Reproducible science: Module1

Launching workshop-Replication and reproducibility

Gbadamassi G.O. Dossa

Xishuangbanna Tropical Botanical Garden, XTBG-CAS

2021/09/13 (updated: 2022-06-27)

Funny video about the reality in reproducible science

Acknowledgements

The contents of this module are based on materials from:

Roger D. Peng's materials

Replication and Reproducibility

Definitions *via* cartoon

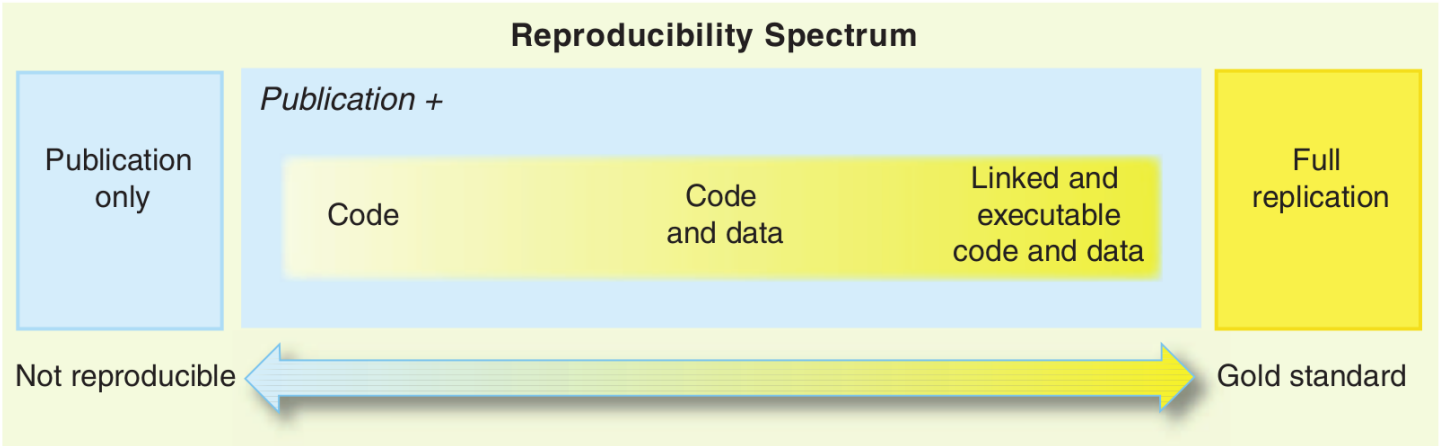
Replication and Reproducibility

Replication

- Aim to verifying a science claim
- Question: "Is this claim true?"
- Gold standard for strengthening scientific evidence
- New investigators, data, methods, laboratories, *etc.*
- Important in policy or decision driving studies

Reproducibility

- Aim to verifying a data analysis
- "Can we trust this analysis?"
- Arguably a minimum standard for any scientific study
- New investigators, **same** data, **same** methods
- Important when replication is impossible



Reproducible spectrum (Peng 2011)

