The Fourth Paradigm: Data-Intensive Scientific Discovery

By KRISTIN M. TOLLE

D. STEWART W. TANSLEY

ANTHONY J. G. HEY

External Research, Microsoft Research, Redmond, VA 98052 USA

The book *The Fourth Paradigm: Data Intensive Scientific Discovery*,[1] contains a series of essays by scientists and computer scientists looking forward five years or more to how different scientific fields are being transformed by the exponential increase in scientific data. Along with this "data revolution" we are also in the midst of a revolution in scholarly communications. As one example, the number of scientific abstracts being deposited in the Medline database corresponds to approximately 1100 papers every day; over 400 000 per year.[2] As Faniel and Zimmerman point out, in many fields, scientists have numerous challenges gaining access to the original data to either check the claims of a scientific paper or to combine that data with other data for further analysis.[3] Smit goes on to suggest that data and the publications should be "wedded."[4] We are now seeing governments and funding agencies looking at ways to increase the value and pace of scientific research through increased or open access to both data and publications. In this point of view article, we wish to look at another aspect of these twin revolutions, namely, how to enable developers, designers and researchers to build intuitive, multimodal, user-centric, scientific applications that can aid and enable scientific research—essentially a

[1]*The Fourth Paradigm: Data-Intensive Scientific Discovery*, T. Hey, S. Tansley, and K. Tolle, Eds., Redmond, VA: Microsoft Research, 2009, ISBN 978-0-9825442-0-4, http://fourthparadigm.org.

[2]http://library.dialog.com/bluesheets/html/bl0154.html.

[3]http://scientificdatasharing.com/wp-content/uploads/2011/01/Faniel_Zimmerman.pdf. Accessed Jan. 25, 2011.

[4]http://www.dlib.org/dlib/january11/smit/01smit.html. Accessed Jan. 25, 2011.

Digital Object Identifier: 10.1109/JPROC.2011.2155130

natural user interface (NUI) to facilitate scientific discovery.

The term NUI refers to finding the best possible way or ways to interact with computing devices. Although existing modes of interaction will not generally be displaced by new gesture and touch technologies, in the future there will be an augmentation and combination of interaction models that optimizes the user's experience for whatever task/purpose and situation/context is relevant.

In order to sift through and filter the data to arrive at a relevant set of information, a scientific NUI system must begin to understand a scientists' *intent* in order to truly enable scientific discovery. Furthermore, the system must be capable of presenting that information in intuitive and personalized views, but also in a "lossless" way that allows for interactive querying of the results.

To capture true "relevance" for an individual scientist we believe one must go beyond current semantic classification systems, which are likely to continue to be too rigid and have difficulty dealing with rapid changes in the nomenclature and terminology ambiguities both between and within disciplines. The problem is also one of scope and scale. Just as one human cannot effectively know everything, a single system, no matter how many compute cores and how much storage is available can ever semantically capture all that can be discovered in today's vast expanse of data—which will one day seem small by the standards of future generations of scientists and computer scientists. Just as we specialize, even within our specializations, so must the computer technology that supports our scientific discoveries. However, specialization at this sublevel is still problematic given the difficulties alluded to above.

A future world of natural scientific interactions will blend existing interaction modes with newer experiences. This will allow for an augmentation and combination of interaction models that optimize the user's experience for a specific task or purpose within a specific situation or context. Determining what combination of interactions are most effective in any specific context and how these can be used to enable personalization and understanding of user intent is a very active area of research for the computer science community.[5]

We have so far described our vision in generalities thus far. We would like to close by describing a scenario where a future scientist will be facilitated in dealing with information overload by relying on NUI.

