Discovering Radio Transients with Machine Learning and Citizen Science

Alex Andersson, with Rob Fender and the ThunderKATs, Chris Lintott and the Zooniverse Team, plus 1000s of volunteers!

Warwick- Machine Learning for Transient Science, Dec 2023

alexander.andersson@physics.ox.ac.uk

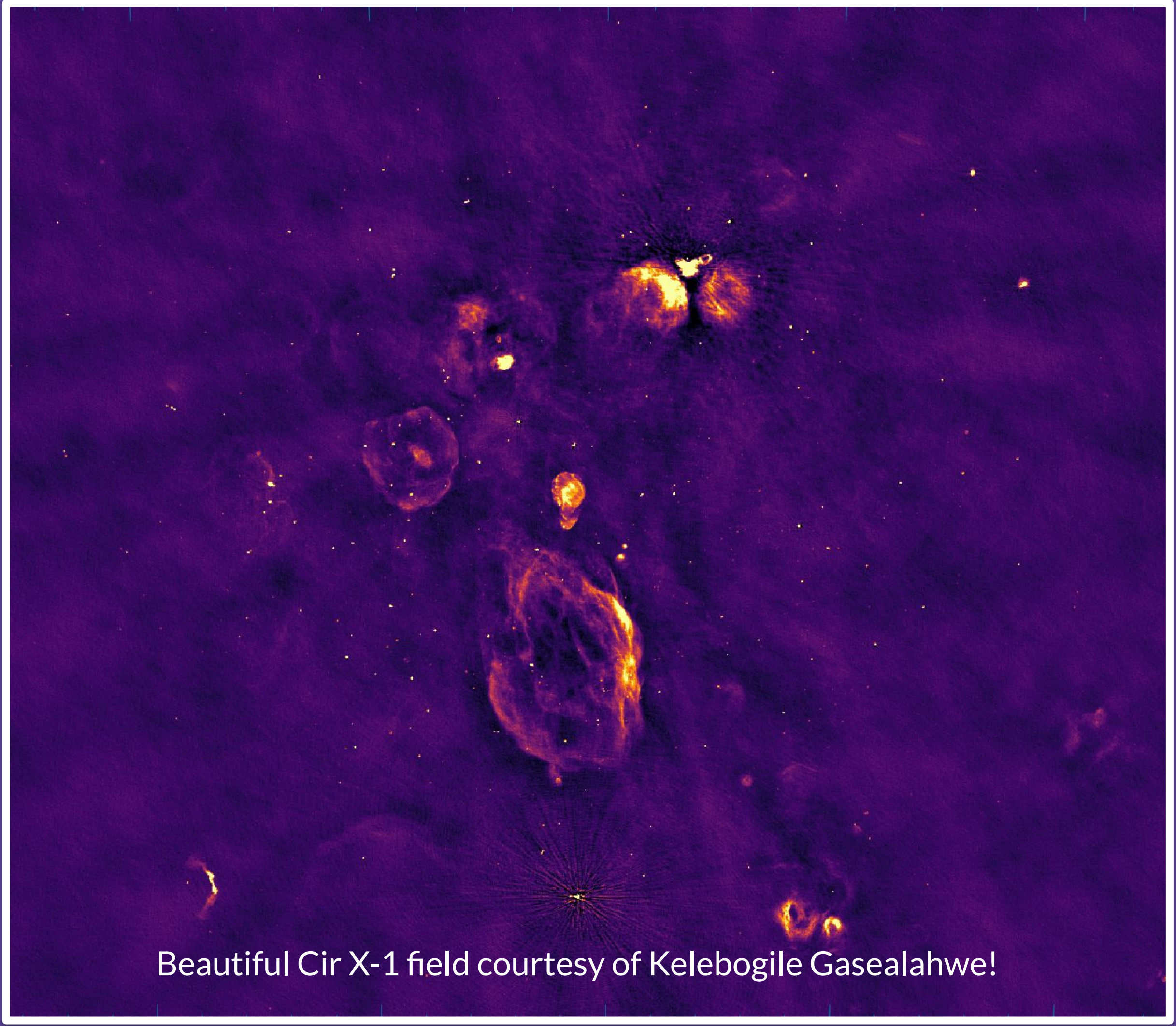
Commensalism

<u>0 hrs additional telescope time</u>

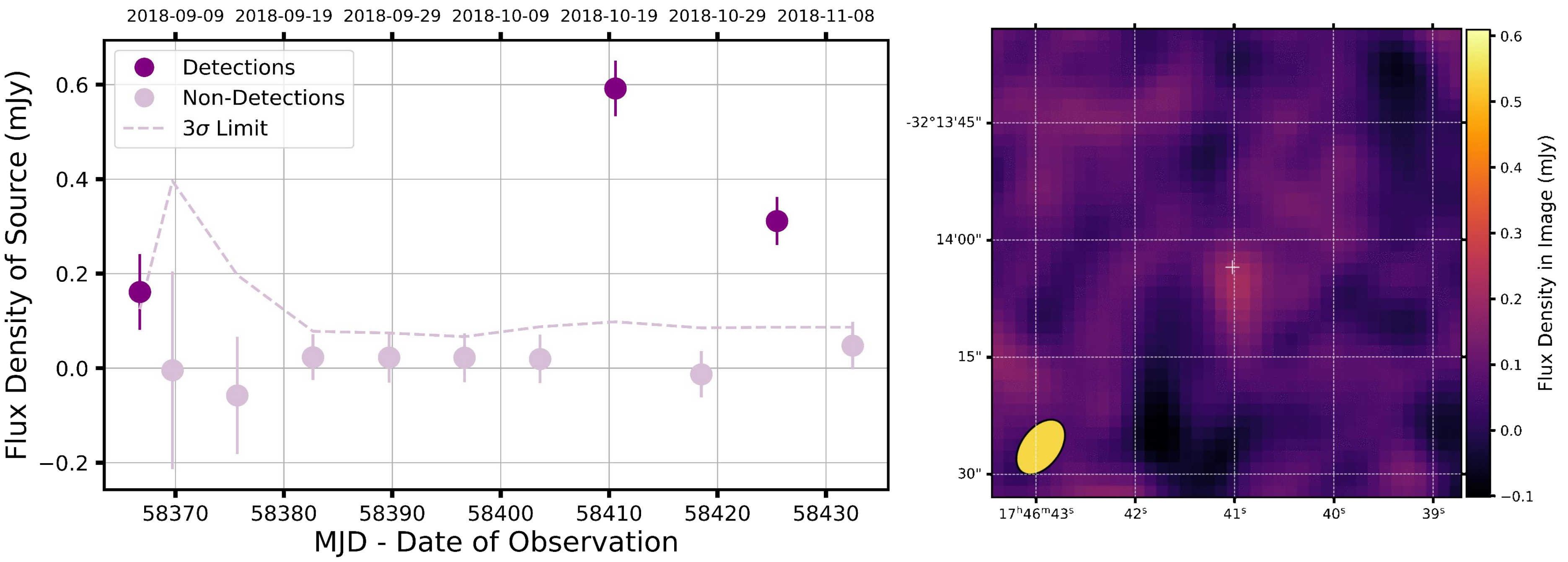
ThunderKAT Survey on MeerKAT (1.3 GHz)

25μJy in 15 mins

(~21 mag)

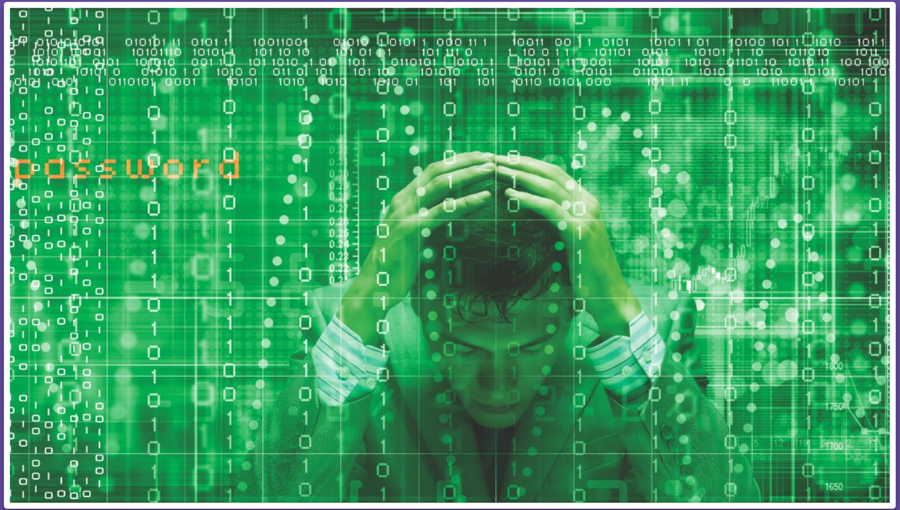


Beautiful Cir X-1 field courtesy of Kelebogile Gasealahwe!



