

# O Bushido do Script

## Manejo, Visualização e Compartilhamento de Dados

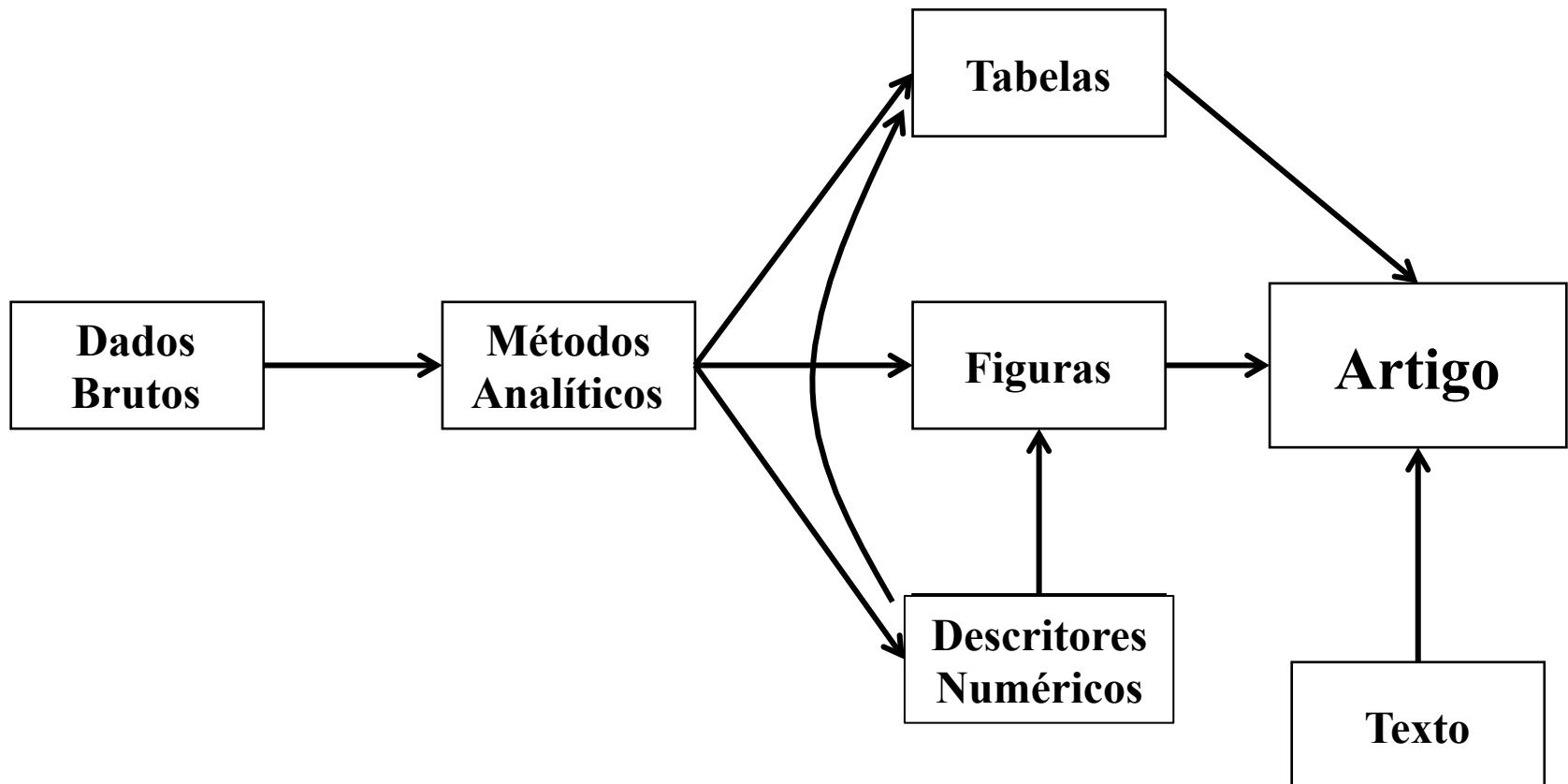
Nicholas A. C. Marino

[nac.marino@gmail.com](mailto:nac.marino@gmail.com)

