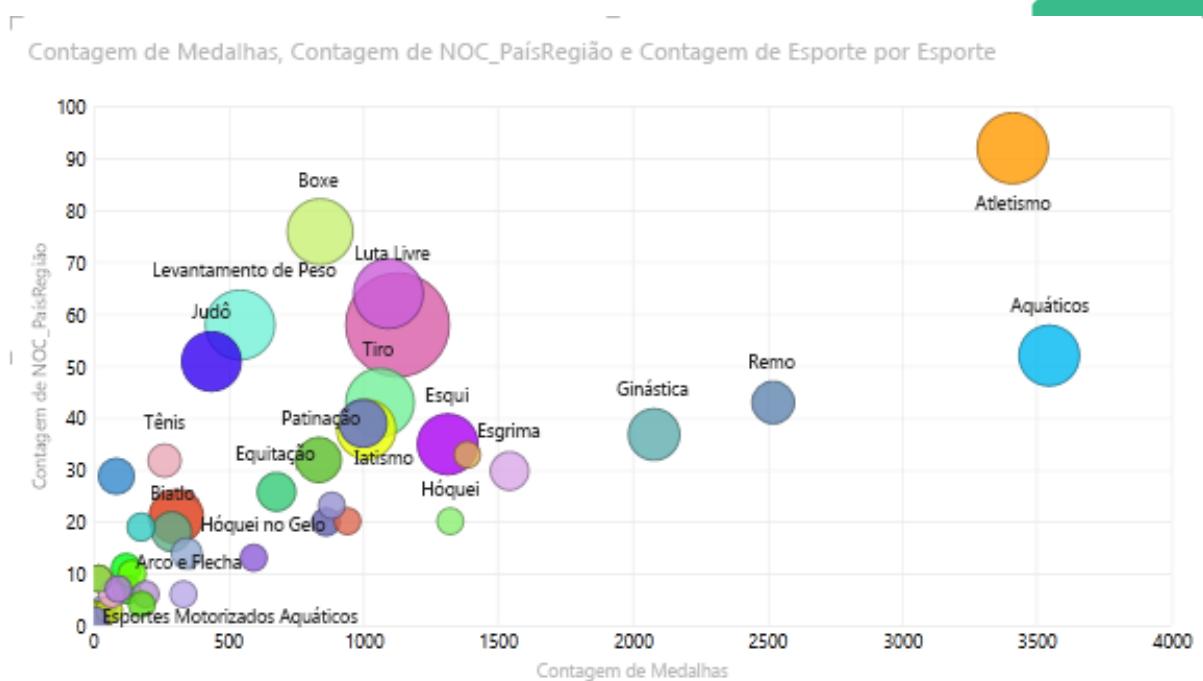
Gráfico de Radar

- Compara valores sobre várias séries coletadas sobre os dados. Esse gráfico permite apresentar várias dimensões ao mesmo tempo com fácil visualização comparativa.



Gráfico de Dispersão

- Análise Bivariada (dados numéricicos);
- Um gráfico de dispersão é uma representação visual útil da relação entre duas variáveis numéricas (atributos);
- Geralmente é elaborado antes de trabalhar uma correlação linear ou um modelo de regressão linear.



Matriz de Dispersão

- Análise Bivariada (dados numéricicos);
- Permite apresentar a dispersão de um conjunto de dados com muitas variáveis.

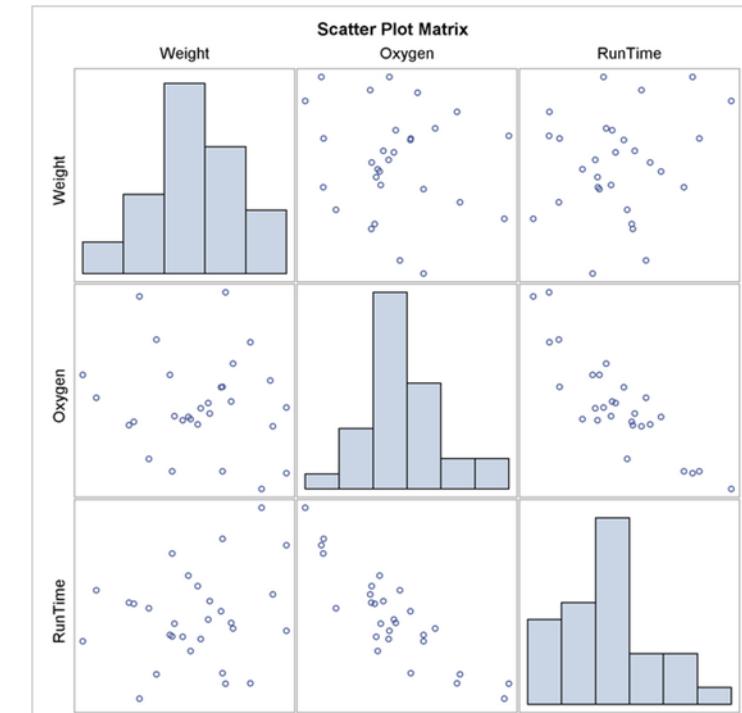
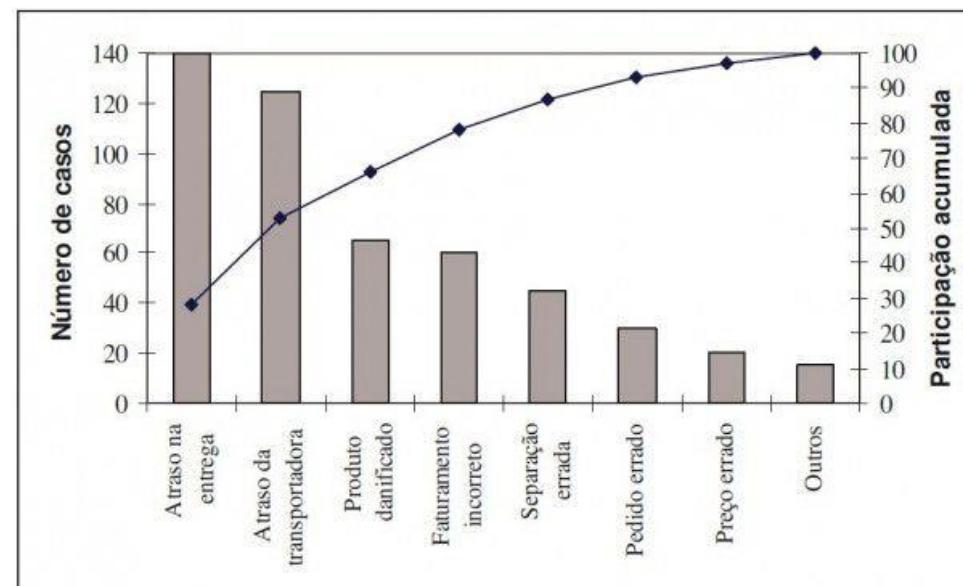


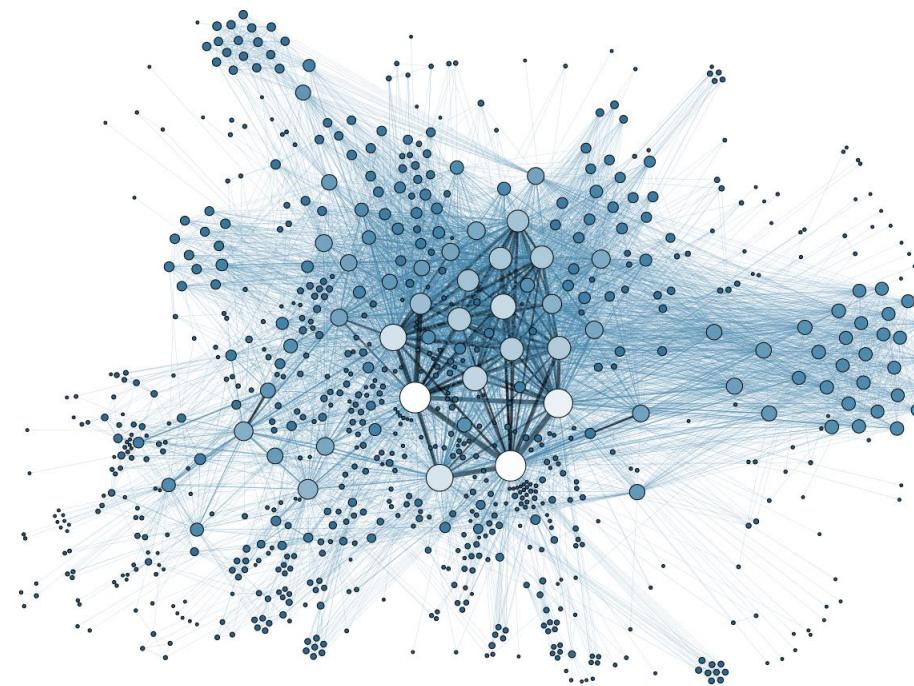
Gráfico de Pareto

- Princípio de Pareto: 80 % das ocorrências vem de 20% das causas
- Permite a visualização e identificação das ocorrências mais importantes, possibilitando a concentração sobre esses.



Grafos

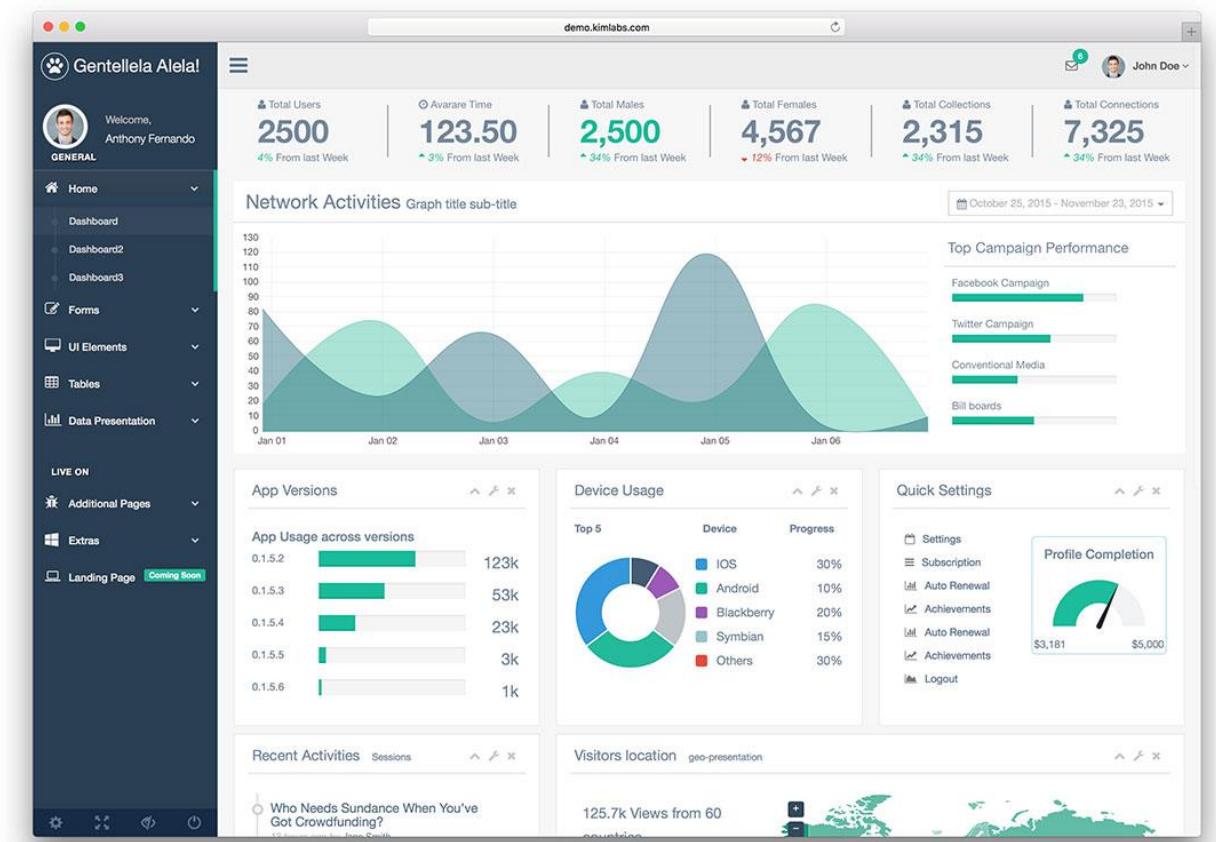
- Permite analisar as relações entre objetos;
- Permite visualizar e analisar redes sociais e sistemas complexos.



Painéis

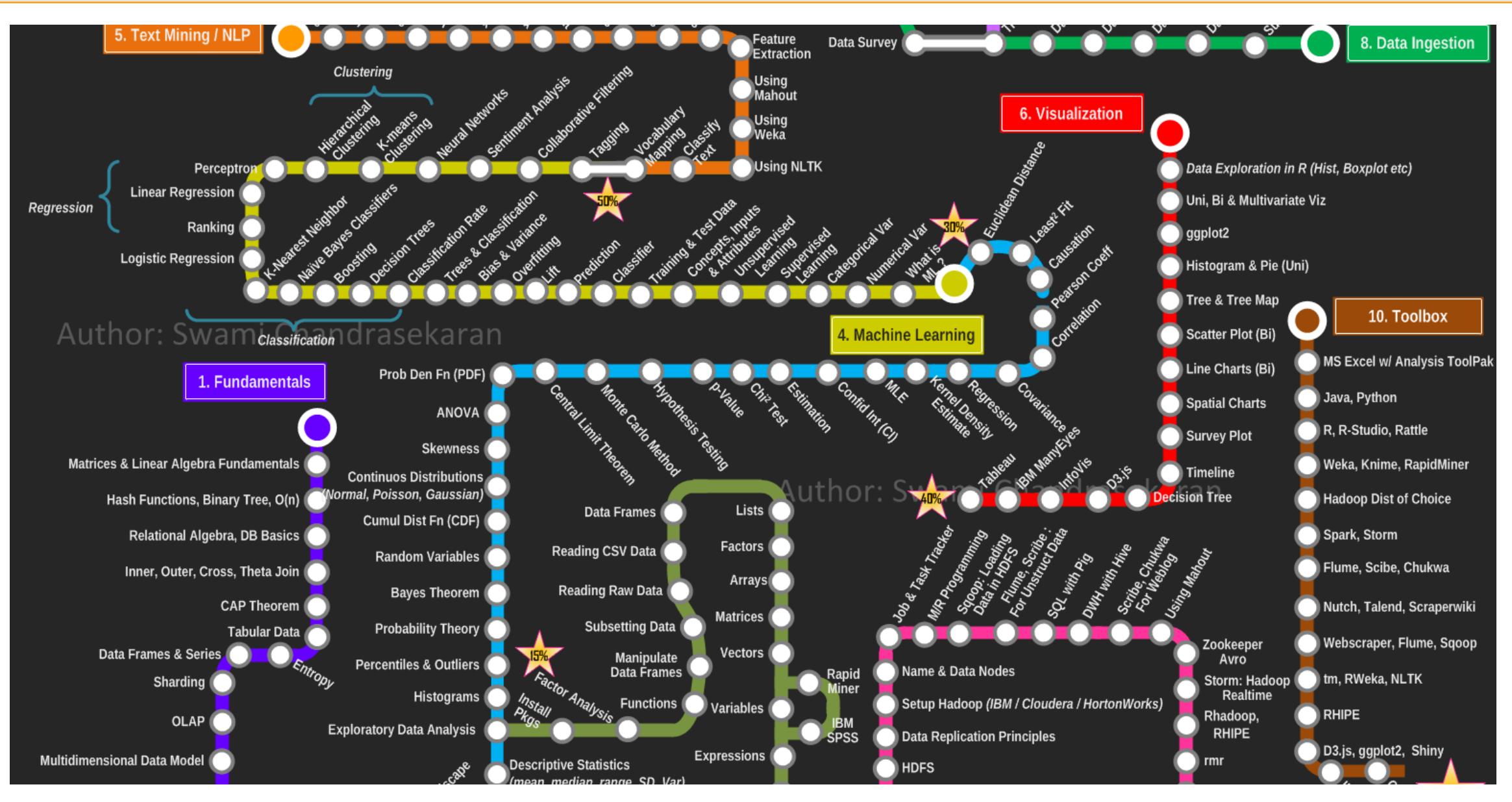
- Fornece uma representação ilustrada do desempenho dos negócios em toda a organização;
- É utilizado para a apresentação de dados e alertas relevantes para a tomada de decisão;
- Apresenta facilidade para interpretar e tirar conclusões.

Painéis



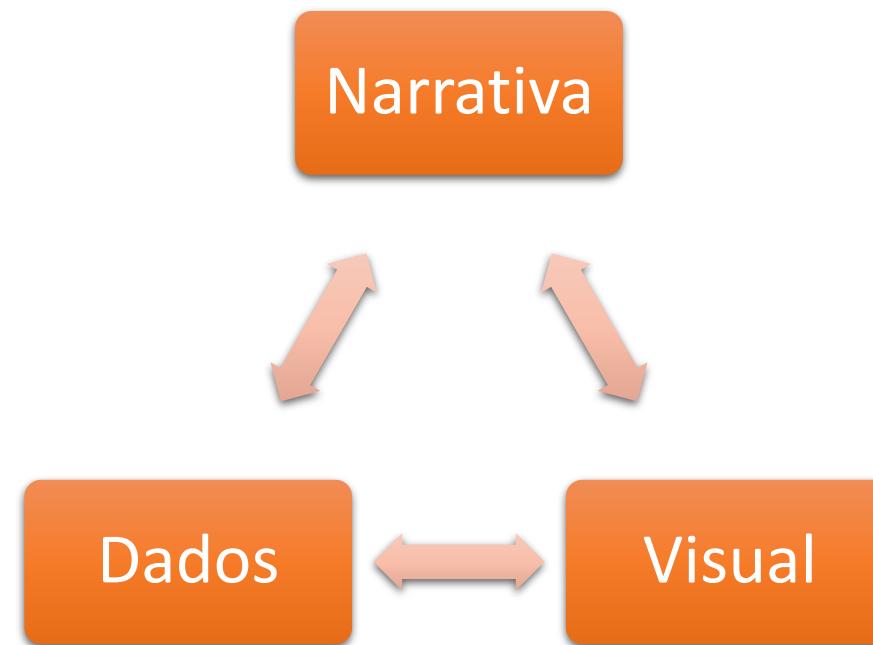
Infográfico

- Ilustrações explicativas sobre um tema ou um assunto;
- Apela para **linguagem visual**;
- Expõe informações de forma clara e objetiva.



Storytelling com Dados

- Habilidade de contar histórias sobre dados de forma interessante, utilizando recursos audiovisuais.



Exemplo de Storytelling:



Medida, Métrica e Indicador

Medida – Quantificação de dados em um padrão e qualidade aceitáveis (exatidão, completude, consistência, temporalidade).

Ex: Comprimento de um material em metros.

Métrica – Extrapolação de medidas, isto é, conclusão com base em dados finitos.

Ex: Número de defeitos (defeitos / total do lote).

Indicador – Representação de forma simples e intuitiva de uma métrica ou medida para facilitar sua interpretação quando comparada a uma referência.

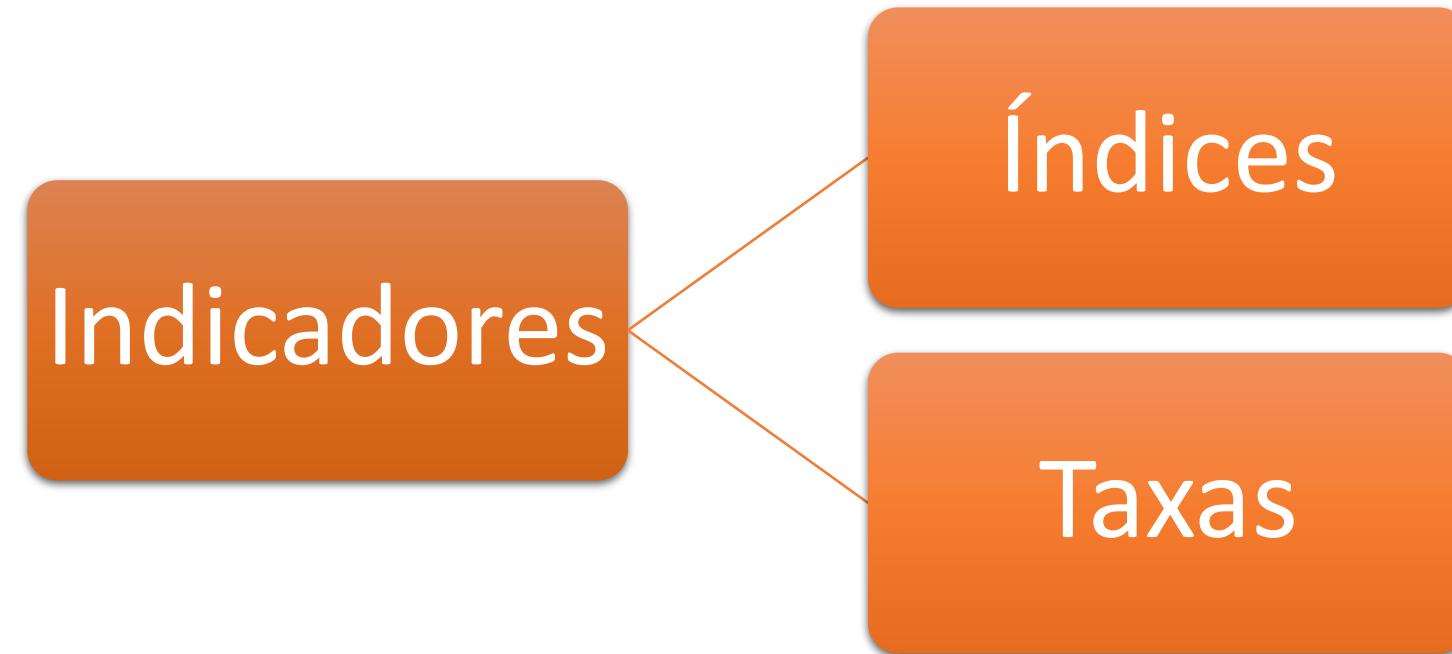
Medida, Métrica e Indicadores

	Definição	Exemplo
Medida	É a associação de uma grandeza numérica a uma característica ou a frequência em que um evento ocorre.	Qual a altura dos cliente que entram na loja? Quantos cliente entraram na loja?
Métrica	É o conjunto de medidas tomadas ao longo de um período utilizando a mesma metodologia de mediação.	Ao longo de um período, dia-a-dia, qual a altura e quantos clientes entram na loja?
Indicador	Representa a variação, geralmente percentual, de uma medida ou métrica em relação a um referencial conhecido.	Conhecendo-se a média mensal de clientes que entram na loja, dizer se em um determinado mês o número de clientes que efetivamente entrou superou ou não a média conhecida.

Indicadores

- Indicador social é uma medida, geralmente estatística, usada para **traduzir quantitativamente um conceito social abstrato** e informar algo sobre determinado aspecto da realidade social para fins de pesquisa ou visando a formulação, monitoramento e avaliação de programas e políticas públicas;
- Para a OCDE, indicador é um parâmetro, ou valor derivado de parâmetros, que indica, fornece informações ou **descreve o estado de um fenômeno área/ambiente**, com maior significado que aquele apenas relacionado diretamente ao seu valor quantitativo;
- Os indicadores podem ser **analíticos** (constituídos de uma única variável: esperança de vida ao nascer, taxa de alfabetização, escolaridade média, etc.) ou **sintéticos** (quando resultantes de uma composição de variáveis, como o IDH).

Categorías de Indicadores



Taxa e Índice

Índice - representa uma medida síntese de dados agregados.

Taxa - representa uma medida que avalia a proporcionalidade de um todo.

Qualidades de um bom indicador

- ✓ Ele tem valor próprio;
- ✓ É capaz de mostrar resultados;
- ✓ O que ele mede é importante;
- ✓ Ele é estatisticamente significativo.



Indicadores mal definidos

- Exemplo: tomar somente a quantidade de reclamações de clientes, mês a mês, ao longo do ano, e verificar que o número absoluto de reclamações cresceu no período não indica, necessariamente, uma piora nos negócios. Está claro que, se a sua empresa efetuar 1.000 vendas em dezembro e ter 10 reclamações de clientes, é uma situação melhor comparado a ter efetuado 100 vendas em janeiro e ter recebido 5 reclamações. Proporcionalmente, o número de reclamações terá caído de 5% ($5/100$) para 1% ($10/1.000$), embora em números absolutos elas tenham dobrado.

Indicadores-Chave de Desempenho

- São métricas consideradas essenciais para **avaliar um processo** sob gestão;
- Suas combinações podem apontar o **sucesso** e a conclusão de um **objetivo estratégico**;
- Chamados de **KPI**.

Exemplos de Indicadores:

- **Lead Time** - Tempo de Duração de um processo.
- **Stock Out** - Número de vezes ou dias que determinado item controlado no estoque chega ao saldo zero.
- **Produtividade Homem/hora** - Número de unidades produzidas por mão de obra escalada na produção.
- **Ociosidade** - porcentagem de tempo que uma máquina, equipe ou planta ficam parados.

Índice

- Medida com propósito de comparação de entidades sobre a derivação de diversas dimensões.



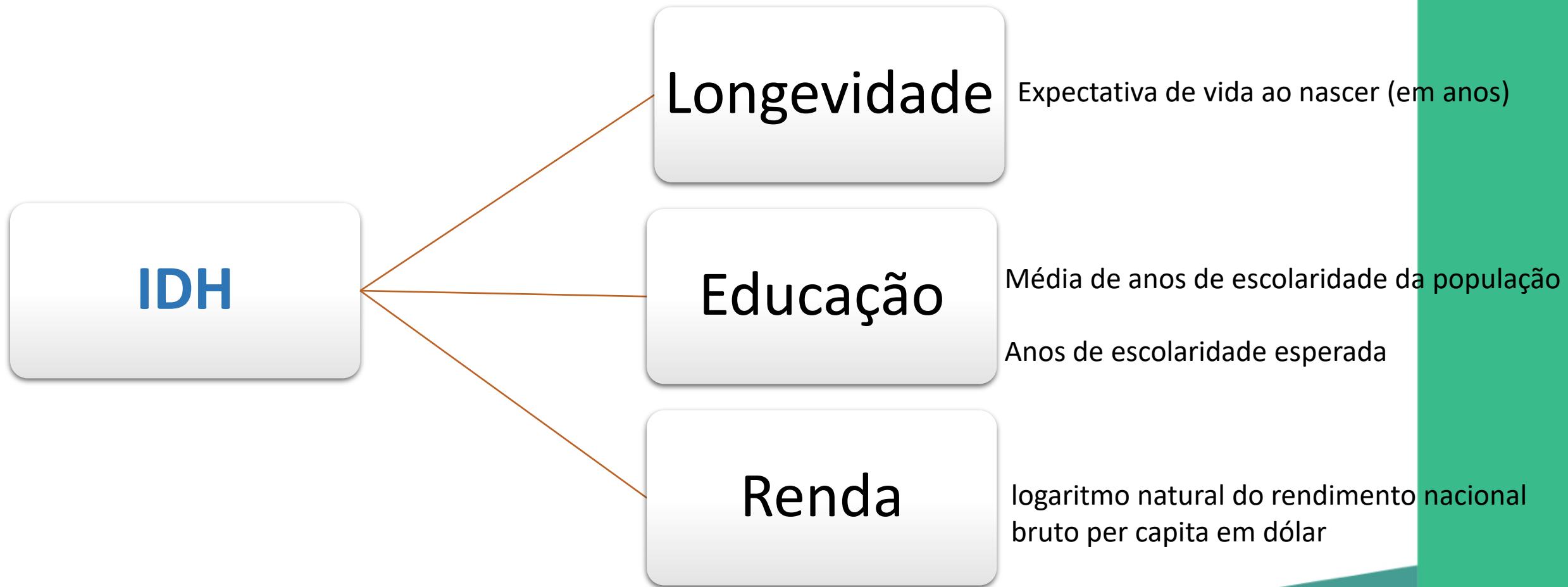
Exemplo de Índice:

Índice de Desenvolvimento Humano (IDH)

- Utilizado pelo Programa das Nações Unidas como medida de desenvolvimento social e econômico dos países, desde 1993;
- É uma medida comparativa usada para classificar os países pelo seu grau de "desenvolvimento humano".

Obs: Não pode ser interpretado como uma medida de "felicidade" ou um indicador do "melhor lugar para se viver".

Dimensões do IDH



Dimensões Conceituais do IDH

- ✓ **Uma vida longa e saudável:** expectativa de vida ao nascer;
- ✓ **O acesso ao conhecimento:** anos Médios de Estudo e Anos Esperados de Escolaridade;
- ✓ **Um padrão de vida decente:** PIB.

Metodologia do Índice

Para permitir a representação dos indicadores de educação, longevidade e renda em um único índice é preciso **normalizar cada dimensão** (transformar em índice entre 0 e 1) pelo valor **máximo** e **mínimo** observado a partir da variável em 187 países, entre 1980 e 2010.

Metodologia do Índice

Os **valores mínimos** são fixados em patamares considerados como de subsistência ou de “zero” naturais, enquanto os **máximos** são aqueles realmente observados, exceto para os anos de escolaridade esperada.

Tabela 1 Máximos e Mínimos das variáveis utilizadas no cálculo do IDH em 2011

Subíndice	Variável	Máximo	Mínimo
Educação	Expectativa de Vida	83,4 Japão (2011)	20,0
	Média de Anos de Escolaridade	13,1 República Tcheca	0
	Anos de Escolaridade Esperado	18 Limitado	0
Renda	PIB per capita(Dólar PPC)	107.721 Qatar	100

Fonte: PNUD

Metodologia para Normalização

Os subíndices de longevidade, educação e renda para cada país são calculados da seguinte forma:

$$Indice = \frac{Valor do país i - Valor mínimo}{Valor máximo - Valor mínimo}$$

Metodologia para Renda

- Suposição: a contribuição da renda para o desenvolvimento humano está sujeita a retornos decrescentes.
- Um real extra de renda, quando a renda é de 10 mil reais, não é um insumo tão importante para o desenvolvimento humano quanto um real extra, quando a renda é de 100 reais.
- Desde 1999 o IDH passou-se a ajustar o indicador renda tomando o seu logaritmo, independentemente do nível de renda. Exemplo:
$$= \frac{\ln(\text{GNIpc}) - \ln(100)}{\ln(75,000) - \ln(100)}$$
- Em comparações internacionais, dado que o poder de compra de US\$ 1 não é o mesmo em países diferentes, os valores dos PIBs per capita devem ser convertidos em dólares pela taxa de câmbio que igualaria o poder de compra do dólar entre os países (paridade do poder de compra - PPC).

Metodologia

O IDH de cada país é obtido pela média geométrica dos subíndices de longevidade, educação e renda. Ponderação: cada um destes indicadores normalizados entra no IDH com o mesmo peso (1/3).

$$IDH = \sqrt[3]{I_{longevidade} \times I_{educação} \times I_{renda}}$$

Natureza dos Dados

Dados Não Estruturados

Dados Semi-Estruturados

Dados Estruturados

Dados Estruturados

- Dados tabulares com estrutura fixa e bem definida;
- Dados organizados em blocos semânticos;
- Baseados em uma estrutura de esquema previamente definido;
- Podem-se realizar consultas com linguagem SQL.

Dados Semiestruturados

- Dados que apresentam uma organização heterogênea;
- Definição da estrutura de esquema flexível;
- A representação do esquema é presente de forma explícita ou implícita.

Exemplos de formatos de dados: XML, JSON.

Formato de dados JSON

- Modelo de armazenamento e transmissões de informações no formato de texto
- Notação de objeto em Java Script;
- Comumente usado em aplicações de Big Data por sua simplicidade em comparação ao XML;
- Suporte em praticamente todas as linguagens de programação.

