STScI | SPACE TELESCOPE SCIENCE INSTITUTE

EXPANDING THE FRONTIERS OF SPACE ASTRONOMY

# Enabling Data Science at MAST

Josh Peek for MAST

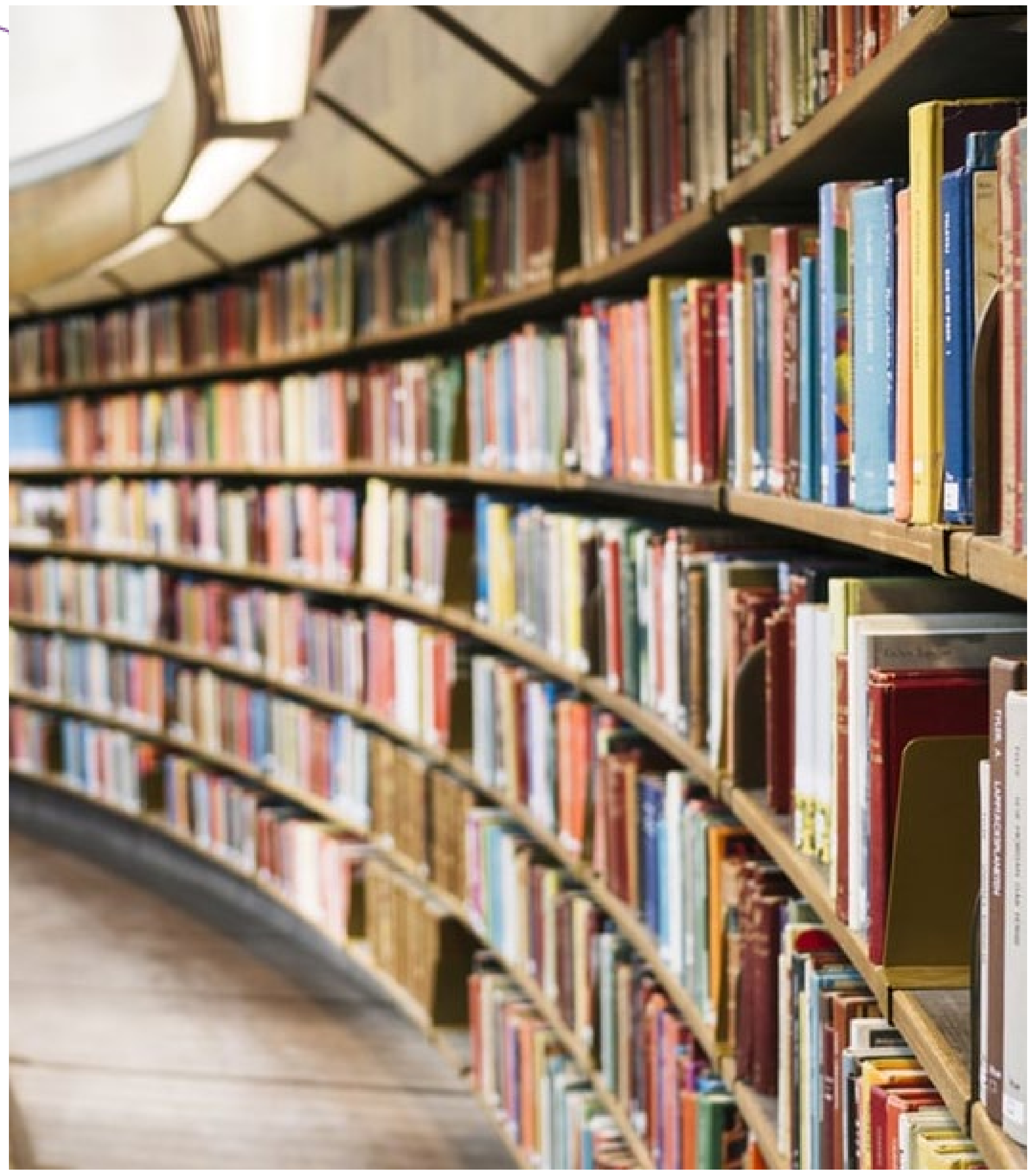
STScI Town Hall

AAS 240, Pasdena

*Data science, including applications of machine learning, will play an increasing role in astronomical research over the coming decade. Incorporating training in this area at the graduate level and beyond will better prepare researchers regardless of whether they pursue careers in astrophysics or in other STEM fields.*

