Reproducible science: Module2

Bridging the gap among actors

Gbadamassi G.O. Dossa

Xishuangbanna Tropical Botanical Garden, XTBG-CAS

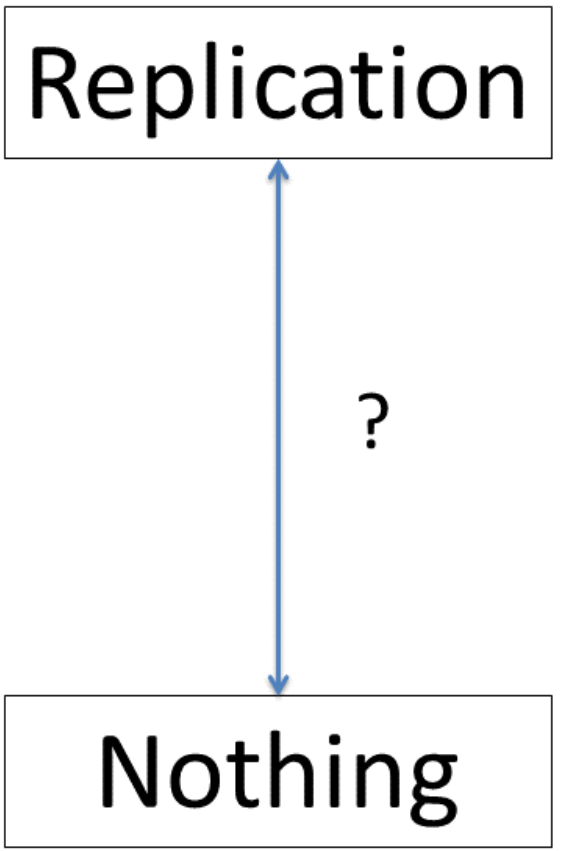
2021/09/20 (updated: 2021-11-02)

Acknowledgements

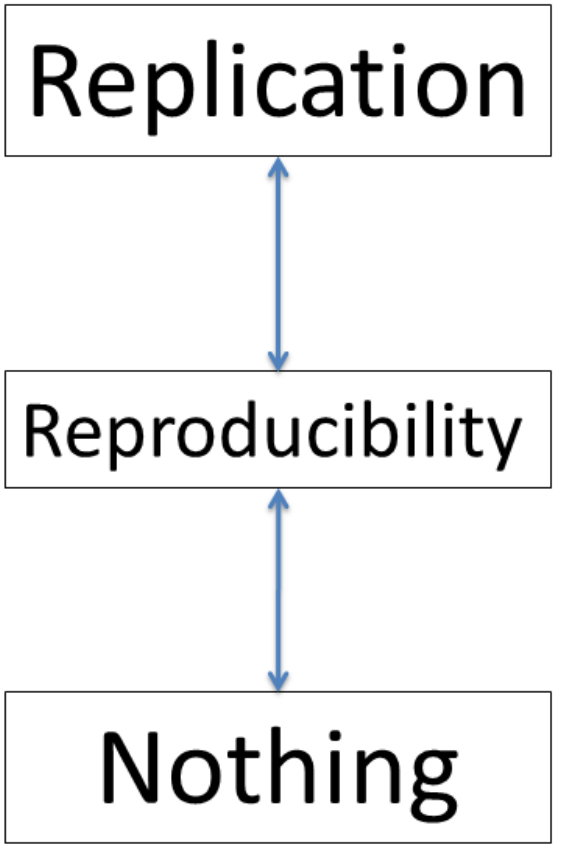
The content of this module are based on materials from:

