NASSP 2019 Master Project Proposal

Data Mining and Virtual Observatory for Classification and Analysis of the Faint Radio Sky

- Primary Supervisor: Dr Khadija El Bouchefry
- Supervisor's Institution: South African Radio Astronomy Observatory/HartRAO
- Supervisor's contact details: khadija@hartrao.ac.za/kelbouchefry@ska.ac.za ; 0765147982
- Co-Supervisor: TBD
- Co-Supervisor's Institution: TBD

☐ Overviews and aims of the research project:

In the last decade a new generation of telescopes and sensors has allowed the production of a very large amount of data and astronomy has become a data-rich science; this transition is often labeled as: "Data Tsunami". The data are not just increasing in size but also in complexity and dimensionality. Astro-informatics is a new field of science which has emerged from this technology-driven progress. Virtual Observatory, Machine Learning, Data Mining and Grid Computing are just a few examples of the new tools available to scientists. The new concept of data infrastructure named Virtual Observatory (VO) offers an ideal basis for dealing with distributed heterogeneous datasets.

The VO is an international astronomical community-based initiative. The main goal of the VO is to enable new science by making the huge amount of data presently on-hand easily accessible to astronomers and providing a new research environment that will enable new possibilities for scientific research based on data discovery, efficient data access and interoperability.

This work is a case study aimed at using VO tools and Data Mining technologies to process data towards the classification of faint radio sources. Multi-wavelength data from X-ray, Ultra-Violet, Optical, and Infrared will be used to analyse and classify the faint radio population into different classes (i.e. AGN, Star-forming, late type, early type, ERO, LRG, etc.). The nature of Optically/X-ray unidentified faint radio sources will also be explored. A further aim is to automate the processing chain.

This work will use the wealth of archival data already available. Radio data will be compiled from FIRST and GMRT surveys. Multi-wavelength data will be queried from SDSS DR14, ROSAT, GALEX, 2MASS and WISE surveys. Examples of Data Mining algorithms that will be used in this project include Decision Tree classifier, Support Vector Machine Classifier, KNN, and Random Forest, etc. This work will be using a number of VO tools (i.e. Topcat, Aladin, VOSpec) and will also make use of some Data Mining Packages like WEKA (Open Source Machine Learning Software) and DAMEWARE (Web Application REsources of DAME (Data Mining & Exploration)).

☐ Special Requirements/Technical Competencies

- Programming experience. Preferably some experience with Python and using IPython/ JupyterLab
- A background in computer science would be an advantage.