The Big Data Problem: Turning Maps into Knowledge

Florian Engert[1,*]
[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA
*Correspondence: florian@mcb.harvard.edu
http://dx.doi.org/10.1016/j.neuron.2014.09.008

In this NeuroView, Engert discusses the challenges for the connectomics field in making insights about brain function from big data.

There has been a great deal of focus in recent years on efforts to map the brain. The ability to record from every neuron in the brain of an awake, and ideally behaving, animal is unquestionably immensely useful. In addition, having a wiring diagram at hand that can be overlaid on such activity maps is probably a dream come true for most systems neuroscientists. Given the vast number of neurons in the brain, however, such systematic analysis could yield enormous reams of data. The same could be said for efforts in the connectomics field to reconstruct structural connections throughout the brain via EM. Here, I argue that "big data" and the oft-discussed challenges inherent to it (e.g., mining, storing, and distributing it) is not the key challenge we face in transitioning from making neural maps to making useful insights into brain function. I would suggest that the essential ingredient that turns a useless map into an invaluable resource is the experimental design employed to gather and analyze the underlying data, and ultimately the thought process, creativity, and ingenuity that went into this design. This is where the hard work is—in formulating precisely the question of what we actually want to know, what an answer would look like, and what kind of insight we can take away from the experiment.

In this essay I will focus on two endeavors that are presently underway in the neurosciences that aim to collect rather large amounts of data: the Open Connectome Project (Burns et al., 2013; Kandel et al., 2013) and the BRAIN initiative (Devor et al., 2013; Kandel et al., 2013; Striedter et al., 2014). While it has been suggested that a critical challenge to be addressed with these initiatives is the issue of "big data" (Brinkmann et al., 2009; Choudhury et al., 2014; Swain et al., 2014), I will make the argument that it will be comparatively small data sets (on the order of a few terabytes at most) that will contain the relevant information and need to be distributed and made available as resources to the community. These small and information-rich data sets will include a description of all the neurons in the brain, their activity, and, ideally, their wiring diagram. The development of the methodologies necessary to generate these data sets is essential, it is important—and it is very difficult to do. But the difficulty lies primarily in developing the right technology. Overcoming these problems is essentially the goal of the BRAIN initiative and, in my opinion, a good place for investing money, energy, and time.

Big Data in Neuroscience?

Big data is a hot topic these days, and it's not surprising that there is discussion in the community about what to do with the data generated by these endeavors. Big data can be defined in many ways, and the continuous increase in computational power leads to a somewhat amorphous concept of what we mean when we talk about big data. For the purposes of this commentary, I will define as big data anything that exceeds the size of a standard laptop hard drive.

It is useful and important to make a definitive distinction between big data and complex data, however, two concepts that frequently get mixed up. The former is just that: big. The latter is complicated, hard to interpret, and—usually—very hard to compress. It also requires the application of mathematical tools and quantitative methods to analyze. Complex data sets, quite often, are not big in the sense of "big data," but they are ubiquitous in modern science.

How Big Is a Connectome?

Let's consider the respective challenges of converting data into information within the connectome project and the BRAIN initiative. Connectomics relies on recovering a circuit diagram by imaging the whole region of interest at the resolution of an electron microscope (EM) (Briggman and Bock, 2012; Kleinfeld et al., 2011; Lichtman and Denk, 2011; Randel et al., 2014). These EM data sets then need to be analyzed by segmentation and reconstruction of the individual neurons, which ultimately allows the identification of all the synaptic connections. The final product is the circuit diagram of the complete network in the volume under scrutiny. The size of the raw data collected in such an enterprise is truly daunting.

Let us look at a few numbers: a mouse brain imaged at 5 nm × 5 nm × 40 nm resolution at a volume of approximately 500 mm^3 would generate a raw data volume of 500 petabyte. Big data, indeed. However, what we want to get out of this volume is the connectivity matrix among the 100 million neurons that a mouse brain contains. If we assume ~1,000 connections for each neuron, the resulting connection matrix contains ~10^{11} entries. Assuming a bit depth of a few bytes, these 10^{11} entries result in a data set of a few hundred gigabytes, which will fit comfortably on an ordinary laptop hard drive. Complex data, but not big. It is true that we haven't yet developed fast, reliable, and efficient segmentation and tracing algorithms to actually do the segmentation and tracing—and as such this particular problem of data compression is far from

being solved. However, the solution to this problem will come most likely out of machine vision research and doesn't quite have the flavor of "big data mining." The task of segmentation and tracing itself is actually quite straightforward; it is easy to formulate and can be accomplished by a trained middle school student (see, for example, Eyewire.org), it's just very hard to implement in computer algorithms at the moment (Jain et al., 2010; Turaga et al., 2010). However, once these algorithms have been developed, whole-brain EM volume data can be reduced and compressed by six orders of magnitude. Not so big data anymore. It is unquestionably important to allocate resources to solve this problem, but it is most likely going to be solved—in the end—by a handful of smart mathematicians and might not really require a national (or international) effort and billions of dollars. Once compressed in this manner—and converted into information—the data sets to be analyzed in the context of systems neuroscience questions will comfortably fit on a flash drive that you can carry in your pocket.

