Democratizing BIG DATA

By Sarah DeWeerdt

SeaFlow, a research instrument developed in the lab of UW School of Oceanography director Ginger Armbrust, analyzes 15,000 marine microorganisms per second, generating up to 15 gigabytes of data every single day of a typical multi-week-long oceanographic research cruise.

UW professor of astronomy Andy Connolly is preparing for the unveiling of the Large Synoptic Survey Telescope (LSST), which will map the entire night sky every three days and produce about 100 petabytes of raw data about our universe over the course of 10 years. (One petabyte of music in MP3 format would take 2,000 years to play.)

What scientists like Armbrust and Connolly have is popularly known as "big data," and as rich and exciting as it can be, big data can also be a big problem.

"Every field of discovery is transitioning from data-poor to data-rich, and the people doing the research don't have the wherewithal to cope with this data deluge," says Ed Lazowska, director of the UW's eScience Institute. "And it's not just the volume of data that's increasing relentlessly, it's the velocity and the variety too - the 3 V's." The result is that many scientists spend more time wrangling data than actually doing science.

The eScience Institute aims to change that by connecting researchers with experts in large-scale data management, data analysis, data visualization, machine learning and related fields. Researchers from across the campus gain the skills and tools they need to work with increasingly enormous data sets, while data scientists advance their own research by grappling with real-world problems.

And now, the eScience team — the core team includes faculty from 12 departments representing five schools and colleges — is poised to scale way up. Last year, the UW won a five-year, $37.8 million grant from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation that will be shared with New York University and the University of California, Berkeley, to foster a data science culture at the three universities.

At the UW, that shift is already underway. The eScience Institute was founded in 2008, modeled in part on the university's Center for Statistics in the Social Sciences, which has been in existence for nearly 15 years.

"Having the history of CSSS, and having the eScience Institute in operation at a modest scale, provided a lot of experience that helped us convince the sponsors to give us this grant," says Bill Howe, associate director of the eScience Institute. "Institutionally and culturally I think we are a bit ahead of the curve."

In addition to the Moore/Sloan Foundation grant, the eScience Institute also received $9.3 million from Washington Research Foundation to aid in hiring faculty and postdoctoral fellows, and $2.8 million from the National Science Foundation to support creation of an interdisciplinary graduate program in data science. "You need to have career paths for people who simultaneously do science and build tools," says Lazowska. "We want to create them, hire them, and reward them."

Real-time Analysis of a Sea of Data

Researchers who have worked with the eScience Institute call the experience transformational.

"Out of these collaborations come not just faster but new ways of thinking about our studies and the kinds of questions we can ask," says Armbrust. "The faster part is great. But that's not as important as the new ways of doing things."

Armbrust sought help from the eScience Institute for dealing with the massive streams of data generated by her lab's SeaFlow instrument. Out of that collaboration came SQLShare, a web-based tool that makes database technology accessible to non-specialists.

A plot of multicolored dots dances on Armbrust's computer screen: data from an ongoing research cruise. The colors represent different types of phytoplankton — the base of the marine food web — shifting in abundance as the ship moves through the water. The research vessel is far out in the Pacific Ocean, but Armbrust sees the data in her Seattle office with just an eight-minute delay.

"Most of the time when you're at sea you're sort of driving blind," Armbrust says. In the past, oceanographers often couldn't analyze their data and identify interesting phenomena in the ocean until days after passing it by — too late to follow up. "Now we're watching things in real time and we can make choices in real time.

Real-time analysis in SQLShare also bolsters collaborations across disciplines. Armbrust recently organized a workshop that brought together 40 oceanographers from different disciplines. They loaded dozens of individual data sets — all collected on the same research cruise — into SQLShare, and during the meeting two data scientists from the eScience Institute typed in queries as ideas for scientific analyses came up in the conversation.

The oceanographers were able to investigate the relationship between zinc and cobalt, for example, or how salinity affects the levels of a certain virus, almost as easily as people look up an actor's filmography or consult Wikipedia on their smartphones during cocktail party conversation. "It was a blast," Armbrust says.

Scaling Up to Astronomical Data Sets

In the Department of Astronomy, Connolly has worked with the eScience Institute to develop algorithms to sort through the massive amounts of data that will come from the LSST project. For example, researchers hope the telescope's 3.2-billion-pixel camera will help detect asteroids; orbiting chunks of rock that can help reconstruct the evolution of the solar system — not to mention identify which ones have the potential to slam into the Earth.

LSST data may enable astronomers to detect up to 10 million asteroids within our solar system. But they won't be able to do it the old-fashioned way. Traditionally, astronomers detect asteroids by rapidly toggling back and forth two pictures of the night sky taken a few hours apart.

"It's really effective — your eye is really good at this," Connolly says. "But it doesn't scale."

Connolly worked with eScience experts to develop algorithms that can detect objects that move like asteroids on LSST images, returning to the same region of the sky every few nights to see what has moved.

"It's like playing the world's largest game of connect-the-dots," Connolly says. "You can answer all of those questions with the same data set, you're just looking at it in different ways. That's where the computer science comes in."

Sharing Data Science Know-How

These questions also reflect widespread issues in computer science — problems like tracking and detection of anomalies. As such, algorithms that are developed to help astronomers do their work can have much wider application. "We've had a number of examples where people in different fields really do have problems that are similar enough that they can utilize the same solutions," Lazowska says.

The institute's algorithm for bringing people together will get a big boost in October, when the former library on the sixth floor of the Physics/Astronomy building is slated to reopen as a campus-wide Data Science Studio: a collaboration space, providing a central spot for like-minded researchers to gather.

As demand increases at the UW for their expertise, eScience Institute data scientists are expanding their reach despite their limited numbers. One strategy is the Data Science Incubation Program, in which researchers across campus pitch their data analysis projects, then send a lab member to collaborate with eScience experts for one or two quarters to build a solution.

"We don't want this to be a magic trick that only computer scientists know how to do," Howe says. "It should be something that everybody can do."