Exploring astronomy data archives at scale using deep learning and crowdsourcing

Sandor Kruk
Data Scientist, European Space Agency

Bruno Merín, Rachana Bhatawdekar, Deborah Baines, Elena Racero, Pablo García Martín, Marcel Popescu, Max Mahlke, Benoît Carry, Ross Thomson, Samet Karadag, Guido de Marchi, Emily Garvin, Claude Cornen, Ben Aussel, Megan Perks, Steven Dillman, Tamina Lund, Mark McCaughrean

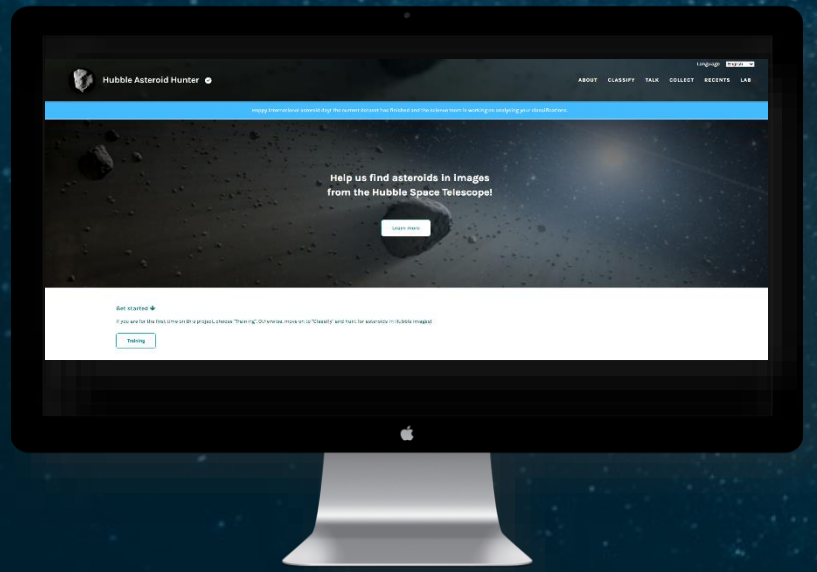
09/05/2023

ESA UNCLASSIFIED – For ESA Official Use Only

AI and crowdsourcing for knowledge discovery

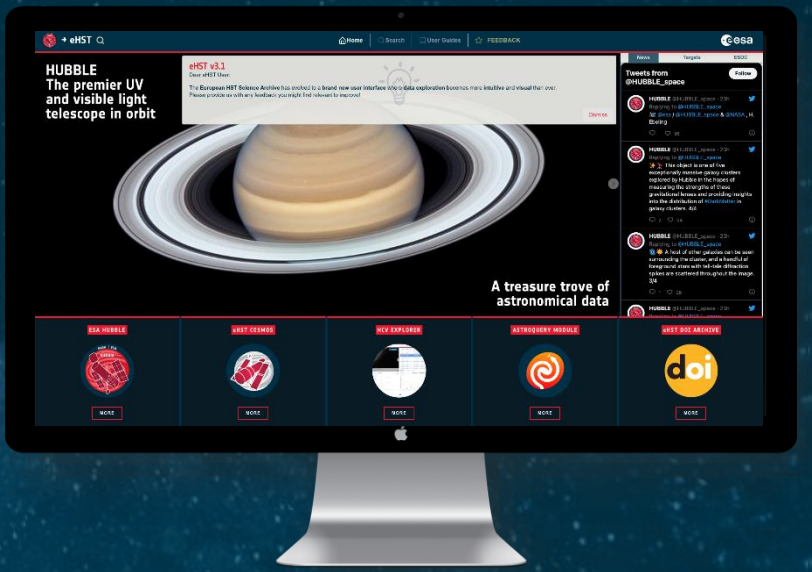
"Knowledge Discovery is the task of processing and analysing astronomical datasets with the aim of extracting new knowledge. This endeavour spans multiple disciplines including visualisation, data access and exploration, machine learning, statistical methods and workflow orchestration."

