

# Biologia e Ciência de Dados

*O que eu tenho haver?*

---

Marília Melo Favalesso



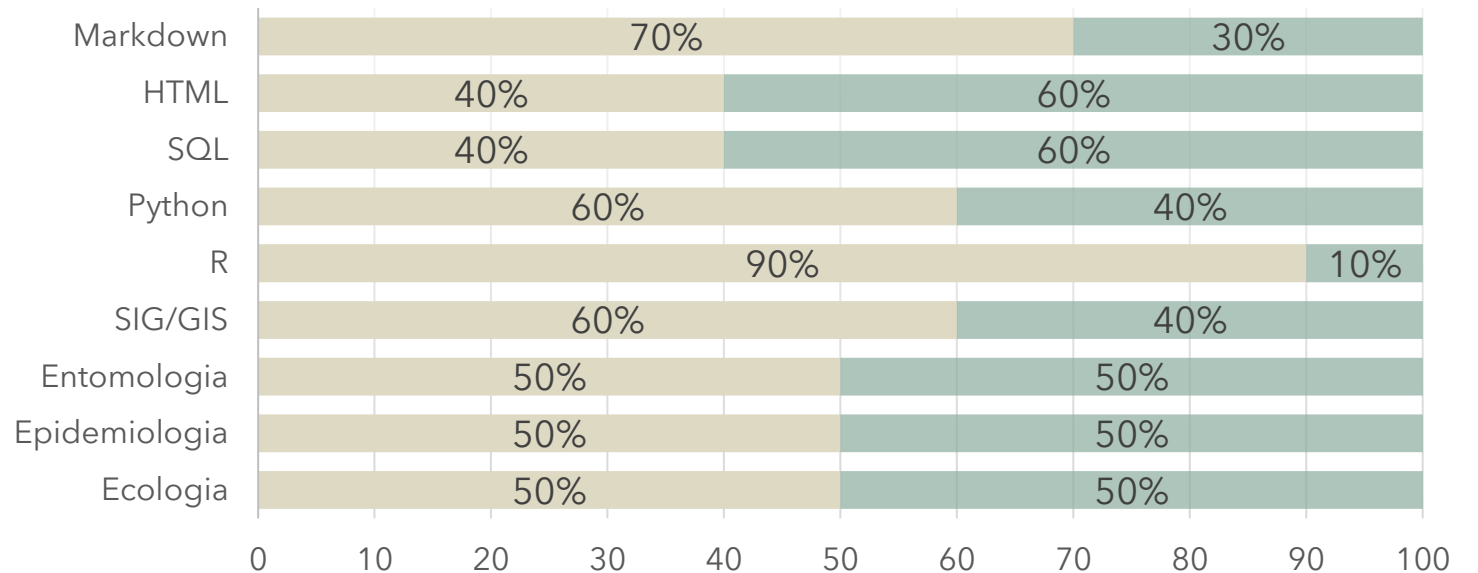
# Marília Melo Favalesso



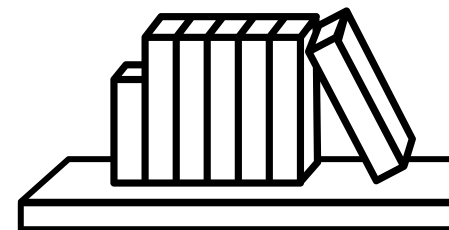
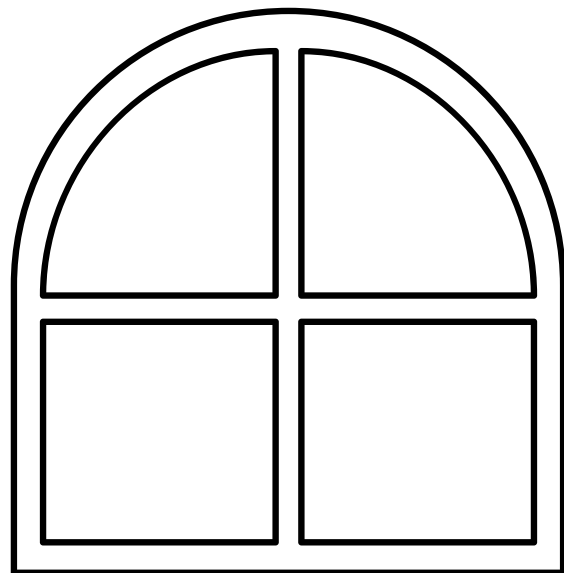
[www.mmfava.com](http://www.mmfava.com)

- ✓ Técnica Ambiental (CEEP – Cascavel – PR)
- ✓ **Bióloga** (UFPR – Palotina – PR)
- ✓ Consultoria e assessoria em Bioestatística (Cascavel – PR)
- ✓ Mestre em Ciências Ambientais (UNIOESTE – Cascavel – PR)
- Doutoranda em Ecoepidemiologia (UBA – Buenos Aires – AR)
- ✓ **Cientista de dados** (Hospital Israelita Albert Einstein – SP)

## Skills



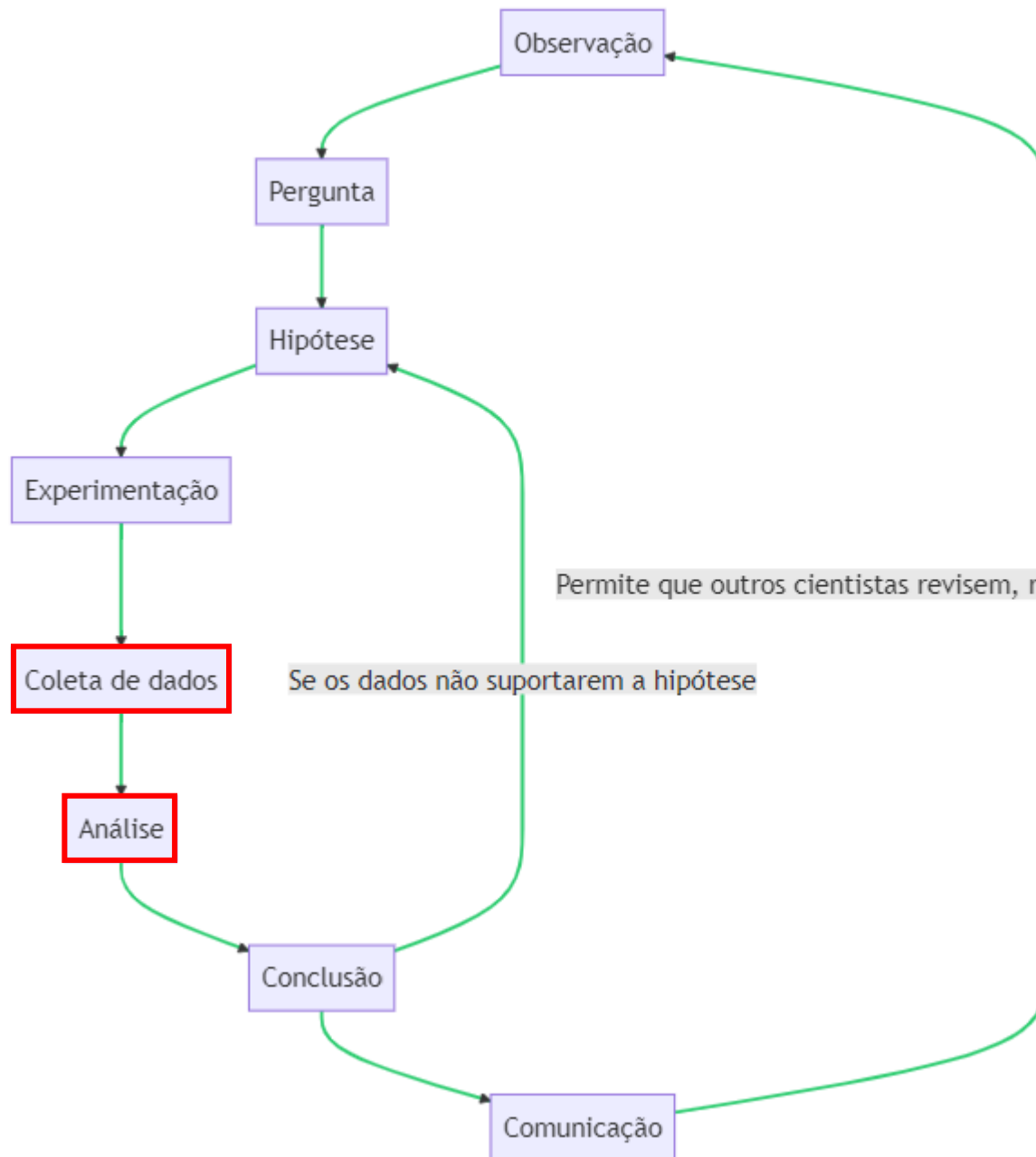




O que você quer “ser”  
quando se formar?

Independentemente do tipo de biólogo@ que você vai ser,  
você com certeza terá que...

