Introduction

and…

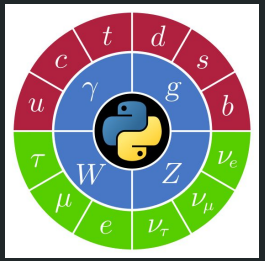
Using HDF5 as an alternative to ROOT files

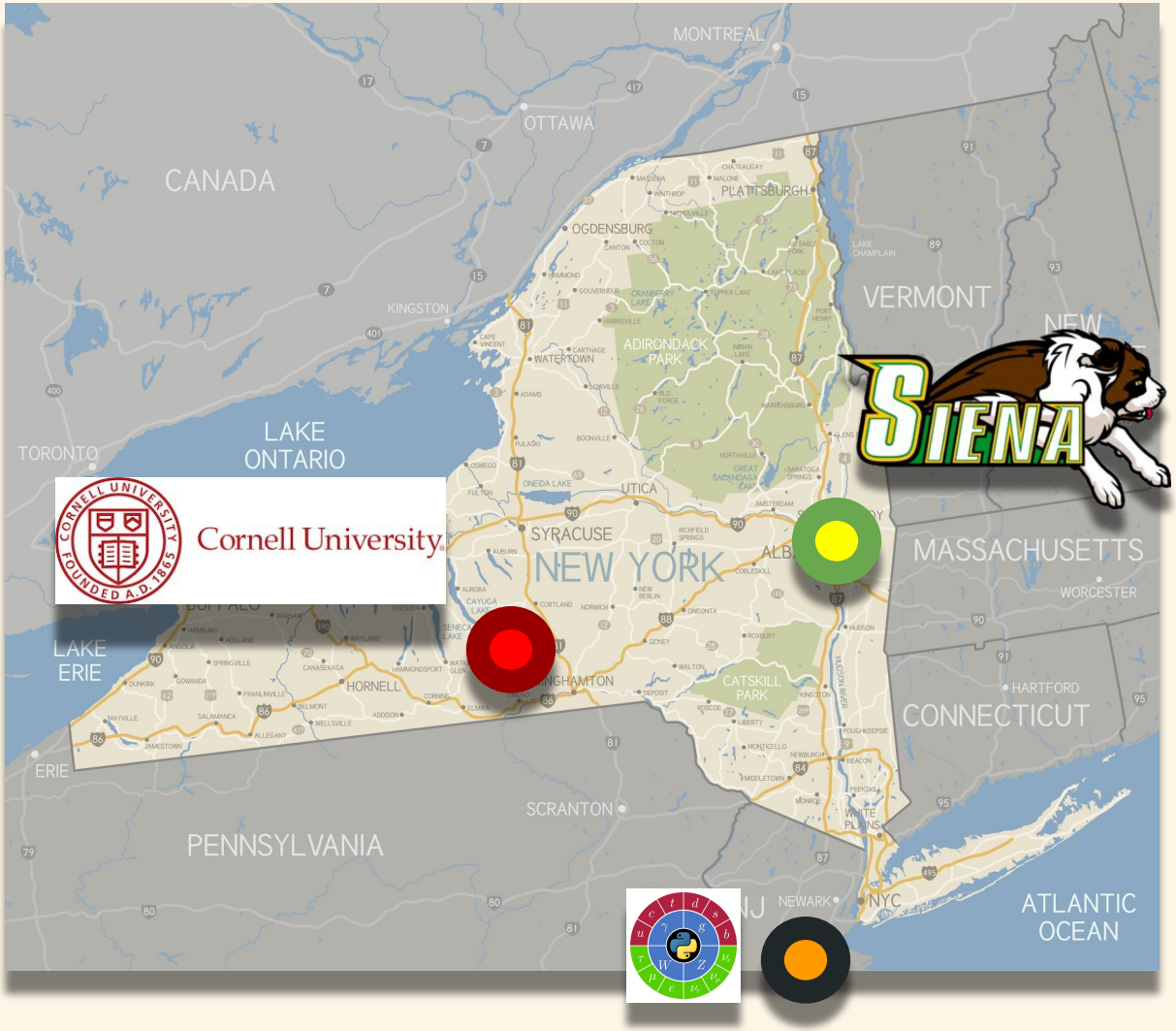
Matt Bellis and Noah Franz

Siena College, Department of Physics and Astronomy

PyHEP.dev workshop, *Princeton University*

7/25/2023





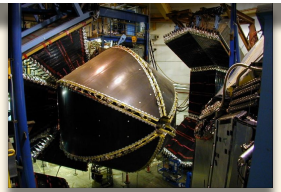
Siena College
Undergraduate-only
institution in upstate-NY

