



# Mini curso de Introdução a Bioinformática

Ms. Prof.<sup>a</sup> Patricia Pedros Estevam Ribeiro



I ENCONTRO  
DE CIÊNCIA E SAÚDE

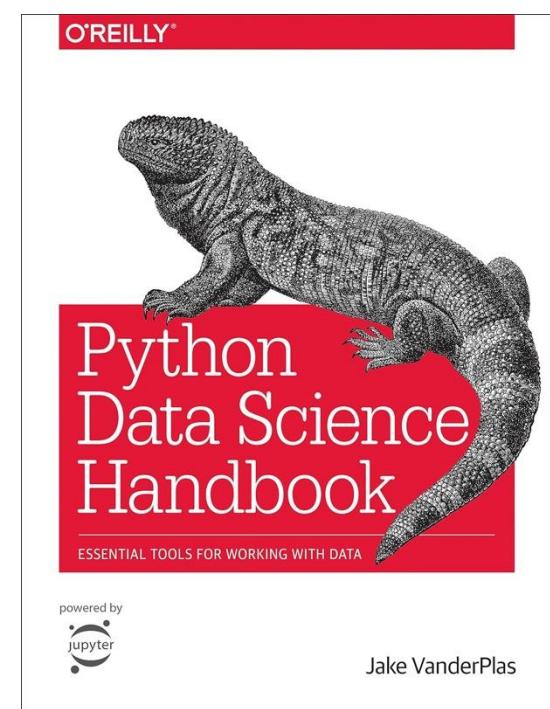
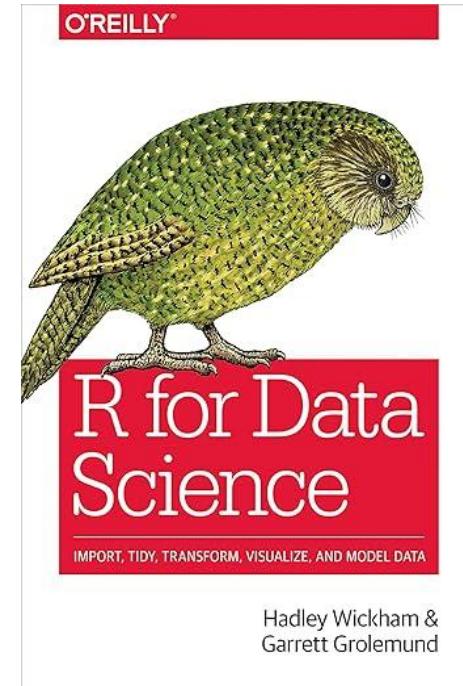
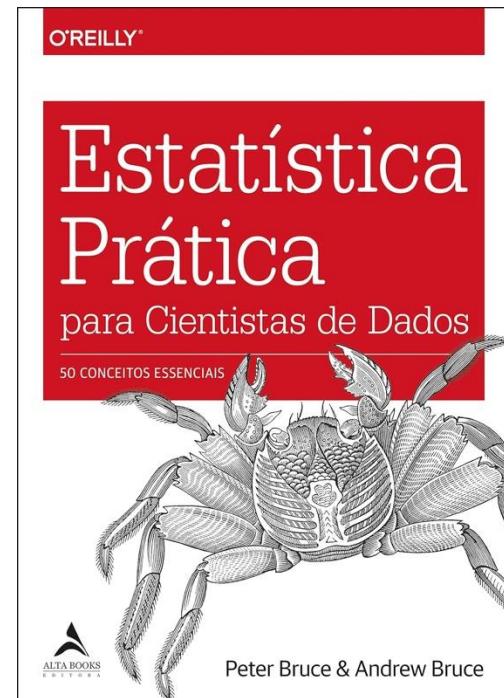
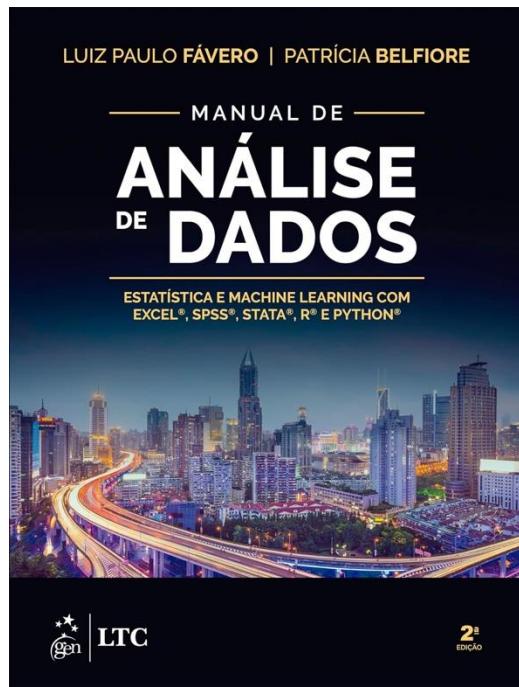
unifeb



PATRICIA PEDROSO ESTEVAM RIBEIRO

- Doutorando na área de Oncologia Molecular  
(Hospital de Câncer de Barretos, com foco em Bioinformática)
- MBA em Data Science e Analytics  
(USP / ESALQ)
- Mestrado em Processamento de Imagens  
(USP / São Carlos)
- Graduação em Engenharia Elétrica com ênfase em computação  
(Unifeb)

# Referências Bibliográficas



## O QUE É BIOINFORMÁTICA?



- A bioinformática é a união da ciência da computação com a biologia molecular.
- Podemos defini-la como uma **especialização da informática voltada ao estudo de técnicas computacionais e matemáticas aplicadas à geração e ao gerenciamento de informações biológicas.**

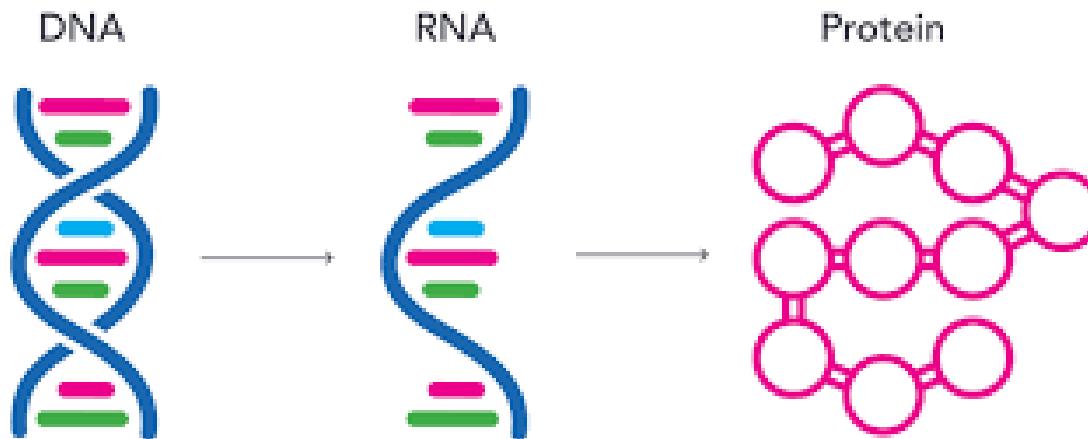
# Para que serve a bioinformática?

- Com o avanço da biotecnologia e da genômica, a bioinformática tornou-se fundamental para a pesquisa biomédica, permitindo a **identificação de padrões em grandes volumes de dados biológicos e contribuindo para descobertas na saúde, na farmacologia e na genética.**



# Para que serve a bioinformática?

- Essa área envolve o desenvolvimento de ferramentas e métodos computacionais para analisar e interpretar **dados biológicos, como sequências de DNA, RNA, proteínas.**

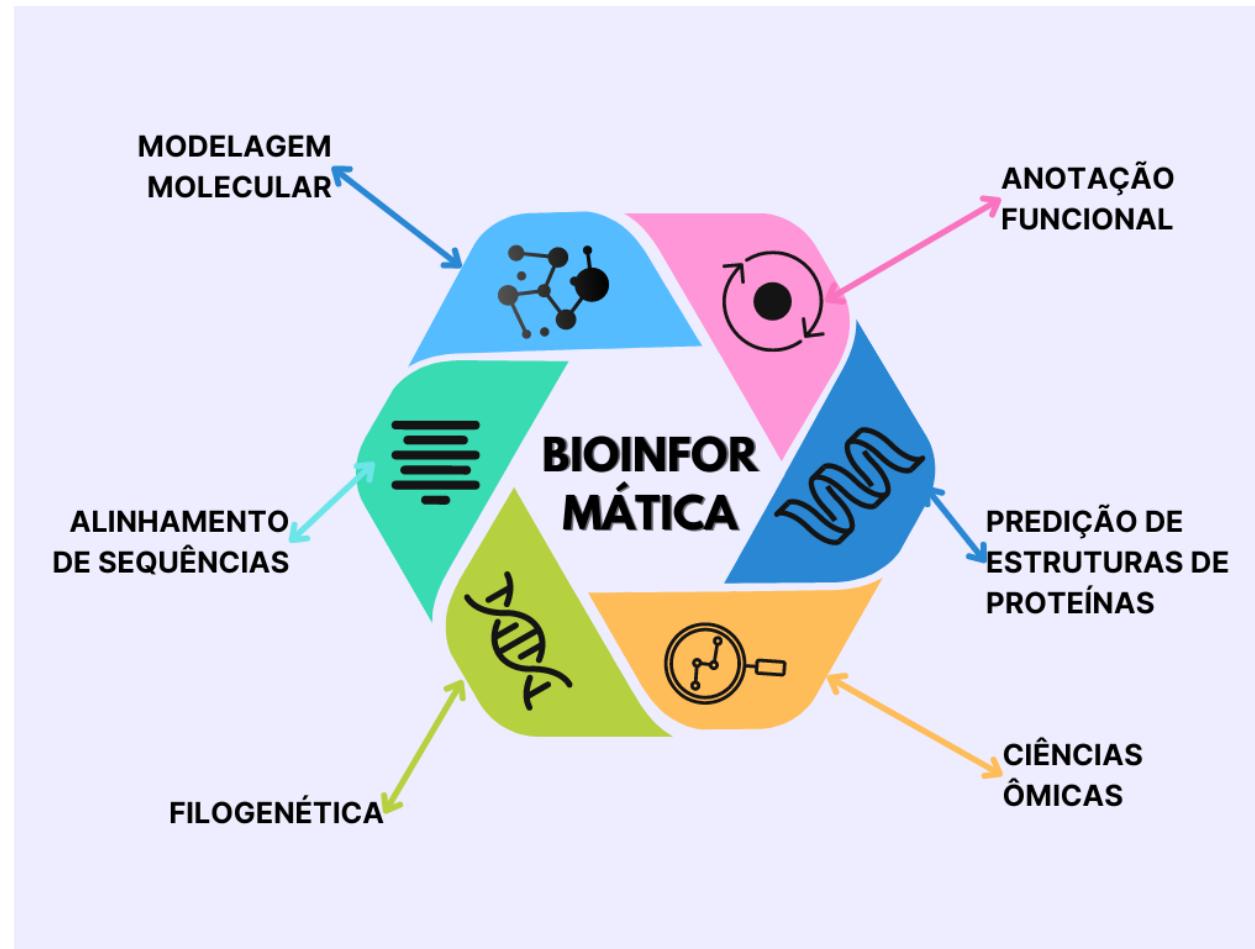


# O Papel do Cientista de Dados na Bioinformática

- **O Cientista de Dados em Bioinformática** aplica técnicas avançadas de biologia computacional, combinando **métodos estatísticos e computacionais** para interpretar grandes volumes de dados biológicos. Sua atuação possibilita descobertas científicas e contribui para o avanço da pesquisa médica, genômica e genética populacional.

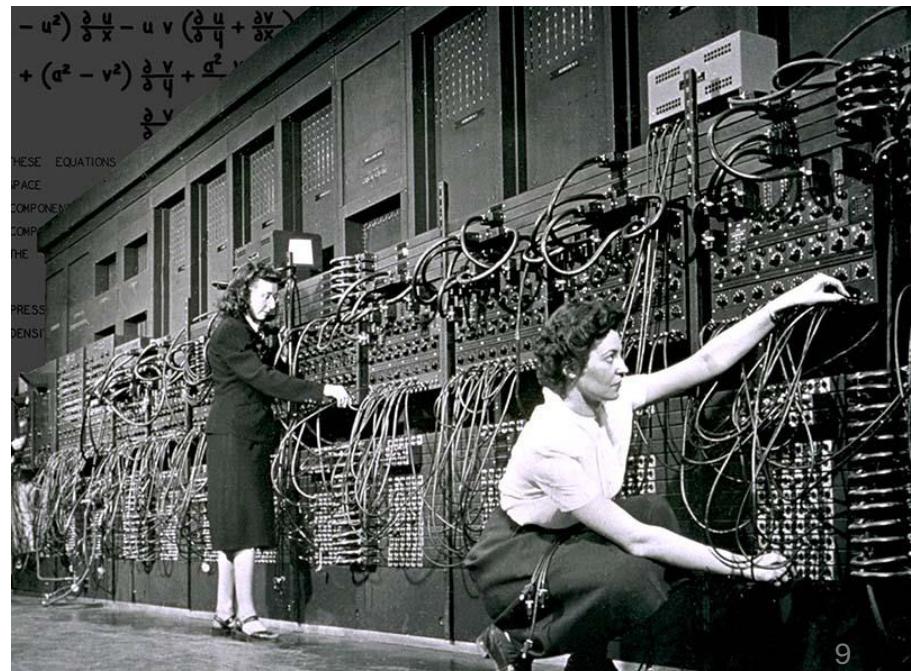


# Áreas de Aplicação da Bioinformática



# Início da Computação

- No contexto da bioinformática, o avanço dos processadores permitiu a análise de grandes quantidades de dados genéticos, viabilizando pesquisas sobre evolução, doenças genéticas e interações moleculares.
- 1950 – 1970
  - Os primórdios da bioinformática:
  - Durante esse período, avanços na computação, como o ENIAC (*Electronic Numerical Integrator and Computer*), apresentado em 1946.



# Início da Computação

- Desde os primeiros algoritmos matemáticos desenvolvidos por cientistas como Alan Turing e John von Neumann, até a criação de computadores modernos, essa tecnologia evoluiu significativamente, impactando diversas áreas do conhecimento, incluindo a biologia.

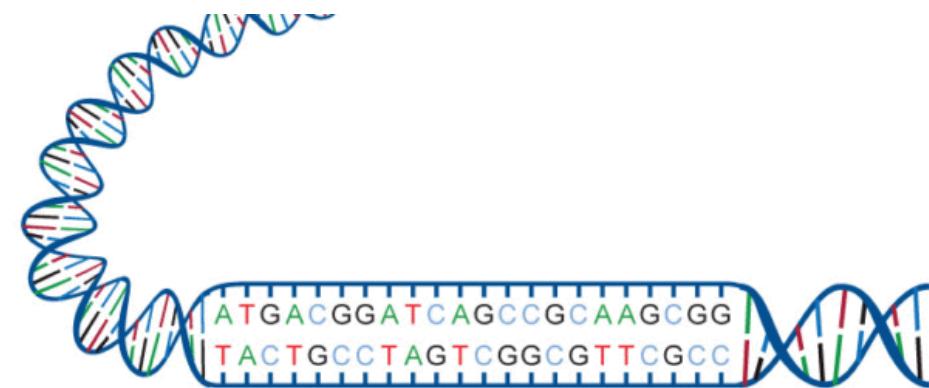


Alan Turing  
1912–1954

John von Neumann  
1903–1957

# A Revolução da Bioinformática e dos Computadores: Sequenciamento de DNA

- Em 1953 **Watson e Crick** desvendaram a estrutura tridimensional do DNA, mas somente na década de 70 que os primeiros métodos para sequenciamento de DNA foram criados.
- O sequenciamento de DNA nada mais é do que a determinação da ordem exata em que os nucleotídeos se encontram ao longo dessa dupla fita.



