'How Much Information?'
Case Studies on Scientific Research at MIT

Stuart Madnick
MacKenzie Smith
Kate Clopeck

Working Paper CISL# 2009-16

June 2009

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

Global Information Industry Center
at the School of International Relations and Pacific Studies
University of California, San Diego

HMI? Case Studies
How Much Information? Case Studies on Scientific Research at MIT

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 of Technology, MIT Libraries

Kate Clopeck, Masters of Science, Technology and Policy Program, MIT

June 2009

The following six case studies describe the data generation, growth, retention, and sharing trends at MIT in the fields of:

- Biological Oceanography
- Chemistry/Chemical Engineering
- Climate Change
- Materials Science and Engineering
- Neuroimaging
- Physics

Contents

Case Study Summary

Background

In order to identify key trends in data generation, growth, retention, and sharing at the Massachusetts Institute of Technology, an MIT team conducted a series of interviews with 29 faculty members from a variety of departments and compiled the key results into six case studies. These case studies focus on 16 faculty members who are currently conducting research in the six following fields: Physics, Biological Oceanography, Neuroimaging, Chemistry/Chemical Engineering, Materials Science and Engineering, and Climate Change.

Data Generation

The amount of data generated by faculty members at MIT varies significantly depending on the research field and the specific goals of the project. The total amount of data generated by the scientists featured in the six case studies is approximately 41,000 terabytes (TB) per year. One tereabyte is 10^{12} bytes, or 1000 gigabytes. Professors in the Physics Department are currently generating the most data with an estimated 20,600 TB of data per year. The Department of Chemistry Instrumentation Facility (DCIF) is currently generating the least amount of data, with an average of 165 gigabytes (GB) per year. The total amount of data generated by the other scientists interviewed are as follows: Biological Oceanography – 130.1 GB/year; Materials Sciences and Engineering – 1.46 TB/year; Neuroimaging – 5.4TB/year; Climate Change – 200 TB/year.

Data Growth

Despite differing research projects and techniques, every professor interviewed had experienced an increase in data generation over the past five years. Today, the amount of data generated by the scientists featured in the case studies was about 5-10 times more data than five years ago. While many scientists hesitated to make predictions for the future, most expect to see similar growth rates in the future. This increase in data production was most often attributed to improvements in experimental methods, instruments, computing, and cheaper data storage.

Data Retention

Despite the large amounts of data being generated, few departments or research labs have data retention policies. While a few of the scientists hired staff members to manage their data, most of the professors that were interviewed left all data retention decisions up to their graduate students. Some scientists, such as faculty members featured in the Neuroimaging and Biological Oceanography case studies, avoid making these decisions by permanently storing all of the data that they generate. Others will delete the majority of their data after they publish the results of a specific project. Two of the scientists featured in the Physics case study are involved in large, multi-university, international research projects, and use a tiered system for data distribution, storage and sharing.

Data backup techniques also varied greatly among the scientists featured in the case studies. The most widely used back-up system was the service offered by MIT. However, many professors preferred using their own, less expensive backup systems, while other did not back up their data at all.

Data Sharing and Reuse

Attitudes towards data sharing and reuse varied among labs or academic field. The biological oceanographers, physicists, and climate change scientists were the most open to sharing their data with scientists at other labs or universities. This willingness to share could be due to the existence of national or international data repositories that make it easier for scientists in these fields to collaborate. For example, 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 this field. As mentioned earlier, many of the physicists interviewed for this project are involved in international collaborations that have a tiered system data sharing.

Biological Oceanography at MIT

Background

Microbial life has been integral to the history and function of life on Earth for over 3.5 billion years. As such, microbes have evolved to be the fundamental engines that drive the cycles of energy and matter on Earth, past and present. Scientists in the field of biological oceanography conduct research in marine ecology by studying relationships among aquatic organisms and their interactions with the environments of the oceans or lakes. This case study highlights the work of three scientists at MIT working in the field of biological oceanography.

One scientist is a Professor in MIT's Departments of Civil and Environment Engineering and Biological Engineering. There are about 18 researchers working at this lab including graduate students, post doctoral researchers, research scientists, visiting scientists, and a computational biologist. This professor is one of the leading scientists in the new, but rapidly growing segment of biological oceanography: marine metagenomics. Unlike traditional microbiology and microbial genome sequencing studies that rely on cultivated cultures, marine metagenomics draws on genetic material recovered directly from environmental samples. Metagenomic data has enabled scientists across disciplines, (e.g., biological engineering, genomics, environmental engineering, etc.) to begin to explore and model the relationship between marine microbes and things like climate change and the ocean's carbon cycle.

The overall goal of this professor's laboratory is "to better describe and exploit the genetic, biochemical, and metabolic potential that is contained in the natural microbial world." Their central focus is on marine systems, due to the fundamental environmental significance of the oceans. These systems are also well suited for enabling development of new technologies, methods, and theory for assessing the gene and genomic content of natural microbial communities without cultivation, quantitatively comparing gene content of different microbial communities based on environmental variables, and developing predictive models that relate community gene content to environmental process. Currently, the lab is engaged in applying contemporary genomic technologies to dissect complex microbial assemblages. While biotic processes that occur within natural microbial communities are diverse and complex, much of this complexity is encoded in the nature, identity, structure, and dynamics of interacting genomes in situ. This genomic information can now be rapidly and generically extracted from the genomes of co-occurring microbes in natural habitats, using standard genomic technologies. This professor is currently involved in three main projects: A Time-Series Project in the Pacific Ocean, A Microbial Observatory and an Oxygen Minimum Zone Project.

The second scientist is also a Professor in MIT's Departments of Civil and Environmental Engineering and Biology. Unlike the first professor who studies a wide range of microbial life, this scientist's research is focused on understanding the biology of one single organism, Prochlorococcus, from the genome level to the global scale. This organism is the dominant primary producer in the oceans, the smallest known phototroph, and the most abundant photosynthetic cell on the planet. Over the past ten years, her laboratory set as their goal to develop Prochlorococcus as a model system for cross-scale systems biology. Her lab consists of 22 graduate students, post-doctoral associates, research scientists, research assistants and MIT undergraduates spanning the fields of biochemistry, genomics, virology, microbial ecology, and oceanography, all united around Prochlorococcus. The goal of this professor's laboratory is to use their studies of Prochlorococcus to gain a better appreciation for the full complexity of Life's properties and processes; not only those that are encoded in an organism's DNA, but also those emerging at higher levels of biospheric organization.

The final scientist is an associate Professor in MIT's Department of Civil and Environmental Engineering. While he is also studying biological oceanography and marine metagenomics, his research includes more computational work than the other two scientists in this case study and aims to develop complementary computational and experimental methods for studying microbial evolution. His laboratory is also much smaller than the other two labs, consisting of only five graduate students and one undergraduate student.

Data Generation

Most of the data generated by scientists in the field of biological oceanography is a combination of observational data (environmental conditions and oceanographic data that describe the water sample sites) and experimental data (DNA sequences). However, as mentioned earlier, there are some scientists who are also using computational models to generate data.

Despite similar types of data, scientists generate this data in a number of different ways. For example, one professor receives samples from ocean cruises that he sequences in his own lab using a high volume DNA sequencer (the ROCHE 454 pyrosequencer). The time-series project provides a good example of the typical data production protocol at his lab. The objective of this research program "is to provide a comprehensive description of the ocean at a site representative of the North Pacific subtropical gyre." Since October 1988, scientists working on this project have been making repeated observations of the hydrography, chemistry and biology of the water column at a station north of Oahu, Hawaii. Cruises are made approximately once per month to the deep-water station. At the station, scientists on the ship will take a water sample from eight different depths, called a profile. These scientists will also take measurements of the thermocline structure, water column chemistry, currents, optical properties, primary production, plankton community structure, and rates of particle export are made on each cruise. They then filter the microbes out of the sample, freeze them, and ship them to the professor at MIT. He will then extract nucleic acids from the microbe sample and perform "pyrosequencing" to determine the DNA sequence. Pyrosequencing is a method of DNA sequencing based on the "sequencing by synthesis" principle that was developed by Mostafa Ronaghi and Pål Nyrén at the Royal Institute of Technology in Stockholm in the 1990s. This method is "based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemiluminescent enzyme. The method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step. The template DNA is immobilized, and solutions of A, C, G, and T nucleotides are added and removed after the reaction, sequentially. Light is produced only when the nucleotide solution complements the first unpaired base of the template. The sequence of solutions which produce chemiluminescent signals allows the determination of the sequence of the template."

This professor's lab performs one pyrosequencing run per microbe sample. Each sample contains 100 Megabase pairs (MBp), which is equivalent to 500,000 DNA sequences. Each week, they will perform about 2-3 pyrosequencing runs. These runs generate approximately 200 Megabytes of "raw data" (actual DNA sequences) per week or about 30 Gigabytes of raw data per year for all three projects.

Once they determine the DNA sequence, his lab then re-formats the data so they can apply different analytic procedures. For example, they may translate the raw DNA sequence into a predicted protein sequence. The number of times that the data can be re-formatted varies depending on the analyses that will be done on the sample. However, the raw data is always at least annotated, so that the researchers can easily identify that specific sequence when searching through their data. During the annotation process, 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 sequence, non-coding regulatory regions, etc). These annotations are linked to both the raw DNA sequence identified and the portion of the coding region. The re-formatting process essentially doubles the amount of data produced by the lab. Therefore, his lab generates approximately 60 Gigabytes of data per year.

In addition to the raw DNA sequences and the re-formatted data, the pyrosequencer also produces images files for each sequence. As described above, the pyrosequencer determines the DNA sequence by adding and then removing solutions of A, C, G, and T nucleotides to the sample. When this solution complements an unpaired base, light will be produced. In order to determine which bases "light-up" the 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, which is used to determine the entire sequence. For a single run you need to create 200 images, which is equivalent to approximately 30 gigabytes. However, once the full DNA sequence is determined, the image files are no longer needed. These images are saved for six months, just in case the researchers want to reprocess them into a sequence again, however, this does not happen very often and the files are eventually deleted.

DNA Sequencing Technology	AB3730	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

Table 1: Cost and Data Production Comparison of DNA Sequencing Technologies

Table 1 shows how the data production rates have changed with the new DNA sequencing technologies. This professor's lab purchased their ROCHE 454 pyrosequencer about one year ago. Prior to the 454, they used an AB3730 capillarity sequencer, which was the standard technology at the time. The Solexa is one of the newer technologies on the market now. As seen below, running the ROCHE454 costs this lab about 15% of the cost of running the AB3730, for the same amount of information.

Generally speaking, five years ago this professor's research group was storing a total 10s -100s MBp of sequence data (that translates to approximately 100s megabytes of data). Now, in four hours, their DNA sequencer produces 100 MBp of data in a single 4 hour run. Another professor at MIT, who uses the newest Solexa DNA sequencer, produces about 100 genomes per year. Each genome uses about 1 gigabyte for raw data storage, resulting in about 100 gigabytes of raw genomes produced per year. Once this data is processed and assembled into sequences, the final amount of data is even larger. Although they plan to keep this sequencing technology for years, this professor predicts that this rate of data production will still increase in the near future due to new types of experiments (metabolomics).