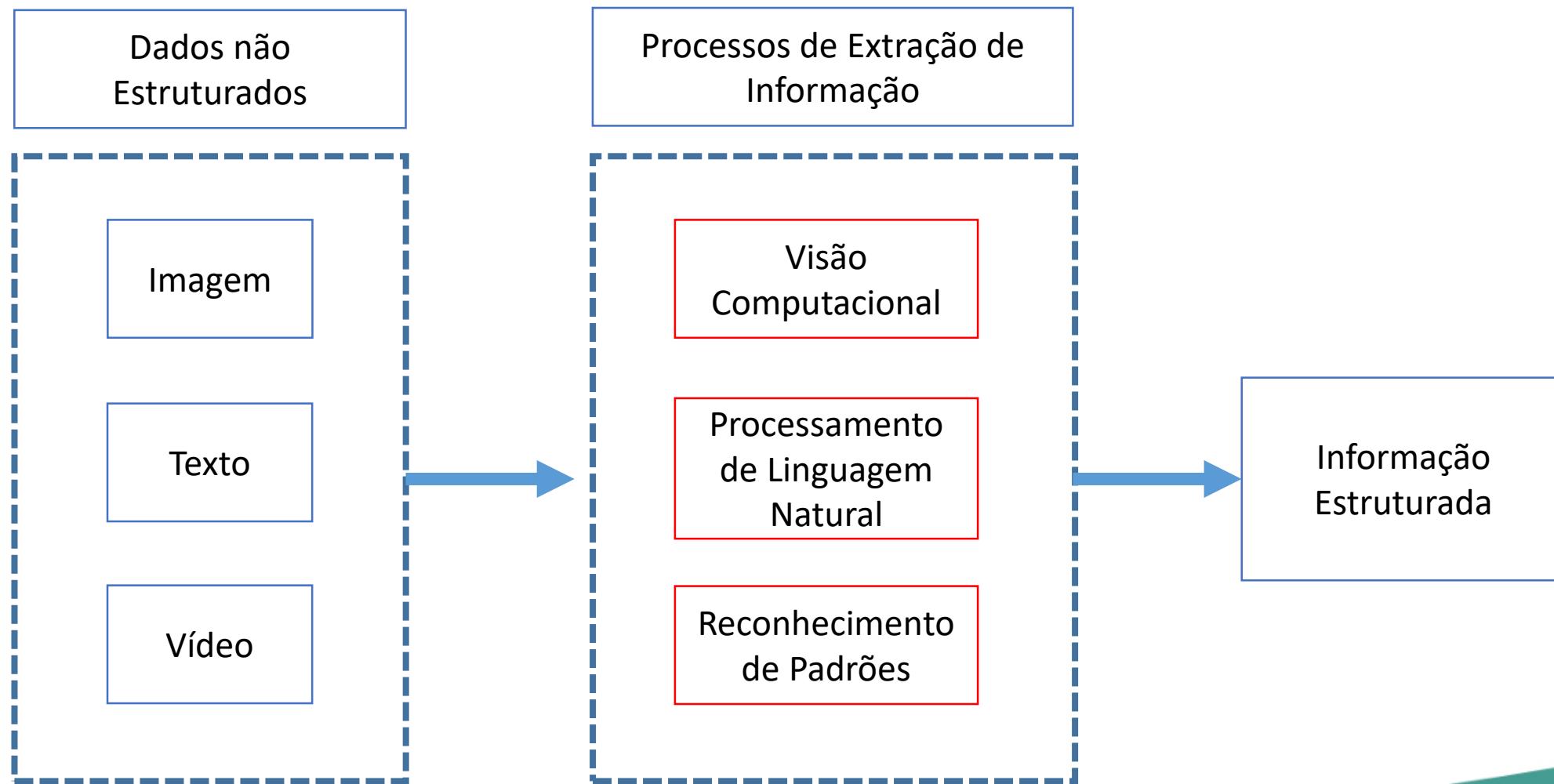
Exemplo JSON

```
{ "Alunos": [  
    { "nome": "João", "notas": [ 8,9,7 ] },  
    { "nome": "Maria", "notas": [ 8,10,7 ] },  
    { "nome": "Pedro", "notas": [ 10,10,9 ] }  
]
```

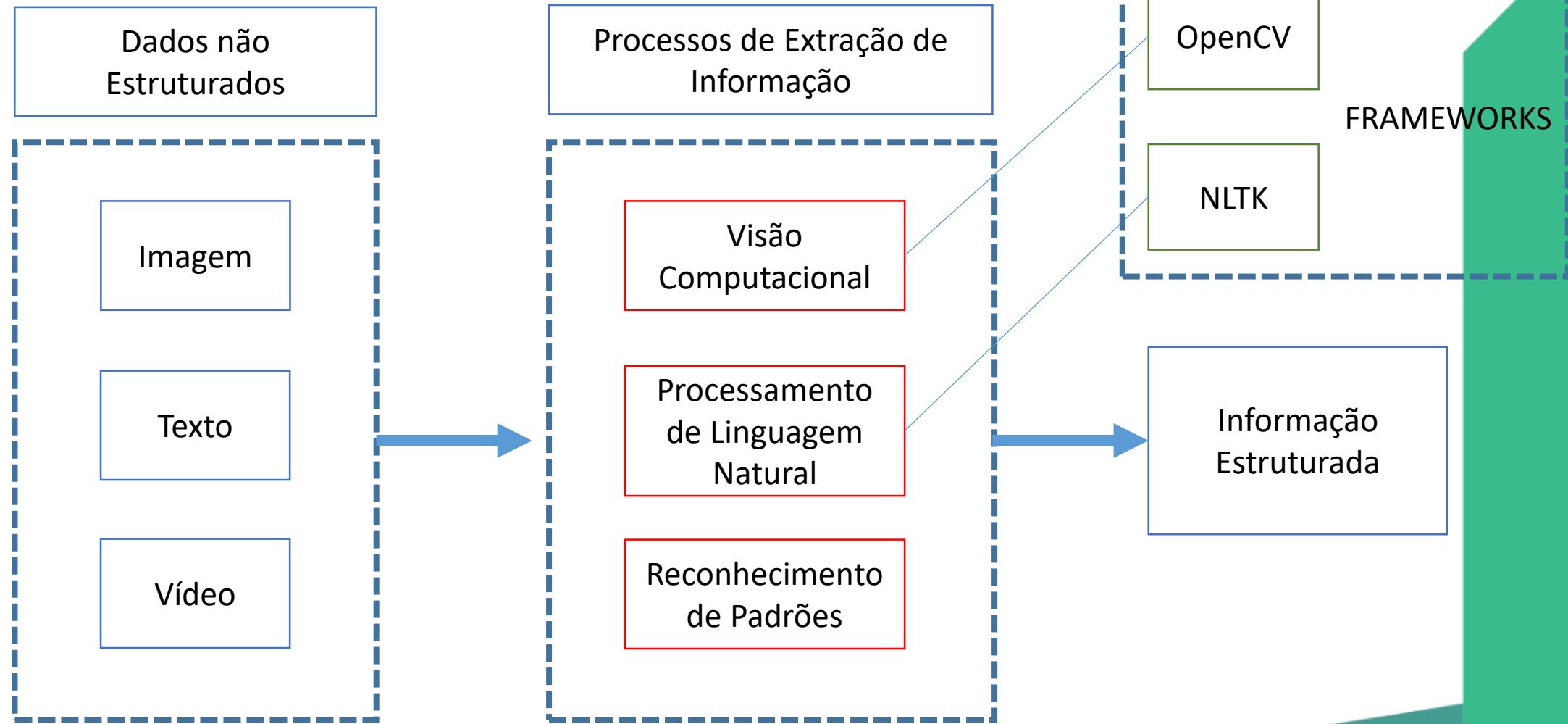
Dados Não Estruturados

- Não possuem uma estrutura definida;
- Normalmente são caracterizados por documentos, textos, imagens, vídeos, entre outros;
- A maior parte dos dados gerados na web e nas organizações seguem esse formato;
- Necessita de pré-processamento para a extração de informações estruturadas para análise.

Extração de Informação



Extração de Informação



Sistemas Gerenciadores de Banco de Dados

- Software responsável por gerenciar o acesso, a manipulação e a organização dos dados;
- Modelos de implementação (relacional, hierárquico, semi-estruturado, chave-valor e etc.)



Características de Transações em SGBD

ATOMICIDADE - A transação deve ter todas as suas operações executadas em caso de sucesso;

CONSISTÊNCIA - A execução de uma transação deve levar o banco de dados de um estado consistente a um outro estado consistente;

ISOLAMENTO - Evitar que transações paralelas interfiram umas nas outras;

DURABILIDADE - Os efeitos de uma transação em caso de sucesso devem persistir no banco de dados.

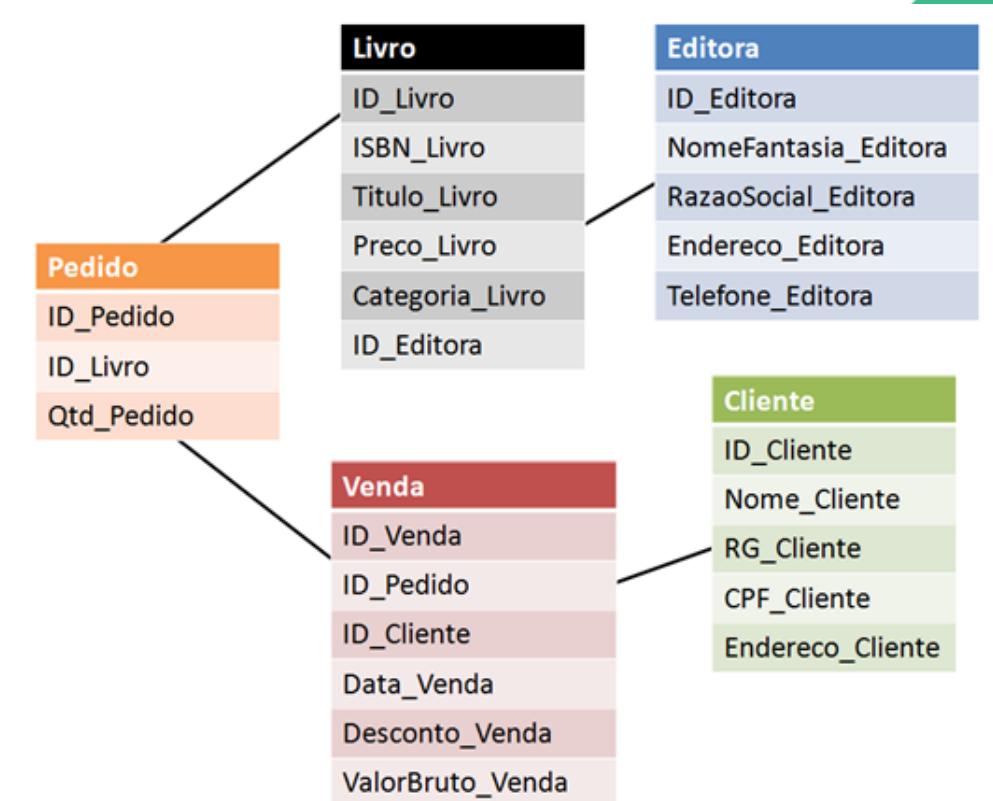
Ontologia

- Representa um conjunto de conceitos de um domínio e os relacionamentos entre esses.
- Indivíduos: os objetos básicos;
- Classes: conjuntos, coleções ou tipos de objetos;
- Atributos: propriedades, características ou parâmetros que os objetos podem ter e compartilhar;
- Relacionamentos: as formas como os objetos podem se relacionar com outros objetos.

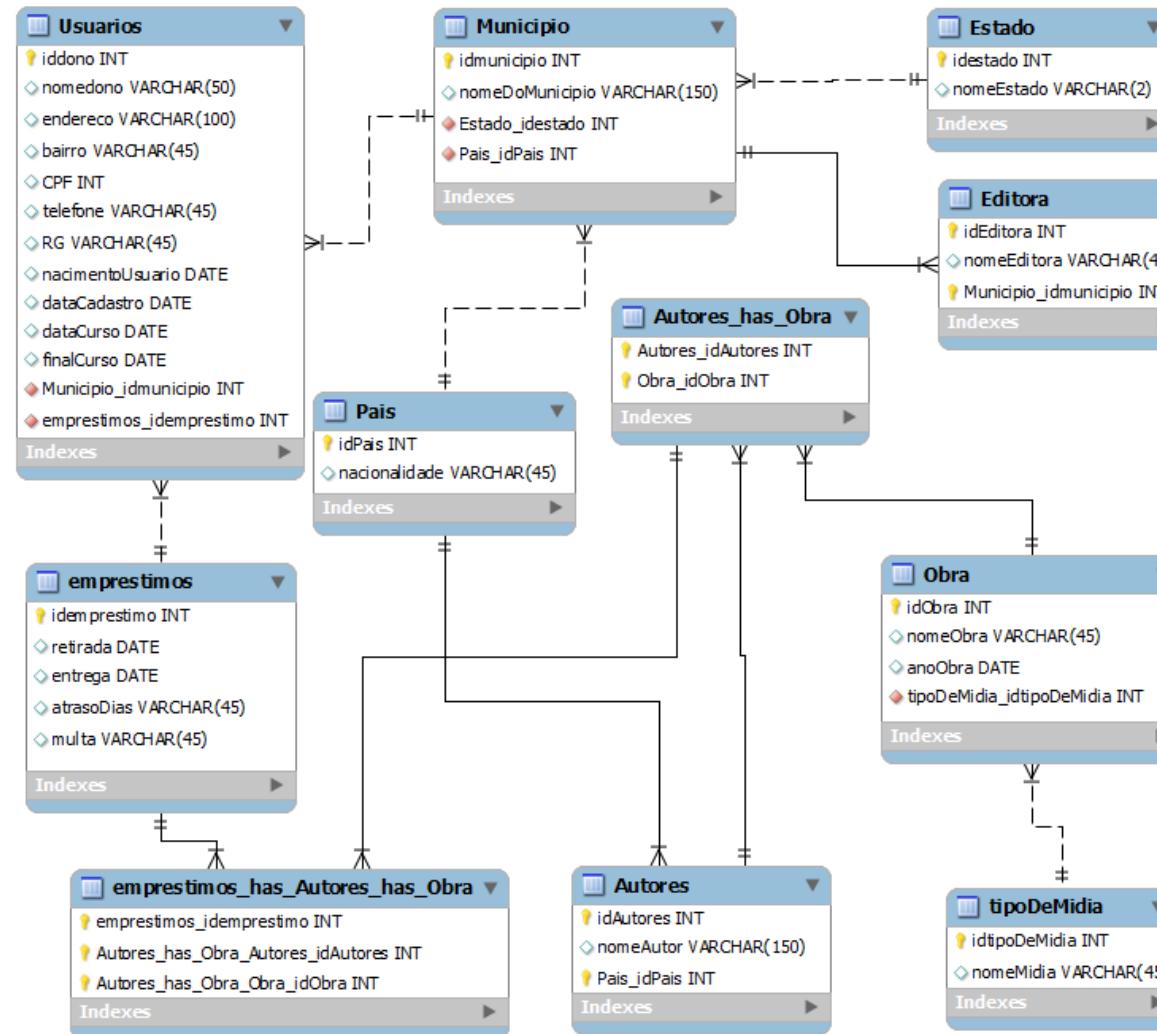
Modelo de Dados

➤ Abstração conceitual que explica as características e comportamento de uma entidade

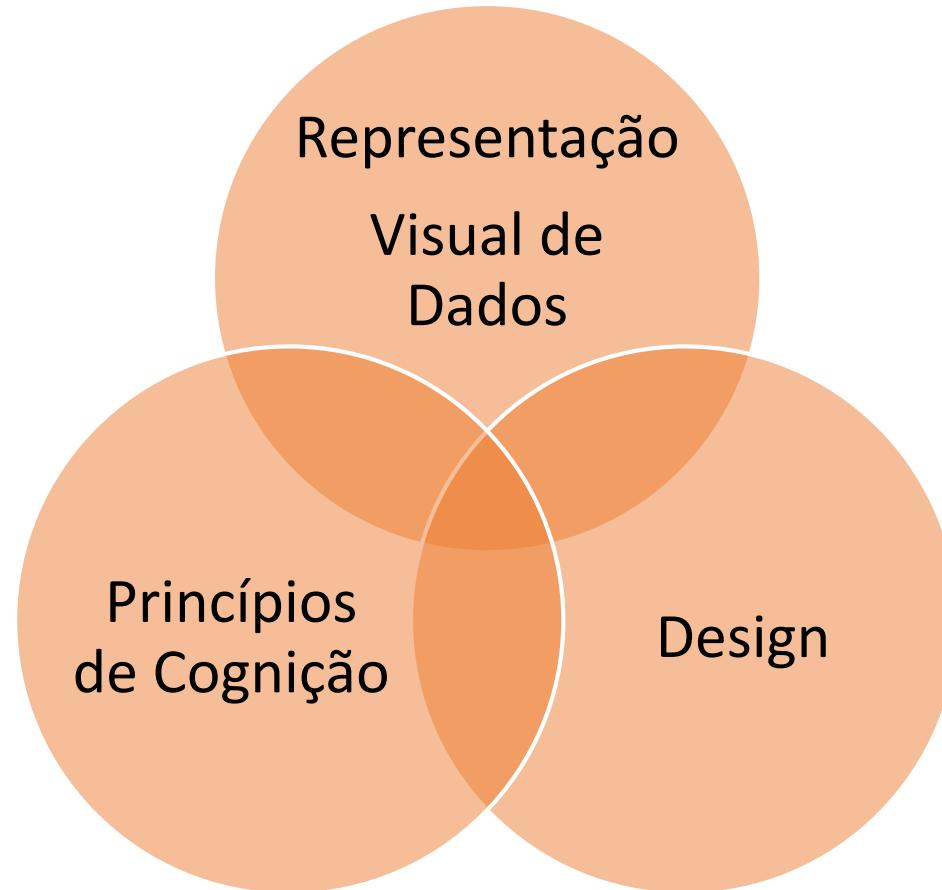
- Relacional;
- Multidimensional.



Modelagem Entidade Relacionamento



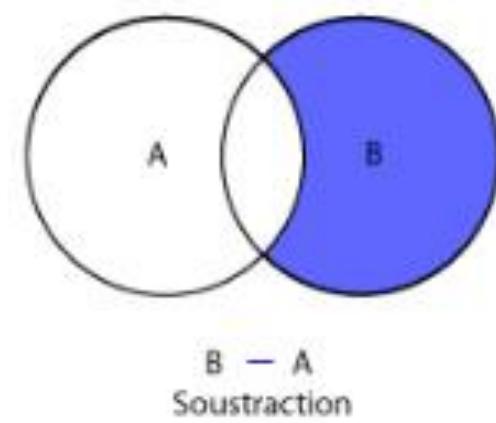
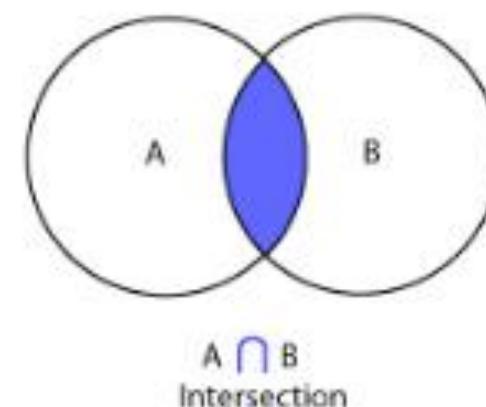
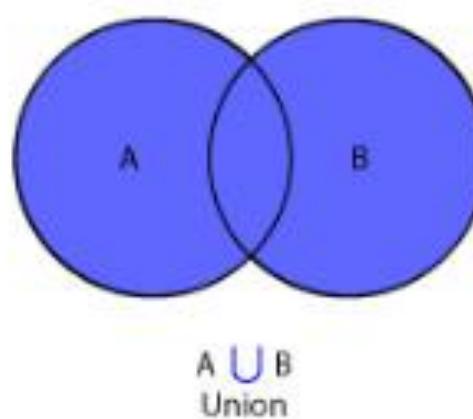
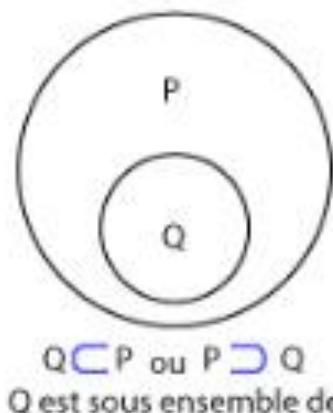
Visual Analytics



Análise de Dados Exploratória

- É uma abordagem para a análise de conjuntos de dados por meio do resumo de suas principais características.
- **Estatística:** box plot, histograma, gráfico de dispersão, Pareto;
- **Business Intelligence:** ferramentas de Data Discovery;
- **Análises de Redes Sociais:** ferramentas de análises de grafos;
- **Mineração de textos:** ferramentas de Visual Text Analytics.

Teoria dos Conjuntos



Álgebra Relacional

- Aplicação de álgebra relacional, lógica de primeira ordem e álgebra de conjuntos em operações sobre conjunto de dados (relações);
- Permite a implementação de mecanismos de consultas em bancos de dados.

Conceitos Básicos

- Relação;
- Túpla;
- Seleção;
- Projeção;
- Junção.

Relação

CPF	Nome	Sexo	Data de Nascimento
000.111.222-33	João de Sousa	M	22/05/1976
111.222.333-44	José da Silva	M	12/02/1980
222.333.444-55	Maria Santos	F	17/08/1991
333.444.555-66	Pedro João	M	03/11/1984
444.555.666-77	Ana Paula	F	29/06/1982
555.666.777-88	Francisco de Assis	M	13/09/1988

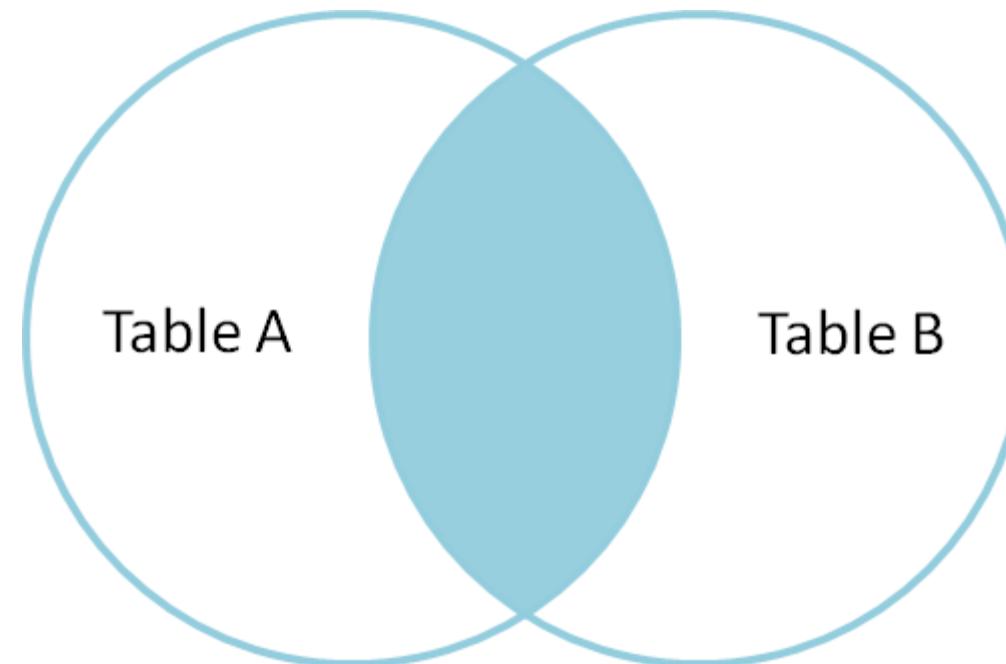
Túpla

CPF	Nome	Sexo	Data de Nascimento
000.111.222-33	João de Sousa	M	22/05/1976
111.222.333-44	José da Silva	M	12/02/1980
222.333.444-55	Maria Santos	F	17/08/1991
333.444.555-66	Pedro João	M	03/11/1984
444.555.666-77	Ana Paula	F	29/06/1982
555.666.777-88	Francisco de Assis	M	13/09/1988

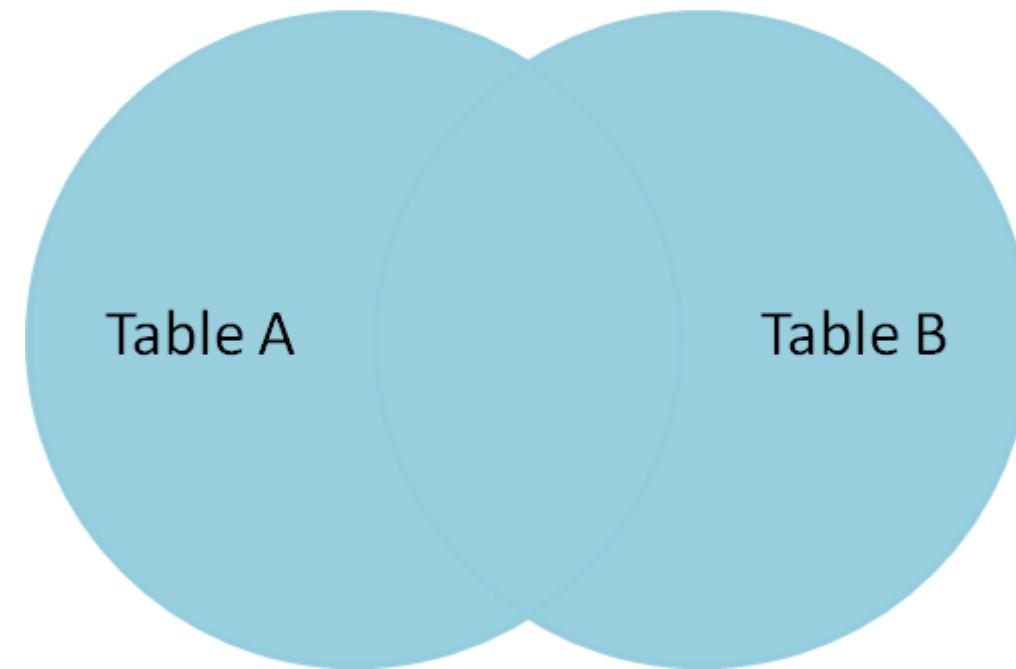
Projeção

CPF	Nome	Sexo	Data de Nascimento
000.111.222-33	João de Sousa	M	22/05/1976
111.222.333-44	José da Silva	M	12/02/1980
222.333.444-55	Maria Santos	F	17/08/1991
333.444.555-66	Pedro João	M	03/11/1984
444.555.666-77	Ana Paula	F	29/06/1982
555.666.777-88	Francisco de Assis	M	13/09/1988

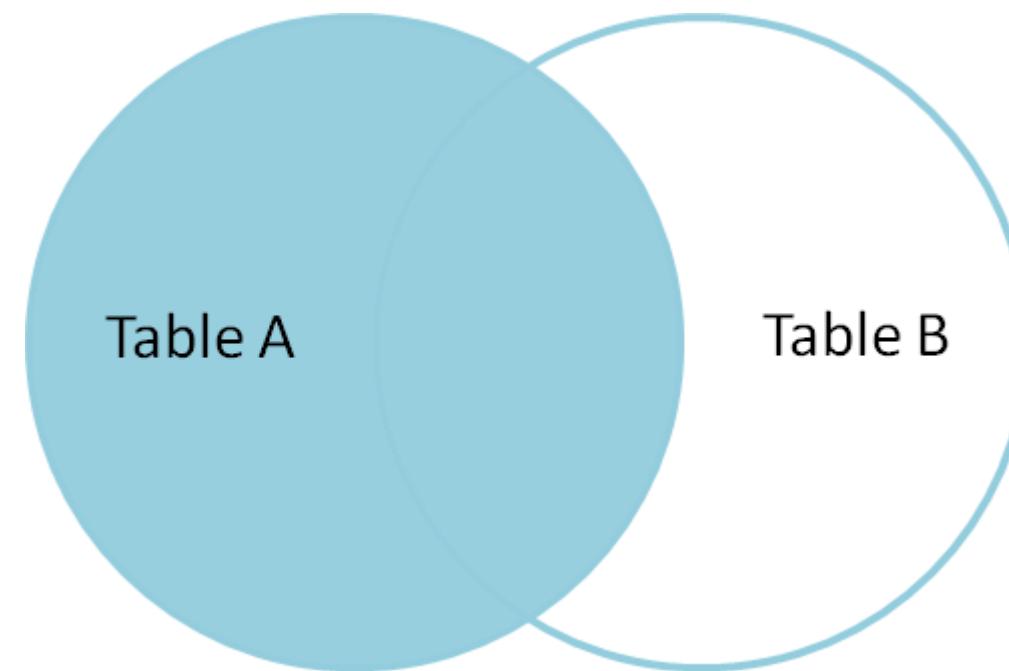
Junção Interna



Junção Externa Completa



Junção Externa Direcionada



Consulta aos Dados - SQL

Linguagem de Consulta Estruturada

SQL ANSI – Interoperabilidade em diversas tecnologias distintas

- ✓ DML - Linguagem de Manipulação de Dados
- ✓ DDL - Linguagem de Definição de Dados
- ✓ **DQL - Linguagem de Consulta de Dados**

Busca por Padrões - Expressões Regulares

- Expressão regular: **método formal de especificar um padrão de texto;**
- **Composição de símbolos, caracteres com funções especiais,** (metacaracteres) que, agrupados entre si e com caracteres literais, formam uma sequência, uma expressão;
- Essa expressão é testada em textos e retorna sucesso caso esse texto obedeça exatamente a todas as suas condições.

Metacaracteres de Expressões Regulares

O circunflexo (^)

Simboliza o início de uma linha, ou seja, ao utilizá-lo associado ao que deve ser retornado, teremos como resultado as linhas que possuem o que especificamos, no início dela;

O Cifrão (\$)

Essa expressão regular procura pela palavra no final da linha, ou seja, é o contrário do metacaractere anterior. Também pode ser usado com o circunflexo (^\$), retornando as linhas em branco;

Metacaracteres de Expressões Regulares

A Lista ([])

Utilizando os colchetes podemos especificar os caracteres que podem aparecer em uma posição. Podemos utilizá-lo para retornar palavras iguais, sejam com a primeira letra maiúscula ou minúscula.

Podemos usar intervalos em lista. Por exemplo, “a-f” é interpretado como “Todas as letras entre a e f”;

O ponto (.)

O ponto é o metacaractere que significa “Qualquer letra”. Ou seja, ao utilizá-lo seguido dos caracteres desejados, serão retornadas as linhas que possuem as letras/número/símbolo informados;

Metacaracteres de Expressões Regulares

As chaves ({})

Utilizando um determinado número entre as chaves, serão retornadas as linhas que possuem a quantidade de caracteres especificados na expressão

As chaves também suportam intervalos, por exemplo “{20, 40}”, procura por linhas que contenham entre 20 e 40 caracteres;

O curinga (.*)

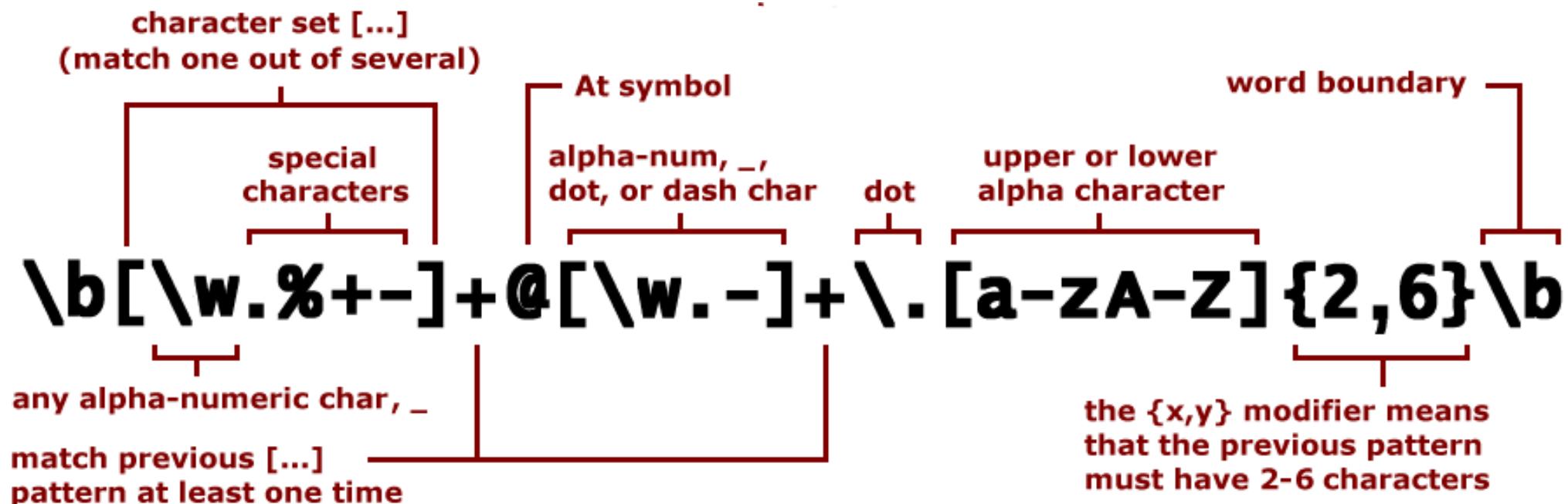
Se quisermos mais de uma condição para as linhas retornadas, podemos usar o curinga;

Metacaracteres de Expressões Regulares

- O ou | (OR)

Se quisermos que uma linha contenha uma condição ou outra, podemos usar o metacaractere citado.

Laboratório Expressões Regulares



Parse: username@domain.TLD (top level domain)

Introdução a Estatística

Metodologia científica para obtenção, organização e análise dos dados



Abusos da Estatística

“Há três espécies de mentiras: mentiras, mentiras deslavadas e estatísticas”

“Os números não mentem; mas os mentirosos forjam números”

“Se torturarmos os dados por bastante tempo, eles acabam por admitir qualquer coisa”

(Benjamin Disraeli)

Dados distorcidos

- Pequenas Amostras;
- Números imprecisos;
- Estimativas por suposição;
- Porcentagem distorcidas;
- Perguntas tendenciosas;
- Más amostras.

Relação Causal

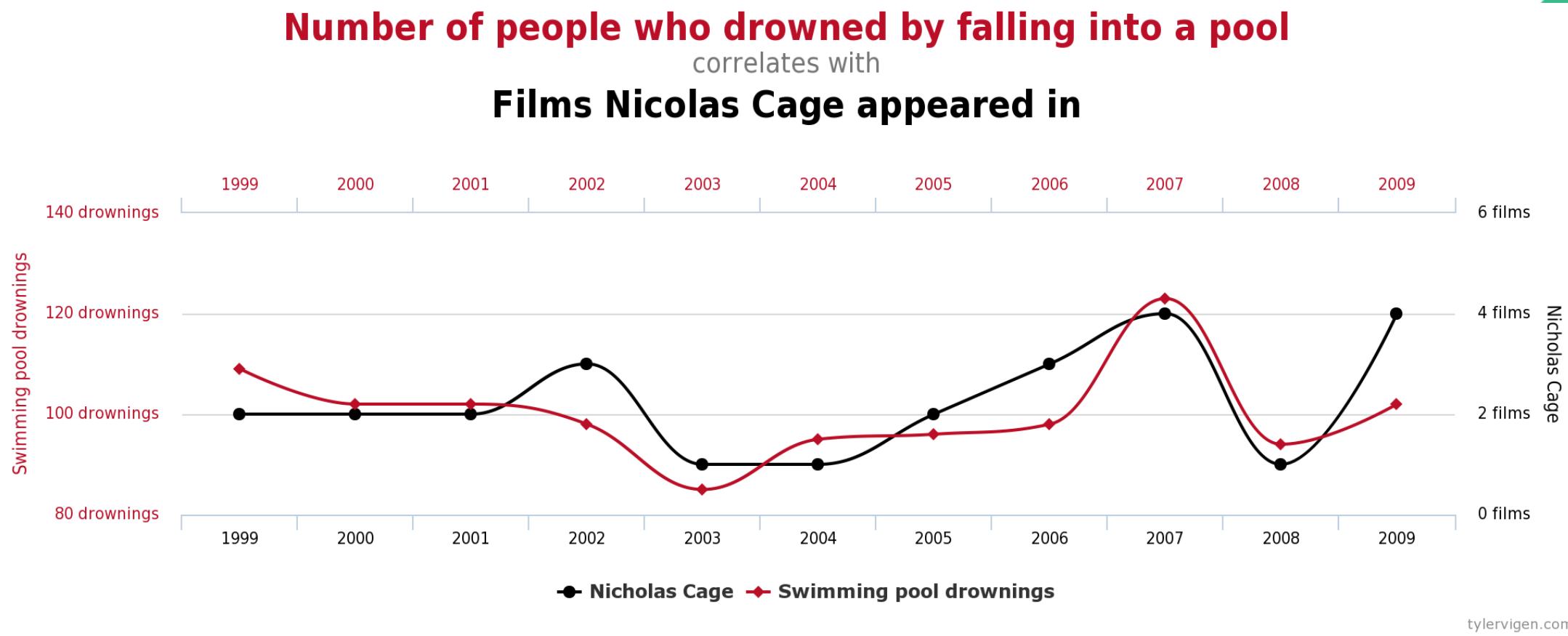
- A Estatística nunca dá certezas, dá apenas probabilidades baseadas em pressupostos;
- Os resultados estatisticamente significativos podem não ter relevância;
- Os resultados da Estatística poderão dar-nos uma ideia sobre a eventual associação entre variáveis, mas nunca nos revela a relação de causalidade.