Example - flaring M dwarf found serendipitously (see Andersson et al. 2022, 2204.03481)

Data Overload



ThunderKAT: ~½ TB of raw data / observation

500 images every year

100s-1000s of sources / image

Plus all the deep imaging fields (LADUMA, MIGHTEE, MHONGOOSE, …)

What about fast-imaging - (3600 images in an 8-hour track)

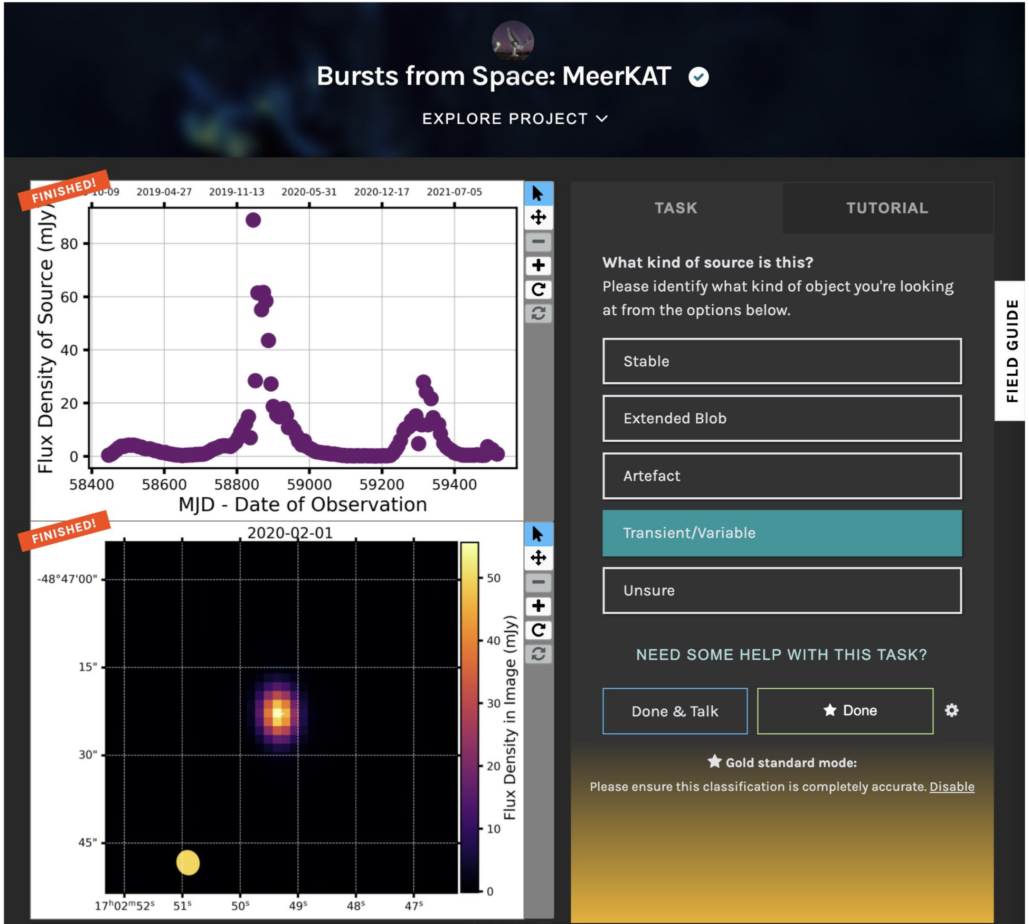
I can't check every single light curve myself!

How do we find transients? Citizen Scientists

We generate light curves and an image of each source and present them to citizen scientists on the Zooniverse.

~9,000 sources to classify, 10 votes / source.

Lots of help available, through a Field Guide, Tutorial, Help Text and Talk forums.



BfS:MKT workflow to classify radio sources.

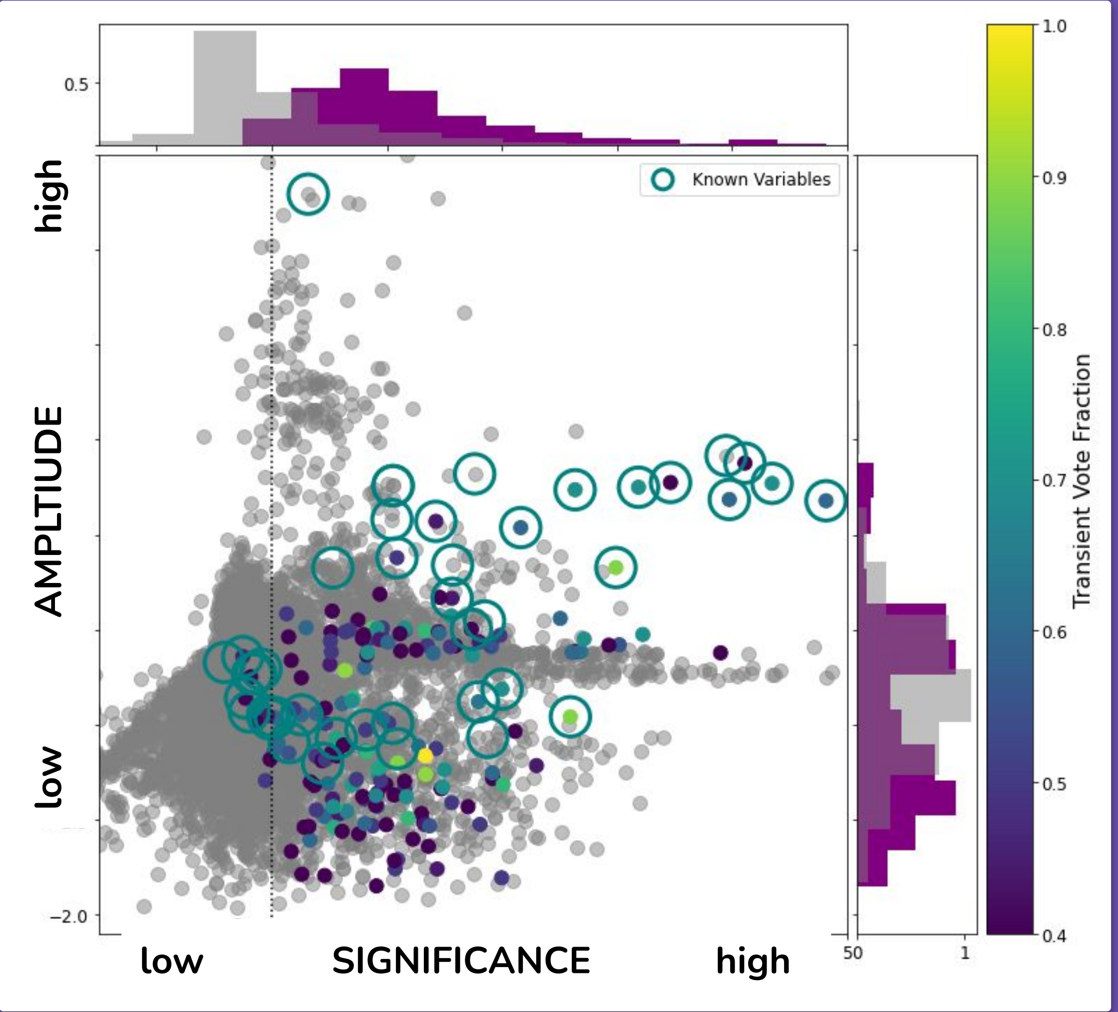
What transients do citizen scientists find?

$$V_\nu \equiv \frac{s_\nu}{\overline{F_\nu}}$$

Quantifies the amplitude of variability

$$\eta_\nu \equiv \chi^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(F_{i,\nu}-\overline{F_\nu}^*)^2}{\sigma_i^2}$$

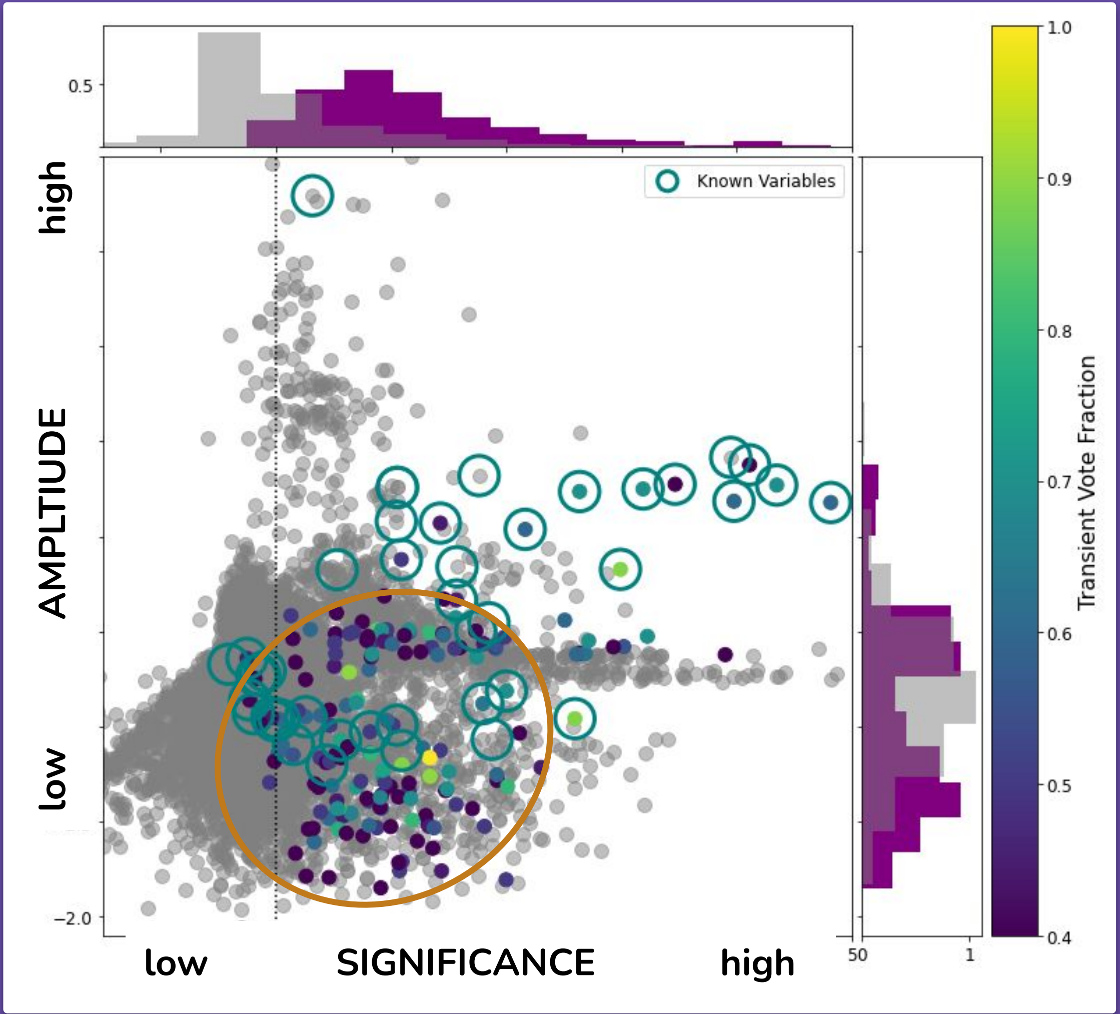
Quantifies the statistical significance of variability, compared to a ‘constant’ source.



