

# Aula 1 - Introdução

Introdução ao aprendizado de máquina - UEMA 2025

---

Thiago S. F .Silva

2025-12-09

**Eu: Thiago Silva**

*Formação:*

- Biologia (UFRN, 2002)
- Sensoriamento Remoto (INPE, 2004)
- Geografia Física (UVic, 2009)

*Interesses:*

- Funcionamento e dinâmica de ecossistemas
- Efeitos das mudanças globais sobre a biota
- Interface entre Ecologia, Computação e Geociências
- “Ecologia Digital”

**Vocês?**

Nunca ouvi falar / Já ouvi falar / Já usei / Uso sempre

- Programação em R
- Programação em outras linguagens
- Estatística descritiva
- Modelos Lineares (regressão/ANOVA)
- Sensoriamento Remoto
- SIG
- Machine Learning

Nunca ouvi falar / Já ouvi falar / Já usei / Uso sempre

- CSV
- JSON
- shapefile
- GeoTiff
- Git
- GitHub

# O que esse curso é sobre?

- Análise **predictiva** de dados
- Introdução ao aprendizado de máquina (*machine learning*)
- Conversa sobre princípios e práticas científicas
- Ponto de partida para o seu próprio aprendizado

# O que esse curso não é sobre?

- Inferência estatística (mas vamos tocar nela várias vezes)
- “IA” - mas vamos contextualizar.

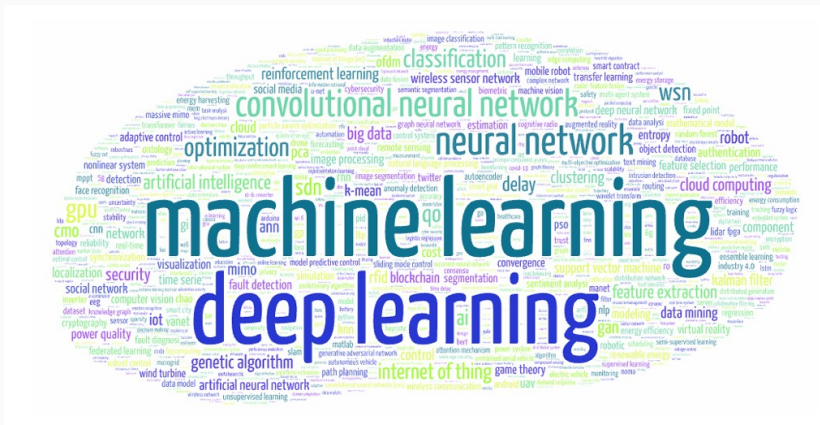
# Ciência de dados / data science ?

Wikipedia:

A *ciência de dados* é um campo **interdisciplinar** que utiliza *estatística, computação, método científico, visualização científica, algoritmos e sistemas* para **extrair ou extrapolar conhecimento** a partir de **dados** potencialmente *ruidosos, estruturados ou não-estruturados*. [2]

A ciência de dados também integra **conhecimento especializado** do domínio de aplicação (ex. ciências naturais, ciências sociais, medicina).

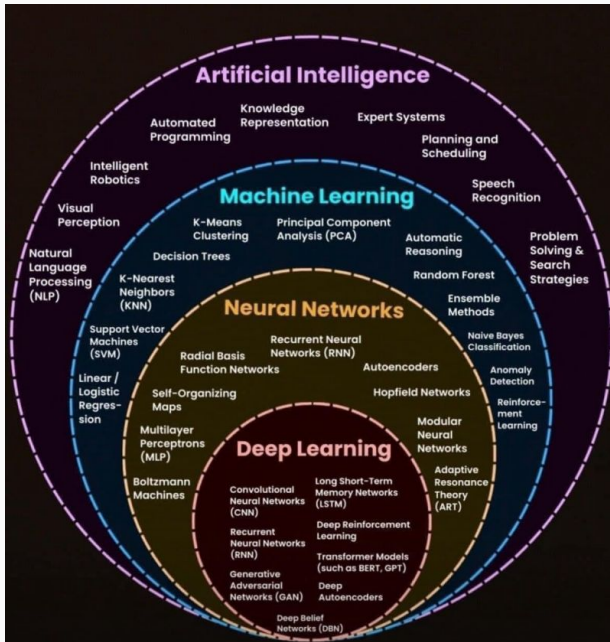
# Machine Learning?



