

# Cyber-enabled Discovery and Innovation

*Allyn Jackson*

The National Science Foundation (NSF) is launching a new initiative called Cyber-enabled Discovery and Innovation (CDI) which will begin at the start of the new fiscal year in October 2007. A budget of US\$52 million has been requested for the first year of the CDI, and the NSF is planning substantial additional investments over the coming five years to bring the program up to a level of US\$250 million by 2012. The NSF's Division of Mathematical Sciences (DMS) is playing a prominent role in CDI: In fiscal 2008 the DMS is slated to receive US\$5.2 million for CDI activities, which is 10 percent of the CDI total for the NSF overall.

At the time of this writing the program had not been fully formulated and a program announcement was not yet available, so exactly what kinds of research CDI will support remains to be seen. How much mathematics gets funded as CDI grows in the coming years depends in part on how researchers in the mathematical sciences interpret the initiative and what kinds of proposals they submit.

## **Emphases of CDI**

"There are some frontiers in science that are only now accessible because of advances in computation, both on the hardware and on the algorithmic side," explained DMS director Peter March. "The impetus behind Cyber-enabled Discovery and Innovation is to get at these frontiers now in a systematic way." To accomplish this, the NSF has identified three main scientific emphases. One is "Knowledge Extraction", which comprises a range of concepts and techniques for organizing, analyzing, and visualizing the massive data sets that have become ubiquitous in scientific research. A

---

*Allyn Jackson is senior writer and deputy editor of the Notices. Her email address is axj@ams.org.*

second is "Complex Interactions", which refers to systems characterized by many scales or components that interact. The interactions could be nonlinear and could produce different kinds of emergent phenomena at different scales. Examples include modeling of protein folding or the flow of information across the Internet. The third theme is "Computational Experimentation", which encompasses simulation and modeling of phenomena or processes that cannot be analyzed through any other means. Examples include modeling of weather and of cosmological phenomena such as the "big bang".

There are two other principal components of CDI that focus on the infrastructure needed to advance the scientific themes of the initiative. One is called "Virtual Environments", which refers to the distributed environments that are needed to allow researchers to collaborate across geographic boundaries and to share data and tools. The other component focuses on education and training of students and researchers to build a coterie of workers conversant with the methods and tools needed for the cyber-based research of the future.

The role of mathematics in CDI is "central", said Tony Chan, who heads the NSF's Mathematical and Physical Sciences Directorate. He pointed out that half of the directorate's CDI funds are going to the DMS. "Many of the intellectual ideas behind CDI are mathematical in nature: computational science, modeling, extracting knowledge from data, etc.," he said. "I hope that the mathematical sciences community will see this as an opportunity to expand its engagement with other disciplines and play a leadership role in this important scientific area. Breakthroughs in this area will depend on fundamental contributions from mathematical scientists." Chan also noted that CDI would be open to supporting theoretical research that has

a potential for a cyber-based application but has not yet been applied in that way.

### Examples in Mathematics

Asked what kind of research might be funded under CDI, March pointed to three quite different examples. The first is work by Gunnar Carlsson of Stanford University and his collaborators, who have used methods of algebraic topology to analyze large data sets. One of the recent projects of this group examined data from a collection of digital pictures of nature scenes. The data came from David Mumford of Brown University and his collaborators, who had developed a method for identifying high-contrast patches in the pictures. If these patches are 3 pixels by 3 pixels, with each pixel represented by a grayscale value, then the data forms a subset of 9-dimensional Euclidean space. A bit of data-massaging reduces the dimension so that the data lies in the 7-dimensional unit sphere in  $R^8$ . They then used a parametrized family of density estimators to obtain a subspace of “frequently occurring”, or highly dense, patches. What does this subspace look like topologically?

Using one value of the estimation parameter, the work by Carlsson et al. revealed that the subspace has first Betti number 1; in other words, it contains a loop. Going back to the data, they found what the loop corresponds to. Some 3 by 3 patches are half dark and half light, with a line separating the light and dark regions. The loop corresponds to the rotation through 360 degrees of the line separating light and dark. Using a smaller value of the estimation parameter, they found two secondary loops, which again could be traced back to the data: One corresponds to patches with a dark band across a light background and the other to patches with a light band across a dark background. Reasoning topologically, they concluded that the subspace of high-contrast patches is well approximated by a two-dimensional Klein bottle sitting in  $S^7$ . This work could have important applications to image-compression, which is something Carlsson said they are actively working on. “We are looking for ways to apply this more precise understanding of the density situation to actual problems in compression,” he said.

The driving principle behind this work, Carlsson noted, is that “topology should be viewed as a tool for geometric pattern recognition. . . . Holes and things like that are patterns in the geometry [of a data set] that take it away from just being a Gaussian blob.” Carlsson and collaborators are also using some of the techniques they have developed to analyze neuroscience data, specifically, data from the visual cortex of monkeys. The work on data sets is now opening up some new and interesting questions in algebraic topology. (The

work of Carlsson and his collaborators is supported by the NSF and by the Defense Advanced Research Projects Agency.)

The second example also centers on large data sets but takes a completely different approach to address a different kind of problem. In 2004 Emmanuel Candès of the California Institute of Technology, together with his postdoc Justin Romberg and with Terence Tao of the University of California at Los Angeles, published a paper establishing a result that went completely against received wisdom and sparked a revolution in signal processing. The received wisdom is contained in theorems of Harry Nyquist and Claude Shannon, which dictate the rate at which one must sample a signal in order to fully reconstruct it. But this approach does not take into account the structure that is present in most data of scientific interest. Such data is typically compressible: Assuming a suitable representation of the data, such as local cosines or wavelets, one usually finds that many coefficients in the representation are close to zero. The sparsity of such a representation permits efficient signal processing and data compression. More surprisingly, what Candès, Romberg, and Tao found is that the compressibility of the signal can be exploited to make the data acquisition process itself more efficient.

