Hossen Teimoorinia

Patrick Dowler

(Canadian Astronomy Data Centre)

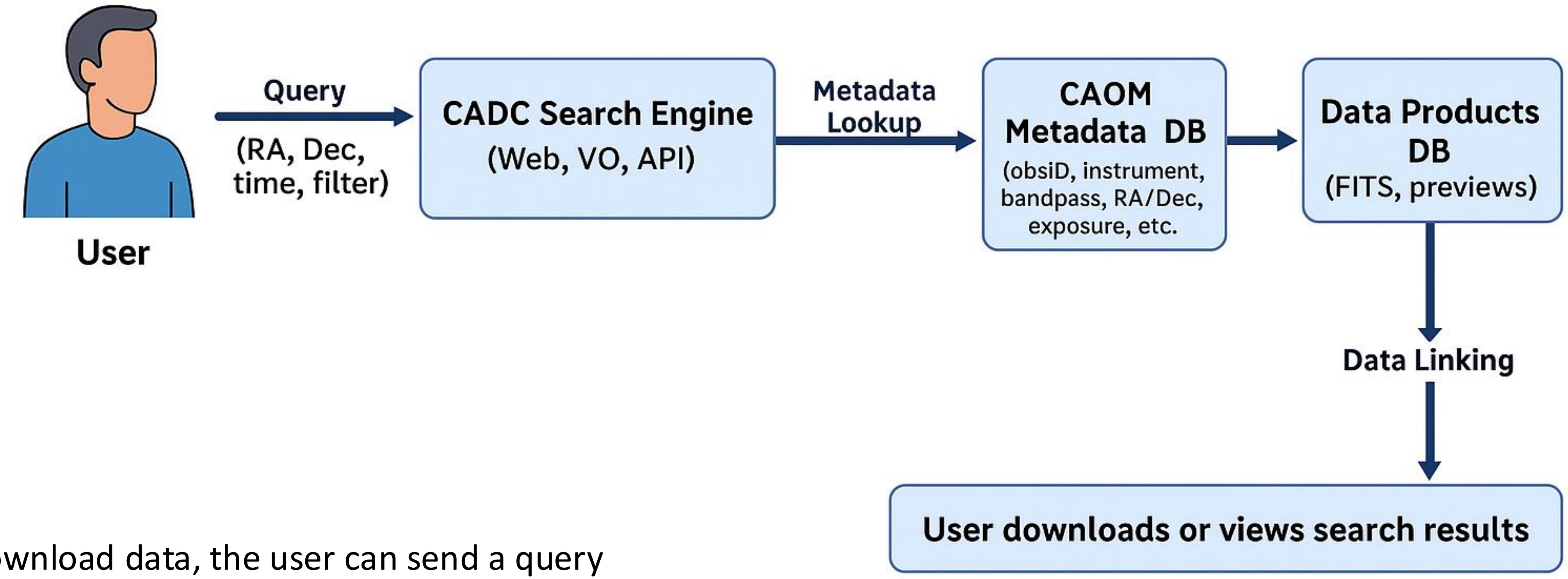
ADASS, Görlitz (DE)

Nov, 2025

CAOM-AI: Content-Based Image Search System

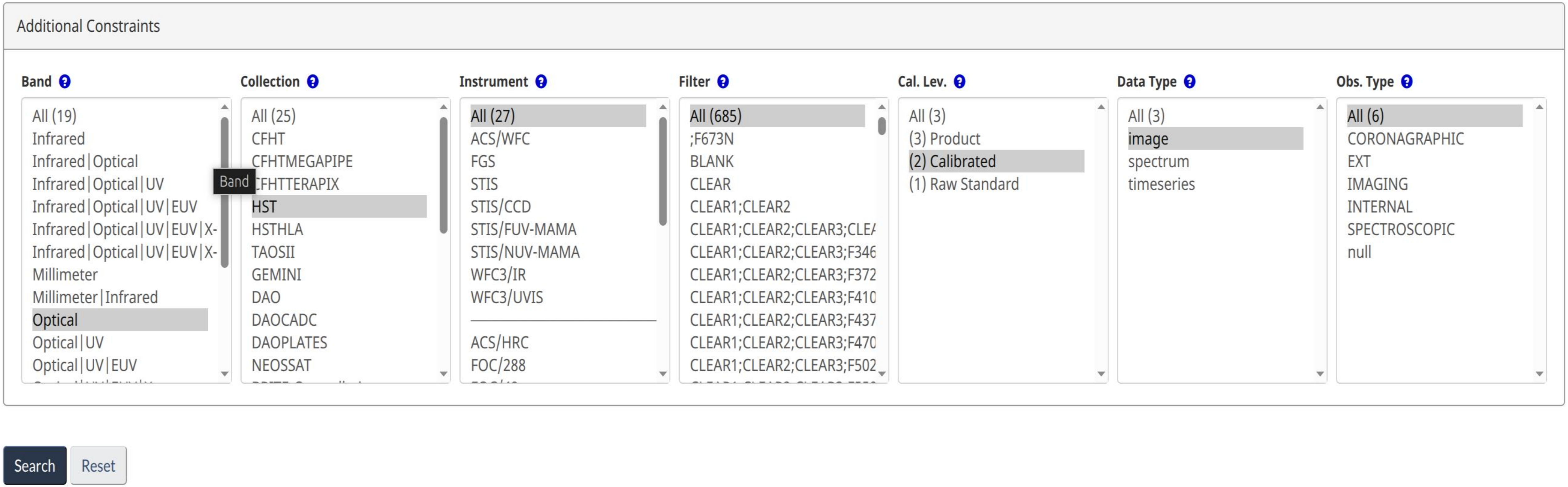
From Unsupervised to Self-Supervised Representation Learning:

Implementing Content-Based Search in Astronomical Data Archives

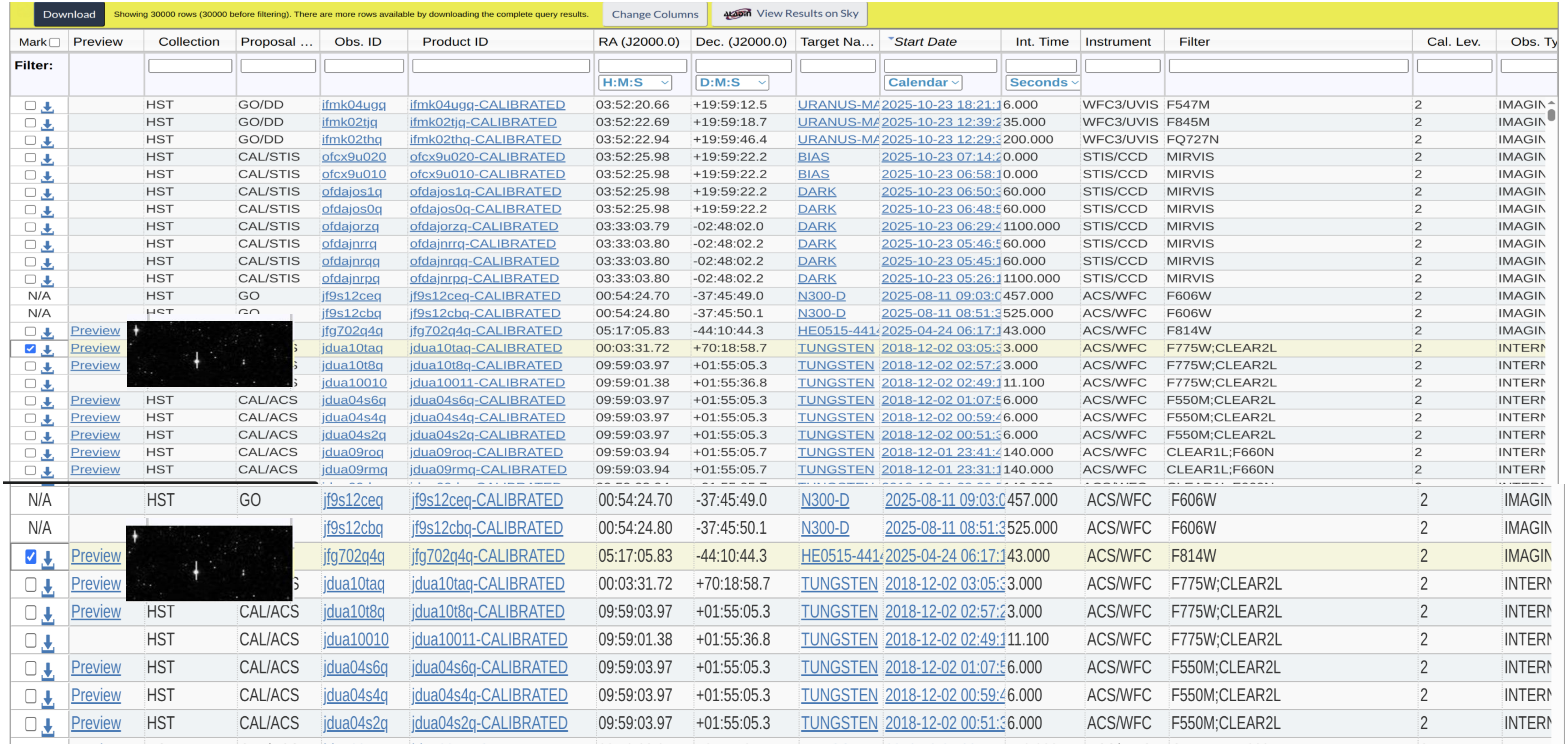


To download data, the user can send a query to the CADC, and the CAOM model can match metadata with relevant images, returning a list of downloadable images.

An example of CADC Advanced Search



This is an example of what you could see in CADC, the advanced search page, choose your metadata like filters, instruments and then click search



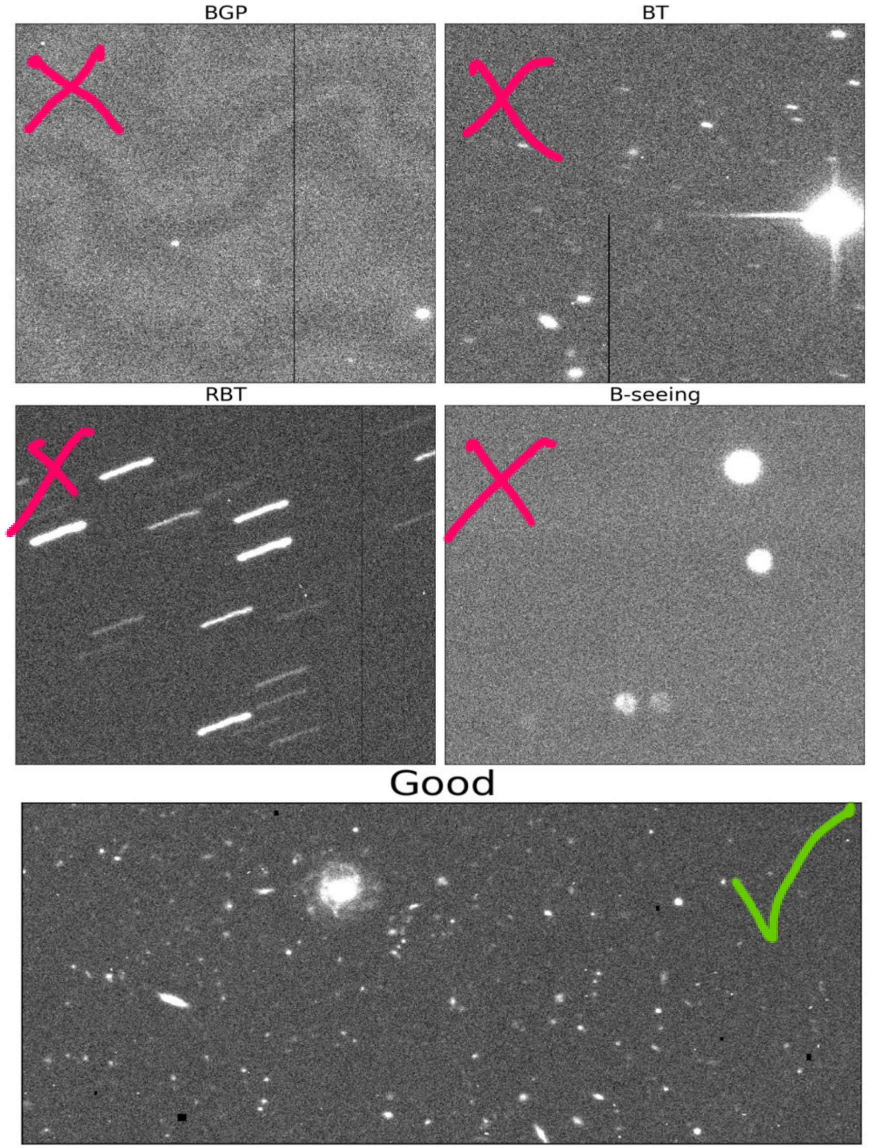
Download | Showing 30000 rows (30000 before filtering). There are more rows available by downloading the complete query results. | Change Columns | View Results on Sky

