

# Introdução a Ciências de Dados

## Aula 1 parte 1: Introdução

Francisco A. Rodrigues  
ICMC/USP  
francisco@icmc.usp.br



# Aula 1: Introdução

- O que é Ciência de Dados
- Problemas e Soluções em Ciência de Dados

# O que é Ciência de Dados

“Ciência de dados (em inglês: data science) é uma área interdisciplinar voltada para o estudo e a análise de dados, estruturados ou não, que visa a extração de conhecimento ou insights para possíveis tomadas de decisão, de maneira similar à mineração de dados.”



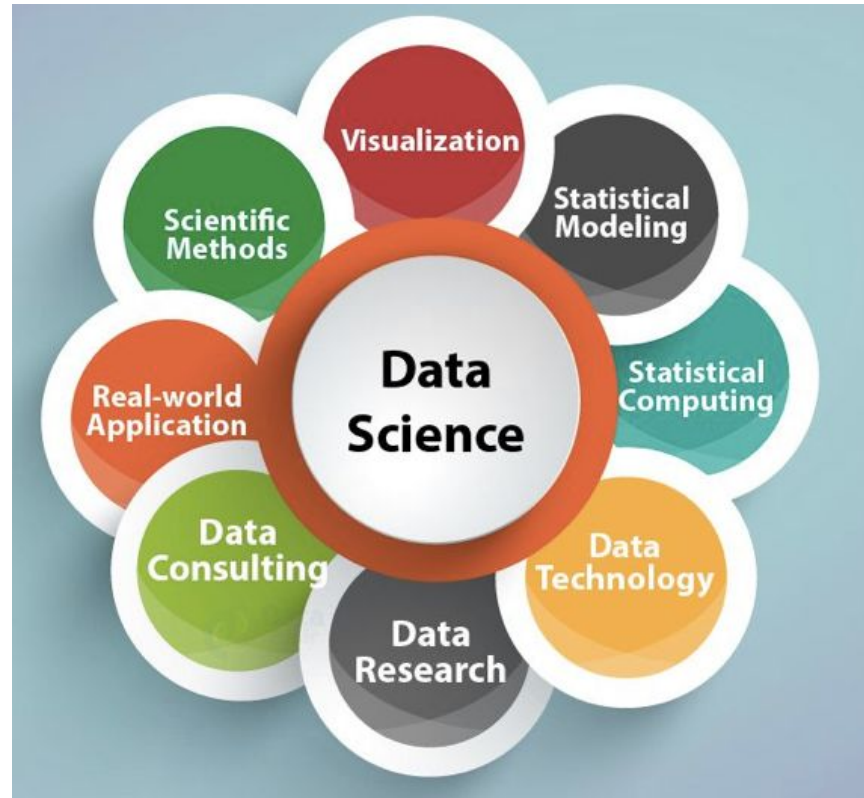
WIKIPÉDIA  
A enciclopédia livre

# O que é Ciência de Dados





# O que é Ciência de Dados



# **Ciência de Dados**

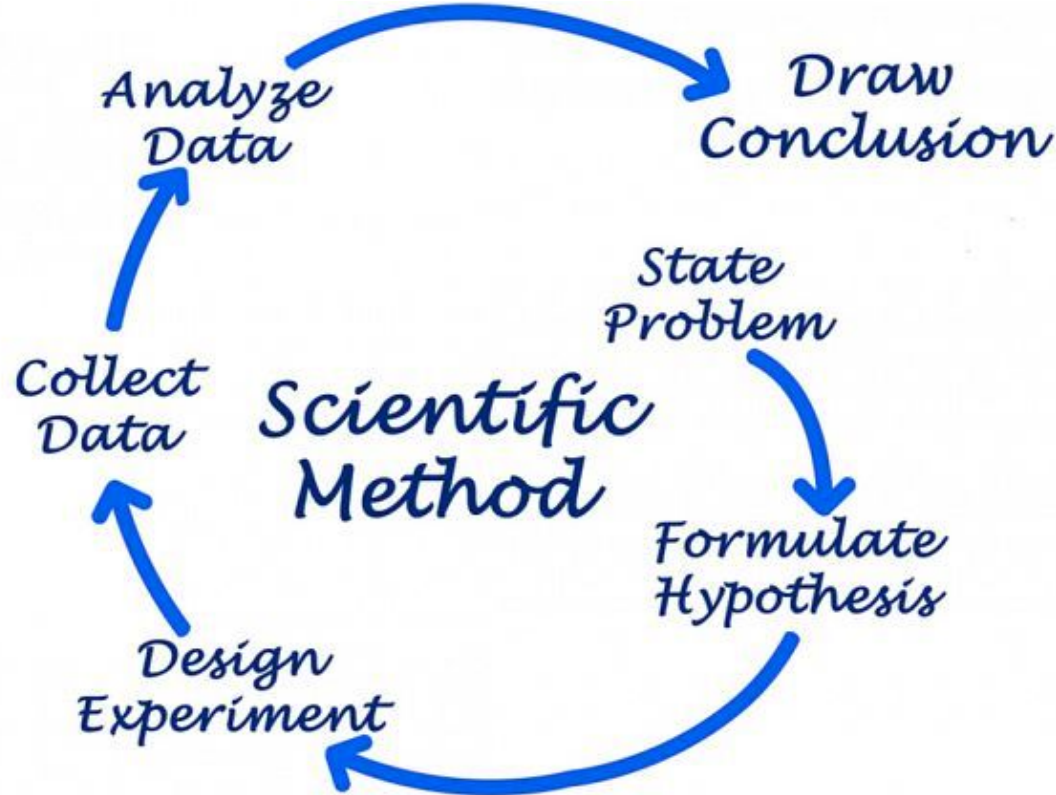
**=**

# **Ciência**

**+**

# **Dados**

# Método científico



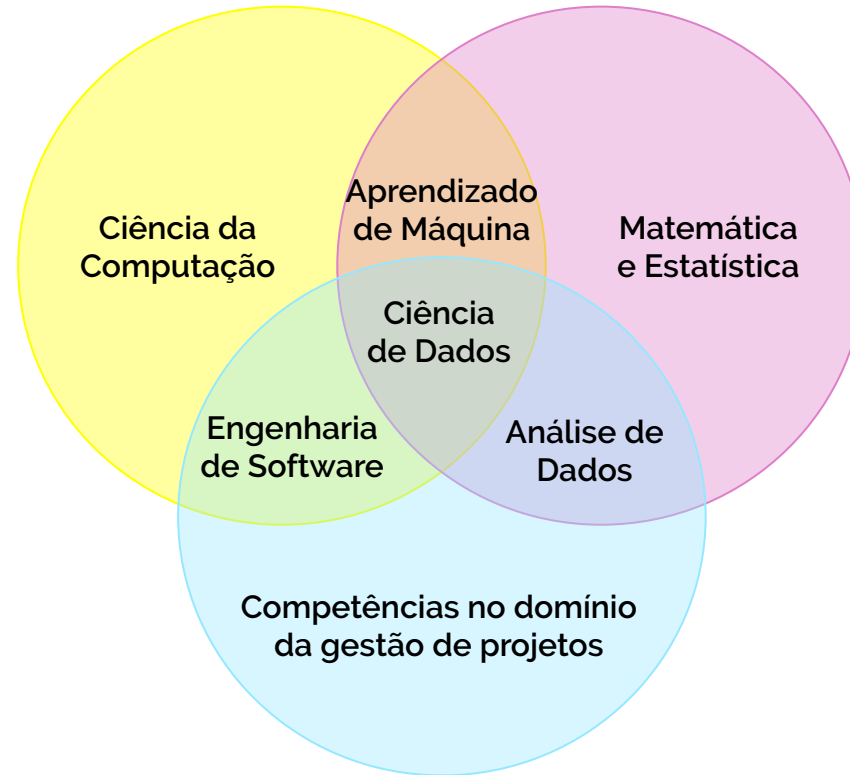
# Método Científico

Every baby knows the  
scientific method!





