Gap Analysis:

Materials Discovery Through Data Science at Advanced User Light Sources

Workshop Report

Los Alamos
NATIONAL LABORATORY
EST.1943

October 3-5, 2018
Santa Fe, New Mexico

Gap Analysis: Materials Discovery through Data Science at Advanced User Light Sources

Workshop, October 3-5, 2018

Hotel Santa Fe | Santa Fe, New Mexico

Role	**Name**	**Organization**
Chair:	Christine Sweeney	Los Alamos National Laboratory
Co-Chairs:	Richard Sandberg	Los Alamos National Laboratory
	Aric Hagberg	Los Alamos National Laboratory
Steering Committee:	John Sarrao	Los Alamos National Laboratory
	James Ahrens	Los Alamos National Laboratory
	John Post	Lawrence Livermore National Laboratory
	Daniela Ushizima	Lawrence Berkeley National Laboratory
	Kerstin Kleese Van Dam	Brookhaven National Laboratory
	Ian Foster	Argonne National Laboratory
	Amedeo Perazzo	SLAC National Accelerator Laboratory
	Tony Rollett	Carnegie-Mellon University
Keynote Speaker:	Mike Dunne	Linac Coherent Light Source, SLAC National Accelerator Laboratory
Plenary Speakers:	Simon Billinge	Columbia University and Brookhaven National Laboratory
	Kevin Jorissen	Amazon Web Services
	Derek Bingham	Simon Fraser University
	Jeffrey Donatelli	Lawrence Berkeley National Laboratory
	Matthew Cherukara	Argonne National Laboratory
	Jana Thayer	SLAC National Accelerator Laboratory
Break-out Session Speakers:	Claudio Mazzoli	Brookhaven National Laboratory
	Diego Casa	Argonne National Laboratory
	Ian Robinson	Brookhaven National Laboratory
	Arianna Gleason	SLAC National Accelerator Laboratory
	Zsolt Jenei	Lawrence Livermore National Laboratory
	Reeju Pokharel	Los Alamos National Laboratory
	Ryan Coffee	SLAC National Accelerator Laboratory
	Turab Lookman	Los Alamos National Laboratory
	James Ahrens	Los Alamos National Laboratory
	Devin Francom	Los Alamos National Laboratory

	Stephan Hruszkewycz	Argonne National Laboratory
	Daniel Allan	Brookhaven National Laboratory
	Logan Ward	Argonne National Laboratory
	Daniela Ushizima	Lawrence Berkeley National Laboratory
Discussion Leads:	Kerstin Kleese Van Dam	Brookhaven National Laboratory
	Richard Sandberg	Los Alamos National Laboratory
	Garth Williams	Brookhaven National Laboratory
	Alex Scheinker	Los Alamos National Laboratory
	Thomas Proffen	Oak Ridge National Laboratory/Spallation Neutron Source
	Peer-Timo Bremer	Lawrence Livermore National Laboratory
	Earl Lawrence	Los Alamos National Laboratory
	Jeffrey Donatelli	Lawrence Livermore National Laboratory
	Amedeo Perazzo	SLAC National Accelerator Laboratory
	Bryce Meredig	Citrine Informatics
	Aric Hagberg	Los Alamos National Laboratory
Strategic Coordinator:	Lucy Maestas	Los Alamos National Laboratory
Meeting Planner:	Peggy Vigil	Los Alamos National Laboratory
Website:	Sarah Haag	Los Alamos National Laboratory
Communications:	Don Montoya	Los Alamos National Laboratory
	Amy Elder	Los Alamos National Laboratory
	Carlos Trujillo	Los Alamos National Laboratory
	Brye Steeves	Los Alamos National Laboratory
	Mike Nudelman	Los Alamos National Laboratory

Role	**Organization**
Workshop Sponsor:	Los Alamos National Laboratory
Workshop Venue:	Hotel Santa Fe, Santa Fe, New Mexico
Water Bottles:	New Mexico Consortium, Los Alamos, New Mexico
Report Online:	www.lanl.gov/2018gapanalysis

Table of Contents

Executive Summary

As revolutionary increases in coherent x-ray flux come online via upgraded technologies at both synchrotrons and x-ray free electron laser facilities, materials scientists are facing radically increased data volumes, much higher data velocity, and data from an ever-increasing number of sources and at larger scales. The challenge is to leverage these data for scientific discovery; therefore, data science is an area of increasing interest in the domain of x-ray light source user facilities. Many recognize that there are gaps between data science, materials, and light sources. Furthermore, many agree that there is a need as a community to clearly identify the gaps and prioritize ideas on how best to bridge them.

Los Alamos National Laboratory (LANL) hosted a workshop entitled: “Gap Analysis: Materials Discovery through Data Science at Advanced User Light Sources” held on October 3-5, 2018, in Santa Fe, New Mexico. The workshop brought together more than 60 invited experts in data science, materials science, and light source experiments from academia, industry organizations, and national laboratories. All of the Department of Energy Office of Science x-ray light source laboratories were represented.

The workshop began with a motivational keynote address from Prof. Mike Dunne of Stanford University and Director of the Linac Coherent Light Source at SLAC National Accelerator Laboratory. His talk focused on data science needs as light sources are upgraded, especially during the upcoming LCLS-II upgrade.

Plenary talks focused on data types, experimental design, mathematics, modeling, and simulation. Breakout sessions covered six experimental techniques: x-ray photon correlation spectroscopy, resonant inelastic x-ray scattering, Bragg coherent diffraction imaging, dynamic x-ray diffraction at high pressures, high pressure diamond anvil cell x-ray diffraction, and high energy diffraction microscopy. Eight additional breakout sessions covered data science techniques: machine learning, modeling and simulation, visual analytics, statistics and emulation, physics and math in data analysis, data management for data science, materials databases, and data mining.

Priority gaps are identified in the following areas:

Gap 1: Tools for exploiting multidimensional and multimodal data. Analysis of data from multiple data types, other experiments, or simulations is often not performed because it can be complex and require theoretical simulation tools and increased computational resources.

Gap 2: Data, algorithm, and software curation. Often little support is provided for data curation in addition to open source, debugged, and efficient software for relevant algorithms. Materials databases are not easy to effectively utilize for data sharing and reuse.

Gap 3: Real-time decision-making analytics and tools. Real-time analytic and decision-making tools have lagged behind advances in experimental facilities and the surge of data volume and velocity. Essential tasks, such as anomaly detection and error reduction, are slipping through the cracks. Those users without access to data science lag behind those with access.

Gap 4: Limited technology literacy of available tools. Programming languages and interfaces that make it easier for scientists to create software are still needed. Software infrastructure provided by user facilities needs to be user-friendly as well as production-quality.

Gap 5: Data reduction/extraction planning and tools. Scientists need tools for compression, streaming data formats, data veto, and more. The community is aware that all data cannot be saved and they must address the gaps in data reduction and extraction planning and tools.

Gap 6: Experimental design aligned to research questions. Analytical software stacks that include simulation, machine learning, and data science are not widely available, yet would provide systematic, principled, and automated support for pre- and in-experiment design. Advanced and semi-automated experimental design are needed to quickly re-plan experiments.

Priority research opportunities (PROs) are identified in the following areas:

PRO 1: Real-time decision-making tools that provide data-informed decisions. Scientists need tools that provide high-level feedback and choices during an experiment. They want recommendations with levels of uncertainty and a rationale, but also be able to drill down.

PRO 2: Codesign of experiments with data analysis, end-to-end simulation and planning. This codesign includes research in statistical emulation and experimental design, beamline simulation for feedback control, and end-to-end experiment simulation and planning.

PRO 3: High-dimensional data visualization and interaction. Research is needed in reconstructing images or structure from vast amounts of data, incomplete data, or noisy data. Capturing the time dimension is important, too. Approximate solutions are also needed.

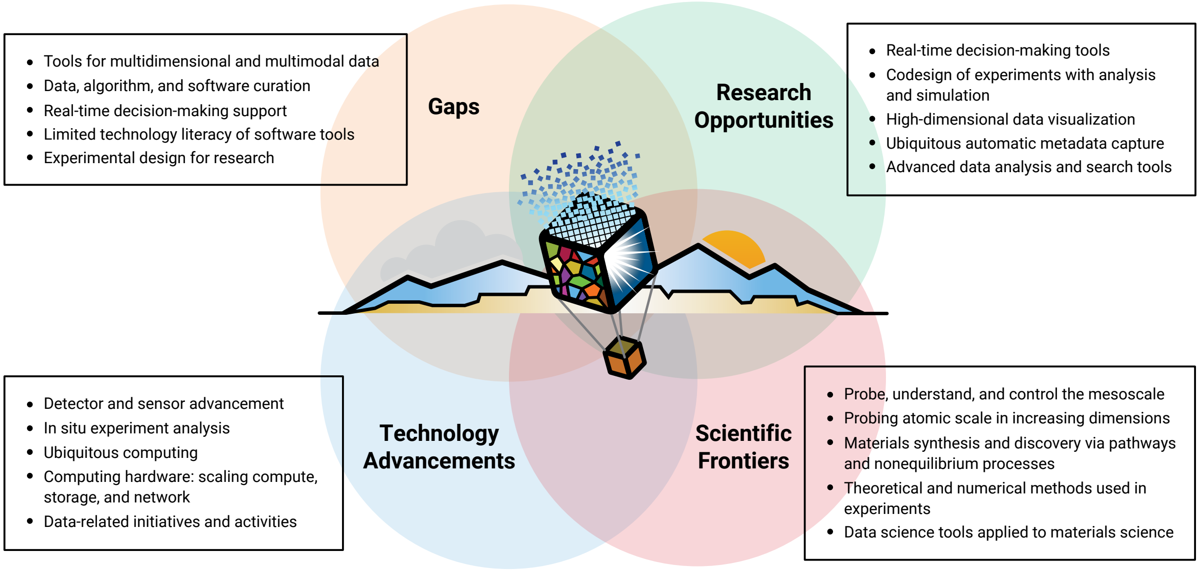
PRO 4: Ubiquitous, automatic metadata capture. Metadata (data that describe other data) support database design and access as well as data science, such as machine learning. Nonintrusive metadata capture requires research in understanding metadata requirements.

PRO 5: Advanced data analysis (multimodal, multiscale) enabled by searchable, open source tools. Experiment data analysis can benefit from different data types used together, previous experiment data, data from related experiments, experiments at different facilities, simulation data, emulated data, and data at different time and length scales.

The following transformational opportunities are identified for the materials science endeavor that are presented by light source experiment facility advances: 1) increasing call to probe, understand, and control the mesoscale, 2) probing the atomic scale in ever-increasing dimensions (three spatial dimensions, time, spectral/energy, and others), 3) materials synthesis and discovery via pathways and nonequilibrium processes, 4) theoretical or numerical models used during experiments, and 5) data science tools applied to materials science.

Five revolutionary technology advancements are identified: 1) detector and sensor advancement, 2) in situ experiment analysis, 3) ubiquitous computing, 4) computing hardware, and 5) data-related initiatives and activities.

Follow-on action items are identified and include outreach, education, funding opportunities, and the development of community resources. The contributions and level of engagement by workshop attendees were remarkable, which is encouraging and bodes well for future activity in this area. The issues raised in this workshop are important to being at the technological edge in an ever-more competitive national and international environment, and important to accelerating scientific discovery that supports academia, industry organizations, and the national mission of the laboratories represented at the workshop.



- Tools for multidimensional and multimodal data
- Data, algorithm, and software curation
- Real-time decision-making support
- Limited technology literacy of software tools
- Experimental design for research

- Real-time decision-making tools
- Codesign of experiments with analysis and simulation
- High-dimensional data visualization
- Ubiquitous automatic metadata capture
- Advanced data analysis and search tools

- Detector and sensor advancement
- In situ experiment analysis
- Ubiquitous computing
- Computing hardware: scaling compute, storage, and network
- Data-related initiatives and activities

- Probe, understand, and control the mesoscale
- Probing atomic scale in increasing dimensions
- Materials synthesis and discovery via pathways and nonequilibrium processes
- Theoretical and numerical methods used in experiments
- Data science tools applied to materials science

Summary graphic for the Gap Analysis workshop.

1 Workshop Objectives

This workshop was largely motivated by the overwhelming challenges posed by recent and future upgrades to x-ray light source user facilities. X-ray light source facilities are experiencing unprecedented increases in available x-ray flux. These upgrades include diffraction-limited storage rings (DLSR) at synchrotrons and x-ray free electron lasers (see Figure 1) [1, 2]. A new generation of synchrotron storage rings are emerging around the world beginning with the MAX IV in Sweden [1] and the upgrade to the European Synchrotron Research Facility (ESRF) in France that are enabled by multi-bend achromat lattices. These DLSR increase average coherent x-ray flux by upwards of 100 times compared to traditional synchrotrons. Additionally, x-ray free electron lasers, beginning with the first hard x-ray free electron laser (XFEL) at the Linac Coherent Light Source (LCLS) at SLAC National Accelerator Laboratory, have increased peak x-ray brightness by approximately 10 billion times compared to past synchrotron sources. These new brilliant XFELs have been rapidly adopted across the world.

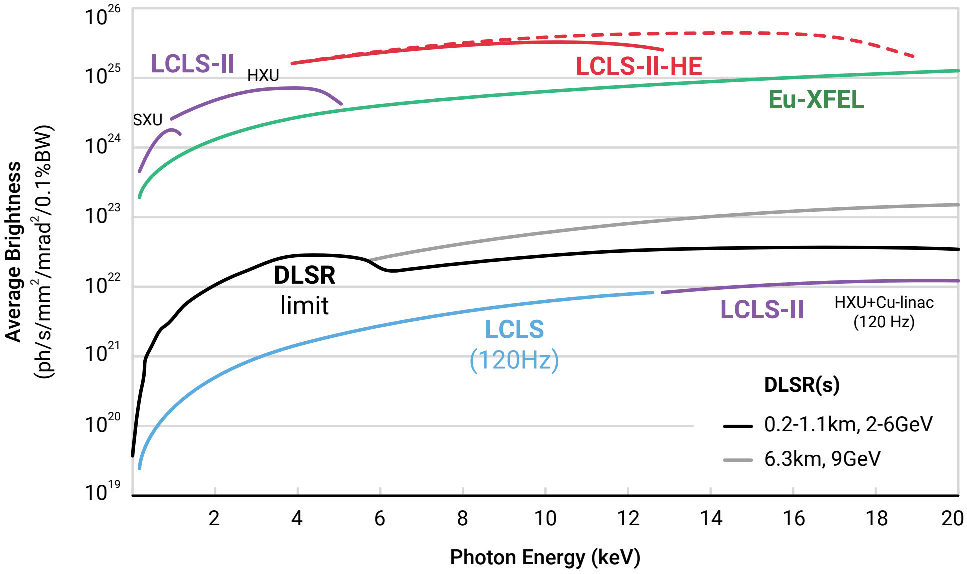


Figure 1. Plot showing current average brightness of current XFELs, DLSR upgrades to synchrotrons, and planned LCLS-II upgrade (from Mike Dunne's keynote presentation).

Although hailed for the scientific discovery, these upgrades have enabled [3], and will continue to enable, considerable research. Advanced supporting technology is necessary to fully utilize the upgraded facilities and the vast amounts of data that will be produced by them. Accelerator technology is far outpacing our ability to use and detect the photons produced and process and store the data produced at these facilities. Experiment output data rates at facilities like the upgraded LCLS-II are projected to reach terabyte-per-second data rates (Figure 2). Data extraction is particularly challenging at these rates.

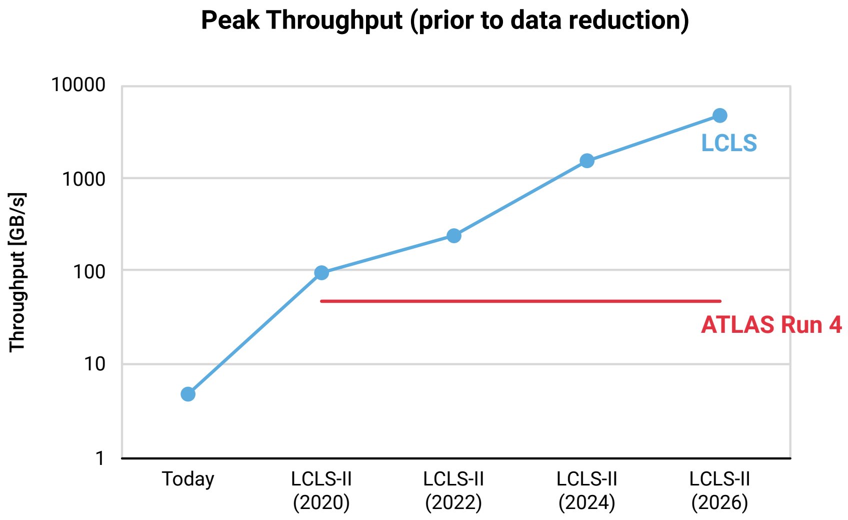


Figure 2. LCLS data rates as new, higher repetition rate upgrades come online. (Courtesy of Amedeo Perazzo).

Take, for example, the challenge of high data rates and volumes. One can illustrate this using the Fisherman's Analogy. The fisherman gets a new net that is better than anything he had before. He catches so much sea life in his net during a drought that he cannot keep them all in the boat or it will sink. He has three options: reject whatever is not a fish (veto), just save the parts of the fish he cares about (feature extraction), or pack them into the boat in the most efficient manner (compression). This analogy summarizes current thought in data reduction and utilization: throw out perceived useless data, extract only useful data, or efficiently save all data via compression. The first two choices risk discarding important science that may not be known until later or requires significant processing to extract from noise. The last choice is rapidly getting beyond our storage capabilities.

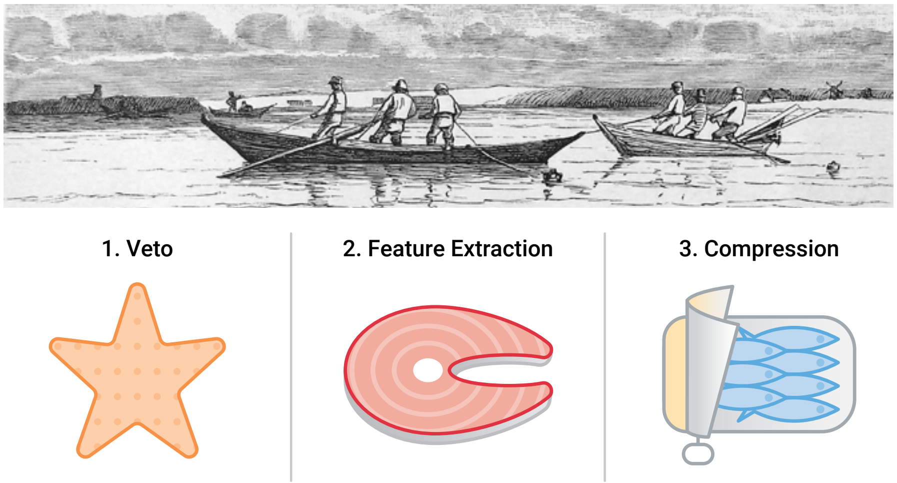


Figure 3. Analogy of fisherman with a new net and how different data reduction and extraction strategies can be implemented.

Another challenge is the sheer volume and variety of experiments that are performed at these light source facilities. Experimental techniques are multiplying as users develop new spectroscopic, diffraction, and imaging methods that capitalize on the increased x-ray coherent flux. These developments are presenting a new challenging paradigm for Big Data science, as compared to the traditional large high energy physics user facilities such as the European Organization for Nuclear Research, known as CERN. As Prof. Simon Billinge of Columbia University and Brookhaven National Laboratory said at the workshop, “We are not 200 scientists working on one experiment, but 200 scientists working on 2,000 experiments.” Therefore, solutions to challenges faced by scientists in this domain, broadly defined as materials science, are usually not one-size-fits-all.

At the same time as light source science is growing and encountering new challenges, computing is also evolving with large-scale and heterogeneous hardware resources as well as data science (Figure 4) that can be leveraged to help meet light source data challenges.