Mark	Preview	Collection	Proposal ...	Obs. ID	Product ID	RA (J2000.0)	Dec. (J2000.0)	Target Na...	Start Date	Int. Time	Instrument	Filter	Cal. Lev.	Obs. Ty
Filter:						H:M:S	D:M:S		Calendar	Seconds				
		HST	GO/DD	ifmk04ugq	ifmk04ugq-CALIBRATED	03:52:20.66	+19:59:12.5	URANUS-MA	2025-10-23 18:21:1	6.000	WFC3/UVIS	F547M	2	IMAGIN
		HST	GO/DD	ifmk02tjq	ifmk02tjq-CALIBRATED	03:52:22.69	+19:59:18.7	URANUS-MA	2025-10-23 12:39:2	35.000	WFC3/UVIS	F845M	2	IMAGIN
		HST	GO/DD	ifmk02thq	ifmk02thq-CALIBRATED	03:52:22.94	+19:59:46.4	URANUS-MA	2025-10-23 12:29:3	200.000	WFC3/UVIS	FQ727N	2	IMAGIN
		HST	CAL/STIS	ofcx9u020	ofcx9u020-CALIBRATED	03:52:25.98	+19:59:22.2	BIAS	2025-10-23 07:14:2	0.000	STIS/CCD	MIRVIS	2	IMAGIN
		HST	CAL/STIS	ofcx9u010	ofcx9u010-CALIBRATED	03:52:25.98	+19:59:22.2	BIAS	2025-10-23 06:58:1	0.000	STIS/CCD	MIRVIS	2	IMAGIN
		HST	CAL/STIS	ofdajos1q	ofdajos1q-CALIBRATED	03:52:25.98	+19:59:22.2	DARK	2025-10-23 06:50:3	60.000	STIS/CCD	MIRVIS	2	IMAGIN
		HST	CAL/STIS	ofdajos0q	ofdajos0q-CALIBRATED	03:52:25.98	+19:59:22.2	DARK	2025-10-23 06:48:5	60.000	STIS/CCD	MIRVIS	2	IMAGIN
		HST	CAL/STIS	ofdajorzq	ofdajorzq-CALIBRATED	03:33:03.79	-02:48:02.0	DARK	2025-10-23 06:29:4	1100.000	STIS/CCD	MIRVIS	2	IMAGIN
		HST	CAL/STIS	ofdajnrrq	ofdajnrrq-CALIBRATED	03:33:03.80	-02:48:02.2	DARK	2025-10-23 05:46:5	60.000	STIS/CCD	MIRVIS	2	IMAGIN
		HST	CAL/STIS	ofdajnrqq	ofdajnrqq-CALIBRATED	03:33:03.80	-02:48:02.2	DARK	2025-10-23 05:45:1	60.000	STIS/CCD	MIRVIS	2	IMAGIN
		HST	CAL/STIS	ofdajnrpq	ofdajnrpq-CALIBRATED	03:33:03.80	-02:48:02.2	DARK	2025-10-23 05:26:1	1100.000	STIS/CCD	MIRVIS	2	IMAGIN
N/A		HST	GO	jf9s12ceq	jf9s12ceq-CALIBRATED	00:54:24.70	-37:45:49.0	N300-D	2025-08-11 09:03:0	457.000	ACS/WFC	F606W	2	IMAGIN
N/A		HST	GO	jf9s12cbq	jf9s12cbq-CALIBRATED	00:54:24.80	-37:45:50.1	N300-D	2025-08-11 08:51:3	525.000	ACS/WFC	F606W	2	IMAGIN
	Preview			jfg702q4q	jfg702q4q-CALIBRATED	05:17:05.83	-44:10:44.3	HE0515-441	2025-04-24 06:17:1	43.000	ACS/WFC	F814W	2	IMAGIN
☑	Preview		S	jdua10taq	jdua10taq-CALIBRATED	00:03:31.72	+70:18:58.7	TUNGSTEN	2018-12-02 03:05:3	3.000	ACS/WFC	F775W;CLEAR2L	2	INTERN
	Preview		S	jdua10t8q	jdua10t8q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 02:57:2	3.000	ACS/WFC	F775W;CLEAR2L	2	INTERN
			S	jdua10010	jdua10011-CALIBRATED	09:59:01.38	+01:55:36.8	TUNGSTEN	2018-12-02 02:49:1	11.100	ACS/WFC	F775W;CLEAR2L	2	INTERN
	Preview	HST	CAL/ACS	jdua04s6q	jdua04s6q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 01:07:5	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
	Preview	HST	CAL/ACS	jdua04s4q	jdua04s4q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 00:59:4	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
	Preview	HST	CAL/ACS	jdua04s2q	jdua04s2q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 00:51:3	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
	Preview	HST	CAL/ACS	jdua09roq	jdua09roq-CALIBRATED	09:59:03.94	+01:55:05.7	TUNGSTEN	2018-12-01 23:41:4	140.000	ACS/WFC	CLEAR1L;F660N	2	INTERN
	Preview	HST	CAL/ACS	jdua09rmq	jdua09rmq-CALIBRATED	09:59:03.94	+01:55:05.7	TUNGSTEN	2018-12-01 23:31:1	140.000	ACS/WFC	CLEAR1L;F660N	2	INTERN

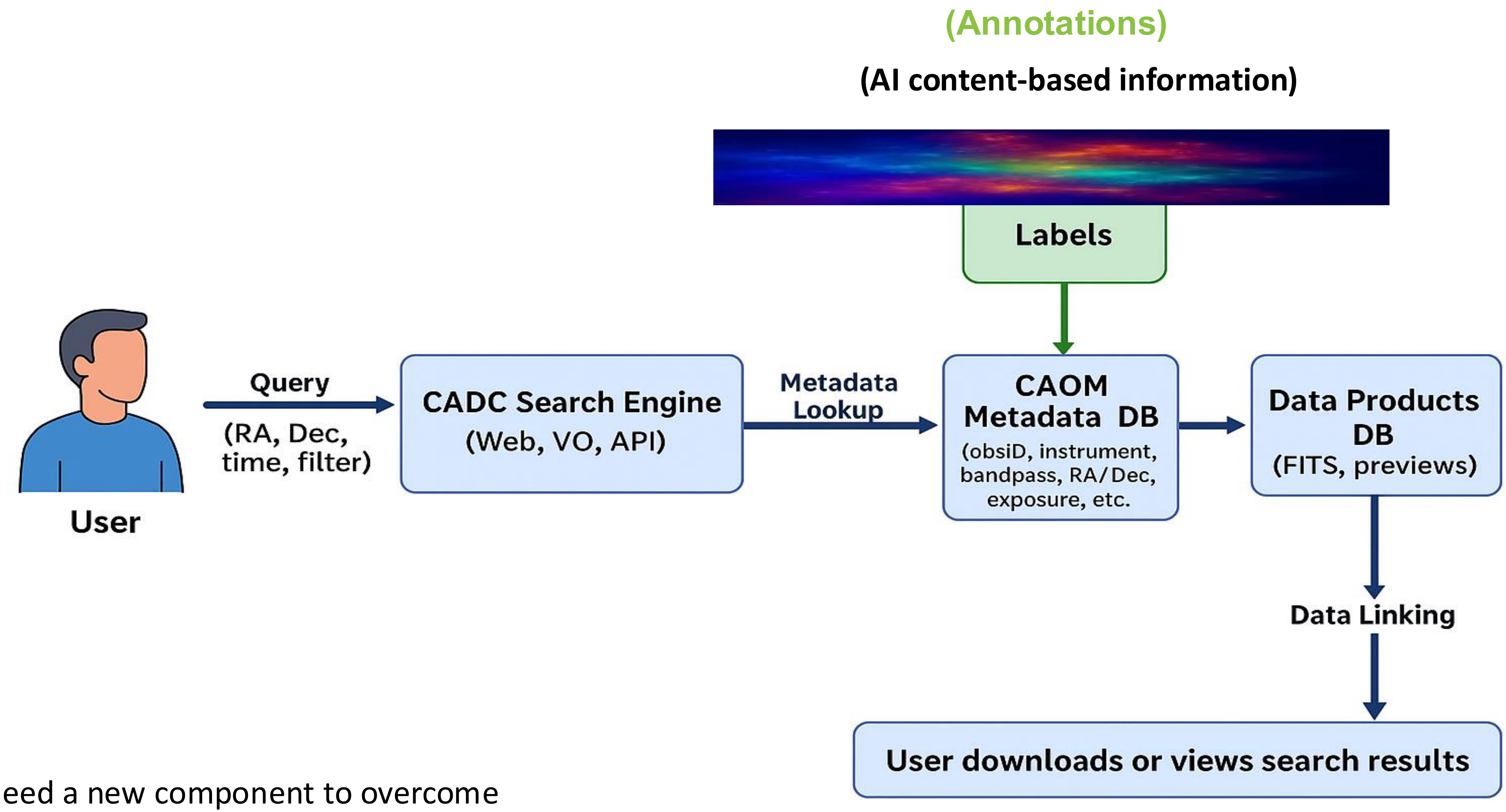
N/A		HST	GO	jf9s12ceq	jf9s12ceq-CALIBRATED	00:54:24.70	-37:45:49.0	N300-D	2025-08-11 09:03:0	457.000	ACS/WFC	F606W	2	IMAGIN
N/A				jf9s12cbq	jf9s12cbq-CALIBRATED	00:54:24.80	-37:45:50.1	N300-D	2025-08-11 08:51:3	525.000	ACS/WFC	F606W	2	IMAGIN
☑	Preview			jfg702q4q	jfg702q4q-CALIBRATED	05:17:05.83	-44:10:44.3	HE0515-441	2025-04-24 06:17:1	43.000	ACS/WFC	F814W	2	IMAGIN
	Preview		S	jdua10taq	jdua10taq-CALIBRATED	00:03:31.72	+70:18:58.7	TUNGSTEN	2018-12-02 03:05:3	3.000	ACS/WFC	F775W;CLEAR2L	2	INTERN
	Preview	HST	CAL/ACS	jdua10t8q	jdua10t8q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 02:57:2	3.000	ACS/WFC	F775W;CLEAR2L	2	INTERN
		HST	CAL/ACS	jdua10010	jdua10011-CALIBRATED	09:59:01.38	+01:55:36.8	TUNGSTEN	2018-12-02 02:49:1	11.100	ACS/WFC	F775W;CLEAR2L	2	INTERN
	Preview	HST	CAL/ACS	jdua04s6q	jdua04s6q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 01:07:5	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
	Preview	HST	CAL/ACS	jdua04s4q	jdua04s4q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 00:59:4	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
	Preview	HST	CAL/ACS	jdua04s2q	jdua04s2q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 00:51:3	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN

This is an example of what you could see. More than 30,000 downloadable images. You can see some of them to explore them, but it is still very challenging without any prior knowledge to know which one should be downloaded

