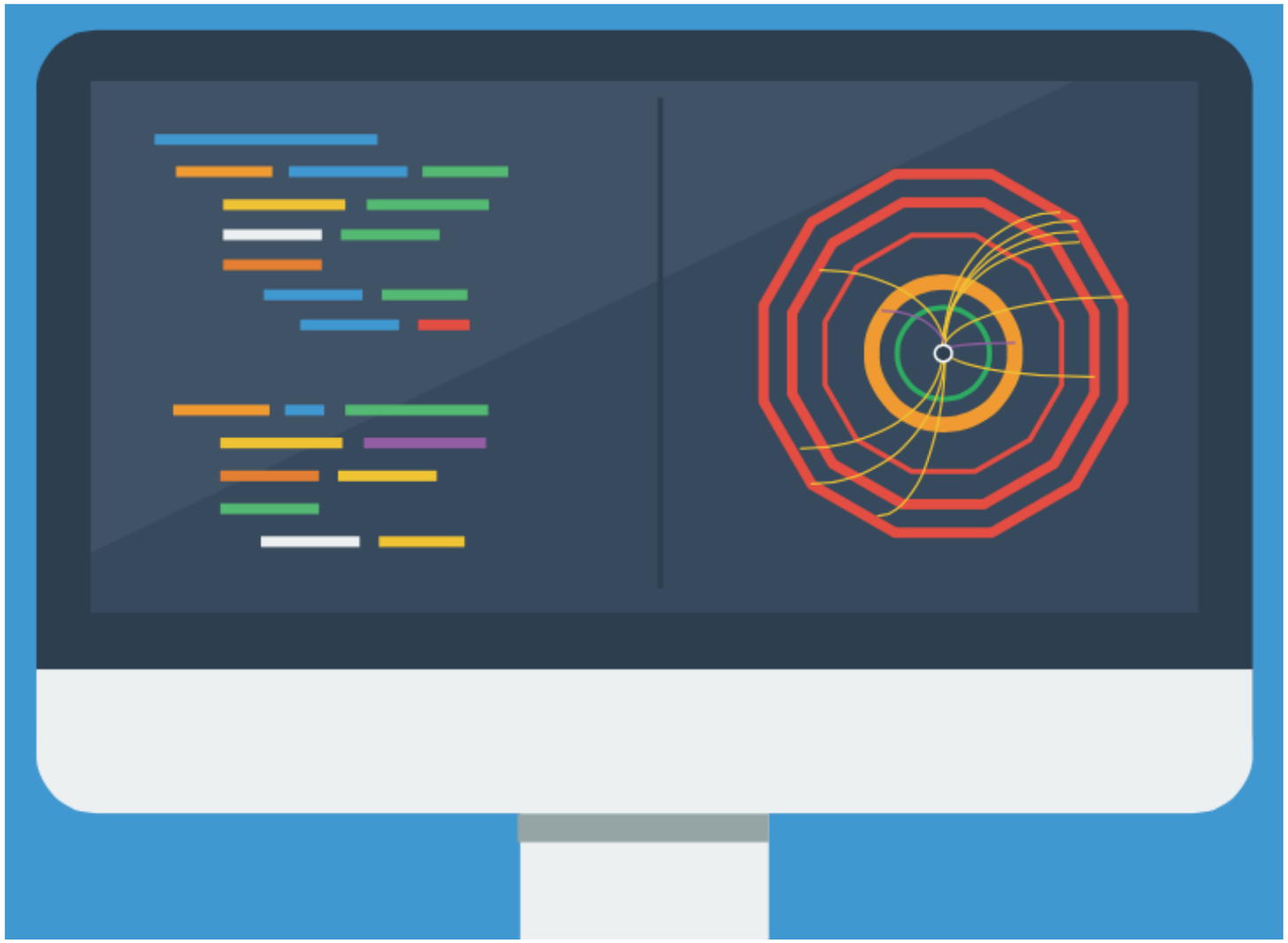


diana hep

http://diana-hep.org

Peter Elmer - Princeton University
Brian Bockelman - UNL
Kyle Cranmer - NYU
Mike Sokoloff - U.Cincinnati

Data Intensive ANAlysis for HEP

- The primary goal of DIANA/HEP is to develop state-of-the-art tools for experiments which acquire, reduce, and analyze petabytes of data.

- DIANA is not a piece of software itself, but a collaborative project to improve and extend analysis tools as sustainable infrastructure for the community.

