

aldeia

# Sejam bem-vindos!



Utilize a nossa redes de wi-fi:

#ALDEIA

utilizando a senha

**REALIZADORES**

# A Aldeia é muito mais que espaço

---

Somos um movimento de desenvolvimento de realizadores.

Temos tudo que realizadores precisam para fazer uma ideia dar certo.

<http://aldeia.cc>

Cursos  
Confrarias  
Coworking  
Offices  
Networking  
Eventos  
Acelerações



# Não passe perrengue

Tem água e café à vontade, e um doce e um salgado para você pegar na hora que quiser.

Temos banheiros nos dois andares da **Cândido**:

- Primeiro andar: atrás da recepção
- Segundo andar: ao lado da escada

E atrás da recepção na unidade **Estação**.

**Se algo não estiver certo, fale com a nossa equipe**

# Faça parte da nossa Tribo

---

Receba os **materiais do curso** e seu **certificado** de participação por meio da nossa comunidade virtual.

Acesse <https://aldeia.cc/chamado> e faça sua solicitação para fazer parte da plataforma, utilizando o e-mail da compra do curso para se identificar.



Tire uma foto deste QR code e vá direto para a página da Tribo

# Curso de Data Science

Charles Adriano dos Santos  
Rafael Roberto Dias



# Pauta

**1 – Apresentação Alunos**

**2 – Professores**

**3 – Agenda**



# Apresentação Alunos

1 – Apresentação Alunos

2 – Professores

3 – Agenda



# Apresentação Alunos

Galera, queremos conhecer vocês!!



Nome



Área de atuação



O que é Data Science pra vocês?



Expectativa com o curso

# Professores

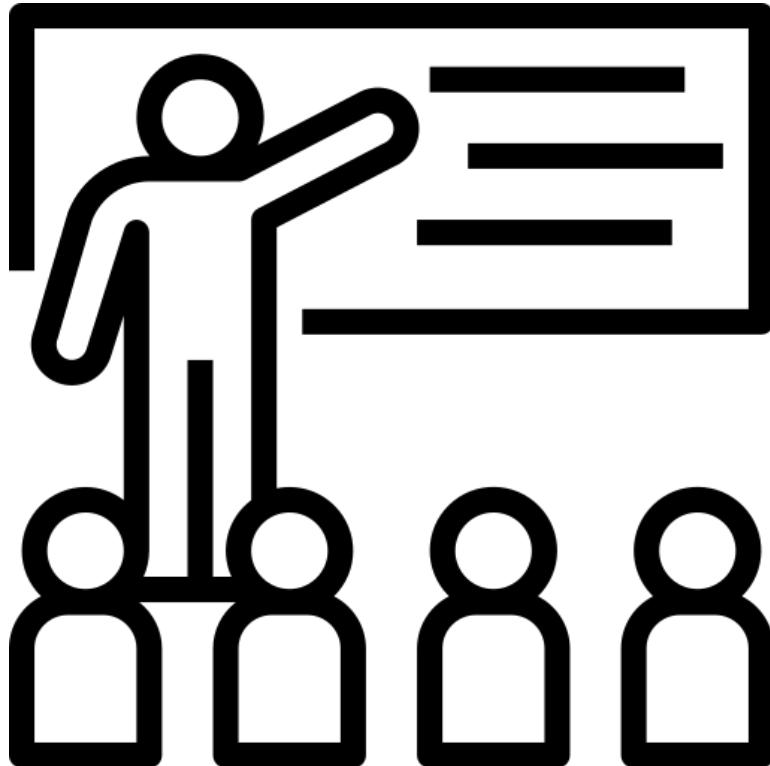
1 – Apresentação Alunos

2 – Professores

3 – Agenda



# Professores



## **Charles Adriano dos Santos**

Bacharel em Análise de Sistemas - PUCPR

Especialista em Engenharia de Software - PUCPR

Especializando em Data Science & Big Data - UFPR

## **Rafael Roberto Dias**

Bacharel em Estatística - UFPR

Especialista em Data Science & Big Data - UFPR

# Agenda

1 – Apresentação Alunos

2 – Professores

3 – Agenda

# Manhã

---

## Horário Assunto

- 09:30 Apresentação e Anseios dos Futuros(as) Cientistas de Dados
- 10:00 Apresentação Equipe
- 10:15 Agenda do Curso
- 10:30 Estrutura do Curso e Objetivos
- 10:45 O que é Data Science and Analytics
- 11:00 Profissão e Carreira e Mercado Atual
- 11:15 Mercado Atual e Projeção
- 11:30 O Trabalho do Cientista de Dados
- 11:45 Matriz de Habilidades do Cientista de Dados

Tarde

---

## Horário Assunto

- 13:30 Desafio Curso → O Desafio da AgroXP Brazil
- 14:00 Desafio Pessoal → Data Set x Machine Learning
- 14:30 Conceitos Estatísticos Aplicados a Ciência de Dados
- 16:00 Conceitos Computacionais Aplicados a Ciência de Dados
- 17:30 Estudo de Caso – Madeira Madeira
- 18:00 Material do Curso e VM do Cientista de Dados



# Data Science



# O que são dados?

The Economist

Topics ▾

Current edition

More ▾

Subscribe

## The world's most valuable resource is no longer oil, but data

*The data economy demands a new approach to antitrust rules*



[Print edition | Leaders >](#)

May 6th 2017



A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A

FORTUNE

LEADERSHIP • ON LEADING

## Intel CEO Says Data is the New Oil



By SUSIE GHARIB June 7, 2018

Brian Krzanich believes big data will dramatically change the world.



# Qual a quantidade de dados são gerados por dia?



# Qual a quantidade de dados são gerados por dia?

## QUANDO BILHÕES QUEREM CONTAR TUDO

Redes sociais concentram momentos da vida da maioria de seus mais de 3 bilhões de usuários ativos. Confira o potencial para negócios e para pesquisas que as redes sociais concentram em seus arquivos (por dia).

**Facebook  
Messenger +  
WhatsApp**  
60 bilhões de mensagens

**Instagram**  
95 milhões de fotos

**YouTube**  
1 bilhão de horas dedicadas aos vídeos da plataforma

**Facebook**  
500 mil novos usuários

**Twitter**  
500 mil tuítes

**Snapchat**  
8 bilhões de visualizações de vídeos

FONTE: BRANDWATCH

# O que é Data Science?

Ramo da Ciência especializada em:

- Coleta
- Armazenamento
- Visualização
- Transformação
- Análise
- Modelagem de Dados

Com foco principal na obtenção de  
subsídios para tomada de **decisões!**



# Profissão, Carreira, Mercado Atual e Projeção

- ✓ A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais

# Profissão, Carreira, Mercado Atual e Projeção

- ✓ A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais
- ✓ Apenas visualizar os dados não atende mais às necessidades das empresas e instituições, a palavra de ordem é: RECOMENDAÇÃO

# Profissão, Carreira, Mercado Atual e Projeção

- ✓ A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais
- ✓ Apenas visualizar os dados não atende mais às necessidades das empresas e instituições, a palavra de ordem é: RECOMENDAÇÃO
- ✓ Este profissional é o responsável por gerar conhecimento para tomada de decisões rápidas e precisas

# Profissão, Carreira, Mercado Atual e Projeção

- ✓ A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais
- ✓ Apenas visualizar os dados não atende mais às necessidades das empresas e instituições, a palavra de ordem é: RECOMENDAÇÃO
- ✓ Este profissional é o responsável por gerar conhecimento para tomada de decisões rápidas e precisas
- ✓ Inclusive, é responsável por automatizar as tomadas de decisões em tempo real (Aprendizado de Máquina & Inteligência Artificial)

# Profissão, Carreira, Mercado Atual e Projeção

- ✓ A Profissão de Data Scientist se faz necessária pela enorme quantidade de dados que são gerados nos dias atuais
- ✓ Apenas visualizar os dados não atende mais às necessidades das empresas e instituições, a palavra de ordem é: RECOMENDAÇÃO
- ✓ Este profissional é o responsável por gerar conhecimento para tomada de decisões rápidas e precisas
- ✓ Inclusive, é responsável por automatizar as tomadas de decisões em tempo real (Aprendizado de Máquina & Inteligência Artificial)
- ✓ Logo, são bem remunerados: [https://www.glassdoor.com.br/Sal%C3%A1rios/data-scientist-sal%C3%A1rio-SRCH\\_K00,14.htm](https://www.glassdoor.com.br/Sal%C3%A1rios/data-scientist-sal%C3%A1rio-SRCH_K00,14.htm)

# Profissão, Carreira, Mercado Atual e Projeção

- **O mercado brasileiro acordou para o valor desta profissão graças as startups**
- São empresas que já nascem 100% digitais, com o DNA perfeito para implantação de metodologias de ciência de dados
- Elas precisam sempre pensar em processos escaláveis que compreendem tomadas de decisão em tempo real
- E as demais? Estão correndo atrás do prejuízo!

## ≡ EXAME

Imposto de Renda Venezuela Previdência Concur

### Por que o Nubank sempre busca cientistas de dados e paga até R\$ 25 mil

O Nubank não exige background de programação para contratar. Confira o que a fintech valoriza e como funciona o trabalho

Por Udacity  
© 30 jun 2018, 09h00



# Profissão, Carreira, Mercado Atual e Projeção

- O mercado brasileiro acordou para o valor desta profissão graças as startups
- São empresas que já nascem 100% digitais, com o DNA perfeito para implantação de metodologias de ciência de dados
- Elas precisam sempre pensar em processos escaláveis que compreendem tomadas de decisão em tempo real
- E as demais? Estão correndo atrás do prejuízo!

## ≡ EXAME

Imposto de Renda Venezuela Previdência Concur

### Por que o Nubank sempre busca cientistas de dados e paga até R\$ 25 mil

O Nubank não exige background de programação para contratar. Confira o que a fintech valoriza e como funciona o trabalho

Por Udacity  
© 30 jun 2018, 09h00



# Profissão, Carreira, Mercado Atual e Projeção

- O mercado brasileiro acordou para o valor desta profissão graças as startups
- São empresas que já nascem 100% digitais, com o DNA perfeito para implantação de metodologias de ciência de dados
- Elas precisam sempre pensar em processos escaláveis que compreendem tomadas de decisão em tempo real
- E as demais? Estão correndo atrás do prejuízo!

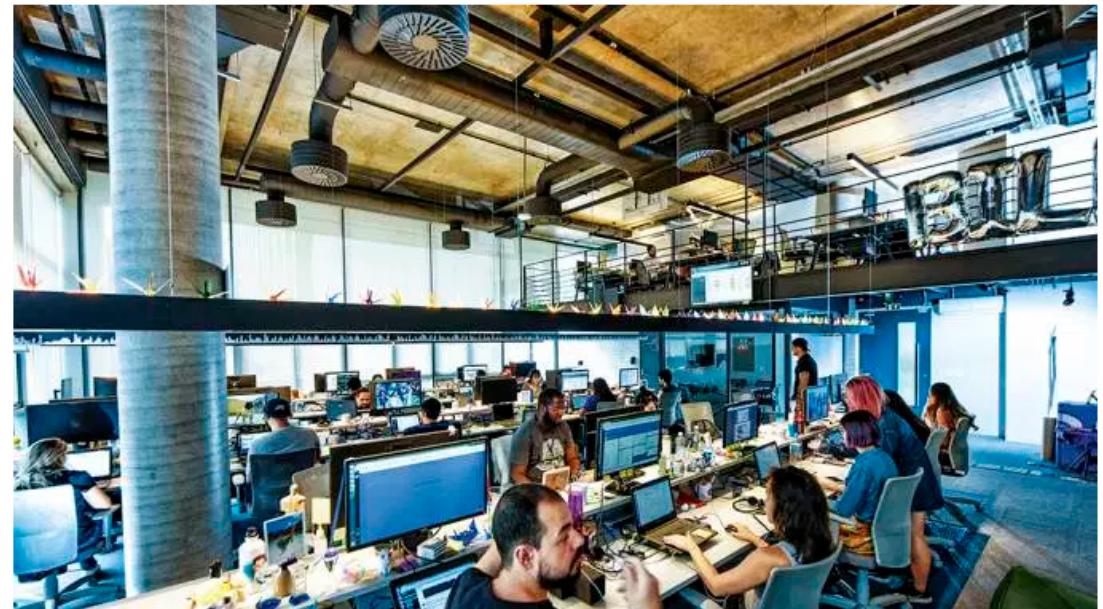
## ≡ EXAME

Imposto de Renda Venezuela Previdência Concur

### Por que o Nubank sempre busca cientistas de dados e paga até R\$ 25 mil

O Nubank não exige background de programação para contratar. Confira o que a fintech valoriza e como funciona o trabalho

Por Udacity  
© 30 jun 2018, 09h00



# Profissão, Carreira, Mercado Atual e Projeção

- O mercado brasileiro acordou para o valor desta profissão graças as startups
- São empresas que já nascem 100% digitais, com o DNA perfeito para implantação de metodologias de ciência de dados
- Elas precisam sempre pensar em processos escaláveis que compreendem tomadas de decisão em tempo real
- E as demais? Estão correndo atrás do prejuízo!

## ≡ EXAME

Imposto de Renda Venezuela Previdência Concur

### Por que o Nubank sempre busca cientistas de dados e paga até R\$ 25 mil

O Nubank não exige background de programação para contratar. Confira o que a fintech valoriza e como funciona o trabalho

Por Udacity  
© 30 jun 2018, 09h00



# O Trabalho do Cientista de Dados

- 1. Definição do problema e levantamento de perguntas a serem respondidas**
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
- 2. Planejamento do processo de Data Science**
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
- 3. Coleta de dados**
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
- 4. Processamento e limpeza dos dados**
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
- 5. Armazenamento dos dados**
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
- 6. Análise de dados**
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. **Construção e validação de algoritmos e modelos**
8. Data Visualization
9. Disseminação da informação
10. Colocar modelo em produção



# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
- 8. Data Visualization**
9. Disseminação da informação
10. Colocar modelo em produção



# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. **Disseminação da informação**
10. Colocar modelo em produção

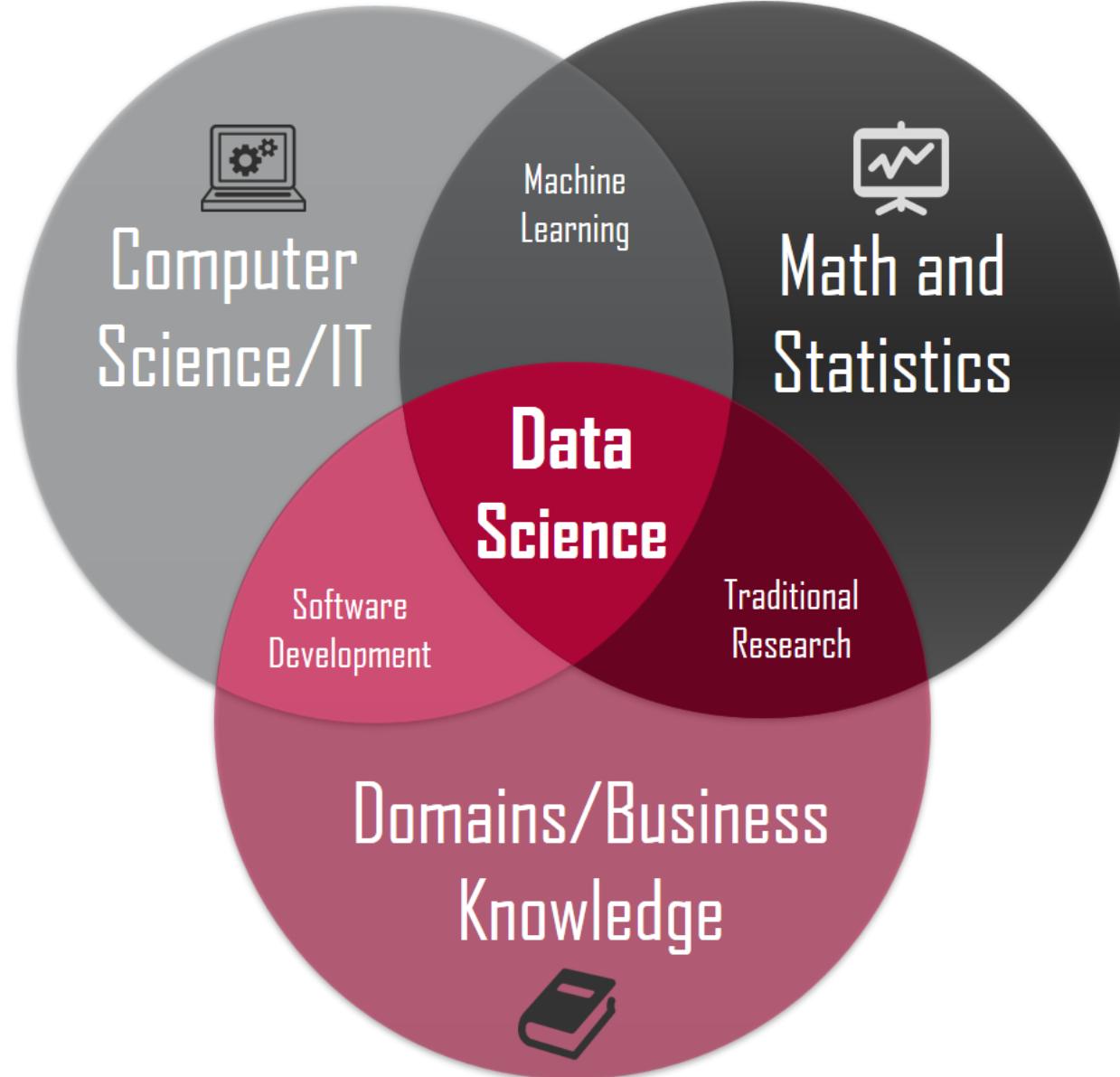


# O Trabalho do Cientista de Dados

1. Definição do problema e levantamento de perguntas a serem respondidas
2. Planejamento do processo de Data Science
3. Coleta de dados
4. Processamento e limpeza dos dados
5. Armazenamento dos dados
6. Análise de dados
7. Construção e validação de algoritmos e modelos
8. Data Visualization
9. Disseminação da informação
- 10. Colocar modelo em produção**



# Habilidades



# Habilidades

Habilidades	Analista de Dados	Engenheiro de Machine Learning	Engenheiro de Dados	Cientista de Dados
Ferramentas de programação	●	●	●	●
Visualização de Dados	●	●	●	●
Conhecimento do Negócio	●	●	●	●
Estatística	●	●	●	●
Data Preparation	●	●	●	●
Machine Learning	●	●	●	●
Engenharia de Software	●	●	●	●
Análise Multivariada	●	●	●	●

● Pouco Importante

● Importante

● Muito Importante



# Desafio Curso → AgroXP Brazil

Sua admissão como Cientista de Dados da empresa **AgroXP Brazil** não foi sem propósito. Esta empresa atua na exportação de alimentos (*commodities*) em geral. No primeiro desafio você recebeu a missão de montar, em três dias, um modelo para recomendar aos diretores da empresa os produtos que deverão ter foco na exportação nos próximos 12 meses.



