Microsoft

AI for Earth Grantee Profile

Dax Soule and Timothy Crone

Furthering oceanographic study with the Microsoft cloud

Summary

The Ocean Observatories Initiative (OOI) Cabled Array collects large quantities of data from the seafloor and overlying ocean environment of the Juan de Fuca tectonic plate in the northeast Pacific—providing a valuable opportunity for researchers and students to learn more about the ocean and seafloor processes. But currently, downloading and processing the data on local computers takes days or even weeks. With funding from the Microsoft AI for Earth program, Dr. Timothy Crone and Dr. Dax Soule are helping make this data more accessible and usable to scientists and students around the globe by building a Microsoft cloud-based system on Pangeo, an open-source platform for big data geoscience.

Increasing access to technology is a passion for Dr. Soule. At the City University of New York (CUNY), he teaches students from very diverse ethnic and economic backgrounds that are often not well-represented in science-related careers. Through the partnership with Microsoft, Dr. Soule and his students now have the cloud-based tools they need to access and work with the OOI Cabled Array data, conducting important research and becoming the next generation of oceanographic scientists.

Furthering oceanographic study with the Microsoft cloud

Our oceans are crucial to life on Earth. From simply providing food to influencing climates and global weather, and even to absorbing increasing amounts of carbon emissions, the oceans support us in many ways. Studying the oceans is thus a vital scientific activity that can help us better understand and deal with the threats of climate change. To that purpose, the Ocean Observatories Initiative (OOI) was established by the National Science Foundation (NSF) to provide long-term, continuous observations from the seafloor to the sea surface from several key locations. With NSF funding and coordination, the OOI developed, constructed, and maintains a network of oceanic infrastructure and sensors for scientific research. The OOI network integrates multiple scales of globally distributed marine observations into one observing system and allows for that data to be freely downloaded over the internet in near-real time.

One part of that network is the OOI Cabled Array, which supplies high power and bandwidth through approximately 900 kilometers of fiber-optic cable to various sensors and instruments at several different locations off the northwestern coast of the US. The array spans the Juan de Fuca tectonic plate, including sensors at Axial Seamount, which is the most magmatically active volcano on the plate's outer ridge, and

provides over 100 real-time streaming datasets on a variety of aspects of the volcanic-marine environment. Collectively, the various geophysical, chemical, and biological sensors and high-definition cameras have generated over 300 terabytes of data and continue to generate about 5 terabytes a month.

Axial Seamount is an area of particular scientific interest as the site is both geologically complex, combining a volcanic hotspot with a mid-ocean ridge, and conveniently located in relatively shallow waters close to the Oregon coast. Through the OOI Cabled Array, scientists are monitoring the relationship between local earthquakes at the seamount caused by volcanic seismicity and the magma chamber inflation. The volcano was first identified through satellite altimetry in the 1970s, and the first remote sensor packages—including seismometers and pressure, tilt, and temperature sensors—were deployed there starting in the late 1980s. With this variety of sensors in place, the 1998 eruption became the first submarine eruption to be observed entirely on site, and the data gathered through that eruption helped to establish a predictable correlation between the inflation of the volcano's magma chamber and its eventual eruption. Further eruptions in 2011 and 2015 provided more data to define the cycle of inflation and eruption.

Building a Pangeo system on the Microsoft cloud

The OOI Cabled Array was designed without cloud computing in mind. The data generated from its sensors is gathered into very large files that must be downloaded locally to be analyzed and used. Not only does this cause long delays for downloading, but also it requires powerful local computers to process the massive amounts of data. Moving, analyzing, and visualizing data from the OOI repository in a non-cloud environment can take days or even weeks. However, Dr. Timothy Crone, a marine geophysicist with the Lamont-Doherty Earth Observatory at Columbia University, has developed a project, funded by a grant from Microsoft's AI for Earth program, to use Azure AI and cloud services and the Azure Open Datasets program to make this data more accessible and more usable to scientists worldwide.

> Making OOI Cabled Array data more accessible and usable helps the scientific community gain better insights on geologic, oceanographic, and climate processes.

The first goal of Dr. Crone's project is to move all of the OOI Cabled Array data into modern cloud-optimized data storage formats hosted in a 500-TB Azure Blob Storage repository, enabling scientists to use the latest open-source data analysis and visualization tools with the data. Using scripts, most of the OOI data can be automatically transferred into the Zarr data format, which can take advantage of out-of-memory processing

pipelines for efficient access and workflows. Crone authored the ABSStore storage class that enables the use of the Zarr format in Azure. Because the OOI video data is already stored in an efficient format (through the ProRes codec), Crone is building a service layer for Azure capable of decompressing images on the fly for scientific analysis. That service layer will work with the PyCamHD utility (also developed by Dr. Crone), which facilitates the extraction of individual frames for analysis without downloading entire files.

With the Cabled Array data ported to Azure, Azure Kubernetes Services and Data Science Virtual Machines (DSVMs) will provide a computing environment for scientists to gain insights on geologic, oceanographic, and climate processes. This specialized computing environment is based on a popular community platform and software ecosystem for big data geoscience called Pangeo. For scientists doing remote computation on massive environmental datasets, Pangeo offers some of the best environmental data analysis and visualization tools available today.

Crone's Azure-enabled Pangeo environment will allow researchers to easily access OOI data and build correlational and predictive models of the relationships between the seafloor lithosphere, the ocean, the atmosphere, and the biospheres among and across these spaces. In particular, with the next eruption of the Axial Seamount predicted to occur by 2022, scientists can focus on the relationships among the temperature, pressure, seismicity, and other aspects of the inflation and eruption cycle, and how those may relate to climate change processes as well.

