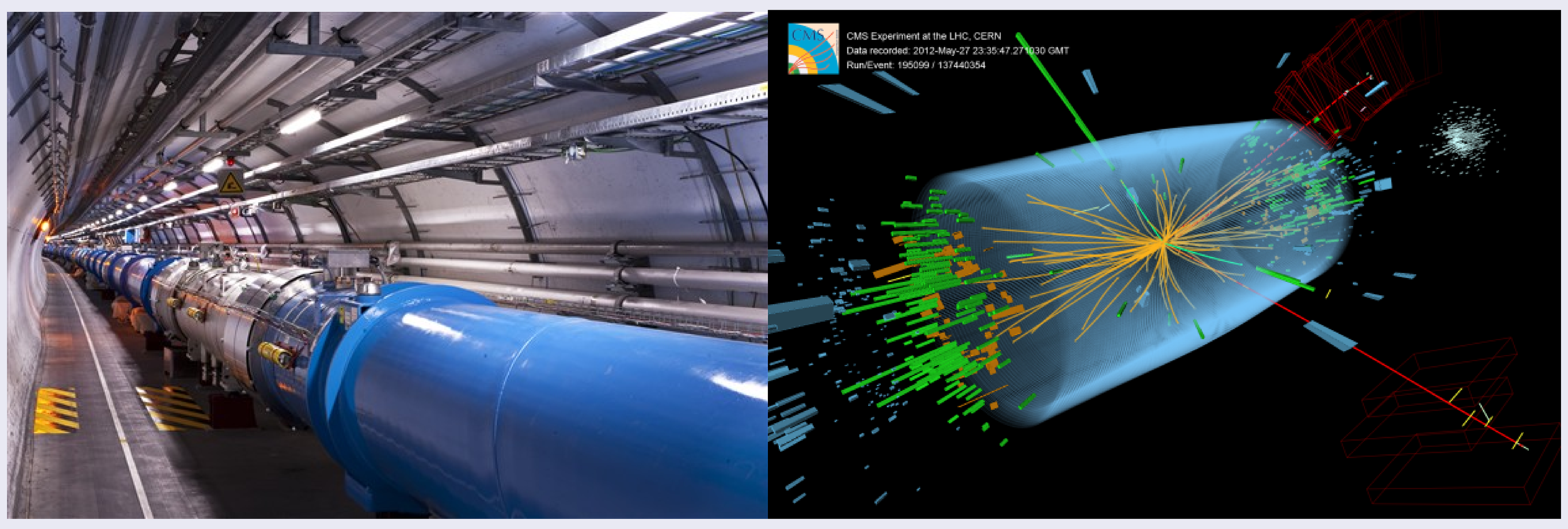
Data-Intensive Analysis for High Energy Physics (DIANA/HEP)

dianahep
http://diana-hep.org

PIs: Peter Elmer (Princeton U.), Brian Bockelman (U.Nebraska-Lincoln), Kyle Cranmer (NYU), Mike Sokoloff (U.Cincinnati)

High Energy Physics (HEP)

The quest to understand the fundamental building blocks of nature, and their interactions, is one of the longest running and most ambitious of human endeavors. Facilities such as the Large Hadron Collider (LHC), where we do our research, represent a huge step forward in our ability to answer these questions. The discovery of the Higgs boson, the observation of exceedingly rare decays of B mesons, and exclusion of countless theories beyond the Standard Model (SM) of particle physics demonstrate that these facilities deliver results. However, the most interesting fundamental physics questions remain wide open, amongst them: What is the dark matter which pervades the universe? Does space-time have additional symmetries or extend beyond the 3 spatial dimensions we know? What is the mechanism stabilizing the Higgs mass from enormous quantum corrections? Are neutrinos, whose only SM interactions are weak, their own anti-particles? Can the theories of gravity and quantum mechanics be reconciled?



The DIANA/HEP Project

The primary goal of DIANA/HEP is to develop state-of-the-art tools for experiments which acquire, reduce, and analyze petabytes of data. Improving performance, interoperability, and collaborative tools through modifications and additions to ROOT and other packages broadly used by the community will allow users to more fully exploit the data being acquired at CERN's LHC and other facilities. The LHC experiments, for example, use nearly 0.5 Exabyte of storage today, and planned upgrades through the 2020s will increase this by more than a factor of 100.

