

# THE CHRONICLE

of Higher Education

## Technology

[Home](#) [News](#) [Technology](#)



May 28, 2010

### **Crowd Science Reaches New Heights**

*By Jeffrey R. Young*

Baltimore

Alexander S. Szalay is a well-regarded astronomer, but he hasn't peered through a telescope in nearly a decade. Instead, the professor of physics and astronomy at the Johns Hopkins University learned how to write software code, build computer servers, and stitch millions of digital telescope images into a sweeping panorama of the universe.

Along the way, thanks to a friendship with a prominent computer scientist, he helped reinvent the way astronomy is studied, guiding it from a largely solo pursuit to a discipline in which sharing is the norm.

One of the most difficult tasks has been changing attitudes to encourage large-scale collaborations. Not every astronomer has been happy to give up those solo telescope sessions. "To be alone with the universe is a very dramatic thing to do," admits Mr. Szalay, who spent years selling the idea of pooling telescope images online to his colleagues.

Today, data sharing in astronomy isn't just among professors. Amateurs are invited into the data sets through friendly Web interfaces, and a schoolteacher in Holland recently made a major discovery, of an unusual gas cloud that might help explain the life cycle of quasars—bright centers of distant galaxies—after spending part of her summer vacation gazing at the objects on her computer screen.

Crowd Science, as it might be called, is taking hold in several other disciplines, such as biology, and is rising rapidly in oceanography and a range of environmental sciences. "Crowdsourcing is a natural solution to many of the problems that scientists are dealing with that involve massive amounts of data," says Haym Hirsh, director of the Division of Information and Intelligent Systems at the National Science Foundation. Findings have just grown too voluminous and complex for traditional methods, which consisted of storing

numbers in spreadsheets to be read by one person, says Edward Lazowska, a computer scientist and director of the University of Washington eScience Institute. So vast data-storage warehouses, accessible to many researchers, are going up in several scholarly fields to try to keep track of the wealth of information.

Persuading scientists to fully embrace the age of big data, though, will require a change in academic reward structures to give new currency to papers with more authors than ever and to scientists who spend their careers crunching other peoples' numbers.

"The culture shift is the sharing of data," says Mr. Lazowska. "And the astronomers have led the way."

#### **Astronomy Rebooted**

Mr. Szalay's unusual career began with a stint as a rock star. While in graduate school in Hungary, he played lead guitar in the band Panta Rhei, which released two albums and several singles in the 1970s. "I wouldn't call ourselves 'stars,' but we were pretty well received," he says, modestly. "We toured Germany, we went to Poland and Czechoslovakia."

Their sound was decidedly nerd rock—lots of plinky synthesizers and broken rhythms. The synthesizers were home-built. "In Communist Hungary you couldn't buy anything—you had to build things on your own," Mr. Szalay says. The willingness to tinker would become a hallmark of his career.

Mr. Szalay left the band to focus on his academic career after landing a postdoctorate position at the University of California at Berkeley, making solitary visits to telescopes as many astronomers did.

He wound up at Johns Hopkins, where he has been for most of the last 23 years.

Then in 1992 came the project that would change his career. Johns Hopkins joined the Sloan Digital Sky Survey project, a computerized snapshot of the heavens.

Mr. Szalay signed up to lead the design and building of the archive, even though he knew nothing about the technology of data storage. His research interests drove his decision to jump in: He was hoping to better understand the Big Bang by looking at the distribution of galaxies in the universe.

"I needed a lot of data that was well organized so I could easily apply statistical tools to it," he said. It was such an enormous task, though, that he promised his departmental colleagues he would devote all

his time to the sky survey and put aside any of his own trips to observatories. "I thought, OK, this is going to be six to eight years, I can deal with it," he said. "It turned out to be 18 years."

#### A Geeky Guide

A couple of years after Mr. Szalay joined the project, a colleague introduced him to Jim Gray, who was a kind of rock star himself—in the computer-science world. *Wired* magazine once wrote that the programmer's work had made possible ATM machines, electronic tickets, and other wonders of modern life.

When Mr. Szalay met him, Mr. Gray was a technical fellow at Microsoft Research and was looking for enormous sets of numbers to place in the databases he was designing.

The two men formed an instant friendship, and decided they had a lot to learn from each other.

"So I taught him astronomy, and he really turned into a very good astronomer—he became a card-carrying member of the community," says Mr. Szalay.

And Mr. Gray taught the astronomer computer science. Mr. Szalay has now published so many papers about his work on databases that he has a joint appointment in the computer-science department at Hopkins.

As the sky survey matured, though, many traditional stargazers remained skeptical.

"The astronomical community did not believe we would ever really make the data public," says Mr. Szalay. The typical practice in the mid-1990s was to guard data because it was so difficult to get telescope time, and scholars did not want to get scooped on an analysis of something they gathered.

One incident demonstrates the mood at the time. A young astronomer saw a data set in a published journal and wanted to reanalyze it, so he asked his colleague for the numbers. The scholar who published the paper refused, so the junior scholar took the published scatterplot, guessed the numbers, and published his own analysis. The original scholar was so upset that he called for the second journal to retract the young scholar's paper.

Mr. Szalay said that astronomers changed their minds once the first big data sets hit the Web, starting with some images from NASA, followed by the official release of the first Sloan survey results in 2000.

"Once they saw the first data release, and they also saw that it was easy to use, I think they started turning around," he said.

And Mr. Szalay and Mr. Gray spoke at many astronomy conferences, presenting a list of 20 questions that could be answered only with large, shared data sets, to try to win support for the approach. They felt they were onto something that would have an impact far beyond astronomy.

"We realized that this is the new way of doing science," said Mr. Szalay. "Computers are becoming a new kind of instrument."

