Introdução à mineração de dados

Disciplina: Mineração de Dados

Prof. Braian Varjão

Agenda

1. Questões existenciais;
2. Knowledge Discovery in Databases;
3. Técnicas de mineração de dados;
4. Fatores críticos para uma mineração de dados eficiente e eficaz;
5. Aplicações;
6. Considerações finais.

Prof. MSc. Braian Varjão

Questões existenciais

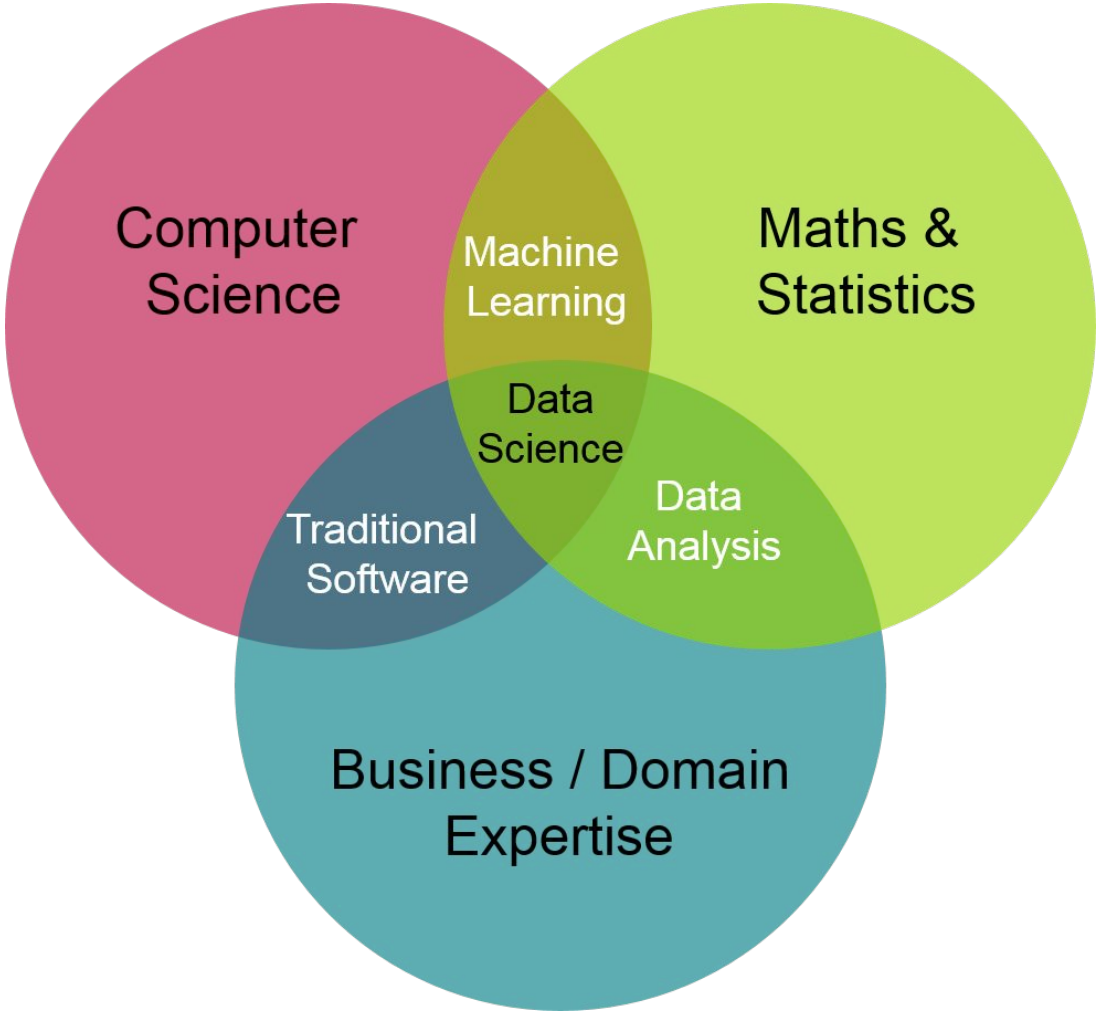
O que é Data Science?

Onde ela estava esse tempo todo?

O que mudou?

Prof. MSc. Braian Varjão

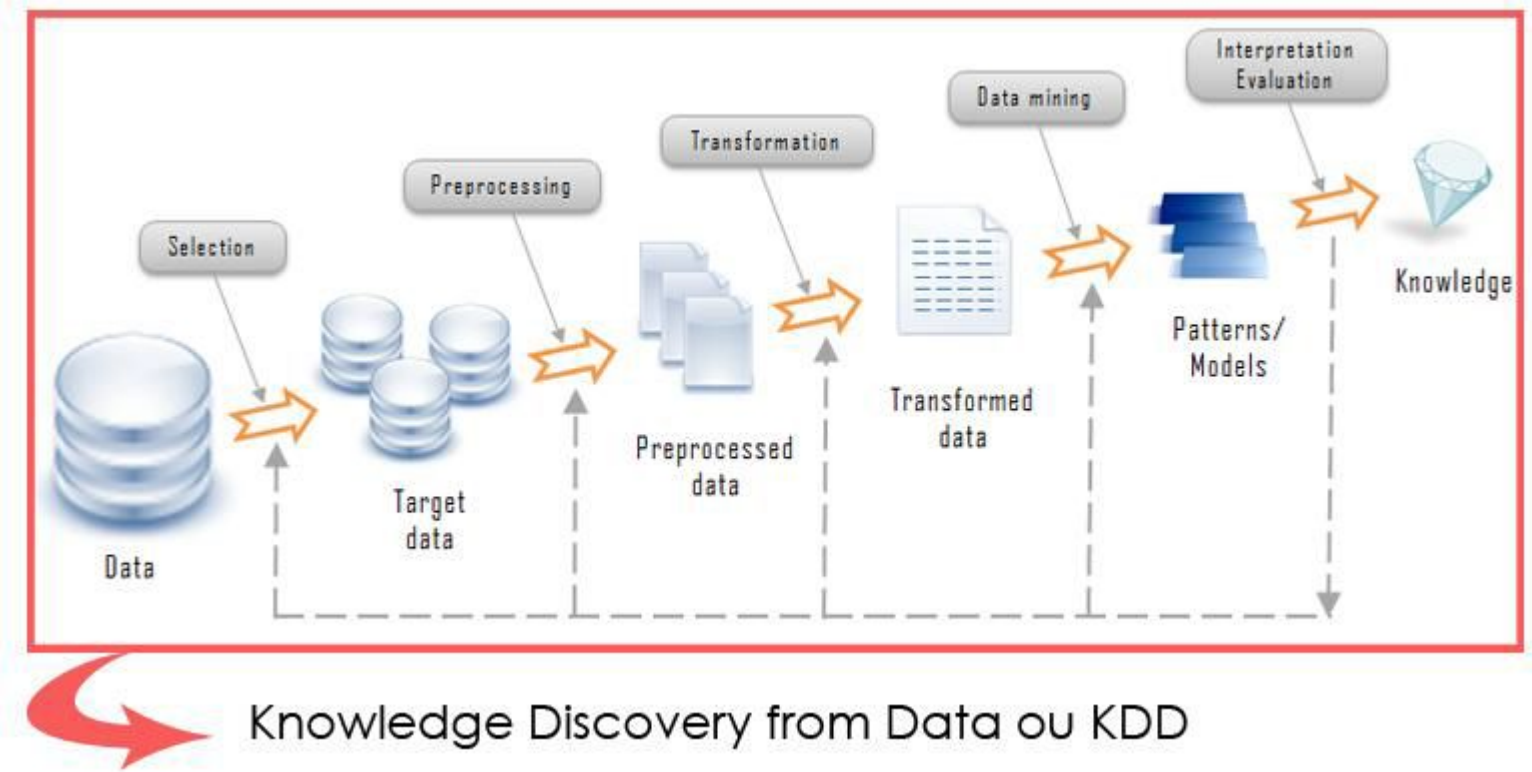
Questões existenciais



Prof. MSc. Braian Varjão

Knowledge Discovery in Databases (KDD)

