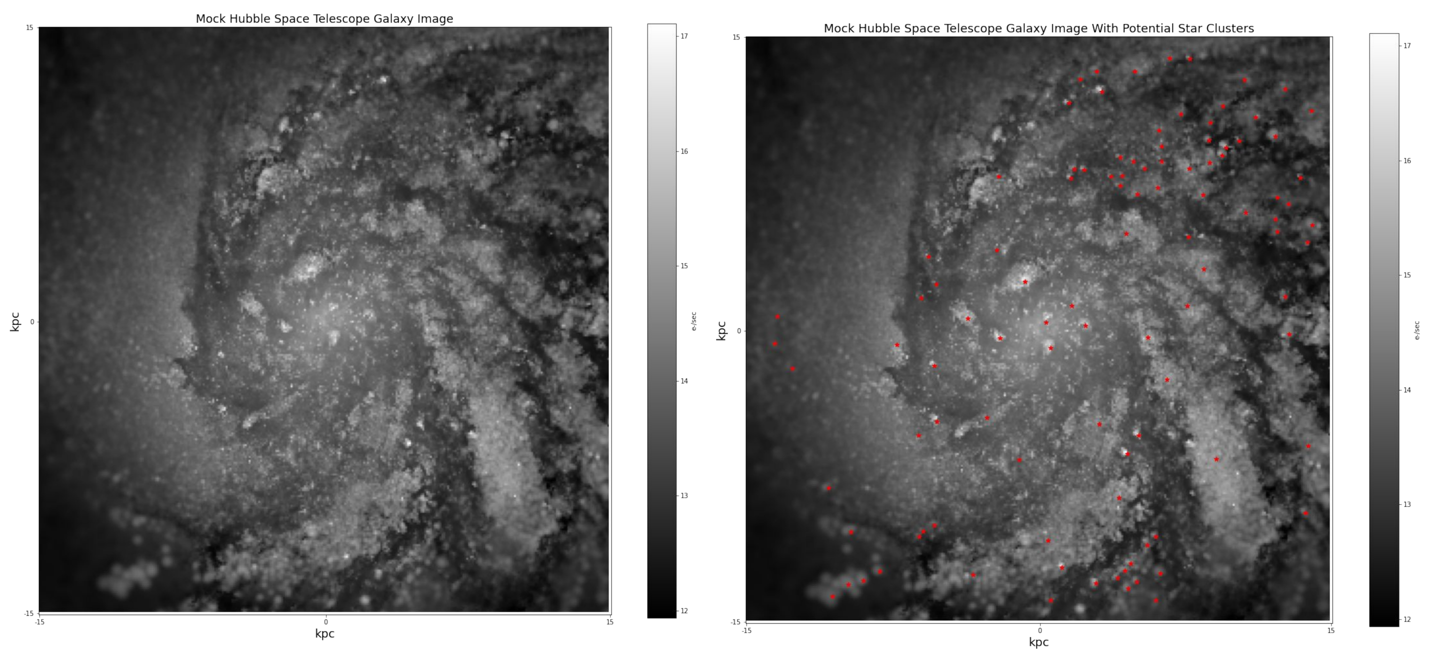
Star Clusters on FIRE: Using Machine Learning to Classify Star Clusters in Synthetic Galaxy Images

Jonah Otto[1 2 3], Sarah Loebman[2], Matt Orr[4], Josh Peek[5], Andrew Wetzel[6], Binod Bhattarai[2], Dylan Benton[2], Micah Oeur[2], Erick Villegas[2]

[1]California State University, Los Angeles, [2]University of California, Merced, [3]Texas Christian University, [4]Rutgers University and the Center for Computational Astrophysics, [5]Space Telecope Science Institute and Johns Hopkins University, [6]University of California, Davis

What are Star Clusters and why do we care?

A star cluster is a collection of stars that formed in the same gas cloud and are the same age. The majority of stars form in clusters, including the Sun. Being able to determine how and when the Sun left its birth cluster could lead to the discovery of its sister stars. These stars that formed in the same environment could provide valuable insight into our solar system's history. More broadly, star clusters are common in star forming galaxies and are the building blocks of galactic disks, making them crucial to understanding how galaxies form and evolve over time.



The FIRE Simulations

Since many stars live for billions of years, observing a single star cluster over its whole lifetime is problematic. So instead, astronomers have turned to using computers to simulate how a galaxy would form and evolve over the life of the universe. We specifically used the FIRE simulations in this work. The FIRE simulations are a collection of galaxy simulations that model how a galaxy evolves from the very early universe to present day. We produced *Hubble* style images of the simulated galaxy (above left). Then, software was used to detect where potential star clusters might be in images (above right). We focused on a single Milky Way-mass galaxy, but FIRE has a wide variety of galaxies.

