

R como ferramenta para projetos em Biologia

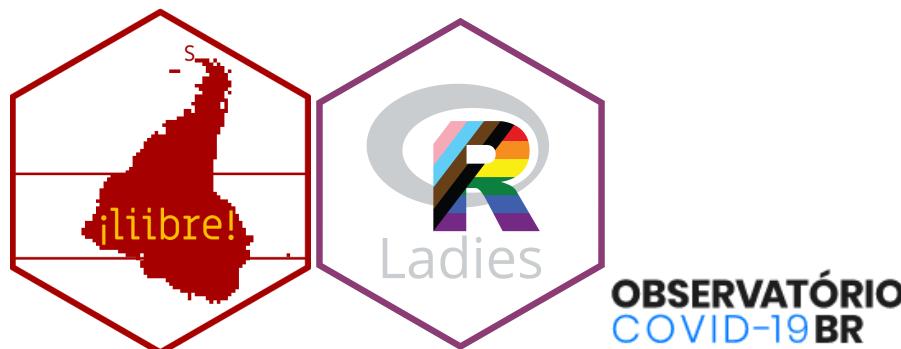
Sara Mortara

Instituto Internacional para Sustentabilidade IIS-Rio

13 de novembro 2020

sobre

- trabalho com modelagem estatística & ecologia de comunidades
- **¡liibre!** laboratório independente de biodiversidade e reproduzibilidade em ecologia
- pacotes **modler** e **coronabr**
- **@RLadiesRio**
- **Observatório COVID-19 BR**



disciplina *Projetos de análise de dados usando R*



Dra. Andrea Sánchez-Tapia **Boas práticas**
em análise de dados (Material [aqui](#))



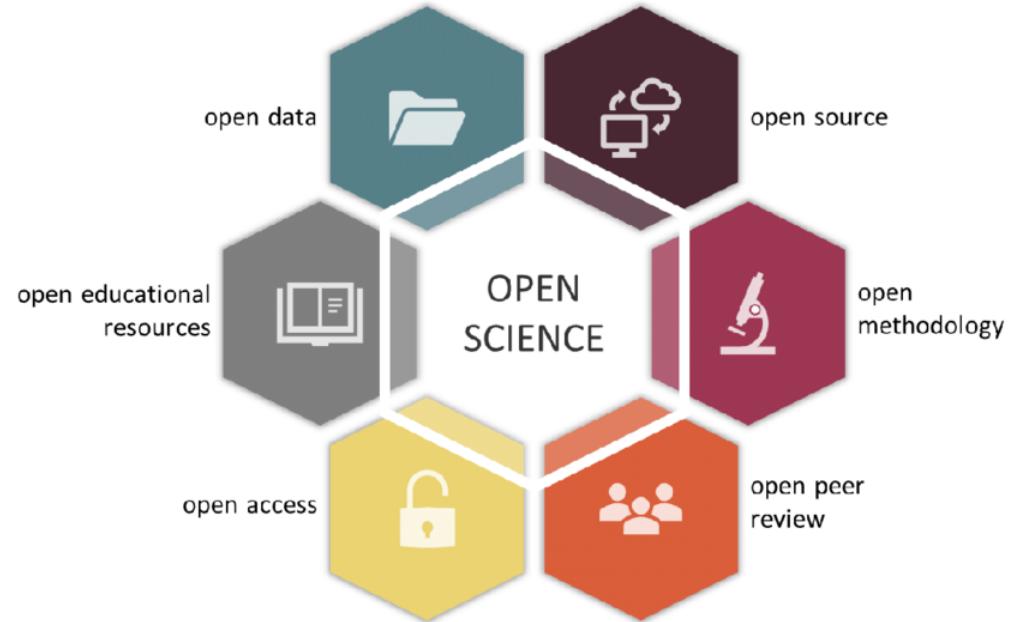
sobre hoje

1. por que usar o R?
2. como usar o R?
3. aplicações em Biodiversidade
4. outras aplicações

por que usar o R?

R de reproduzibilidade

- um dos pilares da **ciência aberta**
- ferramenta baseada em script
- permite a reconstrução dos passos
- cultura coletiva



por que R?

- projeto GNU
 - | I think of free as in free speech, not as in free beer
- FLOSS: **free, libre & open-source**
- script é essencial para reproduzibilidade, mas não a garante
- acessível (em comparação a outras linguagens de programação)
- muito comum na Biologia, Ciência de Dados e em diversas áreas
- **comunidade**

comunidade

The screenshot shows a user interface of the Stack Overflow website. At the top, there is a navigation bar with links for "About", "Products", and "For Teams". A search bar contains the placeholder "Search...". On the right side of the top bar are "Log in" and "Sign up" buttons. Below the navigation bar, there is a sidebar on the left with "Home" and "PUBLIC" buttons, and a "Stack Overflow" button which is highlighted with an orange border. The main content area displays a question titled "How to learn R as a programming language [closed]". Below the title, it says "Asked 10 years, 11 months ago" and "Active 1 year, 6 months ago". It also mentions "Viewed 88k times". To the right of the question is a blue "Ask Question" button.

ambiente para análise

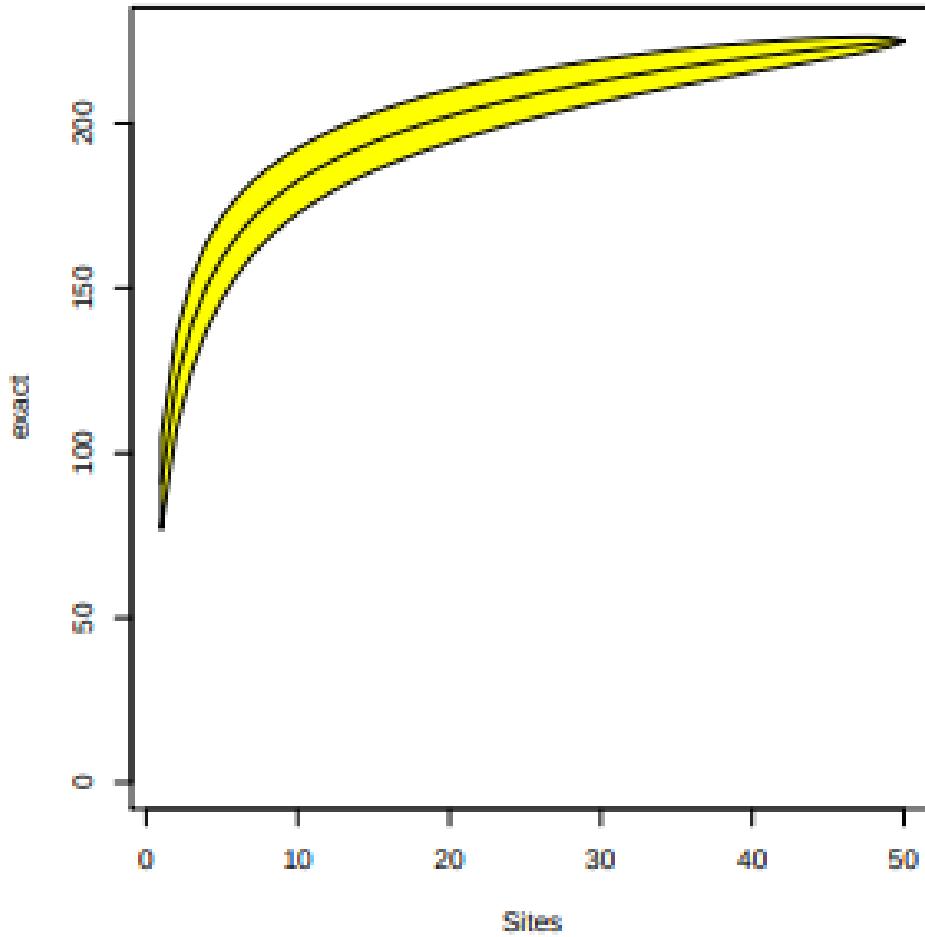
- manipulação de dados
- ampla coleção de ferramentas para análise de dados e criação de gráficos
- linguagem simples e efetiva que inclui condicionais `if{}`, laços `loop{}`, funções definidas pela pessoa usuária `function()`