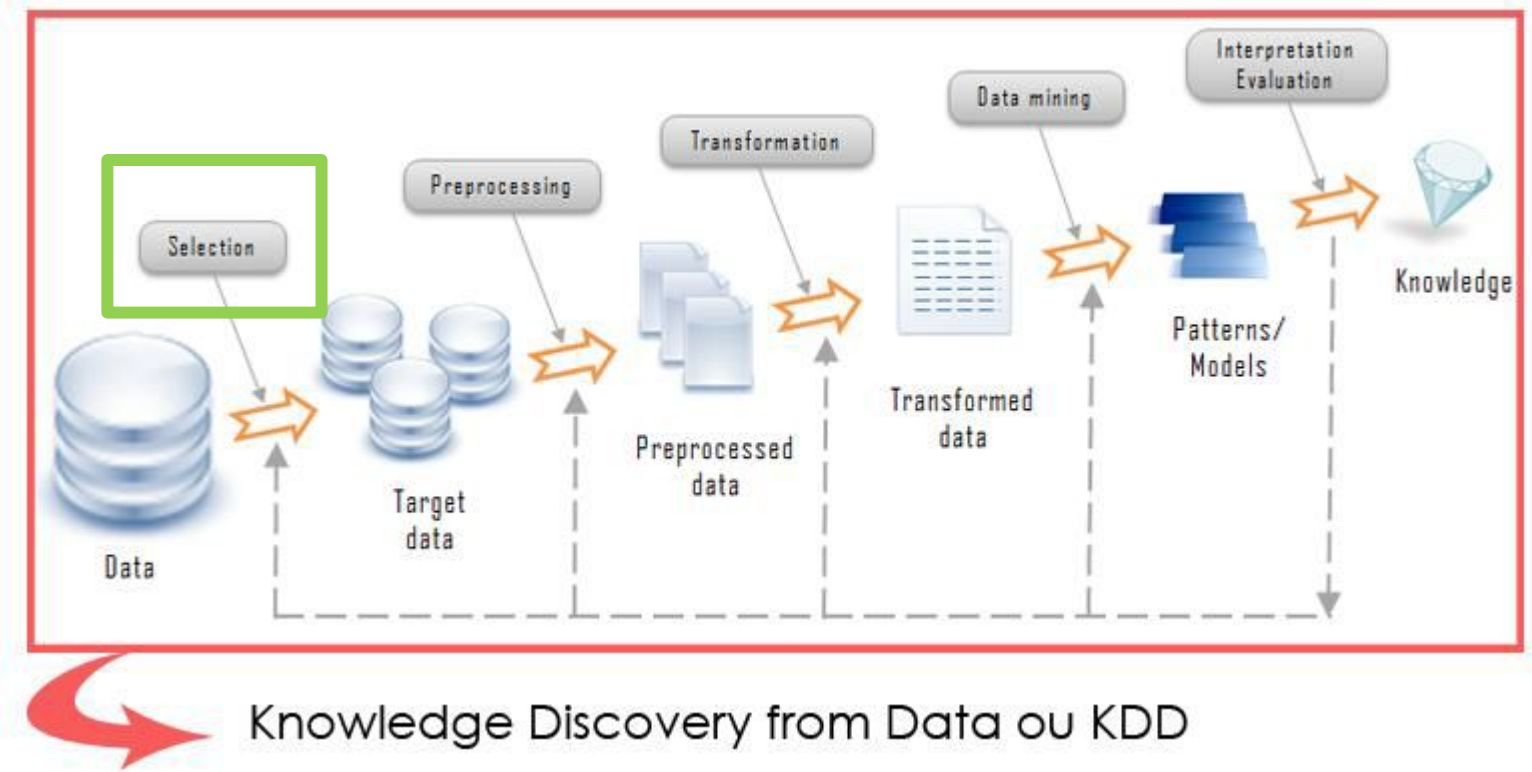
Processo de KDD



Knowledge Discovery from Data ou KDD

Prof. MSc. Braian Varjão

Processo de KDD



Prof. MSc. Braian Varjão

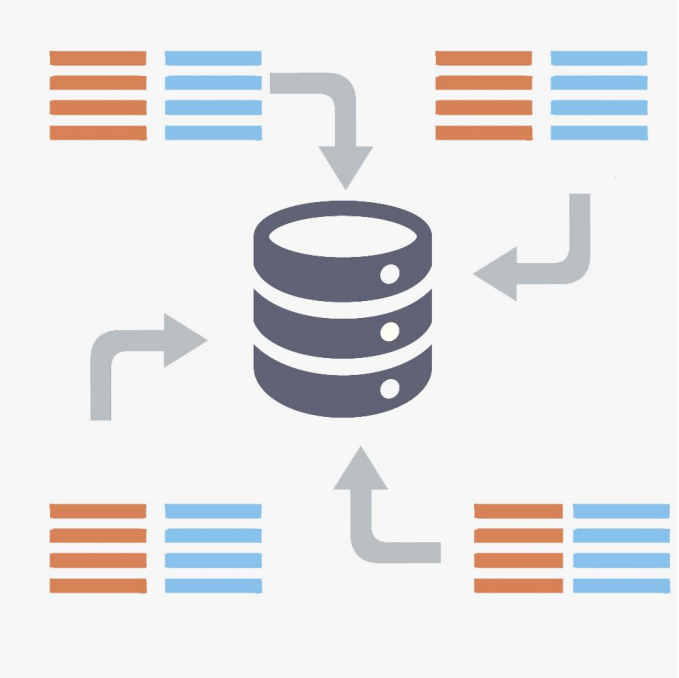
Seleção

Quais dados eu preciso?

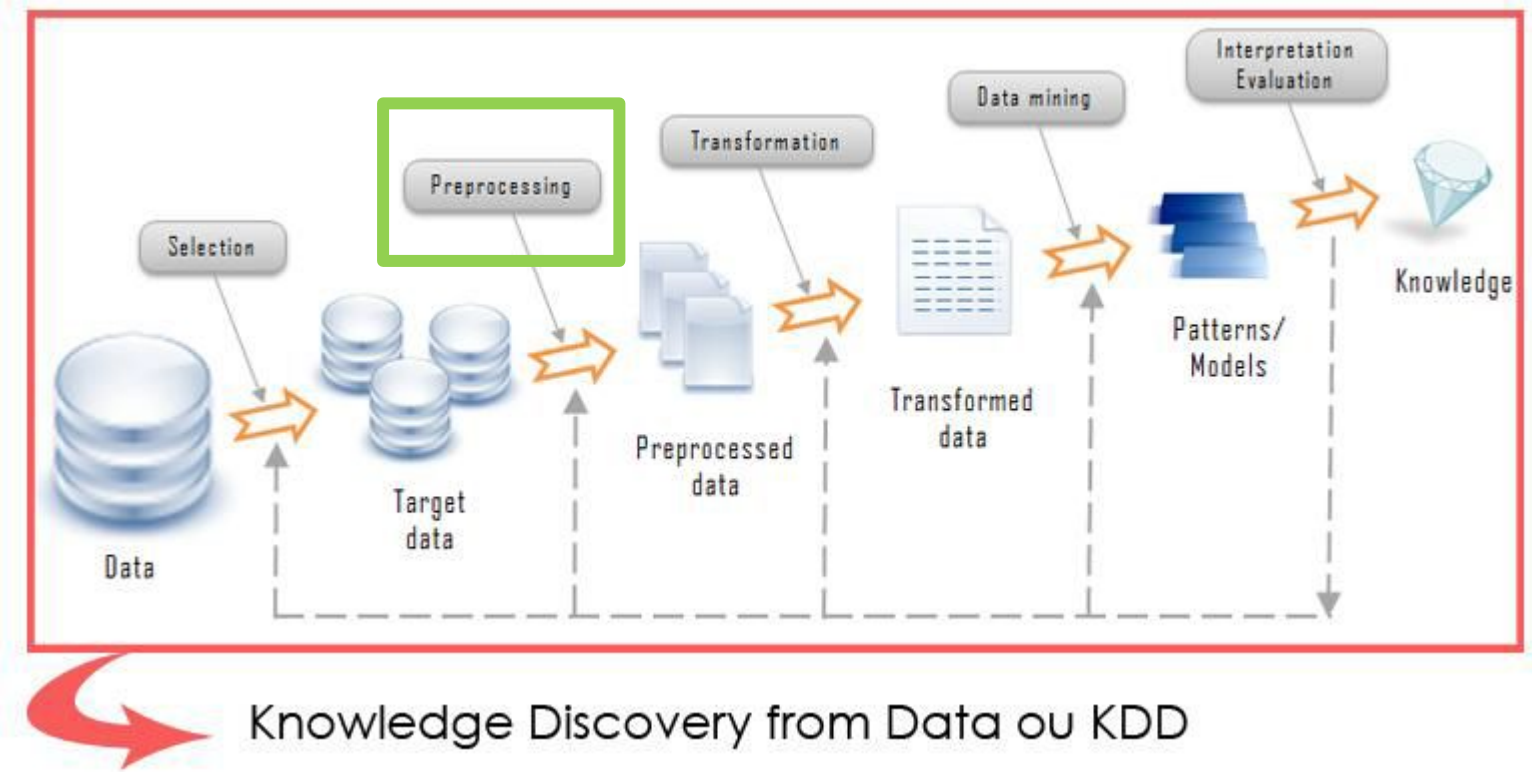
Onde eles estão?

Como obtê-los?

Como integrá-los?