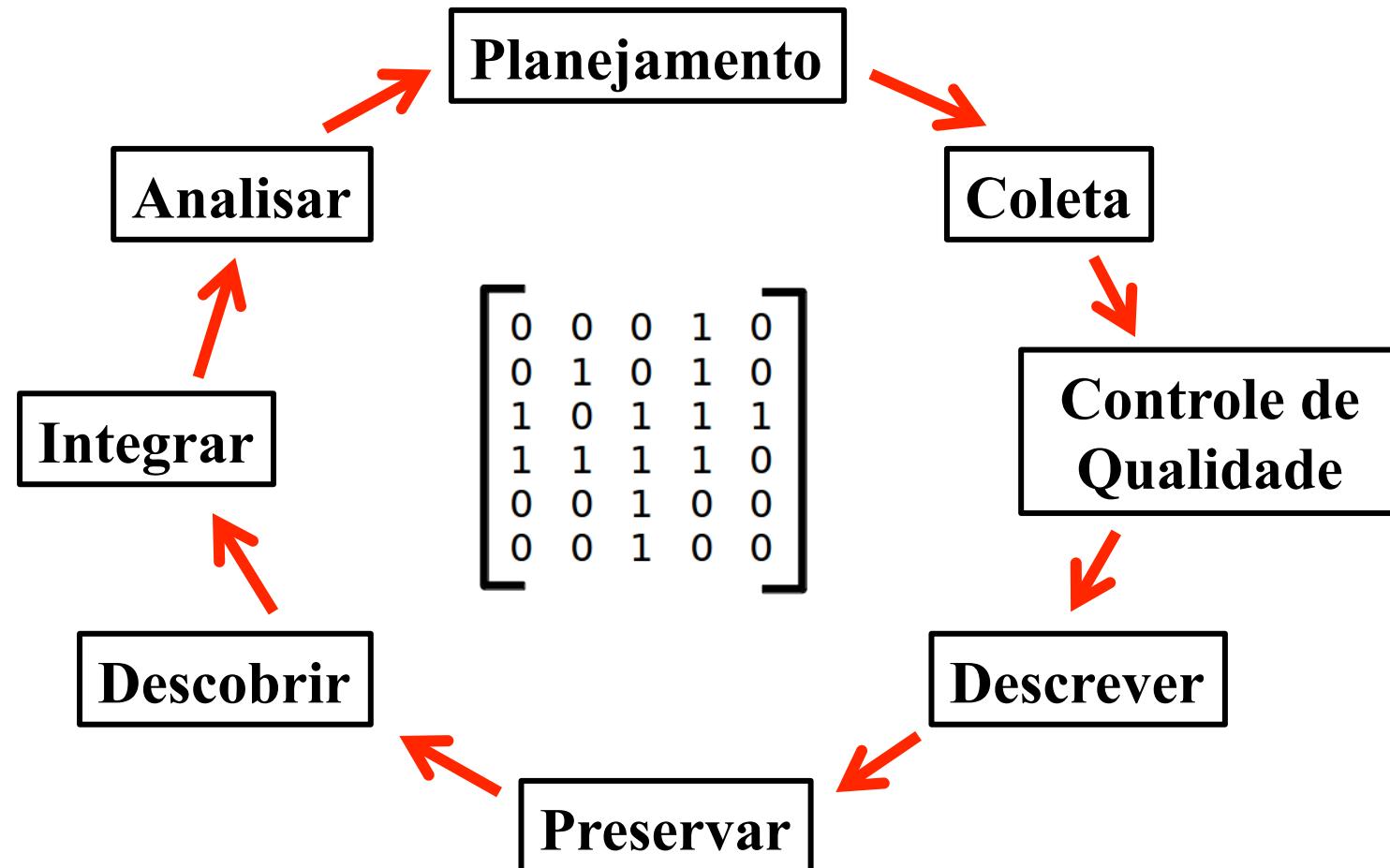
[github.com/nacmarino/compartilhaR](https://github.com/nacmarino/compartilhaR)



# O processo de produção de um artigo científico



# O ciclo de vida dos dados não termina na publicação de um artigo científico



# O compartilhamento de dados é parte atual e fundamental do processo de produção de um artigo científico

- *Open data*: compartilhamento e abertura de dados, pelo menos quando da publicação do trabalho.
  - ✓ Torna o processo de produção científica mais transparente;
  - ✓ Dá mais robustez aos seus achados;
  - ✓ Reduz a possibilidade de fraude na criação de dados;
  - ✓ Na maioria das vezes, *torna público o que é público*: a maior parte da pesquisa científica é financiada com dinheiro público;
  - ✓ A abertura da disponibilidade de dados pode ajudar na conservação de espécies e ecossistemas.

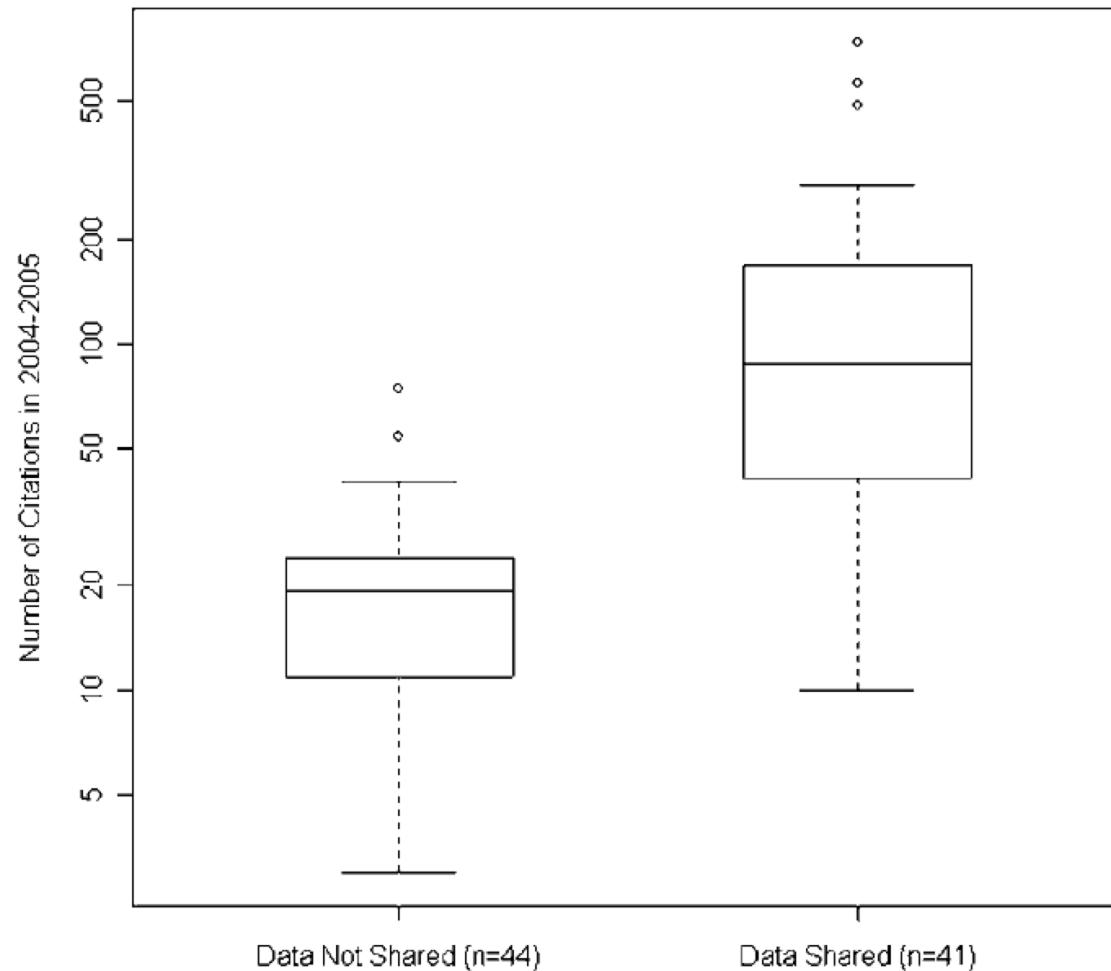
# A sobrevivência dos dados a longo prazo pode ser aumentada através de seu compartilhamento

- A chance de um dado se perder aumenta com o tempo: reinventando a roda.

