Jet Propulsion Laboratory
California Institute of Technology

DSWG Data Science Pilot Program

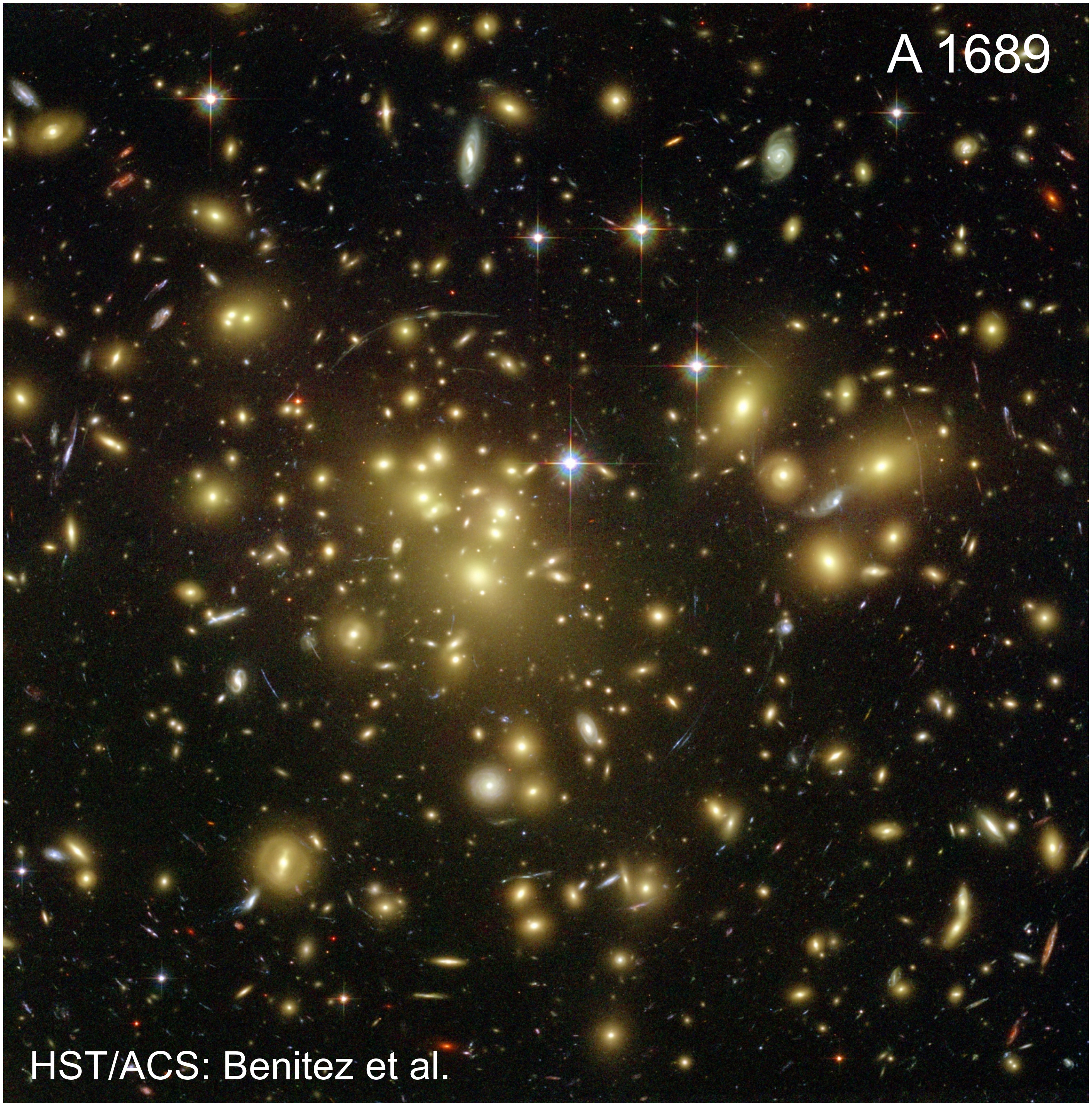