Sample of Simulated FIRE Galaxies

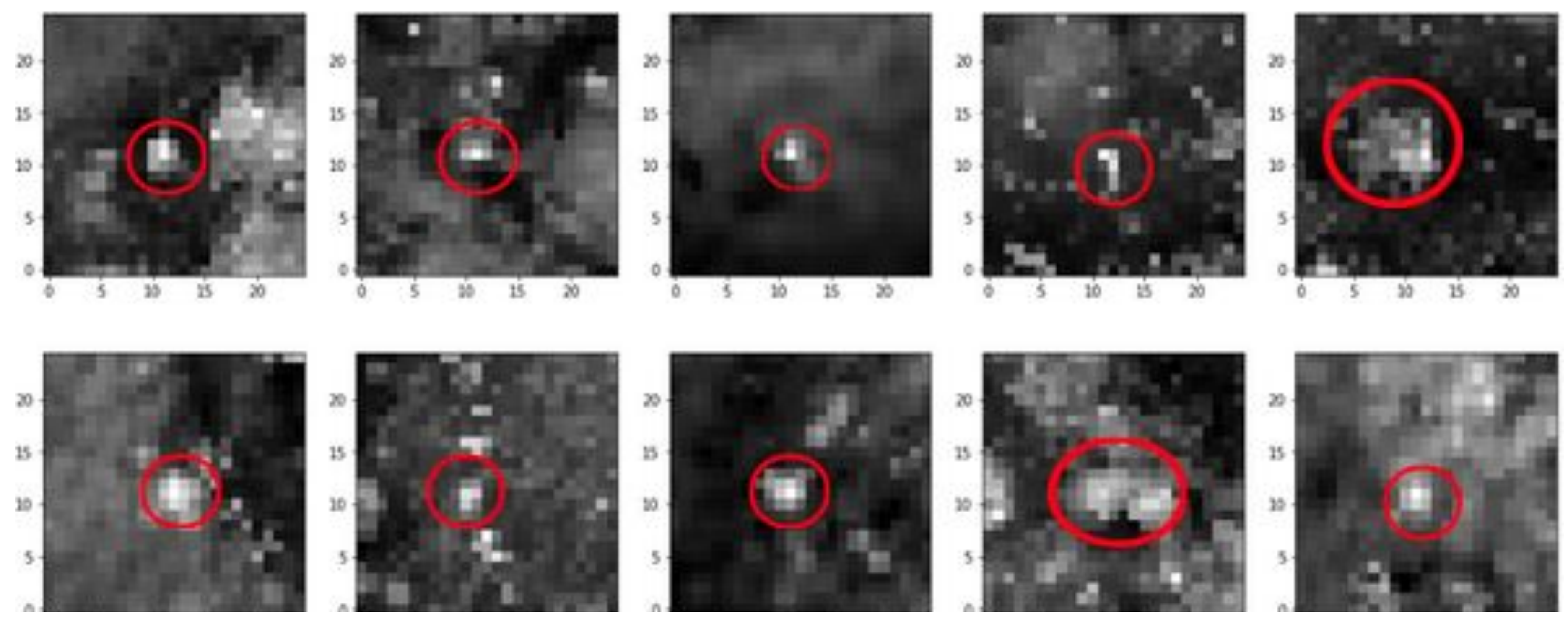
Above: Some images of FIRE galaxies, simulated from purely cosmological initial conditions (laid on a star field purely for visual effect here). (FIRE homepage)

Star clusters are groups of stars that are born in the same place at the same time. These structures are important to understand because most stars are born in a cluster, and clusters form the basic building blocks of galaxies. Watching a cluster form and then break apart would help increase this understanding. However, since many stars live for billions of years, this is impossible to observe. Instead, we use large simulations of galaxies that show many star clusters throughout their whole life cycle. We automated a way to identify and classify potential star clusters to help further our understanding.

The Simulated Star Cluster Catalog

The final product of this project was a catalog of all the star clusters found in the simulated galaxy by our machine learning program. The program works thousands of times faster than humans can and at a similar level of accuracy, classifying over 35,000 potential star clusters in a manner of minutes, while it took our group over 4 hours to do a little over 300 images. Below is a random sample of 10 images that our program said were star clusters. How did we do?

Machine Learning Identified Star Clusters



Machine Learning Program Performance

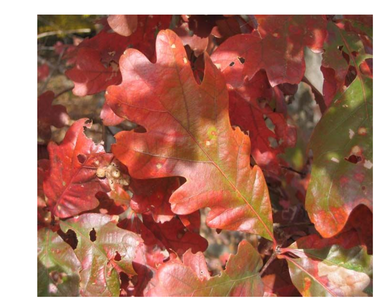
Class 1	Class 2	Class 3	Class 4
2658	10967	5401	16322
1961	4488	9019	19880
3279	4898	9703	20468
1986	7113	8209	18040
1754	6501	12453	14640

The table above shows how our machine learning program classified the ~35,000 running the program from start to finish 5 different times. It's easy to see there is quite a bit of variation in the numbers. Some of this is due to how machine learning programs work, but there our methods also bring in some variation. Due to the small amount of labeled images and the large amount of disagreement of those labels, the program only had 15 class 1-star clusters and 19 class 2-star clusters that we were reasonably confident about to train on. Overall, less than half of our labels were 'strong', meaning at least 4 out of 5 people were in agreement. The rest were in the controversial category.