Exploration: crowdsourcing

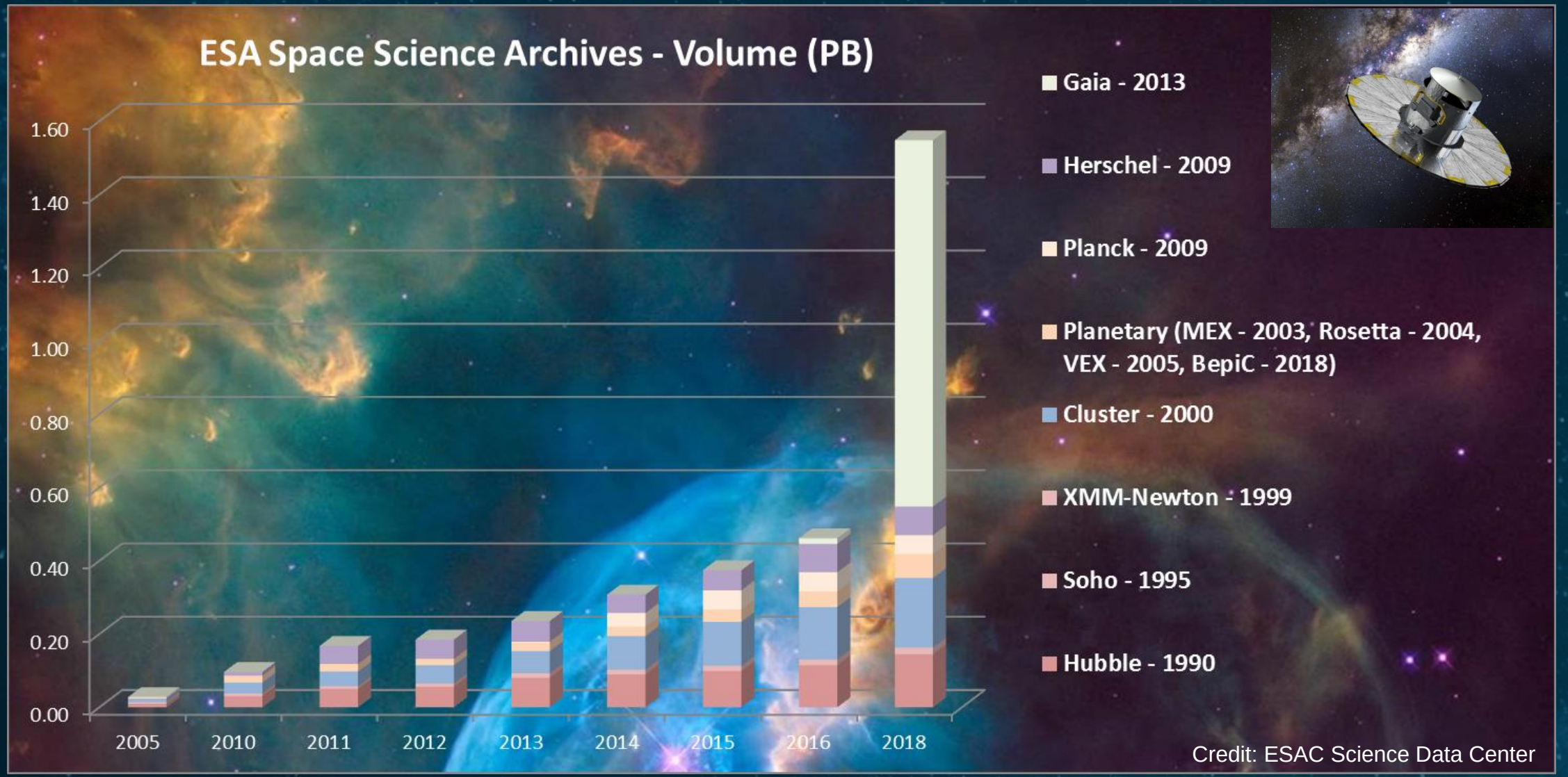


ML: Deep learning

Large astronomical archives: eHST

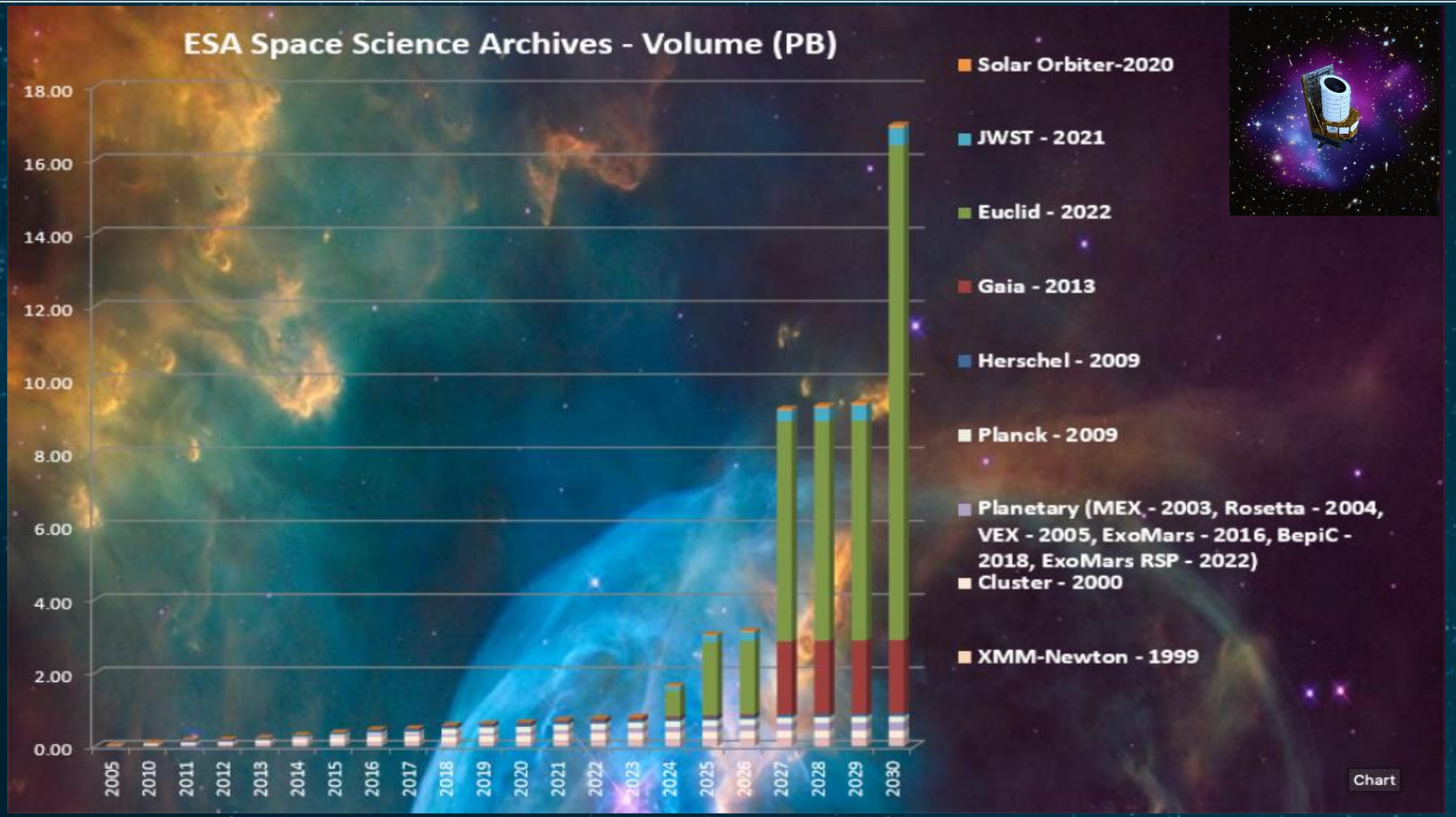


Motivation: increasing data volumes in astronomy

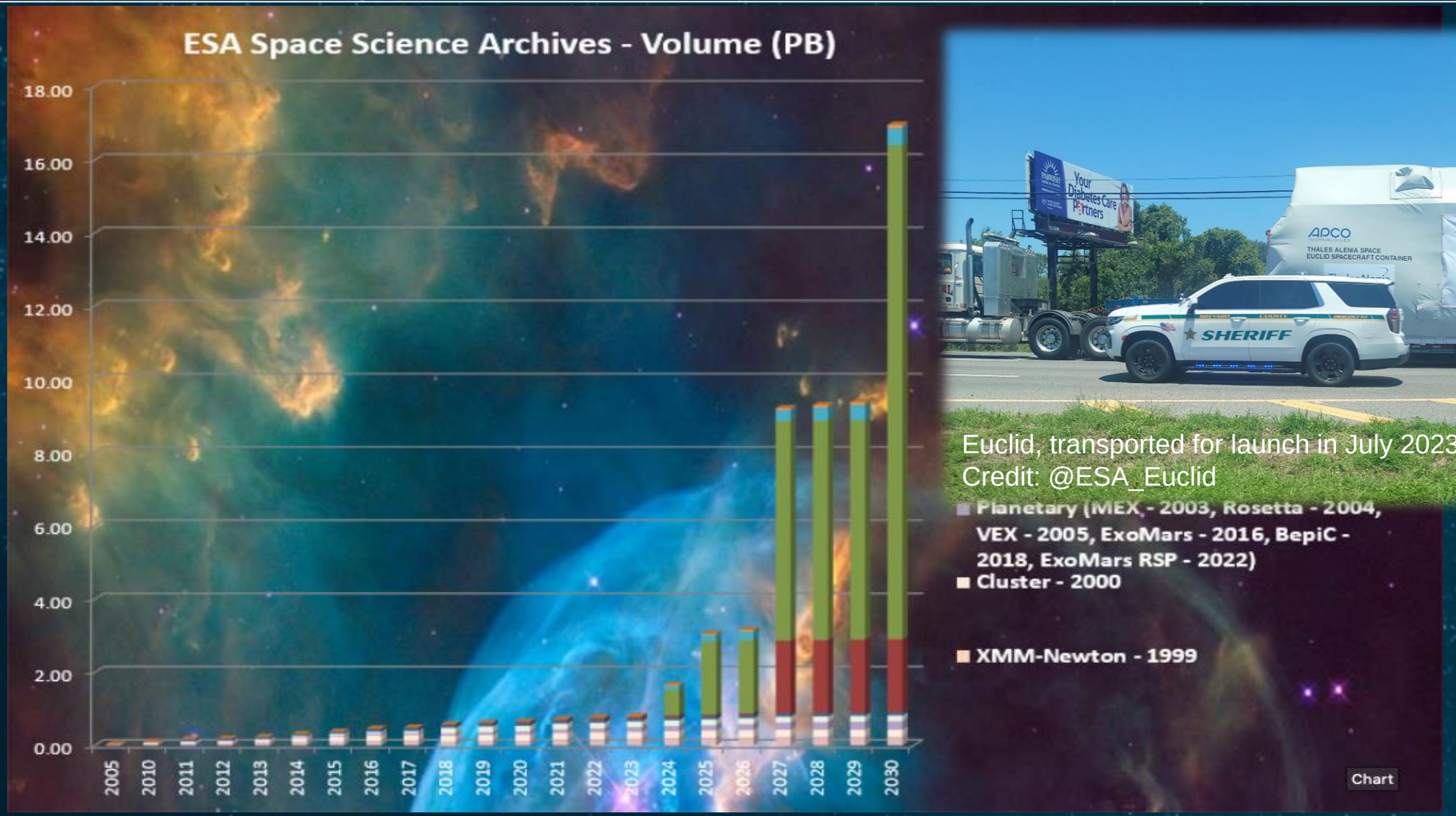


Credit: ESAC Science Data Center

Motivation: increasing data volumes in astronomy

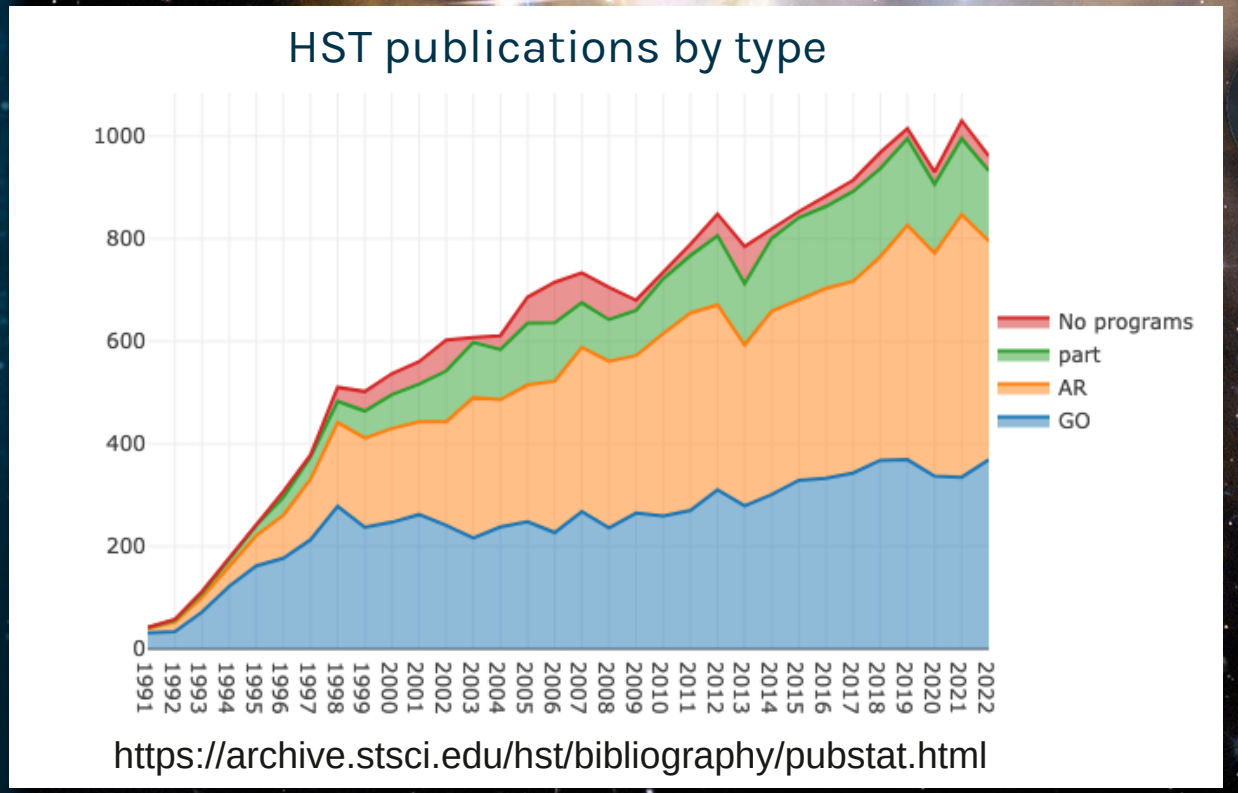


Motivation: increasing data volumes in astronomy



Euclid, transported for launch in July 2023.
Credit: @ESA_Euclid

Archival data – Hubble Space Telescope



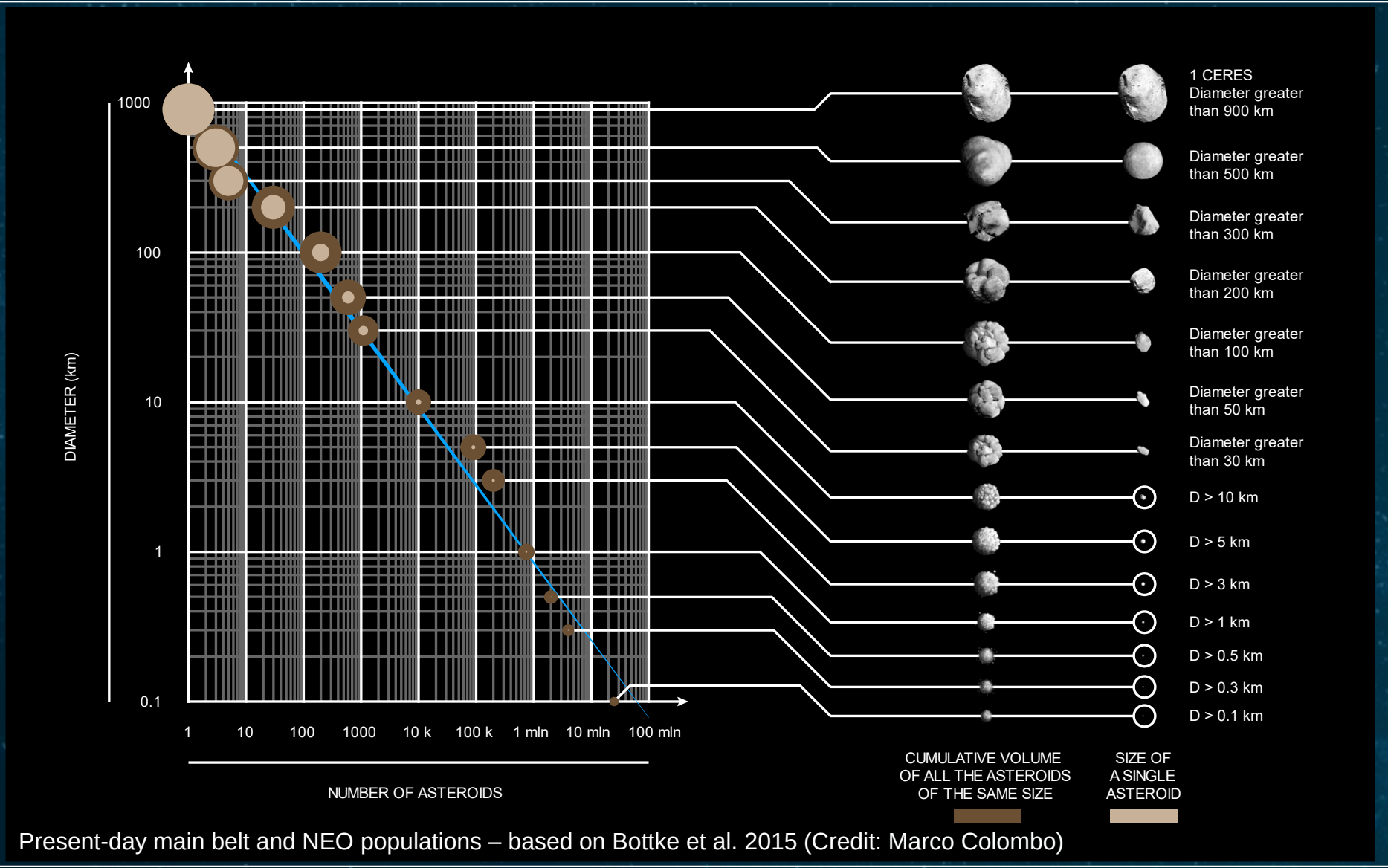
https://archive.stsci.edu/hst/bibliography/pubstat.html

Hubble Asteroid Hunter: Identifying asteroid trails in Hubble Space Telescope images

Kruk et al. 2022, arXiv: 2202.00246

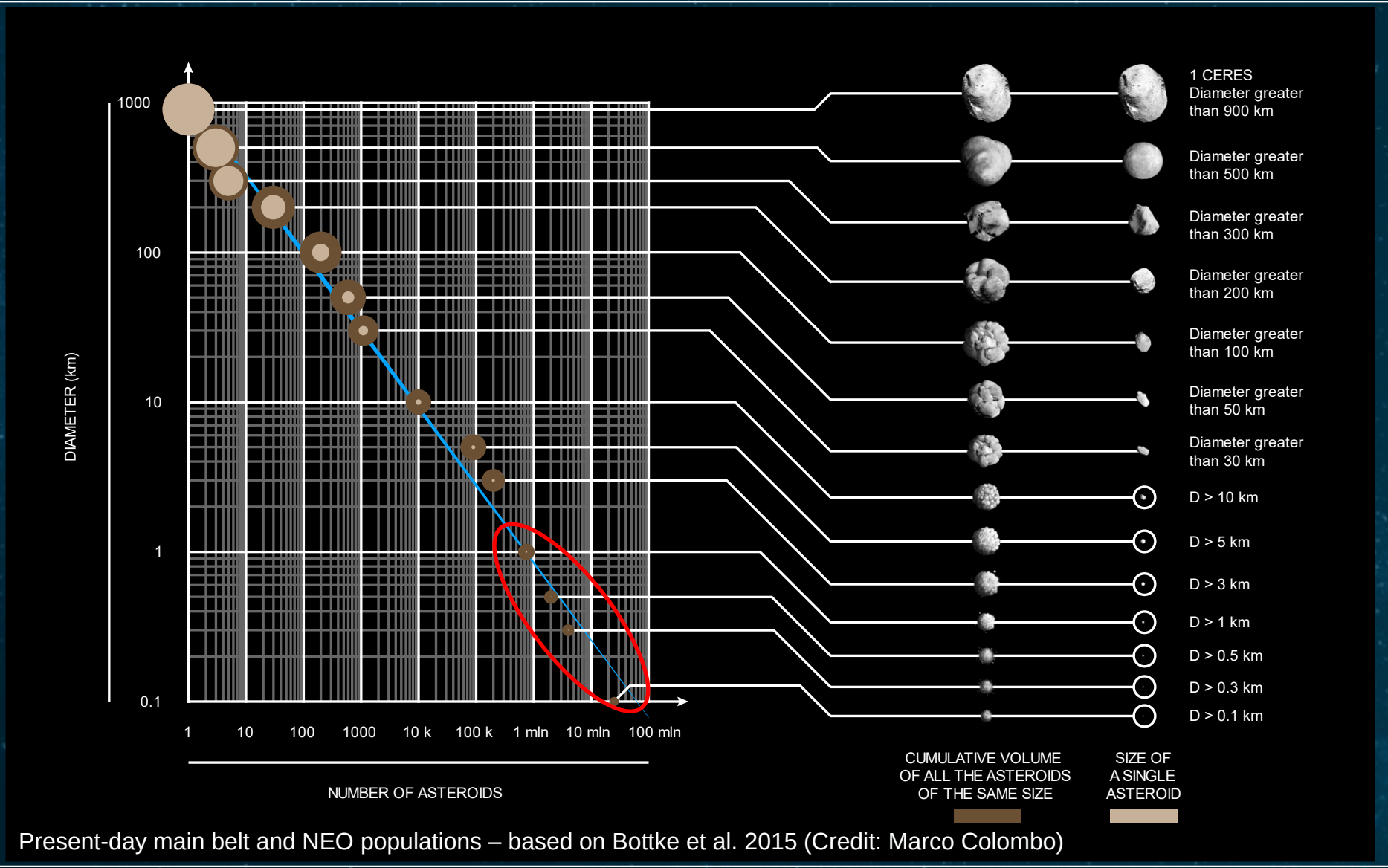
Abell 370, Credit: NASA, ESA/Hubble, B.Sunnquist and J. Mack

Asteroids in our solar system

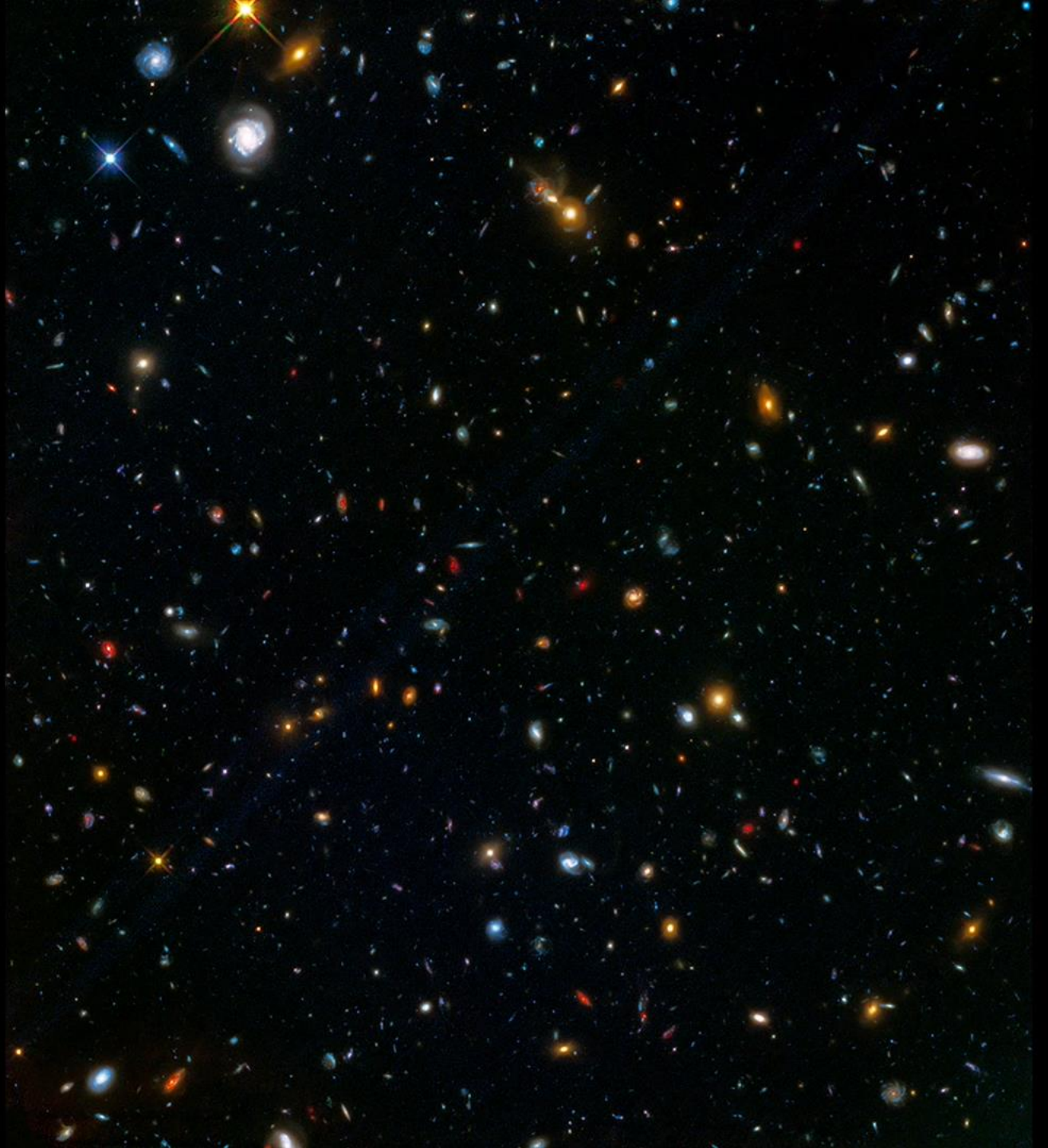


Present-day main belt and NEO populations – based on Bottke et al. 2015 (Credit: Marco Colombo)

Asteroids in our solar system

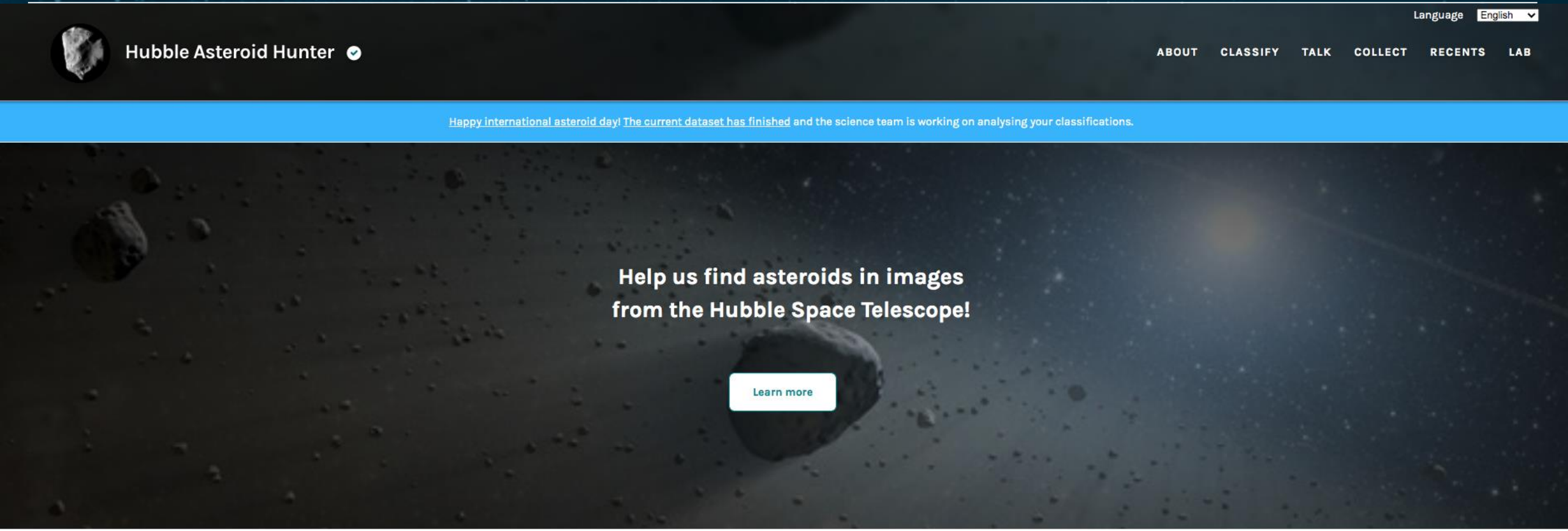


Present-day main belt and NEO populations – based on Bottke et al. 2015 (Credit: Marco Colombo)

