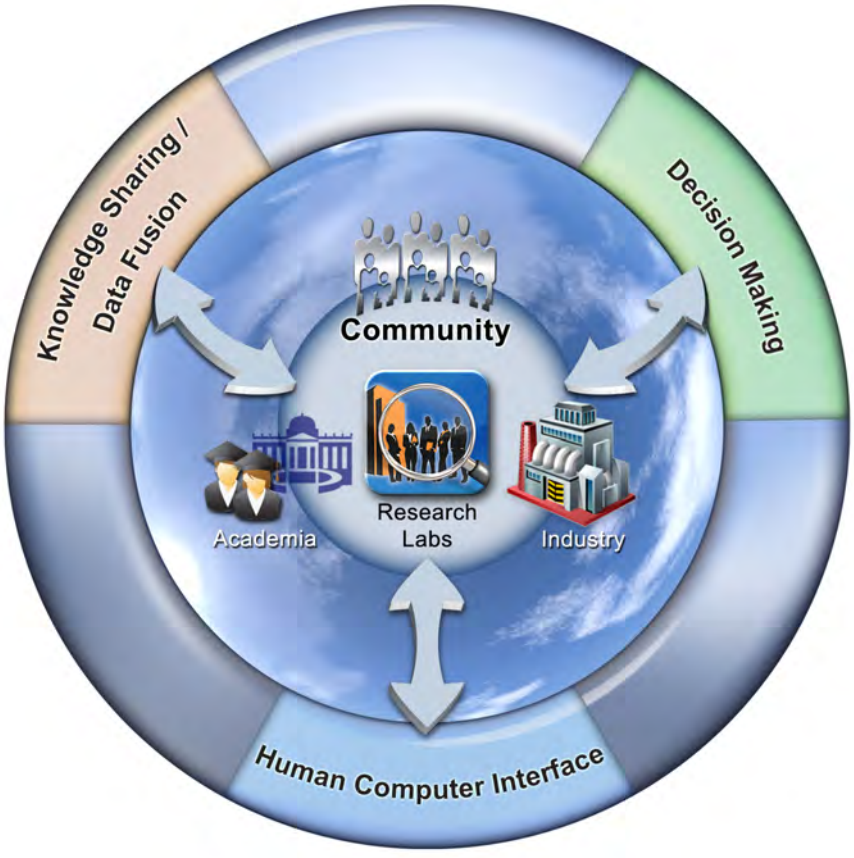
# Accelerating Scientific Knowledge Discovery (ASKD)

## Working Group Report

## August 2013



## Executive Summary

Sustained scientific progress over the next decade and beyond will require new **advanced discovery ecosystems** quite different from the computational and collaborative environments in which most research is performed today. These systems will need to connect increasing numbers of scientists, enable use of data and computational services at unprecedented scales, foster scientific discoveries based on ever more complex cross-disciplinary hypotheses, facilitate the immediate sharing and exchange of existing and emerging knowledge, and provide mechanisms for timely control of and feedback to instruments and simulations. To achieve this goal requires computer science research advances in multiple areas.

We propose here a new research program within the DOE Office of Advanced Scientific Computing Research (ASCR) designed **to shorten significantly the time needed to transform scientific data into actionable knowledge by enabling the dynamic creation of advanced discovery ecosystems.** The overarching goal is to not simply sustain but to accelerate the remarkable pace of scientific knowledge discovery that has defined the 20th century. Hence the name of the proposed program: **Accelerating Scientific Knowledge Discovery.** These advanced discovery ecosystems will provide interconnected communities of scientists with the technologies and infrastructures needed to access and use a diverse set of computational and storage resources and extreme-scale data. The seven research areas identified are:

- Knowledge management: collection, representation, storage, exchange and sharing of large quantities of diverse information.
- Rapid information and knowledge-based response: decision-making mechanisms and support in near real-time.
- Data and knowledge fusion: integration of data and knowledge into consistent, accurate, and useful representation of the same or related real-world objects.
- Dynamic data and information resource collection: discovery, allocation and management mechanisms.
- Composition and execution of end-to-end scientific processes: configurable workflow methodologies spanning heterogeneous communities, applications, and environments.
- Human computer interaction: user interface, access, and interaction through computational environments and mechanisms.
- Trust and attribution: secure access, verification, and acknowledgement of contributions.

In this report, we present an overarching vision for a research program aimed at the grand challenge of accelerating scientific knowledge discovery; summarize a set of supporting goals and science drivers; and list important computer science research challenges that must be addressed to make the vision a reality. This material is based on wide-ranging discussions within and beyond DOE as well as contributions made by participants in an intensive two-day workshop.

The research agenda presented responds to a compelling and urgent set of needs that span numerous DOE scientific communities. We posit that this agenda can best be pursued through a new program within ASCR that will work in close conjunction with, and as an extension to, existing programs and institutes and that will complement domain science-motivated programs with related themes such as SciDAC and KBASE. The focus of the new program would be on foundational computer science research with the goal of providing transformative new methods and approaches. The resulting capabilities when adapted and adopted by domain sciences will enable the new advanced discovery ecosystem.

# 1 Introduction

The translation of scientific results into new knowledge, solutions, policies, decisions and, ultimately, economic, social or environmental benefits, is foundational to all DOE science and energy missions. However, while advances in experimental and computational technologies have led to an exponential growth in the volume, velocity, and variety of data available for scientific discovery, advances in technologies to convert this data into actionable knowledge have fallen far short of what science communities need to deliver timely and immediately impacting outcomes. Acceleration of the scientific knowledge discovery process is essential if DOE scientists are to continue making major contributions in their scientific disciplines.

Responding to this unmet need, the DOE Advanced Scientific Computing Research (ASCR) program convened a working group (WG) to assist in identifying the next generation computer science research themes that can lower the barriers to scientific knowledge discovery. Consulting closely with a diverse set of DOE science communities, that group produced this report. The overarching recommendation is that ASCR create an Accelerating Science Knowledge Discovery (ASKD) program, to research and develop the technologies, frameworks, and theories needed to create the advanced discovery ecosystems that will allow DOE researchers to work effectively with growing data volumes, data heterogeneity, and geographical distribution of computational and storage resources. The report endorses and extends the findings of the Scientific Collaborations for Extreme-Scale Science Workshop Report (SCESS) from December 2011[1], in particular by identifying critical gaps in currently existing research programs.

The material in this report is based on a detailed review of numerous existing DOE science requirements documents as well as interviews with various science community members that were used to identify science challenges facing domain science communities. Ultimately, twenty-one domain science case studies were prepared to summarize these challenges. From these case studies and related sources, computer science challenges that impact one or more of these domain science communities. The WG members then distilled these challenges into common themes corresponding to areas of computer science in which research is most needed. The results and conclusions of this process were vetted at a workshop attended by science community leaders.

The rest of this report summarizes the vision and strategy (§2), provides a brief description of the needed research areas (§3), lists representative science challenge examples gathered from domain scientists (§4), and provides more specifics on the computer science research topics (§6).

