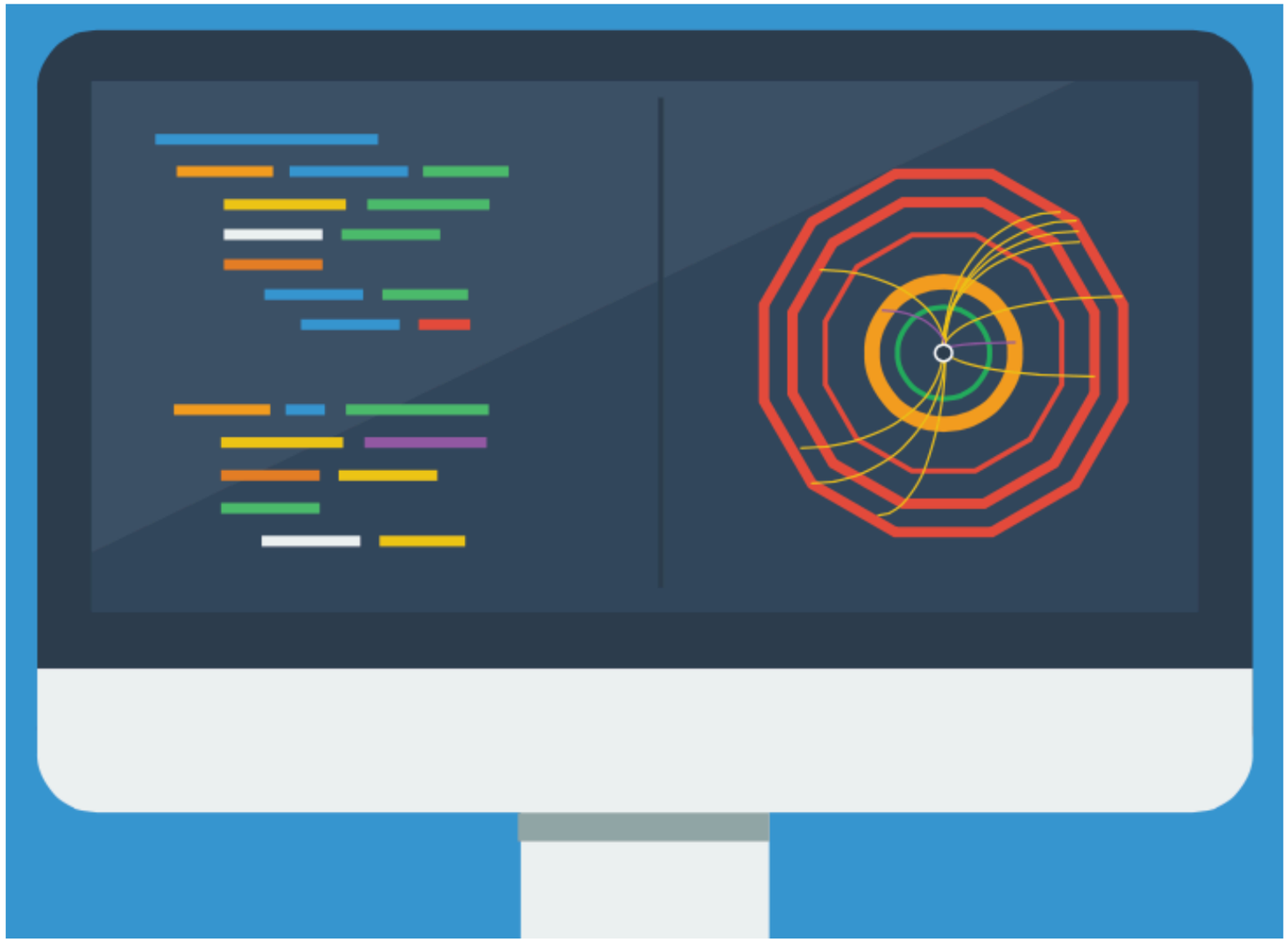


dianahep

http://diana-hep.org

Peter Elmer
Princeton University

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/HEP is part of the 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.

