



MACHINE LEARNING: COLETA DE DADOS

André Gustavo Adami
Daniel Luis Notari

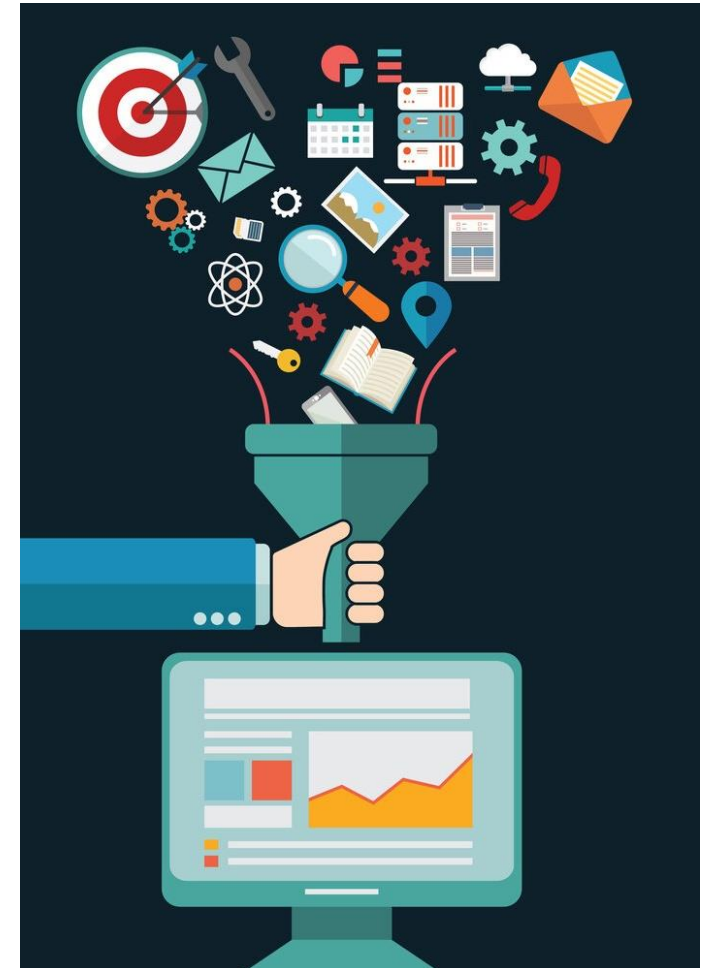
INTRODUÇÃO

Os dados são a base para a construção de soluções de aprendizado de máquina

A qualidade do aprendizado depende da qualidade dos dados

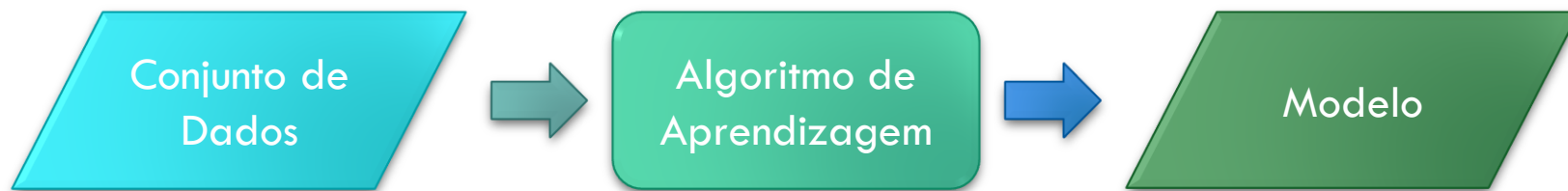
Dados com problemas podem levar a resultados com problemas

Garbage in, garbage out



CONJUNTO DE DADOS

O processo de aprendizagem tem por objetivo produzir um **modelo** matemático do domínio com base em um **conjunto finito de dados (dataset)**



Esse conjunto finito de dados X é uma coleção de **medidas** ou **observações (padrões)** de algum fenômeno

$$X = \{x_1, x_2, \dots, x_N\}$$

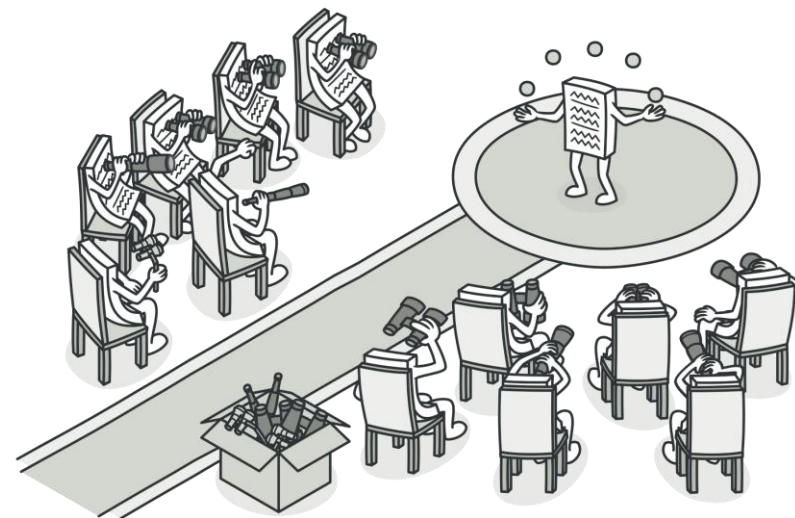
CONJUNTO DE DADOS

Cada entrada x_i no conjunto de dados pode ser definida como um vetor multidimensional com d dimensões

