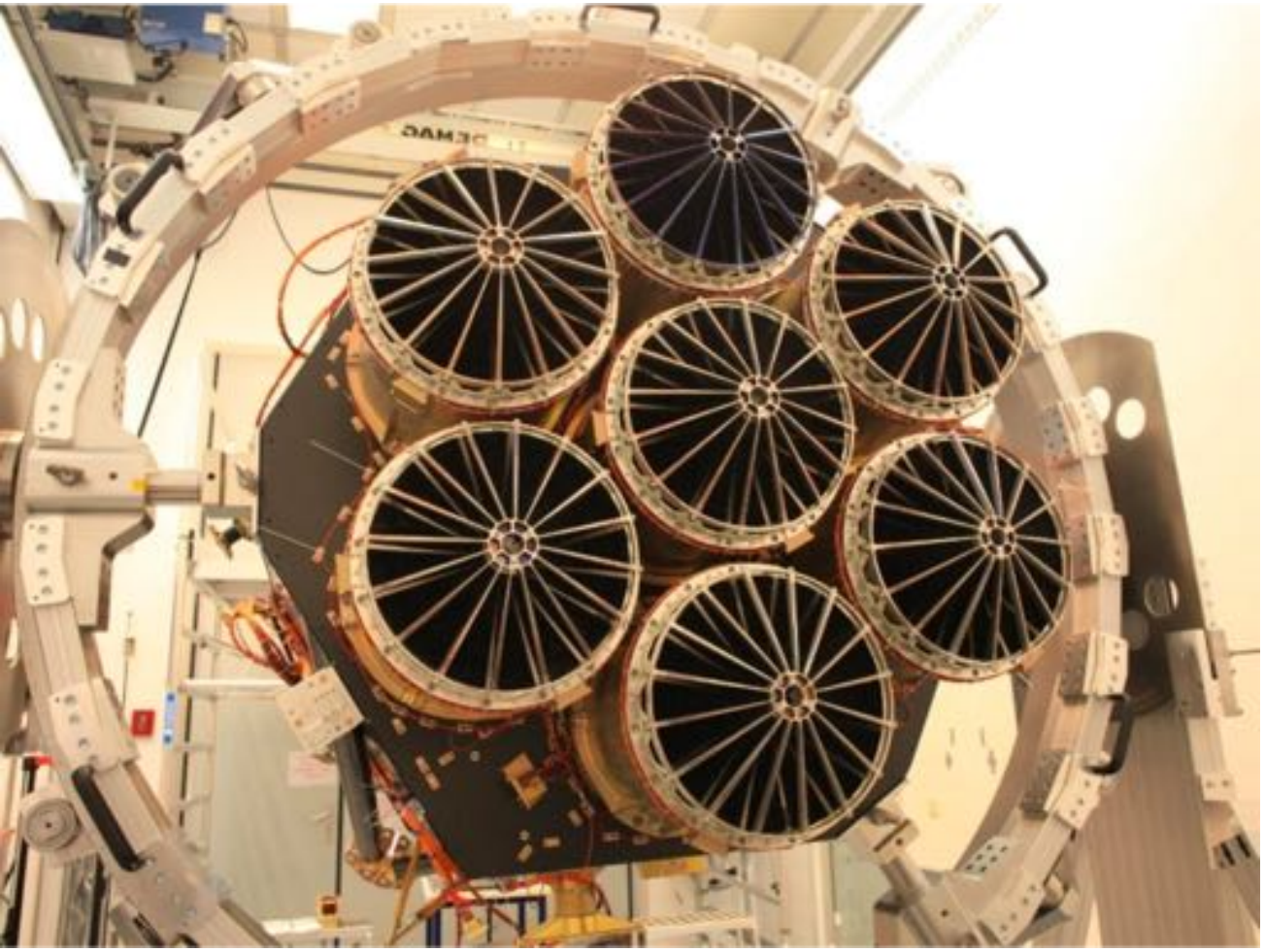
Automated Classification of eROSITA's Transient and Variable Sources

