# Big Data


This special issue of *Computing in Science and Engineering* is devoted to the emerging challenges facing data-intensive scientific computing and the accompanying and highly significant opportunities on the horizon.

With the exponential growth in data acquisition and generation—whether by next-generation telescopes, high-throughput experiments, peta-scale scientific computing, or high-resolution sensors—it's an extremely exciting time for scientific discovery. As a result of these technological advances, the next decade will see even more significant impacts in fields such as medicine, astronomy and cosmology, materials science, social sciences, and climate. Discoveries will likely be made possible with amounts of data previously unavailable (as an example, see work by Alon Halevy and his colleagues.[1])

1521-9615/11/$26.00 © 2011 IEEE
Copublished by the IEEE CS and the AIP

Francis J. Alexander
*Los Alamos National Laboratory*
Adolfy Hoisie
*Pacific Northwest National Laboratory*
Alexander Szalay
*The Johns Hopkins University*

Given space limitations, we focus here on the computational challenges that scientists face in dealing with extremely large datasets; we don't cover the key related topics of data gathering, experimental design, and analysis tools, all of which must be further developed if we're to make optimal use of data-intensive scientific computing.

One of the challenges is shear scale: the data simply can't be moved for analysis. Therefore, the data must be analyzed in situ and/or we must develop methods for extracting a smaller set of relevant data. Among other challenges,

- high-flux, streaming data methods are often required because the analyst might have only one shot at accessing the data;
- the data of interest might be structured, as in the output of a large-scale computer simulation of a turbulent flow, or it might be unstructured, as in a large corpus of text;
- decisions about which data to keep and which to eliminate must be made in real time; and
- we must determine the optimal mix of access versus compute resources and how to organize the data.

An optimal approach to the overall data analysis problem requires close cooperation between computer scientists and domain experts.

## In This Issue

This special issue begins with a list of scientific challenges confronting the US Department of Energy. The DOE has been and continues to be a major player in high-performance computing. Obtaining optimal benefits from significant simulation efforts—including large-scale systems modeling in areas such climate and plasma fusion—and experiment analysis requires the ability to generate and manage large datasets. Moreover, tools must be in place to extract meaning and knowledge from that data.

In "Data-Intensive Science in the Department of Energy: Case Studies and Future Challenges," James P. Ahrens and his coauthors use a case study approach to tease out the common challenges that must be met for projects to succeed. These challenges include network and analysis infrastructure, data from large-scale climate and cosmology simulations, data from x-ray observatories, and neutron scattering data from DOE user facilities such as Argonne National Laboratory's Advanced Photon Source and Oak Ridge National Laboratory's Spallation Neutron Source. Ahrens and his coauthors discuss workflow models and data, software, and architectures required for success.

In his article, "Data-Intensive Scalable Computing for Scientific Applications," Randal E. Bryant explores the scalability requirements for data-intensive scientific computing, both for managing the data and carrying out large-scale numerical calculations with massive datasets. Bryant argues that future data-intensive scientific computing systems will differ considerably from the more traditional HPC systems used by the scientific community today, such as compute intensive, diskless clusters. He also discusses how a new class of highly adaptive and scalable architectures developed and used by Internet-based service companies could come to the rescue.

Finally, in "Extreme Data-Intensive Scientific Computing," Alexander S. Szalay focuses on the challenges faced by—but by no means unique to—the astronomy community where new telescopes coming on line will be generating petabytes of data per day. Szalay discusses the notion of balance characterized via Amdahl's law. He also offers a pathway forward using commodity hardware, given both the practical budgetary constraints that most universities and research centers face these days and the emerging power wall, which is soon to become the limiting factor in deploying high-performance computing.

## Addressing the Challenges

It's important that we embrace and promote a balanced approach to addressing the challenges of data-intensive scientific computing. In addition to the hardware investments required, there's a pressing need to invest in research and development of analysis algorithms. For example, data collection methods often fail to conform to the hypotheses in statistical analyses, and the data might not be independent and identically distributed. This is almost always true for data from experiments and observations of physical systems. There's also a serious need to develop deterministic, scalable algorithms for analysis, as well as randomized algorithms that are robust to almost certain hardware failures. In some cases, there's sufficient data, but scientists are faced with a semantic gap. Such a gap occurs, for example, in analyzing video streams.

There's also the risk that for some problems, *there still might not be enough data* to reach any defensible conclusions. So, as a research community, we shouldn't abandon work on developing analysis tools for extracting knowledge from limited data.

With the availability of all of this easily obtained data, there's a real danger that the scientific method might be side-stepped. We should therefore be especially diligent about adhering to the scientific method during all phases of the research process.[2] Getting more and better quality data certainly is desirable, but data alone is no substitute for the hard work of analysis.

If we asked computational scientists today what would maximize progress in their field, most would still say: more disk space and more CPU cycles. However, the emerging petabytes of data fundamentally change every aspect of scientific discovery: the tools (computer hardware and software), the techniques (algorithms and statistics), and, as a result, the cycle of the scientific method itself, which is greatly accelerated.

## References


1. A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems*, vol. 24, no. 2, 2009, pp. 8–12.
2. E.R. Dougherty, "On the Epistemological Crisis in Genomics," *Current Genomics*, vol. 9, no. 2, 2008, p. 69.



***Francis J. Alexander*** *is the director of the Information Science and Technology Center at Los Alamos National Laboratory. His research interests include statistical mechanics, computational physics, and optimal estimation. He received a PhD in physics from Rutgers University. Contact him at fja@lanl.gov.*

***Adolfy Hoisie*** *is a laboratory fellow and director of the Center for Advanced Architectures at the Pacific Northwest National Laboratory. His research interests include performance analysis and modeling of extreme-scale systems and applications. Hoisie is a recipient of the Gordon Bell Award and other honors for research and teaching excellence. Contact him at adolfy.hoisie@pnnl.gov.*

***Alexander Szalay*** *is the Alumni Centennial Professor of Astronomy and a professor in the Department of Computer Science at The Johns Hopkins University. His research interests are in cosmology, including statistical measures of the spatial distribution of galaxies and galaxy formation. Szalay has a PhD in astrophysics from the Eötvös University, Budapest. He is a corresponding member of the Hungarian Academy of Sciences and a fellow of the American Academy of Arts and Sciences. Contact him at szalay@jhu.edu.*