BIG DATA IN NEUROSCIENCE:

Improving efficiency, efficacy, and transparency in clinical and basic research

Professor Mark Stokes
Head of Attention Group, Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford

Nicholas Myers
Research Associate, St John's College, Oxford University

We have entered the Age of Big Data. Every day, the world generates over two exabytes of information (that's two million 1-TB hard drives, or a video in DVD quality that runs for 100,000 years[1]). From online marketing to particle physics, scientific as well as economic progress critically depends on aggregating, and understanding, *Big Data*. The same is true for the burgeoning field of neuroscience. Understanding how the brain works is without doubt one of the most complex challenges facing science today. Thousands of laboratories around the world are pushing the boundaries of this exciting frontier, generating vast amounts of valuable information each year. While the 20th century was marked by ever-improving techniques for measuring the brain, the next big leap in neuroscience will be driven by improved methods for aggregating, sharing, and understanding Big Data. Such a collective endeavour will depend on the support of science funders, who will need to encourage and reward researchers who share their data openly. Success in this new era of neuroscience will have implications for understanding and treating brain-related disorders, from autism to Alzheimer's disease.

... Understanding how the brain works...

What does big data in neuroscience look like? Frequently, it consists of vast archives of brain images from healthy volunteers and/or patients with neurological or psychiatric disorders. Data banks also exist for maps of brain activity patterns, collected across dozens of facilities and hundreds of individual brains using non-invasive functional imaging (eg functional magnetic resonance imaging: *fMRI*). However, arguably the most valuable data to neuroscience come from individual brain cells that can only be recorded during rare neurosurgical procedures, or in research animals. Individual labs lack the resources to generate enough of this kind of data for rigorous scientific analysis. The full potential of such information can only be realised by sharing research output across labs and between institutes. A recent example is a project that provides anonymized brain scans from over 500 volunteers with autism[2]. Compared to individual studies (which would typically test only 20 to 30 individuals), this database allows fine-grained research at an unprecedented scale that could never be reached without data sharing. Aggregating data therefore promises benefits beyond its constituent parts.

... reward researchers who share their data ...

A major advantage is that large sample sizes make discovery easier. The evidence is simply clearer for generating robust conclusions. For instance, studies scanning the human genome for variations that increase the risk of Alzheimer's disease already involve thousands of individuals. Similarly, the compilation of brain scans from patients with Alzheimer's disease into an openly accessible database has spurred a wealth of discovery (over 400 research articles have made use of the database[3]). However, data sharing is equally important for basic research, where small sample sizes still pose a fundamental problem[4]. More refined theories of brain function will require validation on larger amounts of data. As with the self-correcting nature of websites like Wikipedia, openness inherent to the collective process will also help eliminate errors that inevitably creep into the scientific literature, and will help prevent outright fraud. By increasing the reliability of published research, data sharing can increase trust and reduce wasteful follow-up studies that are based on false leads.

Furthermore, by making data available online, funders of research can save money that would otherwise be spent collecting redundant data. One neuroscience sharing platform alone is estimated to have saved funders 35 million $US (over 166 studies that used their shared data[5]). In addition to saving money, data sharing also democratizes the business

of doing science: valuable resources are no longer under lock and key, but can be accessed by researchers, laboratories, or countries that could otherwise not afford to test their theories. The diversity of a broader collective of scientific minds could help us find new directions and ways of thinking.

... Animal models make a unique and irreplaceable contribution ...

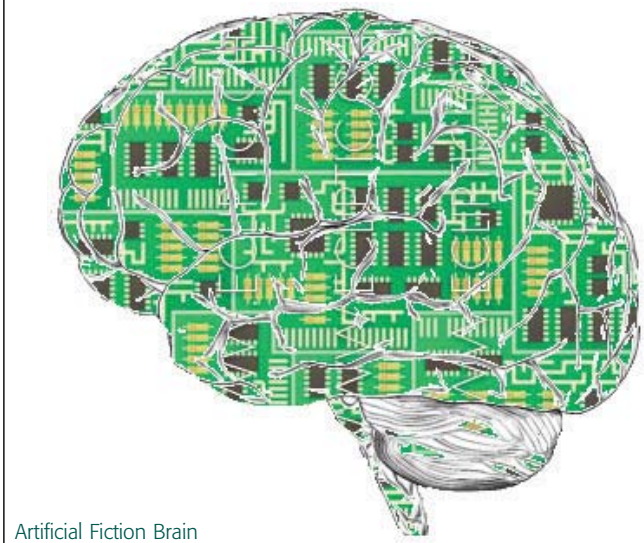
There is also an additional ethical dimension when considering data acquired from experimental animals. Animal models make a unique and irreplaceable contribution to neuroscience, but we have a greater responsibility to share these data, because their maximal re-use will increase the value of each animal life that is lost in the name of science[6]. Secure storage and accessibility of animal data will ensure that they not fall victim to the otherwise inevitable loss of data over time.

