Machine Learning in Astronomy

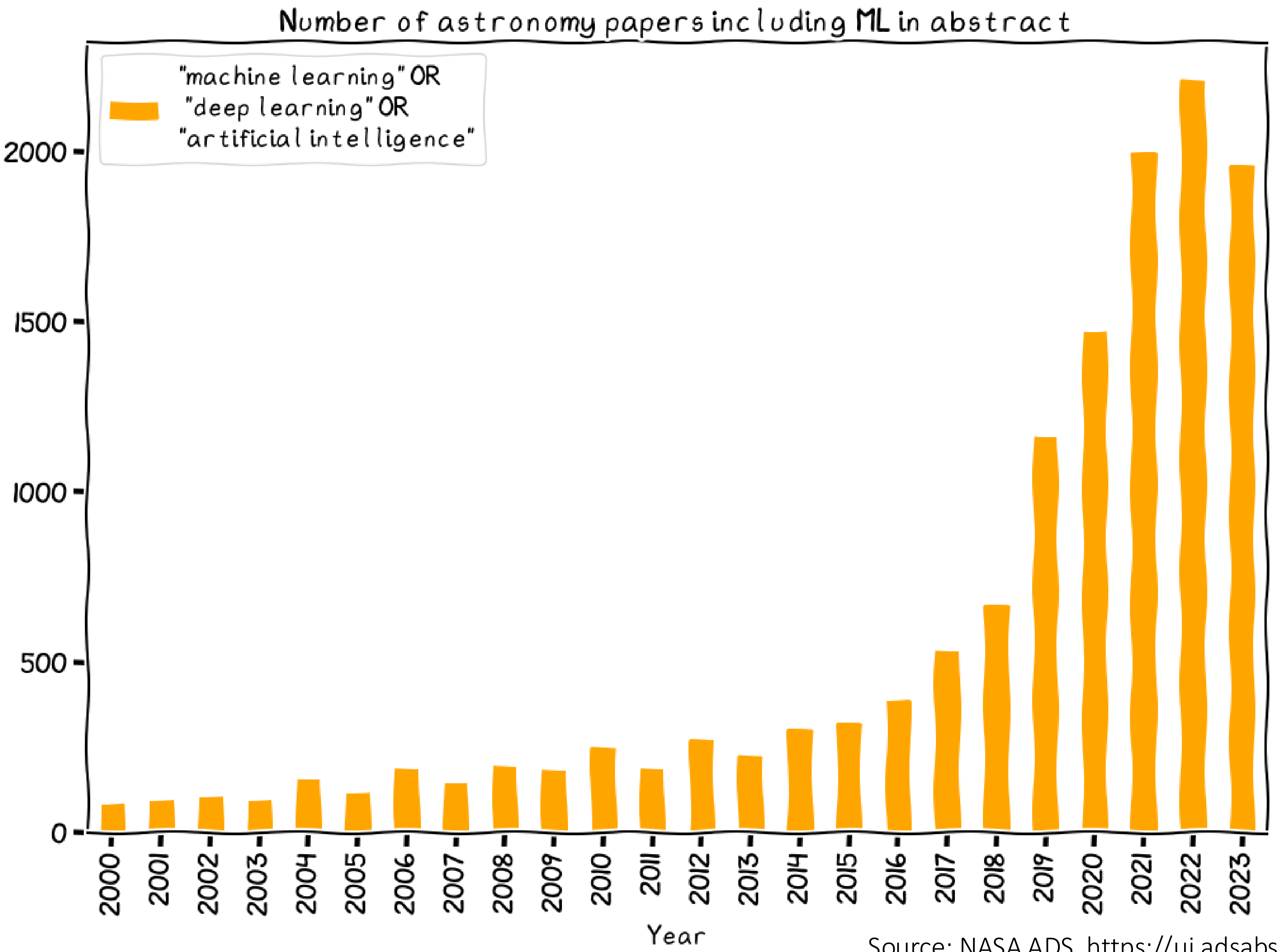
Francisco M. Montenegro Montes
María Zambrano fellow, UCM, IPARCOS

2nd IPARCOS Workshop on Machine Learning
and
Applications to Physics, 18-20 Dec 2023

GOBIERNO DE ESPAÑA
MINISTERIO DE UNIVERSIDADES
Plan de Recuperación, Transformación y Resiliencia
Financiado por la Unión Europea
NextGenerationEU

UNIVERSITAS COMPLVTENSIS MATRITENSIS

IPARCOS



Source: NASA ADS. https://ui.adsabs.harvard.edu/

Explosive increase in the use of ML techniques in the last 5-10 years

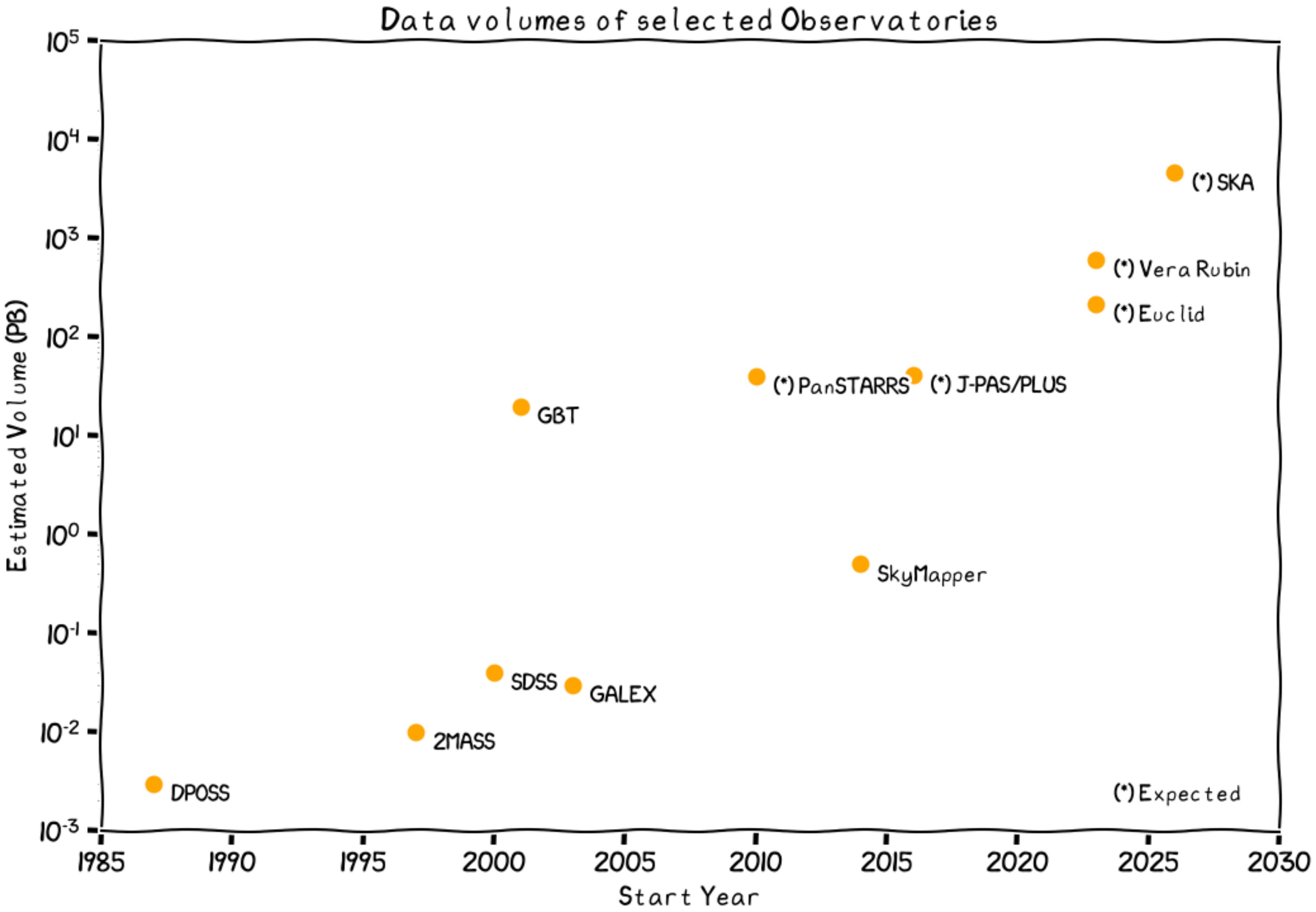
- Bigger and more complex datasets available (open source). ML is almost a necessity. And the future will become more demanding (Rubin, SKA).
- Techniques become more popular and better known.
- Success stories involving citizen science projects.
- Availability of better computing infrastructures (GPUs, cloud services) and more funding for AI-based projects.
- No ethical issues like privacy concerns or biases that may affect other disciplines.