Processo de KDD



Prof. MSc. Braian Varjão

Pré-processamento

Identificação e análise de variáveis

Tratamento de inconsistências

Tratamento de dados faltantes

Tratamento de outliers

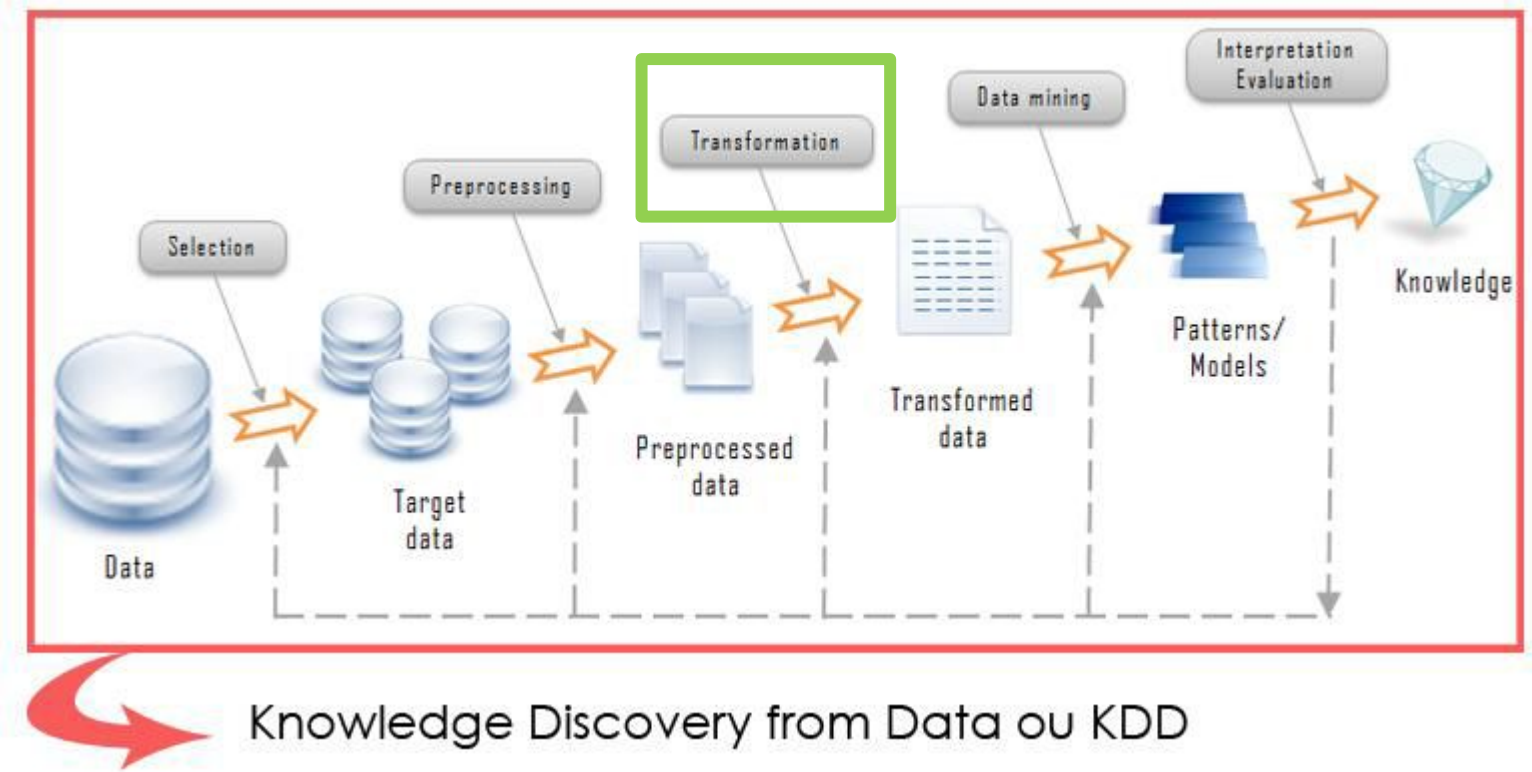
Derivação de atributos

Reamostragem

Redução de dimensionalidade

Prof. MSc. Braian Varjão

Processo de KDD

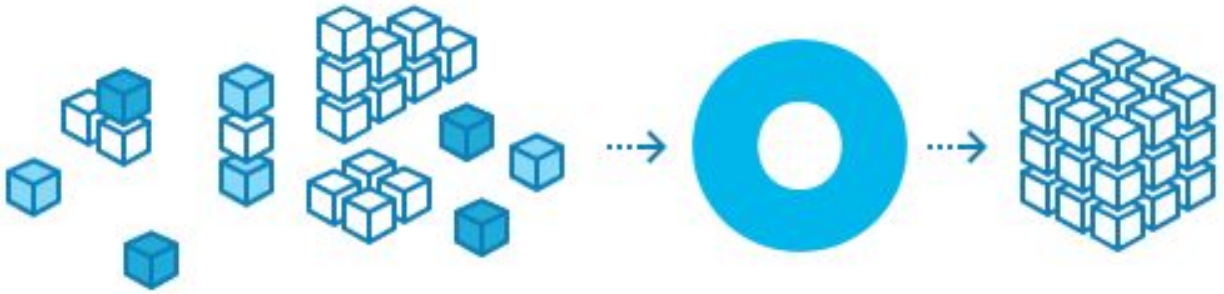


Prof. MSc. Braian Varjão

Transformação

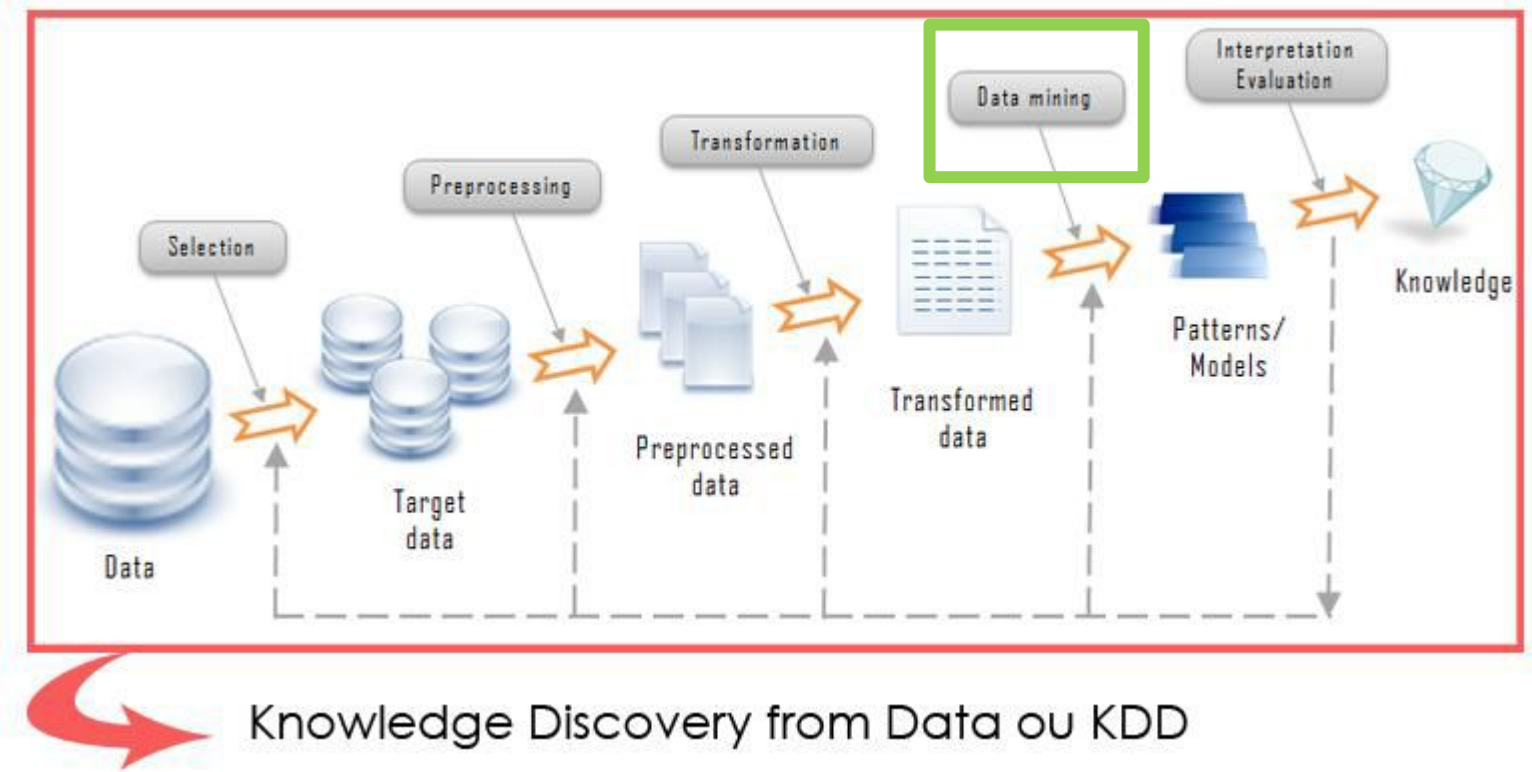
Estruturação e padronização dos dados.

- Transformação de atributos;
- Extração de atributos.



Prof. MSc. Braian Varjão

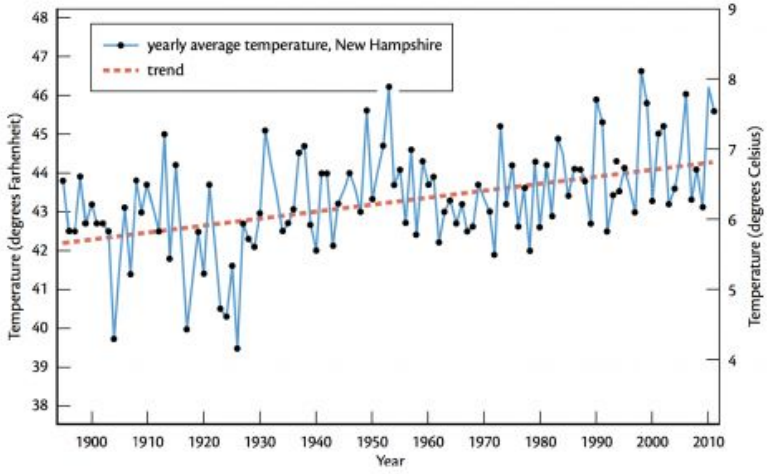
Processo de KDD

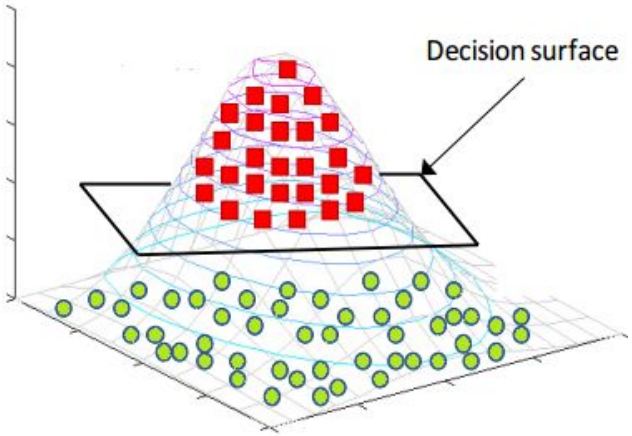


Prof. MSc. Braian Varjão

Mineração

Encontrar padrões nos dados





Banco de dados vs. Mineração de dados

Banco de dados

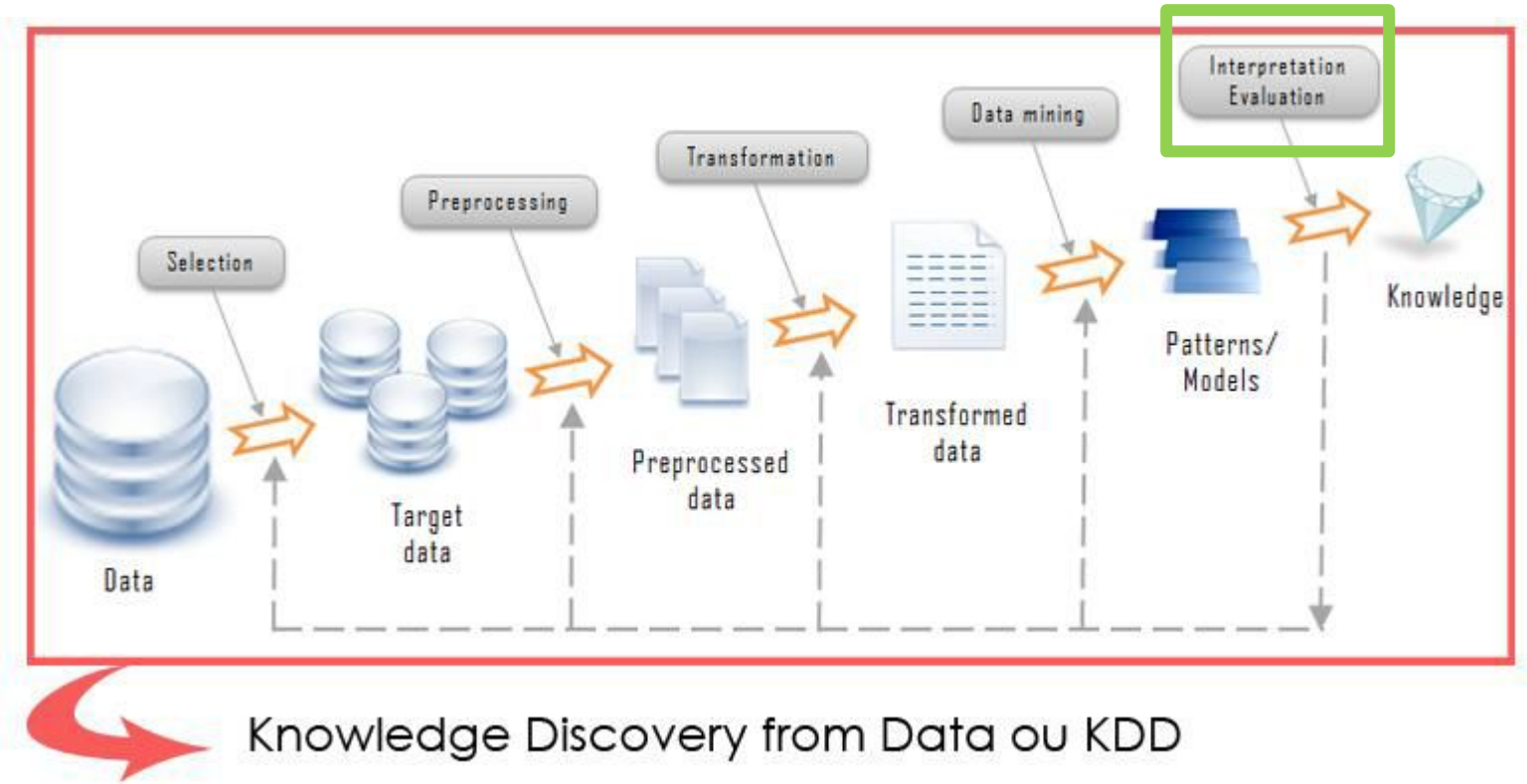
Encontre todos os clientes que vivem em Boa Vista
Encontrar todos os clientes que usam Mastercard
Encontrar todos os clientes que perderam um pagamento