FROM NSF

http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504817

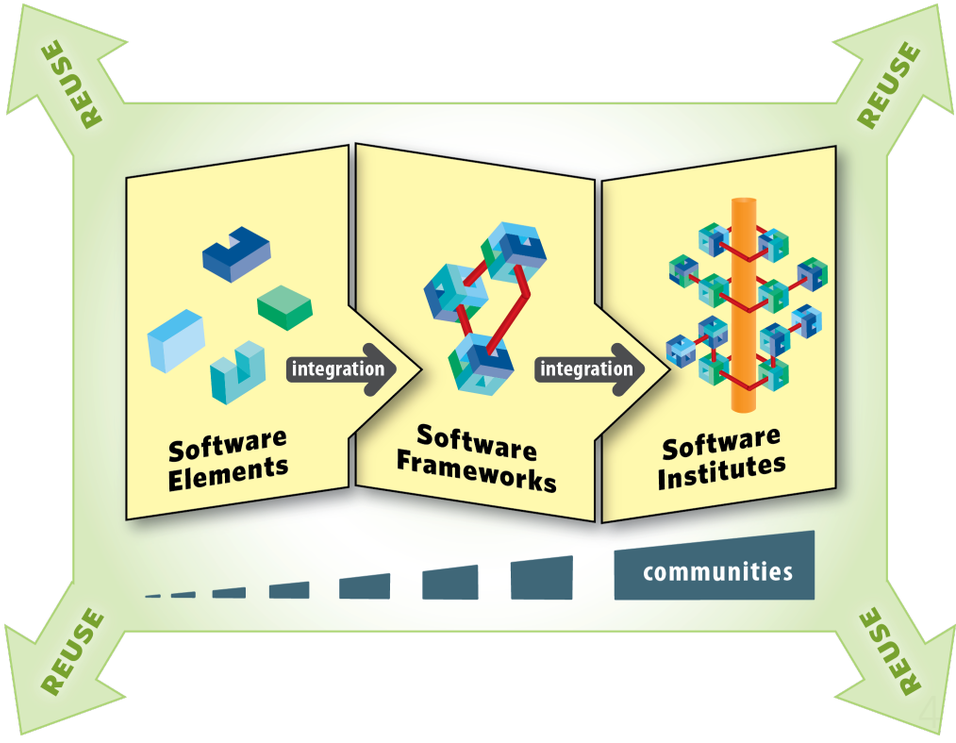
The SI2 program includes four classes of awards:

1. **Scientific Software Elements (SSE)**: SSE awards are Software Elements. They target small groups that will create and deploy robust software elements for which there is a demonstrated need that will advance one or more significant areas of science and engineering.

2. **Scientific Software Integration (SSI)**: SSI awards are Software Frameworks. They target larger, interdisciplinary teams organized around the development and application of common software infrastructure aimed at solving common research problems. SSI awards will result in sustainable community software frameworks serving a diverse community. ← DIANA is an SSI

3. **Scientific Software Innovation Institutes (S2I2)**: S2I2 awards are Software Institutes. They focus on the establishment of long-term hubs of excellence in software infrastructure and technologies that will serve a research community of substantial size and disciplinary breadth.

4. **Reuse**: In addition, SI2 provides support through a variety of mechanisms (including co-funding and supplements) to proposals from other programs that include, as an explicit outcome, reuse of software. Proposals that integrate with previously developed software, either by reference or inclusion, are encouraged. Proposals developing new software with an explicitly open design for reuse may also be considered. The purpose of the Reuse class is to stimulate connections within the broader software ecosystem. The class of reuse awards is currently being developed.



HIGH-LEVEL INTRO

DIANA is about (cross-experiment) analysis tools. Grant runs 2015-2019. We have broad areas of activity and goals:

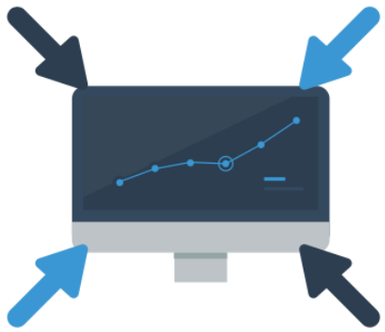
- **performance**: ROOT I/O, vectorization, …
- **interoperability**: scientific python ecosystem, R, hadoop, spark, …
- **collaborative tools & reproducibility**: RooFit workspace, HEPdata, CAP

Approach:

- Specific focus is meant to be coordinated with needs of experiments.
- Of course ROOT sits at the center of the analysis tools ecosystem in HEP, thus are collaborating directly with ROOT team (and others).

