# Thesis Title Here

**Matthew Alger**

A thesis submitted in partial fulfillment of the degree of
Bachelor of Science (Advanced) (Honours) at
The Research School of Computer Science
The Australian National University

August 2016

Typeset in Palatino by TEX and LATEX 2ε.

Except where otherwise indicated, this thesis is my own original work.

Matthew Alger
12 August 2016

To the ANU-themed rubber duck on my bookshelf.
TODO: *Write a real dedication.*

# Acknowledgements

I would like to thank my lucky stars, and the cat, for not eating me.

— ANU Physics Thesis Template

# Abstract

TODO: *Add abstract.*

# Contents

**Acknowledgements** vii

**Abstract** ix

**1 Imagine There's No Chapters** 1
- 1.1 Papers Read . . . 1
- 1.2 The Radio Galaxy Zoo . . . 1
- 1.3 Radio Galaxy Zoo Consensus Labels . . . 1
- 1.4 Radio Cross-identification . . . 1
- 1.5 Training Data . . . 2
- 1.6 Yan Derivation . . . 2
  - 1.6.1 Formulation . . . 2
  - 1.6.2 Expectation-Maximisation . . . 3
    - 1.6.2.1 Expectation . . . 3
    - 1.6.2.2 Maximisation . . . 4
  - 1.6.3 Gradients of the Optimisation Target . . . 6

**2 Introduction** 9
- 2.1 Section . . . 9
  - 2.1.1 Section . . . 9

**3 Galaxies and Active Galactic Nuclei** 11
- 3.1 Active Galactic Nuclei . . . 11
- 3.2 ATLAS: The Australia Telescope Large Area Survey . . . 12

**4 Machine Learning** 15

**5 Radio Cross-identification** 17
- 5.1 Cross-identification as Binary Classification . . . 17
- 5.2 Data Sources . . . 18
  - 5.2.1 Radio Data . . . 18
  - 5.2.2 Infrared Data . . . 19
  - 5.2.3 Radio Galaxy Zoo Consensus Labels . . . 19
  - 5.2.4 Norris Labels . . . 20
- 5.3 Features . . . 20
  - 5.3.1 Astronomical Features . . . 20
  - 5.3.2 Radio Image Features . . . 20
- 5.4 Choosing a Binary Classifier . . . 20

5.5 Results . . . . . 20
5.5.1 Classifying Galaxies . . . . . 20

**6 Active Learning** **23**

**7 Conclusion** **25**

Chapter 1

# Imagine There's No Chapters

## 1.1 Papers Read

- **?**
- **?**
- **?**
- **?**
- **?** (re-reading in progress)
- **?** (probably need to re-read)
- **?** (need to re-read)

## 1.2 The Radio Galaxy Zoo

## 1.3 Radio Galaxy Zoo Consensus Labels

## 1.4 Radio Cross-identification

Each radio object has some associated infrared object called the host galaxy. The cross-identification task is to find the host galaxy given the radio object.

For modelling the distribution, I have chosen to use logistic regression, i.e.

$$p(y(x) = 1 \mid x, r) = \vec{w} \cdot \vec{\phi}(x, r) \tag{1.1}$$

where $\vec{\phi}$ is a feature space mapping dependent on a galaxy and a radio object. The features should represent the galaxy in some way, so I have chosen the following feature space:

$$\vec{\phi}(x, r) = \begin{pmatrix} \vec{\text{flux}}(x) \\ \text{dist}(x, r) \\ \vec{\text{cnn}}(\text{radio}(x)) \end{pmatrix} \tag{1.2}$$

$\vec{\text{flux}}(x)$ is a vector of infrared flux measurements of $x$, which can be obtained from the infrared survey catalogue. $\text{dist}(x, r)$ is the Euclidean distance across the sky between the centre of the $x$ and the centre of $r$. $\vec{\text{cnn}}(m)$ is the output of the convolutional neural network on input image $m$, and $\text{radio}(x)$ is a $0.8' \times 0.8'$ image of the radio sky centred on $x$.

## 1.5 Training Data

The Crowdastro dataset is a set of training data for the binary classification problem described in Section 1.4. The dataset contains features and labels for all objects detected in the WISE infrared survey. The prediction task is to predict the label of an object given its features.

The features are not scaled and have not undergone any feature extraction process. They are the raw fluxes, distances, and radio images described in Section 1.4.

The labels are based on the consensus locations from the Radio Galaxy Zoo, matched to the nearest WISE object. WISE objects matched to a consensus location have the label 1, and all other objects have the label 0. Consensuses are found as described in Section 5.2.3, with consensus location decided by fitting a Gaussian mixture model. The number of Gaussians is found by a grid search minimising the Baysian information criterion.