Você possui os seguintes dados:

- 1) Ministério de Desenvolvimento Indústria e Comércio --> apresenta os dados de TODOS commodities exportados no País desde 1997 até 1 mês atrás (formato .csv)
- 2) Tabelas auxiliares de nomenclatura de produtos com NCM – Nomenclatura Comum do Mercosul (formato .xls)
- 3) Taxa cambial mensal desde 1997 (formato .csv)

# Desafio Curso → AgroXP Brazil

Sua admissão como Cientista de Dados da empresa **AgroXP Brazil** não foi sem propósito. Esta empresa atua na exportação de alimentos (*commodities*) em geral. No primeiro desafio você recebeu a missão de montar, em três dias, um modelo para recomendar aos diretores da empresa os produtos que deverão ter foco na exportação nos próximos 12 meses.



Você possui os seguintes dados:

- 1) [Ministério de Desenvolvimento Indústria e Comércio](#) --> apresenta os dados de TODOS commodities exportados no País desde 1997 até abril 2019(formato .csv)
- 2) Tabelas auxiliares de nomenclatura de produtos com NCM – Nomenclatura Comum do Mercosul (formato .xls)
- 3) Taxa cambial mensal desde 1997 (formato .csv)

# Desafio Curso → AgroXP Brazil

Sua admissão como Cientista de Dados da empresa **AgroXP Brazil** não foi sem propósito. Esta empresa atua na exportação de alimentos (*commodities*) em geral. No primeiro desafio você recebeu a missão de montar, em três dias, um modelo para recomendar aos diretores da empresa os produtos que deverão ter foco na exportação nos próximos 12 meses.



Você possui os seguintes dados:

- 1) [Ministério de Desenvolvimento Indústria e Comércio](#) --> apresenta os dados de TODOS commodities exportados no País desde 1997 até abril 2019 (formato .csv)
- 2) Tabelas auxiliares de nomenclatura de produtos com NCM – Nomenclatura Comum do Mercosul (formato .xls)
- 3) Taxa cambial mensal desde 1997 (formato .csv)

# Desafio Curso → AgroXP Brazil

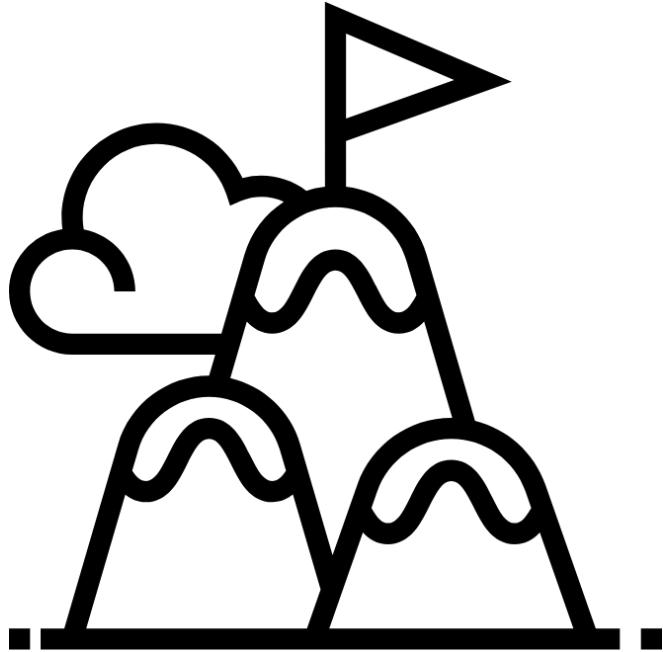
Sua admissão como Cientista de Dados da empresa **AgroXP Brazil** não foi sem propósito. Esta empresa atua na exportação de alimentos (*commodities*) em geral. No primeiro desafio você recebeu a missão de montar, em três dias, um modelo para recomendar aos diretores da empresa os produtos que deverão ter foco na exportação nos próximos 12 meses.



Você possui os seguintes dados:

- 1) [Ministério de Desenvolvimento Indústria e Comércio](#) --> apresenta os dados de TODOS commodities exportados no País desde 1997 até abril 2019 (formato .csv)
- 2) Tabelas auxiliares de nomenclatura de produtos com NCM – Nomenclatura Comum do Mercosul (formato .xls)
- 3) Taxa cambial mensal desde 1997 (formato .csv)

# Desafio Pessoal



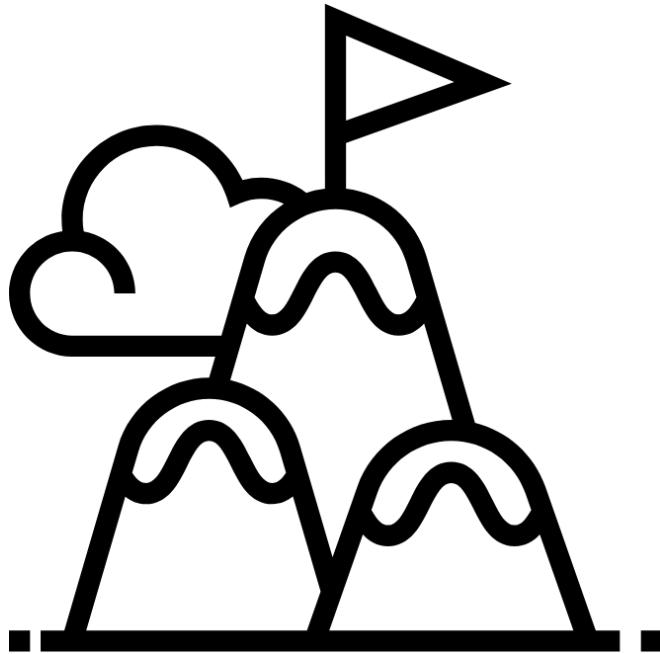
**Machine Learning**

**Apresentação na Aldeia**

**O Melhor Algoritmo Vence**

- Maior Acurácia
- Menor Tempo Processamento

# Desafio Pessoal



Machine Learning

Apresentação na Aldeia

O Melhor Algoritmo Vence

→ Maior Acurácia

→ Menor Tempo Processamento

Você vai **AMARELAR** no Desafio?

# Desafio Pessoal



110101011001



# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

7 – Previsão

8 – Correlação



# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

7 – Previsão

8 – Correlação



# Média

A Média é um cálculo relativamente simples que possui muitas aplicações:

$$\text{Média} = \frac{\text{Soma dos Termos}}{\text{Quantidade dos Termos}}$$

Na estatística a média é chamada de **Esperança**, comumente utilizado em  
Inferências e Previsões

# Média

A tabela abaixo informa a cotação do dólar (moeda estrangeira) durante a última semana:

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$3,79	R\$3,78	R\$3,76	R\$3,80	R\$3,90	R\$3,95	R\$3,86	R\$3,95

# Média

A tabela abaixo informa a cotação do dólar (moeda estrangeira) durante a última semana:

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$3,79	R\$3,78	R\$3,76	R\$3,80	R\$3,90	R\$3,95	R\$3,86	R\$3,95

Bora Calcular???

# Média

A tabela abaixo informa a cotação do dólar (moeda estrangeira) durante a última semana:

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$3,79	R\$3,78	R\$3,76	R\$3,80	R\$3,90	R\$3,95	R\$3,86	R\$3,95

$$\text{Média} = \frac{\text{R\$3,79} + \text{R\$3,78} + \text{R\$3,76} + \text{R\$3,80} + \text{R\$3,90} + \text{R\$3,95} + \text{R\$3,86} + \text{R\$3,95}}{8}$$

$$\text{Média} = \frac{\text{R\$30,79}}{8} \quad \rightarrow \quad \text{Média} = \text{R\$3,85}$$

# Média

E se os valores da cotação fossem os abaixo, a média continua sendo um bom estimador?

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$1,79	R\$3,78	R\$3,76	R\$3,80	R\$15,90	R\$3,95	R\$3,86	R\$3,95

# Média

E se os valores da cotação fossem os abaixo, a média continua sendo um bom estimador?

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$1,79	R\$3,78	R\$3,76	R\$3,80	R\$15,90	R\$3,95	R\$3,86	R\$3,95

1 - Qual é a Média?

2 - Qual a Variação % entre os resultados?

3 - O que pode-se concluir?

# Média

E se os valores da cotação fossem os abaixo, a média continua sendo um bom estimador?

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$1,79	R\$3,78	R\$3,76	R\$3,80	R\$15,90	R\$3,95	R\$3,86	R\$3,95

$$\text{Média} = \frac{\text{R\$1,79} + \text{R\$3,78} + \text{R\$3,76} + \text{R\$3,80} + \text{R\$15,90} + \text{R\$3,95} + \text{R\$3,86} + \text{R\$3,95}}{8}$$

$$\text{Média} = \frac{\text{R\$40,79}}{8} \quad \rightarrow \quad \begin{aligned} 1 - \text{Média} &= \text{R\$5,10} \\ 2 - \text{Var \%} &= 32,46\% \\ 3 - \text{Não Recomendado} \end{aligned}$$

# Mediana

Para o segundo exemplo é indicado utilizar a **mediana**, como é possível verificar abaixo:

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$1,79	R\$3,78	R\$3,76	R\$3,80	R\$15,90	R\$3,95	R\$3,86	R\$3,95

**1º Passo:** Organizar os números em ordem crescente:

R\$ 1,79	R\$ 3,76	R\$ 3,78	R\$ 3,80	R\$ 3,86	R\$ 3,95	R\$ 3,95	R\$ 15,90
----------	----------	----------	----------	----------	----------	----------	-----------

# Mediana

2º Passo: Calcular a média dos dois números do meio:

R\$ 1,79    R\$ 3,76    R\$ 3,78    R\$ 3,80    R\$ 3,86    R\$ 3,95    R\$ 3,95    R\$ 15,90

$$\text{Mediana} = \frac{\text{R\$}3,80 + 3,86}{2} \rightarrow \text{Mediana} = \text{R\$}3,83$$

Qual é a **mediana** dos números abaixo?

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$3,79	R\$3,78	R\$3,76	R\$3,80	R\$3,90	R\$3,95	R\$3,86	R\$3,95

Qual é a **mediana** dos números abaixo?

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$3,79	R\$3,78	R\$3,76	R\$3,80	R\$3,90	R\$3,95	R\$3,86	R\$3,95

$$\text{Mediana} = \frac{\text{R\$}3,80 + 3,86}{2} \rightarrow \text{Mediana} = \text{R\$}3,83$$

A **Moda** é o valor que mais aparece num conjunto de dados

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$3,79	R\$3,78	R\$3,76	R\$3,80	R\$3,90	R\$3,95	R\$3,86	R\$3,95

A Moda é o valor que mais aparece num conjunto de dados

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$3,79	R\$3,78	R\$3,76	R\$3,80	R\$3,90	R\$3,95	R\$3,86	R\$3,95

**Moda = R\$3,95**

# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

7 – Previsão

8 – Correlação



# Outlier

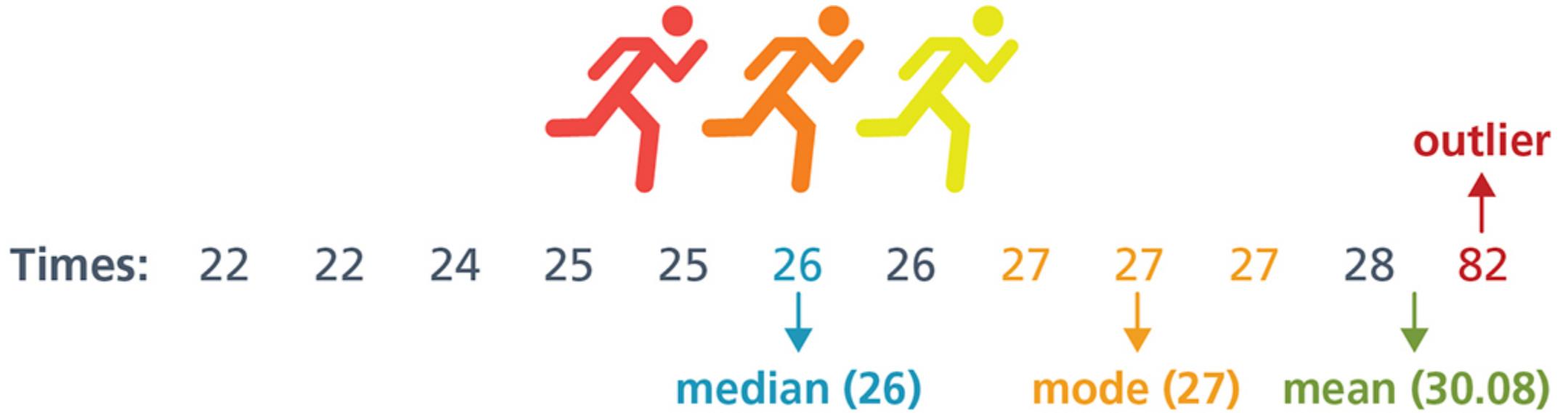
- Em estatística, **outlier**, valor aberrante ou atípico, é uma observação que apresenta um afastamento das demais da série, ou que é inconsistente



- A existência de outliers implica, tipicamente, em prejuízos a interpretação dos resultados dos testes estatísticos aplicados às amostras

