Baker, Shenda, et al., "Data-Enabled Science in the Mathematical and Physical Sciences: Workshop Report"

PRELIMARY DRAFT

Data-Enabled Science in the Mathematical and Physical Sciences

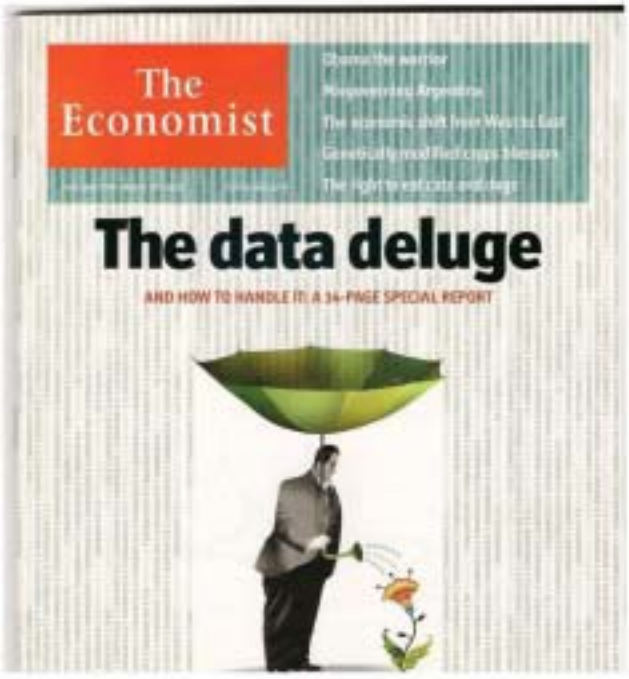
A workshop funded by the National Science Foundation

Held on March 29 – 30, 2010

1

Table of Contents

1. Background and Charge..... 3
2. Executive Summary..... 3
3. Data-Enabled Science and the MPS Divisions..... 7
3.1 Astronomical Sciences..... 7
3.2 Chemistry..... 9
3.3 Materials Research..... 11
3.4 Mathematical Sciences..... 13
3.5 Physics..... 15
List of Participants and Report Authors..... 17
Acknowledgements..... 18
Appendix A – National Study Groups Face the Data Flood..... 18



2

1. Background and Charge

Science has always been data-driven, but what is changing dramatically is the amount of data with which scientists now engage. The Mathematical and Physical Sciences (MPS) community generates much of this data. Major experiments and facilities are now generating petabytes of data per year that must be distributed globally for analysis. Projects already in development will generate much larger volumes at faster rates, approaching an exabyte per week, with exaflop computing capacity needed to perform the analysis.

In addition to this growing number of prodigious data generators, virtually all of science is becoming data-intensive, with increasing size and/or complexity, even at the level of PIs in individual labs. This trend extends beyond MPS disciplines to: biological data; financial, commercial, and retail data; audio and visual data; data assimilation and data fusion; data in the humanities and social sciences; web-based data; and governmental data. Virtually all disciplines need potentially radical new ways to manage this data, as well as major mathematical, statistical, and computational advances to utilize these data sets, if the enormous potential scientific advances are to be realized.

This data-crisis facing science and society has been widely recognized (see, e.g., *The Data Deluge*, in The Economist, Feb. 27, 2010, and the many reports listed in Appendix A). But it is particularly relevant to the MPS community both because of the severe challenge, yet enormous potential reward, inherent in dealing with the data-crisis and because much of the solution will require fundamental advances in the data sciences, of which mathematics and statistics within MPS is a highly prominent part.

Charge: The MPS Workshop on Data-Enabled Science is charged with providing
(1) a high-level assessment of the needs of the MPS communities, including anticipated data generation, capability and inability to mine the data for science, strengths and weaknesses of current efforts, and work on developing new algorithms and mathematical approaches; and
(2) an assessment of the resource requirements for addressing these needs over the next five years.

2. Executive Summary

To realize the extraordinary potential for scientific advance inherent in the data-crisis, two major hurdles need to be overcome: (1) Data Management and (2) Scientific Inference from massive or complex data. We summarize the major issues involved in each hurdle below; details and examples are given in later sections.

Data Management: Handling the enormity of arriving and soon-to-arrive scientific data requires complex and new strategies and understandings. Components of this management include:

- Designing the data collection strategy.
- Collecting the data, from either single or distributed sites.

- Preprocessing (if necessary) to keep only the most essential data.
- Storing the data, with appropriate meta-data to ensure usability.
- Ensuring accessibility of the data by scientists, possibly through layered distribution of the data to multiple sites.
- Providing platforms and software that enable efficient use of the data by scientists, as well as allowing for capture of the scientists' post-processing of the data.
- Ensuring curation and preservation of data.

Scientific Inference from Massive or Complex Data: There are major challenges in producing breakthrough science from massive or complex data. Note that we emphasize complex data in this discussion as well as massive data; what might appear to be of modest size today (e.g. the number of genes in the human genome) can cause as severe inferential difficulties as massive data when consideration is given to complexity (e.g., the need to consider the vast multitude of possible gene networks). A few of the overarching challenges are given here; others are in later sections.

