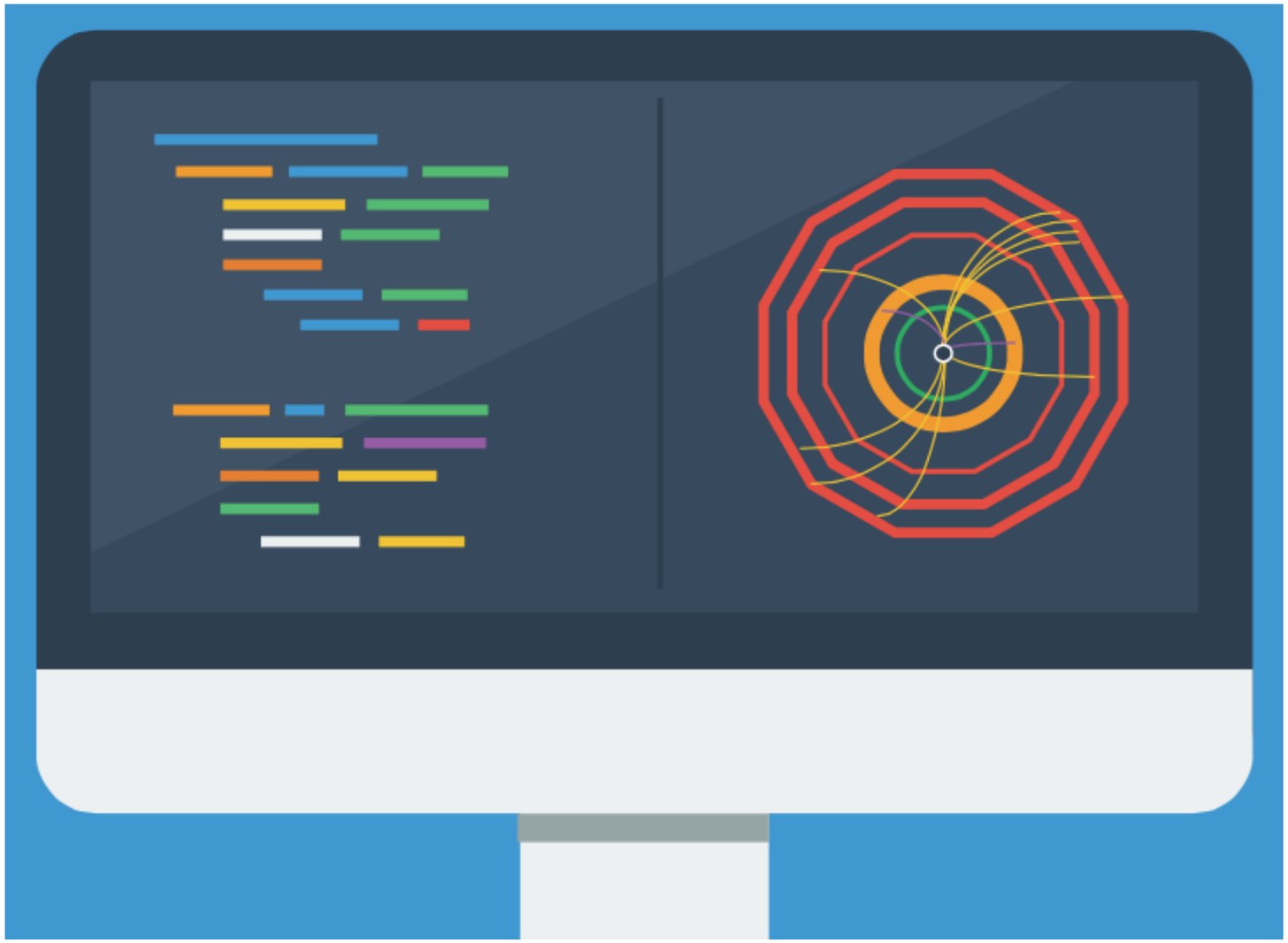


diana hep

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

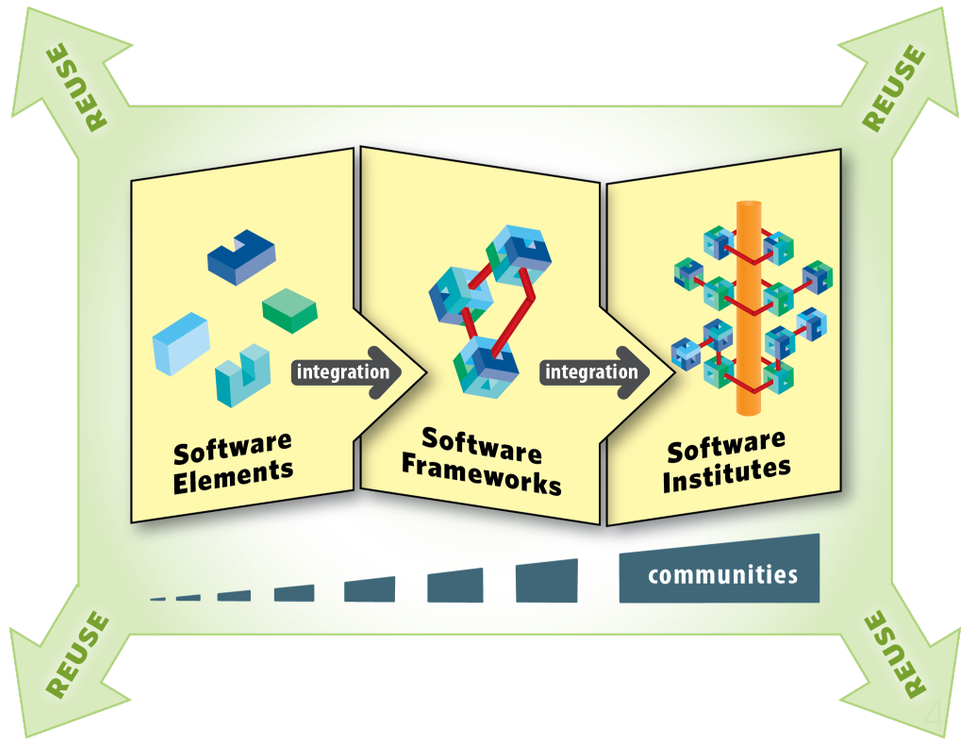
The SI2 program includes four classes of awards:

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

PROJECT TEAM

Peter Elmer (Lead PI) - Princeton, CMS

Brian P. Bockelman (PI) - University of Nebraska-Lincoln, CMS

Kyle Cranmer (PI) - NYU, ATLAS

Michael D. Sokoloff (PI) - Cincinnati, LHCb

Jinyang Li (Senior Personnel) - New York University, Computer Science Department

David Lange - Princeton, CMS co-coordinator Offline Software and Computing

Gilles Louppe - NYU, ATLAS, Machine Learning PhD (former CERN fellow), scikit-learn developer

Jim Pivarski - Princeton, CMS , ROOT interoperability with Hadoop, Spark, etc.

Eduardo Rodrigues - Cincinnati, LHCb, coordinator analysis tools; tracking, trigger, and fitting

Zhe Zhang - Nebraska, Comp. Sci PhD student, improving ROOT IO performance

Chien-Chin Huang - NYU, Comp. Sci. PhD student, RooFit data parallelism, Theano, TensorFlow, etc.

Note: Gilles Louppe started fall 2015, others just started

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.

MACHINE LEARNING: TRAINING AND EXPERTISE

15:00 - 18:45 **Thursday Afternoon Session**

Convener: Maurizio Pierini (CERN)

15:00 **An introduction to machine learning with Scikit-Learn** *2h15'*

This tutorial gives an introduction to the scientific ecosystem for data analysis and machine learning in Python. After a short introduction of machine learning concepts, we will demonstrate on High Energy Physics data how a basic supervised learning analysis can be carried out using the Scikit-Learn library. Topics covered include data loading facilities and data representation, supervised learning algorithms, pipelines, model selection and evaluation, and model introspection.

