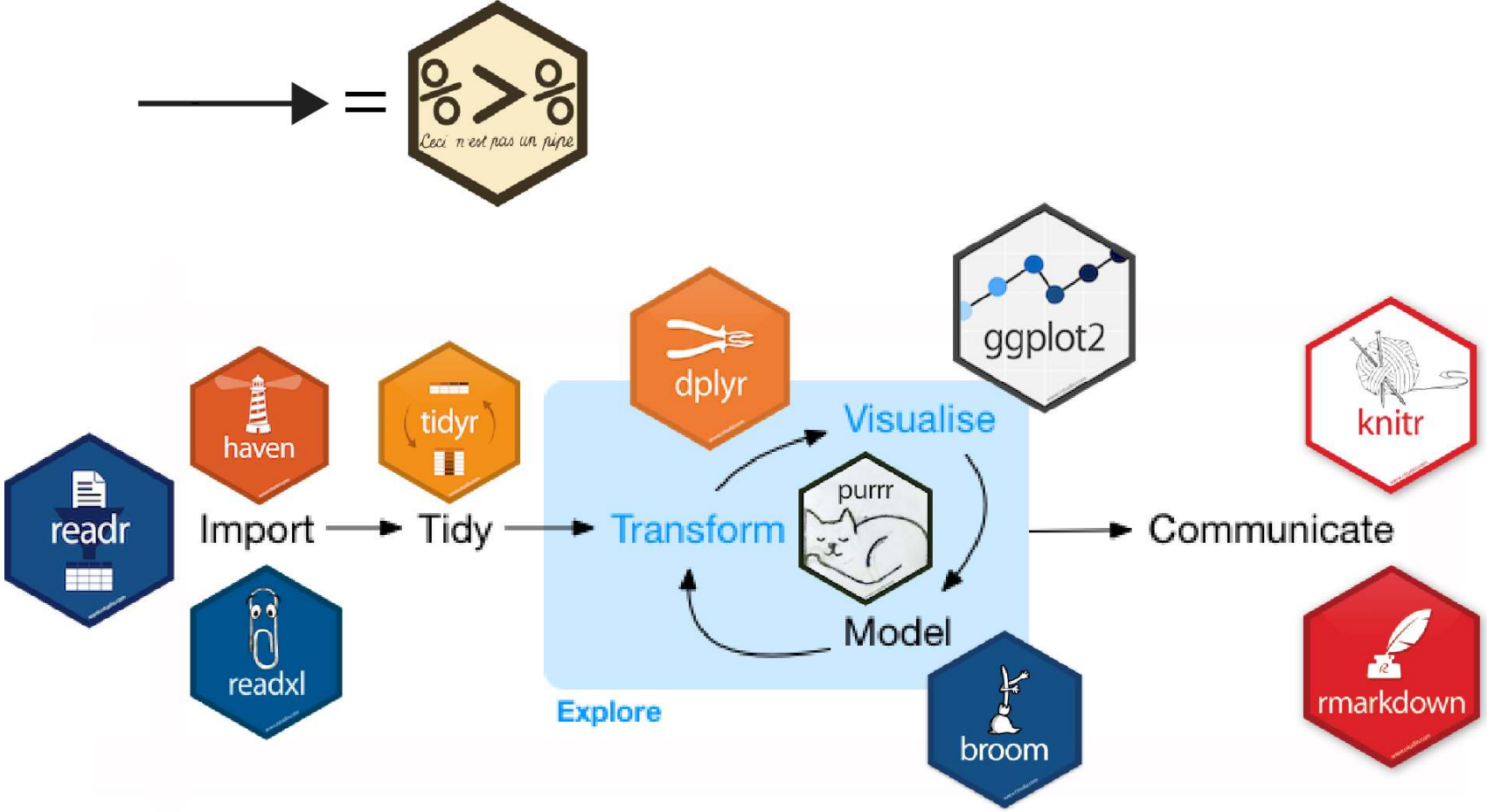
Aula 8
Introdução às ferramentas de reprodutibilidade em pesquisa



data.frames

factors

strings

https://www.tidyverse.org

Qual a importância de tornar o que se faz reprodutível?

- É o cerne da ciência! Tudo o que fazemos deve ser passível de ser produzido e assim confirmado... Ou não
- Lembra daquela história do seu eu passado não responder email?
- Torna o que você faz mais transparente e consequentemente mais pessoas vão ler e potencialmente citar ☺, causando maior impacto

*Se **eu** repetir o **seu experimento** exatamente como **você** descreveu, chegarei aos mesmos resultados?*

PLOS Medicine

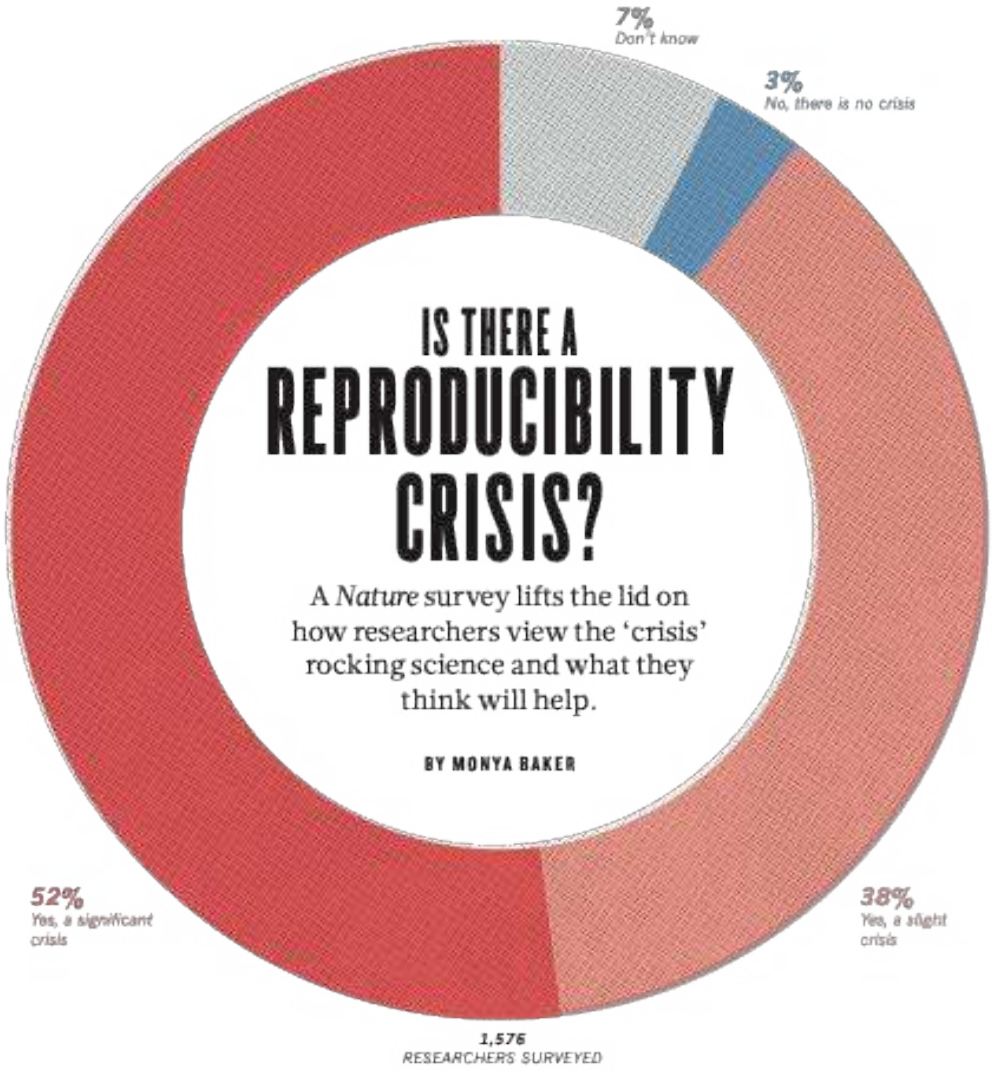
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

"The **probability that a research** claim **is true** may depend on **study power and bias**, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. (...) Simulations show that **for most study designs** and settings, it is **more likely for a research claim to be false than true**."