# Outlier

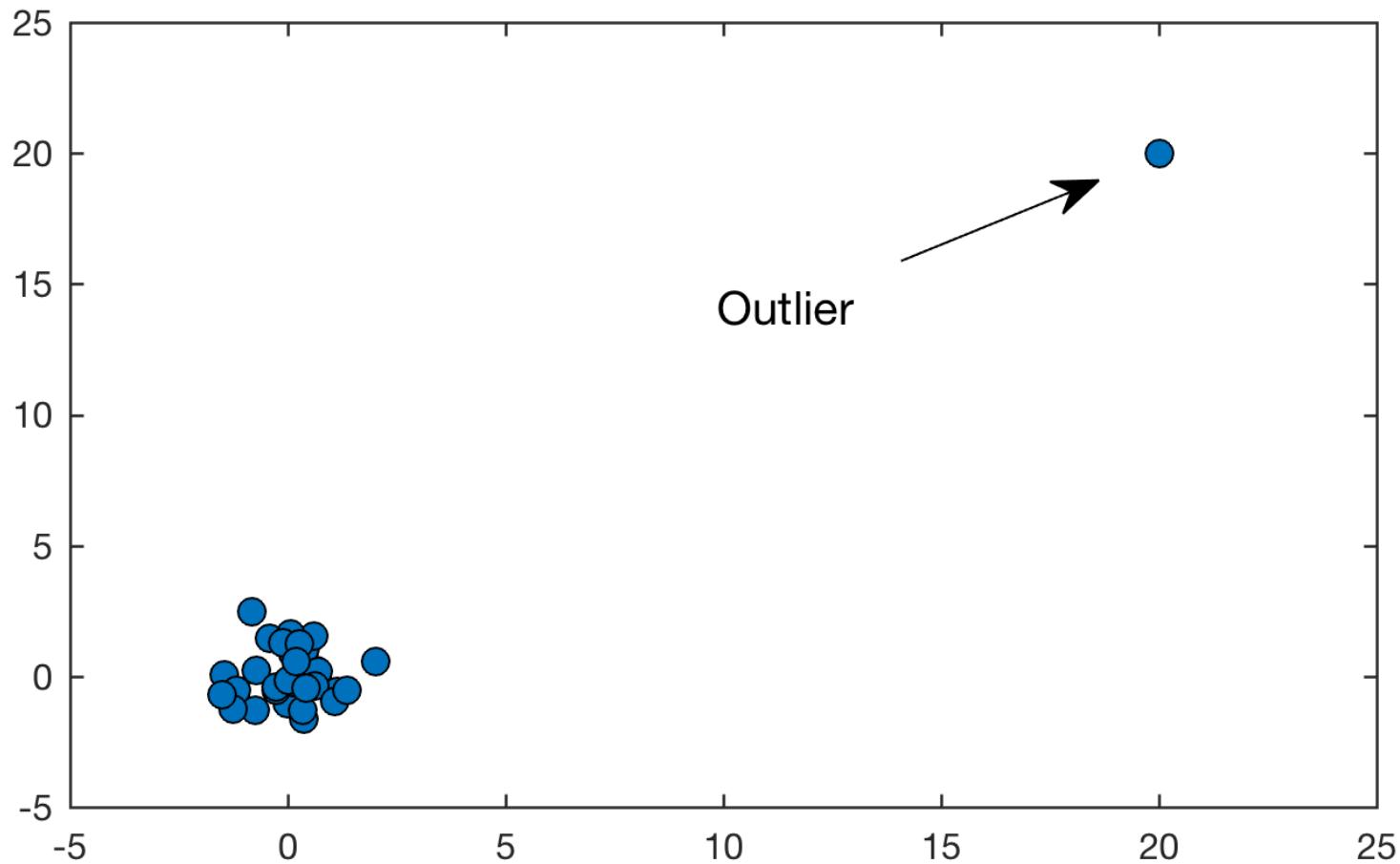
- Em estatística, **outlier**, valor aberrante ou atípico, é uma observação que apresenta um afastamento das demais da série, ou que é inconsistente



- A existência de outliers implica, tipicamente, em prejuízos a interpretação dos resultados dos testes estatísticos aplicados às amostras... **Depende!!!**

# Outlier

- Em muitos casos encontrar o **outlier** é o objetivo do estudo, como por exemplo: detecção de fraudes, encontrar espécies mais resistentes, seleção de atletas de alto desempenho, etc



# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

7 – Previsão

8 – Correlação



# População e Amostra

A **População** é um conjunto de pessoas, itens ou eventos sobre os quais existe interesse em inferir

A **Amostra** é um subconjunto de pessoas, itens ou eventos de uma população que é coletada e analisada para fazer inferências



# População e Amostra

Métodos de **Amostragem Probabilística** são os que selecionam os indivíduos da população de forma que todos tenham as mesmas chances de participar da amostra.

**Erro Amostral** é a diferença entre um resultado amostral e o verdadeiro resultado populacional; tais erros resultam de flutuações amostrais aleatórias



# População e Amostra

A fórmula para cálculo do tamanho da amostra para uma estimativa confiável da **Média Populacional ( $\mu$ )** é dada por:

$$n = \frac{N}{1+Ne^2}$$

**n** = Número de indivíduos na amostra

**N** = Tamanho da População

**e** = Margem de Erro



<https://pt.wikihow.com/Calcular-o-Tamanho-de-uma-Amostra>

# População e Amostra

**Qual é o tamanho de Amostra para uma população de 859.347 indivíduos com uma margem de erro de 1,25% ?**



<https://pt.wikihow.com/Calcular-o-Tamanho-de-uma-Amostra>

# População e Amostra

Qual é o tamanho de Amostra para uma população de 859.347 indivíduos com uma margem de erro de 1,25% ?

$$n = 6.352$$

E para margem de erro de 5%?



<https://pt.wikihow.com/Calcular-o-Tamanho-de-uma-Amostra>

# População e Amostra

Qual é o tamanho de Amostra para uma população de 859.347 indivíduos com uma margem de erro de 1,25% ?

$$n = 6.352$$

E para margem de erro de 5%?



<https://pt.wikihow.com/Calcular-o-Tamanho-de-uma-Amostra>

# População e Amostra

Qual é o tamanho de Amostra para uma população de 859.347 indivíduos com uma margem de erro de 1,25% ?

**n = 6.352**

E para margem de erro de 5%?

**n = 400**



<https://pt.wikihow.com/Calcular-o-Tamanho-de-uma-Amostra>

# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

7 – Previsão

8 – Correlação



# Probabilidade

A palavra **probabilidade** deriva do Latim probare (provar ou testar).

Informalmente, provável é uma das muitas palavras utilizadas para eventos incertos ou desconhecidos , sendo também substituída por algumas palavras como "sorte", "risco", "azar", "chance", "incerteza", "duvidoso", dependendo do contexto.



# Probabilidade

No lançamento de um dado, observa-se os seguintes eventos e suas probabilidades:

A = Obter um número par:

$$A = \{2, 4, 6\} \text{ e } n(A) = 3$$

$$P(A) = 3/6 = 0,5 \text{ ou } 50\%$$

B = Sair um número primo:

$$B = \{2, 3, 5\} \text{ e } n(B) = 3$$

$$P(B) = 3/6 = 0,5 \text{ ou } 50\%$$

C = Sair um número maior ou igual a 5:

$$C = \{5, 6\} \text{ e } n(C) = 2$$

$$P(C) = 2/6 = 0,333 \text{ ou } 33,3\%$$

D = Sair um número natural:

$$D = \{1, 2, 3, 4, 5, 6\} \text{ e } n(D) = 6$$

$$P(D) = 6/6 = 1 \text{ ou } 100\%$$

# Probabilidade

No lançamento de um dado, observa-se os seguintes eventos e suas probabilidades:

A = Obter um número par:

$$A = \{2, 4, 6\} \text{ e } n(A) = 3$$

$$P(A) = 3/6 = 0,5 \text{ ou } 50\%$$

B = Sair um número primo:

$$B = \{2, 3, 5\} \text{ e } n(B) = 3$$

$$P(B) = 3/6 = 0,5 \text{ ou } 50\%$$

C = Sair um número maior ou igual a 5:

$$C = \{5, 6\} \text{ e } n(C) = 2$$

$$P(C) = 2/6 = 0,333 \text{ ou } 33,3\%$$

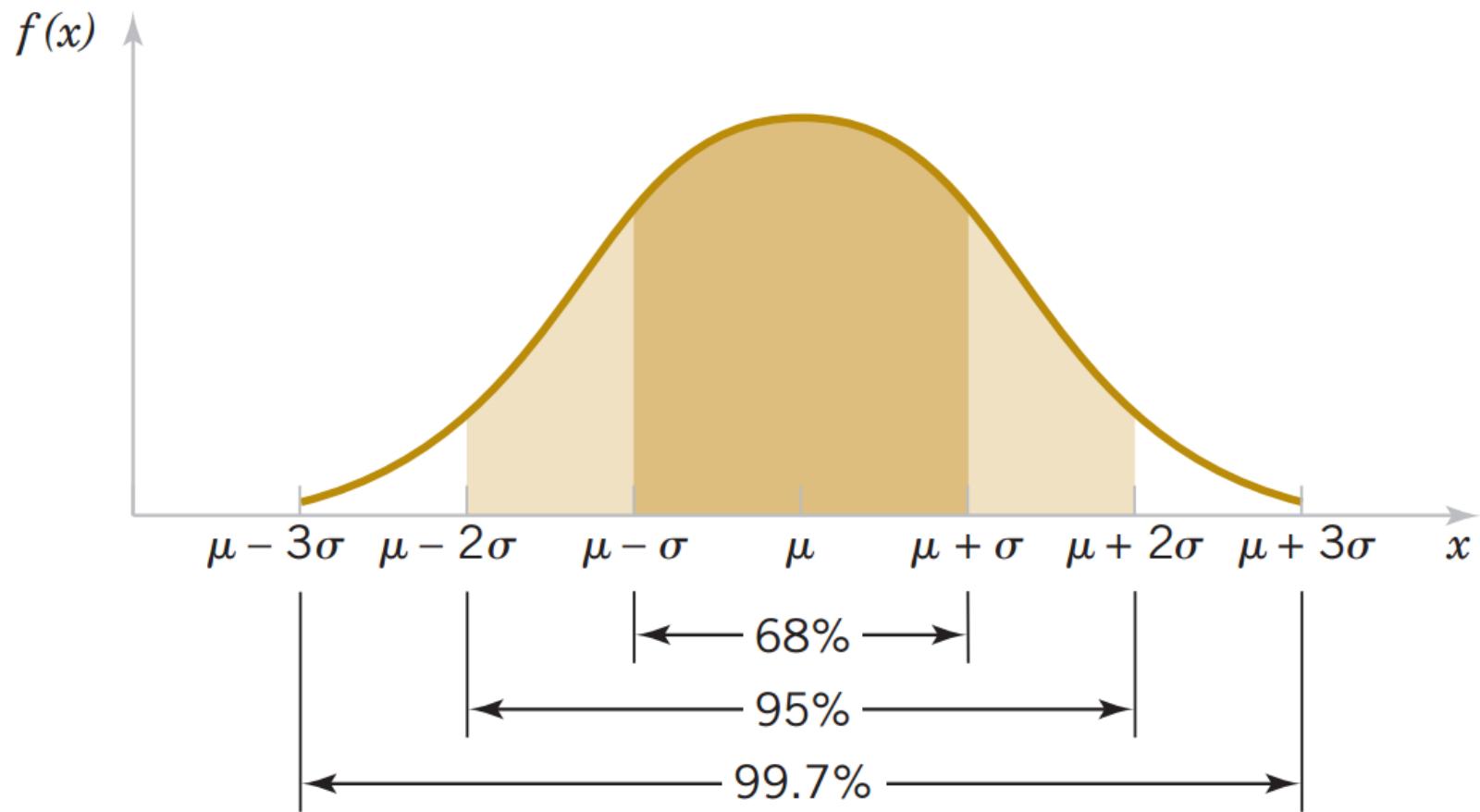
D = Sair um número natural:

$$D = \{1, 2, 3, 4, 5, 6\} \text{ e } n(D) = 6$$

$$P(D) = 6/6 = 1 \text{ ou } 100\%$$

# Probabilidade

A **Distribuição Normal** é uma das distribuições de probabilidade mais utilizadas para modelar fenômenos naturais. Isso se deve ao fato de que um grande número de fenômenos naturais apresenta sua distribuição de probabilidade tão proximamente normal



# Probabilidade

A **Distribuição Normal** é uma das distribuições de probabilidade mais utilizadas para modelar fenômenos naturais. Isso se deve ao fato de que um grande número de fenômenos naturais apresenta sua distribuição de probabilidade tão proximamente normal

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

7 – Previsão

8 – Correlação



# Variáveis: Discreta, Contínua

**Discretas:** características mensuráveis que podem assumir apenas um número inteiro finito ou infinito contável de valores e geralmente são o resultado de contagens.

Exemplos: número de filhos, número de bactérias, quantidade de produtos vendidos

**Contínuas:** características mensuráveis que assumem valores em uma escala contínua, para as quais valores fracionais fazem sentido, usualmente medidas através de algum instrumento. Exemplos: peso (balança), tempo (relógio), pressão arterial, idade.



# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

7 – Previsão

8 – Correlação



# Variância e Desvio Padrão

A variância e o desvio padrão são medidas que dão uma ideia da dispersão de uma distribuição de dados

# Variância e Desvio Padrão

A variância e o desvio padrão são medidas que dão uma ideia da dispersão de uma distribuição de dados

A variância irá mostrar com eficiência a distância existente entre os valores em cada conjunto, mostrando a distância em que o conjunto se encontra com referência ao valor central

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

# Variância e Desvio Padrão

A **variância** e o **desvio padrão** são medidas que dão uma ideia da dispersão de uma distribuição de dados

A **variância** irá mostrar com eficiência a distância existente entre os valores em cada conjunto, mostrando a distância em que o conjunto se encontra com referência ao valor central

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

O **desvio padrão** serve para identificar onde existe um erro na amostragem de dados. Se existir discordância nos dados e assim pode-se substituí-los pela média aritmética do conjunto

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

# Variância e Desvio Padrão

A variância e o desvio padrão são medidas que dão uma ideia da dispersão de uma distribuição de dados

	NOTA	DESVIOS EM RELAÇÃO À MÉDIA	DESVIOS AO QUADRADO
<b>Julia</b>	<b>9,0</b>	<b>3,2</b>	<b>10,24</b>
<b>Marcos</b>	<b>7,0</b>	<b>1,2</b>	<b>1,44</b>
<b>Maria</b>	<b>5,0</b>	<b>-0,8</b>	<b>0,64</b>
<b>Andreza</b>	<b>4,0</b>	<b>-1,8</b>	<b>3,24</b>
<b>Yuri</b>	<b>4,0</b>	<b>-1,8</b>	<b>3,24</b>

# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

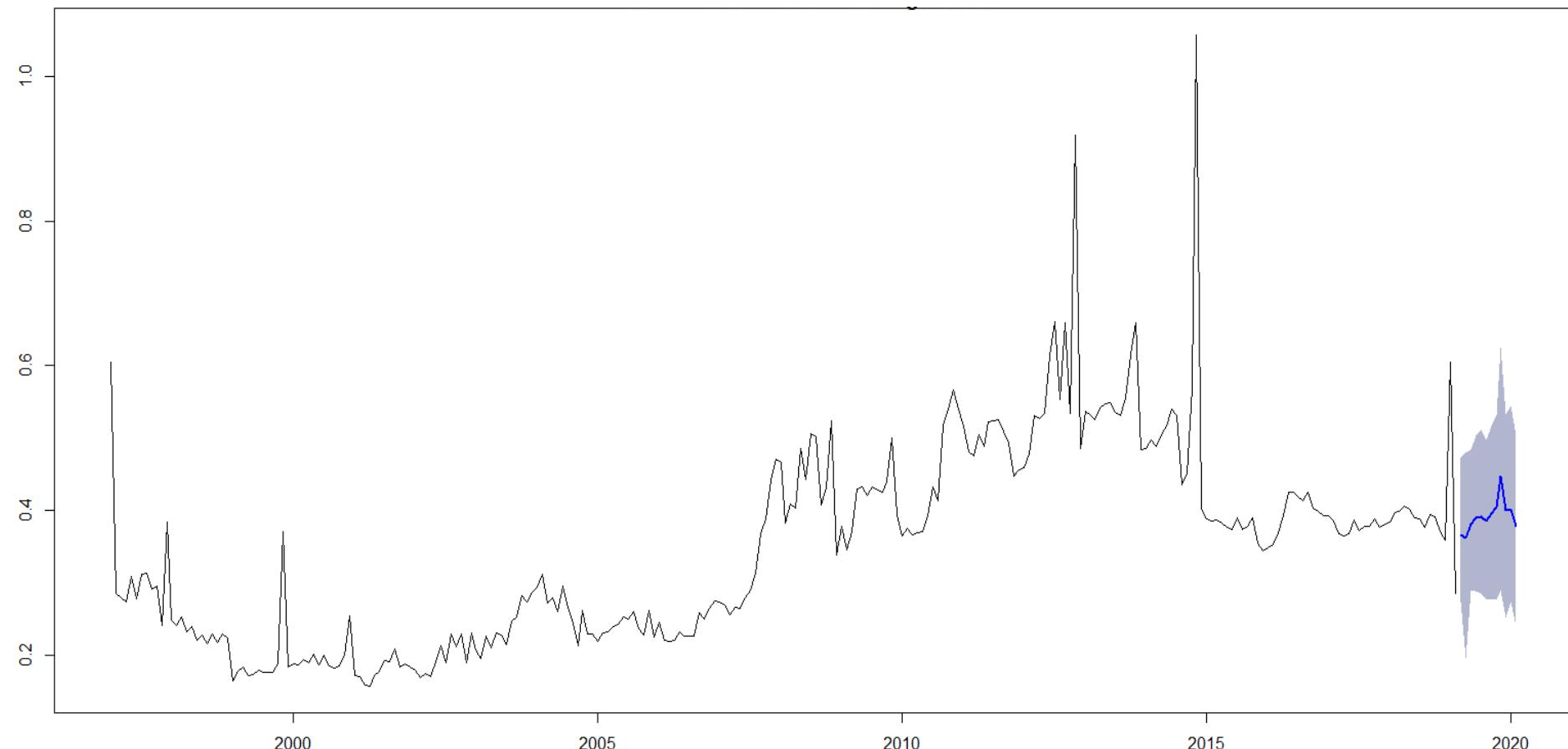
7 – Previsão

8 – Correlação



# Previsão

Previsão é o processo de estimativas em situações de incertezas e evoluiu para a prática do plano de demanda diária para tomada de decisões em negócios



