SARAO 2022 PhD Project Proposal

Section A

1. Project Title

   Enabling Scientific Discovery in Radio Data with Machine Learning

2. Research Area

   Science

3. Academic Level

   PhD

4. Abstract

Radio telescopes such as MeerKAT, ASKAP, LOFAR, VLA and eventually the SKA are capable of producing datasets of unprecedented size and complexity. Coupling large area with high resolution and sensitivity, these enormous datasets represent a rich source of potential scientific discoveries. However, datasets are quickly becoming too large to manually scan and identify anomalous objects that may be scientifically interesting and without automation, such discoveries may be missed.

In this project, we propose to utilise and extend the publicly available software Astronomaly [1] to optimise the search for scientifically interesting anomalies in radio data using machine learning. The project will combine development of new data science techniques, including image manipulation, deep learning and other machine learning approaches, with scientific exploration of publicly available radio data, with a focus on MeerKAT. We will also explore the power of combining multiwavelength data for anomaly detection, particularly radio and optical, with a view of developing a framework for future data from the SKA and the Vera C. Rubin Observatory.

5. Primary Supervisor's Details

a. **Full name:** Dr. Michelle Lochner
b. **Email address:** mlochner@uwc.ac.za
c. **University:** University of the Western Cape/ South African Radio Astronomy Observatory

Section B

1. Scientific Merit

New SKA pathfinder telescopes have been making significant scientific discoveries at a rapid rate owing to their groundbreaking sensitivity, resolution and survey speed. Recent examples of interesting radio sources found in MeerKAT data include X-shaped galaxies [2], unusual synchrotron threads [3] and giant radio galaxies [4]. These amazing discoveries have all been made the traditional way: with a human expert manually inspecting the data.

However as datasets grow to the petabyte scale and catalogues of sources number in the millions, it becomes increasingly difficult to manually inspect all objects and there is a high probability interesting sources may be missed. New automated techniques must be developed to sort through the data and highlight anomalous sources that warrant further inspection.

Astronomaly [1] is a publicly available framework, developed by Lochner & Bassett (2020), that runs machine learning on a wide variety of datasets, including radio images, to detect anomalies. It makes use of an interactive web-based frontend to allow active learning: the use of a small amount of human labels to further improve the algorithm. While Astronomaly has been successfully applied to some MeerKAT data, including the MeerKAT Galaxy Clusters Legacy Survey, significant improvements are required before it can be used broadly on all types of radio images.

The student will focus on the following new developments:

- ➢ Developing an image manipulation approach that allows for automatic resizing of cutouts to allow for very different sized sources (currently a single cutout size must be selected)
- ➢ Developing new feature extraction techniques based on deep learning that are optimised for radio images. These include autoencoders, self-organising maps [5] and deep self-organising maps [6]
- ➢ As well as applying existing algorithms, the student will investigate the merit in implementing new machine learning based anomaly detection algorithms such as one class neural networks, DRAMA [7] and BADAC [8]

An exciting array of data is (or soon will be) publicly available for a student to work with (see below) allowing the real opportunity to discover interesting sources. Additionally, a range of multiwavelength data is available allowing a completely novel approach of multiwavelength anomaly detection. Combining multiwavelength data can also allow interesting avenues of scientific enquiry as we start to analyse entire populations of “interesting objects”.

This project will not stop at developing the techniques, we will also investigate any interesting sources discovered.

The science component of the project will include:

- Analysing interesting sources found making use of existing public multiwavelength data
- If necessary, submitting proposals to MeerKAT open-time calls and other telescopes
- Working with collaborators who are experts in the types of radio galaxies found
- Investigating classes of anomalies at a population level, particularly including optical data (such as DECALs [9]), to learn global properties of anomalous sources such as perturbed radio galaxies, x-shaped radio galaxies, giant radio galaxies etc.

2. Feasibility

This project makes use of existing, publicly available data. By focusing on radio datasets from different telescopes, we will develop more general techniques and will be sensitive to different types of sources. Datasets we will analyse include:

- MeerKAT Galaxy Clusters Legacy Survey (soon to be released)
- MeerKAT DEEP2 data [10]
- The ASKAP RACS survey [11]
- The LOFAR Two-Metre Sky Survey [12]

Additional data to follow-up interesting sources may be required, which we will submit proposals to acquire.

In terms of computational feasibility, by incorporating their work into the existing Astronomaly framework, the student can focus on research rather than developing infrastructure. Astronomaly is a light-weight framework meaning a significant amount of work can be done on a simple laptop. However for very large datasets, the student may require access to supercomputing facilities. As a staff member in the Astronomy group at UWC, Lochner and her students have access to the IDIA supercomputer. Additionally, for any optical or multiwavelength work, Lochner has access to NERSC in the USA, which hosts the DECALS data and will host LSST data in the future.

Rough timeline:

- 2022 - Develop and optimise new machine learning techniques for anomaly detection applied to existing MeerKAT data (primarily the MGCLS as an example dataset)
- 2023 - Apply anomaly detection to datasets listed above, optimising code as necessary. Identify objects of interest and targets of follow-up
- 2024 - Focus on scientific investigation of objects of interest, combining with multiwavelength data and follow-up data as necessary

We expect to write at least one journal article on the application of anomaly detection to radio data and likely a second article focusing on scientifically interesting discoveries.

3. Link to SARAO research priority areas

This project is focused heavily on exploiting existing radio data, primarily MeerKAT. The tools developed have the potential to yield great scientific returns as more MeerKAT data is taken and becomes available, especially to take full advantage of MeerKAT's policy of making data public after a proprietary period. Additionally, this project will develop a valuable and rare skillset in the student of data science and especially machine learning techniques. In the so-called "fourth industrial revolution", South Africa severely lacks people who possess these skills. By developing these skills and tools in South Africa, we place South African scientists at the forefront of scientific discovery with MeerKAT data and places us in an equally good position for when the SKA begins to take data. The multiwavelength element builds on Lochner's strong ties to the USA-lead project the Vera C. Rubin Observatory LSST.

4. Student Requirements

Good programming skills are critical for this project, primarily in python. Experience with machine learning is advantageous, but not essential.

References

[1] https://arxiv.org/abs/2010.11202 (M. Lochner & B. Bassett, 2020)
[2] https://arxiv.org/abs/2005.02723 (W.D. Cotton et al., 2020)
[3] https://arxiv.org/abs/2004.03203 (M. Ramatsoku et al., 2020)
[4] https://arxiv.org/abs/2012.05759 (J. Delhaize, 2020)
[5] https://ui.adsabs.harvard.edu/abs/2007SchpJ...2.1568K (T. Kohonen & T. Honkela, 2007)
[6] https://www.esann.org/proceedings/2019 (F. Forest et al. 2019)
[7] https://arxiv.org/abs/1909.04060 (A. Vafaei Sadr et al. 2019)
[8] https://arxiv.org/abs/1902.08627 (E. Roberts, B.A. Bassett & M. Lochner, 2019)
[9] https://arxiv.org/abs/1804.08657 (A. Dey et al. 2018)
[10] https://arxiv.org/abs/1912.06212 (T. Mauch et al., 2018)
[11] https://arxiv.org/abs/2012.00747 (D. McConnell et al. 2020)
[12] https://arxiv.org/abs/1811.07926 (T. W. Shimwell et al. 2018)