We are an elected committee of users of the national department of energy neutron source facilities, the Spallation Neutron Source and the High Flux Isotope reactor at Oak Ridge National Laboratory, who represent over 1200 users spanning domestic and foreign academic institutions, government laboratories, and US industry. We are both generators and consumers of Big Data, with individual experiments producing datasets too large to analyze on single machines or with existing software tools. We seek to extract scientific meaning and research and development value from the data generated by our experiments. The vast quantities of data produced by the advances in instrumentation and detector technology promise to usher in a new era in our ability to understand and design complex materials (which in turn are tied into ongoing initiatives for "Materials by Design" such as the White House Office of Science and Technology Policy Materials Genome Initiative). At the same time, we are not software or computer aficionados, and do not have the time or resources to devote to dealing with the challenges in the storage, transport, analysis and presentation of the large datasets produced in these experiments.

Our needs, as the scientists and researchers tackling problems from sustainable energy and curing cancer to fundamental science, would best be met if a national big data framework provided appropriate incentives to all parties so that the necessary hardware, software tools, and technical expertise are put in place to allow us to focus on the science and engineering of what we do, not on the manhandling of the data. This involves addressing many issues, including:

1) **Data Analysis Software**: The past forty years has seen the development of a wide range of bespoke software tools for the workup and analysis of data from national facility instrumentations – from codes that simply convert between different possible representations of the data to Rietveld codes for structural modeling. With rare exceptions, these codes were designed and built on the premise that data analysis would occur on a single computer, with manual intervention required to ensure the data input and outputs are in usable formats. While highly effective for their time, these tools are unable to cope with the volumes of data now available. Versions of these bespoke tools that use a cluster or distributed processing approach, with little or no manual intervention, simply do not exist. Yet without them, we as users are simply not able to tackle the science and engineering that we are interested in. A national big data framework must ensure that functional and usable (by non-computer expert) software tools exist.

2) **Data Storage and Transfer**: Today, even relatively modest datasets – 0.1 to 10 terabytes in size – present a significant challenge to move between the national facility and the home institution of the researcher, due to the patchwork nature of networking and storage infrastructure. Costly upgrades of national, facility, and home institution infrastructure are required to solve this issue, but typical federal grants do not provide separate funding resources to accomplish either networking upgrades or install and maintain the requisite data storage (especially at the home institution). To ensure a healthy data lifecycle, additional funds specifically directed to ensuring speedy data transfer and adequate data storage, in addition to those typically allocated for a grant, are required.

3) **Data Origin, Citation, and Preservation**: The data collected only has meaning when there is knowledge of the specimens/materials/etc. measured. Thus for data to be usable by those other than the researchers who originally carried out the measurement, appropriate annotation of measurement conditions (including as much sample information as practical) is critical. Further, it is simply no longer possible to include all of the primary data in a research publication or internal report, nor are there resources provided to ensure long-term availability of the primary

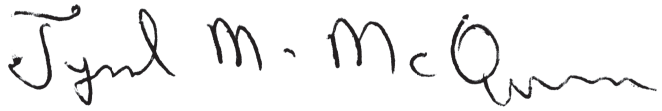
(raw, unprocessed) data. To make the data part of the scientific record, and permit its usability by future researchers, mechanisms to cite raw (and processed) datasets, and preserve for "eternity" are needed. In part, this means moving the data to newer computing systems as they are created.

4) **Data Presentation**: For the integrity of the scientific and engineering process, it is critical that researchers be able to present, in a printed format, the characteristic features of the data that lead to the conclusions drawn from the data. This requires the development of new approaches to the presentation of Big Data analysis results. Achieving this also requires the development of new ways of visualizing and interacting with the large data sets in their full glory, not just in a reduced parameter space.

5) **User Training and Computing Capability Access:** As with the computing revolution before it, there must be adequate training and investment in end-user usability of the software analysis tools for Big Data. Virtually all of our users have little or no programming experience – which locks them out of the tools currently available to handle large datasets in a distributed fashion. Even for our users who are so inclined, often access to the necessary cluster computing facilities is either clunky or incurs a significant time delay, which impedes research progress.

In short, from our perspective as users, any national big data framework must focus on providing seamless access to data and analysis capabilities that do not require a computer science degree to utilize. At the same time, the tools and methods to tackle these issues cannot be developed independently of users – history is littered with examples of data processing software that failed because users were never brought in as active participants in the process and consequently either did not meet users' needs, or were so clunky and counterintuitive as to effectively not exist. Instead, ***any national big data framework must foster active engagement of users at all stages of tool development***, while at the same time ***providing the necessary funding and stewardship*** for the maintenance of the tools so they can be used without distraction to tackle the grand challenges in health, science, and engineering.

Sincerely,



Prof. Tyrel M. McQueen on behalf of the SHUG-EC
Chair, SNS/HFIR User Group Executive Committee (SHUG-EC)
Department of Chemistry
Department of Materials Science and Engineering
Department of Physics and Astronomy
Institute for Quantum Matter
The Johns Hopkins University
https://occamy.chemistry.jhu.edu