- Scalability is a primary concern; much of science today uses 'small data' methodologies for scientific inference, strategies that are ill-equipped for today's massive or complex data. As but one example of the scalability crisis, while many thousands of astronomers (and data scientists) have used the Sloan Digital Sky Survey (SDSS) data collection over the past decade, with over 2000 refereed publications (making it one of the most scientifically productive data repositories in the world), nevertheless still less than 10% of the SDSS imaging data have been retrieved and analyzed by individual scientists. The Large Synoptic Survey Telescope promises to blow this gap wide open, by three orders of magnitude, with the acquisition of one SDSS equivalent amount of imaging data each and every night for 10 years. Without advanced data science (mathematics/statistics, data mining, and machine learning) algorithms and methodologies tuned to and applied to such a data flood, we cannot hope to reap its full scientific discovery potential.
- There will be a dynamic tension between the desirability of broadly useable approaches to data-enabled science – across applications and disciplines – and the frequent need for solutions tailored to a specific setting.
- Mechanisms for transference of methodologies between disciplines is a major need; MPS is well-positioned for this, because mathematics and statistics have traditionally been the major disciplines for effecting such transfer.
- Data-enabled science is not just data exploration and understanding; it is often using the science to provide the insight that unlocks the data. (One cannot find a needle in a haystack without knowing what a haystack is or a needle is.)
- Understanding how to deal with the multiplicity issue – distinguishing a scientific signal from noise, when a large data set is subjected to a massive number of probes – poses a major challenge.
- Fundamental advances in the methodology of data-enabled science often require awareness of the entire spectrum of the problem; from the nature of the data to computational issues (e.g. parallelization) in the final analysis.
- There is frequently a need for real time analysis of the incoming data-stream.

Overall Recommendation on Data-Enabled Science: We urge the MPS Directorate to obtain very significant additional funding to support data-enabled science. This funding could be used for new data-enabled science initiatives or to provide targeted additional support to the MPS Divisions for data-enabled science activities, support that could be applied to individual investigator awards, group grants, centers, and facilities, as the individual Division deems most appropriate.

- Funding of data-enabled science will require the same process care by NSF program officers as funding of interdisciplinary research.
  - Peer reviewers in all MPS review panels should be clearly informed as to the unique evaluation metrics that apply to cross-disciplinary DES research proposals, which bridge both data sciences (including scientific data management, scientific database research, mathematics/statistics, data mining/machine learning, and visualization) and the traditional physical sciences.
  - Dedicated data-enabled science review panels should be utilized when appropriate, certainly at the Divisional level and possibly at the Directorate level.
  - If support is through additional funding to the Divisions, MPS tracking mechanisms should be developed to insure accountability for these targeted funds.
- Funding should be made available for needed Workforce enhancements:
  - Support dedicated Early CAREER awards for young faculty specifically in DES research areas.
  - Support dedicated fellowship programs (graduate and postdoctoral) in DES and Data Science research areas. This would be similar to the NSF Fellowships for Transformative Computational Science using Cyberinfrastructure (CI TraCS: http://www.nsf.gov/pubs/2010/nsf10553/ nsf10553.htm)
  - Support workforce development in careers associated with data handling and understanding.
  - Provide stronger DES research support for scientists working within large data-producing projects during construction, commissioning, and early operations phases. This enables early science results from these facilities specifically from the people who know the facility and its data the best.
  - Provide REU supplements in data-enabled science.
  - Support educational initiatives in data-enabled science, including the training of computational scientists for scientific inference with massive and complex data. (See section 3.1 for numerous concrete suggestions.)

Recommendations on Data Management:

- For facilities, data management is a major (but often unfunded) component of operating costs. As part of the overall NSF strategy of funding facility operating costs, dedicated data management operating funds should be provided and tracked. This should include funding for data management personnel and software development.
- Project proposals which deal with massive data should include a data management plan consistent with the size, collaborative structure and funding scale of the project.
  - The plan should address (as relevant), meta-data, access, long term funding, data storage, computational requirements, and standards.
  - Data Management with massive data requires significant innovation, and new management ideas should be encouraged and supported (recognizing they might

5

fail). Conferences or other vehicles for sharing of data management innovations across facilities and disciplines should be created.

- NSF should continue to seek mechanisms to ensure that data arising from funded NSF projects be made public (in a useable form) within a reasonable time period.
  - Otherwise, reproducibility of science will be at question.
  - Without this mandate, science will lose much of this major resource.

Recommendations on Scientific Inference: The scope of needed fundamental advances in using massive or complex data for scientific inference is enormous. Some of the most urgent needs are listed here. Others can be found in the discipline-specific sections.

- Advances in fundamental mathematics and statistics are needed to provide the language, structure, and tools for many of the needed methodologies for data-enabled scientific inference. (See section 3.4.)
- Algorithmic advances in handling massive and complex data are crucial, including methods of exploiting sparsity (e.g., out of a huge list of proteins, only an unknown few may be active in a particular metabolic process), clustering and classification, data mining and machine learning (including feature detection and information extraction), Bayesian analysis and Markov chain Monte Carlo methodology, anomaly detection, optimization, and many more.
- Potentially major tools for the characterization and interpretation of massive and complex data sets include visualization (visual analytics) and citizen science (human computation or data processing).
- Data assimilation and uncertainty quantification – names given to the interface of data and computer modeling of processes (simulation-enabled science) – requires special focus as the basis of much real-world prediction (e.g., of the effects of climate change).
- Progress in new areas of data-enabled science will require teams consisting of combinations of disciplinary scientists, data-scientists (including mathematicians, statisticians, and machine learners), and computational scientists. Mechanisms for support of such teams are needed; the current mechanism of occasional joint initiatives between divisions is too transient for the future data-enabled science world.