Adam Malyali, MPE

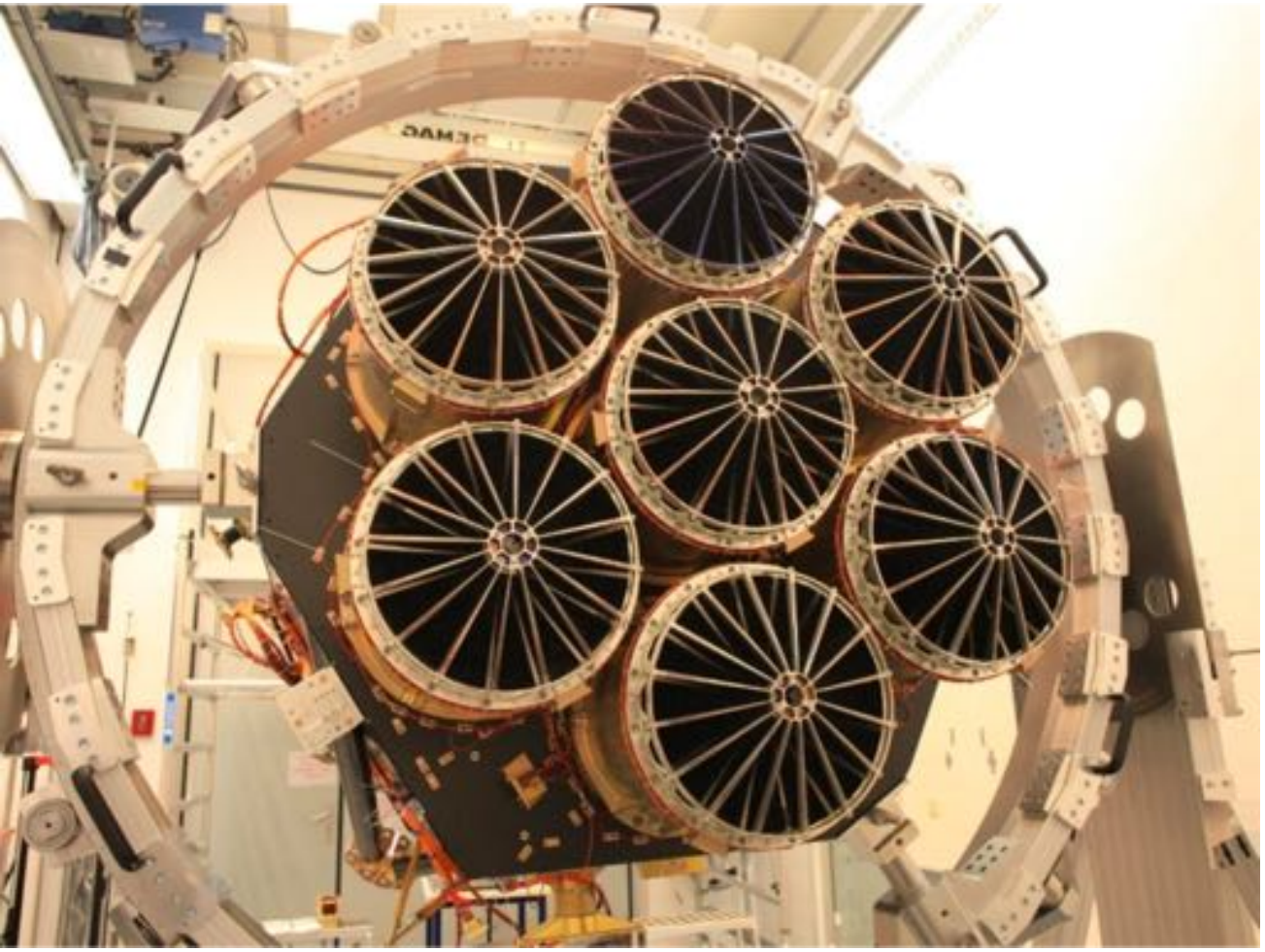
with Arne Rau, Kirpal Nandra

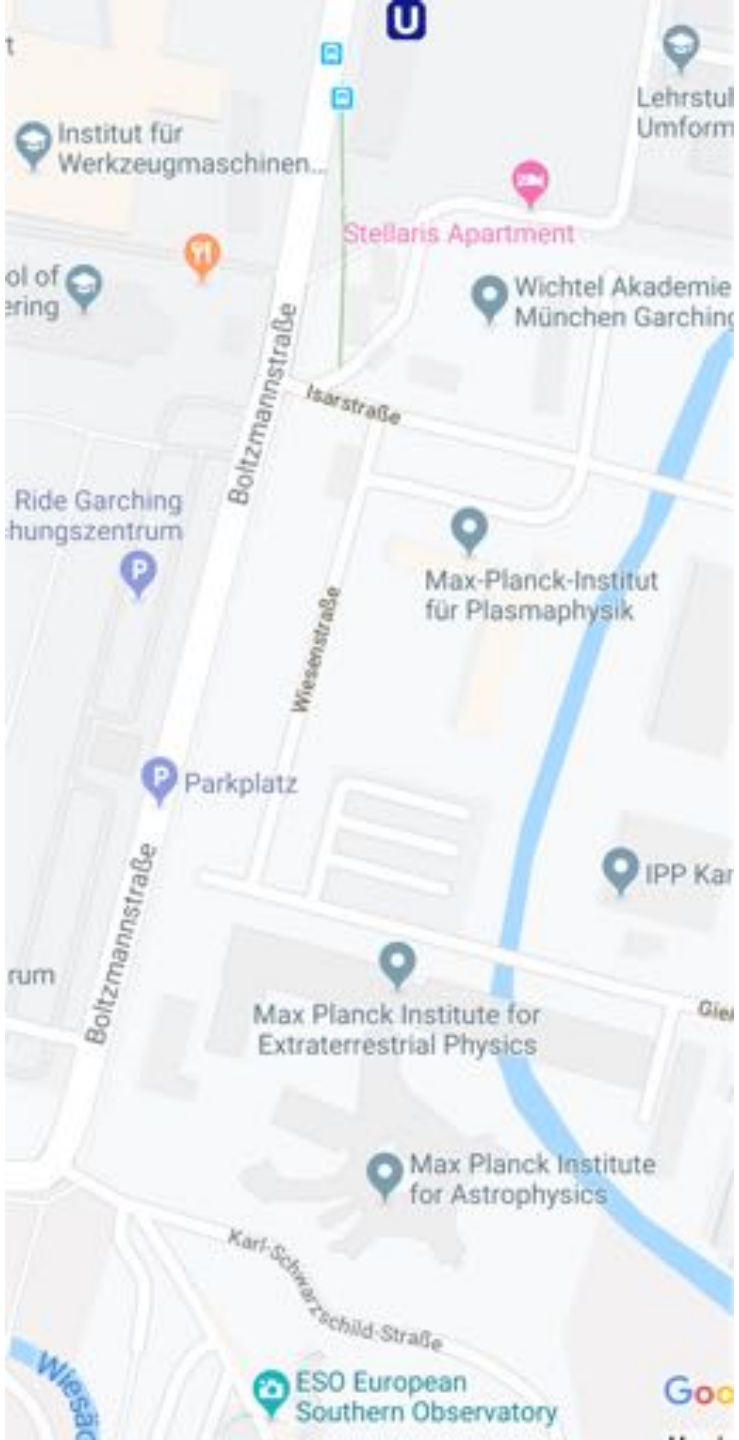
AI in Astronomy, ESO, 23.7.19

eROSITA overview



- New X-ray telescope, operating in 0.2-10keV energy range
- Primary instrument on board Russian-German SRG mission (also ART-XC)
- Data is shared 50/50 between Russian and German consortia




Institut für
Werkzeugmaschinen...
Stellaris Apartment
Wichtel Akademie
München Garching
Isarstraße
Boltzmannstraße
Ride Garching
hungszentrum
Max-Planck-Institut
für Plasmaphysik
Wiesenstraße
Parkplatz
Boltzmannstraße
IPP Kar
Max Planck Institute for
Extraterrestrial Physics
Max Planck Institute
for Astrophysics
Karl-Schwarzschild-Straße
ESO European
Southern Observatory

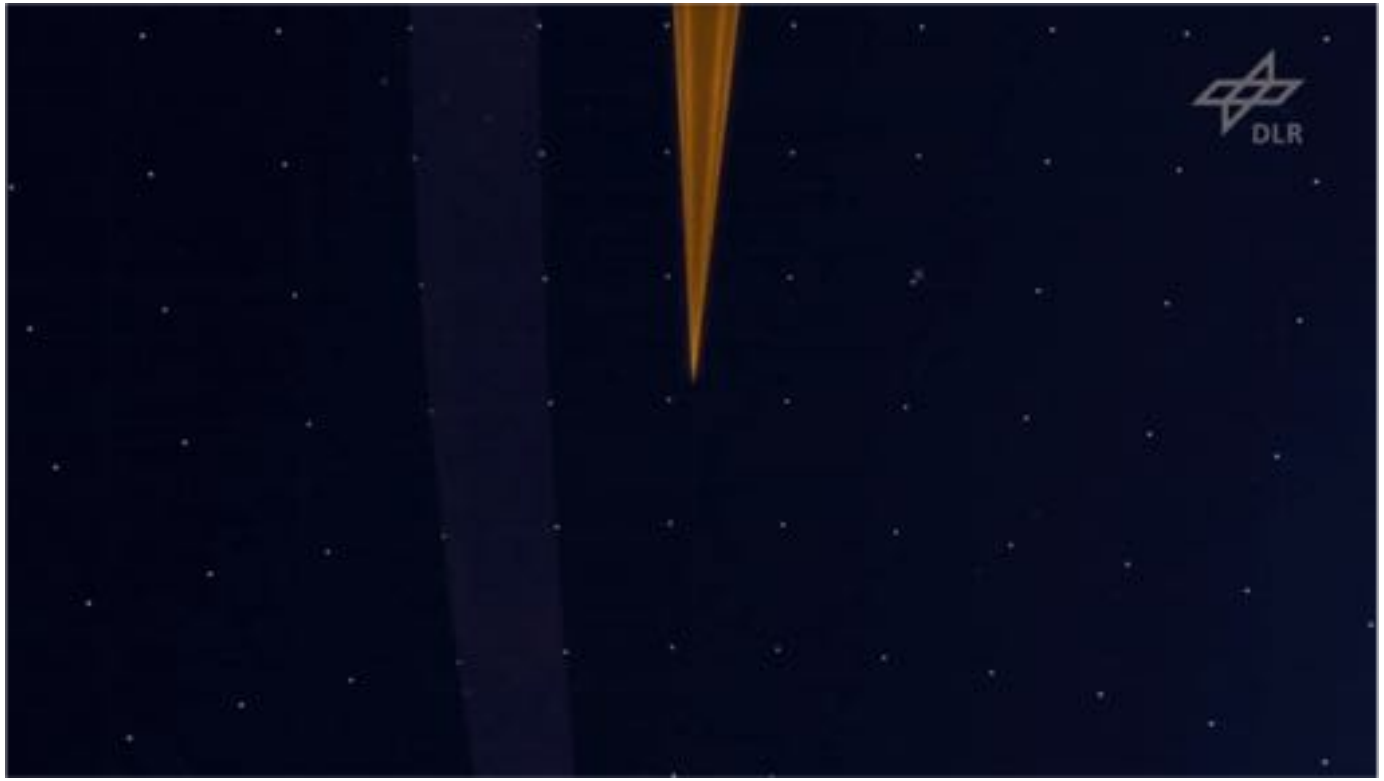
Successful launch on 13/7/19

(last Saturday)

