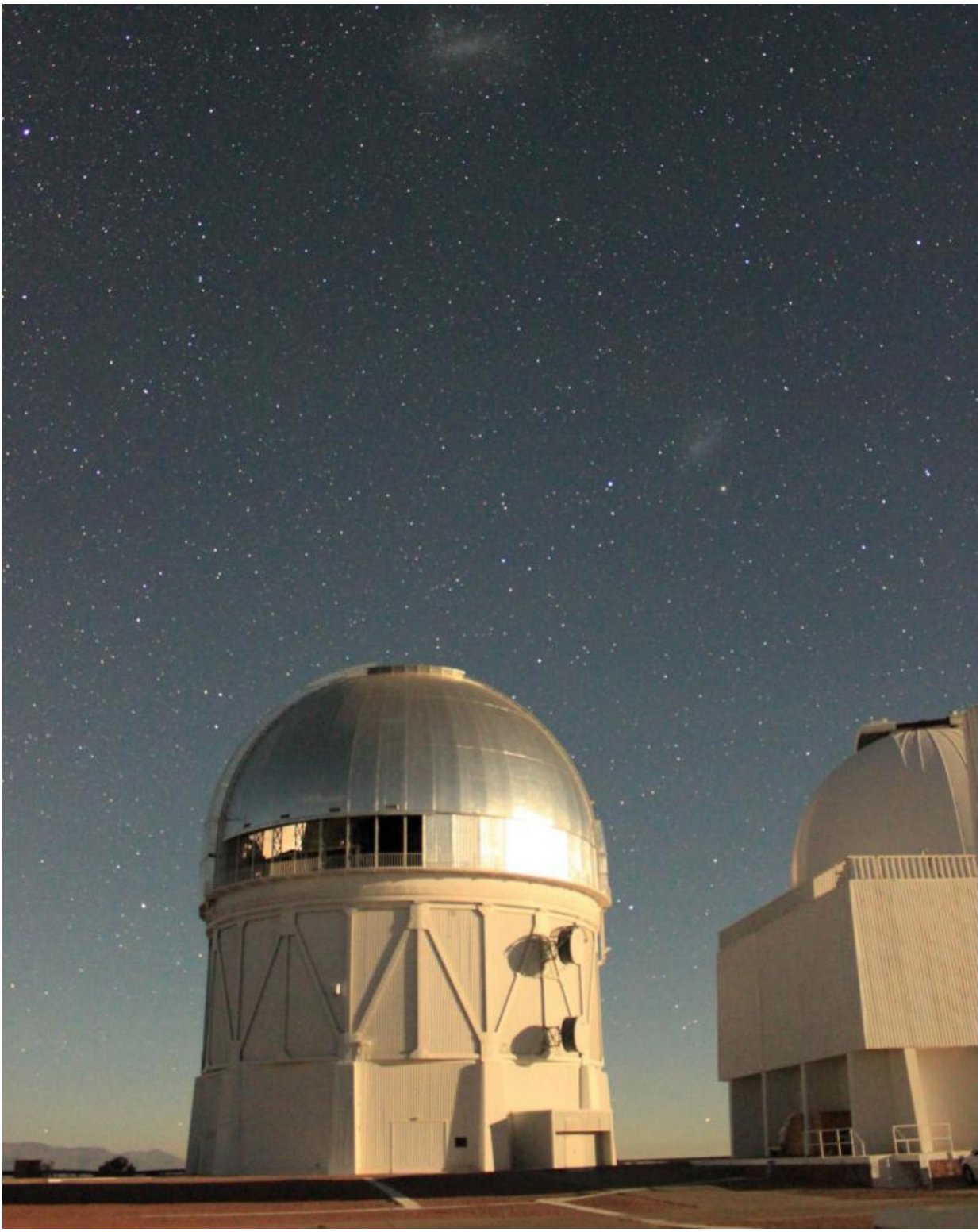
Celeste: A new model for cataloging the universe

September 9 2015



The Victor M. Blanco Telescope at the Cerro Tololo Inter-American Observatory in Chile, where the Dark Energy Camera is being used to collect

image data for the DECam Legacy Survey. The glint off the dome is moonlight; the small and large Magellanic clouds can be seen in the background. Credit: Dustin Lang, University of Toronto

The roots of tradition run deep in astronomy. From Galileo and Copernicus to Hubble and Hawking, scientists and philosophers have been pondering the mysteries of the universe for centuries, scanning the sky with methods and models that, for the most part, haven't changed much until the last two decades.