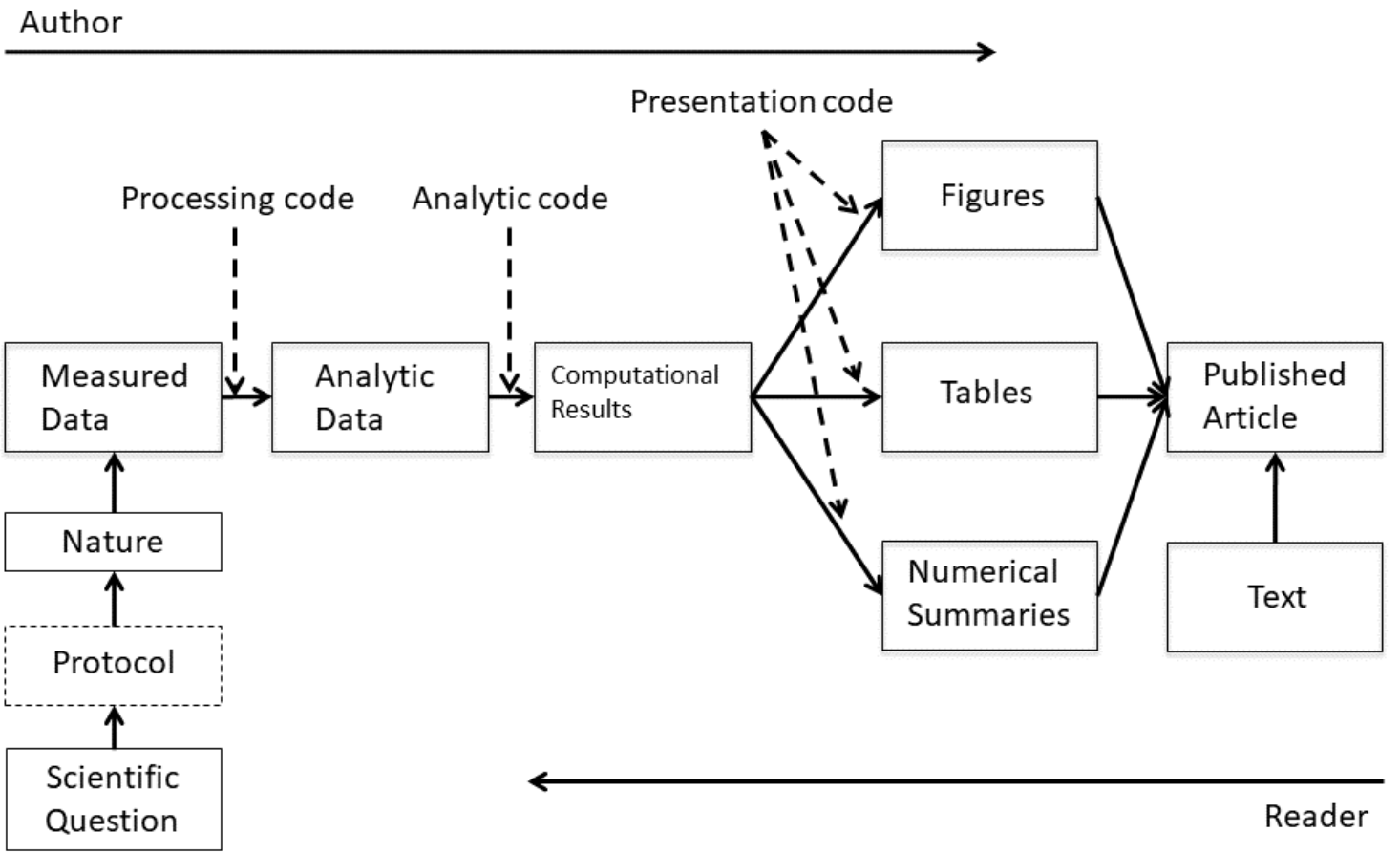
Background: Underlying Trends

- Some studies cannot be replicated: no time, no money, Unique/opportunistic
- Technology is increasing data collection throughput; data are more complex and high-dimensional
- Existing databases can be merged to become bigger databases (but data are used off-label)
- Computing power allows more sophisticated analyses, even on "small" data
- For every field "X" there is a "Computational X"