# 2 Vision

The proposed computer science research will enhance the value of existing DOE science facilities, contribute to the creation of a new generation of user facilities, and accelerate the use of extreme-scale computational facilities. We articulate the overarching vision as follows:

**To shorten significantly the time needed**
**to transform scientific data into actionable knowledge**
**by enabling the dynamic creation of advanced discovery ecosystems.**

The winners of the increasingly global and aggressive 21st Century race for scientific discovery will be those who are more agile in formulating, testing and gaining new insights from hypotheses through physical experiments, computational simulation, and data analysis.

[1] http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Scientific_Collaborations_for_Extreme_Scale_Science_Report_Dec_2011_Final.pdf

DOE manages the world's largest collection of unique scientific user facilities[2] for open science research. Scientists using these facilities need new adaptive mechanisms that appropriately balance automation with expert intervention. They must extract trustable knowledge and discovery in a timely manner against a background of growing data diversity, rates, and volumes as well as human diversity in locale, experience and scientific discipline. Full understanding of objects and processes can often only be gained through a combination of different scientific techniques. Scientists require better methods to formulate and execute discovery processes across facilities and disciplines. Equally important are methods for establishing persistent syntheses of knowledge and results that span across disparate facilities and permit large multidisciplinary teams to work together effectively and efficiently. ASCR's natural role in this domain is to research and develop new theories, mechanisms, tools, and services to accelerate this scientific discovery process.

We envision a program that will support research aimed at:

- Creating new methods for rapid and scalable collaborative analysis, interpretation, modeling, simulation, and prediction of complex phenomena relevant to DOE, to create new sharable insights;
- Constructing a cyber framework that supports increasingly complex real-time analysis and knowledge navigation, integration, and creation processes that ensures timely sharing of knowledge;
- Developing a high-speed, secure, and easily traversed web of scientific user facilities, computational facilities, and data, information, and knowledge resources;
- Demonstrating broad applicability within diverse scientific communities that span DOE and academic R&D institutions, educational partners, and industry;
- Developing measures of productivity and effectiveness within scientific communities to assess the impact of increased collaboration;
- Developing the programmatic structure needed to support an appropriate mix of short (3 year), medium (6 year), and long (12 year) term research activities; and
- Establishing mechanisms for funding the development of responsive sustainable knowledge discovery efforts in an expanding number of scientific communities over time.

In so doing, the program will produce answers to questions like:

- How can we co-design tools and technologies in order to accelerate evidence-based knowledge navigation, production, and dissemination for exascale science?
- How do we enable scientists to work efficiently with disparate and distributed large-volume, potentially streaming, complex sources of data and knowledge?
- How can we deliver immediate and reactive knowledge discovery and creation processes for research teams that span multiple locations?
- How can we develop diagnostic and predictive models to understand the time spent in the science process, and how can we reduce the bottlenecks observed in our predictions?
- How can we understand potential trade-offs when configuring and operating complex reactive real-time solution environments for scientific knowledge discovery?
- How can we rebalance the human, machine and instrument interactions that enable scientists to make the leap from data to knowledge faster?

[2] http://science.energy.gov/~/media/_/pdf/user-facilities/Office_of_Science_User_Facilities_FY_2012_rev1.pdf

- How can we accelerate the assembly of science discovery infrastructures that effectively and efficiently combine scientific instruments, data resources, and computational resources to answer new questions at the forefront of science?

We note that while data, information and knowledge are treated equivalently in some contexts in the end-to-end scientific process, here knowledge is differentiated, in particular, as including the outcome of experience, values, contextual information, and insights, evaluating and incorporating new data, information, and existing knowledge. Knowledge may be codified in formal description languages, captured in text, or represented in other forms.

Successful realization of the vision articulated here will provide a foundation for the more effective use of facilities and resources across the DOE science complex, from leadership-class exascale systems, through multi-user facilities and local community facilities, to individual investigator laptops and smart devices. Long-term sustainability will come from integrating hardened research results into the experimental facility infrastructure and scientific operating procedures, which will in turn stimulate new challenges and research for future continuing innovation.

A large-scale multi-year program will be required to address the identified seven areas of computer science research in sufficient detail to ensure sustained acceleration of the knowledge discovery process. While some domain-focused programs, such as the Systems Biology Knowledgebase (KBase) and the LHC, have done relevant work in some of the defined areas, there was a consensus among report authors and workshop participants that point solutions will not scale to multiple DOE science communities. Thus there is a gap in the coverage of ASCR fundamental computer science research in these areas. Such basic research would augment and extend existing programs in these and other areas such as extreme scale Data and SciDAC, and would provide capabilities in areas needed to generate end-to-end integrated solutions that meet the needs of multiple DOE science communities.

# 3 Research Needed

To define the computer science research challenges the WG members engaged with a wide range of DOE science communities with demonstrated unmet needs for enhanced capabilities. The WG members then reviewed existing reports and conducted interviews with these science communities to articulate their research goals, and captured science community challenges in written form. The challenges identified by the science domains were compared to determine crosscutting needs and drove categorization of computer science research areas to address gaps in capability, usability and scalability. The WG then came to consensus on these shared needs and distilled them down into seven themes:

- Knowledge management: collection, representation, storage, exchange and sharing of large quantities of diverse information, both raw data and more structured knowledge.
- Rapid knowledge-driven response: fully and partially automated decision-making mechanisms capable of leveraging diverse knowledge to provide high quality decisions in near real-time.
- Data and knowledge fusion: integration of data and knowledge into consistent, accurate, and useful representation of the same or related real-world objects.
- Dynamic discovery, allocation, and management mechanisms for the data resources, computational resources, and other resources required to achieve the science goals.
- Composition and execution of end-to-end scientific processes: configurable workflow methodologies spanning heterogeneous communities, applications, and environments.
- Human computer interaction: enabling user interface, access, and interaction through computational environments and mechanisms.
- Trust and attribution mechanisms to enable secure access, verification, and acknowledgement of contributions.

These areas identified are described in more detail below, together with a summary of representative science domain drivers for each area.

## 3.1 Knowledge Acquisition, Management, and Sharing

Needs specific to data and knowledge acquisition extend past beyond simply capturing data from an experiment or simulation. They include data representation and interpretation; synthesis of more structured knowledge from raw data; handling complex unstructured search and access queries; and sharing and exchanging information with peers. The synthesis of knowledge from data will often involve additional inputs of information, insight, and/or actions from one or more persons as part of the scientific process. Knowledge must be captured manually and digitally, along with other digital products, such as data, software, workflows, algorithms, papers to ensure inclusion in, justification for, and validation of the scientific outputs. Thus there is the need to provide and account for, describe and respond to, capture and process, actions and stimuli from humans.

> Increasing need to integrate climate research with research in other fields such as biology, geology and energy, new collaborative approaches are needed not only to share data, but also to acquire, share, and integrate knowledge in a dynamic discovery environment…*Climate example*

## 3.2 Rapid Knowledge-Based Response and Decision Making Mechanisms

Decisions and responses to the occurrence of events or analyses of data are most valuable when they are fast enough to match the natural timing-structures of the science processes: that is, to proceed at or faster than the speed of the evolving phenomena. To maximize returns on expensive investments, reduce the risk of no outcomes at all, and increase the likelihood of new discovery, scientists need to react at the rate their systems operate. Thus, we encounter challenges such as mapping emerging streaming results to immediate knowledge; creating candidate hypotheses from existing knowledge and results that explain emerging results; and evaluating conflicting hypotheses that are possible options. Meeting these goals will allow for systems that can *actively steer scientific processes* taking decisions on the response to both anticipated and unanticipated events, and as a result improve the quality of outcomes and the return on research investments.

