DATA MINING
CLARA INTERFACE

G. Gavalian
(ODU)

OVERVIEW

- Collect the data for all multi-hadron runs. Make them available for instant analysis.

- Provide universal access for all participating universities. Provide software/tools for skimming, analysis and simulation.

- Provide framework for implementing analysis tools, corrections and fiducial cuts.

- Provide interface for fast Monte-Carlo.

DATA STORAGE AND DISTRIBUTION

- The data is stored at ODU on NAS server. Using newly developed DST format using HDF5 library.
- The new DST format allows parallel random access to the data stream which makes parallel multi-process data analysis possible.
- Unique structure of HDF5 allows for indexed data access, which allows reading only events with certain criteria (aka skimming).
- For data distribution CLARA framework is used, with platform running at ODU.
- A GUI interface is developed for data access. Different types of CLARA service are provided for direct data streaming,server side analysis and Monte-Carlo simulations.

PROJECT

- The C++ version of CLARA was underdeveloped since there is no effort spend on expanding the capabilities of SOA architecture.

- We expended the C++ service package to allow multi-process dynamic analysis applications.

- A new file format was employed to allow multi-process analysis applications access the data in parallel mode.

- Random data access was used for data indexing and on-fly skimming of the data, reducing the disk access.

- Multi-RAID data distribution algorithm is currently under development for speeding up the data access. Data will be “smartly” distributed over several RAID disks increasing reading rate.

DATA FORMAT

1 A versatile data model that can represent very complex data objects and a wide variety of metadata.

2 A completely portable file format with no limit on the number or size of data objects in the collection.

3 A software library that runs on a range of computational platforms, from laptops to massively parallel systems, and implements a high-level API with C, C++, Fortran 90, and Java interfaces.

4 A rich set of integrated performance features that allow for access time and storage space optimizations.

5 Tools and applications for managing, manipulating, viewing, and analyzing the data in the collection.

THE PLATFORM

- CLARA is a SOA (Service Oriented Architecture) designed for CLAS12.

- The CLARA platform for DM (Data Mining) runs at ODU with 24 services available.

- DM project is one of the first implementations of working platform for data distribution and analysis.

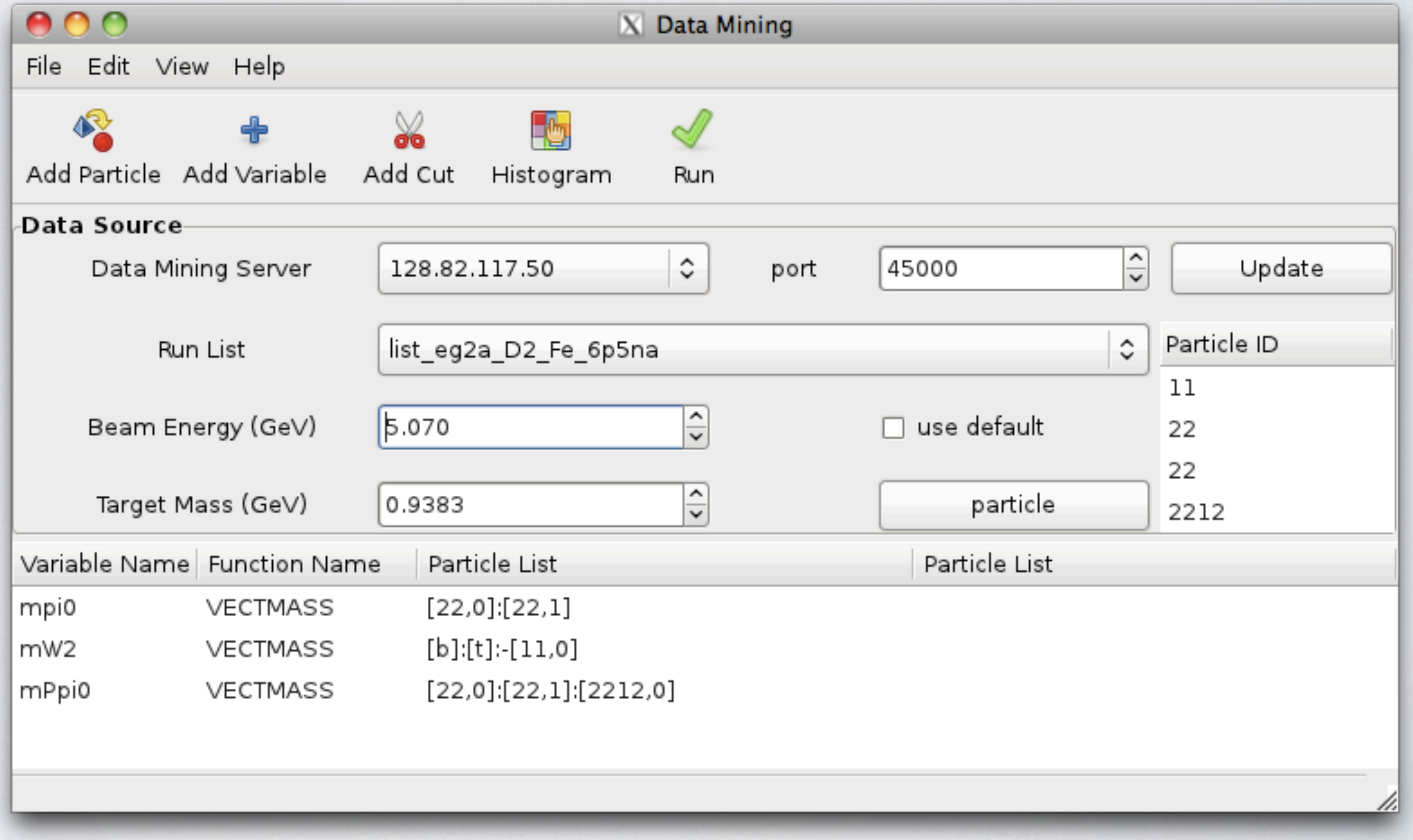
- A lot of improvements/additions were done past month to CLARA to fit the needs of complex multi-process application.

