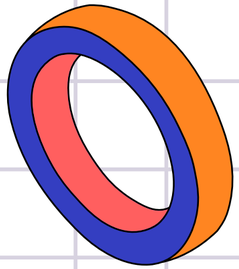
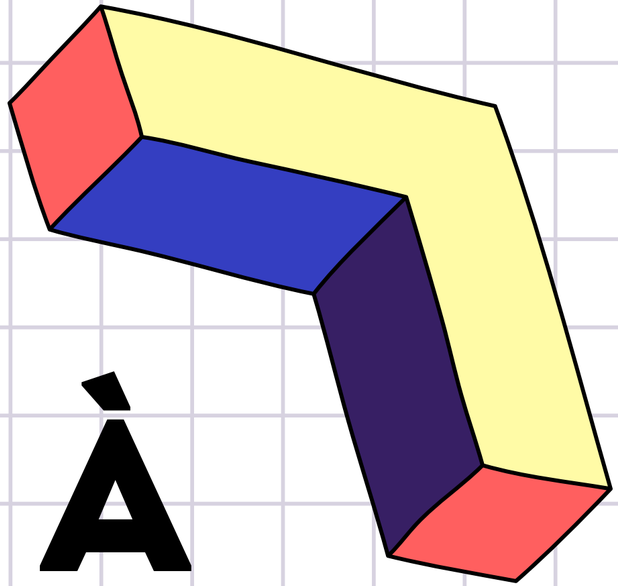
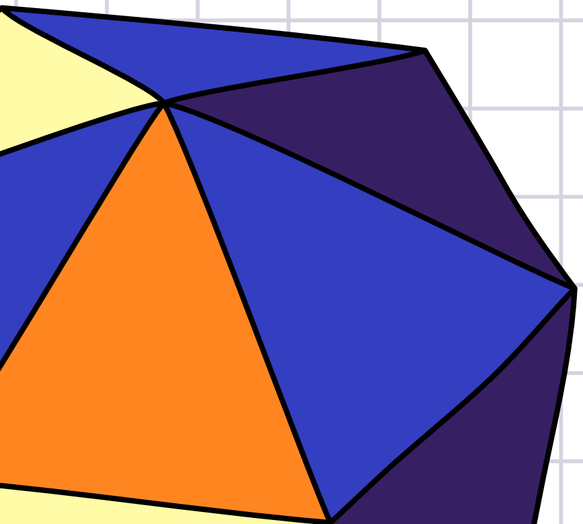
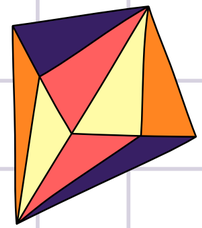
ESCOLA DE INVERNO IESP-UERJ 2023





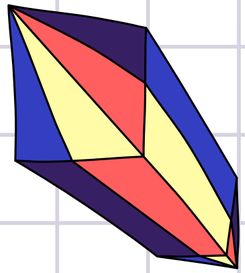
INTRODUÇÃO À CIÊNCIA DE DADOS





PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023



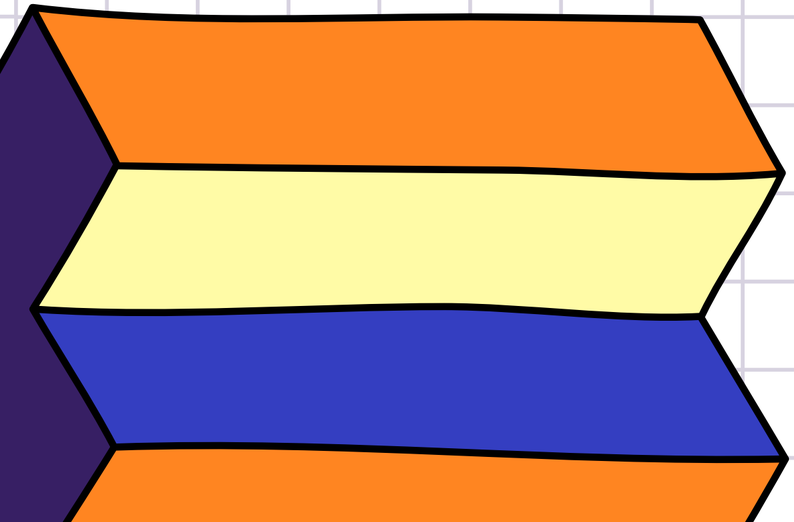
APRESENTAÇÕES

QUEM SOU EU?

- Professor na Escola de Comunicação, Mídia e Informação da Fundação Getulio Vargas (FGV ECMI)
- Doutorando em Ciência Política pelo IESP-UERJ
- Mestre em Ciência Política pelo IESP-UERJ
- Cientista de Dados no Instituto de Estudos da Religião (ISER)
- Estudos na área de partidos, eleições, política russa, religião e política, mídias sociais e modelos de IA generativa

ONDE ME ENCONTRAR?