Deep Learning for Cluster Detection

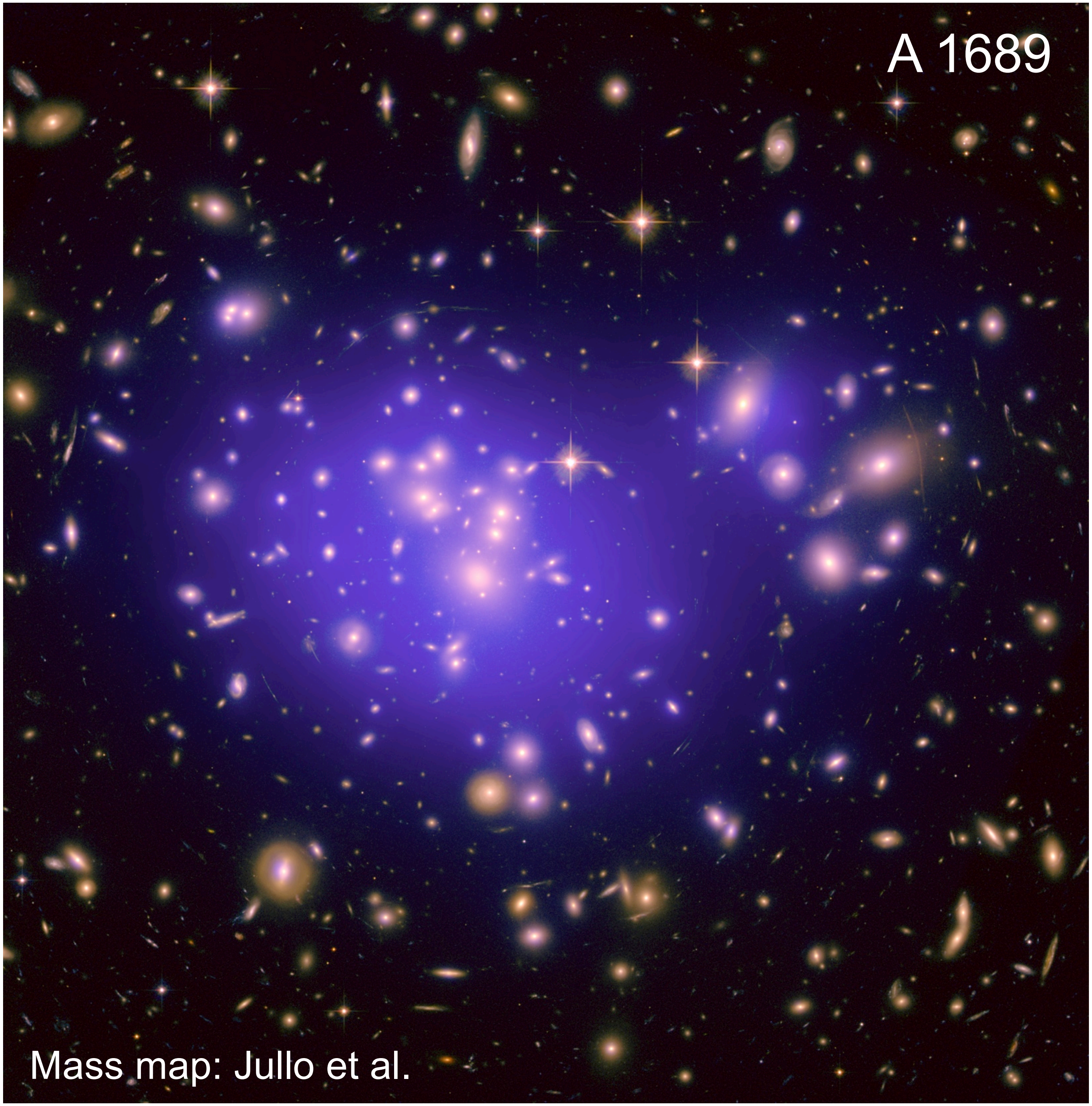
8 November 2018

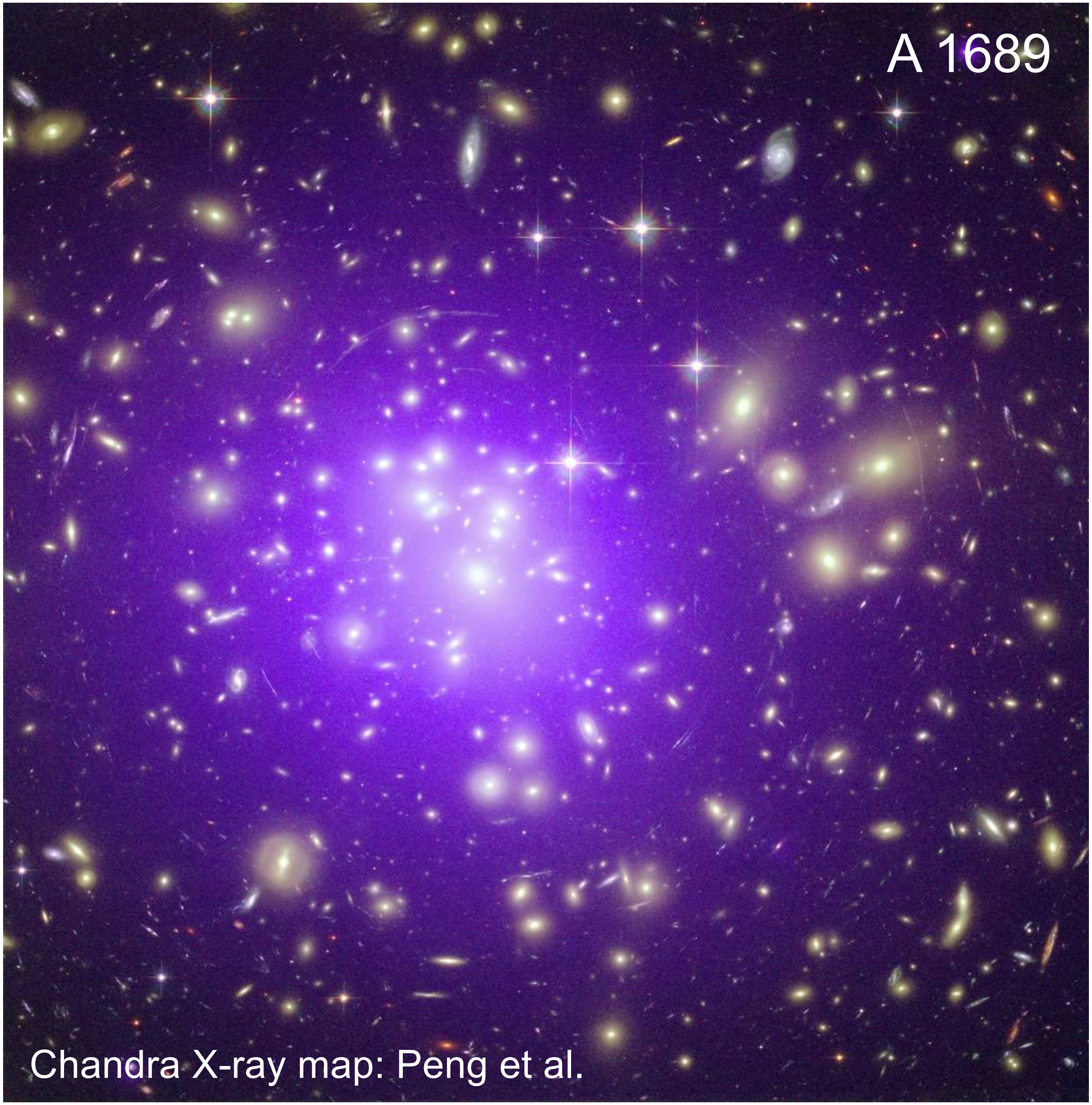
James G. Bartlett (3266)