**Pathways to Discovery in Astronomy and Astrophysics for the 2020s**

# MAST as a library: a place for data, tools, teaching, & gathering




TIME SERIES

INTEGRATED

KNOWLEDGE

ENGINE

TIME SERIES

INTEGRATED

KNOWLEDGE

ENGINE

# The TIKE is live now!

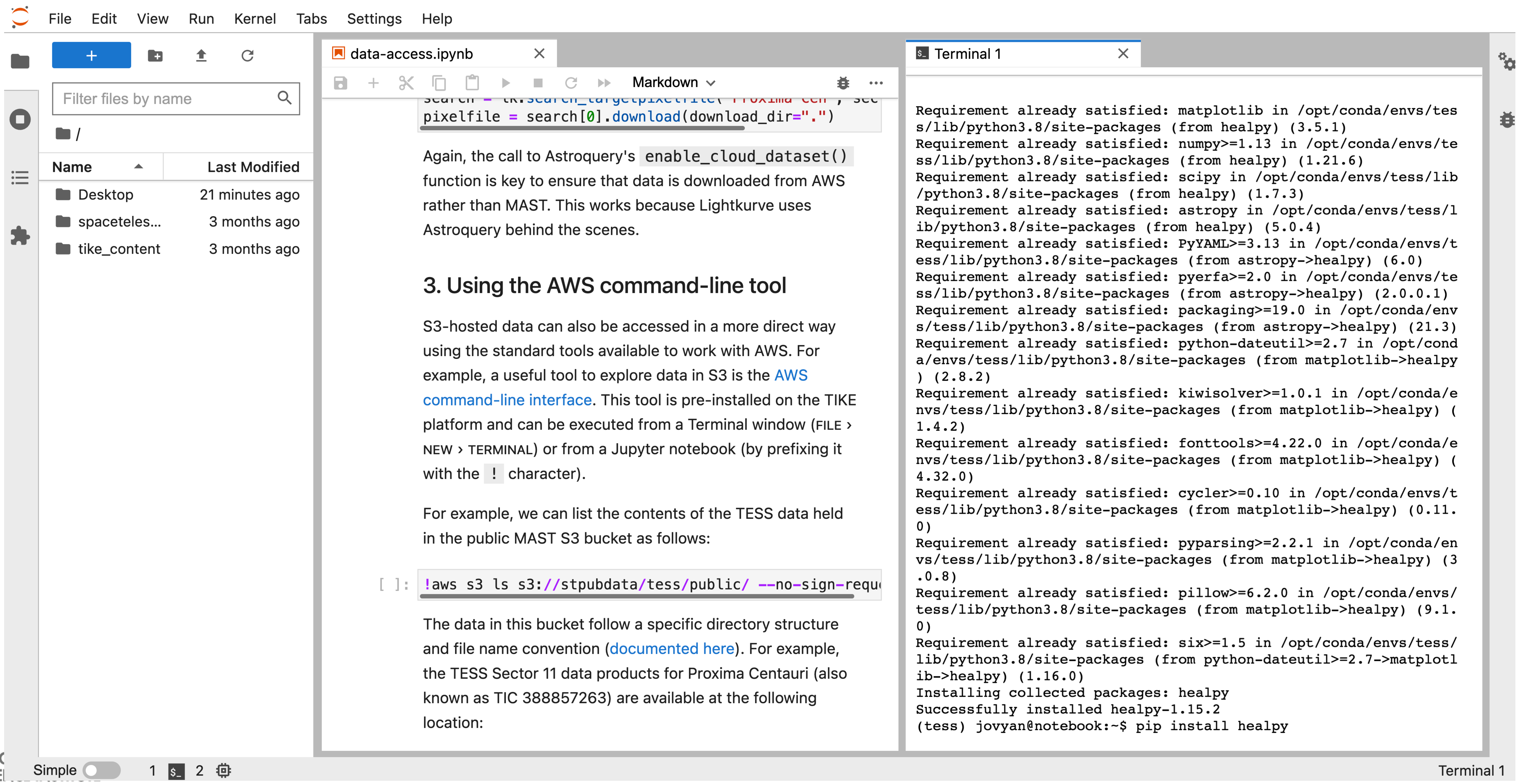
https://timeseries.science.stsci.edu

A JupyterHub platform for exploring TESS, Kepler & K2 timeseries data

Provides tools for accessing MAST AWS Open Data

Full of notebooks and software for this kind of exploration

# The TIKE is a full, free-to-use software environment for everyone



NUMPY, SCIPY, MATPLOTLIB, PANDAS...

ASTROPY, ASTROQUERY, PYVO...