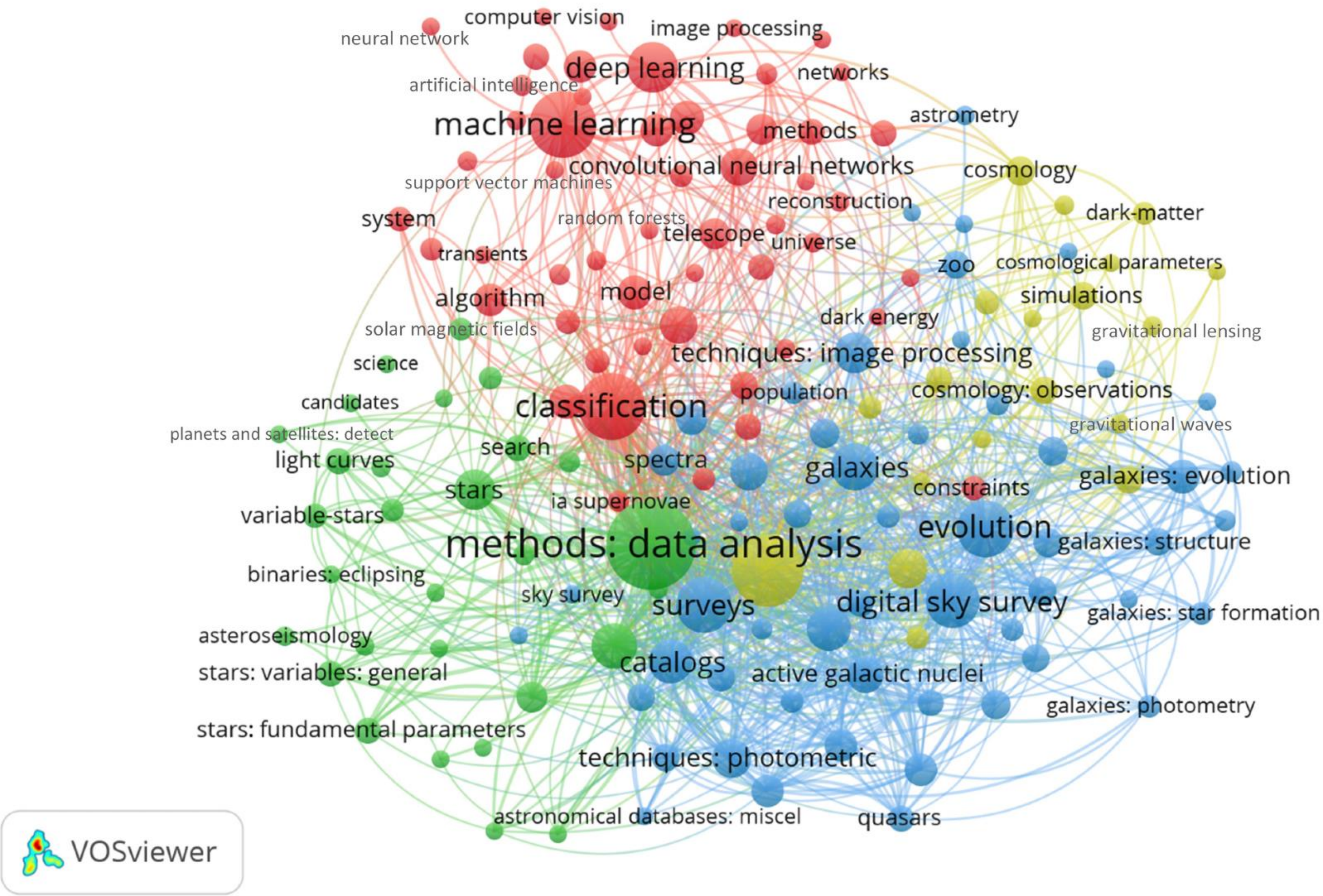
I don't know why physicists insist saying that photons don't occupy volume...

"Observatories' data storage costs will threaten their budgets. Data transport latency and bandwidth will threaten not just budgets, but available technology and human patience".

R. Seaman et al. **2007**. ADASS XVI Meeting

I don't know why physicists insist saying that photons don't occupy volume...





Includes AI and ML techniques applied to the fields of astronomy and astrophysics

Application of AI and ML in the field of stellar analysis.

Study of cosmology using AI and ML methods

Analysis of galaxies by means of AI and ML techniques

Rodríguez et al. 2022. *On the application of machine learning in astronomy and astrophysics: A text-mining-based scientometric análisis.* (~2700 papers published between 1989 and 2021)

Classification: Categories or labels are applied to objects or features. Based on a training set (labeled or unlabeled), the algorithm learns the characteristics that relate an instance to a category. When applied to a new instance, the algorithm assigns the most likely category label.

Regression: Assignment of a numerical value (or values) based on the characteristics that are learnt or otherwise predicted by the machine learning algorithm. As with classification, a training set may be used or the characteristics may be inferred from the dataset.

Clustering: Determine whether an object or a feature is part of (i.e., a member of) something. This might be a physical structure or association—as in the more familiar usage of the term in astronomy as applied to open, globular, or galactic clusters—or a region within an N-dimensional parameter space.

Forecasting: The purpose of the machine learning algorithm is to learn from previous events, and predict or forecast that a similar event is going to occur. There is an implicit time-dependence to the prediction.

Generation and reconstruction: Missing information is created, expected to be consistent with the underlying truth. The cause of the missing information might be due to the presence of noise, processing artifacts, or additional astronomical phenomena, all of which conspire to obscure the required signal.

Discovery: New celestial objects, features, or relationships are identified as a consequence of the application of a ML or AI method.

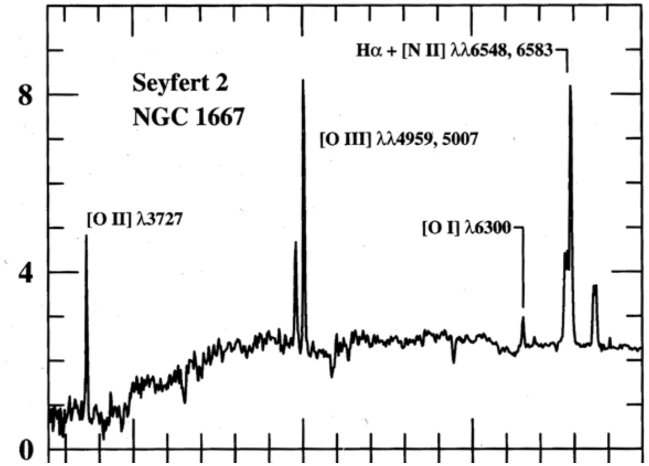
Insight: Moving beyond the discovery of celestial objects, new scientific knowledge is demonstrated as a consequence of applying machine learning or AI. This includes cases where insight is gained into the suitability of applying machine learning, choice of data set, hyperparameters, and comparisons with human-based classification.

Fluke & Jacobs 2019

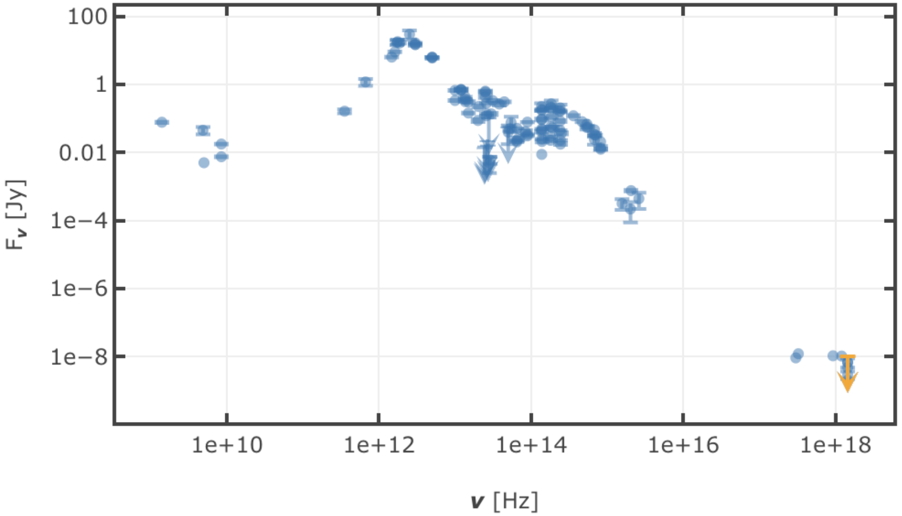
Types of astronomical data where ML is normally applied

