# Identifying galaxies, quasars, and stars with machine learning:

## A new catalogue of classifications for 111 million SDSS sources without spectra

A. O. Clarke, A. M. M. Scaife, R. Greenhalgh, and V. Griguta. (Jodrell Bank Centre for Astrophysics)

Classifying sources is one of the foundations of astronomy. This needs to be done accurately and automated at scale.

Sources seen in the Sloan Digital Sky Survey (SDSS) are either a galaxy, quasar, or star. But only around 3 million of these are labelled accurately.

We built a machine learning model to classify 111 million sources (plotted to the right):

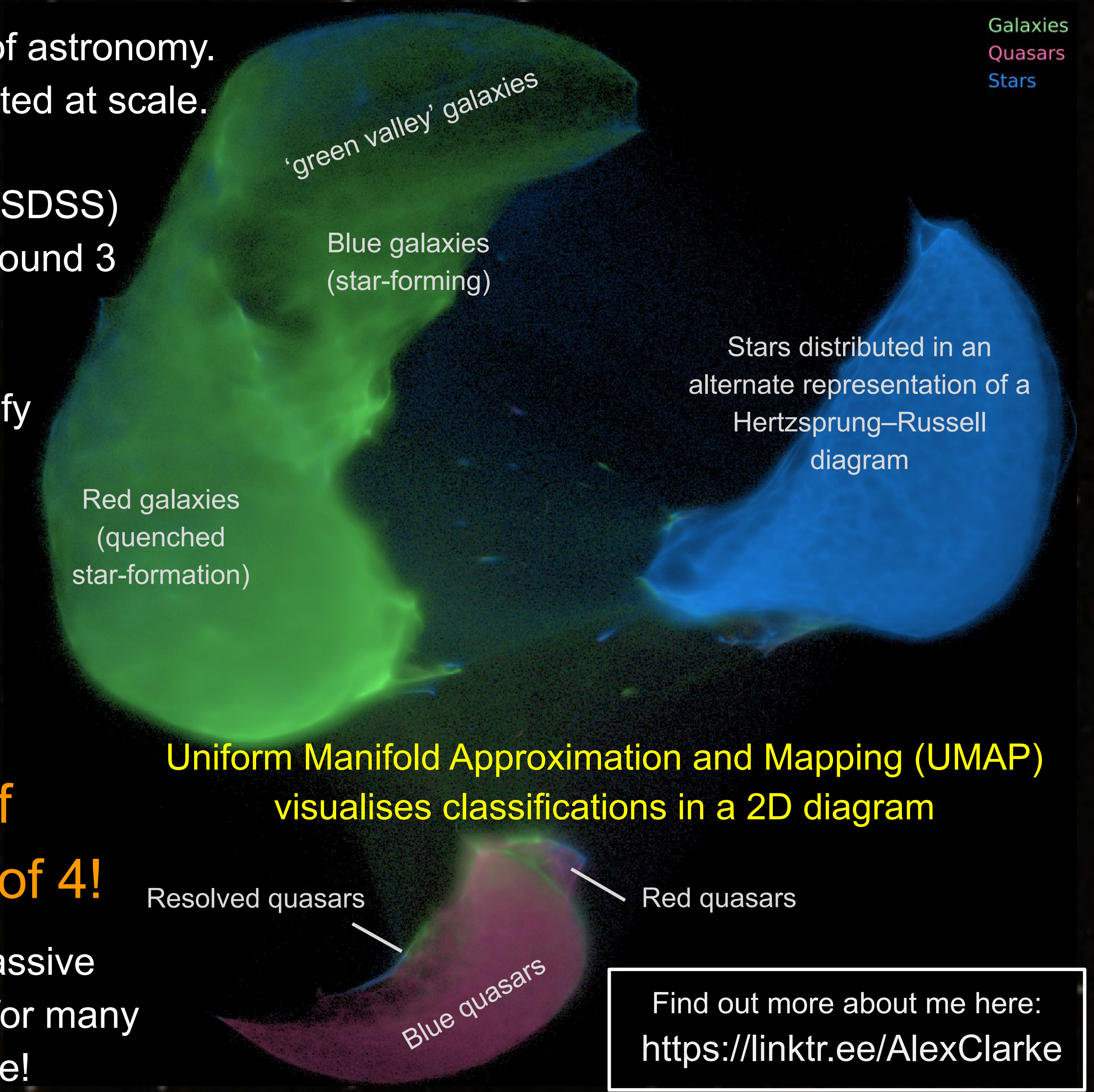
50 417 547 galaxies

2 137 839 quasars

58 840 082 stars

We increased the number of catalogued quasars by a factor of 4!

Quasars are galaxies which host supermassive blackholes at their centre and are essential for many science goals, so we need to find more!