- The C++ interface went through re-design and additions.

- All made possible with intensive help of V.Gurjyan.

GUI TOOLS

- Python powered GUI provides tools for fast analysis of the data with certain pattern of reaction. The analysis are ran on the server and the ntuple is streamed to the client application.



Data Source

Data Mining Server: 128.82.117.50 port 45000

Run List: list_eg2a_D2_Fe_6p5na

Beam Energy (GeV): 5.070 □ use default

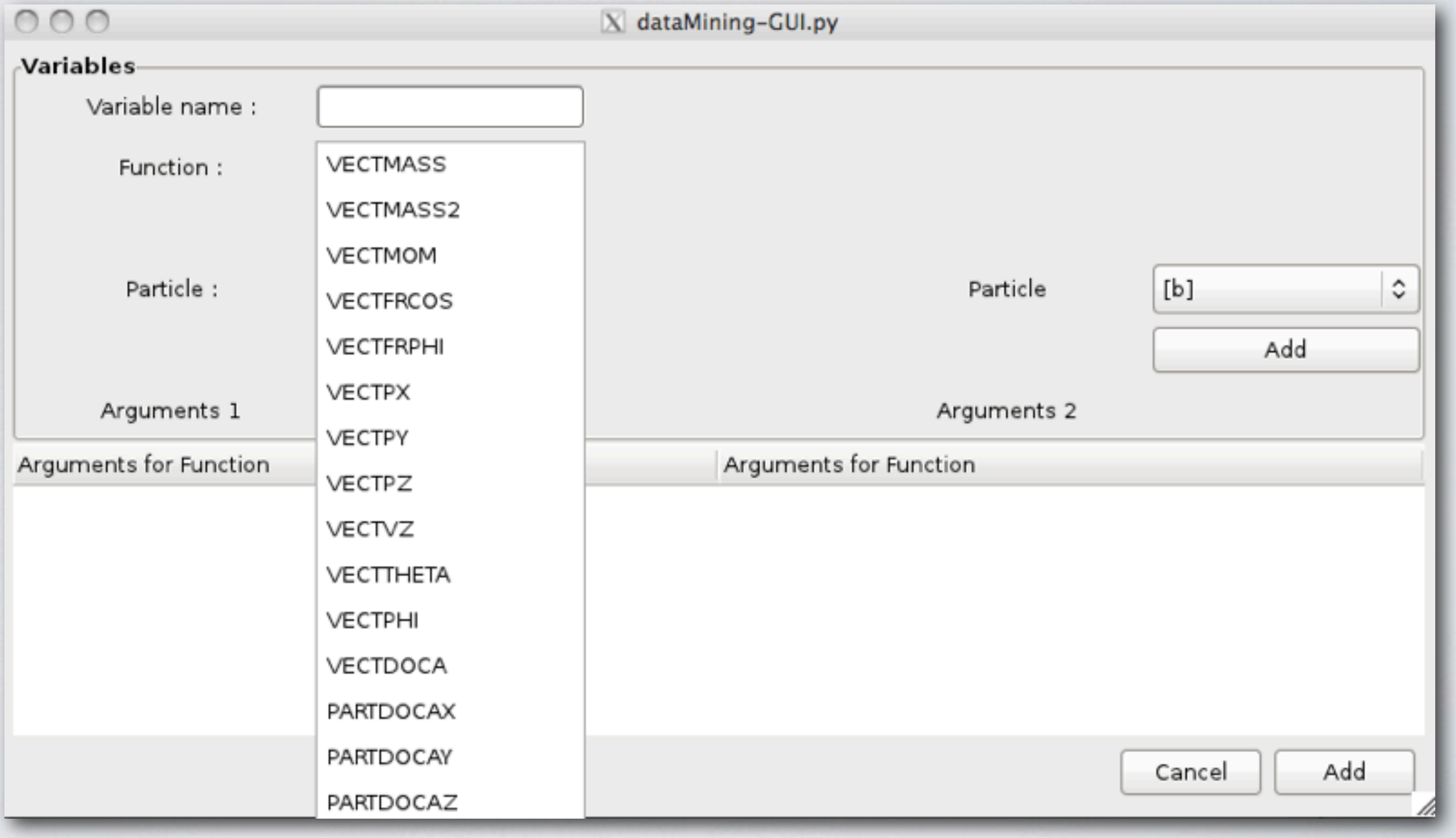
Target Mass (GeV): 0.9383

Particle ID
11
22
22
2212

Variable Name	Function Name	Particle List	Particle List
mpi0	VECTMASS	[22,0]:[22,1]	
mW2	VECTMASS	[b]:[t]:-[11,0]	
mPpi0	VECTMASS	[22,0]:[22,1]:[2212,0]	

GUI TOOLS

- Interface contains standard function set for analyzing the event.
- Abstracted classes allow user to implement their own function sets and submit them to the server.

