DATA SCIENCE

Conteúdo Programático

Prof. Gabriel Resende Machado

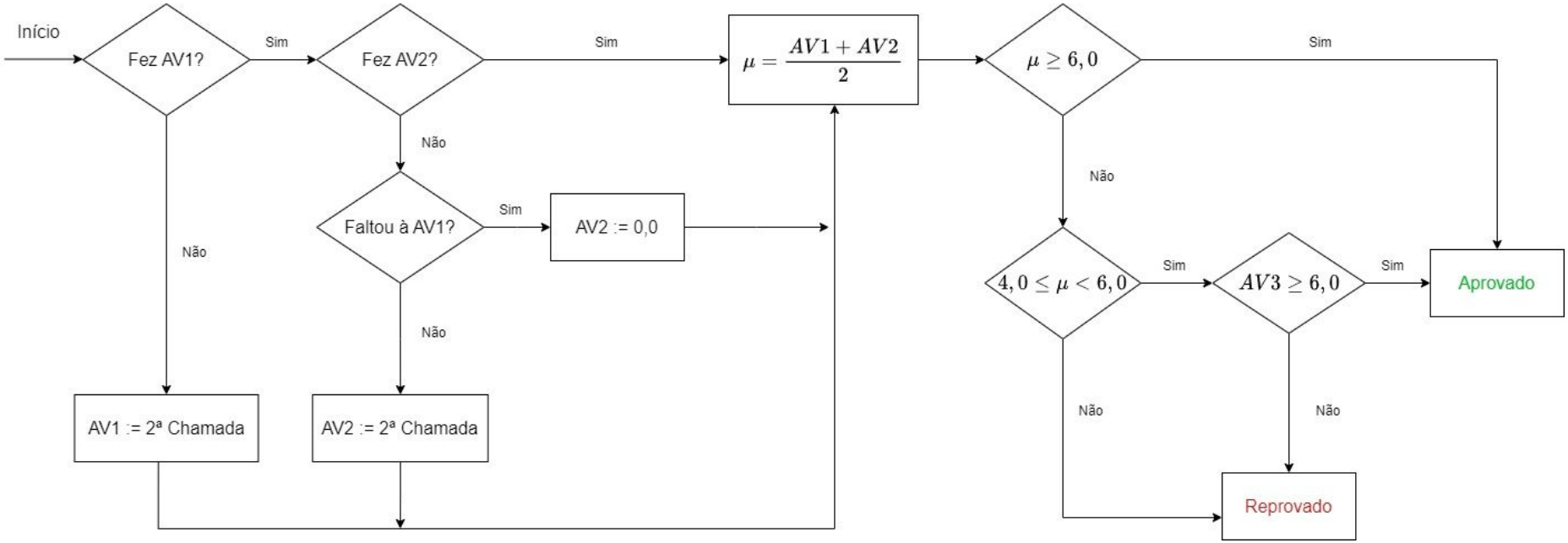
gabriel.rmachado10@gmail.com

https://www.linkedin.com/in/machadogabriel

https://github.com/UNIFESO-Gabriel/data-science

CONTEUDO PROGRAMATICO – *DATA SCIENCE*			
CONTEÚDO NUCLEAR	**AULA**	**DATA**	**TÓPICOS**
Introdução	01	14/08	Introdução à Ciência de Dados e à Linguagem *Python*.
Análise exploratória	02	21/08	Exploração de dados, estatística descritiva e visualização (histograma, distribuição de probabilidade, *boxplots*, valores máximo, mínimo, média, mediana, moda, quantis, percentis, IQR, covariância, correlação); prática em laboratório.
Pré-processamento de dados	03	28/08	Pré-processamento de dados (integração de dados, tratamentos de valores ausentes, *outliers* e duplicidades, conjuntos de dados desbalanceados, transformação de dados e seleção de atributos); prática em laboratório.
Aprendizagem supervisionada I	04	04/09	Algoritmos de aprendizagem supervisionada I (k-NN); prática em laboratório.
Aprendizagem supervisionada II	05	11/09	Algoritmos de aprendizagem supervisionada II (Regressão Logística); prática em laboratório.
Aprendizagem supervisionada III	06	18/09	Algoritmos de aprendizagem supervisionada III (Árvores de Decisão); prática em laboratório.
Avaliação 1 (conteúdo de 14/08 a 18/09)	07	25/09	1ª Avaliação (AV1).
Aprendizagem supervisionada IV	08	09/10	Algoritmos de aprendizagem supervisionada IV (Comitês de Aprendizado, *Random Forests*, técnicas de reamostragem); prática em laboratório.
Aprendizagem não supervisionada I	09	16/10	Algoritmos de aprendizagem não supervisionada I (*k-means* e *k-medoids*); prática em laboratório.
Aprendizagem não supervisionada II	10	23/10	Algoritmos de aprendizagem não supervisionada II (*DBSCAN*); prática em laboratório.
