Tópicos

Avançados II

Aula 6

DATA SCIENCE & MODELAGEM EM TRANSPORTES

AUGUSTO UCHÔA

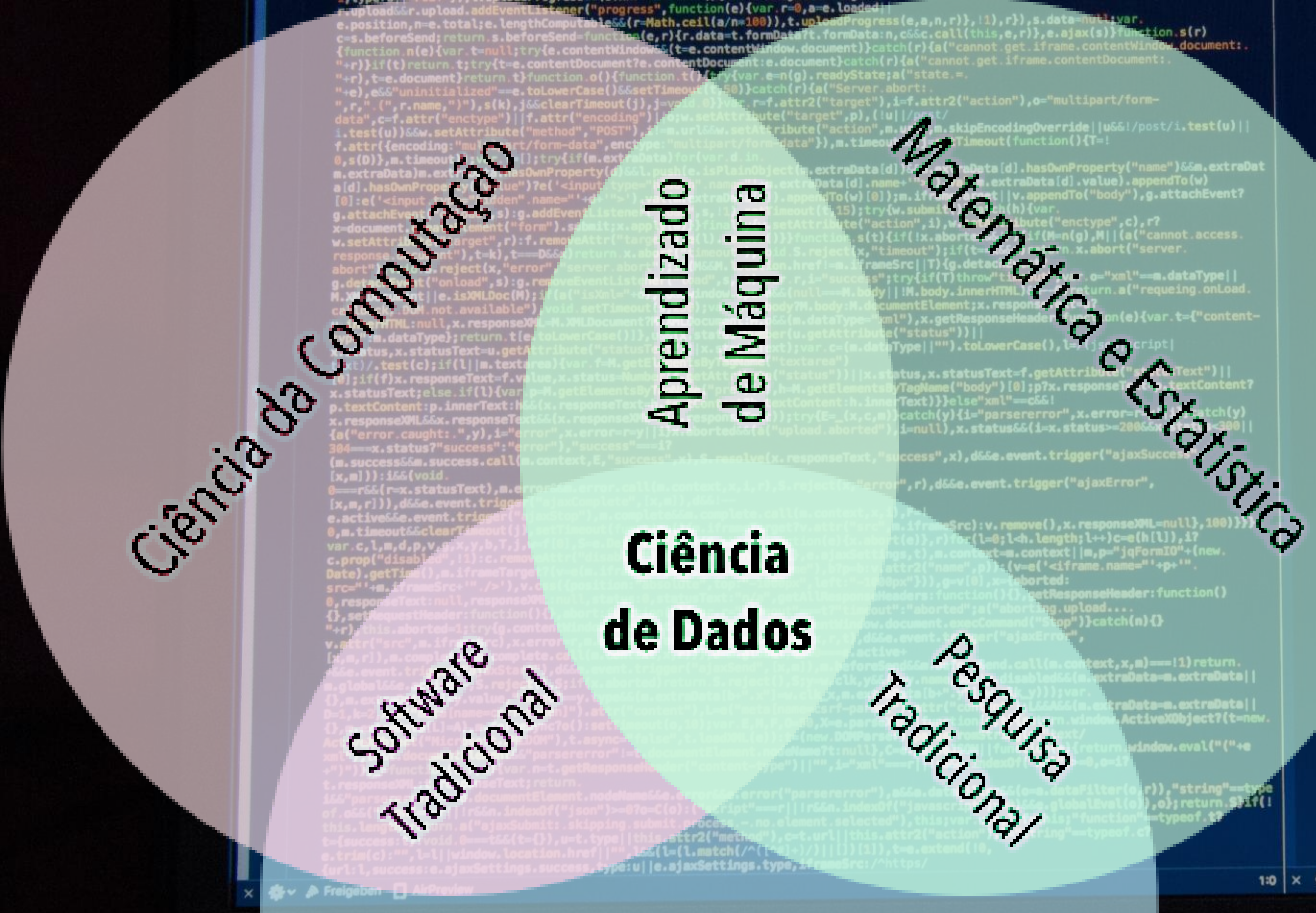
Petran- Programa de Pós-graduação em Engenharia de Transportes

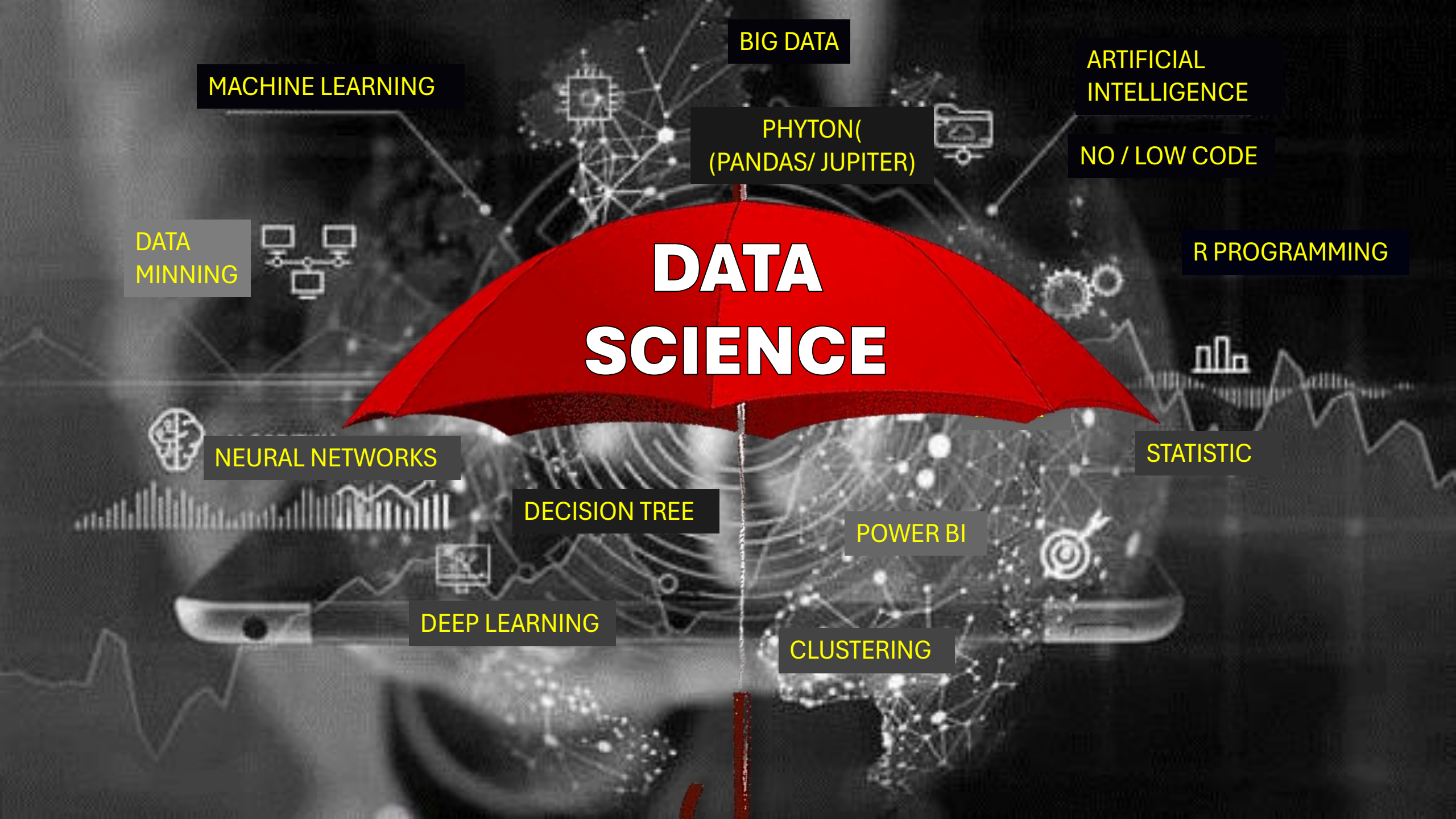
TRILHA DE HOJE:

1. Compreender a interrelação: Transpotation Engineering & Data Science & Machine Learning
2. Compreender os conceitos de: ciência de dados,dados. classificação dos dados, dados brutos, informação, conhecimento e BigData
3. Conhecer as exigências da modelagem neural me termos de características dos dados: formato, dimensão, representatividade, aleatoriedade, normalização e dummyzação
4. Perceber que se pode usar um chatbot alimentado por um modelo de linguagem natural para ajudar no pré-processamento dos seus dados

CIÊNCIA DE DADOS, O QUE É?

“É a área que utiliza diversas técnicas e ferramentas para extrair informação, valor dos dados e auxiliar na tomada de decisões.”




BIG DATA
MACHINE LEARNING
ARTIFICIAL INTELLIGENCE
PHYTON((PANDAS/ JUPITER)
NO / LOW CODE
DATA MINNING
R PROGRAMMING
DATA SCIENCE
NEURAL NETWORKS
STATISTIC
DECISION TREE
POWER BI
DEEP LEARNING
CLUSTERING

O QUE SÃO DADOS?

"São medidas ou fatos individuais, obtidos/coletados por meio de técnicas de mensuração ou por observações diretas/indiretas"

- "tudo que é observável/mensurável/"armazenável " de quaisquer formas, em qualquer meio pode ser considerado dado"

CLASSIFICAÇÃO DOS DADOS

ESTRUTURADOS

organizados, formatados, por exemplo, em formato tabular, o que facilita seu processamento em um banco de dados relacional

NÃO ESTRUTURADOS

desorganizados, sem formato específico, o que cria uma maior dificuldade em seu processamento, por exemplo, áudios, vídeos, imagens, emails e postagens em redes sociais

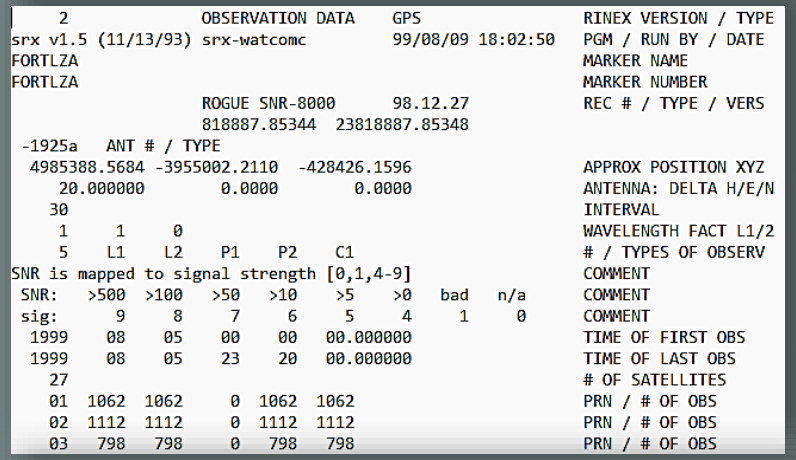
DADOS BRUTOS

São armazenados exatamente como coletados, sem nenhum tratamento anterior.

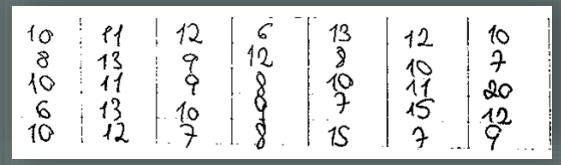
De onde eles vêm?

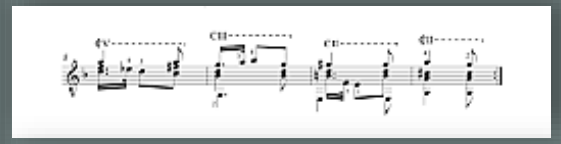
Sensores de todo tipo IOT, contadores de tráfego, imagens de acidentes, sensores deflectométricos em pavimentos, acelômetros, receptores gnss, coordenadas...pode ser qualquer coisa em qualquer formato.

Podem ser caracteres, números, imagens, vídeos, binários, de sorte que quantificações físicas podem ser transformadas em números.



4	6	4	6	2	6	3	6	3	3
3	2	3	5	2	3	1	0	3	3
2	5	1	5	1	0	3	1	2	1
0	5	6	0	4	2	6	2	4	6
1	1	2	3	3	4	2	2	0	4





DADO ≠ INFORMAÇÃO ≠ CONHECIMENTO

"Um dado sozinho, não tem significado, contudo, se estiver associado a outros dados, contextualizados e processados, aí sim, isso é informação, possui significado inteligível"