## 1.6 Yan Derivation

In this section, I elaborate on the derivation of the expectation-maximisation formulae from **?**. Notation etc. is taken from **?**. I only consider the case of binary labels.

### 1.6.1 Formulation

We have $N$ data points $\{\vec{x}_1, \ldots, \vec{x}_N\}$, where $\vec{x}_i \in \mathbb{R}^D$. We also have a set of labels $\{y_1^{(1)}, \ldots, y_n^{(1)}, \ldots, y_1^{(T)}, \ldots, y_N^{(T)}\}$, where $y_i^{(t)} \in \mathbb{Z}_2$ is the (potentially incorrect) binary label assigned to $\vec{x}_i$ by annotator $t$. We want to train a classifier to predict labels of new data points, we want to estimate the groundtruth labels $\{z_1, \ldots, z_N\}$ where $z_i \in \mathbb{Z}_2$, and we want to model the quality of each annotator's labels. Let $\vec{x}$, $y$, and $z$ be random variables representing data points, labels, and groundtruths, respectively. The classification task is then to model $p(z \mid \vec{x})$. Define matrices to represent the data:

$$
\begin{aligned}
X &= [\vec{x}_1^T; \ldots; \vec{x}_N^T] \in \mathbb{R}^{N \times D} \\
Y &= [y_1^{(1)}, \ldots, y_1^{(T)}; \ldots; y_N^{(1)}, \ldots, y_N^{(T)}] \in \mathbb{Z}_2^{N \times T} \\
Z &= (z_1, \ldots, z_N) \in \mathbb{Z}_2^N
\end{aligned}
$$

We assume that annotator labels depend on both the data point and the groundtruth, that annotator labels have annotator-dependent noise, and that annotator labels fol-

low a Bernoulli distribution:

$$p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) = (1 - \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t))^{|y_i^{(t)} - z_i|} \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t)^{1-|y_i^{(t)} - z_i|}$$
$$\eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) = \sigma(\vec{w}_t^T \vec{x}_i - \gamma_t)$$

We use logistic regression to model the posterior distribution:

$$p(z_i = 1 \mid \vec{x}_i, \alpha, \beta) = \sigma(\vec{\alpha}^T \vec{x}_i + \beta)$$

The parameters of the model are $\vec{\theta} = \{\vec{\alpha}, \beta, \vec{w}_1, \ldots, \vec{w}_T, \gamma_1, \ldots, \gamma_T\}$.

### 1.6.2 Expectation-Maximisation

We can find optimum values of the parameters by maximising the log-likelihood $p(Y \mid X, \vec{\theta})$, i.e.

$$\begin{aligned}\vec{\theta}^\star &= \operatorname*{argmax}_{\vec{\theta}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log p(y_i^{(t)} \mid \vec{x}_i, \vec{\theta}) \\ &= \operatorname*{argmax}_{\vec{\theta}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \sum_{z_i=0}^{1} p(y_i^{(t)}, z_i \mid \vec{x}_i, \vec{\theta})\end{aligned}$$

noting that we assume independence between labels of different data points and labels from different annotators. Since $z_i$ are latent variables, we must use expectation-maximisation.

#### 1.6.2.1 Expectation

For the expectation step, we want to evaluate $p(z_i \mid \vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, \vec{\theta})$ for $i = 1, \ldots, N$. We can write this in terms of the parameters:

$$\begin{aligned}p(z_i \mid \vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, \vec{\theta}) &= \frac{1}{A_i} p(z_i, y_i^{(1)}, \ldots, y_i^{(T)} \mid \vec{x}_i, \vec{\theta}) \\ &= \frac{1}{A_i} \prod_{t=1}^{T} p(z_i, y_i^{(t)} \mid \vec{x}_i, \vec{\theta}) \\ &= \frac{1}{A_i} \prod_{t=1}^{T} p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{\theta}) p(z_i \mid \vec{x}_i, \vec{\theta}) \\ &= \frac{1}{A_i} \prod_{t=1}^{T} p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) p(z_i \mid \vec{x}_i, \vec{\alpha}, \beta)\end{aligned}$$

$A_i$ is a normalisation term given by

$$A_i = \sum_{z_i=0}^{1} p(z_i, y_i^{(1)}, \ldots, y_i^{(T)} \mid \vec{x}_i, \vec{\theta}).$$

To simplify notation, let $\tilde{p}(z_i) = p(z_i \mid \vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, \vec{\theta})$.