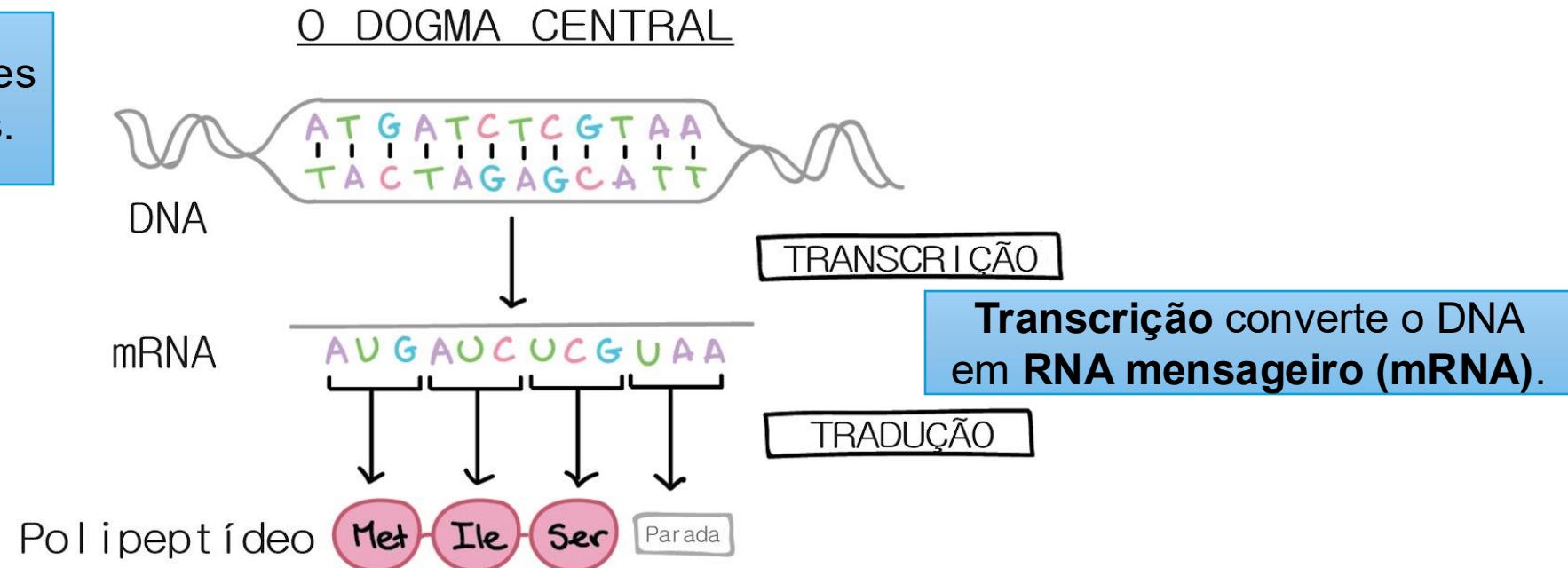
O DNA é composto por duplas fitas de quatro moléculas quimicamente distintas: **adenina (A)**, **guanina (G)**, **citosina (C)** e **timina (T)**.

Para entender como os genes funcionam, é necessário “ler” a ordem em que elas estão dispostas – processo conhecido como sequenciamento de DNA.

# O Dogma Central da Biologia

Descreve o fluxo de informação genética dentro das células

DNA contém as instruções genéticas armazenadas.

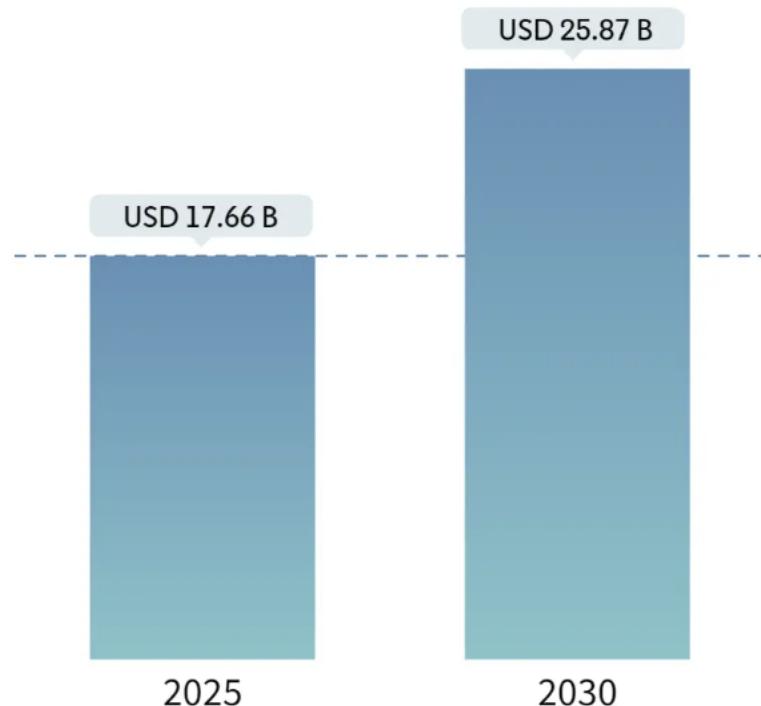


A tradução gera proteínas a partir do mRNA, e a bioinformática auxilia na análise genética para diagnósticos e novos medicamentos

## • Mercado de Bioinformática: Tendências 2025-2030

O relatório com a tendência o crescimento global da bioinformática, abordando tamanho de mercado, tendências e evolução do setor, como **(genoma microbiano, engenharia genética, desenvolvimento de medicamentos, medicina personalizada, ômicas)**. O mercado é segmentado por **produtos e serviços, aplicações e regiões**, com valores expressos em milhões de dólares, destacando seu impacto na pesquisa e inovação.

Bioinformatics Market  
Market Size in USD Billion  
CAGR **7.94%**



Source : Mordor Intelligence



# Banco de Dados Biológico On-line

Banco de Dados	Link	Descrição
NCBI ( <i>National Center for Biotechnology Information</i> )	<a href="https://www.ncbi.nlm.nih.gov/">https://www.ncbi.nlm.nih.gov/</a>	Repositório de dados biomédicos e genômicos, incluindo GenBank, PubMed e BLAST.
EBI ( <i>European Bioinformatics Institute</i> )	<a href="https://www.ebi.ac.uk/">https://www.ebi.ac.uk/</a>	Centro europeu de bioinformática que fornece acesso a dados biológicos, incluindo Ensembl e UniProt.
DNA Data Bank of Japan (DDBJ)	<a href="https://www.ddbj.nig.ac.jp/index-e.html">https://www.ddbj.nig.ac.jp/index-e.html</a>	Banco de dados japonês que coleta e armazena sequências de DNA, parte do INSDC.
Swiss Institute of Bioinformatics (SIB)	<a href="https://www.sib.swiss/">https://www.sib.swiss/</a>	Portal suíço de bioinformática que hospeda recursos como UniProt, Swiss-Prot e Expasy.
Protein Data Bank (PDB)	<a href="https://www.rcsb.org/">https://www.rcsb.org/</a>	Banco de dados global de estruturas tridimensionais de proteínas e macromoléculas biológicas.

# Banco de Dados Biológico On-line

NCBI (The National Center for Biotechnology Information)

<https://ncbi.nlm.nih.gov/>

 National Library of Medicine  
National Center for Biotechnology Information

Todos os bancos ▾  Procurar

Conecte-se

**Página inicial do NCBI**

- [Lista de Recursos \(AZ\)](#)
- [Todos os recursos](#)
- [Produtos Químicos e Bioensaios](#)
- [Dados e Software](#)
- [DNA e RNA](#)
- [Domínios e Estruturas](#)
- [Genes e Expressão](#)
- [Genética e Medicina](#)
- [Genomas e Mapas](#)
- [Homologia](#)
- [Literatura](#)
- [Proteínas](#)
- [Análise de Sequência](#)
- [Taxonomia](#)
- [Treinamento e Tutoriais](#)
- [Variação](#)

**Bem-vindo ao NCBI**

O Centro Nacional de Informações sobre Biotecnologia promove a ciência e a saúde fornecendo acesso a informações biomédicas e genômicas.

[Sobre o NCBI](#) | [Missão](#) | [Organização](#) | [Notícias e Blog do NCBI](#)

**Enviar**  
Depositar dados ou manuscritos em bancos de dados do NCBI



**Download**  
Transfira os dados do NCBI para o seu computador



**Aprender**  
Encontre documentos de ajuda, assista a uma aula ou a um tutorial



**Desenvolver**  
Use APIs e bibliotecas de código do NCBI para criar aplicativos



**Analisar**  
Identifique uma ferramenta NCBI para sua tarefa de análise de dados



**Pesquisar**  
Explore a pesquisa e os projetos colaborativos do NCBI



**Recursos populares**

- [PubMed](#)
- [Estante de livros](#)
- [PubMed Central](#)
- [EXPLOSÃO](#)
- [Nucleotídeo](#)
- [Genoma](#)
- [SNP](#)
- [Gene](#)
- [Proteína](#)
- [PubChem](#)

**Notícias e Blog do NCBI**

Atualizações da taxonomia do NCBI para classificação de vírus  
25 de abril de 2025

A partir de 28 de abril de 2025, em dezembro de 2024, anunciamos diversas

A pré-visualização atualizada da pesquisa de texto completo do PubMed Central já está disponível  
08 de abril de 2025

# Banco de Dados Biológico On-line

Pesquisar NCBI  x Procurar

Resultados encontrados em 28 bases de dados

**GENE**

**EGFR – receptor do fator de crescimento epidérmico**  
*Homo sapiens (humano)*  
Também conhecido como: ERBB, ERBB1, ERRP, HER1, NISBD2, NNCIS, PIG61, mENA  
ID do gene: 1956

[Produtos RefSeq](#) [Ortólogos](#) [Visualizador de dados do genoma](#)

[Novo - Visualize genes em várias espécies](#)

**Sequências RefSeq**

Isso foi ú

Literatura	
Estante de livros	5.961
Malha	26
Catálogo NLM	111
PubMed	138.371
PubMed Central	354.539

Genes	
Gene	11.539
Conjuntos de dados GEO	63.878
Perfis GEO	326.114

Proteínas	
Domínios Conservados	126
Grupos de proteínas idênticos	863
Proteína	27.953
Modelos de Famílias de Proteínas	52
Estrutura	562

Genomas	
Montagem / Genoma	0
Conjuntos de dados do NCBI	0
BioColecções	0
BioProjeto	1.422
Bioamostra	5.422
Nucleotídeo	47.827
SRA	18.850
Taxonomia	0

Clínico	
ClinicalTrials.gov	0
ClinVar	3.999
dbGaP	32
dbSNP	85.415
dbVar	1.185
GTR	236
MedGen	341
OMIM	255

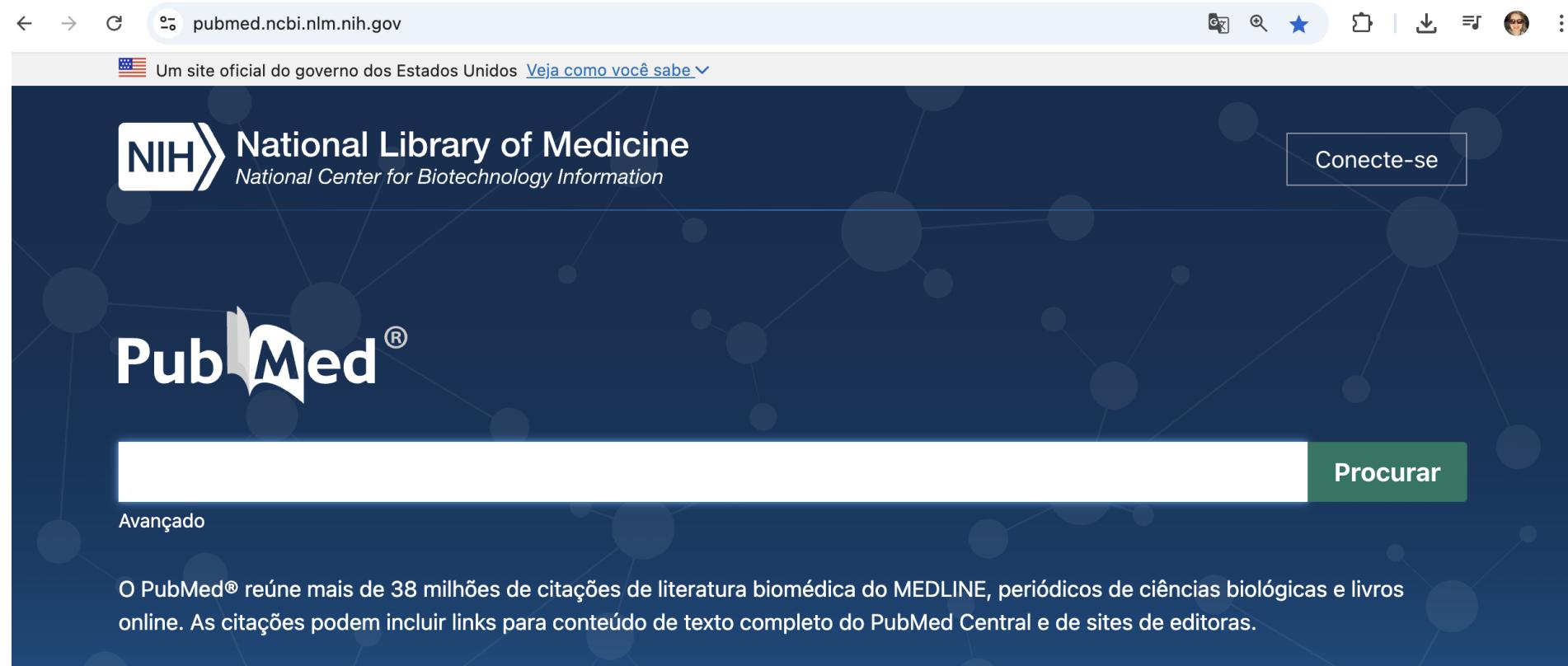
PubChem	
Bioensaios	25.541
Compostos	308
Caminhos	80
Substâncias	1.365

**Conjuntos de dados do NCBI** Todos os dados do genoma reunidos agora estão disponíveis por meio dos conjuntos de dados do NCBI

# Banco de Dados Biológico On-line

NCBI (The National Center for Biotechnology Information)

<https://pubmed.ncbi.nlm.nih.gov/> (Citações de literatura biomédica, periódicos e livros)



# Banco de Dados Biológico On-line

NCBI (The National Center for Biotechnology Information)

<https://pubmed.ncbi.nlm.nih.gov/> (Citações de literatura biomédica, periódicos e livros)

The screenshot shows the PubMed search interface with the query 'lung cancer + machine learning' entered. The results page displays 4,125 entries. The first result is a study titled 'Aprendizado de máquina para diagnóstico, tratamento e prognóstico do câncer de pulmão'. The second result is another study titled 'A inteligência artificial e o aprendizado de máquina na imunoterapia do câncer de pulmão'. Both results are from 2022 and are available as free PMC articles.

MEUS FILTROS PERSONALIZADOS

RESULTADOS POR ANO

DATA DE PUBLICAÇÃO

DISPONIBILIDADE DE TEXTO

Resumo

lung cancer + machine learning

Procurar

Avançado Criar alerta Criar RSS

Guia do usuário

Salvar E-mail Enviar para

Ordenar por: Melhor correspondência Opções de exibição

4.125 resultados Página 1 de 413

Aprendizado de máquina para diagnóstico, tratamento e prognóstico do câncer de pulmão.  
1 Citar Li Y, Wu X, Yang P, Jiang G, Luo Y. Genômica Proteômica Bioinformática. 2022 Out;20(5):850-866. doi: 10.1016/j.gpb.2022.11.003. Compartilhar Epub 2022 Dez 1. PMID: 36462630 Artigo gratuito do PMC. Análise. Abordagens baseadas em aprendizado de máquina desempenham um papel crítico na integração e análise desses conjuntos de dados grandes e complexos, que caracterizaram extensivamente o câncer de pulmão por meio do uso de diferentes perspectivas desses dados acumulados. ... Além disso, nós oi ...

A inteligência artificial e o aprendizado de máquina na imunoterapia do câncer de pulmão.  
2 Citar Gao Q, Yang L, Lu M, Jin R, Ye H, Ma T. J Hematol Oncol. 2023 24 de maio;16(1):55. doi: 10.1186/s13045-023-01456-y. Compartilhar PMID: 37226190 Artigo gratuito do PMC. Análise. Nos últimos anos, a inteligência artificial (IA) baseada em aprendizado de máquina (ML) foi desenvolvida na área de convergência médico-industrial. ... Nesta revisão, as aplicações da IA na predição de PD-L1/TMB, predição de TME e imunoterapia para câncer de pulmão são d...

