nature neuroscience

# The big data challenges of connectomics

Jeff W Lichtman[1,2], Hanspeter Pfister[2,3] & Nir Shavit[4,5]

**The structure of the nervous system is extraordinarily complicated because individual neurons are interconnected to hundreds or even thousands of other cells in networks that can extend over large volumes. Mapping such networks at the level of synaptic connections, a field called connectomics, began in the 1970s with a the study of the small nervous system of a worm and has recently garnered general interest thanks to technical and computational advances that automate the collection of electron-microscopy data and offer the possibility of mapping even large mammalian brains. However, modern connectomics produces 'big data', unprecedented quantities of digital information at unprecedented rates, and will require, as with genomics at the time, breakthrough algorithmic and computational solutions. Here we describe some of the key difficulties that may arise and provide suggestions for managing them.**

The sheer complexity of the brain means that, sooner or later, the data describing brains must transition from something that is rather easily managed to something far less tractable. This transition appears to now be under way. The accumulation of ever bigger brain data is a byproduct of the development of new technologies that provide digitized information about the structural organization and the function of neural tissue. These new collection approaches bring novel data into neuroscience, data that bears on many poorly understood aspects of the nervous system. Fundamental questions such as how learned information is physically stored in the brain, how psychiatric diseases affect brain structure and function, how genetic and environmental interactions influence brain structure and its variability, and how the brain changes over the course of development and aging may be usefully addressed in the coming decades as large data sets (perhaps in the petabyte range) describing high-resolution brain structure and function become available.

Unfortunately, the generation of large data sets is actually the (relatively) easy part. Our own experiences in the nascent field of connectomics indicate that there are many challenges associated with the steps after data acquisition, that is, the process of turning raw image data into a minable map of neural connectivity. We describe some of these challenges and provide a few potential strategies that may help overcome the big data difficulties ahead.

## Toward a theory of connectomics

The field of connectomics is so new that there is no consensus yet about its central aims or even its data acquisition strategies. All practitioners would probably agree, however, that connectomics will generate large amounts of data that concern the fine details of neural circuitry over large volumes; an unprecedented data trove. For most neuroscientists, such data is not only quantitatively, but also qualitatively, different from the kinds of information they have experienced previously. Thus, there is considerable uncertainty about what may be learned from this data, and for some researchers, the lack of a theory of connectomics is a show-stopper.

From our perspective, the uncertainties speak more to the opportunities than the shortcomings of connectomics. This point can be better understood by looking back several centuries to the efforts of the singular genius Galileo Galilei, who designed telescopes to look at the night sky. He was in a predicament that seems analogous to the one that faces neuroscience today. Many phenomena about the heavens were deeply mysterious, and existing theories were impossible to test with the available data. His new imaging technology was, at the outset, less than ideal. Moreover, it was not obvious what he should look at with his telescope or what kind of data he should obtain. And yet, Galileo found ways to use the limited data he could collect to refute the hypothesis that the Earth was the center of the universe. Ultimately, the value of the telescope was less its refutation of an existing theory than that it provided data that went beyond known theoretical frameworks. The discovery of galaxies, the expanding universe and many other phenomena were products of telescopes. These observations required new theories to make sense of them, and these new theories required acquisition of more new data, a virtuous cycle that led to the birth of modern astronomy.

To be sure, it would be much easier to figure out what kinds of data to obtain, and how to mine it, if we already had a well-developed theory of how networks of connections translate into brain function; unfortunately, this is not yet the case. Our own view is that, lacking a clear idea of what parts of the connectomic data trove will ultimately be relevant, we should err on the side of getting too much data rather than just the data that answers a particular question. This view, of course, compounds the big data problem by requiring as much resolution in the data as we can muster.

## Connectomic data

As organ systems go, there is none as physically complicated as the brain: it is far more heterogeneous as a tissue than any other organ. In part, because of this heterogeneity, the organizing principles and even cell types in each part of the brain vary substantially. Understanding

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA. [2]Center for Brain Science, Harvard University, Cambridge, Massachusetts, USA. [3]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. [4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to J.W.L. (jeff@mcb.harvard.edu).

Received 31 July; accepted 10 September; published online 28 October 2014; doi:10.1038/nn.3837

the relationship between the brain's structure and its function also requires attending to design principles over both small and large length scales[1,2]. At the small end, describing the cellular connections mediating neural signaling requires resolving neural tissue at the scale of nanometers. Electron microscopy images at the nanometer level reveal all of the synapses and intracellular details such as the numbers of synaptic vesicles, the size of postsynaptic densities, etc. The diffraction limit of visible light limits its resolution to several hundred nanometers, making standard fluorescence techniques too blurry to resolve synaptic connections, especially if one is trying to see all of the connections in a tissue sample. Although recent developments in fluorescence methods overcome the diffraction limit[3], these super-resolution techniques are not designed to image everything; rather, they gain much of their power by selectively labeling a small subset of molecules, organelles and/or cells. At some point, it may be possible to combine these new optical approaches with methods that label each neuron a different color (for example, Brainbow labeling[4,5]), but there are still technical hurdles to overcome along the way[6]. This is why many of the current connectomic techniques use electron microscopy instead: the short wavelengths of high energy electrons provide sufficient resolution to see the finest details of synaptic connections and the heavy metal stains show all cells and all organelles. Data about the biochemistry of the brain would also be very informative (such as the receptor subtypes at different synapses or the particular ion channels on the membranes of each neuronal process). The only reason that we do not add this data to the electron microscopy images is that we do not yet know how; efforts are now under way to solve this problem[7].

