Harvard Data Science Review • Issue 6.1, Winter 2024

Collective Intelligence and Collaborative Data Science

Joseph Salmon[1,2,3]

[1]**Institut Montpelliérain Alexander Grothendieck (IMAG), University of Montpellier, Montpellier, France,**

[2]**National Centre for Scientific Research (CNRS), Montpellier, France,**

[3]**Institut Universitaire de France (IUF), France**

Published on: May 24, 2024

DOI: https://doi.org/10.1162/99608f92.bf3d6b1d

I congratulate Professor David Donoho (2024) for this insightful article on the current and future state of data science. I cannot agree more with his vision of the field! But I must acknowledge that his research vision, including some early comments on reproducibility (Buckheit & Donoho, 1995), has deeply inspired me in my work over the years. That being said, I would like to share a few thoughts on the importance of collective efforts in data science, a point implicitly raised by Donoho in his article, but that I believe deserves more attention. Indeed, I believe that the collective aspect of data science is fundamental to its current success and development. Yet, it is often overlooked in the media or in the scientific literature, where the focus is rather on individual or 'artificial' achievements. I hope that these examples will help to illustrate the importance of collective efforts in data science and the need for more recognition of such efforts by our community. My thoughts are guided by two examples I am very familiar with, the Benchopt and Pl@ntNet projects.

First, I would like to thank Professor Donoho publicly for the positive feedback on Benchopt (Moreau et al., 2022), a benchmarking tool for optimization and machine learning, which I helped develop to provide frictionless reproducibility (FR). According to the terminology proposed, Benchopt is rooted in "FR2: Re-execution" and "FR3: Challenges," though the leaderboard is not always explicit. Concerning "FR1: Data," Benchopt offers to automatically load data sets in a unified application programming interface (API), but they are in general not hosted or created by the project itself. I would also like to clarify the following point: the project benefited from (French) public funding only for the computing power. As with many open source projects, the initial development was done voluntarily and collaboratively (more than 20 researchers), and initial costs were only indirectly covered by public funding through personal grants (including mine). More importantly, it builds on many other open source projects (NumPy in particular) leveraging a vast community effort. I still believe this kind of hybrid project, at the border between engineering, coding, and research, is essential for the community. Its collaborative nature is cumulative in time and could help reduce the reproducibility crisis (at least in optimization and machine learning) by providing a common ground for comparisons. This aspect is perfectly developed in the article by Donoho, under the name CORA, for Computing on the Digital Research Artifacts. It also echoes the attempts developed in the image-processing community, with the creation of *Image Processing On Line* (*IPOL*), a journal where each article contains a text on an algorithm and its source code, with an online demonstration facility and an archive of experiments. With Benchopt we went a step further by providing a common framework for comparing optimization algorithms for specific problems, and the algorithms are simply available in a unified API in Python (though multiple languages can be considered, such as R or Julia): this allows adding new algorithms and comparing them on a wide range of data and metrics 'frictionlessly.'

It is also important to be aware that such projects are not without costs, especially in long-term maintenance. Worse, they are still underrated in academia and are hard to fund directly. When successful, they can have a significant impact on the community, but their hybrid form is often not evaluated properly by our peers, especially for promoting young researchers (say for hiring positions, grants, etc.). I hope that this will change

in the future with more support from leaders in the field like Professor Donoho. Proper credits and citations, or improved evaluation criteria for such projects could be other ways to improve the situation. Improvements in the peer-review process for software and tools could also be beneficial. Interesting examples include the *Journal of Open Source Software* (*JOSS*) or the Machine Learning Open Source Software track in the *Journal of Machine Learning Research* (*JMLR*), which have been successful in this regard. A more recent example going one step further is *Computo*, leveraging literate programming. *Computo* is a journal created by the French Statistical Society and dedicated to the publication of computational and algorithmic contributions in statistics and machine learning. Note that *Computo* automatically reexecutes the code submitted by the authors, and assesses the quality of the provided code during the review process. I am looking forward to seeing such journals using tools like Benchopt to evaluate the reproducibility of the results submitted.

The second example I would like to share is the Pl@ntNet project. Pl@ntNet (Affouard et al., 2017) is a citizen science project for automatic plant identification through photographs, based on machine learning. Its main artifact is a mobile app that is free, user-friendly, and can be installed within seconds on a standard smartphone. It is a collaborative project that has been running for more than 10 years, collecting a large amount of data, which in turn has been used to train machine learning models. At the end of 2023, this participatory approach resulted in the collection of more than 20 million labeled observations (corresponding to approximately one billion images) by more than six million observers worldwide, belonging to nearly 46,000 species.

Pl@ntNet's success relies on recent improvements in computer vision (especially in deep learning), but also on recent technologies that have made it possible to collect and process large amounts of data: mainly smartphones and the internet, elements also raised by Donoho (2024) in his article. The possibility of gathering and leveraging feedback from millions of users is fairly recent to the human scale, and emerged with the internet: Wikipedia is a famous example of this trend in the natural language processing community. It has led to the development of a new kind of intelligence, which I would call *collective intelligence* rather than 'artificial intelligence,' relying on *collaborative data science*. This collaborative effort is often overlooked in the scientific literature. The creation of popular (labeled) data sets requires wide crowdsourcing efforts, relying either on a cheap workforce (e.g., with Amazon Mechanical Turk) or on voluntary effort (e.g., Pl@ntNet or Wikipedia), neither with much scientific recognition. This is for me the dark matter of collective intelligence, where the collaborative effort of many individuals is used to create a valuable resource for the community.

