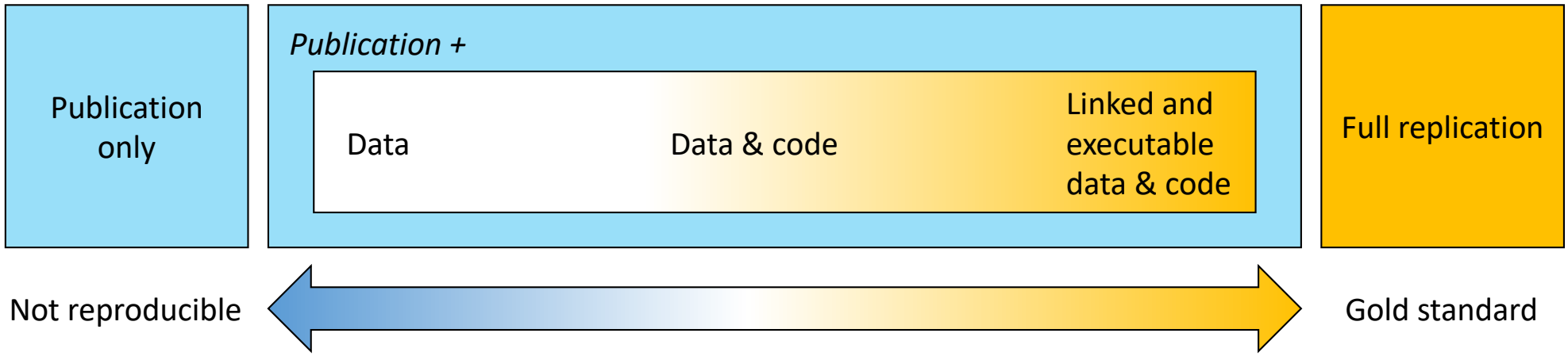
EEB 603 – Reproducible Science



BOISE STATE UNIVERSITY

INSTRUCTOR

- **Name:** Sven BUERKI (he/his)
- **Office:** Science Building (114)
- **Email:** svenbuerki@boisestate.edu
- **Office hours:** By appointment

Class meetings

- **Weekdays and times:** Tuesday Thursday 9-10:15 AM.
- **Location:** In-person in Science Bldg, Rm 149
- Encouraged to bring your personal computer.

TEACHING & RESEARCH ETHOS

- Everyone here is smart; distinguish yourself by being kind.

  Kindness in Science is an inclusive approach that fosters diversity, respect, wellbeing & openness leading to better science outcomes.

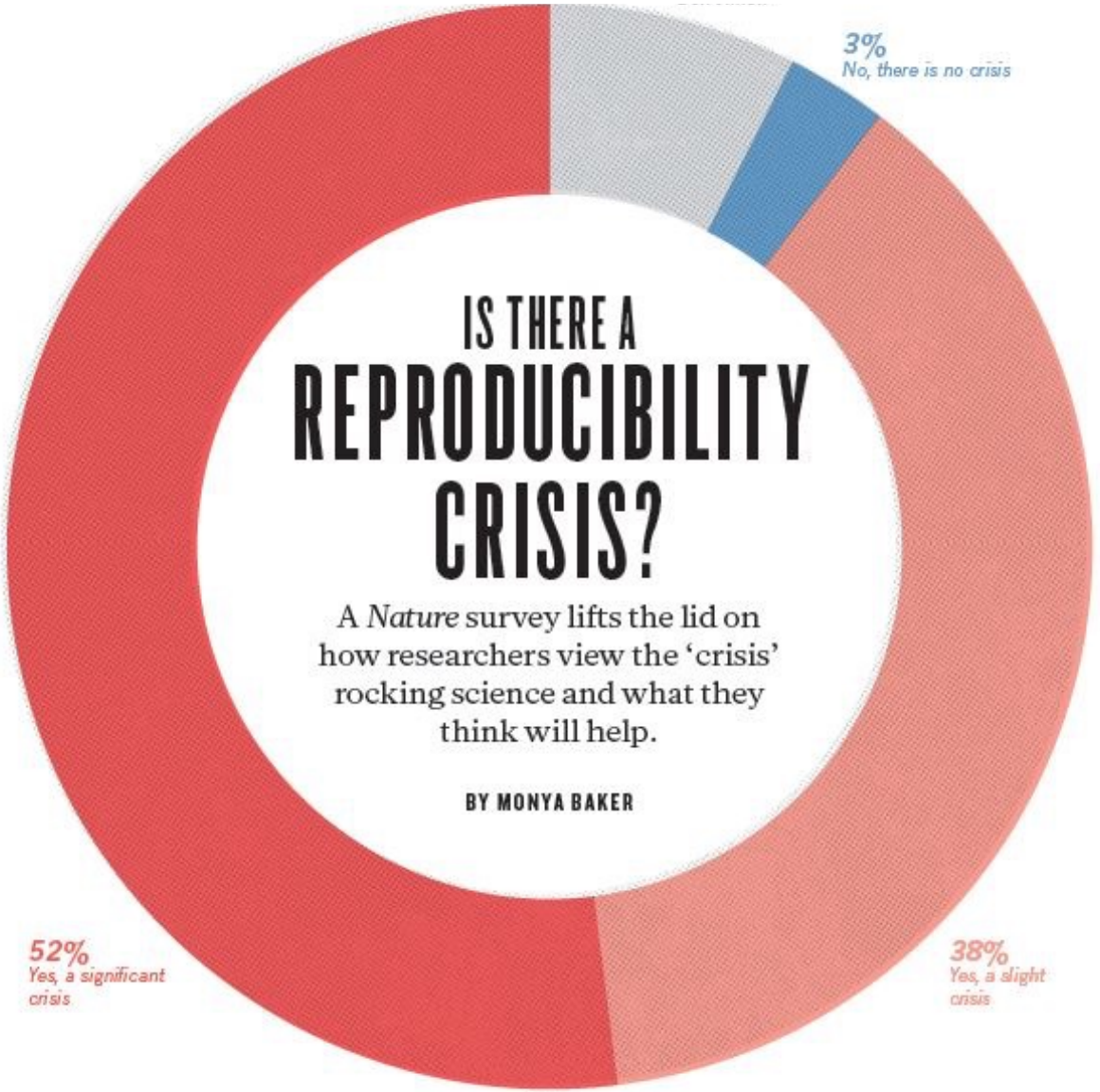
https://doi.org/10.1038/d41586-018-00482-y

Round of introductions

- Who am I?
- What program am I in?
- What do I want to take away from this course?

WHY THIS CLASS?

