Title

Big data from small data: data-sharing in the 'long tail' of neuroscience.

Journal

Nature neuroscience, 17(11)

ISSN

1546-1726

Authors

Ferguson, Adam R
Nielson, Jessica L
Cragin, Melissa H
et al.

Publication Date

2014-11-01

Peer reviewed

Big data from small data: data-sharing in the 'long tail' of neuroscience

Adam R Ferguson[1], Jessica L Nielson[1], Melissa H Cragin[2], Anita E Bandrowski[3] & Maryann E Martone[3,4]

The launch of the US BRAIN and European Human Brain Projects coincides with growing international efforts toward transparency and increased access to publicly funded research in the neurosciences. The need for data-sharing standards and neuroinformatics infrastructure is more pressing than ever. However, 'big science' efforts are not the only drivers of data-sharing needs, as neuroscientists across the full spectrum of research grapple with the overwhelming volume of data being generated daily and a scientific environment that is increasingly focused on collaboration. In this commentary, we consider the issue of sharing of the richly diverse and heterogeneous small data sets produced by individual neuroscientists, so-called long-tail data. We consider the utility of these data, the diversity of repositories and options available for sharing such data, and emerging best practices. We provide use cases in which aggregating and mining diverse long-tail data convert numerous small data sources into big data for improved knowledge about neuroscience-related disorders.

The premise that neuroscience will benefit from routine and universal data sharing has been around since the early days of the Internet. Calls to develop shared data repositories similar to those developed for genomics and protein structure communities were instantiated through the US Human Brain Project in the early 1990s, funded by the US National Institutes of Health (NIH)[1]. Part of the motivation behind this was the idea that an understanding of the brain would require cooperative efforts to integrate information across scales and modalities[2], combining data generated with different techniques practiced across the various disciplines in neuroscience.

Through 2005 (refs. 3,4), the US Human Brain Project funded many software tools and databases for diverse data types, including neuroimaging, microscopy, physiology and computational modeling. As databases and community data repositories for neuroscience have continued to accrue, the Neuroscience Information Framework (NIF, http://neuinfo.org) has been charged with surveying, cataloging and federating public resources since 2008. NIF currently lists hundreds of neuroscience-specific databases comprising millions of records in its resource registry and data federation. Well-known examples of public data in neuroscience include the Allen Brain Atlas, and consortia such as the Alzheimer's Disease NeuroImaging Initiative (ADNI, http://www.adni-info.org/) and the Human Connectome project (http://www.humanconnectomeproject.org/). The utility of such resources is clear, as hundreds of publications have used these data (**Supplementary Table 1**). With the newly funded European Human Brain Project (https://www.humanbrainproject.eu/) and US Brain Research through Advancing Innovative Neurotechnologies (BRAIN) initiative (http://www.whitehouse.gov/share/brain-initiative), the amount of public data for neuroscience will continue to increase.

In the context of astronomy and high energy physics, the aforementioned projects might be termed big science[5] projects, characterized by large, coordinated teams and extensive instrumentation[6]. Although they clearly argue for open data resources in neuroscience, these new initiatives do not address the issue of routine data sharing by neuroscience researchers. The myriad data sets produced by individual small-scale studies have come to be known as long-tail data[6] (**Fig. 1**), as each data set may be small, but they collectively represent the vast majority of scientific data. Historically, raw long-tail data has been treated as a "supplement to the written record of science"[6], rather than a primary research product for formally sharing. Investments in open data repositories, defined as databases or infrastructure that accept data contributions from the community at large for distributed reuse, have been driven by the premise that making such research data available benefits science. Data sharing in the long tail is viewed as essential for increasing transparency, for mitigating against known biases in publication and for increasing data reuse by third parties[6,7]. Yet the value and effect of sharing non-standardized, heterogeneous data sets by neuroscientists across disciplines remains an open question. In this commentary, we review current practices and mechanisms for sharing long-tail neuroscience data. We distinguish long-tail data from big science initiatives such as the Allen Brain Atlas, whose mission is to produce data for the public domain, or large consortia such as ADNI or the Human Connectome Project, in which an agreement is in place to make the data arising from these initiatives publically available (that is, prospective data sharing). We focus instead on the discrete, unique data sets produced during the course of neuroscience research by individual researchers. We address issues such as data-sharing infrastructure, best practices

[1]Brain and Spinal Injury Center, Department of Neurological Surgery, University of California at San Francisco, San Francisco, California, USA. [2]Directorate for Biological Sciences, National Science Foundation, Arlington, Virginia, USA. [3]Center for Research in Biological Structure, University of California at San Diego, San Diego, California, USA. [4]Department of Neuroscience, University of California at San Diego, San Diego, California, USA. Correspondence should be addressed to M.E.M. (mmartone@ucsd.edu).