GITHUB **LINKEDIN** **TWITTER**



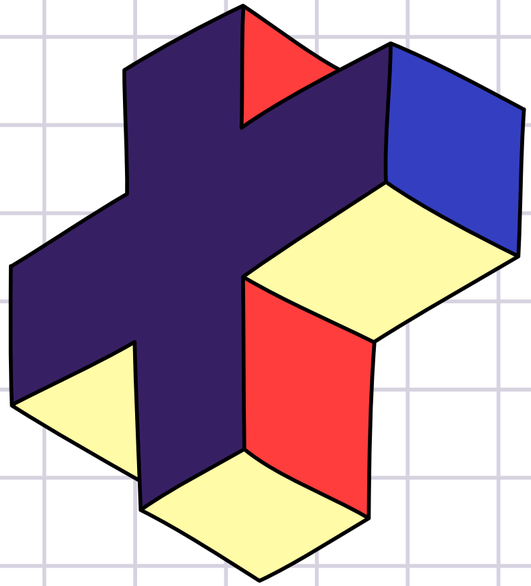
PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

>> Todos os modelos estão errados, mas alguns são úteis"
- George Box

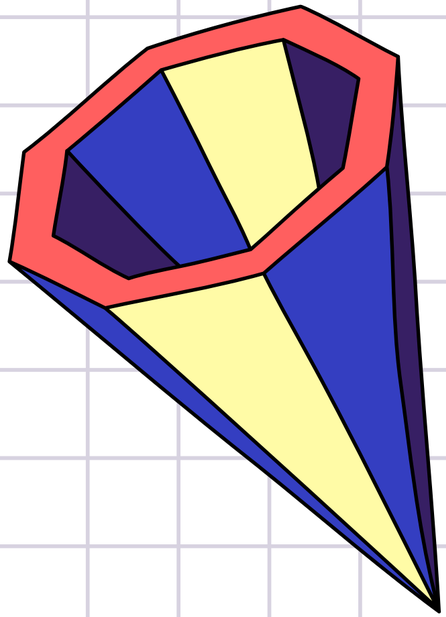
01

ENTENDENDO CIÊNCIA DE DADOS



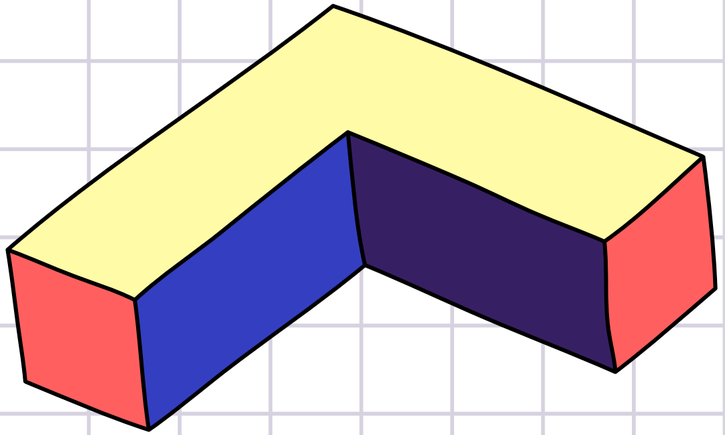
PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023



O QUE É CIÊNCIA DE DADOS?

PROF. MATHEUS C. PESTANA



É UM CAMPO INTERDISCIPLINAR

que envolve o uso de métodos científicos, processos, e sistemas para extrair conhecimento e fazer inferência sobre dados.

UTILIZA TÉCNICAS ORIUNDAS

da estatística, matemática, ciência da computação e ciência da informação para analisar esses dados, resolver problemas e tomar decisões.

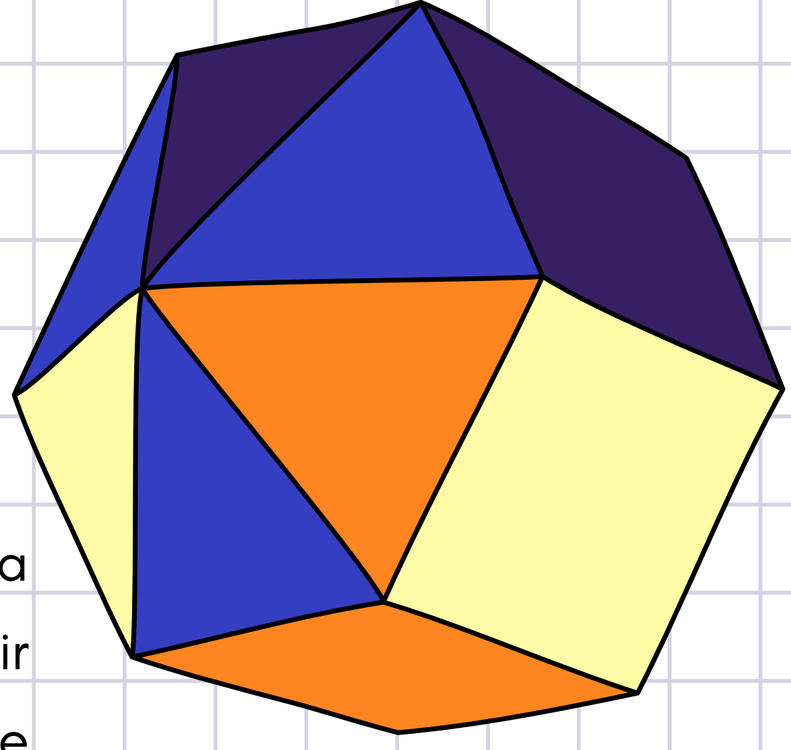
MODELOS ESTATÍSTICOS,

algoritmos e redes neurais são algumas das ferramentas utilizadas para processar, analisar e extrair insights desses dados.

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

DATANATURE

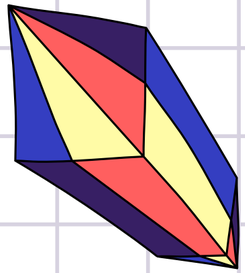


"Ciência de Dados é a ciência de estudo da *datanature* e a ciência do dado em si. Em um nível básico, envolve extrair conhecimento de dados. Como alguns dados da datanature representam coisas reais, o conhecimento adquirido através desses dados pode ser usado para a ciência natural e social. [...] A ciência de dados é a teoria, o método e a tecnologia de estudar a *datanature*.