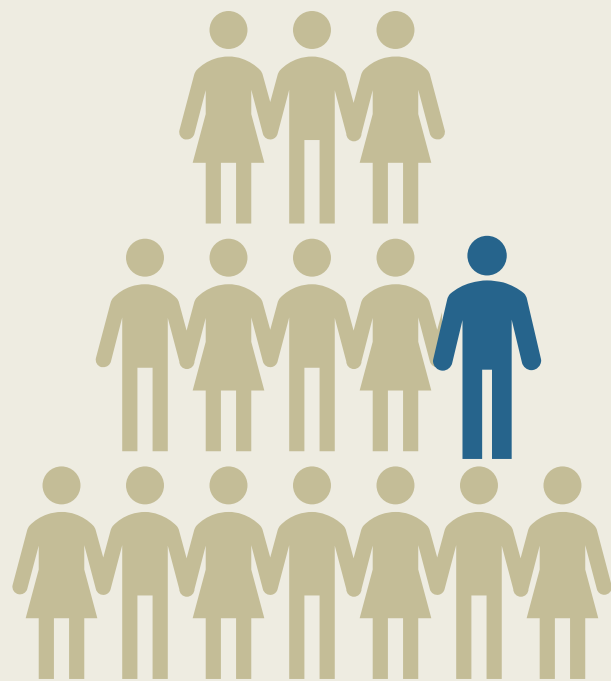
# Coletar e/ou Analisar dados



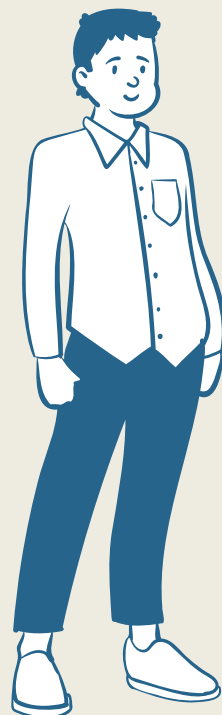


Charles Darwin  
O "paizão" da biologia

# Mas o que são dados?



Unidades observacionais



Unidade observacional



Temperatura: 37°C



Glóbulos brancos: 15g/100ml



Stress: Relaxado



Dengue: Negativo



Idade: 40 anos

Os **dados** são **valores** derivados de qualquer tipo de medição, contagem ou observação obtidos a partir de **unidades observacionais** (menor unidade de medida)

(Callegari-Jacques 2003)

# Dados são Variáveis

37°C



Unidade observacional 1

36,8°C

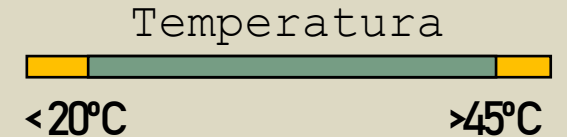


Unidade observacional 2

38°C



Unidade observacional 3



Uma **variável** possui a capacidade de assumir **qualquer valor** dentro de um **conjunto de valores possíveis**



# Tipos de Variáveis

## **Numéricas** (ou quantitativas)

Dados são valores numéricos que expressam quantidades

### **Discretas**

Os dados numéricos discretos possuem valores inteiros; não podem ser fracionados.

**Ex. Número de pacientes**

### **Contínuas**

Dados numéricos contínuos podem ter valores decimais ou fracionários. É possível particionar um dado contínuo infinitamente.

**Ex. Temperatura, Número de glóbulos brancos**

## **Categóricas** (ou qualitativas)

Dados de natureza não-numérica que possuem como valores categorias

### **Nominais**

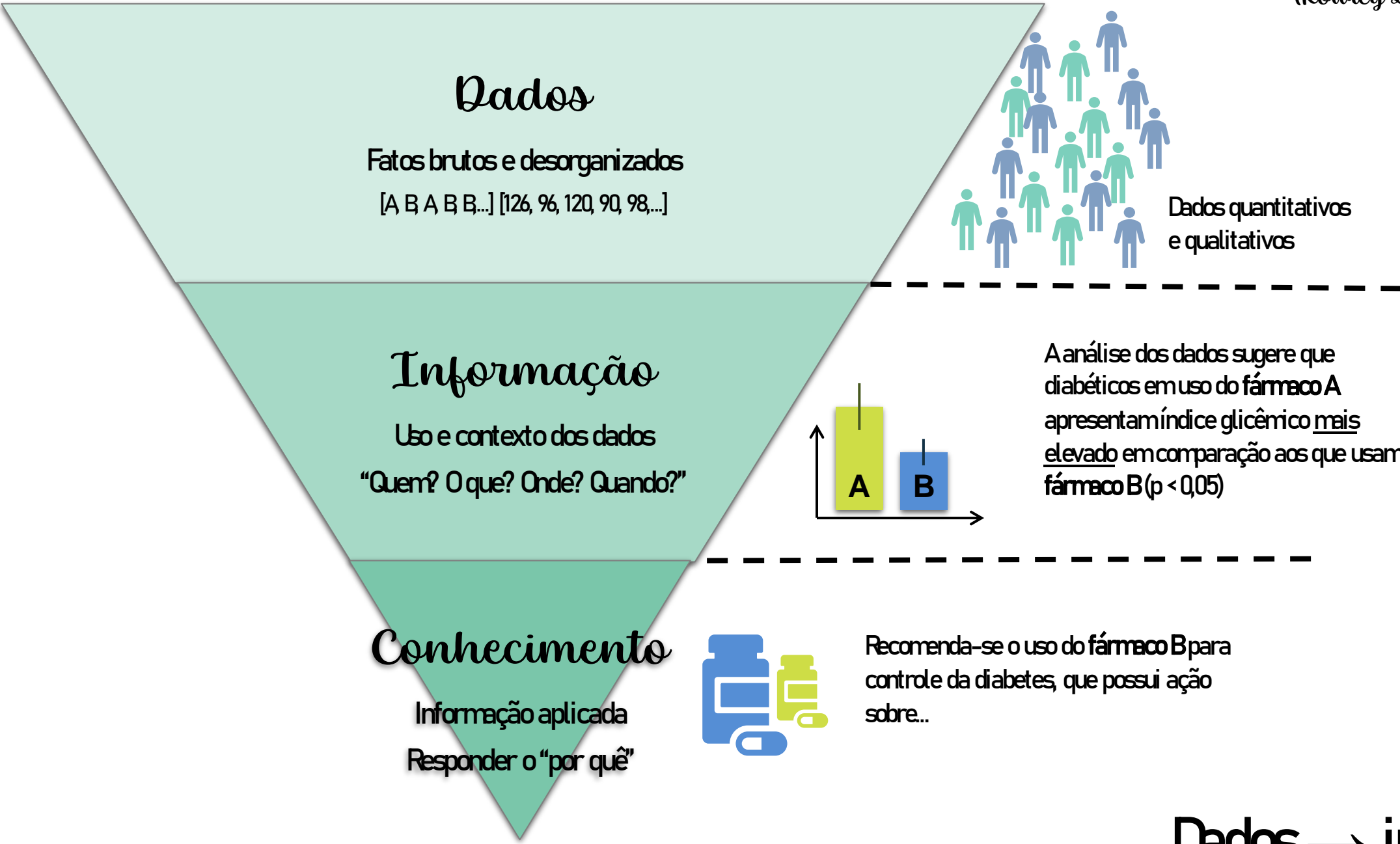
Os dados são categorias sem ordenação.

**Ex. Positivo ou negativo para dengue**

### **Ordinais**

Os dados são categorias que apresentam ordenação por grau de intensidade, desde que ela seja inerente à variável e não imposta por conveniência.

**Ex. Nível de estresse, desde “relaxado” até “estressado”**

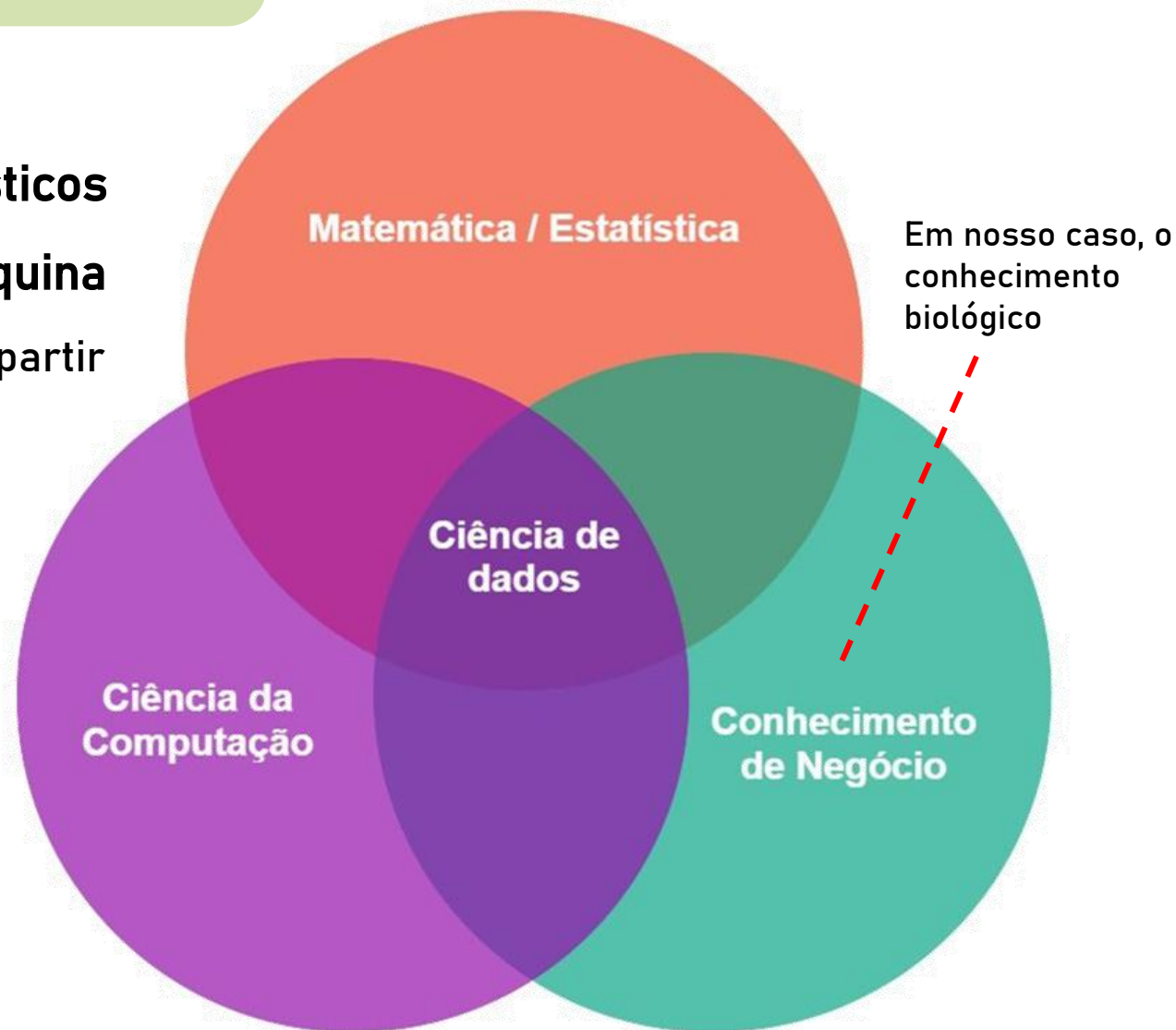


Dados  
Dados → informação

# Ciência de dados

É o campo que utiliza **métodos estatísticos** e **algoritmos de aprendizado de máquina** para **extrair conhecimento e insights** a partir de grandes conjuntos de dados (o tal do Big Data).

