Machine Learning with Crowdsourced Labels on the Radio Galaxy Zoo

Matthew Alger
The Australian National University

April 18, 2016

1 The Radio Galaxy Zoo

In this section, I will introduce the Radio Galaxy Zoo, the problem it aims to solve, and how this relates to my project.

Black holes are masses that are so dense that even light cannot escape their gravitational field. Supermassive black holes are found at the centre of most galaxies, but astronomers don't know a great deal about them or their interactions with the galaxy surrounding them, which is called the *host galaxy.* To find out more, astronomers need to look at many different black holes and their host galaxies. The host galaxies turn out to be particularly important: You can't get a lot of information from looking at a black hole by itself. Even simple things like how big the black hole is need more information, and this information comes from the host galaxy. From this, it's apparent that even if we have observations of black holes, we need to know what galaxy hosts them.

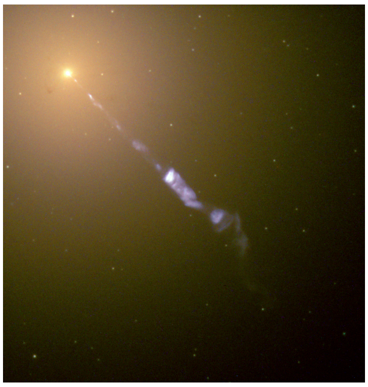
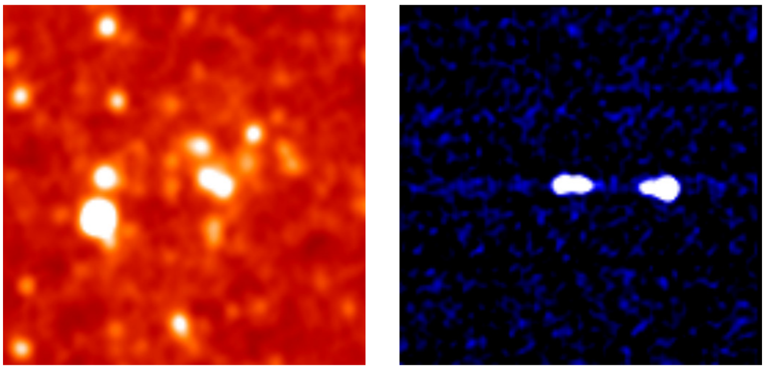


Figure 1: A jet emitted by a black hole located in galaxy M87. *Credit: NASA.*

Black holes can't be seen directly, since they don't emit light. However, black holes draw in huge amounts of nearby matter, and some of this matter is ejected in a *jet* (Figure 1). These jets emit light. For distant supermassive black holes, the light from jets is in the form of radio waves, which we can see using radio telescopes. We call these radio-emitting black holes *radio sources.* Radio surveys of the sky, such as the Australia Telescope Large Area Survey (ATLAS)[5] and Faint Images of the Radio Sky at Twenty-Centimeters (FIRST)[2], provide us with many images of jets emitted by radio sources. Distant galaxies emit infrared light, which we see using infrared telescopes. Infrared surveys of the sky, such as the Wide-Field Infrared Survey Explorer (WISE)[7] and the Spitzer Wide-Area Infrared Extragalactic Survey (SWIRE)[3], provide us with images of these galaxies. So that we can investigate black holes, we need to match radio sources observed in radio images to their host galaxies observed in infrared images. This is called *cross-identification.*

Cross-identification is very difficult. Jets can be very complex, and it may not be clear where the radio source associated with the jets is located. For this reason, cross-identification is usually done by hand[1], but this is impractical: There are already around 2.5×10^6 known radio sources, and future radio surveys such as the Evolutionary Map of the Universe (EMU)[4] and WODAN[6] are expected to find over 10^8 new radio sources[1]. *Radio Galaxy Zoo* is a crowdsourced citizen science project that tasks volunteers



(a) Infrared image.

(b) Radio image.

Figure 2: Infrared and radio images of Radio Galaxy Zoo subject ARG000180p.

with cross-identifying 177461 images of radio sources through a web interface. [TODO(MatthewJA): details] To date[1], 94407 sources have been cross-identified, and Banfield et al.[1] report that when over 75% of volunteers agree on a cross-identification, the volunteers are as good as expert astronomers at the cross-identification task.

Of course, this still isn't enough to cross-identify all of the new radio sources we expect to find. The hope is that the cross-identifications provided by the Radio Galaxy Zoo volunteers can be used as labels for training a supervised machine learning algorithm on the cross-identification task. This hasn't been done before, and is the central task of my honours project.

2 Data Sources

In this section, I will describe the data files used in the project.

Julie has provided me with four files:

- The ATLAS catalogue,
- CDFS radio and infrared images,
- ELAIS radio and infrared images, and
- Radio Galaxy Zoo subjects and classifications.

The ATLAS catalogue describes each object detected in ATLAS. However, the IDs of each object are actually different to the IDs in the images, so it's important to use the name as the primary ID of each object, and not the ID. More information is available in the crowdastro documentation.

The CDFS and ELAIS data sets are formatted identically. Each subject in each set consists of an infrared PNG and FITS, a radio PNG and FITS, and an infrared PNG with the brightness contours of the radio PNG overlayed. There are both 2 × 2 and 5 × 5 patches provided around each subject. The FITS files contain metadata (mostly location data) for each image. Both datasets also include a metadata file that contians the spatial location of each subject.

The Radio Galaxy Zoo subjects each correspond to a radio and infrared image. There are also a number of corresponding classifications; each of these represents one cross-identification by a volunteer of all radio sources in the image.

3 Potential Host Detection

Potential host detection is a subproblem of the cross-identification problem. If we have a list of potential host galaxies (hereafter *potential hosts*), then we can classify each of these as 1 if the potential host is the true host galaxy, and 0 otherwise. This then turns the cross-identification problem into a binary classification task. The question, then, is: How do we detect potential hosts? I have found three possible methods.