# Banco de Dados Biológico On-line

cBIOPortal for Cancer Genomics  
<https://www.cbioportal.org/>

Navegue pelos conjuntos de dados disponíveis e selecione estudos para explorar ou consultar

Lista de todos os estudos, organizados por sistema orgânico

Estudos de pesquisa

O que há de novo @cbioportal 06 de maio de 2025

Dados adicionados consistindo de 4.571 amostras de 10 estudos:

- Adenocarcinoma pancreático (MSK, Nat Med 2024) 2336 amostras
- DNA tumoral circulante do líquido céfalorraquidiano (MSK, Acta Neuropathol Commun 2024) 1007 amostras
- Câncer de ovário (Gray Foundation, Cancer Discov 2024) 567 amostras
- Melanócitos normais (UCSF, Nature 2020) 153 amostras
- Queratinócitos normais de pele humana (UCSF, BioRxiv 2024) 136 amostras
- Fusões BRAF - Coorte de

Leia a última Newsletter do cBioPortal! Assine via:

LinkedIn Grupos do Google

Consultas de exemplo

- Câncer de próstata primário vs. metastático
- Alterações RAS/RAF no câncer colorretal

# Banco de Dados Biológico On-line

The screenshot shows the cBioPortal homepage with a search query for "glioma". A sidebar on the left lists studies by organ system: "Estudos imunogenômicos" (1), "SNC/Cérebro" (22), and "Gliodo mole" (2). The main search results are displayed in a table where each row represents a study. The first row for "MSK-CHORD" is selected, indicated by a checked checkbox. Other studies listed include "Consórcio de Testes Pré-clínicos Pediátricos" (218 samples), "Glioma difuso" (530 samples), and "Glioma (MSK, Clin Cancer Res 2019)" (1004 samples). At the bottom, there are buttons for "Consulta por Gene" and "Explore estudos selecionados".

← → ⌂ cbioportal.org

cBioPortal FOR CANCER GENOMICS

Conjuntos de dados API da Web Tutoriais/Webinars Perguntas frequentes Notícias Visualize seus dados

Reapresentando a Newsletter cBioPortal! Assine via LinkedIn ou Grupos do Google

Consulta Busca rápida

Selezione Estudos para Visualização e Análise:

1 estudo selecionado (514 amostras) Desmarcar tudo

Tipo de dados

glioma X

1. Lista de todos os estudos, organizados por sistema orgânico

2. Selecione a caixa de seleção ao lado do estudo de interesse e clique em “Explorar Estudos Selecionados”

4. Ou clique no botão “PubMed”

3. Ou clique no botão “Ver resumo do estudo”

1 estudo selecionado (514 amostras) Desmarcar tudo

Consulta por Gene OU Explore estudos selecionados

Estudo	Amostras	Detalhes
MSK-CHORD (MSK, Natureza 2024)	42 amostras	<a href="#">Detalhes</a> <a href="#">PubMed</a>
Consórcio de Testes Pré-clínicos Pediátricos (CHOP, Cell Rep 2019)	218 amostras	<a href="#">Detalhes</a> <a href="#">PubMed</a>
Glioma difuso	530 amostras	<a href="#">Detalhes</a> <a href="#">PubMed</a>
Glioma cerebral de baixo grau (TCGA, legado Firehose)	514 amostras	<a href="#">Detalhes</a> <a href="#">PubMed</a>
Glioma cerebral de baixo grau (TCGA, PanCancer Atlas)	693 amostras	<a href="#">Detalhes</a> <a href="#">PubMed</a>
Glioma Difuso (Consórcio GLASS)	444 amostras	<a href="#">Detalhes</a> <a href="#">PubMed</a>
Glioma Difuso (Consórcio GLASS, Nature 2019)	530 amostras	<a href="#">Detalhes</a> <a href="#">PubMed</a>
Glioma Difuso (TCGA, GDC)	1004 amostras	<a href="#">Detalhes</a> <a href="#">PubMed</a>
Glioma (MSK, Clin Cancer Res 2019)		

# Banco de Dados Biológico On-line

**Glioma cerebral de baixo grau (TCGA, PanCancer Atlas)** 

Dados do PanCancer TCGA para Glioma Cerebral de Grau Inferior. Os dados originais estão [aqui](#). As publicações estão [aqui](#). [PubMed](#)

Clique nos símbolos genéticos abaixo ou entre  [Consulta](#)

**Resumo** **Dados clínicos** **Segmentos CN** **Tomografia computadorizada** **Tramas Beta!** **Seleção personalizada** **Gráficos** **Grupos** 

**Selecionado: 514 pacientes | 514 amostras**   

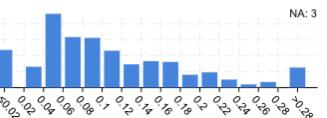
Ajuda da página de estudo

**Tipo de câncer detalhado**

	#	Frequênc
Astrocitoma	194	37.7%
Oligodendrogioma	189	36.8%
Oligoastrocitoma	130	25.3%
Low-Grade Glioma (NOS)	1	0.2%

Procurar...

**Fração do Genoma Alterado**



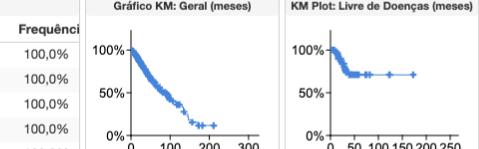
**Contagens de amostras de perfil genómico**

**Perfil Molecular**

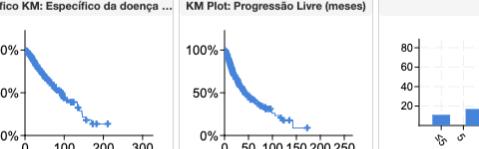
	#	Frequênc
Escores z de expressão de mRNA...	514	100,0%
Mutações	514	100,0%
Número de cópia putativa do níve...	514	100,0%
Variantes estruturais	514	100,0%
Metilação (fusão de HM27 e HM4...	514	100,0%
Escores z de expressão de mRNA...	514	100,0%
Expressão de mRNA, RSEM (Lote...	514	100,0%
Valores de número de cópias do l...	511	99,4%
Alterações putativas no número d...	511	99,4%
Ancestralidade Genética	480	93,4%
Escores z de expressão proteica (...	428	83,3%

Procurar...

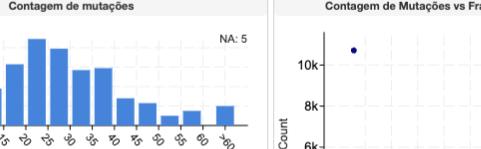
**Gráfico KM: Geral (meses)**



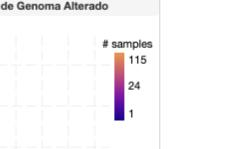
**KM Plot: Livre de Doenças (meses)**



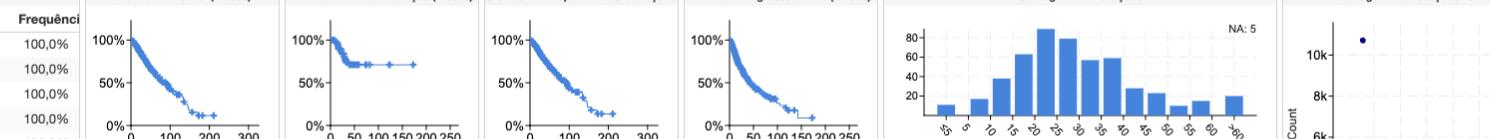
**Gráfico KM: Específico da doença ...**



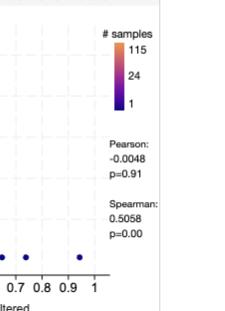
**KM Plot: Progressão Livre (meses)**



**Contagem de mutações**



**Contagem de Mutações vs Fração de Genoma Alterado**



**Genes Mutados (514 amostras perfiladas)**

Gene	# Mut	#	Frequênc
IDH1	395	□ 395	76,8%
TP53	319	□ 249	48,4%
ATRX	218	□ 194	37,7%
CIC	130	□ 108	21,0%
TTN	114	□ 62	12,1%
FUBP1	51	□ 48	9,3%
PIK3CA	46	□ 42	8,2%
ENTALHE 1	49	□ 38	7,4%
MUC16	54	□ 36	7,0%
EGFR	42	□ 35	6,8%
NF1	47	□ 31	6,0%

Procurar...

**Genes Variantes Estruturais (514 amostras perfiladas)**

Gene	#SV	#	Frequênc
PDGFRA	6	□ 5	1,0%
SEPTIN14	6	□ 6	1,2%
CLU	6	□ 6	1,2%
EGFR	5	□ 5	1,0%
FGFR3	5	□ 4	0,8%
ATAD1	4	□ 3	0,6%
NPTXR	4	□ 2	0,4%
QKI	4	□ 4	0,8%
TSPAN31	4	□ 3	0,6%
TACC3	4	□ 3	0,6%
KIF5A	4	□ 4	0,8%

Procurar...

**Genes CNA (511 amostras perfiladas)**

Gene	Citobanda	CNA	#	Frequênc
CDKN2B	9p21.3	<b>HOMDEL</b>	□ 56	11,0%
CDKN2B-A...	9p21.3	<b>HOMDEL</b>	□ 56	11,0%
CDKN2A	9p21.3	<b>HOMDEL</b>	□ 55	10,8%
CDKN2A-A...	9p21.3	<b>HOMDEL</b>	□ 51	10,0%
MTAP	9p21.3	<b>HOMDEL</b>	□ 45	8,8%
EGFR	7p11.2	<b>AMP</b>	□ 39	7,6%
EGFR-AS1	7p11.2	<b>AMP</b>	□ 38	7,4%
VELHO	7p11.2	<b>AMP</b>	□ 37	7,2%
SEC61G-DT	7p11.2	<b>AMP</b>	□ 34	6,7%
SEC61G	7p11.2	<b>AMP</b>	□ 30	5,9%
DMRTA1	9p21.3	<b>HOMDEL</b>	□ 30	5,9%

Procurar...

**Tratamento por Paciente**

Tratamento	#
Radiação 1	□ 299
Temozolomida	□ 255
Bevacizumabe	□ 48
Lomustina	□ 39
Radiação 2	□ 31
Procarbazina	□ 23
Irinotecano	□ 21
Vincristina	□ 16
Carmustina	□ 15
Etoposídeo	□ 12
Tamoxifeno	□ 8

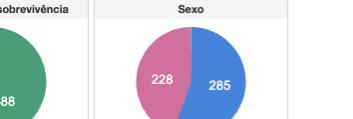
Procurar...

**Tratamento por Amostra (pré/pós)**

Tratamento	Pré / Pós	#
Radiação 1	Pre	□ 293
Temozolomida	Pre	□ 250
Bevacizumabe	Pre	□ 47
Lomustina	Publicar	□ 1
Radiação 2	Pre	□ 31
Procarbazina	Pre	□ 23
Irinotecano	Pre	□ 21
Vincristina	Pre	□ 16
Carmustina	Pre	□ 15
Etoposídeo	Pre	□ 12

Procurar...

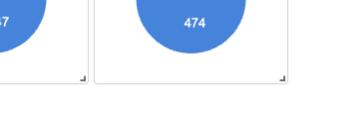
**Status geral de sobrevida**



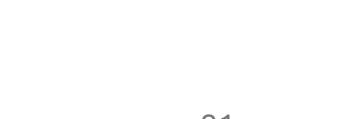
**Sexo**



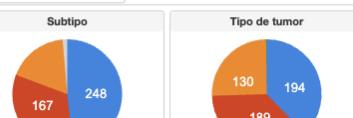
**Categoria Etnia**



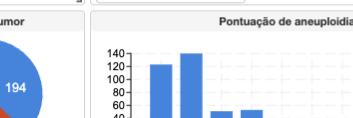
**Categoria de corrida**



**Subtipo**



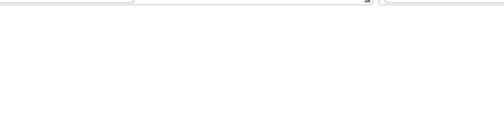
**Tipo de tumor**



**Pontuação de aneuploidia**



**Nascimento a partir da data do diagnóstico patológico inicial**



# Linguagens de programação R e Python



# Linguagens de programação R e Python

- **Linguagem de programação**

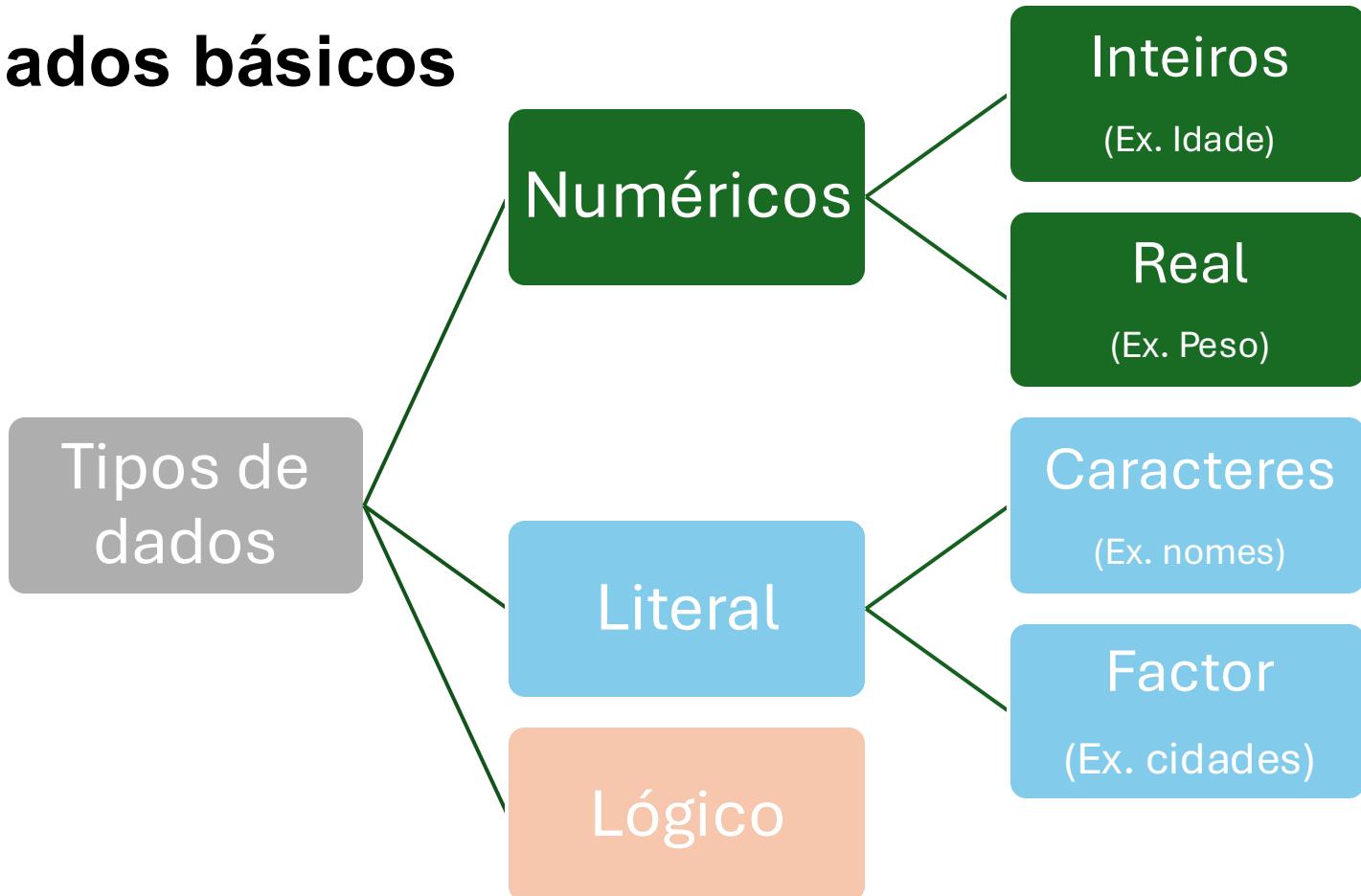
- É um conjunto de regras sintáticas e semânticas para implementação de um código fonte que pode ser compilado e transformado em um programa de computador.

