# An Evaluation of Machine Learning Methods for Morphological Galaxy Classification

## Towards a Cognitive Vision System for Astronomy

Victor Gueorguiev, Julius Stopforth, Roy Henha Eyono

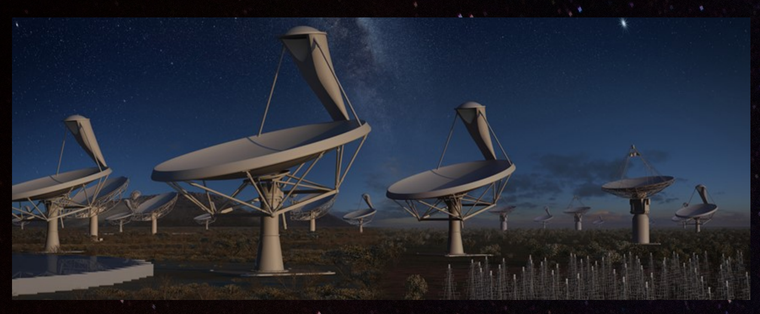
University of Cape Town

## Aim

Build and Evaluate a machine learning testbench of various state-of-the-art approaches to galaxy morphology classification on the Galaxy Zoo dataset. Use this testbench as a foundation to builda cognitive vision system to efficiently and automatically classify galaxy phenomena.

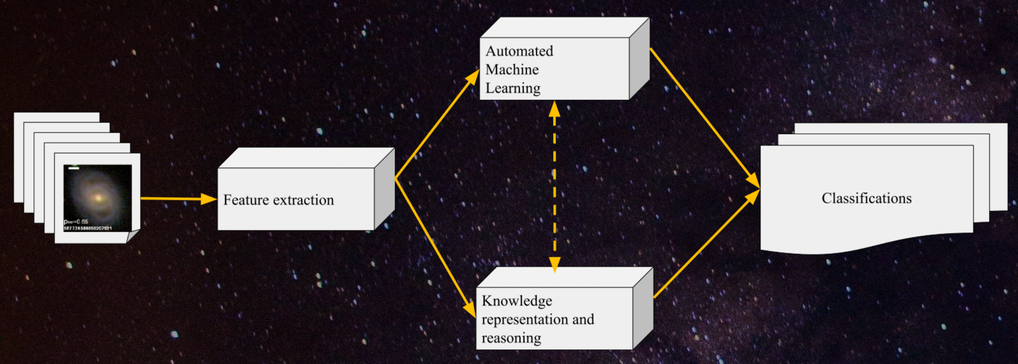
## Background

**South Africa** will soon be one of the leading contributors to the field of radio astronomy. The **Square Kilometre Array (SKA)** is anticipated to produce exabytes of data in a single year, at 2 TB/s during operation.



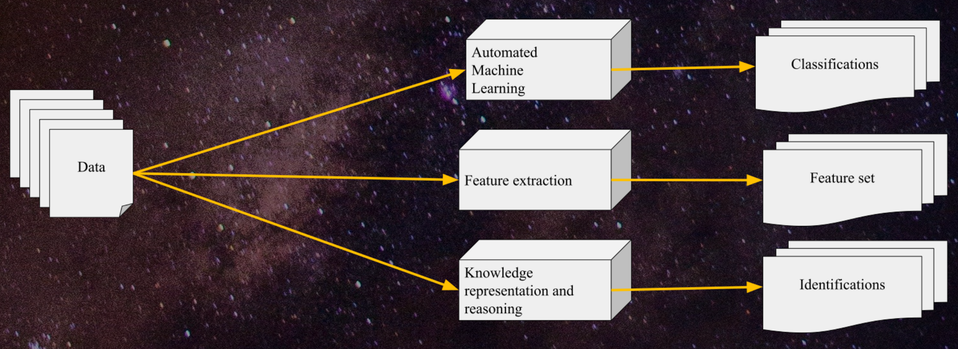
## Cognitive Vision System Structure

**A Cognitive Vision System** or intelligent system, is any system that aims to gather expert knowledge and combine it with predictive capabilities to model and explain future events. Three components will be used in our CVS for astronomy: Feature extraction, Machine learning hyperparameter optimization and knowledge reasoning.




## Components of Our Cognitive Vision System

These components entail a feature selection step for the extraction of a feature representation from a dataset, a machine learning approach harnessed by model selection and hyperparameter optimization techniques, and inference capabilities embodied by expert domain knowledge of astronomy for spatial and temporal reasoning for identification of phenomenon in optical images.




