Analyzing neural data at huge scale

John P Cunningham

A new distributed computing framework for data analysis enables neuroscientists to meet the computational demands of modern experimental technologies.

Neuroscience, like many fields, is experiencing dramatic increases in the quantity and complexity of recorded data, with the goal of similarly dramatic advances in basic scientific understanding and applied biomedical technologies. Critical to realizing these ambitions are the analysis methods and computing infrastructure suitable for this scale of data. Thunder, a software library developed by Freeman *et al.*[1] and described in this issue of *Nature Methods*, offers neuroscientists a distributed data analytics platform that is both fast and scalable to data of massive size.

Neuroscientists have begun to adopt optical[2] and electrical[3] technologies that record up to thousands or millions of distinct signals from the brain simultaneously. Light-sheet microscopy[4] is one popular example capable of producing data sets upwards of 1 terabyte per hour; indeed, three other papers[5–7] using this technology also appear in this issue of *Nature Methods*. With these data, researchers are now testing a variety of hypotheses at the neural circuit and population level[8], many of which require analyses that consider large data sets jointly. However, even with continued improvements in computational power, many data sets already surpass the memory capacity of a single machine, and many data analyses are impractically slow. The future of large-scale neuroscience will require an alternative computing paradigm.

Recognizing this critical need, Freeman *et al.*[1] have developed Thunder for analysis of neural data at a massive scale. Neuroscience is by no means the only field facing this challenge, so Freeman *et al.* have wisely built their software library atop the distributed computing platform Spark[9], an exciting new implementation of the industry-standard MapReduce[10] concept. In a traditional setting, a single computer jointly analyzes a full data set of many neurons (which could also be voxels, electrical channels, etc.), a mode of operation that does not scale. Conversely, in a distributed computing setting, data are split into manageable groups of neurons and partitioned across a cluster of computers (**Fig. 1**), as is available in many institutional facilities and cloud services. The analysis method must be configured to first perform computations that are local to each neuron on the individual computers in parallel and to then transmit a summary of these operations (the 'Reduce' step). Another computer then updates global parameters on the basis of these collected summaries and distributes updated parameters across the cluster (the 'Map' step). Many analysis methods have well-known MapReduce implementations, including a wide range of statistical and machine learning algorithms that iteratively evaluate a model on many data points (for example, calculating the fit of a regression) and then update a model parameter (for example, regression coefficients).

Thunder, via the underlying Spark engine, handles many of the underlying complexities of distributed computing, allowing users to quickly and simply load data, retrieve results, analyze data with existing analyses, and implement methods of their own. Freeman *et al.* made at least four key choices that have resulted in an exceptional computing framework. First, they have included in Thunder many analyses common to neuroscience, including basic statistics, simple regressions, tuning-curve estimations, spatial and temporal matrix factorizations such as the singular value decomposition and independent-component analysis, and more, which allow Thunder to be immediately useful for a wide range of neural data.