70% dos pesquisadores tentaram e falharam em reproduzir o experimento de outro cientista

Behavioral Ecology (2017), 28(2), 348–354. doi:10.1093/beheco/arx003

Invited Ideas

Striving for transparent and credible research: practical guidelines for behavioral ecologists

Malika Ihle,[a] Isabel S. Winney,[a,b,c] Anna Krystalli,[a] Michael Croucher[d]

Trends in Ecology & Evolution

ECOLOGY LETTERS

Ecology Letters, (2016) **19**: 726–728 doi: 10.1111/ele.12610

EDITORIAL **Promoting transparency in evolutionary biology and ecology**

T. H. Parker,[1]* S. Nakagawa,[2] J. Gurevitch,[3] and IIEE (Improving Inference in Evolutionary Biology and Ecology) workshop participants†

Opinion

Transparency in Ecology and Evolution: Real Problems, Real Solutions

Timothy H. Parker,[1,*] Wolfgang Forstmeier,[2] Julia Koricheva,[3] Fiona Fidler,[4] Jarrod D. Hadfield,[5] Yung En Chee,[4] Clint D. Kelly,[6] Jessica Gurevitch,[7] and Shinichi Nakagawa[8]

SORTEE

Society for Open, Reliable, and Transparent Ecology and Evolutionary Biology

https://linktr.ee/sortecoevo

Framework for Open and Reproducible Research Training

FORRT

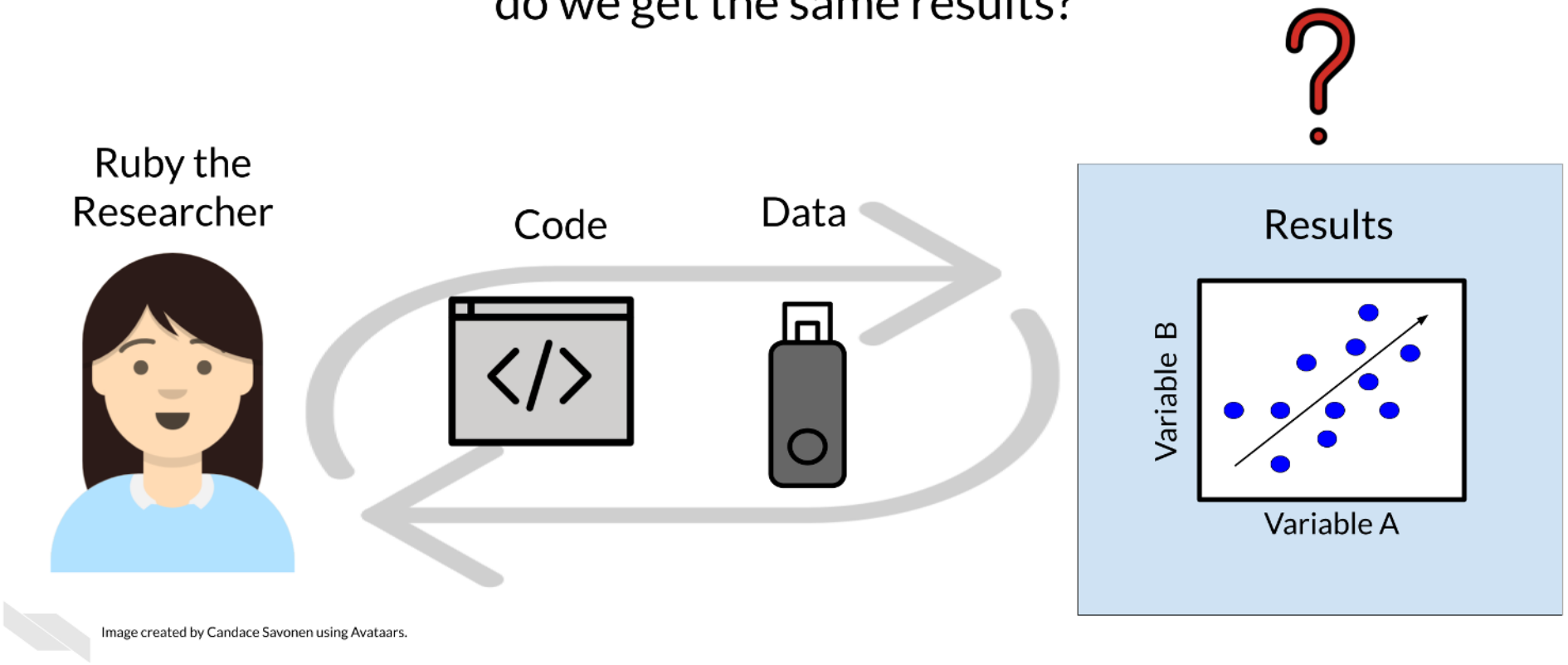
https://forrt.org

Conceitos Importantes

- **Reprodutibilidade:** Mesmos dados + Mesmo código = Mesmos resultados.

- **Replicabilidade:** Novos dados (novo experimento) + Mesmo método = Mesmos resultados (consistência).

Repeatable: keeping everything the same but repeating the analysis - do we get the same results?



https://jhudatascience.org/Adv_Reproducibility_in_Cancer_Informatics/defining-reproducibility.html

Reproducible: using the same data and analysis but in the hands of *another researcher* - do we get the same results?

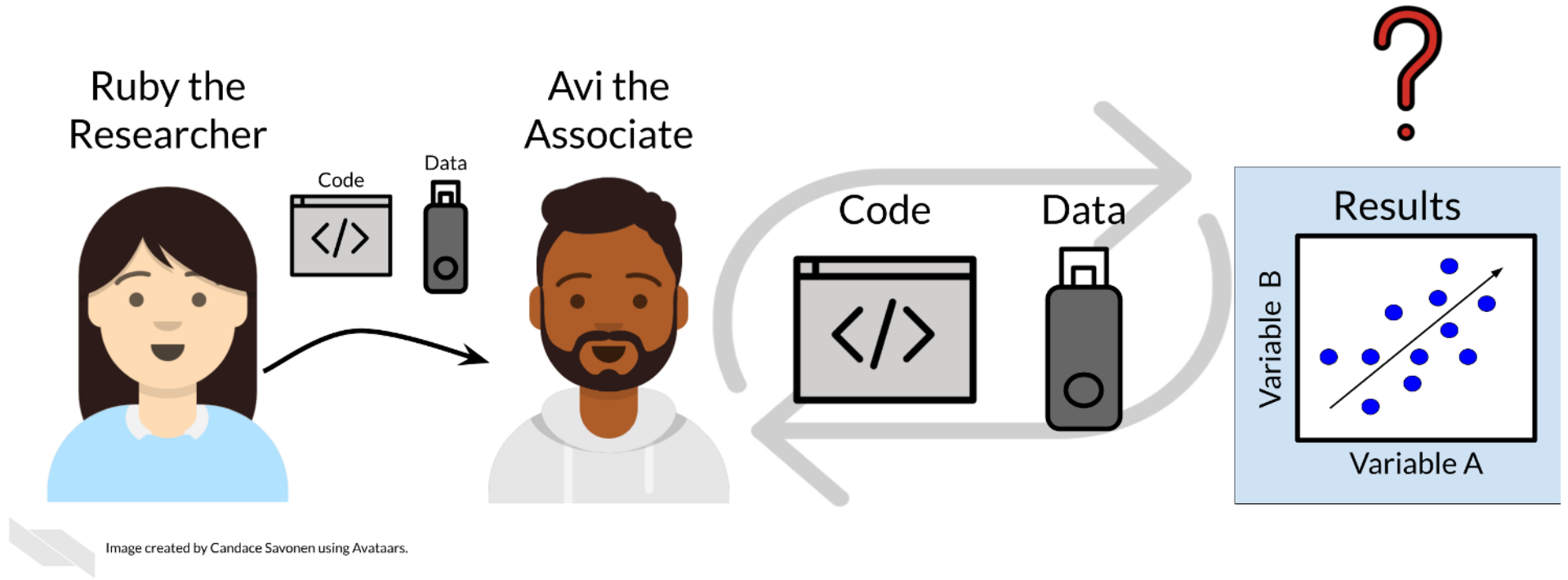


Image created by Candace Savonen using Avataars.

https://jhudatascience.org/Adv_Reproducibility_in_Cancer_Informatics/defining-reproducibility.html

Replicable: with new data do we obtain the same inferences?