Table 1. Breakdown of Data Availability by Year of Publication

Year	No Working E-Mail	No Response to E-Mail	Response Did Not Give Status of Data	Data Lost	Data Exist, Unwilling to Share	Data Received
1991	9 (35%)	9 (35%)	2 (8%)	4 (15%)	1 (4%)	1 (4%)
1993	14 (39%)	11 (31%)	3 (8%)	7 (19%)	0 (0%)	1 (3%)
1995	11 (31%)	9 (26%)	0 (0%)	7 (20%)	2 (6%)	6 (17%)
1997	11 (37%)	9 (30%)	1 (3%)	2 (7%)	3 (10%)	4 (13%)
1999	19 (48%)	13 (32%)	1 (2%)	1 (2%)	0 (0%)	6 (15%)
2001	13 (30%)	15 (35%)	3 (7%)	4 (9%)	0 (0%)	8 (19%)
2003	9 (20%)	20 (43%)	4 (9%)	2 (4%)	0 (0%)	11 (24%)
2005	11 (24%)	14 (31%)	6 (13%)	1 (2%)	0 (0%)	13 (29%)
2007	12 (18%)	31 (47%)	2 (3%)	4 (6%)	1 (2%)	16 (24%)
2009	9 (13%)	34 (49%)	3 (4%)	5 (7%)	6 (9%)	12 (17%)
2011	13 (16%)	29 (36%)	8 (10%)	0 (0%)	7 (9%)	23 (29%)
Totals	131 (25%)	194 (38%)	33 (6%)	37 (7%)	20 (4%)	101 (19%)

# Trabalhos que compartilham seus dados são, geralmente, mais citados

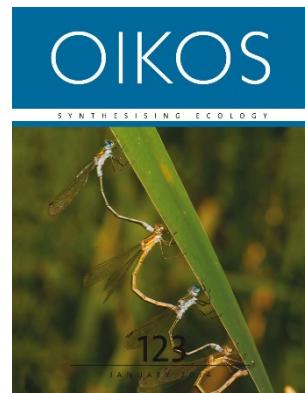
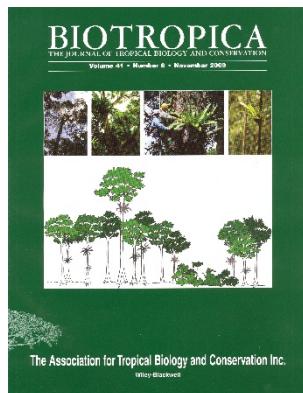
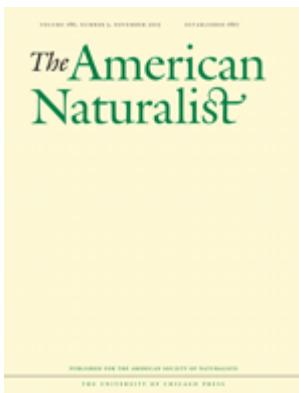


# Você pode não gostar, mas...

- Cientistas no início da carreira tendem a não querer compartilhar seus dados (Tenopir et al, 2014, PLOS One):
  - ✓ “**Posso ter cometido algum erro.**” (mas se está errado, precisa ser consertado de qualquer forma, não?);
  - ✓ “**Alguém pode publicar esses dados antes de mim**” (não é ético, e a academia tem punições severas para pessoas que cometem esse deslize).
  - ✓ “**Se alguém usar meu dado, não vai me dar crédito pela publicação**” (você consegue dar um DOI para praticamente TUDO).

# E por mais que você não queira...

- Revistas na área de ecologia e evolução estão exigindo o compartilhamento de dados para a publicação dos artigos.



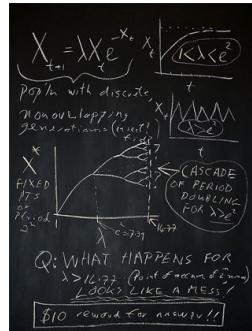
**+ 100 revistas, cadastradas na Dryad Digital Repository**  
(PLOS, Ecology Letters, Behavioral Ecology, Ecography, Ecology, J Ecol, Funct Ecol, Heredity, Evolution...)

# Onde compartilhar os seus dados?

- Fora do Brasil, além das revistas, os órgãos de fomento pedem (1) o compartilhamento dos dados e (2) um plano para armazenamento e disponibilização destes dados em longo prazo.