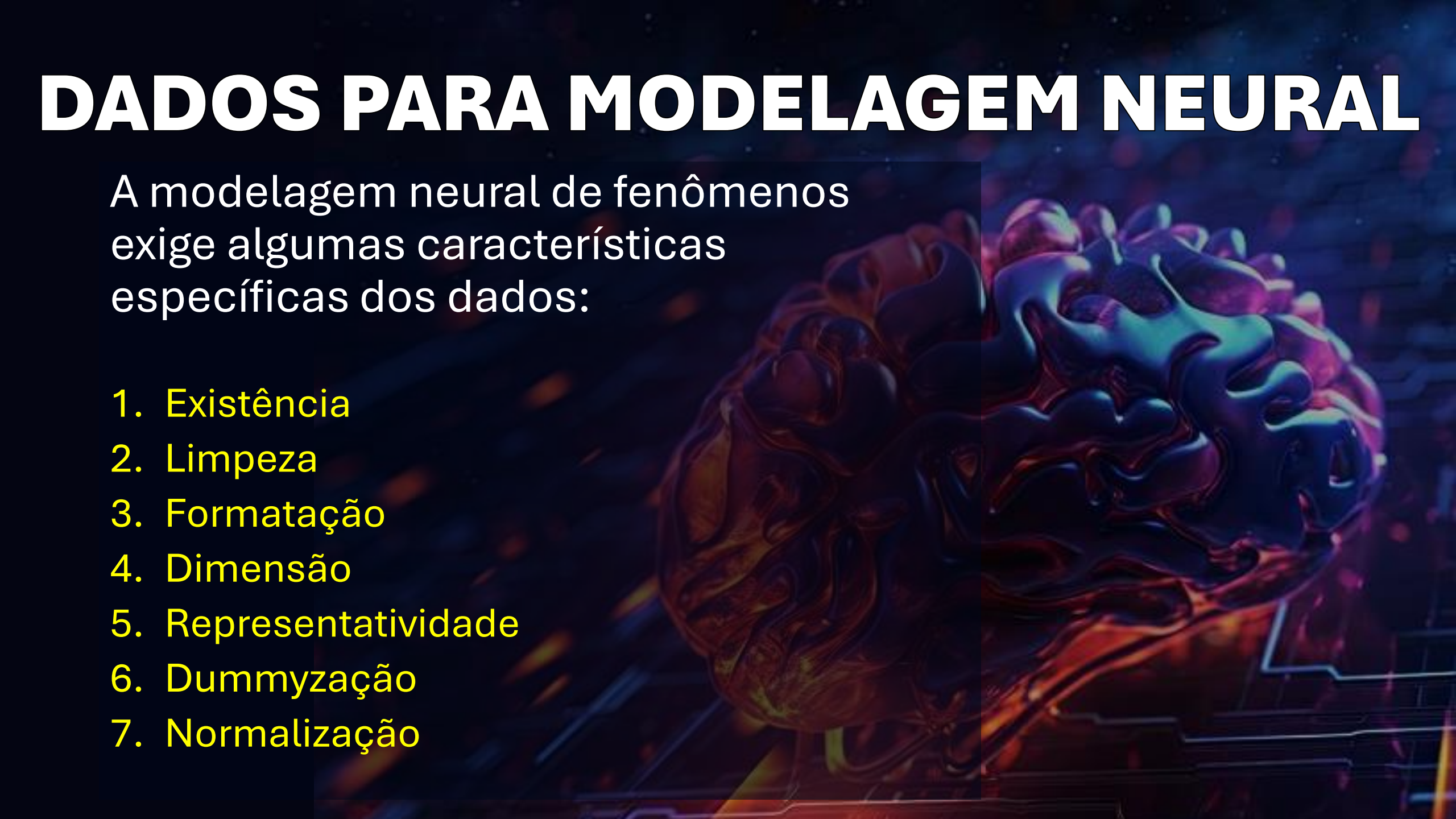
•Via de regra, os dados são coletados, armazenados e, *à posteriori*, processados e analisados a partir de diferentes técnicas, visando-se extrair informações e apoiar a tomada de decisões (isso é conhecimento). Informações são dados com relevância e propósito!

Conjuntos de dados, de diferentes formato, de tamanho extraordinário, colossais, massivos e devido a isso, necessitam de técnicas especiais de armazenamento e tratamento

DADOS PARA MODELAGEM NEURAL

A modelagem neural de fenômenos exige algumas características específicas dos dados:

1. Existência
2. Limpeza
3. Formatação
4. Dimensão
5. Representatividade
6. Dummyzação
7. Normalização



DÊ SEU JEITO!

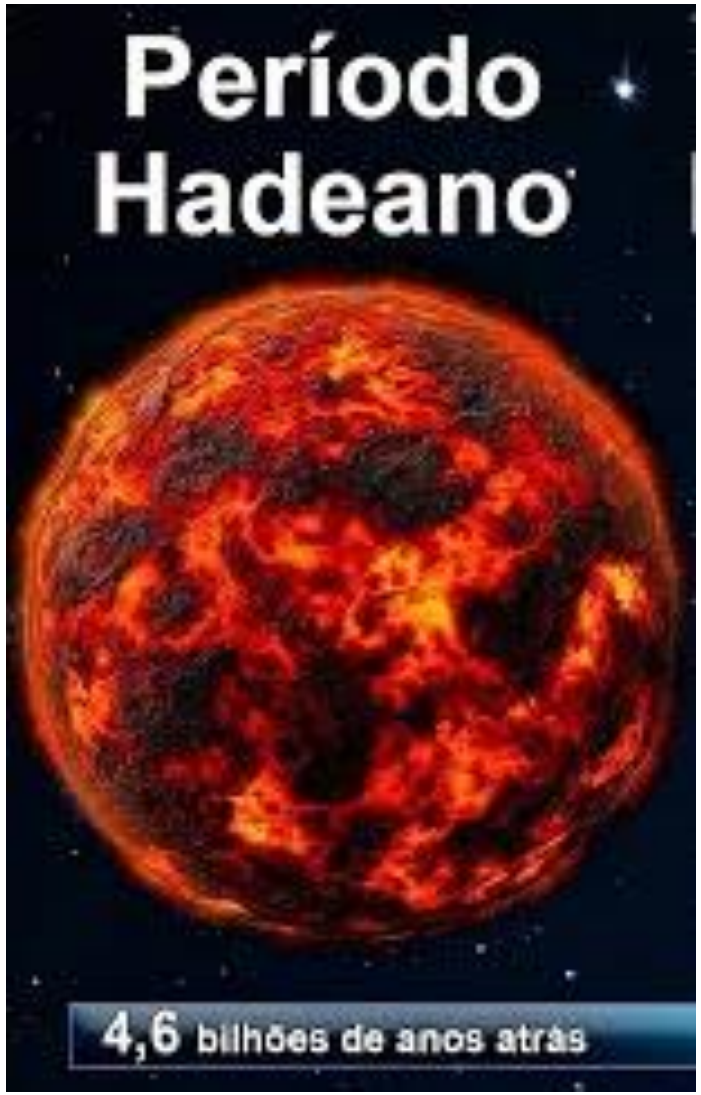
DADOS PRIMÁRIOS: use um método adequado para capturar todas as características de seu fenômeno, seja através de observação ou experimento