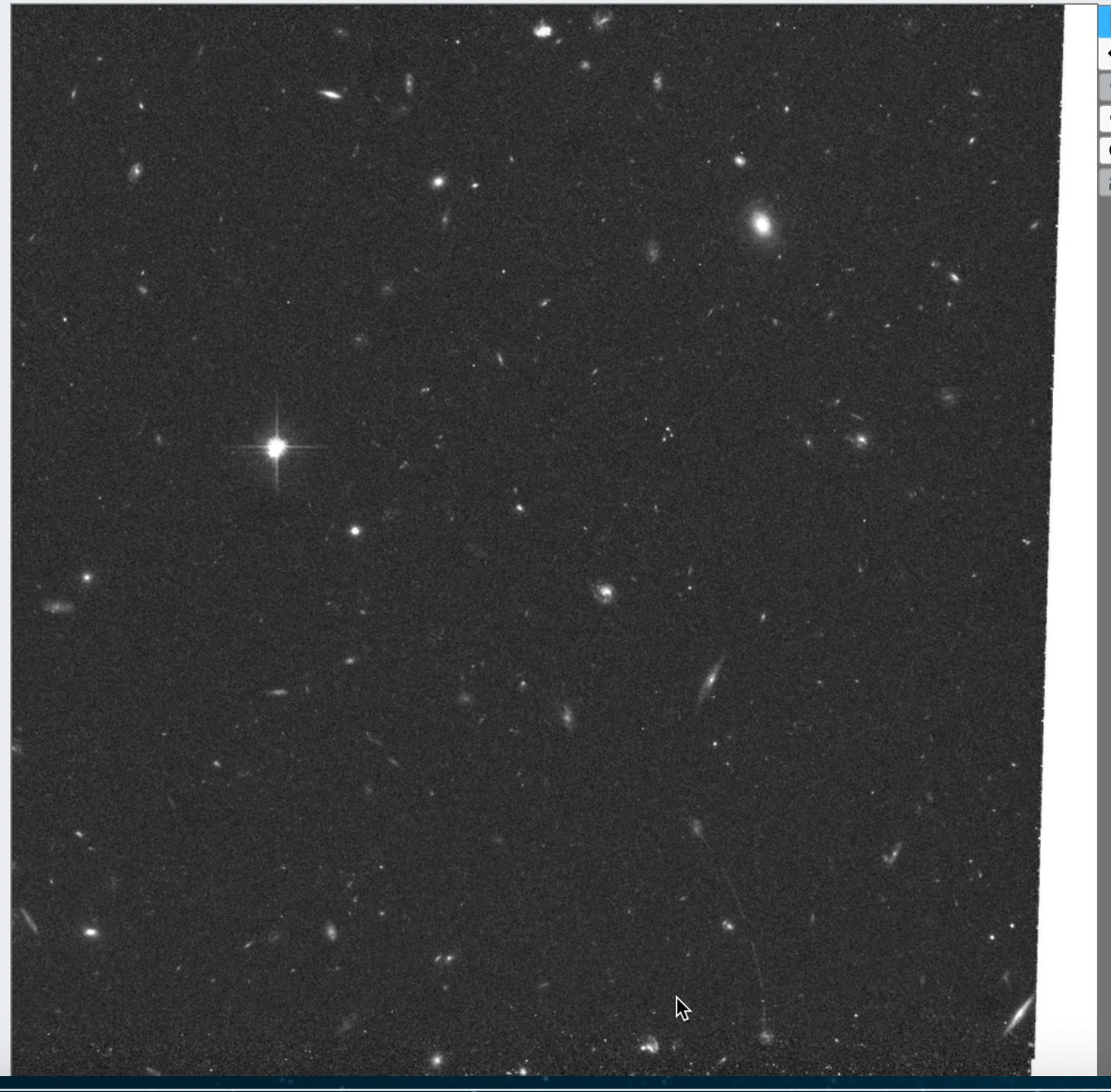


Abell 370 Parallel Field
Credit: NASA, ESA/Hubble

Zooniverse Citizen Science project



www.asteroidhunter.org



TASK TUTORIAL

Is there an asteroid trail visible in the images?

Yes

No

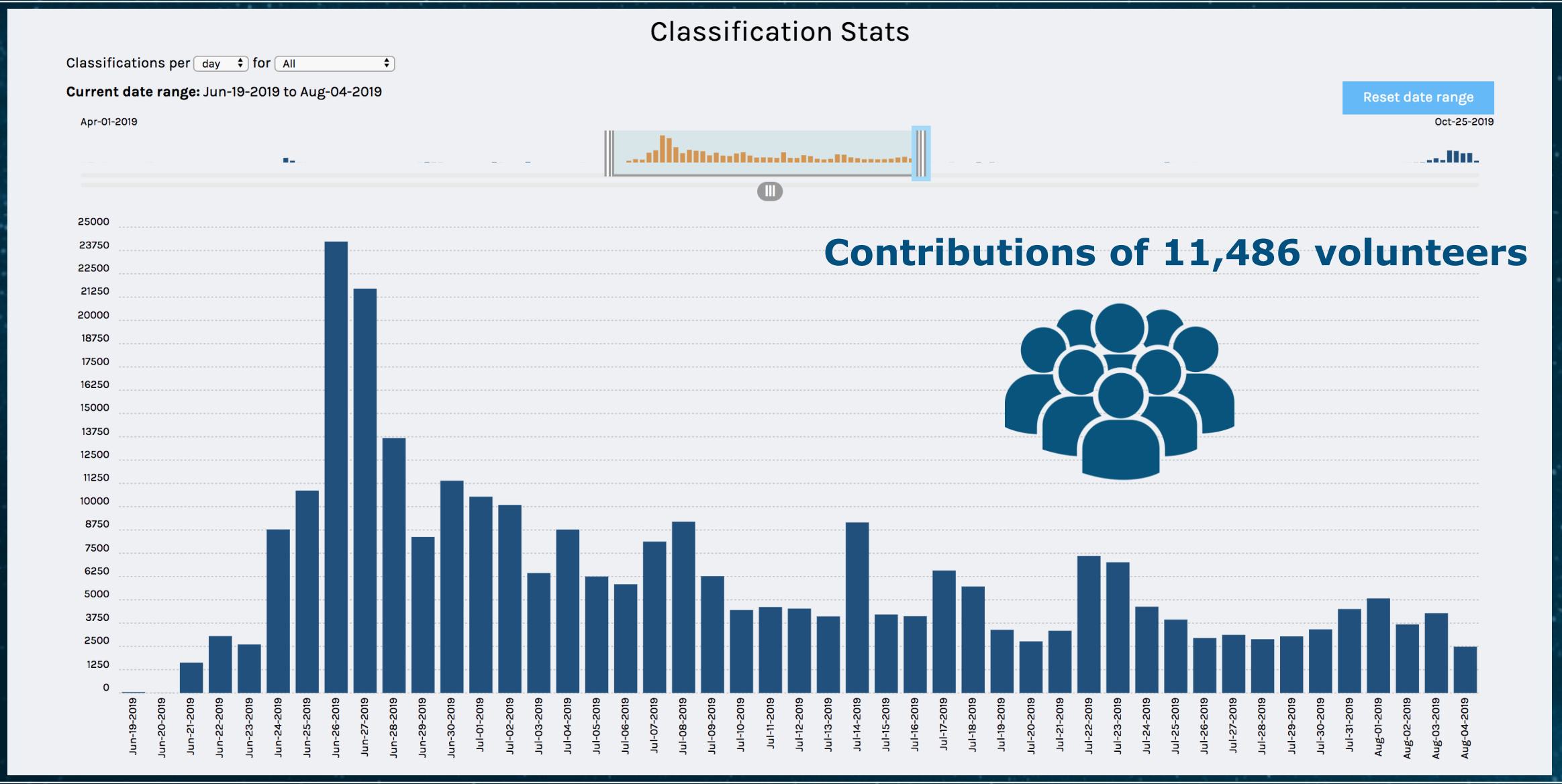
Impossible to tell

NEED SOME HELP WITH THIS TASK?

Done & Talk Done

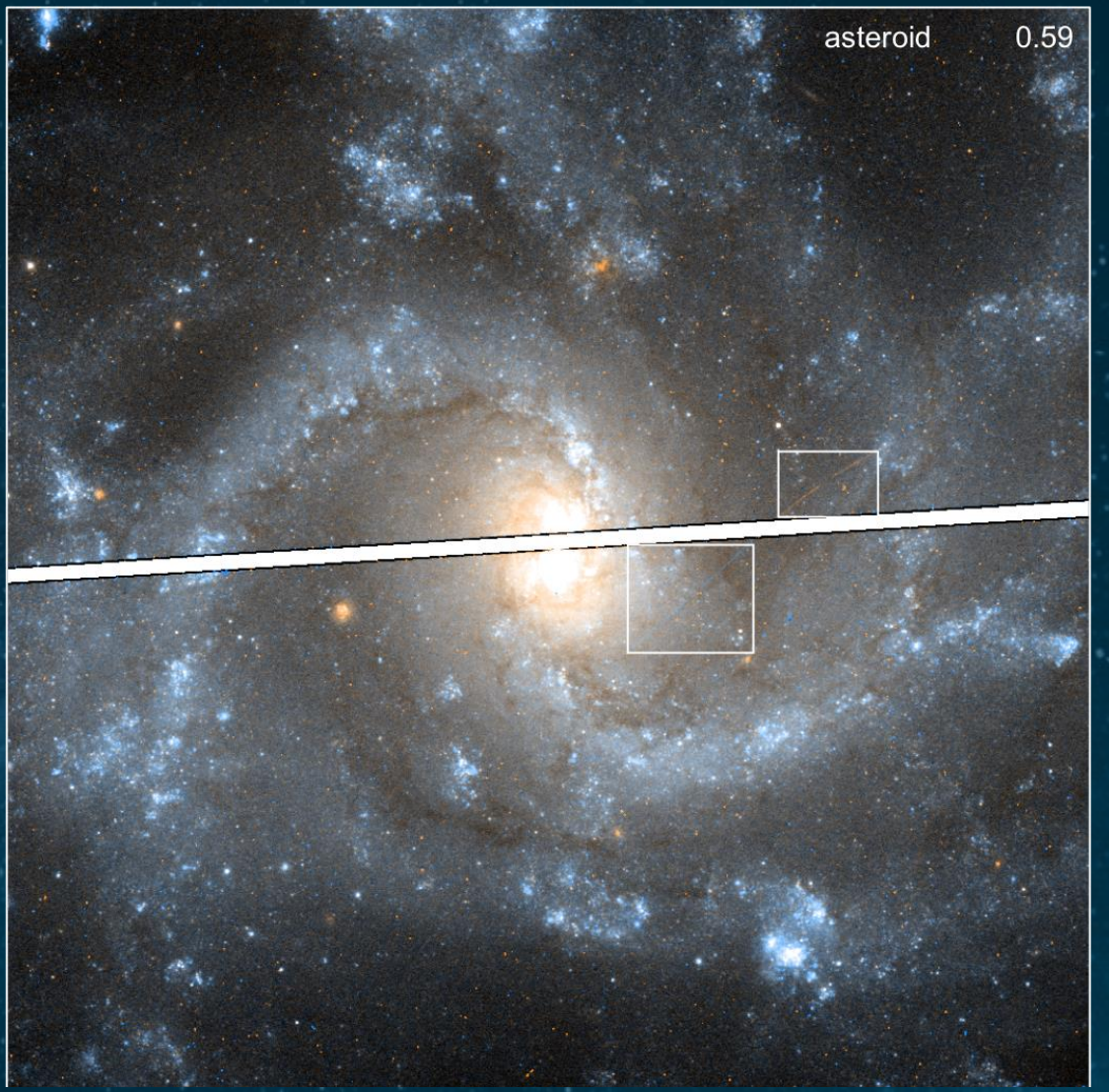
FIELD GUIDE

Volunteer participation in the citizen science project



Scaling up with AutoML to identify asteroid trails

Use automated machine learning (AutoML) on Google Cloud



* In collaboration with Google

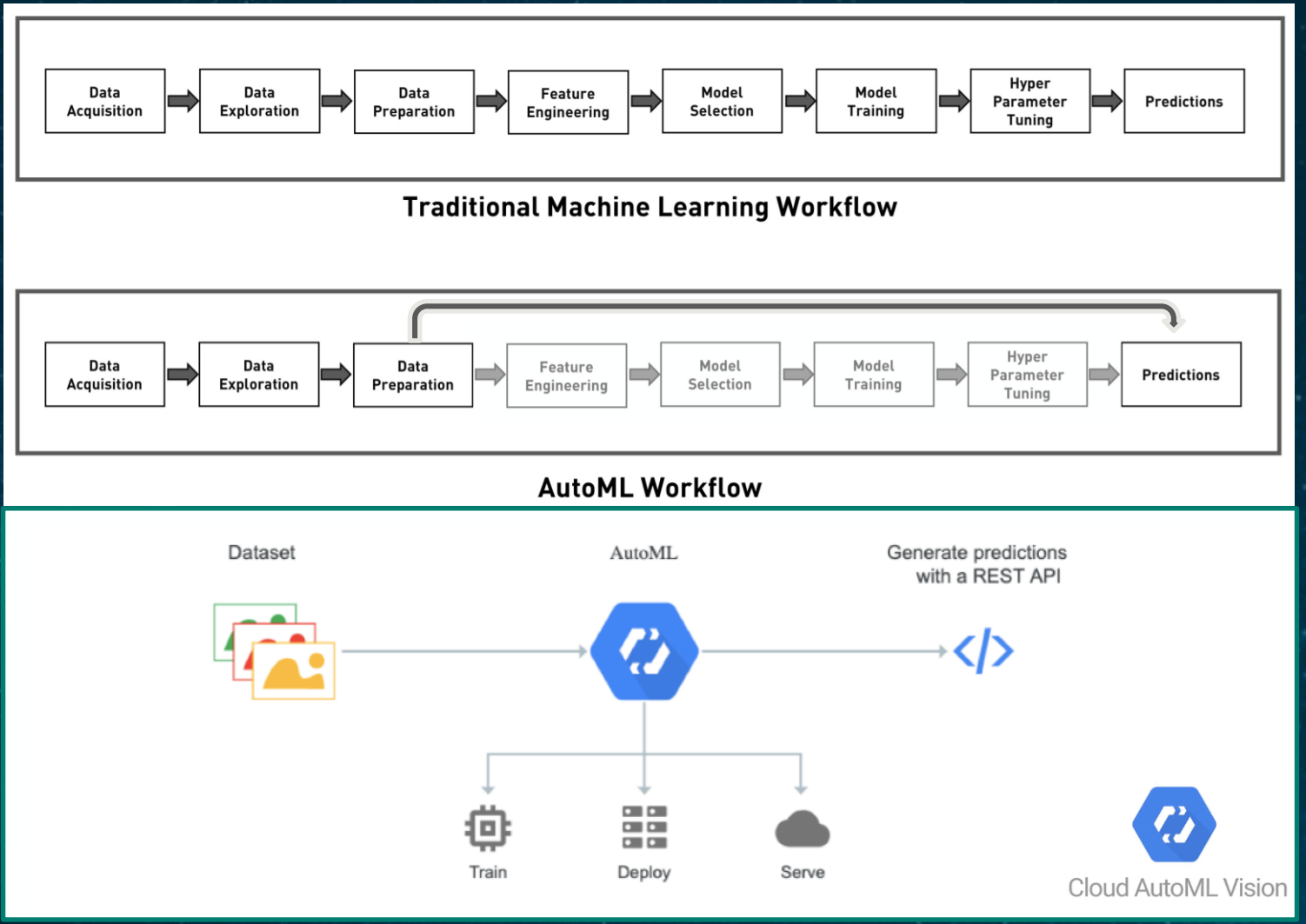
Scaling up with AutoML to identify asteroid trails

