

Biologia e Ciência de Dados

O que eu tenho haver?

Marília Melo Favalesso



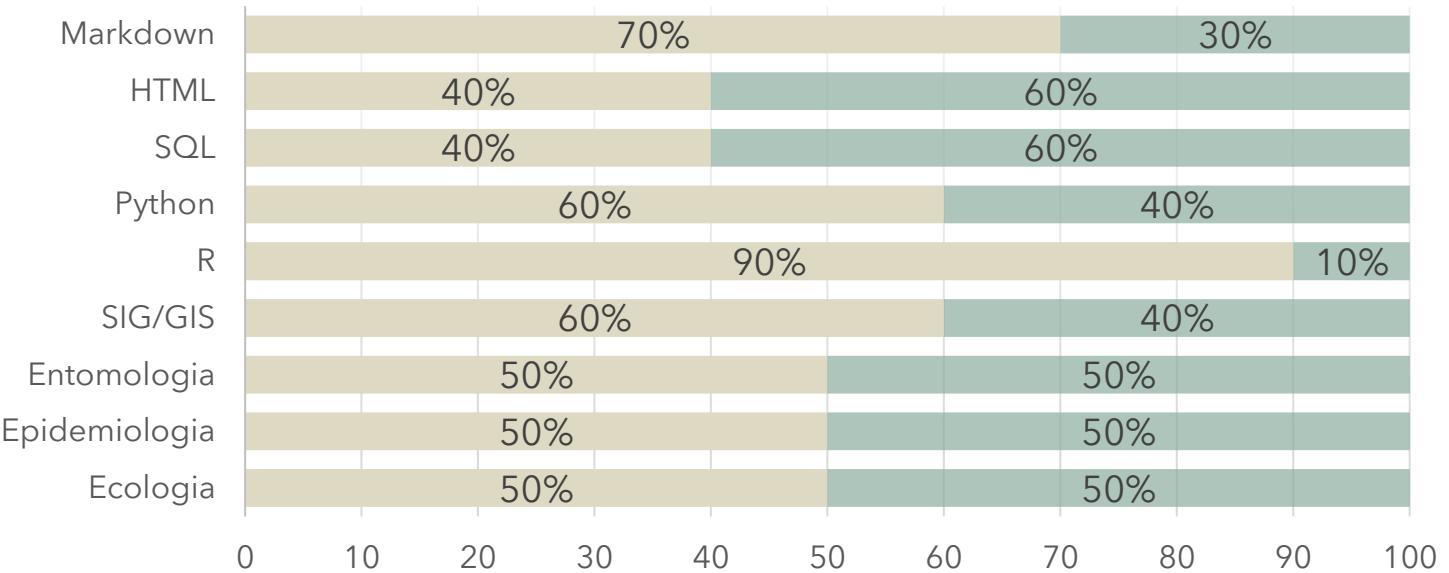
Marília Melo Favalesso



www.mmfava.com

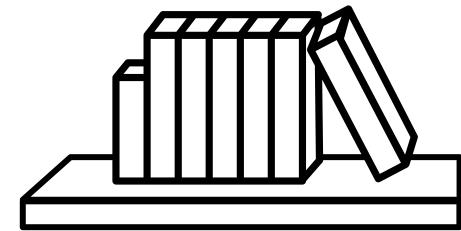
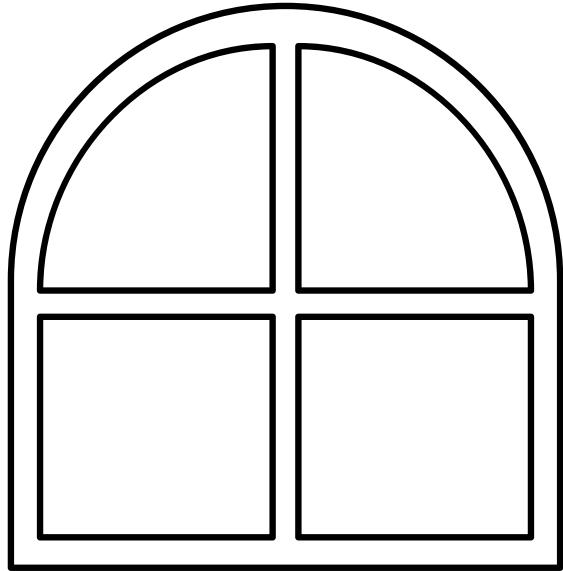
- ✓ Técnica Ambiental (CEEP – Cascavel - PR)
- ✓ Bióloga (UFPR – Palotina - PR)
- ✓ Consultoria e assessoria em Bioestatística (Cascavel - PR)
- ✓ Mestre em Ciências Ambientais (UNIOESTE – Cascavel - PR)
- Doutoranda em Ecoepidemiologia (UBA - Buenos Aires - AR)
- ✓ Cientista de dados (Hospital Israelita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

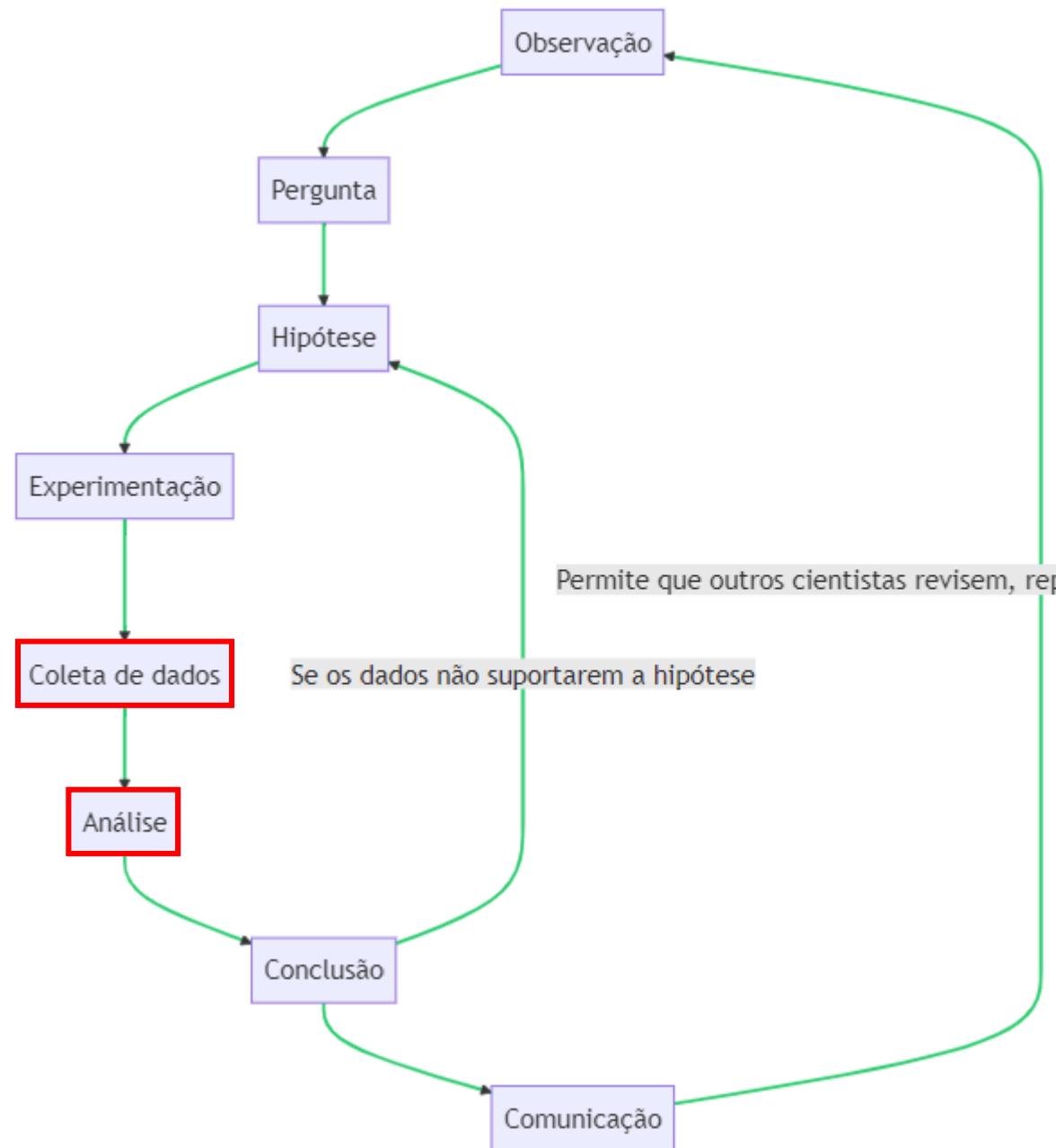
Skills





O que você quer “ser”
quando se formar?

