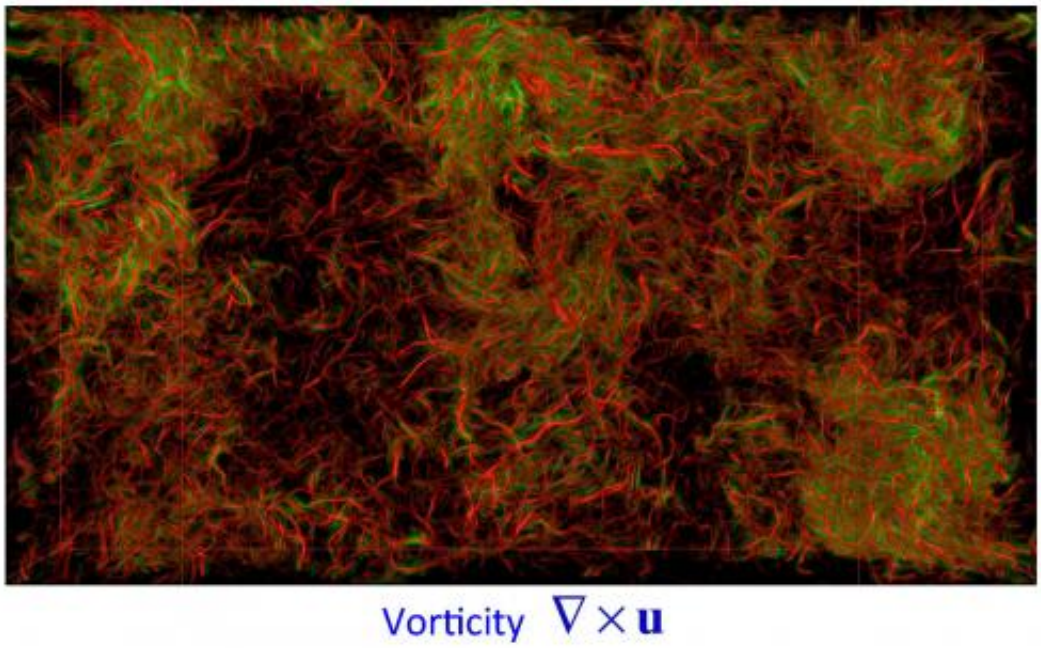
Big Data infrastructure for science

December 5 2014



Tangle of vortices in a snapshot of numerically simulated three-dimensional and isotropic fluid turbulence. Credit: Visualization by Kai Buerger (TU Munchen) based on data from the Johns Hopkins Turbulence Databases.

Big Data comes naturally to science. Every year, scientists in every field, from astronomy to zoology, make tremendous leaps in their ability to generate valuable data.

But all of this information comes at a price. As datasets grow exponentially, so do the problems and costs associated with accessing, reading, sharing and processing them.

A new project called SciServer, supported by the National Science Foundation (NSF), aims to build a long-term, flexible ecosystem to provide access to the enormous data sets from observations and simulation.

"SciServer will help meet the challenges of Big Data," said Alex Szalay of Johns Hopkins University, the principal investigator of the five-year NSF-funded project and the architect for the Science Archive of the Sloan Digital Sky Survey. "By building a common infrastructure, we can create data access and analysis tools useful to all areas of science."

SciServer's heritage: Big Data in astronomy

SciServer grew out of work with the Sloan Digital Sky Survey (SDSS), an ambitious, ongoing project to map the entire universe.

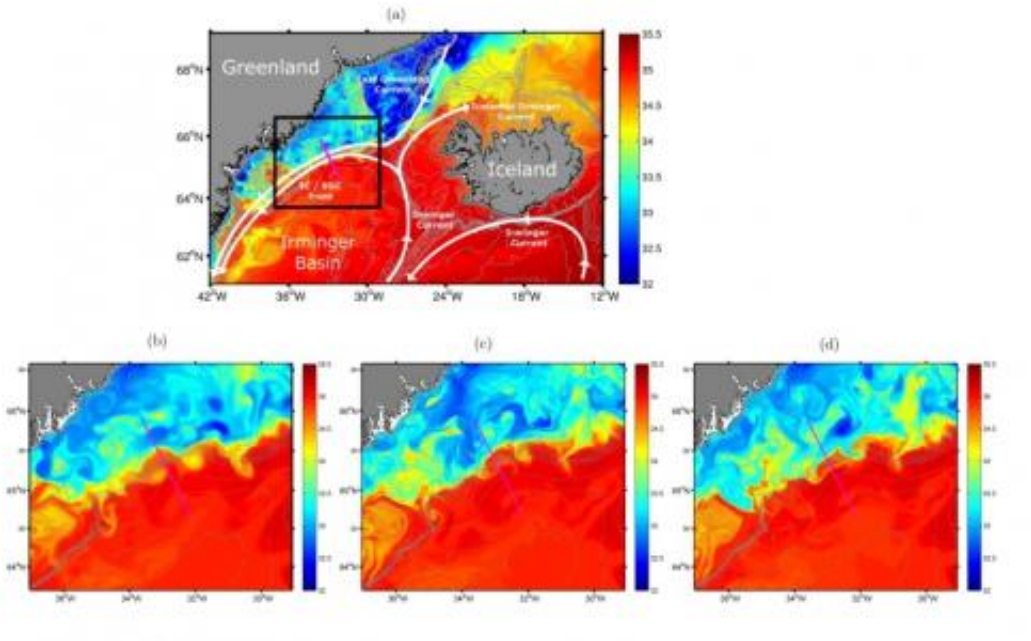
"When the SDSS began in 1998, astronomers had data for less than 200,000 galaxies," said Ani Thakar, an astronomer at Johns Hopkins who is part of the SciServer team. "Within five years after SDSS began, we had nearly 200 million galaxies in our database. Today, the SDSS data exceeds 70 terabytes, covering more than 220 million galaxies and 260 million stars."