$$x_i = \{x_{i1}, x_{i2}, \dots, x_{id}\}$$

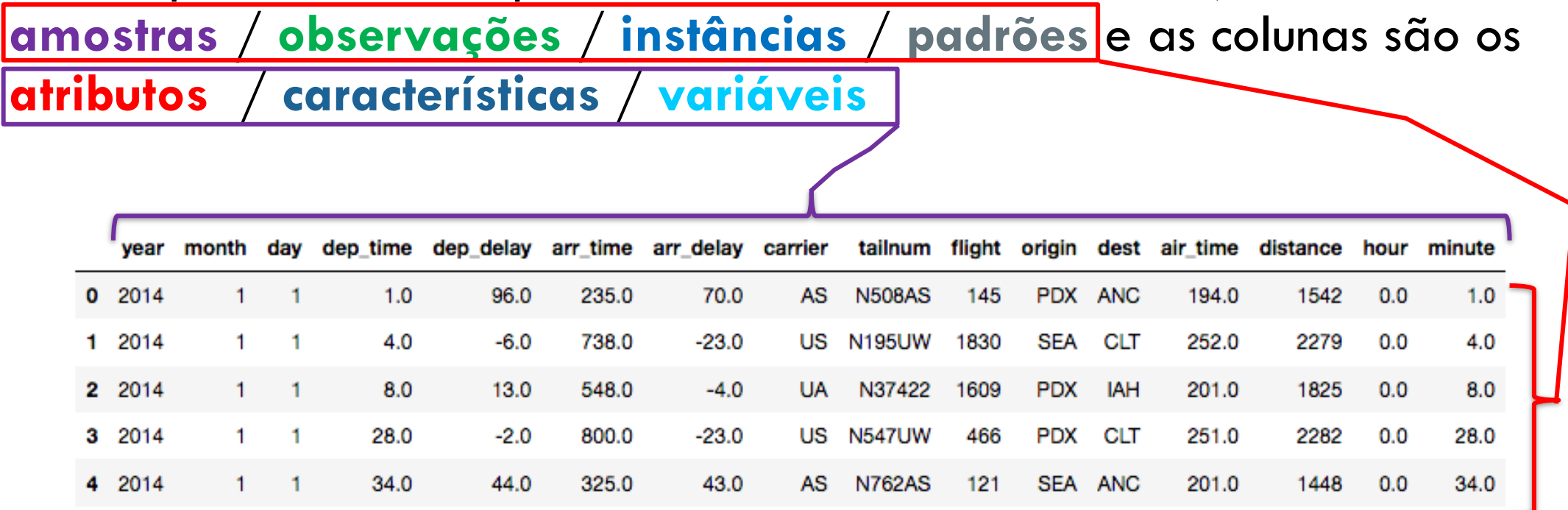
onde cada dimensão representa um **atributo**, **característica** (*feature*) ou **variável**

