

Minicurso

Ciência dos dados na prática usando Python

IFCE Campus Quixadá - 22 de maio de 2019

Prof. Regis Pires Magalhães (regis@insightlab.ufc.br)

www.insightlab.ufc.br

bit.ly/insightface

Sobre mim

- ▶ Ciência da Computação (UECE)
- ▶ Tecnologias Web (UFPI)
- ▶ IFPI
- ▶ Bancos de Dados (UFC)
- ▶ Ciência de Dados (UFC / CNR-Itália)

▶ Educação

▶ Aprender Fazendo

- ◆ Escoteiros
- ◆ Coding DOJO (IFPI)
- ◆ Grupo de Estudos da Maratona de Programação (UFC)
- ◆ Núcleo de Práticas em Informática (UFC)
- ◆ Grupo de Práticas em Ciência de Dados (UFC)
- ◆ Insight Lab (UFC)



Somos Vizinhos!

Campus da UFC em Quixadá



Visitas ao Insight Lab



PRINCIPAIS ÁREAS DE PESQUISA



Ciência de Dados

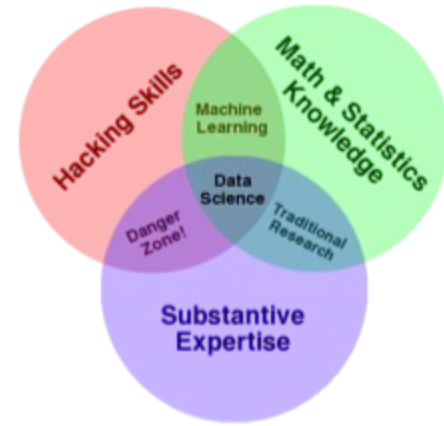
- ▶ É a transformação de dados usando matemática e estatística em insights, decisões e produtos valiosos [Foreman, 2013].
- ▶ É usada nas organizações para ajudar a melhorar seu funcionamento e a criar valor.



Foreman, John W. Data Smart: Using Data Science to Transform Information into Insight. Wiley. 1ª Ed., 2013.

Ciência de Dados

- ▶ O uso de técnicas de análise de dados e aprendizagem de máquina podem levar ao uso muito mais eficiente de recursos.