- **Algoritmo**

- É uma sequencia finita de ações executáveis que visam obter uma solução para o um determinado tipo de problema.

# Linguagens de programação R e Python

## Tipos de dados básicos



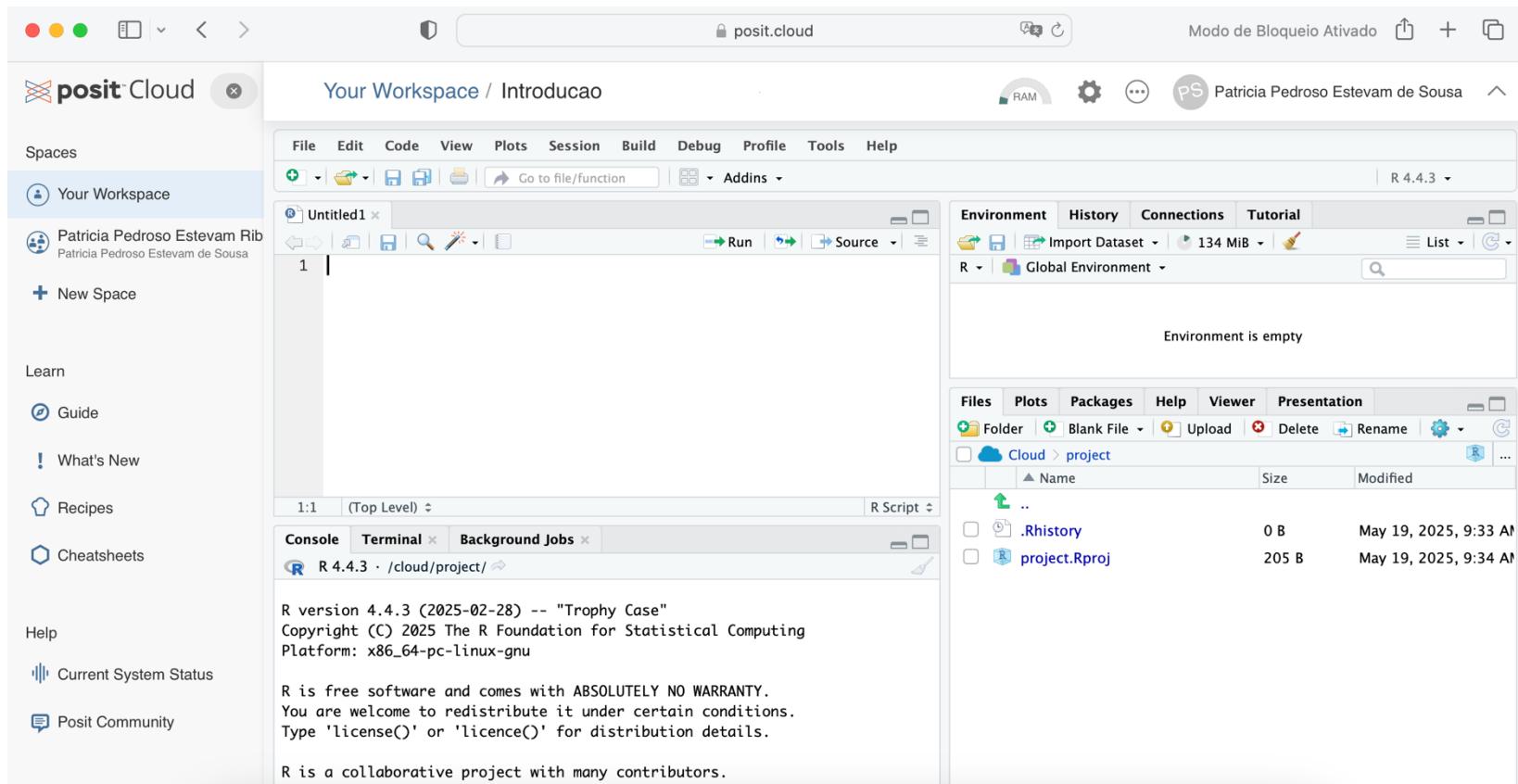
# Linguagens de programação R e Python



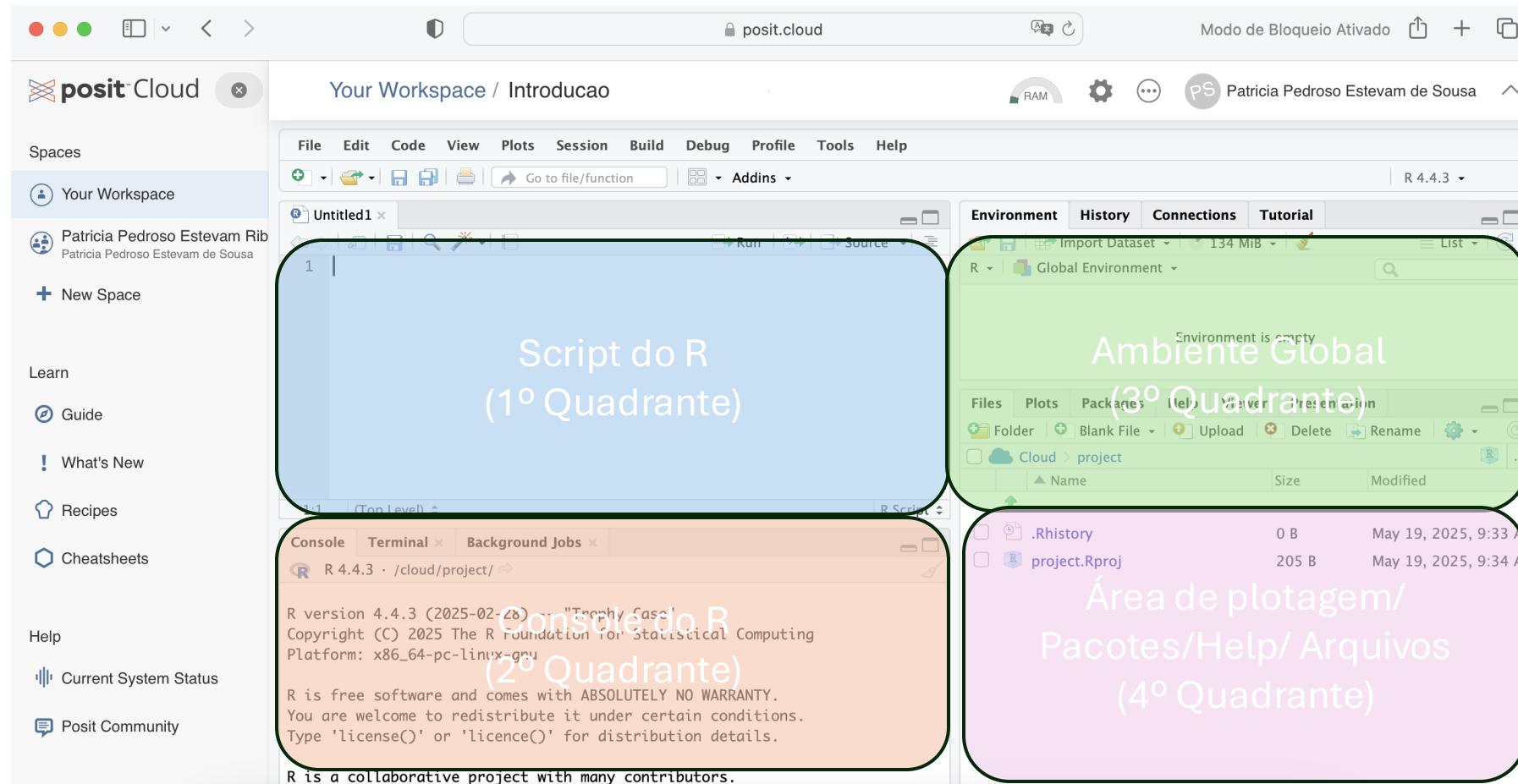
- O R é um software livre de linguagem interativa para computação estatística
- Possui uma vasta biblioteca de funções matemáticas, técnicas simples e sofisticadas para análise e visualização de dados
- RStudio é um Ambiente de Desenvolvimento Integrado (IDE) com várias funcionalidades e gratuito

# Linguagens de programação R e Python

- Programação em R On-line (<https://posit.cloud/>)



# Linguagens de programação R e Python



# Linguagens de programação R e Python

- O R é uma linguagem orientada a objetos, aos quais são atribuídos valores e expressões
- Atribuímos a um objeto de nome x o valor 10 fazendo:

`x <- 10` ou `x = 10`

- O nome do objeto precisa começar com uma letra, pode conter números, ponto, underline. Mas não pode conter símbolos como vírgula, ponto-e-vírgula ou espaço
- As informações ficam armazenadas na memória do computador, podendo ser acessadas, geradas, salvas, apagadas e manipuladas de diversas formas

# Exemplo de programação em R

The image shows a screenshot of the RStudio interface, which includes four main panels:

- Code Editor (Top Left):** Shows an R script named "Untitled1" with the following code:

```
1 x <- c(10, 20, 30, 40)
2 mean(x) # Média
3 sum(x) # Soma
4 length(x) # Tamanho do vetor
5
```
- Environment (Top Right):** Shows the global environment with the variable "x" defined as a numeric vector [1:4] containing 10, 20, 30, and 40.

Name	Type	Length	Size	Value
x	numeric	4	80 B	num [1:4] 10 20 30 ...
- Console (Bottom Left):** Shows the R console output for the same commands:

```
R 4.2.2 · ~/ 
> x <- c(10, 20, 30, 40)
> mean(x) # Média
[1] 25
> sum(x) # Soma
[1] 100
> length(x) # Tamanho do vetor
[1] 4
```
- Global Environment (Bottom Right):** Shows the global environment with the variable "x" defined as a numeric vector [1:4] containing 10, 20, 30, and 40, matching the entry in the Environment panel.

# Principais comandos e funções

Os comandos fundamentais permitem manipular dados e executar análises automatizadas:

- **Estruturas condicionais** (if, elif, else) para tomada de decisão
- **Loops** (for, while) executa uma sequência ou uma ação repetidamente até que uma condição seja atendida.
- **Funções** (def) que organiza e reutiliza o código para tarefas específicas

# Importação de dados

A manipulação de arquivos biológicos é essencial na bioinformática. Importar e processar dados em diversos formatos, como:

- **CSV (*Comma-Separated Values*)**: É um formato de arquivo de texto usado para armazenar dados tabulares, onde cada linha representa um registro e os valores são separados por vírgulas.
- **Txt**: Contém apenas caracteres sem formatação
- **Dbf**: Formato de banco de dados estruturado, originalmente criado para o software dBase
- **FASTQ/FASTA**: São amplamente utilizados na bioinformática para armazenar sequências biológicas
- **Xlsx**: Planilhas do excel.

# Importação de dados

```
# Carrega a biblioteca  
library(readr)
```

```
#Carregar os dados com extensão .csv  
arquivo <- file.choose()
```

```
# Verifica se o R está enxergando o arquivo  
file.exists(arquivo)
```

```
# Carrega os dados do arquivo "Arquivo2.csv"  
dados1 <- read_csv(arquivo)  
dados1
```

```
# Carrega a biblioteca  
library(readr)
```

```
#Carregar os dados com extensão .txt  
arquivo <- file.choose()
```

```
# Verifica se o R está enxergando o arquivo  
file.exists(arquivo)
```

```
# Carrega os dados do arquivo "dadosCusto.txt"  
dados2 <- read.delim(arquivo, sep = " ", header = TRUE)  
dados2
```

# Importação de dados

```
library(rmarkdown)  
  
#Carregar os dados  
arquivo <- file.choose()  
  
# Verifica se o R está enxergando o arquivo  
file.exists(arquivo)  
  
# Lê o arquivo DBF  
dados <- read.dbf(arquivo, as.is = TRUE)  
  
# Visualização dos dados  
head(dados)  
  
#Carregar os dados com extensão .txt  
arquivo <- file.choose()  
  
# Verifica se o R está enxergando o arquivo  
file.exists(arquivo)  
  
# Carrega os dados do arquivo "dadosCusto.txt"  
dados2 <- read.delim(arquivo, sep = " ", header = TRUE)  
  
# Visualização dos dados  
head(dados2)
```

# Linguagens de programação R e Python

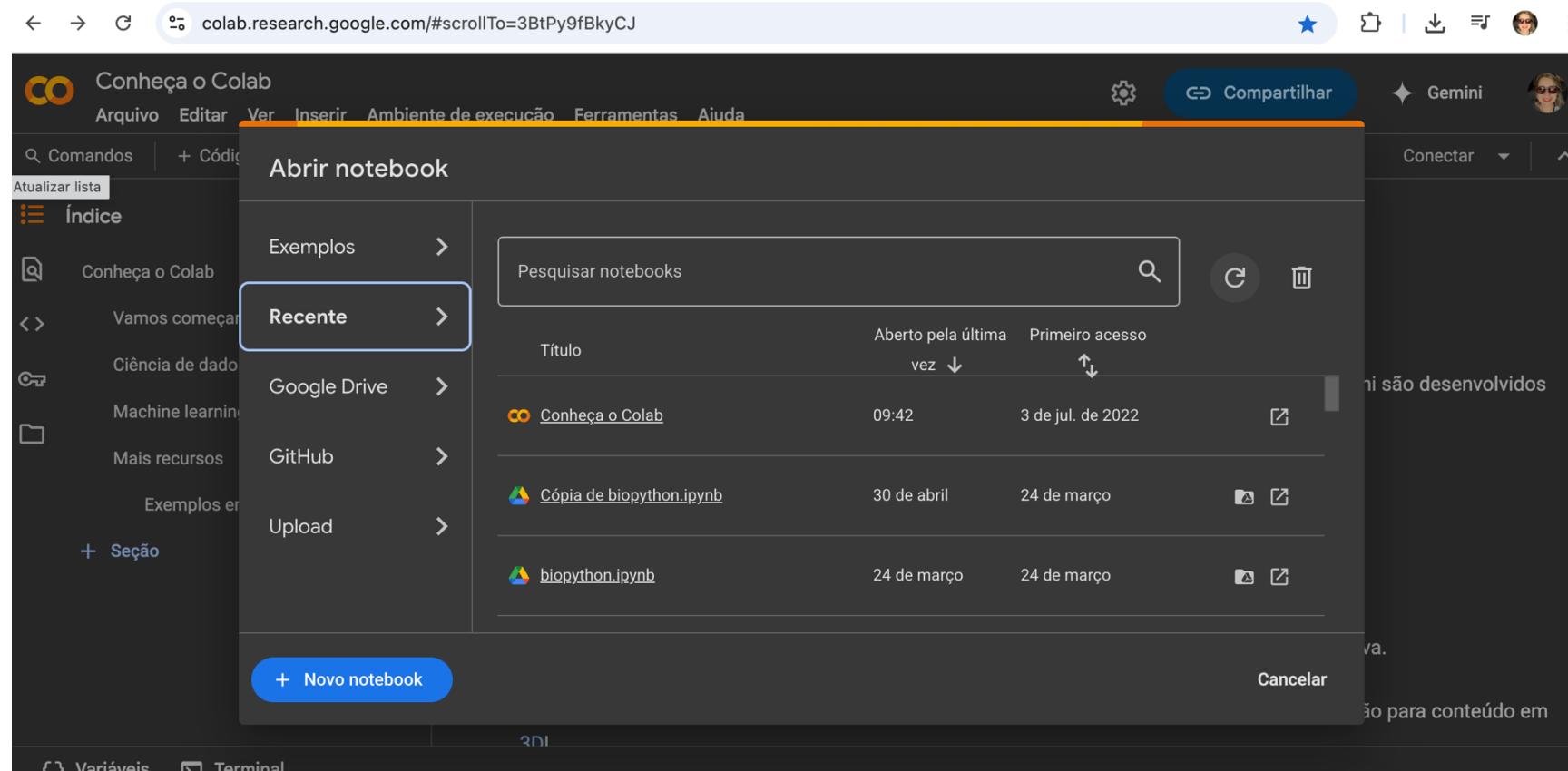


Python é uma linguagem de programação interpretada, de alto nível, e de uso geral, conhecida pela sua sintaxe clara e legível. É utilizada em uma vasta gama de aplicações, desde desenvolvimento web e automação de tarefas, até análise de dados e ciência de dados.