Received 12 May; accepted 17 September; published online 28 October 2014; doi:10.1038/nn.3838

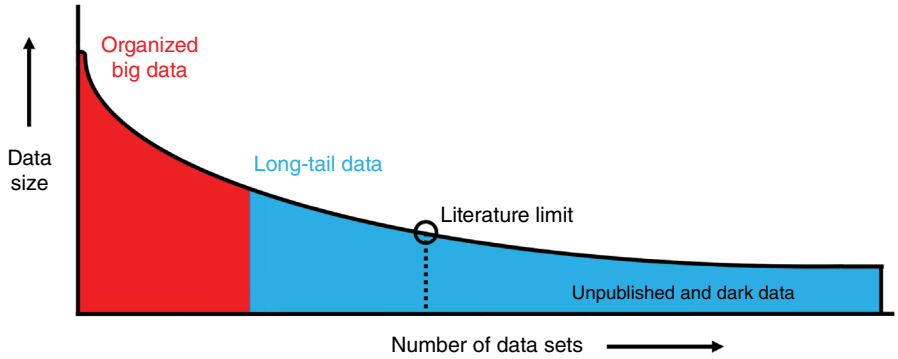

Figure 1 Schematic illustration of long-tail data. Studies that have plotted data set size against the number of data sources reliably uncover a skewed distribution. Well-organized big science efforts featuring homogenous, well-organized data represent only a small proportion of the total data collected by scientists. A very large proportion of scientific data falls in the long-tail of the distribution, with numerous small independent research efforts yielding a rich variety of specialty research data sets. The extreme right portion of the long tail includes data that are unpublished; such as siloed databases, null findings, laboratory notes, animal care records, etc. These dark data hold a potential wealth of knowledge but are often inaccessible to the outside world.

and incentives, and case studies in which sharing long-tail data has yielded clear benefits.

What are long-tail data and why share them?

Long-tail data in neuroscience can be defined as small, granular data sets, collected by individual laboratories in the course of day-to-day research. These data consist of small publishable units (for example, targeted endpoints), as well as alternative endpoints, parametric data, results from pilot studies and metadata about published data (**Fig. 1**). The long-tail of data is also composed of 'dark data', unpublished data that includes results from failed experiments and records that are viewed as ancillary to published studies (for example, veterinary care logs). Although these data may not be considered useful in the traditional sense, data-sharing efforts may illuminate important information and findings hidden in this long tail.

An analysis of the literature provides three historical arguments for increased access to long-tail research data in neuroscience. The first wave of calls for neuroscience data sharing was driven by the computational neuroscience and neuroimaging communities, which were interested in data-integration for modeling brain function[1,8–10]. The imaging community, particularly human neuroimaging, has recently renewed calls for data sharing, driven in part by gaps in what is currently available with a single research center[11,12]. This second call for data sharing emphasizes the development of large aggregated data sets to increase subpopulation sizes for improving analytical potential of participant-level data[13]. There is a third call extending across biomedicine in support of data sharing that shifts the focus from altruism and discovery to issues of transparency, reproducibility and waste[7,14]. Like many communities across biomedicine, neuroscience is grappling with issues of data quality and reproducibility[11,14–16]. Proponents of open data sharing contend that no scientific field is immune from errors or methodological limitations, and making primary data available for re-analysis serves to uncover and correct errors more quickly than our current practices. In a meta-analysis of psychology papers, however, Wicherts *et al.*[17] found that studies with accessible data tended to have fewer errors and more robust statistical effects than those that did not, suggesting that when researchers know that data will be made public, more care is taken in data management and/or reporting.

Driven in part by the current squeeze in funding, biomedical communities are also raising concerns that insufficient data sharing has led to waste across the medical enterprise[7]. Recent estimates indicate that more than 50% of scientific findings do not appear in the published literature and instead reside in file drawers and personal hard drives[7]. This so-called 'file-drawer phenomenon' dominates "the long-tail of dark data"[18]. Lack of publication of dark data undermines the entire scientific research enterprise, leaving an incomplete and biased record[7], needless duplication of scientific efforts (as previous attempts are unknown), and contributing to failures in scientific replication and translation[19]. A recent estimate suggests that over 50% of completed studies in biomedicine go unreported, often because results do not conform to author's hypotheses[7]. These issues have high costs for stakeholders beyond the data producers, from patients and taxpayers to policy makers, and suggest that there are wide-ranging inefficiencies across the current system of scholarly communication.