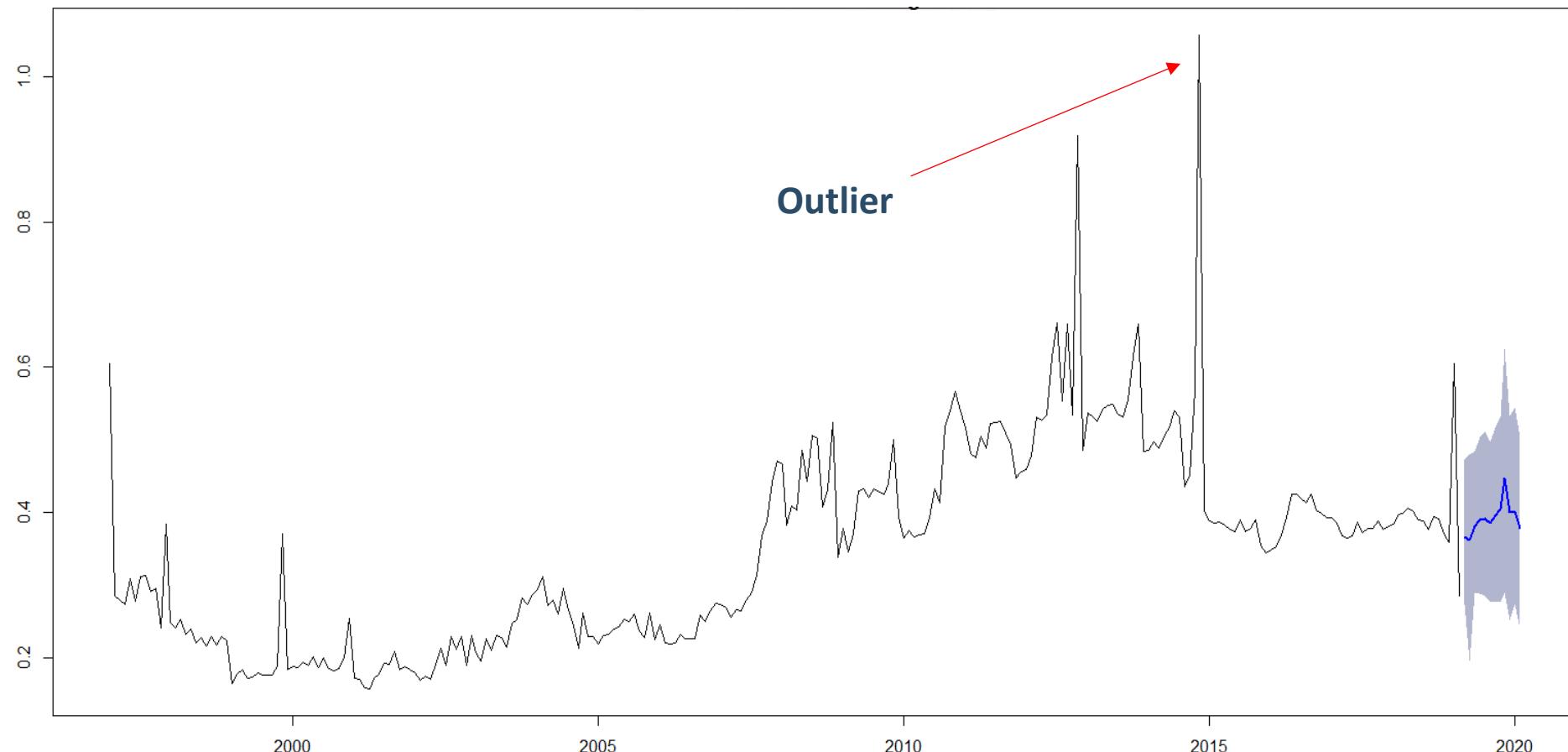
# Previsão

Previsão é o processo de estimativas em situações de incertezas e evoluiu para a prática do plano de demanda diária para tomada de decisões em negócios



# Previsão

Previsão é o processo de estimativas em situações de incertezas e evoluiu para a prática do plano de demanda diária para tomada de decisões em negócios



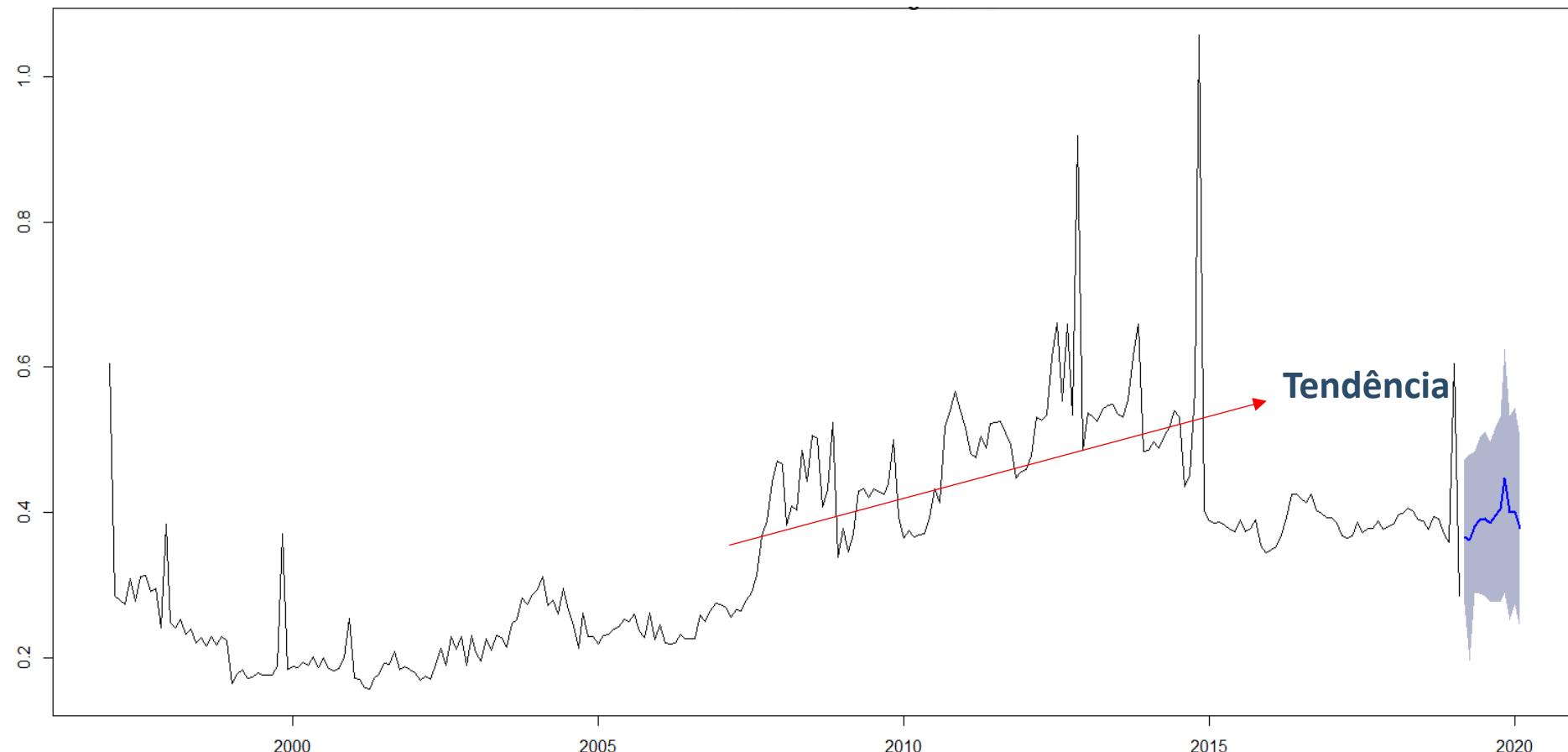
# Previsão

Previsão é o processo de estimativas em situações de incertezas e evoluiu para a prática do plano de demanda diária para tomada de decisões em negócios



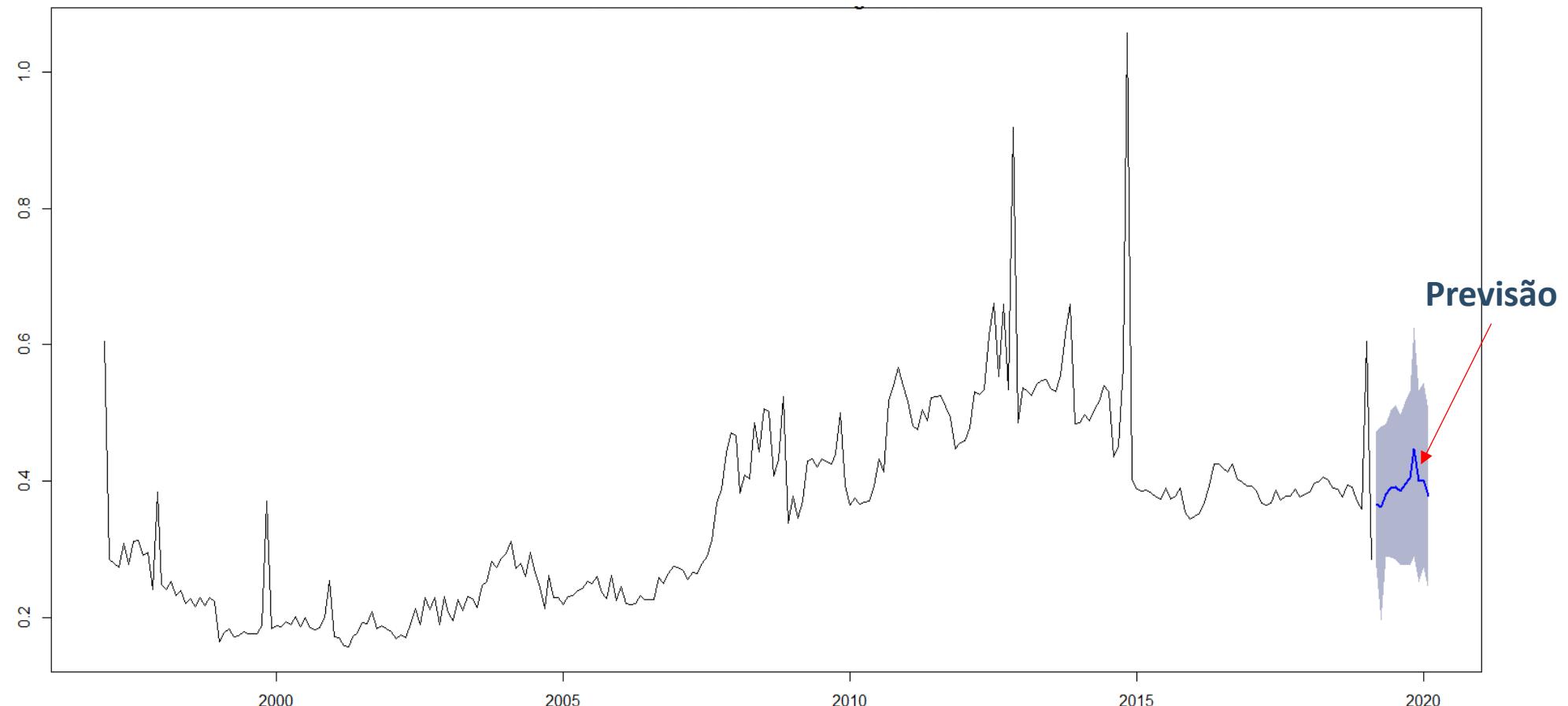
# Previsão

Previsão é o processo de estimativas em situações de incertezas e evoluiu para a prática do plano de demanda diária para tomada de decisões em negócios



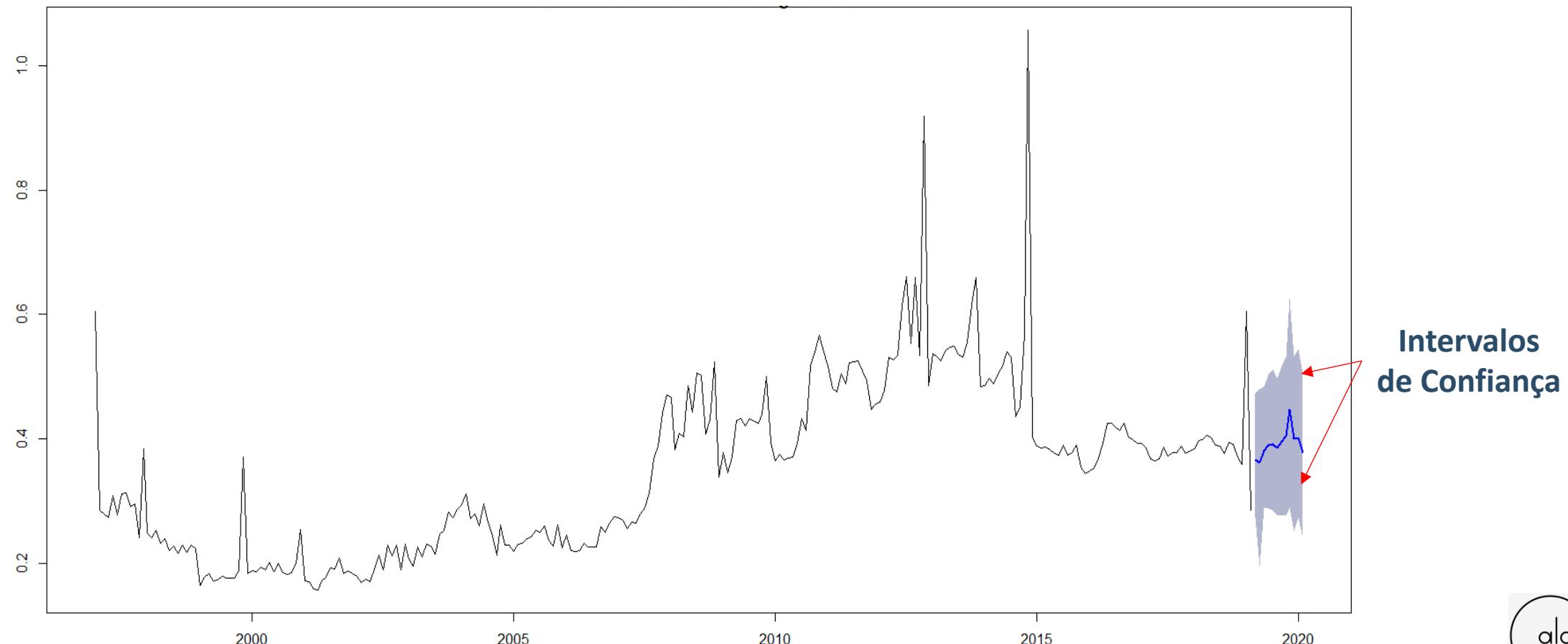
# Previsão

Previsão é o processo de estimativas em situações de incertezas e evoluiu para a prática do plano de demanda diária para tomada de decisões em negócios



# Previsão

Previsão é o processo de estimativas em situações de incertezas e evoluiu para a prática do plano de demanda diária para tomada de decisões em negócios



# Previsão

Façam uma Previsão dos próximos 2 dias de valor de fechamento do dólar:

18/03	19/03	20/03	21/03	22/03	25/03	26/03	27/03
R\$3,79	R\$3,78	R\$3,76	R\$3,80	R\$3,90	R\$3,95	R\$3,86	R\$3,95

Justifique!