Images

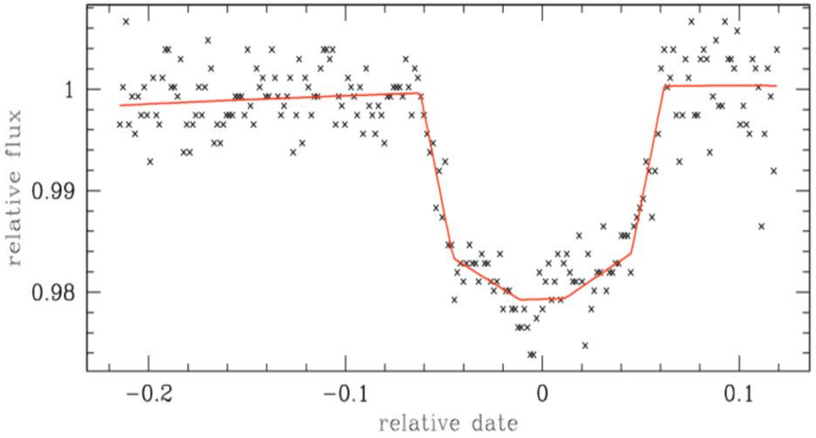
Spectra



Spectral energy distributions



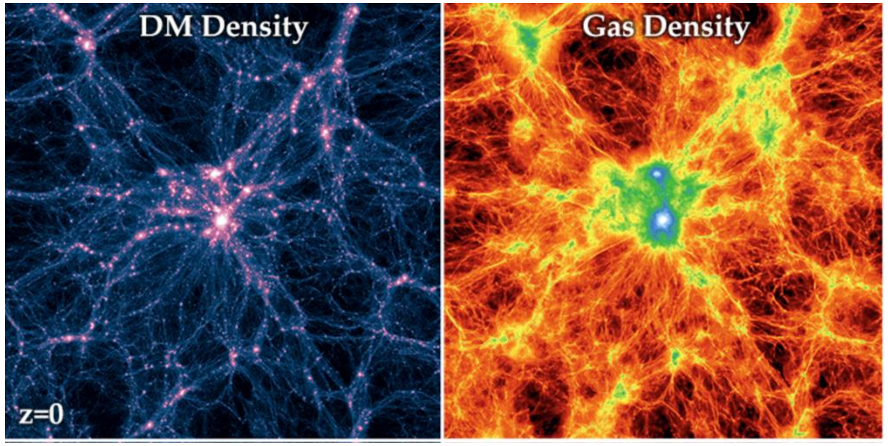
Lightcurves / Time series

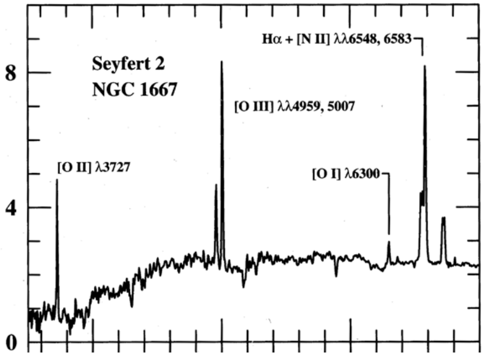


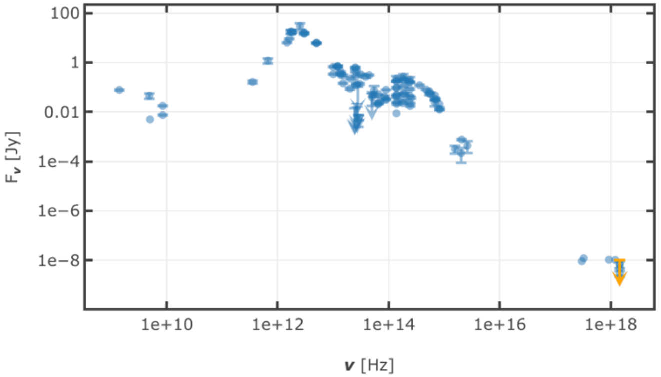
Astronomical catalogues

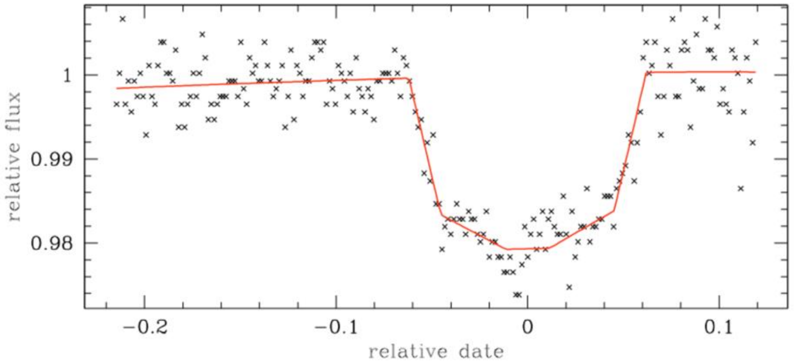
Freq. (MHz)	Name	Year of publ.	RA (h) or l (d)	Dec (°) or b (d)	HalfPower BeamWidth (arcmin)	S_{min} (mJy)	Number of objects
10-25	UTR-2	78-95	0-24	>-13	25..60	10000	1754
38	8C	90/95	0-24	>+60	4.5	1000	5859
74	VLSS	2004	0-24	> −30	1.3	350	32521
80	CUL1	73	0-24	-48,+35	3.7	2000	999
80	CUL2	75	0-24	-48,+35	3.7	2000	1748
82	IPS	87	0-24	-10,+83	27x350	500	1789
151	6CI	85	0-24	<+80	4.5	200	1761
151	6CII	88	8.5-17.5	+30,+51	4.5	200	8278
151	6CIII	90	5.5-18.3	+48,+68	4.5	200	8749
151	6CIV	91	0-24	+67,+82	4.5	200	5421
151	6CVa	93	1.6- 6.2	+48,+68	4.5	300	2229
151	6CVb	93	17.3-20.4	+48,+68	4.5	300	1229
151	6CVI	93	22.6- 9.1	+30,+51	4.5	300	6752
151	7CI	90	(10.5+41)	(6.5+45)	1.2	80	4723

Simulations









Data/method	ANN	CNN	GAN	SVM	DT	RF	DBSCAN	*k*-NN	*k*-M
Image	•	•	•	•	•	•		•	
Spectroscopy	•	•		•		•			•
Photometry	•				•	•	•		•
Light curve		•				•			
Time series	•	•			•	•	•		
Catalogue	•			•	•	•	•	•	
Simulation	•	•	•	•		•			

Notes: The table presents a summary of the types of astronomical data and the algorithms that appeared most regularly. The purpose of the table is to provide a convenient starting point for selecting an algorithm that has been used successfully for each data type.

Abbreviations: ANN, artificial neural network; CNN, convolutional neural network; DBSCAN, density-based spatial clustering of applications with noise; DT, decision tree; GAN, generative adversarial network; *k*-M, *k*-means clustering; *k*-NN, *k*-nearest neighbors; RF, random forest; SVM, support vector machine.

Fluke & Jacobs 2019

Applications at different levels of maturity

- Established
- Progressing
- Emerging

Classification and **forecasting** of solar flares. Segmentation for identification of umbra/penumbra/photosphere in the solar surface.

Identification of candidates to extrasolar planets from stellar lightcurves (Kepler)

Stellar and photometric **classification** of stars, leading to finding new objects of specific types of stars (WR, hot sub-dwarfs, etc).

Classification of galaxies from optical and radio imaging surveys. **Prediction** of physical properties from emission-line spectra. **Identification** of galaxies undergoing a special evolutionary phase as predicted by simulations.

Identification and **classification** of transient objects.

Accurate **estimation** of distance to extragalactic objects from photometric information (photometric redshift).

Identification of systems affected by gravitational lenses in wide-area surveys

Discriminating noise from signal in the detection of gravitational waves.

Applications at different levels of maturity

- Established
- Progressing
- Emerging

Reduction of false detections from the moving objects detection pipelines. **Detection** and **classification** of asteroids.

Assigning morphological types to radio-detected AGNs.
Identification of blazar candidates in catalogues of high-energy sources (Fermi-LAT)
Detecting high-redshift extremely luminous quasars
Discriminating populations of BAL QSOs from non-BAL QSOs

Examination of the output of cosmological simulations to **connect** physical properties of galaxies, dark matter halos and the cosmic environment.
Classification of DM sub-halos. **Assignment** of galaxies to halos in simulations.

Applications at different levels of maturity

- Established
- Progressing
- Emerging

Planetary metheorology: **Classification** of atmospheric features on the Surface of Mars aiming at **predicting** dust storms.
Discovery of previously unknown impact craters.

Study of the ISM in our Galaxy. Spatial or chemical **clustering** of components in atomic and molecular clouds.

Determination of dust reddenning in millions of stars, with application to GAIA data.
Discovery of new open clusters from overdensities in GAIA DR2 data

Identify faults in telescope drive systems that can be tackled in real time with automated expert systems

