Thesis project and overview of the 1st year of PhD

1st seminar of the PhD school in Data Science and Computation

Luca Giommi (Univ. of Bologna and INFN) E-mail: *luca.giommi3@unibo.it*

Supervisors: Daniele Bonacorsi and Claudio Grandi

Istituto Nazionale di Fisica Nucleare

1

- **Who am I?**

I am a PhD student in Data Science and Computation of the XXXIV cycle. I am associated with the National Institute for Nuclear Physics (INFN) and CERN. My research work is within the CMS-Bologna group.

- **What is CERN?**

The European Organization for Nuclear Research (CERN) was born in 1954 and it is based in the northwest suburb of Geneva on the Franco–Swiss border. At CERN is located the world's largest and most powerful particle accelerator called Large Hadron Collider (LHC). Inside the accelerator, two high-energy particle beams travel at close to the speed of light before they are made to collide at four locations, corresponding to the positions of four particle detectors: ATLAS, CMS, ALICE and LHCb.

- **What is CMS?**

The Compact Muon Solenoid (CMS) is a general-purpose detector at the LHC. It has a broad physics programme ranging from studying the Standard Model (including the Higgs boson) to searching for extra dimensions and particles that could make up dark matter.

- **What is CNAF?**

CNAF is the biggest computing facility of INFN. Located in Bologna, since 2003 CNAF has hosted the Italian Tier-1 data center of the World-wide LHC Computing Grid, providing the resources, support and services needed for data storage and distribution, data processing and analysis, and Monte Carlo production.

2

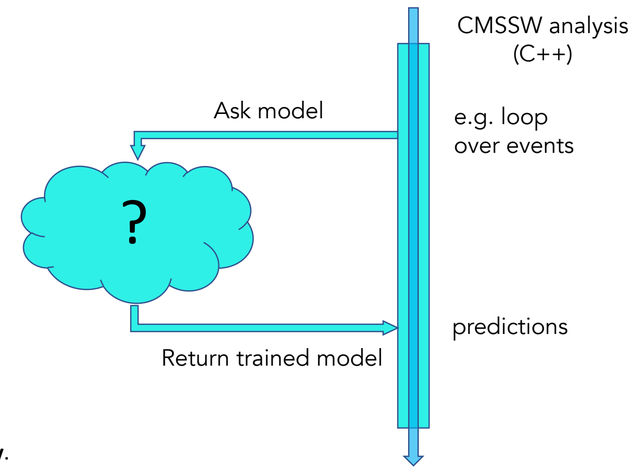
About my thesis project… where we were: master thesis

A full proof-of-concept demonstration of an end-to-end data service to provide trained Machine Learning (ML) models to the CMS software framework (CMSSW) and its usage in Signal/Background (S/B) discrimination in $t\bar{t}$ selection

ML "as a service" for CMS physics

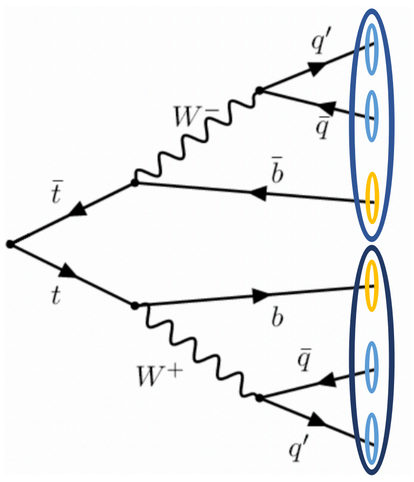
Instead of asking each physicists who wants to exploit ML in their own task to just learn how to do it independently, each user would build a modified data analysis code where **"calls" to an external service** - as simple as calls to functions - would be implemented to return a trained ML model output that could be directly used in the analysis code (e.g. in loops over events) in a streamlined manner.

This hides all the complexity related to the ML machinery.



3

From master thesis…



$t\bar{t}$ use case

the Signal versus Background discrimination in the selection of $t\bar{t}$ events in the all-hadronic topology

Comparison between ML techniques and ROOT MVA

scikit learn + K + TensorFlow

ROOT Data Analysis Framework TMVA

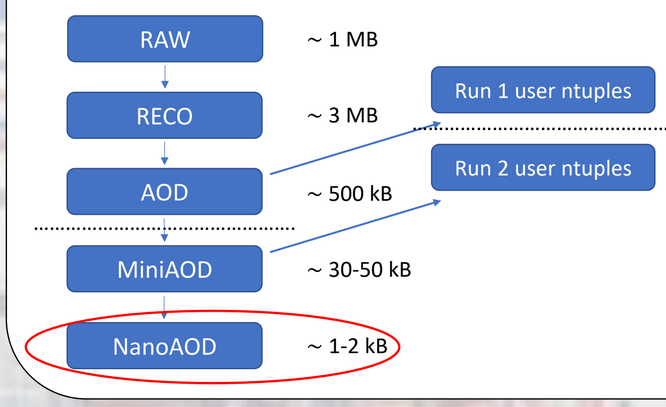
… to PhD thesis project

Exploit the potential of the "Data Streaming Layer"

Use distributed NanoAOD datasets (accessible via xrootd) without intermediate step of converting data into CSV format, pre-processing, etc.

In this way we could easily read (through **uproot**) and directly use O(10 TB) of datasets in the training of the ML models

CMS event data formats



4

MLaaS architecture diagram

It is responsible for reading local and/or remote ROOT files.

It represents a thin wrapper around standard ML libraries such as TensorFlow, Keras and others. It transforms data from ROOT TTree based representation to the flat data format suitable for underlying ML framework and uses it for training purposes.

It refers to the inference part of pre-trained models and can be either tightly integrated within underlying HEP framework or represented as a Service (aaS).