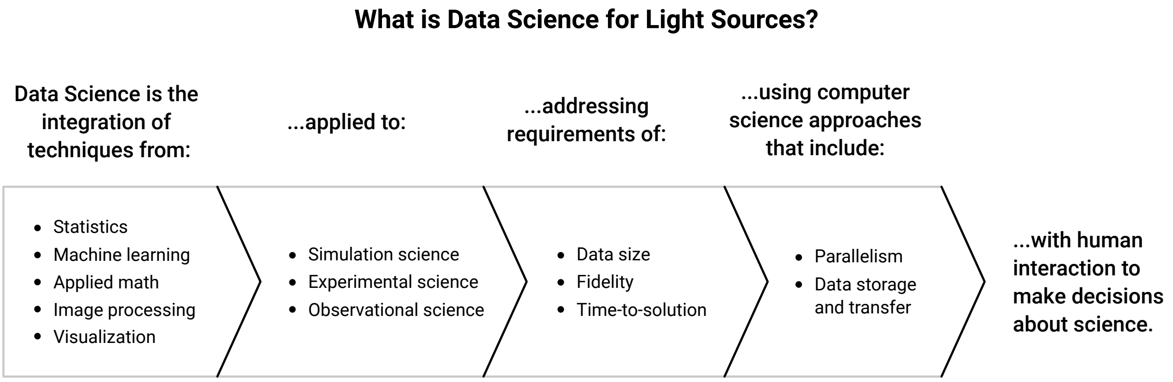

What is Data Science for Light Sources?
Data Science is the integration of techniques from:
Statistics
Machine learning
Applied math
Image processing
Visualization
...applied to:
Simulation science
Experimental science
Observational science
...addressing requirements of:
Data size
Fidelity
Time-to-solution
...using computer science approaches that include:
Parallelism
Data storage and transfer
...with human interaction to make decisions about science.

Figure 4. Data science definition by James Ahrens.

With these light source and computing challenges and also the opportunities in mind, the workshop aimed to look for consistent themes among various light source experimental subcommunities doing research in materials science, such that cross-cutting gaps and priority research areas utilizing data science could be discerned. Talks and discussions focused on six representative experiment types on the first day, 12 data science techniques on the second day, and synthesizing discussions on the last day (see agenda in Appendix D). Keynote and plenary talks brought the discussion focus on key cross-cutting topics, such as cloud, data reduction, data types, experimental design, mathematics, modeling, and simulation. Discussions also touched on general themes such as, “Are we conducting experiments or taking measurements? If it is an experiment, are there different requirements than if we were taking measurements?” The data science topic names were often qualified by when the data science is done (e.g., “pre-experiment machine learning,” or “in-experiment statistics”), so as to make data science solutions more specific and ensure the entire experiment workflow was covered by the workshop.

2 Materials Science Endeavor

Using light source experiment facilities truly presents transformational opportunities for the materials science endeavor. With the newly available x-ray flux, advanced detectors, and new data science tools, materials scientists and engineers are pushing to ever smaller and faster time scales to tackle problems as diverse as materials failure to mimicking photosynthesis in bio-inspired solar cells [4–7]. Additionally, moving beyond an observational materials science capability to a predictive capability where we can design the materials of the future is highlighted in numerous reports [6–13]. Four transformational opportunities for materials science were identified by the workshop.

I. Increasing call to probe, understand, and control the mesoscale. We can define the mesoscale as the gap between the atomic scale where molecular dynamics codes most effectively capture materials behavior and the bulk scale where continuum models dominate [9–10]. A critical aspect of developing a predictive capability for materials science is understanding the role of heterogeneity at the mesoscale frontier via such materials attributes as grain boundaries, defects, impurities, alloys, and materials boundaries. Advanced accelerator-based x-ray light sources stand ready to probe this gap and give materials scientists the tools necessary to develop the fundamental understanding of heterogeneities as well as the mesoscale in order to develop necessary predictive capability [1–3, 6–7].

II. Probing the atomic scale in ever-increasing dimensions (three spatial dimensions, time, spectral/energy, etc.). More often, facility proposals are pursuing increasingly difficult experiments with themes such as rare events, phase transformation kinetics, and chemical pathways [6–7]. As one example, scientists are proposing rapidly probing large volumes (upwards of millimeters cubed) at the atomic scale to look for phase transformation nucleation of materials under extreme physical or chemical environments with x-ray photon correlation spectroscopy or coherent imaging techniques [6–7]. Both of these techniques are data hungry (reading out megapixel arrays of large bit pixels at kilohertz rates) and computationally intensive (multiple Fourier transforms for iterative phase retrieval algorithms). For starters, experimenters must comb through potentially kilohertz to megahertz frame rates of these megapixel arrays, at times on non-reproducible experiments where small details need to be extracted from multimodal data sets (diffraction, imaging, spectroscopic).

III. Materials synthesis and discovery via pathways and nonequilibrium processes. Materials synthesis and discovery are dependent on initial conditions and pathways. As such, in situ and time-resolved results are essential elements. Lacking are the visualization, decision-making, and feature extraction tools necessary to prevent lost opportunities at experiments as illustrated in Figure 5.

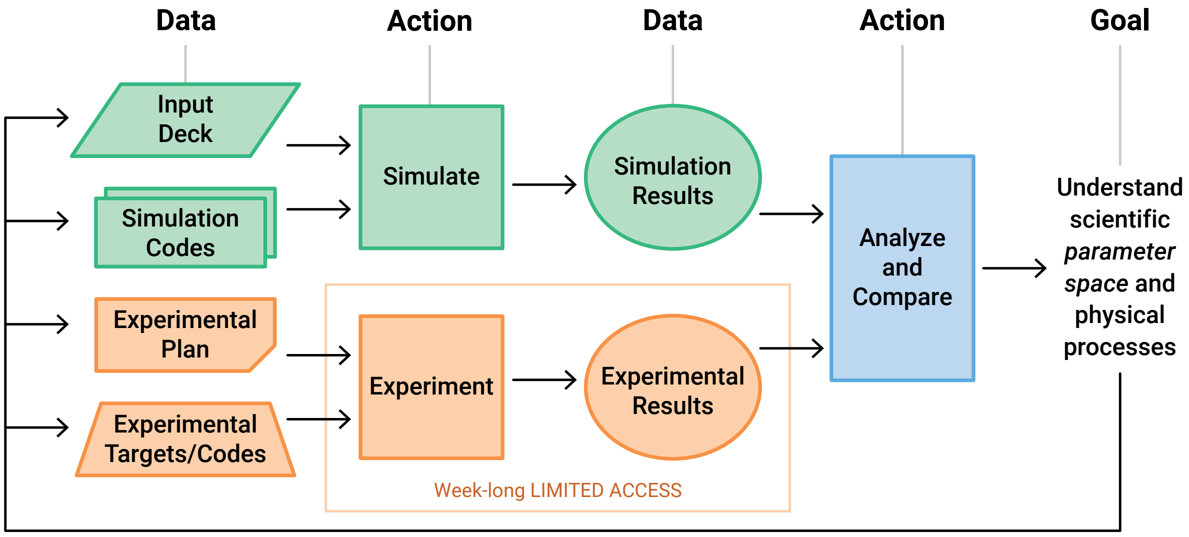
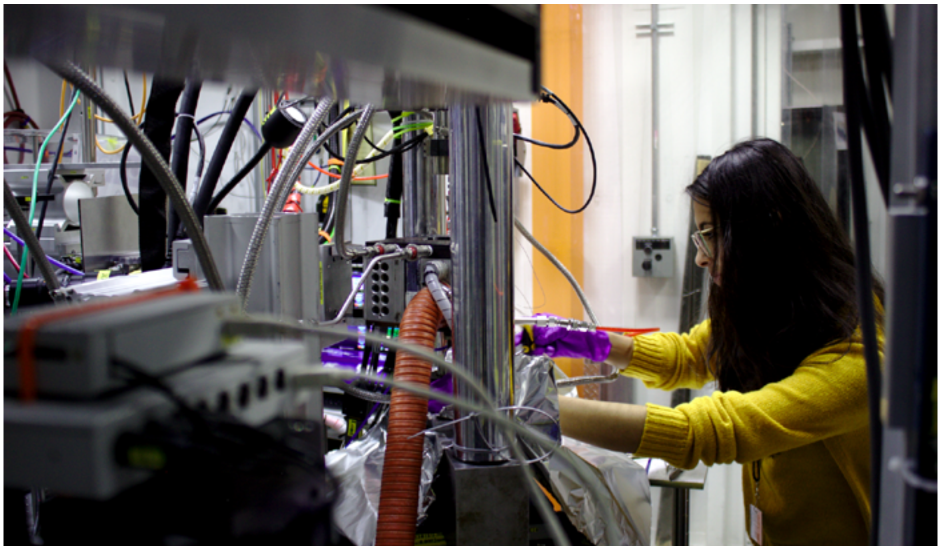


Figure 5. Schematic of the idealized discovery science process at x-ray light sources. Key elements are motivated by and compared to simulations; theory is used to motivate the hypotheses, experimental planning, simulation and experimental results, then real time comparison, data reduction, synthesis, and inference that lead to feedback and iteration upon the simulations and experiments. Ideally, these processes can be conducted in real time at the experiment (Credit: Christine Sweeney, Cindy Bolme, James Ahrens, LANL ASSIST LDRD-DR, 2016).

IV. Theoretical or numerical models used during experiments. Models could be used to guide experiment hypotheses and proposals prior to experiments, but all too often scientists cannot query these models in real time during the experiment. These lost opportunities can take the form of false positives (taking the wrong data believing they are correct), missing the key features of the data (i.e., not getting in the correct parameter regime), and lacking the data needed to make decisions (i.e., wasting time). These gaps are identified in the paragraph below. Developing the ability to query these models in real time would represent a transformational opportunity—especially if these comparisons can guide experiments in real time.

What is desperately needed at advanced x-ray user facilities are the tools increasingly available in the world of data science: visualization, data reduction/synthesis, real-time data reduction tools, real-time decision-making tools, and real-time tools for querying simulations/ emulations. Working smarter by using advanced data science tools is absolutely required because just working harder cannot overcome the challenges imposed by trying to process ever-increasing amounts of data. What is critical is that we take the right data.



Reeju Pokharel (LANL) mounting a sample for diffraction measurements during in-situ heating and subsequent loading of additively manufactured steel at 1-ID beam line at the advanced photon source. Photo courtesy Tom Stockman.

3 Revolutionary Technology Developments

In addition to the revolution in x-ray light sources, the Gap Analysis Workshop was motivated by a number of revolutionary technology developments that are game changers for experimental science. These developments are key to future scientific discoveries; however, they additionally pose challenges to the community in terms of adoption and were addressed by the workshop. Five technology developments were identified by the workshop:

I. X-ray detector advancement. Sophisticated detectors are emerging with increased spatial, temporal, and spectral resolution while aiming to keep up with frame rates in the kilohertz regime. These new detector characteristics will provide better resolution than in the past and enable experiments to keep up with increased beam flux, but will result in vastly increased data volumes (see Figure 6 with new gigahertz rate framing cameras [14]). Detector capability and innovation will be needed for cutting-edge materials science research. Some examples are gated detectors that can handle increased frame rates, detectors with contiguous angular and azimuthal coverage, detectors that assist with texture analysis, detectors with increased dynamic range, and detectors with smaller pixel size.

II. In situ experiment analysis. There is a paradigm shift in the use of simulation to do in situ interrogation of experiments instead of ex situ (post-experiment). Using simulations (or synthetic data) allows us to query the data in real-time to help with the next set of decisions. This mode of inquiry is possible due to computationally advanced high-performance computing (HPC) platforms that can execute ensembles of simulations (many runs with different parameters) that provide ample coverage of the search space of possible answers.

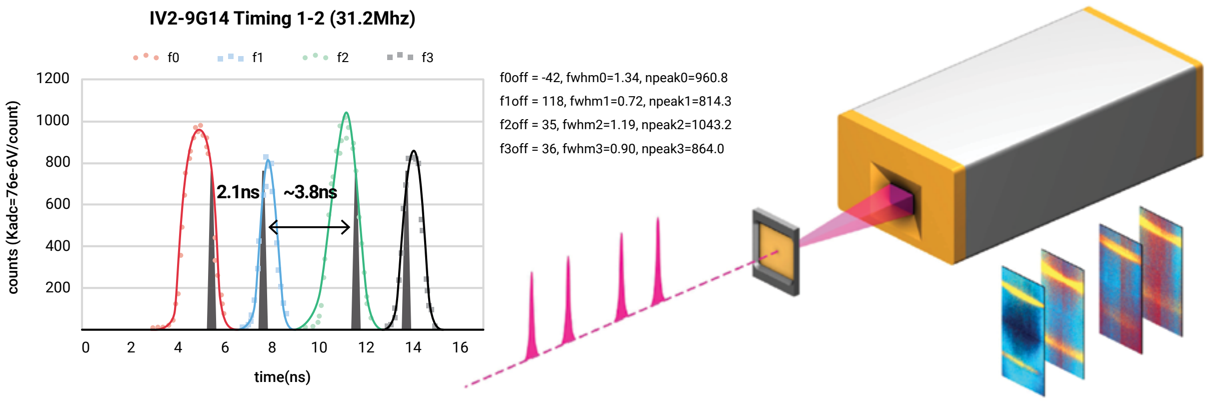


Figure 6: Advances in detector technology are enabling higher data rates and data volumes. For example, the new Icarus detectors from SLAC can enable multiple frame readouts of x-ray pulses separated by a few nanoseconds. As detectors approach the ability to read out at MHz rates continuously, data volumes will approach multiple TB/s [14].

III. Ubiquitous computing. The phrase "ubiquitous computing" refers to increasingly available computing resources—on any device, in any location, and in any format. Ubiquitous computing is coming to the light source experiment world as well. This leads to the emerging possibility of "ubiquitous provenance"—access to provenance (information about what was computed and how) available on any device, in any location, and in any format. Ubiquitous provenance makes it possible to record much more information on experimental settings, conditions, and data processing choices. Provenance will capture everything that has been done to the data from cradle to grave to allow users to track all relevant metadata and processes performed on their data. It opens up many more possibilities on what can be done with the resulting data as well.

IV. Computing hardware. Memory technology is making it easier to store more information and do more computations in situ. Computational accelerators are speeding up calculations. Field programmable gate arrays enable streaming data through custom hardware configured to do common mathematical functions. Custom machine-learning hardware is also appearing. Exascale computing [15] and interfacility workflows are bringing the "big iron" (leadership class supercomputers) to experimental science. They are enabling the coupling of advanced analytics and simulations to live experiments. Supercomputers can be used to generate massive data sets for training of machine-learning models that can be used to do experimental analysis.

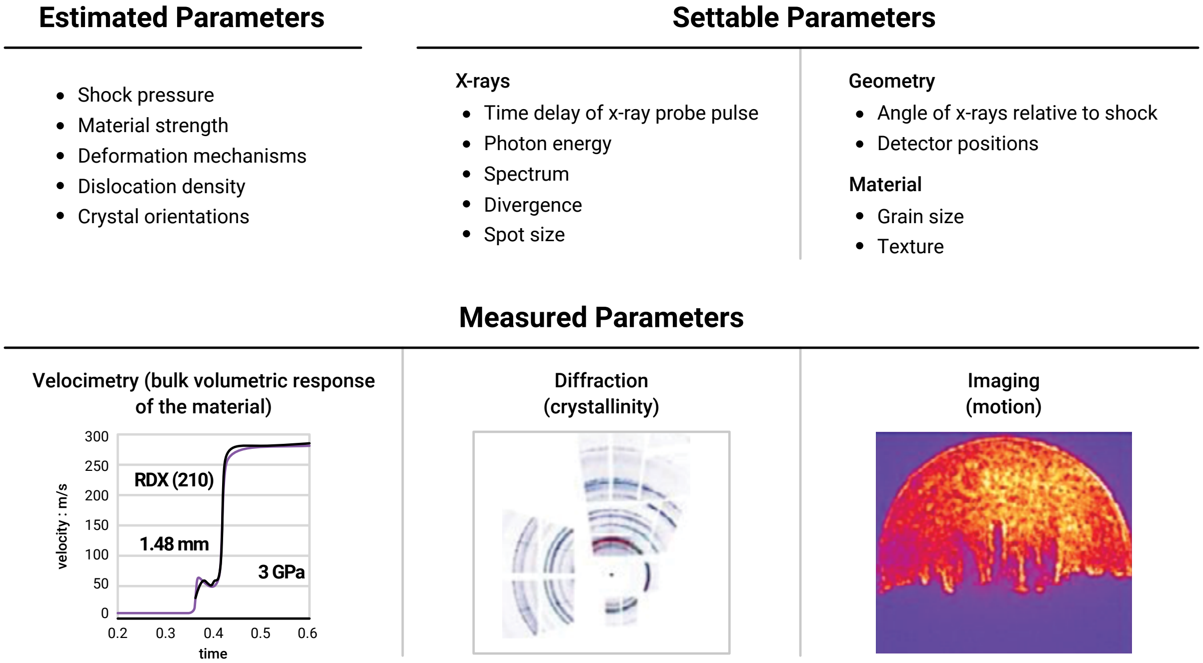
V. Data-related initiatives and activities. A recent United States presidential executive order [16] has put artificial intelligence, machine learning, and data curation on the national agenda. Big data in the arena of massive online data has sparked commercial innovation in cloud computing and cloud-based data science. It is projected that by 2025, 49 percent of the world's stored data will reside in public cloud environments [17]. Companies like Amazon Web Services are coming up with innovative ways [18] to transport large amounts of data to the cloud. It could be possible that in the future, cloud computing could become a more affordable option for scientific purposes.

4 Priority Gaps

Now that we have reviewed the motivating background and rapid advances in x-ray light sources, data volumes, materials science, and enabling technology, we will review the gaps and priority research opportunities identified at the workshop. Workshop attendees converged on six gaps in data science for materials science at light sources. The following subsections describe each gap, why it is important to close this gap, and any barriers and challenges to doing so.

Gap 1: Tools for Exploiting High-Dimensional and Multimodal Data

One of the biggest gaps identified in the workshop was inadequate experimental and analytics tools for exploiting high-dimensional, multimodal data. Experimental scientists make decisions based on many forms of data. For example, for a dynamic compression-related experiment, the materials scientist needs to know the initial state of the material, the phase diagram for the material, and the deformation and transformation mechanism. Data that provides this information may come from diffraction patterns, x-ray images, a variety of visible or x-ray spectroscopies, velocimetry measurements (velocity of the material during a shock), historical phase diagram information, and even synthetic data resulting from simulations. One example from dynamic compression experiments from the LCLS is shown in Figure 7.



Estimated Parameters

- Shock pressure
- Material strength
- Deformation mechanisms
- Dislocation density
- Crystal orientations

Settable Parameters

X-rays

- Time delay of x-ray probe pulse
- Photon energy
- Spectrum
- Divergence
- Spot size

Geometry

- Angle of x-rays relative to shock
- Detector positions

Material

- Grain size
- Texture

Measured Parameters

Velocimetry (bulk volumetric response of the material)

Diffraction (crystallinity)

Imaging (motion)

Figure 7. Example of multidimensional and multimodal data from dynamic compression experiments at the LCLS. (Courtesy of Cindy Bolme, LANL ASSIST LDRD-DR).

Not only do experimentalists use multiple data types from light source experiments, they also want to use data from other kinds of experiments, such as spectroscopy. X-ray diffraction (for atomic structure) and spectroscopy (for electronic properties) using current and next-generation x-ray sources will be key to obtaining synchronized measurement of atomic structure and electronic properties to reveal basic structure–function relationships [13].

Having a specialized tool for each kind of data may be effective in isolation, but in order to view and reason about multiple data types from different sources and with different fidelity, better analytic tools are needed. Analysis of high-dimensional and multimodal data can be quite complex and often requires the involvement of theoretical simulation tools, mathematicians and computational scientists, and increased computational resources. This problem is exacerbated by the variety of experiments being performed and data types possible for experimental data. Some commonalities exist in data types generated by standard sensors and detectors at user facilities; however, users also bring their own equipment, such as custom stages, motors, attenuators, and others, that may change how experiments are conducted and how data coming from standard sensors are interpreted.

Gap 2: Data, Algorithm, and Software Curation

Experimental scientists are faced not just with creating experimental data, which is challenging on its own, but curating the data (labeling, storing, searching for, and accessing their data). For many facilities, little support is provided for data curation. At some facilities, the user is responsible for taking home his or her data on his or her own drive. Additionally, users need to develop and manage code for processing and analyzing the data. For many experiments, open source, debugged, and efficient software for relevant algorithms is not available, leading to unmanaged ad hoc scripts that get modified per experiment run. The lack of curated data and code slows down science because it is harder to build on previous efforts. This can create duplicated efforts by a number of scientists.