Estas entradas são também conhecidas por **amostras** ou **instâncias**



CONJUNTO DE DADOS

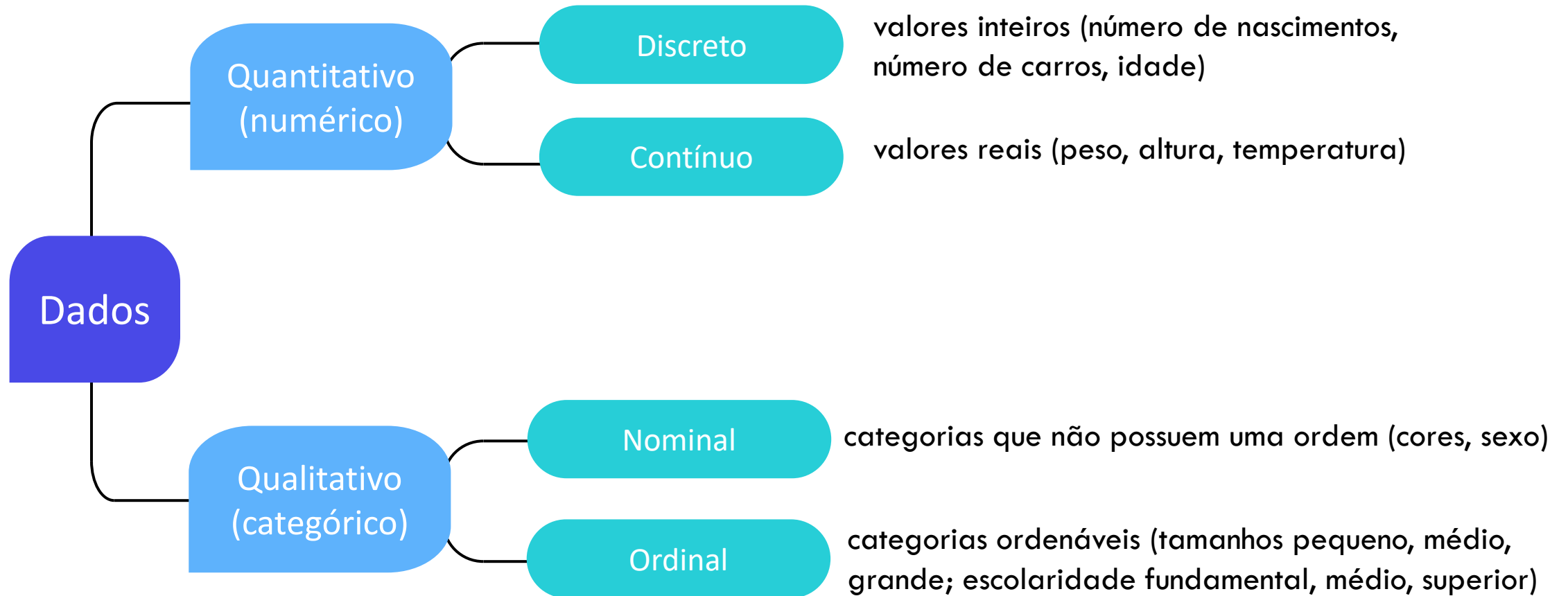
Um conjunto de dados pode ser visto como uma tabela, onde as linhas são as **amostras** / **observações** / **instâncias** / **padrões** e as colunas são os **atributos** / **características** / **variáveis**



	year	month	day	dep_time	dep_delay	arr_time	arr_delay	carrier	tailnum	flight	origin	dest	air_time	distance	hour	minute
0	2014	1	1	1.0	96.0	235.0	70.0	AS	N508AS	145	PDX	ANC	194.0	1542	0.0	1.0
1	2014	1	1	4.0	-6.0	738.0	-23.0	US	N195UW	1830	SEA	CLT	252.0	2279	0.0	4.0
2	2014	1	1	8.0	13.0	548.0	-4.0	UA	N37422	1609	PDX	IAH	201.0	1825	0.0	8.0
3	2014	1	1	28.0	-2.0	800.0	-23.0	US	N547UW	466	PDX	CLT	251.0	2282	0.0	28.0
4	2014	1	1	34.0	44.0	325.0	43.0	AS	N762AS	121	SEA	ANC	201.0	1448	0.0	34.0

Dados que podem ser organizados em tabelas são chamados de **estruturados**

TIPOS DE DADOS



TIPOS DE DADOS

Dados coletados de um conjunto de crianças de uma escola

Cada amostra é composta pelos atributos

- Sexo: valor nominal (f ou m)
- Idade: valor discreto
- Altura: valor contínuo
- Peso: valor contínuo

Sexo	Idade (meses)	Altura (polegadas)	Peso (libras)
f	165	55.5	67.0
f	163	56.5	84.0
f	171	63.0	84.0
f	193	59.8	115.0
f	169	61.5	85.0
m	146	57.5	90.0
m	151	66.3	117.0
m	153	60.0	84.0
m	151	61.0	81.0
m	193	66.3	133.0

CONJUNTO DE DADOS

No caso de classes de problema de **aprendizado supervisionado**, para cada amostra deve existir a saída esperada

Problemas de classificação possuem uma saída do tipo categórica, chamada de **rótulo** (*label*) ou **classe** (daí o nome classificação)

sex	age	Time	Number of Warts	Type	Area	Induration diameter	Treatment Result
2	19	6	2	1	225	8	1
2	32	12	6	3	35	5	0
2	33	6,25	2	1	30	3	1
2	17	5,75	12	3	25	7	1
2	15	1,75	1	2	49	7	0
2	34	11,5	12	1	25	50	0
2	20	7,75	18	3	45	2	1

Problemas de regressão possuem uma saída do tipo numérica

age	sex	bmi	children	smoker	region	charges
19	female	27,9	0	yes	southwest	\$ 16.884,92
28	male	33	3	no	southeast	\$ 4.449,46
33	male	22,705	0	no	northwest	\$ 21.984,47
31	female	25,74	0	no	southeast	\$ 3.756,62
46	female	33,44	1	no	southeast	\$ 8.240,59
37	female	27,74	3	no	northwest	\$ 7.281,51
37	male	29,83	2	no	northeast	\$ 6.406,41

