Overview of the MAPD Workshop
June 3–5, 2008, Rockville, Maryland

The Organizing Committee

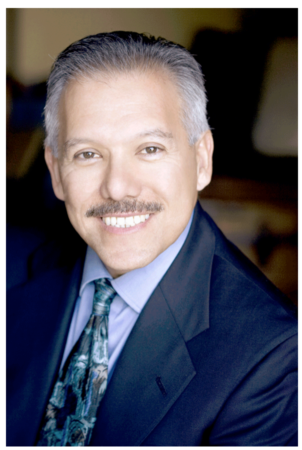
Philip Kegelmeyer
SNL/CA, Chair

Robert Calderbank
Princeton

Terence Critchlow
PNNL

Leland Jameson
NSF

Chandrika Kamath
LLNL



Juan Meza
LBNL

Nagiza Samatova
NC State/ORNL

Alyson Wilson
LANL

Demographics

- 66 attendees: 57 invitees, 9 observers
- 30 from DOE labs, 27 from academia, 6 from DOE ASCR, one each from NSA, Office of Technology Policy, NSF
- 27 attendees with an application focus, 30 with a math focus.

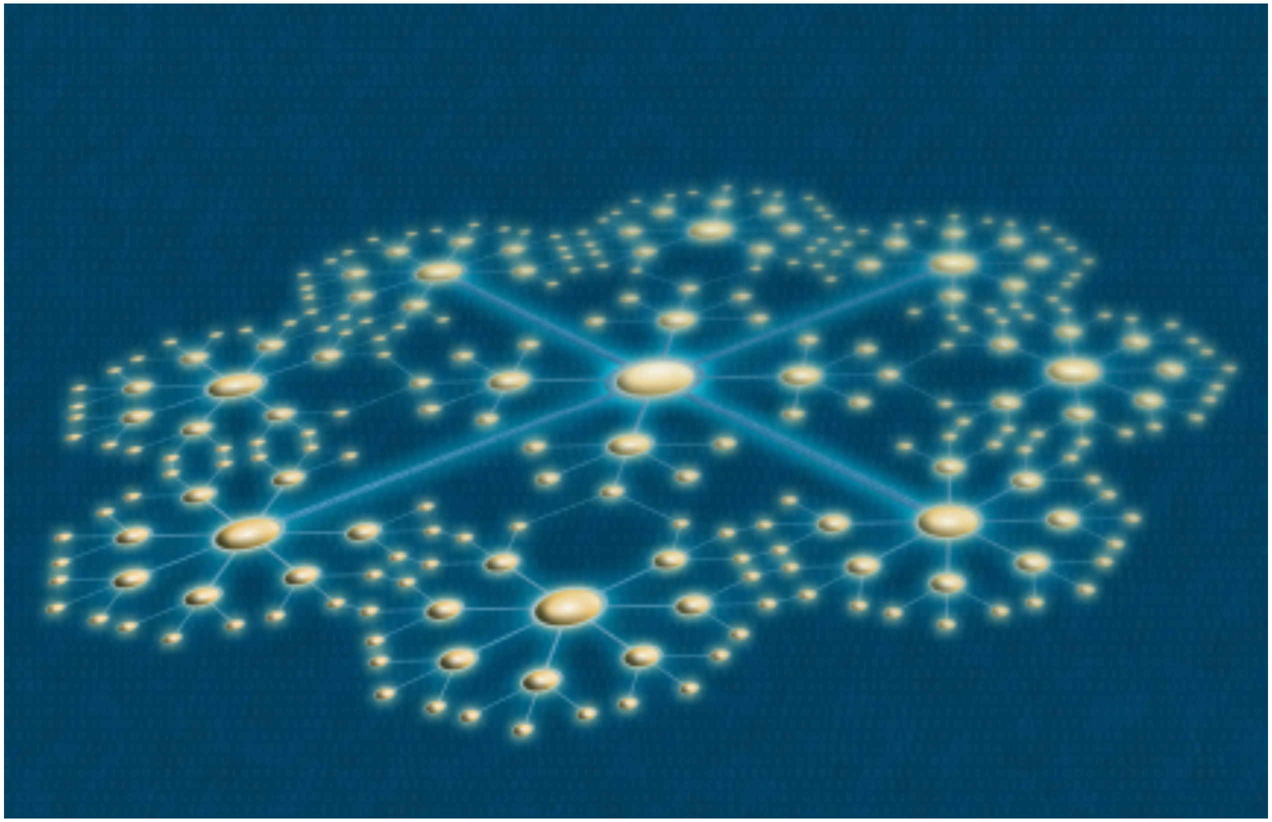
Application attendees		Math discipline attendees	
2	**Astrophysics**	6	Dimensionality Reduction
3	**Biology**	3	Optimization
4	**Earth Systems**	4	Uncertainty Quantification
2	**Nanophysics**	5	Machine Learning
4	**Networks**	4	Network/Graph Analysis
2	Combustion	6	Statistics
5	Cybersecurity	3	Streaming Data
1	Fusion physics		
1	Accelerator physics		
3	Visualization		

Findings: Scalability

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.