First 4 years: the eROSITA All Sky Survey

Orbit about L2

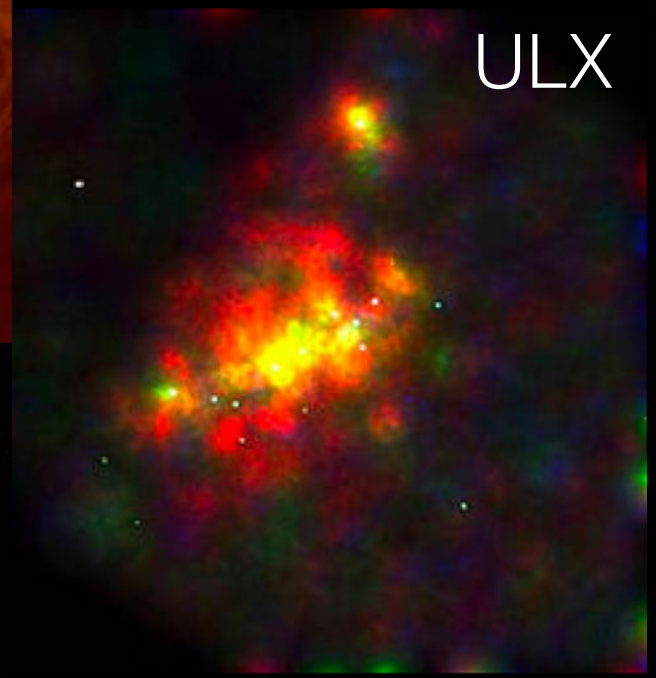
8 successive all sky scans, each lasting 6 months

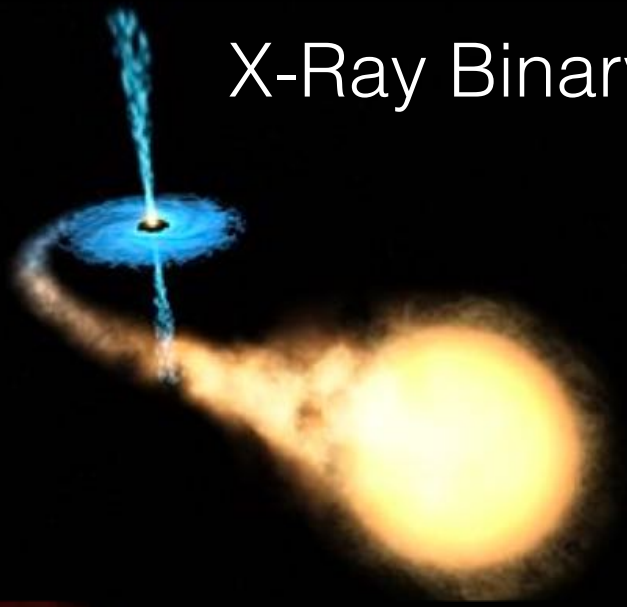


Main scientific design driver:

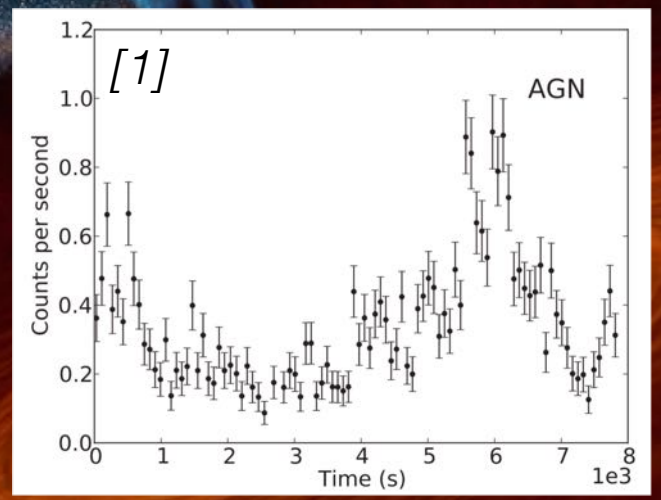
detection of ~100,000 galaxy clusters, and ~3 million AGN

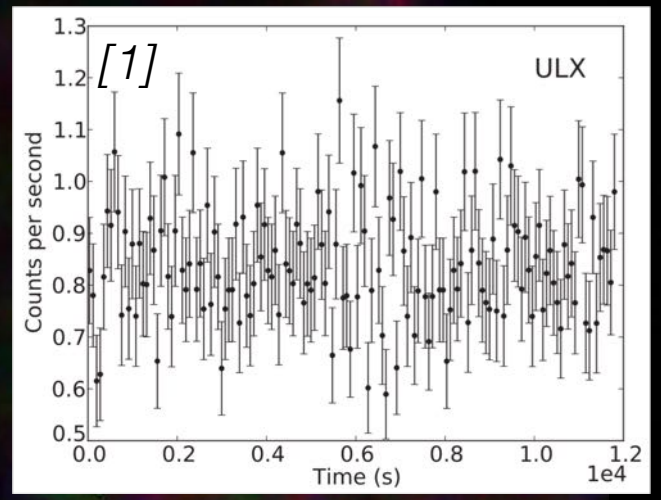
A probe of variability on a range of timescales

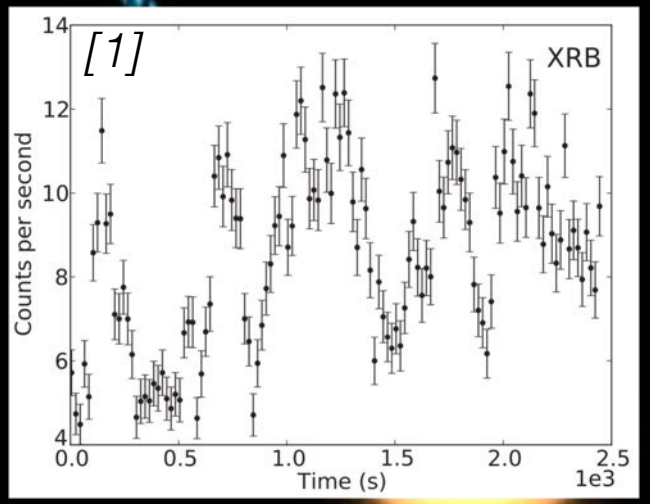


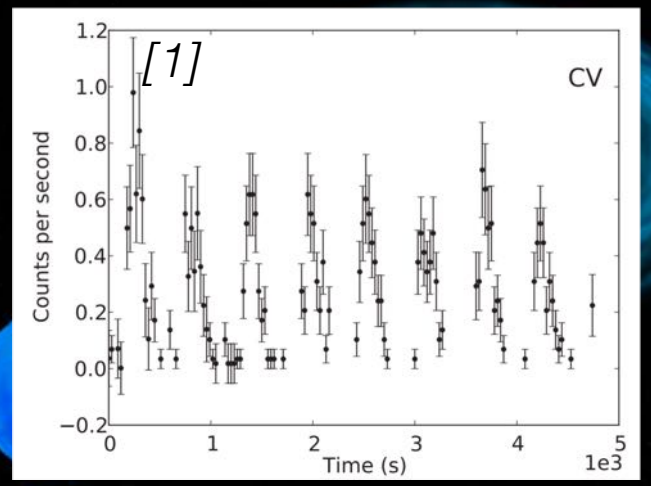


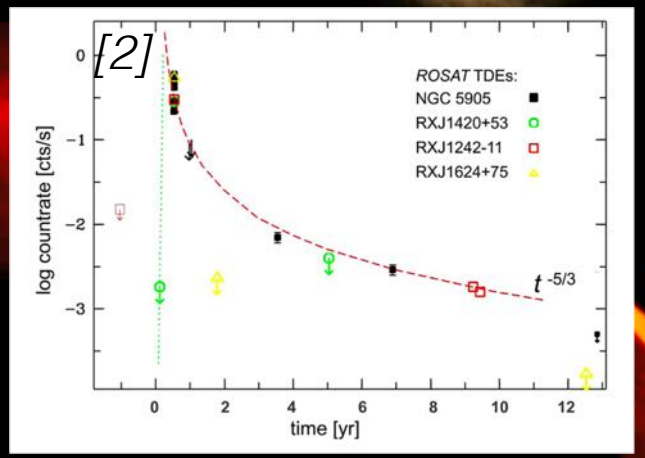
Each source class varies in a unique way