#### 1.6.2.2 Maximisation

For the maximisation step, we want to set

$$\vec{\theta}^{\text{new}} = \underset{\vec{\theta}^{\text{new}}}{\operatorname{argmax}} \sum_{i=1}^{N} Q_i(\vec{\theta}^{\text{new}}, \vec{\theta})$$

where

$$Q_i(\vec{\theta}^{\text{new}}, \vec{\theta}) = \sum_{z_i=0}^{1} p(z_i \mid \vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, \vec{\theta}) \log p(\vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, z_i \mid \vec{\theta}^{\text{new}}).$$

Once again, we need to write this in terms of the parameters.

$$\begin{aligned}
Q_i(\vec{\theta}^{\text{new}}, \vec{\theta}) &= \sum_{z_i=0}^{1} \tilde{p}(z_i) \log p(\vec{x}_i, y_i^{(1)}, \ldots, y_i^{(T)}, z_i \mid \vec{\theta}^{\text{new}}) \\
&= \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) \log p(\vec{x}_i, y_i^{(t)}, z_i \mid \vec{\theta}^{\text{new}}) \\
&= \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) \log(p(y_i^{(t)}, z_i \mid \vec{x}_i, \vec{\theta}^{\text{new}}) p(\vec{x}_i \mid \vec{\theta}^{\text{new}})) \\
&= T \log p(\vec{x}_i \mid \vec{\theta}^{\text{new}}) + \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) \log p(y_i^{(t)}, z_i \mid \vec{x}_i, \vec{\theta}^{\text{new}}) \\
&= T \log p(\vec{x}_i \mid \vec{\theta}^{\text{new}}) + \\
&\quad \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i)(\log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{\theta}^{\text{new}}) + \log p(z_i \mid \vec{x}_i, \vec{\theta}^{\text{new}})) \\
&= T \log p(\vec{x}_i \mid \vec{\theta}^{\text{new}}) + \\
&\quad \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i)(\log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t^{\text{new}}, \gamma_t^{\text{new}}) + \log p(z_i \mid \vec{x}_i, \vec{\alpha}^{\text{new}}, \beta^{\text{new}}))
\end{aligned}$$

Then the maximisation step is

$$\vec{\theta}^{\text{new}} = \underset{\vec{\theta}^{\text{new}}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i)(\log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t^{\text{new}}, \gamma_t^{\text{new}}) + \log p(z_i \mid \vec{x}_i, \vec{\alpha}^{\text{new}}, \beta^{\text{new}}))$$

noting that $T \log p(\vec{x}_i \mid \vec{\theta}^{\text{new}})$ is the same for all $\vec{\theta}^{\text{new}}$ as $x_i$ is observed. To simplify notation, let

$$f(\vec{\theta}) = \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i)(\log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) + \log p(z_i \mid \vec{x}_i, \vec{\alpha}, \beta)).$$

where $\tilde{p}(z_i)$ is evaluated using the old value of $\vec{\theta}$.

### 1.6.3 Gradients of the Optimisation Target

In this section, I derive the gradients of $f$ with respect to the parameters $\vec{\theta}$.

First, we differentiate with respect to $\vec{\alpha}$.

$$\begin{aligned}
\nabla_{\vec{\alpha}} f(\vec{\theta}) &= T \sum_{i=1}^{N} \sum_{z_i=0}^{1} \nabla_{\vec{\alpha}} (\tilde{p}(z_i) \log p(z_i \mid \vec{x}_i, \vec{\alpha}, \beta)) \\
&= T \sum_{i=1}^{N} \sum_{z_i=0}^{1} \tilde{p}(z_i) \nabla_{\vec{\alpha}} \log p(z_i \mid \vec{x}_i, \vec{\alpha}, \beta) \\
&= T \sum_{i=1}^{N} \tilde{p}(z_i = 1) \nabla_{\vec{\alpha}} \log \sigma(\vec{\alpha}^T \vec{x}_i + \beta) + \tilde{p}(z_i = 0) \nabla_{\vec{\alpha}} \log(1 - \sigma(\vec{\alpha}^T \vec{x}_i + \beta)) \\
&= T \sum_{i=1}^{N} \frac{\tilde{p}(z_i = 1)}{\sigma(\vec{\alpha}^T \vec{x}_i + \beta)} \nabla_{\vec{\alpha}} \sigma(\vec{\alpha}^T \vec{x}_i + \beta) - \frac{\tilde{p}(z_i = 0)}{1 - \sigma(\vec{\alpha}^T \vec{x}_i + \beta)} \nabla_{\vec{\alpha}} \sigma(\vec{\alpha}^T \vec{x}_i + \beta) \\
&= T \sum_{i=1}^{N} \Big( \tilde{p}(z_i = 1)(1 - \sigma(\vec{\alpha}^T \vec{x}_i + \beta)) - \tilde{p}(z_i = 0) \sigma(\vec{\alpha}^T \vec{x}_i + \beta) \Big) \vec{x}_i \\
&= T \sum_{i=1}^{N} \Big( \tilde{p}(z_i = 1)(1 - \sigma(\vec{\alpha}^T \vec{x}_i + \beta)) - (1 - \tilde{p}(z_i = 1)) \sigma(\vec{\alpha}^T \vec{x}_i + \beta) \Big) \vec{x}_i \\
&= T \sum_{i=1}^{N} \Big( \tilde{p}(z_i = 1) - \sigma(\vec{\alpha}^T \vec{x}_i + \beta) \Big) \vec{x}_i
\end{aligned}$$