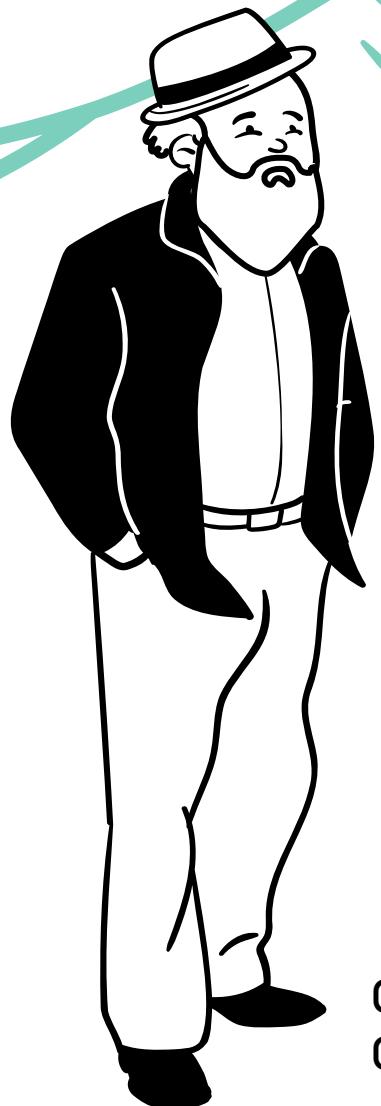




Independentemente do tipo de biólogo que você vai ser,
você com certeza terá que...

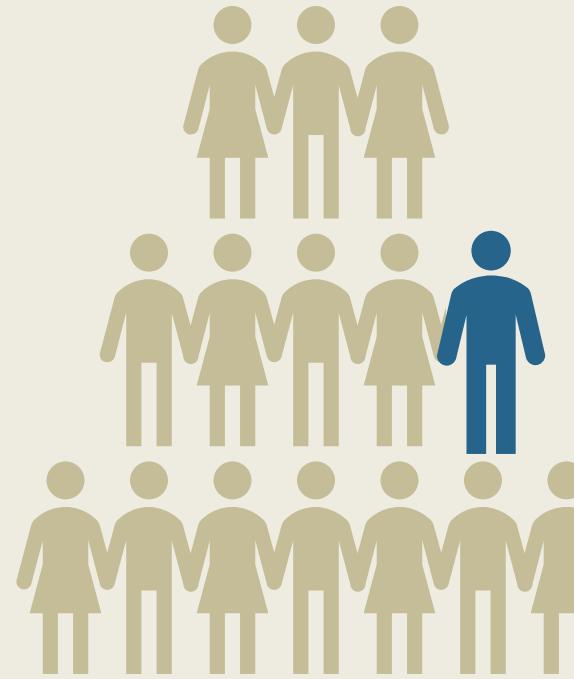
Coletar e/ou Analizar dados



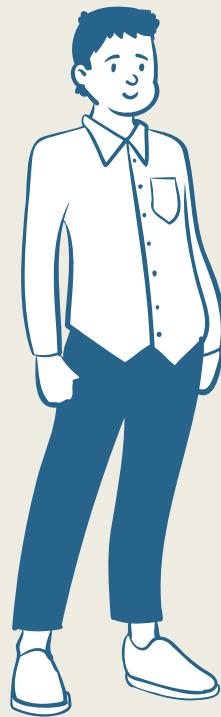


Charles Darwin
O "paizão" da biologia

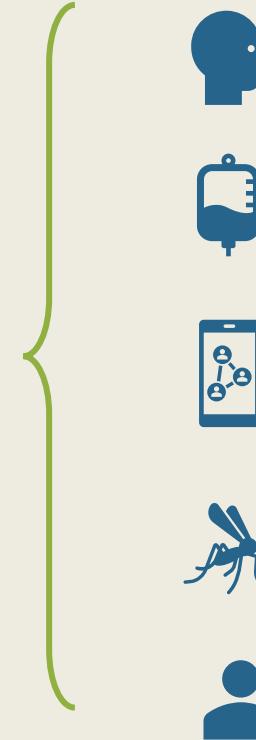
Mas o que são dados?



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)

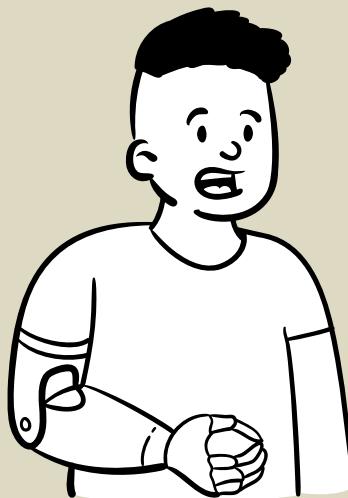
Dados são Variáveis

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

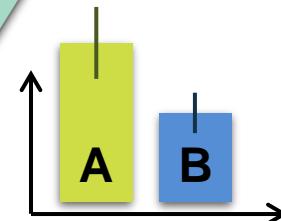
Fatos brutos e desorganizados
[A B A B B...] [126, 96, 120, 90, 98,...]



Dados quantitativos
e qualitativos

Informação

Uso e contexto dos dados
“Quem? O que? Onde? Quando?”



A análise dos dados sugere que
diabéticos em uso do fármaco A
apresentam índice glicêmico mais
elevado em comparação aos que usam o
fármaco B ($p < 0,05$)

Conhecimento

Informação aplicada
Responder o “por quê”

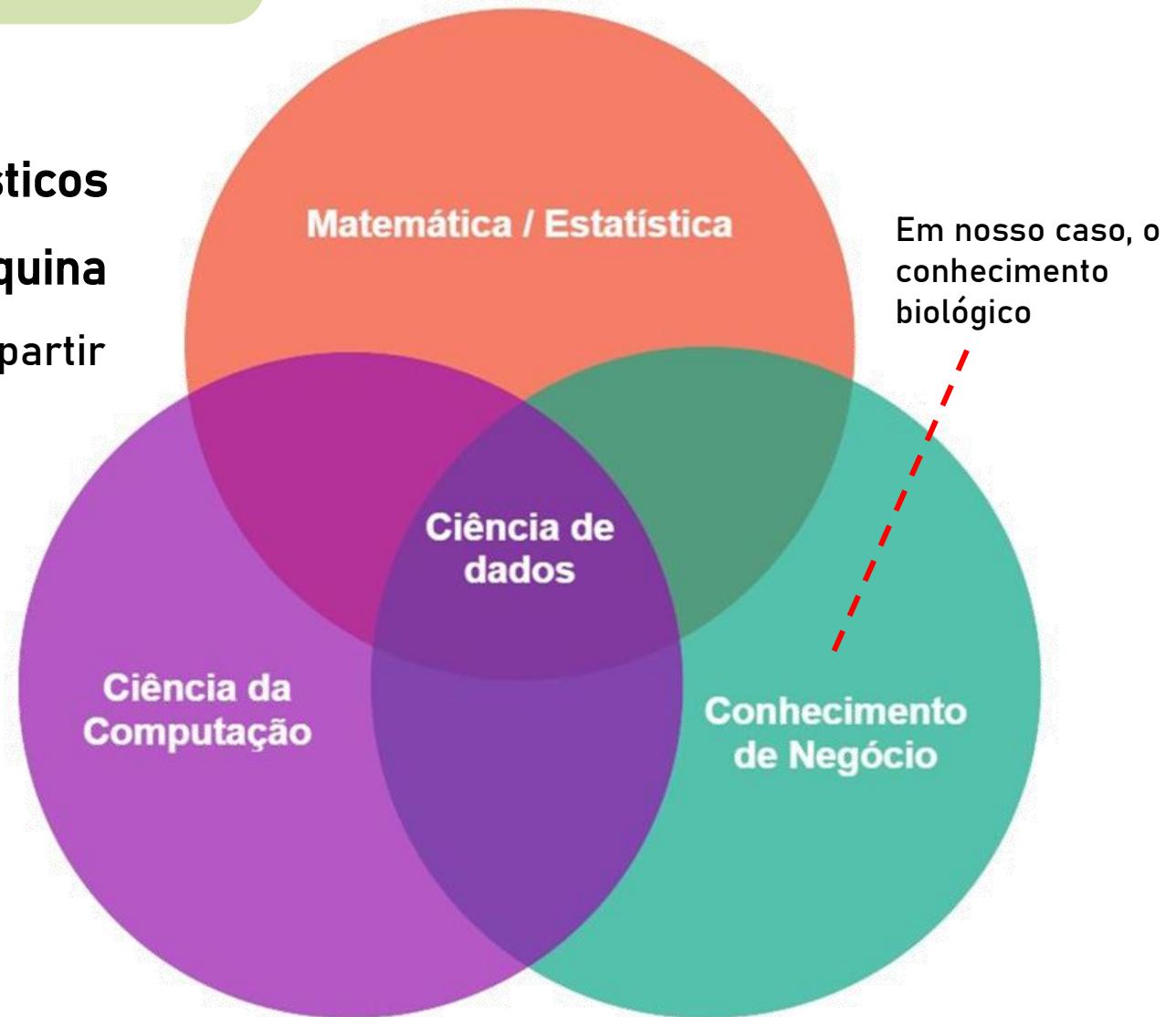


Recomenda-se o uso do fármaco B para
controle da diabetes, que possui ação
sobre...