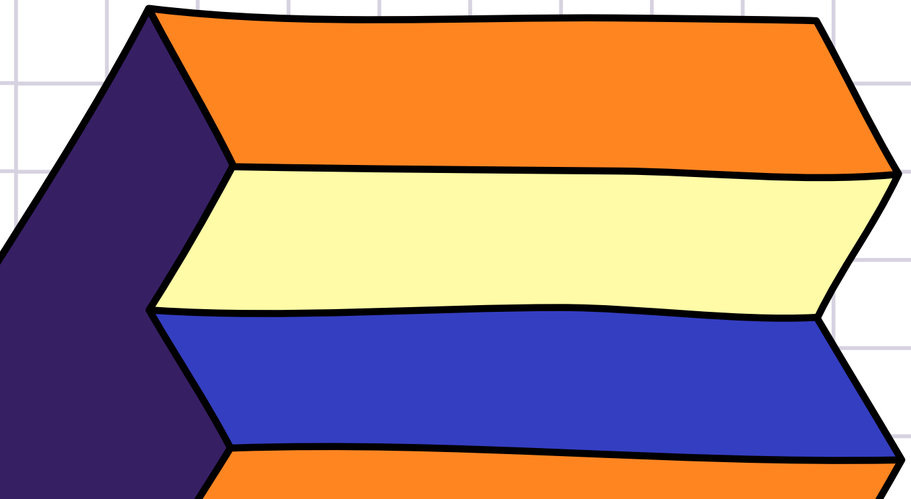
ZHU E XIONG (2015)

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023



SOLUCIONANDO PROBLEMAS



SISTEMAS DE RECOMENDAÇÃO

- Netflix
- Amazon
- MercadoLivre

SISTEMAS BANCÁRIOS

- Sustentabilidade
- Padrão de comportamento
- Análise de crédito

BUSINESS

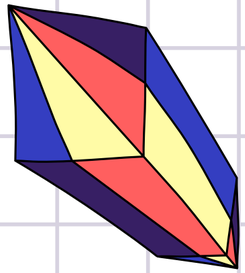
- Expectativa de vendas
- Demanda de insumos

REDES SOCIAIS

- Acompanhamento de métricas
- Padrões de postagem
- Comportamento nas redes
- Identificação de violações

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023



SOLUCIONANDO PROBLEMAS

MEDICINA

- Identificação de tumores

EDUCAÇÃO

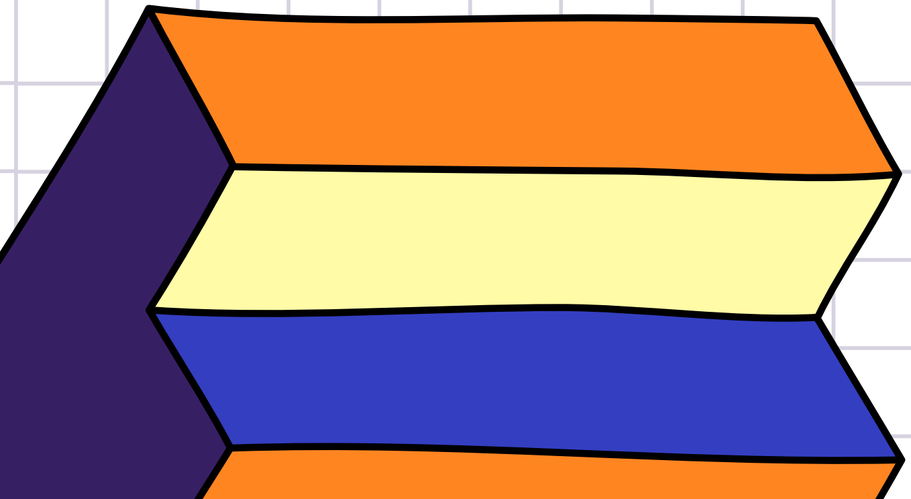
- Padrões de evasão
- Análise de desempenho escolar

POLÍTICAS PÚBLICAS

- Simulações de mudanças em políticas

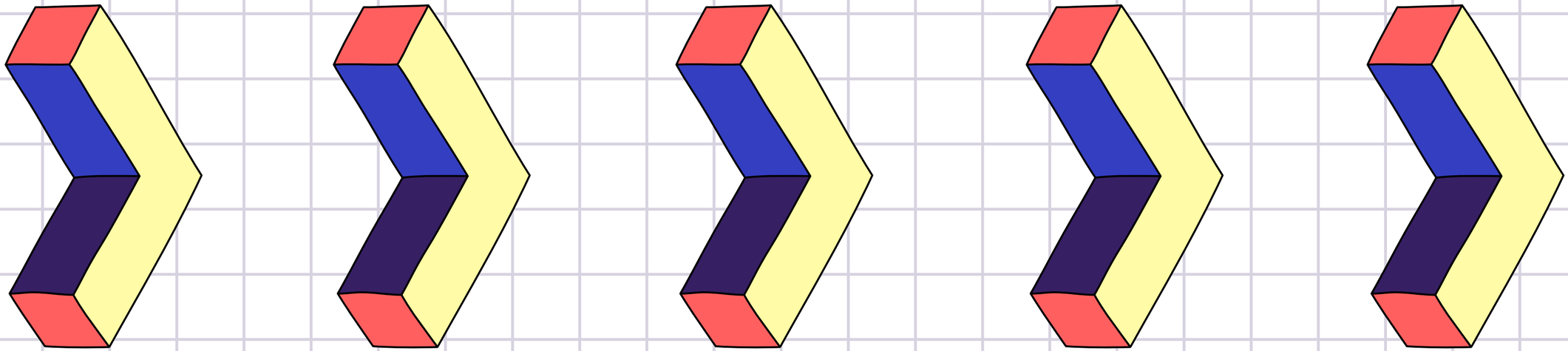
HISTÓRIA

- Leitura e identificação de textos manuscritos antigos
- Reconhecimento de caligrafia



