Putting big data to good use in neuroscience

Terrence J Sejnowski, Patricia S Churchland & J Anthony Movshon

Big data has transformed fields such as physics and genomics. Neuroscience is set to collect its own big data sets, but to exploit its full potential, there need to be ways to standardize, integrate and synthesize diverse types of data from different levels of analysis and across species. This will require a cultural shift in sharing data across labs, as well as to a central role for theorists in neuroscience research.

Big data, the buzz phrase of our time, has arrived on the neuroscientific scene, as it has already in physics, astronomy and genomics. It offers enlightenment and new depths of understanding, but it can also be a bane if it obscures, obstructs and overwhelms. The arrival of big data also marks a cultural transition in neuroscience, from many isolated 'vertical' efforts applying single techniques to single problems in single species to more 'horizontal' efforts that integrate data collected using a wide range of techniques, problems and species. We face five main issues in making big data work for us.

First, data in neuroscience exist at an astonishing range of scales of both space and time. Neuroscientific data are obtained from a wide range of techniques, from patch clamping to optogenetics to fMRI (**Fig. 1**). Most of these techniques are used one at a time. One lab will record spikes from an array of neurons, but not be able to determine which types of neurons they are or how they are connected to other neurons. Another lab will reconstruct the wiring diagram of the same circuit, but without recording data to identify the properties of the reconstructed neurons. In some heroic cases, functional data have been laboriously combined with anatomical reconstructions[1], but rarely if ever in a broad behavioral context.

Different techniques differ also in concepts and vocabularies, in background assumptions and experimental norms. Decision-making, for example, might be studied at the level of populations of single-cell recordings in monkeys or by fMRI in humans or by lesions in rats or by molecular and optical techniques in mice. These differences mean that standardization in neuroscience must be made relative to a technique and that cross-level and cross-technique data integration cannot easily be automated. Standardizing data collected with a single technology is not trivial, making meaningful causal relationships among data sets obtained with very different technologies even more difficult to achieve.

Second, different animal models are used to study different problems: flies, worms, fish, mice, rats, monkeys and humans all have their place. It is often unclear how to extrapolate from worm data to a mammalian nervous system, for example, or from *in vitro* preparations to *in vivo* preparations. Each model has its distinct virtues, and new efforts to integrate information across species and technologies

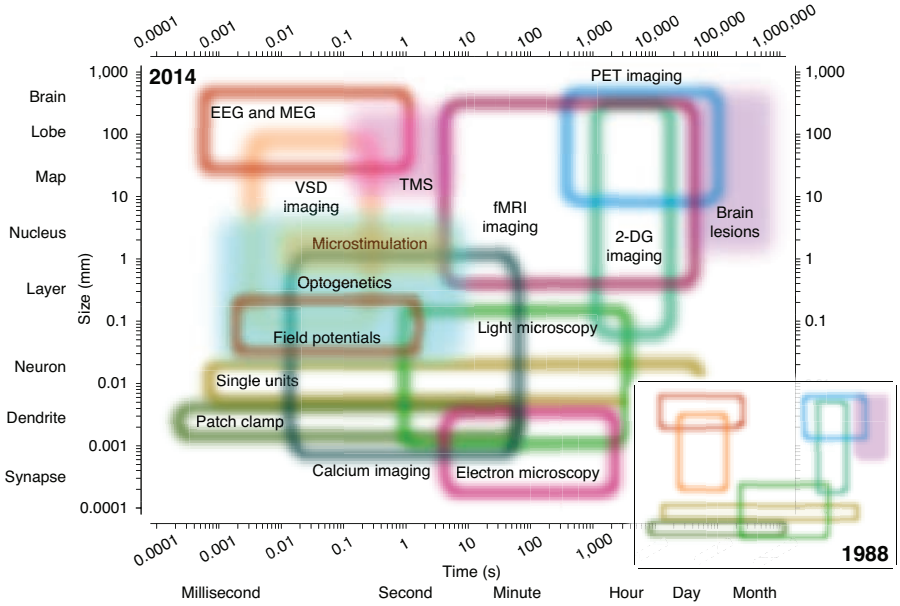

Figure 1 The spatiotemporal domain of neuroscience and of the main methods available for the study of the nervous system in 2014. Each colored region represents the useful domain of spatial and temporal resolution for one method available for the study of the brain. Open regions represent measurement techniques; filled regions, perturbation techniques. Inset, a cartoon rendition of the methods available in 1988, notable for the large gaps where no useful method existed[9]. The regions allocated to each domain are somewhat arbitrary and represent our own estimates. EEG, electroencephalography; MEG, magnetoencephalography; PET, positron emission tomography; VSD, voltage-sensitive dye; TMS, transcranial magnetic stimulation; 2-DG, 2-deoxyglucose.

Terrence J. Sejnowski and Patricia S. Churchland are at the Howard Hughes Medical Institute, the Salk Institute for Biological Studies, La Jolla, California, USA. Terrence J. Sejnowski is also in the Division of Biological Sciences, University of California at San Diego, La Jolla, California, USA, and Patricia S. Churchland is in the Department of Philosophy, University of California at San Diego, La Jolla, California, USA. J. Anthony Movshon is at the Center for Neural Science, New York University, New York, New York, USA.
e-mail: terry@salk.edu

may pay off handsomely. But this will require a deepened appreciation of comparative and evolutionary neurobiology.