Use automated machine learning (AutoML) on Google Cloud

Scalable: training and batch classification on Google Cloud

* In collaboration with Google

Automating ML pipelines



How AutoML works – using Neural Architecture Search

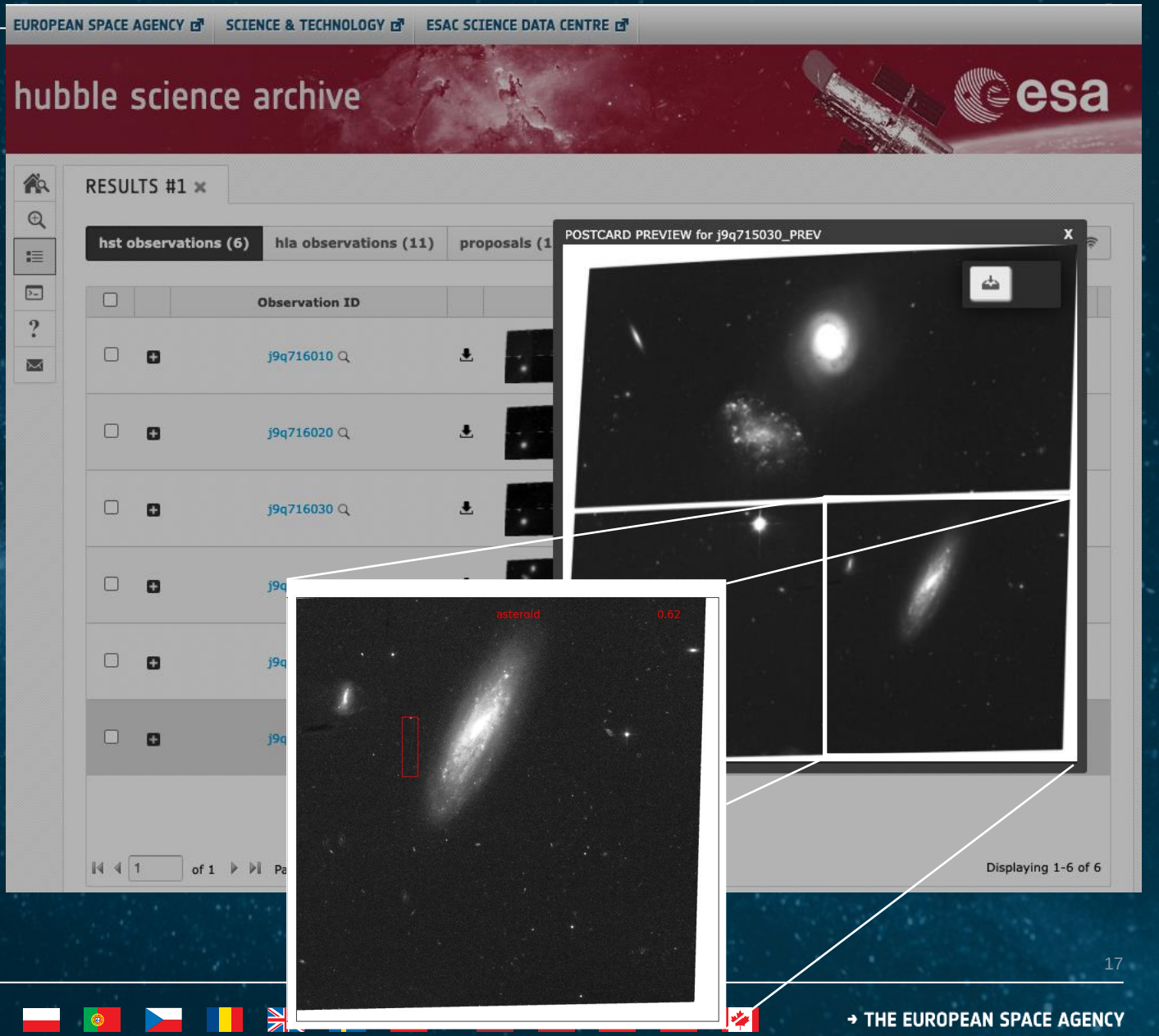
Scaling up with AutoML to identify asteroid trails

Use automated machine learning (AutoML) on Google Cloud

Scalable: training and batch classification on Google Cloud

Classifying the entire HST archive (2002 – 2021) of 37,324 images (x4 = 150k cutouts): 10 hours

It would have taken volunteers 1 year!



* In collaboration with Google

Scaling up with AutoML to identify asteroid trails

Use automated machine learning (AutoML) on Google Cloud

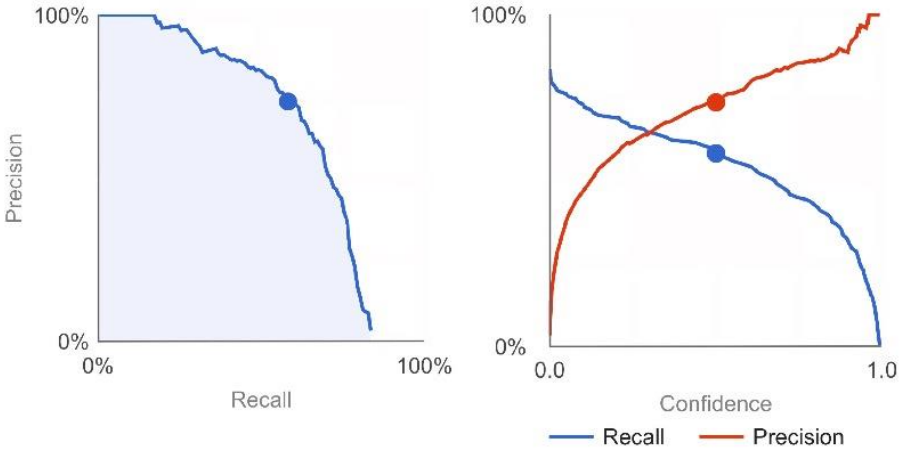
Scalable: training and batch classification on Google Cloud

Classifying the entire HST archive (2002 – 2021) of 37,324 images (x4 = 150k cutouts): 10 hours

It would have taken volunteers 1 year!

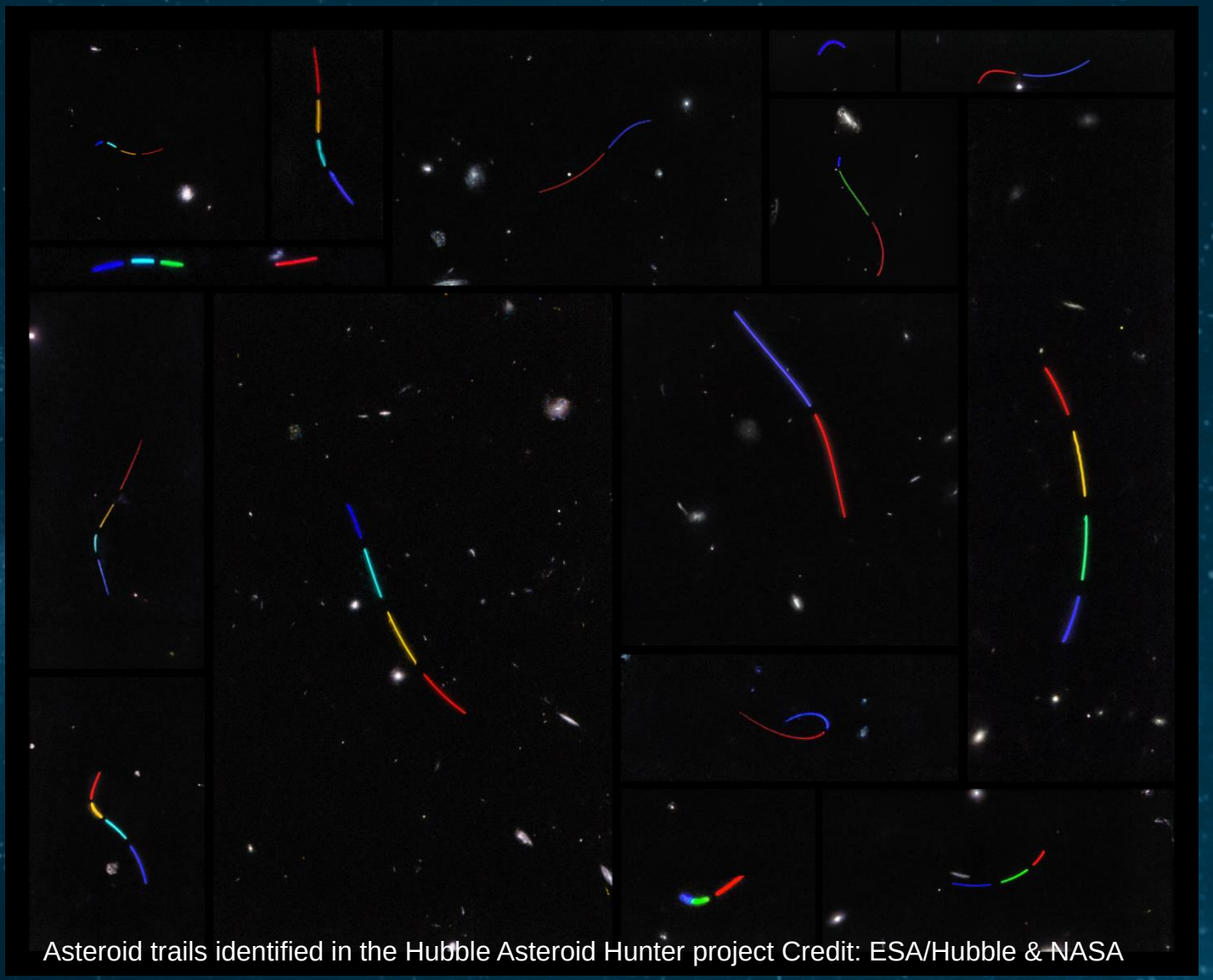
* In collaboration with Google

Performance in identifying asteroid trails



Results: asteroids detected in the HST images

2487 asteroid trails recovered by citizen scientists and by AutoML

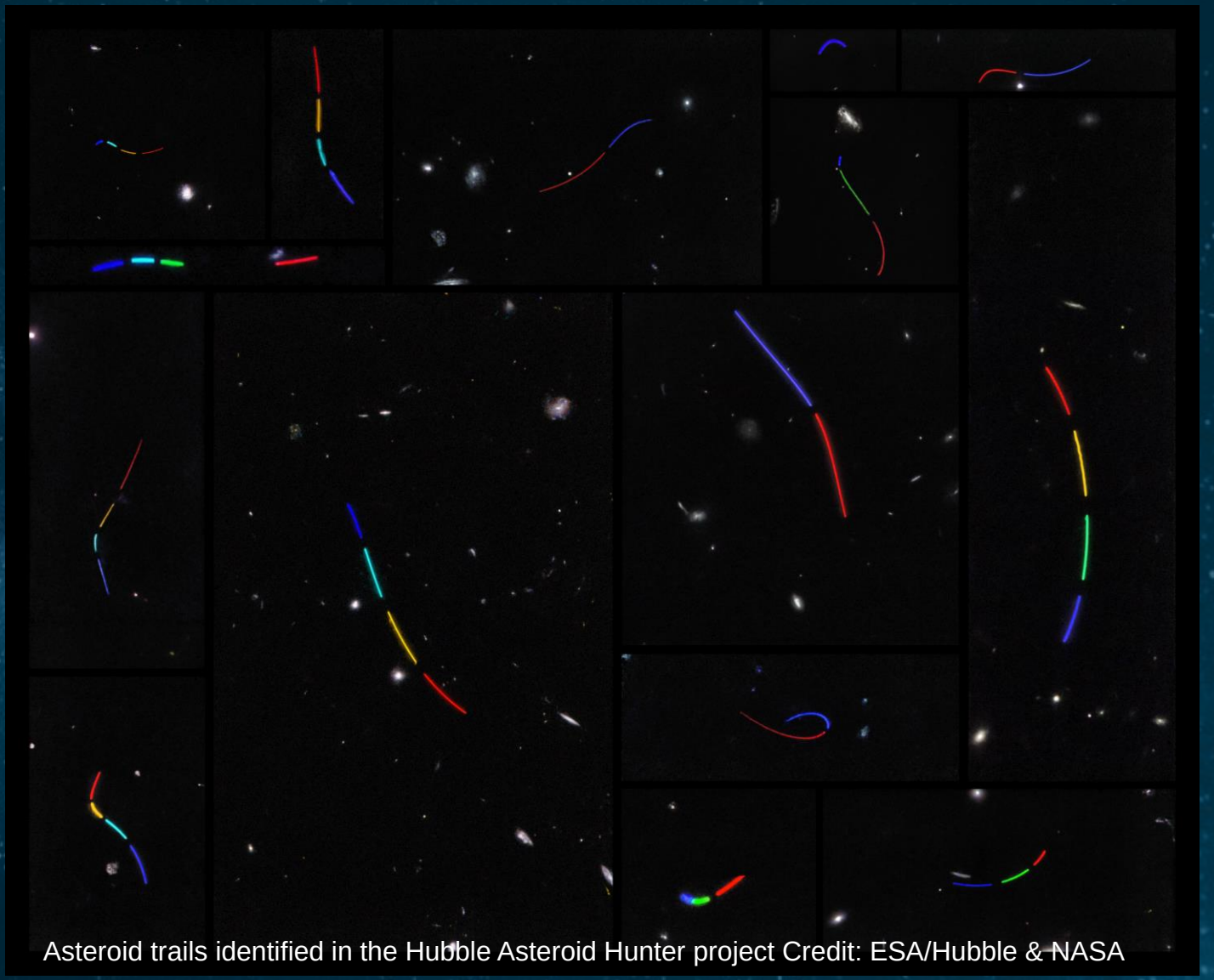


Asteroid trails identified in the Hubble Asteroid Hunter project Credit: ESA/Hubble & NASA

Results: asteroids detected in the HST images

2487 asteroid trails recovered by citizen scientists and by AutoML

1701 asteroids validated by the team



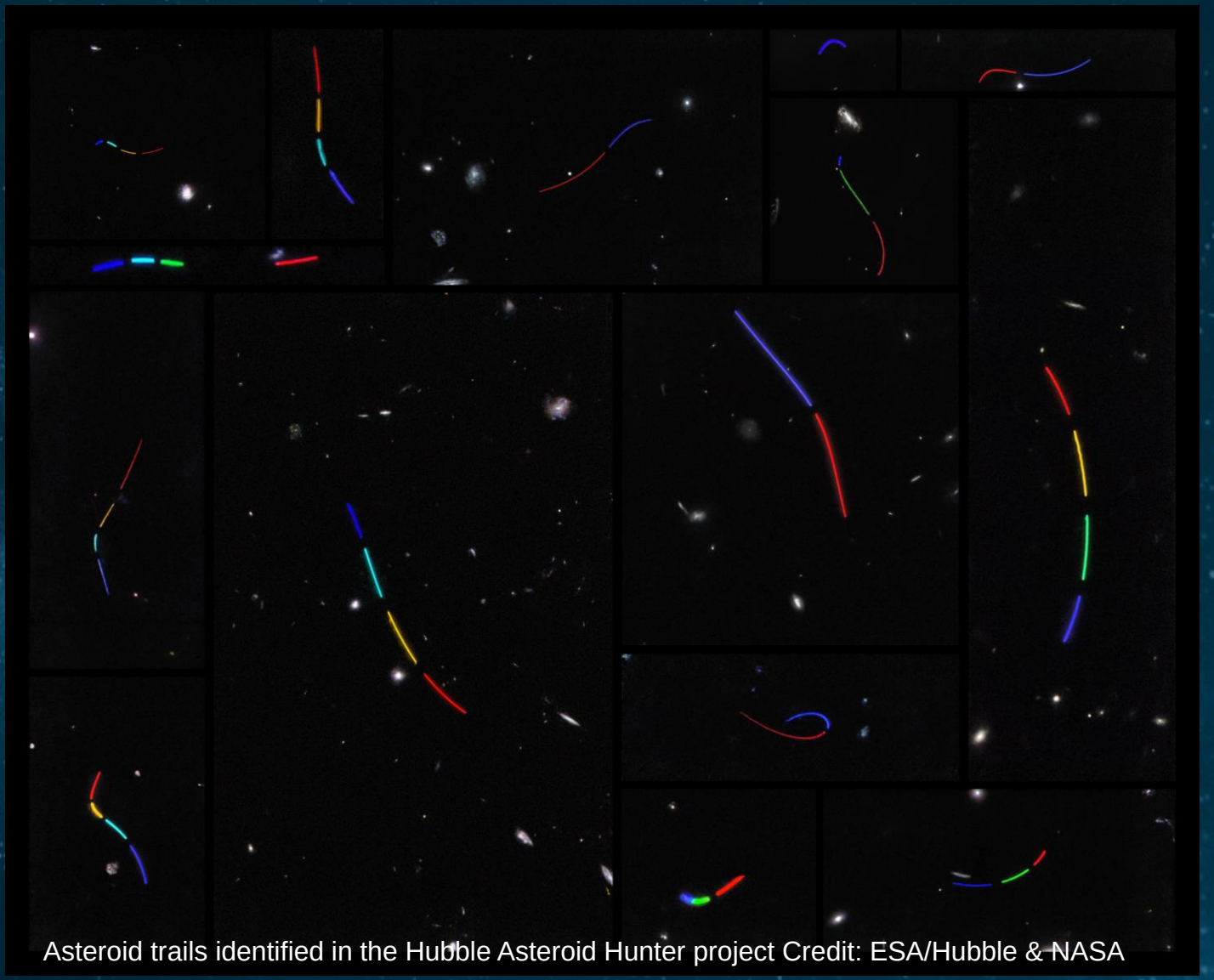
Asteroid trails identified in the Hubble Asteroid Hunter project Credit: ESA/Hubble & NASA

Results: asteroids detected in the HST images

2487 asteroid trails recovered by citizen scientists and by AutoML

1701 asteroids validated by the team

670 asteroids matched with known objects. 95% are Main Belt.