PROF. MATHEUS C. PESTANA

O PROCESSO



COLETA

LIMPEZA

EXPLORAÇÃO

MODELAGEM

DEPLOY

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

ALGUNS TERMOS RECORRENTES

ALGORITMOS

- Conjunto específico de instruções ou regras bem definidas que visa resolver um problema

FUNÇÃO

- Código que executa uma tarefa específica. Em geral, pode receber argumentos e entregar uma saída.

MODELO

- Representação simplificada de um sistema real, buscando simular ou prever o comportamento de um sistema complexo

INTELIGÊNCIA ARTIFICIAL

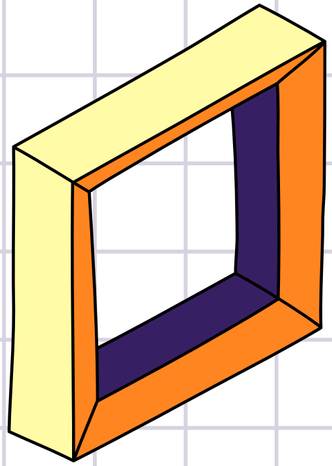
- Ramo da Ciência que busca fazer com que máquinas executem tarefas que exigem "cognição"

MACHINE LEARNING

- Subcampo da IA. Busca criar algoritmos e modelos para máquinas aprenderem a partir dos dados.

DEEP LEARNING

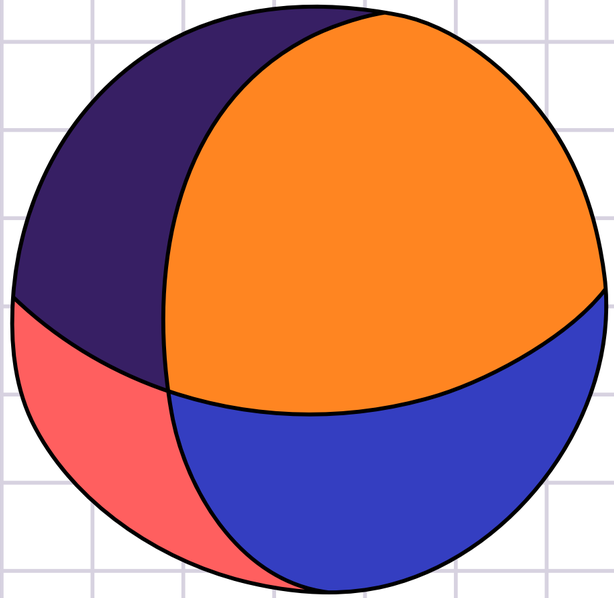
- Subcampo do ML. Usa estruturas que mimetizam um cérebro humano, com muitas camadas, para aprender com mais eficácia.

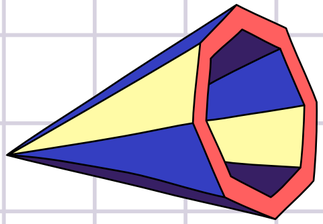


PROF. MATHEUS C. PESTANA

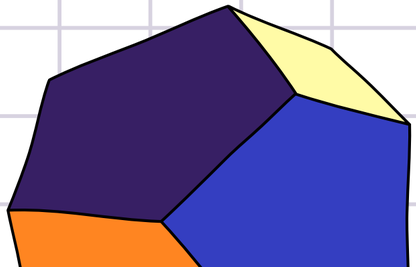
ESCOLA DE INVERNO IESP-UERJ 2023

ALGORITMOS VS. MODELOS



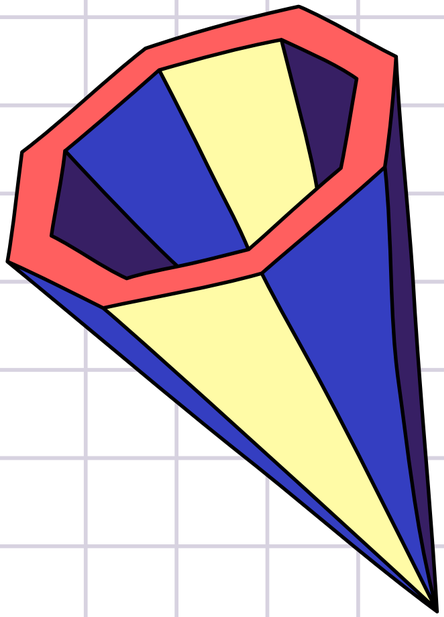


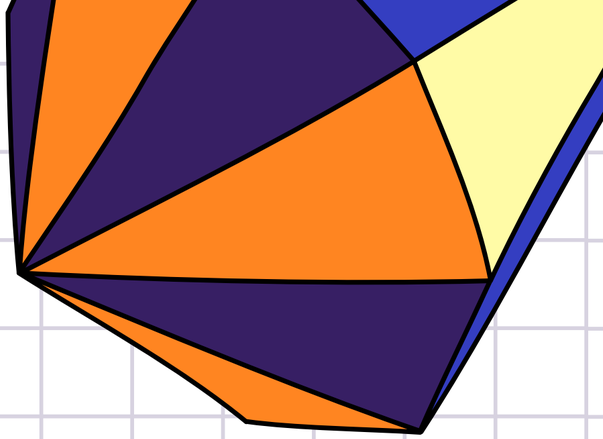
- Um **modelo** é um sistema que busca prever ou explicar algo.
- O **algoritmo** é o processo de aprendizado de máquina usado para treinar esse modelo.
    - Podemos gerar um modelo que classifica posts de redes sociais entre Lulistas e Bolsonaristas pelo léxico utilizado.
    - Para treinar esse modelo, temos várias possibilidades de algoritmos, que farão o ajuste dos dados, e criarão um sistema que poderá prever em cima de dados novos, não vistos anteriormente.



PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023





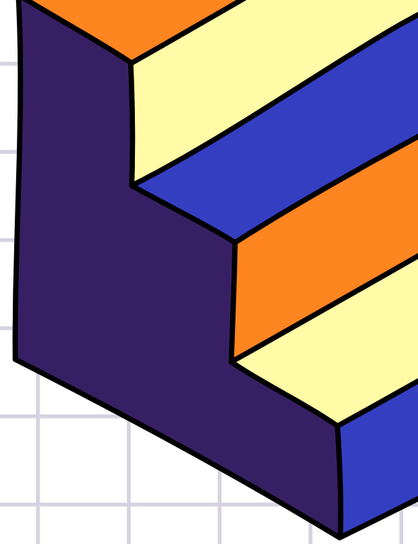
DADOS

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

DADOS

Qualquer tipo de informação que pode ser coletada, armazenada, processada ou transmitida por um dispositivo digital

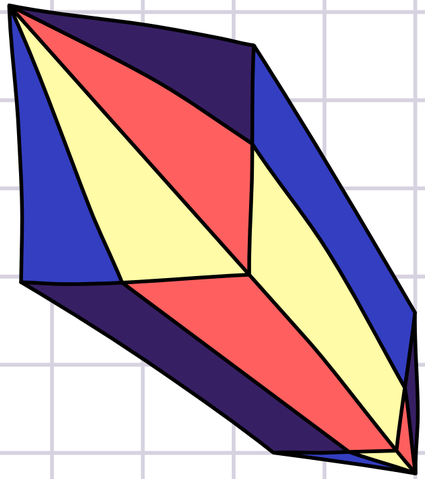


ESTRUTURADOS

Dados tabulares (retangulares, como em tabela), que é exibido em linhas e colunas.
Ex.: Arquivos de Excel, CSV, SQL, etc.

DESESTRUTURADOS

Dados "desorganizados", sem um desenho ou modelo definido.
Ex.: sites, arquivos de áudio e vídeo em pastas, etc.



SEMIESTRUTURADOS

São dados que não estão estruturados, mas tem uma relação/organização em si.
Ex.: Javascript Object Notation (JSON), XML, etc.

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

BOAS PRÁTICAS EM DADOS

CÉLULA

Cada célula tem seu próprio valor. Ex.: "R$ 3500,00"

LINHAS

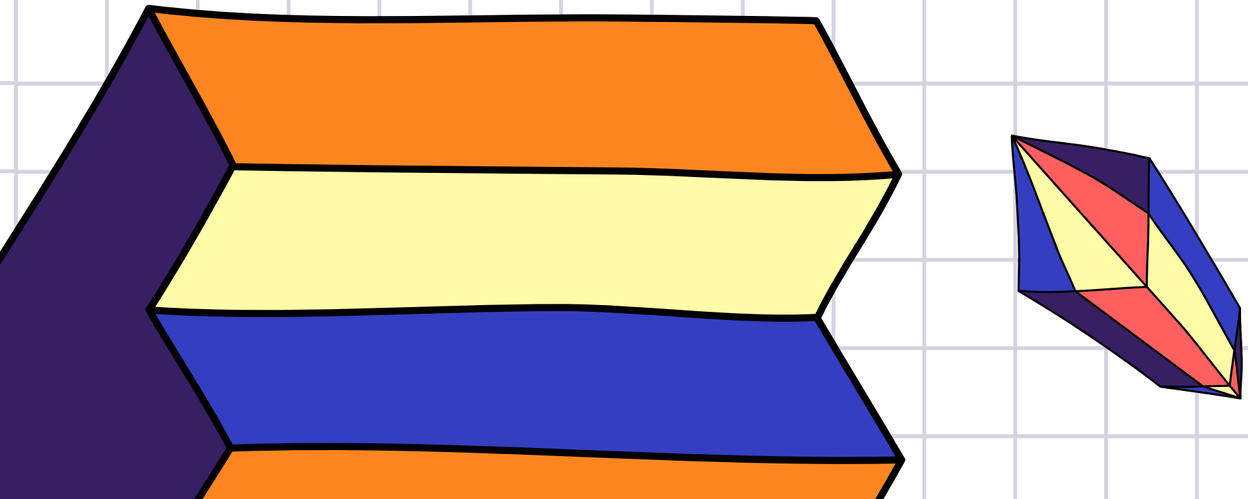
Cada linha, uma observação. Ex.: um indivíduo, com suas características observadas

COLUNAS

Cada coluna, uma variável (característica observada) Ex.: Idade, gênero, nome, salário, cidade natal...