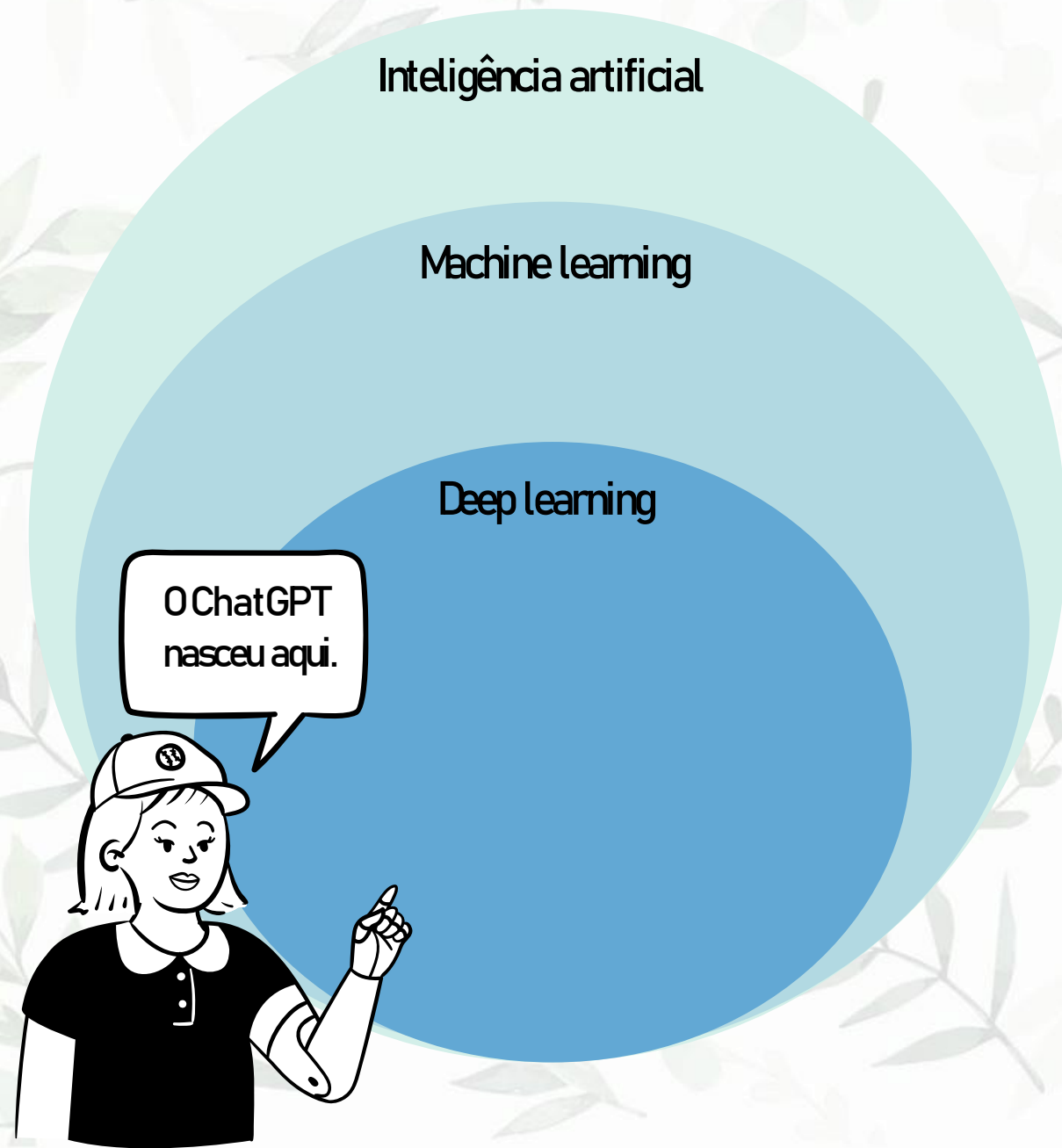
---



**Inteligência artificial:** Ciência dedicada a fazer as máquinas pensarem como humanos.

**Machine learning:** desenvolver algoritmos e técnicas que permitem que o computador aprenda automaticamente a partir dos dados, sem ser explicitamente programado para tarefas específicas.

**Deep learning:** Se baseia em redes neurais artificiais com várias camadas para aprender representações complexas dos dados. Essas redes neurais são projetadas para imitar o funcionamento do cérebro humano, onde cada camada sucessiva extrai características mais abstratas dos dados.



# Estatística

vs.

# Ciência de Dados

- **Inferência estatística** – fazer afirmações sobre uma população baseado em uma amostra representativa.
- Um número mais limitado de preditores.
- Dados amostrados sob rigoroso protocolo de amostragem
- **Conjunto de dados menores**, geralmente coletados de forma direta.
- Modelos paramétricos ou não-paramétricos.
- Foco na **capacidade de predição** e na **interpretabilidade** (a depender do objetivo).

- Busca **extrair insights** e conhecimentos dos dados para resolver problemas e tomar decisões informadas
- **Variabilidade fonte de dados e features** – incluindo dados estruturados e não-estruturados.
- **Big Data** > grande escala, dados de mundo real
- **Ampla e multidisciplinar**. Uma variedade de técnicas, ademais dos estatísticos, como mineração de dados, aprendizado de máquina, deep learning, etc.
- Foco na **capacidade preditiva**.
- Programação, visualização de dados??

# Estadística

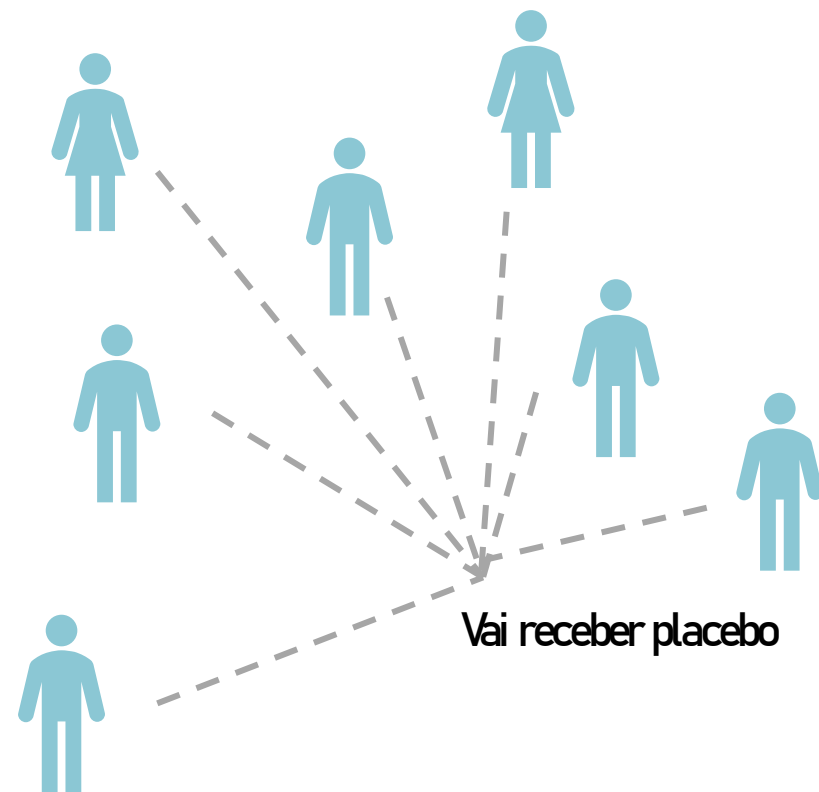
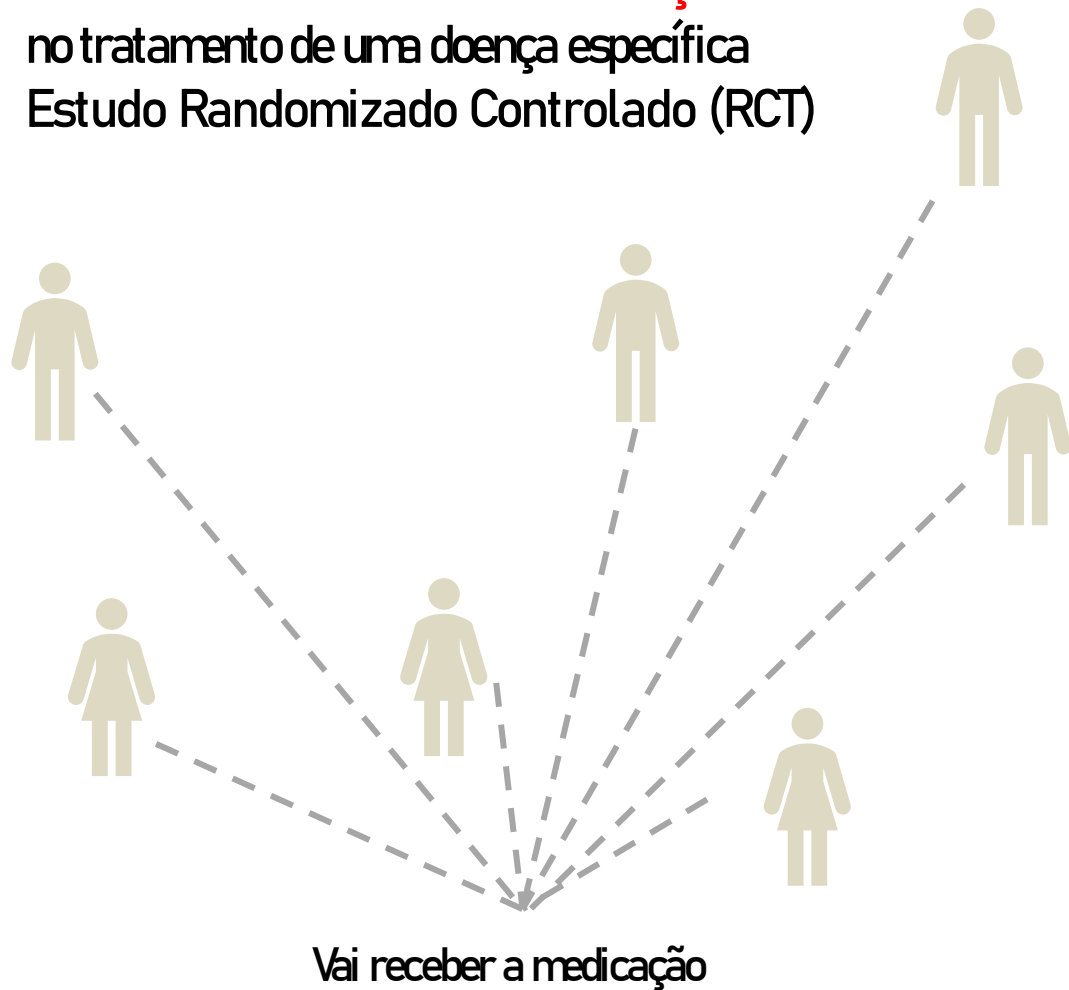
Avaliar o **efeito de uma nova medicação**  
no tratamento de uma doença específica  
Estudo Randomizado Controlado (RCT)



