EDITORIAL

nature neuroscience

Focus on big data

Nature Neuroscience **presents a special focus issue highlighting big data efforts under way in the field.**

The number of big data projects in neuroscience, such as the BRAIN initiative in the United States or the Human Brain Mapping initiative in Europe, has been increasing in recent years. Will such big data efforts become the modus operandi in neuroscience, replacing smaller scale, hypothesis-driven science? How much insight will be gained from such projects? What are the best ways to go about conducting such projects and sharing the data that they produce? In this issue, we present a collection of Reviews, Perspectives and Commentaries discussing different kinds of big data in neuroscience and tackling these tough questions.

Although big data projects are relatively new in the neurosciences, molecular biologists have been undertaking such projects for decades. Given the regional and cell-type heterogeneities inherent in the CNS, neuroscience studies need to factor in potential technical caveats and the nature of their source material when performing genome-wide molecular profiling studies of neurons and glia from intact tissues. On page 1463, Jaehoon Shin and colleagues provide a technical primer on generating large transcriptomic and epigenomic data sets from brain tissue. Continuing on this theme, on page 1476, Ian Maze and colleagues provide a comprehensive overview of the tools available for data analysis of epigenomic data sets, efforts and steps necessary to enable data deposition and sharing, and the challenges of integrating this information with other genome-wide and proteomic approaches to fully understand the neuroepigenome and link it to function. In a third Review, Robert Kitchen and colleagues discuss methods for overcoming the challenges associated with collecting large, high-throughput proteomic data sets and compare the insights provided by proteomics with those derived from transcriptomic and genomic data. The authors suggest some strategies for overcoming cellular heterogeneity and for integrating these varied data modalities.

Unlike collecting molecular data, investigating the connectivity and activity of the nervous system is a challenge unique to the neurosciences. Electron microscopy–based connectomics, while still in its infancy, is already producing data sets of staggering size. Data acquisition rates are similarly colossal. As the field moves toward the high-resolution reconstructions of entire fly and mouse brains, this torrent of data will keep growing, and it is unclear how it will be accommodated by computer systems or distributed to end-users. In a piece on page 1448, Lichtman and colleagues discuss the practical and analytical challenges associated with connectomics big data and propose some solutions to overcome them. Along with efforts to map brain connectivity, many projects are currently focused on recording activity from large groups of neurons, either sequentially or simultaneously. As with connectomics, the challenge of such experiments is not over once the data are collected; analyzing such data to provide insights into specific questions can be incredibly challenging. In a Review on page 1500, Cunningham and Yu describe dimensionality reduction methods that are commonly applied to population activity data. They discuss how to choose appropriate analysis methods and also how to interpret results gained from using them. Ultimately, studies of the connectivity and activity of neurons seek to understand the relationship between these neural data and behavior. On page 1455, Alex Gomez-Marin and colleagues review technological advances that have accelerated the collection and analysis of big behavioral data, but argue that substantial challenges remain in interpreting the results of such efforts. They conclude that large-scale quantitative and open behavioral data may transform neuroscience, but only if coupled with new theoretical frameworks and elegant experimental design.

Although many big data efforts focus on model animals, big data is also being generated from humans. Over the last decade, the amount of openly available and shared neuroimaging data has increased substantially. In their Review on page 1510, Poldrack and Gorgolewski discuss the current state of sharing task-based functional magnetic resonance imaging data and the many challenges it poses. Unlike structural or resting-state data, task-based neuroimaging requires more metadata to annotate how it was collected, making integrating across studies difficult. In addition, functional imaging data is most meaningful in relation to baseline or control measures, making it difficult to standardize and collate. It is clear that sharing data increases power and efficiency, but the authors note that institutional and publishing practices must change to promote it.

Although the products of coordinated big data efforts come in a highly standardized format and can be made readily accessible to all, the large majority of neuroscience data sets are small. When small data sets are shared, their reuse is often stymied by a lack of community data-sharing standards. Such data sets have been referred to as 'long-tail' data and, as explained in a Commentary by Martone and colleagues on page 1442 of this issue, they are a potentially important source of new findings. The authors introduce the benefits and caveats associated with data sharing, describe the existing attitudes toward such initiatives, introduce best practices and offer their views on why and how the field should establish a credit system for sharing long-tail data. Once data are shared, the next challenge comes from how to integrate data sets encompassing a vast range of spatial and temporal scales and a myriad of techniques. In their Commentary on page 1440, Sejnowski and colleagues discuss the problems neuroscientists face when trying to integrate diverse data sets into a coherent understanding of brain function. They argue that this will require a cultural shift in not only sharing data across laboratories, but also making theorists central players in its analysis and interpretation.

Although it's impossible to predict the size of the effect big data will have on the way neuroscience research is done and what progress will be made in understanding the brain, it's clear that the wave of big data is not only coming, it's here to stay. The ultimate success or failure of such efforts will be determined by their ability to be integrated with other types of data and by the insights that they provide. We hope that this focus will provide an overview of the types of big data efforts underway in the neurosciences, including the challenges associated with them, and we look forward to the exciting results that follow.