I. ACCELERATING SCIENCE SCENARIO

Imagine you are a scientist working on several research projects in a specific domain such as liver immunology. You have just arrived at your work place. As you enter the building your thumb resting on a slim cellphone/banking/transportation ID card works biometrically to ensure that it is you holding the card and gives you immediate access to the building, your laboratory, and your office. As you walk into your office you say "log on." Instantly several of your walls become heads up displays that pick up right where you left off the previous day along with today's scheduled meetings. You can see which lab experiments are in progress, get status reports on ones that have completed, or see which are next in the queue to get started.

In another area, you see the changes to one of the papers you are working on with location disparate collaborators. You see any relevant literature that may have surfaced since you last logged in or has become relevant as a result of new laboratory findings or additions to the paper by collaborators. Grant information is also displayed: new ones posted in relevance order, ones you are working on ordered by priority and deadline, and submitted grants with countdown to announcement dates.

As you glance at your schedule you can select the most important task to work on next. You begin working on reviewing lab results engaging with your computer almost as if it were a colleague as it brings up relevant tools and workflows. After you kick off the next set of experiments to be run and give feedback or make notes on the completed experiments, you transfer the paper to a slate for a closer inspection of paper changes and settle in to make updates of your own using speech, gesture, keyboard, pen, and touch as necessary. You add your most recent lab results, insert relevant 2-D or 3-D images and figures to as is required and uplink the data behind these images and figures to the document. When finished you transfer these back to your computer which places it in the cloud with your additions and changes highlighted for your collaborators to review. The computer tells you when your next paper telepresence meeting will take place and confirms your availability. You spend a moment reviewing any new e-mails or information coming from your students to ready yourself for the afternoon lab meeting.

In the afternoon you receive a notification that someone has accessed your online journal publication. Using liver cells from a different disease in a similarly constructed experience they were able to plug in their data to your model and produce interesting results. They have added an addendum to your paper and referenced a live link to their data and the corresponding paper showing these results.

Your computer/office understands your context, priorities, even your preferred level of interactivity at different times and in different locations, responds as you expect, and sets the environment to match your mood and current work mode. The computer environment is ubiquitous and immersive. Everything you need is at your command to start a productive day in the lab.

[5]For example, see "Being Human: Human-Computer Interaction in the year 2020." http://research.microsoft.com/en-us/um/cambridge/projects/hci2020/downloads/BeingHuman_A4.pdf.

Later that evening you arrive at home. Before bed you decide to check in with the computer to get an end of day status or see if any of your collaborators have contacted you regarding the paper changes. You step into your home office and a similar representation of your lab information is available scoped to only those activities you primarily perform at home. Your personal e-mails and contacts are also present. It is the best of both worlds.

We believe that NUIs will be an important aspect of the future of scientific discovery, embracing all aspects of research from the inception of the idea through publishing. Many computer scientists and scientists are already working together to realize this vision.

II. OCEANS OF DATA

After a boating or aircraft accident at sea, the U.S. Coast Guard (USCG) historically has relied on sea current charts and wind gauges to figure out where to hunt for survivors. But thanks to data originally collected by Rutgers University oceanographers to answer scientific questions about earth–ocean–atmosphere interactions, the USCG has a new resource that promises to literally save lives. It is a powerful example that large data sets can drive myriad new and unexpected opportunities and it is an argument for funding and building robust systems to manage and store the data.

There is a revolution underway in oceanography today. Scientists around the world are augmenting the ship-based expeditionary science of the last two centuries with a distributed, observatory-based approach involving instruments, facilities, and networked interactions with other scientists. These efforts, including those sponsored by the National Science Foundation (NSF) Ocean Sciences Division, are focused on routine, long-term measurement of episodic oceanic processes on a wide range of spatial and temporal scales. Such data are crucial to resolving scientific questions related to Earth's climate, geodynamics, and marine ecosystems. However, the same sensors and systems are also yielding valuable data that are being shared and repurposed for government and commercial uses, including energy planning, defense, and even real-time life-and-death challenges such as ocean rescue.