**Figure 1:** Fonte: Xiaochen, Zhou, et al. “DeepDiveAI: Identifying AI Related Documents in Large Scale Literature Data.” arXiv preprint arXiv:2408.12871 (2024).



# Terminologia



# Estatística?

Matemática vs. Estatística?

- O que é Estatística?
- Por que usamos Estatística?

- Exploração
- Confirmação
- Predição

- **Exploração:** sem expectativas prévias, perguntas abertas.
- **Confirmação:** hipóteses pré-estabelecidas
- **Predição:** foco em gerar novos dados

## Exemplo:

- **Exploração:** Quão intenso é o efeito da redução de disponibilidade hídrica sobre o crescimento da espécie de planta *Plantus plantus*?

## Exemplo:

- **Exploração:** Quão intenso é o efeito da redução de disponibilidade hídrica sobre o crescimento da espécie de planta *Plantus plantus*?
- **Confirmação:** A redução em 50% na disponibilidade hídrica esperada para a década de 2100 resultará em redução de pelo menos 30% no crescimento de *Plantus plantus*.

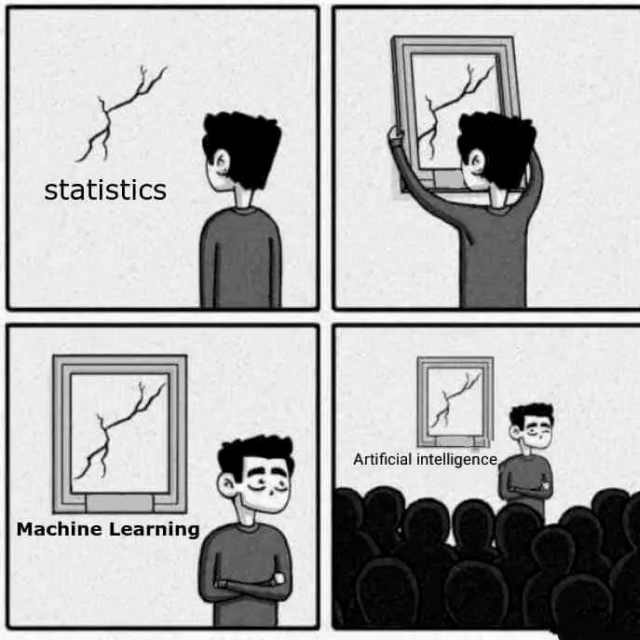
## Exemplo:

- **Exploração:** Quão intenso é o efeito da redução de disponibilidade hídrica sobre o crescimento da espécie de planta *Plantus plantus*?
- **Confirmação:** A redução em 50% na disponibilidade hídrica esperada para a década de 2100 resultará em redução de pelo menos 30% no crescimento de *Plantus plantus*.
- **Predição:** Qual o crescimento esperado para *Plantus plantus* em 2100 sob os cenários A, B e C de mudanças climáticas?

- Utiliza **análise exploratória** para entendimento dos *dados*
- Não se preocupa com **confirmação**.
- Especializado em gerar as melhores **predições** possíveis.