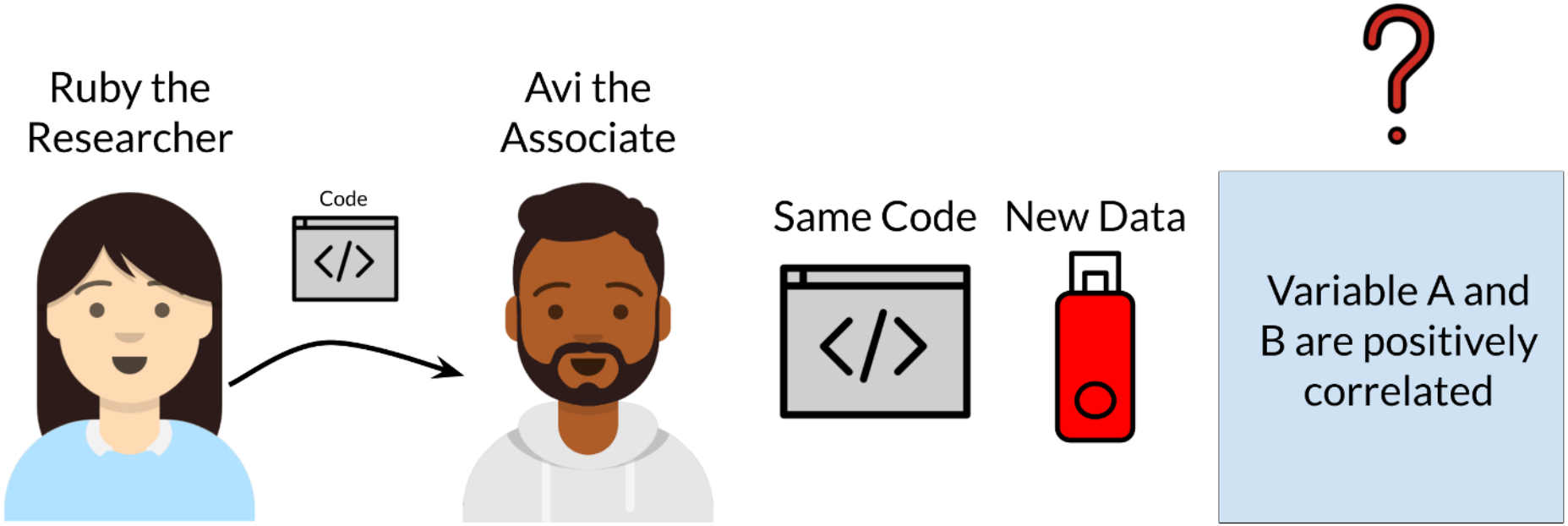
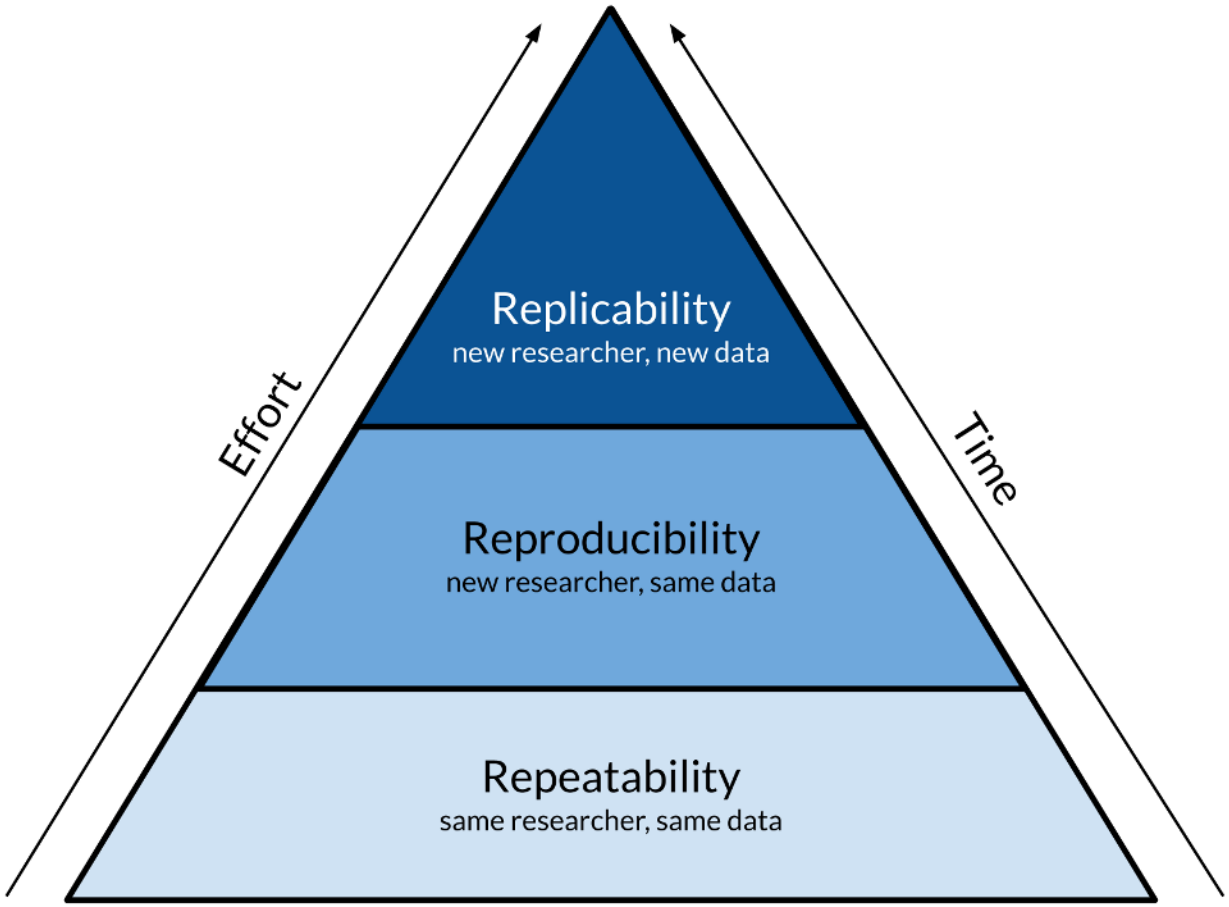


Image created by Candace Savonen using Avataars.

https://jhudatascience.org/Adv_Reproducibility_in_Cancer_Informatics/defining-reproducibility.html



Based off of a figure from Essawy et al, 2020 https://doi.org/10.1016/j.envsoft.2020.104753

https://jhudatascience.org/Adv_Reproducibility_in_Cancer_Informatics/defining-reproducibility.html

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng 2 DECEMBER 2011 VOL 334 SCIENCE

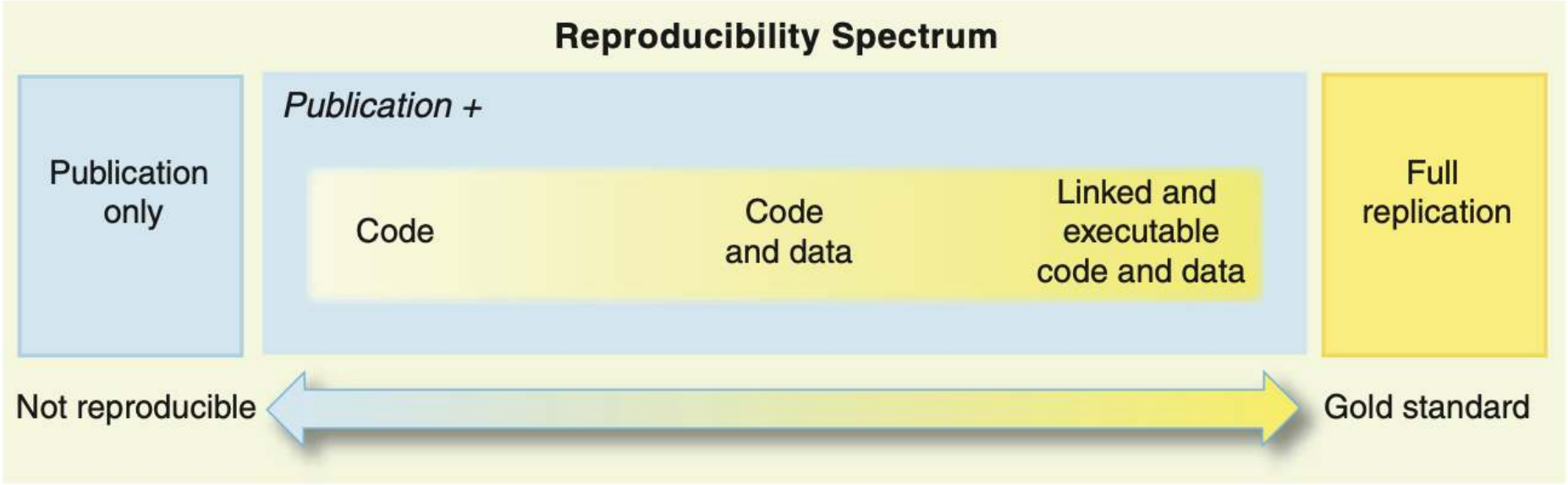


Fig. 1. The spectrum of reproducibility.