#### Lost at Sea

In 2007 tragedy ended their long partnership. Mr. Gray set out from San Francisco on a solo trip on his 40-foot sailboat and did not return.

His friends in computer science and astronomy quickly mobilized what has become a legendary search effort, taking their ideas about crowdsourcing to a new level in the process.

The scientists, along with tech-industry leaders whom Mr. Gray had mentored in the past, offered to help the Coast Guard search the open sea using any technology they could think of. Google executives and others helped provide fresh satellite images of the area. And an official at Amazon used the company's servers to send those satellite images to volunteers—more than 12,000 of them stepped forward—who scanned them for any sign of the lost researcher.

Mr. Szalay and his son, Tamas, wrote software that would make the satellite images clearer and led a parallel analysis with researchers who volunteered via e-mail.

But Jim Gray was never found.

Some of the techniques that the astronomer learned from the search effort, though, have now been incorporated into a Web site that invites anyone to help categorize images from the Sloan Digital Sky Survey.

It's called Galaxy Zoo, and it's led by Chris Lintott, an astronomer at the University of Oxford.

Just click "classify galaxies" on the Galaxy Zoo Web site, and a picture from a telescope appears, along with questions including "Is the galaxy smooth and rounded?" and "Does the galaxy have a mostly clumpy appearance?" Visitors must register and complete a short tutorial before their results are counted. Each image is shown

to at least 10 different people to try to cut down on erroneous classifications. If 80 percent of the crowd agrees on a classification of an image, it sticks. Otherwise, the image might go through the whole process again.

"It's not some fun game online while the scientist do the real work," says Mr. Lintott. "I hope visitors are learning that science is not just something done by people in lab coats in some underground bunkers. Science is something people can get involved in."

The number of volunteers surprised the organizers. "The server caught fire a couple of hours after we opened it" in July 2007, he said, burning out from overuse. More than 270,000 people have signed up to classify galaxies so far.

One of them is Hanny van Arkel, a schoolteacher in Holland, who found out about the site after her favorite musician, Brian May, guitarist for the rock group Queen, wrote about it on his blog.

After clicking around on Galaxy Zoo for a while one summer, she landed on an image with what she describes as a "very bright blue spot" on it. "I read the tutorial and there was nothing about a blue spot," she says, so she posted a note to the site's forums. "I was just really wondering, What is this?"

Her curiosity paid off.

Scientists now believe the spot is a highly unusual gas cloud that could help explain the life cycle of quasars. The Hubble telescope was recently pointed at the object, now nicknamed "Hanny's Voorwerp," the Dutch word for object.

Astronomers have published papers about the discovery, listing Ms. van Arkel as a co-author. "Don't ask me to explain them to you, but I am a co-author of them," she says with a laugh.

Now other disciplines have approached Galaxy Zoo to find out how they can use the approach.

#### Gene Wikis

Astronomy is just one of many disciplines being reshaped by a data explosion. Bioscientists have found that decoding entire genomes also meant cultural shifts for their profession. Again, persuading professors to take the time to share proved to be a challenge.

A case in point is a project to create a genetic road map using the same wiki platform that supports Wikipedia.

It started under the name of GenMAPP, or Gene Map Annotator and Pathway Profiler. Participation rates were low at first because

researchers had little incentive to format their findings and add them to the project. Tenure decisions are made by the number of articles published, not the amount of helpful material placed online. "The academic system is not set up to reward the sharing of the most usable aspects of the data," said Alexander Pico, bioinformatics group leader and software engineer at the Gladstone Institute of Cardiovascular Disease.

In 2007, Mr. Pico, a developer for GenMAPP, and his colleagues added an easy-to-edit Wiki to the project (making it less time-consuming to participate) and allowed researchers to mark their gene pathways as private until they had published their findings in academic journals (alleviating concerns that they would be pre-empting their published research). Since then, participation has grown quickly, in part because more researchers—and even some pharmaceutical companies—are realizing that genetic information is truly useful only when aggregated.

"There's a sort of a call to action in the biology community right now toward sharing data in usable formats and usable ways," says Mr. Pico. But he admits some in the field are still skeptical that sharing will become the norm.

Another gush of data is happening deep in the Pacific Ocean, as a series of thousands of sensors strung along an underwater fiber-optic cable, along with new self-guided mobile sensors that can beam back data, promises to make oceanography the next field to embrace the data revolution and a crowd approach.

Mr. Lazowska, the computer scientist at the University of Washington who focuses on data-driven science, says that at the moment oceanography is "expeditional," meaning that data are hard to come by because only a few organizations can afford the equipment to probe the depths. But new technologies, like those mobile sensors, promise to pipe in more data than scientists can manage without a shared database, like what the Sloan project did for astronomy.

"In oceanography the individual investigator tends to be king or queen—it's individual papers that really determine how one proceeds in the field," said John Orcutt, a professor of geophysics at the Scripps Institution of Oceanography at the University of California at San Diego. "Generally there haven't been big data undertakings in the past, but there are many pressures now that are forcing that change, and I believe we're moving toward a different sort of world."

Major issues remain unresolved. As data continue to grow at an ever-more-rapid pace, more efficient ways to store and process the information will be needed. Computer algorithms will play an increasing role, too, so that robot scientists can do some classifying, perhaps checked by human volunteers.

Mr. Szalay spends much of his time trying to build faster servers to handle all that telescope data.

He's involved with a new project, the National Virtual Observatory, which will link many large telescope data sets that have emerged in recent years.

And he is focused on training the next generation of astronomers to become card-carrying computer scientists—to learn as much about mapping data as mapping the heavens. They will need such training, he argues, to master a new paradigm of science and answer the universe's biggest questions.

Copyright 2011. All rights reserved.

The Chronicle of Higher Education 1255 Twenty-Third St, N.W. Washington, D.C. 20037