> Plasma pulses are taken with 20 minutes in between each 10 s pulse. Geographically dispersed teams take decisions between pulse intervals placing a large premium on rapid data analysis that can be assimilated in near real time …*Fusion Energy example*

## 3.3 Data and Knowledge Fusion

Data and knowledge fusion is the process by which multiple data and knowledge representing the same or similar real-world artifacts are integrated into a consistent, accurate, and useful representation. The growing diversity of data sources and types, and the increasing importance of multi-disciplinary science, makes it increasingly imperative to combine and compare data from different sources, measurement modalities, and time periods. In other cases archived data and knowledge artifacts will be used by new science teams for purposes unrelated to the original data collection purpose. Discovering and extracting these data and knowledge artifacts remains an unmet challenge that a single science community is unprepared to meet.

> Measurements collected at different scales are an ongoing challenge. Integration of measurements with scientific results from disparate domains such as biology and geology is a growing need… *Atmospheric Measurement Science example*

## 3.4 Dynamic Resource Collection, Discovery, Allocation, and Management

The agility to use any appropriate available resource and to ensure that all data needed is dynamically available at that resource is fundamental to future science discoveries. In this context “resource” has a broad meaning and, given our focus on knowledge, includes data and people as well as computing and

Timely knowledge about the availability and state of resources globally need to be used by the scientific processes in order to ensure the adequate throughput needed to keep up with the data analyses;. *High Energy Physics example*

other non-computer based entities: thus, any kind of data—raw data, information, knowledge, etc., and any type of resource—people, computers, storage systems, scientific instruments, software, resource, service, etc. In order to make effective use such resources, a wide range of management capabilities must be provided in an efficient, secure, and reliable manner, encompassing for example collection, discovery, allocation, movement, access, use, release, and reassignment. These capabilities must span and control large ensembles of data and other resources that are constantly changing and evolving, and will often be indeterministic and fuzzy in many aspects.

## 3.5 Composition and Execution of End-to-End Scientific Processes

A scientific process encompasses the activities performed by humans and computers as they work and interact towards a scientific goal. Increasingly ambitious science goals result in processes that are larger, longer, and more complex. These processes increasingly span multiple heterogeneous computing environments and must be reactive to intermediate results and conditions, while also being resilient to errors and defects. Research into and implementation of workflow methodologies are long-standing areas of attention. These methods and technologies must now be expanded to address end-to-end solutions that include multiple applications across multiple systems, multiple timespans from the immediate (milliseconds) to the long-term (weeks, years, decades), and support for interruption and resumption of processes from external, asynchronous stimuli. Research goals here also need to encompass the modeling and design of collaborative science activities in a broad sense.

Challenges include warping small, high resolution image fields and developing pattern recognition tools that allow one to go from spectra per pixel raw data to histograms of elemental distribution by flow models for fluids based on tomographic data of porous materials... *X Ray Imaging at the Nanoscale example*

## 3.6 Human Computer Interaction

Knowledge discovery is ultimately a human activity, with computational tools acting, ideally, to amplify rather than hinder the application of human intelligence and to facilitate rather than hinder human-to-human interactions. The frequently primitive means by which researchers interact today with computational tools, data, and each other was identified as a significant impediment to research progress. Challenges were identified resulting from the need to address a wide range of time frames, from sub-second to multi-year; the complexity of existing user interfaces; an increasing need for rapid, flexible, agile, and extensible interactions; and the increasing geographical, technical and cultural diversity of human collaborations. These challenges are distinctive as success is dependent on a holistic approach that addresses both human and technology needs across diverse scientific domains.

Cell to watershed questions require an ability to curate and analyze many types of data sets. tools for user interfaces to enable a clear understanding of the data at hand is necessary for this community … *Biological and Environmental Science example*.

## 3.7 Trust and Attribution

Information security, privacy, digital rights, and intellectual property (IP) protection (all parts of trust) are frequently essential to the scientific discovery process. So too is the attribution of data or knowledge to the author or process that produced it. Past work has tended to focus on authentication and authorization; future work needs to extend to issues of trust, audit, and attribution for human

Due to the high competitiveness, only particularly promising experiments will have a chance to compete for follow up studies. Appropriate access control and sharing is crucial for the information gained. .. *Linac Coherent Light Source example*

actors, computational processes, and the data and knowledge that they work with. Dynamically established, cross-institutional collaborating teams lead to particular challenges. The ultimate goal must be that knowledge claims resulting from research, even when distributed, multi-disciplinary, collaborative, and extending over long time periods, can be rigorously assessed with respect to truth, accuracy, and provenance.

# 4 Science Drivers

The WG collected science drivers from a wide variety of scientific communities that were then used to inform discussions of computer science challenges. Each science driver is an example provided by that science community with the goal of capturing important challenges of relevance to ASKD. While certainly not complete in their coverage of all DOE science domains and problems, they do provide span a wide range.


Table 1 lists the 21 science drivers and, for each, indicates the computer science areas in which research gaps were identified during discussions with the scientific groups and/or at the workshop. We provide in the following brief descriptions of each science driver. More complete descriptions and pointers to background material are available at http://www.orau.gov/ASKD2013/reference.htm. We group the science drivers into six groups: Climate, Environment, and Biology; Materials; Combustion; Energy Fusion; Nuclear and High Energy Physics; and Multi-Domain Science Facilities.

**Table 1:Science drivers and computer science research areas identified**

| | Knowledge Management | Rapid Decision Making | Data Fusion | End/end Scientific processes | Dynamic Resource Management | Human computer Interaction | Trust and Attribution |
|---|---|---|---|---|---|---|---|
| Atmospheric Radiation Measurement | X | | X | | | | |
| Climate Modeling | X | | X | | X | | X |
| Climate and Ecosystems | X | | X | X | | | |
| Bio-Imaging | X | | X | X | X | X | |
| Multi-Scale Biological & Environmental Analysis | X | | X | X | | | |
| Systems Biology Knowledgebase | X | X | X | | X | X | X |
| Joint Genome Institute | X | X | | | X | | |
| Center for Integrated Nanotechnology | X | X | | | | | |
| Heterogeneous Catalysis | X | X | X | X | X | | |
| Materials Genome | X | X | X | X | X | | |
| Combustion | X | | X | | X | | |
| Fusion Energy | X | X | X | X | | X | X |
| Nuclear Physics | | X | X | X | X | | |
| High Energy Physics | X | X | X | | X | X | |

| | Knowledge Management | Rapid Decision Making | Data Fusion | End/end Scientific processes | Dynamic Resource Management | Human computer Interaction | Trust and Attribution |
|---|---|---|---|---|---|---|---|
| Accelerator Science | | X | | X | | | |
| ALS and APS Light Source | X | X | X | | | | |
| Linac Coherent Light Source | X | X | X | | | | X |
| Spallation Neutron Source | X | X | X | | | | |
| Environmental Molecular Science | X | X | X | | | | |
| Multi Domain Simulations | X | | X | | | X | X |
| X-Ray Imaging at Nanoscales | X | X | X | X | | | |

## 4.1 Climate, Environment, and Biology

**Atmospheric Radiation Monitoring:** The BER Atmospheric Measurement Program operates a variety of fixed and mobile long-term measurement facilities to collect data related to cloud formation, sunlight energy fate, aerosol formation/decay, and aerosol interactions with clouds. The more than 2000 data streams collected 24/7 year round from US and internationally located research sites are continuously quality controlled, analyzed, and processed. Selected data streams are combined with each other and further measurement results collected at other facilities around the world to create Value Added Data products in support of the climate research community. The fusion of atmospheric measurements collected at different scales and by different instruments has been an ongoing challenge for the ARM facility. Looking into the future this challenge will increase significantly as their user community grows, requiring the integration of ARM measurements with scientific results from as disparate domains as biology and geology.