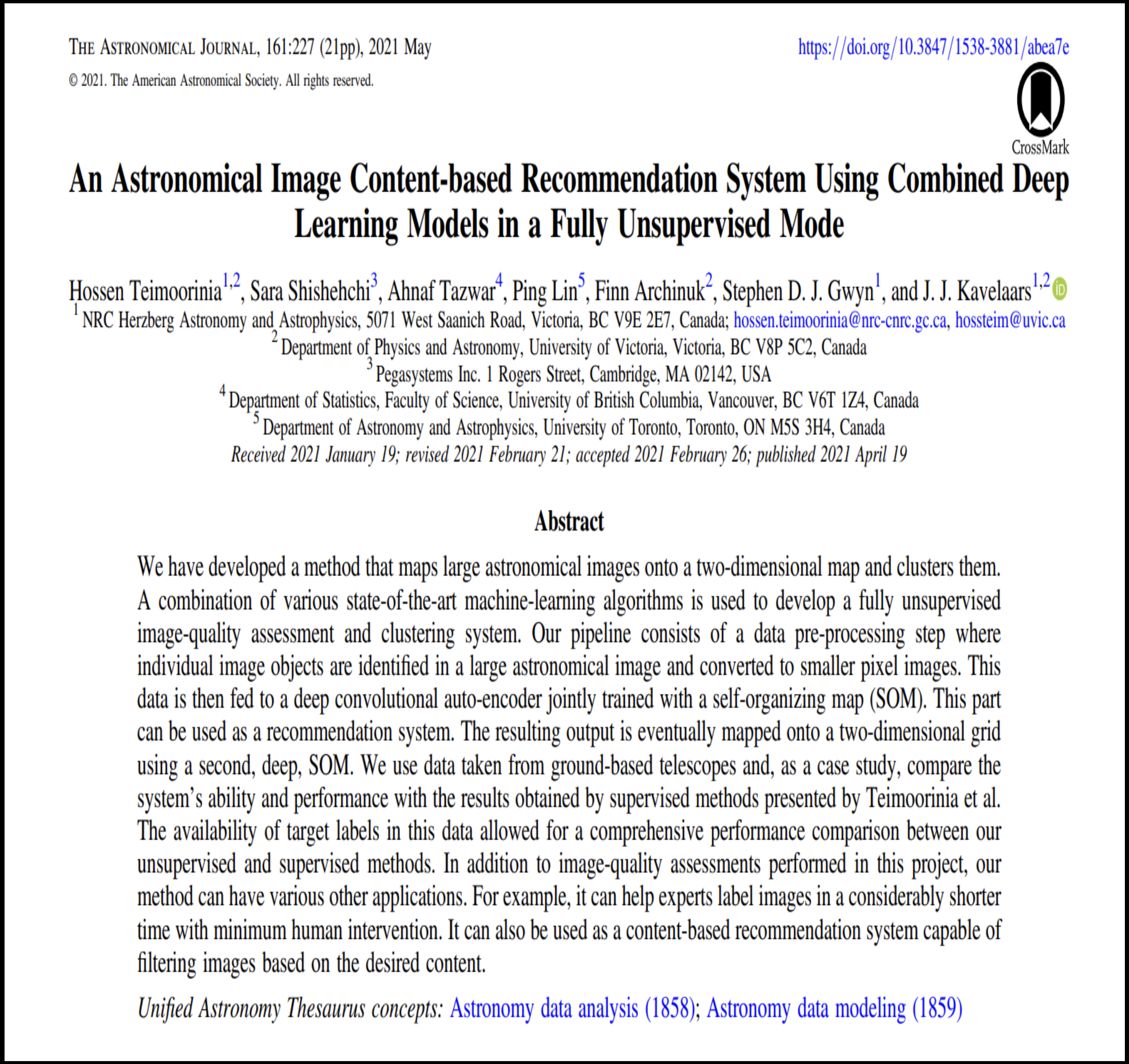
Sometimes you would end up downloading garbage, the images that you don't need



No content-based information.



We need a new component to overcome this problem, and would like to talk about this interpretable component here.


The Astronomical Journal, 161:227 (21pp), 2021 May
https://doi.org/10.3847/1538-3881/abea7e

An Astronomical Image Content-based Recommendation System Using Combined Deep Learning Models in a Fully Unsupervised Mode

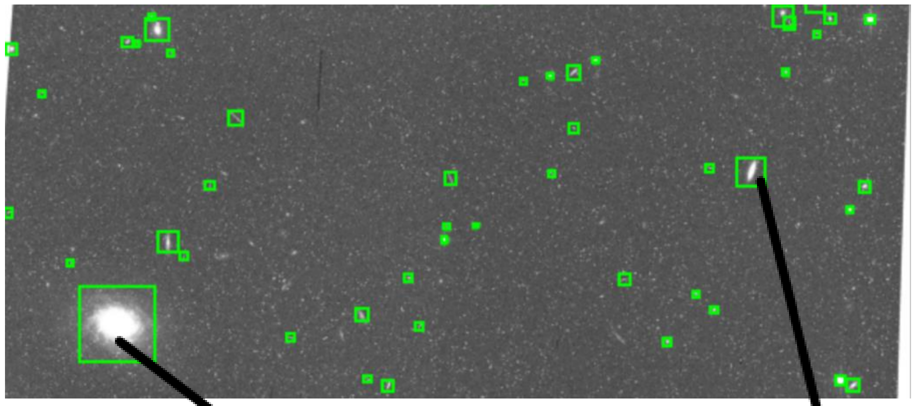
Hossen Teimoorinia[1,2], Sara Shishehchi[3], Ahnaf Tazwar[4], Ping Lin[5], Finn Archinuk[2], Stephen D. J. Gwyn[1], and J. J. Kavelaars[1,2]
[1] NRC Herzberg Astronomy and Astrophysics, 5071 West Saanich Road, Victoria, BC V9E 2E7, Canada; hossen.teimoorinia@nrc-cnrc.gc.ca, hossteim@uvic.ca
[2] Department of Physics and Astronomy, University of Victoria, Victoria, BC V8P 5C2, Canada
[3] Pegasystems Inc. 1 Rogers Street, Cambridge, MA 02142, USA
[4] Department of Statistics, Faculty of Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[5] Department of Astronomy and Astrophysics, University of Toronto, Toronto, ON M5S 3H4, Canada
Received 2021 January 19; revised 2021 February 21; accepted 2021 February 26; published 2021 April 19

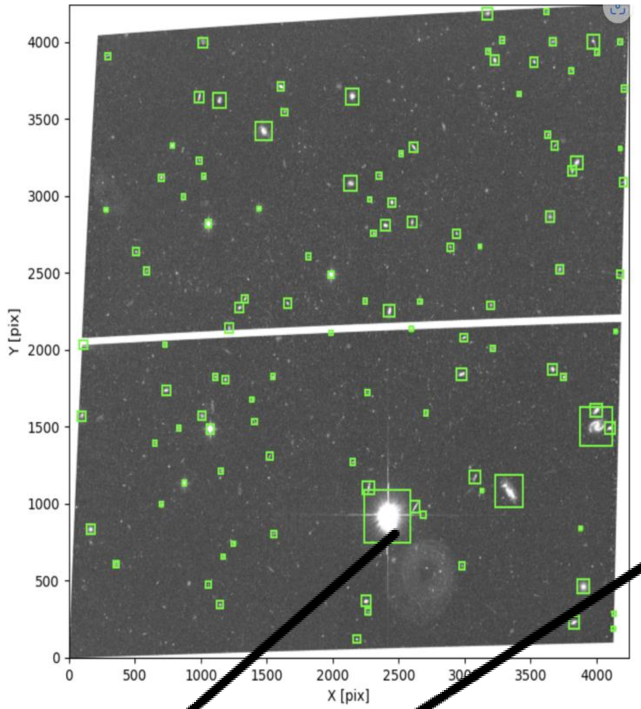
Abstract

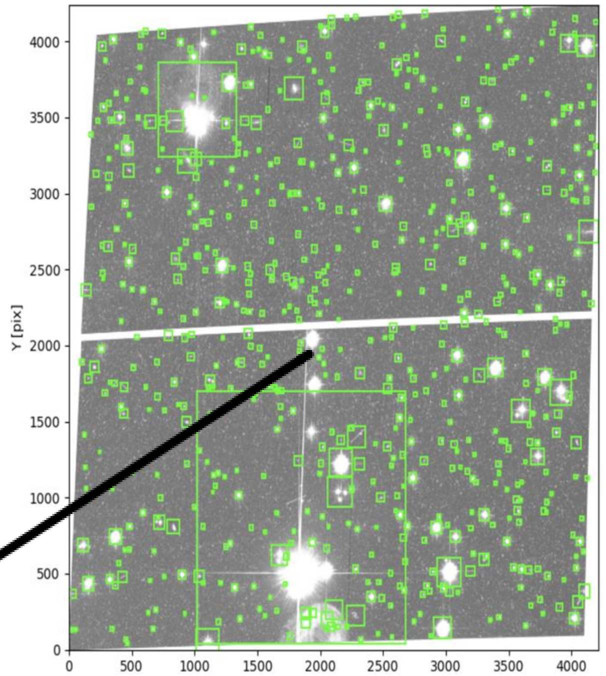
We have developed a method that maps large astronomical images onto a two-dimensional map and clusters them. A combination of various state-of-the-art machine-learning algorithms is used to develop a fully unsupervised image-quality assessment and clustering system. Our pipeline consists of a data pre-processing step where individual image objects are identified in a large astronomical image and converted to smaller pixel images. This data is then fed to a deep convolutional auto-encoder jointly trained with a self-organizing map (SOM). This part can be used as a recommendation system. The resulting output is eventually mapped onto a two-dimensional grid using a second, deep, SOM. We use data taken from ground-based telescopes and, as a case study, compare the system's ability and performance with the results obtained by supervised methods presented by Teimoorinia et al. The availability of target labels in this data allowed for a comprehensive performance comparison between our unsupervised and supervised methods. In addition to image-quality assessments performed in this project, our method can have various other applications. For example, it can help experts label images in a considerably shorter time with minimum human intervention. It can also be used as a content-based recommendation system capable of filtering images based on the desired content.

Unified Astronomy Thesaurus concepts: Astronomy data analysis (1858); Astronomy data modeling (1859)

We addressed the problem during the 2020 pandemic, but we couldn't implement it. We use deep learning and self-organizing maps to show the method. The idea is more or less the same, but we use more advanced models and no Self-Organizing Map in the new work.



