nature neuroscience

Sharing neuroimaging studies of human cognition

John Darrell Van Horn, Scott T Grafton, Daniel Rockmore & Michael S Gazzaniga

After more than a decade of collecting large neuroimaging datasets, neuroscientists are now working to archive these studies in publicly accessible databases. In particular, the fMRI Data Center (fMRIDC), a high-performance computing center managed by computer and brain scientists, seeks to catalogue and openly disseminate the data from published fMRI studies to the community. This repository enables experimental validation and allows researchers to combine and examine patterns of brain activity beyond that of any single study. As with some biological databases, early scientific, technical and sociological concerns hindered initial acceptance of the fMRIDC. However, with the continued growth of this and other neuroscience archives, researchers are recognizing the potential of such resources for identifying new knowledge about cognitive and neural activity. Thus, the field of neuroimaging is following the lead of biology and chemistry, mining its accumulating body of knowledge and moving toward a 'discovery science' of brain function.

The observation that changes in regional cerebral blood flow accompany neural activity during cognition[1–3] has been a boon to the cognitive and brain sciences, most notably through the use of brain mapping technologies such as functional magnetic resonance imaging (fMRI). Current research efforts for imaging the brain 'in action' are underway to rigorously explore the structure and function of cognitive brain processes, thereby characterizing the mental properties that make us uniquely human[4]. The fMRI studies range from the examination of familiar cognitive processes such as human memory and language processing to novel studies of racial threat[5] and the neurofunctional components of humor[6].

This increasing dependence on brain mapping for exploring cognition has led to an unprecedented data explosion that is pressing neuroscientists to manage and analyze data on scales never before imagined. Complete fMRI study data sets now routinely reach several gigabytes in size, with the amount of brain image data collected in some articles[7,8] beginning to rival the current size of many biological and physical science databases[9,10]. What is more, the size of fMRI studies has grown over time, and what is now considered a large fMRI study will seem relatively small within only a few years, as new technological developments occur in scanner physics, engineering and protocol design.

Unfortunately, despite this progress, much of these fMRI data are not readily available to anyone beyond the original research team that collected them. There are several reasons behind the fact that other investigators do not typically get to work with the actual data that went into the heavily processed images appearing in a published article: (i) limitations of publication space on the complete representation of fMRI methods and findings, (ii) the proprietary feelings of investigators against letting others view their data, (iii) the immensity of data set size and (iv) the convention of only reporting tabular representations of activity in individual image voxels. However, given recent success stories from genomics[11] and proteomics[12] for organizing, archiving and mining large amounts of data from their communities, it may come as no surprise that cognitive neuroscientists are now looking to unfettered data sharing and study archiving to better understand these rich collections of dynamic brain data.

Data sharing sociology in neuroimaging

In 2000, with precisely such a goal, we founded the fMRIDC (www.fmridc.org) at Dartmouth College. We sought to facilitate progress in understanding cognitive processes through the collection, archiving and open distribution of neuroimaging data sets in the peer-reviewed literature[13]. We reasoned that there could be several positive outcomes to making the complete study data sets available to others. First, the study findings could be independently confirmed, helping to strengthen the findings drawn by the original authors. Second, new statistical methodologies could be applied to the data, providing novel insights into cognitive processes. Different studies could be compared, possibly identifying unanticipated functional homologies between seemingly different cognitive tasks. Moreover, these studies could be used to train the next generation of neuroscientists by using fMRI data that had already undergone interpretation by those who collected it and had published it in leading journals. We decided to focus on fMRI data from published articles and not to be concerned with unpublished data. This allowed us to focus the enormous chore of collecting and managing the data, as well as to construct an archive that was representative of the field's body of work.

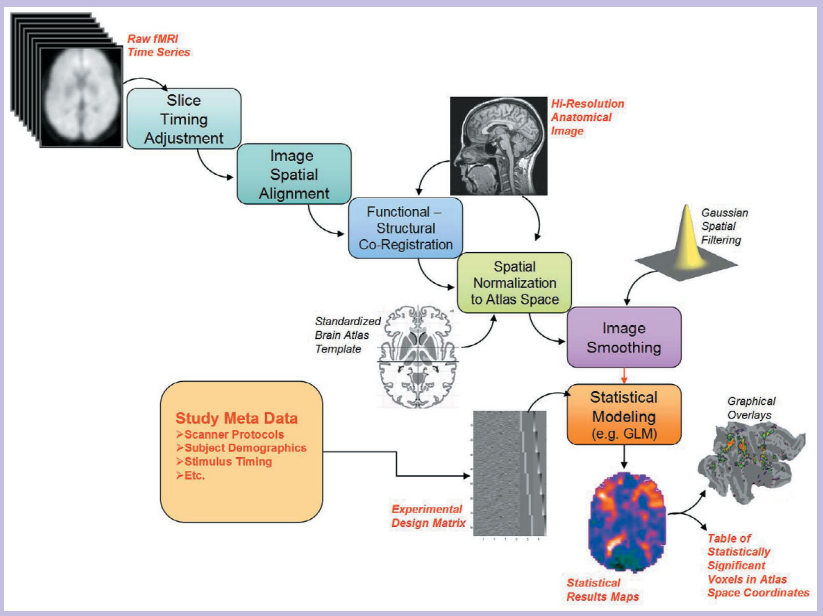
We approached the editors of several leading journals and were pleased by their initial support. To form the first corpus of data sets and accompanying study material, a special issue of the *Journal of Cognitive Neuroscience* (*JOCN*) was published containing a collection of articles from leading laboratories (Vol. 12, Suppl. 2, 2000). The authors of these articles generously provided the raw, processed and results image data along with structural images and study meta-data

All authors are at Dartmouth College, Hanover, New Hampshire 03755, USA. John Darrell Van Horn and Scott T. Grafton are at the Center for Cognitive Neuroscience, the Dartmouth Brain Imaging Center and the fMRI Data Center, Daniel Rockmore is in the Department of Mathematics and the fMRI Data Center, and Michael S. Gazzaniga is at the Center for Cognitive Neuroscience and the fMRI Data Center.
e-mail: John.D.Van.Horn@dartmouth.edu

Published online 27 April 2004; doi:10.1038/nn1231

BOX 1 fMRI data processing pipelines

Much in the way that a biological tissue sample is assayed to ultimately reveal genetic information, neuroimaging time-course data often undergo a number of digital processing operations in preparation for subsequent statistical analyses and the localization of brain activity[56]. This can be envisioned as an fMRI data processing pipeline (see illustration below). Data saved from the MRI scanner are first stored as a collection of two-dimensional (2D) slices in which the spatial phase and frequency of the brain image are represented—often referred to as *k*-space. This data is converted to a 3D image space using a Fourier transform to become what many refer to as the 'raw' image volume data, representing one volume of brain data for each time point in the series. Collected on the order of one volume every 2 s, these individual image time courses alone can grow to several megabytes or more of data. Collected across subjects, with one or more time course per subject, the raw data set for an entire study can grow to several gigabytes. This is typically the earliest form of the brain image data upon which investigators begin their analyses.



Despite efforts to minimize subject movement at the time of scanning, the raw image must be subjected to rigid-body spatial realignment routines to remove motion-related effects. Still other processing steps may be applied, including corrections for slice timing, spatial distortions, physiological effects, etc. The next step is averaged, aligned functional image co-registration with a high-resolution anatomical scan, followed by non-linear spatial image warping of the anatomical into a known stereotactic atlas space. This controls for anatomical variation in brain size and sulcal and gyral patterns across subjects and places the functional data into a common coordinate system. The spatial normalization process is ordinarily done with respect to brain atlases such as that of Talairach and Tournoux[57], which researchers then use to report the statistically significant locations by providing their coordinates as *x*, *y* and *z* values in this space with respect to an imaginary set of orthogonal planes passing through the anterior commissure (the small bundle of fibers connecting the cerebral hemispheres located behind and beneath the genu of the corpus callosum). Often, as a final step before statistical modeling, the resulting images are spatially filtered using a Gaussian smoothing kernel to reduce the effects of misregistration and increase the signal-to-noise ratio of the image time course.

The predominant method of statistical analysis is to use the General Linear Model (GLM) to obtain regression coefficients for predictor variables pertaining to stimulus presentations contained in an experimental design matrix[58]. These coefficients are assessed for statistical significance using voxel-wise Student's *t*-tests, and these are displayed as statistical parametric brain maps. Researchers will often represent regional activation using only the most statistically significant voxel ('volume element') in a cluster of activity and provide tabular summaries of these 'local maxima' in their published articles. So what starts out as gigabytes worth of image data is systematically processed, analyzed, and ultimately compressed for publication into a neat table and several overlay figures of task-related brain activity.

The fMRIDC seeks to capture data from key points in this process (indicated in red) in order to have as complete a set of information as possible so that an independent investigator may begin with the study's raw image data, follow all the same processing steps, and, using the same experimental meta-data, obtain the same statistical results as provided by the original study authors. Thus, this 'digital assay' may be reproduced, verified or altered by others to explore new methods and interpretations of the data.

(*i.e.*, data that describe data, such as subject demographics, scanner protocols and task information).

The neuroimaging and cognitive neuroscience communities responded cautiously. Researchers wondered whether new science was possible from archived fMRI data. Many young scientists were concerned that giving others access to their raw data might undermine their research programs and hurt their chances at career advancement. Established scientists were concerned that fMRI was too new and immature, without fully established methods, to warrant mandatory data sharing. Many worried that the fMRIDC would act as self-appointed data-police, passing judgment on a study's validity or credibility. Some felt that there was little chance that such an effort would acquire any support—or get anyone to contribute their data—and that the technological challenges to archiving data on this scale were insurmountable. As a result, those journal editors who had earlier expressed support, understandably decided to take a wait-and-see approach, holding off implementing any new data-sharing policies until the community had decided how best to proceed. As one of us (M.S.G.) was the editor of *JOCN* at the time, a decision was made to go it alone and see how far we could get.

Community resistance to data sharing is not new in science. Geneticists first blanched at the idea of sharing their DNA sequence data until the US federal government established the National Center for Biotechnology Information (NCBI) and the GenBank[14] archive in 1988. Likewise, the X-ray crystallographic community began considering databases when Richard J. Roberts, a 1993 Nobel

laureate, first urged the formation of a shared database for crystallography. Society newsletters from the mid-1990s show a series of vitriolic exchanges between those in favor and those against. The argument raged on until the governing bodies in the field recognized the benefits that could be gained through data-sharing and began requiring their membership to deposit structures in archives like the Protein Data Bank (PDB)[15]. The PDB effort, begun in the early 1970s by a small collection of like-minded scientists wishing to gather structure data for the purposes of modeling and visualization, is now recognized as the premier world resource for protein data. The journal *Science* decided to require its contributors in the field to deposit their genomic sequence and crystallographic coordinates in such public databases, but allowed the contributors to wait for a year after publication before meeting the requirement[16,17]. Although companies interested in commercially exploiting the data supported the delay, academic scientists wanted quicker access to the information. Now, *Science* and *Nature*[18], as well as *The Proceedings of the National Academy of Science* (*PNAS*)[19,20], require contributors in the field to deposit their material in these databases as a condition of publication. With the potential and promise of these databases becoming clear, the data-hold policies have been shortened or disappeared altogether.

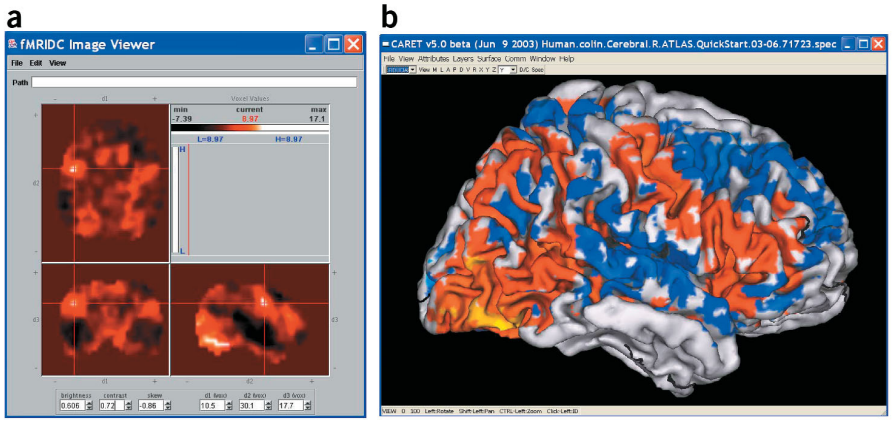


Figure 1 The use of archived neuroimaging data can lead to new findings from the original study data or, simply, novel ways of visualizing previously reported effects. For instance, a recent study[59] (fMRIDC Accession Number 2-2003-113HT) focused on the functional anatomy of ventral temporal cortex and prefrontal cortex during working memory. In the original published article, only two representative slices were shown in the figure from each region of interest. With online availability, the actual fMRI statistical map can be explored using web-enabled image viewers, and brain regions may be inspected for activation patterns not necessarily described by the original study authors—such as browsing the experiment's unthresholded random-effects Student's *t*-test map (**a**) for the main effects of working memory encoding compared against baseline activity. Data may then be requested or downloaded and subjected to novel forms of visualization or subsequent analyses using tools interoperable with the format of image data. For example, a user can view the same results image data from the study in (**a**), but represented on a cortical surface model using the Caret software package from Washington University (**b**). This interoperability of data and software tools to support the published literature is a valuable and much sought-after feature for neuroscience databases.

As the fMRIDC effort has progressed, much of the initial concern from the community about contributing data has significantly diminished, with researchers increasingly being supportive and enthusiastic about contributing their study data to the archive. After recently implementing an online submissions process, whereby authors must indicate their agreement with the policy on sharing data by checking a box in a web form, the *JOCN* has witnessed a doubling of imaging-related papers. Other neuroscience journals are now strongly encouraging authors to contribute their data to the archive[21]. We have also expanded our services beyond simple data curation toward developing fMRI data management software tools that make data sharing, exploration and interoperability with imaging data easier (**Fig. 1**). Finally, many of our neuroimaging colleagues have recognized the potential for how the archive can be used to test new hypotheses about cognitive function, where published study data might serve as valuable control data for their own fMRI studies, and how the data in the archive might be combined to perform large-scale, population-level fMRI analyses. These satisfying outcomes indicate that data archival efforts, like the fMRIDC, are fast becoming an essential resource for neuroscience researchers[22].

Which data should be shared?

One critical question in the development of a database for published fMRI studies has been the question of which data from image processing and analysis pipelines ought to be shared[23,24]. Should we bother with the raw data at all? Are the final results alone not good enough? Why start now when neuroimaging methods are still in their infancy?

Some authors have specifically encouraged "sharing the most valuable data type first"—arguing that tables of neuroimaging summary test statistics and accompanying Talairach atlas spatial coordinates alone possess the greatest scientific value[25]. This point of view maintains that neuroimaging data increase in worth with every step of mathematical and statistical processing. However, the amount of mathematical information contained in a processed data set about the original form of the data remains the same or is reduced by that processing[26]. So although this basic approach is aimed at being practical and parsimonious, sharing only lists of statistical maxima may not be suitable when the goal is to extract the information contained in these routinely large, multi-dimensional and highly structured data sets (extracting all the scientific worth from the data). For instance, researchers wishing to investigate alternative image processing methods or to apply novel time series analytic approaches do not benefit from having only such summary results. Simply collecting the local maxima provided in a published study is highly useful in a bibliographic sense, but by itself, provides only limited insight into the total information that is contained in the raw and processed forms of the image data.

Providing users access into the neuroimaging data processing pipeline itself seems to best satisfy the range of potential uses of archived data[27] (**Box 1**). The fMRIDC asks authors to provide the data from these several key points in their fMRI data processing chain (for example, raw, preprocessed and final-result images plus high-resolution anatomical volumes) along with an accurate description of how the data were processed. Given this information, anyone should be able to apply the identical data processing routines as described by the original study authors, in the same order, and obtain the same brain activity findings. On the other hand, some users may only be concerned with the statistical images that are the end result of an fMRI analysis, wishing to use them in meta-analyses or new forms of data

Table 1 fMRI Data Center archive summary[a]

Number of studies:	70 complete fMRI studies
Number of subjects across studies:	~1,000 individual subjects
Number of functional runs:	>5,000 time series runs
Individual brain image volumes:	>500,000 image volumes
Average fMRI study size:	6 Gigabytes
Maximum fMRI study size:	20 Gigabytes
Overall archive size:	~2.6 Terabytes
Total number of files being managed:	~22 million
Number of data set requests fulfilled:	1,239

[a]As of 3/21/2004

visualization. Other researchers may wish to apply statistical modeling approaches other than those used in the original article; they would therefore be most interested in the preprocessed image data (for example, after spatial realignment and normalization). Finally, still others may be interested in comparing and contrasting multiple study data sets against one another, which would require a uniform data processing pipeline across studies starting from raw data. Therefore, the fMRIDC archives the image data in its rawest form after image reconstruction, after the last stage of preprocessing and before statistical modeling and significance testing. The archive also includes statistical results and parameter images, as well as the accompanying high-resolution structural image data. Intermediate versions of the data are not needed, provided a suitable description of the data processing chain is also provided. This 'multiple entrypoint' approach offers the greatest accessibility across different uses in new science.

Study 'meta-data' are also important to consider because they accurately characterize the study protocols and experimental design matrix information. These include subject demographics, stimulus timing and scanning parameters. One approach to designing a database to store functional neuroimaging data is to use existing well-understood, relational, spatial and object-relational database technology. Several free (MySQL, PostGres) and commercially available (Oracle, DB2) solutions exist. However, for heterogeneous collections of studies from different laboratories, often using differing or new experimental techniques, these formats may be limited and unable to adapt readily as the field evolves.

In recent years, a new method for organizing scientific data has received growing attention[28]. These 'knowledge bases' are databases organized according to an ontology[29], rather than a database schema, and occupy a middle ground between very loose and very rigid data architectures. Ontologies describe the knowledge about a domain using declarative language structures (for instance, an MRI scanner "is a" measurement device; a brain image "has a" data-type). Definitions associate the names of entities in a particular domain (such as classes of items, relations between them, and their functions) with human-readable text describing what the names mean, and provide formal axioms that constrain the interpretation and well-formed use of these terms. Ontologies are advantageous in that they have been developed to handle and search over qualitative information as easily as the more traditional formats deal with quantitative information (for example, see The Cyc Public Ontology; www.cyc.com). These concepts have given rise to the notion of a Semantic Web as an emerging representation of data on the Internet in which information is given well-defined meanings, better enabling people and computers to work in cooperation[30].

These approaches are ideal for encapsulating data from often highly heterogeneous fMRI studies[31]. They can have the structure required for data sharing and reuse, while maintaining the flexibility required for accommodating variations from lab to lab, researcher to researcher, and as the field evolves. Because ontologies are a relatively new idea for scientific databasing, they are often deployed as an add-on to a traditional database, or as a kind of connective tissue to enable a limited data exchange between more rigid formats. This approach makes sense in a context of pre-existing data management tools that need merely to be interconnected. But, more importantly, ontologies can serve as the basis of a canonical representation—the primary format on which data updates are performed, the logical representation and the in-memory format around which tools and user interfaces are built—for the fMRI community. By merging the requirements for fMRI data management with those for neuroimaging data sharing and exchange, the fMRIDC has been making significant progress toward the construction of extensible fMRI study ontologies.

But why begin the task of fMRI databasing right now? Neuroimaging data processing methods are still being perfected, and procedures within the field are still in flux. Shouldn't the brain imaging community wait ten years until methods are better established?

Databasing efforts in genomics and proteomics, though underway for nearly 30 years (in the case of proteomics), have only in the last few years begun to be fully appreciated as the rich scientific resources that they are. Also, though some fields in neuroscience are more than 50 years old, many of the current databasing efforts are less than ten. For instance, researchers in the single- and multi-unit recording community have begun forming neuronal property databases[32] in order to continually collect and contrast novel approaches for looking at their data because computational simulations of neurons and their interactions are an ongoing and exciting research activity[33]. To miss this opportunity to archive valuable, published fMRI data during the rise of cognitive neuroimaging, and not capture key instances of the ever-growing body of work in this field would be unfortunate, indeed.

Financial incentives for sharing primary data

Several US funding agencies are now taking an active role in promoting data sharing, in particular for the neurosciences[34]. The National Institutes of Health (NIH; http://grants.nih.gov/grants/policy/data_sharing) and The National Science Foundation (NSF Social and Behavioral Sciences data sharing policy; www.nsf.gov/sbe/ses/common/archive.htm) have each recognized the benefits of sharing primary research data for advancing science and have implemented policies requiring data sharing. Research funding organizations in other countries, like Great Britain (UK Medical Research Council; www.mrc.ac.uk/strategy-data_sharing_policy), have also instituted scientific data sharing requirements. Each has noted that sharing data reduces unnecessary duplication of data collection, provides added value to research publications, and spreads the cost of doing research across a diverse set of investigators. In a field like neuroimaging research, where data collection can be expensive and study replication is infrequent, permitting researchers to access previously collected data has many economic advantages for funding bodies.

Justifying how researchers will share their data is now expected as part of proposals for research funding. Beginning with the October 1, 2003 receipt date, investigators submitting NIH grant applications seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or to state why data sharing is not possible (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html). In the case of neuroimaging, publicly accessible repositories like the fMRIDC provide a convenient means for researchers to satisfy such requirements.

Compared to positron emission tomographic (PET) scanning, which involves injection of radioactive isotopes, fMRI is reasonably inexpensive, which has been part of its appeal to brain scientists as a research tool. Nevertheless, the costs associated with studies of brain imaging do add up. Many fMRI centers charge investigators between $200 (US) and $700 per hour to use the scanner and its facilities. Additional expenses associated with subject reimbursement ($20–$100 per hour), staff salaries, study sample size, journal of publication and indirect costs all contribute to the final expense of what becomes an article published in a peer-reviewed journal. It is not impossible for a complete fMRI study to have a total cost in the hundreds of thousands of dollars by the time the article describing it appears in print. Formal sharing of data with the neuroscience community through open efforts like the fMRIDC allows a great many more researchers to use the data, thus mitigating its price tag. Secondary analyses of these image data can then be done by other researchers at a smaller cost than the original study. These can be used as a prelude to and justification for new, hypothesis-driven fMRI data collection. Undoubtedly, funding agencies in the US and elsewhere find these sorts of cost savings an attractive feature of data sharing and a better return on taxpayer investment.

Table 2 Notable scientific databases

Name	Year founded	Governing institution(s)/ funding	Current archive status	Notes
Protein Data Bank (www.rcsb.org/pdb/)	1971	Research Collaboratory for Structural Bioinformatics (RCSB; Rutgers, UCSD, NIST); NIGMS, NLM, NSF	Number of Protein Structures: 21,998 DB Size: **15 Gigabytes (GB)**	Expected to grow at a rate of about 30% per year.
GenBank (www.ncbi.nlm.nih.gov/)	1982	National Center for Biotechnology Information (NCBI) – National Library of Medicine (NLM)	Base pairs: 32,528,249,295 Sequences: 25,592,865 Current DB Size: **138 GB (as of 12/2003)**	Continuously updates gene assemblies, incorporates new data, fills in existing gaps, and increases overall accuracy. Released to the public on a regular basis. Holdings approximately double annually.
Visible Human Project (www.nlm.nih.gov/research/visible/visible_human.html/)	1986	Visible Human Project; NLM	High-resolution image sections of two human cadavers Current DB size: 15 GB (male) 40 GB (female) **55 GB Total**	Widely used in education and training. Also used to devise methods for visualization of anatomical data.
Sloan Digital Sky Survey (www.sdss.org/sdss.html)	1993	Astrophysical Research Consortium (ARC); Alfred P. Sloan Foundation, the NASA, NSF, DOE, the Japanese Monbukagakusho and the Max Planck Society	Number of Celestial Objects: 100 million Current DB Size: **15 Terabytes (TB)**	Used for characterizing the distribution of celestial objects within a particular region of the sky. Ultimate storage requirement of 50 TB

Growth in leaps and bounds

Since the first data set was made available, the size of the fMRIDC archive has grown to exceed 2.6 terabytes (1 TB = 1,024 gigabytes (GB)) of data from over 70 different fMRI studies of cognitive function (**Table 1**). These include studies from a range of cognitive domains including investigations of human memory encoding and retrieval[35], the processing of visual information[36], motor representation[37], attention[38] and working memory processes[39].

At present, the fMRIDC has fulfilled over 1,200 study data requests from researchers around the world. From user feedback questionnaires, the data obtained through the fMRIDC is being used for (i) new analyses across data from multiple studies, (ii) teaching purposes (both undergraduate and graduate), (iii) further region-of-interest analyses of brain regions that were not the focus of the original published article, (iv) application of new statistical approaches (*e.g.*, functional connectivity in contrast to the general linear model) and visualization methods and (v) for benchmarking existing neuroimaging software tools.

Prominent databasing efforts in biology and the space sciences (**Table 2**) have routinely required ample data storage to hold their contents, as well as high-end processing power to conduct new analyses of their data. Neuroimaging data sets clearly require high-performance computers (HPC) and copious amounts of data storage capability. The fMRIDC uses HPC hardware and Grid-based software[40] with a view toward large-scale data storage and for supporting large-scale neuroimaging 'mega-analysis' and modeling projects. The HPC configuration includes an array of multiprocessor compute engines, each with an accompanying server to handle individual user access. Multi-terabyte storage currently exists with potential to grow to over 10 TB of spinning disk space. The system also uses hierarchical storage management facilities to interface with a high-volume, near-line storage 100 TB robotic library. Users interested in performing tera-scale levels of fMRI analysis across a number of different studies can use fMRIDC systems with Grid-based technology rather than strictly relying on their own local single-processor machines. The Globus Alliance toolkit (www.globus.org), specifically designed for distributed computing, enables geographically distributed users to share data, computational resources and software tools.

Though modest in contrast to the larger computer systems in the earth (The Earth Simulator Project; www.earthsimulator.edu) and ocean sciences (NOAA; www.cio.noaa.gov/hpcc/), this represents a significant computational effort. As Grid-enabled, distributed users increase to levels comparable to the proteomics (such as the folding@home effort in the molecular modeling world), physical (The GridPP Project for particle physics; www.gridpp.ac.uk/) and mathematical (the ZetaGrid; www.zetagrid.net) sciences, such efforts bring an emerging degree of computing and intellectual power to modeling and visualization of complex cognitively induced patterns of functional neural activity.

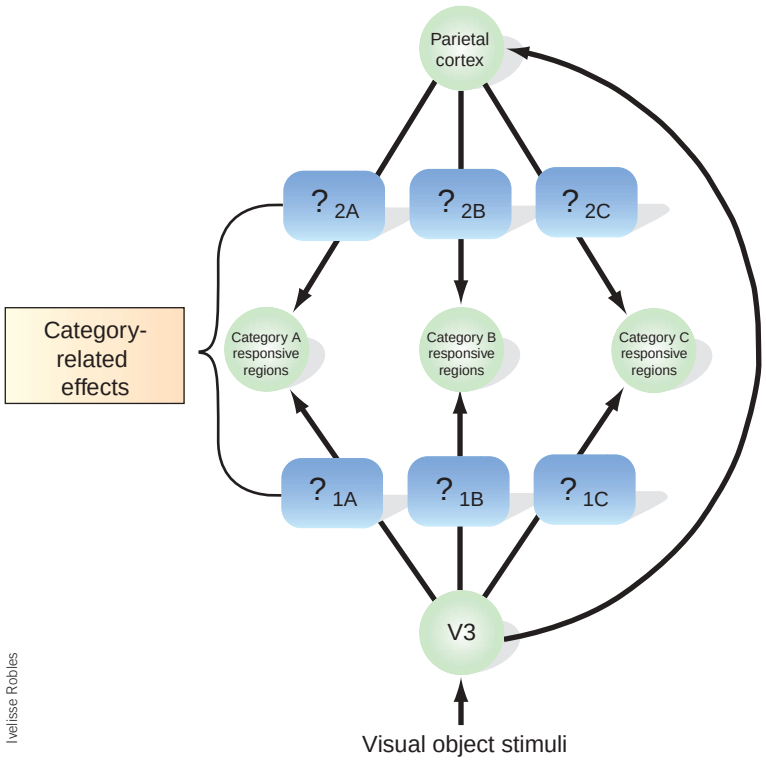


Figure 2 Functional connectivity modeling is one way in which fMRI time-course data may be reused, not to identify individual activation foci, but to explore how these foci are interacting. This connectivity diagram (based on Figure 1 in ref. 49) includes V3, a superior/inferior parietal area, and the category-responsive regions in the occipital and temporal cortices. Inputs to the connectivity model "Visual Object Stimuli" encode the presentation of visual objects (*i.e.*, houses, faces and chairs) and enter the model through V3. Weights on the various connections are determined through a Bayesian framework and indicate the strength of that connection in explaining model variance. Connectivity analyses done on publicly available fMRI data are an interesting means of exploring new hypotheses about fundamental cognitive processes.

Neuroimaging data re-use for understanding cognition

These initial outcomes indicate that data archival efforts like the fMRIDC are well on their way to becoming an essential resource for neuroscientists. But is anyone actually using previously published neuroimaging data to conduct new research? We know of several published studies that have done so. One was a reanalysis of four complete fMRI study data sets to investigate temporal components related to human consciousness[41], even though consciousness was not specifically investigated by the original authors. Another conducted a multivariate discriminant analysis to explore whether the overall pattern of fMRI activity could be used to predict what category of visual stimuli subjects were viewing[42]. A recent study of resting-state or so-called 'default-mode' activity[43] in previously published data from older adults with Alzheimer Disease (AD) and age-matched controls[7] identified reduced resting-state activity in posterior cingulate and hippocampal areas in the AD sample, suggesting a source for commonly observed reduced metabolism in these patients[44].

Two other examples of fMRI data re-use, in particular, are also worth mentioning. Areas of increasing interest to fMRI researchers are functional connectivity analysis[45] and large-scale neural modeling[46], wherein analysis of data extends beyond the identification of discrete activation hot-spots to that of examining the inter-regional patterns of correlation. These patterns are reflective of the degree to which one brain area is functionally correlated with another[47]. This is a useful tool, as the mechanisms of many cognitive phenomena are very complicated due to the huge range of interconnected and compounded processes. Modeling these connections can provide insight into how cognition may be 'wired' and can be used to make testable predictions for use in future fMRI experiments[48]. In a reanalysis of data from the fMRIDC, Mechelli *et al.*[49] estimated neuronal interactions that mediate category effects using a functional connectivity modeling technique called Dynamic Causal Modeling (DCM). They used a Bayesian framework to estimate and make inferences about the influence that one region exerts over another and how this is affected by experimental changes[50] (**Fig. 2**). They concluded that category-specific effects in occipital and temporal regions were mediated by inputs from early visual areas. In contrast, the connectivity from the superior/inferior parietal area to the category-responsive areas was unaffected by the spatial form of the presented stimuli. Their provocative findings indicate that category-specific effects in the occipital and temporal cortices may be driven by bottom-up, as opposed to top-down, mechanisms.

The reproducibility of fMRI-related effects in previously published data has also been explored. Important in measuring statistical power, reproducibility is defined as the extent to which the active status of a voxel remains the same across replicates conducted under the same conditions. Liou *et al.*[51] used an empirically based Bayesian method for estimating blood oxygenation level–dependent (BOLD) effects due to experimental stimuli, the threshold optimization procedure for assigning voxels to the active status, and the construction of reproducibility brain maps. They found that subjects in a study obtained via the fMRIDC seemed to exercise more than one mechanism in responding to visual objects while performing alternately matching and passive tasks. One mechanism appeared to evoke a pattern of BOLD activity in ventromedial temporal areas, in close agreement with what was reported in the original published article. But some subjects showed additional activity in the precuneus and posterior cingulate, suggesting a second cognitive mechanism. Additionally, the latency between the stimulus presentation and the peak of the hemodynamic response function varied considerably among individual subjects according to types of stimuli and experimental tasks. Overall, the patterns of activity were found to be statistically reproducible in at least four out of six subjects involved in the experiment. The analysis by Liou *et al.* strongly suggests that the subjects in this experiment used different cognitive processing strategies, recruiting activity in additional regions to aid object-specific areas, when making their responses to visual stimuli.

What is interesting about the Mechelli and Liou studies is that they each, independently, used the object categorical processing fMRI study of Ishai *et al.*[36], which initially reported differential patterns of activation in response to categorical stimuli in bilateral regions of the ventral and dorsal occipital cortex and in the superior temporal sulcus. But

Table 3 Online neuroscience databases

Database name	Principal modality	Data provided	Availability	Species	Principal funding source[a]	Website URL
BIRN	Microscopy; MRI	Neuroimage data; cell photomicrograph	Limited public access/ greater access to participating BIRN centers	Human, Mouse	NCRR	http://nbirn.net
BRAID	fMRI	Neuroimages	Limited access	Human	N/A	www.rad.upenn.edu/sbia/braid/publications/all.shtml
BrainMapDBJ	PET/fMRI	Neuroimaging results local maxima	Limited public access/ greater access to contributing researchers	Human	NLM	www.brainmapDBJ.org
BREDE	fMRI	VRML, XML	Open	Human	NIMH/HBP, Non-US	http://hendrix.imm.dtu.dk/software/brede
CoCoMac	Single and multi-unit recordings	Neural connectivity data	Open	Non-human	Non-US	www.mon-kunden.de/cocomac
EarLab	Single/multiunit recording of auditory neurons	Cell recording time series	Open	Non-human	NIMH/HBP	http://earlab.bu.edu
ECHBD	PET, fMRI	Neuroimage data	Limited access	Human	N/A	www.dhbr.neuro.ki.se/ECHBD/Database/index.html
fMRIDC	fMRI	Raw, processed, results brain images and study meta-data	Open	Human	NSF, WM Keck, NIMH/HBP	www.fmridc.org
ICBM	PET, MRI, fMRI, EEG, MEG	Neuroimage data	Limited access/ greater access to International Consortium for Brain Mapping (ICBM) centers	Human	NIMH/HBP, NCRR, private funding	www.loni.ucla.edu/ICBM/index.html (see also www.loni.ucla.edu)
SenseLab	Single/multi unit recording	Cell recordings from multiple sources	Open	Non-human	NIMH/HBP	www.senselab.yale.edu
Surface Management System (SuMS)	Structural MRI	Digital cortical surface models	Open	Human and non-human primates	NIMH/HBP	http://brainmap.wustl.edu:8081/sums/index.jsp

[a]Where information was available from the database web site. NCRR, National Center for Research Resources; NLM, National Library of Medicine; NIMH, National Institute of Mental Health; HBP, Human Brain Project (NIMH); NSB, National Science Foundation.

these new examinations explored the data with entirely unique approaches—at once underscoring the importance of the originally reported effects and also offering new insights into the underlying brain systems in visual processing and how fMRI data might be reused to generate new hypotheses for future study. Such secondary analyses provide new perspectives on data that may reveal effects not conceived of by the original authors, nor envisioned by theoreticians.

Neuroinformatics

In response to the sheer volume of data now collected across disciplines, a great deal of neuroscience is rapidly moving beyond its roots as simply an empirically based, hypothesis-testing endeavor, toward computationally based discovery science as well—the examination of large and disparate collections of biological data looking for unseen patterns that might provide clues to underlying mechanisms[52]. Brain imaging researchers must now also be adept computer programmers, statisticians and mathematical modelers in order to fully mine the information present in their data. This has contributed to the emerging role of 'neuroinformatics', a unifying discipline at the nexus of information technology, computer science and the neurosciences[53,54].

A number of databasing efforts have been established to accommodate the interest in sharing data from other neuroscience

domains (**Table 3**). Many of these are supported through the Human Brain Project (www.nimh.nih.gov/neuroinformatics/index.cfm), a multidisciplinary, multi-institute NIH program seeking to promote neuroscience data sharing, cross-cutting collaboration and the development of neuroinformatics tools to be used in processing, visualizing and integrating these rich sources of brain data. Other scientific interactions are occurring also in broad-based, inter-institutional collaborations, aimed at developing nationwide infrastructures for cognitive and computer scientists to share data, computational and software resources. For example, the neuroscientific thrust of the National Partnership for Advanced Computational Infrastructure (NPACI; www.npaci.edu) has for several years encouraged sharing of data and use of the computational and large-scale data storage resources based at the San Diego Supercomputing Center (www.sdsc.edu). Likewise, the Biomedical Informatics Research Network (BIRN; www.nbirn.net; see p. 467–472 this issue[55]) seeks to cross-reference imaging and computational resources between parti-cipating centers.

Collectively, these projects are now at a crossroads where they must look beyond content-building within their respective fields and migrate toward connecting information across databases and associated resources to form a larger, web-enabled, interoperable network of neuroscientific data, tools and knowledge. The *Society for Neuroscience* (SFN) has recognized the wealth of information available in these databases, their potential utility as tools for research and education (see Autumn 2003 *SFN Neuroscience Quarterly* for commentary by former SFN president H. Akil; http://web.sfn.org/content/Publications/NeuroscienceNewsletter/2003fall/message.html), and is taking a leading role in linking these efforts through a common website entitled the Neuroscience Database Portal (http://big.sfn.org/ndg/site). The Human Brain Project also provides a portal that cross-references these and other online sources of data and information (http://ycmi-hbp.med.yale.edu/hbpdb). As these portals develop further and become fully implemented over the next few years, brain researchers may be able to tunnel from the systems level down to the molecular level of spatial resolution, explore neurally related electrical and magnetic field changes across the scalp, and directly perform basic neuroimaging meta-analyses, all via their web browsers.

New databases arise often, while many others go stale over time. It can be difficult for one to know which are best suited to one's needs and have a track record of success. We believe that organizations like the SFN and HBP are best suited to evaluate and rate available databases for content, accessibility, ease of use, longevity, interoperability, and so on, and to direct researchers to sites where they can request as well as contribute data. Although they would probably avoid supporting any particular database, these leading scientific societies can have a very important role by providing a periodically reviewed stamp of approval to those data resources listed in their online portals.

Conclusions

Large-scale brain imaging archival efforts like the fMRIDC have begun to pay significant scientific dividends for cognitive neuroscience. Other databases, too, hold great promise. The involvement of other researchers as well as multiple scientific communities in examining published brain imaging data should be welcomed, not feared, as it will serve to strengthen and improve the inferences and conclusions we can make from our data. As a result of these infrastructural and data resources, novel research, hypotheses and education using existing data can reach across scientific disciplines—engaging workers from other fields to apply sophisticated new tools for data analysis and integration. Bright minds, new points of view and sophisticated tools will do much to organize and help focus ideas about the neural processes represented in these data.

These projects are not trivial, however, requiring a dedicated staff to manage study curation and maintain computer systems. Government mandates to share and archive primary research data cannot be successful unless federal agencies are willing to support the infrastructure needed to make such sharing easy and accessible to others. Further progress will certainly necessitate that funding bodies continue to invest in new fMRI experimentation, but also in supporting study data repositories, thereby ensuring their survival as essential archival and scientific resources.

We expect that, in time, because of the enormous scientific and educational benefits, the sharing of neuroimaging and other brain data will simply be an expected part of publishing in leading periodicals like this one. In so-doing, neuroscience will be able to fully leverage its collected scientific knowledge into a rich understanding of complex brain processes.

ACKNOWLEDGMENTS

The fMRI Data Center is supported by the National Science Foundation (BCS-9978166), the William M. Keck Foundation, and the National Institute of Mental Health Human Brain Project.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/

1. Raichle, M.E. Cerebral blood flow and metabolism. *Ciba. Found. Symp.* 85–96 (1975).
2. Roy, C.S. & Sherrington, C.S. On the regulation of blood supply of the brain. *J. Physiol.* **11**, 85–108 (1890).
3. Phelps, M.E., Hoffman, E.J., Huang, S.C. & Kuhl, D.E. Positron tomography: *in vivo* autoradiographic approach to measurement of cerebral hemodynamics and metabolism. *Acta. Neurol. Scand. Suppl.* **64**, 446–447 (1977).
4. Posner, M.I. & Desimone, R. Beyond images. *Curr. Opin. Neurobiol.* **8**, 175–177 (1998).
5. Richeson, J.A. *et al.* An fMRI investigation of the impact of interracial contact on executive function. *Nat. Neurosci.* **6**, 1323–1328 (2003).
6. Goel, V. & Dolan, R.J. The functional anatomy of humor: segregating cognitive and affective components. *Nat. Neurosci.* **4**, 237–238 (2001).
7. Buckner, R.L., Snyder, A.Z., Sanders, A.L., Raichle, M.E. & Morris, J.C. Functional brain imaging of young, nondemented, and demented older adults. *J. Cogn. Neurosci.* **12**, 24–34 (2000).
8. Simpson, J.R. *et al.* The emotional modulation of cognitive processing: an fMRI study. *J. Cogn. Neurosci.* **12**, 157–170 (2000).
9. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank. *Nucleic Acids Res.* **31**, 23–27 (2003).
10. Berman, H.M. *et al.* The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 899–907 (2002).
11. Collins, F.S. & Mansoura, M.K. The human genome project. *Cancer* **91**, 221–225 (2001).
12. Yarmush, M.L. & Jayaraman, A. Advances in proteomic technologies. *Annu. Rev. Biomed. Eng.* **4**, 349–373 (2002).
13. 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. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356**, 1323–1339 (2001).
14. Benson, D.A. *et al.* GenBank. *Nucleic Acids Res.* **30**, 17–20 (2002).
15. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
16. Wlodawer, A. *et al.* Immediate release of crystallographic data: a proposal. *Science* **279**, 302 (1998).
17. Sussman, J.L. Protein Data Bank deposits. *Science* **282**, 1991 (1998).
18. Anonymous (Opinion). Rules of genome access. *Nature* **404**, 414 (2000).
19. Cozzarelli, N.R. UPSIDE: Uniform Principle for Sharing Integral Data and Materials Expeditiously. *Proc. Natl. Acad. Sci. USA* **101**, 3721–3722 (2004).
20. Cech, T.R. Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences (The National Academies Press, Washington D.C., 2003).
21. Shepherd, G.M. Supporting databases for neuroscience research. *J. Neurosci.* **22**, 1497 (2002).
22. Van Essen, D.C. Windows on the brain: the emerging role of atlases and databases in neuroscience. *Curr. Opin. Neurobiol.* **12**, 574–579 (2002).
23. Toga, A. Neuroimaging databases: the good, the bad, and the ugly. *Nat. Rev. Neurosci.* **3**, 302–309 (2002).

24. Koslow, S.H. Should the neuroscience community make a paradigm shift to sharing primary data? *Nat. Neurosci.* **2**, 863–261 (2000).
25. Fox, P. & Lancaster, J. Mapping context and content: the BrainMap model. *Nat. Rev. Neurosci.* **3**, 319–321 (2002).
26. Cover, T.M. & Thomas, J.A. *Elements of Information Theory* (ed. Schilling, D. L.) (Wiley, New York, 1991).
27. Van Horn, J.D. & Gazzaniga, M.S. Maximizing information content in shared neuroimaging studies of cognitive function. *Databasing the Brain: From Data to Knowledge* (eds. Koslow, S.H. & Subramanian, A.) (John Wiley and Sons, New York, in press).
28. Noy, N.F. & Klein, M. Ontology evolution: not the same as schema evolution. *Technical Report SMI-2002-0926, Stanford Medical Informatics* (2002).
29. Oliver, D.E. *et al.* Ontology development for a pharmacogenetics knowledge base. *Pac. Symp. Biocomput.*, 65–76 (2002).
30. Berners-Lee, T., Hendler, J. & Lassila, O. The semantic web: a new form of web content that is meaningful to computer will unleash a revolution of new possibilities. *Sci. Am.,* **284,** 34-43 (2001).
31. Van Horn, J.D. *et al.* The fMRI data center: software tools for neuroimaging data management, inspection, and sharing. *A Practical Guide to Neuroscience Databases and Associated Tools* (ed. Kotter, R.) 221-235 (Kluwer, Amsterdam, 2002).
32. Mirsky, J.S., Nadkarni, P.M., Healy, M.D., Miller, P.L. & Shepherd, G.M. Database tools for integrating and searching membrane property data correlated with neuronal morphology. *J. Neurosci. Methods* **82**, 105-121. (1998).
33. Ascoli, G.A., Krichmar, J.L., Scorcioni, R., Nasuto, S.J. & Senft, S.L. Computer generation and quantitative morphometric analysis of virtual neurons. *Anat. Embryol. (Berl.)* **204**, 283–301 (2001).
34. Insel, T.R., Volkow, N.D., Li, T.-K., Battey, J.F. & Landis, S.C. Neuroscience networks: data-sharing in an information age. *PLoS Biol.* **1**, 9–11 (2003).
35. Mechelli, A., Gorno-Tempini, M.L. & Price, C.J. Neuroimaging studies of word and pseudoword reading: consistencies, inconsistencies, and limitations. *J. Cogn. Neurosci.* **15**, 260–271 (2003).
36. Ishai, A., Ungerleider, L.G., Martin, A. & Haxby, J.V. The representation of objects in the human occipital and temporal cortex. *J. Cogn. Neurosci.* **12**, 35–51 (2000).
37. Toni, I. *et al.* Multiple movement representations in the human brain: an event-related fMRI study. *J. Cogn. Neurosci.* **14**, 769–784 (2002).
38. Kable, J.W., Lease-Spellmeyer, J. & Chatterjee, A. Neural substrates of action event knowledge. *J. Cogn. Neurosci.* **14**, 795–805 (2002).
39. Rypma, B., Berger, J.S. & D'Esposito, M. The influence of working-memory demand and subject performance on prefrontal cortical activity. *J. Cogn. Neurosci.* **14**, 721–731 (2002).
40. Foster, I. The grid: computing without bounds. *Sci. Am.* **288**, 78–85 (2003).
41. Lloyd, D. Functional MRI and the study of human consciousness. *J. Cogn. Neurosci.* **14**, 818–831 (2002).
42. Carlson, T.A., Schrater, P. & He, S. Patterns of activity in the categorical representations of objects. *J. Cogn. Neurosci.* **15**, 704–717 (2003).
43. Greicius, M.D., Srivastava, G., Reiss, A.L. & Menon, V. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc. Natl. Acad. Sci. USA* published online March 15 (2004).
44. Reiman, E.M. *et al.* Functional brain abnormalities in young adults at genetic risk for late-onset alzheimer's dementia. *Proc. Natl. Acad. Sci. USA* **101**, 284–289 (2004).
45. Ramnani, N. *et al.* Exploring brain connectivity: a new frontier in systems neuroscience. *Trends Neurosci.* **25**, 496–497 (2002).
46. Horwitz, B., Tagamets, M.A. & McIntosh, A.R. Neural modeling, functional brain imaging, and cognition. *Trends Cogn. Sci.* **3**, 91–98 (1999).
47. Buchel, C., Coull, J.T. & Friston, K.J. The predictive value of changes in effective connectivity for human learning. *Science* **283**, 1538–1541 (1999).
48. Arbib, M.A., Billard, A., Iacoboni, M. & Oztop, E. Synthetic brain imaging: grasping, mirror neurons and imitation. *Neural Net.* **13**, 975–997 (2000).
49. Mechelli, A., Price, C., Noppeney, U. & Friston, K. A dynamic causal modelling study on category effects: bottom-up or top-down mediation? *J. Cogn. Neurosci.* **15**, 925–934 (2003).
50. Friston, K.J., Harrison, L. & Penny, W. Dynamic causal modelling. *Neuroimage* **19**, 1273–1302 (2003).
51. Liou, M., Su, H.-R., Lee, J.-D. & Cheng, P.E. Bridging functional MR images and scientific inference: reproducibility maps. *J. Cogn. Neurosci.* **15**, 934–945 (2003).
52. Hood, L. Leroy Hood expounds the principles, practice and future of systems biology. *Drug Discov. Today* **8**, 436–438 (2003).
53. Beltrame, F. & Koslow, S.H. Neuroinformatics as a megascience issue. *IEEE Trans. Inf. Technol. Biomed.* **3**, 239–240 (1999).
54. Koslow, S.H. Opinion: Sharing primary data: a threat or asset to discovery? *Nat. Rev. Neurosci.* **3**, 311–313 (2002).
55. Martone, M.E., Gupta, A. & Ellisman, M.H. e-Neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat. Neurosci.* **7**, 467–472 (2004).
56. Brett, M., Johnsrude, I.S. & Owen, A.M. The problem of functional localization in the human brain. *Nat. Rev. Neurosci.* **3**, 243–249 (2002).
57. Talairach, J. & Tournoux, P. *Co-Planar Stereotactic Atlas of the Human Brain* (Thieme, New York, 1988).
58. Friston, K.J. *et al.* Analysis of fMRI time-series revisited. *Neuroimage* **2**, 45–53 (1995).
59. Druzgal, T.J. & D'Esposito, M. Dissecting contributions of prefrontal cortex and fusiform face area to face working memory. *J. Cogn. Neurosci.* **15**, 771–784 (2003).