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.

Inteligência artificial

Machine learning

Deep learning

O ChatGPT
nasceu aqui.



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 previsão** e na **interpretabilidade** (a depender do objetivo).

- Busca **extrair insights** e conhecimentos dos dados para resolver problemas e tomar decisões informadas
- **Variedade 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??

Estatística

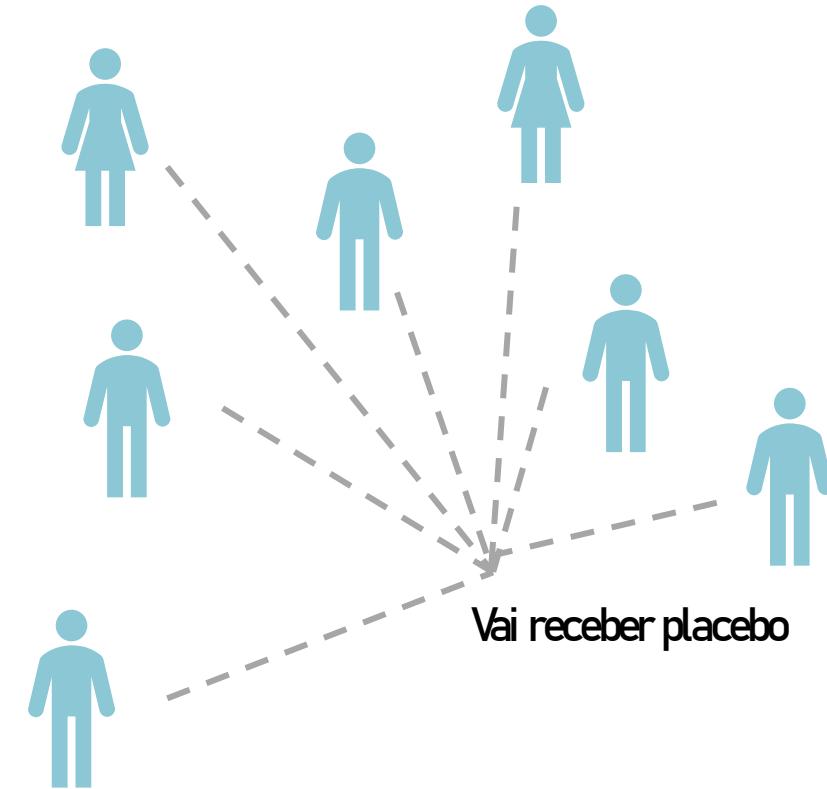
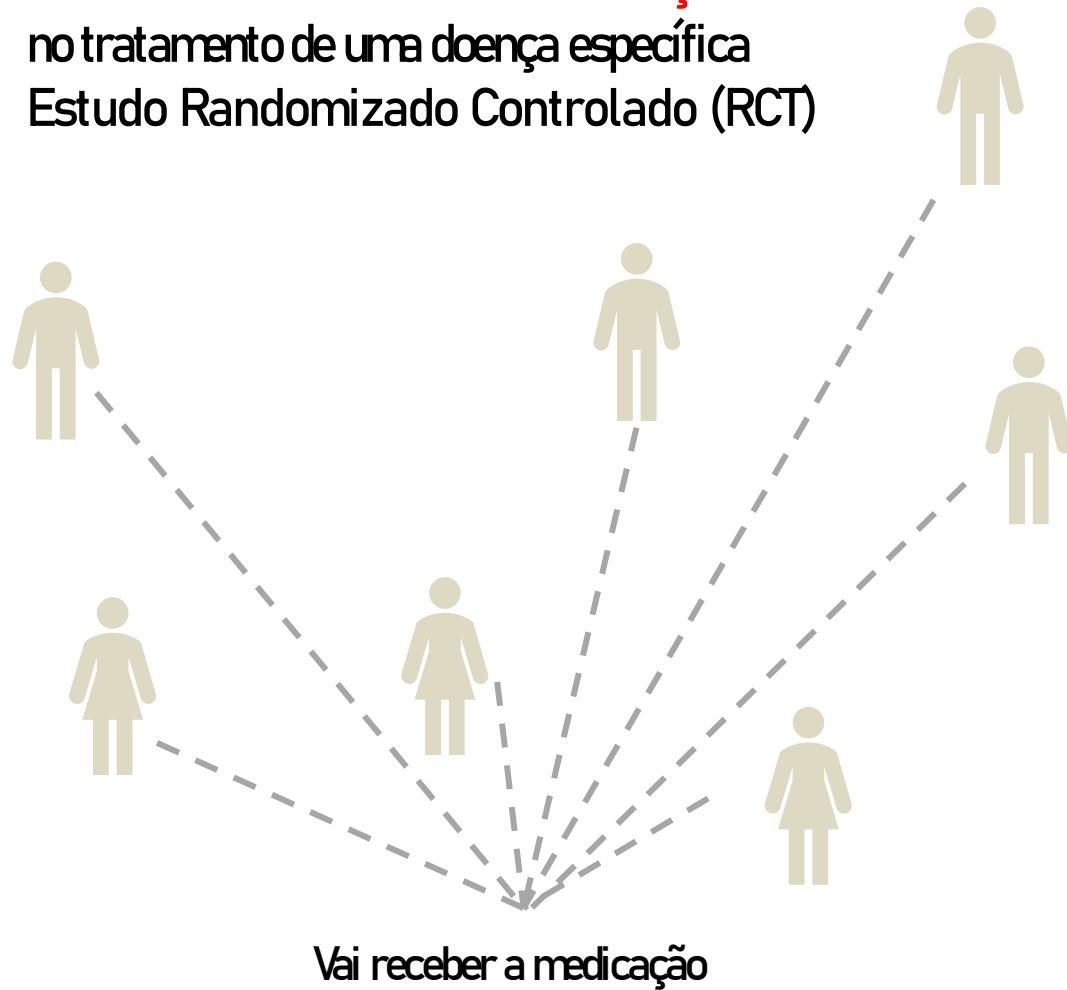
Avaliar o **efeito de uma nova medicação**
no tratamento de uma doença específica
Estudo Randomizado Controlado (RCT)