# Machine learning é estatística?



# Prova oral de estatística

Call:

```
lm(formula = y ~ x, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-377.94	-66.26	-4.97	66.86	281.84

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.229	6.277	0.833	0.40501
x	0.341	0.111	3.071	0.00219 **

---

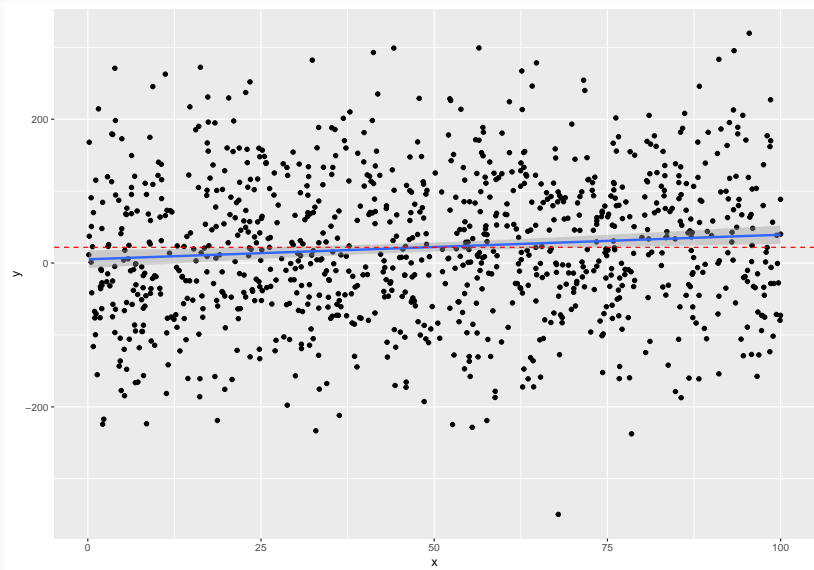
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 100.4 on 998 degrees of freedom

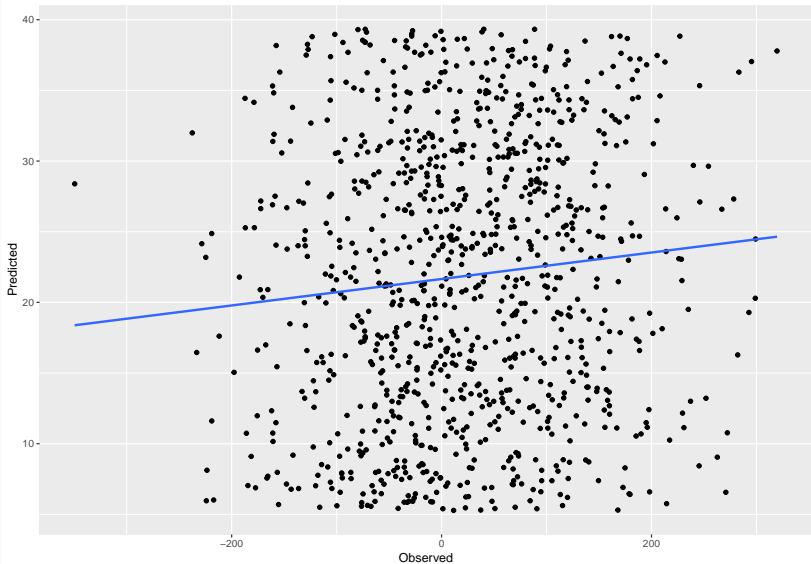
Multiple R-squared: 0.009364, Adjusted R-squared: 0.008372

F-statistic: 9.434 on 1 and 998 DF, p-value: 0.002188

# Gráfico do Modelo



# Observado vs Predito



## Bom modelo ou mau modelo?

- Excelente do ponto de vista *confirmatório*.

## Bom modelo ou mau modelo?

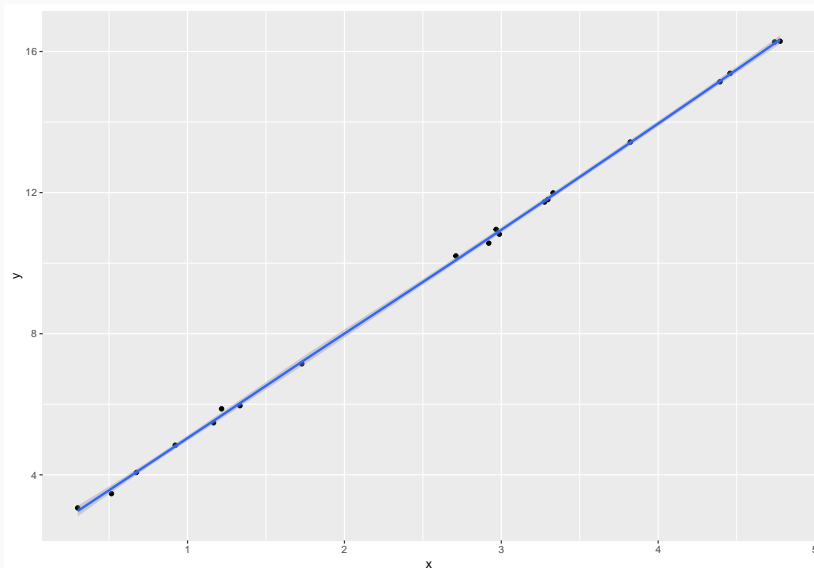
- Excelente do ponto de vista *confirmatório*.
- Péssimo do ponto de vista *preditivo*.

## Erro, erro, erro

Quando estamos preocupados com predição, o que realmente importa é minimizar o **erro**.

# Pouco erro

O modelo abaixo tem bem pouco erro - estamos satisfeitos?





Workflow de uma análise estatística 'tradicional':

1. Preparação dos dados

Workflow de uma análise estatística 'tradicional':

1. Preparação dos dados
2. Análise descritiva

Workflow de uma análise estatística 'tradicional':

1. Preparação dos dados
2. Análise descritiva
3. Visualização de dados

Workflow de uma análise estatística 'tradicional':

1. Preparação dos dados
2. Análise descritiva
3. Visualização de dados
4. Ajuste do modelo (teste estatístico)

Workflow de uma análise estatística 'tradicional':

1. Preparação dos dados
2. Análise descritiva
3. Visualização de dados
4. Ajuste do modelo (teste estatístico)
5. Interpretação dos resultados

Workflow de uma análise preditiva:

1. Preparação dos dados

Workflow de uma análise preditiva:

1. Preparação dos dados
2. Análise descritiva

Workflow de uma análise preditiva:

1. Preparação dos dados
2. Análise descritiva
3. Visualização de dados



Workflow de uma análise preditiva:

1. Preparação dos dados
2. Análise descritiva
3. Visualização de dados
4. **Divisão dos dados em *treinamento* e *teste***

Workflow de uma análise preditiva:

1. Preparação dos dados
2. Análise descritiva
3. Visualização de dados