The approach we take is to generate electron microscopy images of a consecutive sequence of sections of the brain that are automatically cut as 30-nm-thin slices and picked up on a tape substrate by a conveyor belt. Each section is imaged with a scanning electron microscope at a resolution at which each pixel represents a 4 × 4 nanometers region of tissue. At this pixel density, a 1-mm$^2$ brain section requires acquiring a 62.5 gigabyte image. Obtaining throughput speeds commensurate with processing large volumes (tens of thousands of such sections or more) requires that the sectioning and image acquisition be carried out automatically with few interruptions. The sectioning step is far quicker than the imaging step (see below).

There are several alternative methods we could use for obtaining electron microscopy–level connectomic data over large volumes[1,8]. These include block-face methods that use either microtomes[9] or focused ion beams[10] to successively shave a layer of the sample block between acquiring images of its face, and transmission electron microscopy techniques that use camera arrays[11] or single digital detectors[12]. Each imaging approach has particular advantages: block-face techniques give better alignment between sections, saving computation time, whereas transmission techniques can provide better lateral resolution, which may aid in identification of fine structural details, and focused ion beams give thinner sections, which improve segmentation. The current methodological diversity means that connectomic data is not the same from one laboratory to another, which slows the development of analytic tools, as each type of data requires somewhat different processing before it is minable.

With the tools we are using, the current throughput of the image acquisition step is about 1 terabyte, or 16 images, 1 mm$^2$ each, per day. This translates to about 6 years of image data acquisition to complete a cubic millimeter of brain with 33,333 sections that are 30-nm thick. At that rate, even a cerebral cortex of a small rodent such as a mouse (112 mm$^3$, see ref. 13) could take more than 600 years, and a rat (253 mm$^3$, see ref. 14) would take 1,000 years, plus or minus. However, image acquisition rates are undergoing substantial speed-up as a result of, for example, new imaging microscopes that use multiple scan beams to parallelize image acquisition. The first such device was delivered to us this summer from Carl Zeiss. Such speed-up strategies may soon allow a microscope to generate in the range of 1,000 1-mm$^2$ sections per day, allowing images of 1 mm$^3$ to be completed in a little over a month.

Latent in such images is not only the wiring that connects each nerve cell, but also details such as the volume of every synapse, the number of synaptic vesicles at each synapse, the location of every mitochondrion, the glial cell investment of some synapses, the location of every node of Ranvier, etc. However, a number of analytic problems stand between the raw acquired digitized images and having access to this data in a useful form.

**Alignment.** For the tape-based method, each successive digitized image needs to be aligned with the previous and subsequent images. Despite being largely similar, image alignment is challenging because the sections are collected by a conveyor belt and each may rotate a few degrees, or stretch depending on its thickness. Fortunately, because of the high image resolution, alignment is practical, as axons and dendrites are readily visible in cross-section and can be traced from one section to the next. Block-face serial electron microscopy techniques have much better section-to-section alignment; thus, although some small alignment corrections are made, this step is less challenging from a computational standpoint.

**Reconstruction.** A second challenge is that, once the image data is aligned, the sectioned objects must be individuated. In these data sets, the objects are neurons and other cellular entities that are interwoven in the three-dimensional space of the sample tissue. The reconstruction of neural processes as they pass from one section to the next is directly related to the computer vision problem of obtaining a segmentation of an image series, that is, the labeling of pixels in the images according to which neuron or glial cell they belong to.

Although considerable progress has been achieved in computer-based image segmentation in the last few years, reliable automatic image segmentation is still an open problem. Automating the segmentation of connectomic data is challenging because the shapes of neural objects are irregular, branching, non-repeating and intertwined. Moreover, the actual number of different objects and their synaptic interconnections in a volume of brain tissue is unknown and, at the moment, even difficult to estimate or bound. Segmentation of a standard electron microscopy image is further complicated by the fact that the range of pixel intensity values of cell membranes overlaps with that of other organelles. Thus, simple thresholding to find cell boundaries does not work. Finally, in electron microscopy, imaging the lateral resolution of a section is sometimes several-fold finer than its thickness[1]. This anisotropy means that moving in the *z* axis between sections causes an object's membrane outline to move a greater distance than tracing it out in a section (the *x* or *y* axis).