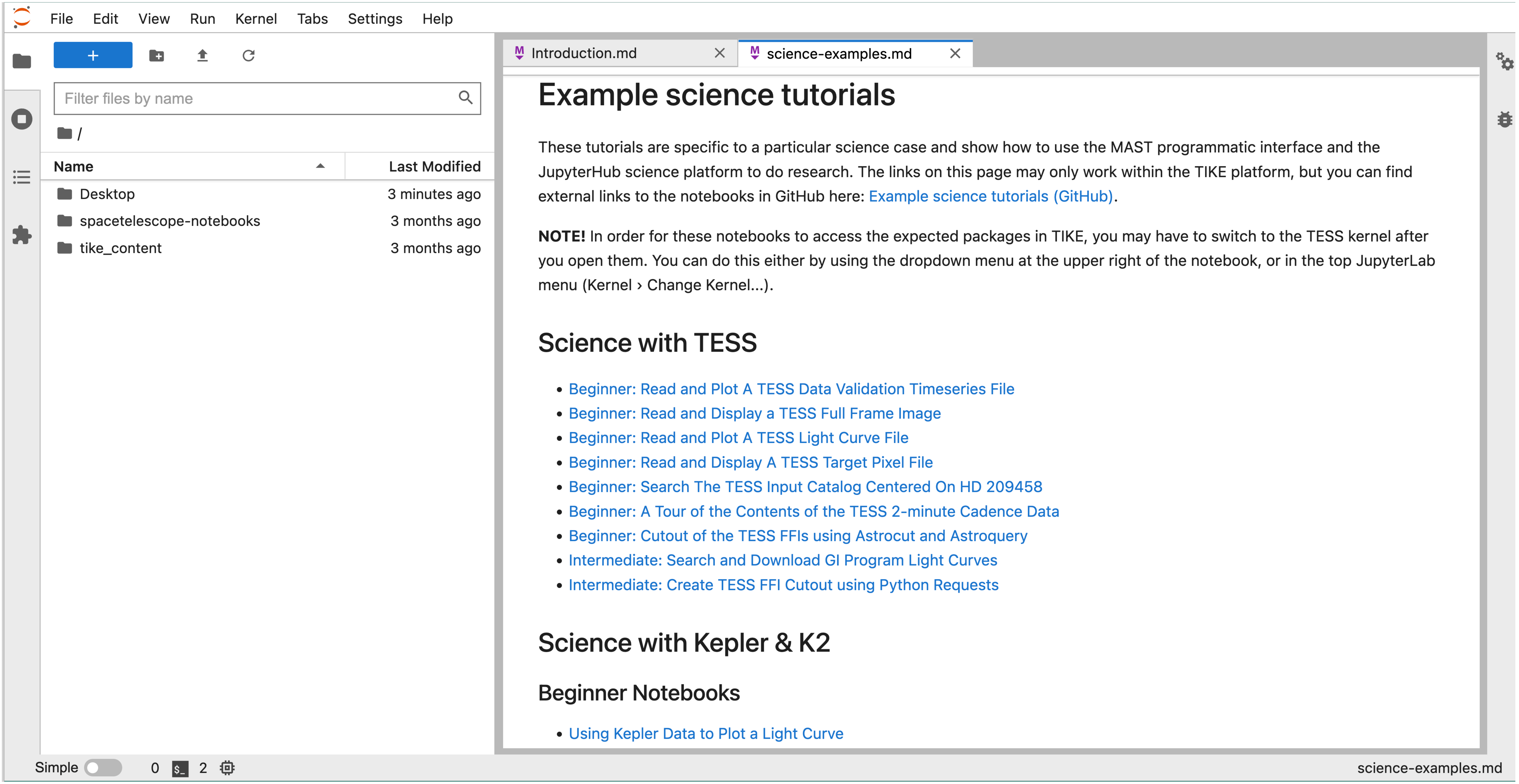
LIGHTKURVE, ASTROCUT, EVEREST...

EMCEE, GEORGE, CELERITE...

TENSORFLOW, SCIKIT-LEARN...

AWSCLI, BOTO3, S3FS...