What transients do citizen scientists find?

Volunteers find variables and transients that we would never otherwise have found.

That's 142 interesting sources, not before seen to be varying or interesting.



Volunteers get the credit!

Volunteers can and do find hundreds of transients and variables!

All the data underlying this work (lots of light curves) is available online (AnderssonAstro on GitHub, or see paper)

<u>**DR2 has just finished!**</u>
LADUMA / ECDFS deep imaging. Fast timescale work to come.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

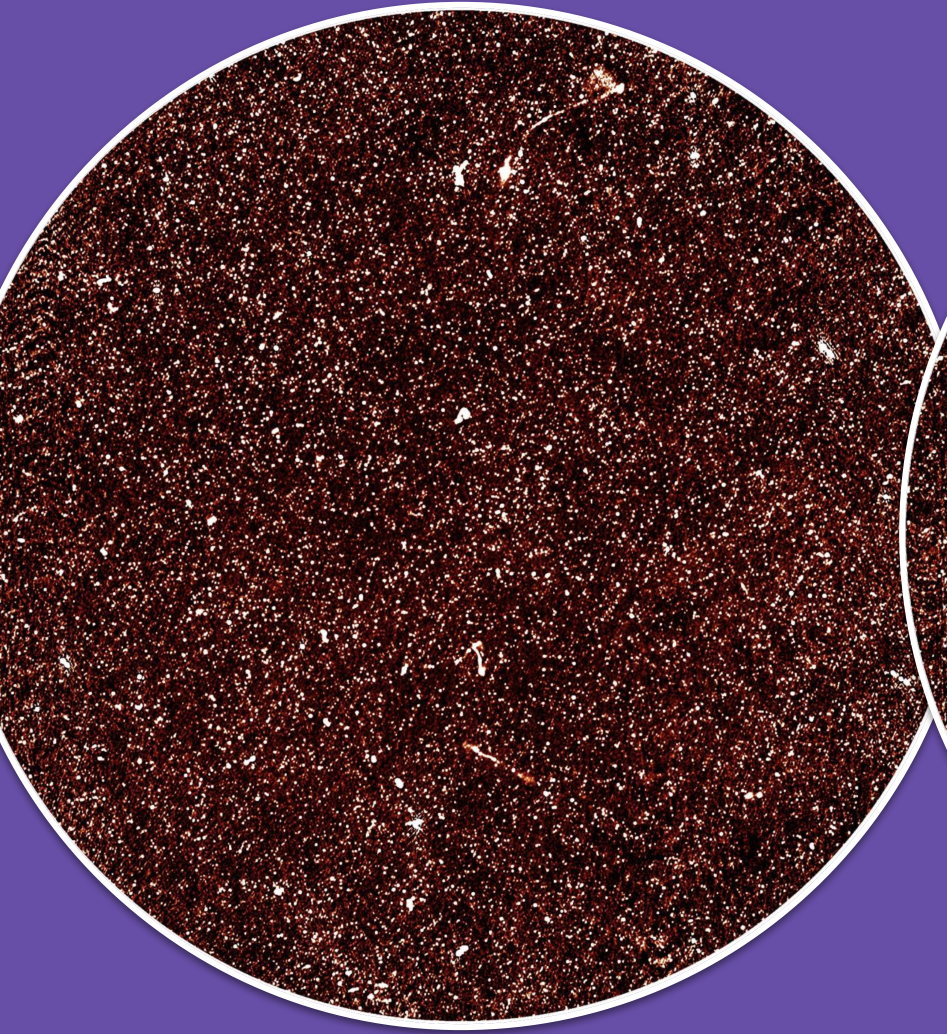
MNRAS **523**, 2219–2235 (2023)
Advance Access publication 2023 May 11

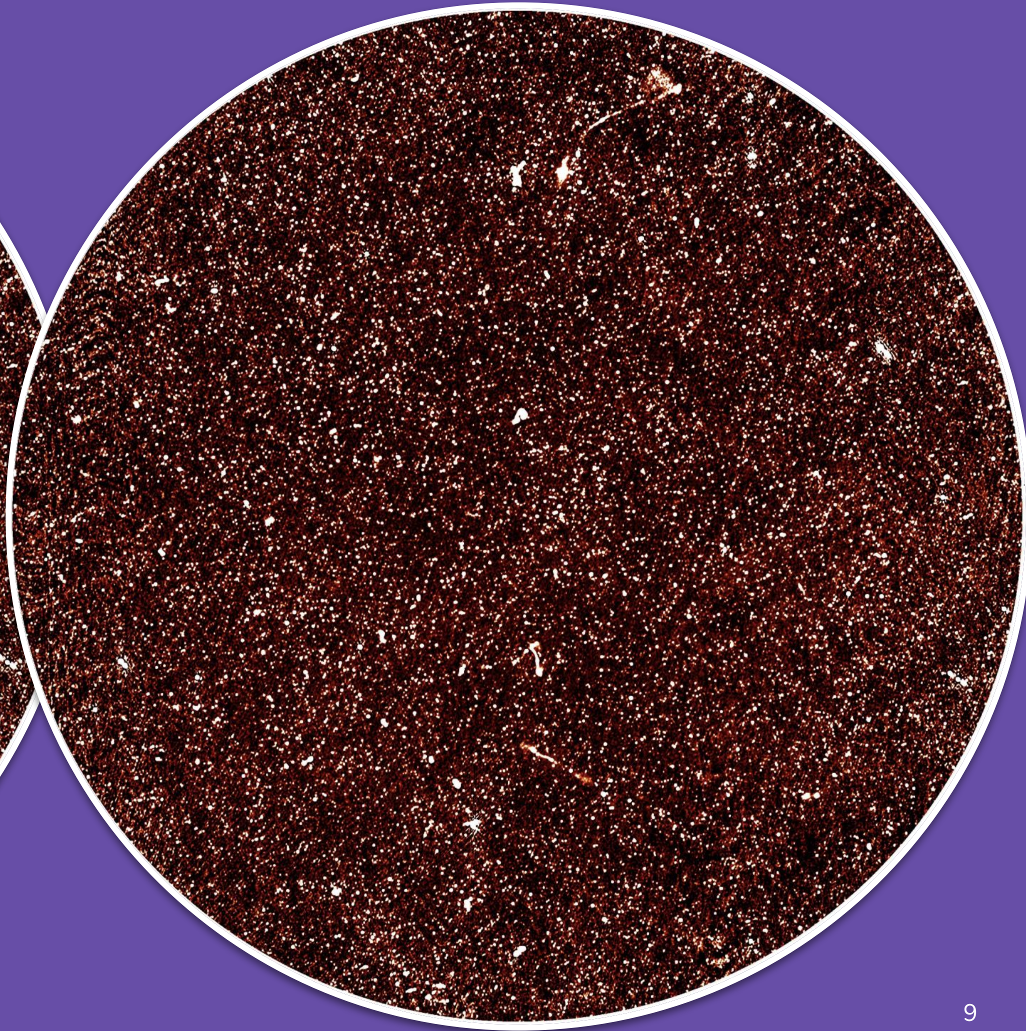
Bursts from Space: MeerKAT – the first citizen science project dedicated to commensal radio transients

Alex Andersson,[1★] Chris Lintott,[1] Rob Fender,[1,2] Joe Bright,[1] Francesco Carotenuto,[1] Laura Driessen,[3,4,5] Mathilde Espinasse,[6] Kelebogile Gasealahwe,[2,7] Ian Heywood,[1,8,9] Alexander J. van der Horst,[10,11] Sara Motta,[12] Lauren Rhodes,[1] Evangelia Tremou,[13] David R. A. Williams,[4] Patrick Woudt,[2] Xian Zhang,[14,15] Steven Bloemen,[16] Paul Groot,[2,7,16] Paul Vreeswijk,[16] Stefano Giarratana,[17,18] Payaswini Saikia,[19] Jonas Andersson,[20] Lizzeth Ruiz Arroyo,[20] Loïc Baert,[20] Matthew Baumann,[20] Wilfried Domainko,[20] Thorsten Eschweiler,[20] Tim Forsythe,[20] Sauro Gaudenzi,[20] Rachel Ann Grenier,[20] Davide Iannone,[20] Karla Lahoz,[20] Kyle J. Melville,[20] Marianne De Sousa Nascimento,[20] Leticia Navarro,[20] Sai Parthasarathi,[20] Piilonen,[20] Najma Rahman,[20] Jeffrey Smith,[20] B. Stewart,[20] Newton Temoke,[20] Chloe Tworek[20] and Isabelle Whittle[20]

Affiliations are listed at the end of the paper

Accepted 2023 April 20. Received 2023 April 27; in original form 2023 March 14





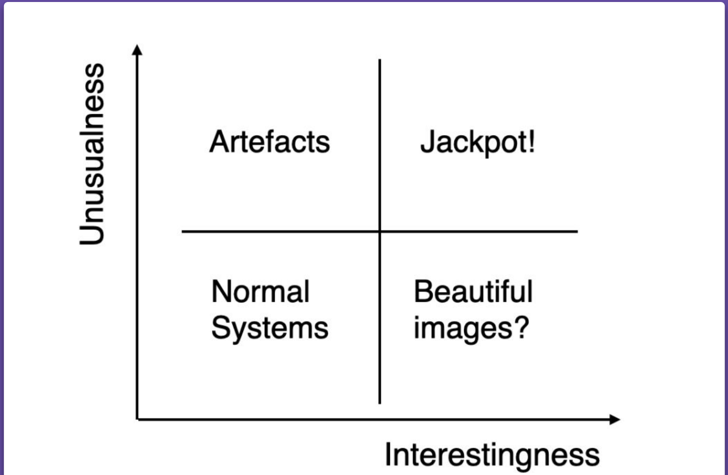
Machine Learning to the rescue!

