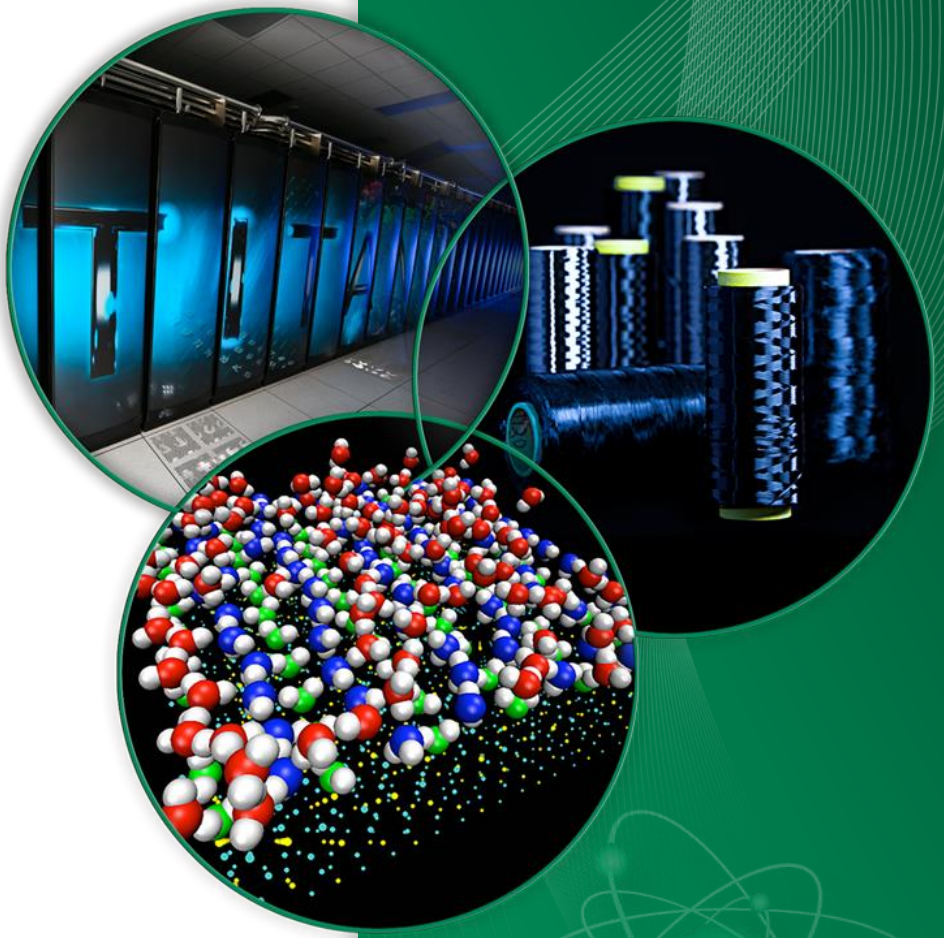
Mathematical and Statistical Methodologies for DOE Data-Centric Science at Scale



Barney Maccabe
Director, Computer Science and Mathematics

May 14, 2013

http://science.doe.gov/grants/pdf/LAB_13-918.pdf

Key Points

- Exploration and understanding of massive science data is emerging as a bottleneck in many DOE science and engineering domains,
  - lack of scalability of the underlying algorithms
  - complexity of the data that needs to be analyzed
- Basic research
  - Novel mathematical and statistical methods, model, and tools
  - For the representation, analysis and understanding of DOE data-centric science at scale
- Data from both
  - DOE advanced experimental facilities and observatories, and
  - data generated from high-resolution simulations on massively parallel computers for complex applications.

Challenges

- Orders of magnitude increase in volume, velocity, and complexity due to improvements in
  - Measurement and sensor technologies
  - Computer architecture
- Beyond volume
  - Heterogeneity, timeliness, and complexity
  - Throughout all phases of the research cycle from generation to scientific insights

Funding

- 3 year program
- Total of $3M / year
- $150K / year – $1M / year per project
- 3–8 projects funded

Areas of Interest

- Data and dimension reduction
- Automated analysis
- Integration of observational data, experimental data, simulation and models

Data and Dimension Reduction

- Data and dimension reduction
  - Novel mathematical approaches and improved methods for data and dimension reduction for extracting pertinent subsets, features of interest, or low-dimensional patterns.
- Spatially and temporarily robust algorithms
  - Algorithms that are robust to spatial and temporal variations in the data. Methodologies for analyzing data, distributed over space and time, with changing local properties that may not statistically resemble the global properties of the data.
- Sampling methodologies
  - Improved sampling methodologies that address the complexity and high-dimensionality of the data.

Automated Analysis

As the **volume, velocity, and variety** of data from experimental facilities increase, and the data collected are **uncertain and imprecise**, advances in automated analysis and algorithm research are required to address these challenges.

- **Real-time anomaly identification**
  - in streaming and evolving data
  - for detecting and responding to interesting phenomena
  - that may be short-lived or urgent.
- **Approximate algorithms** for large, high-dimensional data sets.

  The increasing size of data from experimental and observational facilities may require algorithms that return an approximate result in a short time enabling more accurate end results.

Integration

of observational data, experimental data, simulation and models

- Mathematical and statistical approaches for **combining data of different types and quality**.
  - Techniques should address the challenges that different experiments may be linked across time- or length-scales, or provide different physics insights into a phenomenon.
- Mathematical and statistical techniques to **integrate** large-scale, heterogeneous, high-dimensional data from **observation and/or experiment with simulations and models**.
  - Observational data may be used to guide the construction of the model and to adjust and identify the parameters that define a model. Conversely, simulations or models may be used to understand measurements or observations. New methods and algorithms are needed for data-driven model validation and control and the effective design of experiments, both physical and computational.
- Improved techniques for solution to the **inverse problem**.
  - Inverse problems may be particularly challenging when the volume of data is large and the quality of the data is variable and/or poor. Consideration must be given to the ill-posedness, the computational cost of any solution, and the need to exploit domain information to make the problem tractable.
- **Inference, prediction and reasoning under uncertainty**.

Evaluation (page 10)

1. Advances in applied mathematics and/or statistics basic research:
   - Does the proposed research develop a strong mathematical or statistical foundation for data-centric science?
   - Do the proposed methods and algorithms have the potential to provide a new and/or significant capability for enabling DOE data-centric science over the next decade?
2. Impact for DOE science and engineering challenges:
   - What are the significant or compelling DOE science drivers motivating the proposed research?
   - In the long-term, how will the proposed research impact DOE applications?
3. Applicability to real-world data:
   - Projects should discuss, and preferably, demonstrate applicability to real-world data. Simulated data based on high fidelity models are also acceptable.
4. Scalability:
   - Projects should discuss algorithmic scalability, scalability with regard to the amount of data produced by experiment, observation and/or simulation, and parallel scalability on future computing architectures.

ORNL pre-pre-proposal

- Due May 24th – noon; Yes/No decisions May 28th
- Initial version of pre-proposal
  - Cover sheet
  - Technical narrative
  - (no need for conflict pages)
- Make sure to explicitly hit the four criteria.
- Clearly describe the areas you are addressing.
- What data do you have access to?

Deadlines

Date	Deliverable
May 24 (noon)	Pre-pre-proposals due
May 28	ORNL approval
June 3	Pre-proposals
June 7	Notification
June 12	ORNL organization meeting for successful pre-proposals
June 19	PI with successful pre-proposal meet with Finance Officer before this date.