# The TIKE is loaded with dozens of science examples for TESS & Kepler

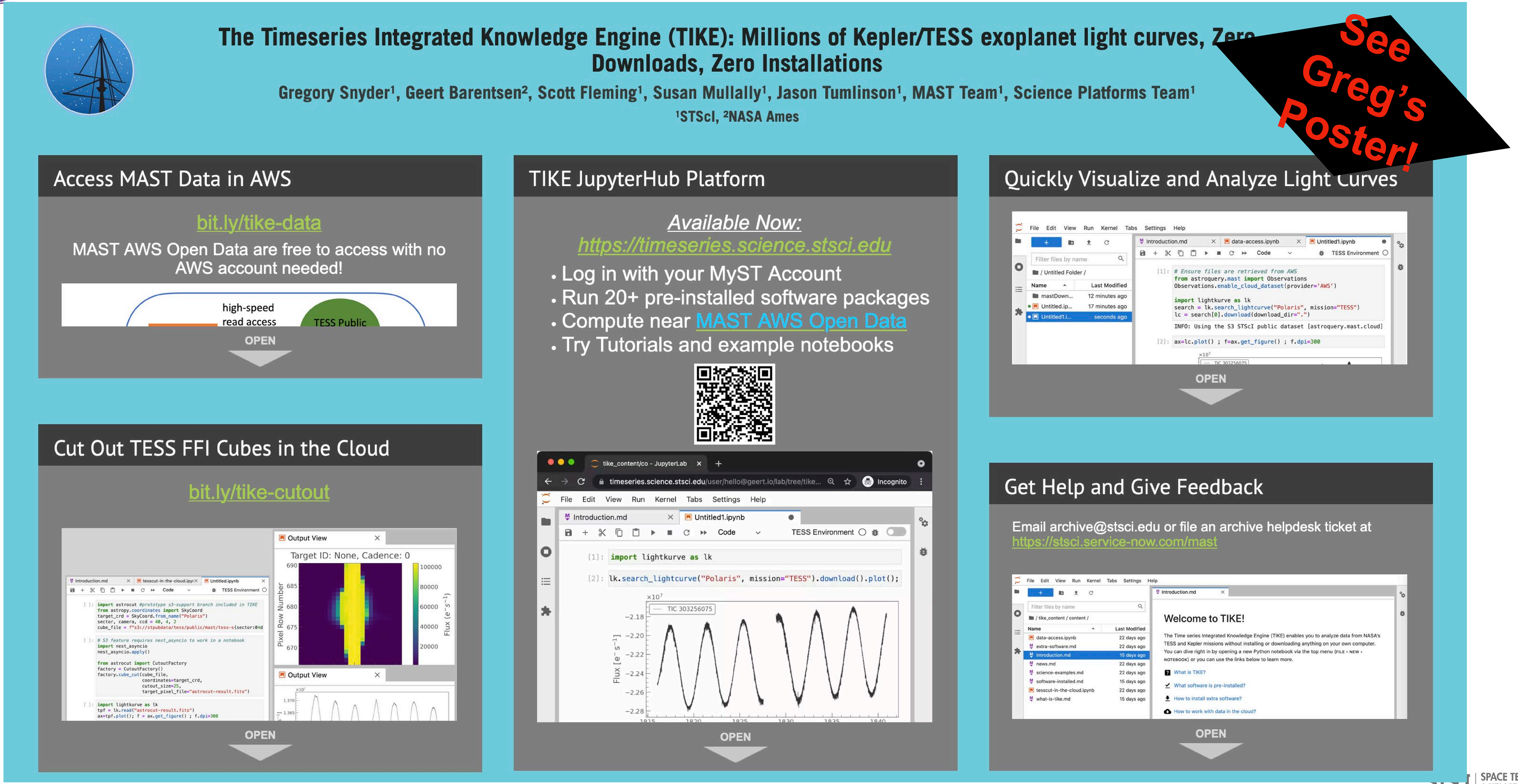


# The TIKE accelerates Big Data analyses with the cloud

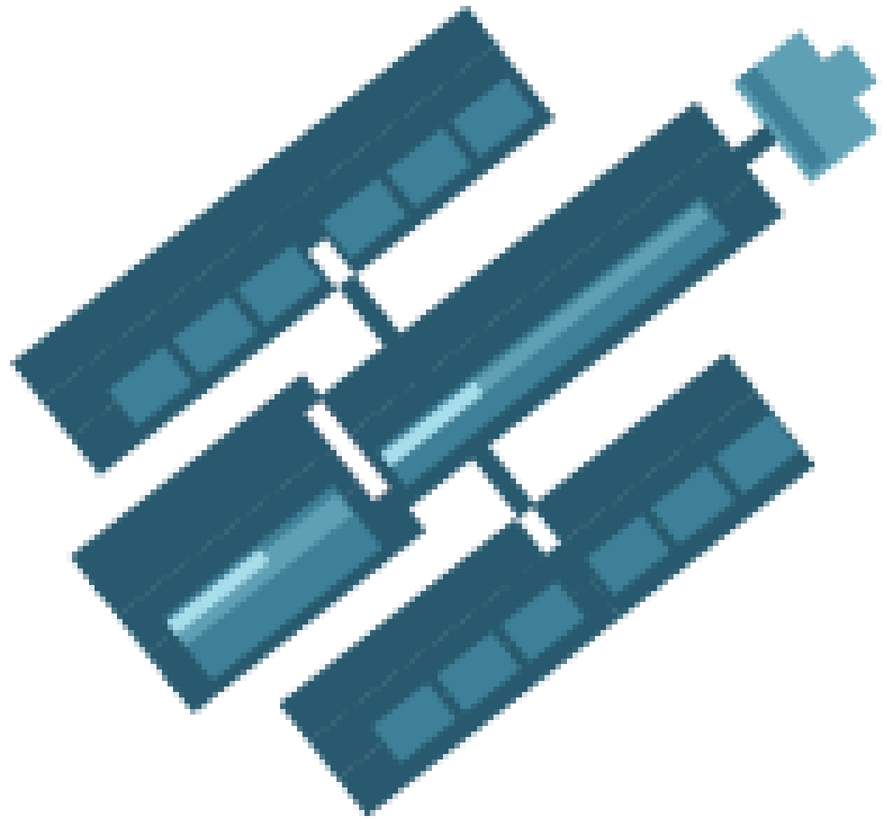
## *Time to retrieve 30 TESS images (FFIs)*

| Where you are | Where you get the data | Time |
|---|---|---|
| Home Wifi | Baltimore | 14 minutes |
| Home Wifi | Cloud | 6 minutes |
| TIKE | Baltimore | 1 minute |
| TIKE | Cloud | **9 seconds** |