Uniform Manifold Approximation and Mapping (UMAP) visualises classifications in a 2D diagram

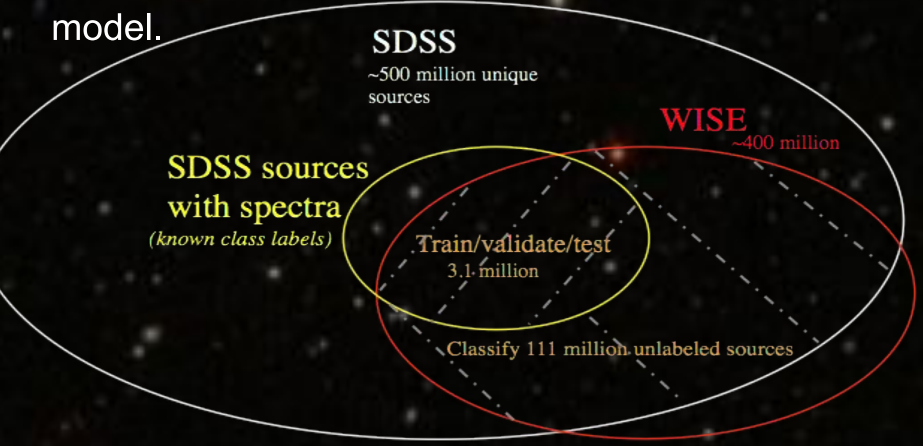
Find out more about me here:
https://linktr.ee/AlexClarke

# Identifying galaxies, quasars, and stars with machine learning:

## A new catalogue of classifications for 111 million SDSS sources without spectra

**Model** We used 3.1 million sources with known labels (from spectra) to train and test a Random Forest using **features** from optical and infrared photometry.

This Venn diagram shows how we selected sources for training and classification. We only used sources cross-matched with the Widefield Infrared Survey Explorer (WISE), as a wider wavelength range reduces bias in the model.


SDSS
~500 million unique sources
WISE
~400 million
SDSS sources with spectra
(known class labels)
Train/validate/test
3.1 million
Classify 111 million unlabeled sources

Half of the 3.1 million labelled sources were used to train a model, and the other half to test it (plotted below).

Accurate source classification is done by taking spectra - this is slow, taking broadband photometry is quick.

### Features used in the model:

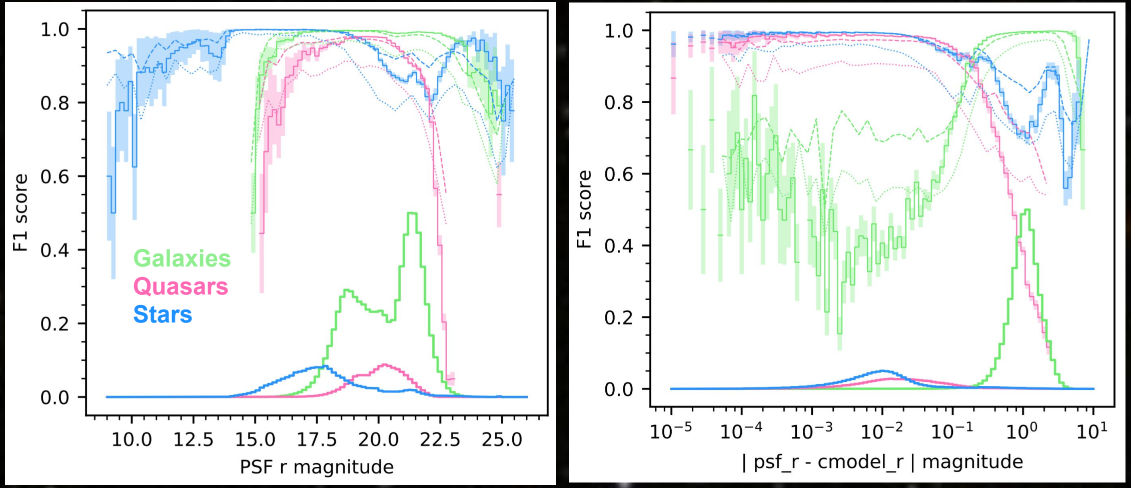
Photometry in 5 SDSS frequency bands ($u, g, r, i, z$)

Measure how resolved the source is by comparing point spread function (PSF) magnitude to a model magnitude

$$resolved_r = |psf_r - cmod_r|.$$

Photometry in 4 WISE frequency bands ($w1, w2, w3, w4$)

In total we have 10 features per source



Shaded region - Error (Wilson interval score)
Dashed lines - Mean class probability
Dotted lines - 1 standard deviation below the mean class probability
Histogram per class is shown in the lower half of the plot peaking at 0.5

