Changing the Equation on Scientific Data Visualization

Peter Fox, *et al.*

Science **331**, 705 (2011);
DOI: 10.1126/science.1197654

With a tug from software to manage genomic data online and a push from publishers unwilling to continue editing and printing the growing volume of gene sequences, a robust data repository for gene sequences was born. Today, after almost 30 years, registering gene sequences and sharing them broadly is the norm and is recognized as fostering one of the greatest scientific revolutions in the past century.

Ecology is poised for a similar transformation. The pull comes from a need for data in synthesis and cross-cutting analysis that is facilitated by the emergence of community metadata standards and federated data repositories that span adjacent disciplines. The push is coming from funding entities that are requiring open access to data, with a dose of urgency engendered by the chronic and acute environmental degradation occurring globally. Furthermore, the rewards for sharing data are increasing. As noted, it is possible to publish peer-reviewed, citable data sets in repositories while giving credit to the data contributors, and there is evidence that published papers that do make available their data are cited more frequently than those that do not (*21*).

We have presented some of the major challenges and emerging solutions for dealing with the vast volume and heterogeneity of ecological data. To accelerate the advance of ecological understanding and its application to critical environmental concerns, we must move to the next level of information management by providing revolutionary new data-management applications, promoting their adoption, and hastening the emergence of communities of practice. Concurrently, we must encourage the growing culture of collaboration and synthesis that has emerged in ecology that is fundamentally altering the scientific method to require comprehensive data sharing, as well as greater reproducibility and transparency of the methods and analyses that support scientific insights.

References and Notes

1. S. Carpenter *et al.*, *Bioscience* **59**, 699 (2009).
2. E. Hackett, J. Parker, D. Conz, D. Rhoten, A. Parker, in *Scientific Collaboration on the Internet*, G. M. Olson *et al.*, Eds. (MIT Press, Boston, 2008), pp. 277–296.
3. T. J. Crone, M. Tolstoy, *Science* **330**, 634 (2010).
4. Florida Coastal Everglades Data Resources, http://fce.lternet.edu/data/FCE.
5. J. W. Tunnell, Q. R. Dokken, M. E. Kindinger, L. C. Thebeau, "Effects of the Ixtoc I oil spill on intertidal and subtidal infaunal populations along lower Texas coast barrier island beaches," in *Proceedings of the 1981 Oil Spill Conference* (American Petroleum Institute, Washington, DC, 1981), pp. 467–475.
6. C. J. Savage, A. J. Vickers, C. Mavergames, *PLoS ONE* **4**, e7078 (2009).
7. P. B. Heidorn, *Libr. Trends* **57**, 280 (2008).
8. W. K. Michener, *Ecol. Inform.* **1**, 3 (2006).
9. M. B. Jones, M. Schildhauer, O. J. Reichman, S. Bowers, *Annu. Rev. Ecol. Evol. Syst.* **37**, 519 (2006).
10. V. S. Chavan *et al.*, *State-of-the-Network 2010: Discovery and Publishing of the Primary Biodiversity Data Through the GBIF Network* (Global Biodiversity Information Facility, Copenhagen, 2010).
11. S. J. Andelman, C. Bowles, M. R. Willig, R. Waide, *Bioscience* **54**, 240 (2004).
12. The Knowledge Network for Biocomplexity, http://knb.ecoinformatics.org.
13. H. White, S. Carrier, A. Thompson, J. Greenberg, R. Scherle, "The Dryad data repository: A Singapore framework metadata architecture in a DSpace environment," in *Proceedings of the International Conference on Dublin Core and Metadata Applications*, J. Greenberg, W. Klas, Eds. (Dublin Core Metadata Initiative and Universitätsverlag Göttingen, Berlin, 2008), pp. 157–162.
14. I. San Gil, V. Hutchison, G. Palanisamy, M. Frame, *J. Libr. Metadata* **10**, 99 (2010).
15. 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* (Association for Computing Machinery, New York, 2008), pp. 1265–1266.
16. J. Madin, J. S. Bowers, M. Schildhauer, M. B. Jones, *Trends Ecol. Evol.* **23**, 159 (2008).
17. P. Buneman, S. Khanna, W.-C. Tan, *Lect. Notes Comput. Sci.* **1974**, 87 (2000).
18. W. Sutherland, A. Pullin, P. Dolman, T. Knight, *Trends Ecol. Evol.* **19**, 305 (2004).
19. P. Missier *et al.*, Linking Multiple Workflow Provenance Traces for Interoperable Collaborative Science., Presentation at WORKS 2010: 5th Workshop on Workflows in Support of Large-Scale Science, IEEE Computer Society, New Orleans, 14 November 2010.
20. T. Vision, *Bioscience* **60**, 330 (2010).
21. H. A. Piwowar, R. S. Day, D. B. Fridsma, J. Ioannidis, *PLoS ONE* **2**, e308 (2007).

22. Supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (grant EF-0553768), the University of California, Santa Barbara, and the State of California. Additional support for M.B.J. was provided by NSF grant OCI-0830944 and for O.J.R. by NSF grant DEB-0444217.

10.1126/science.1197962

PERSPECTIVE

Changing the Equation on Scientific Data Visualization

Peter Fox and James Hendler*

An essential facet of the data deluge is the need for different types of users to apply visualizations to understand how data analyses and queries relate to each other. Unfortunately, visualization too often becomes an end product of scientific analysis, rather than an exploration tool that scientists can use throughout the research life cycle. However, new database technologies, coupled with emerging Web-based technologies, may hold the key to lowering the cost of visualization generation and allow it to become a more integral part of the scientific process.

A critical aspect of the data deluge is the need for users, whether they are scientists themselves, funders of science, or the concerned public, to be able to discover the relations among and between the results of data analyses and queries. Unfortunately, the creation of visualizations for complex data remains more of an art form than an easily conducted practice. What's more, especially for big science, the resource cost of creating useful visualizations is increasing: Although it was recently assumed that data-centric science required a rough split between the time to generate, analyze, and publish data (*1*), today the visualization and analysis component has become a bottleneck, requiring considerably more of the overall effort. This trend will continue to get worse as new technologies for data generation are decreasing in price at an incredible rate (in terms of cost per data generated), whereas visualization costs are falling much more slowly. As a result of these trends, the extra effort of making our data understandable, something that should be routine, is consuming considerable resources that could be used for many other purposes.

A consequence of the major effort for visualization is that it becomes an end product of scientific analysis, rather than an exploration tool allowing scientists to form better hypotheses in the continually more data-intensive scientific process. However, new database technologies and promising Web-based visualization approaches may be vital for reducing the cost of visualization generation and allowing it to become a central piece of the scientific process. As an anecdotal example, consider the papers in the recently published *The Fourth Paradigm*, a collection of invited essays about the emerging area of data-intensive science (*2*). Only one of the more than 30 papers is primarily about visualization needs, but virtually all of the essays include visualizations that show off particular scientific results.

From Presentation

In the computing sciences, visualization has been in the hands of two communities. The first is the

Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180, USA.

*To whom correspondence should be addressed. E-mail: hendler@cs.rpi.edu

human-computer interaction (HCI) community, which has considered visualization to be an important technology to study in its own right. Many of the tools used by scientists were developed by HCI practitioners as spin-offs of more general visualization technologies. The second is the graphics community, which has often been focused on the hardware for creating high-quality visualizations in science and other communities. The work of these research communities has led to very exciting capabilities including large-scale "immersive environments" (*3*), high-end three-dimensional displays, rendering software kits, and visualization libraries, among others.

Visualizations are absolutely critical to our ability to process complex data and to build better intuitions as to what is happening around us. For instance, consider how online display presentations, rather than text-only content, has enriched weather forecasts and allowed anyone to explore data ranging from precipitation to wind fields to temperature. These weather displays generally include the means to provide contextualization of the meteorological data (such as location, time, and key annotations) and to animate it. In addition, as weather data has moved to the Web, interactive visualizations have become more common. For example, through your browser you can localize the reports, look at video images, and link to many other sources.