The potential benefits from sharing long-tail neuroscience data, including dark data, can be exemplified by experiences in the neurotrauma field. Stroke, traumatic brain injury (TBI) and spinal cord injury (SCI) collectively affect over 2.5 million people every year in the US. Given the prevalence of these disorders, substantial public resources have been dedicated to discovery of new therapeutics. Numerous high-profile clinical trials have failed, despite promising published findings from animal models[20]. In response to these failures, neurotrauma communities have dedicated substantial time and resources to standardizing study design at both the preclinical and clinical level[21–23]. For example, the Stroke Treatment and Academic Roundtable (STAIR) standards were implemented in 1999 (updated in 2009) to create a set of guidelines for testing neuroprotective therapies in preclinical stroke models[24,25]. However, reproducibility failures continue, even at the pre-clinical level[16].

More recently, both the TBI and SCI communities engaged in substantial efforts to acquire and harmonize existing data sets through efforts such as IMPACT[26] for human TBI data and VISION-SCI[23] for animal research (**Box 1**). A tangible outcome of these activities is the emergence of new data standards, including a set of common data elements for prospective studies and powerful new prognostic statistical models for predicting neurological recovery. These case studies directly address questions about whether analyzing pooled long-tail data from multiple has value for human health. In the case of the IMPACT study, the answer is clearly yes, as aggregation of data from 43,000 patients has led to development of common data elements (CDEs) in clinical TBI studies and more accurate diagnostic/prognostic statistical models[27] (**Box 1**). These CDEs, in turn, standardize data collection to ensure that new studies produce long-tail data that can be more easily pooled across multiple centers and trials and combined and tested for common features present in TBI. Emerging methods for TBI neuroimaging, genetics and proteomic biomarkers are being further developed as part of newly announced international big-data discovery trials for TBI precision medicine, the CENTER-TBI and TRACK-TBI projects[28,29]. Early results of these international efforts have already identified previously unknown magnetic resonance imaging (MRI) and molecular biomarkers to predict long term neurocognitive outcome after TBI[30,31].

In preclinical SCI, similar wide-scale attempts to harmonize legacy long-tail

Box 1 Successful long-tail data sharing: case studies from translational neurotrauma

After several failed clinical trials for TBI, a multinational consortium launched IMPACT. IMPACT gathered long-tail data from all of the major clinical trials in TBI conducted over the past 20 years into a single database (http://www.tbi-impact.org)[26]. The IMPACT database now contains data from over 43,243 patients with TBI and (re)analysis of these long-tail data has produced 62 publications to date[57], including more accurate prognostic models of outcome. As an example, when data from about 8,700 patients with TBI was mined and reanalyzed, researchers found that combining information from the Glasgow Coma Scale (GCS), pupil reactivity, blood work and CT imaging improved outcome prediction[30]. This drove development of a publically available statistical 'prognostic calculator' (http://www.tbi-impact.org/?p=impact/calc) with unprecedented precision for predicting TBI recovery, providing clear guidance for tailoring patient care. These efforts also contributed to the creation of the NIH and National Institute of Neurological Disorders and Stroke data-reporting standards for TBI, known as the NINDS TBI Common Data Elements (TBI CDEs). These long-tail data sharing efforts provided proof of concept leading to ~$60,000,000 investments by the US and Europe as part of the International Initiative for Traumatic Brain Injury Research (InTBIR). Early results suggest that these long-tail data-sharing efforts will help usher in a new era for TBI precision medicine.

