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

Apurva Mehta and Amber Boehnlein; Staff Scientists, SLAC National Accelerator Laboratory

Ability to probe the chemistry and structure of materials at atomic scale has not only allowed a deeper understanding of fundamental processes in materials, but has illuminated knowledge-driven pathways to search the Materials Genome for novel new materials and devices, and the ability to scale-up and optimize them to address many of our technological challenges. With this vision in mind our nation has, over the last three decades, invested and continually upgraded x-ray, electron and neutron based national user facilities. With the continuing investments in brighter sources and development of faster detectors, the rate of data generation is perpetually accelerating; the pace of new discoveries and deepening scientific understanding, on the other hand, hasn't seen equally dramatic increase. The widening gap arises because the data management infrastructure and advanced data analytics essential to convert data to knowledge hasn't progressed sufficiently to keep pace with the volume as well as the changing modality of data collection. The widening gap between data and knowledge generation is also affecting the quality of the data: it is noisier, of uneven quality, and often does not optimally cover the most significant portion of the parameter space.

The traditional pathway to evidence based new knowledge is organizationally divided into sequential stages of a cyclic process. Traditional data to knowledge cycle begins with hypothesis generation, followed by experimental design, data collection, data quality assessment, and extraction of relationships and trends, which often leads to modification and refinement of hypothesis generation and experimental design. Historically, every stage of this cycle is curated and supervised by humans, often the same team of humans. Human eye and brain are superb at detecting subtle changes and seeing patterns buried in noise, but processing speed is a major limitation. As the speed of data generation increases, the human brain becomes overwhelmed and cannot keep up with on-the-fly curation, certification, and supervision, throttling the pathway from 'Big Data' to new knowledge. Accelerating pace of data generation, therefore, demands development of new tools and protocols, which overcome human limitations and even surpass their abilities, to reliably assess noise and spectral distortions and accurately extract subtle and hidden patterns from large, high dimensional and noisy data sets. Such tools must provide feedback within milliseconds (and with the commissioning of even higher throughput facilities must become even faster) to be useful for restore responsive control of data collection. In conclusion, the new data management software and hardware must be computationally efficient and fast, operate without human supervision and must be integrated back into the process of data generation to allow a knowledge cycle based on Big Data to operate optimally.

Our recommendation is then that National Big Data R&D Initiative must prioritize the development of scalable systems that seamlessly integrate the collection of experimental research data with sophisticated, smart and high-speed data analytics at high volume data (Big Data) generators, such as light sources and other national user facilities. This will insure that the highest quality data is collected under the most relevant experimental conditions and even the information hidden in noisy and distorted regions of data is reliably extracted, leading to accelerated pace of knowledge generation. Federal investment that can integrate big data analytics with big data generation will enable high production of knowledge from the already significant investment of Federal funds in Big Data generating facilities.

Generation of knowledge at an accelerated pace from accelerating pace of data generation also requires close collaboration among researchers who collect experimental data; experts in statistics, image processing and computer science who assess the quality of data and extract hidden relationships and features from that data (data miners); and scientists who develop models based on extracted relationships, and predict yet undiscovered relationships, phenomena, and (device) behavior and functionalities. National investment that will enable integration of data collection and analytics (in a Big Data ecosystem) will effectively bring these currently distinct communities together. Training the next generation of scientists in these emerging cross-disciplinary techniques must explicitly be part of this agenda as well.

It is also critical that this new infrastructure is widely (and easily) accessible to researchers from broad range of scientific disciplines. This will enable diverse communities, from image processing to instrument developers, from electron microscopists to x-ray spectroscopists to learn and build from successes and failures of each other (e.g., x-ray scattering community building upon success of x-ray spectroscopy community, which in turn learns from successes in image processing and compression.)

National user facilities are ideal environment for the development of the proposed new big data R&D infrastructure as they are intellectual hubs and generate a significant fraction of big data in scientific research. Furthermore, their day-to-day operation is overseen by staff scientists who closely interact with the several thousand users from a wide range of disciplines and institutions (from academia to industry) and they themselves come from wide range of backgrounds and collectively have broad range of interests and knowledge.

To design an optimal knowledge generation engine based on Big Data requires three key new technical developments. The new developments include infrastructure to make data transparent and easily accessible by wide range of researchers, researchers beyond just those that generated the data. This necessitate development of common and self-describing data formats and data and meta-data archiving formats; and data storage and retrieval systems which are fast, secure, fail-safe, and easy to access and maintain.

It includes building machinery to assess the quality of the data faster than the rate of data collection; allowing sources of noise, artifacts and distortions to be discovered in real time and when possible to quickly alter the data collection conditions to optimize the data quality.

It includes building analytics tools – algorithms, protocols and approaches – that are robust, reliable, and computationally efficient and applicable to wide range of data as well as combination of different data streams (e.g., electron spectroscopy with optical spectroscopy; x-ray scattering with a chemical probe of reaction kinetics). Two types of data analytics are needed: 1) tools which are based on the knowledge of sources of noise and the detector responses and large population set (which enable use of statistical methods) to allow useful information to be extracted even from noisy and distorted data; 2) and tools that can, with minimal input from a human, reduce the Big Data to a small set of features and relationships that can not only be used to further curate the dataset, but can also be used to modify and refine data collection strategy on-the-fly (for example, collect more and closely spaced data across a phase transition).