BY PAUL McFEDRIES

The Coming Data Deluge

The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration.

—Jim Gray, computer scientist

ACCORDING TO the late Jim Gray (he was lost at sea in 2007), until recently science was largely the product of three interrelated paradigms: experimental, theoretical, and computational. However, the computational paradigm is now generating so much data that a fourth is emerging, one that requires new tools and techniques to store, organize, filter, share, and analyze these massive amounts of data. Gray called this new paradigm **eScience** and characterized it as "IT meets scientists." Whether you're a scientist or a technologist, this new **data-intensive science** is fascinating stuff, and for the neologist, this new field is generating a flood of new words.

In the past, most scientific disciplines could be described as **small data**, or even **data poor**. Most experiments or studies had to contend with just a few hundred or a few thousand data points. Now, thanks to massively complex new instruments and simulators, many disciplines are generating correspondingly massive data sets that are described as **big data**, or **data rich**. Consider the Large Hadron Collider, which will eventually generate about 15 *petabytes* of data per year. A petabyte is about a million gigabytes, so that qualifies as a full-fledged **data deluge**.

And if you thought the complete human genome involved a lot of data, neuroscientists have set themselves the goal of creating a **connectome**, a complete map of the brain's neural circuitry. The problem? According to researchers Michael F. Cohen and Jeff W. Lichtman, an image of a cubic millimeter chunk of the brain would comprise about 1 petabyte of data (at a 5-nanometer resolution). There are about a *million* cubic millimeters of neural matter to map, making a total of about a thousand *exabytes* (an exabyte is about a thousand petabytes), so it qualifies as what Jim Gray once called an **exaflood** of data.

These massive data sets require **massive computation**, and so workaday scientists will have to become **data scientists** who use the latest software and database tools for *data mining*, which is the extraction of patterns and knowledge from large and complex data sets.

Perhaps the biggest data set of all is the collection of actions, choices, and preferences that each person performs throughout the day, which is called his or her **data exhaust**. Using such data for scientific purposes is called **citizen science**. This is **noisy data** in that most of it is irrelevant or even misleading, but there are ways to cull signal.

A good example is Google Flu Trends. In the past, epidemiologists would look for emerging flu outbreaks by laboriously examining physician logs, flu and cold medicine purchases, and other real-world sources. This **syndromic surveillance** has traditionally been too time consuming to be of much use. By contrast, Google Flu Trends examines search queries for flu-related terms, which enables researchers to tease out flu patterns in real time.

With Earth increasingly blanketed by sensors and other devices that provide raw data on the state of the planet, many scientists now envision combining these devices into a distributed, Earth-scale **macroscope** that they can use to "zoom in" on any one of myriad global states.

As all this **eResearch** becomes more sophisticated and more valuable, data scientists are realizing that these humongous data sets need to be shared among multiple scientists, labs, and institutions. We're starting to do a good job of making papers and other research end products more widely available, but what's needed are more **digital data libraries** that store not only documents such as research papers but also the data on which those papers were based. Now all we need is for someone to come up with a Digital Dewey Decimal System to catalog all this data. A Dewey Binary System, perhaps? ❑

GLUEKIT