The New York Times

December 15, 2009

BOOKS ON SCIENCE

A Deluge of Data Shapes a New Era in Computing

By JOHN MARKOFF

THE FOURTH PARADIGM

Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley and Kristin Tolle. Microsoft Research. 252 pages.

In a speech given just a few weeks before he was lost at sea off the California coast in January 2007, Jim Gray, a database software pioneer and a Microsoft researcher, sketched out an argument that computing was fundamentally transforming the practice of science.

Dr. Gray called the shift a "fourth paradigm." The first three paradigms were experimental, theoretical and, more recently, computational science. He explained this paradigm as an evolving era in which an "exaflood" of observational data was threatening to overwhelm scientists. The only way to cope with it, he argued, was a new generation of scientific computing tools to manage, visualize and analyze the data flood.

In essence, computational power created computational science, which produced the overwhelming flow of data, which now requires a computing change. It is a positive feedback loop in which the data stream becomes the data flood and sculptures a new computing landscape.

In computing circles, Dr. Gray's crusade was described as, "It's the data, stupid." It was a point of view that caused him to break ranks with the supercomputing nobility, who for decades focused on building machines that calculated at picosecond intervals.

He argued that government should instead focus on supporting cheaper clusters of computers to manage and process all this data. This is distributed computing, in which a nation full of personal computers can crunch the pools of data involved in the search for extraterrestrial intelligence, or protein folding.

The goal, Dr. Gray insisted, was not to have the biggest, fastest single computer, but rather "to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other." He was instrumental in making this a reality, particularly for astronomy, for which he helped build vast databases that wove much of the world's data into interconnected repositories that have created, in effect, a worldwide telescope.

Now, as a testimony to his passion and vision, colleagues at Microsoft Research, the company's laboratory that is focused on science and computer science, have published a tribute to Dr. Gray's perspective in "The Fourth Paradigm: Data-Intensive Scientific Discovery." It is a collection of essays written by Microsoft's scientists and outside scientists, some of whose research is being financed by the software publisher.

The essays focus on research on the earth and environment, health and well-being, scientific infrastructure and the way in which computers and networks are transforming scholarly communication. The essays also chronicle a new generation of scientific instruments that are increasingly part sensor, part computer, and which are capable of producing and capturing vast floods of data. For example, the Australian Square Kilometre Array of radio telescopes, CERN's Large Hadron Collider and the Pan-Starrs array of telescopes are each capable of generating several petabytes of digital information each day, although their research plans call for the generation of much smaller amounts of data, for financial and technical reasons. (A petabyte of data is roughly equivalent to 799 million copies of the novel "Moby Dick.")

"The advent of inexpensive high-bandwidth sensors is transforming every field from data-poor to data-rich," Edward Lazowska, a computer scientist and director of the University of Washington eScience Institute, said in an e-mail message. The resulting transformation is occurring in the social sciences, too.

"As recently as five years ago," Dr. Lazowska said, "if you were a social scientist interested in how social groups form, evolve and dissipate, you would hire 30 college freshmen for $10 an hour and interview them in a focus group."

"Today," he added, "you have real-time access to the social structuring and restructuring of 100 million Facebook users."

The shift is giving rise to a computer science perspective, referred to as "computational thinking" by Jeannette M. Wing, assistant director of the Computer and Information Science and Engineering Directorate at the National Science Foundation.

Dr. Wing has argued that ideas like recursion, parallelism and abstraction taken from computer science will redefine modern science. Implicit in the idea of a fourth paradigm is the ability, and the need, to share data. In sciences like physics and astronomy, the instruments are so expensive that data must be shared. Now the data explosion and the falling cost of computing and communications are creating pressure to share all scientific data.

"To explain the trends that you are seeing, you can't just work on your own patch," said Daron Green, director of external research for Microsoft Research. "I've got to do things I've never done before: I've got to share my data."

That resonates well with the emerging computing trend known as "the cloud," an approach being driven by Microsoft, Google and other companies that believe that, fueled by the Internet, the shift is toward centralization of computing facilities.

Both Microsoft and Google are hoping to entice scientists by offering cloud services tailored for scientific experimentation. Examples include Worldwide Telescope from Microsoft and Google Sky, intended to make a range of astronomical data available to all.

Similar digital instruments are emerging in other fields. In one chapter, "Toward a Computational Microscope for Neurobiology," Eric Horvitz, an artificial intelligence researcher for Microsoft, and William Kristan, a neurobiologist at the University of California, San Diego, chart the development of a tool they say is intended to help understand the communications among neurons.

"We have access to too much data now to understand what's going on," Dr. Horvitz said. "My goal now is to develop a new kind of telescope or microscope."

By imaging the ganglia of leeches being studied in Dr. Kristan's laboratory, the researchers have been able to identify "decision" cells, responsible for summing up a variety of inputs and making an action, like crawling. Someday, Dr. Horvitz hopes to develop the tool into a three-dimensional display that makes it possible to overlay a set of inferences about brain behavior that can be dynamically tested.

The promise of the shift described in the fourth paradigm is a blossoming of science. Tony Hey, a veteran British computer scientist now at Microsoft, said it could solve a common problem of poor use of graduate students. "In the U.K.," Dr. Hey said, "I saw many generations of graduates students really sacrificed to doing the low-level IT."

The way science is done is changing, but is it a shift of the magnitude that Thomas Kuhn outlined in "The Structure of Scientific Revolutions"?

In his chapter, "I Have Seen the Paradigm Shift, and It Is Us," John Wilbanks, the director of Science Commons, a nonprofit organization promoting the sharing of scientific information, argues for a more nuanced view of data explosion.

"Data is not sweeping away the old reality," he writes. "Data is simply placing a set of burdens on the methods and the social habits we use to deal with and communicate our empiricism and our theory."