Harvard Data Science Review • Issue 6.1, Winter 2024

After Computational Reproducibility: Scientific Reproducibility and Trustworthy AI

Bin Yu[1]

[1]University of California Berkeley, Berkeley, California, United States of America

Published on: May 24, 2024

DOI: https://doi.org/10.1162/99608f92.ea5e6f9a

Introduction

Donoho (2024) proposes a convincing thesis that “frictionless reproducibility” (FR) in data science (which consists of three components: data sharing, code sharing, and benchmark data competition) has been a main driver for the recent AI advances. He also argues that it is data science “singularity” happening, not AI singularity. I found the article thought-provoking, and I very much enjoyed learning the perspectives of the article and the arguments provided for initiatives FR-1–3.

In this discussion, I argue for another important driving force for data science singularity, a large team with good team culture, and I point out that resources are needed for FR to ‘spread like a wildfire.’ Building on computational reproducibility, I advocate for scientific reproducibility and trustworthy AI through our predictability-computability-stability (PCS) framework and documentation. I believe AI trust and safety concerns are accelerated by data science singularity and they need to be addressed urgently. It is hoped that, after data science singularity, quality data, human knowledge, and relevant mathematical theory combined with empirical machine learning (EML) will achieve energy efficient AI development.

Another Important Driving Force for Data Science Singularity: Large Team With Good Team Culture

I agree that data science is an important driving force of AI and that FR-1–3 have been key for data science. However, I do believe there is at least one more equally important, if not more important, driving force for recent data-driven technological advances or data science singularity. That is the large-team human collaboration at a scale of hundreds of people. In other words, the acceleration comes not only from compute advances, but also from large human teams, organized in a company or self-organized. Before ChatGPT, OpenAI worked as a large team of hundreds of people, giving it a real advantage over the more distributed team cultures at other IT companies such as Google and Meta. Since ChatGPT, Google has reorganized in big AI teams of hundreds. Even for benchmark data competitions, the size of a team has increased greatly since the Netflix competition time. Science teams have also grown. For example, in astronomy, papers have been published with 100 to 200 authors. Of course, data sharing and code sharing and collaboration platforms such as Git have been indispensable for such human collaborations. Large-team human collaborations do not always follow from the three components of FR since.

For a large collaboration to succeed, it also needs a good team culture with fair and transparent leadership and a fair and transparent credit-sharing reward system. Some IT companies had had the three components for a long time, but did not produce ChatGPT, possibly due to the lack of large-team collaboration culture—this is not an uncommon belief among machine learning (ML)/AI researchers. In fact, for FR-1–3 to succeed, large teams are often necessary as well. The big-bench benchmark data set (Srivastava et al., 2023) to compare language models contains 2,000-plus tasks, contributed by 450 authors from 320 institutions. The open source and distributed processing system (codebase) for big data, Apache Spark (Sugob, 2018), had already had more than 1,000 contributors in 2018 from 250-plus organizations. The public ENCODE database (National Human Genome Research Institute, 2012) for functional human genomics came from hundreds of researchers in the United States, United Kingdom, Singapore, and Japan.

Resources Are Needed for FR to 'Spread Like a Wildfire'

As alluded to earlier, I completely agree that the three components FR-1–3 are driving forces for the recent advances such as ChatGPT, at least at top IT and other top computational companies. However, I would not think they are frictionless yet, in general, but rather 'less friction.' As much as I wish that "[FR-1]+[FR-2]+[FR-3] may spread across a field like wildfire," according to Donoho, how much of FR-1–3 are practiced differs, depending on the field and resources available.

High-stakes fields such as medicine and cybersecurity still have big obstacles to overcome to implement FR-1–3 due to privacy, national security, and legal concerns. Specifically, data sharing and benchmark data sets are very slow in coming to medicine and cybersecurity. Although progress is being made in medicine with Nightingale Open Science and MIMIC data sets, cybersecurity still suffers from the absence of benchmark data sets. Meanwhile, cyberattacks are already a main form of warfare in business and geopolitical frontiers. Even given a field such as chemistry, the practice of FR-1–3 is probably uneven due to limited human and compute resources in nonresearch universities and smaller companies, despite the examples given in the article. FR-1–3 will really spread like wildfires when FR-1–3 systematically enter our computational science education (e.g., coding, statistics, and data science courses) at the undergraduate level and across different universities with different resource levels, preferably even in high schools. FR-1–3 will really spread like wildfires when there are enough human and compute resources available at different organizations, including different colleges, high schools, companies, and government agencies.

