visualizations. Research scientists need to work more closely with their computing colleagues to make sure that these needs are met and that the development of new analytic methods tied to scientific, as opposed to business, analysis is pursued.

Finally, we must work together to explore new ways of scaling easy-to-generate visualizations to the data-intensive needs of current scientific pursuits. (Figure 2 shows the use of standard projection techniques as a low-cost alternative to expensive high-end technologies.) Although there are scientific problems that do call for specialized visualizers, many do not. By bringing visualization into everyday use in our laboratories, we can better understand the requirements for the design of new tool kits, and we can learn to share and maintain visualization workflows and products the way we share other scientific systems. A side effect may well be the lowering of barriers (such as costs and accessibility) to more sophisticated visualization of increasingly larger data sets, a crucial functionality for today's data-intensive scientist.

References and Notes

1. National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), *Towards a Robust, Agile, and Comprehensive Information Infrastructure for the Geosciences: A Strategic Plan for High-Performance Simulation* (NCAR, UCAR, Boulder, CO, 2000); www.ncar.ucar.edu/Director/plan.pdf.
2. T. Hey, S. Tansley, K. Tolle, Eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft External Research, Redmond, WA, 2009).
3. We use the term "immersive environments" to include a number of high-end technologies for the visualization of complex, large-scale data. The term "Cave Automatic Virtual Environment (CAVE)" is often used for these environments based on the work of Cruz-Neira *et al.* (*13*). A summary of existing projects and technologies can be found at http://en.wikipedia.org/wiki/Cave_Automatic_Virtual_Environment.
4. 2010 Web usage estimate from Internet World Stats, www.internetworldstats.com/.
5. This general term accounts for a number of different research approaches; the Web site http://nosql-database.org/ maintains links to ongoing blogs and discussions on this topic.
6. R. Magoulas, B. Lorica, *Introduction to Big Data, Release 2.0*, issue 11 (O'Reilly Media, Sebastopol, CA, 2009); http://radar.oreilly.com/2009/03/big-data-technologies-report.html.
7. C. Bizer, T. Heath, K. Idehen, T. Berners-Lee, "Linked data on the Web (LDOW2008)," in *Proceedings of the 17th International Conference on World Wide Web*, Beijing, 21 to 25 April 2008.
8. G. Williams, J. Weaver, M. Atre, J. Hendler, *J. Web Semant.* **8**, 365 (2010).
9. R. Lengler, M. J. Epler, A Periodic Table of Visualization Methods, available at www.visual-literacy.org/periodic_table/periodic_table.html.
10. We note that this is also a good example of our general point about the need to integrate visualization into the exploration phase of science; identifying this problem earlier would have enabled substantially higher-quality data collection over the period shown.
11. S. K. Card, J. D. Mackinlay, B. Shneiderman, *Readings in Information Visualization: Using Vision to Think* (Morgan Kaufmann, San Francisco, 1999).
12. A. Perer, B. Shneiderman, "Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis," *ACM Conference on Human Factors in Computing Systems (CHI 2008)*, Florence, Italy, 5 to 10 April 2008.
13. C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, J. C. Hart, *Commun. ACM* **35**, 64 (1992).

10.1126/science.1197654

PERSPECTIVE

Challenges and Opportunities in Mining Neuroscience Data

Huda Akil,[1]* Maryann E. Martone,[2] David C. Van Essen[3]

Understanding the brain requires a broad range of approaches and methods from the domains of biology, psychology, chemistry, physics, and mathematics. The fundamental challenge is to decipher the "neural choreography" associated with complex behaviors and functions, including thoughts, memories, actions, and emotions. This demands the acquisition and integration of vast amounts of data of many types, at multiple scales in time and in space. Here we discuss the need for neuroinformatics approaches to accelerate progress, using several illustrative examples. The nascent field of "connectomics" aims to comprehensively describe neuronal connectivity at either a macroscopic level (in long-distance pathways for the entire brain) or a microscopic level (among axons, dendrites, and synapses in a small brain region). The Neuroscience Information Framework (NIF) encompasses all of neuroscience and facilitates the integration of existing knowledge and databases of many types. These examples illustrate the opportunities and challenges of data mining across multiple tiers of neuroscience information and underscore the need for cultural and infrastructure changes if neuroinformatics is to fulfill its potential to advance our understanding of the brain.

