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 [12], 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 the manuscripts were even encoded in anagrams [12]. 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 [9]. However, there exist 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 prove itself to be right, hence it needs to be *reproducible*. In the 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 [10].

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.

[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.

ML has 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 REPRODUCIBILITY ISSUE AND OPEN SCIENCE

Scientific discoveries suffer of reproducibility due to selective reporting, selective analysis, or insufficient specification to recreate the expected results [1]. In ML, these specifications include tuning parameters for statistical models, which result to be crucial for reproducibility. Furthermore, tuning parameters is a very sensitive issue both in practical applications and in academic studies [5].

Besides there are still many published research outside the field of ML which cannot be reproduced [3, 4, 13], the degree of openness that ML requires in order to have a reproducible research, is making transparency and open science a key point for reproducibility. Howerver, some researchers are not willing to share code and data [8]. In a study conducted by Colleberg and Proebsting [6], even in open access journals, such as Association for Computing Machinery (ACM), only 66% of the experimental papers were backed by code and only 32% of those were easily reproducible. Following the study of Gundersen [8], conducted in other conferences, out of 400 algorithms presented, 54% included pseudo code, 30% included test data, and only the 6% included the source code.

Our interest in this manuscript relies on research data, and ML does not only require the raw dataset to be reproducible. Hence, the research needs to share also *training data*, *validation data*, *test data*, and *results*. The same survey from Grundesen [8], conducted on papers from International Joint Conferences on Artificial Intelligence Organization (IJCAI) and Association for the Advancement of Artificial Intelligence (AAAI), enhanced that only the training set was shared by the majority, with a score of 56%, while test, validation and the results got respectively a low score of 30%, 16%, and 4%.

Being blockchain technology and cryptocurrencies our domain, data are public, hence after performing a research in dataset search engines, such as Dataverse [7] or Google Dataset Search[1], we found the information we needed, openly available. However, experiments involving ML models over large datasets are time consuming and expensive to run. We derive that the main challenges of publishing open research data in ML are the excessive time for training and depositing big data, and the costs of running experiments to produce results and validation sets.

3 CONCLUSIONS

We agree that open science is beneficial in order to conduct good scientific research, and we have seen that to have full reproducibility in ML, research data must be open. To speed up the scientific process, it is fundamental to make available also bad data samples, even if they are often evicted from the research data published. However, even with open data, infrastructure costs and lack of time can hinder reproducibility. Encouraging the use of open research data is not enough to ensure a good use of it. For each subject of study, there should be national guidelines to assist researchers in warranted methods for open research data. Furthermore, we believe that open science methodology and approaches should be part of the young researchers' curriculum.

Even if the openness of research data will speed up the scientific process, while enabling reproducibility, there are still many cases where openness is not possible. For instance, regulatory agencies are generally obligated to protect the commercial value of the data collected by companies [11]. Information contained in datasets should also comply with the normative on privacy preserving, and data anonymization is often recommended. Furthermore, scientific researches commissioned by private companies will benefit in keeping the secrecy on their research data.

Finally, we think that is preferable, for the sake of science and progress, to keep a high level of openness. Disciplines such as ML should be transparent concerning source code and research data. We have shown that closed models or data, can slow down the scientific process, hence the progress.

[1] https://toolbox.google.com/datasetsearch

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] Birattari, M. The problem of tuning metaheuristics as seen from a machine learning perspective.

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

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

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

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

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

[11] National Academies of Sciences, E., Medicine, et al. *Principles and obstacles for sharing data from environmental health research: Workshop summary*. National Academies Press, 2016.

[12] Nielsen, M. *Reinventing discovery: the new era of networked science*. Princeton University Press, 2011.

[13] 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.