Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)
NSF 12-499

Vasant Honavar
Program Director
Information & Intelligent Systems (IIS) Division
Computer and Information Science and Engineering (CISE) Directorate

National Science Foundation

Big Data Research and Development Initiative

- Big Data Senior Steering Group – chartered in spring 2011 under the Networking and Information Technology R&D (NITRD) Program
  - Members from DARPA, DOD OSD, DHS, DOE-Science, HHS, NARA, NASA, NIST, NOAA, NSA, and USGS
  - Co-chaired by NIH and NSF
- White House Big Data Launch – March 29, 2012
- Long-term, National Big Data R&D with four major components:
  - Foundational research to develop new techniques and technologies to derive knowledge from data
  - New cyberinfrastructure to manage, curate, and serve data to research communities
  - New approaches for education and workforce development
  - Challenges and competitions to create new data analytic ideas, approaches, and tools from a more diverse stakeholder population

The Big Data Team

Suzi Iacono, NSF, Karin Remington, NIH
NITRD Big Data Steering Group

- Vasant G. Honavar, NSF - CISE
- Jia Li, NSF - MPS
- Dane Skow, NSF – OCI
- Peter H. McCartney, NSF - BIO
- Doris L. Carver, NSF - EHR
- Eduardo A. Misawa, NSF - ENG
- Eva Zanzerkia, NSF - GEO
- Peter Muhlberger, NSF - SBE
- Vladimir Papitashvili, NSF – OPP

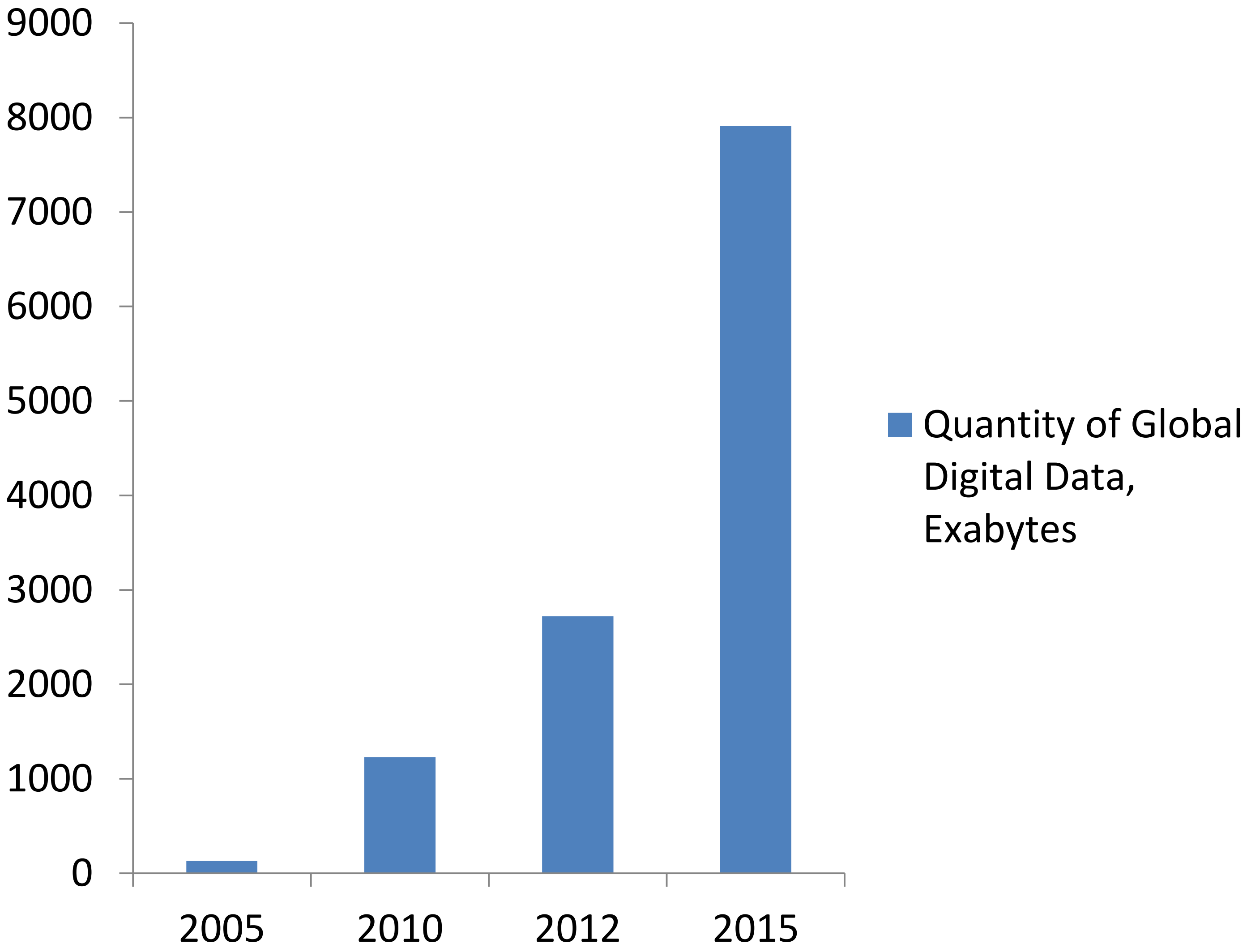
- Peter Lyster, NIH - NIGMS
- Karin A. Remington, NIH - NIGMS
- Jerry Li, NIH - NCI
- Vinay M. Pai, NIH - NIBIB
- Karen Skinner, NIH - NIDA
- Yuan Liu, NIH - NINDS
- Valerie Florance, NIH - NLM
- Vivien Bonazzi, NIH - NHGRI