População de indivíduos que sofrem da doença

# Estatística

Avaliar o efeito de uma nova medicação  
no tratamento de uma doença específica  
Estudo Randomizado Controlado (RCT)

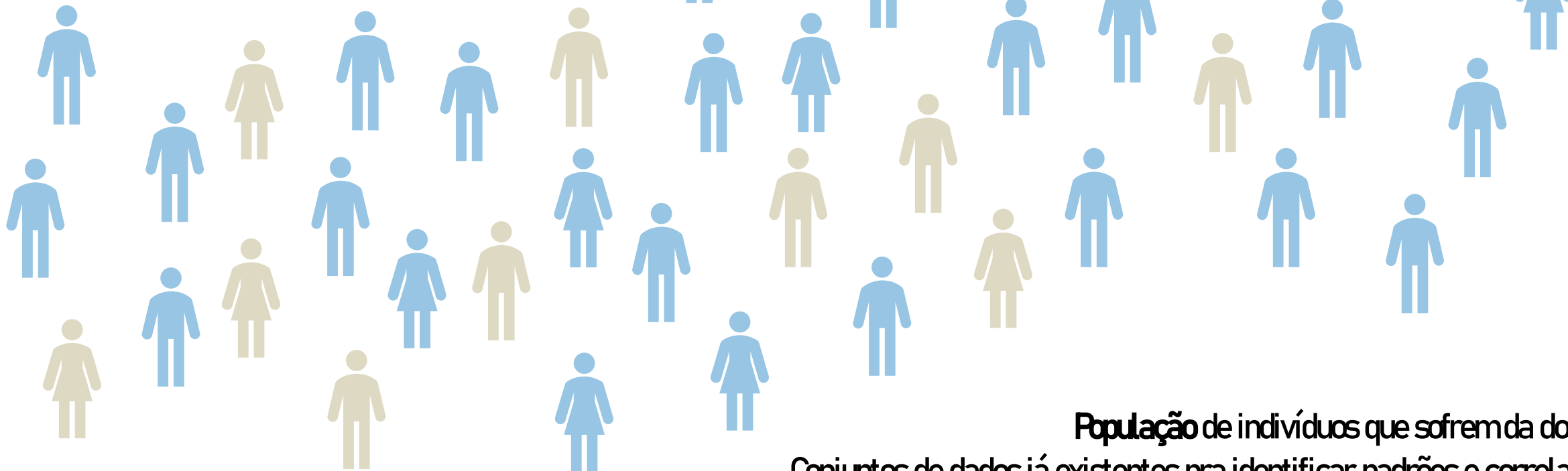


Amostra de indivíduos que sofrem da doença  
Vai ser acompanhado no tempo.

No final do estudo os resultados são analisados  
estatisticamente comparando grupo que recebe a  
medicação vs. Grupo que recebeu o placebo.

# Ciência de Dados

Avaliar o efeito **de uma nova medicação**  
no tratamento de uma doença específica  
Big Data: sem “controle” na amostragem do  
dado



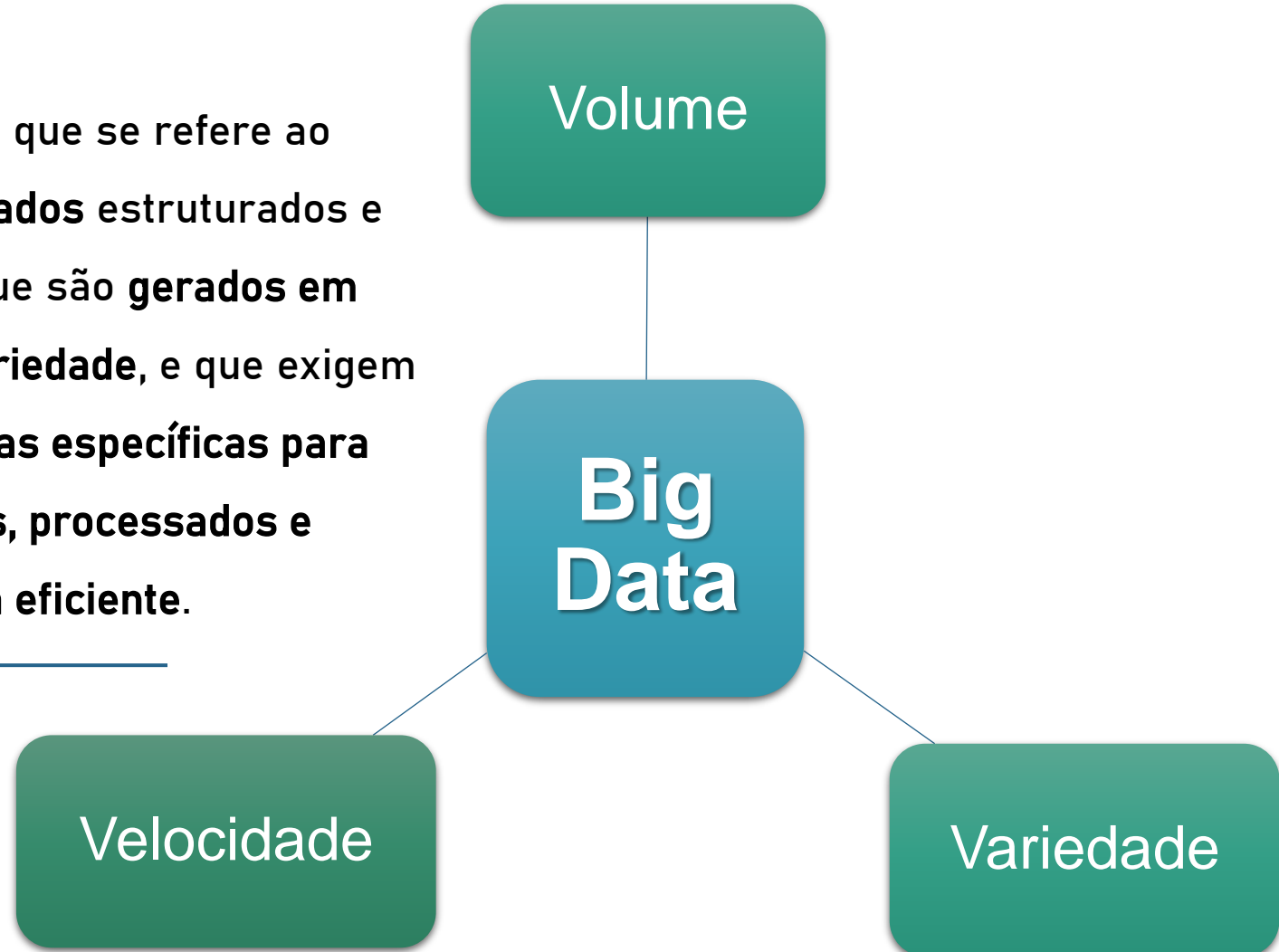
População de indivíduos que sofrem da doença  
Conjuntos de dados já existentes pra identificar padrões e correlações  
Dados do mundo real!!



# Big Data

**Big data** é um termo que se refere ao **grande volume de dados** estruturados e não estruturados, que são **gerados em alta velocidade e variedade**, e que exigem **técnicas e tecnologias específicas** para serem armazenados, processados e analisados de forma eficiente.

