DOI https://doi.org/10.30525/978-9934-26-397-2-12

CURRENT AND FUTURE PROBLEMS OF DATA INTEGRITY, QUALITY, AND SCARCITY

Vasyl Gorbachuk, Tamara Bardadym, Sergiy Osypenko
V. M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine
VGorbachuk@nas.gov.ua

Introduction. While Open Science is motivated by (big) data integrity, the quality scientific data might be exhaustible resources: data is the new oil [1]. A noteworthy analysis is the growing size of datasets used in machine learning (ML) for natural language processing and computer vision. To extrapolate such growth, one can use historical growth rates or start from dataset sizes that are optimal for projected computing resources (budgets) of the future.

Results. A study of data growth, based on estimates of the total stock of unlabeled data available on the Internet over the next decades, indicates a relatively rapid depletion of the stock of high-quality language data (likely by 2026) and a slow depletion of the stock of low-quality language data (likely by 2040) and image data (probably by 2045). Then the current trend of ever-expanding ML models that rely on very large datasets may slow down under modest increases in data quality (efficiency) and new data sources. The main factors that determine the performance of ML models are training data, algorithms, computations. Current understanding of the scaling laws by Open AI [2] and DeepMind [3] suggests that future ML capabilities will depend heavily on the availability of large volumes of data for training large models. EpochAI compiled a database of over 200 training datasets used in ML models and estimated the historical growth rates of datasets for language and vision models [4]. To learn about the limits of such rates (trends) in the future, EpochAI has developed probabilistic models for estimating the total volume of language and image data that will be available during 2022−2100. Based on the predictions of trends in dataset sizes in these models, it is possible to estimate the limits of such trends due to exhaustion of the data available. Data storage as the size of the Internet and the total amount of information available was estimated by the LightWave Networks Research Department and the Mathematics and Cryptography Research Department of AT&T Labs [5], Cyveillance Company [6], University of Berkeley [7].

Conclusion. The future scarcity of high-quality scientific data raises the issues of data values and data markets for quality data as exhaustible resources. On other hand, Science is becoming the new transdisciplinary industry.

References

1. Gorbachuk, V., Gavrilenko, S., Golotsukov, G., Nikolaevska, O., Nikolenko, D., Pustovoit, M. (2022). Infrastructural technologies of big and open data of the National academy of sciences of Ukraine. *Open Science and Innovation in Ukraine* (October 27−28, 2022, Kyiv, Ukraine). Kyiv : Ministry of Education and Science of Ukraine, 80−82.

2. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D. (2020). Scaling laws for neural language models, arXiv:2001.08361

3. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Anne Hendricks, L., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L. (2022). Training compute-optimal large language models, arXiv:2203.15556

4. Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., Ho, A. (2022). Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning, arXiv:2211.04325

5. Coffman, K., Odlyzko, A. (1998). The size and growth rate of the Internet. *First Monday*, 3 (10), https://doi.org/10.5210/fm.v3i10.620

6. Murray, B.H., Moore, A. (2000). *Sizing the Internet. White Paper*. Reston, VA : Cyveillance.

7. Lyman, P., Varian, H. R., Swearingen, K., Charles, P., Good, N., Jordan, L. L., Pal, J. (2003). *How much information?* 2003.

Berkeley, CA: School of Information Management and Systems; University of California at Berkeley.

Key words: datasets, extrapolation, machine learning, Internet, training models.

DOI https://doi.org/10.30525/978-9934-26-397-2-13

DEVELOPING A FAIR EDUCATIONAL ENVIRONMENT AND QUALITY CULTURE AT LESYA UKRAINKA VOLYN NATIONAL UNIVERSITY

Olena Halytska
Lesya Ukrainka Volyn National University
halytska@ukr.net

Introduction. In this abstract, we focus on the experience of my alma mater in promoting and implementing academic integrity among academic staff and higher education seekers.

Results. Thanks to the implementation of the project "Initiative for Academic Integrity and Education Quality", in which our university has been a participant since 2020, a series of initiatives have been launched: regular surveys of students, faculty, and university administration; annual Academic Integrity Week events at faculties since 2020; moderation of workshops, seminars, webinars; the use of programs to check master's theses for textual matches (StrikePlagiarism and Unicheck); lectures on "Academic Writing and Rhetoric" and "Academic Integrity."

A memorable online meeting took place on April 7, 2021, with members of the International Center for Academic Integrity, Camilla Roberts and Ann Domorad, who presented the Center and the fundamental values of academic integrity.

In December 2021, our University made a significant contribution to Ukraine's advancement in the Council of Europe, as Lesya