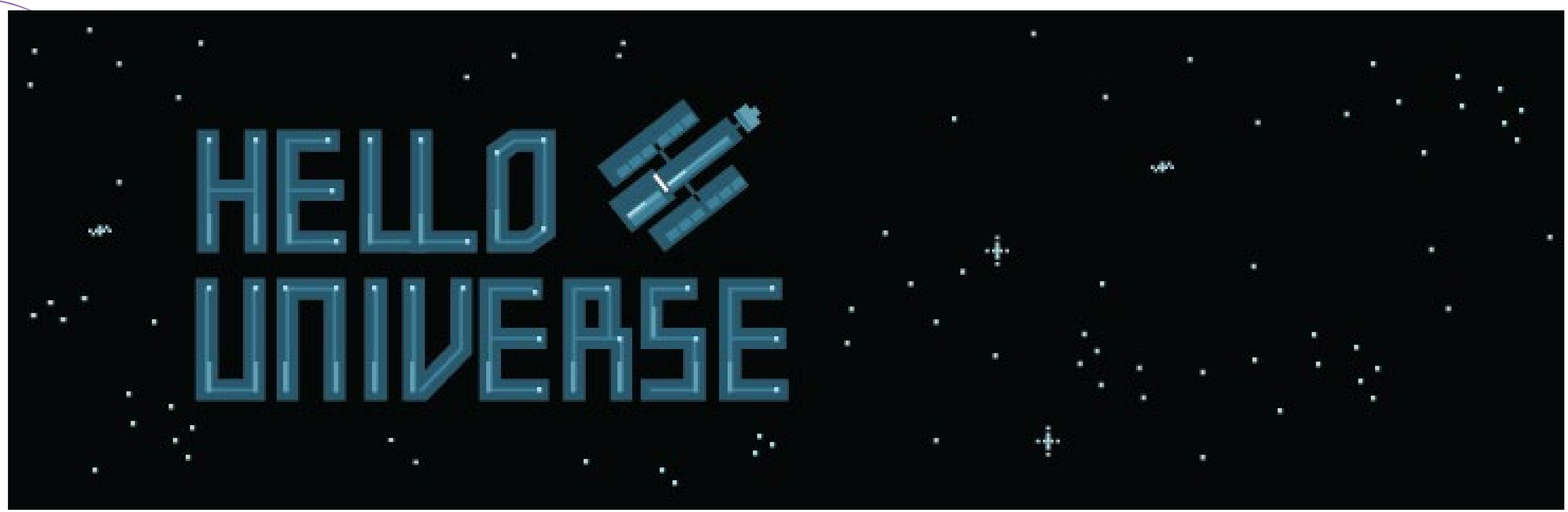
# Learn More about TIKE from the creators



# HELLO UNIVERSE



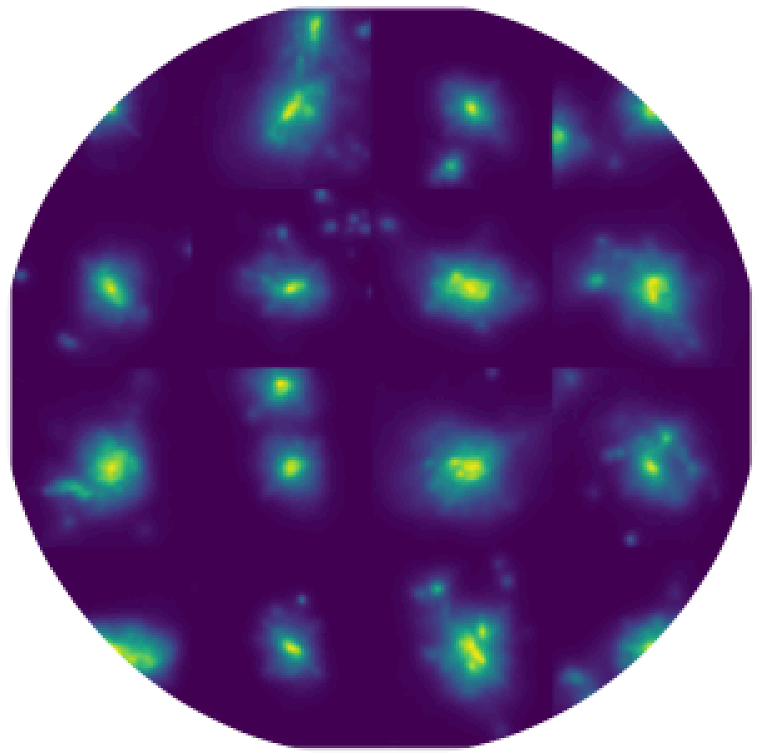
# We developed *Hello Universe* to make ML more accessible in astro




Upcoming large and rich data sets will require astronomers to develop fluency in data science methods. Hello Universe addresses this need by providing:

- Data:reliable **high level science products** (HLSPs) that are appropriate for testing and benchmarking new machine learning algorithms.
- Code: pedagogical, **always working** Jupyter notebooks that provide a step-by-step example of how to apply machine learning (ML) to the data.