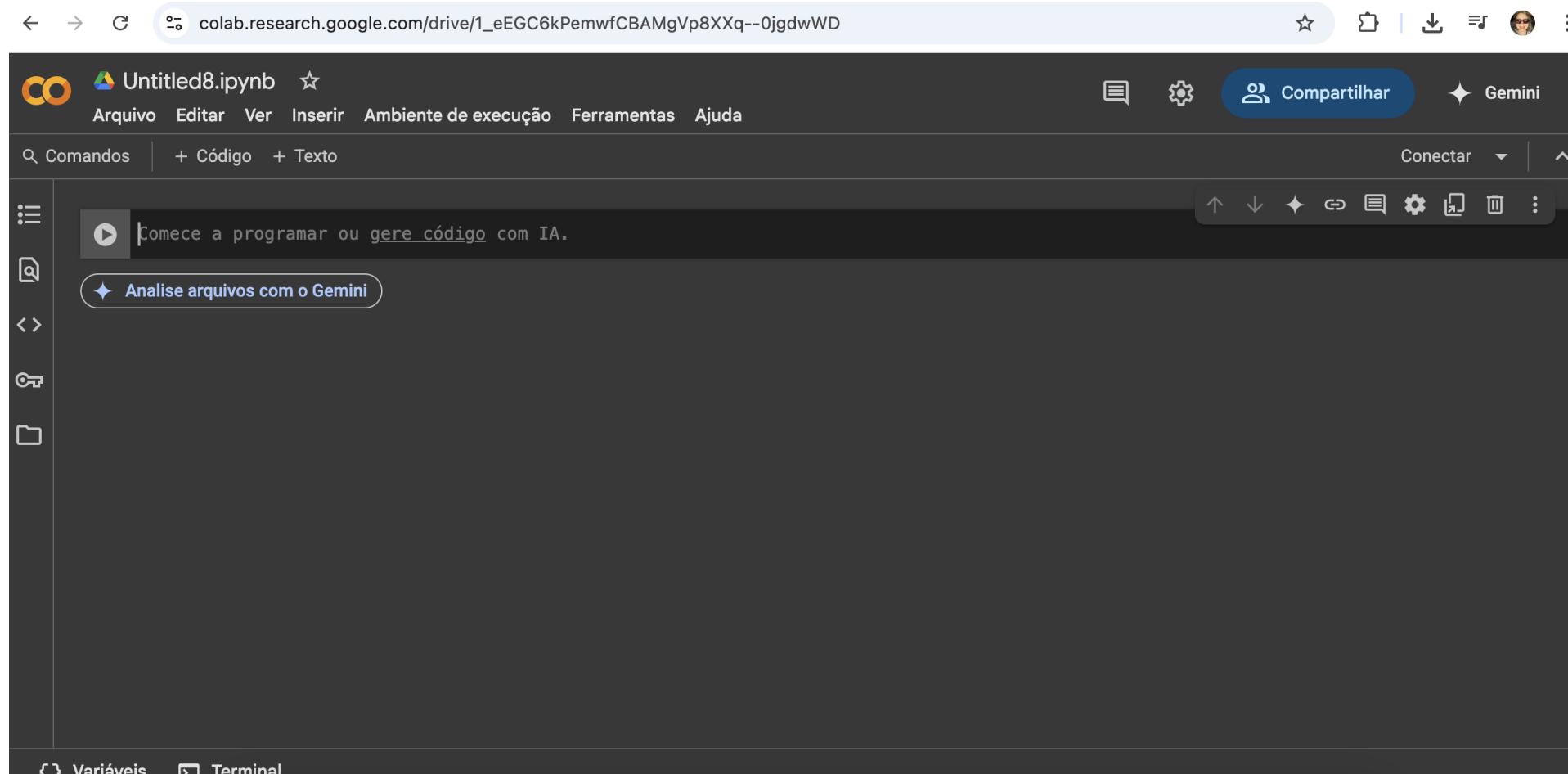
# Linguagens de programação R e Python

Programação em Python usando o Colab - Google

<https://colab.research.google.com/>



# Linguagens de programação R e Python



A bioinformática permite estudar por exemplo quais genes estão ativos em determinadas condições biológicas, como doenças e tratamentos. Métodos comuns incluem:

- **1. RNA-Seq (Sequenciamento de RNA):** Técnica de alto rendimento que permite quantificar a expressão gênica em diferentes condições
- **2. Microarrays de Expressão Gênica:** Método que usa chips de DNA para medir a expressão de milhares de genes simultaneamente.
- **3. SAGE (*Serial Analysis of Gene Expression*):** Técnica que gera pequenos fragmentos de RNA para análise quantitativa da expressão gênica.
- **4. Métodos de Predição de Genes:** Algoritmos que identificam genes ativos com base em padrões de sequência e expressão.

# Exemplo: Sequenciamento de DNA

## Exemplo

Sequenciamento permite determinação da ordem exata dos nucleotídeos (**A, T, C, G**) em uma sequência de DNA.

## Alinhamento de duas sequências v e w

ATCGTAC

ATGTTAT

Avaliando Mutações ou  
polimorfismos

Ferramentas computacionais (como **BLAST** ou **ClustalW**) fazem o alinhamento para encontrar semelhanças e variações.

# Exemplo: Sequenciamento de DNA

[blast.ncbi.nlm.nih.gov/Blast.cgi](http://blast.ncbi.nlm.nih.gov/Blast.cgi)

**National Library of Medicine**  
National Center for Biotechnology Information

**Conecte-se**

**BLAST®**

Lar Resultados recentes Estratégias Salvas Ajuda

**Ferramenta básica de busca de alinhamento local**

O BLAST encontra regiões de similaridade entre sequências biológicas. O programa compara sequências de nucleotídeos ou proteínas com bancos de dados de sequências e calcula a significância estatística.

[Saber mais](#)

**NEWS**

Seg, 17 de março de 2025  
As melhorias incluem a atualização para o GCP Artifact Registry e melhor tratamento do status de conclusão do trabalho no Kubernetes versão 1.30+.

ElasticBLAST 1.4.0 já está disponível! [Mais notícias do BLAST...](#)

**Explosão na Web**

**Nucleotide BLAST**  
nucleotide ▶ nucleotide

**blastx**  
translated nucleotide ▶ protein

**tblastn**  
protein ▶ translated nucleotide

**Protein BLAST**  
protein ▶ protein

[genome.jp/tools-bin/clustalw](http://genome.jp/tools-bin/clustalw)

**Multiple Sequence Alignment by CLUSTALW**

**ETE3 MAFFT CLUSTALW PRRN**

**General Setting Parameters:**  
Output Format: CLUSTAL   
Pairwise Alignment:  FAST/APPROXIMATE  SLOW/ACCURATE

**Enter your sequences (with labels) below (copy & paste):**  PROTEIN  DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

**Or give the file name containing your query**  
Escolher arquivo | Nenhum arquivo escolhido

**Execute Multiple Alignment** **Reset**

**More Detail Parameters...**

**Pairwise Alignment Parameters:**  
For FAST/APPROXIMATE:

# Exemplo: Sequenciamento de DNA

← → G  biopython.org

[Edite esta página no GitHub](#)



# Biopython

Veja também nosso [feed de notícias](#).

## Introdução

Biopython é um conjunto de ferramentas disponíveis gratuitamente para computação biológica, escritas em [Python](#) por uma equipe internacional de desenvolvedores.

Trata-se de um esforço colaborativo distribuído para desenvolver bibliotecas e aplicativos Python que atendam às necessidades do trabalho atual e futuro em bioinformática. O código-fonte é disponibilizado sob a [Licença Biopython](#), que é extremamente liberal e compatível com quase todas as licenças do mundo.

Somos um projeto membro da [Open Bioinformatics Foundation \(OBF\)](#), que cuida do nosso nome de domínio e da hospedagem da nossa lista de e-mails, etc. A OBF costumava hospedar nosso repositório de desenvolvimento, rastreador de problemas e site, mas agora eles estão no [GitHub](#).

Esta página ajudará você a baixar e instalar o Biopython e começar a usar as bibliotecas e ferramentas.

- Ferramentas Python para Biologia Molecular Computacional
- Documentação
- Download
- Listas de discussão
- Notícias
- Contribuidores do Biopython
- Scriptcentral

# Exemplo: Sequenciamento de DNA

```
# instalando o Biopython no Colab
!pip3 install biopython
```

```
Collecting biopython
  Downloading biopython-1.85-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (13 kB)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from biopython) (2.0.2)
  Downloading biopython-1.85-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.3 MB)
    3.3/3.3 MB 30.0 MB/s eta 0:00:00
Installing collected packages: biopython
Successfully installed biopython-1.85
```

```
[3] # importando todo o pacote (testa se ele está instalado)
import Bio
```

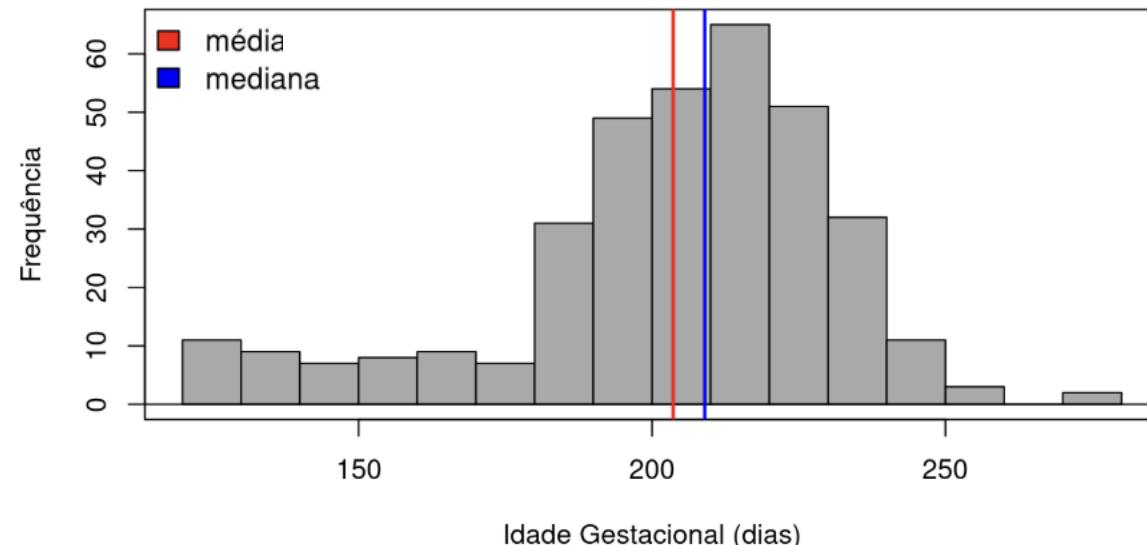
Vamos Praticar?



SequenciamentoDNA.ipynb

# Análise Estatística

- As análises descritivas ajudam a interpretar dados biológicos, revelando padrões e variações dentro de populações genéticas. Alguns conceitos essenciais:



**Média:** Valor médio dos dados, calculado pela soma dos valores dividida pelo número total de observações.

**Mediana:** O valor central de um conjunto de dados ordenado; separa a metade inferior da metade superior.

# Análise Estatística

- **Frequência:** Número de ocorrências de cada categoria ou valor dentro do conjunto de dados.

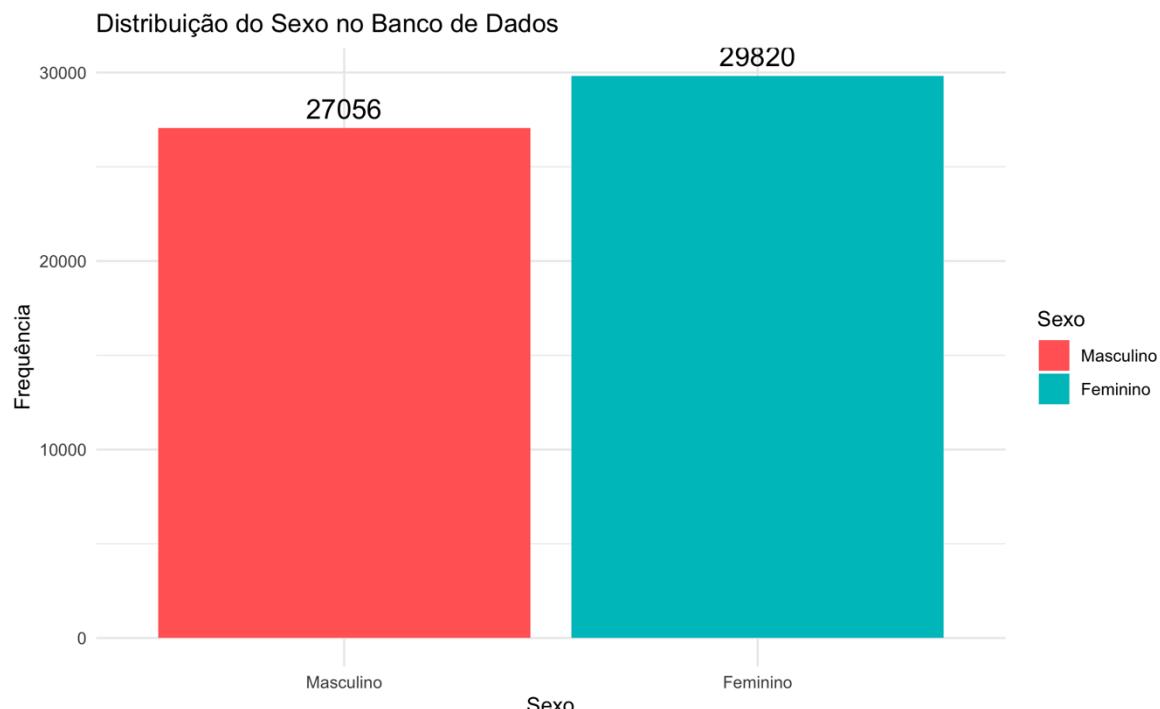


Tabela 1 – Tabela de frequência do sexo  
n=56876

Variáveis	Categoría	N	%
Sexo	Masculino	27056	47,57
Sexo	Feminino	29820	52,43

Figura 1 - Gráfico de frequência do sexo no Banco de Dados.

# Análise Estatística

- **Desvio padrão:** Medida de dispersão que indica o quanto espalhados os dados estão em relação à média.

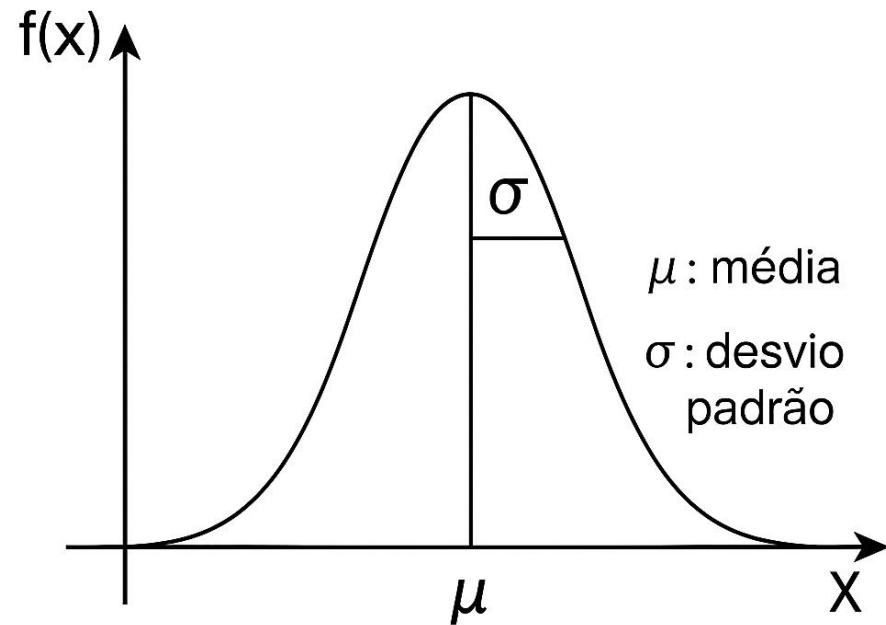
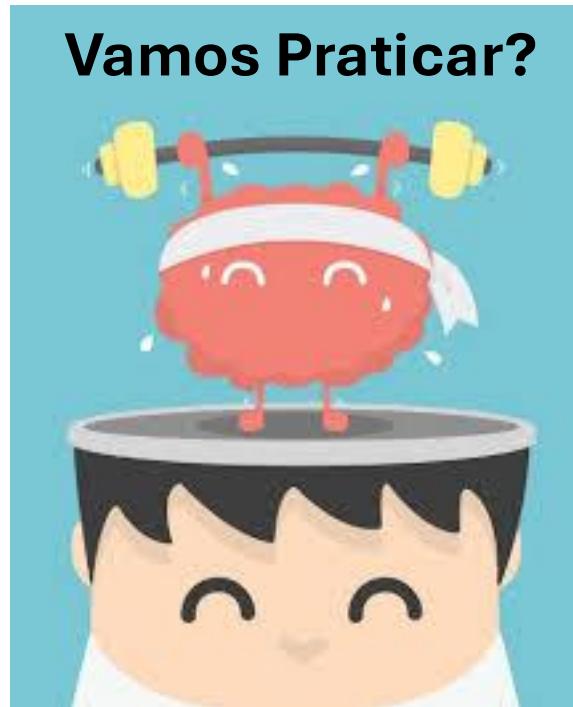


Tabela 2 – Tabela com média da idade dos participantes.