Big Data + Citizen Science Labels => Machine Learning

Used lots when finding and classifying optical transients.

Unsupervised with volunteer votes as ground truth.

Do algorithmic anomalies match volunteer-verified transients?



Credit: Chris L.

Lochner and Bassett 2021

/MichelleLochner/astronomaly

Features

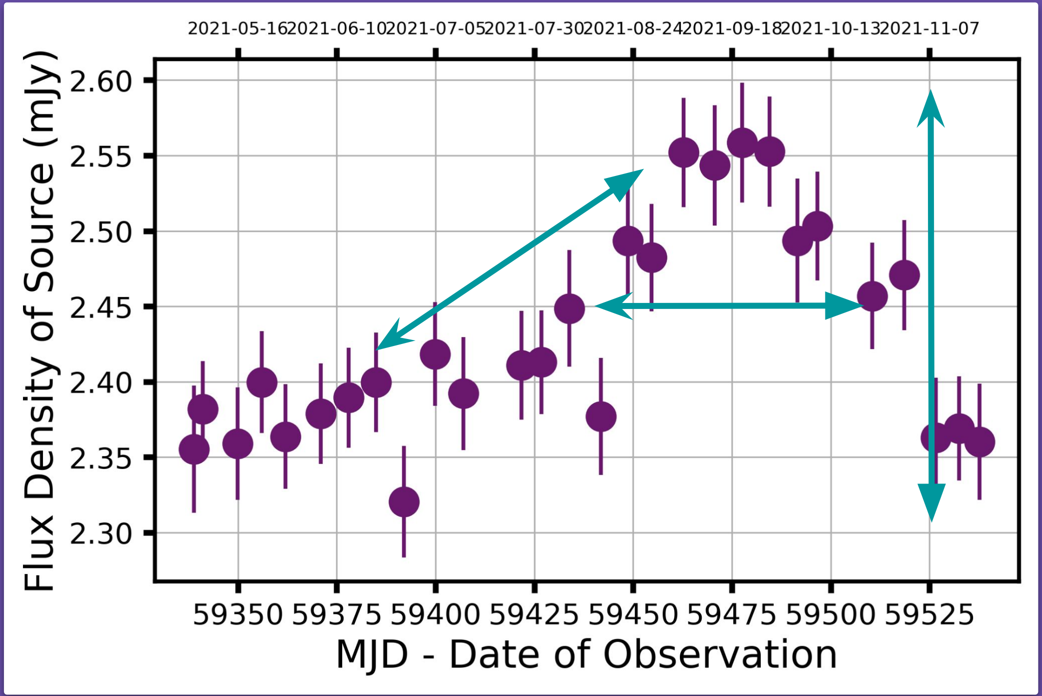
Ways of describing your data.

We use **three** main feature sets:

A 2-parameter ‘fiducial’ set using (η, V).

45 Feets package features (Cabral+ 2018, ascl:1806.001)

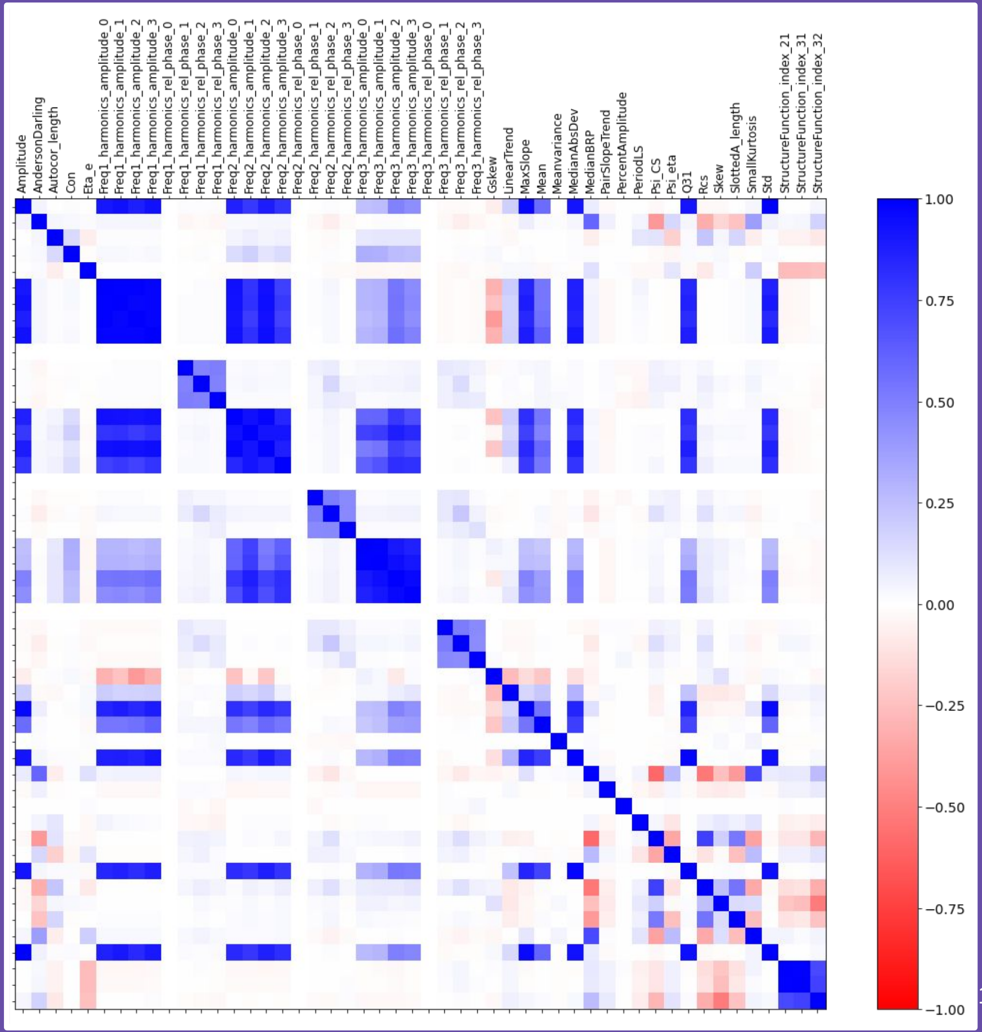
- LC measurements (amplitude, width, slopes)
- Low order Fourier terms
- Statistics e.g. mean, M.A.D, Q31, skew, kurtosis...



/quatrope/feets

Features

Correlation matrix between all of the Feets features.

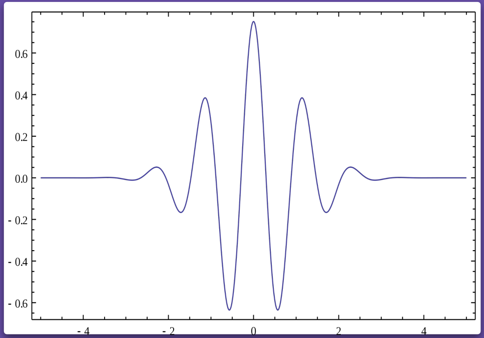


Where things are uncorrelated we can hope they probe different kinds of variability in light curves.

Features

40 Wavelet Features:

$$f(t) = \sum_i a_i \phi_i(t)$$



Gaussian Process Regression to interpolate all our light curves and have them starting at t = 0 days.

We sample these now-uniform light curves every ~week

Decompose into a set of basis functions and coefficients.

Think Fourier decomp, but now our basis function are wavelets.

Lochner et al. 2016

Use principal component analysis to reduce this overload of coefficients to the 40 components that describe the most of the variance.