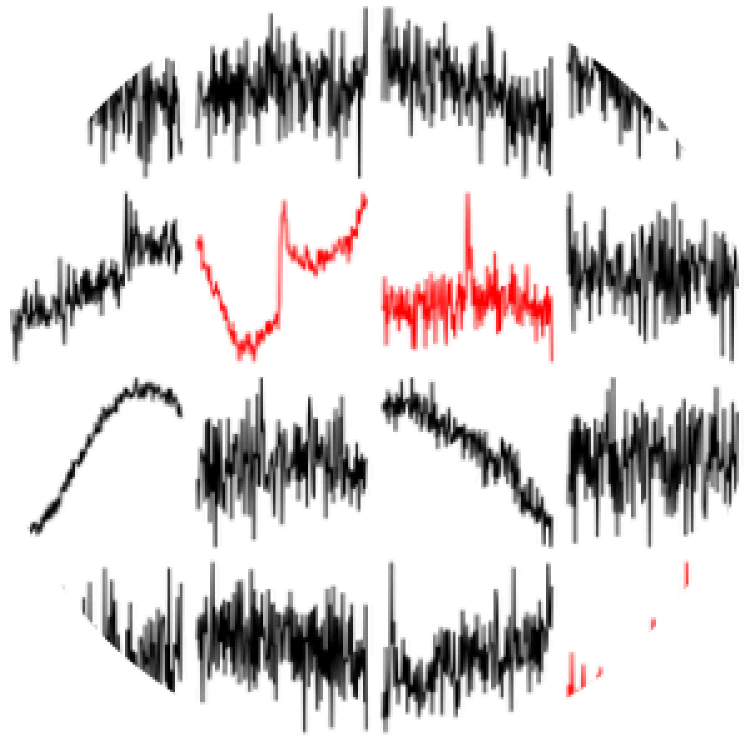
# *Hello Universe* launching today with 4 notebooks + data sets



## Classifying JWST/HST galaxy mergers with CNNs

neural networks | 2d data | classification | overfitting | confusion matrix

Read More



## Classifying TESS stellar flares with CNNs

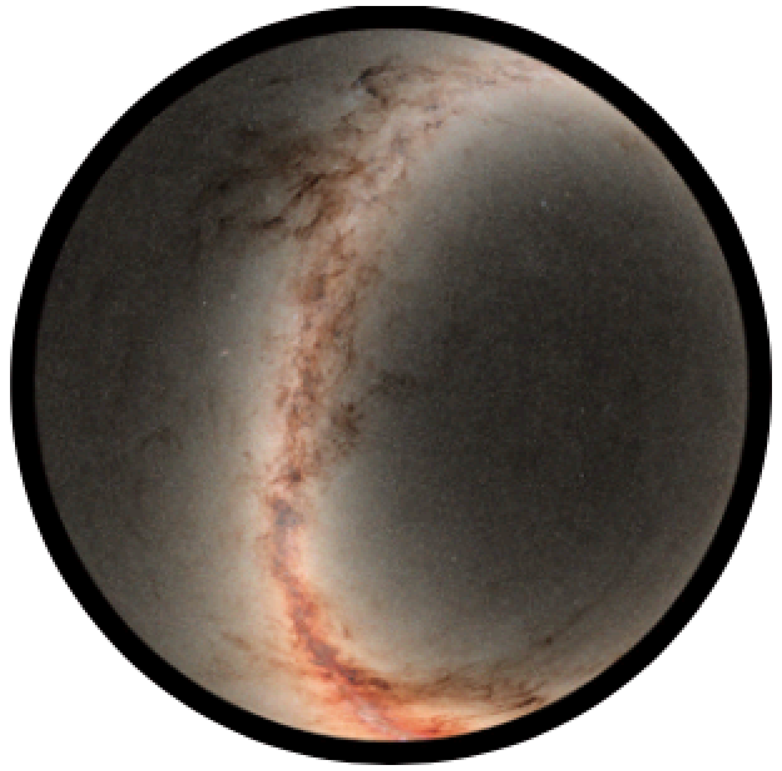
neural networks | 1d data | classification | prediction

Read More

## Predicting 3D-HST redshift with decision trees

decision trees | 1d data | regression | cross-validation

Read More



## Classifying Pan-STARRS with (un)supervised learning

classification | 1d data | PCA | tSNE | k-means | SGD | unsupervised | supervised

Read More

# *Hello Universe* launching today with 4 notebooks + data sets

## Classifying galaxy mergers with JWST/HST and CNNs

This entry explores how to classify merging galaxies vs. non-merging galaxies from multi-wavelength imaging from the James Webb Space Telescope (JWST) and the Hubble Space Telescope (HST) with convolutional neural networks (CNNs). This approach was used by the DeepMerge team (Ciprijanovic et al. 2020) to classify mergers in sythetic observations of simulated galaxies. The same workflow is presented in simplifed format to walk through the construction of the CNN model. The results are then validated and the performance is discussed.