Aprendizagem não supervisionada III	11	30/10	Algoritmos de aprendizagem não supervisionada III (clusterização hierárquica); prática em laboratório.
Aprendizagem não supervisionada IV	12	06/11	Algoritmos de aprendizagem não supervisionada IV (*Apriori*); prática em laboratório.
Avaliação 2 (conteúdo de 09/10 a 06/11)	13	13/11	2ª Avaliação (AV2).
Revisão do conteúdo	14	27/11	Revisão do conteúdo para a Reavaliação.
Matéria toda (referente à **AV1 ou AV2** em que o aluno esteve ausente)	15	04/12	2ª Chamada
Reavaliação **(Matéria toda)**	16	11/12	3ª avaliação.
Encerramento	17	15/12	Entrega de notas.

Avaliações



- **Ambas AV1 e AV2 serão compostas por 10 questões objetivas e 02 discursivas;**
- **A AV1 ou AV2 compõem 40% da nota trimestral. 60% restantes vêm de trabalhos e/ou atividades.**

DATA SCIENCE

AULA 1 - Introdução

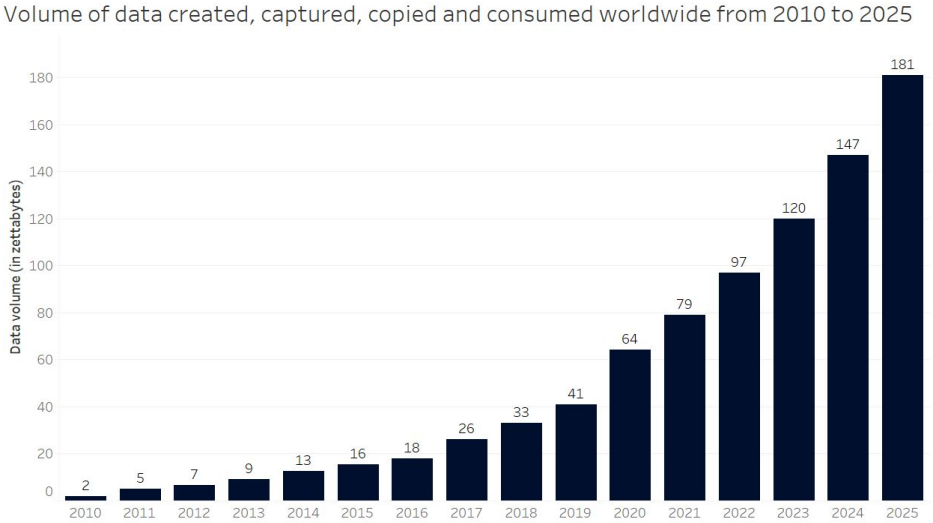
Prof. Gabriel Resende Machado

gabriel.rmachado10@gmail.com

https://www.linkedin.com/in/machadogabriel

https://github.com/UNIFESO-Gabriel/data-science

Dados: o ouro do Século XXI



Disponível em https://shorturl.at/wU248.

Abreviação	Unidade	Valor
b	bit	0 or 1
B	bytes	8 bits
KB	kilobytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes
GB	gigabyte	$1{,}000^3$ bytes
TB	terabyte	$1{,}000^4$ bytes
PB	petabyte	$1{,}000^5$ bytes
EB	exabyte	$1{,}000^6$ bytes
ZB	zettabyte	$1{,}000^7$ bytes
YB	yottabyte	$1{,}000^8$ bytes

- **De acordo com https://shorturl.at/aGMPR, em 2019, por dia, em média:**
  - 500 milhões de *tweets* foram postados;
  - 294 bilhões de *emails* foram enviados;
  - 95 milhões de fotos foram compartilhadas no *Instagram*;
  - 5 bilhões de buscas foram realizadas, sendo 3,5 bilhões no *Google;*
  - 4 *petabytes* de dados foram criados pelo *Facebook.*

1 ZB equivale a 1.073.741.824 HDs

A era do *Big Data*

Adaptado de https://shorturl.at/bdWZ2.