- DIANA is 4 year project, 6-7 FTE spread over 4 universities (Princeton, NYU, U.Cincinnati, U.Nebraska-Lincoln)

DIANA focus areas

- Performance - focus on both CPU- and IO-performance improvements, including use of multi-/manycore technologies

- Interoperability - Work towards better interoperability with the larger scientific software ecosystem, including both the Python ecosystem, Apache Spark, etc. How do we transition to a more sustainable path where new ideas and software developed elsewhere can be more easily used in HEP and our best products can be evaluated by other fields?

- Collaborative Analysis - new tools that build on the concept and emerging practices in HEP that data analysis is a collaborative activity, involving many individuals working within a given experiment, working in different experiments and even between the experimental and theory communities. Focus on new and best practices for data preservation, analysis archival, reproducibility, and open access.

NSF SI2 program

- In the U.S. the National Science Foundation (NSF) has recognized that software is a critical piece of the research (cyber)infrastructure.

- Previously only a by-product of the scientific research program, it requires actual support to grow into powerful sustainable infrastructure.

- "Software Infrastructure for Sustained Innovation" (SI2) program provides support for such projects in universities

- "Allied" funding from beyond the particle physics program

NSF SI2 program

- Not just software development, but part of a larger set of strategic goals:

- Capabilities: Support the creation and maintenance of an innovative, integrated, reliable, sustainable and accessible software ecosystem providing new capabilities that advance and accelerate scientific inquiry and application at unprecedented complexity and scale.
- Research: Support the foundational research necessary to continue to efficiently advance scientific software, responding to new technological, algorithmic, and scientific advances.
- Science: Enable transformative, interdisciplinary, collaborative, science and engineering research and education through the use of advanced software and services.
- Education: Empower the current and future diverse workforce of scientists and engineers equipped with essential skills to use and develop software. Further, ensure that the software and services are effectively used in both the research and education process realizing new opportunities for teaching and outreach.
- Policy: Transform practice through new policies for software addressing challenges of academic culture, open dissemination and use, reproducibility and trust of data/models/ simulation, curation and sustainability, and that address issues of governance, citation, stewardship, and attribution of software authorship.

- Need to build only software, but also better structures for collaboration, career paths, education, etc.

DIANA team - Principal Investigators

- Peter Elmer (Princeton)
  - Many roles in Software/Computing in BaBar and CMS
  - Early involvement in xrootd, etc.
- Mike Sokoloff (Cincinnati)
  - Physics research: flavor analysis on BaBar/LHCb
  - NSF-funded R&D investigations into many/multicore technologies (GooFit prototype, likelihood fitting)

DIANA team - Principal Investigators

- Brian Bockelman (U.Nebraska-Lincoln)
  - Computer Science research faculty
  - Significant involvement in CMS and Tier2 Computing and the Open Science Grid
  - NSF-funded AAA project (xrootd-based data federation)
  - Collaboration on I/O system: initially performance on long-latency systems, leading also to general purpose improvements/contributions

DIANA team - Principal Investigators

- Kyle Cranmer (NYU)
  - Physics research on Atlas
  - RooStats and HistFactory, statistical procedures and Higgs combination
  - RECAST, Data Preservation (NSF-funded DASPOS project), Moore-Sloan Data Science Environment

DIANA team

NYU

- Gilles Louppe hired as NYU research assistant (at CERN)
- Background in Machine Learning, focus on random forests, Major contributor to scikit-learn (Python machine-learning library)
- http://www.montefiore.ulg.ac.be/~glouppe/research.php
- Associate faculty (computer science) Jinyang Li
- Expect a graduate student

University of Cincinnati

- Currently open position for a Sr. Research Associate or Research Scientist

- Development of software tools as part of DIANA

- Core computing work and data-analysis as an LHCb collaborator

- Location: CERN

- Full job ad at: http://diana-hep.org/pages/jobs.html

Princeton

- 2 positions will open (1 DIANA, 1 USCMS, with slightly rotated "basis vectors")

- Expecting job ad will be public in a week or two

- Computational Physicist or Scientific Application Developer, depending on background