### Feature Extraction

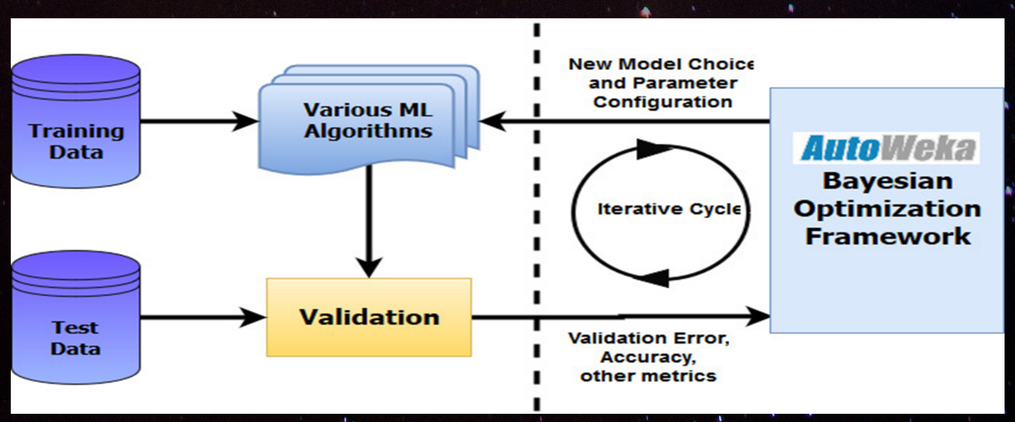
Feature extraction in image processing is a method of transforming large redundant data into a reduced data representation. In our model, we'll perform the WND-CHARM [1] scheme for feature extraction. Weighted neighbor distances using a compound hierarchy of algorithms representing morphology or WND-CHARM is a multi-purpose classifier which can be applied to a variety of image classification tasks without modifications or fine-tuning.

The features extracted from the WND-CHARM scheme will be accompanied by features that describe the brightness, shape, average size and roughness of the galactic images. Hence, amounting to the extraction of approximately 3010 features, of which, approximately 3000 is attributed to the WND-CHARM scheme and the remainder to the additional descriptive features.

### Automated Machine Learning

The machine learning component addresses the problem of combined algorithm selection and hyperparameter optimization (CASH), which aims to both select among a set of algorithms one most appropriate for a given task and set its algorithm parameters to some optimal configuration. Choosing between different machine learning algorithms, and adjusting the manually tunable parameters is a cumbersome task [2] and can automated through optimization means.

The method that is used for CASH and building the machine learning component will be Bayesian Optimization via Auto-WEKA as outlined by Thornton et al [2]. This method is chosen for its efficiency on expensive to evaluate black-box functions such as deep neural networks, and for the few hyperparameters that this method has in its core implementation.




### Knowledge Representation & Reasoning

Knowledge representation and reasoning will be done using a Bayesian network that incorporates both an classification ontology, as well as semantic image features.

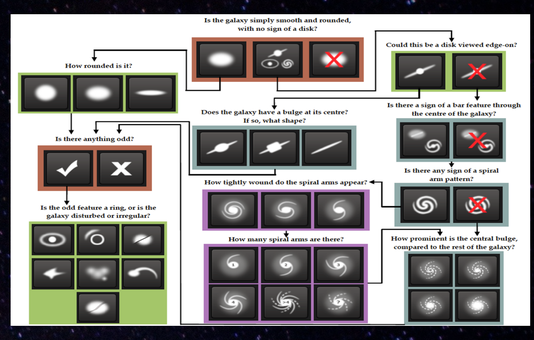
The classification ontology will encapsulate the domain knowledge of astronomy while the semantic image features will relate directly to those of the feature extraction module.

The motivation for using a Bayesian network as opposed to another method such as description logics, is that the Bayesian network not only acts as both the knowledge representation and reasoner, but the behaviour of the reasoning can be changed depending on the direction that the network is traversed.

Specifically, the network is able to reason from given observations (features) to classify items, but if the reasoning is reversed, the network is able to generate queries for specific features of a classification.

## Our Dataset

**The Galaxy Zoo** is an openly available, crowd-sourced labelled dataset using galaxy images from the Sloan Digital Sky Survey. It contains millions of classified galaxies.