Following similar logic, we also obtain the gradient with respect to $\beta$:

$$\frac{\partial f}{\partial \beta} f(\vec{\theta}) = T \sum_{i=1}^{N} \tilde{p}(z_i = 1) - \sigma(\vec{\alpha}^T \vec{x}_i + \beta)$$

We now differentiate with respect to $\vec{w}_t$.

$$
\begin{aligned}
\nabla_{\vec{w}_t} f(\vec{\theta}) &= \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} \tilde{p}(z_i) \nabla_{\vec{w}_t} \log p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) \\
&= \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} \frac{\tilde{p}(z_i)}{p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t)} \nabla_{\vec{w}_t} p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) \\
&= \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} \frac{\tilde{p}(z_i)}{p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t)} \frac{\partial}{\partial \eta_t} p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t) \nabla_{\vec{w}_t} \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) \\
&= \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} -\tilde{p}(z_i) \frac{(1-\eta_t)^{|y_i^{(t)} - z_i| - 1} \eta_t^{-|y_i^{(t)} - z_i|} \left( \eta_t + |y_i^{(t)} - z_i| - 1 \right)}{p(y_i^{(t)} \mid \vec{x}_i, z_i, \vec{w}_t, \gamma_t)} \nabla_{\vec{w}_t} \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) \\
&= \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} -\tilde{p}(z_i) \frac{\eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) + |y_i^{(t)} - z_i| - 1}{(1 - \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t)) \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t)} \nabla_{\vec{w}_t} \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) \\
&= \sum_{i=1}^{N} \sum_{z_i=0}^{1} \sum_{t=1}^{T} -\tilde{p}(z_i) \left( \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) + |y_i^{(t)} - z_i| - 1 \right) \vec{x}_i \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T} -\tilde{p}(z_i = 1) \left( \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) - y_i^{(t)} \right) \vec{x}_i - \tilde{p}(z_i = 0) \left( \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) + y_i^{(t)} - 1 \right) \vec{x}_i \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T} \left( 2\tilde{p}(z_i = 1) y_i^{(t)} - \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) - y_i^{(t)} + 1 - \tilde{p}(z_i = 1) \right) \vec{x}_i \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T} \left( 2\tilde{p}(z_i = 1) y_i^{(t)} - \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) - y_i^{(t)} + \tilde{p}(z_i = 0) \right) \vec{x}_i
\end{aligned}
$$

Similarly, the gradient with respect to $\gamma_t$ is

$$
\nabla_{\gamma_t} f(\vec{\theta}) = \sum_{i=1}^{N} \sum_{t=1}^{T} 2\tilde{p}(z_i = 1) y_i^{(t)} - \eta_t(\vec{x}_i \mid \vec{w}_t, \gamma_t) - y_i^{(t)} + \tilde{p}(z_i = 0)
$$

Chapter 2

# Introduction

## 2.1 Section

### 2.1.1 Section

Chapter 3

# Galaxies and Active Galactic Nuclei

Modern astronomy relies on observations of deep space. Telescopes image the sky in different wavelengths, with different wavelengths carrying different physical meanings. Infrared surveys detect star formation and dust in distant galaxies, and radio surveys detect massive objects called active galactic nuclei. In this section, I introduce the physics of what we see when we look at the sky in these wavelengths, as well as introducing specific radio and infrared surveys relevant to this thesis. I will also discuss the motivation behind cross-identifying active galactic nuclei with their host galaxies as well as the inherent difficulty in doing so, and hence the motivation behind this thesis.

