Split Uproot into two phases: metadata-reading and data-reading

Yehyun Choi (yc2698@cornell.edu)
Mentors: Andres Rios-Tascon, Nick Smith, Ianna Osbonre

The majority of high-energy physics data is stored in .root files, a specialized computer program for data analysis. Its design involving 'TTree' is highly optimized for storing event-based physical data from particle detectors, with regards to data access and scalability. While the data structure itself is useful, the usage of the standard ROOT implementation with C++ can be tricky, from its installation to navigating the C++ language itself. Furthermore, with recent developments in various Python libraries for data analysis and machine learning, it is practical to read and write ROOT files in pure python and NumPy, which Uproot hopes to accomplish. Unlike PyROOT, Uproot does not require the installation of C++ ROOT and allows the user to directly read ROOT data into Numpy arrays.

This project aims to modularize Uproot and improve its performance. The lack of compartmentalization in Uproot's code results in disjoint functions slowing each other down alongside general readability of code. Specifically, we hope to split the metadata-reading phase (for example, getting the user-defined branch names and types or detector/simulation parameters)and data-reading phase, i.e. reading event-level information alongside physical quantities. Working closely with rootfilespec, [1] we expect that for large ROOT files, this modularization will reduce redundant tasks and enhance performance [2]. Furthermore, if time permits, we hope to diagnose hot spots in the process and wrap the appropriate code in Rust, allowing multithreading and faster processing of loops and binary data.

Estimated Timeline:

Week 1 - 3: understand uproot implementation and pre-existing code in rootfilespec; write test functions

Week 4 - 5: with sample ROOT files, write functions to parse their specifications

Week 6-7: source complex sample ROOT files and debug edge cases

Week 8: understand Uproot's structure

Week 9-12: implement rootfilespec in Uproot, write tests and documentation

References:

[1]: https://github.com/nsmith-/rootfilespec/tree/main/src/rootfilespec

[2]: https://indico.cern.ch/event/1338689/contributions/6015924/attachments/2942382/5169952/pivarski-chep2024-python313.pdf