Data Analysis and Visualization at LCLS

Madhu Karra[1], Jason Koglin[1,2]

[1]Stanford University, 450 Serra Mall, Stanford, CA 94305

[2]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA.

[+]Contact: koglin@slac.stanford.edu

Introduction

Every experiment conducted in each LCLS instrument generates massive amounts of data. It is crucial then, to be able to efficiently parse this information in a way that extracts data of utmost relevance. In particular, I worked with tutorial data from CXI and XPP to work towards this goal. This project addresses the need for quick and efficient data analysis and visualization techniques at LCLS. We use an array of statistical tools built into software packages, especially in Python.

Programs

All of my work this summer was done using Python. In particular, I made use of pandas, a Python package that incorporates statistical functionality from R with convenient visualization and plotting methods.

One thing I worked on was figuring out a way to find appropriate Gaussian fits to curves with multiple peaks while subtracting irrelevant noise. I wrote a Python script to find inflection points in graphs and translate their indices to split the graph into several graphs and then do Gaussian fits accordingly. Another thing I worked on was using mathematical plotting tools in Python to generate nice plots. I tested code from run summaries to generate basic plots, then modified them.

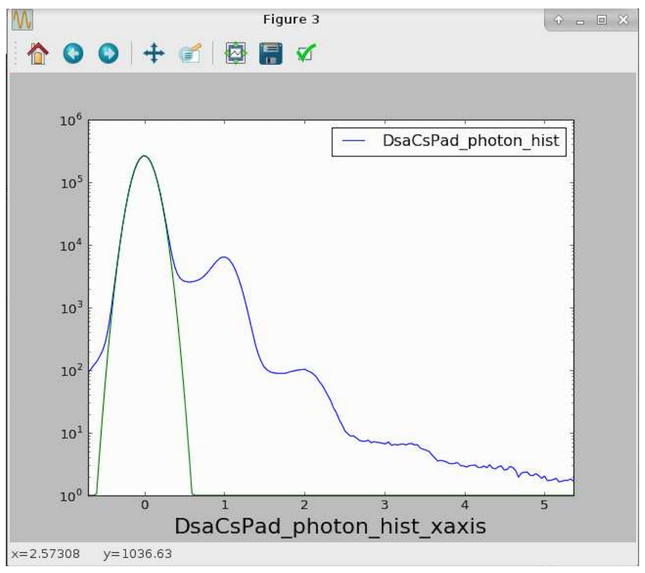

Fig 1. Photon Histogram with first Gaussian fit shown

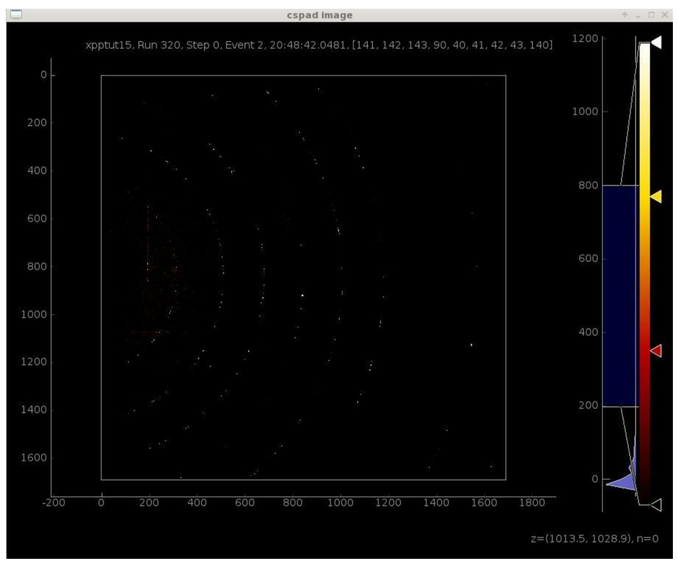

Fig 2. Example plot from xpptut15

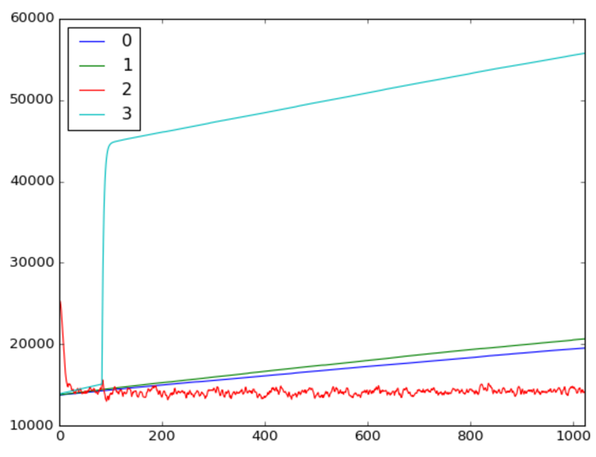

Fig 3. Example plot for waveform data

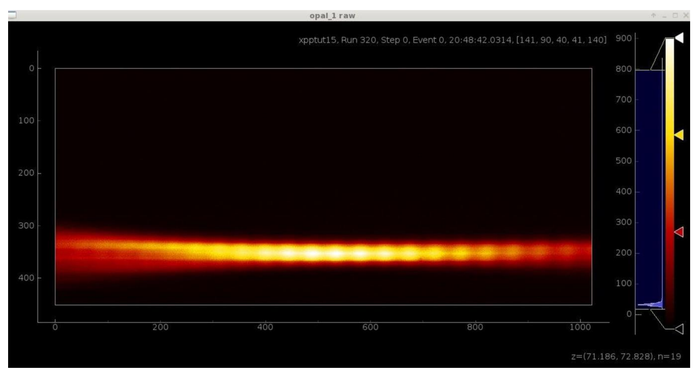

Fig 4. Horizontal projection example

Python Tools

- scikit-learn
- matplotlib
- PyDataSource
- xarray, pandas
- NumPy, SciPy

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a NUMFocus sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to donate to the project.

A Fiscally Sponsored Project of NUMFOCUS OPEN CODE = BETTER SCIENCE

Applications

After figuring out how to manipulate data appropriately, I have been attempting to create run summary reports for most of the publicly available tutorial data to make it easier for users to generate plots.

xpptut15 starting run number	Data origin	Comments
54,59	XPP test DAQ Runs	59 contains epics
101,102,124	cxii0314	xtcav: 101: lasing off, 102: dark, 124: lasing on
140	xppi0813 75-85,87	87 is dark
160	xppc0114 287	
170	xppc0115 270	

Future Directions

There are many directions in which data analysis could be furthered. One such direction would be to consolidate correlations across different groups of data to see which relations are more consistent in the big picture.

Another direction could be to make analysis methods generalizable to different types of data. Machine learning tools could also be applied to data analysis to predict useful information.

Acknowledgments

Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

Date: 09/01/2017