Asteroid trails identified in the Hubble Asteroid Hunter project Credit: ESA/Hubble & NASA

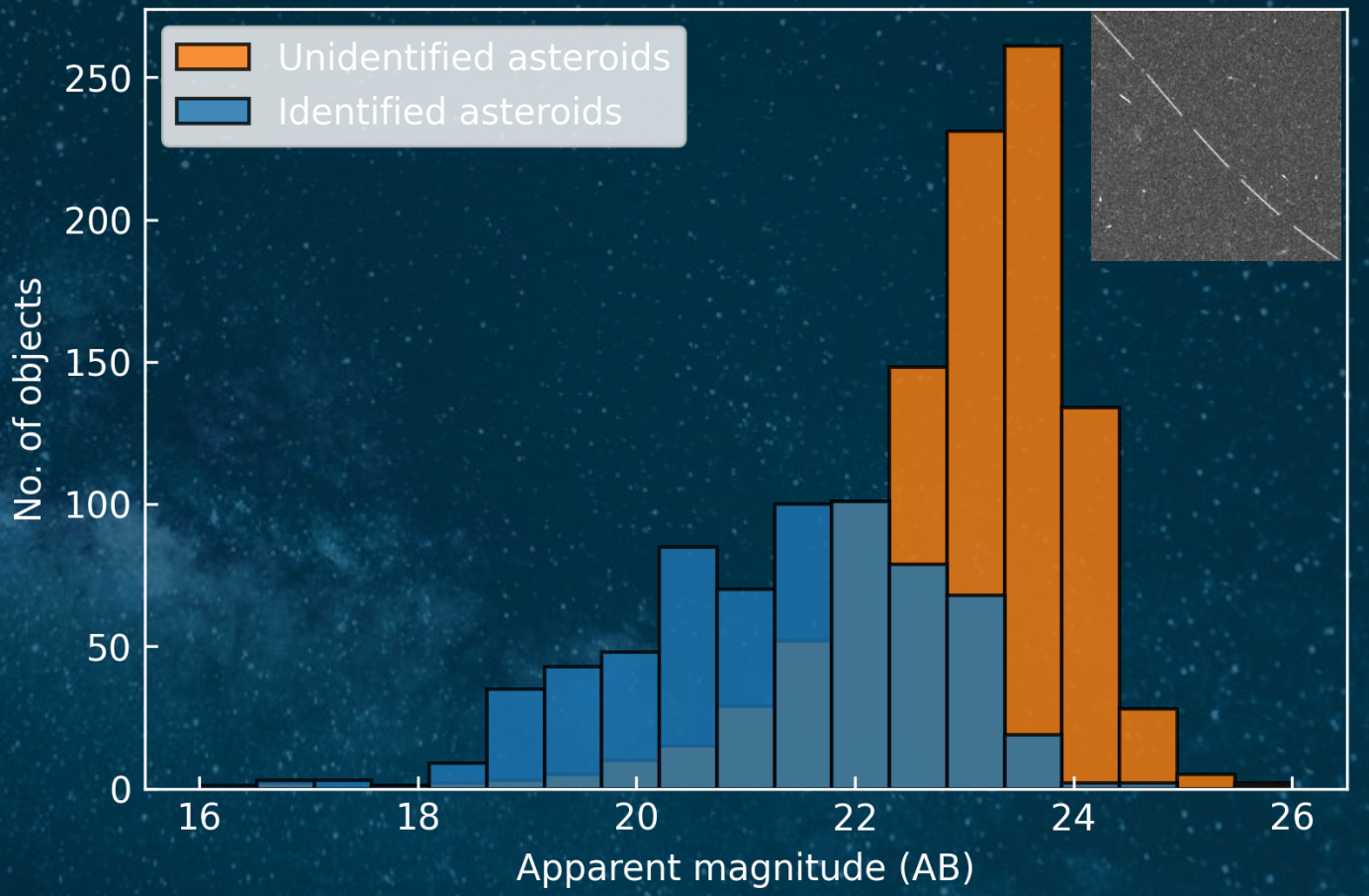
Results: magnitude distribution of detected asteroids

2487 asteroid trails recovered by citizen scientists and by AutoML

1701 asteroids validated by the team

670 asteroids matched with known objects. 95% are Main Belt.

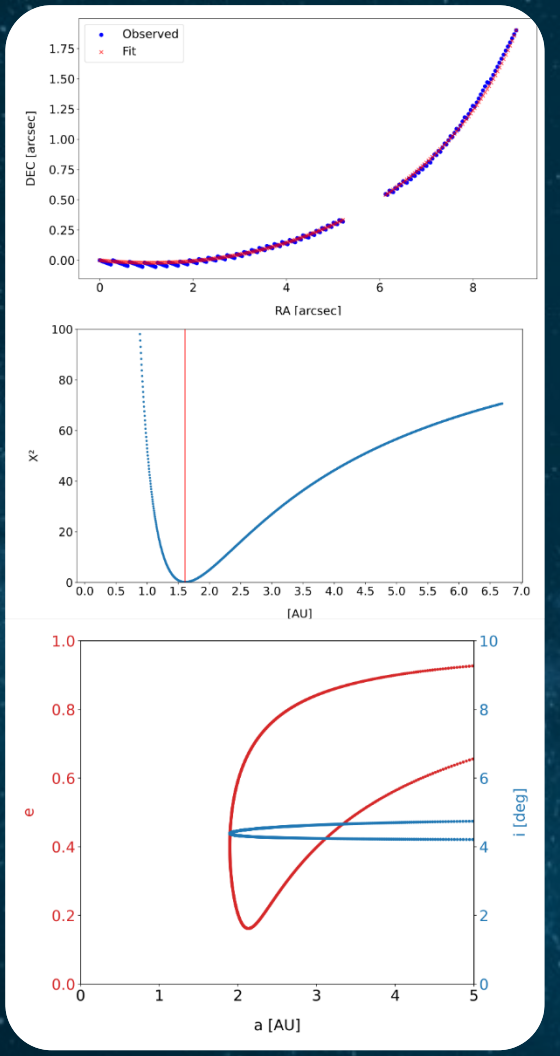
1031 unidentified asteroid trails -- previously unknown asteroids?



Parallax fitting method to study asteroids

Use parallax effect to determine the distance to the asteroids observed with HST

Parallax fitting method



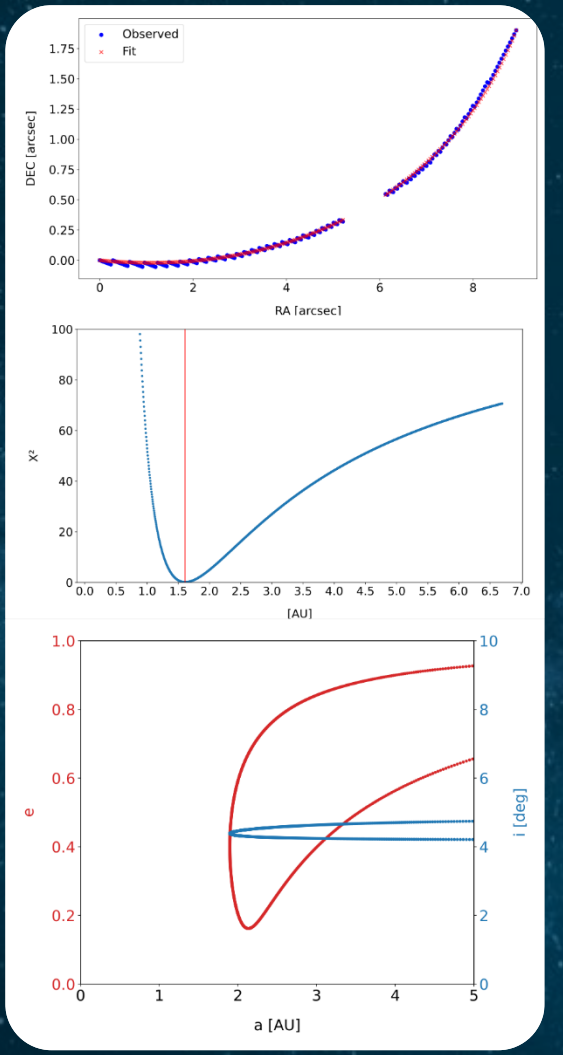
Parallax fitting method to study asteroids

