Summary of 2017 NSLS-II/CFN Users Meeting Workshop

Workshop 1: Multi-dimensional and multi-modal x-ray imaging and analysis

Written by the organizers - Yu-chen Karen Chen-Wiegart, (Stony Brook University; NSLS-II, BNL), Lisa Miller, (NSLS-II, BNL), Juergen Thieme, (NSLS-II, BNL), Yong Chu, (NSLS-II, BNL), Thomas Caswell (NSLS-II, BNL), and Stuart Campbell (NSLS-II, BNL)

Last updated date: 2017 July 4th

Figure 1 – Multi-dimensional and Multi-modal X-ray Imaging & Analysis workshop organizers and attendees at Brookhaven National Laboratory, May 15, 2017.

Overview

Across all wavelengths, multi-dimensional and multi-modal imaging are very powerful techniques advancing research in many different areas of science. Multi-dimensional imaging includes techniques such as tomography, spectroscopic microscopy, or *in situ/operando* imaging, enabling the characterization of samples with structural and chemical sensitivity, in space and time. Multi-modal imaging combines imaging techniques across different length-scales with different contrast mechanisms or with other interaction mechanisms, such as diffraction or scattering. Individually or in combination, these techniques can provide unprecedented analysis of samples on multi length-scales and with enriched information, as illustrated in

Figure 2.

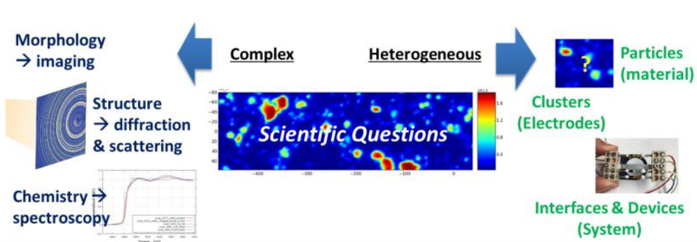

Figure 2 – The need of using multi-modal synchrotron characterization to address scientific questions, and how heterogeneous and complex samples can profit from such a approach

Big challenges for the usefulness of this powerful research approach, however, lie in the handling of the acquired data. The complex nature of multi-dimensional and multi-modal imaging requires the development of specialized and sophisticated tools for data collection and management, analysis, quantification and visualization. Only then can the scientific community can make use of these techniques to the full extent of its capability.

Focusing on x-ray imaging at the synchrotron radiation facility NSLS-II, a one-day workshop on May 15th 2017, brought together experts on four specific, but crucial themes. The aim was to facilitate in-depth discussions on these challenges with the NSLS-II user community. These topics were (1) the grand challenges on multi-dimensional and multi-modal imaging and analysis, (2) the improvement on multi-dimensional and multi-modal data collection and pre-processing, (3) the multi-modal contrast and length scale registration and correlation, and (4) the quantification, visualization and modeling in multi-dimensional and multi-modal imaging. Excellent presentations and lively discussions throughout the workshop highlighted the existing challenges and ways were presented to improve the power of existing tools and to develop new tools to optimize the effectiveness of multi-dimensional and multi-modal x-ray imaging.

The photographs showed the workshop participants (***Figure 1***) and organizing committee (***Figure 3***). The workshop agenda and all the presentations can be found on the workshop webpage: http://usersmeeting.ps.bnl.gov/workshops/workshop.aspx?year=2017&id=137. This article summarizes each of the themes, as discussed below, with a summary and conclusion at the end.

