Accelerating Data-driven Innovation and Scientific Discovery

Robert J. Harrison, rharrison@bnl.gov, (631) 3744-7090
Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973-5000

Brookhaven National Laboratory (BNL) is a multipurpose research institution funded primarily by the U.S. Department of Energy's Office of Science. Located on the center of Long Island, New York, BNL brings world-class facilities and expertise to the most exciting and important questions in basic and applied science—from the birth of our universe to the sustainable energy technology of tomorrow. We operate large-scale facilities for studies in physics, chemistry, biology, applied science, and a wide range of advanced technologies. The success and benefit to the nation of these facilities and associated research programs are predicated upon our continued innovation of new techniques and technologies to manage, mine, and analyze data at the frontier of scientific discovery.

For instance, data-centric computation at BNL is central to discoveries at the CERN Large Hadron Collider (LHC) that is the largest scientific enterprise worldwide in data intensive computing. BNL's ATLAS Tier 1 Center (the largest outside CERN) delivers cost-effective computing to many physics programs, and BNL also provides the PanDA (Production and Distributed Analysis) system that orchestrates ATLAS' data intensive computing at our computing facility and around the world: about 1.3 Exabytes were processed by PanDA in 2013. The ATLAS data set emerging from the first LHC run is currently about 150 PB. BNL physicists at our Relativistic Heavy Ion Collider (RHIC), the only collider now operating in the U.S., leads the world in exploring how the matter that makes up atomic nuclei behaved just after the Big Bang. These and multiple other capabilities make BNL one of the largest Big Data / High Throughput Computing resources and expertise pools in US science.

Data is the lifeblood of all our science programs, not just high-energy and nuclear physics. BNL is involved in all main DOE cosmology experiments. There are two unique features of cosmological datasets: first the information of interest is buried in correlations (e.g., power spectra and bispectra) within the petascale data sets, and second, cosmological datasets offer a great opportunity for serendipitous discovery but this is hard to automate and the data volume vastly exceeds what can be done even with citizen scientists – there are not enough humans around. In Climate, Environment, & Biosciences, BNL seeks to understand the relationships between climate change, sustainable energy initiatives, and the planet's natural ecosystems. For instance, we host the Atmospheric Radiation Measurement (ARM) Climate Research Facility that operates and hosts the data from strategically located in situ and remote sensing observatories to improve the understanding and representation of clouds and aerosols as well as their interactions and coupling with the Earth's surface. BNL plays a lead role in the DOE

Knowledge Base (KBase) that is a big-data bioinformatics collaboration developing tools for storing, interrogating and analyzing the exponentially increasing body of genome and metagenome sequences. A new frontier is emerging at the brand new National Synchrotron Light Source-II (NSLS-II) that uses electrons accelerated around a ring at nearly the speed of light to create beams of light in the x-ray, ultraviolet, and infrared wavelengths. It will generate many thousands of times more data than its predecessor being 10,000 times brighter with much larger/faster detectors and new instruments. Analysis of the generated petascale data will pave the way to discoveries in physics, chemistry, and biology — advances that will ultimately enhance national security and help drive the development of abundant, safe, and clean energy technologies. These are just a few of our interests that also include energy security, smart grids and energy infrastructure, remote sensing, cyber security, urban science, and so on.

In what must government invest to ensure progress towards broad science objectives while fully leveraging the huge economic engine of big data in business and industry that will continue to push technology in most of this space?

- Inter-agency research programs directed towards the intellectual foundations of translating science and engineering data into knowledge. Data is primarily a deep intellectual challenge, not just technology. For instance, new math and algorithms are necessary to efficiently and robustly identify correlations in peta/exascale data sets, to solve the inverse problems underlying advanced experiments, to robustly fuse data from multiple sources, or to identify novel features representing unexpected discoveries.
- Industry-academic-government partnerships are needed to fill in the gap between what is needed and what is feasible with off-the-shelf technologies. Motivating this are the investments currently underway by the Department of Energy in pursuit of exascale high-performance computing. The equivalent program is needed for big scientific data.
- The continuum between big data and big compute needs new technologies to open up new frontiers for discovery and to enable cost-effective reuse of computer infrastructure.
- Workflow and scientific data management platforms capable of handling the volume, velocity, and variety of scientific data that will very soon exceed exabytes.
- Long term and sustainable national plans, avoiding unfunded mandates, for all aspects of cyber infrastructure so that scientists can plan, share and coordinate/leverage investments.
- Aggressive development and development of advanced networking.
- Topical data facilities (such as the BNL RHIC Atlas Computing Facility, RACF) can be applied across multiple science domains leveraging the experience and domain expertise of the scientific/technical staff and the network/computer/storage/software infrastructure.

These are explored in more detail in other BNL position papers: Exascale Data Mining for Visual Data Analytics; Facilitating Big Data Sharing between Institutions; HEP/NP data intensive, high throughput computing programs in the BNL Physics Department; and NSLS-II.