# O que é Ciência de Dados



# Problemas em Ciência de Dados

## Classificação de documentos:



Sports  
Science  
News

# Problemas em Ciência de Dados

## Agrupar imagens similares:



$C_1$



$C_2$



$C_3$



$C_4$



$C_5$

# Problemas em Ciência de Dados

## Mercado de ações:





# Problemas em Ciência de Dados

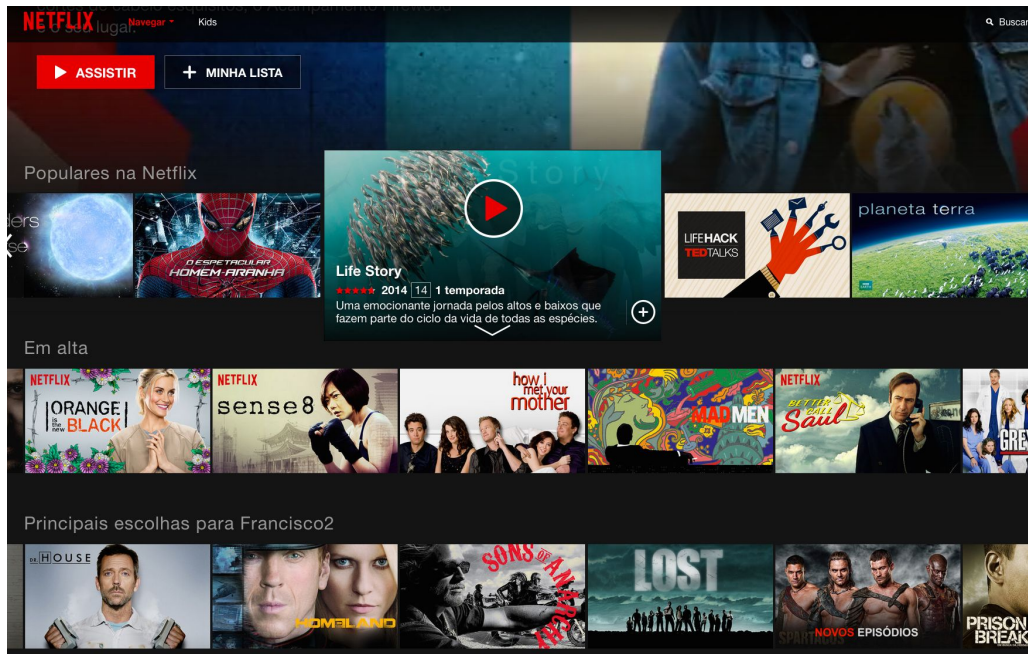
## Recomendação

The screenshot shows the Amazon website interface. At the top, there's a navigation bar with the Amazon logo, a search bar, and links for 'Shop by Department', 'Francisco's Amazon.com', 'Today's Deals', 'Gift Cards', 'Sell', and 'Help'. Below this, a personalized message for 'Francisco's Amazon' says 'You could be seeing useful stuff here! Sign in to get your order status, balances and rewards.' with a 'Sign In' button.

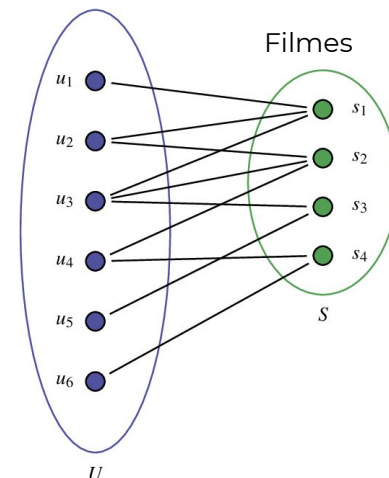
The main content area is divided into two sections: 'Kindle eBooks' and 'Books'. Each section displays a grid of recommended books. The 'Kindle eBooks' section shows books like 'Derivatives Analytics with Python', 'The Count of Monte Cristo', 'Grimm's Fairy Tales', 'Quantum Mechanics', 'Entropy', 'A Student's Guide to...', 'Quantum Field Theory', and 'Statistical...'. The 'Books' section shows similar titles, including 'Quantum Mechanics', 'Entropy', 'Quantum Field Theory', 'Statistical...', 'Quantum Mechanics and Path Integrals', 'Renormalization...', 'Complex Networks', and 'An Introduction to...'. Each book entry includes a cover image, title, author, star rating, and price.