Based on a survey published in *Nature* (2016), 90% of the respondents said that **there is a reproducibility crisis in Science!**



https://doi.org/10.1038/533452a

Retraction Watch

Tracking retractions as a window into the scientific process

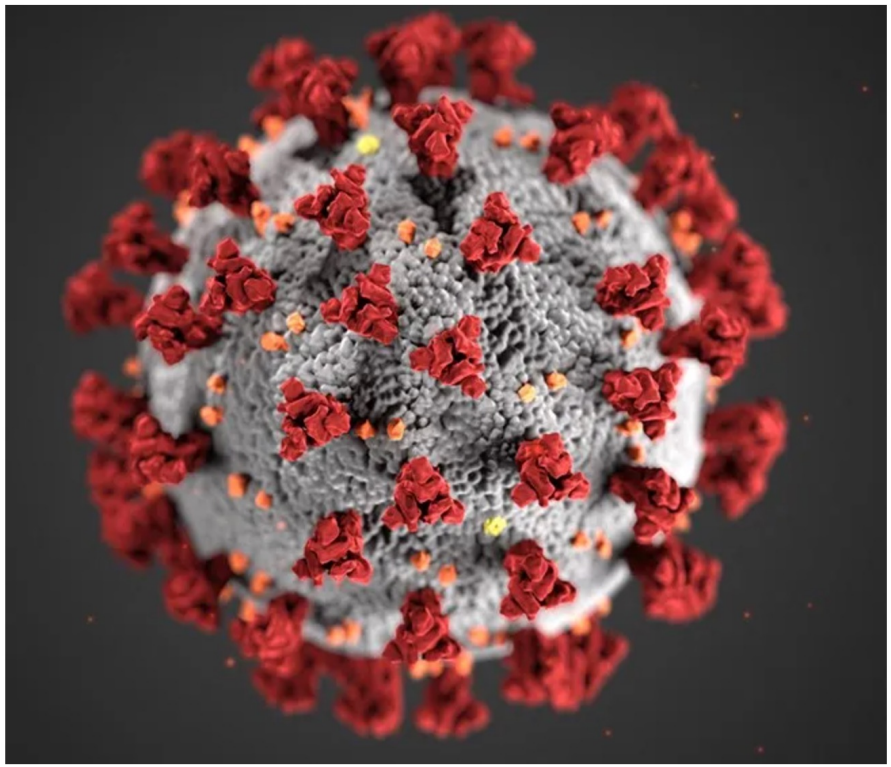
https://retractionwatch.com/

PAGES

- How you can support Retraction Watch
- Meet the Retraction Watch staff
  - About Adam Marcus
  - About Ivan Oransky
- Our Editorial Independence Policy
- Papers that cite Retraction Watch
- Privacy policy
- Retracted coronavirus (COVID-19) papers
- Retraction Watch Database User Guide
  - Retraction Watch Database User Guide Appendix A: Fields

Retracted coronavirus (COVID-19) papers

Fall 2023: 359 papers retracted



via CDC

REPRODUCING CODE IS ALSO AN ISSUE

Analysis | Open Access | Published: 21 February 2022

A large-scale study on research code quality and execution

Ana Trisovic ✉, Matthew K. Lau, Thomas Pasquier & Mercè Crosas

Scientific Data **9**, Article number: 60 (2022) | Cite this article

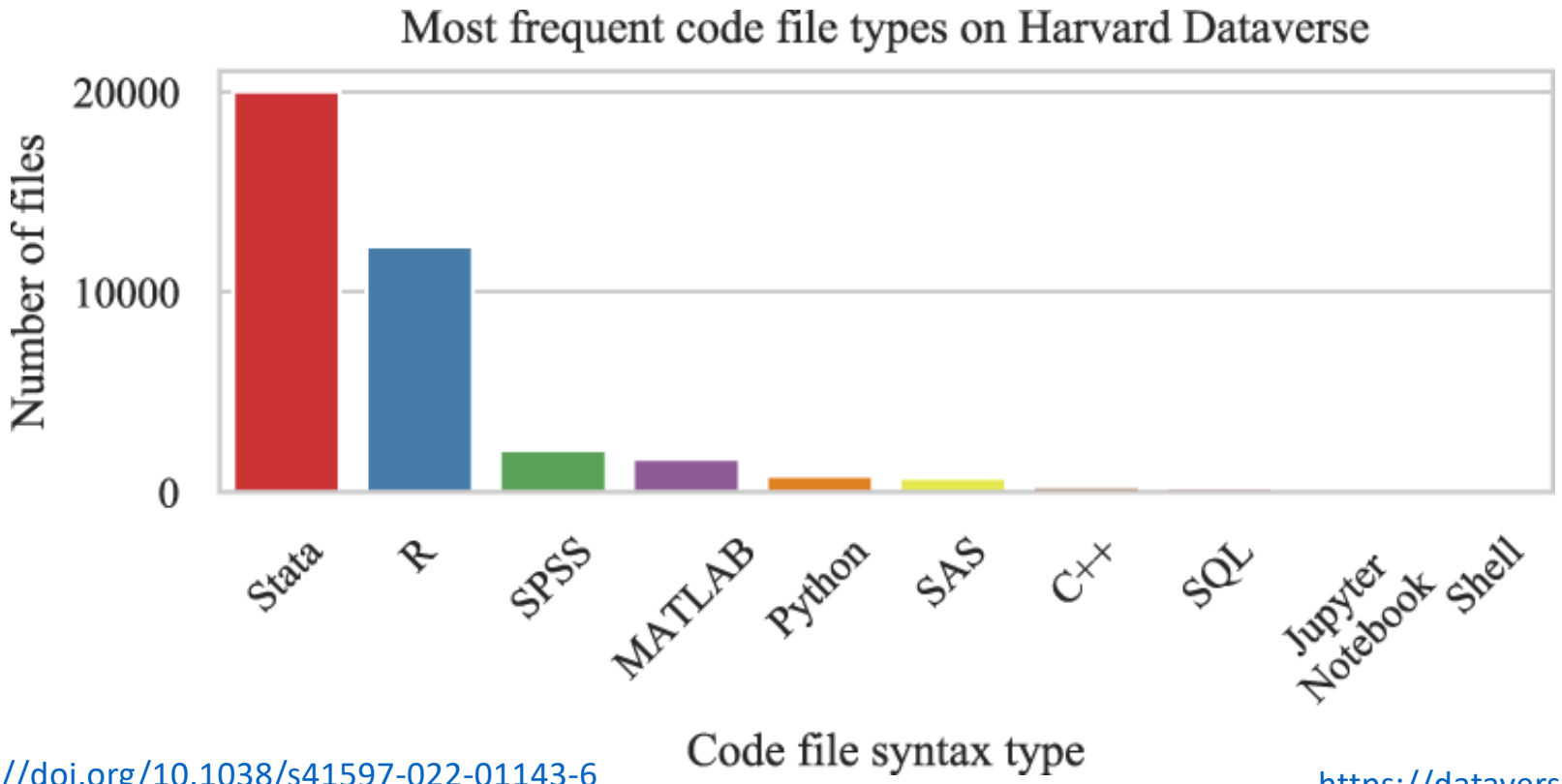
9542 Accesses | **5** Citations | **164** Altmetric | Metrics

