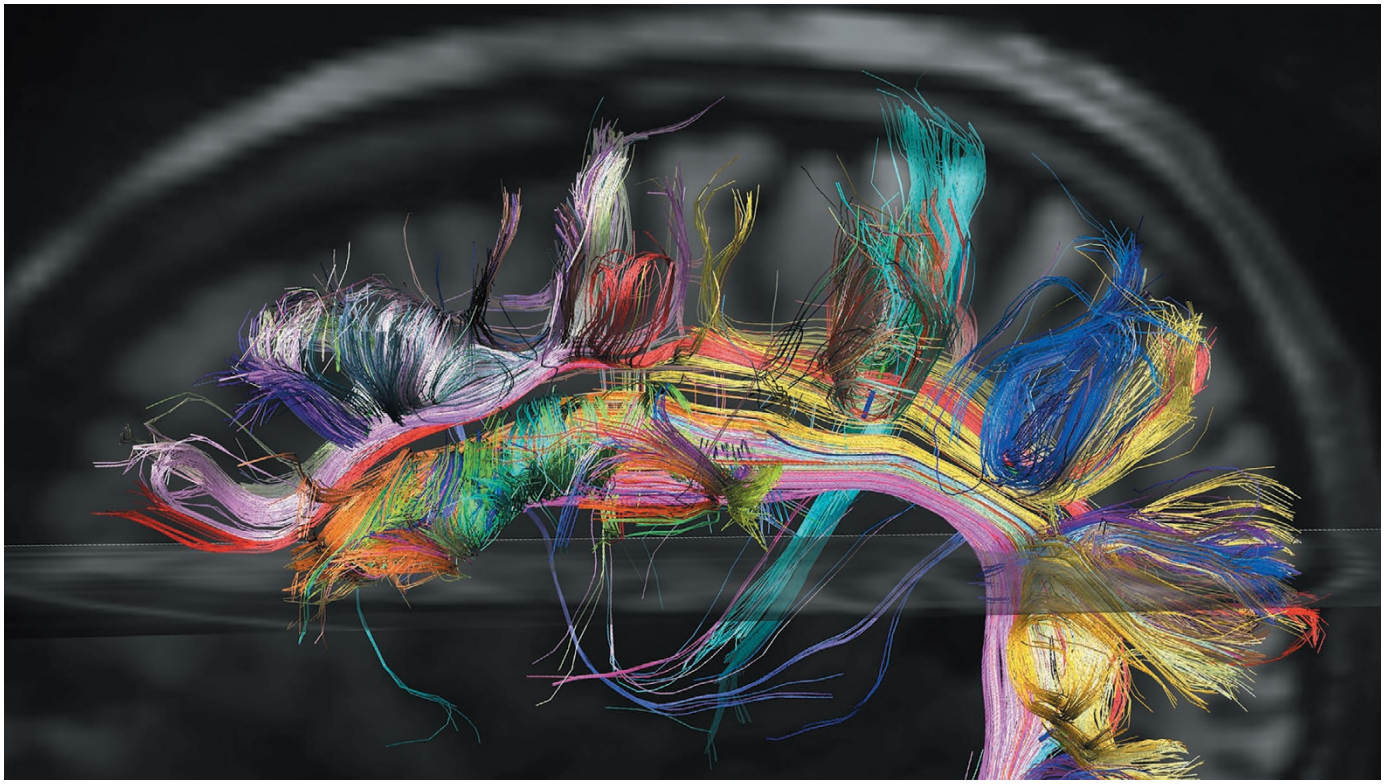
TECHNOLOGY FEATURE

BIG BRAIN, BIG DATA

Neuroscientists are starting to share and integrate data — but shifting to a team approach isn't easy.



VAN WEDEEN/MARTINOS CENTER FOR BIOMEDICAL IMAGING/HARVARD MEDICAL SCHOOL

Diffusion magnetic resonance imaging is just one of many data types that researchers are working out how to handle to bring the brain into focus.

BY ESTHER LANDHUIS

As big brain-mapping initiatives go, Taiwan's might seem small. Scientists there are studying the humble fruit fly, reverse-engineering its brain from images of single neurons. Their efforts have produced 3D maps of brain circuitry in stunning detail.