construção de uma habilidade analítica

Uma das coisas mais importantes que você pode fazer é dedicar um tempo para aprender uma linguagem de programação de verdade. Aprender a programar é como aprender outro idioma: exige tempo e treinamento, e não há resultados práticos imediatos. Mas se você supera essa primeira subida íngreme da curva de aprendizado, os ganhos como cientista são enormes. Programar não vai apenas livrar você da camisa de força dos pacotes estatísticos, mas também irá aguçar suas habilidades analíticas e ampliar os horizontes de modelagem ecológica e estatística.

Tradução um tanto livre de Gotelli & Ellison, 2004. *A Primer of Ecological Statistics*. Sunderland, Sinauer.

motivação: pacote vegan & curva de acúmulo de espécies



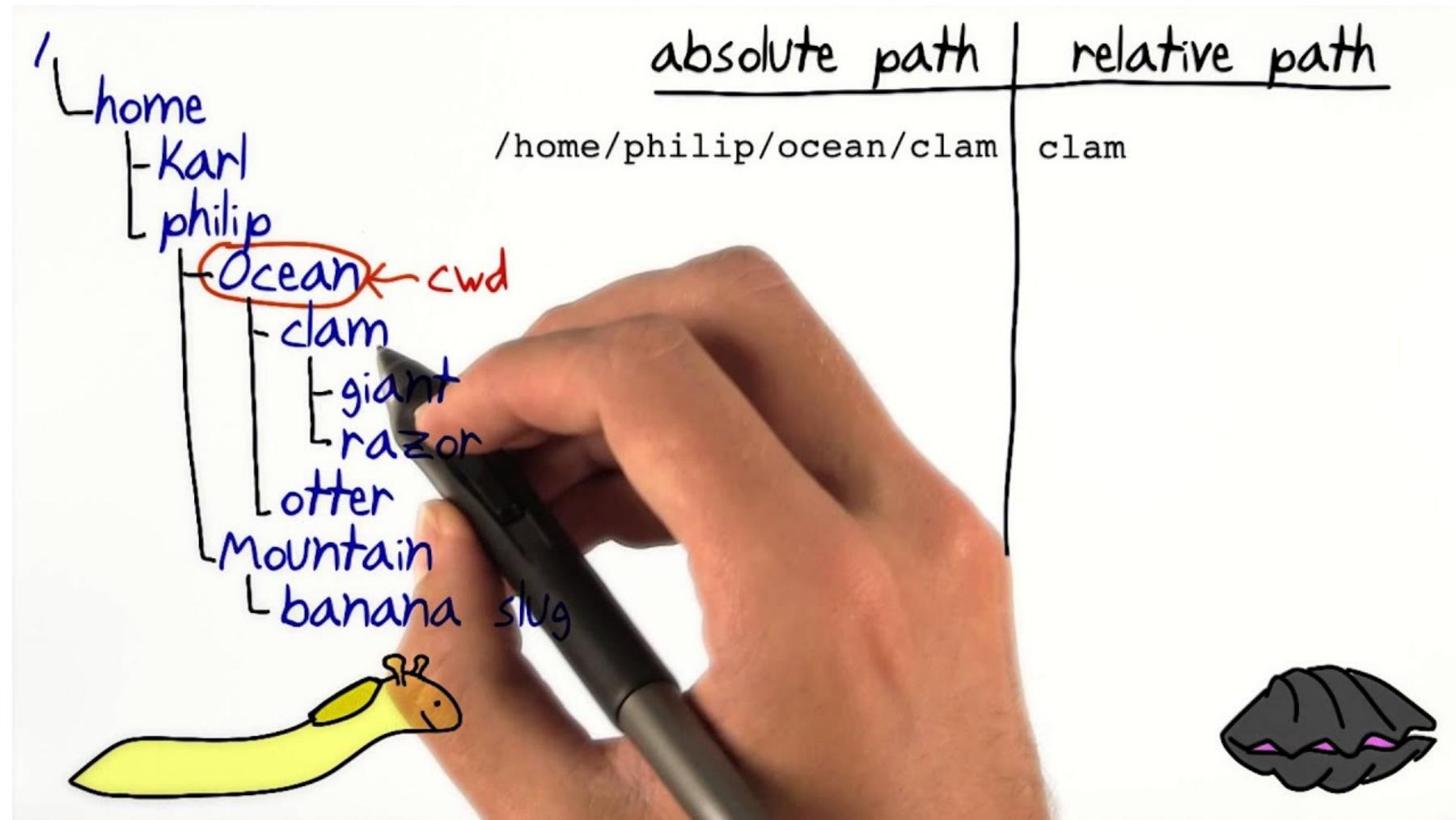
como usar o R?

criando um fluxo de trabalho

- **um projeto, uma pergunta, uma(s) análise(s)**
- cada projeto uma pasta **R_ufrb_demo**
- cada pasta é **autocontida**

```
•  
  └── data/  
  └── docs/  
  └── figs/  
  └── outputs/  
  └── R_ufrb.Rproj  
  └── README.md  
  └── scripts/
```

Caminhos relativos



usando projetos de RStudio

esqueça `setwd()` e conheça Jenny Bryan

If the first line of your R script is

```
setwd("C:\Users\jenny\path\that\only\I\have")
```

I will come into your office and SET YOUR COMPUTER ON FIRE 🔥.

If the first line of your R script is

```
rm(list = ls())
```

I will come into your office and SET YOUR COMPUTER ON FIRE 🔥.



.Rproj define o wd

The screenshot shows the RStudio interface with a project structure defined in a README.md file. The project structure is as follows:

```
1 # Um exemplo de um projeto de análise
2
3 Esta é uma estrutura básica de um projeto.
4
5 ...
6 ... └── data ..... # dados brutos e processados
7 ... └── docs ..... # apresentação
8 ... └── figs ..... # figuras geradas pelos códigos
9 ... └── outputs ..... # outputs gerados pelos códigos
10 ... └── R_ufrb_demo.Rproj ...
11 ... └── README.md ..... # explicação sobre o projeto
12 ... └── scripts ..... # código
13
14 Este projeto contém um exemplo de download de dados de uma espécie, Ziziphys joazeiro, incluindo uma limpeza das coordenadas e um mapa com os pontos de ocorrência.
15
16 ...
```

The Environment pane shows "Environment is empty". The Files pane displays the following files and folders:

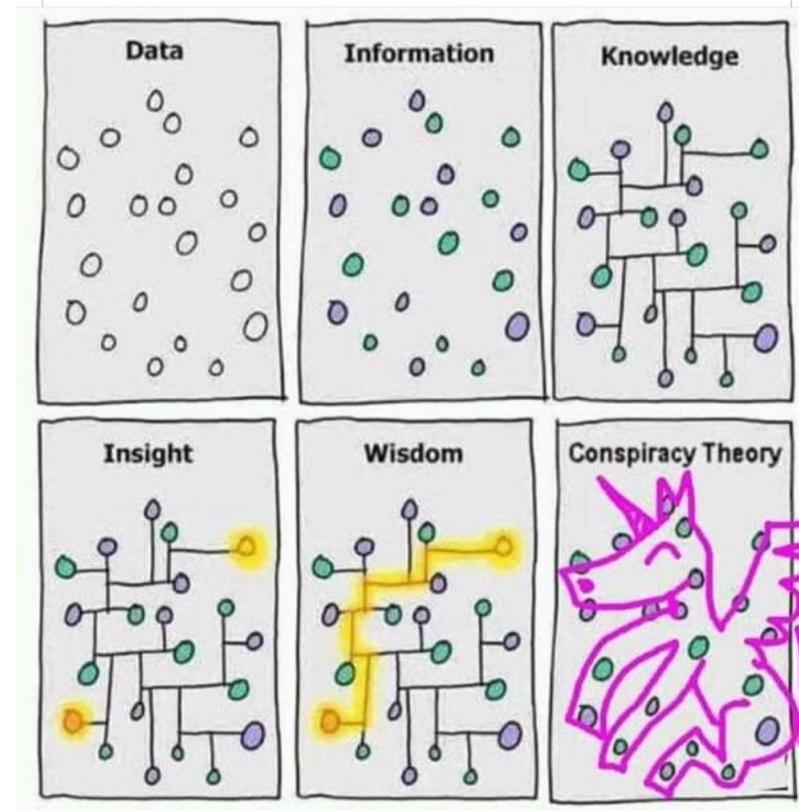
Name	Size	Modified
..		
.gitignore	136 B	Nov 13, 2020, 11:26 AM
data		
docs		
figs		
outputs		
R_ufrb.Rproj	257 B	Nov 13, 2020, 11:22 AM
README.md	1 KB	Nov 13, 2020, 11:21 AM
scripts		

usando caminhos relativos nos scripts

```
1 #·Script·para·criar·mapa·com·os·pontos·de·Z..joazeiro·
2 ·
3 #·Bibliotecas·
4 library(ggplot2)·
5 library(sf)·
6 library(ggspatial)·
7 ·
8 #·Carregando·o·objeto·com·o·limite·do·Brasil·para·o·R·
9 bra<-·readRDS("data/GADM/BRA_0_sf.rds")·
10 ·
11 #·Lendo·o·shapefile·dos·limites·da·caatinga·
12 caa<-·st_read("data/caatinga_border/caatinga_border.shp")·
13 ·
14 #·Lendo·os·dados·das·coordenadas·limpas·
15 coord<-·read.csv("data/Ziziphys_joazeiro_processed.csv")·
```

um projeto de análise

- uma boa pergunta
- dados
- conectando conceitos com dados



**como posso entender a distribuição de uma
espécie a partir de dados de coleções?**

aplicações em biodiveRsidade

dados de biodiversidade

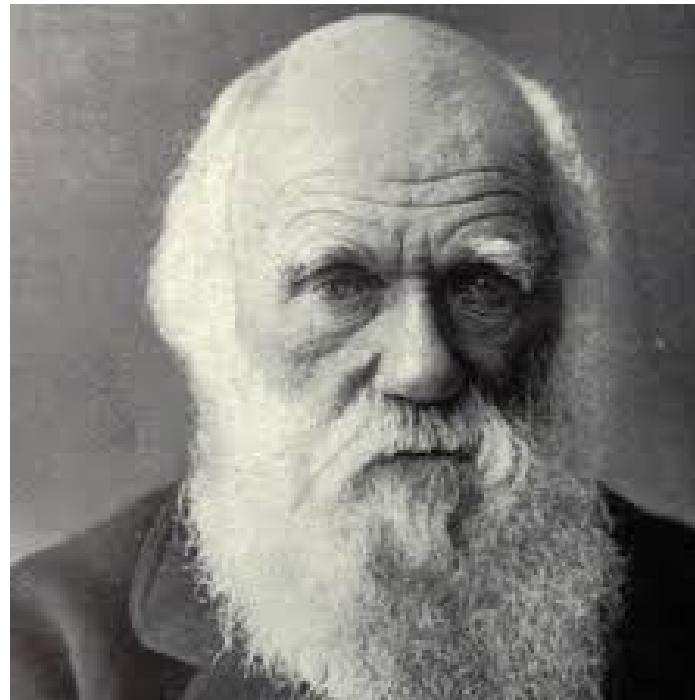
museus & herbários



padronização de dados em biodiversidade

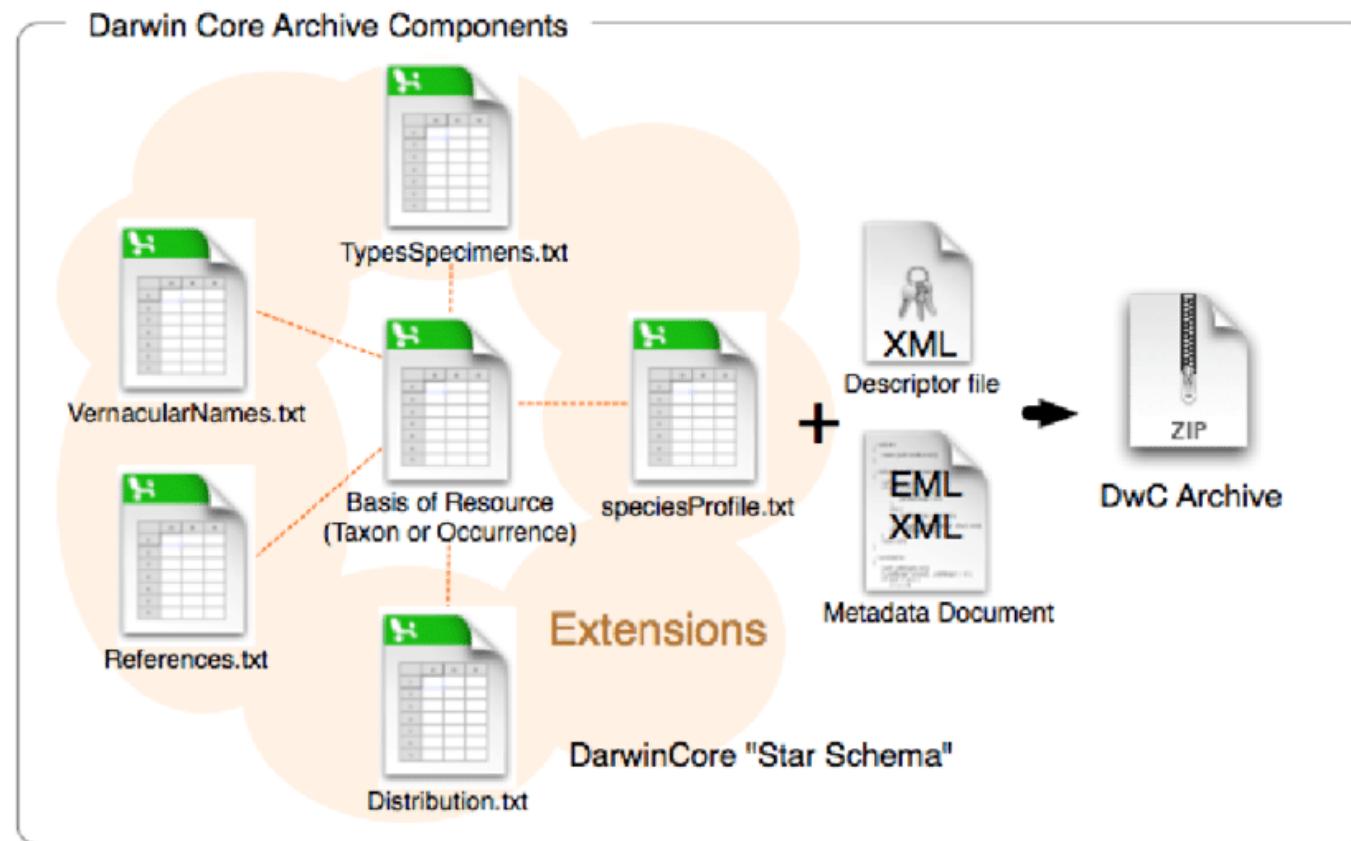
<https://dwc.tdwg.org/>

facilitar o compartilhamento da informação sobre diversidade biológica



padronização de dados em biodiversidade

<https://dwc.tdwg.org/>



como funcionam as bases de dados relacionais

- diferentes dados são organizados em diferentes tabelas
- tabelas são integradas
- identificador comum para cada tabela
- em geral organizadas em **SQL** (*Structured Query Language*)