Abstract

This article presents a study on the quality and execution of research code from publicly-available replication datasets at the Harvard Dataverse repository. Research code is typically created by a group of scientists and published together with academic papers to facilitate research transparency and reproducibility. For this study, we define ten questions to address aspects impacting research reproducibility and reuse. First, we retrieve and analyze more than 2000 replication datasets with over 9000 unique R files published from 2010 to 2020. Second, we execute the code in a clean runtime environment to assess its ease of reuse. Common coding errors were identified, and some of them were solved with automatic code cleaning to aid code execution. We find that 74% of R files failed to complete without error in the initial execution, while 56% failed when code cleaning was applied, showing that many errors can be prevented with good coding practices. We also analyze the replication datasets from journals' collections and discuss the impact of the journal policy strictness on the code re-execution rate. Finally, based on our results, we propose a set of recommendations for code dissemination aimed at researchers, journals, and repositories.

https://dataverse.harvard.edu/

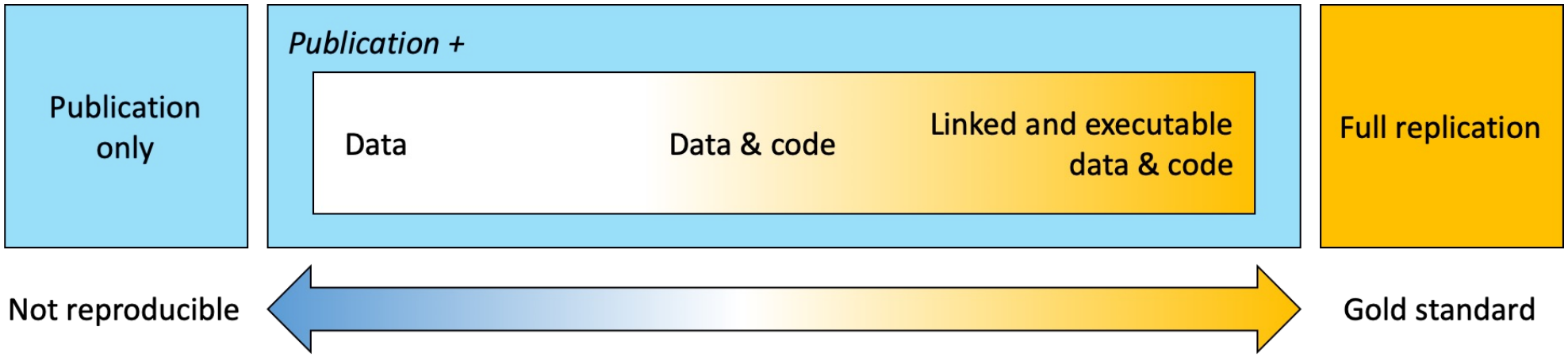
R IS AMONG TOP COMPUTING LANGUAGES USED AND IT IS OPEN SOURCE AND FREE



https://doi.org/10.1038/s41597-022-01143-6

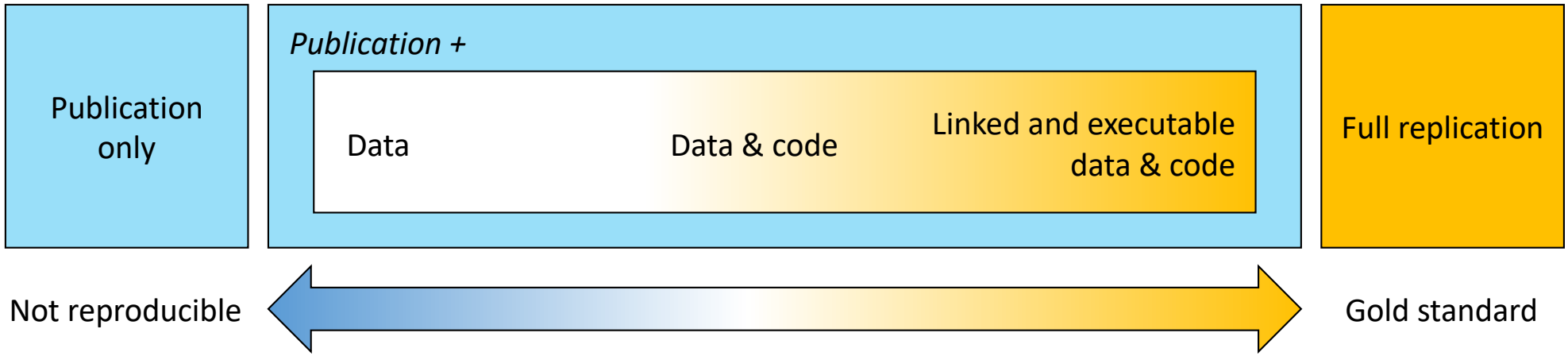
https://dataverse.harvard.edu/

WHY SUCH CRISIS? THE REPRODUCIBILITY SPECTRUM



THE REPRODUCIBILITY SPECTRUM

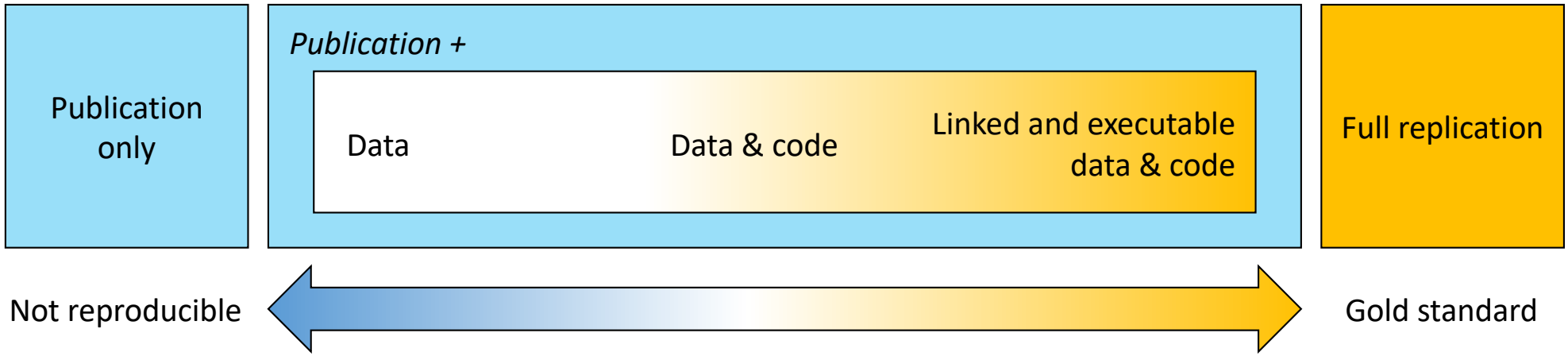
Research is often presented in the form of publications. These documents announce a project's findings, but they are not the research, they are the advertisement part of the research project!



THE REPRODUCIBILITY SPECTRUM

R Markdown
from RStudio

When research (data, code) is separated from its advertisement (publication), it becomes challenging for others to reproduce findings. But this can be solved by…



Code: Covering procedures in the wet and dry labs.

Example: publication & reproducible report

Open Access Article

Genomic Insights into Cultivated Mexican *Vanilla planifolia* Reveal High Levels of Heterozygosity Stemming from Hybridization

by Paige Ellestad [1,*], Miguel Angel Pérez-Farrera [2] and Sven Buerki [1]

[1] Department of Biological Sciences, Boise State University, 1910 University Drive, Boise, ID 83725, USA

[2] Herbario Eizi Matuda, Laboratory of Evolutionary Ecology, Institute of Biological Sciences, Universidad de Ciencias y Artes deChiapas, Libramiento Norte Poniente 1151, Col. Lajas Maciel, Tuxtla Gutiérrez 29039, Mexico

* Author to whom correspondence should be addressed.

