# Selected Research Projects

## Classifying radio phenomena in real time with streaming machine learning (API)

**Intended promotor(s):** *prof. dr. Ralph Wijers (API); IvI promotor*
**Intended co-promotor(s):** *dr. Antonia Rowlinson (API); lab manager*

### Background:

One of the cutting edge areas of astronomy is recognising rare features in data streams in near-real time. This problem has arisen from massive imaging surveys of the sky at many wavelengths (optical, radio, X rays) where the purpose is no longer just to find objects that are there all or most of the time, or to find so-called ‘transient’ objects that appear fleetingly in hindsight. Instead, it is to find transients in near-real time and to signal their presence automatically to other observatories.

These transients are typically multi-messenger events, meaning their emission is seen throughout the electromagnetic spectrum and using other ‘messengers’ including neutrinos and gravitational waves. By combining the multi-messenger observations, we can probe the most extreme physical environments in the Universe. However, we are not monitoring the whole sky all the time with sensitive observatories, so we need to alert them to interesting objects in order to trigger follow-up observations. As these events are often fleeting at different wavelengths, we need to trigger the follow-up in as close to real-time as possible.

To identify transient sources in radio images, we developed the LOFAR Transients Pipeline (TraP; Swinbank et al. 2015). This pipeline receives a set of image cubes from a radio survey and outputs a database of monitored sources. Our team have demonstrated that machine learning strategies can be successfully applied to archival observations processed using TraP with a combination of simulated transients and real datasets (Rowlinson et al. 2019).

How does one apply techniques for rare feature extraction in such a situation? Does one need to fix the algorithm up front, and if so, how does one find optimal algorithms? Or are there ways of applying unsupervised machine learning in even such environments? Is it possible to design algorithms that can handle such situations with sufficiently low latency? A typical example would be to have an image in optical with 100 Mpixel every minute, with several times $10^5$ objects per image. Or to have 10 images per second in radio, 10 Mpix each, with about $10^4$ objects per image.

### Aims:

Automatically classify astrophysical transients, space weather events and terrestrial radio interference in real-time data streams from radio facilities, such as AARTFAAC.

## Approach:

**Feasibility:** Machine learning demonstrated feasible and valuable for filtering in Rowlinson et al. (2019). Manual filtering and development of automated filtering of known features in AARTFAAC data by Mark Kuiack (final year PhD student).

**Challenges:** The nature of our searches means that we cannot always predict the transient behaviour we are searching for, nor all more rare specific instrumental faults in the data stream. The real-time data streams are subject to different types of behaviour which we do not want to confuse with astrophysical transient events. Including radio frequency interference, calibration artefacts, varying noise features, space weather, meteors, airplanes and satellites.

**Available data:** We have a large quantity of AARTFAAC data available on disk for training algorithms (800 hours of 1 second snapshots at 2 (or more) observing frequencies, ~6 million images with ~1000 monitored sources per image) and these data can be labelled with known features as they are currently being characterised by Mark Kuiack (final year PhD student). These images can be used directly or their data products following processing using the TraP.

**Machine learning approach:** Hybrid approach, whereby a strategy is developed using machine learning to label training data for a machine learning algorithm that is able to adapt in real time to a streaming dataset.

### Rough work plan:

WP1. Trial classification of transient phenomena in archival AARTFAAC data using supervised machine learning algorithms, building on existing work in Rowlinson et al. (2019) where appropriate. Use an unsupervised classification algorithm to develop a classification scheme that may identify previously unknown phenomena within a datastream. Compare and contrast to supervised classification.

- Deliverable 1: publication of classification results including the comparison between supervised and unsupervised strategies for this purpose.
- Deliverable 2: publicly published code with documentation

WP2. Determine if an image stream or a catalogue datastream is most effective for AARTFAAC type systems.

- Deliverable 1: publication outlining the image stream approach in contrast to the catalogue strategy from WP1.
- Deliverable 2: publicly published code with documentation

WP3. Utilise the classification scheme to produce a large labelled training dataset to train an initial dataset. This dataset is used to train a machine learning algorithm that is able to adapt in real time to account for changing behaviours in the datastream.

- Deliverable 1: publication outlining the hybrid approach
- Deliverable 2: publicly published code with documentation

WP4. Apply adaptive machine learning strategy in real-time AARTFAAC pipeline and produce reliable transient alerts.

- Deliverable 1: code implemented in AARTFAAC streaming pipeline with documentation

- Deliverable 2: production of near real-time published alerts

## Impact:

**Innovative aspects:** real time radio transient detection, streaming data (1 second cadence), hybrid approach - developing a training set with an unsupervised strategy which is then used to train a supervised algorithm, real-time adaptive machine learning model.

**Added value for radio transients:** we will be able to classify features observed in the radio images in real time, leading to the rapid identification of astrophysical transients and space weather events. This real-time identification enables rapid follow-up of these events at other wavelengths.

**Added value for AI:** innovative combination of using unsupervised strategies with real-time adaptive supervised strategies. Speed of processing in real-time adaptive strategies to manage the datastream.

## Possibly interested external parties:

Space weather.

## References:

Rowlinson et al., 2019, Astronomy & Computing (accepted), arXiv:1808.07781, http://adsabs.harvard.edu/abs/2018arXiv180807781R

Swinbank et al., 2015, Astronomy & Computing, 11, 25