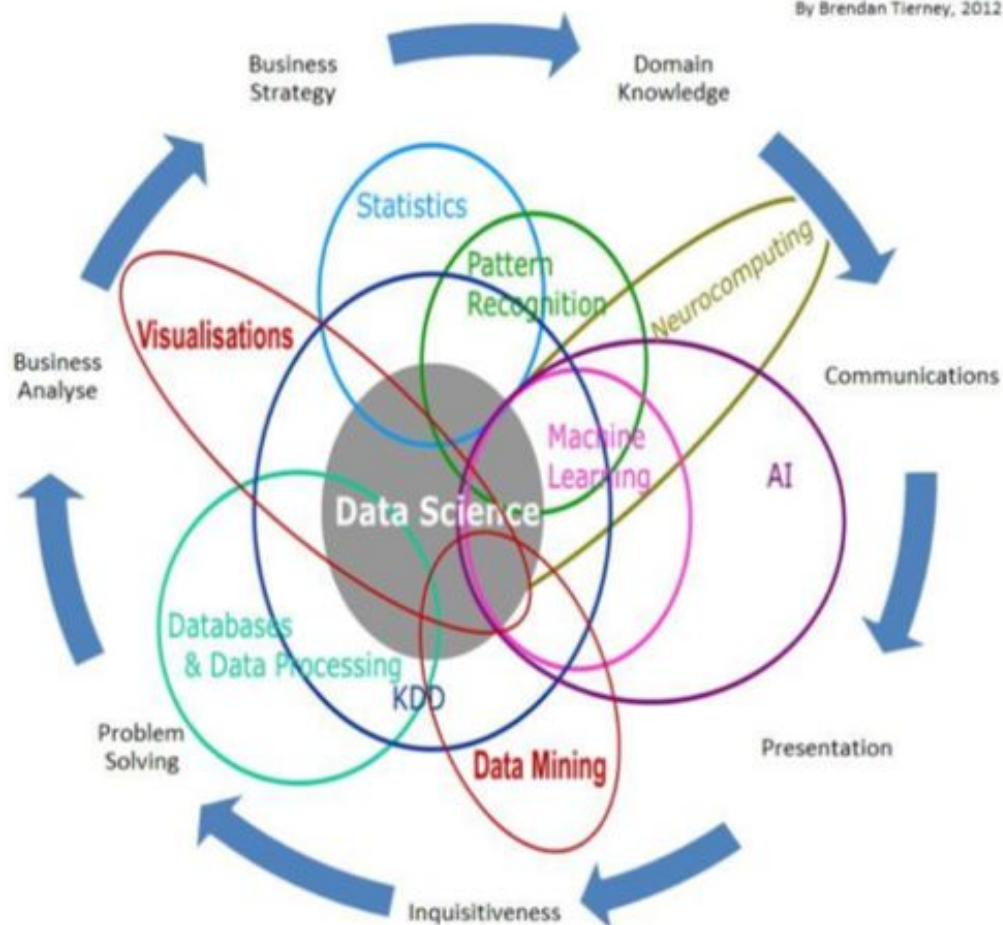
Motivações

- ▶ Barateamento e popularização dos dispositivos móveis e dos mais diversos tipos de sensores para coleta de dados.
- ▶ Aumento do poder de processamento e armazenamento dos computadores.
- ▶ Otimização de recursos em um mercado global cada vez mais competitivo.
- ▶ Amadurecimento, melhoria e disseminação de técnicas de aprendizagem de máquina.



Data Science Is Multidisciplinary

By Brendan Tierney, 2012

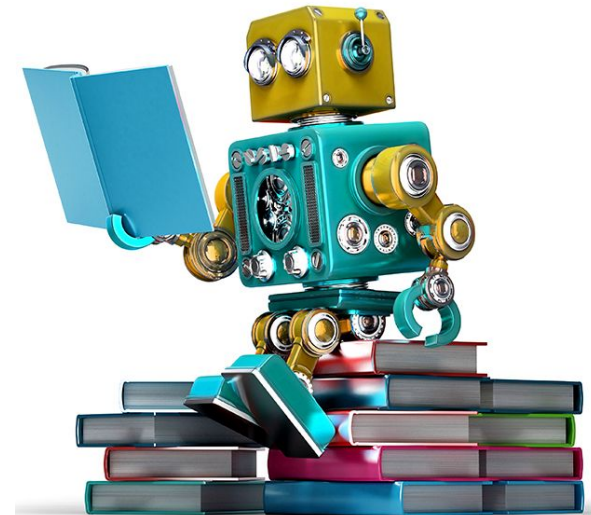


The background image is a blurred office space. In the foreground, there's a desk with a computer monitor, a red desk lamp, a black desk lamp, a small potted plant, and a black cat figurine. In the background, there are more desks with computers and a large window on the right side. The overall tone is professional and modern.

E na prática como a ciência de dados e usada mesmo?

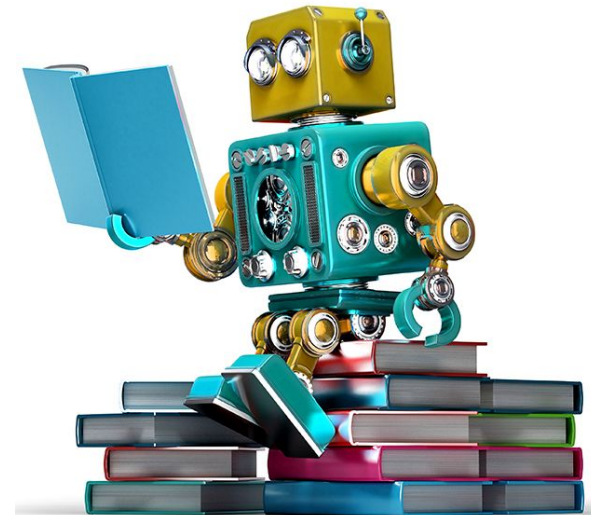
Exemplos de uso de técnicas de aprendizado de máquina

- ▷ Risco de empréstimo / seguro / plano de saúde
- ▷ Preços de imóveis / carros
- ▷ Produtos com maior chance de venda
- ▷ Sistemas de recomendação
- ▷ Evasão de cursos
- ▷ Probabilidade de desenvolver uma doença
- ▷ Sistema de busca
- ▷ Sistemas de segurança
- ▷ Etc.