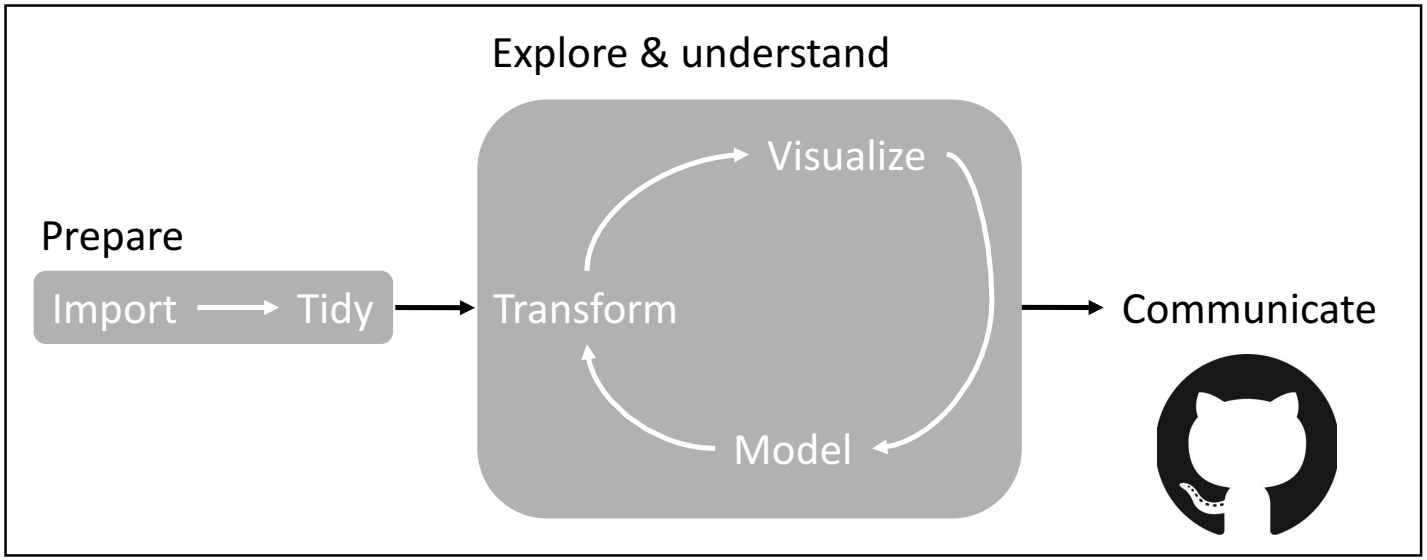
Academic Editor: Pasquale Tripodi

Plants **2022**, *11*(16), 2090; **https://doi.org/10.3390/plants11162090**

https://www.mdpi.com/2223-7747/11/16/2090
https://svenbuerki.github.io/VanillaGenomicsCode/

AIM OF THE COURSE

To provide students with key theoretical and practical **knowledge to gather, store, share, prepare and analyze data as well as communicate results to the scientific community**



STRUCTURE OF THE COURSE

The class is subdivided into three parts:

- PART 1: *The Big Picture*

  Output: key theoretical knowledge allowing students to successfully implement a reproducible approach.

- PART 2: *Bioinformatics for Reproducible Science*

  Output: learn and apply bioinformatic tools required to implement a reproducible workflow. This will be done by using the computing environment implemented in RStudio.

- PART 3: *Apply a Reproducible Approach to your Data*

  Output: develop individual reproducible workflows tailored to thesis project or to data presented in a publication.

Content of the Course & Timetable

PART 1: The Big Picture

1. **Introduction: The reproducibility crisis & prospects to tackle it!**
   - Chapter 1: Introduction to R, RStudio, Markdown (incl. referencing) & User-defined functions.
   - Chapter 2: The reproducibility crisis.
   - Chapter 3: A road map to implement reproducible science in Ecology & Evolution.
   - Chapter 4: Open science and CARE principles.
2. **Getting started: Overview of data workflow & used software**
   - Chapter 5: Data management, Reproducible code.
   - Chapter 6: Getting published & Peer review.

PART 2: Bioinformatics for Reproducible Science

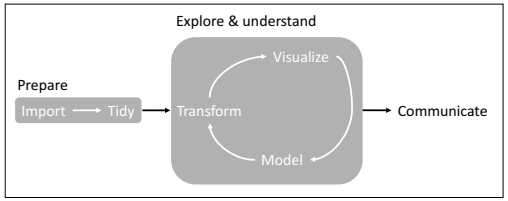
3. **Prepare your data**
   - Chapter 7: Organize and import data with R.
   - Chapter 8: Prepare/tidy data for analyses in R.
4. **Explore & understand your data**
   - Chapter 9: Statistical modelling and *knitr*.
   - Chapter 10: Visualize results with Tables.
   - Chapter 11: Visualize results with Figures (incl. handling phylogenetic trees).
5. **Communicate/Disseminate your results**
   - Chapter 12: Git & GitHub: What are those and how can they help you with your code and communicating your research?

PART 3: Apply a Reproducible Approach to your Data

6. **Students develop and present reproducible workflows applied to their project**
   - Produce individual reports showcasing reproducible workflows tailored to thesis projects.
   - Oral presentations of individual students' projects.

Learning the fundamentals:
Split sessions between:
- Lectures & group discussions
- Students work on tutorials

Bioinformatic labs:
Students design and teach tutorials



Individual projects:
Students produce reproducible workflows tailored to their data

COURSE MATERIAL & RESOURCES

svenbuerki.github.io/EEB603_Reproducible_Science/index.html

Reproducible Science | Home | Timetable | Chapters | Resources

1 Instructor
2 Our class ethos
3 Class details
4 Course goal & description
5 Structure of the course
6 Shared Google Drive
7 Publications & Textbooks
8 The computing tools of reproducible science
9 RStudio Cheat Sheets: A gold mine to design your bioinformatic tutorials
10 R tutorials
11 Assessments
12 Late work policy
13 Engagement expectations
14 What you can expect of me
15 This course was designed with you in mind
16 Academic integrity
17 Student well-being
18 References
19 Appendix 1

EEB 603 – Reproducible Science

BOISE STATE UNIVERSITY

Syllabus

Sven Buerki - Boise State University

2022-08-16

Download pdf version
Raw data on GitHub

1 Instructor

- Name: Sven Buerki
- Office: Science building, office 114 (ground floor).
- Email: svenbuerki@boisestate.edu
- Office hours: By appointment.

