OPINION

# Databasing fMRI studies — towards a 'discovery science' of brain function

*John D. Van Horn and Michael S. Gazzaniga*

Enormous progress has been made over the past decade in the development of neuroimaging technology to study *in vivo* brain function. But as was once the case in genomics, much of the raw functional imaging data that are collected and described in the literature have not been made available to other researchers. The fMRI Data Center aims to archive raw functional imaging data from peer-reviewed publications, making it freely available to researchers from all disciplines to confirm conclusions, test new methods and generate new hypotheses. This bold project hopes to open up new vistas of understanding of complex cognitive processes and usher in the study of 'neuronomics'.

Everybody agrees that GenBank (from The National Center for Biotechnology Information) is a great success. It has become the *de facto* clearing-house for one-stop shopping for genetic information — the central source for genomics, the study of genes in action. The resource archives information on the genomes of assorted organisms from numerous viewpoints, including data on genomic sequence, on expression, on disease, and on taxonomy, and these data are linked to the relevant literature. GenBank has already begun to help catapult our understanding of human genetics by allowing public access to this information for its use in research, in clinical applications[1] and in education[2]. Needless to say, neuroscientists should want, and do need, a similar service to make progress towards the larger goal of understanding the brain and its various functions.

GenBank has also ushered into biology a new way of doing research. Genome informatics is not like other scientific catalogues — for example, the *Handbook of Chemistry and Physics*[3]. Although a huge amount of data has been assembled, the semantic content of the information is still largely unknown. But it will be through the evolution of data-mining tools, and through homology searches, the identification of coding regions and genes, and so on, that information about gene function will be discovered[4]. This has led some biologists, such as Leroy Hood, to suggest that certain fields are moving towards a model of 'discovery science'. As Hood has said, "It's the idea that you take an object and you define all its elements and you create a database of information quite independent of the more conventional hypothesis-driven view."[5]

It is against this backdrop that we have launched an effort to create a database for functional magnetic resonance imaging (fMRI) studies. The first order of business was to construct a framework that would allow scientists easy access to raw data from published, peer-reviewed studies. This aspect of the project is complete, and the fMRI Data Center (fMRIDC) is up and running. The second objective — one that is both challenging and exciting — is to enter the raw data into a database that will allow for the mining of highly heterogeneous and voluminous fMRI data. This 'data-mining' goal involves the use of information-retrieval methodologies that can sift through the huge amount of information that is contained in the MR images to give efficient and interesting responses to queries such as "find all study data that are close to the following study". Such searches, conducted across a broad variety of fMRI study data, might reveal hidden patterns of brain function that might have gone unseen in any one particular study. So, just as geneticists have gone beyond simply considering gene sequence data and are now consumed with the intricacies of genomics, neuroscientists must transcend their fascination with static brain images of basic mental functions and migrate into the study of 'neuronomics' — the examination of the 'complete', dynamic brain and the spatiotemporal interactions of its many systems. Database projects such as the fMRI Data Center will assist in reaching this ultimate goal of brain research.

We hope to benefit from earlier databasing efforts, such as the GenBank experience, and make accessing information on brain imaging as seamless as possible. Furthermore, with our current effort we are clearly dealing with one, small aspect of neuroscientific research. Accordingly, we are compelled to construct our system so that, in the future, it can exchange information with other database efforts[6,7] or be plugged into a larger, perhaps single resource for all of neuroscience. We review our progress in the following sections.

## Progress to date and current usage

Several challenges are faced when building a data-archiving project of the magnitude of the fMRI Data Center. We have discussed these particular issues in greater detail elsewhere[8], but several key problems are worth stressing, given the attention that has recently been given to them[9,10].