**Climate Modeling:** Large climate simulation projects such as the international Climate Model Intercomparison Project (CMIP), the US Community Earth Systems Model (CESM), and the DOE BER Climate Science for a Sustainable Energy Future (CSSEF) involve hundreds of climate scientists. They require the capabilities to create and share complex input data sets, created from sources such as the ARM facility above; execute large-scale simulations, analyze the results across complex ensemble runs; and share their results with the wider community. The Earth System Grid Federation (ESGF) is providing some required capabilities, but with an increasing need to integrate climate research with research in other fields such as biology, geology and energy, new collaborative approaches are needed not only to share data, but also to acquire, share, and integrate knowledge in a dynamic discovery environment.

**Climate and Ecosystems Research:** Developing a predictive understanding of climate change and terrestrial system response and feedback to climate requires the integration and analysis of over 2,000 different biological, geological, and atmospheric data streams spanning multiple formats, different units of measure, different spatial and temporal frequencies and resolutions, and different system organization. A primary challenge is producing scientific outcomes from data fusion; combining a number of diverse data streams, totaling more than 1.5 Petabytes/a year by 2015, depending on different processing algorithms per stream, in an environment where scientists (today) are able to cope with less than an order of magnitude fewer streams (and less 100GB data) per analysis task.

**Multi-Scale Biological and Environmental Analysis:** Microbial and plant communities and their interactions in natural environments regulate the fluxes of most life-critical elements, control the production of food and biofuel feedstock, control water quality, and regulate the flux of major greenhouse gases to the atmosphere. Protecting these resources requires an understanding of and capabilities to handle disparate datasets associated with multiple systems, including microbial,

ecosystem, and climate. An understanding of in-situ microbial functioning in natural communities is evolving via a pairing of community sequencing with larger environmental system observations and experiments. An example includes the BER research ongoing at the Rifle, CO Site, which is exploring how the microbial metagenome influences biogeochemical functioning of the larger system relevant to contaminant transport and carbon cycling and how the metabolic activity is expected to change with climate and land use. Addressing such cell-to-watershed questions requires an ability to curate, integrate, simulate and analyze many different types of datasets - including those associated with molecules, microbes, microbial communities, aqueous geochemistry, groundwater fluxes, sediment physical and mineralogical properties, vadose zone moisture and fluxes, soil properties, organic matter, surface water distribution and fluxes, and vegetation. Providing tools for user interface to enable a clear understanding of the data at hand and the ability to configure remote instrumentation is necessary for this community of scientists.

**Bio-Imaging:** Optical and electron microscopes are, through camera resolution improvements and automation, becoming data-pipelines for computational analysis and are increasingly limited by data to knowledge issues. In these modes of operation a single electron microscope "shot" can generate a 32k x 32k x 20k block of voxel data several terabytes in size. Multi-modal imaging techniques that probe samples with multiple spectroscopies for a greater depth of understanding are increasingly computational, high throughput, and informatic in nature. Image analysis, through machine learning and other scalable computing techniques, can transform the massive amount of pixel data generated into biological objects that inform statistical or geometric models of biology. The 2012 BERAC report identified the need for scalable searchable data services that allow scientists to collaborate in model building and testing. In the space of bio-imaging that will require large scale analysis pipelines that do expert-guided and hands-off conversion of pixel data into shared knowledge based models. Early work by the Protein Atlas, NCEMhub, and Microbial Communities projects inform this trend toward computational bioimaging.

**Systems Biology Knowledgebase:** KBASE is a collaborative science environment in which researchers can find, formulate, and share testable hypotheses across biological research areas including microbes, plants, and their communities. Systems biology is becoming intensely data-driven as researchers strive to better test their inferences and models, which involve large diverse data sources. In the interest of provenance and reproducibility, both data and analysis must be encapsulated into narratives that are reusable and repeatable with ease. For example a search for genetic variations in plant populations might illuminate variations in organism or tissue traits. This search as well as the underlying data and analysis, along with annotations experts, must be made available to distributed teams of system biologists. KBASE is growing to include ever-larger data sets and more detailed analysis. An effort to accelerate the scale and speed of systems biology will require bridges between scalable computational resources and knowledge systems familiar to biologists.

**Joint Genome Institute:** The Joint Genome Institute has undergone a transition from wet chemistry during the Human Genome Project toward a mostly-digital approach to sequencing in which commercially available *short-read* sequencers produce enormous amounts of data sequences that require computational analysis for assembly. Post-analysis steps require databases and science gateways. In the 2010 BERAC Long-Term Vision three main computational research drivers are identified: 1) scalable data analysis paradigms for simulation and experiment in biology, climate, and environment; 2) protocols for data integration across science areas that improve productivity; and 3) software frameworks that deal with complexity inherent in systems biology data in supercomputing environments, by for example automating large ensemble computations and complex pipelines. Access to these resources must be enabled through an easily understandable HCI.

## 4.2 Materials

**Center for Integrated Nanotechnologies:** The Materials Genome Initiative highlights the need for an open innovation environment to accelerate the discovery of new materials and shorten the cycle time from laboratory discovery to commercially viable materials, particularly through effective computational approaches screening for new materials to meet specific functional objectives. In key arenas, such as future energy technologies, those functional characteristics hinge on the electronic and optical excitation characteristics of materials. Due to the heavy computational burden associated with present-day predictive tools for electronic and optical excitations, there are significant practical limitations both to the complexity of the system and the number of accurate calculations that can be performed even for much simpler systems. Both issues present significant impediments to incorporate accurate calculations of excited state properties into the in silico materials discovery effort. Data and knowledge fusion for comparison of experimental and simulation data is necessary but difficult to achieve. Bridging the gap will lead to broad impact across the materials community.

**Heterogeneous Catalysis:** Research in heterogeneous catalysis has established a standard model for analyzing the activity, selectivity, and stability of catalysts. This standard model enables researchers to systematically screen tens of thousands of potential catalytic materials on the computer using atomic-scale simulations. The volume of data has led to 'myriads' of new initiatives to generate and curate massive amounts of data. Ultimately, coordination between initiatives will be essential to store data to enable recognition of properties or correlations using machine-learning techniques. This challenge is largely dependent on the ability the share knowledge.