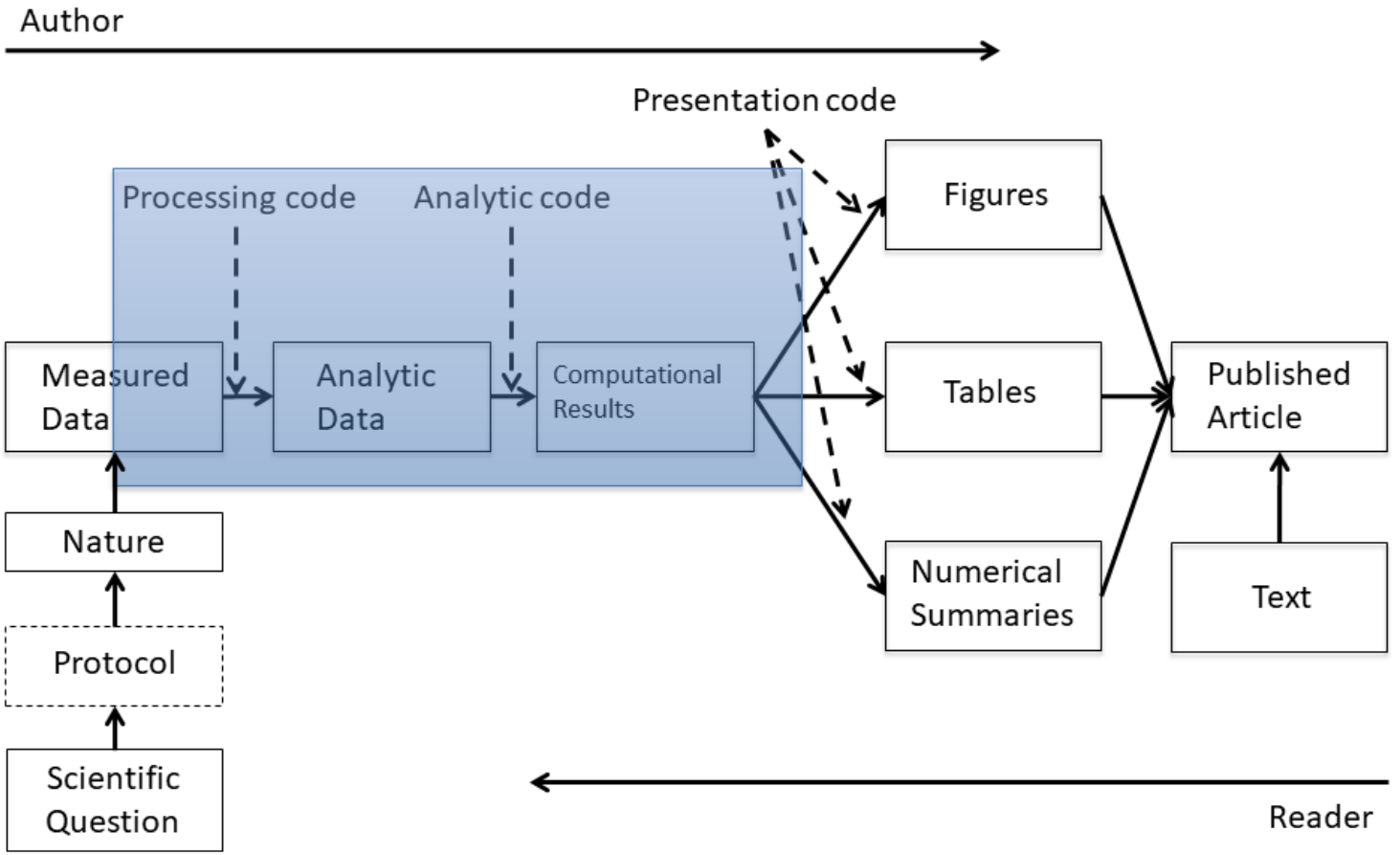
Outstanding problems: Complicated results

- Even basic analyses are difficult to describe
- Heavy computational requirements are thrust upon people without adequate training in statistics and computing
- Errors are more easily introduced into long analysis pipelines
- Knowledge transfer is inhibited
- Results are difficult to replicate or reproduce
- Complicated analyses cannot be trusted

Data science pipeline



Reproducible realm



Out of reproducibility realm

An analysis can be reproducible and still be wrong.

We want to know **"can we trust this analysis?"**

Does requiring reproducibility deter bad analysis?

What we get?

- Transparency
- Data availability;
- Software / Methods availability;
- Improved transfer of knowledge

What we do not get?