Findings: Distributed

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.



Findings: Architectures

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

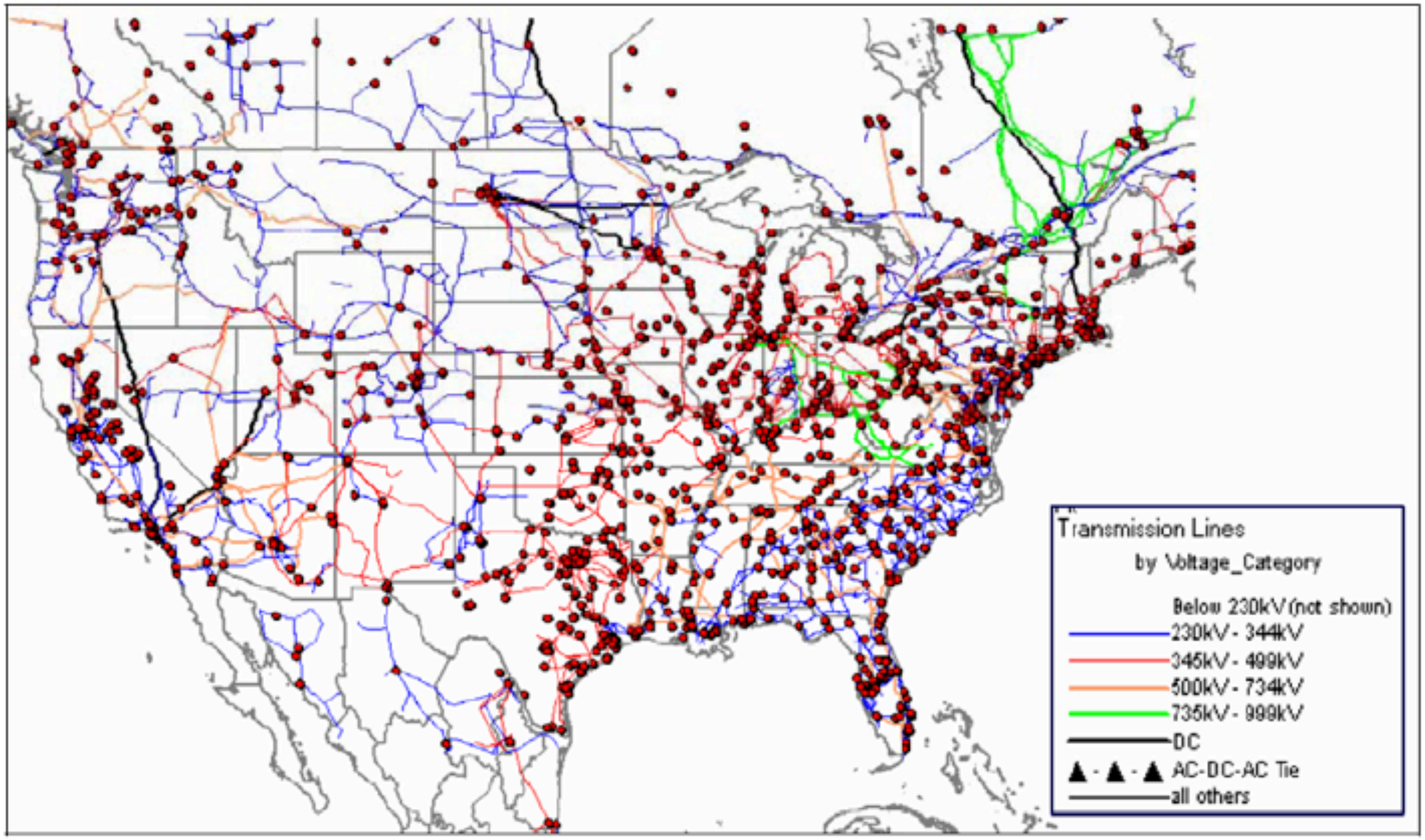
Platform	**Local Memory Access Latency**	**Remote Memory Access Latency**	**Programming Model**	**Type of Remote Access**
Red Storm	Commodity	Medium	MPI	Distributed Memory
XMT	Long	Long	Heavily Multithreaded	Shared Memory
Netezza	Commodity	Short (SPU to Disk), Commodity (Netezza to Netezza)	Augmented 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

Findings: Reduction

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.