Fully curated data are also dependent on provenance (information about how the data were obtained). Lack of provenance affects reproducibility of experimental results. Lack of reproducibility degrades the quality, integrity, and the pace of science. This provenance light or provenance weak approach is becoming less acceptable as funding agencies require data management plans and publishers require authors to submit information that allows others to obtain data and reproduce results. Accessing, aggregating, and comparing data without consistent curation is challenging, if not impossible.

Data curation is supported by materials databases. Just as protein databanks have revolutionized biological sciences, materials databases have the potential for revolutionizing materials science [19–20]. Materials databases are starting to make headway and have great potential for accelerating materials design and education by providing new data and software tools to the research community [19]. Not only can they store data, algorithms, and software, but they can also store both experimental and simulation data, thus leveraging both. In addition, they enable a type of crowd-sourced collaborative science, which is a new direction.

Although materials databases could be a boon to experimental science, gaps in this area still exist. Some gaps in current materials databases are the following: consistent and usable application programming interfaces (APIs), interoperability between different materials databases, insufficient amounts of new experimental data being added to materials databases, insufficient metadata to properly associate with the data, and lack of investment in long-term materials database maintenance and staff. Filling these gaps could help realize the potential of materials databases.

Gap 3: Missed Opportunities During Experiment: Real-Time Decision-Making Analytics and Tools

The lack of automation and of real-time decision-making analytics and tools has led to many missed opportunities for capitalizing on real-time experiment steering. Previously, experiments took longer and it was standard practice for scientists to do many tasks manually. Even with overnight shifts, scientists could not keep up. With upgraded experimental facilities and supporting computing systems, the duration and cadence of experiments is significantly accelerated. At the same time, the volume and velocity of data have surged. Support for higher level analytic and decision-making tools has lagged. The cognitive load to be able to handle the experiment pace with manual processing has become untenable, especially at XFELs. Essential tasks such as anomaly detection and error reduction are slipping through the cracks. Challenges in closing this gap include providing sufficient automation at all parts of the workflow so that there are no remaining bottlenecks, developing decision-making tools that help with many kinds of experiments, not just one in particular, and making this workflow production-quality so that it can be relied upon, trusted, and easy to use. There is a growing gap between user groups who have access to teams of data science collaborators and their tools and those who do not. There is a concern that smaller institutions with limited resources will fall behind and be excluded. Furthermore, light source research on materials involves systematically varying experimental conditions to elicit a material response and your chances of doing this are much greater if you can vary parameters correctly in a single beam time. It is much harder and often impossible to recreate conditions after an experiment shift is over.

Gap 4: Software Literacy (Both Languages and Facility Infrastructure): An Educational Gap

Light source accelerators have long been an engineering challenge and advancements in their technology have been an incredible achievement. As we move into an age where accelerators are more productive than we could have ever imagined decades ago, software needs for users are coming to the fore. Capability in designing, assembling, maintaining, and operating accelerators has been developed; however, a corresponding capability in computation that maximizes the usage of these accelerators is less developed, enough so that it appears to be a gap. Software technologies, such as machine learning, visualization, and other data science techniques, are progressing rapidly, which adds to the difficulty scientists have in becoming more software literate. With the wide variety of experiments and science domains, software needs are diverse, which also make software literacy more challenging.

Scientists from earlier days in light source experimentation are picking up software skills as they can. No longer can an experimental scientist obtain an undergraduate or graduate degree in physics or chemistry without having participated in some sort of software engineering effort. Many take classes in computer science as well. User facilities, although often not as equipped locally with the scale of computing as some national laboratories, are becoming more adept in computing; however, much more progress in this area could be imagined. Programming languages and interfaces that make it easier for scientists to create software are still needed. Software infrastructure and utilities provided by user facilities have been improved, but require more support to be user-friendly and production-quality.

This gap in software literacy, although quite visible to materials experimental scientists, seems to be one that has not received the attention it could at a programmatic level in the form of targeted initiatives that result in sufficiently increased computing software and systems education, hackathons, webinars, and user-friendly documentation. The biological community has a longer history of advanced computation engagement than the materials community, so it seems to be somewhat ahead in this respect. Perhaps this is due to an earlier need for computation in connection with genetics and macromolecular protein crystallography. This has accelerated with high data volumes and rates in experimental serial femtosecond crystallography. We are now seeing that these skills are equally important for materials research as data volumes and rates in this experimental regime reach levels previously encountered by the biological community.

Gap 5: Planning for the Data Tsunami: Data Reduction and Extraction Planning and Tools

The data tsunami, while affecting certain light source facilities (like XFELs) more than others in terms of unprecedented absolute data volumes and velocities, also applies to many other light sources and experiments. This is because “big data” really means data that is much bigger than what has been encountered previously. So even for many small data experiments, when the data are one to two orders of magnitude larger or produced faster than what is typically processed and stored, it presents major challenges for analysis and obtaining the desired results during an experiment. Areas of data growth in light source experiments for materials science include dynamic compression, with a 106 increase in shot rates over the past 20 years. Now, at facilities such as the European XFEL (Eu-XFEL) 2,700 individual x-ray pulses can be delivered separated by 200 nanoseconds arriving in bunches at 10 hertz. Growing data rates are also on the horizon for resonant inelastic x-ray scattering (RIXS), where about 300 kilohertz multichannel detector readout is a possibility in the near future.

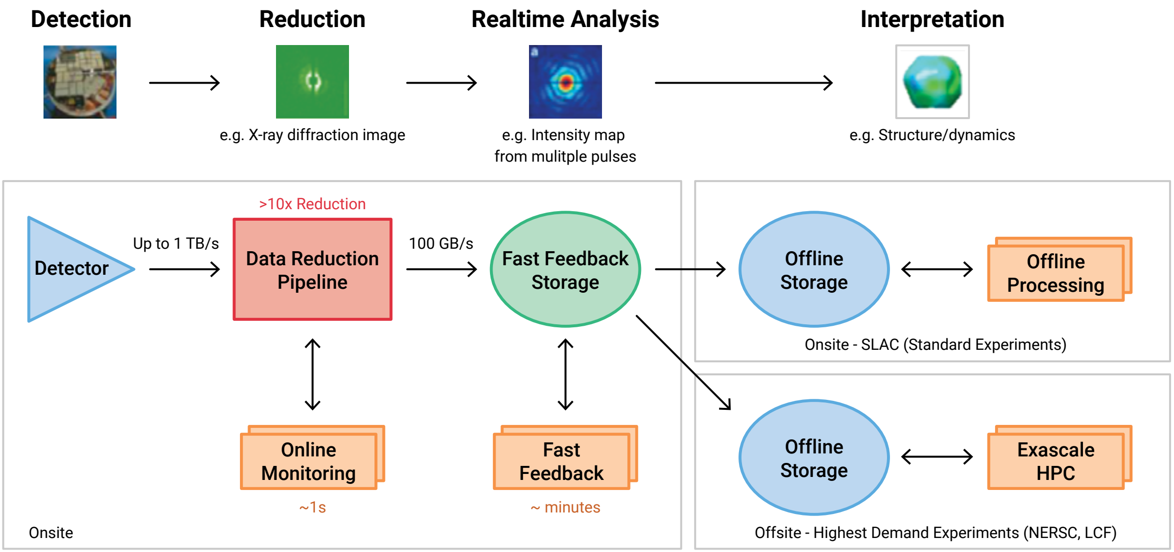


Figure 8. A four-step description of the data flow for LCLS-II and LCLS-HE from the October 2018 report "LCLS Strategic Facility Plan."

Great strides have been made with analysis of current XFEL data generation, as well as projections of future data generation, and many gaps have been identified in tools for compression, streaming data formats, and others. LCLS has formulated a description of data-reduction requirements and a data flow that includes data reduction (Figure 8). Time is of the essence as upgraded facilities come online now and in the next few years. The community is aware that, although a difficult and unpopular task, it is critical for many to take action to address the gaps in data reduction and extraction planning and tools. In an ideal world all data would be saved while at the same time analysis would keep up with experiment. According to the talk from Jana Thayer from LCLS-SLAC, if all the raw data from LCLS-II were stored, data storage costs would top $250 million by 2026! However, without the resources or technology to save, manage, and access all the data that is generated, the pragmatic approach is to manage data reduction in a way that culls out what is important during an experiment, does an on-the-fly analysis where possible, compresses and saves what can be used for post-processing, and discards what is least likely to be of use in future analyses. Even for those small data experiments that are experiencing growth, many of the data-reduction and extraction planning techniques are applicable.

Gap 6: Experimental Design Allied to Research Questions

Experimental plans are usually generated manually, well before an experiment, and are based on sparse historical data, limited simulations, or intuition. However, with data archives, access to advanced simulations, and advanced statistical methods, these experiment designs can be made in a way that utilizes current knowledge and statistically guides the experiment towards areas of less certainty so that the search space of possible parameters is better known. Experimental plans can also guide experiments to areas of the search space that are more likely to produce experimental results in line with desired outcomes. Analytical software stacks that include simulation, machine-learning, and data science technologies are lacking, yet would provide systematic, principled, and automated support for pre- and in-experiment design. Materials experiments tend to have a high level of complexity with usually multiple measurement/diagnostic capabilities and several experiment parameters that require modifications during an experiment. As repetition rates increase, automation will be used for fast analysis. Advanced and semi-automated experiment design will be needed to quickly re-plan the experiment based on data taken so far.

5 Priority Research Opportunities

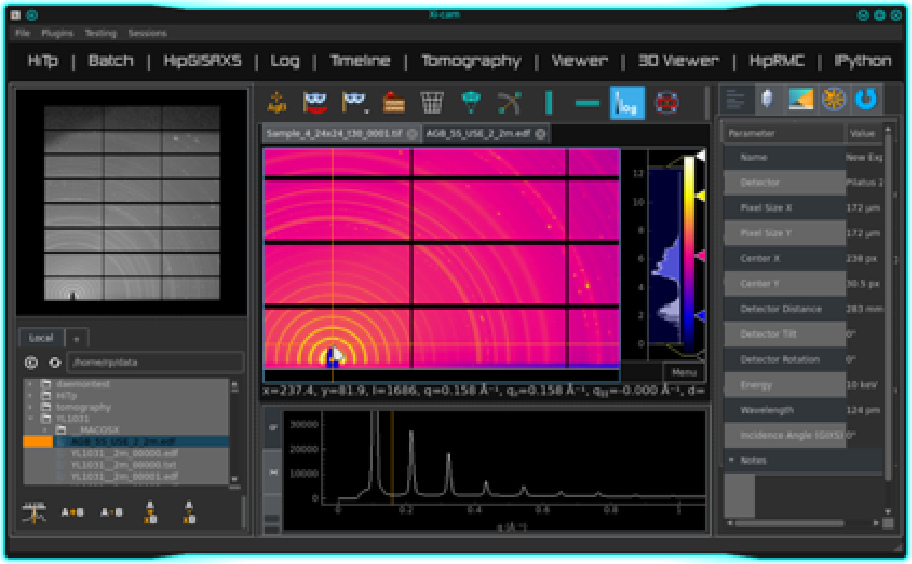
Workshop attendees converged on six priority research opportunities (PROs) for data science for materials science at light sources. The following subsections describe challenges and opportunities that motivate the PRO, the state of the art, a description of the PRO including assumptions and dependencies, the potential impact of the research, and a timeline for when success will impact materials science.

Before diving into the PROs it is useful to notice that each of them is made more challenging due to the increased experiment, detector, data rates, the need to support real-time decision making, the wide variety of experiments/facilities that need to be supported, and the resources available to provide the production-quality software required to earn the trust of experimental scientists. For brevity, these challenges will not be described in depth for each priority research opportunity. However, tackling these challenges will help enable light source science to reach its full potential.

It is also worth noting that although none of these PROs directly address Gap 4, “Software literacy (both languages and facility infrastructure),” it is clear that in order to pursue these priority research directions below and make a scientific impact via data science, software literacy initiatives are needed. Educational initiatives are not inherently PROs, but could be efficiently intertwined with workforce building and research efforts and targeted to support the PROs listed below.

PRO 1: Real-Time Decision-Making Tools: Automation, Data Analytics, Data-Informed Decisions Beyond Hit/Miss, Visualization

To address Gap 3, “Missed Opportunities During Experiment,” real-time decision-making tools are needed. “Opportunities” in this context include being able to better utilize scarce experiment time in any of the following ways: make experiment parameter modification decisions more quickly, re-plan the experiment design based on current data, detect and recover from errors while there is still time, notice anomalies that might lead to a scientific discovery or show that the experiment is leading to an area of less certainty, or realize that enough data has been taken and it is possible to switch from one experimental sample or setup to a new one.


Xi-cam
File Plugins Testing Sessions
HiTp | Batch | HipGISAXS | Log | Timeline | Tomography | Viewer | 3D Viewer | HipRMC | IPython
Local
/home/rp/data
Sample_4_24x24_t30_0001.tif
AGB_5S_USE_2_2m.edf
Menu
q (Å⁻¹)

x=237.4, y=81.9, I=1686, q=0.158 Å⁻¹, q_r=0.158 Å⁻¹, q_z=-0.000 Å⁻¹, d=

Parameter	Value
Name	New Exp
Detector	Pilatus 2
Pixel Size X	172 μm
Pixel Size Y	172 μm
Center X	238 px
Center Y	30.5 px
Detector Distance	283 mm
Detector Tilt	0°
Detector Rotation	0°
Energy	10 keV
Wavelength	124 pm
Incidence Angle (GIXS)	0°
Notes	

Figure 9. Xi-CAM is a general platform for experiment data analysis, management, and visualization. (Image courtesy the LBNL CAMERA team [21]).

Again, the biology domain is further along in the use of real-time decision-making tools for popular light source experiments. During serial femtosecond crystallography experiments, for example, scientists are able to see in real time the development of an electron density map of a protein. They are then able to change samples as soon as they get the desired resolution. Real-time generic diffraction analysis tools are also commonly available at advanced user light sources; however, they do not give domain-specific analytics that are necessary for high-level decision making.

For this research opportunity, tools that do a first-generation graph of experiment data are not sufficient to aid the scientist in obtaining the highest level of information that enables fast decision making. What is needed are tools that get to higher levels or meta-levels, where experiment data are shown in aggregate and in a specific way that presents the experimentalist with a few select options along with associated uncertainties and/or likelihoods for success. Tools must allow for drill-down when necessary so that the scientist can see the supporting data when desired. These tools need to be able to incorporate multimodal data where necessary (e.g., experimental, science simulation, beamline simulation, historical data, current data, data from related materials, data from other experimental facilities, and others.). Targeted usage of machine learning, data science, statistics, image processing advanced mathematics, and algorithms will be necessary. Visual decision making with a human in the loop needs to also be backed up by rationale for the recommendations and associated uncertainties. Workflows that allow simulations, reinforcement learning, and big data analysis to be done on the fly are also needed here. Tools such as Xi-CAM (Figure 9) are an exciting start for this kind of analysis [21]. Although dependencies exist on research in multimodal analysis and efficient workflows, work in this priority research direction will impact science immediately.

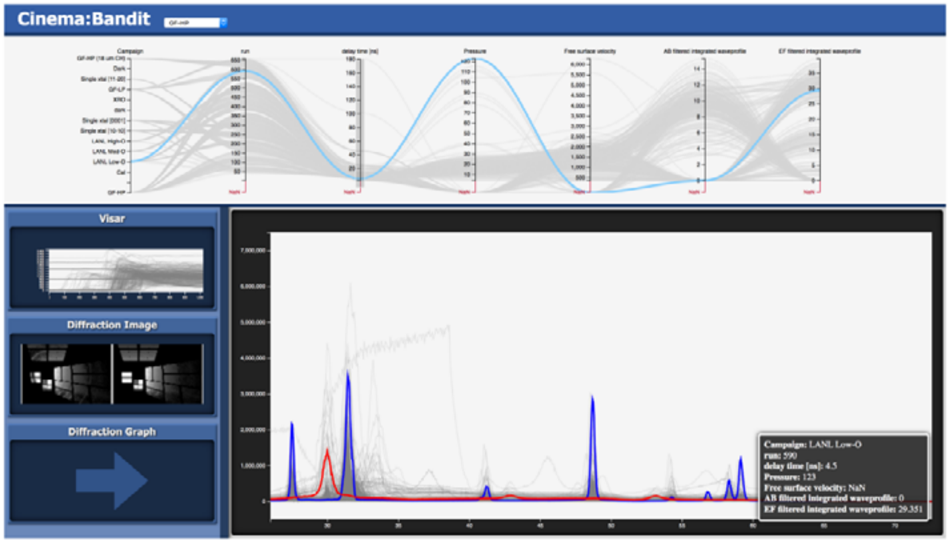
PRO 2: Codesign of Experiments with Data Analysis, End-to-End Planning, and Design of the Experiment

To address Gap 6, “Experimental Design Allied to Research Questions” and analytical software stack, research is needed to assist scientists in viewing their experiment holistically in terms of all the possible experiment runs that can be done and various parameters. Opportunities here are being more systematic about planning an experiment, going in with quantitative metrics on what is known and not known, (machine) learning from available experimental and simulation data to extrapolate to unknown data areas and plan accordingly, and being agile in re-planning, because the plans and learning are modifiable and automatable. A unique challenge here is the high-dimensionality of the input parameter space beyond what users can manage manually. The state of the art is that this area is in its infancy—few capabilities exist in this area. For example, it was difficult to find statisticians familiar with light source data and simulations to attend the workshop. Only recently (Winter 2018) the first Machine Learning for Particle Accelerators workshop was held, sponsored by SLAC. However, if statistical experimental design tools such as emulators can be developed, they can provide the needed real-time feedback necessary for success (See the feedback arrow in Figure 5).

Research in codesign of experiments with data analysis includes research in statistical emulation [22], statistical experimental design, beamline simulation, beamline emulation for possible feedback control, and end-to-end simulation of an experiment along with its analysis. End-to-end planning of the experiment includes parameter selection, data reduction planning, data analysis on the fly, data transfer speed planning, metadata capture, real-time decision-making tools, loop back to control, and post-processing. It depends on the availability of historical data and beamline simulation technology as well as the availability of experiment metadata. This is an ambitious priority research area. It will be a while before all the dependencies can be put into place for it to make an impact scientifically because there are so many dependent, connected aspects. However, initial efforts in this area, as presented by Simon Billinge, are already revolutionizing materials discovery. Future efforts will greatly accelerate these advances.

PRO 3: High-Dimensional Data Visualization and Interaction

Light source experiment parameters and the resulting experiment data are both of higher dimension—high enough to be beyond human cognition without tools. New tools are needed to address Gap 1, “Tools for Exploiting High-Dimensional and Multimodal Data.” Challenges here are reconstructing images as needed from vast amounts of data, incomplete data, or noisy data. Time is another dimension that can make the visualization more challenging, but provides opportunities in making movies of experimental data. The current state of the art includes a number of advanced reconstruction algorithms and tools developed at Lawrence Berkeley National Laboratory (LBNL) [23] and Argonne National Laboratory (ANL) [24], as well as new tools becoming available, such as Los Alamos National Laboratory’s (LANL’s) Cinema [25, 26] (see Figure 10) and Ensight’s NEXUS [27] for allowing exploration of data along different dimensions.



Campaign: LANL Low-O
run: 590
delay time [ns]: 4.5
Pressure: 123
Free surface velocity: NaN
AB filtered integrated waveprofile: 0
EF filtered integrated waveprofile: 29.351

Figure 10. Example of a visualization tool for N-dimensional data from dynamic compression experiments at LCLS entitled “Cinema:Bandit” [25, 26].

Research in this area requires a continued push in advanced mathematics, algorithms, and machine learning to help with approximating solutions, or to help with making reconstructions faster. Usability is also an area of research here, because these tools will involve human interaction and require care in design so that they are intuitive, have low cognitive load when used, and be generally useful for a number of experiments. Flexibility in visual analytics is also important so that tools flex with the level of complexity desired as it evolves during an experiment. Tools should not just involve data analytics, but also physical models that help explain the data obtained. Visualization tools with a common application programming interface (API) or graphical user interface that can be used pre-experiment, during the experiment, and post-experiment are also an area for research, so as to reduce the number of tools an experimental scientist needs to learn and use.

PRO 4: Capturing Metadata Non-intrusively, Both Automatically and Manually, Prior to the Experiment

Capturing metadata non-intrusively addresses Gap 2, “Data, Algorithm, and Software Curation.” Additionally, it supports most of the other PROs and the reproducibility of experiments. Metadata, which is data that describes other data, are essential to support activities, such as storage of data in databases. Metadata also allow access to various types of data science that rely on structured data, such as machine learning. Examples of metadata that need to be captured include information on x-ray parameters (photon energy, intensity, and others), sample parameters (type, temperature, preparer, and more), detector parameters (exposure time, binning, and region of interest), experimental parameters (pump or bias information, delay time, temperature, and other environmental parameters), and even information about the experimenters.