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

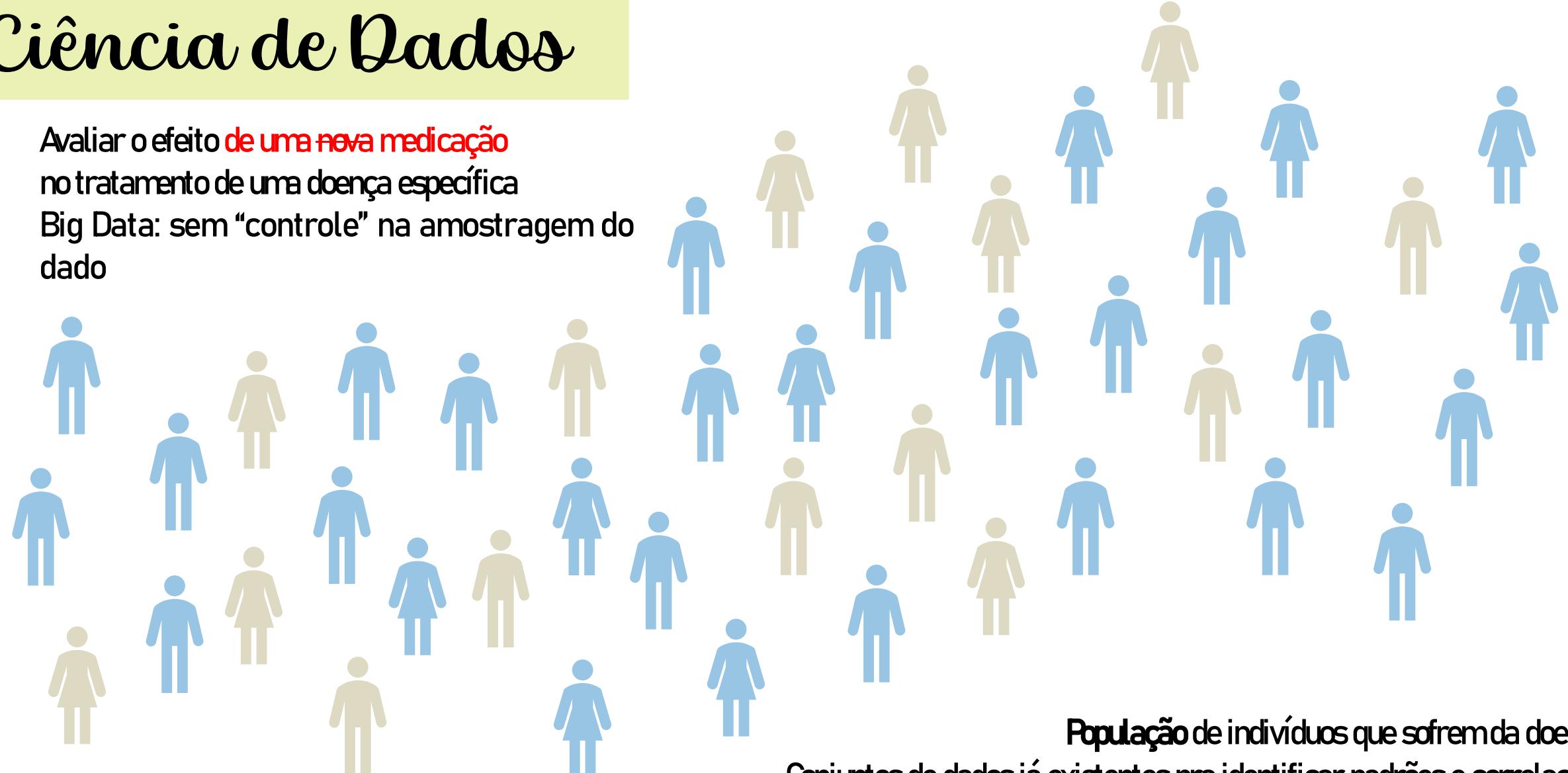
Nº final do estudo os resultados são analisados
estatisticamente comparando grupo que recebe a
medicação vs. Grupo que recebeu o placebo.

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.

Volume

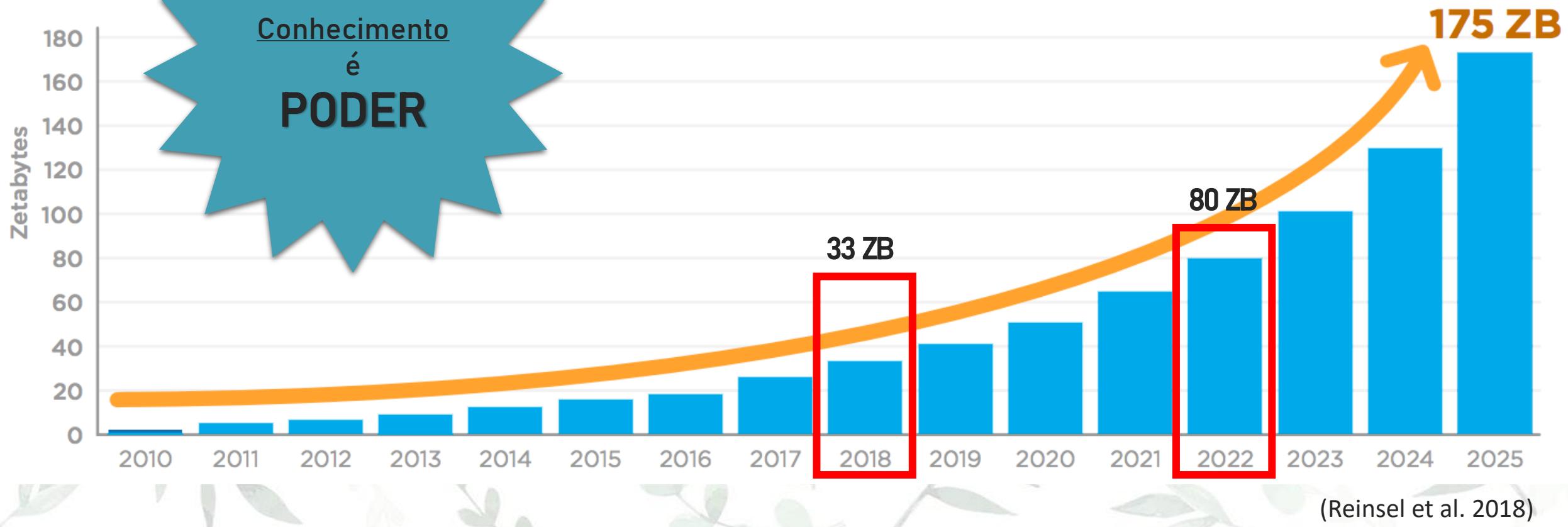
Big
Data

Velocidade

Variedade

Big Data Volume

Conhecimento
é
PODER



Big Data Volume

2025 = 175zb

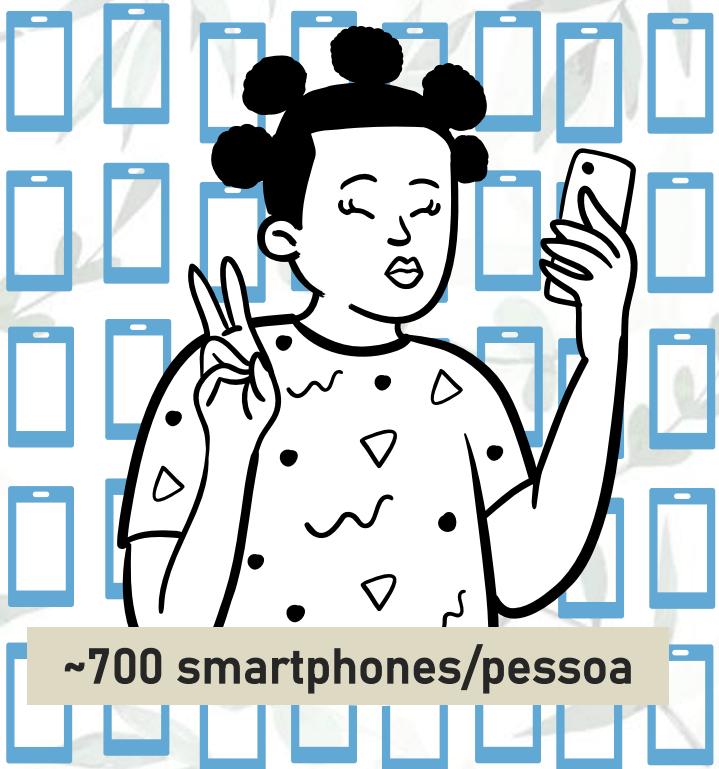
1 zettabyte
=
1 trilhão de gigabytes!

(Reinsel et al. 2018; Roser et al. 2019)



~5,5 trilhões de smartphones

(smartphone médio = 32gb)