# Conceitos Estatísticos

1 – Média, Mediana e Moda

2 – Outlier

3 – População e Amostra

4 – Probabilidade

5 – Variáveis: Discreta, Contínua

6 – Variância e Desvio Padrão

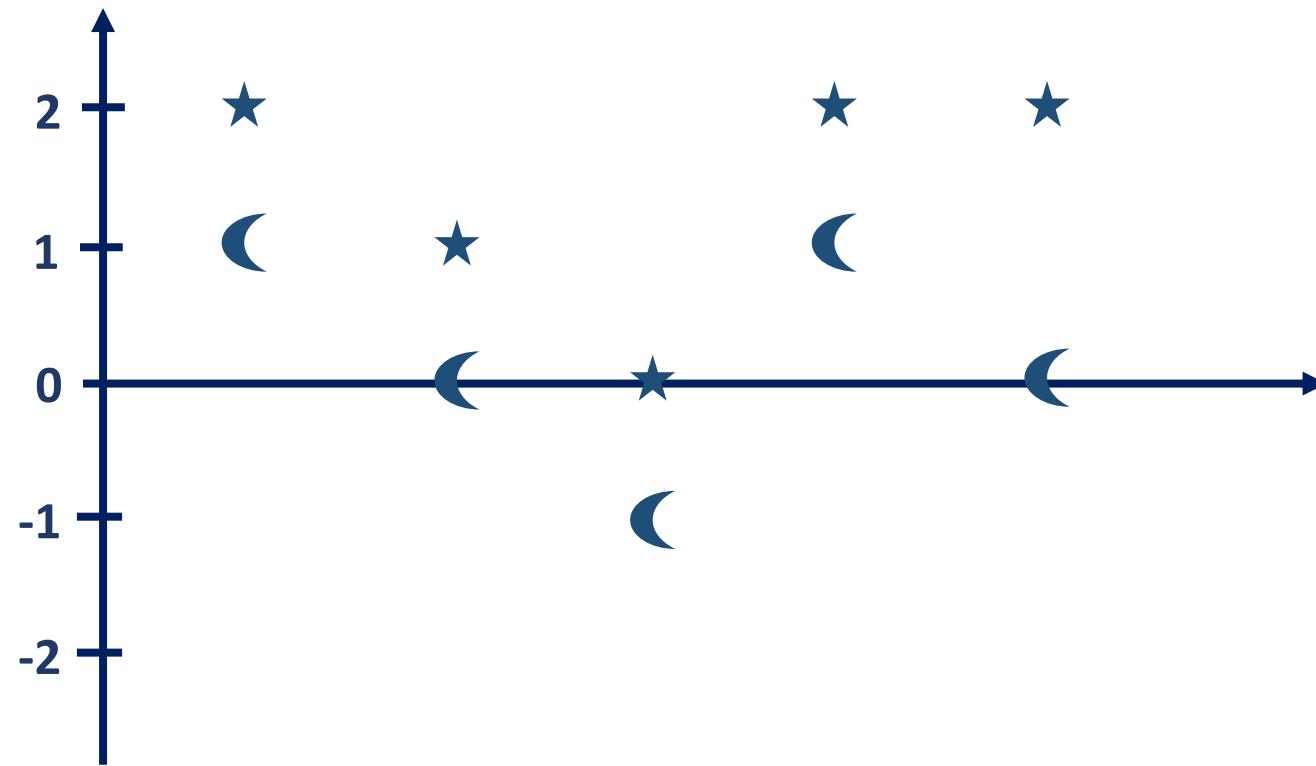
7 – Previsão

8 – Correlação



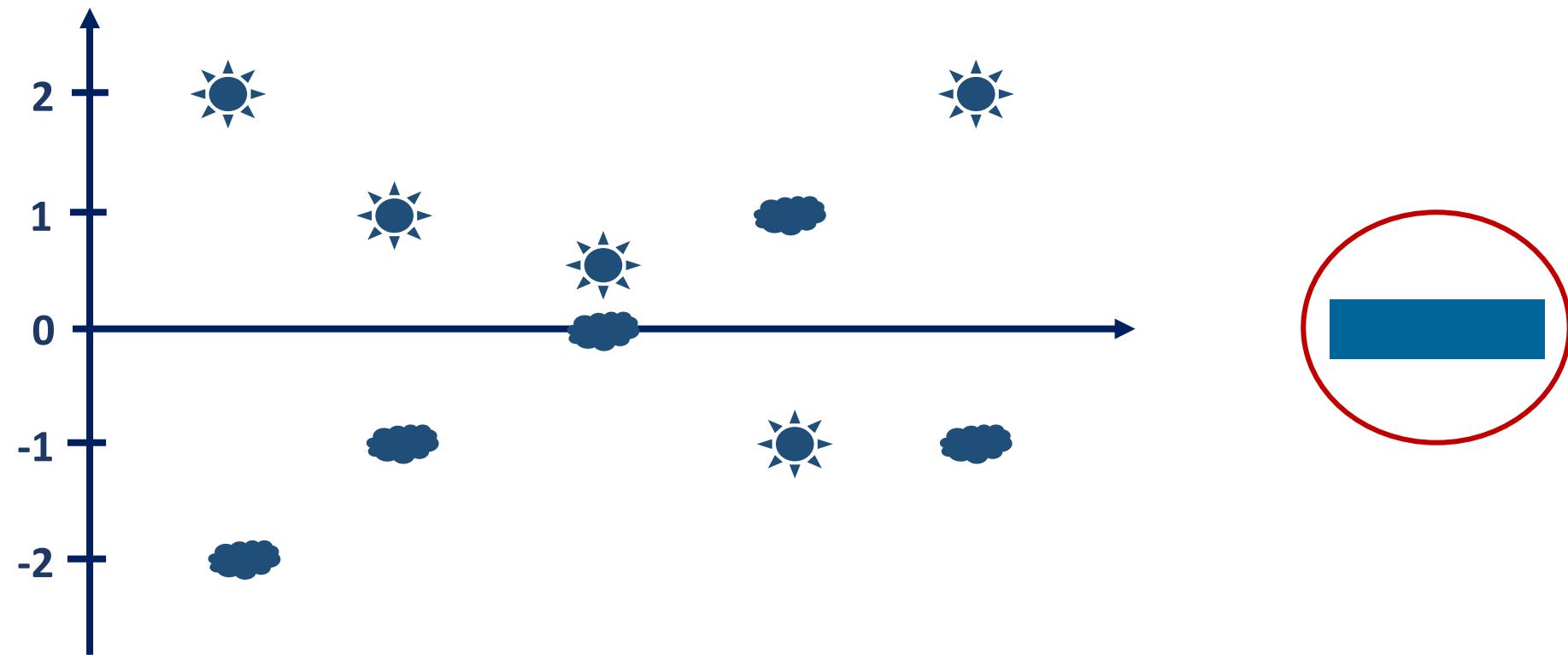
# Correlação

Correlação, dependência ou associação é qualquer relação estatística (causal ou não causal) entre duas variáveis, como por exemplo, a correlação entre o céu ensolarado e poucas nuvens.



# Correlação

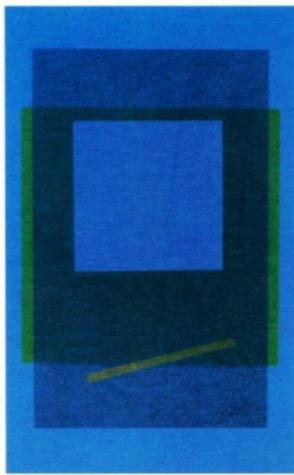
Correlação, dependência ou associação é qualquer relação estatística (causal ou não causal) entre duas variáveis, como por exemplo, a correlação entre a estatura dos pais e a estatura dos filhos



# Referências

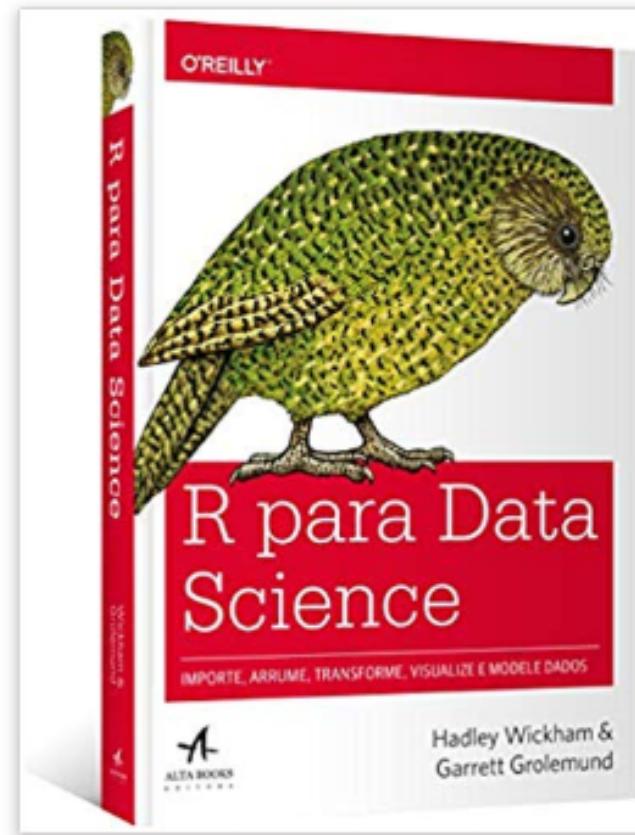
NOÇÕES DE PROBABILIDADE  
E ESTATÍSTICA

Marcos Nascimento Magalhães  
Antonio Carlos Pedroso de Lima



edusp

**Noções de  
Probabilidade e  
Estatística (Português)**  
Capa Comum – 5 jun  
2013  
por [Marcos  
Nascimento  
Magalhães](#) (Autor), [Ant  
onio Carlos Pedroso de  
Lima](#) (Autor)



**R Para Data  
Science (Português)** Ca  
pa Comum – 7 fev  
2019  
por [Hadley  
Wickham](#) (Autor), [Garr  
ett  
Grolemund](#) (Autor), [Sa  
mantha  
Batista](#) (Tradutor)

# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning



# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning



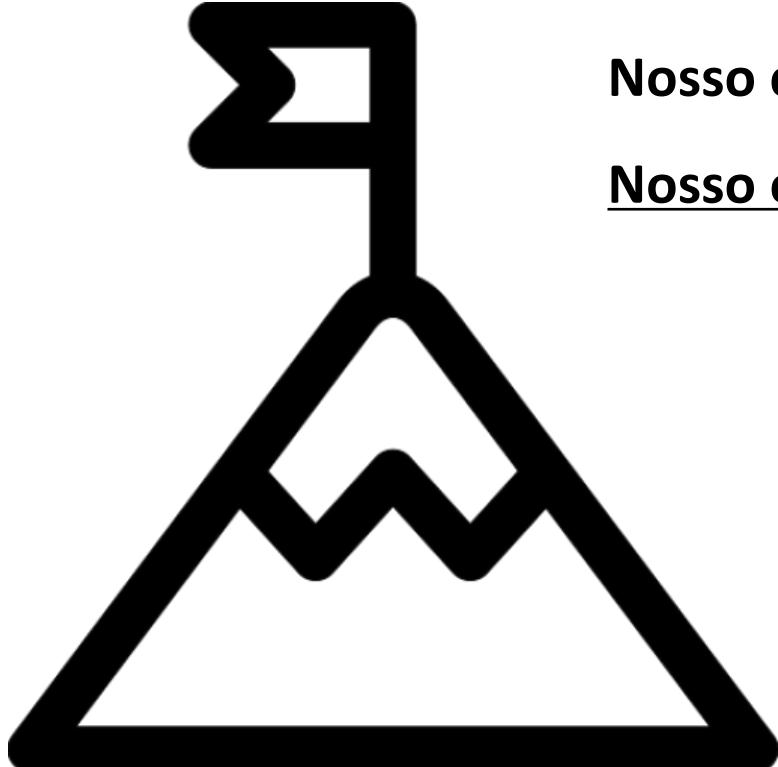
# Objetivo

Graduação área de Tecnologia da Informação → de 3 a 5 anos

Especializações em TI → de 1 a 2 anos

Nosso curso → 32h

Nosso objetivo neste tema?



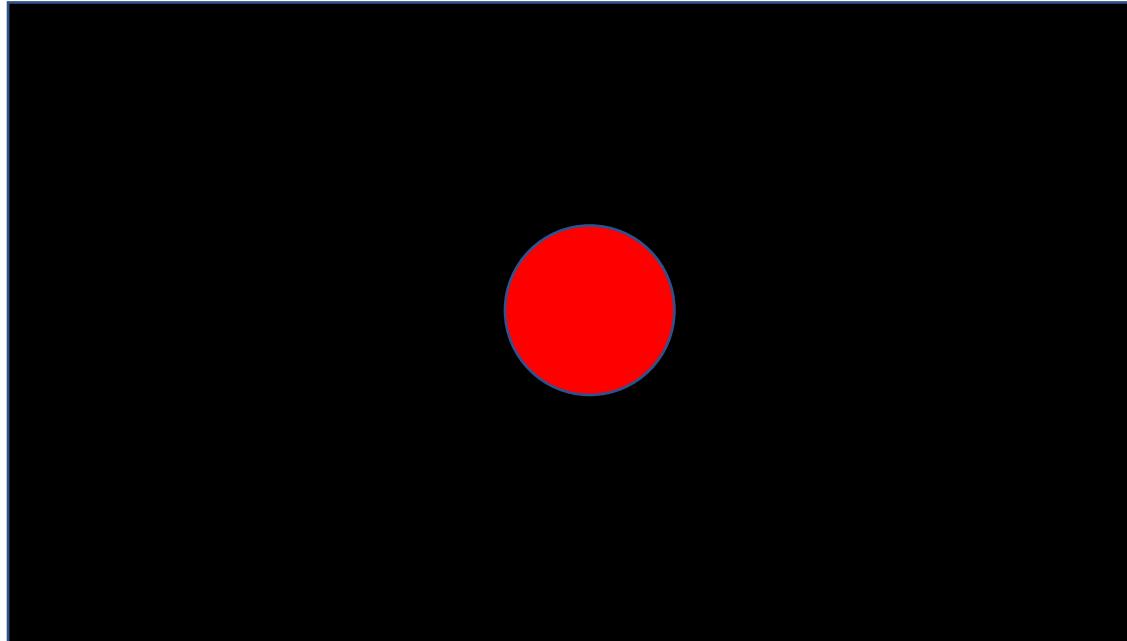
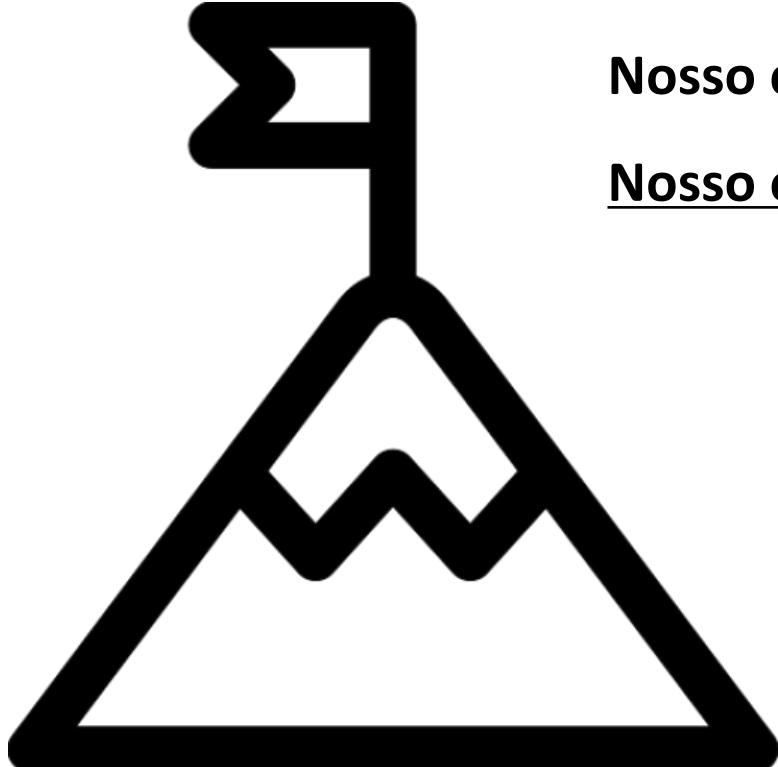
# Objetivo

Graduação área de Tecnologia da Informação → de 3 a 5 anos

Especializações em TI → de 1 a 2 anos

Nosso curso → 32h

Nosso objetivo neste tema?



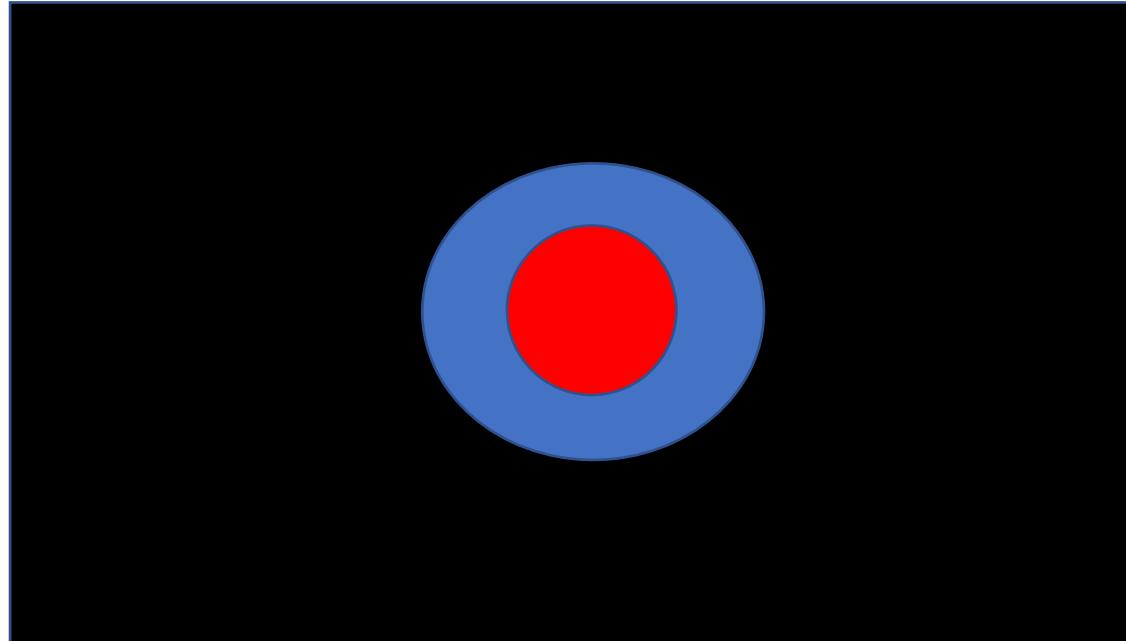
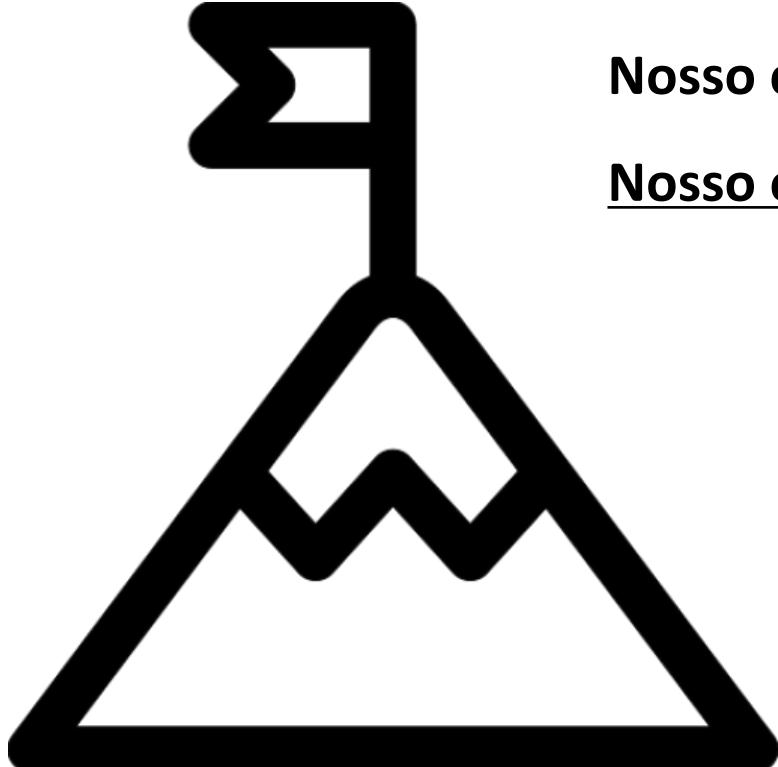
# Objetivo

Graduação área de Tecnologia da Informação → de 3 a 5 anos

Especializações em TI → de 1 a 2 anos

Nosso curso → 32h

Nosso objetivo neste tema?