Deciphering the workings of the brain is the domain of neuroscience, one of the most dynamic fields of modern biology. Over the past few decades, our knowledge about the nervous system has advanced at a remarkable pace. These advances are critical for understanding the mechanisms underlying the broad range of brain functions, from controlling breathing to forming complex thoughts. They are also essential for uncovering the causes of the vast array of brain disorders, whose impact on humanity is staggering (*1*). To accelerate progress, it is vital to develop more powerful methods for capitalizing on the amount and diversity of experimental data generated in association with these discoveries.

The human brain contains ~80 billion neurons that communicate with each other via specialized connections or synapses (*2*). A typical adult brain has ~150 trillion synapses (*3*). The point of all this communication is to orchestrate brain activity. Each neuron is a piece of cellular machinery that relies on neurochemical and electrophysiological mechanisms to integrate complicated inputs and communicate information to other neurons. But no matter how accomplished, a single neuron can never perceive beauty, feel sadness, or solve a mathematical problem. These capabilities emerge only when networks of neurons work together. Ensembles of brain cells, often quite far-flung, form integrated neural circuits, and the activity of the network as a whole supports specific brain functions such as perception, cognition, or emotions. Moreover, these circuits are not static. Environmental events trigger molecular mechanisms of neuroplasticity that alter the morphology and connectivity of brain cells. The strengths and pattern of synaptic connectivity encode the "software" of brain function. Experience, by inducing changes in that connectivity, can substantially alter the function of specific circuits during development and throughout the life span.

A grand challenge in neuroscience is to elucidate brain function in relation to its multiple layers of organization that operate at different spatial and temporal scales. Central to this effort is tackling "neural choreography": the integrated functioning of neurons into brain circuits, including their spatial organization, local, and long-distance connections; their temporal orchestration; and their dynamic features, including interactions with their glial cell partners. Neural choreography cannot be understood via a purely reductionist approach. Rather, it entails the convergent use of analytical and synthetic tools to gather, analyze, and mine information from each level of analysis and capture the emergence of new layers of function (or dysfunction) as we move from studying genes and proteins, to cells, circuits, thought, and behavior.

The Need for Neuroinformatics

The profoundly complex nature of the brain requires that neuroscientists use the full spectrum of tools available in modern biology: genetic,

[1]The Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, MI, USA. [2]National Center for Microscopy and Imaging Research, Center for Research in Biological Systems, University of California, San Diego, La Jolla, CA, USA. [3]Deparment of Anatomy and Neurobiology, Washington University School of Medicine, St. Louis, MO 63110, USA.

*To whom correspondence should be addressed. E-mail: akil@umich.edu

cellular, anatomical, electrophysiological, behavioral, evolutionary, and computational. The experimental methods involve many spatial scales, from electron microscopy (EM) to whole-brain human neuroimaging, and time scales ranging from microseconds for ion channel gating to years for longitudinal studies of human development and aging. An increasing number of insights emerge from integration and synthesis across these spatial and temporal domains. However, such efforts face impediments related to the diversity of scientific subcultures and differing approaches to data acquisition, storage, description, and analysis, and even to the language in which they are described. It is often unclear how best to integrate the linear information of genetic sequences, the highly visual data of neuroanatomy, the time-dependent data of electrophysiology, and the more global level of analyzing behavior and clinical syndromes.

The great majority of neuroscientists carry out highly focused, hypothesis-driven research that can be powerfully framed in the context of known circuits and functions. Such efforts are complemented by a growing number of projects that provide large data sets aimed not at testing a specific hypothesis but instead enabling data-intensive discovery approaches by the community at large. Notable successes include gene expression atlases from the Allen Institute for Brain Sciences (*4*) and the Gene Expression Nervous System Atlas (GENSAT) Project (*5*), and disease-specific human neuroimaging repositories (*6*). However, the neuroscience community is not yet fully engaged in exploiting the rich array of data currently available, nor is it adequately poised to capitalize on the forthcoming data explosion.

Below we highlight several major endeavors that provide complementary perspectives on the challenges and opportunities in neuroscience data mining. One is a set of "connectome" projects that aim to comprehensively describe neural circuits at either the macroscopic or the microscopic level. Another, the Neuroscience Information Framework (NIF), encompasses all of neuroscience and provides access to existing knowledge and databases of many types. These and other efforts provide fresh approaches to the challenge of elucidating neural choreography.