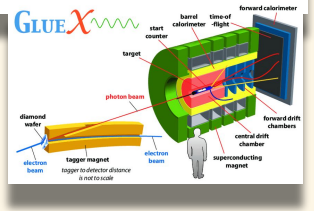
~3500 students

US-CMS constitution allows smaller institutions to partner with R1s, in our case Cornell University

CMS
Compact Muon Solenoid

Jefferson Lab
Thomas Jefferson National Accelerator Facility



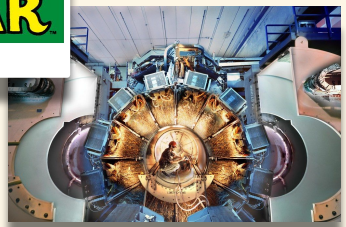


2-4 GeV γ - p

PAW++ → ROOT (1999), C/C++, some prolog, Ruby, Makefiles, Minuit, partial wave analysis, amplitude analysis

Baryon resonances and exotic mesons

BABAR
™ and © Helvene, All Rights Reserved

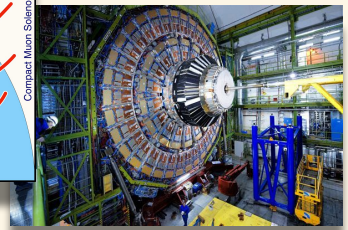


10 GeV e^+e^-

Tcl, gmake, srtpath, some C++, xrootd, PyROOT, python, RooFit, bsub

Baryon-number violation

CMS
Compact Muon Solenoid



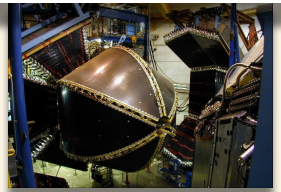
7-13 TeV p-p

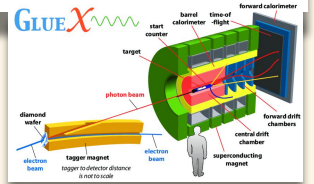
CMSSW, FWlite, scram, crab, condor, slurm, python, numpy/matplotlib/pandas

uproot, awkward, coffea

Top-quark, LLP

Jefferson Lab
Thomas Jefferson National Accelerator Facility

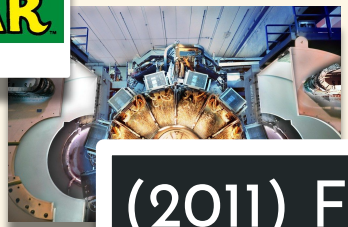




2-4 GeV γ - **p**

PAW++ → ROOT (1999), C/C++, some prolog, Ruby, Makefiles, Minuit, partial wave analysis, amplitude analysis
Baryon resonances and exotic mesons

BABAR

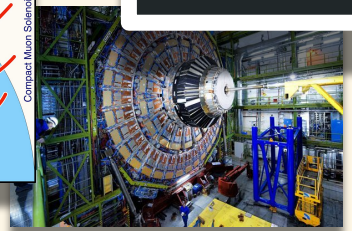


10 GeV **e^+e^-**

Tcl, gmake, srtpath, some C++, xrootd, PyROOT, python, RooFit, bsub

(2011) Frustrations with ROOT → Exploring new file formats / no-OO / iMinuit

CMS
Compact Muon Solenoid

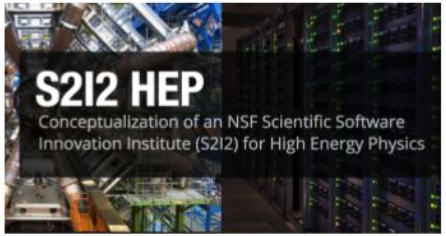


7-13 TeV **p-p**

slurm, python, numpy/matplotlib/pandas

uproot, awkward, coffea
Top-quark, LLP

Data Intensive Analysis & Visualization



S2I2 HEP/CS Workshop
May 1-3, 2017, Princeton, University

Jim Pivarski
Princeton University

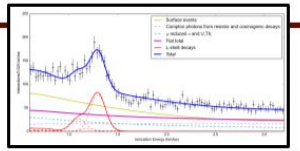
Lightning Talks...