CARACTERÍSTICAS DE DADOS “BONS”

Acurácia: medidas e características devem refletir corretamente o que está sendo observado

Relevância: deve relacionar diretamente ao fenômeno sendo estudado

Representativo: dados devem ser escolhidos apropriadamente para refletir o que está sendo estudado

Bem definido: o significado dos dados devem ser definido de forma não ambígua

Completo: dados devem incluir todas as medidas e características potencialmente relevantes

Granular: dados devem ter um intervalo e detalhe suficiente para capturar a variabilidade dos dados

DADOS NÃO ESTRUTURADOS

A maioria dos dados disponíveis são não estruturados, isto é, não possuem uma organização que permita que seja aplicado diretamente algoritmos de aprendizagem

- Texto (blogs, tweets, e-mails, mensagens, documentos, ...)
- Sinais de sensores (imagem, áudio, vídeo, acelerômetro, ...)

Conjunto de dados de produções linguísticas são chamados de **corpus** (plural **corpora**)

Dados não estruturados devem ser convertidos em dados estruturados (dados que podem ser armazenados de forma tabular)

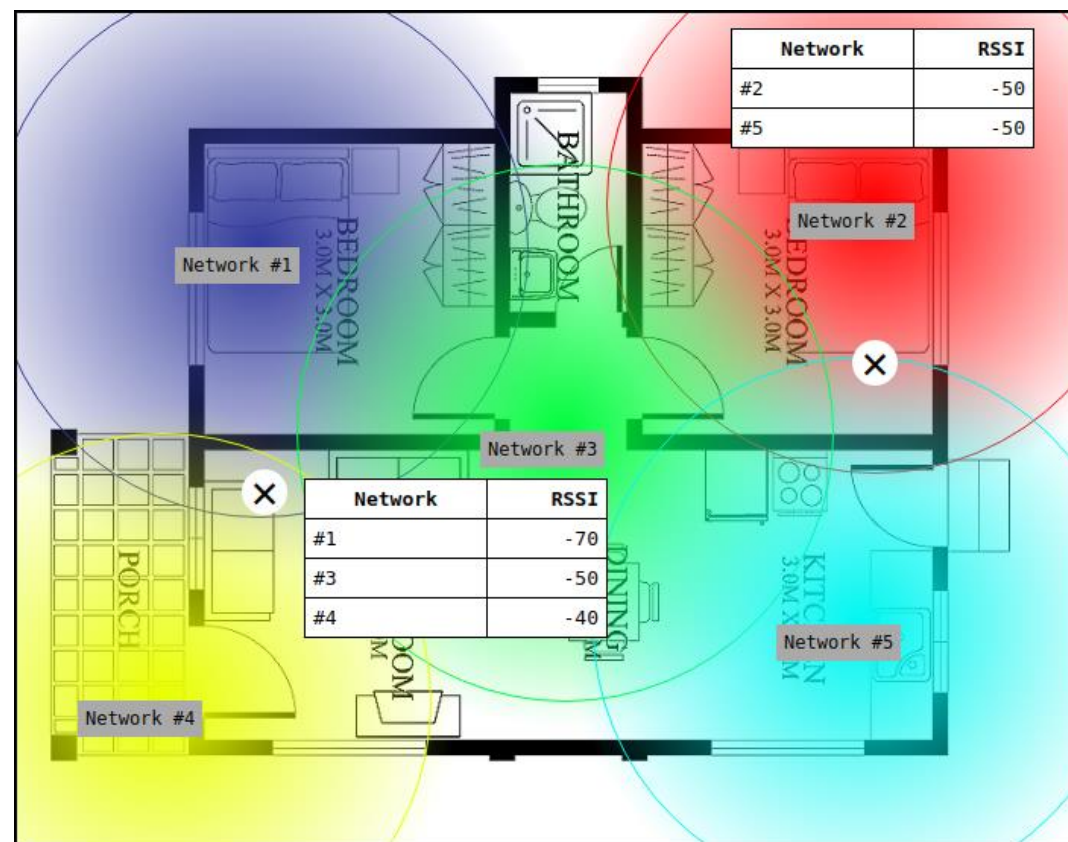


DADOS NÃO ESTRUTURADOS

Em determinadas situações, a conversão de dados não estruturados em estruturados é realizado por meio do rearranjo dos dados