- **Dados produzidos a partir do aumento da digitalização e de avanços tecnológicos em diversos setores;**
- *** Se dividem principalmente em dados (i) estruturados, (ii) semi-estruturados e (iii) desestruturados.**

A necessidade por Cientistas de Dados

Data Scientist: A Hot Job That Pays Well

Since December 2013, data science postings have rocketed 256%—more than tripling. Houston, San Francisco offer the best salaries for data scientists.

By Andrew Flowers **January 17, 2019**

Back in 2012, the Harvard Business Review called data scientist "the sexiest job of the 21st Century." Six years later, the job has only grown sexier. More employers than ever are looking to hire these skilled digital data jockeys. And while interest from job seekers is growing, Indeed research shows job postings are growing even faster. There may not be enough skilled applicants, so bargaining power in this explosively-growing data scientist job market likely remains with job seekers.

Job postings statistics tell the story. Data scientist postings as a share of all postings on Indeed jumped a full 31% in December 2018, compared with the same period the year before. Yet, that was just another solid year in the spectacular and steady rise in data science jobs on Indeed. Since December 2013, postings have rocketed 256%—more than tripling.

Why is this job growing like gangbusters? It's because employers use data scientists to solve all sorts of problems. In essence, data scientists are tasked to take raw data and use programming, visualization, and statistical modeling to extract insights, according to the Bureau of Labor Statistics (BLS).

Disponível em https://shorturl.at/vwLMN.

OLHAR DIGITAL NOTÍCIAS VÍDEOS CIÊNCIA E ESPAÇO DICAS E TUTORIAIS EDITORIAS OFERTAS

Olhar Digital > Pro > Levantamento mostra aumento de mais de 450% nas vagas para cientistas de dados

PRO

Levantamento mostra aumento de mais de 450% nas vagas para cientistas de dados

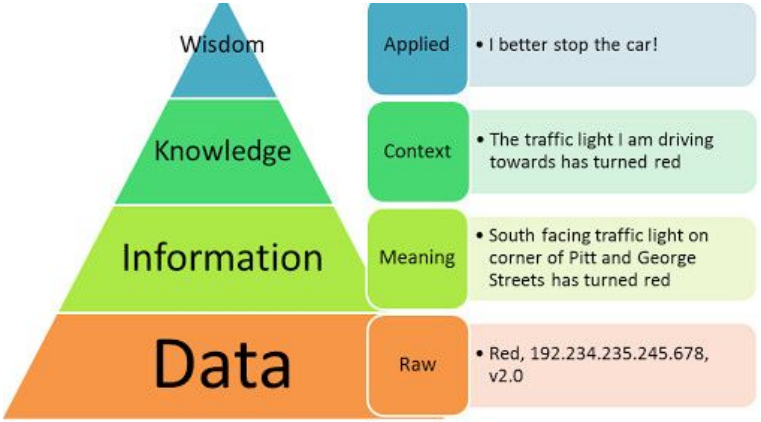
Lauro Lam | 24/06/2021 13h42

Disponível em https://shorturl.at/eixW5.

Mas afinal, o que é Ciência de Dados?

- A Ciência de Dados envolve princípios, processos e técnicas para a compreensão de fenômenos a partir da **análise automatizada** de dados. [1, p. 4];

- O papel da Ciência de Dados é **transformar** grandes quantidades de dados, que as empresas coletam diariamente, em **informações** compreensíveis e úteis por meio de Análise Estatística e do uso de técnicas de **Inteligência Artificial (IA),** como o **Aprendizado de Máquina (AM)** [2].