The challenge is that data are highly dependent upon the conditions under which they are generated. Usually, these dependencies are recorded as metadata. When it comes to experimental data, it can be very difficult, if not impossible, to record of every possible factor involved in generating the data. Even then, for future access and sharing, consensus must be reached on metadata labels and how the metadata will be stored. Currently, most metadata are not recorded except for the most high-level metadata entered into spreadsheets or logbooks during an experiment, or extremely low-level metadata from the data acquisition system.

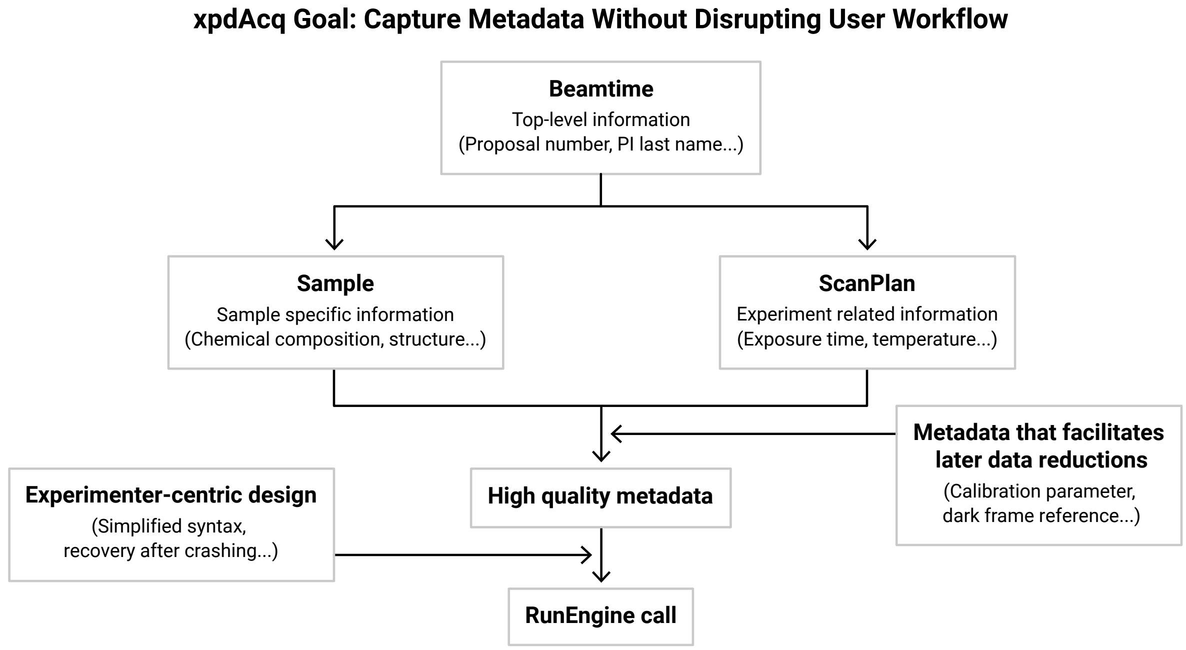

xpdAcq Goal: Capture Metadata Without Disrupting User Workflow
Beamtime
Top-level information
(Proposal number, PI last name...)
Sample
Sample specific information
(Chemical composition, structure...)
ScanPlan
Experiment related information
(Exposure time, temperature...)
Metadata that facilitates later data reductions
(Calibration parameter, dark frame reference...)
Experimenter-centric design
(Simplified syntax, recovery after crashing...)
High quality metadata
RunEngine call

Figure 11. Proposed method for capturing metadata by workflow provenance engineering proposed by Billinge et al. (Source: Billinge workshop presentation.)

Automatic and manual metadata capture in a nonintrusive way requires research in understanding metadata requirements. For automation-related metadata, what is needed to perform the experiments at the beamline? Some possibilities are the minimum viable metadata related to experiment conditions, metadata needed just to analyze the data, and metadata needed to link the experiment with the analysis. For domain/sample-related metadata, what data are needed for the larger context of the data—the material or the science data? Metadata are highly reliant on how the data will be used and what is needed for reproducibility.

PRO 5: Multimodal, Multiscale, Holistic Analysis via Open Data Resources and Search Tools

Increasingly, scientists are interested in using multimodal experimental data related to materials, which include different data types used together, previous data, data from related experiments, experiments at different facilities, simulation data, and emulated data. Also, there is interest in data at different scales. This could be length scale of nanoscale versus macroscopic dimension. It could also be short- and long-time scale for a dynamic process for the material or combining data at different scales into a holistic analysis. Current state of the art in multimodal, multiscale, holistic analysis is that it is just emerging and materials databases are also just emerging [19]. They are not yet in wide use, however.

This is a challenging area of research, not just due to data fusion aspects, but also data availability. In order to successfully combine data, access to appropriate data is required. Open data resources will be increasingly needed as well as tools to search over data. Automated searches in the form of bots can help bring data to the analysis. Bots are programs that run automated tasks over the data, either internet data or other data sources. This PRO is in many ways dependent on PRO 4, "Capturing Metadata Non-intrusively," because well-abstracted metadata will be useful for open data resources and searchability tools.

6 Summary and Suggested Activities

In addition to the gaps and priority research opportunities identified, a number of supporting activities were suggested by the participants:

- Follow-on workshop(s): A general follow-on to this workshop or a workshop exploring a particular gap or research opportunity.
- Activities that support the educational gap: Internships, summer schools, courses, hackathons.
- Funding opportunities (Laboratory Directed Research and Development [LDRD], Basic Energy Sciences [BES], Advanced Scientific Computing Research [ASCR], and others.) that mix people from backgrounds, such as science, computer science, statistics, and mathematics, so that they learn each other's languages.
- Proposal process: Make building multidisciplinary teams a criterion for success and make data analysis readiness part of the criterion.
- Community platforms, such as data stores, code sharing sites, and more.
- Periodic town hall meetings either via video conferencing or in person at user meetings to discuss and brainstorm gaps and research topics and engage in community building.

This workshop on analyzing the gap between the rapidly advancing light sources and the availability and application of advanced data science tools demonstrated the urgent need to apply data science tools to x-ray materials studies. Exciting opportunities and challenges were identified and discussed. To all the participants it was obvious that these opportunities need to be addressed with the above activities or the full potential of materials discovery and innovation at advanced x-ray light sources will never be realized.

7 Acknowledgements

This workshop was sponsored by Los Alamos National Laboratory. The idea for the workshop was originally conceptualized by LANL's Richard Sheffield in 2017.

Hotel Santa Fe hosted the workshop and many of its participants. We thank the hotel for its hospitality and for accommodating our specific workshop needs.

New Mexico Consortium provided water bottles, which were much appreciated by all attendees.

Thank you to the steering committee for helping to bring the workshop idea to fruition, those at LANL who helped in the planning and execution of the workshop, and to all who traveled to Santa Fe to attend, contributed to discussions, gave presentations, and helped write this report.

8 References

1. M. Eriksson, J. F. van der Veen, and C. Quitmann, "Diffraction-limited storage rings—a window to the science of tomorrow," J. Synch. Radiat. 21, 5: 837–842 (2014).

2. E. Weckert, "The potential of future light sources to explore the structure and function of matter," *IUCrJ.* 2, 2: 230–245 (2015).

3. C. Bostedt et al., "Linac Coherent Light Source: The first five years," *Rev. Mod. Phys.* 88, 1: 015007, Mar. 2016.

4. BESAC, Science for Energy Technology: Strengthening the Link between Basic Research and Industry (2010).

5. Bergmann, Uwe et al., Science and Technology of Future Light Sources: A White Paper. SLAC-R-917. (2008).

6. R. Schoenlein, P. Abbamonte, F. Abild-Pedersen, P. Adams, M. Ahmed, F. Albert, R. A. Mori, A. Anfinrud, A. Aquila, M. Armstrong, and others, "New science opportunities enabled by LCLS-II x-ray lasers," SLAC Rep. SLAC (2015).

7. Advanced Photon Source Upgrade Project Preliminary Design Report (2017).

8. BESAC, Directing Matter and Energy: Five Challenges for Science and the Imagination (2007).

9. J. Hemminger, G. Crabtree, and J. L. Sarrao, From Quant to Continuum: Opportunities for Mesoscale Science (DOE-BESAC Report) (2012).

10. J. C. Hemminger, J. Sarrao, G. Crabtree, G. Flemming, and M. Ratner, Challenges at the Frontiers of Matter and Energy: Transformative Opportunities for Discovery Science (2015).

11. W. House, Materials Genome Initiative for Global Competitiveness (2011), (June).

12. The Minerals Metals & Materials Society (TMS), 'Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering' (TMS, 2017).

13. https://science.energy.gov/~/media/bes/pdf/reports/2018/Ultrafast_x-ray_science_rpt.pdf

14. P. A. Hart et al., "First x-ray test of the Icarus nanosecond-gated camera," in *X-Ray Free-Electron Lasers: Advances in Source Development and Instrumentation V*, 2019, 11038: 27.

15. The Exascale Computing Project. https://www.exascaleproject.org/

16 United States, Executive Office of the President, Executive order 13859: Maintaining American Leadership in Artificial Intelligence. 19 Feb. 2019. Federal Register. 84, 31, 19 Feb. 2019: 3967–3972. https://www.hsdl.org/?abstract&did=821398.

17. Reinsel, D., Grantz, J., Rydning, J., Data Age 2025 (IDC White Paper, 2017), https://www.seagate.com/our-story/data-age-2025.

18. AWS Snowmobile. https://aws.amazon.com/snowmobile/.

19. Jain A. et al. (2018), The Materials Project: Accelerating Materials Design Through Theory-Driven Data and Tools. In: Andreoni W., Yip S. (eds.) *Handbook of Materials Modeling*. Springer, Cham.

20. Citrine Informatics: The AI Platform for Materials Development, https://citrine.io/

21. R. J. Pandolfi et al., "Xi-cam: a versatile interface for data visualization and analysis," *J. Synch. Radiat.* 25, 4: 1261–1270, Jul. 2018.

22. D. J. Walters, A. Biswas, E. C. Lawrence, D. C. Francom, D. J. Luscher, Fredenburg, D. A., K. R. Moran, C. M. Sweeney, R. L. Sandberg, J. P. Ahrens and C. A. Bolme. Bayesian calibration of strength parameters using hydrocode simulations of symmetric impact shock experiments of Al-5083. *Journal of Applied Physics*. 124, 20: 205105. (LA-UR-18-20884 DOI: 10.1063/1.5051442).

23. The Center for Advanced Mathematics for Energy Research Applications (CAMERA). https://www.camera.lbl.gov/.

24. D Gürsoy, F De Carlo, X Xiao, C Jacobsen, J. Synch. Radiat. 21, 5: 1188–1193. TomoPy: a

framework for the analysis of synchrotron tomographic data

25. James Ahrens, Sébastien Jourdain, Patrick O'Leary, John Patchett, David H. Rogers, and Mark Petersen. An image-based approach to extreme scale in situ visualization and analysis. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '14). IEEE Press, Piscataway, NJ, USA: 424–434, 2014. DOI: https://doi.org/10.1109/SC.2014.40.

26. Daniel Orban, Divya Banesh, Cameron Tauxe, Christopher M. Biwer, Ayan Biswas, Ramon Saavedra, Christine Sweeney, Richard L. Sandberg, C. A. Bolme, James Ahrens and David Rogers, "Continuous work flow and exploration of disparate data types using Cinema:Bandit: A toolset for beamline science demonstrated on XFEL shock physics experiments," submitted to *J. Synch. Radiat.* (2019).

27. ANSYS Ensight NEXUS. https://nexusdemo.ensight.com/docs/html/Introduction.html.

9 Appendices

Appendix A: Abbreviations, Acronyms, Initialisms

Appendix B: Workshop Website

Appendix C: List of Participants

Appendix D: Workshop Agenda

Appendix E: Keynote, Plenary, and Lightning Talk Abstracts

Appendix F: Experiment Sessions

Appendix G: Data Science Techniques

Appendix H: Summary Table of Experiment Requirements

Appendix I: Summary of Prior Work

Appendix J: Experiment, Measurement, and Characterization

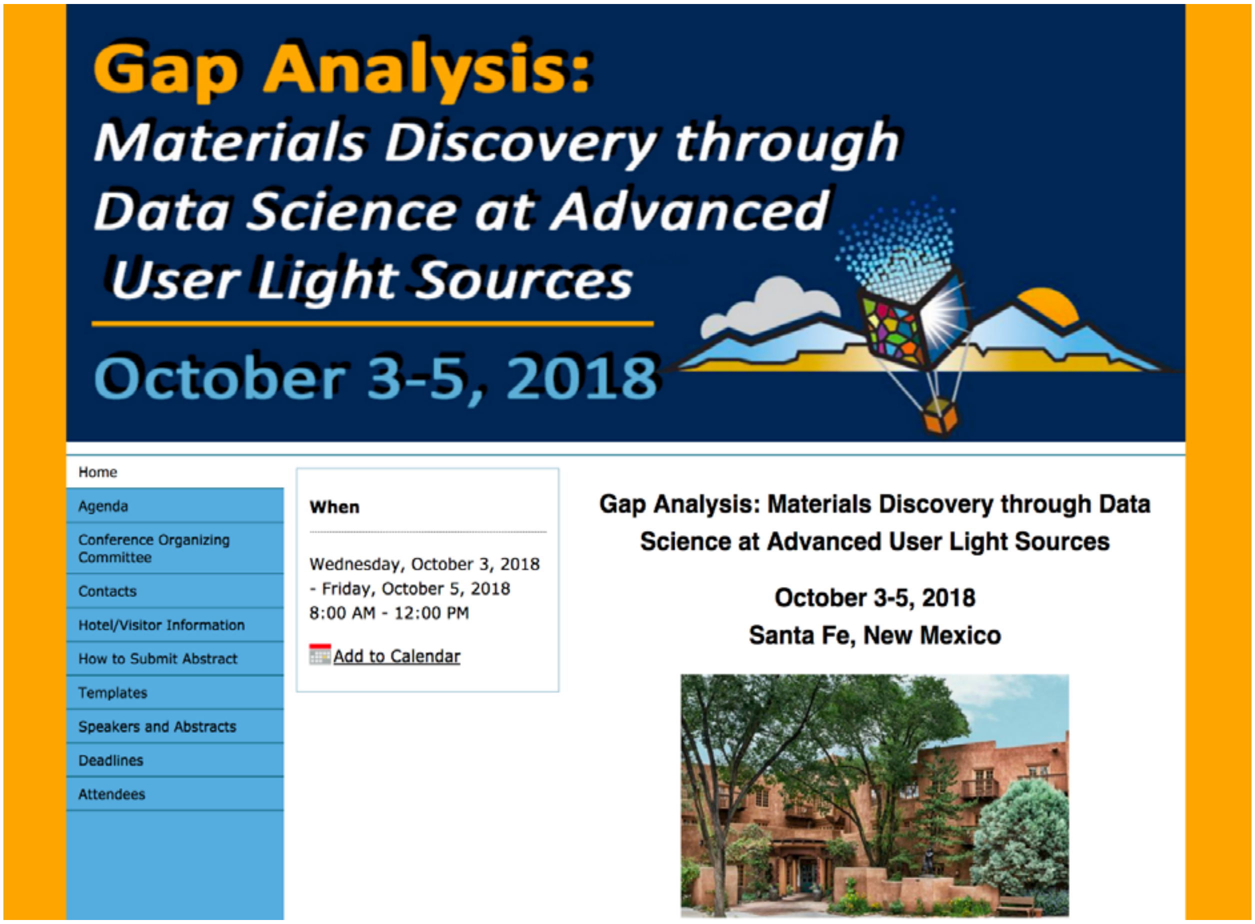
Appendix A: Abbreviations, Acronyms, Initialisms

ANL	Argonne National Laboratory
API	application programming interface
APS	advanced photon source
APS-U	advanced photon source upgrade
ASCR	Advanced Scientific Computing Research
AWS	Amazon Web Services
BCDI	Bragg coherent diffractive imaging
BES	Basic Energy Sciences
BESAC	Basic Energy Sciences Advisory Committee
BNL	Brookhaven National Laboratory
CDI	coherent diffraction imaging
CDI NN	coherent diffraction imaging neural networks
CPA	coherent potential approximation
DAC	diamond anvil cell
dDAC	dynamic diamond anvil cell
DFT	density functional theory
DLSR	diffraction-limited storage rings
DOE	Department of Energy
ECB	Extreme Conditions Beamline
Eu-XFEL	European XFEL
HPC	high performance computing
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LCLS	Linac Coherent Light Source
LDRD	Laboratory Directed Research and Development
LLNL	Lawrence Livermore National Laboratory
MCMC	Markov chain Monte Carlo

MDF	Materials Data Facility
MDI	materials data infrastructure
NIST	National Institute of Standards and Technology
NSLS-II	National Synchrotron Light Source II
PRO	priority research opportunities
RIXS	resonant inelastic x-ray scattering
SLAC	Stanford Linear Accelerator Center
SNL	Sandia National Laboratory
XFEL	x-ray free electron laser
XPCS	x-ray photon correlation spectroscopy
XRD	x-ray diffraction

Appendix B: Workshop Website

www.lanl.gov/2018gapanalysis



Gap Analysis:
Materials Discovery through Data Science at Advanced User Light Sources

October 3-5, 2018

When

Wednesday, October 3, 2018 - Friday, October 5, 2018
8:00 AM - 12:00 PM

Gap Analysis: Materials Discovery through Data Science at Advanced User Light Sources

October 3-5, 2018
Santa Fe, New Mexico

Appendix C: List of Participants

Ahrens, James LANL	Kleese van Dam, Kerstin BNL
Bethel, Wes LBNL	Lawrence, Earl LANL
Billinge, Simon BNL	Lin, Meifeng BNL
Bingham, Derek Simon Fraser	Lookman, Turab LANL
Biswas, Ayan LANL	Mazzoli, Claudio BNL
Biven, Laura DOE/SC/ASCR	Meredig, Bryce Citrine Informatics
Bremer, Peer-Timo LLNL	Mertes, Kevin LANL
Campbell, Stuart BNL	Nelson Weker, Johanna SLAC
Casa, Diego APS/ANL	Pateras, Anastasios New Mexico State University
Cherukara, Mathew ANL	Perazzo, Amedeo LCLS
Coffee, Ryan SLAC	Pokharel, Reeju LANL
Cyr, Eric SNL	Post, John LLNL
Daniel, Allan BNL	Pouchard, Line BNL
Di, Zichao (Wendy) ANL	Proffen, Thomas ORNL
Donatelli, Jeffrey LBNL	Robinson, Ian BNL
Dunne, Mike SLAC	Rollett, Anthony Carnegie Mellon University
Dutta, Soumya LANL	Sanchez, Marcos SNL
Dwaraknath, Shyam LBL	Sandberg, Richard LANL
Eggert, John LLNL	Sariaydin, Selin, ANL
England, Troy SNL	Scheinker, Alexander LANL
Francom, Devin LANL	Schold, Elijah New Mexico State University
Fuoss, Paul SLAC	Schram, Malachi PNNL
Gleason, Arianna SLAC/Stanford	Sheffield, Richard LANL
Graziani, Carlo ANL	Sweeney, Christine LANL
Hagberg, Aric LANL	Tang-Kong, Robert Stanford Synchrotron Radiation Lighsource
Hexemer, Alexander LBNL	Thayer, Jana LCLS/SLAC
Hruszkewycz, Stephan ANL	Ushizima, Daniela LBNL/UC Berkeley
Jorrissen, Kevin Amazon Web Services	Vogt, Stefan ANL
Jenei, Zsolt LLNL	Ward, Logan ANL
Jha, Shantenu BNL	Williams, Garth BNL
Jorrissen, Kevin Amazon Web Services	

Appendix D: Workshop Agenda

Los Alamos National Laboratory **Agenda**

NNSA National Nuclear Security Administration

Final 2

Gap Analysis: Materials Discovery through Data Science at Advanced User Light Sources
Hotel Santa Fe
October 3-5, 2018