As part of the NSF's Software Infrastructure for Sustained Innovation (SI2) program, DIANA is concerned with the overarching goal of transforming innovations in research and education into sustained software resources that are an integral part of the cyberinfrastructure.

HTTP://DIANA-HEP.ORG



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

Reproducible Analyses

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

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.

Training

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

Design by Eamonn Maguire, CERN fellow, HEPdata developer

DIANA Team

Project Team

- Peter Elmer (Lead PI) - Princeton University, Department of Physics
- Brian P. Bockelman (PI) - University of Nebraska-Lincoln, Department of Computer Science and Engineering
- Kyle Cranmer (PI) - New York University, Department of Physics & Center for Data Science
- Michael D. Sokoloff (PI) - University of Cincinnati, Department of Physics
- Jinyang Li (Senior Personnel) - New York University, Computer Science Department
- David Lange - Princeton University, Department of Physics
- Gilles Louppe - New York University, Department of Physics & Center for Data Science
- James Pivarski - Princeton University, Department of Physics
- Eduardo Rodrigues - University of Cincinnati, Department of Physics
- David Abdurachmanov - University of Nebraska-Lincoln, Department of Computer Science and Engineering
- Zhe Zhang - University of Nebraska-Lincoln, Department of Computer Science and Engineering (Ph.D. Student)
- Chien-Chin Huang - New York University, Computer Science Department (Ph.D. Student)
- Lukas Heinrich - New York University, Department of Physics (Ph.D. Student)

Associated team members

- Vassil Vassilev - Princeton University & ROOT Intel Parallel Computing Center

Collaborators

- The ROOT team at CERN and Fermilab
- CMS Big Data Project: Oliver Gutsche, Matteo Cremonesi, Nhan Tran, Jim Kowalkowski, and Saba Sehrish - Fermilab
- Histogrammar: Alexey Svyatkovskiy - Princeton University
- Scikit-HEP: Noel Dawe - University of Melbourne, Vanya Belyaev - ITEP, and Sasha Mazurov - University of Birmingham
- Spark-ROOT: Viktor Khristenko - University of Iowa
- Scope-aaS: Jin Chang and Igor Mandrichenko - Fermilab
- Scope-GPU: Roger Rusack and Peter Hansen - University of Minnesota
- Carl: Juan Pavez, Cyril Becot - New York University, Lukas Heinrich - New York University
- Scikit-Optimize: Manoj Kumar - New York University, Tim Head, Noel Dawe - University of Melbourne

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, University of Washington
- Fernando Pérez - 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

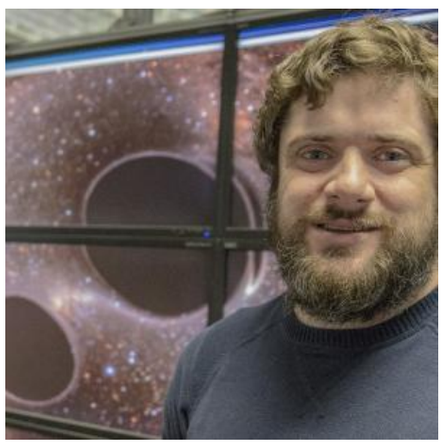
DIANA team - Principal Investigators

- 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

Gilles Louppe - NYU

- **Bio:** computer science background, post-doc in machine learning, scikit-learn core developer

- **Goals:**

- development of machine learning software and applications to high energy physics data

  - ongoing projects: carl (likelihood-free inference toolbox), scikit-optimize (user friendly toolbox for black box optimization)

- machine learning research targeted to high energy physics use cases

  - ongoing projects: likelihood-free inference with classifiers, ATLAS projects, etc.

- education: various courses, tutorials and talks already given on machine learning and related software.

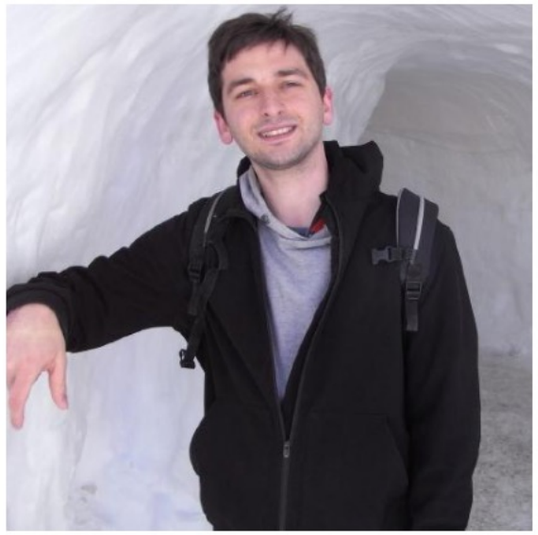
Eduardo Rodrigues - U.Cincinnati

- **Bio:** Physicist, on LHCb since early (2002). Mostly worked on physics and software. Had roles of responsibility such as Coordinator of the Physics Analysis Software Project, Convener of Physics Working Group on Charmless b-hadron Decays, Vertex Detector Software Coordinator, etc. (going backwards in time).