[1] Lo et al. (2014)
[2] Komossa et al. (2015)

What we're developing

The eVSC (*eROSITA Variable Source Classifier*):

- an automated multi-class classification tool, for variable and transient sources
- In early stages after detection
- Using a supervised machine learning approach
- Based on eROSITA data products (eg. 'colours', lightcurves…)

Why?

- Reduce wasted time with human classification
- Identify sources that require prompt multi-wavelength follow-up

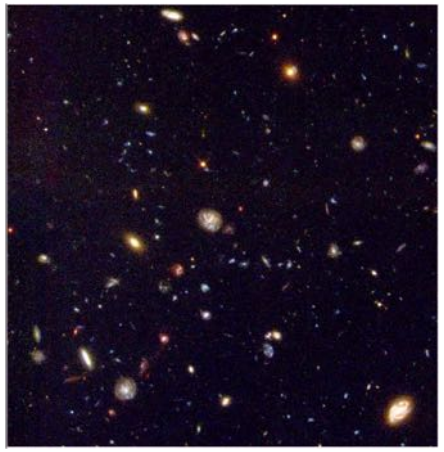
Generating a training dataset, without real data

Supervised ML requires a training dataset, but no eROSITA data yet. Leaves two main options:

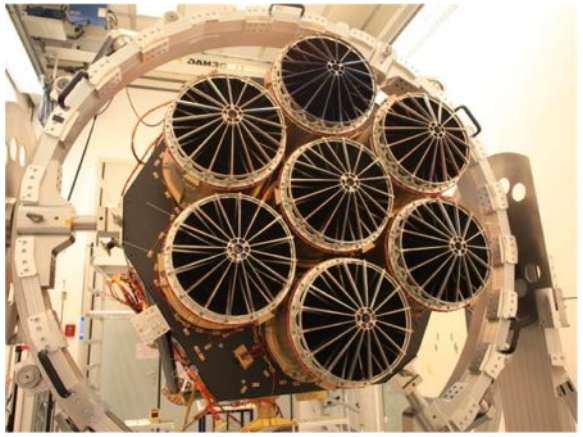
1. Mock light curves:
   - *eg. From simulations, but no robust models of variability for each source class*

2. Train on labelled X-ray datasets, from different surveys:
   - *eg. From 3XMM data products (spectra, lightcurves). But vastly different time sampling, different spectral response info…*

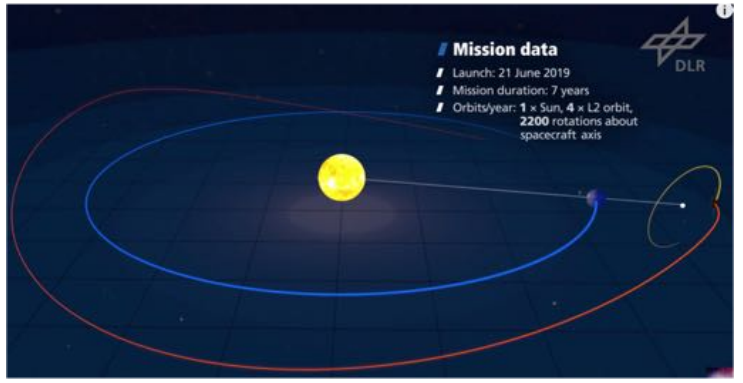
Our approach: simulate eROSITA observations of labelled 3XMM sources



Sky model



eROSITA calibration files



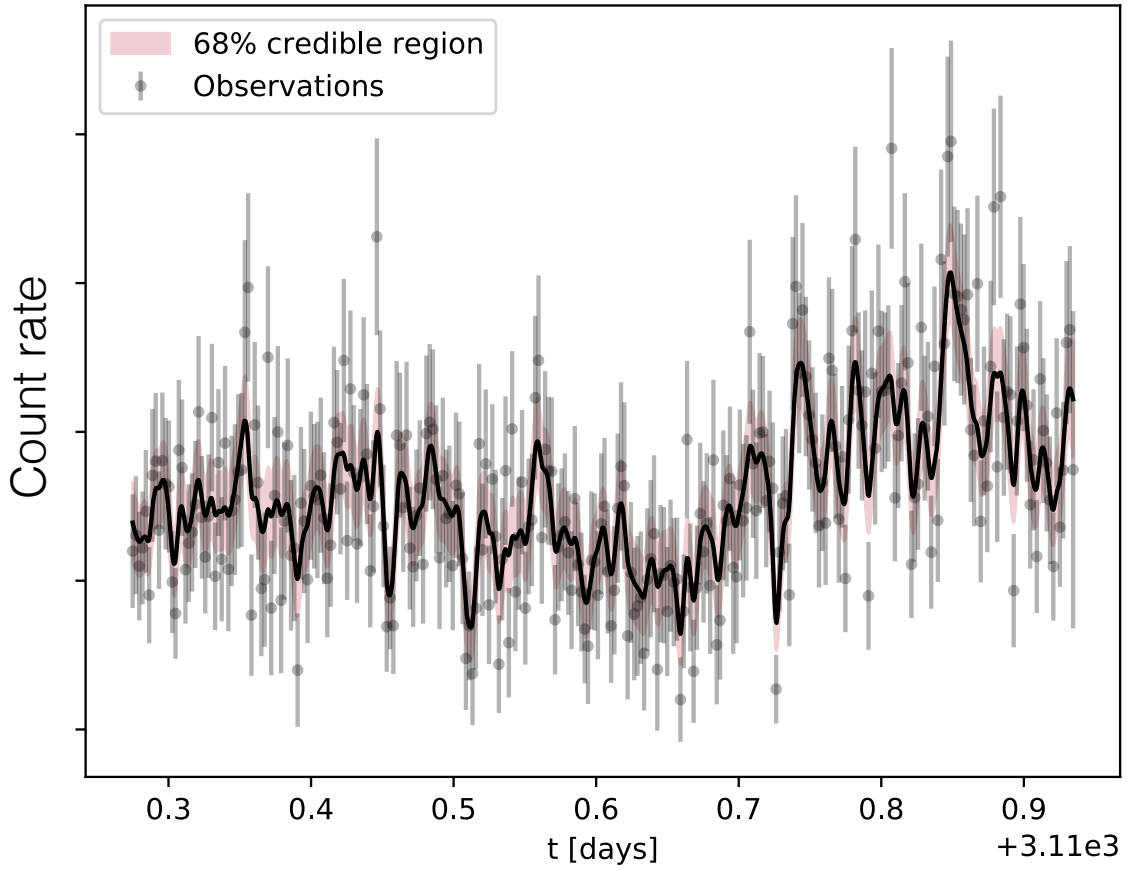
eROSITA scan strategy

SIXTE simulator: obtain eROSITA event files → eROSITA analysis software: obtain data products → Feature computation, machine learning…