Wednesday, October 3, 2018
Hotel Santa Fe, KIVA Conference Rooms A and B

8:00 am – 8:05 am	*Welcome, Committee Introductions,* *Working Breakfast*	*Christine Sweeney, LANL*
8:05 am – 8:15 am	*Formal Welcome, Opening Conference Remarks*	*John Sarrao, LANL, PADSTE*
8:15 – 9:05 am	Keynote Address: Opportunities and Challenges	Mike Dunne, SLAC/LCLS
9:05 – 9:35	Charge to Participants and Conference Approach	Christine Sweeney, LANL
9:35 – 10:20	Plenary Session 1 – Data Types	Simon Billinge, BNL
10:20 – 10:50	*Group Photo and Break*	
10:50 – 11:10	Experiment Talk No. 1 – X-ray Photon Correlation Spectroscopy (XPCS)	Claudio Mazzoli, NSLS, BNL
11:10 – 11:30	Experiment Talk No. 2 – Resonant Inelastic X-ray Scattering (RIXS)	Diego Casa, APS, ANL
11:30 – 12:00	Discussion	Kerstin Kleese Van Dam, BNL
12:00 – 1:00	*Working Lunch – Plenary Session 2: Industry/Cloud/Big Data*	*Kevin Jorissen, AWS*
1:00 – 1:45	Plenary Session 2 – Experimental Design	Derek Bingham, Simon Fraser
1:45 – 2:05	Experiment Talk No. 3 – Bragg Coherent Diffraction Imaging (BCDI)	Ian Robinson, BNL
2:05 – 2:25	Experiment Talk No. 4 – Dynamic X-ray Diffraction at High Pressures (Dynamic XRD)	Arianna Gleason, SLAC
2:25 – 2:55	Discussion	Richard Sandberg, LANL
2:55 – 3:10	*Break*	
3:10 – 3:30	Experiment Talk No. 5 – High Pressure Diamond Anvil Cell X-ray Diffraction (XRD w/ DAC)	Zsolt Jenei, LLNL
3:30 – 3:50	Experiment Talk No. 6 – High Energy Diffraction Microscopy (HEDM)	Reeju Pokharel, LANL
3:50 – 4:20	Discussion	Garth Williams, BNL

Institutional Host:	John Sarrao, PADSTE	Agenda Revised: 10/02/18
Technical Host:	Christine Sweeney, CCS-7, 505-606-0195	
Protocol Contact:	Peggy Vigil, GAP, 505-667-8448; Cell: 505-699-2195	

Gap Analysis: Materials Discovery through Data Science at Advanced User Light Sources
Hotel Santa Fe
October 3-5, 2018

Thursday, October 4, 2018 - Hotel Santa Fe, KIVA Conference Rooms A, B and C

8:00 am – 8:15 am	*Welcome to Day 2, Announcements, Working Breakfast, Charge for today's agenda*	*Christine Sweeney, LANL*
8:15 – 9:05 am	Plenary Session 3 - Mathematics	Jeffrey Donatelli, LBNL
9:05 – 10:00	Plenary Session 4 - Modeling and Simulation	Matthew Cherukara, APS
10:00 – 10:15	*Break*	

Parallel Breakout Sessions, KIVA Conference Rooms A, B and C

10:15 – 11:05 – Parallel Session – Kiva A and B

Pre-experiment Briefing A – Machine Learning Pre-experiment	Ryan Coffee, SLAC
Discussion of Machine Learning Pre-experiment	Alex Sheinker, LANL

10:15 – 11:05 – Parallel Session – Kiva C

Pre-experiment Briefing B – Modeling and Simulation	Turab Lookman, LANL
Discussion of Modeling and Simulation	Thomas Proffen, ORNL/SNS

11:10 – 12:00 – Parallel Session – Kiva A and B

In-experiment Briefing A – Visual Analytics, Decision Making	James Ahrens, LANL
Discussion of Visual Analytics, Decision Making	Timo Bremer, LLNL

11:10 – 12:00 – Parallel Session – Kiva C

In-experiment Briefing B – Statistics and Emulation	Devin Francom, LANL
Discussion of Statistics and Emulation	Earl Lawrence, LANL

12:00 - 1:00	*Working Lunch – Plenary Session – Data Reduction*	*Jana Thayer, SLAC*
1:00 – 1:20	Reports from Morning Sessions	Discussion Moderators
1:20 – 3:00	Lightning Talks	Various Speakers
3:00 – 3:15	*Break*	

Parallel Breakout Sessions, KIVA Conference Rooms A, B and C

3:15 – 4:05 – Parallel Session – Kiva A and B

Post-experiment Briefing A – Physics and Math in Data Analysis	Stephan Hruszkewycz, ANL
Discussion of Physics and Math in Data Analysis	Jeffrey Donatelli, LBNL

3:15 – 4:05 – Parallel Session – Kiva C

Post-experiment Briefing B – Data Management for Data Science	Daniel Allan, BNL
Discussion of Data Management for Data Science	Amedeo Perazzo, LCLS

LOS ALAMOS NATIONAL LABORATORY **Agenda** NNSA

Gap Analysis: Materials Discovery through Data Science at Advanced User Light Sources
Hotel Santa Fe
October 3-5, 2018

Thursday, October 4, 2018 (CONTINUED)
Hotel Santa Fe, KIVA Conference Rooms A, B and C

4:05 – 4:55 – Parallel Session – Kiva A and B

Systems Briefing A – Materials Database	Logan Ward, ANL
Discussion of Materials Database	Bryce Meredig, Citrine

4:05 – 4:55 – Parallel Session – Kiva C

Systems Briefing B – Data Mining	Daniela Ushizima, LBNL
Discussion of Data Mining	Aric Hagberg, LANL

4:55 – 5:15	Reports from afternoon sessions	Discussion Moderators
5:15 – 5:30	Preview of Day 3	Christine Sweeney, LANL
5:30	Participants - Dinner on your own	
7:30	Writing Groups – on your own	

LOS ALAMOS NATIONAL LABORATORY **Agenda** NNSA

Gap Analysis: Materials Discovery through Data Science at Advanced User Light Sources
Hotel Santa Fe
October 3-5, 2018

Friday, October 5, 2018
Hotel Santa Fe, KIVA Conference Rooms A, B, C and Library

8:00 am – 8:15 am	*Welcome to Day 3, Announcements, Working Breakfast, Charge for today's agenda*	*Christine Sweeney, LANL*
8:15 – 9:15	Discuss Priority Research Directions	All
9:15 – 10:30	Finalize Writing Priority Research Directions	All
10:30 – 10:45	*Break*	
10:45 – 11:45	Report Back – **Room Kiva A/B**	All
11:45 – 12:00	Wrap-up and Closing Comments	Christine Sweeney, LANL
12:00	Conference ends	

Appendix E: Keynote, Plenary, and Lightning Talk Abstracts

1.1 Keynote and Plenary Abstracts

1.1.1 Keynote: The Bright Future of X-Ray Science

Mike Dunne, LCLS/SLAC

The past decade has seen the emergence of x-ray free-electron lasers (XFELs) as a powerful new tool for studying the world at the atomic and molecular scale [1], with applications to quantum materials, catalytic chemistry, the science of extreme conditions, and structural biology, to name a few [2]. These facilities provide ultrashort x-ray pulses with a peak brilliance over nine orders of magnitude higher than synchrotron sources—allowing us to capture atomic-level detail on femtosecond timescales using a wide range of coherent imaging and spectroscopy tools.

This field is now entering another step-change, with the repetition rate of the sources increasing by many orders of magnitude to provide high average power beams that can track rare and transient phenomena, or study heterogeneous systems with stochastic properties, isolated defects or buried interfaces [3]. Repetition rates will increase from about100 hertz to 1 megahertz, in which each pulse is capable of delivering a multi-megapixel image of high dynamic range and rich scientific content. Our challenge is to develop x-ray cameras, sample delivery systems, and data acquisition and analysis tools to take full advantage of these remarkable new sources.

Billions of dollars are being invested in the U.S., Europe, and Asia to construct these next-generation user facilities. Data generation could be over 1 terabyte per second, and will require real-time reduction and analysis of that data to dynamically guide the experiments. This will entail coordinated development of intelligent data extraction and compression techniques that are compatible with the preservation of high-fidelity information. Methods are needed to pipe these data to exascale computers with numerical tools to take full advantage of accelerated architectures. Artificial intelligence algorithms need to be developed and validated that can help identify and interpret critical information in the data, along with user-friendly tools to enable utilization by a broad scientific community.

The scientific impact of success will be profound, providing a wealth of fundamental scientific insights that will benefit societal priorities from human health to renewable energy technology, national security, and our understanding of the cosmos.

This talk will provide an overview of the current state of the field and the pace and nature of the developments underway in the U.S. and around the world. It will present examples of the scientific opportunities and the approaches to data analysis that are being explored.

This is a rapidly growing field in which entirely new approaches to data analysis and facility optimization are required. As such, it is ideally suited to those new to the field and those who look forward to a career full of new challenges and broad impact.

References
[1] W. White, A. Robert, M. Dunne, J. Synchrotron Rad. 22, 472–476 (2015) and references therein.
[2] C. Bostedt et al., Reviews of Modern Physics 88, 1 (2016).
[3] M. Dunne, Nature Review Materials (2018), doi:10.1038/ s41578-018-0048-1.

1.1.2 Plenary: Sequential Experimental Design for Model Calibration

Derek Bingham, professor, Department of Statistics and Actuarial Science, Simon Fraser University

In many branches of physical science, when the complex physical phenomena are either too expensive or too time-consuming to observe, deterministic computer codes are often used to simulate these processes. Nonetheless, true physical processes are also observed in some disciplines. It is preferred to integrate both the true physical process and the computer model data for a better understanding of the underlying phenomena. In this talk, methodology is presented for selecting optimal experimental or simulation trials designs based on integrated mean squared error that help us capture and reduce prediction uncertainty as much as possible. The aim is to use this methodology within a fast model calibration framework so that the methodology can be deployed in real time.

1.1.3 Plenary: Data-Driven 4D X-Ray Imaging of Nanoscale Dynamics

Mathew Cherukara, Assistant Scientist, Center for Nanoscale Materials, Argonne National Laboratory

Observing the dynamic behavior of materials can reveal insights into the response of materials under nonequilibrium conditions of pressure, temperature, and mechanical load. Such insights into materials response under nonequilibrium are essential to design novel materials for catalysis, low-dimensional heat management, piezoelectrics, and other energy applications. However, material response under such conditions is challenging to characterize especially at the nano to mesoscopic spatiotemporal scales. Time-resolved coherent diffraction imaging (CDI) is a unique technique that enables three-dimensional imaging of lattice structure and strain dynamically. In this talk, I will present some examples of our recent work on imaging and modeling of phonon transport and lattice dynamics in nanomaterials under a variety of external stimuli. I will highlight the use of experimentally informed models that leverage large-scale computational resources available at Argonne. These experimentally informed models were used to provide information complementary to the imaging experiment, and at spatio-temporal scales inaccessible to the experiment.

With the APS upgrade (APS-U) it will become possible to image at similar resolution to what is attainable today with significantly shorter acquisition times, or conversely acquire significantly higher resolution data using the same acquisition times as today. Either route will create challenges associated with data/memory that will necessitate unconventional approaches to image recovery. I will describe our work in the use of deep generative neural networks (CDI NN) in accelerating the analysis of, and potentially increasing the robustness of image recovery from x-ray diffraction data. Once trained, CDI NN is thousands of times faster than traditional phase retrieval algorithms used for image reconstruction from coherent diffraction data, opening up the prospect of real-time 3-D imaging at the nanoscale.

1.1.4 Plenary: Impact of Data Reduction on Data Science at LCLS-II

Jana Thayer, Department Head for LCLS Data Systems, LCLS/SLAC National Accelerator Laboratory

Data systems for the future LCLS will need to handle very high data throughputs ranging from hundreds of gigabytes per second to over a terabytes per second. Even assuming a favorable cost and density scaling for network, storage, and processing technologies over the next few years, the costs associated with moving, recording, and processing these amounts of data are prohibitive. We believe that applying on-the-fly data reduction effectively manages the costs of the data systems and mitigates processing times. We present a description of the LCLS-II Data System and associated inline data reduction pipeline that provides a configurable set of tools, including feature extraction, lossless data compression, and event veto to reduce the volume of data written to disk while still producing the same analytical results. Our strategies for reducing the data in a highly changeable operations environment and the effect of these strategies on the design of the data system and on the results of the data analysis will be discussed.

1.1.5 Plenary: Amazon Cloud Resources as Part of Scientific Workflows

Kevin Jorissen, Business Development Manager, Scientific Compute (SC) group, Amazon Web Services

This talk will get scientists thinking about how they can benefit from the virtually limitless resources Amazon Web Services (AWS) offers them for computing, storage, and data analytics. We will start with the general principles that make the public cloud a good match for today’s research challenges, from shortening the time to prove or disprove a new idea, to collaborating on massive data sets. Next, we will review examples of current and upcoming research in the cloud across several science domains from astronomy to genomics, demonstrating the benefits of new technologies and of scale. We’ll then visit key cloud services that will be the building blocks of most scientists’ work: using AWS Elastic Compute Cloud (EC2) for setting up personal compute clusters with hundreds of scientific applications; using AWS Batch and AWS Lambda Serverless to create automated pipelines to analyze incoming data; and Amazon SageMaker to democratize machine learning, so that applied scientists who’ve used a few Python scripts for analysis can suddenly take advantage of massive GPU clusters and optimized deep learning frameworks to train highly accurate and publishable models. We'll close out with some of the ways AWS is engaging with research science and how to get started. Finally, there'll be time for the community to identify its greatest challenges that AWS could help with—from handling unprecedented volumes of detector output, to giving applied scientists more power to analyze their measurements, or any other needs. The speaker has a background in computational materials science and worked for years on the FEFF code for x-ray absorption spectroscopy at the University of Washington.

1.1.6 Plenary: CAMERA: New Mathematics for Enabling the Next Generation of Experimental Science

Jeffrey Donatelli, Computational Research Scientist, Mathematics Group, Deputy Director of the Center for Advanced Mathematics for Energy Research Applications (CAMERA), Lawrence Berkeley National Laboratory

Recent technological advancements have allowed new experiments to make more sensitive measurements, collect more data, and study more complex phenomena than ever before. However, accurately and efficiently analyzing the data from these experiments is increasingly becoming a major bottleneck in enabling new scientific advancements. There is a dire need to develop new targeted mathematics and algorithms to accelerate data processing and analysis, handle increasing amounts of data, improve accuracy and resolution, and model new kinds of physical behavior. In this talk, I will provide an overview of several new mathematical tools being developed at the Center for Advanced Mathematics for Energy Research Applications (CAMERA) aimed at tackling these challenges. In particular, I will describe new machine learning techniques that can greatly decrease the number of learning parameters and training data requirements, a modular approach to modeling and reconstructing molecular structure from noisy, complex, and incomplete data, and tools for enabling autonomous and real-time experiments.

1.1.7 Plenary: Materials Discovery Is More Than Materials Prediction: The Role of Light-Sources in Addressing the Materials Synthesis Data Gap

Simon Billinge, Professor of Materials Science and Engineering and Applied Physics and Applied Mathematics at Columbia University and Scientist at Brookhaven National Laboratory

The discovery of new materials involves prediction of novel materials with useful properties. However, beyond that, we need also to reliably predict the synthesis recipe that we will use to make those novel materials. This is an extremely difficult challenge because, especially for inorganic materials, there currently is no computational physics model for synthesis, which is an inherently kinetic, non-thermodynamic and irreversible process. I will describe how we can take a machine learning approach to the problem, but critically we are missing experimental synthesis data for the algorithms to be trained on. There is a critical role for light-sources, coupled to advanced computation, to play in rectifying this problem. I will discuss current progress, but mostly the possibilities and challenges.

1.2 Lightning Talks

- Simultaneous Sensing Error Correction and Tomographic Inversion Using an Optimization-Based Approach, Zichao (Wendy) Di
- Provenance-Enabled Sample Measurements for Multi-Modal Analysis and Predictive Synthesis, Line Pouchard
- Data and Computing Needs for X-Ray Imaging at National Synchrotron Light Source II (NSLS-II), Garth Williams
- Adaptive Machine Learning for Automatic Control of Particle Accelerator Beams, Alexander Scheinker
- Current and Future Plans for the Advanced Light Source, Alexander Hexemer
- Joint Ptycho-Tomography Reconstruction Through Alternating Direction Method of Multipliers, Selin Sariaydin
- A Stream Processing Framework for High-Performance Computing, Shantenu Jha and Andre Luckow
- Designing Characterization into Materials Discovery with the Materials Project, Shyam Dwaraknath
- Software Platforms to Enable Materials Data Science at Scale, Bryce Meredig
- More with Less, Bridging Gaps from Measurement to Discovery, Robert Tang-Kong

Appendix F: Experiment Sessions

The Gap Analysis workshop focused on six key experiments of interest for materials science. For each of these experiments, a domain expert gave a talk with discussion afterwards. The following subsections summarize these talks and discussions and address the materials science discoveries enabled by the experiments, experiment setup, experiment data requirements, data analyses and associated simulations, data science workflow, and opportunities and challenges for these experiments with upgraded light sources.

1 X-ray Photon Correlation Spectroscopy (XPCS)

Claudio Mazzoli, speaker; Kristin Kleese Van Dam, moderator; Christine Sweeney, author

X-ray Photon Correlation Spectroscopy [1,2] is used to investigate system dynamics down to nanometer and atomic length scales. X-ray Photon Correlation Spectroscopy (XPCS) is an ideal tool for observing the equilibrium dynamics of atomic-scale fluctuations that occur near phase transitions, which helps with understanding of phase transition dynamics of magnetic, ferroelectric and ferroelastic materials, dynamics and local ordering of glassy materials, and nonequilibrium dynamics of soft condensed matter glassy systems.

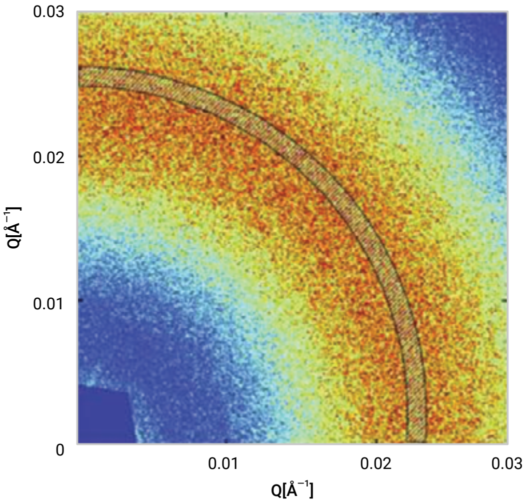


Figure A.F.1. Speckle pattern from porous silica glass (Vycor). Dark (blue) colors correspond to regions of low intensity. The shaded area (highest intensity) corresponds to a region of interest for analysis [1].

When coherent light is scattered from a disordered system, the scattering pattern presents a peculiar grainy appearance also known as speckles, as illustrated in Figure A.F.1. These speckles originate from the exact position of all scatterers within the system under investigation. XPCS characterizes the temporal fluctuations in speckle patterns. From these fluctuations, insight into the dynamic behavior of the system can be revealed.

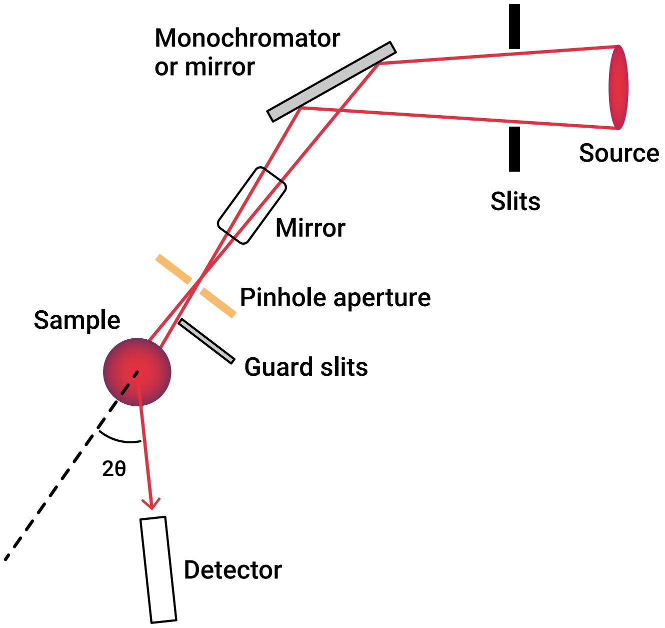


Figure A.F.2. Example setup for XPCS measurements.

The setup of a typical XPCS experiment is shown in Figure A.F.2. A perfect crystal monochromator or a short mirror are located away from the source in horizontal reflection geometry. A second mirror is installed in vertical reflection geometry downstream of the first mirror (monochromator) in order to reduce the harmonic content. Beam-selecting pinholes of different diameters are installed downstream of the second mirror.