Connectomes: Macroscopic and Microscopic

Brain anatomy provides a fundamental three-dimensional framework around which many types of neuroscience data can be organized and mined. Decades of effort have revealed immense amounts of information about local and long-distance connections in animal brains. A wide range of tools (such as immunohistochemistry and in situ hybridization) have characterized the biochemical nature of these circuits that are studied electrophysiologically, pharmacologically, and behaviorally (*7*). Several ongoing efforts aim to integrate anatomical information into searchable resources that provide a backbone for understanding circuit biology and function (*8*–*10*). The challenge of integrating such data will dramatically increase with the advent of high-throughput anatomical methods, including those emerging from the nascent field of connectomics.

A connectome is a comprehensive description of neural connectivity for a specified brain region at a specified spatial scale (*11*, *12*). Connectomics currently includes distinct subdomains for studying the macroconnectome (long-distance pathways linking patches of gray matter) and the microconnectome (complete connectivity within a single gray-matter patch).

The Human Connectome Project. Until recently, methods for charting neural circuits in the human brain were sorely lacking (*13*). This situation has changed dramatically with the advent of noninvasive neuroimaging methods. Two complementary modalities of magnetic resonance imaging (MRI) provide the most useful information about long-distance connections. One modality uses diffusion imaging to determine the orientation of axonal fiber bundles in white matter, based on the preferential diffusion of water molecules parallel to these fiber bundles. Tractography is an analysis strategy that uses this information to estimate long-distance pathways linking different gray-matter regions (*14*, *15*). A second modality, resting-state functional MRI (R-fMRI), is based on slow fluctuations in the standard fMRI BOLD signal that occur even when people are at rest. The time courses of these fluctuations are correlated across gray-matter locations, and the spatial patterns of the resultant functional connectivity correlation maps are closely related but not identical to the known pattern of direct anatomical connectivity (*16*, *17*). Diffusion imaging and R-fMRI each have important limitations, but together they offer powerful and complementary windows on human brain connectivity.

To address these opportunities, the National Institutes of Health (NIH) recently launched the Human Connectome Project (HCP) and awarded grants to two consortia (*18*). The consortium led by Washington University in St. Louis and the University of Minnesota (*19*) aims to characterize whole-brain circuitry and its variability across individuals in 1200 healthy adults (300 twin pairs and their nontwin siblings). Besides diffusion imaging and R-fMRI, task-based fMRI data will be acquired in all study participants, along with extensive behavioral testing; 100 participants will also be studied with magnetoencephalography (MEG) and electroencephalography (EEG). Acquired blood samples will enable genotyping or full-genome sequencing of all participants near the end of the 5-year project. Currently, data acquisition and analysis methods are being extensively refined using pilot data sets. Data acquisition from the main cohort will commence in mid-2012.

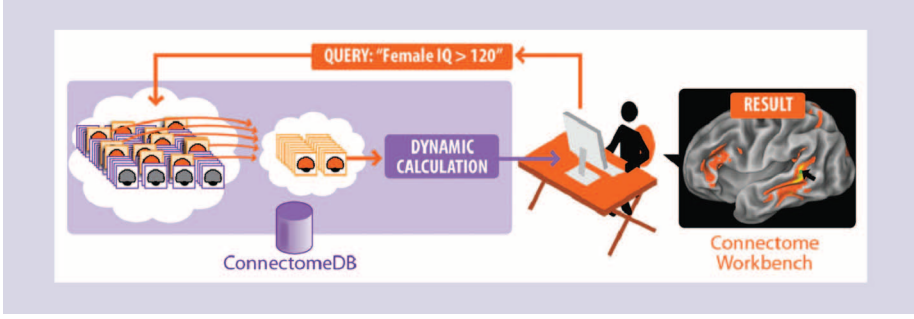

Fig. 1. Schematic illustration of online data mining capabilities envisioned for the HCP. Investigators will be able to pose a wide range of queries (such as the connectivity patterns of a particular brain region of interest averaged across a group of individuals, based on behavioral criteria) and view the search results interactively on three-dimensional brain models. Data sets of interest will be freely available for downloading and additional offline analysis.

Neuroimaging and behavioral data from the HCP will be made freely available to the neuroscience community via a database (*20*) and a platform for visualization and user-friendly data mining. This informatics effort involves major challenges owing to the large amounts of data (expected to be ~1 petabyte), the diversity of data types, and the many possible types of data mining. Some investigators will drill deeply by analyzing high-resolution connectivity maps between all gray-matter locations. Others will explore a more compact "parcellated connectome" among all identified cortical and subcortical parcels. Data mining options will reveal connectivity differences between subpopulations that are selected by behavioral phenotype (such as high versus low IQ) and various other characteristics (Fig. 1). The utility of HCP-generated data will be enhanced by close links to other resources containing complementary types of spatially organized data, such as the Allen Human Brain Atlas (*21*), which contains neural gene expression maps.