~700 smartphones/pessoa

Big Data
Velocidade



102.911 buscas



1s



3.153.985 enviados

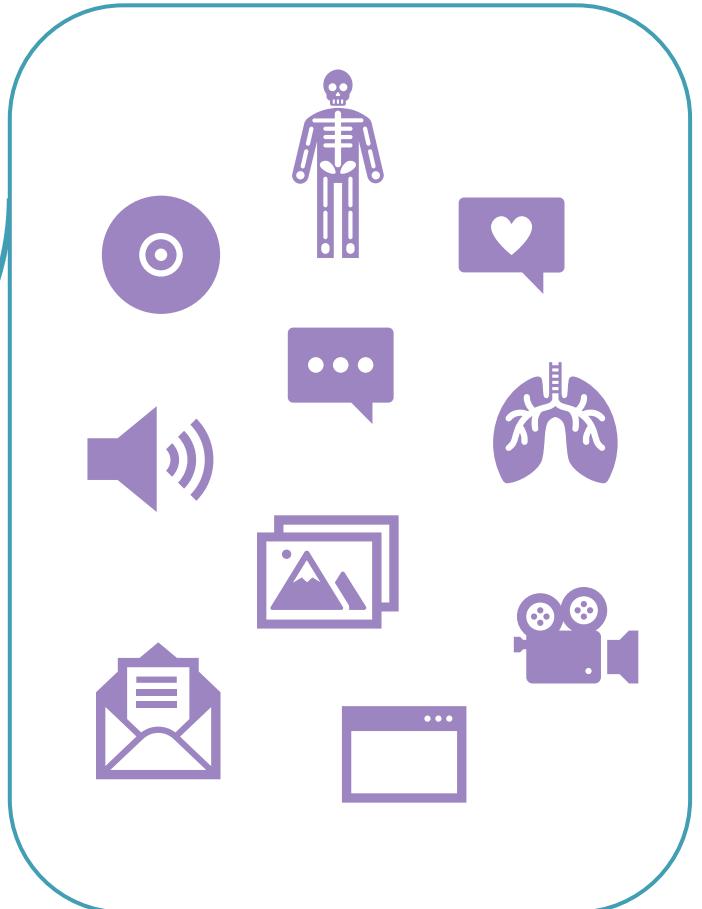
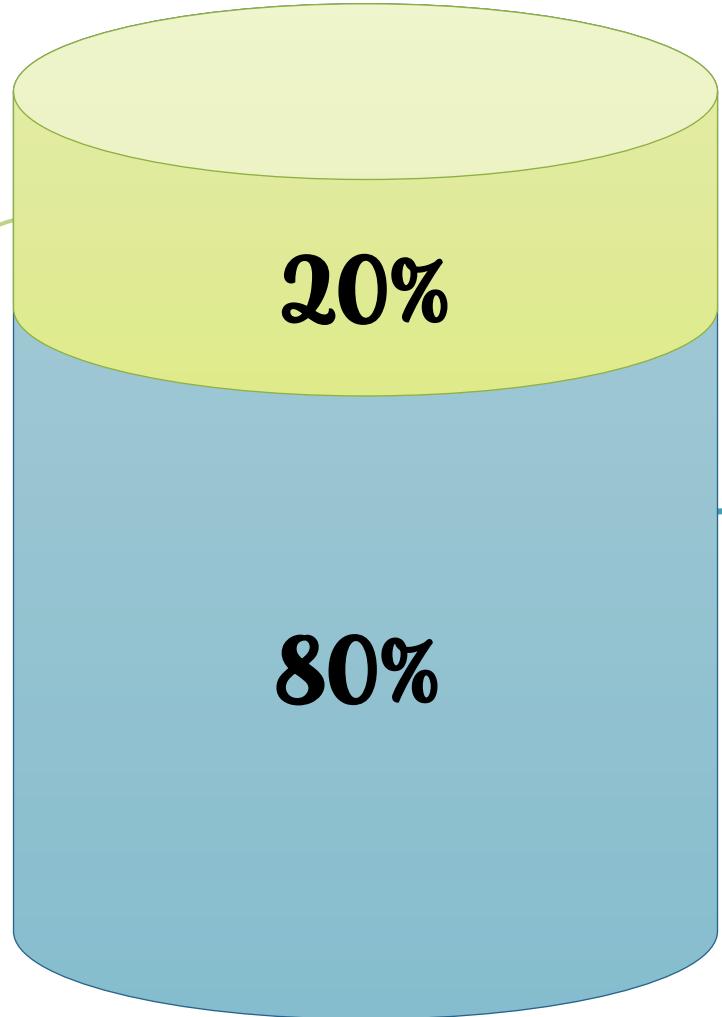
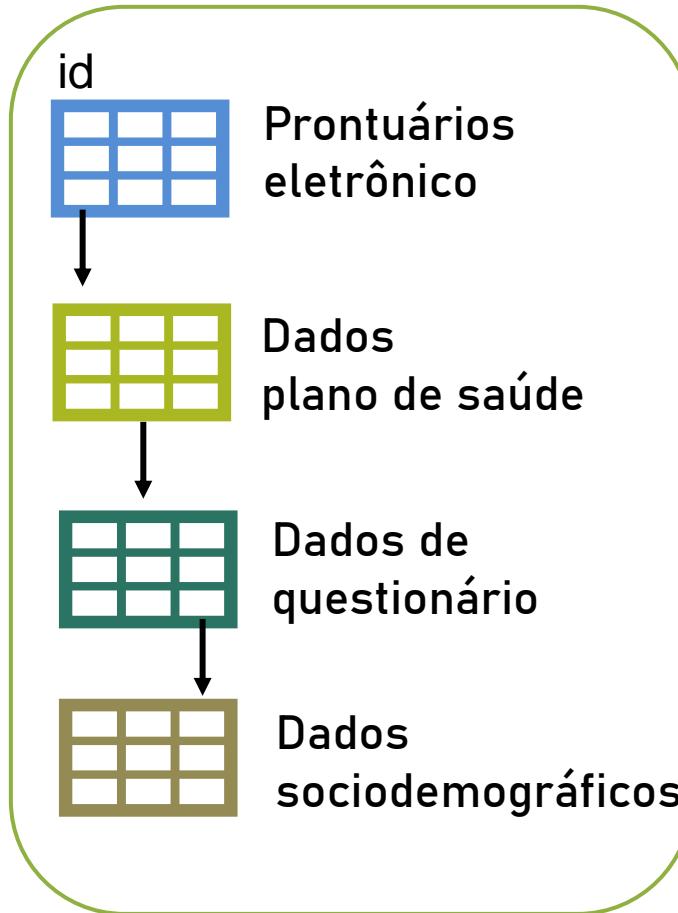


95.961 views

Internet live stats (07/06/2022)

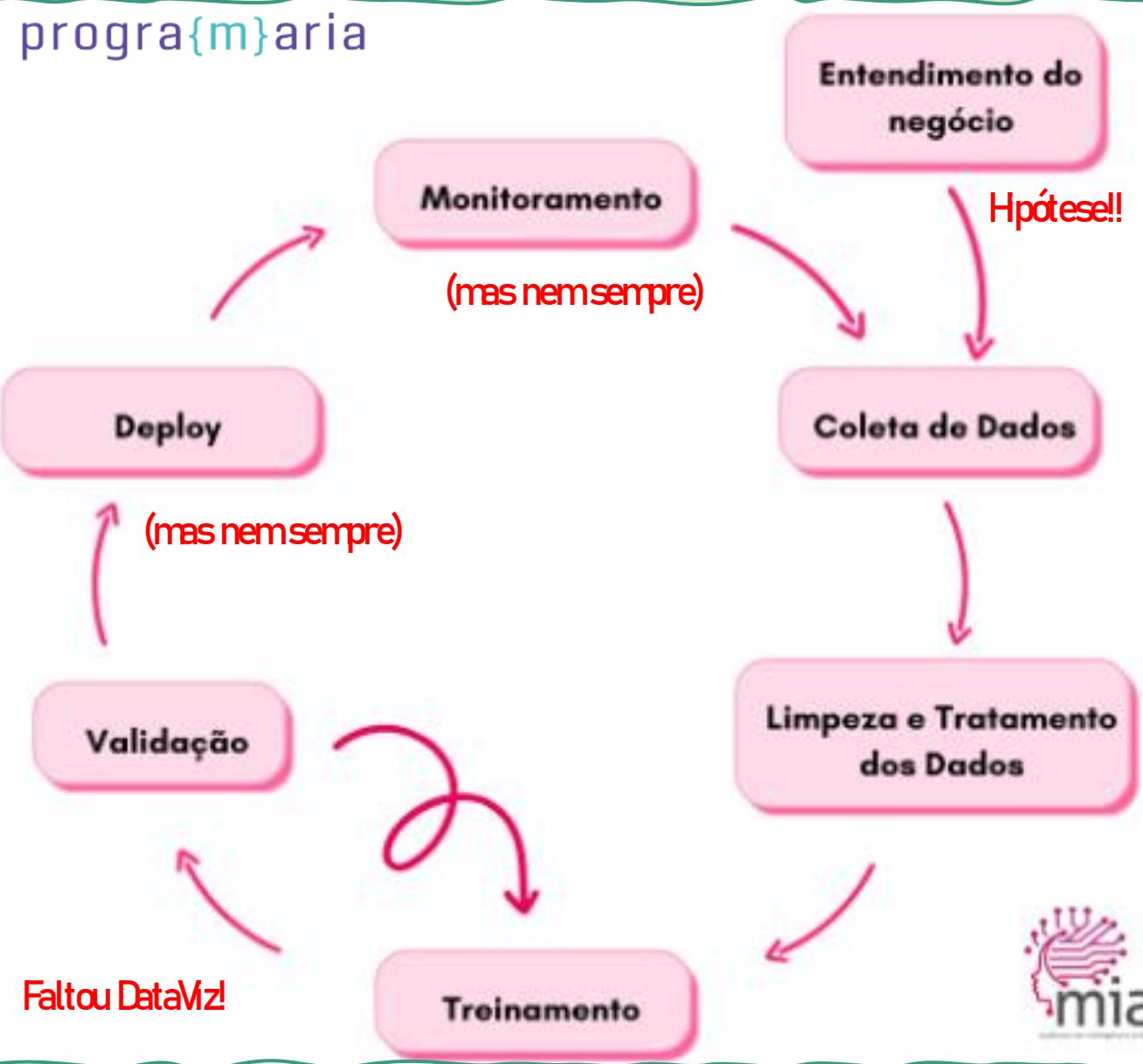
Big Data

Variedade



Ciclo de Vida da Ciência de Dados

progra{m}aria



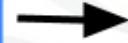
Ciclo de Vida da Ciência de Dados

Importar



Arrumar

(Armazenar os dados
consistentemente)