At Rutgers University's Coastal Ocean Observation Lab, scientists have been collecting high-frequency radar data that can remotely measure ocean surface waves and currents. The data are generated from antennas located along the eastern seaboard from Massachusetts to Chesapeake Bay. The network was built bit-by-bit to answer very specific scientific questions, such as determining the precise physical river flows of the Hudson River when it empties into the Atlantic Ocean in order to track its impact on the marine food chain. However, over time the data, and this research group's willingness to share it, are also providing previously unobtainable information for an array of users.

The New Jersey Board of Public Utilities (BPU), for example, is interested in developing an offshore wind farm industry. It turns out that surface currents serve as a proxy for sea breezes that can be localized and measured with high accuracy. BPU is using these historical data to plan the placement of equipment and project the energy likely to be captured.

The Department of Homeland Security (DHS) realized that Rutgers' radar data also include the echoes of ships. Not only can the historical data allow DHS to analyze past shipping patterns and detect changes, but the technology holds the promise for what is called "over the horizon" ship detection that cannot be collected any other way. For example, DHS is looking to use the data to focus security checks on vessels that have not reported their location.

Perhaps the most dramatic sharing involves the USCG. The Rutgers' experiments sought to identify highly accurate, real-time ocean circulation patterns. "Now, instead of developing a search box that could be as big as the state, they can integrate our data and get actual currents and decrease search areas for survivors," explains Oscar Schofield, professor of Bio-Optical Oceanography at Rutgers. "That raises the probability of faster rescue and higher survival rates."

One of the group's frustrations today, unfortunately, is the lack of funding to design and support long-term preservation of data. A large fraction of the data the Rutgers team collects has to be thrown out because there is no room to store it and no support within existing research projects to better curate and manage the data. "I can get funding to put equipment into the ocean, but not to analyze that data on the back end," says Schofield.

III. RAPID DATA SHARING SPEEDING: QUEST FOR ALZHEIMER'S BIOMARKERS

The promise of speeding up vital biomedical research by better and faster sharing of data is becoming real in an innovative Alzheimer's disease research partnership between the private sector and the National Institutes of Health. Called the Alzheimer's Disease Neuroimaging Initiative (ADNI), it was launched in 2004 specifically to improve clinical trials for the dread neurological condition. One reason Alzheimer's research is so difficult is that researchers lack good biomarkers to track disease progression. In fact, the disease still can only be definitively diagnosed by brain biopsy after death.

ADNI attacks that challenge in several ways. First, it combines data from several volunteer subject groups and several diagnostic methods, including spinal fluid analysis, magnetic imaging resonance (MRI) scans, and positron emission tomography (PET). These tests are periodically performed on 800 volunteers on the spectrum from completely healthy, through mild impairment, to patients with clinically diagnosed Alzheimer's. As

some volunteers progress from healthy to mild impairment or impairment to full-blown Alzheimer's, the hope is to find biomarkers that can more faithfully track the progression of the disease.

Not only can the data from the 14 different centers involved in the initiative be combined and compared, but it is also highly significant that the data are typically made publicly available within a week of being collected. Those two factors are catalyzing the energy of neuroscientists both at those centers and around the world. Hundreds of scientists have made tens of thousands of downloads from the ADNI website, and, of several dozen papers that have so far been published using ADNI data, a significant number were authored by researchers who are not even directly funded by the project. Scientists say the rapid sharing is motivating them to analyze and publish data more quickly, and companies are incorporating the data into their clinical trials for promising treatments.

Increasingly, scientists believe that Alzheimer's disease pathology may be present 10 or even 20 years before the onset of dementia. Three studies published by non-ADNI researchers suggest they are zeroing in on biomarkers that may help identify which patients are likely to progress to Alzheimer's disease. That could be highly significant in a devastating disease where earlier detection is considered key to better treatments. ■

Acknowledgment

The authors would like to thank the attendees of the MSR Faculty Summit 2010 NUI workshop and many colleagues in Microsoft Research.