Of course, many of these issues arise throughout science, engineering and society. They are also of NSF-wide importance and of importance to numerous other agencies and the nation. We here primarily highlight MPS issues in data-enabled science, while recognizing that solutions to the overall problem may well require a coordinated national (and international) effort.

We also note that MPS developments in data-enabled science will likely be major drivers of solutions to data-enabled science problems in general. The data management methodologies arising from major MPS facilities and the fundamental breakthroughs for scientific inference from massive or complex data that arise through mathematics, statistics, and other MPS disciplines will have major impact in other sciences and society.

6

3. Data-Enabled Science and the MPS Divisions

3.1 Astronomical Sciences

While there are a plethora of astronomical research projects for which the access to and understanding of large-scale data is critical, exploration of the time-domain is perhaps the most revolutionary. Facilities now in operation and others planned for the coming decade will observe the night sky systematically, with a cadence never before achieved. At this level of sampling virtually all stars in our Galaxy become non-stationary, and many will be discovered to be variable in ways not previously known. Other variable, episodic, and transient events—supernovae, novae, accreting black holes, gamma-ray bursts, gravitational microlensing events, extrasolar planetary transits, incoming asteroids, trans-Neptunian objects—will be recorded at rates 100-1000 times higher than in the past.

In order to make sense of the 10^3 to 10^5 detections of transients per night, and to aid other observers in assessing the need for and priority of follow-up observations, analysis and probabilistic classification of events will have to be highly automated. A combination of advanced machine learning technologies with immediate access to extant, distributed, multi-wavelength data will be needed to make these assessments and to construct event notices to be autonomously distributed to robotic observatories for near-real-time follow-up.

The scientific implications of these capabilities span all areas of astrophysics: planet formation and the prevalence of extrasolar planetary systems, stellar evolution and the structure and history of our Galaxy, galaxy formation and evolution, active galaxy phenomena (quasars, blazars, Seyfert galaxies), the distribution of dark matter in galaxies and clusters of galaxies, and the very nature of the cosmos on the largest scales. The most important and exciting astronomical discoveries of the coming decade will rely on research and development in data science disciplines (including data management, access, integration, mining, and analysis algorithms) that enable rapid information extraction, knowledge discovery, and scientific decision support for real-time astronomical research facility operations.

Specific Astronomy Data-Enabled Science Recommendations:

1. Data management
    a. Support core facilities at adequate level so that data processing and data management are not eroded by other operational requirements.
    b. Incorporate data management planning from the outset
    c. Protect data management budgets from hardware cost overruns
    d. Manage data close to source of expertise, recognizing that data management is inherently distributed and that data centers will vary in scale
    e. Adopt community-wide standards for metadata to facilitate discovery, access, integration, and re-use
        i. International VO standards for data collections
        ii. Standard access protocols
        iii. Management of virtual data spaces
        iv. Authentication/authorization as needed

7

f. Close the gaps in astronomical data archiving
   i. Engage private observatories to establish coherent, community-accessible archive facilities, especially in cases where private facilities accept NSF support for instrumentation and/or facility augmentation
   ii. Capture high-level data products associated with peer-reviewed publications and manage as community data resource with VO-compliant access
   iii. Develop strategies for long-term curation and preservation of survey data (e.g., SDSS), perhaps in collaboration with NSF DataNet programs
   iv. Support creation of advanced data products from archival collections (source catalogs, cross-matched source identifications, parameter extraction for specific types of astronomical objects)
   v. Establish programs for digitization of legacy data collections; the photographic record (images, spectra) is on the verge of being lost

2. Analysis and visualization
   a. Invest in new software and databases aimed at exploitation of large and distributed data collections
   b. Modernize widely used tools, with built-in access to distributed data through VO service standards
   c. Support algorithm development related to large/distributed data and scale-up existing algorithms
      i. Clustering and classification methods
      ii. Bayesian statistical analysis and Monte Carlo Markov chain approaches
      iii. Visualization of large, many-dimensional data sets
   d. Support interdisciplinary resources – molecular spectral line databases (astronomy, molecular chemistry), atomic spectral line databases (astronomy, atomic physics)
   e. Support collaborative research with industry that utilizes emerging technologies for data-intensive science (e.g., the recent NSF-Microsoft MOU for data-intensive cloud computing: http://www.nsf.gov/pubs/2010/ nsf10027/nsf10027.jsp).
   f. Support collaborations among astronomers, statisticians, mathematicians, and computer scientists. The NIH program in informatics is a successful model of the kind of research objectives that would be useful in astronomy: http://grants.nih.gov/grants/guide/pa-files/PA-06-094.html.