After FR1-3, Scientific Reproducibility and Trustworthy AI Through PCS Framework and Documentation

As Donoho (2024) states clearly, the reproducibility discussed is computational reproducibility, which is different from scientific reproducibility (or replicability; National Academies of Sciences, Engineering, and Medicine, 2019). Computational reproducibility is necessary for the lamer, but not sufficient. Scientific reproducibility (or replicability) and trustworthy AI are naturally the next phase of data science singularity. Scientific reproducibility (or replicability) is about whether a scientific conclusion holds when a different lab collects data, possibly with a different tuning of a similar instrument, with similar but different sample preparations, and by different human researchers using domain knowledge and making judgment calls in the process. It needs empirical evidence and a serious check of reality and stability or robustness to the many reasonable perturbations in a data science life cycle (DSLC), for which domain knowledge and human judgment calls are also necessary. Scientific reproducibility (or replicability) and trustworthy AI require that the entire DSLC be examined (including problem formulation, data cleaning, and modeling choices through human judgment calls and domain knowledge), not just computational reproducibility, which is crucial to DSLC.

In fact, uncertainty quantification (or instability or irreproducibility) needs to be assessed for scientific reproducibility (or replicability) and trustworthy AI. The goal of science is trustworthy knowledge. Computational reproducibility serves this purpose and is a prerequisite. In medicine, the goal is trustworthy diagnoses and devices and safe and effective treatments and drugs for patients. In business, the goal is trust and safety of their products. Computational reproducibility in data science is thus also a prerequisite for data-driven medical AI and data-driven business AI. When broadly interpreted, scientific or trustworthy reproducibility is also the next phase for medical AI and business AI.

Together with my team, I have developed the PCS framework for veridical (truthful) data science toward scientific reproducibility and trustworthy AI (Yu, 2023; Yu & Barter, in press; Yu & Kumbier, 2020). FR-1–3 are necessary for "C" under PCS. "C" includes data-inspired simulations that also must adhere to FR-1–3. This is because data-inspired simulations must share data used, codes for model checking and simulation, and the simulated data, which could be used for benchmark data challenges as has happened already in ACIC (Atlantic Causal Inference Conference) data competition in 2022 (Mathematica, 2022; see also Carvalho et al., 2019). Moreover, PCS emphasizes documentation to record human judgment calls in a DSLC. Documentation is implicit in FR-1–3 since computational reproducibility's FR-1–3, code, data set, and benchmark data competition, all require adequate and accessible documentation. Good documentation for FR-1–3 describes human judgment calls to a certain extent (e.g., for codebase choices and data collection process and pre-processing choices for data sharing and benchmark competitions).

In the example of single-cell data analysis at the cutting edge of modern biology, FR-1–3 are adhered with Seurat. However, researchers might not use Seurat in R, and instead use Scanpy (2022) in Python with similar functionalities, depending on programming language choice. Let us assume one uses Seurat v.5; there are still many hyperparameters in Seurat that a researcher experiments with and makes judgment calls on. Depending on the aim of the data analysis, such hyperparameters could be about which highly variable genes to use, how many principal components to take, and which resolution in clustering to select.

The stability principle, or "S" in PCS, advocates an examination of the impact on the final scientific conclusions of human judgment calls (generally called perturbations in PCS) in the entire DSLC. Such perturbations include hyperparameter choices in Seurat or Python-Scanpy and choice of R-Seurat vs. Python-Scanpy. In fact, there are already five versions of Seurat. It would be an interesting scientific reproducibility and PCS exercise to check whether different versions of Seurat lead to the same scientific conclusions on the same data set studying the same biological problem. When such reasonable perturbations to the DSLC is deemed to lead to unstable or qualitatively different scientific conclusions, "S" in PCS advocates adding stability to the DSLC to achieve enough stability for the data conclusions to be stable and trustworthy or become scientific knowledge, while keeping “P” or prediction performance as much as possible (with “P” as a surrogate for reality check). Stability can be added through, for example, improving data collection process, standardization of hyper-parameter choices in a software package (or aggregation of results from different good packages), and improving data science or machine learning methods.

Data Science Acceleration Necessitates an Acceleration of Concerns for AI Trust and Safety

As made clear by the article under discussion, FR-1–3 has propelled an acceleration in data science. At the same time, I would argue that this acceleration or dazzling speed has also brought about more concerns about trust and safety for data-driven scientific knowledge, and about data-driven medical AI and business products. An evaluation process is indispensable to ensure trust and safety, but it takes time, such as in the FDA drug approval process. With enough human, compute, and financial resources, it is possible to speed up this evaluation process as we have seen in recent COVID-19 vaccine developments (from the traditional 18 months to 6 months).

However, there has not been enough human, compute, and financial resources to evaluate AI products in medicine, especially given the accelerated speed of their developments. The FDA process for AI medical algorithms is just beginning or not fully developed to say the least. An ever-expanding amount of investment money is pouring into acceleration of AI product development. It remains unclear whether a matching acceleration is possible in resources and where the resources will come from to ensure safe and trustworthy AI products.