Overview

- Galaxy clusters are an important tool for astrophysical studies
  - Cluster abundance is a standard cosmological probe; e.g., dark energy
  - Clusters are valuable laboratories for studying structure and galaxy formation; e.g., relation between galaxies, gas and dark matter
- Cluster catalogs are key mission products: *Planck*, *Euclid Spitzer*, ROSAT, WISE, WFIRST…
- Data science program
  - Does machine learning (ML) offer a new and better way to detect galaxy clusters?


A 1689
HST/ACS: Benitez et al.


A 1689
Mass map: Jullo et al.


A 1689
Chandra X-ray map: Peng et al.

Context

- By mass, clusters consist of
  - 80% dark matter
  - 15% hot gas (T~5keV)
  - 5% galaxies
- They are detectable from radio to X-ray bands
- Current cluster detection methods
  - Depend on observation band
  - Rely on simplified models of the cluster signal; e.g., spherical spatial-frequency filters
  - Often use linear filtering or likelihood evaluation of cluster properties, both based on the simple models.

Motivation

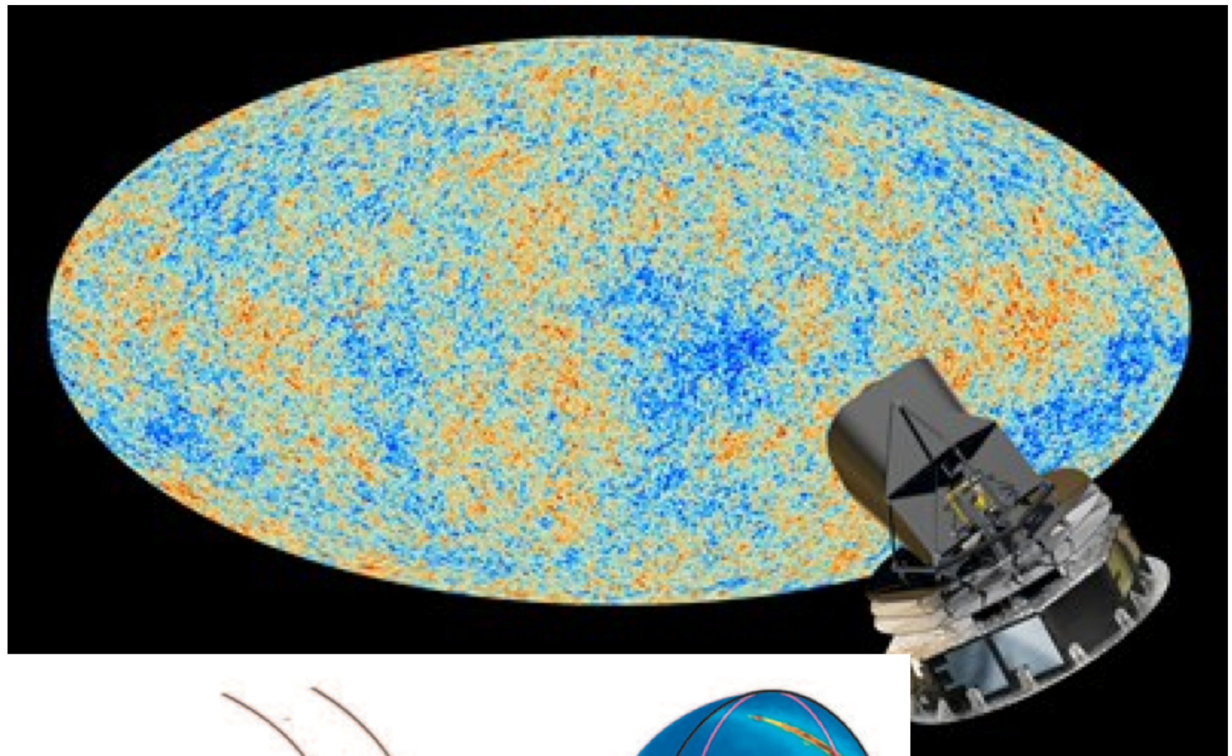
- Clusters are complex, extended objects with multiple components
- ML is capable of developing detection methods based on more complex cluster models through training
- ML can learn to use potentially powerful non-linear methods
- Better modeling could lead to better cluster detection methods
  - Higher catalog completeness and reliability