[1] March 2016

The first method is to find all local maxima in the infrared image of a subject, and then merge neighbouring local maxima. This requires preprocessing to reduce noise. The end result is still quite noisy and tends to pick up 200 − 500 potential hosts per subject. This is explored in notebook 2.

The second method is to use a library such as scikit-image to detect blobs in the infrared image. This is more robust though it tends to merge galaxies that are very close together. This tends to find around 50 − 80 potential hosts per subject. This is explored in notebook 10.

The third method is to consult the WISE object catalogues for the area covered by the subject, and use these objects. This will probably be a good method, since it will be in agreement with other astronomical results that use WISE data, but I haven't tried it yet.

4 Crowd Click Consensus — Majority Vote

It is often the case that not every volunteer agrees on where the radio source is located for a given subject[2]. It's useful to know how often people disagree, and if they do disagree, how much. If it turns out that people don't disagree often, or they disagree a lot but there's always a clear majority, we can just pick the most common radio source location as the "true" location.

This is explored in notebook 8.

TODO(MatthewJA): Write up results from notebook 8.

References

[1] JK Banfield, OI Wong, KW Willett, RP Norris, L Rudnick, SS Shabala, BD Simmons, C Snyder, A Garon, N Seymour, et al. Radio galaxy zoo: host galaxies and radio morphologies derived from visual inspection. *Monthly Notices of the Royal Astronomical Society*, 453(3):2326–2340, 2015.

[2] R. H. Becker, R. L. White, and D. J. Helfand. The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters. *Astrophysical Journal*, 450:559, September 1995.

[3] Carol J Lonsdale, Harding E Smith, Michael Rowan-Robinson, Jason Surace, David Shupe, Cong Xu, Sebastian Oliver, Deborah Padgett, Fan Fang, Tim Conrow, et al. SWIRE: The SIRTF wide-area infrared extragalactic survey. *Publications of the Astronomical Society of the Pacific*, 115(810):897, 2003.

[4] R. P. Norris, A. M. Hopkins, J. Afonso, S. Brown, J. J. Condon, L. Dunne, I. Feain, R. Hollow, M. Jarvis, M. Johnston-Hollitt, E. Lenc, E. Middelberg, P. Padovani, I. Prandoni, L. Rudnick, N. Seymour, G. Umana, H. Andernach, D. M. Alexander, P. N. Appleton, D. Bacon, J. Banfield, W. Becker, M. J. I. Brown, P. Ciliegi, C. Jackson, S. Eales, A. C. Edge, B. M. Gaensler, G. Giovannini, C. A. Hales, P. Hancock, M. T. Huynh, E. Ibar, R. J. Ivison, R. Kennicutt, A. E. Kimball, A. M. Koekemoer, B. S. Koribalski, Á. R. López-Sánchez, M. Y. Mao, T. Murphy, H. Messias, K. A. Pimbblet, A. Raccanelli, K. E. Randall, T. H. Reiprich, I. G. Roseboom, H. Röttgering, D. J. Saikia, R. G. Sharp, O. B. Slee, I. Smail, M. A. Thompson, J. S. Urquhart, J. V. Wall, and G.-B. Zhao. EMU: Evolutionary Map of the Universe. *PASA*, 28:215–248, August 2011.

[5] Ray P Norris, José Afonso, Phil N Appleton, Brian J Boyle, Paolo Ciliegi, Scott M Croom, Minh T Huynh, Carole A Jackson, Anton M Koekemoer, Carol J Lonsdale, et al. Deep atlas radio observations of the chandra deep field-south/spitzer wide-area infrared extragalactic field. *The Astronomical Journal*, 132(6):2409, 2006.

[6] Huub Röttgering, Jose Afonso, Peter Barthel, Fabien Batejat, Philip Best, Annalisa Bonafede, Marcus Brüggen, Gianfranco Brunetti, Krzysztof Chyży, John Conway, Francesco De Gasperin, Chiara Ferrari, Marijke Haverkorn, George Heald, Matthias Hoeft, Neal Jackson, Matt Jarvis, Louise Ker, Matt Lehnert, Giulia Macario, John McKean, George Miley, Raffaella Morganti, Tom Oosterloo, Emanuela Orrù, Roberto Pizzo, David Rafferty, Alexander Shulevski, Cyril Tasse, Ilse van Bemmel, Bas Tol, Reinout Weeren, Marc Verheijen, Glenn White, and Michael Wise. Lofar and apertif surveys of the radio sky: Probing shocks and magnetic fields in galaxy clusters. *Journal of Astrophysics and Astronomy*, 32(4):557–566, 2012.

[2] It's also that case that not every volunteer agrees on which radio emissions are associated with the same source, but we're not sure how to deal with that right now, so I'm ignoring it.

[7] E. L. Wright, P. R. M. Eisenhardt, A. K. Mainzer, M. E. Ressler, R. M. Cutri, T. Jarrett, J. D. Kirkpatrick, D. Padgett, R. S. McMillan, M. Skrutskie, S. A. Stanford, M. Cohen, R. G. Walker, J. C. Mather, D. Leisawitz, T. N. Gautier, III, I. McLean, D. Benford, C. J. Lonsdale, A. Blain, B. Mendez, W. R. Irace, V. Duval, F. Liu, D. Royer, I. Heinrichsen, J. Howard, M. Shannon, M. Kendall, A. L. Walsh, M. Larsen, J. G. Cardon, S. Schick, M. Schwalm, M. Abid, B. Fabinsky, L. Naes, and C.-W. Tsai. The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *The Astronomical Journal*, 140:1868–1881, December 2010.