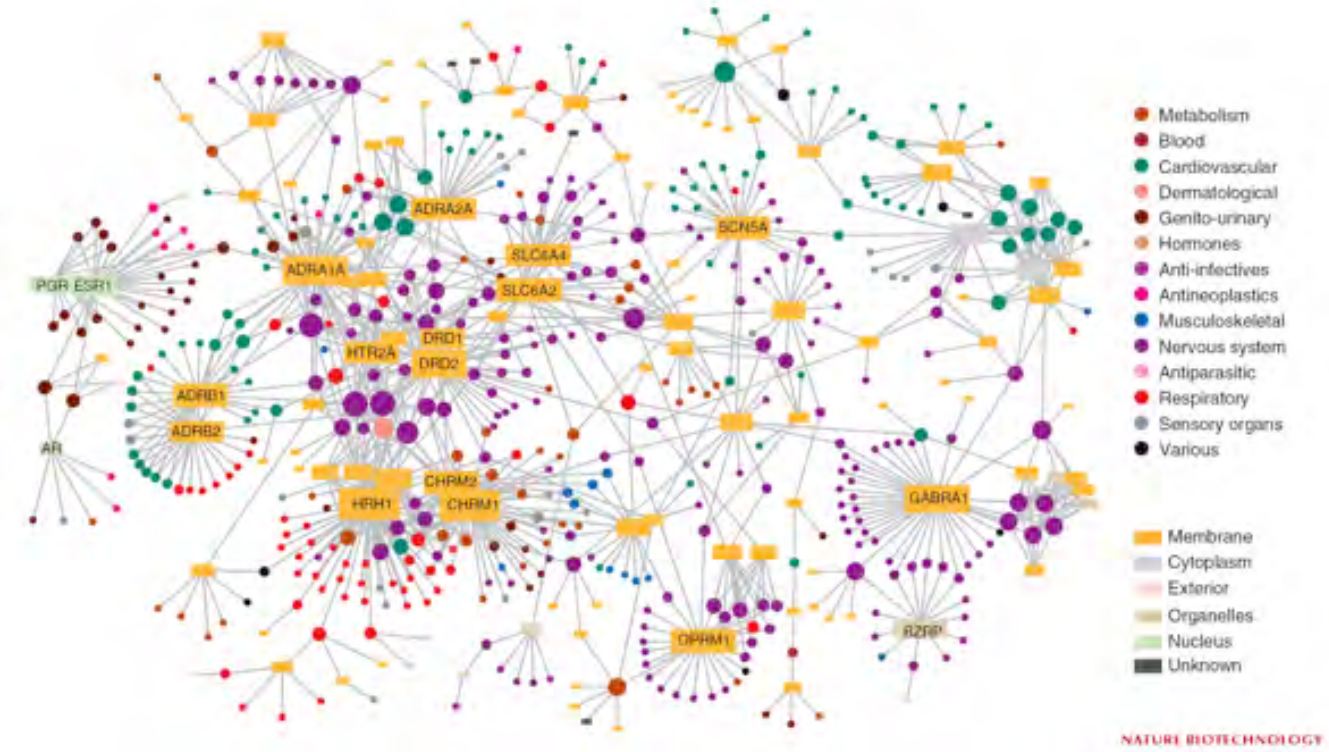
Mathematics for Analysis of Petascale Data

Report on a Department of Energy Workshop

June 3–5, 2008



Organizers and Authors:

Philip Kegelmeyer, Chair	Sandia National Laboratories
Robert Calderbank	Princeton University
Terence Critchlow	Pacific Northwest National Laboratory
Leland Jameson	National Science Foundation
Chandrika Kamath	Lawrence Livermore National Laboratory
Juan Meza	Lawrence Berkeley National Laboratory
Nagiza Samatova	North Carolina State University/Oak Ridge National Laboratory
Alyson Wilson	Los Alamos National Laboratory

,

Cover Figure: Drug-target network (DT network). The DT network is generated by using the known associations between FDA-approved drugs and their target proteins. Circles and rectangles correspond to drugs and target proteins, respectively. A link is placed between a drug node and a target node if the protein is a known target of that drug. The area of the drug (protein) node is proportional to the number of targets that the drug has (the number of drugs targeting the protein). Color codes are given in the legend. Drug nodes (circles) are colored according to their Anatomical Therapeutic Chemical Classification, and the target proteins (rectangular boxes) are colored according to their cellular component obtained from the Gene Ontology database.

From: *Drug-Target Network* , Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-László Barabási and Marc Vidal, Nature Biotechnology, Volume 25, Number 10, October 2007.

Contents

1 Executive Summary

"Mathematics for Analysis of Petascale Data" was the topic of a June 3–5, 2008 workshop organized on behalf of the Applied Mathematics Program of the Office of Advanced Scientific Computing Research (ASCR), Office of Science, Department of Energy.

The workshop was motivated by the realization that DOE scientists must address the challenges posed by petascale data sets. Extracting scientific knowledge from these massive data sets has become both increasingly difficult and increasingly necessary, as computer systems have grown larger and experimental devices more sophisticated. Mathematical methods have long been the mainstay of the analysis of such scientific data. Unfortunately, many existing methods fail to provide adequate robustness, scalability, and combinatorial tractability when applied to petascale data.

The workshop therefore engaged mathematical scientists and applications researchers to identify the next-generation mathematical techniques needed to meet the challenges posed by petascale data.

The application domains addressed were: astrophysics, biology, nanoscale chemistry and physics, networks, earth and climate systems modeling, fusion physics, accelerator physics, cybersecurity, combustion, and visualization. The organizing themes for the mathematics breakout sessions were: statistics, optimization, uncertainty quantification, machine learning, network and graph analysis, analysis of streaming data, and data reduction (which included dimension reduction, feature extraction, and topological methods).

For each of the application domains, the workshop participants identified a prioritized list of specific challenges faced by the analysis of data from that field. These challenges were considered from the perspective of each of the mathematics disciplines, generating a prioritized list of specific gaps that must be addressed in response.

The resulting discussions were vigorous, detailed, and generated many pages of notes and summaries. These were integrated and distilled to arrive at the primary themes and findings of the workshop, which assert the need to:

- **Adapt analysis methods to the requirements of scientific petascale data**
  - Algorithms must be re-engineered to scale with the size of the data, which is often independent of the number of model parameters, and so may require parallel, single-pass, or subsampling methodologies.
  - Petascale data sets are too large to easily move, yet are often non-uniformly distributed, requiring algorithms that come to the data, rather than vice versa, and can adapt to the lack of a global perspective on the data.
  - New algorithms should make effective use of new computer architectures that are being developed for data analysis.

- **Develop new mathematics to extract novel insights from complex data**
  - Analysis of petascale data in their raw form is often infeasible. Instead, improved methods for data and dimension reduction are needed to extract pertinent subsets, features of interest, or low-dimensional patterns.
  - Using data to make predictions or discover new science requires methods to build and evaluate appropriate models of large-scale, heterogeneous, high-dimensional data.
  - Interpreting results from petascale data requires methods for analyzing and understanding uncertainty, especially in the face of messy and incomplete data.
  - Near real-time identification of anomalies in streaming and evolving data is needed in order to detect and respond to phenomena that are either short-lived (e.g., supernova onset) or urgent (e.g., power grid disruptions).
- **Support a research environment that recognizes the challenges of petascale data analysis**
  - Effective mathematics research requires infrastructure, recognition of contributions, and incentives for efficient exchange and persistent storage of knowledge, both internally, within the mathematics community, and externally, to the application communities.
  - Mathematics is a critical part of the path from data to decision making; it leverages and completes investments in networking, architectures, and visualization.

There are a welter of challenges facing the application domains that matter to the DOE, as scientists gear up to handle petascale data and beyond. We believe that substantial and sustained support for the findings above will allow the Applied Mathematics Program to address those challenges in a focused and significant fashion.

2 Workshop Overview

"Mathematics for Analysis of Petascale Data" was the topic of a June 3–5, 2008, workshop organized on behalf of the Applied Mathematics Program of the Office of Advanced Scientific Computing (ASCR), Office of Science, Department of Energy.

The workshop was motivated by the realization that DOE scientists must address the challenges posed by petascale data sets. Extracting scientific knowledge from these massive data sets has become both increasingly difficult and increasingly necessary, as computer systems have grown larger and experimental devices more sophisticated. Mathematical methods have long been the mainstay of the analysis of such scientific data. Unfortunately, many existing methods fail to provide adequate robustness, scalability, and combinatorial tractability when applied to petascale data.

The workshop therefore engaged mathematical scientists and applications researchers to identify the next-generation mathematical techniques needed to meet the challenges posed by petascale data. Specific objectives were to:

- understand the needs of various scientific domains,
- delineate appropriate mathematical methods,
- determine the current capabilities of these methods, and
- identify the gaps that must be addressed to enable the effective analysis of large, complex data sets in the next five to ten years.

To this end a broad variety of participants were aggressively recruited. There were sixty-six attendees. Roughly half came from a background in DOE applications, and the other half came from a background in mathematics. Similarly, roughly half came from academia, and the other half from DOE laboratories, with a smattering of other government organizations (the National Science Foundation, the National Security Agency) mixed in.

The workshop was organized around a few plenary discussions and a slew of focused break-out sessions. The detailed organization is presented in Sections 4 and 5, and the full agenda is in Appendix B.

The central goal of the entire workshop was to facilitate the conversations necessary to address the objectives above. To that end, specific discussion questions (see Appendices C and D) were circulated to all attendees a week before the workshop. Notes from each breakout session were immediately uploaded to a dedicated file sharing site, so that subsequent sessions could easily access and build on the discussions that had come before.[1]

[1]The report, agenda, and other conference details are available at `http://www.orau.gov/mathforpetascale`. The raw notes and the summaries from the breakout sessions will be available at `http://drop.io/mapd` until May 1, 2009. Thereafter, search the ASCR home page, `http://www.er.doe.gov/ascr`.

3 DOE Context for the Workshop

To place the structure and findings of this workshop in context, it is likely worthwhile to review why the DOE would sponsor such a workshop in the first place. The motivation stems from the DOE Strategic Plan[2], particularly Theme 3, "Scientific Discovery and Innovation", whose aim is "Strengthening U.S. scientific discovery, economic competitiveness, and improving quality of life through innovations in science and technology". Two of the many aims cited in the expansion of that theme are to "Increase financial support for innovation-enabling research" and "Strengthen the ties between the basic research and applied mission programs in Departmental planning."