Jim Pivarski - Princeton

- **Physics background:** 5 years of QCD studies with the CLEO Collaboration and 5 years of commissioning and early exotica with CMS Run I. Deeply involved in alignment of both detectors (muon alignment of CMS).

- **Industry background:** 5 years as a data science consultant, helping small and large companies with data analysis techniques and Big Data software. Created a language-agnostic standard for encoding data mining models that is being adopted by the industry (http://dmg.org/pfa).



Lukas Heinrich

My research:

ATLAS BSM Physics Searches. Lead developer of RECAST. Working with analysis teams to capture and define their workflows for use RECAST and large-scale reintrepretation campaigns.

Application of the findings from analysis preservation to projects in Machine Learning.

Trigger Analysis Tools Coordinator in ATLAS

My expertise is:
Triggering Systems, Workflow automation, analysis preservation and reinterpretation. Containers.

A problem I'm grappling with:
How to enable analysis teams to efficiently/easily capture their know-how/code/workflows to maimize the utility of individual analyses

I've got my eyes on:
New analysis models/patterns that go beyond sending batch jobs. Learning from other communities

I want to know more about:
How we can move Machine Learning application further upstream to e.g. simulation.

recast NYU DASPOS dianahep CERN Analysis Preservation

David Lange - Princeton

- Many software roles over the years in BaBar and CMS
- Original co-author of EvtGen
- CMS offline/computing co-coordinator until Sept. 2016

David Abdurachmanov - U.Nebraska-Lincoln

- Started 1 July 2016. For DIANA he focuses on performance in ROOT (I/O, etc.), data compression among other things.

- Has been working on alternative processors (ARM, Power8) in CMS, as well as x86 and power efficiency, significant experience with compilers, ports, etc.

- [There is also an Intel-funded hire at Princeton, Vassil Vassilev, who will collaborate with DIANA on ROOT performance.]

DIANA FELLOWS

Each year, 4 DIANA Graduate Fellows will each spend 3 months intensively developing tools in conjunction with collaborating institutions.

- call for applications will go out soon

Similarly, a DIANA Undergraduate Fellow will work 10 - 12 weeks during the summer, either developing or using data-intensive tools.

DIANA topical meetings

- A forum for presentations and discussion about analysis techniques and analysis tools, of relevance to the broader HEP community

- These meetings are meant to explore near and long term possibilities, ideas and collaborations. We hope to engage people from a number of experiments and from beyond HEP.

- In the steady state we expect approx. 2 meetings per month. If you have ideas, please contact us or bring them up in the meetings.

https://indico.cern.ch/category/7192/

DIANA topical meetings (Monday 17:30GVA)

DIANA

The DIANA/HEP project focuses on improving performance, interoperability, and collaborative tools through modifications and additions to ROOT and other packages broadly used by the HEP community.

July 2016

- 18 Jul DIANA Meeting - The Julia Language

June 2016

- 20 Jun DIANA Meeting - Analysis script language
- 13 Jun DIANA Meeting - Python/ROOT interoperability
- 06 Jun DIANA Meeting - Data and Software Preservation

May 2016

- 23 May DIANA Meeting - MC generation and numerical integration on GPUs
- 16 May DIANA Meeting - MC generation and numerical integration on GPUs (CANCELLED, MOVED to 23 May)
- 02 May DIANA Meeting - Recasting Searches for New Physics using Reproducible Workflows