Many other data visualizations on the Web are becoming increasingly more sophisticated and interactive, while at the same time becoming easier to generate thanks to the prevalence of open Application Programming Interfaces and interactive Web-based visualization tool kits. Applications now allow the rapid generation of maps, charts, timelines, graphs, word clouds, search interfaces, RSS feeds, and many others capabilities. Additionally, owing to Web-based linking technologies, the visualizations change as the data changes (because they can make use of direct links to the data), which drastically reduces the effort to keep the visualizations timely and up to date. These "low-end" visualizations are often used in business analytics, open government data systems, and media infographics, but they have generally not been used in the scientific process.

Despite the increasing prevalence of these techniques on the Web, we are often looking at tables of numbers, best-fit curves, or other analytic results rather than being able to use visual means when we interact with the complex scientific data in many fields. Many of the visualization tools that are available to scientists do not allow live linking as do these Web-based tools. Once the visualization is created, it is no longer tied to the data, so that it becomes an immutable information product—as the data changes, the visualization is no longer up to date.

There are many reasons why this is the case, but two of them really dominate the situation. The first is that collecting scientific data is often difficult and instrument-specific. As a result, most scientific data is created in a form and organization that facilitates its generation rather than focusing on its eventual use. The second is the scale of scientific data. The need for more and better data, as well as the continuing increases in our ability to design new data collection devices have continually kept scientists at the leading edge of data users. Thus, data collection, with the use of traditional databases, has mostly focused on the efficiency of query-based retrieval of the collected data, rather than on data exploration.

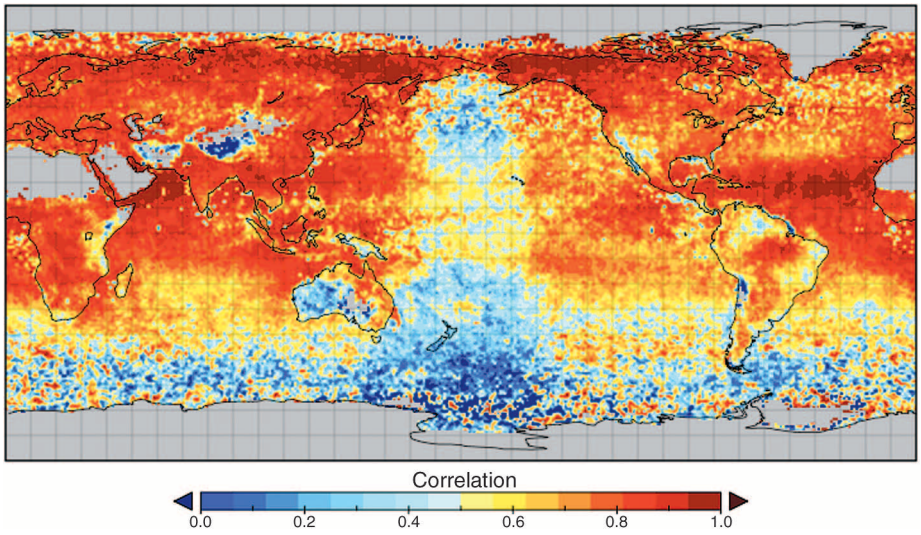

Fig. 1. Correlation of aerosol optical depth for the same instrument [the Moderate Resolution Imaging Spectroradiometer (MODIS)] operating on different satellites (Terra and Aqua) for the year 2008. The visualization reveals that there is a zero-correlation anomaly centered on the date line with the shape of an orbital character. [Image courtesy of NASA/Goddard Space Flight Center (G. Leptoukh)]

The data-scaling problem is particularly exacerbated by the difficulty in linking data from multiple instruments or sources together. Both the speed and precision of data retrieval have often been dependent on the quality of the data models being used, and modeling integrated data from many sources is generally exponentially harder than developing the model for any single source. The challenge is that many of the major scientific problems facing our world are becoming critically linked to the interdependence and interrelatedness of data from multiple instruments, fields, and sources. Consider, for example, climate modeling or translational biology, which increasingly require a systems-level perspective of our world, in the former, and our interactions with that world, in the latter. Although scientists with a deep understanding of their own data can often make sense of the raw data or that from field-specific analytic tools, getting the same sort of intuitions from data generated by different instruments or in a different part of an interdisciplinary study is much harder. Thus, where the need for exploration via visualization grows in importance for these large-scale interdisciplinary efforts, the technical issues of collection and scaling have traditionally worked against that end.

