SCIENCE PRACTICE

Sailing on an Ocean of 0s and 1s

James P. Collins

When the development of theory outpaces data, scientists often find that new ideas cannot be tested for lack of tools or technology. Researchers in genomics, astronomy, and many other active areas of science face a different challenge: Gathering data is so easy and quick that it exceeds our capacity to validate, analyze, visualize, store, and curate the information. *The Fourth Paradigm* addresses this challenge—and the opportunity it presents.

The Fourth Paradigm
Data-Intensive Scientific Discovery

Tony Hey, Stewart Tansley, and Kristin Tolle, Eds.

Microsoft Research, Redmond, WA, 2009. 286 pp. Paper, $46. ISBN 9780982544204. PDF at http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Tony Hey, Stewart Tansley, and Kristin Tolle (computer scientists at Microsoft Research) have grouped the essays into four sections: Earth and Environment, Health and Wellbeing, Scientific Infrastructure, and Scholarly Communication. Examples from health care and the environment dominate but are leavened with some cases from the physical and social sciences. The editors' thesis is that although empirical, analytical, and simulation methods have provided answers to many questions, a new scientific methodology driven by data-intensive problems is now emerging—the "fourth paradigm."

Some contributors discuss software and hardware advances that would help scientists cope with the data deluge. Most consider how research practices will be transformed by computational thinking—scientific instruments powered by computers in ways that make them a universal intellectual amplifier. Hardware and, especially, software will facilitate "cross-domain ventures that accelerate discovery, highlight new connections, and suggest unforeseen links that will speed science forward." For example, while analyzing the relationship between variables x and y, background software will seamlessly link to diverse databases. A prompt may then suggest looking at how variable z relates to x, to y, or to x and y. A word of caution: in this work flow, scientists may fall into the trap of understanding fewer details of an analysis if they accept uncritically such prompts as offered.

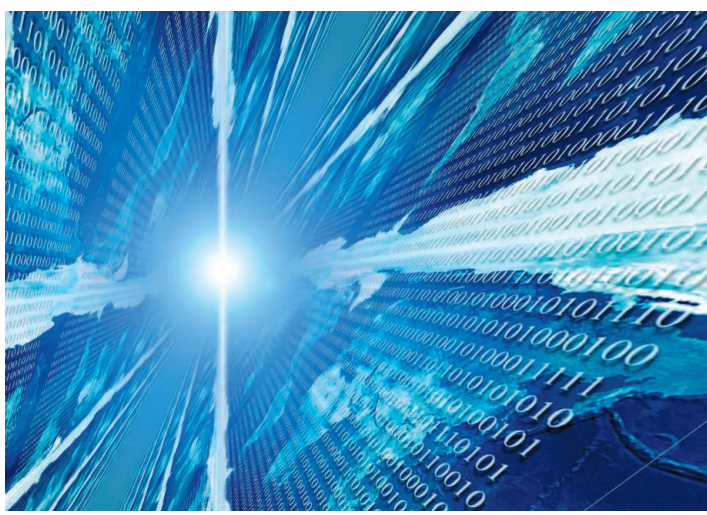
Backing away from a paradigm shift in the sense of Thomas Kuhn, John Wilbanks characterizes cyberinfrastructure advances as new tools cutting across the sciences. It will be up to historians and philosophers of science to wrestle with the questions: In what if any sense will computational thinking transform the process of scientific discovery? Will data-intensive science cause an incremental change in how science is done? Or will the changes be a true step function?

Many of the authors see a big new step in networking: researcher to researcher and also lab notebooks to archived databases and published results. A network of investigators is, of course, not novel—think of Darwin and his correspondents. But until recently, even scientists in a network usually conducted experiments and wrote papers alone or with only a few others. Now dozens to hundreds of researchers may participate in projects.

At its best, this environment will facilitate what we might call open source innovation, in which advances have a sociotechnical component. Individual labs will routinely reach beyond their walls, and organizations will strive to create and sustain collaborative, distributed networks of investigators. But hurdles will appear. The toughest changes ahead involve training, institutional organization, and the social context of research. A radical restructuring of the culture of scientific work will be needed to integrate biological, physical, and social sciences and engineering; move across the science-technology interface; foster systems thinking; support flexible and interdisciplinary approaches to problem-solving; integrate knowledge creation and knowledge use; and balance individual with group achievement. Fostering innovation will need to become a core institutional value.

The editors' collective enthusiasm about data-intensive science leads to an occasional odd statement, such as "Science is becoming increasingly dependent on data." More precisely, solving some modern science problems requires very large data sets. A justifiable excitement about large reference data collections has caused some to talk about the end of theory because extensive databases will support "hypothesis-neutral research." Paul Ginsparg counters this argument nicely in his chapter on a data-centric world: "Science aims to produce far more than a simple mechanical prediction of correlations; instead, its goal is to employ those regularities extracted from data to construct a unified means of understanding them *a priori*." Data mining simply to predict trends confuses the goals of phenomenological modeling and theory development.

Although most authors foresee a bright future, even the visionaries admit we are some distance from computer systems that seamlessly link large numbers of related but disjoint information sources. *The Fourth Paradigm* also offers a vision including few caveats associated with ethics, privacy, or cybersecurity. No lessons are drawn from Aldous Huxley's *Brave New World* or George Orwell's *1984*. Breathtaking advances in the sciences must be placed in a larger societal context by drawing on the law, humanities, and arts.