Flora do Brasil 2020

The screenshot shows the homepage of the Flora do Brasil 2020 website. At the top left is the Reflora logo. To its right is the text "FLORA DO BRASIL 2020 - ALGAS, FUNGOS E PLANTAS". On the far right are language ("PT") and login buttons.

The main content area features a search bar with fields for "Nome", "Descrição", and "Forma de Vida e Substrato", each with dropdown menus and checkboxes. Below these are sections for "Imagens" and "Abrangência Geográfica".

A navigation menu at the top includes links for "Página Inicial", "Equipe", "Instituições", "Condição Atual dos Táxons", "Publicações Relacionadas", "Acesso aos Dados", and "Notícias".

A central text box titled "Flora do Brasil 2020" provides a detailed summary of the project's history and objectives:

No ano de 2010, o Brasil conseguiu cumprir a Meta 1 estabelecida pela Estratégia Global para a Conservação de Plantas (GSPC-CDB), com a publicação do Catálogo de Plantas e Fungos do Brasil (veja Publicações Relacionadas acima) e com o lançamento da primeira versão online da Lista de Espécies da Flora do Brasil. Este marco para a botânica brasileira só foi possível devido ao empenho de mais de 400 taxonomistas, brasileiros e estrangeiros, que trabalharam em uma plataforma, onde as informações sobre a nossa flora eram incluídas e divulgadas em tempo real. O projeto "Lista do Brasil", como ficou popularmente conhecido, foi encerrado em novembro de 2015, com a publicação de cinco artigos e suas respectivas bases de dados (veja Acesso aos Dados acima). Com grande entusiasmo apresentamos, em 2016, o novo sistema do projeto da Flora do Brasil 2020, que objetiva cumprir a Meta 1 estabelecida pela GSPC-CDB para 2020, com a divulgação de descrições, chaves de identificação e ilustrações para todas as espécies de plantas, algas e fungos conhecidos para o país. O projeto Flora do Brasil 2020 é parte integrante do Programa Reflora e está sendo realizado com o apoio do Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr). Conta no momento com 938 pesquisadores trabalhando em rede para a elaboração das monografias. Esses pesquisadores também são responsáveis por informações nomenclaturais e distribuição geográfica (abrangência no Brasil, endemismo e Domínios Fitogeográficos), além de incluirem dados valiosos sobre formas de vida, substrato e tipos de vegetação para as espécies monografadas. Os resultados das buscas nesta página também incluem informações sobre as espécies ameaçadas da nossa flora (devido à cooperação com o Centro Nacional de Conservação da Flora) e possibilitam acesso ao Index Herbariorum (devido à cooperação do The New York Botanical Garden). Além dessas informações, os usuários também podem ter acesso a imagens de exsicatas, inclusive de tipos nomenclaturais, provenientes tanto do Herbário Virtual Reflora, como do INCT Herbário Virtual da Flora e dos Fungos; bem como a imagens de plantas vivas e de ilustrações científicas, sendo todas as imagens incluídas pelos especialistas de cada grupo.

Acima, em "Condição Atual dos Táxons", você pode saber quais famílias e/ou gêneros já estão sendo monografados e quais ainda estão disponíveis. Caso você seja um taxonomista de formação e tenha interesse em participar deste projeto, envie um e-mail para o nosso contato indicando o grupo taxonômico de interesse para receber maiores informações.

Neste momento, são reconhecidas **49344** espécies para a flora brasileira (nativas, cultivadas e naturalizadas), sendo **4794** de Algas, **35741** de Angiospermas, **1569** de Briófitas, **5722** de Fungos, **113** de Gimnospermas e **1405** de Samambaias e Licofitas.

Get data How-to Tools Community About

SEARCH OCCURRENCES | 99,398,922 WITH COORDINATES

TABLE GALLERY MAP TAXONOMY METRICS DOWNLOAD

Occurrence status

Everything Present Absent

'Absent' is applied to an occurrence record when a survey of a taxon at a specific time and place encounters no specimens

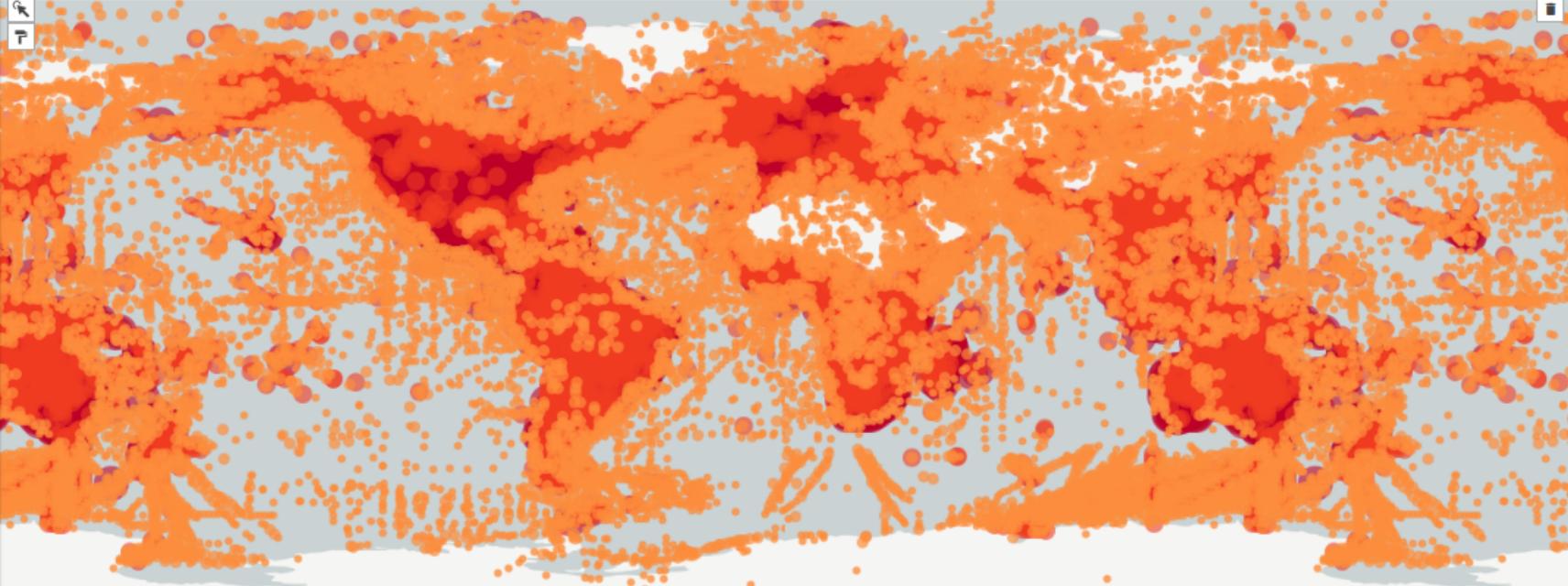
License

Scientific name

Basis of record

<input type="checkbox"/> Observation	18.911.086
<input type="checkbox"/> Machine observation	12.438.475
<input type="checkbox"/> Human observation	1.341.780.308
<input type="checkbox"/> Material sample	32.805.860
<input type="checkbox"/> Literature	537.286
<input checked="" type="checkbox"/> Preserved specimen	178.548.828
<input type="checkbox"/> Fossil specimen	11.328.578
<input type="checkbox"/> Living specimen	1.661.177
<input type="checkbox"/> Unknown	18.714.000

