Ciência de Dados

Prof. Dr. José Eduardo Storopoli
josees@uni9.pro.br

Orientação Geral

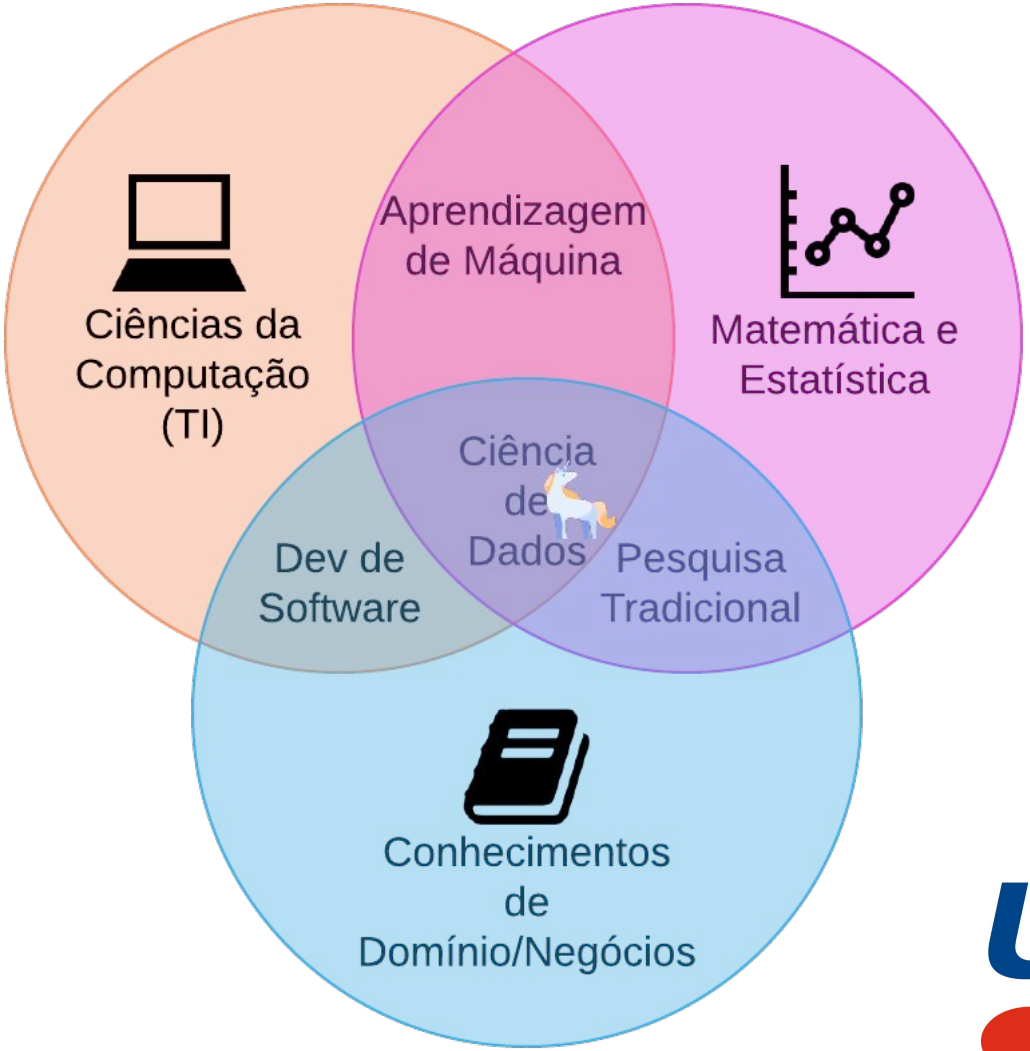
- **Conteúdo da Disciplina**: https://www.github.com/storopoli/ciencia-de-dados
- **Horário de Aula**
  - Chamada - Início 09:00
  - Chamada - Final 11:30
- **Avaliação**:
  - 1 Trabalho em Grupo
  - 1 Prova Digital Individual
  - **Nota**: $\frac{Trabalho + Prova}{2}$