Data rates for each XPCS experiment at National Synchrotron Light Source II (NSLS-II) at BNL typically reach tens to hundreds of gigabytes per hour and several terabytes per day [3]. Data are both structured and unstructured. In a typical experiment, 20,000 speckle patterns are recorded.

As the scatterers fluctuate over time, the resultant speckle intensity will change and time autocorrelation functions of the speckle fluctuations will reveal the scattering vector-dependent timescales associated with the motion of the scatterers. In a streaming fashion, data are analyzed via a statistical analysis of the variation of the speckle pattern as a function either of elapsed time or of an externally varied quantity, such as magnetic field, electric field or temperature. Scikit-Beam offers an XPCS data analysis package [4]. Map-reduce schemes are also being used to conduct the analysis in parallel on HPC resources [5].

Facility upgrades that increase coherent flux will enable nanosecond-resolution studies of nanometer-scale fluctuations and, potentially in certain instances, time resolution to 100 picoseconds for correlations detected with the duration of a single x-ray pulse [6]. This increase will help match the timescale of the XPCS measurements to the physics of the system. XPCS studies at higher photon energies will provide access to the interior of complex materials and to resonant edges.

1. G. Grübel, A. Madsen, A. Robert, X-Ray Photon Correlation Spectroscopy (XPCS), Soft Matter Characterization, edited by R. Borsali & R. Pecora, 953–995. Heidelberg, Springer.
2. A. Robert, Measurement of self-diffusion constant with two-dimensional X-ray photon correlation spectroscopy, J. Appl. Cryst. 40, (2007), S34–S37
3. Sameera K. Abeykoon, Meifeng Lin, Kerstin Kleese Van Dam, "Parallelizing x-ray photon correlation spectroscopy software tools using python multiprocessing," Scientific Data Summit (NYSDS) 2017 New York, 1–10, 2017.
4. Scikit-Beam. https://github.com/scikit-beam/scikit-beam.
5. Khan, F., Narayanan, S., Sersted, R., Schwarz, N. & Sandy, A. (2018). J. Synchrotron Rad. 25, 1135–1143.
6. Early Science at the Upgraded Advanced Photon Source. October 2015.

2 Resonant Inelastic X-ray Scattering (RIXS)

Diego Casa, Speaker and Author; Kristin Kleese Van Dam, Moderator

Resonant inelastic x-ray scattering focuses on correlated electron system properties of both technological and fundamental importance, i.e., superconductivity, quantum computing, exotic ground states, and others. Resonant Inelastic X-ray Scattering (RIXS) can use very small (10 micrometers) samples and extreme/in situ/in operando conditions. RIXS experiments are element- and orbital-specific. The instrument setup is mostly not a subject of the experiment; typically, the setup involves the choice of sample environment, energy resolution, and measurement parameter space.

Presently, experiment data requirements are modest, but will increase significantly and will benefit from solutions developed for other more data-intensive techniques. During the experiment, data analysis includes simple peak fitting, background subtraction, and visualizations. The pre- and post-experiment uses simulated RIXS spectra from cluster density functional theory (DFT) and coherent potential approximation (CPA) calculations.
The data science workflow is currently disjointed, per above. This dilates the turnaround time from conception to final analysis to many months. Scientists do not currently have the capacity to make data-informed decisions during the experiment other than the presence or absence of obvious spectral features.

Some opportunities with upgraded light sources are to upgrade the data infrastructure to accommodate faster readout (approximately 300 kilohertz) from multichannel detectors with on-the-fly processing. The advent of time-resolved pump-probe experiments at XFELs could possibly provide live simulation support (e.g., DFT, CPA) and statistical tests (e.g., primary component analysis [PCA]) for efficient counting times.

3 Bragg Coherent Diffraction Imaging (BCDI)

Ian Robinson, speaker and author; Richard Sandberg, moderator

Bragg coherent diffraction imaging (BCDI) provides opportunities for discovery of small-scale core-shell systems, with possibility of one phase stabilizing the other. BCDI is able to track these changes on the nanoscale and suggests opportunities for scaling up to macroscopic dimensions. There are some examples known already of oxide hidden phases living only nanoseconds and an electronically ordered excited phase of fullerene C60 lasting only femtoseconds. These are ripe for materials discovery in the time domain.

There are experimental challenges right now to reach the low temperatures where quantum materials become interesting. This is because of extreme sensitivity to vibrations. There are also challenges to perform BCDI under pressure. Here, the problem is the complex refraction properties of diamonds and beryllium when they are deformed under pressure, they turn into lenses that introduce strong near-field distortions of the x-ray optical wavefronts entering and leaving the sample. Dynamic compression at an XFEL facility may be a way to mitigate the problem.

Experiment data requirements are not too stringent for synchrotron operations because data sets are usually under (256)3 and rarely bigger than (512)3. XFEL data are a different situation because it is necessary to filter the data stream post facto. Therefore, keeping all diffraction shots is required to eliminate the bad ones, unless a simple online veto can be implemented. Data tend to be fairly sparse, so compression should be effective.

Simulation is always valuable to save time with the measurements, especially with scarce facilities like XFELs. Analysis is done through iterative phase retrieval algorithms requiring hundreds to thousands of fast Fourier transforms, so even for modest memory sizes (typically between 2563 and 5123), parallelized codes are needed for processing on multiple cores.

Complete BCDI experiments could be performed at an XFEL in one hour. We need to consider the access model in an altogether different way; once it is accepted that the throughput of an XFEL is inherently faster than the human thought process, we must simply accept that thinking during beamtime is no longer allowed. Instead of the synchrotron model of massively parallel access with up to 80 beamlines fed simultaneously, we must evolve towards operating with different levels of time multiplexing.

For materials science applications, one could do simple experiments that do not involve setup changes of the machine beyond those which can be fully automated. A standard pump–probe single-shot diffraction experiment could be performed at fixed x-ray wavelength with a fixed detector bank at a standard distance from the sample. In order to examine a wide variety of samples, the detector must span a wide enough range of Bragg angles. Rough estimates suggest that that 90 percent of experiments can be done at a 0.138 nanometer wavelength (9 kiloelectron volts) and 2 theta = 35 degrees. This follows from the lattice spacings of most materials being set by the ionic radius of common elements like oxygen. The number of pixels must be large enough to oversample the diffraction from crystals up to some maximum size, which we take to be 1micrometer. Bragg diffraction from crystals of this size will come close to saturating the dynamic range of an analog detector, which can be made as high as 105. This requires a detector of 2,500 pixels in the radial direction and as many banks as possible around the circumference of the Debye–Scherrer cone.

4 Dynamic X-Ray Diffraction at High Pressures (Dynamic XRD)

Arianna Gleason, speaker and author; Richard Sandberg, moderator

Understanding the processes that dictate physical properties in condensed matter, such as strength, elasticity, plasticity, and the kinetics of phase transformation/crystallization, requires studies at the relevant length-scales (e.g., interatomic spacing and grain size) and time-scales (e.g., phonon period). The material science discoveries enabled by ultrafast x-ray diffraction (XRD) combined with a dynamic driver include: i) phase transition kinetics, ii) phase transformation pathway, and iii) then digest the results of (i) and (ii) and contextualize it to generate a predictive model for designing materials functionality.

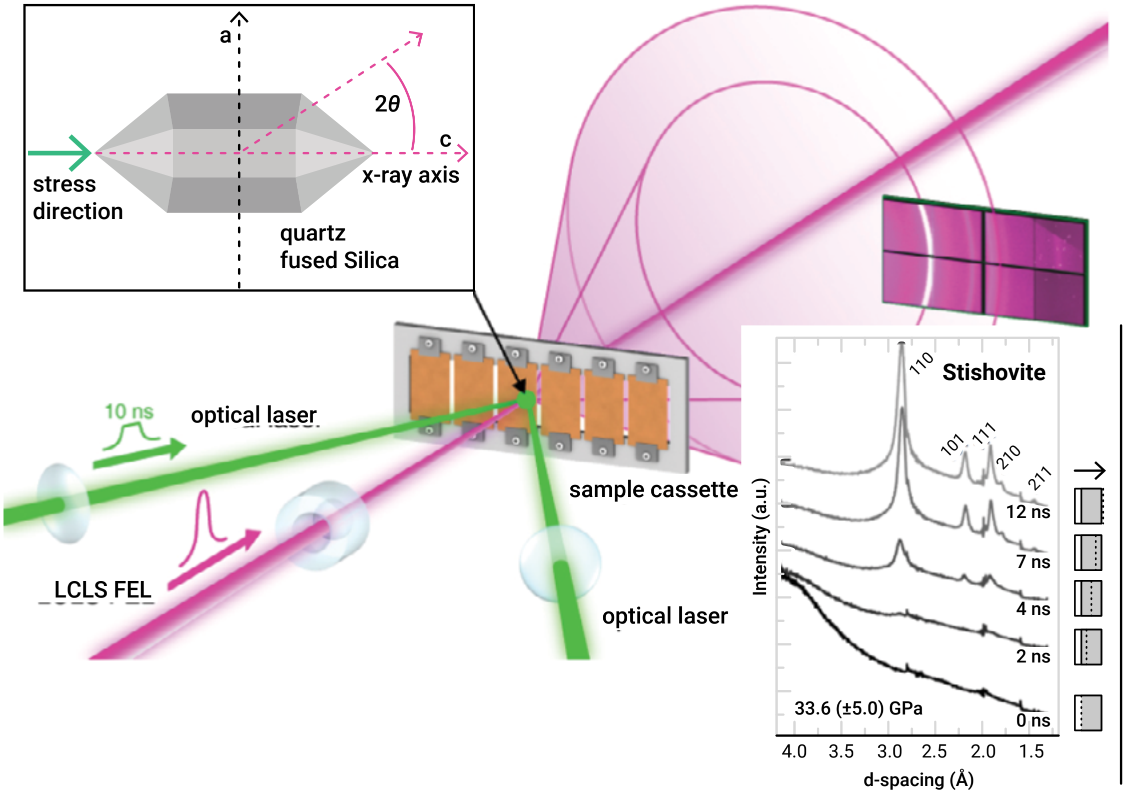


Figure A.F.3. Schematic of dynamic x-ray diffraction experiment to understand the high pressure phases of materials. This experiment was to study high pressure phases of quartz conducted at the LCLS-MEC hutch [1, 2].

Experiments performed at the Matter in Extreme Conditions end-station at the Linac Coherent Light Source, SLAC, combine a laser-driven dynamic compression pump and x-ray free electron laser (XFEL) probe to explore nonequilibrium transformation pathways and mechanisms. A schematic of these pump-probe experiments is shown in Figure A.F.3 [1, 2].

We sequentially collect XRD patterns on downstream detectors (CSPADs) during the passage of a shock wave. Each XRD pattern is collected at a different probe time with respect to compression and release waves, therefore providing time-resolved lattice-level structural information as a function of pressure and temperature. This allows us to map out pressure-temperature-time-phase for a material.

Currently, dynamic XRD data volumes at LCLS are 10-100s gigabytes per run where a run is collected in seconds to minutes. This will increase by orders of magnitude with the LCLS-II/HE upgrades. Similarly, data rates today are at 10 gigabytes per second, which will also increase by orders of magnitude with the LCLS-II/HE upgrades. An experiment duration is typically one to five shifts, where a shift is 12 hours of beamtime. Standard data format for diffraction is often in tiff format. Data fusion is absolutely needed—meaning assimilation between data types like diffraction + imaging + spectroscopy + velocimetry + real-time visualization. Prediction information and streaming analyses are also needed.

To stay at the frontier of dynamic-XRD and materials science in general, there are five tasks that should be implemented to guarantee success in materials innovation and prediction:

- Pre-experiment planning
  -Synthetic data for structure prediction; XRD integrated traces
  -X-ray source/beamline modeling
- Real-time crystallography toolkit
  -New peak/features for phase assignment; symmetry, space group
  -Phase diagram visualization
- Real-time transformation/deformation mechanism visualization
  -Twinning prediction
  -Texture prediction
  -Mosaicity prediction
- Synergy of data sets
  -XRD + spectroscopy + imaging for new materials properties correlations
- Leveraging time-resolution
  -Kinetics models (e.g., JMAK)
  -Nucleation mode tied to MD/DFT and synthetic data

One of the biggest opportunities and challenges for dynamic-XRD in the future will be detector capability/innovation. There are several components needed for cutting-edge materials science research:

- Increased frame rates: gated detectors
- Contiguous angular and azimuthal coverage
  -Less tiling (mitigate dead space)
  -Larger active area (goal: 8 centimeters by 8 centimeters)
  -Mitigate background fluctuations
- Texture analysis
- Diffuse scatter for melt/amorphous structure (amorphous-amorphous transitions; short-range order via PDF to get nearest-neighbor information)

- Increased dynamic range
- Smaller pixel size (goal: 10-25 micrometers)
  -Higher fidelity diffraction character

References

[1] A. E. Gleason et al., “Ultrafast visualization of crystallization and grain growth in shock-compressed SiO2,” Nat. Commun. 6: 8191, Sep. 2015.
[2] A. E. Gleason et al., “Time-resolved diffraction of shock-released SiO2 and diaplectic glass formation,” Nat. Commun. 8, 1: 1481, Dec. 2017.

5 High Pressure Diamond Anvil Cell X-Ray Diffraction (XRD with DAC)

Zsolt Jenei, speaker and author; Garth Williams, moderator

Many materials in both nature and technological manufacturing are produced under dynamic conditions, such as the ejecta from a meteoritic impact or the formation of a liquid metal (metallic glass) by rapid quenching. Probing these processes is therefore essential to help us understand our natural environment and to improve technologically relevant materials. The short time scales of these processes offer the possibility of significant deviation from conventional equilibrium phase studies, such as the existence of meta-stable phases. In addition, studies of pressure-induced phase transitions hold the promise of uncovering rich new physics and phenomena, such as the rate-dependent morphology previously observed in water. The dynamic diamond anvil cell (dDAC) plus third-generation synchrotron enable studies of dynamic phenomena in material occurring on the scale of few hundred microseconds (phase transition, diffusion, deformation, crystal growth).

The experimental setup builds on general-purpose high pressure beamline equipment. With the additional use of piezo-actuator driven compression, where the compression drive can be arbitrarily tailored to the users’ desire by defining the 1-10 volt driving voltage. We have the dDAC implemented at both HPCAT, Sector 16, at the Advanced Photon Source and the Extreme Conditions Beamline (ECB), P02.2 at PETRA III, DESY. ECB has the advantage of a pair of GaAs LAMBDA detectors working at a 2 kilohertz frequency; shifting these by a half period gives us an effective time resolution of 250 microsecond.

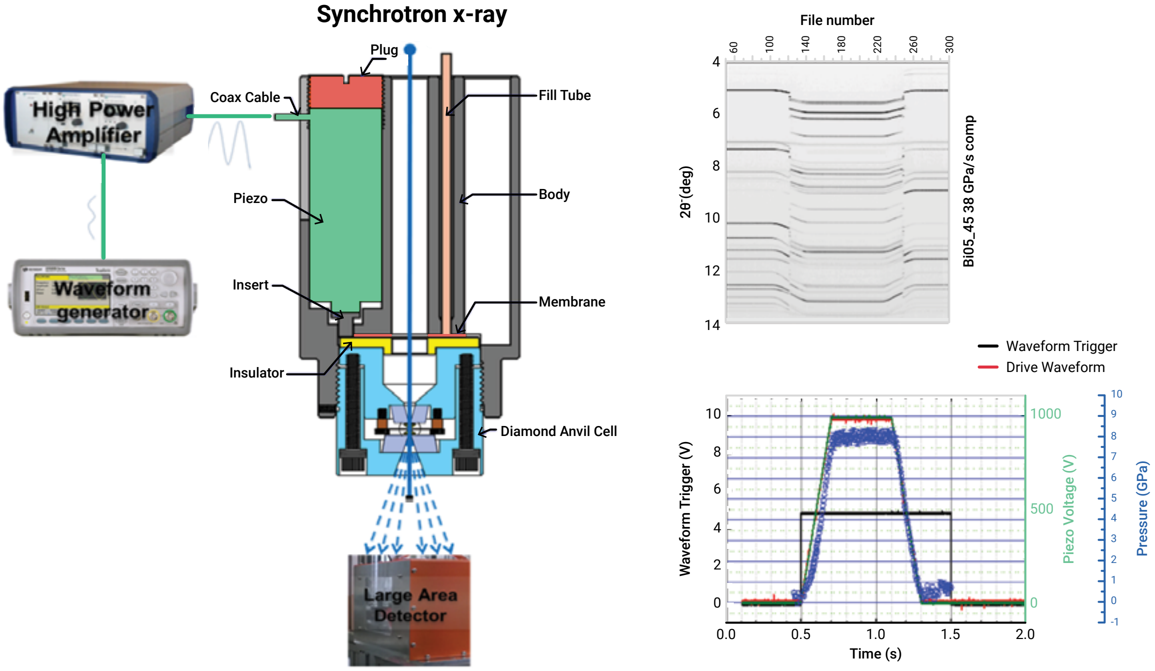


Figure A.F.4. Dynamic Diamond Anvil Cell (dDAC) setup with waveform generator, and amplifier connected to the piezo actuators driving the DAC. X-ray diffraction data are collected by either one or two large area detectors. Right panel on top shows 300 stacked integrated patterns collected in less than two seconds.

At current third generation x-ray light sources, the data acquisition rate at a typical experiment is about 6,000-10,000 diffraction files per day.

Data analysis basic steps:

- Convert nexus files to tiffs (missing software to integrate XRD patterns directly from nexus);
- Integrate the diffraction patterns;
- Fit peaks to determine pressure, crystal structure, equation of state (EOS).

As it is, currently we do not have the tools to evaluate data in real time. Preliminary analysis would have to focus on pressure evolution as a function of time (as shown in Figure A.F.4 in the bottom right panel.)

Upgraded light sources are a great opportunity to expand understanding of phase transition kinetics, with brighter, more coherent probing x-rays, faster detectors, and new techniques. However, from the standpoint of data quantity, this presents great challenges that need to be addressed to take full advantage of the presented possibilities. For example, in the very near future at the EU-XFEL HED station the XRD data acquisition rate is expected to reach 3,500 diffraction patterns per second; this will increase the amount of data by one to two orders of magnitude.

6 High Energy Diffraction Microscopy

Reeju Pokharel, speaker and author; Garth Williams, moderator

Crystallographic grains, defects, and interfaces at the mesoscale are known to influence material properties and performance of polycrystals. Third and fourth generation light sources have opened up immense opportunities for 3-D mesoscale science. High-energy x-ray diffraction microscopy (HEDM) is one such technique developed at the third generation synchrotron source that affords nondestructive characterization of 3-D polycrystalline materials at high spatial and orientation resolutions (approximately 2 microseconds and approximately 0.01 degree). A single sample can be measured and remeasured multiple times under different thermo-mechanical conditions, enabling microstructure evolution studies. Such data sets are invaluable for dynamic mesoscale model development and validation.

HEDM experiments are performed in a transmission geometry, where a monochromatic x-ray beam illuminates a cross-section or a volume of the sample and also sets of grains that satisfy the Bragg condition produce a diffracted beam on a flat panel detector located at 5-10 millimeters (near-field setting) or 1-2 meters (far-field setting) away from the sample. The sample is rotated about the axis perpendicular to the incoming beam and diffraction data are collected to ensure diffraction from all the grains in the illuminated volume are recorded. This is crucial for accurate data inversion.

Extracting 3-D microstructure information from a measured diffraction pattern is an inverse problem. Like any inverse problem, it is highly under-determined. Therefore, usually a large amount of redundant data from multiple rotation angles are required to accurately reconstruct microstructure from diffraction patterns.

HEDM data inversion entails determining crystal orientations of grains in a polycrystalline sample that gave rise to the observed diffraction spots on the 2-D detector. Current approach performs forward diffraction simulations of the experiment and a brute force search is conducted to determine crystal orientations and locations that produce the diffraction images that best match the observed patterns. Instrument geometry of the experiment needs to be known before performing the diffraction simulations of the sample. Therefore, a calibration sample, usually a single crystal, is used for determining the instrument parameters.

A major limitation faced by HEDM is that the reconstruction of 3-D material microstructures from diffraction data is an extremely lengthy process, sometimes requiring up to a week on many thousand core computer clusters for a single sample. This highly limits the possibility of any feedback during a typical beam line experiment, which would be very useful for steering and driving experiments, especially under dynamic conditions. This problem will only exacerbate with increasing data collection rates with upgraded light sources. Therefore, there is a growing need for rapidly analyzing diffraction data during in situ or ex situ experiments.