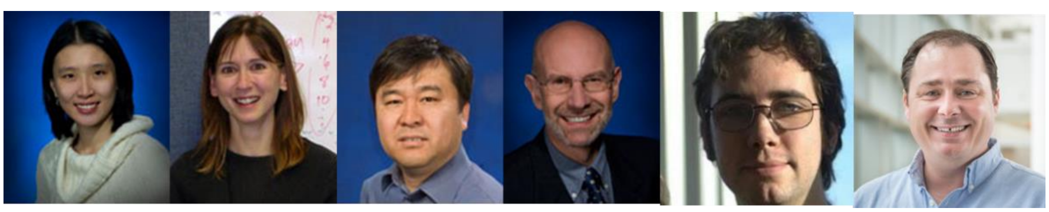

Figure 3 – Workshop organizers – (left to right) Yu-chen Karen Chen-Wiegart, (Stony Brook University; NSLS-II, BNL), Lisa Miller, (NSLS-II, BNL), Juergen Thieme, (NSLS-II, BNL), Yong Chu, (NSLS-II, BNL), Thomas Caswell (NSLS-II, BNL), and Stuart Campbell (NSLS-II, BNL)

Theme I. Grand challenges on multi-dimensional and multi-modal imaging and analysis

The workshop was kicked off with presentations by Stuart Campbell and Mark Basham from the scientific data analysis groups at NSLS-II and the Diamond Light Source, respectively. Each provided a perspective on the grand challenges and approaches taken at each of their facilities. Not surprisingly, many similarities were identified and discussed further throughout the day.

Issues were identified in areas of user access (e.g. computer accounts, proposal logistics, and scheduling beam time), sample preparation, data collection, processing of multimodal data, and statistical analysis / data mining of the multimodal data. Much of the further discussion focused on the data structure, viewing, and analysis. In both talks, the speakers emphasize that, in order to be successful, these issues must be addressed *prior to implementing multimodal imaging approaches*, rather than trying to retrofit data after they are collected. Basham noted that there should be no need for the creation of "dark data", which is inevitable without an early consideration of these grand challenges.

Emphasis was put on the quality of the metadata, ensuring that the appropriate data are stored and made accessible across technique/instrument/software platforms. Campbell showed an example of an interactive data browser developed at NSLS-II that enables searching of the metadata fields in order to generate a plot of any particular aspect of the data. Of course this requires that data architecture is accessible and interpretable across techniques, platforms, and even different facilities.

Basham made the point that multimodal imaging is a big data problem, combining the volume, complexity, variety, and speed at which data are now collected. These data require special attention in order to reduce the volume and/or parallelize the processing without losing the valuable multimodal information obtained by combining techniques.

In the area of software for data analysis and visualization, the point was made by both Campbell and Basham that many powerful tools exist today, but a multi-facility effort is needed to decide what software to adopt and to make data compatible. Basham showed an example of Savu, which is a multimodal tomography analysis tool developed by Diamond scientists and collaborators but used at a number of other facilities.

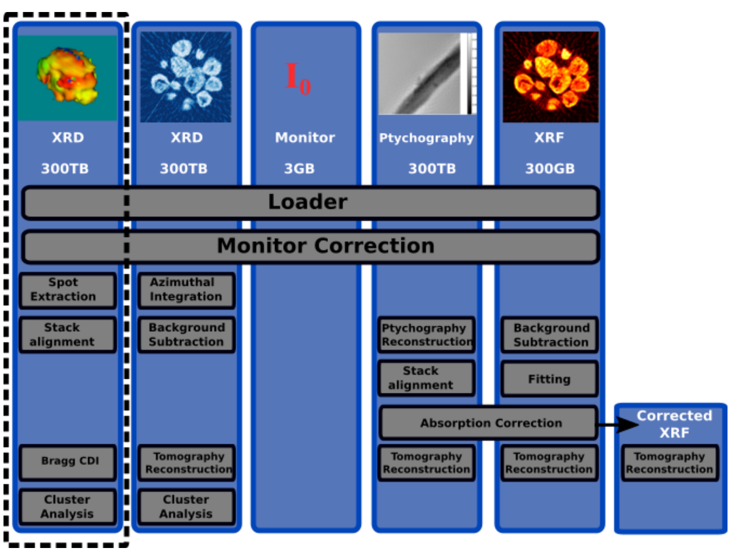

Figure 4 – Summary of a multimodal imaging activity at Diamond (image credit: Mark Basham)

A related software issue its location and availability, i.e. should it be installed locally on a user's computer, or should it be accessible for use across the network. It was agreed that a hybrid approach is likely to be the most successful, although the future directions are aimed at a more network-driven approach.