Transformar

(Criar novas variáveis e
agregações)

Visualizar

(Surpreende, mas não é
escalável)

Modelar

(É escalável, mas não
surpreende)

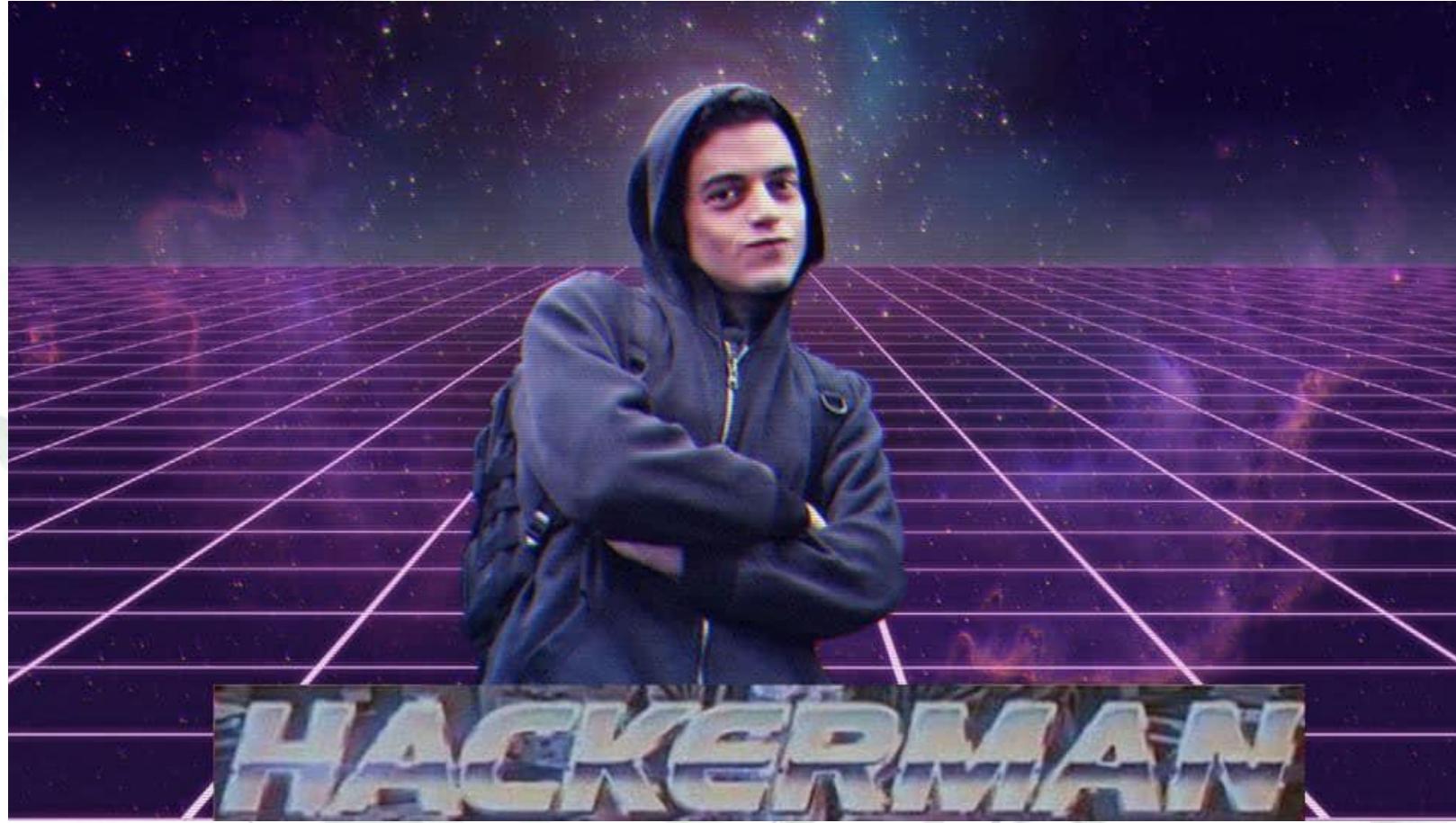
Comunicar

Automatizar



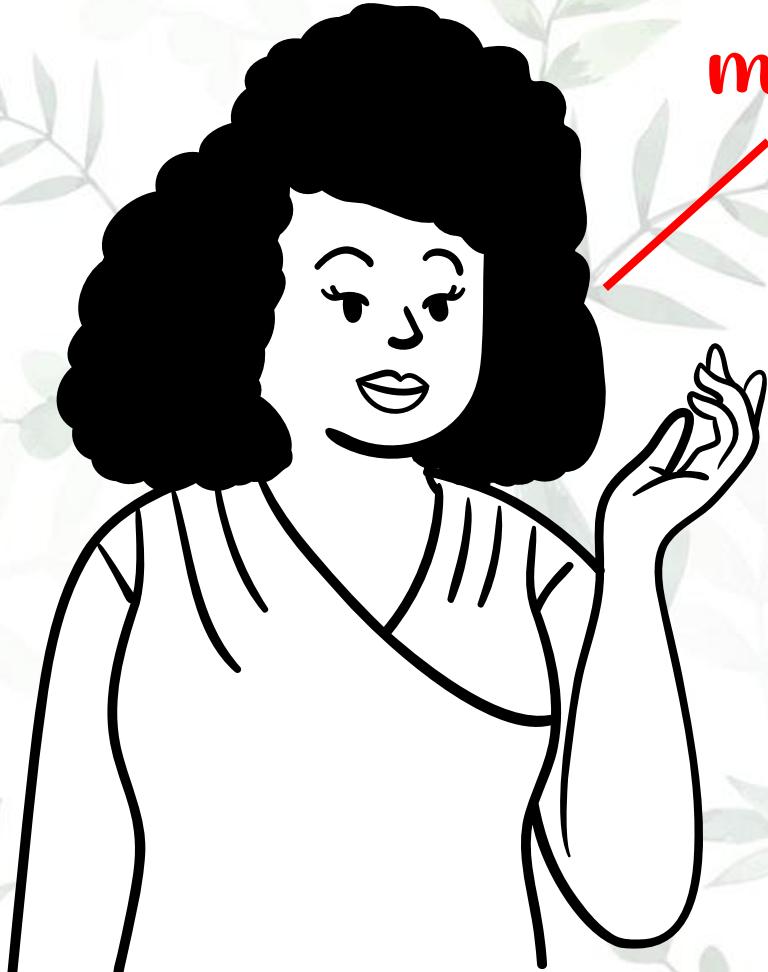
Faltou a hipótese/entendimento do negócio

Programação



Linguagem de programação

É como “falar” o idioma das máquinas!



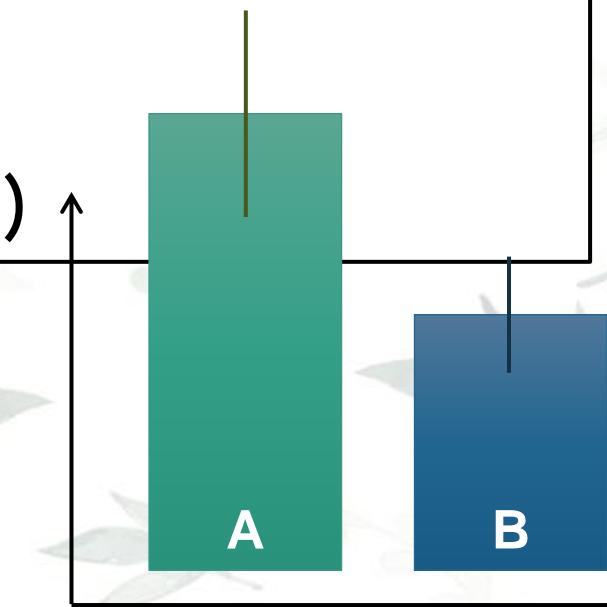
As linguagens de programação fornecem uma sintaxe e uma semântica que permitem aos programadores escreverem código compreensível para humanos e, ao mesmo tempo, compreensível para a máquina.