CLEAR REVERSE



ferramenta IPT do GBIF

IPT: The Integrated Publishing Toolkit

*A free open source software tool used to publish and share biodiversity datasets through the
GBIF network.*





INTEGRATED PUBLISHING TOOLKIT^(IPT)

free and open access to biodiversity data

JARDIM BOTANICO
DO RIO DE JANEIRO

Brazil Flora G (2020): Brazilian Flora 2020 project - Projeto Flora do Brasil 2020. v393.262. Instituto de Pesquisas Jardim Botanico do Rio de Janeiro. Dataset/Checklist. doi:10.15468/1mtkaw

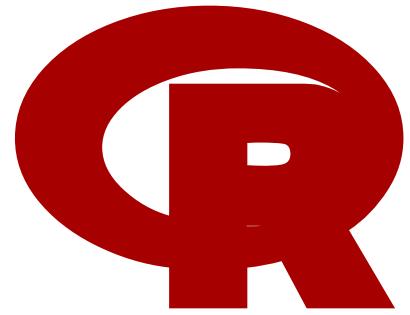
- 49.343 espécies
- 136.314 taxa
- informação sobre: distribuição, habitat, endemismo, referência ...

acessando por meio do R

- pacote `rgbif`
- pacote `flora`
- pacote `Rocc`

| Transforming science through open data, software & reproducibility

- conjunto de pacotes para dados abertos
- equipe de bioinformática muito ativa
- bases de dados, limpeza taxonômica ...
- pacotes **rgbif, taxize, taxview** ...



o quanto conhecemos da espécie?

Ziziphus joazeiro Mart.

boas práticas

- organizar bem suas pastas
- scripts **modulares**
- não misturar dados brutos com dados processados!
- **documentação** é essencial

task views - Environmetrics

CRAN Task View: Analysis of Ecological and Environmental Data

Maintainer: Gavin Simpson

Contact: ucfagls at gmail.com

Version: 2020-08-30

URL: <https://CRAN.R-project.org/view=Environmetrics>

Introduction

This Task View contains information about using R to analyse ecological and environmental data.

The base version of R ships with a wide range of functions for use within the field of environmetrics. This functionality is complemented by a plethora of packages available via CRAN, which provide specialist methods such as ordination & cluster analysis techniques. A brief overview of the available packages is provided in this Task View, grouped by topic or type of analysis. As a testament to the popularity of R for the analysis of environmental and ecological data, a [special volume](#) of the *Journal of Statistical Software* was produced in 2007.

task views - Phylogenetics

CRAN Task View: Phylogenetics, Especially Comparative Methods

Maintainer: Brian O'Meara

Contact: omeara.brian at gmail.com

Version: 2020-06-16

URL: <https://CRAN.R-project.org/view=Phylogenetics>

The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical patterns along phylogenetic trees. This task view describes R packages that implement a variety of different comparative phylogenetic methods. This is an active research area and much of the information is subject to change. One thing to note is that many important packages are not on CRAN: either they were formerly on CRAN and were later archived (for example, if they failed to incorporate necessary changes as R is updated) or they are developed elsewhere and have not been put on CRAN yet. Such packages may be found on GitHub, R-Forge, or authors' websites.

ouTRas aplicações

pacote coronabr

coronabr 0.1.0  funções como usar mais exemplos e mapas ▾ sobre 

Download de dados de COVID-19 no Brasil

coronabr é um pacote de R para fazer *download* e visualizar os dados dos casos diários de coronavírus (COVID-19) disponibilizados por diferentes fontes:

- Ministério da Saúde;
- Brasil I/O;
- Johns Hopkins University



Nosso objetivo

O nosso objetivo é facilitar o acesso aos dados de diferentes fontes, usando ferramentas de acesso aberto e que permitam reprodutibilidade.

O código é aberto. Entre em [como usar](#) para um exemplo de como utilizar o pacote. Compartilhe.

Fazemos ciência aberta, democrática e reprodutível. Este é um trabalho em desenvolvimento. Para entender como contribuir, clique [aqui](#).

Aviso!

Links

Browse source code at
<https://github.com/lilibre/coronabr/>
Report a bug at
<https://github.com/lilibre/coronabr/issues>

License

GPL (>= 3)

Community

[Contributing guide](#)

Developers

Sara Mortara
Author, maintainer 

Andrea Sánchez-Tapia
Author 

Karlo Guidoni Martins
Author 

dados, responsabilidade & contexto

- dados deveriam ser abertos e acessíveis

Trânsparência COVID-19 OPEN KNOWLEDGE BRASIL

- nem toda análise que **pode** ser feita, **deve** ser feita
- cada dado diz respeito a uma pessoa
- para COVID-19 e SRAG: **subnotificação** & **atraso**
- inconsistência com dados reportados em diferentes escalas: município, estado, país

Macapá-AM

```
# Script para download de dados de covid-19 no município de Macapá

# Para instalar o pacote use:
#remotes::install_github("libre/coronabr")

# 1. bibliotecas #####
library(coronabr)
library(ggplot2)

# 2. download #####
## dados macapa-ap usando o geocode IBGE
ma <- get_corona_br(filename = "macapa",
                     dir = "data/",
                     ibge_cod = "1600303")
```

entendendo os dados

```
# 3. inspeção dos dados #####
head(ma[, c(3, 4, 9:18)])
```

```
##      date epidemiological_week last_available_confirmed last_available_confirmed_per_100k_inhabitants last_available_date
## 1 2020-03-20                  12                      1                           0.19497 2020-03-20
## 2 2020-03-21                  12                      1                           0.19497 2020-03-21
## 3 2020-03-22                  13                      1                           0.19497 2020-03-22
## 4 2020-03-23                  13                      1                           0.19497 2020-03-23
## 5 2020-03-24                  13                      1                           0.19497 2020-03-24
## 6 2020-03-25                  13                      2                           0.38994 2020-03-25
##   last_available_death_rate last_available_deaths order_for_place place_type state new_confirmed new_deaths
## 1                      0                  0                 1     city    AP        1        0
## 2                      0                  0                 2     city    AP        0        0
## 3                      0                  0                 3     city    AP        0        0
## 4                      0                  0                 4     city    AP        0        0
## 5                      0                  0                 5     city    AP        0        0
## 6                      0                  0                 6     city    AP        1        0
```