## How good is the model?

F1 score is a performance metric assessing true positives (TP), false negatives (FN) and false positives (FP). We can measure these per class (right table) and as a function of variables such as *magnitude* or *resolved*$_r$ (left plots) to see where the model is strongest and weakest.

| $F_1$ score | | |
|---|---|---|
| Galaxy | Quasar | Star |
| 0.991 | 0.952 | 0.978 |

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Find out more about me here:
https://linktr.ee/AlexClarke

# Identifying galaxies, quasars, and stars with machine learning:

## A new catalogue of classifications for 111 million SDSS sources without spectra

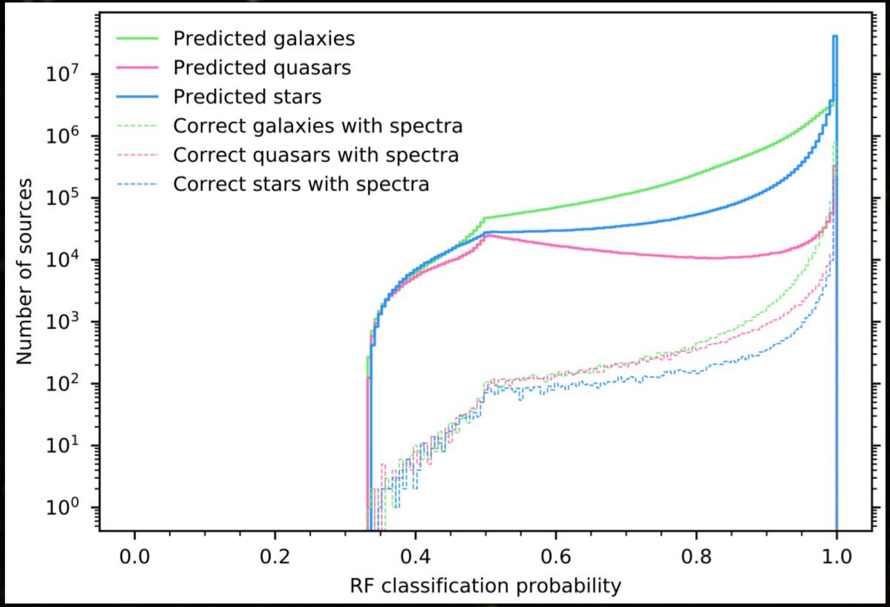
### Classifying new sources with our model

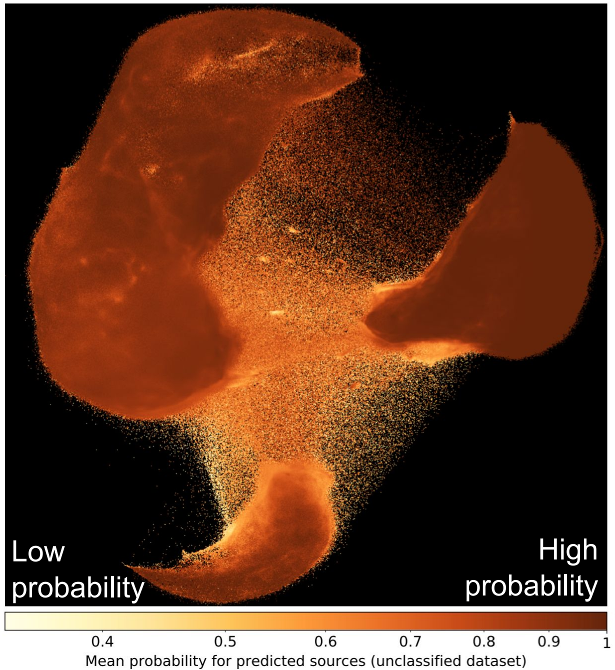
The F1 scores from the test data tell us how the model will perform on unseen data. We show that F1 scores correlate with classification probabilities returned by the Random Forest (see the paper). For new sources, the classification probabilities allow us to evaluate the confidence of the classifications without spectroscopic truth labels (plotted to the right).

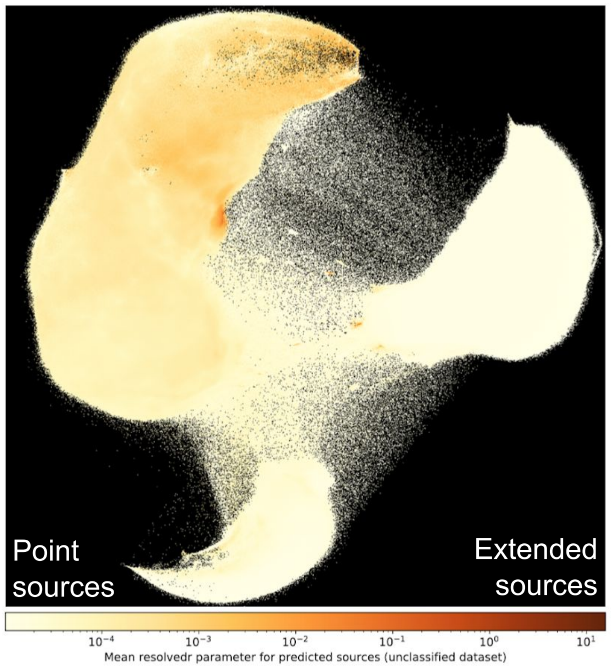
A spectroscopic follow-up survey could target quasars we have identified that have high classification probabilities.

