Fact Sheet: Big Data Across the Federal Government
March 29, 2012

Below are highlights of ongoing Federal government programs that address the challenges of, and tap the opportunities afforded by, the big data revolution to advance agency missions and further scientific discovery and innovation.

DEPARTMENT OF DEFENSE (DOD)

Data to Decisions: The Department of Defense (DOD) is "placing a big bet on big data" investing $250 million annually (with $60 million available for new research projects) across the Military Departments in a series of programs that will:

- Harness and utilize massive data in new ways and bring together sensing, perception and decision support to make truly autonomous systems that can maneuver and make decisions on their own.
- Improve situational awareness to help warfighters and analysts and provide increased support to operations. The Department is seeking a 100-fold increase in the ability of analysts to extract information from texts in any language, and a similar increase in the number of objects, activities, and events that an analyst can observe.

To accelerate innovation in Big Data that meets these and other requirements, DOD will announce a series of open prize competitions over the next several months.

Defense Advanced Research Projects Agency (DARPA)

The ***Anomaly Detection at Multiple Scales (ADAMS)*** program addresses the problem of anomaly-detection and characterization in massive data sets. In this context, anomalies in data are intended to cue collection of additional, actionable information in a wide variety of real-world contexts. The initial ADAMS application domain is insider-threat detection, in which anomalous actions by an individual are detected against a background of routine network activity.

The ***Cyber-Insider Threat* (*CINDER*)** program seeks to develop novel approaches to detect activities consistent with cyber espionage in military computer networks. As a means to expose hidden operations, CINDER will apply various models of adversary missions to "normal" activity on internal networks. CINDER also aims to increase the accuracy, rate and speed with which cyber threats are detected.

The ***Insight*** program addresses key shortfalls in current intelligence, surveillance and reconnaissance systems. Automation and integrated human-machine reasoning enable operators to analyze greater numbers of potential threats ahead of time-sensitive situations. The Insight program aims to develop a resource-management system to automatically identify threat networks and irregular warfare operations through the analysis of information from imaging and non-imaging sensors and other sources.

The ***Machine Reading*** program seeks to realize artificial intelligence applications by developing learning systems that process natural text and insert the resulting semantic representation into a knowledge base rather than relying on expensive and time-consuming current processes for knowledge representation require expert and associated knowledge engineers to hand craft information.

The ***Mind's Eye*** program seeks to develop a capability for "visual intelligence" in machines. Whereas traditional study of machine vision has made progress in recognizing a wide range of objects and their properties—what might be thought of as the nouns in the description of a scene—Mind's Eye seeks to add the perceptual and cognitive underpinnings needed for recognizing and reasoning about the verbs in those scenes. Together, these technologies could enable a more complete visual narrative.

The ***Mission-oriented Resilient Clouds*** program aims to address security challenges inherent in cloud computing by developing technologies to detect, diagnose and respond to attacks, effectively building a "community health system" for the cloud. The program also aims to develop technologies to enable cloud applications and infrastructure to continue functioning while under attack. The loss of individual hosts and tasks within the cloud ensemble would be allowable as long as overall mission effectiveness was preserved.

The ***Programming Computation on Encrypted Data (PROCEED)*** research effort seeks to overcome a major challenge for information security in cloud-computing environments by developing practical methods and associated modern programming languages for computation on data that remains encrypted the entire time it is in use. By manipulating encrypted data without first decrypting it, adversaries would have a more difficult time intercepting data.

The ***Video and Image Retrieval and Analysis Tool* (*VIRAT*)** program aims to develop a system to provide military imagery analysts with the capability to exploit the vast amount of overhead video content being collected. If successful, VIRAT will enable analysts to establish alerts for activities and events of interest as they occur. VIRAT also seeks to develop tools that would enable analysts to rapidly retrieve, with high precision and recall, video content from extremely large video libraries.