GitHub

https://github.com/svenbuerki/EEB603_Reproducible_Science

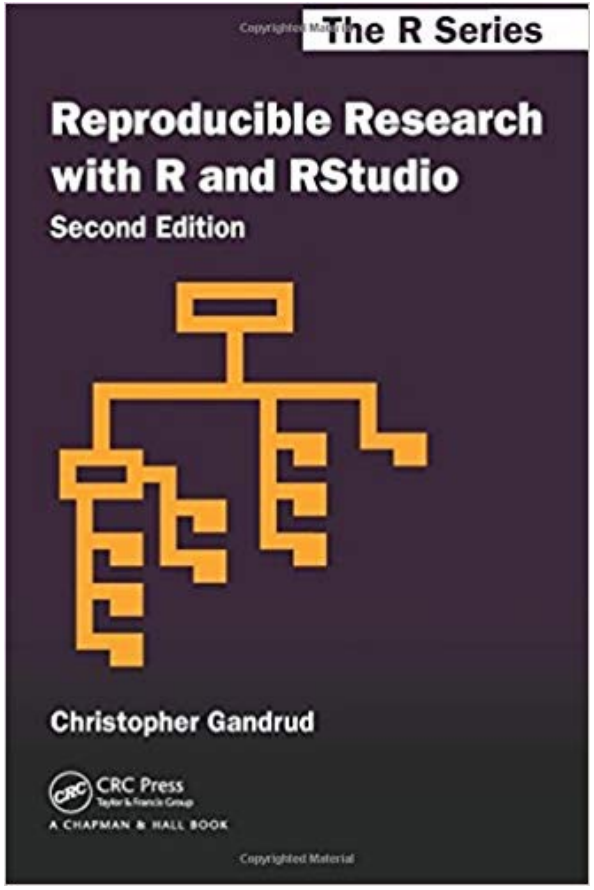
2 Our class ethos

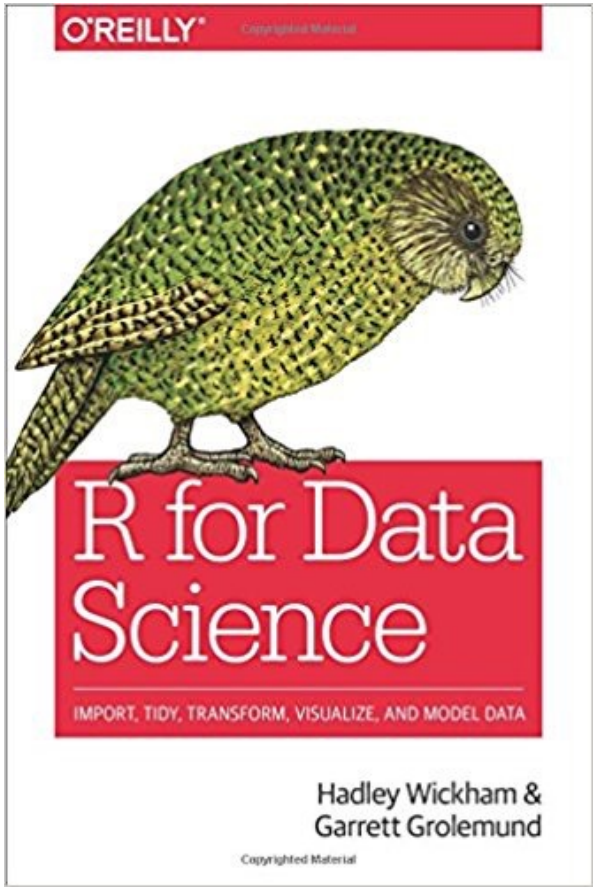
> Everyone here is smart; distinguish yourself by being kind.
>
> — Kindness in Science is an inclusive approach that fosters diversity, respect, wellbeing & openness leading to better science outcomes.

See class website for full list

PUBLICATIONS & TEXTBOOKS

https://bookdown.org/





BIOINFORMATIC TOOLS

The main bioinformatic tools covered in this class are:

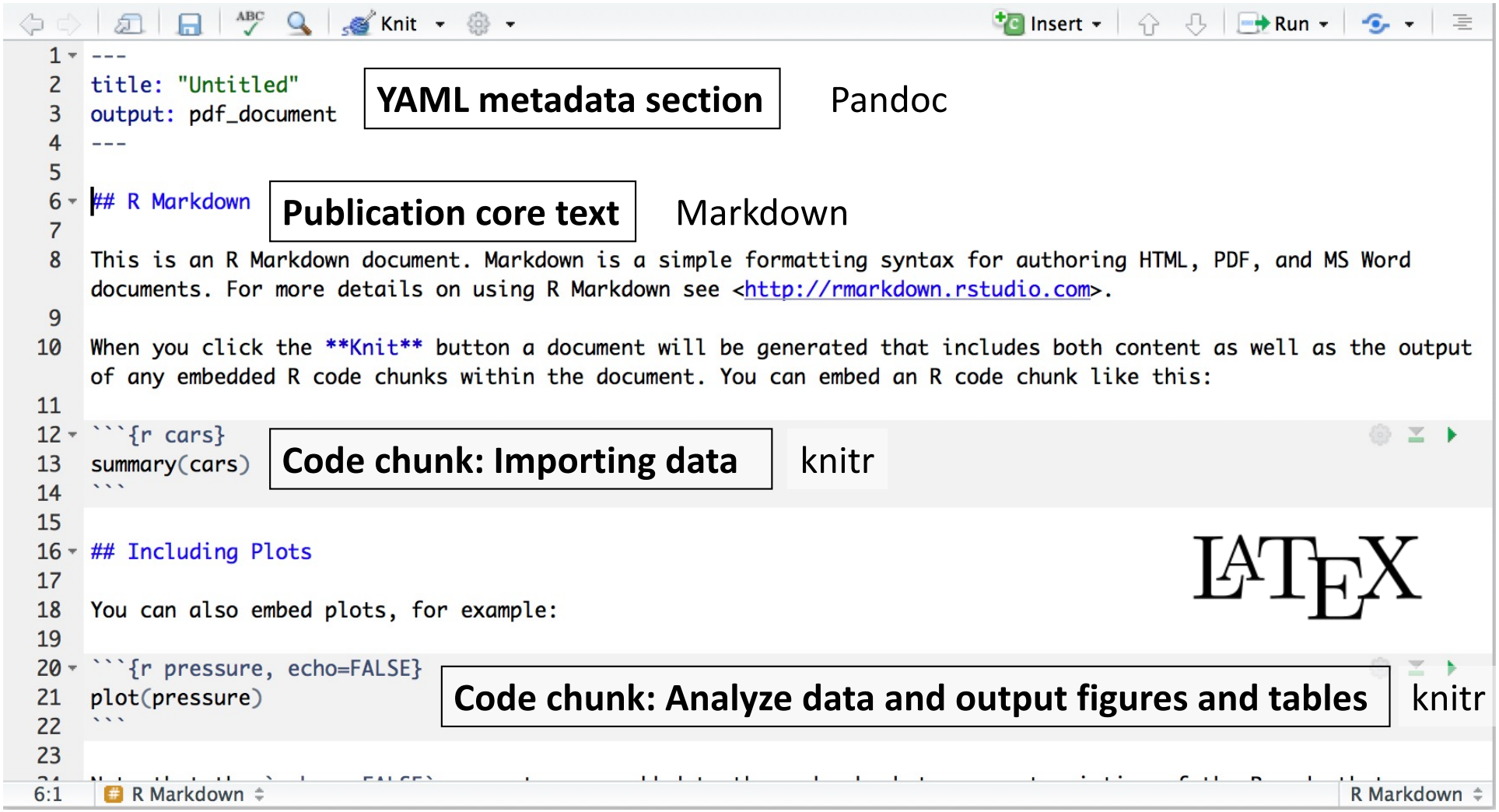
- **R** statistical language: **gather and analyze data**.
- **Markdown** markup language: **create documents** (slideshows, articles, books, webpages) **for presenting your findings**.
- ***knitr*** and ***rmarkdown*** R packages: **dynamically tiding, gathering, analyzing, and presenting your data into one document.** This is the core of the process allowing data reproducibility. Several additional R packages will be required for this course.
- **RStudio**: **framework implementing all tools in one place**.