Researchers need only a computer mouse and web browser to home in on individual cells and zoom back out to intertwined networks of nerve bundles. The wiring diagrams look like colourful threads on a tapestry, and they're clear enough to show which cell clusters control specific behaviours. By stimulating a specific neural circuit, researchers can cue a fly to flap its left wing or swing its head from side to side — feats that roused a late-afternoon crowd in November at the annual meeting of the Society for Neuroscience in San Diego, California.

But even for such a small creature, it has taken the team a full decade to image 60,000 neurons, at a rate of 1 gigabyte per cell, says project leader Ann-Shyn Chiang, a neuroscientist at the National Tsing Hua University in Hsinchu City, Taiwan — and that's not even half of the nerve cells in the *Drosophila* brain. Using the same protocol to image the 86 billion neurons in the human brain would take an estimated 17 million years, Chiang reported at the meeting.

Other technologies are more tractable. In July 2016, an international team published a map of the human brain's wrinkled outer layer, the cerebral cortex[1]. Many scientists consider the result to be the most detailed human brain-connectivity map so far. Yet, even at its highest spatial resolution (1 cubic millimetre), each voxel — the smallest distinguishable element of a 3D object — contains tens of thousands of neurons. That's a far cry from the neural connections that have been mapped at single-cell resolution in the fruit fly.

"In case you thought brain anatomy is a solved problem, take it from us — it isn't," says Van Wedeen, a neuroscientist at Massachusetts General Hospital in Charlestown and a principal investigator for the Human Connectome Project (HCP), a US-government-funded global consortium that published the brain map.

So it goes in the world of neurobiology, where big data is truly, epically big. Despite advances in computing infrastructure and data transmission, neuroscientists continue to grapple with their version of the 'big data' revolution that swept the genomics field decades ago.

But brain mapping and DNA sequencing are different beasts. A single neuroimaging data set can measure in the terabytes — two to three orders of magnitude larger than a complete mammalian genome. Whereas geneticists know when they've finished decoding a stretch of DNA, brain mappers lack clear stopping points and wrestle with much richer sets of imaging and electrophysiological data — all the while wrangling over the best ways to collect, share and interpret them. As scientists develop tools to share and analyse ever-expanding neuroscience data sets, however, they are coming to a shared realization: cracking ▶

the brain requires a concerted effort.

Scientists can chart the brain at multiple levels. The HCP seeks to map brain connectivity at a macroscopic scale, using magnetic resonance imaging (MRI). Some labs are mapping neural tracks at a microscopic level, whereas others, such as Chiang's, trace every synapse and neural branch with nanoscale precision. Still others are working to overlay gene-expression patterns, electrophysiological measurements or other functional data on those maps. The approaches use different methods — but all create big data (see 'Big data by the numbers').

HOW BIG?

In part, this is because the brain, no matter the species, is so large and interconnected. But it also stems from the cells' unwieldy dimensions. A mammalian neuron's main extension — its axon — can be 200,000 times as long as its smallest branches, called dendrites, are wide. If a scale model were built such that spaghetti strands represented the dendrites, the neuron itself would be more than one-third of a kilometre long, or four American-football fields.

In the lab, researchers chart each neuron by tracing its thousands of projections through stacks of hundreds of overlapping brain-slice images. Light-based microscopy affords 0.25–0.5-micrometre resolution, which is sufficient to trace the main body of an individual neuron. But to reveal synapses — the minute signalling junctions through which electrical or chemical signals flow — nanometre-resolution electron microscopy is required. Higher resolution means smaller fields of view and so more pictures. And more pictures mean more data.

"We're not dealing with megabytes anymore, or even gigabytes," says Arthur Toga, who leads the Laboratory of Neuro Imaging at the University of Southern California in Los Angeles. "We're dealing with terabytes. Just getting it from one place to another is an issue" — 2 terabytes of data would fill the hard drive of many desktop computers.

Chiang's fruit-fly team combed through a terabyte of images to reconstruct 1,000 nerve cells — less than 1% of the *Drosophila* brain. And to map the human cerebral cortex, HCP researchers analysed 6 terabytes of MRI data from 210 healthy young adults, says Kamil Ugurbil, the HCP's co-principal investigator at the University of Minnesota in Minneapolis. Labs can download those data from the project's website or, for larger data sets, order 8-terabyte hard drives for US$200 apiece.