The ***XDATA*** program seeks to develop computational techniques and software tools for analyzing large volumes of semi-structured and unstructured data. Central challenges to be addressed include scalable algorithms for processing imperfect data in distributed data stores and effective human-computer interaction tools that are rapidly customizable to facilitate visual reasoning for diverse missions. The program envisions open source software toolkits for flexible software development that enable processing of large volumes of data for use in targeted defense applications.

DEPARTMENT OF HOMELAND SECURITY (DHS)

***The Center of Excellence on Visualization and Data Analytics (CVADA*),** a collaboration among researchers at Rutgers University and Purdue University (with three additional partner universities each) leads research efforts on large, heterogeneous data that First Responders could

use to address issues ranging from manmade or natural disasters to terrorist incidents; law enforcement to border security concerns; and explosives to cyber threats.

DEPARTMENT OF ENERGY (DOE)

The Office of Science

The Office of Advanced Scientific Computing Research (ASCR) provides leadership to the data management, visualization and data analytics communities including digital preservation and community access. Programs within the suite include widely used data management technologies such as the Kepler scientific workflow system; and Storage Resource Management standard; a variety of data storage management technologies, such as BeSTman, the Bulk Data Mover and the Adaptable IO System (ADIOS); FastBit data indexing technology (used by Yahoo!); and two major scientific visualization tools, ParaView and VisIt.

The High Performance Storage System (HPSS) is software that manages petabytes of data on disks and robotic tape systems. Developed by DoE and IBM with input from universities and labs around the world, HPSS is used by digital libraries, defense applications and a range of scientific disciplines including nanotechnology, genomics, chemistry, magnetic resonance imaging, nuclear physics, computational fluid dynamics, climate science, etc., as well as Northrop Grumman, NASA and the Library of Congress.

Mathematics for Analysis of Petascale Data addresses the mathematical challenges of extracting insights from huge scientific datasets and finding key features and understanding the relationships between those features. Research areas include machine learning, real-time analysis of streaming data, stochastic nonlinear data-reduction techniques and scalable statistical analysis techniques applicable to a broad range of DOE applications including sensor data from the electric grid, cosmology and climate data.

The Next Generation Networking program supports tools that enable research collaborations to find, move and use large data: from the Globus Middleware Project in 2001, to the GridFTP data transfer protocol in 2003, to the Earth Systems Grid (ESG) in 2007. Today, GridFTP servers move over 1 petabyte of science data per month for the Open Science Grid, ESG, and Biology communities. Globus middleware has also been leveraged by a collaboration of Texas universities, software companies, and oil companies to train students in state-of-the-art petroleum engineering methods and integrated workflows.

The Office of Basic Energy Sciences (BES)

BES Scientific User Facilities have supported a number of efforts aimed at assisting users with data management and analysis of big data, which can be as big as yerabytes (10^{12} bytes) of data per day from a single experiment. For example, the *Accelerating Data Acquisition, Reduction and Analysis* (ADARA) project addresses the data workflow needs of the Spallation Neutron Source (SNS) data system to provide real-time analysis for experimental control; and the *Coherent X-ray Imaging Data Bank* has been created to maximize data availability and more efficient use of synchrotron light sources.

The Data and Communications in Basic Energy Sciences workshop in October 2011 sponsored by BES and ASCR identified needs in experimental data that could impact the progress of scientific discovery.

The Biological and Environmental Research Program (BER), ***Atmospheric Radiation Measurement (ARM) Climate Research Facility*** is a multi-platform scientific user facility that provides the international research community infrastructure for obtaining precise observations of key atmospheric phenomena needed for the advancement of atmospheric process understanding and climate models. ARM data are available and used as a resource for over 100 journal articles per year. Challenges associated with collecting and presenting the high temporal resolution and spectral information from hundreds of instruments are being addressed to meet user needs.

The Systems Biology Knowledgebase **(Kbase)** is a community-driven software framework enabling data-driven predictions of microbial, plant and biological community function in an environmental context. Kbase was developed with an open design to improve algorithmic development and deployment efficiency, and for access to and integration of experimental data from heterogeneous sources. Kbase is not a typical database but a means to interpret missing information to become a predictive tool for experimental design.

