Project Summary

Elements: Improving tools based on data-description standards for gigabyte-scale data sets

Overview: The PI proposes to improve a set of data-analysis tools that require a description of the data format - rather than requiring a particular data format - to handle the gigabyte-scale datasets common in nuclear physics. The long-term goal of this work is to increase the accessibility of science analysis for both active researchers and science students.

In pursuit of this goal, this proposal has three objectives: (1) Improve existing software that provides access to data in any user-described format so that it works well with gigabyte-scale datasets. (2) Build community adoption and support of these tools. (3) Increase access to analysis of scientific data sets for undergraduate students, early-career researchers, and scientists who do not have access to dedicated software support. These objectives will be met as follows:

(1) Improving standards-based analysis software: Scientists who wish to analyze custom-format data can already do so with existing open-source software, e.g. kaitai-struct. However, this software has poor performance for files larger than a gigabyte, limiting its usefulness to the nuclear physics community. The PI proposes to improve this software to allow responsive analysis of gigabyte-scale datasets in python. This work is feasible thanks to libraries built and maintained by the IRIS-HEP collaboration that provide an intuitive interface to data structures optimized for rapid access to large, event-based data sets.

(2) Building community: The PI proposes yearly workshops to bring together developers and scientists for training, user feedback, and development of this software.

(3) Increasing Access: To fully participate in nuclear physics research, students must have extensive scientific computing skills. Mentor networks help some students through this maze but are not equally available. The PI requests funding for undergraduate students to develop and test documentation and training materials for the standards-based analysis tools and for fundamental scientific computing concepts.

Intellectual merit: This work supports the science effort of the Super Cryogenic Dark Matter Search, an experiment that seeks to better understand the nature of dark matter. This work supports additional high-priority science in the nuclear physics community such as the origin of the elements in the universe, fundamental symmetry testing, and radiation monitoring for national security.

Broader impacts: The PI proposes this work because it would be immediately useful to her own work on dark matter and because software that can provide easy access to any user-described data format would meet an urgent need of communities dealing with gigabyte-scale data. The PI requests funds to develop and field-test documentation for this software to maximize its usefulness to the community.

Right now, the necessity for local development of analysis code restricts participation in science analysis. With a small set of well-documented community tools based on data-description standards, along with open-access material introducing the computing concepts required to use those tools, the PI hopes to increase research access to a much wider group of individuals and increase the time expert scientists can spend doing science.