**Materials Genomics:** Advances in energy storage, structural alloys, and photocatalysis are increasingly being met through simulation surveys across wide spaces of candidate materials. The Materials Project (materialsproject.org) has 3500 science users of a database developed for materials discovery that has shown successful application to areas such as energy storage, functional electronic, and other emerging materials science challenges. Barriers to expanding its success to other areas of materials research are software for small teams to compose and execute methods of O(10^6) tasks, standard and interoperable interfaces so diverse applications can act on the same data, databases that are analytics friendly, making the data more web accessible, methods that improve finding gaps and inconsistencies in large complex data sets and community-wide authentication and access. Query languages for materials properties, error detection in highly variable data sets, and accessible machine learning at scale will advance the MGI research agenda.

## 4.3 Combustion

**Combustion Research:** The turbulent combustion community requires software tools and infrastructure for broad access to large data sets from high-fidelity exascale simulations and from large experimental data sets and to enable the collaborative development and application of methodologies and algorithms for quantitative comparison of multi-dimensional experimental data and simulations. End-to-end workflows are needed to orchestrate large-scale data movement, data transformations, and analysis and visualization of large, turbulent-combustion computations and experiments. These workflow tools need to operate in situ in a large-scale computation or experiment, in-transit as data is streamed off the scientific instrument or computer, or as a post analysis step. The in situ and in-transit workflows are required for computational steering and to reduce the amount of salient data written to persistent storage for archival purposes. The infrastructure needs to provide digital archiving of key experimental and simulation benchmark validation data sets for model assessment and capturing the results of model calculations, with ease of access, to better document progress and avoid duplication. Curation and metadata tools are required to capture the community establishment of best practices, methods of analysis, baseline data for experiments and simulations, all based on quality metrics and uncertainty quantification (UQ) methods where possible. The

chemistry sub-community needs cyberinfrastructure to develop comprehensive and reduced combustion mechanisms from curated data and models needed to capture relevant combustion chemistry and accurately simulate target observables in turbulent reacting flow simulations.

## 4.4 Fusion Energy

**Fusion Energy Sciences:** The current generation of magnetic fusion experiments operate in a pulsed mode where in any given day, 25–35 plasma pulses are taken with approximately 10 to 20 minutes in between each ~10 s pulse. Throughout the experimental session, hardware/software plasma control adjustments are made as required by the science. These adjustments are debated and discussed among the experimental team within the roughly 20-minute, between-pulse interval. This mode of operation places a large premium on rapid data visualization and analysis that can be assimilated in near real-time by a geographically dispersed research team. Looking toward the future, ITER is an internationally sponsored burning plasma experiment under construction in France and is the next major step in the program aimed at proving the scientific viability of controlled fusion as a practical energy source. The ability to share complex visualizations and applications among remote participants is a necessary adjunct for scientific dialogues. Interpersonal communications media needs to be enriched via remote sharing of displays and applications among researchers. Distributed shared displays will be an important element supporting remote control rooms and need to operate in the high latency environment inherent to long distance communication.

## 4.5 Nuclear and High Energy Physics

**Experimental Nuclear Physics:** The RHIC and LHC Nuclear Physics science programs generate extremely complex datasets at scales that are expected to grow to even larger samples, and reach the exa-scale regime, as the experiments pursue an aggressive set of upgrades. Upgrades will increase data acquisition rates (by a factor of two) and also add newer detector data types, increasing the need for real-time decision making based on analysis of multiple, diverse streams of data. New data flows will allow more sophisticated filters (High Level Trigger or HLT, pattern recognition) and require better data reduction and repacking methods, two examples being fast online tracking and “pile-up” rejection at the source. The experiments are also considering moving detector calibration processes closer to the data-taking, where managing the knowledge generated—based on real-time first pass track reconstruction in HLT and collision vertex reconstruction—leads to improved selection of collision events with the highest potential for key physics measurements. The prospect of agile framework with processing “blocks” moving from online to offline suggests the need for asynchronous IO and novel interchangeable data representations.

**HEP Energy Frontier:** Current Large Hadron Collider upgrades will increase data collection rates up to 10KHZ (from the ~400 HZ in 2012) and also an order of magnitude complexity per “event.” Today’s worldwide analysis engine, serving several thousand scientists, will have to be commensurately extended in the cleverness of its algorithms, the automation of the processes, and the reach (discovery) of the computing, to enable scientific understanding of the detailed nature of the Higgs boson. As two different examples: 1) timely knowledge about the availability and state of resources globally need to be used by the scientific processes in order to ensure the adequate throughput needed to keep up with the data analyses; and 2) the approximately forty different analysis methods used to investigate the detailed characteristics of the HIGGS boson (many using machine learning techniques) must be combined in a mathematically rigorous fashion to have an agreed upon publishable result.

## 4.6 Multi-Domain Science Facilities

The following facilities provide a variety of research infrastructure capabilities to the DOE science communities, they are hereby not focused on the support of a particular science domain, but many of their instruments will be used by a multitude of science areas, often in rapid succession.

**Accelerator Science;** Researchers supported by the DOE Offices of High Energy Physics, Nuclear Physics, and Basic Energy Sciences, DOE NNSA, and NSF are now exploring advanced accelerator concepts and advanced beam manipulation systems that could revolutionize the next generation of particle accelerators. Real-time or near-real-time access to distributed HPC resources will make it possible to quickly analyze data and perform parallel simulations to help guide experiments. The combination of experiment with timely large-scale simulation would provide an otherwise inaccessible window by which we can gain insight into the physics of these systems where particle beams, lasers, plasmas, and radiation are all interacting simultaneously.

**Light Source Science:** The ALS and APS support a large number of distinct beamline stations with varying imaging capabilities used by 20,000 scientists worldwide. It is expected that in the next 1-2 years, Advanced Light Source (ALS) scientists will see their data volumes grow from 65 TB/year to 1.9 PB/year, driven in particular by rapidly evolving detector technologies. Likewise APS produces 150 TB/month. Some of their instruments are expected to reach data rates of up to 20GB/s. The light source facilities are facing two major challenges in the coming years, how to enable their users to acquire and share their gained knowledge effectively against the background of these rising data rates, and optimized support for experiments by providing rapid access to analyzed results and knowledge to inform experimental steering. Beamline science will be advanced by CS research that makes these huge data volumes tractable for analysis, especially in providing prompt analysis pipelines that provide computational insight back to the beamline to improve the experiment while it is in progress.

**Linac Coherent Light Source:** Quite different from traditional laser facilities and light sources, the Linac Coherent Light Source (LCLS) is providing a unique five-days long time scale on which design and simulation support of as-shot conditions and database searches of diffraction patterns must occur to help with real-time decision-making during the experiment. Simulations and searches that occur after the five-day long experiment will only be of limited use because of the long time period one must expect between experiments. Computer science research that enables the rapid analysis of diverse data in support of the decisions is needed. In addition, due to the high competitiveness, only particularly promising and successful experiments will have a chance to compete for follow-up studies, for which research into trust and attribution is needed to ensure appropriate access and sharing of the information gained. End-to-end processes are critical to demonstrate the feasibility of the experimental concepts, the best use of beam time during data collection and post experimental analysis to confirm the concepts.