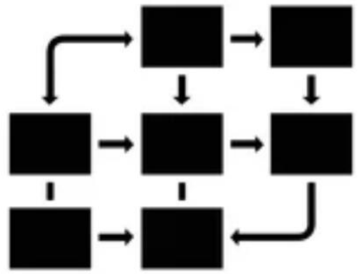


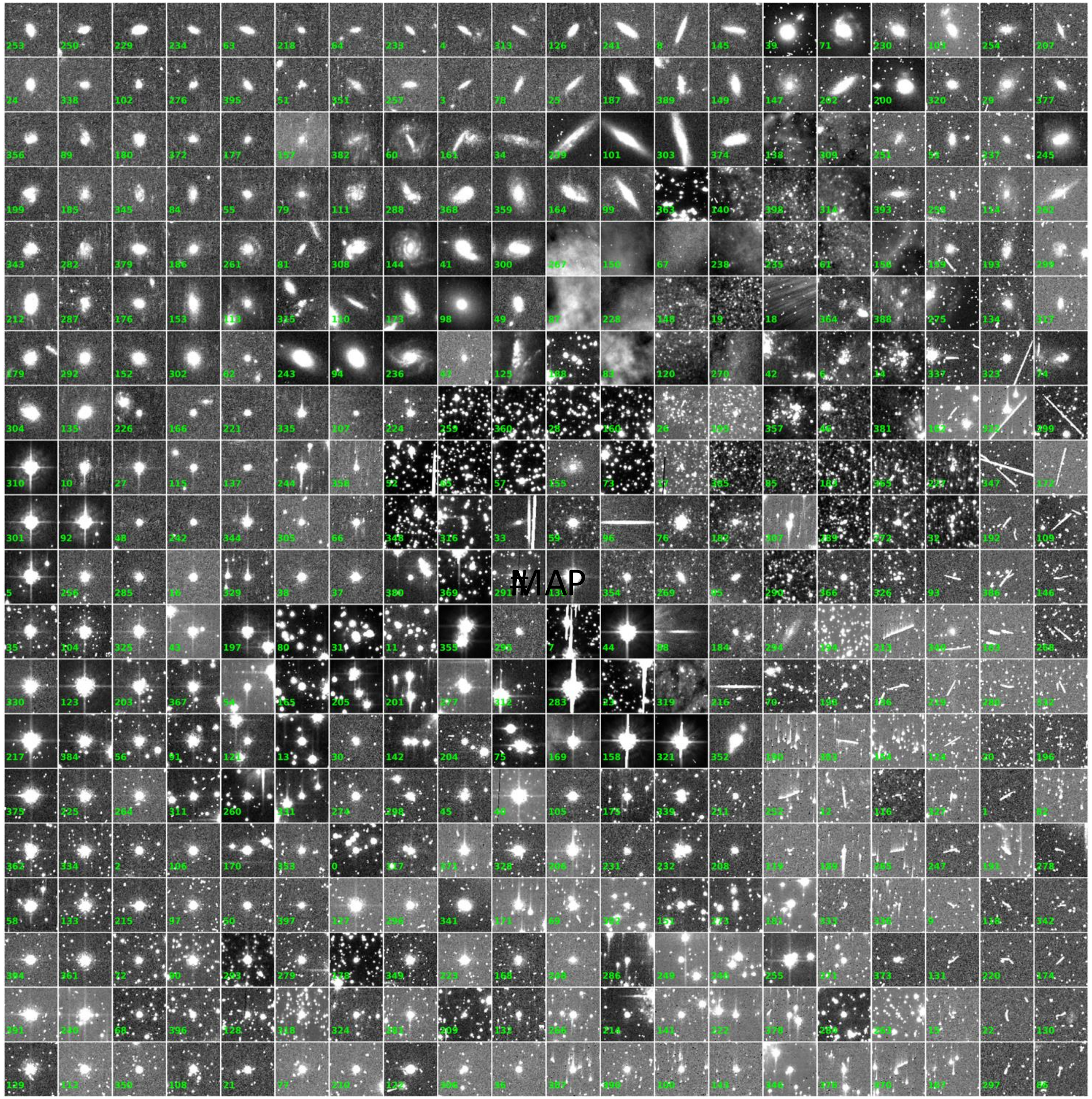


The main idea is simple: focus on detectable objects within images. Cut them out, gather them in one place, and use a pipeline of models to analyze and organize these cutouts

4,000,000 random cutouts from 20000 random images

A pipeline of models

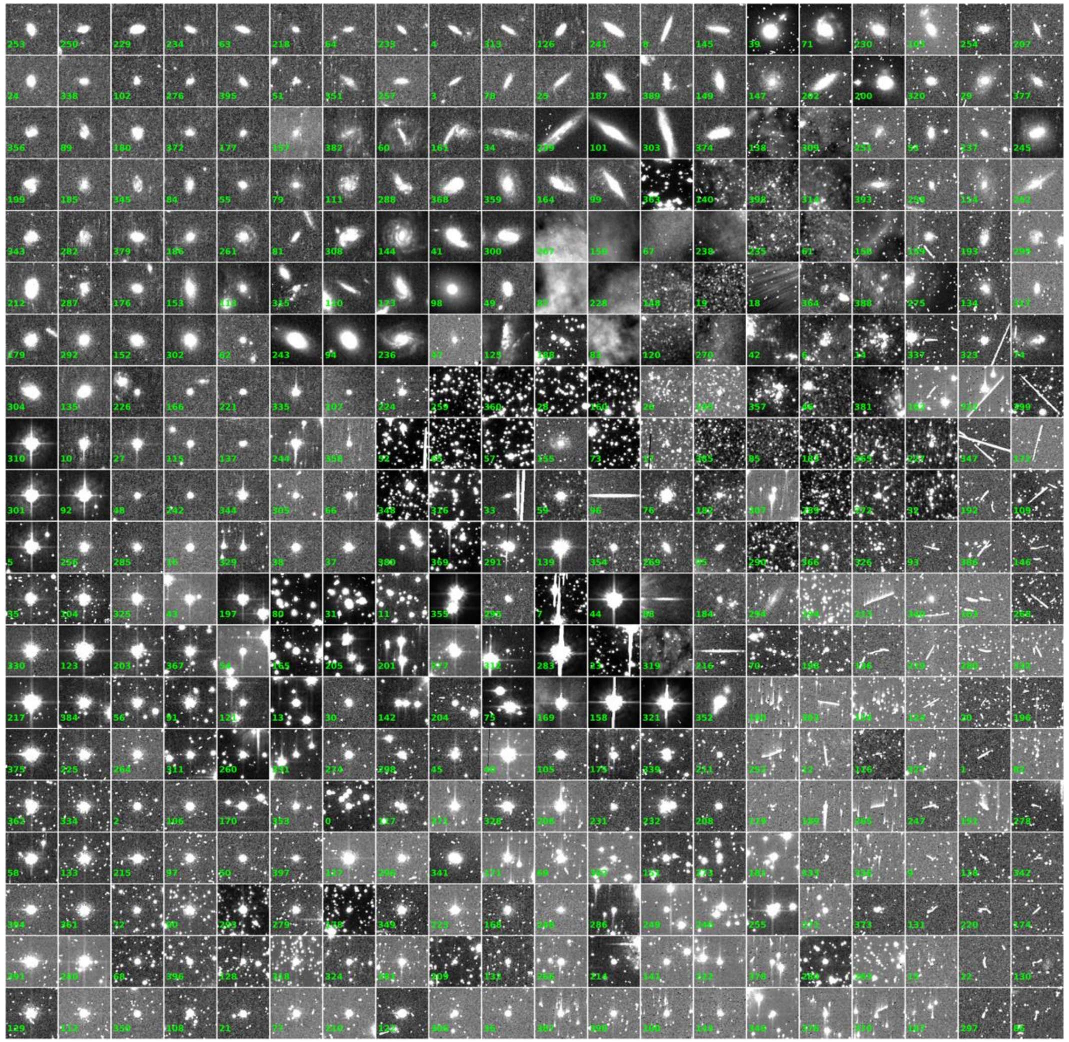




MAP #1:
the model #1, for finding the **objects** of interest

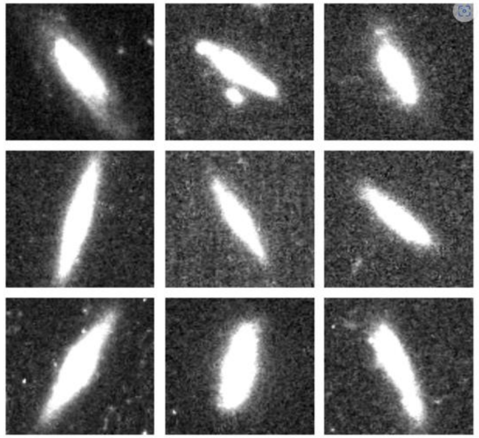
This slide shows the result of that pipeline. It arranges all the cutouts into a map with 400 nodes or groups. Each node represents a typical object type found in the survey. We call this Map 1.
Next, we need to verify the reliability of this map.

The model that can cluster detected objects in images in 400 groups

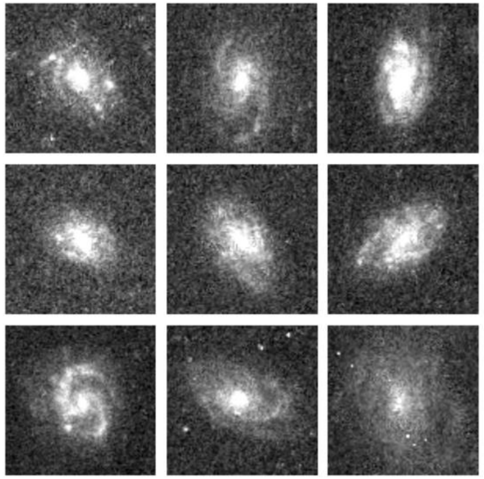


Once the model performs well, we can use it to create the content-based component required for CADC. Now, let's see how that content information is actually built.

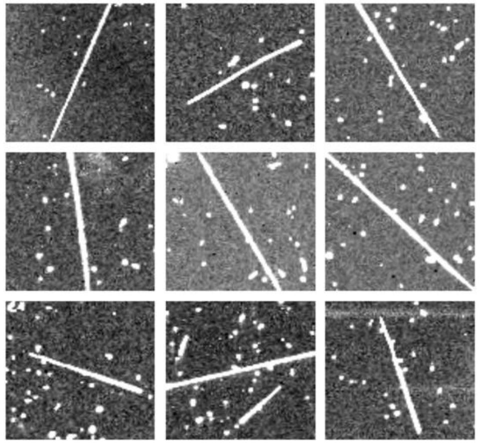
9 samples of group 101



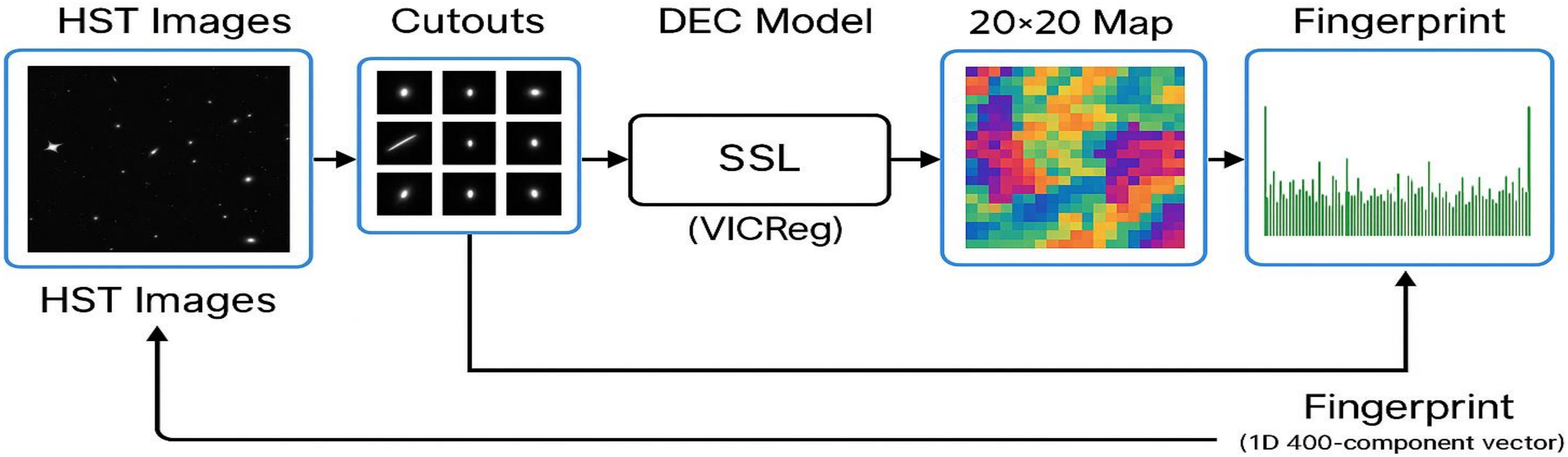
9 samples of group 144



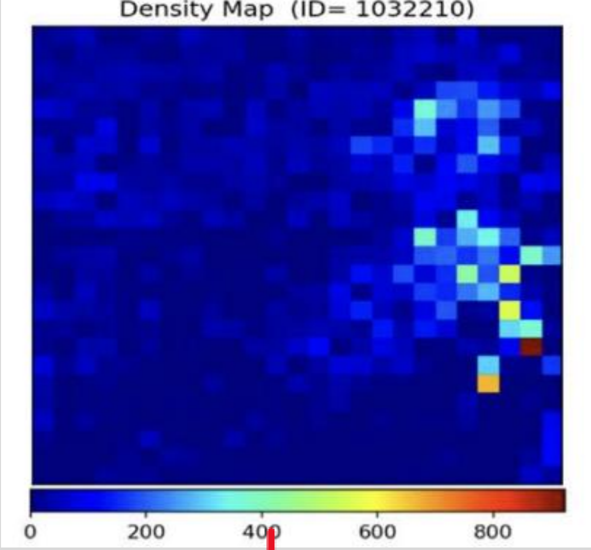
9 samples of group 399

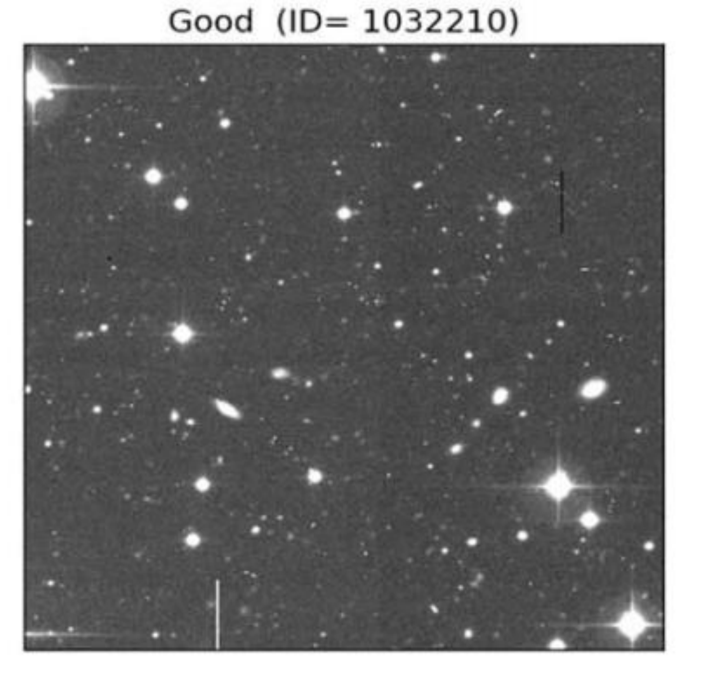


This slide shows how we create that content information.The cutout objects of an image are fed to the model, which produces a histogram or vector of 400 components.Each component represents the number of similar objects in a single node on Map 1.We call this vector the fingerprint, or FP, of an image.

