Build your own Virtual Research Environment for Reproducible Research

Georgios Kaklamanos

17.07.2017

Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen
Georg-August-Universität Göttingen

REPRODUCIBLE RESEARCH

“Hypothetical” Scenarios

- Submitted a paper
  - Review process takes months
  - Requests to modify parameters / figures
  - Need to run / modify code from 6 months ago
- Found a paper which could help significantly in your research
  - No source code
  - No data

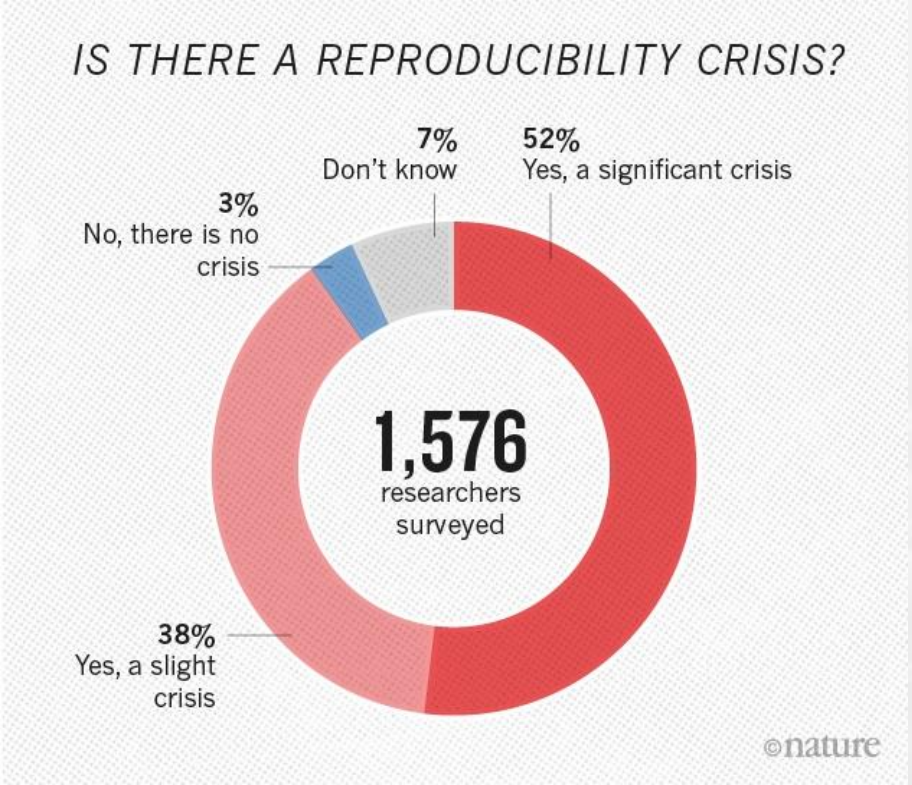
“Hypothetical” Scenarios

- Received a zip archive with files and code from a previous student at your lab
  - Student has left
  - Code doesn’t compile
  - No documentation
  - Program is critical to continue the project…
- These are global reproducibility problems

Crisis of Reproducibility

Nature Survey denotes a reproducibility crisis

- 70% of researchers
  - failed to reproduce other scientists experiments
- 50% of researchers
  - failed to reproduce own experiments



Source: [1]

Crisis of Reproducibility

How does it look across the fields?

- Phycology, 2006 study [2]
  - 249 data sets from American Psychology Association (APA) empirical articles
  - 73% of contacted authors did not respond with their data over a 6-month period.
- Cancer Research, 2012 study [3]
  - 47 out of 53 medical research papers were irreproducible (90%)
- Applied Computer Science, 2014 study [4]
  - 613 papers
  - 102 had code that could build and run

How did we get here?

Novelty

Researcher != Inventor

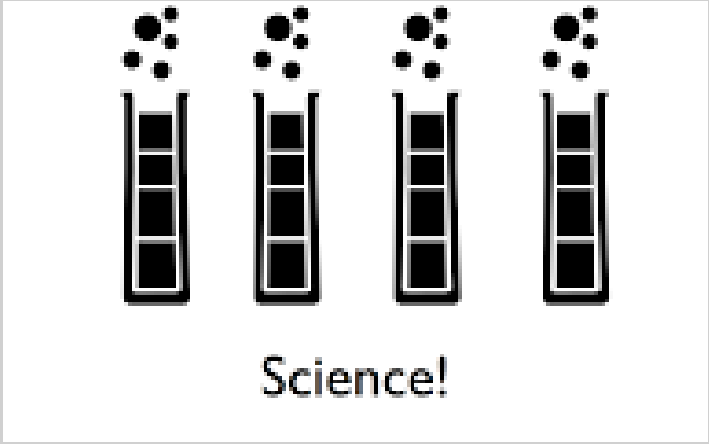
“Nullius in Verba”

“ It is an expression of the determination of Fellows to withstand the domination of authority and to verify all statements by an appeal to facts determined by experiment. ”

Royal Society

What is Replication

- Ultimate Standard for scientific evidence
- Ability to:
  - reproduce findings and conduct studies
  - with independent
    - Investigators
    - Data
    - Analytical methods
    - Laboratories
    - Instruments