Fen Zhao, AAAS, NSF - CISE

Outline

- Welcome – Suzi Iacono and Karin Remington
- Webinar – Vasant Honavar
  - Background
    - Data Deluge
    - Research Opportunities and Challenges
    - BIGDATA Program in Context
  - BIGDATA Solicitation
    - Scope
    - Research Thrusts
    - Proposal Types and Deadlines
    - Proposal preparation and submission
    - Proposal Review Process
  - Q & A – Please email your questions to bigdata@nsf.gov

Data Deluge



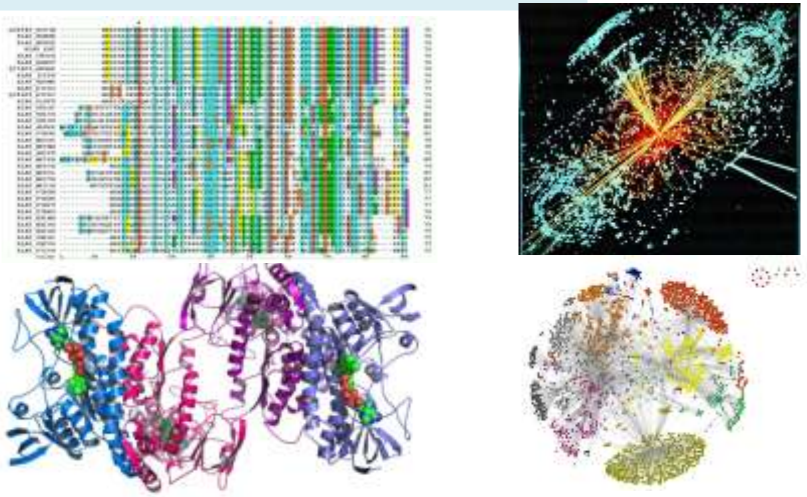
Source: EMC/IDC Digital Universe Study, 2011

Dealing with Data

- Big data is not just about volume or rate of acquisition, but also
  - Heterogeneity/diversity
    - Multiple levels of granularity
    - Multiple media and modalities
    - Scientific disciplines
  - Complexity
    - Uncertainty
    - Incompleteness
    - Representation