entendendo os dados

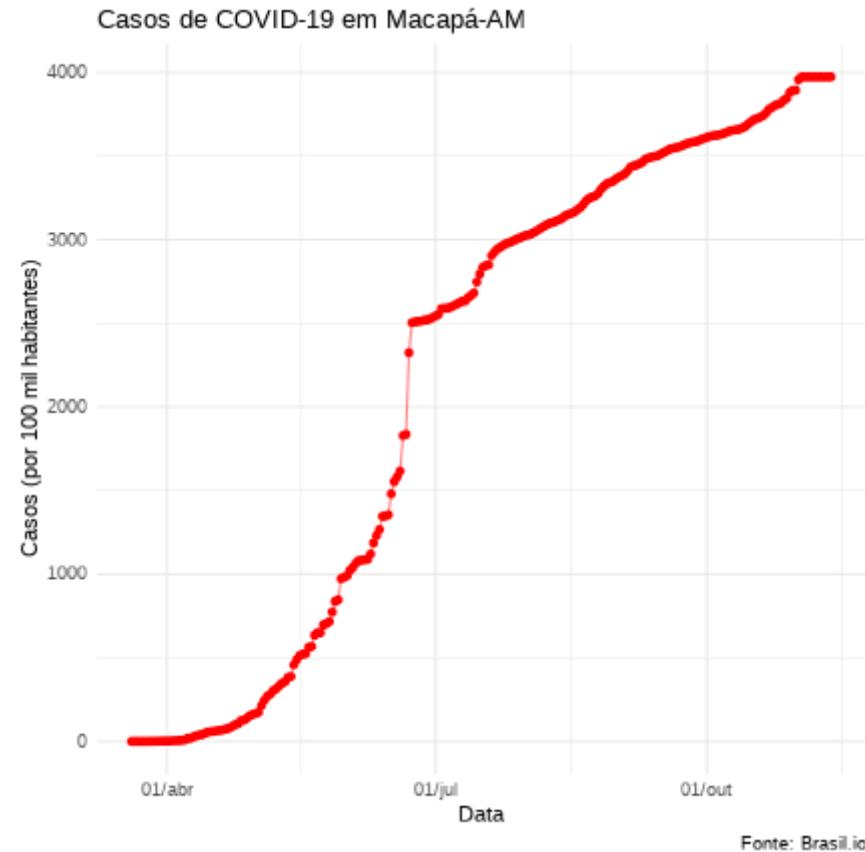
```
tail(ma[, c(3, 4, 9:18)])
```

```
##           date epidemiological_week last_available_confirmed last_available_confirmed_per_100k_inhabitants last_available_date
## 233 2020-11-07                  45                 20375                         3972.494 2020-11-02
## 234 2020-11-08                  46                 20375                         3972.494 2020-11-02
## 235 2020-11-09                  46                 20375                         3972.494 2020-11-02
## 236 2020-11-10                  46                 20375                         3972.494 2020-11-02
## 237 2020-11-11                  46                 20375                         3972.494 2020-11-02
## 238 2020-11-12                  46                 20375                         3972.494 2020-11-02
##   last_available_death_rate last_available_deaths order_for_place place_type state new_confirmed new_deaths
## 233          0.0255             520            233     city    AP       0        0
## 234          0.0255             520            234     city    AP       0        0
## 235          0.0255             520            235     city    AP       0        0
## 236          0.0255             520            236     city    AP       0        0
## 237          0.0255             520            237     city    AP       0        0
## 238          0.0255             520            238     city    AP       0        0
```

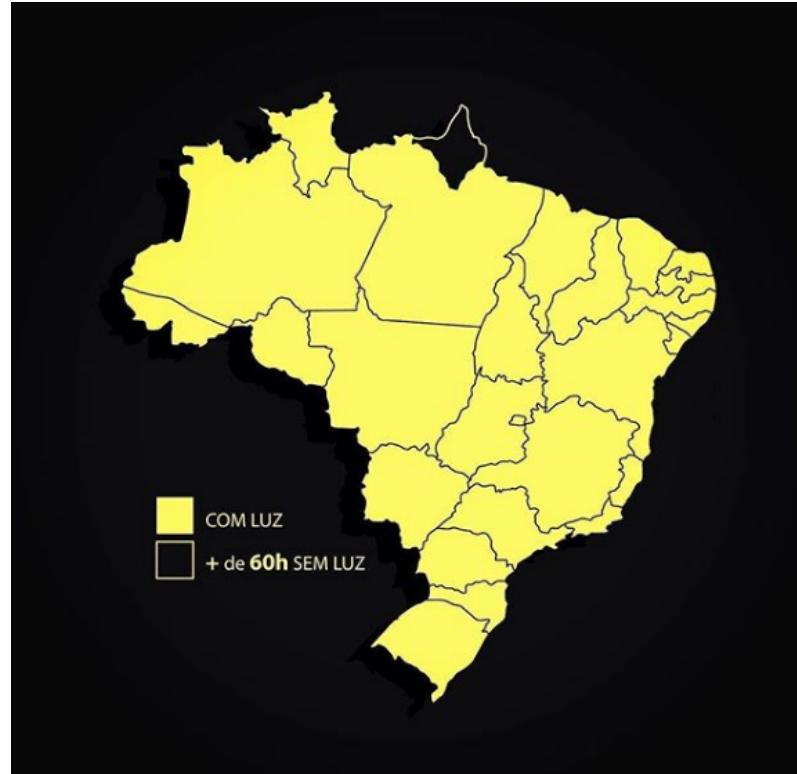
criando um gráfico com ggplot2

```
# 4. Fazendo um gráfico simples ####
ggplot(ma, aes(x = date,
                y = last_available_confirmed_per_100k_inhabitants)) +
  geom_line(color = "red",
            alpha = .5) +
  geom_point(color = "red") +
  scale_x_date(date_labels = "%d/%b") +
  labs(x = "Data",
       y = "Casos (por 100 mil habitantes)",
       caption = paste0("Fonte: Brasil.io"),
       title = "Casos de COVID-19 em Macapá-AM") +
  theme_minimal()
```

nosso gráfico



apagão



© @designativista

casos não reportados



Alcinéa Cavalcante
@alcinea

...

Covid - 216 novos casos no Amapá; 234 pacientes internados, sendo 68 em UTI alcinea.com/saude/covid-21... via @alcinea

Covid – 216 novos casos no Amapá; 234 pacientes internado...
O Governo do Amapá atualizou nesta quarta-feira, 11, o boletim informativo sobre a situação do novo coronavírus no estado. ...
alcinea.com