PLOS ONE

OPEN ACCESS

Citation: Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F (2018) Questionable research practices in ecology and evolution. PLoS ONE 13(7): e0200303. https://doi.org/10.1371/journal.pone.0200303

Editor: Jelte M. Wicherts, Tilburg University, NETHERLANDS

Received: March 21, 2018

Accepted: June 22, 2018

Published: July 16, 2018

RESEARCH ARTICLE

Questionable research practices in ecology and evolution

Hannah Fraser[1]*, Tim Parker[2], Shinichi Nakagawa[3], Ashley Barnett[1], Fiona Fidler[1,4]

- HARKing (Hypothesizing After Results are Known)
- Manipulação, fabricação, e seleção de dados
- P-hacking
  - Testar diversas variáveis/relações e reportar apenas a que deu "significante", ignorando as outras
- Más condutas em publicações
  - File drawer
- Reporte seletivo (dados, modelos, procedimentos de manuseio de dados etc)

O problema do valor de *P*

- O valor-p mede a incompatibilidade dos dados com a Hipótese Nula, **não** a probabilidade da hipótese ser verdadeira
- $P \leq 0{,}050$ não é um selo de verdade
  - Importância de calcular e reportar tamanho do efeito e tamanho amostral (graus de liberdade)

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

ASA Statement on Statistical Significance and *P*-Values

1. Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data**

- Os valores de p não medem a probabilidade de a hipótese estudada ser verdadeira, nem a probabilidade de os dados terem sido produzidos apenas por acaso.
- As conclusões científicas e as decisões empresariais ou políticas não devem basear-se apenas no fato de um valor p ultrapassar um determinado limiar.
- Por si só, o valor *P* não fornece uma boa medida de evidência em relação a um modelo ou hipótese.
- **A inferência adequada exige relatórios completos e transparência**

FORUM
FORUM
FORUM

FORUM is intended for new ideas or new ways of interpreting existing information. It provides a chance for suggesting hypotheses and for challenging current thinking on ecological issues. A lighter prose, designed to attract readers, will be permitted. Formal research reports, albeit short, will not be accepted, and all contributions should be concise with a relatively short list of references. A summary is not required.

The "file drawer problem" of non-significant results: does it apply to biological research?

Oikos 1996

Ryan D. Csada, 82 Straub Cres., Regina, SK, Canada S4T 656. – Paul C. James, Wildlife Branch, Saskatchewan Environment and Resource Management, 3211 Albert St., Regina, SK, Canada S4S 5W6. – Richard H. M. Espie,

OPINION
OPINION

Opinion is intended to facilitate communication between reader and author and reader and reader. Comments, viewpoints or suggestions arising from published papers are welcome. Discussion and debate about important issues in ecology, e.g. theory or terminology, may also be included. Contributions should be as precise and references should be kept to a minimum. A summary is not required.

Oikos 1997

Is there a "file drawer problem" in biological research?

Vincent Bauchau, Netherlands Inst. of Ecology, P.O. Box 40, NL-6666 ZG Heteren, The Netherlands

“Researcher’s degrees of freedom”

As decisões que tomamos durante a análise (remover este outlier? transformar log?) afetam o resultado da análise e conclusões biológicas

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

2014 American Scientist

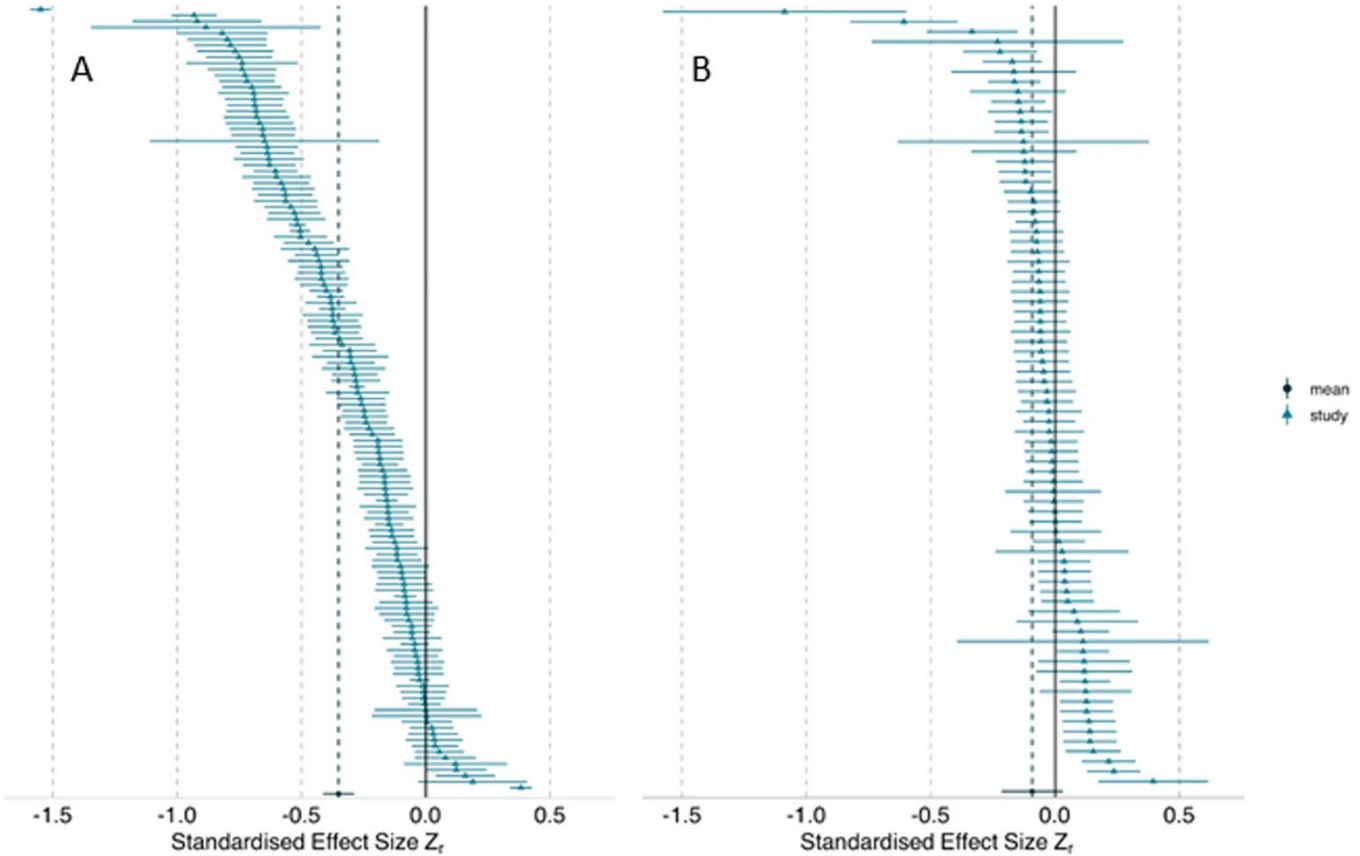
Andrew Gelman and Eric Loken

Gould et al. BMC Biology (2025) 23:35
https://doi.org/10.1186/s12915-024-02101-x

REGISTERED REPORT Open Access

Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology

Elliot Gould[1], Hannah S. Fraser[2], Timothy H. Parker[3*], Shinichi Nakagawa[4], Simon C. Griffith[5], Peter A. Vesk[1], Fiona Fidler[2], Daniel G. Hamilton[6], Robin N. Abbey-Lee[7], Jessica K. Abbott[8],



Conservation Biology

Timothy Parker [1], Hannah Fraser [2], and Shinichi Nakagawa [3]

Editorial

Making conservation science more reliable with preregistration and registered reports