While the other biological and civil engineering professor is also working with sequence data, she does not sequence samples from the environment. Instead, she receives the raw sequencing data, also called "raw reads" and works to assemble them into a genome. This professor receives the raw reads from larger laboratories. They are sent to her lab as text files, and are therefore not large files. The final assembled genome will be approximately 1.5 Megabase pairs (or approximately 1.5 megabytes of data).

In addition to assembling genome sequences from environmental Prochlorococcus samples, researchers in her lab conduct both physiological experiments (e.g. growth rate of cells as a function of some environmental variable, virus infection experiments, and experimental evolution) and experiments that involve culturing the Prochlorococcus cells in a controlled environment. There are two main types of culturing experiments that they perform: micro array experiments and proteomic experiments. The goal of the micro array experiment is to observe the mRNA present in different conditions. Researchers at this professor's lab will compare the results of these experiments to the sequence data and determine another round of gene annotation. There are 12 different strains of Prochlorococcus, each with 2,000 different genes (resulting in 24,000 genes). By comparing the results of the micro array experiments for different strains, this professor can learn which strains share similar genes. This process is called "comparative genomics."

The proteomic experiments are similar to the micro array experiments except they are used to detect peptides instead of mRNA. By comparing the proteomic data to the micro array data, this professor can learn how the Prochlorococcus cell at the mRNA level differs from the cell at the protein level. The proteomic experiments are relatively new for her lab and therefore, her researchers and students are still learning how to interpret their results.

Due to new experimental techniques (such as the microarray experiments), faster and cheaper sequencing technologies, and increased funding, this professor is now producing about five times as much data as she was 5 years ago. In the future, she predicts a both rapid decrease in the cost and an increase in speed of DNA sequencing technologies which could cause her lab to produce about 10-20x more data by 2014.

Metadata

The metadata associated with biological oceanography research differs based on the type of raw data that it is associated with. For example, two of the professor's metadata is the oceanographic data from where their samples were taken. This data is collected and recorded by the scientists who take the ocean samples. Examples of this 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. Again, the time-series program provides a good example of how one professor accesses this metadata. Once a frozen microbe sample arrives at his lab, he can log onto the project's website, enter the date that the sample was taken, and download all of the metadata associated with the sample.

Another professor deals with two types of metadata. For the raw reads that she receives, she needs metadata that describes how the Prochlorococcus strain was isolated, where it was isolated, what the optimal temperature is, the natural habitat of the organism, the ecotype of the organism, and the name of the person who sequenced the sample. This information can be hard to track down, but it usually exists in an excel spreadsheet created by the scientist who originally sequenced the sample. For the experiments that she conducts in the lab, this professor's metadata includes the experimental conditions, what strain was used, and the temperature. This data will be linked to the raw data from the micro array or proteomic experiments (however, if they perform a micro array and a proteomic experiment on the sample, the results from these experiments will not be linked to each other or to the raw genome data). The bioinformatics specialist in this lab is in the process of developing a centralized database to control all of the lab's metadata.

Data Retention

There is no centralized storage system for biological oceanographers at MIT. Additionally, there are no standard data retention or data back up policies. The data generated at one professor's lab is stored on RAID arrays at two different computational clusters on the MIT campus. The smaller of the two clusters is located in the same building of his lab, and the other is located at a building across campus where this professor rents space from MIT's Information Services and Technology (IS&T) Department. He currently rents three racks of space for this cluster. This professor built both of these clusters when he first came to MIT in 2004 and has been adding to them (by buying more RAID arrays) ever since.

Both the raw data and the formatted data are stored on each computer. As mentioned earlier, the image files created by the pyrosequencer are kept for about six months, but are then deleted to make room for new sequence data. All other data is stored forever. Between the two clusters, this lab has a capacity of about 40-50 terabytes. Compared to other major labs in the country and around the world that have multiple sequencers running samples throughout the year, this laboratory generates a small number of DNA sequences. As mentioned earlier, the lab is only generating about 60 gigabytes of data per year for all three projects. However, they also import all of the datasets posted by other labs on the National Center for Biotechnology Information (NCBI) GenBank database to use for comparative analysis (either comparing their data and the places it comes from to similar datasets or looking at the same gene sequence in different environments). They import approximately 50 times the amount of data they generate. The researchers in this professor's laboratory keep the NCBI data stored on the lab's computers because they frequently need it for their analyses and it takes too long to look up, download it from GenBank, and re-format it every time they need it. Currently, they are using about 10% of the total capacity (4-5 terabytes). The GenBank database is described in more detail in the data sharing section below. This lab does not use a formal back-up program for their data since they store all of their data at both computational clusters. They also deposit as much of their data as they can to national data repositories like the NCBI and CAMERA databases (see next section).

Another professor deletes his intermediate data files after about six months (after the scientific papers they were generated for are published) but keeps the computer code needed to reproduce these files from the raw data. Although not all of the data produced in his lab is backed up, the key data is moved to servers (outside the group) with periodic backups.

The other professor's laboratory has 1 terabyte of storage at building 48 that is managed by their bioinformatics specialist and it is not backed up. Most of the researchers in her lab keep their important data in their own personal computers, and use their own back up methods. The laboratory's final sequence data (including annotations) is kept on a database that was developed by this specialist and is backed up through MIT's back up service. This entire database is only about 200 gigabytes and includes all of the published or publishable data). In addition to the storage at their laboratory, this professor's research team also has shared space on the MIT DARWIN cluster. Each DARWIN user is given 2 terabytes on the cluster, however they are only given 200 gigabytes of guaranteed back up. Therefore, most of the researchers at her lab do not use this cluster for storage. The bioinformatics specialist is in charge of all data retention decisions. The laboratory has not yet reached their data storage capacity, so most researchers just keep everything that they generate.

Data Sharing and Reuse

Every researcher or lab that publishes a paper in the field of genomics or metagenomics must upload his or her 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 (~7.8 terabytes of data). From 1982 to the 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.

All three of the scientists mentioned in this case study contribute to and draw from the NCBI GenBank database. NCBI places no restrictions on the use or distribution of the GenBank data. While this database provides useful information to the field of genomics, only data that are linked with published papers are deposited to GenBank. Additionally, this data has marginal utility because GenBank only contains flat files (i.e. no metadata is associated with the sequences). Although scientists can submit data that has not been published yet, not many scientists are doing this because it very arduous to submit to NCBI. This is due to the GenBank standards and formats, which made sense when gene sequencing began two decades ago, but are no longer appropriate.

In addition to the GenBank database, there have been a number of other data repositories for biological oceanography data. Two of these professors either use or plan to use the Community Cyber Infrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA). CAMERA is a new project that aims "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. One professor is a member of the scientific advisory board. CAMERA went public on March 17, 2007 and is currently in the "data catch-up" phase (i.e. gathering all of the metagenomic data that is already available to the public). Once this phase is complete, 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 national data repositories for the fields of biological oceanography, genomics and metagenomics include the European Bioinformatics Institute (http://www.ebi.ac.uk/), the Joint Genome Institute (http://www.jgi.doe.gov/) and MICROBES online (http://www.microbesonline.org/).

Biological oceanography data can constantly be re-used for different analyses. As mentioned earlier, researchers in one professor's lab will often run both micro array and proteomic experiments the same sequence and/or strain and perform comparative analysis. Another professor also re-uses his data quite frequently. He will often go back to look at previously unknown chunks of DNA data and apply new tools that have been developed for a particular type of molecule. In fact, one recent breakthrough in the field of marine metagenomics was discovered this way when a type of rhodopsin derived from bacteria was discovered through the genomic analyses of naturally occurring marine bacterioplankton. Rhodopsins are light-absorbing pigments that are formed when retinal (vitamin A aldehyde) binds together integral membrane proteins (opsins). Rhodoposins are currently known to belong to two distinct protein families: visual rhodopsins and archaeal rhodopsins. These two protein families show "no significant sequence similarity and may have different origins". In the year 2000 (when the results of this study were first published), no rhodoposin-like sequences had been reported in members of the domain Bacteria. By going back and studying previous unknown parts of DNA data with new analyses, researchers were able to demonstrate 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 some relatives of the proteorhodopsin-containing bacteria use CO_2 as a carbon source, these results "suggest the possibility of a previously unknown phototrophic pathways that may influence the flux of carbon and energy in the ocean's photic zone worldwide".

While data re-use can often result in new discoveries like the rhodopsin derivation, it can also lead to some problems because once a gene expression is published, it is hard to record how it is evolving. This is a frequent problem for one of the professor's research group.

Key Trends and Indicators for Data Growth

Based on the work of the three biological oceanographers highlighted in the case study, we have identified the following trends and indicators for data growth:

1. Data generation will rapidly increase with new improvements in gene sequencing technologies. Five years ago, the state of the art DNA sequencers were only generating about 72 kilobytes of data per run, now they are producing about 720 Megabytes per run. If this trend continues, then sequencers in 2014 should be producing about 7 gigabytes per run.

2. New experimental techniques have also led to an increase in metagenomic data production.

3. Metadata is extremely important to the field of biological oceanography and especially in marine metagenomics. Types of metadata range from oceanographic data describing sample sites (for raw sequence data), to experimental and lab conditions (for cultured samples).

4. Every researcher or lab that publishes a paper in the field of genomics or metagenomics must upload his or her 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 (~ 7.8 terabytes of data). From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.

Chemistry and Chemical Engineering at MIT

Background

The Department of Chemistry at the Massachusetts Institute of Technology has over 30 faculty members who teach and conduct research on a variety of subjects including biological chemistry, inorganic chemistry, organic chemistry, physical chemistry, environmental chemistry, materials chemistry and nanoscience. One scientist interviewed, for example, is currently studying quantum chemistry in an effort to develop new methods to make reliable predictions about chemical phenomena. Currently, his lab is focused on physical chemistry topics such as electron transfer, electron dynamics, electron spins, and molecular magnetism.

Many faculty members of this department conduct experiments at the Department of Chemistry Instrumentation Facility (DCIF). This NSF-funded facility's function is to maintain state-of-the-art major analytical instruments in order to support the ongoing research programs within the MIT Chemistry Department. Currently, four permanent staff members provide instrument training, maintenance, repair, and applications assistance to well over four hundred users. The lab houses seven Nuclear Magnetic Resonance (NMR) spectrometers, one Electronic Paramagnetic Resonance (EPR) spectrometer, one high-resolution Fourier Transform mass spectrometer, a Gas Chromatograph mass spectrometer, a polarimeter, a Bruker Omniflex MALDI-TOF, and a Fourier Transform Infrared (FT-IR) spectrometer.