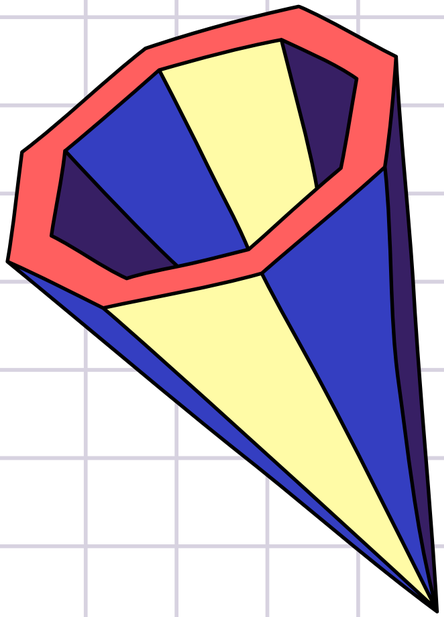
TABULAR/RETANGULAR

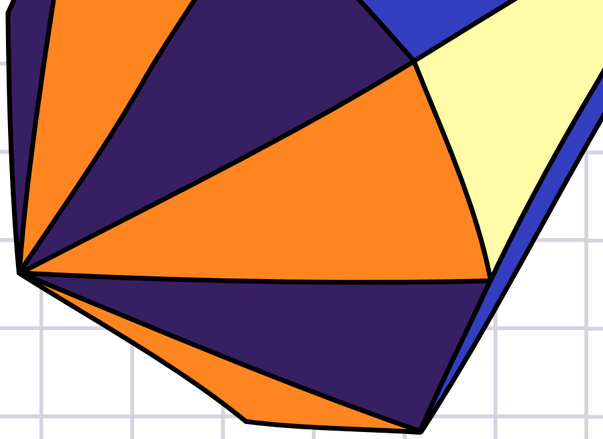
O dado é retangular, ou seja, todos os casos tem algum dado (ou não-dado) em todas as variáveis



PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

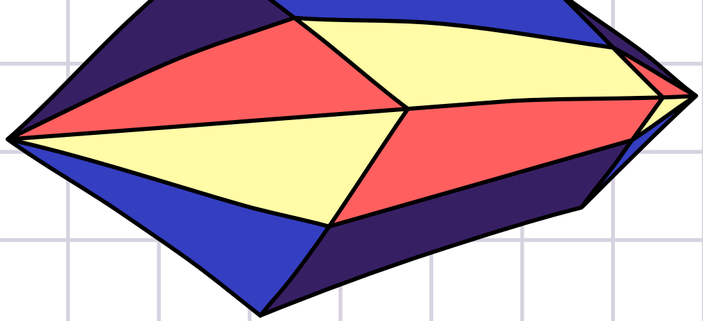




VARIÁVEIS

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023



NUMÉRICAS

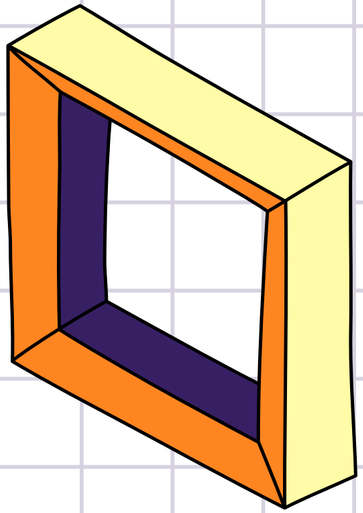
Quantitativas, representam uma medida. Podem ser **contínuas**, assumindo qualquer valor em um intervalo específico, ou **discretas**, assumindo apenas valores inteiros.

CATEGÓRICAS

Qualitativas, e descrevem um atributo ou uma característica. Podem ser **nominais**, quando são características sem ordem ou prioridade, ou **ordinais**, quando tem uma ordem interna.

BINÁRIAS

Considera apenas dois valores: 0 ou 1.



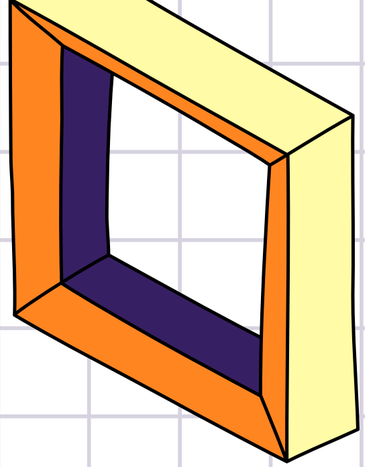
VARIÁVEIS

"Toda e qualquer característica, número ou quantidade que pode ser mensurada ou contada"

"Tudo aquilo que pode variar"

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023



DEPENDENTES (RESPOSTA)

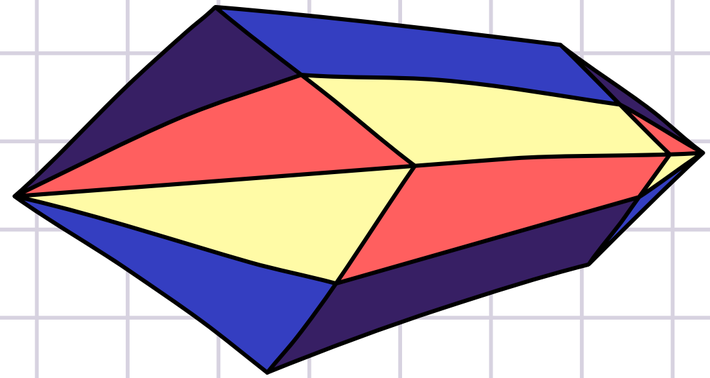
Variáveis que queremos **prever** ou **estimar**. São dependentes pois **DEPENDEM** de outros fatores (outras variáveis).