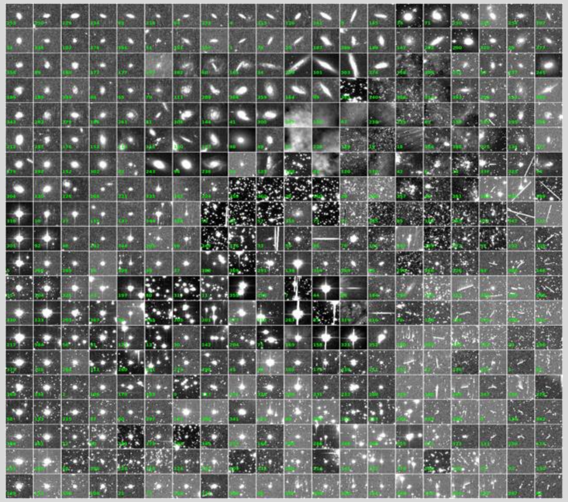


The two-dimensional distribution of objects (existing in the input image) on the model

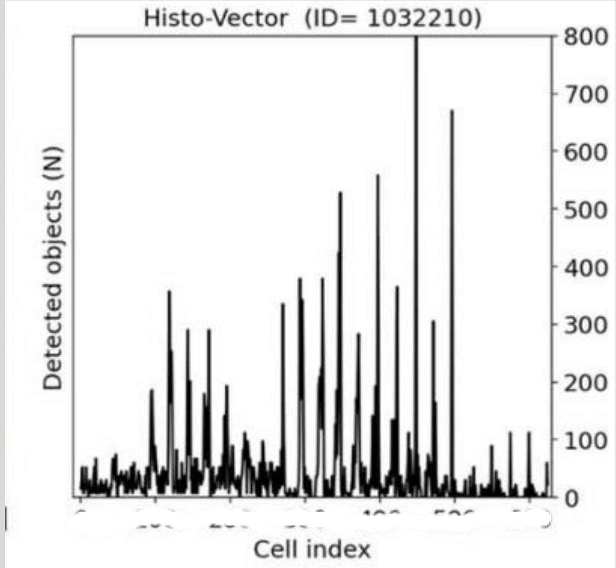




The input Image

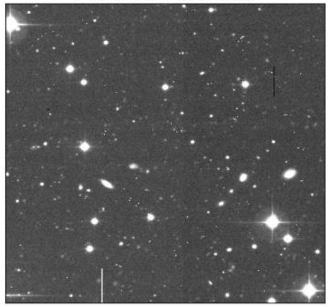


The model

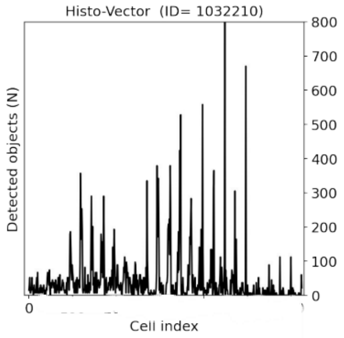


One dimensional distribution version (the fingerprint)

Good (ID= 1032210)

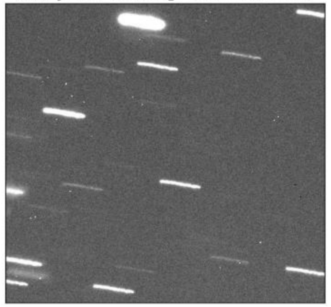


1000 × 1000
(Pixels/values)

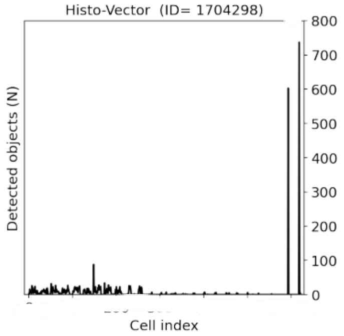


A fingerprint with 400 values

Really Bad Tracking (ID= 1704298)

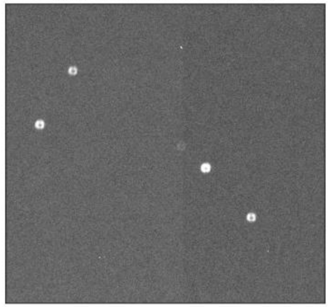


2000 × 2112
(Pixels/values)

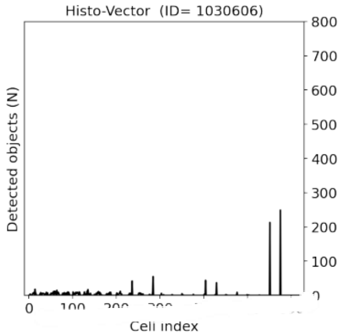


A fingerprint with 400 values

Bad Focus (ID= 1030606)

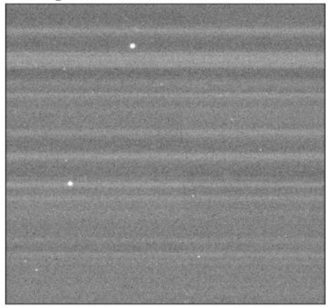


2644 × 2112
(Pixels/values)

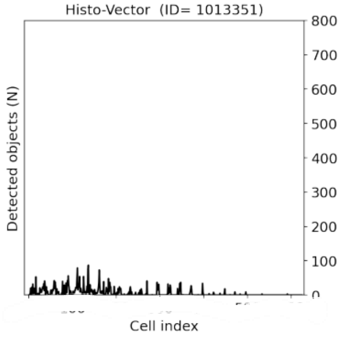


A fingerprint with 400 values

Background Problem (ID= 1013351)



4644 × 4112
(Pixels/values)

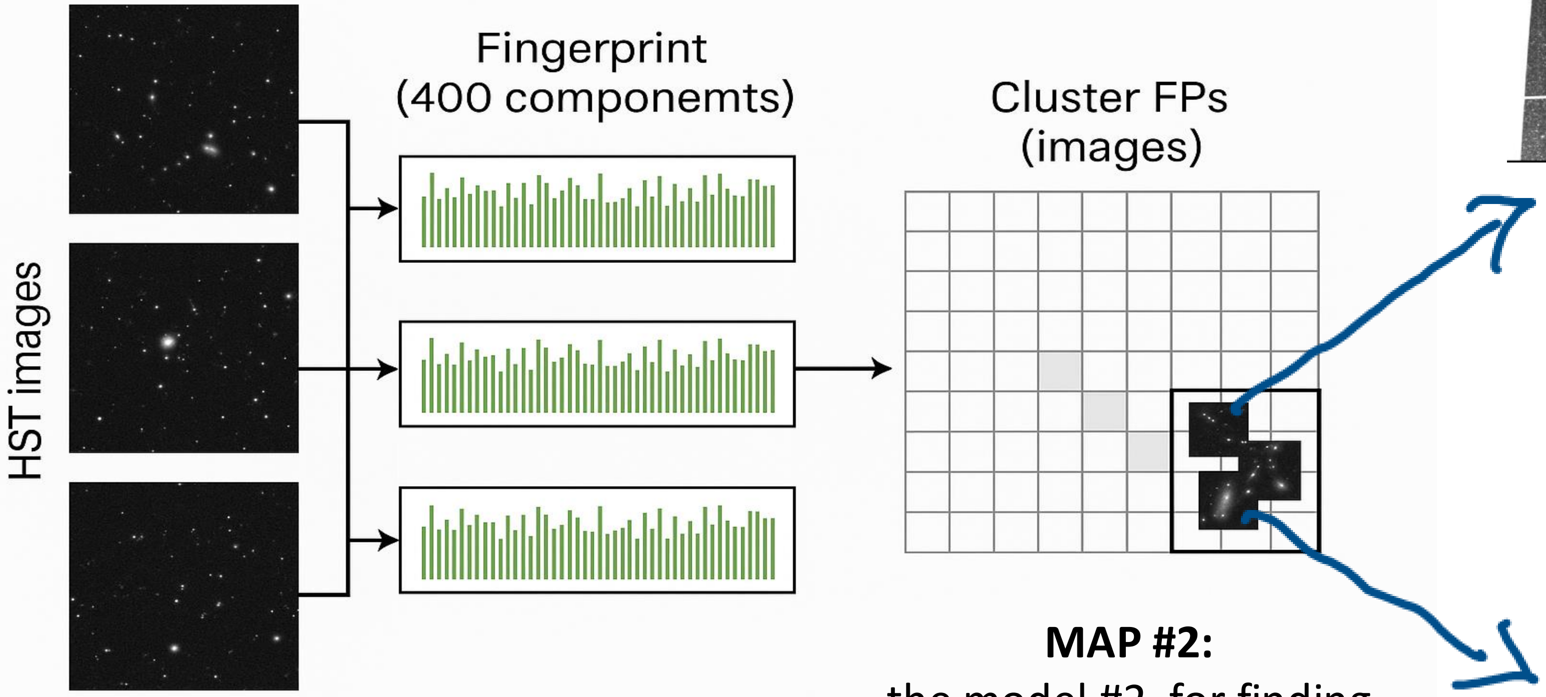


A fingerprint with 400 values

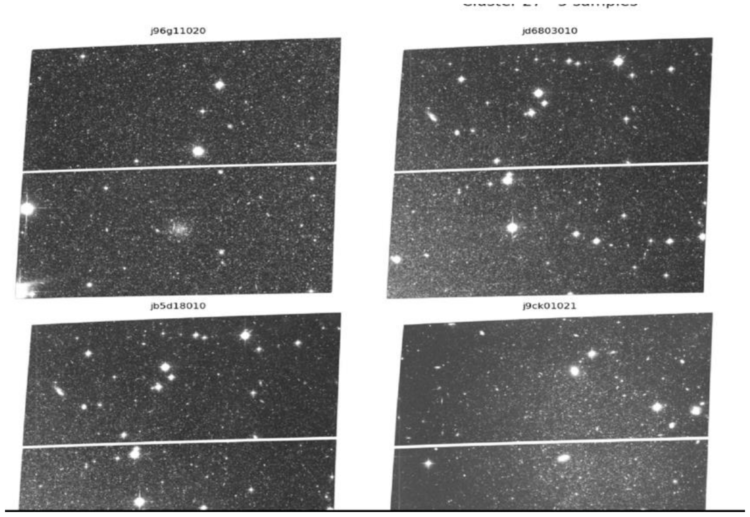
Fingerprints are tiny vectors, representing images that are ~ 10000 times smaller than the original.