R MARKDOWN: THE GOLD STANDARD FOR REPRODUCIBLE SCIENCE

This workflow integrates 4 programming languages:

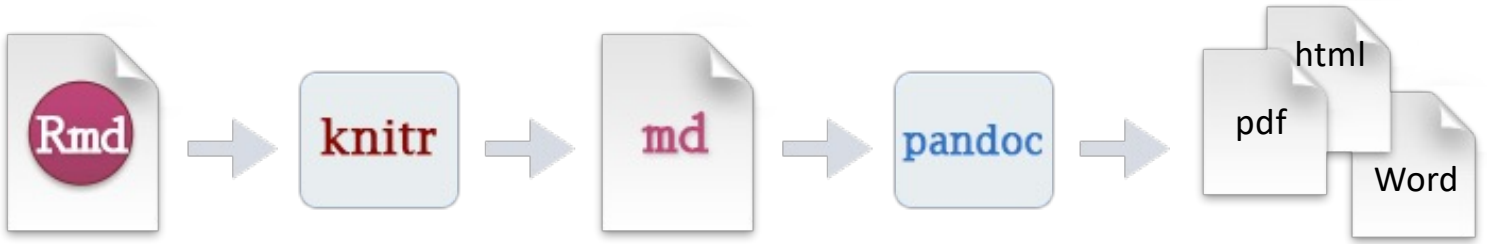
1. R (interpreted by *knitr*)
2. Markdown (*rmarkdown*)
3. YAML[1] (interpreted by Pandoc[2])
4. LaTeX (only if output is pdf)

[1] YAML (YAML Ain't Markup Language) is a human friendly data serialization standard for all programming languages.
[2] If you need to convert files from one markup format into another, Pandoc is your swiss-army knife.


Knit
Insert
Run
---
title: "Untitled"
output: pdf_document

R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
```{r cars}
summary(cars)
```
Including Plots
You can also embed plots, for example:
```{r pressure, echo=FALSE}
plot(pressure)
```
YAML metadata section
Pandoc
Publication core text
Markdown
Code chunk: Importing data
knitr
LaTeX
Code chunk: Analyze data and output figures and tables
knitr
6:1
R Markdown
R Markdown