Associações Espúrias

- São resultados de associações entre variáveis estatisticamente significantes que não refletem sentido causal;
- Muito comum em análises de Big Data que possui uma grande diversidade de informação gerando coincidências para associações espúrias.

<http://www.tylervigen.com/>

Exemplo



Conceitos Iniciais

População – Conjunto de indivíduos ou objeto que apresentam uma característica em comum.

Censo – Coleção de dados relativos a todos os elementos da população.

Amostra – Parte representativa da população.

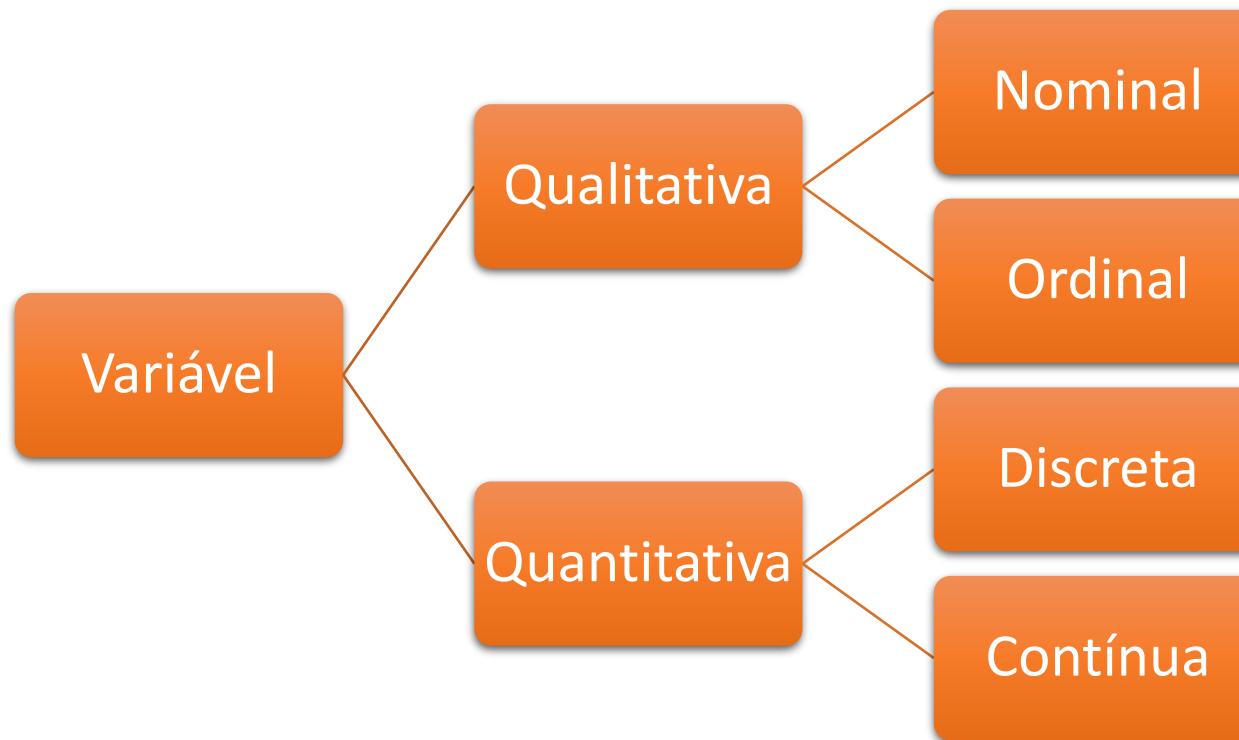
Conceitos Iniciais

Estatística Descritiva – Trata-se da coleta, organização, descrição dos dados, cálculo e interpretação de coeficientes.

Estatística Inferencial – Trata-se da análise e a interpretação dos dados, associado a uma margem de incerteza.

Conceito de variável

➤ Atributo mensurável que tipicamente varia entre indivíduos.



Variáveis Qualitativas

Qualitativa Nominal – os valores representam atributos ou qualidades mas não tem uma relação de ordem entre eles;

Ex: Sexo, Grupo sanguíneo, Raça

Qualitativa Ordinal – os valores representam atributos ou qualidades mas incluem uma relação de ordem;

Ex: Classe Social, Grau de Instrução

Variáveis Quantitativas

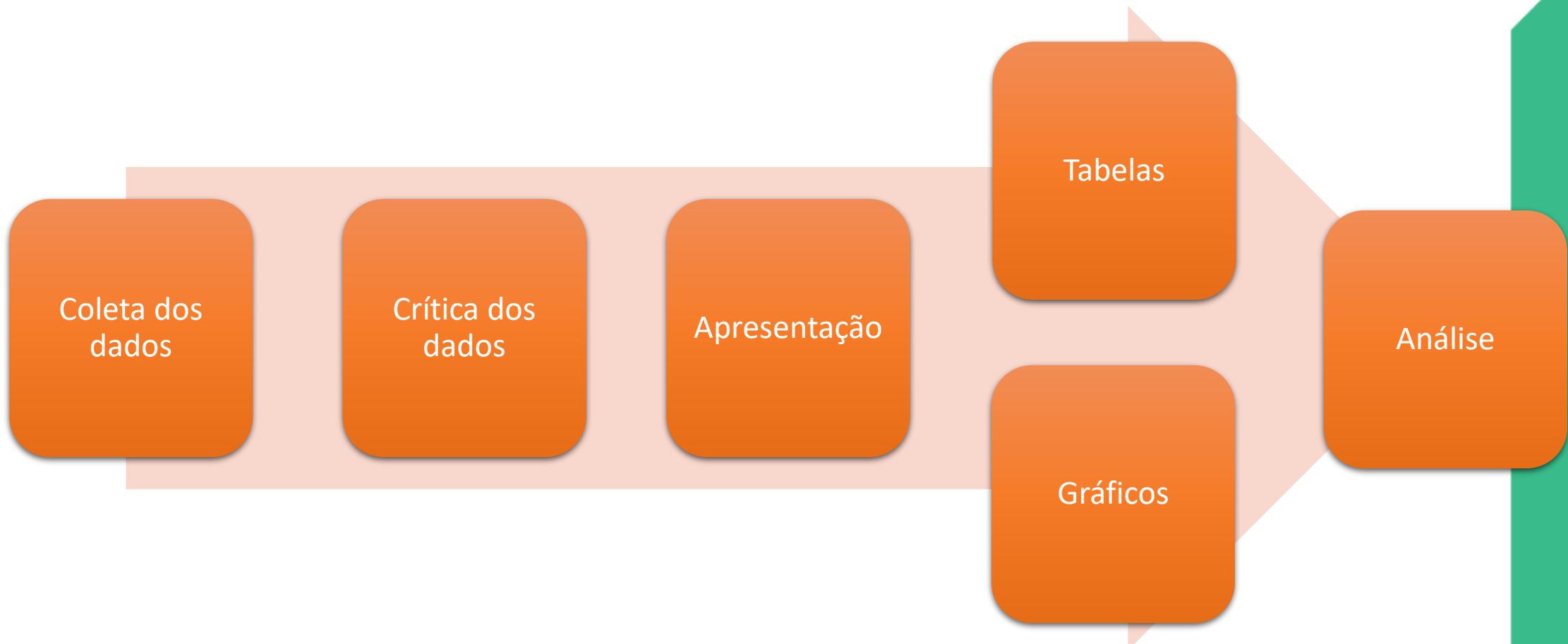
Quantitativa Contínua – os valores são medidos numa escala métrica e onde todos os valores fracionários são possíveis;

Ex: Altura, Peso, Temperatura

Quantitativa Discreta – os valores resultam de um conjunto finito, enumerável de valores possíveis.

Ex: Número de Filhos

Processo de Estatística Descritiva



Coleta dos dados

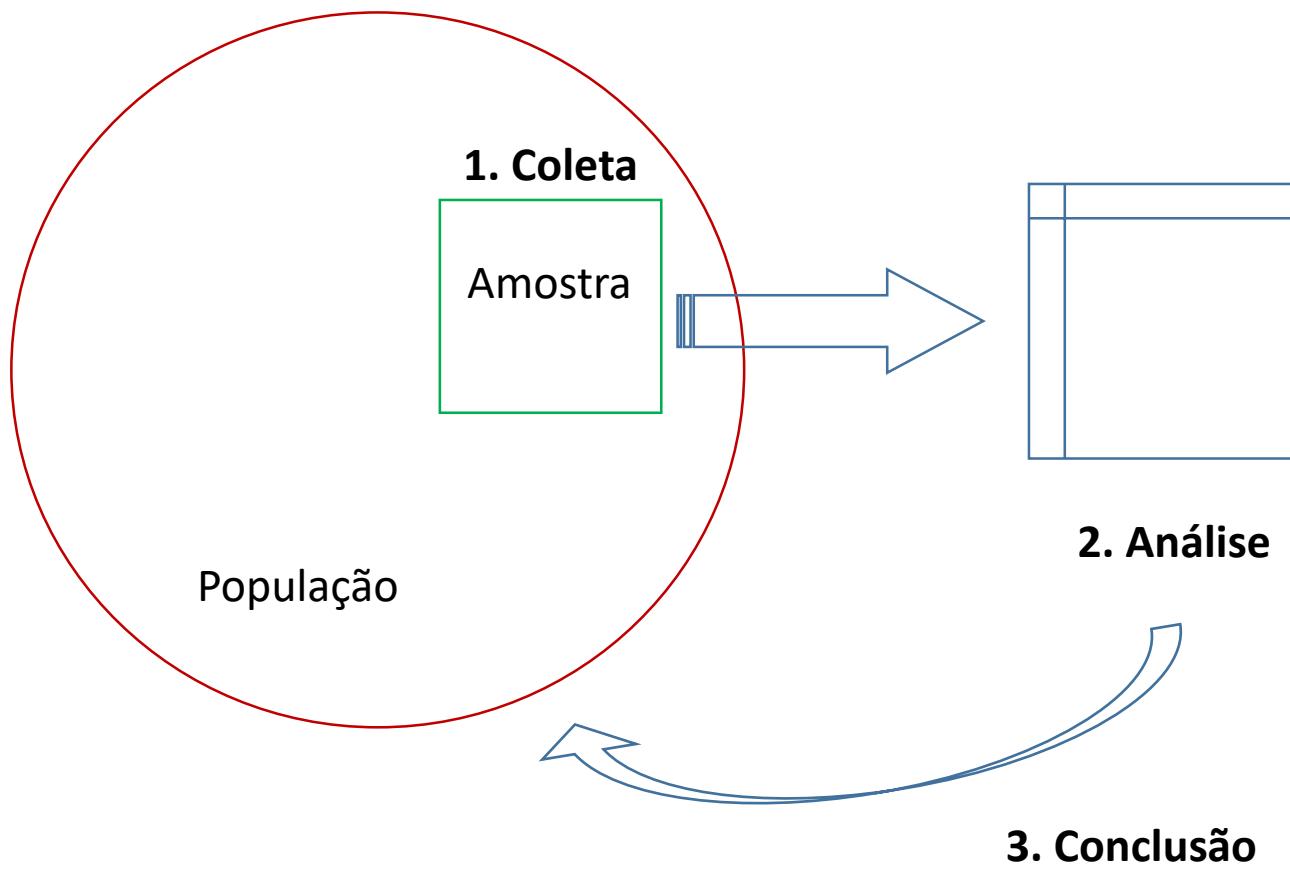
Coleta direta – Dados obtidos da fonte originária;

Coleta indireta – Dados inferidos a partir dos elementos conseguidos pela coleta direta.

Apresentação dos dados

- Organização do conjunto de dados de maneira prática e racional;
- Séries Estatísticas;
- Apresentação por meio de tabelas e gráficos.

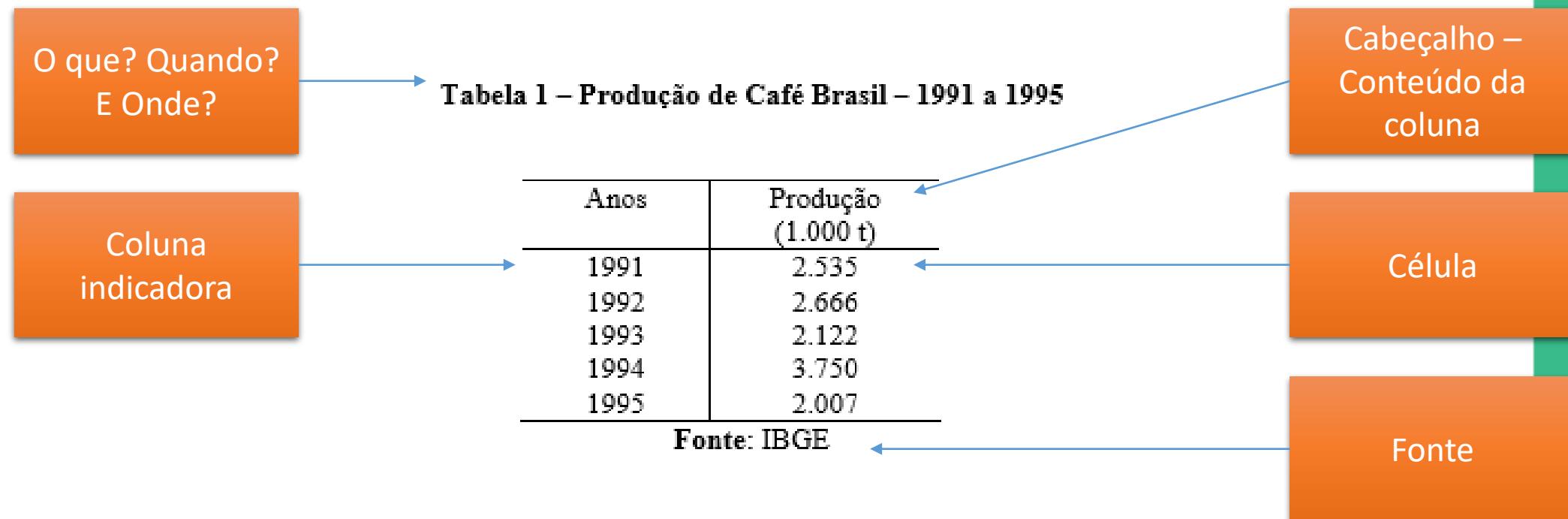
Visão Geral da Estatística



Técnicas de Amostragem



Normas para construção de tabelas (IBGE)



Sinal Convencional

- Dado numérico igual a zero
- ... Quando não temos dados
- ? Quando temos dúvida na informação
- 0 Quando o valor for muito pequeno

Séries Estatísticas

- Coleção de dados estatísticos referidos a uma mesma ordem de classificação quantitativa

Série Temporal/Histórica - O fator cronológico é o caráter variável;

Série Geográfica - O local do fenômeno é o caráter variável;

Série Específica/Categórica – O caráter variável é o fenômeno.

Dados Brutos

- Conjunto de dados numéricos obtidos após a crítica dos valores coletados. Exemplo:

24 23 22 28 35 21 23 23 33 34

25 21 25 36 26 22 30 32 25 26

33 34 21 31 25 31 26 25 35 33

ROL

- É o arranjo dos dados brutos em ordem de frequência crescente ou decrescente. Exemplo:

21 21 21 22 22 23 23 24 24

25 25 25 25 26 26 28 30 31

31 32 33 33 33 34 35 35 36

```
dados <- c(21,21,21,22,22,23,23,24,24,25,25,25,25,26,26,28,30,31,31,32,33,33,33,34,35,35,36)
```

Frequência Absoluta (F_i)

- É o número de vezes que o elemento aparece na amostra, ou o número de elementos pertencentes a uma classe.

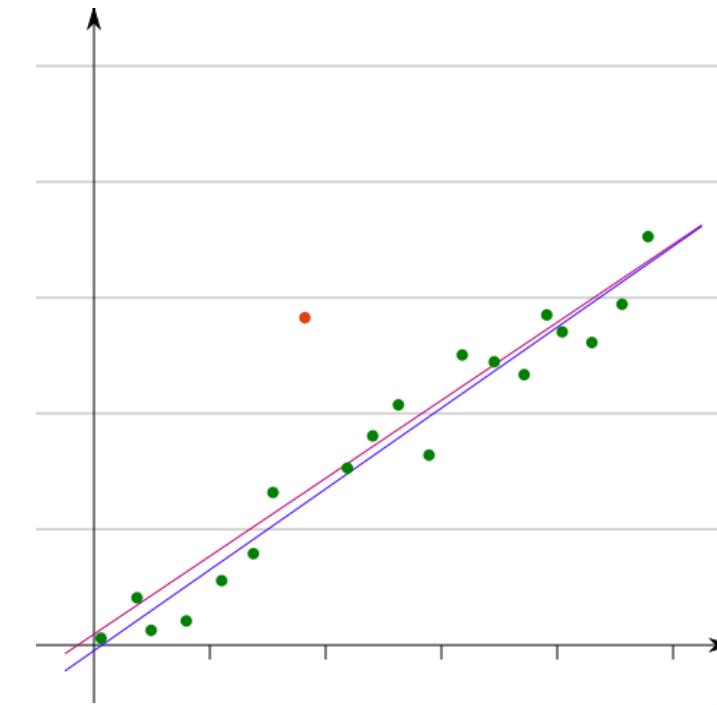
$$F_{(21)} = 3$$

- O comando `table(dados)` irá retornar a distribuição de frequências inseridas no banco de dados.

<code>dados</code>	21	22	23	24	25	26	28	30	31	32	33	34	35	36
	3	2	2	2	4	2	1	1	2	1	3	1	2	1

Outlier

- **Valor discrepante ou atípico.** É uma observação que apresenta um grande afastamento das demais da série (que está "fora" dela), ou que é inconsistente. Implica em prejuízos para interpretação dos resultados dos testes estatísticos aplicados às amostras. Exemplo:



Medidas de Tendência Central

➤ São valores que representam o conjunto de dados. Exemplos:

- Média;
- Moda;
- Mediana;
- Separatriz ou Quantil.

Média

- Corresponde ao somatório de todos os valores obtidos dividido pela quantidade de valores extraídos. A média é altamente influenciada por outliers. Exemplos:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

`mean(dados)`

`> 27.5`

Moda

➤ É o valor que ocorre com maior frequência em uma série de valores.

Pode ser encontrada pelo comando `table()` ou pelo comando:

```
subset(table(dados), table(dados)==max(table(dados)))
```

> 25

4

Mediana

➤ É o **valor central de um conjunto de dados ordenados**. Também pode ser chamada de 2º quartil, 5º decil, 50º percentil, etc. Vale ressaltar que a Mediana não é influenciada por outliers e os dados precisam estar em rol.

$$M_{ímpar} = \text{posição } \frac{n + 1}{2}$$

$$M_{par} = \text{posição entre } \left(\frac{n}{2}\right) \text{ e } \left(\frac{n}{2} + 1\right)$$

`median(dados)`
`> 26`

Medidas de Separatrizes ou Quantis

- São pontos estabelecidos em intervalos regulares a partir da Função Distribuição Acumulada (FDA) de uma variável aleatória. Dividem os dados ordenados em **q** subconjuntos de dados de dimensões iguais.
- Como regra geral, pode-se utilizar o comando `quantile()` para os quartis, decis e percentis. Basta inserir no comando um vetor dos valores percentuais desejados.

Medidas de Separatrizes ou Quantis

```
quantile(dados, seq(primeiroquantil, últimoquantil, distância))
```

quartil

```
quantile(dados, seq(0.25, 0.75, 0.25))
```

decil

```
quantile(dados, seq(0.1, 0.9, 0.1))
```

percentil

```
quantile(dados, seq(0.01, 0.99, 0.01))
```

Medidas de Dispersão

- Indicam se os valores estão relativamente próximos um dos outros ou separados em torno de uma medida de posição: a média.
- Amplitude Total;
- Variância;
- Desvio Padrão;
- Coeficiente de Variação.

Amplitude Total (Range)

- É a diferença entre o maior e menor valor observado.

$$A = X_{max} - X_{min}$$

$\text{max}(\text{dados}) - \text{min}(\text{dados})$

> 15

Variância

- Baseia-se nos desvios em torno da média, indicando "o quanto longe" em geral os seus valores se encontram da média.

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

`var(dados)`

`> 25.02`

Desvio-Padrão

➤ Ele mostra o quanto de variação ou "dispersão" existe em relação à média. Um baixo desvio padrão indica que os dados tendem a estar próximos da média, um desvio padrão alto indica que os dados estão espalhados por uma gama de valores. É a raiz quadrada da variância

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

`sd(dados)` ou `sqrt(var(dados))`
`> 5.002`

Coeficiente de Variação

- É uma medida de dispersão relativa, empregada para estimar a precisão de experimentos e representa o desvio-padrão expresso como porcentagem da média. Sua principal qualidade é a capacidade de comparação de diferentes distribuições

$$C_v = \frac{s}{\bar{x}}$$

`sd(dados)/mean(dados)`
`> 0.18`

Histograma

- É a representação gráfica, em colunas, de um conjunto de dados previamente tabulado e dividido em classes uniformes. A construção de histogramas tem caráter preliminar em qualquer estudo e é um importante indicador da distribuição de dados

`hist(dados)`

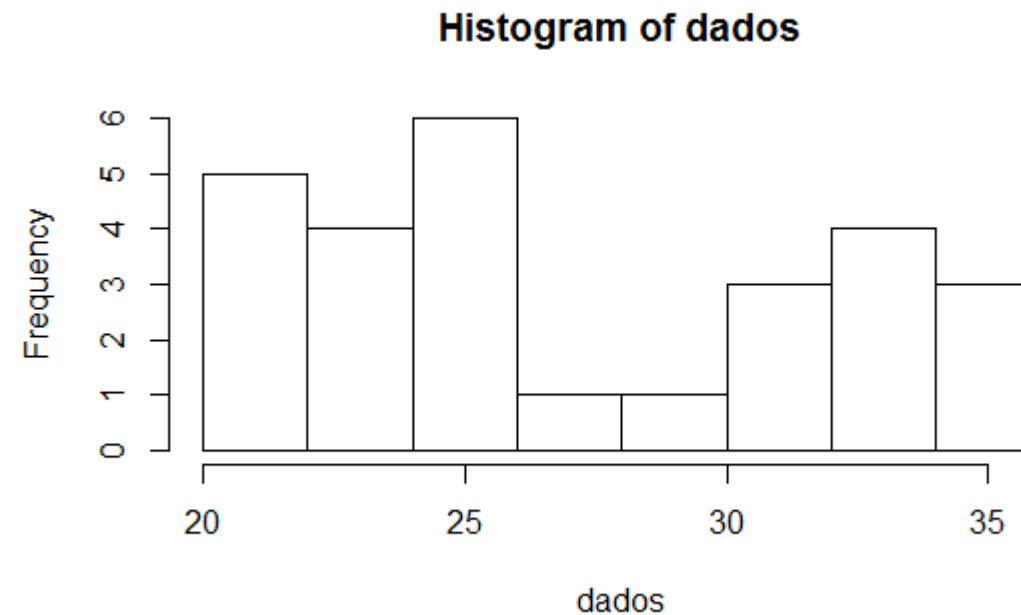
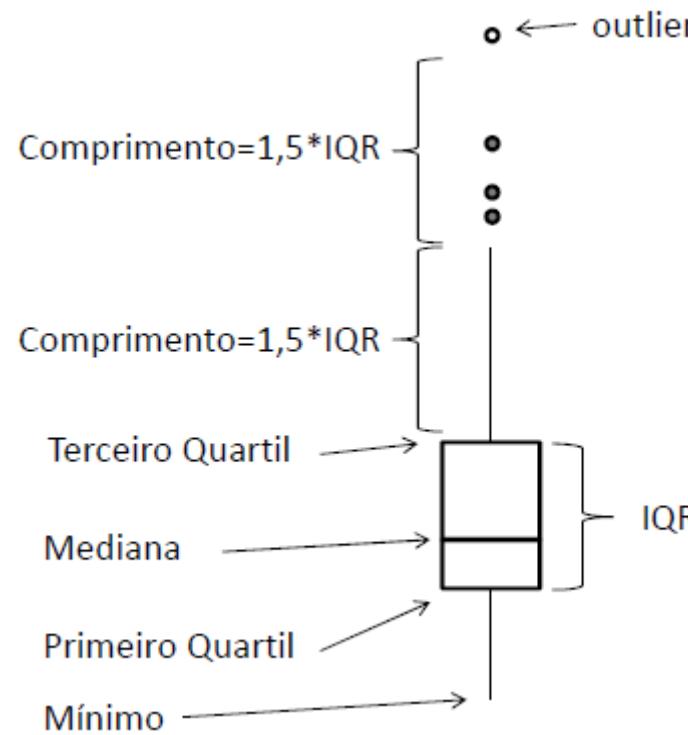
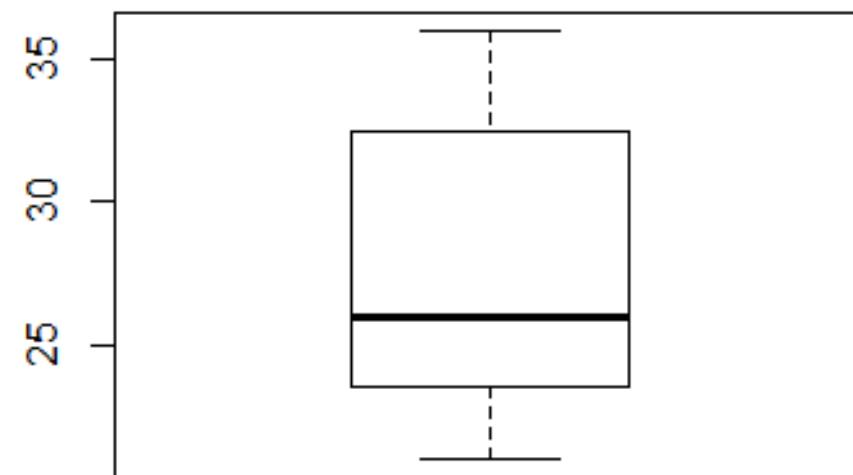


Gráfico de Caixa (Box Plot)

- Permite avaliar visualmente o comportamento de um conjunto de dados



boxplot(dados)



Séries Temporais

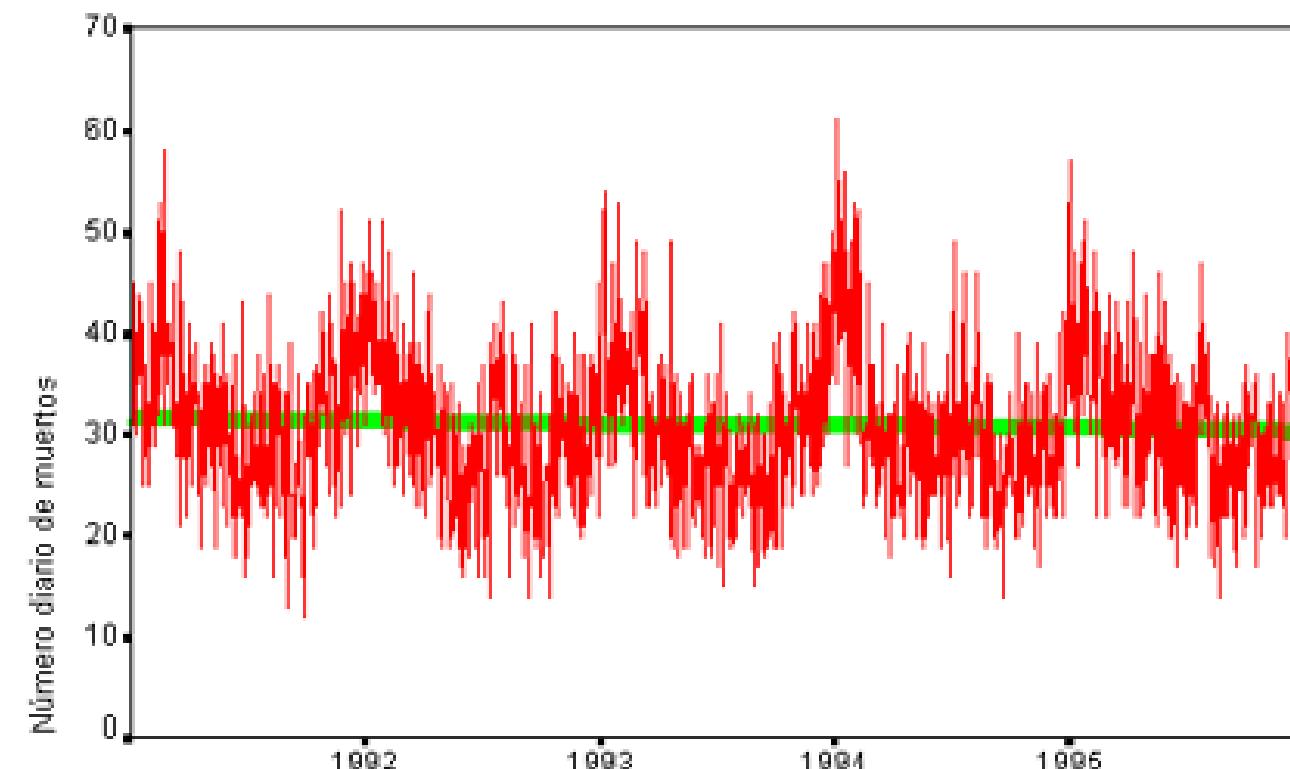
➤ “Uma série temporal é um **conjunto de observações ordenadas no tempo**, registrado em períodos regulares”.

Exemplos:

- Valores diários de poluição em uma cidade;
- Acidentes ocorridos nas rodovias da cidade de São Paulo durante um mês.

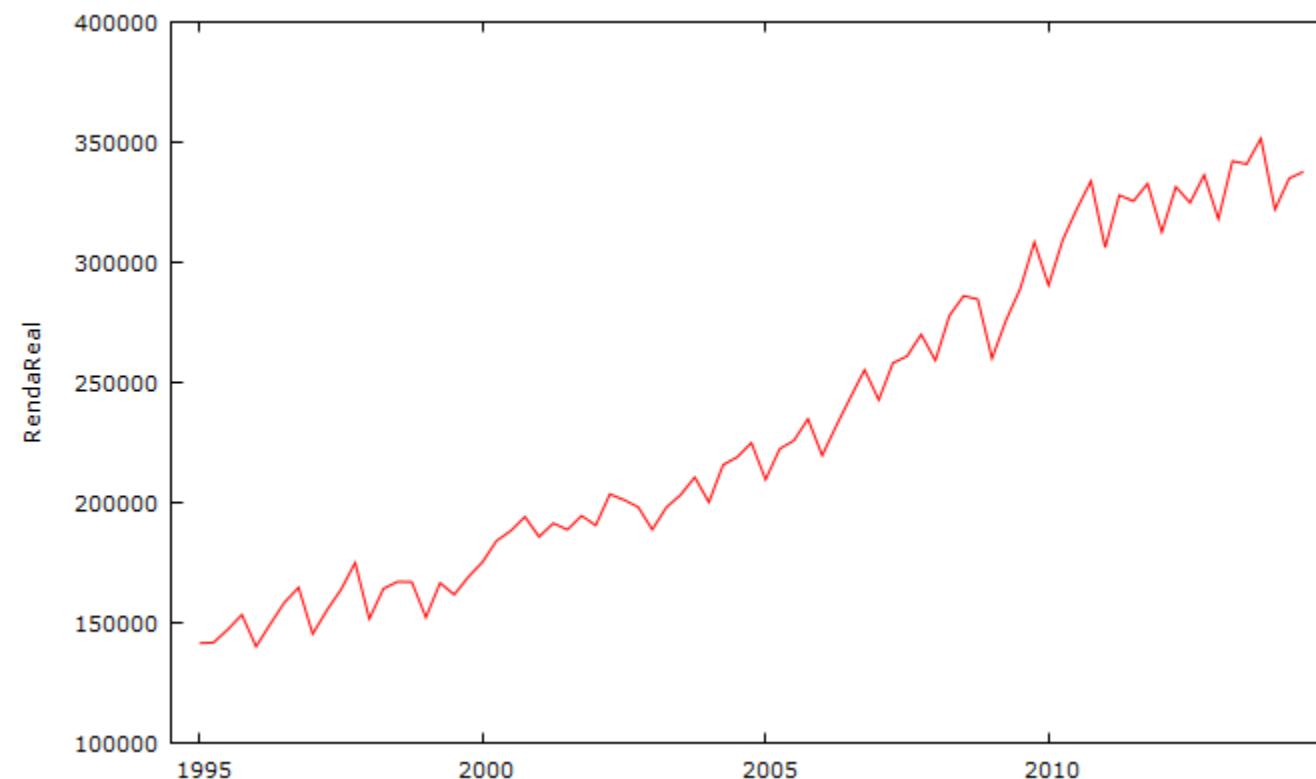
Característica de uma Série Temporal

Série Temporal Estacionária – quando a série se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável.