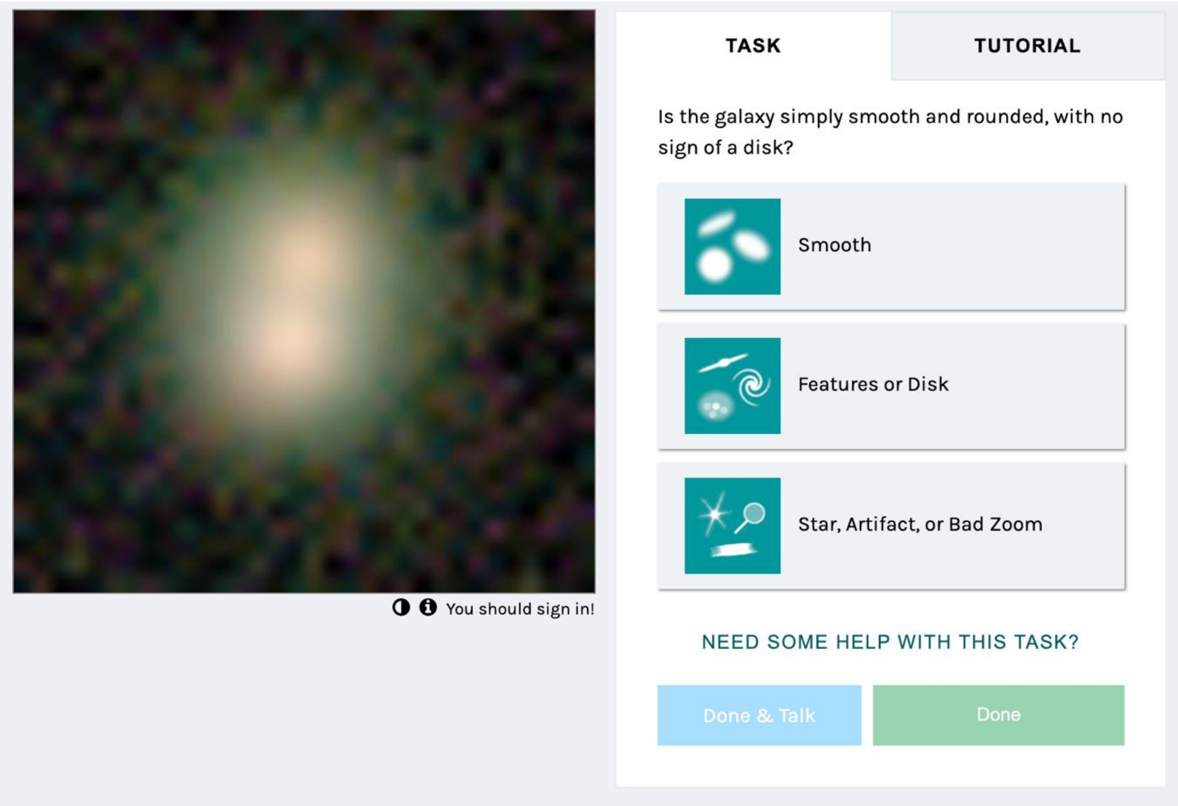
A few success stories…

Astronomy Research + Citizen Science

+

Pattern recognition and classification + Citizen Science

Galaxy Zoo @ Zooniverse



Goal: Train a model to classify morphologicaly galaxies based on their images

Datasets (> 300.000 galaxies):

- Sloan Digital Sky Survey (APO, US, 2.5m)
- Dark Energy Camera Legacy Survey (DECaLS) (V. Blanco Telescope, Chile, 4m)
- Hawaii H2O Survey (Subaru Telescope, US, 8.2m)
- Cosmic Evolution Early Research Science (CEERS) with JWST (Space, 6.5m)

Methodology:

- Train a classifier with labelled data
- Labels are put by thousands of volunteers with no specific field knowledge

Pattern recognition and classification + Citizen Science

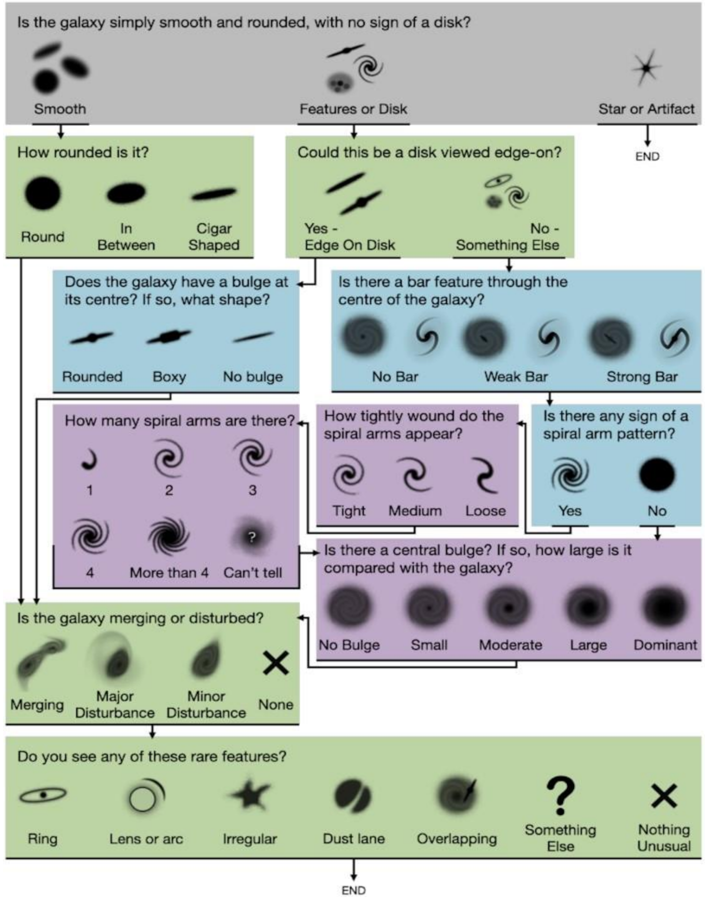


Figure 4. Classification decision tree for GZD-5, with new icons as shown to volunteers. Questions shaded with the same colours are at the same level of branching in the tree; grey have zero-dependent questions, green one, blue two, and purple three.

- 7.5M classifications(!) from which 140.000 get > 30 classifications
- Robust classifications
- Some work to be done to flag misclassifications
- This enabled producing >75 publications 2008-2023

Physical properties of speficic types of Galaxies

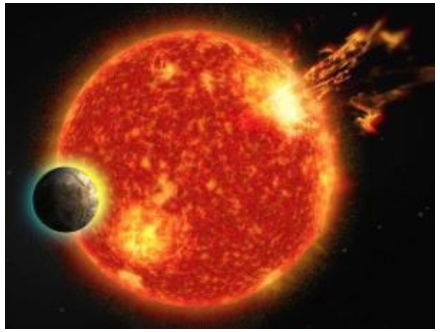
Morphology as a function of the environment

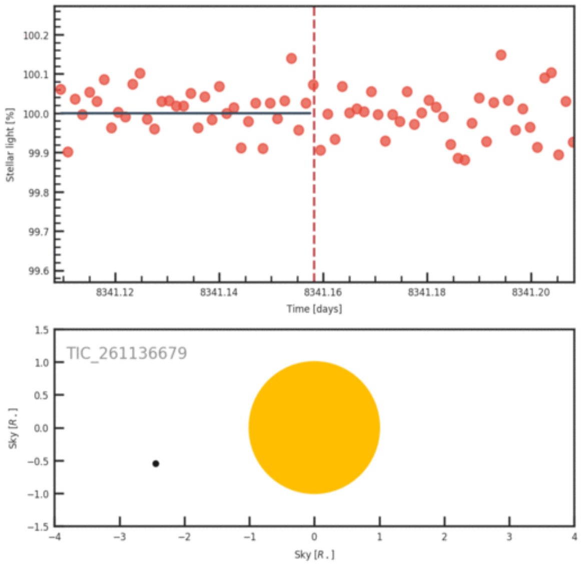
Statistics of orientation of Galaxies (spin)

Comparision with morphologies at other wavelengths

Pattern recognition and classification + Citizen Science

Planet Hunters TESS @ Zooniverse





Goal: Identify exoplanets by using the transit method.

Data: Lightcurves observed with dedicated space missions:
Kepler or TESS

Method:
- Build a classifier that automatically identifies potential candidates
- Volunteers help to identify tricky or borderline patterns, suggesting a classification (variable star, data glitch, potential planet)

Results:
- More than 100 new planetary systems identified in Kepler data

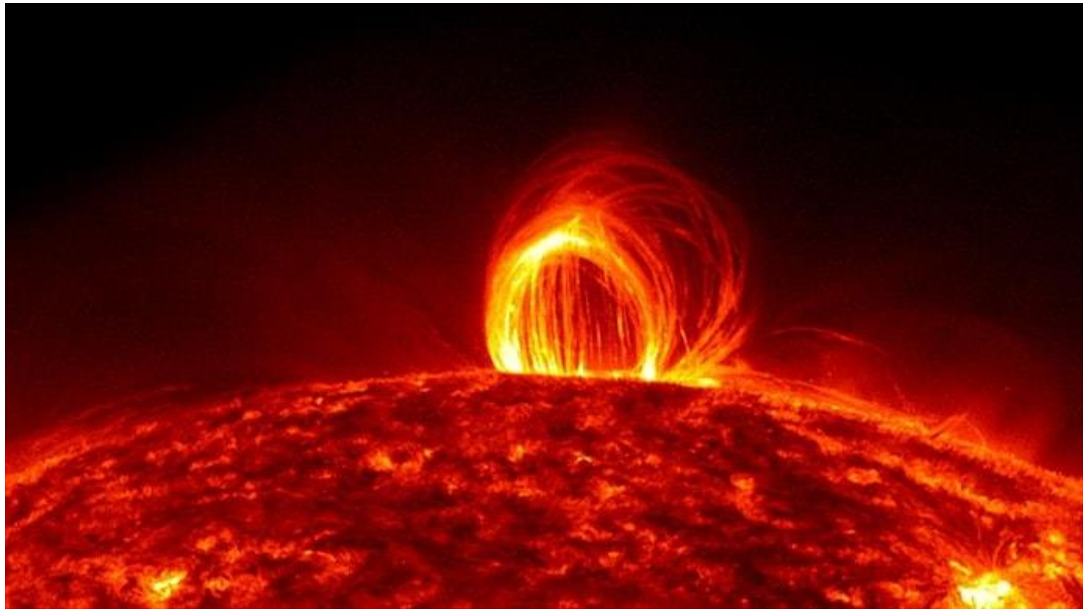
Classification and forecasting of solar flares