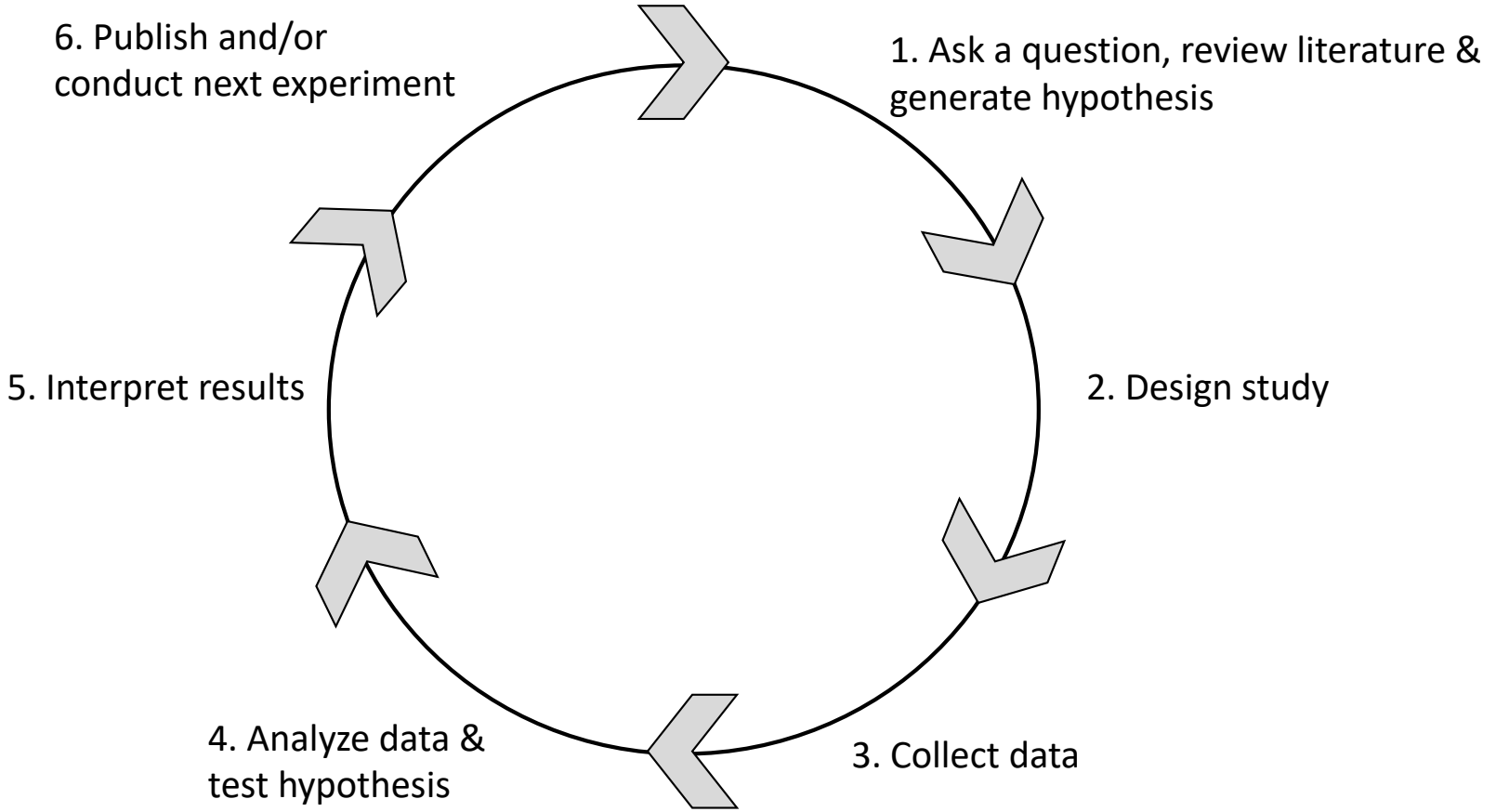
R MARKDOWN WORKFLOW



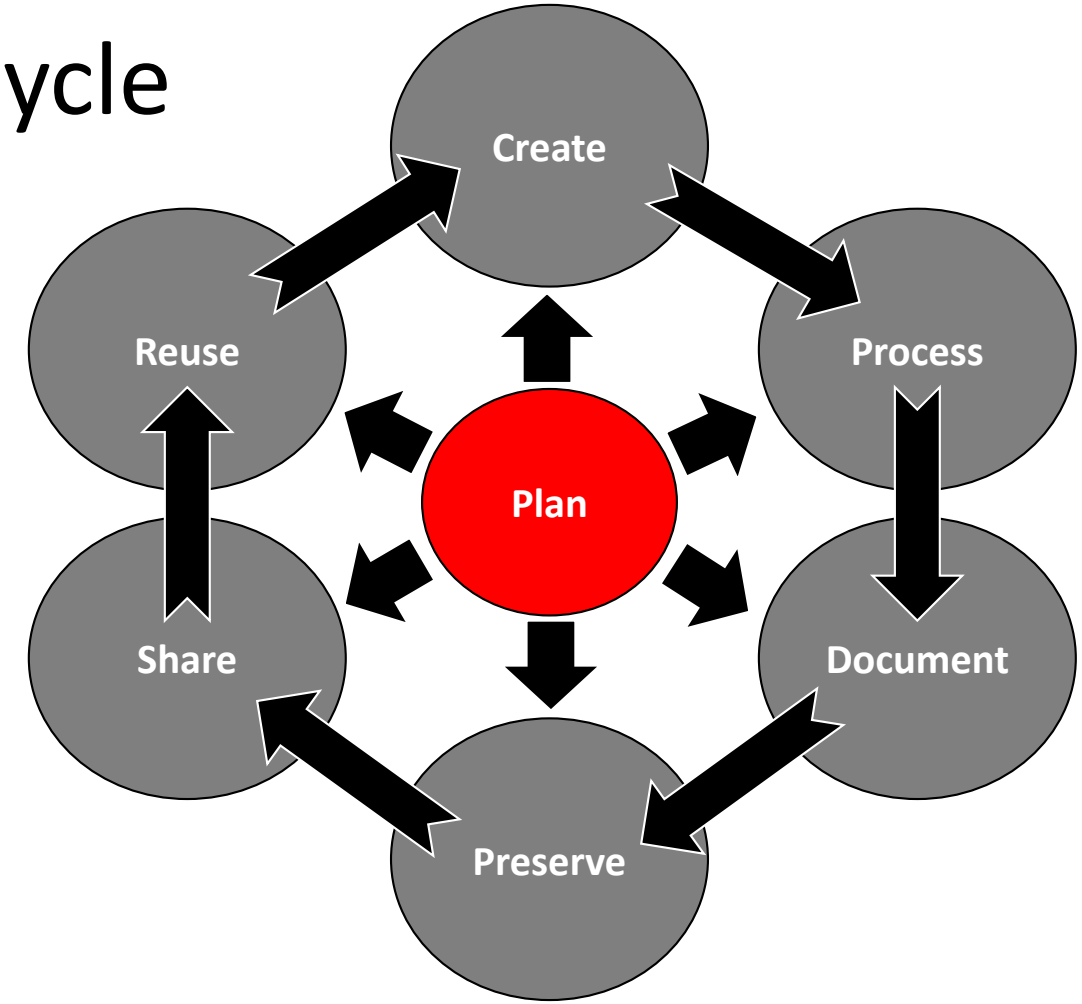
Executes code chunks and creates new markdown document including code and outputs

Responsible for creating the finished format and converting into final format

THE SCIENTIFIC PROCESS



The data life-cycle



Data dissemination & Open Science

https://www.nature.com/articles/s41597-021-00892-0

Data dissemination & Open Science

Ecology and Evolution
Open Access

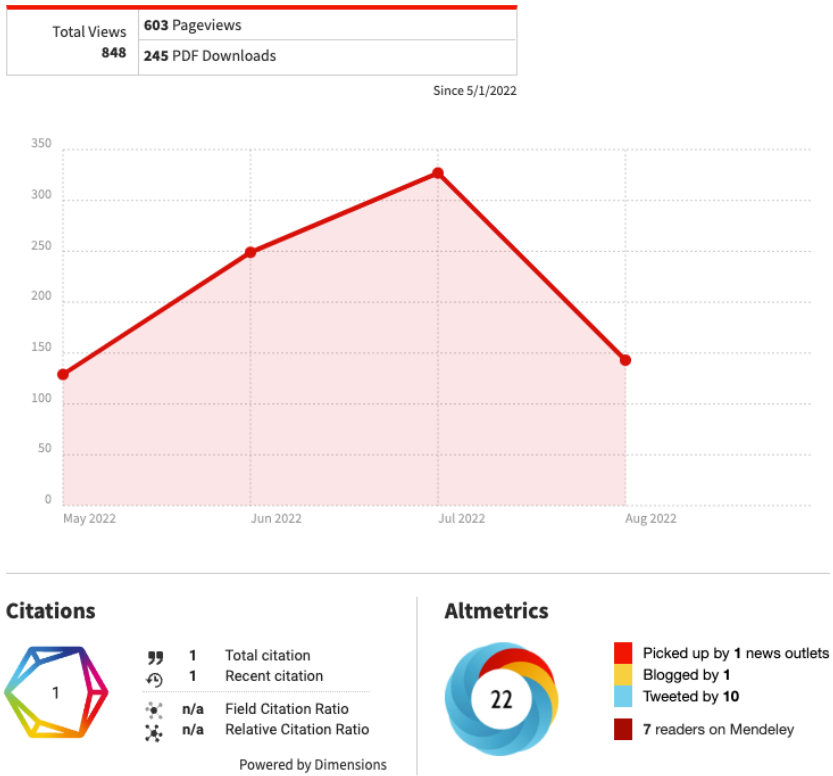
Editors-in-Chief: Allen Moore, University of Georgia, USA; Andrew Beckerman, University of Sheffield, UK; Marcus Lashley, University of Florida, USA; Chris Foote, John Wiley & Sons, UK; Gareth Jenkins, John Wiley & Sons, UK, and Zhaoxue Ma (Zoe Ma), John Wiley & Sons, Shanghai