Microconnectomes. Recent advances in serial section EM, high-resolution optical imaging methods, and sophisticated image segmentation methods enable detailed reconstructions of the

microscopic connectome at the level of individual synapses, axons, dendrites, and glial processes (*22*–*24*). Current efforts focus on the reconstruction of local circuits, such as small patches of the cerebral cortex or retina, in laboratory animals. As such data sets begin to emerge, a fresh set of informatics challenges will arise in handling petabyte amounts of primary and analyzed data and in providing data mining platforms that enable neuroscientists to navigate complex local circuits and examine interesting statistical characteristics.

Micro- and macroconnectomes exemplify distinct data types within particular tiers of analysis that will eventually need to be linked. Effective interpretation of both macro- and microconnectomic approaches will require novel informatics and computational approaches that enable these two types of data to be analyzed in a common framework and infrastructure. Efforts such as the Blue Brain Project (*25*) represent an important initial thrust in this direction, but the endeavor will entail decades of effort and innovation.

Powerful and complementary approaches such as optogenetics operate at an intermediate (mesoconnectome) spatial scale by directly perturbing neural circuits in vivo or in vitro with light-activated ion channels inserted into selected neuronal types (*26*). Other optical methods, such as calcium imaging with two-photon laser microscopy, enable analysis of the dynamics of ensembles of neurons in microcircuits (*27*, *28*) and can lead to new conceptualizations of brain function (*29*). Such approaches provide an especially attractive window on neural choreography as they assess or perturb the temporal patterns of macro- or microcircuit activity.

The NIF

Connectome-related projects illustrate ways in which neuroscience as a field is evolving at the level of neural circuitry. Other discovery efforts include genome-wide gene expression profiling [for example, (*30*)] or epigenetic analyses across multiple brain regions in normal and diseased brains. This wide range of efforts results in a sharp increase in the amount and diversity of data being generated, making it unlikely that neuroscience will be adequately served by only a handful of centralized databases, as is largely the case for the genomics and proteomics community (*31*). How, then, can we access and explore these resources more effectively to support the data-intensive discovery envisioned in *The Fourth Paradigm* (*32*)?

Tackling this question was a prime motivation behind the NIF (*33*). The NIF was launched in 2005 to survey the current ecosystem of neuroscience resources (databases, tools, and materials) and to establish a resource description framework and search strategy for locating, accessing, and using digital neuroscience-related resources (*34*).

The NIF catalog, a human-curated registry of known resources, currently includes more than 3500 such resources, and new ones are added daily. Over 2000 of these resources are databases that range in size from hundreds to millions of records. Many were created at considerable effort and expense, yet most of them remain underused by the research community.

Clearly, it is inefficient for individual researchers to sequentially visit and explore thousands of databases, and conventional online search engines are inadequate, insofar as they do not effectively index or search database content. To promote the discovery and use of online databases, the NIF created a portal through which users can search not only the NIF registry but the content of multiple databases simultaneously. The current NIF federation includes more than 65 databases accessing ~30 million records (*35*) in major domains of relevance to neuroscience (Fig. 2). Besides very large genomic collections, there are nearly 1 million antibody records, 23,000 brain connectivity records, and >50,000 brain activation coordinates. Many of these areas are covered by multiple databases, which the NIF knits together into a coherent view. Although impressive, this represents only the tip of the iceberg. Most individual databases are underpopulated because of insufficient community contributions. Entire domains of neuroscience (such as electrophysiology and behavior) are underrepresented as compared to genomics and neuroanatomy.

Ideally, NIF users should be able not only to locate answers that are known but to mine available data in ways that spur new hypotheses regarding what is not known. Perhaps the single biggest roadblock to this higher-order data mining is the lack of standardized frameworks for organizing neuroscience data. Individual investigators often use terminology or spatial coordinate systems customized for their own particular analysis approaches. This customization is a substantial barrier to data integration, requiring considerable human effort to access each resource, understand the context and content of the data, and determine the conditions under which they can be compared to other data sets of interest.