Overcoming challenges to science research and education

Another goal of this project is to reach more students, and a wider variety of students, by taking advantage of the cloud. Dr. Crone has already begun this outreach in collaboration with Dr. Dax Soule and his students in the School of Earth & Environmental Sciences at Queens College, an undergraduate-focused institution that is part

> The Azure platform helps Soule to help a wider diversity of students to become the next generation of scientists.

of the City University of New York (CUNY). When Dr. Soule started at Queens College as a tenure-track lecturer, his position came with no budget or resources for research—and research was vital to pursuing a professorship. However, what he did have was free access through the internet to the vast amounts of data from the OOI. In addition to introducing Soule's team to cloud computing and the Microsoft AI for Earth program, Crone has provided direct mentorship to both Soule and his students, working one-on-one with students on their projects. Soule credits Crone as the catalyst for adopting Azure Services and driving Soule's

capabilities with its resources. With a grant from Microsoft's AI for Earth program, that access to the Cabled Array data not only helps Soule's career, but his students' education as well.

The student body of Queens College is as diverse as New York City itself; its 19,000 members come from many different ethnic and cultural backgrounds around the world. Many of these ethnic groups are underrepresented in STEM academic programs and professions. Additionally, many of the students come from economically challenged situations, with more than half reporting annual household incomes under $30,000 and receiving need-based financial aid. Such students also historically have limited opportunities to enter STEM fields. Dr. Soule considers it part of his role to help these students overcome their challenges to get their degrees and careers in science.

Opening opportunities to students and scientists through the cloud

With the cloud-based AI services and virtual machines provided by Azure, Dr. Soule can set up online environments for conducting data analysis. That opens up the opportunity to involve students in his research—the students don't need the latest and fastest (and expensive) computers with lots of storage to hold and process the data, they just need an internet connection and Azure access. And through the OOI data providing virtual access to the ocean, the students can gain practical experience in scientific research while

Dax Soule working in the lab with his students. Credit: Andy Poon, Queens College

simultaneously improving their quantitative reasoning, computer literacy, and technical capability to deal with ever-expanding Earth science datasets. In this way, the Azure platform helps Soule help a wider diversity of students to become the next generation of scientists.

Beyond his own classroom, Dr. Soule is also involved in additional programs to engage other college-level teachers in exploring the opportunities presented by the OOI Cabled Array data. The OOI Data Labs workshops focus on using OOI data in undergraduate teaching of introductory oceanography themes and concepts. Meanwhile, Project EDDIE seeks to develop flexible classroom modules using large, publicly available, digital data for undergraduate students in biology, geology, and environmental science, as well as provide the associated professional development needed to ensure their effective use. Both workshop series aim to reach a wide variety of faculty ranging from tribal colleges, historically black universities and Hispanic-serving institutions to top-ranked research universities. As the OOI data will be hosted on Azure, a natural extension of this work will be to make the participants aware of Azure Cloud Services and the AI for Earth Grant Program.

Going forward

Dr. Soule and Dr. Crone plan to extend their research in ways that highlight both the transformative science and the exciting pedagogical opportunities that Microsoft Azure cloud computing resources make possible. For example, the team plan to explore the relationship between volcanic eruption cycles and climate change. In the next three years, the Axial seamount is expected to erupt, releasing a flux of carbon dioxide. This creates an opportunity to study the significance of eruptions relative to the global budget of greenhouse gasses. As the program matures, the team hopes to host the entire data stream from the Cabled Array on Azure, giving other scientists the ability to apply cutting-edge analysis algorithms on measurements that characterize numerous processes occurring from deep within the ocean all the way into the atmosphere. “Our hope is that the broader scientific community leverages this data to gain insights into Earth, ocean, and climate processes—resulting in not only more accurate climate models, but also explorations into how Azure AI and machine learning can be used to build system-level predictive models,” explains Dr. Soule.

About Dax Soule

Dax Soule is an assistant professor in the School of Earth & Environmental Sciences at Queens College, an undergraduate-focused institution that is part of the City University of New York (CUNY) system in New York City. After leaving Texas A&M University and spending a decade working in car sales, Soule decided to return to school with the intent to finish his degree in history. Instead, by chance he was inspired to pursue a career in oceanography, earning his undergraduate degree in geophysics at Texas A&M and then completing doctoral studies in oceanography at the University of Washington. Today, his research uses seismic tomography to explore the structure of the ocean crust near mid-ocean ridge-spreading centers. A lecturer at Queens College since 2016, Dr. Soule was promoted to assistant professor in 2018.

About Timothy Crone

Timothy Crone is a marine geophysicist with the Lamont-Doherty Earth Observatory at Columbia University. He is currently studying the nature of fluid flow variability within seafloor hydrothermal systems, with the goal of understanding how flow variations affect hydrothermal fluxes and subseafloor biological production. Crone uses numerical models and laboratory studies to investigate these processes and develops new observational techniques and instrumentation to verify model predictions. Since earning his doctorate degree in oceanography at the University of Washington in 2007, Dr. Crone has worked at the Lamont-Doherty Earth Observatory, rising from postdoctoral research fellow to his current position as Lamont Associate Research Professor in January 2018.

Resources

Websites

Ocean Observatories Initiative main site
Cabled Array—OOI site page about the Cabled Array
Interactive Oceans—UW home site for the OOI Cabled Array

Press

Dax Soule – Using the OOI to Build Paths for Success in His Students and His Research