



# Global Information Industry Center

at the School of International Relations and Pacific Studies

University of California, San Diego

## Neuroimaging at the Martinos Imaging Center The Scientific Data Flood: A Case Study of “How Much Information?”

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

### **Abstract:**

This case study provides a detailed look into the Martinos Imaging Center, a collaborative effort among the Harvard-MIT Division of Health Sciences and Technology, the McGovern Institute for Brain Research, Massachusetts General Hospital, and Harvard Medical School. Researchers at the Martinos Center study the human brain in three interrelated areas: perception, cognition and action. They use different imaging technologies, the main one is Magnetic Resonance Imaging (MRI), to study different aspects of the brain. Each MRI session produces a total of 3.6 gigabytes of human image data per subject. The Center sees 1500 subjects a year, generating approximately 5.4 terabytes of data. For all practical purposes, all of this data is saved indefinitely, as MRI scans are expensive, time consuming, and almost impossible to identically reproduce as the same subjects cannot be used again. The Center’s current rate of data generation will increase as scanner hardware and software improves. It is anticipated that improvements over the next five years will increase the size of subject data sessions by a factor of 10 (3.6 gigabytes of data currently, 36 gigabytes of data per session in five years). The case concludes with observations on why there is a lack of data sharing in the field at present, and references some embryonic efforts to develop network platforms for sharing neuroimaging data. Other papers examine other labs at MIT.

- Executive Briefing:** a summary of one or more research projects with preliminary findings for a non-academic audience.
- Research Report:** a completed report drawing on one of more research projects that presents study data, findings and management implications.
- Case Study:** an in-depth description of a firm’s approach to an information management issue.
- Research Article:** an academic research paper with sections on hypotheses tested, methods and data, analysis, findings and references.

**Contents**

<b>Background .....</b>	<b>3</b>
<b>Data Generation .....</b>	<b>3</b>
<b>Metadata .....</b>	<b>7</b>
<b>Data Retention.....</b>	<b>8</b>
<b>Data Sharing and Reuse .....</b>	<b>9</b>
<b>Key Trends and Indicators for Data Growth .....</b>	<b>10</b>
<b>About the HMI? Program .....</b>	<b>11</b>
<b>Acknowledgements .....</b>	<b>11</b>

## Neuroimaging at the Martinos Imaging Center

### The Scientific Data Flood: A Case Study of “How Much Information?”

Stuart Madnick, MacKenzie Smith, and Kate Clopeck, MIT

June 2009

#### 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<sup>1</sup>.

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

---

1 Novelline, Robert. *Squire's Fundamentals of Radiology*. Harvard University Press. 5th edition. 1997. ISBN 0674833392.

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 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

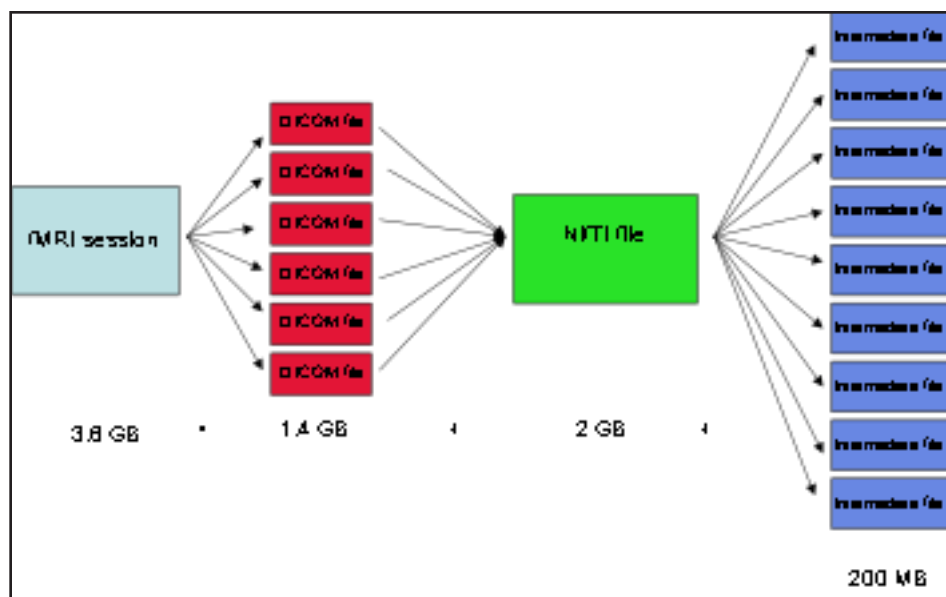


Figure 1: Files produced by 1 fMRI session, with one monolithic 4D NIFTI file

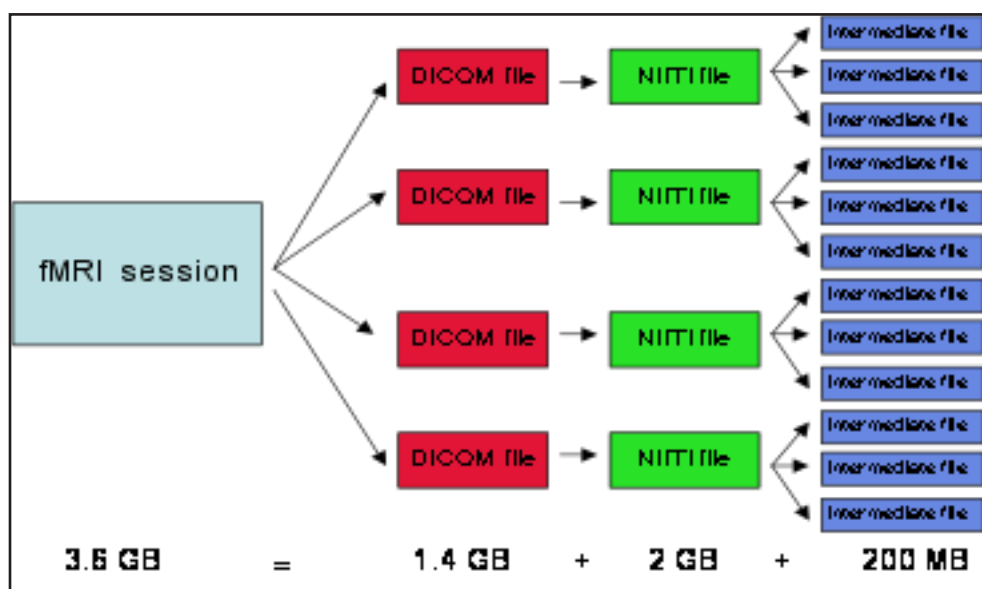


Figure 2: Files produced by 1 fMRI session, with one to one DICOM to NiftI mapping

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 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 Imaging 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 File 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 scalable 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 backup. 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.

## 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](mailto:rbohn@ucsd.edu)

Jim Short, Research Director, [jshort@ucsd.edu](mailto:jshort@ucsd.edu)

Pepper Lane, Program Coordinator, [pelane@ucsd.edu](mailto:pelane@ucsd.edu)