Now a Berkeley Lab-based research collaboration of astrophysicists, statisticians and computer scientists is looking to shake things up with Celeste, a new statistical analysis model designed to enhance one of modern astronomy's most time-tested tools: sky surveys.

A central component of an astronomer's daily activities, surveys are used to map and catalog regions of the sky, fuel statistical studies of large numbers of objects and enable interesting or rare objects to be studied in greater detail. But the ways in which image datasets from these surveys are analyzed today remains stuck in, well, the Dark Ages.

"There are very traditional approaches to doing astronomical surveys that date back to the photographic plate," said David Schlegel, an astrophysicist at Lawrence Berkeley National Laboratory and principal investigator on the Baryon Oscillation Spectroscopic Survey (BOSS, part of SDSS) and co-PI on the DECam Legacy Survey (DECaLS). "A lot of the terminology dates back to that as well. For example, we still talk about having a plate and comparing plates, when obviously we've moved way beyond that."

Surprisingly, the first electronic survey—the Sloan Digital Sky Survey

(SDSS)—only began capturing data in 1998. And while today there are multiple surveys and high-resolution instrumentation operating 24/7 worldwide and collecting hundreds of terabytes of image data annually, the ability of scientists from multiple facilities to easily access and share this data remains elusive. In addition, practices originating a hundred years ago or more continue to proliferate in astronomy—from the habit of approaching each survey image analysis as though it were the first time they've looked at the sky to antiquated terminology such as "magnitude system" and "sexagesimal" that can leave potential collaborators outside of astronomy scratching their heads.

It's conventions like these in a field he loves that frustrate Schlegel.

"There's a history of how the data are used in astronomy, and the language and terminology reflect a lot of the problems," he said. "For example, the magnitude system—it is not some linear system of how bright objects are, it is an arbitrary label dating back thousands of years. But you can still pick up any astronomy paper and they all use the magnitude system."

When it comes to analyzing image data from sky surveys, Schlegel is certain existing methods can be improved upon as well—especially in light of the more complex computational challenges expected to emerge from next-generation surveys like DECaLS and higher-resolution instruments like the Large Synoptic Survey Telescope (LSST).

"The way we deal with data analysis in astronomy is through 'data reduction,'" he said. "You take an image, apply a detection algorithm to it, take some measurements and then make a catalog of the objects in that image. Then you take another image of the same part of the sky and you say 'Oh, let me pretend I don't know what's going on here, so I'll start by identifying objects, taking measurements of those objects and then make a catalog of those objects.' And this is done independently for

each image. So you keep stepping further and further down into these data reduction catalogs and never going back to the original image."

A Hierarchical Model

These challenges prompted Schlegel to team up with Berkeley Lab's MANTISSA (Massive Acceleration of New Technologies in Science with Scalable Algorithms) project, led by Prabhat from the National Energy Research and Scientific Computing Center (NERSC), a DOE Office of Science User Facility. "To tackle this grand challenge, we have engaged leading researchers from UC Berkeley, Harvard, Carnegie Mellon and Adobe Research," said Prabhat.

The team spent the past year developing Celeste, a hierarchical model designed to catalog stars, galaxies and other light sources in the universe visible through the next generation of telescopes, explained Jeff Regier, a Ph.D. student in the UC Berkeley Department of Statistics and lead author on a paper outlining Celeste presented in July at the 32nd International Conference on Machine Learning. It will also enable astronomers to identify promising galaxies for spectrograph targeting, define galaxies they may want to explore further and help them better understand Dark Energy and the geometry of the universe, he added.

"What we want to change here in a fundamental way is the way astronomers use these data," Schlegel said. "Celeste will be a much better model for identifying the astrophysical sources in the sky and the calibration parameters of each telescope. We will be able to mathematically define what we are solving, which is very different from the traditional approach, where it is this set of heuristics and you get this catalog of objects, then you try to ask the question: mathematically what was the problem I just solved?"