# Problemas em Ciência de Dados

## Recomendação



Usuários



# Problemas em Ciência de Dados

## Diagnóstico médico:

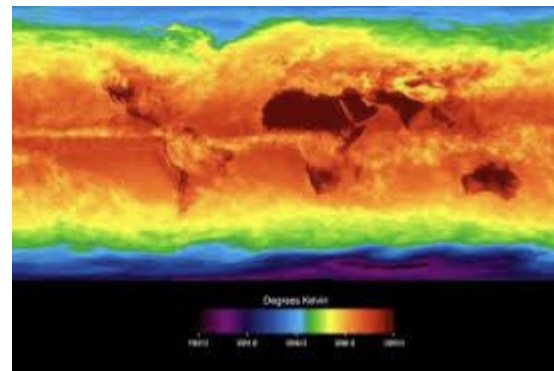


pixabay.com



# Problemas em Ciência de Dados

## Predição de secas:



[pixabay.com](https://pixabay.com)



# Problemas em Ciência de Dados

## Agricultura:



[pixabay.com](https://pixabay.com)

# Problemas em Ciência de Dados

## Operações Bancárias:



[pixabay.com](https://pixabay.com)

# Problemas em Ciência de Dados

## Detecção de Fraudes:



[pixabay.com](https://pixabay.com)

# Problemas em Ciência de Dados

## E-commerce

- Identificação de clientes
- Recomendação de produtos
- Análise de avaliações de produtos

## Bancos

- Detecção de fraudes
- Modelagem de risco de crédito
- Mercado futuro

## Medicina

- Análise de dados médicos
- Descoberta de novas drogas
- Bioinformática

## Transporte

- Carros autônomos
- Sistema de monitoramento
- Segurança

## Finanças

- Segmentação de usuários
- Decisões estratégicas
- Análise de risco

## Agricultura

- Uso de pesticidas
- Previsão das safras
- Planejamento de lavouras



# Tarefas

- Classificação
- Regressão
- Agrupamento
- Regras de associação
- Visualização



# Aprendizado supervisionado

**Modelos preditivos:** função que, dado um conjunto de exemplos rotulados, constrói um estimador.

$$y = f(X, \theta) + \epsilon$$

## Classificação

- Rótulos nominais (conjunto discreto e não ordenado de valores)
  - Ex. {doente, saudável}, {bom pagador, mau pagador}, {iris setosa, iris versicolor, iris virginica}
- Estimador é chamado **classificador**.

## Regressão

- Rótulos contínuos (conjunto infinito ordenado de valores)
  - Ex. peso, temperatura, vazão de água.
- Estimador é chamado **regressor**.

Estimadores podem ser vistos como funções.

# Classificação

**Definição formal:** Dado um conjunto de observações:

$$D = \{\mathbf{X}, \mathbf{y}, i = 1, \dots, N\}$$

- **f** representa uma função desconhecida (função objetivo).
- Essa função mapeia as entradas nas saídas correspondentes.
- O algoritmo preditivo aprende a aproximação, que permite estimar valores de **f** para novos valores de **X**.

$$y_i = f(X_i, \theta) + \epsilon_i$$

## Classificação

$$y_i \in \{C_1, C_2, \dots, C_n\}$$

# Classificação

## Classificação de documentos:



Sports  
Science  
News



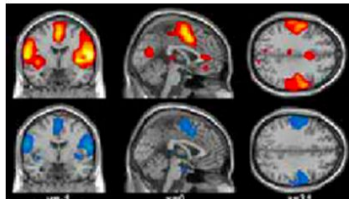
# Classificação

## Identificação de Fake News

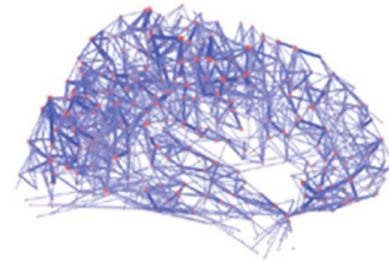


# Classificação

## Exemplo



Data Mining



# Classificação

## Produtos Bancários:



[pixabay.com](https://pixabay.com)

# Regressão

**Definição formal:** Dado um conjunto de observações:

$$D = \{\mathbf{X}, \mathbf{y}, i = 1, \dots, N\}$$

- **f** representa uma função desconhecida (função objetivo).
- Essa função mapeia as entradas nas saídas correspondentes.
- O algoritmo preditivo aprende a aproximação, que permite estimar valores de **f** para novos valores de **X**.

