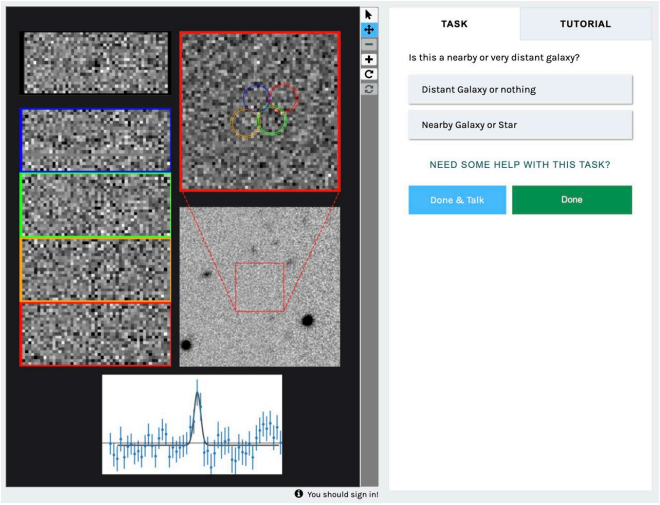
Citizen science, supercomputers and AI

7 January 2022, by Aaron Dubrow



Screenshot from the 'Dark Energy Explorers' citizen science app that lets non-experts differentiate real galaxies from false positives, in the process training a machine learning model to help search for dark energy. Credit: Karl Gebhardt, UT Austin

Citizen scientists have helped researchers discover new types of galaxies, design drugs to fight COVID-19, and map the bird world. The term describes a range of ways that the public can meaningfully contribute to scientific and engineering research, as well as environmental monitoring.

As members of the Computing Community Consortium (CCC) recently argued in a Quadrennial Paper, "Imagine All the People: Citizen Science, Artificial Intelligence, and Computational Research," non-scientists can help advance science by "providing or analyzing data at spatial and temporal resolutions or scales and speeds that otherwise would be impossible given limited staff and resources."

Recently, citizen scientists' efforts have found a new purpose: helping researchers develop machine learning models, using labeled data and algorithms, to train a computer to solve a specific task.

This approach was pioneered by the crowdsourced astronomy project Galaxy Zoo, which started leveraging citizen scientists in 2007. In 2019, researchers used labeled data to train a neural network model to classify hundreds of millions of unlabeled galaxies.

"Using the millions of classifications carried out by the public in the Galaxy Zoo project to train a neural network is an inspiring use of the citizens science program," said Elise Jennings, a computer scientist at Argonne Leadership Computing Facility (ALCF) who contributed to the effort.

TACC is supporting a number of projects—from identifying fake news to pinpointing structures in danger during natural hazards—that use citizen science to train AI models and enable new scientific successes.

Tinder for galaxies

The Hobby-Eberly Telescope Dark Energy Experiment, or HETDEX, is the first major experiment to search for evolution in dark energy. Based at the McDonald Observatory in West Texas, it looks deeper into the past than ever before to determine with great accuracy how fast the universe is accelerating.

The experiment relies on being able to identify the location, distance, and redshift of tens of millions of galaxies. But Karl Gebhardt, a professor of Astronomy at The University of Texas at Austin (UT Austin) and lead scientist on the project, faced a problem. The computational algorithms were having difficulty separating real target galaxies from false positives.

Strangely enough, humans can detect the difference easily. So, working with graduate students Lindsay House and Dustin Davis, and data scientist Erin Mentuch Cooper, they created a

citizen science app called 'Dark Energy Explorers' to train a machine learning algorithm to assist in the process.

Individuals with minimal training can look at spectral lines and images of point sources and swipe left or right, depending on whether they believe it is a real galaxy or something else such as an artifact of the algorithm or a speck of dust on the sensor. The app has jokingly been called "Tinder for Galaxies," Gebhardt says. To date, citizen scientists have made almost 2 million classifications and more are needed.