Electrophysiology studies have also become computationally demanding. Today, researchers routinely record hundreds of neurons at a time. Soon, it will be thousands; in five years, hundreds of thousands, says Alexandre Pouget, a neuroscientist at the University of Geneva in Switzerland. "That's the kind of leap we'll go through."

And those data come in multiple formats. Brain activity can appear as peaks amid squiggles on electrophysiological charts, or as green flashes of calcium ions moving in and out of neurons. On those green images, other fluorescent hues can indicate which neurons are sending and receiving signals. And researchers can collect these data as subjects navigate mazes, find food or watch flashing dots on a screen.

If you record 20 minutes of neural activity in a mouse brain, you produce about 500 petabytes of 'flickering', in which nerve-cell firing is represented as changes in pixel values, says neuroscientist Florian Engert of Harvard University in Cambridge. But "nobody cares about pixels. People are interested in which neurons connect to which others, and when they fire." By isolating each neuron and assigning time-stamps as they fire, he says, you can shrink the data set to a more manageable 500 gigabytes.

"The information content in raw data is mostly irrelevant," says Engert. He draws an analogy with genome sequencing: before they had automated sequencers, researchers read DNA as ordered patterns of bands on polyacrylamide gels exposed to X-ray film. Now, computer algorithms convert those bands to a sequence of Gs, As, Ts and Cs — the bases that make up the DNA strands — and no one saves the original images. Similarly, Engert says, brain scientists should "focus not on curating and distributing raw data, but rather on developing algorithms" to encode the information using fewer bits. Ideally, he says, such algorithms would enable the microscopes that collect the data to compress them as well.

The idea is sensible, but could prove challenging for the brain, in part because of mathematics. To determine protein structure using X-ray crystallography, for example, there's a "really clean theoretical model" — a series of equations that relates specific characteristics of a protein to quantifiable features in its diffraction pattern, says Greg Farber, who manages the US National Institute of Mental Health (NIMH) data archive in Rockville, Maryland. To work out the 3D structure, "you'd just measure the intensities of the spots. You don't need to keep the many, many other pixels of data on that film," he says.

"In case you thought brain anatomy is a solved problem, take it from us — it isn't."

Neuroscientists have no comparable model — no map that associates neural connectivity and activity with behaviour, memory or cognition. Given the brain's immense intricacy, Farber says, the problem "is not that we have too much data, but that we don't have nearly enough for the complexity we're trying to address".

The issue of "not enough" data resonates with Julie Korenberg, a systems neuroscientist who studies neurodevelopmental disorders at the University of Utah in Salt Lake City. A common assumption about such diseases is that changes in genes skew protein expression in certain neurons, which in turn alters the brain's wiring to cause characteristic behavioural deficits. MRI can detect gross neuroanatomic changes, such as enlarged brain areas. But subtler changes require higher-resolution approaches, such as confocal or electron microscopy. But these imaging data are represented in completely different formats, and there's no way to switch between the two: once scientists zoom in to the level of single cells, they cannot pan out again to see those cells in the context of the whole brain.