Accordingly, this workshop was structured to first investigate the DOE's applied mission needs, and then to determine what innovative research would be required to support those needs.

Furthermore, this workshop had the freedom, and the duty, to focus specifically on the applied mission needs posed by petascale data. It was enabled to do so by recent efforts relevant to the Applied Mathematics Program in the DOE ASCR office, efforts which tackled other dimensions of this problem space.

An important example is a May 2008 report entitled "Applied Mathematics at the U.S. Department of Energy: Past, Present, and a View to the Future"[3], which addressed the full range of science and engineering challenges that the DOE faces, and identified broad, long-term advances in mathematics necessitated by those challenges.

Another example is a recent and more tightly focused report on "Mathematical Research Challenges in Optimization of Complex Systems"[4]. This current report is similarly focused, on the mathematical issues generated by scale rather than complexity.

Of course, this analysis research does not proceed in a vacuum. In addition to mathematical analysis, DOE ASCR supports research[5] in networking, architectures, and visualization, to address the other critical components in the path from "data to decision".

And to broaden the scope by one final step, DOE ASCR's work provides significant support to the DOE Office of Science's[6] research programs in basic energy sciences, biological and environmental sciences, materials and chemical sciences, climate change, geophysics, genomics, and life sciences.

[2] `http://www.cfo.doe.gov/strategicplan/mission.htm`

[3] `http://brownreport.siam.org`

[4] Hendrickson, B. A., and Wright, M. H., (Eds.); *Mathematical Research Challenges in Optimization of Complex Systems*, U.S. Department of Energy — Office of Science Workshop Report, December 2006, `http://www.sc.doe.gov/ascr/Research/AM/ComplexSystemsWorkshopReport.pdf`

[5] `http://www.sc.doe.gov/ascr/`

[6] `http://www.sc.doe.gov/`

4 Application Domains

The first goal of the workshop was to "understand the needs of various scientific domains" pertinent to the DOE. This understanding was the motivating context for all consideration of mathematical methods, gaps, and suggested research.

Accordingly, the workshop hosted ten applications-focused breakout sessions, each intended to generate a prioritized list of the specific challenges faced in the analysis of data from each application domain. (See Appendix C.) There were also five plenary presentations, to help build a common understanding among all attendees.

The domains addressed were: astrophysics, biology, nanoscale chemistry and physics, networks, earth and climate systems modeling, fusion physics, accelerator physics, cybersecurity, combustion, and visualization. (Visualization, unusually but usefully, was treated as an application that could use math support, rather than as an analysis method). Below we present a brief summary of the issues highlighted by discussions around each of these application domains.

4.1 Astrophysics

Astrophysics is the study of the physics of the universe, from the smallest to largest scales humans can observe. It is undergoing an extraordinary transition from a data-starved to a data-swamped discipline; the observational data obtained in the next decade alone will supercede everything accumulated over the preceding four thousand years of astronomy. For example, the Large Synoptic Survey Telescope (LSST), in Figure 1, will be on-line within a decade and will generate multiple terabyte raw data sets each night and more than seven petabytes of reduced data each year. This is fifty times more data than the forty terabytes collected in the entire life of the Sloan Digital Sky Survey, which is the largest and most recent astronomical data collection project.

Realizing the full scientific potential of these observations will require correspondingly precise predictions from our theoretical models. Given the physical complexity of the systems involved, obtaining such predictions necessarily requires ever more detailed numerical modeling, and simultaneously generates an equivalent quantity of simulation data. From the detailed theoretical predictions made possible by complex simulations to the precise reference points obtained from painstaking analyses of the new observations, the development of astrophysics in the new millennium will be regulated by our computational and analysis capabilities.

The challenges to these capabilities are many. Observational data must be analyzed in real time to identify unusual events with few false positives, followed by a determination of the best use of resources to acquire the necessary data to study these events further. Other analysis is done off-line, and astronomers may go back to old data looking for additional information available about an object observed in the new data. This implies a need for distributed data mining as well as fusion of data from different surveys.

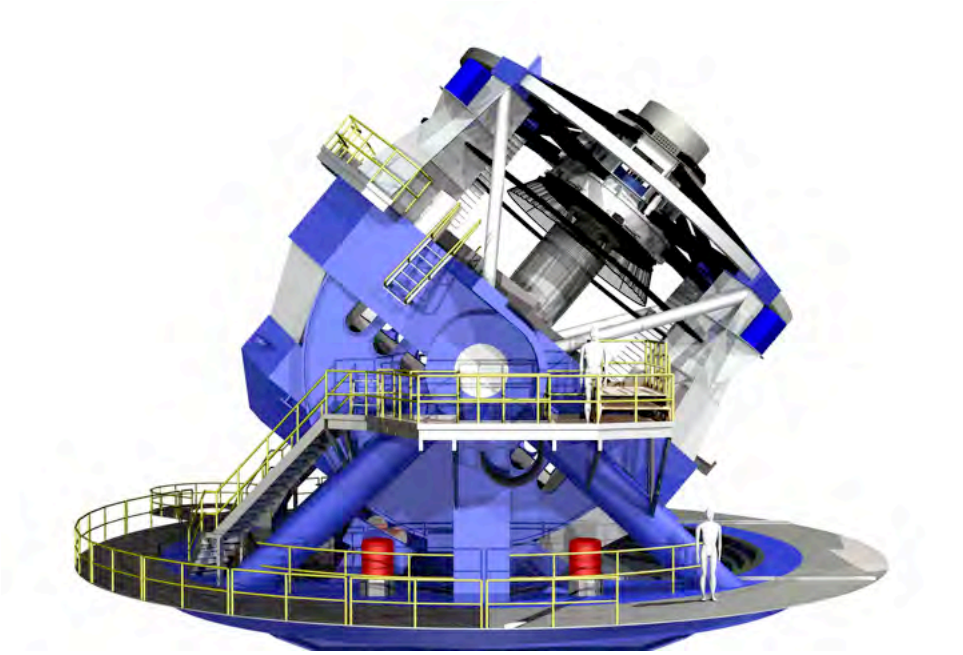


Figure 1: *The Large Synoptic Survey Telescope, in construction on the Cerro Pachon mountain in Chile, will generate 30 terabytes of data a night, every night, for ten years.*

4.2 Biology

Biology questions in the petascale computing era will drive challenges that include explaining, at a mechanistic level, interactions both within and between organisms, and developing predictive models focusing on large-scale, complex interactions such as within populations or between communities and their environments. Such predictive modeling and simulation of the dynamics of biological systems requires data-driven model building, in contrast to "first principles" simulations, where the underlying models are described by a system of equations.

The individual data sets used to build these models are currently comparatively modest in size, but they are already both challenging and dramatically growing. A single human genome is only about three gigabytes, but eventually the genome of every living human will be extracted and stored. Even now the 1000 Genome Project is attempting to sequence, and map the variations in, a thousand people from around the world; in the near future, proteomic analysis will easily generate petascale data for model building. Furthermore, the relevant data are extremely heterogeneous; as an example, Figure 2 indicates the wide variety of data types required for drug discovery.

Data-driven model construction is often considered as a combinatorial optimization problem, where a search for a model or enumeration of all feasible models with given properties is being sought. However, many existing mathematical, statistical and combina-

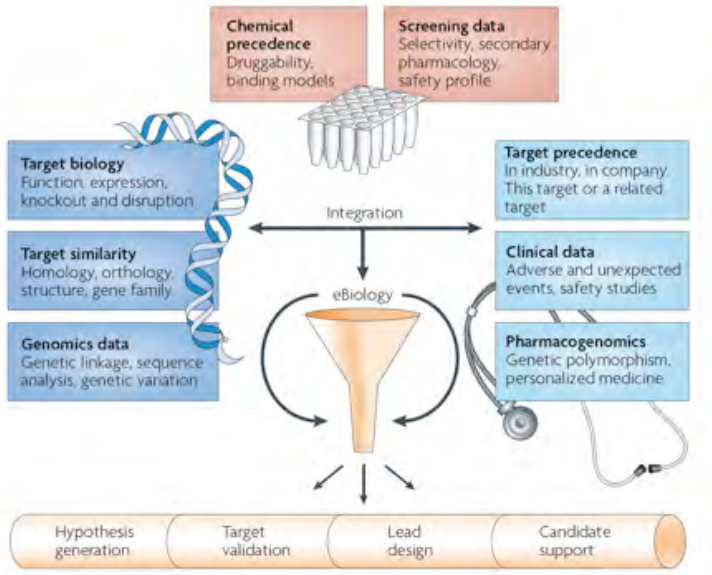


Figure 2: *The wide variety of data sources needed for drug discovery illustrates the need for the ability to analyze heterogeneous data. From* High-throughput electronic biology: mining information for drug discovery*, William Loging, Lee Harland and Bryn Williams-Jones, Nature Reviews|Drug Discovery, Volume 6, March 2007*

torial methods for building such predictive models often fail to meet the required demands of data size, heterogeneity, dimensionality, and uncertainty.

Meeting these challenges at scale will require multiscale, hierarchical, predictive mathematical and statistical models, algorithms for robust, high-throughput sequence assembly, integrated analysis of heterogeneous data (including sequence and sequence-derived data, experimental data, modeling results, imaging data, and biomolecular interaction networks), and rapid, responsive methods that can, for instance, track features in high resolution dynamic images in order to guide medical procedures in real time.

4.3 Nanoscale Chemistry and Physics

The design of materials with prescribed properties is one of the major challenges in materials science and chemistry. To this end, scientists must understand the behavior of small numbers of molecules or atoms (i.e. nano ensembles such as proteins, clusters, interfacial films) and predict physical properties from first principles. Such computationally enabled first principles design would help to address problems in fluids, biofuels, energy storage and generation, catalysis and drug delivery.

Individual data sets are not currently large; at large experimental facilities, data sets rarely exceed one gigabyte per day per user. But there are a large number of users, and

the number of users is expected to grow by at least an order of magnitude. Furthermore, the field depends on real-time data analysis, such as the ability to set up on-the-fly investigations that have 15-minute turn-around, in order to adjust experimental parameters such as pressure and temperature. Computationally generated data sets are rapidly increasing in size as well; simulation of the valence charge density for a nanoparticle is rapidly approaching 1 TB.