To address the terminology problem, the NIF has assembled an expansive lexicon and ontology covering the broad domains of neuroscience by synthesizing open-access community ontologies (*36*). The Neurolex and accompanying NIFSTD (NIF-standardized) ontologies provide definitions of over 50,000 concepts, using formal languages to represent brain regions, cells, subcellular structures, molecules, diseases, and functions, and the relations among them. When users search for a concept through the NIF, it automatically expands the query to include all synonymous or closely related terms. For example, a query for "striatum" will include "neostriatum, dorsal striatum, caudoputamen, caudate putamen" and other variants.

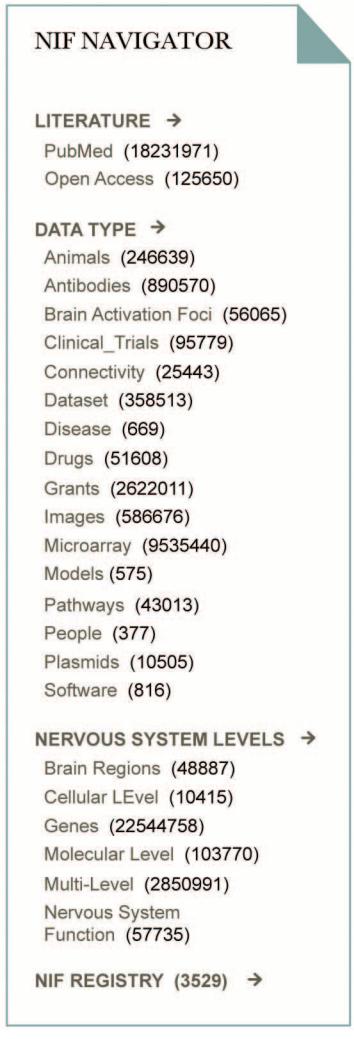

Fig. 2. Current contents of the NIF. The NIF navigation bar displays the current contents of the NIF data federation, organized by data type and level of the nervous system. The number of records in each category is displayed in parentheses.

Neurolex terms are accessible through a wiki (*37*) that allows users to view, augment, and modify these concepts. The goal is to provide clear definitions of each concept that can be used not only by humans but by automated agents, such as the NIF, to navigate the complexities of human neuroscience knowledge. A key feature is the assignment of a unique resource identifier to make it easier for search algorithms to distinguish among concepts that share the same label. For example, nucleus (part of cell) and nucleus (part of brain) are distinguished by unique IDs. Using these identifiers in addition to natural language to reference concepts in databases and publications, although conceptually simple, is an especially powerful means for making data maximally discoverable and useful.

These efforts to develop and deploy a semantic framework for neuroscience, spearheaded by

the NIF and the International Neuroinformatics Coordinating Facility (*38*), are complemented by projects related to brain atlases and spatial frameworks (*39*–*41*), providing tools for referencing data to a standard coordinate system based on the brain anatomy of a given organism.

Neuroinformatics as a Prelude to New Discoveries

How might improved access to multiple tiers of neurobiological data help us understand the brain? Imagine that we are investigating the neurobiology of bipolar disorder, an illness in which moods are normal for long periods of time, yet are labile and sometimes switch to mania or depression without an obvious external trigger. Although highly heritable, this disease appears to be genetically very complex and possibly quite heterogeneous (*42*). We may discover numerous genes that impart vulnerability to the illness. Some may be ion channels, others synaptic proteins or transcription factors. How will we uncover how disparate genetic causes lead to a similar clinical phenotype? Are they all affecting the morphology of certain cells; the dynamics of specific microcircuits, for example, within the amygdala; the orchestration of information across regions, for example, between the amygdala and the prefrontal cortex? Can we create genetic mouse models of the various mutated genes and show a convergence at any of these levels? Can we capture the critical changes in neuronal and/or glial function (at any of the levels) and find ways to prevent the illness? Discovering the common thread for such a disease will surely benefit from tools that facilitate navigation across the multiple tiers of data—genetics, gene expression/epigenetics, changes in neuronal activity, and differences in dynamics at the micro and macro levels, depending on the mood state. No single focused level of analysis will suffice to achieve a satisfactory understanding of the disease. In neural choreography terms, we need to identify the dancers, define the nature of the dance, and uncover how the disease disrupts it.

Recommendations

Need for a cultural shift. To meet the grand challenge of elucidating neural choreography, we need increasingly powerful scientific tools to study brain activity in space and in time, to extract the key features associated with particular events, and to do so on a scale that reveals commonalities and differences between individual brains. This requires an informatics infrastructure that has built-in flexibility to incorporate new types of data and navigate across tiers and domains of knowledge.