Aplicações na engenharia

- ▷ Melhorias no gerenciamento de processos;
- ▷ Predição de eventos;
- ▷ Monitoramento e análise de saúde das construções;
- ▷ Análises de impactos ambientais;
- ▷ Projeto de melhores sistemas de transporte;
- ▷ Análise de desempenho de materiais e dispositivos;
- ▷ Melhoria de processos;
- ▷ ...





<https://serenatadeamor.org/>

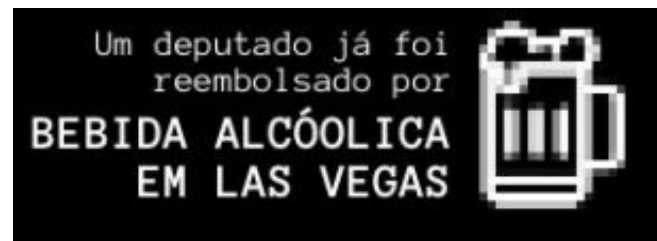
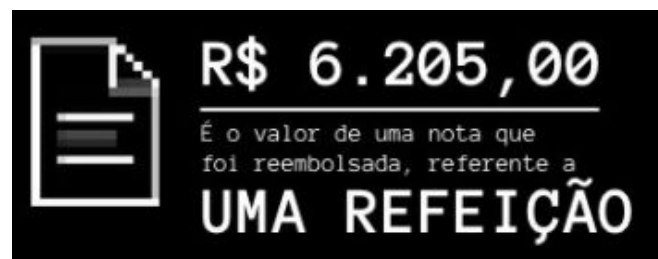
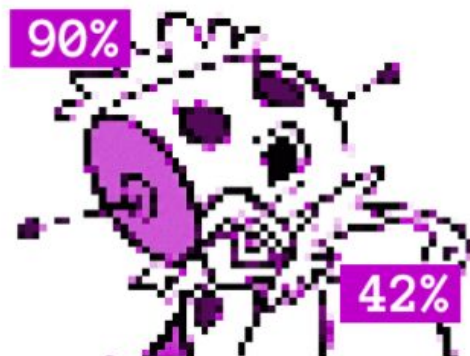
<https://medium.com/data-science-brigade>

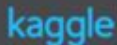

Brazilian group develops an AI to help in public expenditures monitoring. Rosie, the robot's name, found more than 8.000 suspicious reimbursements from Brazilian congresspeople.


Despite all the corruption related news coming from Brazil, there is a movement for transparency in the country. Several bills signed in the last few years put Brazil in the top of transparency rankings worldwide, specially when our former president Dilma Rousseff signed in 2011 the Access of Information Law, a Brazilian version for the american FOIA (Freedom of Information Act), which completed 50 years in 2016. It makes open data compulsory for all public bodies. Something similar has happened in some




A Operação Serenata de Amor criou a **Rosie** – uma inteligência artificial capaz de analisar cada pedido de reembolso dos deputados e identificar a probabilidade de ilegalidade.




 Search kaggle 

Competitions Datasets Kernels Discussion Jobs 

 Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

 Kaggle · 6,174 teams · 3 years to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [More](#) [Submit Predictions](#)

[Overview](#)

[Description](#) [Evaluation](#) [Frequently Asked Questions](#) [Tutorials](#)


Start here if...

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.



Competition Description


The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.



[Competitions](#)
[Datasets](#)
[Kernels](#)
[Discussion](#)
[Learn](#)
...







