Big Data R&D Initiative

Mhyron Gutmann
Directorate for the Social, Behavioral and Economic Sciences
National Science Foundation

Digital Preservation 2012
July 25, 2012

Image Credit: Exploratorium
Advances in information technologies are transforming the fabric of our society and data represents a transformative new currency for science, engineering, education and commerce.
Smart Sensing, Reasoning and Decision

**Environment Sensing**
- Percepts (sensors)
- Agent (Reasoning)
- Actions (controllers)
- Pervasive

**Emergency Response**
- Situation Awareness: Humans as sensors feed multi-modal data streams

**People-Centric Sensing**
- Personal Sensing
- Public Sensing
- Social Sensing

**Social**

**Informatics**
- Temperature
- Light, microphone

**Smart Health Care**
- ECG
- Blood pressure
- SpO₂, GSR
- Accelerometer

Credit: Photo by US Geological Survey

Source: Sajal Das, Keith Marzullo
New Paradigms for Communications

1988

Remarkable Pace of Innovation

1988

Today

MOBILE

SOCIAL NETWORKS

BLOGS

EMAIL

VOIP

VIDEO
Communications Volume & Traffic Diversity

**VoIP**

663M registered Skype users in 2011. Represents 20% of long distance minutes world-wide. If Skype were a carrier, it would be the 3rd largest in the world (behind China Mobile and Vodaphone). Largest provider of cross-border communication.

**Video**

Recent estimates as high as 60% of internet traffic is video and music sharing; 35 hours of new videos are uploaded every minute in 2011; 2 billion views per day.

**Twitter**

Currently 175 million registered users.

**Broadband**

20% of global internet users have residential broadband; 68% in US subscribe to broadband.

**Mobile**

5.3 billion mobile phone subscribers; 85% of new handsets will be able to access the mobile web; 1 in 5 has access to fast service, 3G or better; IM, MMS, SMS expected to exceed 10 trillion message by 2013.
Data Deluge

• Science gathers data at an ever-increasing rate across all scales and complexities of natural phenomena

• Sloan Digital Sky Survey in 2000, collected more data in its 1st few weeks than had been amassed in the entire history of astronomy
  – Within a decade, over 140 terabytes of information collected
  – The proposed Large Synoptic Survey Telescope (3.3 gigapixel digital camera) will generate 40 terabytes of data nightly

• By 2015, the world will generate the equivalent of approximately 93 million Libraries of Congress

• Estimated 40 exabytes of unique new information generated worldwide in 2010

• Only 5% of the information created is “structured” in a standard format of words or numbers; the rest are from cameras, smart phones, etc.
How Big is Big?

• “Big Data”: “Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”

-McKinsey Global Institute, Big data: the next frontier for innovation, competition, and productivity, May 2011.
...Not Just Volumes of Data

The science of big data is not just about volumes and velocity of data, but also

– Heterogeneity and diversity
  • Levels of granularity
  • Media formats
  • Scientific disciplines

– Complexity
  • Uncertainty
  • Incompleteness
  • Representation types
Why is Big Data Important?

- Transformative implications for commerce and economy
- Potential for addressing some of the society’s most pressing challenges
- Critical to accelerating the pace of discovery in almost every science and engineering discipline
Paradigm Shift: from Hypothesis-driven to Data-driven Discovery

http://www.sciencemag.org/site/special/data/
http://www.economist.com/node/15579717
The Age of Data: From Data to Knowledge to Action

- **Data-driven discovery** is revolutionizing scientific exploration and engineering innovations.

- **Automatic extraction of new knowledge** about the physical, biological and cyber world continues to accelerate.

- Multi-cores, concurrent and parallel algorithms, virtualization and advanced server architectures will enable **data mining and machine learning**, and **discovery and visualization of Big Data**.
Future SBE research: Technology and data drivers

- Scale: More data from more sources (environmental, sensor, administrative, survey, commercial, usage, and so on)
- Density (merge, overlap, georectify)
- Tools (statistics, GIS, network analysis, modeling, scenarios)
- Granularity (fMRI, administrative, commercial and behavioral level)
- Greater access to and demand for high performance computational resources
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 federate data
• Utility of data limited by our ability to interpret and use it
  • Extract and visualize actionable knowledge
  • Evaluate results
• Large and linked datasets may be exploited to identify individuals
  • Design management and analysis with built-in privacy preserving characteristics
A Complex Policy Setting

• Researchers want data.
• Public policy requires access to data.
• Public policy also requires protection of privacy and intellectual property and other sensitive information.
• Much more to be done: Policy on data management and data access
A National Imperative

PCAST calls on the Federal government to increase R&D investments for collecting, storing, preserving, managing, analyzing, and sharing the increasing quantities of data.

