Big Data and NITRD

George O. Strawn

Director, National Coordination Office for NITRD

Caveat auditor

The opinions expressed in this talk are those of the speaker, not the U.S. government

Outline

- What is NITRD?
- USG and Open Data
- Big Data
- NITRD and Big Data

NITRD Networking and

ITRD IT R&D

CIC computing, info and comm

HPCC and communication

HPC high-performance computing

NITRD and the NCO

- NITRD: an interagency program to enhance coordination and collaboration of the IT R&D that is performed and supported by Federal agencies

- National Coordination Office: provides support for the NITRD Program, reports to OSTP, and interfaces for NITRD with OMB, GAO, Congress, etc.

NITRD Member Agencies

Department of Commerce (2)

Department of Defense (5)

Department of Energy (3)

Department of Health and Human Services (3)

Department of Homeland Security

Environmental Protection Agency

National Archives and Records Administration

National Aeronautics and Space Agency

National Reconnaissance Office

National Science Foundation

National Security Agency

NITRD PCAs
(program component areas)

- Cyber Security and Information Assurance
- High-End Computing (R&D and I&A)
- High Confidence Software and Systems
- Human Computer Interaction and Info Mgmt
- Large Scale Networking
- Social, Economic, and Workforce Implications
- Software Design and Productivity

NITRD SSGs
(senior steering groups)

- Big Data -> HCI&IM
- CPS -> HCSS
- Cybersecurity -> CSIA
- Health IT R&D ->
- Wireless Spectrum R&D -> LSN

FY 2012 Budget Estimates

	HECia	HECrd	CSIA	HCIim	LSN	HCSS	SDP	SEW	Total
NSF	250	103	98	292	122	85	78	110	1,138
DoD	211	49	145	111	112	36	30		694
NIH	222	18		215	12	10	54	22	553
DOE	317	92	34		74	4	16	6	543
DARPA		75	223	138	53				489
NIST	14	5	47	15	8	6	4	1	100
NASA	61			14	1	18	9		103
DHS			43		1		3		47
AHRQ				25	1				26
NOAA	19				2		1		22
DOEnnsa	9	5						4	18
EPA	3			3					6
NARA				1					1
Total	1,107	347	590	814	385	158	196	143	3,739

USG and Data

- *Open Access* to usg data becomes the default (http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf)

- *Public Access* to Federally funded science results (journal articles *and* science data) required of all agencies funding more than $100M per year (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)

Big Data

A term applied to data whose size (volume), rate of acquisition (velocity) or complexity (variety) is beyond the ability of commonly used software tools to capture, manage, and/or process within a tolerable elapsed time.

Why now for Big Data?

- "Moore's laws" for disks, sensors, networks and CPUs
- Disk storage cost has gone from a dollar per *byte* to less than a dollar per *ten gigabytes* today. A dollar per terabyte soon?
- Sensors: cheap remote sensing, video surveillance, environmental sensing, scientific instruments (not necessarily cheap), etc
- The Internet: billions of gigabytes and growing rapidly

Volume: big data requires big computing

- These days, supercomputers aren't actually bigger: they're broader (thousands of *tightly* coupled cpu's)

- Server farms are *loosely* coupled cpu's (thousands of servers)

- Big volume data resides on supercomputers or server farms (or at least on clusters)

Volume: big data requires new database architectures

- Relational database architecture doesn't scale
- NoSQL databases limit functionality and do scale
- Eg, BigTable, Document- and Column-oriented databases, Graph databases

Velocity: fast big data

- Success with OLTP (*parallel* online transaction processing) such as google search and amazon ordering, but sensor input (IoT) poses a bigger challenge

- Need "smart sensors" like the LHC, which generates a petabyte of data per second but "only" saves a petabyte per month (take the processing to the data if a good model exists)

Variety: diverse big data

- The *interoperability of heterogeneous data* is a major big data challenge

- The "long tail" of many small data sets requires metadata to enable interoperability

- *Semantic Medline* (the creation and use of semantic metadata with Medline) portends the a new mode of discovery from scientific text

NITRD's
Big Data Initiative

- Core Technologies
- Domain Research Data
- Challenges/Competitions
- Workforce Development

Core Technologies

- Collection, Storage and Management of Big Data
- Data Analytics
- Data Sharing and Collaboration

Domain Research Data

- Astronomy, Virtual Observatory
- data.gov
- Earth Observation Systems
- Genomics
- Materials Genome
- Nano S&T, Nanohub
- NSF projects such as DataOne, DataNet
- Particle Physics, LHC

Challenges/Competitions

- Engage a broader public

Workforce Development

- Data Science, BigData degrees

Next NITRD steps

- Creating a science of big data?
- From mining knowledge to directing action?
- Bringing together diverse communities
- Enhancing big data education and training

Statistics and Big Data

(from a National Academies study in 2013)

1. Basic statistics
2. Generalized N-body problem
3. Graph-theoretic computations
4. Linear algebraic computations
5. Optimization
6. Integration
7. Alignment problems

What the future may hold

- Data intensive science appears to be revolutionary science

- Data analytics and other big data services are major opportunities for business and government

- Big Data may also be the basis of new services for people, perhaps as significant as the Web, Google and Facebook