Conservation Biology, Volume 33, No. 4, 747–750

DOI: 10.1111/cobi.13342

Received: 3 April 2023 | Accepted: 4 April 2023

DOI: 10.1002/ece3.10023

Ecology and Evolution Open Access WILEY

EDITORIAL

Registered Reports: A new chapter at *Ecology & Evolution*

nature ecology & evolution

Perspective

https://doi.org/10.1038/s41559-024-02433-5

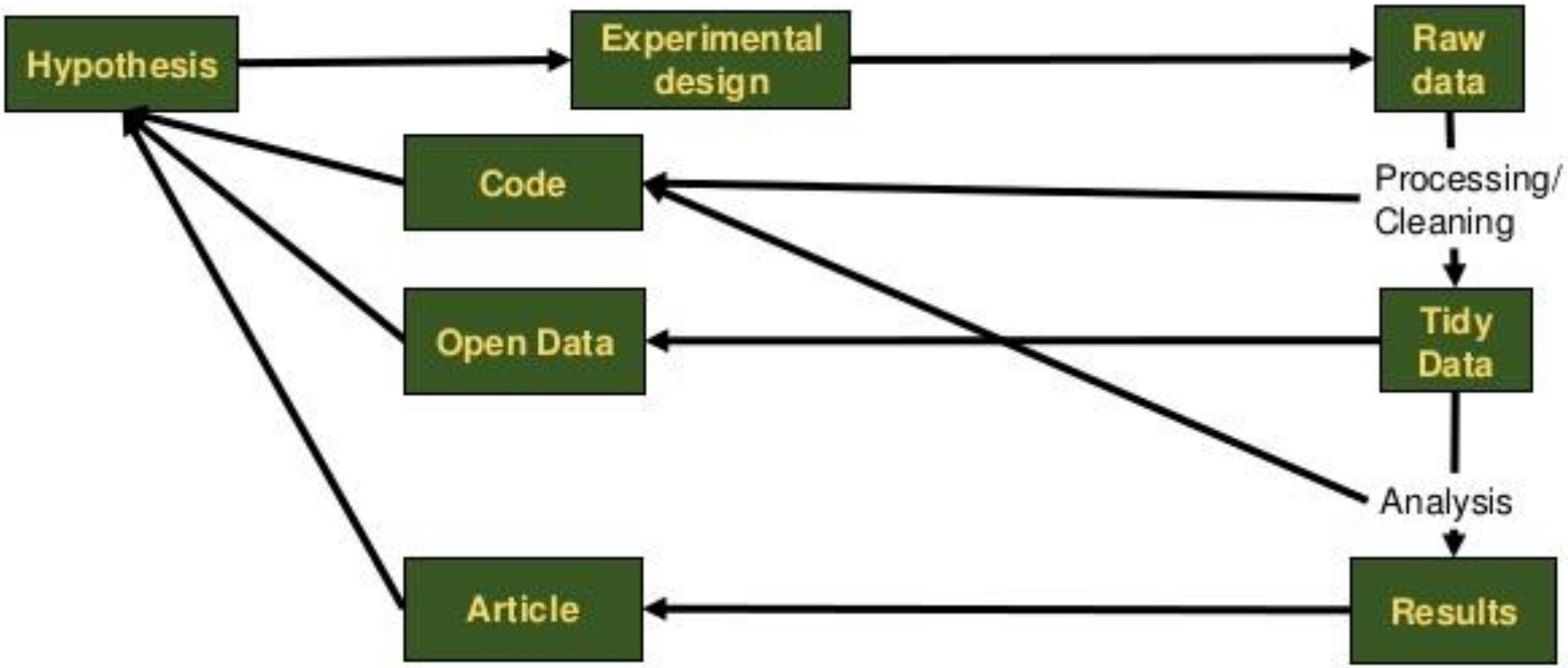
Supporting study registration to reduce research waste

Received: 5 October 2023

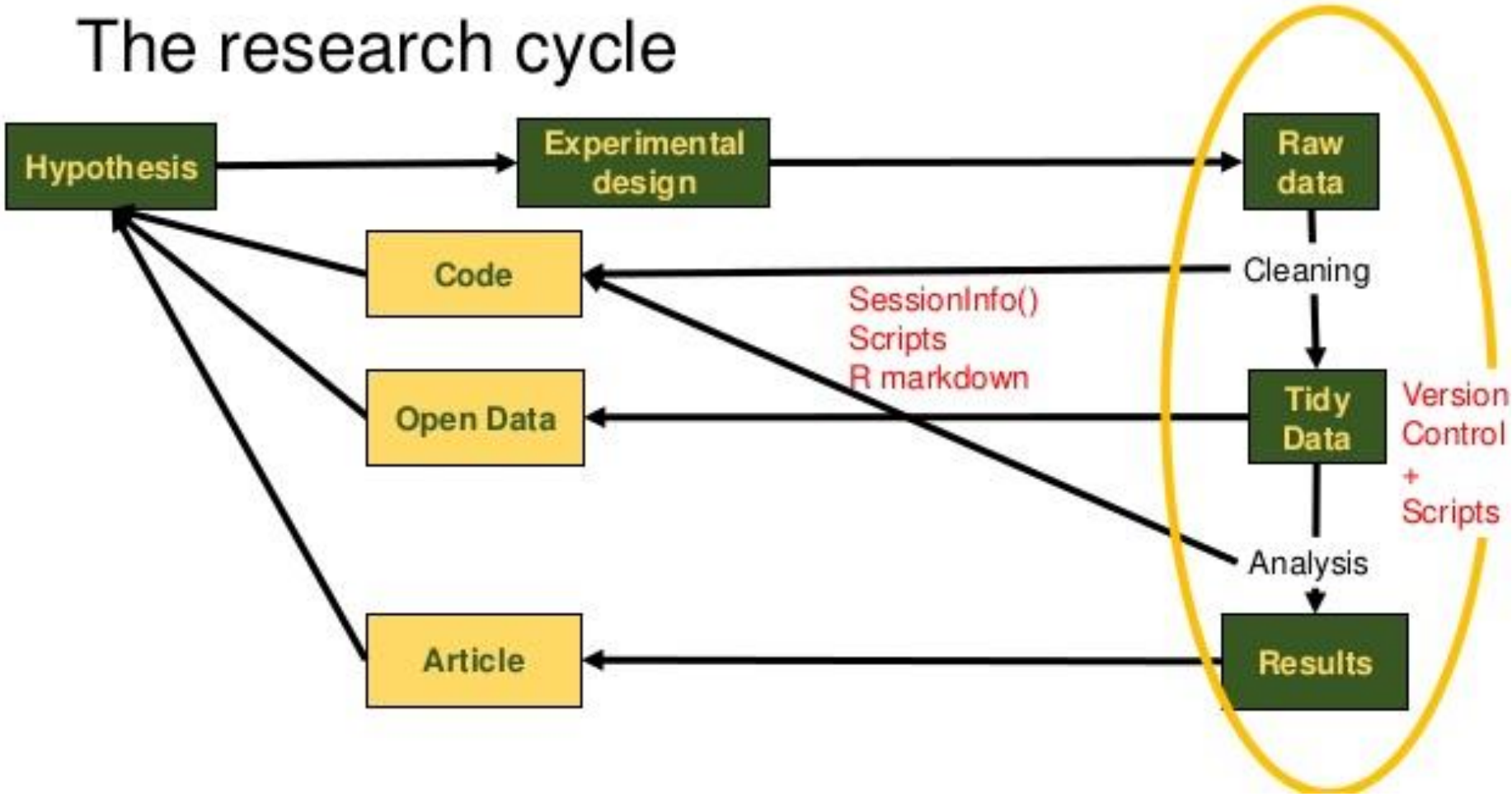
Accepted: 8 May 2024

Marija Purgar [1], Paul Glasziou [2], Tin Klanjscek [1], Shinichi Nakagawa [3,4] & Antica Culina [1,5]

The research cycle



The research cycle



Mas com o que devemos nos preocupar especificamente?


replaceAll(",", " ", a); a =
return a.split(" "); } $("#unique").
array_from_string($("#fin").
").val(), c = use_unique(array_
").val())); if (c < 2 * b - 1) {
), this.trigger("click"); }
!= a[b] && " " != a[b] ||
logged").val(); c = array
c.length;b++) { -1 != a.
; for (b = 0;b < c.length;
$("#User_logged").val(a)
.click(function

“Your best collaborator is yourself six months from now, and your past self doesn’t answer emails!”

Ética e Transparência em Pesquisa

- Diretrizes RSS e ASA
- **Dados Abertos (Open Data):** Depositar dados em repositórios (Dryad, Figshare).
- **Código Aberto:** Disponibilizar o script da análise.
- **Reportar incerteza:** Sempre reportar Intervalos de Confiança, não apenas valores-p.
- **Não dicotomizar:** Evitar frases "O efeito existe" apenas porque p=0.049 e "O efeito não existe" porque p=0.051.