Administration’s 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, NASA, NIST, NOAA, NSA, and USGS
  - Co-chaired by NIH and NSF
Big Data Launch

• Federal Big Data R&D Initiative launched by White House OSTP on March 29, 2012 at AAAS

• Federal Announcements:
  • NSF – Subra Suresh
  • NIH – Francis Collins
  • USGS – Marcia McNutt
  • DoD – Zach Lemnios
  • DARPA Ken Gabriel
  • DOE – William Brinkman

NSF Strategy to Address Big Data

- 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
- New types of inter-disciplinary collaborations, grand challenges, and competitions

Policy
Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIG DATA)

Foundational research to extract knowledge from data

Foundational research to advance the core techniques and technologies for managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets.

Image Credit: Jurgen Schulze, Calit2, UC-San Diego

Cross-Directorate Program: NSF Wide
Multi-agency Commitment: NSF and NIH
## BIG DATA Research Thrusts

### Collection, Storage, and Management of “Big Data”
- Data representation, storage, and retrieval
- New parallel data architectures, including clouds
- Data management policies, including privacy and access
- Communication and storage devices with extreme capacities
- Sustainable economic models for access and preservation

### Data Analytics
- Computational, mathematical, statistical, and algorithmic techniques for modeling high dimensional data
- Learning, inference, prediction, and knowledge discovery for large volumes of dynamic data
- Data mining to enable automated hypothesis generation, event correlation, and anomaly detection
- Information infusion of multiple data sources

### Research in Data Sharing and Collaboration
- Tools for distant data sharing, real time visualization, and software reuse of complex data sets
- Cross disciplinary model, information and knowledge sharing
- Remote operation and real time access to distant data sources and instruments

<table>
<thead>
<tr>
<th>Credit: Fermilab Photo</th>
<th><img src="image.png" alt="Image" /></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\int \frac{x^2 + 5}{x^3 - 2x - 3} , dx$</td>
<td><img src="math.png" alt="Math Expression" /></td>
</tr>
</tbody>
</table>
Earthcube

- EAGER awards announced as part of White House Big Data
- Integrates geosciences data and high-performance computing technologies in an open, adaptable and sustainable framework to enable transformative research and education in Earth System Science
- Innovative Model: Community designed, community owned, community governed
- Interdisciplinary research:
  - Building and sustaining “new” communities
  - Workshops to bring together (GEO, SBE, CISE) communities
  - EAGER awards to seed new research
NSF Announcements

• **Dear Colleague Letters:**
  – *Data Citation to the Geosciences Community* to encourage transparency and increased opportunities for the use and analysis of data sets: [http://www.nsf.gov/pubs/2012/nsf12058/nsf12058.jsp](http://www.nsf.gov/pubs/2012/nsf12058/nsf12058.jsp)
  – *Data-Intensive Education-Related Research Funding Opportunities* 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: [http://www.nsf.gov/pubs/2012/nsf12060/nsf12060.jsp](http://www.nsf.gov/pubs/2012/nsf12060/nsf12060.jsp)

• **Expeditions-in-Computing award:**
Big Data to Address National Priorities

Health & Wellbeing
Environment & Sustainability
Emergency Response & Disaster Resiliency
Manufacturing, Robotics, & Smart Systems
Secure Cyberspace
Transportation & Energy
Education and Workforce Development
Social Networks Solving Complex Problems

Networks of human minds are taking citizen science to a new level

In 2011, players of Foldit helped to decipher the crystal structure of the Mason-Pfizer monkey virus (M-PMV) retroviral protease, an AIDS-causing monkey virus. Players produced an accurate 3D model of the enzyme in just ten days. The problem of how to configure the structure of the enzyme had stumped scientists for 15 years.
Vannevar Bush’s Vision of the Memex

*Innovations for access to and interacting with information*

1945

Today
Big Opportunities for the Future

• Our investments in research and education have returned exceptional dividends to our nation.

• Scientific discovery and technological innovation are at the core of our response to national and societal challenges – from environment, energy, transportation, sustainability and healthcare, to cyber security and national defense.

• Many of tomorrow’s breakthroughs will occur at the intersections of diverse disciplines.
Thanks!

mgutmann@nsf.gov
Smart Health & Wellbeing

*Transforming healthcare knowledge, delivery, and quality of life through IT*

**Paradigm Shift:** transforming healthcare from reactive and hospital-centered to preventive, proactive, evidence-based, person-centered and focused on wellbeing rather than disease.