And even as it stands, less than 50% of the useful experimental data is fully analyzed. One of the major current barriers is the lack of parallel data mining and visualization tools to combine disparate sets of experimental and computer modeling data. This leads to difficulty in generating self-consistent models that exhibit structure and dynamics property relationships as well as emergent behavior. Furthermore, taking full advantage of these predictive models requires tools for comparing data sets, both theoretical and experimental. Here, data mining, image recognition and compression, and pattern formation techniques are clearly needed.

4.4 Power and Communication Networks

Network infrastructure is widely recognized to be vulnerable to cascading failure and attack. In most operational networks, the cost of the people and systems that manage the network exceeds the cost of the underlying equipment, and more than half of network outages are caused by operator error. What makes network operations so challenging is that the spatio-temporal information required to manage traffic must be distilled from the traffic itself, and that traffic often has both petabyte scale and extraordinary variability. This is well illustrated by Internet traffic (Figure 3): the sizes of IP flows range from a few packets to order 100,000 packets; durations last from milliseconds to seconds to minutes and beyond; link bandwidth varies from 56 Kbps dial-up connections to Mbps DSL lines to Gbps backbone links, and so on.

Yet understanding these networks is critical, and has immediate operational consequences. An example is the ground-breaking discovery that Internet traffic is long-range dependent and (asymptotically) self-similar, which elevated network measurement to an active operational role within the network infrastructure. Those measurements, collected passively or actively at end-hosts, routers, and edge devices and proxies, are currently fed back and used to improve performance, resilience, and robustness, and to enable new applications and services.

Simulation is the core technique of network analysis, as most of the relevant networks are too large, or too critical, to experiment on directly. So one great challenge is the building of accurate, analyzable network models. These models must be able to dynamically assimilate whatever network measurements are available, and model in depth all potential interactions, including feedback loops.

Yet in order to be tractable, networks with hierarchical, complex, non-linear relationships are sometimes mapped onto very simple graph models, in a fashion that fails to

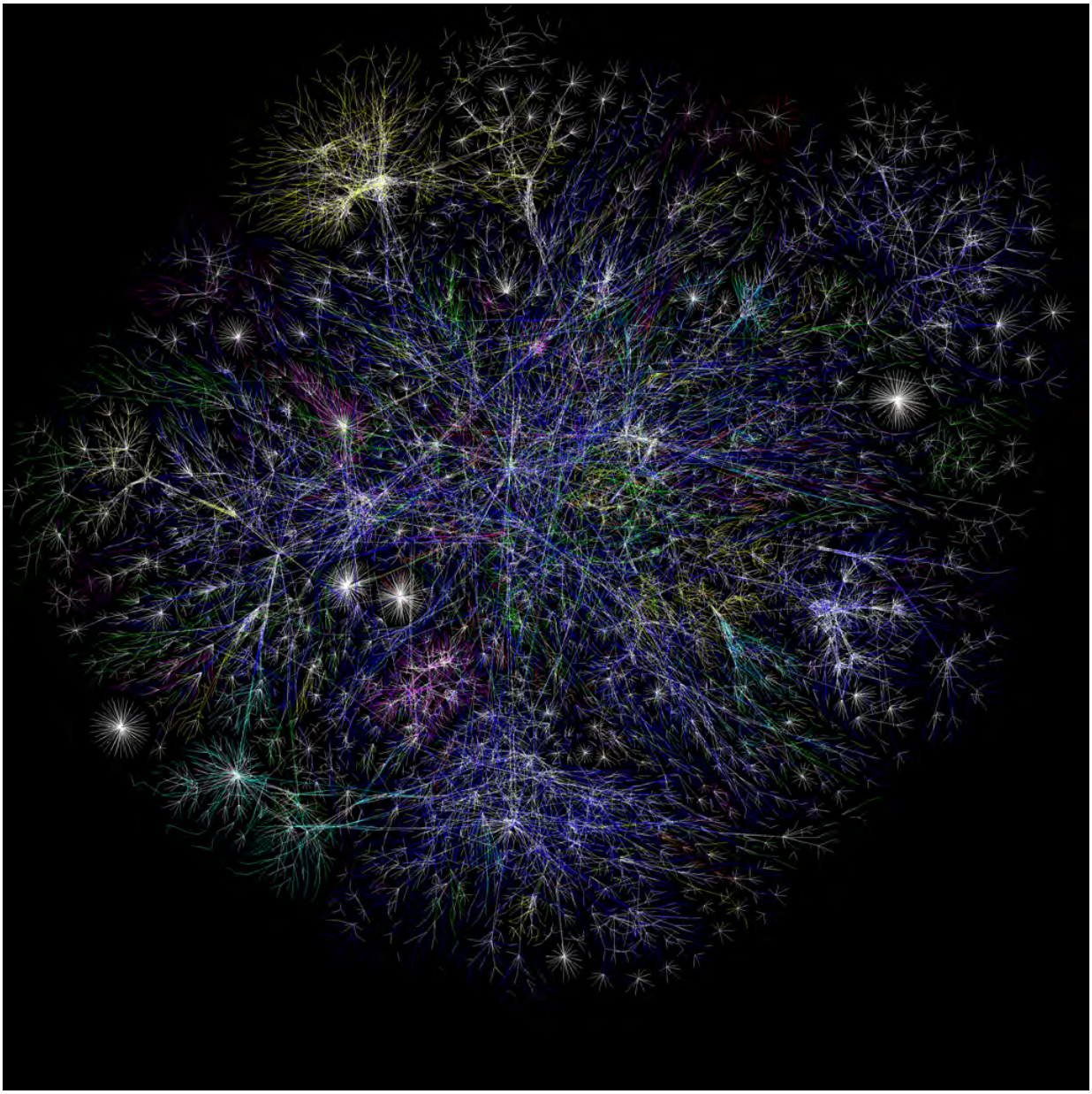


Figure 3: *A partial map of the Internet from January 15, 2005, illustrating the heterogeneity of communication path, delay, and source that must be incorporated into any model of this network. Each line is a communication path drawn between nodes representing IP addresses. The length of the lines are indicative of the delay between those two nodes. Lines are color-coded according to their corresponding RFC 1918 domain allocation. Image generated by Matt Britt from data provided by Opte Project.*

capture the numerous heterogeneous attributes of the real problem. There is therefore a need for new methods capable of operating directly on the complex graph representations, and for new metrics and analysis techniques to evaluate the fidelity of these models, against simplified versions and against "the real world".

A related need is for advances in the scalable use of current network analysis. Many of the best tools have only NP-hard implementations and fail to scale. The lack of scalability is particularly troubling when the uncertainty, or noise, in the data from which these networks are derived is taken into account. This uncertainty often calls for the study of many perturbed versions of the original network; but if a single network resists analysis, then analyzing an *ensemble* of perturbations is simply infeasible.

4.5 Earth and Climate Systems Modeling

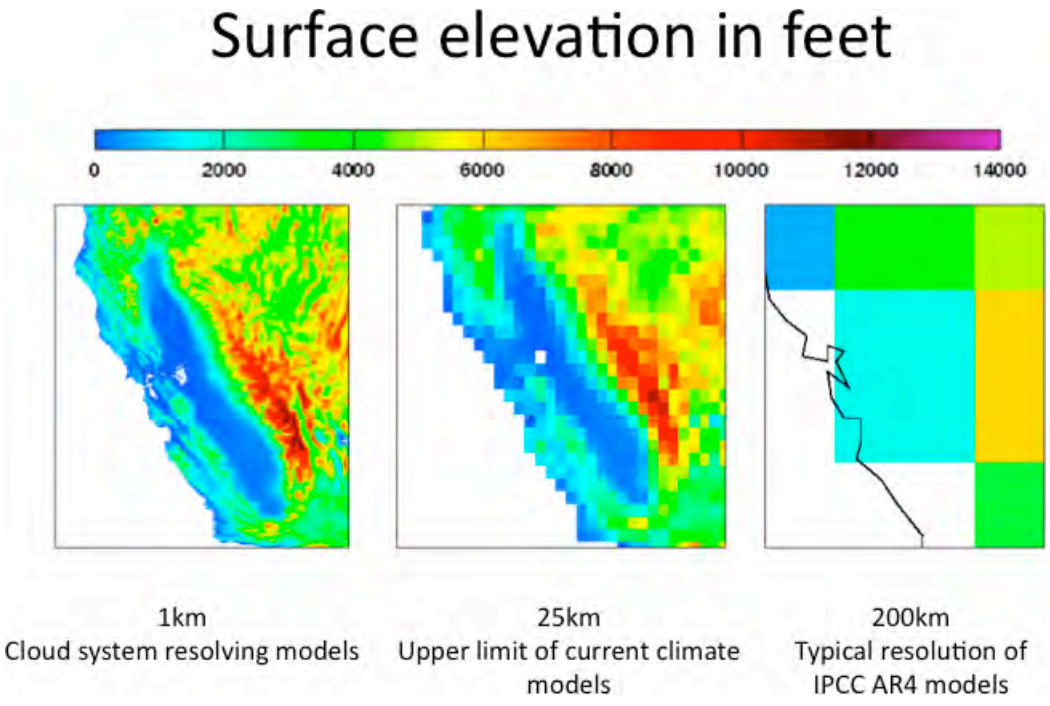


Figure 4: *Increasing resolution (and data sizes) in climate simulations results in increased accuracy in modeling cloud coverage. Courtesy Michael Wehner.*

Earth systems modeling seeks to develop a comprehensive model of the earth's past, present and future climate states. To this end, an enormous amount of data is collected or simulated. There is 110 terabytes of data at the Program for Climate Model Diagnosis and Intercomparison (PCMDI) repository at LLNL, representing 10,000 years of ensemble calculations using 25 different models. Furthermore, in pursuit of ever more accurate analysis, climate models continuously increase their resolution (Figure 4), and so also the

amount of data they generate; it is estimated that the 5th Intergovernmental Panel on Climate Change (IPCC) assessment will produce 1.6 to 5 petabytes of data. According to climate scientists there already is too much data to analyze, and in many cases the analysis of the data is more complex and takes more time to run than the original simulation.