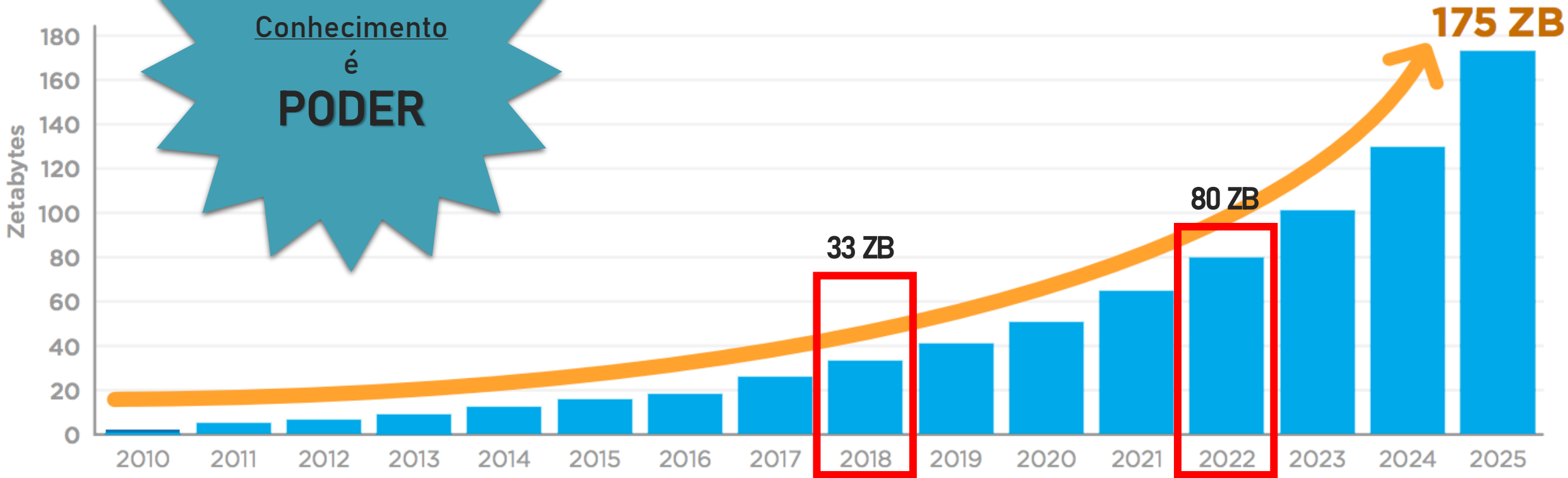
---



# Big Data Volume



Conhecimento  
é  
**PODER**



(Reinsel et al. 2018)

# Big Data Volume

2025 = 175zb

1 zettabyte  
=  
1 trilhão de  
gigabytes!



~5,5 trilhões de  
smartphones

(smartphone médio = 32gb)