PLOS | COMPUTATIONAL BIOLOGY

EDITORIAL

Ten Simple Rules for Digital Data Storage

Edmund M. Hart[1]*, Pauline Barmby[2], David LeBauer[3], François Michonneau[4,5], Sarah Mount[6], Patrick Mulrooney[7], Timothée Poisot[8], Kara H. Woo[9], Naupaka B. Zimmerman[10], Jeffrey W. Hollister[11]

Alguns pontos importantes

- Passo 3 ”Deixe os dados brutos, brutos”
- Passo 4 ”Armazene dados em formatos abertos”
  - Vai que a empresa dona do formato vai à falência

GitHub

protocols.io

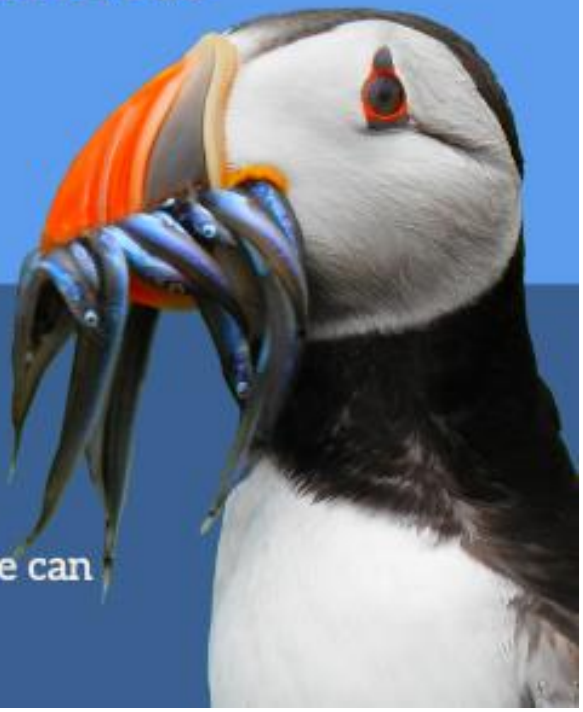
figshare

zenodo

BRITISH ECOLOGICAL SOCIETY

Better Science Guides

Data management



Through science we can

Alguns pontos importantes

- Passo 3 ”Deixe os dados brutos, brutos”
- Passo 4 ”Armazene dados em formatos abertos”
  - Vai que a empresa dona do formato vai à falência
- Passo 5 ”Dados devem ser tratados para análise”

A

Untidy Data

species	habitat	weight	length	latitude/longitude	date
Alligator mississippiensis	swamp	431 lb	4 ft 2	29.531,-82.184	Sept 15, 2015
Puma concolor	forest	125 lb	2.2m	29.125,-81.682	08/10/2015
Ursus americanus	forest	88 kg	133 cm	N29°7'30"/W81°40'55.2"	07-13-2015

B

Tidy Data

meta-data

data

species_code	date	station_code	weight_kg	length_cm
TSN 551771	2015-09-15	1	196	127
TSN 55247	2015-08-10	2	57	220
TSN 180544	2015-07-13	2	88	133

station_code	habitat	latitude	longitude
1	swamp	29.531	-82.184
2	forest	29.125	-81.682

species_code	class	genus	species
TSN 551771	Reptilia	Alligator	mississippiensis
TSN 55247	Mammalia	Puma	concolor
TSN 180544	Mammalia	Ursus	americanus

A Solução Tecnológica: Programação Literária

- **O problema do "Clique":** Análises feitas clicando em menus (Excel, SPSS) são difíceis de auditar. "Em que ordem você clicou?"
- **A solução do Código (R/Julia):** O script é a receita exata.
- **Ferramentas:** R Markdown ou Quarto (R) / Pluto.jl (Julia).
- **Vantagem:** Texto, Código e Resultados no mesmo documento. Se os dados mudam, o relatório atualiza sozinho.

Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Better Science Guides

Reproducible code

Through science we can

Como tornar o meu código reprodutível?

- 1a regra: comente!
- Dê nomes que façam sentido aos objetos
- Seja organizado! Faça com que o seu código adira a algum ”código de estilo”
- Teste o seu código antes de distribui-lo.
- Disponibilize o seu código em um repositório público
- Evite usar vários scripts para uma mesma análise

Check-list

- Fiz o desenho experimental antes de coletar os dados?
- Reportei todos os testes que fiz, ou só os que "deram certo"?
- Meu código roda no computador do colega?
- Estou confundindo significância estatística com relevância biológica?

Introdução ao
** Tidyverse **
Documentação com rmarkdown (quarto) + knitr

O que é e para que serve o markdown e knitr?

What is Markdown?

"Markdown is a text-to-HTML conversion tool for web writers. Markdown allows you to write using an easy-to-read, easy-to-write plain text format, then convert it to structurally valid XHTML (or HTML)."

- John Gruber, creator of Markdown

What is R Markdown?

- R markdown is the integration of R code with markdown
- Allows one to create documents containing "live" R code
- R code is evaluated as part of the processing of the markdown
- Results from R code are inserted into markdown document
- A core tool in **literate statistical programming**