BUILDING A BRIDGE

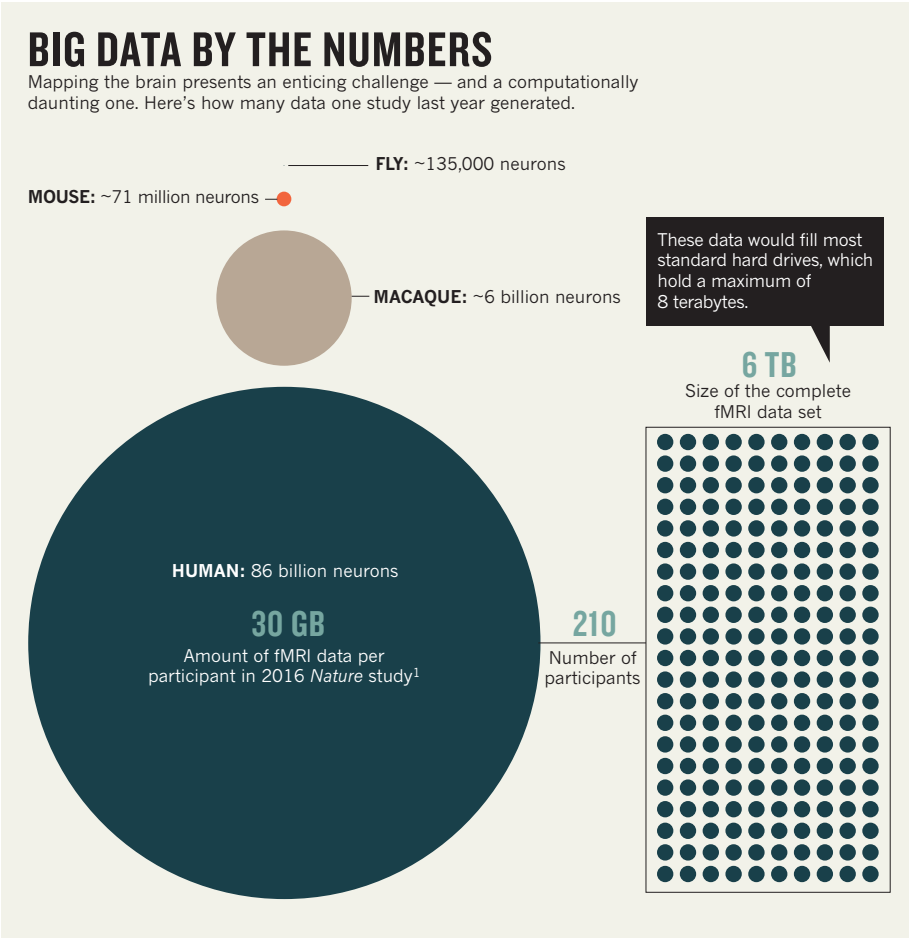
For the past 17 years, Korenberg and her colleagues have been working to bridge that gap by mapping the limbic system in macaques. These primates have 6 billion neurons in their brains, as compared to the human brain's 86 billion. But among research models, macaques are our nearest relative — much closer than a mouse or fruit fly.

Korenberg's team is developing a 3D coordinate system to align various types of neuroimaging data in the macaque brain, from whole-brain MRI connectivity to single-cell confocal data and, for some areas, subcellular resolution with electron microscopy. They're creating "a system that allows you to pick a point on one image and look at the same point at another resolution", says Janine Simmons, who heads the NIMH's Affect, Social Behavior and Social Cognition Program, which partially funds Korenberg's project. It's similar to Google Earth, Simmons says — for example, you can zoom from 40× directly to 1×, but cannot necessarily access in-between magnification scales.

Mapping the entire macaque limbic system using a 20× confocal lens will require massive data sets — well over 600 terabytes per animal. So far, the team has collected about 100 terabytes of data, accessible from a network-attached storage device that combines local 30-terabyte servers with cloud storage. The researchers can address some questions using downsized data sets and a good laptop, Korenberg says. But manipulating large 3D confocal data sets requires special workstations, and even so, the rendering of a single tiled image is slow.

Nevertheless, the work, yet to be published, "has the potential to be a major advance in the field of connectomics", says Patrick Hof, a neuroanatomist at the Mount Sinai School of Medicine in New York City who has previously collaborated with Korenberg. For instance, says Korenberg, the data could help scientists to link genes that seem important in certain mental disorders, such as schizophrenia or autism, to specific brain-wiring abnormalities.

As scientists push the limits of what is possible, they are creating computational pipelines to handle the expanding workflow, and new tools — such as Thunder and BigDataViewer — to share and visualize the resulting data. But it will take more than tool development to ease neuroscientists' data woes. A culture shift is also required. It's hard "getting people to let go of



their data", says Russell Poldrack, a psychologist at Stanford University in California who uses neuroimaging to study learning and memory. It could be "a generational thing", he says: millennials are "much more into sharing code and data than my generation". Poldrack worries that top minds might leave the field out of frustration with science "not aligning with the values they think it should have".