After enough of these determinations are made, Gebhardt will use TACC's machine learning-centric Maverick supercomputer to train the galaxy detection model. The analysis will map over 1 million target galaxies and determine the rate of cosmic acceleration.

Labels to save lives

Another prime example of citizen science is the "Building Detective for Disaster Preparedness" project developed by the SimCenter of UC Berkeley. It invites the public to identify specific architectural features of buildings, like roofs, windows, and chimneys. These labels are then used to train additional AI modules for the researchers' citywide simulations of natural hazard events.

The project, hosted on the citizen science web portal Zooniverse, has been an unqualified success. "We launched the project in March and within a couple of weeks we had a thousand volunteers, and 20,000 images annotated," said Charles Wang, assistant professor in the College of Design, Construction and Planning at the University of Florida and lead developer of a suite of AI tools called BRAILS—Building Recognition using AI at Large-Scale.

The "Building Detective For Disaster Preparedness" project in Zooniverse invites citizen scientists to label data that helps train the BRAILS tool. Credit: SimCenter, UC Berkeley

BRAILS applies deep learning—multiple layers of algorithms that progressively extract higher-level features from the raw input—to automatically classify features in millions of structures in a city. Architects, engineers, and planning professionals can use these classifications to assess risks to buildings and infrastructure, and they can even simulate the consequences of natural hazards.

"To successfully tackle pressing scientific and societal challenges, we need the complementary capabilities of both humans and machines," the CCC authors wrote. "The Federal Government could accelerate its priorities on multiple fronts through judicious integration of citizen science and crowdsourcing with artificial intelligence (AI), Internet of Things (IoT), and cloud strategies."

Biases and bad data

There are challenges, of course, to datasets generated by citizen scientists or other amateurs (paid or volunteer). Matt Lease, an associate professor in the School of Information at UT Austin, employs crowdsourced labor for AI training. He also studies the dynamics of these human-computer interactions.

Lease recently paid non-professionals to label whether or not a tweet should be considered hate speech, and used this data to train a hate speech

classification model. His team has similarly collected data from crowd workers about whether articles were fake news, which they used to train a prediction model.

Lease said he believes data is potentially the most under-valued aspect in developing accurate AI models (He fleshes this perspective in a recent arxiv article that will appear in the March/April issue of *ACM Interactions.)*

"Research to improve models is often prioritized over research to improve the data environments in which models operate, even though mismatches between datasets and the real-world can lead to significant modeling failures in practice," he said. "Improvements in prediction accuracy from better data can exceed improvements from better models."

He pointed to a recent study that showed that the ten most cited AI data sets are riddled with label errors. "Data quality is crucial to ensure that AI systems can accurately represent and predict the phenomenon it is claiming to measure," he said.

However, sometimes the biases themselves can be gleaned from studying the datasets and can suggest better ways to collect data. "There have been findings that hate speech detection models may be biased against African-American speech," said Lease. "Just as companies should hire diverse workers to create products incorporating diverse perspectives, so too should AI data be labeled by diverse workers so that AI models learned from data will similarly reflect diverse perspectives."

Probing the limits of citizen science

Ben Goldstein, a Ph.D. candidate at UC Berkeley, is writing a dissertation motivated by the question: what kinds of information can we get out of the wealth of citizen science biodiversity data available?

Goldstein and his collaborators Sara Stoudt and Perry de Valpine are comparing iNaturalist to eBird data to estimate which species are over- or under-reported relative to a baseline.

Goldstein was awarded an allocation by the NSF-funded Extreme Science and Engineering Discovery Environment to use Jetstream, a national science and engineering cloud co-located at TACC and Indiana University, for the study.

"We argue that this 'overreporting index' captures human preference," he said. "We use it to identify which species and traits—size, color, rarity—are perceived as charismatic." They published the results of their study in Biorxiv.

Citizen science is as old as science itself, and yet it has more tricks to teach us, if we can learn to harness it properly. By employing cutting edge computational tools, citizen science is poised to add even more value to the traditional scientific enterprise.

Provided by Texas Advanced Computing Center