Generating a sky model

For each source in sky model, we specify:

- A position
- Flux
- Spectrum
  - Bayesian spectral analysis
- Light curve
  - *Via Gaussian process regression of 3XMM lightcurves*



eROSITA All Sky Survey Time Sampling

- Per all sky scan (6 months), a point source will have a minimum of:
  - *6 ~ 40s visits every 4 hours, for one day (roughly 90% of sky).*

- As the source position gets nearer to the North/South Ecliptic Pole, each source is still observed every 4 hours, but number of consecutive days of being visited increases
  - *eg. a source near the SEP will be visited every 4 hours over a 30 day period.*

Very limited labelled dataset for all sky survey

Unfortunately, the combination of:

- XMM lightcurve coverage *(being too short)*
- Simulation of realistic eROSITA observations

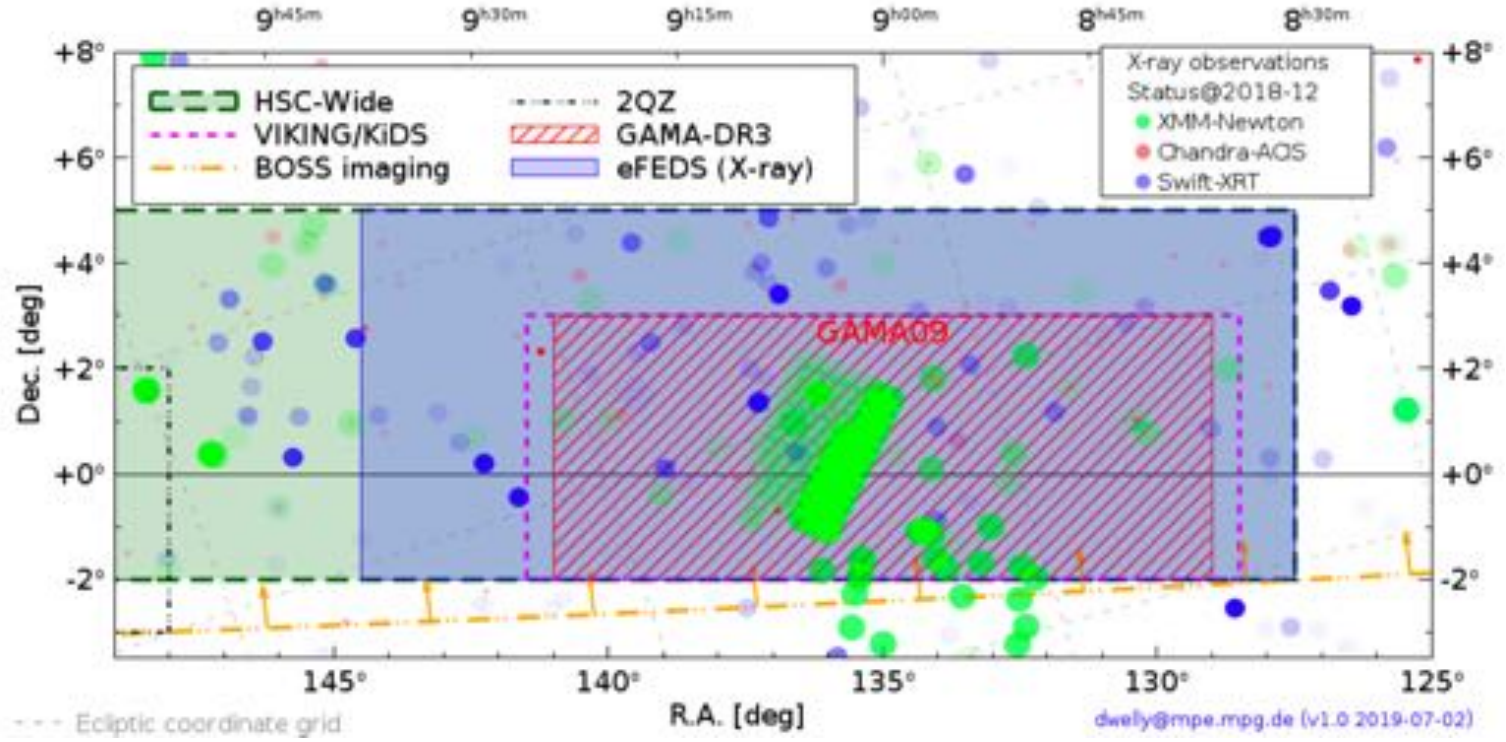
reduces size of our training set to only ~120 sources, with 7 source classes *(for the all sky survey)*

We are 'reduced' to evolve our classifiers during survey using active learning:

- Initially based on searching for counterparts to sources in order of decreasing uncertainty in classification
- Later with more complex cost functions

What else can we do to prepare for ML classification of eROSITA sources?

Variable source classification in a pre-all sky survey field (‘eFEDS’)



Prior to All-Sky Survey, eROSITA will run a smaller survey ‘eFEDS’

- Different scan strategy- smaller cut to training set size

Variable source classification in eFEDS

~ 600 eROSITA-fied XMM sources

(as observed during eFEDS, not the all sky survey)

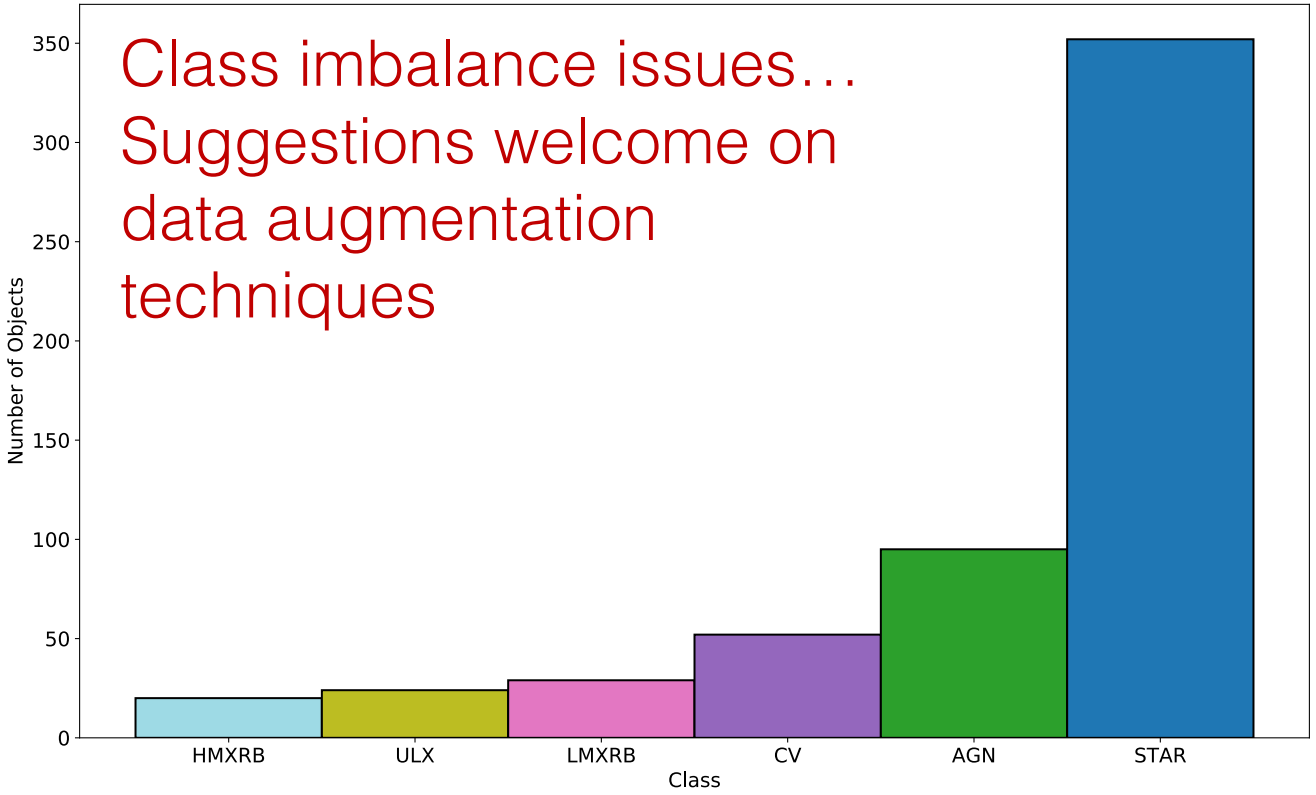
Train a random forest classifier on features computed from these