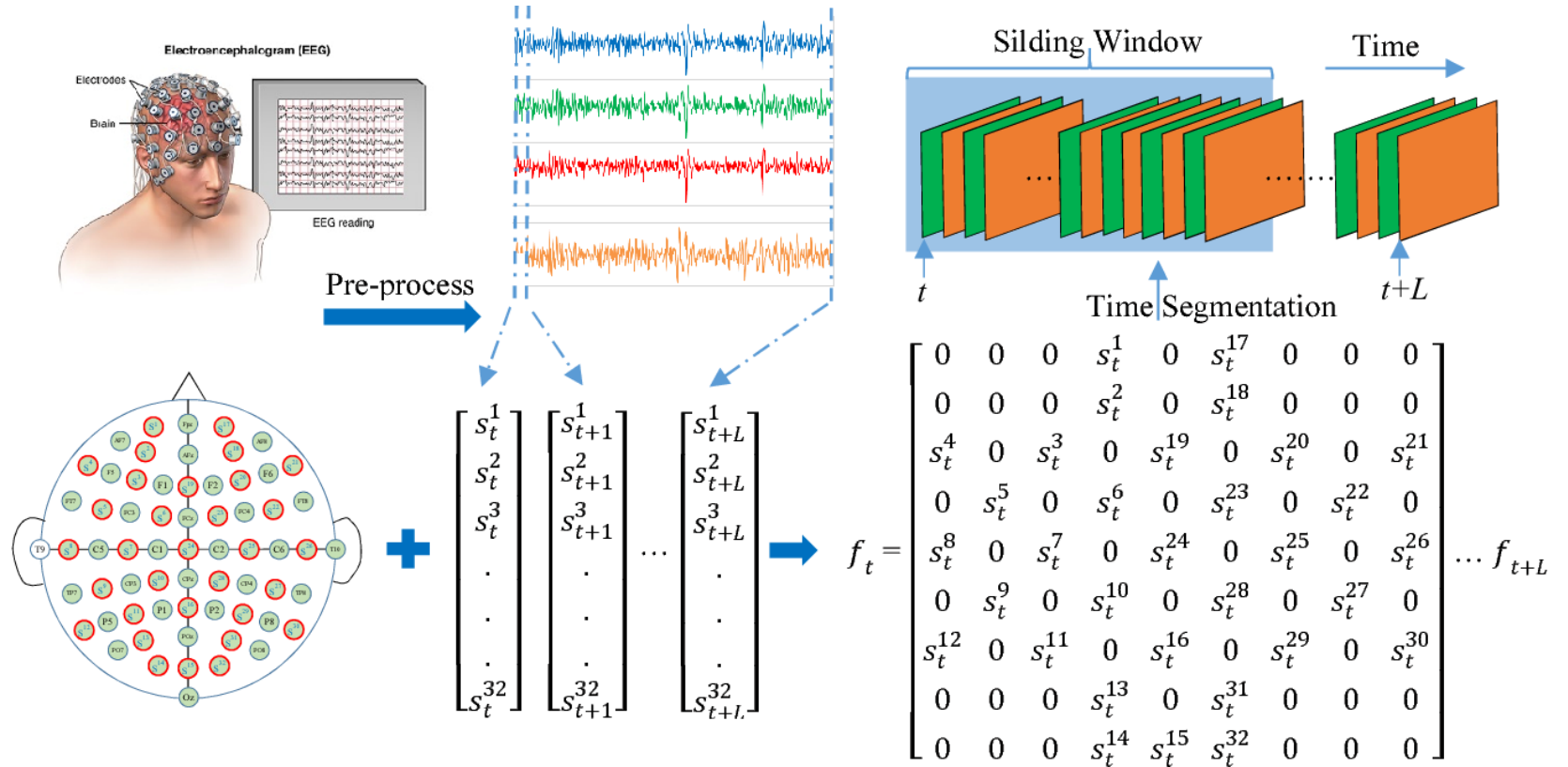
Localização ou posicionamento indoor utiliza o **indicador de intensidade do sinal recebido** (RSSI) dos pontos de acesso ou roteadores para determinar a localização de uma pessoa ou objeto

- O conjunto de RSSIs dos ponto de acesso ou roteador produz uma amostra multidimensional em um instante de tempo t



DADOS NÃO ESTRUTURADOS

Em reconhecimento de emoções baseado em sinais de encefalografia, a cada instante de tempo t , as amostras de 32 canais de eletrodos são organizadas de forma matricial para produzir uma amostra bidimensional



Y. Yang, Q. Wu, M. Qiu, Y. Wang and X. Chen, "Emotion Recognition from Multi-Channel EEG through Parallel Convolutional Recurrent Neural Network," *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, 2018, pp. 1-7.

