Unsupervised clustering of the Extra-galactic Universe

Research Field
Radio Astronomy

Project Suitability
Summer
3rd Year
Honours

Project Supervisor
Dr Natasha Hurley-Walker
nhw@icrar.org

Co-Supervisors
Dr Tim Galvin

Modern all-sky radio surveys represent a big data challenge, one where traditional approaches informed by physical expectations may no longer be the most appropriate. The GaLactic and Extragalactic All-sky MWA eXtended (GLEAM-X) project has observed the sky south of Dec +30 across a frequency range of 72 – 230 MHz. When finished it is expected to detect in excess of a million objects across twenty intermediary frequencies. Collectively, these objects and their properties represent a massive multi-dimensional dataset that is difficult to interact with and efficiently mine for meaningful scientific outcomes. This is especially true for rare or previously unseen objects that are buried beneath the more typical sources.

This project will explore how unsupervised clustering methods may best be applied and exploited to create a structured framework to allow objects detected by GLEAM-X to be categorised and explored efficiently. Methods regularly applied in the machine learning community (including auto-encoders, generative adversarial networks, t-SNE etc.) will attempt to organize the contents of the GLEAM-X outputs into a scheme that neatly separates objects into fundamental or regressed classes. These methods are especially powerful as they do not necessarily have to incorporate expectations specified by assumed physical models, thereby avoiding potential bias or constraints inadvertently introduced. When finished, the applied approach would be capable of compressing all object properties into a simple two- or three-dimensional embedding to be made available for exploration. A major outcome of this data-driven approach to modelling is the ability to identify outliers, which in this scenario could be a set of exceptionally rare or previously unseen objects.

Aims of the project:

i. Develop appropriate clustering methods using existing public catalogues (e.g. GLEAM);
ii. Expand the method to the new GLEAM-X data;
iii. Search for and characterize reference objects such as radio galaxies;
iv. Explore outlying populations to discover new populations of sources.

This project is well suited to a student with a good programming background, an interest in developing new computational techniques, and good organisational skills.

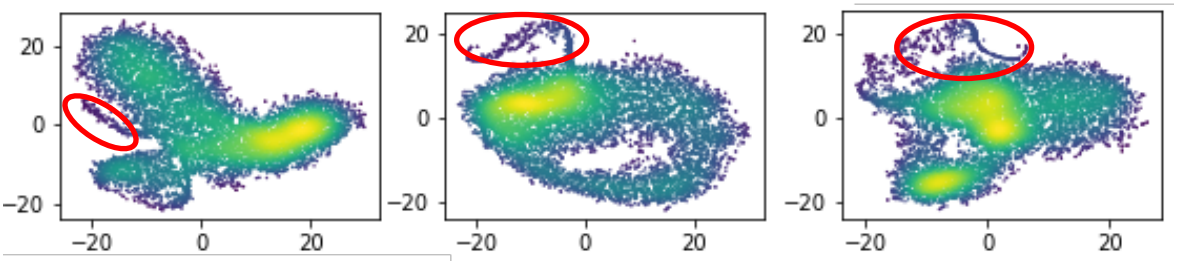

Figure 1 A naïve example of 30,000 GLEAM sources (21 input features) being projected to a lower dimensional embedding (3 dimensions) using t-SNE. Potential interesting populations are highlighted with red circles.