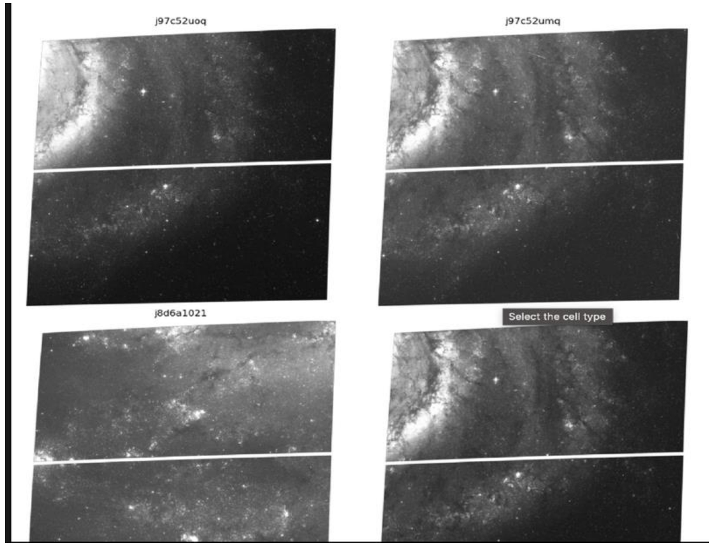
This slide shows more examples, comparing fingerprints with their corresponding large images. No matter what the original image size is, each image is converted into a 400-dimensional vector. These vectors are perfect for clustering in machine learning.We use them to build a second map, Map 2, where each node represents similar images.

Here you can see the clustering process.Similar fingerprints fall into the same node, so each node on Map 2 contains similar FPs, or equivalently, similar images. Now imagine using this in a large-scale search.

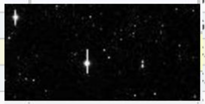


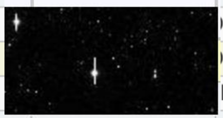
MAP #2:
the model #2, for finding the **images** of interest





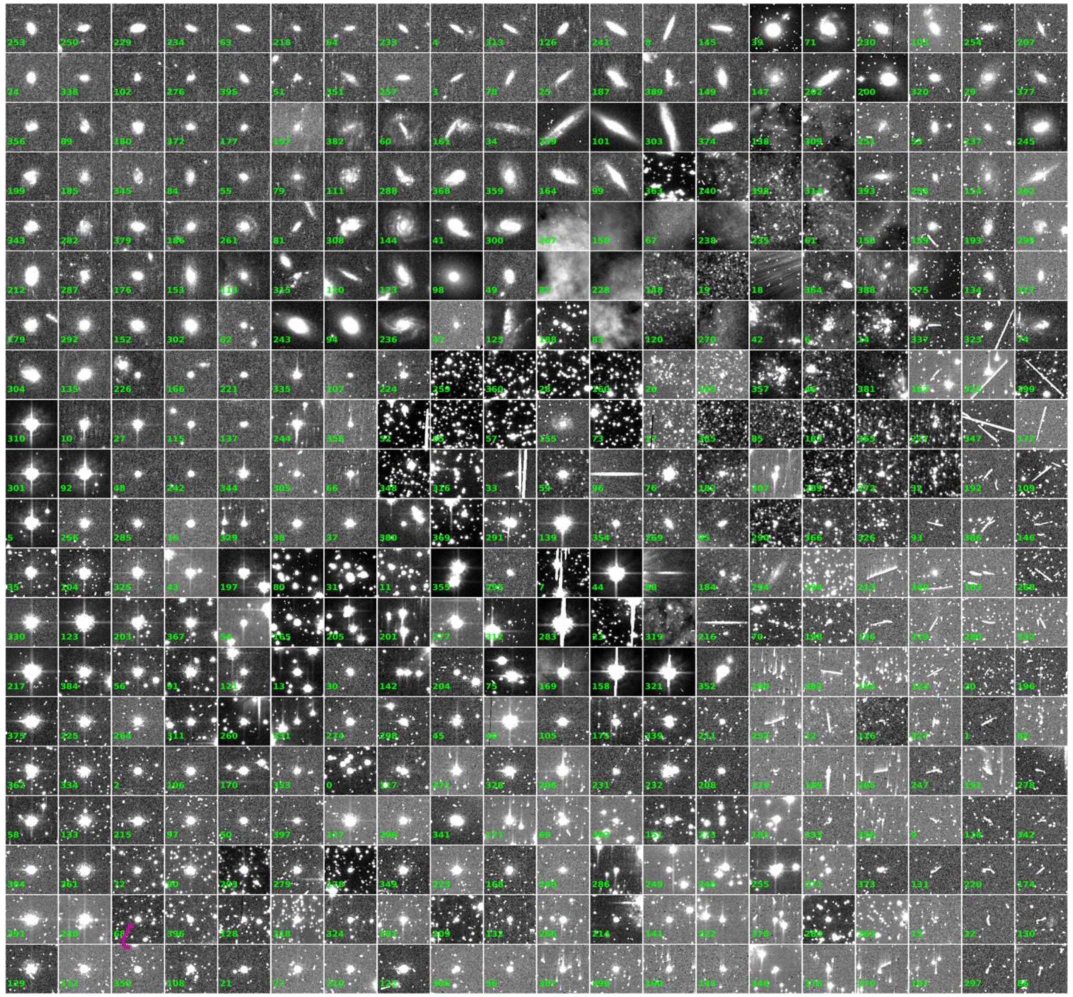
Download | Showing 30000 rows (30000 before filtering). There are more rows available by downloading the complete query results. | Change Columns | View Results on Sky

Mark	Preview	Collection	Proposal ...	Obs. ID	Product ID	RA (J2000.0)	Dec. (J2000.0)	Target Na...	Start Date	Int. Time	Instrument	Filter	Cal. Lev.	Obs. Ty
Filter:						H:M:S	D:M:S		Calendar	Seconds				
☐		HST	GO/DD	ifmk04ugq	ifmk04ugq-CALIBRATED	03:52:20.66	+19:59:12.5	URANUS-MA	2025-10-23 18:21:1	6.000	WFC3/UVIS	F547M	2	IMAGIN
☐		HST	GO/DD	ifmk02tjq	ifmk02tjq-CALIBRATED	03:52:22.69	+19:59:18.7	URANUS-MA	2025-10-23 12:39:2	35.000	WFC3/UVIS	F845M	2	IMAGIN
☐		HST	GO/DD	ifmk02thq	ifmk02thq-CALIBRATED	03:52:22.94	+19:59:46.4	URANUS-MA	2025-10-23 12:29:3	200.000	WFC3/UVIS	FQ727N	2	IMAGIN
☐		HST	CAL/STIS	ofcx9u020	ofcx9u020-CALIBRATED	03:52:25.98	+19:59:22.2	BIAS	2025-10-23 07:14:2	0.000	STIS/CCD	MIRVIS	2	IMAGIN
☐		HST	CAL/STIS	ofcx9u010	ofcx9u010-CALIBRATED	03:52:25.98	+19:59:22.2	BIAS	2025-10-23 06:58:1	0.000	STIS/CCD	MIRVIS	2	IMAGIN
☐		HST	CAL/STIS	ofdajos1q	ofdajos1q-CALIBRATED	03:52:25.98	+19:59:22.2	DARK	2025-10-23 06:50:3	60.000	STIS/CCD	MIRVIS	2	IMAGIN
☐		HST	CAL/STIS	ofdajos0q	ofdajos0q-CALIBRATED	03:52:25.98	+19:59:22.2	DARK	2025-10-23 06:48:5	60.000	STIS/CCD	MIRVIS	2	IMAGIN
☐		HST	CAL/STIS	ofdajorzq	ofdajorzq-CALIBRATED	03:33:03.79	-02:48:02.0	DARK	2025-10-23 06:29:4	1100.000	STIS/CCD	MIRVIS	2	IMAGIN
☐		HST	CAL/STIS	ofdajnrrq	ofdajnrrq-CALIBRATED	03:33:03.80	-02:48:02.2	DARK	2025-10-23 05:46:5	60.000	STIS/CCD	MIRVIS	2	IMAGIN
☐		HST	CAL/STIS	ofdajnrqq	ofdajnrqq-CALIBRATED	03:33:03.80	-02:48:02.2	DARK	2025-10-23 05:45:1	60.000	STIS/CCD	MIRVIS	2	IMAGIN
☐		HST	CAL/STIS	ofdajnrpq	ofdajnrpq-CALIBRATED	03:33:03.80	-02:48:02.2	DARK	2025-10-23 05:26:1	1100.000	STIS/CCD	MIRVIS	2	IMAGIN
N/A		HST	GO	jf9s12ceq	jf9s12ceq-CALIBRATED	00:54:24.70	-37:45:49.0	N300-D	2025-08-11 09:03:0	457.000	ACS/WFC	F606W	2	IMAGIN
N/A		[illegible]	[illegible]	jf9s12cbq	jf9s12cbq-CALIBRATED	00:54:24.80	-37:45:50.1	N300-D	2025-08-11 08:51:3	525.000	ACS/WFC	F606W	2	IMAGIN
☐	Preview		[illegible]	jfg702q4q	jfg702q4q-CALIBRATED	05:17:05.83	-44:10:44.3	HE0515-441	2025-04-24 06:17:1	43.000	ACS/WFC	F814W	2	IMAGIN
☑	Preview	[illegible]	[illegible]	jdua10taq	jdua10taq-CALIBRATED	00:03:31.72	+70:18:58.7	TUNGSTEN	2018-12-02 03:05:3	3.000	ACS/WFC	F775W;CLEAR2L	2	INTERN
☐	Preview	[illegible]	[illegible]	jdua10t8q	jdua10t8q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 02:57:2	3.000	ACS/WFC	F775W;CLEAR2L	2	INTERN
☐		[illegible]	[illegible]	jdua10010	jdua10011-CALIBRATED	09:59:01.38	+01:55:36.8	TUNGSTEN	2018-12-02 02:49:1	11.100	ACS/WFC	F775W;CLEAR2L	2	INTERN
☐	Preview	HST	CAL/ACS	jdua04s6q	jdua04s6q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 01:07:5	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
☐	Preview	HST	CAL/ACS	jdua04s4q	jdua04s4q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 00:59:4	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
☐	Preview	HST	CAL/ACS	jdua04s2q	jdua04s2q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 00:51:3	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
☐	Preview	HST	CAL/ACS	jdua09roq	jdua09roq-CALIBRATED	09:59:03.94	+01:55:05.7	TUNGSTEN	2018-12-01 23:41:4	140.000	ACS/WFC	CLEAR1L;F660N	2	INTERN
☐	Preview	HST	CAL/ACS	jdua09rmq	jdua09rmq-CALIBRATED	09:59:03.94	+01:55:05.7	TUNGSTEN	2018-12-01 23:31:1	140.000	ACS/WFC	CLEAR1L;F660N	2	INTERN