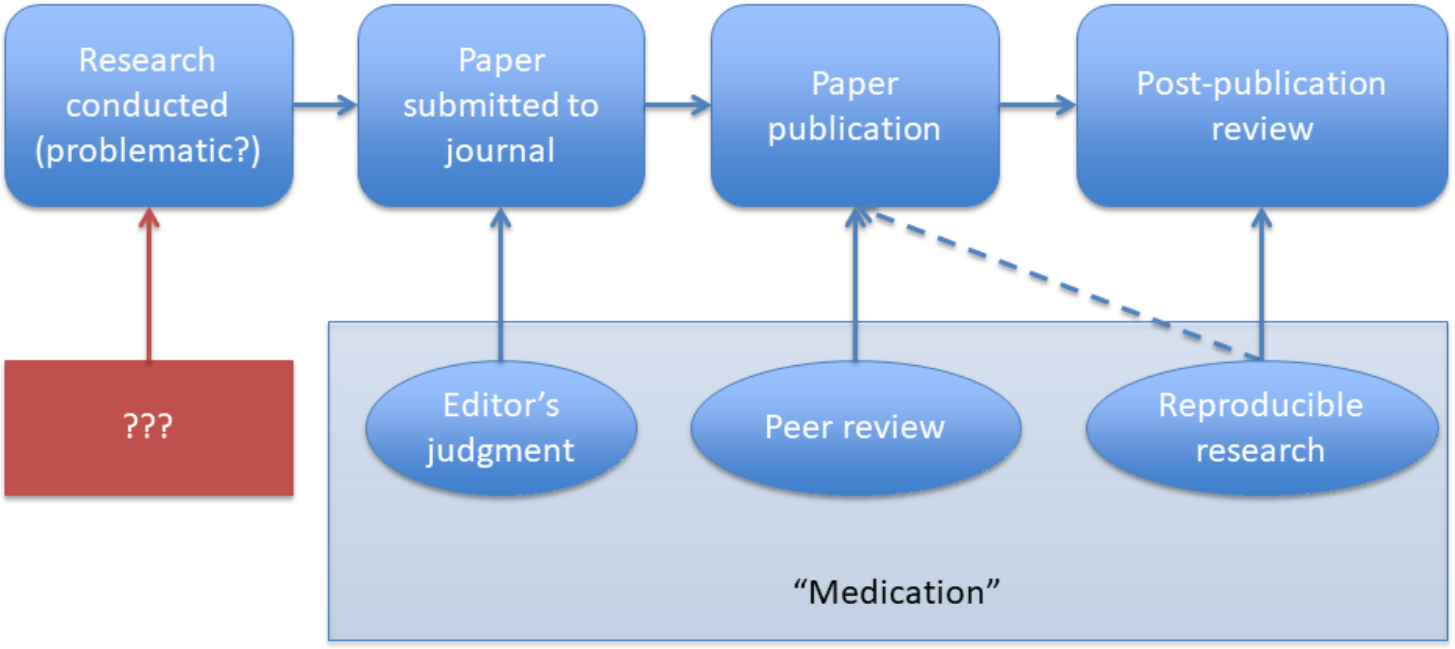
- Validity of results;
- Correctness of the analysis

Reproducibility assumption

The premise of reproducible research is that with data/code available, people can check each other and the whole system is self-correcting.

- Addresses the most "downstream" aspect of the research process post-publication;
- Assumes everyone plays by the same rules and wants to achieve the same goals (i.e., scientific discovery).

Reproducibility as preventive measure



Who reproduces research?

- For reproducibility to be effective as a means to check validity, someone needs to do something:

    1. Re-run the analysis;

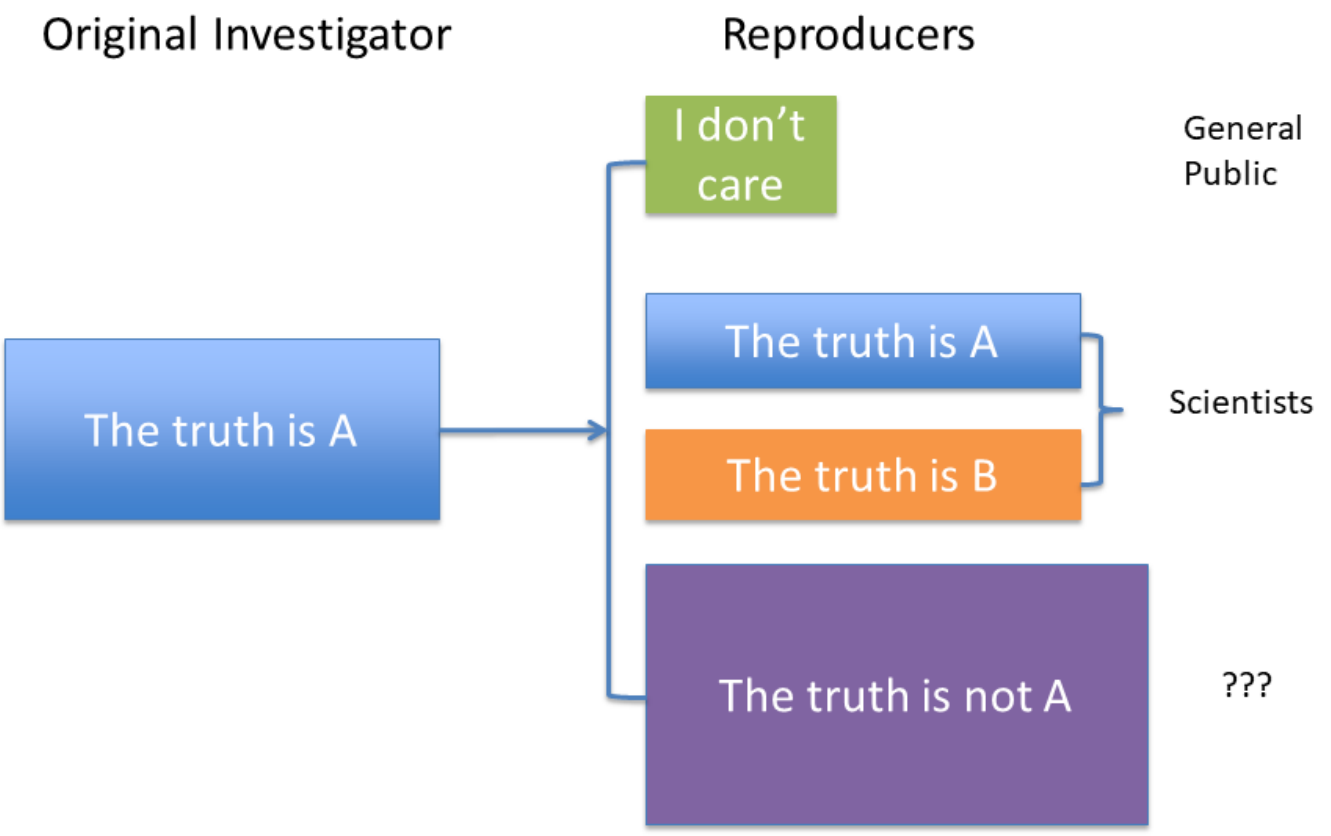
    2. Check results match;

    3. Check the code for bugs/errors

- Try alternate approaches; check sensitivity The need for someone to do something is inherited from traditional notion of replication

- Who is "someone" and what are his/her goals?

Reproducers' map



Reproducibility story so far

- Reproducibility brings transparency (wrt code+data) and increased transfer of knowledge;
- A lot of discussion about how to get people to share data;
- Key question of "can we trust this analysis?" is not addressed by reproducibility;
- Reproducibility addresses potential problems long after they've occurred ("downstream");
- Secondary analyses are inevitably coloured by the interests/motivations of others

