Response to Request for Input (RFI)-National Big Data R&D Initiative

Center for Nanoscale Materials
Argonne National Laboratory
9700 S. Cass Ave., Argonne, IL 60439 USA

Contacts:

Ian McNulty
X-ray Microscopy Group Leader, Senior Scientist
mcnulty@anl.gov, 640-252-2882

Katie Carrado Gregar
CNM User & Outreach Programs Manager
kcarrado@anl.gov, 630-252-7968

Big-Data experience: Significant quantities of scientific data are being generated at the Center for Nanoscale Materials by both experimental and theoretical approaches. The highest data rates reach 1 TB on time scales of an hour; several activities at CNM generate this scale of data within a few days. We anticipate that these data rates will increase by an order of magnitude over the next five years. Nearly all of this data requires extensive analysis, and in an increasing number of cases, high performance computing resources, to extract scientifically useful information from it. In addition to producing Big Data on a sustained basis, CNM is actively planning how to manage its current data explosion and correspondingly increasing needs over the next few years. We co-organized a workshop addressing several of these issues at the 2013 CNM/EMC/APS User Meeting titled, "Driving Discovery: Visualization, Data Management, and Workflow Techniques", and participated in a meeting held on 25-26 October, 2014 in Melbourne, Australia, titled "Big Data X-ray Microscopy Workshop". We are participants in a joint ASCR/BES-funded project as well as ongoing Big Data-related collaborations with staff at Lawrence Berkeley National Laboratory and Brookhaven National Laboratory.

We addressed the NITRD submission questions as follows:

1. What are the gaps that are not addressed in the Visions and Priority Actions document?

Response: The principal gap in the Visions and Priority Actions document that impacts our community now, and that is expected to impact it into the future, is analysis between acquisition of data and its reduction into publishable and usable form. While far from ideal, we are better at storing the raw data as it is generated, and organizing processed and reduced data into a form suitable for dissemination, then we are at keeping up with the analysis necessary to process and reduce the raw data (the "Analysis Bottleneck").

2. From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

Response: The National User Facilities such as the Nanoscale Science Research Centers and Light Sources have a significant Big Data problem: a growing fraction of the prodigious amount of data they produce is not getting analyzed. The primary sources of this flood of data are multispectral, megapixel detectors with high frame-rates (up to 10 Kframes/s), substantially brighter (100-200x) x-ray sources, and increasingly larger-scale (up to petascale) computational capabilities being developed to analyze and simulate more complex (multidimensional) problems and atomistic (more than 1 million atom) systems. From a facility utilization standpoint it is highly inefficient to take days to months to analyze a data set acquired in minutes. Guiding the course of a research program or making critical decisions on the time scale of the data acquisition becomes impossible under these conditions. Solutions to the Analysis Bottleneck- new analysis approaches, paradigms and infrastructure for keeping up with the data being collected - will have the highest impact on our community, and provide the best possible utilization of the considerable investment in our National User Facilities such as the CNM.

3. What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem?

Response: The game-changer for our community would be a concerted investment into infrastructure and methods to resolve the Analysis Bottleneck. This, more than any other investment, will create the best possible opportunity space for scientific discovery. This requires investment well beyond than a brute-force approach, e.g. just adding more computational capability, data storage space, and faster computer networks. We believe a more effective approach for our community, in addition to hardware investments, would be to engender greater awareness among our users and staff through workshops and seminars about the opportunities and challenges presented by this bottleneck, to support R&D that addresses these opportunities and challenges through more effective data-sharing, algorithmic, and analysis methods, and to encourage partnerships with other laboratories and groups facing similar challenges so as to address them together.

4. How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

Response: The most effective ways that government can enable new partnerships in our user facility environment is to promote awareness about the challenges and opportunities of Big Data, to support longer term as well as pilot grants via proposal calls for development of education, new approaches and infrastructure to manage Big Data more effectively, and to encourage inter-laboratory (e.g., between CNM and the Molecular Foundry) as well as cross-divisional collaboration (e.g. between DOE-BES and DOE-ASCR).

5. A short explanation of why you feel your contribution/ideas should be included in the strategic plan.

Response: As a National User Facility we represent the needs of a broad community of scientists that will benefit from improvements in managing Big Data, leading to greater scientific productivity and facility utilization.

Additional information: We surveyed the CNM staff and User Community to estimate current and future data generation rates, and to consider what aspects of our local Big Data environment pose the greatest bottlenecks to scientific productivity at CNM. Table 1 summarize our findings. A 1-5 scale was used to rate the significance of the data bottlenecks, where 1 is most and 5 is least significant.

Questions	Response 1	Comments	Response 2	Comments	Response 3	Comments	Response 4	Comments	Response 5	Comments	Response 6	Comments
1. What typical data volumes per experiment and data rates per day do you have currently? Please quantify these in gigabytes (GB) and GB/day.	1-3 GB/day	Currently we are generating roughly 1-3 GB/day with our current set ups, but I see this increasing in the near future.	10 GB/day		0.01 GB/expt, 0.2 GB/day		1 GB/day	Use is not daily.	Up to 1000 GB/expt, 200 GB/day		Up to 16K GB/expt, 2K GB/day	We already have a major problem analyzing, storing, and transporting these data.
2. What data volumes and rates do you expect next year as your needs and technology advance? Please quantify these in gigabytes (GB) and GB/day.	Not indicated	We are moving towards more nanoscale mapping and STXM type experiments at APS Sector 26. We don't know what our data sizes will be but they will likely larger than we currently handle, especially as x-ray imaging moves towards the 1 nm range.	20 GB/day		0.02 GB/expt, 0.4 GB/day		1-5 GB/day	This incorporates using a new high sensitivity, high speed camera.	2000 GB/expt, 500 GB/day		32K GB/expt, 4K GB/day	
3. What data volumes and rates do you expect in 5 years? Please quantify these in gigabytes (GB) and GB/day.	Not indicated	See #2	50 GB/day		0.01 GB/expt, 2 GB/day		5 GB/day	On a rather regular basis, but again, not every day.	10K GB/expt, 2K GB/day		100K GB/expt, 12K GB/day	
4. Which of the following issues pose the greater bottleneck to extracting the most science from your data? Please choose one or more issues and rank them on a scale of 1-5 (1 being worst):												
(a) data acquisition (such as due to slow detectors)	5	This is not the rate limiting step in our experiments.	N/A		N/A		1		5		5	
(b) gaining access to your data (via networks, data stores, etc.)	3-4	Gaining access to my data has been problematic, but not so much.	2		N/A		2		3		2	
(c) early-stage processing	2	Early stage data processing is where I see the bottleneck occurring. It can take some time to go from the raw data to some form of usable information in regards to scanning XRF measurements.	1		N/A		3		2		1	
(d) final processing and analysis for publication	4	Final processing has been made easier with the development of software like MAPS at the APS. This is no longer a bottleneck.	N/A		N/A		3		1		3	

Table 1. Results of data generation rates and bottlenecks at CNM.