**Spallation Neutron Source:** Neutron facilities like many other multi-science research facilities are over subscribed and are not able to provide their capabilities to as many scientists as they would like. A key factor in this is the time each experiment takes. The expense of the Spallation Neutron Source could serve more experiments and produce more science if the data being acquired is analyzed in near-real-time to determine if enough data has been collected to achieve the scientific goal of the experiment and thus data taking can be stopped. To achieve this goal new techniques would be needed to enable multiple collaborators to compose and execute real-time analytics, fuse data, present the results and understand the errors in the data being collected in real time. Applications that are easily configurable need to be designed and made available to enable workflow processes.

**Environmental Molecular Sciences:** The Environmental Molecular Science Laboratory (EMSL) is a national user facility that provides a collection of over 100 state-of-the-art experimental instruments

as well as high performance computing capabilities. A distinctive focus is enabling of research that requires the integration of results across different experimental capabilities (multi-modal) and computational methods. Future requirements will see an extended need for knowledge discovery, data management and sharing across facilities within BER, including the provision of seamless research environments across EMSL, JGI and the Systems Biology Knowledgebase. In addition EMSL is further developing its experimental capabilities, including for example new world leading capabilities in Transmission Electron Microscopy. These new instruments will not only deliver data at much higher rate (e.g. DTEM will push data rates from 100 images /day to 1,000,000 images/sec), but also require a new analytical environment, that allows for rapid decision-making and control based on the acquired data and knowledge to enable optimized experimental design and steering.

**Multi-Domain Simulations**: The interplay between simulation and experiments at DOE facilities is increasingly important for scientific discovery. As simulations of physical phenomena become increasingly precise and accurate, there is wider applicability to the design of facilities, detectors, and experiments, which in turn can drive subsequent improvement of the simulations. A recent example is results from the Fermi Glast Space telescope, which demonstrate that expanding shock waves from ancient supernovas are the origin of cosmic rays observed on earth. Recent simulations have demonstrated that filamentary structures in high-energy density plasmas can be driven in laser-plasma interactions and lead to the formation of collisionless shocks with properties similar to the ones that are believed to occur in astrophysics. There is a need to be able to access simulations models and experimental data from remote sites. This requires remote access by means of a well-understood interface. CS research into fusion of data from the simulations and instruments is for validation and improvement of the models, as well as ensure that the appropriate knowledge is accessible and correctly attributed.

**X-Ray Imaging at the Nanoscale:** Real-life materials are heterogeneous down to the nanoscale, so an increasing number of x-ray investigations of material properties are being combined with 2D and 3D imaging: spectroscopy, diffraction, fluorescence trace element analysis, and so on. Consider the example of looking at how titanium dioxide nanoparticles (used as potential cancer therapeutic agents) enter into cell types and into specific organelles into cells: one would like to scan a several millimeter area of a tissue to get an overview, and then take a higher resolution view of cells which have taken up the nanoparticles in a particular way, and then obtain a nanoscale view of specific targeting in organelles. Challenges include warping small, high resolution image fields into the whole object view when there may be different imaging modalities (with different field distortions, contrast, and signal to noise) involved, and developing pattern recognition tools that allow one to go from spectra per pixel raw data to histograms of elemental distribution by automatically-identified cell type, or flow models for fluids based on tomographic data of porous materials.

# 5 Realizing the Benefits from Research Outcomes

We make two recommendations that the collaborative scientific community regards as important to addressing perennial issues in any research program designed to impact scientific practice:

The program must support transitional mechanisms for bridging from pure research to implementation, production deployment, and sustained support for the methods and solutions that result from that research; and

The program should support mechanisms for capturing, preserving, and communicating knowledge gained (and lessons learned) across projects, domains, and generations.

# 6 Specific Research Topics

We provide in this section a more detailed list of specific computation and computer science research topics that emerged from discussions with DOE science domain communities. Each topic listed here is viewed as relevant to the vision "**to shorten significantly the time needed to transform scientific data into actionable knowledge by enabling the dynamic creation of advanced discovery ecosystems."** The list is not complete, indeed, we expect items to evolve over time as the scientific community's experience deepens and lengthens. We group these research topics according to whether they are common across science domains or specific to some subset of those domains.

## 6.1 Common Across Science Domains

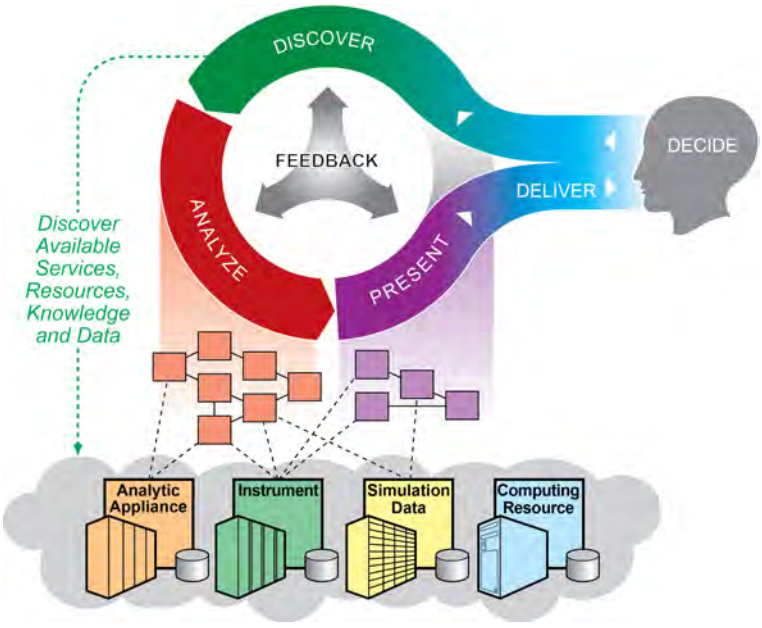



**Figure 3: The Knowledge Discovery Cycle**

### 6.1.1 Knowledge Gathering, Managing, and Sharing

*Federated Search:* Mechanisms for finding and validating diverse aspects of knowledge and data across a globally distributed set of administrative and scientific domains.

*Knowledge representation:* Development of representations that support the acquisition, access, search, evolution, and curation of large federated knowledge collections. Representations and meta-data to provide for trusted seamless long-term access to shared knowledge. Semantics for diverse knowledge to be described, characterized, and reused by new – short term or long term - defined scientific processes.

*Time to Knowledge:* Models and analyses that describe the processes involved in generating knowledge from data. Simulation methods that use such models to identify areas in which improvements can be made.

*Traceability:* Development of infrastructures and processes that give the information about the history of knowledge and react to the deposition of new knowledge. Methods for tracking the origins and characteristics of vast amounts of disparate and dispersed knowledge.

### 6.1.2 Rapid Knowledge-Based Response and Decision Making

*Automated multi-representational translation and mapping:* Efficient automated methods for mapping high-data-rate streaming results from instruments and computational processes to higher-level knowledge representations.

*Evaluation, ranking and validation across diverse hypotheses:* Approaches for the rapid evaluation and ranking of candidate hypothesis, and for resolving conflicts between conflicting hypotheses.