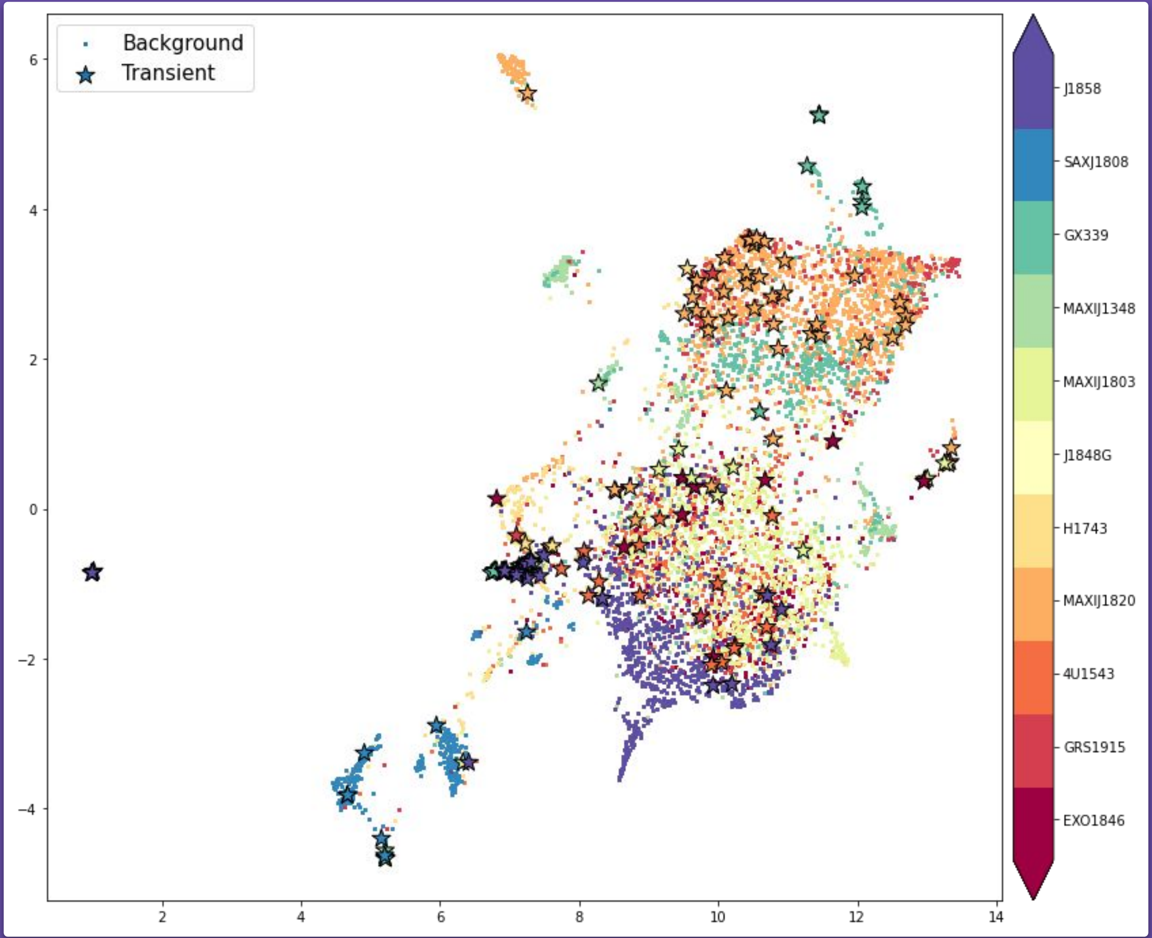
/LSSTDESC/snmachine

UMAP - Feets

Project our 45D data down to 2D and see what kinds of structure exist.

Global structure determined by differences between fields (cadence, noise, systematics)

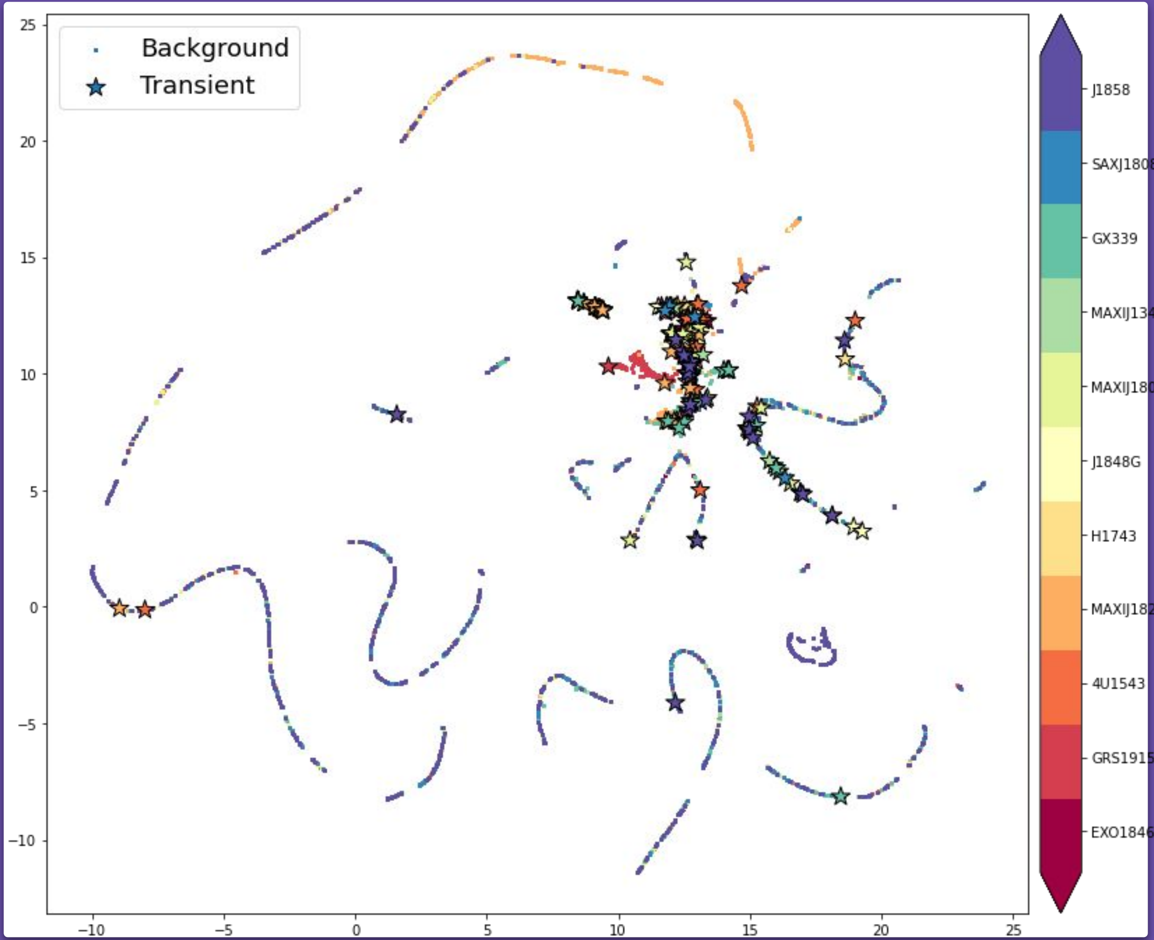
Aim of the game: separate the transients from the rest.



UMAP - Wavelets

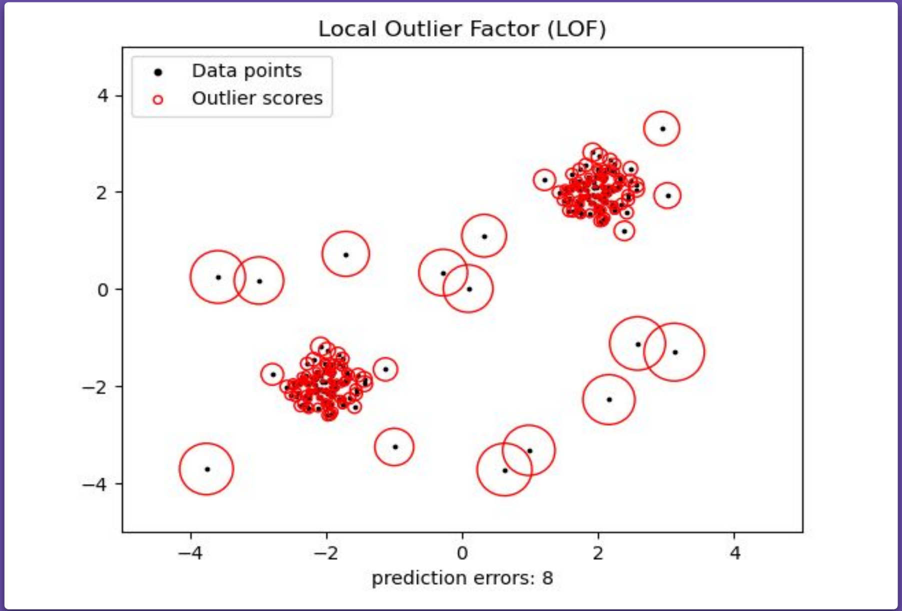
Project our 40D data down to 2D and see what kinds of structure exist.

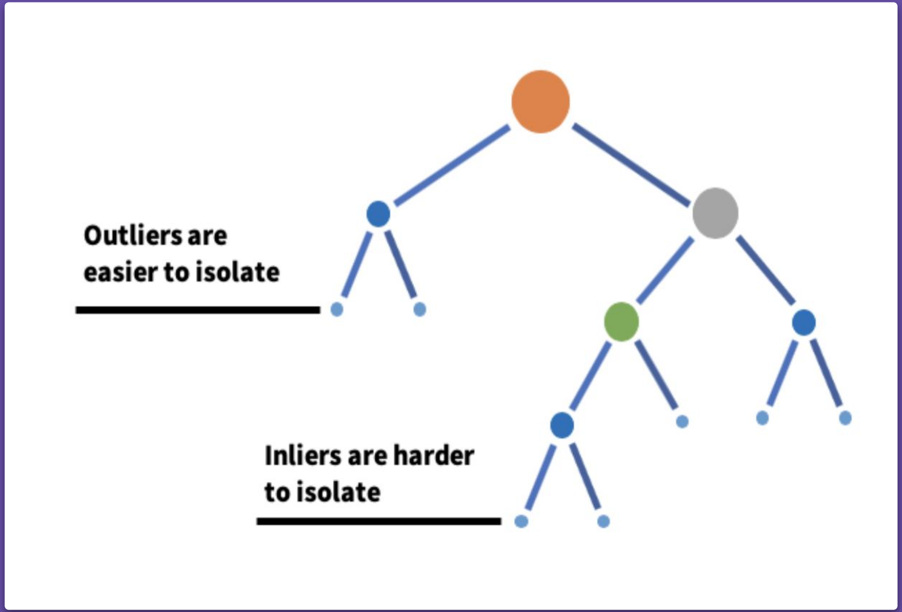
Again a field-dependence, with longer fields in these fun wiggles.



Aim of the game: separate the transients from the rest.