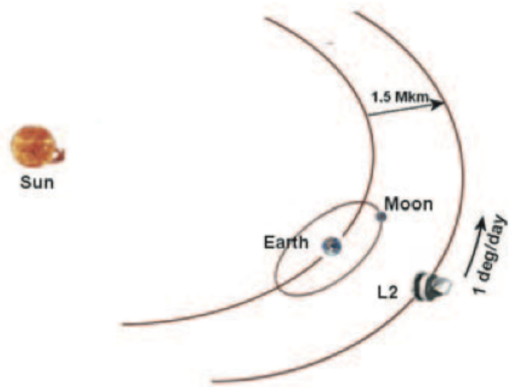
Approach

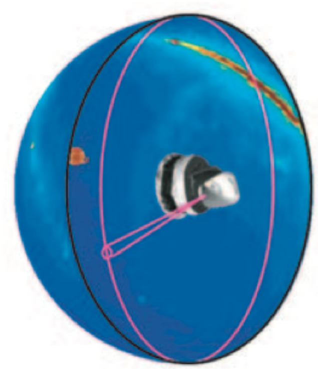
- Development on MacBook using Google's TensorFlow
- Simulated data
  - Start with *Planck* mission
  - Catalog pre-selected using multi-matched filter (MMF, spatial-frequency filter standard *Planck* pipeline)
  - **Step 1:** ML applied to filter output across 32 spatial scales. Look for additional information not used in simple filter S/N cut.
  - **Step 2:** ML directly applied to nine Planck frequency maps. Use deep convolution neural network (CNN).
  - **Step 3**: ML applied to optical/NIR galaxy catalogs
- Move production to AWS platform.

The *Planck* Mission

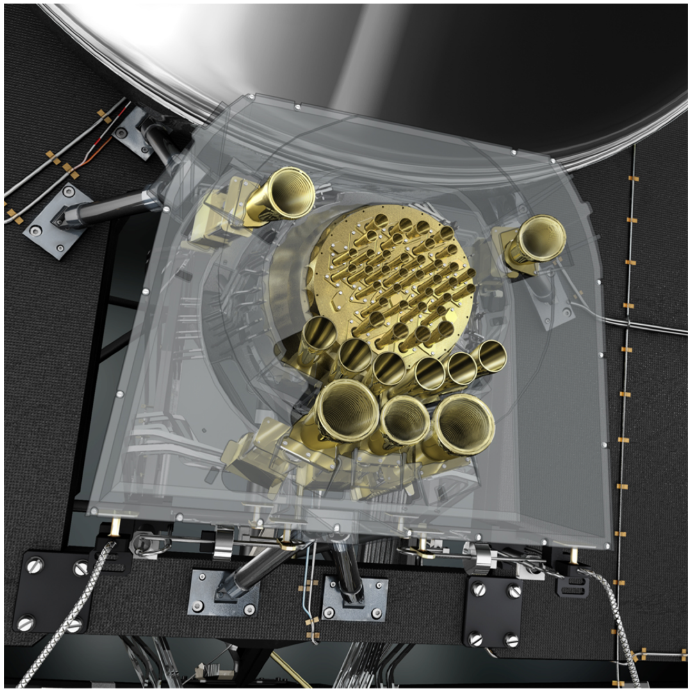
- Launch: 14 May 2009
- Early results: 1/2011
- HFI coolant exhausted: 1/2012 – 29 months of science data
- 1st data release: 3/2013 – 15.5 months
- LFI shutdown: 10/2013
- Deorbit L2: 10/2013
- 2nd data release: 10/2014 – full data set
- Final data analysis: 2018

