https://doi.org/10.1038/s43588-023-00499-y

Of data and transparency

Check for updates

While the increasing availability of data creates unprecedented research opportunities, it is important to understand the provenance of these datasets to ensure reliable data-driven conclusions.

The importance of the ever-growing amount of data being generated in the 21st century cannot be overstated. Most of the recent developments in computational science are deeply intertwined with this increasing data availability, as a plethora of computational tools and methods have been developed with the ultimate goal of analyzing and deriving new insights from existing datasets. It goes without saying that the popularity and progress of machine learning has taken the role of data to the next level. Undoubtedly, data has become a first-class citizen within many, if not all, areas of research.

With great power comes great responsibility though. Data is produced and collected in many different ways, and it is critical to be discerning when identifying which datasets should be used and when determining how they should be applied to a research project. Notably, understanding how these datasets were created and their potential biases goes a long way toward obtaining trustworthy research output.

There are different factors that must be considered before using a dataset. Some repositories may contain limitations that researchers must be aware of depending on the types of insights they need to derive from the data. As an example, the Materials Project contains computed information on known and predicted materials, but it is not always clear if the molecules or structures are experimentally validated, meaning that some of the molecules may not be synthesizable and chemically valid[1]. Some public data repositories for biological sequencing data have also been shown to include misclassified sequences due to errors in the metadata submitted by users, contamination errors in the biological samples, and limitations of computational methods[2]. Data repositories may also contain obsolete information or have been discontinued[3].

It is also well-established that datasets are not exempt from human bias. For instance, research has shown that datasets used for facial recognition are skewed towards certain races and genders[4]. In a particularly shocking instance, ImageNet — a dataset of human annotated photographs extensively used by researchers for developing computer vision algorithms — was found to have images labeled as 'loser', 'alcoholic' and even with racial slurs, which likely reflected the bias of the individuals labeling the figures[5]. Medical imaging datasets, which are now largely used for cancer classification and detection, have also shown alarming gender imbalance. Models trained on these medical imaging datasets have then gone on to show lower performance when tested on underrepresented groups[6]. Most ecological data repositories also suffer from geographical bias that arises due to researchers from specific areas majoritively contributing to these repositories[7]. Increasingly, fields such as wildlife conservation in ecology also use crowdsourcing for the development of data repositories, such as eMammal and Zooniverse, and these come with their own associated biases due to observer bias and contributions from non-researchers[8].

It is worth noting that researchers have put in concerted efforts to mitigate some of the biases that exist in these datasets. For instance, in medical datasets, biases observed with respect to gender and races can be potentially alleviated by oversampling from demographics that are underrepresented[9]. Further, checklists such as PROBAST[10] can be used to determine the level of bias that the use of these datasets entails. Several advances have also been made in the algorithmic space by the development of methods and constraints with the goal of ensuring that algorithms achieve equitable decision-making. Some of these constraints, however, may inadvertently worsen outcomes for individuals in marginalized groups, as discussed in a Perspective by Sharad Goel and colleagues. Needless to say, there is still a lot of work to be done in this area.

To ensure transparency when producing and using these datasets, data provenance is key. Those contributing to data repositories should ideally provide thorough documentation on how their datasets were collected and what their recommended uses are, along the lines of datasheets for datasets[11]. Researchers should adhere to the FAIR guiding principles by ensuring that their data is findable, accessible, interoperable, and reusable[12]. It is also imperative to recognize and document all of the known biases in these datasets, so that others are aware of the potential risks. But the responsibility is not only on those who create the datasets: those who use the data should also make sure that they understand that data to the best of their abilities. In addition, researchers should comply with certain guidelines when it comes to data citation[13]. For instance, datasets that are regularly updated may have version numbers associated with them that should be included in the citations.

A model is only as good — performance-wise, but also in terms of being unbiased and reproducible — as the data it is fed. While we are living in unprecedented times for data-driven research, we should take a step back and ensure that we are not just blindly collecting and using data.

Published online: 25 July 2023

References

1. Horton, M. K., Dwaraknath, S. & Persson, K. A. *Nat. Comput. Sci.* **1**, 3–5 (2021).
2. Bagheri, H., Severin, A. J. & Rajan, H. *Bioinformatics* **36**, 4699–4705 (2020).
3. Imker, H. J. *Data Sci. J.* **19**, 8 (2020).
4. Wu, W., Protopapas, P., Yang, Z. & Michalatos, P. In *12th ACM Conference on Web Science* 106–114 (Association for Computing Machinery, 2020).
5. Crawford, K. & Paglen, T. *AI & Soc.* **36**, 1105–1116 (2021).
6. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. *Proc. Natl Acad. Sci.* **117**, 12592–12594 (2020).
7. Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. *Ecol. Inform.* **19**, 10–15 (2014).
8. Caravaggi, A. et al. *Conserv. Sci. Pract.* **2**, e239 (2020).
9. Gosain, A. & Sardana, S. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* 79–85 (IEEE, 2017).
10. Wolff, R. F. et al. *Ann. Intern. Med.* **170**, 51 (2019).
11. Gebru, T. et al. *Commun. ACM* **64**, 86–92 (2021).
12. Wilkinson, M. et al. *Sci. Data* **3**, 160018 (2016).
13. Cousijn, H. et al. *Sci. Data* **5**, 180259 (2018).