In addition, Celeste has the potential to significantly reduce the time and

effort that astronomers currently spend working with image data, Schlegel emphasized. "Ten to 15 years ago, you'd get an image of the sky and you didn't even know exactly where you were pointed on the sky. So the first thing you'd do is pull it up on the computer and click around on stars and try to identify them to figure out exactly where you were. And you would do that by hand for every single image."

Applied Statistics

To alter this scenario, Celeste uses analytical techniques common in machine learning and applied statistics but not so much in astronomy. The model is fashioned on a code called the Tractor, developed by Dustin Lang while he was a post-doctoral fellow at Princeton University.

"Most astronomical image analysis methods look at a bunch of pixels and run a simple algorithm that basically does arithmetic on the pixel values," said Lang, formerly a post-doc in cosmology at Carnegie Mellon and now a research associate at the University of Toronto and a member of the Celeste team. "But with the Tractor, instead of running fairly simple recipes on pixel values, we create a full, descriptive model that we can compare to actual images and then adjust the model so that its claims of what a particular star actually looks like match the observations. It makes more explicit statements about what objects exist and predictions of what those objects will look like in the data."

The Celeste project takes this concept a few steps further, implementing statistical inference to build a fully generative model to mathematically locate and characterize light sources in the sky. Statistical models typically start from the data and look backwards to determine what led to the data, explained Jon McAuliffe, a professor of statistics at UC Berkeley and another member of the Celeste team. But in astronomy, image data analysis typically begins with what isn't known: the locations and characteristics of objects in the sky.

"In science what we do a lot is take something that is hard and try to decompose it into simpler parts and then put the parts back together," McAuliffe said. "That's what is going on in the hierarchical model. The tricky part is, there are these assumed or imagined quantities and we have to reason about them even though we didn't get to observe them. This is where statistical inference comes in. Our job is to start from the pixel intensities in the images and work backwards to where the light sources were and what their characteristics were."

So far the group has used Celeste to analyze pieces of SDSS images, whole SDSS images and sets of SDSS images on NERSC's Edison supercomputer, McAuliffe said. These initial runs have helped them refine and improve the model and validate its ability to exceed the performance of current state-of-the-art methods for locating celestial bodies and measuring their colors.

"The ultimate goal is to take all of the photometric data generated up to now and that is going to be generated on an ongoing basis and run a single job and keep running it over time and continually refine this comprehensive catalog," he said..

The first major milestone will be to run an analysis of the entire SDSS dataset all at once at NERSC. The researchers will then begin adding other datasets and begin building the catalog—which, like the SDSS data, will likely be housed on a science gateway at NERSC. In all, the Celeste team expects the catalog to collect and process some 500 terabytes of data, or about 1 trillion pixels.

"To the best of my knowledge, this is the largest graphical model problem in science that actually requires a supercomputing platform for running the inference algorithms," Prabhat said. "The core methods being developed by Jon McAuliffe, Jeff Regier and Ryan Giordano (UC Berkeley), Matt Hoffman (Adobe Research) and Ryan Adams and Andy

Miller (Harvard) are absolutely key for attempting a problem at this scale."

The next iteration of Celeste will include quasars, which have a distinct spectral signature that makes them more difficult to distinguish from other light sources. The modeling of quasars is important to improving our understanding of the early universe, but it presents a big challenge: the most important objects are those that are far away, but distant objects are the ones for which we have the weakest signal. Andrew Miller of Harvard University is currently working on this addition to the model, which couples high-fidelity spectral measurements with survey data to improve our estimates of remote quasars.

"It may be a little surprising that up to now the worldwide astronomy community hasn't built a single reference catalog of all the light sources that are being imaged by many, many different telescopes worldwide over the past 15 years," McAuliffe said. "But we think we can help with that. This is going to be a catalog that will be incredibly valuable for astronomers and cosmologists in the future."

Provided by Lawrence Berkeley National Laboratory

Citation: Celeste: A new model for cataloging the universe (2015, September 9) retrieved 27 April 2024 from https://phys.org/news/2015-09-celeste-universe.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.