Mineração de dados

Encontrar os clientes que provavelmente perderão um pagamento (**classificação**)
Agrupar os clientes com hábitos de compra mais simples (**agrupamento**)
Listar os itens que são frequentemente comprados com bicicletas (**associação**)
Encontre clientes "incomuns" (**descoberta de anomalias**)
Quantos quilos de sorvete devo produzir no próximo mês? (**regressão**)
Quais são os clientes que mais consomem energéticos? (**caracterização**)

Prof. MSc. Braian Varjão

Processo de KDD



Knowledge Discovery from Data ou KDD

Prof. MSc. Braian Varjão

Avaliação

Análise de utilidade e confiabilidade

Prof. MSc. Braian Varjão

Técnicas de Mineração de Dados

Técnicas de mineração de dados

- Classificação;
- Regressão;
- Agrupamento;
- Descrição;
- Associação;
- Predição.

Prof. MSc. Braian Varjão

Técnicas de mineração de dados

- **Classificação;**
- Regressão;
- Agrupamento;
- Descrição de classes;
- Associação;
- Predição.

Prof. MSc. Braian Varjão

Classificação

Relationship

"Dog"

Relationship

"Cat"

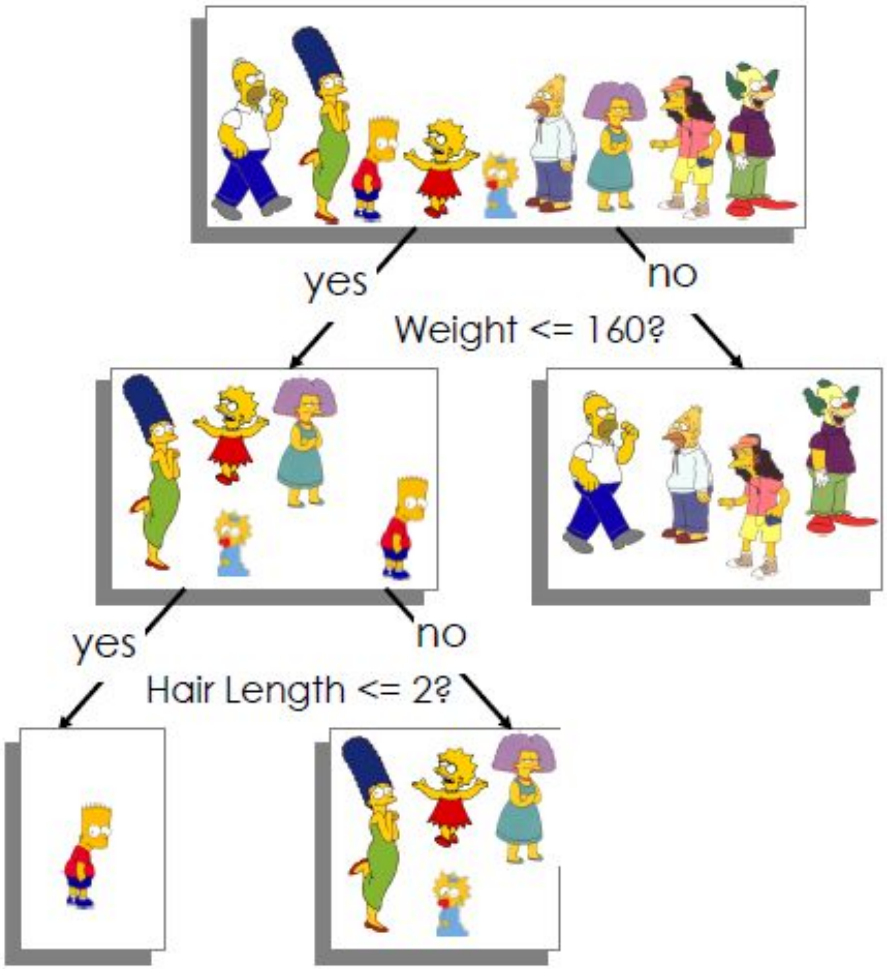
Prof. MSc. Braian Varjão

Classificação

Pessoa	Tamanho do Cabelo	Peso	Idade	Classe
Homer	0	250	26	M
Marge	10	150	34	F
Bart	10	90	10	M
Lisa	6	78	8	F
Maggie	4	20	1	F
Abe	1	170	70	M
Selma	8	160	41	F
Otto	10	180	38	M
Krusty	6	200	45	M
Comic	**8**	**290**	**38**	**?**

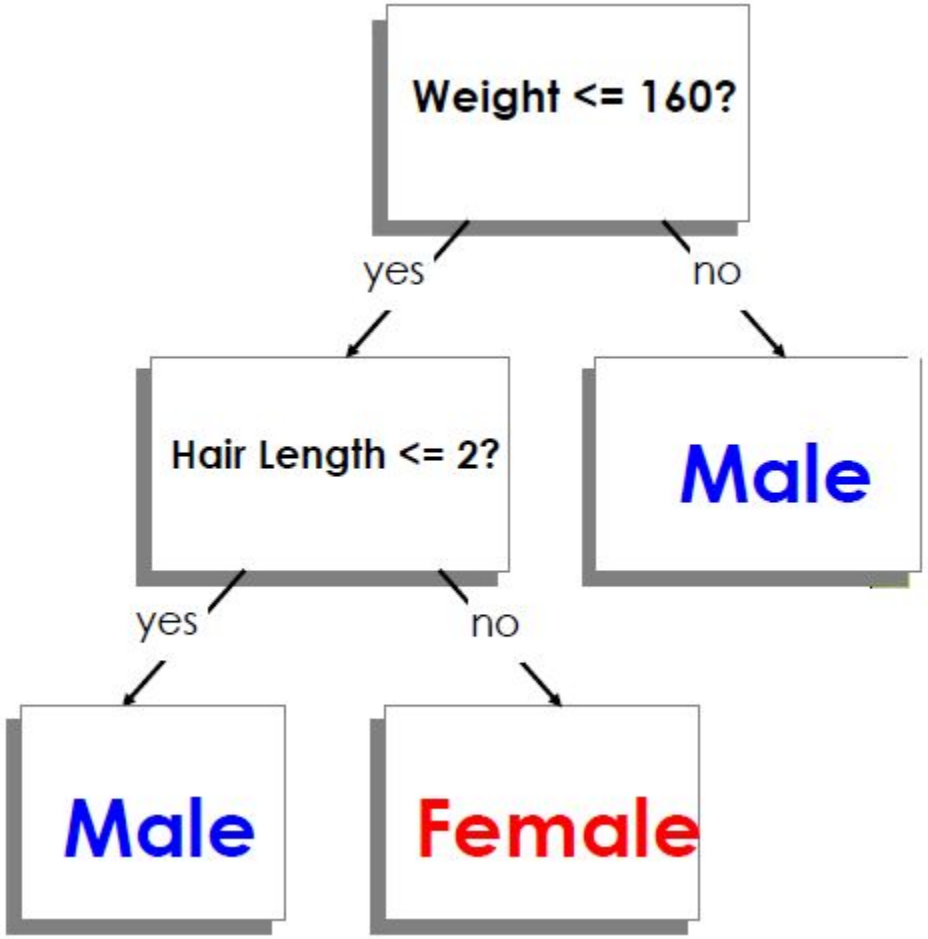
Prof. MSc. Braian Varjão

Classificação



Prof. MSc. Braian Varjão

Classificação

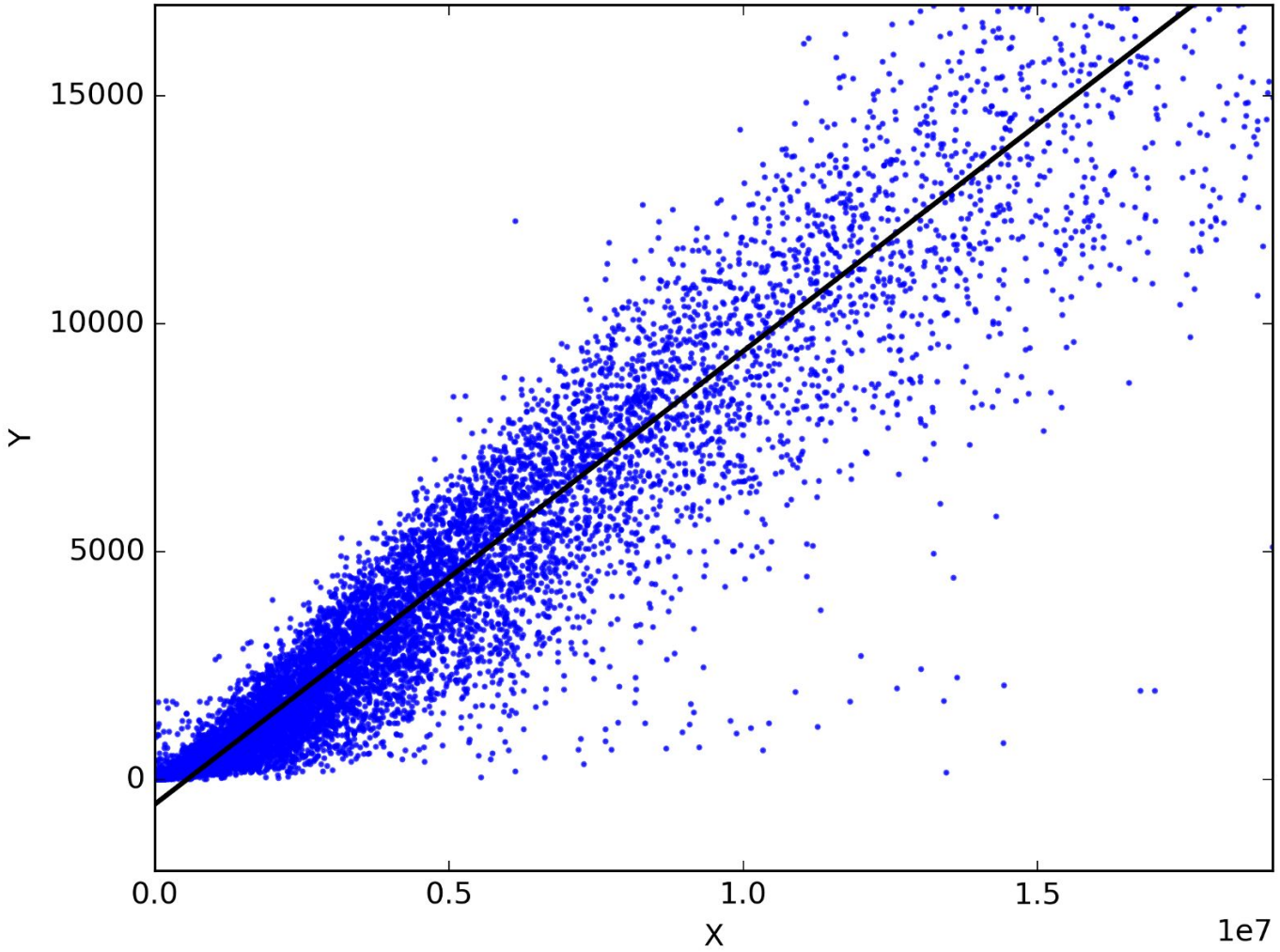


Técnicas de mineração de dados

- Classificação;
- **Regressão;**
- Agrupamento;
- Descrição;
- Associação;
- Predição.