Data Deluge

Scientific Data

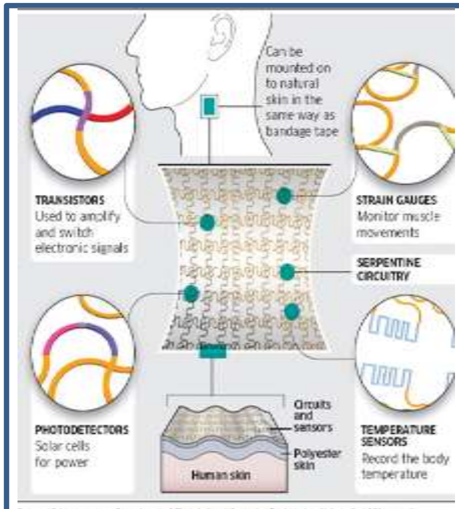


Digital Media

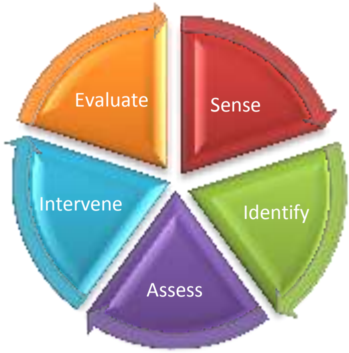
Human Sensors

Source: Sajal Das, Keith Marzullo

Health Care



Source: Sciencemag.org, Department of Electrical and Computer Engineering, University of Wisconsin



Opportunities

- Big Data presents unprecedented opportunities to
  - Accelerate scientific discovery and innovation
  - Lead to new fields of inquiry that would not otherwise be possible
  - Improve decision making
  - Understand human and social processes
  - Promote economic growth
  - Improve health and quality of life

Dealing with Data

**Picture of
The Economist
magazine pending
Image permission
approval**

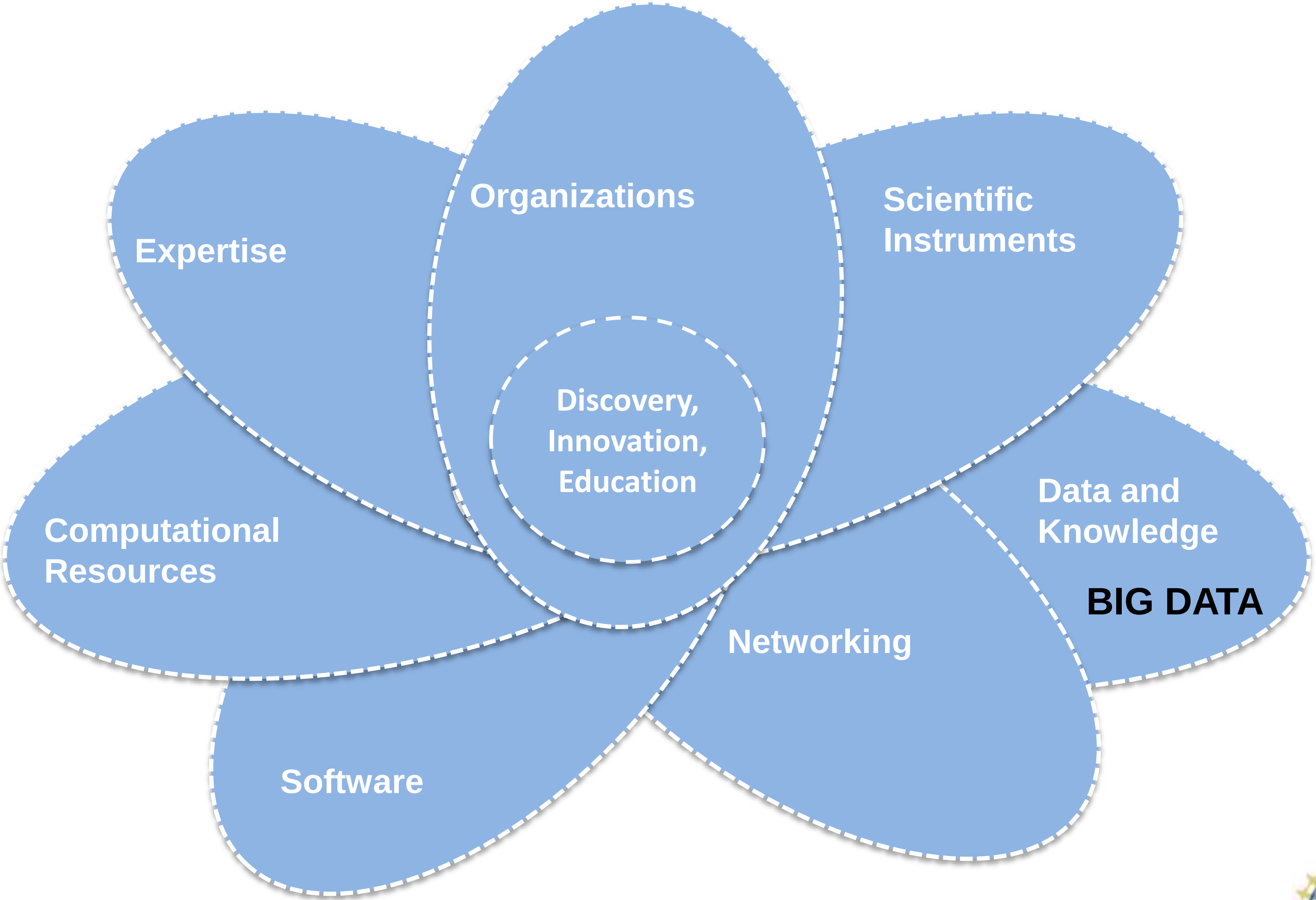
**Picture of The
Fourth Paradigm
Pending Image
Permission
approval**