4. **Divisão dos dados em *treinamento* e *teste***
5. Ajuste do modelo

Workflow de uma análise preditiva:

1. Preparação dos dados
2. Análise descritiva
3. Visualização de dados
4. **Divisão dos dados em *treinamento* e *teste***
5. Ajuste do modelo
6. **Otimização e validação do modelo**

Workflow de uma análise preditiva:

1. Preparação dos dados
2. Análise descritiva
3. Visualização de dados
4. **Divisão dos dados em *treinamento e teste***
5. Ajuste do modelo
6. **Otimização e validação do modelo**
7. **Teste do modelo**

Por que o curso é baseado em *R*?

- Porque todo profissional nas ciências naturais precisa saber o mínimo de programação.
- Fácil de instalar e usar.
- Criado especificamente para análise de dados.
- 'Padrão' nas ciências naturais.

Mas e o tal do *Python*?

- Sim, é a principal linguagem para data science.
- Mais complicada pra instalar e começar a usar.
- Menos popular no mundo acadêmico.
- É uma linguagem criada para uso geral, e por isso muito ‘maior’ do que o R.
  - No. Pacotes no CRAN: 23035
  - No. Pacotes no PyPI: 709358

## Entrando no “Tidyverso”

**Tidyverse:** um conjunto de pacotes do R que ‘redefiniu’ a linguagem e a maneira de trabalhar, com ênfase em ciência de dados.

<https://tidyverse.org/>

O conceito central do tidyverse é o operador '*pipe*' (cano), que conecta operações sequenciais:

`%>%` ou `|>` a partir do R 4.0



# Um exemplo

```
summary(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300
Median :5.800	Median :3.000	Median :4.350	Median :1.300
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500

Species

setosa	:50
versicolor	:50
virginica	:50

```
sel <- iris[iris$Sepal.Length > 6,]  
sel <- sel[, "Petal.Width"]  
res <- mean(sel)  
res
```

```
[1] 1.847541
```

```
# Ou tudo em uma linha  
res <- mean(iris[iris$Sepal.Length > 6, "Petal.Width"])  
res
```

```
[1] 1.847541
```

# Com pipe

```
library(tidyverse)
res <- iris %>%
  filter(Sepal.Length > 6) %>%
  summarise(mean_pw = mean(Petal.Width))
res
```

```
      mean_pw
1 1.847541
```

**Perguntas?**

---

### Gerenciamento e visualização de dados

- Inclui práticas no R
- Se possível trazer seu computador com R, RStudio e os pacotes básicos já instalados (programa de curso que eu envie)