The Office of Fusion Energy Sciences (FES)
The Scientific Discovery through Advanced Computing (SciDAC) partnership between FES and the office of Advanced Scientific Computing Research (ASCR) addresses big data challenges associated with computational and experimental research in fusion energy science. The data management technologies developed by the ASCR – FES partnerships include high performance input/output systems, advanced scientific workflow and provenance frameworks, and visualization techniques addressing the unique fusion needs, which have attracted the attention of European integrated modeling efforts and ITER, an international nuclear fusion research and engineering project.

The Office of High Energy Physics (HEP)
The Computational High Energy Physics Program supports research for the analysis of large, complex experimental data sets as well as large volumes of simulated data—an undertaking that typically requires a global effort by hundreds of scientists. Collaborative big data management ventures include *PanDA* (Production and Distributed Analysis) Workload Management System and *XRootD* , a high performance, fault tolerant software for fast, scalable access to data repositories of many kinds.

The Office of Nuclear Physics (NP)
The US Nuclear Data Program (USNDP) is a multisite effort involving seven national labs and two universities that maintains and provides access to extensive, dedicated databases spanning several areas of nuclear physics, which compile and cross-check all relevant experimental results on important properties of nuclei.

The Office of Scientific and Technical Information (OSTI)

OSTI, the only U.S. federal agency member of DataCite (a global consortium of leading scientific and technical information organizations) plays a key role in shaping the policies and technical implementations of the practice of data citation, which enables efficient reuse and verification of data so that the impact of data may be tracked, and a scholarly structure that recognizes and rewards data producers may be established.

DEPARTMENT OF VETERANS AFFAIRS (VA)

Consortium for Healthcare Informatics Research (CHIR) develops Natural Language Processing (NLP) tools in order to unlock vast amounts of information that are currently stored in VA as text data.

Protecting Warfighters using Algorithms for Text Processing to Capture Health Events (ProWatch): Efforts in the VA are underway to produce transparent, reproducible and reusable software for surveillance of various safety related events. ProWatch is a research-based surveillance program that relies on newly developed informatics resources to detect, track, and measure health conditions associated with military deployment.

AViVA is the VA's next generation employment human resources system that will separate the database from the business applications and from the browser-based user interface. Analytical tools are already being built upon this foundation for research and ultimately support of decisions at the patient encounter.

Observational Medical Outcomes Project is designed to compare the validity, feasibility and performance of various safety surveillance analytic methods.

Corporate Data Warehouse (CDW) is the VA program to organize and manage data from various sources with delivery to the point of care for a complete view of the disease and treatment for individuals and populations.

Health Data Repository is standardizing terminology and data format among health care providers and notably between the VA and DOD, allowing the CDW to integrate data.

Genomic Information System for Integrated Science (GenISIS) is a program to enhance health care for Veterans through personalized medicine. The GenISIS consortium serves as the contact for clinical studies with access to the electronic health records and genetic data in order that clinical trials, genomic trials and outcome studies can be conducted across the VA.

Million Veteran Program is recruiting voluntary contribution of blood samples from veterans for genotyping and genetic sequencing. These genetic samples support the GenISIS consortium and will be attributed to the "phenotype" in the individual veteran's health record for understanding the genetic to disease states.

VA Informatics and Computing Infrastructure provides analytical workspace and tools for the analysis of large datasets now available in the VA, promoting collaborative research from anywhere on the VA network.

HEALTH AND HUMAN SERVICES (HHS)

Center for Disease Control & Prevention (CDC)

BioSense 2.0 is the first system to take into account the feasibility of regional and national coordination for public health situation awareness through an interoperable network of systems, built on existing state and local capabilities. BioSense 2.0 removes many of the costs associated with monolithic physical architecture, while still making the distributed aspects of the system transparent to end users, as well as making data accessible for appropriate analyses and reporting.

