

How Much Information?

July Webinar

July 22, 2009

S. Madnick, M. Smith

How Much Information? Program
Contact: Pepper Lane at pelane@ucsd.edu
hmi.ucsd.edu





Agenda

Welcome & Introductions – R. Bohn & J. Short

10:00 – 10:10

HMI? Case Studies on Scientific Research
at MIT – S. Madnick & M. Smith

10:10– 10:50

Closing Discussion

10:50 – 11:00

"HMI? Case Studies on Scientific Research at MIT" Webinar

July 2009

Stuart Madnick

John Norris Maguire Professor of Information Technology, MIT Sloan School of Management & Professor of Engineering Systems, MIT School of Engineering



Mackenzie Smith

Associate Director for Technology,
MIT Libraries



Context

- **Science and engineering are key drivers for the U.S. and developed world economies**
- **About 5 million scientists and engineers are employed in the U.S.**
- **About 60 million worldwide**
- **Very little known about the quantity of their data and how it is produced, managed, shared, reused, and preserved**
 - That is the focus and purpose of this study
- **Big challenge/opportunity to leverage data better**

Goal and Method of Study

- **Goal:** To identify key trends in data generation, growth, retention, and sharing of scientific data amongst various research groups at MIT
- **Method:** Conducted a study of several data-intensive departments at MIT and compiled the key results into six case studies addressing issues of types and volumes of data as well as some future projections
- **Case studies:** focus on research in the following fields: Physics, Biological Oceanography, Neuroimaging, Chemistry/Chemical Engineering, Materials Science and Engineering, and Climate Change.

Key Topics and Questions Addressed

- **What is meant by scientific data**
- What are the ways scientific research data is created, and how has that changed over time
- **How big are “scientific databases” and how is “big” measured**
- How fast are scientific databases growing both in size and quantity and how are people addressing that growth
- **How much data replication (by multiple organizations) and redundancy (for backup and performance purposes) is going on**
- How are decisions about retention (what portions, duration, purposes, and at what cost) and disposal of scientific data made
- **What are novel data flow, data re-use, and data usage patterns of scientific research data**
- In what ways can these collections of scientific data change the way science is conducted and lead to scientific breakthroughs
- **What symbiotic relationships exist (e.g., do existing collections of scientific data accelerate the generation of new data)**

Process

- Interviews with 29 faculty members from a variety of departments at MIT
- Key results into six case studies:
 - based on 16 faculty members
- Case studies
 - **Physics**
 - **Biological Oceanography**
 - **Neuroimaging**
 - Chemistry/Chemical Engineering
 - Materials Science and Engineering
 - Climate Change

Summary of Key Findings - Data Generation

- **Total amount of data generated annually in the case studies is 41,391 terabytes (41 petabytes)**
 - Physics Department generates the most data with average of 20,600 terabytes (20.6 petabytes) per year
- **The total amount of data generated by the other scientists interviewed are as follows:**
 - Biological Oceanography – 160 GB
 - Neuroimaging – 5.4TB
 - Chemistry Instrumentation Facility – 165 GB
 - Materials Sciences and Engineering – 1.46 TB
 - Climate Change – 200 TB

Summary of Key Findings - Data Growth

- Each case study experienced a significant increase in data generation over the past 5 years
- About 5-10 times more data than 5 years ago
- Most expect to see similar growth rates in the future.
- Increase in data most often attributed to:
 - **Instruments** (generate more detailed data)
 - **Improvements in experimental methods** (e.g., computer simulations and analyses)
 - Computing capabilities
 - Cheaper data storage

Summary of Key Findings - Retention & Backup

Data Retention

- Very few had explicit data retention policies
- Most data retention decisions up to their graduate students
- Some scientists permanently store all of their data (so far)
- Others delete the majority of their data after they publish the results of a specific project
- Some are involved in very large, multi-university, international research projects, and use a tiered system for data distribution, storage and sharing.
- Some contribute their data to central repositories (more later)

Data Backup

- Data backup techniques also varied greatly
- The most widely used back-up system was MIT's TSM service
- Many used their own, less expensive backup systems
- Others did not back up their data at all

Summary of Key Findings – Sharing & Reuse

- **Data sharing and reuse varied**
- **The biological oceanographers, physicists, and climate change scientists were most open to sharing their data**
 - Could be due to the existence of national or international data repositories that make it easier for scientists in these fields to collaborate.
 - E.g., the biological oceanographers contribute to and download from NCBI's Genbank database
 - although it is better suited for geneticists and is not specifically tailored for the needs of biological oceanography
 - Many of the physicists are involved in international collaborations that have a tiered system data sharing

MIT Department of Physics

- **Over 120 faculty members**
- **4 major research divisions:**
 - Astrophysics
 - Atomic, Condensed Matter, and Plasma Physics * (*largest*)
 - Experimental Nuclear and Particle Physics
 - Theoretical Nuclear and Particle Physics
- **Affiliated with over 20 research centers, e.g.,**
 - MIT–Harvard Center for Ultracold Atoms
 - Plasma Science and Fusion Center
 - Fermi National Accelerator Laboratory (Fermilab)
 - European Organization for Nuclear Research (CERN)