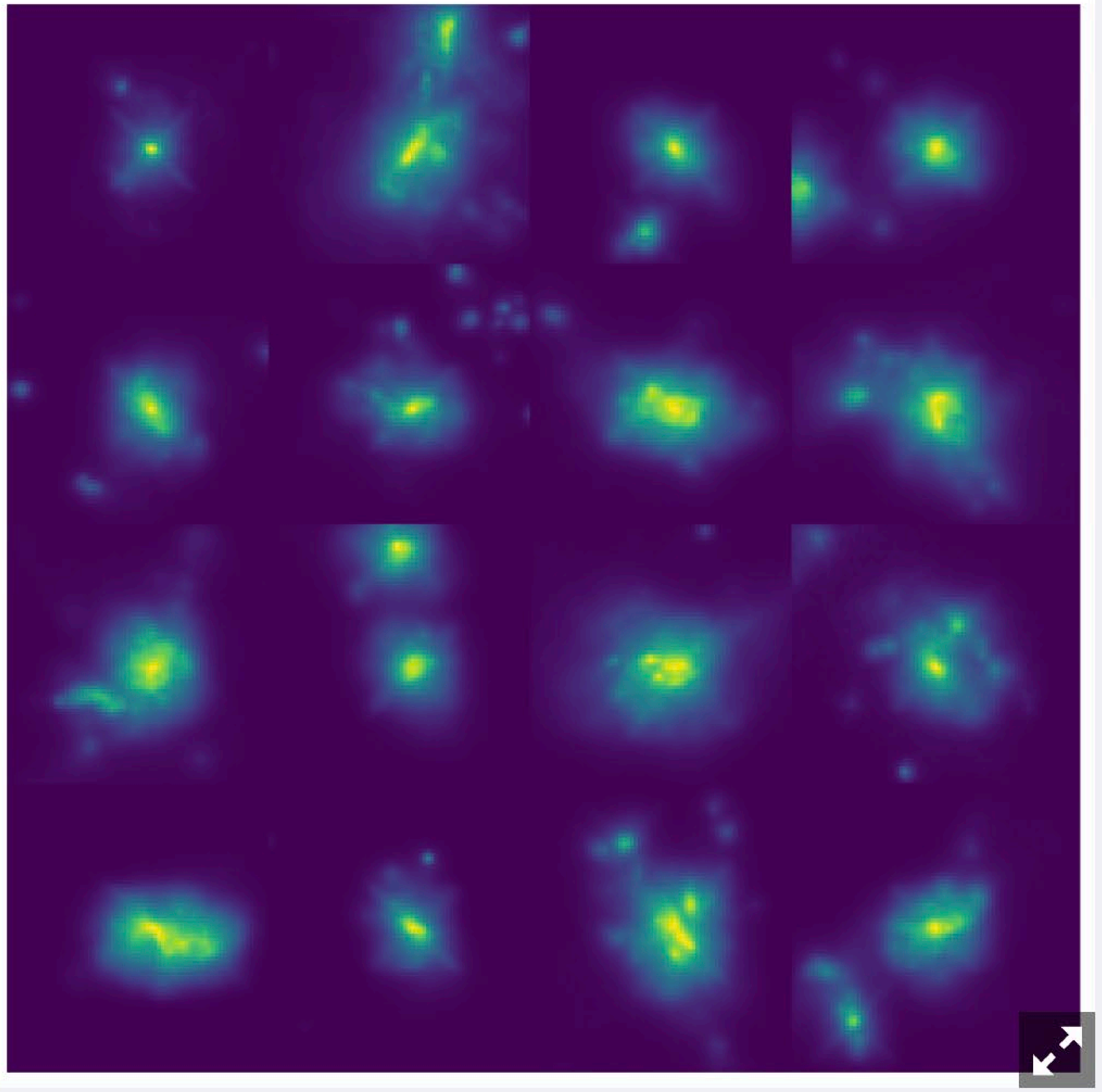
**Data:** The DeepMerge HLSP

**Notebook:** Classifying JWST-HST galaxy mergers with CNNs

**Released:** 2022-06-12

**Updated:** 2022-06-12

**Tags:** classification, deep learning, 1d data



Example simulated galaxy images (JWST NIRCam F356W) from DeepMerge (Ciprijanovic et al. 2020).

# *Hello Universe* is already helping astronomers new to ML

See Jonah's Poster!

## Star Clusters on FIRE: Classifying Star Clusters in Synthetic Galaxy Images using Machine Learning

UCMA

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

[1]California State University, Los Angeles [2]University of California, Merced [3]Rutgers University and the Center for Computational Astrophysics [4]Space Telescope Science Institute and Jo[hns] Hopkins University [5]University of California, Davis

### Producing Synthetic Galaxy Images

0:00 / 0:42

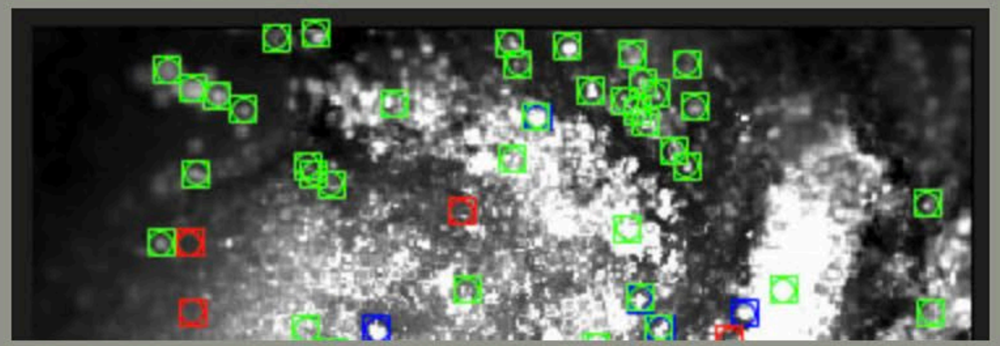
Below is an example synthetic galaxy image, we focused on galaxy m12i of the *Latte* suite in this study.