It has been said that "nothing in neuroscience makes sense except in the light of behavior"[2]. Traditionally, neuroscientists have restricted the range and richness of behavioral measurements to keep the collection and interpretation of correlated data from neurons manageable. This strategy constrains our understanding of how the brain supports the full range of behaviors. Big data is making it possible to record from the same set of neurons while the subject engages in a much richer set of behaviors. Behavioral research will greatly benefit from the application of machine learning techniques that allow fully automated analysis of behavior in freely moving animals[3–5]. The challenge is to discover the causal relationships between big neural data and big behavioral data.

Third, as things stand in neuroscience, integration of functional data is mainly tackled by individual labs and by those with whom they collaborate. Such a strategy of 'every tub on its own bottom' depends on individuals to absorb information, communicate with others in the same subfield, and otherwise keep up. Meetings, lab visits, publications, review articles and so forth have been the mainstay of this form of integration. Although powerful and productive and a source of innovation, this style has limits. With increases in numbers of laboratories and publications, it is hard for individuals to keep up with the latest technology and harder still to keep data from slipping into oblivion, including data whose significance can be appreciated only later when the science catches up with the technology. This will require a cultural shift in the way that data are shared across labs.

Note too that this kind of integration is essentially vertical, in the sense that it is largely directed toward one particular problem, going up and down the organizational levels on that problem. Horizontal integration of data across a range of problems—for example, learning, decision-making, perception, emotion and motor control—is even harder to achieve in one laboratory. There is just too much data for one laboratory to get its collective head around.

A goal of the BRAIN Initiative[6] is to record and manipulate a large number of neurons during extended, behavioral experiments, to identify the neurons recorded from, to reconstruct the circuit that gave rise to the activity, and to relate the combined data to behavior—all in the same individual. Although this may seem like a pie-in-the-sky experiment, it is within reach in some species, such as the nematode worm *Caenorhabditis elegans*, whose neuronal connectivity is already known, and the transparent larval zebrafish, where it is possible to record simultaneously from most of its 100,000 or so neurons. To accomplish these ambitious goals will take teams of closely coordinated researchers with complementary expertise.

Fourth, as data sets grow and become more complex, it will become more and more difficult to analyze and extract conclusions. In the worst case scenario, the data may not be reducible to simpler descriptions. Here we need to rely on new approaches to analyzing data in high-dimensional spaces using pattern-searching algorithms that have been developed in statistics and machine learning.

To illustrate, consider the project of Vogelstein *et al.*[7], whose aim was to understand in *Drosophila* larvae the causal role of each of 10,000 neurons in producing a simple behavior in the animal's repertoire, such as turning or going backwards. Drawing on over 1,000 genetic lines and using optogenetic techniques to stimulate individual neurons in each line, they generated a basic data set consisting of correlations between stimulated identified neurons and a behavioral output. (Notice that the data set would have been far more massive had they stimulated neurons two or three or more at a time.) To find patterns in their huge accumulation of correlational data, they fed the data to an unsupervised learning program, which yielded a potential understanding of links between neurons and behavior. Correlational data could enhance understanding of the connectional structure to address questions of circuitry. Nevertheless, the methodological significance of the project is that is shows how new tools can be put to work to find patterns in data obtained from networks of neurons, patterns that emerge only from using new analytic tools on very large data sets.

The statistical design of these experiments will be critical to insure that data sets are carefully calibrated, are of sufficient power to admit analysis, and can be used by other researchers who want to ask different questions. This is not an easy process and requires a level of planning and quality control that goes beyond most exploratory experiments that are undertaken in most laboratories[8]. Here again, a modest cultural change can make a large impact.

Fifth, at some point along the Baconian rise of ever larger and more complex data sets, a deeper understanding should emerge from the accumulated knowledge, as it has in other areas of science. What we have today is a lot of small models that encompass limited data sets. These models are more descriptive than explanatory. Theory has been slow in coming. One obstacle is that sometimes theorists do not clearly convey what they propose, perhaps because they seek safety in needlessly complex mathematics or because they are too remote from the experimental base to undergird their theoretical ideas. Any of these issues can detract from productive ideas. This can change.

What we contemplate are modest cultural changes, wherein some neuroscientists are mainly theorists, with appropriate grant support to make the research feasible. The term "theorist" enjoys an uneven reputation in neuroscience, but serious scholars with this portfolio do now exist, although they tend to be in short supply. We need to cultivate a new generation of computationally trained researchers who are aware of the richness of data and can draw on knowledge from many laboratories, courageous enough to make judicious simplifications and to have their ideas tested, and imaginative enough to generate interesting, testable large-scale ideas.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Bock, D.D. *et al. Nature* **471**, 177–182 (2011).
2. Shepherd, G.M. *Neurobiology*. 8 (Oxford Univ. Press, 1988).
3. Dankert, H., Wang, L., Hoopfer, E.D., Anderson, D.J. & Perona, P. *Nat. Methods* **6**, 297–303 (2009).
4. Falkner, A.L., Dollar, P., Perona, P., Anderson, D.J. & Lin, D. *J. Neurosci.* **34**, 5971–5984 (2014).
5. Wu, T. *et al. IEEE Trans. Syst. Man Cybern. B Cybern.* **42**, 1027–1038 (2012).
6. National Institutes of Health. BRAIN 2025: a scientific vision. http://www.nih.gov/science/brain/2025/ (2014).
7. Vogelstein, J.T., Park, Y. & Ohyama, T. *Science* **344**, 386–392 (2014).
8. Mountain, M. *Phys. Today* **67**, 8–10 (2014).
9. Churchland, P.S. & Sejnowski, T.J. *Science* **242**, 741–745 (1988).