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", "Índice glicêmico") >  
Agrupar_dados_por_coluna("Medicamento") >  
Calcular(Média = média("Índice glicêmico")) >  
Gráfico_de_barras(y = Média, x = Medicamento)
```



Carreiras em dados

Cientista de dados

Cria modelos e análises em cima dos dados.

Possui conhecimento estatístico e de programação.

É uma necessidade de grande empresas ou em projetos específicos.

Python/R

Engenheira/Engenheiro de dados

É responsável pela manutenção da infraestrutura.

Desenvolve código.

Não tem conhecimento específico sobre o domínio.

Preocupa-se com a **manutenção do código**

Analista de dados

É responsável por tirar insights dos dados (ex: Por que estamos perdendo mercado na região X?)

Faz análises de negócio.

Possui conhecimento de negócio.

Excel/SQL/BI

Engenheira/Engenheiro de analytics

Extrai e transforma os dados para análise.

Desenvolve o Data Warehouse.

Possui conhecimento de negócio e programação.

Interage com os analistas e engenheiros de dados.

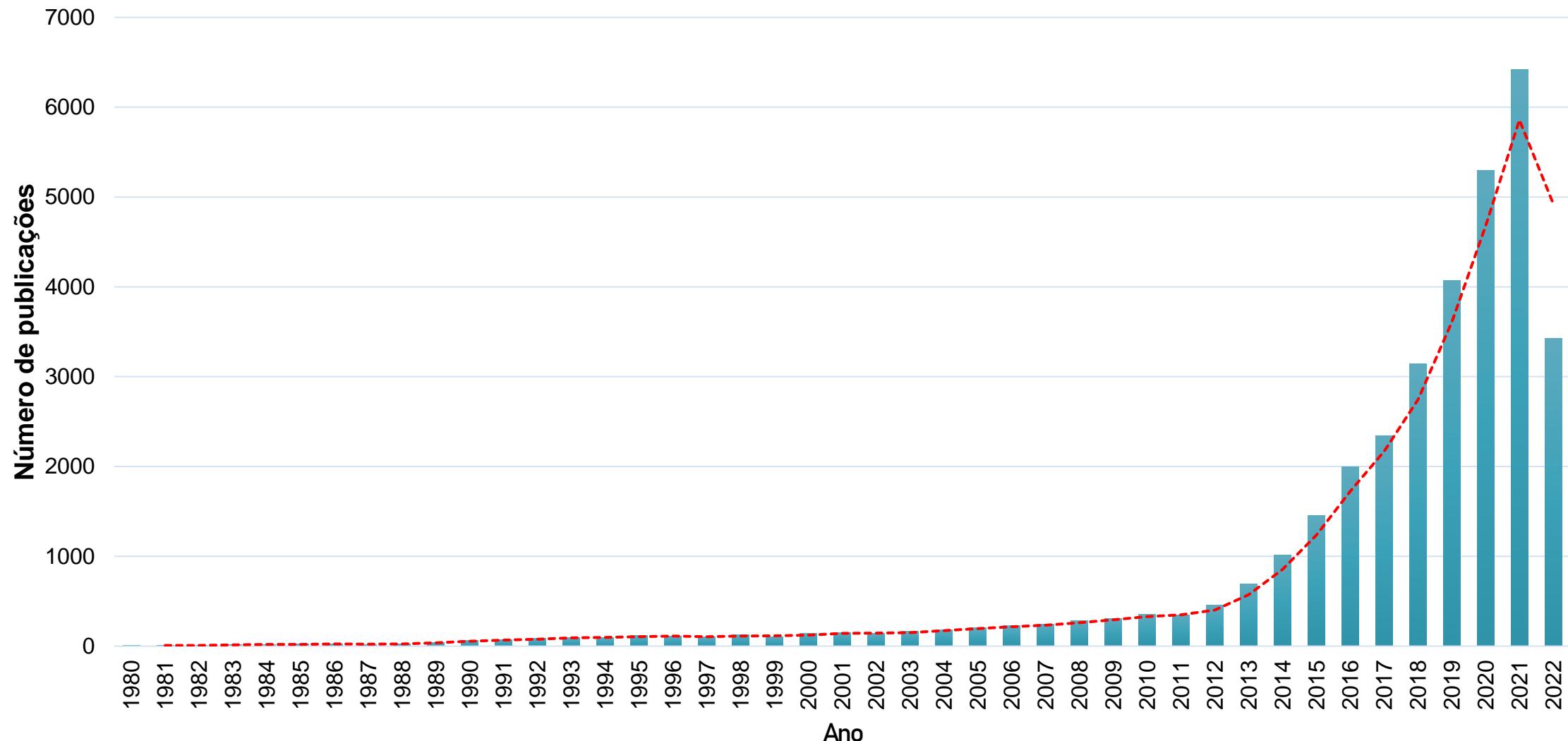
SQL/DBT/BI



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)



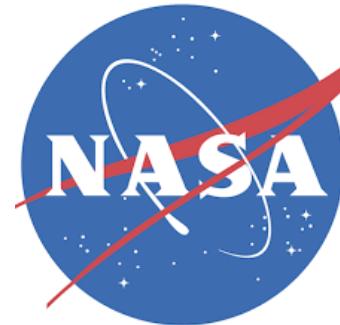
Big Data

Velocidade | Variedade | Volume

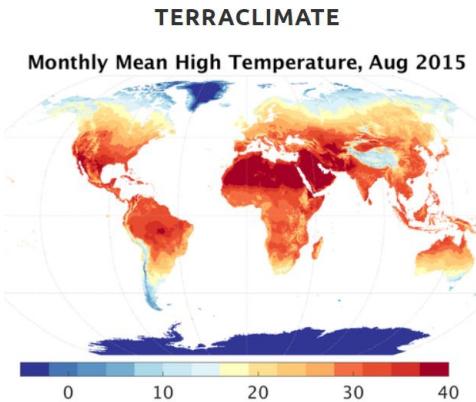


WikiAves

PDB
PROTEIN DATA BANK



FishBase



OBIS
OCEAN BIOGEOGRAPHIC
INFORMATION SYSTEM



iNaturalist



CALIFORNIA