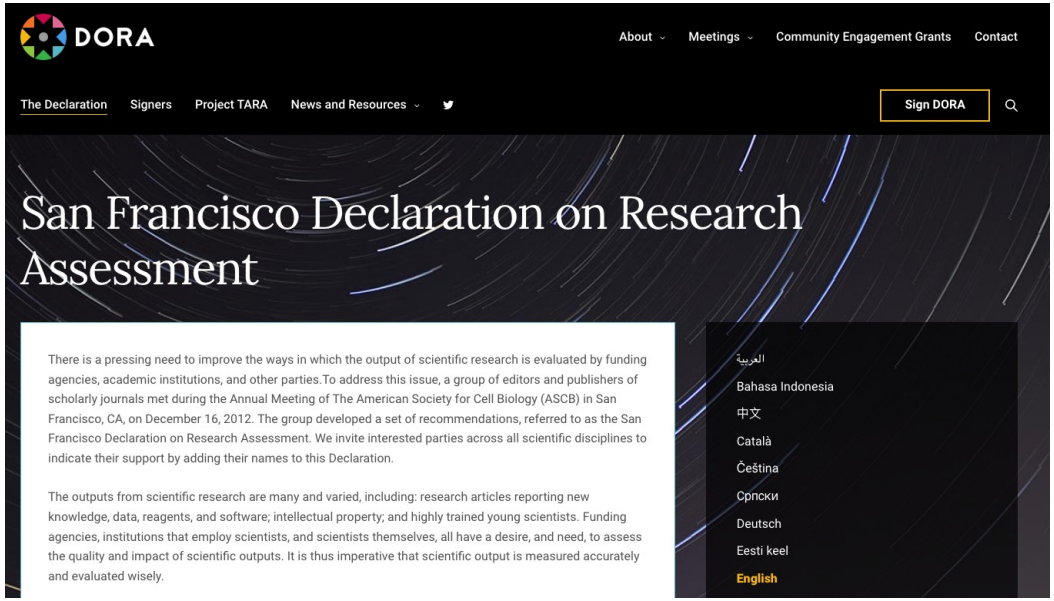
Online ISSN: 2045-7758

https://creativecommons.org/licenses/by/4.0/

Changing our perspective on research impact

Metrics beyond impact factor





CLASS ASSESSMENTS

- Students will be graded based on the following four tasks:
  1. **Produce a bioinformatics tutorial** focusing on a chapter from PART 2 (150 points).
  2. **Teach a bioinformatics lab** (spread across 2 sessions) (100 points).
  3. **Produce an individual report** on thesis project/publication (200 points).
  4. **One oral presentation** on thesis project/publication (100 points).

- TOTAL: **550 points** (see Syllabus for grading scale).

BIOINFORMATIC TUTORIAL (150 POINTS)

- Students will be working in groups (or alone) and assigned a chapter of PART 2 to produce a bioinformatic tutorial.
- Tutorials have to:
  - ✓Be written in R Markdown as implemented in RStudio (Chapter 1).
  - ✓Focus on developing a suite of exercises aiming at gaining key bioinformatic skills specific to each chapter.
  - ✓Be submitted to the instructor 1 week in advance for correction and then uploaded onto the Google Drive (see Timetable).
- Instructor will provide an introduction to R Markdown (Chap. 1) and allocate time to get accustomed to markup language and RStudio.
- See Syllabus for list of relevant publications and textbooks to develop tutorials.

BIOINFORMATIC TUTORIAL (150 POINTS)

- Tutorials should include:
  - ✓A short introduction highlighting the theory and aims of the tutorial.
  - ✓A section on R package requirements with instructions on how to install those and their dependencies (see Syllabus).
  - ✓A section introducing the data (dataset) used to support these exercises (and how to download those).
  - ✓A references section and links to manuals of R packages.
  - ✓Commented R code necessary to guide users through the exercises as well as some knowledge on expected outputs (more details on this topic in Chapters 5 & 6).

TEACHING BIOINFORMATIC TUTORIAL (100 POINTS)

- Students prepare 10-20 minutes presentations providing general guidelines to complete tutorials.

- Students should support their peers in completing tutorials (by e.g. answering questions).

- Instructor will also be circulating in class and answering questions, but students are leading the teaching of the bioinformatic labs.

- Students will be graded according to their abilities to teach their tutorials and answer questions. The instructor might also use student's feedback to grade this test.

INDIVIDUAL REPORT (200 POINTS)

- Students work alongside instructor to develop reproducible workflows specific to their thesis projects.

- In cases where students don't yet have a clear idea on their thesis project, they will work with instructor to identify a publication that can serve as basis for their individual project.

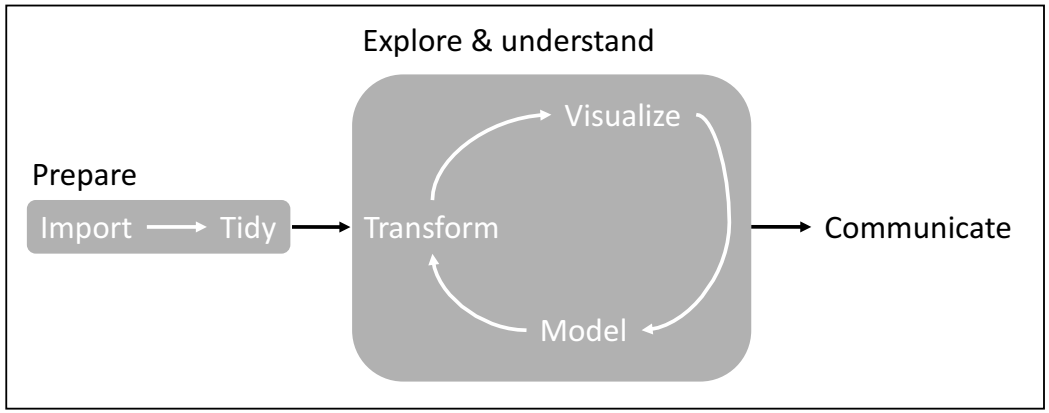
- Reports are written in RStudio (same as tutorial) and should include a list of references. This latter aims at supporting methodological decisions taken in the report and increasing transparency.

INDIVIDUAL REPORT (200 POINTS)

- Students investigate the following *prior* elements:
  - ✓What kind of data are already published/available (and which of those are relevant to your topic)?
  - ✓Where are those published data deposited?
  - ✓Can you reproduce the analyses based on published data?
  - ✓What types of data are or will be produced during your thesis project?
  - ✓What are the specificity of your data (in term of storage, sharing, etc.)?
  - ✓What are the publication standards in your field?

INDIVIDUAL REPORT (200 POINTS)

- Students develop the following core processes:
  - ✓A data management plan specific to your research covering the following stages of the data life cycle: Create, Process, Document, Preserve, Share and Reuse.
  - ✓ A reproducible code to perform the following tasks to your data:



PRESENTATION: INDIVIDUAL REPORT (100 POINTS)

- Prepare 9-minute presentations.
- 5 minutes are devoted to questions and 1 minute for transitioning to the next speaker (see next slide for speaker order and dates).
- Students share their screen and classes will be recorded (URLs will be posted on Google Drive).
- Students are not required to upload their presentations on Google Drive (**BUT** students will have to upload their reports in Personal_reports folder).
- Students can directly use output created by RMarkdown (e.g. the HTML page) for their oral presentation or prepare slides with Powerpoint or Google Slides.

PRESENTATION: INDIVIDUAL REPORT (100 POINTS)

- Students have “carte blanche” to prepare the content of their presentations, but they should cover the following aspects:
  - ✓What is the general context of the study.
  - ✓What is the central scientific question.
  - ✓What are your objectives and if possible working hypothesis.
- You are encouraged to talk about the challenges that you are facing to make your study reproducible and transparent.

ORAL PRESENTATION (100 POINTS)

- Each student will have to present their report during final week.
- The presentation should follow the same structure as the report.
- Presentation should not exceed 15 minutes (+ 5 minutes for questions).