



A New Path for Science?

MARK R. ABBOTT
Oregon State University

THE SCIENTIFIC CHALLENGES of the 21st century will strain the partnerships between government, industry, and academia that have developed and matured over the last century or so. For example, in the United States, beginning with the establishment of the National Science Foundation in 1950, the nation's research university system has blossomed and now dominates the basic research segment. (The applied research segment, which is far larger, is primarily funded and implemented within the private sector.)

One cannot overstate the successes of this system, but it has come to be largely organized around individual science disciplines and rewards individual scientists' efforts through publications and the promotion and tenure process. Moreover, the eternal "restlessness" of the system means that researchers are constantly seeking new ideas and new funding [1, 2]. An unexpected outcome of this system is the growing disconnect between the supply of scientific knowledge and the demand for that knowledge from the private and government sectors [3, 4]. The internal reward structure at universities, as well as the peer review system, favors research projects that are of inherent interest to the scientific community but not necessarily to those outside the academic community.



NEW DRIVERS

It is time to reexamine the basic structures underlying our research enterprise. For example, given the emerging and urgent need for new approaches to climate and energy research in the broad context of sustainability, fundamental research on the global climate system will continue to be necessary, but businesses and policymakers are asking questions that are far more interdisciplinary than in the past. This new approach is more akin to scenario development in support of risk assessment and management than traditional problem solving and the pursuit of knowledge for its own sake.

In climate science, the demand side is focused on feedback between climate change and socioeconomic processes, rare (but high-impact) events, and the development of adaptive policies and management protocols. The science supply side favors studies of the physical and biological aspects of the climate system on a continental or global scale and reducing uncertainties (e.g., [5]). This misalignment between supply and demand hampers society's ability to respond effectively and in a timely manner to the changing climate.

RECENT HISTORY

The information technology (IT) infrastructure of 25 years ago was well suited to the science culture of that era. Data volumes were relatively small, and therefore each data element was precious. IT systems were relatively expensive and were accessible only to experts. The fundamental workflow relied on a data collection system (e.g., a laboratory or a field sensor), transfer into a data storage system, data processing and analysis, visualization, and publication.

Figure 1 shows the architecture of NASA's Earth Observing System Data and Information System (EOSDIS) from the late 1980s. Although many thought that EOSDIS was too ambitious (it planned for 1 terabyte per day of data), the primary argument against it was that it was too centralized for a system that needed to be science driven. EOSDIS was perceived to be a data factory, operating under a set of rigorous requirements with little opportunity for knowledge or technology infusion. Ultimately, the argument was not about centralized versus decentralized but rather who would control the requirements: the science community or the NASA contractor. The underlying architecture, with its well-defined (and relatively modest-sized) data flows and mix of centralized and distributed components, has remained undisturbed, even as the World Wide Web, the Internet, and the volume of online data have grown exponentially.