5

Why NanoAOD?

- It is a flat ntuple format with only standard data types (e.g. int, float, vectors), so it is simple to export to modern machine learning frameworks
- Many CMS analysis are switching to the new 1kb/event format

Goals

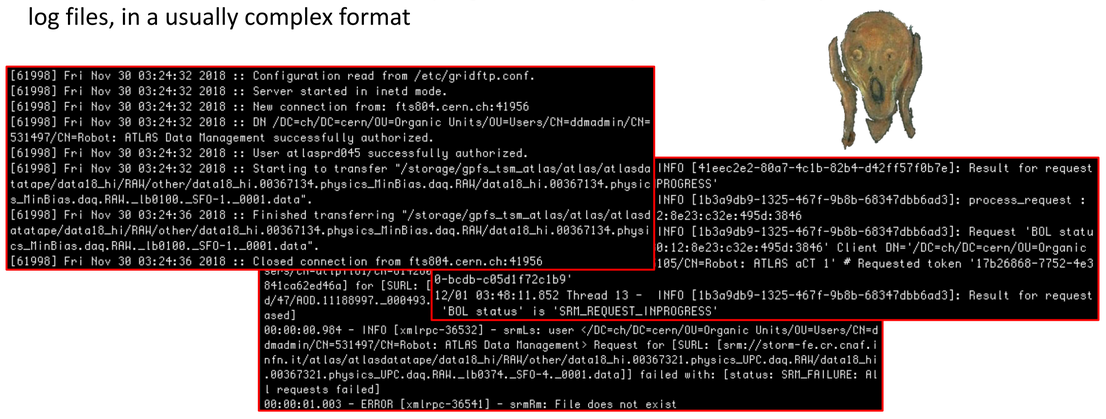
- We don't aim to reproduce an entire analysis with all details, but to demonstrate its feasibility using NanoAOD ROOT files and the MLaaS4HEP framework.
- Performance benchmarks (CPU vs GPU vs TPU, and various versions) for the training phase

6

Project at CNAF

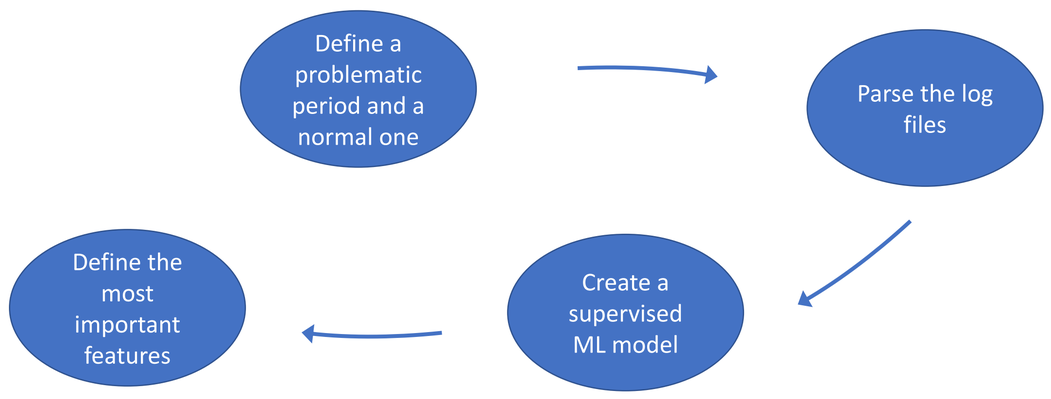
In order to increase efficiency and to remain competitive in the long run, CNAF is launching various activities aiming at implementing a global predictive maintenance solution for the site. Because of efficient storage systems are one of the key ingredients of Tier-1 operations, at CNAF an exploratory work started by investigating logs from the StoRM service.

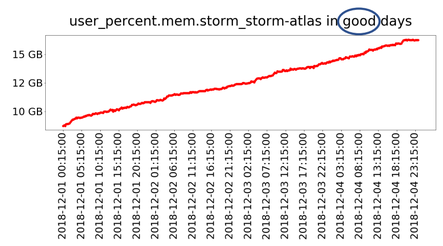
Information about the status and the progress of the requests managed by the service is stored in log files, in a usually complex format

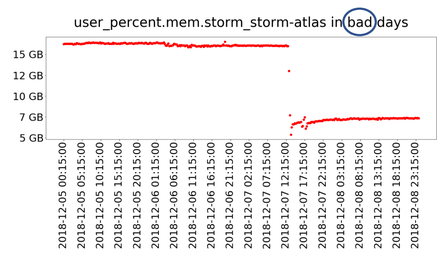


7

Work presented at the ISGC International Conference (Taipei, April 2019)







8

Overview of the 1st year of PhD

Compulsory CFU = 28 over 24 expected Free CFU = 29 over 16 expected

Attendance at conferences and schools	Location
International Symposium on Grids & Clouds (ISGC) 2019	Taipei (Taiwan)
3rd IML Machine Learning Workshop	CERN (Geneva, CH)
Workshop CCR	La Biodola (Italy)
LHCb/CERN & Microsoft Azure OpenHack	Trento (Italy)
Third international School on Open Science Cloud (SOSC)	Bologna (Italy)

References

[1] L. Giommi, *Prototype of ML "as a Service" for CMS Physics in Signal vs Background discrimination*, Master thesis and *PoS LHCP2018 (2018) 093*
[3] V. Kuznetsov, *Machine Learning as a Service for HEP*, arXiv:1811.04492 [hep-ex]
[4] MLaaS github page, https://github.com/vkuznet/MLaaS4HEP
[5] L. Giommi et al, *Towards Predictive Maintenance with ML at the INFN-CNAF computing centre*. Submitted to the proceedings of ISGC 2019, Taipei