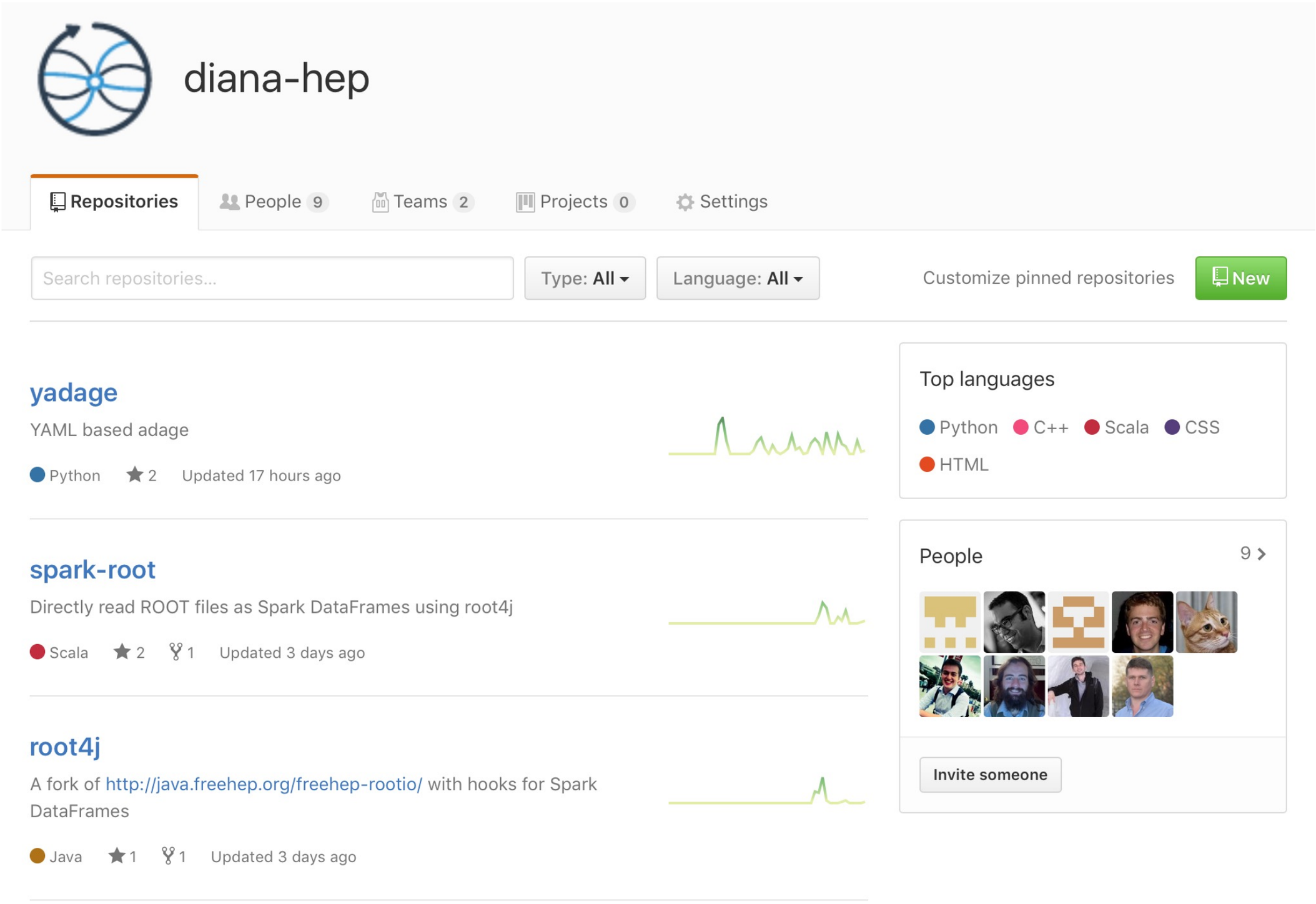
April 2016

- 25 Apr DIANA Meeting - Histogramming in Map-Reduce Systems
- 18 Apr DIANA Meeting - HEPData and Mathematica/ROOT
- 11 Apr DIANA Meeting - Bayesian Optimisation