*Subject anonymity and protection.* In compliance with US federal regulation 45 CFR 46 on human subject protection, we had to address the requirement that any and all information that can be used to identify a subject must be removed from the data, while maintaining its experimental integrity. This is accomplished in a two-pronged effort. First, contributing authors remove identifiers before submitting study information to the Data Center. Then the Data Center checks for identifiers that might have been missed by the authors, while removing other potential identifying subject descriptors. Beyond the obvious need to preserve the anonymity of subject data, neuroimaging data involve the additional consideration that reconstructed, high-resolution images of the head can, in principle, be used as a subject identifier. That is to say, the structural data can be reconstructed in three dimensions, a surface be fit to the data and the contour of the subject's face be revealed. To remove this possibility, the high-resolution images are stripped of facial features, thereby removing the possibility that the identity of the subject be directly determined by three-dimensional anatomical reconstruction. Authors from outside the United States are encouraged to contribute their functional imaging data to this publicly available repository only after carefully considering their country's policy on the sharing of data from human subjects.

*Quality control.* Ensuring the quality of the contents of this neuroscience resource involves active communication between the Data Center personnel and the authors of the study. Authors themselves enter most information concerning their own study, thereby avoiding the possibility of misrepresentation by the Data Center. Data Center personnel crosscheck with the authors any questions about appropriate parameter values that have been entered by the

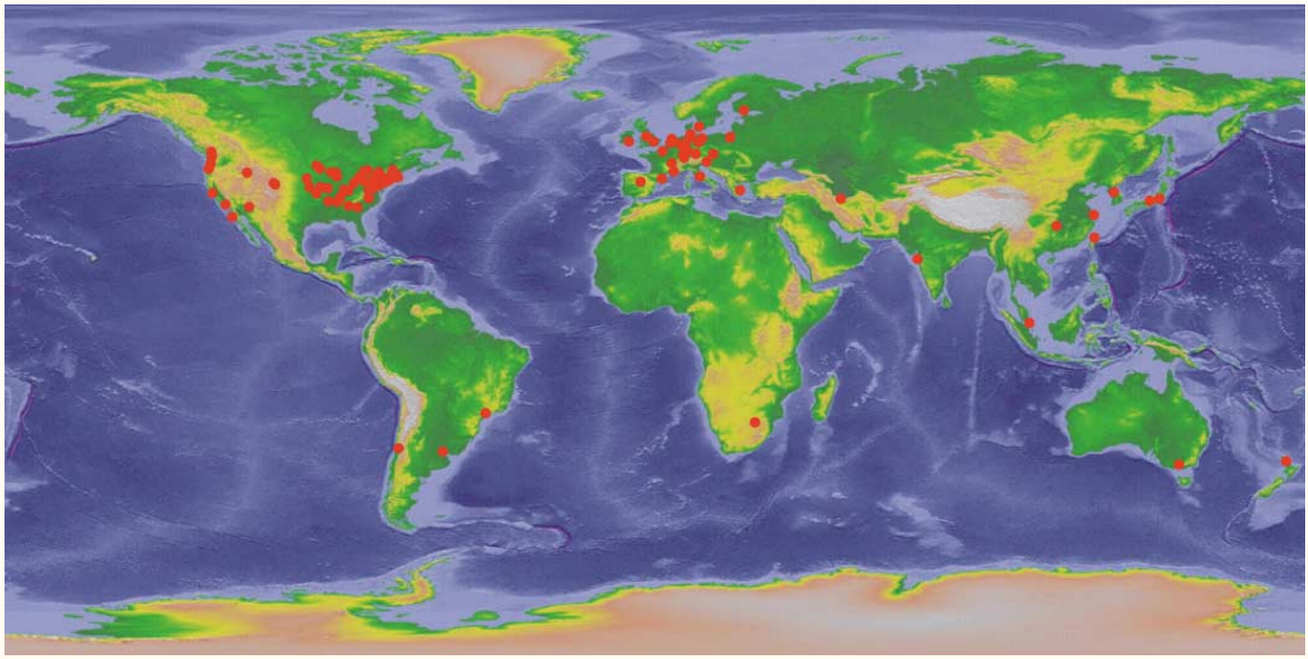

Figure 1 | **Response to the fMRI Data Center.** Since January 2001, the functional magnetic resonance imaging (fMRI) Data Center has fulfilled more than 300 requests from researchers in more than 25 countries around the world for data sets of published fMRI studies that are available through its web site. This impressive response within a relatively short space of time is one of several 'figures of merit' for measuring the success of the Data Center effort. This response indicates the enthusiasm of the international scientific community for the availability of raw published data sets for the purposes of secondary examination and confirmation of results, as well as for education. Researchers visit the Data Center web site and use a MEDLINE-inspired search interface to inspect the holdings of the fMRI study data archive. Studies of interest to the researcher can be selected and then requested. Compact disks with the raw, pre-processed and statistical image data, along with a document describing all aspects of the subject, MR scanner and experimental session protocols, are provided to the requesting researchers free of charge. Information-retrieval technology is being developed to cluster and sort studies of interest to researchers on the basis of predefined and user-defined search parameters. We anticipate even greater use of the Data Center as more studies become available.

authors. Documents that describe the study protocols are provided to authors for comment before the study becomes officially released to the public. Only the authors of the study have the power to change the description document after it has been officially released. Even then, the original information that is contained in the document is preserved in the form of annotations so that others can see how the information was updated.

*Author rights and researcher responsibilities.* Two further issues that have been of particular concern to the neuroscience community are the rights of the original authors and the responsibilities of researchers that use previously published data[11]. First, to allow contributing authors more time to carry out further analyses and to prepare subsequent manuscripts on the same data, an optional 'data hold' can be requested from the Data Center. This hold is effective from the date of publication, which should give the authors ample time for further analyses.

Researchers that receive study data from the Data Center are, at a minimum, expected to cite the work of the original authors and the database accession number in any newly published work. Clearly, publishing new findings from previously published study data without proper referencing of the original work is on par with plagiarism and would not be tolerated by the scientific community. Authorship invitations from the authors of any new research article to those of the original work is an interesting idea for ensuring that appropriate credit is given, but it might be difficult for the community to agree on how this should best be put into practice. We expect that the neuroscientific community will monitor these issues, ensuring that the data are used in an acceptable manner and that the appropriate credit is given to original authors.

*Use of the fMRIDC archive.* The fMRI Data Center has, so far, experienced a solid response from the international scientific community: more than 300 data requests have been fulfilled from researchers in more than 25 countries around the world (FIG. 1). As of 18 February 2002, a total of 54 fMRI studies have been contributed to the archive, with 21 data sets contributed in 2001 alone. The current archive contains study data from researchers in 11 countries and from four separate peer-reviewed journals. At present, only a single journal — the *Journal of Cognitive Neuroscience* — makes the contribution of fMRI study data a prerequisite for publication, although several other journals are encouraging their authors to do so. Sixteen complete studies are now available for immediate shipping at no cost to the interested researchers. Several more are under data hold, to be made available six months after the date of publication of the original author's research article. The total archive now contains more than 300,000 image files, requiring over 300 GB of disk storage. Encouraged by these numbers, which have been generated in just the brief time that the Data Center has been open, we anticipate a continuing increase in the number of scientists that use the Data Center as a resource for accessing complete fMRI studies.

*Content of the data archive.* Essential to the breadth and depth of any successful data archive of this type is not only the inclusion of the data as it was presented in its published and heavily processed form, but also the deposition of the actual raw imaging data (FIG. 2). With its inclusion, other researchers could reconstruct the data-processing stream, confirm the

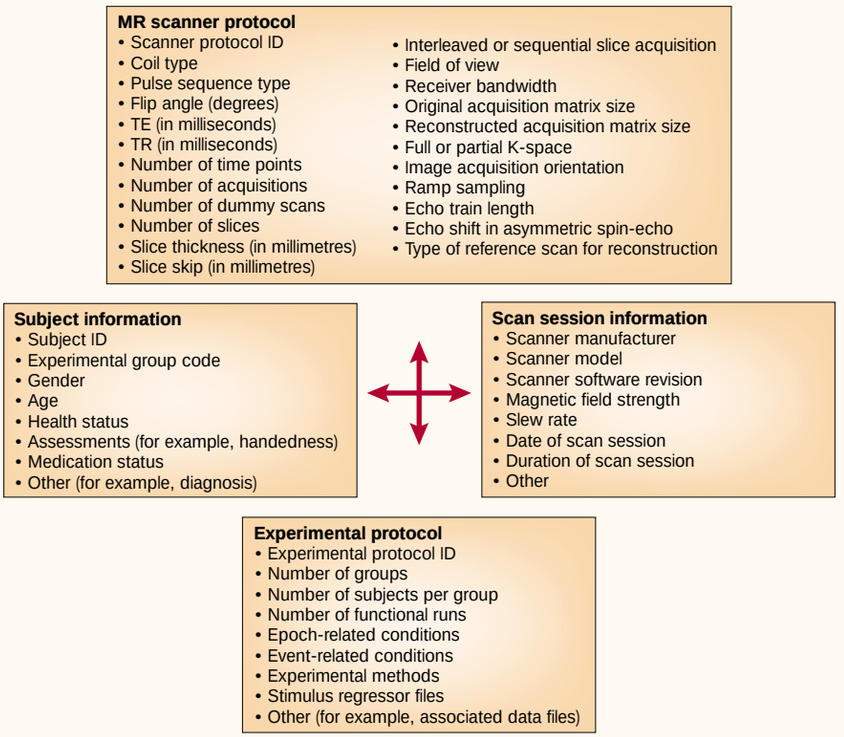


Figure 2 | **Basic fMRI study information collected for the fMRI Data Center archive.** Functional magnetic resonance imaging (fMRI) studies of the brain are associated with numerous study variables concerning the experimental paradigm, the parameters of the MRI scanner, variable settings during the functional and structural scan sessions, as well as information pertaining to the individiual subjects. The 'meta-data' descriptors from these protocols are often missing or incompletely presented in the original published article. The fMRI Data Center asks authors to supply complete information for the above variables so that they can be provided to researchers who request the raw study data. These data serve to describe each study as completely as possible, facilitating future 'meta-' and 'mega-' analyses of the raw study image data. TE, echo time; TR, repetition time.

findings of the original authors, or investigate new methods and draw their own, possibly different, conclusions. Authors are asked to provide the raw fMRI time series and high-resolution anatomical images after reconstruction from the MR scanner. Raw imaging data — the spectrum of frequences that constitute the MRI signal — are recorded in 'K-space'; the final MRI image is an inverse Fourier transform of K-space data. However, due to proprietary issues with scanner manufacturers about the conversion of K-space data into image data, the Data Center is not able to accept raw K-space images. Authors are also asked to provide the version of the imaging data, after preprocessing, at the stage just before being subjected to statistical analysis, as well as all statistical output images. Armed with the details of the original author's data-processing stream, researchers that obtain the data should be able to recreate the same processing stream, carry out the reported statistical comparisons, and arrive at the same quantitative answers as the original researchers. More importantly, researchers should be able to examine different data pre-processing paths and observe the effects of different processing choices on the interpretation of the results (for instance, the use of an alternative image-registration routine). All in all, the principal motivation is to store as much data from the study as possible.

"… neuroscientists must transcend their fascination with static brain images of basic mental functions and migrate into the study of 'neuronomics' — the examination of the 'complete', dynamic brain and the spatiotemporal interactions of its many systems."

*Database organization for efficient mining.* Unlike in genomics, in which data acquisition is the product of a more process-oriented approach, studies of cognitive function involve a range of sensory modalities, stimulus types and subject responses. Also, for as many different fMRI researchers as there are in the field, it seems that there are different scanner protocols and subsequent data-processing streams. Related to this issue are the many image file formats that researchers prefer for storing their image data. It is with this array of approaches to fMRI experimentation that the Data Center has adopted several guiding tenets in the organization of its core database. First, the database should be able and flexible enough to represent the broadest range of possible fMRI experimental paradigms. Second, the database is organized hierarchically, with the study itself at the highest level. Third, in addition to the high-level descriptive data of the study, meta-data characterizations for, and pointers to, all neuroimaging data and time series are represented in the database to facilitate the broadest possible space over which accurate but efficient searches can be made. And fourth, the database should be extensible; that is, it should be able to incorporate new studies, scans or time-course information, as it becomes available.

With these guidelines in place, we are constructing the fMRI Data Center database to accommodate the wide range of study types, and are creating sets of information-retrieval tools for pattern matching and clustering on the basis of the study meta-data or image data[12] (see below). Meta-analyses and examination of study parameters that influence the effects that are reported in structural as well as functional studies[13,14] will follow naturally. Proximity searches and meta-analyses will lead researchers to conduct 'mega-analyses', in which many subjects from separate studies can be analysed *en masse*; the sample size will be many times larger than is typical of any single, published fMRI study. By archiving both complete study descriptive data and raw functional imaging data, we believe that this approach goes well beyond the storing of only the reported statistical local maxima.

*New approaches for meta-data specification.* The advantage of maintaining a publicly available archive of raw functional data over one that is based simply on recording statistical local maxima is the ability to submit the study data themselves to clustering and meta-analysis. Accompanying this issue is the most fundamental challenge in dealing with the

sheer volume of data in these studies: creating intelligent and informative summary identifiers that facilitate accurate and efficient database searches.

Work is now underway to collect sets of descriptive and statistically based meta-data that characterize not only the individual images that make up an fMRI time series but, more importantly, the entire raw fMRI time courses. Rather than re-analysing the data in the same manner as in the original study, these measures involve consideration of both the temporal and frequency domains of the image time courses, with a view to summarizing as much information as possible with the fewest number of parameters. It is the collection of these common measures across fMRI studies that will allow useful searches of data in the archive.

Examples of such metrics include simple mean and variance images as measures of the distribution of the data, as well as voxel-based level-crossing (the number of signal excursions across some threshold) and frequency-spectrum measures that characterize signal oscillatory behaviour and periodicity. In particular, the last two measures capture the dynamic nature of the fMRI time course, but do not include information about the experimental paradigm under investigation. However, these measures will be sensitive to task-related modulation of the fMRI signal and show spatial patterns that are similar to those obtained through more traditional statistical tests (FIG. 3). These points are important, as we are interested in summarizing all of the data, not just those at voxel locations that the authors of the original research article chose to report. There might be additional aspects of the image data that are of interest to other researchers. In other words, it is essential that the data be allowed to speak for themselves. Once obtained, the information in these summary images can be represented in bit-wise fashion, further compressed, and regularities be identified, allowing inspection of the 'distance' between data sets to be carried out simply and efficiently.

But no single descriptive statistical measure will ever completely characterize the signal dynamics of data from all fMRI paradigms. Having several key measures of the signal parameters for the blood oxygen level dependent (BOLD) fMRI time courses in a study will allow the measurement of data similarity and permit further data-mining approaches. This, in turn, will lead to more precise and statistically more powerful cross-study re-analyses, potentially revealing patterns of brain activity that were not reported in the original research articles.

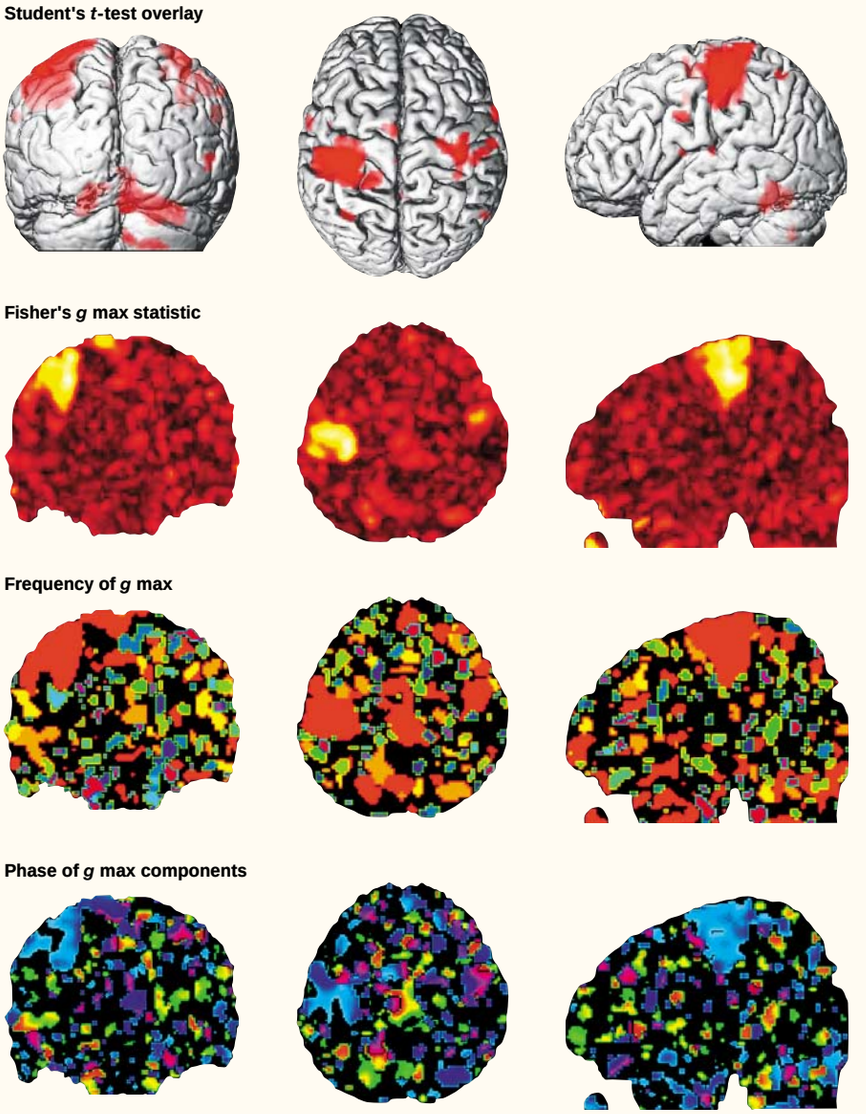


Figure 3 | **Descriptive statistical images.** An example of several descriptive statistical images that help to summarize the dynamics of functional magnetic resonance imaging (fMRI) time-series data. For this, we used echo planar image (EPI) data from a single male subject, collected by a GE Advance 1.5 Tesla scanner (repetition time (TR) = 2.0 s, echo time (TE) = 500 ms, field of view = 24 cm, slices = 27). The subject performed an epoch-related motor task in which he rotated an object in his hand for fifteen seconds followed by fifteen seconds of rest and so on for four-and-a-half minutes. The top row shows a traditional overlay map of the Student's *t*-statistic, which is typically used in fMRI analysis, thresholded at $P<0.001$. Here, the test statistic is determined by contrasting periods of motor task activity with periods of rest using the Statistical Parametric Mapping software package (Wellcome Department of Cognitive Neurology, London, UK). The second row depicts a map of Fisher's *g* max statistic at each voxel of the same motor task data. This measure is computed as the peak frequency magnitude standardized over total signal power. In this image, brighter-coloured areas represent a high proportion of signal power contained at a single spectral frequency. Related to this, the third and fourth rows show the frequency at which the maximum value of *g* was obtained (from 0 to $2\pi$) and the phase offset of the frequency values (from $-\pi$ to $\pi$), respectively. In the former, false-coloured red areas pertain to regions in which frequency was comparatively low and darker-coloured areas to regions in which frequencies were relatively high. In the latter, blue areas indicate the phase closest to zero, with green being negative and red being positive. Note that only the *t*-test image procedure involves explicitly including information that pertains to the paradigm being investigated. It is clear that the motor and premotor brain areas that are shown to be statistically significant by the *t*-test analysis are the same as those indicated as having a high percentage of the signal power in the raw time series, and is also reflected in frequency component and relative phase. These three, purely descriptive measures, together with others, such as level crossings and the higher-order distributional moments, can be used to summarize and cluster data across studies in order, for instance, to identify interesting features about the physiological basis of motor function.

## Current challenges

Indexing the data from these rich studies into a coherent database framework is clearly an enormous undertaking. Through the guidelines that we discussed above, indexing must occur at several levels. Effort is now underway on the intelligent cataloguing of study, individual-image and time-series levels of information in the database. Moreover, we look forward to developing further image meta-data measures, as well as receiving feedback from the community to which meta-data characterizations are the most appropriate, meaningful and successful in identifying interesting study clusters.

A further challenge is to deal with the considerable size of the data archive. For instance, the typical size of an fMRI study that is contributed to the Data Center has, so far, been in the order of 3.26 GB. But there are several examples that exceed 15 GB of image data. At the current rate that the Data Center is receiving study data, we expect to fill one terabyte of disk space over the coming year and to double these holdings every nine months. It is unrealistic, however, to expect that the trend in the size of neuroimaging studies will ever be towards smaller data sets. In fact, stronger field magnets, higher image resolution and other advances in scanner technology are already on the horizon, and will undoubtedly result in the acquisition of still larger data sets. Simple calculations show that the amount of storage space that is required for archiving published functional imaging experimental data in the not-too-distant future will be well into the petabyte range. Clearly, the data storage issue associated with archiving functional neuroimaging data is a serious one, not to mention the computational challenges of attempting to carry out analyses on such an archive.

To address these issues, we are now investigating means of 'near-line' and off-line data storage, as well as means of providing more suitable computing resources to researchers that are interested in carrying out large-scale analyses. But this does not address the question of how best to manage and provide access to data stored off-line. It is a considerable problem for any large data archive, particularly for one that maintains neuroimaging data. This problem will require the examination of existing hierarchical approaches to data storage[15], as well as the examination of new methods for large-scale archive management that are under consideration in other fields[16]. The Data Center is even now liaising with computer science experts to address this issue. They and others are excited about the research challenge that this enormous data-management problem poses and hope to develop methods that have a scope beyond the realm of just the storage of neuroscientific data.

## Conclusions

Throughout the past decade, an explosion of research on *in vivo* human brain function and in computer information technology has turned into a reality what once was considered an untenable idea — the creation of a database for raw fMRI time-series data. Just as the sequencing of genetic information has given way to the excitement of functional genomics and proteomics[17], we are quickly approaching a time when the localizationist approach to brain mapping will give way to the study of evolving, plastic, whole-brain neural systems. The archiving and databasing of raw fMRI data will help lead to a time when such study is possible — extending hypothesis-based study into discovery-based science of the brain.

But the archiving of neuroimaging data is a small part of a larger community that is dedicated to indexing neuroscientific data from numerous other experimental modalities on other spatiotemporal scales[18–20]. The fMRI Data Center effort can provide insights into the BOLD response to cognitive tasks, but will not be able to shed light on, for example, neuronal spike-train patterns. So, to move towards our goal of understanding the brain, raw data archives of these other, rich forms of data will be required and will need to be linked to allow understanding at multiple levels of granularity. However, the pursuit of this multifaceted form of research will be possible only through data sharing, the use of neuroinformatics, databasing and information-retrieval technologies[21]. We look forward to working with those who build such data archives to create a web of neuroscientific knowledge that might help us to attain a more complete understanding of how the brain works.

Time will tell when the fMRI Data Center database will attain its full potential. When it does, we expect it to become an essential resource for researchers and educators alike, changing the way neuroscience is conducted.

*John D. Van Horn and Michael S. Gazzaniga are at the Center for Cognitive Neuroscience, Dartmouth College, 6162 Moore Hall, Hanover, New Hampshire 03755, USA.*
*Correspondence to J.D.V.H.*
*e-mail: John.D.Van.Horn@dartmouth.edu*

DOI: 10.1038/nrn788

1. Collins, F. S. & McKusick, V. A. Implications of the Human Genome Project for medical science. *JAMA* **285**, 540–544 (2001).
2. Magee, J., Gordon, J. I. & Whelan, A. Bringing the human genome and the revolution in bioinformatics to the medical school classroom: a case report from Washington University School of Medicine. *Acad. Med.* **76**, 852–855 (2001).
3. Lind, D. R. (ed.) *CRC Handbook of Chemistry and Physics* (CRC, New York, 2001).
4. Collins, F. S. The Human Genome Project and the future of medicine. *Ann. NY Acad. Sci.* **882**, 42–55; discussion 56–65 (1999).
5. Interview with L. Hood in *Technology Review* [online] (cited 18 Feb 2002) <http://www.technologyreview.com/articles/qao901.asp> (2001).
6. Mazziotta, J. C., Toga, A. W., Evans, A. C., Fox, P. T. & Lancaster, J. L. Digital brain atlases. *Trends Neurosci.* **18**, 210–211 (1995).
7. Toga, A. W. & Thompson, P. M. Maps of the brain. *Anat. Rec.* **265**, 37–53 (2001).
8. Van Horn, J. D. *et al.* The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. *Phil. Trans. R. Soc. Lond. B* **356**, 1323–1339 (2001).
9. Governing Council of the Organization for Human Brain Mapping. Neuroimaging databases. *Science* **292**, 1673–1676 (2001).
10. Chicurel, M. Databasing the brain. *Nature* **406**, 822–825 (2000).
11. Editorial. Whose scans are they, anyway? *Nature* **406**, 443 (2000).
12. Goutte, C., Hansen, L. K., Liptrot, M. G. & Rostrup, E. Feature-space clustering for fMRI meta-analysis. *Hum. Brain Mapp.* **13**, 165–183 (2001).
13. Van Horn, J. D. & McManus, I. C. Ventricular enlargement in schizophrenia. A meta-analysis of studies of the ventricle:brain ratio (VBR). *Br. J. Psychiatry* **160**, 687–697 (1992).
14. Hopfinger, J. B., Buchel, C., Holmes, A. P. & Friston, K. J. A study of analysis parameters that influence the sensitivity of event-related fMRI analyses. *Neuroimage* **11**, 326–333 (2000).
15. Ghandeharizadeh, S., Ierari, D. J. & Zimmerman, R. in *Computing the Brain* (eds Arbib, M. A. & Grethe, J. S.) 265–284 (Academic, San Diego, 2001).
16. Szalay, A. & Gray, J. The world-wide telescope. *Science* **293**, 2037–2040 (2001).
17. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
18. Bloom, F. E. The multidimensional database and neuroinformatics requirements for molecular and cellular neuroscience. *Neuroimage* **4**, S12–S13 (1996).
19. Miller, P. L. *et al.* Integration of multidisciplinary sensory data: a pilot model of the human brain project approach. *J. Am. Med. Inform. Assoc.* **8**, 34–48 (2001).
20. Kotter, R. Neuroscience databases: tools for exploring brain structure–function relationships. *Phil. Trans. R. Soc. Lond. B* **356**, 1111–1120 (2001).
21. Koslow, S. H. Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neurosci.* **3**, 863–865 (2000).

## Acknowledgements

The fMRI Data Center represents the work of several outstanding personnel. We would like to recognize the contributions of D. Rockmore, J. Aslam, P. Kostelec, D. Rus, J. Grethe, J. Woodward, W. Starr and A. Tilden. We would also like to thank D. Smith, B. Donald, and S. Grafton for their helpful comments on this article. This work is supported by the National Science Foundation, the William M. Keck Foundation and the National Institute of Mental Health.

## Online links

FURTHER INFORMATION
Encyclopedia of Life Sciences: http://www.els.net/
bioinformatics | biological data centres | brain imaging: localization of brain functions | brain imaging: observing ongoing neural activity | ethics of research: protection of human subjects | magnetic resonance imaging | mining biological databases
fMRI Data Center: http://www.fmridc.org/
GenBank: http://www3.ncbi.nlm.nih.gov/Genbank/
MIT Encyclopedia of Cognitive Sciences: http://cognet.mit.edu/MITECS/
magnetic resonance imaging
US federal regulation 45 CFR 46: http://ohsr.od.nih.gov/mpa/45cfr46.php3
**Access to this interactive links box is free online.**