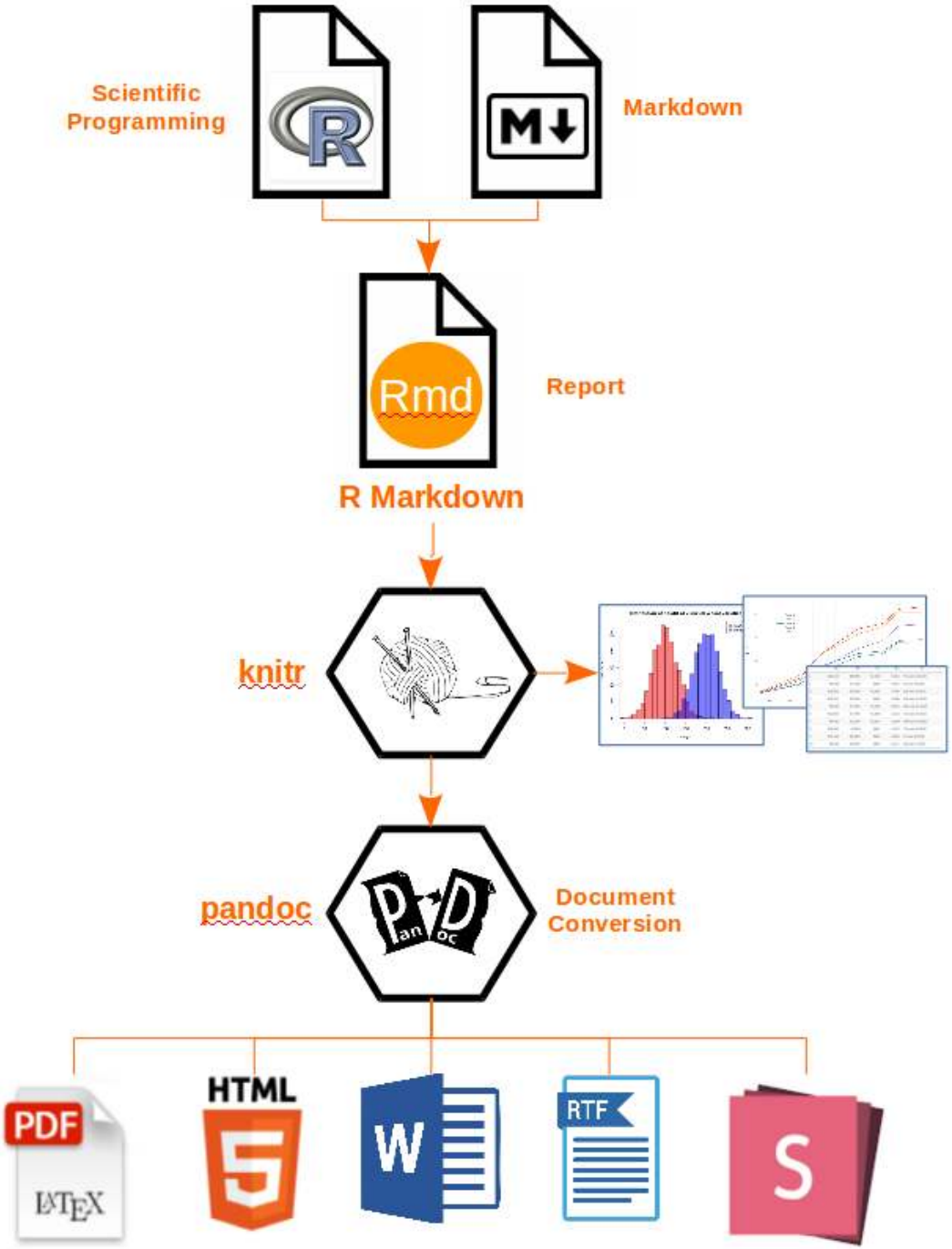
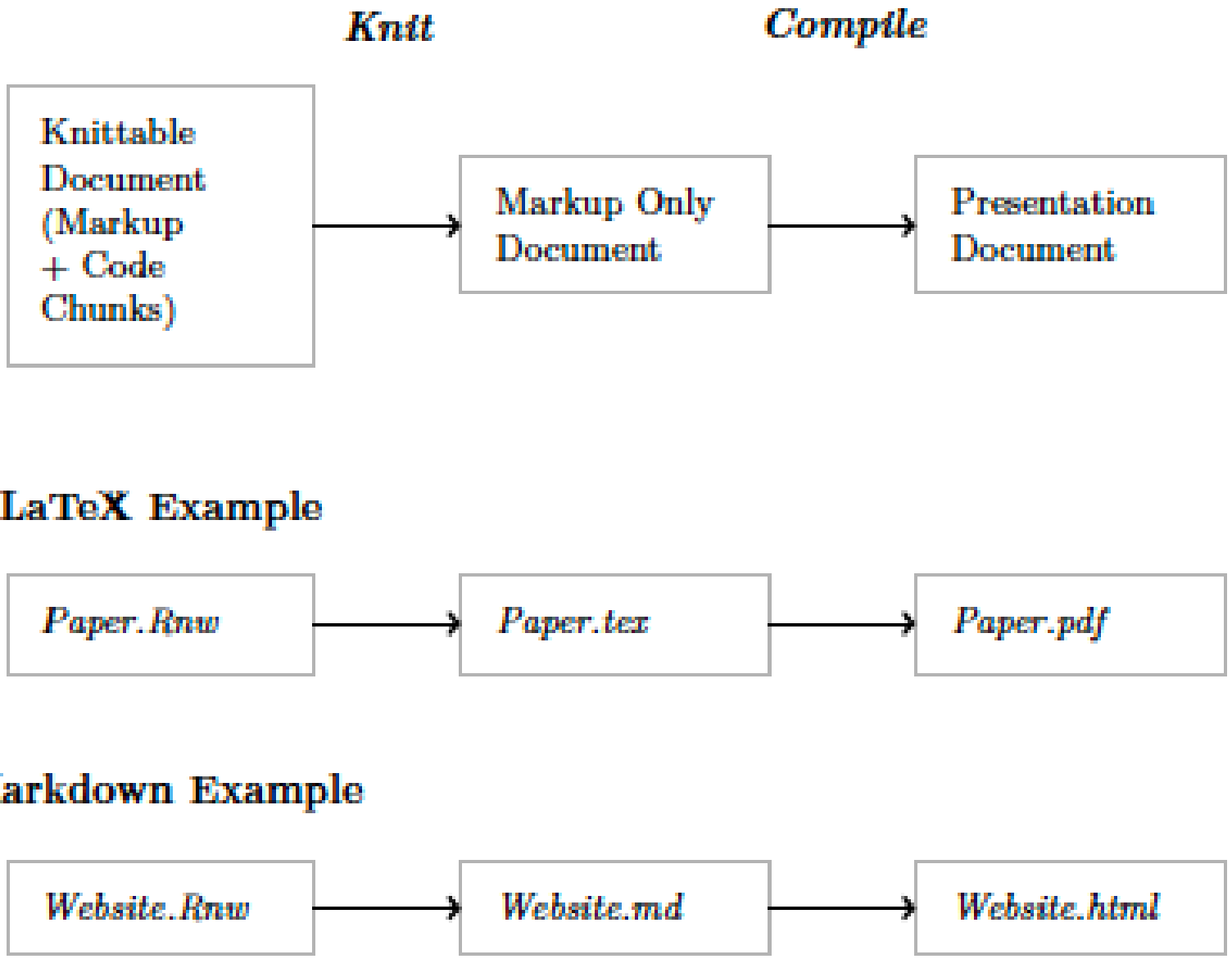

Scientific
Programming
Markdown
Rmd
Report
R Markdown
knitr
pandoc
Document
Conversion
PDF
LaTeX
HTML
RTF