Another point resonates with elements raised by Donoho: this is the possible positive feedback loop that could occur in data science, leading to singularity. For instance, through the creation of the Pl@ntNet-300K data set[1] (Garcin et al., 2021), a short version of the Pl@ntNet data set, we aimed at offering the ecology community a data set so (empirical) machine learners could come and help to improve the learned models tailored for plant identification. Similar attempts are currently being made to extend the data set and expose its crowdsourced nature, providing one of the first crowdsourced data sets, and not simply a processed version. In practice, this means sharing labels from multiple users for each image, from expert to beginner ecologists.

In conclusion, my experience with these two projects aligns with the frictionless reproducibility vibes described by Donoho, and I am glad to see that the community is moving in the right direction by acknowledging such efforts. With this evolution, new challenges are emerging: how to evaluate, encourage, support, and fund such projects[2] and how to make them more sustainable are some of the questions that need to be addressed. More maintenance (for stronger tools / long-term support), more documentation (for welcoming a wider audience), more collaborative data (for solving more complex problems), and more recognition for such efforts are needed. Adopting open source models can also accelerate the innovation process by enabling broader collaboration and fostering knowledge sharing. This can also have a huge economic impact, and companies that rely on open source business models can benefit from this approach by attracting talent and developing products more rapidly and effectively.

Possible actions to encourage the whole scientific community to embrace frictionless reproducibility include:

- accepting collaborative/software papers in conferences on par with regular papers,
- promoting software development in the profiles of young researchers or even targeting it for job openings,
- specifically funding these aspects at the institutional level,
- incorporating these themes into master's training or summer schools,
- proposing more complementary doctoral activities[3] focused on code development to engage doctoral candidates,
- funding and organizing more hackathons or coding sprints to foster collaboration and sharing of knowledge, leveraging real-life interactions,
- et cetera

I hope this article will help raise awareness of these issues and encourage the community to address them robustly.

Acknowledgments

I would like to emphasize the importance of work and discussions with my colleagues Alexandre Gramfort, Thomas Moreau, Mathurin Massias, and Alexis Joly. They have been instrumental in the development of the ideas presented here. I would also like to thank Mathurin Massias for his feedback on a preliminary version of this text and for suggesting possible actions to accelerate frictionless reproducibility in academia.

Disclosure Statement

Joseph Salmon has no financial or non-financial disclosures to share for this article.

References

Affouard, A., Goëau, H., Bonnet, P., Lombardo, J.-C., & Joly, A. (2017). Pl@ntNet app in the era of deep learning. *ICLR: International Conference on Learning Representations*. https://openreview.net/pdf?id=eLYinD0TtIt

Buckheit, J. B., & Donoho, D. L. (1995). WaveLab and reproducible research. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and statistics* (pp. 55–81). Springer New York.

Donoho, D. (2024). Data science at the singularity. *Harvard Data Science Review*. Advanced online publication. https://doi.org/10.1162/99608f92.b91339ef

Garcin, C., Joly, A., Bonnet, P., Affouard, A., Lombardo, Chouet, M., Servajean, M., & Salmon, J. (2021). Pl@ntNet-300K: A plant image dataset with high label ambiguity and a long-tailed distribution. In J. Vanschoren and S.-K. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*. Curran Associates. https://openreview.net/pdf?id=eLYinD0TtIt

Moreau, T., Massias, M., Gramfort, A., Ablin, P., Charlier, B., Bannier, P.-A., Dagréou, M., Dupré la Tour, T., Durif, G., Dantas, C. F., Klopfenstein, Q., Larsson, J., Lai, E., Lefort, T., Malézieux, B., Moufad, B., Nguyen, T. B., Rakotomamonjy, A., Ramzi, Z., . . . Vaiter, S. (2022). Benchopt: Reproducible, efficient and collaborative optimization benchmarks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems 35 - 36th Conference on Neural Information Processing Systems, NeurIPS 2022* (pp. 25404–25421). Curran Associates.

Footnotes

1. https://zenodo.org/records/5645731 ↩

2. A recent effort on that road was proposed by the French National Research Agency with the Thématiques Spécifiques en Intelligence Artificielle (TSIA) call. ↩

3. Such activities are called ‘missions complémentaires’ in France, and are possible alternatives to teaching duties for PhD students. ↩

References

↩

- Buckheit, J. B., & Donoho, D. L. (1995). WaveLab and reproducible research. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and statistics* (pp. 55–81). Springer New York.

↩

- Donoho, D. (2024). Data science at the singularity. *Harvard Data Science Review*. Advanced online publication. https://doi.org/10.1162/99608f92.b91339ef

↩

- Garcin, C., Joly, A., Bonnet, P., Affouard, A., Lombardo, J.-C., Chouet, M., Servajean, M., Lorieul, T., & Salmon, J. (2021). Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution. In J. Vanschoren & S.-K. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*. https://openreview.net/pdf?id=eLYinD0TtIt ↩
- Moreau, T., Massias, M., Gramfort, A., Ablin, P., Charlier, B., Bannier, P.-A., Dagréou, M., Dupré la Tour, T., Durif, G., Dantas, C. F., Klopfenstein, Q., Larsson, J., Lai, E., Lefort, T., Malézieux, B., Moufad, B., Nguyen, T. B., Rakotomamonjy, A., Ramzi, Z., . . . Vaiter, S. (2022). Benchopt: Reproducible, efficient and collaborative optimization benchmarks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems 35 - 36th Conference on Neural Information Processing Systems, NeurIPS 2022* (pp. 25404–25421). Curran Associates.

↩