The process of segmentation is easiest to appreciate by looking at an electron microscopy section. **Figure 1a** shows a small part of a 30-nm-thick section of cerebral cortex. This section contains the cross sections of many different axons, dendrites and glial cell processes. Using an automatic segmentation algorithm[15], all of these cross sections can be 'colored in' to form a saturated reconstruction (**Fig. 1b**). The quality of the automatic segmentation can be appreciated by comparing a computationally generated image (**Fig. 1c**) with a human-traced segmentation of the same section (**Fig. 1d**). Although many objects are segmented similarly by both methods, it is obvious that, in the automatic method, objects are segmented into multiple objects (split errors) or erroneously combined (merge errors) where

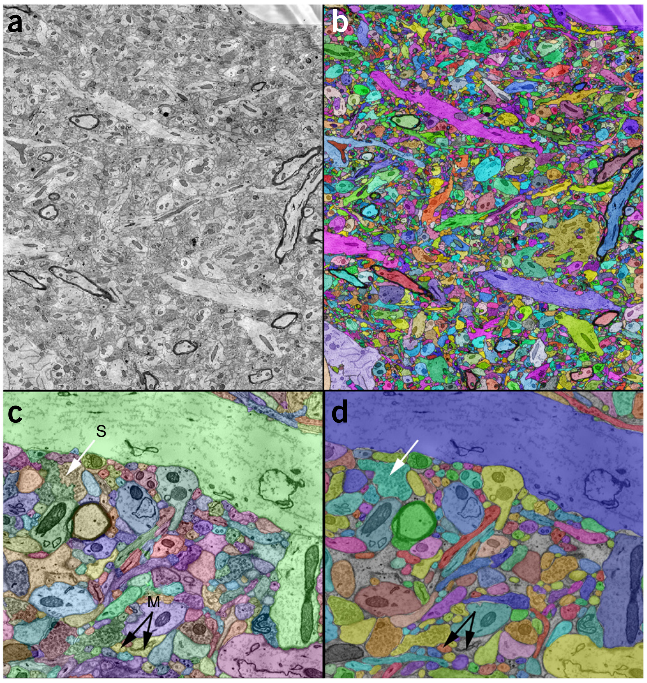


**Figure 1** Segmenting brain images. (**a**) Shown is an electron micrograph of a small part of a 30-nm-thick section of mouse cerebral cortex. Even though the region shown is only 40 × 20 µm and less than 1,000th of the area of 1 $mm^2$, it contains the cross-sections of more objects than are practical to identify by the human eye. (**b**) Automatic computer-based methods, however, can attempt to completely segment such data (that is, saturate the segmentation). The result of the segmentation is shown as an image overlay with a distinct color (ID) for each cellular object. In subsequent sections (not shown), the same color is used for the same objects. (**c**,**d**) Higher magnification image of a small part of the automatically segmented image from a subsequent section (**c**) shows that the automatic method makes errors when compared to human tracing (**d**). The white arrow labeled S in **c** shows that a vesicle-filled axonal profile is split into several compartments, whereas the same axonal profile in **d** is labeled correctly as one object (compare the white arrow pairs in **c** and **d**). The black pair of arrows labeled M in **c** show two objects that have been merged into one by the automatic segmentation algorithm, whereas the objects are correctly labeled as separate in **d** (compare the black arrow pairs in **c** and **d**).

a human tracer would not make the same mistakes. At present, the error rates of automatic segmentation are on the order of one merge or split mistake per cubic micrometer of tissue volume, which is clearly inadequate for reliable segmentations. Further algorithmic work is needed, as well as perhaps new tissue preparation, tissue labeling and imaging strategies that provide better raw image data to work with. Notably, two phenomena mitigate the seriousness of most of the errors. First, errors of omission are more common than actual crossed wires. Second, the errors are mostly confined to the finest terminal processes of axons (terminal single-synapse branches) and dendrites (single spines).

The greatest challenge facing connectomics at present is to obtain saturated reconstructions of very large (for example, 1 $mm^3$) brain volumes in a fully automatic way, with minimal errors, and in a reasonably short amount of time. The human visual system has little trouble identifying objects and tracing them in subsequent sections. Indeed, attempts to use human segmentation to do connectomic work can succeed provided a large number of people (thousands) can be persuaded to work on a data set[1,16]. Our own efforts suggest that human tracers rarely disagree (<0.1% of connectivity) when tracing the same data set independently and quickly reach consensus when results are compared (N. Kasthuri and D. Berger, unpublished data). This consensus, of course, does not mean that human tracing is without error; it is just that, at this moment in time, we have no way to detect the inaccuracies. However, even if human tracings were perfect, the data size would eventually require the equivalent of millions of human tracers[17], and the rapid data acquisition rates that will soon be available will make it necessary to do the reconstruction at the acquisition rate of the microscope and to keep the computation close to where the data is being acquired, which will require new computational strategies.

There is currently some difference of opinion as to how close research is to generating fully automatic segmentation with accuracies that approach that of human tracers. For example, some experts believe that the gap between human-based segmentation and machine-based approaches "is unlikely to close soon"[17]. They liken the problem to that of recognizing human cursive handwriting by machine[18]. Because no two handwriting styles are identical, very large training sets from many individuals are typically required for automatic methods, and these methods do not yield good results in the end.