- Location: CERN or FNAL preferred, but negotiable

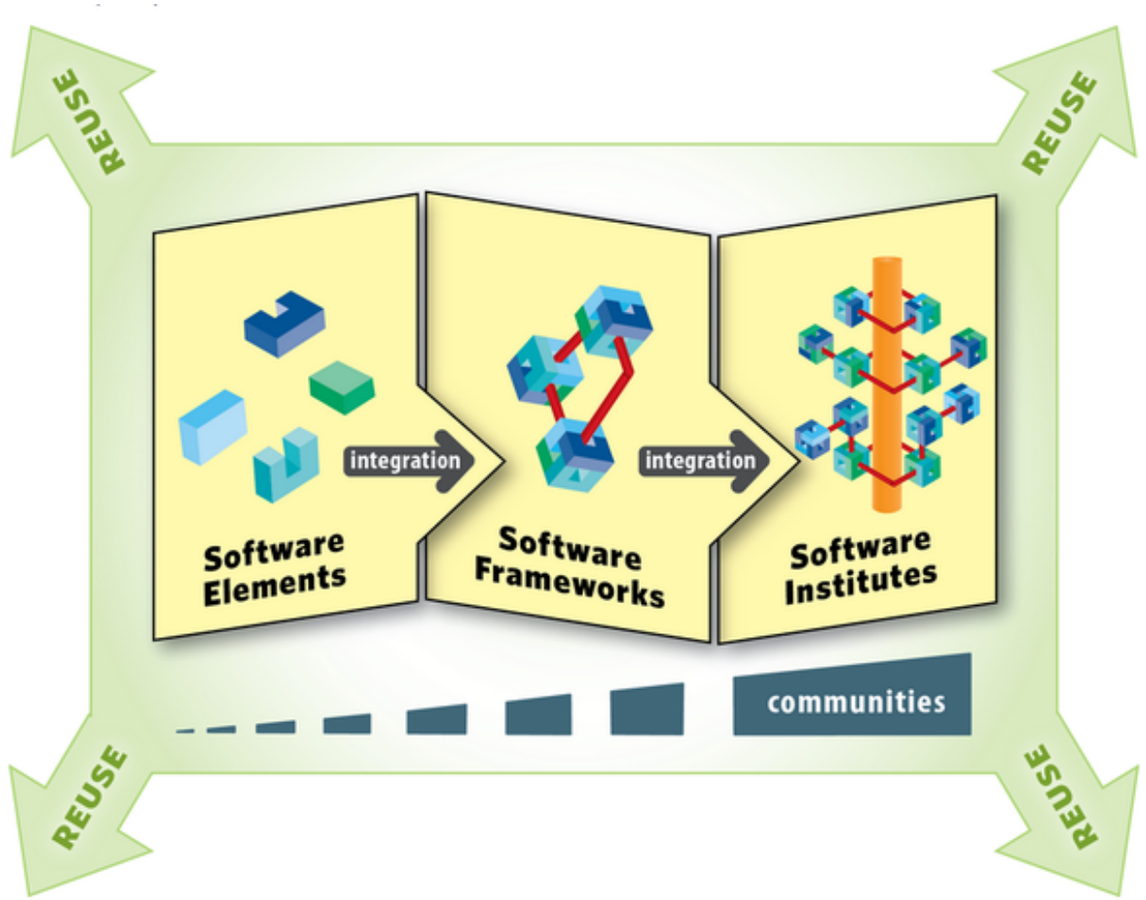
Univ. Nebraska - Lincoln

- Expect the position to open soon
- Position will focus 100% on DIANA
- Possible opportunity as a long-term hire with the Nebraska team
- Location: Nebraska
- Expect also a graduate student

Where does this go?

- DIANA is a 4-year project, then what?

- The NSF SI2 program envisions also a broader long-term evolution towards software ecosystems supporting research communities



NSF SI2 - Software Institutes

- DIANA is a "Software Framework" project (SI2-SSI)

- M. Sokoloff, M.Neubauer (U.Illinois) and I have also made a proposal for (pending) funding for the "conceptualization" of a "Software Institute"

- If funded, this will likely lead to a series of more focused HSF-branded workshops in 2016, to produce a community white paper and (specifically for the U.S. university community) a strategic plan

- Next step would be proposal for a 5-year "Software Institute" implementation project

Summary

- DIANA/HEP is a new collaborative project focused on the development of analysis tools

- ROOT and its ecosystem are a key part of analysis in HEP

- We are looking forward to a fruitful collaboration with the ROOT, HEP and wider scientific software communities in the coming years

- The DIANA project is supported by National Science Foundation grants ACI-1450310, ACI-1450319, ACI-1450323, and ACI-1450377.