After the presentations, follow up discussion with the workshop participants reemphasized the need to put resources toward these grand challenges up front in order to achieve the most impactful outcome. Suggested follow-up action items from this section are:

- *Pilot studies across techniques and facilities are necessary for testing different modes of user access, sample preparation, and data collection.*
- *And for data handling, processing, and analysis, the participants agreed that there is a strong need for more collaboration amongst software developers, where the final product is a GUI-driven, user-friendly set of tools that become familiar to the user community regardless of the technique, instrument, or technique used for the multimodal imaging study.*

Theme II. Improvement on multi-dimensional and multi-modal data collection and pre-processing

Quantitative analysis of 3D data set requires accurate data processing, so that the reconstructed volumetric data is not altered by reconstruction artifacts. Patrick La Riviere from University of Chicago gave a lecture on novel reconstruction algorithms and acquisition strategies for x-ray fluorescence computed tomography or XRF-CT. Fluorescence x-rays, particularly those with low emission energies, are absorbed in the sample before measured by a detector. The amount of absorption depends on local density and composition of the sample. When uncorrected, the reconstructed data exhibit systematic artifacts, which are more serious for lower energy XRF signals. Patrick presented an iterative solution with statistical penalization which works well with noisy data. He also presented novel methods of acquiring XRF tomography dataset without scanning but using a concept of pinhole imaging (***Figure 5***).

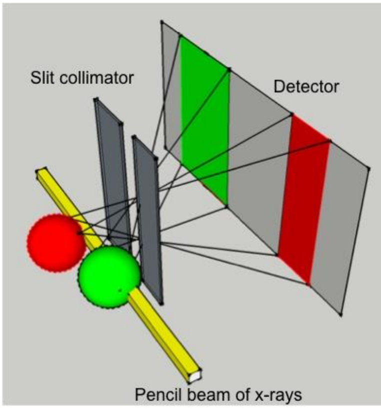

Figure 5 – slit imaging with pencil beam illumination (image credit: Patrick La Riviere)

Doga Gursoy from the Advanced Photon Source presented challenges in aligning 3D dataset for nanotomography. With higher imaging resolutions, imperfection of sample rotation motion or run-out errors contribute to significant challenges with greater magnitudes. Without correction, the reconstructed results exhibit significant streak artifacts and degradation of resolution. Doga presented an iterative solution for aligning 3D dataset, correcting translational errors. Doga also discussed other types of errors including frame-cut, partial-loss, and stretch and blurring, which are more important for scanning microscopy imaging. He discussed a novel approach of solving these problems implementing frame work of optical flow.

There is one major suggested follow-up item came out from the follow discussion:

- *The workshop participants recognized importance of sharing data so that new developed algorithms can be tested with different dataset from different types of samples. A cross-facility collaboration will be beneficial, via effort such as TomoBank (https://aps.anl.gov/tomobank) established by APS (**Figure 6**).*

Figure 6 – TomoBank, 16,759 files associated with 76 datasets from all major synchrotron facilities around the world were deposited onto this shared data bank to facilitate collaboration across research disciplines and institutions.

(image credit: Doga Gursoy)

Theme III. Multi-modal contrast & length scale registration and correlation.

The motivation of this theme is that, image registration is a critical component in being able to fully exploit the power of performing multi-modal experiments. Without it, it is impossible to correlate different modalities in any qualitative or quantitative way.

