

Cap.1 - Conceitos e Ferramentas Fundamentais



Prof. MSc. Renzo P. Mesquita

Objetivos

- Compreender o que é o Big Data;
- Entender o que é o campo da Ciência de Dados (Data Science) e qual o papel da Análise de Dados dentro dela;
- Entender como a linguagem de programação Python pode ajudar neste processo de forma mais profissional e simplificada.



Capítulo 1

Conceitos e Ferramentas Fundamentais

- 1.1. O que é o Big Data?;*
- 1.2. Conceito de Ciência de Dados (Data Science);*
- 1.3. Casos de uso da Ciência de Dados;*
- 1.4. Variedade de Dados (Data Varieties);*
- 1.5. Análise de Dados;*
- 1.6. Formatos de Arquivos de Dados (Data Formats);*
- 1.7. Fontes de Dados (Data Sources);*
- 1.8. Por que usar Python?*



1.1. O que é o Big Data?

Neste exato momento, o mundo está sendo inundado por novos dados oriundos de:

Computadores pessoais

Smartphones

Câmeras

Sensores de TODOS os tipos

Navegação na Internet

Compras online

Vídeos

Redes Sociais

Imagens

Pesquisa em mecanismos de buscas

Wearables

**E INFINITAS
OUTRAS
FONTES..**

- Tsunamis de dados estruturados, não estruturados ou semiestruturados estão sendo produzidos por atividades que acontecem tanto no mundo real como virtual;
- **Bem-vindo ao mundo do BIG DATA!**



Prezi

1.2. Conceito de Ciência de Dados (Data Science)

Porém, mesmo com a geração desta grande quantidade de dados, muitos devem estar se perguntando:

*Mas e aí? Qual o objetivo de todos estes dados?
Qual a razão de gerá-los e guardá-los?*

Já houve tempos que isso era uma preocupação, ou seja, onde guardar tudo isso e para que?

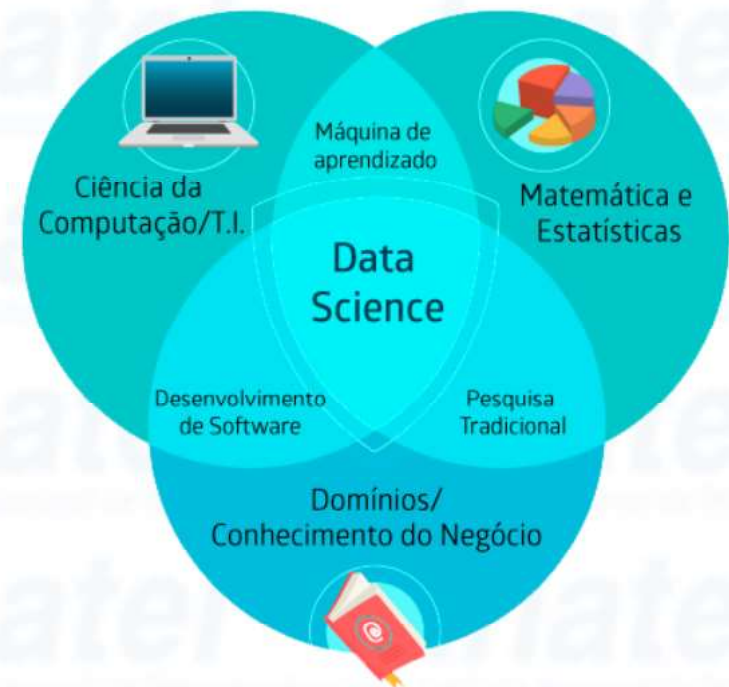
- Porém, nos últimos tempos o cenário reverteu completamente;
- Dados, de todas as quantidades e espécies, têm se tornado o "novo petróleo do mundo e combustível para o futuro";
- Conhecimentos sólidos em áreas específicas, matemática e computação têm dado origem aos chamados Cientistas de Dados (ou Data Scientists), capazes de extrair insights (conclusões) valiosos dos dados com o objetivo de melhorar o andamento de negócios, investimentos, da saúde, da qualidade de vida das pessoas e de infinitas outras outras possibilidades.



1.2. Conceito de Ciência de Dados (Data Science)

A Ciência de Dados (Data Science) é o domínio científico dedicado à descoberta do conhecimento por meio da Análise de Dados, podendo se criar Modelos e Algoritmos capazes de encontrarem padrões nestes dados e até "prever o futuro" se baseando neles.

Tomar uma boa decisão nem sempre é algo fácil, por isso, a Ciência de Dados é uma área de estudo multidisciplinar, que engloba conhecimentos de **Negócios**, **Matemática e Computação**.



1.3. Casos de Uso da Ciência de Dados

"Bradesco cria sistema antifraudes analisando logs gerados por sensores em caixas eletrônicos". Como resultado, conseguiu reduzir de 10 mil para 5 o número de incidentes diários;



"Airbnb se torna maior empresa hoteleira da atualidade, mas sem possuir nenhum hotel". Grande parte disto se deve à utilização de um modelo completamente orientado a dados (data-driven) para tomada de decisões;

"Nike, a gigante dos artigos esportivos, adquiriu a empresa Celect" para reunir e tratar dados de seus clientes com o objetivo de identificar tendências e adaptar seus produtos de acordo com as demandas do mercado".



1.4. Variedade de Dados (Data Varieties)

O ideal seria que todos os dados a serem analisados já estivessem em repositórios organizados, mas...