Speaker: Dr. Gilles Louppe (CERN)

Jupyter notebook | Recording

Pitfalls of evaluating a classifier's performance in high energy physics applications
ALEPH workshop, NIPS, Montréal, Canada
December 11, 2015
https://github.com/glouppe/talk-aleph-workshop2015 [Materials]
An introduction to machine learning with Scikit-Learn
Data Science at LHC, Switzerland
November 12, 2015
https://github.com/glouppe/tutorial-sklearn-dslhc2015 [Materials]
Classification with a control channel: Don't cheat yourself
CERN, Switzerland
October 5, 2015
https://github.com/glouppe/talk-classification-control-channel [Materials]
Scikit-Learn tutorial
AstroHack Week, New York, USA
September 30, 2015
https://github.com/AstroHackWeek/AstroHackWeek2015 [Materials]
Understanding Random Forests
CERN, Switzerland
September 21, 2015
https://github.com/glouppe/talk-pydata2015 [Materials]
An introduction to Machine Learning with Scikit-Learn
CERN, Switzerland
April 23, 2015
https://github.com/glouppe/tutorial-sklearn-lhcb [Materials]

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).

- **Goals:** Make sure the software meets the physicists requirements, and physicists get to use the best tools available and/or being developed. Particular interest in machine-learning related software out there. Keen on tutorials, e.g. gave in 2012 a course on RooFit.

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).

- **DIANA focus:** (1) integrating physics analyses with Big Data software, (2) introducing physicists to more high-level and functional styles of data analysis, and (3) creating tools that bridge both worlds.

David Lange - Princeton

- Many software roles over the years in BaBar and CMS
- Original co-author of EvtGen
- Currently CMS offline/computing co-coordinator until Sept. 2016
- Near term DIANA goals: Investigating interoperability between python and root packages

TBN - U.Nebraska-Lincoln

- One staff position is still being filled...

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

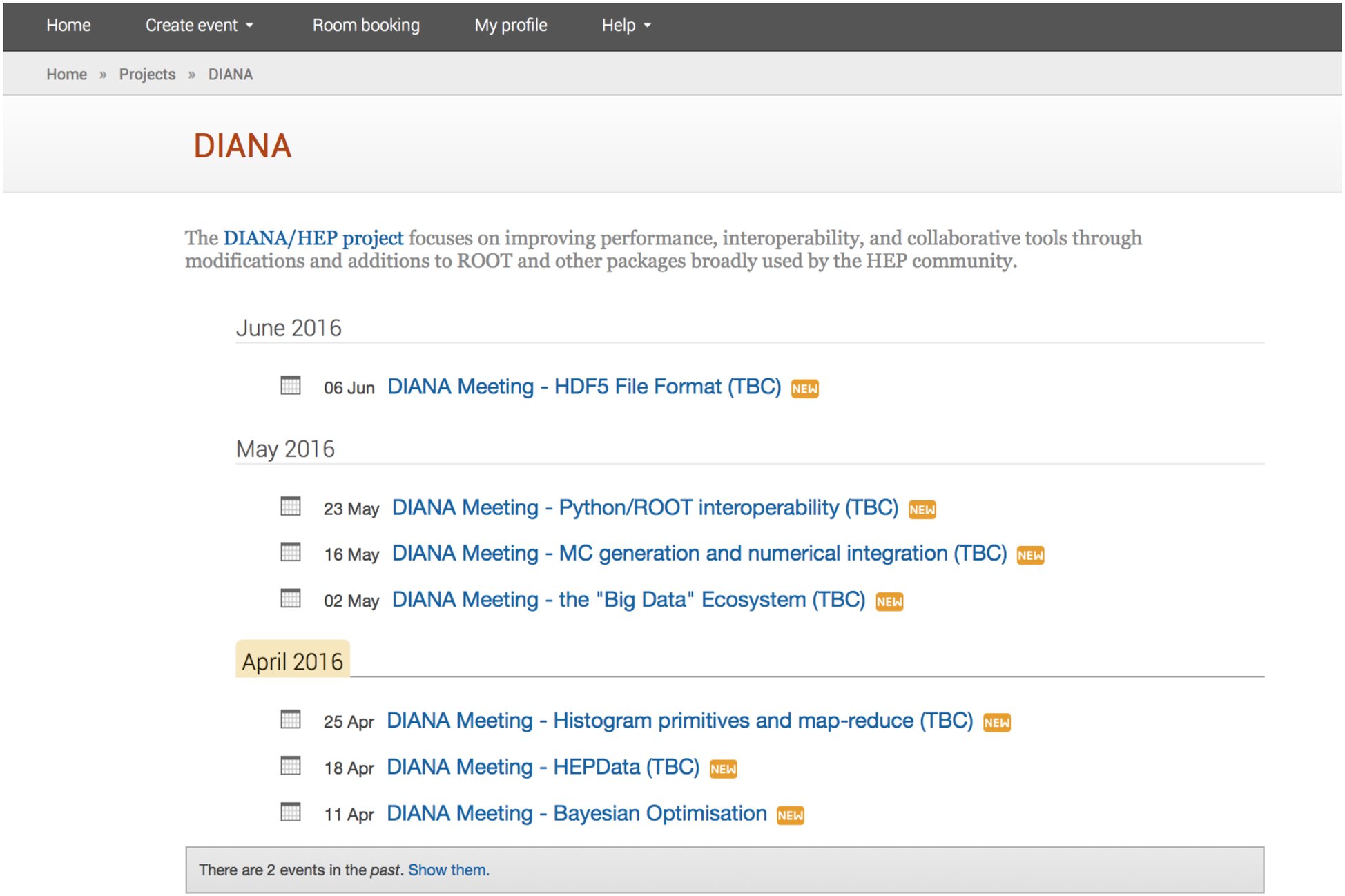
- A forum for presentations and discussion about analysis techniques and analysis tools, of relevance to the broader HEP community