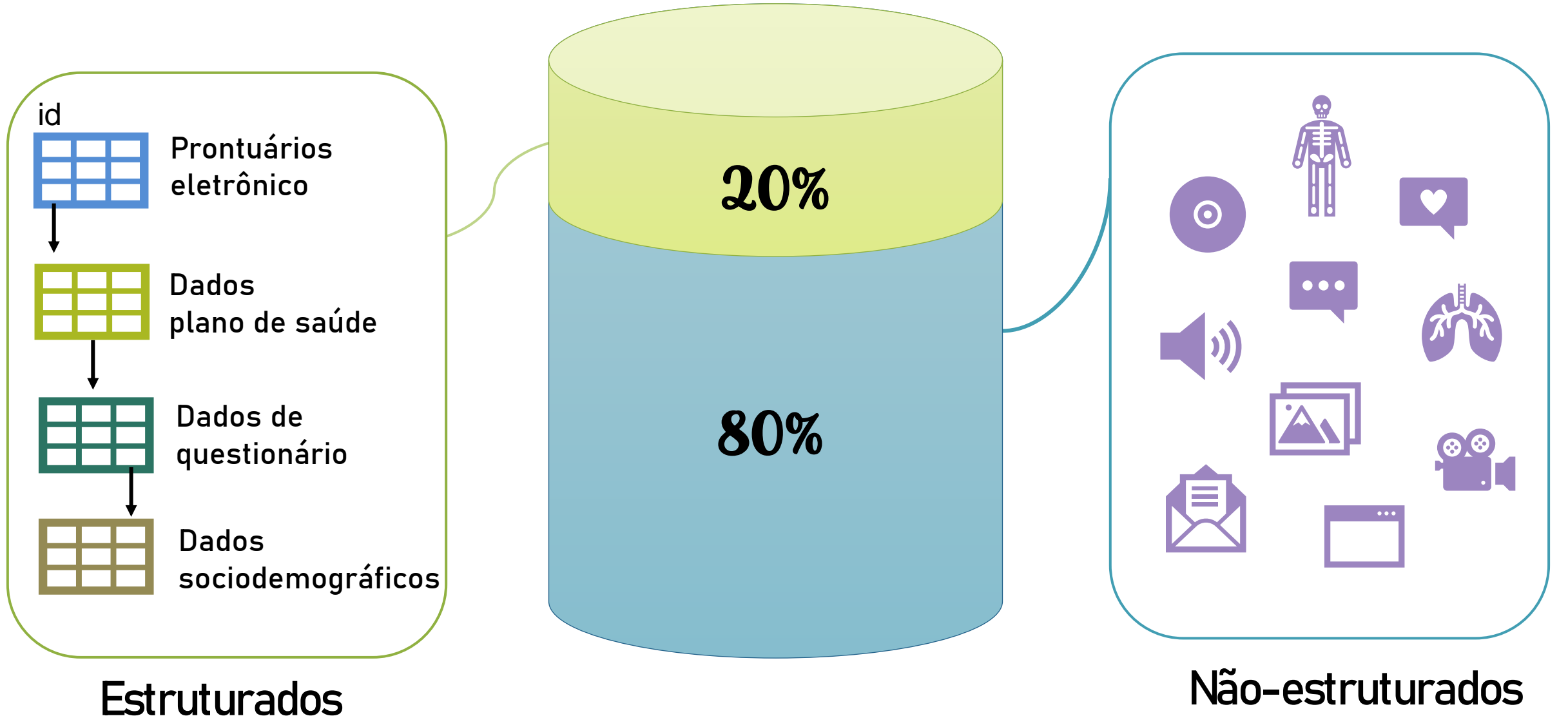
~700 smartphones/pessoa

# Big Data Velocidade

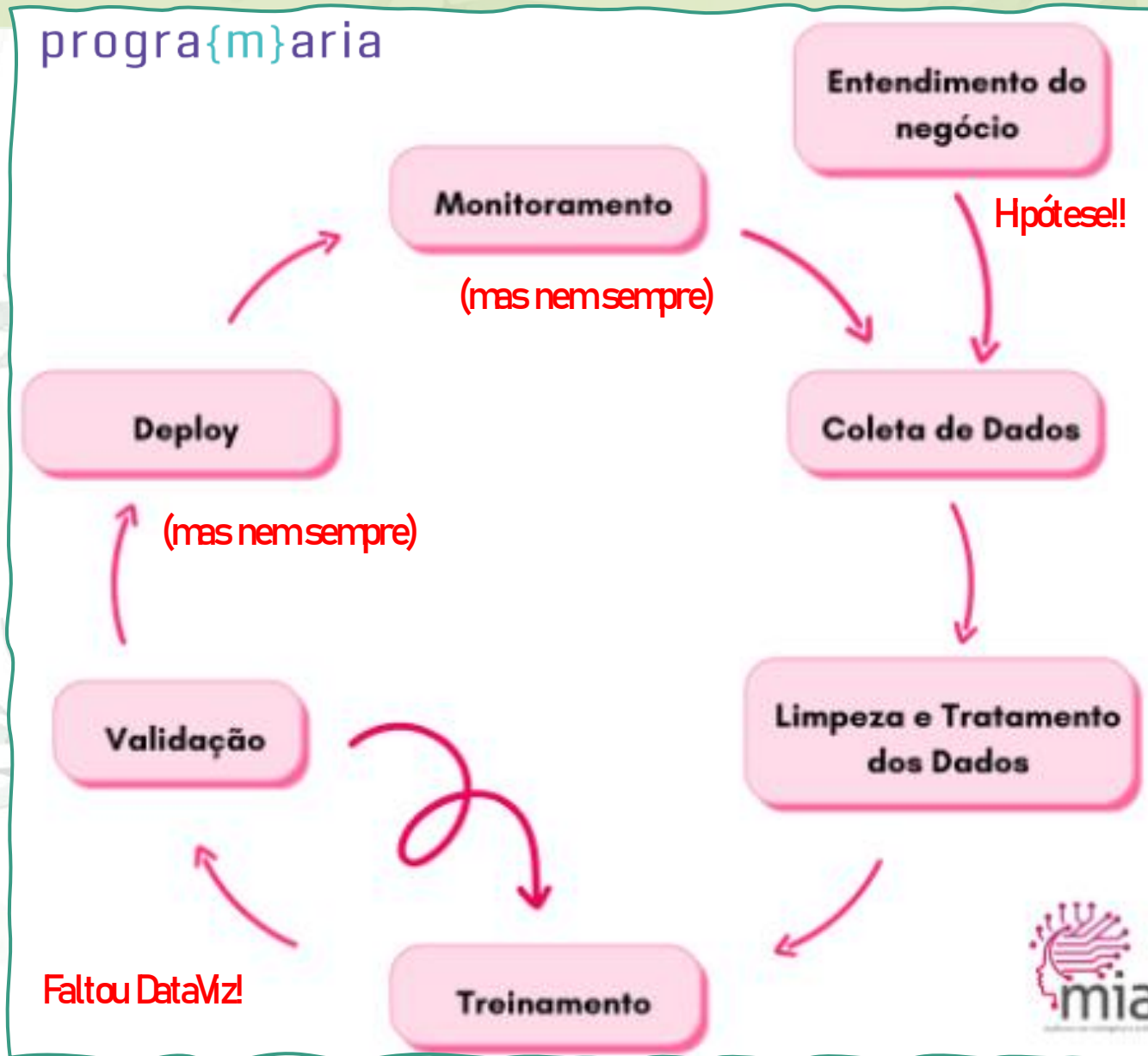


# Big Data Variedade

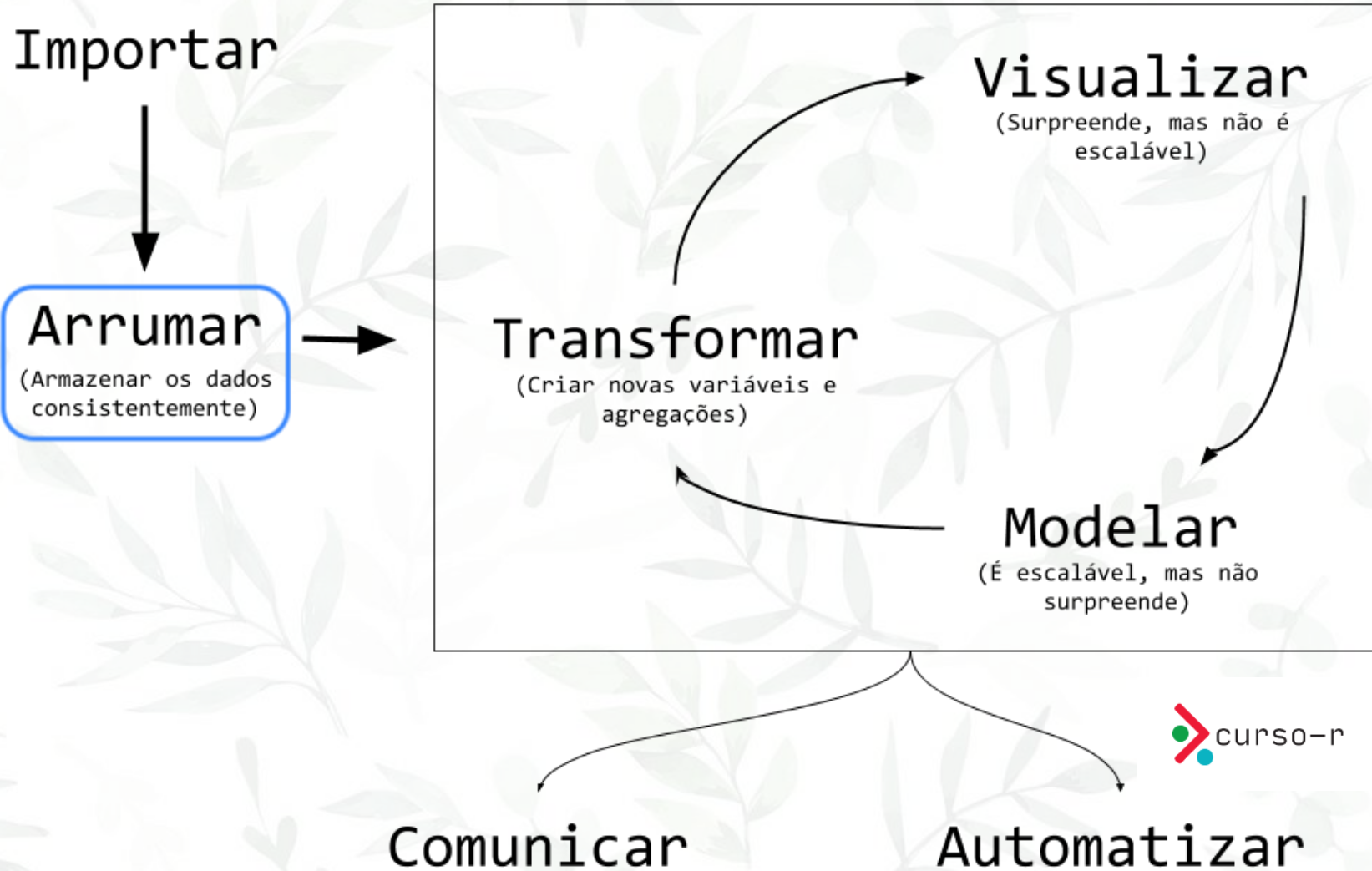
(Pierson 2015; Maissenhaelter et al. 2018)



# Ciclo de Vida da Ciência de Dados



# Ciclo de Vida da Ciência de Dados



Faltou a hipótese/entendimento do negócio

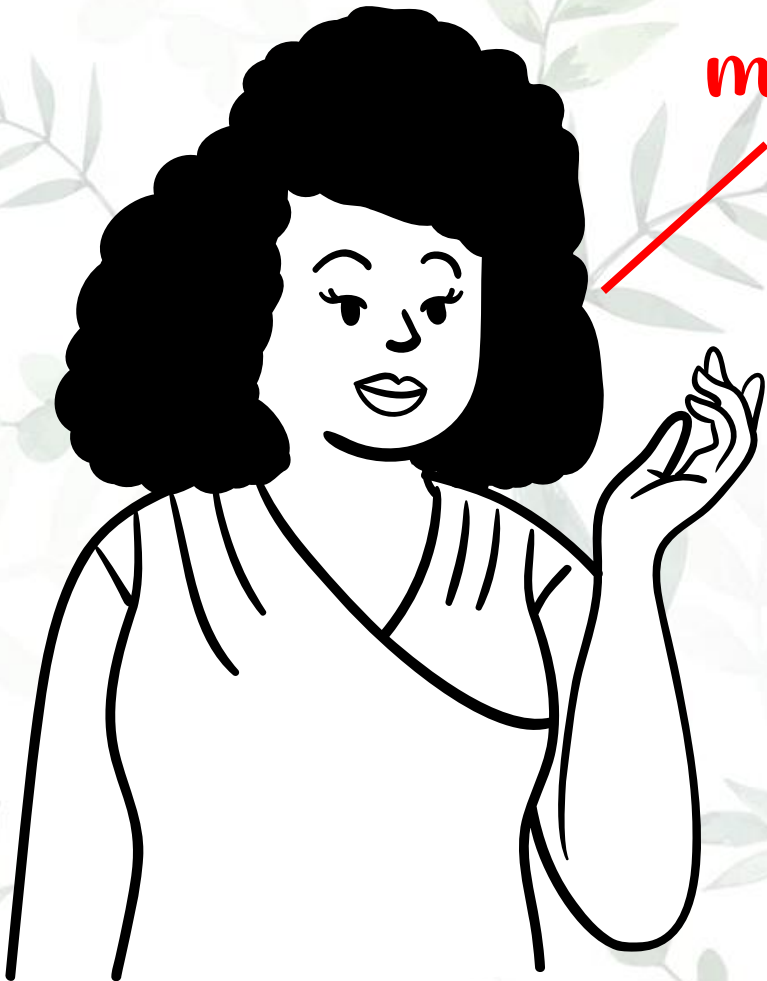
# Programação





# Linguagem de programação

É como “falar” o idioma das máquinas!



As linguagens de programação fornecem uma **sintaxe** e uma **semântica** que permitem aos programadores escreverem código compreensível para humanos e, ao mesmo tempo, **compreensível para a máquina**.

O código escrito em uma linguagem de programação é traduzido para instruções que a máquina pode entender e executar.

# Programação

## Linguagem de programação

## Código

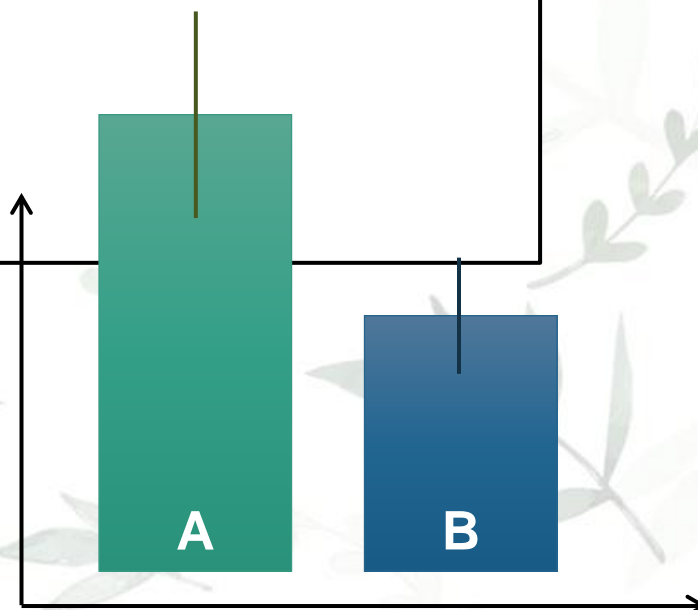
```
Importar(Planilha_de_dados) >
```

```
Selecionar_as_colunas("Medicamento", "Indice glicêmico") >
```

```
Agrupar_dados_por_coluna("Medicamento") >
```

```
Calcular(Média = média("Indice glicêmico")) >
```

```
Gráfico_de_barras(y = Média, x = Medicamento)
```