- **DADOS SECUNDÁRIOS:** certifique-se que a fonte é confiável e que o método usado para obtenção dos dados foi adequado, assim como seu recorte espaço-temporal

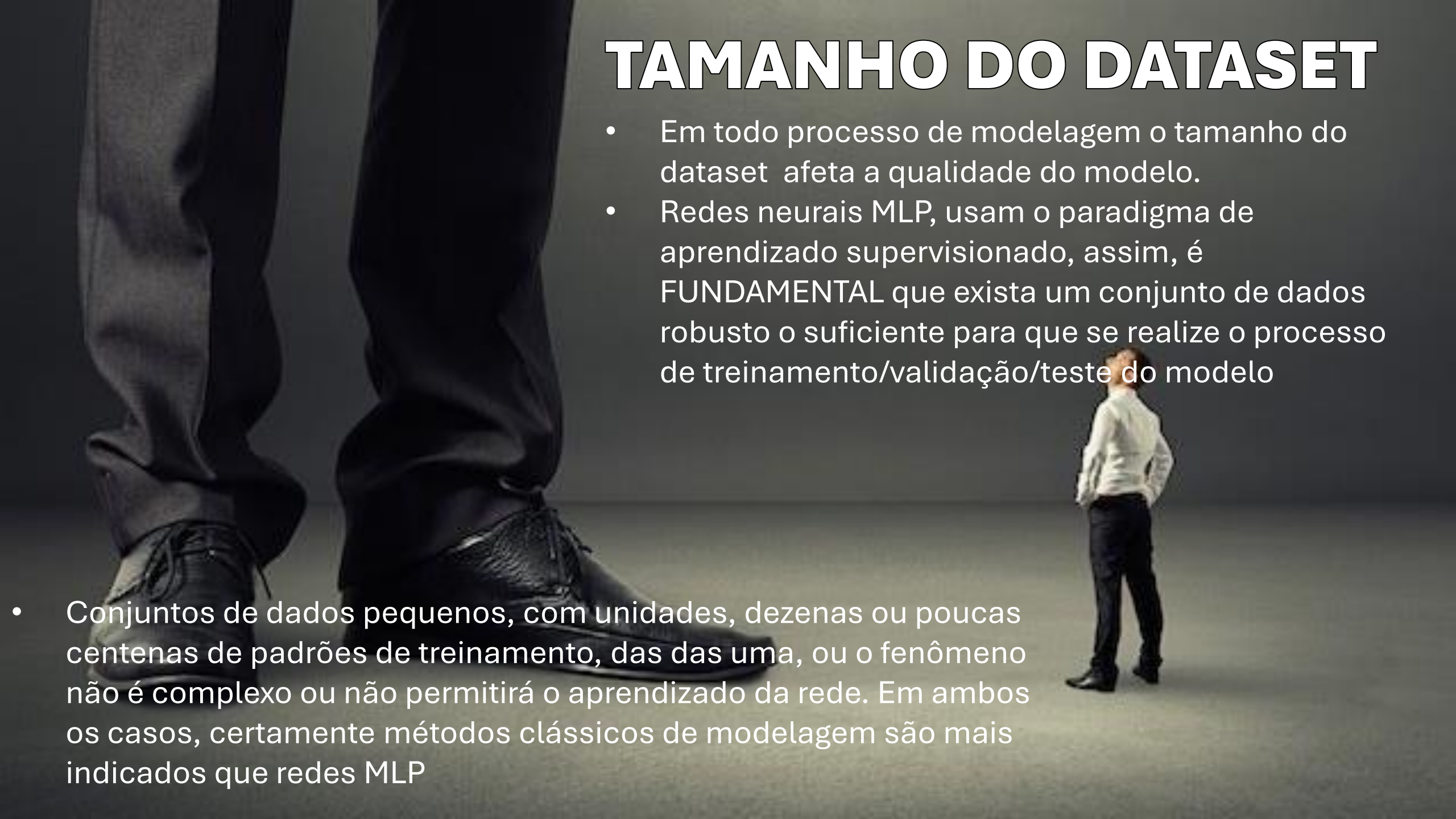
ORNANIZE-OS, LIMPE-OS

- Não basta existirem, os dados precisam estar estruturados, organizados e limpos (ruídos), mas tome cuidado com essa limpeza, quem disse que o que tu consideras sujeira não faz parte, de fato do fenômeno?

Se os seus dados foram coletados por algum órgão municipal, estadual ou federal, ou até alguma empresa privada, há uma chance enorme deles existirem na forma de relatórios ou formulários em papel, guardados em caixas porta arquivo, é urgente que sejam digitalizados e sem ampliar erros de escrita, transcrição e outros que já devem existir em um processo de coleta tão hadeano



Analógico nem mais as vovós do zap



TAMANHO DO DATASET

- Em todo processo de modelagem o tamanho do dataset afeta a qualidade do modelo.
- Redes neurais MLP, usam o paradigma de aprendizado supervisionado, assim, é FUNDAMENTAL que exista um conjunto de dados robusto o suficiente para que se realize o processo de treinamento/validação/teste do modelo

- Conjuntos de dados pequenos, com unidades, dezenas ou poucas centenas de padrões de treinamento, das das uma, ou o fenômeno não é complexo ou não permitirá o aprendizado da rede. Em ambos os casos, certamente métodos clássicos de modelagem são mais indicados que redes MLP