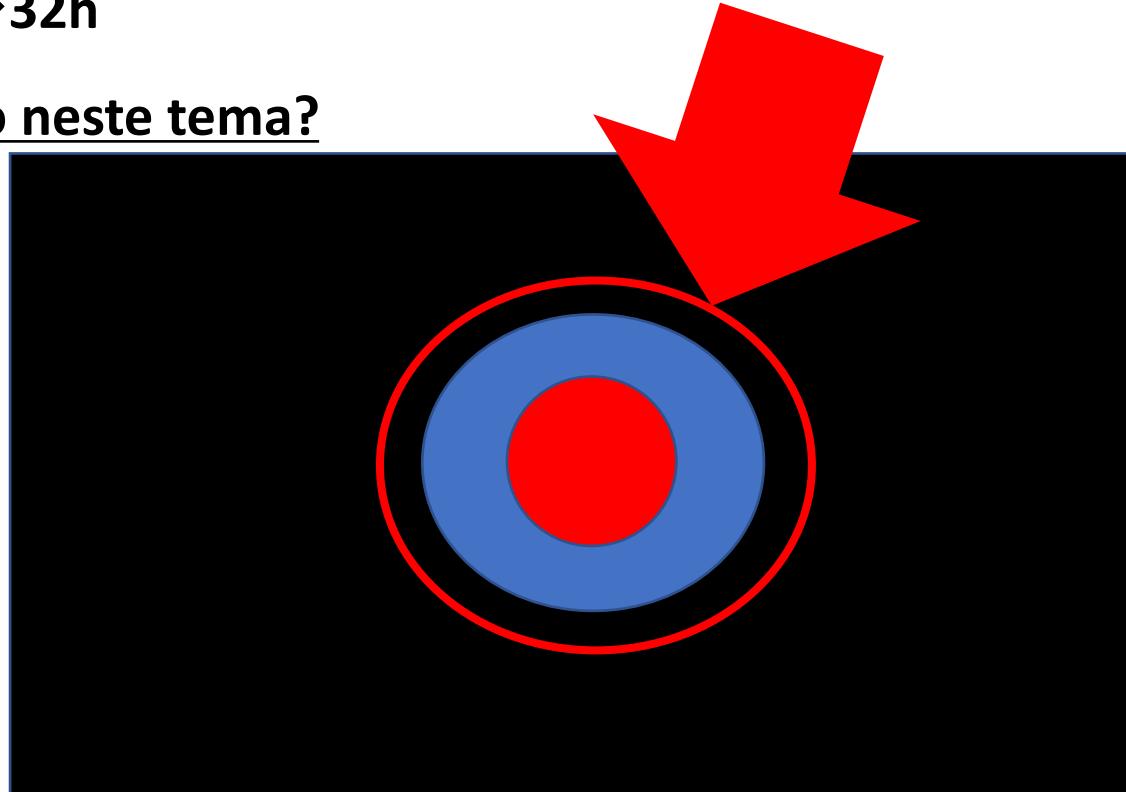
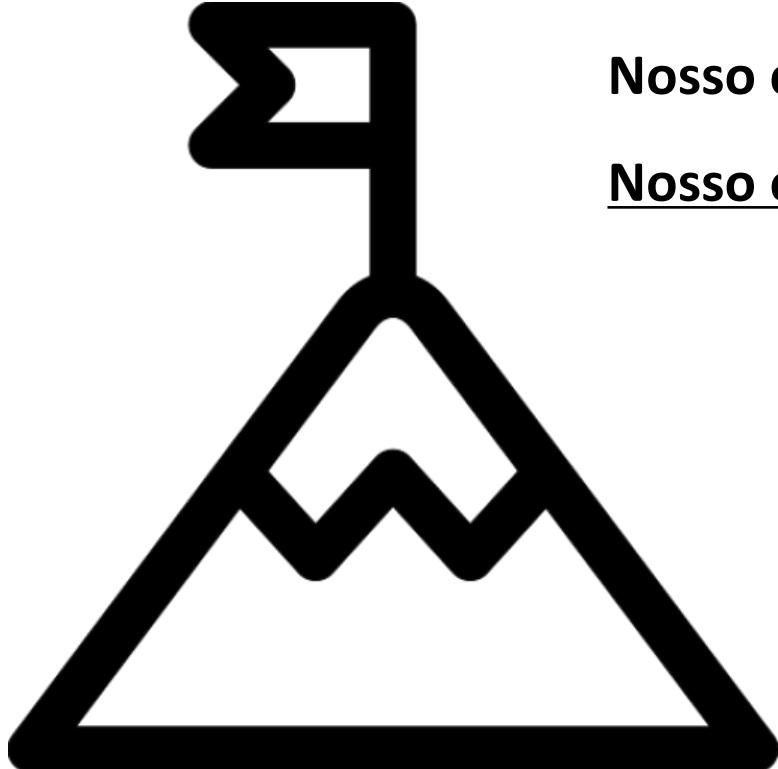
# Objetivo

Graduação área de Tecnologia da Informação → de 3 a 5 anos

Especializações em TI → de 1 a 2 anos

Nosso curso → 32h

Nosso objetivo neste tema?



# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

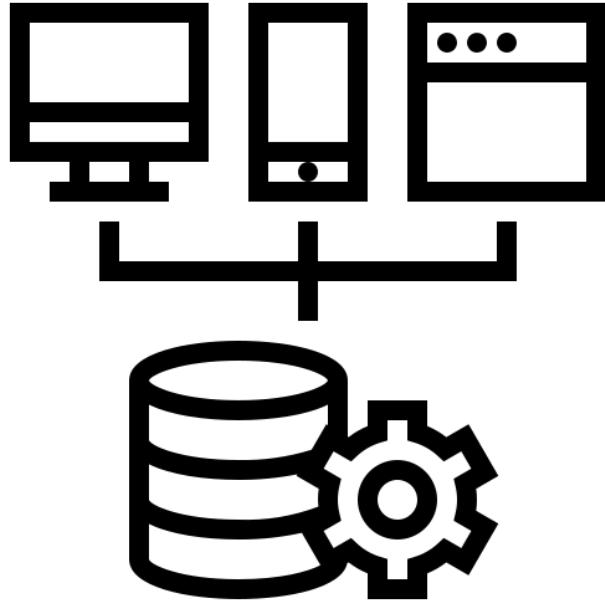
8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning

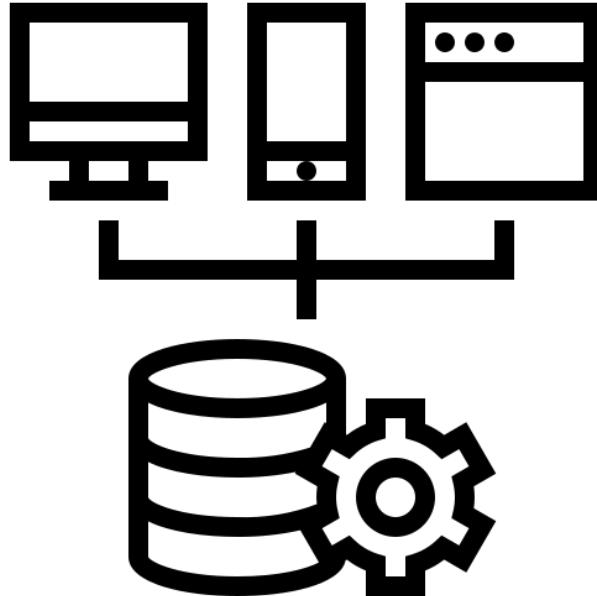


# Ciência da Computação



Para você, o que é a **CIÊNCIA DA COMPUTAÇÃO?**

# Ciência da Computação



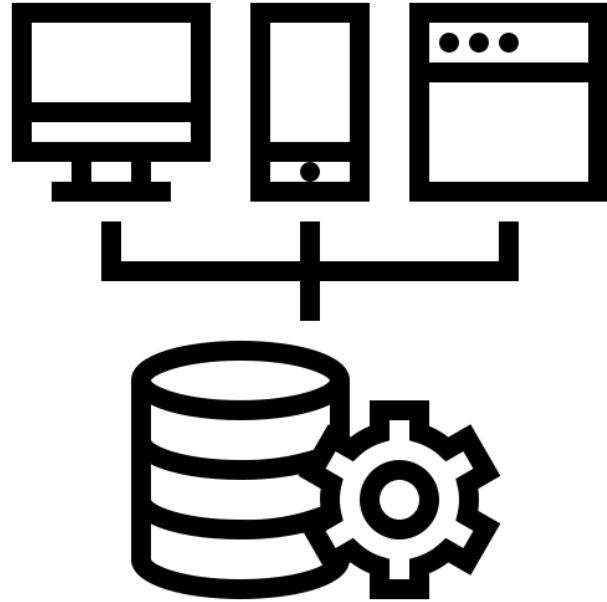
Para você, o que é a **CIÊNCIA DA COMPUTAÇÃO?**

*Ciência da computação tem tanto a ver com o computador como a Astronomia com o telescópio, a Biologia com o microscópio, ou a Química com os tubos de ensaio. A Ciência não estuda ferramentas, mas o que fazemos e o que descobrimos com elas.*

[Edsger Dijkstra](#), cientista de computação holandês.

[https://pt.wikipedia.org/wiki/Edsger\\_Dijkstra](https://pt.wikipedia.org/wiki/Edsger_Dijkstra)

# Ciência da Computação - Definição



Para você, o que é a **CIÊNCIA DA COMPUTAÇÃO?**

Utilizar instrumentos computacionais para, com métodos, automatizar e exponenciar a resolução de problemas de processamento de dados.

**Treinados para resolver problemas!**

# Ciência da Computação - História



Desde os primórdios até hoje em dia...

**Registro de dados** conforme os grupos sociais foram surgindo (“Matei 3 tigres este mês, e você?”).

A história da computação acompanha o crescimento das civilizações, onde se faziam necessárias técnicas e instrumentos para operações matemáticas (como o Ábaco criado há ~5.500 anos atrás).

**Mas o que é o DADO?**

# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

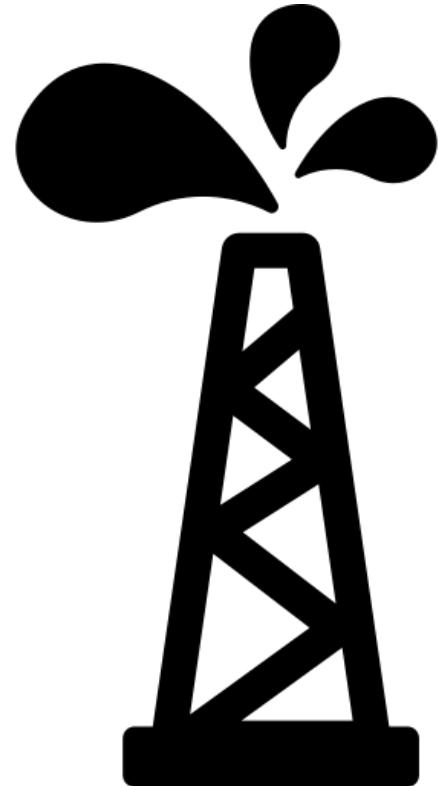
8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning

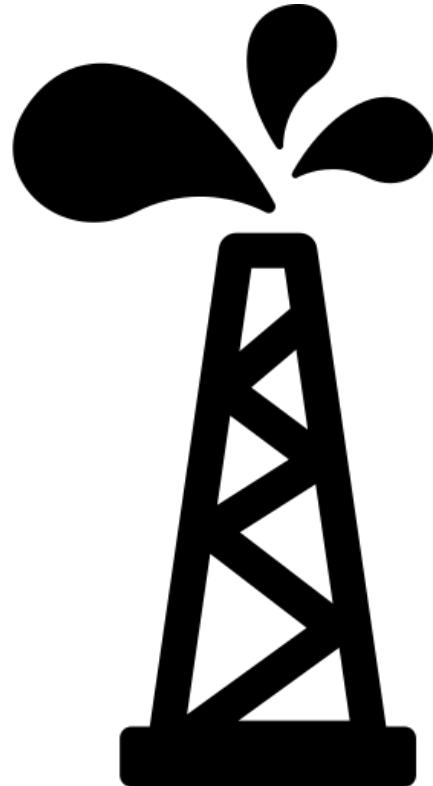


# Ciência da Computação – História dos Dados



O dado (no singular) é mesmo o novo petróleo?

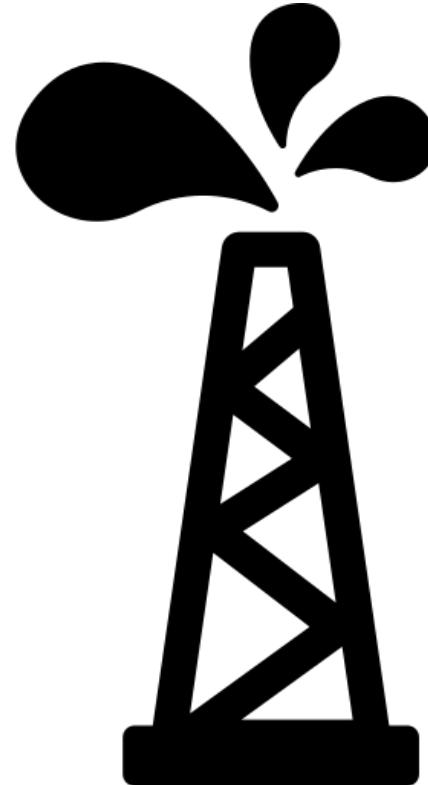
# Ciência da Computação – História dos Dados



O dado (no singular) é mesmo o novo petróleo?

**O dado é a matéria-prima. Sem ser “refinado” não gera valor.**

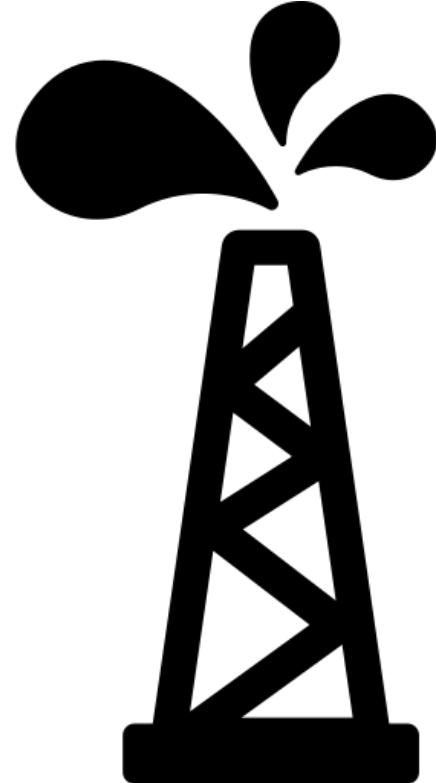
# Ciência da Computação – História dos Dados



Dados refinados, trabalhados geram **INFORMAÇÃO**

Em computação temos estudos referente a **teoria da informação**... ela vem estudando relatos de 30mil anos atrás, do homem primitivo, buscando se comunicar, expressar de forma a ser compreendido seus pensamentos internos. Todos estes “dados” utilizados por nós buscam, de forma encadeada, transmitir uma informação a respeito de um tema.