# Carreiras em dados

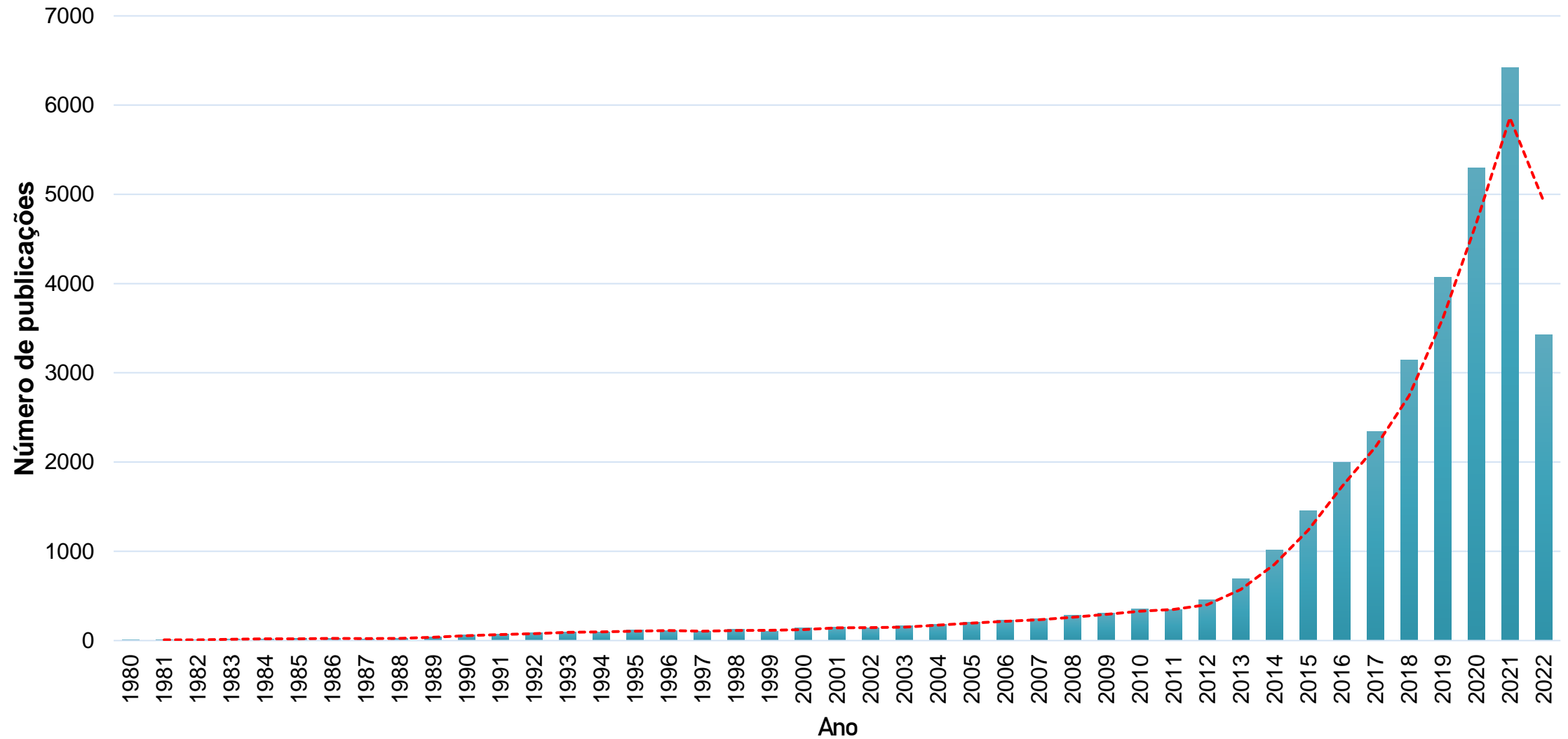
Cientista de dados	Engenheira/Engenheiro de dados	Analista de dados	Engenheira/Engenheiro de analytics
<p>Cria modelos e análises em cima dos dados.</p> <p>Possui conhecimento estatístico e de programação.</p> <p>É uma necessidade de grande empresas ou em projetos específicos.</p> <p><b>Python/R</b></p>	<p>É responsável pela manutenção da infraestrutura.</p> <p>Desenvolve código.</p> <p>Não tem conhecimento específico sobre o domínio.</p> <p>Preocupa-se com a <b>manutenção do código</b></p>	<p>É responsável por tirar insights dos dados (ex: Por que estamos perdendo mercado na região X?)</p> <p>Faz análises de negócio.</p> <p>Possui conhecimento de negócio.</p> <p><b>Excel/SQL/BI</b></p>	<p>Extrai e transforma os dados para análise.</p> <p>Desenvolve o Data Warehouse.</p> <p>Possui conhecimento de negócio e programação.</p> <p>Interage com os analistas e engenheiros de dados.</p> <p><b>SQL/DBT/BI</b></p>



Muito legal Marília, muito bacana...

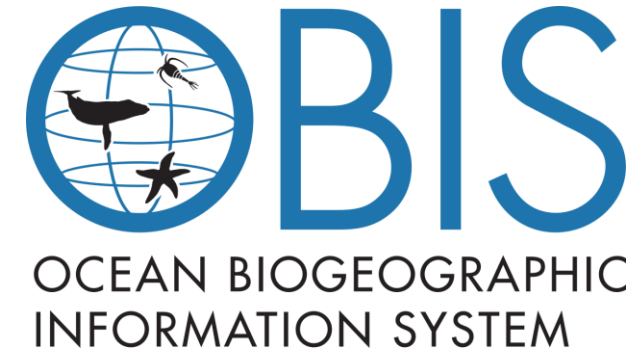
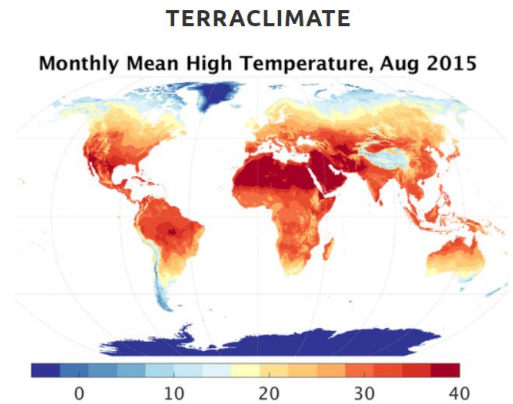
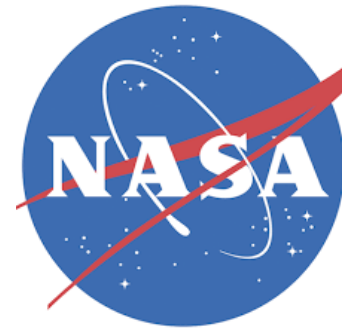
Mas eu sou biólogo@ o **que eu tenho que ver com isso?**

# Publicações associadas ao termo “Big Data” no PubMed (1980-2022)



# Big Data

Velocidade | Variedade | Volume



# iNaturalist



Data Papers | [Free Access](#)

**Atlantic butterflies: a data set of butterfly communities from the Atlantic Forest of South America**

Data Papers | [Free Access](#)

**ATLANTIC ANTS: a data set of ants in Atlantic Forests of South America**

Data Papers | [Free Access](#)

**ATLANTIC POLLINATION: a data set of pollination interactions with nectar-feeding vertebrates in the Atlantic Forest of South America**

Data Paper | [Free Access](#)

**ATLANTIC MAMMALS: a data set of assemblages of medium- and large-sized mammals of the Atlantic Forest of South America**

Joice Iamara-Nogueira ✉, Natália  
Ana Maria Rui, Andréa C. Araujo

First published: 22 November 2023

Corresponding Editor: William

Data Papers | [Free Access](#)

**ATLANTIC BIRD TRAITS: a data set of bird morphological traits from the Atlantic forests of South America**

authors ▾



## Brazil burned an area equivalent to Colombia and Chile combined between 1985 and 2022

More than 185 million hectares were consumed by fire. Each year, the area burned in Brazil is equivalent to that of Suriname

[Read more](#)



## Satellite images reveal the 5 municipalities that deforest the most in the 9 Caatinga states

Deforestation in the biome advanced 70% from 2020 to 2021

[Read here](#)

## Brazil gains 1.7 million hectares of water by 2022, but continues to dry up

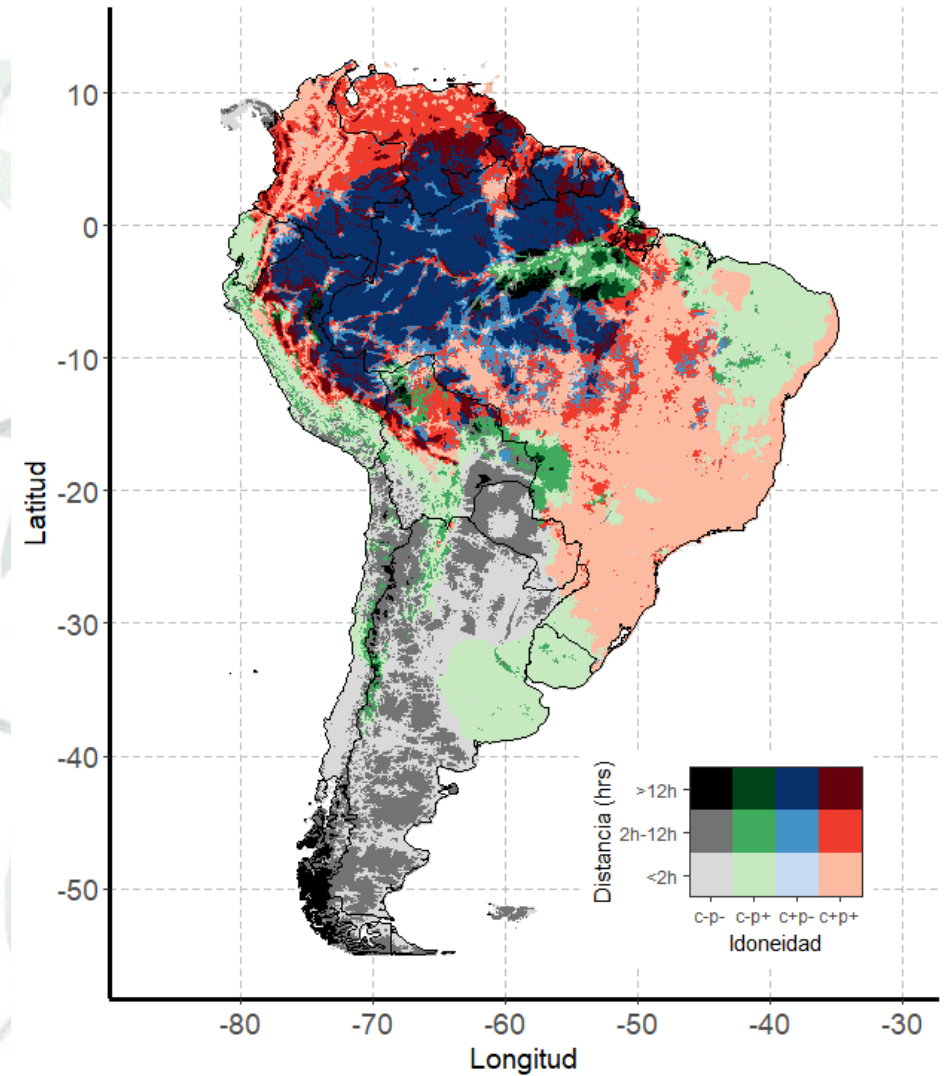
Pantanal continues as the biome with the largest reduction in water surface area