The Johns Hopkins team created several online tools for accessing SDSS data. For instance, using the SkyServer website, anyone with a web browser can navigate through the sky, getting detailed information about stars or searching for objects using multiple criteria. The site also includes classroom-ready educational activities that allow students to learn science using cutting-edge data.

To allow users—scientists, citizen scientists, students—to run longer-term analyses of the Sloan data, they created CasJobs, an online workbench where registered users can run queries for up to eight hours

and store results in a personal "MyDB" database for later analysis.

With each new tool, the community of users grew, leading to more and more scientific discoveries.



Remote sensing data are important in many branches of science, such as physical oceanography, meteorology, and climatology. This illustration shows such data: sea surface salinity off the East Greenland Shelf (from a paper by Magaldi and Haine, in press).

The problem: data without infrastructure

One major challenge in managing and extracting value from Big Data is simply preserving the data as file formats change and scientists retire. Another challenge is that most datasets are stored in an ad hoc manner with insufficient metadata for describing how the data should be

interpreted and used. Yet another challenge is unequal access to data and expertise among researchers.

Even when individual datasets are well-preserved, the difficulty of combining data for joint analysis means that researchers miss opportunities for new insights. The result is that scientists work inefficiently and miss chances to grow their research projects in new directions.

A variety of projects have developed approaches to preserving and managing datasets, but providing easy access so all researchers can compare, analyze and share them remains a problem. The SciServer team has spent the last two decades addressing these problems, first in astronomy and then in other areas of science.

From SkyServer to SciServer: the new approach

Led by Szalay, the team began work on SciServer in 2013 with funding from NSF's Data Infrastructure Building Blocks program.

Set to launch in phases over the next four years, SciServer will deliver significant benefits to the scientific community by extending the infrastructure developed for SDSS astronomy data to many other areas of science.

"Our approach in designing SciServer is to bring the analysis to the data. This means that scientists can search and analyze Big Data without downloading terabytes of data, resulting in much faster processing times," Szalay said. "Bringing the analysis to the data also makes it much easier to compare and combine datasets, allowing researchers to discover new and surprising connections between them."

Szalay and his team are working in close collaboration with research

partners to specify real-world use cases to ensure that the system will be most helpful to working scientists. In fact, they have already made significant progress in two fields: soil ecology and fluid dynamics.

To help ease the burden on researchers, the team developed "SciDrive," a cloud data storage system for scientific data that allows scientists to upload and share data using a Dropbox-like interface. The interface automatically reads the data into a database, and one can search online and cross-correlate with other data sources.

SciServer will extend this capability to a new citizen science project called GLUSEEN (Global Urban Soil Ecological & Educational Network), which aims to gather worldwide distributed data on soil ecology across a range of climatic conditions. SciDrive will offer extensive new collaborative features and will allow individuals to connect remote sensor measurements to weather and other datasets that are available from external worldwide providers.

"Our approach with SciDrive and citizen science immediately will be useful to many other areas of science where datasets managed by individual researchers must be combined with larger publicly-available datasets," said Szalay.

SciServer also has a major initiative underway to develop an "open numerical laboratory" for the access and processing of large simulation databases. Working with the Turbulence Simulation group at Johns Hopkins, they are developing a pilot system to integrate data sets and processing workflows from simulation of turbulence into SciServer.

As the SciServer system becomes more mature, the team will expand to benefit other areas of science including genomics—where researchers must cross-correlate petabytes of data to understand entire genomes—and connectomics—where researchers explore cellular

connections across the entire structure of the brain. These collaborations will be spread over a five-year period from 2013 to 2018, and will allow SciServer to be incrementally architected and developed to support its growing capabilities.

"Our conscious strategy of 'going from working to working'—building tools by adapting existing, working tools—is a key factor in ensuring the success of our project," Szalay said. "The tools we build will create a fully-functional, user-driven system from the beginning, making SciServer an indispensable tool for doing science in the 21st century."

More information: M.G. Magaldi, T.W.N. Haine, "Hydrostatic and nonhydrostatic simulations of dense waters cascading off a shelf: the East Greenland case," *Deep-Sea Research I*, DOI: 10.1016/j.dsr.2014.10.008

Provided by National Science Foundation