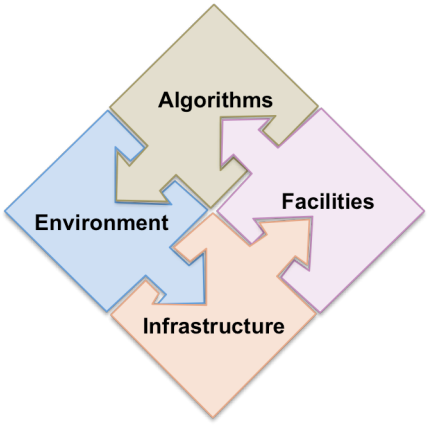
Enabling Big Data in Science

Mihai Anitescu anitescu@mcs.anl.gov
Charlie Catlett catlett@anl.gov
Ian Foster foster@anl.gov
Salman Habib habib@anl.gov
Mark Hereld hereld@anl.gov
Paul Hovland hovland@mcs.anl.gov
Jed Brown jedbrown@mcs.anl.gov
Kate Keahey keahey@mcs.anl.gov
Nicola Ferrier nferrier@anl.gov
Michael Papka papka@anl.gov
Robert Ross rross@mcs.anl.gov
Rajeev Thakur thakur@mcs.anl.gov
Tom Peterka tpeterka@mcs.anl.gov
Venkat Vishwanath venkatv@mcs.anl.gov
Stefan Wild wild@anl.gov
Mike Wilde wilde@mcs.anl.gov
Justin Wozniak wozniak@mcs.anl.gov

Technology advances have led to tremendous increases in the amount of science data generated from experimental facilities, observational platforms, and computational simulation.

The techniques used to manage, understand, and share these data have not kept pace. As a result, the scientific discovery process is being impeded, with engineers at facilities cobbling together systems to manage these data, scientists wrestling with antiquated tools to analyze data, and collaborators struggling to share data and knowledge with one another.



Four components of a successful data intensive science strategy.

A successful solution to the challenges of big data in science must encompass four key areas:

- State of the art algorithms
- High productivity environments
- Robust software infrastructure
- Capable facilities

Algorithms

With the explosion of data generated by experiments, observational platforms, and simulations, new techniques are required to efficiently reduce data, to manage data with uncertainty and/or with variable quality, to integrate scientific knowledge into analysis, and to operate at very large scale, on a variety of architectures. Our goal is to develop novel models and algorithms for analyzing science data in highly parallel and resource-constrained environments, keeping "science in the loop". This approach will bring together math advances, CS techniques, and domain knowledge to remove scalability barriers of classical algorithms, produce the highest quality results, improve interactivity for prototyping, and enable automation of complex analysis.

Examples of ongoing activities include the activity by S. Wang et al. in the analysis of fluorescence microscopy data from the APS and work by T. Peterka et al. in the analysis of

cosmology simulation data for the purpose of studying gravitational lensing and morphology of cosmic structure. Both of these activities highlight our approach of engaging local domain science staff in early stages of R&D to validate approaches while solving real problems.

Environments

Data science involves complex, resource-intensive, end-to-end processes that may span computers, buildings, facilities, and countries. Systems enabling design and execution of the complex workflows that describe the data science process are beyond state of the art in most cases. Our goal is to provide an end-to-end environment for data intensive science that provides effective views of scientists' data, facilitates automation of common tasks, and executes these tasks efficiently on large scale data resources.

One example of activity in this area is our enhancement and deployment of the Galaxy Platform for use in genome sequencing, proteomics, and Earth science applications. Through these deployments we have gained increasing understanding of the needs of small teams of scientists, pointing the way towards general tools for larger communities. Another aspect of this problem is in orchestrating complex scientific workflows, and our Swift parallel scripting language is creating new capabilities for executing complex workflows on a variety of platforms, including leadership computing systems.

Software Infrastructure

Software infrastructure is critical to the effective management of data in many science projects, yet there is a huge disparity between science projects in terms of their infrastructure, its maturity, and its capabilities. The ultimate success of data intensive science in the DOE rests on our collective ability to create reusable infrastructure that support numerous, diverse teams. Our goal is to develop a shared, large-scale, heterogeneous data science infrastructure that is capable of supporting many science domains at the multi-institution level.

Our activities in the KBase project inform us as to some of the requirements of multi-institution software infrastructure, for example highlighting the need for scheduling of computation based on data locality as well as the capabilities of specific sites.

Facilities

Finally, rapidly increasing data volumes and subsequent complexity of analysis are already creating situations where science teams simply do not have the hardware to unlock the knowledge buried in their data. Leading-edge data intensive science facilities will provide the data analysis and storage capabilities, as well as the connectivity to instruments and other sites, that can enable science communities to pursue discovery via data intensive science.

In conclusion, our objective is to foster the growth of a vigorous data science community. We see these four areas as necessary and complementary components of a "data science substrate" that will accelerate discovery in the sciences by allowing researchers to work with data in terms of familiar abstractions, through convenient interfaces, and with state of the art analysis techniques.