Característica de uma Série Temporal

Série Temporal Não Estacionária - quando a série se desenvolve no tempo apresentando uma tendência de inclinação.



Decomposição da Série Temporal

Tendência - descreve o comportamento da variável retratada na série temporal a longo prazo. (utilizar em previsões, removê-la da série para facilitar a visualização das outras componentes, identificar o nível da série crescente ou decrescente);

Ciclos - Os ciclos são caracterizados pelas oscilações de subida e de queda nas séries, de forma suave e repetida, ao longo da componente de tendência;

Sazonalidade - são as oscilações de curto prazo, que ocorrem sempre dentro de um período específico, e que se repetem sistematicamente;

Análise de Séries Temporais

“O objetivo da análise de séries temporais é identificar padrões não aleatórios na série temporal de uma variável de interesse, e a observação deste comportamento passado pode permitir fazer previsões sobre o futuro, orientando a tomada de decisões”.

- Exemplo:
 - Qual será demanda de sorvete em períodos de verão?
 - Qual será o IPCA para os próximos meses?

Modelos para Analises de Séries Temporais

- Box-Jenkins:
 - ARMA
 - ARIMA
- Suavização Exponencial;
- Regressão Linear.

Metodologia Box-Jenkins - ARMA

- Um modelo ARMA é a combinação de dois modelos o AR(Autoregressive) e o Moving Average).
- **AR:** Um modelo **AR** expressa uma série temporal como uma função linear dos seus valores passados.

$$y(t) = a(1)*y(t-1) + e(t)$$

- **MA:** O modelo *média móvel* (**MA**) é uma forma do modelo **ARMA** em que a série temporal é tomada como uma média móvel (pesos desiguais) de uma série de choques aleatórios $e(t)$.

$$y(t) = e(t) + c(1)*e(t-1)$$

- **ARMA:** Por incluir ambos os tipos de termos de defasagens, chegamos ao que é chamado de *média móvel auto-regressiva*, ou modelos **ARMA**.

$$y(t) = d + a(1)*y(t-1) + e(t) - c(1)*e(t-1)$$

Metodologia Box-Jenkins - ARIMA

- **ARIMA** significa Autoregressive – Integrated - Moving Average. A letra "I" (Integrado) indica que a modelagem da série temporal a transformará numa série estacionária. **ARIMA** representa três tipos diferentes de modelos: Ele pode ser um modelo **AR** (*autoregressivo*), ou um modelo **MA** (*moving average*), ou um modelo **ARMA** que inclua ambos os termos AR e MA.
- O propósito da modelagem **ARIMA** é estabelecer uma relação entre o valor presente de uma série temporal e seus valores passados de modo que as previsões possam ser feitas somente com base nos valores passados.
- **Exigências do Modelo:**
 - ✓ Estacionariedade:
 - ✓ Diferenciação:

Metodologia de Suavização Exponencial

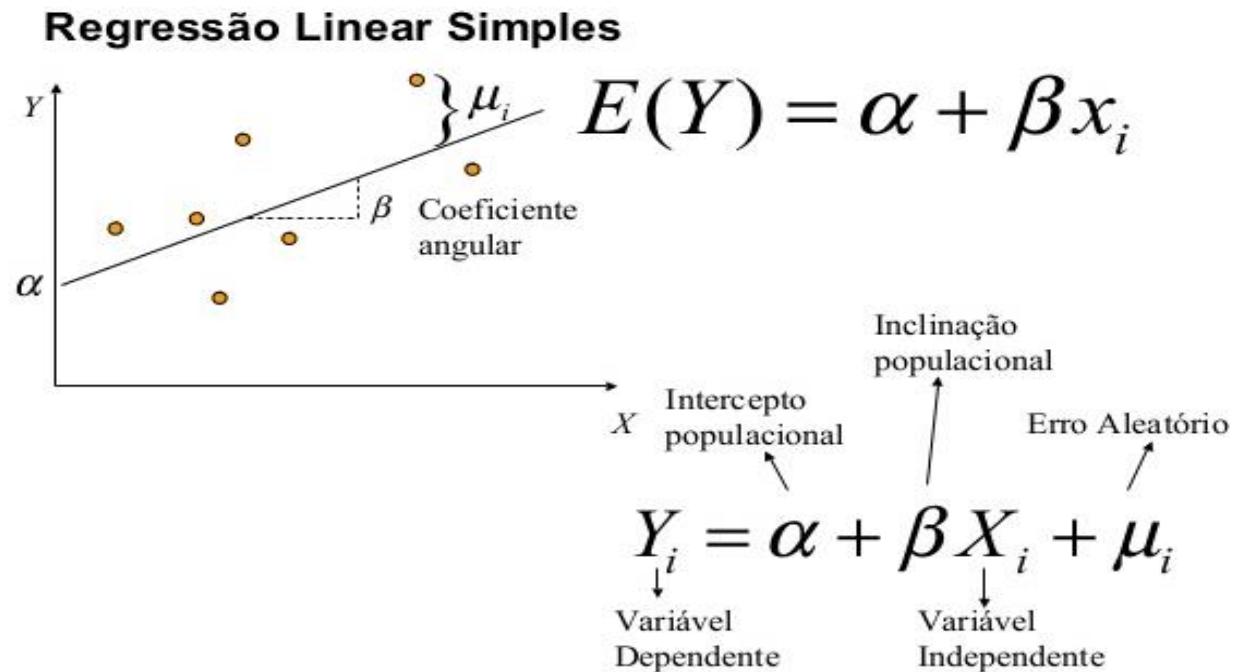
- O método de Suavização Exponencial é utilizado para realizar projeções em séries temporais. É um método que tem por característica aplicar pesos desiguais aos valores passados da série temporal, sendo que os pesos decaem de forma exponencial.
- **Suavização Exponencial Simples (SES):** Modelos sem Tendência e sem Sazonalidade.

$$\hat{Y} = \alpha Y_{t-1} + (1 - \alpha) \hat{Y}_{t-1}$$

- **Suavização Exponencial de Holt (SEH):** Modelos com Tendência e sem Sazonalidade.
- **Suavização Exponencial de Holt -Winters (SEHW):** Modelos com Tendência e com Sazonalidade.

Metodologia de Regressão Linear

- Os modelos de regressão linear relacionam uma variável dependente ou variável de resposta Y, a uma ou mais variáveis explicativas ou independentes X.



Metodologia de Regressão Linear

➤ Os objetivos de um modelo de regressão são:

- Estudar a relação entre a variáveis, para testar se existi causalidade (linear) entre elas;
- Ex: Consumo X Renda.
- Possibilitar análises de cenários;
- Ex: Caso a renda aumente em 1 unidade quanto aumenta o consumo em determinada região.
- Permitir uma eventual previsão da variável dependente.
- Ex: Com o aumento do salario mínimo em 8% o consumo ira aumenta x% no ano que vem.

Frameworks de Análise de dados em Python

- **NumPy** – Suporte a funções matemáticas para trabalhar com vetores e matrizes;
- **SciPy** – Suporte a funções matemáticas avançadas para aplicação em ciência e engenharia;
- **Matplotlib** – Biblioteca para a produção de gráficos complexos;
- **Pandas** – Biblioteca de alto-nível para a estrutura e analise de dados;
- **Statsmodels** – Permite explorar os dados com modelos e testes estatísticos;
- **Scikit-learn** – Suporte a funções de machine learn para data mining.

Linguagem R

- O R é uma plataforma destinada a cálculos estatísticos e construção de gráficos.
- Existe uma grande quantidade de bibliotecas disponíveis com funções específicas de diversas áreas de conhecimento.

Python x R

	Python	R
Usabilidade	Codificação e tratamento de erros mais fácil.	Modelos estatísticos em poucas linhas de código.
Padronização	Mais bem definido.	Maior flexibilidade nas funcionalidades.
Há Interoperabilidade entre as duas linguagens.		
Tratamento de dados	Não é específica para análise de dados, porém, está em constante evolução no aspecto.	Grande número de pacotes com modelos, fórmulas e testes estatísticos.
Pontos positivos	Indicada para criação de scripts e automatização de regras para mineração de dados.	Indicada para prototipagem e para análise estatística.
	Fácil integração em um fluxo de trabalho de produção de desenvolvimento.	Possibilita a análise de diferentes tipos de estatística.
Pontos negativos	Não é tão completo para análise estatística.	É mais difícil de integrar a um fluxo de desenvolvimento de produção.
	Curva de aprendizado mais acentuada.	Mais adequado para tarefas do tipo “consultoria”.
Propósitos	Permite a utilização da linguagem por pessoas com diferentes objetivos.	Desenvolvida por estatísticos, para estatísticos.
Desempenho	Alto desempenho/velocidade.	Baixo desempenho/Velocidade.

Descoberta de Conhecimento em Dados - KDD

- O termo KDD, Knowledge Discovery, in Databases foi formalizado em 1989 em referência ao amplo conceito de procurar conhecimento a partir de base de dados;
- O KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões comprehensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados;
- A Mineração de Dados é uma etapa dentro do processo de KDD.

Processo de KDD

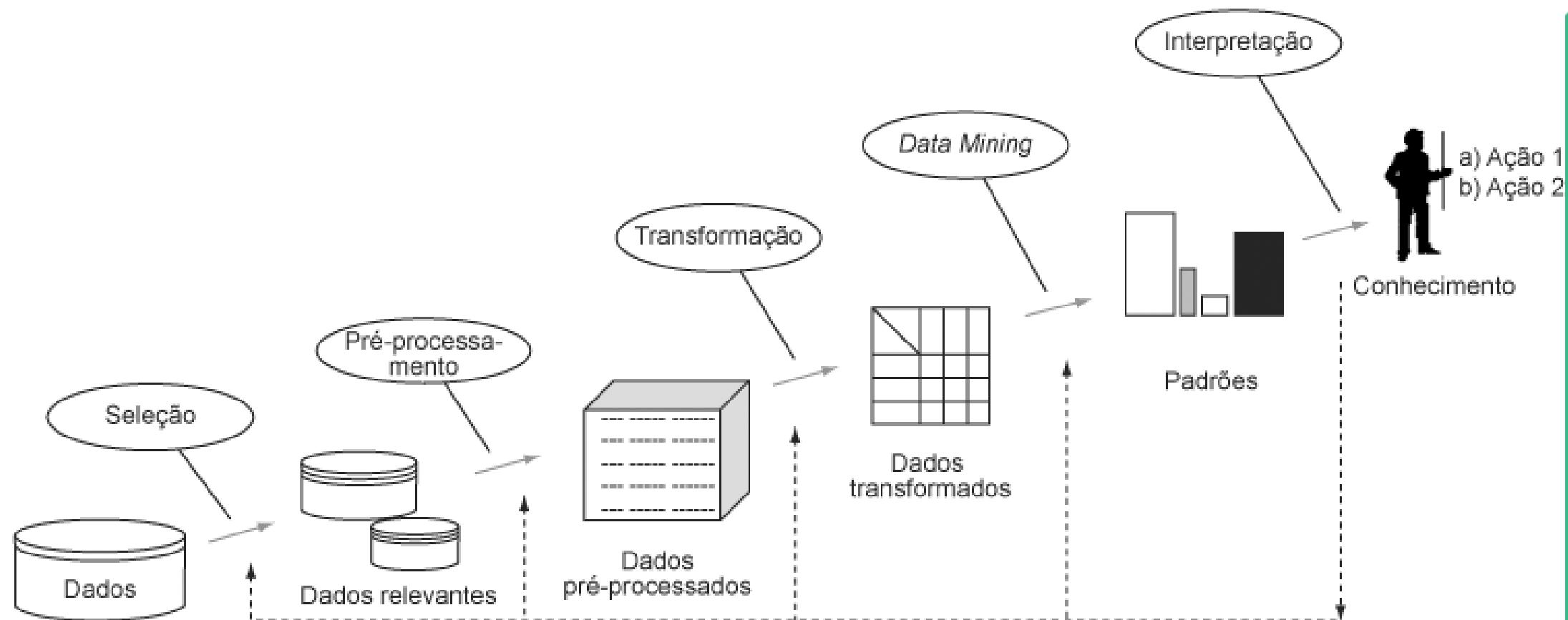


Figura 1. Etapas do processo *KDD* (Fayyad et al. (1996)).

Mineração de Dados

- É a prática de pesquisar automaticamente em grandes conjuntos de dados para descobrir padrões e tendências que estão além de simples análises
- A Mineração de Dados utiliza algoritmos matemáticos sofisticados para segmentar os dados e prever a probabilidade de eventos futuros baseados em comportamentos históricos

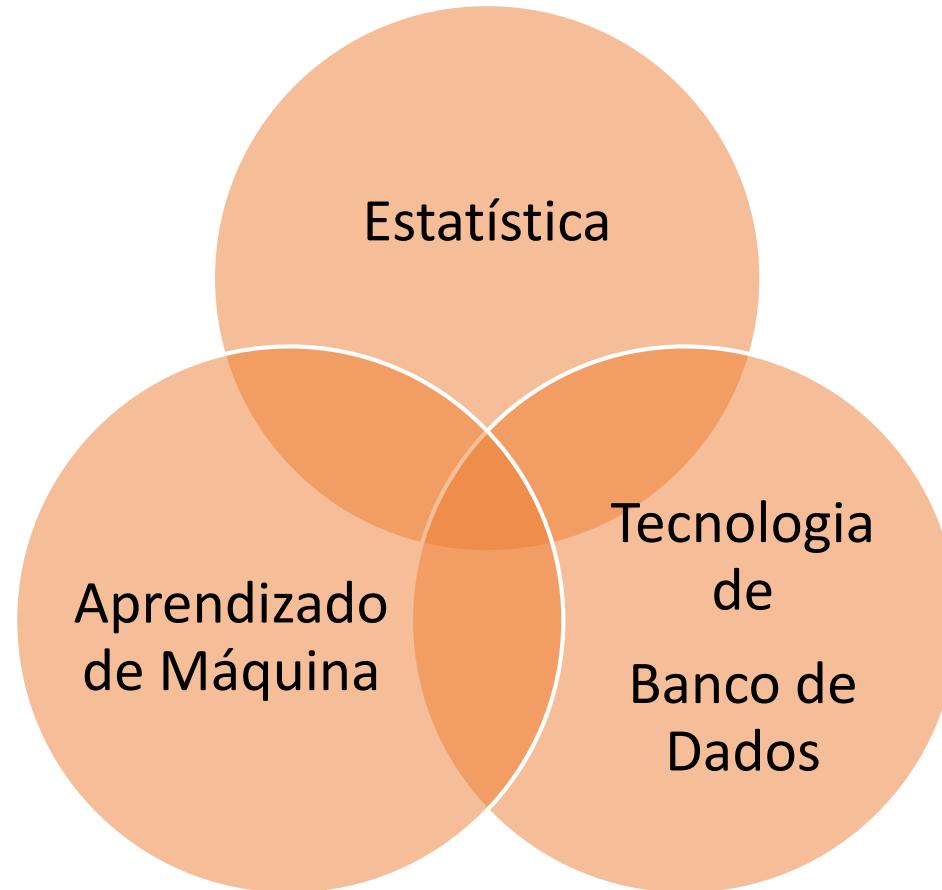
Principais propriedades da Mineração de Dados

- ✓ Descoberta automática de padrões;
- ✓ Previsão de probabilidade sobre resultados;
- ✓ Criação de informação açãoável;
- ✓ Foco em grandes bases e conjuntos de dados;

Aplicações de Mineração de Dados

- ✓ Análise de tendências de compra;
- ✓ Marketing direcionado;
- ✓ Detecção de fraudes.

Áreas de Conhecimento da Mineração de Dados



Aplicações em Tipos de Dados

- Tabelas;
- Banco de Dados;
- Data Warehouses;
- Dados Transacionais;
- Grafos e Redes;
- Dados Espaciais;
- Multimídia;
- Web.

Limitações da Mineração de Dados

- A Mineração de Dados é uma ferramenta poderosa que pode ajudar a encontrar padrões e relacionamentos implícitos nos dados, mas não realiza o trabalho sozinha;
- Não elimina a necessidade de conhecer bem o seu negócio ou seus dados, muito menos de entender de métodos analíticos.
- A Mineração de Dados encontra informações escondida em seus dados, mas não te diz o valor da informação para a sua organização;
- Os padrões encontrados podem ser muito diferentes dependendo da forma como você formula os problemas.

Para obter resultados significantes, você deve aprender como perguntar as perguntas certas



"Would you tell me, please, which way I ought to go from here?"

"That depends a good deal on where you want to get to," said the Cat.

"I don't much care where--" said Alice.

"Then it doesn't matter which way you go," said the Cat.

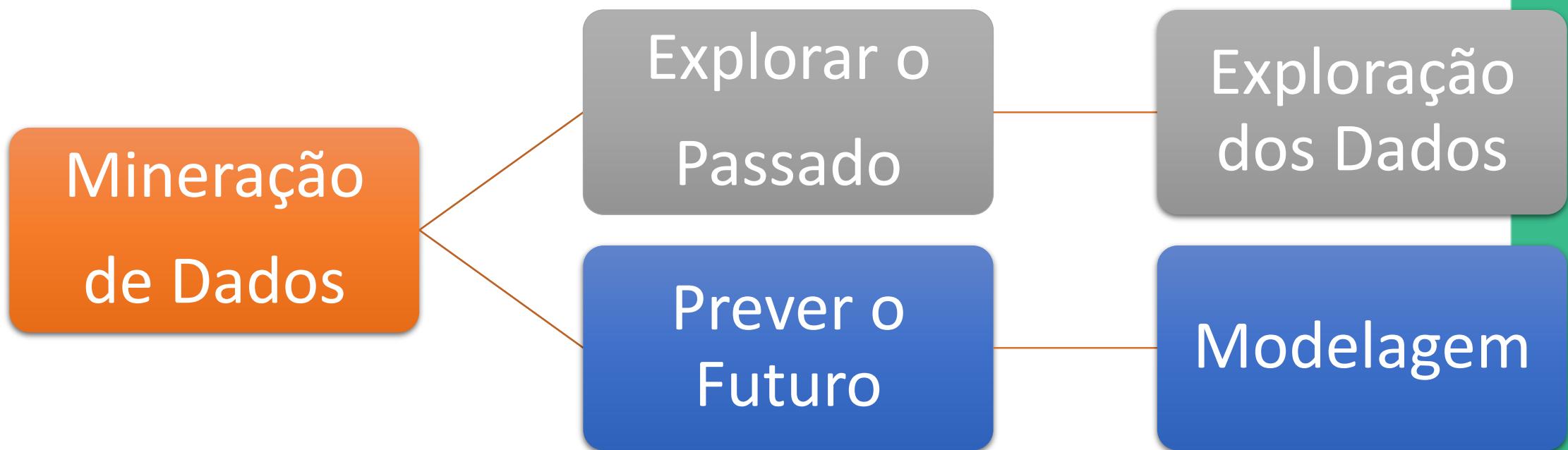
--so long as I get SOMEWHERE," Alice added as an explanation.

"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Os algoritmos são muito sensíveis a características específicas dos dados

- **Outliers:** Dados com valores muito diferentes dos valores típicos encontrados em seu conjunto de dados;
- Atributos irrelevantes;
- Codificação dos dados.

Abordagem Conceitual



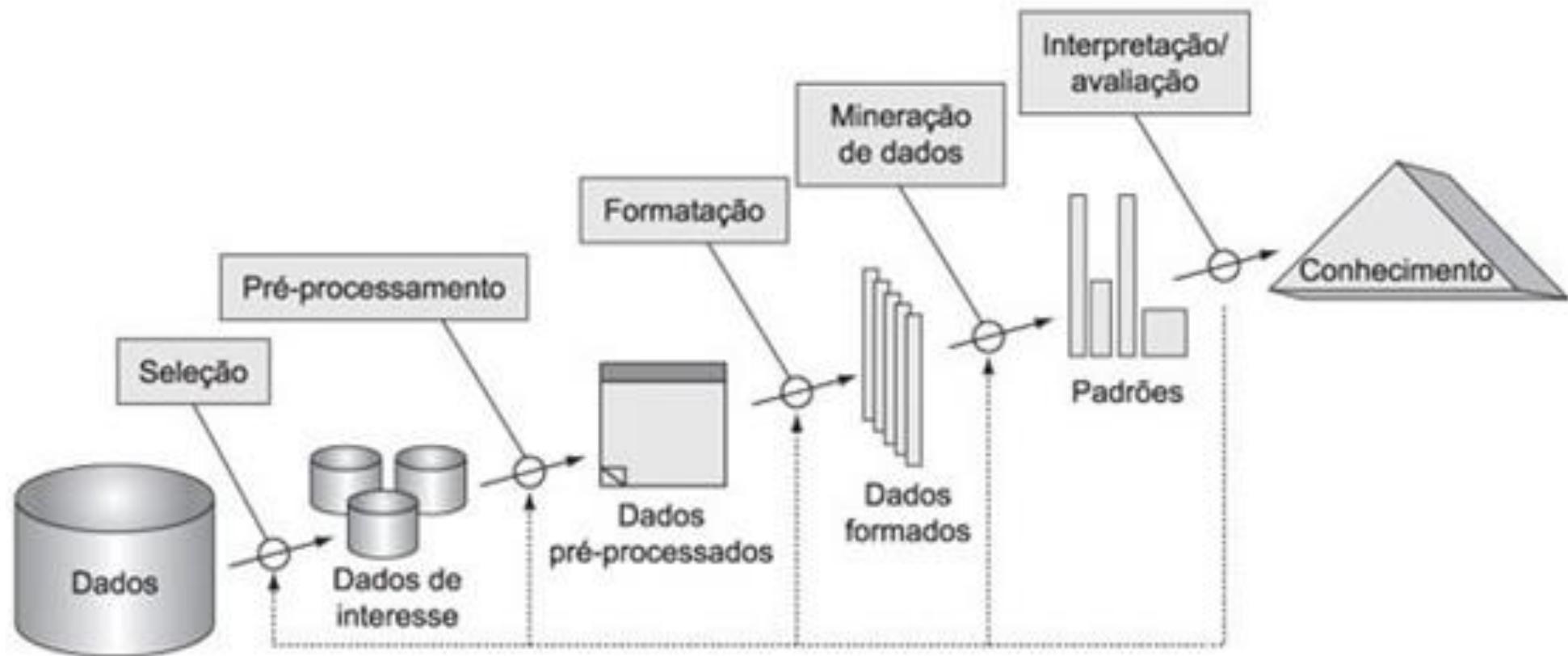
Exploração dos Dados

- A Exploração dos dados permite descrever os dados por meio de técnicas de estatística e visualização;
- Exploramos os dados para realçar aspectos importantes dos dados e subsidiar análises mais complexas.

Modelagem Preditiva

- É o processo pelo qual um modelo é criado para prever um resultado.

Processo Genérico de Mineração de Dados



Metodología SAS SEMMA

Sample

Explore

Modify

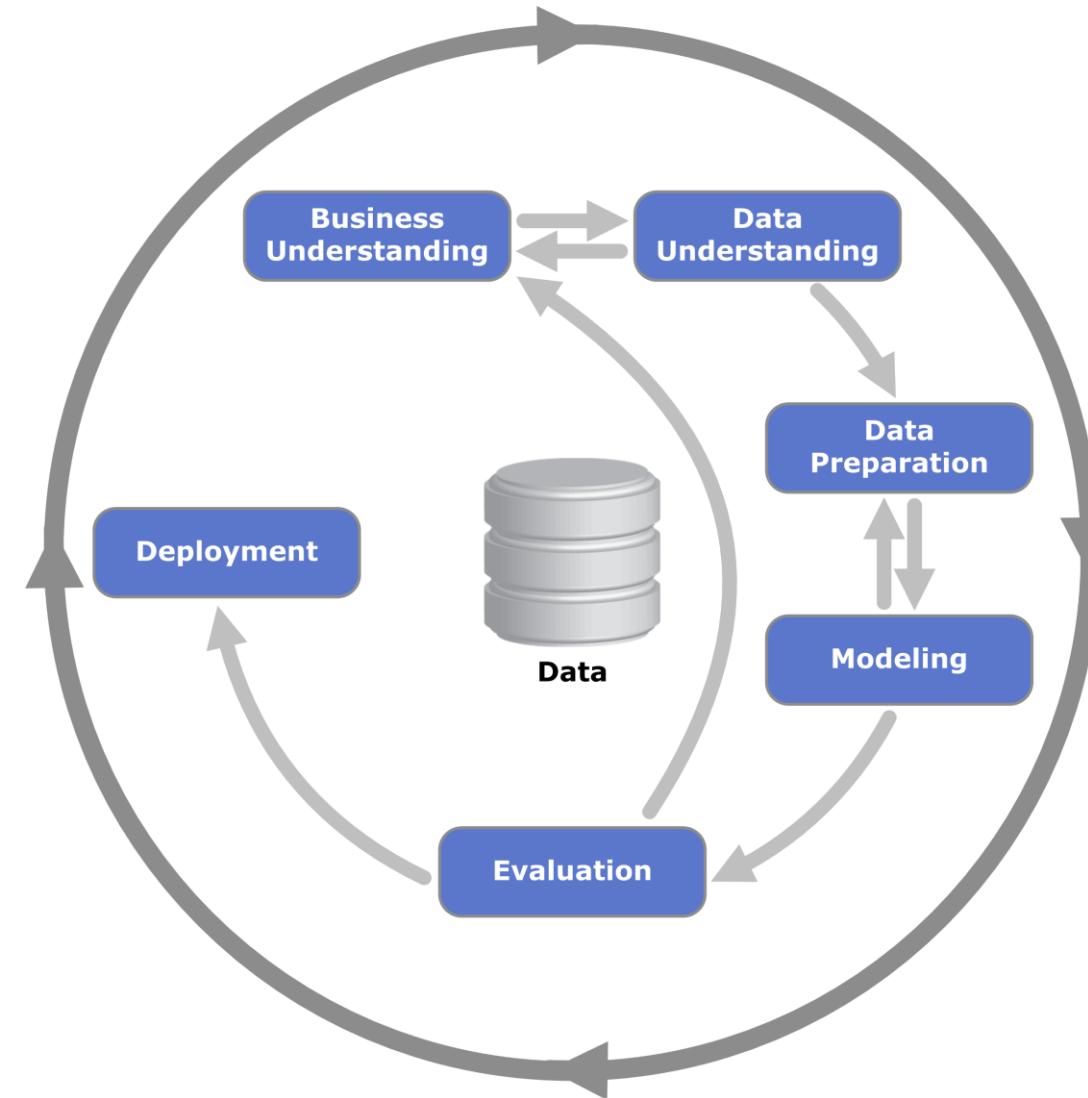
Model

Access

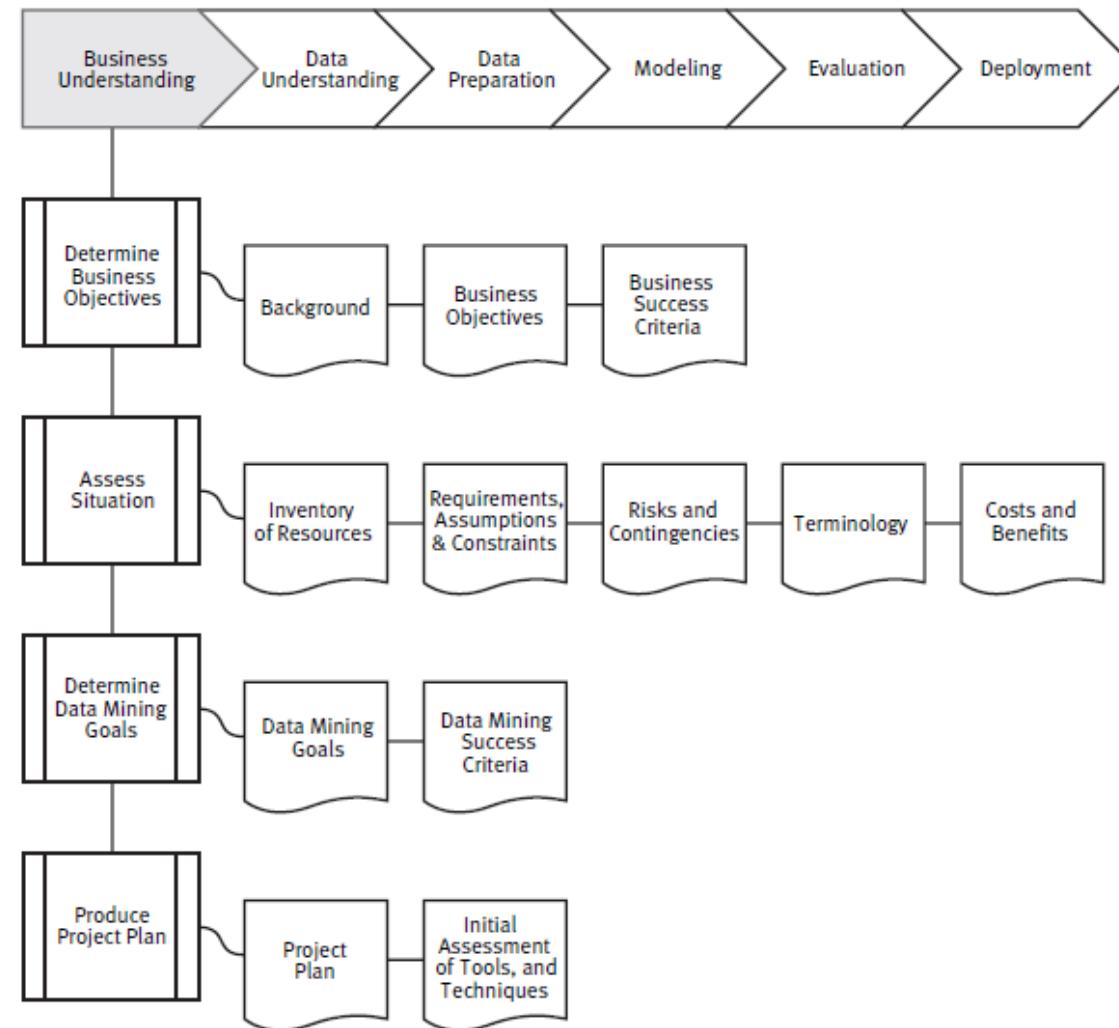
Processos do SEMMA

- **Sample.** Processo de amostragem onde é selecionado os dados para o modelo;
- **Explore.** Entendimento dos dados, relacionamento entre variáveis, anormalidades e etc. apoiados com técnicas de visualização de dados;
- **Modify.** Métodos para selecionar, criar e transformar variáveis para o modelo;
- **Model.** Aplicação de técnicas de Mineração de Dados nas variáveis preparadas para a previsão de resultados;
- **Assess.** Avalia os resultados e confiabilidade do modelo.