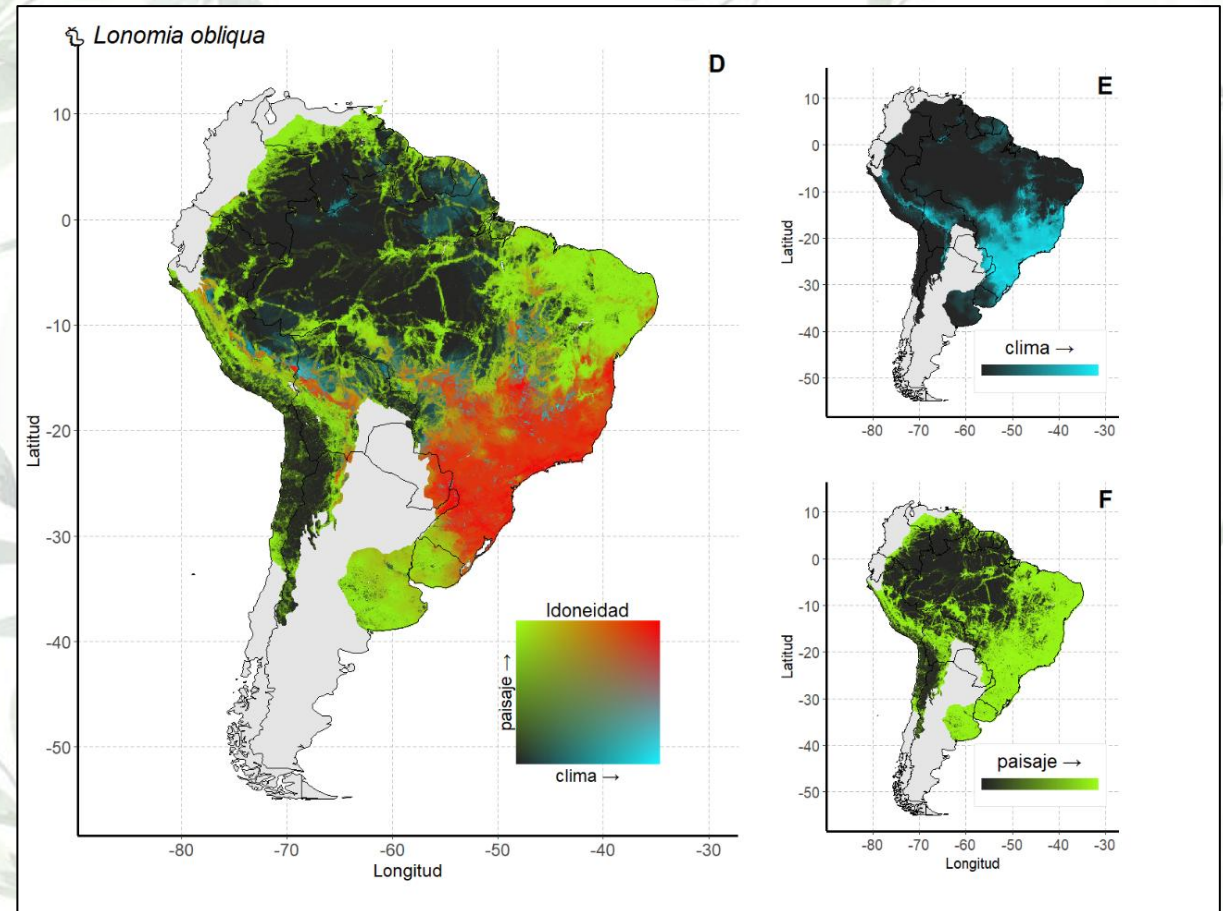
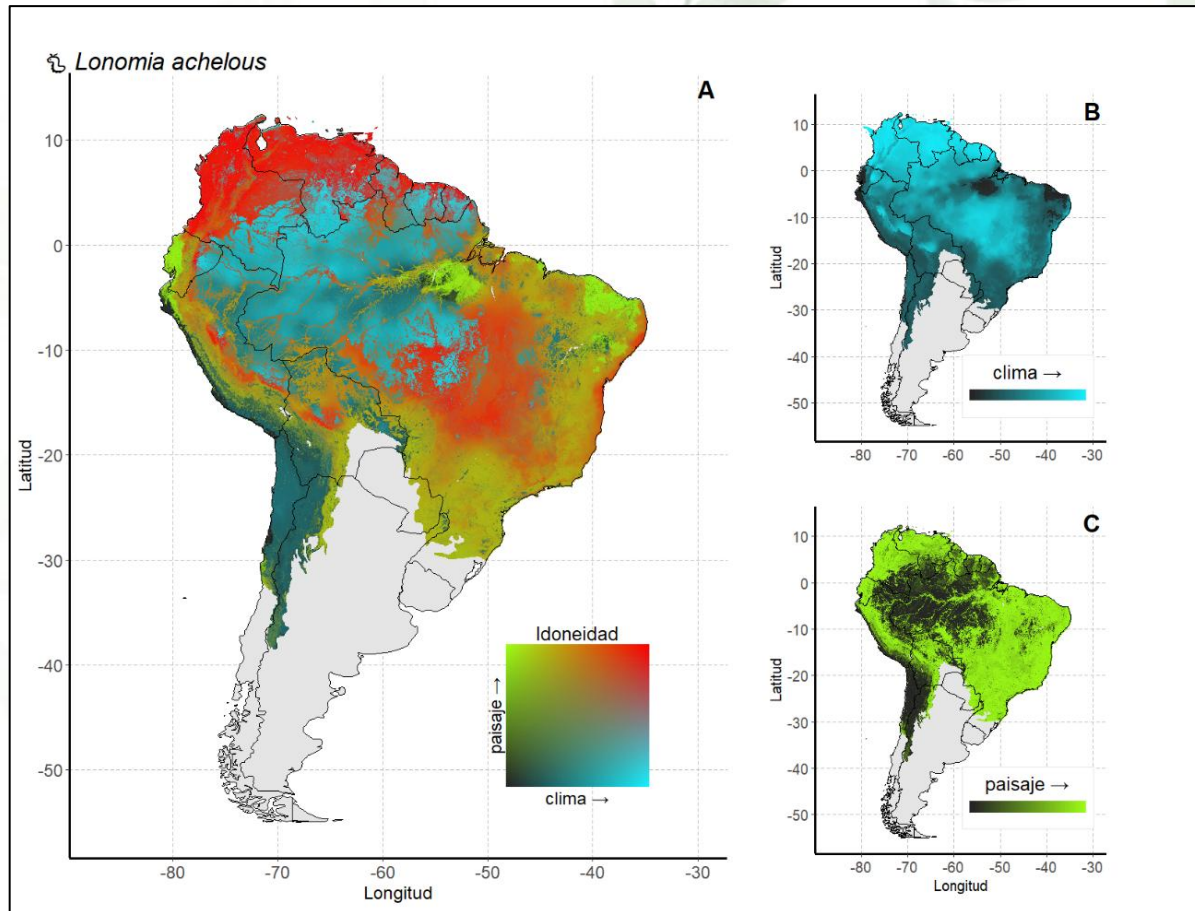
# ¡Cuidado donde tocas!

## Un mapa de riesgo para el lononismo en Sudamérica.



# ¡Cuidado donde tocas!

## Un mapa de riesgo para el Lononismo en Sudamérica.



# OpenAI offers \$100,000 grants for ideas on AI governance

By Greg Bensinger ▾

May 25, 2023 6:20 PM GMT-3 · Updated a day ago



ARTIFICIAL  
INTELLIGENCE



# ChatGPT

Global Grand Challenges BILL & MELINDA GATES Foundation

ABOUT PARTNERSHIPS **CHALLENGES** AWARDED GRANTS GRANT OPPORTUNITIES NEWS Q

### Catalyzing Equitable Artificial Intelligence (AI) Use

SHARE THIS [Twitter](#) [Facebook](#) [LinkedIn](#) [Email](#)

**APPLY FOR THIS OPPORTUNITY →**


**INITIATIVE**  
Grand Challenges

**DATE OPEN**  
May 22, 2023, 4:00 am PDT

**DEADLINE**  
Jun 05, 2023, 11:30 am PDT

**SUPPORTING MATERIALS**

- [Request for Proposal](#)
- [Rules and Guidelines](#)
- [Application Instructions](#)
- [Addendum to the RFP](#)
- [Budget Template](#)
- [Frequently Asked Questions \(FAQs\)](#)



Background

The background of the entire image is a repeating pattern of watercolor-style green leaves and branches. The leaves are in various shades of green, from light and pale to dark and vibrant. The branches are thin and dark green. The overall style is soft and naturalistic.

# Videos

Obrigada



[marilia.melo.favalesso@gmail.com](mailto:marilia.melo.favalesso@gmail.com)

[www.mmfava.com](http://www.mmfava.com)