http://www.sciencemag.org/site/special/data/ http://www.economist.com/node/15579717

Examples of Research Challenges

- More data is being collected than we can store
  - Analyze the data as it becomes available
  - Decide what to archive and what to discard
- Many data sets are too large to download
  - Analyze the data wherever it resides
- Many data sets are too poorly organized to be usable
  - Better organize and retrieve data
- Many data sets are heterogeneous in type, structure, semantics, organization, granularity, accessibility …
  - Integrate and customize access to federated data
- Utility of data limited by our ability to interpret and use it
  - Extract and visualize actionable knowledge
  - Evaluate results

BIG DATA Initiative in Context: NSF Cyber-infrastructure for 21st Century (CIF21) Vision



BIGDATA Solicitation in Context

- BIGDATA solicitation is one component of a national big data initiative
  - Focus: research on core techniques and technologies
- Additional BIGDATA opportunities
  - Computational and Data-enabled Science and Engineering (NSF)
    - CDS&E-MSS: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504687
    - CDS&E-ENG: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504770
  - Big data infrastructure projects (NSF)
    - DIBBs: http://www.nsf.gov/pubs/2012/nsf12557/nsf12557.htm
  - Education and workforce development efforts (NSF)
    - IGERT-CIF21: http://www.nsf.gov/pubs/2012/nsf12555/nsf12555.htm
  - Complex multi-disciplinary grand challenge problems
  - Prizes and competitions
- Other Related Opportunities
  - NSF Core Program Solicitations:
    - CISE/IIS: http://www.nsf.gov/pubs/2011/nsf11556/nsf11556.htm
  - BISTI: Biomedical Information Science and Technology Initiative (NIH) http://www.bisti.nih.gov/funding/index.asp
  - Additional solicitations and dear colleague letters (NSF): http://www.nsf.gov/cif21

BIGDATA Solicitation

- BIGDATA seeks proposals that develop and evaluate core techniques and tools within three thrust areas:
  - Data management, collection and storage (DCM)
  - Data analytics (DA)
  - E-science collaboration environments (ESCE)

Data management, collection and storage (DCM)