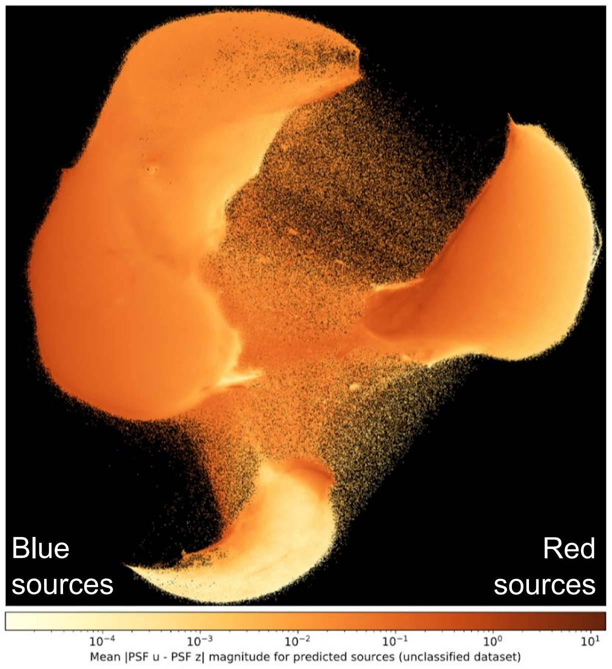
The plots below use Uniform Manifold Approximation and Mapping (UMAP) to reduce the number of dimensions from 10 to 2, allowing us to visualise the distribution of the classes, and correlations with features/variables.

35.1 million galaxies (70%), 0.72 million quasars (34%), and 54.7 million stars (93%) have classification probabilities greater than 0.9









To maintain clarity when plotting 111 million data points we used DataShader, which bins sources per pixel and colours it in proportion to the average value (plots to the left) or in proportion to the number count (image on the 1st slide)

Find out more about me here:
https://linktr.ee/AlexClarke

# Identifying galaxies, quasars, and stars with machine learning:

## A new catalogue of classifications for 111 million SDSS sources without spectra

A. O. Clarke, A. M. M. Scaife, R. Greenhalgh, and V. Griguta. (Jodrell Bank Centre for Astrophysics)

Automated classification methods will be essential for current and next generation astronomical surveys. We hope this work has shown you the potential of machine learning in astronomy and provided inspiration for your own research.

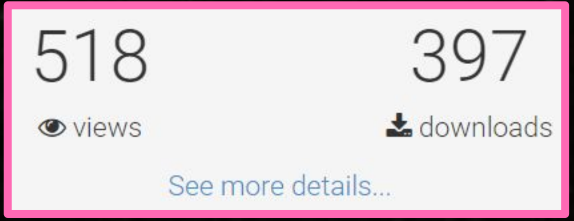
The paper: https://arxiv.org/abs/1909.10963

The code: https://github.com/informationcake/SDSS-ML

The data: https://www.doi.org/10.5281/zenodo.3459293

We want to promote open science practices in research. We ensured our result was reproducible by providing all the code and data. Each is given a Digital Object Identifier (DOI), enabling the code and data to be cited along with the paper. This also allows different versions to be tracked publicly, from submission to a journal, to the published result, and any future updates.

Our data is available on **Zenodo**, making both our catalogue and cleaned training data citable via the DOI. We can also track views and downloads for each version - citations are not the only important impact metric.



518

👁 views

397

downloads

See more details...

Our code is available on **Github**, which also has an associated DOI, making our code citable in case anyone wants to use parts of it.

As a bonus, in case of civilisation collapse and the loss of all digital information, our Github repository is now stored on film in a vault in the Arctic :)

**Arctic Code Vault Contributor**

**Alex Clarke** contributed code to 1 repository in the **2020 GitHub Archive Program**:

informationcake/SDSS-ML

Thank you for taking the time to read my poster. Any questions or feedback is very welcome: a.clarke@skatelescope.org

This work was done at the University of Manchester, in collaboration with professor Anna Scaife and undergraduate students Robin Greenhalgh and Vlad Griguta. I am currently a regional centre scientist at the Square Kilometer Array (SKA) headquarters at Jodrell Bank.

Find out more about me here:
https://linktr.ee/AlexClarke