Appendix G: Data Science Techniques

The Gap Analysis workshop focused on eight key data science techniques of interest for light source experiments involving materials science. For each of these techniques, an expert gave a talk with discussion afterwards. The following subsections describe the materials science discoveries enabled by data science techniques, details of the technique, technique requirements, experiments and associated simulations using the technique, the workflow that uses the technique, and opportunities and challenges for these techniques with upgraded light sources.

1 Pre-Experiment Machine Learning

Ryan Coffee, speaker and author; Alexander Sheinker, moderator

One of the opportunities that high rate sources and detectors unlock is the ability to rapidly scan multidimensional parameter spaces. The dimensionality is growing for both synchrotron sources as spectral imaging, for example, and for FEL sources as dynamics imaging. Allowing users to visualize in real time how their particular representation of the physical observable is being explored will enable not only data compression to be tailored to the information the user actually wants, but it will also allow her or him to make decisions about what regions of parameter space require more statistics.

Techniques that represent the error landscape could make suggestions to users regarding regions of parameter exploration.

Container image repositories that are consistent across user facilities and simulation hosting leadership scale computing facilities. Actually, given the new user facility model for the Large Synoptic Survey Telescope (LSST) (see Carlo discussion), could the light sources encourage a consistent hashing of simulations that could be also added to the related experimental metadata?

Simulations could also be served via container and use a similar image ID hash to track simulation provenance as envisioned for the facility experiment data.

The workflow would be that the user has sample data or simulation data provided by the facility or community HPC cloud. He or she builds a Docker container on a home institution machine from an image sourced from the facility image repository that connects to the sample data. That image is the sandbox image used for exploring with simulation or sample data which interesting data representations are showing the effect the user is hoping to see. The user is then encouraged to port the sandbox model (via git or another system) to a technique-specific container, e.g., lcls-tmo/electron_spectroscopy-latest. This image then embeds its image ID hash into the model that the user develops inside the container. When the user submits the proposal, he or she could include an image that demonstrates how that model sufficiently transforms the sample raw data into compressed and physically meaningful pre-processed data. This also shows how the proposed level of uncompressed raw data can be used to validate the model while staying within the data rate budget of the facility. If the proposal is granted time, then the model container will be trimmed into a lite model and receive a new image hash and be compiled against the beamline computer hardware or data reduction stack. When the experiment runs, the user provided models, or stock facility models, will have their hashes added to the metadata that follows all raw and reduced data files produced in the experiment.

The opportunity with upgraded sources and detectors is that users can place the statistics of shot accumulation into meaningful regions of parameter space. This will happen since the user will have a more refined picture of the information contained in the data and will more densely store that information along with the algorithms that were used to compress the raw data.

2 Pre-Experiment Modeling and Simulation

Turab Lookman, speaker and author; Thomas Proffen, moderator

As experimental facilities become more complex and offer deeper insight into what makes materials work, a change from a cook-and-look approach still used at many scattering facilities needs to make room for hypothesis-driven experiments. This will require the ability to create and run materials models (e.g., atomistic, first principles, and others) as well as being able to evaluate the signal expected from the experimental instrument selected. This will not only allow researchers to demonstrate feasibility of an experiment, but also allow easier real-time monitoring during the experiment.

Pre-experiment needs will also extend to experiment planning, including possibly suggesting materials to be investigated based on gaps in data sets used for machine learning related to a desired application, for example.

Details of pre-experiment modeling and simulation will depend on the science area and particular instrument, good materials, and instrument models with well-established ranges of applicability. In other words, you need to know where the models work and do not work.

Requirements will depend on the science area, instrument details, and the level of data science/technique expertise expected from the user and/or available from the instrument team. The added complexity of pre-experiment is the fact that access might be needed before the users are on site and able to interact in person with the beamline team. Some access to pre-experiment capabilities will also be needed as part of experiment planning, in other words, before an approved proposal exists.

The challenge for simulations might be matching the fidelity and modeling approach to each proposed experiment. However, this is not decoupled from post-experiment modeling needs.

3 In-Experiment Visual Analytics and Decision Making

James Ahrens, speaker; Peer-Timo Bremer, moderator and author

As discussed above, we expect the data rates to increase drastically with new facilities and capabilities. While that represents fantastic opportunities for new discoveries it also constitutes the danger of wasting a significant amount of resources through problems with experiment setup, suboptimal configurations, and more. In the past, these kinds of problems could often be corrected on the fly, for example by manually checking the first couple of shots and adjusting the experimental setup accordingly. However, with the increased data rates these types of check are becoming increasingly difficult or even impossible. Instead, we need a new approach that enables near real-time autonomous and/or user-guided decision making from massive data streams. Both the autonomous and human-in-loop approaches will likely share a joint data management framework capable of providing near real-time access to the experimental data as well as sufficient computing resources. The difference will be that the autonomous branch will rely on pre-determined models and statistics to automatically adjust the parameter within some pre-defined set of possibilities. The human-in-loop approach on the other hand will provide more generic analysis capabilities to the experts, ideally allowing a wide variety of analysis approaches to be executed on the fly. These will aim for the truly unexpected occurrences in which an experiment produces unforeseen results and/or problems. For the latter, visualization and/or visual analytics approaches will be crucial to enable both data exploration as well as to provide intuitive insights into otherwise black-box models.

The necessary techniques to enable decision making, automatic or otherwise, will include a wide variety of statistical, feature extraction, and machine learning approaches, in particular those applicable to complex high-dimensional data. For the automatic analysis, the focus will be on techniques able to ingest large quantities of data quickly as well as those that provide uncertainty bounds to confidently drive an autonomous system. On the other hand, the human-in-the-loop approaches will rely on interactive visual analytics to quickly explore data as it is being collected. The goal will be to maximally support the intuitions and prior knowledge of expert users in adjusting a running experiment. Both branches of decision making will rely on an underlying data management and analysis framework that can quickly and flexibly deliver subsets of data to the analysis and/or visualization pipeline.

The key challenges will be the size of the data as well as the black-box nature of many current analysis routines. The size of the data will stress not only the underlying storage infrastructure, but also will overwhelm virtually all existing analysis and visualization approaches. For example, few interactive visualizations will scale beyond tens of thousands of samples, neither in terms of computational power nor in terms of their visual encoding. This means we are not able to interactively draw and select from a scatter plot with millions of points, nor will such a plot show anything past an uninterpretable, dense set of points. Similar problems exist with many advanced models, i.e., fitting traditional Gaussian Process models to millions of points is not feasible.

Another, often overlooked challenge in decision making is the black-box nature of the underlying models. In order to adjust an expensive and carefully planned experiment on-the-fly users will require a high confidence in whatever model or predictor they are being offered. In post-processing, such confidence often comes from careful re-analysis, cross-validation, and others. However, this type of analysis is not feasible for real-time decisions with partial data. Few users will comfortably adjust parameters based on a computational model that appears opaque and to some extent arbitrary. Instead, we need more interpretable models and ideally, visualization to provide users with the confidence to make decisions. This will include explicit representations of the uncertainty in the relevant models as well as new visual representations of otherwise abstract data, i.e., more interpretable dimension reductions, topologically accurate representations, and more.

4 In-Experiment Statistics and Emulation

Devin Francom, speaker and author; Earl Lawrence, moderator

Understanding some of the more complex aspects of an experiment via simulation is typical in modern science, including at advanced user light sources. Experimental results are used to assess and validate simulators that encode current scientific understanding. Combining simulated and experimental data can be powerful, but requires careful treatment of uncertainty. Statistical approaches to solving inverse problems (i.e., inferring simulator parameters based on experimental data) aim to characterize uncertainty via probability. Bayesian statistical methods allow for seamless propagation of uncertainty. These approaches can be used for both cheap and expensive simulators, though an expensive simulator often necessitates the building of a fast statistical surrogate for the simulator, called an emulator. Emulators need to be fast to evaluate, accurate for any combination of possible simulator input, and sufficiently flexible to capture complexities of the simulator. In many cases, a good emulator can be built based on hundreds to thousands of simulations.

Emulators are often built on reduced dimension summaries of high dimensional simulator output. When experimental data can be similarly reduced, inverse problem inference can take place in a reduced dimension space. Dimension reduction is typically a problem-specific bottleneck in this case. A second bottleneck is the usual approach to inference, which is based on Markov chain Monte Carlo (MCMC). This approach can be too slow when the analysis results are intended for use in planning the next experiment. This sequential design problem is a third area of interest at these experimental facilities.

One in-experiment question we are trying to answer is what the optimal next experiment should be (sequential experimental design). Some other examples of statistics activities in-experiment are dimension reduction, Gaussian process modeling for emulation, MCMC, and sequential Monte Carlo for posterior sampling when solving the inverse problem, sequential experimental design. The in-experiment problem can require nearly real-time solving of inverse problems. Although it is important to answer the science questions, it is not as important in-experiment. Scientists are highly interested in determining if the data are bad in-experiment and steering the experiment parameters so that they get the data they need for scientific discovery, even if the discovery is later.

Statistical methods and emulation are also useful in the pre-experiment and post-experiment stages. For instance, an emulator can be very useful pre-experiment for performing sensitivity analysis. Solving inverse problems is typically a post-experiment problem.

5 Post-Experiment Physics and Math in Data Analysis

Stephan Hruszkewycz, speaker; Jeffrey Donatelli, moderator and author

With upcoming increases in energy and coherence, light sources will enable a number of new diffraction experiments that will require a host of new physics to model, as well as new advanced algorithms to solve the associated inverse problem of determining material properties from the experimental data. Furthermore, in many cases, these problems can become highly ill-posed, and thus additional physical constraints will need to be incorporated in order to uniquely and stably solve the inverse problem.

Many of the inverse problems described in the previous section can be described as extensions of the classical phase problem in coherent diffractive imaging (CDI). The most common way to solve the phase problem in CDI is through iterative projection methods, which requires one to derive computationally efficient operators to project model quantities to be consistent with various constraints in the inverse problem. Several extensions of these traditional iterative projection methods have been developed to target more complex inversion problems, e.g., multi-tiered iterative phasing, and model various uncertainties in specific experiments, such as angular uncertainties and pixel binning/blurring in Bragg coherent diffractive imaging.

Alternative solutions to the inverse problem may also include gradient-based methods, automatic differentiation, and black-box approaches. As the forward problem associated to the experiment becomes more complicated, these alternatives may become more appealing as they bypass the need to derive the mathematically complex projection operators that may be needed in the iterative projection approach.

The iterative projection methods will require that the inverse problem has an appropriate mathematical decomposition that lends itself to the development of computationally efficient projection operations that constrain the reconstructed model to be consistent with the data and satisfy any appropriate physical constraints.

The gradient-based approach requires that the gradient information of the forward model can be efficiently computed. The automatic differentiation route does not require the analytic derivation of the gradient, which is instead automatically computed if the forward model can be expressed in terms of a series of elementary arithmetic operations and elementary functions. The black-box approach only requires that the forward model can be computed, but may need much more computational resources in order to arrive at a solution.

One example of an experiment that is using physics and math is that of Bragg coherent diffractive imaging (BCDI), where one is able to use the data to deduce the strain within a crystal.

In order to tackle the potential ill-posed nature of this new class of complex inverse problems, it may become necessary to incorporate additional physical constraints on the reconstructed models. This will require both physical insight to deduce what these constraints may be and new mathematics to efficiently express and impose these constraints on the physical model.

As the inverse problems become more complicated, as we try to reconstruct more complex models, and as we add new physical constraints to the system, carefully validating the solution becomes crucial, but also more challenging. New validation methods will need to be designed to ensure that the calculated results are real and implied by the data, and not simply the result of over constraining the system with model constraints.

It will be critical to determine whether the best route to solving the inverse problems from these new experiments is to focus on new fundamental mathematics, efficiently utilizing new hardware, incorporating new physical models, or all of the above. To this end, it will be important to establish a dialogue between domain scientists, mathematicians, computer scientists, and physicists in order to determine the best path forward.

6 Post-Experiment Machine Learning

Daniella Ushizima, speaker and author; Aric Hagberg, moderator

From industry to national laboratories, shape and structural properties of new compounds imaged through advanced instrumentation at light sources are used to measure the function and resilience of new materials. Advances in imaging for the design and investigation of materials have been remarkable. As an example, the growth of x-ray brilliance was 18 orders of magnitude in 5 decades, and extremely quick snapshots have enabled description of dynamic systems at the atomic scale. What drastically changed is the frequency with which these data modality are collected and used as a key scientific record, which is unprecedented. One of the main challenges is how to couple increasing data rate experiments to new Machine Learning methods in support of more automated analytical tasks for scientific discovery.

Recent efforts in machine learning applied to data representation and structural fingerprints have streamlined sample sorting and ranking, including the identification of special materials configurations from large databases. Methods such as convolutional neural networks have allowed automated characterization of abstract pictures, such as scattering patterns, based on prototypes stipulated by experts, or simulated at leadership computing facilities. Such characterizations or signatures enable image similarity search with real-time feedback in million-size image collections.

The ability to survey samples more broadly allied to computational algorithms to compare millions of samples not only offers unique opportunities for deeper scientific interpretation of experiments, but also imposes hurdles, such as availability of storage, data transfers, large memory footprint, and intensive computation. Other key requirements that must be included in future strategic planning for post-experimental data:

- Policies for data generated at national user facilities regarding representation, sharing, and data storage;
- Policies and/or rules for raw private data storage in contrast to compressed public data sets.

One example of capability is the use of content-based image retrieval software tools that allow users to search for scientific images in a faster and more intuitive way. Also known as recommendation systems, such categorization tools can accelerate retrieval, inspection, and curation of scientific images produced at light sources:

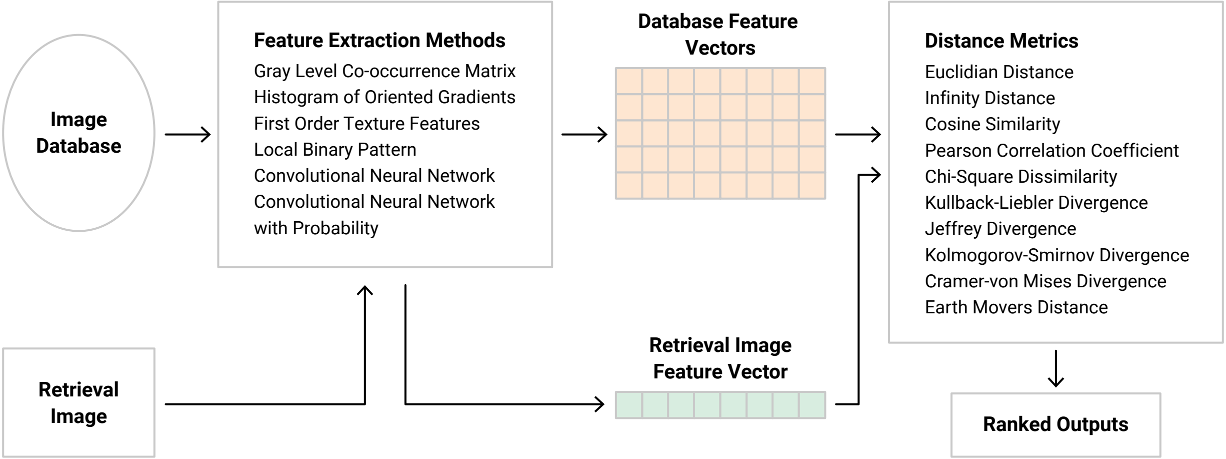
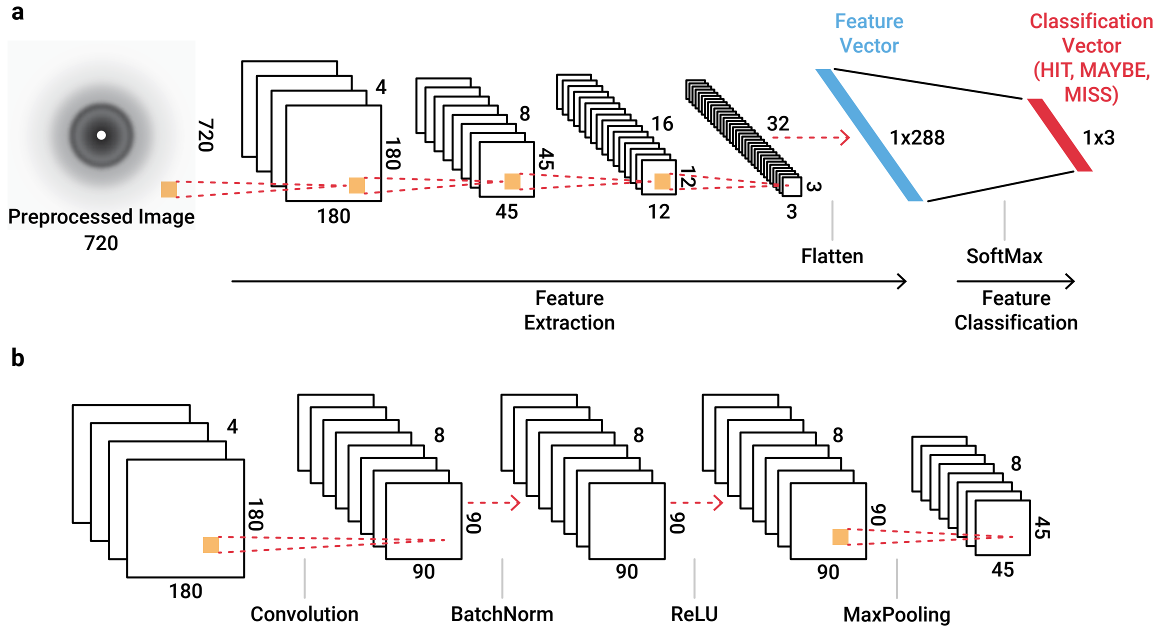


Image Database
Feature Extraction Methods
Gray Level Co-occurrence Matrix
Histogram of Oriented Gradients
First Order Texture Features
Local Binary Pattern
Convolutional Neural Network
Convolutional Neural Network with Probability
Database Feature Vectors
Distance Metrics
Euclidian Distance
Infinity Distance
Cosine Similarity
Pearson Correlation Coefficient
Chi-Square Dissimilarity
Kullback-Liebler Divergence
Jeffrey Divergence
Kolmogorov-Smirnov Divergence
Cramer-von Mises Divergence
Earth Movers Distance
Retrieval Image
Retrieval Image Feature Vector
Ranked Outputs

Figure A.G.1. Example of scientific image recommendation system: pyCBIR is a Python search engine funded by DOE ASCR to solve DOE imaging facilities' problems [1].

Data compression post experiment supported by DOE ASCR in the service of DOE imaging facilities: [2]



The proposed CNN for screening X-ray images consists of four sets of convolution/batch normalization/rectification (ReLU)/ downsampling (MaxPooling) layers. (a) Numbers and the sizes of the output feature response maps produced by each of the filter layers. The resulting feature map is concatenated (flattened) into a one-dimensional feature vector, and fed into the feature classification stage. (b) Details of the particular sequence of filters that is responsible for the transformation from the 4x180x180 image stack to the 8x45x45 stack.

There are opportunities and challenges with upgraded light sources, including:

- What will happen when in situ processing is not possible or limited?
- What will raw data storage of 140 petabytes per year cost?
- Scientists using machine learning need to understand the weight of their influence and limitations of their wisdom.

References

[1] Araújo, Flávio Henrique & Silva, Romuere & Medeiros, F.N.s & D. Parkinson, Dilworth & Hexemer, Alexander & M. Carneiro, Claudia & Ushizima, Daniela. (2018). “Reverse image search for scientific data within and beyond the visible spectrum.” Expert Systems with Applications. 109. 35-48. 10.1016/j.eswa.2018.05.015.
[2] Ke TW, Brewster AS, Yu SX, Ushizima D, Yang C, Sauter NK. “A convolutional neural network-based screening tool for x-ray serial crystallography.” J Synchrotron Radiat. 2018 May;25(Pt 3) 655-670. doi:10.1107/S1600577518004873. PMID: 29714177; PMCID: PMC5929353.

7 Materials Database

Logan Ward, speaker and author; Bryce Meredig, moderator

Materials data infrastructure (MDI) (see TMS Report “Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering”) and databases of past experimental and simulation results have the potential to supply critical input data to data science methods. For example, a database of thousands of x-ray diffraction measurements conducted on different compounds would enable a machine learning-based prediction of the results of such a measurement on a new compound. While several notable databases of materials characterization results and simulations exist (e.g., CSD, ICDD, OQMD, materials project), most newly generated research data are not made available to the community for re-use, representing a missed opportunity for data science.