Giba

Lead Data Scientist at Ople.ai
Curitiba, State of Paraná, Brazil
Joined 7 years ago · last seen in the past day


Followers 5971
Following 23



**Competitions
Grandmaster**

[Home](#)
[Competitions \(138\)](#)
[Kernels \(57\)](#)
[Discussion \(627\)](#)
[Datasets \(1\)](#)
[Followers \(5,971\)](#)
[Contact User](#)
[Follow User](#)

Competitions Grandmaster




Current Rank	Highest Rank
2 of 109,619	1

45	32	25
----	----	----

- Santander Value Prediction...** **1st**
 9 months ago · Top 1%
of 4484
- Melbourne University AES/...** **1st**
 2 years ago · Top 1%
of 478
- Western Australia Rental Pr...** **1st**
 3 years ago · Top 2%
of 59

Kernels Master




Current Rank	Highest Rank
42 of 92,612	28

4	7	8
---	---	---

- The Property by Giba** **375** votes
 10 months ago
- kernel10cd7d598b** **87** votes
 a month ago
- Giba CountVectorizer :-D** **73** votes
 10 months ago

Discussion Master



Current Rank	Highest Rank
7 of 94,081	6

62	77	305
----	----	-----

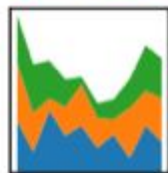
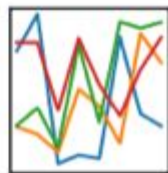
- Data Scientist Hero** **418** votes
 3 years ago
- 1st PLACE - WINNER SOL...** **370** votes
 4 years ago
- The Data "Property"** **291** votes
 10 months ago

Ferramentas de Trabalho



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

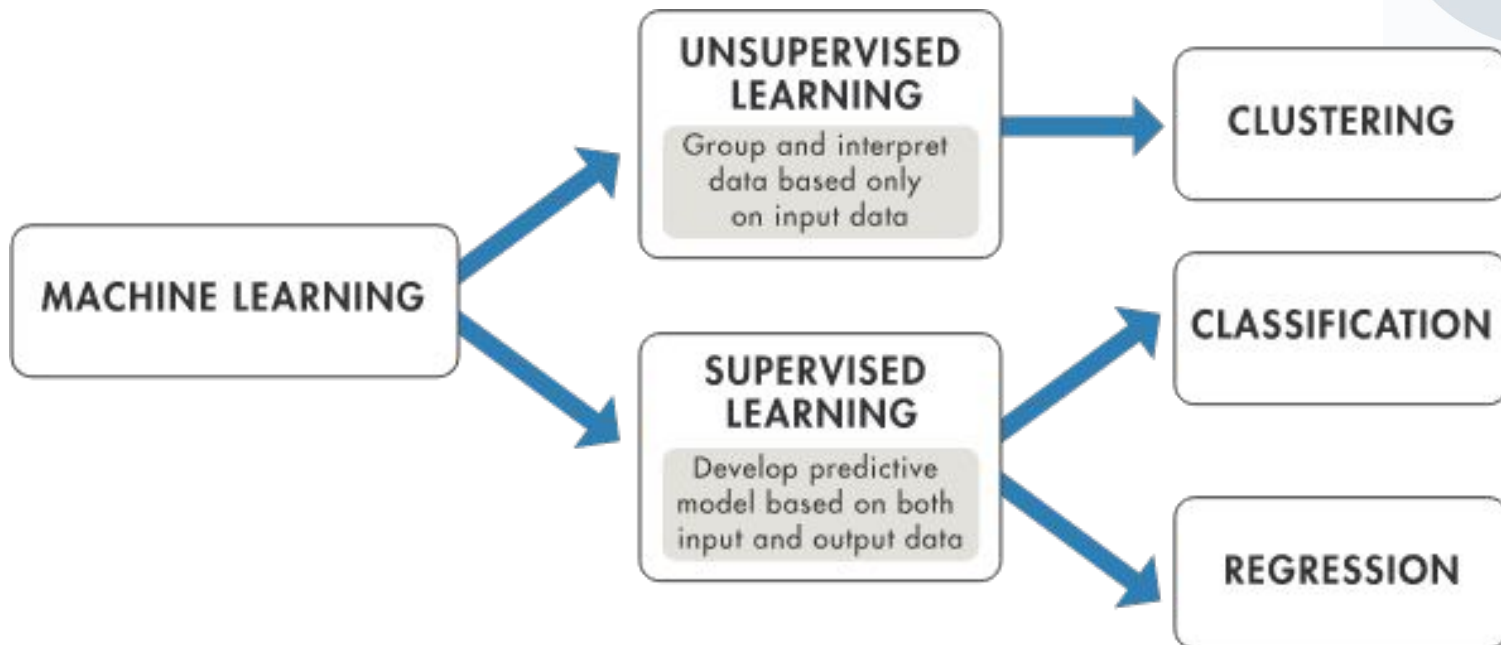
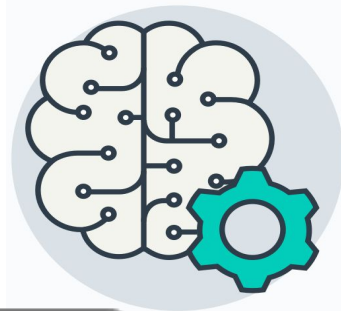