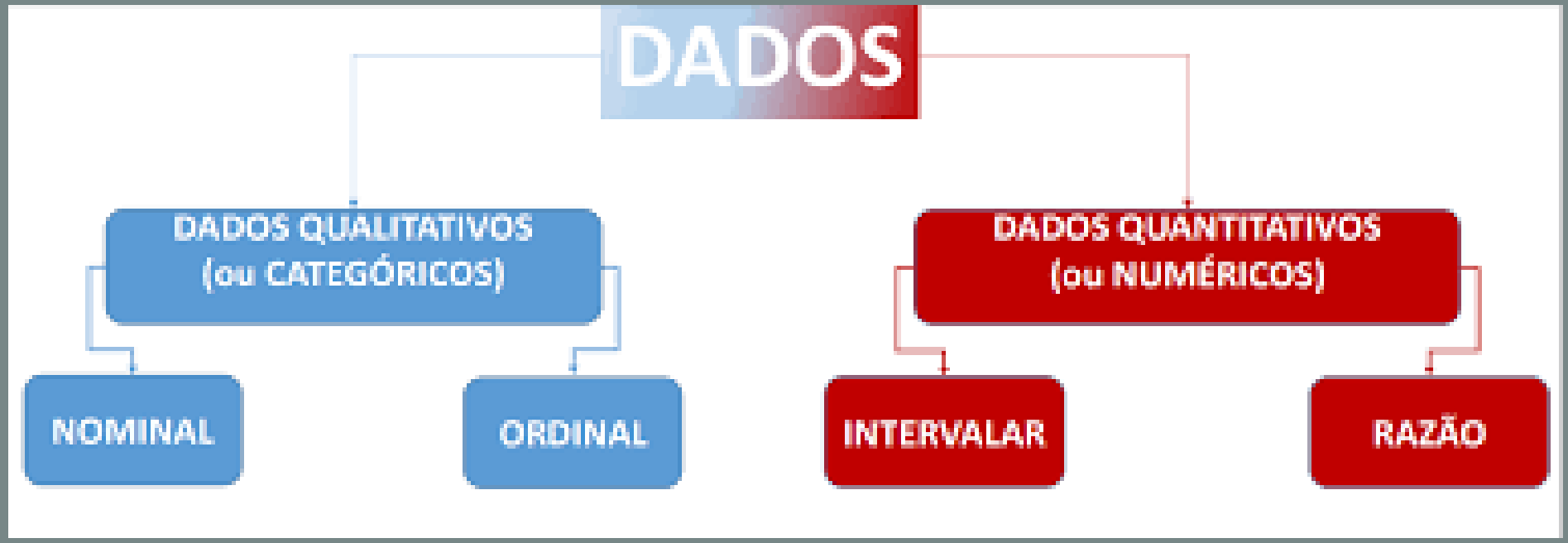
REPRESENTATIVIDADE

Tão importante quando o tamanho. Talvez até mais, é a REPRESENTATIVIDADE do dataset. Isso significa que os dados tem de capturar as principais características do fenômeno a ser modelado.
Equilíbrio de classes, recorte temporal e espacial são atributos desejáveis ao conjunto de dados para modelagem neural

- Por exemplo, pretende-se estimar as características geotécnicas dos solos do estado do Ceará e, quer seja por negligência, ignorância ou má fé,realiza-se um esforça amostral apenas na região metropolitana de Fortaleza, será que pode-se generalizar o modelo criado para todo o estado?

- Por exemplo, imaginem que se deseja modelar o comportamento de usuários de transporte coletivo do município de Fortaleza e, quer seja por negligência, ignorância ou má fé, coleta-se dados apenas de residentes no bairro de Meireles. Será que podemos extrapolar o mesmo comportamento para todos os outros bairros de Fortaleza?

DUMMY VARIABLES

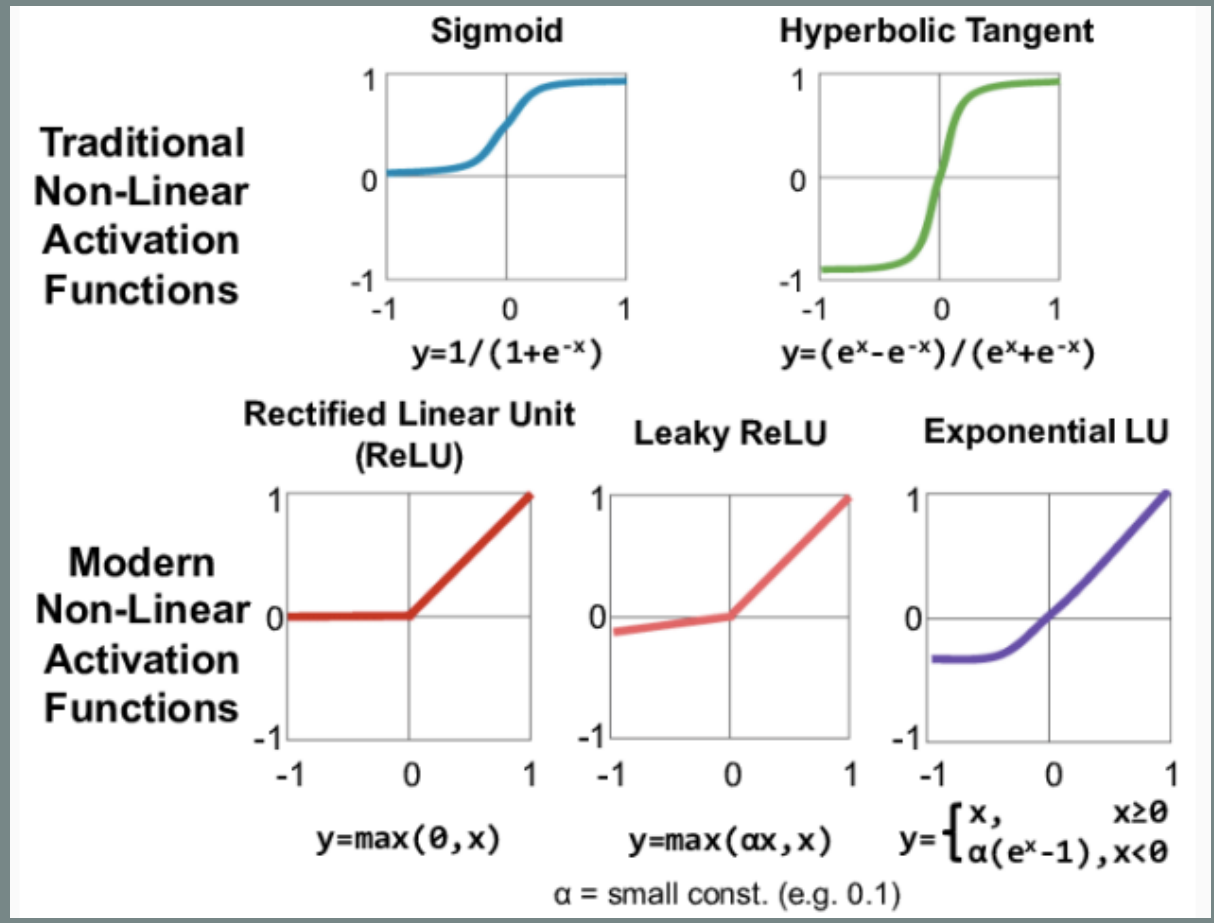


É um processo de converter dados categóricos, como cores ou tipos de animais, em números para que a rede neural possa entendê-los. Isso envolve atribuir um número único a cada categoria.