$$y_i = f(X_i, \theta) + \epsilon_i$$

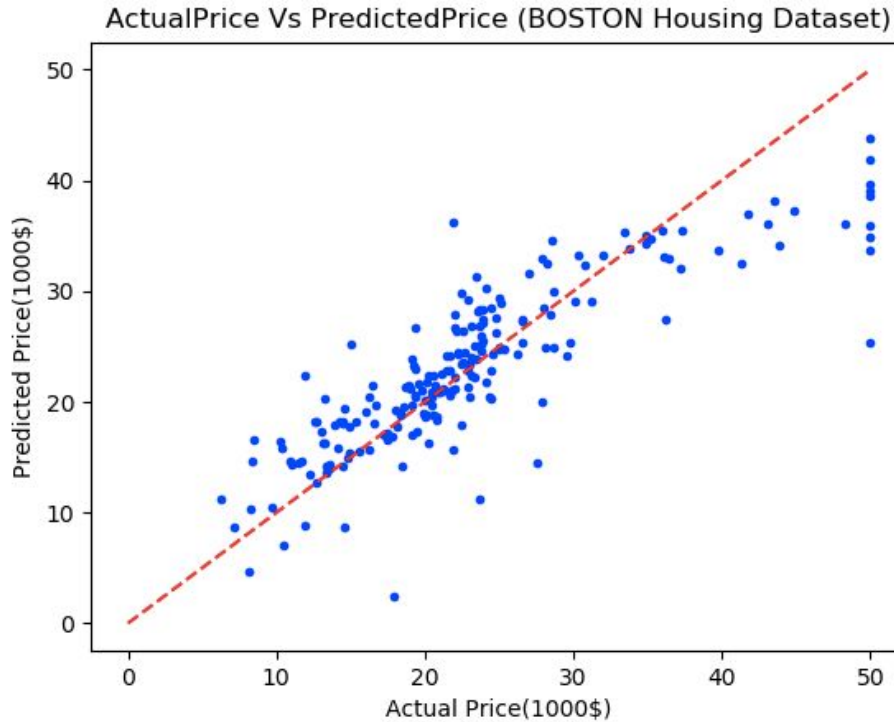
## Regressão

$$y_i \in \mathbb{R}$$



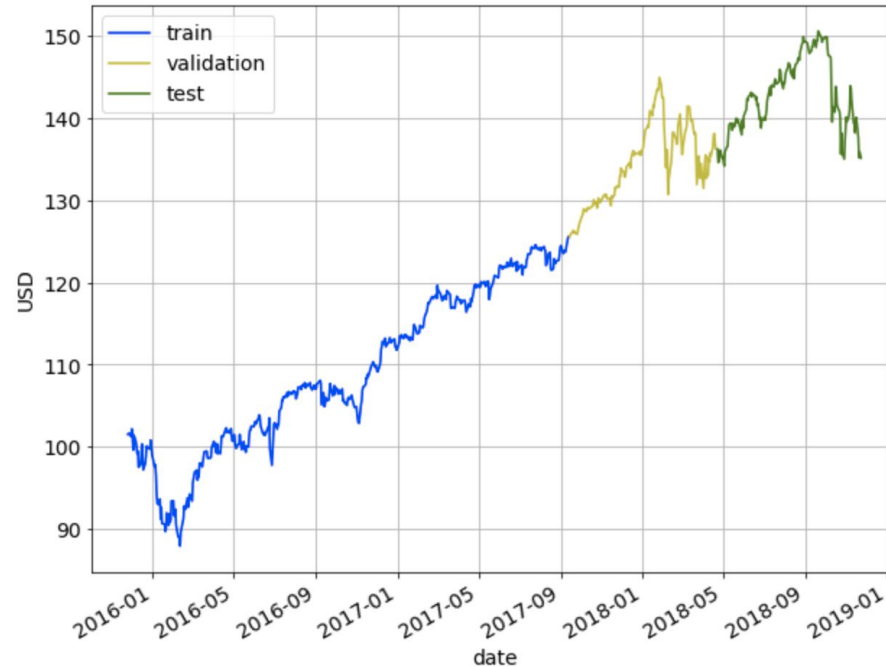
# Regressão

## Exemplo:



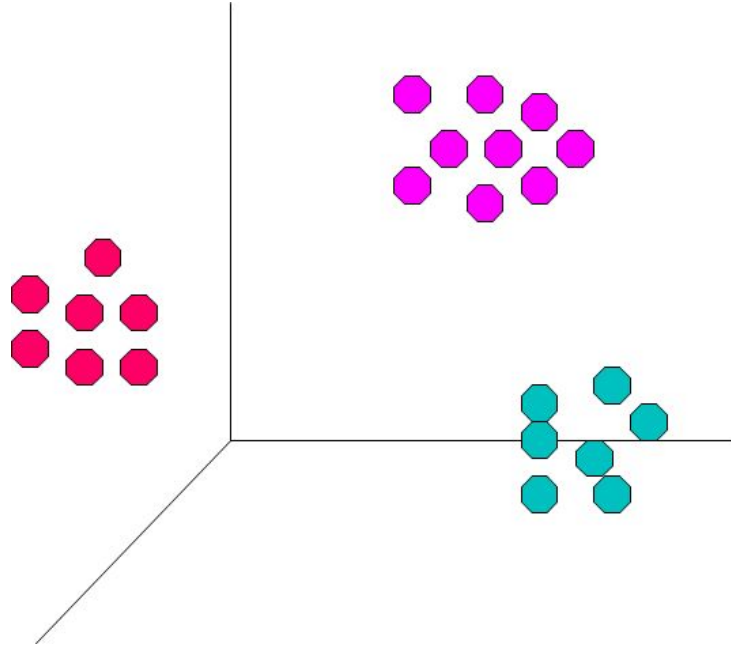
# Regressão

## Exemplo: Previsão de séries temporais



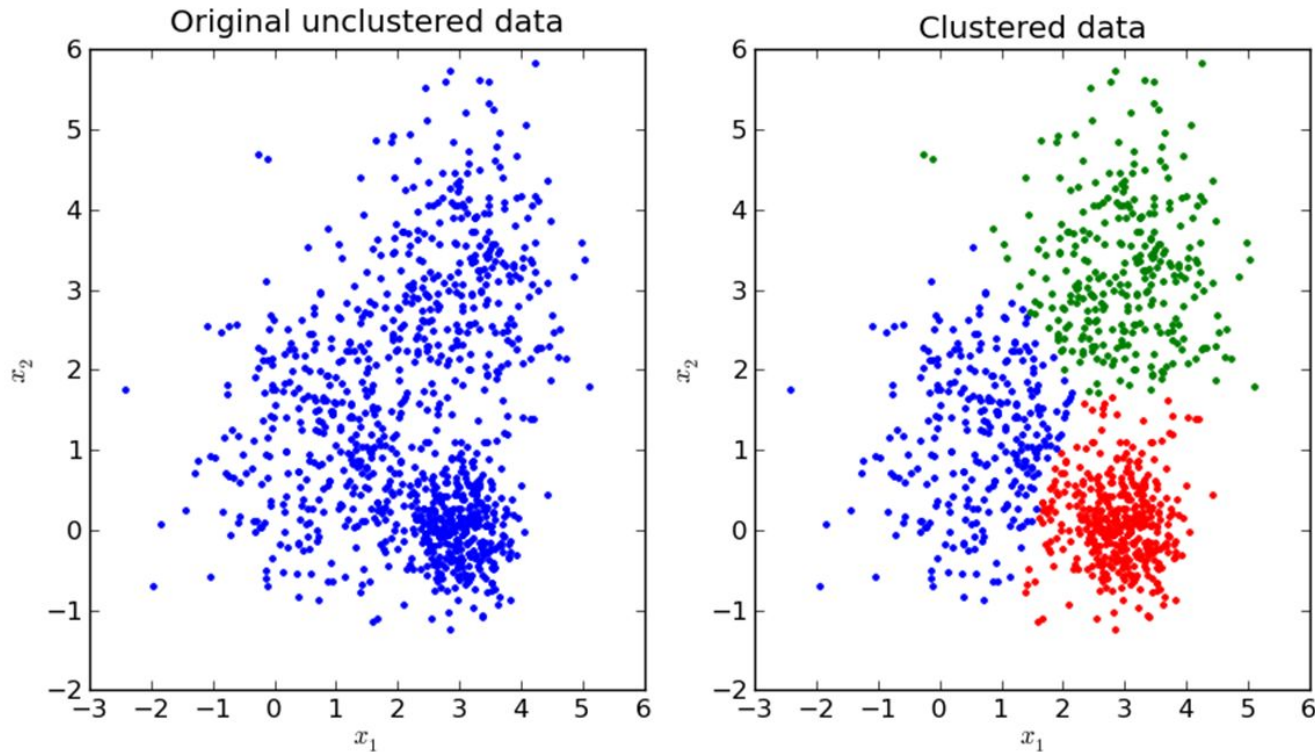
# Agrupamento

**Objetivo:** Agrupar as observações de forma que a similaridade entre objetos no mesmo grupo seja máxima e a entre grupos seja mínima.



# Agrupamento

Exemplo:





# Agrupamento

## Agrupar imagens similares:



$C_1$