ACADEMY OF
SCIENCES



NATIONAL
GEOGRAPHIC

GBIF

NCBI

DATASUS
Departamento de Informática do SUS



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 interactions between nectar-feeding vertebrates and flowering plants in the Atlantic Forest of Brazil

Joice Iamara-Nogueira , Naiara M. R. de Souza, Ana Maria Rui, Andréa C. Araújo, and others

First published: 22 November 2018

Corresponding Editor: William J. S. Reis



Data Paper | Free Access

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



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](#)

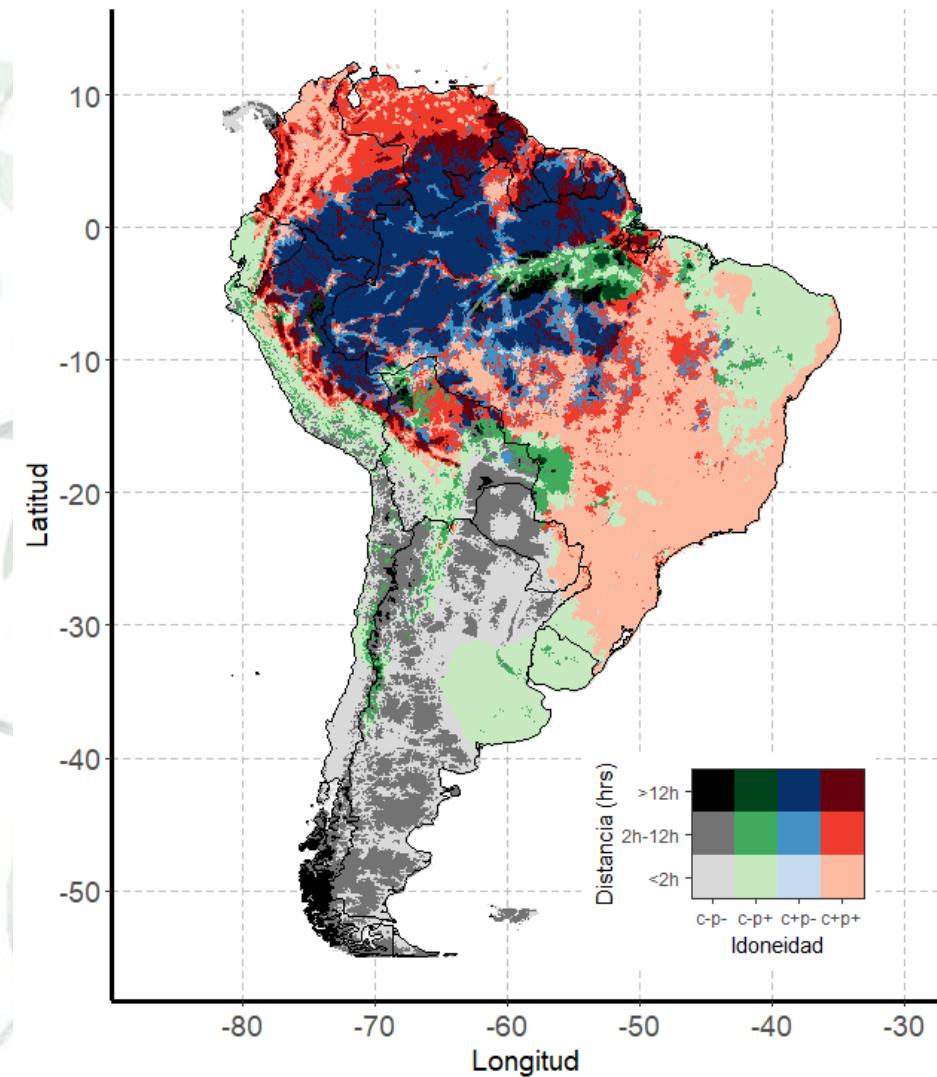


Brazil gains 1.7 million hectares of water by 2022, but continues to dry up

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

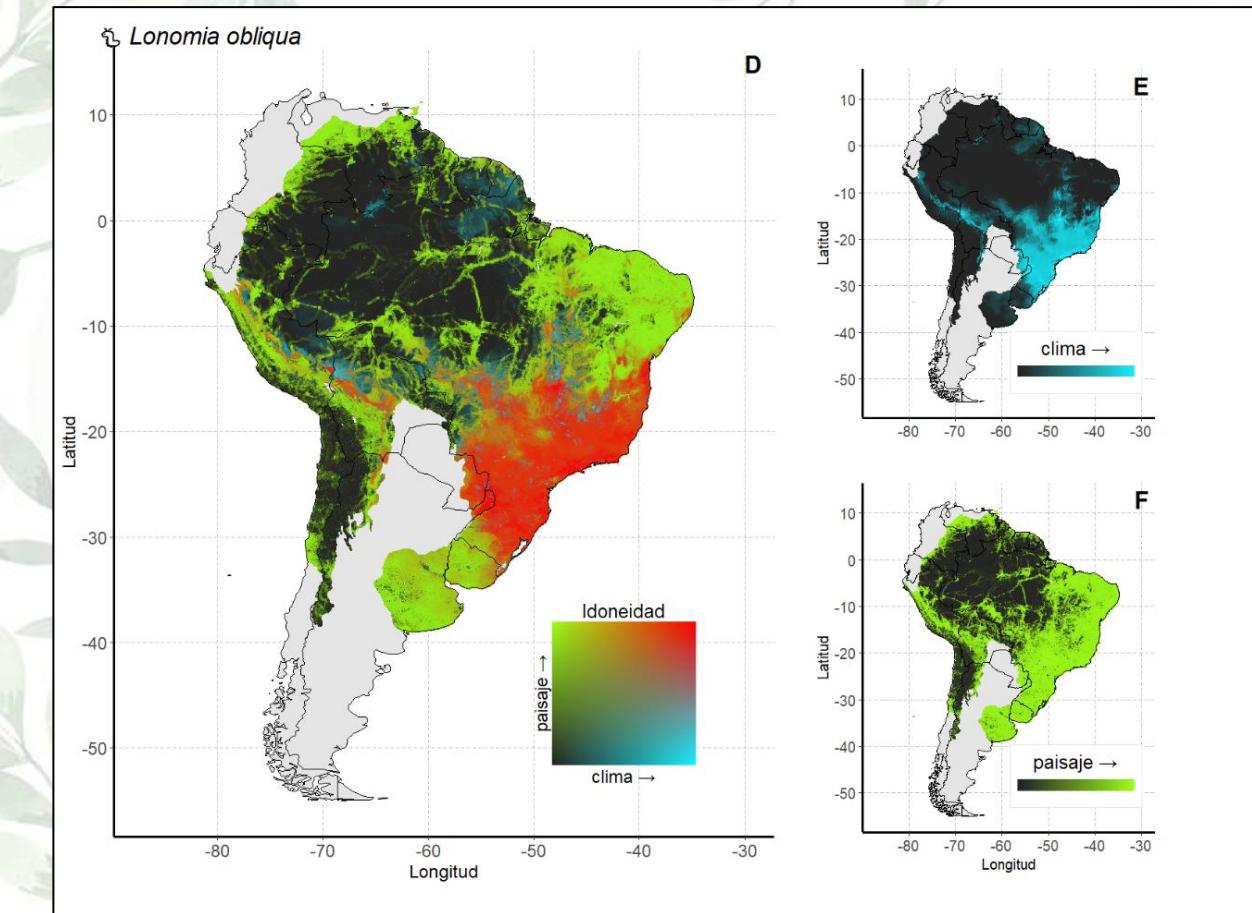
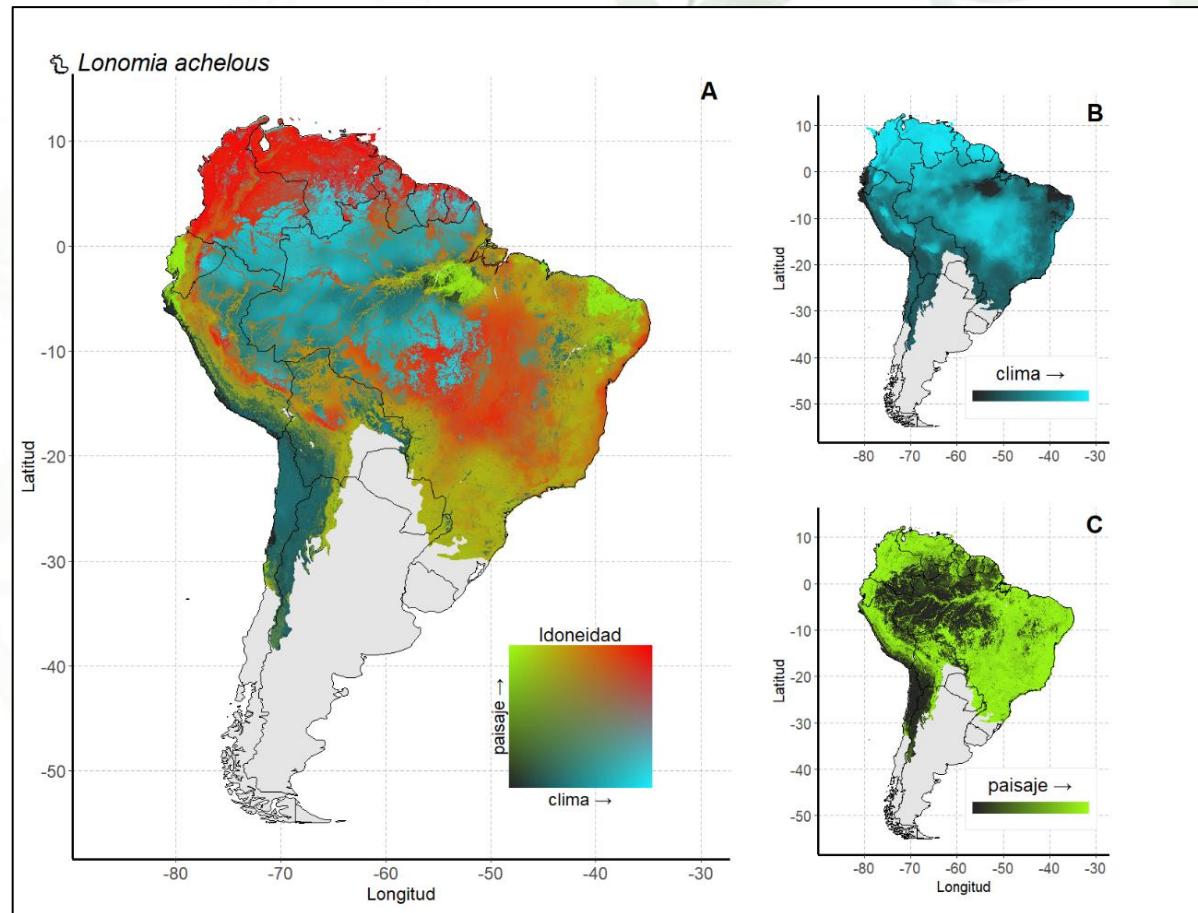
¡Oíだo donde tocas!

Un mapa de riesgo para el lonamismo en Sudamérica.



¡Cuidado donde tocas!

Un mapa de riesgo para el lonamismo 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

BILL & MELINDA
GATES foundation

Global Grand Challenges

ABOUT PARTNERSHIPS CHALLENGES AWARDED GRANTS GRANT OPPORTUNITIES NEWS



Catalyzing Equitable Artificial Intelligence (AI) Use

SHARE THIS



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



ChatGPT



ARTIFICIAL INTELLIGENCE



Vídeos

Obrigada



marilia.melofavalesso@gmail.com
www.mmfava.com