Constructing a ROOT-less workflow with Python and HDF5

- Matt Bellis (*Siena College/Cornell*) - CMS
  - Will ROOT be used for HL-LHC? Or any other experiments 10-20 years in the future?
  - If not, what will we use for fileIO? Development environment? Language/libraries?
  - Need test cases now to see what works and what doesn't. Maybe ROOT ***is*** the right answer!
  - Should we write code to harness maximum benefits of language, rather than writing C-like Python or Python-like C?
  - Should we minimize inheritance to maximize sustainability?
- Related work: Particle Physics Playground (outreach), lichen (wrapper to matplotlib), iminuit)

R&D for alternative workflows

- Lichen - wrapper to matplotlib



- Iminuit - Python wrapper to iminuit (from Piti Ongmongkolkul)
- NEW HDF5 project - replace ROOT files with HDF5/h5py.
  - Early performance looks promising compared to PyROOT!
  - Not using any classes/OO...just to push on this idea
  - Still limited to outreach files now.
  - Will convert and test analysis-grade ROOT files this summer

Survey of data formats and conversion tools

Jim Pivarski

Princeton University – DIANA

May 23, 2017

diana hep

What matters is what you'll use it with

diana hep

ROOT	is the best way to access petabytes of HEP data and use tools developed in HEP
HDF5	is the best way to use tools developed in other sciences, particularly R, MATLAB, HPC
Numpy	is the best way to use the scientific Python ecosystem, particularly recent machine learning software
Avro et al	is the best way to use the Hadoop ecosystem, particularly streaming frameworks like Storm
Parquet	is the best way to use database-like tools in the Hadoop ecosystem, such as SparkSQL
Arrow	is in its infancy, but is already a good way to share data between Python (Pandas) DataFrames and R DataFrames

https://indico.cern.ch/event/613842/contributions/2585787/attachments/1463230/2260889/pivarski-data-formats.pdf

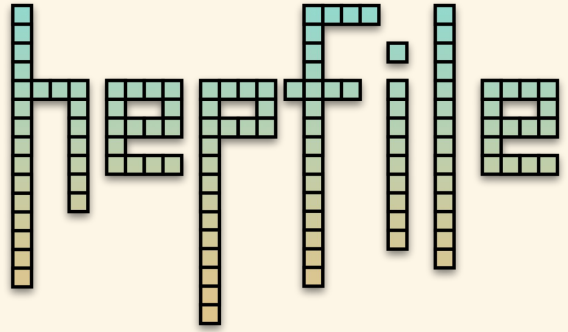
- (2016) proof-of-principle of how to "pack" data in HDF5 files
  - **H5py**
  - Language-agnostic

- (2017) MVP of interface
  - First I wrote down what I would want to type as an analyst to do things...then I tried to figure out how to make that happen

- (2018-2019) Presented **h5hep** at meetings (April APS, HSF JLab)