Evidence-based data analysis

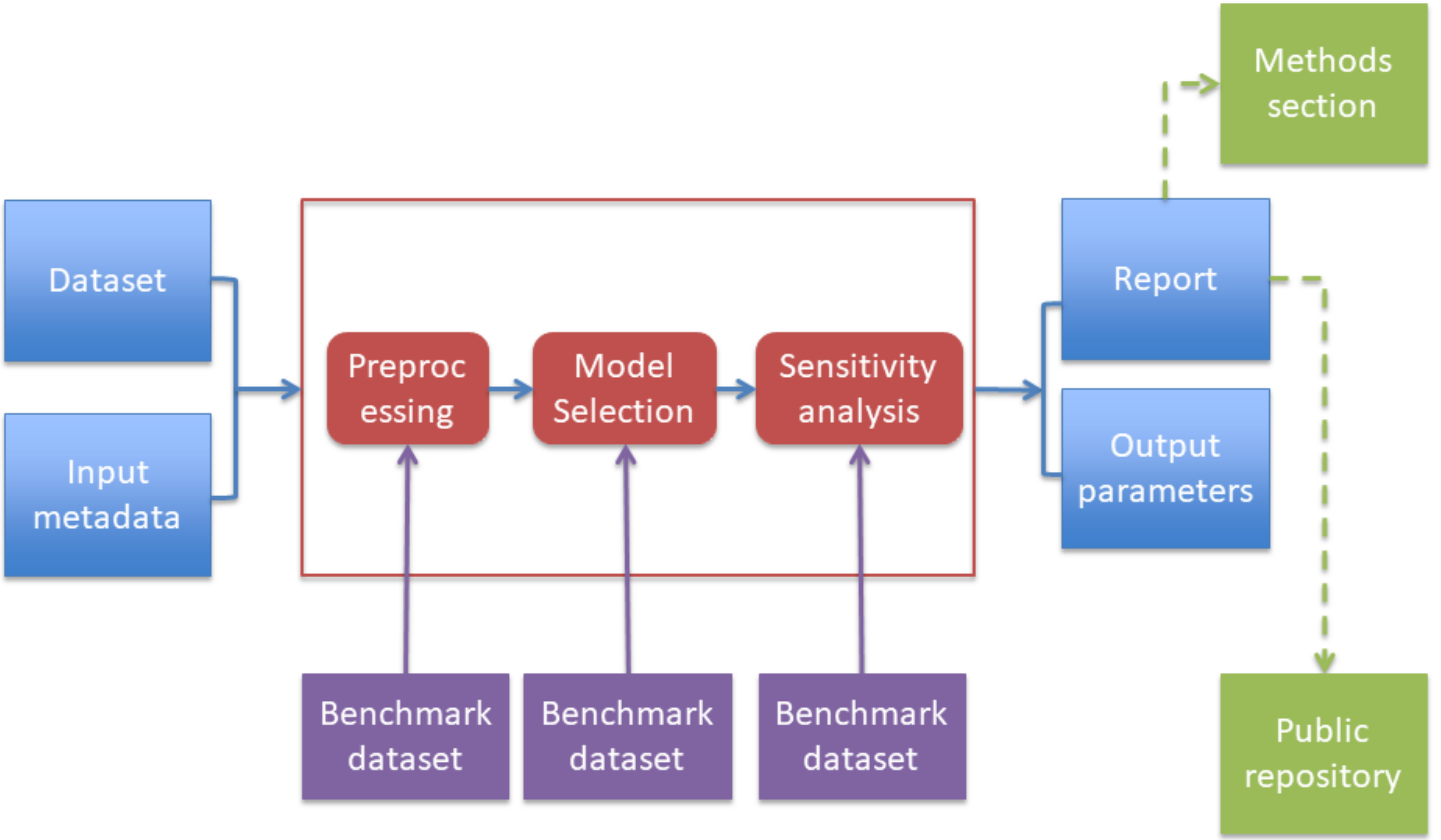
- Most data analyses involve stringing together many different tools and methods;
- Some methods may be standard for a given field, but others are often applied ad hoc;
- We should apply thoroughly studied (via statistical research), mutually agreed upon methods to analyze data whenever possible;
- There should be evidence to justify the application of a given method

Evidence-based data analysis 2

- Create analytic pipelines from evidence-based components - standardize it;
- A Deterministic Statistical Machine;
- Once an evidence-based analytic pipeline is established, we shouldn't mess with it

  -(Analysis with a "transparent box");
- Reduce the "researcher degrees of freedom";
- Analogous to a pre-specified clinical trial protocol.

Desired data analysis map



Summary

- Reproducible research is important, but does not necessarily solve the critical question of whether a data analysis is trustworthy;
- Reproducible research focuses on the most "downstream" aspect of research dissemination;
- Evidence-based data analysis would provide standardized, best practices for given scientific areas and questions;
- Gives reviewers an important tool without dramatically increasing the burden on them;
- More effort should be put into improving the quality of "upstream" aspects of scientific research

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.