FIGURE 3.4
The *knitr* Process



Gandrud 2013

RStudio

rmarkdown + knitr

bookdown

- Books
- Journal articles
- Long form documents

Slide presentations

- ioslides
- slidy
- Beamer
- PowerPoint
- reveal js

xaringan

Document + code

- HTML document
- PDF
- LaTeX
- Interactive notebooks
- Dashboards

HTML

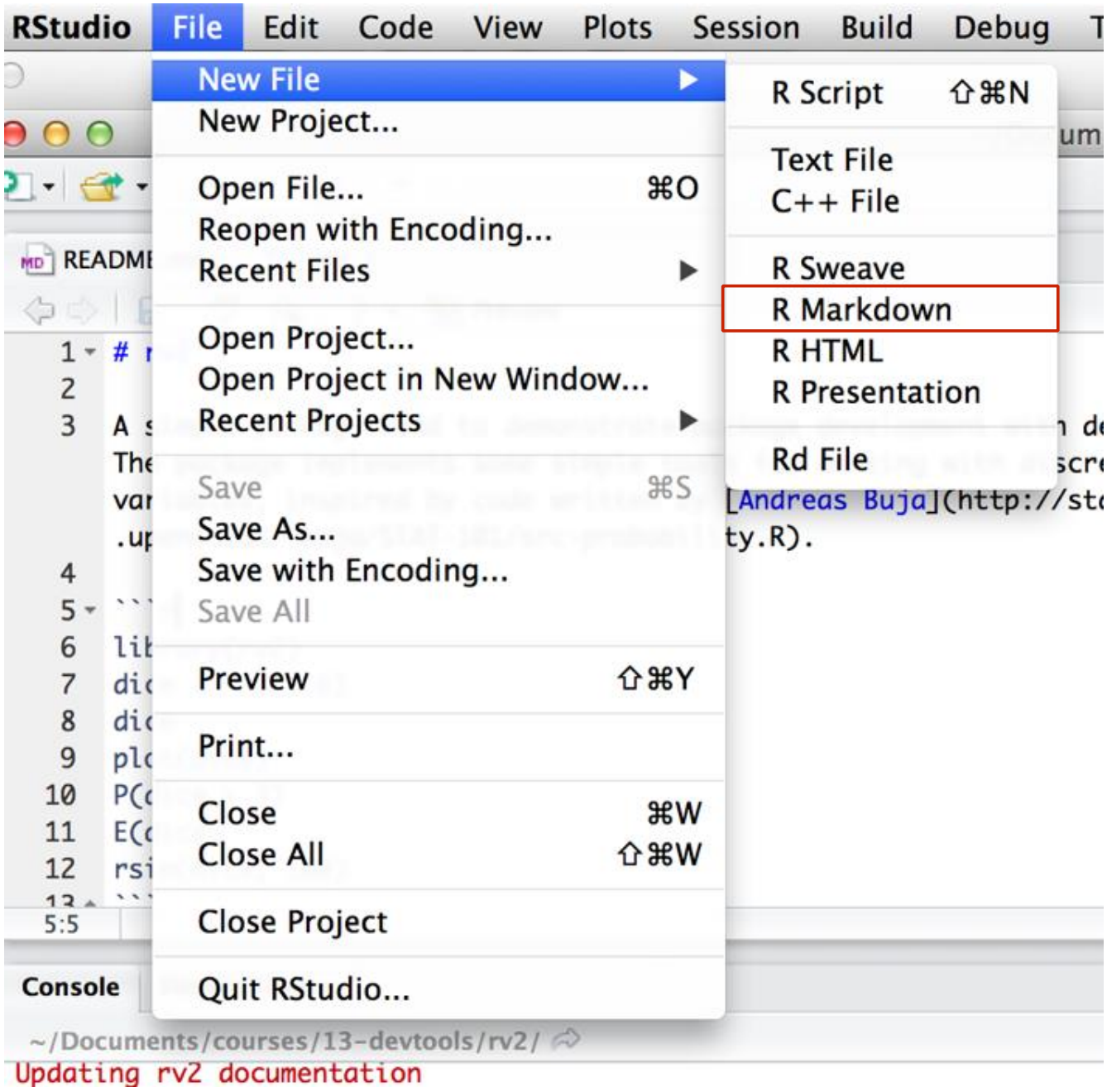
PDF

Adobe

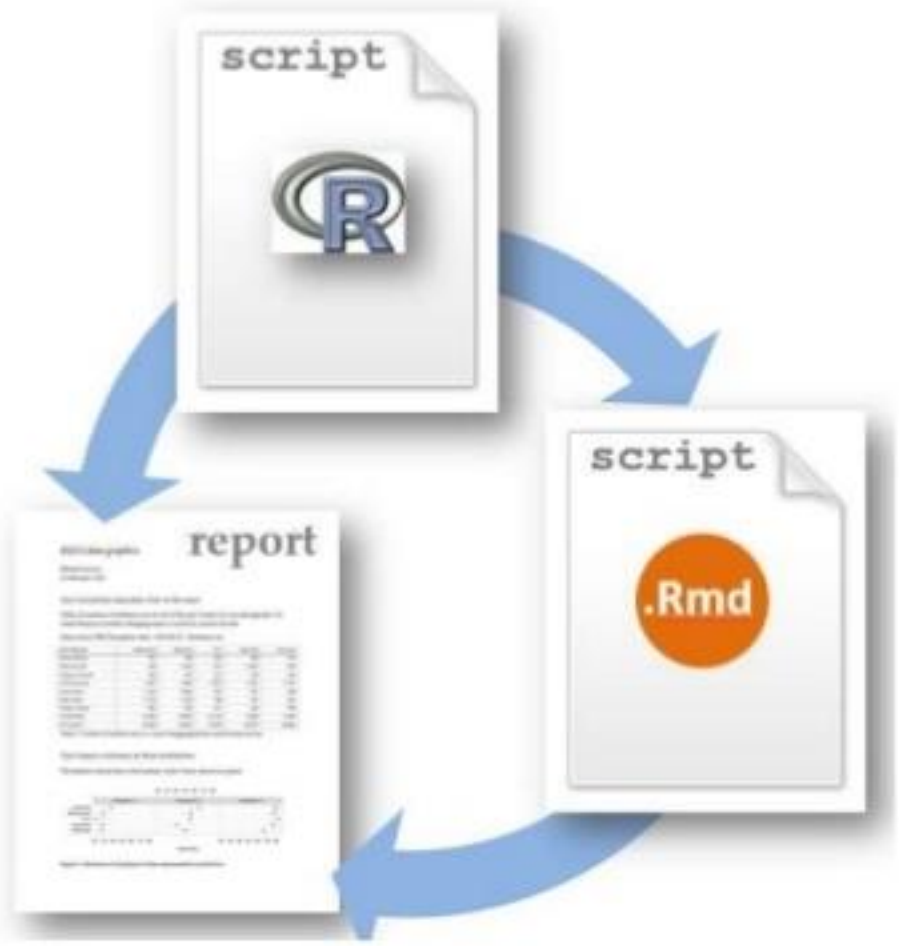
LaTeX

blogdown

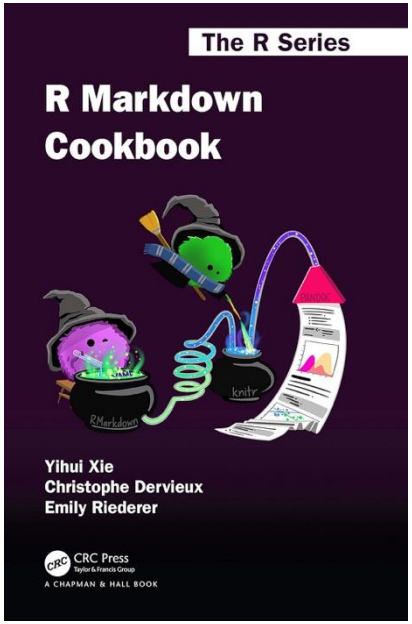
- Websites/blogs

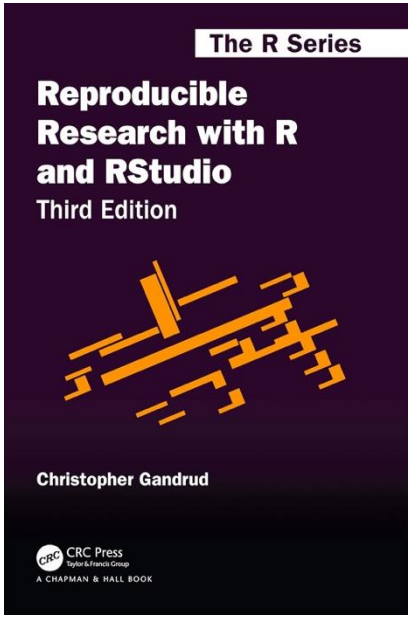

RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
New File
New Project...
Open File... ⌘O
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Window...
Recent Projects
Save ⌘S
Save As...
Save with Encoding...
Save All
Preview ⇧⌘Y
Print...
Close ⌘W
Close All ⇧⌘W
Close Project
Quit RStudio...
R Script ⇧⌘N
Text File
C++ File
R Sweave
R Markdown
R HTML
R Presentation
Rd File
README
5:5
Console
~/Documents/courses/13-devtools/rv2/
Updating rv2 documentation

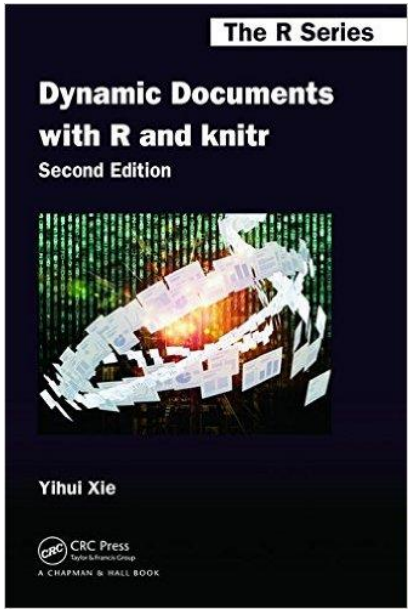
Passos para tornar seu fluxo de análise de dados reprodutível

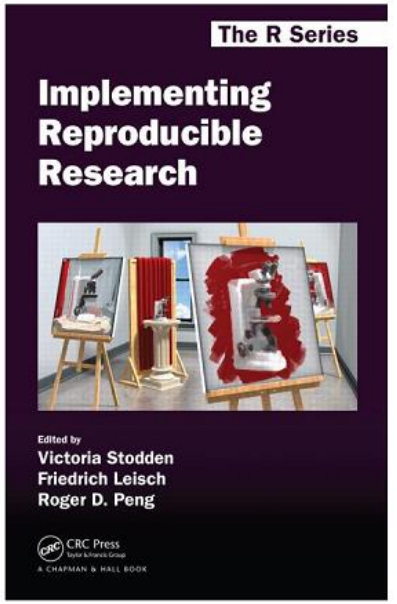


````
# This is a top level heading

This is some text. Make text _italic_ with single  underscores (or stars). Make it **bold** with double stars (or underscores). Here is a [link to a markdown guide](http://bit.ly/19fAexE).

* This is a list

* This is another item

```R
# Some R code
f <- function() x + 1
```

## This is a secondary heading

You can also do `inline code`, numbered lists and quotes and more.
````


The R Series
R Markdown Cookbook
Yihui Xie
Christophe Dervieux
Emily Riederer
CRC Press
A CHAPMAN & HALL BOOK


The R Series
Reproducible Research with R and RStudio
Third Edition
Christopher Gandrud
CRC Press
A CHAPMAN & HALL BOOK


The R Series
Dynamic Documents with R and knitr
Second Edition
Yihui Xie
CRC Press
A CHAPMAN & HALL BOOK


The R Series
Implementing Reproducible Research
Edited by
Victoria Stodden
Friedrich Leisch
Roger D. Peng
CRC Press
A CHAPMAN & HALL BOOK


YES!
AGORA BORA BOTAR A MÃO NA MASSA!

Exercícios

- Introdução ao Markdown, rMarkdown e knitr usando o RStudio
- Exercício 1
  - Passar alguns scripts que usamos para o formato markdown, explicando os comandos, exibindo os gráficos e por fim compilar esse documento pra formato .html

Exercícios

- Como utilizar LLMs para acelerar a geração de um arquivo R markdown/Quarto
- Exercício 2
  - Peça ao LLM: Crie um template de R Markdown que inclua uma seção de importação de dados, uma análise exploratória com gráficos e uma tabela formatada com os resultados de um modelo linear

… E fim de papo!

Obrigado! e por favor respondam o feedback, é importante!

diogo.provete@ufms.br | diogoprovete.weebly.com

Setor de Ecologia – INBIO