Variável	Média	Desvio Padrão
Idade	61	16,417

# Exemplo Prático - Análise Estatística

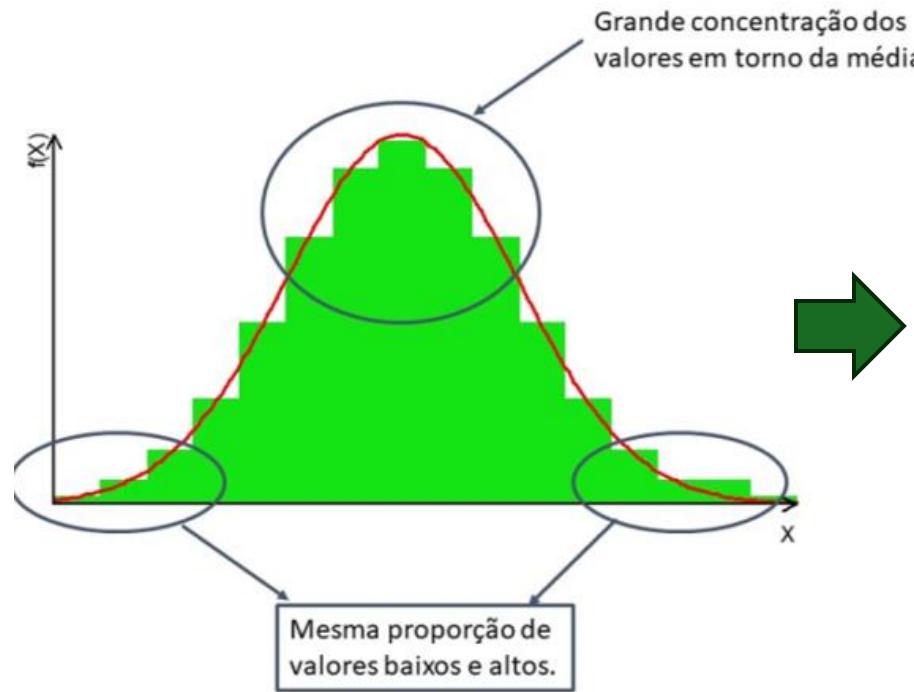
- Aula Pratica 2.R



# Testes estatísticos

- Os testes estatísticos são utilizados para determinar se há diferenças significativas entre grupos biológicos.
- **Primeira passo é verificar se os dados são normais?**
- Dados normais são aqueles que seguem uma distribuição normal, também chamada de **distribuição de Gauss** ou **curva em formato de sino**.

# Testes estatísticos



Distribuição normal

# Testes estatísticos

- **Teste de normalidade**

Verifica se os dados seguem uma distribuição normal

- **Teste de Shapiro-Wilk** → Para amostras pequenas ( $n < 50$ )
- **Teste de Kolmogorov-Smirnov** → Para amostras grande.

**$p > 0,05$**  → Os dados seguem distribuição normal (aceitamos  $H_0$ ).

**$p < 0,05$**  → Os dados não seguem distribuição normal, e devemos usar testes não paramétricos.

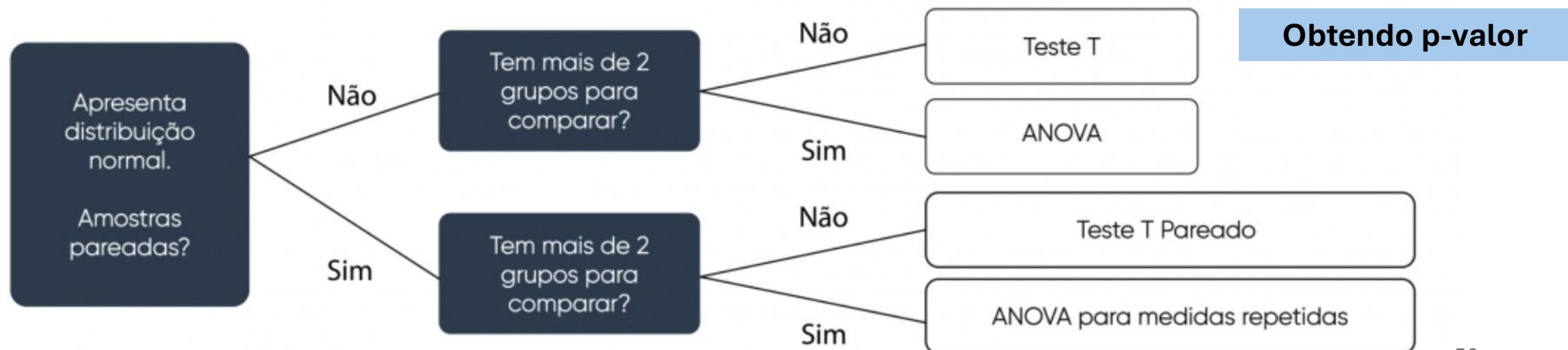
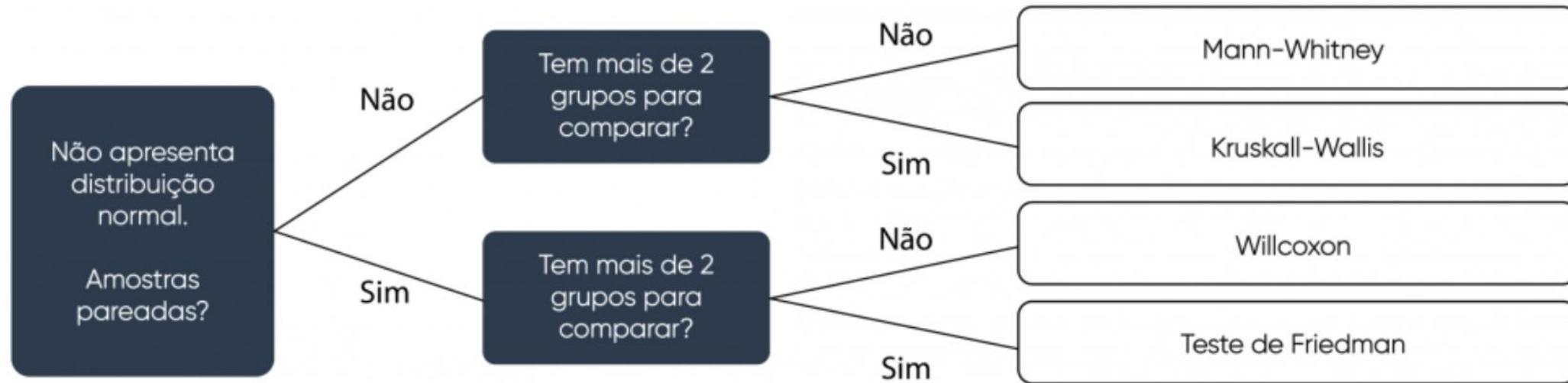
# Testes estatísticos

- **Teste de hipótese**

É um método estatístico usado para determinar se há diferença significativa entre grupos ou variáveis, baseado em uma amostra de dados.

**p < 0,05** → Existe diferença significativa entre os grupos, rejeitamos  $H_0$ .

**p > 0,05** → Não há evidências para rejeitar  $H_0$ , os grupos são estatisticamente iguais.



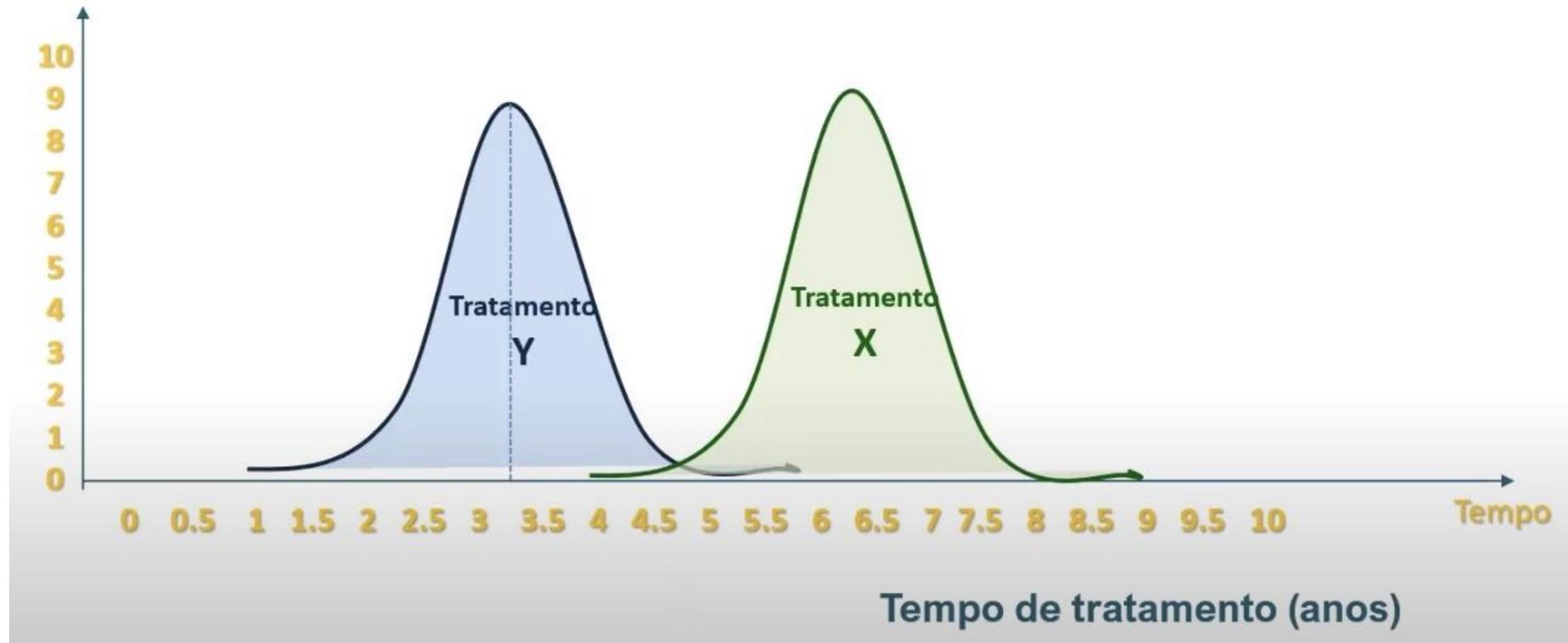
# Testes estatísticos



# Testes estatísticos

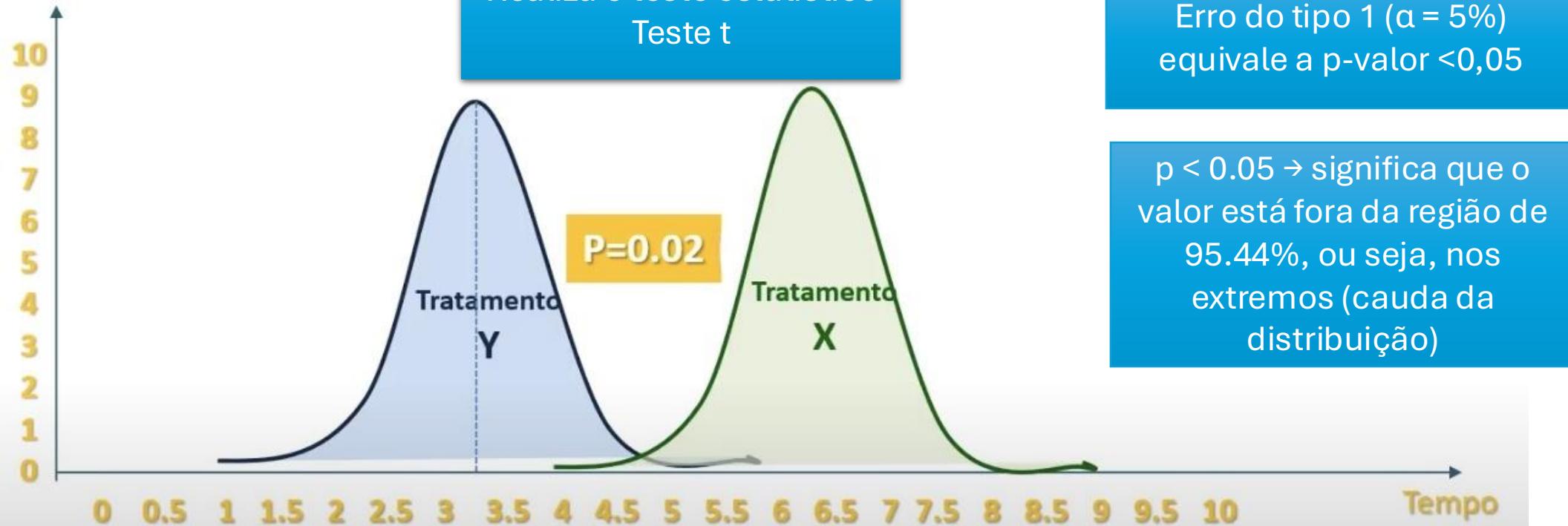


# Testes estatísticos



# Testes estatísticos

Realiza o teste estatístico  
Teste t



Erro do tipo 1 ( $\alpha = 5\%$ )  
equivale a p-valor <0,05

$p < 0.05 \rightarrow$  significa que o  
valor está fora da região de  
95.44%, ou seja, nos  
extremos (cauda da  
distribuição)

$p=0,02$  é menor que  $p<0,05$ , sendo assim  
demostrando uma diferença significativa

