# Extreme-scale computing for new instrument science

Ian Foster

Department of Computer Science, The University of Chicago, Chicago, IL, USA
Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA
Computation Institute, University of Chicago and Argonne National Laboratory

Rapidly evolving data rates, analysis methods, and science processes within the experimental and observational science communities (see Figure 1) are driving dramatic increases in computational requirements—increases so large that these communities will soon require exascale-class computational environments to be productive. Instrument science thus has the potential to greatly expand the impact of exascale technologies.

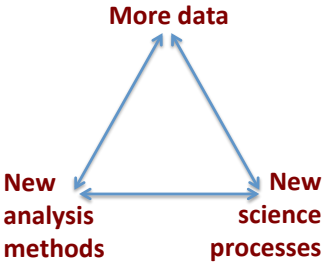


**Figure 1: Major drivers of new instrument science**

The science communities in question work with exceptionally powerful new scientific instruments such as light sources, tokomaks, telescopes, accelerators, and genome sequencers. These communities face a range of big data challenges, as innovations in sensor technologies increase data volumes and velocities at rates often greater than Moore's Law. But the true challenges faced by instrument sciences—and the reason why exascale technologies are so important to their future—relate to the need to transform instrument output (and other data) into actionable knowledge and then act on that knowledge in human-useful timeframes. The scale of both these challenges and the associated opportunities mean that these science communities are increasingly embracing team science methods, in which communities collaborate to develop, extend, and apply substantial knowledge bases—an approach that until recently was limited to a few big-science communities such as high energy physics.

Consider, for example, the use of a light source such as Argonne's Advanced Photon Source to study the internal structure of candidate battery materials. Today, the typical science process followed in such experiments involves sequential steps of material synthesis, data collection, and data analysis—a process that can easily take months from start to finish. With exascale technologies, it becomes possible to imagine far more rapid science processes in which, for example, knowledge bases constructed from past experiments, the literature, and simulation models are used to flag "interesting" features in data as it is generated; instrument output is assimilated, as an experiment is running, into a simulation model that is then used to guide future data collection; and an evolving integrated knowledge base is used to guide the choice of future experiments.

The successful realization of such scenarios requires innovations in not only the hardware and software technologies typically considered by the exascale community [1, 2] but also in three areas relating more specifically to scientific knowledge:

1. *knowledge management and fusion*, to permit the rapid integration of large quantities of diverse data and the transformation of that data into actionable knowledge;
2. *rapid knowledge-based response*, to enable the use of large knowledge bases to guide fully or partially automated decisions within data-driven research activities; and
3. *human-centered science processes*, to enable rapid specification, execution, and guidance of science processes that will often span many resources and engage many participants.

Rich bodies of work exist on the development and use of knowledge-based methods within many fields. However, the unique characteristics of modern instrument science pose unique challenges relating, for example, to data volume, variety, and complexity, and the need to balance the complementary capabilities of human experts, computational methods, and sensor technologies.

The integration of simulation and instrument represents an important example of emerging new science processes within instrument science. In general terms, the opportunity is this: computational simulations capture the best available, but imperfect, theoretical understanding of reality; data from instruments provide the best available, but imperfect, observations of reality. Confronting one with the other can help advance knowledge in a variety of ways.

The Discovery Engines for Big Data project at Argonne is pursuing opportunities in this area, with problems from cosmology and materials science as application drivers. As illustrated in Figure 2, the work in cosmology is comparing virtual skies, constructed from large-scale simulations, with the "real sky" as revealed by digital sky surveys, with the goal of constraining potential theories of dark energy, while in materials science, simulations are being used to determine which defect structures within disordered materials may best explain observed diffuse scattering data. In both cases, large-scale computation is required to process instrument data and to prepare the simulated realities with which observations are compared. In the materials science application, the opportunity exists to use feedback from simulation-observation comparisons to guide experimentation.

These and other application projects within DOE and elsewhere are generating a growing awareness of the opportunities inherent in new science processes based around the knowledge-based integration of large quantities of sensor data, large-scale computation, and human expertise. With care, these opportunities can serve as an important driver for emerging exascale technologies.

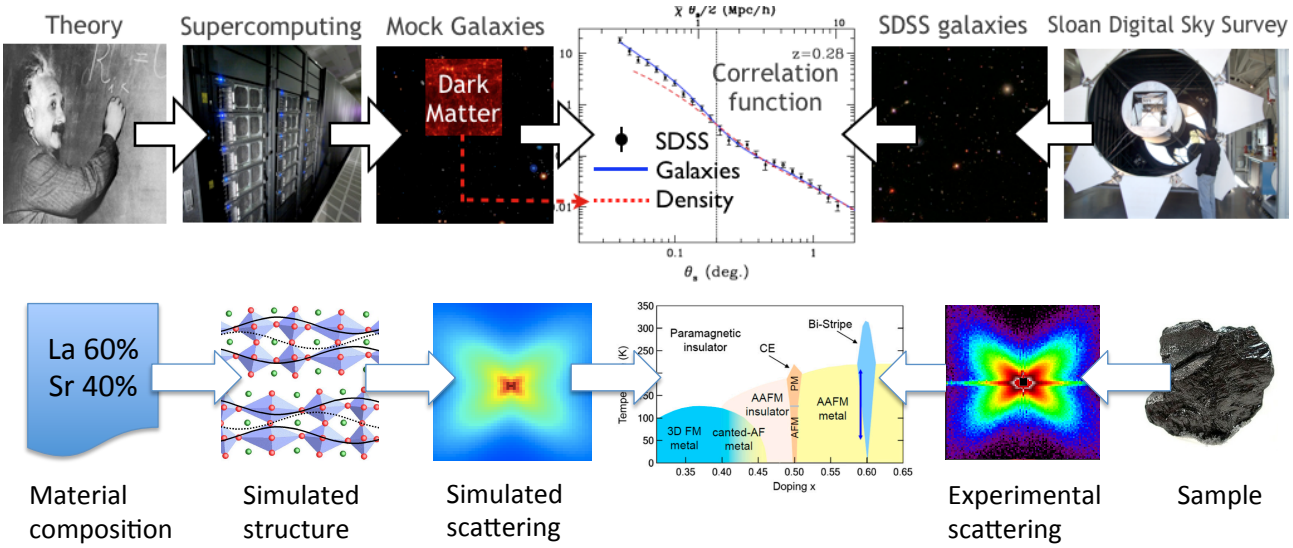


**Figure 2: Two examples of new science processes that couple instrument and simulation. Above: using digital sky survey data to constrain models of dark matter (source: Salman Habib, Argonne). Below: using simulation models to infer defect structure in disordered materials (source: Ray Osborn, Argonne).**

## Acknowledgments

I am grateful to Rich Carlson, Barbara Jennings, Scott Klasky, Kerstin Kleese-Van Dam, Ruth Pordes, and David Skinner for many insightful discussions on these topics in the context of the U.S. Department of Energy (DOE)'s Accelerating Scientific Knowledge Discovery group. This research was supported in part by DOE under contract DE-AC02-06CH11357.

## References

1. Dongarra, J., et al., *The international exascale software project roadmap.* International Journal of High Performance Computing Applications, 2011. **25**(1): p. 3-60.
2. Chen, J., et al., *Synergistic Challenges in Data-Intensive Science and Exascale Computing.* DOE ASCAC Data Subcommittee Report, Department of Energy Office of Science, 2013.