The NIF currently provides a platform for integrating and systematizing existing neuroscience knowledge and has been working to define best practices for those producing new neuroscience data. Good planning and future investment are needed to broaden and harden the overall framework for housing, analyzing, and integrating future neuroscience knowledge. The International Neuroinformatics Coordinating Facility (INCF) plays an important role in coordinating and promoting this framework at a global level.

But can neuroscience evolve so that neuroinformatics becomes integral to how we study the brain? This would entail a cultural shift in the field regarding the importance of data sharing and mining. It would also require recognition that neuroscientists produce data not just for consumption by readers of the conventional literature, but for automated agents that can find, relate, and begin to interpret data from databases as well as the literature. Search technologies are advancing rapidly, but the complexity of scientific data continues to challenge. To make neuroscience data maximally interoperable within a global neuroscience information framework, we encourage the neuroscience community and the associated funding agencies to consider the following set of general and specific suggestions:

1) Neuroscientists should, as much as is feasible, share their data in a form that is machine-accessible, such as through a Web-based database or some other structured form that benefits from increasingly powerful search tools.

2) Databases spanning a growing portion of the neuroscience realm need to be created, populated, and sustained. This effort needs adequate support from federal and other funding mechanisms.

3) Because databases become more useful as they are more densely populated (*43*), adding to existing databases may be preferable to creating customized new ones. The NIF, INCF, and other resources provide valuable tools for finding existing databases.

4) Data consumption will increasingly involve machines first and humans second. Whether creating database content or publishing journal articles, neuroscientists should annotate content using community ontologies and identifiers. Coordinates, atlas, and registration method should be specified when referencing spatial locations.

5) Some types of published data (such as brain coordinates in neuroimaging studies) should be reported in standardized table formats that facilitate data mining.

6) Investment needs to occur in interdisciplinary research to develop computational, machine-learning, and visualization methods for synthesizing across spatial and temporal information tiers.

7) Educational strategies from undergraduate through postdoctoral levels are needed to ensure that neuroscientists of the next generation are proficient in data mining and using the data-sharing tools of the future.

8) Cultural changes are needed to promote widespread participation in this endeavor. These ideas are not just a way to be responsible and collaborative; they may serve a vital role in attaining a deeper understanding of brain function and dysfunction.

With such efforts, and some luck, the machinery that we have created, including powerful computers and associated tools, may provide us with the means to comprehend this "most unaccountable of machinery" (*44*), our own brain.