We believe that the problem of segmenting the wiring of, say, the cerebral cortex of an adult mouse, may be more constrained. Using the handwriting metaphor, deciphering neural wiring in one species may be closer to deciphering the cursive script of one individual, as opposed to that of coming up with a generic solution for all individual writing styles. As complicated as it may seem, the neural wiring of one species, at one age and maybe in one part of the nervous system, has a limited and ultimately identifiable set of geometries, and many constraints. To be sure, there will probably always be places that do not fit the general pattern, and more computationally intensive techniques may need to be used in such instances (see below). However, we suspect that the exceptionally difficult parts will be a small fraction of the overall data set.

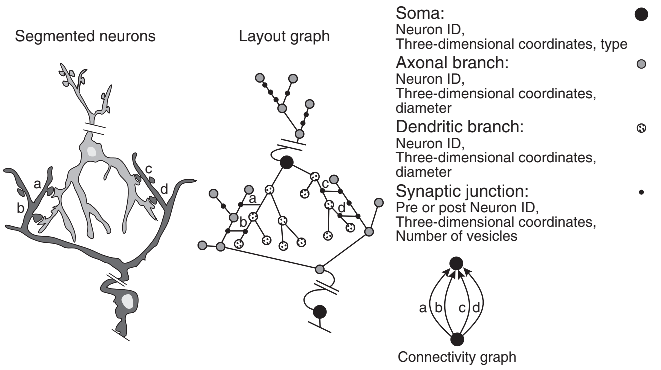
Our optimism at getting automatic solutions to segmentation is deeply rooted (ironically) in our view about how human brains go about solving difficult tasks. Once the arduous task is learned (riding a bicycle, reading, etc.), the brain does its work with little cognitive effort: the 'automatic' solutions become efficiently embedded in the structure (probably the wiring diagram) of the brain.

**Feature detection.** In addition to segmentation, another computer vision challenge in connectomics is the detection of important subcellular features such as mitochondria, synaptic vesicles, and the pre- and postsynaptic specializations at various kinds of synapses. Although the automatic detection of cellular boundaries is a necessary part of image segmentation, additional methods need to be employed to detect these subcellular features. Recently developed synapse detection approaches[19] show promise, but, as with segmentation, error rates and analysis time need to be reduced for practical use in big data volumes.

**Graph generation.** Finally, the reconstructed data needs to be turned into a form that represents the wiring diagram. The scientific value of the resulting connectivity graph may depend on how much of the original data is retained. For example, anatomical details such as the location of synapses along a dendrite, sizes of synapses or the caliber of nerve cell processes may be valuable in using the connectivity data to simulate function in a circuit. On the other hand, to make the connectivity easily minable, some substantial data reduction must occur (see below).

**Figure 2** shows an example of what a wiring diagram resulting from the reconstruction process might look like. The image on the left is

**Figure 2** Transformation of segmented data into a connectivity graph. Left, in this schematic, the segmented axon of the dark gray neuron is found to establish four synapses on the dendritic spines of the light gray neuron (a–d). Middle, the segmented data is transformed into a layout graph that keeps notation of the location of every axonal and dendritic branch point and every synaptic junction. The layout graph has far less data than the segmented images that were used to generate it. Circles of different sizes and shading represent identification tags (IDs). Right, ultimately, the connectivity can be graphed quite simply, as shown here.



a schematic depiction of two neuronal objects reconstructed from a serial electron microscopy data set. For the purpose of illustration, we include four synaptic junctions between the neurons. The neurons, after reconstruction, were transformed into a layout graph representation whose nodes are dendritic or axonal branches, with associated information such as their three-dimensional coordinates, diameter, cell type, etc., and edges between them that could be marked with the direction of signal flow, etc. The edges also contain lists of associated synaptic junction nodes, each with information such as three-dimensional location, number of vesicles, etc. Synaptic junction nodes can have one or more incoming or outgoing edges connecting them to the corresponding synaptic junction nodes of other neurons. Finally, from the layout graph, we can generate a connectivity graph whose nodes are neurons and whose edges represent synaptic junctions between them. The edges can be tagged with information such as connection strength or delay time, which are derivable from the layout graph given biophysical knowledge about neural signaling. Obviously, with serial electron microscopy, we only extract static information, but nothing prevents the incorporation of dynamic properties collected from other modalities. Ultimately, such a graph would contain a listing (ontology) of the types of things the data contains. This list is an essential feature for analysis of the regularities in the data. An example of this kind of analysis for the 302 neurons in the worm *C. elegans* hints at the way these kinds of data will look[20].

The layout graph itself could be stored using data structures for representing three-dimensional data. Such data structures, such as Skip Oct-trees[21], allow representation of the locations of objects in three dimensions so that they can be queried quickly. They can, for example, be used to answer queries such as how many neuronal somas are located in the volume defined by the dendritic arbor of a given neuron or which dendritic spines are within a certain distance to a given axonal branch. The software implementations of these data structures will be highly 'concurrent', that is, allowing multiple processors to share the data to do searches or analysis simultaneously and with little overhead[22].