Por exemplo, se você estiver trabalhando com cores, pode atribuir o número 1 para vermelho, 2 para azul e assim por diante. Dessa forma, a rede neural pode usar esses números como entradas para aprender a partir dos dados categóricos.

NORMALIZAÇÃO

Sim, recomenda-se normalizar os dados, independente se usar uma função de ativação tradicional ou moderna
As funções de ativação sigmoid logística ou a tangente hiperbólica tem como saída apenas valores entre 0 e 1, valores altos neste tipo de função farão com que a saída do neurônio seja sempre 1, levando os pesos da sinapse a uma saturação, sendo assim, a rede nunca irá convergir.



A normalização ajuda a evitar problemas de explosão do gradiente e acelera o treinamento, tornando a otimização mais estável. No entanto, lembre-se de que a necessidade de normalização pode depender da natureza dos dados e do projeto específico, mas como regra geral, é uma boa prática realizar a normalização ao treinar redes neurais

COMO NORMALIZAR OS DADOS?

Normalização Z-Score (Padronização)

Transforma os dados para que tenham média zero e desvio padrão unitário:

$$x_{norm} = \frac{x - \mu}{\sigma}$$

- x_{norm} é o valor normalizado.
- x é o valor original.
- μ é a média dos valores no conjunto de dados.
- σ é o desvio padrão dos valores no conjunto de dados.

Normalização Min-Max

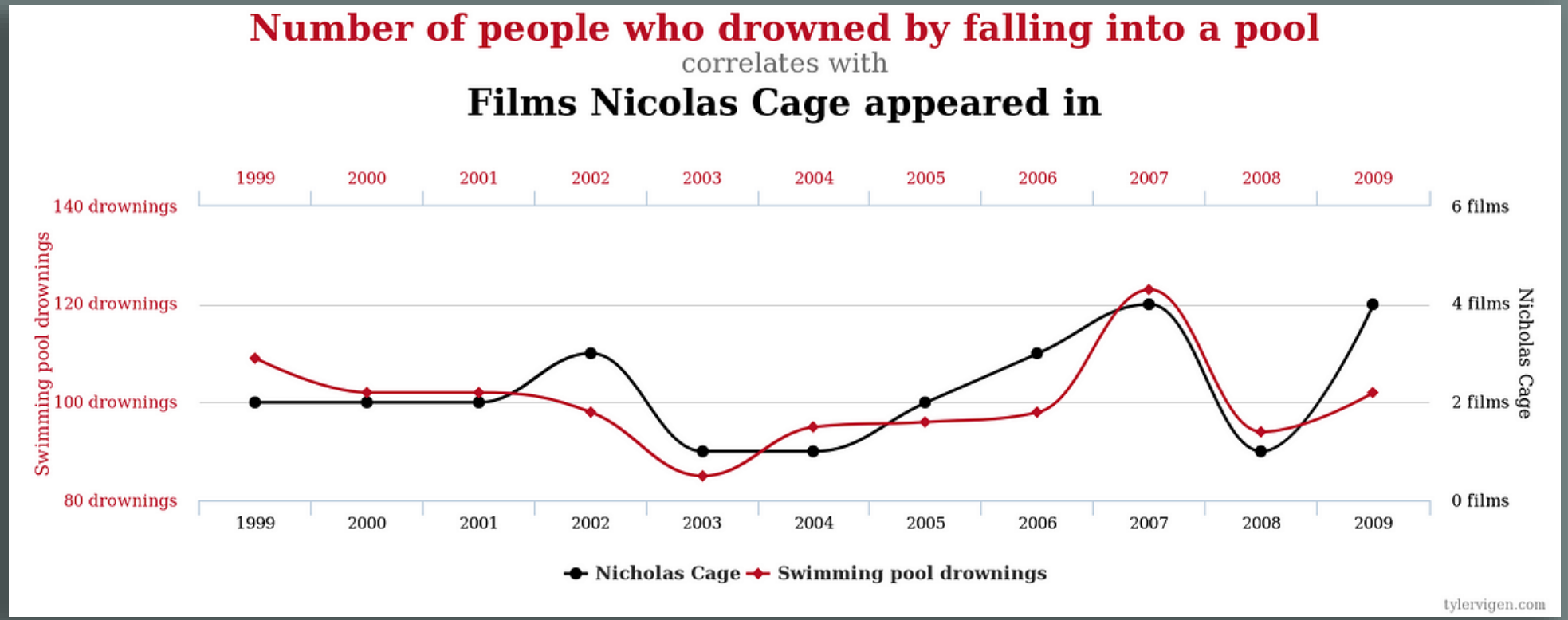
Dimensiona os dados para um intervalo específico, geralmente entre 0 e 1 ou -1 e 1:

$$x_{norm} = \frac{x - \min(X)}{\max(X) - \min(X)}$$

- x_{norm} é o valor normalizado.
- x é o valor original.
- $\min(X)$ é o valor mínimo no conjunto de dados.
- $\max(X)$ é o valor máximo no conjunto de dados.

ANÁLISE EXPLORATÓRIA DAS VARIÁVEIS

Antes do pré-processamento dos dados, já deve ter sido realizada uma análise estatística das variáveis disponíveis no Dataset, a fim de se perceber indícios de causalidade, bem como de multicolinearidade. Além disso é importante lembrar que:



CROSS-VALIDATION

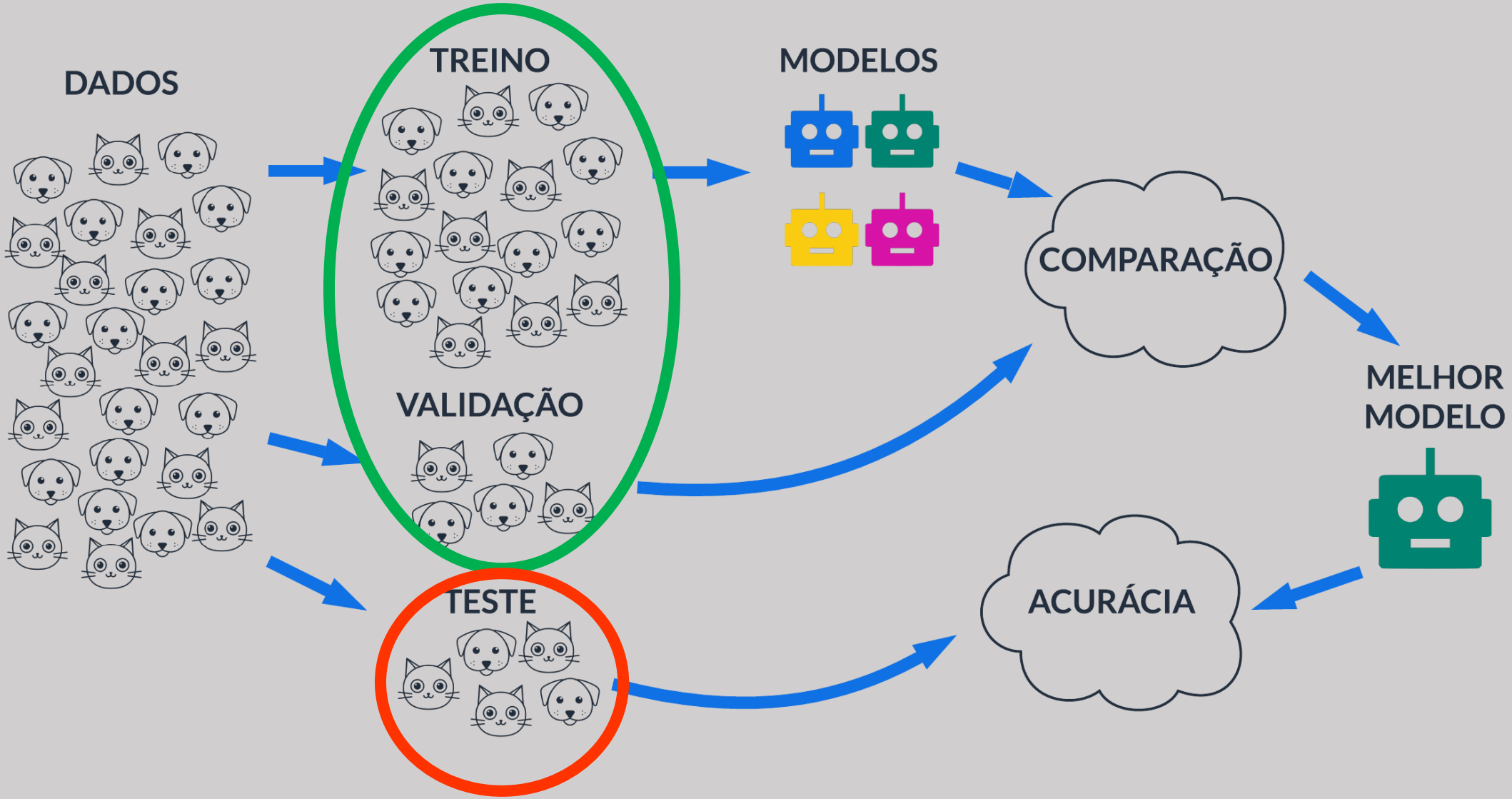
é uma técnica usada na estatística e no aprendizado de máquina para avaliar o desempenho de um modelo estatístico ou de aprendizado de máquina, especialmente quando se trabalha com conjuntos de dados limitados. Ela ajuda a estimar quão bem o modelo generaliza a partir dos dados de treinamento para novos dados não vistos.

A ideia principal da validação cruzada é dividir o conjunto de dados em duas partes: uma parte para treinamento e outra para teste. A divisão é feita de maneira que o modelo seja treinado em uma parte dos dados e testado na outra. No entanto, a validação cruzada vai além de uma única divisão; ela repete esse processo várias vezes, cada vez com uma divisão diferente dos dados. O resultado é uma avaliação mais robusta do desempenho do modelo.

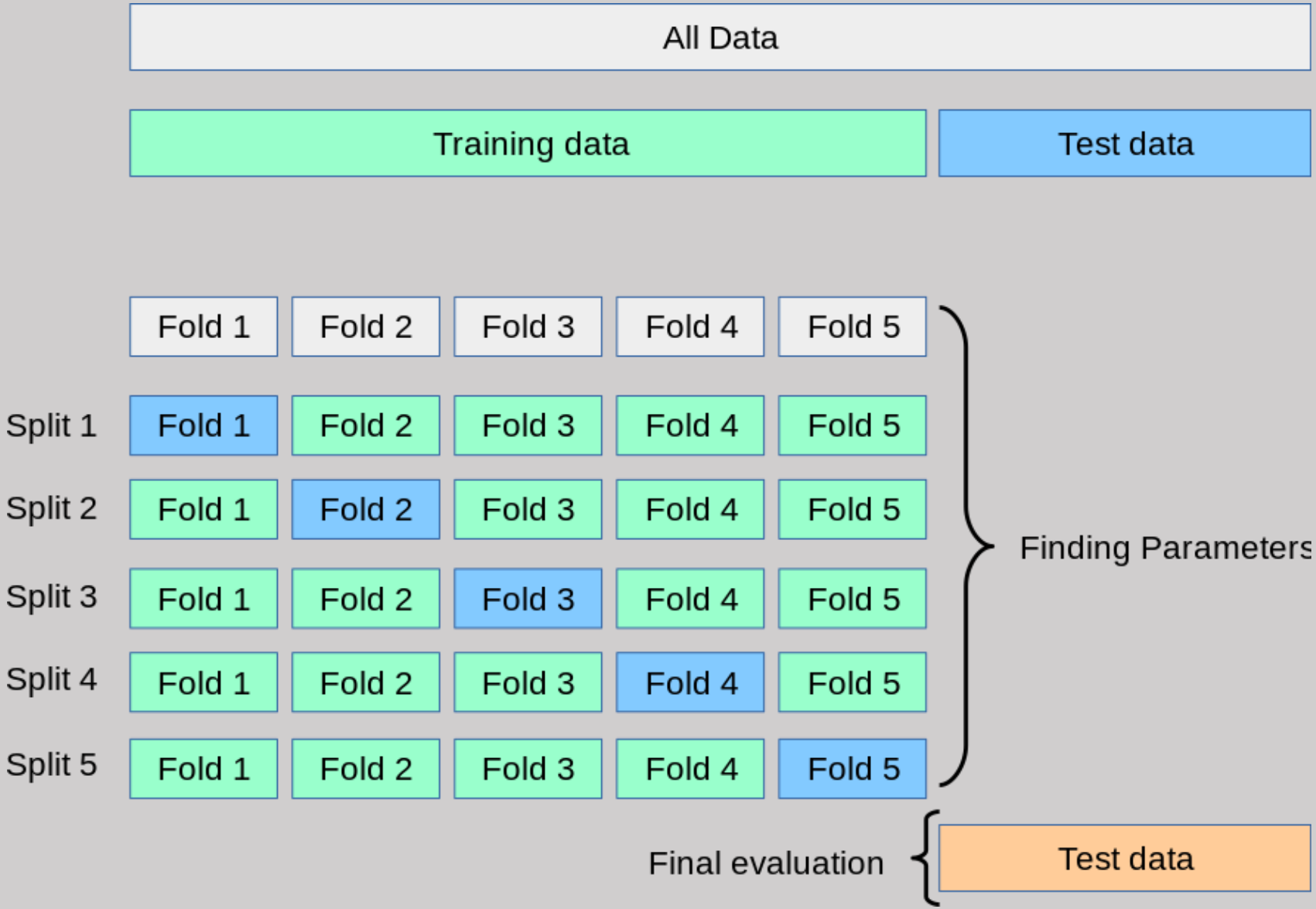
Existem várias técnicas de validação cruzada, sendo a validação cruzada k-fold uma das mais comuns.