UNINOVE

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** econômicos, financeiros e sociais, estruturados e não-estruturados, que visa a extração de conhecimento, detecção de padrões e/ou obtenção de ***insights*** para possíveis tomadas de decisão. (DHAR, 2013)

UNINOVE


Ciências da Computação (TI)
Aprendizagem de Máquina
Matemática e Estatística
Ciência de Dados
Dev de Software
Pesquisa Tradicional
Conhecimentos de Domínio/Negócios

UNINOVE

Por quê Ciência de Dados está um febre?

Poder Computacional

- Apollo 11 - 32kb RAM
- iPhone 11 Pro - 4Gb RAM
- 4Gb = 4,000,000kb

Big Data

- Acumulado até 2013 - 4,400Eb
- Até 2025 - 463Eb/dia
- 1 exabyte (Eb) = 1,000,000 terabyte

- Perceptron - 1957
- Regressão Logística - 1838
- SVM - 1963
- Redes Neurais - 1943

UNINOVE

O que podemos fazer com ciência de dados e aprendizagem de máquina?

Hastie, Tibshirani, & Friedman (2009)

Fridman (2018)

Etapas da Ciência de Dados

- Exploração dos Dados
- Geração de Hipóteses
- Teste das Hipóteses

UNINOVE

Aonde encontrar Dados?

- *Google* Dataset Search: https://datasetsearch.research.google.com
- *Kaggle* Datasets: https://www.kaggle.com/datasets
- data.world: https://data.world
- *Mendeley* Data: https://data.mendeley.com
- Academic Torrents: http://academictorrents.com

UNINOVE

Aonde encontrar Dados sobre Brasil?

- Portal Brasileiro de Dados Abertos: http://dados.gov.br
- Brasil.io Datasets: https://brasil.io/datasets
- Awesome Brazil Data: https://github.com/juliohm/awesome-brazil-data

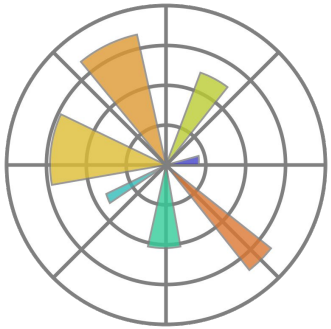
UNINOVE

Ferramentas

Bibliotecas Python - *open source*

pandas

XGBoost



statsmodels

PyTorch

TensorFlow

scikit learn

NumPy

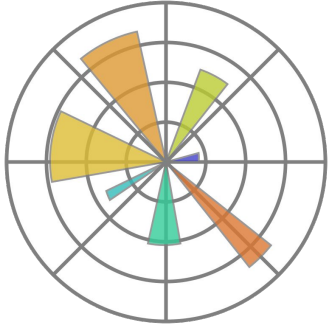
UNINOVE

Pandas

- Criado em 2008 por Wes McKinney
- Importação de dados como *DataFrame*
- CSV, Texto, Excel, SQL, HDF5, JSON, Feather, *Google Big Query*, SAS, SPSS, Stata, HTML
- Transformação e combinação de dados: *Wrangling*, *subsetting*, *groupby*, etc.
- Séries Temporais
- Lançada v1.0 em Janeiro de 2020

UNINOVE

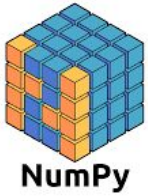
Matplotlib



- Lançada em 2003
- API orientada a objetos para gerar gráficos 2D
- Inspirada nos gráficos de MATLAB
- Pegada “Pythonesca”
- Criação dos gráfico com base em adição de objetos até obter o gráfico final