In all of these steps, from tissue sectioning, image acquisition, alignment, segmentation, feature detection and graph analysis, a central mantra is 'dehumanizing' the pipeline, in the sense that all of these steps will scale and be more efficient when there is less human involvement. Although we are making progress in getting humans out of the workflow, we still have much more to do. One of the ironies of connectomics is that humans are especially good at these kinds of tasks (ones that require manual dexterity such as ultrathin sectioning or image analysis such as segmentation) because of the way our brains are wired. If we knew how brains worked, it might be far easier to develop the tools that would allow us to automate all these processes and learn how brains are wired!

## Big data challenges of connectomics

In the eyes of many, the term big data is synonymous with the storage and analysis of massive collections of digital information. The term big refers to the size of the input sets, typically ranging in the tens or even hundreds of terabytes, and their arrival rates of several tens or hundreds of gigabytes per second.

**Data size.** In connectomics, the size of the input set is at the high end of the big data range, and possibly among the largest data sets ever acquired. As already mentioned, images at resolutions of several nanometers are needed to accurately reconstruct the very fine axons, dendrites and synaptic connections. Thus, acquiring images of a single cubic millimeter of a rat cortex will generate about 2 million gigabytes or 2 petabytes of data. A complete rat cortex, including some white matter, might require 500 mm$^3$ and would produce about an exabyte (1,000 petabytes) of data at the aforementioned resolution. This amount is far beyond the scope of storage that can be handled by any system today (as a reference point, consider that the database system of the Walmart department store chain, one of the largest in the world, manages a few petabytes of data). A complete human cortex, ~1,000-fold larger that of a rodent, will require a zetabyte (1,000 exabytes) of data, an amount of data approaching that of all the information recorded globally today.

**Data rate.** The new electron microscope that we have started using will have a staggering throughput approaching several terabytes of data per hour, placing it at the far end of the big data rate spectrum. This rate, if matched with appropriate reconstruction algorithms (still a big if), will allow us to process a cubic millimeter of rodent brain in about 800 h (2,000 terabytes, 2.5 terabytes per h). However, at this rate, a complete mouse cortex will still require at least a decade if only one machine is doing the imaging, and it will take a professional lifetime to complete a rat cortex. That speed could be multiplied if the task were distributed to multiple labs with imaging microscopes working in parallel; ten multibeam microscopes working in parallel could do a mouse cortex in a year at this high resolution. This leads us to two conclusions.

The first is that, without further speed-up in image acquisition, it will not be possible to acquire the complete connectome of a human cortex, and although reconstructing a small mammalian cortex is not out of the question, it will be a major undertaking. We must therefore consider the possibility of reconstructions of neuronal substructures as opposed to whole brains and hope that testing these substructures will reveal enough modularity and regularity to allow deduction of interesting general organizational principles and overall function.

The second is that whatever algorithm is used to extract the connectomics graph from the image data will eventually have to work on the fly, at the pace of the microscope that generates this data. For starters, storing and processing the data later will require storage of the size of the input data, which is petabytes in size even for 1 mm$^3$. This is because a smaller storage buffer, less than the size of the whole data, would eventually overflow[1]. One can store 2 petabytes of data in a rack of disk drives, or a wall of tapes, but this approach does not seem to scale well (a few cubic millimeters will require a room, a mouse brain will require a dozen rooms). The bigger problem is that even if we do store the data for later processing at a slower rate, how much of a slowdown can we reasonably tolerate? If the data were generated in a month, a tenfold slowdown would mean processing it in about a year. Perhaps this is reasonable for the first cubic millimeter, but seems unreasonable as a general technological approach, as researchers will want to apply it to a growing number of samples without spending a year of their life waiting for each (for example, generating connectivity graphs for 10 1-mm$^3$ regions in a cortex would take a laboratory a decade of computation to complete).

**Computational complexity.** The goal of many big data systems is more than to simply allow storage and access to large amounts of data. Rather, it is to discover correlations within this data. These correlations, the desired outputs of big data algorithms, are typically many times smaller than the original input sets that they are derived from and, notably, can often be extracted without computing on the entire data set.

In computer science, this relation between the sizes of the input and output of a computational problem is captured using the notion of asymptotic complexity. Consider the computational problem of determining, in a given country, the correlation between the life expectancy of a smoker and the decade of their life when they began smoking. Whether we are talking about Andorra, with a population of $N = {\sim}10^5$, Cyprus with $N = {\sim}10^6$ or India with $N = {\sim}10^9$, the correlation, that is, the output of the computation, would remain ten pairs of numbers; these outputs do not grow as $N$, the size of the input set, grows. In computer science terms, one says that the size of the output set is asymptotically constant in (that is, compared to) the size of the input set.

For many big data correlation problems, the asymptotic size of the output is constant, implying that there is no limitation to developing algorithms on the basis of sampling approaches. These algorithms select a subset of the input set, either randomly or based on some simple-to-test criteria, and then compute an approximation of the output on the basis of this smaller subset. For our above example, there are statistical tests, based on small population samples, that quite accurately estimate the degree of correlation between the life expectancy and age of smoking onset.

In connectomics on the other hand, the graph we need to extract from the microscope images grows in direct proportion to the size of this data set because we are not sampling; rather, we are extracting a representation of the entire data set. In computer science terms, one says that the size of the output graph, and thus the computation to extract it, are at best asymptotically linear in (that is, compared to) the size of the input set. In this sense, our problem is much harder than many traditional big data problems and implies that, as we improve our microscopy techniques and increase the rate of data generation, we will have to proportionally increase the pace of our reconstruction computation.

**Parallel computing.** Fortunately, there is also some good news. Consider the cortex of a small mammal. It may contain about 20 million neurons and perhaps 10,000-fold more synapses. In a straightforward graph representation, these 20 million neurons would require 20 million nodes and 200 billion edges, representing the synapses. We should think of the overall graph data structure, even with additional data, as being proportional to the number of edges. Thus, assuming that each annotated edge or node requires about 64 bytes (a couple of 'cache lines' of data), we are talking about 8 terabytes of data to represent a small mammal's complete cortical connectome graph. Such a graph can be wholly contained in the memory of today's server systems. This data reduction is essential: operations executed in the computer's memory are orders of magnitude faster than those requiring access to disk, implying that, once the graph is in memory, it can be analyzed and used for simulations efficiently. The initial target of collecting a cubic millimeter, one thousandth of small mammalian cortex, would only require 8 gigabytes of memory, so graphs of even 10 or 20 such volumes would be quite easy to store in memory and analyze extensively.

So how does one tackle such a big input set with an output that grows linearly with the set? We think the first goal is coming up with a way of extracting the graph (along with the previous steps of alignment, segmentation and feature detection) at the speed of the microscope's image acquisition, that is, at a rate of terabytes per hour. We are hopeful that we can do this with a moderately sized system that parallelizes the most costly steps, segmentation and feature extraction because these computations are what are sometimes called embarrassingly parallel. To understand this term, consider that the speed-up obtained by parallelizing a given computation is governed by Amdahl's Law[23], which says that the time savings attainable by parallelization are limited only by the fraction of code that remains sequential (because it is hard or even impossible to parallelize). The sequential steps are mostly involved with inter-processor coordination and communication. Fortunately, image computations, such as segmentation and feature detection, can be divided among many different processors, each dedicated to a small contiguous image region in a prefixed grid and requiring little communication or coordination between the processors. It is therefore embarrassingly easy to reduce the sequential part of the code to a fraction small enough to allow a speed-up proportional to the number of physical processors applied to the computation.

**Compute system.** But speed-up is not just about computation. One must also take into account the time to move the data into the computer before computing on it. Although our multiprocessor machines have sufficient input/output bandwidth to receive the data, there is a question as to how the data will be moved from the microscope to the machines. At an image collection rate of 2.5 terabytes an hour, at present at least, it is infeasible to reliably transmit the data to a remote computation site. Placing the computer system near the microscope solves this transfer bandwidth bottleneck, allowing us to move the immense volume of data onto processors at the rate that it is being collected. Unlike many other big data problems, in our case, the data is also being processed in real time, and thus moving the computation to the data, that is, in the vicinity of the microscope, becomes absolutely essential.

We estimate that 500 standard 4-core processors, each operating at 3.6 GHz, would be adequate to keep pace with the data generation of the microscope. At this pace, the system can deliver about 10,000 computational instructions[2] per image pixel coming off the microscope. If we equip each processor with 16 gigabytes of local memory, we would have a total of 8 terabytes memory capacity, that is, sufficient memory for about 3 h worth of microscope traffic. This amount of memory would more than suffice to store and analyze 1 h worth of image data while simultaneously inputting new data and outputting the raw image data as backup onto disk (and eventually to tape). The whole computer infrastructure would likely cost less than $1 million. Adding more machines to speed up the computation is only a matter of money.

**A heterogeneous hierarchical approach.** Our goal is to complete the alignment, segmentation, feature detection and graphing within a budget of ~10,000 instructions per pixel (again, we note that increasing this budget is a matter of money). The automatic reconstruction is likely the most costly of these steps. For example, we have experimented with deep learning algorithms that deliver improved accuracy in segmentation results. However, such algorithms require on the order of 1 million instructions per pixel and therefore cannot be applied to the workflow as described. Moreover, the results of these programs are still inadequate because they require human proofreading to correct mistakes (**Fig. 1**).

Instead, we consider a heterogeneous hierarchical approach that will combine bottom-up information from the image data, with top-down information from the assembled layout graph, to dynamically decide on the appropriate computational level of intensity to be applied to a given sub-volume. This approach mimics the way humans go about segmentation, constantly looking forward and backward across sections, and with different resolutions. This approach might initially apply the lowest cost computations to small volumes to derive local layout graphs. These sub-graphs will be tested for consistency and merged at the graph level. If discrepancies are found, then more costly computation on the image data will be applied locally to resolve the inconsistencies so as to allow merging. The more extensive computation will involve, among other things, using information from the layout graph and its failed consistency tests to better segment the problematic region. This process will continue hierarchically, growing the volume of merged segments and continually testing at the layout graph level.

We believe that establishing a set of rules that constrain the results at both the lower (segmented image data) and higher (layout graph) levels will be useful to detect and ultimately resolve errors in the wiring diagram. This correction method could potentially occur without having to apply the most time-consuming computations to entire large volumes. For example, a typical problem that a human easily detects is a split error where a neural process changes ID (that is, color) from one section to the next (**Fig. 1c**,**d**). Our higher level graph checking techniques could flag the sites at which such an error may have occurred, allowing us to apply more costly, but accurate, segmentation approaches to small sub-volumes of the data set. Computers are exceptionally good at detecting inconsistencies in high-level layout graphs, for example, detecting orphaned neural processes or merge errors. Once detected, time-consuming and computer-costly segmentation techniques could then be applied locally to resolve the wiring issue. The combination of top-down graph rules and bottom-up image segmentation properties could allow a fully automatic suite of methods to reduce the image data to an accurate network graph. To be sure, we are far from this ideal. But as we have argued above, the only hope for large connectomic volumes would seem to be the continued improvement of fully automated techniques. Perhaps this is unwarranted optimism, but we believe that technology always improves with time and that this set of problems is no different from any other technical challenge.

**Data management and sharing.** There are many reasons why the original image data should be maintained. One obvious reason is that the needs of different investigators might mean that the extraction of a layout graph for one will not suit the needs of another, whereas the original image data had information that will. This requires solving two problems: the maintenance of the original data despite its large size and the developing of means for sharing it among laboratories that are geographically distributed.

For both of these problems, one must be aware of the data transfer rates available today. For connectomics, we would argue that online data transfer rates today are simply too slow to allow moving the data to remote laboratories far from the data source. Current achievable data rates between distant sites are, at best, 300 megabits per second (a commercially available optical fiber connection), which would translate to about 1.75 years to transfer 2 petabytes of data.

Given this limitation, when planning the storage of data at the petascale level, there are few viable options: store it locally on disk or on tape, or transfer it to a remote mass-storage site piecemeal, via disk or tape delivered by a courier. As can be expected, each approach has its own technical drawbacks and cost limitations. None of them seem to scale in a way that would make them viable beyond the first few peta-bytes of data.

Genomics shows how powerful it is to provide investigators access to sequence data that they had no part in acquiring. In connectomics, the same is true. Online accessibility is therefore the obvious approach to sharing data in this day and age. But given the above-mentioned online transfer rates, sharing is relegated to either transferring online only the output of the data analysis (that is, layout graphs) or defaulting to non-online transfer in the form of disks or tapes to central sharing sites.

In the end, storing petabytes of image data requires large capital investments that may be hard to justify unless there is commercial value in the data. The much smaller reconstructed layout graph is easier to deal with. However, at the moment, there is no consensus on what data such a graph should keep and what can be spared (for example, do you want the location of glia in the connectomic data?). Imagine, however, a data repository that is connected by fast local lines to a powerful image processing computer system that can generate segmentation and layout graphs. This opens up the possibility that scientists 'order-up' a layout graph that has the features they particularly are interested in studying. For this to work, there would have to be unprecedented cooperation and coordination between laboratories and sufficient capital investment and cost-sharing strategies to keep such an effort going. One such effort that is already under way is the Open Connectome Project[24].

## Conclusions

Connectomics is a nascent, data-driven field with parallels to the far more developed biological discipline of genomics[25] that serves both to test existing hypotheses and generate new ones. In connectomics, both the petabytes of original image data and the terabytes of reconstructed layout graphs can be considered to be digital versions of the brain[26]. The prospects for the success of this field depend on how easily this digital brain can be mined.

One of the biggest obstacles to many high-minded projects is money, and connectomics is no exception. Generating, storing and transforming brain tissue into layout graphs is very expensive. It is possible that this field will succeed only if a clear commercial or human health application becomes evident. The catch here is that a significant investment must be made before it will be possible to know the value of this data to society or business. Fortunately, there are potentially many avenues of commercial value in connectomics, ranging from treating brain diseases to applying the lessons learned from connectome graphs to making computers smarter. Already there are efforts under way to generate neuromorphic hardware and software and to apply lessons from cortical circuits to machine learning. Our foray into this effort suggests that, at this point in time, success will absolutely require biologists, engineers and computer scientists, working on an equal footing through the many challenges of transforming real brain into a useful digital form.

Lastly, it is important not to forget the challenges that are beyond the horizon. The outputs of our big data effort will be connectome graphs, which, even for the cortex of a small mammal, can reach several terabytes in size. It is perhaps poetic that in analyzing these smaller graphs, we will find ourselves once again faced with another, completely uncharted, big data problem.

