David Lange's Overview of Projects

David Lange

Princeton – DIANA

December 16, 2016

Focused on increasing engagement of CMS analysis groups.

- Python+ROOT-based analysis widespread in CMS (e.g. FWLite, HEPPY framework).

Expanding and modernizing python toolkit in CMS/CMSSW.

- Expanding suite of scientific Python tools: e.g. Pandas, scikit-learn, Jupyter, ML toolkits, etc.
- Expanding suite of HEP-developed tools: e.g. Histogrammar, rootpy, root_numpy, etc.
- Moving CMS to pip-based python package management.
  - Already makes package version management trivial.
  - Eventual goal is to more generally distribute python stack via pip install (or equivalently as a condo channel) as alternative installation scheme for analysis users.

Jim Pivarski's Overview of Projects

Jim Pivarski

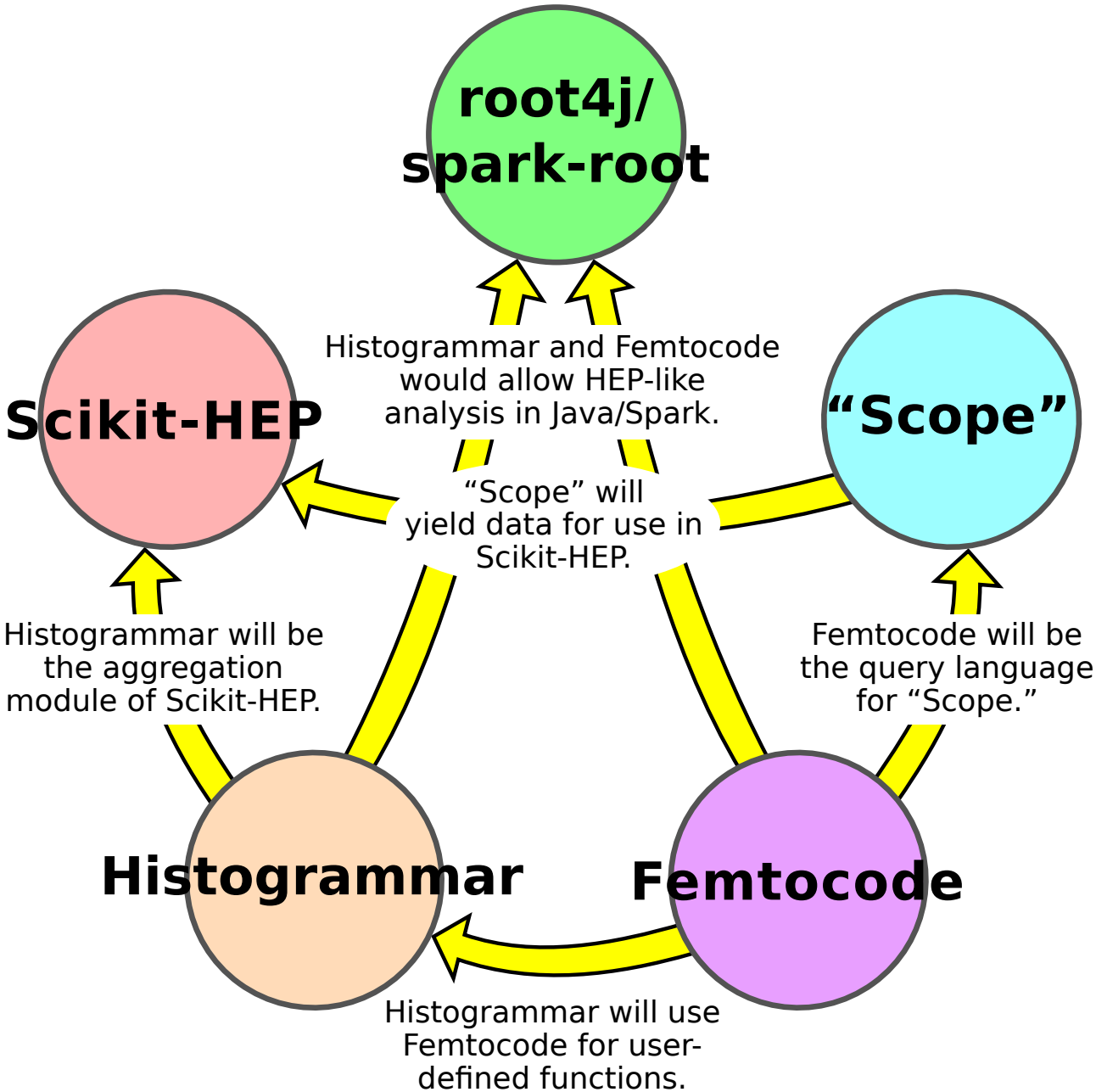
Princeton – DIANA

December 16, 2016

My goals

1. To build bridges between the HEP software ecosystem and the big data ecosystems— Scientific Python and Hadoop/Spark— so that HEP data can easily flow between them.
   - **Scikit-HEP:** reorganize rootpy, root_numpy, Ostap and maybe others into a Pythonic layer between HEP and Scientific Python.
     with Eduardo, David, Noel Dawe, Vanya Belyaev, and Sasha Mazurov
   - **root4j/spark-root:** pure-Java ROOT I/O for Spark integration.
     with Viktor Khristenko, for Oliver Gutsche and Matteo Cremonesi
   - **"Scope:"** NoSQL database/server for interactive analysis.
     with Jin Chang and Igor Mandrichenko
2. To build HEP-friendly APIs in those big data ecosystems that allow us to perform our analyses when we get there.
   - **Histogrammar:** functional API and metalanguage for aggregation.
     with Alexey Svyatkovskiy for Oliver Gutsche and Matteo Cremonesi
   - **Femtocode:** query language for "Scope" and Spark DataFrames.
     with Peter Hansen, for Jin Chang and Igor Mandrichenko