The text has a few rough patches. Some are the inevitable consequence of melding 71 contributors and 36 chapters. Others are interesting, as the unevenness conveys a sense of excitement and creation. Even the term "fourth paradigm" is defined in several ways. The authors provide a view from the leading edge, and the front is often ragged, incomplete, and in construction. We get a bottom-up view of change in progress. Many will enjoy this book: historians of science interested in the dynamics of change, philosophers wrestling with the nature of science, social scientists studying how disciplines evolve, and natural and physical scientists who want fresh ideas.

The reviewer is at the School of Life Sciences, Arizona State University, Post Office Box 874501, Tempe, AZ 85287–4501, USA. E-mail: jcollins@asu.edu

Published by AAAS

The Fourth Paradigm is dedicated to and reflects the vision of the late Jim Gray of Microsoft Research, who envisioned "a world of scholarly resources—text, databases, and any other associated materials—that were seamlessly navigable and interoperable." Gray loved sailing. Sailors, of course, guide a vessel by reacting to the nearest swell and wave. But the ocean also affords a chance to scan the horizon in anticipation of the future, to see what's ahead and imagine what's just out of view. The individual essays—and *The Fourth Paradigm* as a whole—give readers a glimpse of the horizon for 21st-century research and, at their best, a peek at what lies beyond. It's a journey well worth taking.

10.1126/science.1186123

The memory of the Internet. Collectively, server farms house hundreds of exabytes of information.

CREDIT: BOB SACHA/CORBIS

COMPUTERS AND SOCIETY

Programming to Forget

William Dutton

The growing centrality of the Internet is leading to initiatives to archive, curate, and otherwise preserve digital collections. While experts and resources are focused on these problems of remembering, Viktor Mayer-Schönberger tells us that there is also a virtue in forgetting.

Delete
The Virtue of Forgetting in the Digital Age

by Viktor Mayer-Schönberger

Princeton University Press, Princeton, NJ, 2009. 251 pp. $24.95, £16.95. ISBN 9780691138619.

Delete begins with an anecdote about a student, Stacy, who is denied her teaching certificate because a colleague discovered an old photo of Stacy wearing a pirate hat and drinking alcohol—one she had posted on a social networking Web site. As the author put it: "The Internet remembered what Stacy wanted to have forgotten." Similar stories of individuals compromised by information stored on the Internet or related devices are numerous. Mayer-Schönberger (a legal scholar at the National University of Singapore) diagnoses the problem, explains its growing importance, and answers the question "What can be done?"

The book's central argument is that in the analog world of yesterday, forgetting was the default position. It was somewhat harder to remember than to forget, so unless we put effort into it, such as in taking notes or storing text, information disappeared. "Not anymore." In tomorrow's digital world, the default will be remembering. The efficiency of remembering is gaining ground because of the lower costs of memory devices and the accuracy of digital technologies, which can replicate content endlessly, creating the potential for a future of "perfect remembering."

Taken to its logical conclusion, this capability could create a dystopian scenario of self-censorship that moves beyond contemporary conceptions of a surveillance society. Building on Bentham's notion of the panopticon, digital memory is extending the "mechanism of panoptic control" into the past. However, Mayer-Schönberger argues, this problem can be addressed through a variety of legal and technical initiatives, such as creating a means for users to place an expiration date on information they post.

The book offers a provocative counter to prevailing neologisms about information wanting to be shared. Our circumstances are far more complicated, with many not wanting all information to be remembered. In developing a clear line of reasoning behind his argument, Mayer-Schönberger draws evidence from multiple disciplines, bringing together considerations from the neurosciences, computation, and networking technology as well as from law, policy, and literature. It is rare but wonderful for an expert on digital technology to glean from works of major literary figures with the same ease as he discusses shared memory devices and Vannevar Bush's "memex." His book also stands out in being truly international, anchored in European and Asian examples as firmly as North American legal and policy cases.

Mayer-Schönberger's focus on a single issue—remembering and forgetting as enabled by digital technology—enables him to address some familiar subjects, such as the history of the communications revolution, in a fresh and engaging way. Moreover, his style is accessible and clearly targeted beyond his academic peers to reach an audience engaged by the issue rather than the technology or the law. That said, readers will learn about technology, law, and other fields as his narrative unfolds.

Most important, Mayer-Schönberger's focus illustrates a major turn in debates about the Internet. Since the dot-com bubble of the late 1990s, most research on the Internet has dealt primarily with its use and impact in the broadest sense. Will the technology become a routine aspect of everyday life and work? This book takes the Internet's role in society as a given and concentrates on the critical design features that make it easier for machines to remember than to forget. Refocusing on key design issues, as the author does, will enable social and policy research to contribute more to shaping the future Internet.

The author's diagnosis of the problem raises questions. Hasn't memory always been long-term for some people, such as those forced out of a community that will not forget a major transgression? In such cases, the Internet is transforming the geography of memory more than extending its longevity. Also, if individuals can delete their past, will we face Orwellian issues over the rewriting of history?

Mayer-Schönberger's discussion of potential remedies is less convincing than his exposition of the problem. He admits as much in suggesting that his solutions are less than perfect. For example, expiration dates will be difficult to realize, given the distributed nature of the Web, and might cause other problems. Since the book's publication, new applications have been released that enable a text message, for example, to vanish after being read or on a specified date. However, employing software to erase messages can create unwarranted suspicion, or we might lose information we later want to retrieve.

Even if I am not completely convinced of the problem or solutions, *Delete* is well placed to accomplish the author's aim: "to commence a wide-ranging, open, and intense discussion about forgetting, and how we can ensure that we'll remember its importance in our digital future." There is no better source for fostering an informed debate on this issue.

10.1126/science.1187723

The reviewer is at the Oxford Internet Institute, University of Oxford, 1 St. Giles', Oxford OX1 3JS, UK. E-mail: william.dutton@oii.ox.ac.uk