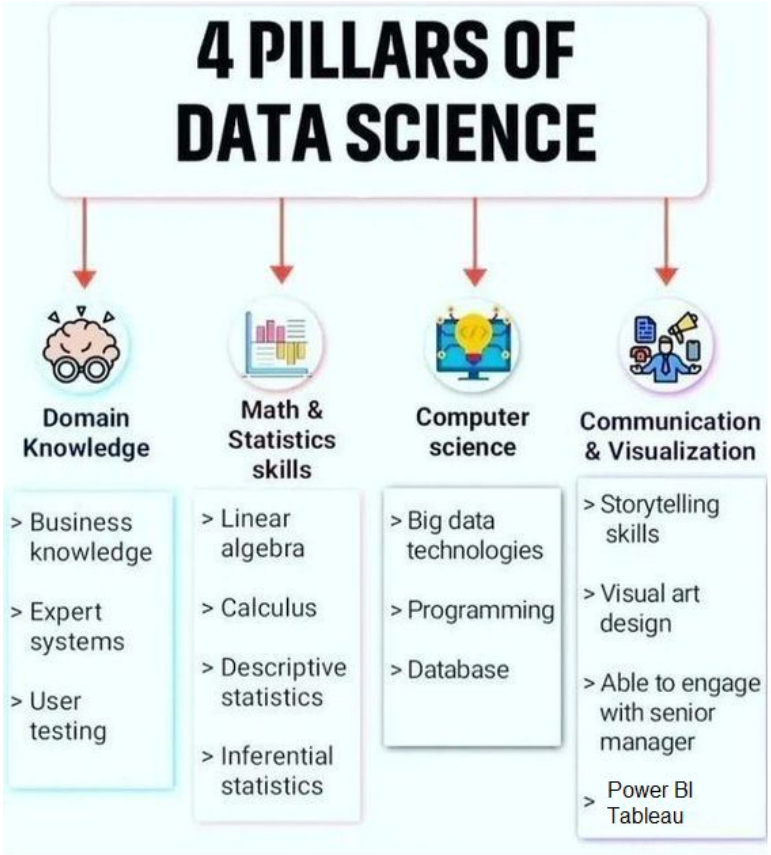


Disponível em https://shorturl.at/wyEGL.

[1] PROVOST, Foster; FAWCETT, Tom. Data Science for Business: What you need to know about data mining and data-analytic thinking." O'Reilly Media, Inc.", 2013.

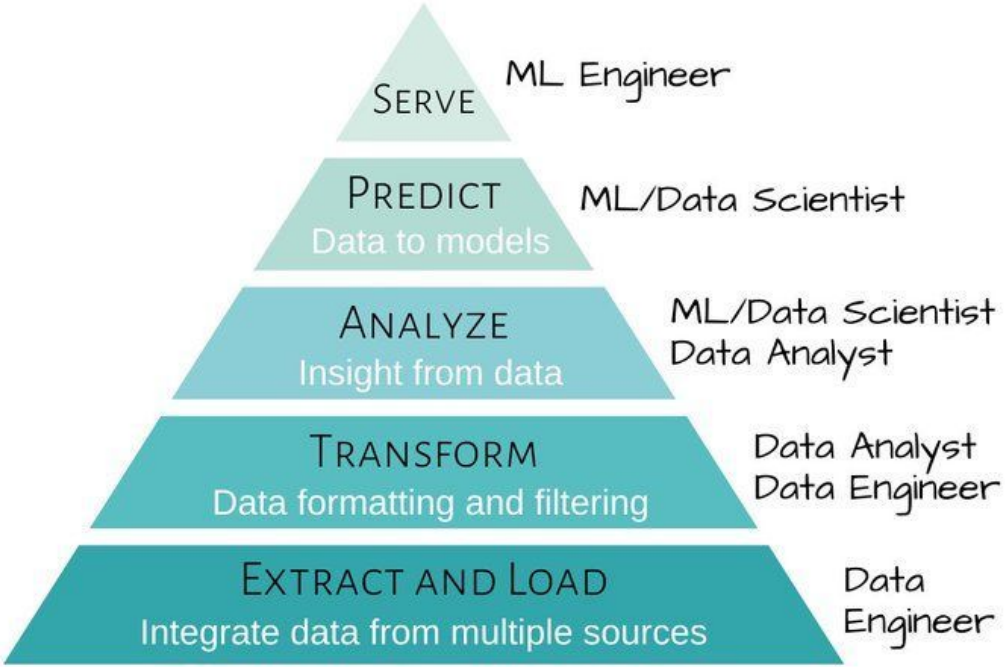
[2] Adaptado de https://shorturl.at/pyXY6.