Physics - Data Generation

- **Most data generated is experimental data**
 - Mostly conducted at larger MIT-affiliated laboratories
 - Run continuously for months, or even years, producing hundreds of MB of data per second
- **E.g., the heavy ion group of the Compact Muon Solenoid (CMS) detector experiment at CERN**
 - A CMS experiment runs for 9 months every year, generating data continuously at the rate of about 300-400 MB per second (approximately 8,165 TB per year)
 - Raw data is then processed 2-3 times per year, which triples the amount of data
 - The raw data is also combined with simulation data (which might be as much as about 1-2 TB of data per week)
 - As a result, one 9-month CMS experimental run will generate approximately 40,824 TB (40 PB) of data
- **The Physics department as a whole is estimated to be generating over 1,900,000 TB of data each year (= 1,900 PB or 1.9 EB)**

Physics - Data Growth

- **Expectations regarding CMS experimental data generation**
 - Remain constant for the next three years
 - Then increase steadily by a factor of two each year as scientists improve their methods for data collection and processing
- **Thus, the CMS detector experiment alone will produce about 98,000 TB (98 PB) of experimental (raw and processed) data annually by 2014 . . . and continue to increase**

Physics - Metadata

- **Metadata is a crucial component to data use and reuse**
- **In physics, standards for recording and saving metadata are either non-existent or only defined within research groups**
- **The gravitational-wave observatory experiment use an electronic logbook to record their metadata, which includes:**
 - Experimental conditions (e.g., start time, end time, data collection channels)
 - Records of external noise, e.g., when a plane flies overhead (need to note the date, time, and a description of this event)
 - Some details can be subjective and often vary based on which technician is recording the information at the time
- **The CMS detector experiments and simulations include:**
 - Configuration of the experiment
 - Beam energy used
 - Physics processes selected to simulate
 - This metadata is uploaded to the central CERN database in Europe, and linked up with the raw experimental data

Physics – Data Retention

- Due to its huge size, storage and retention of physics data is a challenge
- Usually, only a small amount of actual experimental data is kept at MIT
- A tiered approach used: different amounts of raw data stored at multiple facilities
- The entire CMS research collaboration is broken into three tiers:
- **Tier 0 sites**: where all of the raw experimental data is stored forever.
 - No processed data is stored at tier 0.
 - Raw data is copied, divided and distributed to “tier 1” sites to be processed.
- **Tier 1 sites**: must permanently store both the portion of the raw data that they receive from tier 0, as well as the processed data that they produce
 - The processed data is copied, divided, and distributed to all of the “tier 2” sites
- **Tier 2 sites**: only receive data from one tier 1 site (the MIT lab is a tier 2 site)
- **Being a tier 2 site, there is no permanent CMS data storage at MIT**
 - MIT provides space for users to analyze the portions of the CMS experimental data that they receive from their tier 1 site.
 - About 500 users from MIT and other local institutions that have space at MIT
 - Each user is provided with 1 TB of storage. Data is replicated in a RAID array but not backed up.
 - Users usually request about 100 TB of data from the tier 1 site at a time and then filter down to about 1 TB before beginning their analysis

Physics – Data Sharing and Re-Use

- **Data sharing among physicists is very common, and often essential**
- **The extent of data sharing differs depending on the research group**
 - Some scientists only share data within their research centers and rarely share raw data outside their collaboration
 - One scientist's group has an agreement with a sister project in Europe and frequently shares raw experimental data with them
- **Regardless of data sharing practices, most physicists agree that their data can be re-used and re-analyzed often**
 - One scientist explained that he re-uses the same raw data hundreds of times and often re-integrates old data into new analyses
- **Since raw data can be used for such a long period of time, the tiered data storing structure is extremely useful**
 - It allows researchers in smaller labs to have access to the raw data without having to permanently store it

Biological Oceanography

- **What is Biological Oceanography?**
 - Microbial life is integral to function of life and climate
 - Microbes are the fundamental engines that drive the cycles of energy and matter on Earth
 - **Biological Oceanography** research studies marine ecology and the relationships among aquatic organisms and the environments of the oceans or lakes
 - As a result Biological Oceanography has very diverse data needs

Further specialization: Marine Metagenomics

- **Traditional microbiology and microbial genome sequencing studies rely on cultivated cultures**
- **Marine Metagenomics draws on genetic material recovered directly from environmental samples (i.e., from the ocean)**
- **Metagenomic data is used by scientists across multiple disciplines, e.g.,**
 - biological engineering
 - Genomics
 - Environmental engineering
 - Climate: study relationship between marine microbes and phenomena like the ocean's carbon cycle