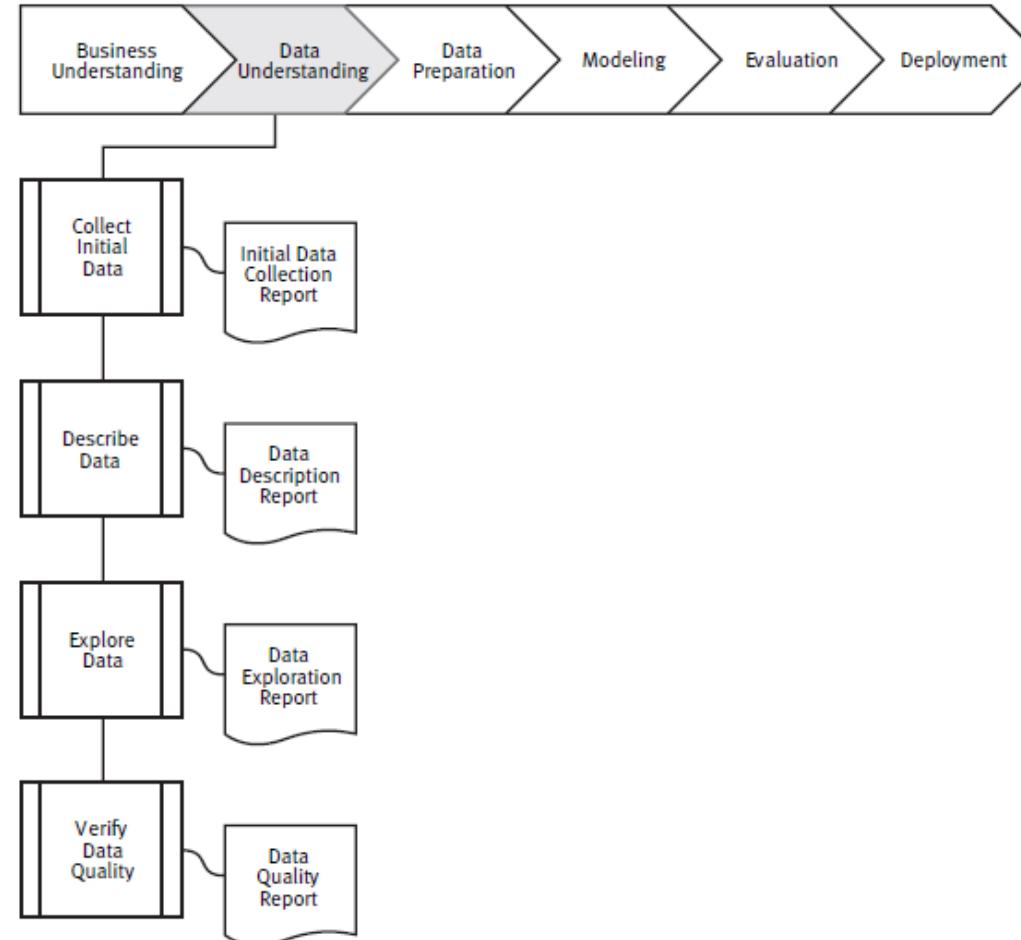
CRISP-DM



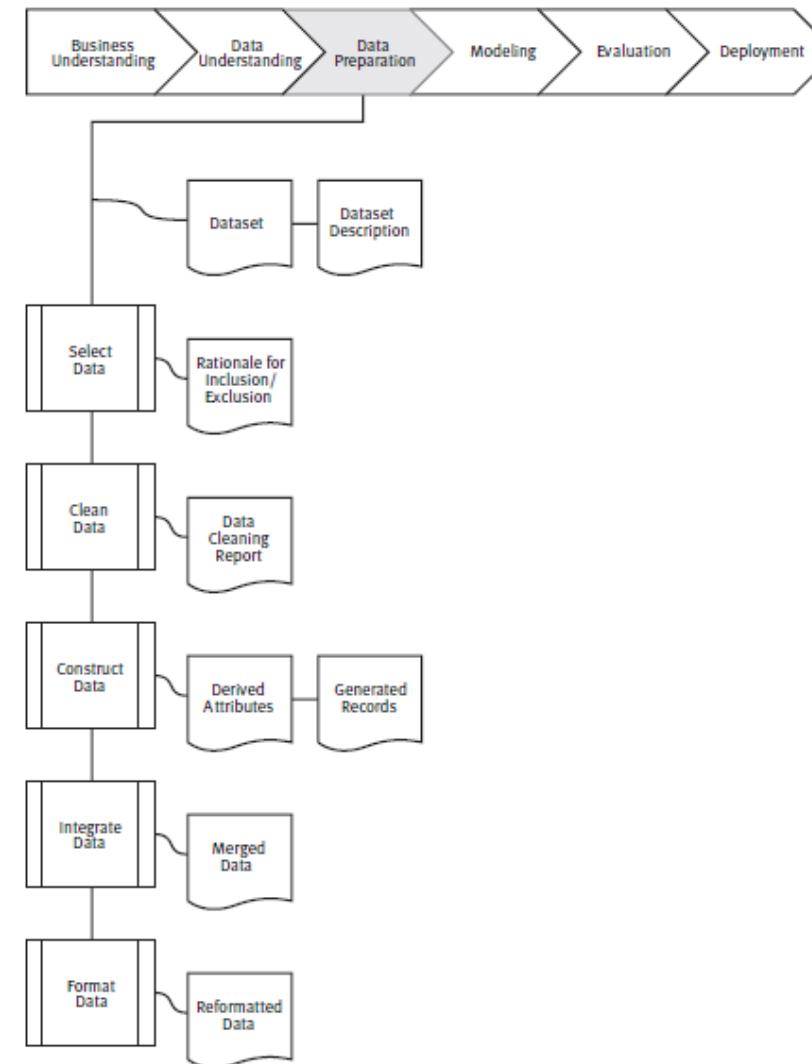
CRISP-DM 1 - Entendimento da Organização



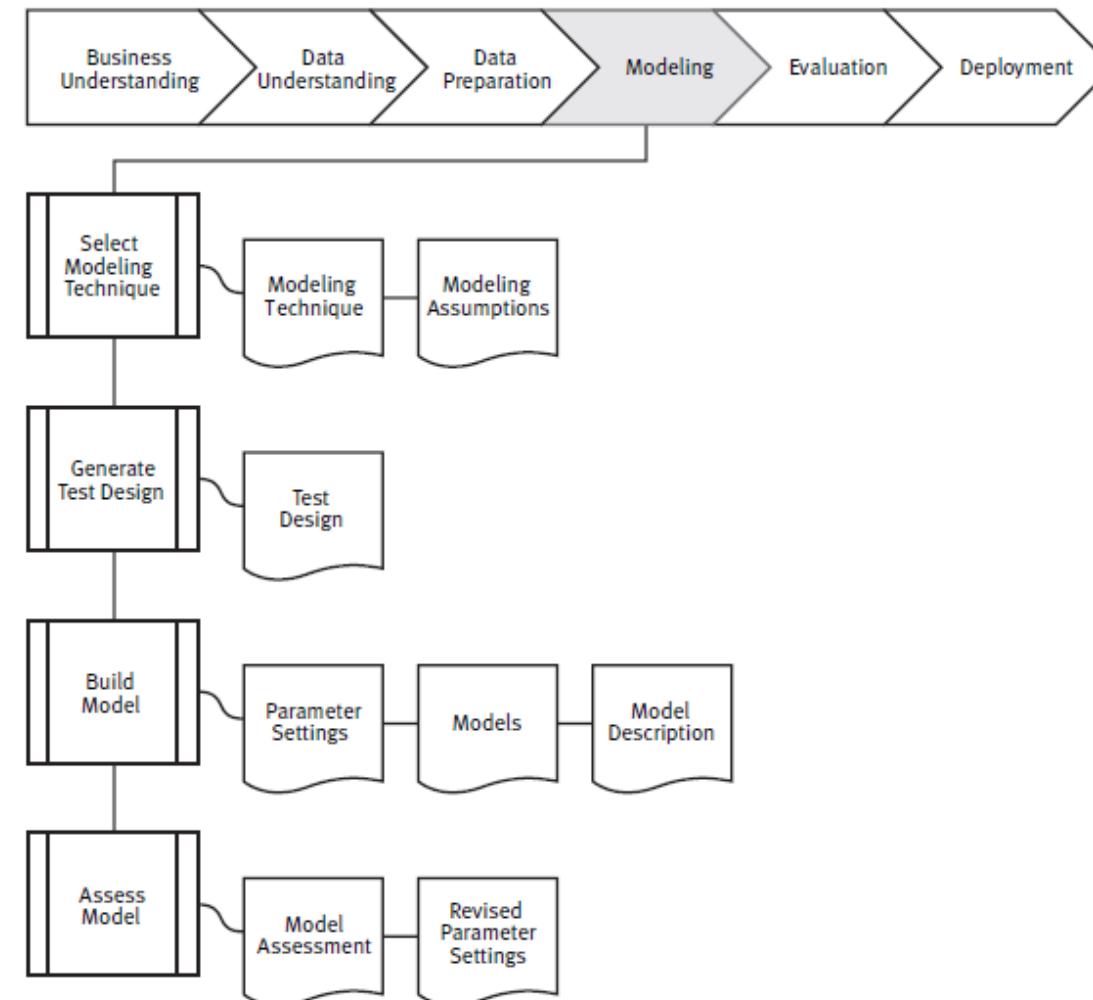
CRISP-DM 2 - Entendimento dos Dados



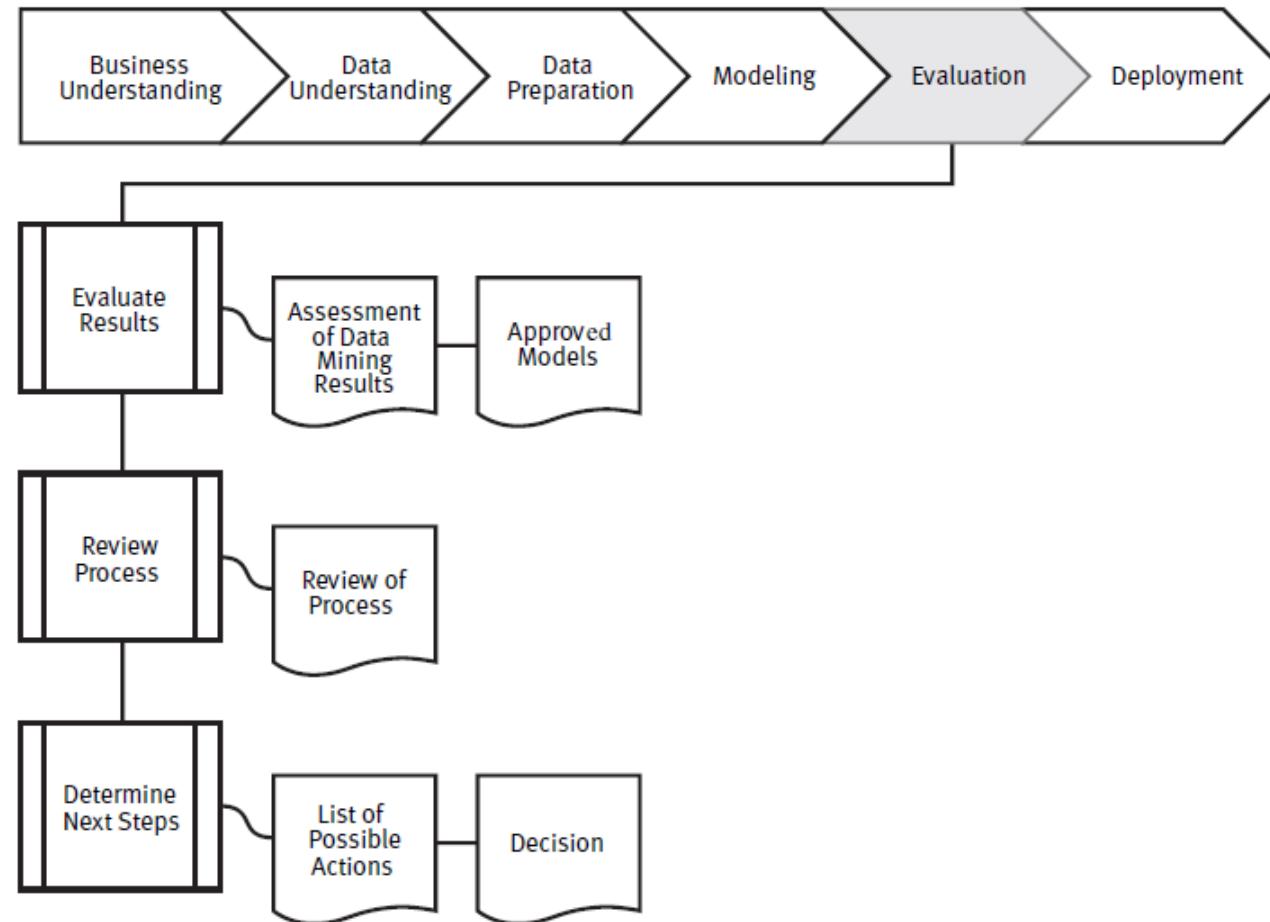
CRISP-DM 3 - Preparação dos Dados



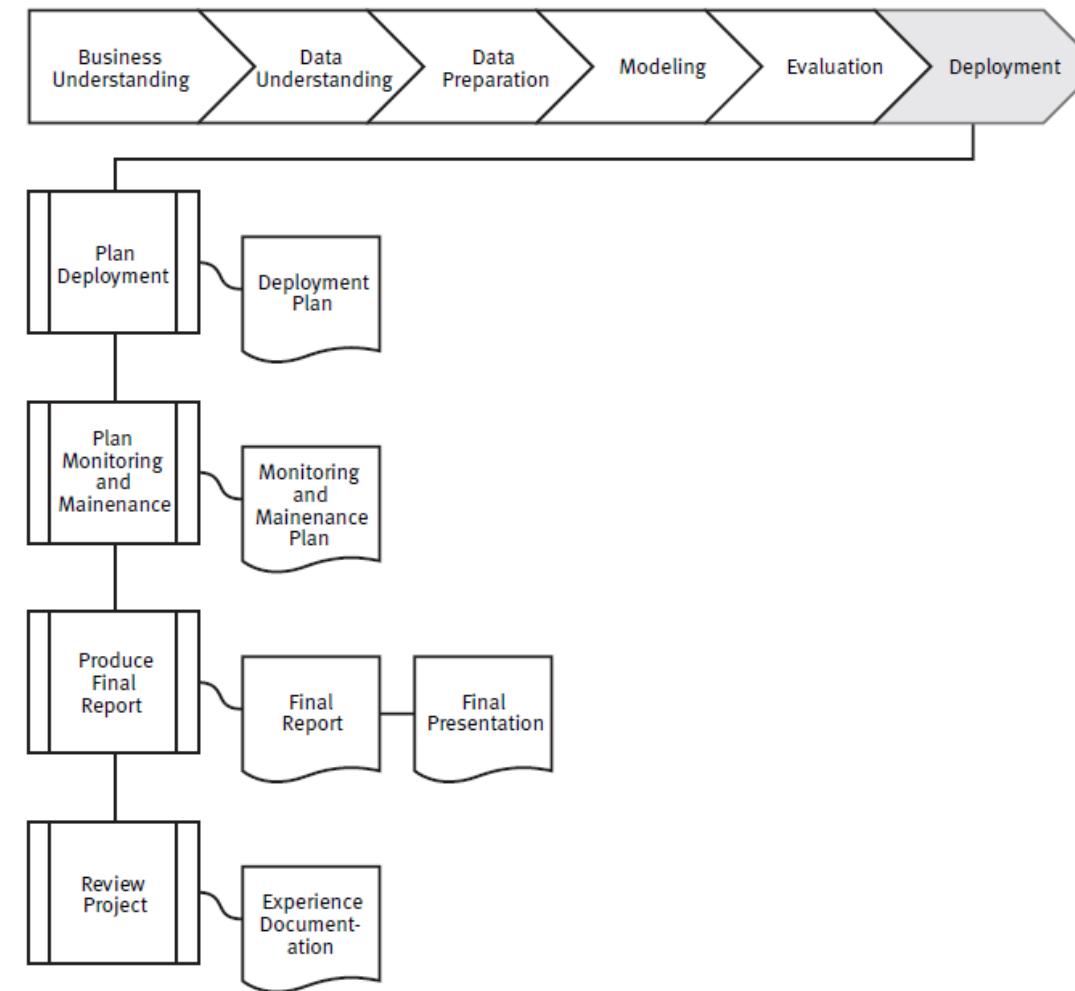
CRISP-DM 4 - Modelagem



CRISP-DM 5 - Avaliação

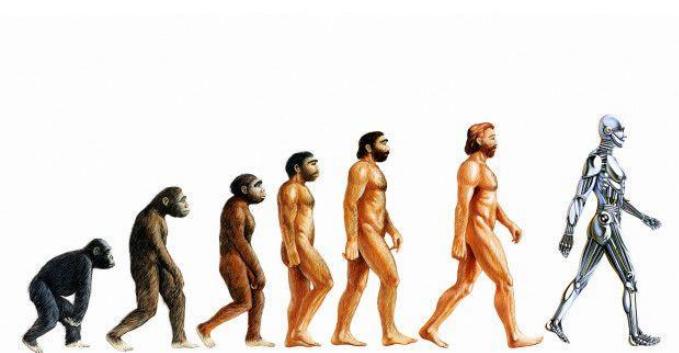


CRISP-DM 6 - Implantação



Aprendizado de Máquina

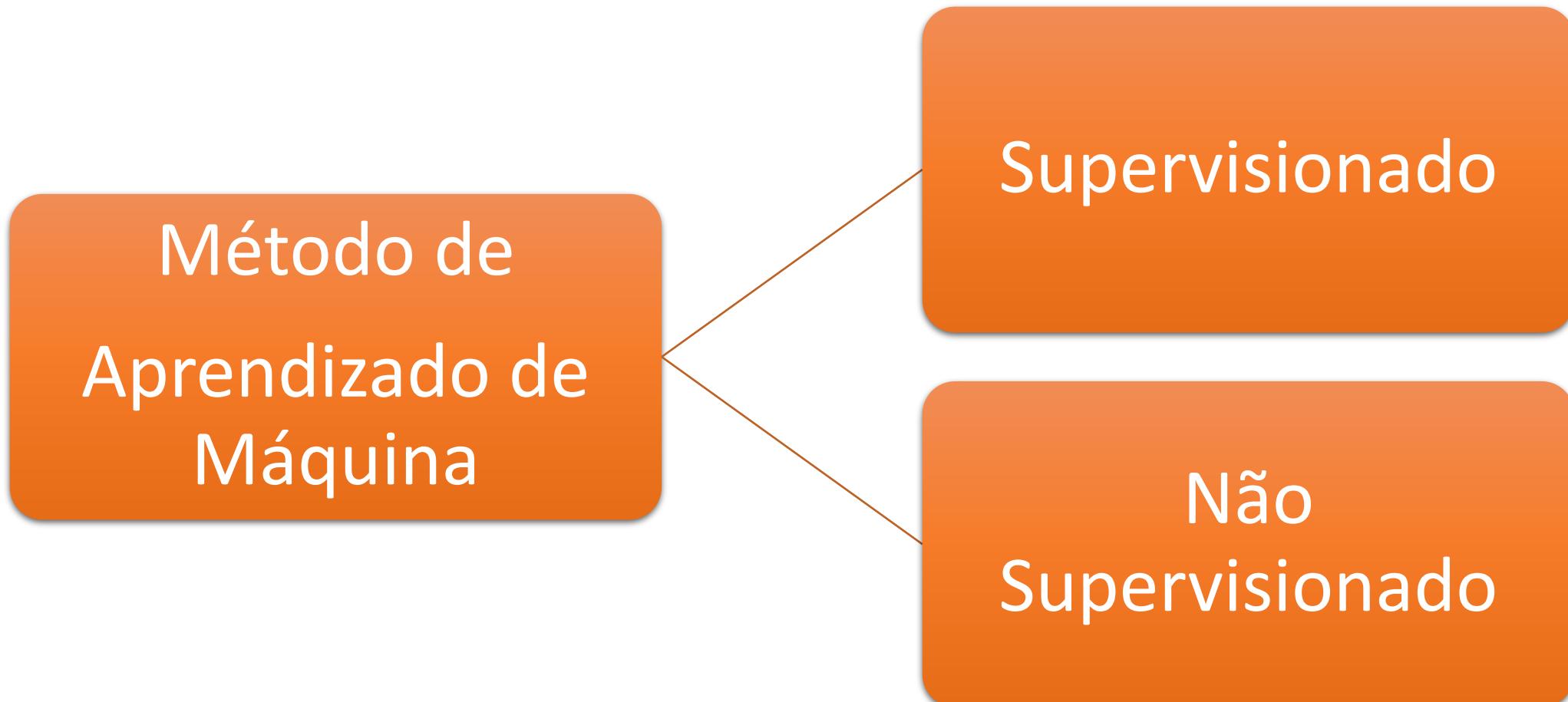
- Subcampo de Inteligência Artificial;
- Estudo de algoritmos e técnicas computacionais que permitem o aperfeiçoamento e otimização da execução de tarefas.



Aplicações de Aprendizado de Máquina

- Visão Computacional;
- Mineração de Dados;
- Processamento de Linguagem Natural.

Aprendizado de Máquina



Aprendizado de Máquina Supervisionado

- O processo de aprendizado é direcionado por uma variável dependente previamente conhecida;
- Explica o comportamento de uma variável alvo como uma função de um conjunto de variáveis dependentes ou preditores;
- Geralmente resulta em modelos preditivos;
- A construção de um modelo supervisionado implica na etapa de treinamento;

Aprendizado de Máquina Não Supervisionado

- Lida com situações onde não existe um resultado previamente conhecido para guiar o algoritmo na construção do modelo;
- Pode ser utilizado para propósitos descritivos;
- Pode ser utilizado também em previsões.

Funções em Mineração de Dados

- Caracterização e Discriminação;
- Padrões frequentes, associação e correlação;
- Classificação e Regressão;
- Agrupamentos;
- Detecção de anomalias.