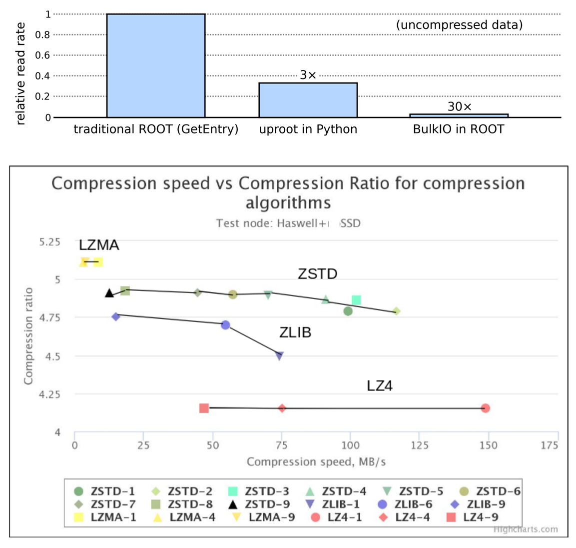
The HEP Analysis Software Ecosystem

ROOT (https://root.cern.ch) is home for most community analysis software developed in particle physics and related fields. Begun at CERN in 1995, it provides a sophisticated data format and serialization technology as well as key software tools for data modeling, likelihood fitting, statistics and multivariate data analysis. It also has a broader range of functionalities, not strictly tied to the data-intensive aspects, including interactive C++ analysis, histogramming, graphics, math libraries, image manipulation, and tools for distributed computing. Given the challenges from technology evolution and analysis complexity, DIANA/HEP is building on and improving these community libraries, moving other existing software elements into community libraries, and developing additional new tools. We are also working to strengthen and grow the wider HEP software development community through a program of documentation, benchmarking, and education in order to build sustainable collaborations which extend beyond the DIANA/HEP project.

Project Status

Improved Performance

To reduce the time to scientific discovery and to enable more in-depth analyses, we are increasing the rate of access to ROOT data files. This includes streamlined access to simpler data types (uproot and BulkIO) and faster compression algorithms (LZ4 and ZSTD). These efforts have already provided factors-of-several improvements.

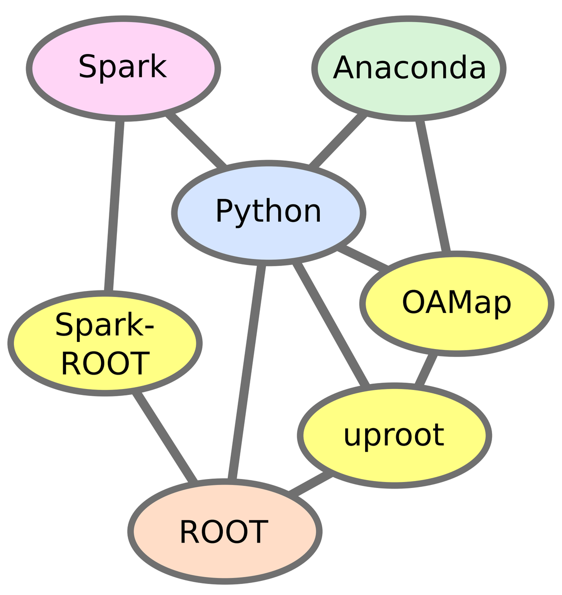


New Statistical Techniques

We are developing new tools and methods for high-level statistical analysis in particle physics. Our activities include research and tools for simulator-based inference (Carl), machine learning for particle physics (Scikit-Optimize), high-level software for efficient numerical computations, and education efforts in these respective domains.

Bridging to Big Data

"Big Data" software in industry, such as the Spark and scientific Python ecosystems, both complement and reproduce functionality of software traditionally developed within HEP. To give physicists more options and reduce maintenance burdens within HEP, DIANA is building bridges between HEP software and the Big Data ecosystems: Spark-ROOT to Spark and uproot/OAMap to Numpy, Numba, and Dask.



High-level Tools

Computing issues can get in the way of a focus on physics, especially for new students who must learn both at the same time. We are therefore striving to present HEP analysis with higher-level interfaces. Scikit-HEP incorporates HEP techniques in Pythonic idioms, uproot provides access to ROOT data as Numpy and Pandas abstractions, and OAMap compiles object-centric user code into fast array operations.

Scikit HEP

uproot

Project Team

- Peter Elmer (Lead PI) - *Princeton Univ., Dept. of Physics*
- Brian P. Bockelman (PI) - *Univ. of Nebraska-Lincoln, Dept. of Computer Science and Engineering*
- Kyle Cranmer (PI) - *New York Univ., Dept. of Physics & Center for Data Science*
- Michael D. Sokoloff (PI) - *Univ. of Cincinnati, Dept. of Physics*
- Jinyang Li (Senior Personnel) - *New York Univ., Computer Science Dept.*
- David Lange - *Princeton Univ., Dept. of Physics*
- Vincent Croft - *New York Univ., Dept. of Physics*
- James Pivarski - *Princeton Univ., Dept. of Physics*
- Eduardo Rodrigues - *Univ. of Cincinnati, Dept. of Physics*
- Oksana Shadura - *Univ. of Nebraska-Lincoln, Dept. of Computer Science and Engineering*
- Zhe Zhang - *Univ. of Nebraska-Lincoln, Dept. of Computer Science and Engineering* (Ph.D. Student)
- Chien-Chin Huang - *New York Univ., Computer Science Dept.* (Ph.D. Student)

Advisory Board

- Amber Boehnlein - *CIO, Thomas Jefferson National Accelerator Facility*
- Katherine Copic - *Director of Growth, Insight Data Science*
- Jacob VanderPlas - *Director of Research, Physical Sciences, eScience Institute, Univ. of Washington*
- Fernando Perez - *Staff Scientist, Data Science and Technology Division, Lawrence Berkeley National Laboratory; Associate Researcher, Berkeley Institute for Data Science, UC Berkeley.*
- Attanagoda Santha - *Architect, Fannie Mae*

Acknowledgement

This project is supported by National Science Foundation grants ACI-1450310, ACI-1450319, ACI-1450323, and ACI-1450377. Any opinions, findings, conclusions or recommendations expressed in this material are those of the developers and do not necessarily reflect the views of the National Science Foundation.

Online version of this poster with clickable links: https://goo.gl/Xvuy2P

NSF