Img source: http://blog.abegong.com/2013/12/replication-is-only-hope-for-science.html

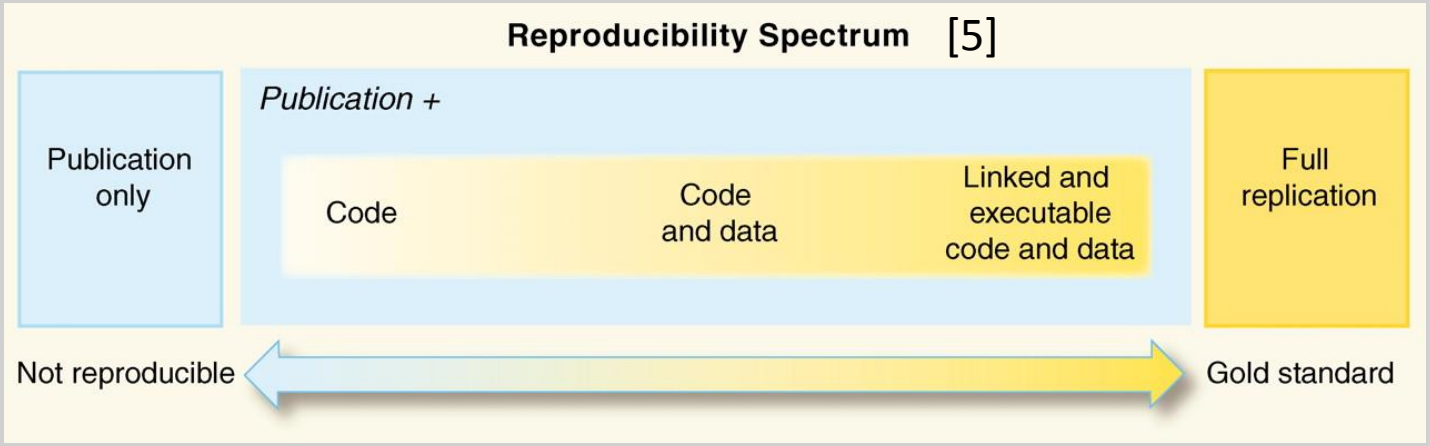
Replication Problems

- Not all studies can be easily replicated
- Unique Cases
  - Astronomic Observations
- Time Constrains
  - Studies that span decades
- Infrastructure
  - Big Data / HPC access
- Costs

Img source: https://www.nature.com/news/reproducibility-the-risks-of-the-replication-drive-1.14184

What is Reproducibility

- Bridges the gap between replication and stand-alone study
- Uses same code / data / methodology
  - Validate findings



Replication VS Reproducibility

- Focus:
  - Validity of Scientific Claim
- Asks:
  - “Is the claim true?”
- Reproduce results:
  - new investigators
  - New data, methods,
- Ultimate Standard for strengthening scientific evidence

- Focus:
  - Validity of data analysis
- Asks:
  - Can we trust this analysis?
- Reproduce Results
  - New investigators
  - Same data, methods,
- A minimum standard for any scientific work

Reproducibility Benefits

Among others

- Individuals
  - Easier to Reproduce your Research
  - Easier to onboard new researchers on the group
  - Easier to share research with other researcher

Reproducibility Benefits

Among others

- Community
  - Easier to disseminate results
  - Increase public trust in science
  - Able to assess the procedure of the analysis, not only the final outcome

Reproducibility Facets

- We can make our research reproducible by focusing on these aspects:
  - Documentation
  - Organization
  - Automation
  - Dissemination

INTRODUCTION TO JUPYTER

Connecting to VMs

- All of you should have gotten information about how to connect to your VMs
  - Username
  - Password
  - IP
- Windows Users: Putty
- Linux / MacOS Users: Terminal

Workshop Repository

- The material of the workshop is stored in GitHub and served under this website

https://gwdg.github.io/ssgoe2017/

Gitter

- A messaging platform
- Focused on developers using GitHub
- There is a room for the workshop repository
- You can access from Gitter button on the workshop website

Etherpad

- To share notes during the workshop

https://etherpad.gwdg.de/p/ssgoe2017

Q & A

References

References

- [1]: Monya Baker, 1.500 scientists lift the lid on reproducibility, Nature, https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
- [2]: The poor availability of psychological research data for reanalysis. Wicherts, Jelte M.; Borsboom, Denny; Kats, Judith; Molenaar, Dylan American Psychologist, Vol 61(7), Oct 2006, 726-728. http://dx.doi.org/10.1037/0003-066X.61.7.726
- [3]: Drug development: Raise standards for preclinical cancer research, C. Glenn Begley & Lee M. Ellis, Nature 483, 531–533 (29 March 2012) doi:10.1038/483531a
- [4]: Collberg, Christian, et al. "Measuring reproducibility in computer systems research." Department of Computer Science, University of Arizona, Tech. Rep (2014).
- [5]: Roger D. Peng, Reproducible Research in Computational Science, Science 02 Dec 2011: Vol. 334, Issue 6060, pp. 1226-1227 DOI: 10.1126/science.1213847