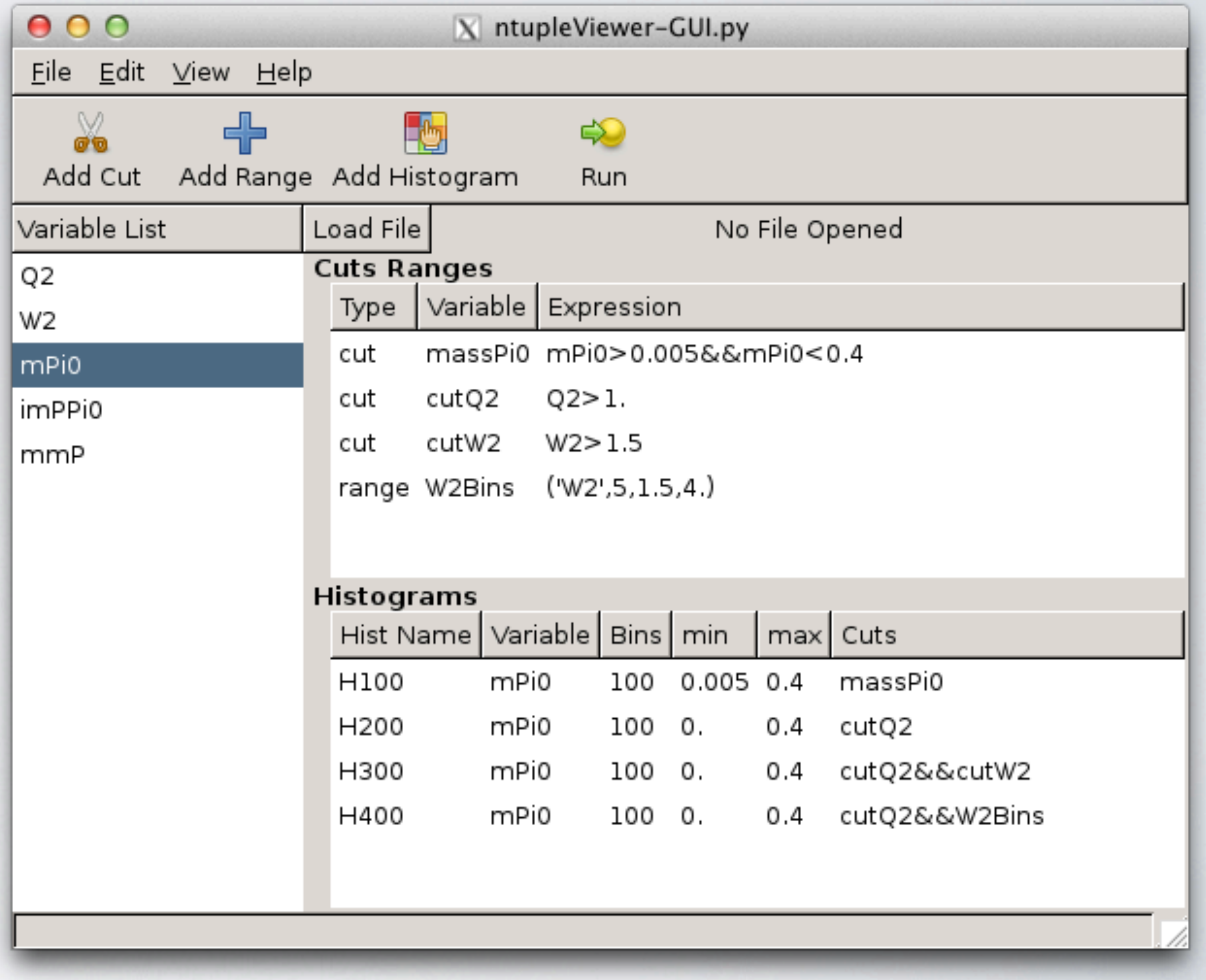


GUI TOOLS

- The streamed ntuple is also in HDF5 format. For convenience there are tools provided for converting them into ROOT.

- Histogram GUI provides tools for making ROOT histogram file with selected cuts and ranges.

- Provides capability to easily create histograms for bins in any variable for multiple variables.



Variable List

Q2

W2

mPi0

imPPi0

mmP

Cuts Ranges

Type	Variable	Expression
cut	massPi0	mPi0>0.005&&mPi0<0.4
cut	cutQ2	Q2>1.
cut	cutW2	W2>1.5
range	W2Bins	('W2',5,1.5,4.)