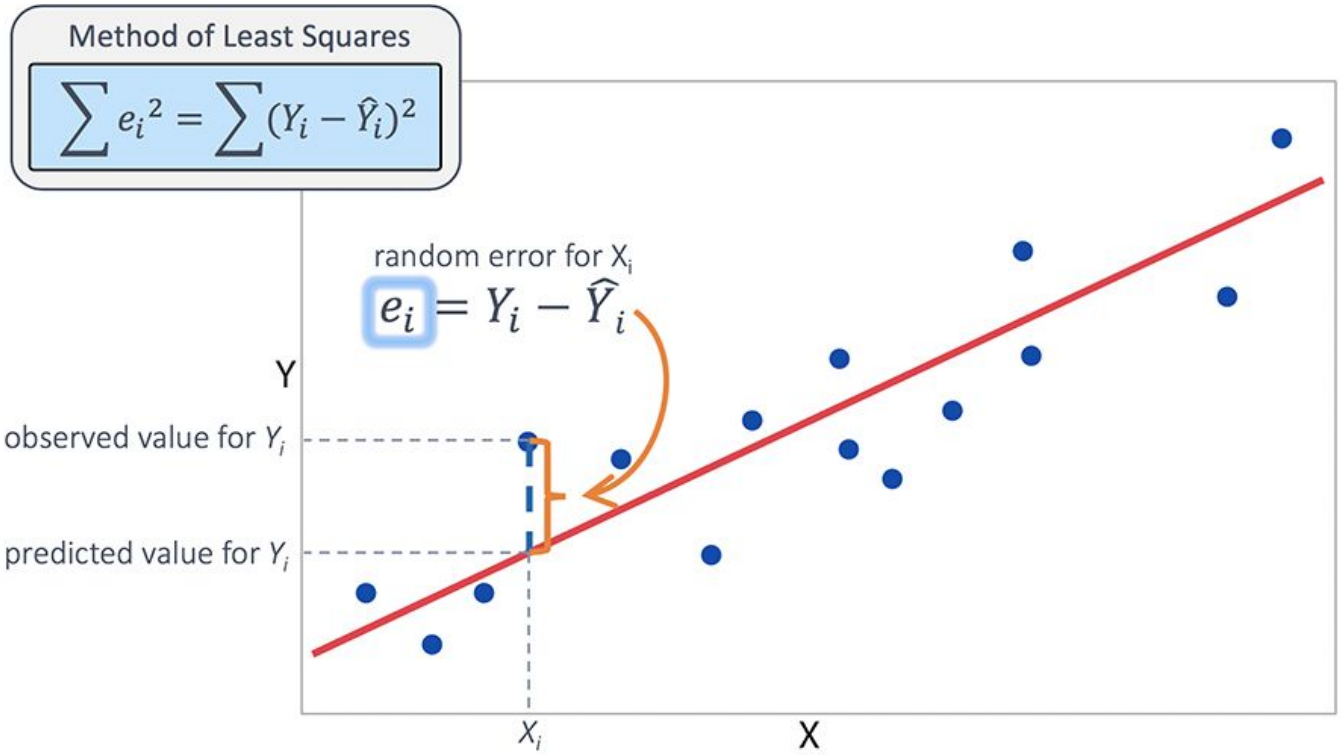
Prof. MSc. Braian Varjão

Regressão



Prof. MSc. Braian Varjão

Regressão



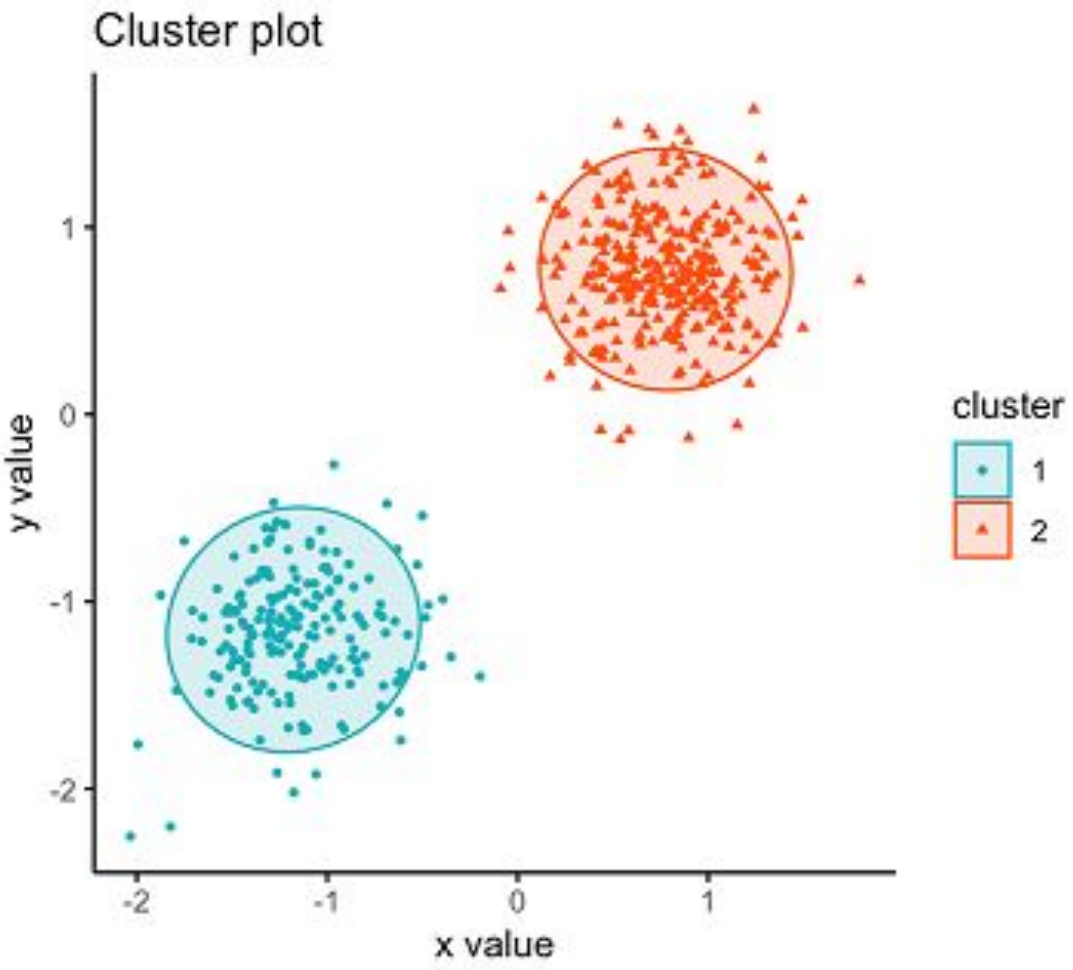
Prof. MSc. Braian Varjão

Técnicas de mineração de dados

- Classificação;
- Regressão;
- **Agrupamento;**
- Descrição;
- Associação;
- Predição.

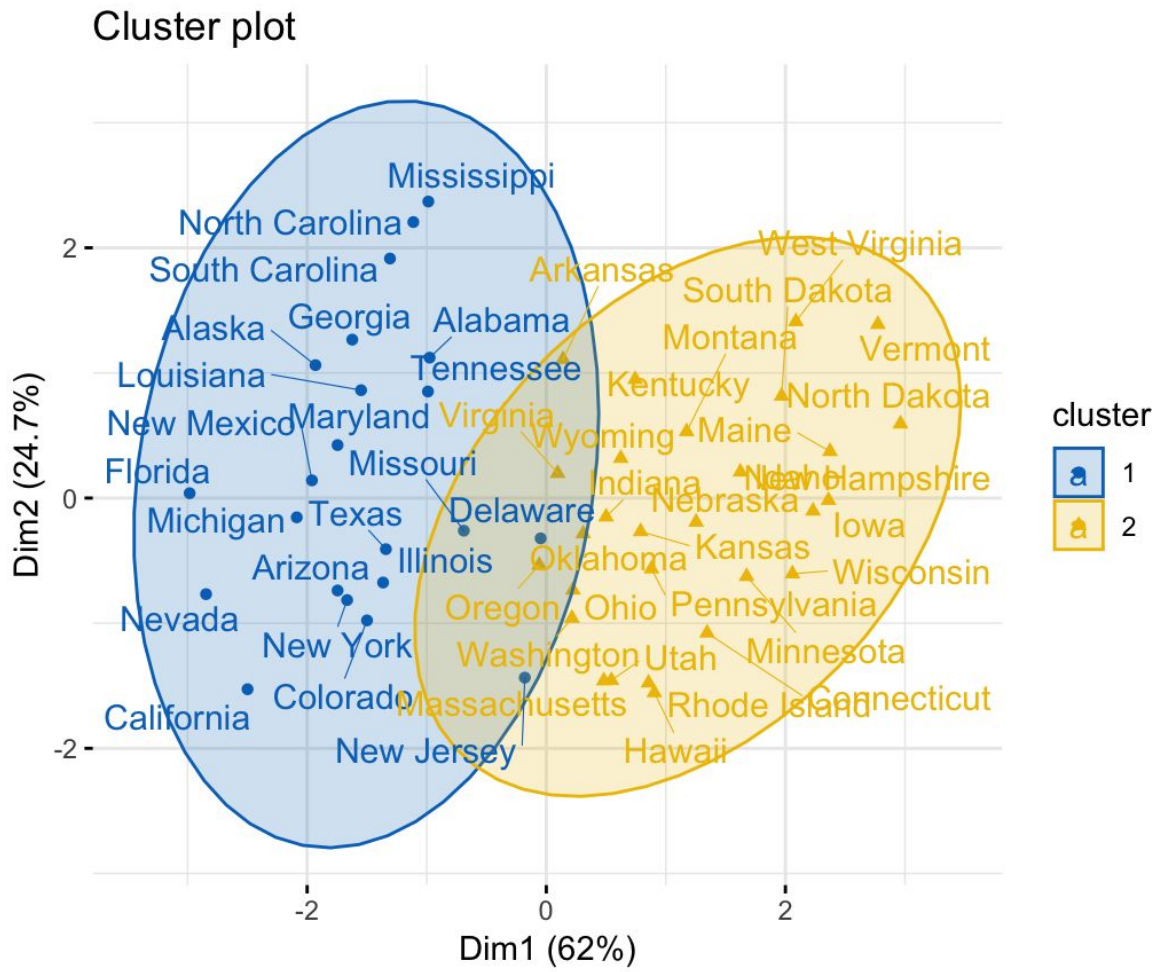
Prof. MSc. Braian Varjão

Agrupamento



Prof. MSc. Braian Varjão

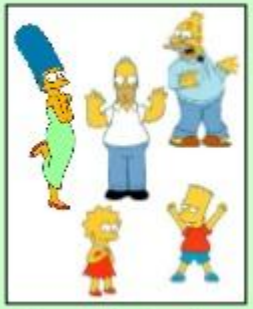
Agrupamento



Agrupamento

What is a natural grouping among these objects?