<table>
<thead>
<tr>
<th>Research Thrusts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Health Information Infrastructure</td>
</tr>
<tr>
<td>Informatics and Infrastructure</td>
</tr>
</tbody>
</table>

Cross-Directorate Program: CISE, ENG, and SBE
Era of “Big Data” in Healthcare

• Large volumes of data currently collected
  EHRs and PHRs
  Multi-scale and multi-source
  During hospitalizations
  For safety and diagnosis
  On an out-patient basis
  Typically event monitors
  Via ubiquitous mobile sensors
  Behavior, physiology, environment
  As part of clinical studies
  To evaluate safety and efficacy
  From growing body of scientific knowledge
  In biomedical research literature

• Gigabits/patient/day
  High sampling rates
  Multiple signals

• Accumulating data is getting easier, but using data is hard
The ability to acquire, aggregate and mine clinical, scientific, behavioral data will create an unprecedented amount of high quality data from individuals and population

Enabling evidence-based medicine, early diagnoses, personalized assessments and care
Data to Knowledge to Decision

**Reasoning under uncertainty**

1. New patient-centric decision support tools for diagnosis and treatment through integration of biomedical knowledge and clinical data with health records

2. Discovery of causal relationships and predictive indicators for individual and population health – better understanding of behavioral, genetic and environmental causes

3. Potential impact on discovery and clinical trial for new drugs and medical devices – faster, less expensive with more predictable outcomes

4. Rapid coordinated response to infectious disease outbreaks and natural/man-made disasters
Secure and Trustworthy Cyberspace (SaTC)

Securing our Nation’s cyberspace

- **New interdisciplinary program** that aims to support fundamental scientific advances and technologies to protect cyber-systems from malicious behavior, while preserving privacy and promoting usability.

- **Scholarship for Service (SFS)** will increase the number of qualified students entering the fields of information assurance and cybersecurity. Of over 1500 funded through the program, over 1100 have been placed in Federal agencies.

Cross-Directorate Effort: CISE, ENG, EHR, MPS, OCI, and SBE
SaTC Perspectives

**Research Opportunities**

**Trustworthy Computing Systems**
- Perspective aims to provide scientific basis for designing, building and operating cyber-infrastructure with improved resilience and resistance
- Support for both theoretical and experimental approaches
- Investigation of tradeoff among trustworthy properties

**Social, Behavioral & Economic**
- Perspective includes research at individual, group, organizational, market and societal levels, identifying risks and exploring solution feasibility
- Understanding attack or defense behaviors to develop more effective strategies and solutions
- *Cyber economic incentives including metrics and models*

**Transition to Practice**
- Perspective addresses the challenge of moving from research to practice
- Focus on later stages of R&D activities including evaluation and experimental deployment
- Software required to be released under open software license
National Robotics Initiative (NRI)

*Developing the next generation of collaborative robots to enhance personal safety, health, and productivity*

A nationally concerted cross-agency program to provide U.S. leadership in science and engineering research and education aimed at the development and use of cooperative robots that work alongside people across many sectors.

**Research Thrusts**

- Fundamental research in robotics science & engineering
- Understanding the long term social, behavioral, and economic implications across all areas of human activity
- Use of robotics to facilitate and motivate STEM learning across the K-16 continuum

Cross-Directorate Program: CISE, EHR, ENG, and SBE

Multi-agency Commitment: NSF, NASA, NIH, USDA
Cyberlearning: Transforming Education

*Improving learning by integrating emerging technologies with knowledge from research about how people learn*

**Goals:**
- Understand how people learn in technology rich environments
- Design and study ways in which innovative technologies and tools can promote learning and support assessment
- Prototype new technologies and integrate them into learning environments

Cross-Directorate Program: CISE, EHR, OCI, SBE

DO-IT Center, University of Washington, Seattle
Networked Society

Computing technologies and human societies co-evolve, transforming each other in the process

- We are increasingly becoming a networked society
- Networks of human minds are taking citizen science to a new level - new methods for problem solving.
- Access to technology and information is enhancing our cognitive and physical capabilities.
- This trend will be accelerated by advances in:
  - social informatics
  - assistive technologies
  - augmented reality
  - robotics
  - crowd sourcing
  - learning technologies
  - natural language understanding
  - vision and perception
  - artificial intelligence
  - machine learning
  - information retrieval
Social Networks Solving Complex Problems

*Networks of human minds are taking citizen science to a new level*

In 2011, players of Foldit helped to decipher the crystal structure of the Mason-Pfizer monkey virus (M-PMV) retroviral protease, an AIDS-causing monkey virus. Players produced an **accurate 3D model** of the enzyme **in just ten days**. The problem of how to configure the structure of the enzyme had **stumped scientists for 15 years**.
Vannevar Bush’s Vision of the Memex

Innovations for access to and interacting with information
Augmented Human Capabilities

Converging technologies for enhancing performance and quality of life

Synergistic combination of emerging technologies from information, cognition, nanotechnology, and materials will improve the quantity and quality of our labor and thought; it will sustain and enhance our function and quality of life diminished by age or injury; and it will improve personal performance with augmented cognition and strength.
**Big Opportunities for the Future**

- Our investments in research and education have returned exceptional dividends to our nation.

- Scientific discovery and technological innovation are at the core of our response to national and societal challenges – from environment, energy, transportation, sustainability and healthcare, to cyber security and national defense.

- Many of tomorrow’s breakthroughs will occur at the intersections of diverse disciplines.