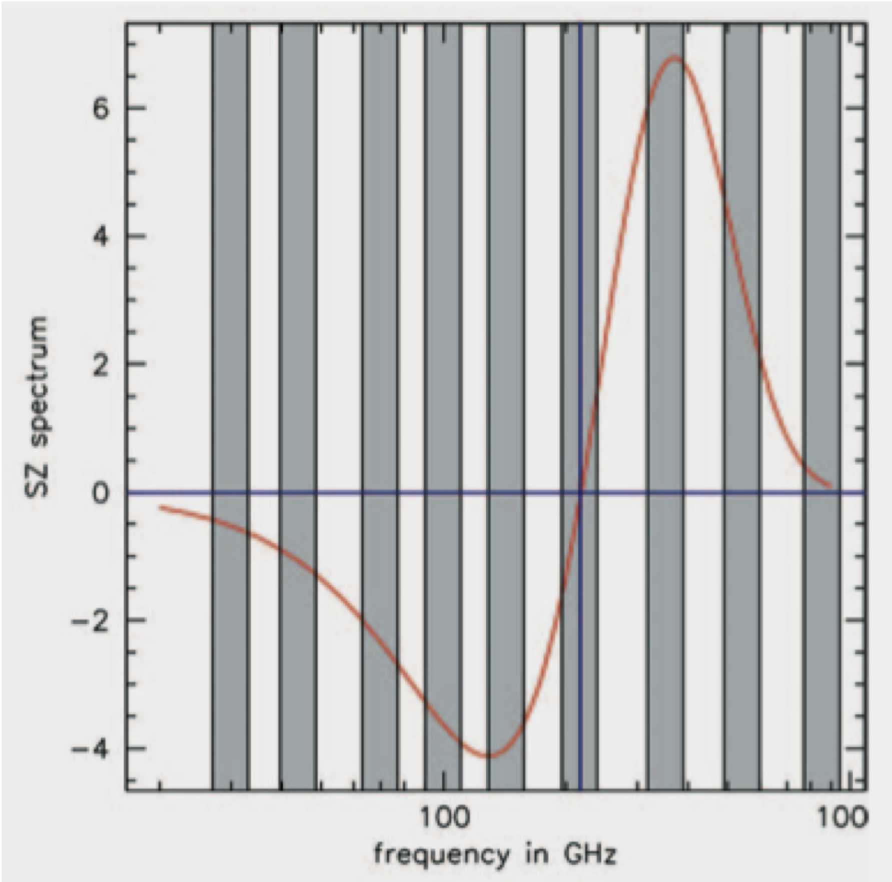






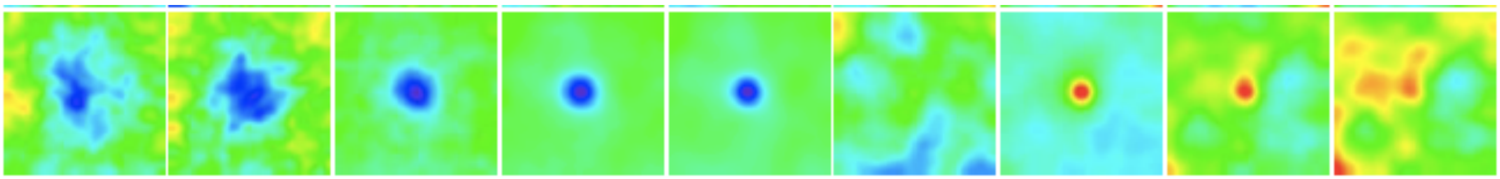
The *Planck* Instruments





- HFI: 100, 143, 217, 353, 545, 857 GHz
- LFI: 30, 44, 70 GHz

Detecting Clusters with *Planck*

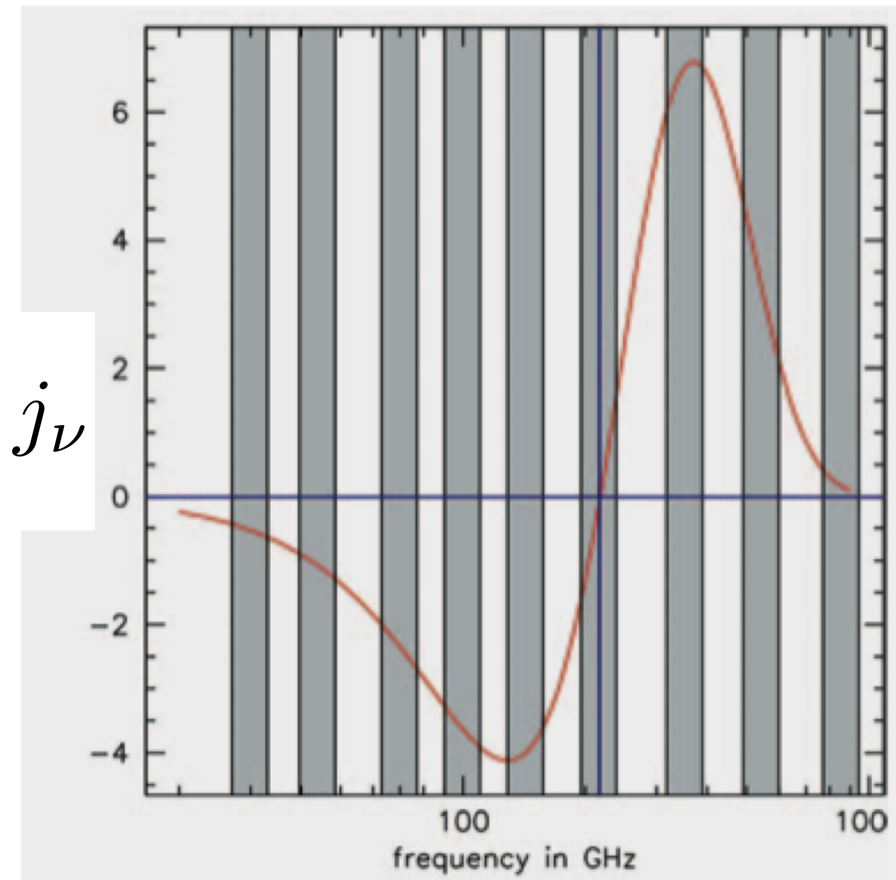


We detect clusters via the Sunyaev-Zeldovich (SZ) effect with a multi-frequency matched filter (*Melin et al. 2006)*

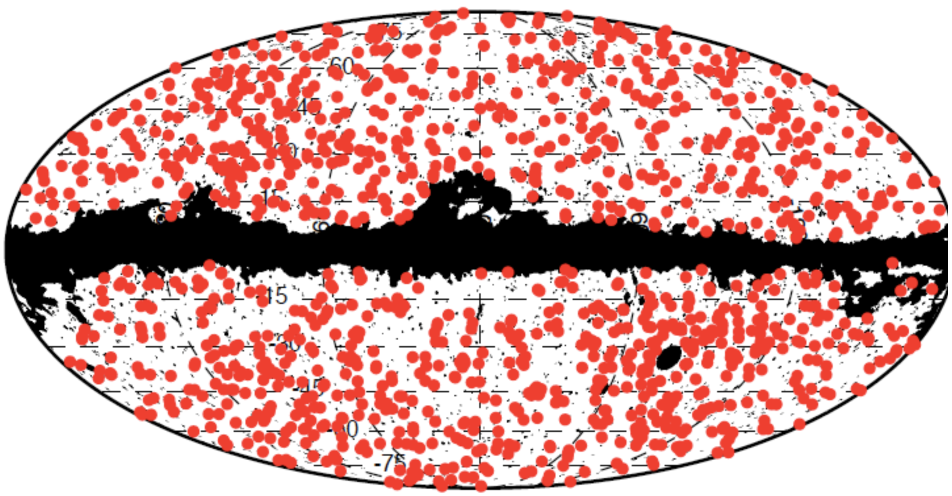
$$\Delta i_\nu(\hat{n}) = y(\hat{n}) j_\nu$$

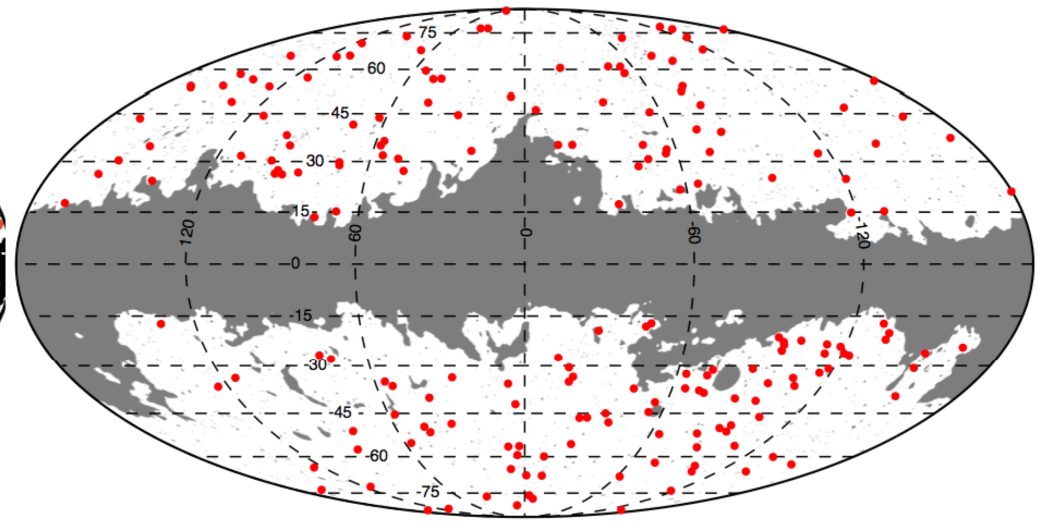
$$y = \int_{\rm los} dl \, \frac{kT_{\rm e}}{m_{\rm e}c^2} n_{\rm e}\sigma_{\rm T}$$

$$Y_{500} = 4\pi \int_0^{R_{500}} r^2 dr \; y(\hat{n})$$



The *Planck* 2013 SZ Catalogs





PSZ1

- 1227 sources
- 861 confirmed clusters / 683 known / 178 new confirmed
- 366 candidate new clusters
- S/N>4.5

Cosmology Sample

- 189 confirmed clusters
- S/N>7
- All but one with redshifts

Approach

- Development on MacBook using Google's TensorFlow
- Simulated *Planck* data
  - Catalog pre-selected using multi-matched filter (spatial-frequency filter standard *Planck* pipeline)
  - **Step 1:** ML applied to filter output across 32 spatial scales. Look for additional information not used in simple filter S/N cut.
  - **Step 2:** ML directly applied to nine Planck frequency maps. Use deep convolution neural network (CNN).
  - **Step 3**: ML applied to optical/NIR galaxy catalogs
- Move production to AWS platform.

