Rapid development of science and technology causes an astonishing increase in volume and complexity of generated data. This growth can be observed in various areas, such as computer science, medicine, but also in basic science like particle physics and astronomy. Nowadays, a single dataset can contain many petabytes of information – hence the term *big data* which has appeared in the literature in recent years. The proliferation of *big data* poses new challenges to be addressed within data processing units, such as data quality assurance and visualisation.

One of the best examples of existing frameworks that generate massive amounts of data is a computer system used in the research facility of European Organisation for Nuclear Research - CERN. The key component of the research done at CERN is the Large Hadron Collider (LHC) – one of the most fruitful research collaboration projects in the world. Recently, LHC has gained international attention thanks to the discovery of famous Higgs boson, the missing link in the Standard Model of Particle Physics. This discovery led to awarding the Nobel Prize in Physics in 2013 to two physicists – François Englert and Peter Higgs – for their work postulating the existence of this particle in the 60s last century, over 50 years before the particle was observed.

Beside the above mentioned experiment, LHC also provides massive amounts of data within the frames of other experiments. The third most prominent experiment in terms of research efforts is ALICE - A Large Ion Collider Experiment. The group from Warsaw University of Technology is involved in the ALICE scientific collaboration and we envision a close collaboration with them throughout the project. ALICE is specifically designed for the study of heavy-ion collisions which is believed to simulate the conditions existing seconds after the Big Bang.

The works proposed in this project will be carried out in collaboration with CERN, precisely because of the massive amounts of data generated during experiments such as ALICE, that have to be processed, monitored and visualised. In the case of ALICE, almost 24 PB of data have been collected since 2009. This environment can be therefore considered a model scenario for a computer system that processes *big data*, i.e. large volumes of highly complex and ever changing data. More precisely, the data collected in the ALICE experiment contain hundreds of thousands of parameters coming from variety of sensors in each of the 18 subdetectors of the experiment. To monitor the quality of the collected data, several field experts inspect visually statistical distributions of registered events, and the visualisation systems they use display only a small portion of the collected data.

In this project, we propose to extend the capabilities of the data quality assurance system using machine learning algorithms and create a new toolkit of visualisation methods. We claim that machine learning can lead to significant reduction in the manual work performed by human experts and focused on tedious comparisons of data histograms. Building up on the historical data gathered during past experiments, we plan to leverage the existing classification and regression methods such as neural networks and boosting methods to monitor the quality of data generated in the experiments. To perform the analysis of the quality in real time we envision developing novel, more efficient methods that could potentially be used also outside of CERN.

Furthermore, the new visualisation systems proposed in this project will allow for more effective presentation of the experiment results as well as the environment conditions, such as 3D distributions of magnetic and electric fields within the detectors. We plan to implement visualisation methods based on virtual reality tools, such as 3D goggles and 3D displays. This approach significantly extends currently existing visualisation methods that rely on 3D to 2D data mappings and, therefore, lead to tremendous oversimplification of data structure. Moreover, the resulting visualisation system can also be useful for searching through the datasets for undiscovered patterns.

Finally, the visualisation tools developed within this project can be employed to communicate with the general public and help them understand complex physical nature of experiments run at LHC. These tools can also be used by students at schools and universities during science classes to explain physical effects of complex phenomena in simple and comprehensible manner.