Os pilares da Ciência de Dados



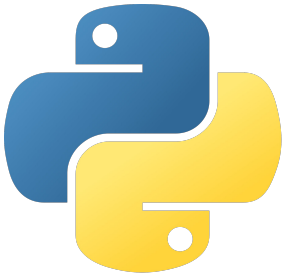
Adaptado de https://dailyinfographic.com/the-4-pillars-of-data-science.

Hierarquia de *jobs* na área de dados



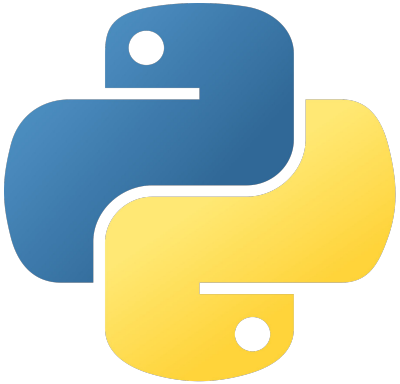
Disponível em https://shorturl.at/mEMN4.

Ferramentas utilizadas ao longo do curso



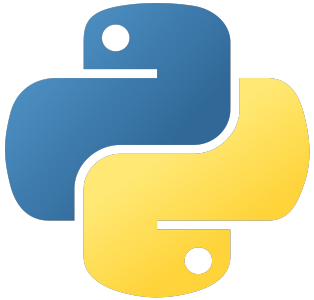
Sobre o *Python* I

- Linguagem multiparadigma, interpretada e dinamicamente tipada;
- Desenvolvida a partir das linguagens C/C++;
- Conhecida por sua sintaxe amigável e de “fácil aprendizado”;
- Versátil e multifuncional: possui grande variedade de pacotes que estendem suas funcionalidades;
- Principal linguagem utilizada em projetos relacionados à *Data Science.*



Sobre o *Python* II

- Nem tudo são flores: *Python* é uma linguagem conhecida também pela **sua lentidão**;
- Versões mais recentes do *Python* vêm tentando otimizar o interpretador, diminuindo o tempo de execução do código;
- Devido à lentidão, outras linguagens no ramo de *Data Science* podem ser adotadas, como *Julia* e *Go*;
    - Vale ressaltar aqui o surgimento da linguagem *Mojo**, um possível sucessor do *Python* a médio-longo prazo.

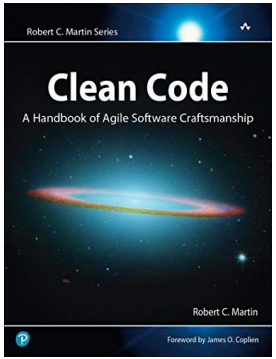


Mojo 🔥

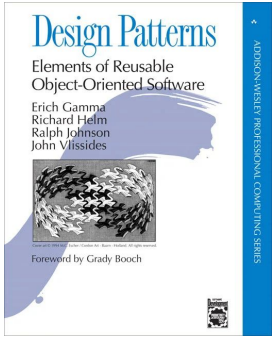
* https://www.modular.com/mojo.

Python: Boas práticas de programação

"Any fool can write code that a computer can understand. Good programmers write code that humans can understand." – **Martin Fowler**



- **Tenha empatia:** pratique a escrita de um código legível, documentado e manutenível. No futuro, alguém irá te agradecer, incluindo você mesmo!
- ***For loops* devem ser evitados ao máximo.** Use as funções *built-in* do *Python* e de suas bibliotecas;
- Leia a documentação oficial do Python de boas práticas: PEP 8*;
- Leia livros e consuma materiais que auxiliem nas boas práticas de programação. Cientistas de Dados também são programadores!



* https://peps.python.org/pep-0008.

Python - *Overview* da Linguagem

DATA SCIENCE

AULA 1 - Introdução
Duvidas e/ou perguntas?

gabriel.rmachado10@gmail.com

https://www.linkedin.com/in/machadogabriel

https://github.com/UNIFESO-Gabriel/data-science