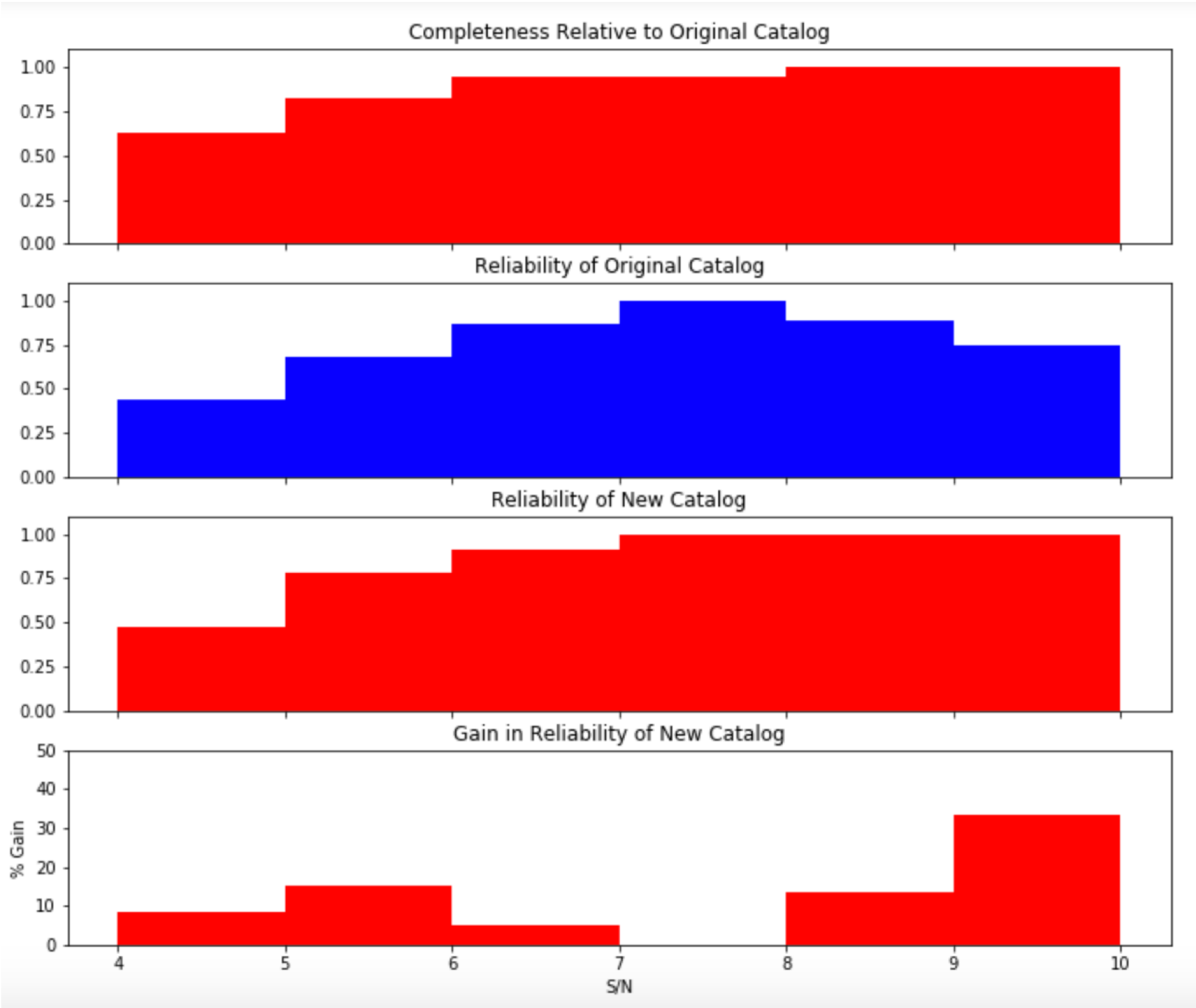
Step 1: Use multi-scale information from MMF

- Multi-matched filter (MMF) assumes a spherical spatial template characterized by the filter scale
- It then linearly combines the filtered maps with weights that extract the cluster signal with minimum variance
- The weights are calculated based on a local noise estimation.