### 6.1.3 Data and Knowledge Fusion

*Access to integrated heterogeneous, multi-scale, and diverse knowledge across science domains:* Including long-term curation, versioning, annotation, and provenance enabling repeatability and traceability.

*Comparative analytics and validation:* Mechanisms to interpret and act on (analyze and reanalyze many times) integrated or federated data sets of already acquired and synthesized diverse data with data being newly acquired. Analytic models for creating and publishing validated and trustable results from the combination of heterogeneous data sets whose quality and uncertainties are then identified.

*Data knowledge visualization:* Interfaces and display methods to understand errors in, and trust factors of, visualization.

*Dynamic and automated translation services:* Methods to describe and in real-time combine diverse, distributed, multi-scale, and changing data as it is acquired (in flight), including measurement, assessment, propagation and management of errors across data sources and the results of computation.

### 6.1.4 Dynamic Resource Collection, Discovery, Allocation, and Management

*Federated semantic discovery:* Interfaces, protocols and environments that support access to, use of, and interoperation across federated sets of resources governed and managed by a mix of different policies and controls that interoperate across streaming and "at rest" data sources. These include: models, algorithms, libraries, and reference implementations for a distributed non-hierarchical discovery service; semantics, methods, interfaces for life-cycle management (subscription, capture, provenance, assessment, validation, rejection) of heterogeneous sets of distributed tools, services and resources; a global environment that is robust in the face of failures and outages; and flexible high-performance data stores (going beyond schema driven) that scale and are friendly to interactive analytics

*Globally optimized dynamic allocation of resources:* These need to take account of the lack of strong consistency in knowledge across the entire system.

*Minimization of time-to-delivery of data and services:* Not only to reduce the time to delivery of the data or service but also allow for a predictive capability, so scientists can deal with uncertainty in real-time decision making processes.

*Resource description and understanding:* Distributed methods and implementations that allow resources (people, software, computing) to publish varying state and function for use by diverse clients. Mechanisms to handle arbitrary entity types in a uniform and common framework – including complex types such as heterogeneous data, incomplete and evolving information, and rapidly changing availability of computing, storage and other computational resources. Abstract data streaming and file-based data movement over the WAN/LAN and over on exascale architectures to allow for real-time, collaborative decision making for scientific processes.

### 6.1.5 Composition and Execution of End-to-End Scientific Processes

*Community based job management and workflows:* Methods and techniques for the optimization of multiple, inter-disciplinary processes. Optimization of community based long-running massively parallel scientific processes that include hundreds to thousands of steps and are subscribed to by many individuals acting as a coherent entity. This is of particular relevance to workflows that must run across exascale

computing resources and other separately managed computing resources. In such situations, as reliable workflow management, checkpointing, and preemption are key.

*Management of both static and dynamic processes and environments:* Composition of scientific processes that incorporate previously generated or acquired knowledge ("parked", stored) as well as data and knowledge at the time it is modified through current analyses and knowledge (dynamic). This is of particular relevance to simulations generated and stored on exa-scale computing resources to drive simulation surveys. Reliable workflow management, checkpoint, preemption, detection of errors and their sources and capture, attribution and reactions to their occurrence.

*Optimization*: of multiple, collaborative scientific processes.

### 6.1.6 Human Computer Interaction

*Control of scientific processes:* Human-computer interface frameworks that support rapid-prototyping of analytics to dynamically adjust the methods being applied to data.

*Dynamic interaction mechanisms* to re-balance and optimize interfaces and interactions.

*Experience-based responsive user interfaces:* Configurable and flexible user interfaces that can be tailored by the individual scientist, by their "signatures."

*Improved interfaces and interactions:* Application GUIs and web APIs that make science data easily accessible through standards, formats, and new APIs.

*Generation and management of candidate hypotheses:* New methods for the creation of candidate hypotheses from existing, distributed knowledge and results that can be applied to emerging data.

### 6.1.7 Trust and Attribution

*Differential privacy frameworks:* Semantics, protocols, management and methods for distributed trust and authorization services, tools including understanding of heterogeneous layers of risk and protection, principles of granularity for sharing partial datasets and knowledge.

*Federated identity management mechanisms:* Models and algorithms, available through a universal interface, that support users and groups to manage and access data and do all their needed work independent of the native identity management system of each/any resource used. Methods for the dynamic creation, life-cycle and close-down of trustable environments in support of distributed collaborating teams.

*Software models, prototypes and implementations that allow prediction and validation of trust models to be implemented:* Operational models that describe how the system works under both normal and abnormal conditions and how trust models are created, verified, and destroyed. Mathematical models that describe the operational performance of this distributed system and to provide metrics to evaluate the operation of methodologies. New general-purpose techniques to protect in flight or static data efficiently.

## 6.2 Community Specific Research Topics

For each of for the areas listed above we also identified, during gathering of information from the science communities and subsequent discussions, an initial set of community-specific research topics, each needed by a sub-set of the communities. These topics are noted and mapped to domains in the following tables and helped in validating the focus of the research areas given above.

**Knowledge Gathering, Managing and Sharing**

| | Climate Environment Biology | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|
| *Ingest and access large amounts of disparate scientific* knowledge. Development of common open repositories where the user community can specify many different attributes to and access their store data easily. | X | | | | X | X |

| | Climate Environment Biology | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|
| *Interoperability:* Interfaces that ensure a high degree of interoperability at format and semantic level. | | | | X | X | |
| *Anomaly detection, data QA, collaborative data enhancement:* Algorithms and collaborative methods that improve knowledge discovery, finding gaps and inconsistencies in large complex data sets. | X | X | | X | | |

**Rapid Information, Knowledge Based Response and Decision Making Mechanisms**

| | Climate Environment Biology | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|
| *Continuous-loop feedback and control:* Feedback during data acquisition for data quality assurance that permits real-time steering of in-situ computational and physical experiments. | | X | | X | X | X |
| *Rapid interpretation and response mechanisms:* The ability to interpret results and verify new insights within the context of existing scientific knowledge against the background of growing data volumes and rates. | X | X | X | X | | X |
| *Geographically distributed interpretation and response mechanisms* to identify and synthesize data/knowledge in support of decision making. Includes indirect information such as when my data/software/resource has been accessed; is changed, or when a retraction occurs; Knowledge of credit for a result due to the chain of provenance. | X | X | X | X | | |
| *In-flight methods for calculating trade-offs:* Mechanisms to explore results of trade-off in accuracy of calculations to decision-making time. | | | | X | | |

**Data and Knowledge Fusion**

| | Climate Environment Biology | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|
| *Multidimensional model navigation and analysis:* Means to provide immediate understanding and actionable outputs from comparison of diverse data types generated from experiment and simulation. Infrastructure to provide a suite of choices and actionable items to apply to an experiment (e.g., when to conclude it) from diverse data and knowledge. Means to integrate in-place scientists' knowledge with that of the remote scientists. | X | X | X | X | X | X |
| *Across scales and disciplines:* for real-time decisions, data must be compared across multiple scales, and comparatively analyzed "in-near-real-time" to understand if disruptions will occur (saving millions of $'s). Algorithms that provide trustable and explainable selection (reduction) in the amount of data to be used for validation. Handling of fuzzy information (e.g. > 31, between 2 and 5, etc.). | X | | | X | | X |

