WORKSHOP #6

Data Access and Machine Learning at NSLS-II – A Tutorial

Stuart Campbell, Dan Allan, Phillip Maffettone, Maksim Rakitin, Dan Olds, and Andi Barbour

Are you curious about recent changes in compute infrastructure at NSLS-II? Do you want to get a jumpstart on your upcoming experiment? Are you wondering why some people seem obsessed with python? Join us in a collaborative series designed by beamline and developer staff to learn more about data access and usage for science at NSLS-II. In our second session, we will introduce machine learning (ML) in the context of NSLS-II experiments and show how ML may be used with practical examples. We close with a sneak peak of future AI/ML and support tools at NSLS-II.

- Python knowledge is not required
- Attendance of both sessions is not required, but we encourage it

Hands on Tutorials

Monday, May 23, 2022	**1:00 p.m. -- 5:00 p.m.**	
Start Time (ET)	**Title**	**Speaker (Affiliation)**
1:00 p.m.	Overview of NSLS-II Computational Infrastructure and Plans	Stuart Campbell (NSLS-II, BNL)
1:20 p.m.	**How to Get Your Data Out of NSLS-II**	Dan Allan (NSLS-II, BNL)
2:00 p.m.	**à la carte data analysis at jupyter.nsls2**	Andi Barbour (NSLS-II, BNL)
2:30 p.m.	Web Applications to Access Tiled Data	Marcus Hanwell (NSLS-II, BNL)
2:50 p.m.	---------- *Session Break* ----------	----------
3:20 p.m.	**Case Studies in Data Access**	Dan Allan (NSLS-II, BNL)
3:40 p.m.	**Basic Powder Diffraction Data Inspection in Jupyter**	Dan Olds (NSLS-II, BNL)
4:05 p.m.	Tiled is useful and you should use it! – An AIMM use case	Juan Marulanda Arias (NSLS-II, BNL)
	---------- *Lightening Talks* ----------	----------
4:40 p.m.	Data Workflows with Prefect	Abigail Giles (NSLS-II, BNL)
4:50 p.m.	Uploading to Tiled	Juan Marulanda Arias (NSLS-II, BNL)
5:00 p.m.	Matplotlib Update	Thomas Caswell (NSLS-II, BNL ; matplotlib's lead developer)

Wednesday, May 25, 2022	1:00 p.m. -- 5:00 p.m.	
1:00 p.m.	Introduction to applying AI/ML at NSLS-II	Phillip Maffettone (NSLS-II, BNL)
1:25 p.m.	**Unsupervised Learning Methods for Rapid Data Analysis @ NSLS-II**	Dan Olds (NSLS-II, BNL)
1:55 p.m.	**Anomaly Detection with ML and Scalar Time Series**	Andi Barbour (NSLS-II, BNL)
2:25 p.m.	Integrating Simulation, Analysis, and Controls Workflows to Support Scientific User Facilities	Nathan Cook (RadiaSoft LLC)
2:30 p.m.	---------- *Session Break* ----------	----------
2:45 p.m.	**XAFS Data Validation by Supervised Learning**	Bruce Ravel (NIST)
3:15 p.m.	**Anomaly Detection for High-Dimensional Data with Autoencoders**	Tatiana Konstantinova (NSLS-II, BNL)
	---------- *Lightening Talks* ----------	----------
3:50 p.m.	Removing Single Crystal Bragg Spots from Scattering Data with Unsupervised Machine Learning Algorithms	Danielle Alverson (University of Florida)
4:00 p.m.	24/7 access to your virtual beamline with Sirepo	Maksim Rakitin (NSLS-II, BNL)
4:10 p.m.	Optimize on the Fly with AI	Thomas Morris (NSLS-II, BNL)
4:20 p.m.	AI Driven Decision Making at NSLS-II	Phillip Maffettone (NSLS-II, BNL)
4:30 p.m.	Ultrafast Transmission X-ray Imaging Aided by Machine Learning	Mingyuan Ge (NSLS-II, BNL)

Abstracts to Follow on the Next Page

Abstracts

SESSION I – Tutorial Talks

Overview of NSLS-II Computational Infrastructure and Plans (15 minutes)

Stuart Campbell (NSLS-II, Brookhaven National Lab)

> This talk will present a brief overview of the computational infrastructure at the NSLS-II. Recent changes and updates to systems and infrastructure will be highlighted. The future roadmap and plans will also be presented.