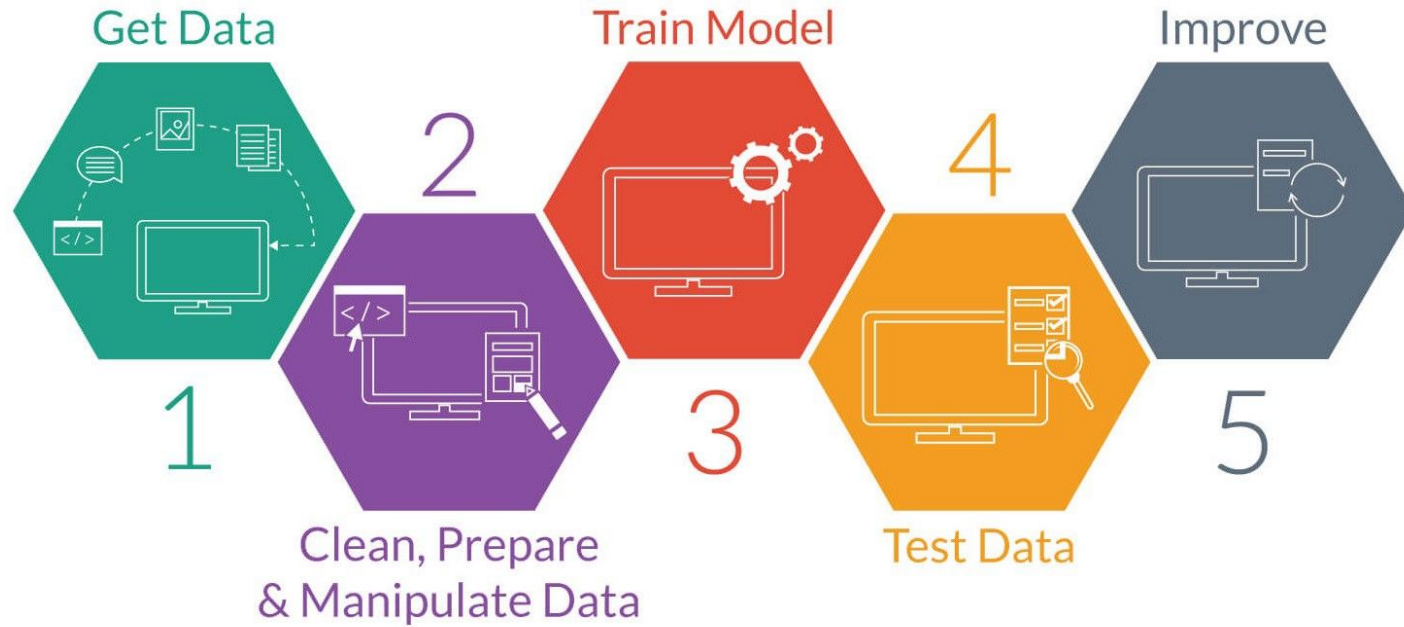
matplotlib



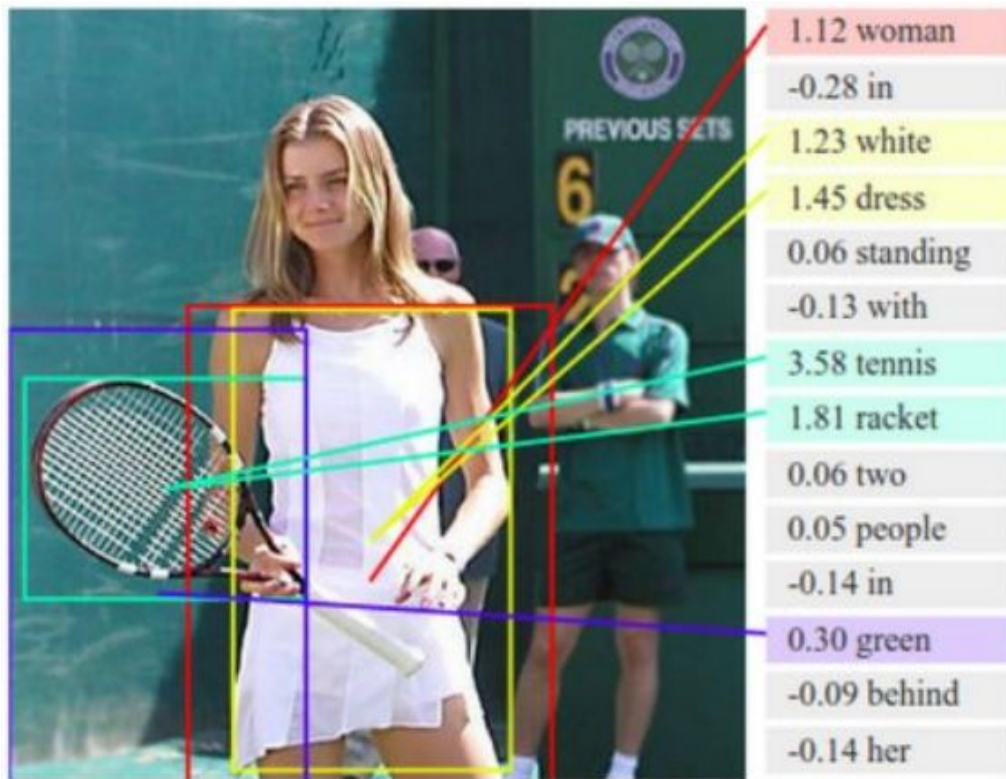
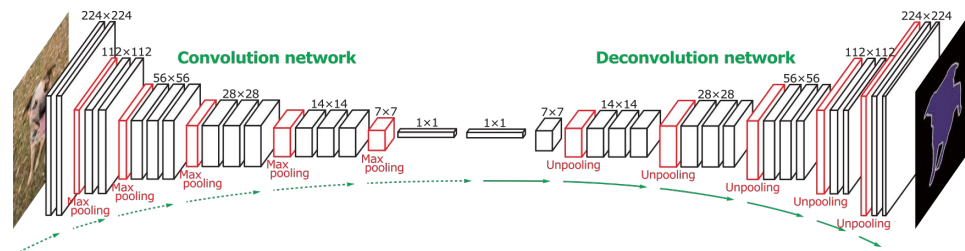
machine learning in Python







Deep Learning



Image,
automatically
annotated by Deep
Learning.

Ferramentas colaborativas online

- Google Colab

- <http://colab.research.google.com>



The screenshot displays the Google Colaboratory web interface. At the top, there's a header with the Colab logo and the text 'Hello, Colaboratory'. Below this is a menu bar with options: File, Edit, View, Insert, Runtime, Tools, and Help. A secondary bar contains icons for adding code, text, cells, and options to copy to drive or discard changes. On the right, it shows 'CONNECTED' status and an 'EDITING' mode indicator. The main content area features a 'Welcome to Colaboratory!' message, explaining that it's a Jupyter notebook environment in the cloud. It mentions that notebooks are stored on Google Drive and are free to use. Below the text, there's a code cell with a single line: `print 'Hello, Colaboratory!'`. The output of this cell is 'Hello, Colaboratory!'. Further down, a text block explains that Colab allows executing TensorFlow code in the browser. It provides an example of adding two matrices, showing the mathematical equation:
$$\begin{bmatrix} 1. & 1. & 1. \\ 1. & 1. & 1. \end{bmatrix} + \begin{bmatrix} 1. & 2. & 3. \\ 4. & 5. & 6. \end{bmatrix} = \begin{bmatrix} 2. & 3. & 4. \\ 5. & 6. & 7. \end{bmatrix}$$
. At the bottom, another code cell is shown with Python code using TensorFlow and NumPy to perform this matrix addition:

```
[ ] import tensorflow as tf
import numpy as np

with tf.Session():
    input1 = tf.constant(1.0, shape=[2, 3])
    input2 = tf.constant(np.reshape(np.arange(1.0, 7.0, dtype=np.float32), (2, 3)))
    output = tf.add(input1, input2)
    result = output.eval()
```