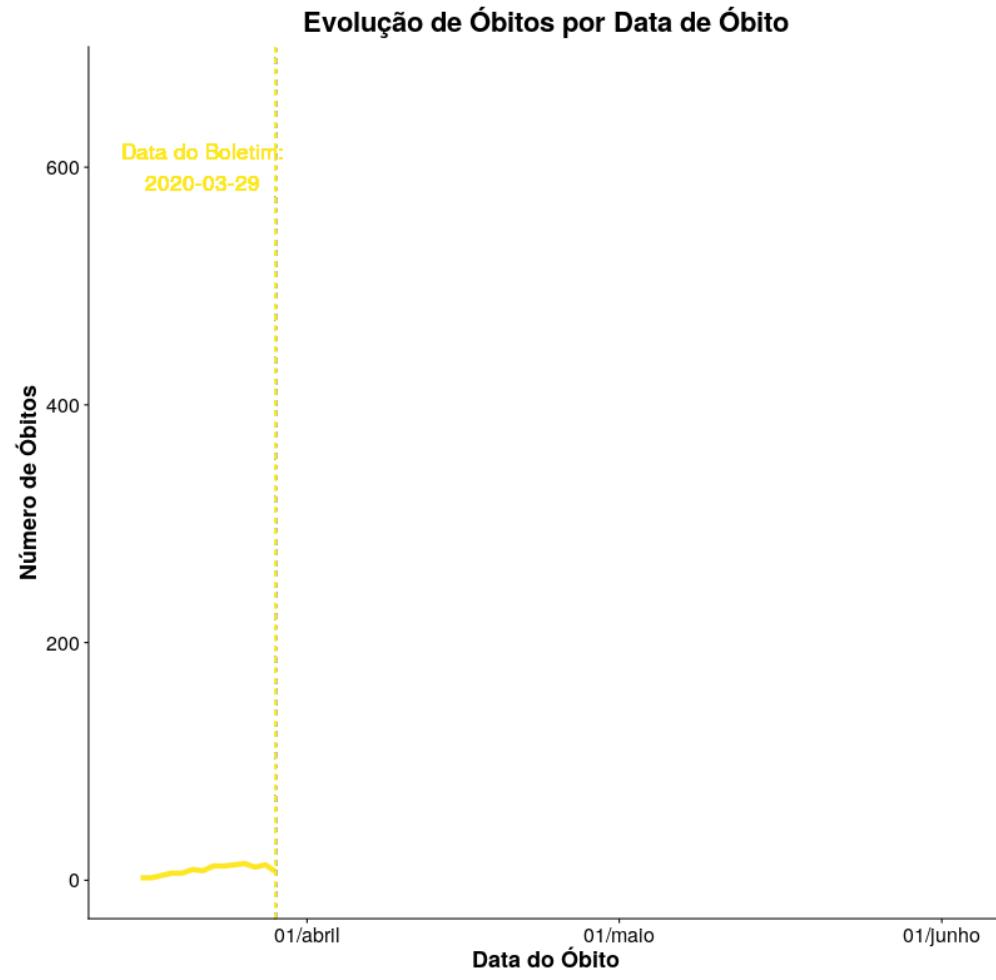
9:32 PM · 11 de nov de 2020 · Twitter Web App

 @alcinea

ressalvas em relação aos dados brutos

1. **apagão**
2. **subnotificação**
3. **atraso** na entrada dos dados no sistema - tanto para casos como óbitos

como vemos o atraso



Observatório Covid-19BR

OBSERVATÓRIO
COVID-19
BR

Visão Geral

Estados Municípios DRS - SP Cenários Análises Florianópolis Perguntas Comuns Informações Técnicas Reportagens Sobre

VISÃO GERAL

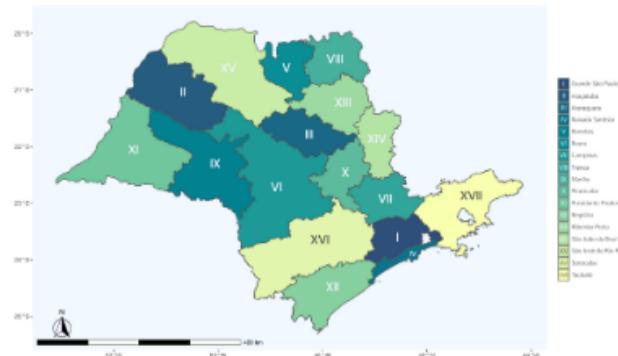
Última atualização: 20:21 · 03/11/2020

Conteúdo novo!

 COVID-19: A solução passa pelo SUS
Confira a versão estendida do artigo de opinião publicado originalmente na Folha de São Paulo.



Departamentos Regionais de Saúde de São Paulo
Acompanha a situação por DRS - SP

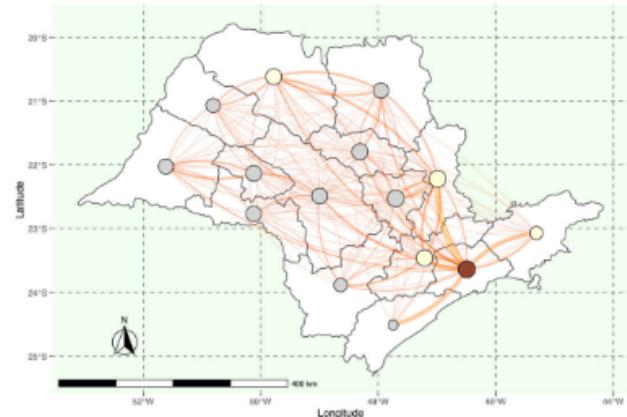


Realizamos um acompanhamento independente da situação epidemiológica em



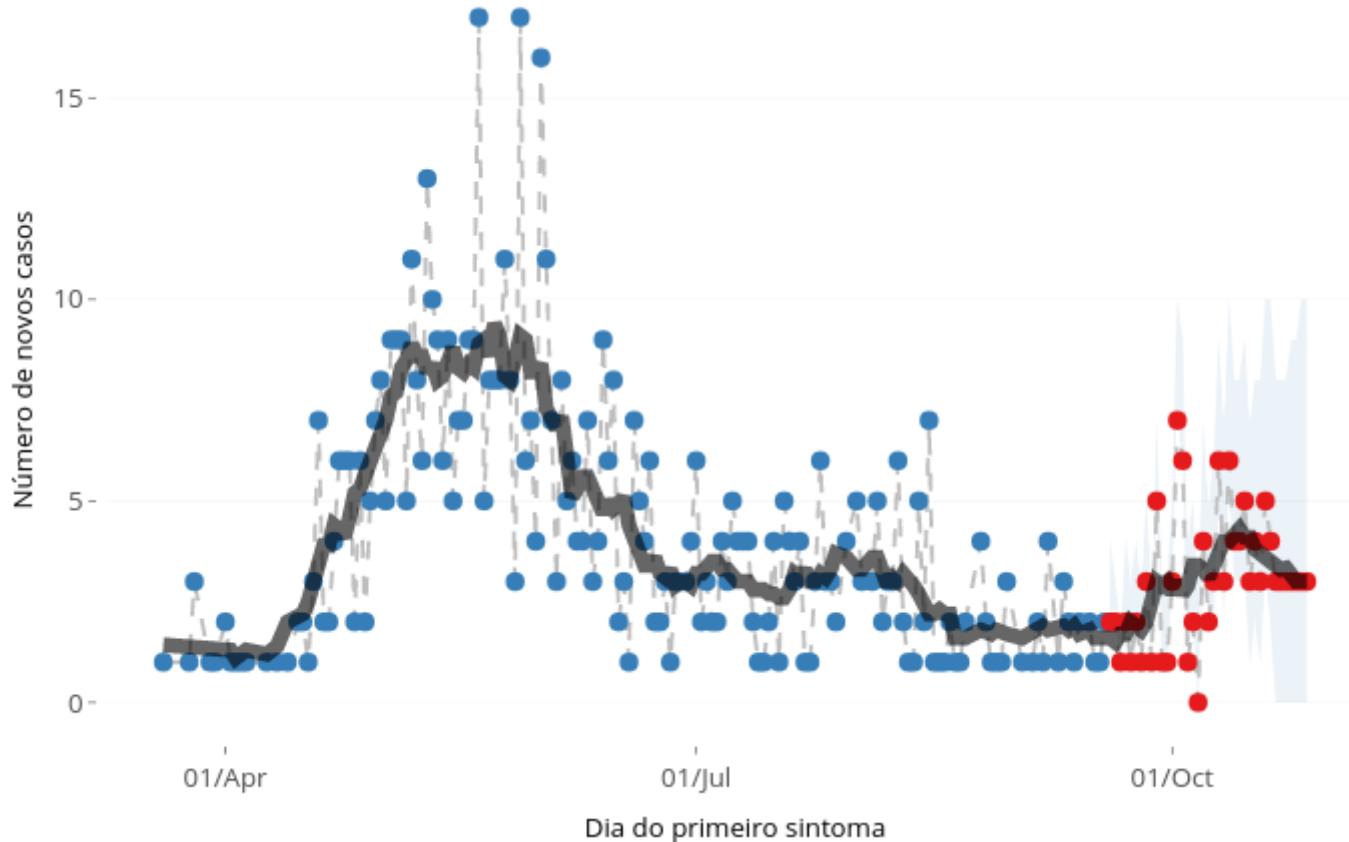
Dinâmicas de contágio 2

E como as epidemias se espalham por dentro de um estado?