Findings: Models

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

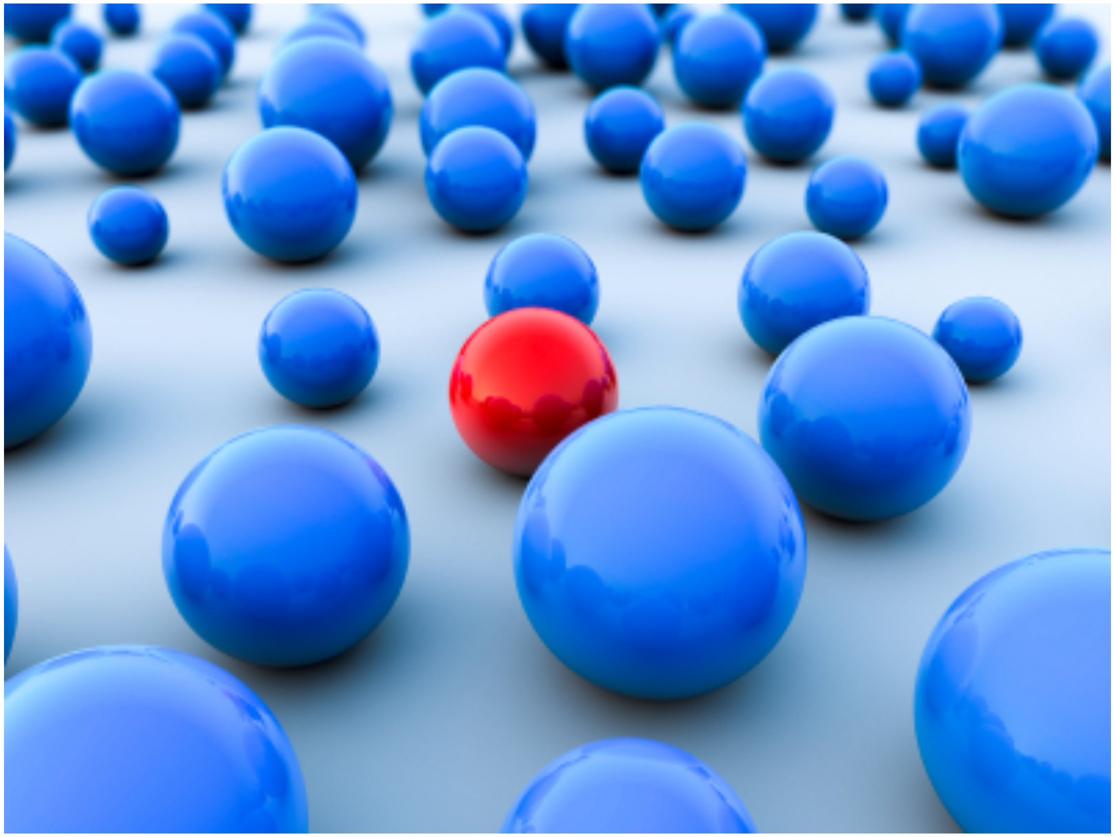


Findings: Uncertainty

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

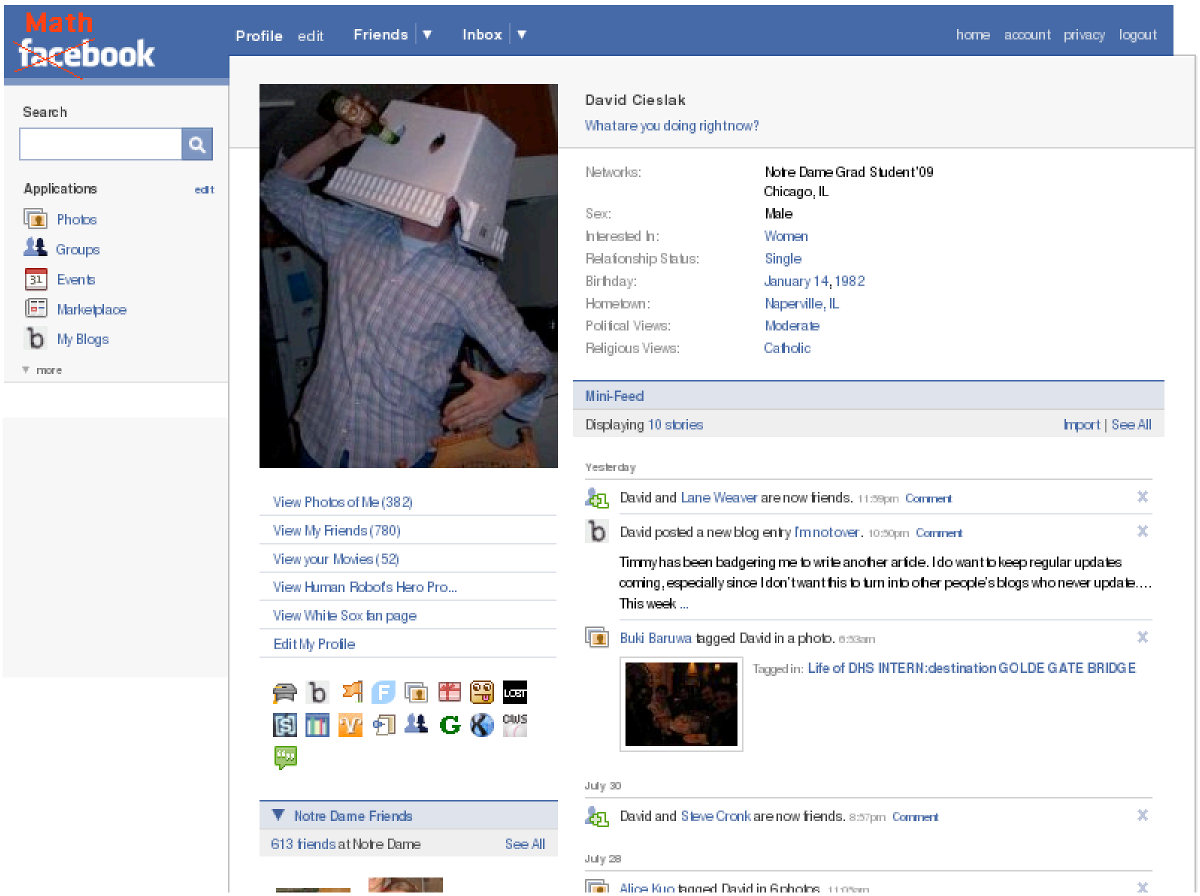
Findings: Outliers

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).



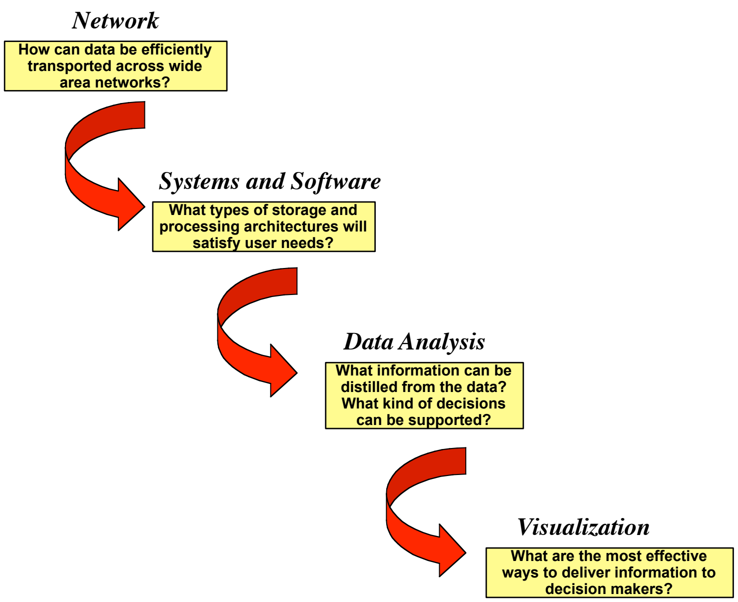
Findings: Culture

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.



Findings: Mathematics is Critical

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



The Data Waterfall illustrates the path from data to decision.

Resources

- Conference materials: `http://www.orau.gov/mathforpetascale`
- Breakout notes and summaries (until May 1, 2009): `http://drop.io/mapd`
- Organizing committee:
  - Philip Kegelmeyer: wpk@sandia.gov
  - Robert Calderbank: calderbk@math.princeton.edu
  - Terence Critchlow: terence.critchlow@pnl.gov
  - Leland Jameson: ljameson@nsf.gov
  - Chandrika Kamath: kamath2@llnl.gov
  - Juan Meza: meza@hpcrd.lbl.gov
  - Nagiza Samatova: samatovan@ornl.gov
  - Alyson Wilson: agw@lanl.gov

Summary

- **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.