UNINOVE

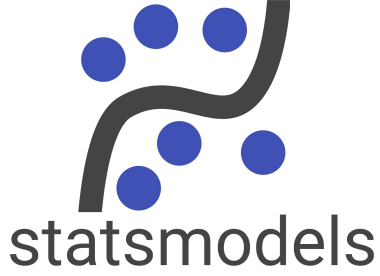
NumPy



- Lançada em 2006 (Numeric 1995)
- Pacote básico de computação científica em Python
- Algebra linear: vetores, matrizes e arrays
- Operações de algebra linear: Decomposições, eigenvalues, operações etc.
- Geração de número randômicos
- Transformações *Fourier*
- Utiliza *BLAS* e *LAPLACK* para velocidade e eficiência

UNINOVE

Statsmodel



- Criado em 2009 (*Google Summer of Code*)
- Modelos Estatísticos Frequentistas
- Integrado com pandas
- Computação com NumPy
- Gráficos em matplotlib
- Usa fórmulas estilo R
  Y ~ X + Z + X*Z

UNINOVE

Stan

- Criado em 2012 (*Columbia University - Statistics*)
- Estatística Bayesiana com *Markov Chain Monte Carlo* (MCMC)
- *Hamiltonian Monte Carlo* (HMC)
- *Approximate Bayesian Inference with Variational Inference* (ADVI)
- Motor em C++
- *Wrappers* em Python, R, julia, MATLAB

UNINOVE

Scikit-Learn

- Criado em 2007 (*Google Summer of Code*)
- Principal Biblioteca de *Machine Learning* (ML)
- Classificação, Regressão, Clusterização, Redução de Dimensão, Seleção de Modelos e Pré-processamento de dados
- SVM, *Random Forests*, *Gradient Boosting*, *k-means* etc.
- Integração com matploblib, pandas, numpy

UNINOVE

XGBoost XGBoost

- Criado em 2014
- Biblioteca de *Gradient Boosting*
- Escrito em C++
- *Wrappers* em Python, R, julia, Ruby
- Roda em GPU - NVIDIA CUDA

UNINOVE

XGBoost XGBoost

O queridinho das competições Kaggle

UNINOVE

TensorFlow

- Criado pelo Google em 2017
- Principal Biblioteca de Deep Learning usada pela indústria (praticantes)
- Roda em GPU - NVIDIA CUDA
- Keras Nativo

UNINOVE

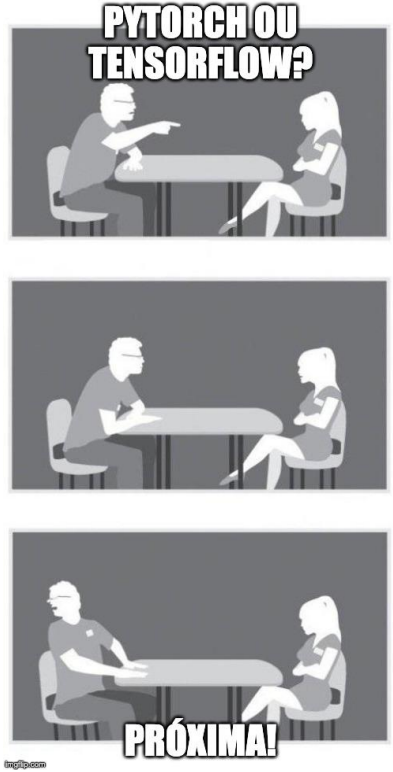
PyTorch

- Criado pelo Facebook em 2018
- Principal Biblioteca de Deep Learning usada pela academia (pesquisadores)
- Roda em GPU - NVIDIA CUDA