Machine Learning - Automate Complex Tasks

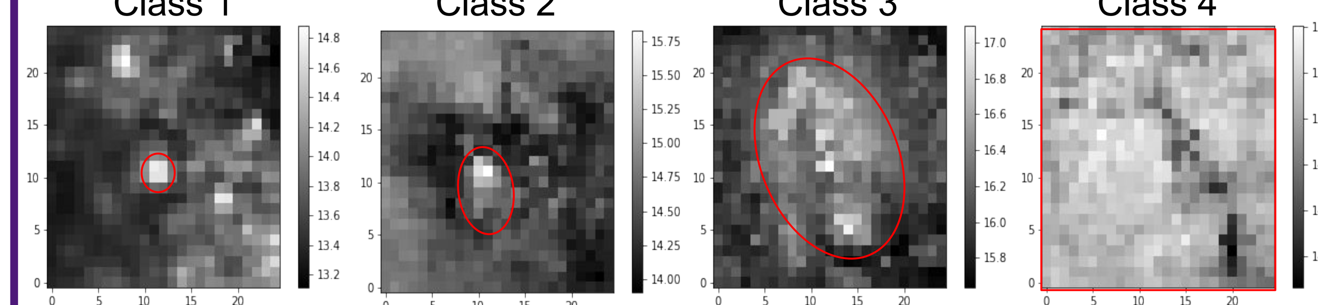
At its most basic, machine learning is a computer program that can improve itself by performing the same task over and over. This is important in astrophysics because huge amounts of data are being generated through both simulations and observations, so much that it would take people way too long to look through it all, so we use computers to do it for us. We used a particular machine learning program called a neural network. Neural networks are excellent at classification problems, such as determining what type of leaf you have. The idea is you know how to tell leaves apart by how they look. If you can break that down into characteristics such as length, width, etc, you can teach a computer to do it much faster than a human can. In our case, we applied this idea to whether an object in our simulated images was a star cluster or not.

Feature Vector:
Length: 10cm
Width: 8cm
# of Points: 21
Sharp Points: yes
Smooth Edges: no



Training Set Labels

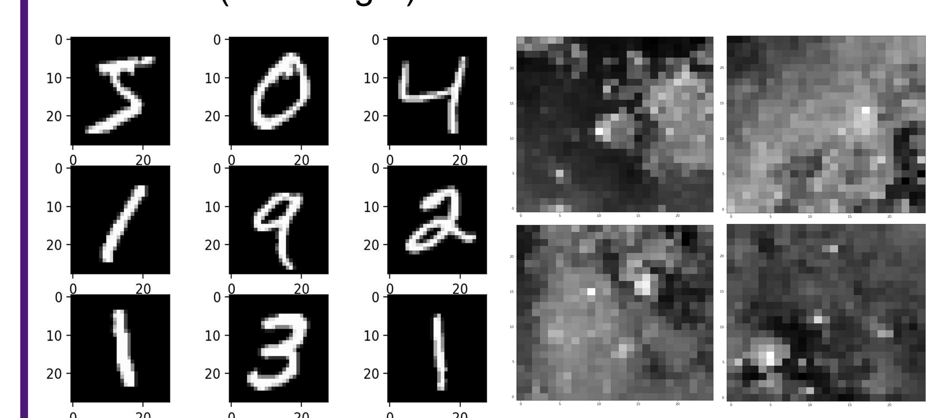
Every classification problem that uses machine learning needs a dataset, in our case a set of potential star cluster images, that is labeled. For the star cluster problem this could be as simple as labeling each image as either 'star cluster' or 'not a star cluster'. We ended up using a different system that labels star clusters based on their shape, but the important part is classes 1 and 2 are star clusters and classes 3 and 4 are not star clusters. Below are examples of each of the 4 classes of star clusters.



A group of 5 people independently looked at ~300 of this small images which are cut out from the larger galaxy around a potential cluster and decided whether it was a star cluster or not. This became the dataset used to train our machine learning program.

This is a difficult classification problem

When creating labels for a machine learning program, ideally what the correct answer should be clear to everyone. For example, the MNIST integer dataset (below left) is a group of hand drawn numbers from 0-9. When looking at these numbers, everyone can agree which number each image represents. On the other hand, the 5 people classifying our star cluster images disagreed completely sometimes (below right).



Each of these images were split 3-2 for cluster vs non-cluster.

When we don't know what is our is not a star cluster, the computer program will not perform any better. While our group had the disadvantage of inexperience, even experts struggle. A study showed the same set of images to a group of star cluster experts at different times. The experts only agreed with themselves ~70% of the time. What is and is not a star cluster is not well-defined, particularly when the individual stars within the cluster are not able to be distinguished.