Semi-supervised learning (SSL) is a branch of machine learning focusing on using not only labeled data samples, but also unlabeled ones, in an effort to decrease the need for labeled data and thus allow using machine learning even when labeling large amounts of data would be too costly. Despite its quick development in the recent years, there are still issues left to be solved before it can be broadly deployed in practice. One of those issues is class distribution mismatch. It arises when the unlabeled data contains samples not belonging to the classes present in the labeled data. This confuses the training and can even lead to getting a classifier performing worse than a classifier trained on the available data in purely supervised fashion.

We designed a filtration method called Unfavorable Data Filtering (UDF) which extracts important features from the data and then uses a similarity-based filter to filter the irrelevant data out according to those features. The filtering happens before any of the SSL training takes places, making UDF usable with any SSL algorithm. To judge its effectiveness, we performed many experiments, mainly on the CIFAR-10 dataset. We found out that UDF is capable of significantly improving the resulting accuracy when compared to not filtering the data, identified basic guidelines as to when to use it and discovered an important property of the class distribution mismatch issue.