3. Archival research
   a. Support PI-based archival research programs through program solicitations focused on use of archival data
      i. Archive-enabled research stands on equal footing to new observations
      ii. Archive research draws on both ground-based and space-based observations
      iii. NSF/NASA co-sponsorship?

4. Community workshops, communication, professional outreach
   a. Support annual community workshops that focus on DES, Data Science, Informatics, and Large Science Database Projects (e.g., LHC, LSST, LIGO, OOI, NEON), in order to develop the field, share lessons learned, offer workforce development

8

opportunities, and provide a venue for educating the scientific community in DES research.

5. Education and public outreach
   a. Work with EHR to support STEM education research programs that focus on the development of curricula and educational programs at the intersection of physical sciences and data sciences. Support for programs that (a) demonstrate the pedagogical value of introducing the reuse and analysis of scientific data in inquiry-based STEM learning, (b) promote computational and data literacy across the STEM curriculum, and (c) encourage education research in the science of learning from large data sets (http://serc.carleton.edu/usingdata/).
   b. Mandate an outreach component in all major projects and facilities – reward innovative public uses of mission/project data (e.g., Citizen Science). Support construction of infrastructure that facilitates the development, sharing, and transparent reuse of data products that have pedagogical value and that serve a broad public audience, not just professional researchers.
   c. Fund the development of digital libraries that provide a permanent repository of data science curricula materials (and data sets vetted for education use) for different core science as a mechanism for easy transfer of DES knowledge, data-centric lesson plans, and MPS-related science results to both informal and formal education venues.
   d. Fund informal science education and human computation initiatives that extend the discovery potential of large science data sets (e.g., through Science@Home or Citizen Science activities).
   e. Fund the development of data science software tools (for data access, manipulation, measurement, mining, analysis, and visualization) for use in informal and formal education.

3.2 Chemistry

Data Enabled Science (DES) uses techniques in statistics and high performance computing to analyze complex data sets and extract features of scientific interest. These complex data sets can be very large data sets from single experiments or large collections of data from several sources. In these cases, visualization techniques and data mining procedures have the potential to dramatically increase the rate of scientific discovery.

Although chemistry and materials science typically generate small scale data sets compared to fields such as astronomy and high energy physics, many experiments are beginning to generate single-run data sets that cannot be easily analyzed by conventional techniques. These experiments are usually multidimensional and involve coupling a high throughput chemical analysis technique, like mass spectrometry or broadband spectroscopy, to an excitation source such as a laser. These multi-dimensional techniques are often required to analyze complex sample mixtures or to examine reactivity as a function of deposited energy. Current techniques in the combustion and reaction dynamics fields, such as Multiplexed Photo-ionization Mass Spectrometry, are generating single-experiment data sets on the order of 50 GB that would benefit significantly from statistically robust visualization methods.

Several other areas of chemical and materials research are also producing large data sets that will continue to increase in size and complexity. In particular, molecular dynamics simulations in biochemistry and materials science generate large scale computational data from single laboratory studies. The use of graphics processing units in computational chemistry, for example, has led to simulations that produce terabytes of computational output per day. There are also large data sets in related fields of science, such as radio astronomy, that contain molecular information that require new analysis tools to extract the chemically useful information.

Finally, several of the industries that employ chemistry and material science Ph.D.'s are rapidly pursuing DES strategies to decrease product development cycles. Providing research experiences for graduate students will become increasingly important for preparing young scientists for the future workforce. Therefore, despite the "single laboratory" tradition of chemistry and material science research, issues in DES are already significant in chemistry and will continue to gain importance.

Special Needs for Chemistry and Material Science

As noted above, chemistry and materials science tend to perform research in a single-laboratory model. Increasingly, each individual laboratory is generating large scale data sets through either computational chemistry, large user facilities (such as SLAC, NIST or ORNL) or high throughput laboratory methods. However, the potentially greater opportunity for DES in these fields is the combination of research data from all groups in a research discipline. For example, a unified spectroscopic database from emerging high throughput spectroscopy methods based on frequency comb spectroscopy and direct digital spectroscopy could have a major impact on related fields of astronomy, environmental science, and analytical chemistry that rely on chemical identification by spectroscopy. Efforts are already underway in the computational chemistry community to create common data bases to permit reuse of these results (examples include iOpenShell (Krylov), the Structural Database (Johnson)). Unified collections of individual data sets in materials science and drug discovery could significantly increase the rate of discovery and add increased value to the individual laboratory data collections. The concept of unified data sets from whole communities of chemistry represents a major shift in the single-laboratory culture where data is often closely guarded.

A special area of DES with great potential in chemistry and material science fields is the combination of laboratory or facility measurements and computational chemistry to provide real-time chemical analysis. Many experiments in chemistry rely on theoretical analysis or computational simulation to interpret the experimental data. In almost all cases in chemistry, these tasks are performed separately and often by different research groups. For example, an experimental group will collect the data set and send it to a collaborator in computational chemistry for analysis. The possibility of closing this loop in real-time would make it possible to optimize experimental conditions in a single experimental run and, therefore, greatly decrease the time required to perform the crucial experiment to reveal the important chemistry. On-the-fly analysis methods are also needed to realize the full potential of new techniques like broadband spectroscopy using frequency combs or digital electronics. Spectrometers based on direct digital spectroscopy will soon be capable of measurement throughput of about 1 TB/hour

(spectrum acquisitions rates of 300 spectra/s with 1 million data points per spectrum). Coupling high performance computing to concurrent measurements could be used to perform on-line spectral analysis in high throughput analytical systems to enable library-free chemical detection and create systems that provide "sample in – structure out" real-time analysis.