Finding anomalies in our data - Astronomaly





Both return a list of our sources from most-to-least anomalous

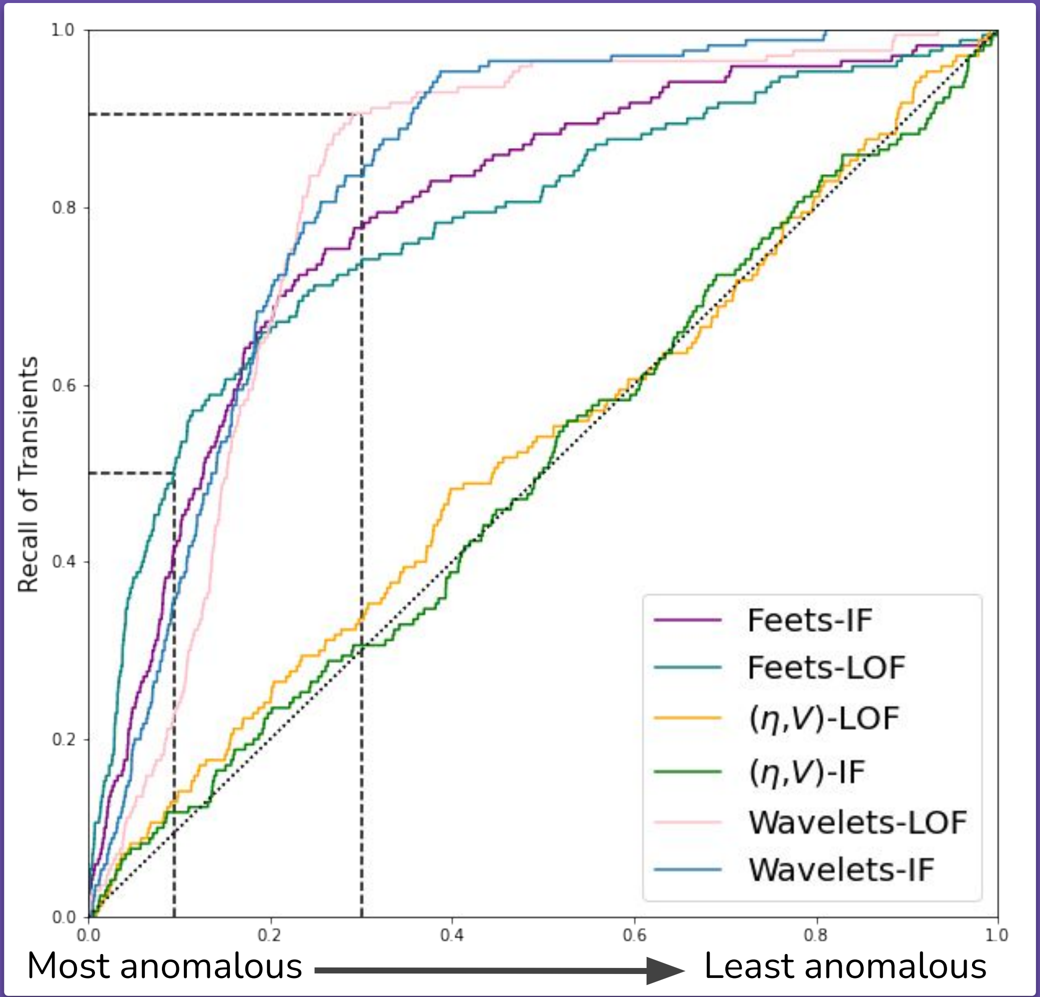
How quickly do we recover our known variables?

Well! We can recover 50% of transients in the top 10% of the data.

Choice of model matters a little.

Choice of feature set matters a lot.

All of this is fast (<10 mins) on a desktop. Think real-time backends...

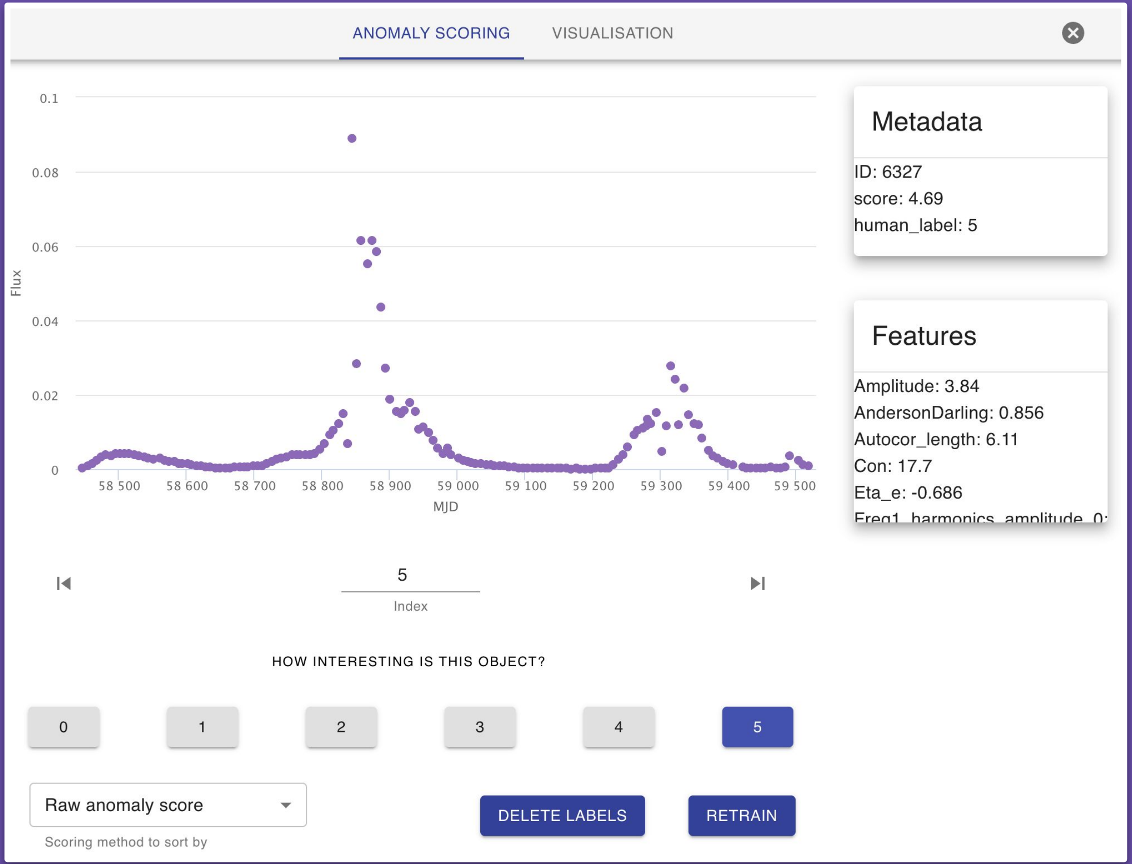


Active Learning

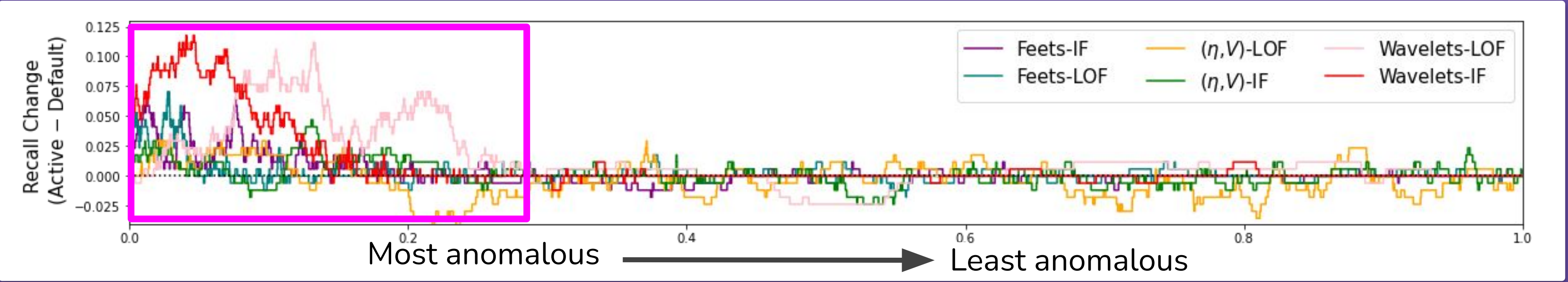
No guarantee that algorithmically anomalous sources are the same as ‘interesting’.

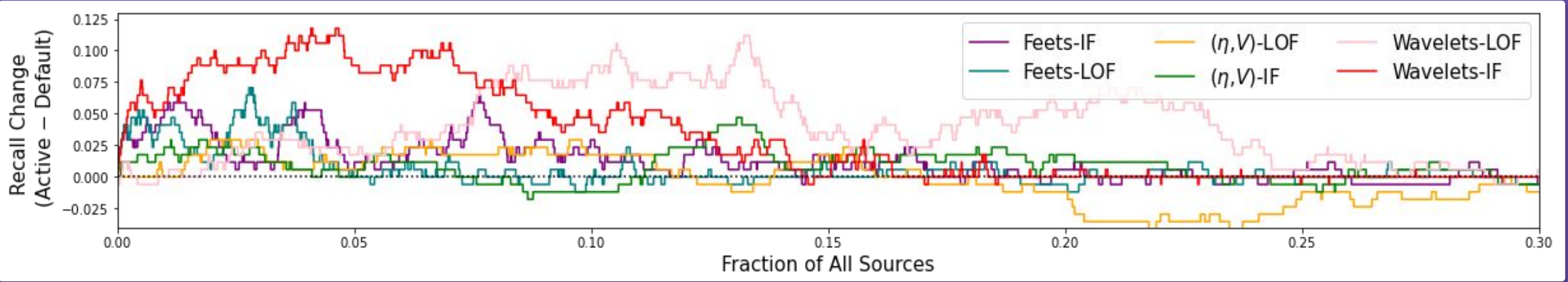
I can go through ~1-2% of the data and label in a subjective way.

How does this improve our recall of transients?



Improvement





Active learning does better for all feature-model pairs - particularly ‘up front’ for astronomers to see, send for follow-up, citizen science inspection etc.

Conclusions

Citizen scientists find scores of interesting light curves for us to follow up.

