

Objectives and motivation

This project aims to implement new Analysis Grand Challenge versions with ROOT's modern analysis interface RDataFrame (RDF) [1]. The Analysis Grand Challenge (AGC) [2] is organized by the Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP) [3] to address the computing challenges of the HL-LHC.

The AGC has two major pieces [4]:

1. specification of a physics analysis using Open Data that captures relevant workflow aspects encountered in physics analyses performed at the LHC,
2. implementation demonstrating the successful execution of this physics analysis at scale.

The physics analysis task we will focus on in this project is a top-antitop production cross-section measurement with 2015 CMS Open Data [5]. The previous RDF implementation [6] included data reading, event selection, and plotting observables with systematic variations.

The main updates that the latest AGC versions bring are the following:

- switching to input datasets with a more realistic schema, namely CMS NanoAOD
- handling systematic variations with new dedicated package `correctionlib` [7]
- adding machine learning inference to the analysis pipeline

Milestones

1. Studying new AGC versions
2. AGC 1.0.0 version implementation with NanoAOD data schema
3. Implementation of the ML aspects of AGC version 2.0
4. Integration of the `correctionlib`-based systematic variations

Timelines

The anticipated duration of the project is three months.

- Week 1

Studying AGC version 1.0.0

- Week 2

Implementation of AGC version 1.0.0 with RDataFrame, switching to NanoAOD data schema

- Week 3

Studying AGC version 2.0.0

- Week 4

Partial implementation of AGC version 2.0.0: cuts and selection criterias redefinition.

- Weeks 5-6

Familiarizing with the existing ML model

- Weeks 7-8

RDF implementation of boosted decision trees, adding ML to the analysis pipeline

- Weeks 9-10

Performance measurements

- Weeks 11-12

Spillover tasks. Writing documentation. Presenting my findings at the ROOT team meeting and IRIS-HEP Fellow meeting at the end of the internship.

References

1. ROOT's RDataFrame documentation. URL:
https://root.cern/doc/master/classROOT_1_1RDataFrame.html
2. IRIS-HEP. Analysis Grand Challenge. URL:
<https://iris-hep.org/projects/agc.html>
3. IRIS-HEP. URL:
<https://iris-hep.org>
4. Analysis Grand Challenge documentation. URL:
<https://agc.readthedocs.io/en/latest/index.html>
5. 2015 CMS Open Data. URL:
<https://cms.cern/news/first-cms-open-data-lhc-run-2-released>
6. RDF analysis implementation. URL:
<http://github.com/root-project/analysis-grand-challenge/>
7. Correctionlib package. URL:
<https://github.com/cms-nanoAOD/correctionlib>