Potential DCM research areas include, but are not limited to:

- New data storage, I/O, architectures
- Efficient archiving, storing, indexing, retrieving, and recovery
- Streaming, filtering, compressed sensing, sufficient statistics
- Automatic data annotation
- Data discovery, workflows, provenance
- Advanced data architectures
- Data validity, integrity, consistency, uncertainty management
- Languages, tools, methodologies and programming environments

Data Analytics (DA)

Potential DA research areas include, but are not limited to:

- Scalable Machine learning, statistical inference, and data mining
- Predictive modeling, hypothesis generation and automated discovery
- New algorithms, programming languages, data structures for data analytics
- Data-driven high fidelity simulations
- Information extraction from unstructured, multimodal data
- Scalable and interactive data visualization
- Extraction and integration of knowledge from massive, complex, multi-modal, or dynamic data
- Data analytics under processing, memory, storage, access, energy, constraints

E-science collaboration environments (ESCE)

Potential DA research areas include, but are not limited to:

- Automated and Interactive Discovery Processes
- Scientific Workflows
- Novel collaboration tools
- Data, knowledge, and model sharing
- Remote operation, scheduling, and real-time remote access to instruments and data resources

NIH BIGDATA Priorities

BIGDATA core technologies and tools for

- Self-sustaining automated approaches to archiving, mining, retrieving, and analyzing diverse biomedical or behavioral research data
- Analysis of structural and functional connectomes
- Analysis of social media for understanding local, regional, national and global health
- Mapping the current state of biomedical research landscape
- Collaborative, integrative analyses of disparate data from multiple clinical research projects and clinical trials
- Predictive modeling of primary biological and pathological driving factors and processes underlying disease and treatment response
- Analysis of large volume of patient data for real-time individualized and optimal diagnosis and treatment plans
- *In silico* science to generate or test hypotheses
- Interactive publications that provide access to data and enable data reuse and reanalysis

National Priorities

- Proposals may, but are not required to, focus on core techniques and technologies needed in areas of national priority
  - Advanced Manufacturing
  - Health IT
  - Emergency response and preparedness
  - Clean energy
  - Cyberlearning
  - Material genome
  - National security

What proposals are good fits for the BIGDATA solicitation?

- The focus of this solicitation is on core scientific and technological advances (e.g., in computer and information sciences and engineering, mathematics, or statistics) needed to take advantage of available data sets to accelerate discovery
- Proposals may address research challenges within one or more of the three thrusts:
  - Data management, collection and storage (DCM)
  - Data analytics (DA)
  - E-science collaboration environments (ESCE)

What proposals are not good fits for the BIGDATA Solicitation?

- Proposals that focus primarily on
  - Implementing existing techniques or technologies
  - Applying existing techniques (e.g., machine learning, statistical analyses) to specific data sets
  - Developing databases to serve specific scientific communities using existing database technologies

Proposal Submission and Review

- All proposals shall be submitted to NSF
- All proposals are reviewed by panels according to
  - Standard NSF merit review criteria and
  - Solicitation-specific review criteria
    - Data Management and Software Sharing Plan
    - Capacity building plan
    - Evaluation plan
    - Coordination plan (required for mid-scale proposals)
- Panels will include panelists from the NSF and NIH PI communities as appropriate
- PIs may target a specific agency sponsorship only if they have
  - Communicated with a program officer from that agency; and
  - Received permission or instruction to do so

Proposals that are not targeted to a specific agency will be considered for funding by NSF or NIH

Review Criterion: Intellectual Merit

- **Intellectual Merit** (encompasses ***all*** of the following)
- How important is the proposed activity to advancing knowledge and understanding within its own field or across different fields?
- How well qualified is the proposer to conduct the project?
- To what extent does the proposed activity suggest and explore creative, original, or potentially transformative concepts?
- How well conceived and organized is the proposed activity?
- Is there sufficient access to resources?

