The potential of benchmarks for the social sciences: Lessons learned and future work

Keywords: benchmarking;challenge; validation framework; mass collaboration; common task method

Extended Abstract

Social scientists aim at creating explanations of the world. For each social phenomenon, scientists have proposed a myriad of theories to explain its working mechanisms. Traditionally, we then translate each theory into a statistical model and assess the significance of the coefficients of the model. However, this predominant focus on null hypothesis testing prevents us from answering a crucial question: how can we compare models to determine which one works better under which circumstances? Without a framework to compare different models and approaches, we are unable to evaluate conflicting theories (Watts, 2017) or monitor progress in the social sciences. Here, we argue that *benchmarking* can be used in the social sciences as such a standard frame of reference.

For computer scientists, benchmarks provide a standardized way to validate predictive models and have been shown to lead to major breakthroughs, e.g., benchmarking has facilitated the rapid adoption of deep learning in a range of fields (Russakovsky et al., 2015).

In short, a benchmark (or benchmarking challenge) is when various teams of participants are invited to predict a particular outcome using a training dataset which includes the target variable (i.e., the ground truth) and a range of predictors. The performance of the teams' models is evaluated by predicting values of the target variable in a holdout dataset which the participants do not have access to. The result is one or multiple leaderboard(s) presenting a ranking of the prediction performance of the models based on a predefined set of metrics. The described benchmarking setup is illustrated in Figure 1.

While benchmarking challenges are popular in data science, machine learning, and other fields, there have been very few benchmark challenges in the social sciences. The most well known social science benchmark to date is the Fragile Families Challenge (Salganik et al, 2019). In this challenge, participants had to predict life outcomes at age 25 using early life predictors from a birth cohort survey. A key finding of this project was that machine learning approaches were not more effective at predicting these life outcomes than human designed approaches, using this benchmark data. This is an important finding, since it was shown at scale. However, we argue that it is important to carefully consider benchmark design. The Fragile Families Challenge could be seen as an initial benchmark providing insight into how follow up benchmarks on this topic should be designed. For example, the predictors used as benchmark data had already been filtered through the theoretical lens of the survey design, with the only concepts captured being those that the survey designers had already identified as being theoretically pertinent. This might have introduced a bias towards human designed approaches, which would be interesting to investigate in a follow up benchmark.

We contend that benchmarks in the social sciences have large potential for answering long standing questions in the field and demonstrating the value of machine learning approaches for the social sciences. However, careful consideration should be given to benchmark design in terms of data, metrics, and ground truth.

While we outline this potential, as well as specific areas and questions where a benchmarking approach could provide significant progress, we also contend that realizing this potential requires a drastic recalibration of data infrastructure and workflows, and that until this is achieved the value of benchmarking for social research will remain unknown. For benchmarking to have the possibility of reaching its potential in the social sciences, broader spectrum data is likely required and the 'feature selection' component needs far greater degrees of freedom. When such greater degrees of freedom are introduced, the marginal added value of machine learning and automated workflows over human driven model design will become evident.

Whilst we are unable to fully test this assertion with existing data infrastructure, we demonstrate the potential of such an approach through a benchmarking challenge utilizing administrative data in the Netherlands and the ODISSEI infrastructure. The aim of the challenge was to predict precarious employment (defined using a combination of income level and contract type) in 2020 using predictors from 2010 or earlier. The challenge made use of the administrative microdata available from Statistics Netherlands (CBS). This data provides thousands of data points on each individual, on a diverse range of measures, and can be operationalized in a myriad of ways. The use of CBS microdata was based around the idea that existing benchmarking challenges in the social sciences had used data with a limited spectrum, such as the Fragile Families data, which was collected via surveys. The relational data structures of the administrative data at CBS greatly increase the degrees of freedom for researchers attacking a particular challenge and may therefore be better suited to the use of machine learning and benchmarking approaches.