Some of the types of analyses that earth and climate scientists cannot currently undertake include: 1) data assimilation for ocean, carbon cycle, and other long-lived greenhouse gases, 2) global climate dependencies, 3) fractional attributable risk, 4) feature extraction, 5) understanding parameter sensitivities and uncertainties, and 6) data mining for specific features. Analysis methods that scale well on large parallel architectures are clearly required even today. In addition, scientists need the ability to data mine across different data repositories, data reduction techniques that can quantify the loss incurred, compression techniques that preserve the correct distribution tails, and methods for quantification of uncertainty.

4.6 Fusion Physics

One of the outstanding problems in fusion energy is the prediction of plasma energy confinement time from first principles; this is a critical step on the road to a fusion reactor. Modeling of a fusion plasma requires the simulation of a complex system with 10^{22} particles, nonlocal (electromagnetics) and anisotropic (strong magnetic) effects, and even regions of stochastic behavior. Since these problems cannot be simulated directly, many phenomena must be approximated. The main problem then is to extract higher level quantities from the simulation results.

The scale of the data complicates that extraction, even at current data sizes. A single recent thirty hour fusion plasma simulation on a Cray XT4 (with 31,000 CPUs) resulted in a four terabyte data set. Similar problems arise in inertial confinement fusion, where simulations to model physical phenomenon such as Richtmeyer-Meshkov instabilities are so large as to fill up an entire machine. Simulation data must also be compared to experimental data in order to validate the results, but a “small” experimental data set can still be approximately one terabyte per shot.

Yet even those data are not enough. Fusion scientists are limited by the amount of data that simulations can write out, mainly because the read and write speeds on current supercomputers are too slow. Therefore, methods to pre-analyze the data and only write out what is needed are critical. In particular, scientists are especially interested in finding regions where the electromagnetic fields become stochastic, in understanding what happens “out in the tails”. Scalable tools for dimension reduction, comparison of large data sets, and parameter scans are also essential. Finally, comparison to experimental data sets poses its own set of analysis challenges, due to severe differences in representation; experimental data tends to have fewer control variables, coarser resolution in time and space, and even to be only two-dimensional.

4.7 Accelerator Physics

Accelerator physics is concerned with some of the most fundamental questions in physics, such as the origin of mass and why we are made up of matter rather than anti-matter. To answer these questions, physicists have designed and operated large accelerators such as the Relativistic Heavy Ion Collider (RHIC) and the Large Hadron Collider (LHC). The RHIC today generates roughly a petabyte of data each year, and starting in 2008 the LHC is expected to generate fifteen petabytes a year. Furthermore, within the High Energy Nuclear Physics (HENP) community, large data volumes are correlated with large collaborations, which brings requirements such as the distributed analysis of distributed data.

A major analysis issue in accelerator physics is finding extremely rare events in the presence of a dominant background signal. This leads to challenges in understanding those background signals, in reconstruction of material structure (crystallography, proton radiography for understanding tampering with nuclear stockpiles) and in real-time monitoring. An example is identifying tracks from pixels/hits, and "connecting the dots" between detectors at different planes. However, as the sizes of the data sets grow and the energy regimes change, some of the data challenges will involve developing efficient event reconstruction methods for those new regimes, techniques for simulating the collisions of strongly-coupled many-body systems, identifying jets, excess mass calculations with rare events, and reconstruction of complicated structures from diffraction patterns.

4.8 Cybersecurity

Cybersecurity is a relatively young field, and covers a number of disparate elements. These include designing or redesigning systems to be more secure and resilient (against both attacks and mistakes), developing methods for detecting and countering threats to the system, and thinking about how to operate in the face of a mix of intelligent adversaries with unknown capabilities.

Cybersecurity applications are also already deluged, every second, with far more data than they can handle, not least because every single packet that travels the Internet could conceivably be pertinent to the security of your system. Furthermore, scientific progress in cybersecurity will require not only observational data, but also modeling, and thus simulation data, and often on the same scale as the real-world networks being modeled. Even with small models, the need to run them repeatedly in order to consider responses and map out possible futures, will run the data requirements into the petascale range.

Additionally, these models are by no means straightforward to build, as they include network structure (with associated devices or infrastructures) and the activities and communications of the people, processes, and devices that use the network.

Accordingly, large scale modeling, and, to some extent, emergent behavior modeling, were identified as the cybersecurity petascale mathematics problem of highest impor-

tance. Four areas that were ranked as very important were: security measurement and quantification, distributed real-time analysis of data, trade-offs between accuracy and real-time response, and integration of analysis over widely varying scales of time and place.

4.9 Combustion

It is estimated that 85% of the U.S. energy needs are provided through the burning of fossil fuels. Combustion science therefore has implications for energy efficiency, energy security, and reduction of emissions and greenhouse gases. These are some of the primary motivations for identifying and developing more efficient combustion processes through the use of simulation models, as in Figure 5. The main goal is to develop predictive mod-

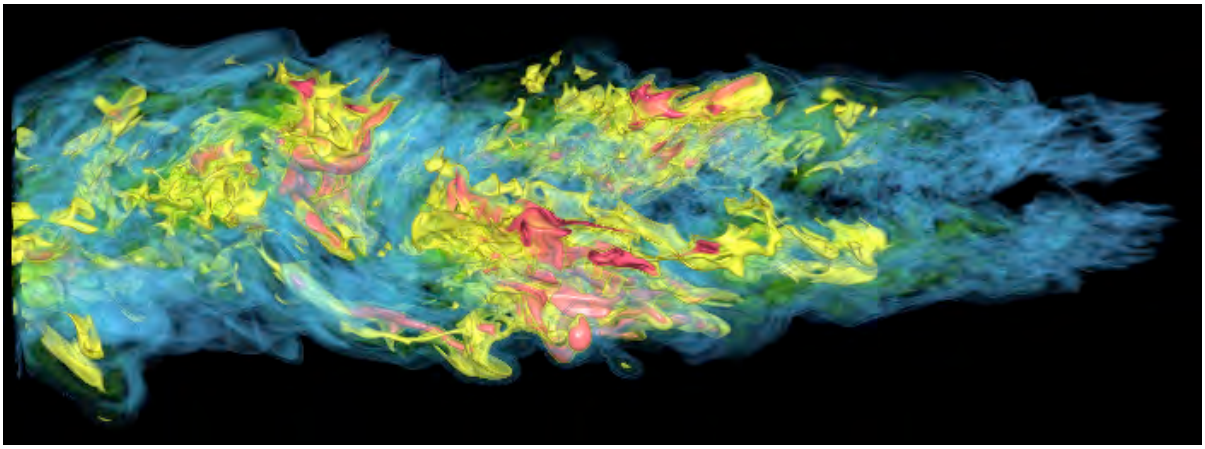


Figure 5: *Accurate modeling of complex physical phenomena requires enormous amounts of data. An example is this simulation of a lifted turbulent autoigniting ethylene/air jet flame (Reynolds no. 10,000), where formaldehyde is a marker of pre-ignition upstream of the lifted flame base. The simulation required 3.5 million CPU-hours on 30,000 CrayXT4 quad-core processors at ORNL, and produced 35 terabytes of data. Courtesy Kwan-Liu Ma, SciDAC Ultrascale Visualization Institute.*

els to better understand underlying combustion processes such as turbulence-chemistry interactions in alternative fuels, extinction and reignition, flame propagation, stratified combustion, flame stabilization in autoignitive flows, formation of soot, and emissions.

Today, a typical combustion simulation can generate an aggregate of 30–100 terabytes of data. A simulation will usually write out 100–300 restart files, with each file being on the order of 250 gigabytes. In the near future, as the *resolutions* used in simulations increase, this number is expected to reach 400 TB per simulation. In addition, the data sets will also increase due to an increase in the *complexity* of the fuels being studied. Currently, some of the time regions are manually analyzed, and automated analysis tools are used to analyze the rest.

However, the available analysis tools are already insufficient. Methods that are defined for smooth domains work well with a given error tolerance for small data sets, but as the

data set size increases the error growth becomes unacceptable. In addition, many other analysis algorithms scale poorly with the size of the data. Yet other types of analysis are completely unavailable, such as parallel topological segmentation and temporal tracking of scalar and vector features in a multi-scale turbulent reactive flow setting, or uncertainty quantification for turbulent reacting flows.

4.10 Visualization

The science of visualization is the study of how to explore and present data so as to facilitate understanding and insight. Though many of its concerns seem mathematical in nature — e.g., finding a "good" reduced-dimension projection of data — its methods are always affected (or afflicted, depending on your perspective) by human psychology. That is, "good" here doesn't mean "minimal error", it means "useful", and the two aren't always the same.

Visualization methods face all the same data scale concerns that the underlying science domains face and then some: visualization is also interactive, and so also has strict *real-time* requirements.

One of the dominant themes that emerged from the workshop discussions is the need to address noise, variability, and uncertainty in data; and therefore, visualization science is faced with the challenge of *representing* these qualities, and the results of analyses about them, in a salient fashion. Another theme was the increasing prevalence of data that is physically distributed in a fashion that is convenient for computers but irrelevant, or injurious, to human insight; visualization methods wish to present this information as if it were indeed monolithic. Finally, visualization, unlike some other domains, requires methods that can explicitly trade the speed of an analysis against its accuracy. Visualization is often characterized by periods of "dwell", when fine detail and the ability to drill down are required, and "travel", when one is quickly scanning through the data, hunting for salient regions.

Visualization scientists thus need improved methods for both data and dimension reduction, for subsampling, for "guided tours", and for uncertainty characterization. Also useful would be scalable techniques for feature extraction, machine learning, and novelty detection in general, to help detect the data worth visualizing. There is also a need for new modeling methods, to facilitate multiresolution and multidimensional visualization of all sorts of data: geometric, abstract, even relational.

5 Mathematics Research Findings

The workshop began with two days of breakout discussions designed to elicit both the data challenges faced by the applications (as reviewed in the previous section) and the gaps in the current range of mathematical tools suggested by those challenges.

The organizing themes for the mathematics breakout sessions were:

- Statistics
- Optimization
- Uncertainty quantification
- Machine learning
- Network and graph analysis
- Analysis of streaming data
- Data reduction (which includes dimension reduction, feature extraction, and topological methods).

The borders between these themes are porous and fuzzy; they are by no means a disjoint partition of the relevant methods, nor an attempt at an elegant, minimal, categorization of mathematics. Rather, the organizing committee started with a brainstorming exercise in which we listed a slew of pertinent and specific math methods that have mattered, or were likely to matter, to DOE applications. Only after a huge list of such methods was generated did we attempt to loosely combine and organize them; that process suggested the umbrella categories listed above.