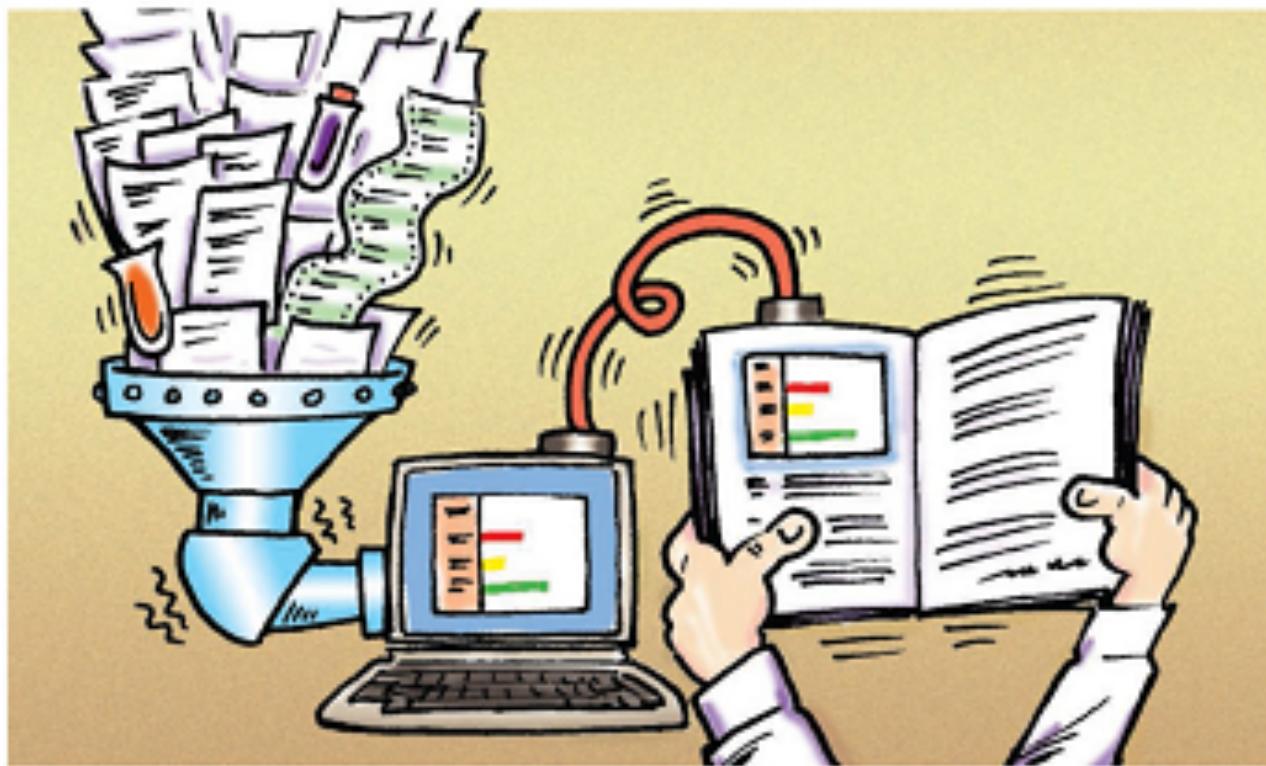
# O compartilhamento de dados coloca a Ecologia na Era da Informação e avança a síntese do conhecimento



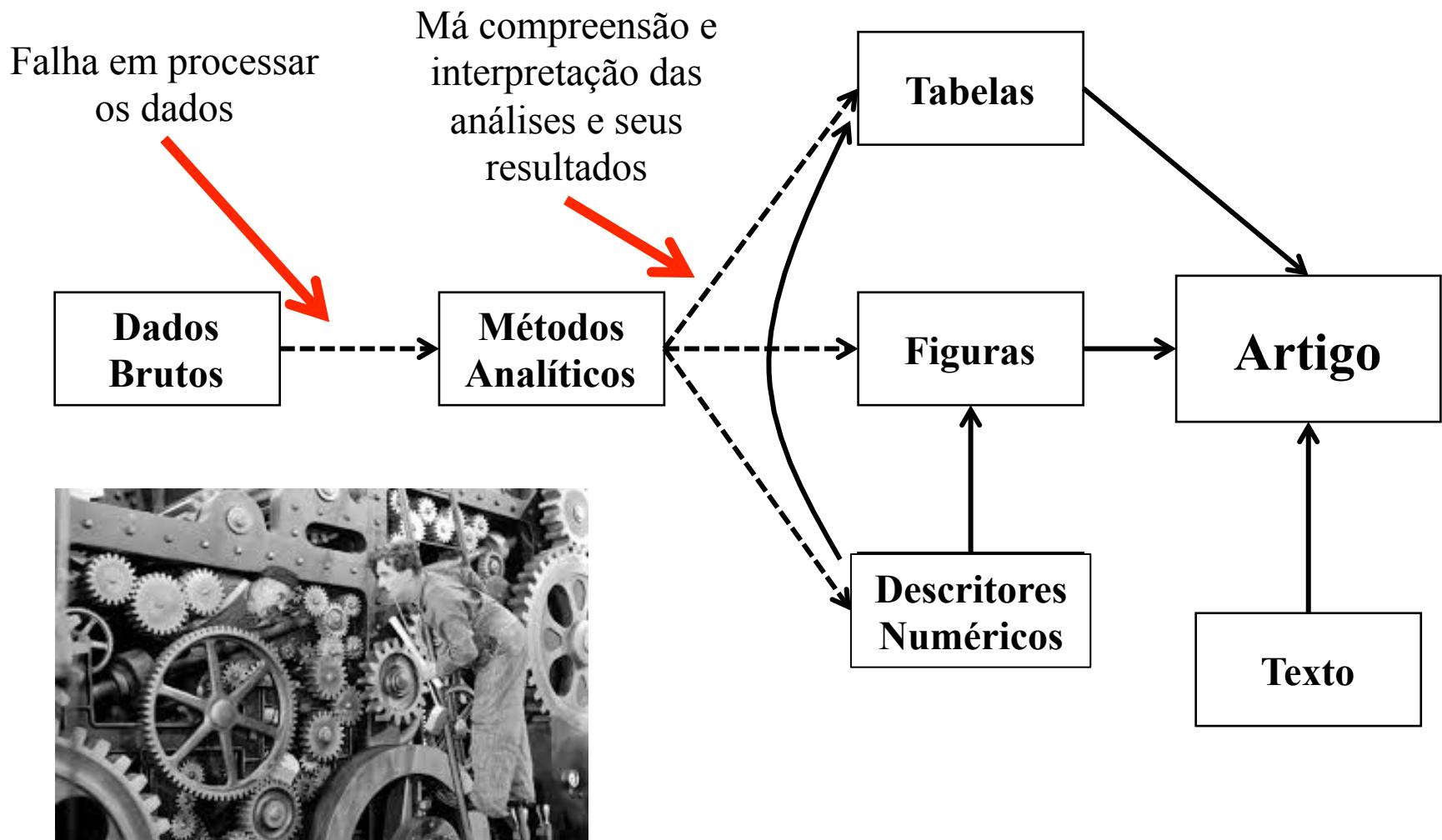
## Pesquisa intensiva de dados (Data-intensive research)

- Grande parte das informações que precisamos estão disponíveis *online*;
- Muitas técnicas e ferramentas para a análise de dados estão disponíveis;
- Grandes avanços podem vir da integração destes enormes volumes de dados.