DADOS NÃO ESTRUTURADOS

Em outras situações, os dados passam por uma transformação tal que uma nova representação é produzida

Esta transformação, conhecida por **extração de características**, produz uma versão dos dados mais compacta (e estruturada) e mais relevante ao problema

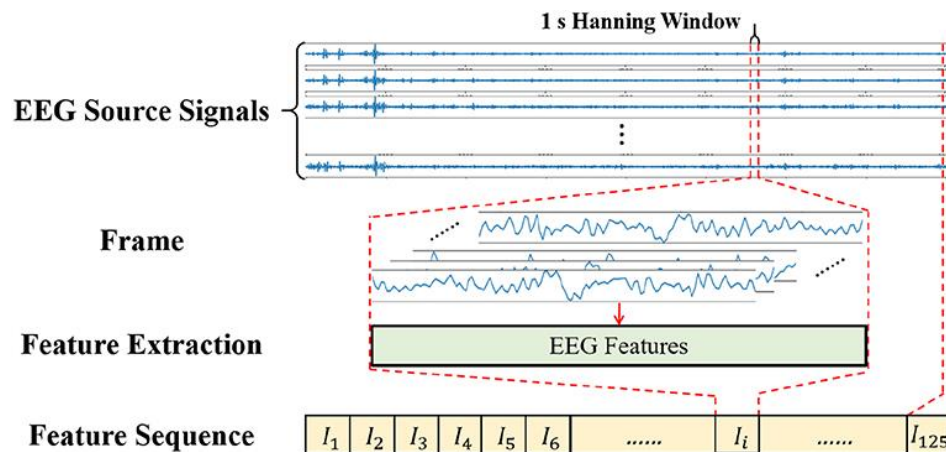
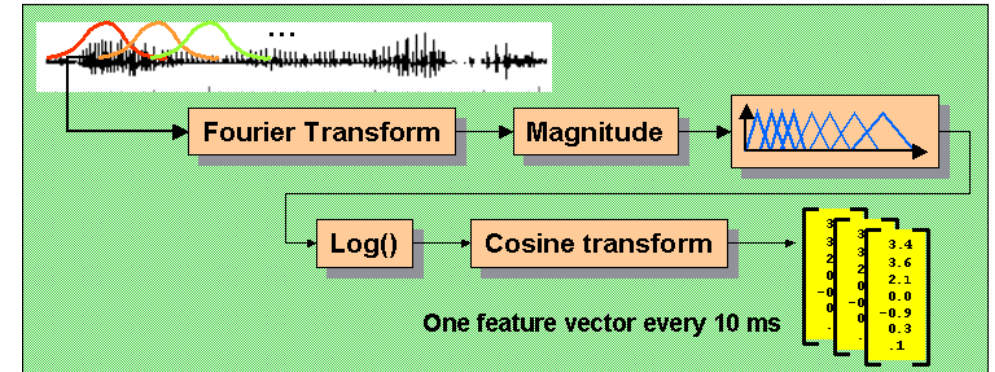
Alguns métodos, como as redes neurais convolucionais, realizam o processo de extração de características e classificação

DADOS NÃO ESTRUTURADOS

Sinais de voz são amostrados a 8 ou 16 kHz (8000 a 16000 amostras por segundo)

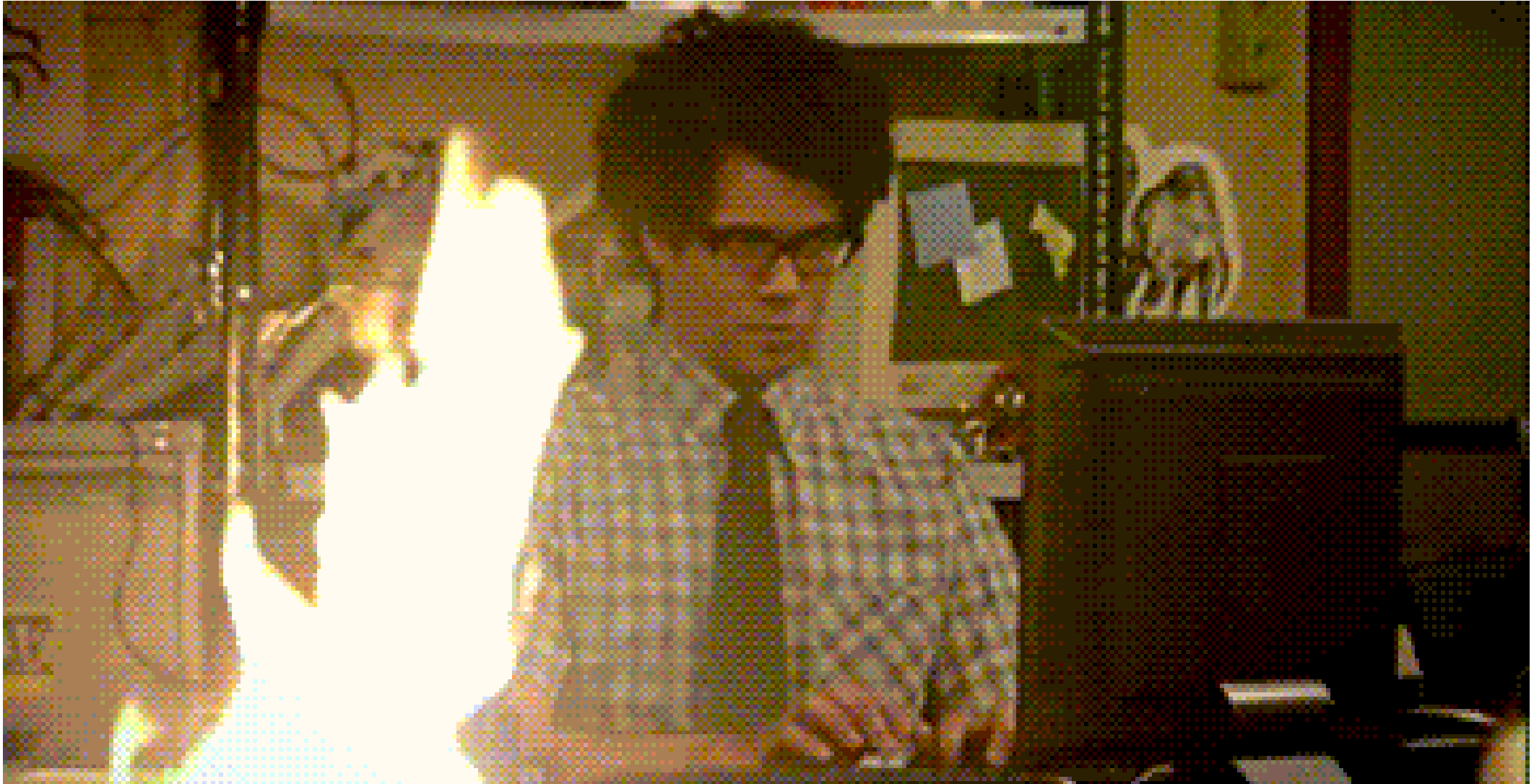
Além da quantidade de dados gerada, informações irrelevantes (silêncio, ruído de fundo / canal) são capturadas