Use parallax effect to determine the distance to the asteroids observed with HST

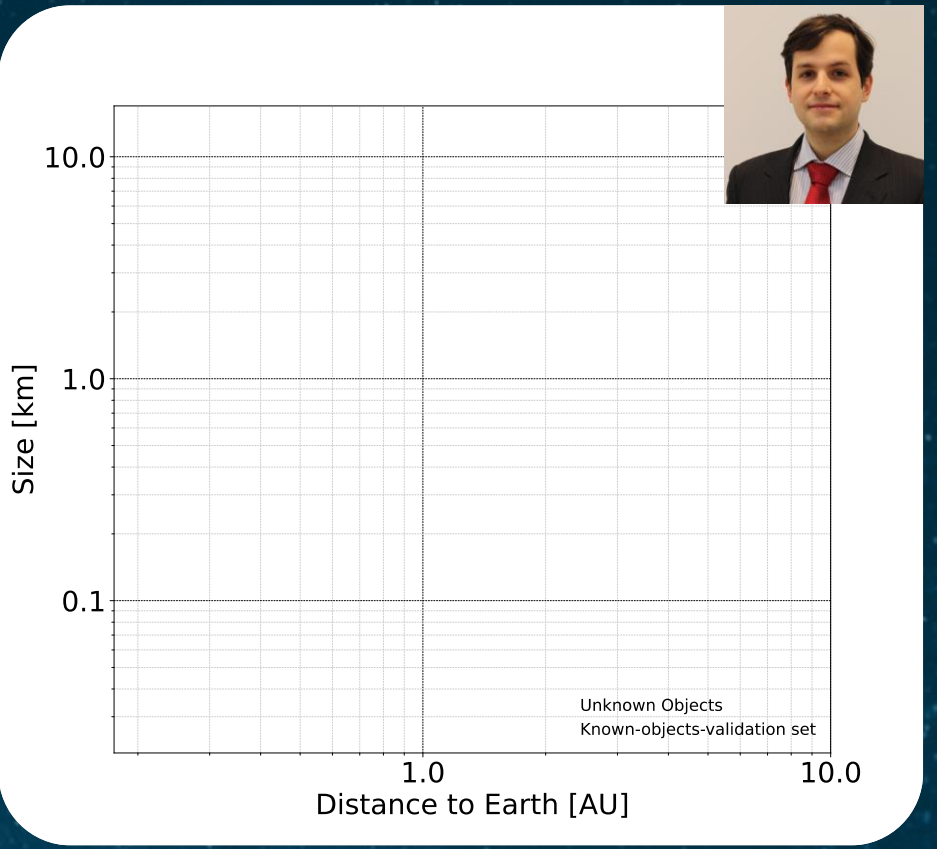
Sizes of new asteroids < 1 km

Improve knowledge of size distribution of small Main Belt asteroids

Parallax fitting method



Size distribution of detected asteroids

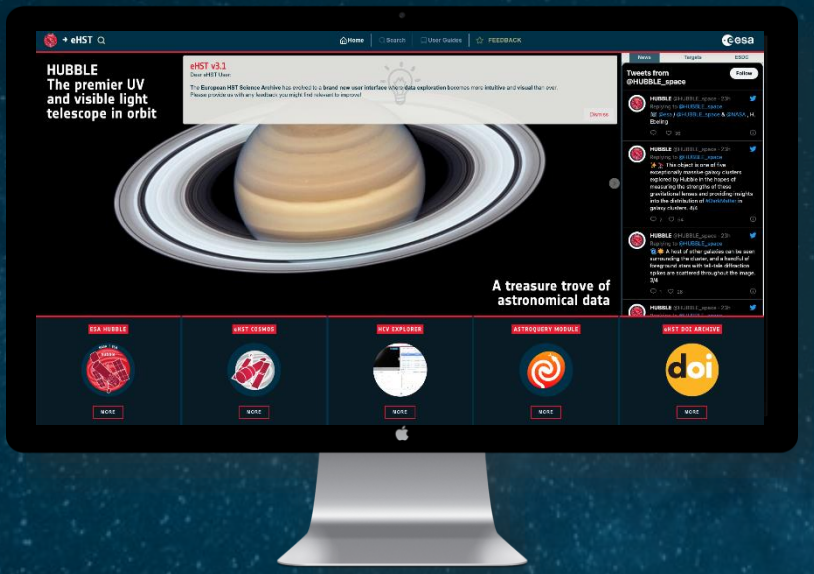


García Martín et al., submitted

Machine learning for data mining in astronomy

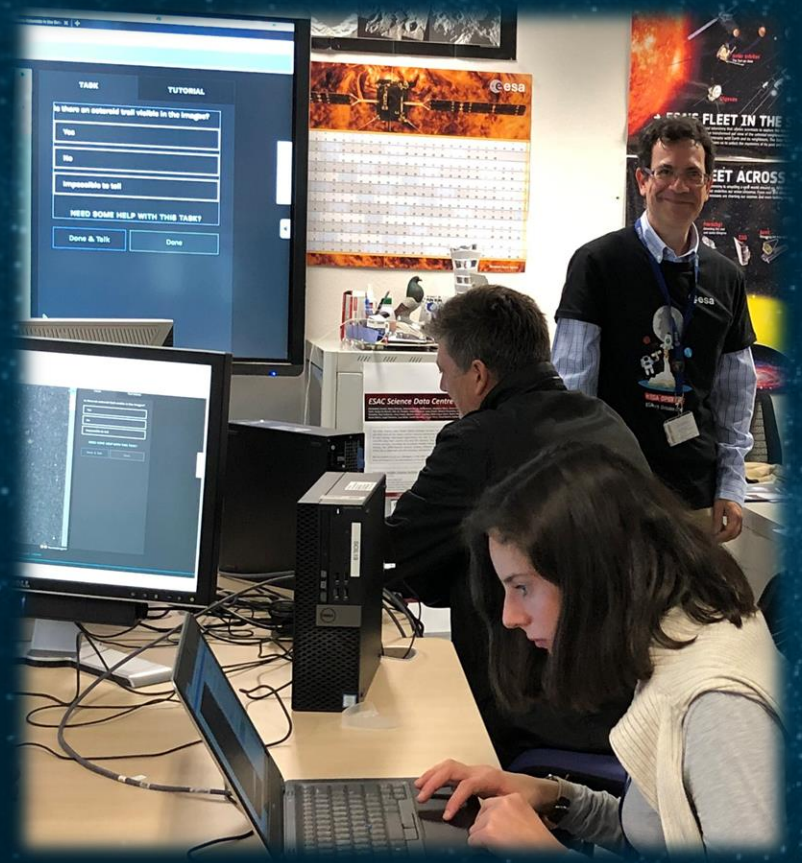
ML: Deep learning on Cloud

Large astronomical archives: eHST



Asteroid discovery

It is mportant to keep humans in the loop



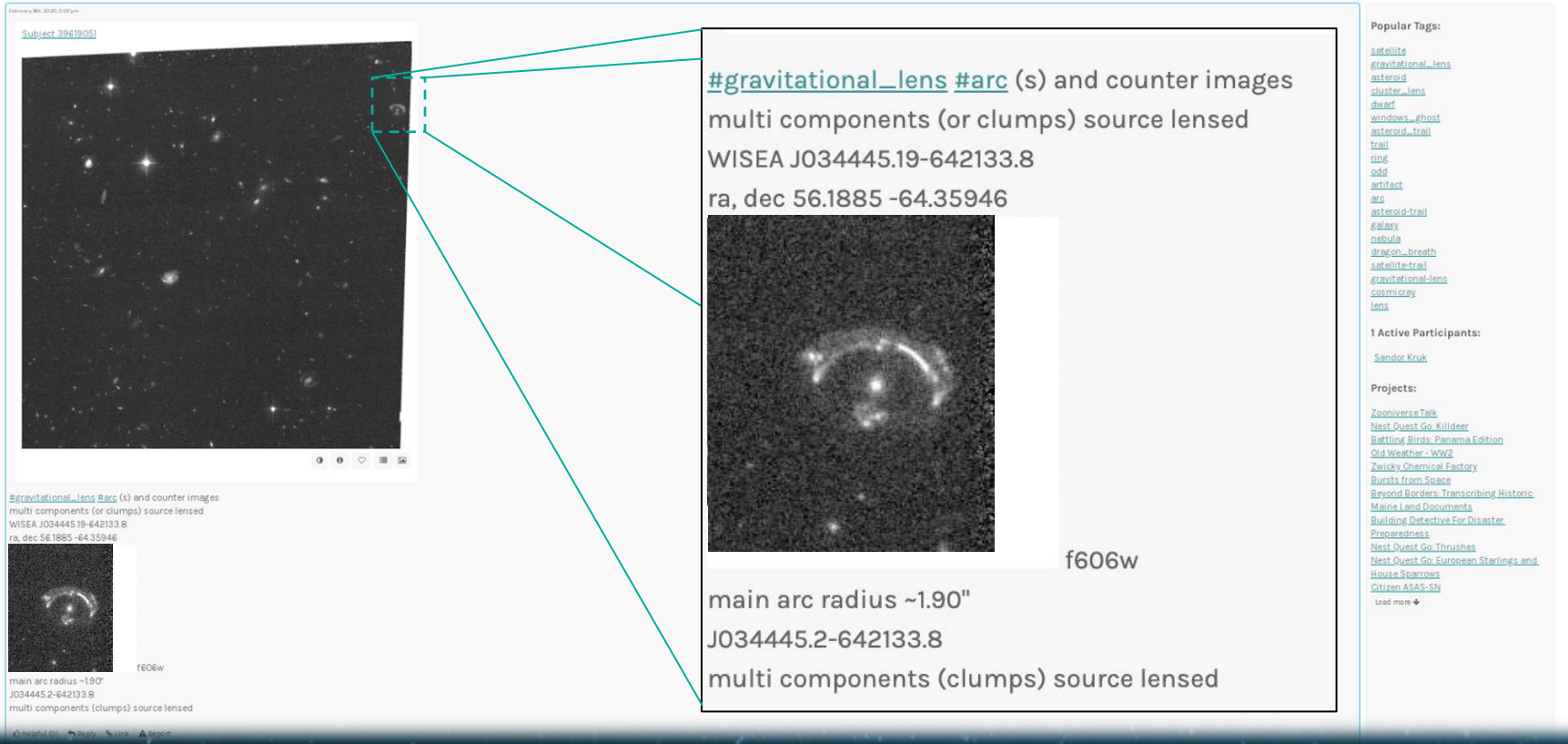
ESAC Open Day, Spain. Credit: C. Arviset

Citizen scientists on Dutch TV for finding new asteroids on www.asteroidhunter.org

Conversations about data: project forum ‘Talk’

https://www.zooniverse.org/projects/sandorkruk/hubble-asteroid-hunter/talk/

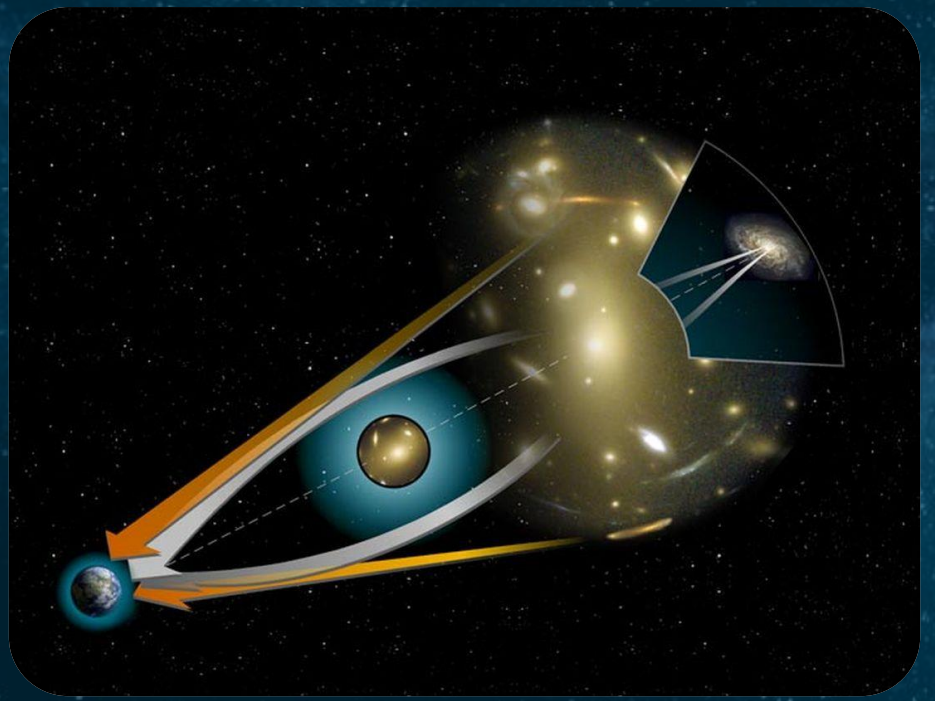
Zooniverse project forum ‘Talk’



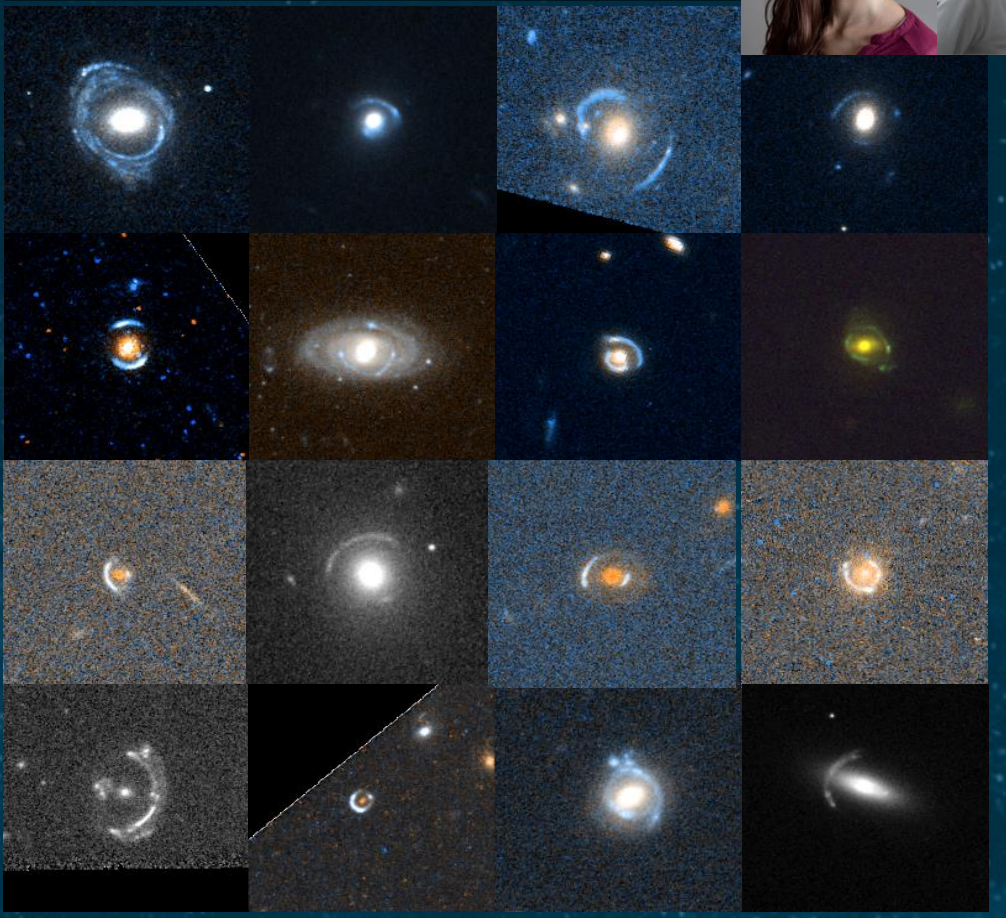
Serendipitous findings: strong gravitational lenses

Having human eyes on the data can lead to new discoveries:

Discovery of 198 new
strong gravitational lenses



New strong gravitational lenses identified in Hubble images



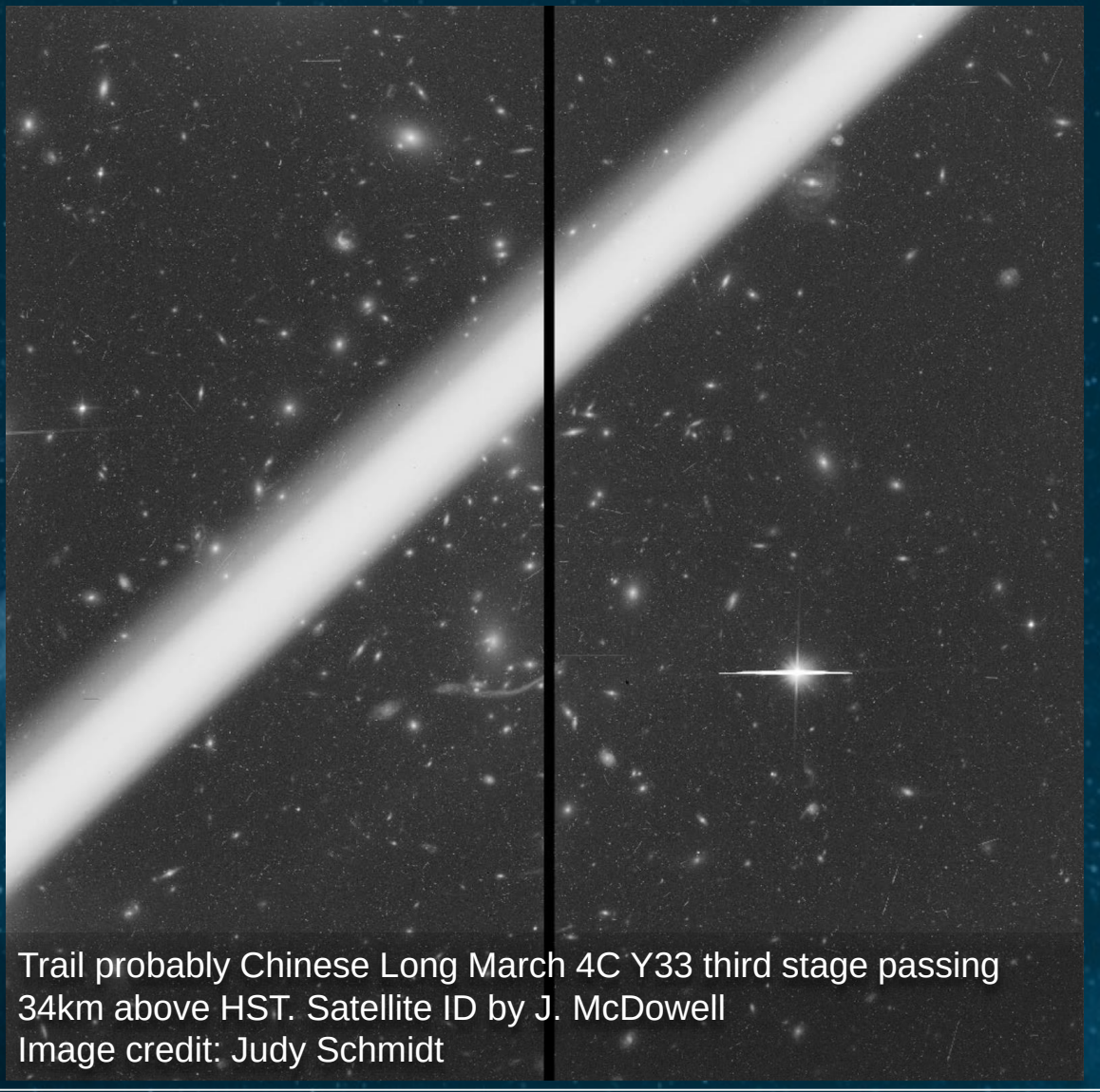
Identifying strong gravitational lenses in HST images with crowdsourcing. E. Garvin and C. Cornen (Garvin et al. 2022, arXiv: 2207.06997)

Sereindipitous findings II

Asteroids observed serendipitously in image of Frontier Fields cluster Abell 370. Credit: NASA, ESA/Hubble

Sereindipitous findings II: satellites in HST images

Asteroids observed serendipitously in image of Frontier Fields cluster Abell 370. Credit: NASA, ESA/Hubble



Trail probably Chinese Long March 4C Y33 third stage passing 34km above HST. Satellite ID by J. McDowell
Image credit: Judy Schmidt

Hubble Asteroid Hunter Talk

Search or enter a #tag

Subject 37652046

Comments:

December 16th 2019, 11:51 am

Not clear what the grey line across the image could be. Perhaps a common #satellite ?

