

# 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 openness in propagation of 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 summarized in papers. After the advent of information technology in last decades of twentieth century, to guarantee reproducibility, disciplines such as Computer Science (CS) required, in addition to mathematics, the logic or abstraction which explains algorithms (*pseudo code*).

In twenty-first century, a specific branch of CS, called Artificial Intelligence (AI), developed new techniques based on big data analysis and statistical inference, where it is possible to train

machines and algorithms to solve particular tasks without being given specific instructions. The latter is also known as Machine Learning (ML). With ML, the complexity and amount of data used to formulate theorems and proofs has drastically changed the way of doing research. In ML, mathematical formulas and pseudo code are not enough to guarantee reproducibility, and research data, source code, and model's set up became fundamental information in order to conduct good science.

In this manuscript we want to enhance the main challenges related to reproducibility in ML, and how much 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 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.