Replication failures in the SCI research community[16] have given rise to grass-roots preclinical data-sharing efforts known as Minimum Information about an SCI experiment (MIASCI)[22] and Visualized Syndromic Information and Outcomes for Neurotrauma-SCI (VISION-SCI)[23]. Multivariate (re)analysis of long-tail data are revealing multidimensional syndromic patterns that translate across SCI injury models, laboratories and species[23]. For example, by combining data from multiple studies in cervical SCI models and performing multivariate statistical analysis, we identified a previously unknown set of overlapping measures of motor recovery that co-vary with gray matter and white matter lesion pathology both in rats[32] and between rats and monkeys[23]. Statistical correction for this multivariate relationship revealed that coordinated weight bearing during locomotion is more sensitive to transection injuries, whereas stride length during locomotion is more sensitive to contusive injuries in the spinal cord[32]. By identifying conserved features expressed in multidimensional (syndromic) space, long-tail data sharing is now helping to screen for mechanistically precise therapeutic effects conserved across models and species to accelerate translation.

data from dozens of laboratories, including unpublished data and 'background data'[6] (for example, animal care records), are helping to develop a more complete picture of SCI by deriving the computationally defined SCI syndromic space[23,32]. As the STAIR experience shows, it is difficult to completely standardize and control for initial conditions in models across laboratories. Thus, each individual animal study provides an incomplete glimpse into the syndrome across the full spectrum of possible injury conditions and outcome metrics. By piecing these studies together and harnessing big-data analytics to look across multiple endpoints, both traditionally reported and those residing in the file drawers (for example, postoperative and veterinary care logs), we can characterize the complex network of interactions of motor, sensory, autonomic and inflammatory responses following SCI. Big-data analytics on SCI long-tail data have uncovered pathophysiological targets that not only translate between injury paradigms[32], but also between species[23] (**Box 1**). Thus, in the neurotrauma field, aggregation of existing data is allowing us look across the full spectrum of research results to both improve our prospective data gathering efforts and make inroads into the complexity of nervous system disorders.

Potential caveats of data sharing

The above case studies support the arguments that the scientific enterprise benefits when data are shared and point to a role for data repositories and curators in aggregating and harmonizing these data to support meaningful re-use. However, it remains controversial whether sharing of long-tail data is uniformly beneficial to science. If transparency and reuse are considered benefits, what are the drawbacks?

Researchers often cite the fear that re-analysis of poor quality data sets or even good data sets by non-experts will lead to a flood of bad science in the literature[33]. Although this is certainly a concern, advocates of data sharing point to increased access to additional human capital available for analyzing data in new ways. We also know that the current literature, as evidenced by the lack of reproducibility, is rife with examples of poor data, poor experimental design and faulty analysis. Another objection concerns the costs associated with managing, hosting and curating data. As these activities are largely supported by research dollars, they are viewed as having a potentially negative effect on research funding. However, this concern must be balanced against the years of failed translational and clinical studies and the cost of generating new data[34].

But as science is a human enterprise, arguments for and against data sharing often focus less on perceived benefits to science as a whole and more on the effect on individual researchers. Does data sharing benefit or harm scientific careers? Interviews and studies of attitudes toward data sharing clearly show that many researchers perceive the latter to be true[35]. An oft-stated reason for not sharing data is the desire to continue to mine the data and the fear of being scooped if the data are made public[36]. Historically, scientists have had a proprietary relationship with their data and, until recently, few have challenged this relationship[37]. Even among researchers willing to share their data, the time and resources required to prepare the data for use by others represent major disincentives. Scientists dedicate enormous time preparing papers for publication, as they serve as their primary index of career success. Metrics such as citation rates quantify the impact of these publications using a well-developed citation system. But no such metrics or norms exist for reuse of data. Given that data are considered to be supplements to the scientific record[6] rather than primary products of research to be credited, cited and tracked, researchers must weigh the time and resources required for preparing data for release relative to the benefits they are likely to accrue.

Data sharing may lead to the fear that others will uncover errors in the data or question the validity of the analysis[35]. Unlike the open source software community, where error correction is encouraged and welcomed, uncovering errors in scientific data may be perceived as an attack on a researcher's reputation. In the hypercompetitive environment of biomedicine, such attacks may lead to hard feelings, finger pointing and a competitive disadvantage. In recognition of such potential abuses, data sharing has contributed to the advocacy for development of normative practices as to how researchers raise issues of errors in a manner that encourages open dialog. Replication etiquette[38] might include contacting the data contributor and inviting them to review findings or perhaps co-author such a study when new findings result, or to help reveal and correct errors in a collaborative, collegial fashion. A recent example of this approach comes

from neurophysiology, wherein re-analysis of pooled data from the CRCNS repository (http://crcns.org) by a third party yielded a high-impact paper reporting that distributed hippocampal local field potentials encode a rat's position in space[39]. The original data donors were directly engaged during re-analysis and served as co-authors (F. Sommer, personal communication), demonstrating that data sharing can benefit data donors and data end-users alike.