Matthew McCormick from Kitware (https://www.kitware.com/) gave an interesting overview of multi-dimensional and multi-modal image registration methods and how these techniques have been incorporated into Kitware products. He gave an overview of some of the multi-modal challenges including different length scales, non-corresponding intensities and imperfect systems that introducve noise and artifacts. Matt then went on to describe some of the potential solutions both from an an algorithmic and software tools point of view. The Insight Toolkit (ITK) which provides an open source suite of tools for performing image analysis, such as registration and segmentation of multi-dimensional data. He then gave an overview of Tomviz which is a cross platform, open source application for the processing, visualization and analysis of tomographic data (***Figure 7***).

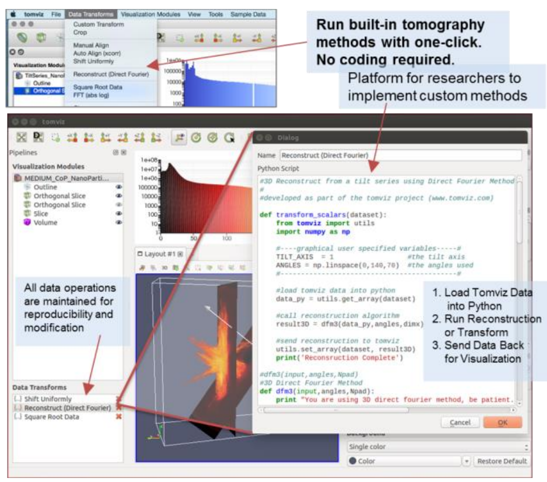

Figure 7 – Open Source: Tomviz Python operators and scientific packages, with built-in Python scripting integration (image credit: Matthew McCormick)

Yijin Liu from SSRL presented his thoughts on data mining in correlative multi-modal X-ray microscopy. He pointed out that the point of big data mining is to ultimately arrive at new scientific findings. Multi-modal is important because scientifically interesting phenonenon exists at multiple length scales, and spatial resolution is very important to investigate this correctly. He then went on to describe several real world problems, such as CO_2 sequestration and industrial catalysis. Yijin described the wide range of imaging platforms that are available at SLAC/Stanford together with some use cases that have made use of them to deliver real world scientific results (***Figure 8***). He concluded by pointing out that the number of publications produced is not scaling with the amount of diffraction patterns collected.

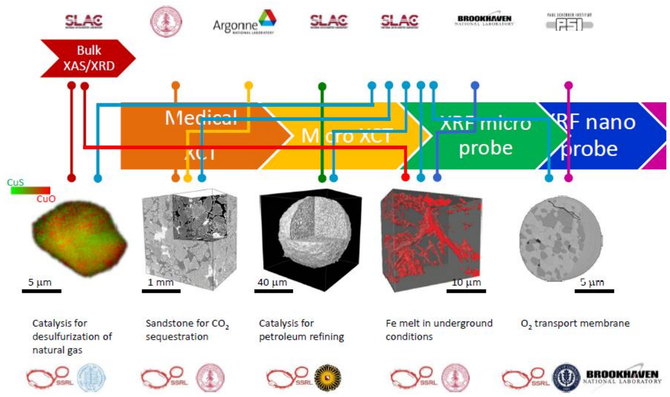

Figure 8 – Scientific cases for correlative X-ray microscopy across different length scales (image credit: Yijin Liu)

In the follow- up discussions, there was discussion on how companies such as Kitware can collaborate with Universities and Laboratories, this ranges from joint grants to running on-site workshops. It was clear from the scientific cases presented from the SSRL that a lot can currently be achieved in the multi-modal space by a single PI, but it is not done in a coordinated manner and is not easily accessible for everyone to routinely do. How do we collaborate to close the 'gap' between having cutting edge algorithms available, but not necessarily implemented for the end-user? The consensus was that the facility has a crucial role to play here as it has the unique position to be able to bridge between the computational science and user science worlds. The problem of recording the workflow and provenance of the data, was raised as an important topic in today's growing climate of promoting reproducible science. The importance of making sure any solution produced by facilities was open source and available by all users was raised.