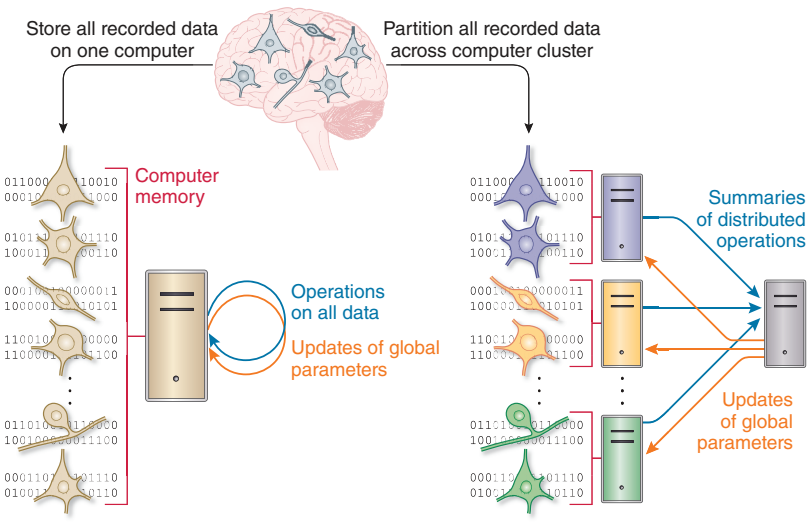

Figure 1 | Data sets recorded from the brain are growing in scale and require scalable analysis methods. The same data (different neurons with binary data) are analyzed on a single computer (left) and on a cluster with Thunder (right). When data grow past the memory capacity (left, red) of a single computer, or when computational time becomes impractically long, distributing this data across a computer cluster (right) confers massive gains in storage and computational efficiency.

John P. Cunningham is in the Department of Statistics, Columbia University, New York, New York, USA.
e-mail: jpc2181@columbia.edu

Second, their paper[1] thoroughly benchmarks Thunder, with comparisons across cluster sizes and to a stand-alone workstation implementation, enabling researchers to see the striking gains available with a distributed version of their analysis methods. Third, through Spark's programming interface, Thunder abstracts away most details of the distributed computing engine and offers a particularly simple front end (in Python) that should promote use. Fourth, Thunder is entirely open source and tightly integrated with Spark, choices that confer usability advantages, growing community support and notable performance gains over distributed computing alternatives (for example, some Hadoop projects).

The open question for Thunder is that of its reception and use by neuroscientists, as obstacles to adoption do exist. First, users must deploy Thunder across a computing cluster, which in many cases requires involvement of system administrators. Second, despite the best-in-class abstractions used by Thunder and Spark, distributed computing still requires an adjustment in algorithmic thinking and the use of programming languages outside the toolkit of many neuroscientists. Freeman *et al.* have done an excellent job of minimizing these obstacles by including demonstration websites, automatic install scripts, sample code, sample data sets and more.

At a broader level, the importance of large-scale computational strategies should not be underestimated. In April 2013, US President Obama announced the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative as one of the grand challenges for the 21st century. Parallels have been drawn to the Human Genome Project, which, although different from the BRAIN Initiative in many ways, shares the key similarity that large-scale data analysis was essential[11]. The US National Institutes of Health working group on the BRAIN Initiative stated seven key goals for this grand challenge, of which two were "large-scale monitoring of neural activity" and "development of new theoretical and data analysis tools"[12]. Accordingly, the question seems to be not if, but when, advanced computational frameworks will become the norm for neuroscience. Thunder stands as an important first step in this direction.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Freeman, J. *et al. Nat. Methods* **11**, 941–950 (2014).
2. Kerr, J.N.D. & Denk, W. *Nat. Rev. Neurosci.* **9**, 195–205 (2008).
3. Kipke, D.R. *et al. J. Neurosci.* **28**, 11830–11838 (2008).
4. Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M. & Keller, P.J. *Nat. Methods* **10**, 413–420 (2013).
5. Vladimirov, N. *et al. Nat. Methods* **11**, 941–950 (2014).
6. Amat, F. *et al. Nat. Methods* **11**, 951–958 (2014).
7. Mickoleit, M. *et al. Nat. Methods* **11**, 919–922 (2014).
8. Cunningham, J.P. & Yu, B.M. *Nat. Neurosci.* doi:10.1038/nn.3776 (in the press).
9. Zaharia, M. *et al.* in *Proc. Netw. Syst. Des. Implement.* **9** (USENIX, 2012).
10. Dean, J. & Ghemawat, S. *Commun. ACM* **51**, 107–113 (2008).
11. Collins, F.S., Morgan, M. & Patrinos, A. *Science* **300**, 286–290 (2003).
12. Bargmann, C. *et al.* Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group Report to the Advisory Committee to the Director, NIH http://www.nih.gov/science/brain/2025/index.htm (NIH, 2014).