- 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/

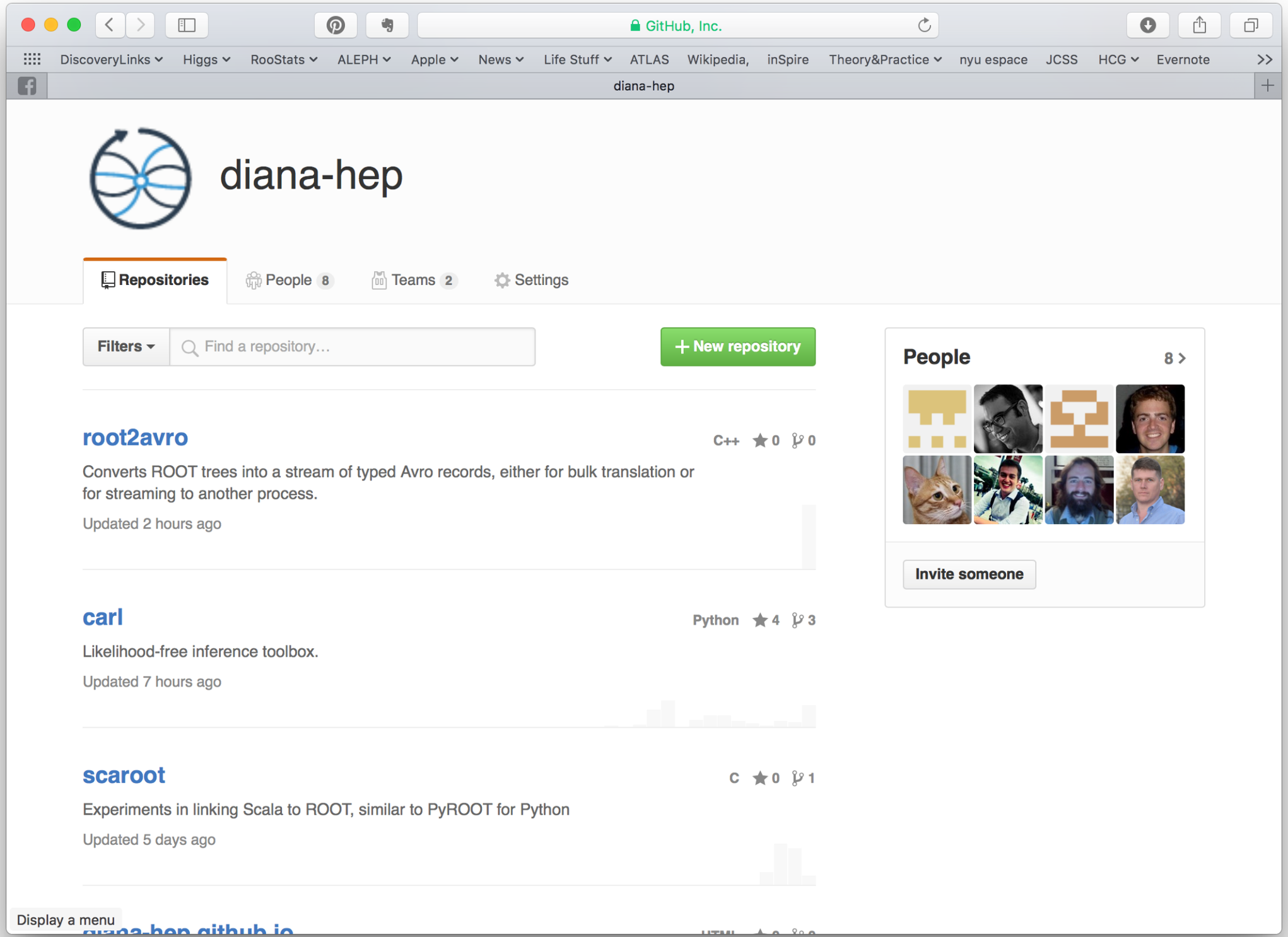
DIANA topical meetings (Monday 17:30GVA)



