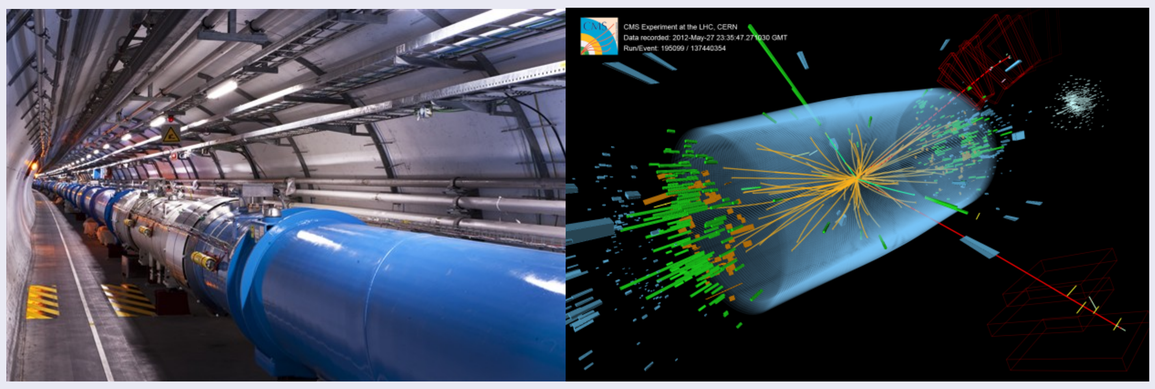
Data-Intensive Analysis for High Energy Physics (DIANA/HEP)

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

http://diana-hep.org

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 experiments 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? Planned and running HEP experiments aim to answer these questions over the next 20 years.



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 Large Hadron Collider (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.

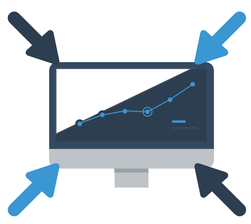
The HEP Analysis Software Ecosystem

ROOT (`https://root.cern.ch`) is the *de-facto* 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 of our science, including interactive C++ analysis, histogramming, graphics (2D and 3D), math libraries (matrix algebra), image manipulation, and tools for distributed computing. Despite many pioneering and innovative features, the components are seen as too coupled, and limited by design decisions taken 20 years ago. Given the challenges from technology evolution and analysis complexity, we are at a point in the software lifecycle where large changes are needed, much as ROOT replaced an earlier generation of FORTRAN-based tools (PAW, HBOOK). DIANA/HEP is building on and improving these community libraries, moving other existing software elements into community libraries, and developing additional new tools.

Project Goals

Collaborative Analyses

Establish infrastructure for a higher-level of collaborative analysis, building on the successful patterns used for the Higgs boson discovery and enabling a deeper communication between the theoretical community and the experimental community.



Interoperability

Improve the interoperability of HEP tools with the larger scientific software ecosystem, incorporating best practices and algorithms from other disciplines into HEP.

Faster Processing

Increase the CPU and IO performance needed to reduce the iteration time so crucial to exploring new ideas

Better Software

Develop software to effectively exploit emerging many- and multi-core hardware. Promote the concept of software as a research product.

Reproducible Analyses

Streamline efforts associated to reproducibility, analysis preservation, and data preservation by making these native concepts in the tools.

Training

Provide training for students in all of our core research topics.

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* (From Jan. 2016)
- Gilles Louppe - *New York Univ., Dept. of Physics & Center for Data Science* (From Oct. 2015)
- James Pivarski - *Princeton Univ., Dept. of Physics* (From Jan. 2016)
- Eduardo Rodrigues - *Univ. of Cincinnati, Dept. of Physics* (From Feb. 2016)
- 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.