On the third day, after all the breakout sessions, the workshop attendees re-gathered as a whole; first, to hear summaries from the various sessions, and then to look for common themes and to extract the major findings, which follow.

It will be evident that these findings are not organized around specific mathematics disciplines. One of the primary realizations to emerge from the workshop is that many of the gaps in capability are broad, pervasive, and apply across mathematics fields. Another realization is that these findings are not, and could not be, entirely disjoint, as there are often multiple useful perspectives on the same core idea.

We hope that findings presented in this cross-cutting fashion will be accessible to both policy makers and mathematicians, and further, may lead mathematicians to useful insight regarding how problems they are facing manifest and are addressed in other fields.

5.1 Scalability

Finding: Algorithms must be re-engineered to scale with the size of the data, which is often independent of the number of model parameters, and so may require parallel, single-pass, or subsampling methodologies.

Applications are hindered by analysis that doesn't scale. As the size of scientific data grows, examples from across a broad set of disciplines make it clear that many of

the current algorithms for analyzing data are inadequate. In accelerator physics, more efficient event reconstruction methods will need to be developed to analyze the data coming from the next generation of accelerators such as the Large Hadron Collider. Kalman filters, which are commonly used today, are computationally expensive and are mainly used to refine already detected particle tracks. For earth systems modeling, analysis tools such as the ones contained in the Climate Data Analysis Tools[7] or NCAR Command Language[8] need to be parallelized to be able to run at the scale required for the increased fidelity simulations envisioned. Before even attempting analyses of combustion data sets, scientists frequently ask how to determine if a particular tool will scale to the large data sizes from their simulations. Fusion scientists will also require scalable tools for dimension reduction, comparison of petabyte data sets, and parameter scans over the data. In nanoscience, the lack of parallel data mining and visualization tools to combine disparate sets of experimental and computer modeling data is hindering the ability to generate self-consistent models that contain structure and dynamics property relationships as well as emergent behavior.

The need for scalable algorithms is apparent across the range of mathematics disciplines:

Machine learning. Classification, clustering, association rules, anomaly detection, tracking, and time-series analysis all require scalable algorithms. Particularly attractive would be methods that can operate with a single pass through the data, and/or are able to handle distributed data in situ, as the data volume is such that data access is expensive.

Statistics. Because of the need to understand uncertainty quantification and sensitivity analysis, statistical methods will also be vital. Scalable algorithms for improved sampling, such as adaptive sampling methods that can still mimic the statistical properties of the original large data sets, will need to be developed.

Optimization: Optimization under uncertainty, which underlies many of the analysis problems in both the application areas and the enabling technologies, will also need to be adapted for large data sets. In particular, areas that need further research include methods for multi-scale optimization and both linear and nonlinear dimension reduction. Given the large dimensions of the problems, it is also possible that the notion of optimality may have to be re-defined or that methods for sub-sampling the computation of the objective or constraint functions will need to be developed.

5.2 Distributed Data

Finding: Petascale data sets are too large to easily move, yet are often non-uniformly

[7] http://www-pcmdi.llnl.gov/software-portal/cdat

[8] http://www.ncl.ucar.edu

distributed, requiring algorithms that come to the data, rather than vice versa, and can adapt to the lack of a global perspective on the data.

Petascale data are frequently distributed data. As computational science addresses existing gaps in knowledge, it moves on to increasingly complex questions. A side effect of this is that much more information is usually required to answer these new questions, and that information is often distributed across a variety of heterogeneous, multi-modal, petascale data stores.

Distributed data call for a variety of new ideas. And once distributed, data tend to stay that way. Petascale data sets have a property similar to inertia — they are relatively cheap to store, and to keep moving, but the transitions between those two states are expensive in time and hardware. Therefore the trend is more and more towards moving the analysis to the data, rather than the other way around. And in turn, this means that the analysis methods will have access only to local neighborhoods of information.

While analyzing those neighborhoods in interesting ad hoc ways is critical to advancing scientific discovery, accomplishing this requires significant advances in data analysis and integration capabilities. New mathematics methods need to be developed and evaluated to operate in this distributed environment.

In situ processing. When large-scale data are generated by a simulation, and the desired analysis is known in advance, it may be beneficial to perform the analysis immediately, as part of the simulation run, instead of as a post-processing step. In this case, only the derived (post analysis) data need to be recorded. Since the overhead of saving and reading the raw simulation data is completely avoided, this can substantially reduce the overall analysis cost. Research into techniques that allow this type of dynamic, high-performance data analysis across a variety of simulation codes could dramatically increase the efficiency of computational science.

Active storage processing. In cases where the raw information needs to be stored, it may be possible to push some of the analysis to the storage device. This process dramatically reduces the amount of information transferred from the storage device to more traditional compute nodes, at the cost of additional algorithmic complexity. Thus, new mathematics methods, utilizing the limited processing capabilities found on storage devices and combining these data subsets to efficiently generate results, are required.

Non-stationary data. Many analysis methods assume a global understanding of the data. At the very least, they assume that all the data resemble, statistically, the locally accessible data. This assumption is often violated by distributed data, especially data that are distributed over space or time. Methods are required which can detect, adapt to, and assess the impact of this non-stationarity.

Error estimation. Echoing a theme that recurs throughout this report, a critical aspect of distributed data analysis is the estimation of error in the computations. We must understand what is lost by operating on distributed data, compared to an (impossible) global analysis of the data. Though the development of approximation (known error) versus heuristic (unknown error) algorithms is highly challenging, it is nonetheless crucial;

we need distributed algorithms that can capture and characterize uncertainty.

5.3 Architectures

Finding: New algorithms should make effective use of new computer architectures that are being developed for data analysis.

Architectures optimized for simulation are not necessarily optimized for data analysis. Over the past decade, investment in scientific computing hardware and software has overwhelmingly focused on scientific simulations from "first principles", as can be seen in astrophysics, climate modeling, fusion, and materials. The result is that high-end computing hardware is relatively well configured to run scientific simulation software and to store the massive amount of data that it produces. However, fundamental differences exist between sorts of computations appropriate to "first principles" simulations and that appropriate to data analysis, especially at extreme scale. Architecture experts are thus developing next generation hardware architectures suitable for large-scale data analytics; a sampling of such architectures illustrating their varied properties and strength is in Table 1.

Platform	**Local Memory Access Latency**	**Remote Memory Access Latency**	**Programming Model**	**Type of Remote Access**
Red Storm	Commodity	Medium	MPI	Distributed Memory
XMT	Long	Long	Heavily Multithreaded	Shared Memory
Netezza	Commodity	Short (SPU to Disk), Commodity (Netezza to Netezza)	Augmented SQL	Custom Query
Commodity Cluster	Commodity	Long	MPI or PGAS	Distributed Memory
SMP	Commodity	Short	Lightly Multithreaded	Shared Memory
Multithreaded SMP (e.g., SUN Niagara)	Commodity	Short	Moderately Multithreaded	Shared Memory

Table 1: *Current and new platforms vary dramatically in programming model, remote access methods, and memory handling. This means an algorithm optimized and practical for one platform may be entirely unsuited for another. Table courtesy of Richard Murphy, Sandia National Laboratories.*

We must develop the mathematics and analysis algorithms that take the best advantage of these new architectures.

New data access patterns call for new algorithms. Fundamental differences in data context and access patterns exist among the data production and data analysis steps. Data generation generally proceeds from one time step to the next. In contrast, data analysis often requires the global context of the data, across multiple time steps, to discover not only short-range but also long-range multi-scale relationships of the phenomena under study.

Furthermore, there will undoubtedly be more applications with unstructured meshes, adaptive meshes, and structures derived from graphs. All of these will have substantially different memory access patterns than the majority of today's applications. For example, it is well known that algorithms for highly unstructured graphs are difficult to parallelize efficiently on today's distributed memory machines. The graph algorithms have high communication requirements, high degree nodes (leading to load balancing challenges), and a lack of locality that leads to poor performance within the memory hierarchies.

Parallelization everywhere. The trend in high performance computing towards multi-core and many-core architectures is noteworthy in that many, if not all, computations will need to be parallelized in the future. In addition, high performance computers have deep memory hierarchies with highly non-uniform memory access times. This will create a demand for mathematical algorithms and tools that are adapted to these computer architectures.

5.4 Data and Dimension Reduction

Finding: Analysis of petascale data in their raw form is often infeasible. Instead, improved methods for data and dimension reduction are needed to extract pertinent subsets, features of interest, or low-dimensional patterns.

As we have seen, many of the application domains important to the DOE are bedeviled by data that are high volume, high dimensional, or both.

Scalable, flexible data reduction. Many of the tools in the current mathematical analysis toolkit are extremely useful, but are inapplicable to petascale data because they do not scale. So if the tools can't come to the data, then perhaps the data, suitably slimmed, can come to the tools. Accordingly, one identified need is for scalable tools for data reduction, for extracting relevant subsets, compressed representations, or just the features of interest.

One example is compressed sensing, a novel application of constrained l_1 minimization that permits the exact reconstruction of sparse signals from what might have seemed to be incomplete measurements. Though promising, to be applicable to petabyte data it will require fast scalable algorithms for l_1 minimization. Another example is feature extraction; most existing methods are serial, by implementation and often by nature.

Note also that the features of interest may not be known or specifiable in advance. Accordingly, there is a need for methods for the identification and extraction of latent features, and for the detection of multivariate extremes.

Quantification of induced error. Equally important, however, are methods to provide estimates of the accuracy or uncertainty of the reduced data products. That is, a compression method must be able to indicate the nature and degree of loss in its compressed result, a feature detection method must be able to report its sensitivity and specificity, and so on. Constructing such metrics is not straightforward, not least because some currently useful metrics do not scale. For instance, one might want to know how a clustering method has been affected by reduction of the data. As data grow in size, it likely also grows in the number of clusters N_C, but the best and simplest cluster comparison metric requires an unscalable $O(N_c^2)$ calculations.

Reduction under constraints. Assuming that suitable accuracy metrics can indeed be established, a particularly attractive property of any reduction method would be the ability to adjust the balance between accuracy and degree of reduction. But accuracy is just one of many possible desired properties; other relevant criteria might be smoothness, positivity, lossless preservation of pre-identified regions, explicability, and so on. So an identified need is for techniques that enable reduction under constraints.