The Astrophysical Journal, 843:104 (14pp), 2017 July 10
https://doi.org/10.3847/1538-4357/aa789b

Predicting Solar Flares Using *SDO*/HMI Vector Magnetic Data Products and the Random Forest Algorithm

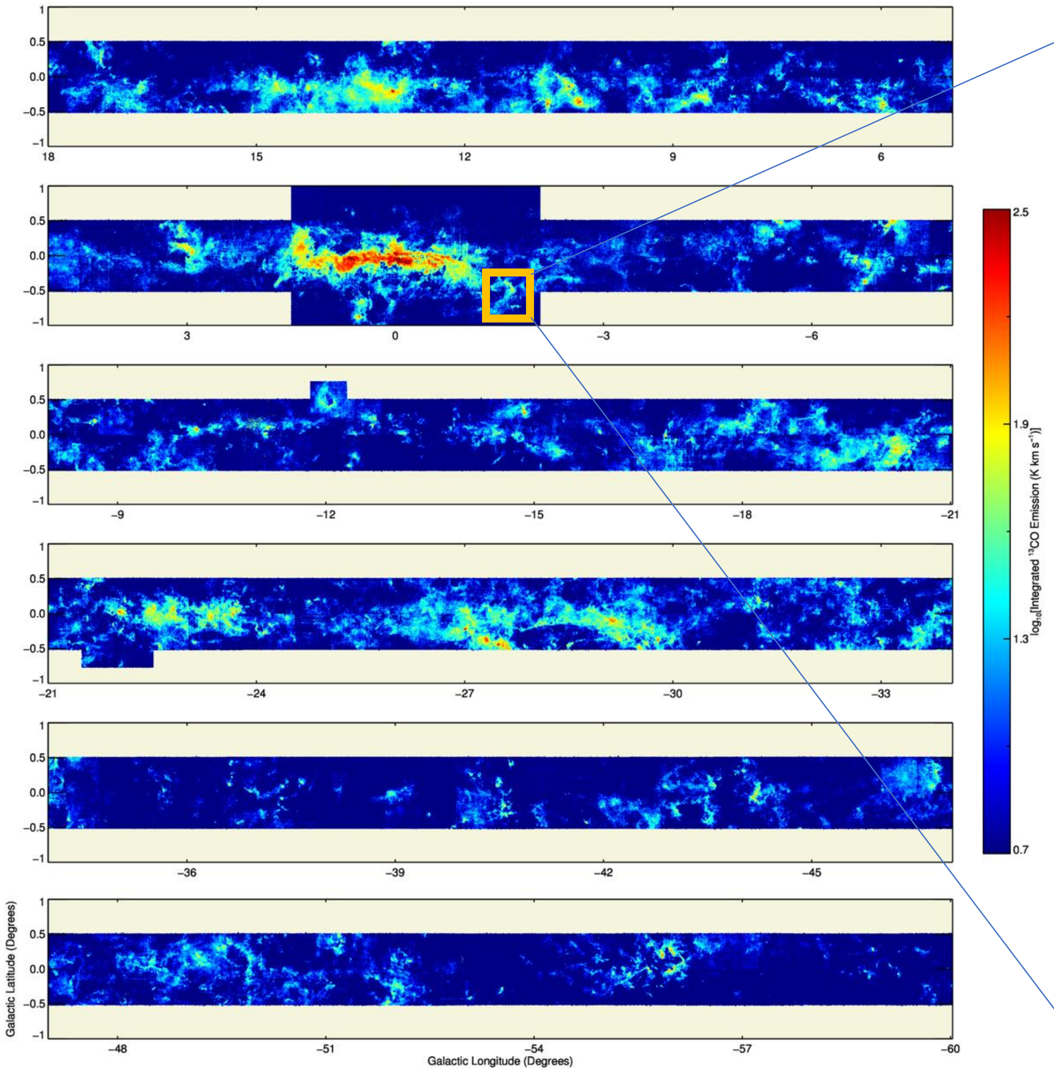
Chang Liu[1,2,3], Na Deng[1,2,3], Jason T. L. Wang[4], and Haimin Wang[1,2,3]
[1] Space Weather Research Laboratory, New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982, USA
chang.liu@njit.edu, na.deng@njit.edu, haimin.wang@njit.edu
[2] Big Bear Solar Observatory, New Jersey Institute of Technology, 40386 North Shore Lane, Big Bear City, CA 92314-9672, USA
[3] Center for Solar-Terrestrial Research, New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982, USA
[4] Department of Computer Science, New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982, USA; jason.t.wang@njit.edu
Received 2016 May 6; revised 2017 June 7; accepted 2017 June 7; published 2017 July 11

- Catalog of actual solar X-ray flares happened between 2010 and 2016
- Categorization of the solar flare emerging regions in 4 categories (B, C, M, X) according to their intensity.
- Building a data base with a set of physical parameters extracted from the HMI instrument at the Solar Dynamics Observatory (SDO) at the beginning of each flare.
- Train a RF classifier with these data to predict the occurrence of the different categories of flares.

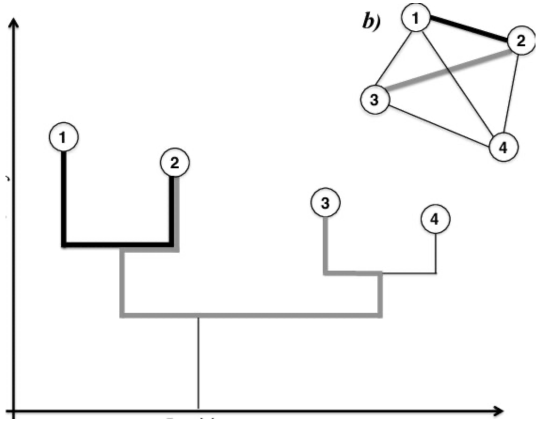


SEDIGISM*

*Structure, Excitation and Dynamics of the Inner Galactic Interstellar Medium



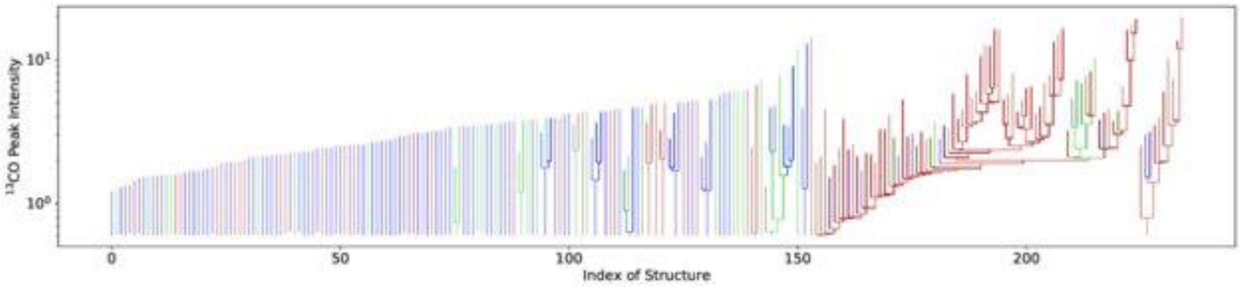
F. Schuller et al, A&A 601, A124 (2017)