Clustering is subjective



Simpson's Family

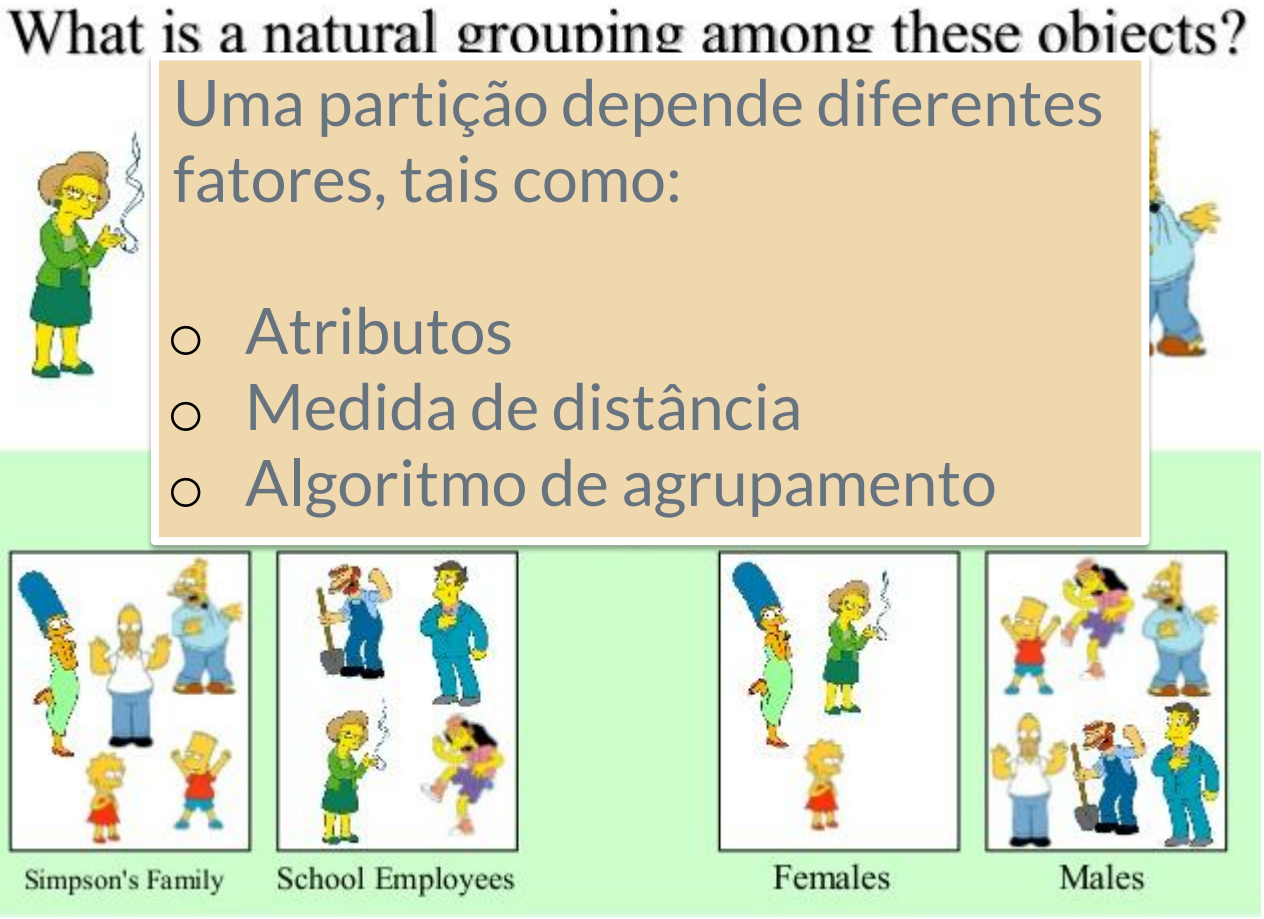
School Employees

Females

Males

Prof. MSc. Braian Varjão

Agrupamento



Uma partição depende diferentes fatores, tais como:

- Atributos
- Medida de distância
- Algoritmo de agrupamento

Prof. MSc. Braian Varjão

Técnicas de mineração de dados

- Classificação;
- Regressão;
- Agrupamento;
- **Descrição;**
- Associação;
- Predição.

Prof. MSc. Braian Varjão

Descrição de classes/conceitos

Descrever classes e conceitos em função dos dados de forma resumida e precisa.

- Caracterização;
- Discriminação.

Prof. MSc. Braian Varjão

Descrição de classes

Caracterização

Sumarização das características de uma determinada classe em função dos atributos.

Prof. MSc. Braian Varjão

Descrição de classes

Caracterização

Sumarização das características de uma determinada classe em função dos atributos.

Clientes que gastam mais de 5 mil reais/ano em compras

Clientes entre 40 e 50 anos, empregados e com alta taxa de crédito

Prof. MSc. Braian Varjão

Descrição de classes

Discriminação

Sumarização as características de uma determinada classe que a distingue de uma outra ou de um conjunto de classes.

Prof. MSc. Braian Varjão

Descrição de classes

Discriminação

Sumarização as características de uma determinada classe que a distingue de uma outra ou de um conjunto de classes.

- Eleitores de Bolsonaro:
  - Palavras frequentes: “mito”, “petralha”.
- Eleitores do Haddad:
  - Palavras frequentes: “Lula” e “bolsominion”.

Prof. MSc. Braian Varjão

Técnicas de mineração de dados

- Classificação;
- Regressão;
- Agrupamento;
- Descrição;
- **Associação;**
- Predição.

Prof. MSc. Braian Varjão

Associação

Descoberta de *regras de associação* que apresentam padrões de coocorrência de valores de atributos em uma base.

Prof. MSc. Braian Varjão

Associação

É confiável afirmar que clientes que compram cerveja também compram amendoim?

Transação 1	Leite, pão
Transação 2	Pão, manteiga
Transação 3	Cerveja, amendoim
Transação 4	Leite, pão, manteiga
Transação 5	Pão

Prof. MSc. Braian Varjão

Associação

Suporte

Dada uma regra A -> B, a sua medida de suporte é a porcentagem de transações da base que contém os elementos A e B, indicando a relevância da mesma.

Prof. MSc. Braian Varjão

Associação

Confiança

Dada uma regra A -> B, a confiança representa a porcentagem de transações em que, ocorrendo A, também ocorre B, indicando a validade da regra.

Prof. MSc. Braian Varjão

Associação

Quais itens costumam ser comprados juntos?

Cliente que compra leite, compra pão.

Suporte: 2/5 = 40%;

Confiança: 2/2 = 100%

Transação 1	Leite, pão
Transação 2	Pão, manteiga
Transação 3	Cerveja, amendoim
Transação 4	Leite, pão, manteiga
Transação 5	Pão

Prof. MSc. Braian Varjão

Associação

Quais itens costumam ser comprados juntos?

Cliente que compra leite, cor

Suporte: 2/5 = 40%;

Confiança: 2/2 = 100%

Atenção:

A ordem dos tratores altera o viaduto.

Transação 1	Leite, pão
Transação 2	Pão, manteiga
Transação 3	Cerveja, amendoim
Transação 4	Leite, pão, manteiga
Transação 5	Pão

Associação

Quais itens costumam ser comprados juntos?

Cliente que compra pão, compra leite.

Suporte: 2/5 = 40%;

Confiança: 2/4 = 50%

Transação 1	Leite, pão
Transação 2	Pão, manteiga
Transação 3	Cerveja, amendoim
Transação 4	Leite, pão, manteiga
Transação 5	Pão

Prof. MSc. Braian Varjão

Exercício 1

Cite 2 padrões frequentes que poderiam ser minerados considerando o banco de dados abaixo:

Transação 1	Pão, leite, queijo, presunto, desodorante, feijão
Transação 2	Achocolatado, pão, leite
Transação 3	Cebola, laranja, salsa, manga
Transação 4	Carne, presunto, ovos, queijo, pão
Transação 5	Chocolate, pipoca, refrigerante, leite
Transação 6	Caneta, bala, fralda, queijo, leite, pão

Exercício 1 - Resposta

Compra (Cliente, **Pão**) => Compra (Cliente, **Leite**)
[suporte: **50%** confiança: **75%**]

Compra (Cliente, **Queijo**) => Compra (Cliente, **Presunto**)
[suporte: **33%** confiança: **67%**]