Modelo e Algoritmos

Situação Problema

Classificação da Função

Método

Algoritmo

K-means

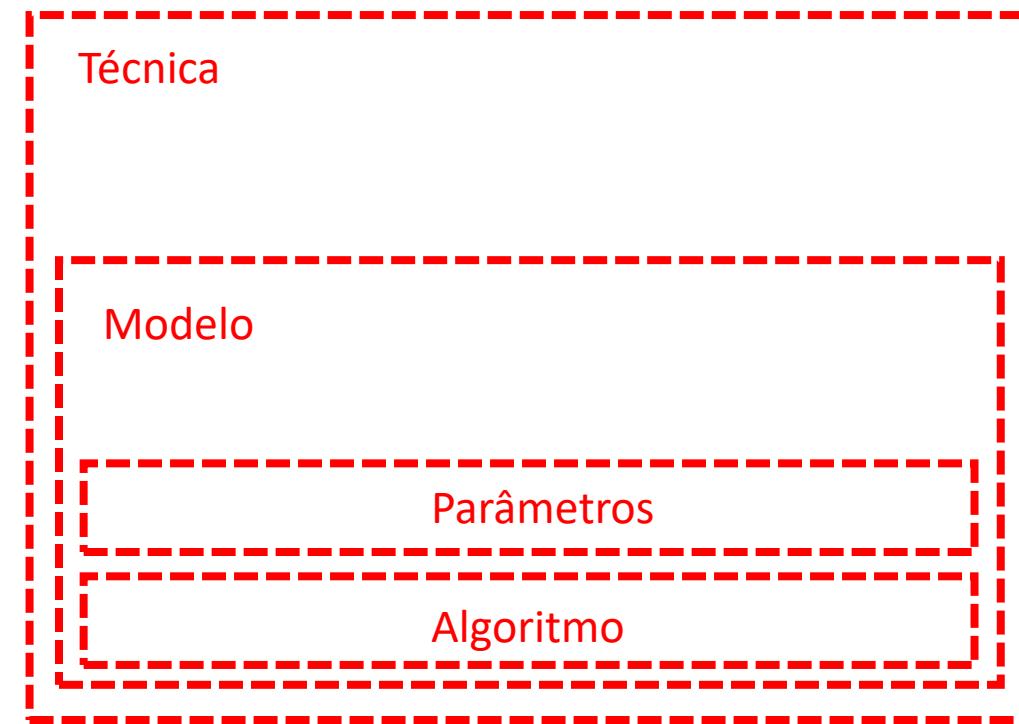
K-NN

SVM

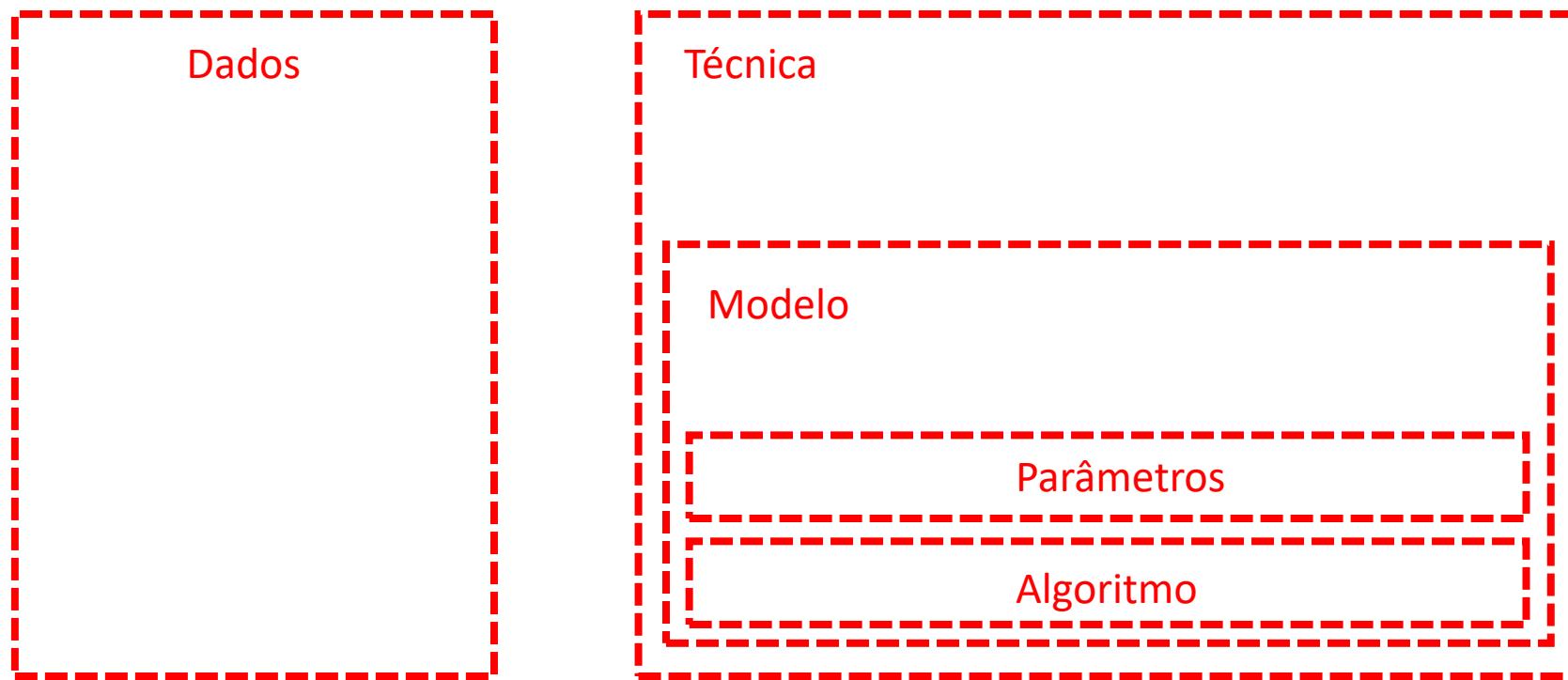
APRIORI

NAIVE BAYES

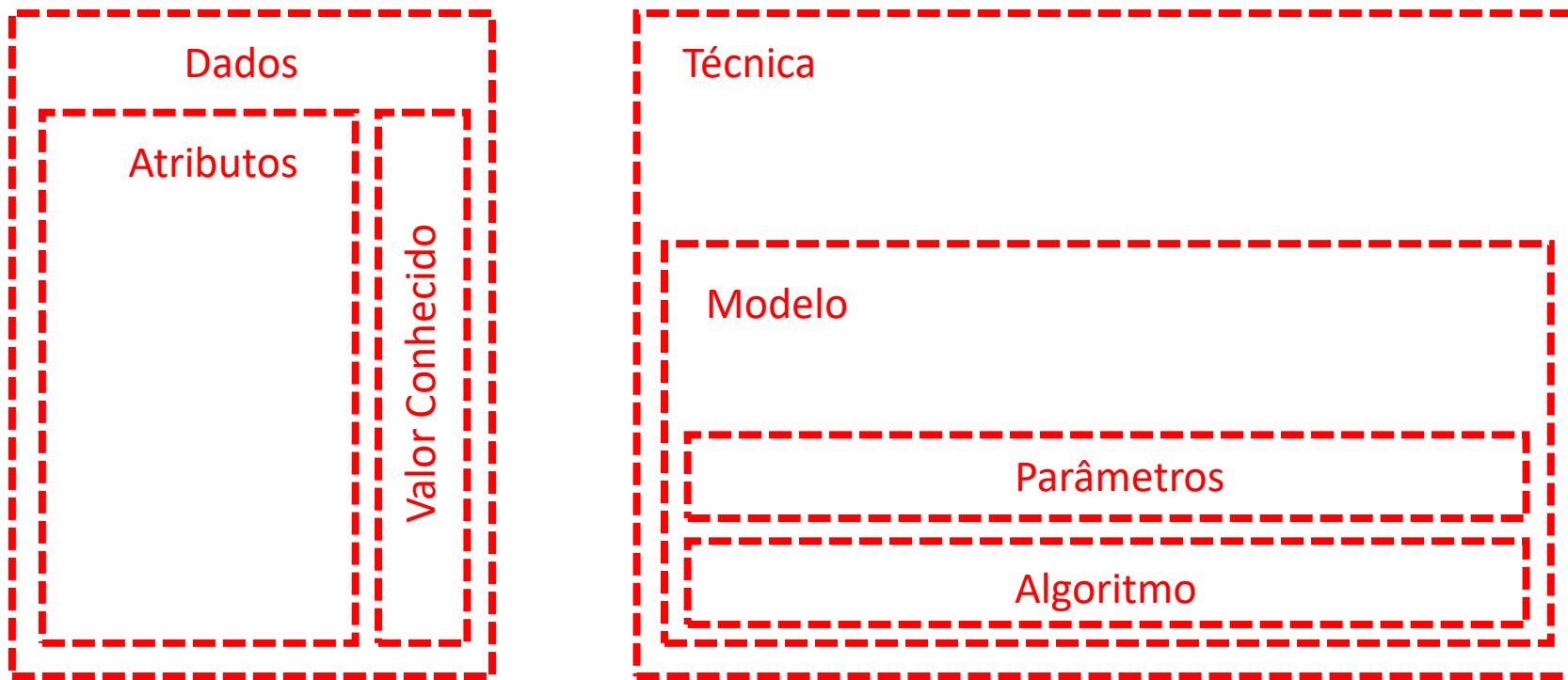
Modelo e Algoritmos



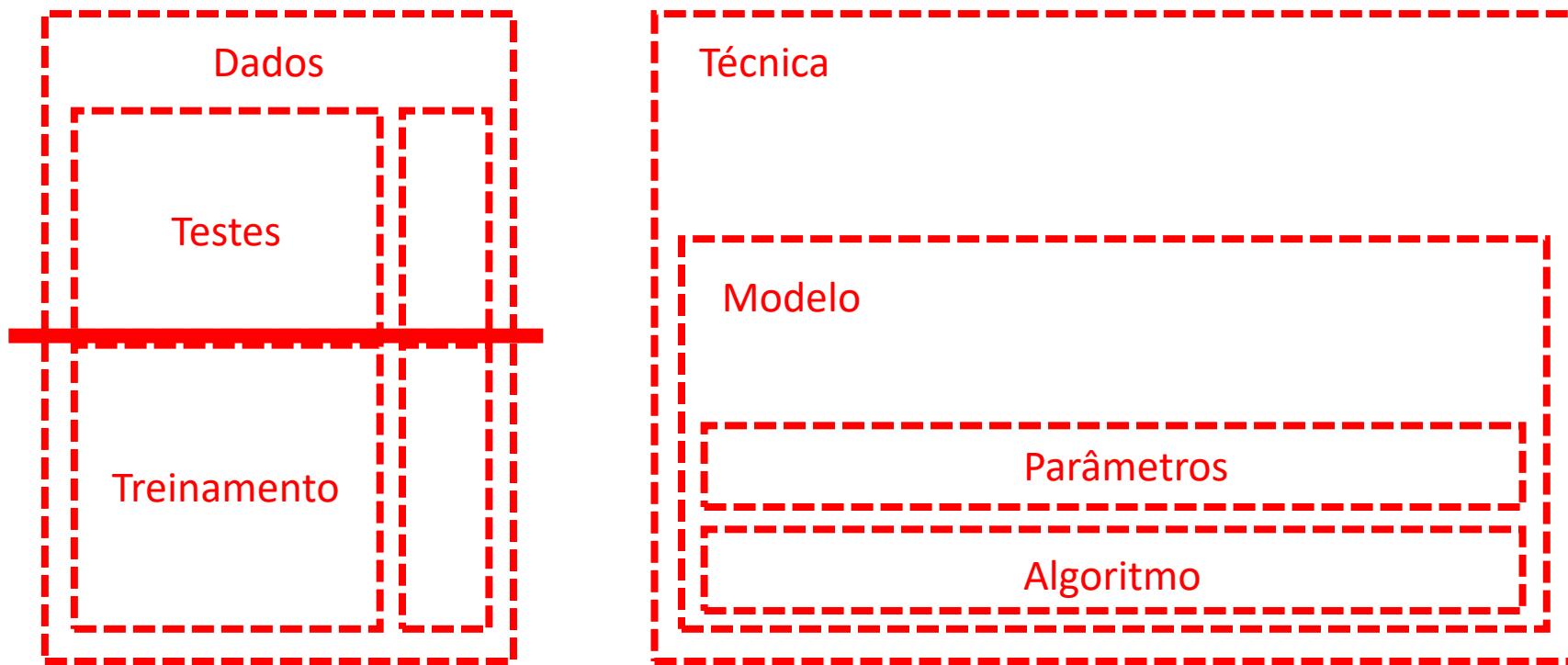
Modelo e Algoritmos



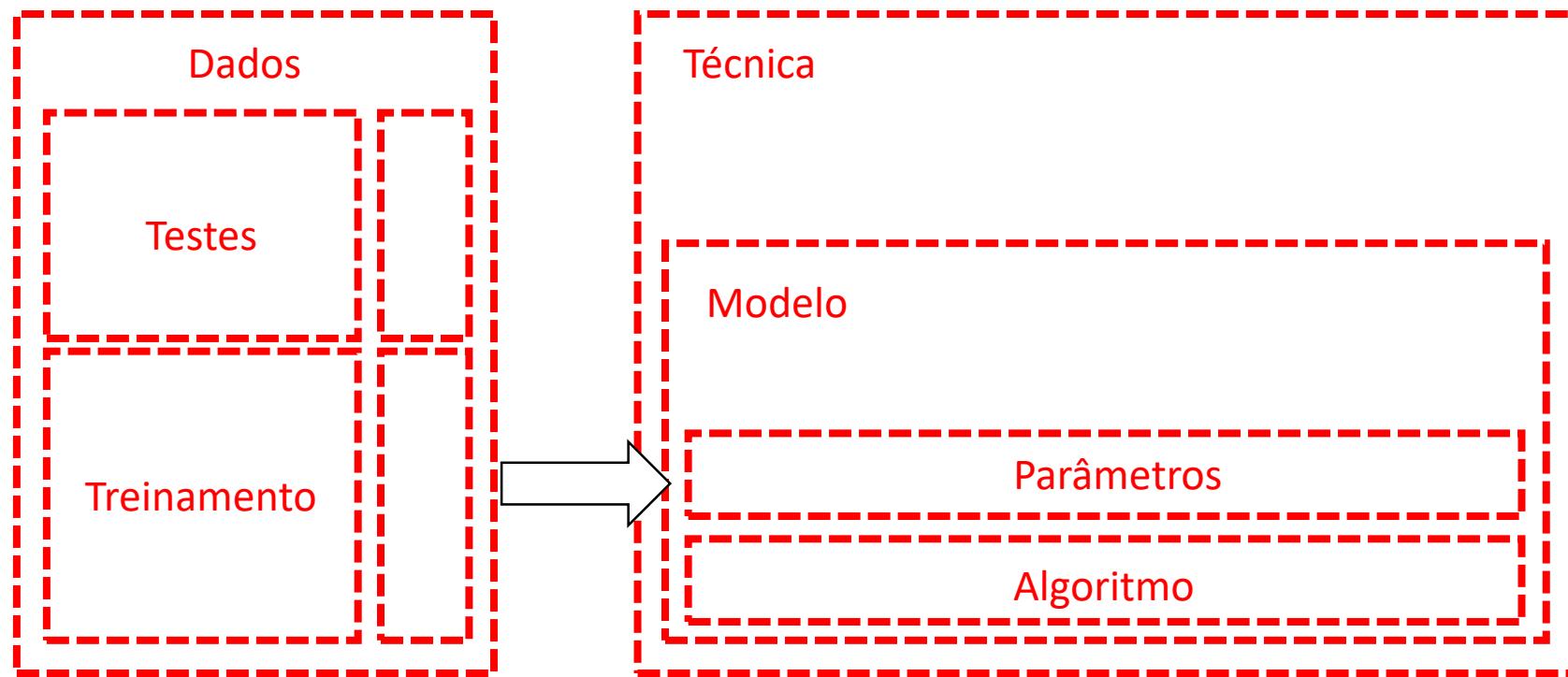
Modelo e Algoritmos



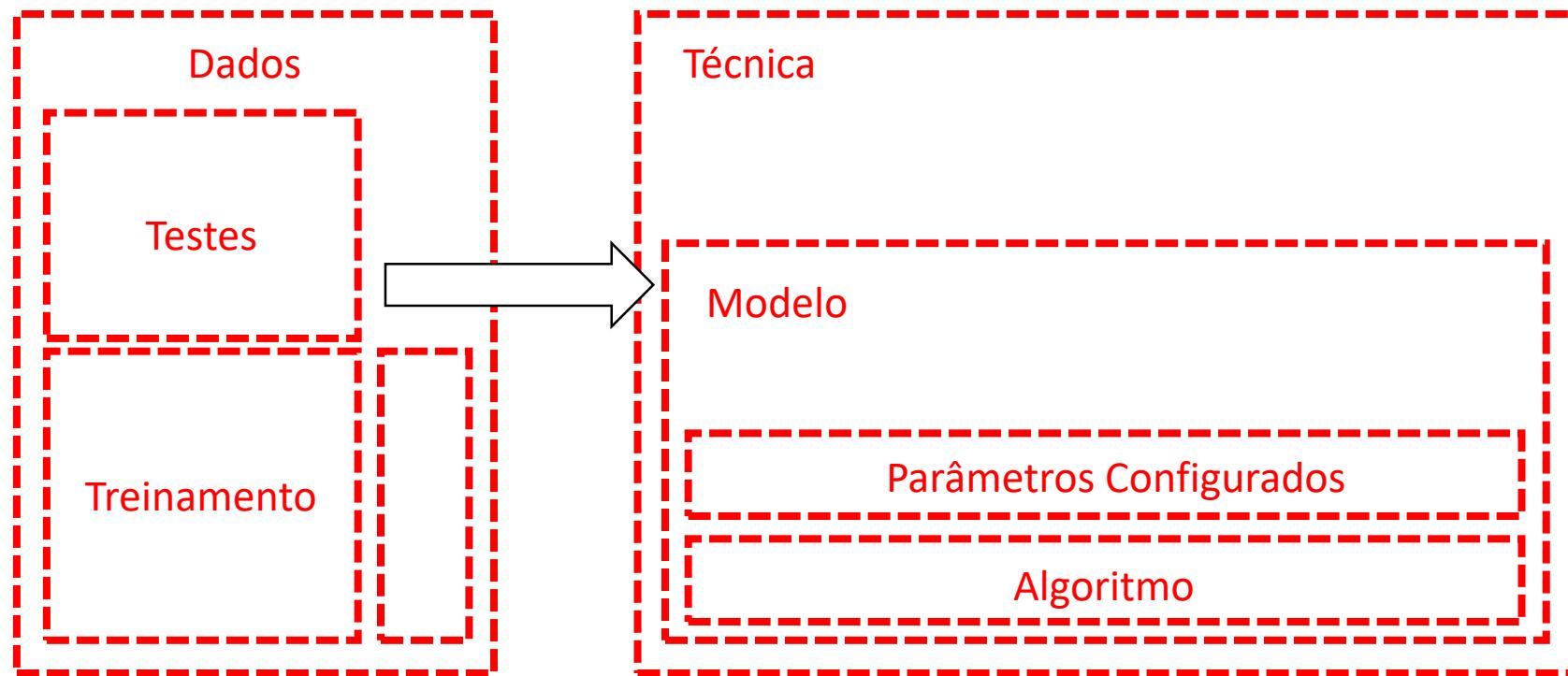
Modelo e Algoritmos



Modelo e Algoritmos



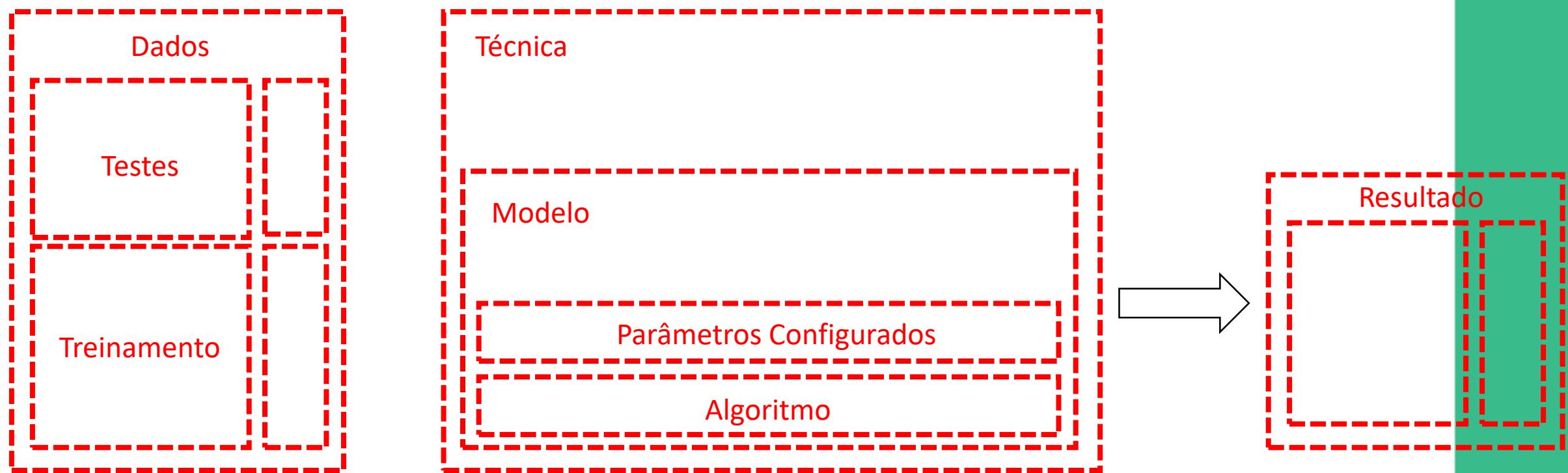
Modelo e Algoritmos



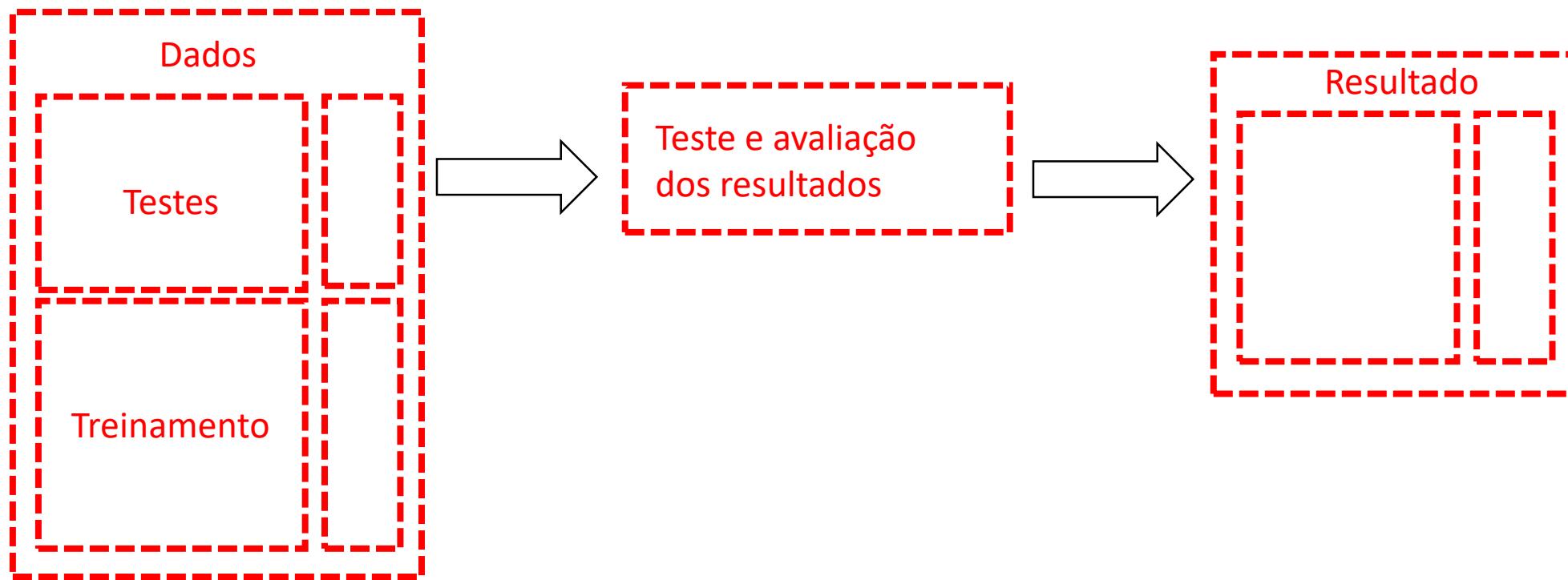
Modelo e Algoritmos



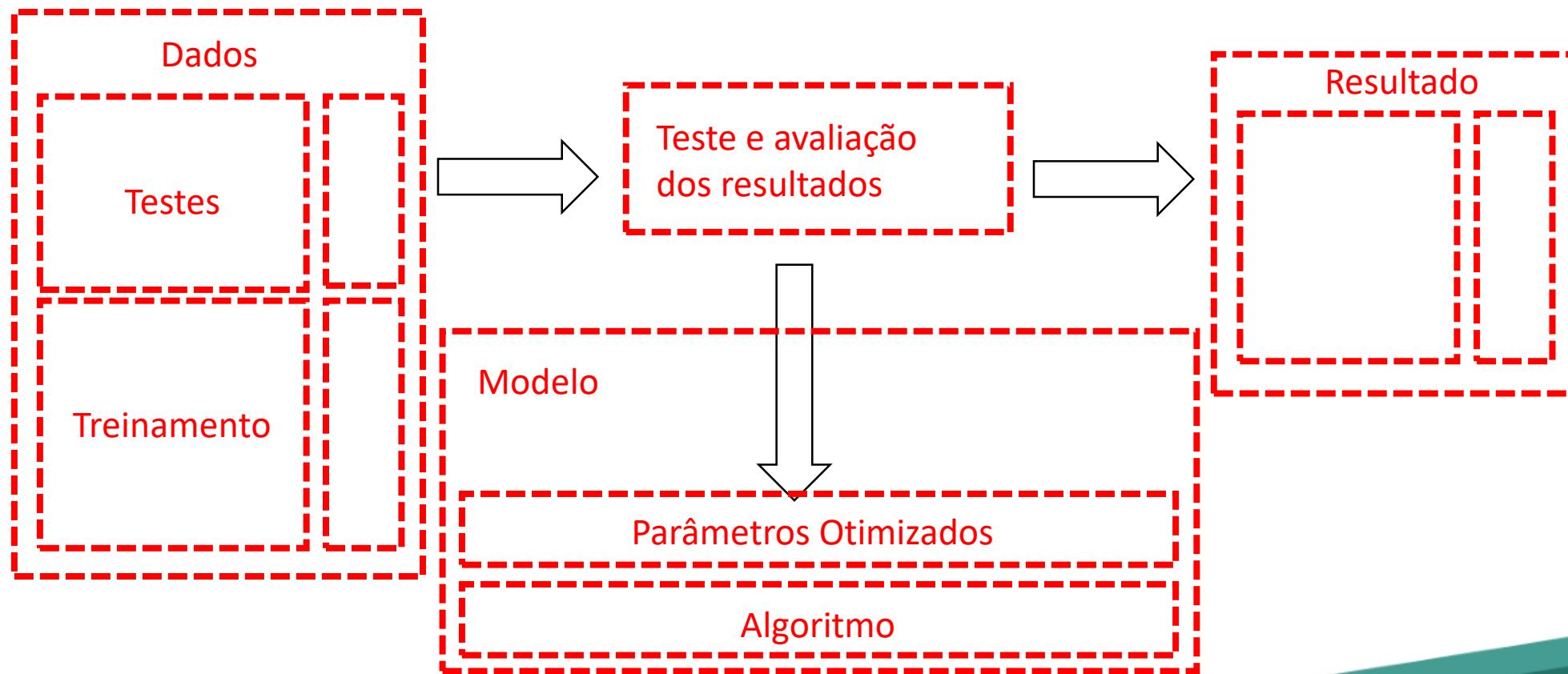
Modelo e Algoritmos



Modelo e Algoritmos



Modelo e Algoritmos



Matriz de Correlação

Attributes	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Temperature	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_Oil	0.736	-0.774	1	-0.042	0.848	0.381
Num_Occupants	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_Size	0.201	-0.214	0.381	-0.023	0.307	1

Análise de Padrões Frequentes

✓ Itens -> Transações

✓ Sequencial ->

Cliente compra notebook -> Compra câmera -> Compra memory card

✓ Subestruturas -> grafos combinados, estruturas

Análise de Associação

- O modelo permite a descoberta de itens que são **frequentemente encontrados juntos em uma mesma transação**;
- **Aplicação:** Marketing, Logística.

Análise de Associações

Comprar(x,computador) -> comprar(x,software)

| suporte=1% e confiança=50% |

Suporte Mínimo =

Confiança Mínima =

Análise de Relevância

- Precede as atividades de Regressão e Classificação;
- Identifica os atributos significativos com poder de predição sobre uma variável alvo.

Regressão

➤ Função de Mineração de Dados para prever uma **variável alvo numérica continua**

Exemplo: Estimar o valor de uma casa baseado em sua localização, número de quartos, tamanho do lote e outros fatores

- Vendas;
- Margem de lucro;
- Taxa de mortalidade;
- Temperatura.

Conjunto de Dados de Treino

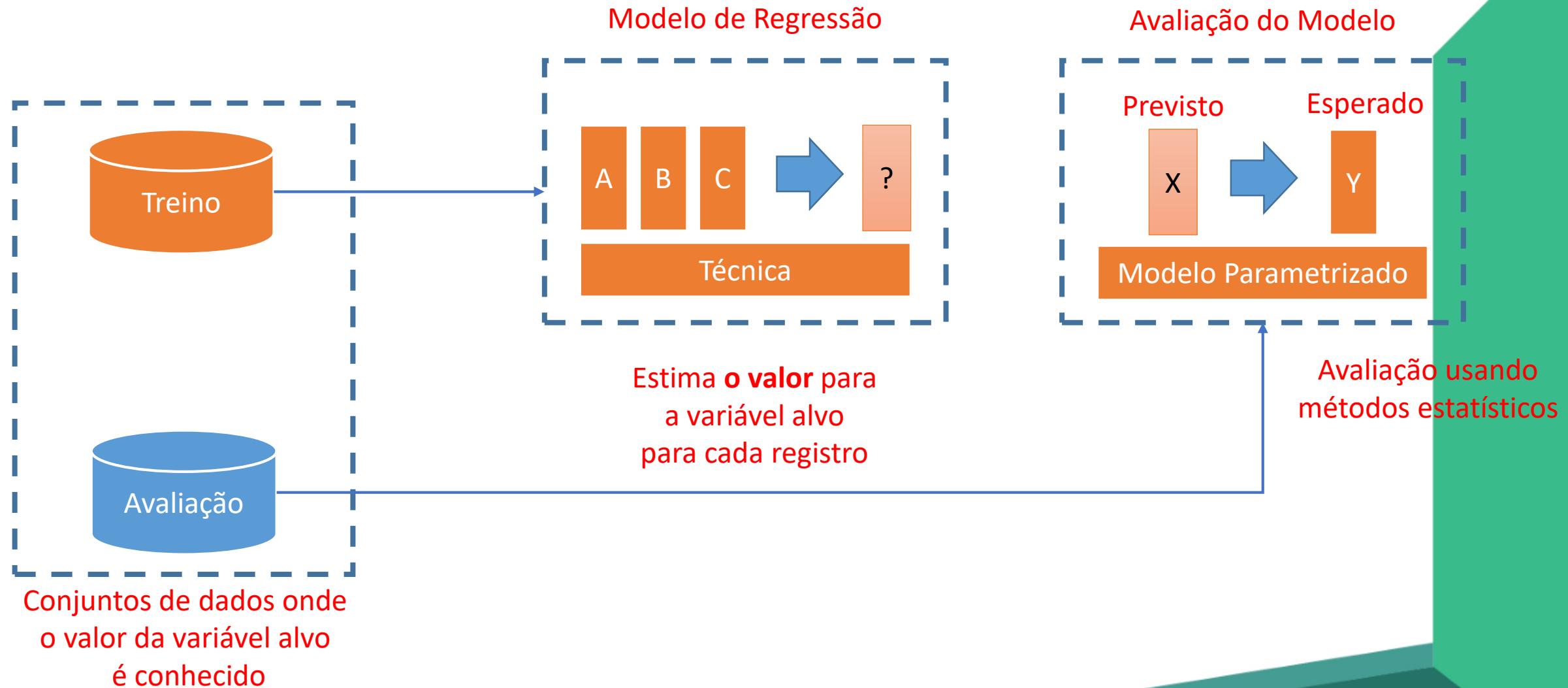
- A atividade de Regressão inicia com um conjunto de dados onde os valores da variável alvo é conhecido;
- Um modelo que estima o valor de casas pode ser desenvolvido baseado em observações de diversas casas sob um período de tempo.

Variável Alvo

Ex.: Estimar o **valor de uma casa** baseado em sua
localização, número de quartos, tamanho do lote, proximidade de comércios

Preditores

Modelo de Regressão



Aplicações de Regressão

- Análise de Tendências;
- Planejamento de Negócios;
- Marketing;
- Previsão Financeira;
- Previsão de Séries Temporais.

Como funciona?

- Uma análise de regressão busca determinar os valores de parâmetros para uma função que melhor represente os dados onde o valor da variável alvo é conhecido

Coeficientes de regressão
(parâmetros)

Variável dependente

$$y = F(x, \theta) + e$$

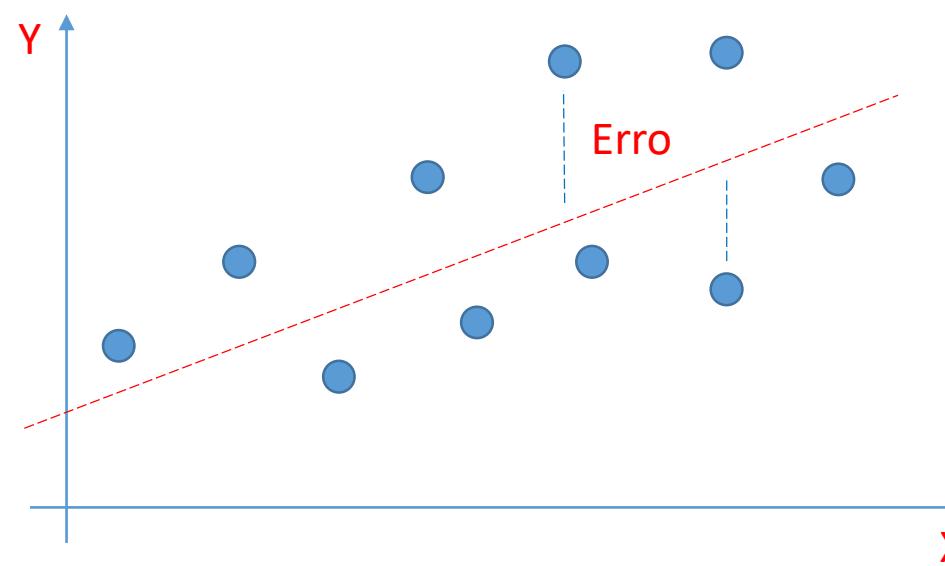
Erro residual

Preditores ($x_1, x_2, x_3, \dots, x_n$)

O processo de treinamento para o modelo
busca encontrar os valores de parâmetros
que minimizem a medida de erro

Regressão Linear

- A técnica de regressão linear é aplicada quando o relacionamento entre os preditores e o alvo pode ser aproximado com uma linha reta



$$y = \theta_1 + \theta_2 x + e$$

Regressão Linear Multivariada

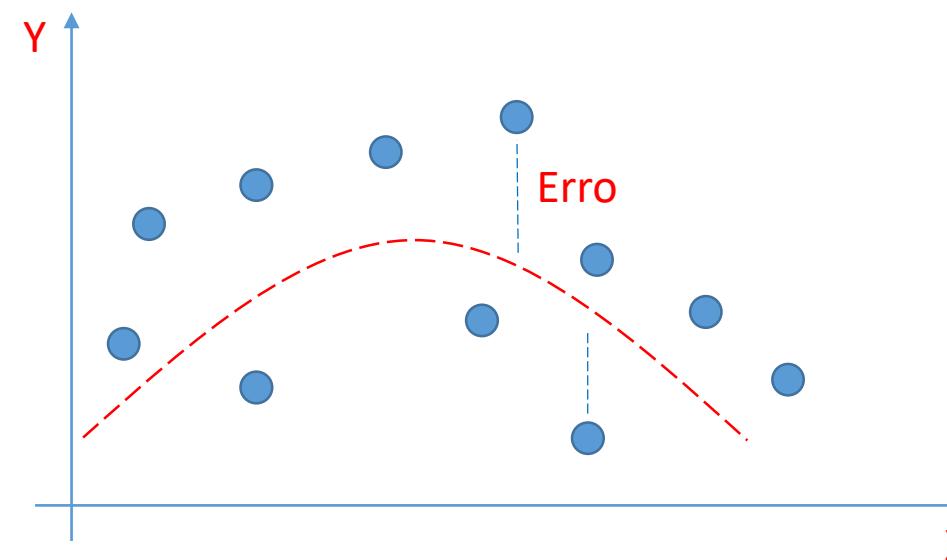
- A técnica de regressão multivariada refere-se a aplicação com dois ou mais preditores;
- Não pode ser visualizada em um gráfico 2D, mas pode ser representada expandindo a equação para vários preditores.

$$y = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \dots + \theta_n x_{n-1} + e$$

Coeficientes

Regressão Não Linear

- A técnica de regressão não linear é aplicada quando o relacionamento entre x e y **não pode ser aproximado por uma reta**.



Testes do Modelo de Regressão

- Um modelo de regressão pode ser testado comparando os valores previstos com os valores conhecidos utilizando um conjunto de dados de testes

Métricas de testes - Medem a amplitude e magnitude do erro

- Raiz do Erro Quadrático Médio;
- Erro médio absoluto .

Raiz do Erro Quadrático Médio - REQM

Representa a raiz da média do quadrado da distância de um ponto para a linha apropriada

Mede a amplitude do erro

$$\text{REQM} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Medida de precisão mais sensível a erros.

O valor zero indica uma previsão perfeita e este valor aumenta conforme aumenta a diferença entre valores de previsão e observação (sempre positivo)

Erro Médio Absoluto - EMA

- Representa a média do valor absoluto dos resíduos (erros);
- Ajuda a conceituar a magnitude do erro expresso na mesma unidade dos dado. Exemplo:

$$\text{EMA} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Algoritmos para Regressão

- Otimizados para conjunto de dados com alta dimensionalidade (muitos atributos)
- Modelos Lineares Generalizados;
- Máquina de Vetor de Suporte.

Classificação

- Função de Mineração de Dados Supervisionada para prever uma **variável alvo categórica**;
- Permite atribuir a uma **categoria** um item em uma coleção de dados .

Ex. Categorizar um risco de crédito em baixo, médio ou alto

- ✓ Segmentação de Clientes
- ✓ Marketing
- ✓ Análise de Crédito

Tipos de Classificação

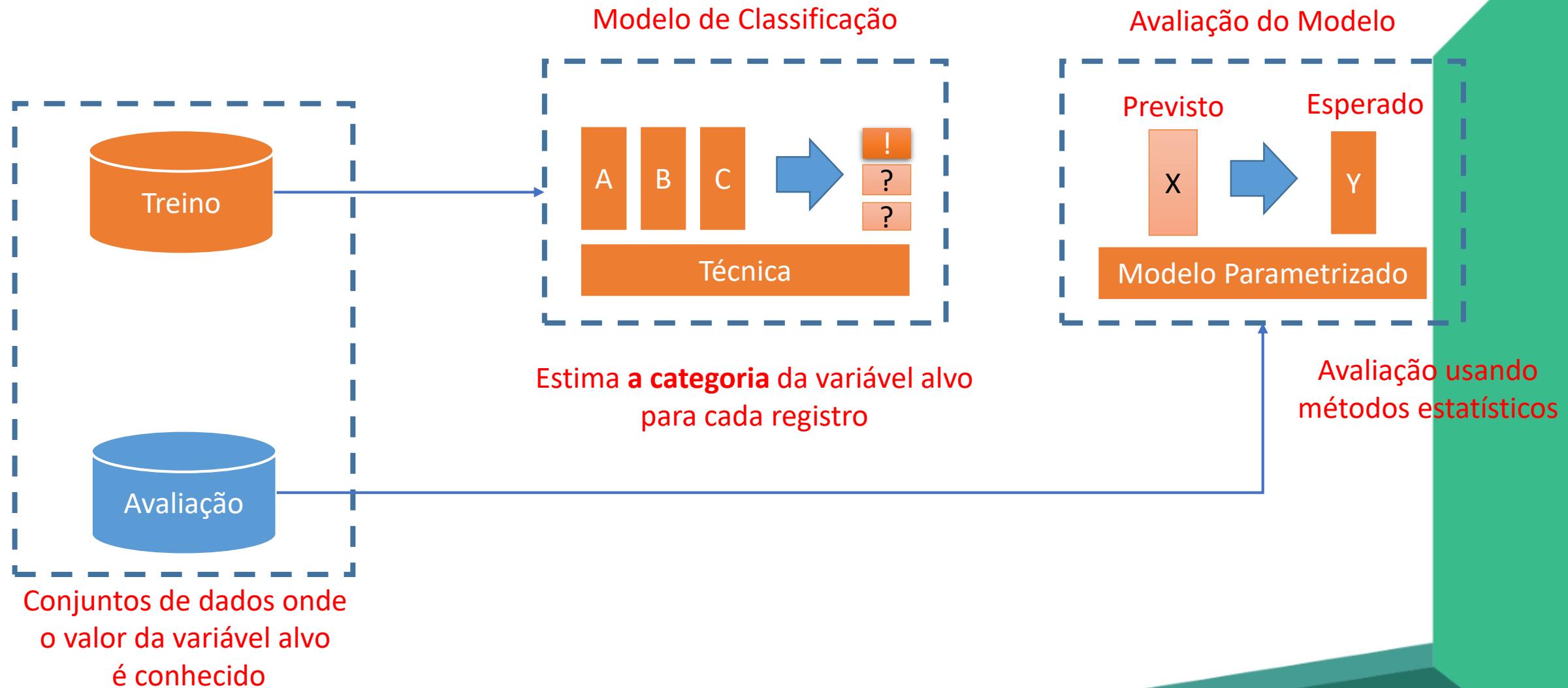
- **Binária** - O valor da variável alvo assume somente dois valores;

Ex. Filtro de Spam (Spam, E-mail Normal)

- **Multi-classes** - O valor da variável pode assumir mais de dois valores.

Ex. Classificação do Conteúdo do Email (Noticia, Pessoal, Compras e etc.)

Modelo de Classificação



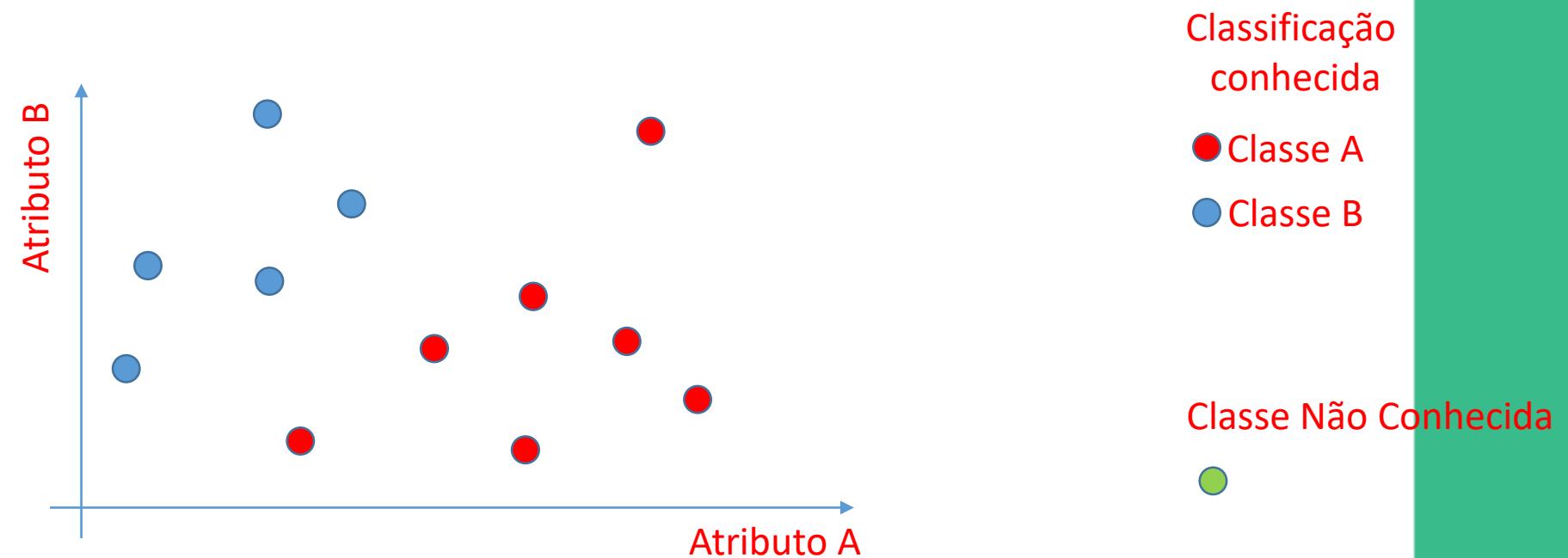
Matriz Confusão

- A matriz de confusão exibe o numero de previsões corretas e incorretas realizadas pelo modelo comparada as classificações do conjunto de dados de testes

		Classe Prevista	
		SPAM = 1	NÃO SPAM = 0
Classe Real	SPAM = 1	512	25
	NÃO SPAM = 0	10	725

Algoritmos de Classificação K-NN

- k-NN é um tipo de algoritmo de aprendizagem de máquina baseada em exemplos



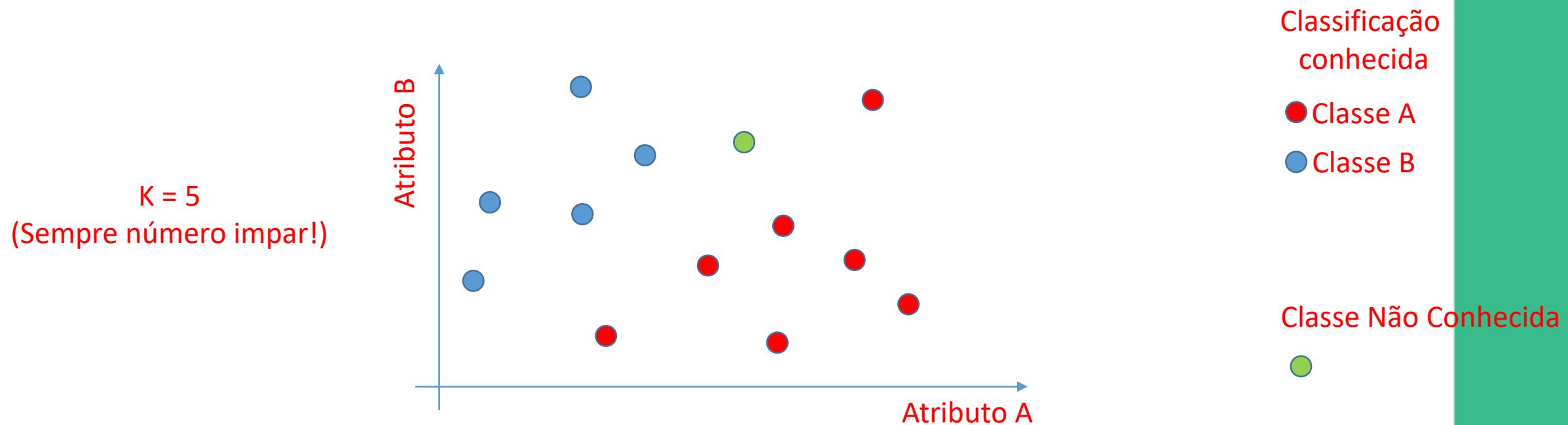
Algoritmos de Classificação K-NN

- k-NN é um tipo de algoritmo de aprendizagem de máquina baseada em exemplos



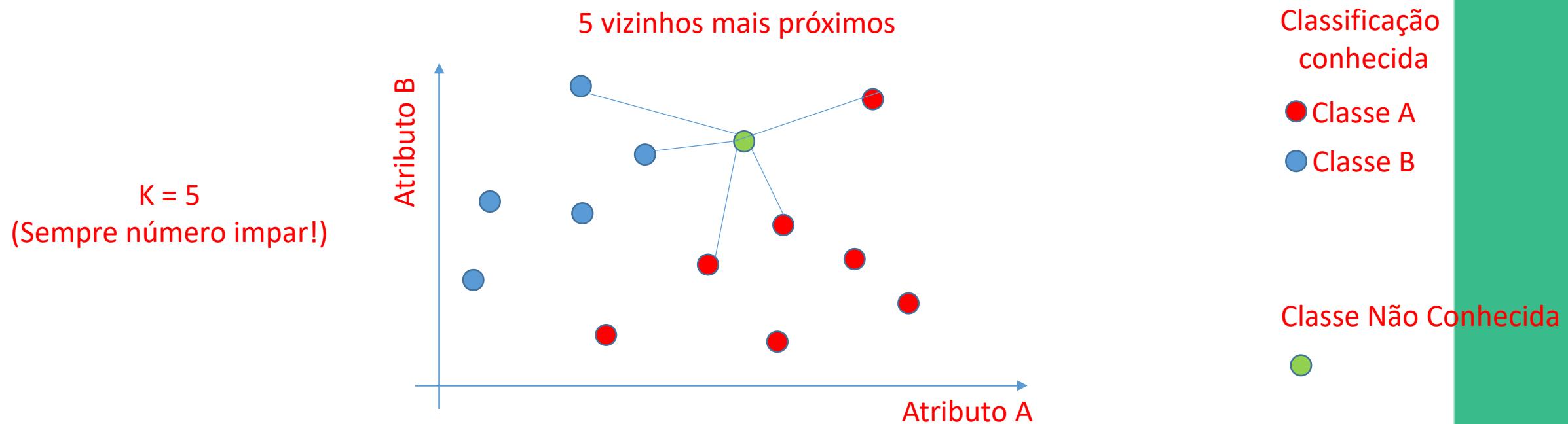
Algoritmos de Classificação K-NN

- k-NN é um tipo de algoritmo de aprendizagem de máquina baseada em exemplos.



