Supporting Sustained Learning in Data Science with an Open Science Platform

February 7, 2019

Data-centric disciplines like Machine Learning and Data Science have become major research areas in the field of computer science. The data-centric, quantitative, and empirical nature of Machine Learning, the increasing complexity of experimental setups and number of established research lead to a high demand for supporting tools. However, the development of current research processes and tools could not keep pace with the rapid advancement of the disciplines. This led to a "crisis of reproducibility". Replicability, comparability and understandability require structured approaches and comprehensible presentation. In our view to establish an open and transparent research process the implementation and enforcement of core open science principles is mandatory. Research paper based knowledge sharing reaches its limitation as soon as procedures are to be compared on details, which are omitted to fit the papers format. While a wide range of functionalities is offered by the open source community, the ecosystem lacks a comprehensive system supporting the full stack of Machine Learning research. To tackle these challenges we propose a platform capable of tracking and managing experimental execution on a fine grained level. One of the main pillars of the platform design is to enable the generation of Structured Results, which provide the necessary qualities for proper tool support. Structured Results store input data, metrics as well as information about single decisions on instance level of Machine Learning experiments. Additionally the platform should support a link between publications and their relevant formation properties, as well as feature storage of semantically enriched data in an resource description framework knowledge base. The **P**latform for M**a**chine Learning and **D**ata Science **Re**poducibility (PaDRe) is being implemented abiding to the Open Science Process Model for Machine Learning (OSPMML). We developed the OSPMML to bring together open science principles, like the sharing of research artefacts as first class citizens and Machine Learning requirements in a common description. Our project aims to enhance the whole process from a State-of-the-Art analysis to the quality control serving as managing environment. By means of formalising domain knowledge incorporating algorithms, empirical conclusions, and task definitions we hope to further facilitate automatic evaluation, setup recommendations, and accessibility of research results across ecosystems.