We will share lessons learned from this experience and elaborate on how we will use them when organizing a new benchmark focusing on fertility prediction in the coming months. We will emphasize the importance of a relevant research problem that is high on the research agenda of many researchers in the field. It is also crucial to clearly define a purpose statement for the benchmark and state how the data, metrics, and ground truth address the purpose statement, as this helps to provide insights into how the benchmark and its outcomes can help to answer the research question at hand.

Our overarching goal is to provide a framework for the future implementation of benchmarking challenges within the social sciences.

References

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the Special Collection on the Fragile Families Challenge. Socius: Sociological Research for a Dynamic World, 5, 237802311987158. https://doi.org/10.1177/2378023119871580

Watts, D. J. (2017). Should social science be more solution-oriented?. Nature Human Behaviour, 1(1), 0015. https://doi.org/10.1038/s41562-016-0015

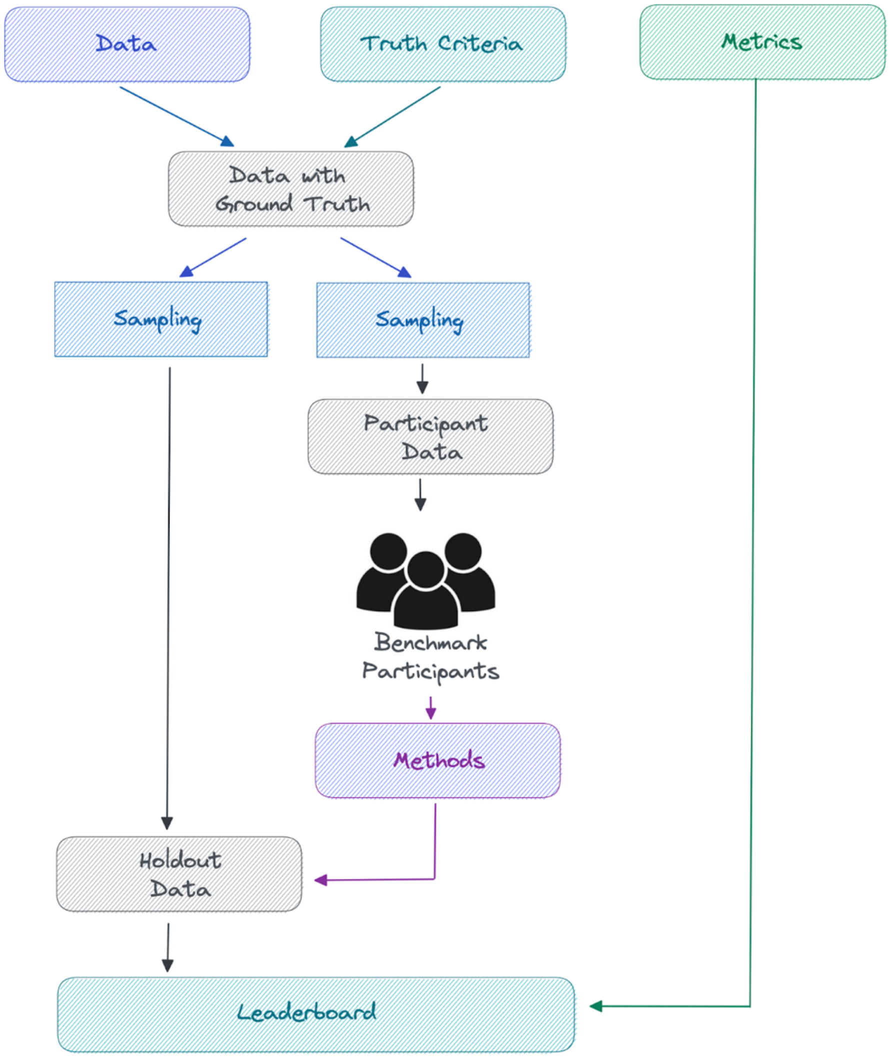

Figure 1: An illustration of a benchmark setup