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 scientific priority of research 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 technology has solved scientific priority issue to a certain extent. Many scientists are in favor of open science, for instance Merton argues that knowledge-creation is more efficient if scientists work together, and it is morally binding on a professional scientist [7]. However, there exists research data, journals, and papers which are not openly accessible. In some cases research data is not available due to privacy concerns. Research data collected by for-profit organizations is held in secrecy, for safeguarding commercial interests. Furthermore, pay-walled journals limit access of scientific publications. All this hinders the propagation of open science.

Besides seeking knowledge, science needs also to constantly proof itself to be right, hence it needs to be *reproducible*. In sixteenth and seventeenth centuries, scientists such as Newton or Galileo could guarantee reproducibility in physics by using mathematical formulas. In the late twentieth century, in the field of Computer Science (CS), *pseudo code*[1] helped reproducibility.

In twenty-first century, Machine Learning (ML) rapidly gained popularity [8]. ML relies on new techniques based on big data analysis and statistical inference. With ML it is possible to train machines to solve particular tasks without being given specific instructions. ML has

[1] pseudo code is the logic or abstraction which explains algorithms.

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

drastically changed the way of doing research [2], so that mathematical formulas and pseudo code are not enough to guarantee reproducibility, while research data, source code, and model's set up, became fundamental information in order to reconstruct the same outcomes of an experiment.

In this manuscript we want to enhance the main challenges related to reproducibility in ML. More importance should be given to open source code and open research data, in order to perform good science. Excessively closed research data can lead to reproducibility issues, especially when big data analysis is involved.

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 ML.

In modern science, even before Artificial Intelligence (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 [3, 4, 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 [5], 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] Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine 16*, 7 (2008), 16–07.

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

[4] 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.

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

[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] Mjolsness, E., and DeCoste, D. Machine learning for science: state of the art and future prospects. *science 293*, 5537 (2001), 2051–2055.

[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.