Google group: diana-hep@googlegroups.com

GITHUB ORGANIZATION

https://github.com/diana-hep



FROM NSF

http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504817

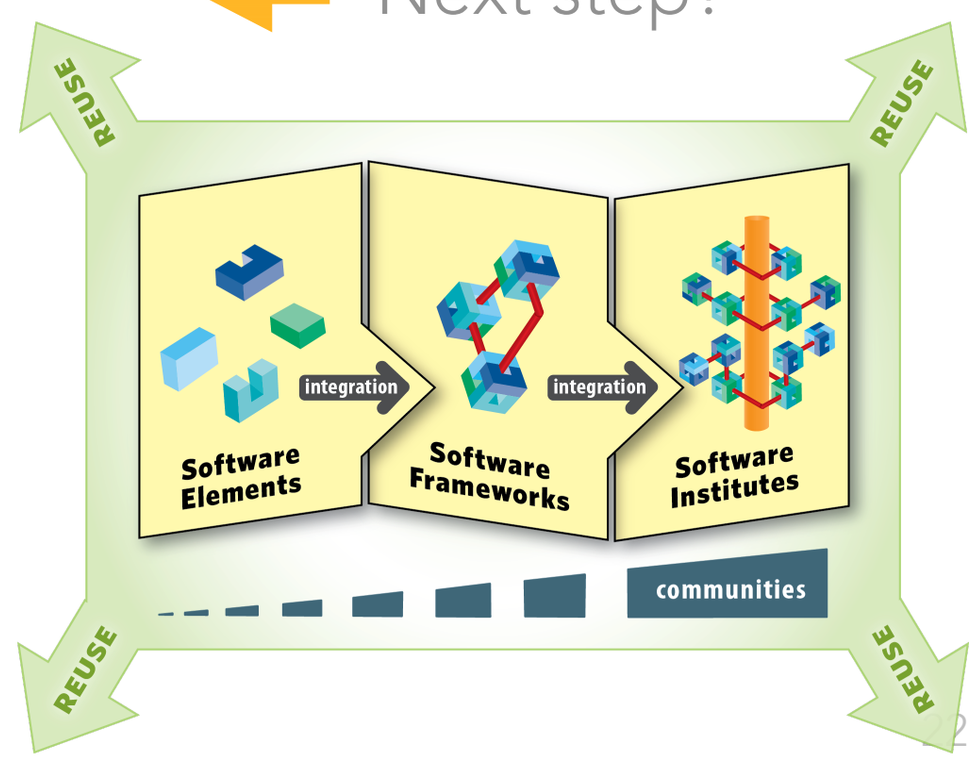
The SI2 program includes four classes of awards:

1. **Scientific Software Elements (SSE)**: SSE awards are Software Elements. They target small groups that will create and deploy robust software elements for which there is a demonstrated need that will advance one or more significant areas of science and engineering.

2. **Scientific Software Integration (SSI)**: SSI awards are Software Frameworks. They target larger, interdisciplinary teams organized around the development and application of common software infrastructure aimed at solving common research problems. SSI awards will result in sustainable community software frameworks serving a diverse community.

3. **Scientific Software Innovation Institutes (S2I2)**: S2I2 awards are Software Institutes. They focus on the establishment of long-term hubs of excellence in software infrastructure and technologies that will serve a research community of substantial size and disciplinary breadth. Next step?

4. **Reuse**: In addition, SI2 provides support through a variety of mechanisms (including co-funding and supplements) to proposals from other programs that include, as an explicit outcome, reuse of software. Proposals that integrate with previously developed software, either by reference or inclusion, are encouraged. Proposals developing new software with an explicitly open design for reuse may also be considered. The purpose of the Reuse class is to stimulate connections within the broader software ecosystem. The class of reuse awards is currently being developed.



Advanced software plays a fundamental role for large scientific projects - from designing experimental instruments to acquiring, reducing, and analyzing the resulting data.

The primary goal of the S2I2-HEP conceptualization project is to prepare a strategic plan for a potential NSF Scientific Software Innovation Institute (S2I2) to develop software for experiments taking data in the "High-Luminosity Large Hadron Collider" (HL-LHC) era in the 2020s. In addition, we are working with the HEP Software Foundation to prepare a larger HEP Community White Paper (CWP) describing a global roadmap for HEP Software and Computing R&D for the 2020s. To this end we are organizing a number of workshops between Fall 2016 and Summer 2017.

Please join the Google Group for updates.

Upcoming Events:

- 23-26 Jan, 2017 - HEP Software Foundation Workshop
  - *University of California at San Diego / San Diego Supercomputer Center*
  - Indico page
- 20-22 Mar, 2017 - IML Topical Machine Learning Workshop (includes CWP session)
  - *CERN* (plus possible Vidyo videconference)
  - Indico page

Links

NSF SI2

DIANA/HEP

HEP Software Foundation

Community White Paper

Molecular Sciences Software Institute

Science Gateways Community Institute