Suggested follow-up action items from this section are:

- *Facilities to put effort into coordinating their multi-modal efforts to make them accessible to everyone., and*
- *Investigate how facilities and companies can work together to provide better solutions for users*

Theme IV. Quantification, visualization and modeling in multi-dimensional and multi-modal imaging.

As beamlines collect a greater volume and variety of data it is becoming more important to be able to visualize mulit-scale multi-modal data simultaneously, to automatically extract insight, and to be able to find 'similar' data to drive discovery.

Mike Marsh from Object Research Systems, Inc introduced Dragonfly (http://www.theobjects.com/dragonfly/), an extendable, propritary GUI for multi-scale image analysis.

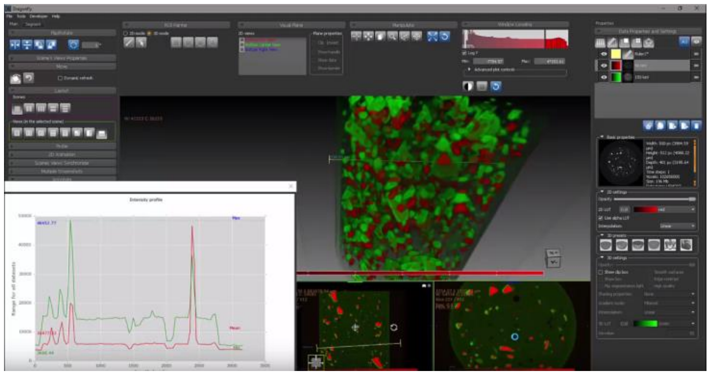

Figure 9 – Dragonfly features tomography data processing and visualization capabilities including Multi-channel Image handling (overlay and fusion in 2D/3D) (image credit: Mike Marsh, Object Research Systems, Inc)

Underlying the GUI there is a collection of c++ data structures, with the application logic written in python 3. The data can can be visualized and operated on both programattically and interactively in the GUI. As all operations in the GUI map directly on to operations on the underlying data structures it is possible to capture macros from user input for later use and for reproducibility. Because the application logic is written in python, it can be directly extended to make use of the existing scientific python eco-system. Out of the box Dragonfly provides tools for high-quality 3D renderings and natural 3D selections tools for manual and smart, machine learning based, segmentation. Multiple data sets can be brought into the workspace, aligned, and simultaneously segmented (***Figure 9***). ORS is working an developing an exchange where users can share custome plugins. Dragonfly is currently windows only, but a Linux beta is coming summer of 2017.

Daniela Ushizima from Lawrence Berkeley National Laboratory told us about a collection of tools to help scientist discover relevant, but hidden information in image data. This is both from extracting quantitative information from a specific class of samples and from aggregating across many different samples and experiments to find hidden similarities. Several examples of applying machine-learning to recognizing patterns across measurements of the same class of sample were presented. In one case it was possible to set up fast pipelines that, in near-realtime identify micro-fracture events from microCT measurements. On the macro scale, pyCIBIR, a tool for reverse image search was introduced. Building on the existing python eco-system pyCIBIR uses machine learning to query a database of images with an input image. By enabling scientist to find other measurements like the one they are currently considering which may illuminate scientific connection which would have otherwise been missed.