Another area of need for chemists is a lack of standard and compelling visualization tools. High performance computing tools and software that provide visualization of chemical models, processes and structures should be developed. NSF should provide funds to support both the people who develop the computational interfaces and software as well as the hardware to handle the data manipulation. Much of the massive data generated with local and facility instrumentation is collected in phase space and frequency, and needs to be converted into real space and real time. With appropriate software and algorithms, visualization of the real structure and dynamic modes and patterns emerging from the data can be observed and interpreted. In addition, science is better communicated to the public and as an educational tool through visual representation of interpreted data.

Specific Chemistry Data-Enabled Science Recommendations:

The NSF should develop funding opportunities that provide incentives for research communities in chemistry and materials science to reach agreements on data sharing protocols, including data formats and associated meta data. These programs will need to include continued support for curating and validating the data collections so that users within the research community and outside the direct community trust the content, security, and future accessibility of the collection. Additional support to develop discipline-specific software tools, perhaps through collaborative research opportunities in math, statistics, and computer-related disciplines, to navigate and mine the data sets will also be required. Instrument development that emphasizes real-time data analysis and visualization through the integration of high-performance computing with state-of-the-art instrumentation should be encouraged. The NSF should also support interdisciplinary educational opportunities that train chemistry and material science students in data-related fields to better prepare them for future opportunities in industry and government positions.

3.3 Materials Research

The frontiers of computational materials science research, supported within the Condensed-Matter and Materials Theory and Materials Chemistry areas, aer driven by *Data-Enabled Science* (DES). DES within materials community constitutes a necessary "fourth paradigm" within the now-standard theory, experiment, and computational simulations paradigm defining our modern research and discovery. While the community has extensive efforts in a number of challenges, supported by various NSF programs, in high-performance computing, algorithmic developments, computer-architecture utilization, e.g., GPU and vector accelerators, DES is at the heart of critical-need materials development and of challenges in understanding of complex materials systems. Large-data sets and data-mining critical information from that data (e.g., intrinsic correlations between structure and property) are increasingly important in materials science and engineering, and increasingly necessary for breakthroughs. Managing, storing, sharing, utilization and visualizing these data from diverse materials areas require new approaches and

new developments in cyberinfrastructure, and, especially, a huge cultural change within the community and from other critical communities that will have great impact on DES success in the materials community, such as critical computer science experts in the database research and architecture arenas. In addition, although materials data often is more heterogeneous than other areas, the materials community can benefit in DES from advances made in other communities, such as biomedical database (see, e.g., http://wtec.org/sbes) and the Sloan Digital Sky Survey (http://www.sdss.org), as well as from tools developed to describe, manage, archive, and disseminate data, such as MatDL Pathway (http://www.matdl.org), an effort that, nonetheless, did not solve *workflow issues*, and the materials community's data remains an afterthought. Other critical areas are data provenance and data security, while providing an open resource for NSF-supported science efforts.

Currently, standard workflow is a bottleneck to progress; namely, there is limited sharing of data and data products. Data is provided on "need-to-know" basis, peer-to-peer sharing difficult (learning curve between groups), no meaningful relationships between files and data products (need for meta-data and workflow), data lost over time (storage and management) or unable to be found or searched except by person who generated them (unusable but existing data).

There has been a vision developing over recent years, referred to as Integrated Computational Materials Engineering (ICME) in recent NAE reports, where computationally-driven materials developments is a core activity of material scientists and engineering in coming decades, along with standard experimentally-driven materials engineering. As such, both data from computation and simulation research and experiment are critical. Certainly, there may no "one-stop" solution for the entire community. However, even having research groups with similar applications and data needed could provide a "local community" effort with much more robust data access and management with useful tools to enhance DES for their entire community (shared resource and development). Overall, most of the materials community desired an easy, searchable access to full research product anytime and from anywhere, so as to provide collaborations with seamless and protected sharing of data and metadata. Data repositories require new advances in cybersecurity and large-scale networking for geographically disperse collaborations.

Thus, from the NAE report, the ICME cyberinfrastructure will be the enabling framework for DES and Discovery, including libraries of materials models, experimental data, software tools, datamining tools. To accomplish this task, the creation of accepted taxonomy, informatics technology, as well as materials databases with open access is essential. "Knowledge Bases" are the key to capture, curate and archive information to succeed with the vision for ICME.

To accomplish these needs, the cultural must be changes, as there is no culture for massive datasharing, and no incentives from funding agencies for sharing. Multi-agencies issues, as opposed to NIH model, means that funding and coordination are modest for needed cyberinfrastructure (database, security, curation, etc.). In addition, the culture to support cross-disciplinary developments for DES in materials science is critical. For example, recent funding calls within NSF certainly permitted database develop efforts. However, reviewers from the database research and architecture within computer-science often found the database research "not groundbreaking", while acknowledging that the impact on the DES materials side would be significant, effectively killing any funding possibility. Changing the mindset and the cultural to

permit cross-disciplinary support for DES in materials science based on coordinated developments with critical computer science research, which are often extraordinarily useful for real DES but not "not groundbreaking database research", is a critical need for success.