Although many discussions on incentives for data sharing focus on the harm done to the researcher sharing the data, fewer have focused on the harm done to researchers when data are not shared. Economic estimates indicate that lack of transparency and data inaccessibility in biomedicine cost tens of billions of dollars annually worldwide[7]. But there is likely a human cost as well. How many young scientists or graduate students are derailed by trying to replicate studies that essentially reported cherry-picked results or results based on faulty data or tools? One author[19] refers to these findings as "occasional happy mistakes" and provides an anecdote about a frustrated graduate student who might not have attempted to replicate a finding had all the original data and results been available. Although difficult to quantify, conversations with colleagues and our own experiences suggest that such avoidable dead ends exact a human cost, discouraging scientists and perhaps driving some of them from science altogether.

A matter of incentives: how credit may be given for data sharing

Surveys on data-sharing practices[6,35,40,41] find evidence of peer-to-peer sharing, where researchers barter for something in exchange for data, such as authorship or good will of colleagues. The incentives here are personal and controlled, and time requirements are minimized using a simple file transfer. But preparing data for hosting in a repository often involves more effort and the benefits from reuse of these data may not directly benefit the contributor. For example, statistics on reuse of data from the Cell Centered Database[42], a database of high resolution imaging data (now part of the Cell Image Library), reveal that the majority of published studies come from computational scientists reusing the data for creation of models or algorithms for image analysis and segmentation[43]. This result is not surprising given that computational scientists may lack the skills or instruments for acquiring such data *de novo*. To this end, data producers have to expend considerable time and effort to make data discoverable and useful to third parties. Reuse of these data clearly benefits the third party, who gets a publication, and the resource provider, as reuse provides justification for further funding. Without a system of citation and reward for data reuse, the benefits to the data contributor are unclear.

Creating systemized incentives for data sharing to the individual contributor will be critical for making such practices routine in neuroscience. Development of a scholarly system for credit attribution for data, equivalent to our current system for literature citations, is underway. Currently, two main approaches are receiving attention. The first is the launch of 'data papers' or full 'data journals', that is, journals that are designed to describe a data set rather than an analysis of data. Data papers are designed to serve two purposes: to provide sufficient metadata and description of data such that it can be reused, and to co-opt our current paper-based reward system to credit researchers who make data available. Data papers require that data be deposited into a managed repository and that a stable identifier such as a digital object identifier (DOI) be assigned as a handle for identifying data[44]. A set of guidelines has emerged for data papers specifically to promote data sharing in neuroscience[45], with an exemplar data paper in the journal *Gigascience*[46] and linked data deposited in the OpenFMRI repository (OpenfMRI: ds000114; https://openfmri.org/dataset/ds000114).

The second approach is the creation of a citation and tracking system for data sets themselves, independent of whether a data paper appears. This developing citation system does not require the production of a separate paper, but supports formal referencing in articles. This system has the advantage of making data machine-readable, improving tracking and mining of data citations. There are currently several standards and principles for data citation, including the CODATA/ITSCI Taskforce on Data Citation[47] and the Joint Declaration of Data Citation Principles[48]. Task forces are underway in several communities, including the Research Data Alliance[49] and FORCE11, to develop recommendations for a data citation system. Although above solutions create mechanisms through which researchers can gain credit for data, research communities themselves will have to determine the relative value of a given data set in terms of academic promotions, funding and careers.

Some funding bodies, such as the NIH, have successfully instituted targeted data-sharing requirements, requiring communities to deposit data in a shared repository as a condition of funding. Notable examples include the National Database on Autism Research (NDAR) and the Federal Interagency TBI Research (FITBIR) informatics system. These focused efforts have implemented standards and tools for tracking compliance and have sustained intramural support from the NIH, US Department of Defense Congressionally Directed Medical Research Program and the US Army Medical Research and Materiel Command, among others. Coupled with support mechanisms, this infrastructure provides a model for sustained long-tail data sharing.

Data sharing in the neuroscience community: attitudes and best practices

Given the current reward system and the perceived disincentives to data sharing, do we have any evidence that neuroscientists are ready to share data? We believe that the answer is yes, although an examination of current repositories and communities yields some interesting observations on when, where and how. A sampling of community data repositories reveals that most public neuroscience data repositories are minimally populated relative to the total amount of data produced and the number of laboratories producing them (**Supplementary Table 1**). The minimal population suggests that a wide-scale culture of routine long-tail data sharing does not yet exist. Nevertheless, considerable variability exists across these resources, with some, for example, NeuroMorpho.org, NDAR and Cell Image Library, being well-populated with contributions from multiple laboratories (**Supplementary Table 1**). Searching FigShare for 'neuroscience' also returns thousands of data sets. This finding suggests that, in some communities, data sharing is occurring through third party repositories.