**A velocidade da produção de dados é muito maior do que a nossa capacidade de processá-los**



# O processo de produção de um artigo científico tornou-se uma linha de produção – e com muitas falhas!



# Algumas das consequências destas falhas

12 MAY 2016 | VOL 533 | NATURE | 147

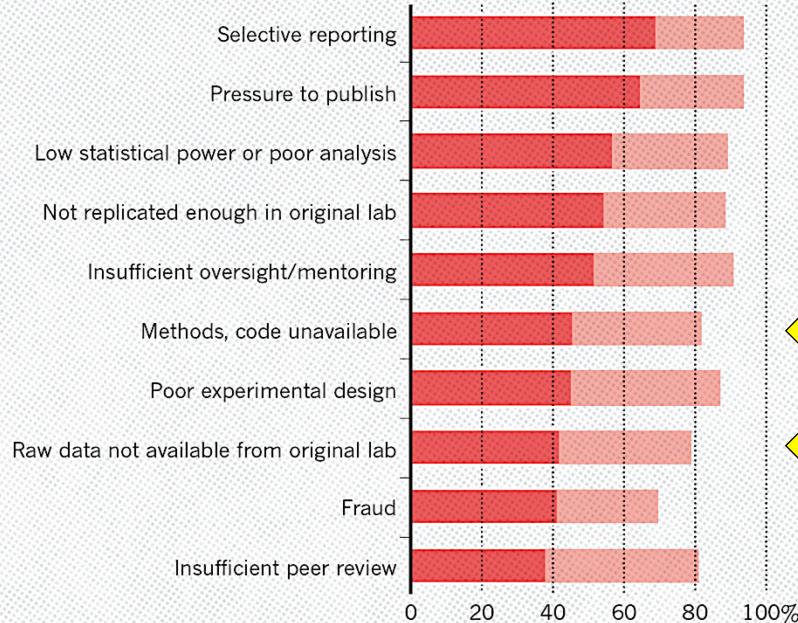
## The pressure to publish pushes down quality

Scientists must publish less, says Daniel Sarewitz, or good research will be swamped by the ever-increasing volume of poor work.

### WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

- Always/often contribute
- Sometimes contribute



## IS THERE A REPRODUCIBILITY CRISIS?

A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.