→ To make classifications of these

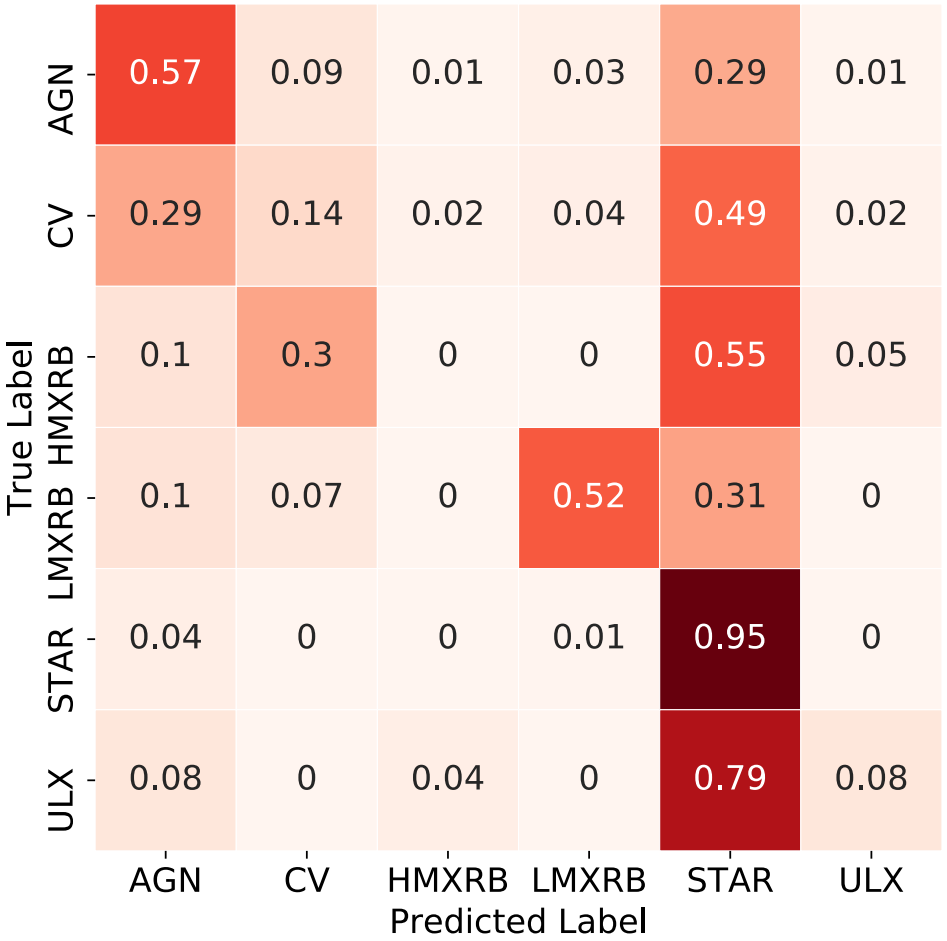
N eFEDS variables

Variable source classification in eFEDS

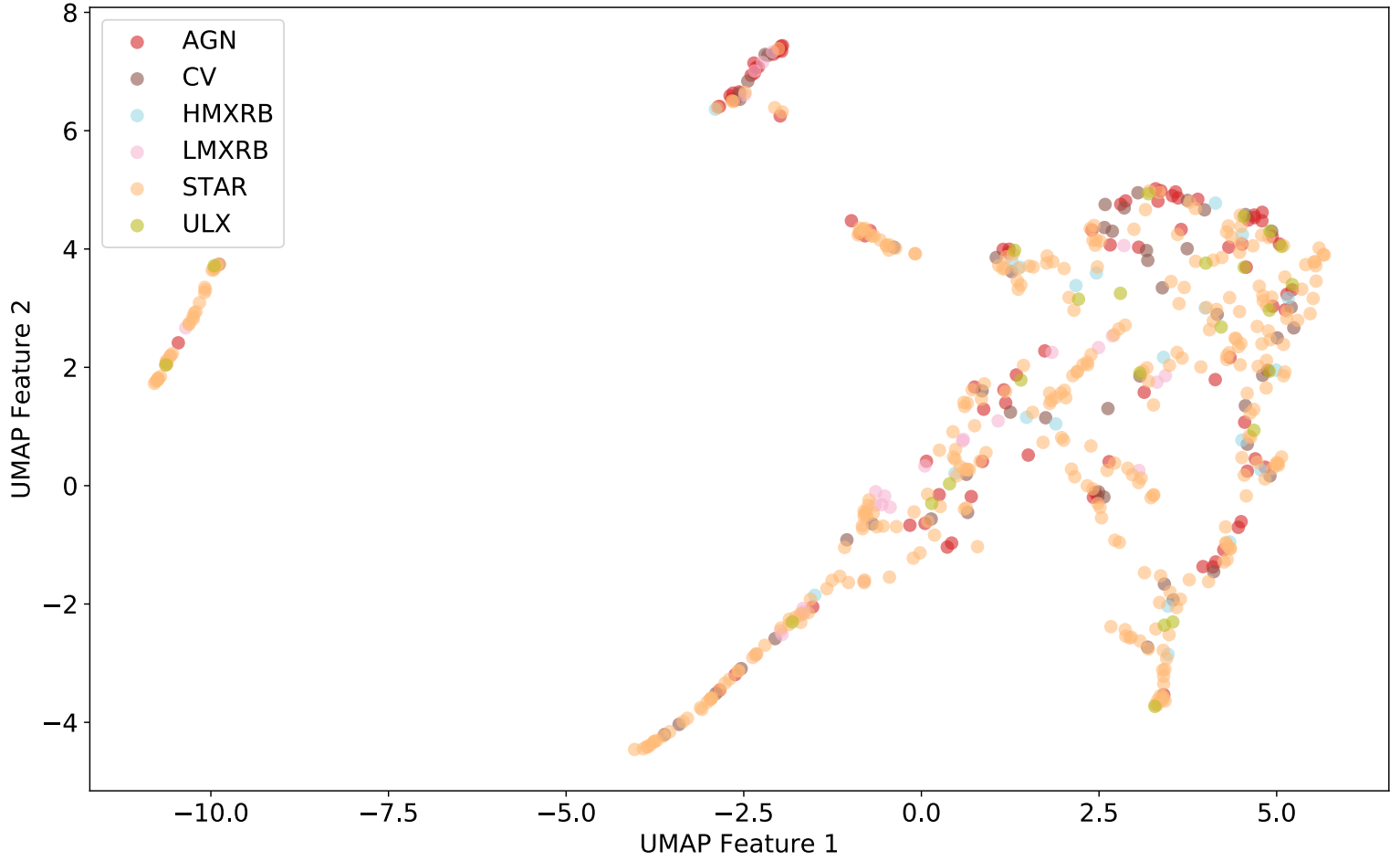
~ 600 eROSITA-fied XMM sou



Basic classification performance

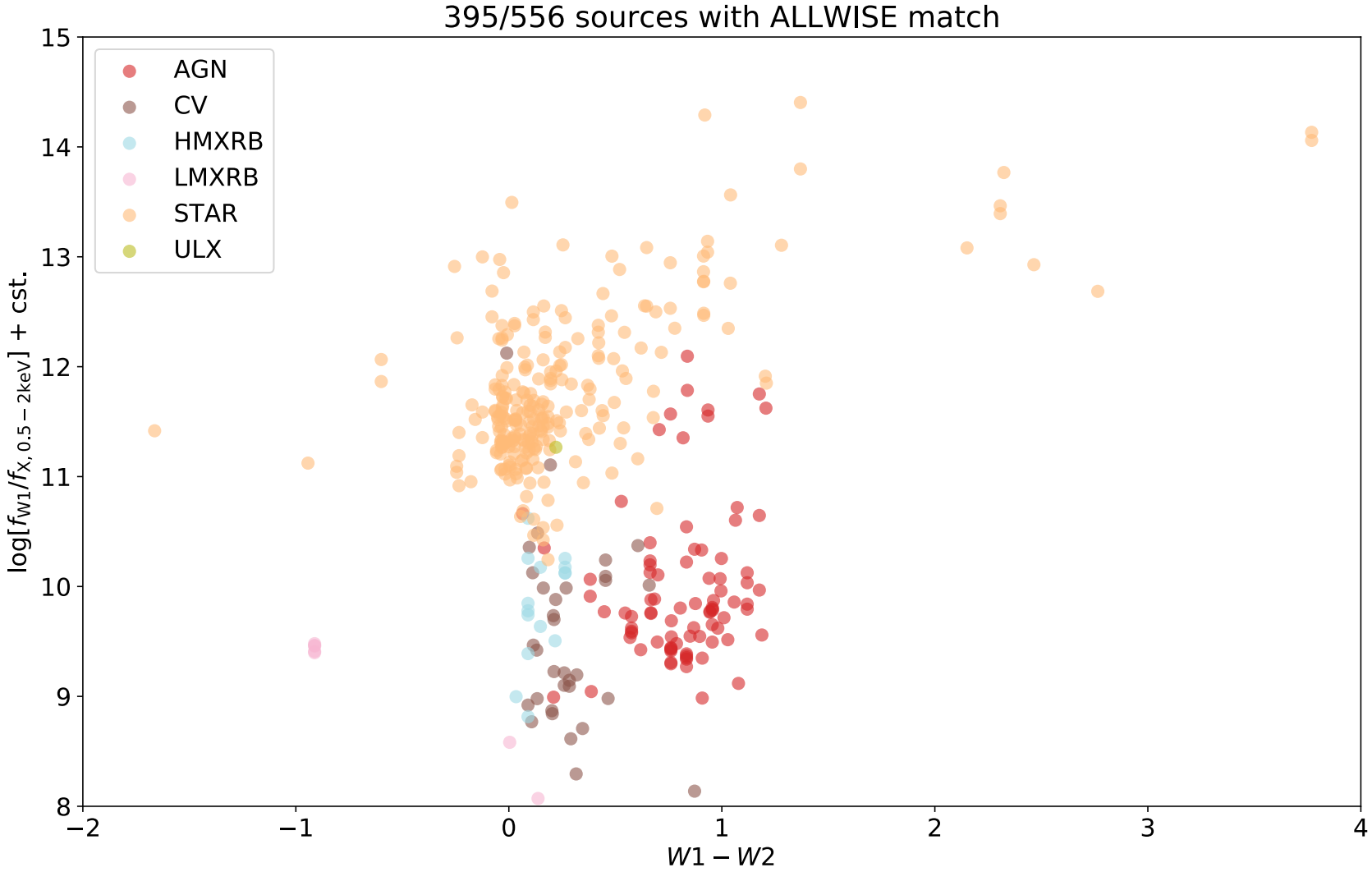


Normalised confusion matrix

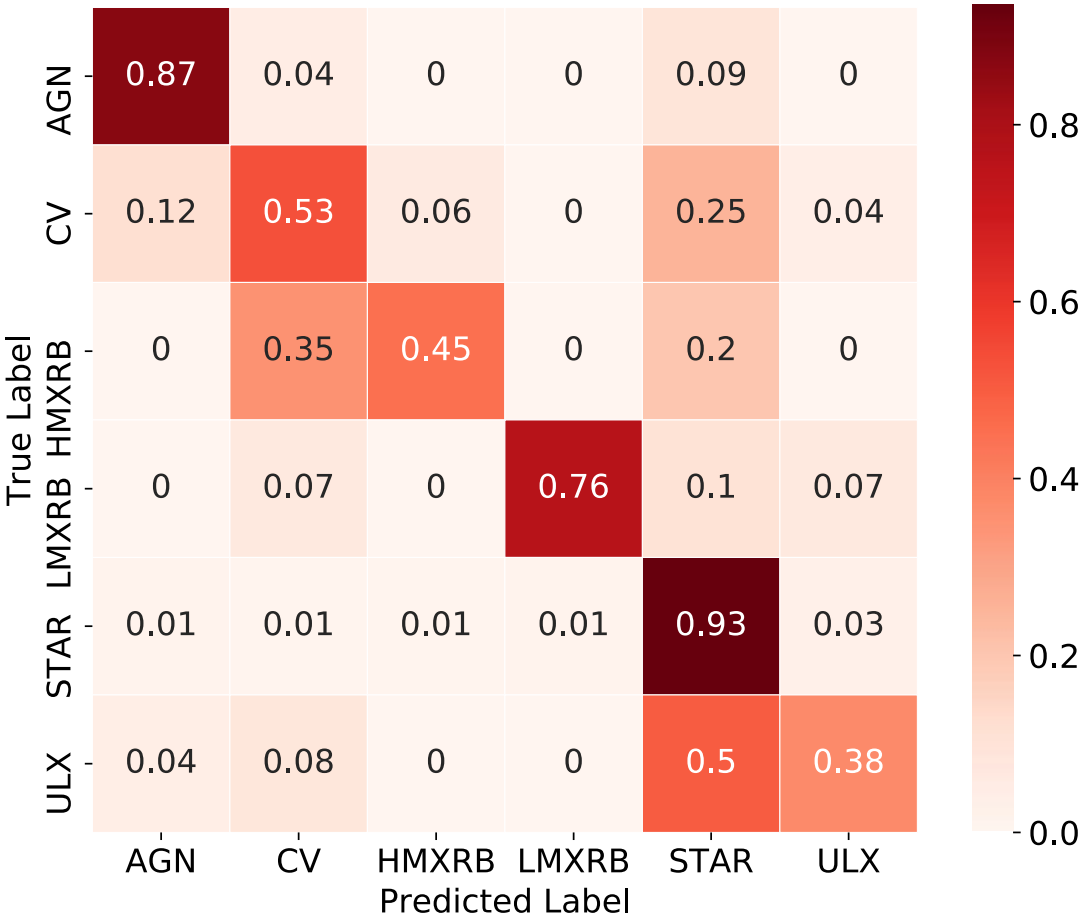


UMAP visualisation of structure

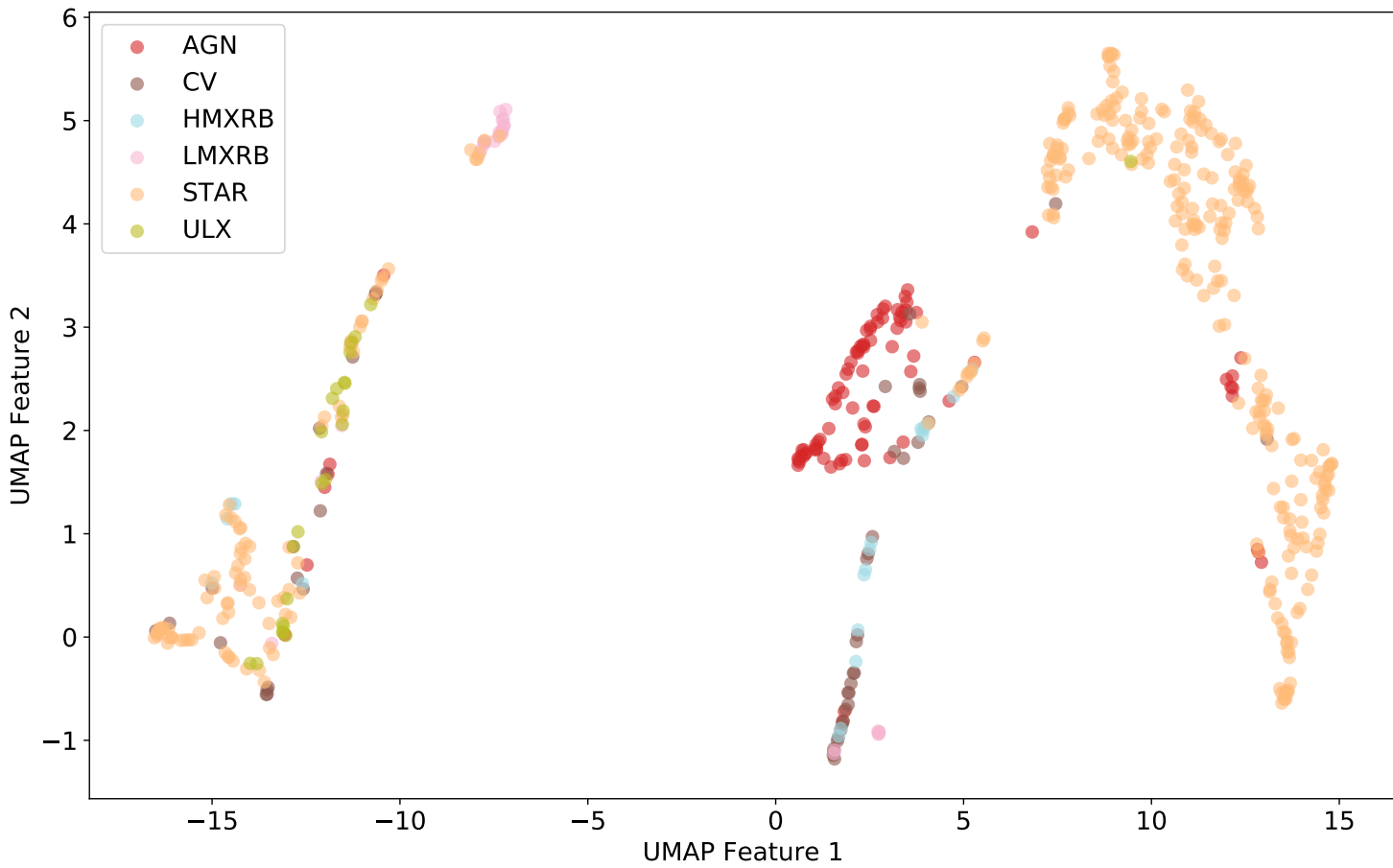