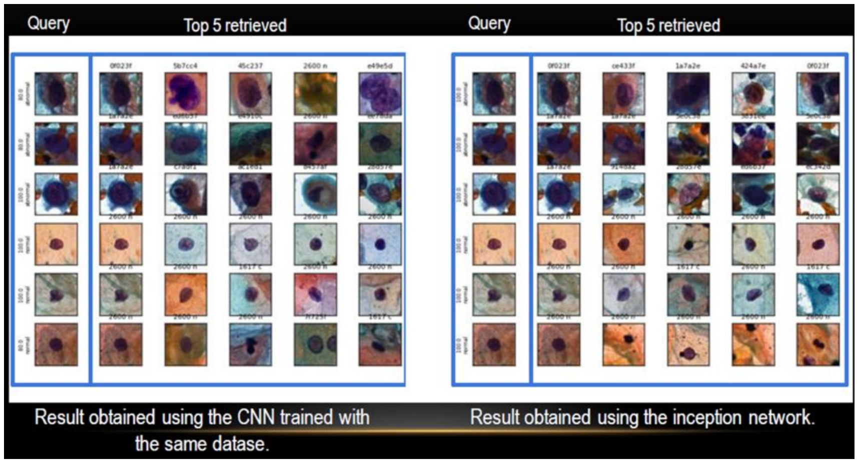

Figure 10 – An example shows pyCIBIR uses machine learning to query a cell image database of images with input cell images (image credit: Daniela Ushizima)

Discussion focused on the importance of users being able to explore their data, both at the small scale where all of the relevant data can be brought into one workspace and on macro scale to find related data sets. There was again discussion of the need for real data to feed machine learning methods.

The suggested follow-up action items are

- *Collaborate with third party packages (pyCIBIR, Paraview, Dragonfly and Tomviz), particularly via established / published documents from the NSLS-II databroker / bluesky so that the staff and user community at NSLS-II can profit from the suite of advanced data analysis and visualization tools readily available.*
- *Opportunities on 'training' machine learning algorithm with the big data from NSLS-II were recognized. For instance, we can improve the data search by 'feeding' pyCIBIR with data from the high-throughput NSLS-II beamlines.*

Final Discussions

In the final discussions, the software developers expressed a need to be able to access more data in order to develop the algorithms; particularly with the new machine learning type of algorithm, 'training the algorithm' is an important component and can only be done with accessing to the data. On the flip side, the beamline expressed the needs to gain help in analyzing the data. A mode to enable sharing the data between these two groups would be highly desirable. The discussion in the group went on to discuss about this donating data, which are already published or without significant scientific value, and also about having a data depository with data descriptor for the meta-data contents. Currently a TomoBank has been setup by scientists at the Advanced Photon Sciences, where data can be provided voluntarily; software developers can then download the data, with fully citable DOI numbers.

It was also recognized that a collaborative efforts on analyzing scientific data can be mutually beneficial. Cross-disciplinary collaboration can be established – it would be helpful if users can reach out to computer science communities, who would like to process 'real' data. It was recognized that staff at the synchrotron facility is in a good position to bride these two groups. It is important to address how to motivate data sharing, and how to capture future data collection meta-data, with a general purpose? It was mentioned that it would be important to provide the users with motivation to share their data to have their challenges solved. Open access journal and data, open sources software, the ability to have citation and recognitions of data and data providers would be beneficial.

Concluding Remarks

The motivation in pushing the development in this case can come from multiple directions: end users drive by having challenging science cases that require multi-modal, multi-scale and multi-dimensional approach. Computer scientists drive the development with their leading algorithms and computational architecture. On the instrument side, scientists at synchrotron beamlines also push technical capabilities that need better analysis and characterization tools.

From the presentation and discussions of the workshop, to tackle the challenges on multi-scale, multi-modal, and multi-dimensional, we need a "Hybrid Approach". This approach requires collaboration across academia, government laboratories and industry. In each of this unit, people with different roles need to work together, such as end-users, tool developers/engineers, and computer/math scientists. We have seen successful stories coming from individual projects led by single PIs, however, to make a broader impact, efforts need to be coordinated either facility-wide or even cross-facility.

The approaches to take in solving some of the technical challenges may also be 'hybrid'. For instance, data access might need to be both remote and local, depending on the use case. The data analysis tool development can come from commercially available software, open source software developed by service companies, and developed by the staff from the facilities or user groups. We have seen successful collaboration formed in an organized way or in an organic way. Keeping the contributions to the field diverse may be the key to continue the active development and research of the field. Overall, the multi-modal, multi-dimensional x-ray imaging analysis remains challenging, but with an exciting future.