Algoritmos de Classificação K-NN

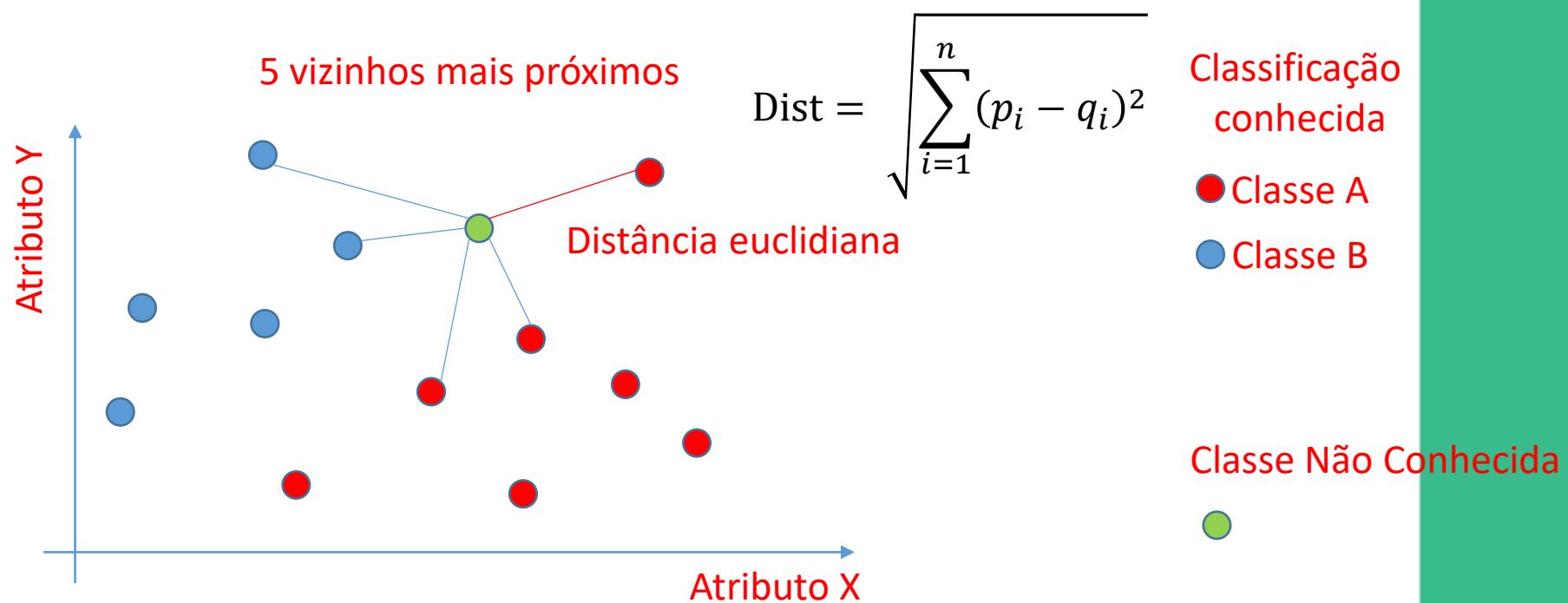
- k-NN é um tipo de algoritmo de aprendizagem de máquina baseada em exemplos.



Algoritmos de Classificação K-NN

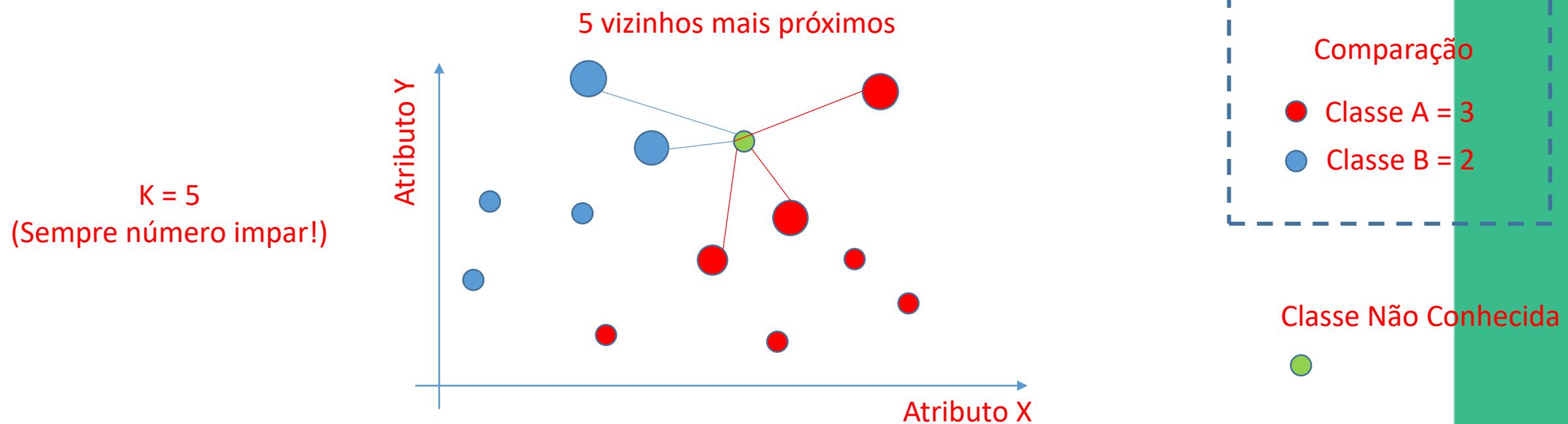
- k-NN é um tipo de algoritmo de aprendizagem de máquina baseada em exemplos

K = 5
(Sempre número ímpar!)



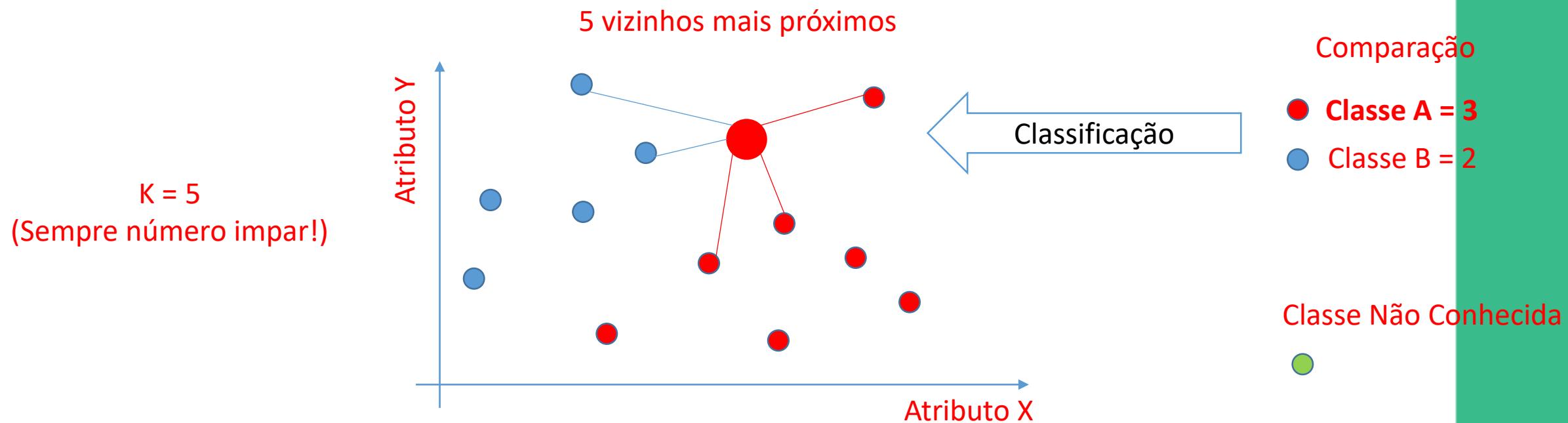
Algoritmos de Classificação K-NN

- k-NN é um tipo de algoritmo de aprendizagem de máquina baseada em exemplos



Algoritmos de Classificação K-NN

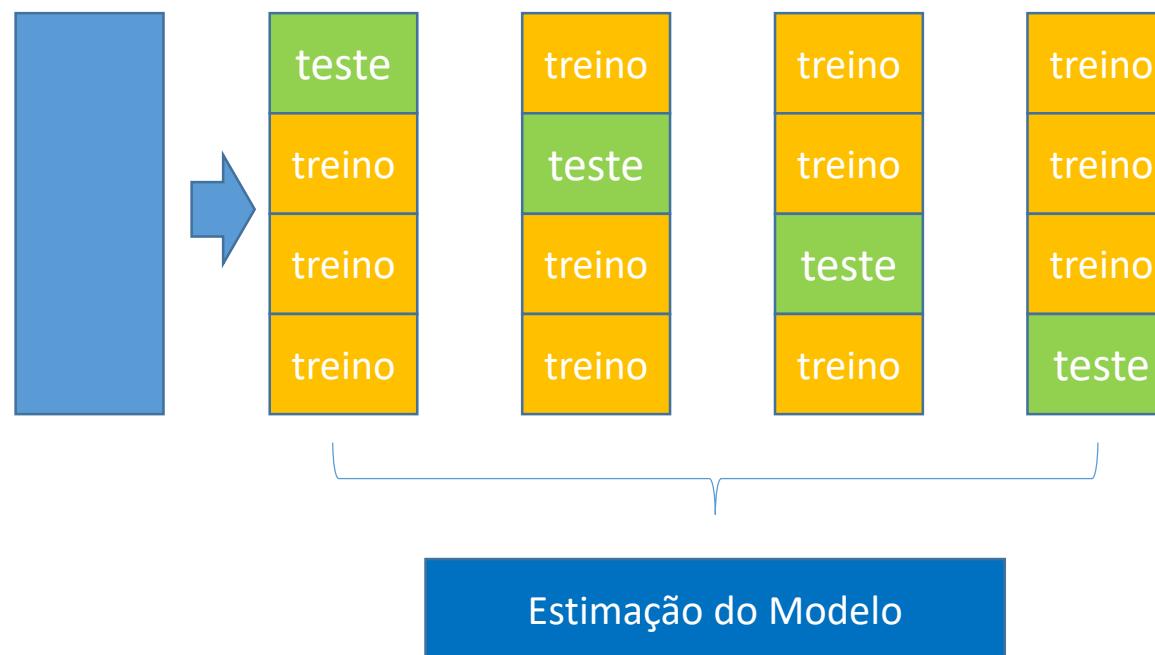
- k-NN é um tipo de algoritmo de aprendizagem de máquina baseada em exemplos



Validação Cruzada

- Técnica para avaliar um **modelo de predição** a partir de um conjunto de dados

Conjuntos de dados onde
o valor da variável alvo
é conhecido



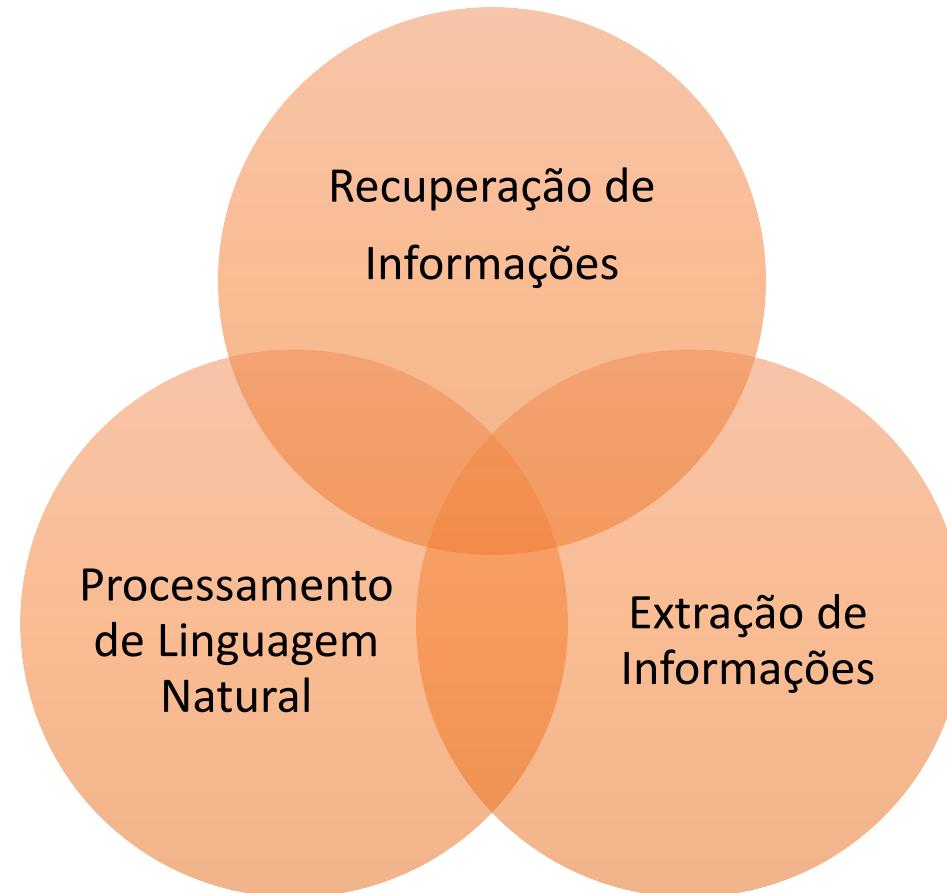
Agrupamentos

- O modelo permite **encontrar agrupamentos naturais** nos dados baseados em critérios de similaridade nos elementos do conjunto de dados;
- **Aplicação:** segmentação demográfica, segmentação de clientes.

Algoritmo K-Means



Mineração de Textos



Dados Gerados em Formato de Texto

- 85 % das informações de negócio são gerados em formato de texto;
- O principal conteúdo da web é texto;
- Notícias, Facebook, Twitter ...

Análise de Dados Não Estruturados

- A informação armazenada em formato de textos (dados não estruturados) **não podem ser processadas pelos métodos tradicionais** que lidam com texto como simples cadeias de caracteres;
- Os paradigmas tradicionais de programação baseados em lógica possuem grande dificuldade em capturar as características de **incerteza e ambiguidade** nos textos.

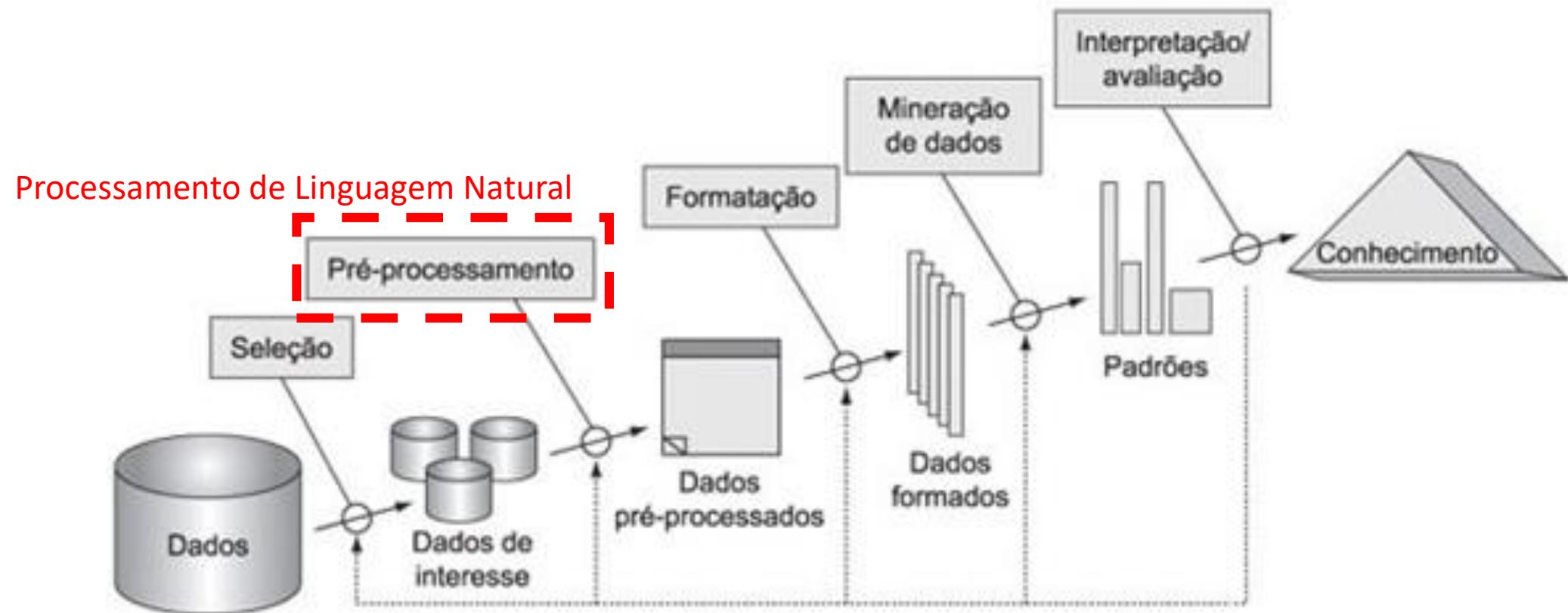
Abordagens de Mineração de Textos

- **Mineração de Textos = Extração de Informações** – Assume que a Mineração de Textos essencialmente corresponde a extração de informações (fatos) de textos;
- **Mineração de Textos = Processo de KDD** – Parte do processo de KDD especializada em encontrar informação em grandes coleções de texto;

Abordagens de Mineração de Textos

- **Mineração de Textos = Especialização de Mineração de Dados –** Aplicação de algoritmos e métodos dos campos de aprendizado de máquina e estatística aplicada a textos para encontrar padrões úteis. Métodos de processamento de linguagem natural e representação são utilizadas como etapa de pré-processamento.

Processo de Mineração de Textos



Aplicações e Usos da Mineração de Textos

- Análise de Fraudes;
- Filtro de Spam;
- Extração de Conceitos;
- Classificação automática de documentos;
- Avaliação de Opinião.

Pré-Processamento do Texto

- Estruturar o texto em um formato apropriado para aplicação de algoritmos de Mineração de Dados;
- Explora as estruturas sintáticas e semânticas do texto.

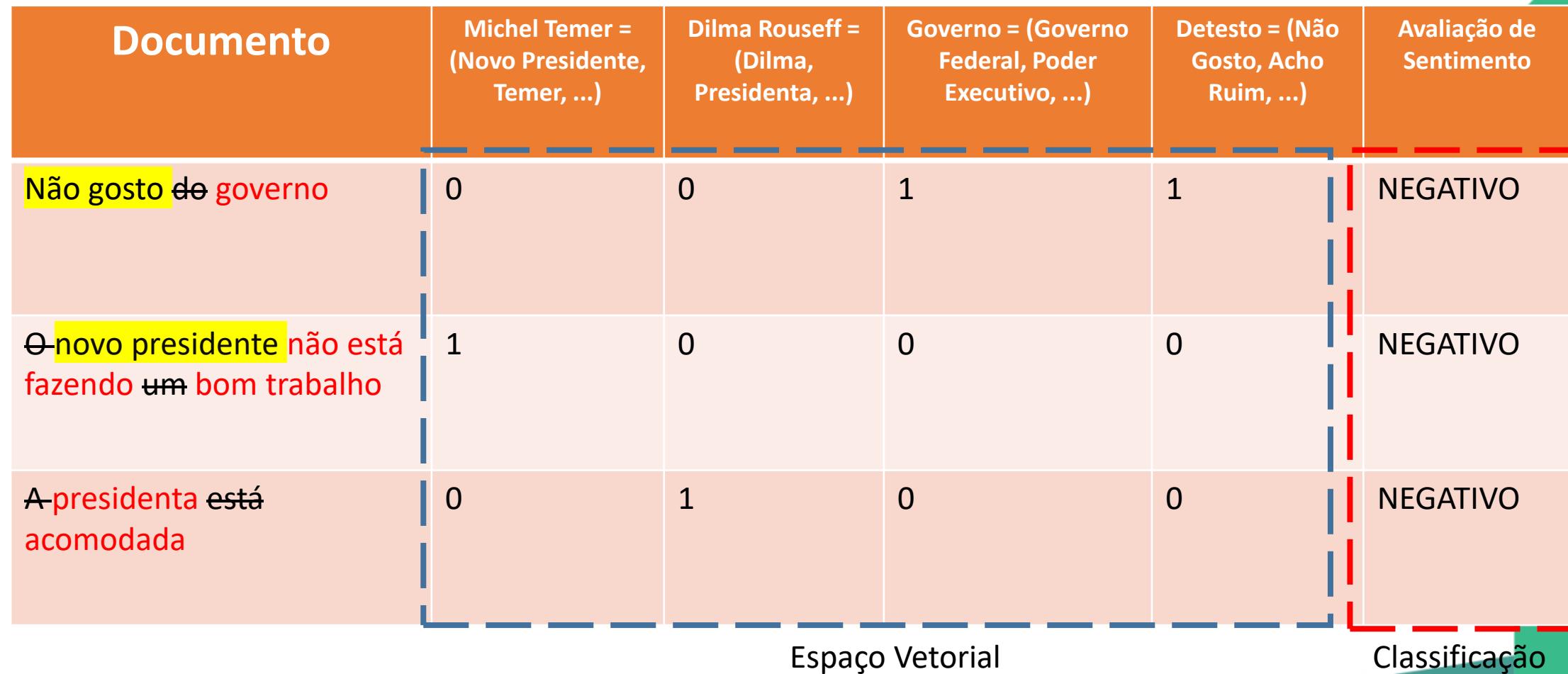
Modelo de Representação BOW

- BOW - Bag of Words;
- Abordagem onde o documento pode ser representado por um conjunto de palavras;
- Pode não preservar a estrutura sintática do texto;
- Permite abordagens de modelos probabilísticos, lógico e espaço vetorial.

Modelo de Representação BOW

Documento	Michel Temer = (Novo Presidente, Temer, ...)	Dilma Rouseff = (Dilma, Presidenta, ...)	Governo = (Governo Federal, Poder Executivo, ...)	Detesto = (Não Gosto, Acho Ruim, ...)	Avaliação de Sentimento
Não gosto do governo	0	0	1	1	NEGATIVO
O novo presidente não está fazendo um bom trabalho	1	0	0	0	NEGATIVO
A presidente está acomodada	0	1	0	0	NEGATIVO

Modelo de Representação BOW



Decodificação do texto

- Decodificação da sequência de caracteres de um documento textual.
Prioriza a transformação de um documento em formato proprietário para codificações padrões de texto (ASCII, UTF8)
- Html;
- Pdf;
- Doc.

Tokenização

Processo para a segmentação da sequência de termos em unidade semânticas menores

Abordagens:

- ✓ Palavras
- ✓ Conceitos
- ✓ Frases
- ✓ Sintagmas Nominais

Definição de Conceitos Básicos

Corpus = Conjunto de dados textuais, coleção de textos

Exemplo:

Conjunto de Notícias, Feed de Twitter, Feedback de usuários de um serviço

Definição de Conceitos Básicos

Documento = Unidade do Corpus

Exemplo:

Uma notícia específica, uma mensagem do Twitter, um feedback de usuários para um serviço específico

Definição de Conceitos Básicos

Termo = Menor unidade semântica que designa um conceito

- Pode ser composto por diversas definições a partir de um Thesaurus

Exemplo:

R(Presidente) = (Presidenta, Dilma, Dilma Rouseff, ...)

Definições Formais

D = Conjunto de Documentos

Dicionário – Conjunto T dos diferentes termos t que correm em D

$$T = \{t_1, \dots, t_m\}$$

Definições Formais

- Vetor de termos de um dado documento.

$$\vec{t_d} = (\text{tf}(d, t_1), \dots, \text{tf}(d, t_m))$$

Definições Formais

O Centroide de um conjunto X de vetores de termos é definido pela média do vetor de termos

$$\vec{t}_x = \frac{1}{|X|} \sum_{\vec{t}_d \in X} \vec{t}_d$$

Definições Formais

Subconjunto de Termos

$$tf(d, T') = \sum_{t \in T'} tf(d, t)$$

Técnicas de Redução da Dimensionalidade

- Abordagens para reduzir o tamanho do dicionário e a dimensionalidade da descrição dos documentos de uma coleção.
 - Minimizar a matriz esparsa;
 - Melhorar a representação dos conceitos.

Filtros

- Permitem a redução da dimensionalidade por meio da remoção de palavras.
- Filtro baseado em tamanho;
- Filtro baseado em dicionário;
- Stopwords – Palavras com pouco conteúdo de informação (artigos, conjunções, preposições, etc).

Lemmatization

- Processo linguístico de agrupar diferentes formas de uma palavra (lexema), preservando a sua unidade morfológica do lema
- Depende do conhecimento das regras morfológicas das construções das palavras;
- Sensível à linguagem do texto.

Stemming

- Processo para reduzir uma palavra para o seu radical, removendo o sufixo.
- Algoritmo de Poter;
- Métodos de aproximação;
- N-gram.

Estratégia de seleção por Entropia

- A seleção dos termos é baseadas em sua entropia relativa, ou ganho de informação;
- Permite a seleção de um numero fixo de termos para uma melhor performance com técnicas de agrupamentos.

$$W(t) = 1 + \frac{1}{\log_2 |D|} \sum_{d \in D} P(d, t) \log_2 P(d, t)$$

com $P(d, t) = \frac{tf(d, t)}{\sum_{l=1}^n tf(d_l, t)}$

Modelo Vetorial

- Representa os documentos em um espaço m -dimensional;
- Apesar de não utilizar informações semânticas explícitas. É um Modelo eficiente para a análise de grandes coleções de documentos;
- Cada documento é descrito como um vetor de características, permitindo que os documentos possam ser comparados com operações de vetores.

$$w(d) = (x(d, t_1), \dots, x(d, t_m))$$

Estratégia para elemento do vetor

- Binário;
- Contagem;
- Esquema de pesos.

TF-IDF

- Melhora a performance, onde o peso reflete a importância de um termo em um coleção de documentos.

$$w(d, t) = \text{tf}(d, t) \times \text{idf}(t)$$

TF-IDF

- Fator de normalização

$$idf(d, t) = \log\left(\frac{N}{n_t}\right)$$

$$w(d, t) = \frac{tf(d, t)\log\left(\frac{N}{n_t}\right)}{\sqrt{\sum_{j=1}^m tf(d, t_j)^2 (\log\left(\frac{N}{n_{t_j}}\right))^2}}$$

Cálculo de Similaridade

- Produto interno dos vetores (vetores normalizados)

$$S(d_1, d_2) = \sum_{k=1}^m w(d_1, t_k) \times w(d_2, t_k)$$

Distância Euclidiana

- Aplicação para vetores normalizados

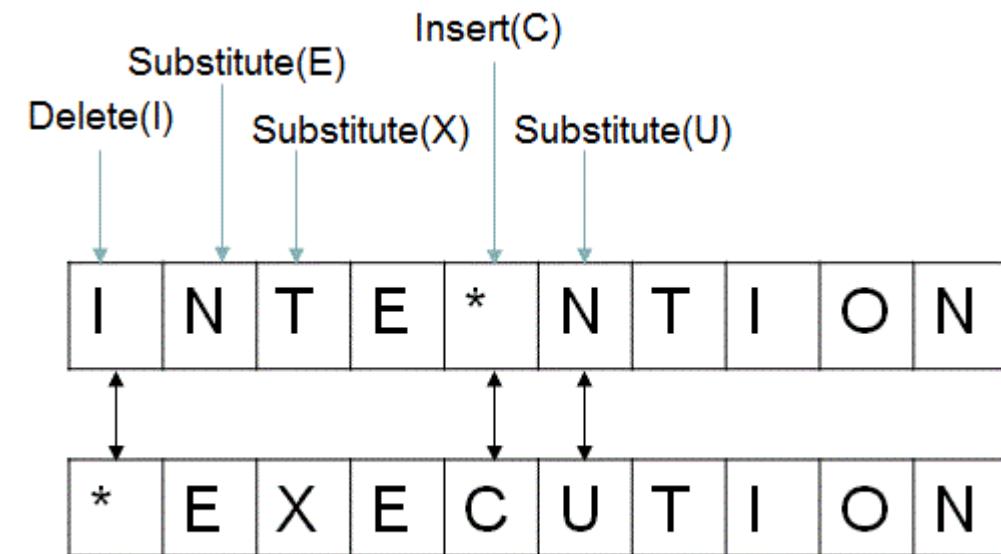
$$dist(d_1, d_2) = \sqrt{\sum_{k=1}^m |w(d_1, t_k) - w(d_2, t_k)|^2}$$

Edit Distance

- Permite quantificar o quanto duas cadeias de caracteres são diferentes

Distância de Levenshtein

Aplicações
Correção ortográfica
Bioinformática



Processamento de Linguagem Natural

- Part-of-Speech Tagging;
- Text Chunking;
- Word Sense Disambiguation;
- Parsing.