OPEN

### Source Extractor finds Potential Clusters

0:00 / 1:17

Below is an example of potential star clusters identifed using the Friends of Friends algorithm, Source Extractor (Bertin & Arnouts 1996) and visually.

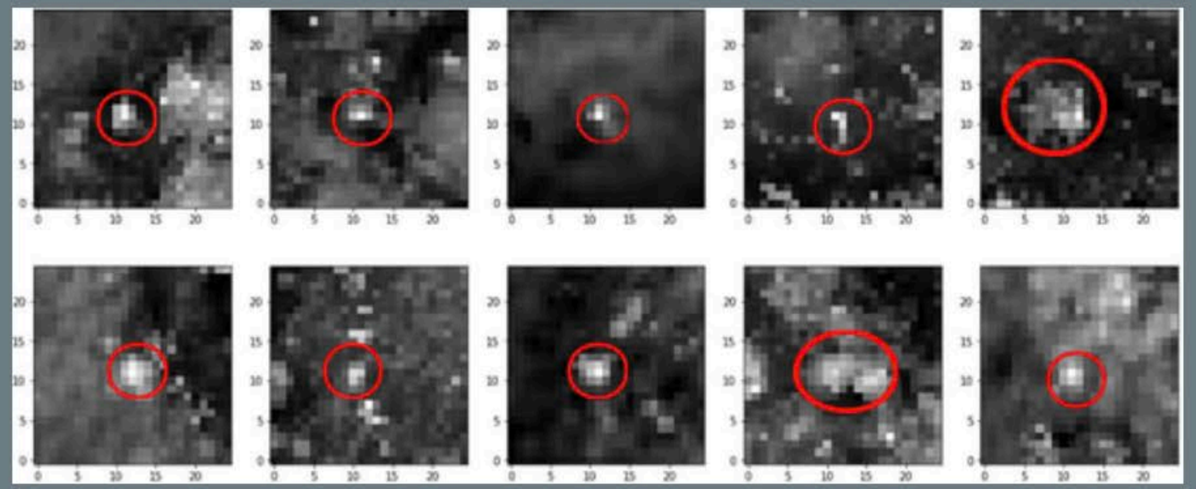


OPEN

### Synthetic Galaxy Star Cluster Catalog

0:00 / 1:20

We use machine learning to successfully classify over 35,000 potential star clusters identified in synthetic galaxy images, thousands of times faster than humans could at a similar level of accuracy.



Above are a random sample of class 1 star clusters identified in m12i

| Class 1 | Class 2 | Class 3 | Class 4 |
|---|---|---|---|
| 2658 | 10967 | 5401 | 16322 |
| 1961 | 4488 | 9019 | 19880 |
| 3279 | 4898 | 6703 | 20468 |
| 1986 | 7113 | 8209 | 18040 |
| 1754 | 6501 | 12453 | 14640 |

The above table shows how our CNN classified the ~35,000 potential star clusters generated from all 555 synthetic images.

OPEN

### Human Generated Training Set Labels

0:00 / 1:26

A CNN trained on human generated labels can only be as accurate as the labels it is given.

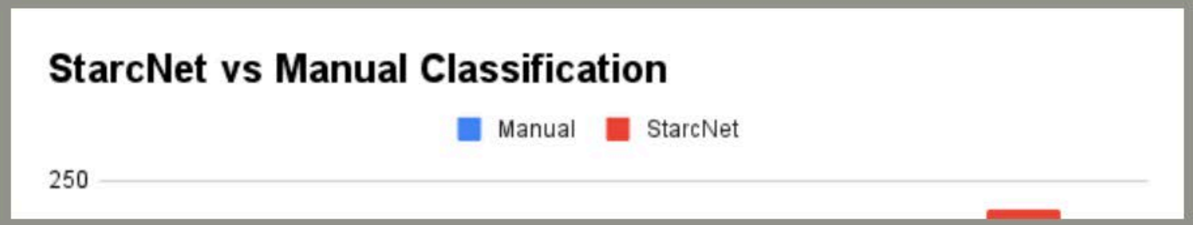
Following Grasha et al. 2015, each potential star cluster indentified by Source Extractor could be labeled as 1 of 4 classes.

OPEN

### Convolutional Neural Network

0:00 / 1:20

We developed our own CNN because StarcNet (Perez et al. 2021), a publicly available CNN architecture, was designed with observational data in mind.




OPEN