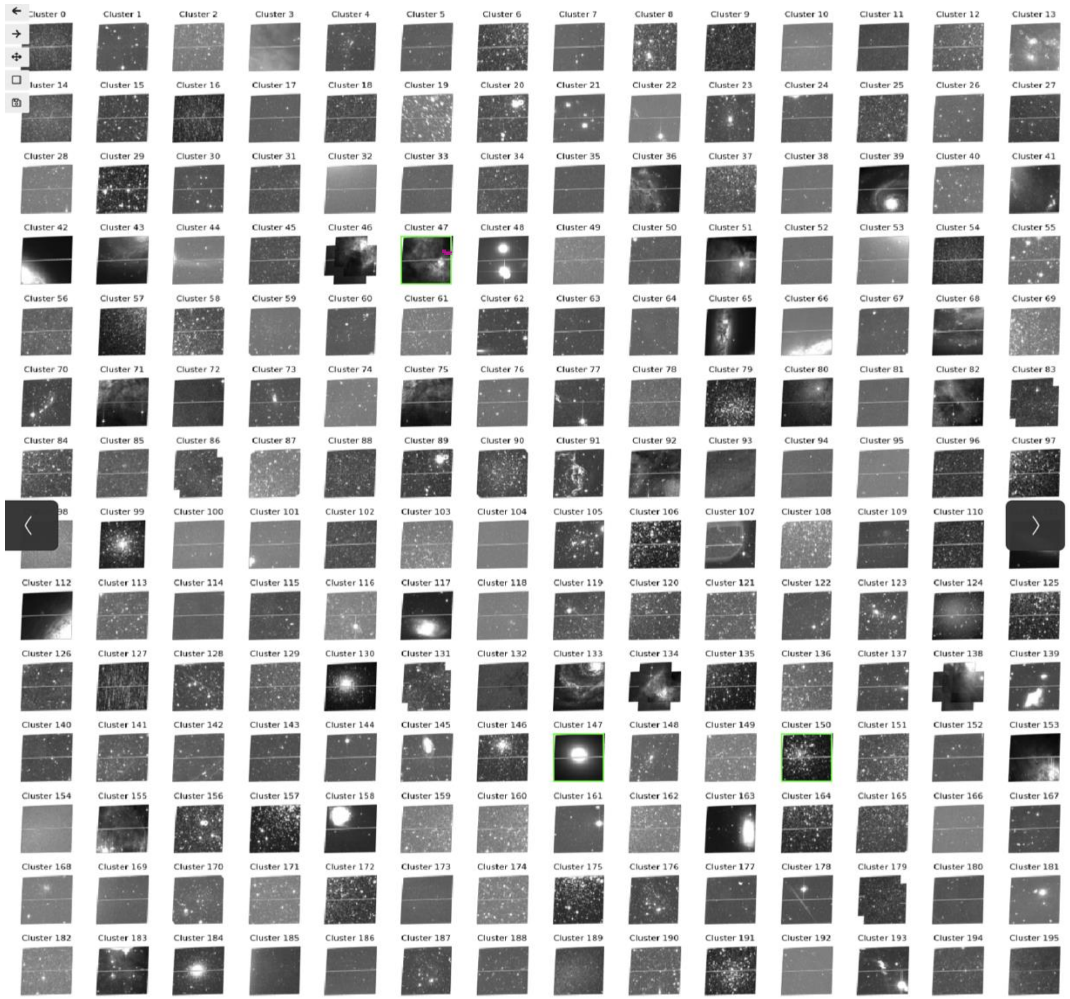
Mark	Preview	Collection	Proposal ...	Obs. ID	Product ID	RA (J2000.0)	Dec. (J2000.0)	Target Na...	Start Date	Int. Time	Instrument	Filter	Cal. Lev.	Obs. Ty
N/A		HST	GO	jf9s12ceq	jf9s12ceq-CALIBRATED	00:54:24.70	-37:45:49.0	N300-D	2025-08-11 09:03:0	457.000	ACS/WFC	F606W	2	IMAGIN
N/A			[illegible]	jf9s12cbq	jf9s12cbq-CALIBRATED	00:54:24.80	-37:45:50.1	N300-D	2025-08-11 08:51:3	525.000	ACS/WFC	F606W	2	IMAGIN
☑	Preview	[illegible]	[illegible]	jfg702q4q	jfg702q4q-CALIBRATED	05:17:05.83	-44:10:44.3	HE0515-441	2025-04-24 06:17:1	43.000	ACS/WFC	F814W	2	IMAGIN
☐	Preview	[illegible]	[illegible]	jdua10taq	jdua10taq-CALIBRATED	00:03:31.72	+70:18:58.7	TUNGSTEN	2018-12-02 03:05:3	3.000	ACS/WFC	F775W;CLEAR2L	2	INTERN
☐	Preview	HST	CAL/ACS	jdua10t8q	jdua10t8q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 02:57:2	3.000	ACS/WFC	F775W;CLEAR2L	2	INTERN
☐		HST	CAL/ACS	jdua10010	jdua10011-CALIBRATED	09:59:01.38	+01:55:36.8	TUNGSTEN	2018-12-02 02:49:1	11.100	ACS/WFC	F775W;CLEAR2L	2	INTERN
☐	Preview	HST	CAL/ACS	jdua04s6q	jdua04s6q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 01:07:5	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
☐	Preview	HST	CAL/ACS	jdua04s4q	jdua04s4q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 00:59:4	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN
☐	Preview	HST	CAL/ACS	jdua04s2q	jdua04s2q-CALIBRATED	09:59:03.97	+01:55:05.3	TUNGSTEN	2018-12-02 00:51:3	6.000	ACS/WFC	F550M;CLEAR2L	2	INTERN

Imagine a massive search where you checkmark one image.All other similar images in the same node on Map 2 can be automatically selected. This makes large searches much easier to handle.

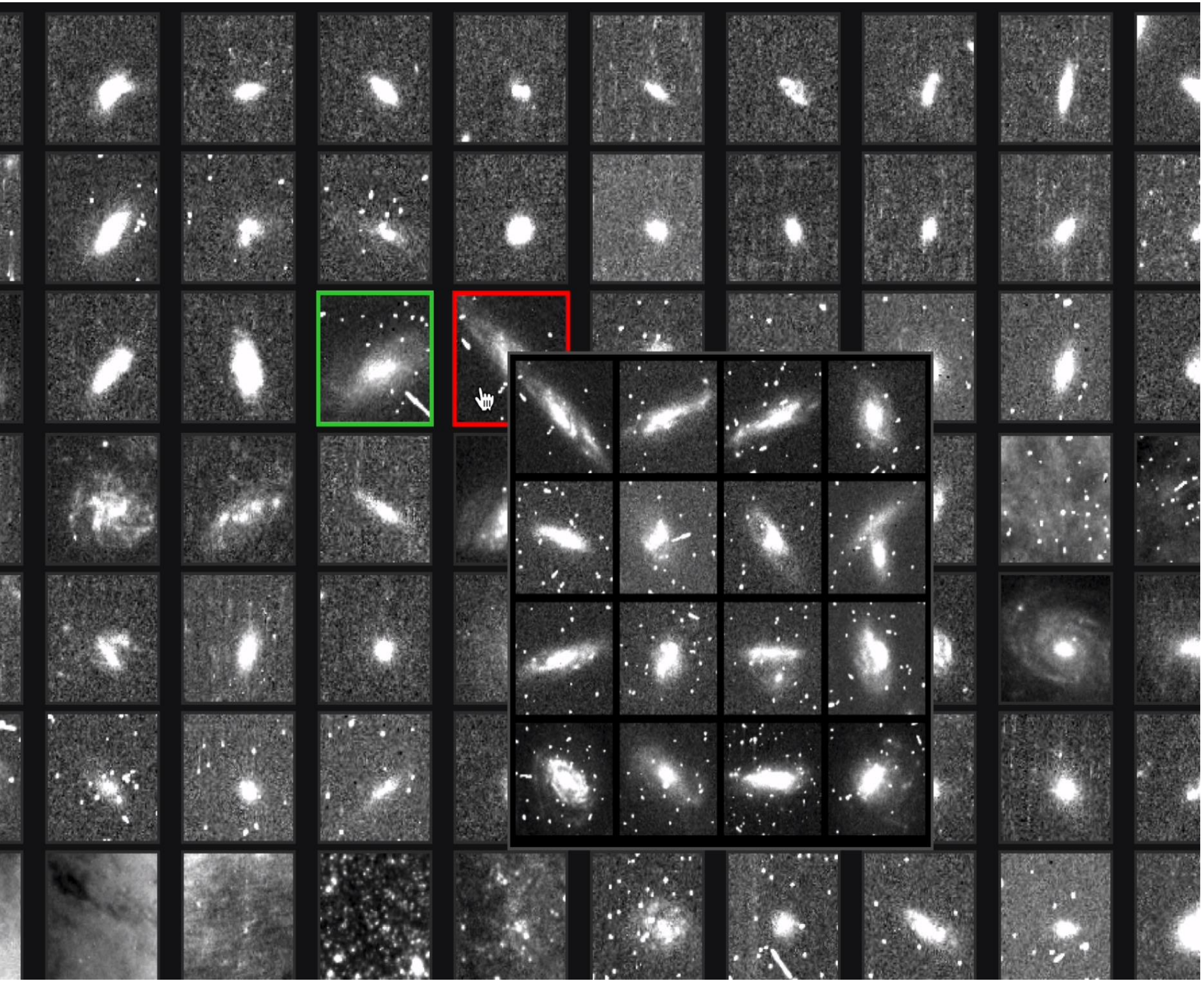
So now we have two maps to help download images:From Map 1, you can select objects that interest you, and from Map 2, you can select similar fields or images. To make it even more user-friendly, we built interactive versions of these maps.



MAP #1 for choosing the **objects** of interest



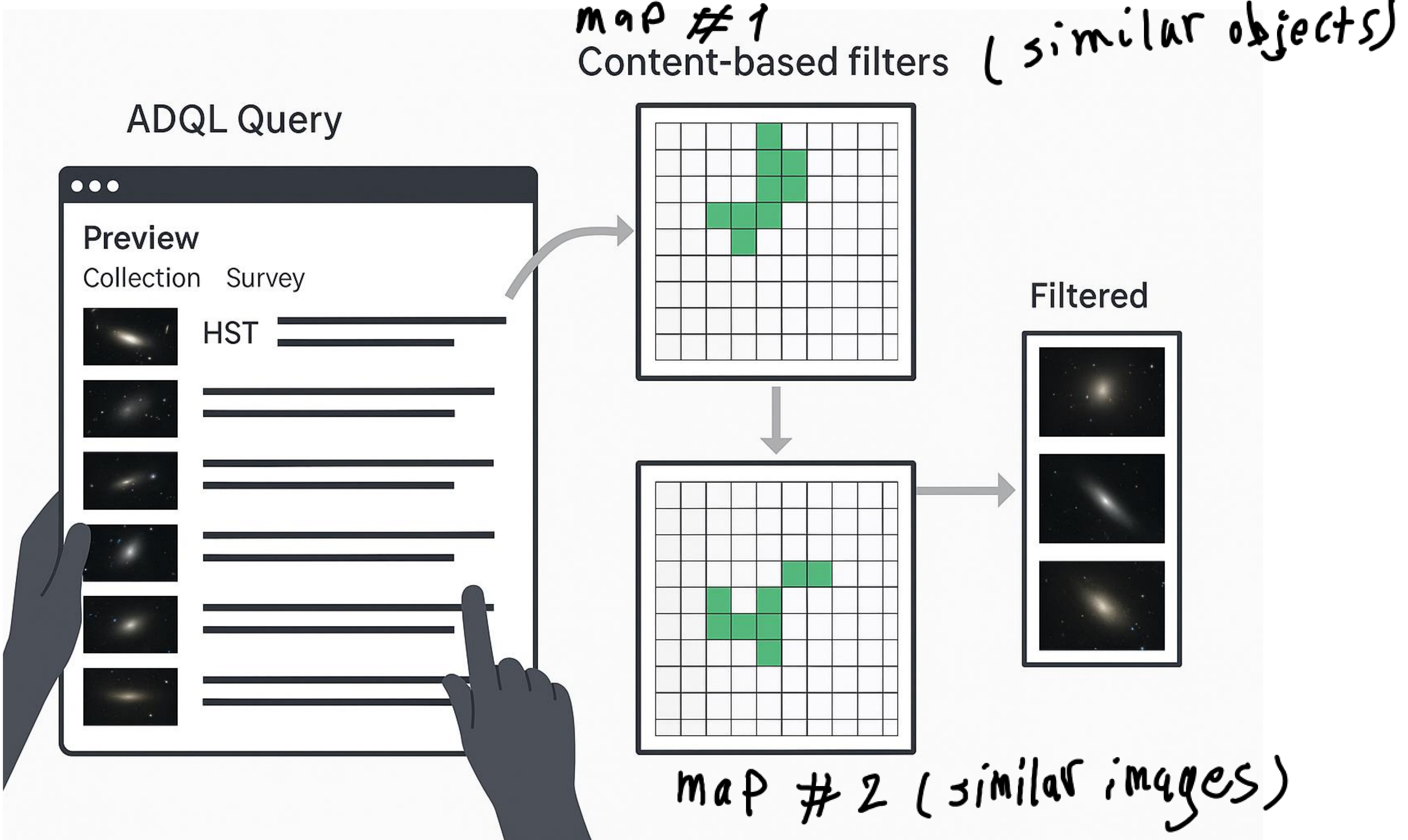
MAP #2 for choosing the **images** of interest



We created interactive maps that make this process simple.Users can hover over a node to see a sample of its content and statistics.You can include objects by marking nodes green, or exclude them by marking them red. Next, let's look at how all of this fits together in the full system.