Dados podem ser de três tipos:

ESTRUTURADOS

Organizados e representados com uma estrutura rígida, a qual foi previamente planejado para armazenar dados, como por exemplo, um Banco de Dados Relacional;

SEMIESTRUTURADOS

Não possui uma estrutura pré-planejada ou formal de dados, mas contém tags ou outros marcadores para separar e identificar elementos. Arquivos XML, JSON e CSV, por exemplo, podem guardar dados deste tipo;

NÃO ESTRUTURADOS

Dados sem padrão ou estrutura. Arquivos como imagens, vídeos, gráficos e fotos, isolados ou misturados.



1.5. Análise de Dados

A Análise de Dados é uma das etapas mais importantes da Ciência de Dados e foco deste nosso curso.

Geralmente, organizada em 5 etapas, que são:



1.6. Formatos de Arquivos de Dados (File Formats)

Não importa onde os dados estejam armazenados, um Analista de Dados deve ser capaz de escrever comandos capazes de extraí-los de diferentes formatos de arquivos.

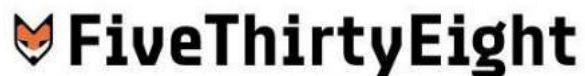
São INÚMEROS os formatos disponíveis para extração de dados valiosos, porém, os MAIS POPULARES são:



1.7. Fontes de Dados (Data Sources)

De nada adianta querer se basear em um modelo orientado a dados para tomada de decisões sem o suporte de bons conjuntos de dados (Data Sets).

Existem conjuntos de dados de todos os tipos, sejam públicos ou privados, mais gerais ou específicos. Alguns exemplos de locais onde podemos conseguir bons Data Sets, principalmente públicos:

The Kaggle logo, featuring the word "kaggle" in a bold, blue, sans-serif font with a small trademark symbol, set against a dark gray rectangular background.The Google Dataset Search logo, with the word "Google" in its multi-colored font above the words "Dataset Search" in a gray, sans-serif font.The FiveThirtyEight logo, featuring a small orange fox head icon to the left of the text "FiveThirtyEight" in a bold, black, sans-serif font.The dados.gov.br logo, featuring a 3D cube made of smaller cubes in green, yellow, and blue, followed by the text "dados.gov.br" in a blue, sans-serif font with a registered trademark symbol.The DATA.GOV logo, featuring a small American flag icon to the left of the text "DATA.GOV" in a blue, sans-serif font, all enclosed within a thin gray rectangular border.

1.8. Por que usar Python?

Python é a principal linguagem de programação voltada para Ciência de Dados.

São inúmeros os fatores que fazem do Python a linguagem ideal para Data Science, mas o grande diferencial se encontra nas bibliotecas otimizadas e poderosas que a mesma oferece para Análise de Dados. Dentre as muitas, destacam-se as seguintes e foco deste curso:

- **NumPy**

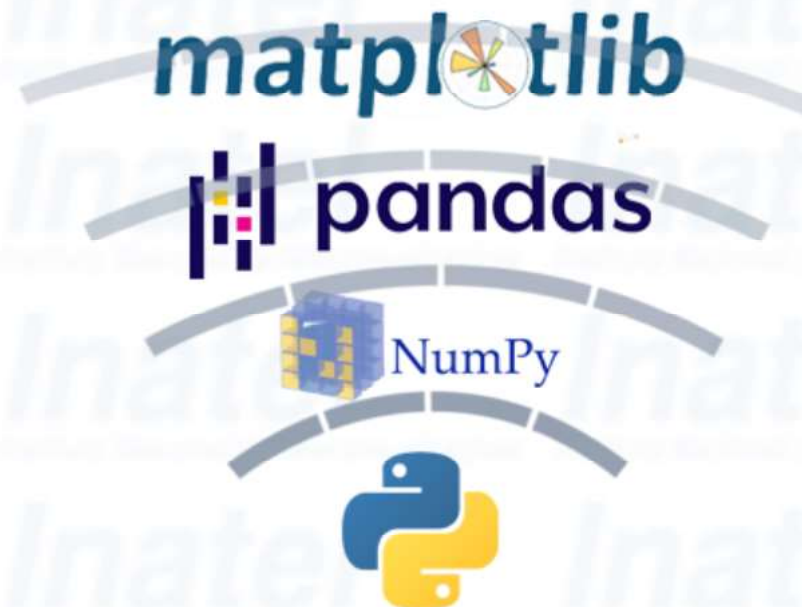
Biblioteca fundamental utilizada para se trabalhar com Arrays Multidimensionais no Python, além de oferecer um grande conjunto de funções e operações para realização de cálculos, dos mais simples aos mais complexos;

- **Pandas**

Principal biblioteca do Python para manipulação e análise de dados. Oferece estruturas e operações para manipular tabelas numéricas e séries temporais.

- **Matplotlib**

Principal biblioteca de plotagem bidimensional do Python, com inúmeros recursos para criação de diagramas e gráficos para facilitar a apresentação e análise de dados.



1.8. Por que usar Python?

Exercício 1 - Be Ready for Python!

Antes de iniciar nossos estudos, é importante que seja instalado os softwares adequados para a nossa disciplina. Por isso, é importante que você instale e configure na sua máquina as seguintes ferramentas (sempre que possível, as versões mais recentes):

1. Ambiente Python 3;

<https://www.python.org/downloads/>

2. IDE PyCharm Community;

<https://www.jetbrains.com/pycharm/download/>



obs: nesta disciplina o professor utilizará do Sistema Operacional Windows.

FIM DO CAPÍTULO 1



Próximo Capítulo
Fundamentos de Python