SCIMES segmentation algorithm based on graph theory

D. Colombo et al. (2015). MNRAS, 454, 2067

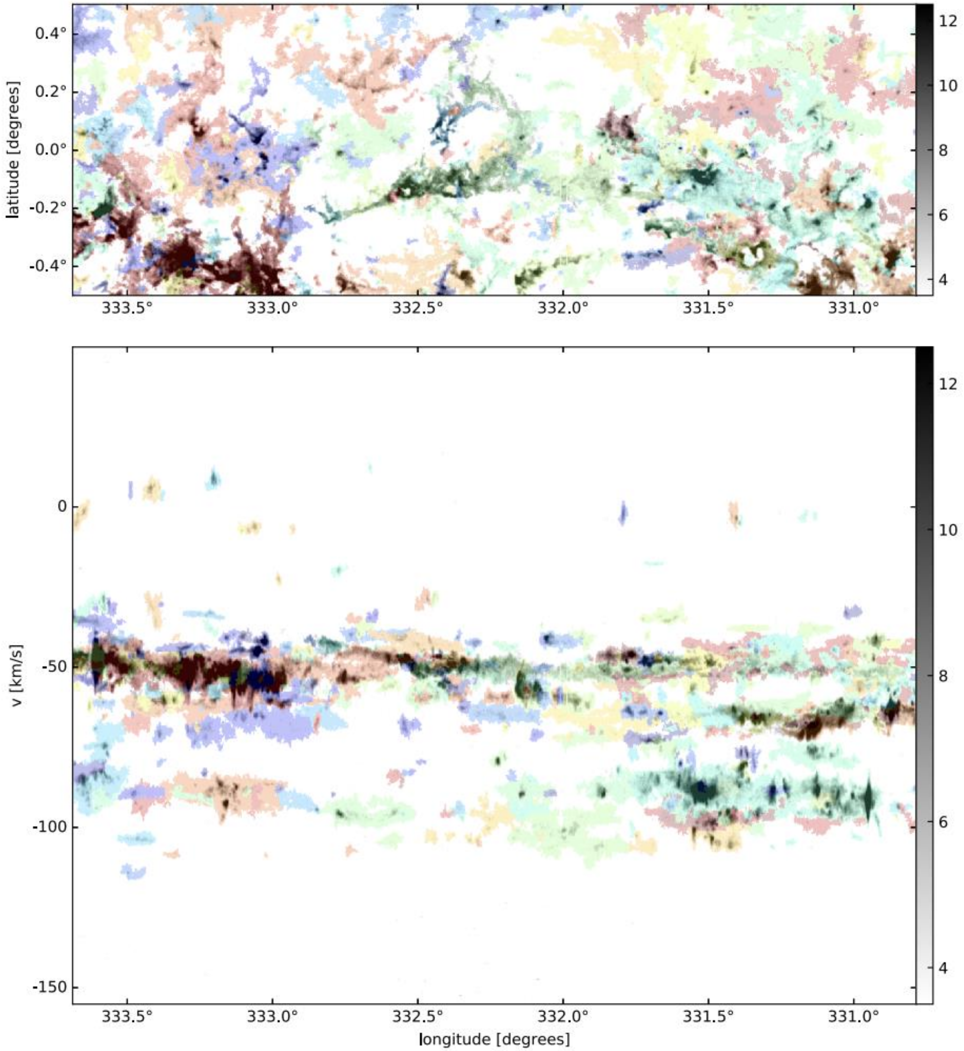
Dendrogram for G305



P. Mazumbar et al. (2023) A&A 656, A101 (2021)

SEDIGISM*

*Structure, Excitation and Dynamics of the Inner Galactic Interstellar Medium

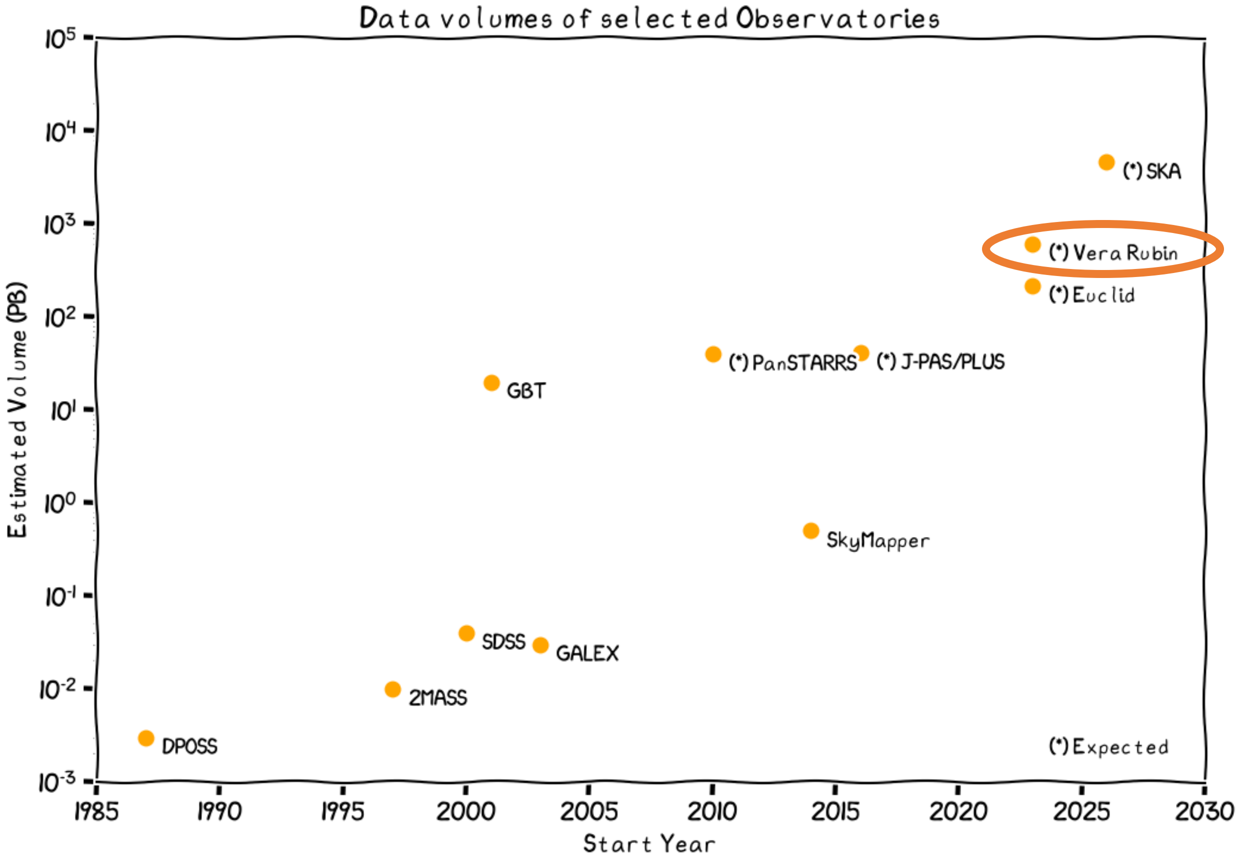


A. Duarte-Cabral et al. (2021). MNRAS, 500, 3027

- Catalog of >10.000 molecular clouds identified, most of them with determination of distances
- Determination of the physical properties of the clouds.
- Study of the distribution and dynamics of clouds as a function of the arm structures known in our Galaxy

What is coming…

LSST: Massive data processing in quasi real time



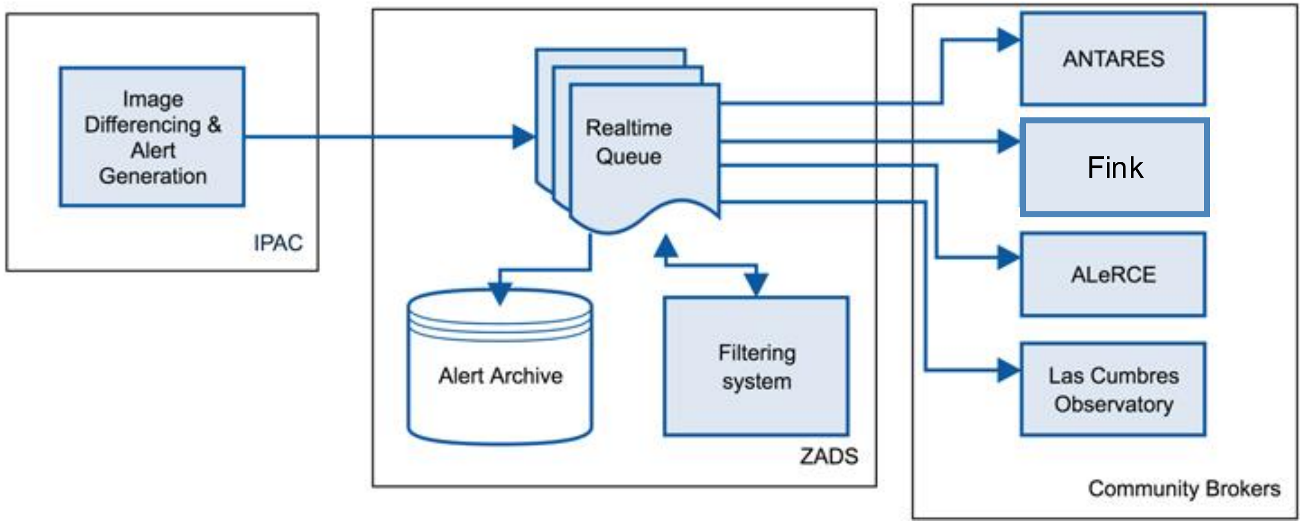
LSST: Massive data processing in near real time

VERA C. RUBIN
OBSERVATORY

- ❑ Cover all the visible sky every 2-3 nights
- ❑ Exhaustive study of the transient sky. What has changed?
- ❑ About 10 million of alerts per night (20.000 alerts per minute)
- ❑ Latency of alert: 60 seconds
- ❑ Classification of alerts are crucial to decide on the workflow to follow
    - ❖ Supernova?
    - ❖ NEO?
    - ❖ Variable star?
    - ❖ Micro lensing?
    - ❖ New cool stuff?