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:

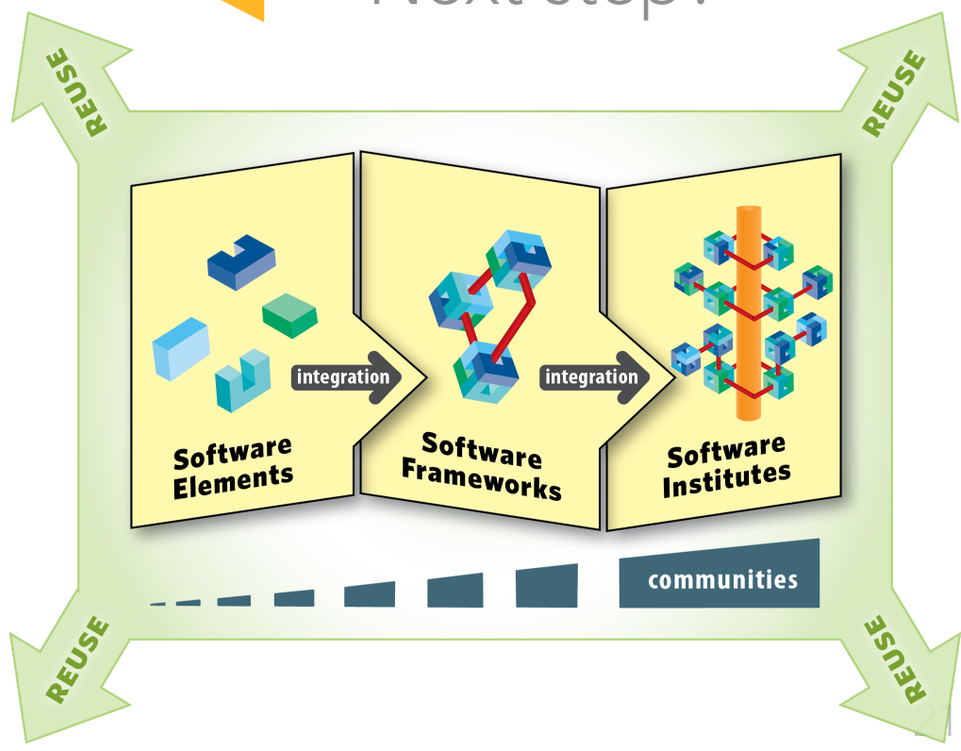
1. **Scientific Software Elements (SSE)**: SSE awards are Software Elements. They target small groups that will create and deploy robust software elements for which there is a demonstrated need that will advance one or more significant areas of science and engineering.

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.