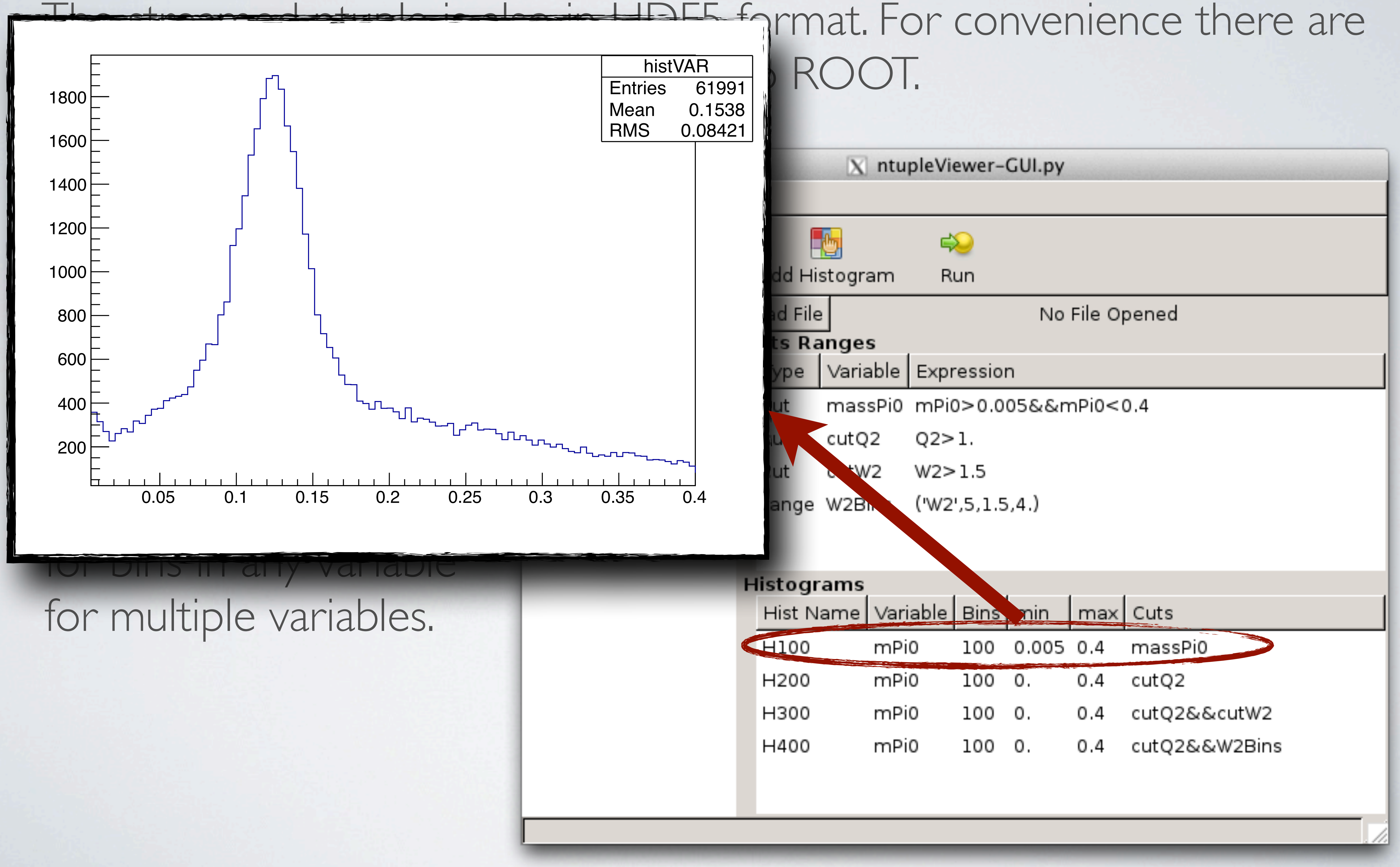
Histograms

Hist Name	Variable	Bins	min	max	Cuts
H100	mPi0	100	0.005	0.4	massPi0
H200	mPi0	100	0.	0.4	cutQ2
H300	mPi0	100	0.	0.4	cutQ2&&cutW2
H400	mPi0	100	0.	0.4	cutQ2&&W2Bins

GUI TOOLS

The streamed ntuple is also in HDF5 format. For convenience there are ROOT.

for bins in any variable for multiple variables.



ts Ranges

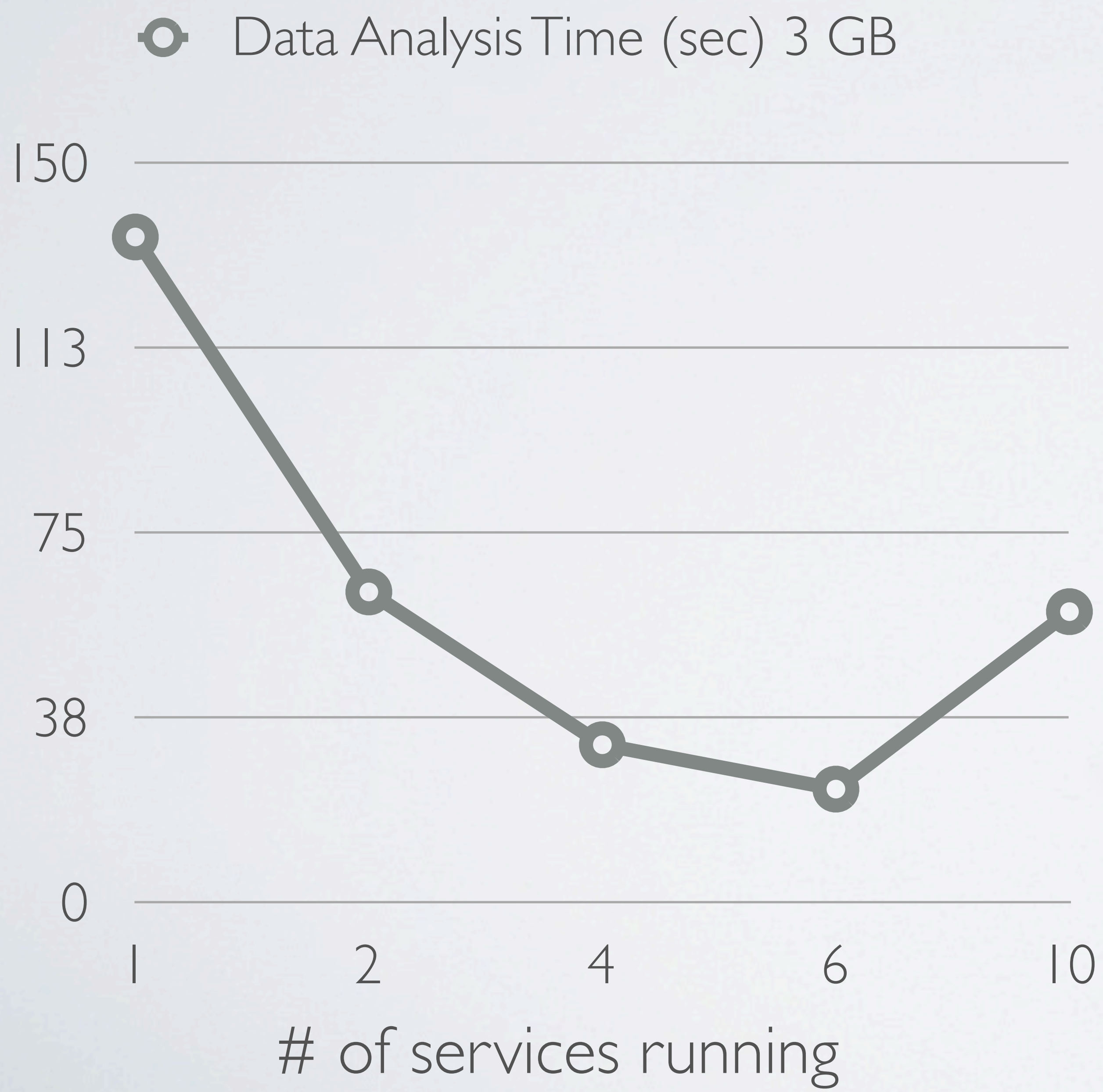
ype	Variable	Expression
ut	massPi0	mPi0>0.005&&mPi0<0.4
u	cutQ2	Q2>1.
ut	cutW2	W2>1.5
ange	W2Bi	('W2',5,1.5,4.)

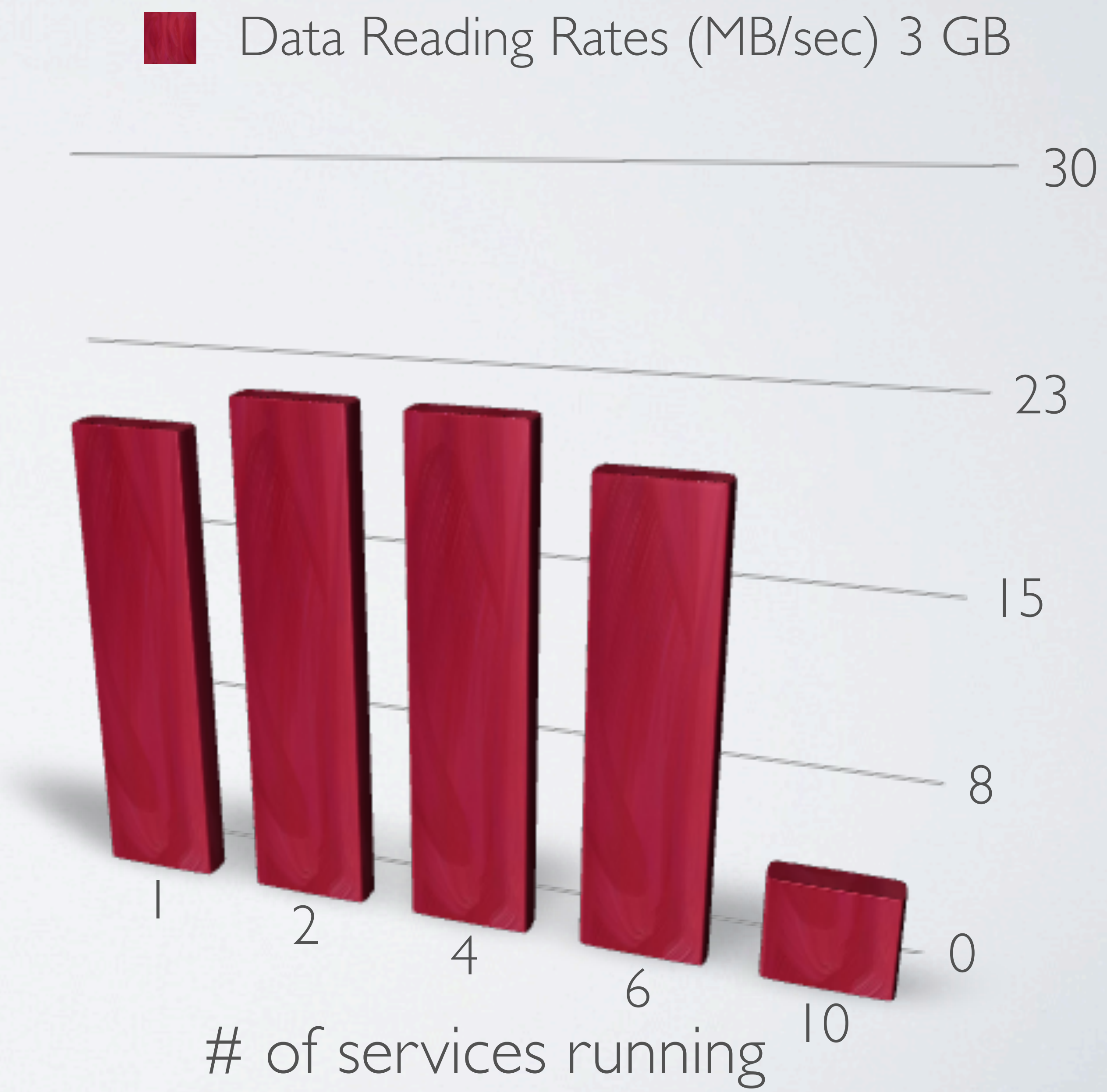
Histograms

Hist Name	Variable	Bins	min	max	Cuts
H100	mPi0	100	0.005	0.4	massPi0
H200	mPi0	100	0.	0.4	cutQ2
H300	mPi0	100	0.	0.4	cutQ2&&cutW2
H400	mPi0	100	0.	0.4	cutQ2&&W2Bins

DATA ANALYSIS RATES

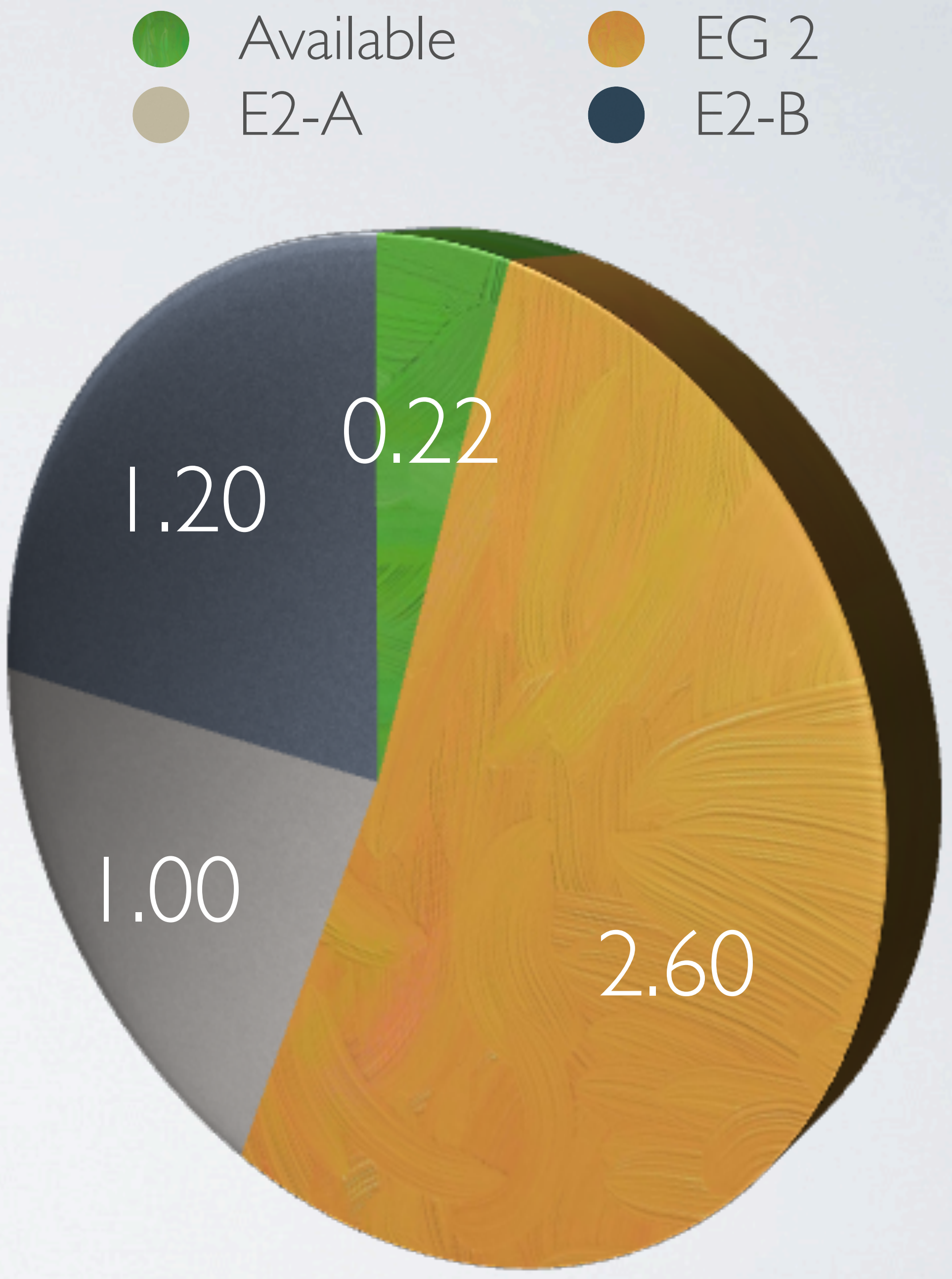
- Test results running CLARA platform on ODU server and client program on JLAB machines. Analysis of eg2 data with good electron selection. Reading data from a single disk.





EXISTING DATA

- Available converted data consists of EG2 data set with Fe/D2 target.
- Planning to have all of EG2 data set transferred to ODU farms in 1.5 month.
- E2-A and E2-B data sets will follow.
- Data is sorted by target, beam energy and beam intensity.



Data Size in TB

SUMMARY

- The framework is designed ground up and requires no additional libraries to run.
- It provides abstracted classes for interacting with a CLAS event. Allowing implementation of alternative EVENT construction, Momentum Corrections and Fiducial cuts.
- The library provides full functionality of the tools using Python wrappers.
- Easy Python scripting for writing CLARA clients to interact with the ODU platform for data download and data analysis.
- GUI interface for easy navigation through available data and analysis methods.

COMING SOON

TO A COMPUTER NEAR YOU