Roger D. Peng's materials

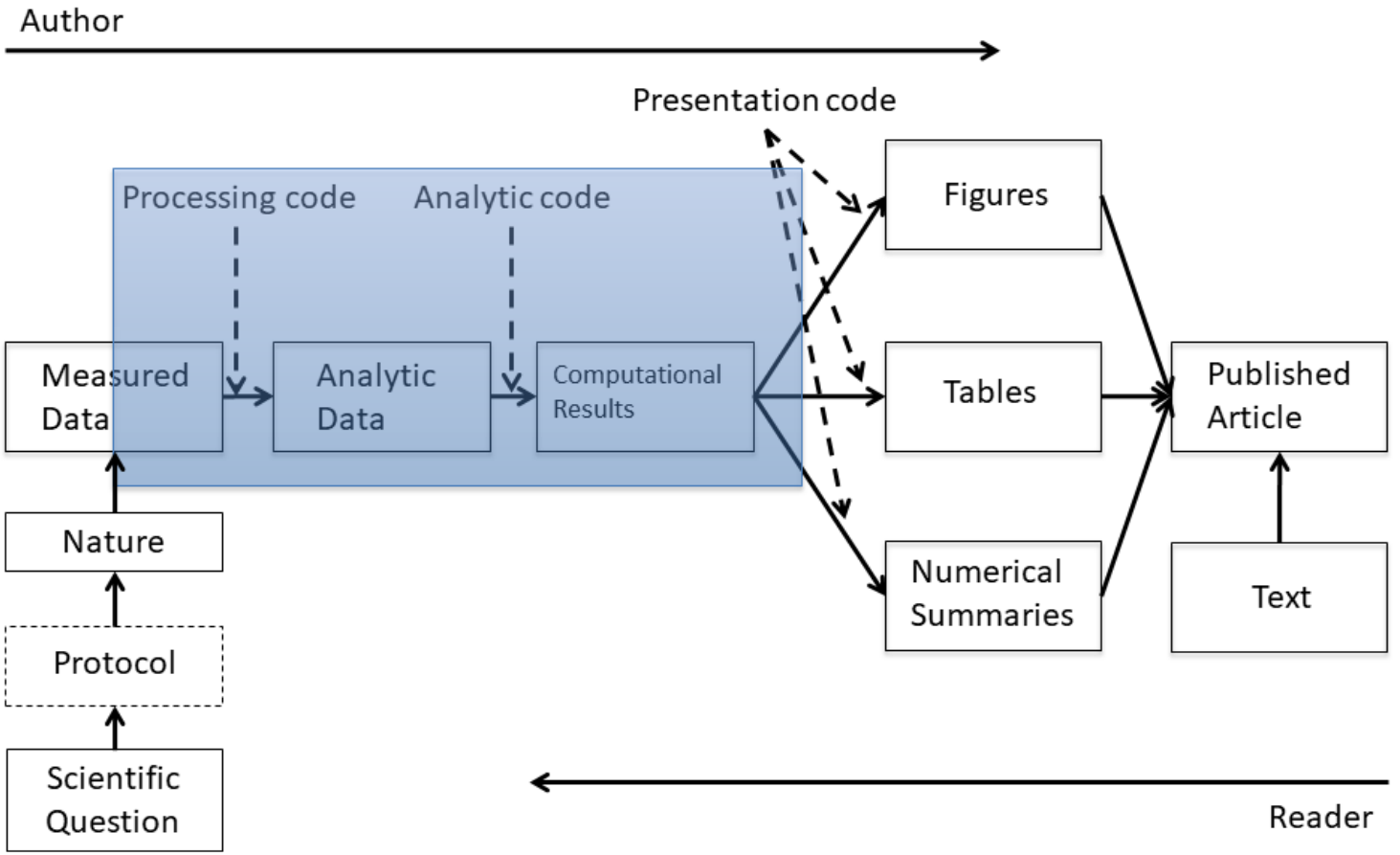
Where is the Gap?



How can we bridge the gap?



Research pipeline



Please examine again this pipeline.

Recent developments in reproducible research

Data Replication & Reproducibility

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

What do we need for reproducible research?

- Analytic data are available;
- Analytic code are available;
- Documentation of code and data;
- Standard means of distribution.

Who are the players?

- Authors:

  -> Want to make their research reproducible

  -> Want tools for RR to make their lives easier (or at least not much harder)

- Readers:

  -> Want to reproduce (and perhaps expand upon) interesting findings

  -> Want tools for RR to make their lives easier

Existing chalenges

- Authors must undertake considerable effort to put data/results on the web (may not have resources like a web server);
- Readers must download data/results individually and piece together which data go with which code sections, etc.;
- Readers may not have the same resources as authors;
- Few tools to help authors/readers (although toolbox is growing!).