# Ciência da Computação – História dos Dados

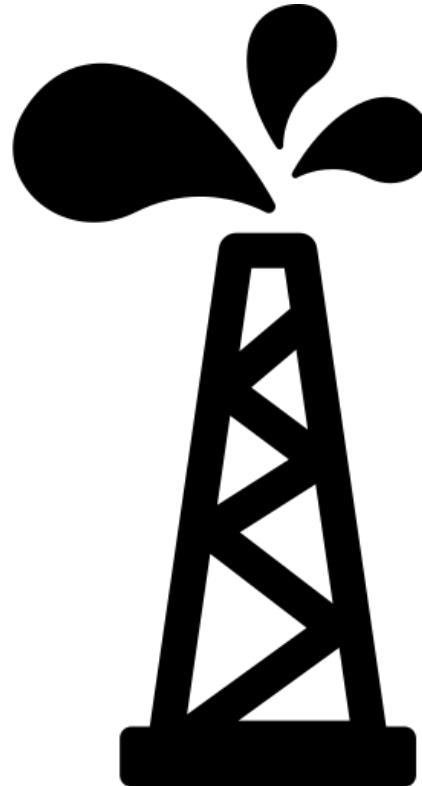


Porém a parte nobre dos Dados não é apenas se tornar informação, mas sim um conjunto de informações que gere **CONHECIMENTO**



Dados → Informação → Conhecimento

# Ciência da Computação – História dos Dados



## CONHECIMENTO É O NOVO ATIVO DE FATO

28/03/2019 - 15h51 - ATUALIZADA ÀS 15h51 - POR ÉPOCA NEGÓCIOS ONLINE

### McDonald's investe US\$ 300 milhões para adquirir startup de IA e big data

Primeira mudança após a compra da Dynamic Yield deverá ser vista nos drive thrus da rede de fast food

### Indústria 4.0 pode economizar R\$ 73 bilhões ao ano para o Brasil

Os ganhos de eficiência produtiva correspondem a uma economia de R\$ 31 bilhões

Fernando Rotta | 20/12/2017

Guerra comercial custou bilhões de dólares aos EUA e à China em 2018, diz economista

Disputa atingiu setores como automobilístico, tecnologia e, acima de tudo, agricultura

Reuters  
20/12/2018 - 13:05 / Atualizado em 29/12/2018 - 13:34



## Está chegando a era dos super-humanos. E eles são chineses

Tiago Cordeiro, especial para a Gazeta do Povo [22/03/2019] [11:09]

# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

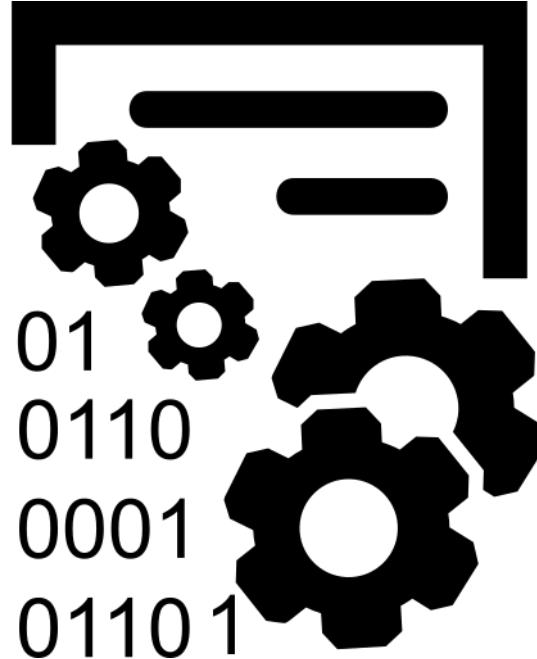
8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning

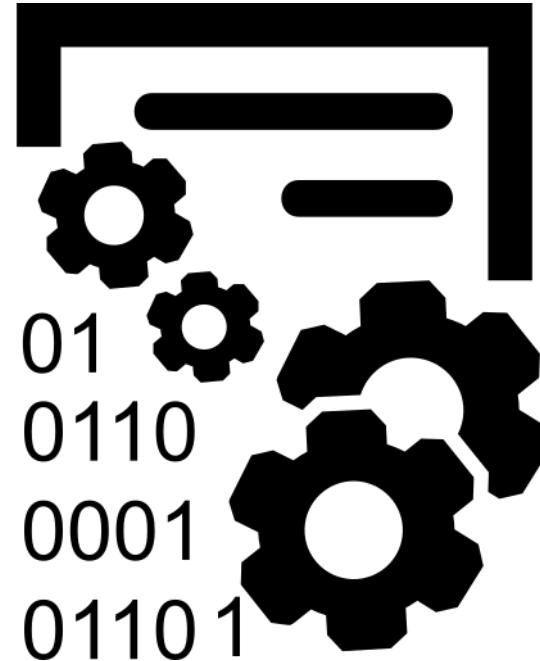


# Ciência da Computação – Sistemas Binários



E como o dado é refinado em um instrumento computacional?

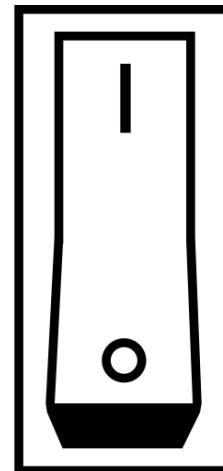
# Ciência da Computação – Sistemas Binários



E como o dado é refinado em um instrumento computacional?

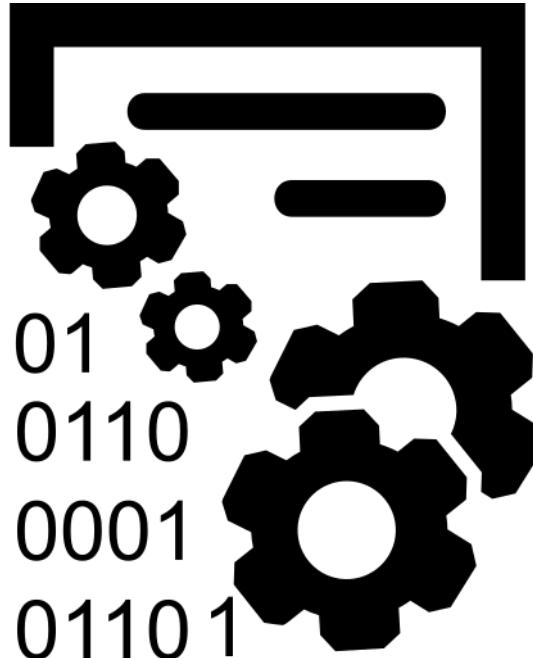
On/Off

Ligado/Desligado



Ou, na notação binária ZERO ou UM

# Ciência da Computação – Sistemas Binários



Através deste sistema binário temos representação de qualquer dado. O decimal 4 por exemplo representa-se por LIGADO, Desligado, Desligado:

Número Decimal	Representação Binária
0	0
1	1
2	10
3	11
4	100

# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

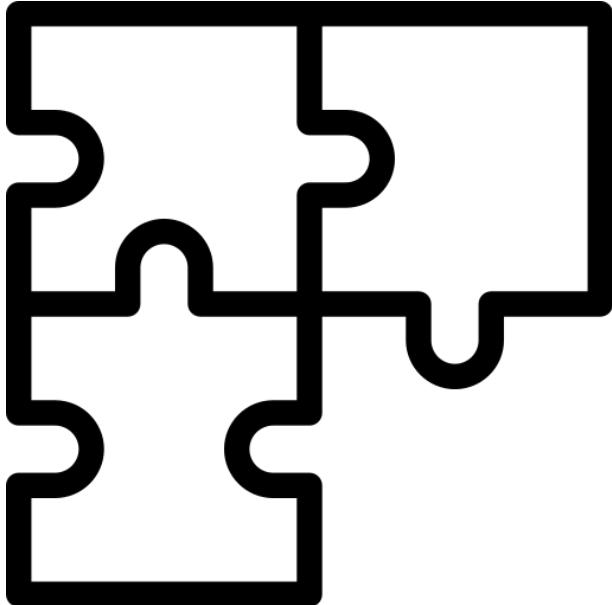
8 – Algoritmos

9 – Cluster e Cloud

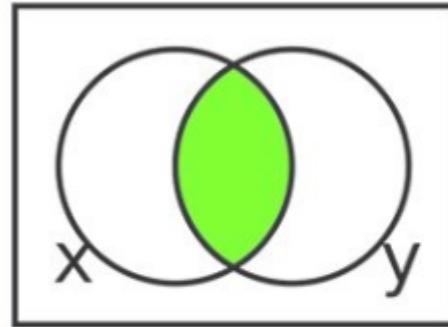
10 – Machine Learning



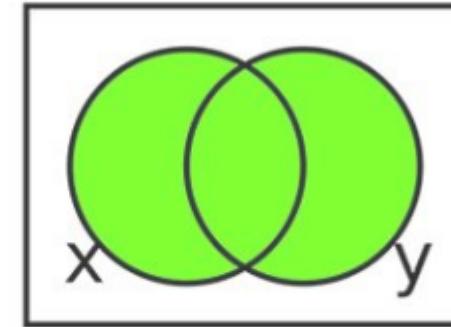
# Ciência da Computação – Lógica Booleana



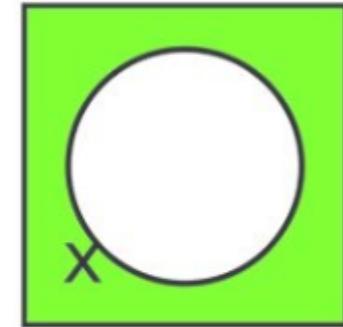
Criada em 1847 George Boole cria o conceito de lógica simbólica (base para sistemas computacionais)  
3 operadores principais {AND, OR, NOT}



$$x \wedge y$$



$$x \vee y$$



$$\neg x$$

# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

8 – Algoritmos

9 – Cluster e Cloud

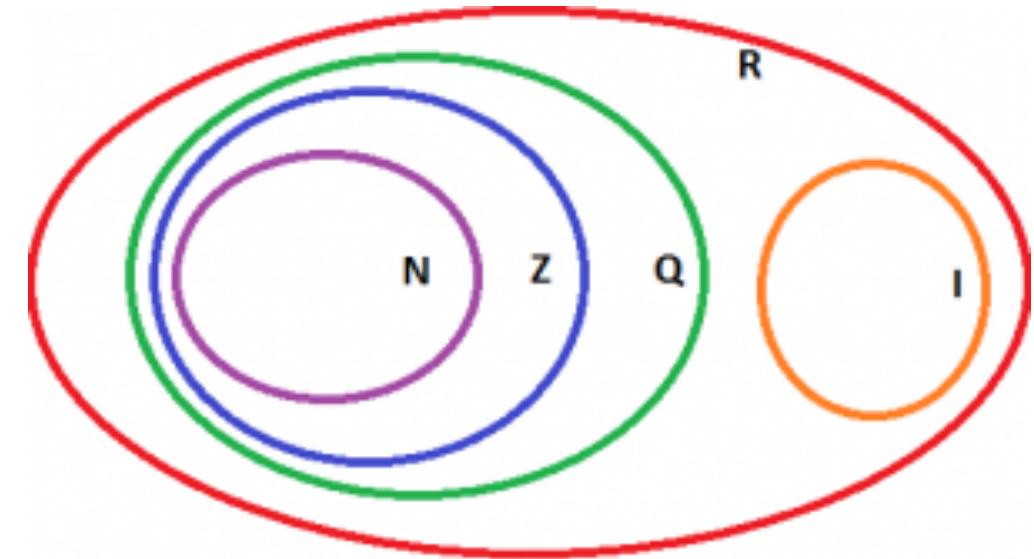
10 – Machine Learning



# Conjuntos & Matrizes

Os **Conjuntos** são coleções de elementos:

- Conjuntos dos Números Naturais
- Conjunto dos Números Inteiros
- Conjunto dos Números Racionais
- Conjunto dos Números Irracionais
- Conjunto dos Números Reais

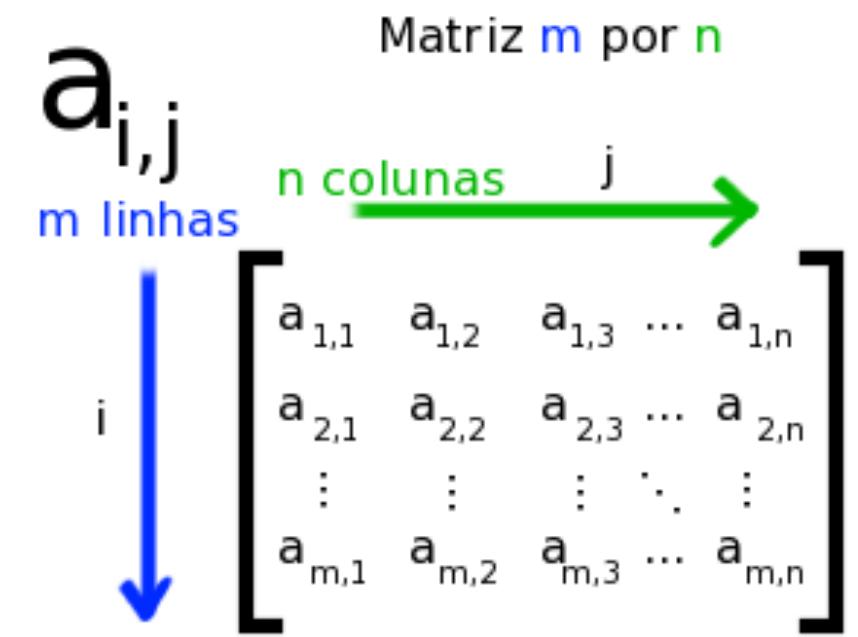


Conjunto Planetas Sistema Solar = {Mercúrio, Vênus, Terra, Marte, Júpiter, Saturno, Urano}

# Conjuntos & Matrizes

Matrizes são Valores apresentados em forma tabelas com linhas e/ou colunas:

- Representações de Matrizes
- Operações de Matrizes
- Matriz Inversa
- Sistemas de Equações - Resolução por Matrizes



# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

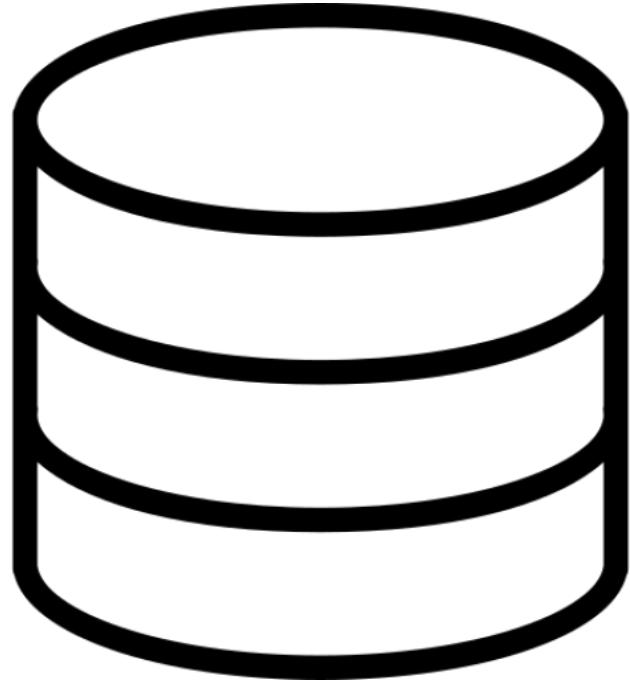
8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning



# Ciência da Computação – Banco de Dados



E tudo junto agora (teoria de conjuntos, sistemas booleanos, matrizes, teoria da informação)

Vossa Majestade o **Banco de Dados**

**Armazenar dados estruturados e não estruturados  
(conforme o SGDB, Sistemas Gerenciador de Banco de  
Dados)**

Veremos a frente no curso, em mais detalhes, o nosso BD (para os íntimos) utilizando o SGDB PostgreSQL

# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning



# Ciência da Computação – Algoritmos



A esposa do programador disse:

- Vá ao mercado e traga um litro de leite. Se tiver ovos, traga seis.

O programador voltou com seis litros de leite, então sua esposa disse:

- Programador! Porquê você trouxe seis litros de leite?

E o Programador respondeu:

- Eles tinham ovos...

# Ciência da Computação – Algoritmos

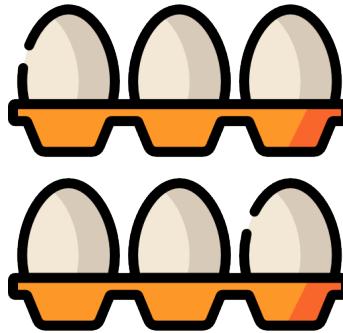


E um conjunto sequencial de passos para resolução de um problema. Utiliza os operados lógico booleanos e estrutura condicionais e de repetição.