## 3.1 Active Galactic Nuclei

Many galaxies contain a supermassive black hole in their centre. These black holes accrete matter from the surrounding galaxy in an accretion disk TODO: *figure*. The accretion process emits huge amounts of light through different physical processes. These light-emitting black holes are called active galactic nuclei (AGNs). AGNs can be extremely bright, emitting up to $10^{39}$ J of energy every second — nearly a thousand times more energy than our entire galaxy emits (**?**). AGNs are found throughout the universe: The closest known AGN is Centaurus A (shown in Figure 3.1, with $z \approx 0.0018$, and AGNs have been detected up to redshifts of $z \approx 7$ TODO: *citation needed*.

Many AGNs produce *jets* from their accretion disk. Jets are long, thin streams of matter such as the one shown in Figure 3.2. These jets can be very long, with "giant" AGNs emitting jets up to 1 Mpc in length TODO: *citation needed — maybe* **??**. The process through which jets are emitted is currently unknown TODO: *citation needed*.

Electrons in jets produce *synchrotron radiation*, a form of radiation emitted by charged particles as they accelerate at relativistic speeds in a magnetic field TODO: *citation needed*. This radiation is emitted in radio wavelengths, and so AGNs emitting this kind of radiation are called *radio AGNs*. It is radio AGNs that are the focus of this thesis, and from this point on, "AGN" will refer only to radio AGNs.

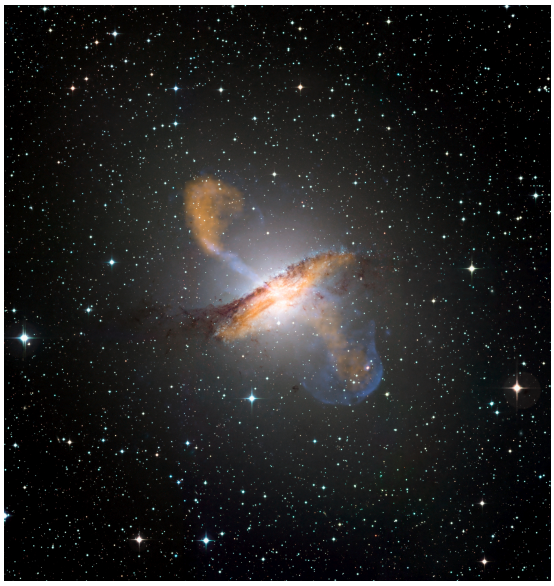

**Figure 3.1:** Centaurus A, a relatively close radio active galactic nuclei. *Image: ESO/WFI (Optical); MPIfR/ESO/APEX/A.Weiss et al. (Submillimetre); NASA/CXC/CfA/R.Kraft et al. (X-ray)*

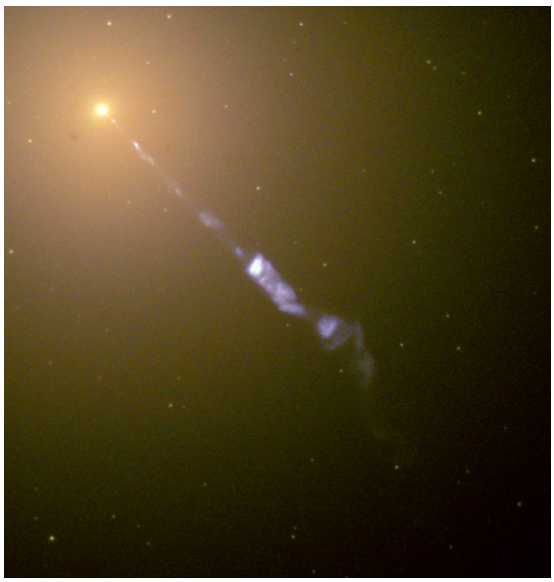

**Figure 3.2:** M87, a giant elliptical galaxy with a jet. *Image: NASA and The Hubble Heritage Team (STScI/AURA)*

## 3.2 ATLAS: The Australia Telescope Large Area Survey

The Australia Telescope Large Area Survey (ATLAS) is a deep TODO: *define?* survey of two small areas of the sky in radio wavelengths, which aims to help understand the evolution of early galaxies (**?**). The Australia Telescope Compact Array was used to image the Chandra Deep Field South (CDFS) and the European Large Area ISO