A extração de características é realizada periodicamente produzindo amostras relevantes para o problema



Abordagem similar é realizada em sinais de eletroencefalografia a fim de capturar características das ondas cerebrais mais relevantes para o reconhecimento de emoções

PREOCUPADO COM DADOS?

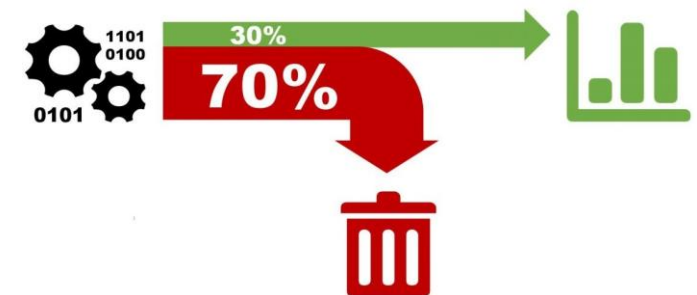


PREOCUPADO COM DADOS?

Em algumas situações, os dados já estão disponíveis, mas **não utilizados**

- Deve-se entender como é o processo de aquisição a fim de avaliar a qualidade e se adquire as informações necessárias (e.g., sensor trabalha em baixas frequências enquanto que a necessidade é de altas frequências)

Um estudo realizado pela AT Kearney em conjunto com o Fórum Mundial Econômico mostra que 70% dos dados de produção coletados por indústrias não são utilizados para gerar algum valor



DE ONDE VEM DADOS “RUINS”

Erro de projeto de coleta: pesquisa, experimentos ou instrumentação mal projetada, especificações erradas dos tipos de dados ou formatos e falta de uma definição clara dos dados



Erros de coleta: falta de um processo claro de coleta, falta de monitoramento e verificação do processo, falta de checagem ou validação dos dados coletados, manipulação tendenciosa dos dados, erros de gravação (falta de padronização, representação)

Erros pós coleta: problemas de armazenamento dos dados, rotulamento errado ou inconsistente

“É impressão minha, mas está vindo um cheiro estranho destes dados?”

DE ONDE VEM DADOS “RUINS”

O *Amazon Rekognition* é um serviço que automatiza a análise de imagem e vídeo com aprendizado de máquina

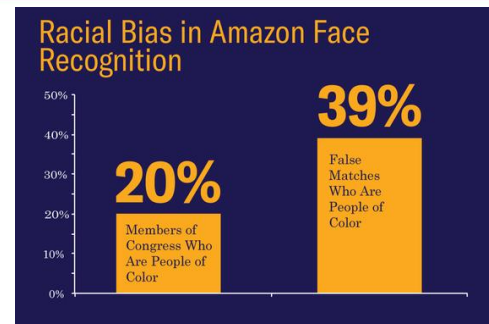
Após testar o *Rekognition* em 2018, a Associação Americana de Liberdades Cíveis (*American Civil Liberties Union – ACLU*) alegou que o sistema tinha um viés racial

- **28 dos membros** do congresso americano foram erroneamente identificados como **criminosos**
- **39% das identificações errôneas** eram pessoas **não caucasianas**