Networked phylogenomics for bacteria and outbreak ID. CDC's Special Bacteriology Reference Laboratory (SBRL) identifies and classifies unknown bacterial pathogens for effective, rapid outbreak detection. Phylogenomics, the comparative phylogenetic analysis of the entire genome DNA sequence, will bring the concept of sequence-based identification to an entirely new level in the very near future with profound implications on public health. The development of a SBRL genomic pipeline for new species identification will allow for multiple analyses on a new or rapidly emerging pathogen to be performed in hours, rather than days or weeks.

Center for Medicare & Medicaid Services (CMS)

A data warehouse based on Hadoop is being developed to support analytic and reporting requirements from Medicare and Medicaid programs. A major goal is to develop a supportable, sustainable, and scalable design that accommodates accumulated data at the Warehouse level. Also challenging is developing a solution complements existing technologies.

The use of XML database technologies are being evaluated to support the transactional-intensive environment of the Insurance Exchanges, specifically to support the eligibility and enrollment processes. XML databases potentially can accommodate Big Tables scale data, but optimized for transactional performance.

Using administrative claims data (Medicare) to improve decision-making: CMS has a current set of pilot projects with the Oak Ridge National laboratories that involve the evaluation of data visualization tools, platform technologies, user interface options and high performance computing technologies--aimed at using administrative claims data (Medicare) to create useful information products to guide and support improved decision-making in various CMS high priority programs.

FOOD AND DRUG ADMINISTRATION (FDA)

A Virtual Laboratory Environment (VLE) will combine existing resources and capabilities to enable a virtual laboratory data network, advanced analytical and statistical tools and capabilities, crowd sourcing of analytics to predict and promote public health, document management support, tele-presence capability to enable worldwide collaboration, and basically make any location a virtual laboratory with advanced capabilities in a matter of hours.

NATIONAL ARCHIVES & RECORDS ADMINISTRATION (NARA)

The Cyberinfrastructure for a Billion Electronic Records (CI-BER) is a joint agency sponsored testbed notable for its application of a multi-agency sponsored cyber infrastructure and the National Archives' diverse 87+ million file collection of digital records and information now active at the Renaissance Computing Institute. This testbed will evaluate technologies and approaches to support sustainable access to ultra-large data collections.

NATIONAL AERONAUTICS & SPACE ADMINISTRATION (NASA)

NASA's Advanced Information Systems Technology (AIST) awards seek to reduce the risk and cost of evolving NASA information systems to support future Earth observation missions and to transform observations into Earth information as envisioned by NASA's Climate Centric Architecture. Some AIST programs seek to mature Big Data capabilities to reduce the risk, cost, size and development time of Earth Science Division space-based and ground-based information systems and increase the accessibility and utility of science data.

NASA's Earth Science Data and Information System (ESDIS) project, active for over 15 years has worked to process, archive, and distribute Earth science satellite data and data from airborne and field campaigns. With attention to user satisfaction, it strives to ensure that scientists and the public have access to data to enable the study of Earth from space to advance Earth system science to meet the challenges of climate and environmental change.

The Global Earth Observation System of Systems (GEOSS) is a collaborative, international effort to share and integrate Earth observation data. NASA has joined forces with the U.S. Environmental Protection Agency (EPA), National Oceanic and Atmospheric Administration (NOAA), other agencies and nations to integrate satellite and ground- based monitoring and modeling systems to evaluate environmental conditions and predict outcomes of events such as forest fires, population growth and other developments that are natural and man- made. In the near-term, with academia, researchers will integrate a complex variety of air quality information to better understand and address the impact of air quality on the environment and human health.

A Space Act Agreement, entered into by NASA and Cray, Inc., allows for collaboration on one or more projects centered on the development and application of low-latency, "big data" systems. In particular, the project is testing the utility of hybrid computers systems using a highly integrated non-SQL database as a means for data delivery to accelerate the execution of modeling and analysis software.