Interactive MAP 1

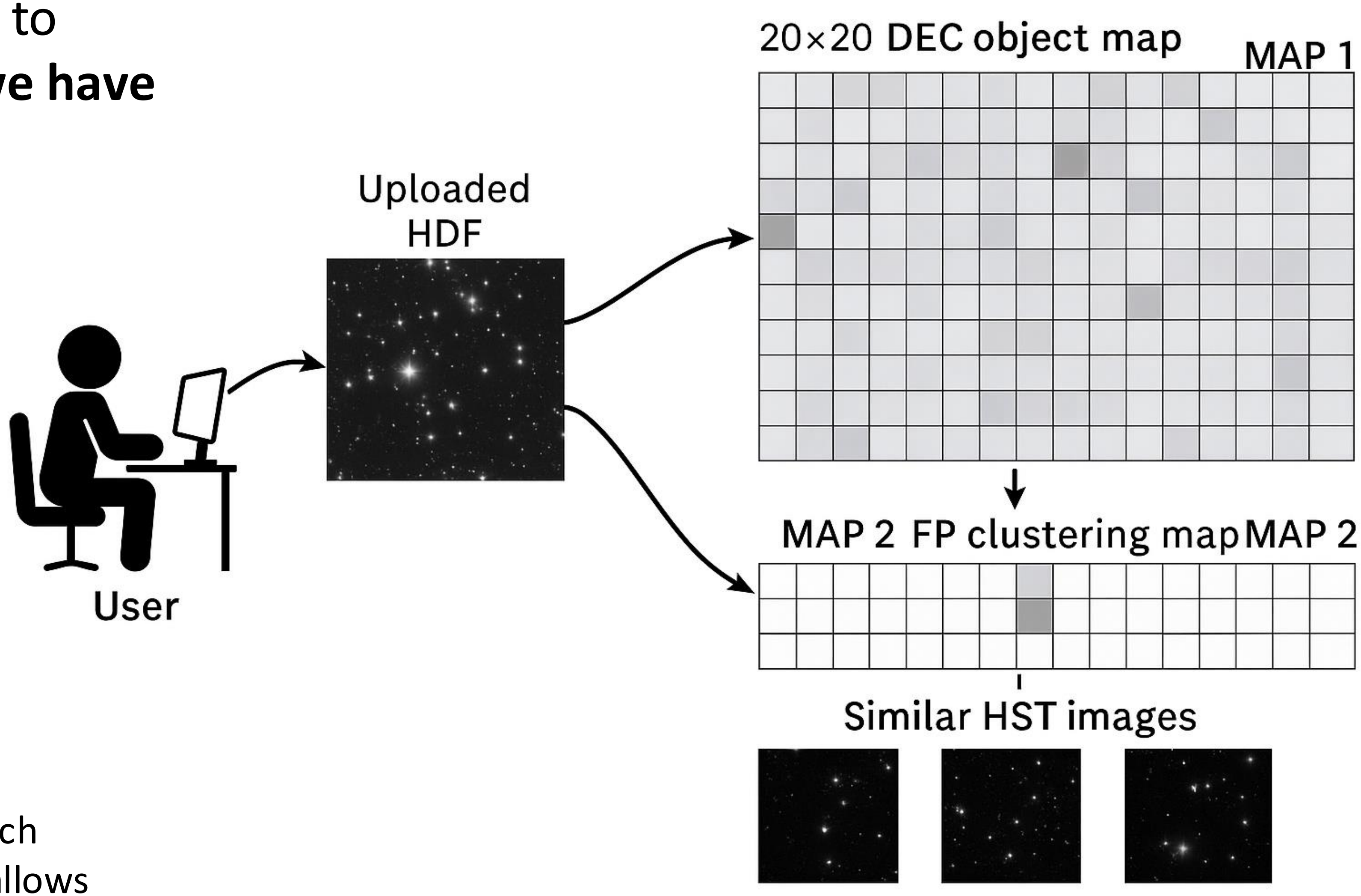
Shaylin Thahdani
Shavon Thahdani
(U of Ottawa)



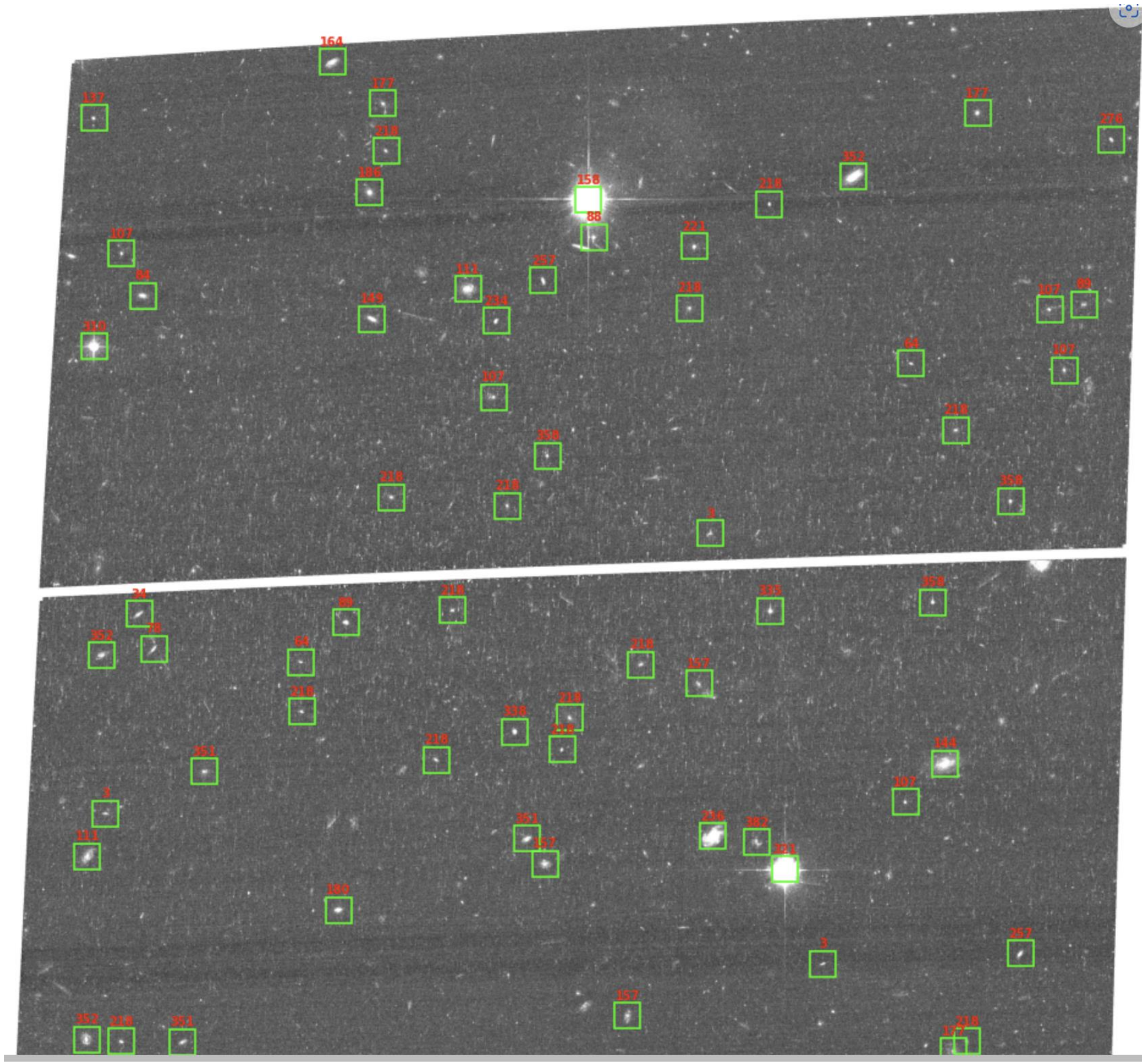
map #1
Content-based filters
(similar objects)
ADQL Query
Preview
Collection
Survey
HST
Filtered
map #2 (similar images)

Another application of it is to

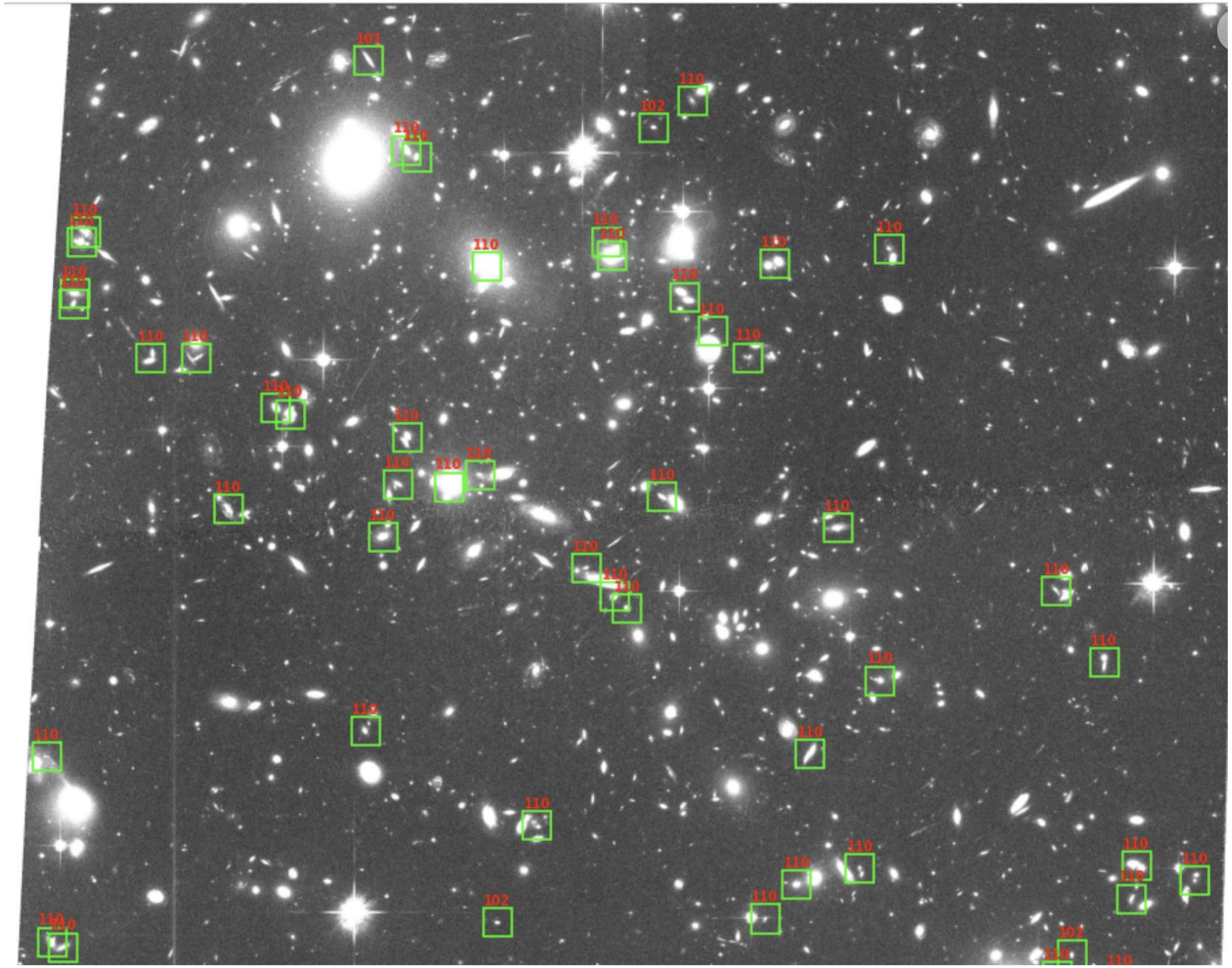
find similar images to those we have



Also, a user can upload an image.The pipeline can create its fingerprint, which then falls into a node on Map 2. This allows the system to suggest similar images from the archive.It’s a powerful potential application for future use.



The model can also annotate all detected objects in an image, assigning each one to its matching node on Map 1. This gives us automatic, large-scale labeling and classification.And we can even customize it for specific science goals.



Also, it can list images in a collection in CADC with a desired shape and character, such as mergers.

Pat (**Patrick Dowler** from CADC) can show you some examples of searches and the interactive maps, if you are interested.

Thanks!