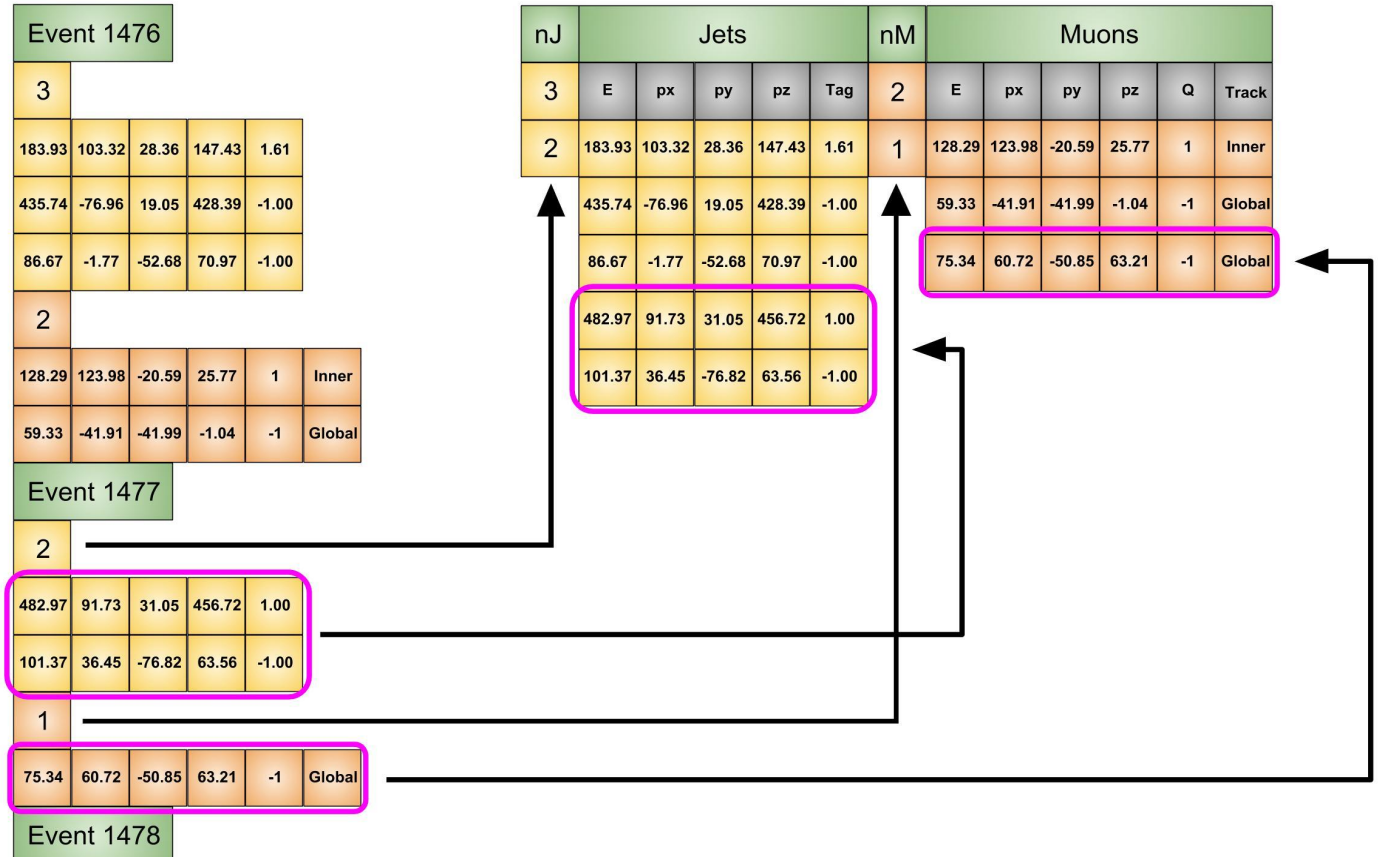
- (2021) DIANA undergrad fellowship (Matt Dreyer) → **hepfile**

- (2023) Major development and documentation with **Noah Franz**
  - v1.0 next week!



https://hepfile.readthedocs.io/en/latest/usage.html

hepfile data storage



hepfile - usage

```
hepfile.create_group(my_data, 'my_group', counter = 'my_counter')
```

```
hepfile.create_dataset(my_data, 'my_dataset', group = 'my_group', dtype = str)
```

```
hepfile.create_dataset(my_data, ['data1', 'data2'] , group = 'my_group')
```

```
for i in range(5)
    my_bucket['my_group/my_dataset'] = 'yes'
    my_bucket['my_group/data1'] = 1.0
    my_bucket['my_group/data2'] = 2.0

my_bucket['my_unique'] = 3

hepfile.pack(my_data, my_bucket)
```

```
hepfile.write_to_file('my_file.hdf5', my_data)
```