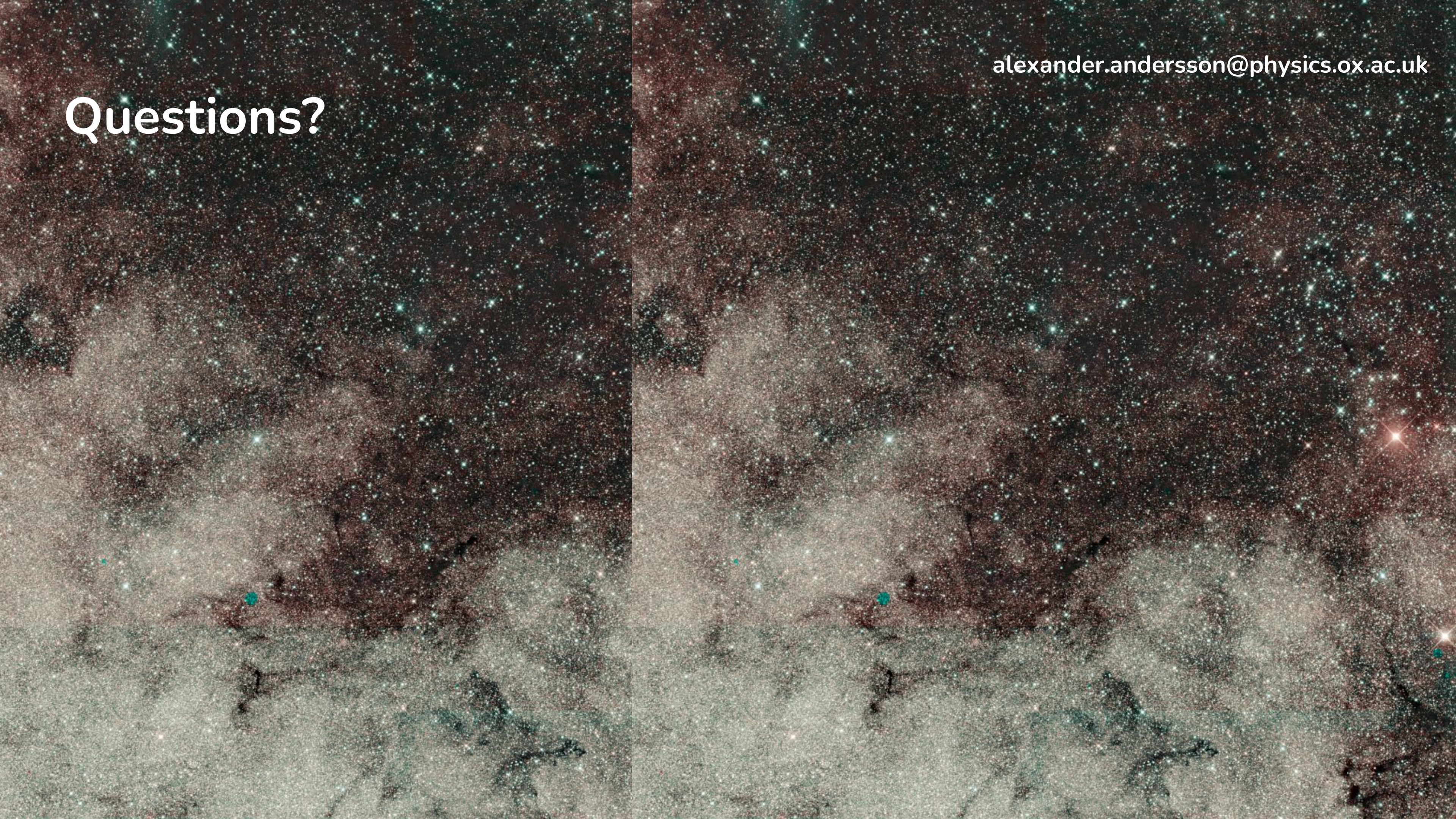
Anomaly detection techniques can find transients (anomalies) quickly, with active learning showing huge improvement.

If you're interested in finding anomalies in large haystacks of data, I'd love to chat!

I'm handing in my PhD thesis in ~5 months - currently in the job market!

alexander.andersson@physics.ox.ac.uk

2204.03481, 2304.14157, ++



Questions?

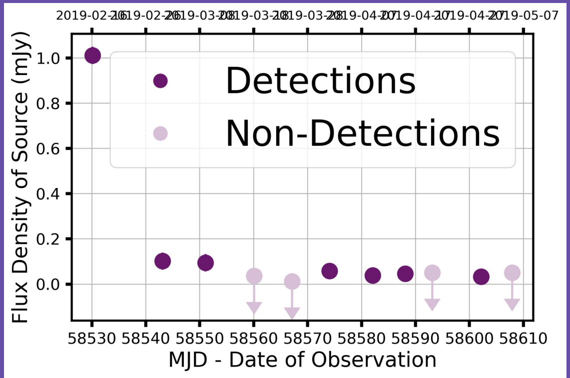
alexander.andersson@physics.ox.ac.uk

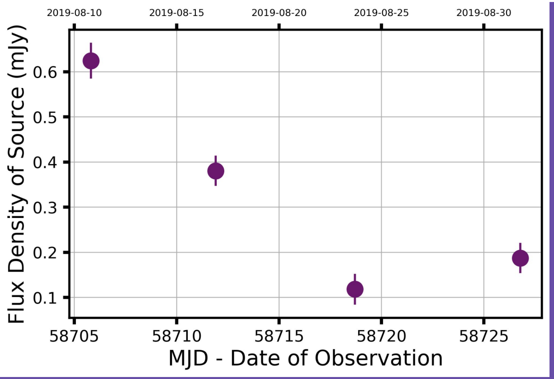
XRBs and LTVs we missed

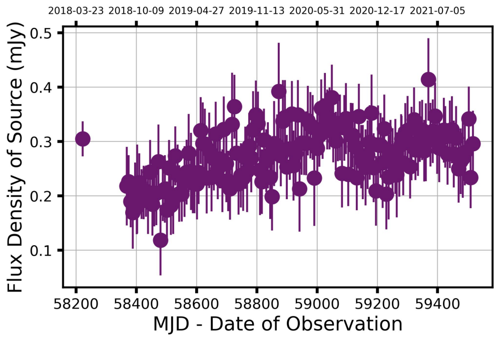
We miss Swift J1858 (see Rhodes et al. 2022) because the figure generated looks bad.

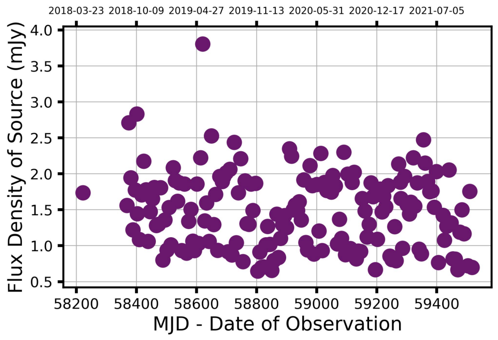
We miss SAX J1808 (Gasealahwe et al. in prep) because volunteers like long light curves.

We miss some LTVs (Driessen et al. 2022) because volunteers prefer ‘clean’ light curves over scattered ones and because of the different number of epochs, TraP parameters, use of a deep first epoch







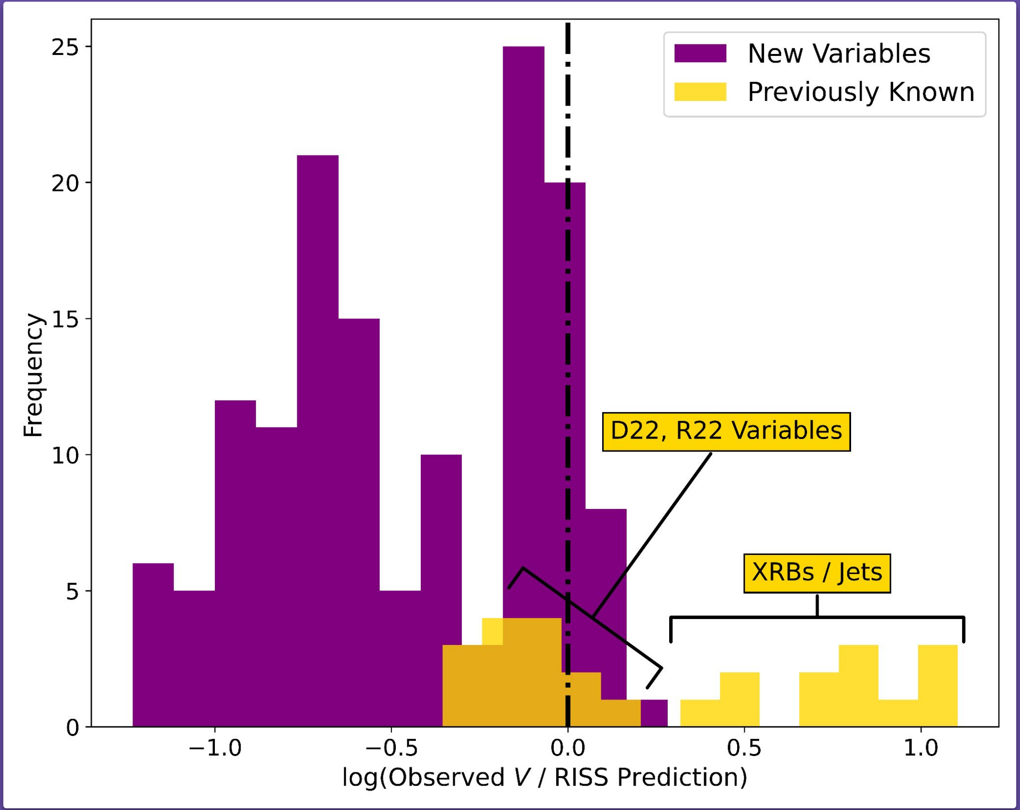


alexander.andersson@physics.ox.ac.uk

What are these radio-variables?

We use Hancock et al (2019)'s code to calculate the predicted Refractive Interstellar Scintillation for a line of sight through the MW towards all of our variables.