In working with doctors on magnetic resonance imaging data, Candès came up with an algorithm intended to help reduce artifacts in the data. But in tests of the algorithm, a strange thing happened: the algorithm reconstructed the image exactly every time. When an algorithm “always spits out the truth, you have to wonder what’s going on,” Candès said. This was the starting point of his paper with Romberg and Tao. For their method to work, the data must be sparse when represented in a basis using one kind of waveforms (say, wavelets). Then, by sampling using a different kind of waveform (say, the Fourier transform) that are incoherent with the initial waveforms, one can sample in an extremely parsimonious way—meaning that very few samples are needed to reconstruct the original image. Their paper used optimization theory, combinatorics, and a good deal of probability theory, particularly random matrix theory. Indeed, there is an analogy between their result and results in theoretical computer science on randomized algorithms that can be used to solve—with errors as small as one likes—problems that are thought to be in NP.

The notion that one could reconstruct a signal from such minimal sampling was a big surprise. The new field of “compressive sampling” or “compressed sensing” has attracted huge attention within pure and applied mathematics, signal processing, and computer science and has even reached the engineering level, where it is influencing the design of MRI systems and analog-to-digital

converters. One group of engineers at Rice University pushed the idea of compressive sampling to the limit and developed a digital camera that has just one pixel. This innovation was widely reported in the media and was highlighted in the annual report “Ten Emerging Technologies” in the March/April 2007 issue of the Massachusetts Institute of Technology magazine *Technology Review*. Remarked Candès, “It’s very surprising to see this transfer from theorems... to hardware so quickly.”

The third example of ongoing mathematical work that could be funded under CDI is that of the Atlas team, headed by Jeffrey Adams at the University of Maryland (the other principal investigators are Dan Barbasch, John Stembridge, Peter Trapa, and David Vogan). The team’s goal is to solve a major open problem in Lie group theory, namely, the question of classifying the irreducible unitary representations of a semisimple Lie group. Although research over the past several decades has not produced a conceptual picture that would answer this question, enough theory has been built up that it is thought to be possible, for any given Lie group, to compute its unitary representations. The main purpose of the Atlas project is to create a computer program that will take a Lie group  $G$  as input and return a list of its unitary representations; a secondary purpose is to make available software that can be used for group theory research. A program that would compute the unitary dual of  $G$ , or even decide whether a given representation is unitary, would be extremely useful in a variety of mathematical areas where unitary representations arise, particularly number theory and quantum mechanics.

The work of the Atlas team became known worldwide when the American Institute of Mathematics, which provides support for the project and serves as the sponsoring institution for the project’s NSF grant, mounted a publicity campaign to bring media attention to an Atlas milestone that focused on the exceptional Lie group  $E_8$ . This milestone consisted of the calculation of the irreducible admissible representations of the split real form of  $E_8$  and of its Kazhdan-Lusztig-Vogan polynomials. The algorithms and software for carrying out this feat were created by the mathematician-computer scientist Fokko du Cloux (who passed away in November 2006) and bring the use of computing in mathematics research to a new level of sophistication. The work on  $E_8$  is one step along the way to the ultimate goal of understanding which representations of Lie groups are unitary. The Atlas team found that the  $E_8$  calculation deepened their mathematical understanding and also provided a crucial test-bed for computational methods.

Mathematics institutes in the U.S. have in recent years organized many events that reflect the

themes of CDI. One example is the graduate summer school “Intelligent Extraction of Information from Graphs and High Dimensional Data”, held in July 2005 at the Institute for Pure and Applied Mathematics at the University of California at Los Angeles. The school focused on mathematics problems arising in such applications as automated feature extraction, face and shape recognition, image analysis, graph mining, social and transactional networks, and robust network design. A workshop called “Mathematical Issues in Stochastic Approaches for Multiscale Modeling” was held in May of this year as part of a longer program in dynamical systems at the Mathematical Sciences Research Institute in Berkeley. This workshop centered on computational and analytical techniques for studying complex physical, biological, geophysical, and environmental systems, in which stochastic methods are playing a prominent role. The work of Candès and others is the topic of the short course “Compressive Sampling and Frontiers in Signal Processing”, held in June of this year at the Institute for Mathematics and its Applications at the University of Minnesota. This same area of research is also the focus of this summer’s AMS von Neumann Symposium, “Sparse Representation and High Dimensional Geometry”.

### It’s Not Just Hardware

The examples given above are by no means the only ones, March pointed out. But they do illustrate some essential ways in which mathematics plays a role in cyber-enabled research: in creating new approaches to interpreting data, in fashioning new tools for handling and visualizing data, and in attacking theoretical problems that are inaccessible by conventional means. In these examples “the mathematics comes through in the way that mathematics usually does: in a clear conception of the problem, clear organization of the line of attack, and also innovation in the algorithmic side of things,” March explained. “It’s not just an issue of hardware, although one needs that. The issue bears on the algorithmic side. How the computation is conceived and organized plays a crucial role. This is an essentially mathematical issue, independent of whatever the goal of the computation is aimed at. So I think the role that mathematics and statistics are going to play in CDI is on the algorithmic side, helping to organize essential calculations so that they can be done effectively.”

*Information about CDI can be found in the NSF’s fiscal year 2008 budget request at the URL [http://www.nsf.gov/about/budget/fy2008/pdf/39\\_fy2008.pdf](http://www.nsf.gov/about/budget/fy2008/pdf/39_fy2008.pdf).*