How Big Is an Activitome?

If we consider recording all the spikes in all the neurons of the brain, we can envision a similar compression. If we achieve such large-scale recording through some technology based on volume imaging (point- or sheet-scanning, spatial light modulation, etc.) coupled with genetically encoded activity indicators (GCaMPxx or voltage-sensitive protein), we are initially faced with similarly big data volumes: a mouse brain contains 500×10^9 cubic micron pixels (filling a volume of $\sim$500 mm^3), and if we want to record all of them for 20 min (1,000 s) at 1000 Hz, we again have 500 petabytes of raw data. Here, however, the initial compression is much more straightforward: you isolate all the cell bodies (100 million) and find the timestamps of all the fluorescence intensity spikes. With the assumption that all the neurons fire at an average rate of 5 Hz through the recording time period (probably an upper estimate since many neurons might be silent), we again end up with a data volume of 500 gigabytes. Quite manageable. Here, the mathematical tools to do this compression are more or less already in place. Segmentation of neuronal cell bodies and isolation of spikes from fluorescent traces is presently made difficult only by signal-to-noise problems. If the signals are large, this is easily done with the help of standard and ubiquitously available software.

Thus, in both cases, the size of the relevant data volumes can be reduced from hundreds of petabytes to a few hundred gigabytes, and this can be done by relatively straightforward analysis pipelines that are—at least intellectually—very straightforward. Furthermore, this data reduction will eventually be done on the fly, i.e., during the acquisition of the raw data, and will probably be achieved with dedicated hardware in the form of custom-designed coprocessors. Raw data sets might be very large, but once converted into information, the volumes aren't big data anymore.

Large-Scale, Small-Scale: A Question of Style

I've argued that the big data in question could, with appropriate analysis and technological developments, be relatively easily compressed into information, albeit complex. But the big data still must be gathered. So what's the best approach to collecting the data that will give us an unprecedented view into brain function? One could envision either large-scale, industrial data collection or the traditional small-scale, individual lab approach. Here, I will briefly discuss the potential contributions of both.

Whole brain imaging will greatly facilitate the identification and localization of essential neural subnetworks related to a behavioral context under scrutiny. The product or "deliverable" of whole brain imaging will then be a small and spatially identified subset of neurons that shows correlated activity with all—or any—aspect of the behavioral context. This is probably more useful than any other way of labeling subsets of cells if the goal is to decipher the roles of circuits in generating behavior. It offers an attractive and complementary approach to labeling neurons with genetic methods like enhancer trapping. The catch is that whole brain imaging has to be integrated into the experimental context and it has to be designed and optimized for the specific project. As such, it needs to be turned into a readily available technology for all laboratories and accessible on the small scale.

The issues are slightly different for connectomics, which has the goal of generating complete wiring diagrams that—ideally—can and should be overlaid onto previously acquired functional maps. Such an enterprise will require concerted and large-scale efforts and indeed might best be accomplished by industrially organized science at the more corporate level. Indeed, in recent years several voices have been raised that argue—occasionally quite convincingly—for neuroscience to move from tinkering in individual laboratories to industrial-scale research that allows for the many challenges to be tackled systematically and in a properly organized fashion.

I propose that there is equal space and opportunity for both: corporate-style/industrial-size science as well as the individual, small-scale, cottage industry style. Connectomics is clearly an example that is begging to be turfed out to a contract research organization (CRO), equipped with a park of various electron microscopes, where fixed brains can be automatically sectioned, mounted, imaged, and even segmented. Several successful service industries come to mind that all started out as relatively small-scale operations in individual laboratories and that are now used routinely by almost every laboratory in the world.

Sequencing services are being used ubiquitously around the world, yet the technology certainly started as some form of cottage industry by the likes of Sanger and colleagues. Oligonucleotide synthesis as well as protein sequencing is another powerful technology that quickly made it into a service industry. The generation of transgenic mice—a job that used to soak up a large part of a PhD thesis—is now in most cases outsourced to CROs. It is frequently observed that even the outsourcing of graduate student supervision occurs, in this case to thesis advisory committees and/or postdoctoral fellows.

