The Fourth Paradigm: Data-Intensive Scientific Discovery by Tony Hey, Stewart Tansley, and Kristin Tolle. Redmond, WA: Microsoft Research, 2009. 284pp. ISBN 978-0982544204.

In a pivotal speech given to the National Research Council on January 11, 2007, a leading expert in computer science named Jim Gray unveiled his vision of the fourth-paradigm, an impending shift in scientific research. Since its inception, Gray argues that scientific methodology has evolved into three distinct archetypes: empiricism, theory, and computation. First, within an empirical framework, science was a technique to systematically characterize and organize natural phenomena, such as in the development of taxonomic descriptions or chemical terminology. Next, experimentation gave way to a theoretical framework where scientific hypotheses could be linked to expected outcomes in a reliable manner. Finally, with the present use of computer technology, scientific data is being obtained computationally without performing a single experiment on the bench top, and in some cases is surpassing the practical limitations of traditional experimentation.

In *The Fourth Paradigm*, evidence is presented to support Gray's notion that scientific methodology is entering an entirely new phase that involves data-intensive practices. Termed "eScience," this fourth paradigm unites theory, experimentation, and computation, and leads to changes in the way science is funded, communicated, and published. *The Fourth Paradigm* is an edited volume of short essays written by academics and experts, and is organized into four major scientific areas: Earth and Environment, Health and Wellbeing, Scientific Infrastructure, and Scholarly Communication. Using examples within these categories, eScience is described by its effects on different aspects of data management. This review intends to highlight some of the works that specifically pertain to data capture, analysis, and sharing.

The amount of scientific data being created each year by scientists is growing excessively. For example, in the article *Beyond the Tsunami*, Southan and Cameron provide a sobering glimpse into the life sciences, where whole genomes and other biological data are being submitted to international databases. According to the authors, over 160 million genetic base sequences, or a staggering 250 billion nucleotides, have been submitted to international databases for storage as of 2009. Additionally, advances in technology are allowing the rate of data capture to continue accelerating. In *A 2020 Vision for Ocean Science*, Dalaney and Barga highlight how ocean scientists have been able to unify robotics and high-speed communication to develop enormous sensor arrays within the ocean that operate 24/7. With thousands of sensors, this giant ocean laboratory is

expected to generate petabytes of data that will need to somehow be accessed and visualized in real time.

In *Gray's Laws: Database-Centric Computing in Science*, Szalay and Blakeley state that one of the most fundamental challenges of eScience is simply being able to store, access, and process all of this information. Although some projects have been successful thus far using current computer and database technologies, they have required project-specific customization and have not easily scaled to larger datasets. As an alternative to these existing technologies, the authors propose that data-intensive science should adopt nodal computer architectures that are composed of many small "cyberbricks." In an effort to overcome the performance limitations of classical systems, each cyberbrick must be a self-contained processing unit with its own dedicated storage and networking elements. On a more fundamental level, Largus and Gannon contend in *Multicore Computing and Scientific Discovery* that a viable solution may also lie with the use of parallel programming, a technique that allows time-consuming processes to be easily tackled in a divide-and-conquer fashion.

While the storage and retrieval of data is important, the most significant aspect of data-intensive research from a scientist's perspective is the ability to reduce data into something that is both manageable and accurate. In fact, it is often the case that the relatively small datasets of today can already be too large to fully comprehend, and approximations must be applied to reach a dependable conclusion. Large datasets can only benefit science if new methods are developed to digest the massive amounts of information. One intriguing solution that was discussed was the use of artificial intelligence (AI). In *A Unified Modeling Approach to Data-Intensive Healthcare*, Buchan et al. portray a scenario where an epidemiologist uses AI to battle children's asthma. The example clearly illustrates how computerized methods could streamline and accelerate important medical discoveries. Moreover, it intimately describes the relationship between researchers and computers within the fourth paradigm.

Although AI seems very promising, a skeptical scientist may note that its success relies on the development of standards and protocols for recording medical and scientific databases. In fact, this lack of standardization has become a burden to researchers, often hindering rather than helping the flow of scientific discovery. In *The Impact of Workflow Tools on Data-Centric Research*, Goble and de Roure suggest that a solution lies in the development of scientific workflow tools. These software suites are envisioned to ease the duties of research by automating routine tasks such as managing data capture, validating data, and performing data-mining techniques. The authors provide a number of example workflow systems that are currently being used, such as Microsoft's Trident workflow package operated by the Pan-STARRS astronomical survey that

validates 30 terabytes of telescope data each year. In a world of competing standards and formats, these tools may become essential for the application of computerized analytical techniques.

Standard data practices are also suggested to improve researcher collaboration. In *Healthcare Delivery in Developing Countries*, Robertson et al. provide a striking example with the application of NxKM technology in developing countries. This state-of-the-art platform involves a knowledge base developed by health experts, a medical diagnostic engine, and a cell-phone interface that is designed to administer multi-choice questions regarding patient information, symptoms, and location. Diagnostic information is collected from local village workers, but analyzed by teams of professionals remotely. Through collaborative efforts, more advanced treatment options can be explored thus advancing both science and humanity.

Although *The Fourth Paradigm* continues on to describe elements within the emerging field of eScience, the examples provided in this review clearly demonstrate that researchers are tackling important questions regarding the management, analysis, and sharing of big data. However, as data-intensive science is just beginning to discover its role in many disciplines, it is important to note that many of the examples presented in *The Fourth Paradigm* are largely speculative. While the ideas certainly provide a glimpse into an exciting future for science, they tend to overlook the excitement felt by many scientists of today. More specifically, the essays presented fail to describe of how computerized methods will affect the *thrill* of human scientific discovery. The curiosity invoked by science's mysteries and the subsequent thirst for understanding seems to become obsolete qualities of data-intensive scientific research. Ultimately, when computers begin to surpass human ability to comprehend science and when intense computational methods are necessary to analyze data, it becomes disheartening for generations of scientists who find beauty in understanding nature's complexities using solely the power of the human mind. Nonetheless, *The Fourth Paradigm* can certainly be considered to provide a sobering and vivid vision of how science must cope with the emergence of big data.

Reviewer

Clinton J Regan is a PhD student in the Department of Chemistry and Chemical Engineering at the California Institute of Technology, where he is currently developing new biophysical techniques to understand protein structure and function. He is specifically targeting neuronal channels found in the brain that have been implicated in a wide array of neurological disorders, such as Parkinson's disease, Alzheimer's disease, and schizophrenia.