The challenges of high dimensional data. High dimensional data complicate the application of both human intuition and our current analytic tools. Humans have enough difficulty visualizing in three dimensions. For tens of dimensions, techniques such as the "grand tour" have helped, but those methods break down in the face of hundreds or thousands of dimensions. Similarly, high dimensional data complicate the use of many methods in mathematical disciplines such as machine learning, linear algebra, and optimization. All of which pose a serious mathematical challenge: how can we reliably detect and classify low-dimensional patterns embedded in high-dimensional spaces, while not being defeated by the curse of dimensionality?

5.5 Models

Finding: Using data to make predictions or discover new science requires methods to build and evaluate appropriate models of large-scale, heterogeneous, high-dimensional data.

In many application domains, there is a need to build various sorts of models. Some models are predictive, using examples of several known categories of objects to build a model that can predict the category of a new object. Others are descriptive, grouping similar objects together and then determining a description for each group. Other models are simplified versions of complex systems, such as a computer network, the power grid (Figure 6), or a computer simulation of a complex phenomena.

Models must integrate huge, heterogeneous, high-dimensional data. Regardless of the type of model built, several application domains are integrating a variety of

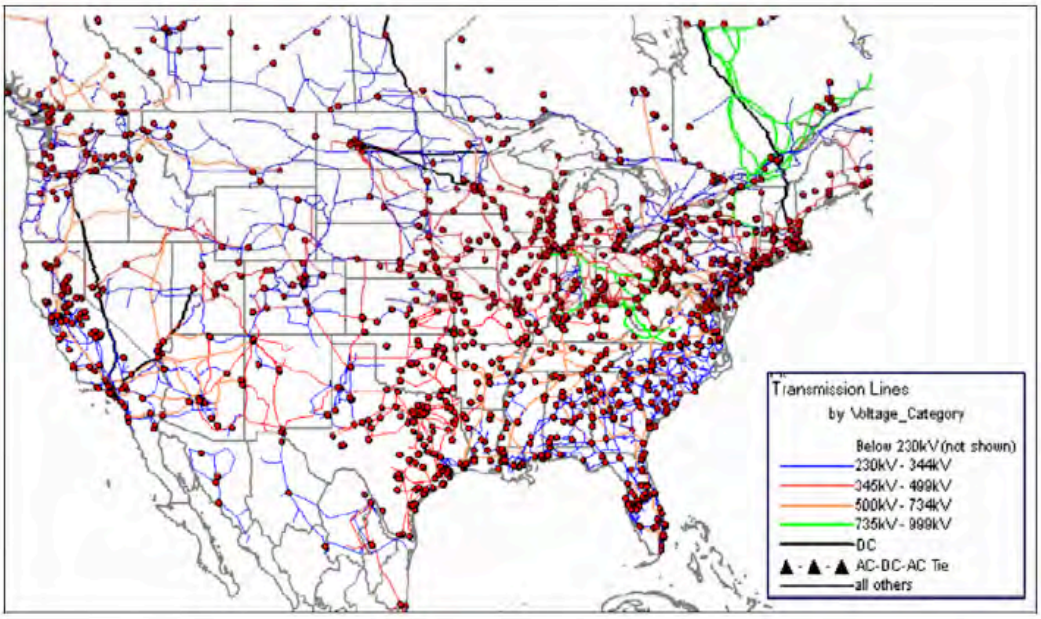


Figure 6: *The US electric transmission system is an example of a network which is both critical and extremely challenging to model accurately. Courtesy North American Reliability Corporation.*

information sources into their models, resulting in data that are both voluminous and heterogeneous. The variables of interest in a power grid could be sampled at different rates, the images in an astronomical survey could come from different types of telescopes, and a biology application may need to combine categorical data with numerical data. In addition, the data used to build the models is often high-dimensional, with thousands of features describing a microarray experiment or hundreds of sensors monitoring a fusion experiment. To fully exploit all the available data and thus build the most accurate models, there is a need for algorithms that can handle the heterogeneity and high-dimensionality of the data.

Model validation is critical. The models built from the data must of course be scientifically meaningful. Validation of the models therefore plays an important role, and models that are interpretable and robust to small changes in the data are often preferred. An associated mathematical challenge is building an accurate model when the data are of poor quality (due to noise from the sensors and other sources), have missing values (due, e.g., to inoperational sensors), or are unbalanced (with many examples from one category and few from the category that is of interest). Techniques that allow the incorporation of, and validation against, even such messy domain knowledge would result in more accurate and meaningful results.

Network models are complicated by heavy-tailed distributions. Finally, there is a need to develop a new mathematical foundation for network science, one grounded in em-

pirical data and enabling design and control of efficient, stable and secure networks. The issue is that we lack fundamental understanding of high-variability network phenomena, and therefore we lack models that are adequate for the purpose of network engineering and control. We need mathematical methods that capture in a parsimonious manner key characteristics commonly associated with high variability. One such characteristic is the 80-20 rule, which states that while most values of a heavy-tailed distribution are small, the probability of an extremely large event is non-negligible. Another is the lack of a typical value; mathematical means and statistical averages are non-informative in cases that include heavy-tailed distributions where variances are either very large or, in fact, infinite, and variance estimates converge either very slowly or fail to converge altogether.

5.6 Uncertainty

Finding: Interpreting results from petascale data requires methods for analyzing and understanding uncertainty, especially in the face of messy and incomplete data.

Or: we don't know how to do error bars at petascale.

The quantitative understanding of uncertainty is essential when predictions are used to inform decisions and policy. Uncertainty characterization casts a broad net that encompasses both the assessment of confidence in predictions and the analysis of messy data. While neither of these issues is unique to petascale data analysis, both require the development of new methods and techniques to address the volume and complexity of the data.

Characterizing predictive uncertainty. Predictive uncertainty is associated with the combined effects of limitations in sensitivity and accuracy of physical measurements, incomplete understanding of the underlying physical processes, the complexity of coupling different physical processes across large scales, and the numerical errors associated with simulations of complex models. The mathematical challenges in characterizing predictive uncertainty include choosing appropriate uncertainty representations and analyzing the effects of uncertainty on predictions and model calibration.

Visualizing uncertainty. Visualization is a key tool in the characterization of predictive uncertainty, as it provides both researchers and decision-makers with insights about high-dimensional input and output spaces. For petascale data, analytical tools that guide visualization — for example, by automatically identifying interesting low-dimensional projections of high-dimensional data — are critical. Can scalable versions of visualization tools like exploratory data analysis and grand tour/projection pursuit be developed?

Design of physical and computational experiments. At any scale, the fundamental challenge of understanding variability and uncertainty with high-dimensional, complex models is the limited number of runs available. Given the explosion of observational and simulation technology that offers many options for data collection and generation, researchers must answer the question of how to design experiments, both physical and computational, to optimally collect data. The traditional design of experiments is con-

cerned with allocating trials within a single, typically physical, experiment. As computation has emerged as an experimental tool, new methods have been developed to optimize data collected from computational experiments as well. The mathematical challenge is how to combine and scale existing methods to make accurate predictions, understand uncertainty, and allow sequential choice of the next set of runs to make.

Sensitivity analysis. Sensitivity analysis is the organized study of how the output of a model responds to variations in model inputs (parameters, initial and boundary conditions) and in the model itself. Inputs to a model are subject to many sources of uncertainty, variability, and measurement error. The model itself may be subject to uncertainty arising from incomplete information or poor understanding of the physical processes and driving forces. There are no general methodologies for treating the effects of variation and uncertainty in high-dimensional parameter spaces that include strong interdependencies between parameters and large-scale ranges in parameter values. Sensitivity analysis of a complex system typically benefits from a fusion of mathematical and statistical techniques, yet mathematical frameworks for combining such approaches in a unified fashion do not yet exist.

Optimization under uncertainty. Often, analysis of petascale data results in optimization problems. After obtaining a solution, a scientist usually wishes to know how the result varies with uncertainty in both the model and the input parameters to the model. These problems are usually formulated as either stochastic programming problems or robust optimization problems. Both of these are relatively new areas for optimization researchers and limited mostly to linear and quadratic functions. Another example is machine learning; a support vector machine can be formulated as a linear programming robust optimization problem with uncertainty in the data. It is well known, however, that small errors in the input data can make optimal solutions highly infeasible. Stochastic programming problems have similar issues, not the least of which is the large number of realizations that must be computed to obtain a solution. In order to analyze petascale data, several challenges will have to be addressed including specifying the problem and the uncertainty in a computationally tractable form and extending existing methods to the more general nonlinear case.

5.7 Outliers

Finding: Near real-time identification of anomalies in streaming and evolving data is needed in order to detect and respond to phenomena that are either short-lived (e.g., supernova onset) or urgent (e.g., power grid disruptions).

Streaming data are common, and challenging. A common task in many of the applications considered during the workshop is the identification of unusual events or anomalies in streaming data that are collected over time. Data from an astronomical telescope are monitored as it is collected so that any unusual phenomenon can be identified quickly and resources devoted to study it further. Similarly, data from sensors on a power grid or a

physics experiment may indicate the increased likelihood of imminent failure, which, if not identified on time, could be catastrophic. The size of the data, and the rate at which they are acquired, are obvious challenges to the near-real-time analysis of streaming data.

Defining and detecting anomalies. There is thus a need for anomaly detection algorithms that can process multiple data streams, sampled at different rates and times, from different sensors, with potentially missing values. A particular challenge is to define what constitutes an anomaly, so that we do not miss any real anomalies and, at the same time, minimize the false positives. This challenge is complicated by the fact that the term "anomaly" is here intended to stand in for a variety of related concepts, including "outlier", "novelty", and/or "change".

Non-stationarity and spatial data. Furthermore, these are not the only challenges in the analysis of streaming data. In many problems, the data are dynamic and non-stationary, with statistical characteristics that change over time. If the definition of what constitutes an anomaly also changes with time, a phenomenon called concept drift, then the identification of the anomalies becomes even more difficult. Finally, some data are effectively "streaming", but spatially, rather then temporally. That is, it is distributed over space and is non-stationary because the underlying data model changes with location, as in geographic information systems (GIS) data or scientific simulation data distributed across processors. It is still "streaming" in the sense that the data are too voluminous to visit more than once, but the spatial organization also means that there is not a single preferred path through the data, which poses its own complications.