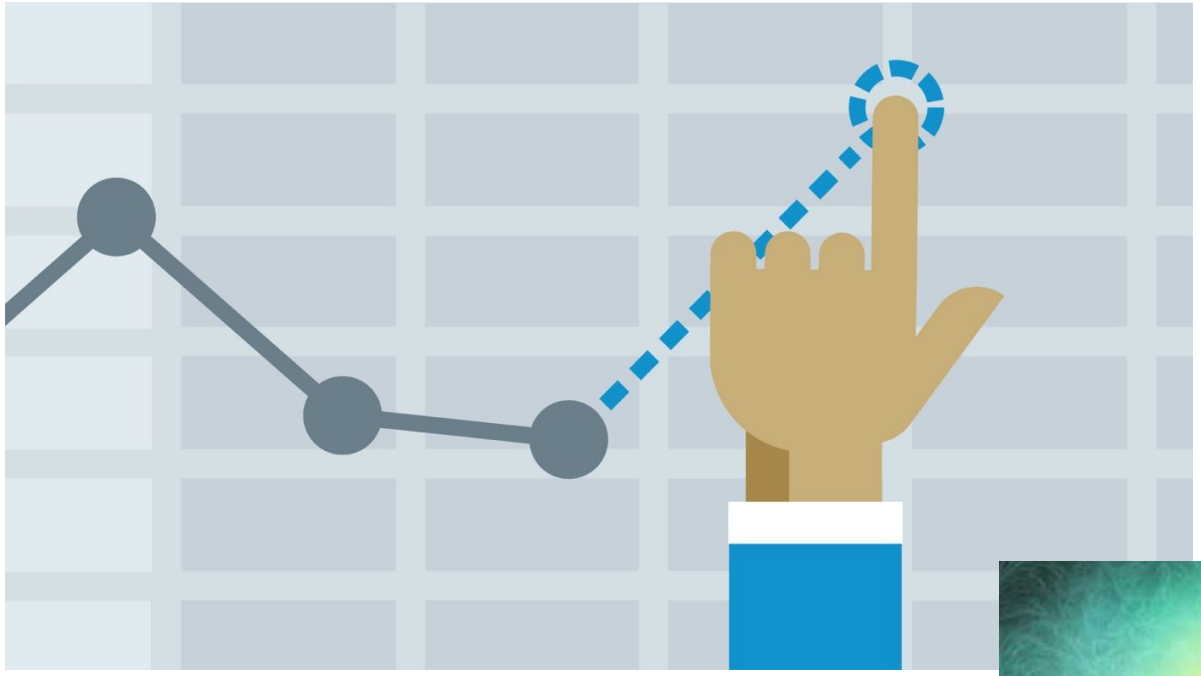
Transação 1	Pão, leite, queijo, presunto, desodorante, feijão
Transação 2	Achocolatado, pão, leite
Transação 3	Cebola, laranja, salsa, manga
Transação 4	Carne, presunto, ovos, queijo, pão
Transação 5	Chocolate, pipoca, refrigerante, leite
Transação 6	Caneta, bala, fralda, queijo, leite, pão

Técnicas de mineração de dados

- Classificação;
- Regressão;
- Agrupamento;
- Descrição;
- Associação;
- **Predição.**

Prof. MSc. Braian Varjão

Predição



Exercício 2

Qual dessas técnicas você usaria se tivesse acesso ao banco de dados da sua rede social predileta?

Com qual objetivo?

Prof. MSc. Braian Varjão

Fatores Críticos para uma Mineração de Dados Eficiente e Eficaz

Fatores críticos para uma mineração eficiente e eficaz

- As características da base de dados influenciam todos os resultados;
- Conhecer os dados;
- Complexidade vs. Erro;
- Validar resultados;
- Investigar erros;
- **Utilidade do conhecimento obtido.**

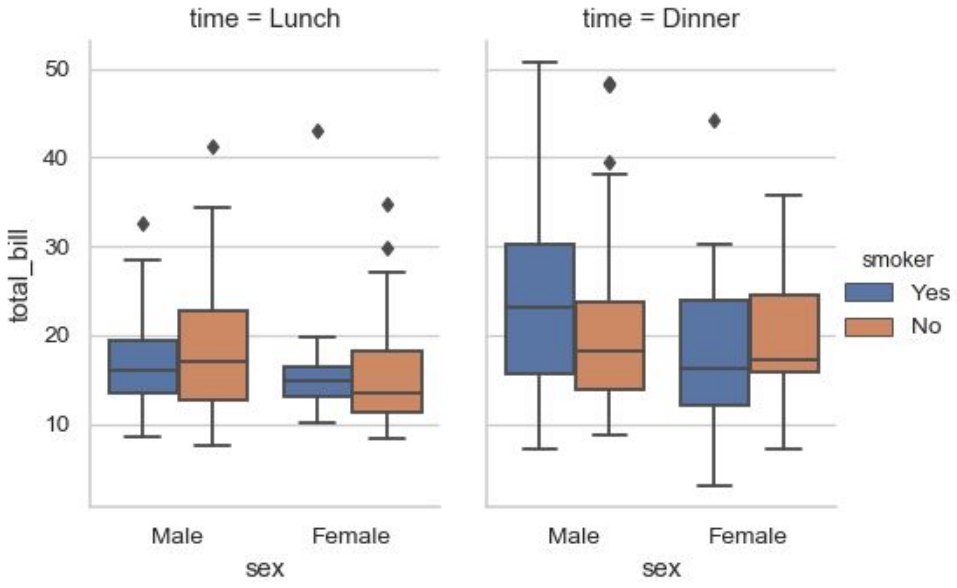
Prof. MSc. Braian Varjão

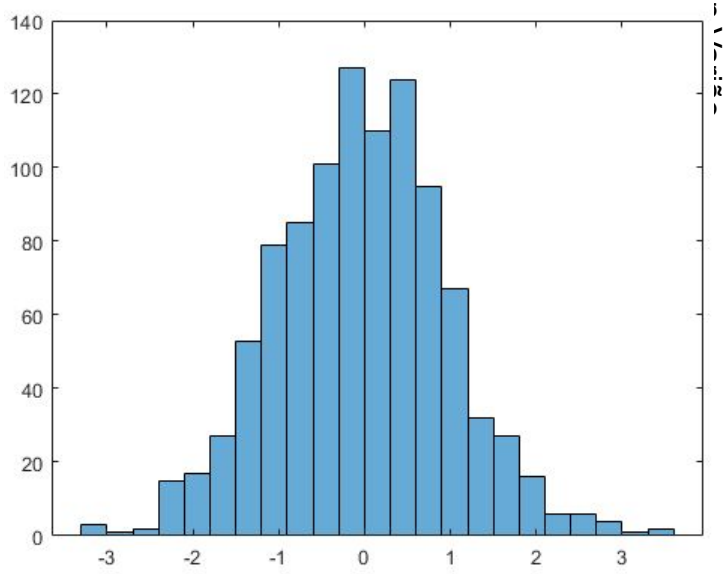
Características da base de dados

- Representatividade da amostra;
- Desbalanceamento;
- Alta dimensionalidade;
- Esparsidade;
- Pré-processamento.

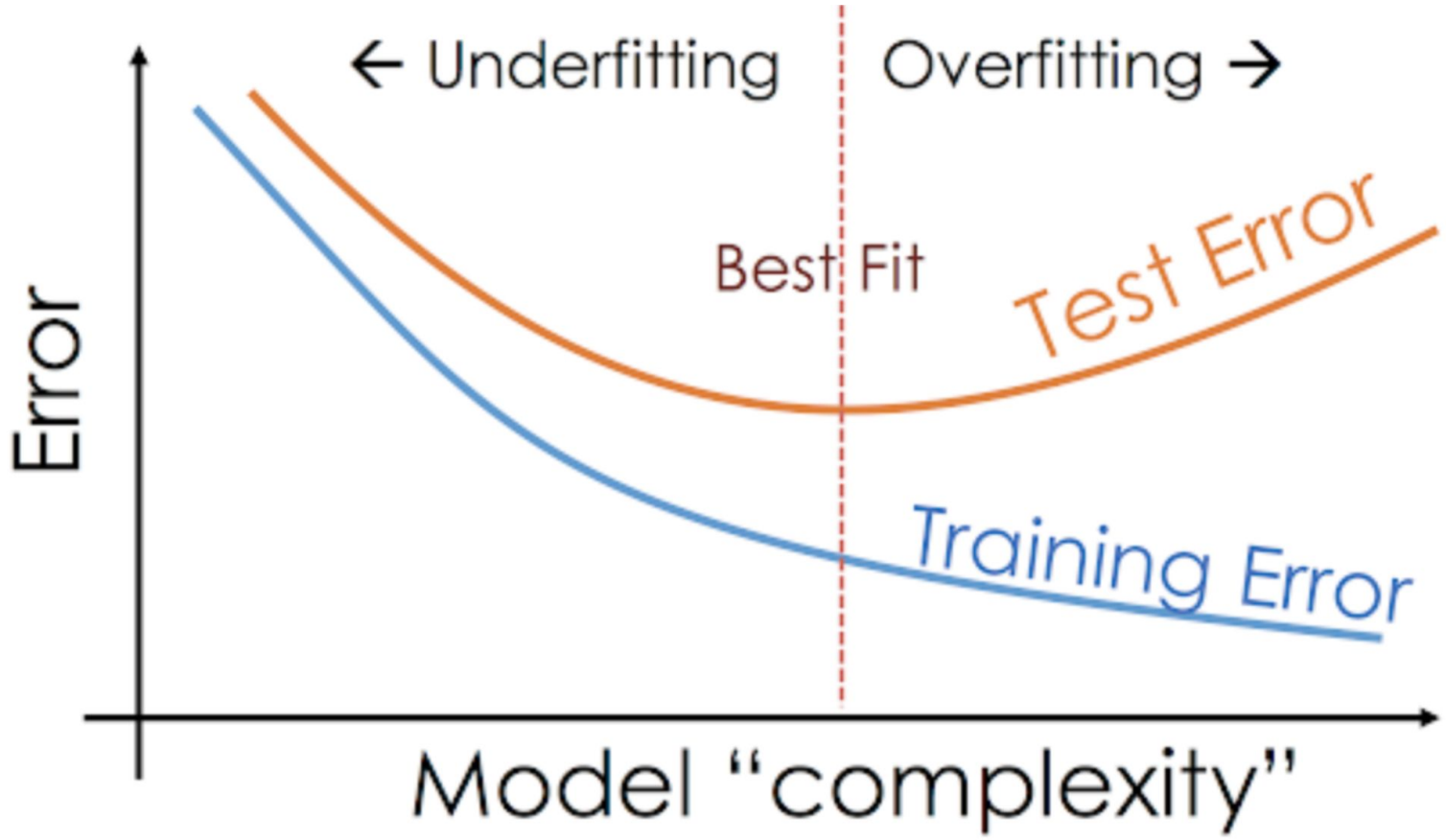
Conhecer os dados

- O que significam os atributos?
- Como eles se relacionam?
- Como eles estão distribuídos?