LSST prototype: Zwicky Transient Facility (ZTF)

Several brokers becoming specialized in classification of alerts by ZTF



ALeRCE Automatic Learning for the Rapid Classification of Events

https://alerce.science/

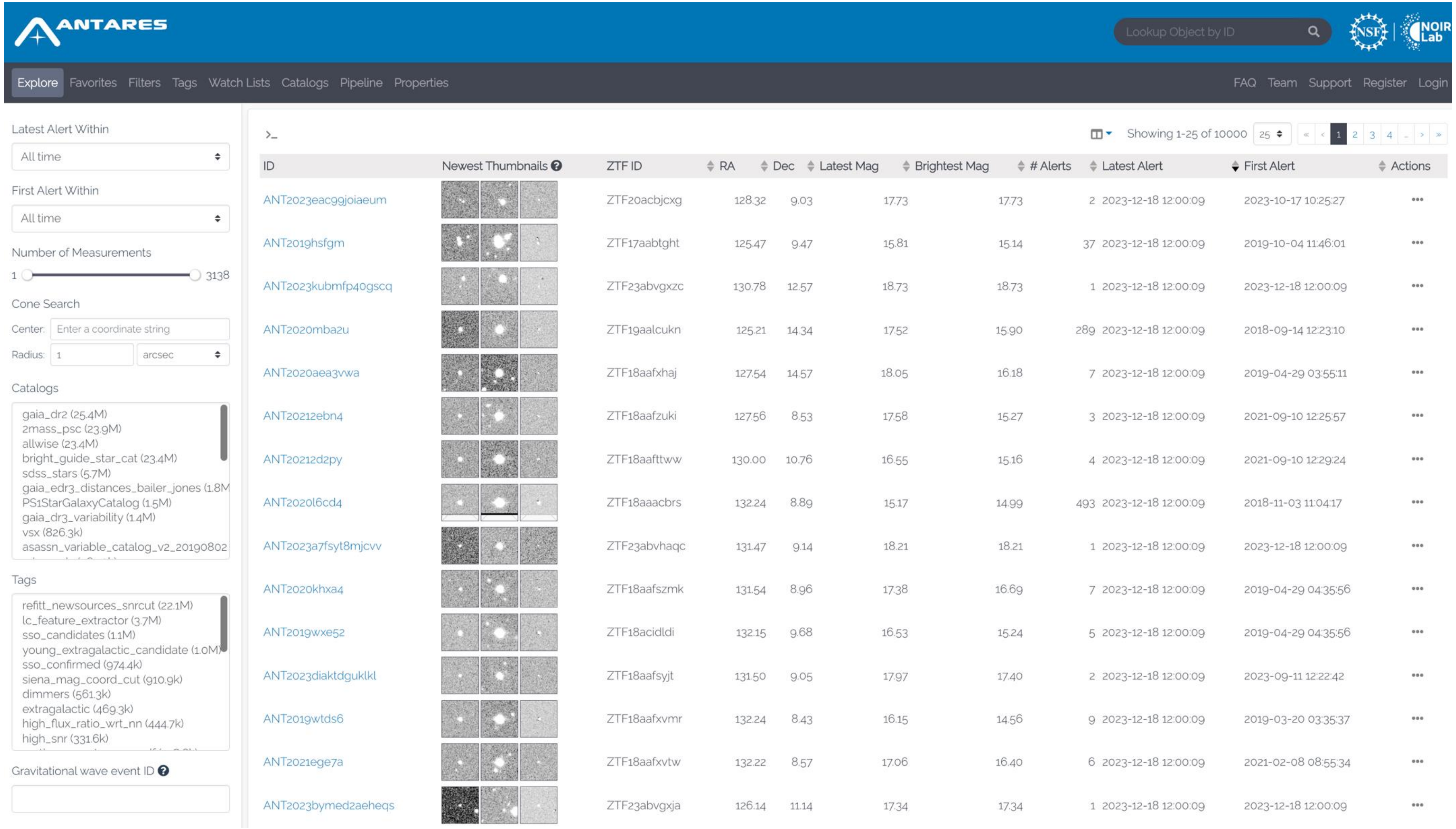
ANTARES

https://antares.noirlab.edu/

FINK

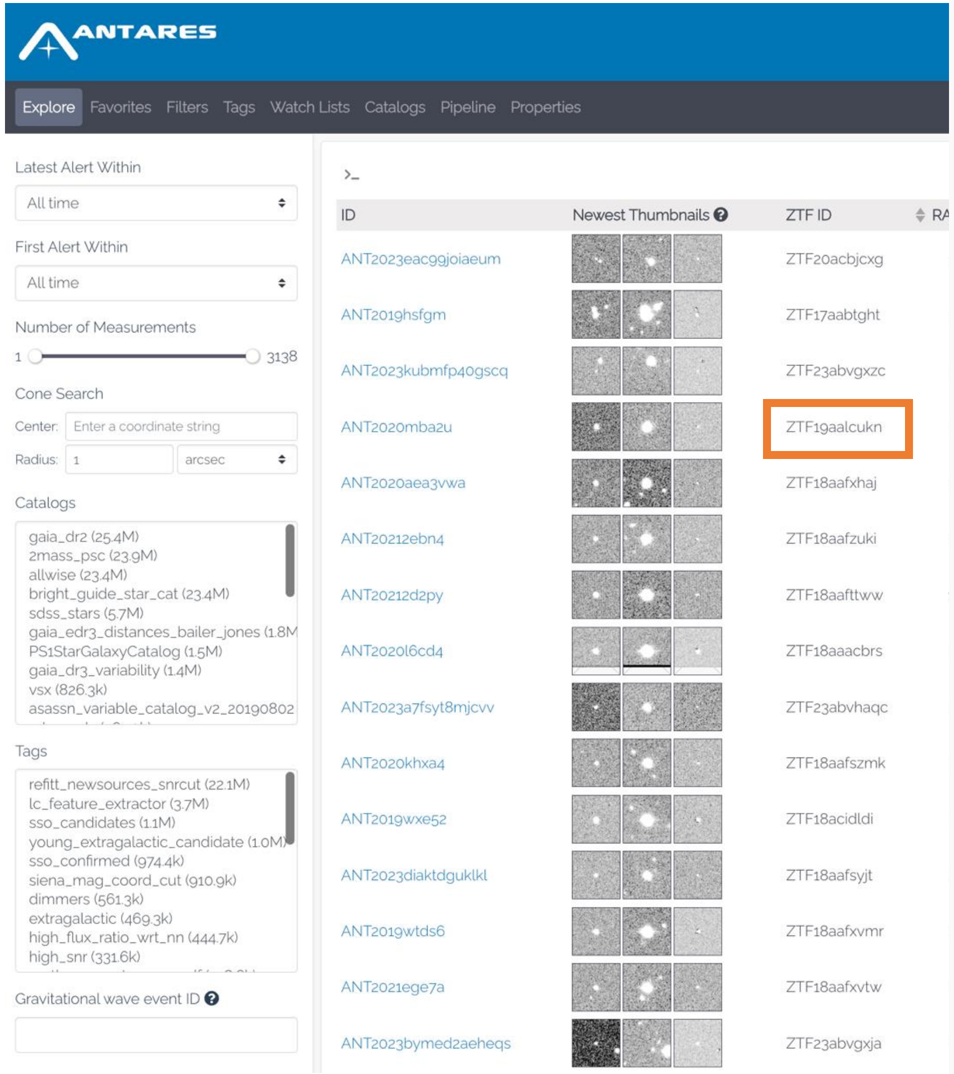
https://fink-broker.org/

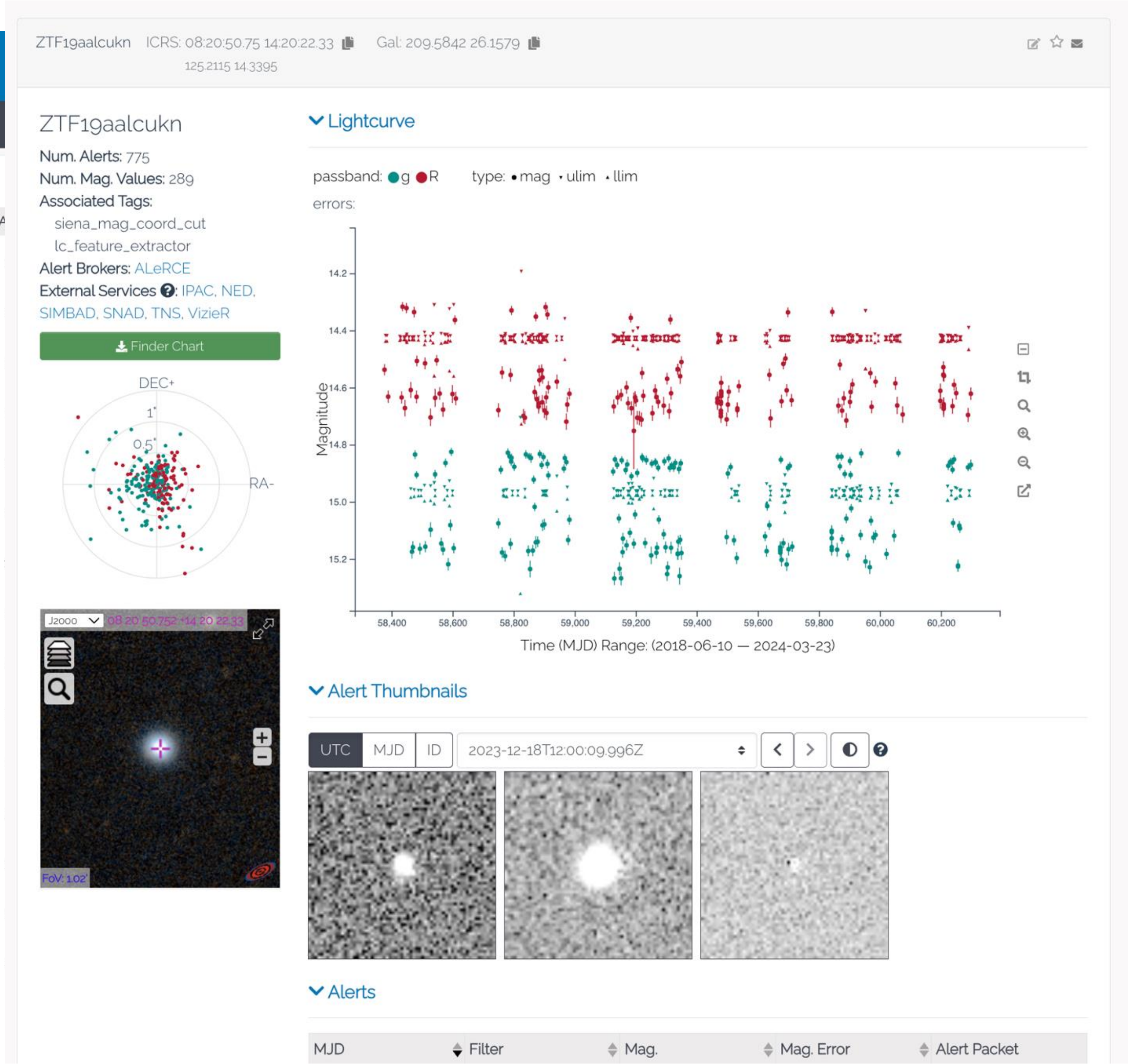
LSST prototype: Zwicky Transient Facility (ZTF)