BY MONYA BAKER

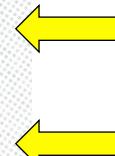
52%  
Yes, a significant crisis

1,576  
RESEARCHERS SURVEYED

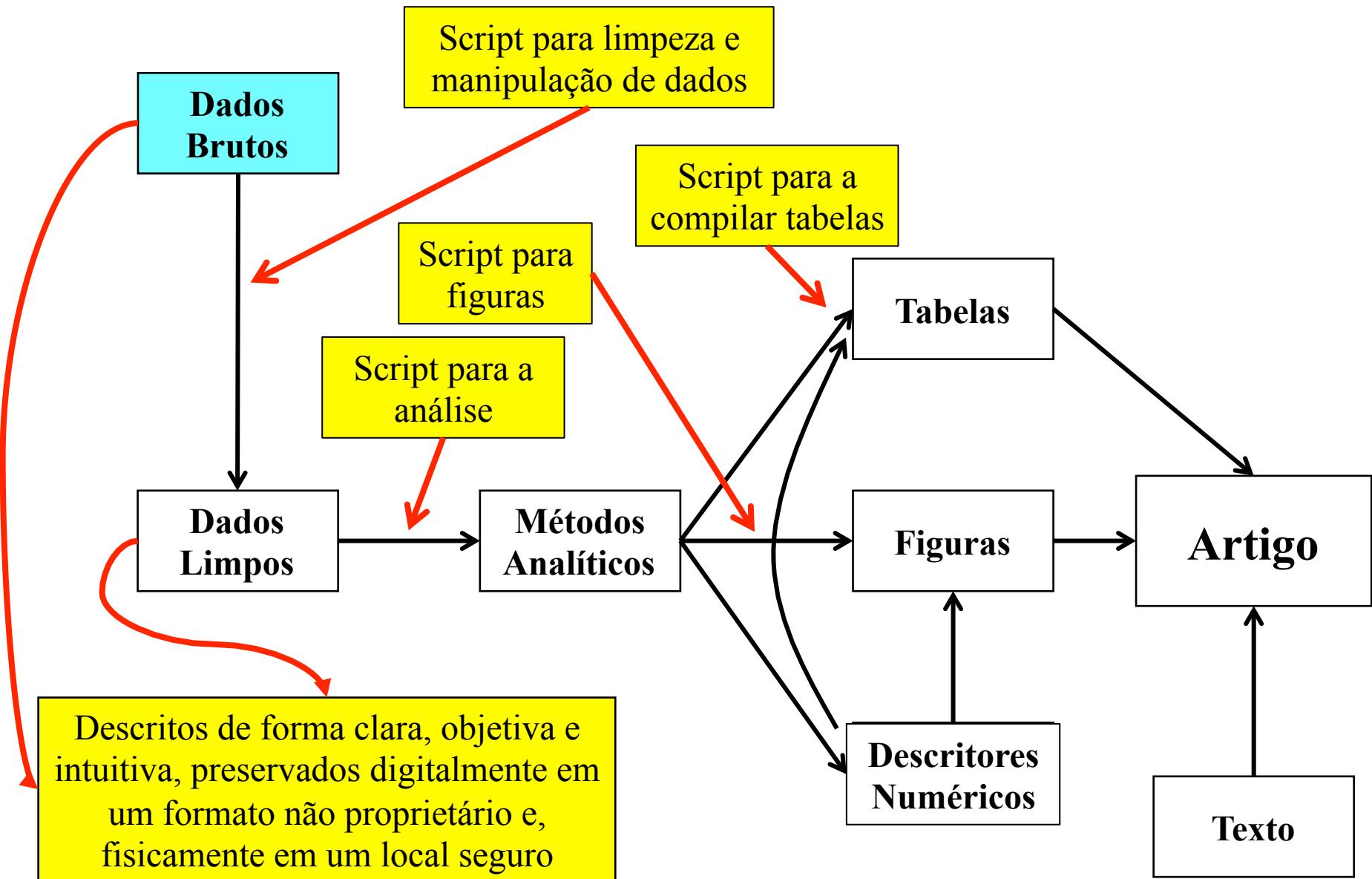
38%  
Yes, a slight crisis

3%  
No, there is no crisis

7%  
Don't know



# Como este curso se encaixa neste contexto?



**Tudo parece um prego quando temos um martelo novo**

## The Law of the Relational Database

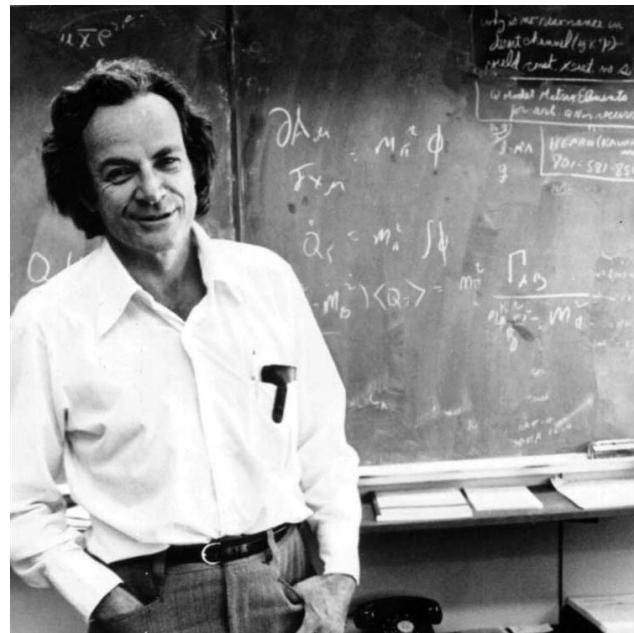


By HikingArtist.com

If the only tool you have is a relational database,  
everything looks like a table.

**“Para todo o homem é dada uma chave para os portões do céu.  
A mesma chave abre os portões do inferno.  
E assim é também com a ciência.”**

(Richard P. Feynman)



- Ter a chave mas não saber como usá-la pode fazer um grande estrago;
- Mas saber como usar a chave e não tê-la é inútil.
- É preciso ter cuidado e balancear esse trade-off.

**Acredito que a mesma lógica vale ao aprender os métodos quantitativos e a linguagem de programação R.**

Algumas dicas para o gerenciamento  
do seu *script* (baseado em experiências  
reais)

# Use a tecnologia ao seu favor, mas não se torne escravo dela

- Não perca tempo utilizando uma linguagem de programação para fazer uma tarefa simples quando você poderia ter feito a mesma coisa e em menos tempo usando outra ferramenta.
- Não perca tempo aprendendo a fazer algo em uma linguagem de programação que só vai te ajudar uma única vez – a não ser que seja um hobbie.



**Minimize a sua intervenção no código cada vez que você  
for fazer alguma coisa**



**NUNCA USE file.choose() PARA LER UM ARQUIVO NO R**

```
fitoplancton <- read.table("processed data/fitoplancton.txt", header = TRUE)
zooplankton <- read.table("processed data/zooplankton.txt", header = TRUE)
bentos <- read.table("processed data/bentos.txt", header = TRUE)
abioticos <- read.table("processed data/abiotic.txt", header = TRUE)
species <- readxl::read_excel("raw data/species traits.xlsx")
```

# Automatize tudo o que você puder, mas sempre confira o que o seu código está fazendo e gerando

- Utilize a linguagem de programação para realizar toda e qualquer manipulação, limpeza e processamento dos dados brutos, ao invés de fazer isso manualmente.
- Programe suas próprias funções quando existir a necessidade de fazer uma tarefa repetitiva (*e.g.*, um *loop* para executar um mesmo processamento em pedaços diferentes dos dados);
- Entretanto, sempre confira se os resultados de uma função ou se o produto final de uma sequência de processamentos são aqueles que você espera – antes de fazer a linguagem de programação salvar estes *outputs* no seu computador.

# **Antes de começar a escrever um código, pense no que você quer e quais serão as tarefas envolvidas**

1. Carregar/simular dados;
2. Carregar pacotes que vou usar;
3. Criar um código para as categorias;
4. Transformar o formato da tabela;
5. Fazer um cálculo;
6. Desenhar um gráfico
7. Salvar o gráfico.

# Torne se código mais legível

- Quebre linhas quando houver muita coisa escrita e use o *indent* (Control + I, no Windows; Command + I, no MAC).

```
# this function calculates the dissimilarity between pairs of sites, but taking into account differ
# in species richness in each comparison
# in order to improve the estimate of the rarefied dissimilarity, this rarefaction should be run fo
# number of times, from which we take the average dissimilarity between pairs of communities after
# .reps runs
beta.rich <- function(.dados, .reps){
  reps <- .reps|
  #create an empty data frame where you are going to store each pairwise dissimilarity value for ea
  jac0 <- as.data.frame(matrix(0, ncol=nrow(.dados), nrow=nrow(.dados)), row.names = row.names(.dad
  #create an empty data frame where you are going to sum up the values for each pairwise comparisor
  jac1 <- as.data.frame(matrix(0, ncol=nrow(.dados), nrow=nrow(.dados)), row.names = row.names(.dad
  for(k in 1:reps) {
    for(j in 1:nrow(.dados)) {
      for(i in j:nrow(.dados)) {
        #for sites that match their species richness, there is no need of rarefaction in the comparis
        if (sum(.dados[i,]) == sum(.dados[j,])) {
          r1 <- as.numeric(vegdist(.dados[c(i,j),], method="jac"))
          jac0[i,j] <- r1
        } else if (sum(.dados[i,]) > sum(.dados[j,])) { #if one sites is richer than the other, app
```