NASA's Planetary Data System (PDS) is an archive of data products from NASA planetary missions, which has become a basic resource for scientists around the world. All PDS-produced products are peer-reviewed, well-documented, and easily accessible via a system of online catalogs that are organized by planetary disciplines.

The Multimission Archive at the Space Telescope Science Institute (MAST), component of NASA's distributed Space Science Data Services, supports and provide to the astronomical community a variety of astronomical data archives, with the primary focus on scientifically related data sets in the optical, ultraviolet, and near-infrared parts of the spectrum. MAST archives and supports several tools to provide access to a variety of spectral and image data.

The Earth System Grid Federation is a public archive expected to support the research underlying the International Panel on Climate Change's Fifth Assessment Report to be completed in 2014 (as it did for the Fourth Assessment Report). NASA is contributing both observational data and model output to the Federation through collaboration with the DOE.

NATIONAL INSTITUTES OF HEALTH (NIH)

National Cancer Institute (NCI)

The Cancer Imaging Archive (TCIA) is an image data-sharing service that facilitates open science in the field of medical imaging. TCIA aims to improve the use of imaging in today's cancer research and practice by increasing the efficiency and reproducibility of imaging cancer detection and diagnosis, leveraging imaging to provide an objective assessment of therapeutic response, and ultimately enabling the development of imaging resources that will lead to improved clinical decision support.

***The Cancer Genome Atlas* (TCGA)** project is a comprehensive and coordinated effort to accelerate understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. With fast development of large scale genomic technology, the TCGA project will accumulate several petabyte of raw data by 2014.

National Heart Lung and Blood Institute (NHLBI)

The Cardiovascular Research Grid (CVRG) and ***the Integrating Data for Analysis, Anonymization and Sharing (iDASH)*** are two informatics resources supported by NHLBI which provide secure data storage, integration, and analysis resources that enable collaboration while minimizing the burden on users. The CVRG provides resources for the cardiovascular research community to share data and analysis tools. iDASH leads development in privacy-preserving technology and is fostering an integrated data sharing and analysis environment.

National Institute of Biomedical Imaging and Bioengineering (NIBIB)

The Development and Launch of an Interoperable and Curated Nanomaterial Registry, led by the NIBIB institute, seeks to establish a nanomaterial registry, whose primary function is to provide consistent and curated information on the biological and environmental interactions of well-characterized nanomaterials, as well as links to associated publications, modeling tools, computational results and manufacturing guidances. The registry facilitates building standards and consistent information on manufacturing and characterizing nanomaterials, as well as their biological interactions.

The Internet Based Network for Patient-Controlled Medical Image Sharing contract addresses the feasibility of an image sharing model to test how hospitals, imaging centers and physician practices can implement cross-enterprise document sharing to transmit images and image reports.

As a ***Research Resource for Complex Physiologic Signals,*** PhysioNet offers free web access to large collections of recorded physiologic signals (*PhysioBank*) and related open-source software (*PhysioToolkit*). Each month, about 45,000 visitors worldwide use PhysioNet, retrieving about 4 terabytes of data.

The Neuroimaging Informatics Tools and Resource Clearinghouse (NITRC) is a NIH blueprint project to promote the dissemination, sharing, adoption and evolution of neuroimaging informatics tools and neuroimaging data by providing access, information and forums for interaction for the research community. Over 450 software tools and data sets are registered on NITRC; the site has had over 30.1 million hits since its launch in 2007.

The Extensible Neuroimaging Archive Toolkit (XNAT) is an open source imaging informatics platform, developed by the Neuroinformatics Research Group at Washington University, and widely used by research institutions around the world. XNAT facilitates common management, productivity and quality assurance tasks for imaging and associated data.

The Computational Anatomy and Multidimensional Modeling Resource The Los Angeles Laboratory of Neuro Imaging (LONI) houses data bases that contain imaging data from several modalities, mostly various forms of MR and PET, genetics, behavior, demographics and other data. The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a good example of a project that collects data from acquisition sites around the U.S., makes data anonymous, quarantines it until quality control is done (often immediately) and then makes it available for download to users around the world in a variety of formats.