Exemplo de algoritmo (ida do programador ao mercado):

```
Vá ao Mercado  
TemOvos <- Verdadeiro  
Leite <- 1  
Ovos <- 1  
Cesta <- []  
Cesta <- Leite  
SE TemOvos Então  
    Cesta <- Leite * 6
```

# Ciência da Computação – Algoritmos



Exemplo de algoritmo – Agora com pedido correto:

```
Vá ao Mercado  
TemOvos <- Verdadeiro  
Leite <- 1  
Ovos <- 1  
Cesta <- []  
Cesta <- Leite  
SE TemOvos Então  
    Para Contador De 1 até 6 Faça  
        Cesta <- Cesta + Ovos
```

# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning



# Ciência da Computação – Cluster e Cloud

Computação em Nuvem é a distribuição de serviços de computação – servidores, armazenamento, bancos de dados, redes, software, análises, inteligência e muito mais pela Internet (“a nuvem”), proporcionando inovações mais rápidas, recursos flexíveis e economia na escala



# Ciência da Computação

1 – Objetivo

2 – Ciência da Computação

3 – Dados

4 – Sistemas Binários

5 – Lógica Booleana

6 – Conjuntos & Matrizes

7 – Banco de Dados

8 – Algoritmos

9 – Cluster e Cloud

10 – Machine Learning



# Machine Learning

**Machine Learning:** ocorre quando um algoritmo ou um programa de computador é utilizado para melhorar as tarefas desempenhadas por humanos com base na experiência

## Supervisionado

**Regressão:** Prever os resultados ou tentativa de mapeamento de variáveis para função contínua

**Classificação:** Prever os resultados ou tentativa de mapeamento de variáveis de entrada em categorias distintas

## Não Supervisionado

Permite abordar problemas com pouca ou nenhuma ideia dos resultados

Esta abordagem é utilizada para agrupar os dados com base em relações entre eles, ou para detectar tendências. O desafio aqui é que não há feedback com base nos resultados da previsão

## Reforço

Maximizar o resultado baseado em um recompensa.

# Machine Learning

**Machine Learning:** ocorre quando um algoritmo ou um programa de computador é utilizado para melhorar as tarefas desempenhadas por humanos com base na experiência

## Supervisionado

**Regressão:** Prever os resultados ou tentativa de mapeamento de variáveis para função contínua

**Classificação:** Prever os resultados ou tentativa de mapeamento de variáveis de entrada em categorias distintas

## Não Supervisionado

Permite abordar problemas com pouca ou nenhuma ideia dos resultados

Esta abordagem é utilizada para agrupar os dados com base em relações entre eles, ou para detectar tendências. O desafio aqui é que não há feedback com base nos resultados da previsão

## Reforço

Maximizar o resultado baseado em um recompensa.

# Machine Learning

**Machine Learning:** ocorre quando um algoritmo ou um programa de computador é utilizado para melhorar as tarefas desempenhadas por humanos com base na experiência

## Supervisionado

**Regressão:** Prever os resultados ou tentativa de mapeamento de variáveis para função contínua

**Classificação:** Prever os resultados ou tentativa de mapeamento de variáveis de entrada em categorias distintas

## Não Supervisionado

Permite abordar problemas com pouca ou nenhuma ideia dos resultados

Esta abordagem é utilizada para agrupar os dados com base em relações entre eles, ou para detectar tendências. O desafio aqui é que não há feedback com base nos resultados da previsão

## Reforço

Maximizar o resultado baseado em um recompensa.

# Machine Learning

**Machine Learning:** ocorre quando um algoritmo ou um programa de computador é utilizado para melhorar as tarefas desempenhadas por humanos com base na experiência

## Supervisionado

**Regressão:** Prever os resultados ou tentativa de mapeamento de variáveis para função contínua

**Classificação:** Prever os resultados ou tentativa de mapeamento de variáveis de entrada em categorias distintas

## Não Supervisionado

Permite abordar problemas com pouca ou nenhuma ideia dos resultados

Esta abordagem é utilizada para agrupar os dados com base em relações entre eles, ou para detectar tendências. O desafio aqui é que não há feedback com base nos resultados da previsão

## Reforço

Maximizar o resultado baseado em um recompensa.

# Machine Learning

**Machine Learning:** ocorre quando um algoritmo ou um programa de computador é utilizado para melhorar as tarefas desempenhadas por humanos com base na experiência

## Supervisionado

**Regressão:** Prever os resultados ou tentativa de mapeamento de variáveis para função contínua

**Classificação:** Prever os resultados ou tentativa de mapeamento de variáveis de entrada em categorias distintas

## Não Supervisionado

Permite abordar problemas com pouca ou nenhuma ideia dos resultados

Esta abordagem é utilizada para agrupar os dados com base em relações entre eles, ou para detectar tendências. O desafio aqui é que não há feedback com base nos resultados da previsão

## Reforço

Maximizar o resultado baseado em um recompensa.

# Estudo de Caso

MadeiraMadeira - O Maior Hom X

https://www.madeiramadeira.com.br

madeiramadeira

Busque aqui tudo para sua casa...

Quer ajuda? Olá, entre ou cadastre-se

Departamentos Por Ambiente Móveis Kits & Conjuntos Frete Grátis SP Outlet Veja mais

fechamês Julho :)

com até 70% OFF + 12x sem juros

SALA de ESTAR

A partir de R\$ 97,75

CONFIRA AQUI

Podemos Ajudar?

ESSA É A SUA CHANCE!

Descontos imperdíveis com hora pra acabar!

01 DIAS 22 HORAS 34 MINUTOS 16 SEGUNDOS

CONFIRA

37% OFF 44% OFF 43% OFF 26% OFF 31% OFF

Chat

The screenshot shows the homepage of the MadeiraMadeira website. At the top, there's a navigation bar with links for 'Departamentos', 'Por Ambiente', 'Móveis', 'Kits & Conjuntos', 'Frete Grátis SP', 'Outlet' (which is highlighted in red), and 'Veja mais'. Below the navigation is a large blue promotional banner for a 'fechamês' (month-end) sale in July. The banner features two images of living room furniture: a white TV unit with a black chair and a dark wood TV unit with a shelf above it. The text on the banner includes 'com até 70% OFF + 12x sem juros' (with up to 70% off + 12x interest-free) and 'SALA de ESTAR A partir de R\$ 97,75' (Living Room starting from R\$ 97,75). Below the banner is a call-to-action button labeled 'CONFIRA AQUI'. To the left of the banner, there's a vertical blue sidebar with the text 'Podemos Ajudar?'. At the bottom of the page, there's a light gray box containing the text 'ESSA É A SUA CHANCE!' (This is your chance!) and 'Descontos imperdíveis com hora pra acabar!' (Irresistible discounts with time to end!). It also features a timer showing '01 DIAS 22 HORAS 34 MINUTOS 16 SEGUNDOS' (1 day 22 hours 34 minutes 16 seconds) and a 'CONFIRA' button with a downward arrow. Below the timer are five small images of different furniture pieces with their respective discount percentages: 37% OFF, 44% OFF, 43% OFF, 26% OFF, and 31% OFF. On the right side of the page, there's a blue 'Chat' button. The overall design is clean and modern, with a focus on promoting discounts and encouraging immediate action.

# Pauta

**1 – Material do Curso**

**2 – Virtual Machine do  
Cientista de Dados**



# Pauta

1 – Material do Curso

2 – Virtual Machine do  
Cientista de Dados



# Material do Curso



## Best Practice → Versionamento

Utilização de repositório (geralmente na cloud) para sincronizar e criar versões de arquivo.

É uma boa prática quando estamos gerando scripts / códigos fontes.

Soluções mais difundidas utilizadas:

SVN → <https://subversion.apache.org/>

GiT → <https://git-scm.com/>

**Nosso repositório! → <https://github.com/datasiencealdeia/201909.git>**

**Apresentações, Scripts, Dados Utilizados e etc, sempre atualizados**

# Material do Curso



## Principais Comandos GiT

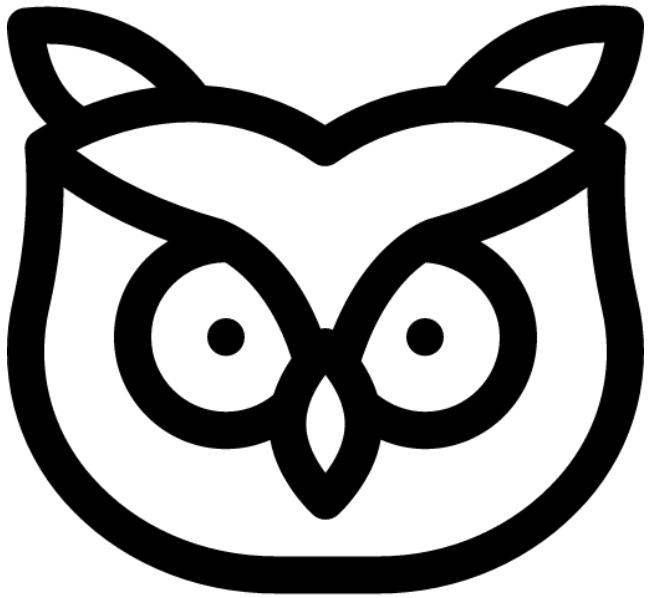
<https://woliveiras.com.br/posts/comandos-mais-utilizados-no-git/>

O que usaremos no curso:

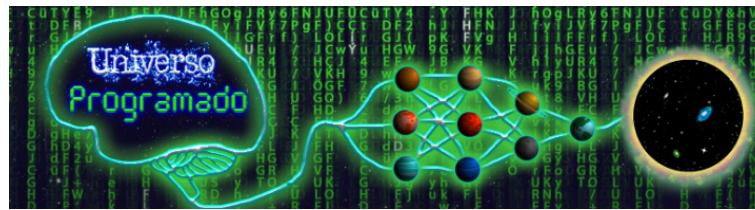
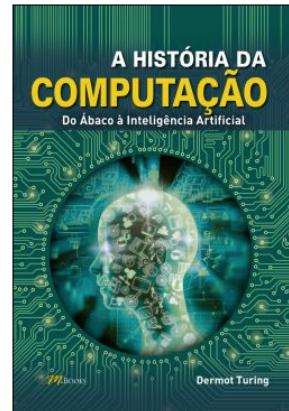
*git pull* (baixar versão atualizada do repositório para vm)

*git status* (status do repositório local da vm)

# Quero Saber Mais...



Ciência e Dados | Data Science for Professionals



**INTRODUÇÃO A VERSIONAMENTO  
DE CÓDIGO E CONHECENDO O GIT**



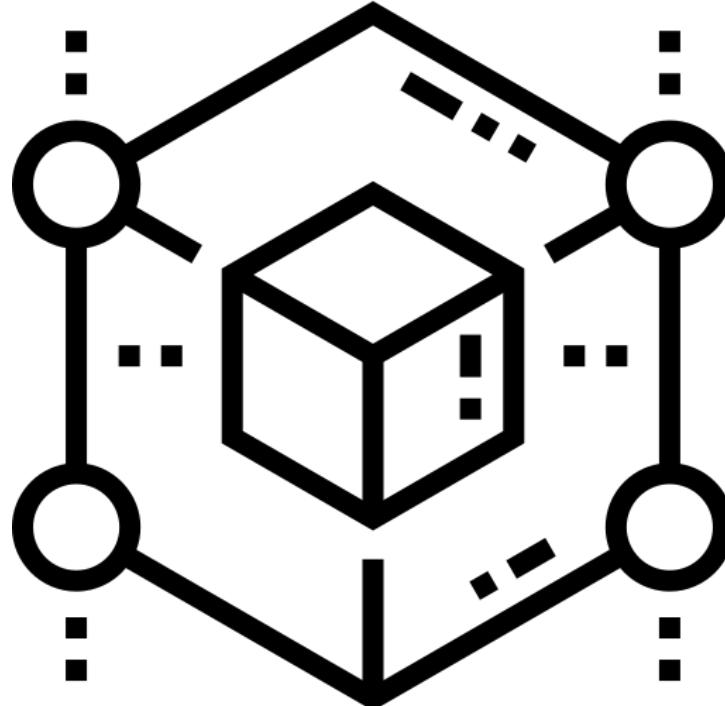
# Pauta

**1 – Material do Curso**

**2 – Virtual Machine do  
Cientista de Dados**



# Virtual Machine



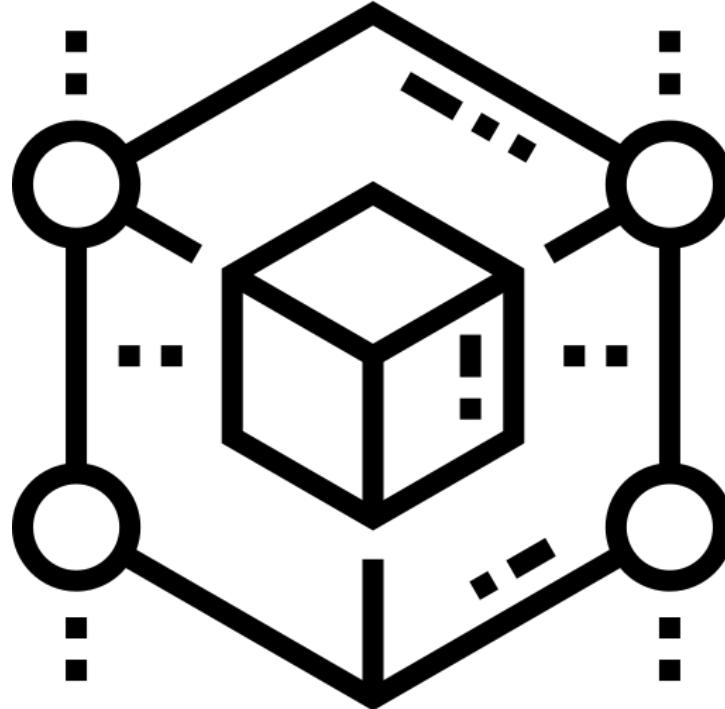
Conceito da Virtualização de Ciência da Computação

Software que simula um computador (máquina virtual) utilizando recursos do computador que está instalado (máquina host)

Pode ter configuração dimensionada com simples configuração: Memória RAM, HD, Placa de Rede e etc. (Sempre limitada a configuração física do computador host)

**Utilizado na prática nas máquinas cluster/cloud**

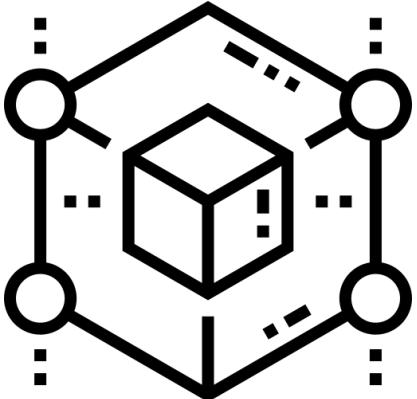
# Virtual Machine



## Configuração da Nossa Virtual Machine (VM)

- Sistema Operacional: Ubuntu 16.04.02 LTS 32 Bits
- 4 Gb de Memória RAM
- 50 Gb de HDD
- Softwares Embarcados:
  - ETL → Pentaho 5.0.1
  - Banco de Dados → PostgreSQL 10.7
  - Linguagens → R e Python 3
  - IDE Versionamento → GitEye 2.2

# Virtual Machine



**1) Download do Virtual Box (versão 6.0.4)**

<https://www.virtualbox.org/wiki/Downloads>

**2) Download da nossa VM (~7 Gb)**

[https://drive.google.com/open?id=1OVr0BVaok5cfAKJ8XinzFY0dF-W\\_RpQK](https://drive.google.com/open?id=1OVr0BVaok5cfAKJ8XinzFY0dF-W_RpQK)

**3) Executar o Virtual Box**

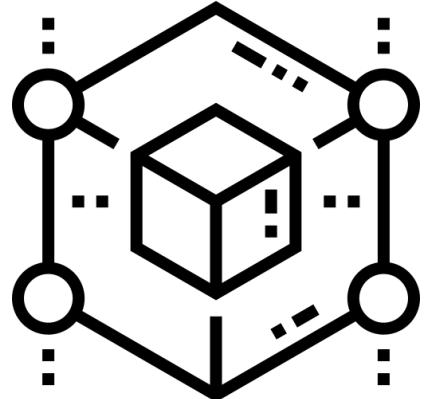
Opção Arquivo > Importar Appliance > Escolher o arquivo baixado anteriormente para importar (DataScience.ova)

user: ds

passwd: ds2019Xpto

**IMPORTANTE!!! A VM será a base para os exercícios das próximas aulas**

# Virtual Machine



Dúvidas?

Momento Help Desk VM

# Obrigado !

 Charles Adriano dos Santos  
 charles.a.santos@caelis.it  
 chadri  
 41 99144 6663

 Rafael Roberto Dias  
 rafael.dias@madeiramadeira.com.br  
 rafael-roberto-dias-00b39123  
 41 99672 7170