*indent*

# Torne se código mais legível

- Adicione comentários e explicações sobre o que você está fazendo em cada linha de comando ou bloco do script – seu eu do futuro agradece!

```
# this function calculates the dissimilarity between pairs of sites, but taking into account differ
# in species richness in each comparison
# in order to improve the estimate of the rarefied dissimilarity, this rarefaction should be run fo
# number of times, from which we take the average dissimilarity between pairs of communities after
# .reps runs
beta.rich <- function(.dados, .reps){
  reps <- .reps
  #create an empty data frame where you are going to store each pairwise dissimilarity value for ea
  jac0 <- as.data.frame(matrix(0, ncol=nrow(.dados), nrow=nrow(.dados)), row.names = row.names(.dad
  #create an empty data frame where you are going to sum up the values for each pairwise comparisor
  jac1 <- as.data.frame(matrix(0, ncol=nrow(.dados), nrow=nrow(.dados)), row.names = row.names(.dad
  for(k in 1:reps) {
    for(j in 1:nrow(.dados)) {
      for(i in j:nrow(.dados)) {
        #for sites that match their species richness, there is no need of rarefaction in the comparis
        if (sum(.dados[i,]) == sum(.dados[j,])) {
          r1 <- as.numeric(vegdist(.dados[c(i,j),], method="jac"))
          jac0[i,j] <- r1
        } else if (sum(.dados[i,]) > sum(.dados[j,])) { #if one sites is richer than the other, app
```

# Torne seu código mais legível

- Organize seu código em seções! (Control + R, no Windows; Command + R, no MAC).

```
# limpando ambiente -----
rm(list=ls(all=TRUE))

# carregando pacotes -----
library(tidyverse)
library(readxl)

# carregando dados -----
## dados de biomassa
biomassa <- list(respiracao = read_excel(path = "data/raw data/Biomassa - Respiracao.xlsx", na = "NA"),
                  excrecao = read_excel(path = "data/raw data/Biomassa - Excrecao.xlsx", na = "NA")) %>%
  bind_rows(.id = "ensaio")
biomassa

## dados taxonomicos
dados <- list(excrecao = read_excel(path = "data/raw data/Excrecao.xlsx", na = "NA"),
              respiracao = read_excel(path = "data/raw data/Respiracao.xlsx", na = "NA")) %>%
  bind_rows(.id = "ensaio")
dados

# limpando ambiente ▾
```

# Um script, um objetivo

- Cada script deve fazer apenas uma única tarefa: limpar dados **ou** gerar os dados para análise **ou** rodar as análises de dados **ou** compilar dados para tabelas de resultados **ou** gerar figuras **ou** fazer download de arquivos,...