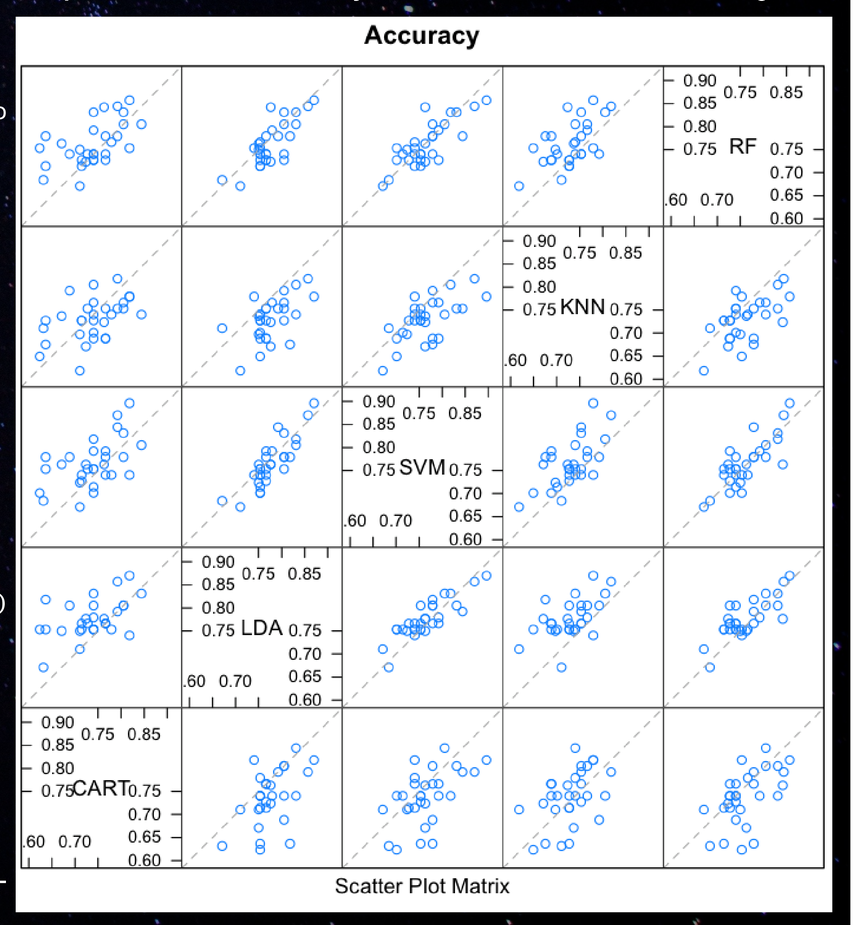

## Evaluating of Our Cognitive Vision

Evaluations will be focused on how the model fairs on astronomical datasets other than the Galaxy Zoo dataset. Cognitive Vision Systems are scarce in the field of astronomy, hence the motivation to conduct our evaluations from dataset to dataset. The ability of the model to generalize will be tested through this means. The overall cognitive vision system is tested via k-fold cross validation for the machine learning components using RMSE and WND-CHARM metrics for feature selection. Knowledge reasoning is evaluated together with the ML

## Machine Learning Test Bench

**As part of the evaluation of the hyperparameter optimization and model selection component of the project, a machine learning library was set up and tested on the Galaxy Zoo dataset.** This machine learning library is open-source and freely-available. It is constructed as follows: Image data was first analysed using Principle Component Analysis (PCA) to extract features that would be fed into the various machine learning algorithms. The extracted features from PCA are then fed into various machine learning classifiers. These are:

- K-Nearest Neighbours (KNN)
- Convolutional Neural Network (CNN)
- Classification and Decision Trees (CART)
- Support Vector Machines (SVM)
- Linear Discriminant Analysis (LDA)




Results from a many-fold cross validation are shown in the scatter plot matrix below . These display plots of matrices between different methods. It shows that many of the methods agree well with a CNN approach and can thus be used in an ensemble approach, but not many agree with the KNN or decision trees. Agreeent is described by the R coefficient.

[1] Nikita Orlov, Lior Shamir, Tomasz Macura, Josiah Johnston, D Mark Eckley, and Ilya G Goldberg. 2008. WND-CHARM: Multi-purpose image classification using compound image transforms. Pattern recognition letters 29, 11 (2008), 1684–1693

[2] Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013, August). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 847-855). ACM.

science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRI