Neuroinform
DOI 10.1007/s12021-008-9009-y

Data Sharing for Computational Neuroscience

Jeffrey L. Teeters · Kenneth D. Harris · K. Jarrod Millman · Bruno A. Olshausen · Friedrich T. Sommer

Received: 9 January 2008 / Accepted: 10 January 2008

Abstract Computational neuroscience is a subfield of neuroscience that develops models to integrate complex experimental data in order to understand brain function. To constrain and test computational models, researchers need access to a wide variety of experimental data. Much of those data are not readily accessible because neuroscientists fall into separate communities that study the brain at different levels and have not been motivated to provide data to researchers outside their community. To foster sharing of neuroscience data, a workshop was held in 2007, bringing together experimental and theoretical neuroscientists, computer scientists, legal experts and governmental observers. Computational neuroscience was recommended as an ideal field for focusing data sharing, and specific methods, strategies and policies were suggested for achieving it. A new funding area in the NSF/NIH Collaborative Research in Computational Neuroscience (CRCNS) program has been established to support data sharing, guided in part by the workshop recommendations. The new funding area is dedicated to the dissemination of high quality data sets with maximum scientific value for computational neuroscience. The first round of the CRCNS data sharing program supports the preparation of data sets which will be publicly available in 2008. These include electrophysiology and behavioral (eye movement) data described towards the end of this article.

Keywords Computational neuroscience · Data sharing · Electrophysiology · Online data repositories · Computational models · Hippocampus · Sensory systems · Eye movements · Visual cortex

J. L. Teeters · K. J. Millman · B. A. Olshausen · F. T. Sommer (✉)
Redwood Center for Theoretical Neuroscience & Helen Wills Neuroscience Institute, University of California, Berkeley, 3210F Tolman Hall, MC 3192, Berkeley, CA 94720-3192, USA
e-mail: fsommer@berkeley.edu

K. D. Harris
Center for Molecular and Behavioral Neuroscience, Rutgers, The State University of New Jersey, 197 University Avenue, Newark, NJ 07102, USA

Introduction

Many aspects of brain function that are currently unexplained could potentially be understood if a culture of data sharing were established in neuroscience. The fields of computational neuroscience and systems neuroscience seek to elucidate the information processing strategies employed by neural circuits in the brain. In spite of intense investigation in these areas over many decades, fundamental problems are still unsolved. For example, the question of how brains can perceive and navigate so robustly, even under rich and highly variable real-world conditions, or the question of how sensation and action interact, or how brain function relies on concerted neural activity across scales, remain complete mysteries. Tackling these and other types of questions will require that a number of recent technologies be combined, in particular:

- New experimental techniques for acquiring rich high-dimensional physiological and anatomical data of the behaving brain.
- New computational approaches for integrating data across modalities and levels of analysis.
- New data mining and data integration techniques for extracting meaningful and useful information from complex interacting systems.

Hybrid approaches using various combinations of expertise in these technologies extend and complement the traditional "single lab" approach and thereby promise new scientific discovery (Insel et al. 2004). The most flexible interface to allow researchers to combine different technologies is public data sharing. Thus, it is crucial for neuroscience to make data available to scientific communities rather than sequestering them in individual labs (Gardner et al. 2003; Kennedy 2006).

To explore the best strategies for promoting data sharing in neuroscience a workshop was organized at the University of Maryland on June 6th and 7th, 2007. The workshop brought together five different groups of researchers and experts: (1) Principal investigators from experimental neuroscience labs presented data sets which they could prepare for data sharing. (2) Principal investigators from computer science and theoretical neuroscience labs shared their experiences with scientific databases (e.g. neuronal morphologies, gene expressions and linguistic data). (3) Representatives of national and international organizations to coordinate research in neuroscience. (4) An expert provided guidance about general legal issues of data and resource sharing. 5) Observers from federal funding agencies, NSF and NIH.

The following four sections summarize the assessments and recommendations that resulted from the discussions at the workshop. The fifth section describes the new CRCNS funding area to support data sharing that was guided in part by these recommendations.

Objectives, Obstacles and Opportunities

How to Target the New Data Sharing Activity?

The successful examples for sharing neuroscience resources that already exist focus on specific data types, for instance, cell morphologies (NeuroMorpho.org; Liu and Ascoli 2007), cortical connectivity (CoCoMac.org; Koetter 2004), neuroimaging (http://www.fmridc.org; Van Horn and Ishai 2007) and also open source analysis tools (Chronux.org), to give just a few examples. In general, workshop participants assessed that the willingness to share data is influenced by the following factors:

- How difficult are the data to collect?
- Will sharing compromise individual research programs? (Are the shared data part of a lengthy ongoing study? Have the data been mined locally?)
- How will funding agencies acknowledge contributions?
- How much time must data contributors devote to supporting the shared data?
- How difficult are the data to interpret? Might an investigator who was not involved in the experimental design or data collection make critical errors when using the data?

Any new data sharing initiative should focus on data types that are inadequately covered by existing initiatives. However, rather than being restricted to a particular data type, the workshop participants recommended that the new data sharing activity be guided by an overarching objective: to foster and advance computational neuroscience. Of course, for practical reasons, the activity has to initially focus on a few types of neuroscience resources. The workshop identified data from electrophysiology and from microscopic imaging of neural activity as resource types whose public availability is still vastly underrepresented—though at least one pioneer effort exists (NeuroDatabase.org; Gardner 2004). The broader availability of such data can be expected to be of direct impact to computational neuroscience and neuroscience in general. Another important argument for focusing on neurophysiology data is that they can often be used to address multiple questions. Data that have been collected with one question in mind often turn out to be highly valuable to address other questions. Examples are: (1) Hippocampus recordings for mapping place fields were the basis for high-profile papers addressing questions concerning temporal organization of neural codes (Harris et al. 2003; Harris et al. 2002). (2) Paired recordings using extracellular and intracellular electrodes originally collected for detecting dendritically generated action potentials provide ground truth for testing and comparing spike-sorting techniques (Harris et al. 2000).

Impediments to Data Sharing

In practice, there are social and technical obstacles for creating public databases in neuroscience (Ascoli 2006). Currently, experimenters have reason to worry that sharing data could work against them and has no clear benefit. There is no established mechanism to provide credit for sharing data, and, conversely, in competitive situations shared data could even be used unfairly by reviewers in confidential paper reviews. Further, experimentalists are concerned about misinterpretation of their data by individuals who are insufficiently familiar with the experimental procedures. Theorists often underestimate the intricacies of interpreting experimental data appropriately. Also, it is important to realize that large gaps in language and ways of thinking between experimenters and theorists must be overcome. Many theorists believe that complex and seemingly disparate processes are based on a small set of basic principles. In contrast, biologists are aware of myriads of details that are necessary to fully understand the experimental component of the work.

A principal technical obstacle to data sharing in neuroscience is that there is currently no standard for specifying the metadata required for understanding a data set.

Recommended Strategies for New Data Sharing Activities

A central question discussed at the workshop was how data sharing in neuroscience could be achieved efficiently. There was broad consent that sharing should be voluntary and should obey the rule that whoever collected the data has priority in determining the rules for its use. In particular, the contributing lab should determine when—and to what extent—the data is made available, and the conditions under which it can be used.

Two endeavors are critical to advance data sharing in neuroscience: (1) the development of generally applicable database techniques and (2) pragmatic focal approaches for data sharing of particular types of data. There are already a number of successful focal activities of data sharing in neuroscience (see first paragraph in this section). The expansion of such focused efforts to more data types will have a high impact in the near future. Initiating and shaping communities that share certain data types can help define the specifications of database technology required for organization of neuroscience data across data types.

Data must be very well documented to reduce the danger of misinterpretation. Such documentation will require substantial work from the experimenter. Thus, the workshop participants concluded that funding mechanisms for data contributors are necessary to promote data sharing in neuroscience.

Impacts on Research and Teaching

All workshop participants agreed that there are significant unexploited scientific and educational opportunities to be gained from sharing experimental data, analysis and modeling tools. In particular, the following impacts were identified:

- The analysis of experimental data gathered in an individual lab by multiple research groups with different perspectives is expected to significantly enhance understanding in neuroscience.
- New scientific insights are expected if experimental data from several labs are combined and subjected to meta analyses that are unfeasible with individual data sets.
- Increased transparency, reproducibility and comparability of neuroscience results. Labs may be able to cross-check their own results against those of other labs in a more quantitative manner than would be possible based on journal publications.
- Comprehensive repositories for neuroscience data and methods will be invaluable for neuroscience education.
- Availability of web-based teaching infrastructures for students and investigators. Such infrastructures could consist of tutorials including data and teaching material made by biologists, as well as bibliographies of review articles and links to relevant websites.
- Creation of test beds for improving and benchmarking methods for analyzing and modeling neuroscience data could accelerate progress in software tools critical for basic neuroscience and clinical research.

Impact on Productivity in the Neuroscience Community

Many types of neurophysiology data sets, such as multi-channel recordings in rat hippocampus, are typically acquired in one or two days by a team of highly trained and experienced experimenters. The analysis of even a single of these very large data sets can take years. At the same time, there is a growing community of theorists in neuroscience that are trained in analysis methods but that have no direct access to experimental data. The workshop participants concluded that this mismatch of skills and resources is one important reason that existing data sets are underexploited. At the same time too much theoretical expertise is spent on models that are not guided by data and often only of limited relevance to neuroscience. Thus, the desirable and most optimal resource allocation in neuroscience is to enlist more theorists with varied theoretical skills in analyzing and modeling experimental data sets. Public sharing of experimental data would offer the fastest and most flexible interface for organizing and optimizing this "resource allocation." First, it allows theorists to shift towards working on real data. Second, it improves the exploitation of existing experimental data. Third, it encourages new collaborations between theorists and experimental labs. Lastly, it creates new educational opportunities for students at institutions without direct access to neuroscience labs.

Recommended Activities and Cautions

The workshop participants identified important activities that the planned data sharing activity should include:

The central services and infrastructure should provide:

1) *Service to resource contributors*: The major role of an infrastructure for data sharing is to lessen the burden on contributors to make their data/resources available. This includes making it easy to convert and upload data, safeguarding data, and relieving contributors from having to provide support to individual users.
2) *Service to data users*: The repository web site should offer a means of indexing and flexible online-access

and visualization. Further, each resource should have its individual facilities for user groups, such as a FAQ list and discussion board to enable problem resolution based on user interactions.

3) *Ontologies*: Experience gained from preparing data sets for sharing should be used to create and refine adequate ontologies for describing the data and metadata.
4) *Monitoring of usage*: Tracking and evaluating the usage is a very important measure for credit assignment to the resource contributors. Further, it is an important indicator to steer the activities in data sharing.
5) *Expandable infrastructure*: Although the initial scope is limited, it is important to set up the infrastructure, such as data formats, data transmission and browsing methods, and web resources so that they may be expanded and enhanced over time.

Additional resources (in addition to research resources):

6) *Teaching tools*: Explanatory tutorials for the shared resources will not only allow potential users to get oriented and acquainted with the data, they are also online facilities that can substantially improve education in neuroscience.
7) *Challenges and competitions*: To encourage new interdisciplinary approaches in neuroscience, contributions should be encouraged that can be used for organization of competitions and challenges. Good existing examples of such activities are the *Berkeley Prediction Challenge* (J. Gallant, F. Theunissen, http://neuralprediction.berkeley.edu/) and the Lausanne competition (W. Gerstner, A. Roth, F. Schuermann, R. Jolivet, http://icwww.epfl.ch/~gerstner/QuantNeuronMod2007/challenge.html) for predicting single neuron behavior. Such contributions should raise interesting questions and/or should offer benchmark environments for evaluating and comparing different solutions (e.g., data sets with some ground truth for comparing spike sorting algorithms).

Recommended strategies and cautions:

8) *Responding to market demands*: The data sharing activity should enable demand-driven scientific foci in the area of computational neuroscience. Thus, the main goal of the program should be to create an interactive market place for resources of particular significance for the field rather than creating a large repository for everything.
9) *Funding mechanism for data contributions*: To select contributions for funding, it was recommended to use a reviewing process involving outside reviewers to judge quality and importance.
10) *Avoid trying to be all things to all people*: Today, no single initiative for sharing resources over the web can be all encompassing. It will be important that the planned data sharing activity closely ties in with other neuroscience initiatives. This includes the exchange with similar efforts (Gardner 2004) and the involvement with national (eg. Neural Information Framework by NIH—neurogateway.org, and Neuroscience Database Gateway by SfN—ndg.sfn.org), as well as international coordinating efforts (e.g. International Neuroinformatics Coordination Facility in Stockholm; Bjaalie and Grillner 2007).
11) *Answer the needs of many*: Experiences with current activities (e.g. competitions described under item 7, above) suggest that on-line availability alone is insufficient to foster broad use of a neurophysiology data set. Therefore, user-friendliness and attentiveness to user feedback are essential.
12) *Seek dialogue with scientific community*: The workshop discussed the initial procedures to start the effort in data sharing in neuroscience. As this effort is developing, feedback from the relevant communities will be crucial for success.
13) *Sensitivity to specific problems with data sharing in neuroscience*: The specific problems brought up were (1) the difficulty of communicating all relevant metadata about experimental conditions without direct interaction between contributors and data users, and (2) confidentially issues connected with drawing public attention to animal studies.
14) *Miscellaneous concerns*: Software is scalable, humans are not. Human costs to generate, share, access data. Security and privacy: think about security up front! Keep distribution/legal issues in mind (copyrighted stimulus material/disclaimer). Be aware of HIPAA, etc. Software used to setup the repository should be open source and documented.

Organization of Data Sharing Activity

How can the above activities be accomplished? It is reasonable that most of the service requirements described under items (1) to (5) should be provided by a small core team of theoretical neuroscientists and programmers. Thus *core services* are being developed to implement and administer standardized procedures that can be used across resources that are shared in the program.

Further, for steering and evaluation of the activities (items 8–14) a *governance board* will be established. The board will include noninvolved experts in the field, experimental and theoretical neuroscientists, some active resource contributors, representatives of related activities and other experts. It is important that the board not only

represents neuroscience but different scientific communities. The board should guide important decisions concerning the core services. Conversely, the experience gathered in this data sharing project should be brought back to the relevant communities in workshops and publications. The proposed organization and interactions between the entities making up the data sharing activity is illustrated in Fig. 1.

Specific Methods for Data Sharing Activity

The Metadata of an Experiment

Neuroscience experiments involve the generation and manipulation of large quantities of both raw and processed data. Without knowing the precise acquisition parameters and the experimental conditions the raw data are meaningless. Moreover, only when keeping track of the exact history of the analytic steps can an analysis result be correctly interpreted. All these *metadata* are typically scattered and too often lost. The ability to reanalyze one's own data depends critically on access to the totality of this disparate, information; the problem compounds when sharing the data. In the data preparation process the description of metadata will probably take the majority of the work of the data contributor. The metadata should include:

- Description of experimental conditions/experimental paradigms etc.
- Species, age of the animal, etc.
- Surgical procedures
- Recording technique (electrode type, clamp method, etc.)
- Positions of recording electrodes
- Information of recorded cells, such as cell anatomy, type and laminar position
- Tools/procedures used to process the data (e.g., type and parameters of spike sorting)
- Information on how stimuli were generated
- Information on how the timing of the stimuli is correlated with the recordings
- For cells in the visual system, if available, the size and location (eccentricity) of stimuli in retinal coordinates.
- Known caveats about use of the data (identified artifacts, etc.)
- Links to publications

Organizing and Handling the Data

As mentioned earlier, a unified data model for handling metadata in neuroscience is still an open research problem. Thus, the workshop participants recommended a pragmatic approach to handle metadata. Information about experimental conditions can be handled in a database with state-of-the-art features. In particular, the database should allow for object-oriented schemes including inheritance and aggregation, and allow specifying default settings (e.g., conditions) that are automatically inherited. Information that cannot be fully organized in a database should be described in text form, perhaps supplemented by markups (XML). Existing schemes (for example Gardner 2004—http://brainml.org/) should be extended and customized as needed.

To organize raw data and metadata of experiments in a clear fashion it will be advantageous to use a hierarchical data format, such as HDF5. Hierarchical data formats enable organizing not only the primary experimental data

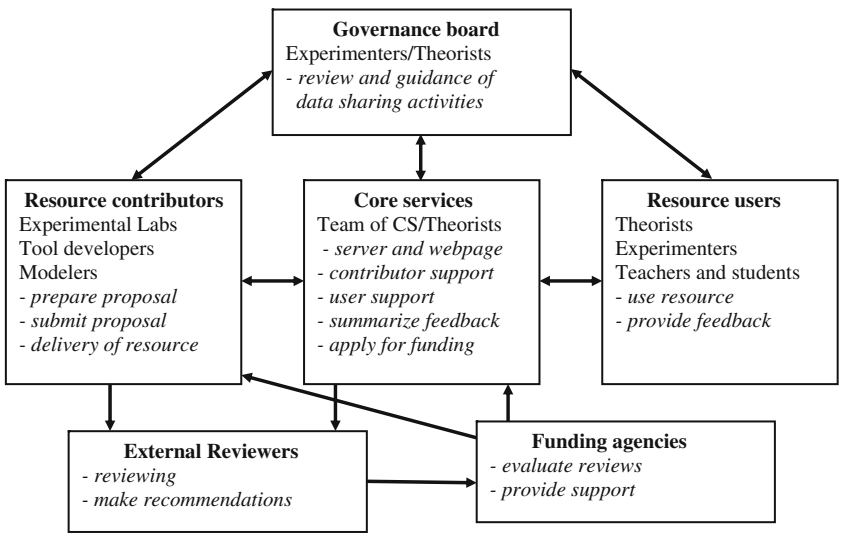

Fig. 1 Recommended organization of data sharing activity

but also the results from analyses of the data: Such secondary data sets can be put on levels in the hierarchy that reflect the number of analysis steps necessary to produce them. Further, HDF5 is supported by data access protocols, which provide flexible, partial online access for browsing of the data. This kind of functionality is important for user-friendly data access.

Version control, such as CVS or SVN (subversion) should be used for all available resources to allow tracking changes.

The core services should establish the infrastructure so that even very large data sets can be shared. For data sets that are too large to be transferred over the Internet a service should be provided for mailing hard disks (collaboration with Google's Palimpsest Project).

Validation of Contributions

Proposals for contributions of data or other resources should be reviewed by experts in the field. These experts should be asked to evaluate the proposed contribution with respect to scientific quality and relevance.

The course of each resource contribution should involve several stages of quality control. First, the contributing lab must carefully screen and analyze the data before selecting them for data sharing. Second, if data conversion is involved, the core services should provide configurable tools for the bidirectional conversion between the data formats. After a conversion cycle (to the new format and back) the data providers should verify—by screening and by data analysis using their own tools—that the conversion process did not corrupt the data. Third, the core services should check the consistency of the data.

Levels of Control for Data Contributors

The level of control that data contributors will seek over studies of their data will vary widely. Some labs may allow publications about their data without requesting any direct involvement, simply requesting proper attribution of the data. This form of data sharing has proven to be most effective in other fields, such as Bioinformatics (e.g. Human Genome Database—http://www.gdb.org) and Linguistics (e.g. Linguistic Data Consortium—http://www.ldc.upenn.edu). In these fields, the lack of direct authorship of the contributor seems usually be compensated by credit for having created widely used data. Currently, there is no precedent but it seems reasonable to assume that such compensation would also work in neuroscience, as the impact of high-quality contributed data is realized and appreciated by the research community.

In some cases, labs may wish to control the publications involving their data. One concern is the possibility that other researchers might discover errors in earlier publications using the data. It was generally agreed that this reason, though understandable, is illegitimate: Although potentially troublesome in the short term, such discovery would be important for science and would also be in the long-term interest of the contributing lab. Two important reasons for some limited form of control were brought up in the workshop discussion. The first reason was the danger that miscommunication about meta-information could lead other researchers to publish erroneous conclusions. It was argued that peer review for publishing papers should in principle take care of this problem. However, some experimenters felt that this is not enough protection because even a single incident of data misinterpretation could have harmful effects on the reputation of the lab. The second reason was the fact that the contributed data are still used by the lab and therefore all publications using these data should be somehow coordinated.

A "soft" way for contributors to retain control over publications involving their data would be to request that researchers using their data contact their lab before publication. Although this might work in practice, non-observance of the request can not be legally enforced. As was clarified by an expert on intellectual property, there is little legal basis for exerting and enforcing control over published data. *As a general rule, data does not fall under US copyright protection*.

A stronger way for contributors to retain some control over publications about their data is to release them partially and to keep some data that can be used to cross-validate predictions about the data. The remaining data can be requested from the lab on a case-by-case basis. It was agreed that partial releases are still useful, if the released data is enough to enable interesting studies. The disadvantage of partial releases is that the lab stays coupled to the data-sharing process and must remain responsive to user requests.

To facilitate this, the repository should be designed to support releasing the data in varying degrees: Release all data for research activities; release large fractions but withhold some for cross-validation; release partial data for teaching purposes and release only metadata with a contact mechanism to gain more access.

Evaluating Utility to the Community

The usage of individual data sets in the repository should be monitored and analyzed, for example using *Google Analytics*. The usage statistics and other user feedback is

important for assigning credit to contributors of resources. Further, it can be used to estimate the utility to the community and steer future efforts of the sharing activity.

CRCNS Data Sharing Program

The Collaborative Research in Computational Neuroscience (CRCNS) is a joint program of NSF and NIH that, since 2002, has supported integration of theoretical and experimental neuroscience through collaborative research projects typically involving two to five senior investigators. CRCNS has recently begun to offer funding for a new class of proposals focused on the sharing of data and other resources (http://www.nsf.gov/pubs/2008/nsf08514/nsf08514.htm).

NSF solicited comments on data sharing from the research community in March 2007. Responses from the community suggested that while sharing of data, code, stimuli, and other resources are all highly desirable, sharing of experimental data represented a particularly acute need, which was not likely to be filled without leadership from within the research community, support by funding agencies, and a concerted, organized effort. A first round of data sharing proposals was awarded in August 2007, and the first shared data resources will be made available publicly in March 2008. A broader call for proposals for data sharing and corpora development was included in a new release of the CRCNS solicitation in November 2007, to support the preparation and deployment of data, software, code bases, stimuli, or other resources that would be useful to a broad community of researchers.

This section summarizes the data sharing projects that were supported in the first round of reviews, and the core services that will be provided to support both data contributors and data users.

Data Sets Supported by the CRCNS Program in 2007

Hippocampus data (Buzsáki lab, Rutgers University) Physiological and anatomical data from the rat hippocampus, including (1) recordings from hippocampal CA1 neurons during open field foraging, (2) simultaneous intracellular and extracellular in vivo recordings from CA1 pyramidal cells and histological identities of those neurons, (3) quantitative information on the cellular connectivity of the hippocampal formation, and (4) axonal reconstruction data from *in vivo* preparations. Anatomical and physiological data will be cross-annotated to facilitate browsing and integration, and provided in a form that is compatible with widely used simulators. It is anticipated that these data will be useful for developing anatomically and physiologically realistic neural networks and understanding emergent behavior of neuronal populations, in particular, the mechanisms of memory.

Recordings to explore sensory coding A team of investigators from UC Berkeley—Yang Dan, Tim Blanche, Jack Gallant, and Frederic Theunissen—will make several data sets available, each exploring different aspects of sensory coding: (1) cortical slice data acquired in order to examine the effects of complex spike trains in the induction of long-term synaptic modification; (2) recordings of primary visual cortical neurons made during stimulation with complex stimuli, white noise, and natural images; (3) recordings from visual area V4 during stimulation with parametrically varying bars, rings and gratings; (4) recordings from visual areas V1, V2, and V4 during stimulation with a rapid dynamic sequence of gratings; (5) recordings of neurons at three levels of the avian auditory system during stimulation with complex synthetic and natural sounds; and (6) large-scale neuronal recordings from primary visual cortex made with multi-site electrode arrays that allow simultaneous recording from more than a hundred single units at once. It is anticipated that these data will be useful for the study of spatial and temporal neural coding, nonlinear receptive field properties, learning rules, hierarchical processing strategies, and other aspects of the analysis of complex sensory information.

Multi-unit recordings in primary visual cortex (Dario Ringach Lab, UCLA) The data are single- and multi-unit recordings from primary visual cortex, obtained using either standard microelectrodes or micro-machined electrode arrays. Both spontaneous and stimulus driven activity are available in a number of different conditions, including standard receptive field characterizations (e.g., orientation tuning, spatial and temporal frequency tuning) and more specific experiments such as sub-space receptive field mapping and natural image sequences. Data from micro-machined electrode arrays also include local field potentials and surface EEG. It is anticipated that these data will be useful for studies of visual processing, population coding, and retinotopy, and that the large-scale high-dimensional data will be well suited for exploration by novel machine learning and statistical methods.

Data and tutorial on intracellular recordings in sensory areas The data are intracellular (whole-cell patch) recordings obtained in vivo from visual, auditory, somatosensory, and motor areas of the neocortex by the laboratories of Judith Hirsch, USC; Anthony Zador, CSHL; Michael DeWeese, UC Berkeley and Michael Brecht, Humboldt

University Berlin. These data include not only spikes but also membrane voltages or currents generated by synaptic connections and intrinsic membrane channels. In addition to providing data, the investigators will develop tutorial materials describing recording methods, stimulus paradigms, and issues relevant to the interpretation of intracellular recordings. It is anticipated that this pooled data set will be useful for those wishing to study a particular sensory modality as well as those who hope to understand common features of neocortical function. It will also be of great value for the development of new methods of data analysis.

Eye movement data (Laurent Itti Lab, USC) The data are recordings of eye movements of subjects watching video clips under natural free viewing conditions. Data will be made available in both raw and processed forms, along with the corresponding video stimuli. Code will be provided for calibration of traces. Software, training data, and validation data will be provided to facilitate the development of prediction algorithms. These data were originally collected for development of an information-theoretic model of visual saliency and visual attention. It is anticipated that they will be useful for a broad range of questions in neuroscience, cognitive psychology, and computer vision. Saliency maps and raw feature maps tied to the information-theoretic model will also be made available, to allow users interested in quantifying which low-level visual features may more strongly attract human attention and gaze to easily perform quantitative analyses.

Core Services Supporting the CRCNS Data Sharing Program

The core facility (Fritz Sommer lab, Redwood Center for Theoretical Neuroscience, UC Berkeley) will provide access to shared resources in a manner that scales up to large data sets. Services will be designed to lessen the burden on contributors to make their data or other resources available and to optimize the ability of the user community to identify and use those resources. Community- and market-oriented mechanisms will be developed to identify resources of particular significance for the field, and to solicit feedback from relevant communities.

Summary and Conclusions

This article summarizes the views of experimental and theoretical neuroscientists, experts in computer science, governmental observers and a legal expert, who convened in June 2007 to discuss ways to advance public data sharing for computational neuroscience. The workshop revealed that although some types of neuroscience data have become publicly available, neurophysiology data are still scarce but are essential for improving computational techniques and models. Thus, it was concluded that an initial focus on neurophysiology data is warranted. Conditions under which labs would be willing to share such data were discussed, and the required services, infrastructure and support to ensure data sharing with such focus is successful were outlined. It was recommended that a small core team should provide basic services for data contributors and data users, and it should be responsible for building the data repository. It was also suggested that a governance board consisting of independent scientists and experts connected to the fields of neuroscience and computer science should direct and guide the data sharing activity. Finally, the NSF funded CRCNS data sharing program has been implemented on the basis of recommendations derived from the workshop, and brief descriptions of data sets that will be made publicly available in 2008 were provided. Our hope is that this program will spark broad interest from experimental labs seeking funding for data sharing activities, as well as from theorists in need of high quality experimental data.

Acknowledgements The authors thank the other participants of the workshop: Giorgio Ascoli (George Mason University); György Buzsáki (Rutgers University); Mike Hasselmo (Boston University); Judith Hirsch and Laurent Itti (both from University of Southern California); John Hogenesch (University of Pennsylvania); Terran Lane (University of New Mexico); Dario Ringach (University of California, Los Angeles); Raphael Ritz (International Neuroinformatics Coordination Facility, Stockholm); and Tim Blanche, Yang Dan, Jack Gallant (all from University of California, Berkeley). Thinh Nguyen (Science Commons, Cambridge, MA) provided legal expertise. Government observers (from various agencies) were: Christopher Greer, Dennis Glanzman,Yuan Liu, Peter Lyster, Michael Marron, Rae Silver and Ken Whang. Most of the ideas in this article have emerged from discussions at the workshop. Tim Blanche and Martin Rehn provided helpful feedback on drafts of this article. The authors thank numerous other researchers who provided thoughtful responses to an initial request for comments on data sharing issued by NSF in March 2007. The work described in this article was supported by NSF Grant 0636838 to K. D. Harris, Rutgers University, and by NSF Grant 0749049 to F. T. Sommer, UC Berkeley.

References

Ascoli, G. A. (2006). The ups and downs of neuroscience shares. *Neuroinformatics*, *4*, 213–215.

Bjaalie, J. G., & Grillner, S. (2007). Global neuroinformatics: The international neuroinformatics coordination facility. *Journal of Neuroscience*, *27*, 3613–3615.

Gardner, D. (2004). Neurodatabase.org: networking the microelectrode. *Nature Neuroscience*, *5*, 486–487.

Gardner, D., Toga, A. W., Ascoli, G. A., Beatty, J. T., Brinkley, J. F., Dale, A. M., et al. (2003). Towards effective and rewarding data sharing. *Neuroinformatics*, *1*, 289–294.

Harris, K. D., Csicsvari, J., Hirase, H., Dragoi, G., & Buzsaki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature*, *424*, 552–556.

Harris, K. D., Henze, D. A., Csicsvari, J., Hirase, H., & Buzsaki, G. (2000). Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *Journal of Neurophysiology*, *84*, 401–414.

Harris, K. D., Henze, D. A., Hirase, A., Leinekugel, X., Dragoi, G., Czurko, A., et al. (2002). Spike train dynamics predicts theta-related phase precession in hippocampal pyramidal cells. *Nature*, *417*, 738–741.

Insel, T. R., Volkow, N. D., Landis, S. C., Li, T. K., Battley, J. F., & Sieving, P. (2004). Limits to growth: Why neuroscience needs large-scale science. *Nature Neuroscience*, *7*, 426–427.

Kennedy, D. N. (2006). Where's the beef? Missing data in the information age. *Neuroinformatics*, *4*, 271–273.

Koetter, R. (2004). Online retrieval, processing and visualization of primate connectivity data from the CoCoMac database. *Neuroinformatics*, *2*, 127–136.

Liu, Y., & Ascoli, G. A. (2007). Value added by data sharing: Long-term potentiation of neuroscience research. *Neuroinformatics*, *5*, 143–145.

Van Horn, J. D., & Ishai, A. (2007). Mapping the human brain: New insights from fMRI data sharing. *Neuroinformatics*, *5*, 146–153.