The Computer-Assisted Functional Neurosurgery Database develops methods and techniques to assist in the placement and programming of Deep Brain Stimulators (DBSs) used for the treatment of Parkinson's disease and other movement disorders. A central database has been developed at Vanderbilt University (VU), which is collaborating with Ohio State and Wake Forest universities to acquire data from multiple sites. Since the clinical workflow and the stereotactic frames at different hospitals can vary, the surgical planning software has been updated and successfully tested.

NIH Biomedical Information Science and Technology Initiative (BISTI) Consortium for over a decade has joined the institutes and centers at NIH to promote the nation's research in Biomedical Informatics and Computational Biology (BICB), promoted a number of program announcements and funded more than a billion dollars in research. In addition, the collaboration has promoted activities within NIH such as the adoption of modern data and software sharing practices so that the fruits of research are properly disseminated to the research community.

NIH Blueprint

The Neuroscience Information Framework (NIF) is a dynamic inventory of Web-based neuroscience resources: data, materials, and tools accessible via any computer connected to the Internet. An initiative of the NIH Blueprint for Neuroscience Research, NIF advances neuroscience research by enabling discovery and access to public research data and tools worldwide through an open source, networked environment.

The NIH Human Connectome Project is an ambitious effort to map the neural pathways that underlie human brain function and to share data about the structural and functional connectivity of the human brain. The project will lead to major advances in our understanding of what makes

us uniquely human and will set the stage for future studies of abnormal brain circuits in many neurological and psychiatric disorders.

NIH Common Fund

The National Centers for Biomedical Computing (NCBC) are intended to be part of the national infrastructure in Biomedical Informatics and Computational Biology. The eight centers create innovative software programs and other tools that enable the biomedical community to integrate, analyze, model, simulate, and share data on human health and disease.

Patient Reported Outcomes Measurement Information System (PROMIS) is a system of highly reliable, valid, flexible, precise, and responsive assessment tools that measure patient–reported health status. A core resource is the Assessment Center which provides tools and a database to help researchers collect, store, and analyze data related to patient health status.

National Institute of General Medical Sciences

The Models of Infectious Disease Agent Study (MIDAS) is an effort to develop computational and analytical approaches for integrating infectious disease information rapidly and providing modeling results to policy makers at the local, state, national, and global levels. While data need to be collected and integrated globally, because public health policies are implemented locally, information must also be fine-grained, with needs for data access, management, analysis and archiving.

The structural genomics initiative advances the discovery, analysis and dissemination of three-dimensional structures of protein, RNA and other biological macromolecules representing the entire range of structural diversity found in nature to facilitate fundamental understanding and applications in biology, agriculture and medicine. Worldwide efforts include the NIH funded Protein Structure Initiative, Structural Genomics Centers for Infectious Diseases, Structural Genomics Consortium in Stockholm and the RIKEN Systems and Structural Biology Center in Japan. These efforts coordinate their sequence target selection through a central database, TargetDB, hosted at the Structural Biology Knowledgebase.

The WorldWide Protein Data Bank (wwPDB), a repository for the collection, archiving and free distribution of high quality macromolecular structural data to the scientific community on a timely basis, represents the preeminent source of experimentally determined macromolecular structure information for research and teaching in biology, biological chemistry, and medicine. The U.S. component of the project (RCSB PDB) is jointly funded by five Institutes of NIH, DOE/BER and NSF, as well as participants in the UK and Japan. The single databank now contains experimental structures and related annotation for 80,000 macromolecular structures. The Web site receives 211,000 unique visitors per month from 140 different countries. Around 1 terabyte of data are transferred each month from the website.

The Biomedical Informatics Research Network (BIRN), a national initiative to advance biomedical research through data sharing and collaboration, provides a user-driven, software-based framework for research teams to share significant quantities of data – rapidly, securely and privately – across geographic distance and/or incompatible computing systems, serving diverse research communities.