To Exploration

Although many of the largest data sources in the world remain scientific databases, the 20 years of exponential growth in the World Wide Web have brought a number of other players to the massive database table. The backend data technologies that power large Web companies like Google and Facebook and massive online communities such as World of Warcraft are generating huge and diverse data collections. These companies and communities have been pushing new approaches to databases. The wealth of data coming from the interactions of the estimated 1.9 billion users of the Web (*4*) is requiring these and other hubs to rapidly expand and deploy new capabilities, bring new data resources online quickly, and link together extremely large numbers of diverse and difficult-to-model data sources. In these enterprises, the traditional, model-based approaches to data integration, which are still used in most scientific efforts, has given way to new approaches such as NoSQL (*5*), "big data," (*6*) and scalable linked data (*7*, *8*).

Driven by the rapid pace of change on the Web, these new data approaches let visualization be integrated into the analytics process, which allows for the rapid understanding of these vast data holdings. The ability to understand, for example, the constantly changing connectivity of the social network underlying Facebook is crucial to the site's ability to remain responsive to its rapidly growing user community. Adding real-time search capabilities to Google has required a deep understanding of the "twitter sphere" and an ability to track it as it changes and grows.

The capabilities being seen in the Web domain may hold the key to breaking the scientific visualization bottleneck. These new approaches come with two key capabilities: (i) easy-to-use, low-end tools that will allow scientists to rapidly generate visualizations to explore hypotheses and (ii) scalable tools for creating and curating "high-end" visualizations, tied to new approaches in data collection and archiving, making it possible to develop and maintain existing visualizations at lower cost.

However, these tools also create a number of research challenges that the scientific community must tackle. First, new approaches are needed for determining how best to visualize particular kinds of scientific data. A strong start in this direction can be seen in the "Periodic Table of Visualization Methods" developed by Lengler and Epler (*9*), which shows a number of visualization techniques organized by the type of data (or processes) they apply to and the complexity of their application. Additionally, discussion is beginning to move from general principles of effective visualization (*10*) to much more specific advice of use for scientists, such as how best to merge particular kinds of statistics with visualizations (*11*).

A second challenge is to create and maintain data and information provenance for visualizations; often these are key to understanding and fixing data errors. As an example, consider the data shown in Fig. 1, which depicts Earth observation results from two satellites. It becomes immediately clear from this visualization that something odd is happening in the middle of the presentation, where the displayed data quality is clearly different than that elsewhere in the diagram. Understanding the cause of this error, however, requires knowing which observations came from which satellite, the orbital characteristics, and even how a "day" is defined for the data products, but more importantly, knowing when a combination of these factors leads to the artifact displayed (which turns out to be due to overpass time differences; in this case the correlated values are defined up to 22 hours apart at the date line). When known, these time differences can be accounted for and a new and corrected visualization created (*12*).

A related and scientifically critical challenge is the need to ensure that fitness for purpose is well explained in systems that can generate a wide variety of visual analysis products. Factors such as data and information quality, bias, and contextual relevance rarely make their way into visual representations, but they must. There are major efforts under way to meet this set of challenges, but research efforts are still required for a scalable and Web-enabled solution.

Another challenge follows from the desirable features that these visualizations are linked to the underlying data and can change dynamically as the data changes, and that these visualizations can be interactive in a Web context. Though powerful for exploring data, these features drive us toward visualizations that are primarily quantitative, as distinct from ones that may be interpretative or nonrealistic (for instance, a cartoon or a purposefully distorted view). In presenting scientific results, particularly outside of a traditional scientific context, these visualizations can be extremely powerful, but they generally require creative or artistic efforts that are beyond the range of current computational capabilities. Finding ways to couple the fully or semiautomated approaches of data analytics with the more creative, human-based methods so that change in the former can be exploited to help maintain the latter is a clear and emerging challenge for future publication and presentation of scientific results.