Surveys and studies of data sharing across science also indicate that attitudes in the research community toward routine data sharing are not uniformly negative, but are varied and in flux. For example, the neuroimaging community has undergone a substantial change in attitude about data sharing since the early 2000s, when the *Journal of Cognitive Neuroscience* requirement to deposit fMRI data into the fMRI Data Center prompted a loud and angry response[12]. By 2014, the ADNI and Human Connectome[50] projects have made large amounts of neuroimaging data available. But beyond these large, institutionally coordinated consortia, there are notable proponents of long-tail data sharing in the neuroimaging community. Grass-roots projects are making data sets freely available for re-use, including the 1000 Functional Connectomes, now known as the International Neuroimaging Data-sharing Initiative (INDI)[51], and the NeuroImaging Tools and Research Clearinghouse (NITRC, http://nitrc.org) lists 89 data resources, whereas NIF lists 56 databases for data-sharing in neuroimaging.

Box 2 Data-sharing best practices for long-tail data

Discoverable. Data must be modeled and hosted in a way that they can be discovered through search. Many data, particularly those in dynamic databases, are considered to be part of the 'hidden web', that is, they are opaque to search engines such as Google. Authors should make their metadata and data understandable and searchable, (for example, use recognized standards when possible, avoid special characters and non-standard abbreviations), ensure the integrity of all links and provide a persistent identifier (for example, a DOI).

Accessible. When discovered, data can be interrogated. Data and related materials should be available through a variety of methods including download and computational access via the Cloud or web services. Access rights to data should be clearly specified, ideally in a machine-readable form.

Intelligible. Data can be read and understood by both human and machine. Sufficient metadata and context description should be provided to facilitate reuse decisions. Standard nomenclature should be used, ideally derived from a community or domain ontology, to make it machine readable.

Assessable. The reliability of data sources can be evaluated. Authors should ensure that repositories and data links contain sufficient provenance information so that a user can verify the source of the data.

Useable. Data can be reused. Authors should ensure that the data are actionable, for example, that they are in a format in which they can be used without conversion or that they can readily be converted. In general, PDF is not a good format for sharing data. Licenses should make data available with as few restrictions as possible for researchers. Data in the laboratory should be managed as if it is meant to be shared; many research libraries now have data-management programs that can help.

Developing a tracking and reward system for data will require that researchers themselves pay more attention to managing and sharing their data[44,52]. Although private data sharing via an individual laboratory webpage or as supplementary information to a publication is convenient, public repositories maintained by independent organizations can better ensure that the long-tail data sets are maintained, archived, searchable and visible. Although there is a perception that we lack data repositories for the types of data neuroscientists produce, NIF and other registries in fact list over 350 data repositories that cover a variety of data types. We have also seen the emergence of generic repositories like FigShare and Dryad that can accommodate multiple data types. More importantly, each community in neuroscience will need to agree and adopt best practices and standards so that long-tail data are transparent and informative to others (**Box 2**).

Next steps in sharing long-tail data

From the above discussion, we see that a basic infrastructure, a set of best practices and a system of citation for sharing research data are starting to take shape in neuroscience. With these tools, funding agencies, scientific journals and research institutions could have a more active role in developing policies about when, where, what and how data should be shared. Given the early stages of big-data science and the extreme variations in data set size and type, we don't think that a one-size-fits-all policy or infrastructure will work. In our modern networked world, it is unnecessary for all data to be in a single location; rather, the development of stable identifiers and reference systems for data sets allows dynamic indices to be maintained that connect required pieces together. Thus, different types of data, even if they derive from a single study, may be deposited into different resources as a cost-effective solution. For example, institutional repositories might be used for much of the data produced during a study, and especially dark data, whereas community repositories might host more specialized or curated subsets. Identifiers for data and a system of data tracking would also allow funding agencies and journals to monitor compliance with data-sharing policies, something that is currently difficult to do. And finally, communities will have to develop the normative practices to reuse data in a cooperative and ethical manner, as well as procedures that attribute and credit data contributors appropriately.