UNINOVE

TensorFlow

PyTorch



UNINOVE

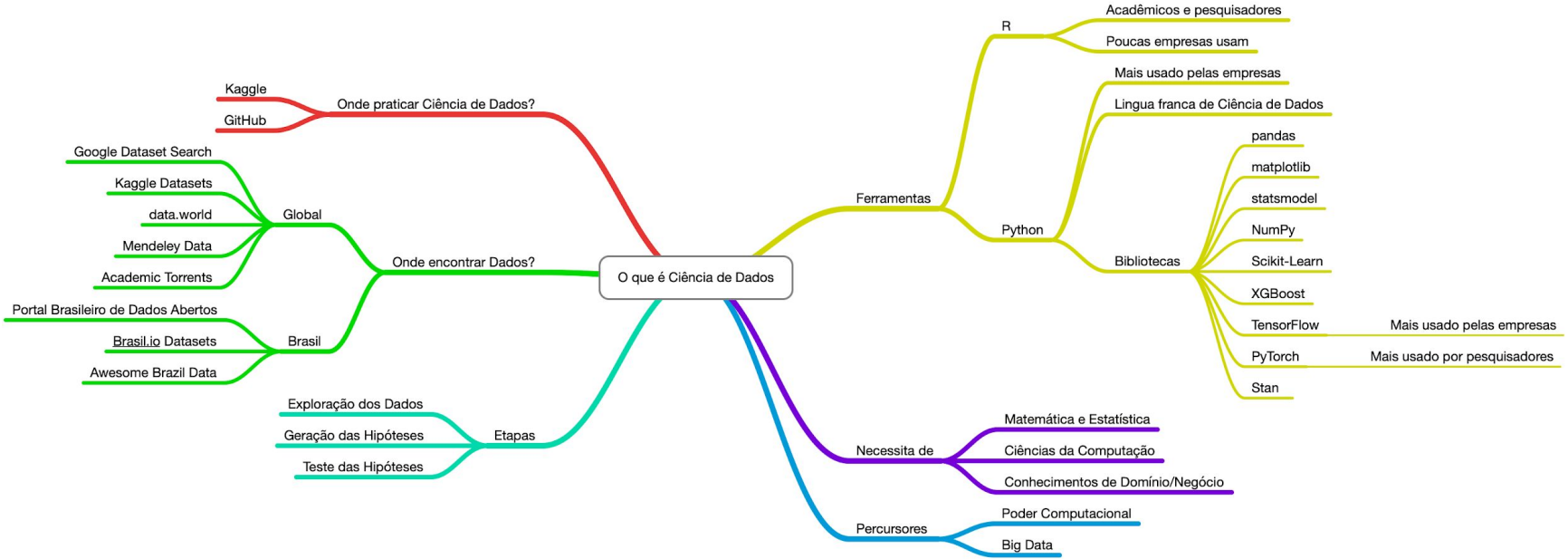
Onde praticar Ciência de Dados?

Conteúdo

- INTRODUÇÃO: CIÊNCIA DE DADOS - O QUE É
- PYTHON: OPERADORES ARITMÉTICOS, RELACIONAIS E LÓGICOS
  - DESVIO CONDICIONAL
  - ESTRUTURAS DE REPETIÇÃO, ESTRUTURA DE DADOS, MÉTODOS E FUNÇÕES, REQUISIÇÕES WEB, EXPRESSÕES REGULARES
- PYTHON: VETORES E MATRIZES (NUMPY)
- ESTATÍSTICA: MÉTODO ESTATÍSTICO E AMOSTRAGEM
- ESTATÍSTICA: ANÁLISE EXPLORATÓRIA DOS DADOS E REGRESSÃO
- ESTATÍSTICA: ANÁLISE DE SÉRIES TEMPORAIS
- PYTHON: ANÁLISE DE DADOS (PANDAS)
- VISUALIZAÇÃO DE RESULTADOS (MATPLOTLIB)
- APRENDIZADO DE MÁQUINA (SCIKIT-LEARN)

UNINOVE

Mapa Conceitual



UNINOVE

Referências

DHAR, Vasant. Data science and prediction. **Communications of the ACM**, v. 56, n. 12, p. 64-73, 2013.

FRIDMAN, Lex. “**MIT Human-Centered Autonomous Vehicle**”. YouTube video, 7:39. September 29, 2018. https://youtu.be/OoC8oH0CLGc.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. T**he elements of statistical learning: data mining, inference, and prediction**. Springer Science & Business Media, 2009.

UNINOVE