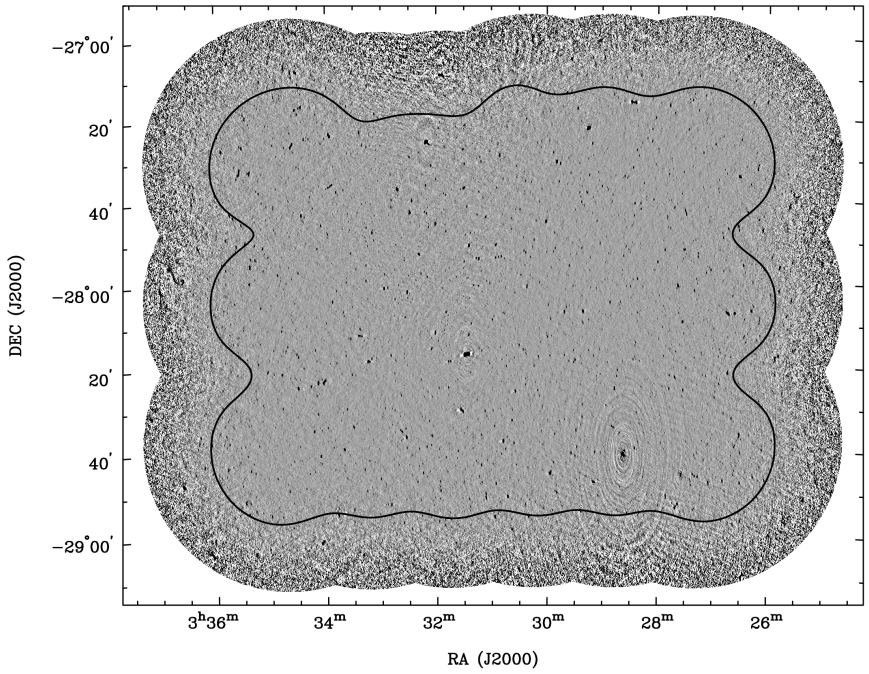


**Figure 3.3**: ATLAS observations of CDFS. Reproduced from **?**.

.

Survey - South 1 (ELAIS-S1) fields. These fields are areas of the sky with few nearby objects, meaning that observations in these fields are of very old, distant objects. These fields were chosen because they are the two fields imaged in the Spitzer Wide-area Infrared Extragalactic Survey (SWIRE) visible from the southern hemisphere. SWIRE produced high-resolution infrared and optical images of the fields, allowing all objects detected in the ATLAS radio images to be compared with their infrared and optical counterparts.

ATLAS is considered a pilot survey for the Evolutionary Map of the Universe (EMU), an upcoming radio survey of the entire southern sky at resolutions 45 times higher and angular resolutions 4.5 times better than the benchmark NRAO VLA Sky Survey (**?**). EMU will image the same radio frequencies as ATLAS at the same resolutions, so tools and methods developed to process and interpret ATLAS data are expected to work well on the data produced by EMU. EMU is expected to detect around 70 million radio objects, compared to the 2.5 million currently known (**?**).

ATLAS provides both a catalogue of detected radio objects and a radio image of the CDFS and ELAIS-S1 fields. The CDFS image covers a total area of 3.7 $\text{deg}^2$ and the ELAIS-S1 image covers a total area of 2.7 $\text{deg}^2$. The CDFS image is shown in Figure 3.3. The catalogue is a list of all objects detected in the images with a peak or integrated flux more than 5 times the background noise levels. Each object has an associated survey identifier, an International Astronomical Union name, a position on the sky of the peak flux, a peak flux density, an integrated flux density, an angular size, whether the object is extended or compact, and a spectral index, as well as

uncertainties associated with each measurement (**?**).

1. something about ATLAS, FIRST, EMU, Meerkat–MIGHTEE, WODAN
2. something about SWIRE, WISE

Chapter 4

# Machine Learning

1. something about machine learning, classification, regression
2. something about active learning, pool-based active learning

Chapter 5

# Radio Cross-identification

In this chapter, I develop a machine learning approach to the radio cross-identification problem. First, I will discuss different ways to train and use a classifier for the task, in particular framing the cross-identification problem as an object localisation problem. Then, I will discuss the available training data and how I chose to process it. Finally, I will present results against a dataset of expert labels, and compare these results to those found by other methods.

## 5.1 Cross-identification as Binary Classification

Given a radio object, we want to locate the host galaxy containing the AGN emitting that object. In general, there may be multiple hosts associated with one radio object (such as in Figure 5.1), but we make the assumption that there is only one. This is the same assumption made by Radio Galaxy Zoo TODO: *cite:rgz-analysis-github(?)*.