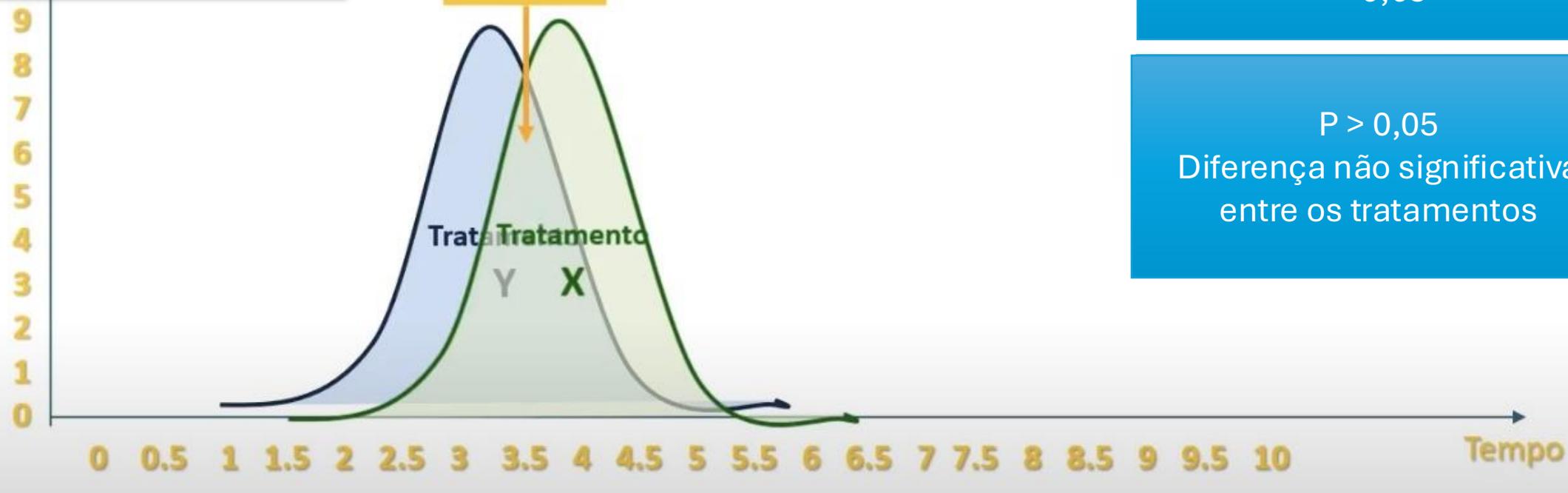
**Tempo de tratamento (anos)**

Indicando que o tratamento X teve um resultado  
diferente do Y

Um outra análise

# Testes estatísticos

Realiza o teste estatístico

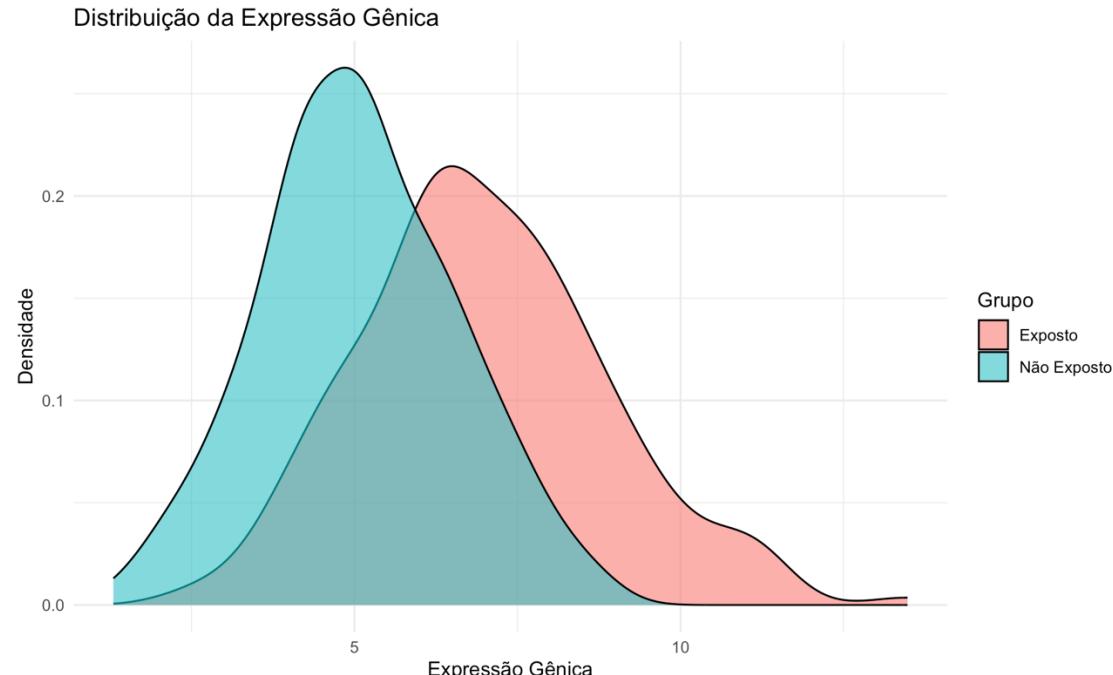
**P=0.60**

p obtido é maior que p-valor  
 $<0,05$

$P > 0,05$   
Diferença não significativa  
entre os tratamentos

# Visualização de Dados

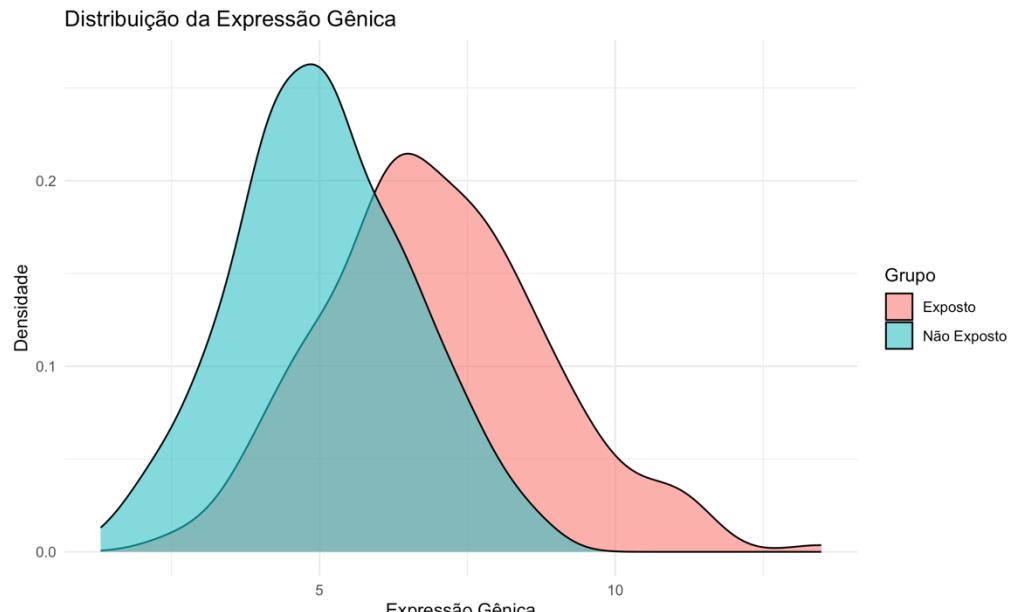
As técnicas de visualização de dados refere-se à apresentação gráfica desses dados, proporcionando uma compreensão intuitiva das relações, tendências e padrões.



# Visualização de Dados

## Gráfico de Densidade

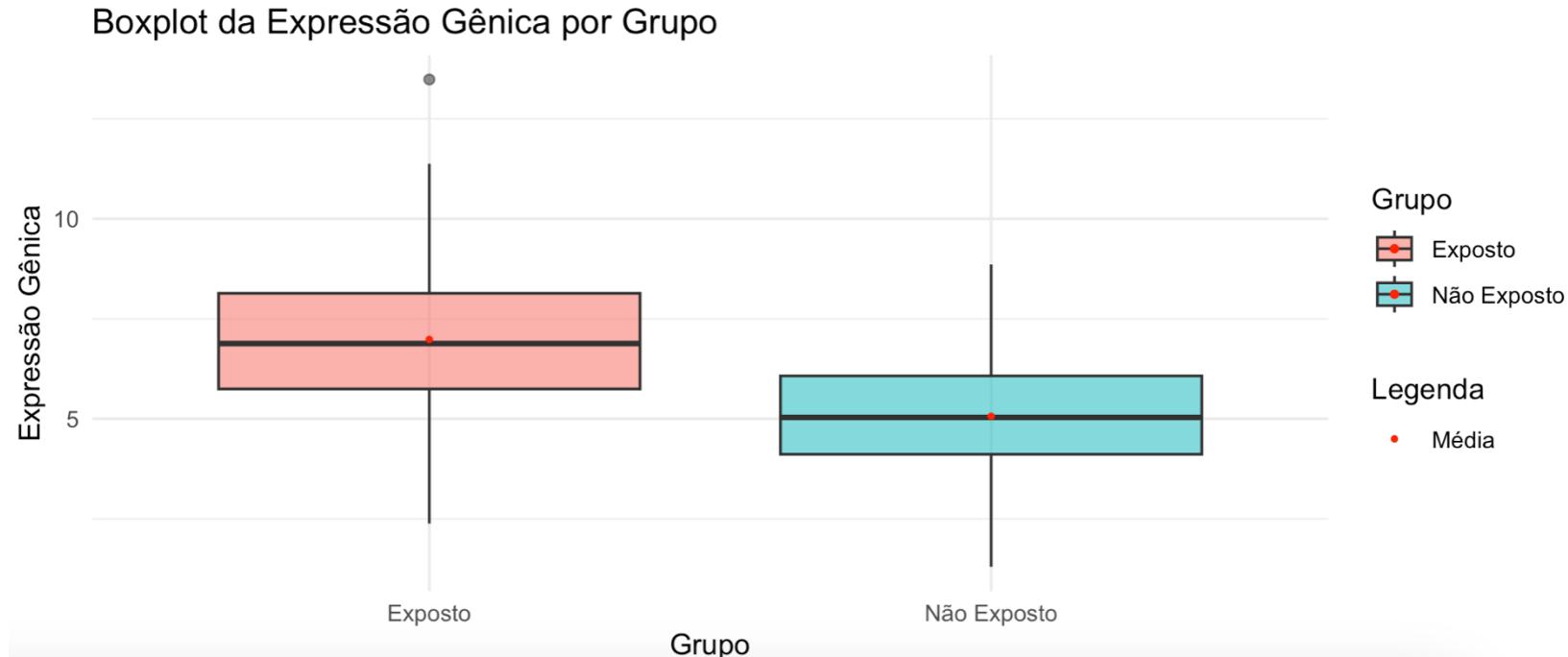
Utilizado para analisar a distribuição de valores em um conjunto de dados biológicos. Exemplo: distribuição de expressão de genes em diferentes amostras.



# Visualização de Dados

## Gráfico Box Plot

Mostra a variação dos dados e possíveis outliers. É útil para comparar níveis de expressão gênica entre grupos de pacientes.

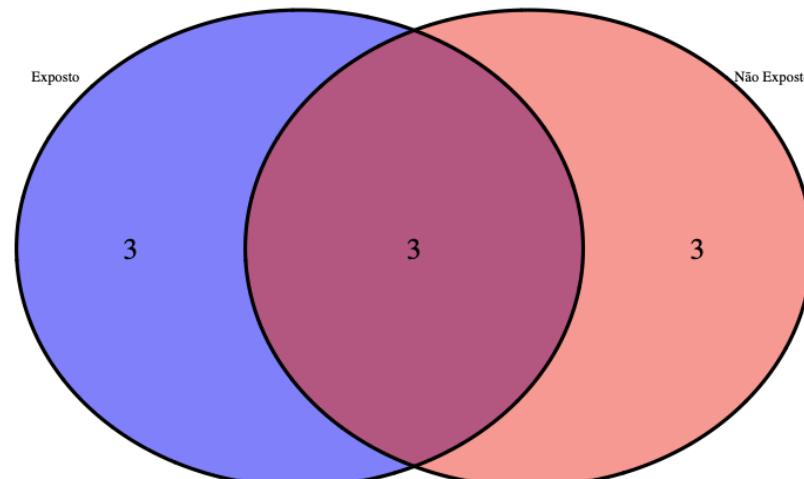


# Visualização de Dados

## Gráfico Venn

- Usado para visualizar interseções entre conjuntos de dados.  
Exemplo: genes comuns entre diferentes organismos ou condições experimentais.

Diagrama de Venn – Expressão Gênica



# Visualização de Dados

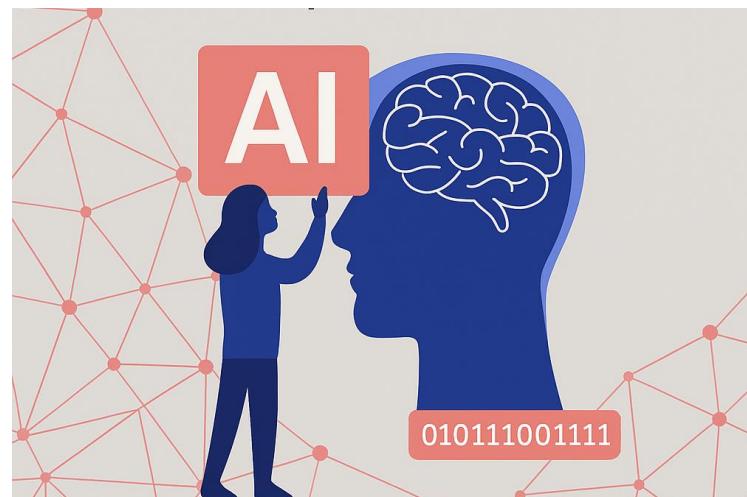
- Aula Pratica 3.R



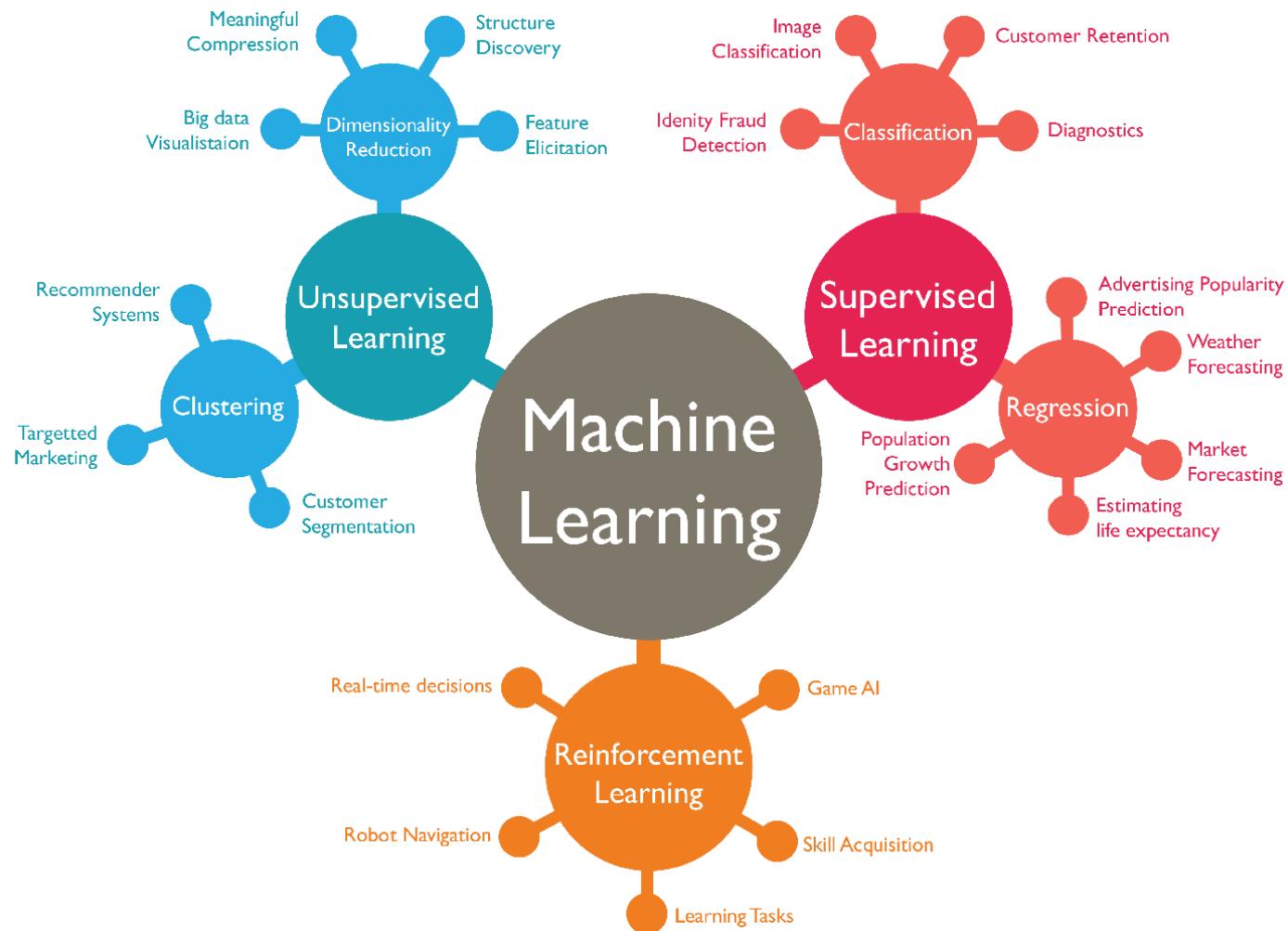
# Aprendizado de Máquina

- **O que é Aprendizado de Máquina?**

Aprendizado de Máquina (*Machine Learning*) é um ramo da inteligência artificial que permite que os computadores aprendam padrões a partir de dados, sem serem explicitamente programados para cada tarefa.