It is particularly challenging to evaluate a constantly changing public-facing product such as ChatGPT since the source code is not open and ChatGPT today might be different from ChatGPT tomorrow. An independent evaluation of a public-facing AI product dictates that its producer needs to freeze a version to allow adequate testing or scrutiny (assuming evaluation resources are available) and before its release to the public. It is worth noting that the European Union and the United States both recently have made progress toward AI laws or regulations. The next few years will bear witness to the process of giving teeth to these AI laws and regulations. I believe FR-1–3, team efforts, PCS, transparent documentation of the DSLC for an AI algorithm development, and independent cohort validation are key components to ground them in the context of medical and other domain problems.

What Is Next: Quality Data, Human Knowledge, and Relevant Mathematical Theory Combined With EML

Two other important factors for scientific reproducibility and trustworthy AI are data quality control and human knowledge. The latter includes human inputs through reinforcement learning and external knowledge augmentation to generative models, for example. These factors are believed to be critical for ChatGPT in addition to the large-team culture at OpenAI. For the former, data cleaning is key since it could be a substantial source of uncertainty (Yu, 2023), and it is not yet frictionless. The first step toward frictionless data cleaning is to assess the uncertainty or instability arising from different reasonable data-cleaning choices (made with domain knowledge) and to employ the "S" in PCS to add stability and increase data quality, while documenting everything.

Last, but not least, I would like to bring up the topic of energy consumption associated with FR-1–3, data science, and deep learning or AI. FR-1–3 became possible due to the advances of computing technology, which depends on serious energy consumption. An article by Thompson et al. (2022) asserts that deep learning developments are "reliant on increases in computing power [....] Extrapolating forward this reliance reveals that progress along current lines is rapidly becoming economically, technically, and environmentally unsustainable." One effective way to meet this energy consumption challenge is to develop efficient algorithms for deep learning and beyond. *Relevant* mathematical theory can be very useful for building efficient algorithms when combined with EML, as in papers by Yang (2019) and colleagues (Yang & Hu, 2021; Yang et al., 2022) and a recent work from my group (Hayou et al., 2024) to substantially speed up pretraining and fine-tuning deep learning models, respectively. Hayou et al. (2024)would not have been possible without Amazon Web Services (AWS) compute credits through an Amazon Research Award.

Disclosure Statement

Partial support is gratefully acknowledged from NSF grants DMS- 2209975, 2015341, NSF grant 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI), the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639, and NSF grant MC2378 to the Institute for Artificial CyberThreat Intelligence and OperatioN (ACTION).

Acknowledgments

I would like to thank Maria Brbic, Jonah Langlieb, Sandrine Dudoit, and Ziad Obermeyer for helpful discussions.

References

Mathematica. (2022). *ACIC Competition*. https://acic2022.mathematica.org/#

Carvalho, C., Feller, A. Murray, J., Woody, S., & Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*, *5*(2), 21–35.

Donoho, D. (2024). Data science at the singularity. *Harvard Data Science Review*, *6*(1). https://doi.org/10.1162/99608f92.b91339ef

Hayou, S., Ghosh, N., & Yu, B. (2024). LoRA+: *Efficient low rank adaptation of large models*. ArXiv. https://doi.org/10.48550/arXiv.2402.12354

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press. https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science

National Human Genome Research Institute. (2012). *ENCODE data describes function of human genome.* National Institutes of Health. https://www.genome.gov/27549810/2012-release-encode-data-describes-function-of-human-genome

Srivastava, A., Rastogi, Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lwekowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A.,...Wu, Z. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=uyTL5Bvosj

Sugob, S. (2018, December 28). Apache Spark — The largest open source project in data processing. *Medium*. https://medium.com/hiredevops-org/apache-spark-the-largest-open-source-project-in-data-processing-403c35028208

Scanpy. (2022, March 31). *Scanpy – Single-Cell Analysis in Python*. https://scanpy.readthedocs.io/en/stable/

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2022). *The computational limits of deep learning*. ArXiv. https://doi.org/10.48550/arXiv.2007.05558

Yang, G. (2019). *Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation*. ArXiv. https://doi.org/10.48550/arXiv.1902.04760

Yang, G., & Hu, E. J. (2021). Tensor programs IV: Feature learning in ifinite-width neural networks. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 11727-11737). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v139/yang21c.html

Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, D., Pachocki, J., Chen, W., & Gao, J. (2022). *Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer*. ArXiv. https://doi.org/10.48550/arXiv.2203.03466

Yu, B. (2023). What is uncertainty in today's practice of data science? *Journal of Econometrics*, *237*(1), Article 105519.

Yu, B., & Barter, R. (in press). *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press. https://vdsbook.com/

Yu, B., & Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences*, *117*(8), 3920–3929.