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots



By [Jacob Snow](#), Technology & Civil Liberties Attorney, ACLU of Northern California
JULY 26, 2018 | 8:00 AM



PARA SOLUÇÕES REAIS, DADOS REAIS

Em 2013, a IBM associou-se com o Centro de Câncer MD Anderson da Universidade do Texas para desenvolver um novo sistema de “Assessoria Expert de Oncologia”

MD Anderson Taps IBM Watson to Power "Moon Shots" Mission Aimed at Ending Cancer, Starting with Leukemia

Big Data Insights to Help Accelerate Translation of Cancer-Fighting Knowledge to Cutting Edge Medical Practices

HOUSTON - 18 Oct 2013: The University of Texas MD Anderson Cancer Center and IBM (NY)

O objetivo era **curar câncer!!**

PARA SOLUÇÕES REAIS, DADOS REAIS

Em fevereiro de 2017, a Forbes anuncia que a parceria foi cancelada após não atingir os seus objetivos e ter gasto em torno de **US\$ 62 milhões**

EDITORS' PICK | Feb 19, 2017, 03:48pm EST

MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine

Análises mostraram que Watson chegou a sugerir aos médicos que administrassem a um paciente de câncer com sangramento severo uma droga que agravaria o sangramento

Verificou-se que o software foi treinado em um **pequeno número** de **hipotéticos pacientes** com câncer, em vez de dados de pacientes reais!

<https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/>

PREOCUPADO COM DADOS?

Causa	Efeito
Muitos dados	Rotulamento de dados é tedioso, vagaroso, custoso e sujeito a erros (subjetividade), limita os modelos que podem ser utilizados
Muita informação	Modelos mais complexos, demanda coletar mais dados
Poucos dados/informação	Modelo não aprende as relações necessárias para generalizar
Representatividade	Viés (<i>bias</i>) do modelo para a classe majoritária ou informação incompleta
Causa	Efeito

COLETAR?

Planeje a coleta de dados de tal maneira que cubra todas as fontes de variação (condições de funcionamento) e situações (classes)

- Deve produzir uma quantidade “suficiente” para permitir a correta aprendizagem e generalização

Evitar ou diminuir dados tendenciosos (*bias*) para evitar modelos errados e possíveis conclusões erradas

- Bias de amostragem (não representativo)
- Bias de expectativa (resultado definido antes da pesquisa)
- Bias de conveniência (mais fácil obter um tipo de dado)
- Bias de medição (utilização de sensores com falha ou mal regulados)

COLETAR?

Dependendo dos dados que devem ser utilizados para o desenvolvimento da solução, o processo de coleta pode implicar em instalação de novos sensores

- Deve-se avaliar quais são as limitações dos sensores (resolução, distorção, ruídos, latência, entre outros) com respeito ao problema

Deve-se preocupar com o rotulamento dos dados (custo e tempo)

- Erros ou inconsistências no rotulamento contaminam os dados

REPOSITÓRIOS

Socrata

OpenML Home

Registry of Open Data on AWS

UCI Machine Learning Repository

StatLib---Datasets Archive (cmu.edu)

data.world | The Cloud-Native Data Catalog

Data Sources on the Web . MRAN (revolutionanalytics.com)

Kaggle: Your Machine Learning and Data Science Community

Datasets for Data Mining, Data Science, and Machine Learning – Kdnuggets



REPOSITÓRIOS

API Spreadsheets

Datasets - Data.gov

Harvard Dataverse

All OpenMV.net Datasets

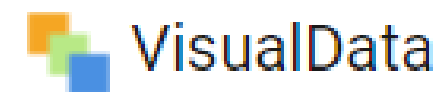
Zenodo - Research. Shared.

DASHlink - Resources (nasa.gov)

Conjuntos de dados - Portal Brasileiro de Dados Abertos

Datasets | Research | Canadian Institute for Cybersecurity | UNB

Prognostics Center of Excellence - Data Repository (nasa.gov)



CONCLUINDO

Um conjunto de dados estruturado é uma coleção de **amostras** / **observações** / **instâncias** / **padrões**, as quais são compostas por **atributos** / **características** / **variáveis**, organizadas de forma tabular

- A saída esperada (rótulos para a classificação e valores numéricos para regressão) também acompanha cada amostra em determinados problemas

Dados podem ser divididos em qualitativos (nominal e ordinal) e quantitativos (discreto e contínuo)

- Algoritmos geralmente trabalham sobre valores quantitativos, alguns em qualitativos e poucos em ambos simultaneamente

CONCLUINDO

Dados devem ser relevantes para o problema, representativos do domínio e apresentar boa acurácia na medição

Existem mais situações em torno da concepção, coleta e tratamento dos dados que possam afetar a qualidade

- Remediar não pode ser uma opção!

Geralmente, dados não estruturados são transformados (extração de características) em estruturados para que os algoritmos de aprendizagem sejam aplicados