How to Get Your Data Out of NSLS-II

Dan Allan (NSLS-II, Brookhaven National Lab)

à la carte data analysis at jupyter.nsls2

Andi Barbour (NSLS-II, Brookhaven National Lab)

We will use jupyter notebooks to explore a variety of x-ray scattering data collected with bluesky. The main goal is to demonstrate (with hands-on examples) how jupyter, tiled and python can help support different aspects of analysis workflows, human or computer-based.

Web Application to Access Tiled Data

Marcus Hanwell (NSLS-II, Brookhaven National Lab)

> A single-page web application consumes RESTful web interfaces to offer an interactive experience within the web browser. The development of a simple, single-page web application will be discussed, and how it interacts with the RESTful interfaces offered by the Tiled server. The basic structure of the application will be shown along with a short demonstration of the application on some representative data.

Case Studies in Data Access

Dan Allan (NSLS-II, Brookhaven National Lab)

Basic Powder Diffraction Data Inspection in Jupyter

Dan Olds (NSLS-II, Brookhaven National Lab)

> This tutorial will cover some of the basic functionality of reading reduced powder diffraction datasets into jupyter via Tiled, and simple data visualization and manipulation.

Tiled is useful and you should use it! – An AIMM use case

Juan Marulanda Arias (NSLS-II, Brookhaven National Lab)

The AIMM project is a joint and collaborative effort between beamline scientist from different National Labs. Our goal is to enable and accelerate scientific discovery by leveraging large, complex multimodal datasets generated across BES synchrotron facilities. We are developing different tools to create easy experiences for beamline users. In this case, we are putting together the data access service of Tiled and the interactive experience of ipywidgets.

SESSION I – Lightening Talks

Data Workflows with Prefect

Abigail Giles (NSLS-II, Brookhaven National Lab)

To handle the data processing needs of the NSLS-II beamlines, we can use Prefect, which is a workflow management tool that allows users to run and monitor data pipelines. By using Prefect, we can keep track of any processing failures, we can get notified when the data is ready to be analyzed, and we can validate that we can access the data. In this talk we are going to take a look at how we can utilize Prefect to automate both data processing and moving data into proposal directories.

Tiled – The writer module

Juan Marulanda Arias (NSLS-II, Brookhaven National Lab)

Tiled is constantly growing. In this presentation, we are showing a new feature of Tiled that is currently in the works. The writer module will allow the user to pass/upload new data to a running Tiled server. Tiled will run a strict validation process of the data structure in the background while the user only needs to worry few simple parameters.

Matplotlib Update

Thomas Caswell (NSLS-II, Brookhaven National Lab and Matplotlib Lead Developer)

This talk will briefly cover new features added in recent Matplotlib releases (3.3, 3.4, 3.5) and the upcoming future release (3.6). There will be a brief discussion on the APIs of Matplotlib, when to use them, and how to make your own.

SESSION II – Tutorial Talks

Introduction to applying AI/ML at NSLS-II

Phillip Maffettone (NSLS-II, Brookhaven National Lab)

Here, we will introduce some core concepts of machine learning to familiarize the attendees with the vocabulary and workflows of ML. This will prepare users who are unfamiliar with ML to engage with the remaining tutorials of the session. Lastly, we a framework to integrate your own algorithms into experiments accomplished at NSLS-II with Bluesky.

Unsupervised Learning Methods for Rapid Data Analysis @ NSLS-II

Dan Olds (NSLS-II, Brookhaven National Lab)

Unsupervised learning methods are an attractive approach for rapid, model free analysis of data streams. In this tutorial, users will test different unsupervised learning approaches on data from the PDF beamline at NSLS-II including hierarchical clustering, Principle Component Analysis (PCA), Non-negative Matrix Factorization (NMF), and Constrained Matrix Factorization (CMF).

Anomaly Detection with ML and Scalar Time Series

Andi Barbour (NSLS-II, Brookhaven National Lab)

Data rates at X-ray facilities are becoming too large for human researchers to verify every single data set is "normal" or collected under normal conditions. Many scatter measurements require optical and sample stability. X-ray Photon Correlation Spectroscopy (XPCS), in particular requires, more than stable flux, as the source positions plays a role in contaminating the results. We discuss our approach to detecting anomalies with a semi-supervised approach using common supervised outlier detection models: Local Outlier Detection (LOD), Elliptical Envelope (EE), and Isolation Forest (IFT). Data are generated at the CSX beamline.