We don't want to give the impression that all of the challenges in routine data sharing have been addressed. As yet, perhaps the biggest unknown is who will pay for it all. Thus far, funding agencies, institutions, publishers and researchers have almost engaged in a game of 'hot potato', with each passing the burden for sustainability of such resources on to someone else. Although a wide range of revenue models exist across the digital content marketplace, there are many uncertainties about whether these will work for publicly funded basic research that is years away from affecting human health, that is, where the immediate value of the data are unknown. Databases such as NDAR pass the cost for data curation, ingestion and storage to the data providers, who must include these funds in their grant application. Such models work when deposition of data is a condition of funding, but are an uncertain revenue stream without this mandate. What is clear is that the accommodation of data as a primary product of research will require new funding models and market options be explored for both the content and the services these resources provide. For example, recently, the idea of data persistence insurance"[53] was proposed; although the viability of this proposal is uncertain, the call for creative engagement with the commercial sector for scientific data is timely.

Despite these many challenges, emerging evidence suggests that long-tail neuroscience data is of value. Individual data sets can be reanalyzed for new insights and multiple data sets can be aggregated and meaningfully analyzed when databases are broadly populated and well-curated, enabling researchers to ask questions across the data space that are not addressable with a single study[23,54,55]. Integrating curated data across scales of analysis through data links and data federation engines will continue to accelerate neuroscience data–driven discovery[56]. Although big science is expected to produce big data, the scientific community already has vast and not yet fully archived big data, particularly when we consider all of the long-tail data that have been sitting in desk drawers in every laboratory and office. It's time that we take advantage of all that long-tail data have to offer.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the NIF staff, especially B. Ozyurt for his text mining expertise and tools that contributed substantially to **Supplementary Table 1**. The Neuroscience Information Framework is supported by a contract from the NIH Neuroscience Blueprint HHSN271200800035C via the National Institute on Drug Abuse. VISION-SCI is supported by NIH grants NS067092 (A.R.F.) and NS079030 (J.L.N.), and the Craig H. Neilsen foundation (A.R.F.) and Wings for Life foundation (A.R.F). This material is based on (M.H.C.) work supported while serving at the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in