$C_2$



$C_3$



$C_4$



$C_5$

# Regras de associação

Dado um conjunto de transações, onde cada uma contém um número de itens de uma dada coleção,  
produzir regras de dependência para predizer um item baseado na ocorrência de outros itens.

ID	Itens
1	Pão, Café, Leite
2	Cerveja, Pão
3	Cerveja, Café, Fralda, Leite
4	Cerveja, Pão, Fralda, Leite
5	Fralda, Leite, Café

**Regras Descobertas:**

**{Leite} -> {Café}**  
**{Fralda, Leite} -> {Cerveja}**

# Regras de Associação

## Recomendação

The screenshot shows an Amazon.com page with a user profile for "Francisco's Amazon". The page displays a list of recommended books, categorized into "Kindle eBooks" and "Books".

**Kindle eBooks**

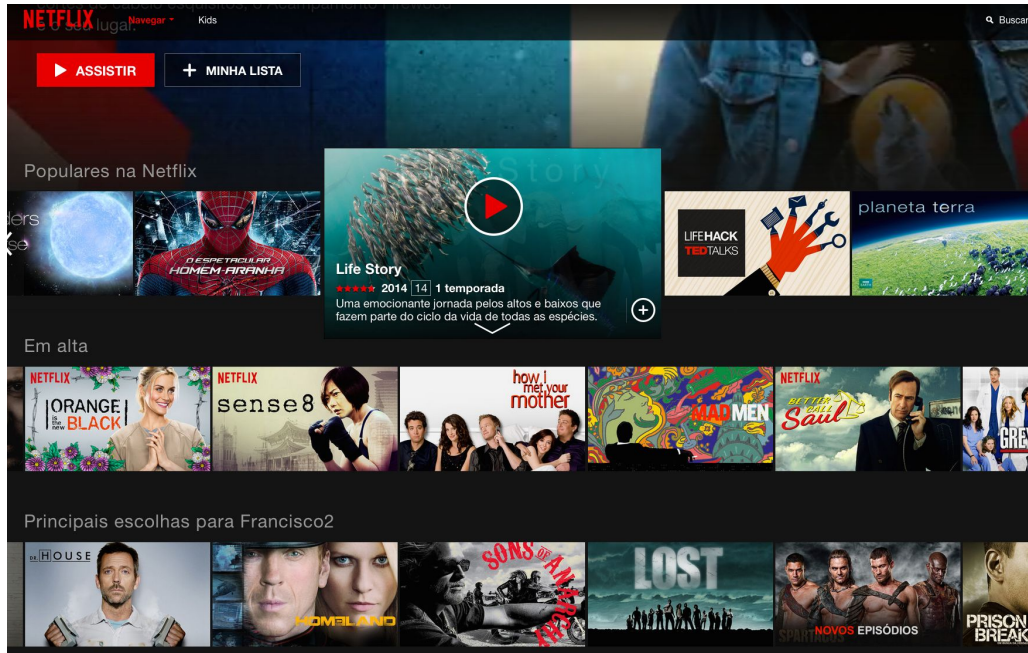
Book Title	Author	Price	Rating
Derivatives Analytics with Python	Yves Hilpisch	\$52.79	★★★★★ (116)
The Count of Monte Cristo	Alexandre Dumas	\$0.99	★★★★★ (193)
Grimm's Fairy Tales	Jacob Grimm	\$0.99	★★★★★ (67)
Quantum Mechanics: A Student's Guide to Entropy	Leonard Susskind	\$10.49	★★★★★ (67)
Quantum Field Theory	Tom Lancaster	\$31.84	★★★★★ (43)
Statistical ...	Werner Krauth	\$49.42	★★★★★ (5)