References and Notes

1. World Health Organization, *Mental Health and Development: Targeting People With Mental Health Conditions As a Vulnerable Group* (World Health Organization, Geneva, 2010).
2. F. A. Azevedo *et al.*, *J. Comp. Neurol.* **513**, 532 (2009).
3. B. Pakkenberg *et al.*, *Exp. Gerontol.* **38**, 95 (2003).
4. www.alleninstitute.org/
5. www.gensat.org/
6. http://adni.loni.ucla.edu
7. A. Bjorklund, T. Hokfelt, Eds., *Handbook of Chemical Neuroanatomy Book Series*, vols. 1 to 21 (Elsevier, Amsterdam, 1983–2005).
8. http://cocomac.org/home.asp
9. http://brancusi.usc.edu/bkms/
10. http://brainmaps.org/
11. http://en.wikipedia.org/wiki/Connectome
12. O. Sporns, G. Tononi, R. Kotter, *PLoS Comput. Biol.* **1**, e42 (2005).
13. F. Crick, E. Jones, *Nature* **361**, 109 (1993).
14. H. Johansen-Berg, T. E. J. Behrens, *Diffusion MRI: From Quantitative Measurement to in-vivo Neuroanatomy* (Academic Press, London, ed. 1, 2009).
15. H. Johansen-Berg, M. F. Rushworth, *Annu. Rev. Neurosci.* **32**, 75 (2009).
16. J. L. Vincent *et al.*, *Nature* **447**, 83 (2007).
17. D. Zhang, A. Z. Snyder, J. S. Shimony, M. D. Fox, M. E. Raichle, *Cereb. Cortex* **20**, 1187 (2010).
18. http://humanconnectome.org/consortia
19. http://humanconnectome.org
20. D. S. Marcus, T. R. Olsen, M. Ramaratnam, R. L. Buckner, *Neuroinformatics* **5**, 11 (2007).
21. http://human.brain-map.org/
22. K. L. Briggman, W. Denk, *Curr. Opin. Neurobiol.* **16**, 562 (2006).
23. J. W. Lichtman, J. Livet, J. R. Sanes, *Nat. Rev. Neurosci.* **9**, 417 (2008).
24. S. J. Smith, *Curr. Opin. Neurobiol.* **17**, 601 (2007).
25. H. Markram, *Nat. Rev. Neurosci.* **7**, 153 (2006).
26. K. Deisseroth, *Nat. Methods* **8**, 26 (2011).
27. B. F. Grewe, F. Helmchen, *Curr. Opin. Neurobiol.* **19**, 520 (2009).
28. B. O. Watson *et al., Front. Neurosci.* **4**, 29 (2010).
29. G. Buzsaki, *Neuron* **68**, 362 (2010).
30. R. Bernard *et al., Mol. Psychiatry*, 13 April 2010 (e-pub ahead of print).
31. M. E. Martone, A. Gupta, M. H. Ellisman, *Nat. Neurosci.* **7**, 467 (2004).
32. *The Fourth Paradigm: Data Intensive Scientific Discovery*, T. Hey, S. Tansler, K. Tolle, Eds. (Microsoft Research Publishing, Redmond, WA, 2009).
33. http://neuinfo.org
34. D. Gardner *et al.*, *Neuroinformatics* **6**, 149 (2008).
35. A. Gupta *et al.*, *Neuroinformatics* **6**, 205 (2008).
36. W. J. Bug *et al.*, *Neuroinformatics* **6**, 175 (2008).
37. http://neurolex.org
38. http://incf.org
39. http://www.brain-map.org
40. http://wholebraincatalog.org
41. http://incf.org/core/programs/atlasing
42. H. Akil *et al.*, *Science* **327**, 1580 (2010).

43. G. A. Ascoli, *Nat. Rev. Neurosci.* **7**, 318 (2006).
44. N. Nicolson, J. Trautmann, Eds., *The Letters of Virginia Woolf, Volume V: 1932–1935* (Harcourt, Brace, New York, 1982).
45. Funded in part by grants from NIH (5P01-DA021633-02), the Office of Naval Research (ONR-N00014-02-1-0879), and the Pritzker Neuropsychiatric Research Consortium (to H.A); and the HCP (1U54MH091657-01) from the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research (to D.C.V.E.). The NIF is supported by NIH Neuroscience Blueprint contract HHSN271200800035C via the National Institute on Drug Abuse to M.E.M. We thank D. Marcus, R. Poldrack, S. Curtiss, A. Bandrowski, and S. J. Watson for discussions and comments on the manuscript.

10.1126/science.1199305

PERSPECTIVE

The Disappearing Third Dimension

Timothy Rowe[1] and Lawrence R. Frank[2]

Three-dimensional computing is driving what many would call a revolution in scientific visualization. However, its power and advancement are held back by the absence of sustainable archives for raw data and derivative visualizations. Funding agencies, professional societies, and publishers each have unfulfilled roles in archive design and data management policy.

Three-dimensional (3D) image acquisition, analysis, and visualization are increasingly important in nearly every field of science, medicine, engineering, and even the fine arts. This reflects rapid growth of 3D scanning instrumentation, visualization and analysis algorithms, graphic displays, and graduate training. These capabilities are recognized as critical for future advances in science, and the U.S. National Science Foundation (NSF) is one of many funding agencies increasing support for 3D imaging and computing. For example, one new initiative aims at digitizing biological research collections, and NSF's Earth Sciences program will soon announce a new target in cyberinfrastructure, with a spotlight on 3D imaging of natural materials.

Many consider the advent of 3D imaging a "scientific revolution" (*1*), but in many ways the revolution is still nascent and unfulfilled. Given the increasing ease of producing these data and rapidly increased funding for imaging in nonmedical applications, a major unmet challenge to ensure maximal advancement is archiving and managing the science that will be, and even has already been, produced. To illustrate the problems, we focus here on one domain, volume elements or "voxels," the 3D equivalents of pixels, but the argument applies more broadly across 3D computing. Useful, searchable archiving requires infrastructure and policy to enable disclosure by data producers and to guarantee quality to data consumers (*2*). Voxel data have generally lacked a coherent archiving and dissemination policy, and thus the raw data behind thousands of published reports are not released or available for validation and reuse. A solution requires new infrastructure and new policy to manage ownership and release (*3*).