**Capitulo 1 da tese.RScript**



**Limpando dados brutos.RScript**



**Analises estatisticas.RScript**

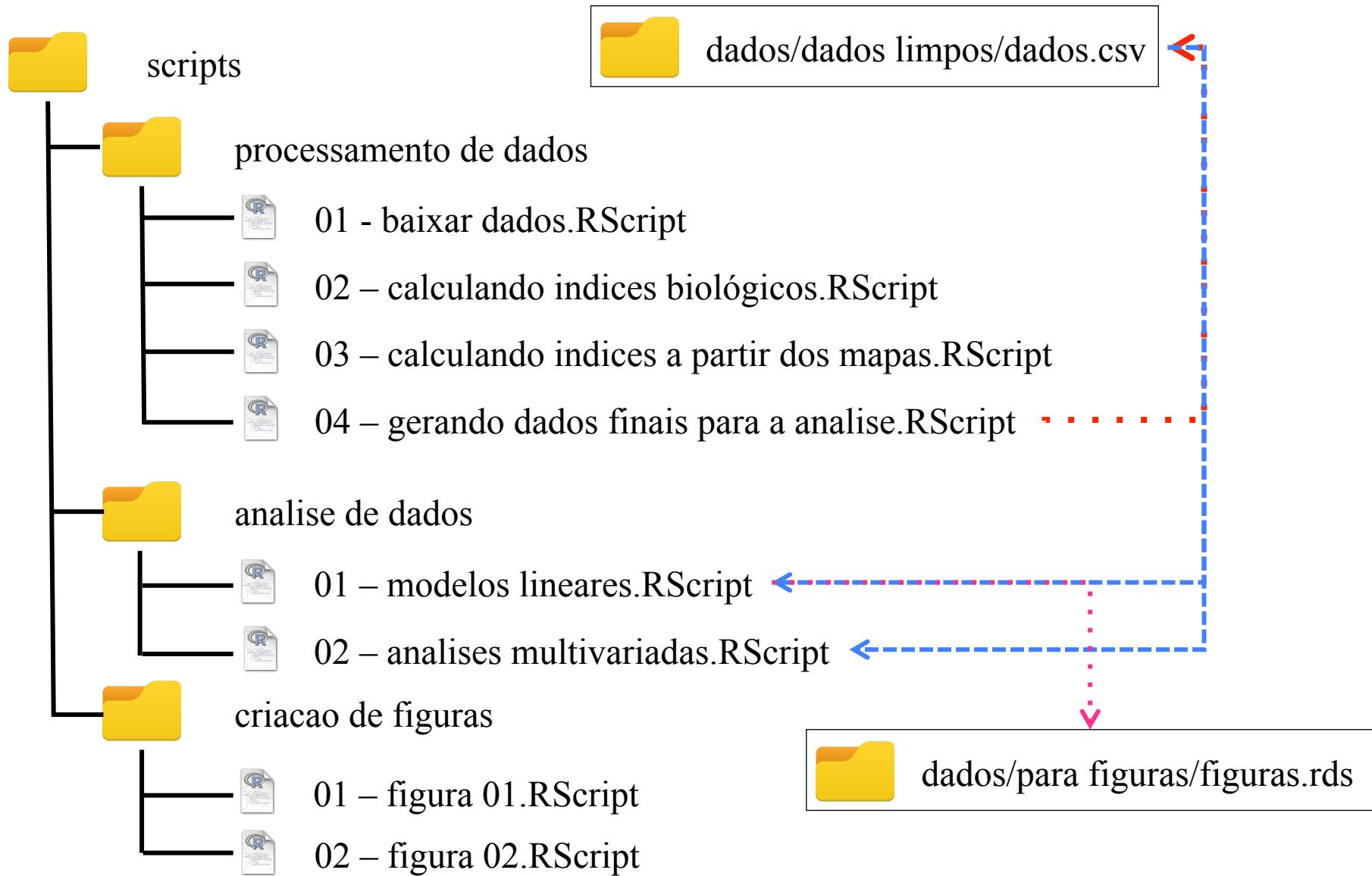


**Limpando tabelas de resultados brutos.RScript**

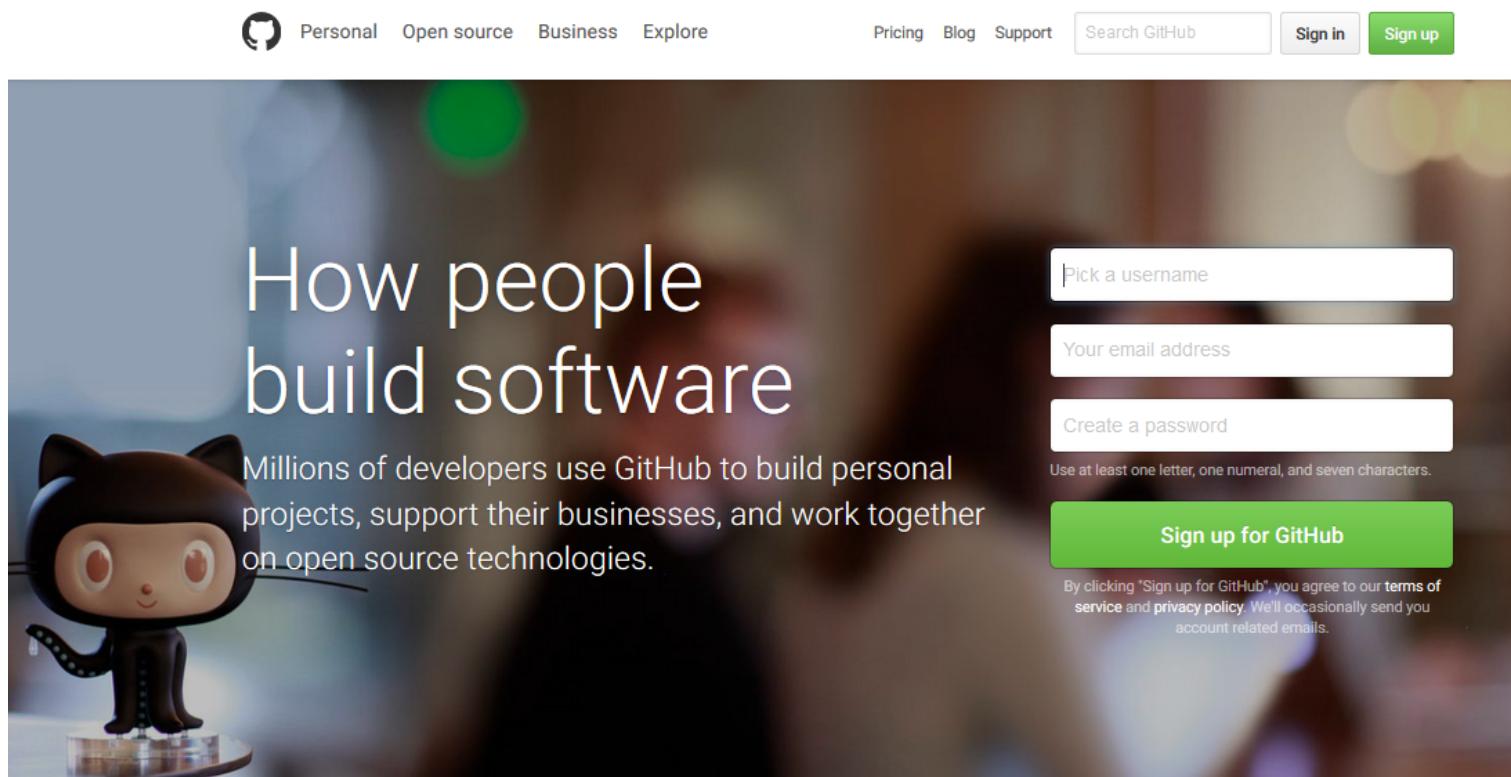


**Gerar figuras.RScript**

# Organize seus scripts no computador



# Utilize um sistema de controle de versão para gerenciar seus scripts, dados e manuscritos



The image shows the GitHub sign-up page. At the top, there is a navigation bar with links for Personal, Open source, Business, Explore, Pricing, Blog, and Support. There is also a search bar labeled "Search GitHub" and buttons for "Sign in" and "Sign up". The main heading on the page is "How people build software". Below it, there is a sub-headline: "Millions of developers use GitHub to build personal projects, support their businesses, and work together on open source technologies." To the left of the text, there is a small image of the GitHub cat figurine. On the right side, there are three input fields for "Pick a username", "Your email address", and "Create a password". Below the password field, a note says "Use at least one letter, one numeral, and seven characters." A large green "Sign up for GitHub" button is located at the bottom right. A small note below the button states: "By clicking 'Sign up for GitHub', you agree to our [terms of service](#) and [privacy policy](#). We'll occasionally send you account related emails."

# Resumindo

- Tão importante quanto seus resultados é como você chegou neles;
- O compartilhamento de informações (dados e código) é fundamental para o progresso científico;
- Você vai precisar compartilhar os dados do seu trabalho;
- Existem padrões e recomendações para que seu dados e seu código possam ser compartilhados de forma adequada, garantindo o uso futuro por outras pessoas e por você mesmo.
- Você precisa treinar seus colaboradores nessa cultura.