**Composition and Execution of End-to-End Scientific processes (across heterogeneous environments)**

| | Climate Environment | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|

| | Biology | | | | | |
|---|---|---|---|---|---|---|
| *Domain specific abstractions:* Ontologies and software that enables small teams to engage in simulation surveys that demands O($10^6$) tasks to be marshaled through an HPC batch queue. Real time image analysis software and algorithms for ~100,000's of images for use on HPC computing. | X | X | | X | | X |
| *Flexible, resilient, and rapidly reconfigurable runtime environments:* Prioritization of critical work. | | | | X | | |

**Dynamic Resource Collection, Discovery, Allocation and Management**

| | Climate Environment Biology | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|
| *Component based software algorithms:* that provide efficient, robust, easy to specify unstructured deep searches for information in a distributed environment. | X | | | X | | |
| Mechanisms and software frameworks that enable community data inputs to configure and drive large ensembles of tasks and management of jobs. | X | X | | X | X | X |
| *Arbitrary, non-deterministic management of objects:* Description and delivery of arbitrary object level data entities (smaller than files) from distributed data stores to scientific codes | | | | X | | X |
| *Life-Cycle Artifacts:* Software frameworks that support life-cycle provisioning and support of flexible, resilient, high throughput and rapidly reconfigurable runtime job management and execution environments. Enabling discovery of appropriate tools and results | | | | X | X | X |
| *New techniques to work with deep memory hierarchies:* to manage and share information efficiently | | | | X | | |

**Human Computer Interaction**

| | Climate Environment Biology | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|
| *Domain specific interactions:* Improved designs and principles for human-computer interaction that support dynamically composable interfaces | X | | | X | X | X |
| *Flexible, dynamic interactions:* Methods to easily present and modify information about the capabilities of an interactive interface for both non-experts and experts. | X | | X | X | X | |

**Trust and Attribution**

| | Climate Environment Biology | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|
| *Data Intention:* Methods and languages for describing and adhering to intellectual property in systems where not all data is openly available nor open to all members of the groups who access the data repositories. | | | | | | X |
| *Across institutions and communities:* Libraries and repositories that allow for community-wide authentication and access, without the end user having to understand the full system. | | X | | X | | |
| *Dynamic Provenance and Accreditation*: New research techniques that support encoding and including the provenance within the streams of data. These are | X | X | | X | X | X |

| | Climate Environment Biology | Materials | Combustion | Fusion Energy | Physics | Multi-Domain |
|---|---|---|---|---|---|---|
| modified as the data is transformed into knowledge to indicate what scientist did with the data as well as point to and provide access to their code. This can provide a framework that allows the researchers to trust decisions made during complex collaborations. | | | | | | |
| *Data Proxies:* Explore means to encode errors in the data as well as the algorithms used to generate it and define levels of trust based on the characteristics of these errors. The goal is to allow scientist to understand their level of trust - interpret and understand - related to prior decisions made by collaborators or other researchers. | | X | | X | | X |

# 7 Metrics

The ASKD research program will need to define metrics to measure success and engage in continuous evaluation of progress against these metrics. The definition of such metrics was viewed as outside the scope of this report. While measuring the speed of discovery will be desirable, it may be difficult to do in practice. One potential focus for metrics is the extent to which research outcomes drive changes in design and implementation of the computing and software systems used for scientific discovery. Another is the extent to which scientists and developers of software and computing solutions attribute their success to the findings and new methodologies developed as a result of the ASKD program. The cycle of research into new techniques followed by their adoption into active scientific programs will surely result in the identification of new or extended challenges based on experience and increased understanding of the outcomes. The nature and quality of these new challenges is a third potential area for metrics.

# 8 Acknowledgements, Bibliography and Contributors

We would like to acknowledge the support for the workshop and ASKD web site provided by the ORISE team, Keri Cagle and Linda Severs. They were of particular help in arranging the venue and logistics for the July workshop.

The bibliography for this report encompasses the reference reports and papers posted at the workshop web site at http://www.orau.gov/ASKD2013/reference.htm. The workshop included a survey from attendees that was used to further input to the set of research areas described above.

Table 2 lists the scientists providing input, participants in the ASKD Workshop in Washington in July 2013, members of the WG (*'d) and other contributors to this report.

**Table 2: Contributors to this report**

| *Name* | *Organization* |
|---|---|
| Eric Ackerman | Sandia National Laboratories |
| Deb Agarwal* | Lawrence Berkeley National Laboratory |
| Nick Berente | University of Georgia |
| Amber Boehnlein* | SLAC |
| Richard Carlson* | DOE/ASCR |
| Jacqueline Chen | Sandia National Laboratories |
| Jim Costa | Sandia National Laboratories |
| Paramvir Dehal | Lawrence Berkeley National Laboratory |

| *Name* | *Organization* |
|---|---|
| Narayan Desai | Argonne National Laboratory |
| Michael Ernst* | Brookhaven National Laboratory |
| Ian Foster* | Argonne National Laboratory/University of Chicago |
| Peter Fox | Rensselaer Polytechnic Institute |
| Siegfried Glenzer | SLAC |
| Robert Harrison | Brookhaven National Laboratory |
| Barbara Helland | DOE/ASCR |
| Susan Hubbard | Lawrence Berkeley National Laboratory |
| Jens Hummelshoej | SLAC |
| Chris Jacobsen | Argonne National Lab/Northwestern University |
| Barbara Jennings* | Sandia National Laboratories |
| Andrzej Joachimiak | Argonne National Laboratory |
| Kate Keahey | Argonne National Laboratory |
| Carl Kesselman | USC/Information Sciences Institute |
| Scott Klasky* | Oak Ridge National Laboratory |
| Kerstin Kleese van Dam* | Pacific Northwest National Laboratory |
| Tahsin Kurc | Emory University |
| Jerome Lauret | Brookhaven National Lab |
| Miron Livny | University Wisconsin-Madison |
| Arthur Maccabe | Oak Ridge National Laboratory |
| Lucy Nowell | DOE/ASCR |
| Shyue Ping Ong | University of California San Diego |
| Ray Osborn | Argonne National Laboratory |
| Manish Parashar | Rutgers University |
| Valerio Pascucci | University of Utah |
| Ruth Pordes* | Fermilab |
| Jeff Porter | LBNL |
| Thomas Proffen | Oak Ridge National Laboratory |
| Rob Roser | FNAL |
| Robert Ross | Argonne National Laboratory |
| Robert Ryne | Lawrence Berkeley National Laboratory |
| David Schissel | General Dynamics |
| David Skinner* | Lawrence Berkeley National Laboratory |
| Bill Tang | Princeton University |
| Nancy Washton | Pacific Northwest National Laboratory |
| Dean N. Williams | Lawrence Livermore National Laboratory |
| Matthew Wolf | Georgia Tech |
| Barbara Helland | Department of Energy |
| Rich Carlson* | Department of Energy |
| Lai-Yung Leung | Pacific Northwest National Laboratory |
| Paul Zschack | Brookhaven National Laboratory |
| Dula Parkinson | Lawrence Berkeley National Laboratory |
| Nagiza Samatova | North Carolina State University |