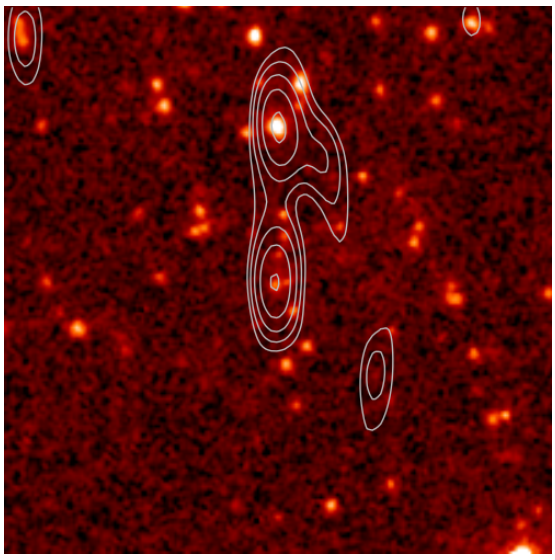

**Figure 5.1:** A radio object (ARG0003ra1) with two host galaxies. This radio object is actually two radio objects that have been incorrectly detected as one, and there is one host galaxy for each object.

We can interpret this as an object localisation problem. As input, we have an image of the radio sky, and we want to locate a host galaxy in this image. A common way to find an object in an image is by using a sliding window. A fixed-size image patch centred on each pixel is taken as a feature representation of that pixel. This is then used as input to a classification model which outputs a probability for each pixel, with higher probabilities corresponding to higher likelihood of the object being located at that pixel. The pixel with the highest rating is considered the location of the object. This approach can be improved by intelligently selecting candidate pixels and only testing these. For the cross-identification problem, we can use galaxy locations as candidate pixels, with galaxies found in infrared surveys such as WISE and SWIRE. Additionally, astronomical measurements such as flux may be taken as additional features for each candidate pixel, giving more information to the classifier.

This approach can be formalised as follows. Consider a set $\mathcal{X}$ of candidate host galaxies, and a radio object $r$ that we want to assign a host galaxy. Let $y : \mathcal{X} \to \{0, 1\}$ represent whether a given $x \in \mathcal{X}$ is the host galaxy associated with $r$. If we assume that a radio object has exactly one associated host galaxy, then there exists exactly one $x \in \mathcal{X}$ such that $y(x) = 1$, and for all other $x \in \mathcal{X}$, $y(x) = 0$. The cross-identification task then amounts to modelling $p(y(x) = 1 \mid x, r)$. Once this distribution is modelled, the host galaxy associated with $r$ is given by

$$\text{host}(r) = \underset{x}{\text{argmax}}\ p(y(x) = 1 \mid x, r). \tag{5.1}$$

Ideally, $\mathcal{X}$ is the set of all galaxies. This is clearly intractable, so as an approximation we use a catalogue of infrared objects near the radio object of interest, taken from an infrared survey. We also make the assumption that the host galaxy is within $1'$ of the radio object — while this doesn't hold in general, systems larger than $1'$ are rare and require human insight to discover (**?**).

## 5.2 Data Sources

To learn the cross-identification distribution (Equation 5.1), we need a set of radio objects, a set of candidate host galaxies, and a set of existing cross-identifications for training. I have chosen to use the ATLAS radio survey for radio objects and the WISE survey for candidate host galaxies. The training cross-identifications are based on the Radio Galaxy Zoo.

### 5.2.1 Radio Data

The ATLAS survey (**?**) TODO: *talk about this in the astro chapter* provides both a catalogue of detected radio objects and a radio image of the CDFS and ELAIS-S1 fields.

The catalogue contains around 2400 objects in the CDFS field, all of which have been labelled by expert astronomers (**?**), by prior algorithms for automated cross-identification (**?**), and by Radio Galaxy Zoo volunteers (**?**). The ELAIS-S1 field does not have Radio Galaxy Zoo classifications as it was not included by the Radio Galaxy

Zoo researchers. The existence of expert, algorithmic, and crowdsourced labels makes ATLAS-CDFS an excellent set of radio objects for developing and testing machine learning approaches to cross-identification. Additionally, ATLAS is considered a pilot for the upcoming Evolutionary Map of the Universe survey, which will provide the majority of radio objects needing cross-identification in the future (**?**).

TODO: *talk about ATLAS image in the astro chapter*

The catalogue contains the name and position of each radio object, as well as various metadata such as the noise level and the flux. I have chosen to ignore the metadata and only use the position information. The positions are used to extract a $5' \times 5'$ from the full ATLAS image of the CDFS field, centred on each radio object. These image patches represent the radio object in the object localisation problem.

### 5.2.2 Infrared Data

Both the SWIRE and WISE surveys have imaged the CDFS field in infrared wavelengths. Of these, WISE is more commonly used, so I have chosen to use WISE as a source of candidate host galaxies. Like ATLAS, WISE provides both a catalogue of detected objects as well as images.

