IRIS-HEP Summer 2023 Fellow Program Project Proposal

# Predict CMS data popularity to improve its availability for physics analysis

**Applicant:** Andrii Len

**Mentors:** Dmytro Kovalskyi, Rahul Chauhan, Hasan Ozturk

**Project duration:** 12 Weeks

**Proposed start date:** 3 July 2023

## Introduction and Project Description

The CMS [1] data management team is responsible for distributing data among various computing centers globally. However, due to limited disk space at these centers, it becomes necessary to dynamically manage the data available on disk. When data is not readily available on disk, users have to wait for it to be retrieved from permanent tape storage [2], causing delays in data analysis and hindering scientific productivity. To overcome this challenge, this project aims to develop a tool that utilizes machine learning [3] algorithms to predict which data should be retained on disk based on current usage patterns. By leveraging machine learning techniques, it should be possible to analyze historical data usage patterns and identify trends or patterns that indicate the popularity or likelihood of future data access. This predictive capability allows the CMS data management team to proactively allocate disk space to the data that is more likely to be requested, ensuring faster access for users and minimizing the need for retrieving data from tape storage.

The tool that is to be created will involve several key components:

1) Data Collection: Consists of the gathering of raw data from various sources such as a summary of user analysis jobs. Data should include which datasets are accessed, when they are accessed, the frequency of access, etc. This data will serve as the training set for our Machine Learning models.

2) Feature Engineering [4]: Identifying relevant features or characteristics that can help predict data popularity. These features could include variables such as the size of the dataset, the historical access pattern, the type of data, or any other relevant metadata associated with the dataset.

3) Machine Learning Algorithms: We will need to use Recurrent Neural Networks [5] (LSTM, Transformers etc) to train on the collected historic data and predict the popularity of datasets. I will need to experiment with different algorithms to find the most accurate and efficient solution.

4) Model Evaluation: Assessing the performance of our machine learning models using appropriate evaluation metrics. This step ensures that our predictions are reliable and can guide data management decisions effectively.

5) Integration and Deployment: Integrating the developed tool into the existing CMS data management infrastructure so that it can continuously monitor and update data popularity

predictions.

## Software Deliverables

In this work, we will use Python for Machine Learning, especially the Tensorflow, PyTorch and scikit-learn packages to build, optimize and train our models.

## Preliminary Timeline

**Week 1-3** Aggregating the raw data to extract data usage information. Transferring obtained data into the Python environment and getting familiar with it.

**Week 4-7** Learning about Feature Engineering and using the gained knowledge in practice.

**Week 8-9** Experimenting with different models and testing them with small datasets.

**Week 10** Training model with the full CMS data. Predict the probability that a dataset will be accessed in the next month.

**Week 11-12** Adding mechanisms that will allow the algorithm to monitor real-time trends with a corresponding predictions improvement and summarising results as a short report.

## References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[2] Natalia Ratnikova | WLCG pre-GDB Storage | CMS staging from tape.

[3] Alex Smola and S.V.N. Vishwanathan, "Introduction to Machine Learning", Cambridge University Press 2008.

[4] Alice Zheng & Amanda Casari, "Feature Engineering for Machine Learning", Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[5] Tingwu Wang, Machine Learning Group, University of Toronto, "Recurrent Neural Network", For CSC 2541, SPORTANALYTICS