In addition to the Chemistry Department, MIT has a separate department of Chemical Engineering. Chemical Engineers at MIT conduct research in areas of chemistry, biology, and physics, and have made significant contributions to the fields of medicine, biotechnology, microelectronics, advanced materials, energy, consumer products, manufacturing, and environmental solutions. For example, one scientist in this department conducts research in areas of metabolic engineering, biochemical engineering, bioprocess engineering, and synthetic biology to harness the synthetic power of biology to build "microbial chemical factories." Her current efforts are focused on the development of tools and methodologies for novel biosynthetic pathway design and the investigation of gene dosage effects on the physiology and productivity of engineered microbes.

For this case study, three scientists were interviewed, including two from the Department of Chemistry and one from Chemical Engineering.

Data Generation

As mentioned earlier, scientists in the Departments of Chemistry and Chemical Engineering conduct research in various subject areas from quantum chemistry to biotechnology. The amount of data generated by each scientist varies based on the goals of the specific research project and the instruments used to generate data. One way to estimate the total amount of data generated by scientists in these two departments is by the data produced at the Department of Chemistry's Instrumentation Facility (DCIF).

The DCIF is open to faculty members and research groups at MIT, as well as members of other academic institutions and of industry in the area. Over the course of a year, about 60 research groups (over 400 users) actively use the DCIF. Industrial customers use about 12% of the total "instrument use time", while the remaining 88% is used by groups from academic institutions (MIT research groups account for 84% of "instrument use time" by academic institutions).

The majority of the data generated at the DCIF is produced by the Nuclear Magnetic Resonance (NMR) spectrometers. NMR is a phenomenon that occurs when the nuclei of certain atoms are immersed in a static magnetic field and exposed to a second oscillating magnetic field. Not all nuclei experience this phenomenon – it depends on whether the protons in the nucleus possess a property called spin. The spin of a proton is like a magnetic moment vector, which causes the proton to behave like a magnet with a north and south pole. When the proton is placed in an external magnetic field, the spin vector of the particle aligns itself with the external field, just like a magnet would[1].

Spectroscopy is the study of the interaction of electromagnetic radiation with matter. Nuclear magnetic resonance spectroscopy is the use of the NMR phenomenon to study physical, chemical, and biological properties of matter. NMR spectroscopy is routinely used by chemists to study chemical structures using simple one-dimensional techniques. Other NMR techniques include: two-dimensional techniques to determine the structure of more complicated molecules, time domain techniques to probe molecular dynamics in solutions, and solid state NMR spectroscopy to determine the molecular structure of solids.

Figure 1 is a schematic representation of the major systems of a NMR spectrometer. At the top of this diagram is the NMR spectrometer's super conduction magnet. This magnet is one of the most expensive components of the nuclear magnetic resonance spectrometer system and produces the static magnetic field necessary for all NMR experiments. The shim coils (which are located immediately within the bore of the magnet) are for homogenizing the magnetic field produced by the super conduction magnet. Within the shim coil is the probe, which contains RF coils. The sample is positioned within the RF coil of the probe. These RF coils serve two purposes: 1.) To produce the second, oscillating magnetic field, which is necessary to rotate the spins of the sample during NMR experiments; and 2.) To detect the signal from the spins within the sample.

As shown in Figure 1, the instrument's computer controls all components of the spectrometer. The operator of the spectrometer gives input (i.e. RF frequency, the width and shape of the RF electromagnetic pulses for the oscillating magnetic field) to the computer through a console terminal with a mouse and keyboard. Some spectrometers also have a separate small interface for carrying out some of the more routine procedures on the spectrometer. A pulse sequence is selected and customized from the console terminal. The operator can see spectra on a video display located on the console and can make hard copies of spectra using a printer.

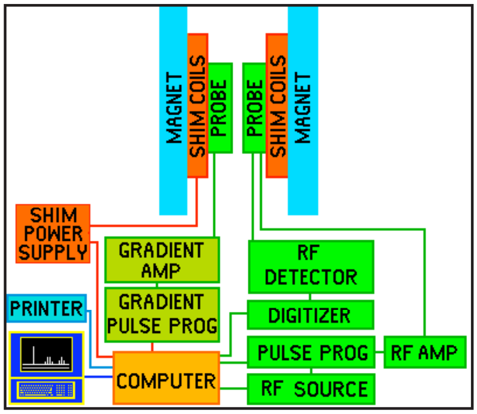


Figure 1: Schematic representation of the major systems of a nuclear magnetic resonance spectrometer[1]

NMR spectrometers produce spectral data. Figure 2 illustrates an example of a low resolution NMR spectrum. The number of peaks in the spectrum is equal to the number of different environments the hydrogen atoms are in. The ratio of the areas under the peaks is the ratio of the number of hydrogen atoms in each of these environments. The amount of splitting indicates the number of hydrogens attached to the carbon atom or atoms "next-door." The number of sub-peaks in a cluster is one more than the number of hydrogens attached to the "next-door" carbon(s). Figure 2 shows the NMR spectrum for C4H8O2. The three peaks indicate that there are three different environments for the hydrogens. The hydrogens in those three environments are in the ratio 2:3:3. Since there are 8 hydrogens altogether, this ratio represents a CH2 group and two CH3 groups. The CH2 group at about 4.1 ppm is a quartet, which means that it is "next-door" to a carbon with three

[1] Hornak, Joseph P. The Basics of NMR. The Rochester Institute of Technology (2002).

hydrogens attached - a CH3 group. The CH3 group at about 1.3 ppm is a triplet. Therefore, this group must be "next-door" to a CH2 group. The CH3 group at about 2.0 ppm is a singlet. That means that the carbon "next-door" is not attached to any hydrogen atoms[2].

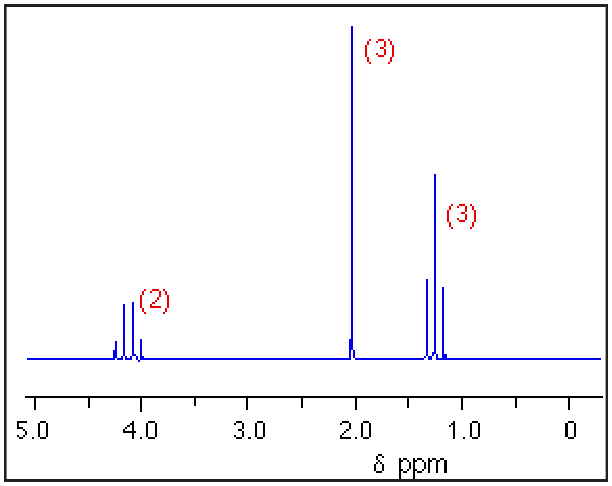
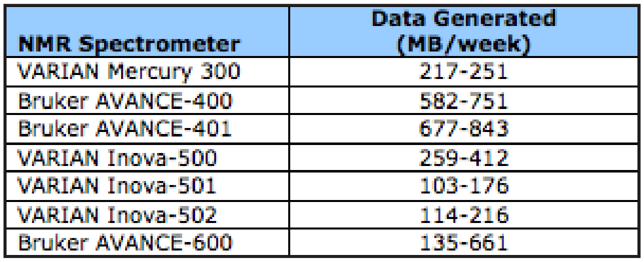


Figure 2: Low resolution NMR spectrum for C4H8O2.

In total the 7 NMR spectrometers instruments at the DCIF generate a maximum of about 3.3 gigabytes of data per week or approximately 165 GB of raw data per year (see table 1). The biggest data generator is the Bruker AVANCE 400 MHz NMR spectrometer with Spectro Spin superconducting magnet. This instrument generates approximately 677-843 MB per week. Although the other instruments at the facility do generate data, it is insignificant when compared to the amount produced by the NMR spectrometers (about 1/10th of the data).

Table 1: Data generated by the NMR spectrometers at MIT's DCIF



NMR Spectrometer	Data Generated (MB/week)
VARIAN Mercury 300	217-251
Bruker AVANCE-400	582-751
Bruker AVANCE-401	677-843
VARIAN Inova-500	259-412
VARIAN Inova-501	103-176
VARIAN Inova-502	114-216
Bruker AVANCE-600	135-661

The amount of data generated at the DCIF has increased over time. This increase is due to an increase in the use of the instruments, not a change in the instruments themselves. Therefore, the best way to gauge this increase is by examining the billing statements, which describe the amount of minutes the facility bills each month. In 2003, the facility billed an average of 99.5 kilominutes per month and a total of 1194 kilominutes for the year. In 2008, the monthly average increased to 115 kilominutes and the total billed time increased to 1382 kilominutes. Since there is a limit to the amount of time that the instruments can be used, the data generated at the DCIF has the potential to plateau (assuming that the facility does not add any new instruments). However, based on current trends, the facility does not expect to reach this plateau in the next five years, and predicts the increase in billed time to be similar to the increase over the past five years. This would result in a monthly average of 133 billed kilominutes and an annual total of 1595 billed kilominutes in the year 2014.

Although many faculty members in the Chemistry Department use the DCIF, others may conduct experiments at different facilities, either at MIT or other collaborating institutions. Additionally, there are a number of scientists in this department who generate data from models and simulations and do not conduct experiments at all. These scientists often generate a lot of "intermediate" or "temporary" data while their models run, but only need to save a small amount of this data in the end. For example, the scientist in this department studying quantum chemistry generates approximately 32 gigabytes of temporary data per run but then condenses that data into about 1 megabyte of what he considers "output data." The temporary data usually contains more information than the research group is able to analyze or store, so they sift through and keep only the information needed for the given project. For the quantum chemist, the research data is usually the x, y, and z coordinates of an atom in a certain chemical situation, while the temporary data would usually include "everything you would want to know about the atom." Although he does not generate large

[2] http://www.chemguide.co.uk/analysis/nmr/highres.html

amounts of data, the data production has increased over time. Five years ago he stored approximately 100 gigabytes of data, and today he has about 1 terabyte of total data stored. In the future, he predicts that data generation will continue to increase as computers get faster and storage prices decrease.

Metadata

Like the raw experimental or model data, the metadata generated by chemists and chemical engineers at MIT varies based on the research project. For example, the metadata for the chemist mentioned in the previous section is mostly descriptors of the accuracy of the model used. This information is stored in text files that are usually about 1 kilobyte per project. This metadata is often stored in lab notebooks as well.

Data Retention

There are no formal data retention policies for the Chemical Engineering or Chemistry Departments. Each scientist decides how to store, manage and back up his or her own data. One chemical engineer, who generates High Performance Liquid Chromatography (HPLC) and mass spectrometry data, will store all of her data in two locations: on the HPLC instrument, and on her students' personal computers. Each student is in charge of his or her own data. Lab members use MIT's central back up service to back up the data on their personal computers, but not the data on the instrument computers. While MIT's back up service does offer nightly back up, most students are not constantly connected to the MIT network and need to manually back-up their data on their own schedule. In addition to the data storage on computers, this scientist also keeps hard copies of the data that she personally generates. First she records all of the data in her lab notebook, and then she prints out copies of all of the data generated by an experiment. While she believes in creating paper back ups, not all of the students in her lab practice this technique.

While the Chemical Engineering and Chemistry departments as a whole do not have specific data retention policies, since their disk space is limited the DCIF is trying to institute a five-year data retention policy for the facility. Currently, this facility has approximately 1.5 terabytes of disk space on the main lab computers. If they get close to capacity, then the facility will delete the oldest files. If the new five-year retention policy were implemented, then the facility would automatically delete any data that has been stored on the main lab computers for more than five years.

The DCIF backs up all of the data that is generated in their lab onto DVDs. Each week they copy two weeks worth of data (so each week's data is eventually copied twice). The DVDs are kept onsite. Based on the data generation estimates discussed in the previous section, this facility backs up about 330 GB of data per year.

Data Sharing and Reuse

While many chemists and chemical engineers from different laboratories do share data there are no widely used national data repositories. For the experiments run by the chemical engineer mentioned in the previous section, the experimental conditions are often more important than the results. She will often call her colleagues at different universities or labs and ask about their experimental conditions.

Another scientist in the Department of Chemistry sometimes uses the National Institute of Standards and Technology's Computational Chemistry Comparison and Benchmark Database. This database is a collection of experimental and ab initio thermochemical properties for a selected set of molecules. The goals of this data collection are to: 1) provide a benchmark set of molecules for the evaluation of ab initio computational methods; and 2) allow the comparison between different ab initio computational methods for the prediction of thermochemical properties. The thermochemical values included in the CCCBDB are enthalpies of formation, entropies, heat corrections (integrated heat capacity), data needed to compute thermochemical properties (such as geometries, rotational constants, vibrational frequencies, barriers to internal rotation, and electronic energy levels), and additional computed properties (such as atomic charges, electric dipole moments, quadrupole moments, polarizabilities, and HOMO-LUMO gaps)[3].

Key Trends and Indicators for Data Growth

Although the different researchers in the Departments of Chemistry and Chemical Engineering study a range of different topics, there are several key trends and indicators for data growth that we have identified:

1. MIT's Department of Chemistry Instrumentation Facility generates approximately 165 GB of data per year, and an additional 330 GB of data backups. The majority of this data is generated by the seven Nuclear Magnetic Resonance (MR) spectrometers. Of these seven NMR spectrometers, the biggest data generator is the Bruker AVANCE 400 MHz NMR spectrometer with Spectro Spin superconducting magnet. This instrument generates approximately 677-843 MB per week.

2. 12% of the total "instrument use time" at the DCIF is from industrial customers. The remaining 88% is used by groups from academic institutions (MIT research groups account for 84% of "instrument use time" in this category).

3. In addition to the scientists using the DCIF, there are also many chemist and chemical engineers that generate data with models instead of instruments. While these scientists can generate large amount of data while the model is running, most of this data is the temporary outputs of calculations and is either deleted or significantly condensed before the model is finished running.

4. Data retention and back-up policies vary depending on the preference of the specific researcher. Some scientists will delete their data after publishing a paper, while others keep everything until their storage capacity is reached (and then delete the oldest files). Others will keep their data forever and buy more storage if they reach their current storage capacity.

[3] http://cccbdb.nist.gov/

Climate Change at MIT

Background

In recent years, our society has become more aware of the delicate balance of the Earth system, and has devoted much time and energy to debates over how best to ensure a sustainable future for the planet. The Earth System Initiative (ESI) is predicated on the notion that, to be meaningful, these debates must be informed by reliable scientific data regarding the evolution and current state of our planet. ESI scientists and engineers marshal their efforts around four broad research themes:

- System Characterization
- System Organization
- Evolutionary Processes
- Human Impacts

The Earth System Initiative facilitates the development of large-scale research efforts in key areas of Earth system science and engineering. In December 2006, the Darwin Project, the first example of such an undertaking, was launched.

The Darwin Project is an ESI initiative to advance the development and application of novel models of marine microbes and microbial communities, identifying the relationships of individuals and communities to their environment, connecting cellular-scale processes to global microbial community structure.

For this case study, three scientists in MIT's Department of Earth, Atmospheric and Planetary Science were interviewed. These research scientists focus on the large scale modeling of the physical and biological processes in the global oceans. To do this, they build large numerical simulations that are constrained with observational ocean data.

Data Generation

Numerical models produce about 90% of the data generated by scientists in the Earth Science Initiative within the MIT Department of Earth, Atmospheric and Planetary Science, and particularly for the Darwin Project. The remaining 10% is observational data that is recorded by NASA satellites or oceanographers. There are about 20-25 people who work on the Darwin Project at MIT. They post all of their numerical models on the project website for others in the field to download and use. The primary research estimates that there are another two hundred people around the world who download these models to use in a variety of different ways. The group at MIT communicates with these other scientists through the web, and often collaborates on projects.

Over the last year, this project has generated about 200 terabytes of data. The majority of the data is from high resolution calculations that model the processes (both physical and biological) occurring in a certain area of the ocean. One high-resolution calculation will run for about 1-2 months and will produce about 60 terabytes of data. However, the amount of data produced is very dependent on the specific processes that are being modeled.

The amount of data produced by these models has increased over time. However, this increase is largely due to better storage technologies, not changes in the models. Five years ago, the research group was "theoretically capable" of generating just as much data as today, but they did not have enough storage to handle the size of the data files. As the storage hardware continues to improve and become less expensive, the amount of data that is generated by improvement to the resolution of the models will continue to increase. According to the primary research scientist, mathematically speaking, there is "no upper bound." Since 2003, the data generated by these researchers has increased by a factor of 100. In the next five years they predict that it will increase by another factor of 100 as the computer infrastructure continues to become less expensive and more widespread. Based on these predictions, the group could produce about 20 petabytes of data in 2014.

Metadata

Metadata for this project's research includes the grid (area of the ocean) that the model is using, the physical fields that are produced by the model configuration, and the biological fields being modeled. Like other areas of scientific research, the amount of metadata is very small in comparison to the experimental or model data produced, but is critical to making use of the primary data.

Data Retention

The data generated by this group of researchers is stored on their computational cluster's file system. The cluster is a collection of 750 hard drives, with a certain amount of redundancy in case a drive fails. Currently, they do not have the capacity to do a redundant back up of the entire cluster. Instead, they use national facilities to back up important data. One facility that they frequently use is the NASA-Ames Lab, which has an archive system to which they can transfer data over the network. It is relatively simple (although sometimes time consuming) to re-run a model, so data can also be reproduced if it is lost.

The storage capacity of his team's cluster is 500 TB. They have had this storage technology for about 18 months. Before purchasing this hardware, the team was storing all of their data the NASA-Ames facility, which had 100 terabytes of storage available for them to use, however the transfer time was very slow (approximately 1 terabyte per day).

This research group saves the source code for all of their model configurations, but has no other data retention policies. The individual scientist running a model will usually decide what data to keep, and what to delete. Five years after a project is completed, only about 10% of the data from that project is still available. The research group employs a computer system administrator to manage the computational cluster, however he only informs the researchers when they are close to their storage capacity limit. He does not make any decisions about data retention.

Data Sharing and Reuse

As mentioned earlier, this research team posts all of their computational models online and is open to sharing their data with other researchers in this field. They also use the NASA-Ames facility to archive important data. However, the NASA-Ames archive is only shared with immediate collaborators and is not publicly accessible.

In addition to sharing their data with other researchers, the research group is constantly re-using their own data, mainly to re-run models to test for reproducibility of results, or to re-analyze data with different models.

Key Trends and Indicators for Data Growth

Several key trends and indicators for data growth can be identified for the Earth Science Initiative on Climate Change:

1. The amount of data currently generated is more than 200 terabytes per year (based on one large project within the Initiative).
2. Increases in computing power and storage capacity have caused the amount of data generated to increase by a factor of 100 over the past five years. If hardware trends continue, in the next five years data production could reach 20 petabytes annually.
3. Like other scientific areas, although metadata is important to research projects, the amounts generated are not significant.
4. Data retention decisions are up to the individual researchers and retention is not a major concern today. Retention is constrained by the volume of data produced, and is facilitated by use of national data archiving facilities (managed by NASA, in this case).
5. Data sharing and reuse are commonplace in Climate Change research, and would be facilitated by improved data storage and archiving capabilities.

Materials Science and Engineering at MIT

Background

There are 41 faculty members in MIT's Department of Materials Science and Engineering, covering a wide range of expertise that includes both theoretical and applied research, with interests spanning the entire materials cycle from mining and refining of raw materials, to production and utilization of finished materials, and finally to disposal and recycling. For example, one scientist studies the coupling phenomenon that occurs at materials interfaces. By exploring coupling at the fundamental force and length scales of atoms and molecules, she looks for commonalities among materials ranging from metallic crystals to living biological cells that her research group can exploit for human advantage in sensing, actuating and transduction applications. Another scientist studies how materials change at the atomic level by applying external stimuli like plastic deformation, bombardment by energetic ions, or exposure to rapidly varying temperatures. By understanding how materials respond to these stimuli at an atomic level, this scientist hopes to create strategies for designing materials with desired properties from the atomic scale up.

For this case study, two researchers from the Department of Materials Science and Engineering were interviewed during the spring semester of 2009.

Data Generation

Like most other scientific or engineering fields of study, the amount of data generated by materials scientists and engineers depends on the specific research goals and the experimental or computational techniques employed by the individual researcher.

The scientist studying materials interfaces generates two kinds of data: experimental data and data from simulations. The experimental data is generated by different instruments run by open source code developed by researchers in the lab. One example of the type of instrument used is an atomic force microscope, which provides pictures of atoms on or in surfaces by scanning a fine ceramic or semi-conductor tip over that surface. This tip is set at the end of a cantilever beam that will deflect as the tip is either repelled or attracted to the surface, and the magnitude of that deflection is captured by a laser and plotted, providing the scientists with the resolutions of the surface topography[4]. Other examples of instruments include indentures (machines that pull materials) and optical microscopes.

The resulting file generated by one of these instruments is about 10 megabytes and includes all of the raw experimental data, as well as the details describing the operating parameters. This raw data are typically images (see Figure 1). This scientist runs about two of these experiments per day, resulting in approximately 7 gigabytes of raw data each year. The lab then analyzes the raw experimental data, which generates another 10 megabytes per experiment. As a result, the lab generates a total of approximately 14 gigabytes of experimental data per year.

The simulations run in this lab generate the bulk of the data. These simulations are based on calculations made from full electronic models of the materials being studied. The closer the simulation is to the scale of the electronic model, the more storage it requires. For a typical simulation, lab members would investigate approximately 100,000 atoms and generate about 5 gigabytes of raw data that describes the structural and functional states of the atoms during the simulation. A whole study would require about 30 simulations and generate 150 gigabytes of data. This data generation phase lasts about three months and is followed by six months of data analysis. Unlike the experimental data analysis which doubles the amount of data generated, the simulation data analysis only generates an extra gigabyte of data per study. At any given time, the group is conducting about three different simulation

[4] http://www.che.utoledo.edu/nadarajah/webpages/whatsafm.html

studies resulting in approximately 450 gigabytes of simulation data generated each year.

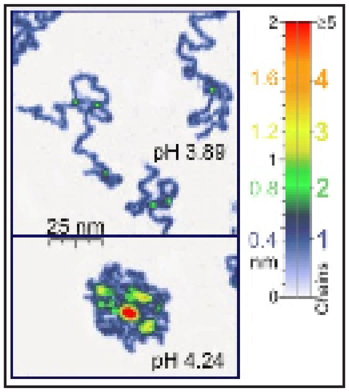


Figure 1: Single molecules of poly(2-vinylpyridine) recorded using an AFM operating in tapping mode under water media of different pH[5].

In total, this research lab is currently generating approximately 460 gigabytes of experimental and simulation data each year. This is an order of magnitude more data than was generated five years ago due to an increase in research volume (i.e. personnel, funding, and improvements in computing power, speed and storage). The scientists has only been at the Institute for five years and this increase is typical for new faculty. This order of magnitude increase in data generation has mostly been in the simulation data; the amount of experimental data has only doubled in the past five years. This is because the instruments used to generate the experimental data are very expensive and are only replaced every 8-10 years. In the future, the scientist predicts that the biggest change will be the need for data storage from the simulations because the output from these simulations is increasing at a dramatic rate. The amount of experimental data, on the other hand, will probably only double in the next five years due to increased personnel, not data density.

Unlike the scientist described above, almost all of the data generated by the scientist studying material change is computational model data. However, he does exchange data with experimental scientists with a goal of connecting his modeling research and the experimental research in the field. He is particularly interested in iron beam analysis, nuclear reaction analysis, and image-based experiments.

As mentioned earlier, this scientist's research is focused on how materials change when subjected to external stimuli. A project begins by constructing an atomic-scale model of the material of interest. The model is then run through different simulations of external stimuli (i.e. extreme temperature changes, plastic deformation, etc.) and the outputs of those simulations are a series of "snapshots" of how the atomic system looks at different times or states. The research group then analyzes the state of this system and decides whether the initial model inputs were relevant. They then iterate this process and, over time, begin to notice the important elements in the model. The group then performs targeted simulations on those elements. Over the course of this process, the scientist compares the outputs of the simulations with experimental data.

Since this process is so iterative, the types of simulations run throughout one research project (and therefore, the amount of data generated) can greatly vary. For example, the lab may run a simulation that only lasts a split second in order to get a feel for how the simulation and model works. On the other hand, one simulation could also run for weeks or months on large-scale supercomputer at a national center. These large simulations, however, do not necessarily generate a lot of data. For example, if the researcher is only interested in learning about the pressure profile of a material over time under certain conditions, a hundred million-atom simulation will only produce a few kilobytes of data. However, if he is interested in studying the whole state of a system at every time step, a much "smaller" simulation (fewer atoms) could generate terabytes, or even petabytes of data.

After the simulation is run, the research group will perform analysis that could increase the amount of data by a factor of ten. However, the amount

[5] Roiter and Minko, S. AFM Single molecule experiments at the solid-liquid interface: in situ conformation of adsorbed flexible polyelectrolyte chains. Journal of the American Chemical Society, vol. 127, no. 45, pp. 15688-15689 (2005).

of analysis data can vary based on the goals of the specific project. For example, 90% of the scientist's PhD research data was analytical (only 10% was raw data from simulations) while his postdoc project mostly consisted of raw simulation data (25% was analytical data). In general, the smaller the simulation, the more detailed the analysis because larger simulations produce too much data to analyze in detail.

Despite the potential to generate large amount of data, this scientist's projects have produced only modest amounts. During his PhD project he generated a total of 200 gigabytes of data and for his postdoc project he produced approximately 1 terabyte. He produced this relatively small amount of data because, for these two particular projects, he tried to avoid big simulations that would have required reserving time on national supercomputers and instead worked with the smallest possible computer systems.

Metadata

Like the raw data, different scientists define materials science and engineering metadata differently. For example, the scientist does not believe that any of the information generated by his research projects should be considered "metadata" because everything is valuable to his research. He considers the models he develops, the conditions of the simulations, and all other parameters to be "data" not "metadata."

The other scientist, who runs both experiments and simulations, will generate metadata from her instruments (describing the date, time, temperature and other experimental conditions), and from her simulations (explaining the version of the software, the number of atoms in the simulation, date, etc). The experimental metadata is either recorded in lab notebooks or included in the instrument's output file. The simulation metadata is included in the header of the output files. Neither type of metadata is very large and is insignificant in size when compared to the raw data generated. The importance of the metadata varies depending on the sophistication of the experiment or simulation. The more complex methodologies, the more important the metadata becomes.

Data Retention

There are no common data retention policies for the Department of Materials Science and Engineering and therefore, each faculty members tackles the issue differently. For example, one scientist's research data is stored on her students' and postdocs' personal computers. The data is replicated twice during the data production process (it is kept on the computer it was generated on, and on the computer that it is analyzed on). The data is only backed up after a publication has been submitted, and no automated backup system is used. Each student in her research group is in charge of the data that they generate. The first author listed on the final publication is responsible for backing up the data onto a CD or portable hard drive. Each year, this scientist assigns one of her students the task of sorting through the data on the lab's cluster and deleting data from students who are no longer with the group (assuming that the data has been published and, therefore, backed up).

Another scientist stores all his data on his personal computer and external hard drives. One of these drives has automatic back up. His total storage capacity is 2 terabytes. This capacity will increase in the near future (before the end of the year) because he is in the process of setting up a new computational cluster, which will have 15 terabytes of storage. This scientist is in charge of all his own data management, including data retention decisions. He does not delete any of his data, and would rather buy more storage then delete old data. If he used a national supercomputer center for a simulation then all of the data generated during that simulation is also stored at the cluster where it was generated. There are no data retention policies at these clusters, but there are strict rules about data sharing.

Data Sharing and Reuse

There are no national data repositories for data sharing in materials science. However, scientists may share data on a lab-by-lab basis. Both of the scientists described above share their data a few times a year with colleagues at other universities or research institutions. One of the scientists will usually share models but not the data generated by running the model through simulations. The other scientist will often share the outputs of the simulations. In addition to sharing data and models, both scientists also reuse their own data by running new analyses on older models or raw data from previous projects.

Key Trends and Indicators for Data Growth

Although the different researchers in the Department of Materials Science and Engineering study a range of different topics, there are several key trends and indicators for data growth that we have identified:

6. The amount of data currently generated in the Department of Materials Science and Engineering is approximately 32 terabytes per year (based on scientists currently in this department, and assuming larger research groups and more funding opportunities allow researchers to generate 75% more data than assistant professors, and full professors to generate twice as much data as assistant professors).

7. Increases in computing power and speed have caused the amount of data generated in this field to increase by an order-of-magnitude in the past five years. As computer get faster and storage get cheaper, data generation will increase exponentially in the next five years.

8. Although metadata is important to research projects, materials scientists and engineers to not generate significant amounts.

9. Data retention decisions are often up to the individual researchers. Retention policies vary by lab.

10. There are no commonly used national data repositories, but individual scientists are open to data sharing and may share their data (raw experimental data, simulations outputs, and computational models) with colleagues at other laboratories, universities, or research institutions.

Neuroimaging at the Martinos Imaging Center

Background

The Martinos Imaging Center is a collaboration among the Harvard-MIT Division of Health Sciences and Technology (HST), the McGovern Institute for Brain Research, Massachusetts General Hospital, and Harvard Medical School. The center opened in 2006 and provides one of the few places in the world where researchers can conduct comparative studies of the human brain and the brains of differing animal species.

There are 12 principle investigators working at the Martinos Imaging Center. While each PI's research project is distinct, they all share core interests in three interrelated research areas: perception, cognition and action. For example, one scientist aims 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. To do so, he examines brain-behavior relations across the life span, from children through the elderly. Another scientist's lab focuses specifically on the cognitive and neural processes that support working and long-term memory. Participants in her research are healthy young adults (e.g. MIT students), healthy older adults, and patients with neurological diseases (e.g. amnesia, Alzheimer's and Parkinson's diseases). The overall goal of the research conducted at the Martinos Center is "to meet one of the great challenges of modern science – the development of deep understanding of thought and emotion in terms of their realization of the brain."

In addition to the scientists at the Martinos Imaging Center, there are also a number of other researchers at MIT who use neuroimaging techniques in their work. For example, one research scientist at the Research Laboratory of Electronics combines behavioral and neuroimaging studies to explore the processes underlying speech production and perception. For this case study, two scientists at the Martinos Imaging Center and one researcher at the Research Laboratory of Electronics were interviewed.

Data Generation

Magnetic Resonance Imaging (MRI) is a medical technique used to produce images of the internal structure and function of the body. MRI scanners use a magnetic field to align the nuclear magnetization of (usually) hydrogen atoms in water in the body. They then systematically alter this alignment using radio frequency (RF) fields. As a result, the hydrogen nuclei produce a rotating magnetic field, which is detectable by the scanner. By manipulating this signal with additional magnetic fields, enough information is generated to construct an image of the body[6].

There are two types of brain images that are studied by researchers at the Martinos Center: structural magnetic resonance images (structural MRI), which document the brain anatomy, and functional magnetic resonance images (fMRI), which document brain physiology. fMRI measures the hemodynamic response (i.e. the process that occurs when blood releases oxygen to active neurons at a faster rate than inactive neurons to provide them with energy) 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, and therefore, the hemodynamic response in the brain to activity results in magnetic signal variation, which can be detected by an MRI scanner. In order to perform an fMRI scan, the machine must also acquire structural scans. The Martinos Center contains three sunken bays for the magnets used in fMRI. Two of these bays house actual MRI machines and one is reserved for a next-generation technology that the MIT community of researchers will help develop.

One bay holds a new 3 Tesla Siemens Tim Trio 60

[6] Novelline, Robert. Squire's Fundamentals of Radiology. Harvard University Press. 5th edition. 1997. ISBN 0674833392.

cm whole-body fMRI machine. Tesla refers to the strength of the magnet, and 3 Tesla is as strong as considered safe and practical for people. While this is considered an fMRI machine, it also has EPI, MR angiography, diffusion, perfusion, and spectroscopy capabilities for both neuro and body applications. The visual stimulus system for fMRI studies uses a Hitachi (CP-X1200 series) projector. The image is projected through a wave-guide and is displayed on a rear projection screen (Da-Lite).

The second bay has a higher power 9.4 Tesla MRI for animal studies. This machine provides higher resolution images, which can then provide insights into areas to be explored in human studies. For example, such animal scans led to the discovery that the frontal cortex is involved in working memory. In addition, MIT researchers investigating the role of specific genes in brain functions can use the imaging center to literally see the difference that genetic manipulations in animals produce.

The image datasets produced by the fMRI machines are Digital Imaging and Communications in Medicine (DICOM) files. DICOM is a standard for handling, storing, printing and transmitting medical images, which includes both a file format definition and a network communications protocol. DICOM enables the integration of scanners, servers, workstations, printers, and network hardware from multiple manufacturers into a picture archiving and communication system (PACS) and 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. A run is distinguished by the kind information the researcher wants. There are anatomical runs, which are high-resolution scans of the anatomy of the brain; there are functional runs, which are low-resolution images of the hemodynamic state of the brain over time; and there are other special-purpose runs, like 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 run generates a series of images. The number of images can vary depending upon how long the run lasted. Therefore, each session results in hundreds or thousands of DICOM images. The average session will produce 1.4 gigabytes of DICOM images.

Each DICOM file contains a metadata section and a data section. There are about a dozen image types stored as DICOMs. Examples include blood-oxygen-level-dependant (BOLD) images and diffusion tensor images (DTIs). Each image type represents something different. For example, different types of electromagnetic pulse sequences (different tissues are sensitive to different pulse sequences, so different pulse sequences are used).

After the scan is complete, analysis packages make copies of the DICOMS and then convert them to a different file format for storage. One common format is the Neuroimaging Informatics Technology Initiative (NIfTI). Unlike the DICOM standard, which attempts to address the general requirements of digital imaging in diagnostic and therapeutic healthcare environment, the NIfTI standard was developed and implemented by neuroscientists to meet the specific needs of their discipline. While the DICOM standard has a large, clinically focused storage overhead and relatively complex specifications for multi-frame MRI and spatial registration, NIfTI is relatively simple format that has low storage overhead, resolves some immediate format problems in the fMRI community and is not difficult for developers to learn and use.

The NIfTI format allows you to either coalesce all the files for one session into one monolithic 4D file (see Figure 1), to keep a series of separate 3D files, or to keep a one-to-one mapping from DICOM to NIFTI (see Figure 2). After the DICOM files are copied and converted to NIfTI files, various software packages transform the NIfTI files into "intermediate files". There are 8-9 "intermediate data files" for each NIfTI file. Examples of intermediate files include slice-timing corrected NIfTIs, motion corrected NIfTIs, realigned NIfTIs, smoothed NIfTIs, and normalized NIfTIs. These transformations lead to a lot of wasted disk space because there are 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 (shown in Figure 2).

The Martinos Imaging Center sees about 30 human subjects/week (1500/year). Each subject has one session and each session produces a total of 3.6 gigabytes of data. The scanner is booked all year (approximately 50 weeks). Therefore, the center is generating a total about 5.4 terabytes of human image data each year (this estimate includes fMRI scans and the structural MRI scans required to

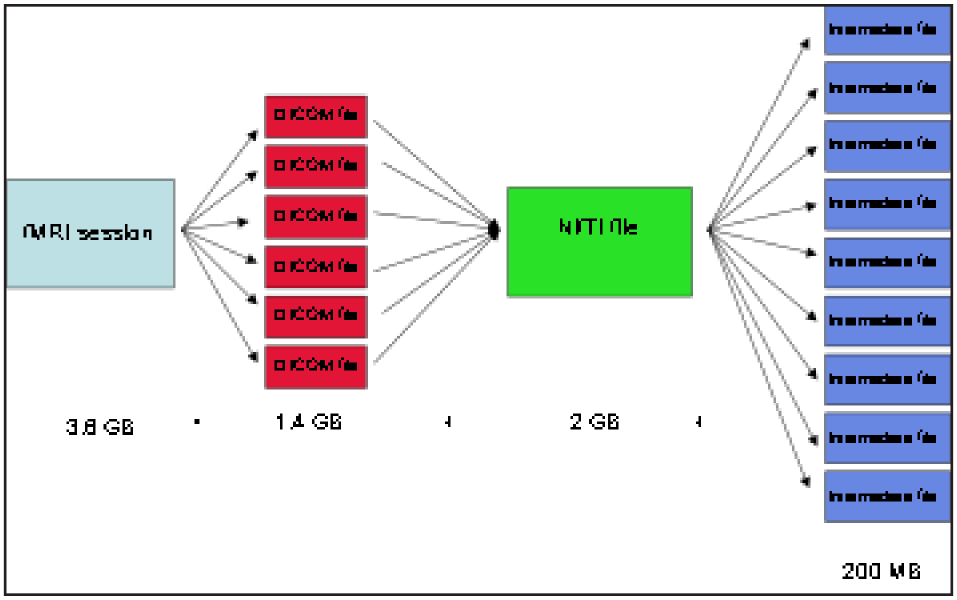


Figure 1: Files produced by 1 fMRI session, with one monolithic 4D NIfTI file

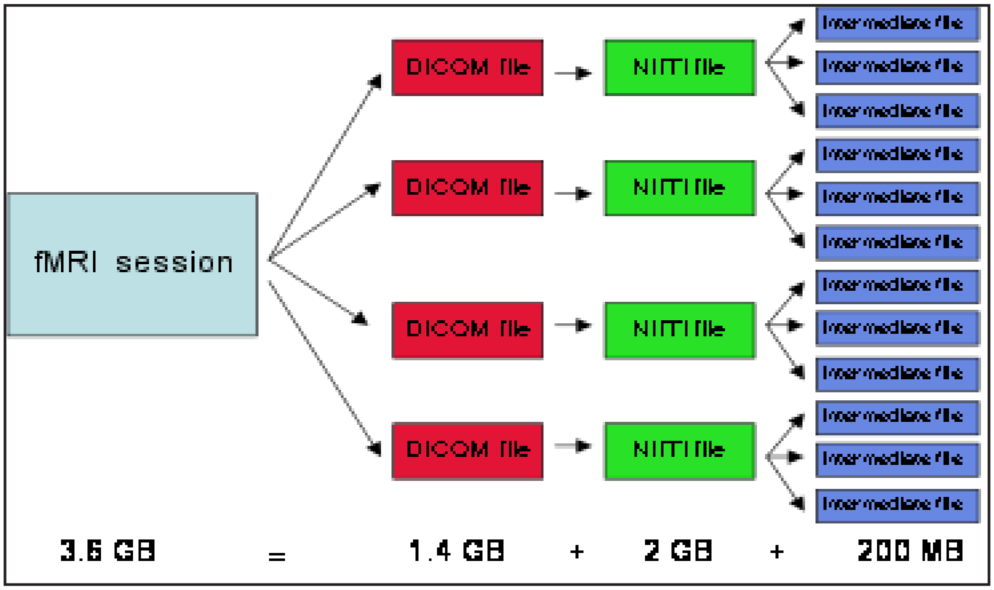


Figure 2: Files produced by 1 fMRI session, with one to one DICOM to NIfTI mapping

perform the fMRI scans).

Each fMRI scan is very expensive. It costs about $550/hr for the scanner time (including the staff). In addition to scanner time, there is the cost of recruiting, screening and compensating subjects. The total cost per subject is approximate $750-$1,000.

Although the majority of data generated by researchers at the Martinos Center are fMRI and structural MRI images, many researchers combine these images with additional data about the subject in order to fully understand what they observe in the brain. For example, one scientist often collects demographic information, health histories, behavioral data and genetic information from her subjects. However, the amount of non-image data is significantly smaller than the MRI image data. As a result, this scientist performs behavioral experiments on many more subjects than she is able to scan. In 2008, she gathered behavioral data on 260 subjects (five per week) but performed only 52 MRI scans (about one per week). After post-processing and analysis, this resulted in a total of 230 gigabytes of image data (structural MRI and fMRI) and only a few hundred kilobytes of behavioral, demographic and genetic data. If each of the other 11 researchers at the Martinos Center generated a similar amount of non-image data, this would only result in about 2-4 gigabytes of data per year (an insignificant amount when compared to the 5.4 Terabytes of image data produced each year).

Since the amount of non-image data is relatively small when compared to the amount of MRI scan data, the data generation growth rate for the field of neuroimaging depends on the fMRI scanners. Since they only have one scanner, and the scanner is booked for the entire year, the amount of data generated each year at the Martinos Center has remained relatively constant. However if there were more scanners, they would be able to increase the number of subjects per week and therefore produce more data.

Like the scientist described above, the research scientist in the Research Laboratory for Electronics also uses a combination of behavioral data and fMRI images in his effort to study speech. There are four sets of behavioral data that he generates. The first set is the experimental protocol itself. This includes the information that this scientist aims to acquire with the behavioral experiment and the methods for acquiring that information (i.e. the scripts used for the behavioral experiment). The other three sets of data are generated during the experiment itself. Subjects are exposed to a stimuli will be asked to respond on a keyboard, which generates the first set of data. Video and audio data is also recorded throughout the experiment, generating the second and third experimental datasets. During the behavioral experiment, the video camera is directed at the lower half of the subject's mouth. This video serves as back-up to the audio data. After the experiment, this video data is immediately saved to DVDs and is only watched by the scientists if he hears something abnormal in the audio recording. The most important sets of data from the behavioral experiments are the typed responses and the audio data. For a given session with a subject, this researcher will generate about 50-100 megabytes of audio and typed response data. His research group conducts about 20-30 sessions per year resulting in a maximum of 3 gigabytes of audio and typed response data. These files are then processed, which increases the size of the datasets to about 1 gigabyte per subject, or approximately 30 gigabytes of audio and typed response data per year. As mentioned earlier, the video data generated during the behavioral experiment is saved straight to DVD and is not processed. Each session generates about 8 gigabytes of video data (two DVDs). Therefore, this scientist's research laboratory generates approximately 240 gigabytes of video data each year. Overall, his behavioral experiments generate a total of 270 gigabytes of raw and processed data.

This scientist then gathers neuroimage data which he will combine with the results of the behavioral studies. He will typically obtain both structural and functional MRIs. All of his scans are run at the Martinos Imaging Center. Each subject generates a total of about 1 gigabyte of neuroimage data. The structural images will be about 16 megabytes while the fMRI images will be about 984 megabytes. There are approximately 20 neuroimaging subjects per year (these are different subjects than those in the behavioral experiments), resulting in 20 gigabytes of raw neuroimage data per year. The image data is then analyzed, which drastically increases the size of the dataset. After analysis, the structural image datasets will increase to about 200-240 megabytes (including the original 16 megabyte raw data file) and the fMRI datasets will double in size to about 2 gigabytes. As a result, the total amount of raw and processed neuroimage data generated by this scientist is approximately 67 gigabytes.

The rate of data generation will increase as the hardware and software on the scanners improve. One scientist predicts that in five years the state of the art fMRI scanners will have more channels for data acquisition, which could increase the size of the files produced by each scan session by a factor of 10. As mentioned earlier, the Martinos Center has a third sunken bay reserved for next generation technologies. If the center were to purchase a new scanner with the predicted technology improvements, then in 2014 the Martinos Imaging Center could produce approximately 60 terabytes of data (assuming that the scanners are booked all year).

In addition to a new fMRI scanner, the Martinos Center could also purchase a number of different technologies that could increase the amount of data produced each year. For example, one scientist's lab will soon have Electroencephalography (EEG) technology. This technology measures the electrical signals recorded at the surface of the scalp. Although EEG's have lower spatial resolution than fMRI, they have higher temporal resolution and are widely used in the field of neuroimaging. The McGovern Institute has also raised funds towards acquiring a Magnetoencephalography, or MEG, machine at MIT. This technology is similar to the EEG but based on magnetic rather than electric signals. The MEG has better spatial resolution than the EEG and also detects signals that are orthogonal to those of the EEG. The McGovern Institute hopes to add the MEG capability in the next 2-3 years.

Metadata

There are approximately 170 fields of metadata associated with each fMRI scan, which are kept in header files for each scanner. Examples of fMRI metadata include:

- Name of Principal Investigator
- Scanner manufacturer
- Information about the actual scan sequence including patient position, nucleus being imaged, and repetition time. This information is necessary for comparing images from two different scans "apples to apples."
- Subject demographic information

Typically, it is the MR physicist, not the neuroscientists, who use the metadata associated with each image in their research. This is because most of the metadata describes the specifics of the fMRI scan sequence, not the subject being scanned. Therefore, it is used to help replicate scans, but not in the data analysis.

In addition to scan sequence information, the subject's demographic information is also extremely important metadata for the researchers at the Martinos Center. However, the amount of demographic information needed depends on the goals of the faculty member's research. For example, one scientist only needs a small amount of information about his subjects: mainly age, gender, and "handedness" (i.e. what hand the patient writes with, left or right). Another scientist needs her subjects to submit an entire "patient fact sheet" describing their medical history, education history, drinking and smoking habits, and the geographic areas where they have lived. For her research, the MRI scans are useless without this metadata.

NIfTI is the de-facto standard for neuroimaging data. It defines the standard set of header information that should exist for neuroimaging data. However, it is the DICOM files that contain most of the metadata, not the NIfTI files.

Although this metadata is important to the research conducted at the Martinos Center, the amount generated is small compared to the size of the image files. As mentioned earlier, one of the scientists who needs much more metadata than the other, only generates a few hundred kilobytes of this data each year.

Data Retention

There is no centralized data storage system for the Martinos Imagining Center. One scientist's lab shares a storage system with three other PIs at the Center. Although the focus of their research differs, each of these four scientists recognized the importance of data storage and decided to jointly acquire the system. A research specialist is in charge of managing this system.

This specialist's data storage system involves a server that uses a Network Files System (NFS) to allow the scientists to share files over the network. The server is physically connected to many different RAID arrays. Although this storage technique is not as sophisticated as Network Area Storage (NAS) or Storage Area Networks (SAN) solutions, it is less expensive and the group has a constrained budget. Other scientists at the imaging center employ a variety of different storage techniques ranging from high performance storage clusters to "Mac mini" laptops with no backup.

The shared storage system was implemented in January of 2008. Since then, the four scientists have stored about 25 terabytes between their labs. About 3 terabytes came from existing data that the scientists had previously stored on various computers. Therefore, 22 terabytes have been generated since January 2008, or about 2.2 terabytes/month (about ½ terabyte/scientist/month). Before the system was implemented, each scientist stored their data on their own computer.

The capacity of the current storage system is 44 terabytes. The research specialist in charge of it predicts that the system will reach capacity by the end of the year. Once this happens, the lab hopes to move to a more scaleable storage application. One possible storage application is the NetApp, which uses Network Area Storage. This new system would be much faster, and more reliable and fault tolerant than the current system.

The shared storage system uses MIT's central backup service for backup, which they selected because it is affordable, relatively easy to use, and the lab does not have to maintain any of the hardware. Every evening the lab performs an incremental backup of all of their data, sending it over the MIT network to a secure server located on campus. The rest of the researchers at the Martinos Center use a variety of different methods for storing their data. One scientist, for example, keeps all of her MRI data on a server in a local hospital. This server has a 2 terabyte capacity, is backed up every day, and is managed by an IT department at the hospital. Additionally, this scientist makes copies of all of her DICOM files on CDs, which she keeps at MIT (each scan fills about two CDs). All of her non-image data is kept at the Martinos Center on a Mac G4. This computer has a 500 gigabyte storage capacity and is managed by her graduate students. Like the other system mentioned, this scientist uses MIT's central backup service to back up the data she stores at MIT. However, she also keeps hard copies of all of the patient fact sheets on campus.

Despite differing storage techniques, researchers at the Martinos Center seem to share 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. As mentioned earlier, the cost per subject is about $750-$1,000. Additionally, although an experiment can be reproduced if the data was lost, the lab could not use the same subject because they could have memorized the visual stimuli.

While he can save some of his image data at the Martinos Center, the researcher at the Research

Laboratory of Electronics has his own data storage hardware and policies. He has a new 8 TB server located in his lab that is used by all of the research groups in his building. This server currently has about 2 TB of data stored on it. For backup he uses a combination of local backup on external discs and the MIT backup system. Sometimes, he will also burn all of his data on DVDs for a third form of back up. Like the scientists at the Martinos Center, this researcher has not deleted any of his old data and will buy more storage if he reaches capacity (instead of deleting old data). His oldest is stored on tapes.

Data Sharing and Reuse

While data sharing across labs, institutions, and disciplines is limited in the field of neuroimaging, data is commonly reused within labs. There are two major data analysis packages used by neuroscientists: Statistical Parametric Mapping (SPM) and Free Surfer. Each package is based on a different philosophy on how the brain works, and while they result in the same kinds of answers, they get there in a different way. Neuroscientists will often use one of these analysis packages to analyze their data, and then re-run it through the other package later on to compare results.

Data is also reused to perform voxel-based morphometry (VBM). VBM measures change in brain anatomy over time and are typically used to study dysfunction. In a clinical setting, VBM is done by looking at images of the same brain over time. From an epidemiology standpoint scientists take 100, 10,000, or 1,000,000 brain images and partition them according to characteristics (sex, hometown of subject, etc.). They then use all of the images to create an "average brain." For example, if a scientist is interested in learning how emissions from a factory affected the brains of people living near by, they could take 1,000 brain scans from people living in the area and morph them into one average brain for people living by the factory for that year. They would then repeat this process over time (but not necessarily with the same subjects) to see how the average brain from that geographic area changes.

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 in light of new findings and imaging and analysis techniques. One major reason for this lack of data sharing is the sheer size of the datasets. Another reason is that many scientists in this field are protective of their data and are not open to sharing with other labs. Traditionally, neuroscientists have taken the "single lab" approach to research and have not been motivated to provide data to researchers outside of their local community. Many of the fundamental aspects of brain function, such as the questions of how brains can perceive and navigate so robustly, how sensation and action interact, or how brain function relies on concerted neural activity across scales, remain unsolved due to this lack of data sharing.

Despite the general disinterest in data sharing in this field, some research groups have started to develop platforms or networks for sharing neuroimaging data. For example, one faculty member of the Martinos Center for Biomedical Imaging has been working with a team of programmers and scientists from across the United States to develop an open source software platform designed to facilitate management and exploration of neuroimaging and related data called the Extensible Neuroimaging Archive Toolkit (XNAT). The Biomedical Informatics Research Network (BIRN), a "geographically distributed virtual community of shared resources," has also developed a database for sharing neuroimaging data. This database is called the "human imaging database." However, it only has datasets from four subjects available. Furthermore, the data from each of those subjects is stored and catalogued in different ways, making it unusable.

Key Trends and Indicators for Data Growth

Although the different researchers at the Martinos Imaging Center have different research goals, and are interested in different metadata, the majority of the data generated at this center is produced by their fMRI scanner, and therefore key trends and indicators for data growth can be identified.

1. As long as the Martinos Imaging Center only has one fMRI scanner for human subjects, the amount of data generated will remain relatively constant at 5.4 terabytes per year.

2. The rate of data generation will increase as the hardware and software on the scanners improve. In 5 years, the state of the art fMRI scanners will have more channels for data acquisition, which could increase the size of the files produced by each scan session by a factor of 10. If the Martinos Center were to purchase a new scanner in five years (and continued to use the scanner they already have), then in 2014 the Martinos Imaging Center could produce approximately 60 terabytes of data (assuming that the scanners are booked all year).

3. Data generation will also increase as the Martinos Center purchases different neuroimaging technologies. In the next 2-3 years, the center plans to have both Electroencephalography (EEG) and Magnetoencephalography (MEG) capabilities.

4. The amount of metadata needed depends on the faculty member's specific research. However, even scientists who need a relatively large amount of metadata still only generate a few hundred kilobytes each year.

5. While there are no official data retention standards, most researchers at the Martinos Center save all of their image data permanently. This is because fMRI scans are expensive, time consuming, and almost impossible to identically reproduce (the same subject could not be used again).

6. In general, scientists in the field of neuroimaging are reluctant to share data with other laboratories. However, they typically reuse their own data.

The Department of Physics at MIT

Background

The Massachusetts Institute of Technology's Department of Physics has been a national resource since the turn of the 20th century. This department is home to over 120 faculty members who conduct research on a wide variety of subject areas ranging from cosmology to string theory. These faculty members are divided into four major research divisions: Astrophysics; Atomic, Condensed Matter, and Plasma Physics; Experimental Nuclear and Particle Physics; and Theoretical Nuclear and Particle Physics. The largest of these divisions is Atomic, Condensed Matter and Plasma Physics, which spans a broad range of activities in physics, including atomic physics, optics, condensed matter experiment and theory, biophysics experiment and theory, and plasma physics. 41 faculty members and approximately 50% of the graduate students in the department conduct research in this Division. In addition to the laboratories located on campus, MIT is affiliated with over 20 other research centers and facilities, including the MIT–Harvard Center for Ultracold Atoms, the Plasma Science and Fusion Center, the Fermi National Accelerator Laboratory (Fermilab), the European Organization for Nuclear Research (CERN), the Brookhaven National Laboratory: High Flux Beam Reactor and the Laser Interferometer Gravitational-Wave Observatory (LIGO). These affiliated centers and facilities employee staff scientists who work together with MIT faculty and graduate students, as well as other universities, on joint research projects. For this case study, three scientists were interviewed from the Department of Physics, Divisions of Astrophysics and Experimental Nuclear and Particle Physics.

Data Generation

While there are approximately 30 theoretical physicists at MIT, the majority of the faculty members in the Physics Department are experimental, and therefore, most of the data generated in the department is experimental data. Many of these experiments are conducted off campus at one of the larger MIT-affiliated laboratories and run continuously for months, or even years, producing hundreds of megabytes of data per second. For example, one scientist is working with the heavy ion group of the Compact Muon Solenoid (CMS) detector experiment at CERN, located in Switzerland and one of the world's largest and most respected centers for scientific research. This scientist's research group is made up of 20 high energy or nuclear physicists who receive and analyze data produced at the CMS detector. In addition to this group, more than two thousand other scientists collaborate in CMS, coming from 155 institutes in 37 countries.

A CMS experiment at this facility runs for nine months every year, writing data continuously at the rate of about 300-400 Megabytes per second (approximately 8,165 terabytes per year). This raw data is then processed 2-3 times per year, which triples the amount of data produced by the experiment. In addition to this processing, the raw data is also combined with simulation data generated at computer centers around the world, including the lab at MIT. When the data from each of these centers is combined, the total amount of simulation data produced is about the same as the original CMS experiment (i.e. the simulations generate about 1-2 terabytes of data per week). As a result, one nine-month CMS experimental run will generate approximately 40,824 terabytes of data.

Another example of physics research being conducted at MIT is the work being done by scientists at a gravitational-wave observatory (LIGO). The purpose of this observatory is to detect cosmic gravitational waves and to develop gravitational-wave observations as an astronomical tool. The facility consists of two separate installations within the United States, operated in unison as a single observatory. This observatory is available for use by the world's scientific community, and is a vital member of a developing global network of gravitational wave observatories.

The MIT scientist is part of an international scientific collaboration, a growing group of approximately 600 researchers at roughly 40 institutions working to analyze the data from the observatory and other detectors, and working toward more sensitive future detectors.

This second scientist's research group at MIT consists of ten researchers working to analyze the data produced at the observatory. Experiments at the observatory run continuously for years at a time and produce about 1 terabyte of data per day. Since his work began on this project, the observatory has generated approximately 1,095 terabytes (1 petabyte) of data. This data was produced by five years of "half-run" experiments followed by one two-year long experiment. Like the data produced by the CERN collaboration, the data analysis and processing steps increase the total amount of data produced by the gravitational-wave experiments. However, unlike the data processing and simulation steps at the European facility, that more than triple the total amount of data generated by the experiment, the analysis done at the observatory only increases the amount of data by about 10%. Therefore, the gravitational-wave observatory has generated approximately 1,204 terabytes of data over the past seven years.

The amount of data produced has been steadily increasing over time. For example, 5 years ago MIT was not a computer center for the European facility and was not producing any data simulations. As a result, the first scientist's research group was producing about 10 times less data. In the future, the group plans to improve the capabilities of their computing system, allowing them to produce about 1.5 times more data each year. According to this growth plan, in 2014 this group alone will be generating about 540 terabytes of simulation data and all of the collaboration's computer centers combined will be generating 61,230 terabytes (approximately 60 petabytes) of simulation data.

Despite the high growth rate for the simulation data, the amount of raw experimental data produced by the CMS detector will not increase as quickly. This is because the detector technology will not change for at least another ten years. The scientist predicts that the rate of raw experimental data generation will remain constant for the next three years and then increase steadily by a factor of two each year for the next ten years as scientists improve their methods for data collection and processing. At this rate, the CMS detector experiment will be produce about 98,000 terabytes of experimental (raw and processed) data in 2014.

The growth rates seen and predicted by this scientist and his colleagues seem to be typical for physics labs at and/or affiliated with MIT Even smaller projects, like the work done by one scientist at a linear accelerator lab, have experienced large increases in data generation capabilities over the past five years. Similar to the CMS detector, experiments at this lab can run continuously for up to three years. However, these experiments typically generate much less experimental data, averaging about 1 terabyte of data per week. Despite the smaller amount of data produced, this professor's group has experienced similar data generation growth rates to the professor working with the CMS detector data. He estimates that the experiments conducted at the linear accelerator currently generate anywhere from 5-10 times more data than they were generating five years ago. This professor does not attribute this increase to changes in experimental instruments (they typically use the same instrument for a number of years), but instead to improvements in computing power and data storage technologies, which allow researchers to run their experiments for longer periods of time and to perform more advanced analyses. If the amount of data generated by the other 93 experimental physicists at MIT is similar to the amounts produced by the scientists interviewed, the Physics department as a whole is generating over 1,900,000 terabytes of data each year.

Metadata

Like many other scientific fields, metadata is often a crucial component to physics research at MIT. However, standards for recording and saving metadata are either non-existent or only defined within the specific research group. For example, researchers at the gravitational-wave observatory use an electronic logbook to record their metadata. This metadata includes experimental conditions (like start time, end time, data collection channels) as well as records of external noise that could affect the output of the gravitational wave detector. For example, when a plane flies overhead while an experiment is running, the lab technician will note the date, time, and a description of this event in the electronic logbook. The details regarding the event can be subjective and often vary based on which technician is recording the information at the time.

Metadata is also important for scientists with CMS detector data. However, the amount of metadata produced is extremely small when compared to the amount of raw and processed experimental or simulation data (about 10-30 kilobytes per experiment or simulation). Examples of metadata for a CMS simulation run by this scientist include the configuration of the experiment, the beam energy used and the physics processes selected to simulate. The metadata is uploaded to the central CERN database in Europe, where it is linked up with raw experimental data.

Data Retention

Due to the large amounts of data generated by long-running physics experiments, data storage and retention becomes a challenge for both individual researchers and larger laboratory facilities. Usually, only a small amount of actual experimental data is kept at MIT. Instead, research groups use a tiered approach, where different amounts of raw data are stored at multiple facilities depending on their capacity and research goals.

The distributed data storage used by the European collaboration is a good example of the methods used by many MIT physicists and their affiliated laboratories for data analysis, storage, retention, and sharing. The entire CMS research group at this collaboration is broken into three tiers. One facility is considered “tier 0” where all of the raw experimental data is stored forever. No processed data is stored at tier 0. The raw data is then copied, divided and distributed to the “tier 1” sites to be processed. The 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 tier 1 sites are located all around the world. Next, the processed data is copied, divided, and distributed to all of the “tier 2” sites. The MIT lab is considered a tier 2 site. The tier 2 sites will only receive data from one tier 1 site unless they contact a different site and request their data.

Since it is at the bottom of the tier structure, there is no permanent storage at MIT. Instead, the university provides space for users to analyze the portions of the CMS experimental data that they receive from their tier 1 site. There are approximately 500 users from MIT and many other local institutions that have space here at MIT. Each user is provided with 1 terabyte of storage. Their data is replicated in a RAID array but not backed up. Users will usually request about 100 terabytes of data from the tier 1 site at a time and then filter down to about 1 terabyte before beginning analysis.

The scientist working the gravitational-wave observatory’s lab is also considered a “tier 2” site in the observatory’s data storage and analysis structure, however more data is stored locally at MIT then was true of the CMS data. This scientist has 10 RAID arrays with about 2-3 terabytes of storage each (i.e. a total storage capacity of about 300 terabytes). Additionally, members in the lab have local drives with about 200-300 gigabytes of short-term storage to use during their data analysis. Since it is considered a tier 2 site, none of the raw data used by the lab is backed up. However, the RAID arrays do provide redundancy. The main raw data backup is located on the West Coast at a “tier 1” site. The gravitational-wave observatory’s community is involved in all data retention decisions. The collaboration has established committees that oversee how data analysis is done at each site, and decides what data to delete if more storage is needed. Additionally, the data derived by this scientist and his colleagues at MIT (i.e. the output of the analysis conducted at his lab) is archived on tapes.

Data Sharing and Reuse

As the examples in the previous section shows, data sharing among physicists is very common, and often essential. However the extent of data sharing differs depending on the research group. For example, two of the scientists only share data within their research centers and rarely share raw data with scientists outside their collaboration. Another scientist's group has signed 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. For example, one scientist explained that he could re-use the same raw data hundreds of times and often re-integrates old data into new analyses. Since the raw data can be used for such a long period of time, the tiered data storing structure is extremely useful, allowing researchers in smaller labs to have access to the raw data without having to permanently store it.

Key Trends and Indicators for Data Growth

While each physicist at MIT has distinct research goals and methods for dealing with data, we have identified several trends and indicators for data growth.

7. The majority of the faculty members in MIT's Physics Department are experimental physicists. These physicists are often affiliated with large international or inter-institutional research centers and perform their experiments off campus.

8. Based on interviews with faculty members in this department, MIT-affiliated research laboratories and centers are currently generating data at a rate of 1,900,000 terabytes of data each year.

9. While improvements in experimental instruments can cause a jump in data production every 5-10 years, these instruments take years to develop and are not replaced often. Instead, the steady increase in data generation over time can be attributed to faster computing power and cheaper data storage hardware, which allow researchers to run their instruments for longer periods of time and to perform better data analysis.

10. The rate of data generation has been steadily increasing over time. Experimental physicists at MIT are currently producing about 5-10 times more data than they were five years ago.

11. Based on historical data and predicted improvements in computing power and storage, the rate of data generation for MIT physicists in 2014 will be approximately 11,400,000 terabytes per year.

12. While metadata is a critical to understanding the experimental conditions, the amount of metadata produced and stored is insignificant when compared to the amount of raw experimental data generated.

13. Many of the MIT-affiliated labs and centers use tiered data storage and sharing structure where all of the raw data is stored permanently at the "tier 0" site, and then divided among other tiers for redundancy and analysis.

14. Data sharing among physicists working at the same laboratory is common; however sharing among scientists at different labs is rare.

MIT physicists commonly re-use raw data from the same experiment multiple times and often re-integrated into new analyses as they are developed.

About the HMI? Program

The How Much Information? (HMI?) research program is a multi-discipline, multi-university project, formed to investigate the nature of data and information generated and used by individuals and enterprises. The program is sponsored by seven companies, including AT&T, Cisco, IBM, Intel, LSI, Oracle, and Seagate, and involves multiple research universities. The Principal Investigator is Prof. Roger Bohn and the Research Director is Dr. James Short, at UC San Diego's Global Information Industry Center (http://giic.ucsd.edu). Founded in 1960, the University of California, San Diego is one of the nation's most accomplished research universities, widely acknowledged for its local impact, national influence and global reach.

Acknowledgements

This case study is the product of industry and university collaboration in applied research. We are grateful for the support of our industry partners, sponsor liaisons, university research partners, and administrative staff at the University of California, San Diego.

Financial support for HMI? research and the Global Information Industry Center is gratefully acknowledged. Our sponsors are:

AT&T

Cisco Systems

IBM

Intel

LSI

Oracle

Seagate Technology

Additional support was provided by the Alfred P. Sloan Foundation of New York.

Questions about this research may be addressed to the Global Information Industry Center at the School of International Relations and Pacific Studies, UC San Diego:

Roger Bohn, Principal Investigator, rbohn@ucsd.edu

Jim Short, Research Director, jshort@ucsd.edu

Pepper Lane, Program Coordinator, pelane@ucsd.edu