Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility

Victoria Stodden, Matthew Krafczyk, Adithya Bhaskhar

[1] - V. Stodden, M. Krafczyk, and A. Bhaskhar, “Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility”, P-RECS'18 Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems, 10.1145/3214239.3214242

Computational Reproducibility

- Claerbout Reproducibility [2]
  - Can another person reproduce published figures, tables and numbers using the same code/data.
- Benefits
  - Decreased overhead
    - Onboarding new group members
    - Collaborating with other groups
    - Reusing old code

[2] - J. Buckheit and D. Donoho. 1995. WaveLab Architecture. Technical Report. Stanford University, http://www-stat.stanford.edu/wavelab

Reproducibility in Various Fields

- Economics (1986)[3]: Only 15% of datasets documented
- Clinical (2005)[4]: 45/49 claimed effective, 20% replicated
- Epidemiology (2005) [5]: 93% did not detail data proc.
- Psychology (2006) [6]: 25.7% of datasets obtained
- Genetics (2009) [7]: Only 8 of 18 could be even partially replicated
- Cell Biology (2012) [8]: 6 of 53 biotech papers held up to scrutiny
- Economics (2015) [9]: 16 of 67 articles not replicated
- Psychology (2015) [10]: 83% had a larger effect in the original study
- Science (2018) [11]: 44% gave artifacts, 26% reproduced

[3] - Dewald et. al. “Replication in empirical economics: The journal of money, credit and banking project”, The American Economic Review, 76, no. 4, 1986
[4] - Ioannidis “Contradicted and initially stronger effects in highly cited clinical research”, JAMA, 294, no. 2, 2005
[5] - Peng et. al “Reproducible epidemiologic research”, American Journal of Epidemiology, 163, no. 9, 2006
[6] - Wicherts et. al “The poor availability of psychological research data for reanalysis”, American Psychologist, 61, 2006
[7] - Ioannidis et. al. “Repeatability of published microarray gene expression analyses”, Nature Genetics, 41, no. 2, 2009
[8] - Begley et. al. “Drug Development: Raise standards for preclinical cancer research”, Nature, 483, no. 7391, 2012
[9] - Chang et. al. “Is economics research replicable? sixty published papers from thirteen journals say ”usually not”,” Finance and Economics Discussion Series, 2015
[10] - Open Science Collaboration, “Estimating the reproducibility of psychological science,” Science, vol. 349, no. 6251, 2015
[11] - Stodden et. al. “An empirical analysis of journal policy effectiveness for computational reproducibility”, PNAS, 115, no 11

Computational Physics?

- Anecdotally:
  - Physicists are interested in ‘scientific reproducibility’
  - Some subfields are better about computational reproducibility than others.
- The Journal of Computational Physics
  - 306 articles from 10 issues. (322 - 331)
  - Classify articles based on information they contained
  - Email authors requests for necessary code and data
  - Attempt to reproduce article results with 4 hours of time.
- Classification
  - No discussion in the article, and no artifacts made available: 180 (58.8%)
  - Some discussion of artifacts, none made available: 109 (35.6%)
  - Some artifacts made available: 17 (5.6%)

Acquiring and using Code/Data

- Acquiring
  - 6 articles provided digital artifacts
  - 298 email requests for code/data were sent
  - Received code/data from 49 articles
  - Had code/data for 55 articles.
- Using
  - No Progress - 5 (9.1%)
  - Build Only - 2 (3.6%)
  - Run - 30 (54.5%)
  - Partial Replication - 18 (32.7%)
  - Full Replication - 0 (0%)

Reproducibility Evaluation

- Reproducible without wholesale re-implementation?
  - Difficult or impossible - 32 (58.1%)
  - Requires some work - 18 (32.7%)
  - Small Tweaks - 5 (9.1%)
  - Little to no difficulty - 0 (0%)
- Some ICERM [12] criteria
  - Necessary run parameters given - 47 (86%)
  - Documented use and licensing rights - 16 (29%)
  - Code publicly available with no download requirements - 15 (27%)
  - Details about test environment given - 13 (24%)
  - Some type of version control is used - 11 (20%)
  - Precise functions given with settings - 6 (11%)

[12] - Stodden et. al., “Setting the default to reproducible,” ICERM Workshop 2013 Report, https://icerm.brown.edu/topical_workshops/tw12-5-rcem/icerm_report.pdf

Discussion

- Discussion
  - Code and data are currently not easily available.
  - Code and data which was available was never enough to reproduce articles fully within 4 hours of researcher time
    - libraries released with the article's method implemented but lacking test cases and input data such as initial conditions
    - parameter specifications were missing
    - code had evolved to a new version
    - missing function definitions
    - visualization code was missing (or proprietary)
    - artifacts were provided for some but not all claims in the article

Future Work

- Future Work
  - Publications containing more details about our methods are forthcoming.
  - Extended reproduction efforts for collected code/data.
  - Production of compendia demonstrating proper packaging for some of these articles
  - How do we know how well a given work was replicated?
  - More specific guidelines to Journals