The WISE catalogue contains names, positions, and astronomical measurements for all detected infrared objects. I used the positions to generate features for each candidate host, and the raw flux measurements as additional features. Feature generation is described in Section 5.3. I chose to use the fluxes as they have the most physical meaning TODO: *citation needed*.

I have chosen to not use image data from WISE. Preliminary experiments showed that there was not much information gained from the use of infrared images, and using infrared images made the classifier more complex. Future work may use infrared images to improve the classifier.

### 5.2.3 Radio Galaxy Zoo Consensus Labels

TODO: *move lots of this to the astro section*

Radio Galaxy Zoo volunteers are first asked to select combinations of radio objects that correspond to one radio source, and are then asked to select the location of the corresponding host galaxy (**?**). Each radio subject is labelled by multiple volunteers. These labels are then collated as follows. First, the most common combination of radio objects is selected, and all labels that have a different combination are discarded. This radio combination is called the consensus radio combination. Then, the density of host location labels is estimated using Gaussian kernel density estimation (KDE), and the highest density location is selected. This is called the consensus host location. The consensus host location is then matched to the nearest infrared object.

An alternative way to find the consensus host location is by using a clustering algorithm such as $k$-means. Host locations are clustered and the cluster containing the most locations is taken to represent the consensus; the consensus location is then the mean of the cluster. This is faster and more robust than using KDE, but requires

$k$ to be known. $k$ can be estimated using an algorithm such as PG-means (**?**) or by choosing $k$ to minimise some information criterion TODO: *cite:sklearn*. The consensus labels for the data associated with this thesis were found in this way, fitting a Gaussian mixture model to the host locations with the number of Gaussians chosen to minimise the Bayesian information criterion.

Repeated volunteer labelling helps to reduce noise in the labels. This is necessary as the volunteers are not experts, and may incorrectly label the subject. The hope is that the majority of volunteers will correctly label subjects, which seems to be the case for radio subjects where more than 75% of volunteers agree (**?**). The number of times a radio subject is shown to different volunteers is called the redundancy. The redundancy is 5 if the subject is a compact source, and 20 for all other sources. These numbers were chosen based on the redundancy levels of the original Galaxy Zoo project [Banfield, personal communication] TODO: *Is this how to cite personal communication? Alternatively, is this written down somewhere?*. Since labels with radio combinations that disagree with the consensus are discarded, the redundancy is usually lower in practice when finding the host location. This can lead to very low redundancy input to KDE, causing KDE to fail. This failure can usually be caught, but the existing solution in this case is to take the mean of all host locations. This is not the consensus host location in general. Another effect is that since more complex sources have higher levels of disagreement in the radio combination stage, more complex sources have more discarded volunteer labels, and thus lower redundancy — so more complex sources have more noise in their labels.

The training data are generated by assigning each infrared object a binary label: 0 if the object is not associated with a radio object, and 1 if the object is associated with a radio object.

### 5.2.4 Norris Labels

## 5.3 Features

Each candidate potential host galaxy has an associated set of features. Some of these features are extracted from the radio image patch centred on the galaxy, and some of these features come directly from the infrared survey.

### 5.3.1 Astronomical Features

### 5.3.2 Radio Image Features

## 5.4 Choosing a Binary Classifier

## 5.5 Results

### 5.5.1 Classifying Galaxies

TODO: *Make this writing considerably less terrible.*

The infrared objects were partitioned into a testing set and a training set as follows. First, the radio objects were partitioned into a radio testing set and a radio training set, with the radio testing set containing 452 objects and the radio training set containing 1811 objects. Infrared objects were then added to the infrared testing set if they were within $1'$ of a radio object in the radio testing set, and added to the infrared training set otherwise. This was done because infrared objects that were close together had overlapping radio features, and thus random partitioning would result in the training set containing many of the features found in the testing set. The partitioning resulted in 5922 objects in the testing set and 18218 objects in the training set.

For each infrared object, features and labels were generated as described in Sections 5.3 and 5.2.3, respectively. Note that the labels were sourced from the Radio Galaxy Zoo. A logistic regression classifier was then trained on the training set using binary cross-entropy loss, with loss of each data point weighted based on the frequency of its label.

The classifier was then used to classify the objects in the testing set. The labels were compared to those found by **?** as described in Section 5.2.4, resulting in 80.14% balanced accuracy.

TODO: *ROC/precision–recall curves*

Chapter 6

# Active Learning

In this chapter I'll talk about applying active learning to the RGZ data.

Chapter 7

# Conclusion

TODO: *Write a conclusion chapter.*

TODO: *Verify that I didn't break anything by using natbib.*