A Practical Taxonomy of Reproducibility for Machine Learning Research

Rachael Tatman
Kaggle
@rctatman
rachael@kaggle.com

Jake VanderPlas
eScience Institute
University of Washington
jakevdp@uw.edu

Sohier Dane
Kaggle
------
sohier@kaggle.com

Background

Why care about reproducibility?

- It's good science, but it also helps ML practitioners
- Research code is sample code; if people can't get it to run they can't apply your findings
- Reproducible code has a broader impact than non-reproducible code

"Reproducible" here means achieving the same results/output as the original paper using the same data*

* See [7] & [8] for a discussion of terminology: in many fields this is instead called "replicable"

Related work & contribution

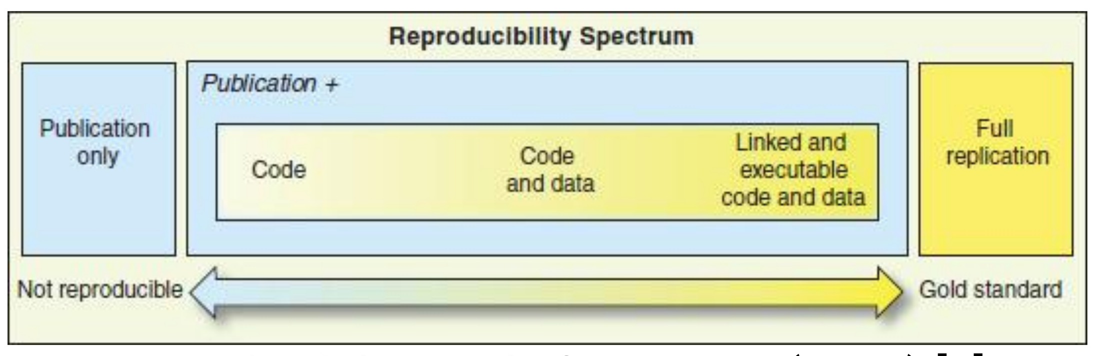
- There's been a lot of discussion of the importance of reproducibility [1,2,3,4,5,6]
- Here, we offer a practical framework for evaluating the reproducibility of a project and tips for improving reproducibility

As a scale

- Reproducibility is a spectrum, not a binary
- We propose an updated version of Peng (2011's) reproducibility scale (see below) with three levels (see sidebar →)
- The less time a reproducer needs to spend on a project, the more reproducible it is

Low Reproducibility

- Only sharing the paper
- Was standard before the ability to easily share code & data
- **Example**: Gelly & Silver 2007 [9]



Reproducibility scale from Peng (2011) [1]

Medium Reproducibility

- Sharing both code and data (if data was used, it should be anonymized & shared)
- Currently the most common way of sharing research code
- Still requires substantial time investment to get environment set up (need to account for versions and subversions of requirements)
- **Example**: "Lost relatives of the Gumbel trick", Balog et al 2017 [10]
- Some tips for improving medium reproducibility research:
    a. Separate preprocessing, modeling & evaluation and distribute data and code for each step
    b. Document the original environment
    c. Ensure that your code and data are licensed for reuse (see Morin et al. [11])

High Reproducibility

- Sharing data, code, and the environment needed to run the code
- Three options for sharing executable environments (in order of decreasing time commitment)
    a. Virtual machines
    b. Containers
    c. Hosted notebooks\scripts
- **Example**: "Understanding Black-box Predictions via Influence Functions", Koh & Laing 2017 [13]

High Reproducibility Options for Sharing Research

Virtual machines

- Code, data, all dependencies and a complete OS
- Excellent reproducibility, but files can be very large and slow to spin up
- **Popular options:** VirtualBox, VMware

Containers

- Code, data and all the dependencies needed to run the code in a single portable format
- Use the OS of the local system, so can be difficulties moving between OS's
- **Popular option**: Docker

Hosted notebooks\scripts

- Allows reproduction from a browser
- Generally faster and easier to run (don't require large downloads or set up time)
- Most services don't provide enough free compute to reproduce very computationally intensive studies
- **Commercial options**: Kaggle Kernels, Google Colaboratory, Amazon SageMaker, IBM Watson Studio, Azure Notebooks
- **Not-for-profit options:** MyBinder, PanGeo, Codalab

Other Considerations

Too few researchers share their code/data ☹

- <40% of papers from NIPS 2017 shared their code
- There were high reproducibility papers, through, like Liu et al [14]

Link rot is a big problem

- Code that is shared does not always remain available
- 20% of links in NLP papers were deprecated within 5 years [15]

Reproducibility Taxonomy

High:

- Code shared
- Data shared
- Environment shared

Medium:

- Code shared
- Data shared

Low:

- Finished paper only
- No code or data shared

[1] Roger D Peng. Reproducible research in computational science. Science , 334(6060):1226–1227, 2011.
[2] Victoria Stodden, Jonathan Borwein, and David H Bailey. Setting the default to reproducible in computational science research. SIAM News , 46(5):4–6, 2013.
[3] Victoria Stodden, Marcia McNutt, David H Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A Heroux, John PA Ioannidis, and Michela Taufer. Enhancing reproducibility for computational methods. Science , 354(6317):1240–1241, 2016.
[4] Justin Kitzes, Daniel Turek, and Fatma Deniz. The practice of reproducible research: case studies and lessons from the data-intensive sciences . Univ of California Press, 2017.
[5] Jonathan B Buckheit and David L Donoho. Wavelab and reproducible research. In Wavelets and statistics , pages 55–81. Springer, 1995.
[6] Jon F Claerbout, Martin Karrenbach, et al. Electronic documents give reproducible research a new meaning. In 1992 SEG Annual Meeting . Society of Exploration Geophysicists, 1992.
[7] Hans E Plesser. Reproducibility vs. replicability: a brief history of a confused terminology. Frontiers in neuroinformatics , 11:76, 2018.
[8] B Marwick, A Rokem, and V Staneva. Assessing reproducibility, 2017.
[9] Sylvain Gelly and David Silver. Combining online and offline knowledge in UCT. In Proceedings of the 24th international conference on Machine learning , pages 273–280. ACM, 2007.
[10] Matej Balog, Nilesh Tripuraneni, Zoubin Ghahramani, and Adrian Weller. Lost relatives of the Gumbel trick. In 34th International Conference on Machine Learning (ICML) , August 2017.
[11] Andrew Morin, Jennifer Urban, and Piotr Sliz. A quick guide to software licensing for the scientist-programmer. PLoS computational biology , 8(7):e1002598, 2012.
[12] Roger D Peng, Francesca Dominici, and Scott L Zeger. Reproducible epidemiologic research. American journal of epidemiology , 163(9):783–789, 2006.
[13] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In International Conference on Machine Learning , pages 1885–1894, 2017.
[14] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In Advances in Neural Information Processing Systems , pages 700–708, 2017.
[15] Margot Mieskes. A quantitative study of data in the NLP community. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing , pages 23–29, 2017.