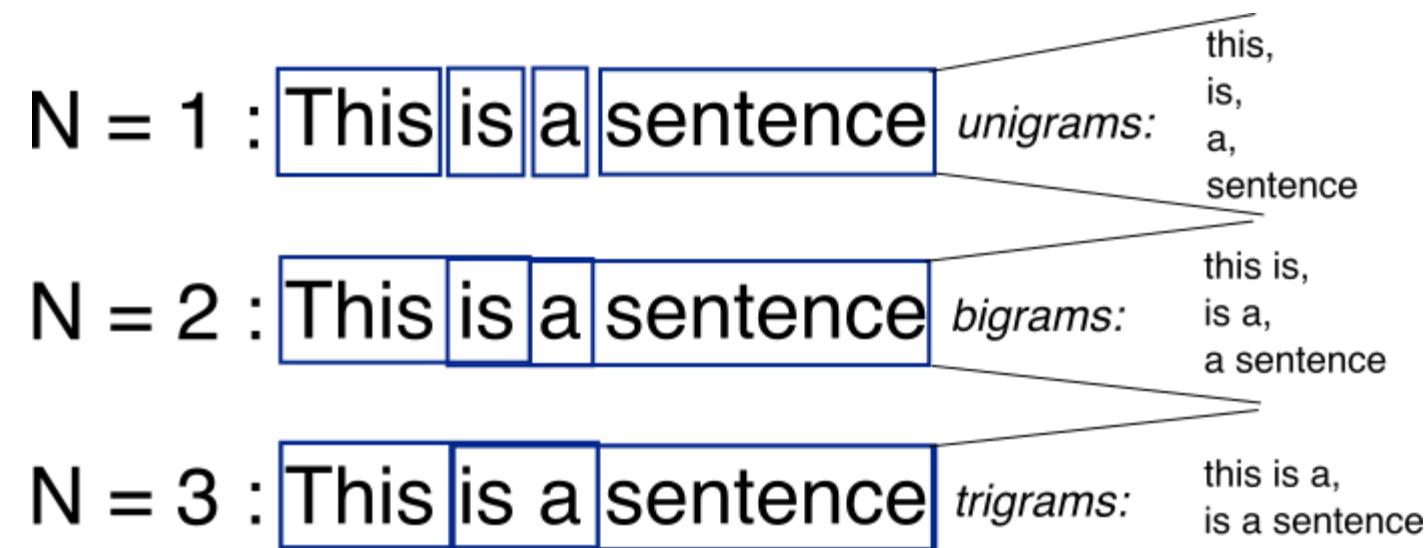
Part-Of-Speech Tag

- Determina a classe da palavra para cada termo (verbo, substantivo, adjetivo, etc);
- Depende do conhecimento da língua.

```
>>> import(nltk)  
>>> text = word_tokenize("And now for something completely different")  
>>> nltk.pos_tag(text)  
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'),  
('completely', 'RB'), ('different', 'JJ')]
```

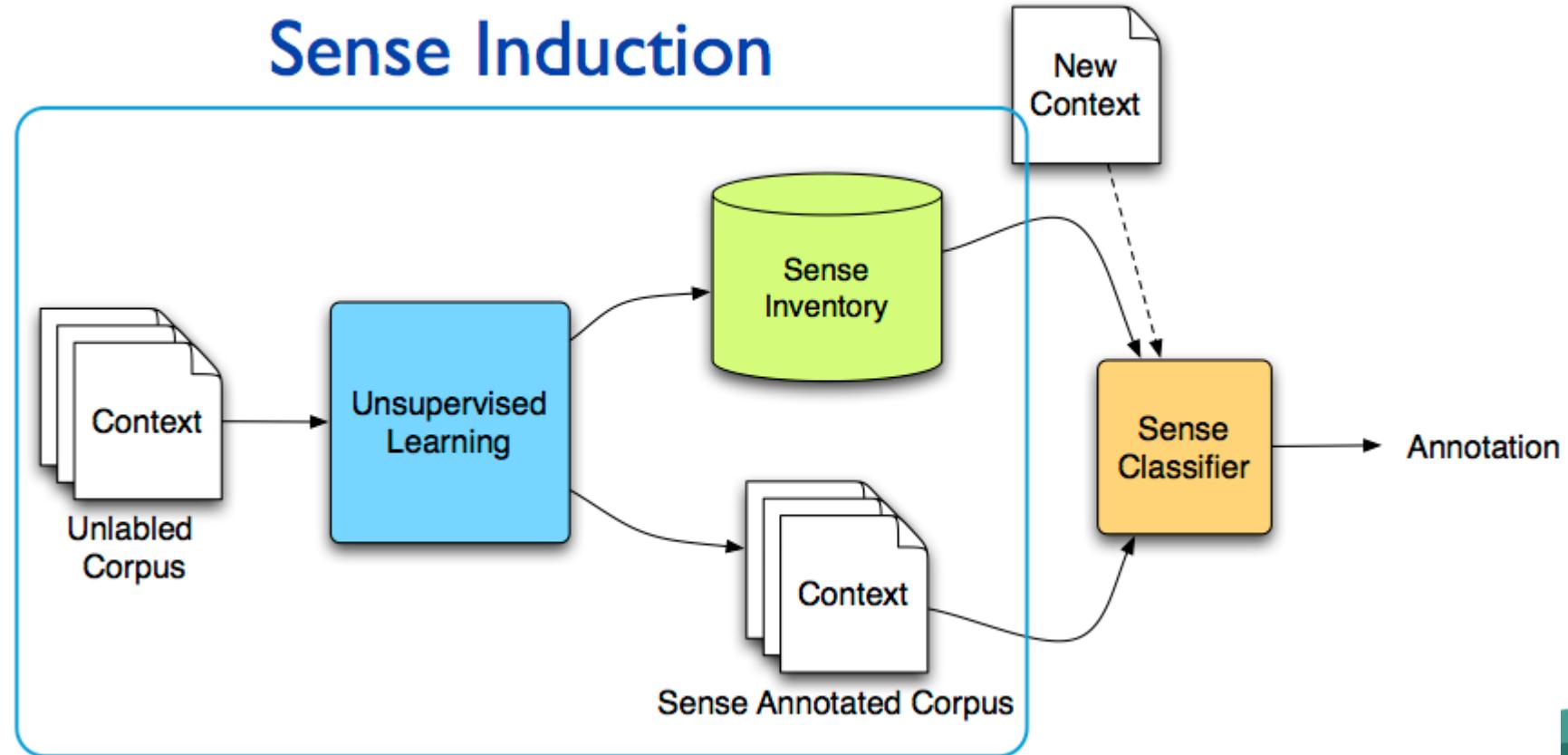
Text Chunking

- Permite agrupar palavras adjacentes em uma sentença com uma unidade semântica (classificador);
- N-gram



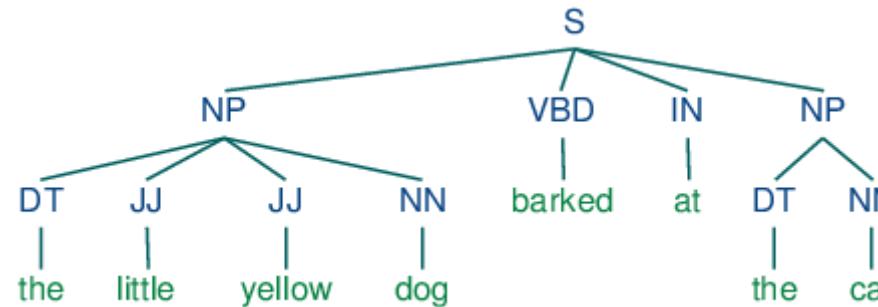
Word Sense Disambiguation

- Técnica para resolver a ambiguidade no significado de palavras ou frases



Parsing

- Produz a árvore de derivação sintática de uma sentença com a relação de cada palavra com as outras e sua função na sentença (sujeito, objeto, predicado)



Métodos de Mineração de Dados para textos

Classificação

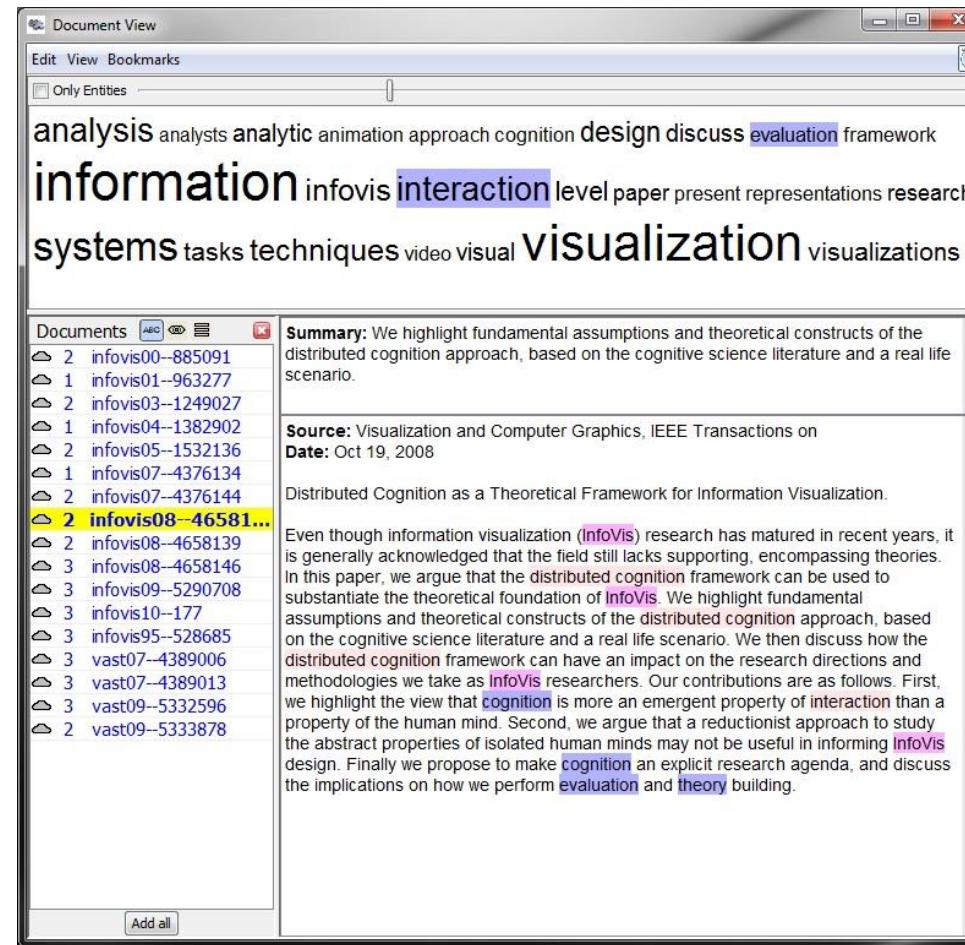
- Categorizar automaticamente documentos por conteúdo dado um conjunto de treinamento.
- Algoritmos;
- KNN;
- Arvore de Decisão.

Métodos de Mineração de Dados para textos

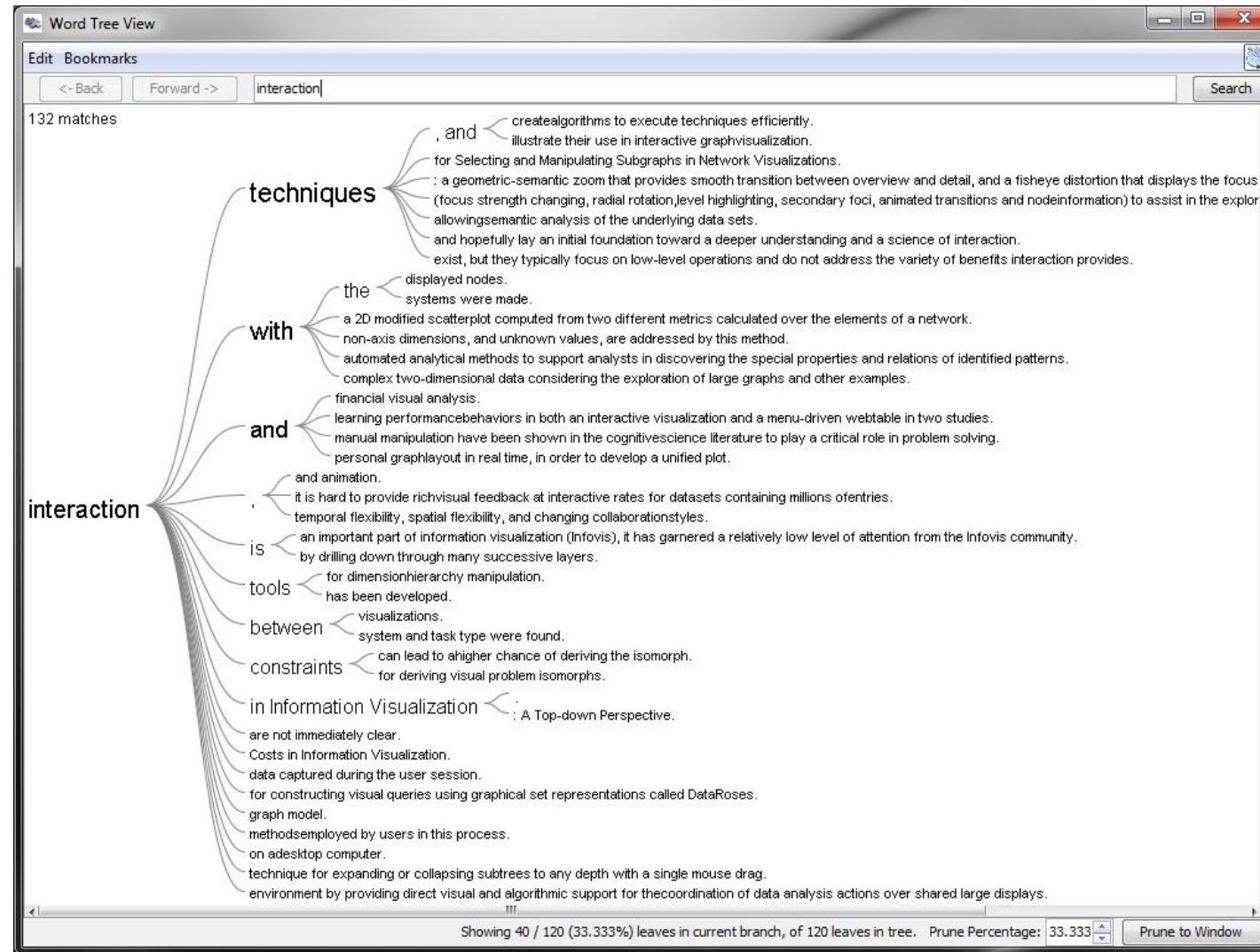
Agrupamentos

- Agrupar documentos com conteúdo similar
 - K-means;
 - Mapas auto organizados;
 - Expectation Maximization.

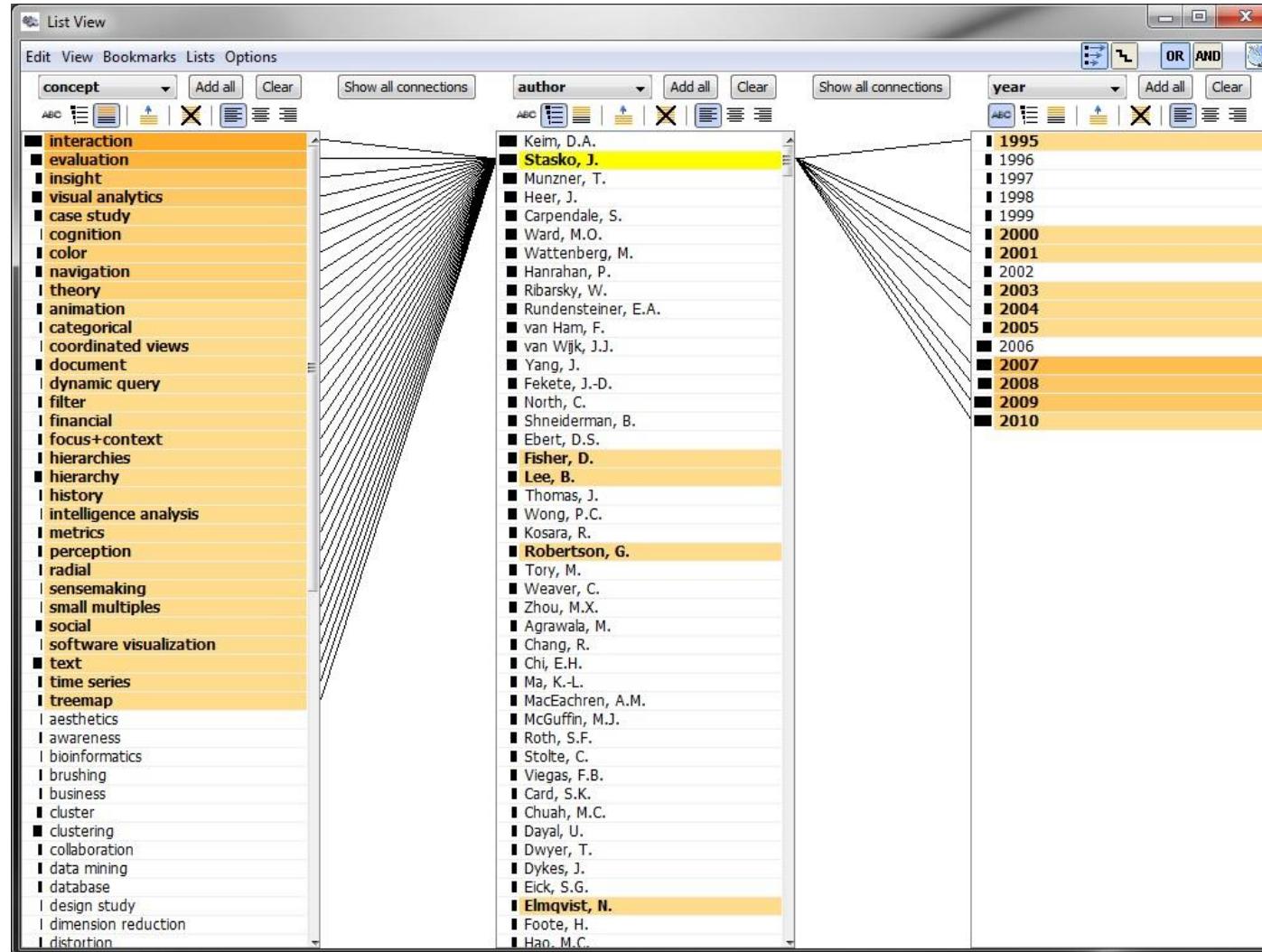
Métodos de Visualização de Textos



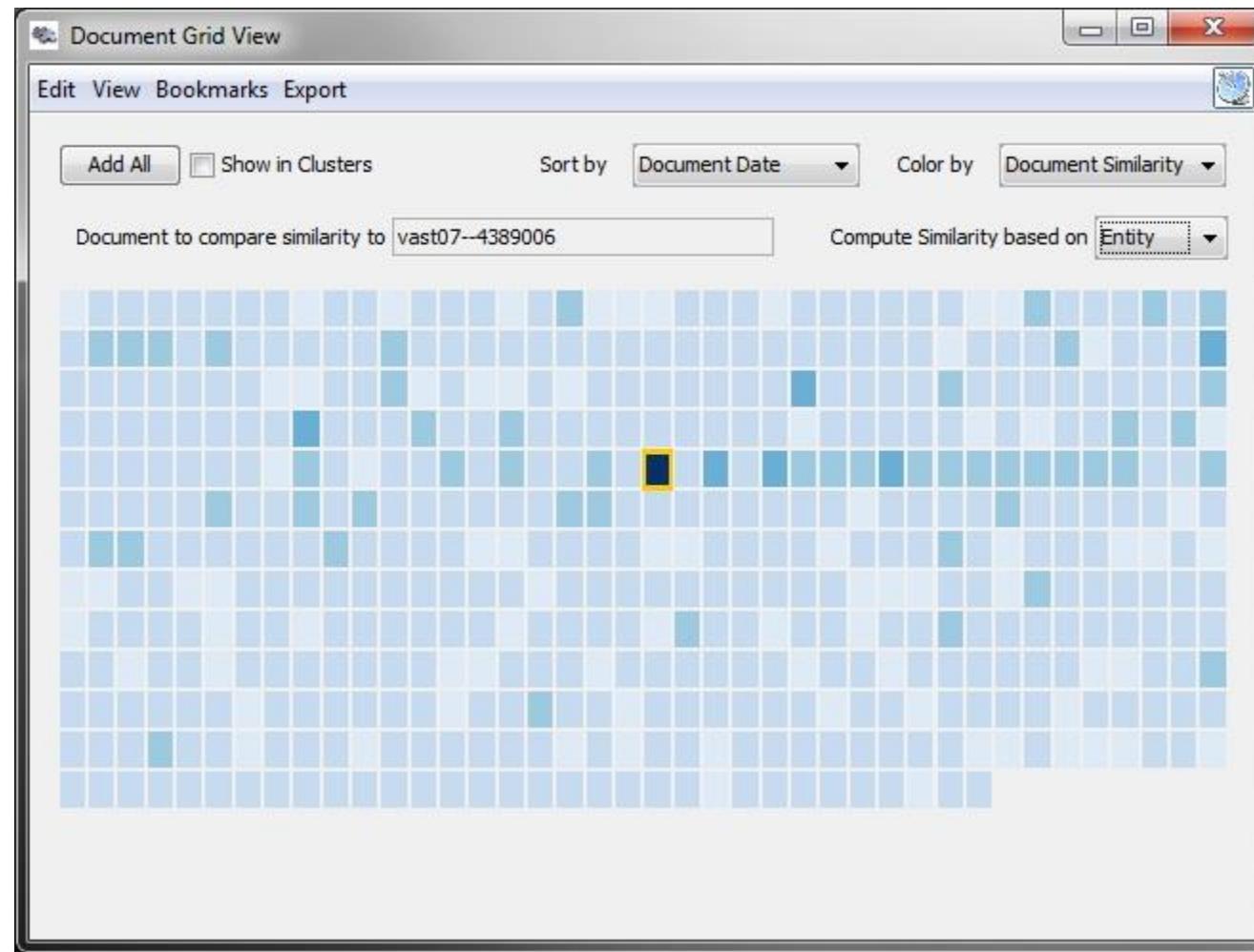
Métodos de Visualização de Textos



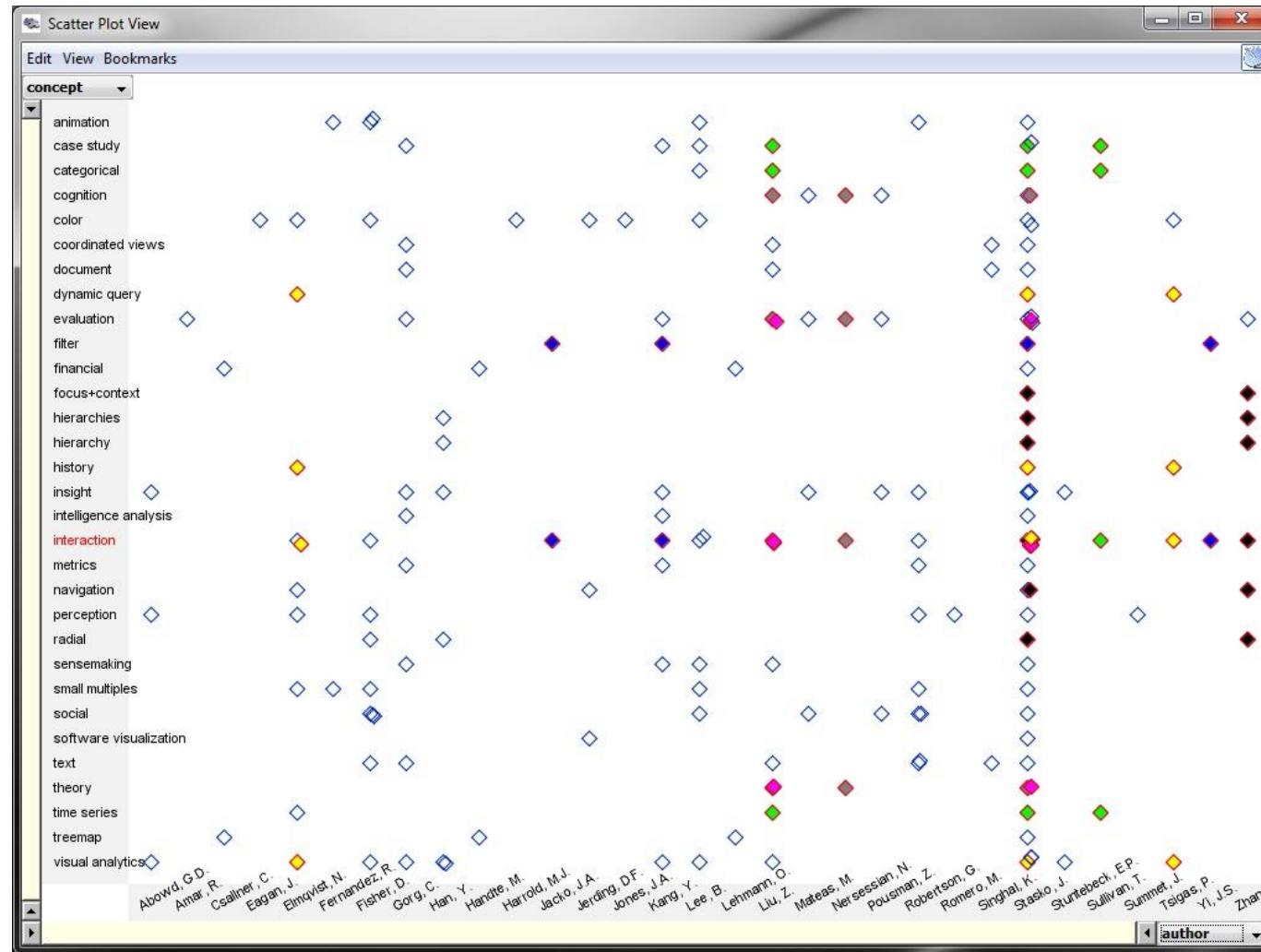
Métodos de Visualização de Textos



Métodos de Visualização de Textos



Métodos de Visualização de Textos

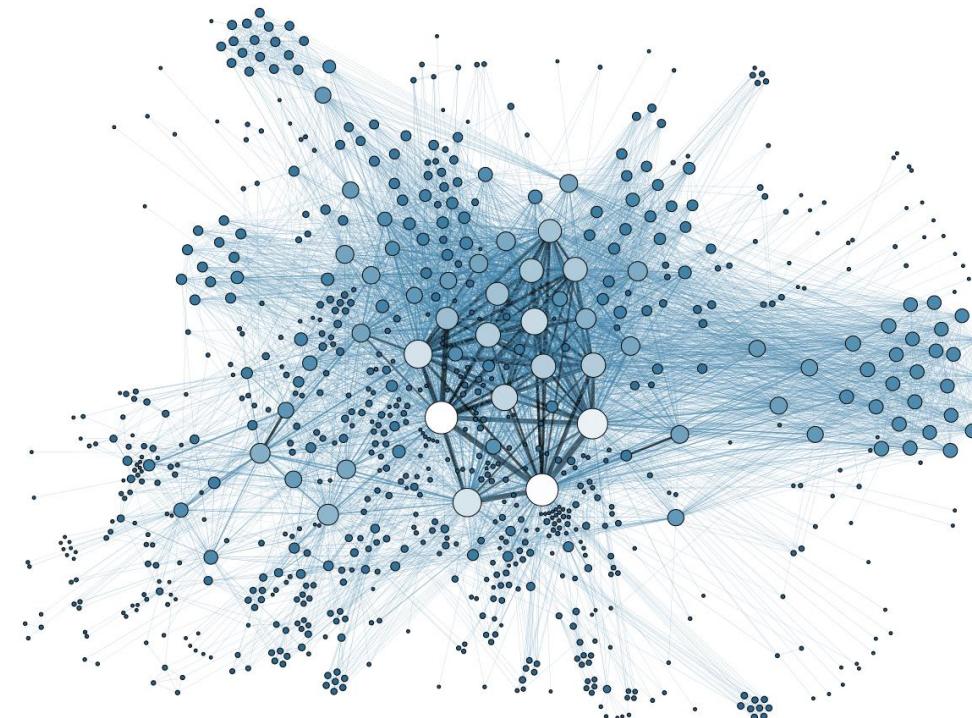


Ferramentas para Análise de Textos

- GATE;
- NLTK;
- Jigsaw;
- Rapid Miner;
- Wordnet.

Análise de Redes Sociais

- Uma rede social é uma estrutura composta por pessoas ou organizações que compartilham valores e objetivos comuns e estão conectadas por vários tipos de relações



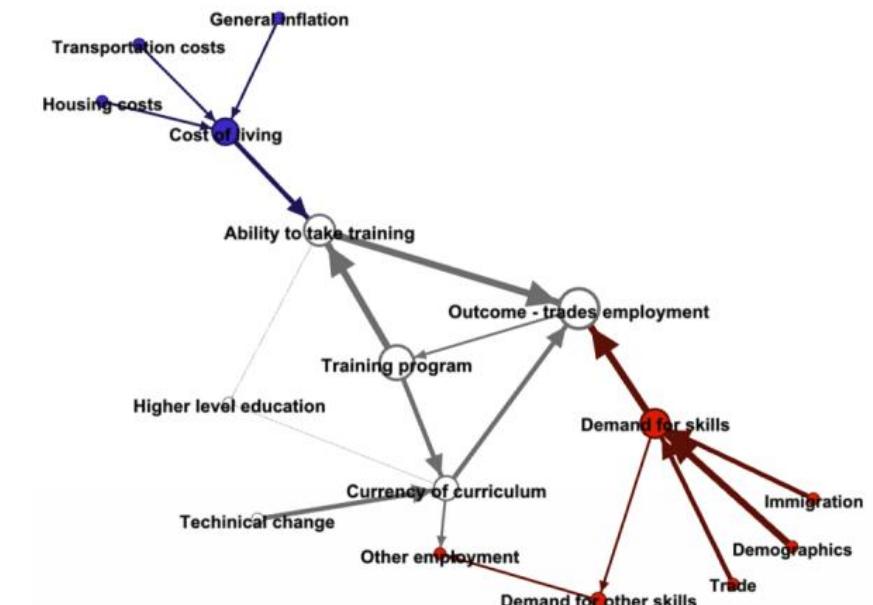
Análise de Redes Sociais

Conceito de Sociologia:

- Redes são **sistemas abertos**, possibilitando relacionamentos horizontais e não hierárquicos entre os participantes.

Elementos de Uma Rede

- Nós ou atores;
- Vínculos;
- Fluxos de Informação (unidirecional ou bidirecional).



Análise de Redes Sociais

- É um processo para mapear e estudar redes de relacionamento entre as pessoas, grupos, organizações, entre outros, por meio de métricas e visualização de grafos.
- Abordagem da Sociologia, da Psicologia Social e da Antropologia.

Métricas de Conexão

- **Similaridade** – Permite estudar as relações sociais, identificando quais atores comportam semelhanças entre si e quais comportam diferenças. As semelhanças podem ser de diversas ordens: idade, gênero, sexo etc;
- **Multiplexidade** – Visão relacionada aos vários tipos de relações entre os atores e que implica a existência de mais de um tipo de ligação. Essas ligações podem ser associadas à força do relacionamento;
- **Reciprocidade** – Métrica que permite medir a reciprocidade da relação/interação entre dois nós;
- **Propinquidade** – Mede a tendência de um nó para ter relações com os nós que lhe são geograficamente mais próximos.

Métricas de Distribuição

- **Ponte** – Uma ponte traduz-se por ser uma ligação entre dois nós ou dois grupos de nós, conforme a teoria dos grafos;
- **Centralidade** – Engloba diversas métricas que permitem identificar e quantificar a importância de um nó ou um grupo de nós em uma rede;
- **Densidade** – Permite medir o número de ligações diretas existentes mediante a um número total de ligações possíveis;
- **Distância** – Permite medir o número total de passos de um extremo a outro na rede ou entre dois nós numa rede;

Métricas de Distribuição

- **Vazio Estrutural** – Métrica que permite identificar a inexistência de ligações entre dois nós numa rede;
- **Força de Ligação (dos Laços)** – Define-se pela combinação de vários fatores: tempo, intimidade, intensidade emocional e reciprocidade (mutualidade). As ligações consideradas fortes estão associadas a hemofilia, propinquidade e transitividade. Já ligações consideradas fracas estão normalmente associadas às pontes.

Métricas de Segmentação

- **Grupos** – Identificar quais são os tipos de grupos e sua similaridade dentro da rede;
- **Coeficiente de Agrupamento (*Clustering*)** – Permite medir o grau pelo qual os nós tendem a agrupar-se (formando *clusters* ou aglomerados);
- **Coesão** – Métrica que permite medir o grau em que os nós se encontram diretamente ligados entre si por meio de ligações coesas.

Visual Analytics

- É o campo de pesquisa que permite o raciocínio analítico por meio de interfaces visuais interativas;
- As ferramentas permitem a manipulação de estruturas, formas e cores para revelar padrões implícitos nos dados.

Deep Learning

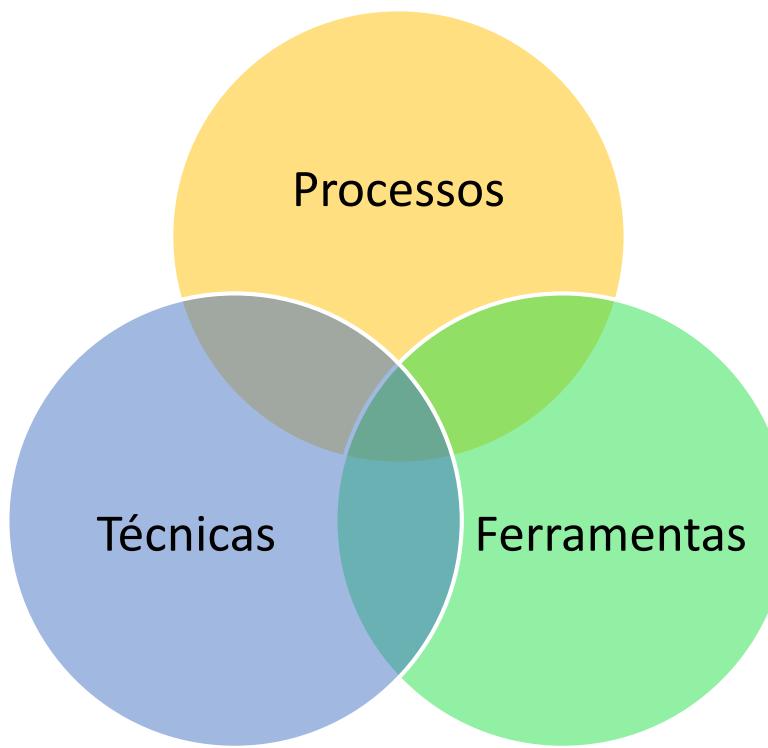
- Conjunto de técnicas avançadas que permite projetar sistemas inteligentes que aprendem a partir de conjuntos de dados complexos e de grande escala;
- Aplica conceitos do estado-da-arte de Redes Neurais;
- Permite que máquinas possam “aprender a interpretar”;
- TensorFlow da Google.

Ética

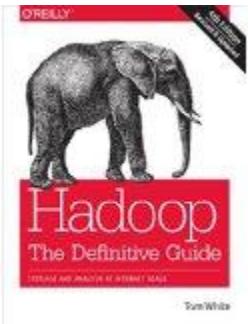
- As pessoas ainda não compreendem o poder dos dados e não conseguem julgar como seus dados deveriam ser usados;
- O real valor do Big Data não está disponível para as pessoas, somente para grandes corporações e governos;
- Segurança x Liberdade.

Agradecimentos

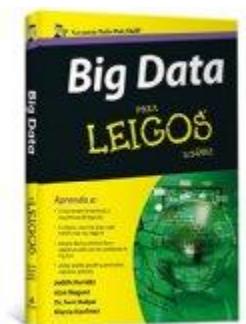
Próximos passos:



Referências

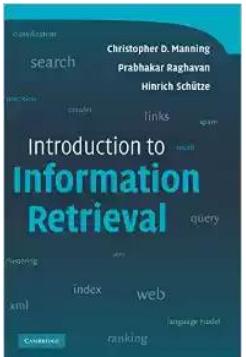


Hadoop: The Definitive Guide, por Tom White.

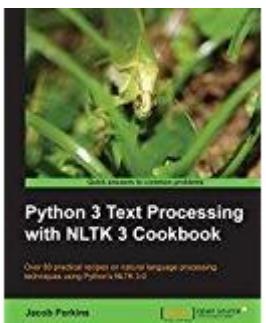


Big Data Para Leigos, por Judith Hurwitz e Alan Nugent.

Referências

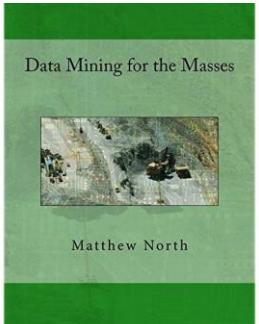


Introduction to Information Retrieval, por Christopher D. Manning.

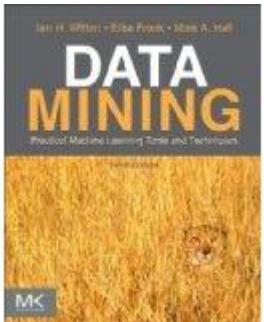


Python 3 Text Processing with NLTK 3 Cookbook, por
Jacob Perkins.

Referências

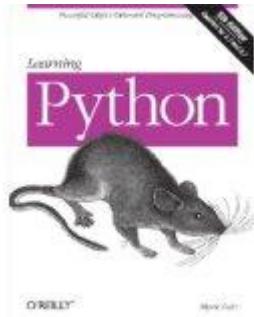


Data Mining for the Masses, por Matthew North.

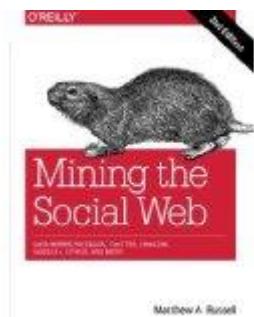


Data Mining: practical Machine Learning Tools and Techniques, por Ian H. Witten e Eibe Frank.

Referências



Learning Python, por Mark Lutz.



Mining the Social Web: data Mining Facebook, Twitter, LinkedIn, Google+, Github, and More, por Matthew A. Russell.