5.8 Culture

Finding: Effective mathematics research requires infrastructure, recognition of contributions, and incentives for efficient exchange and persistent storage of knowledge, both internally, within the mathematics community, and externally, to the application communities.

This finding reflects the importance of DOE support for evolving the current environment into one that will support efficient exchange of knowledge. To meet the challenges of petascale data analysis, researchers need mechanisms for efficiently accessing the current state-of-the-art across a wide variety of mathematics disciplines. There are many essentially off-the-shelf tools available for collaboration and research networking, but there is still nonetheless a need to actually deploy such an infrastructure, and to inspire the cultural changes to make the best use of it.

Barriers and incentives for social networks. As the computational and analysis environments become increasingly complex, large groups of scientists and mathematicians need to easily identify and utilize relevant work from other domains, often in ways unanticipated by the original researchers. Moving from traditional inter-disciplinary project teams, in which relatively small groups of people located at the same site work together, these groups will be much larger, distributed across many universities and DOE labs. In

addition, these individuals will be working on different projects, with disparate goals and deliverables.

As a result, incentives are key. It is critical that math researchers both share their results with each other *and* work closely with the application communities. This will happen only if there is an external driver, such as community recognition, meeting a sponsor requirement, or joint funding. The programmatic motivation for ensuring this work is shared is that the algorithms, techniques, and implementations developed by specific researchers are inherently useful to a variety of applications, and thus the efficiency of the overall research community greatly benefits from their dissemination.

Social networking features. The ultimate goal of this infrastructure would be a social networking capability for the mathematics and applications research communities, beyond that provided by personal web pages. The aim is to support and encourage dissemination of algorithms and implementations, discussion between colleagues, and feedback. In particular, this infrastructure needs to support: flow of ideas across domains, recognition and traceability of contributions, dissemination of research result and code, and mechanisms for encouraging collaboration. A final important aspect is the creation of community benchmarks, and this is another context in which working with the application communities is key, since it is important that these benchmarks both reflect real applications problems and support mathematical analysis.

5.9 Mathematics Is Critical

Finding: Mathematics is a critical part of the path from data to decision making; it leverages and completes investments in networking, architectures, and visualization.

The overall challenge of data at petascale is twofold: first, to understand what information is necessary to go from data to decision; second, to understand how that information is obtained and applied. The data waterfall (see Figure 7) captures both the necessary competencies and interdependencies, and the perspective that it is not possible to deliver insight if a single link is missing.

The data waterfall is not domain specific, and once in place it amplifies the value of domain expertise by accelerating the time to go from data to decision. The DOE Office of Science has made key investments in networking, architectures, and visualization; similarly investing in mathematics and analysis will complete the chain, and multiply the return from these investments.

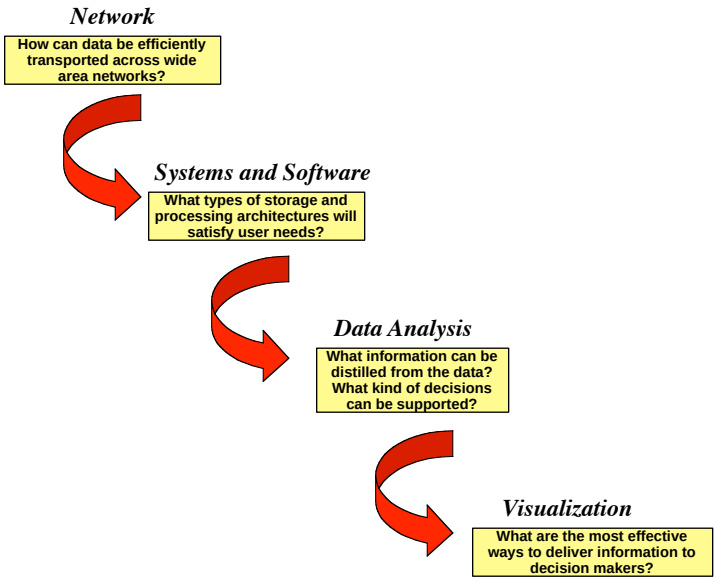


Figure 7: *The Data Waterfall illustrates the path from data to decision.*

6 Summary of Findings

To recap, the primary themes and findings of the workshop assert the need to:

- **Adapt analysis methods to the requirements of scientific petascale data**
  - Algorithms must be re-engineered to scale with the size of the data, which is often independent of the number of model parameters, and so may require parallel, single-pass, or subsampling methodologies.
  - Petascale data sets are too large to easily move, yet are often non-uniformly distributed, requiring algorithms that come to the data, rather than vice versa, and can adapt to the lack of a global perspective on the data.
  - New algorithms should make effective use of new computer architectures that are being developed for data analysis.
- **Develop new mathematics to extract novel insights from complex data**
  - Analysis of petascale data in their raw form is often infeasible. Instead, improved methods for data and dimension reduction are needed to extract pertinent subsets, features of interest, or low-dimensional patterns.
  - Using data to make predictions or discover new science requires methods to build and evaluate appropriate models of large-scale, heterogeneous, high-dimensional data.

  - Interpreting results from petascale data requires methods for analyzing and understanding uncertainty, especially in the face of messy and incomplete data.
  - Near real-time identification of anomalies in streaming and evolving data is needed in order to detect and respond to phenomena that are either short-lived (e.g., supernova onset) or urgent (e.g., power grid disruptions).

- **Support a research environment that recognizes the challenges of petascale data analysis**
  - Effective mathematics research requires infrastructure, recognition of contributions, and incentives for efficient exchange and persistent storage of knowledge, both internally, within the mathematics community, and externally, to the application communities.
  - Mathematics is a critical part of the path from data to decision making; it leverages and completes investments in networking, architectures, and visualization.

There are a welter of challenges facing the application domains that matter to the DOE, as scientists gear up to handle petascale data and beyond. We believe that substantial and sustained support for the findings above will allow the Applied Mathematics Program to address those challenges in a focused and significant fashion.

A Workshop Attendees

Participant	Institution	Role
Ghaleb Abdulla	LLNL	Scribe
Deb Agarwal	LBNL	Moderator
Mine Altunay	Fermi	Scribe
Kirk Borne	George Mason University	Speaker, Scribe
Kevin Bowyer	University of Notre Dame	Scribe
David Brown	LLNL	
Emery Brown	MIT/Mass. Gen. Hospital	Keynote Speaker
Robert Calderbank	Princeton University	Organizing Committee, Moderator
Rick Chartrand	LANL	
Bill Collins	LBNL	Speaker
Terence Critchlow	PNNL	Organizing Committee, Moderator, Scribe
Jim Davenport	BNL	Scribe
Ian Dobson	University of Wisconsin	
Danny Dunlavy	SNL	
Deb Frincke	PNNL	Moderator
Roger Ghanem	University of Southern California	Scribe
Alex Gray	Georgia Institute of Technology	Speaker, Scribe
Larry Hall	University of South Florida	Moderator
Forrest Hoffman	ORNL	
Mac Hyman	LANL	Scribe
Leland Jameson	National Science Foundation	Organizing Committee, Moderator
Cathy (Yu) Jiao	ORNL	Scribe
Ken Joy	UC Davis	
Chandrika Kamath	LLNL	Organizing Committee, Moderator
George Karniadakis	Brown University	
George Karypis	University of Minnesota	Speaker, Scribe
Philip Kegelmeyer	SNL	Organizing Committee Chair
Jon Kettenring	Drew University	Moderator
Siddhartha Khaitan	Iowa State University	
Tammy Kolda	SNL	
Vipin Kumar	University of Minnesota	Speaker
Ryan Lance	National Security Agency	
Duncan Temple Lang	UC Davis	Scribe
John (JZ) Larese	University of Tennessee/ORNL	Moderator

Participant	Institution	Role
Steven Lee	DOE ASCR	
Sven Leyffer	ANL	Moderator
Jim McCalley	Iowa State University	Moderator
Juan Meza	LBNL	Organizing Committee, Moderator, Scribe
Dave Morrison	BNL	Moderator
Habib Najm	SNL	Moderator, Scribe
Thomas Ndousse-Fetter	DOE ASCR	
Peter Nugent	LBNL	Moderator
Chris Oehmen	PNNL	Moderator
Shmuel Oren	UC Berkeley	
George Ostrouchov	ORNL	Moderator, Scribe
Bruce Palmer	PNNL	
Valerio Pascucci	LLNL/UC Davis	Scribe
Donald Petravick	Fermi	
Dan Rokhsar	DOE Joint Genome Institute	Speaker
Chuck Romine	OSTP	
Steve Sain	NCAR	Scribe
Nagiza Samatova	NCSU/ORNL	Organizing Committee, Moderator
Thomas Schulthess	ORNL	Speaker
David Scott	Rice University	Moderator
Devinder Sivia	Rutherford Appleton Laboratory	Speaker
Mitch Smooke	Yale University	Moderator
Linda Sugiyami	M.I.T.	Moderator
Charles Tong	LLNL	
Susan Turnbull	DOE ASCR	
Richard Wagener	BNL	
Homer Walker	DOE ASCR	Welcoming Speaker
Michael Wehner	LBNL	Speaker
Paul Whitney	PNNL	Moderator
Alyson Wilson	LANL	Organizing Committee, Moderator
Andrew Wilson	SNL	
Philip Yu	University of Illinois of Chicago	

See Appendix B for detailed notes on the titles of the speaker's talks and on who moderated and scribed for which session.

A key to the institutions:

ANL	Argonne National Laboratory
ASCR	DOE Advanced Scientific Computing Research
BNL	Brookhaven National Laboratory
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory
Fermi	Fermi National Accelerator Laboratory
NCAR	National Center for Atmospheric Research
OSTP	Office of Science and Technology Policy
ORNL	Oak Ridge National Laboratory
PNNL	Pacific Northwest National Laboratory
SNL	Sandia National Laboratories

B Workshop Agenda

Tuesday, June 3, 2008:

7:00–8:00 Registration and Continental Breakfast

8:00–8:20 Welcome. *State of DOE Mathematics Program, Vision for Mathematics for Petascale Data*, Homer Walker, DOE

8:20–8:30 Agenda and process description. Philip Kegelmeyer, Sandia National Laboratories.

8:30–9:30 Application focus: Astrophysics.

- Application domain challenges and issues. Speaker: Kirk Borne, George Mason University. Title: "*Data Science Challenges from Distributed Petascale Astronomical Sky Surveys*".
- Mathematics commentary. Speaker: Alex Gray, Georgia Tech. Title: "*Computational Mathematics for Large-Scale Data Analysis*".