Whole brain imaging, on the other hand, is difficult to envision as an industrial-scale, massively parallel high-throughput operation. The main reason for this is that such an operation usually requires a clear final product, a deliverable that can be quantitatively described, priced,

benchmarked, and specified by intermediate milestones. These features seem quite feasible in the context of generating connectomes but appear to be ludicrous in the context of whole brain imaging. What would such a product look like? Here, clearly the deliverable is the technology and not the final data set, and as such the aims of the BRAIN initiative are perfectly aligned with these objectives.

Looking to the Future

Once the data are collected and compressed into information, the question becomes how best to turn this information into knowledge. The challenge in the neurosciences will be to come up with good questions and intelligent experimental assays—assays that ultimately will have to be anchored in behavior and that will have to give answers to questions of how specific behaviors are generated by the nervous system. For excellent specific examples, it is useful to go further back in the history of neuroscience and consider stories like the jamming avoidance reflex (JAR) of the weakly electric fish (Heiligenberg, 1991) and the generation of rhythmic activity in the somatogastric ganglion of the lobster (Marder et al., 2014; O'Leary and Marder, 2014).

New technologies that allow us to identify and isolate the neuronal subtypes that are actually involved in a specific task will of course be an important boon to this enterprise, and they will undoubtedly speed up the collection of necessary data. However, I doubt that these new technologies will lead to a paradigm shift or a fundamentally new way of doing neuroscience. The name of the game will always be to think carefully and deeply about how behavioral features can emerge out of neuronally implemented algorithms, and ideally these ideas ought to germinate and take shape well before we actually start generating data, be it big or small.

REFERENCES

Briggman, K.L., and Bock, D.D. (2012). Curr. Opin. Neurobiol. *22*, 154–161.

Brinkmann, B.H., Bower, M.R., Stengel, K.A., Worrell, G.A., and Stead, M. (2009). J. Neurosci. Methods *180*, 185–192.

Burns, R., Roncal, W.G., Kleissas, D., Lillaney, K., Manavalan, P., Perlman, E., Berger, D.R., Bock, D.D., Chung, K., Grosenick, L., et al. (2013). The Open Connectome Project Data Cluster: Scalable Analysis and Vision for High-Throughput Neuroscience. Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM), 27. http://arxiv.org/abs/1306.3543.

Choudhury, S., Fishman, J.R., McGowan, M.L., and Juengst, E.T. (2014). Front Hum Neurosci *8*, 239.

Devor, A., Bandettini, P.A., Boas, D.A., Bower, J.M., Buxton, R.B., Cohen, L.B., Dale, A.M., Einevoll, G.T., Fox, P.T., Franceschini, M.A., et al. (2013). Neuron *80*, 270–274.

Heiligenberg, W. (1991). Neural Nets in Electric Fish. (Cambridge, MA: MIT Press).

Jain, V., Seung, H.S., and Turaga, S.C. (2010). Curr. Opin. Neurobiol. *20*, 653–666.

Kandel, E.R., Markram, H., Matthews, P.M., Yuste, R., and Koch, C. (2013). Nat. Rev. Neurosci. *14*, 659–664.

Kleinfeld, D., Bharioke, A., Blinder, P., Bock, D.D., Briggman, K.L., Chklovskii, D.B., Denk, W., Helmstaedter, M., Kaufhold, J.P., Lee, W.C., et al. (2011). J. Neurosci. *31*, 16125–16138.

Lichtman, J.W., and Denk, W. (2011). Science *334*, 618–623.

Marder, E., O'Leary, T., and Shruti, S. (2014). Annu. Rev. Neurosci. *37*, 329–346.

O'Leary, T., and Marder, E. (2014). Science *344*, 372–373.

Randel, N., Asadulina, A., Bezares-Calderón, L.A., Verasztó, C., Williams, E.A., Conzelmann, M., Shahidi, R., and Jékely, G. (2014). Elife, e02730. Published online May 27, 2014. http://dx.doi.org/10.7554/eLife.02730.

Striedter, G.F., Belgard, T.G., Chen, C.C., Davis, F.P., Finlay, B.L., Güntürkün, O., Hale, M.E., Harris, J.A., Hecht, E.E., Hof, P.R., et al. (2014). Brain Behav. Evol. *83*, 1–8.

Swain, J.E., Sripada, C., and Swain, J.D. (2014). Behav. Brain Sci. *37*, 101–102.

Turaga, S.C., Murray, J.F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., and Seung, H.S. (2010). Neural Comput. *22*, 511–538.