National Library of Medicine
Informatics for Integrating Biology and the Bedside (i2b2), seeks the creation of tools and approaches that facilitate integration and exchange of the informational by-products of healthcare and biomedical research. Software tools for integrating, mining and representing data that were developed by i2b2 are used at more than 50 organizations worldwide through open source sharing under open source governance.

Office of Behavioral and Social Sciences (OBSSR)
The National Archive of Computerized Data on Aging (NACDA) program advances research on aging by helping researchers to profit from the under-exploited potential of a broad range of datasets. NACD preserves and makes available the largest library of electronic data on aging in the United States.

Data Sharing for Demographic Research (DSDR) provides data archiving, preservation, dissemination and other data infrastructure services. DSDR works toward a unified legal, technical and substantive framework in which to share research data in the population sciences.

A Joint NIH - NSF Program
The Collaborative Research in Computational Neuroscience (CRCNS) is a joint NIH-NSF program to support collaborative research projects between computational scientists and neuroscientists that will advance the understanding of nervous system structure and function, mechanisms underlying nervous system disorders and computational strategies used by the nervous system. In recent years, the German Federal Ministry of Education and Research has also joined the program and supported research in Germany

NATIONAL SCIENCE FOUNDATION (NSF)

Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) is a new joint solicitation between NSF and NIH that aims to advance the core scientific and technological means of managing, analyzing, visualizing and extracting useful information from large, diverse, distributed and heterogeneous data sets. Specifically, it will support the development and evaluation of technologies and tools for data collection and management, data analytics, and/or e-science collaborations, which will enable breakthrough discoveries and innovation in science, engineering, and medicine - laying the foundations for U.S. competitiveness for many decades to come.

Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) develops, consolidates, coordinates, and leverages a set of advanced cyberinfrastructure programs and efforts across NSF to create meaningful cyberinfrastructure, as well as develop a level of integration and interoperability of data and tools to support science and education.

CIF21 Track for IGERT. NSF has shared with its community plans to establish a new CIF21 track as part of its Integrative Graduate Education and Research Traineeship (IGERT) program. This track aims to educate and support a new generation of researchers able to address

fundamental Big Data challenges concerning core techniques and technologies, problems, and cyberinfrastructure across disciplines.

Data Citation, which provides transparency and increased opportunities for the use and analysis of data sets, was encouraged in a dear colleague letter initiated by NSF's Geosciences directorate, demonstrating NSF's commitment to responsible stewardship and sustainability of data resulting from federally funded research.

Data and Software Preservation for Open Science (DASPOS) is a first attempt to establish a formal collaboration of physicists from experiments at the LHC and Fermilab/Tevatron with experts in digital curation, heterogeneous high-throughput storage systems, large-scale computing systems, and grid access and infrastructure. The intent is to define and execute a compact set of well-defined, entrant-scale activities on which to base a large-scale, long-term program, as well as an index of commonality among various scientific disciplines.

Digging into Data Challenge addresses how big data changes the research landscape for the humanities and social sciences, in which new, computationally-based research methods are needed to search, analyze, and understand massive databases of materials such as digitized books and newspapers, and transactional data from web searches, sensors and cell phone records. Administered by the National Endowment for the Humanities, this Challenge is funded by multiple U.S. and international organizations.

EarthCube supports the development of community-guided cyberinfrastructure to integrate data into a framework that will expedite the delivery of geoscience knowledge. NSF's just announced, first round of EarthCube awards, made within the CIF21 framework, via the EArly Concept Grants for Exploratory Research (EAGER) mechanism, are the first step in laying the foundation to transform the conduct of research in geosciences.

Expeditions in Computing has funded a team of researchers at the University of California (UC), Berkeley to deeply integrate algorithms, machines, and people to address big data research challenges. The combination of fundamental innovations in analytics, new systems infrastructure that facilitate scalable resources from cloud and cluster computing and crowd sourcing, and human activity and intelligence will provide solutions to problems not solvable by today's automated data analysis technologies alone.