this material are those of the author(s) and do not reflect the views of the National Science Foundation.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Huerta, M.F., Koslow, S.H. & Leshner, A.I. *Trends Neurosci.* **16**, 436–438 (1993).
2. Roysam, B., Shain, W. & Ascoli, G.A. *Neuroinformatics* **7**, 1–5 (2009).
3. National Institutes of Health. NIH Program Announcement NOT-MH-05–014, http://grants.nih.gov/grants/guide/notice-files/NOT-MH-05-014.html (2005).
4. Shepherd, G.M. *et al. Trends Neurosci.* **21**, 460–468 (1998).
5. Weinberg, A.M. *Science* **134**, 161–164 (1961).
6. Wallis, J.C., Rolando, E. & Borgman, C.L. *PLoS ONE* **8**, e67332 (2013).
7. Chan, A.W. *et al. Lancet* **383**, 257–266 (2014).
8. Ascoli, G.A., Donohue, D.E. & Halavi, M. *J. Neurosci.* **27**, 9247–9251 (2007).
9. Gardner, D. *et al. Neuroinformatics* **6**, 149–160 (2008).
10. Gardner, D. *et al. Neuroinformatics* **1**, 289–295 (2003).
11. Boline, J., Lee, E.F. & Toga, A.W. *Front. Neurosci.* **2**, 100–106 (2008).
12. Van Horn, J.D. & Gazzaniga, M.S. *Neuroimage* **82**, 677–682 (2013).
13. Perrino, T. *et al. Perspect. Psychol. Sci.* **8**, 433–444 (2013).
14. Poline, J.B. & Poldrack, R.A. *Front. Neurosci.* **6**, 96 (2012).
15. Poldrack, R.A. *et al. Front. Neuroinform.* **7**, 12 (2013).
16. Steward, O., Popovich, P.G., Dietrich, W.D. & Kleitman, N. *Exp. Neurol.* **233**, 597–605 (2012).
17. Wicherts, J.M., Bakker, M. & Molenaar, D. *PLoS ONE* **6**, e26828 (2011).
18. Heidorn, P.B. *Libr. Trends* **57**, 280–299 (2008).
19. Mueck, L. *Nat. Nanotechnol.* **8**, 693–695 (2013).
20. Sena, E.S., van der Worp, H.B., Bath, P.M., Howells, D.W. & Macleod, M.R. *PLoS Biol.* **8**, e1000344 (2010).
21. Fawcett, J.W. *et al. Spinal Cord* **45**, 190–205 (2007).
22. Lemmon, V.P. *et al. J. Neurotrauma* **31**, 1354–1361 (2014).
23. Nielson, J.L. *et al. J. Neurotrauma* doi:10.1089/neu.2014.3399 (31 July 2014).
24. Fisher, M. *et al. Stroke* **40**, 2244–2250 (2009).
25. Kwon, B.K., Hillyer, J. & Tetzlaff, W. *J. Neurotrauma* **27**, 21–33 (2010).
26. Marmarou, A. *et al. J. Neurotrauma* **24**, 239–250 (2007).
27. Maas, A.I. *et al. J. Neurotrauma* **28**, 177–187 (2011).
28. Manley, G.T. & Maas, A.I. *J. Am. Med. Assoc.* **310**, 473–474 (2013).
29. Yue, J.K. *et al. J. Neurotrauma* **30**, 1831–1844 (2013).
30. Steyerberg, E.W. *et al. PLoS Med.* **5**, e165 (2008).
31. Yuh, E.L. *et al. Ann. Neurol.* **73**, 224–235 (2013).
32. Ferguson, A.R. *et al. PLoS ONE* **8**, e59712 (2013).
33. Turner, C.F. *et al. Database (Oxford)* **2011**, bar043 (2011).
34. Turner, J.A. *et al. Front. Neuroinform.* **4**, 10 (2010).
35. Tenopir, C. *et al. PLoS ONE* **6**, e21101 (2011).
36. Roche, D.G. *et al. PLoS Biol.* **12**, e1001779 (2014).
37. Boulton, G., Rawlins, M., Vallance, P. & Walport, M. *Lancet* **377**, 1633–1635 (2011).
38. Bohannon, J. *Science* **344**, 788–789 (2014).
39. Agarwal, G. *et al. Science* **344**, 626–630 (2014).
40. Cragin, M.H., Palmer, C.L., Carlson, J.R. & Witt, M. *Philos. Trans. A Math. Phys. Eng. Sci.* **368**, 4023–4038 (2010).
41. Halavi, M., Hamilton, K.A., Parekh, R. & Ascoli, G.A. *Front. Neurosci.* **6**, 49 (2012).
42. Martone, M.E. *et al. J. Struct. Biol.* **138**, 145–155 (2002).
43. Fernandez, J.J. *BMC Bioinformatics* **10**, 178 (2009).
44. Goodman, A. *et al. PLoS Comput. Biol.* **10**, e1003542 (2014).
45. Gorgolewski, K.J., Margulies, D.S. & Milham, M.P. *Front. Neurosci.* **7**, 9 (2013).
46. Gorgolewski, K.J. *et al. Gigascience* **2**, 6 (2013).
47. Klein, T. *et al. Data Sci. J.* **12**, 1–9 (2013).
48. The Future of Research Communications and e-Scholarship (FORCE11). Joint Declaration of Data Citation Principles–FINAL, https://www.force11.org/datacitation (2013).
49. Research Data Alliance. Research data sharing without barriers, https://rd-alliance.org/group/data-citation-wg.html (2014).
50. Van Essen, D.C. *et al. Neuroimage* **80**, 62–79 (2013).
51. Mennes, M., Biswal, B.B., Castellanos, F.X. & Milham, M.P. *Neuroimage* **82**, 683–691 (2013).
52. The Royal Society. Science as an open enterprise, https://royalsociety.org/policy/projects/science-public-enterprise/Report/ (2012).
53. Kennedy, D.N. *Neuroinformatics* **12**, 361–363 (2014).
54. Costa L.F., Zawadzki, K., Miazaki, M., Viana, M.P. & Taraskin, S.N. *Front. Comput. Neurosci.* **4**, 150 (2010).
55. Hansen, M.B., Jespersen, S.N., Leigland, L.A. & Kroenke, C.D. *Front. Integr. Neurosci.* **7**, 31 (2013).
56. Martone, M.E., Gupta, A. & Ellisman, M.H. *Nat. Neurosci.* **7**, 467–472 (2004).
57. Maas, A.I. *et al. Lancet Neurol.* **12**, 1200–1210 (2013).

CFI statement:
M.E. Martone is the principal investigator of the Neuroscience Information Framework. A.E. Bandrowski is the NIF Project Leader. A.R. Ferguson, J.L. Nielson and M.H. Cragin are not affiliated with NIF.