Step 1: Results

- Preparation of simulations and mock catalogs with truth tables to produce training, validation and test sets
- Application and studies of TensorFlow models (ANN) to *Planck* filter output
- Fed vector with 32 elements (one for each spatial scale) to ANN.
- Demonstration that ANN models can improve catalog reliability.
  - Global reliability of pre-selected catalog: 62%
  - TensorFlow models trained on 10% of pre-selected catalog improve to 75-80% global reliability with varying loss of completeness

Second run on Model A

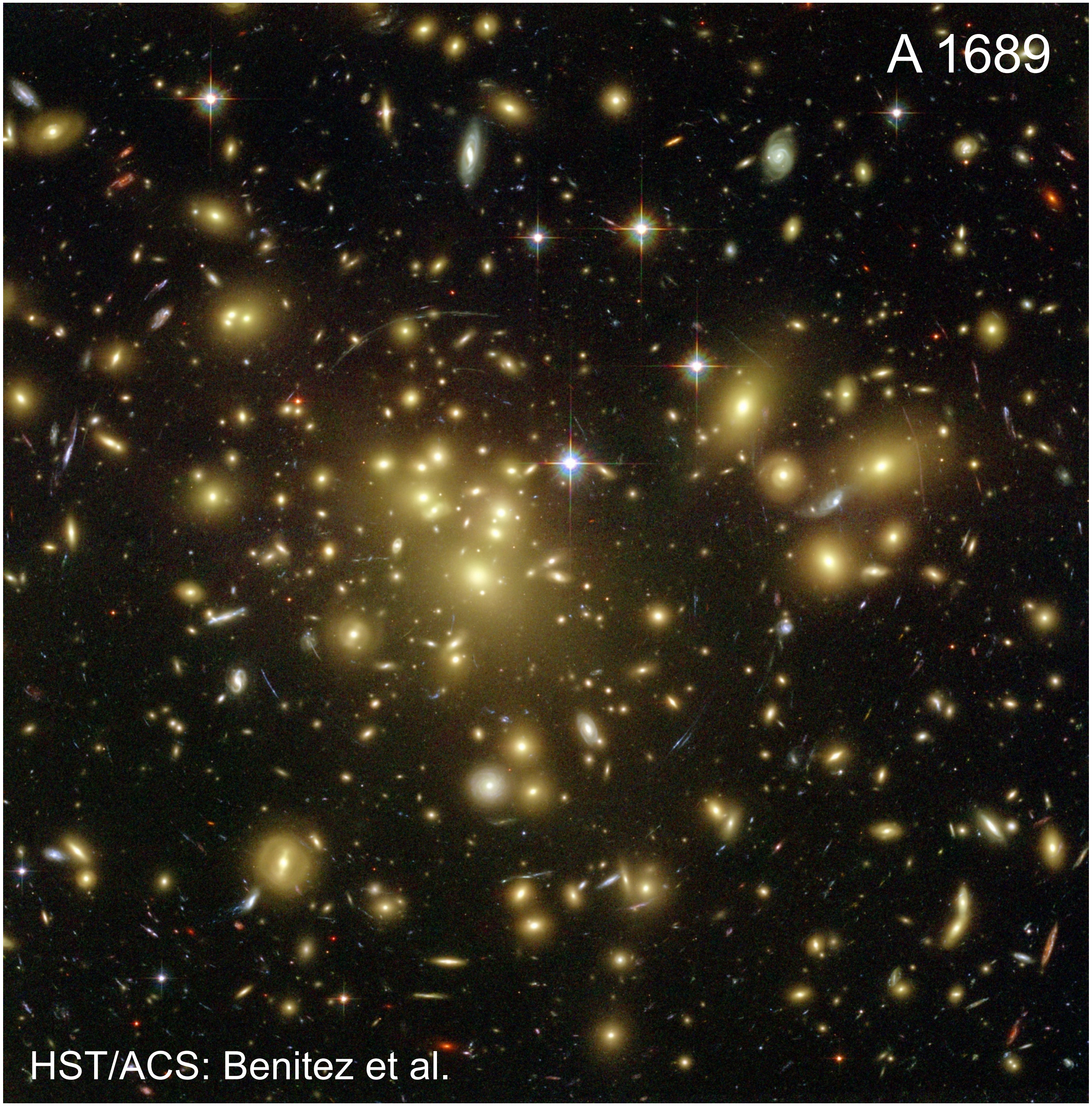


Step 2: Working directly with maps

- Work directly with nine frequency maps
  – Multichannel image (34x34 pixels)
  – Simulated clusters and false detections from MMF
- I'm stuck here
  – Tried different CNN models
  – Seems to overfit training set
  – No improvement on test set: essentially random results
  – Moving to larger images to enable better noise estimation.

Step 3: Move on to optical/infrared images

- Multichannel (e.g., six color bands)
- Find clusters based on galaxy properties
- Clusters are more complex when viewed by their member galaxies
- Greatest potential for gains from ML


A 1689
HST/ACS: Benitez et al.

Conclusions

- There is an indication that ML can improve cluster detection, at least in the simple case of Step 1.
- Trying larger images to improve performance in Step 2.
- Important characteristic of our use case: we are noise-dominated.

Jet Propulsion Laboratory
California Institute of Technology

Jet Propulsion Laboratory
California Institute of Technology

Backup Slides