Responses to challenges: in reality

- Authors

  -> Just put stuff on the web

  -> (Infamous) Journal supplementary materials

  -> There are some central databases for various fields (e.g. biology, ICPSR)

- Readers

  -> Just download the data and (try to) figure it out

  -> Piece together the software and run it

Literate (statistical) programming

- An article is a stream of text and code;
- Analysis code is divided into text and code chunks;
- Each code chunk loads data and computes results.
- Presentation code formats results (tables, figures, etc.);
- Article text explains what is going on;
- Literate programs can be weaved to produce human-readable documents and tangled to produce machine-readable documents.

Literate (statistical) programming 2

Literate programming is a general concept that requires:

- A documentation language (human readable);
- A programming language (machine readable):

    1. Sweave uses LaTeX and R as the documentation and programming languages
    2. Sweave was developed by Friedrich Leisch (member of the R Core) and is maintained by R core

Sweave's main web site.

Sweave Limitations

Sweave has many limitations:

- Focused primarily on LaTeX language, a difficult to learn markup language used only by weirdos;
- Lacks features like caching, multiple plots per chunk, mixing programming languages and many other technical items;
- Not frequently updated or very actively developed.

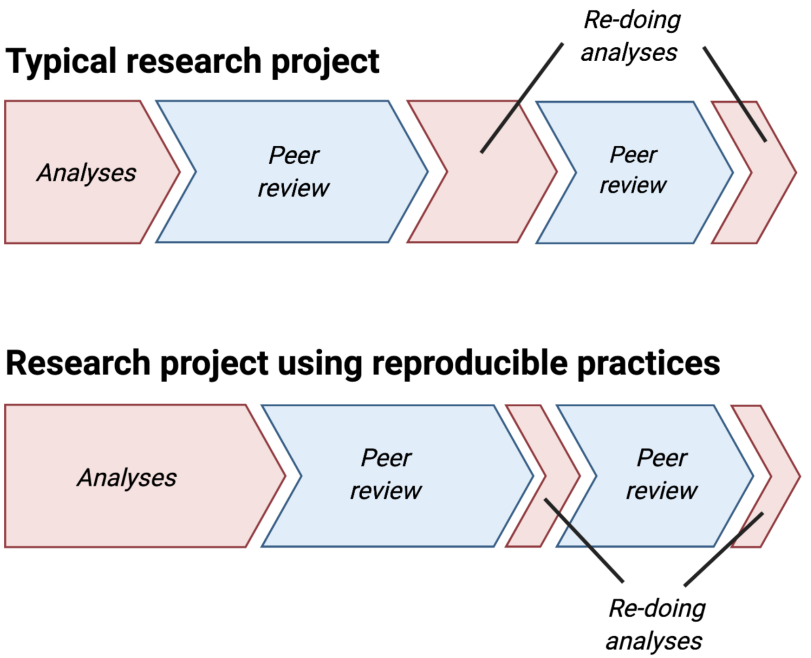
So is there any alternative to Sweave?

Alternative to Sweave: knitr

- Yes, there is knitr is an alternative (more recent) package to Sweave developed by **Yihui Xie**;
- Brings together many features added on to Sweave to address limitations;
- knitr uses R as the programming language and variety of documentation languages LaTeX, Markdown, HTML;
- knitr was developed by Yihui Xie (while a graduate student in statistics at Iowa State);
- See *knitr* for more and in lectures later.

Does reproducibility consume more time?

You don't *lose* time doing reproducible science, you just *relocate* where you spend it



Reproducibilit equals effecient use of time

Summary

- Reproducible research is important as a minimum standard, particularly for studies that are difficult to replicate;
- Infrastructure is needed for creating and distributing reproducible documents, beyond what is currently available;
- There is a growing number of tools for creating reproducible documents.

Thank you for listening!

Any questions now or email me at **dossa@xtbg.org.cn**

Slides created via the R package **xaringan**.

The chakra comes from remark.js, **knitr**, and R Markdown.