# Aprendizado de Máquina



# Aprendizado de Máquina Supervisionado

É onde o modelo aprende a partir de um conjunto de dados rotulados. Ou seja, ele recebe exemplos já classificados e tenta identificar padrões para fazer previsões futuras com base nesses dados.

## Aprendizado de Máquina Supervisionado

Dados usados para construir o modelo



Predição

Maçã

Rótulos

Maçãs



Modelo



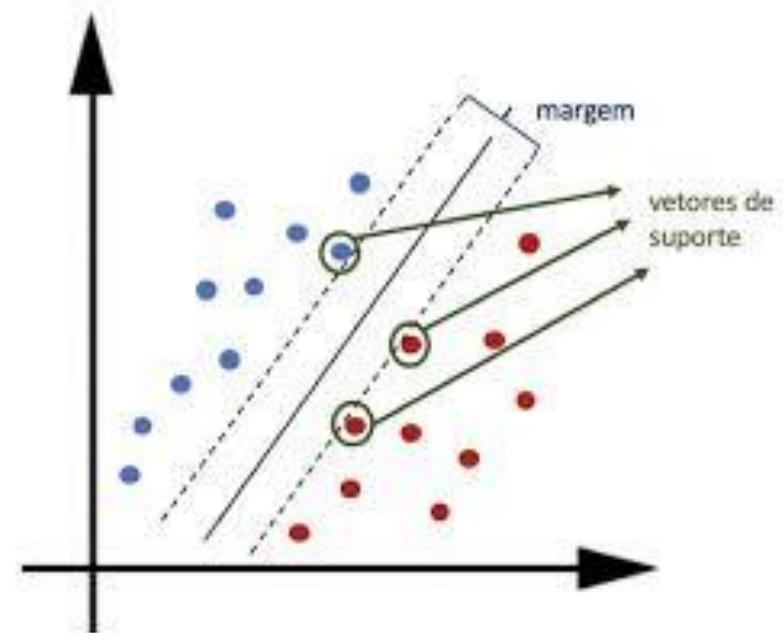
Dados novos sem rótulo

(estes são os que queremos prever o que são)

# Aprendizado de Máquina Supervisionado

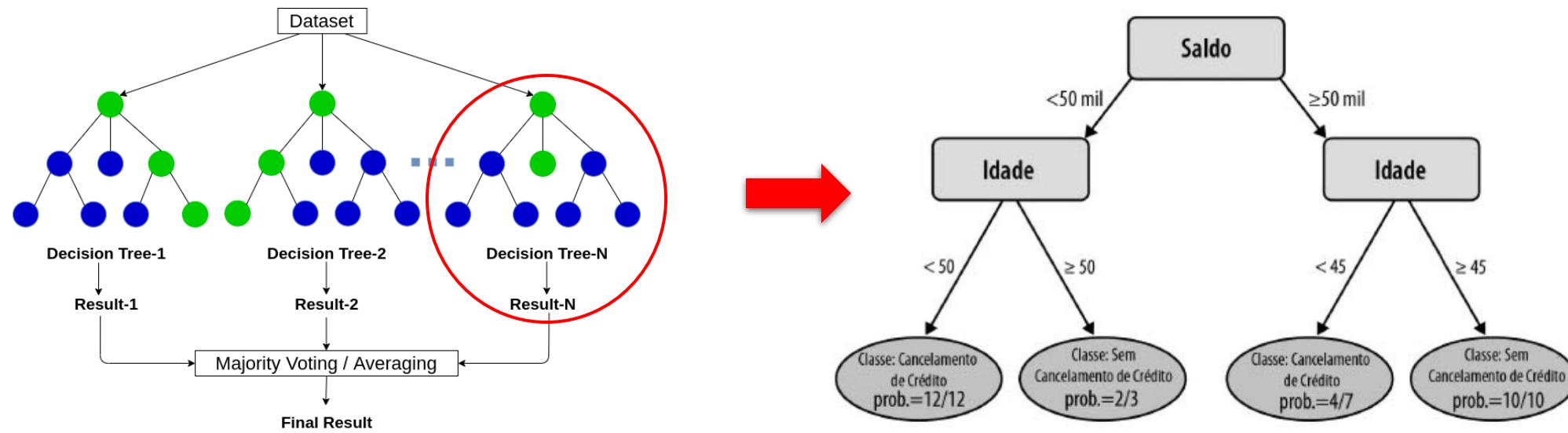
- **SVM (*Support Vector Machine*):**

É um algoritmo supervisionado que cria hiperplanos para separar dados em classes, maximizando a distância entre esses planos e os pontos mais próximos, conhecidos como vetores de suporte, para otimizar a precisão da classificação.



# Aprendizado de Máquina Supervisionado

- **Random Forest**
- É um algoritmo que combina várias árvores de decisão treinadas em subconjuntos aleatórios para gerar variações mais precisas e robustas, reduzindo o risco de *overfitting* e sendo eficazes para dados complexos e de alta dimensionalidade.

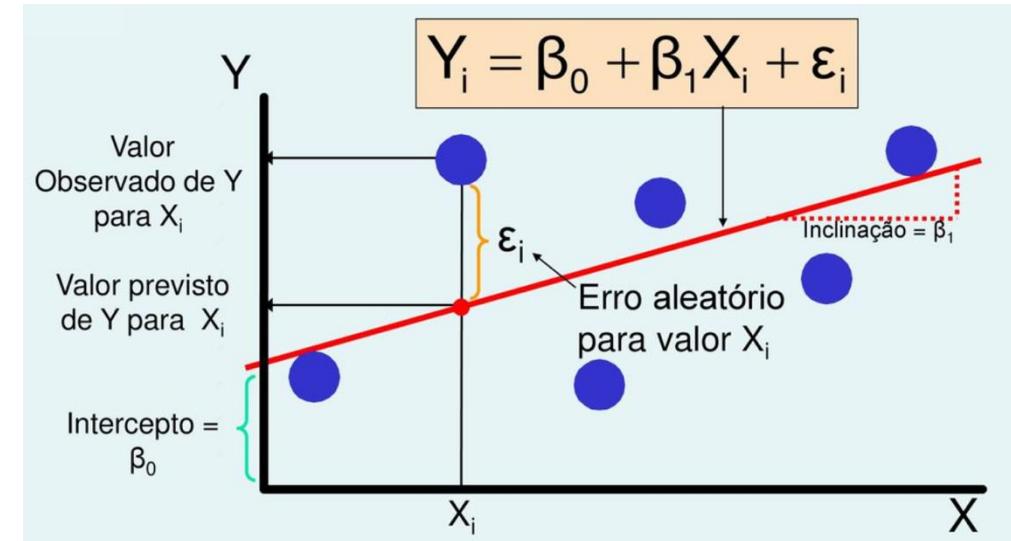


# Aprendizado de Máquina Supervisionado

- Regressão Linear

A regressão linear é usada para modelar a relação entre uma variável dependente (y) e uma variável independente (x). Ela se baseia na equação de uma linha reta:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



**y<sub>i</sub> (Variável dependente):** O valor que queremos prever ou explicar,

**B<sub>0</sub> (Intercepto):** Representa o valor y quando X = 0,

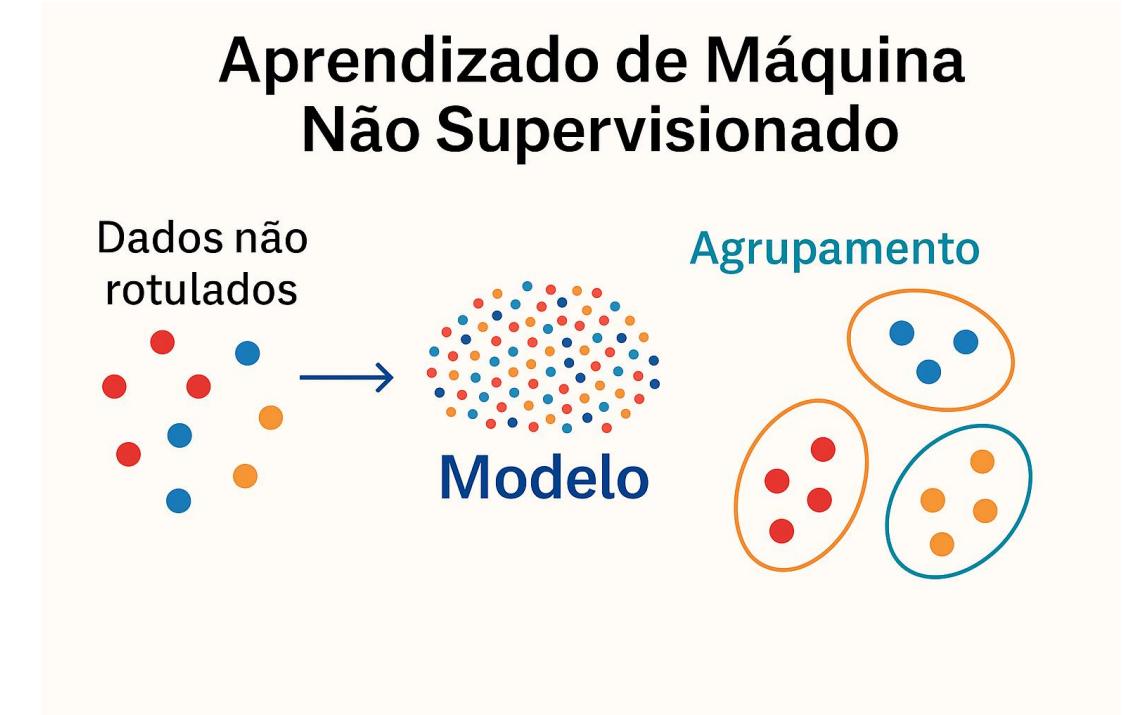
**B<sub>i</sub> (Coeficiente angular):** Representa o aumento esperado em y quando aumenta uma unidade),

**x<sub>i</sub> (Variável independente):** O valor que influencia y.

**E<sub>i</sub> (Erro aleatório ou resíduo):** Diferença entre o valor previsto pelo modelo e o valor real.

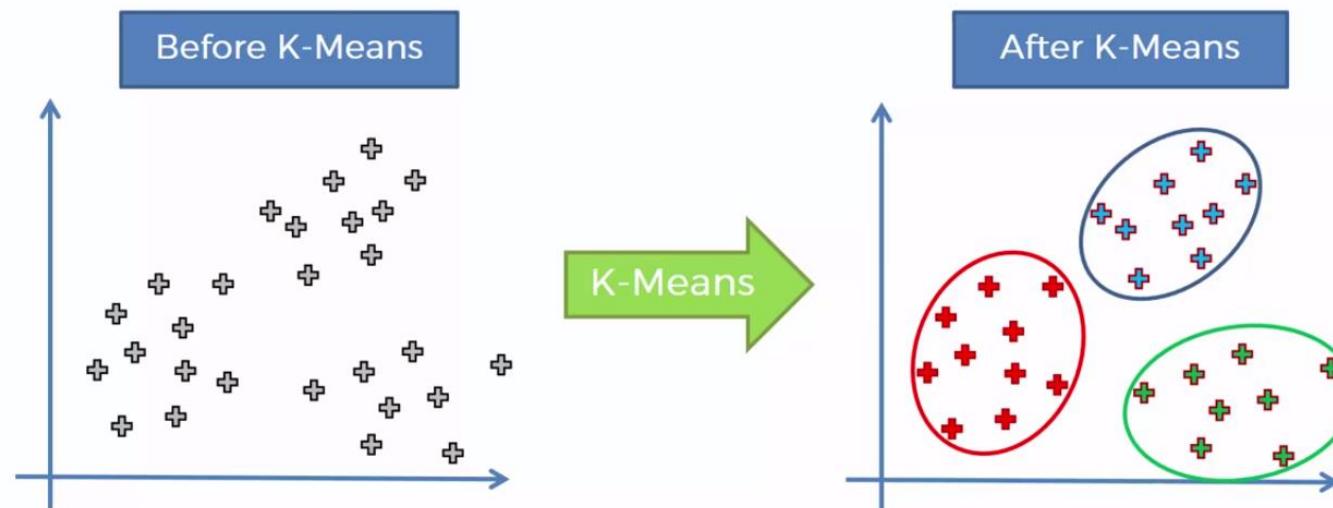
# Aprendizado de Máquina Não Supervisionado

Se caracterizam pela capacidade de agrupar dados com base em suas características semelhantes, sem a necessidade de rótulos pré-existentes.



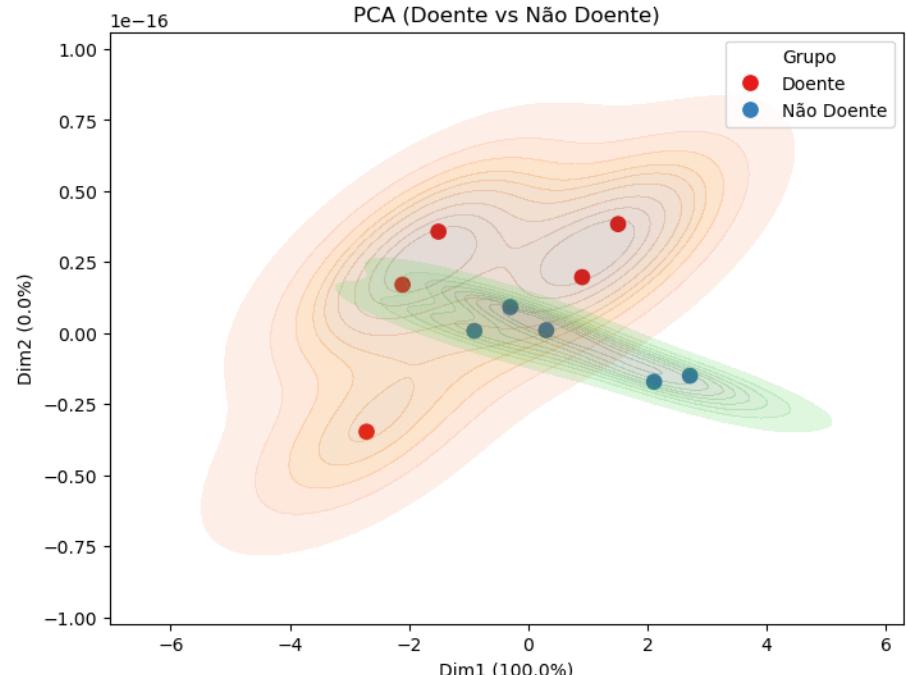
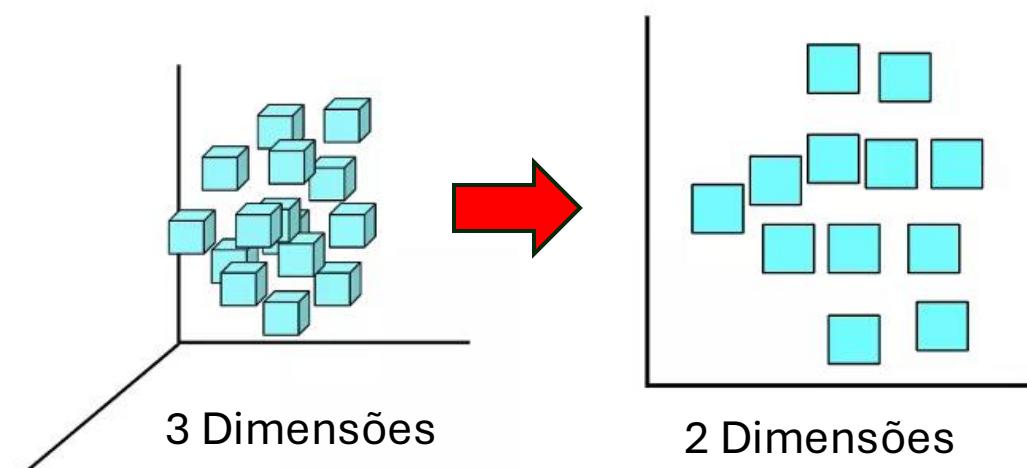
# Aprendizado de Máquina Não Supervisionado

- **K-Means:** agrupa amostras biológicas com base em similaridades



# Aprendizado de Máquina Não Supervisionado

- PCA (*Principal Component Analysis*): A análise de componentes Principais reduz a dimensionalidade de dados para facilitar a interpretação



# Aprendizado de Máquina

- Aula Pratica 4.ipynb



# Obrigada!



Dra.patriciapedroso



patriciapedrosoestevam@gmail.com