**Books**

Book Title	Author	Price	Rating
Quantum Mechanics: A Student's Guide to Entropy	Leonard Susskind	\$10.49	★★★★★ (67)
The Count of Monte Cristo	Alexandre Dumas	\$0.99	★★★★★ (193)
Grimm's Fairy Tales	Jacob Grimm	\$0.99	★★★★★ (67)
Quantum Field Theory	Tom Lancaster	\$31.84	★★★★★ (43)
Statistical ...	Werner Krauth	\$49.42	★★★★★ (5)
Quantum Mechanics and Path Integrals	Richard P. Feynman	\$12.71	★★★★★ (31)
Renormalization ...	W. D. McComb	\$51.99	★★★★★ (5)
Lectures on Complex ...	P. S. N. Dorogovtsev	\$35.43	★★★★★ (2)
An Introduction to ...	Terrell L. Hill	\$9.95	★★★★★ (29)

# Regras de Associação

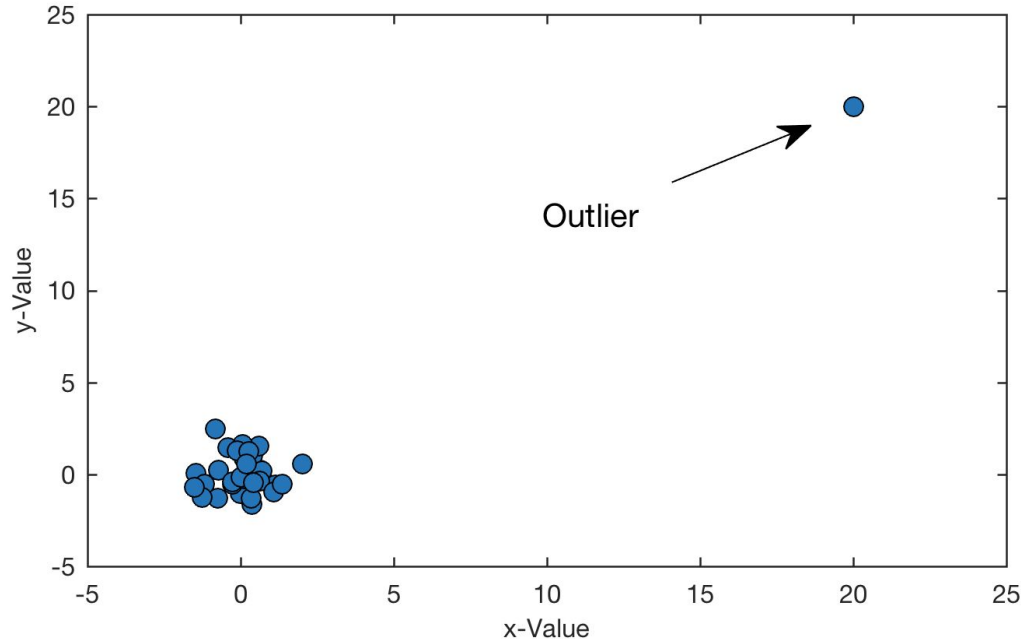
## Recomendação





# Detecção de outliers

Os outliers são dados que se diferenciam drasticamente de todos os outros, são pontos fora da curva.



# Detecção de Outliers

## Detecção de Fraudes:



[pixabay.com](https://pixabay.com)

# Outras tarefas

- Visualização
- Classificação semi-supervisionada
- Aprendizado por reforço

# Sumário

- **O que é Ciência de Dados**
- **Problemas e Soluções em Ciência de Dados**



# Leitura Complementar

- Tan, Steinbach, Karpatne, Kumar, **Introduction to Data Mining**, Pearson, 2013 (capítulo 1)