MIT Research Collaborations in Biological Oceanography

- **MIT has multiple research collaborations, such as:**
 - Time-Series Project in the Pacific Ocean
 - Microbial Observatory
 - Oxygen Minimum Zone Project
- **Specific MIT research goals include:**
 - Understand community gene content to environmental process
 - Study the biology of one single organism from the genome level to the global scale - the smallest known phototroph, and most abundant photosynthetic cell on the planet
 - Develop computational and experimental methods for studying microbial evolution

Biological Oceanography – Data Generation

- **Three types of data:**
 - Observational data (e.g. environmental conditions and oceanographic data of water sample sites)
 - Experimental data (e.g. DNA sequences)
 - Computational model data output

Example of experimental data generation for North Pacific subtropical gyre

- **Since October 1988, monthly observations of the hydrography, chemistry and biology of a water column north of Oahu, Hawaii, such as:**
 - thermocline structure
 - water column chemistry
 - currents
 - optical properties
 - primary production
 - plankton community structure
 - rates of particle export

Pyrosequencing of DNA

- **One pyrosequencing run per microbe sample**
 - Each sample contains 100 Megabase pairs (Mbp)
 - Equivalent to 500,000 DNA sequences.
- **Each week, perform 2-3 pyrosequencing runs**
 - Generating approximately 200 MB of “raw data” (actual DNA sequences) per week
 - About 30 GB of raw data per year for across all projects.
- **Use high volume DNA sequencer**
 - Such as, the ROCHE 454 pyrosequencer

Re-formatting of pyrosequencing data

- **Pyrosequencing data is re-formatted several ways, e.g.,**
 - Raw DNA sequence translated into a predicted protein sequence
 - Raw data annotated so researchers can easily identify the specific sequence when searching through their data.
 - The DNA sequence letters in the raw data (i.e. A, T, G, or C) are translated into words that have a functional meaning (i.e. ribosomal RNA sequences, peptide sequences, function RNA sequences, non-coding regulatory regions, etc).
 - Annotations are linked to both the raw DNA sequence identified and the portion of the coding region
- **The re-formatting process usually doubles the amount of data produced by the lab**
 - Thus generating approximately 60 GB of data per year

Image files created

- **The pyrosequencer also produces images files for each sequence**
 - Pyrosequencer determines the DNA sequence by adding and then removing solutions of A, C, G, and T nucleotides to the sample
 - When solution complements an unpaired base, light is produced.
 - Sequencer takes a picture of the sample each time the nucleotide solution is added
 - The result is a time-series set of images of the sample that is used to determine the entire sequence.
- **For a single run, 200 images are created, which is equivalent to approximately 30 GB**
- **Once the full DNA sequence is determined the image files are no longer needed**
 - These images are only saved for six months in case the researchers want to reprocess them into a sequence again

Changes in DNA Sequencing Technology over time

DNA Sequencing Technology	AB3730	Current ROCHE454	Solexa
Year Launched	2002	2005	2008
Data Generated/Run	72 KiloBytes	200 MegaBytes	720 MegaBytes
Cost per Megabase pair	\$694	\$120	\$7
AB3730 work equivalent	-	100x AB3730/day	300x AB3730/day

Some Impacts:

- 5 years ago one group annually stored about 10-100 Mbp of sequence data using the AB3730
 - **Now, in 4 hours, 100 Mbp of data is produced in a single run of Roche 454**
- Another scientist uses the newer Solexa DNA sequencer to produce about 100 genomes per year, each about 1 GB for raw data storage
 - **Thus, now creates 100 GB per year**

Other Types of Analyses

- There are also culturing experiments performed
- **Microarray experiments are to observe the mRNA present in different conditions**
 - 12 different strains of the organism being studied, each with 2,000 different genes (resulting in 24,000 genes)
 - By comparing the results of the microarray experiments for different strains, the scientist can determine which strains share similar genes – a process called “comparative genomics”
- **Proteomic experiments detect peptides instead of mRNA**
 - By comparing the proteomic data to the microarray data, it is possible to determine how the organism’s cell at the mRNA level differs from cell at the protein level.
- **Due to a rapid decrease in the cost and an increase in speed of DNA sequencing, it is expected that about 10-20x increase in data produced within 5 years**

Metadata

- **Oceanographic metadata include:**
 - depth (m), temp of water (degrees C), salinity, chlorophyll concentration (micrograms/kg), biomass (micromoles/kg), dissolved oxygen concentration (micromoles/kg), oxygen (micromoles/kilogram), cell counts, and pigmentation information
- **Sequencing-related metadata include:**
 - what strain was used, how the organism's strain was isolated, where it was isolated, temperature, the natural habitat of the organism, the ecotype of the organism, the name of the person who sequenced the sample

Data Retention & Backup – Three cases

- **No centralized storage system, no standard data retention or data backup policies**
- **1. Data generated is stored on RAID arrays at two different clusters** . Image files created by the pyrosequencer are kept about 6 months. All other data is stored forever (at present). The lab has a capacity of about 40-50 terabytes.
 - Compared to other major labs, this lab generates a small number of DNA sequences.
 - They import significant data from the National Center for Biotechnology Information (NCBI) GenBank database for comparative analysis – about 50 times the amount of data they generate. The NCBI data is stored locally for frequent analyses and to avoid time to find and download it and re-format it.
 - Currently, they are using about 10% of their total storage capacity (4-5 terabytes). There is no formal backup program. They also deposit much of their data to national data repositories like the NCBI and CAMERA
- **2. Lab has 1 TB of storage and is not backed up.** Most researchers keep important data on their PCs, and use their own back up methods. The lab's final sequence data (including annotations) - about 200 GB - is backed up using MIT's central backup service.
 - Team has access to 2 TB of storage on the MIT Darwin Project's cluster; but since only 200 GB of guaranteed back up, most do not use the Darwin cluster.
 - Since the lab has not reached their storage capacity, most just keep everything that they generate.
- **3. More computational, has cluster with about 1.4 TB of storage.** Usually delete computed data after about six months (i.e. after papers published), but keep the computer code needed to reproduce from the raw data. Not all of the data is backed up, but key data is moved to servers outside the group, with periodic backups.

Data Sharing and Reuse

- **By policy, every researcher or lab that publishes a paper in the field of genomics or metagenomics uploads his or her gene sequence data to the NCBI GenBank Database**
 - GenBank[®] is the National Institutes of Health (NIH) genetic sequence database, an annotated collection of all publicly available DNA sequences
- There are approximately 85,759,586,764 base pairs in the traditional GenBank division (i.e. approximately 7.8 TB of data)
- From 1982 to present, the number of bases in GenBank has doubled approximately every 18 months
- GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA Databank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI.
 - These three organizations exchange data on a daily basis

Some NCBI GenBank Limitations

- Usually only data that are linked with published papers are deposited to GenBank
- GenBank only contains flat files (i.e., no metadata is associated with the sequences)
- Scientists can submit data that has not yet been published, not most do not because it is difficult to submit data to NCBI
 - The GenBank required standards and formats, which made sense when gene sequencing began two decades ago, are now considered outdated

Other databases for the fields of biological oceanography, genomics and metagenomics

- **The Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) is a new project**
 - Aims to “to serve the needs of the microbial ecology research community by creating a rich, distinctive data repository and a bioinformatics tools resource that will address many of the unique challenges of metagenomic analysis”
 - CAMERA started March 2007 and is in the “data catch-up” phase (i.e. gathering all metagenomic data that is already available to the public)
 - CAMERA hopes to become a data deposition site for the metagenomic community
 - CAMERA currently has 50-60 metagenomic datasets posted in the data repository that have been added in an ad-hoc fashion
- **Other related national data repositories include the European Bioinformatics Institute, the Joint Genome Institute, and MICROBES online**

Benefits of Data Reuse

- **Biological oceanography data can be reused for different analyses**
- **Some run both microarray and proteomic experiments on the same sequence and/or strain, in order to perform comparative analysis**
- **Others go back to look at previously unidentified sections of DNA data and applying new tools that have been developed for a particular type of molecule**
- **One recent breakthrough: a type of rhodopsin (derived from bacteria) was found through the genomic analyses of naturally occurring marine bacterioplankton**
 - Rhodopsins are light-absorbing pigments formed when retinal (vitamin A aldehyde) binds together integral membrane proteins (opsins)
 - Rhodopsins were known to belong to two distinct protein families: visual rhodopsins and archaeal rhodopsins and these two protein families showed “no significant sequence similarity and may have different origins”
 - By studying previous parts of DNA data with new analyses, researchers found that archael-like rhodopsins are “broadly distributed among different taxa, including members of the domain *Bacteria*,” and that a “previously unsuspected mode of bacterially mediated light-driven energy generation may commonly occur in oceanic surface waters world wide”
 - Since relatives of proteorhodopsin-containing bacteria use CO₂ as a carbon source, these results “suggest the possibility of previously unknown phototrophic pathways that may influence the flux of carbon and energy in the ocean’s photic zone worldwide”

Neuroimaging at the Martinos Imaging Center

- **A collaboration of Harvard-MIT division of Health Sciences and Technology (HST), the McGovern Institute for Brain Research, Massachusetts General Hospital, and Harvard Medical School**
- **Opened in 2006**
- **Researchers conduct comparative studies of the human brain and the brains of differing animal species**
- **Three interrelated research areas: perception, cognition and action; e.g.,**
 - To understand principles of brain organization that are consistent across individuals, and those that vary across people due to age, personality, and other dimensions of individuality by examining brain-behavior relations across the life span, from children through the elderly.
 - Cognitive and neural processes that support working and long-term memory by studying healthy young adults, healthy older adults, and patients with neurological diseases (e.g. amnesia, Alzheimer's and Parkinson's diseases).

Neuroimaging - Data Generation sources

- **There are two types of Magnetic Resonance Imaging MRI techniques used to produce images of the internal structure and function of the body (with focus on the brain)**
 - **Structural magnetic resonance images (structural MRI)** document the brain anatomy
 - **Functional magnetic resonance images (fMRI)** document brain physiology
- **fMRI measures the hemodynamic response to indicate the area of the brain that is active when a subject is performing a certain task.**
 - Oxygenated and deoxygenated blood has different magnetic susceptibilities
 - The hemodynamic response in the brain to activity results in magnetic signal variation, detected by MRI scanner
 - To perform an effective fMRI scan, must also acquire structural scans

Neuroimaging - Data Generation Technologies

- **Two types of technology used**
- **3 Tesla Siemens Tim Trio 60 cm whole-body fMRI machine**
 - Tesla refers to the strength of the magnet
 - 3 Tesla is as strong as considered safe and practical for people
 - Also capable for EPI, MR angiography, diffusion, perfusion, and spectroscopy for both neuro and body applications.
 - The visual stimulus system for fMRI studies uses a Hitachi (CP-X1200 series) which projects image through a wave-guide and is displayed on a rear projection screen (Da-Lite).
- **A higher power 9.4 Tesla MRI used for animal studies**
 - Provides higher resolution images, which can then provide insights into areas to be explored in human studies.
 - Animal scans led to the discovery that the frontal cortex is involved in working memory
 - The role of specific genes in brain functions can be investigated to see the difference that genetic manipulations in animals produce

Neuroimaging - Data Generation Formats

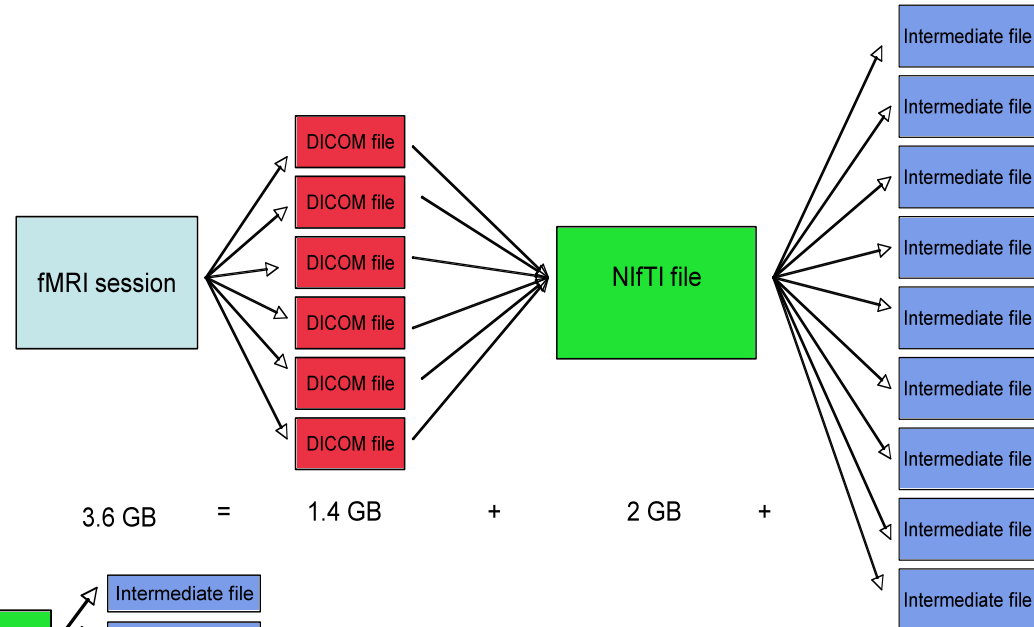
- **fMRI machines produce Digital Imaging and Communications in Medicine (DICOM) files**
 - DICOM is a standard for handling, storing, printing and transmitting medical images
 - DICOM standard has been widely adopted by hospitals and medical researchers worldwide
- **Each visit by a subject to the scanner is called a “session,” which is composed of multiple “runs”**
 - A run is a series of whole-brain volumes across a time course.
 - Anatomical runs are high-resolution scans of the anatomy of the brain
 - Functional runs are low-resolution images of the hemodynamic state of the brain over time
 - Special-purpose runs include DTI (diffusion tensor imaging), localizers (quick scans to help the scanner operator line up the landmarks in the brain with the scanner's field orientation).
- **Each session results in hundreds or thousands of DICOM images**
 - The average session will produce 1.4 GB of DICOM images

Neuroimaging – Data Conversions

- **Software convert the DICOMS to different file formats for storage**
- **Neuroimaging Informatics Technology Initiative (NifTI) is a common format, developed by neuroscientists to meet their specific needs**
 - DICOM standard has a large, clinically focused storage overhead and complex specifications for multi-frame MRI and spatial registration
 - NifTI is relatively simple format with low storage overhead, resolves some format problems in the fMRI community, and not difficult to learn and use
- **With NifTI, either (1) coalesce all the files for one session into one monolithic 4D file or (2) keep a one-to-one mapping with DICOM**
 - See diagram on next slide
- **Also, software packages transforms the NifTI files into “intermediate files”**
 - There are 8-9 “intermediate data files” for each NifTI file
 - such as slice-timing corrected NifTIs, motion corrected NifTIs, realigned NifTIs, smoothed NifTIs, and normalized NifTIs
 - Transformations lead to a lot of wasted disk space because so many types of intermediate files
 - Typically, each DICOM file maps into one NifTI file, and then each NifTI file maps into one or more intermediate files

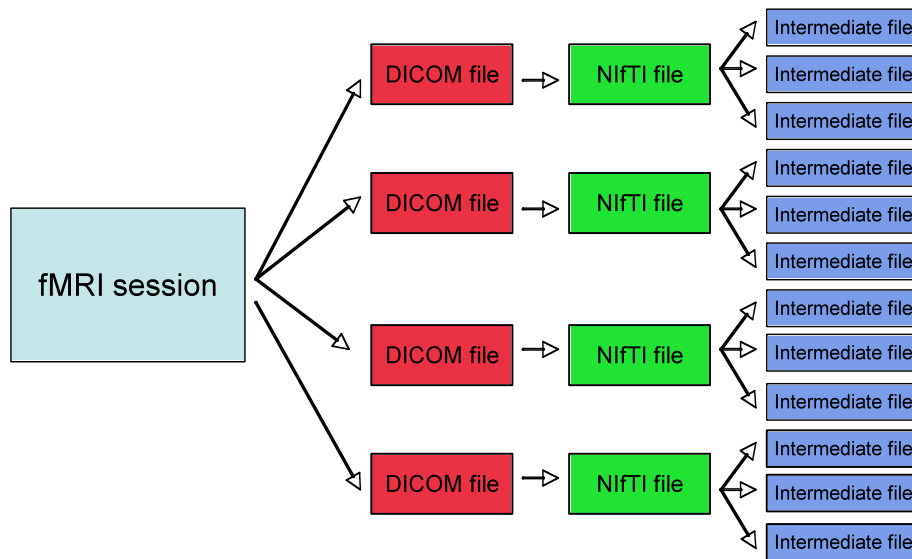
DICOM – NIfTI – Intermediate Files

**monolithic 4D
NIfTI file**



$$3.6 \text{ GB} = 1.4 \text{ GB} + 2 \text{ GB} +$$

200 MB



$$3.6 \text{ GB} = 1.4 \text{ GB} + 2 \text{ GB} + 200 \text{ MB}$$

**one to one DICOM
to NIfTI mapping**

Neuroimaging - Data Generation Quantities

- **The Martinos Imaging Center sees about 30 human subjects/week (1500/year)**
 - Each subject has one session which produces a total of 3.6 GB of data
 - Thus, a total about 5.4 TB of human image data is generated each year
 - this includes fMRI scans and the related structural MRI scans
- **Although the majority of data generated are fMRI and structural MRI images, many combine these images with additional data about the subject**
 - E.g., demographic information, health histories, behavioral data and genetic information
 - The amount of non-image data is significantly smaller than the MRI image data

Neuroimaging – Future Estimates of Data Generation Rate

- **The rate of data generation increases as the hardware and software on the scanners improve**
- **Estimated that in 5 years, fMRI scanners will have more channels for data acquisition, will increase the size of the files by a factor of 10**
- **In addition, will add a number of different technologies, such as:**
 - Electroencephalography (EEG) technology measures the electrical signals recorded at the surface at of the scalp. EEG's have lower spatial resolution than fMRI, but have higher temporal resolution and are widely used in the field of neuroimaging
 - Magnetoencephalography (MEG) is similar to the EEG but based on magnetic rather than electric signals. MEG has better spatial resolution than the EEG and also detects signals that are orthogonal to those of the EEG

Neuroimaging – Data Retention

- **There is no centralized data storage system for the Martinos Imaging Center**
- **One scientist's lab shares a RAID storage system with three other PIs at the Center**
 - Since Jan 2008 they have stored about 25 TB
 - The four generate about 2.2 TB/month (about ½ TB/scientist/month)
 - The capacity of current storage system is 44 TB, which be will reached by the end of the year
- **All the groups have similar data retention policies: they do not delete any of their image data and plan to keep buying as much storage as they need**
 - This is largely due to the high scan cost per subject (about \$750-\$1,000)
 - Additionally, the lab could not repeat experiment with the same subject because they could have memorized the visual stimuli

Neuroimaging – Data Backup

- **Many different approaches to backup, such as:**
- **The storage system shared by the 4 scientists uses MIT's central backup service for backup**
 - selected because it is affordable, relatively easy to use, and lab does not have to maintain any of the hardware
- **Another scientist uses multiple methods:**
 - Keeps all of her MRI data on a server in a local hospital which has a 2 TB capacity, backed up every day, and managed by an IT department at the hospital.
 - Makes copies of all of her DICOM files on CDs which are kept at MIT (each scan fills about two CDs)
 - Uses MIT's central backup service to back up the data at MIT
 - Also keeps hard copies of all of the patient fact sheets on campus

Neuroimaging - Data Reuse

- At present, data sharing across labs, institutions, and disciplines is limited
- But, data is commonly reused within labs
 - Multiple types of analysis on their data
 - E.g., Data is reused to perform voxel-based morphometry (VBM) to measure change in brain anatomy over time and are typically used to study dysfunction
 - VBM is done by looking at images of the same brain over time
 - Scientists take 100, 10,000, or 1,000,000 brain images and partition them according to characteristics (sex, hometown, etc) to create an “average brain.”
 - This process is repeated over time (not necessarily with the same subjects) to see how the average brain from that characteristic (e.g., geographic area) changes

Neuroimaging - Data Sharing

- **Currently, there is no widely used system for distribution and sharing of brain imaging datasets across institutions, or across disciplines**
 - This reduces the chance for future re-analysis
- **One major reason is the size of the datasets**
- **Another reason is that many scientists are protective of their data and are not open to sharing with other labs (“single lab” concept)**
- **Fundamental aspects of brain function remain unsolved due to this lack of data sharing**
 - such as the questions of how brains can perceive and navigate, how sensation and action interact, or how brain function rely on concerted neural activity across scales
- **Some research groups have started to develop platforms or networks for sharing neuroimaging data, such as:**
 - The Extensible Neuroimaging Archive Toolkit (XNAT)
 - The Biomedical Informatics Research Network (BIRN), a “geographically distributed virtual community of shared resources,” has a database for sharing neuroimaging data
 - However, it only has datasets from four subjects available
 - Furthermore, the data from each of those subjects is stored and catalogued in different ways limiting its usefulness

MIT DataSpace Project

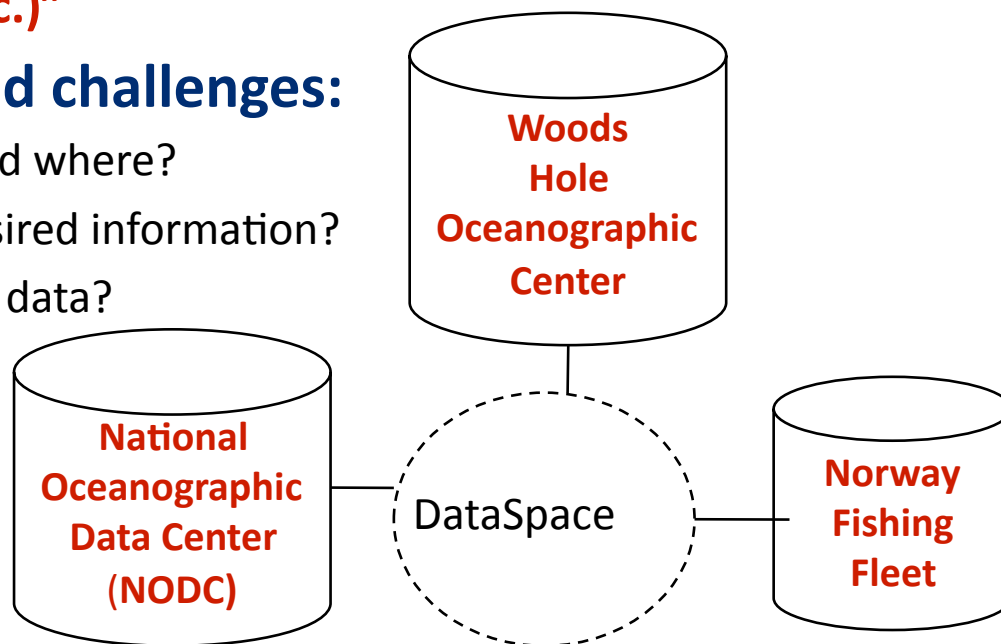
- From the *Summary and Vision* section of proposal:
...data management and long-term curation of scientific data that accommodates multiple, heterogeneous data from a variety of distributed locations ...
- **Proposed Sponsor**
 - National Science Foundation (NSF) as part of its DataNet Initiative
- **Collaborators**
 - Universities: MIT, DSpace Foundation, Georgia Tech, MIST (Abu Dhabi), Oregon State, Rice
 - Corporate: Google Labs, HP Labs, Science Commons, EMC Innovation Network
- **Schedule** (estimate)
 - May 2009: Final Proposal submitted to National Science Foundation (NSF)
 - October 2009: Grant approval go/no-go received
 - January 2010: Project begins (pending grant approval)

Key DataSpace Objectives

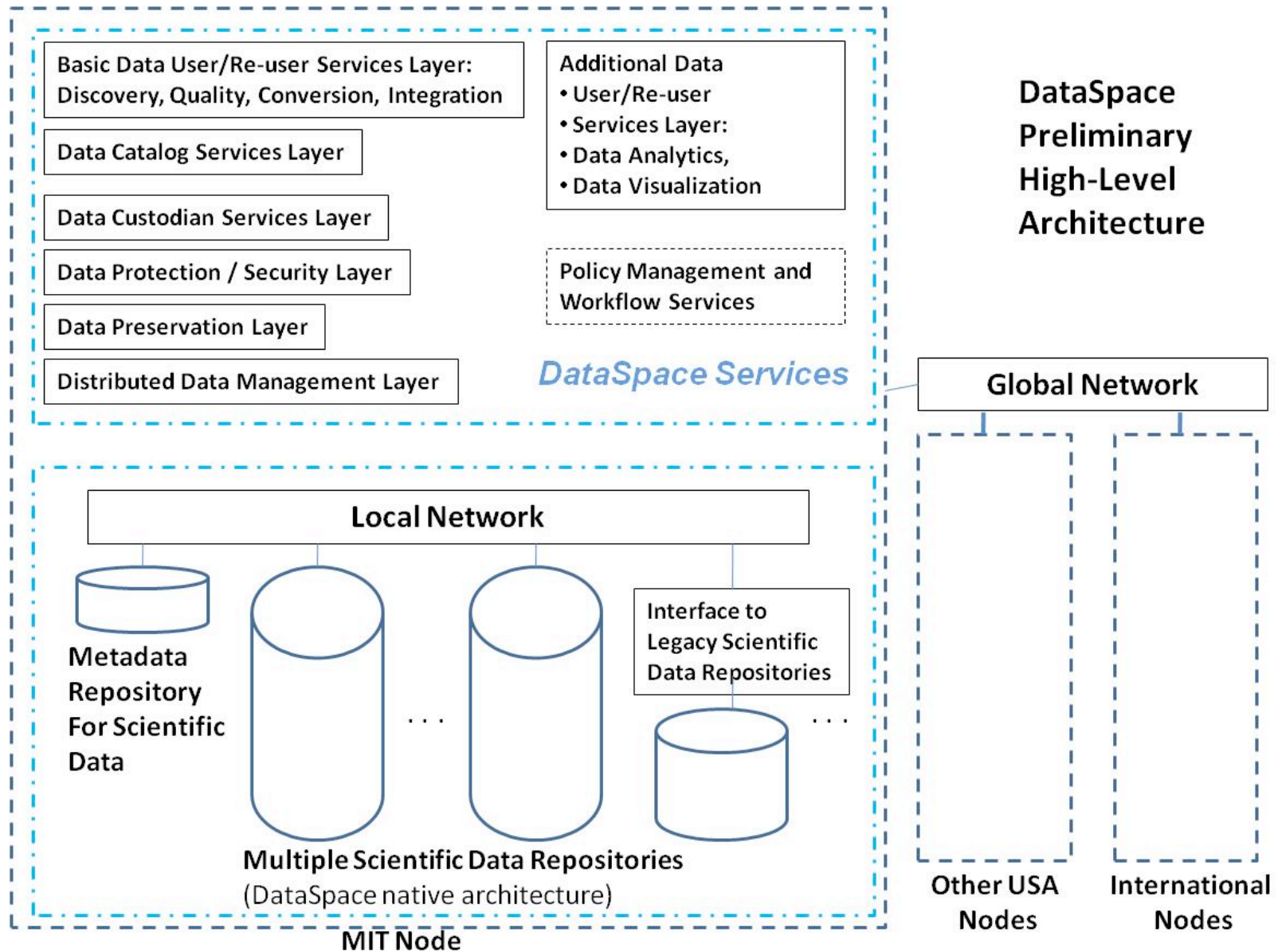
- **Data protection and security**
- **Data discovery and data semantics**
- **Data quality**
- **Data analytics**
- **Data interoperability and integration**
- **Data conversion**
- **Data analysis and visualization**
- **Data storage and preservation**
- **Distributed policy management**

Simplified Example of DataSpace Application

- All kinds and sources of scientific data accessible and “known” by DataSpace (e.g. Oceanography data)
 - National Oceanographic Data Center (NODC) , Woods Hole Oceanographic Center, Norwegian Fishing Fleet, etc.
- Example Query:
 - “Get salt and temperature measurements for ocean area around Martha’s Vineyard over past 20 years ... and provide it to me in my context (i.e., Fahrenheit, etc.)”
- Some DataSpace issues and challenges:
 - What is Martha’s Vineyard – and where?
 - Which data sources contain desired information?
 - What is “context” of the stored data?
 - What is “my context”?
 - How extract desired data and convert to desired context
 - Must also honor any security and privacy conditions



Proposed DataSpace Architecture



Further HMI? Ideas

- **Completed 6 case studies**
 - Good start and interesting trends
 - But that is only small sample
- **Interviewed only small sample of scientists in each case study area**
- **Data characteristics still emerging**
- **Important to look at evolution of instruments and their impact more carefully**
- **High Performance Computing is another important angle**
- **Still far from having a precise “number” for HMI?**
- **Possible collaborations and opportunities related to DataSpace**



Q&A