So most of our new variables' variability can be explained by scintillation through the MW.

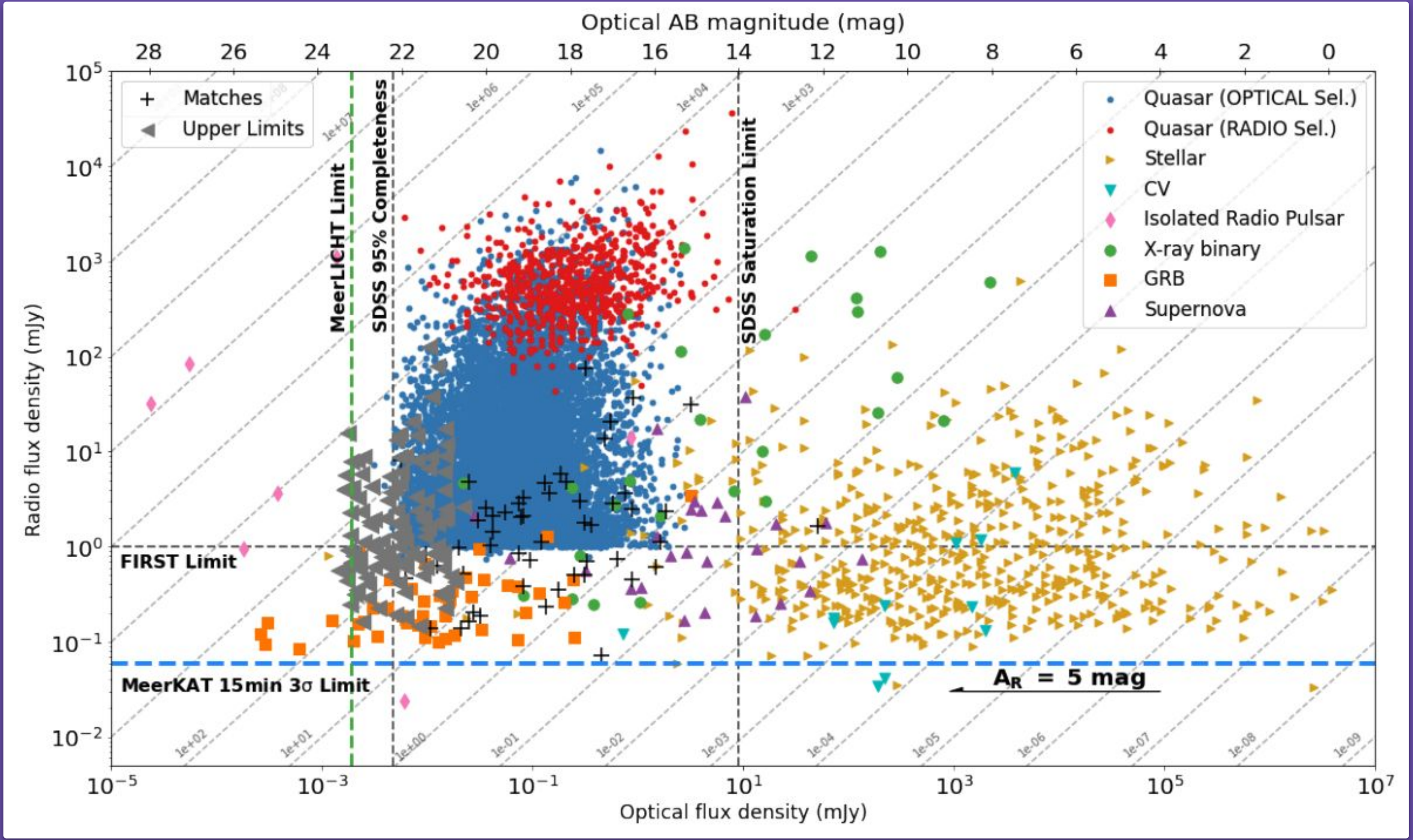


What are these radio-variables?

MeerLICHT data

26 counterparts

Mostly extragalactic



Plot adapted from Stewart et al. 2018, showing radio vs optical flux for known classes of radio variables.

Example 1: OH Maser star

Citizen scientists found a radio-variable OH maser star V1362 Aql / OH 30.1-0.7

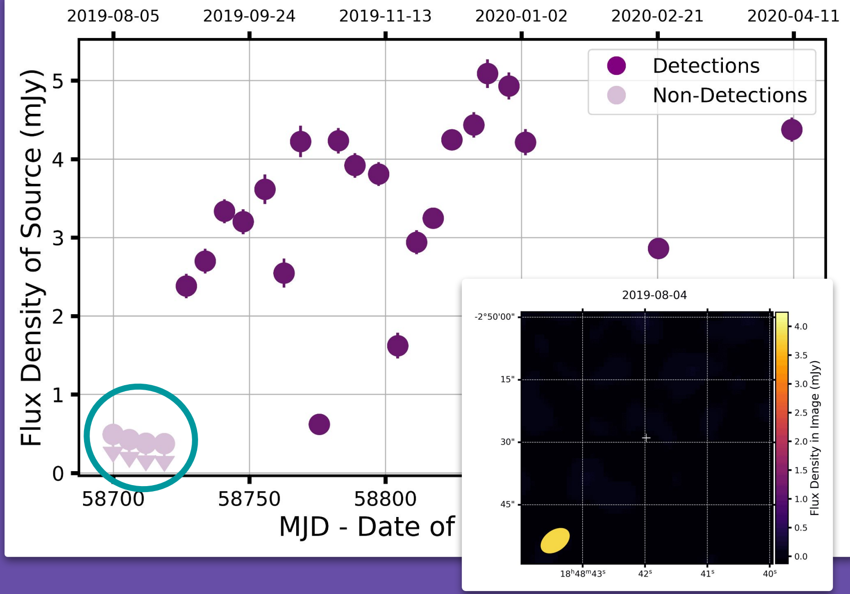
Very bright in IR (W3 = -0.6 mag; Cutri et al. 2013
 66 Jy at 70 μm; Olguin et al. 2015)

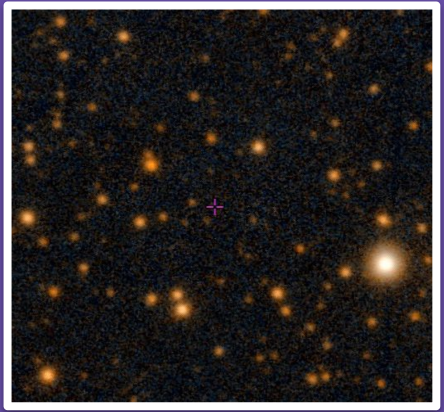
Classified as a Mira variable.

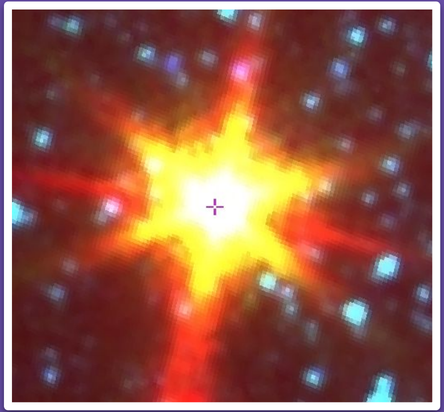
Maser variations expected to be in phase with stellar pulsations.

Not detected in first 4 observations?

Known to be in a binary (Decin et al. 2019)







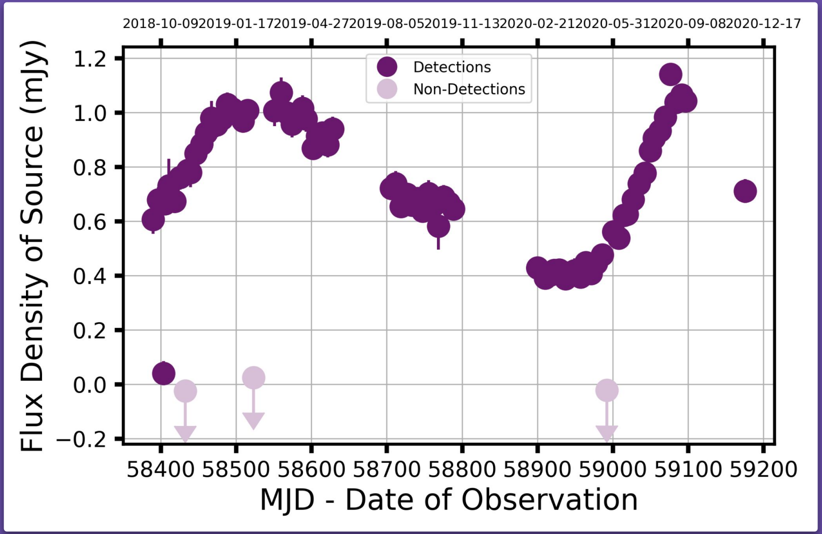
Example 2: Variable AGN?

Spotted on our Talk forum and 7/10 said transient/variable.

VLASS counterpart with flux of ~2mJy.

No counterpart at any other wavelength (Galactic latitude b ~10deg)

We estimate a contribution from scintillation as well as intrinsic variability.




RagIsADoll
@RagIsADoll

January 22nd 2022, 10:06 pm

This #periodic-variable #variable is *soooo* perfect I raked through the meta-data to see if it was a simulation, like some projects toss in to 'test' participants
This is so crazy that this is real, it's just *too* perfect
Bravo, space!

Helpful (0) Reply Link Report

Example 3: Scintillating Pulsar

B1845-01 (J1848-0123) found in the field around globular cluster Glimpse C01.

DM ~ 160 pc cm^-3
Galactic latitude |b| < 5deg

Predicted RISS > observed modulation.

Large free electron content being traversed through the Plane, results in high DM and low-ampltiude, long timescale scintillation.

(see other interesting image plane pulsars from MKT)