**ACKNOWLEDGMENTS**
Support is gratefully acknowledged from the US National Institute of Mental Health Silvio Conte Center (1P50MH094271 to J.W.L.), the US National Institutes of Health (NS076467 to J.W.L. and 2R44MH088088-03 to H.P.), the National Science Foundation (OIA-1125087 to H.P., CCF-1217921 to N.S., CCF-1301926 to N.S., IIS-1447786 to N.S., and IIS-1447344 to H.P. and J.W.L.), a Department of Energy Advanced Scientific Computing Research grant (ER26116/DE-SC0008923 to N.S.), Nvidia (H.P.), Oracle (N.S.) and Intel (N.S.).

**COMPETING FINANCIAL INTERESTS**
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Helmstaedter, M. Cellular-resolution connectomics: challenges of dense neural circuit reconstruction. *Nat. Methods* **10**, 501–507 (2013).
2. Lichtman, J.W. & Denk, W. The big and the small: challenges of imaging the brain's circuits. *Science* **334**, 618–623 (2011).
3. Hell, S.W. Far-field optical nanoscopy. *Science* **316**, 1153–1158 (2007).
4. Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56–62 (2007).
5. Cai, D., Cohen, K.B., Luo, T., Lichtman, J.W. & Sanes, J.R. Improved tools for the Brainbow toolbox. *Nat. Methods* **10**, 540–547 (2013) Epub 2013 May 5.
6. Lakadamyali, M., Babcock, H., Bates, M., Zhuang, X. & Lichtman, J. 3D multicolor super-resolution imaging offers improved accuracy in neuron tracing. *PLoS ONE* **7**, e30826 (2012).
7. O'Rourke, N.A., Weiler, N.C., Micheva, K.D. & Smith, S.J. Deep molecular diversity of mammalian synapses: why it matters and how to measure it. *Nat. Rev. Neurosci.* **13**, 365–379 (2012).
8. Peddie, C.J. & Collinson, L.M. Exploring the third dimension: volume electron microscopy comes of age. *Micron* **61**, 9–19 (2014).
9. Denk, W. & Horstmann, H. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS Biol.* **2**, 329 (2004).
10. Knott, G., Marchman, H., Wall, D. & Lich, B. Serial section scanning electron microscopy of adult brain tissue using focused ion beam milling. *J. Neurosci.* **28**, 2959–2964 (2008).
11. Bock, D.D. *et al.* Network anatomy and *in vivo* physiology of visual cortical neurons. *Nature* **471**, 177–182 (2011).
12. Briggman, K.L. & Bock, D.D. Volume electron microscopy for neuronal circuit reconstruction. *Curr. Opin. Neurobiol.* **22**, 154–161 (2012).
13. Schüz, A. & Palm, G. Density of neurons and synapses in the cerebral cortex of the mouse. *J. Comp. Neurol.* **286**, 442–455 (1989).
14. Korbo, L. *et al.* An efficient method for estimating the total number of neurons in rat brain cortex. *J. Neurosci. Methods* **31**, 93–100 (1990).
15. Kaynig, V. *et al.* Large-scale automatic reconstruction of neuronal processes from electron microscopy images. Preprint at http://arxiv.org/abs/1303.7186 (2013).
16. Kim, J.S. *et al.* Space-time wiring specificity supports direction selectivity in the retina. *Nature* **509**, 331–336 (2014).
17. Plaza, S.M., Scheffer, L.K. & Chklovskii, D.B. Toward large-scale connectome reconstructions. *Curr. Opin. Neurobiol.* **25**, 201–210 (2014).
18. Bunke, H. & Varga, T. Off-line roman cursive handwriting recognition: digital document processing. *Adv. Pattern Recognit.* **2007**, 165–183 (2007).
19. Becker, C., Ali, K., Knott, G. & Fua, P. Learning context cues for synapse segmentation. *IEEE Trans. Med. Imaging* **32**, 1864–1877 (2013).
20. Varshney, L.R., Chen, B.L., Paniagua, E., Hall, D.H. & Chklovskii, D.B. Structural properties of the *Caenorhabditis elegans* neuronal network. *PLoS Comput. Biol.* **7**, e1001066 (2011).
21. Eppstein, D., Goodrich, M.T. & Sun, J.Z. The skip quadtree: a simple dynamic data structure for multidimensional data. *Proc. 21st Ann. Symp. Comput. Geom.* 296–305 (ACM, New York, 2005).
22. Herlihy, M. & Shavit, N. *The Art of Multiprocessor Programming (Revised Edition)* (Morgan Kaufmann Publishers, San Francisco, California, 2012).
23. Amdahl, G.M. Validity of the single processor approach to achieving large-scale computing capabilities. *AFIPS Conf. Proc.* **30**, 483–485 (1967).
24. Burns, R. *et al.* The Open Connectome Project Data Cluster: scalable analysis and vision for high-throughput neuroscience. *Proc. 25th Int. Conf. Sci. Stat. Database Manag.* **27**, 1–11 (2012).
25. Lichtman, J.W. & Sanes, J.R. Ome sweet ome: what can the genome tell us about the connectome? *Curr. Opin. Neurobiol.* **18**, 346–353 (2008).
26. Morgan, J.L. & Lichtman, J.W. Digital tissue. in *Cellular Connectomics.* (eds. Helmsteder, M. & Brigmann, K.) (Academic Press, in the press).