Given these benefits, it may be surprising that the majority of data in biology, and in neuroscience in particular, are not shared publicly. Data sharing

... Who owns the data ...

platforms are still the exception to the rule, and only cover particular areas (such as Alzheimer's disease). There is a number of hurdles holding back data sharing in the mainstream.

Privacy is an issue for human studies. Often there is no explicit consent from participants to share data, and obtaining consent retroactively is not feasible. Moreover, while databases typically strip scans of any information that links them with an individual, there is no guarantee that they could not be identified in the future by comparison with other publicly available data. Furthermore, group identification is possible even with anonymized data, and could lead to group discrimination based on shared neurological properties[7]. Future studies must consider ethical solutions that are pragmatic but respect individual privacy rights. One solution is to appoint a *data steward*. This custodian would be entrusted with the shared data, and grant access to it on a case-by-case basis, ensuring that data are not mined excessively or for nefarious reasons, and that they remain as anonymous as possible.



Artificial Fiction Brain

A bigger challenge may be changing the mind-set of researchers to adopt a culture of open data. At the moment, many researchers mistrust others' use of their data, and suspect that a researcher who was not involved in the initial measurements might not know how to analyse their data correctly[8]. This issue cannot be resolved over night, but once again, trained stewards could help build trust in appropriate use. For instance, they could track who accesses the shared data, and for what purpose, but also act as a consultant on correct and incorrect uses of particular data sets.

There is a related issue of intellectual property. Who owns the data, and the discoveries that arise from them? Data generators currently lack an incentive to share: as long as their valuable data are kept private, they keep exclusive access without fear of another lab publishing similar results first. In a system that rewards novel discoveries, not data sharing, it is not surprising that some researchers feel reluctant to change their ways. This problem has been recognized by funders[9], and incentives for data sharing have appeared (such as specialized data publications that can be cited by users). However,

... group discrimination based on shared neurological properties ...

these incentives are so new that they lack an agreed-upon value (and don't ensure tenure or further grants). Moreover, universities can also raise issues of intellectual property, especially when financial rewards from the translation of basic research findings into drugs or therapies are at stake[10]. How future benefits from shared data are awarded to all data generators is in urgent need of discussion.

Funding in neuroscience needs to start taking a Big Data perspective. Breaking up funds into too many small projects, without a strategy to aggregate those data sets properly and make them accessible, is wasteful and will slow down scientific discovery. By appointing stewards to coordinate and encourage data sharing, and by explicitly rewarding individual researchers and departments for sharing, funders and public policy can help effect a cultural shift in neuroscience, creating an open research community that is more than the sum of its parts.

References

1. The Four V's of Big Data.*www.ibmbigdatahub.com/infographic/four-vs-big-data*
2. Autism Brain Imaging Data Exchange. *fcon_1000.projects.nitrc.org/indi/abide/*
3. ADNI Publications.*adni.loni.usc.edu/news-publications/publications*
4. Button, K. S. *et al*. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci*. **14**, 365–376 (2013).
5. Stretching NIH Research Dollars Further. *www.hhs.gov/idealab/projects-item/stretching-nih-research-dollars/*
6. National Centre for the Replacement, Refinement & Reduction of Animals in Research. *www.nc3rs.org.uk*
7. Mittelstadt, B. D. & Floridi, L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Sci Eng Ethics1–39* (2015).
8. Rathi, V. *et al*. Sharing of clinical trial data among trialists: a cross sectional survey.*BMJ* **345**, e7570–e7570 (2012).
9. Incentives and culture change for data access. *www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/WTP056496.htm*
10. Check Hayden, E. Alzheimer's data lawsuit is sign of growing tensions. *Nature* **523**,265–265 (2015).