3.4 Mathematical Sciences

The era of data-enabled science (DES) opens up exciting research frontiers for the Division of Mathematical Sciences, even as it poses enormous challenges. The challenges can be classified into at least three broad categories: (1) extending existing theory and algorithmic techniques to new scales and new applications, where current methods become bottlenecks, (2) developing new theoretical approaches and algorithms and demonstrating them on benchmark problems, (3) collaborating on real-world applications with domain experts in science, engineering, and policy making, where the availability of new types and quantities of data offers the hope of scientific breakthroughs.

There are many fresh technical results in basic disciplines such as linear algebra (e.g., tensor orthogonal decompositions), approximation theory and harmonic analysis (e.g., sparsity and customized basis sets), and statistics (e.g., the revival in Bayesian analysis) relative to the research agenda discussed herein, but technical details are not featured at the high level of this discussion. Some key concepts are low-dimensional representation of formally high-dimensional data sets, low complexity algorithms that are much less expensive in storage requirements and running time than traditional algorithms (even sublinear in data set size) while maintaining sufficient accuracy, and once-through streaming of the input data set.

Data-enabled science has been called "fourth paradigm" in apposition to the historically dominant paradigms for scientific discovery, engineering design, and decision support of theory and experiment, and the recently rapidly developed "third paradigm" of simulation. Theory and simulation are based on physical models that can be mathematized. Experimentation is model-driven. In contrast, some data-intensive approaches effectively predict outputs of a system *without* the requirement of models representing the dynamics of the system, which makes these approaches very interesting for frontier science. Of course, there are deep mathematical models underlying discovery techniques for data sets that make this predictive power possible, even if the system dynamics are unknown. Such approaches depend upon large volumes of data (system history) and are increasingly interesting as humans collect data from sensors, satellites, sophisticated experiments, and records of their own activity. The statistical and mathematical tools underlying machine learning and dimension reduction techniques of all kinds must be percolated into lower levels of the curriculum, to train data proficient scientists in anticipation of a profound shift of research resources into data-enabled science in the future.

The value of data on a "per byte" basis often increases with the availability of more data for context. Overlays of different types of data (e.g. correlation of multiple measurements in experiments, of multiple diagnostics in medicine, or of multiple indices in geographical information systems) offer insights that are not available from the same data considered separately. Discrete mathematics can play a key role here, in terms of information retrieval and associative databases.

13

Data-enabled science is interesting on its own, but even more interesting in combination with simulation-enabled science. The latter is limited by modeling errors (among other limitations) while data-based methods are limited by observational or experimental error (among other limitations), which can be profound in leading edge scientific experiments in which the signals of interest are weak or rare in the midst of noise. Together, through methods like data assimilation and parameter inversion, these two ugly parents can have a beautiful child, the limitations of each being reduced by being taken together. Moreover, real-time data-enabled scientific discovery can be aided by the simulation informing the experimental or observational process about where to concentrate effort (optimal sensor placement). This synergism is rarely exploited today because of the two worlds, are disconnected in terms of practitioners, software-hardware interfaces, and the compute-intensiveness of doing the assimilation and steering.

A major challenge for mathematical scientists is to winnow massive data sets and represent them sparsely, for computing and storage purposes. Sometimes, loss in compression cannot be tolerated for scientific or legal reasons, but raw large-scale data sets can often be reduced by orders of magnitude in bulk without negative implications and there is a premium on performing this reduction and working in the "right basis" for many reasons, as we become deluged by data. The acquisition cost of large-scale computers is in the data memory and the operation cost of large-scale machines is in moving the data around, not manipulating it arithmetically. Moreover, I/O rates lag processing rates, putting an operational premium on minimizing I/O beyond the budgetary premiums.

The Division of Mathematical Sciences has natural partners beyond the scientific divisions of the MPS Directorate, in other parts of the Foundation and beyond. Other research-intensive agencies (e.g., DoD, DOE, NASA, NIH, NIST) and mission agencies (e.g., AHRQ, BEA, BJS, BLS, BTS, Census, EIA, EPA, IRS, NASS, NCES, NCHS, OMB) are awash in data that need to be gathered, curated, archived, turned into useful information, and applied. Needed from DMS are abstractions, algorithms, and software tools to: characterize and improve data quality, to trade off cost and data quality, to link multiple databases, and to analyze. In some instances, privacy and confidentiality are major concerns. DMS researchers can contribute to tools to handle legacy data and new forms of data (audio, images, video). DMS researchers must also be involved in developing means of quantifying uncertainty, and means of communicating uncertainty to the public and to policy makers. The mathematics of risk analysis must be developed to accompany the emergence of data-enabled science

While considerable opportunities present themselves for mathematicians and statisticians to embed themselves in applications, long-term curiosity-driven research in data science must also be encouraged. History shows that the fruits of curiosity-driven research in the mathematical sciences are plucked by applications, at unpredictable intervals following their invention. Outside of the scientific realm, information management has grown to be a $100 Billion business, so spinoffs from data-enabled discovery can lead to huge multipliers in competitiveness.