Projeto Ciência de Dados na Prática (cdp)

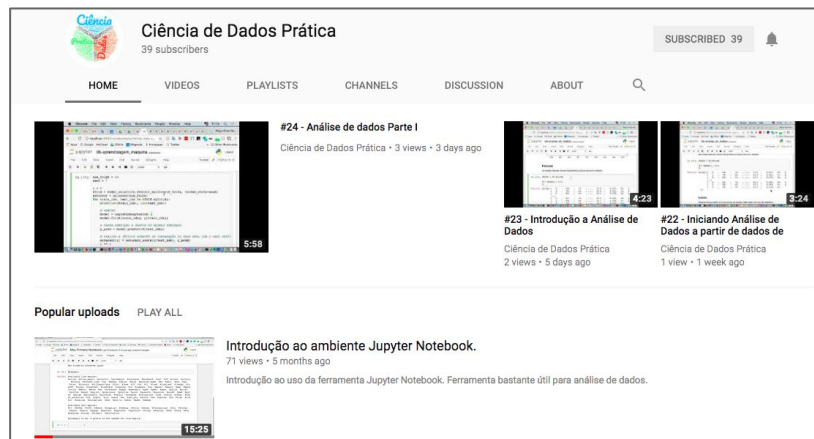
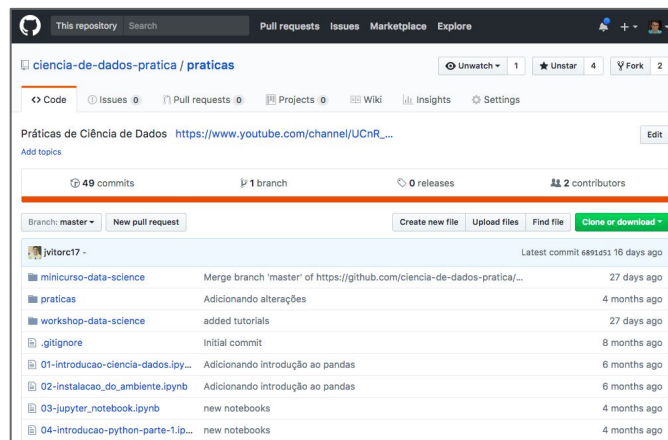
Material didático gratuito

- **Facebook (notícias)**
 - <http://bit.ly/cdpface>
- **Dicas e Referências**
 - <http://bit.ly/cdprefs>
- **GitHub (código)**
 - <http://bit.ly/cdpgithub>
- **YouTube (vídeos)**
 - <http://bit.ly/cdpvideos>

Encontros presenciais:

Quartas 12:30 às 13:30

UFC Quixadá - Sala 1 - Bloco 4



Agenda

1. Introdução (estamos aqui)
2. Instalando tudo que precisamos com **Anaconda** e interagindo com **Jupyter**
3. Programação fácil e produtiva com **Python**
4. Matemática rápida e eficiente com **Numpy**
5. Visualização com **Matplotlib**
6. Manipulação de dados com **Pandas**
7. Aprendizado de Máquina com **Scikit Learn**
8. Conclusões

OBRIGADO!

Dúvidas?

Você pode nos encontrar em:

E-mail: **regis@insightlab.ufc.br**

Site: **www.insightlab.ufc.br**

Facebook: **bit.ly/insightface**