A validação cruzada é particularmente útil para detectar problemas de superajuste (overfitting), pois força o modelo a generalizar a partir de diferentes partes dos dados. Ela também ajuda a evitar a dependência da divisão inicial dos dados, que pode afetar os resultados da validação.

DADOS DE TREINAMENTO, VALIDAÇÃO E TESTE



VALIDAÇÃO CRUZADA: MÉTODO K-FOLD



PULGA ATRÁS DA ORELHA?

1. Esse negócio tá ficando muito enrrolado!
2. Será que consigo fazer esse pré-processamento dos meus dados?
3. ????

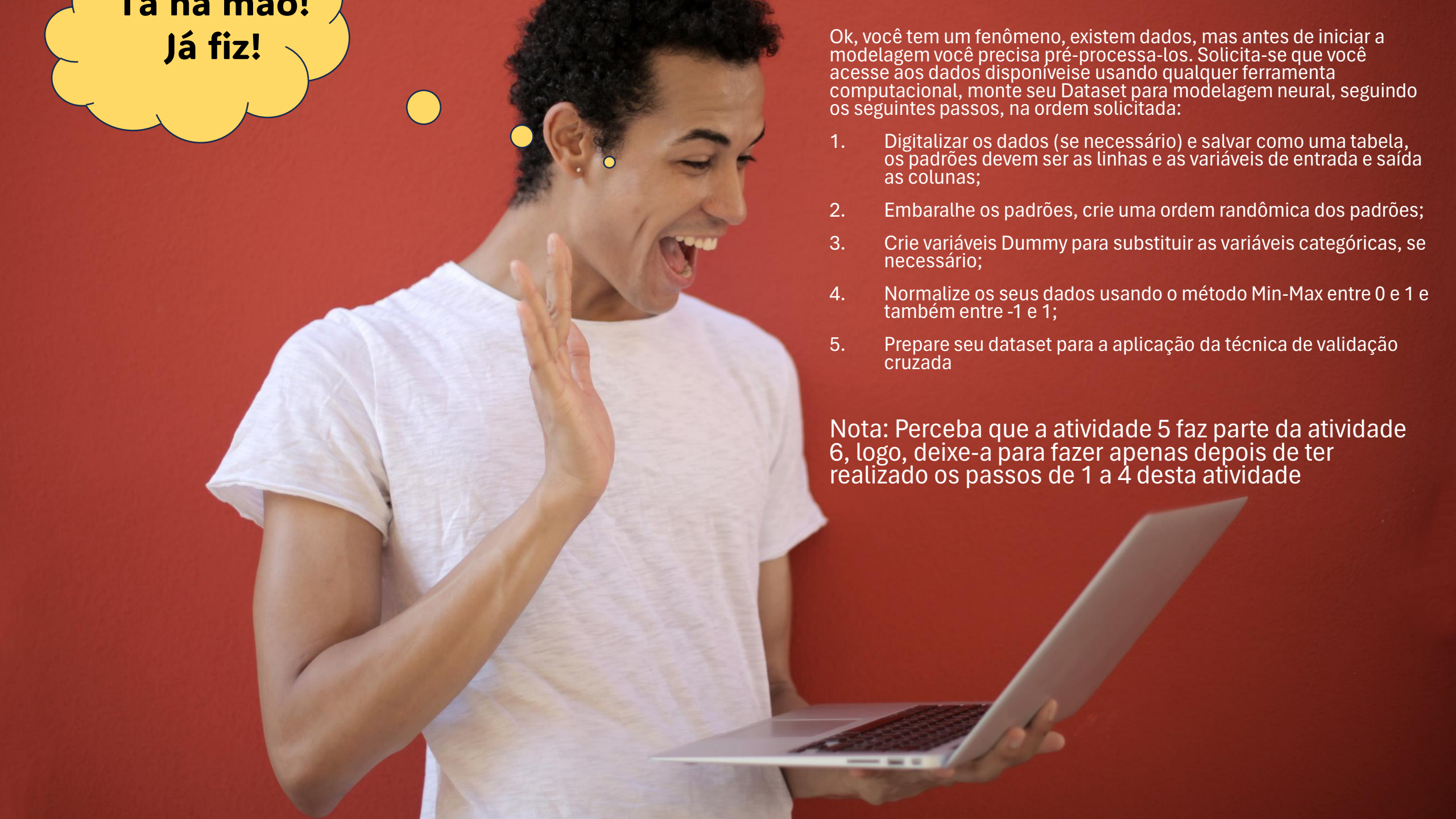
PONTOS CHAVE

A interrelação entre enga de transportes & Ciência de dados e Aprendizado de máquina

Conceitos de ciêcia de dados, dado, informação, conhecimento, classificação dos dados e Bigdata

Principais características dos dados para modelagem neural

O pré-processamento dos dados para modelagem neural



Ok, você tem um fenômeno, existem dados, mas antes de iniciar a modelagem você precisa pré-processa-los. Solicita-se que você acesse aos dados disponíveise usando qualquer ferramenta computacional, monte seu Dataset para modelagem neural, seguindo os seguintes passos, na ordem solicitada:

1. Digitalizar os dados (se necessário) e salvar como uma tabela, os padrões devem ser as linhas e as variáveis de entrada e saída as colunas;
2. Embaralhe os padrões, crie uma ordem randômica dos padrões;
3. Crie variáveis Dummy para substituir as variáveis categóricas, se necessário;
4. Normalize os seus dados usando o método Min-Max entre 0 e 1 e também entre -1 e 1;
5. Prepare seu dataset para a aplicação da técnica de validação cruzada

Nota: Perceba que a atividade 5 faz parte da atividade 6, logo, deixe-a para fazer apenas depois de ter realizado os passos de 1 a 4 desta atividade


A SEGUIR, CENAS DO
PRÓXIMO
CAPÍTULO
Como usar o ChatGPT como co-piloto em minha modelagem neural?