INDEPENDENTES (EXPLICATIVA)

Qualitativas, e descrevem um atributo ou uma característica. Podem ser **nominais**, quando são características sem ordem ou prioridade, ou **ordinais**, quando tem uma ordem interna.

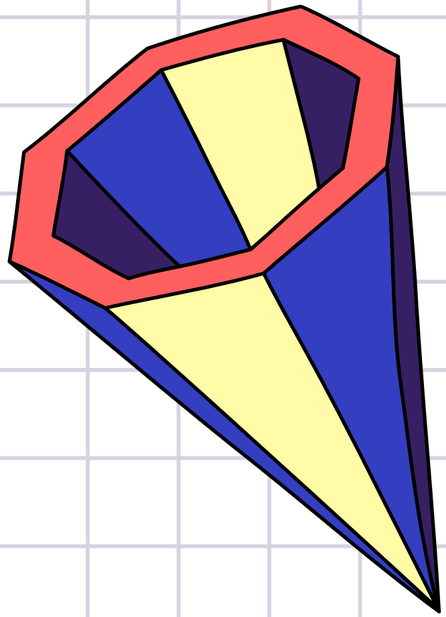
VARIÁVEIS

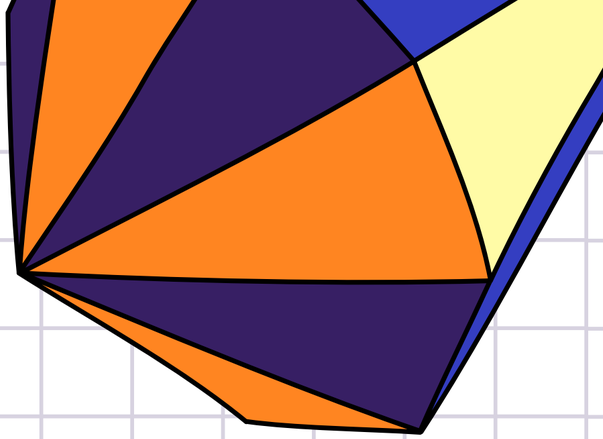
"Toda e qualquer característica, número ou quantidade que pode ser mensurada ou contada"
"Tudo aquilo que pode variar"



PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

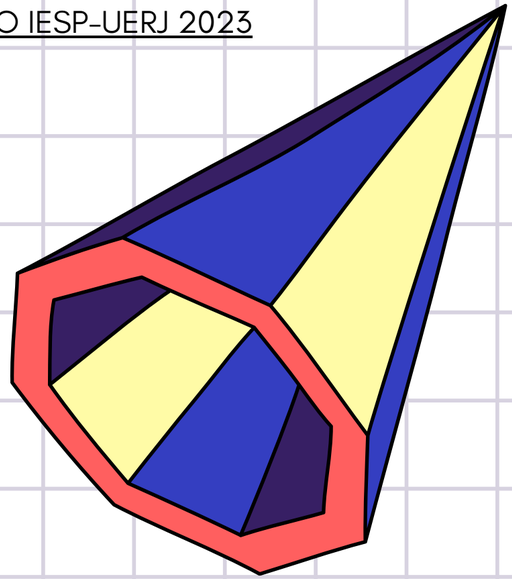




PYTHON & COLAB

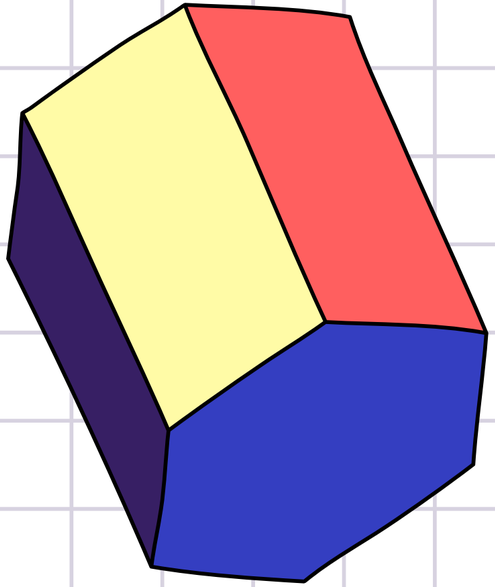
PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023



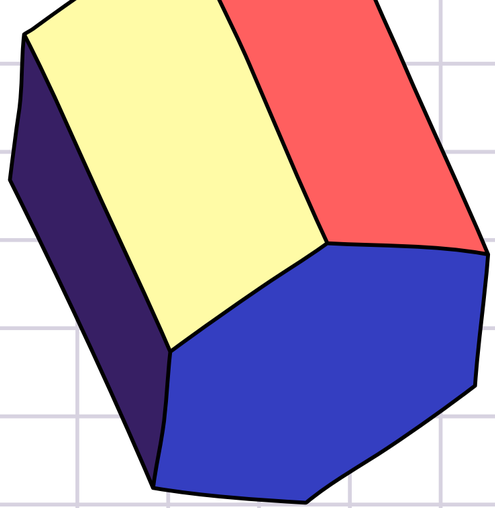
PYTHON

É uma linguagem de programação criada em 1991, mas sendo aprimorada continuamente por ser *open-source*, assim como o R.
Contudo, o Python é mais versátil, não tendo foco somente estatístico.
Com Python, em geral, temos mais possibilidades, principalmente no mundo da inteligência artificial.
É hoje a linguagem mais utilizada do mundo.



PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

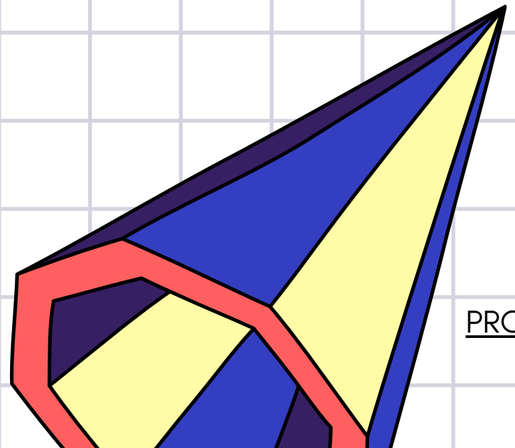


PYTHON

Com o Python, podemos:

- Trabalhar análises de dados
- Construir mapas
- Criar gráficos interativos
- Criar painéis e *dashboards*
- Gerar modelos estatísticos que buscam analisar um fenômeno
- Criar jogos
- Desenvolver softwares
- Trabalhar com modelos de *machine learning* e redes neurais
- Desenvolver robôs e automações de tarefas
- Testar sistemas de segurança

... Dentre outros

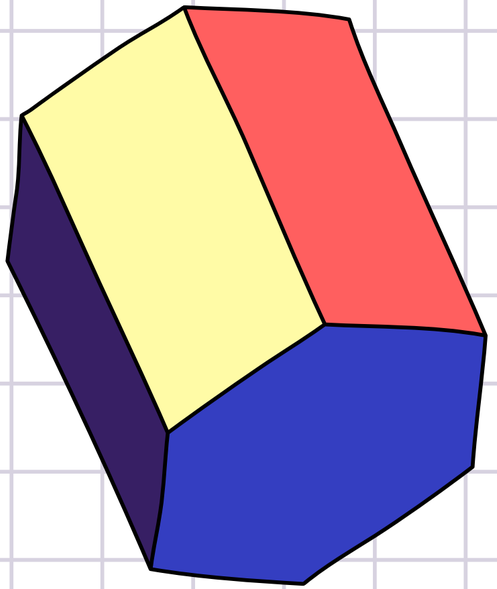


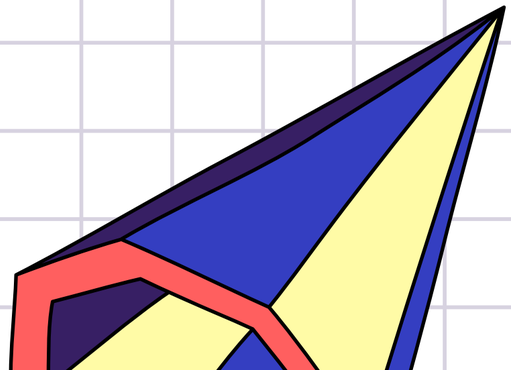
PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

GOOGLE COLAB

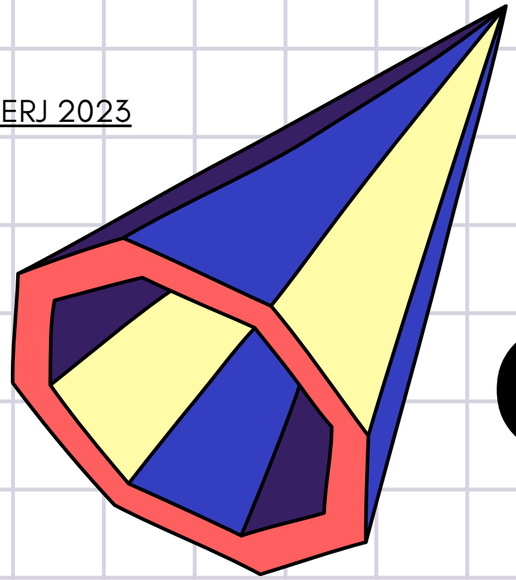
É uma plataforma de nuvem gratuita que oferece suporte ao Jupyter notebook e ao ambiente de execução em nuvem para máquinas virtuais, executando códigos em Python ou em R.
Os notebooks do Colab são armazenados no Google Drive e podem ser compartilhados facilmente com colegas de trabalho ou amigos.

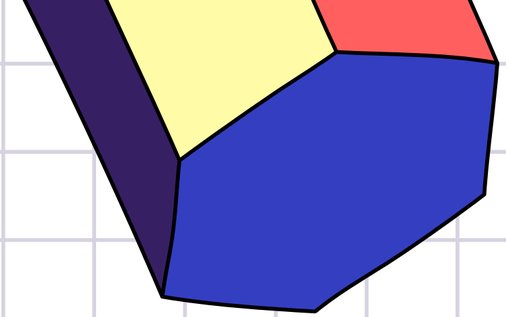




PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023





GOOGLE COLAB

O Google Colab é um *notebook*, que roda por trás um sistema denominado Jupyter, no qual é possível desenvolver e programar em Python por blocos de código, que rodam dentro de células. É como um relatório, no qual o resultado de um código é exibido abaixo dele. Isso nos permite desenvolver uma análise que pode ser reproduzida por qualquer um, obtendo-se os mesmos dados, sempre.

Com o Colab, também podemos aproveitar o uso de placas de vídeo poderosíssimas, que nos permitem um processamento mais rápido, sobretudo em tarefas que usam inteligência artificial.

PROF. MATHEUS C. PESTANA

ESCOLA DE INVERNO IESP-UERJ 2023

MÃOS À OBRA!

HTTPS://COLAB.RESEARCH.GOOGLE.COM

PROF. MATHEUS C. PESTANA