Fauxwise

@Fauxwise

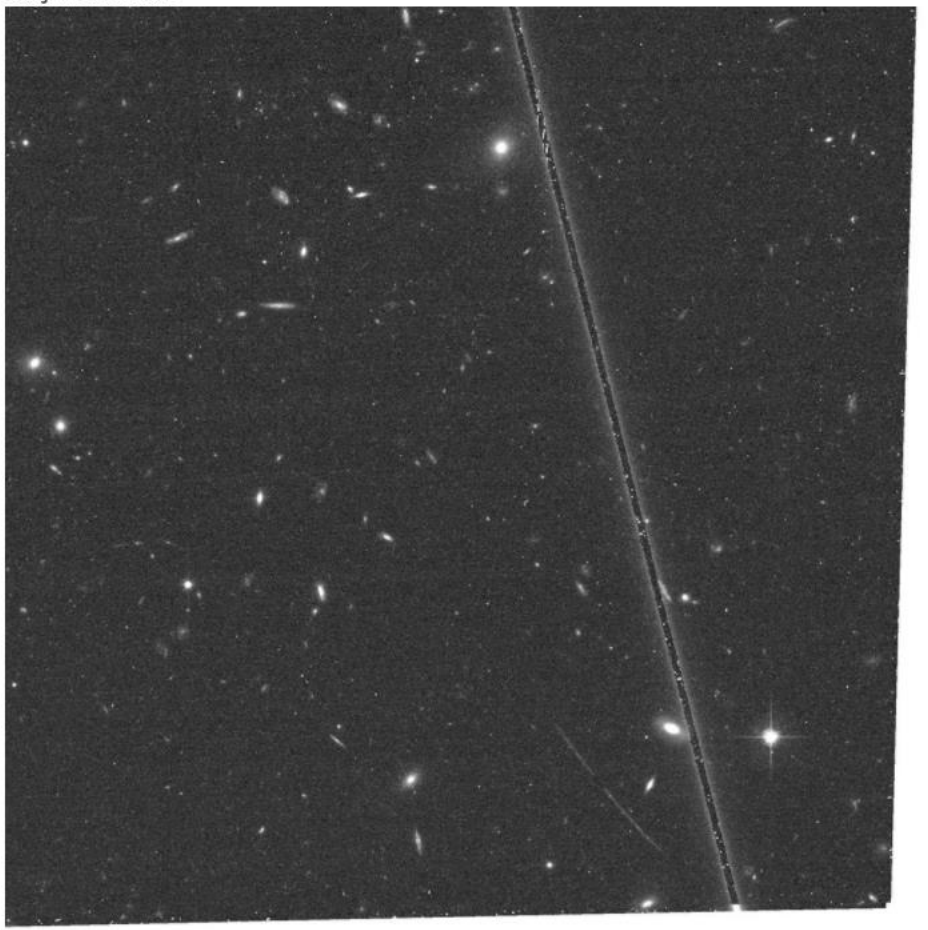
Hubble Asteroid Hunter

Happy international asteroid day! The current d

Hubble Asteroid Hunter Talk

Search or enter a #tag

Subject 42950351



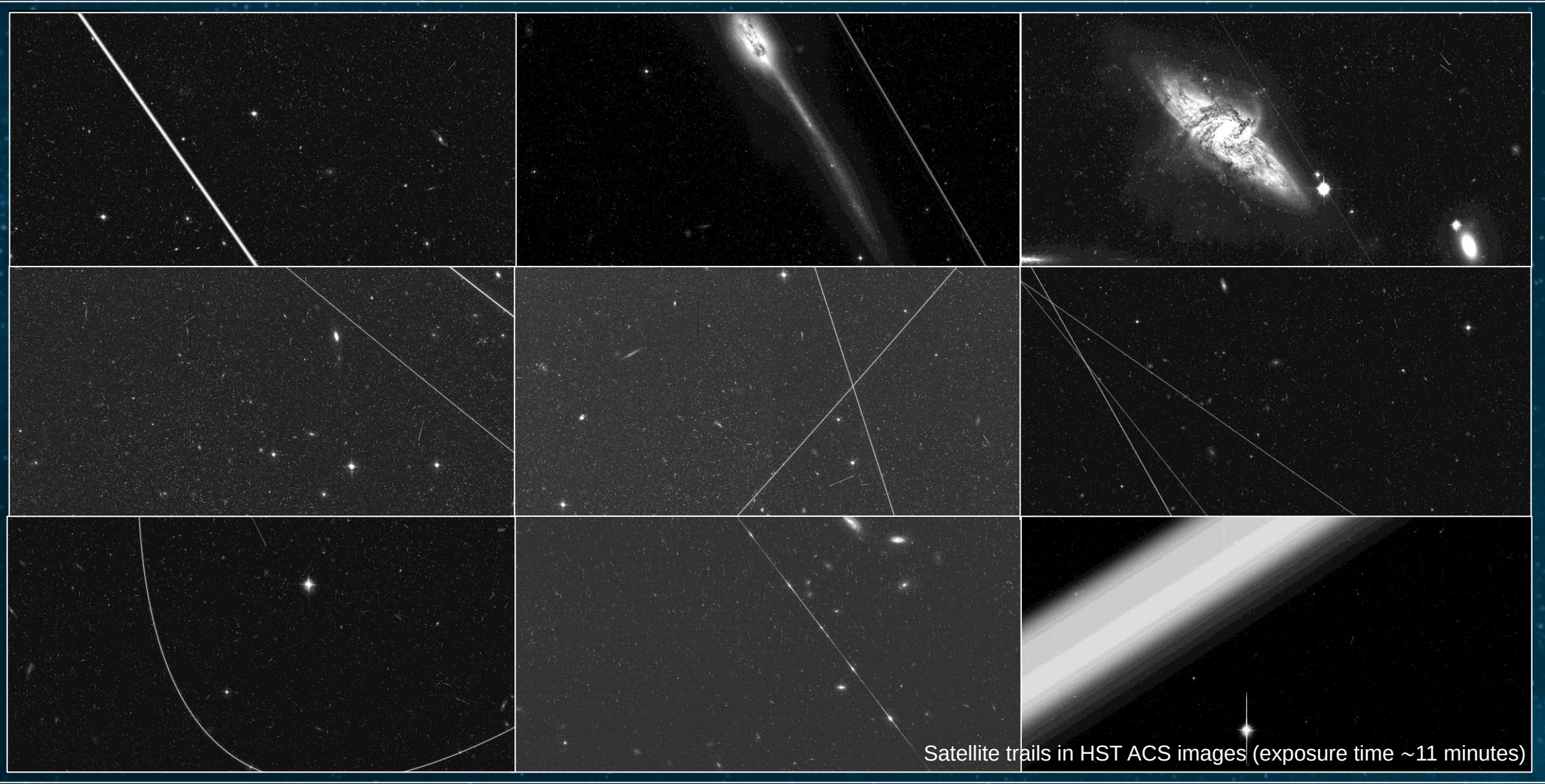
Travaglino

@Travaglino

April 29th 2020, 10:31 pm

#asteroid and #satellite

Satellite trails identified with AutoML in Hubble images



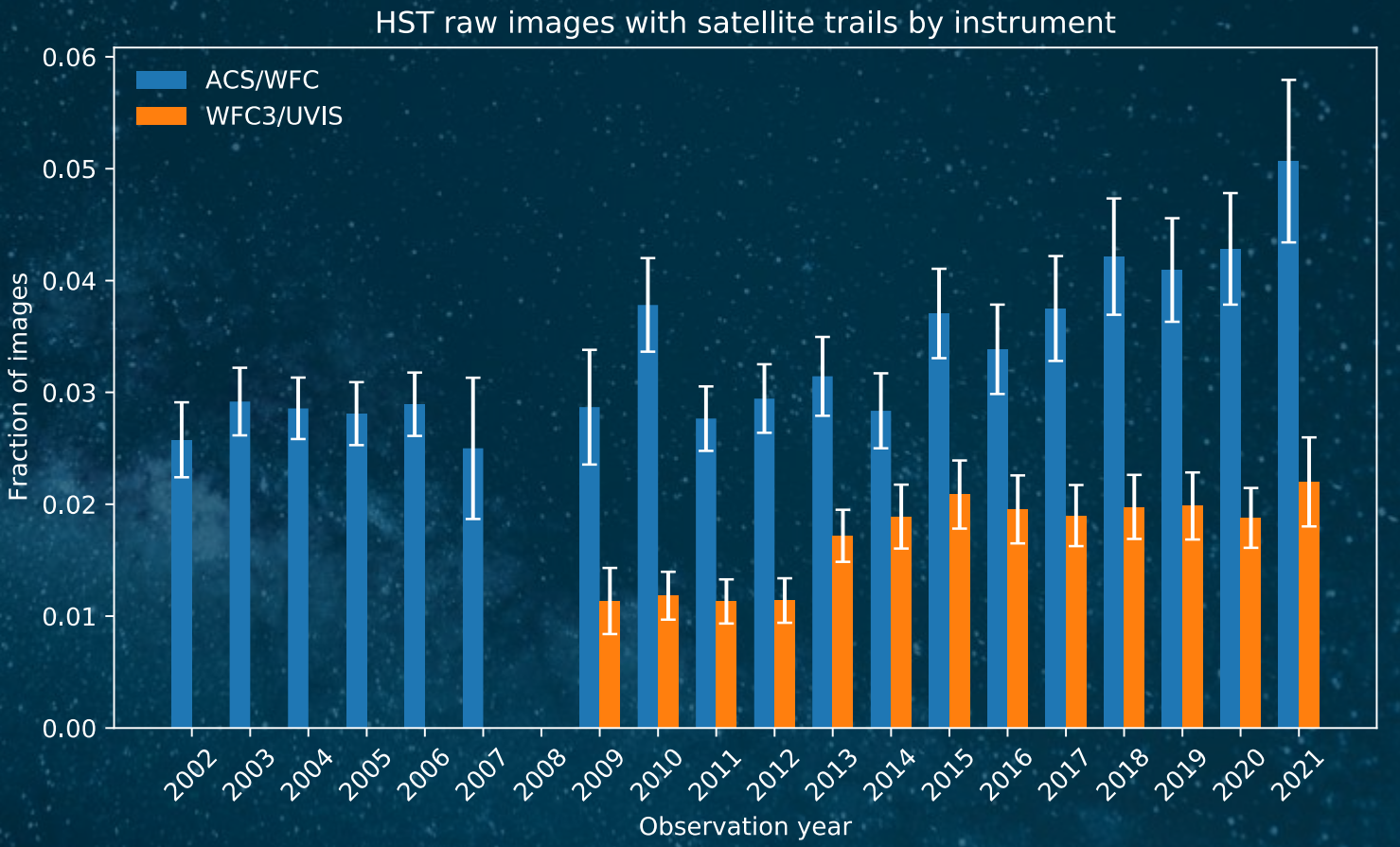
Satellite trails in HST ACS images (exposure time ~11 minutes)

Fraction of HST images crossed by satellites

2.7% of individual exposures of 11 min crossed by satellites (3.2% ACS and 1.7% of WFC3 impacted)

Fraction of images crossed by satellites ~ doubled in the last 20 years, matching the increase in satellite numbers

ACS FoV = 202”x202”
WFC3 FoV = 160”x160”



Kruk et al. 2023, *The impact of satellite trails on Hubble Space Telescope observations*

Take home messages:

Artificial intelligence and crowdsourcing

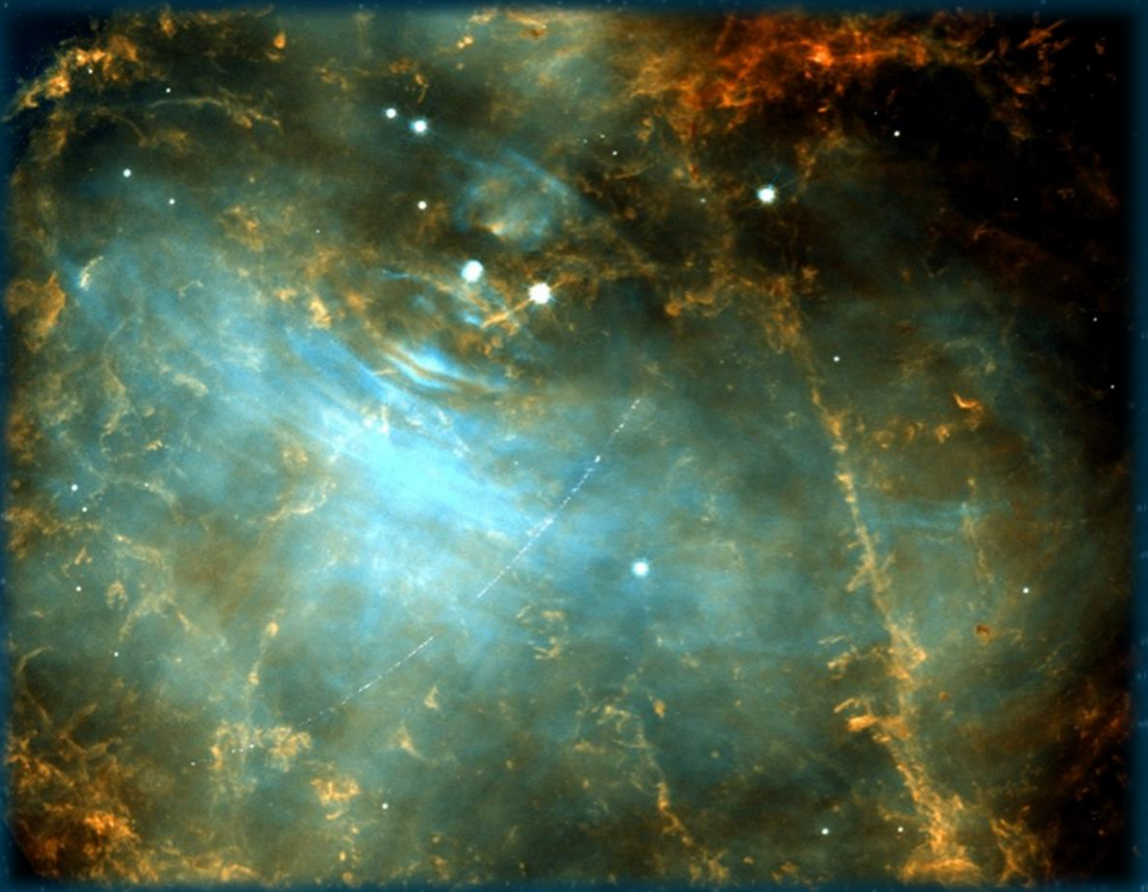
New tools are needed to analyse and mine the increasingly large datasets. Human (e.g. crowdsourcing) and machine collaboration is important to create AI- ready datasets and avoid the garbage-in garbage-out problem.

Importance of conversations about data

Project forums provide the ideal setup for conversations between scientists and volunteers. It can lead to new discoveries: strong lenses and artificial satellites.

Necessity of data platforms

Google Cloud provided the framework to train deep learning algorithms and predict the presence of asteroids in HST images.



Foreground asteroid passing in front of the Crab Nebula, identified in the Hubble Asteroid Hunter project. Credit: ESA/Hubble & NASA, M. Thévenot

Take home messages:

Artificial intelligence and crowdsourcing

New tools are needed to analyse and mine the increasingly large datasets. Human (e.g. crowdsourcing) and machine collaboration is important to create AI- ready datasets and avoid the garbage-in garbage-out problem.

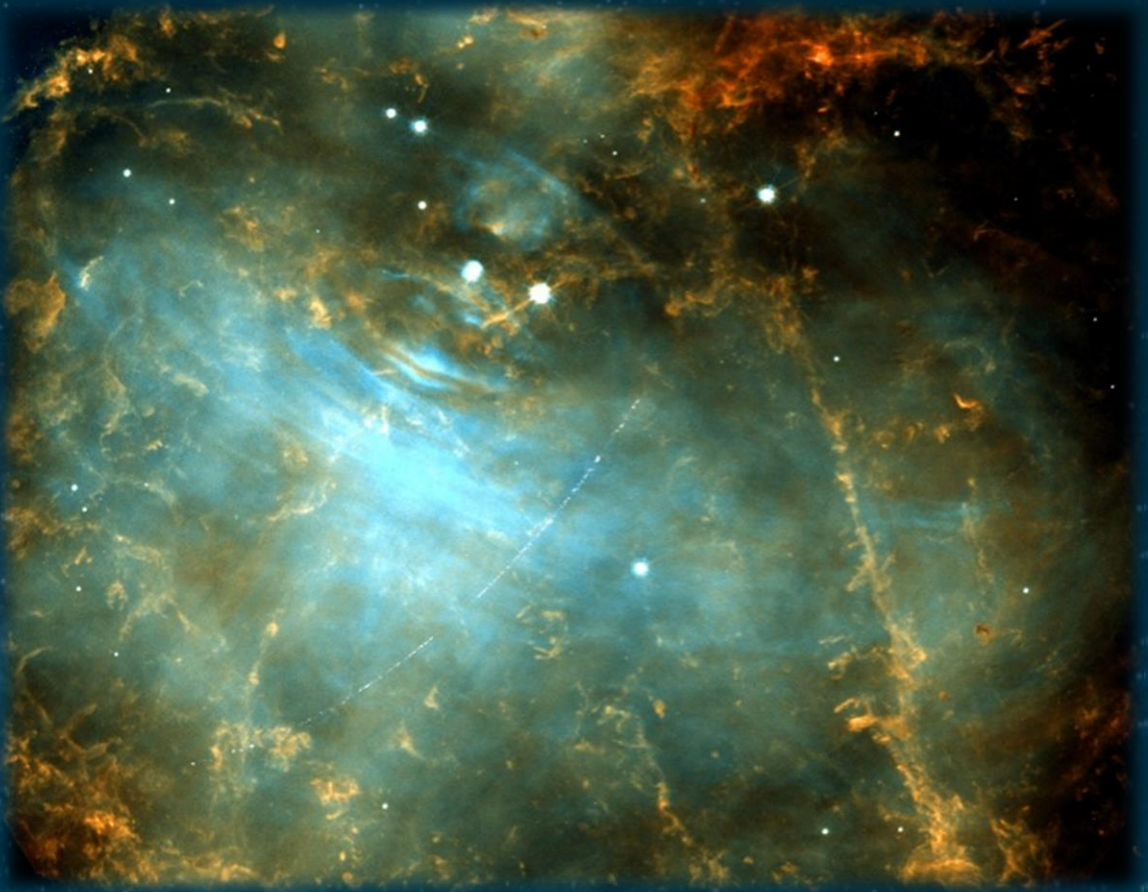
Importance of conversations about data

Project forums provide the ideal setup for conversations between scientists and volunteers. It can lead to new discoveries: strong lenses and artificial satellites.