But slowly, attitudes are shifting — first those towards software, then data. Conventionally, neuroimaging labs spend a lot of time downloading and installing the same beta software, "hacking through various software malfunctions and computing bottlenecks, writing redundant chunks of code and implementing their own data-management solutions to tackle the same problems", says David Grayson, a neuroscience PhD student at the University of California, Davis. Worse, many non-research tasks are relegated to students, postdocs and young investigators, who tend to be tech-savvy, but "did not sign up to be sys-admins", Grayson says.

The International Neuroinformatics Coordinating Facility (INCF), a non-profit organization based in Stockholm, was created in 2005 to develop and promote standards, tools and infrastructure for brain researchers around the globe. A few years later, the United States launched the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) as a platform for sharing neuroimaging computational tools. Back then "no one was even thinking about sharing data, only software", says Nina Preuss, a programme manager for the NITRC, headquartered in Washington DC.

That changed in late 2009, when researchers at the Nathan S. Kline Institute for Psychiatric Research in Orangeburg, New York, released resting-state functional MRI (fMRI) data into the NITRC from more than 1,200 volunteers, collected for the 1000 Functional Connectomes Project (FCP). These were just pooled raw data — yet within a few weeks, NITRC users had downloaded the data set 700 times. "There was such a pent-up demand for data people could freely download and play with," says Preuss.

Download numbers soared to the thousands once the authors had cleaned up the fMRI data and made them searchable. After the data were published[2], the paper logged more than 1,000 downloads in the first 2 weeks. In the same year, the first paper by independent authors — who had downloaded the consortium's fMRI data for their own analyses, but weren't involved in collecting it — was also published[3].

Since the HCP made its first data set available in March 2013, dozens of outside labs have published papers analysing the project's data. In total, the HCP has released some 50 terabytes of brain-imaging data on more than 1,000 people, says Jennifer Elam, an outreach coordinator for the project at the Washington University School of Medicine in St. Louis, Missouri.

Few smaller-scale projects release their data, however — possibly because they don't have to. A few journals require all data supporting published findings to be made available to the community, but by and large, data sharing is not incentivized. There is "no strong impetus" to do that bit of extra work, says Grayson.

The conventional academic model doesn't help. Researchers typically develop hypotheses and work on their own ideas independently of peers in their group. In such an environment, research does not drive people together — it pulls them apart, says Hongkui Zeng of the Allen Institute for Brain Science in Seattle, Washington. "You need to distinguish yourself. To establish your identity in the field, you have to do something different from others."

Zeng joined the Allen Institute in 2006 in search of a culture change: the institute sets out ambitious five-year goals that require teams to work collaboratively and systematically, driving a project to completion rather than piecemeal, as can happen in individual labs.

When it comes to the brain, 'complete' can be a moving target. But so, too, is the neuroscience toolset. During his Society for Neuroscience talk, Chiang lamented that it's taken ten years to map half the fly brain. Working with physicists at Taiwan's Academia Sinica, Chiang's team has started to use a technique called synchrotron X-ray tomography to boost data-acquisition speed dramatically.

"It took less than 10 minutes to image a fly brain containing thousands of Golgi-stained single neurons," says Chiang, whose crew is now trying the method in mice and pigs. They plan to integrate confocal and X-ray images on a single platform from which scientists can download data. "With synchrotron X-ray imaging, mapping the human connectome at single-neuron resolution is now more realistic," Chiang says. How easy it will be to meld the maps with other data remains to be seen. ■

Esther Landhuis *is a freelance science writer in the San Francisco Bay Area of California.*

1. Glasser, M. F. *et al. Nature* **536,** 171–178 (2016).
2. Biswal, B. B. *et al. Proc. Natl Acad. Sci. USA* **107,** 4734–4739 (2010).
3. Tomasi, D. & Volkow, N. D. *Proc. Natl Acad. Sci. USA* **107,** 9885–9890 (2010).

CORRECTION

The Technology Feature 'Metabolomics: Small molecules, single cells' (*Nature* **540,** 153–155; 2016) erroneously stated that Matthias Heinemann was a former postdoc in Renato Zenobi's lab. Although he worked with Zenobi, Heinemann was a postdoc in another lab at the time. Also, Heinemann's background was in biochemical engineering, not analytical chemistry.