In summary, mathematics and statistics lie at the intersection of all quantitative fields engaged in DES, through the power of their abstractions, and they swiftly convey breakthroughs in one field

into related ones. Individuals in DMS are often involved with the frontiers of DES both internally to the discipline and in interdisciplinary contexts. Growing numbers in the mathematical sciences community wish to be involved in DES problems, which has led to some of the most innovative, prize-winning developments in mathematics and statistics in recent years and some of the greatest fun. Impediments to be addressed by MPSAC could include the difficulty of securing postdoctoral funding (especially in statistics) and limited opportunities for interdisciplinary engagements as co-PIs on project proposals to NSF, since projects that are truly collaborative may present particular challenges to review panels.

3.5 Physics

Large data sets are a familiar component of physics research. In recent years, LIGO has acquired about two petabytes of data. With the Large Hadron Collider (LHC) reaching interesting beam energies, particle physics is preparing for the impending data tsunami which will generate about 700 MB of data per second. And this does not include simulated data, which could easily double or triple the data rate.

These big experiments are not the only data-enabled physics, however. The ability to simulate complex physical systems is also advancing rapidly. The output of these simulations will grow in size and complexity as more physics is included in the simulations. Moreover, single investigator experimental programs can easily acquire large amounts of data and many would benefit from better algorithm, software, and even data sharing formats.

The scientific pay-off of these data-intensive projects is bounded by the ability to process and analyze the data at the rate they are acquired.

Case Study I: Gravitational-wave Astronomy (LIGO)

The scientific pay-off of LIGO is bounded by the ability to process and analyze the data at the rate they are acquired. Over the past decade, LIGO has acquired 2 petabytes of data. The scientific collaboration adopted an hierarchical grid model for data storage and computation in which raw data is archived in Tier-0 data centers and centrally aggregated to a Tier 1 from which reduced data is moved to Tier-2 (regional compute centers) and Tier-3 (university compute centers). A similar structure has been adopted by the LHC experiments.

Over the next five years, the Advanced LIGO instrumentats (aLIGO) will be installed and begin operating. LIGO has partnered with Virgo, a French-Italian gravitational-wave detector project, and with GEO, a British-German detector project, to form a global network of gravitational-wave detectors. The goals of aLIGO are to test relativistic gravity and to develop gravitational-wave detection as an astronomical probe. aLIGO operations will span the transition from rare detections to routine astronomical observations. In stable operations, aLIGO will generate about 1 PB of raw data per year which needs to be replicated between the geographically distributed observatories and the compute centers at the same rate as it is acquired. A number of processed data products are planned including reduced data sets for scientific analysis, event databases, and astronomical alerts when transient events are identified. Robust online and offline data handling

and analysis capabilities are required. Pipelines generating transient alerts & data quality information within seconds of data acquisition are also needed. Careful attention must be paid to interfaces between control/diagnostic systems, data acquisition systems, and processing systems to ensure robust operations of the low-latency system. The data will be re-processed offline for transients including deeper searches, enhanced data quality generation, searches for continuous and stochastic signals, parameter estimation, and simulations.

To achieve the science goals, four aspects of data processing and analysis must be supported: 1) storage and compute resources including both hardware and personnel, 2) development, enhancement and support of middleware and services including data discovery and replication, database of events and data quality, authentication/authorization, monitoring, 3) development, enhancement and support of software to provide access and core algorithms, 4) development and prototyping algorithms and pipelines to identify signals to identify correlations with the environment and auxiliary systems. This requires support of discipline specific scientists, mathematicians, statisticians, and computational scientists.

Case Study II: Large Hadron Collider

On March 30, 2010, with the first 7 TeV proton-proton collisions at the LHC, high energy physics entered an era in which data sets are expected to grow to more than 10 PB/year within a few years. Particle physicists are now, in effect, running two sets of experiments simultaneously: one to search for new physics that could change our view of nature and the other to test whether or not the newly created cyber infrastructure, the Worldwide LHC Computing Grid (WLCG), works effectively under highly stressed real-world situations. The goal of the WLCG is to provide physicists controled, and timely, access to approximately 100,000 processors, housed in 170 computer centers in 40 countries.

In a typical analysis in high energy physics, physicists compare observations with background models that have been validated using real data. In addition, the same data may be compared with various models of potential new physics. These signal models typically depend on several parameters. For example, the simplest supersymmetric (SUSY) models require the specification of 5 to 6 parameters, θ, in order to define the models completely and thereby allow for prediction of the expected signal $s = f(\theta)$. In dealing with such models, physicists are faced with at least two problems: 1) the function $f(\theta)$ is typically not known explcitly, but only implicitly through semi-analytical calculations that involve simulation, and 2) to test such models effectively, analyses need to be optimized at multiple parameter points θ. This requires the simulation, at each parameter point, of hundreds of thousands to millions of proton-proton collision events. In the simplest cases, each of these optimized analyses would be applied to the real data yielding N events that satisfy certain cuts. Even for a simple count-based analysis, such as we are describing, which reduces the raw data to a set of (correlated) counts $\{N\}$ and the associated set of background estimates $\{B\}$, the computational burden of performing scientific inference for a multi-dimensonal parameter θ is very large, especially if Bayesian methods are used. Moreover, the entire procedure, in principle, must be repeated for every class of models to be tested. At present the software codes to execute such analyses are developed by teams of physicists in ways that may be not be optimal in terms of resources needed and the timeliness of results. New algorithms will be needed to scale up, or more likely replace, existing practice.