Review Criterion: Intellectual Merit

- **Intellectual Merit** (encompasses ***all*** of the following)

- significance
- investigator(s)
- innovation
- approach
- environment

Review Criterion – Capacity Building (CB)

- Each BIGDATA proposal must include a description of at least one capacity building (broader impacts) activity
- Examples of CB activities:
  - Education and outreach
  - Broadening participation
  - Cost models
  - Data management models
  - Community data standards
  - Access policies
  - Economic sustainability models
  - Communication strategies
- See http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

Evaluation Plan

Each proposal must include a plan to evaluate the techniques and technologies developed through for example,

- Applications of the technology to specific domains
- Assessments of effectiveness

The evaluation plan should be appropriate for:

- The nature of the research activities involved
- The size and scope of the project

Data Sharing Plan

- The types of data, software, curriculum materials, and other materials to be produced in the course of the project
- Standards to be used for data and metadata format and content
- Policies for access and sharing
- Policies and provisions for re-use, re-distribution, and the production of derivatives
- Plans for archiving and for preservation of access

Software Sharing Plan

- The software should be freely available to researchers and educators in the non-profit sector
- The terms of availability should permit
  - The dissemination and commercialization of enhanced or customized versions of the software, or its incorporation into other software packages
  - Further development by other groups
  - Modification to the source code and sharing of modifications
- An applicant
  - Is responsible for creating the original and subsequent official versions of software
  - May consider proposing a plan to manage and disseminate the improvements or customizations of their tools and resources by others

Mid-scale proposals: Coordination plan (CP)

- Must include
  - The specific roles of the collaborating PIs, Co-PIs, other Senior Personnel and paid consultants at all organizations involved
  - How the project will be managed across institutions and disciplines
  - Specific coordination mechanisms for cross-institution and/or cross-discipline scientific integration e.g., workshops, graduate student exchange, project meetings, videoconferencing and other communication tools, software repositories
  - Specific references to the budget line items that support coordination
- Coordination plan may use two additional pages (beyond the 15 pages) in the Project Description
- Mid scale proposals that do not include a coordination plan will be returned without review

Proposal Types and Deadlines

NSF: http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

- Mid-scale projects:
  - Due June 13, 2012 (5 p.m. proposer's local time):
  - Typically three or more investigators
  - Budgets up to $1000,000 (total) per year for up to 5 years
- Small projects:
  - Due July 11, 2012 (5 p.m. proposer's local time)
  - Typically one or two investigators
  - Budgets up to $250,000 (total) per year for up to 3 years

Note: NIH has comparable budget limits

NIH: http://grants.nih.gov/grants/guide/notice-files/NOT-GM-12-109.html

How many awards are anticipated?

- Up to $25 million will be invested in proposals submitted in response to this solicitation, subject to availability of funds, during 2012-2013.
- An estimated fifteen to twenty projects will be funded by NSF and/or NIH during FY 2012 and FY 2013 subject to availability of funds.
- See http://grants.nih.gov/grants/guide/notice-files/NOT-GM-12-109.html for NIH-Specific guidance.

How does one apply?

- Follow the instructions provided in:
  - The solicitation http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm
  - The Proposal and Award Policies and Procedures Guide http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/
  - NIH http://grants.nih.gov/grants/guide/notice-files/NOT-GM-12-109.html
- Consult
  - FastLane FAQ and Grants.gov FAQ: www.fastlane.nsf.gov
  - Your institution's Sponsored Research Office

Questions and Answers

- We will answer selected questions sent through email
- Answers to all questions will be included in the FAQ
- Further Questions?
  - BIGDATA FAQ: http://www.nsf.gov/pubs/2012/nsf12070/nsf12070.jsp
  - Email bigdata@nsf.gov

Credits

- Copyrighted material used under Fair Use. If you are the copyright holder and believe your material has been used unfairly, or if you have any suggestions, feedback etc., please contact: ciseitsupport@nsf.gov
-
- The inclusion of a logo does not express or imply the endorsement by NSF of the entities' products, services or enterprises