MDI has three key components:

- software (the operating system of the MDI);
- computer hardware; and
- data sets.

To take the National Institute of Standards and Technology (NIST)-funded Materials Data Facility (MDF) as an example, the enabling software is built on Globus, a domain-agnostic research data management platform. MDF uses Amazon Web Services and the National Center for Supercomputing Applications (NCSA) for computation and for data storage. Finally, the data sets available on MDF are sourced from other publicly available platforms (e.g., Crystallographic Open Database), as well as contributors from the research community.

Meeting the aforementioned software and compute requirements involves making design choices aimed at specific research use cases. Unsurprisingly, the various MDI platforms in use today have made different design choices to optimize for often-distinct research goals. For example, Bluesky at BNL has beamline applications in mind. MDF, thanks to its Globus roots, is well suited for storing and transferring very large research data sets across different materials research areas. Citrine Informatics' Open Citrination platform is designed to enable users to efficiently train machine learning models on their data and then implement a sequential learning workflow. Given that the various MDI platforms have unique strengths and weaknesses, we see an important opportunity for these platforms to move toward interoperability so that end users can easily access the benefits of multiple platforms.

While the specific software and hardware requirements of a materials database vary depending on its scale of usage, the requirements are significant for every application. Developing a usable database with a functional user interface and application programming interface (API) requires a skillset uncommon for scientists and months to years of effort, along with potentially millions of dollars of investment. Hardware requirements for databases do benefit from the economies of scale for cloud computing, and the adoption of cloud resources by academics has become evident. Even after the construction of a successful database, adapting the website to new requirements and maintaining the underlying systems is also an effort that should not be overlooked. In short, creating and running any materials database is an effort that requires long-term investment on behalf of a dedicated staff.

There are two main use models for materials databases: publishing/organizing data, and using existing data for new science. The first step in either model is finding the appropriate database for particular data. For publication, the main effort is ensuring the data are usable by others by adding metadata, documenting files and formats, and more. Many databases offer tools for aiding this process (e.g., schemas and parsers for common file types), and there are now journals that focus on describing published data (e.g., Scientific Data and Data in Brief). Using published data involves first finding the data and then shaping it into a format that works with the desired analysis tools, the focus of work by R. Seshadri and T. Sparks. Databases that further facilitate analysis by providing APIs are currently rare, but those that do have been employed frequently in data-driven materials efforts, such as matminer Python library and the Materials Project.

There is a rich set of opportunities and challenges for databases with current and future light sources. The foremost is providing training suitable for the diverse and often novice user base at beamlines. Education is a first and critical step for adoption. Adoption of data publication has its own well-known challenges due to the limited incentives and high barriers for publishing open data. The distributed, federated nature of beamlines coupled with increasingly large data sets will require distributed and scalable infrastructure and complicate the data discovery challenge. A final challenge is financially sustaining MDI, which is complicated by the need for specialized staff, uncertainty regarding long-term funding models for software, and costs that scale with usage.

8 Data Management

Daniel Allan, speaker and author; Amedeo Perazzo, moderator

As light sources, and user facilities in general, manage a great volume and variety of materials science data, it is important to remove friction around metadata capture, data management, and data analysis. Scientists must be able to easily connect with existing open-source tools and standards.

The BlueSky Project (developed as a collaboration between multiple facilities including NSLS-II, LCLS-II, and APS-U) supports data management workflows that span across beamlines and computing facilities to provide seamless remote access of user data and analysis resources for post-experiment data analysis. It achieves this by emphasizing data interfaces independent of on-disk formats, working in a common language (Python), and sharing tools wherever practical.

Interfacility data management requires common interfaces (ranging from APIs to data) and streamlined authentication (e.g., avoid multiple hops and facility-specific protocols). Additionally, facilities should adopt common data lifecycle models. The specifics may vary from facility to facility, but there ought to be common models. The model should include the ability to store the data at the facility for some guaranteed amount of time (the actual time is facility dependent); the ability, if desired, to easily move data off site at a high rate over, for example, ESnet; the ability to surge, seamlessly as possible, to HPC if required; and the ability to easily create and automatically submit to analysis pipelines (or tweak existing ones).

The BlueSky Project targets experimental science in all domains from the lab scale to the facility scale. It has been used at nearly all beamlines at NSLS-II and has growing adoption at LCLS-II and APS-U. Prominent examples include a live tomographic reconstruction experiment at APS, which leverages the streaming-friendly architecture that BlueSky propounds.

BlueSky captures experimental metadata, sample metadata, information about scientific intent, and bureaucratic information. It encodes all of this alongside the data into a specified but flexible schema and makes it available behind a programmatic interface. The schema can easily be composed with any existing standards. The data can be accessed programmatically or exported to any desired format.

As light sources become brighter and detectors become larger and faster, light sources are generating greater data velocity and volume. This exposes the data variety problem at user facilities, which stands out compared to other fields, such as astronomy or climate science. User facilities manage a large and changing collection of instruments. Techniques straddle a wide span of data rates, structures, and access patterns, and they employ a mix of well-established data processing procedures and original, improvised techniques.

The expanding open-source scientific software ecosystem, particularly in Python, presents a good opportunity to collaborate between facilities and across domains to address this problem.

Appendix H: Summary Table of Experiment Requirements

Area	Scattering	Spectroscopy	Imaging	Scattering	Scattering	Imaging
Name	X-ray photon correlation spectroscopy (XPCS)	Resonant inelastic x-ray scattering (RIXS)	Bragg coherent diffraction imaging (BCDI)	Dynamic x-ray diffraction at high pressures (dynamic XRD)	High pressure diamond anvil cell x-ray diffraction (XRD with DAC)	High energy diffraction microscopy (HEDM)
Acronym	XPCS	RIXS/IXS	BCDI	Dynamic XRD	HPXRD	HEDM
Purpose	To study fluctuations in materials structure (phase, domains, impurities) as they move around.	To study solid state perturbations due to vibrations, magnetic ordering (phonons, magnons, and others).	To conduct nanometer scale imaging of strain inside materials to understand damage, chemical reactions, and more.	To study dynamic high-pres sure phases and shocked states of materials.	To study high-pressure phases of materials	To image solid (polycrystalline) materials under strain to study damage.
Synchrotron/ XFEL	Both	Both	Both, time resolved stroboscopic at XFEL	Both, mostly XFEL	Synchrotron, will come to EuXFEL	Synchrotron almost exclusively
Experiment Duration	Very fast (single shot at XFEL) to hours	Hours to days (long integrations, low signal)	Minutes to hours	Single shot (XFEL)	Fast (seconds to minutes)	Hours to days
Data Types	Detector images	Mostly 1-D spectra in time series (1-D rows)	Detector images	Detector images	Detector images	Detector images
Data Rate	100 hertz currently, up to 10 kilohertz at LCLS-II	Slow, minutes to hours per 1-D spectra	120 hertz (XFEL) to a few seconds	Currently 0.1 hertz to 10 hertz at EuXFEL	100 hertz currently, up to 10 kilohertz at LCLS-II	10–1 hertz
Data Accuracy	Accurate if fitting is done correctly	Accurate	Accurate if oversampling is good	Accurate if detector calibration is known	Accurate if detector calibration is known	Accurate if data density is good
Data Fusion	Not usually	Maybe	Not usually; can be combined with XRF	Yes, if combined with imaging	Not usually	Yes, combined with XCT
Time to Solution Needed	Days to months afterwards; lots of analysis	Real-time interpretation	Need iterative phase retrieval, typically a minimum of tens of minutes to days	Can be real time if orientation is figured out	Can be real time if orientation is figured out	Need lengthy reconstruction; days to months
Prediction Needed to Proceed	No, but helpful	Yes, or data are confusing	No, but helpful	Yes, to understand phases and Rietveld Refinement	Yes, to understand phases and Rietveld Refinement	Yes, if wanted to understand phases
Associated Simulation	Materials dynamics (finite element or molecular dynamics (MD))	Yes, materials band structure modeling, finite element, or MD	Materials dynamics (finite element, continuum, or MD). MD can be a very powerful way to couple to iterative phase retrieval.	Materials dynamics (finite element, continuum, or MD)	Materials dynamics (finite element, continuum. or MD)	Materials dynamics (finite element, continuum, or MD)
Typical Analysis Needed	Complicated: normalization of frame, compute intensity correlation pixel by pixel with time-delayed frame typically over a region of interest then compute g2 and fit to decorrelation	Normalization/ background subtraction only	Data processing and iterative phase retrieval	Detector orientation and Rietveld Refinement	Detector orientation and Rietveld Refinement	HEDM analysis (iterative fitting of orientations back into sample plane and some sort of tessellation to fill in spaces between; sometimes near field grain shape seeding)

Appendix I: Summary of Prior Workshop Reports

Many workshops, symposia, and conferences have been held and a number of reports have been written covering various areas of materials science, data science, and computing for experiments at advanced user light sources. With the large number and diversity of light source experiments, the number of data analytics challenges is very large.

Fifteen workshop reports were reviewed. They each covered a subset of five categories (data science, computing systems, simulation, materials science, and light sources). Although computing systems is not a big focus of our workshop, it seemed appropriate to note the reports covering this area.

One of the reports covered all areas to some degree:

- "Challenges at the Frontier of Matter and Energy: Transformative Opportunities for Discovery Science," 2015 Report from the Basic Energy Sciences Advisory Committee (BESAC) to the U.S. Department of Energy Office of Science. [1]

A few reports covered data science and light sources:

- "Basic Research Needs for Innovation and Discovery of Transformative Experimental Tools;" [2]
- "BES Exascale Requirements Review;" [3]
- "Data and Communications in BES Creating a Pathway for Scientific Discovery." [4]

General requirements for computing plus data science relating to experimental science was comprehensively covered by:

- "Management, Analysis and Visualization of Experimental and Observational Data: The Convergence of Data and Computing." [5]

Three sections on light source facilities feature details of beamlines and science use cases and list their computing impediments, gaps, needs, and challenges. These sections are very useful resources for our workshop.

Data science as it relates to materials is thoroughly covered in:

- "ASM Materials Data Analytics: A Pathfinding workshop;" [6]
- "Workshop on Artificial Intelligence Applied to Materials Discovery and Design;" [7]
- "Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering;" [8]
- Also, note the recent workshop (no report): "ICFA Beam Dynamics Mini-Workshop: Machine Learning Applications for Particle Accelerators." [16]

It is notable that DOE has funded most of the above highlighted reports and it is very encouraging to see that a number of organizations are attuned to this research area and are communicating the challenges and opportunities well through the focus of each of the workshops. However, the reports do not have a focused look at the nexus of data science, materials science, and light source experiments, which is where our workshop lies.

See below the table of workshop reports listing highlights as they relate to this workshop and listing areas of topic coverage (data science, computing systems, simulation, materials science, and light sources).

Report Name	Year	Sponsor	Data Science	Computing Systems	Simulation	Materials Science	Light Sources
"Basic Research Needs for Innovation and Discovery of Transformative Experimental Tools" [2]	2006	DOE Office of Science	X	X			X
"BES Exascale Requirements Review" [3]	2015	DOE ASCR and BES	X	X			X
"Future Platform Workshop Report" [9]	2017	DOE ASCR and BES	X	X			X
"ASM Materials Data Analytics: A Pathfinding Workshop" [6]	2015	ASM International's Computational Materials Network and Ohio State University and funded by NIST	X	X		X	
"Workshop on Artificial Intelligence Applied to Materials Discovery and Design" [7]	2017	DOE	X			X	
"BES Roundtable Opportunities for Basic Research at the Frontiers of XFEL Ultrafast Science" [10]	2017	BES			X	X	X
"Challenges at the Frontier of Matter and Energy: Transformative Opportunities for Discovery Science" [1]	2015	BES advisory committee	X	X	X	X	X
"Computational Materials Science and Chemistry: Accelerating Discovery and Innovation through Simulation-based Engineering and Science" [11]	2010	DOE		X	X	X	
"Basic Research Needs for Materials Under Extreme Environments" [12]	2007	BES			X	X	X
"Data and Communications in BES Creating a Pathway for Scientific Discovery" [4]	2012	BES	X	X	X		X
"New Research Opportunities in Dynamic Compression Science" [13]	2012	WSU Institute for Shock Physics			X	X	X
"Next Generation Photon Sources for Grand Challenges in Science and Energy" [14]	2009	BES				X	X

Report Name	Year	Sponsor	Data Science	Computing Systems	Simulation	Materials Science	Light Sources
“Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering” [8]	2017	TMS	X	X		X	
“Opportunities for New X-Ray Sources to Shed Light on New Functional Materials” [15]	2016	LANL				X	X
“Management, Analysis and Visualization of Experimental and Observational Data: The Convergence of Data and Computing” [5]	2015	ASCR/DOE	X	X			X

Report Name	Highlight
“Basic Research Needs for Innovation and Discovery of Transformative Experimental Tools” [2]	Identifies Priority Research Direction 4 as integration of experiment, theory and computation (real-time steering, multimodal data analysis, and integration of simulation)
“BES Exascale Requirements Review” [3]	Section 3.6 covers computing and data challenges at BES facilities (streaming analysis, multimodal analysis of results from different instruments, data curation, accelerator simulation). The focus is on exascale.
“Future Platform Workshop Report” [9]	Focus on future platforms (systems, storage, network, resource management, data and frameworks) for applications. Light sources are one application area (APS) out of the four covered. Data breakout summary was short; mentions machine learning, data representation, decision making, reproducible analysis, hypothesis creation.
“ASM Materials Data Analytics: A Pathfinding Workshop” [6]	Materials data analytics focus with priority areas: uncertainty, data sharing, multiscale optimization, decision support, extracting info from publications.
“Workshop on Artificial Intelligence Applied to Materials Discovery and Design” [7]	Priority areas: Common data formatting, integrating multiscale models. R&D pathways, data availability, data management (DM), database management, uncertainty quantification, validity of models, connections between models, artificial intelligence, data extraction, data fusion, materials discovery.
“BES Roundtable Opportunities for Basic Research at the Frontiers of XFEL Ultrafast Science” [10]	Focus on three priority research opportunities (electron motion within molecule, novel quantum phases, and rare events and intermediate states). Cross-cutting opportunities are multimodal ultrafast measurements and advances in theory of dynamical processes far from equilibrium.
“Challenges at the Frontier of Matter and Energy: Transformative Opportunities for Discovery Science” [1]	Very high level. Chapter 5 is “Advances in Modeling, Math, Algorithms, Data and Computing, Inverse Problems.” Chapter 6: Advances in Imaging Capabilities Across Scales.
“Computational Materials Science and Chemistry: Accelerating Discovery and Innovation through Simulation-based Engineering and Science” [11]	Mostly covers materials science and chemistry research, but has two subsections on simulation and its role in accelerating the development of materials and chemical processes. Recommends an integration of experimental capabilities with theoretical and computational modeling.
“Basic Research Needs for Materials Under Extreme Environments” [12]	Focus is primarily on materials research. One cross-cutting research theme is taking advantage of predictive theory and simulation to design and predict the properties and performance of new materials required for extreme environments.
“Data and Communications in BES Creating a Pathway for Scientific Discovery” [4]	Link experimental user facility needs with advances in data analysis and communications. Integrate theory and analysis, move analysis closer to experiment, match DM with capabilities of detectors.
“New Research Opportunities in Dynamic Compression Science” [13]	Highlights simulation and computation as an important parallel area, key to the success of dynamic compression science. “Extending the time scales of such simulations to experimentally observable processes in materials is generally challenging... An important overarching goal ... is to perform experiments on the time and length scales of numerical simulations and to bridge this knowledge gap.”

Report Name	Highlight
“Next Generation Photon Sources for Grand Challenges in Science and Energy” [14]	Scientific descriptions of photon science drivers, descriptions of facility capabilities, cross-cutting challenges (including imaging). Emphasis on how to control experiments to do science connected with grand challenges.
“Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering [8]	Makes several recommendations. Strengthen Materials Data Infrastructure (MDI) core in repository, registry, and tool development. Sustain and grow MDI-dedicated funding. Create, execute, and monitor incentive mechanisms. Develop demonstration project and cross-disciplinary collaboration to accelerate adoption of the MDI. Contains a nice summary of previous workshops and reports in this area.
“Opportunities for New X-Ray Sources to Shed Light on New Functional Materials” [15]	Summary of a workshop. Combining the unique aspects of the proposed MaRIE XFEL with the ability to perform in situ measurements and subject materials to extreme and/or hazardous dynamic environments will enable a host of novel studies, particularly on materials such as plutonium that are difficult or impossible to study elsewhere.
“Management, Analysis and Visualization of Experimental and Observational Data: The Convergence of Data and Computing” [5]	Comprehensive report on experimental and observational data challenges (mathematical aspects of data analysis, software engineering and software infrastructure, visual data exploration and analysis, operating systems, runtime and architecture, service facilities, data management workflow, storage, metadata and provenance, data curation). Case studies of Advanced Light Source, Advanced Photon Source, and Linac Coherent Light Source are included.

[1] https://science.energy.gov/~/media/bes/besac/pdf/Reports/Challenges_at_the_Frontiers_of_Matter_and_Energy_rpt.pdf

[2] https://science.energy.gov/~/media/bes/pdf/reports/2017/BRNIDTET_rpt_print.pdf

[3] https://science.energy.gov/~/media/bes/pdf/reports/2017/BES-EXA_rpt.pdf

[4] https://science.energy.gov/~/media/bes/pdf/reports/2015/Data_and_Communications_in_Basic_Energy_Sciences_rpt.pdf

[5] https://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.pdf

[6] https://www.asminternational.org/documents/10192/25925847/ASM+MDA+Workshop+Report+Final.pdf/0e29644e-a439-4928-a07a-8718817a46e4

[7] https://www.energy.gov/sites/prod/files/2018/03/f49/AI%20Applied%20to%20Materials%20Discovery%20and%20Design_Workshop%20Summary%20Report.pdf

[8] http://www.tms.org/Publications/Studies/Materials_Data_Infrastructure/Materials_Data_Infrastructure.aspx?hkey=d228f86c-e269-49a2-a638-395285b760e4

[9] http://press3.mcs.anl.gov/futureplatform/files/2017/11/FOAP-Roadmap-Report.pdf

[10] https://science.energy.gov/~/media/bes/pdf/reports/2018/Ultrafast_x-ray_science_rpt.pdf

[11] https://science.energy.gov/~/media/bes/pdf/reports/files/Computational_Materials_

Science_and_Chemistry_rpt.pdf

[12] https://science.energy.gov/~/media/bes/pdf/reports/files/muee_rpt_print.pdf

[13] https://dcs-aps.wsu.edu/documents/2016/04/dcs_user_workshop_report.pdf

[14] https://science.energy.gov/%7E/media/bes/pdf/reports/files/Next-Generation_Photon_Sources_rpt.pdf

[15] https://www.lanl.gov/science-innovation/science-facilities/marie/_assets/docs/workshops/opportunities-new-x-Ray-sources.pdf

[16] https://conf.slac.stanford.edu/icfa-ml-2018

Appendix J: Experiment, Measurement, and Characterization

Discussion at the workshop identified different scientific workflows in terms of user science visits, classifying them as experiments, measurements, and characterizations. This classification is useful in understanding how experimental data can be used and help shape an ecosystem of tools that might address different classes of usage.

Experiment: putting a sample in without much knowledge of what to look for and using ad hoc setups and data acquisition. More prevalent in the past, or when the technique is still being developed actively. Very hard to plan for due to high variability and uncertainty in data requirements, analysis, and simulations. Can be viewed as hunting, and while not an efficient use of beamtime, can lead to innovation and is very good for training users.

Measurement: knowing exactly what you are doing and how. Usually requires a lot of preparation and (forward) simulations, and the experimental setup is mature and well-known, while the models might not be. Can produce a lot of data and relies on quick in situ feature identification and so-called “go on-quit” decisions. Examples of measurements include those produced by spectroscopic techniques such as RIXS and XPCS that measure the spectrum of absorption and transmission of light from a material.

Characterization: while the technique is well established, the sample features are not known in detail. It generates a lot of data, and it relies on inversion algorithms and can also benefit most from machine learning. Examples: ptychography, single shot particle diffraction, screening time (where you check sample suitability, see what is in a sample you made before starting the experiment).

Los Alamos
NATIONAL LABORATORY
EST. 1943