KNOWLEDGE DISCOVERY

A cautionary tale from the machine scientist

Machine reading and knowledge extraction methods can be used to mine the scientific literature and reveal the direction and robustness of discoveries. Such efforts now point to the importance of independent tests of reported claims.

Luís A. Nunes Amaral

During the Enlightenment, European elites discussed all the most recent books, plays and scientific discoveries at salons (Fig. 1). Nowadays, that would be an impossible task. Just keeping up with the ever-expanding multitude of experimental approaches, reagents and computational analysis protocols used for probing interactions among human genes is a herculean task. In this issue, Belikov, Rzhetsky and Evans[1] come to the rescue of the overwhelmed scientist and salon participant by demonstrating an automated machine scientist[2–5] that reads scientific research papers, extracts information about scientific claims, aligns them with high-throughput experiments, and provides a Bayesian update of current knowledge.

The study by Belikov et al.[1] is made possible by the availability of massive digital archives, and machine reading and extraction tools. Specifically, they use two previously developed approaches to extract claims about gene interactions from the literature. The first, GeneWays, contains about 496,000 unique claims about 313,000 unique interaction triplets — source gene, target gene and action — extracted from the full text of nearly 200,000 PubMed publications. For the second, Literome, they consider about 259,000 unique claims about 144,000 unique interaction triplets extracted from the abstracts of nearly 220,000 PubMed publications.

However, it is not just a question of adding positive and negative evidence on a ledger to determine whether a claim is valid. The meta-research literature clearly demonstrates that social and institutional factors have crucial roles in deciding which claims are tested, and having been tested, which results of those tests are published[6,7]. Indeed, the growth of both the scientific literature and the complexity of research methods has been accompanied by an increase in concerns about the reproducibility of published methods, results, and inferences[8]. Thus, Belikov et al.[1] test the literature claims against Library of Integrated Network-Based Cellular Signatures (LINCS) experimental results obtained on 77 cell lines, using various perturbation types, durations and dosages.

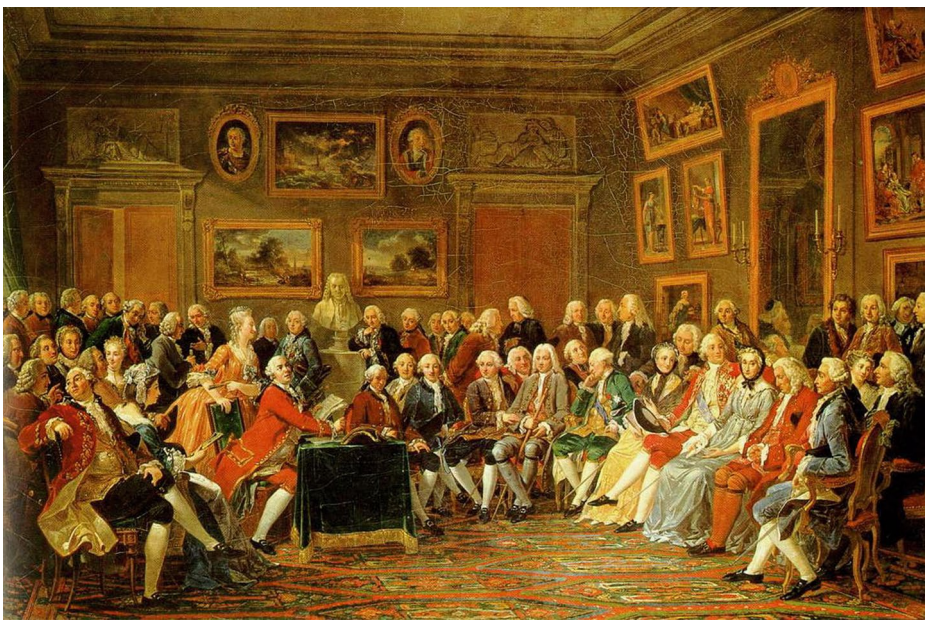

Fig. 1 | Enlightenment salon. Reading of Voltaire's tragedy of the Orphan of China in the salon of Marie Thérèse Rodet Geoffrin, by Anicet Charles Gabriel Lemonnier (1812). Credit: Art Collection 2 / Alamy Stock Photo

By connecting the reproducibility of a claim to characteristics of the authors of the publication from which the claim is extracted, Belikov et al.[1] can identify the characteristics of research communities that report claims that are more reproducible. As expected, increasing the number of researchers, communities and institutions studying a given phenomenon increases our knowledge about it. Perhaps less expected, a lack of independence between investigators, communities, institutions and previous knowledge markedly decreases claim robustness.

But maybe the latter should be expected too. Previous research from some of these authors reported that more centralized research communities are more likely to produce non-reproducible results[9]. So, maybe the main lesson to be taken from Belikov et al.[1] is not about how a machine scientist can help human scientists to keep up with the literature but how it can help us to identify strategies to counter worrisome trends in science.

As the immunologist Bruno Lemaitre observed a few years ago[10], "[p]ower struggles and ego battles are […] quite prevalent in the academic world, notably in our so-called elite institutions". The emergence of battle lines and greater inequality in the apportioning of funding is resulting in big-name driven big science. At present, science is looking more and more like the tech world; the focus is on extravagant rewards for the already fortunate, the constant replacement of technologies[11], and the exploitation of

armies of individuals occupying starting positions without any prospects of advancement[12].

So, what can be done about this situation? Some provocative ideas for actions are to cap the research funding that a single researcher can control and the number of manuscripts a researcher can submit annually. There could be limits on the length of time over which someone can receive funding on a given research direction. And rules on what constitutes authorship on a scientific paper could be more strictly implemented. Making use of the abilities of machine scientists such as developed by Belikov et al.[1] could help to identify the best approaches. ❐

Luís A. Nunes Amaral ✉

Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA.
✉*e-mail: amaral@northwestern.edu*

Published online: 10 May 2022
https://doi.org/10.1038/s42256-022-00491-7

References

1. Belikov, A. V., Rzhetsky, A. & Evans, J. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-022-00474-8 (2022).
2. King, R. et al. *Nature* **427**, 247–252 (2004).
3. Schmidt, M. & Lipson, H. *Science* **324**, 81–85 (2009).
4. Evans, J. & Rzhetsky, A. *Science* **329**, 399–400 (2010).
5. Guimerà, R. et al. *Sci. Adv.* **6**, eaav6971 (2020).
6. Goodman, S. N., Fanelli, D. & Ioannidis, J. P. A. *Sci. Transl. Med.* **8**, 341ps12 (2016).
7. Nissen, S. B., Magidson, T., Gross, K. & Bergstrom, C. T. *eLife* **5**, e21451 (2016).
8. Stoeger, T., Gerlach, M., Morimoto, R. I. & Amaral, L. A. N. *PLoS Biol.* **16**, e2006643 (2018).
9. Danchev, V., Rzhetsky, A. & Evans, J. A. *eLife* **8**, e43094 (2019).
10. Lemaitre, B. *FEBS J.* **284**, 875–882 (2017).
11. Amaral, L. A. N. https://amaral.northwestern.edu/blog/how-do-we-tell-whether-scientific-technology-has-r (2019).
12. Alberts, B., Kirschner, M. W., Tilghman, S. & Varmus, H. *Proc. Natl. Acad. Sci. USA* **111**, 5773–5777 (2014).

Competing interests

The author declares no competing interests.