9:30–10 Break

10–11 Application focus: Networks

- Application domain challenges and issues. Speaker: George Karypis, University of Minnesota.
- Mathematics commentary. Speaker: *also* George Karypis, University of Minnesota.
- Title of the combined talk: "*Drug and Probe Discovery and its Mathematical Challenges*".

11–12 Application focus: Nano/Chemistry

- Application domain challenges and issues. Speaker: Thomas Schulthess, Oak Ridge National Laboratory. Title: "*Understanding emergent properties in nanoscale systems: anticipating the role of petascale computing and data analysis.*"
- Mathematics commentary. Speaker: Devinder Sivia, ISIS/RAL. Title: "*Data analysis in condensed matter science*".

12–1 Lunch

1–2:30 First Application Breakout Sessions. Objective: go into detail, listing the mathematics challenges and needs of each domain. Capture in a short list to be distributed to the mathematics breakouts.

1. *Astrophysics*. Moderator: Peter Nugent, Lawrence Berkeley National Laboratory. Scribe: Kevin Bowyer, Notre Dame.

2. *Networks*. Moderator: Jim McCalley, Iowa State University. Scribe: George Karypis, University of Minnesota.
3. *Fusion Physics*. Moderator: Linda Sugiyami, MIT. Scribe: Juan Meza, Lawrence Berkeley National Laboratory.
4. *Combustion*. Moderator: Mitch Smooke, Yale. Scribe: Habib Najm, Sandia National Laboratories.
5. *Cybersecurity*. Moderator: Deb Frincke, Pacific Northwest National Laboratory. Scribe: Mine Altunay, Fermi National Accelerator Laboratory.

2:30–3 Break

3–4:30 First Mathematics Breakout Sessions. Objective: Extract, list, and prioritize the math challenges gleaned from the problem statement presentations.

1. *Statistics, Part 1*. Moderator: Jon Kettenring, Drew University. Scribe: George Ostrouchov, Oak Ridge National Laboratory.
2. *Data Reduction, Part 1: Dimension Reduction.* Moderator: Lee Jameson, National Science Foundation. Scribe: Cathy Jiao, Oak Ridge National Laboratory.
3. *Uncertainty Quantification, Part 1*. Moderator: Habib Najm, Sandia National Laboratories. Scribe: Roger Ghanem, University of Southern California.
4. *Optimization Part 1*. Moderator: Sven Leyffer, Argonne National Laboratory. Scribe: Duncan Temple Lang, University of California at Davis.
5. *Graph/Network Analysis Part 1*. Moderator: Paul Whitney, Pacific Northwest National Laboratory. Scribe: Mac Hyman, Los Alamos National Laboratory.
6. *Machine Learning, Part 1*. Moderator: Larry Hall, University of South Florida. Scribe: Kirk Borne, George Mason University.

4:30–5 Regroup and process check. Anything to tweak for tomorrow? Any emergent topics for one of the uncommitted breakouts? Philip Kegelmeyer, Sandia National Laboratories.

Wednesday, June 4, 2008:

7:00-7:45 Continental Breakfast

7:45–8:00 Agenda and process review. Philip Kegelmeyer, Sandia National Laboratories.

8–9 Application focus: Biology.

- Application domain challenges and issues. Speaker: Dan Rokhsar, University of California at Berkeley. Title: *"Genomes, Genetics, and Diversity"*.

- Mathematics commentary. Speaker: Vipin Kumar, University of Minnesota. Title: *"Application of Association Patterns Mining Techniques to Genomic Data"*.

9–10 Application focus: Earth System Modeling

- Application domain challenges and issues. Speaker: Bill Collins, University of California Berkeley and Lawrence Berkeley National Laboratory. Title: *"Extreme Climate Change: Scaling Laws, and Scale Invariance".*
- Mathematics commentary. Speaker: Michael Wehner, Lawrence Berkeley National Laboratory. Title: *"Challenges in the analysis of petascale climate model output data sets".*

10–10:30 Break

10:30–12 Second Application Breakout Sessions. Objective: go into detail, listing the mathematics challenges and needs of each domain. Capture in a short list to be distributed to the mathematics breakouts.

1. *Biology.* Moderator: Chris Oehmen, Pacific Northwest National Laboratory. Scribe: Ghaleb Abdulla, Lawrence Livermore National Laboratory.
2. *Nanoscale Chemistry and Physics.* Moderator: John Larese, University of Tennessee. Scribe: Jim Davenport, Brookhaven National Laboratory.
3. *Accelerator Physics.* Moderator: Dave Morrison, Brookhaven National Laboratory. Scribe: Alex Gray, Georgia Tech.
4. *Visualization.* Moderator: Valerio Pascucci, Lawrence Livermore National Laboratory. Scribe: Terence Critchlow, Pacific Northwest National Laboratory.
5. *Earth System Modeling.* Moderator: Deb Agarwal, LBNL. Scribe: Duncan Temple Lang, University of California at Davis.

12–1 Lunch

1–2:30 Second Mathematics Breakout Sessions. Objective: Extract, list, and prioritize the math challenges gleaned from the problem statement presentations, 4 parallel sessions.

1. *Streaming data analysis.* Moderator: Terence Critchlow, Pacific Northwest National Laboratory. Scribe: Ghaleb Abdulla, Lawrence Livermore National Laboratory.
2. *Statistics, Part 2.* Moderator: Alyson Wilson, Los Alamos National Laboratory. Scribe: Steve Sain, National Center for Atmospheric Research.
3. *Data Reduction, Part 2: Feature Extraction and Tracking.* Moderator: George Ostrouchov, Oak Ridge National Laboratory. Scribe: Roger Ghanem, University of Southern California.

2:30–3 Break

3–4:30 Last Mathematics Breakout Sessions. Objective: Extract, list, and prioritize the math challenges gleaned from the problem statement presentations, 6 parallel sessions.

1. *Uncertainty Quantification, Part 2.* Moderator: David Scott, Rice University. Scribe: Steve Sain, National Center for Atmospheric Research.
2. *Optimization Part 2.* Moderator: Juan Meza, Lawrence Berkeley National Laboratory. Scribe: Juan Meza, Lawrence Berkeley National Laboratory.
3. *Graph/network Analysis, Part 2.* Moderator: Robert Calderbank, Princeton University. Scribe: Kevin Bowyer, Notre Dame.
4. *Machine Learning, Part 2*. Moderator: Chandrika Kamath, Lawrence Livermore National Laboratory. Scribe: Terence Critchlow, Pacific Northwest National Laboratory.

4:45–5 Regroup and process check. Anything to tweak for tomorrow? Any issues to be addressed tonight to ease tomorrow's reports? Philip Kegelmeyer, Sandia National Laboratories.

6:00 Conference Dinner, in the Regency Room of the Hilton Rockville.

7:00 Keynote presentation, Prof. Emery Brown of MIT and Harvard Medical School, *"Dynamic Signal Processing Analysis of Brain Function."*

Thursday, June 5, 2008:

7:00-8:00 Continental Breakfast

8:00-10:00: Reports and discussion from the seven mathematics discipline areas.

10:00–10:30 Break

10:30–12 General discussion, compilation of major themes and conclusions, drafting of report outline.

Noon: Adjourn.

1–4: Writing session for organizing committee.

C Questions for Applications Breakouts

Each application breakout was tasked to answer one primary question:
Given the discussion prompted by the questions below, what is a prioritized list of the specific challenges faced in the analysis of data from this field?

Q1: What are the high level science motivations for work in this field?

Q2: What are some of the different science questions addressed in this field?

Q3: Does this domain involve simulations? Experiments? Both? If both, how do simulation and experiment influence each other?

Q4: Are there any real-time requirements for the analysis of this domain's data? What are they?

Q5: Are there any time vs accuracy trade-offs for the analysis of this domain's data? What are they?

Q6: Are there any other sorts of pertinent trade-offs required for the analysis of this domain's data?

Q7: How much of the data currently being collected or generated is analyzed? How much of it needs to be?

Q8: If some of the relevant data is going unanalyzed, why is that? What are the barriers to analysis?

Q9: How big is the data now? How big will it be in 5, 10 15 years?

Q10: What are the kinds of analysis this domain's scientists would like to do, but cannot, due to lack of tools or techniques?

Q11: Given what can be foreseen in 5, 10, 15 years, what analysis tools and techniques are working now but will soon break down? What will be the first analysis tool to break? What will be the most critical one to break?

Q12: What other analysis issues are important but are not being brought out by the above questions?

D Questions for Mathematics Breakouts

Each mathematics breakout was tasked to answer two primary questions[9]:

Primary: Given the discussion prompted by questions Q1–Q8 below,

A) *What is a prioritized list of the specific gaps faced by this field in attempting to address the aggregate of application data challenges?*

B) *For each gap, are there any recommendations to make that are more specific than "fill the gap"?*

Q1: From the perspective of this math discipline, when considering the listed challenges, are there common themes across application domains? Can the challenges be clustered?

Q2: The application break-outs have ranked their challenges by priority. Can we rank them by difficulty, from the perspective of our methods? Are there high-payoff, low effort lines of attack? Conversely, are there challenges that are just so challenging, or impossible, that addressing them would be fruitless?

Q3: What are the qualities by which our methods are evaluated? (Robustness? Stability? Interpretability? What?) Which of those qualities are threatened or undermined by the challenges?

Q4: How do our methods change in the face of whether the data is simulated, experimental, or observational? Are we better at handling one sort of data than other? Does that point at a gap worth considering?

Q5: What are the scalability properties of our methods? Are they appropriate for the challenges?

Q6: What are the metrics by which our methods are evaluated? How does the choice of metric influence the methods? Do the data challenges motivate new metrics? (As an example, in machine learning, F-measure is a much better performance metric for imbalanced data than overall accuracy.)

Q7: Similarly, does the nature of the challenges affect or undermine our ability to put performance bounds on our methods?

Q8: What other math issues are relevant to delineating appropriate approaches and identifying gaps, but are not being brought out by the above questions?

[9] Read "methods" as shorthand for "the mathematical methods, approaches, and techniques of this session's domain", and read "challenges" as shorthand for "the aggregate of data analysis needs, requirements, or challenges heard so far, particularly from applications discussed in the current day".