Focused Research Group, stochastic network models. Researchers are developing a unified theoretical framework for principled statistical approaches to network models with scalable algorithms in order to differentiate knowledge in a network from randomness. Collaborators in biology and mathematics will study relationships between words and phrases in a very large newspaper database in order to provide media analysts with automatic and scalable tools.

Ideas Lab. NSF released a dear colleague letter announcing an Ideas Lab, for which cross disciplinary participation will be solicited, to generate transformative ideas for using large datasets to enhance the effectiveness of teaching and learning environments.

Information Integration and Informatics addresses the challenges and scalability problems involved in moving from traditional scientific research data to very large, heterogeneous data, such as the integration of new data types models and representations, as well as issues related to data path, information life cycle management, and new platforms.

The Computational and Data-enabled Science and Engineering (CDS&E) in Mathematical and Statistical Sciences (CDS&E-MSS), created by NSF's Division of Mathematical Sciences (DMS) and the Office of Cyberinfrastructure (OCI), is becoming a distinct discipline encompassing mathematical and statistical foundations and computational algorithms. Proposals in this program are currently being reviewed and new awards will be made in July 2012.

Some ***Research Training Groups (RTG)*** and ***Mentoring through Critical Transition Points (MCTP***) relate to big data. The RTG project at the UC Davis addresses the challenges associated with the analysis of object-data-- data that take on many forms including images, functions, graphs, and trees--in a number of fields such as astronomy, computer science, and neuroscience. Undergraduates will be trained in graphical and visualization techniques for complex data, software packages, and computer simulations to assess the validity of models. The development of student sites with big data applications to climate, image reconstruction, networks, cybersecurity and cancer are also underway.

The Laser Interferometer Gravitational Wave Observatory (LIGO) detects gravitational waves, previously unobserved form of radiation, which will open a new window on the universe. Processing the deluge of data collected by LIGO is only possible through the use of large computational facilities across the world and the collective work of more than 870 researchers in 77 institutions, as well as the Einstein@Home project.

The Open Science Grid (OSG) enables over 8,000 scientists worldwide to collaborate on discoveries, including the search for the Higgs boson. High-speed networks distribute over 15 petabytes of data each year in real-time from the Large Hadron Collider (LHC) at CERN in Switzerland to more than 100 computing facilities. Partnerships of computer and domain scientists and computing facilities in the U.S. provide the advanced fabric of services for data transfer and analysis, job specification and execution, security and administration, shared across disciplines including physics, biology, nanotechnology, and astrophysics.

The Theoretical and Computational Astrophysics Networks (TCAN) *program* seeks to maximize the discovery potential of massive astronomical data sets by advancing the fundamental theoretical and computational approaches needed to interpret those data, uniting researchers in collaborative networks that cross institutional and geographical divides and training the future theoretical and computational scientists.

NATIONAL SECURITY AGENCY (NSA)

Vigilant Net: A Competition to Foster and Test Cyber Defense Situational Awareness at Scale will explore the feasibility of conducting an online contest for developing data visualizations in the defense of massive computer networks, beginning with the identification of best practices in the design and execution of such an event.

Combining Cybersecurity and Big Data The Intelligence Community (IC) has identified a set of coordination, outreach and program activities to collaborate with a wide variety of partners throughout the U.S. government, academia and industry, as well as make its perspective accessible to the unclassified science community.

The NSA/CSS Commercial Solutions Center (NCSC) hosting vendor capabilities presentations that showcase new commercial technology developments that meet the strategic needs of NSA/CSS and the national security community.

UNITED STATES GEOLOGICAL SURVEY (USGS)

The USGS ***John Wesley Powell Center for Analysis and Synthesis*** just announced eight new research projects for transforming big data sets and big ideas about earth science theories into scientific discoveries. At the Center, scientists collaborate to perform state-of-the-art synthesis to leverage comprehensive, long-term data.