Specific Needs of the Physics Community

A. Need for data storage and computational facilities: The experimental gravitational-wave and particle physics communities have developed an hierarchical grid model for data storage and computation in which raw data is archived in Tier-0 data centers and reduced data is moved to Tier-2 (regional compute centers) and Tier-3 (university compute centers). This hierarchical distribution of data and computing resources is an extremely effective way of insuring the data can be easily accessed and used by the physicists. It is clear that a similar hierarchical approach is needed to support the simulation community which requires a range of computational facilities that allow rapid prototyping and debugging in addition to the larger compute centers which provide the resources for high-resolution and large scale simulations. Ideally, there would be a seemless migration from rapid prototyping to the execution of a large-scale analysis. This is not the case at present.

B. Need for support personnel: The processing and analysis of large data sets requires software and services to allow scientists to extract the maximum scientific pay-off. Among the activities that need to be supported are authentication and authorization services, help desk support, software build and test facilities, monitoring of storage and computational resources, data replication and movement, data and metadata capture services, data mining tools and visualization. To deliver high quality, enabling products requires a combination of discipline specific scientists, software engineers, and programmers. For large experiments, a reasonable rule of thumb is that support for these activities requires about 10% of the operating costs of the effort. It is important to note that the full release of data which have been processed to remove artifacts goes beyond this scope and may require an additional 10% of the operating costs to support.

C. Need for support of interdisciplinary research activities: Algorithm and application development needs vary according to the specific activity being undertaken. With the explosion of data from experiments and simulations, there is an urgent need for collaborations between physicists, mathematicians, statisticians and computer scientists.

LIST OF PARTICIPANTS AND REPORT AUTHORS

Shenda Baker (Harvey Mudd College)
James Berger, Organizer (Statistical and Applied Mathematical Sciences Institute)
Patrick Brady (University of Wisconsin-Milwaukee)
Kirk Borne (George Mason University)
Sharon Glotzer (University of Michigan)
Robert Hanisch (Space Telescope Science Institute)
Duane Johnson (University of Illinois UC)
Alan Karr (National Institute of Statistical Sciences)
David Keyes (KAUST and Columbia University)
Brooks Pate (University of Virginia)
Harrison Prosper (Florida State University)

17

ACKNOWLEDGEMENTS

We are grateful to Leland Jameson (MPS/DMS) and Celeste Rohlfing (MPS/OMA) for their very considerable assistance in the planning and operation of the workshop.

Appendix A – National Study Groups Face the Data Flood

Several national study groups have issued reports on the urgency of establishing scientific and educational programs to face the data flood challenges, including:

1. National Academies report: *Bits of Power: Issues in Global Access to Scientific Data,* (1997) downloaded from http://www.nap.edu/catalog.php?record_id=5504
2. NSF report: *Knowledge Lost in Information: Research Directions for Digital Libraries,* (2003) downloaded from http://www.sis.pitt.edu/~dlwkshop/report.pdf
3. NSF report: *Cyberinfrastructure for Environmental Research and Education,* (2003) downloaded from http://www.ncar.ucar.edu/cyber/cyberreport.pdf
4. NSF Atkins Report: *Revolutionizing Science & Engineering Through Cyberinfrastructure: Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure,* (2003) downloaded from http://www.nsf.gov/od/oci/reports/atkins.pdf
5. NSB (National Science Board) report: *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century,* (2005) downloaded from http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf
6. NSF report with the Computing Research Association: *Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda,* (2005) downloaded from http://www.cra.org/reports/cyberinfrastructure.pdf
7. NSF report: *The Role of Academic Libraries in the Digital Data Universe,* (2006) downloaded from http://www.arl.org/bm~doc/digdatarpt.pdf
8. National Research Council, National Academies Press report: *Learning to Think Spatially,* (2006) downloaded from http://www.nap.edu/catalog.php?record_id=11019
9. NSF report: *Cyberinfrastructure Vision for 21st Century Discovery,* (2007) downloaded from http://www.nsf.gov/od/oci/ci_v5.pdf
10. JISC/NSF Workshop report on Data-Driven Science & Repositories (2007) downloaded from http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf
11. DOE (Department of Energy) report: *Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale,* (2007) downloaded from http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/DOE-Visualization-Report-2007.pdf
12. DOE report: *Mathematics for Analysis of Petascale Data Workshop Report,* (2008) downloaded from http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/PetascaleDataWorkshopReport.pdf
13. NSTC Interagency Working Group on Digital Data report: *Harnessing the Power of Digital Data for Science and Society,* (2009) downloaded from http://www.nitrd.gov/about/Harnessing_Power_Web.pdf
14. National Academies report: *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age,* (2009) downloaded from http://www.nap.edu/catalog.php?record_id=12615