ID	Newest Thumbnails	ZTF ID	RA	Dec	Latest Mag	Brightest Mag	# Alerts	Latest Alert	First Alert	Actions
ANT2023eacggjoiaeum		ZTF20acbjcxg	128.32	9.03	17.73	17.73	2	2023-12-18 12:00:09	2023-10-17 10:25:27	•••
ANT2019hsfgm		ZTF17aabtght	125.47	9.47	15.81	15.14	37	2023-12-18 12:00:09	2019-10-04 11:46:01	•••
ANT2023kubmfp40gscq		ZTF23abvgxzc	130.78	12.57	18.73	18.73	1	2023-12-18 12:00:09	2023-12-18 12:00:09	•••
ANT2020mba2u		ZTF19aalcukn	125.21	14.34	17.52	15.90	289	2023-12-18 12:00:09	2018-09-14 12:23:10	•••
ANT2020aea3vwa		ZTF18aafxhaj	127.54	14.57	18.05	16.18	7	2023-12-18 12:00:09	2019-04-29 03:55:11	•••
ANT2021zebn4		ZTF18aafzuki	127.56	8.53	17.58	15.27	3	2023-12-18 12:00:09	2021-09-10 12:25:57	•••
ANT2021zd2py		ZTF18aafttww	130.00	10.76	16.55	15.16	4	2023-12-18 12:00:09	2021-09-10 12:29:24	•••
ANT2020l6cd4		ZTF18aaacbrs	132.24	8.89	15.17	14.99	493	2023-12-18 12:00:09	2018-11-03 11:04:17	•••
ANT2023a7fsyt8mjcvv		ZTF23abvhaqc	131.47	9.14	18.21	18.21	1	2023-12-18 12:00:09	2023-12-18 12:00:09	•••
ANT2020khxa4		ZTF18aafszmk	131.54	8.96	17.38	16.69	7	2023-12-18 12:00:09	2019-04-29 04:35:56	•••
ANT2019wxe52		ZTF18acidldi	132.15	9.68	16.53	15.24	5	2023-12-18 12:00:09	2019-04-29 04:35:56	•••
ANT2023diaktdguklkl		ZTF18aafsyjt	131.50	9.05	17.97	17.40	2	2023-12-18 12:00:09	2023-09-11 12:22:42	•••
ANT2019wtds6		ZTF18aafxvmr	132.24	8.43	16.15	14.56	9	2023-12-18 12:00:09	2019-03-20 03:35:37	•••
ANT2021ege7a		ZTF18aafxvtw	132.22	8.57	17.06	16.40	6	2023-12-18 12:00:09	2021-02-08 08:55:34	•••
ANT2023bymed2aeheqs		ZTF23abvgxja	126.14	11.14	17.34	17.34	1	2023-12-18 12:00:09	2023-12-18 12:00:09	•••

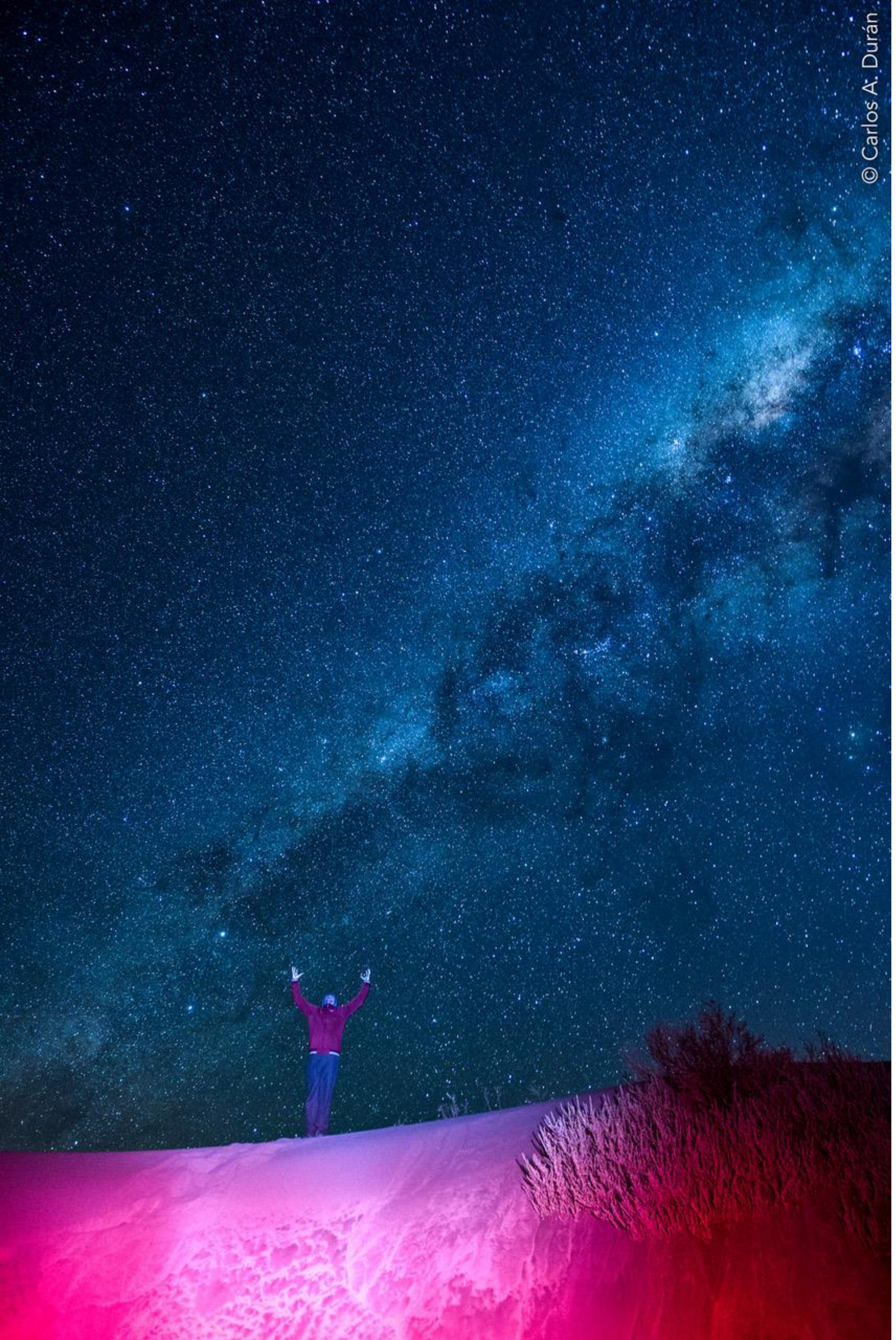
LSST prototype: Zwicky Transient Facility (ZTF)





Bibliography

- Dalya Baron (2019). *Machine Learning in Astronomy: A practical overview.* arXiv:1904.07248v1
- Christopher J. Fluke & Colin Jacobs (2019). *Surveying the reach and maturity of machine learning and artificial intelligence in astronomy*. WIREs Data Mining Knowl Discov. 2020;10:e1349. https://doi.org/10.1002/widm.1349
- Željko Ivezić, Andrew J. Connolly, Jacob T. Vanderplas & Alexander Gray (2020). *Statistics, data mining and machine learning in astronomy: A practical Python guide for the analysis of survey data (2nd Ed)*. Cambridge University Press.
- José V. Rodríguez, Ignacio Rodríguez-Rodríguez, Wai Lok Woo (2022). *On the application of machine learning in astronomy and astrophysics: A text-mining-based scientometric analysis*. WIREs Data Mining Knowl Discov. 2022;12:e1476. https://doi.org/10.1002/widm.1476
- Marc Huertas-Company & François Lanusse (2023). *The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys.* Publications of the Astronomical Society of Australia, 40, e001. doi:10.1017/pasa.2022.55
- Michael J. Smith & James E. Geach (2023). *Astronomia ex machina: a history, primer and outlook on neural networks in astronomy*. R. Soc. Open Sci. 10: 221454. https://doi.org/10.1098/rsos.221454
- XXX Canary Islands Winter School on Astrophysics (2018). *eBook XXX Winter School: summaries, talks, references, and tutorials*. https://meetings.iac.es/winterschool/2018



Thank you!