Finally, once these visualization products become routine, their management becomes a critical part of the scientific process, which means we must develop techniques that can maintain the visualization products throughout their life cycle. If we think that analysis pipelines are opaque, what about visualization pipelines? It is necessary to the future of scientific problem solving that we can find means to open these visualization pipelines to provide suitable data provenance so that the visualizations can be maintained and reused across our ever-widening and increasingly interdisciplinary scope of scientific problems.

As modern information technologies increasingly allow scientists to take advantage of rapid visualizations to provide for better understanding of what our data tells us about the problems we are solving, we can stop thinking of visualization as a necessary evil at the end of the scientific pipeline and use it as a tool in data comprehension. Similar to businesses that have come to use data analytics as a means to keep pace with a changing world, scientists must explore how better data sharing and visualization technologies will allow them to do the same. This requires scientists to use visualization tools earlier in the process and to document the relations between the data and the visualizations produced.

At the top end of the visualization spectrum, and for those that can afford and maintain them, substantial capabilities are being applied to explore and understand large data sets. At the lower end, when data size is not a limiting factor, we see an improving set of capabilities suitable for scientific use based primarily on Web-based visualization services. In between, however, at a scale where increasingly more scientists work, we do not have routine and scalable capabilities for visualizing the data and information sources we need to advance science. What can be done?

First, we must work with tool designers to make sure that visualizations are sharable during the entire life span of the scientific process. As one example of this, there are a number of standards for both the graphics and the metadata involved in sharing visualizations, but very few of these are supported in current scientific application tools.

Second, there has been little work in the standardization of the workflow and linking technologies needed specifically for high-end scientific

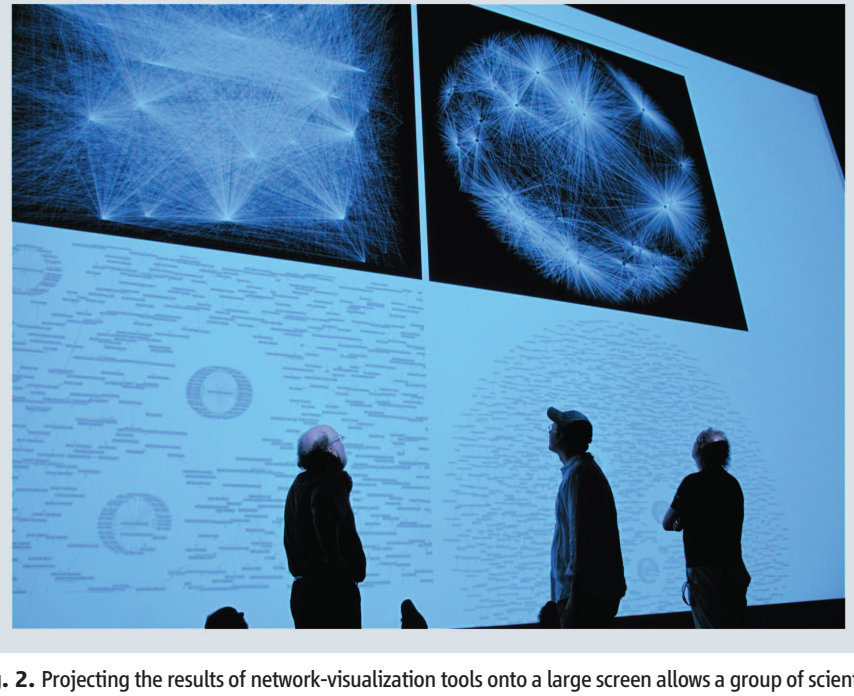

Fig. 2. Projecting the results of network-visualization tools onto a large screen allows a group of scientists to explore the relations among a large number of data elements without the specific need for expensive visualization tools. By using simple visualization techniques such as these, results can be shared early in the research process, rather than waiting to use special-purpose visualization technologies as an end product of scientific analysis.

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