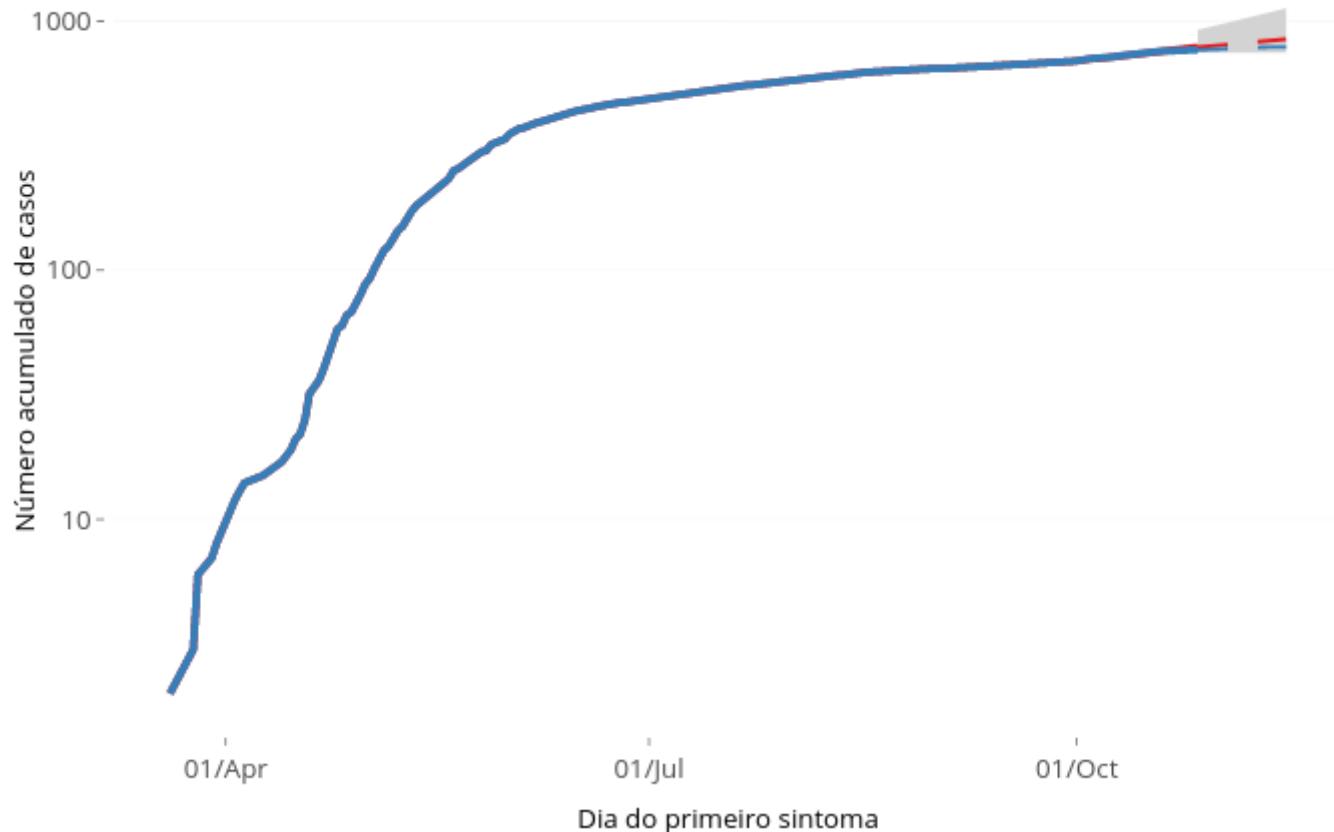


Neste relatório, associamos dados da rede de

correção por nowcasting



comparação entre notificados e estimados



outros tipos de receitas com bookdown

O que você precisa saber



Sobre as receitas

Receitas

Biscoitos

Geléias e Compotas

Outros Pratos Salgados

Tortas e Bolos Doces

Receitas

Biscoitos

Biscoitinhos de Cerveja do Natal



Receita de Tia Elisa



"Eu gosto mais daqueles que ficam mais escuros, mas isto é preferência individual" Tio Valerio

"Todas as famílias saiam do Natal com suas preciosas latinhas, decoradas pela Tia Elisa com papel de presente e forradas com papel manteiga. Era um jeito de levar o Natal para casa" Julia

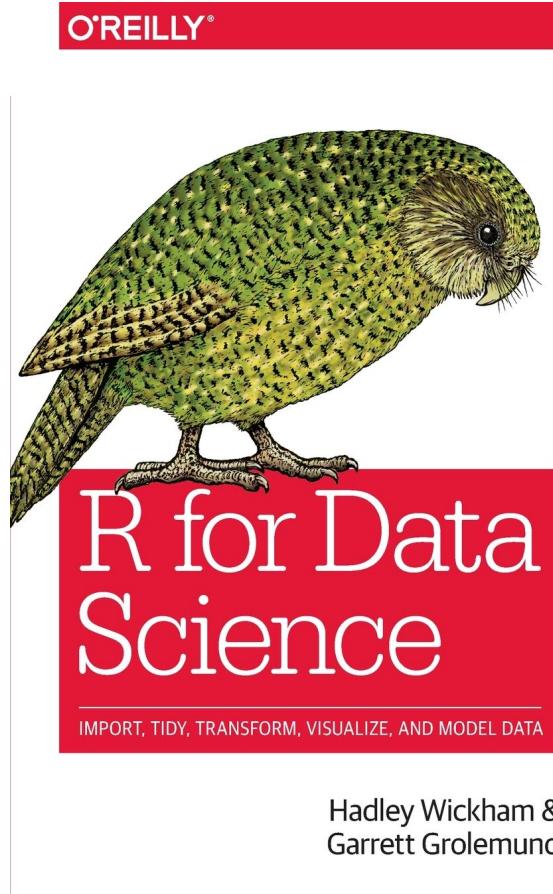
É *trabalhosa!* Tempo de preparo de 3 horas.

Rendimento: Para cada lata (porção familiar do Natal), entre 250 e 300 g de farinha

recursos

- Projetos de análise de dados usando R
- Data cleaning tools. ENM Course 2020
- Canal de Youtube R-Ladies+ Rio
- Canal de Youtube R-Ladies+ BH
- Jenny Bryan. Project Oriented Workflow
- Escrevendo manuscritos acadêmicos usando R Markdown
- Page Piccinini. Curso R - R & git setup.

recursos



Disponível em: <https://r4ds.had.co.nz/>

para ter em mente

- pensar no fluxo de trabalho
- R: reproduzibilidade e responsabilidade
- os dados não falam por si só
- entender o contexto é essencial para fazer uma boa análise
- manter-se libre!



obrigada!

 saramortara@gmail.com

 @MortaraSara

  jliibre!

 saramortara