S2I2 CONCEPTUALIZATION

We have been working with the NSF in the US to understand possible long term "upgrade" paths for HEP software, using the HL-LHC as the primary science driver. In particular we have been looking at the NSF's "Software Infrastructure for Sustained Innovation (SI2)" program:

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

Within this program, **there is the possibility of proposing a "S2I2 Software Institute"** with potential funding at the level of $3-5M/year for 5-10 years.

The S2I2 proposal process envisions a two-step process: a conceptualization phase (1 year) and an implementation phase (the 5 year program mentioned above). Mike Sokoloff, Mark Neubauer and I submitted the proposal for the "conceptualization" phase, essentially a series of planning workshops during 2016/2017.

http://cern.ch/elmer/s2i2-2015-nsf-proposal.pdf

NSF STRATEGIC PLAN FOR HL-LHC COMPUTING

News from NSF

Denise Caldwell

Division Director
Division of Physics

With Input from Program Directors: Jim Shank; Brian Meadows; Jean Cottam; Jim Whitmore; Keith Dienes

HEPAP 6 April 2

Also investing in HEP "computational R&D" e.g. a few PIF awards,

AAA, DASPOS

"Enabling High Energy Physics at the Information Frontier Using GPUs and Other Many/Multi-Core Architectures"

"Particle Tracking at High Luminosity on Heterogeneous, Parallel Processor Architectures"

Computing and Cyberinfrastructure at NSF

Priority area of CIF21 (Cyberinfrastructure Framework for 21st Century Science, Engineering and Education)

Close collaboration with Division of Advanced Cyberinfrastructure (ACI)

Projects: OSG, DASPOS

Funding opportunity within Division:
CDS&E (Computation and Data-Enabled Science and Engineering)

Funding opportunity led by ACI: SI2 (Software Infrastructure for Sustained Innovation)

Opportunities to address computing challenges facing the LHC

Contact Bogdan Mihaila with questions

MORE ON NSF STRATEGIC PLAN FOR HL-LHC COMPUTING

Software Infrastructure for Sustained Innovation (SI)^2

Rajiv Ramnath and Dan Katz

Program Directors
Software Cluster
Division of Advanced Cyberinfrastructure
Directorate for Computer and Information Science and Engineering

rramnath@nsf.gov
dkatz@nsf.gov

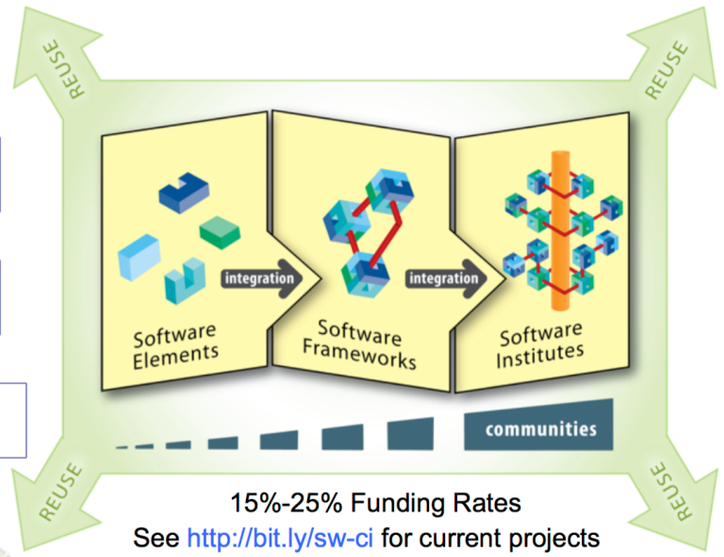
Version: 2/15/16 20:49

ACI Flagship - Software Infrastructure for Sustained Innovation (SI2)

Elements: $500K/3 years

Frameworks: $1M/ year 3-5 years

Institutes: $3-$5m/ year 5-10 years



15%-25% Funding Rates
See http://bit.ly/sw-ci for current projects

Also: EAGERS, RAPIDs and Workshops to target areas of opportunity

Related news: Intel IPCC on ROOT

- I made a proposal 2 months ago for an Intel Parallel Computing Center (IPCC) focused on "code modernization" in ROOT Math and I/O

- The same program currently funds some GeantV effort here at CERN

- Last week they told me that they will fund 1FTE of effort (1+1 year project)