hepfile provides methods to pack data efficiently into an HDF5 file

It also provides dictionary definitions that act like a TTree object

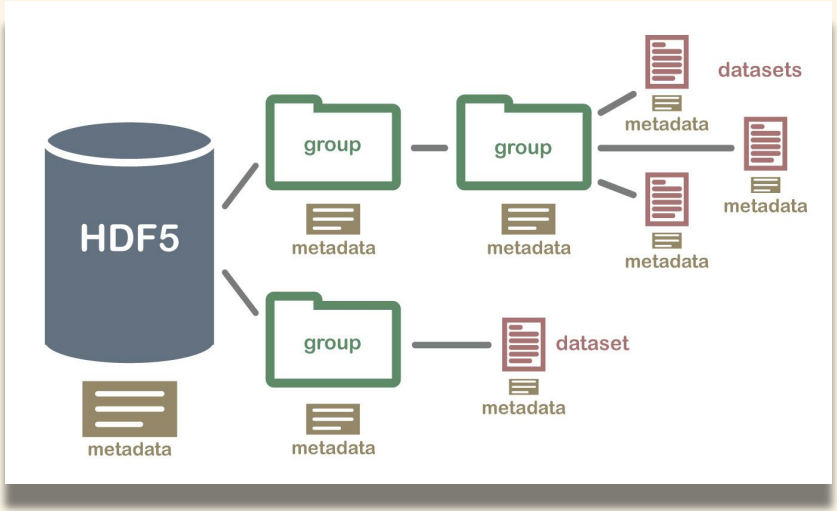
Data is read into memory, but user can pull out subsets from the file

New!

- Can write **awkward** arrays and read data back out as **awkward** arrays/records, regardless of how it was written
- Can write arbitrary header information
- Can return full dataset or individual events as **pandas** dataframes. *Useful for pedagogy!*

Tested it out with non-HEP data

- Database-like info
- Simple histograms
- FITs files (different-sized images)
- FASTA files



Since 2014, built and maintained Particle Physics Playground outreach site

Data from CMS, BaBar, CLEO, Icecube

Simple python tutorials (high school, undergrad)

Data is **h5hep**, soon to be **hepfile**

I have been using **h5hep** extensively for CMS and BaBar analysis for intermediate stages

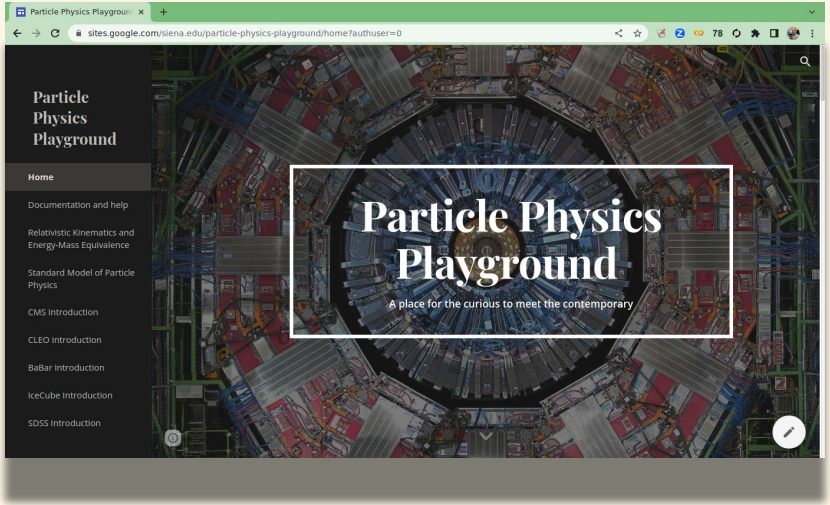
Data is read into memory wholesale…fast!

Writing w/compression can be slow

Tested with small nanoAOD files

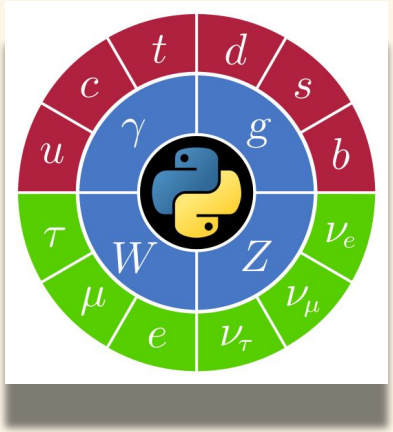
Read out data with Julia just for kicks

FEEDBACK WELCOME!



https://sites.google.com/siena.edu/particle-physics-playground/home

Other interests



Education in the classroom

Training new collaborators (both undergraduate and graduate)

Simplifying workflows.. I do almost all analysis at Siena and I am spread *thin*

Open Data

Documentation

Strawman datasets for fitters (yesterday's discussion)

- https://github.com/mattbellis/datasets-for-testing-fitting-frameworks

I've learned so much in the last few days!

Thank you!