Adding in multi-wavelength information



Classification performance with multi-wavelength features



Normalised confusion matrix



UMAP visualisation of structure

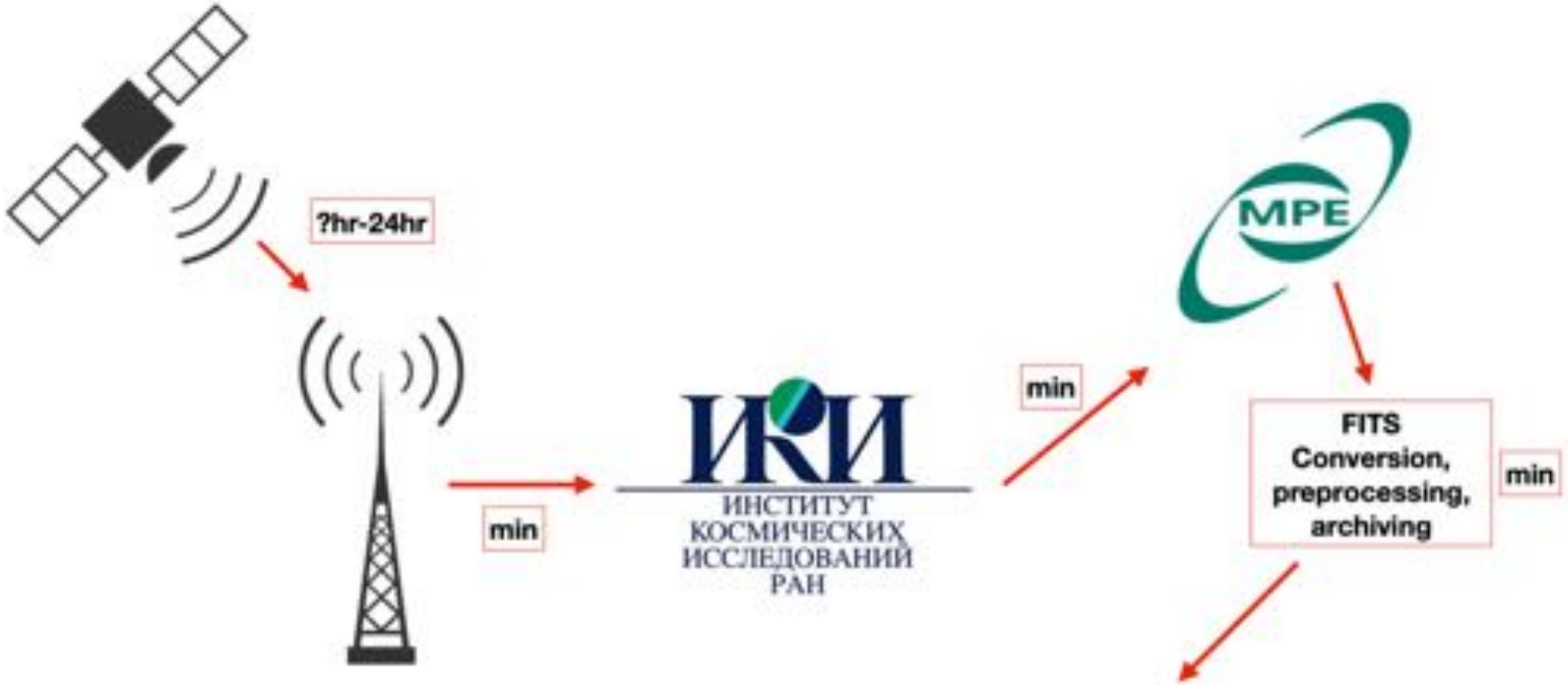
Outlook

- For eROSITA's all sky survey, a lack of training data pushes us strongly towards evolving classifiers through active learning based approach.

- Multi-wavelength features will be vital for eROSITA transient and variable source classification

Special thanks to: Julian Wolfe, Dalya Baron, Emille Ishida, Mara Salvato, Andrea Merloni

eROSITA data download and products



Credit: Jörn Wilms, Ingo Kreykenbohm, Philipp Weber, Sebastian Falkner, the NRTA and TDA teams

Example differences between lightcurves

A relatively long 3XMM lightcurve, example approx. eROSITA sampling times shown in blue.