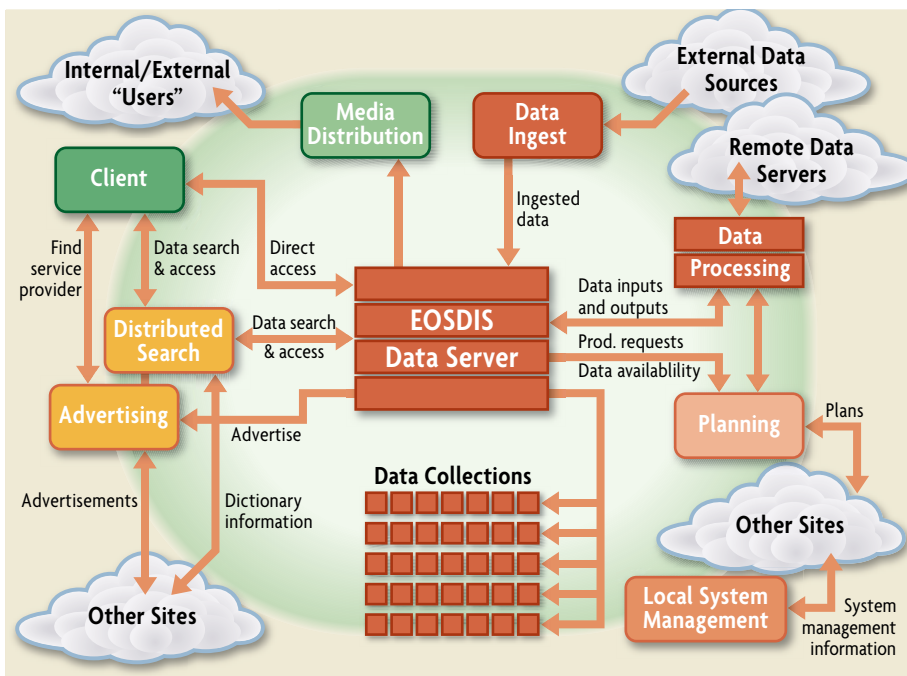


FIGURE 1. NASA's Earth Observing System Data and Information System (EOSDIS) as planned in 1989.

THE PRESENT DAY

Today, the suite of national supercomputer centers as well as the notion of “cloud computing” looks much the same as the architecture shown in Figure 1. It doesn't matter whether the network connection is an RS-232 asynchronous connection, a dial-up modem, or a gigabit network, or whether the device on the scientist's desktop is a VT100 graphics terminal or a high-end multicore workstation. Virtualized (but distributed) repositories of data storage and computing capabilities are accessed via network by relatively low-capability devices.

Moore's Law has had 25 years to play out since the design of EOSDIS. Although we generally focus on the increases in capacity and the precipitous decline in the price/performance ratio, the pace of rapid technological innovation has placed enormous pressure on the traditional modes of scientific research. The vast amounts of data have greatly reduced the value of an individual data element, and we are no



longer data-limited but insight-limited. “Data-intensive” should not refer just to the centralized repositories but also to the far greater volumes of data that are network-accessible in offices, labs, and homes and by sensors and portable devices. Thus, data-intensive computing should be considered more than just the ability to store and move larger amounts of data. The complexity of these new datasets as well as the increasing diversity of the data flows is rendering the traditional compute/data-center model obsolete for modern scientific research.

IMPLICATIONS FOR SCIENCE

IT has affected the science community in two ways. First, it has led to the *commoditization* of generic storage and computing. For science tasks that can be accomplished through commodity services, such services are a reasonable option. It will always be more cost effective to use low-profit-margin, high-volume services through centralized mechanisms such as cloud computing. Thus more universities are relying on such services for data backup, e-mail, office productivity applications, and so on.

The second way that IT has affected the science community is through radical *personalization*. With personal access to teraflops of computing and terabytes of storage, scientists can create their own compute clouds. Innovation and new science services will come from the edges of the networks, not the commodity-driven datacenters. Moreover, not just scientists but the vastly larger number of sensors and laboratory instruments will soon be connected to the Internet with their own local computation and storage services. The challenge is to harness the power of this new network of massively distributed knowledge services.

Today, scientific discovery is not accomplished solely through the well-defined, rigorous process of hypothesis testing. The vast volumes of data, the complex and hard-to-discover relationships, the intense and shifting types of collaboration between disciplines, and new types of near-real-time publishing are adding pattern and rule discovery to the scientific method [6]. Especially in the area of climate science and policy, we could see a convergence of this new type of data-intensive research and the new generation of IT capabilities.

The alignment of science supply and demand in the context of continuing scientific uncertainty will depend on seeking out new relationships, overcoming language and cultural barriers to enable collaboration, and merging models and data to evaluate scenarios. This process has far more in common with network gaming than with the traditional scientific method. Capturing the important elements of



data preservation, collaboration, provenance, and accountability will require new approaches in the highly distributed, data-intensive research community.