Technological advancements in both medical and industrial scanners have increased the resolution of 3D volumetric scanners to the point that they can now volumetrically digitize structures from the size of a cell to a blue whale, with exquisite sensitivity toward scientific targets ranging from tissue properties to the composition of meteorites. Voxel data sets are generated by rapidly diversifying instruments that include x-ray computed tomographic (CT) scanners (Fig. 1), magnetic resonance imaging (MRI) scanners (Fig. 2), confocal microscopes, synchrotron light sources, electron-positron scanners, and other innovative tools to digitize entire object volumes.

CT and MRI have evolved across the widest range of applications. CT is sensitive to density. Its greatest strength is imaging dense materials like rocks, fossils, the bones in living organisms and, to a lesser extent, soft tissue. The first clinical CT scan, made in 1971, used an 80 by 80 matrix of 3 mm by 3 mm by 13 mm voxels, each slice measuring 6.4 Kb and taking up to 20 min to acquire. Each slice image took 7 min to reconstruct on a mainframe computer (*4*). In 1984, the first fossil, a 14-cm-long skull of the extinct mammal *Stenopsochoerus*, was scanned in its entirety (*5*), signaling CT's impact beyond the clinical setting and in digitizing entire object volumes. The complete data set measured 3.28 Mb. With a mainframe computer, rock matrix was removed to visualize underlying bone, and surface models were reconstructed in computations taking all night. By 1992, industrial adaptations brought CT an order-of-magnitude higher resolution (high-resolution x-ray computed tomography, HRXCT) to inspect much smaller, denser objects. A fossil skull of the stem-mammal *Thrinaxodon* (68 mm long) was scanned in 0.2-mm-thick slices measuring 119 Kb each. Scanning the entire volume took 6 hours, and the complete raw data set occupied 18.4 Mb (*6*). The scans revealed all internal details reported earlier, from destructive mechanical serial sectioning of a different specimen, and pushed

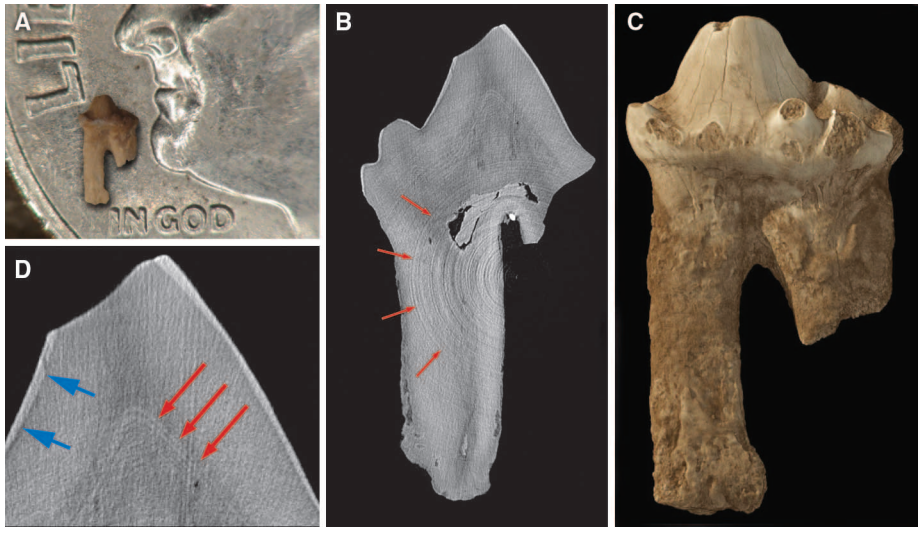

Fig. 1. (**A**) Photomicrograph of fossil tooth (*Morganucodon* sp.). (**B**) MicroXCT slice, 3.2-μm voxel size; arrows show ring artifact, a correctable problem with raw data but an interpretive challenging with compressed data. (**C**) Digital 3D reconstruction. (**D**) Slice through main cusp; red arrows show growth bands in dentine, and blue arrows mark the enamel-dentine boundary, both observed in *Morganucodon* for the first time in these scans.

[1]Jackson School of Geosciences, The University of Texas, Austin, TX 78712, USA. [2]Department of Radiology, University of California, San Diego, CA 92093, USA.

*To whom correspondence should be addressed. E-mail: rowe@mail.utexas.edu (T.R.); lfrank@ucsd.edu (L.R.F.)