Reproducibility Crisis with Computer Science

CANDIDATE NUMBER – 2, Universitetet i Tromsø, The Arctic University of Norway.

1 INTRODUCTION

Before the advent of scientific journals, science was more oriented towards a closed approach, so that only a restricted number of people could have access to it. Many scientists, including Galileo, Newton, Kepler, and Hooke, made their discoveries into something they could profit, and often papers were even encoded in anagrams. Following this closed science logic, it was difficult to identify ownership and priority of scientific discoveries, like for instance the debate on whom first discovered calculus between Leibniz and Newton.

In modern times, scientific research is moving towards openness and free knowledge, and ownership and priority issues are not the main concern anymore. However, there are still research data, journals, and paper which are not accessible to all. Data is not open due privacy and secrecy by for-profit organizations, and papers are not open because they are published in pay-walled journals. Merton arguments in favor of open science are that, knowledge-creation is more efficient if scientists work together, and it is morally binding on the professional scientist [7].

Besides seeking knowledge and discoveries, science needs also to constantly proof itself to be right, hence it needs to be reproducible. If scientists such as Newton, Kepler or Galileo, could guarantee the reproducibility of their discoveries by mathematical proofs, all included in one single paper, in modern science, and especially after the advent of computers and information technology, it is not possible anymore. The complexity and amount of data used to formulate theorems and proofs has drastically changed the way of doing research, having manuscripts, source code, and research data, as a fundamental part of the whole scientific discovery, in order to have a reproducible experiment.

Author's address: Candidate Number – 2, Universitetet i Tromsø, The Arctic University of Norway. Tromsø, 9019.

Being our field of research Computer Science (CS), and in particular application of Artificial Intelligence (AI) on big data in blockchain technology and cryptocurrencies, we want to enhance the main challenges of doing good scientific research when research data and source code are not openly accessible. Excessively closed research data can lead to reproducibility issues, especially when AI and big data analysis are involved. We will show in this manuscript the importance of open source code and open research data when it comes to the scope of making good science.

2 OPEN DATA AND SOURCE CODE

Information technology and big data digitalization together with statistical inference, open up a complete new way of doing research, where it is possible to train machines and algorithms to solve particular tasks without being given specific instructions; the so called Machine Learning (ML).

In modern science, even before AI and ML, scientific discoveries suffered of reproducibility due to selective reporting, selective analysis, or insufficient specification of the necessary condition to obtain the reproducible results [1], and as Gundersen [6] mentions, even if reproducibility is a cornerstone of science, there is a large amount of published researches which cannot be reproduced [2, 3, 9].

In CS, while developing algorithms, it suffices to share an abstraction of the code which mathematically explain its logic (*pseudo code*), in order to enable reproducibility. In AI and ML it is a bit more complex. In order to generate a ML model, big amount of data is needed, this data can be manipulated, normalized, and maybe just a variable change in the optimization function can trigger the outcome of the experiment. Reproducibility in CS and ML then becomes fragile, and because of this, there is more need of data and code openness. However, some researchers are not willing to share code and data [6] and in a study conducted by Colleberg and Proebsting [4], even in open access journals, such as Association for Computing Machinery (ACM), only 66% of the experimental papers were backed by code. Furthermore, only 32% of those were easily reproducible, enhancing even more the reproducibility crisis that CS and ML are facing.

3 CONCLUSIONS

REFERENCES

[1] Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., Babel, M., Bahnik, S., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F., and Penuliar, M. Estimating the reproducibility of psychological science. *Science 349* (08 2015).

[2] Begley, C. G., and Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature 483*, 7391 (2012), 531.

[3] Begley, C. G., and Ioannidis, J. P. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research 116*, 1 (2015), 116–126.

[4] Collberg, C., and Proebsting, T. A. Repeatability in computer systems research. *Commun. ACM 59*, 3 (Feb. 2016), 62–69.

[5] Crosas, M. The dataverse network®: an open-source application for sharing, discovering and preserving data. *D-lib Magazine 17*, 1 (2011), 2.

[6] Gundersen, O. E., and Kjensmo, S. State of the art: Reproducibility in artificial intelligence. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[7] Merton, R. K. Science and technology in a democratic order. *Journal of legal and political sociology 1*, 1 (1942), 115–126.

[8] Nakamoto, S., et al. Bitcoin: A peer-to-peer electronic cash system. *Working Paper* (2008).

[9] Prinz, F., Schlange, T., and Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery 10*, 9 (2011), 712.