Necessity of data platforms

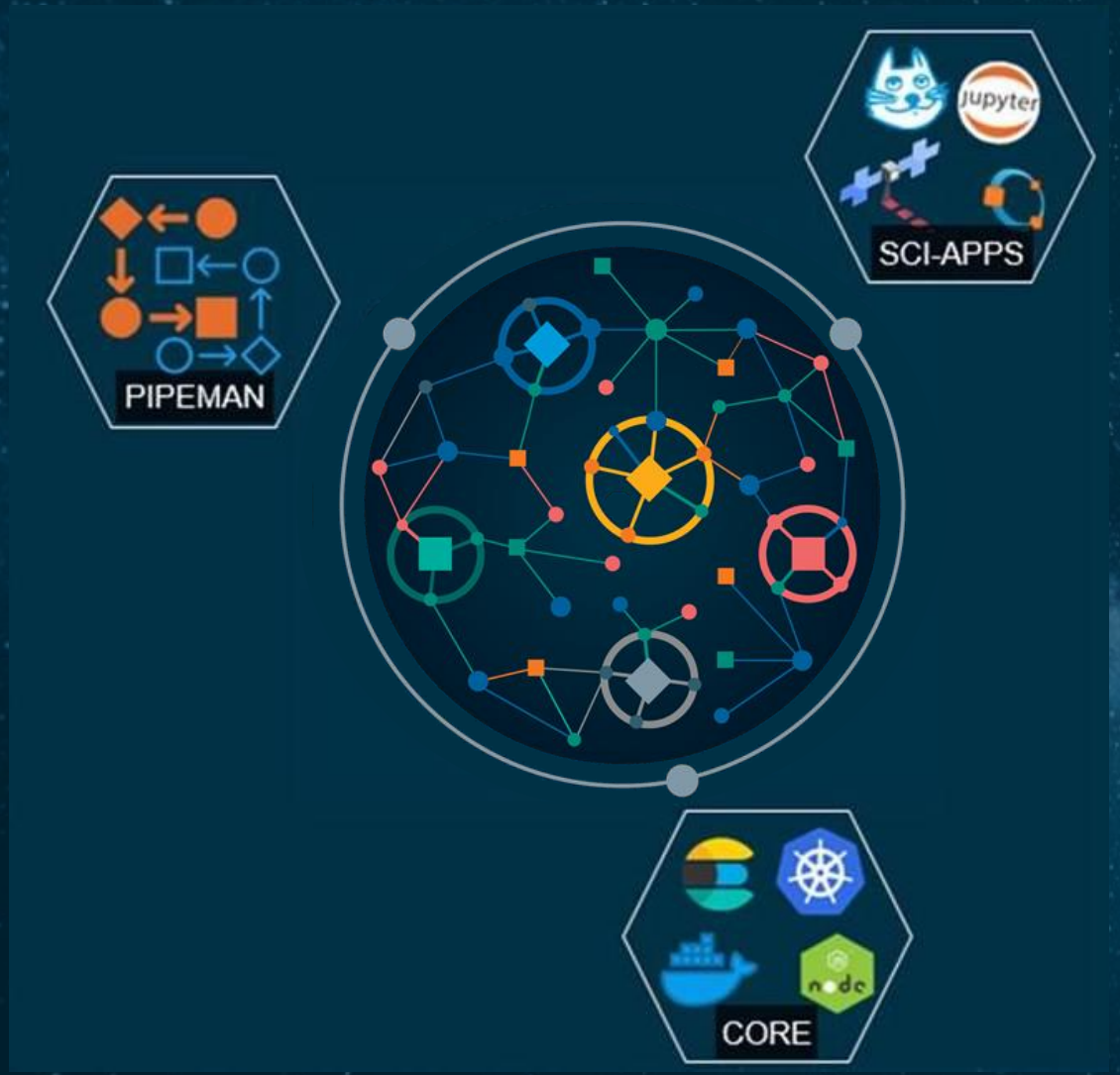
Google Cloud provided the framework to train deep learning algorithms and predict the presence of asteroids in HST images.

…but data processing and data transfer overheads are significant

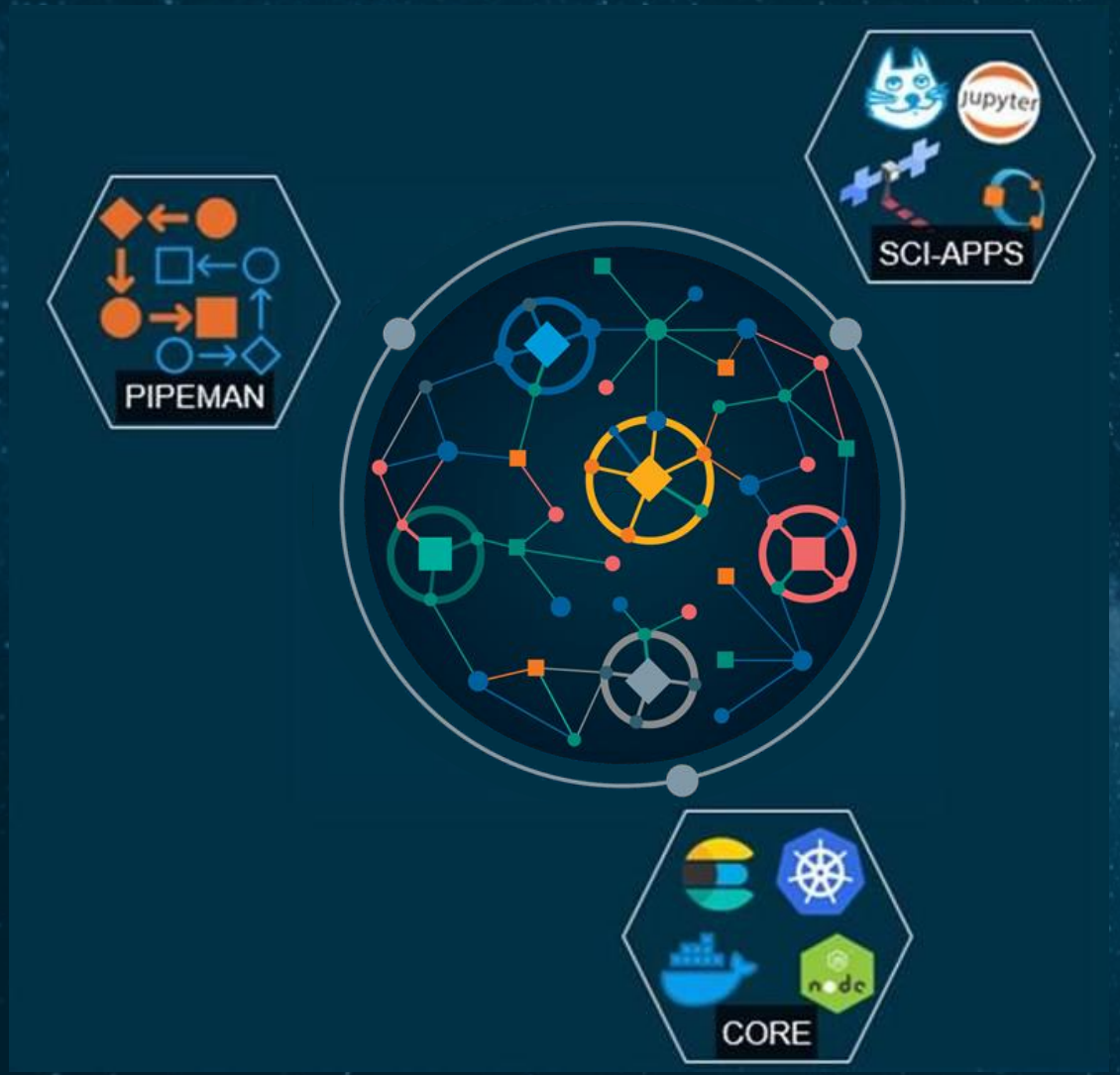


Foreground asteroid passing in front of the Crab Nebula, identified in the Hubble Asteroid Hunter project. Credit: ESA/Hubble & NASA, M. Thévenot

Science platforms – ESA Datalabs

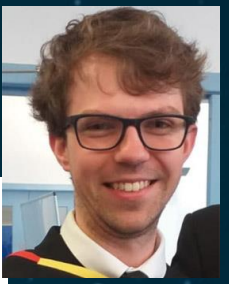


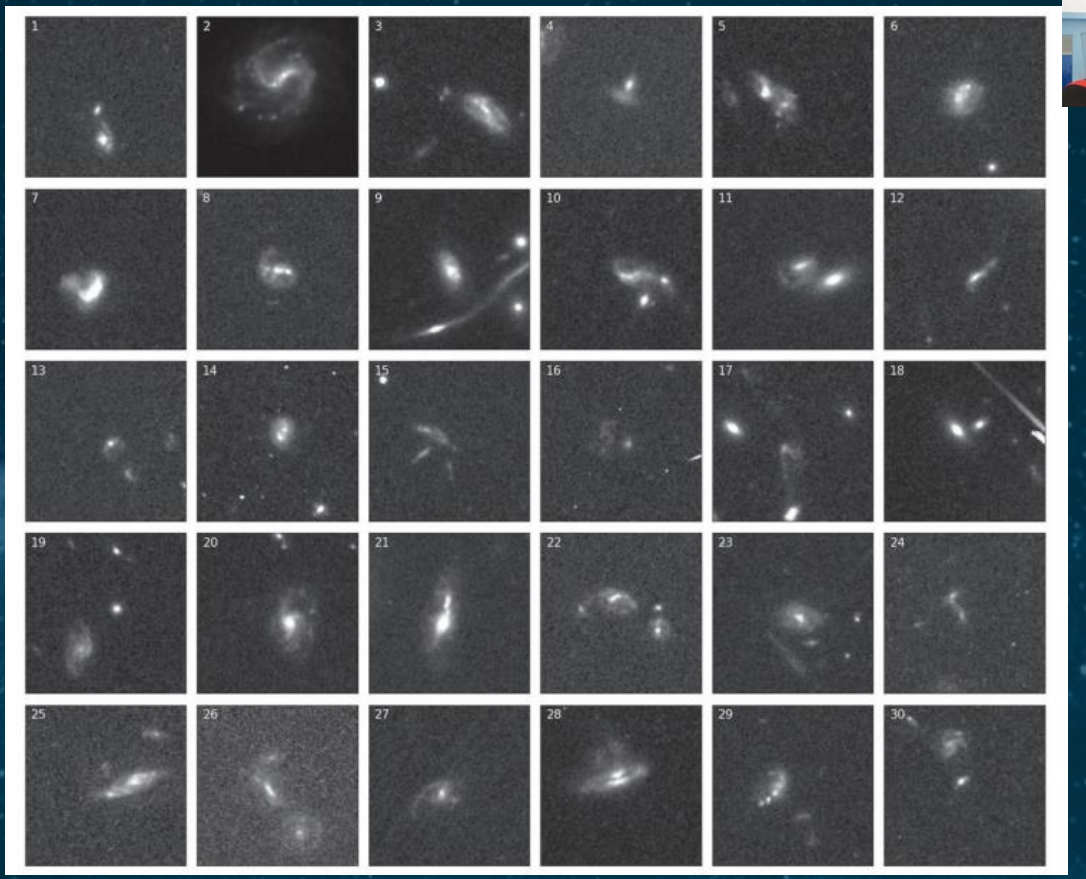
Science platforms – ESA Datalabs



One example use case of ESA Datalabs:

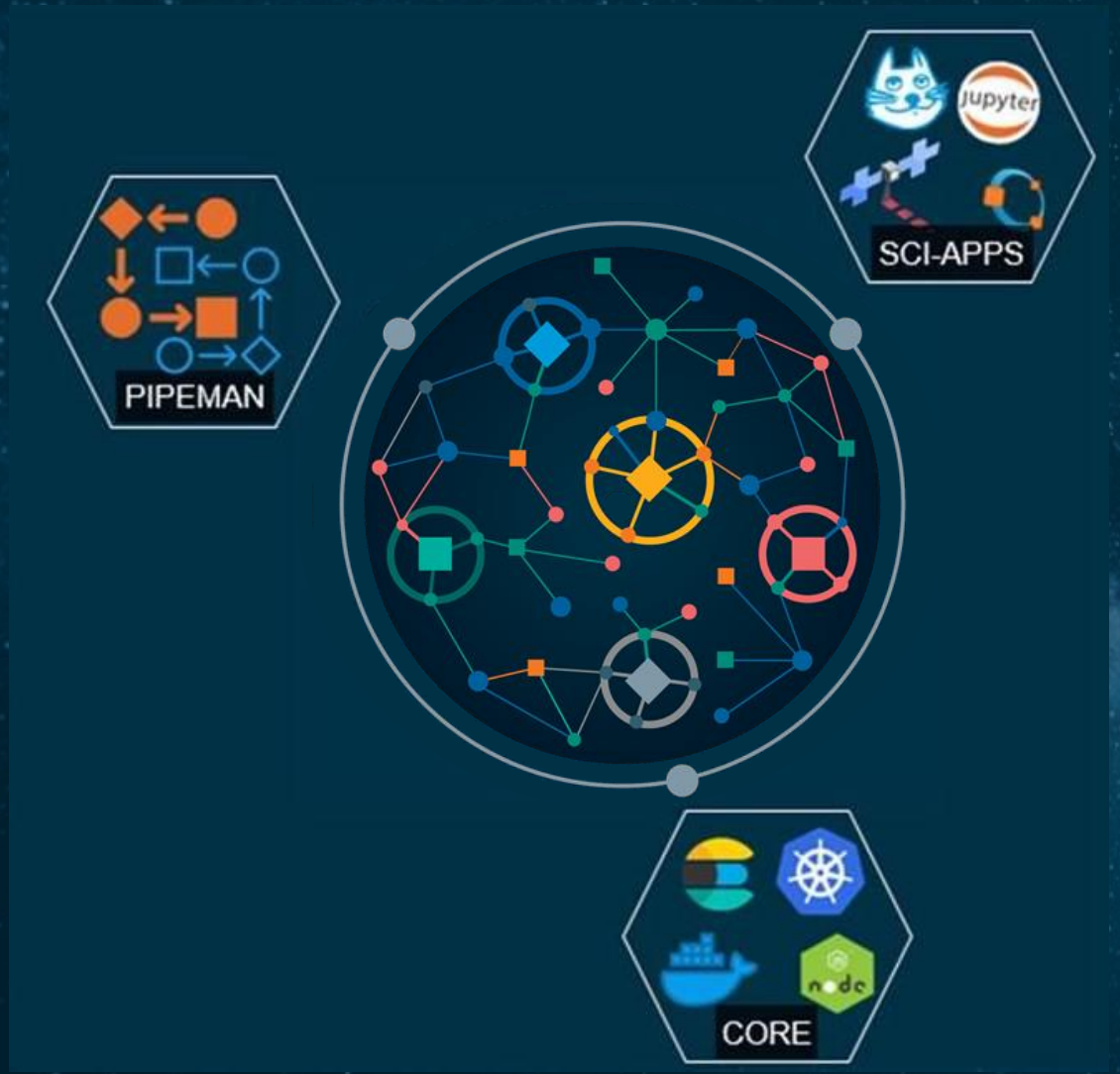
Harnessing the Hubble Space Telescope Archives: A Catalogue of 21,926 Interacting Galaxies





O'Ryan et al. 2023, arXiv:2303.00366

Science platforms – ESA Datalabs



Session II

Time: Thursday 11th May 14:00–15:30 CEST **[session #15]**

Speaker	Title
Vicente Navarro ESA	ESA Datalabs, an Open Digital Platform for Innovation and Collaboration in Space Science
Mario Juric University of Washington	HiPSCat: extending HiPS to for highly scalable large-scale catalog analyses
Robert Nikutta NOIRLab	NOIRLab's Astro Data Lab science platform
Brian Major CADC	CANFAR Science Platform and the IVOA
Gregory Dubois-Felsmann Rubin observatory	Rubin observatory and the IVOA
Dave Morris Edinburgh University	Gaia Data Mining platform
All	Discussion