Complexidade vs. Erro



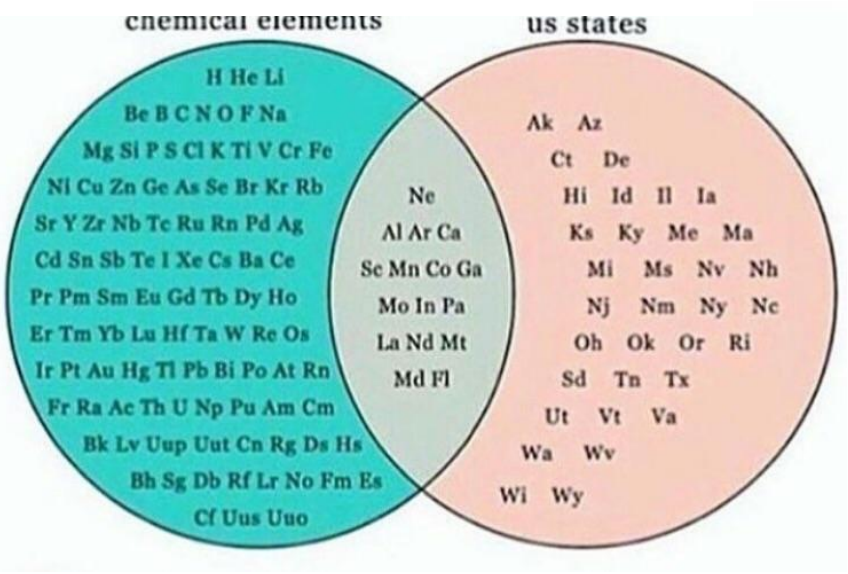
Validar resultados

- Definição de métricas coerentes;
- Capacidade de generalização;
- Comparação de resultados entre diferentes métodos e abordagens.

Investigar erros

Não ignorar erros de predição

Utilidade do conhecimento obtido



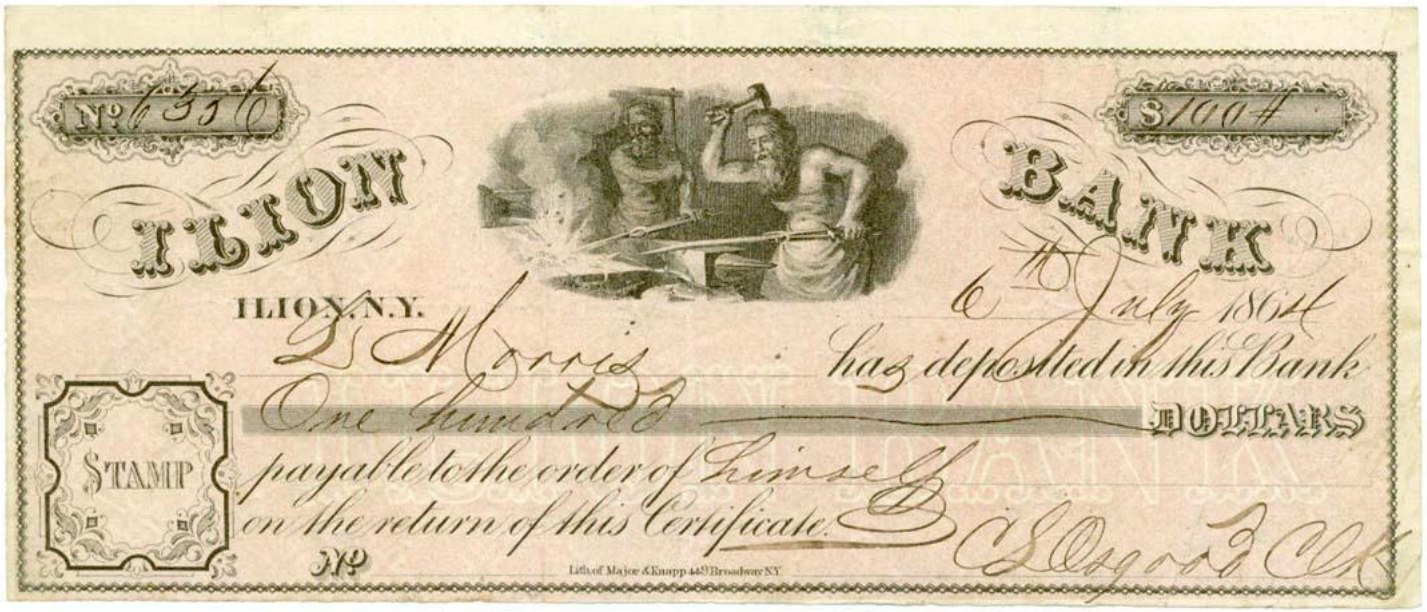
Prof. MSc. Braian Varjão

Aplicações

Filtragem de currículos

Prof. MSc. Braian Varjão

Verificação de assinaturas



☑ TRUE
☐ FALSE

Prof. MSc. Braian Varjão

Qualidade de atendimento

Prof. MSc. Braian Varjão

Análise *Market Basket*

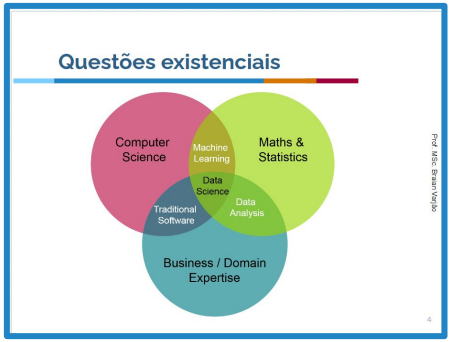
Prof. MSc. Braian Varjão

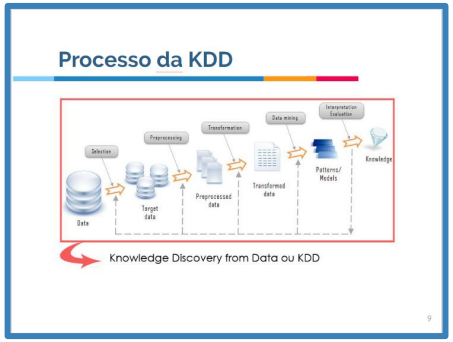
Prevenção de epidemias

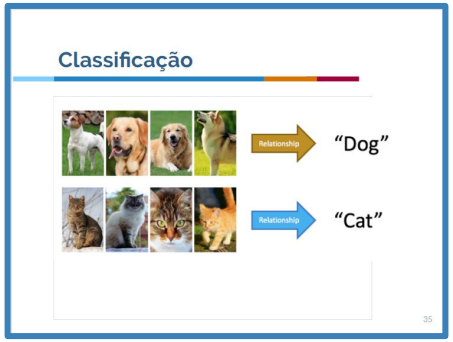
Prof. MSc. Braian Varjão

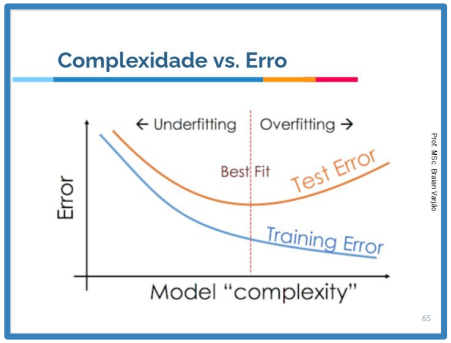
Considerações Finais

Resumo da aula



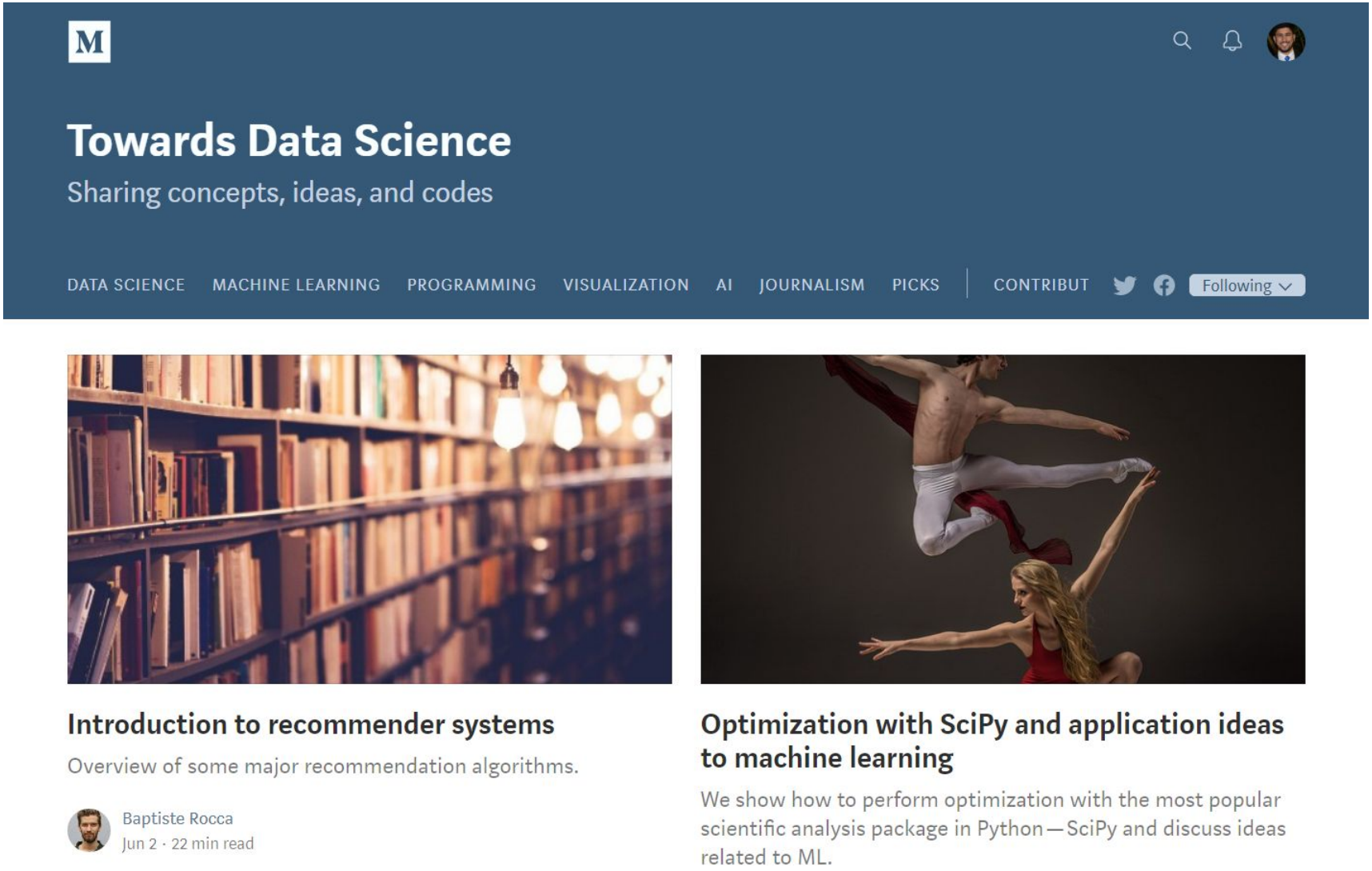






Prof. MSc. Braian Varjão

Fique por dentro!



Prof. MSc. Braian Varjão

Dúvidas?