Integrating Simulation, Analysis, and Controls Workflows to Support Scientific User Facilities

Nathan Cook (RadiaSoft LLC)

RadiaSoft develops and maintains the Sirepo scientific gateway to support an array of community software and analysis tools, featuring a browser-based GUI and an embedded JupyterHub instance. We will highlight capabilities designed to address challenges faced by users and facility scientists in simulating X-ray beamlines, analyzing data, and making informed controls decisions, with examples specific to the NSLS-II.

XAFS Data Validation by Supervised Learning

Bruce Ravel (National Institute of Standards and Technology)

At BMM, we try to use all 24 hours of every operations day. That level of efficiency requires recognizing when things have gone awry with an experiment and triggering a response from beamline staff. Using a corpus of successful and unsuccessful XAFS measurements made at BMM, we have trained a simple machine agent to identify when a measurement does not look like an XAFS spectrum. This validation tool is not much more complex than the iris flower classification problem, and it is easily integrated into regular operations.

Anomaly Detection for High-Dimensional Data with Autoencoders

Tatiana Konstantinova (NSLS-II, Brookhaven National Lab)

Anomaly detection methods are the most efficient for data with few dimensions. However, many of experiments at NSLS-II produce high-dimensional signal representations: images, spectra, diffraction patterns, etc. Determining similarity between such signal becomes computationally expensive and suffers from the effect called 'curse of dimensionality'. A solution to this problem is finding efficient method of dimensionality reduction that retains the essential information about the signal structure. In this talk, I will demonstrate the application of an autoencoder model for compression of two-time photon correlation functions (X-ray Photon Correlation Spectroscopy) and successive application of anomaly detection algorithms. The principles of the model development will be illustrated with an accompanying Jupyter notebook.

SESSION II – Lightening Talks

Removing Single Crystal Bragg Spots from Scattering Data with Unsupervised Machine Learning Algorithms

Danielle Alverson (University of Florida)

Thin-film materials are used in many electronic applications, and for structure-property analysis utilizing X-ray scattering techniques, the thin films are deposited on amorphous substrates. But, for a more accurate representation and characterization of the material, these measurements should be conducted on single-crystal substrates; however, high-intensity Bragg spots from the single crystal substrate leave the scattering data quality unsuitable for structure analysis. Therefore, by using Non-Negative Matrix Factorization and hierarchical clustering algorithms, sorting and separating the thin film and single crystal substrate signals will yield useable scattering data from the thin-film material to determine its atomic structure.

24/7 access to your virtual beamline with Sirepo

Maksim Rakitin (NSLS-II, Brookhaven National Lab)

Sirepo is a user-friendly simulation interface allowing for modeling of the X-ray sources and beamlines using simulation codes such as Synchrotron Radiation Workshop (SRW) and Shadow3. The Sirepo-Bluesky library allows using Sirepo as a virtual detector for the Bluesky data collection framework, enabling automated scanning of various parameters of the optics in simulations and recording of the image data of the resulting intensity distributions. That data becomes available via the Databroker API in the same fashion as the data from physical detectors on the beamlines and enables the use of the datasets with AI/ML algorithms.

Optimize on the Fly with AI

Thomas Morris (NSLS-II, Brookhaven National Lab)

Beamline alignment is often a cumbersome and time-intensive task, due to the many degrees of freedom and the high degree of sensitivity to misalignment of each optical element. As the beam propagates in a nonlinear way which is difficult to model analytically, it is a natural target for machine learning, which can aid in the auto-alignment of the optical system. Starting with the TES beamline (8-BM) as a proof-of-concept, we hope to produce a tool that can optimize any beamline on the fly and in integration with current NSLS-II Bluesky environments.

AI driven decision making at NSLS-II

Phillip Maffettone (NSLS-II, Brookhaven National Lab)

This talk will focus on how adaptive and reinforcement learning can be applied at NSLS-II. We will discuss two examples where these approaches have made an impact, and how to integrate new techniques into experiments at NSLS-II.

Ultrafast transmission X-ray imaging aided by machine learning

Mingyuan Ge (NSLS-II, Brookhaven National Lab)

In this talk, we will present a machine-learning-assisted imaging analysis that enables ultrafast fly-scan nanotomography. With this approach, we achieved 10 seconds nanotomogrpahy with sub-50 nm resolution, which is about 5-10X faster than the current state-of-the-art measurements.