Instead of well-defined data networks and factories coupled with an individually based publishing system that relies on peer review and tenure, this new research enterprise will be more unruly and less predictable, resembling an ecosystem in its approach to knowledge discovery. That is, it will include loose networks of potential services, rapid innovation at the edges, and a much closer partnership between those who create knowledge and those who use it. As with every ecosystem, emergent (and sometimes unpredictable) behavior will be a dominant feature.

Our existing institutions—including federal agencies and research universities—will be challenged by these new structures. Access to data and computation as well as new collaborators will not require the physical structure of a university or millions of dollars in federal grants. Moreover, the rigors of tenure and its strong emphasis on individual achievement in a single scientific discipline may work against these new approaches. We need an organization that integrates natural science with socioeconomic science, balances science with technology, focuses on systems thinking, supports flexible and interdisciplinary approaches to long-term problem solving, integrates knowledge creation and knowledge use, and balances individual and group achievement.

Such a new organization could pioneer integrated approaches to a sustainable future, approaches that are aimed at understanding the variety of possible futures. It would focus on global-scale processes that are manifested on a regional scale with pronounced socioeconomic consequences. Rather than a traditional academic organization with its relatively static set of tenure-track professors, a new organization could take more risks, build and develop new partnerships, and bring in people with the talent needed for particular tasks. Much like in the U.S. television series *Mission Impossible*, we will bring together people from around the world to address specific problems—in this case, climate change issues.

MAKING IT HAPPEN

How can today's IT enable this type of new organization and this new type of science? In the EOSDIS era, it was thought that relational databases would provide the essential services needed to manage the vast volumes of data coming from the EOS satellites. Although database technology provided the baseline services needed for the standard EOS data products, it did not capture the innovation at the edges of the system where science was in control. Today, semantic webs and ontologies are



being proposed as a means to enable knowledge discovery and collaboration. However, as with databases, it is likely that the science community will be reluctant to use these inherently complex tools except for the most mundane tasks.

Ultimately, digital technology can provide only relatively sparse descriptions of the richness and complexity of the real world. Moreover, seeking the unusual and unexpected requires creativity and insight—processes that are difficult to represent in a rigid digital framework. On the other hand, simply relying on PageRank¹-like statistical correlations based on usage will not necessarily lead to detection of the rare and the unexpected. However, new IT tools for the data-intensive world can provide the ability to “filter” these data volumes down to a manageable level as well as provide visualization and presentation services to make it easier to gain creative insights and build collaborations.

The architecture for data-intensive computing should be based on storage, computing, and presentation services at every node of an interconnected network. Providing standard, extensible frameworks that accommodate innovation at the network edges should enable these knowledge “ecosystems” to form and evolve as the needs of climate science and policy change.

REFERENCES

- [1] D. S. Greenberg, *Science, Money, and Politics: Political Triumph and Ethical Erosion*. Chicago: University of Chicago Press, 2001.
- [2] National Research Council, *Assessing the Impacts of Changes in the Information Technology R&D Ecosystem: Retaining Leadership in an Increasingly Global Environment*. Washington, D.C.: National Academies Press, 2009.
- [3] D. Sarewitz and R. A. Pielke, Jr., “The neglected heart of science policy: reconciling supply of and demand for science,” *Environ. Sci. Policy*, vol. 10, pp. 5–16, 2007, doi: 10.1016/j.envsci.2006.10.001.
- [4] L. Dilling, “Towards science in support of decision making: characterizing the supply of carbon cycle science,” *Environ. Sci. Policy*, vol. 10, pp. 48–61, 2007, doi: 10.1016/j.envsci.2006.10.008.
- [5] Intergovernmental Panel on Climate Change, *Climate Change 2007: The Physical Science Basis*. New York: Cambridge University Press, 2007.
- [6] C. Anderson, “The End of Theory,” *Wired*, vol. 16, no. 7, pp. 108–109, 2008.

¹ The algorithm at the heart of Google’s search engine.