They're all connected



Meta-strategy

- Diverse suite of projects with interconnections:
  - Not everything has to work, but they're all better if they do.
- Focusing on building developer communities before user communities.
  - Aiming for 1–2 users in the beta-testing stage, but as many developers as there are use-cases.
  - This includes developers outside of HEP, if possible.
  - Planning on using focus groups of users to guide design, rather than reacting with release-early-release-often.

Detailed descriptions: Histogrammar

Aggregate data from a variety of backends, plot it in a variety of frontends.

histo·*grammar*
/histō,ˈgɹæm.ər/

Composable aggregators that are good for parallelization, particularly good for functional frameworks like Spark, as well as conceptual and performance advantages for PyROOT.

Status:

- Successful beta-tester stage; user-ready but adoption is slow.
- Introduced to physicists in many one-on-one sessions, to industry in high-profile talks. Everyone was excited while we talked, but few followed up.
- Missed ROOT release because ROOT wasn't ready to include Python subpackages. Will target Scikit-HEP instead.
  (And possibly add a C++ Histogrammar to Enrico Guiraud's or Brian's functional chains.)

Detailed descriptions: Scikit-HEP

rootpy, root_numpy, and Ostap are widely-used Pythonic layers for ROOT.

I suggested combination into a single project to build on existing momentum, probably won't be a major contributor (except for adding Histogrammar).

Also shifting focus from "Python and ROOT" to "Python and HEP."

Note: logo suggestion has not been agreed upon.

Status:

- Everyone's enthusiastic and we discuss progress in regular meetings by Skype and Slack.
- Mostly refactoring/relicensing the original projects, preparing for the big merger.

Detailed descriptions: root4j/spark-root

Follow-up to my early work this year converting ROOT to Avro for Oli & Matteo's analysis and replaces earlier plans of connecting ROOT to Apache Arrow.

File conversion	Direct reading
Makes the bookkeeping problem worse, defeating the main benefit of Spark.	C++ ROOT via JNI is buggy, but the old pure-Java package works (with a little effort).

Status:

- Viktor Khristenko has completely taken over the codebase, and is actively using it on his own thesis. I just need to make sure it works for other use-cases.
- Directly juxtaposed with Danilo & Enric (ROOT Team)'s ROOT-in-Spark. Trying to apply both to the same use-case: Marc Dunser (CMS). Having trouble getting Marc's attention.

Detailed descriptions: Femtocode

I've been thinking about a more abstract numerical language (an "executable chalkboard") for a long time. See earlier PFA project.

After months of talking about it, gauging interest and prior art (inside and outside of HEP), I've found a niche that really needs it: big data pulls. See Monday's talk.

Status:

- Started implementing, off and on, right after CHEP in October. Progress is pretty good, considering.
- I'm finding it hard to describe what Femtocode is and what it's about.
- Peter Hansen has signed on to write the GPU backend, but interactions have been slow. (Femtocode doesn't strictly need a GPU backend, though I think it would be a great target.)

Detailed descriptions: "Scope" (name?)

Early on, I thought HEP needed an Ibis/Impala/Kudu/Drill-type thing to make plots in seconds, rather than private skims in weeks. Oli's interest in Spark won out.

Using Spark's DataFrames, I came to realize that the optimization these systems provide wouldn't be available to non-flat ntuple analysis without Femtocode-like extensions.

Meshed with Jin & Igor's NoSQL LDRD project by convergence. Now they'll need Femtocode with a Sep. 2018 final due date.

Status:

- It's hard for Igor to do much without minimally working Femtocode, but we've been converging on the shape of the project. Progress picked up this week.
- Using off-the-shelf parts wherever possible: NoSQL databases.

Conclusions

Overall goal: widely used code and some publications.

Not much to show yet:

- Histogrammar adoption should be higher than it is.
  - I did spend time evangelizing it, to the detriment of other work.
  - Response has been excitement without application.
  - It will be important in later projects (Femtocode, Scope), so it's not a loss to set it aside for a while.
- All other projects are too early to judge.
- Missed the first round of HEP-computing conference deadlines because they're early in the year and I wasn't ready back then. Gearing up for next year's cycle.
  - On Nan Niu (Cincinnati)'s suggestion, I'm going to submit my work on focus groups to the SE4Science Workshop.
  - Will be working with Brian on a "languages for HEP" white paper; haven't started.
  - Aiming for 1–3 of HPDC, IEEE Big Data, ACAT, CHEP, Supercomputing, PEARC.